CONTENTS

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Course File

21CS644 Data Science and Visualization

VI Sem A 2023-24

Faculty In-charge
KAVITHA K S
Assistant Professor
Dept. of Computer science and Engineering
K S School of Engineering & Management, Bangalore

K. S. SCHOOL OF ENGINEERING AND MANAGEMENT

VISION

To impart quality education in engineering and management to meet technological, business and societal needs through holistic education and research.

MISSION

K.S. School of Engineering and Management shall,

- Establish state-of-art infrastructure to facilitate effective dissemination of technical and Managerial knowledge.
- Provide comprehensive educational experience through a combination of curricular and Experiential learning, strengthened by industry-institute-interaction.
- Pursue socially relevant research and disseminate knowledge.
- Inculcate leadership skills and foster entrepreneurial spirit among students.

Department of Computer Science and Engineering

VISION

To produce quality Computer Science professional, possessing excellent technical knowledge, skills, personality through education and research.

MISSION

Department of Computer Science and Engineering shall,

- Provide good infrastructure and facilitate learning to become competent engineers who meet global challenges.
- Encourages industry institute interaction to give an edge to the students.
- Facilitates experimental learning through interdisciplinary projects.
- Strengthen soft skill to address global challenges.

VI Semester

DATA SCIENCE AND VISUALIZATION			
Course Code	21CS644	CIE Marks	50
Teaching Hours/Week (L:T:P: S)	3:0:0:0	SEE Marks	50
Total Hours of Pedagogy	40	Total Marks	100
Credits	03	Exam Hours	03

Course Learning Objectives

CLO 1. To introduce data collection and pre-processing techniques for data science
CLO 2. Explore analytical methods for solving real life problems through data exploration techniques
CLO 3. Illustrate different types of data and its visualization
CLO 4. Find different data visualization techniques and tools
CLO 5. Design and map element of visualization well to perceive information

Teaching-Learning Process (General Instructions)

These are sample Strategies, which teachers can use to accelerate the attainment of the various course outcomes.

1. Lecturer method (L) need not to be only a traditional lecture method, but alternative effective teaching methods could be adopted to attain the outcomes.
2. Use of Video/Animation to explain functioning of various concepts.
3. Encourage collaborative (Group Learning) Learning in the class.
4. Ask at least three HOT (Higher order Thinking) questions in the class, which promotes critical thinking.
5. Adopt Problem Based Learning (PBL), which fosters students' Analytical skills, develop design thinking skills such as the ability to design, evaluate, generalize, and analyze information rather than simply recall it.
6. Introduce Topics in manifold representations.
7. Show the different ways to solve the same problem with different circuits/logic and encourage the students to come up with their own creative ways to solve them.
8. Discuss how every concept can be applied to the real world - and when that's possible, it helps improve the students' understanding.

Module-1

Introduction to Data Science

Introduction: What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? – Datafication, Current landscape of perspectives, Skill sets. Needed Statistical Inference: Populations and samples, Statistical modelling, probability distributions, fitting a model.

Textbook 1: Chapter 1

Teaching-Learning Process	1. PPT – Recognizing different types of data, Data science process 2. Demonstration of different steps, learning definition and relation with data science

Module-2

Exploratory Data Analysis and the Data Science Process

Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA, The Data Science Process, Case Study: Real Direct (online realestate firm). Three Basic Machine Learning Algorithms: Linear Regression, k-Nearest Neighbours (k- NN), k-means.

Textbook 1: Chapter 2, Chapter 3

Teaching-Learning Process	1. PPT –Plots, Graphs, Summary Statistics 2. Demonstration of Machine Learning Algorithms

Module-3	
Feature Generation and Feature Selection Extracting Meaning from Data: Motivating application: user (customer) retention. Feature Generation (brainstorming, role of domain expertise, and place for imagination), Feature Selection algorithms. Filters; Wrappers; Decision Trees; Random Forests. Recommendation Systems: Building a User-Facing Data Product, Algorithmic ingredients of a Recommendation Engine, Dimensionality Reduction, Singular Value Decomposition, Principal Component Analysis, Exercise: build your own recommendation system. **Textbook 1: Chapter 6**	
Teaching-Learning Process	1. PPT – Feature generation, selection 2. Demonstration recommendation engine
Module-4	
Data Visualization and Data Exploration **Introduction:** Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization **Comparison Plots:** Line Chart, Bar Chart and Radar Chart; **Relation Plots:** Scatter Plot, Bubble Plot , Correlogram and Heatmap; **Composition Plots:** Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram; **Distribution Plots:** Histogram, Density Plot, Box Plot, Violin Plot; **Geo Plots:** Dot Map, Choropleth Map, Connection Map; What Makes a Good Visualization? **Textbook 2: Chapter 1, Chapter 2**	
Teaching-Learning Process	1. Demonstration of different data visualization tools.
Module-5	
A Deep Dive into Matplotlib Introduction, Overview of Plots in Matplotlib, **Pyplot Basics:** Creating Figures, Closing Figures, Format Strings, Plotting, Plotting Using pandas DataFrames, Displaying Figures, Saving Figures; **Basic Text and Legend Functions:** Labels, Titles, Text, Annotations, Legends; **Basic Plots:**Bar Chart, Pie Chart, Stacked Bar Chart, Stacked Area Chart, Histogram, Box Plot, Scatter Plot, Bubble Plot; **Layouts:** Subplots, Tight Layout, Radar Charts, GridSpec; **Images:** Basic Image Operations, Writing Mathematical Expressions **Textbook 2: Chapter 3**	
Teaching-Learning Process	1. PPT – Comparison of plots 2. Demonstration charts
Course Outcomes At the end of the course the student will be able to: CO 1. Understand the data in different forms CO 2. Apply different techniques to Explore Data Analysis and the Data Science Process CO 3. Analyze feature selection algorithms & design a recommender system. CO 4. Evaluate data visualization tools and libraries and plot graphs. CO 5. Develop different charts and include mathematical expressions.	
Assessment Details (both CIE and SEE) The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 40% of the maximum marks (20 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures not less than 35% (18 Marks out of 50) in the semester-end examination (SEE), and a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together **Continuous Internal Evaluation:** Three Unit Tests each of **20 Marks (duration 01 hour)**	

1. First test at the end of 5th week of the semester
2. Second test at the end of the 10th week of the semester
3. Third test at the end of the 15th week of the semester

Two assignments each of **10 Marks**

4. First assignment at the end of 4th week of the semester
5. Second assignment at the end of 9th week of the semester

Group discussion/Seminar/quiz any one of three suitably planned to attain the COs and POs for **20 Marks (duration 01 hours)**

6. At the end of the 13th week of the semester

The sum of three tests, two assignments, and quiz/seminar/group discussion will be out of 100 marks and will be **scaled down to 50 marks**

(to have less stressed CIE, the portion of the syllabus should not be common /repeated for any of the methods of the CIE. Each method of CIE should have a different syllabus portion of the course).

CIE methods /question paper has to be designed to attain the different levels of Bloom's taxonomy as per the outcome defined for the course.

Semester End Examination:

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the subject **(duration 03 hours)**

1. The question paper will have ten questions. Each question is set for 20 marks. Marks scored shall be proportionally reduced to 50 marks
2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), **should have a mix of topics** under that module.

The students have to answer 5 full questions, selecting one full question from each module

Suggested Learning Resources:

Textbooks

1. Doing Data Science, Cathy O'Neil and Rachel Schutt, O'Reilly Media, Inc O'Reilly Media, Inc, 2013
2. Data Visualization workshop, Tim Grobmann and Mario Dobler, Packt Publishing, ISBN 9781800568112

Reference:

1. Mining of Massive Datasets, Anand Rajaraman and Jeffrey D. Ullman, Cambridge University Press, 2010
2. Data Science from Scratch, Joel Grus, Shroff Publisher /O'Reilly Publisher Media
3. A handbook for data driven design by Andy krik

Weblinks and Video Lectures (e-Resources):

1. https://nptel.ac.in/courses/106/105/106105077/
2. https://www.oreilly.com/library/view/doing-data-science/9781449363871/toc01.html
3. http://book.visualisingdata.com/
4. https://matplotlib.org/
5. https://docs.python.org/3/tutorial/
6. https://www.tableau.com/

Activity Based Learning (Suggested Activities in Class)/ Practical Based learning

Demonstration using projects

K. S. SCHOOL OF ENGINEERING AND MANAGEMENT
BENGALURU-560109
TENTATIVE CALENDAR OF EVENTS: VI EVEN SEMESTER (2023-2024)
SESSION: APRIL TO JULY 2024

Week No.	Month	Day						Days	Activities
		Mon	Tue	Wed	Thu	Fri	Sat		
1	APR/ MAY	**29***	30	**1 H**	2	3	**4 DH**	4	**29* -Commencement of VI sem** 1- May Day
2	MAY	6	7	8	9	**10 H**	**11**	5	10 - Basava Jayanthi 11- Friday Time Table
3	MAY	13	14	15	16	17 TA	**18 DH**	5	
4	MAY	20	21	22	23	24	25TA	6	25- Monday Time Table
	MAY/ JUNE	**27 T1**	**28 T1**	**29 T1**	30	31	**1 DH**	5	29 - First Faculty Feed Back
6	JUNE	3 BV	4ASD	**5* FFB1**	6	7	8	6	8- Monday Time Table
7	JUNE	10	11	12	13	14	**15 DH**	5	
8	JUNE	**17 H**	18	19	20	21	22	5	17- Bakrid 22- Wednesday Time Table
9	JUNE	24	25	26TA	**27T2**	**28T2**	**29T2**	6	
10	JULY	1BV	2ASD	3* FFB2	4	5	**6 DH**	5	3* - First Faculty Feed Back
1	JULY	8	9	10	11	12	13	6	13- Friday Time Table
12	JULY	15	16	**17 H**	18	19TA	**20 DH**	4	17- Last Day of Moharam
13	JULY	**22T3**	**23T3**	**24T3**	**25LT**	**26LT**	**27LT**	6	
14	JULY	**29LT**	30	**31***				3	**31* - Last Working day**
Total No of Working Days : 71									

Total Number of working days (Excluding holidays and Tests)=58

H	Holiday
BV	Blue Book Verification
T1,T2,T3	Tests 1,2,3
ASD	Attendance & Sessional Display
DH	Declared Holiday
LT	Lab Test
TA	Test attendance

Monday	12
Tuesday	12
Wednesday	11
Thursday	11
Friday	12
Total	**58**

SIGNATURE OF PRINCIPAL
Dr. K. RAMA NARASIMHA
Principal/Director
K S School of Engineering and Management
Bengaluru - 560 109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BENGALURU-560109

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SESSION: 2023-2024(EVEN SEMESTER)

(w. e. f : 22/4/2024)

INDIVIDUAL TIME TABLE

Class: IV 'A & B' &' VI A ' **Faculty Name: Mrs. Kavitha K S**

DAY	8.40-9.35	9.35-10.30	10.30 -10.45	10.45 -11.40	11.40-12.35	12.35-1.20	1.20 -2.10	2.10-3.00	3.00-3.50
MONDAY	Analysis & Design of Algorithms Lab Batch - A2					LUNCH BREAK			DSV (VI A)
TUESDAY	Analysis & Design of Algorithms Lab Batch - A1							DSV (VI A)	
WEDNESDAY	Analysis & Design of Algorithms Lab Batch - B2				DSV (VI A)				
THURSDAY	Analysis & Design of Algorithms Lab Batch - B1								
FRIDAY			Tea Break	DSV (VI A)				Sports	
SATURDAY	AS PER CALENDAR OF EVENTS								

CODE	SUBJECT	Hours /Week	
21CS644	Data science and Visualization	4	Mrs. Kavitha K S
BCSL404	Analysis & Design of Algorithms Lab	6	
BPEK459	Physical Education (PE) (Sports and Athletics)	2	
18CSI85	Internship	1	
18CSS84	Technical Seminar	1	
18CSP83	Project Work Phase-II	1.5	

Time-table Coordinator

HOD

Head of the Department

Department of Computer Science & Engineering
K.S. School of Engineering & Management
Bengaluru-560109

Dr. K. RAMANARASIMHA
Principal/Director
K S School of Engineering & Management
Bengaluru - 560 109

KSSEM

K. S. SCHOOL OF ENGINEERING AND MANAGEMENT, B

DEPARTMENT OF COMPUTER SCIENCE AND EN(

SESSION: 2023-2024 (EVEN SEMESTER)

VI Sem A Section Student List

Sl. No	USN	Name of the Student
1	1KG21CS001	ABBURI PALLAVI
2	1KG21CS002	ABHIJEET DAS
3	1KG21CS003	ABHILASH B R
4	1KG21CS004	ABHISHEK V
5	1KG21CS005	AKHILA A
6	1KG21CS006	AKSHATHA R GOWDA
7	1KG21CS007	ALLU CHINNI KRISHNA
8	1KG21CS008	AMITH C SURI
9	1KG21CS009	AMOGH A
10	1KG21CS010	ANKITHA VENKATESH
11	1KG21CS011	ANKUSH GOWDA K
12	1KG21CS012	ARPITHA S
13	1KG21CS013	ASHWINI C
14	1KG21CS014	B NAYANA
15	1KG21CS015	BATTA PREETHI
16	1KG21CS016	BEEGALU SRINIVAS AKHIL
17	1KG21CS017	BHARATH GOWDA J
18	1KG21CS018	BHAVANA D
19	1KG21CS019	BHAVANA S
20	1KG21CS021	C SUSMITHA
21	1KG21CS022	CHALLA HARI KISHORE NAIDU
22	1KG21CS023	CHANDAN TAVANE
23	1KG21CS024	CHINMAIY P
24	1KG21CS026	DARSHAN R

25	1KG21CS027	DEEKSHITHA R
26	1KG21CS028	DEEPAK ATHRESH R
27	1KG21CS029	DEVDAS
28	1KG21CS030	DHAKSHITHA A
29	1KG21CS031	DHANUSH G P
30	1KG21CS032	DHANUSH U S
31	1KG21CS033	DHARINI
32	1KG21CS034	DHEERAJ D RAIKAR
33	1KG21CS035	DIBYAJYOTI SAHU
34	1KG21CS036	DINESH J L
35	1KG21CS037	DIVYA H U
36	1KG21CS038	DIVYA P
37	1KG21CS039	GEOFFREY SAMUEL
38	1KG21CS040	GONGATI RAGHU
39	1KG21CS041	GORANTLA DIVYA SREE
40	1KG21CS042	GURUJALA BHARATH
41	1KG21CS043	GURURAJ B
42	1KG21CS044	HANOCH CHRISTIAN R
43	1KG21CS045	HARSHITH K P
44	1KG21CS046	HARSHITHA D G
45	1KG21CS047	HITESH A REDDY
46	1KG21CS048	JAHANAVI S
47	1KG21CS050	K J PRAKRUTHI
48	1KG21CS051	K NITHISH
49	1KG21CS052	K ROHITH
50	1KG21CS053	KARTHIK G
51	1KG21CS054	KASTURI POORNIMA CHOWDARY
52	1KG21CS055	KISHOR KUMAR L
53	1KG21CS056	KUNALA GANESH

54	1KG21CS057	KUSHMITHA T A
55	1KG21CS058	KUSUMA B
56	1KG21CS059	L LAVANYA
57	1KG21CS060	LAKSHA SENTHILKUMAR
58	1KG21CS061	M POOJA
59	1KG21CS062	M SURABHI
60	1KG21CS063	MADINENI BHUVANA
61	1KG22CS400	AKSHAY U
62	1KG22CS401	BALAJI N
63	1KG22CS402	BHAVANA M
64	1KG22CS403	DHANUSH R
65	1KG22CS404	DHANUSHREE A
66	1KG22CS405	KIRAN KUMAR

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BENGALURU - 560109

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024 (EVEN SEMESTER)

LESSON PLAN

NAME OF THE STAFF: Mrs. KAVITHA K S

SUBJECT CODE/NAME: 21CS644 / Data Science & Visualization

SEMESTER/SEC/YEAR: VI/ B /III

ACADEMIC YEAR: 2023-2024

Sl. No.	Topic to be covered	Mode of Delivery	Teaching Aid	No. of Periods	Cumulative No. of Periods	Proposed Date	Engaged Date
	MODULE 1						
1	Introduction: What is Data Science? Big Data and Data Science hype – and getting past the hype	L+ D	BB+LCD	1	1	29/04/2024	13/5/24
2	Why now? – Datafication	L+ D	BB+LCD	1	2	30/04/2024	14/5/24
3	Current landscape of perspectives	L+ D	BB+LCD	1	3	03/05/2024	15/5/24
4	Skill sets	L+D	BB+LCD	1	4	06/05/2024	16/5/24
5	Needed Statistical Inference: Populations and samples	L+ D	BB+LCD	1	5	07/05/2024	17/5/24
6	Statistical modelling	L+D	BB+LCD	1	6	08/05/2024	22/5/24
7	Probability distributions	L+ D	BB+LCD	1	7	11/05/2024	24/5/24
8	Fitting a model	L+ D	BB+LCD	1	8	13/05/2024	25/5/24
9	Tutorial	L+D	BB+LCD	1	-	14/05/2024	27/5/24
	MODULE 2						
10	Basic tools (plots, graphs and summary statistics) of EDA	L+D	BB+LCD	1	9	15/05/2024	28/5/24
11	Basic tools continued	L+D+PS	BB+LCD	1	10	17/05/2024	3/6/24

12	Philosophy of EDA	L+D	BB+LCD	1	11	20/05/2024	4/6/24
13	The Data Science Process	L+D+PS	BB+LCD	1	12	21/05/2024	5/6/24
14	Case Study: Real Direct (online real estate firm)	L+D+PS	BB+LCD	1	13	22/05/2024	7/6/24
15	Three Basic Machine Learning Algorithms: Linear Regression	L+D	BB+LCD	1	14	24/05/2024	8/6/24
16	k-Nearest Neighbours (k- NN)	L+D	BB+LCD	1	15	25/05/2024	10/6/24
17	k-means	L+D	BB+LCD	1	16	31/05/2024	11/6/24
18	Tutorial	L+D+PS	BB+LCD	1	-	03/06/2024	12/6/24
MODULE 3							
19	Extracting Meaning from Data: Motivating application: user (customer) retention	L+D	BB+LCD	1	17	04/06/2024	14/6/24
20	Feature Generation (brainstorming, role of domain expertise, and place for imagination	L+ D	BB+LCD	1	18	05/06/2024	18/6/24
21	Feature Selection algorithms.	L+D+PS	BB+LCD	1	19	07/06/2024	19/6/24
22	Filters; Wrappers; Decision Trees; Random Forests.	L+D+PS	BB+LCD	1	20	08/06/2024	21/6/24
23	Recommendation Systems: Building a User-Facing Data Product	L+D	BB+LCD	1	21	10/06/2024	24/6/24
24	Algorithmic ingredients of a Recommendation Engine, Dimensionality Reduction	L+D	BB+ LCD	1	22	11/06/2024	25/6/24
25	Singular Value Decomposition, Principal Component Analysis,	L+D	BB+ LCD	1	23	12/06/2024	26/6/24
26	Exercise: build your own recommendation system.	L+D	BB+LCD	1	24	14/06/2024	28/6/24
27	Tutorial	L+D+PS	BB+LCD	1	-	18/06/2024	29/6/24
MODULE 4							
28	Introduction: Data Visualization, Importance of Data Visualization,	L+D	BB+LCD	1	25	19/06/2024	2/7/24
29	Data Wrangling, Tools and Libraries for Visualization	L+D+PS	BB+LCD	1	26	21/06/2024	2/7/24

30	Comparison Plots: Line Chart, Bar Chart and Radar Chart;	L+D	BB+LCD	1	27	22/06/2024	3/7/24
31	Relation Plots: Scatter Plot, Bubble Plot, Correlogram and Heatmap	L+D+PS	BB+LCD	1	28	24/06/2024	5/7/24
32	Composition Plots: Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram;	L+D+PS	BB+LCD	1	29	25/06/2024	10/7/24
33	Distribution Plots: Histogram, Density Plot, Box Plot, Violin Plot;	L+D+PS	BB+LCD	1	30	26/06/2024	12/7/24
34	Geo Plots: Dot Map, Choropleth Map, Connection Map;	L+D+PS	BB+LCD	1	31	01/07/2024	13/7/24
35	What Makes a Good Visualization?	L+D+PS	BB+LCD	1	32	02/07/2024	15/7/24
36	Tutorial	L+D+PS	BB+LCD	1	-	03/07/2024	15/7/24
	MODULE 5						
37	Introduction, Overview of Plots in Matplotlib,	L+D	BB+LCD	1	33	05/07/2024	16/7/24
38	Pyplot Basics: Creating Figures, Closing Figures, Format Strings, Plotting,	L+D+PS	BB+LCD	1	34	08/07/2024	19/7/24
39	Plotting Using pandas DataFrames, Displaying Figures, Saving Figures	L+D	BB+LCD	1	35	09/07/2024	22/7/24
40	Basic Text and Legend Functions: Labels, Titles, Text, Annotations, Legends;	L+D	BB+LCD	1	36	10/07/2024	22/7/24
41	Basic Plots: Bar Chart, Pie Chart, Stacked Bar Chart, Stacked Area Chart,	L+D+PS	BB+LCD	1	37	12/07/2024	23/7/24
42	Histogram, Box Plot, Scatter Plot, Bubble Plot;	L+D	BB+LCD	1	38	13/07/2024	23/7/24
43	Layouts: Subplots, Tight Layout, Radar Charts, GridSpec;	L+D+PS	BB+LCD	1	39	15/07/2024	24/7/24
44	Images: Basic Image Operations, Writing Mathematical Expressions	L+D	BB+LCD	1	40	16/07/2024	27/7/24
45	Tutorial	L+D	BB+LCD	1	40	19/07/2024	30/7/24
46	Revision	L+D+PS	BB+LCD	10	40	30/07/2024 31/07/2024	—

	Week	Remarks
Assignment 1	4th Week - 24/05/2024	Mode of Assignment – Written Assignment
Assignment 2	9th Week- 29/06/2024	

Total No. of Lecture Hours = 40
Total No. of Tutorial Hours =07

Course In charge **IQAC Coordinator** **HOD CSE** **Principal**

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BENGALURU - 560109

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024(EVEN SEMESTER)

Data Science & Visualization (21CS644)

Question Bank-1

Module -1

1. Define Data Science. Explain the Drew Conway's Venn diagram of data science.
2. Illustrate the Data science profile and explain the work of Data Scientist in academia and industry.
3. Explain the details about Big data and Data Science hype? Why now?
4. Make use of the concept of Datafication explain with examples.
5. Explain the following concepts with examples

   i) Statistical inference ii) Population iii) Samples

   iv) Types of data v) Random Variable
6. Discuss fitting of a model. Explain overfitting with example.
7. Explain Data Scientist from the perspective of Academia and Industry.
8. Illustrate Probability Distribution.
9. Discuss fitting of a model. What is overfitting?
10. List the 3 factors of Big Data Revolution. Discuss N=ALL and n=1 assumptions with examples.

Module -2

1. Interpret Exploratory Data Analysis with example.
2. Draw and explain the Data Science Process and also discuss Data Scientist role in this process with suitable diagram.
3. **Explain** the connection of Data Science to scientific method.
4. In the case study of RealDirect, **explain** how RealDirect makes money.

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BENGALURU - 560109

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024(EVEN SEMESTER)

Data Science & Visualization (21CS644)

Question Bank-2

Module -2

1. Compare clustering and classification.
2. Give an example to explain k-NN.
3. Give an example to explain the K-means algorithm. Also give the advantages and disadvantages.
4. Discuss the advantages and issues in k-means algorithm.
5. Demonstrate Linear Regression with an example.
6. In the case study of RealDirect, demonstrate how RealDirect makes money.

Module -3

1. List and explain the problems with Nearest Neighbors.
2. What is crowdsourcing? Describe the Kaggle Model.
3. List the Feature selection methods. Explain each in detail.
4. What is Dimensionality Problem? Illustrate Principal Component Analysis.
5. What is Dimensionality Reduction? Illustrate Singular Value Decomposition(SVD).
6. Define bagging. Explain Random Forest with an example diagram.
7. Explain decision tree for chasing dragons. Develop a R script for the same.
8. Demonstrate the Feature selection with Chasing Dragons User Retention example.

Course In charge

Head of the Department

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109

DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING

SESSION: 2023-2024 (EVEN SEMESTER)

DATA SCIENCE AND VISUALIZATION (21CS644)

Question Bank-3

Module-4&5

1. **Define** data wrangling. **Explain** With a neat diagram the steps involved in the data wrangling process.
2. **Explain** the overview of Plots in Matplotlib and also discuss the Anatomy of a Matplotlib Figure.
3. **Discuss** scatter plots for single group, multiple groups and also explain scatter plot with marginal histograms with a neat diagram.
4. **Explain** the following with a matplotlib script and suitable diagram:
   i. Annotations ii. Legends
5. **Interpret** how the Density plot is different from Histogram plots. **Compare** Box plots and Violin plots with a neat suitable diagrams.
6. Explain the following with syntax, plots and **develop** a python code for
   i) Bar chart(for single category and subcategories)
   ii) Pie chart(for water usage) iii) Histogram iv)2D Histogram
7. **Draw** an example graph to discuss various Geo plots.
8. **Develop** a python code to demonstrate Subplots, Tight layouts and Gridspec.
9. **Explain** the data wrangling process with an example of employee engagement.
10. Briefly **explain** the operations of creating figures and closing figures performed using pyplot.
11. **Discuss** Heatmaps and its variant in detail.
12. With an example for each **explain** the following:
    i.Labels ii. Titles
13. **Determine** the differences between Stacked Bar Charts and Stacked Area Charts.
14. **Demonstrate** the following with an example:
    i.Scatter Plot ii. Bubble Plot
15. **Draw** an example graph for each of the violin plots with single variable, multiple variables and multiple categories.
16. What are the basic operations for designing an image? **Illustrate** Loading Images and Saving Images.

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BENGALURU - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CO-PO Mapping

Course: DATA SCIENCE AND VISUALIZATION				
Type: Professional Elective Course			Course Code: 21CS644	
No of Hours				
Theory (Lecture Class)	Tutorials	Practical/Field Work/Allied Activities	Total/Week	Total hours of Pedagogy
3	0	0	3	40
Marks				
CIE	SEE		Total	Credits
50	50		100	3

Aim/Objectives of the Course

1. To introduce data collection and pre-processing techniques for data science
2. Explore analytical methods for solving real life problems through data exploration techniques
3. Illustrate different types of data and its visualization
4. Find different data visualization techniques and tools
5. Design and map element of visualization well to perceive information

Course Learning Outcomes
At the end of the course the student will be able to:

CO1	**Explore** the data in different forms and pre-processing techniques for data science.	**Applying (K3)**
CO2	**Apply** different techniques to Explore Data Analysis and the Data Science Process	**Applying (K3)**
CO3	**Interpret** feature selection algorithms & design a recommender system.	**Applying (K3)**
CO4	**Make use of** data visualization tools and libraries and plot graphs.	**Applying (K3)**
CO5	**Develop** different charts and include mathematical expressions	**Applying (K3)**

Syllabus Content

MODULE 1 : Introduction to Data Science: Introduction: What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? – Datafication, Current landscape of perspectives, Skill sets. Needed	**CO1** 8 hrs PO1-3

Statistical Inference: Populations and samples, Statistical modelling, probability distributions, fitting a model. **LO:** At the end of this session the student will be able to 1. Recognizing different types of data, Data science process. 2. Understand the different steps, learning definition and relation with data science.	PO2-3 PO3-3 PO4-2 PO6-1 PO10-2 PO11-1 PO12 -2 PSO1-3 PSO2-3
MODULE 2 : Exploratory Data Analysis and the Data Science Process Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA, The Data Science Process, Case Study: Real Direct(online realestate firm). Three Basic Machine Learning Algorithms: Linear Regression, k-Nearest Neighbours (k- NN), k-means. **LO:** At the end of this session the student will be able to 1. Plots, Graphs, Summary Statistics 2. Understand Machine Learning Algorithms 3. Analyze Case Study: Real Direct	**CO2** 8 hrs. PO1-3 PO2-3 PO3-3 PO4-2 PO5-3 PO6-1 PO10-2 PO11-1 PO12-2 PSO1-3 PSO2-3
MODULE 3: Feature Generation and Feature Selection: Extracting Meaning from Data: Motivating application: user (customer) retention. Feature Generation (brainstorming, role of domain expertise, and place for imagination), Feature Selection algorithms. Filters; Wrappers; Decision Trees; Random Forests. Recommendation Systems: Building a User-Facing Data Product, Algorithmic ingredients of a Recommendation Engine, Dimensionality Reduction, Singular Value Decomposition, Principal Component Analysis, Exercise: build your own recommendation system. **LO:** At the end of this session the student will be able to 1. Understand and Analyze Feature generation, selection 2. Analyze recommendation engine 3. Build their own recommendation engine	**CO3** 8 hrs PO1-3 PO2-3 PO3-3 PO4-2 PO5-3 PO6-1 PO10-2 PO11-1 PO12-2 PSO1-3 PSO2-3
MODULE 4: Data Visualization and Data Exploration **Introduction:** Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization **Comparison Plots:** Line Chart, Bar Chart and Radar Chart; **Relation Plots:** Scatter Plot, Bubble Plot , Correlogram and Heatmap; Composition Plots: Pie	**CO4** 8 hrs PO1-3 PO2-3

Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram; **Distribution Plots:** Histogram, Density Plot, Box Plot, Violin Plot; **Geo Plots:** Dot Map, Choropleth Map, Connection Map; What Makes a Good Visualization? **LO:** At the end of this session the student will be able to 1. Understand the different data visualization tools and libraries. 2. Plot different types of plots.	PO3-3 PO4-2 PO5-3 PO6-1 PO10-2 PO11-1 PO12-2 PSO1-3 PSO2-3
MODULE 5: A Deep Dive into Matplotlib Introduction, Overview of Plots in Matplotlib, **Pyplot Basics:** Creating Figures, Closing Figures, Format Strings, Plotting, Plotting Using pandas DataFrames, Displaying Figures, Saving Figures; **Basic Text and Legend Functions:** Labels, Titles, Text, Annotations, Legends; **Basic Plots:** Bar Chart, Pie Chart, Stacked Bar Chart, Stacked Area Chart, Histogram, Box Plot, Scatter Plot, Bubble Plot; **Layouts:** Subplots, Tight Layout, Radar Charts, GridSpec; **Images:** Basic Image Operations, Writing Mathematical Expressions **LO:** At the end of this session the student will be able to 1. Understand Plots in Matplotlib 2. Develop different charts and include mathematical expression	**CO5** 8hrs PO1-3 PO2-3 PO3-3 PO4-2 PO5-3 PO6-1 PO10-2 PO11-1 PO12-2 PSO1-3 PSO2-3
Text Books 1. Doing Data Science, Cathy O'Neil and Rachel Schutt, O'Reilly Media, Inc O'Reilly Media, Inc, 2013 2. Data Visualization workshop, Tim Grobmann and Mario Dobler, Packt Publishing, ISBN 9781800568112	
Reference Books (specify minimum two foreign authors text books) 1. Mining of Massive Datasets, Anand Rajaraman and Jeffrey D. Ullman, Cambridge University Press, 2010 2. Data Science from Scratch, Joel Grus, Shroff Publisher /O'Reilly Publisher Media 3. A handbook for data driven design by Andy krik	
Useful Websites • https://nptel.ac.in/courses/106/105/106105077/ • https://www.oreilly.com/library/view/doing-data-science/9781449363871/toc01.html	

- http://book.visualisingdata.com/
- https://matplotlib.org/
- https://docs.python.org/3/tutorial/
- https://www.tableau.com/

Teaching and Learning Methods

1. Lecture class: 40 hrs

Assessment Details (both CIE and SEE):

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 40% of the maximum marks (20 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures not less than 35% (18 Marks out of 50) in the semester-end examination (SEE), and a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together

Continuous Internal Evaluation:

Three Unit Tests each of **20 Marks** (duration 01 hour)

1. First test at the end of 5th week of the semester
2. Second test at the end of the 10th week of the semester
3. Third test at the end of the 15th week of the semester

Two assignments each of **10 Marks**

4. First assignment at the end of 4th week of the semester
5. Second assignment at the end of 9th week of the semester Group discussion/Seminar/quiz any one of three suitably planned to attain the COs and POs for 20 Marks (duration 01 hours)
6. At the end of the 13th week of the semester

The sum of three tests, two assignments, and quiz/seminar/group discussion will be out of 100 marks and will be scaled down to 50 marks (to have less stressed CIE, the portion of the syllabus should not be common /repeated for any of the methods of the CIE. Each method of CIE should have a different syllabus portion of the course).

CIE methods /question paper has to be designed to attain the different levels of Bloom's taxonomy as per the outcome defined for the course.

Semester End Examination:

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the subject **(duration 03 hours)**

1. The question paper will have ten questions. Each question is set for 20 marks. Marks scored shall be proportionally reduced to 50 marks

2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), **should have a mix of topics** under that module.

The students have to answer 5 full questions, selecting one full question from each module

CO to PO Mapping

PO1: Science and engineering Knowledge **PO2:** Problem Analysis **PO3:** Design & Development **PO4:**Investigations of Complex Problems **PO5:** Modern Tool Usage **PO6:** Engineer & Society	**PO7:**Environment and Society **PO8:**Ethics **PO9:**Individual & Team Work **PO10:** Communication **PO11:**Project Mgmt. & Finance **PO12:** Lifelong Learning

PSO1: Understand fundamental and advanced concepts in the core areas of Computer Science and Engineering to analyze, design and implement the solutions for the real world problems.

PSO2: Utilize modern technological innovations efficiently in various applications to work towards the betterment of society and solve engineering problems.

CO	PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO 1	PSO 2
21CS 4	K-level														
CO1	K3	3	3	3	2	-	1	-	-	-	2	1	2	3	3
CO2	K3	3	3	3	2	3	1	-	-	-	2	1	2	3	3
CO3	K3	3	3	3	2	3	1	-	-	-	2	1	2	3	3
CO4	K3	3	3	3	2	3	1	-	-	-	2	1	2	3	3
CO5	K3	3	3	3	2	3	1	-	-	-	2	1	2	3	3

Course In charge

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

IQAC Coordinator

Principal

Dr. K. RAMA NARASIMHA
Principal/Director
K S School of Engineering and Management
Bengaluru - 560 109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024 (EVEN SEMESTER)

FIRST ASSIGNMENT

Degree	:	B.E	Semester	:	VI
Branch	:	CSE	Course Code	:	21CS644
Course Title	:	Data Science and Visualization	Max Marks	:	10
Date	:	17/05/2024	Last Date for submission	:	24/05/2024

Q No.	Questions	Marks	K-Level	CO mapping
1	**Explain** the details about Big data and data science hype? Why now? **Define** Data Science. **Explain** the current landscape with a neat diagram.	1	Understanding K2	CO1
2	**Illustrate** the Data Science Profile with suitable diagrams.	1	Applying K3	CO1
3	**Make use of** the concept of Datafication explain with examples.	1	Applying K3	CO1
4	i) **Differentiate** between Big data and Data science. ii) **Explain** Data Scientist from the perspective of Academia and Industry.	1	Understanding K2	CO1
5	**Explain** Statistical thinking in the age of Big Data. **Give an example** to explain the following concepts i) Statistical inference ii) Population iii) Samples iv)Types of data v)Random Variable	1	Applying K3	CO1
6	**Illustrate** Probability Distributions with diagrams.	1	Applying K3	CO1
7	**Discuss** fitting of a model. Explain overfitting with example.	1	Understanding K2	CO1
8	**Interpret** Exploratory Data Analysis with example.	1	Applying K3	CO2

9	**i)Draw** a neat diagram and explain the Data Science Process. **ii) Relate** the role of Data Scientist in the Data Science Process with a neat diagram.	**1**	**Applying K3**	**CO2**
10	**i)Explain** the connection of Data Science to scientific method. **ii)**In the case study of RealDirect, **Explain** how RealDirect makes money.	**1**	**Understanding K2**	**CO2**

Course In charge

HOD

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024 (EVEN SEMESTER)

SECOND ASSIGNMENT

Degree	:	**B.E**	**Semester**	:	**VI**
Branch	:	**CSE**	**Course Code**	:	**21CS644**
Course Title	:	**Data Science and Visualization**	**Max Marks**	:	**10**
Date	:	**29/06/2024**	**Last Date for submission**	:	**07/07/2024**

Q No.	Questions	Marks	K-Level	CO mapping
1	**Demonstrate** Linear Regression with an example.	1	Applying K3	CO2
2	a. **Illustrate** the KNN algorithm with an example. b. **Discuss** the advantages and issues in k-means algorithm.	1	Applying K3	CO2
3	a. **Compare** clustering and classification. b. **Give an example** to **explain** the K-means algorithm.	1	Applying K3	CO2
4	**Explain** the different distance metrics used in k-NN.	1	Understanding K2	CO2
5	a. What is crowdsourcing? **Describe** the Kaggle Model. b. **List** the Feature selection methods. **Explain** each in detail.	1	Applying K3	CO3
6	a. **Demonstrate** the Feature selection with Chasing Dragons User Retention example. b. **Distinguish** between Feature Selection and Feature Extraction.	1	Applying K3	CO3
7	a. **Explain** Decision Tree with an example. **Explain** Entropy. b. **Define** bagging. **Explain** Random Forest with an example diagram.	1	Understanding K2	CO3

8	a. **Write** a note on Recommendation Engine. b. **List** and **explain** the problems with Nearest Neighbors.	1	**Applying K3**	**CO3**
9	What is Dimensionality Problem? **Illustrate** Principal Component Analysis.	1	**Applying K3**	**CO3**
10	**Use** examples to **demonstrate** Singular Value Decomposition.	1	**Applying K3**	**CO3**

Course In charge

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

KSSEM

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESSION: 2021-2022 (EVEN SEMESTER)
I SESSIONAL TEST QUESTION PAPER
SET-A

USN | | | | | | | | | | |

Degree	:	B.E	Semester	:	IV A&B
Branch	:	Computer Science and Engineering	Course Code	:	18CS45
Course Title	:	Object Oriented Concepts	Date	:	06/07/2022
Duration	:	90 Minutes	Max Marks	:	30

Note: Answer ONE full question from each part.

Q No.	Question	Marks	K - Level	CO mapping
	PART-A			
1(a)	**Classify** the differences between C and C++ programming language.	5	Understanding K2	CO1
(b)	**Explain** function prototyping with example.	5	Understanding K2	CO1
(c)	**Explain** copy constructor and **Develop** a c++ program using copy constructor.	5	Applying K3	CO2
	OR			
2(a)	**Explain** 'this' pointer with example.	5	Understanding K2	CO1
(b)	**Demonstrate** scope resolution operator with example.	5	Understanding K2	CO1
(c)	**Classify** the characteristics and different types of constructors. **Develop** a C++ program using parameterized constructor.	5	Applying K3	CO2
	PART-B			
3(a)	**Explain** friend function and friend class with suitable example.	5	Understanding K2	CO1
(b)	**Explain** the concept of passing object as argument to function.	5	Understanding K2	CO1
(c)	**Define** a Constructor. **Develop** C++ program by using default constructor.	5	Applying K3	CO2
	OR			
4(a)	**Explain** inline member functions and constant member functions with suitable example.	5	Understanding K2	CO1
(b)	**Explain** mutable and static data member with suitable example.	5	Understanding K2	CO1
(c)	**Explain** array of objects. **Develop** a C++ program using array of objects.	5	Applying K3	CO2

Course Incharge **HOD CSE** **IQAC- Coordinator** **Principal**

HOD
Dept. of Computer Science & Engineering
K.S. School of Engineering & Management
Bangalore-560 062

DR. K. RAMA NARASIMHA
Principal/Director
K. S. School of Engineering and Management
Bengaluru - 560 109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESSION: 2023-2024 (EVEN SEMESTER)
I SESSIONAL TEST QUESTION PAPER
SET- B

USN | | | | | | | | | | |

Degree	:	**B.E**	**Semester**	:	**VI A&B**
Branch	:	**Computer Science and Engineering**	**Course Code**	:	**21CS644**
Course Title	:	**Data Science and Visualization**	**Date**	:	**30/05/2024**
Duration	:	**60 Minutes**	**Max Marks**	:	**20**

Note: Answer ONE full question from each part.

Q No.	Question	Marks	K - Level	CO mapping
	PART-A			
1(a)	**Define** Data Science. **Explain** the current landscape with a neat diagram.	5	Understanding K2	CO1
(b)	**Give examples** to **explain** the concept of Datafication.	5	Applying K3	CO1
	OR			
2(a)	**Explain** Data Scientist from the perspective of Academia and Industry.	5	Understanding K2	CO1
(b)	**Illustrate** Probability Distribution.	5	Applying K3	CO1
	PART-B			
3(a)	**Discuss** fitting of a model. What is overfitting?	5	Understanding K2	CO1
(b)	**Interpret** Exploratory Data Analysis with example.	5	Applying K3	CO2
	OR			
4(a)	**List** the 3 factors of Big Data Revolution. **Discuss** N=ALL and n=1 assumptions with examples.	5	Understanding K2	CO1
(b)	With a neat diagram, **Demonstrate** the Data Science Process.	5	Applying K3	CO2

Course Incharge | HOD | IQAC- Coordinator | Principal

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

Dr. K. RAMA NARASIMHA
Principal/Director
K S School of Engineering and Management
Bengaluru - 560 109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESSION: 2023-2024 (EVEN SEMESTER)
II SESSIONAL TEST QUESTION PAPER
SET- A

USN | | | | | | | | | | |

Degree	:	B.E	Semester	:	VI A&B
Branch	:	Computer Science and Engineering	Course Code	:	21CS644
Course Title	:	Data Science and Visualization	Date	:	09/07/2024
Duration	:	60 Minutes	Max Marks	:	20

Note: Answer ONE full question from each part.

Q No.	Question	Marks	K - Level	CO mapping
	PART-A			
1(a)	**Demonstrate** Linear Regression with an **example**.	5	Applying K3	CO2
(b)	**List** the Feature selection methods. **Explain** each in detail.	5	Understanding K2	CO3
	OR			
2(a)	**Give an example** to **explain** the K-means algorithm and also discuss its issues and advantages.	5	Applying K3	CO2
(b)	**Define** bagging. **Explain** Random Forest with an **example** diagram.	5	Understanding K2	CO3
	PART-B			
3(a)	**Illustrate** the K-NN algorithm with suitable **example**.	5	Applying K3	CO2
(b)	What is Dimensionality Reduction? **Illustrate** Singular Value Decomposition(SVD).	5	Applying K3	CO3
	OR			
4(a)	In the case study of RealDirect, **demonstrate** how RealDirect makes money.	5	Applying K3	CO2
(b)	**Explain** decision tree for chasing dragons. **Develop** a R script for the same.	5	Applying K3	CO3

Course Incharge | HOD | IQAC-Coordinator | Principal

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

Dr. K. RAMA NARASIMHA
Principal/Director
K S School of Engineering and Manage
Bengaluru - 560 109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESSION: 2023-2024 (EVEN SEMESTER)
III SESSIONAL TEST QUESTION PAPER
SET-A

USN										

Degree	:	**B.E**	**Semester**	:	**IV A&B**
Branch	:	**Computer Science and Engineering**	**Course Code**	:	**21CS644**
Course Title	:	**Data Science & Visualization**	**Date**	:	**30/07/2024**
Duration	:	**60 Minutes**	**Max Marks**	:	**20**

Note: Answer ONE full question from each part.

Q No.	Question	Marks	K - Level	CO mapping
	PART-A			
1(a)	**Define** data wrangling. **Explain** With a neat diagram the steps involved in the data wrangling process.	5	Understanding K2	CO4
(b)	**Explain** the overview of Plots in Matplotlib and also discuss the Anatomy of a Matplotlib Figure.	5	Understanding K2	CO5
	OR			
2(a)	**Discuss** scatter plots for single group, multiple groups and also explain scatter plot with marginal histograms with a neat diagram.	5	Understanding K2	CO4
(b)	**Explain** the following with a matplotlib script and suitable diagram: i. Annotations ii. Legends	5	Understanding K2	CO5
	PART-B			
3(a)	**Discuss** how the Density plot is different from Histogram plots. **Compare** Box plots and Violin plots with a neat suitable diagrams.	5	Applying K3	CO4
(b)	Explain the following with syntax, plots and **develop** a python code for i) Bar chart(for single category and subcategories) ii) Pie chart(for water usage) iii) Histogram iv)2D Histogram	5	Applying K3	CO5
	OR			
4(a)	**Draw** an example graph to discuss various Geo plots.	5	Applying K3	CO4
(b)	**Develop** a python code to demonstrate Subplots, Tight layouts and Gridspec.	5	Applying K3	CO5

Course Incharge **HOD** **IQAC- Coordinator** **Principal**

CBCS SCHEME

USN | | | | | | | | | | |

21CS644

Sixth Semester B.E. Degree Examination, June/July 2024
Data Science and Visualization

Time: 3 hrs. Max. Marks: 100

Note: ***Answer any FIVE full questions, choosing ONE full question from each module.***

Module-1

1 a. Define the Data Science. What are the skill sets required for Data Scientists? **(08 Marks)**

b. Explain the following concepts :
i) Datafication ii) Statistical modeling. **(12 Marks)**

OR

2 a. Discuss the "Fitting Model" in Data Science. **(08 Marks)**

b. Explain the following concepts with suitable examples :
i) Population ii) Samples. **(12 Marks)**

Module-2

3 a. Briefly explain the Exploratory Data Analysis (EDA) process with an example. **(08 Marks)**

b. Explain the working of the K – Nearest Neighbour (K – NN) algorithm, with a suitable example. **(12 Marks)**

OR

4 a. Briefly explain the Data Science Process with a neat diagram. **(08 Marks)**

b. Explain the working of the K – mean algorithm with a suitable example. **(12 Marks)**

Module-3

5 a. What is Feature Extraction? Discuss the Principal Component Analysis (PCA) with example. **(08 Marks)**

b. What is a Decision tree and how does the decision tree play a role in feature selection? Explain the implementation with suitable example. **(12 Marks)**

OR

6 a. What is the Feature selection? Discuss the need for feature selection of classification of feature selection algorithms. **(08 Marks)**

b. What is Random Forest? Analyze the working of the random forest algorithm, with an example. **(12 Marks)**

Module-4

7 a. What is Data Visualization and the importance of Data Visualization? What makes a good visualization? **(08 Marks)**

b. Explain the following concepts :
i) Scatter Plot ii) Histogram iii) Box Plot. **(12 Marks)**

OR

8 a. Define "Data Wrangling" and Identify the steps in the flow of the data wrangling process with a neat diagram to measure employee engagement. **(08 Marks)**

b. Explain the following concepts :
i) Choropleth Map ii) Line chart iii) Bubble chart. **(12 Marks)**

Module-5

9 a. Apply a pie chart for water usage. To understand the reason behind it, generate a visual representation of water usage using a pie chart in matplotlib library.

Usage	Clothes washer	Leak	Other	Toilet	Shower	Faucet
Percentage (%)	17	12	8	24	20	19

(10 Marks)

b. What is the Visualization layout in matplotlib? Explain to create subplots and grid space using matplotlib with suitable examples. **(10 Marks)**

OR

10 a. Explain the Vertical and Horizontal bar chart with parameters. Create a vertical bar chart with suitable example using matplotlib library. **(10 Marks)**

b. What is the "Legend" function and Basic text direction? How to implement legend and basic text in the "Stacked bar chart"? **(10 Marks)**

* * * * *

Model Question Paper with effect from 2021(CBCS Scheme)

USN | | | | | | | | | | |

Sixth Semester B.E. Degree Examination
Data Science and Visualization

TIME: 03 Hours **Max. Marks: 100**

Note: 01. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE.**

		Module -1	*Bloom's Taxonomy Level	COs	Marks
Q.01	a	What is Data Science? Explain.	L2	CO 1	10
	b	Explain in Datafication.	L2	CO 1	10
		OR			
Q.02	a	Explain statistical Inference	L2	CO 1	10
	b	Explain the terms with example: 1) Population 2) Sample	L2	CO 1	10
		Module-2			
Q. 03	a	Explain the Data science Process with a neat diagram.	L2	CO 2	10
	b	Which Machine Learning algorithm to be used when you want to express the mathematical relationship between two variables? Explain.	L3	CO 2	10
		OR			
Q.04	a	Explain Exploratory Data Analysis	L2	CO 2	10
	b	Which Machine Learning algorithm to be used when you have bunch of objects that are already classified and based on which other similar objects that haven't got classified to be automatically labelled? Explain.	L3	CO 2	10
		Module-3			
Q. 05	a	Explain feature selection algorithms and selection criterion.	L2	CO 3	10
	b	Define Feature Extraction. Explain different categories of information.	L2	CO 3	10
		OR			
Q. 06	a	Explain Random Forest Classifier.	L2	CO 3	10
	b	Explain Principal Component Analysis.	L2	CO 3	10
		Module-4			
Q. 07	a	What is the need of Data Visualization? Explain its importance.	L2	CO 4	10
	b	Explain Data Wrangling with a neat diagram.	L2	CO 4	10
		OR			
Q. 08	a	Explain composition plots with diagram.	L2	CO 4	10
	b	Explain i) Tools and libraries used for visualization. ii) Data Representation.	L2	CO 4	10
		Module-5			
Q. 09	a	Explain Plotting Using pandas DataFrames, Displaying Figures and Saving Figures in Matplotlib.	L2	CO 5	10
	b	Explain formatting of strings and Plotting in Matplotlib.	L2	CO 5	10
		OR			
Q. 10	a	Explain the following with respect to Matplotlib. Labels, Titles, Text, Annotations, Legends. Subplots	L2	CO 5	10
	b	Explain basic image operations of Matplotlib.	L2	CO 5	10

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by each unit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN | | | | | | | | | | |

Sixth Semester B.E. Degree Examination
DATA SCIENCE AND VISUALIZATION

TIME: 03 Hours **Max. Marks: 100**

Note: 01. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

		Module -1	Bloom's Taxonomy Level	COs	Marks
Q.01	a	What is data science? List and explain skill set required in a data science profile.	L2	CO1	6
	b	Explain Probability Distribution with example.	L2	CO1	6
	c	Describe the process of fitting a model to a dataset in detail.	L2	CO1	8
		OR			
Q.02	a	Explain with neat diagram the current Landscape of data science process.	L2	CO1	6
	b	Explain population and sample with example.	L2	CO1	6
	c	What is big data? Explain in detail 5 elements of bigdata.	L2	CO1	8
		Module-2			
Q. 03	a	What is Machine Learning? Explain the linear regression algorithm.	L2	CO2	6
	b	Explain K-means algorithm with example.	L2	CO2	6
	c	Describe philosophy of EDA in detail.	L2	CO2	8
		OR			
Q.04	a	Explain the data science process with a neat diagram.	L2	CO2	6
	b	Explain KNN algorithm with example.	L2	CO2	6
	c	Develop a R script for EDA.	L3	CO2	8
		Module-3			
Q. 05	a	Explain the fundamental differences between linear regression and logistic regression.	L2	CO3	6
	b	Explain selecting an algorithm in wrapper method.	L2	CO3	6
	c	Explain decision tree for chasing dragon problem.	L3	CO3	8
		OR			
Q. 06	a	Briefly explain alternating Least squares methods.	L2	CO3	6
	b	Explain different selecting criterion in feature selection.	L2	CO3	6
	c	Explain dimensionality problem with SVD in detail.	L3	CO3	8
		Module-4			
Q. 07	a	Define data visualization and explain its importance in data analysis.	L2	CO4	6
	b	Describe different types of plots in comparison plots.	L2	CO4	6
	c	Plot the following i) density plot ii) box plot iii) violin plot iv) bubble plot	L3	CO4	8
		OR			
Q. 08	a	Describe the process of data wrangling and its significance in data visualization.	L2	CO4	6
	b	Explain the variants of bar chart with example.	L2	CO4	6
	c	Explain different types of plots in relation plots.	L2	CO4	8
		Module-5			
Q. 09	a	Develop a code for labels, titles in matplotlib.	L3	CO5	6
	b	Write code for basic pie chart.	L3	CO5	6

	c	Explain with neat diagram Anatomy of a Matplotlib Figure and Plotting data points with multiple markers.	L2	CO5	8
		OR			
Q. 10	a	Describe the process of creating a box plot in Matplotlib. with suitable programming example.	L2	CO5	6
	b	Apply code for scatter plot on animal statistics using matplotlib.	L3	CO5	6
	c	Develop a code for bar chart, pie chart in matplotlib.	L3	CO5	8

CBCS SCHEME

USN | | | | | | | | | | |

20SCS21

Second Semester M.Tech. Degree Examination, July/August 2022
Data Science

Time: 3 hrs. Max. Marks: 100

Note: Answer any FIVE full questions, choosing ONE full question from each module.

Module-1

1 a. Define Data Science. Explain the Venn diagram of Data Science. **(08 Marks)**
b. Explain the Data Science Profile. **(06 Marks)**
c. Explain the work of the Data Scientist in academia and industry. **(06 Marks)**

OR

2 a. What is Datafication? Explain with examples. **(06 Marks)**
b. Explain the following concepts with examples:
(i) Statistical inference
(ii) Population
(iii) Samples
(iv) Types of data **(10 Marks)**
c. Explain the Probability Distribution. **(04 Marks)**

Module-2

3 a. Explain Exploratory Data Analysis with example. **(10 Marks)**
b. Briefly explain Data Science Process with a neat diagram. **(10 Marks)**

OR

4 a. Explain the Linear Regression technique in brief. **(10 Marks)**
b. Explain K-Nearest Neighbors Algorithm. **(10 Marks)**

Module-3

5 a. Why Linear Regression and K-NN are poor choices for filtering spam? Discuss. **(10 Marks)**
b. Explain the Naïve Bayes Algorithm for Filtering Spam with example. **(10 Marks)**

OR

6 a. Describe scraping the web with API's and other tools. **(10 Marks)**
b. Explain Laplace Smoothing. **(05 Marks)**
c. Compare Naïve Bayes Algorithm with K-NN algorithm. **(05 Marks)**

Module-4

7 a. Explain and construct Decision Tree with an example. **(10 Marks)**
b. Write the short notes on:
(i) Feature selection criteria
(ii) Random Forest
(iii) The three Primary Methods of Regression
(iv) The Kaggle model **(10 Marks)**

OR

8 a. Explain Singular Value Decomposition. (05 Marks)
b. Describe the problems with the Nearest Neighbor in recommendation system. (05 Marks)
c. Explain Principal Component Analysis. (10 Marks)

Module-5

9 a. What is a Social Network? List and explain the characteristics of Social Network. (05 Marks)
b. Explain the Social Network Clustering Methods. (05 Marks)
c. Explain Girvan-Newman algorithm with example. (10 Marks)

OR

10 a. Explain the Neighborhood properties in graphs. (10 Marks)
b. Find the Normalized cuts for the following below graph. Fig.Q10. (05 Marks)
c. Find the Laplacian Matrix for the following below graph G, Fig.Q10. (05 Marks)

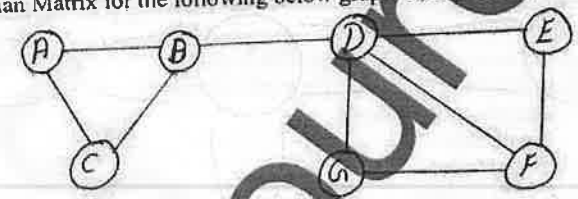


Fig.Q10

* * * * *

MODULE-1

Module 1 Syllabus	**Introduction:** What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? – Datafication, Current landscape of perspectives, Skill sets needed **Statistical Inference:** Populations and samples, Statistical modelling, probability distributions, fitting a model.

Handouts for Session 1: What is data Science

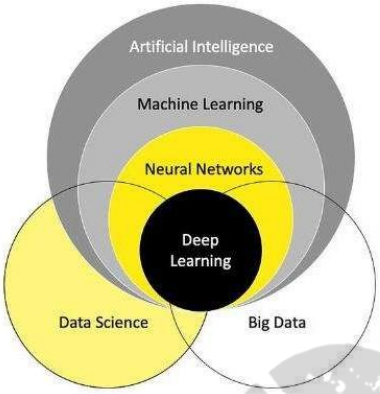
1.0 Data Science

- Data science is an interconnected field that involves the use of statistical and computational methods to extract insightful information and knowledge from data.

All data disciplines in a nutshell

- **Data science** is the broad scientific study that focuses on making sense of data.Ex: Release of movie on Christmas or New year.
- **Data mining** is commonly a part of the data science pipeline. But unlike the latter, data mining is more about techniques and tools used to unfold patterns in data that were previously unknown and make data more usable for analysis.
- Ex: finding the study of movie released and the profit gained in two years.
- **Machine learning** aims at training machines on historical data so that they can process new inputs based on learned patterns without explicit programming, meaning without manually written out instructions for a system to do an action.
- **Deep learning** is the most hyped branch of machine learning that uses complex algorithms of deep neural networks that are inspired by the way the human brain works. DL models can draw accurate results from large volumes of input data without being told which data characteristics to look at.
- Ex: Imagine you need to determine which Movie generate positive online reviews on your website and which cause the negative ones. In this case, deep neural nets can extract meaningful characteristics from reviews and perform sentiment analysis.
- **Artificial intelligence** is a complex topic. But for the sake of simplicity, let's say that any real-life data product can be called AI.

- Let's stay with our fishing-inspired example. You want to buy a certain model fishing rod but you only have a picture of it and don't know the brand name. An AI system is a software product that can examine your image and provide suggestions as to a product name and shops where you can buy it. To build an AI product you need to use data mining, machine learning, and sometimes deep learning.



Questions:

1.Define Data Science.

2.Explain Data Disciplines in detail with example.

3.Define Machine Learning

4.Define Data Mining.

5.Define AI with example.

Handouts for Session 2: Big Data and Data Science Hype

1.1 Big Data and Data Science Hype

Data science is an interconnected field that involves the use of statistical and computational methods to extract insightful information and knowledge from dataThe hype surrounding big data and data science has been quite significant in recent years, and it's not without reason. So, what is eyebrow-raising about Big Data and data science? Let's count the ways:

1. There's a lack of definitions around the most basic terminology. What is big data? What does data science mean? What is the relationship between big data and data science? Is data Science, the science of big data? Is data science the only the stuff happening in big tech companies? Is big data referred to as cross

discipline(astronomy, finance, tech etc) and Data science takes place only in tech? How big is Big in Big Data? In short there is a lot of ambiguity!

2. There is a distinct lack of respect for researchers in academia and industry labs who have been working on this kind of data for years and whose work is based on decades or centuries of work by statisticians, computer scientists, mathematicians, engineers and scientists of all types. There is a lot of Media Hype about this topic which makes it seem like Machine Learning algorithms were just discovered last week and data was never big until Google came along. This is not true. Many techniques and methods we are using now and the challenges we are facing are a part of the evolution of everything that's come before.
3. The hype is crazy—people throw around tired phrases straight out of the height of the pre-financial crisis era like "Masters of the Universe" to describe data scientists, and that doesn't bode well.There are new and exciting things happening as well, but one must respect the things that came before and lead to what is happening today. But the unreal hype has just increased the Noise to Signal Ratio! The longer the hype goes on, the more of us will get turned off by it and this will lead to people missing out the good benefits of data science under all the hype! The terms have lost their basic meaning and now are too ambiguous, thus today they seem meaningless.
4. Statisticians already feel that they are working on the Science of Data and are having a sense of Identity Theft. However Data Science is not just a Rebranding of statistics or machine learning. It is a field by itself, unlike how media makes it sound like it is just statistics or machine learning in the industry context.
5. Data Science may not be science as most people say, but it definitely is more of a craft!

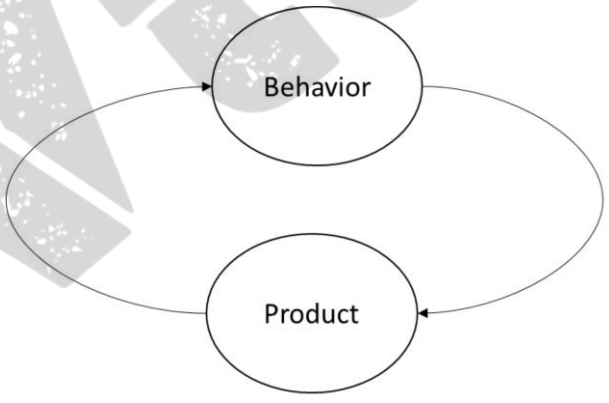
1.2 Getting Past the Hype

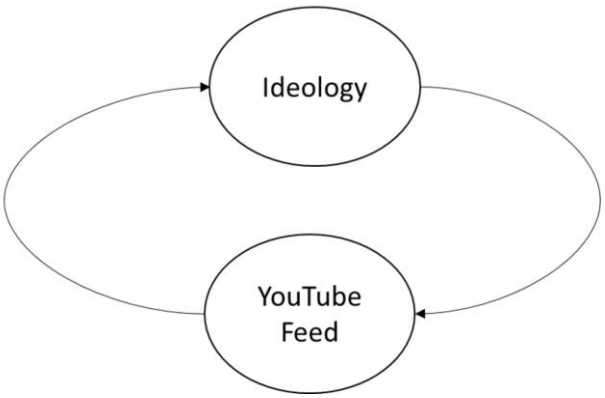
- When you transition from academia to industry, you realize there is a gap between things learned in college and what you do on the job – Industry-Academia Gap. Why does it have to be that way? Why are academic courses out of touch with reality?
- General Experience of a data scientist is that at their job they have access to a larger body of knowledge and methodology as well as a process – Data Science

Process which has foundations in both statistics and computer science. It is something new, fragile and a nascent idea which is at a real risk of being rejected prematurely due to all the unrealistic hype

1.3 Why Now data is required?

- There is availability of massive amount of data from many aspects of our lives. There is abundance of inexpensive computing power.All our activities online such as Shopping, communication, news Consumption, Music preferences, Search records, expression of opinions are tracked.
- There is Datafication of our offline behaviour as well such as Finance Data, Medical Data, Bioinformatics and social welfare. From the online and offline data put together there's a lot to learn about behavior and we are as a species There is a growing influence of data in most sectors and most industries, In some cases the amount collected is considered to be BIG and in some cases it is not.
- The massiveness of data makes it very interesting as well as challenging. Often data itself becomes the building blocks of data products – Amazon Recommendation systems, Facebook Friend recommendation, Movie recommendation on Netflix, Music Recommendation on Spotify.
- In Finance it is Credit Ratings, Trading Algorithms, Risk Assessment etc. In E-Learning – Dynamic personalized learning and assessments – Khan Academy and Knewton. In Government it is creation of Policies based on Data. This is the beginning of a culturally saturated feedback loop





- Behavioral Data of people changes the product. The Product changes the behavior of the people. Large scale data processing, increased memory and bandwidth and cultural acceptance of technology in our lives makes it possible unlike a decade ago. Considering the impact of this feedback loop, we should seriously start

thinking how it is being conducted along with the ethical and technical responsibilities for the people responsible for the process.

Questions

1.Explain the Reasons for Data Science Hype

2.Explain in detail the requirement of data

Handouts for Session 3: Datafication and Current Landscape

1.4 Datafication

- Datafication is a process of taking all aspects of life and turning them into data – Kenneth Neil Cukier and Viktor Mayer Schoenberger (Rise of Big Data, 2013).
- Everything we do online or otherwise ends up recorded for later examination in data storage units or for sale – Facebook isn't free.
- We are the product. Google's AR Glasses Datafies gaze, Twitter Datafies Stray Thoughts LinkedIN Datafies Professional Networks. Consider the importance of datafication with respect to people's intentions about sharing their own data. We are being datafied.
- Our actions are being datafied.The spectrum of this ranges from us gleefully taking part in social media experiment we are proud of to all out surveillance and stalking. When we "Like" something online, we are intending to be datafied or at least we should expect to be When we browse the web, we are unintentionally or at least passively being datafied via cookies we may or may not be aware of.
- When we are walking around on the streets, we are being datafied by Cameras. Once we datafy things, we can transform their purpose and turn the information into new forms of value.
- Who is "We"? It is usually modelers and entrepreneurs profiting from getting people to buy stuff. What kind of "Value"? Increased efficiency through automation.

1.5 The Current Landscape – Skillsets Needed

- Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics. But data science is not merely hacking—because when hackers finish debugging their Bash one-liners and Pig scripts, few of them careabout non-Euclidean distance metrics And data science is not merely

statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.Data science is the civil engineering of data.

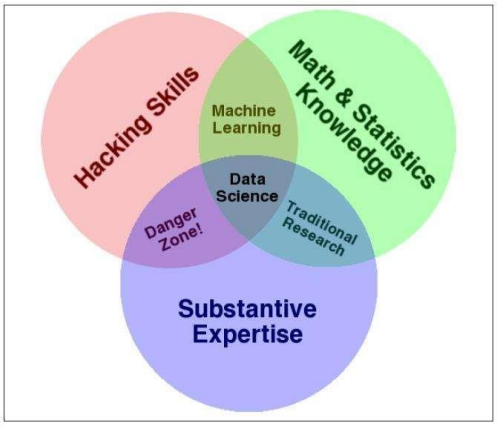


Figure 1-1. Drew Conway's Venn diagram of data science

- Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible.
- So, the statement is essentially saying that while hackers may be proficient at writing code and solving technical problems, they may not necessarily have the depth of knowledge or interest in the mathematical and statistical concepts that are crucial to data science.
- While statisticians may excel in theoretical aspects of data analysis, they may lack the programming skills necessary to handle real-world data effectively.
- In summary, while statistics is an important component of data science, data science encompasses a broader set of skills and activities beyond statistical analysis, including programming, data manipulation, and machine learning.

Questions:

1.Explain the process of datafication in details.

2. Explain the Current landscape of DataScience Process

Handouts for Session 4: Skill Sets

1.6 Data Science Profile

Expertise in the following fields is a requirement

- **Computer Science:** Data science relies heavily on programming and computational tools to manipulate, analyze, and visualize data. Proficiency in programming languages like Python, R, or SQL is essential for data acquisition,

cleaning, and analysis. Additionally, knowledge of algorithms and data structures enables efficient processing of large datasets.

- **Math:** Mathematics forms the foundation of data science. Concepts from calculus, linear algebra, and discrete mathematics are used for understanding and implementing machine learning algorithms, statistical modeling, and optimization techniques.
- **Statistics:** Statistics provides the framework for making inferences and predictions from data. Understanding probability theory, hypothesis testing, regression analysis, and sampling methods is crucial for analyzing data, assessing model performance, and drawing meaningful conclusions.
- **Machine Learning:** Machine learning algorithms are at the heart of data science, enabling systems to learn from data and make predictions or decisions. Data scientists need to understand various machine learning techniques, such as supervised learning (e.g., linear regression, decision trees), unsupervised learning (e.g., clustering, dimensionality reduction), and deep learning (e.g., neural networks).
- **Communication and Presentation Skills:** Data scientists must effectively communicate their findings and insights to stakeholders, which requires strong verbal and written communication skills. They should be able to translate complex technical concepts into layman's terms and craft compelling narratives. Presentation skills involve creating visually appealing and informative presentations or reports to convey the results of data analyses.
- **Data Visualization:** Visualizing data is essential for exploring patterns, trends, and relationships within datasets. Data scientists use tools like matplotlib, ggplot2, or Tableau to create meaningful visualizations that aid in understanding and interpreting data. Effective data visualization enhances communication and decision-making processes.
- **Extensive Domain Expertise:** Domain knowledge is critical for contextualizing data and understanding the specific challenges and opportunities within a particular industry or field. Data scientists with extensive domain expertise can identify relevant variables, interpret results in the appropriate context, and develop actionable insights tailored to the needs of stakeholders.
- In summary, proficiency in computer science, math, statistics, machine learning, communication, data visualization, and domain expertise are all essential requirements for success in the field of data science.

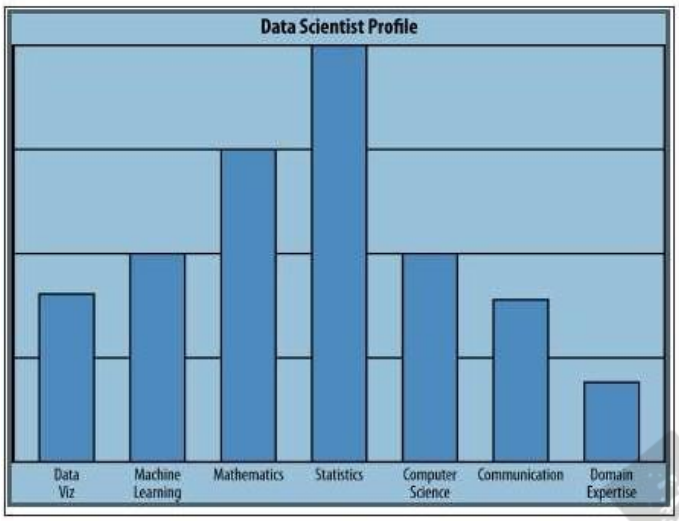
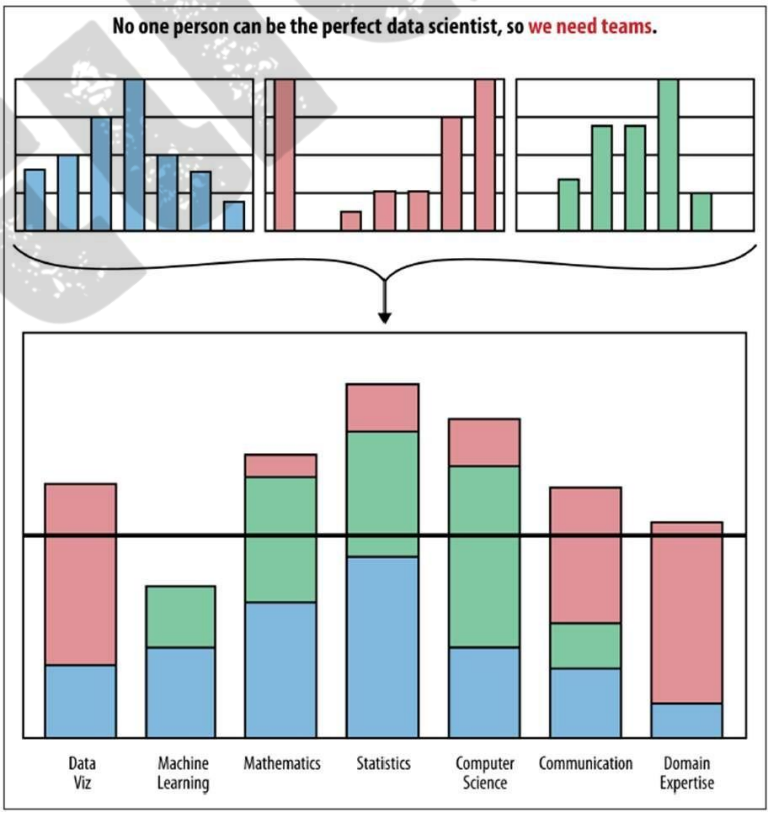


Figure 1-2. Rachel's data science profile, which she created to illustrate trying to visualize oneself as a data scientist; she wanted students and guest lecturers to "riff" on this—to add buckets or remove skills, use a different scale or visualization method, and think about the drawbacks of self-reporting

- As we mentioned earlier, a data science team works best when different skills (profiles) are represented across different people, because nobody is good at everything. It makes us wonder if it might be more mworthwhile to define a "data science team"—as shown in Figure 1-3—than to define a data scientist.



1.7 What is Data Scientist

- Here we discuss regarding how data scientist is defined in terms of academics and industry.

In Academia:

- For the term "data science" to catch on in academia at the level of the faculty, and as a primary title, the research area needs to be more formally defined. An academic data scientist is a scientist, trained in any of the academic fields, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real world problem.

In Industry:

- A chief data scientist sets the data strategy of the company, which involves a variety of things: setting everything up from the engineering and infrastructure for collecting data and logging, to privacy concerns, to deciding what data will be user-facing, how data is going to be used to make decisions, and how it's going to be built back into the product. They will also manage a team of engineers, scientists and analysts and communicate with leadership across the company including the CEO, CTO, and product leadership. They should also be concerned with patenting innovative solutions and setting research goals.They will be concerned with patenting innovative solutions and setting research goals. In a general sense a Data scientist is someone:
- More generally, a data scientist is someone who knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning, as well as being human. She spends a lot of time in the process of collecting, cleaning, and munging data, because data is never clean. This process requires persistence, statistics, and software engineering skills—skills that are also necessary for understanding biases in the data, and for debugging logging output from code.
- Once they get the data into shape, a crucial part is exploratory data analysis, which combines visualization and data sense. They will find patterns, build models, and algorithms—some with the intention of understanding product usage and the overall health of the product, and others to serve as prototypes that ultimately get baked back into the product. They may design experiments, and which is a critical part of data driven decision making. They will communicate

with team members, engineers, and leadership in clear language and with data visualizations so that even if her colleagues are not immersed in the data themselves, they will understand the implications.

Questions:

1.Explain the Skill Sets needed for the Data Scientist Profile.

2. Explain the Data Scientist in terms of Academics and Industry in detail.

Handouts for Session 5: Statistical Inference

1.8 Statistical Inferencing

- As we commute to work on subways and in cars, as our blood moves through our bodies, as we're shopping, emailing, procrastinating at work by browsing the Internet and watching the stock market, as we're building things, eating things, talking to our friends and family about things, while factories are producing products, this all at least potentially produces data.
- Imagine spending 24 hours looking out the window, and for every minute, counting and recording the number of people who pass by. Or gathering up everyone who lives within a mile of your house and making them tell you how many email messages they receive every day for the next year. Imagine heading over to your local hospital and rummaging around in the blood samples looking for patterns in the DNA. That all sounded creepy, but it wasn't supposed to. The point here is that the processes in our lives are actually data-generating processes.
- Data represents the traces of the real-world processes, and exactly which traces we gather are decided by our data collection or sampling method. You, the data scientist, the observer, are turning the world into data.
- Once you have all this data, you have somehow captured the world, or certain traces of the world. But you can't go walking around with a huge Excel spreadsheet or database of millions of transactions and look at it and, with a snap of a finger, understand the world and process that generated it. So you need a new idea, and that's to simplify those captured traces into something more comprehensible, to something that somehow captures it all in a much more concise way, and that something could be mathematical models or functions of the data, known as statistical estimators. This overall process of going from the

world to the data, and then from the data back to the world, is the field of **statistical inference**.

- It is a discipline that is concerned with the development of procedures, methods and theorems that allows the extraction of meaning and information from the data generated by stochastic processes.

World→ Data→ World – Statistical Inferencing

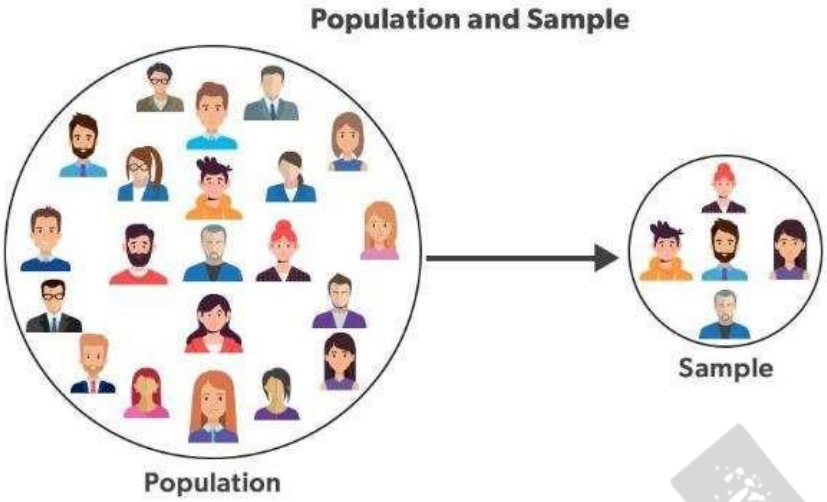
Questions:

1.Explain Statistical Inference with Example.

Handouts for Session 6: Population and Samples

1.9 Population and Samples

- **Population** is any set of objects or units (tweets, photographs, stars). If the characteristics of all the objects can be measured or extracted, then it is a complete set of observations (N).
- A single observation can include a list of characteristics of the object. A subset of the units of size (n) considered to make observations and draw conclusions and make inferences about the population is called a sample. Taking a subset may introduce biases into the data and distort it.
- **Example:** Suppose you are conducting research on smartphone usage habits among teenagers in a specific city. Your population comprises all teenagers aged 13-18 living in that city, which could number in the tens of thousands. Due to logistical constraints and the difficulty of reaching every teenager in the city, you opt to use a sample of 500 teenagers randomly selected from different schools within the city. This sample will participate in interviews or surveys to provide insights into their smartphone usage patterns, preferences, and behaviors.
- Population refers to the total set of observation. say, you are looking for the average height of men. Here population is the set of all men in the world.



1.10 Population and Samples of Big Data

- In the age of big data, we still need to take sample because sampling solves some engineering problems. How much data is needed depends on the goal. For analysis or inferencing there is no need to store the data all the time. For serving purpose you may need it all the time in order to render correct information.
- ***Bias:*** If a sample of data is observed, it may have an inherent bias in it and the data may be representative of only that subset and not of the entire population, thereby any conclusion or inference drawn from it should not be extended to the entire population. – Tweet Pre-Hurricane Sandy and Post-Hurricane Sandy.
- If the tweets immediately before hurricane sandy is analyzed, one would infer that most people went supermarket shopping.If the tweets immediately after hurricane sandy is analyzed, one would infer that most people went partyingMost tweets were from New Yorkers, who are heavy tweeters and not from New Jerseyans. Coastal New Jerseyans were worried about house collapsing etc. and did not have time to tweet
- If only the limited data was studied, the only conclusion one would draw is what the Hurricane Sandy was like for a subset of twitter users (who are not a representative of the whole population) and would infer that the hurricane was not that bad

- ***Types of Data***
    - Traditional – numerical, categorical and binary
    - Text – Emails, Tweetys, Reviews, News Articles

- Records – User-Level Data, Timestamped event data, JSON-Formatted Log Files
- Geo-Based Location Data – Housing Data
- Network
- Sensor Data
- Images

- New Data requires new strategies for sampling. If a Facebook user-level data aggregated from timestamped event logs is analyzed for a week, can any conclusions be drawn that is relevant next week or year? How to sample From a network and preserve the complex network structure? Many of these questions are open research questions

1.11 BIG Data

- BIG is a moving target – When the size of the data becomes a challenge we refer to it as big. BIG is when you cannot fit all the data on one machine. BIG data is a cultural phenomenon. It is characterized by The Vs – Volume, Variety, Velocity, Value, Validity and Veracity.

Elements of Big Data

- **Volume** – Data Measured in terms of petabytes and exabytes (1mn TB), made possible by reduction in cost of storage devices
- **Velocity** – The fast arrival speed of data and increase in data volume. Powered by IoT and High Speed Internet
- **Variety** – Form – Many forms of data – Text, graph, audio, video maps, composite (Video with audio). Function – Human Conversations, Transaction records, old archive data. Source of Data – Open/public data, social media data, multimodal data
- **Veracity** – Aspect like conformity to the facts, truthfulness, believability, and confidence in data – Error sources – technical, typographical and human
- **Validity** – Accuracy of the data for talking decisions or for any other goals
- Value - the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.

1.12 Big Data Can Mean Big Assumptions

- Big Data revolution consists of three things:
    - Collecting and using a lot of data rather than small samples
    - Accepting messiness in your data
    - Giving up on knowing the causes
- N=all very often is not a good assumption and misses the things we should consider the most.
- For Example: Election Day Polls – Even if we poll everyone who leaves the polling stations, we still don't count people who decided not to vote. And these maybe the people we need to understand the voting problems! Recommendations received on Netflix may not be good because people who bother to rate the shows may different taste, leading to a skew in the recommendation system towards the taste of the people who rated.
- Data is not Objective. Data does not speak for itself. Example: Algorithm for hiring – Consider an organization that did not treat female employees well. So when deciding to compare men and women with same qualifications, data showed that women tend to leave more often, get promoted less often and give more negative feedback on the environment than men. The automated model based on this data will likely hire a man over a woman if a man and a woman with the same qualification turned up for the interview. Ignoring Causation can be a flaw rather than a feature and add to historical problems rather than address them. Data is just a quantitative representation of the events of our society
- The n=1 ,assumption is considered for Sample size of 1.For a single person, we can actually record a lot of information We might even sample from all the actions they took in order to make inferences about them This is used in User-Level Modeling.

Questions:

1.Explain Population with example

2.Explain Sample with example

3.Explain the different types of data in population and sampling of big data.

4.Explain in detail 5 elements of big data.

5.Explain the Big Data Revolution in detail.

Handouts for Session 7: Modeling

1.13 Modeling

- Humans try to understand the world around them by representing it in different ways called Models. Statisticians and data scientists capture the uncertainty and randomness of data-generating processes with mathematical functions that express the shape and structure of the data itself. A model is an attempt to understand and represent the nature of reality through a particular lens, be it architectural, biological, or mathematical. It is an artificial construction where all extraneous detail has been removed or abstracted. Attention must always be paid to these abstracted details after a model has been analyzed to see what might have been overlooked.
- In the case of a statistical model, we may have mistakenly excluded key variables, included irrelevant ones, or assumed a mathematical structure which is far from reality.
- A Model is an attempt to understand the population of interest and represent that in a compact form which can be used to experiment, analyze, study and determine cause-and-effect and similar relationships amongst the variables under study in the population.

1.14 Statistical Modeling

- Statistical Modeling is an expression of relationship such as What comes first? What influences what? What causes what? What's a test of that? in terms of mathematical expressions that will be general enough that they have to include parameters, but the parameter values are not yet known.
- Other people prefer pictures and will first draw a diagram of data flow, possibly with arrows, showing how things affect other things or what happens over time. This gives them an abstract picture of the relationships before choosing equations to express them.

- For example: If there are two columns x and y of data and there is a linear relationship between them then we can represent it as $y = (a x + b)$. Where a and b are parameters whose values are not yet known. Below figure depicts the Prediction of salary based on the experience of employee

How to build a model?

- To build a model we start with Exploratory Data Analysis (EDA) which includes making plots, building intuition for a particular dataset. It involves plotting histograms and looking at scatter plots to get a feel for the data. Representative functions are written down.
- It start with a simple linear function and see if it makes sense. If it does not make sense then understand why and see what representative function would make more sense and keep building up the complexity (Eg: Go Parabolic after Linear).Write down complete sentences and try to express the words as equations & code. Simple plots may be easier to interpret and understand.
- A Trade off may usually be required during Modeling. A Simple model may get you 90% of the way & may take few hours to build, whereas a complex model may get you up to 92% and may take months to build.
- Example

Python3

```
import pandas as pd
import numpy as np
# read datasdet using pandas
df = pd.read_csv('employees.csv')
df.head()
```

Output:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services

Questions:

1.Define Modeling.Explain how to build a model with example.

2.Expalin statistical Modeling with Example

Handouts for Session 6: Probability Distributions and Fitting a Model

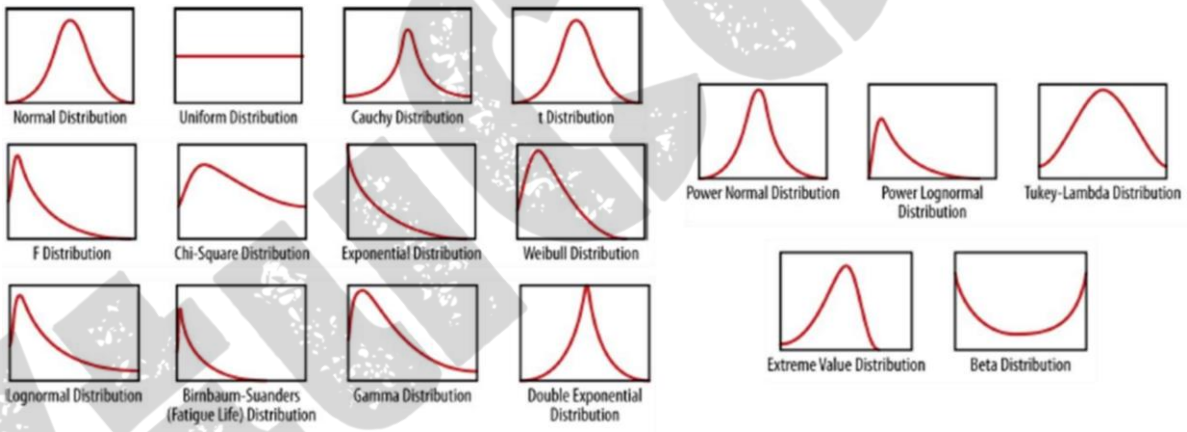
1.14 Probability Distributions

- Probability Distributions are foundations of statistical models.
- Probability distributions are fundamental concepts in statistics and probability theory. In the context of data science, understanding probability distributions is crucial for modeling and analyzing data.
- **Probability Distribution:** A probability distribution describes the likelihood of each possible outcome of a random variable. It assigns probabilities to different values that the variable can take.
- Example – Normal (Gaussian) Distribution, Poisson Distribution, Weibull Distribution, Gamma Distribution, Exponential Distribution. Natural Processes tend to generate measurements whose empirical shape could be approximated by mathematical functions with a few parameters that could be estimated from the data. Not all processes generate data that looks like a named distribution, but many do. These functions can be as building blocks of our models.

- It's beyond the scope of the book to go into each of the distributions in detail, but we provide them in Figure below as an illustration of the various common shapes, and to remind you that they only have names because someone observed them enough times to think they deserved names. There is actually an infinite number of possible distributions. They are to be interpreted as assigning a probability to a subset of possible outcomes, and have corresponding functions. For example, the normal distribution is written as:

Normal Distribution

- $N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
    - μ is mean
    - σ is standard Deviation



- Data tends to be around a central value with no bias on left or right. It is a Symmetric Distribution appearing as a Bell-shaped curve distribution. Mean and Median Controls where the distribution is centered. σ controls how spread out the distribution is. This is the general function form. Specific real-world phenomenon have actual numbers as value which can be estimated from data.
- A random variable denoted by x or y can be assumed to have a corresponding probability distribution p(x) which maps to a positive real number. In order to be a probability density function, we're restricted to the set of functions such that if we integrate p(x) to get the area under the curve, it is 1, so it can be interpreted as probability.

- Example: Let x be the amount of time until the next bus arrives. x is a random variable because there is variation and uncertainty in the amount of time until the next bus. Suppose we know that the time until the next bus has a probability density function of $p(x) = 2e^{-2x}$. If we want to know the likelihood of the next bus arriving in between 12 and 13 minutes, then we find the area under the curve between 12 and 13 by $\int_{12}^{13} 2e^{-2x}$ How do we know that the distribution is correct? We can conduct an experiment where we show up at the bus stop at a random time, measure how much time until the next bus, and repeat this experiment over and over again. Then we look at the measurements, plot them, and approximate the function. Because we are familiar with the fact that "waiting time" is a common enough real-world phenomenon that a distribution called the exponential distribution has been invented to describe it, we know that it takes the form $p(x) = \lambda e^{-\lambda x}$.
- **Joint probability** is the probability of two events occurring simultaneously. Example: washing the car and raining
- **Conditional probability** is the probability of one event occurring in the presence of a second event
- .Example: Total there are 2 Blue marble and 3 Red Marble, If a blue marble was selected first there is now a 1/4 chance of getting a blue marble and a 3/4 chance of getting a red marble. If a red marble was selected first there is now a 2/4 chance of getting a blue marble and a 2/4 chance of getting

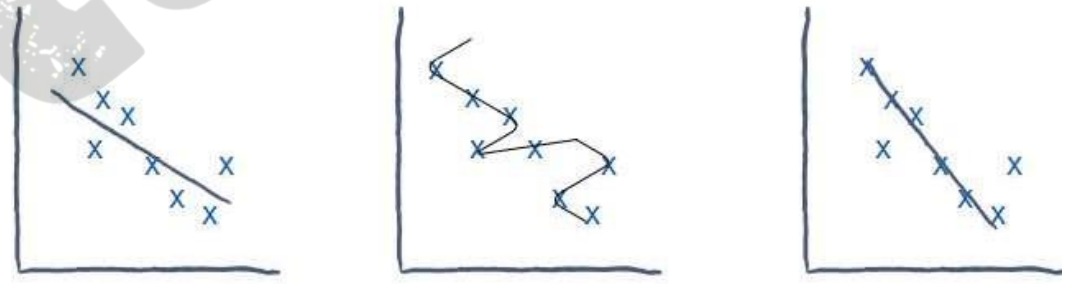
1.15 Fitting a Model

- Fitting a model means estimate the parameters of the model using the observed data. The data is used as evidence to help approximate the real world-mathematical process that generated the data. A **good model fit** refers to a model that accurately approximates the output when it is provided with unseen inputs.
- Fitting the model often involves optimization methods and algorithms such as maximum likelihood estimation, to help get the parameters. When you estimate the parameters, they are actually estimators, meaning they themselves are functions of the data.
- Fitting the model is when you start actually coding: your code will read in the data, and you'll specify the functional form that you wrote down on the piece of paper.

- Then R or Python will use built-in optimization methods to give you the most likely values of the parameters given the data. Initially you should have an understanding that optimization is taking place and how it works, but you don't have to code this part yourself—it underlies the R or Python functions.
- The process involves running an algorithm on data for which the target variable ("labeled" data) is known to produce a machine learning model. Then, the model's outcomes are compared to the real, observed values of the target variable to determine the accuracy.
- **Overfitting** is the term used to mean that you used a dataset to estimate the parameters of your model, but your model isn't that good at capturing reality beyond your sampled data. Overfitting occurs when a model learns the training data too well, including its noise and outliers, to the extent that it performs poorly on unseen data

Causes of Overfitting:

- Complex Models: Models with too many parameters relative to the size of the training data can capture noise instead of underlying patterns.
- Insufficient Data: When the amount of training data is limited, complex models may find patterns where none exist due to randomness.
- Feature Overfitting: Including irrelevant features or too many features in the model can lead to overfitting.
- Lack of Regularization: Regularization techniques, such as L1 and L2 regularization, help prevent overfitting by penalizing overly complex models.



Underfitted model Overfitted model Fitted model

Questions

1.Explain Probability Distribution with example

2.Explain Overfitting and causes of overfitting with example

MODULE-2

Module 2 Syllabus	**Exploratory Data Analysis and the Data Science Process:** Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA, The Data Science Process, Case Study: Real Direct (online real estate firm). **Three Basic Machine Learning Algorithms:** Linear Regression, k-Nearest Neighbours (k- NN), k-means

Handouts for Session 1: Exploratory Data Analysis and the Data Science Process: Basic tools (plots, graphs and summary statistics) of EDA

2.1 Exploratory Data Analysis (EDA)

- **"Exploratory data analysis"** is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there
- Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- It is the First step towards building a model.The understanding of the problem you are working on is changing as you go. – Thereby "Exploratory"

Basic Tools – Plots, Graphs, Summary Statistics

- Method of systematically going through the data, plotting distributions of all variables (using box plots), plotting time series of data, transforming variables, looking at all pairwise relationships between variables using scatterplot matrices, and generating summary statistics for all of them.
- At the very least that would mean computing their mean, minimum, maximum, the upper and lower quartiles, and identifying outliers.
- EDA is about understanding the data and gaining intuition, understanding the shape of it and connecting the understanding of the process that generated the data to the data itself.

Questions:

1.Define EDA.

2.Explain the basic tools involved in EDA

Handouts for Session 2: Philosophy of EDA

2 Philosophy of EDA

- Gain Intuition about the data, Make comparisons between distributions, sanity checking (ensuring data is on the expected scale and format), missing data analysis, outlier analysis and summarize it.
- In the context of data generated from logs, EDA helps with the debugging process. Patterns found in the data could be something actually wrong with the logging process that needs fixing. If you never go to the trouble of debugging, you'll continue to think your patterns are real.
- EDA helps to ensure that the product is performing as intended.
- The insights drawn from EDA can be used to improve the development of algorithms.
- Example: Develop a ranking algorithm that ranks content shown to the users. – Develop a notion of "Popular"
- Before deciding how to quantify popularity (no. of clicks, most commented, average etc.) the behaviour of the data needs to be understood.

Exercise: EDA

There are 31 datasets named nyt1.csv, nyt2.csv,…,nyt31.csv, which you can find here: https://github.com/oreillymedia/doing_data_science.

Each one represents one (simulated) day's worth of ads shown and clicks recorded on the New York Times home page in May 2012. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number impressions, number clicks, and loggedin.

We use R to handle these data. It's a programming language designed specifically for data analysis, and it's pretty intuitive to start using. Code can be written based on the following logic,

- **Reading Data**: Loading a dataset from a URL.
- **Categorization:** Creating age categories based on the 'Age' variable.

- **Summary Statistics:** Generating summary statistics for the dataset and for age categories.
- **Visualization:** Creating histograms and boxplots to visualize data distribution.
- **Click-Through Rate (CTR):** Calculating and visualizing the click-through rate.
- **Creating Categories:** Creating a new column 'scode' to categorize data based on impressions and clicks.
- **Converting to Factor:** Converting the newly created column into a factor.
- **Summary Table:** Generating a summary table for impressions based on the created categories.

```
# Reading Data
data1 ← read.csv(url("http://stat.columbia.edu/~rachel/datasets/nyt1.csv"))
# Displaying the first few rows of the dataset
head(data1)
# the cut() function is used to create age categories (agecat) based on the 'Age' variable.
data1$agecat ← cut(data1$Age, c(-Inf, 0, 18, 24, 34, 44, 54, 64, Inf))

# Generating summary statistics for the dataset
summary(data1)
# Installing and loading the doBy package for data manipulation
install.packages("doBy")
library("doBy")
#the siterange function computes and returns a vector containing the following summary #statistics of a given vector
siterange <- function(x){c(length(x), min(x), mean(x), max(x))}
# Generating summary statistics for Age grouped by age categories
summaryBy(Age ~ agecat, data = data1, FUN = siterange)
# Generating summary statistics for Gender, Signed_In, Impressions, and Clicks #grouped by age categories
summaryBy(Gender + Signed_In + Impressions + Clicks ~ agecat, data = data1)
# Installing and loading the ggplot2 package for data visualization
install.packages("ggplot2")
```

```
library(ggplot2)

# Creating a histogram to visualize the distribution of Impressions across age #categories

ggplot(data1, aes(x = Impressions, fill = agecat)) + geom_histogram(binwidth = 1)

# Creating a boxplot to visualize the distribution of Impressions within each age #category

ggplot(data1, aes(x = agecat, y = Impressions, fill = agecat)) + geom_boxplot()

# Creating a new column to indicate whether there are impressions or not

data1$hasimps ←cut(data1$Impressions, c(-Inf, 0, Inf))

# Generating summary statistics for Clicks grouped by the presence or absence of #impressions

summaryBy(Clicks ~ hasimps, data = data1, FUN = siterange)

# Creating density plots to visualize the click-through rate distribution across age #categories

ggplot(subset(data1, Impressions > 0), aes(x = Clicks / Impressions, colour = agecat)) + geom_density()

# Creating density plots for click-through rate, filtering out cases where there are no #clicks

ggplot(subset(data1, Clicks > 0), aes(x = Clicks / Impressions, colour = agecat)) + geom_density()

# Generating a boxplot to visualize the distribution of Clicks within each age #category

ggplot(subset(data1, Clicks > 0), aes(x = agecat, y = Clicks, fill = agecat)) + geom_boxplot()

# Creating a density plot to visualize the distribution of Clicks across age categories

ggplot(subset(data1, Clicks > 0), aes(x = Clicks, colour = agecat)) + geom_density()

# Creating a new column to categorize the data based on impressions and clicks

data1$scode[data1$Impressions == 0] ← "NoImps"

data1$scode[data1$Impressions > 0] ←"Imps"

data1$scode[data1$Clicks > 0] ← "Clicks"

# Converting the newly created column into a factor

data1$scode ← factor(data1$scode)

# Generating a summary table for impressions based on the created categories and other #variables

etable ←summaryBy(Impressions ~ scode + Gender + agecat, data = data1, FUN = clen)
```

Questions:

1.Explain EDA and explain the steps involved in EDA

2.Write R script for demonstrating EDA

Handouts for Session 3: Data Science Process

2.3 The Data Science Process

- The real world where different types of data is generated. Inside the Real World are lots of people busy at various activities. Some people are using Google+, others are competing in the Olympics; there are spammers sending spam, and there are people getting their blood drawn. Say we have data on one of these things.
- Raw data is recorded. Lot of aspects to these real word activities are lost even when we have that raw data. Real world data is not clean. The raw data is processed to make it clean for analysis. We build and use data munging pipelines (joining, scraping, wrangling). This done with Python, R, SQL Shell scripts.
- Eventually data is brought into a format with columns.

  name | event | year | gender | event time

- The EDA process can now be started. During the course of the EDA we may find that the data is not actually clean as there are missing values, outliers, incorrectly logged data or data that was not logged.
- In such a case, we may have to collect more data or we can spend more time cleaning the data (Imputation). The model is designed to use some algorithm (K-NN, Linear Regression, Naïve Bayes, Decision Tree, Random Forest etc) Model Selection depends on type of problem being addressed – Prediction, Classification or a basic description problem.
- Alternatively, our goal may be to build or prototype a "data product" such as a spam classifier, search ranking algorithm or a recommendation system.The key difference here that differentiates data science from statistics here is that, the data product is incorporated back into the real world and users interact with it and that generates more data, which creates a feedback loop.

- A Movie Recommendation system generates evidence that lots of people love a movie. This will lead to more people watching the movie – feedback loop
- Take this loop into account in any analysis you do by adjusting for any biases your model caused. Your models are not just predicting the future, but causing it!

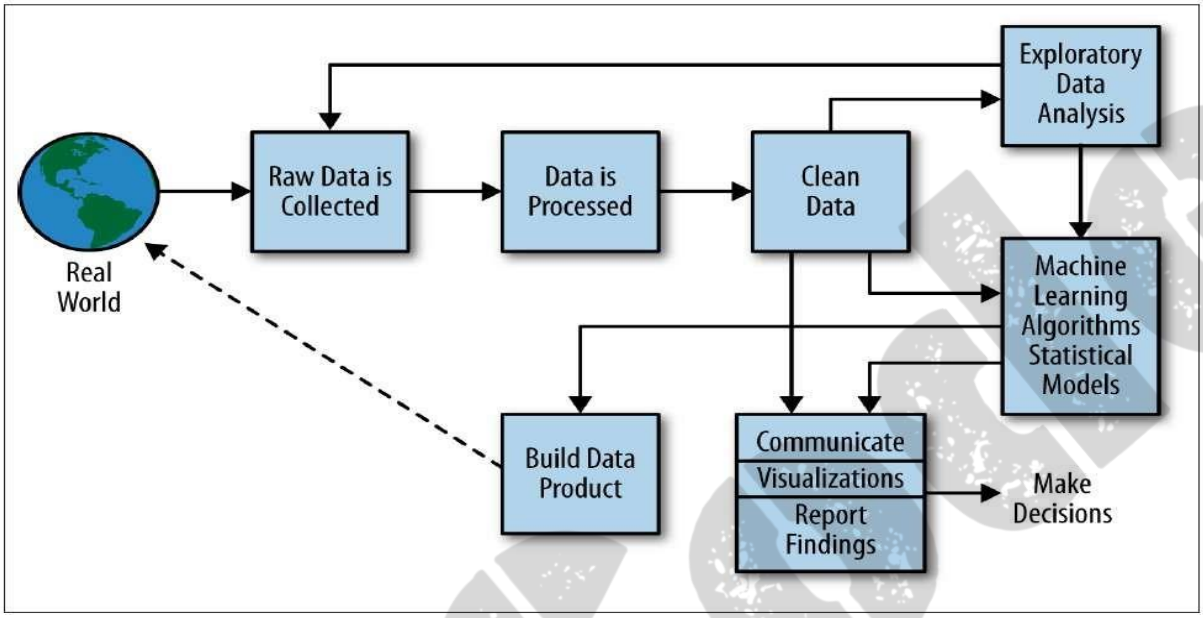


Figure 1: The Data Science Process

2.4 A Data Scientist's Role in this Process

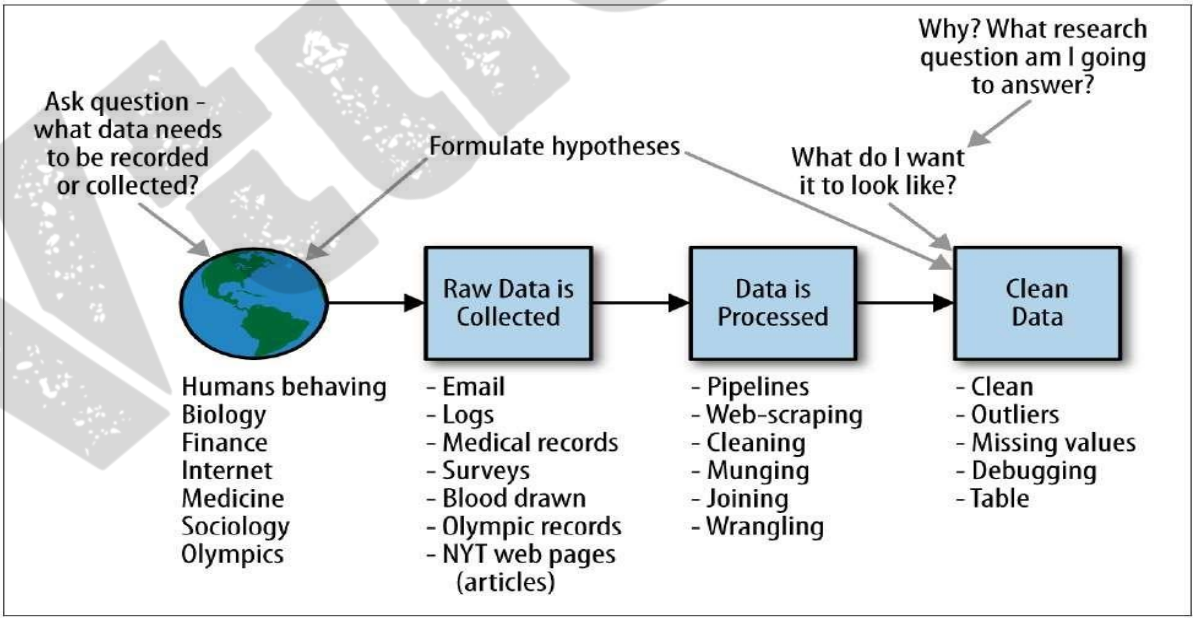


Figure 2: The Data Scientist's Role

- A Human Data Scientist has to make the decisions about what data to collect, and why.
- That person needs to be formulating questions and hypotheses and making a plan for how the problem will be attacked.
- Let's revise or at least add an overlay to make clear that the data scientist needs to be involved in this process throughout, meaning they are involved in the actual coding as well as in the higher-level process, as shown in Figure 2.
- Connection with the Scientific Method:
    - ✓ Ask a question.
    - ✓ Do background research.
    - ✓ Construct a hypothesis.
    - ✓ Test your hypothesis by doing an experiment.
    - ✓ Analyze your data and draw a conclusion.
    - ✓ Communicate your results.
- Not every problem requires one to go through all the steps, but almost all problems can be solved with some combination of the stages.

Questions:

1.Explain the process involved in Data Science.

2.Explain in details the role of Data Scientist.

Handouts for Session 4: Case Study, Three Basic Machine Learning Algorithm

2.5 Case Study: RealDirect

- **Goal**: Use all the accessible real estate data to improve the way people buy/sell houses
- **Problem Statement**: Normally people sell their homes about once every 7 years with the help of professional brokers and current data.
- Brokers are typically free-agents and guard their data aggressively and the really good ones have a lot of experience (i.e slightly more data than the inexperienced brokers).

- Solution by RealDirect:
    - ✓ Hire a team of licensed real estate agents who work together and pool their knowledge.
    - ✓ Provide an interface for sellers, giving them useful data driven tips on how to sell their house.
    - ✓ Uses the interaction data to give real time recommendations on what to do next.
- The team of brokers also become data experts learning to use information collecting tools to keep tabs on new and relevant data or to access publicly available data.
- Publicly available data is old and has a 3-month lag between a sale and when the data about the sale is available.
- RealDirect is working on real-time feeds on when people start searching for a home and what the initial offer is, the time between offer and close and how people search for a home online
- Good information helps both buyer and seller.

➨ **How does RealDirect Make Profits?**

- Subscription to sellers – about $395 a month to access the selling tools
- Sellers can use RealDirect's agents at a reduced commission typically 2% of the sale instead of the usual 2.5% or 3%
- The data pooling enables RealDirect to take smaller commission as it is more optimized and therefore gets more volume
- The site is a platform for buyers and sellers to manage their sale or purchase process.
- There are statuses for each person on site: active, offer made, offer rejected, showing, in contract, etc.
- Based on Status different actions are suggested by the platform.
- Key issues that a buyer might care about—nearby parks, subway, and schools, as well as the comparison of prices per square foot of apartments sold in the same building or block.
- This is the kind of data they want to increasingly cover as part of the service of RealDirect.

2.6 Algorithms

An algorithm is a procedure or set of steps or rules to accomplish a task. Algorithms are one of the fundamental concepts in, or building blocks of, computer science: the basis of the design of elegant and efficient code, data preparation and processing, and software engineering.

With respect to data science, there are at least three classes of algorithms one should be aware of:

1. Data munging, preparation, and processing algorithms, such as sorting, MapReduce, or Pregel. These algorithms are characterized as data engineering.
2. Optimization algorithms for parameter estimation, including Stochastic Gradient Descent, Newton's Method, and Least Squares.
3. Machine learning algorithms - largely used to predict, classify, or cluster.

Machine Learning Algorithms

Machine learning algorithms that are the basis of artificial intelligence (AI) such as image recognition, speech recognition, recommendation systems, ranking and personalization of content— often the basis of data products—are not usually part of a core statistics curriculum or department.

Three Basic Algorithms

Many business or real-world problems that can be solved with data are classification and prediction problems when expressed mathematically. Those models and algorithms can be used to classify and predict.

The key challenge for data scientists isn't just knowing how to implement statistical methods, but rather understanding which methods are appropriate based on the problem and underlying assumptions.

It's about knowing when and why to use certain techniques, considering factors like the nature of the problem, data characteristics, and contextual requirements.

Questions:

1. Explain in Detail the process involved in the case study of Real Direct.

Handouts for Session 5: Linear Regression Algorithm

2.7 Linear Regression Algorithm

- Linear regression is a common statistical method used to show the mathematical relationship between two variables.
- It assumes a linear connection between an outcome variable (also called the response variable, dependent variable, or label, like sales) and a predictor variable (also called an independent variable, explanatory variable, or feature, like advertising spend). Essentially, it helps us understand how changes in one variable can predict changes in another.
- Sometimes, it makes sense that changes in one variable correlate linearly with changes in another variable. For example, it makes sense that the more umbrellas you sell, the more money you make.

- **Example 1.** Suppose you run a social networking site that charges a monthly subscription fee of $25, and that this is your only source of revenue.
- Each month you collect data and count your number of users and total revenue. You've done this daily over the course of two years, recording it all in a spreadsheet.You could express this data as a series of points.
- Here are the first four:

$$S = \{(x, y) = (1,25), (10,250), (100,2500), (200,5000)\}$$

- From the given data it can be observed that y =25x. Which shows that,
  - There's a linear pattern.
  - The coefficient relating x and y is 25.
  - It seems deterministic.

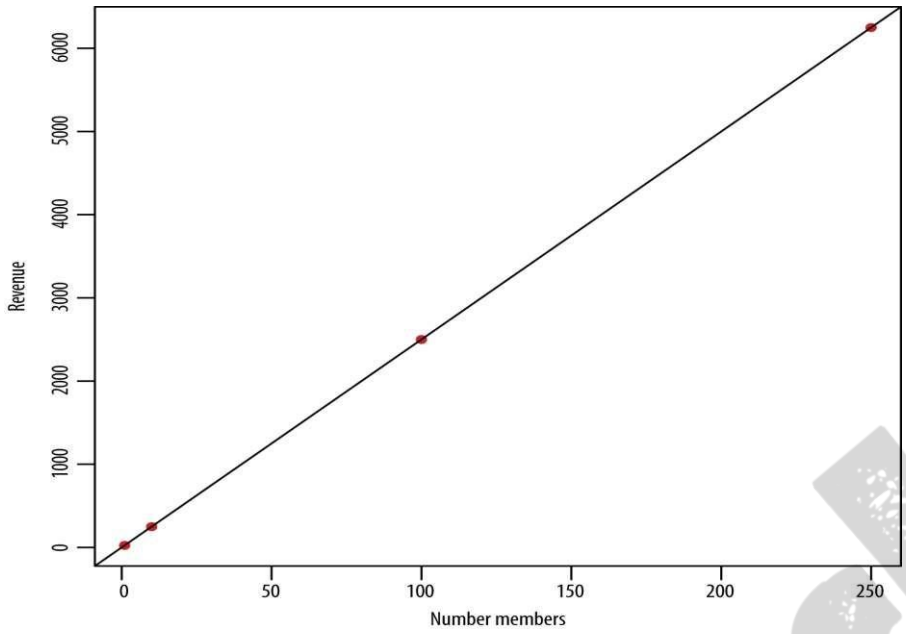


Figure 3: An observed linear pattern

- **Example 2 – User Level Data:** The dataset, keyed by user, contains weekly behavior data for hundreds of thousands of users on a social networking site, with columns like *total_num_friends, total_new_friends_this_week, num_visits, time_spent, number_apps_downloaded, number_ads_shown, gender, and age.* During exploratory data analysis (EDA), a random sample of 100 users was used to plot pairs of variables, such as **total_new_friends** vs. **time_spent**. The business goal is to forecast the number of users to promise advertisers, but the current focus is on building intuition and understanding the dataset. The first few rows are listed below :

total_new_friends	**time_spent**
7	276
3	43
4	82
6	136
10	417
9	269

- When plotted the graph looks like below,

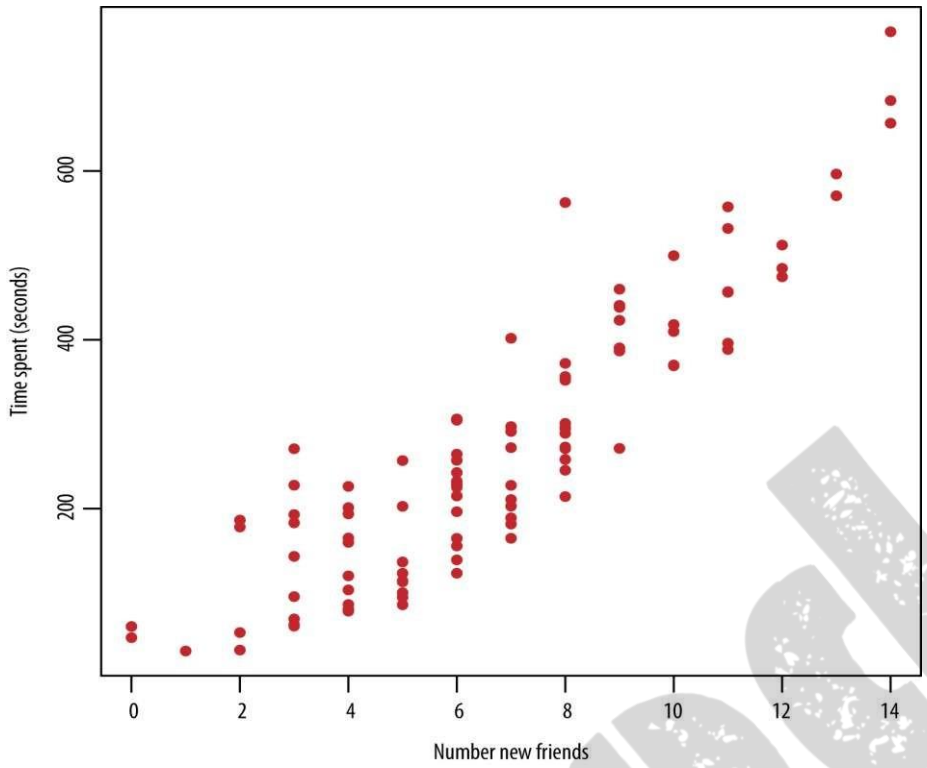


Figure 4: Dataset Plotted

- There seems to be a linear relationship between the number of new friends and the time spent on the social networking site, suggesting that more new friends lead to more time spent on the site.
- This relationship can be described using statistical methods like correlation and linear regression. Although there is an association between these variables, it is not perfectly deterministic, indicating that other factors also influence the time users spend on the site.
- **Start by writing something down**

  To model the relationship, capture the **trend** and **variation**. Start by assuming a linear relationship between variables. Focus on the trend first, using linear modeling to describe how the number of new friends relates to time spent on the site.

- There are many lines that look more or less like they might work, asshown in below figure

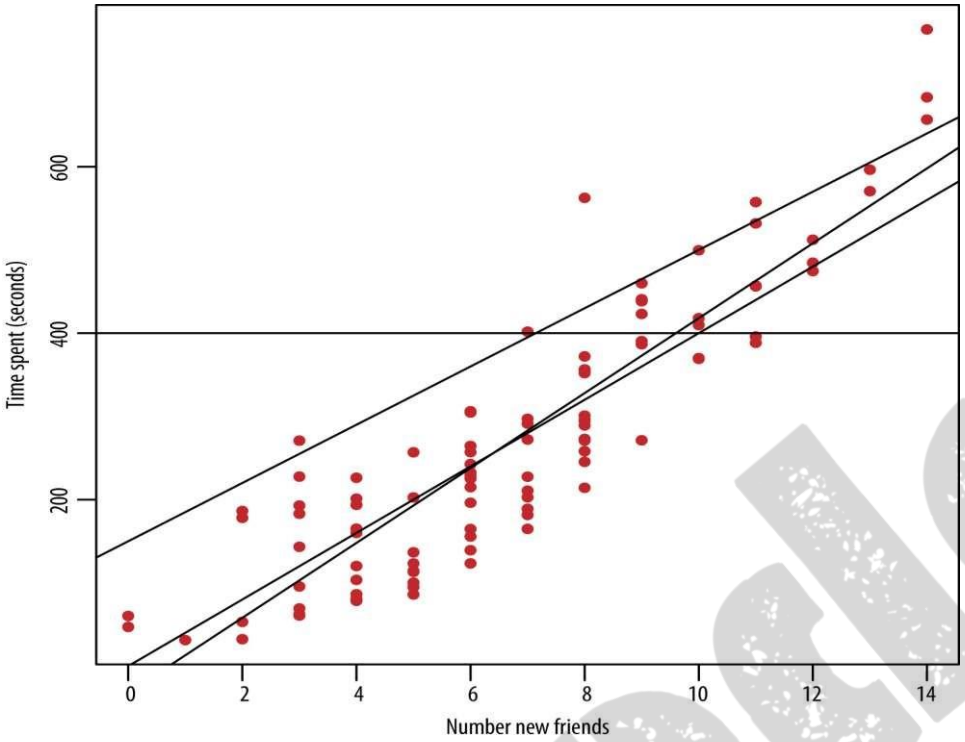


Figure 5: Which line is the best fit?

- To begin modelling the assumed linear relationship $y=\beta 0+\beta 1x$ the task is to find the optimal values for $\beta 0$(intercept is the value of y when x=0) and $\beta 1$(slope of the line- how much y changes for a unit change in x) using the observed data $x1,y1,x2,y2,...,xn,yn$. This model can be expressed in matrix notation as $y=x\cdot\beta$. The next step involves fitting this model to the data.

- **Fitting the model:** In **linear regression**, the goal is to **calculate coefficients $\boldsymbol{\beta}$** by finding the line that minimizes the average distance between all data points and the line itself. **This is achieved by minimizing the sum of squared residuals, representing the vertical distances between data points and the line.**
- Linear regression aims to minimize the sum of the squares of the vertical distances between predicted $\hat{y}_i$s and observed y-values. This minimization reduces prediction errors and is known as **least squares estimation**.

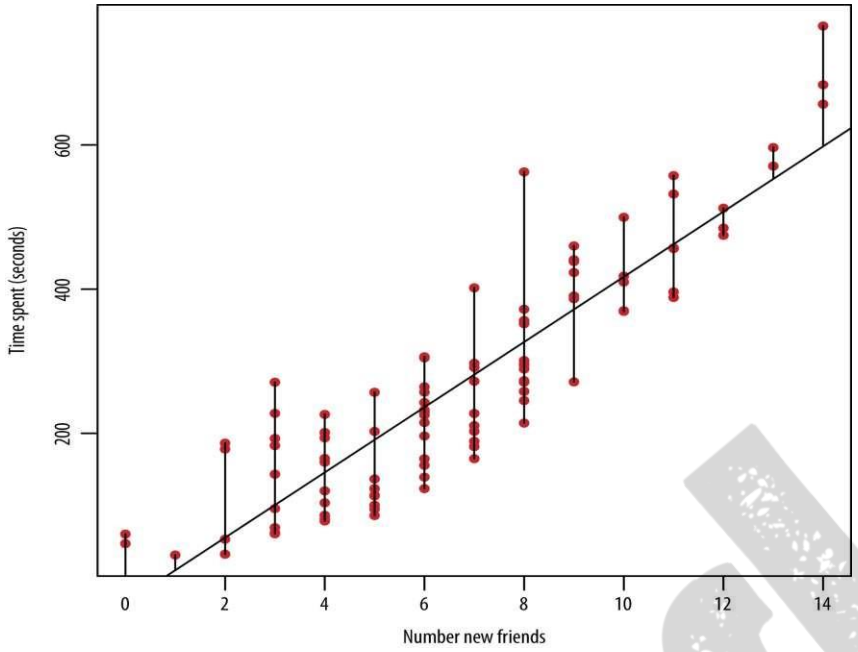


Figure 6: The line closest to all the points

- To find the optimal line, the "residual sum of squares" (RSS), denoted as RSS(β), is defined as the sum of the squared vertical distances between observed points and any given line. It's represented as:

$$RSS(\beta) = \sum_i (y_i - \beta x_i)^2$$

Here, i ranges over the various data points. This function of β needs to be optimized to find the optimal line.

- To minimize $RSS(\beta) = (y - \beta x)^t (y - \beta x)$, differentiate it with respect to β and set it equal to zero, then solve for β. This results in:

$$\hat{\beta} = (x^t x)^{-1} x^t y$$

- This formula gives the vector β that minimizes the RSS.

The "hat" symbol ($\hat{\beta}$) indicates that it's the estimator for β. Since the true value of β is unknown, we use the observed data to compute an estimate using this estimator.

- To fit the linear regression model and obtain the β coefficients in R, you can use the **lm()** function with a simple one-liner. For example:

model <- lm(y ~ x)

This line of code creates a linear regression model named **model** where **y** is the response variable and **x** is the predictor variable. The **lm()** function in R automatically calculates the β coefficients for the linear model based on the provided data.

- The following line of code fits a linear regression model where **time_spent (y)** is the response variable and **total_new_friends (x)** is the predictor variable, using the data provided. This model will estimate the relationship between the number of new friends and the time spent on the social networking site.

Code:

```
# Perform linear regression
model <- lm(y ~ x)
model
coefs <- coef(model)

# Plot the data and regression line
plot(x, y, pch=20, col="red", xlab="Number new friends", ylab="Time spent (seconds)")

abline(coefs[1], coefs[2])

# Plot histogram of time spent
hist(my_dataset$time_spent, breaks = 10, col = "blue",  xlab = "Time spent (seconds)", ylab = "Frequency",main = "Histogram of Time Spent on the Site")
```

output:

When you display model it gives the following

```
  Call:
  lm(formula = y ~ x)

  Coefficients:
        (Intercept)                  x
```

-32.08 45.92

And the estimated line is $y = -32.08+45.92x$, which can be rounded to $y = -32+46x$, a nd the corresponding plot looks like below,

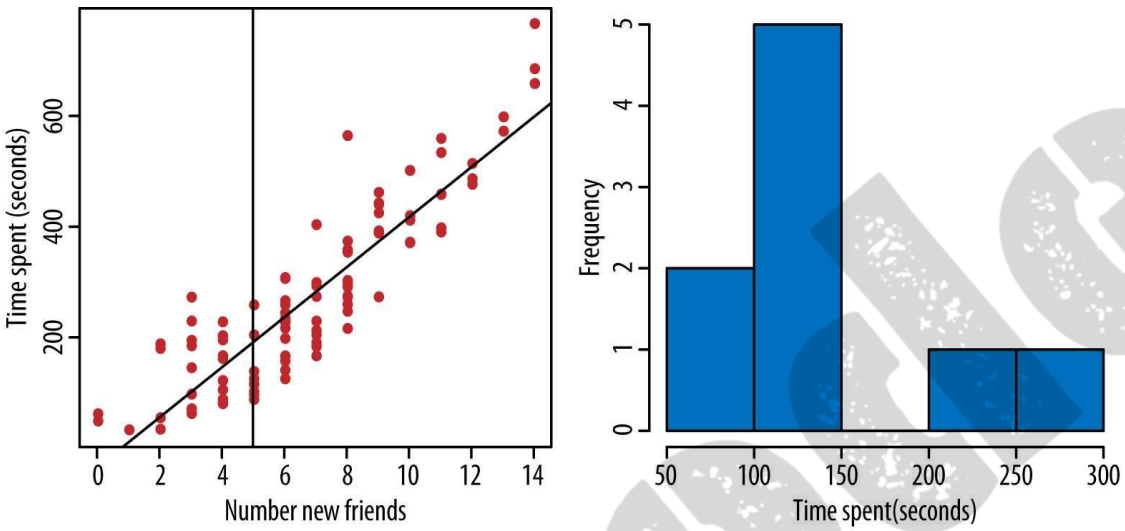


Figure 7: Left: Model fitting Right: Time spent (5 new friend)

The graph on the right shows,

1. **X-axis (Time spent)**: The x-axis represents the range of time spent by users on the social networking site, divided into intervals or bins. Each bin represents a range of time spent (e.g., 0-100 seconds, 101-200 seconds, and so on).
2. **Y-axis (Frequency)**: The y-axis represents the frequency or count of users falling within each bin. It shows how many users spent a particular amount of time on the site, as indicated by the height of the bars.

If a new x-value of 5 came in, meaning user had five new friends, how confident are we in the output value of

32.08+45.92*5 = 195.7s

- To address the question of confidence, you need to extend your model to account for variation. While you've modeled the trend between the number of new friends and time spent on the site, you haven't yet modeled the variation. This means that you wouldn't claim that everyone with five new friends spends exactly the same amount of time on the site.

2.8 Extending beyond least squares

- Now that you have a simple linear regression model down (one output, one predictor) using least squares estimation to estimate your βs, you can build upon that model in three primary ways,
    1. Adding in modeling assumptions about the errors
    2. Adding in more predictors
    3. Transforming the predictors

1. Adding in modeling assumptions about the errors

- When applying the model to predict y for a specific x value, the prediction lacks the variability present in the observed data.
- See on the right-hand side of **Figure 7** that for a fixed value of x =5, there is variability among the time spent on the site. This variability can be shown in the model as,

$$y = \beta_0 + \beta_1 x + \epsilon$$

- The new term ϵ, also known as **noise** or the **error term**, represents the unaccounted variability in the data. It reflects the difference between the observed data points and the true regression line, which can only be estimated using the regression coefficients $\hat{\beta}s$. It is the difference between observations and true regression line. Assumption is that noise is Normally Distributed e~N(0, σ^2). P(y|x) ~N(β1x + β2, σ^2) is the conditional distribution of y given x.
- Eg: Among the set of people who had five new friends this week, the amount of the time they spent on the website had a normal distribution with a mean of β1*5 + β2 and a variance of σ^2, and you're going to estimate your parameters β1, β2,σ from the data.
- Measure the residual, how far the observed points are from the estimated line

$e_i = y_i - \hat{y} = y_i - (\beta 1 x_i + \beta 2)$ for i= 1 to n. The variance of e is : $\frac{\sum_i e_i^2}{n-2}$: Divide by n-2 produces an unbiased estimator. This is called the Mean Squared Error – It captures how much the predicted value varies from the actual value.

- **Evaluation Metrics:** Next we need to find how best the model we have built so we need to use different **evaluation metrics** to measure it. So we have R-squared, p-values, Cross-validation.

  i)R-squared

  $$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

  This can be interpreted as the proportion of variance explained by our model. MSE getting divided by Total Error is the proportion of variance unexplained by the model.1-variance unexplained gives the variance explained by the model.

ii)p-values: In Statistical hypothesis testing, the P-value or sometimes called probability value, is used to observe the test results or more extreme results by assuming that the null hypothesis (H0) is true.P-value is also used as an alternative to determine the point of rejection in order to provide the smallest significance level at which the null hypothesis is least or rejected. It is expressed as the level of significance that lies between 0 and 1, and if there is smaller p-value, then there would be strong evidence to reject the null hypothesis. If the value of p-value is very small, then it means the observed output is feasible but doesn't lie under the null hypothesis conditions (H_0).The p-value of 0.05 is known as the level of significance (**α**). Usually, it is considered using two suggestions, which are given below:

- **If p-value>0.05:** The large p-value shows that the null hypothesis needs to be accepted.
- **If p-value<0.05:** The small p-value shows that the null hypothesis needs to be rejected, and the result is declared as statically significant.

iii)Cross-Validation: Divide the data into a training set and a test set: 80% in the training and 20% in the test. Fit the model on the training set. Then look at

the mean squared error on the test set and compare it to that on the training set. Make this comparison across sample size as well. If the mean squared errors are approximately the same, then the model generalizes well and there is no overfitting.

2. Other Models for Error Terms

- The mean squared error is an example of what is called a loss function. This is the standard one to use in linear regression because it gives us a pretty nice measure of closeness of fit. It has the additional desirable property that by assuming that are normally distributed, we can rely on the maximum likelihood principle. There are other loss functions such as one that relies on absolute value rather than squaring. It's also possible to build custom loss functions specific to your particular problem or context

3. Transforming the predictors

- What we just looked at was simple linear regression, one outcome or dependent variable and one predictor. But we can extend this model by building in other predictors, which is called *multiple linear regression*:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \epsilon.$$

- The R code would be

```
model<-lm(y~x_1+x_2+x_3)

model<-lm(y~x_1+x_2+x_3+ x_2*x_3)
```

- Make scatterplots of y against each of the predictors as well as between the predictors, and histograms of y|x for various values of each of the predictors to help build intuition. To evaluate the model we can use R-squared, p-values, and using cross validation with training and testing sets.
- Sometimes the relationship between may not be linear. The relationship maybe polynomial in nature. It may be possible sometimes that the assumption made is a linear relationship, but the real relationship is quadratic. Acquiring more data can be helpful in this regard.

$$y= (a_0x + a_1x^2 + a_2x^3 + b)+e$$

Questions:

1.Explain the linear Regression with example.

2.Explain Linear Regression Model with R script.

3.Explain the three primary ways to extend the linear regression beyond the least squares

Handouts for Session 6: kNN Algorithm

2.9 k-Nearest Neighbors (k-NN) Algorithm

- K-Nearest Neighbors (K-NN) is an algorithm employed for automatically labeling unclassified objects based on their similarity to already classified ones in a dataset. For instance, it could be applied to classify data scientists as "best" or "worst", individuals as "high credit" or "low credit," restaurants by star ratings, or patients as "high cancer risk" or "low cancer risk," among various other applications.

- The intuition behind K-Nearest Neighbors (K-NN) is to identify the most similar items based on their attributes, examine their labels, and assign the unclassified item the majority label. In the case of a tie, one of the tied labels is randomly selected.

- In the context of movie ratings, K-Nearest Neighbors (K-NN) allows you to predict the rating of an unrated movie, such as "Data Gone Wild," by analyzing its attributes like length, genre, number of comedy scenes, number of Oscar-winning actors, and budget. By comparing these attributes with those of already rated movies, the algorithm identifies the most similar movies and assigns a rating based on the collective ratings of its nearest neighbors, enabling predictions without watching the movie.

- To automate the process, **two key decisions** are essential: defining the measure of similarity or closeness between items and utilizing this measure to identify the most similar items, known as neighbors, to an unrated item. These neighbors contribute their "votes" towards the classification or labeling of the unrated item.

- The **second decision** involves determining the number of neighbors to consider for voting, denoted as "k." As a data scientist, you'll choose this value, which dictates the extent of influence from neighboring items on the classification or labeling of the unrated item.

- Consider a dataset consisting of the age, income, and a credit category of high or low for a bunch of people and you want to use the age and income to predict the credit label of "high" or "low" for a new person.
- **For example**, here are the first few rows of a dataset, with income represented in thousands:

age	income	credit
69	3	low
66	57	low
49	79	low
49	17	low
58	26	high
44	71	high

plot people as points on the plane and label people with an empty circle if they have low credit ratings.

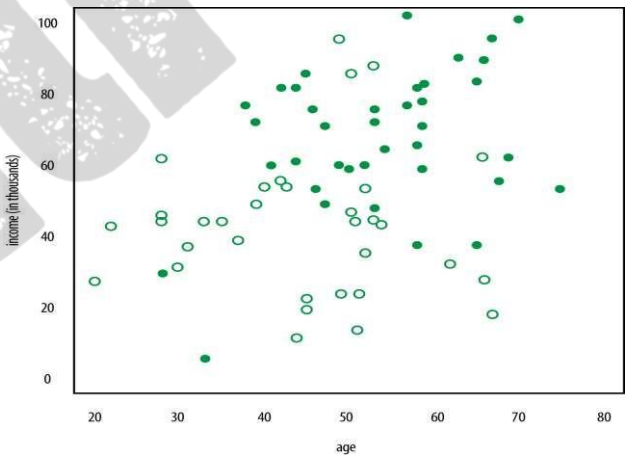


Figure 8:Credit rating as a function of age and income

- What if a new guy comes in who is 57 years old and who makes $37,000? What's his likely credit rating label?

- Given the credit scores of other individuals nearby, what credit score label do you propose should be assigned to him? Let's use K-Nearest Neighbors (K-NN) to automate this process.

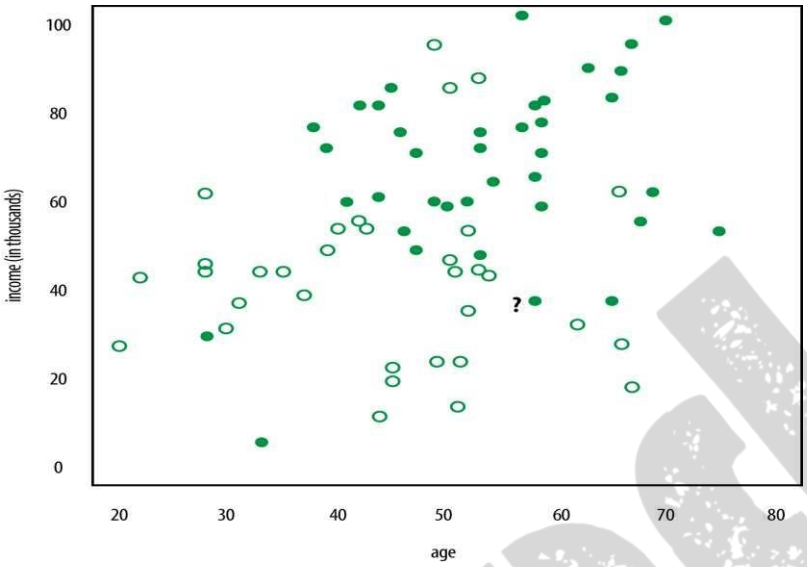


Figure 9: What about the guy?

2.10 Overview of the kNN process:

1. Decide on your similarity or distance metric.
2. Split the original labeled dataset into training and test data.
3. Pick an evaluation metric.
4. Run k-NN a few times, changing k and checking the evaluation measure.
5. Optimize k by picking the one with the best evaluation measure.
6. Once you've chosen k, use the same training set and now create a new test set with the people's ages and incomes that you have no labels for, and want to predict. In this case, your new test set only has one lonely row, for the 57-year-old.

1. Similarity or distance metrics

- Similarity or distance metrics can be employed to quantify the similarity between data points. Definitions of "closeness" and similarity vary depending on the context. There are many more distance metrics available to you depending on your type of data.
- In our scenario, determine a metric (e.g., Euclidean distance) to measure the similarity between individuals based on their age and income. Euclidean distance is a good go-to distance metric for attributes that are real-valued.

1. **Cosine Similarity**

   Also can be used between two real-valued vectors, x and y, and will yield a value between –1 (exact opposite) and 1 (exactly the same) with 0 in between meaning independent.

$$\cos\left(\vec{x}, \vec{y}\right) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \, \|\vec{y}\|}$$

2. **Jaccard Distance or Similarity**

This gives the distance between a set of objects—for example, a list of Cathy's friends A= {Kahn,Mark,Laura, .} and a list of Rachel's friends B= {Mladen,Kahn,Mark, . .} —and says how similar those two sets are

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

3. **Mahalanobis Distance**

Also can be used between two real-valued vectors and has the advantage over Euclidean distance that it takes into account correlation and is scale-invariant.

$$d\left(\vec{x}, \vec{y}\right) = \sqrt{\left(\vec{x} - \vec{y}\right)^T S^{-1} \left(\vec{x} - \vec{y}\right)}$$

4. **Hamming Distance**

   Can be used to find the distance between two strings or pairs of words or DNA sequences of the same length. The distance between olive and ocean is 4 because aside from the "o" the other 4 letters are different. The distance between shoe and hose is 3 because aside from the "e" the other 3 letters are different. You just go through each position and check whether the letters the same in that position, and if not, increment your count by 1.

5. **Manhattan**

   This is also a distance between two real-valued k-dimensional vectors.

$$d\left(\vec{x}, \vec{y}\right) = \Sigma_i^k \left|x_i - y_i\right|$$

where i is the ith element of each of the vectors.

2.Training and test sets for k-NN

- In machine learning, the typical process involves two phases: **training and testing.**
- During **training**, a model is created and trained using labeled data to learn patterns and relationships.
- In the **testing** phase, the model's performance is evaluated using new, unseen data to assess its effectiveness in making predictions or classifications.

- In K-Nearest Neighbors (k-NN), the training phase involves reading the labeled data with "high" or "low" credit points marked. During testing, the algorithm attempts to predict the labels of unseen data points using the k-NN approach without prior knowledge of the true labels, evaluating its accuracy in the process.
- To accomplish this, a portion of clean data from the entire dataset needs to be reserved for the testing phase. Typically, about 20% of the data is randomly selected and set aside for testing purposes.
- Sample R code to prepare Train and Test set,

```
n.points <- 1000 # number of rows in the dataset
sampling.rate <- 0.8

# we need the number of points in the test set to calculate
# the misclassification rate
num.test.set.labels <- n.points * (1 - sampling.rate)

# randomly sample which rows will go in the training set
training <- sample(1:n.points, sampling.rate * n.points,
                   replace=FALSE)
train <- subset(data[training, ], select = c(Age, Income))
# define the training set to be those rows

# the other rows are going into the test set
testing <- setdiff(1:n.points, training)
# define the test set to be the other rows
test <- subset(data[testing, ], select = c(Age, Income))

cl <- data$Credit[training]
# this is the subset of labels for the training set
true.labels <- data$Credit[testing]
# subset of labels for the test set, we're withholding these
```

3. Pick an evaluation metric

- An evaluation metric is a measure used to assess the performance of a machine learning model. Evaluation metrics are not always straightforward or universal, as different scenarios may require prioritizing certain types of errors over others. For instance, false

negatives might be more critical than false positives in certain applications. Collaborating with domain experts to design an evaluation metric tailored to the specific requirements of the problem at hand can be essential.

- For **example**, if you were using a classification algorithm to predict whether someone had cancer or not, you would want to minimize false negatives (misdiagnosing someone as not having cancer when they actually do), so you could work with a doctor to tune your evaluation metric.

  - **Accuracy**: ***Ratio of the number of correct labels to the total number of labels***, and the **misclassification rate**, which is just **1–accuracy**. Minimizing the misclassification rate then just amounts to maximizing accuracy.
  - **Sensitivity (*true positive rate* or *recall)*:** sensitivity is here defined as the probability of correctly diagnosing an ill patient as ill.
  - **Specificity (*true negative rate*):** specificity is here defined as the probability of correctly diagnosing a well patient as well. There is also the **false positive rate** and the **false negative rate**, and these don't get other special names.
  - **True Positive (TP):** Instances that are actually positive and are correctly classified as positive by the model.
  - **False Positive (FP):** Instances that are actually negative but are incorrectly classified as positive by the model.
  - **True Negative (TN):** Instances that are actually negative and are correctly classified as negative by the model.
  - **False Negative (FN):** Instances that are actually positive but are incorrectly classified as negative by the model.

4. Run k-NN and checking the evaluation measure

- Once we know, distance measure and evaluation metric. Apply K-Nearest Neighbors to classify individuals in the test set based on the majority label among their nearest neighbors.
- Calculate the misclassification rate to evaluate model performance.All this is done automatically in R, with just this single line of R code:

```
knn (train, test, cl, k=3)
```

5. Optimize k by picking the one with the best evaluation measure.

- To choose k, run k-nn a few times, changing k, and checking the evaluation metric each time.
- When you have binary classes like "high credit" or "low credit," picking k to be an odd number can be a good idea because there will always be a majority vote, no ties. If there is a tie, the algorithm just randomly picks.

```
# we'll loop through and see what the misclassification rate
# is for different values of k
for (k in 1:20) {
   print(k)
   predicted.labels <- knn(train, test, cl, k)
   # We're using the R function knn()
   num.incorrect.labels <- sum(predicted.labels != true.labels)
   misclassification.rate <- num.incorrect.labels /
                             num.test.set.labels

      print(misclassification.rate)
   }
```

Here's the output in the form (k, misclassification rate):

```
k  misclassification.rate
1, 0.28
2, 0.315
3, 0.26
4, 0.255
5, 0.23
6, 0.26
7, 0.25
8, 0.25
9, 0.235
10, 0.24
```

- So let's go with k =5 because it has the lowest misclassification rate, and now k=5 can be applied to the guy who is 57 with a $37,000 salary.

```
> test <- c(57,37)
> knn(train,test,cl, k = 5)
[1] low
```

The output by majority vote is a low credit score when k = 5.

- The k-NN algorithm is an example of a nonparametric approach. It operates without modeling assumptions about the underlying data-generating distributions and does not involve estimating any parameters.

  But still there were some assumptions:

- Data is in some feature space where a notion of "distance" makes sense.
- Training data has been labeled or classified into two or more classes.
- You pick the number of neighbors to use, k.
- The assumption is that the observed features and labels are associated. The evaluation metric will show the algorithm's labeling performance. Adding features and tuning *k* can improve the model, but there's a risk of overfitting.

- Both linear regression and k-NN are examples of "supervised learning," where you've observed both x and y, and you want to know the function that brings x to y.

Questions:

1.Explain the kNN with example.

2.Explain the overview of kNN process in detail

3.Explain the evaluation metric process in kNN

4. Explain the Training and test Phases of kNN

Handouts for Session 8: k-means Algorithm

2.10 k-means Algorithm

- K-means is an unsupervised learning technique that aims to define the correct answer by identifying clusters within the data.Consider some user level data and assume each row of your dataset corresponds to a user as follows: age, gender, income, state, household, size.
- The goal is to segment users, a process known as segmenting, stratifying, grouping, or clustering the data. All these terms refer to finding similar types of users and grouping them together.
- To see why an algorithm like this might be useful, let's bucket users using handmade thresholds. You may have 10 age buckets, 2 gender buckets, and so on, which would result in $10 \times 2 \times 50 \times 10 \times 3 = 30{,}000$ possible bins, which is big. Moreover, this data existing in a five-dimensional space where each axis corresponds to one attribute. Each user would then live in one of those 30,000 five-dimensional cells. This makes it impossible to build a different marketing campaign for each bin.

- This is where k-means comes into picture where, k-means is: a clustering algorithm where k is the number of bins. k-means algorithm looks for clusters in d dimensions, where d is the number of features for each data point.

2.10.1 k-means algorithm

Algorithm

Randomly pick k centroids (or points that will be the center of your clusters) in d-space, ensuring they are near the data but distinct from one another.

- Assign each data point to the closest centroid.
- Move the centroids to the average location of the data points assigned to them.
- Repeat the previous two steps until the assignments don't change or change very little.

- One has to determine if there's a natural way to describe these groups once the algorithm completes. At times, you may need to make slight adjustments to k a few times before obtaining natural groupings. This is an example of unsupervised learning because the labels are not known and are instead discovered by the algorithm.

K-means has known issues:

1.Choosing k is more art than science, bounded by $1 \leq k \leq n$, where n is the number of data points.

2.Convergence issues may arise, with the algorithm potentially falling into a loop and failing to find a unique solution.

Interpretability can be problematic, with results sometimes being unhelpful or unclear.

K-means advantages:

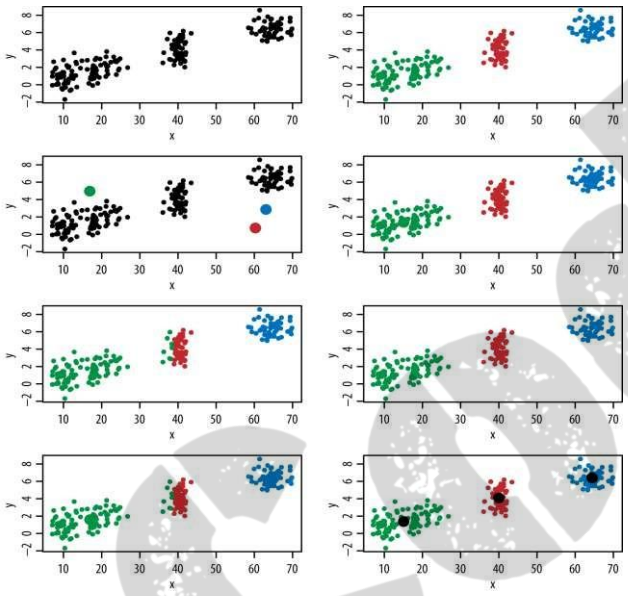
1.k-means is pretty fast (compared to other clustering algorithms),

2.There are broad applications in marketing, computer vision (partitioning an image).

3.Can be a starting point for other models.

2D version:

- Consider a simpler example than the five-dimensional one previously discussed. Suppose there is data on users, including the number of ads shown to each user (impressions) and the number of times each user clicked on an ad (clicks).
- Clustering in two dimensions; look at the panels in the left column from top to bottom, and then the right column from top to bottom.
- In practice, k-means is just one line of code in R:



- Clustering in two dimensions; look at the panels in the left column from top to bottom, and then the right column from top to bottom.
- In practice, k-means is just one line of code in R:


```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                     "MacQueen"))
```

- The kmeans function in R has several parameters:
- **x:** The dataset, which should be a numeric matrix or data frame where each row represents an observation and each column represents a feature.
- **centers:** Specifies the number of clusters (k) to create or the initial cluster centers. If an integer, it indicates the number of clusters; if a matrix, each row represents an initial cluster center.
- **iter.max:** The maximum number of iterations allowed. The default is 10. This parameter controls how long the algorithm will run before stopping.

- **nstart**: The number of random sets of initial cluster centers. The algorithm will run nstart times with different initial centers and return the best solution based on the total within-cluster sum of squares. The default is 1.
- **algorithm:** Specifies the algorithm to use for clustering. Options include:
    - **"Hartigan-Wong" (default):** The standard algorithm proposed by Hartigan and Wong.
    - **"Lloyd":** Also known as the standard k-means algorithm.
    - **"Forgy":** Another version of the k-means algorithm.
    - **"MacQueen":** A variation of the k-means algorithm.

  Example:
- kmeans(x, centers = 3, iter.max = 10, nstart = 1, algorithm = "Hartigan-Wong")
- This example runs the k-means clustering algorithm on the dataset x, aiming to create 3 clusters, with a maximum of 10 iterations, using 1 random start, and employing the Hartigan-Wong algorithm.

Questions:

1.Explain K-means Algorithm in detail.

2.Explain the advantages and disadvantages of K-means

Module 3- Feature Generation and Feature Selection

Extracting Meaning from Data

Module 3 Syllabus	**Feature Generation and Feature Selection** **Extracting Meaning from Data:** Motivating application: user (customer) retention. Feature Generation (brainstorming, role of domain expertise, and place for imagination), Feature Selection algorithms. Filters; Wrappers; Decision Trees; Random Forests. Recommendation Systems: Building a User-Facing Data Product, Algorithmic ingredients of a Recommendation Engine, Dimensionality Reduction, Singular Value Decomposition, Principal Component Analysis, Exercise: build your own recommendation system.

Handouts for Session 1: Motivating Application: User Retention

3.1 Motivating application: user (customer) retention

- Suppose an app called Chasing Dragons charges a monthly subscription fee, with revenue increasing with more users.
- However, only 10% of new users return after the first month.
- **To boost revenue**, **there are two options**: **increase the retention rate of existing users** or **acquire new ones.**
- Generally, retaining existing customers is cheaper than acquiring new ones.
- Focusing on retention, a model could be built to predict if a new user will return next month based on their behavior this month.
- This model could help in providing targeted incentives, such as a free month, to users predicted to need extra encouragement to stay.
- **A good crude model:** Logistic Regression – Gives the probability the user returns their second month conditional on their activities in the first month.
- User behavior is recorded for the first 30 days after sign-up, logging every action with timestamps: for example, a user clicked "level 6" at 5:22 a.m., slew a dragon at 5:23 a.m., earned 22 points at 5:24 a.m., and was shown an ad at 5:25 a.m. This phase involves collecting data on every possible user action.
- User actions, ranging from thousands to just a few, are stored in timestamped event logs.
- These logs need to be processed into a dataset with rows representing users and columns representing features. This phase, known as **feature generation**, involves brainstorming potential features **without being selective**.

- The data science team, including game designers, software engineers, statisticians, and marketing experts, collaborates to identify relevant features.

  Here are some examples:

  - ✓ Number of days the user visited in the first month
  - ✓ Amount of time until second visit
  - ✓ Number of points on day j for j=1, . . .,30 (this would be 30 separate features)
  - ✓ Total number of points in first month (sum of the other features)
  - ✓ Did user fill out Chasing Dragons profile (binary 1 or 0)
  - ✓ Age and gender of user
  - ✓ Screen size of device
- Notice there are redundancies and correlations between these features; that's OK.

- To **construct a logistic regression model** predicting user return behavior, the initial focus lies in attaining a functional model before refinement. Irrespective of the subsequent time frame, classification ci=1 designates a returning user. The logistic regression formula targeted is:

$$\text{logit}\left(P\left(c_i = 1 \middle| x_i\right)\right) = \alpha + \beta^{\tau} \cdot x_i$$

- Initially, a comprehensive set of features is gathered, encompassing user behavior, demographics, and platform interactions.
- Following data collection, **feature subsets must be refined for optimal predictive power** during model scaling and production.
- Three main methods guide feature subset selection: **filters**, **wrappers**, and **embedded methods**.
- Filters independently evaluate feature relevance, wrappers use model performance to assess feature subsets, and embedded methods incorporate feature selection within model training.

Questions

1. Define Customer Retention?
2. What are different relevant features of the Chasing Dragon app?
3. How to boost the revenue of Chasing Dragon application.

Handouts for Session 2: Feature Generation or Feature Extraction

3.2 Feature Generation or Feature Extraction

Feature generation, **also known as feature extraction**, is the process of transforming raw data into a structured format ***where each column represents a specific characteristic or attribute (feature) of the data***, and ***each row represents an observation or instance***.

- This involves identifying, creating, and selecting meaningful variables from the raw data that can be used in machine learning models to make predictions or understand patterns.
- This process is both an art and a science. Having a domain expert involved is beneficial, but using creativity and imagination is equally important.
- Remember, **feature generation is constrained by two factors**: the feasibility of capturing certain information and the awareness to consider capturing it.

Information can be categorized into the following buckets:

- ***Relevant and useful, but it's impossible to capture it.***
  Keep in mind that much user information isn't captured, like free time, other apps, employment status, or insomnia, which might predict their return. Some captured data may act as proxies for these factors, such as playing the game at 3 a.m. indicating insomnia or night shifts.
- ***Relevant and useful, possible to log it, and you did.***
  The decision to log this information during the brainstorming session was crucial. However, mere logging doesn't guarantee understanding its relevance or usefulness. The feature selection process aims to uncover this information.
- ***Relevant and useful, possible to log it, but you didn't.***
  Human limitations can lead to overlooking crucial information, emphasizing the need for creative feature selection. Usability studies help identify key user actions for better feature capture.
- ***Not relevant or useful, but you don't know that and log it.***
  Feature selection aims to address this: while you've logged certain information, unknowing its necessity.
- ***Not relevant or useful, and you either can't capture it or it didn't occur to you.***

Feature Generation or Feature Extraction.

Questions:

1.Define Feature Generation.

2.Define Feature Extraction.

3.Define Feature Generation. Explain how information can be categorized in feature generation in detail.

Handouts for Session 3: Feature Selection: Filters and Wrappers

3.3 Feature Selection Algorithms

Feature selection involves identifying the most relevant and informative features from a dataset for building predictive models.

1. Filters:

- ✓ Filters prioritize features based on specific metrics or statistics, such as correlation with the outcome variable, offering a quick overview of predictive power of .
- ✓ However, ***they may ignore redundancy*** and ***fail to consider feature interactions***, potentially resulting in correlated features and limited insight into complex relationships.
- ✓ By treating the features as independent, it is not taking into account possible interactions or correlation between features.
- ✓ However in some cases 2 redundant features can be more powerful when used together and appear useless when considered individually.

2. Wrappers

- ✓ Wrapper feature selection explores subsets of features of a predetermined size, seeking to identify combinations that optimize model performance.
- ✓ However, as the number of potential combinations grows exponentially with the number of features.

  The number of possible size k subsets of n things, called $\binom{n}{k}$.
- ✓ This exponential growth of possible feature subsets can lead to overfitting.

- ✓ In **wrapper** feature selection, **two key aspects** require consideration:
- ✓ **first**, ***the choice of an algorithm for feature selection***, and
- ✓ **second**, ***the determination of a selection criterion or filter*** to find out the usefulness of the chosen feature set.

A. Selecting an algorithm

Stepwise regression is a category of feature selection methods which involves systematically adjusting feature sets within regression models, typically through **forward selection**, **backward elimination**, or **a combined approach**, to optimize model performance based on predefined selection criteria.

Forward selection:

Forward selection *involves* ***systematically adding features to a regression model one at a time*** ***based on*** ***their ability to improve model performance according to a selection criterion****. This iterative process* ***continues until further feature additions no longer enhance the model's performance****.*

Backward elimination:

Backward elimination *begins with* ***a regression model*** ***containing all features.*** *Subsequently,* ***one feature is systematically removed at a time,*** *the feature whose removal makes the biggest improvement in the selection criterion****. Stop removing features*** ***when removing the feature makes the selection criterion get worse.***

Combined approach:

The ***combined approach*** *in feature selection* ***blends forward selection and backward elimination*** *to strike a balance between* ***maximizing relevance*** ***and minimizing redundancy****. It* ***iteratively adds and removes features based on their significance and impact on model fit,*** *resulting in a subset of features optimized for predictive power.*

B. Selection criterion

The choice of selection criteria in feature selection methods may seem arbitrary. To address this, experimenting with various criteria can help assess model robustness. Different criteria may yield diverse models, necessitating the prioritization of optimization goals based on the problem context and objectives.

R-squared

R-squared can be interpreted as the proportion of variance explained by your model.

$$R^2 = 1 - \frac{\Sigma i \left(yi - \widehat{yi}\right)^2}{\Sigma i \left(yi - \bar{y}\right)^2}$$

p-values

In regression analysis, the interpretation of p-values involves assuming a null hypothesis where the coefficients (βs) are zero. A low p-value suggests that observing the data and obtaining the estimated coefficient under the null hypothesis is highly unlikely, indicating a high likelihood that the coefficient is non-zero.

AIC (Akaike Information Criterion)

Given by the formula **2k−2ln(L)** , where k is the number of parameters in the model and **ln(L)** is the "maximized value of the log likelihood." **The goal is to minimize AIC.**

BIC (Bayesian Information Criterion)

Given by the formula **k*ln(n) −2ln(L)**, where k is the number of parameters in the model, n is the number of observations (data points, or users), and ln(L) is the maximized value of the log likelihood. The goal is to minimize BIC.

Entropy

Entropy is a measure of disorder or impurity in the given dataset.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Questions:

1.Define Feature Selection.

2.Explain Filter Method

3.Explain Wrapper Method.

4.Explain selecting an algorithm in wrapper method.

5.Explain the different Selecting Criteria in feature selection.

Handouts for Session 4: Decision tree

3. Embedded Methods: Decision Trees

i)Decision Trees

- ✓ Decision trees are a popular and powerful tool used in various fields such as machine learning, data mining, and statistics. They provide a clear and intuitive way to make decisions based on data by modelling the relationships between different variables.
- ✓ A decision tree is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions.
- ✓ Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test, and each leaf node corresponds to a class label or a continuous value. In the context of a data problem, a decision tree is a classification algorithm.
- ✓ For the Chasing Dragons example, you want to classify users as “Yes, going to come back next month” or “No, not going to come back next month.” This isn’t really a decision in the colloquial sense, so don’t let that throw you.
- ✓ You know that the class of any given user is dependent on many factors (number of dragons the user slew, their age, how many hours they already played the game). And you want to break it down based on the data you’ve collected. But how do you construct decision trees from data and what mathematical properties can you expect them to have?

- ✓ But you want this tree to be based on data and not just what you feel like. Choosing a feature to pick at each step is like playing the game 20 Questions really well. You take whatever the most informative thing is first. Let's formalize that—we need a notion of "informative."
- ✓ For the sake of this discussion, assume we break compound questions into multiple yes-or-no questions, and we denote the answers by "0" or "1." Given a random variable X, we denote by p(X)=1 and p(X) =0 the probability that X is true or false, respectively.

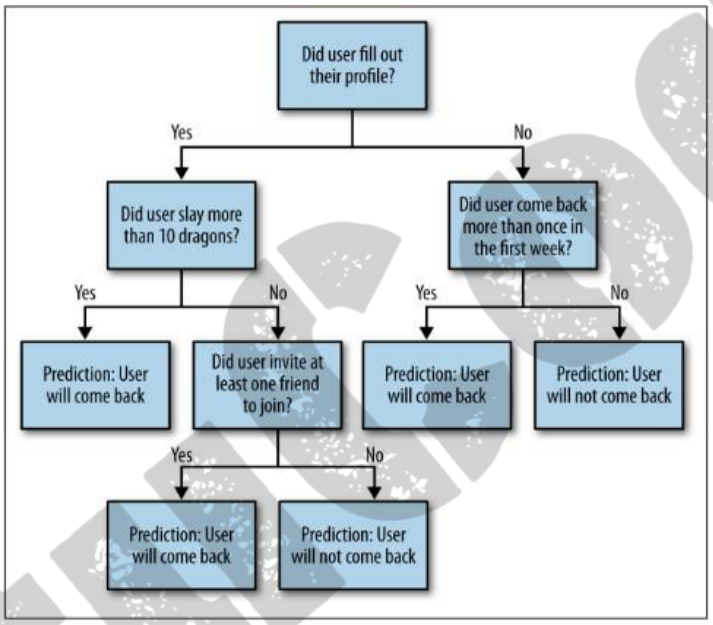


Figure 7-4. Decision tree for Chasing Dragons

Entropy

- Entropy is a measure of disorder or impurity in the given dataset.
- In the decision tree, messy data are split based on values of the feature vector associated with each data point.
- With each split, the data becomes more homogenous which will decrease the entropy. However, some data in some nodes will not be homogenous, where the entropy value will not be small.The higher the entropy, the harder it is to draw any conclusion.

- When the tree finally reaches the terminal or leaf node maximum purity is added.
- For a dataset that has C classes and the probability of randomly choosing data from class, i is Pi. Then entropy E(S) can be mathematically represented as

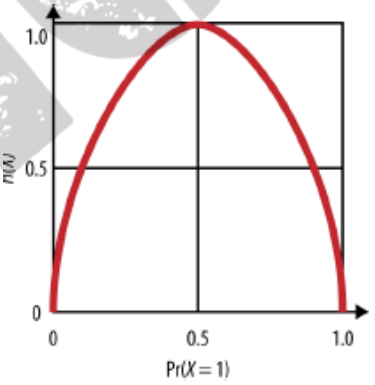
$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

- To quantify what is the most "informative" feature, we define entropy– effectively a measure for how mixed up something is—for X

$$H(X) = - p(X =1) \log_2(p(X =1)) - p(X =0) log_2(p(X =0))$$

when p(X =1)= 0 or p (X =0) =0, the entropy vanishes i.e. if either option has probability zero, the entropy is 0(pure).As p(X =1) =1− p(X =0) , the entropy is symmetric about 0.5 and maximized at 0.5.

- In particular, if either option has probability zero, the entropy is 0. Moreover, because p (X =1) =1− p (X =0) , the entropy is symmetric about 0.5 and maximized at 0.5, which we can easily confirm using a bit of calculus. The Below Figure shows the picture of that.



- **Example:**

If we have a dataset of 10 observations belonging to two classes YES and NO. If 6 observations belong to the class, YES, and 4 observations belong to class NO, then entropy can be written as below.

$$E(S) = -(P_{yes}\log_2 P_{yes} + P_{no}\log_2 P_{no})$$

Pyes is the probability of choosing Yes and *Pno* is the probability of choosing a No. Here *Pyes* is 6/10 and *Pno* is 4/10.

$$E(S) = -(6/10 * \log_2 * 6/10 + 4/10 * \log_2 * 4/10) \approx 0.971$$

If all the 10 observations belong to 1 class then entropy will be equal to zero. Which implies the node is a pure node.

$$E(S) = -(1 \log_2 1) = 0$$

If both classes YES and NO have an equal number of observations, then entropy will be equal to 1.

$$E = -(\tfrac{5}{10} * \log_2 \tfrac{5}{10} + \tfrac{5}{10} * \log_2 \tfrac{5}{10}) = -2(0.5 \log_2 0.5) = 1$$

- Entropy is a measurement of how mixed up something is.If X denotes the event of a baby being born a boy, the expectation is to be true or false with probability close to 1/2, which corresponds to high entropy, i.e., the bag of babies from which we are selecting a baby is highly mixed.
- If X denotes the event of a rainfall in a desert, then it's low entropy.In other words, the bag of day-long weather events is not highly mixed in deserts.
- X is the target of our model. So, X could be the event that someone buys something on our site.
- Which attribute of the user will tell us the most information about this event X needs to be determined
- Information Gain, IG(X,a), for a given attribute a, is the entropy we lose(reduction in entropy, or uncertainty) if we know the value of that attribute
- IG (X,a) =H(X) − H (X|a).
- H(X|a) can be computed in 2 steps:
- For any actual value a_0 of the attribute a we can compute the specific conditional entropy H(X|$a = a_0$) as:
- H(X|$a = a_0$) = -p(X=1| $a = a_0$) log_2(p(X=1| $a = a_0$))
  - p(X=0| $a = a_0$) log_2(p(X=0| $a = a_0$))

  And then we can put it all together, for all possible values of a, to getthe conditional entropy H(X|a) :

  H(X|a)=$\sum_{a_i} p(a = a_i)$. H(X|$a = a_i$)

Decision Tree Algorithm

- The Decision Tree is built iteratively.

- It starts with the root. - You need an algorithm to decide which attribute to split on; e.g., which node should be the next one to identify.
- The attribute is chosen in order to maximize information gain.
- Keep going until all the points at the end are in the same class or you end up with no features left. In this case, you take the majority vote.
- The Tree can be pruned to avoid overfitting. - cutting it off below a certain depth.
- If you build the entire tree, it's often less accurate with new data than if you prune it.

- **Example:**

  - Suppose you have your Chasing Dragons dataset. Your outcome variable is Return: a binary variable that captures whether or not the user returns next month, and you have tons of predictors.

```
# Load necessary library

#Loads the rpart library, which is used for recursive partitioning and regression trees.

library(rpart)

# Read the CSV file

chasingdragons <- read.csv("chasingdragons.csv")

setwd("F:/college/data science-2021 scheme/dscience")

# Grow the classification tree

#Builds a classification tree model to predict the Return variable using the other variables (profile, num_dragons, num_friends_invited, gender, age, num_days) as predictors. The method="class" specifies that this is a classification tree.

model1 <- rpart(Return ~ profile + num_dragons + num_friends_invited + gender + age + num_days, method="class", data=chasingdragons)
```

Display the results

#Prints the complexity parameter (CP) table, which helps in understanding the performance of the model and in pruning the tree.

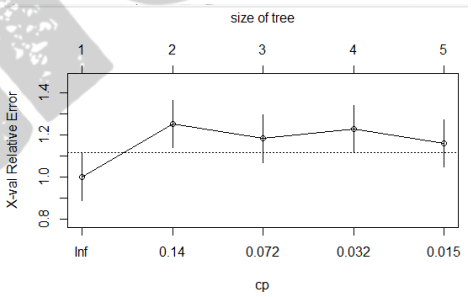
```
printcp(model1)
```

```
Variables actually used in tree construction:
[1] gender             num_days           num_friends_invited

Root node error: 44/100 = 0.44

n= 100

        CP nsplit rel error xerror    xstd
1 0.181818      0   1.00000 1.0000 0.11282
2 0.113636      1   0.81818 1.1591 0.11361
3 0.045455      2   0.70455 1.1136 0.11361
4 0.022727      3   0.65909 1.1136 0.11361
5 0.010000      4   0.63636 1.1818 0.11355
```

Visualize cross-validation results

plot of the cross-validation results for the complexity parameter. The plotcp function helps visualize how the model's error changes with the complexity of the tree, which aids in selecting the optimal tree size.

```
plotcp(model1)
```



Detailed summary of the model

#The summary function provides comprehensive information about the model, including splits, nodes.

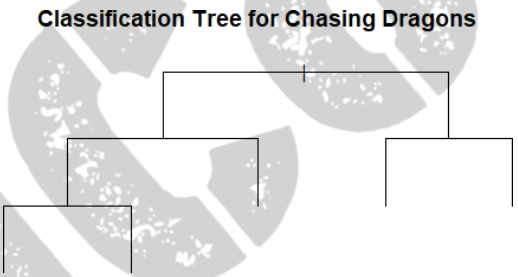
```
summary(model1)
```

#sample summary output:

```
Variable importance
num_friends_invited              gender            num_days                 age
                 41                  27                  15                   9
        num_dragons
                  7

Node number 1: 100 observations,    complexity param=0.1818182
  predicted class=No   expected loss=0.44  P(node) =1
    class counts:    56    44
   probabilities: 0.560 0.440
  left son=2 (62 obs) right son=3 (38 obs)
  Primary splits:
```

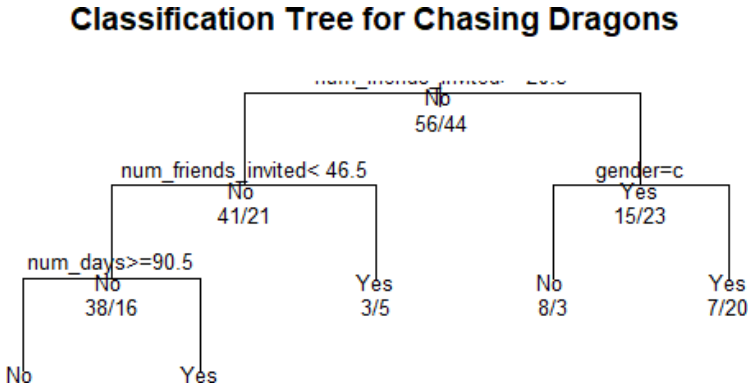
Plot the classification tree.The plot function visualizes the tree structure, and the uniform=TRUE argument makes the branch lengths uniform. The main argument specifies the title of the plot.

```
plot(model1, uniform=TRUE, main="Classification Tree for Chasing Dragons")
```



Add text labels to the tree plot. The text function annotates the tree plot with information about the nodes. The use.n=TRUE argument includes the number of observations at each node, all=TRUE ensures all nodes are labeled, and cex=.8 sets the size of the text labels.

```
text(model1, use.n=TRUE, all=TRUE, cex=.8)
```



Handling Continuous Variables

- In case of continuous variables, the correct threshold of a value needs to be determined to consider it as a binary variable.
- Example: A User's number of Dragon Slays can be partitioned into categories such as "less than 10" and "at least 10". Now we have a binary variable case.
- In this case, it takes some extra work to decide on the information gain because it depends on the threshold as well as the feature.
- Instead of a single threshold, bins of values can be created for the attribute. – Depends on situation

Questions:

1. Define Decision tree.
2. Explain Decision Tree for Chasing Dragons Problem.
3. Suppose you have your Chasing Dragons dataset. Your outcome variable is Return: a binary variable that captures whether or not the user returns next month, and you have tons of predictors. Write a R Script using decision tree algorithm for the above scenario.
4. Write Decision Algorithm in detail.

Handouts for Session 5: Random forest

ii)Random Forest

- Random forests generalize decision trees with bagging,otherwise known has Bootstrap Aggregating.

- Makes the models more accurate and more robust, but at the cost of interpretability
- But easy to specify – 2 Hyperparameters: Number of Trees (N) in the forest and Number of Features (F) to randomly select for each tree
- A bootstrap sample is a sample with replacement, which means we might sample the same data point more than once.We usually take to the sample size to be 80% of the size of the entire (training) dataset, but of course this parameter can be adjusted de-pending on circumstances. This is technically a third hyper parameter of our random forest algorithm.
- To construct a random forest, you construct N decision trees as follows:
  - For each tree, take a bootstrap sample of your data, and for each node you randomly select F features, say 5 out of the 100 total features.
  - Then you use your entropy-information-gain engine as described in the previous section to decide which among those features you will split your tree on at each stage.
- Note that you could decide beforehand how deep the tree should get,or you could prune your trees after the fact, but you typically don't prune the trees in random forests, because a great feature of random forests is that they can incorporate idiosyncratic noise.

- **Algorithm**
  - Select random K data points from the Training Set
  - Build a Decision Tree based on the selected K points
  - Select the Number of Trees to build and repeat Steps 1 & 2
  - For a new data point (test data point) makes each of the trees predict the class for the data point. And assign the new data point the average across all the predicted classes

Questions

1.Define Random Forest.

Explain Random forest Algorithm.

Handouts for Session 6: User Retention

3.4 User Retention: Interpretability Vs. Predictive Power

- Assume a model predicts well, but can you find the meaning in the model?
- Example: The prediction could be "the more the user plays in the first month, the more likely the user is to come back next month". This is obvious and not very helpful when doing the analysis
- It could also tell that showing them ads in the first 5 minutes decreases their chances of coming back, buts its ok to show ads after the first hour. This would give an insight not to show ads in the first 5 minutes
- To study this more, you really would want to do some A/B testing, but this initial model and feature selection would help you prioritize the types of tests you might want to run.
- Features that are associated with the user's behaviour are qualitatively different from the features associated with one's own behaviour.
- If there's a correlation of getting a high number of points in the first month with players returning to play next month, does that mean if you just give users a high number of points this month without them playing at all, they'll come back – No!
- It's not the number of points that caused them to come back, it's that they're really into playing the game which correlates with both their coming back and their getting a high number of points.
- Therefore, do feature selection with all variables, but then focus on the ones you can do something about conditional on user attributes.

Questions:

1.Explain the User Retention in Detail.

Handouts for Session 7: User Retention, Dimensionality Reduction, SVD

3.5 A Real-World Recommendation Engine

- Recommendation engines are used all the time—what movie would you like, knowing other movies you liked? What book would you like, keeping in mind past purchases?
- There are plenty of different ways to go about building such a model, but they have very similar feels if not implementation. To set up a recommendation engine, suppose you have users, which form a set U; and you have items to recommend, which form a set V.
- We can denote this as a bipartite graph (shown again in Figure below) if each user and each item has a node to represent it—there are lines from a user to an item if that user has expressed an opinion about that item.
- Note they might not always love that item, so the edges could have weights: they could be positive, negative, or on a continuous scale (or discontinuous, but many-valued like a star system).
- The implications of this choice can be heavy but we won't delve too deep here for us they are numeric ratings.
- Next up, you have training data in the form of some preferences—you know some of the opinions of some of the users on some of the items. From those training data, you want to predict other preferences for your users. That's essentially the output for a recommendation engine.
- You may also have metadata on users (i.e., they are male or female, etc.) or on items (the color of the product). For example, users come to your website and set up accounts, so you may know each user's gender, age, and preferences for up to three items.
- You represent a given user as a vector of features, sometimes including only metadata—sometimes including only preferences (which would lead to a sparse vector because you don't know all the user's opinions) and sometimes including both, depending on what you're doing with the vector.
- Also, you can sometimes bundle all the user vectors together to get a big user matrix, which we call U, through abuse of notation.

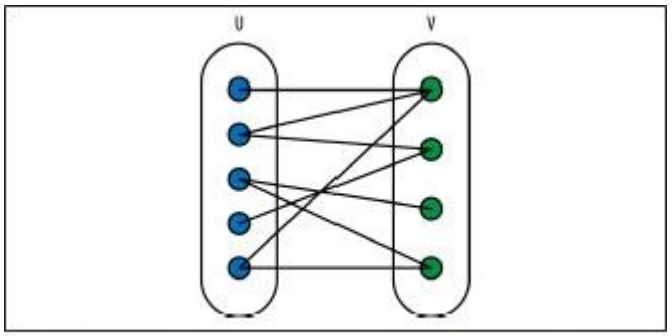


Figure 8-1. Bipartite graph with users and items (television shows) as nodes

3.6 The Dimensionality Problem

- As we increase the dimension (Number of features), the accuracy of the system increases up to a certain limit. Beyond the limit, the accuracy starts to decline.
- Solution - **Dimensionality Reduction:** This does not mean removing features, but rather transform the data into a different perspective.
- Our goal is to build a model that has a representation in a low dimensional subspace that gathers "taste information" to generate recommendations. So we're saying here that taste is latent but can be approximated by putting together all the observed information we do have about the user.

3.6.1 Singular Value Decomposition (SVD)

- **Rank of a Matrix:** The rank of a matrix A is the maximum number of linearly independent row vectors or column vectors in the matrix. It is denoted as rank(A). In case of SVD the rank of A is the number of non-zero singular values in the diagonal matrix S.
- **Latent Features:** These are the hidden patterns or factors in the data. For a user-item matrix, they represent underlying user preferences and item characteristics.
- Importance of Latent Features: The singular values in S indicate the importance of these features.
- A larger singular value means the corresponding latent feature is more significant in capturing the structure of the data.
- Given an m×n matrix X of rank k: $\mathbf{X=USV^T}$

- **U**: m×k matrix (Left Singular Vectors) that contains user latent features. Each row corresponds to a user, and each column represents a latent feature. For example, a row might capture a user's preference for genres like action or comedy.
- **S**: k×k diagonal matrix with singular values. These values indicate the importance of each latent feature. Larger values correspond to more significant features.
- **V:** k×n matrix (Right Singular Vectors) Contains item latent features. Each row corresponds to an item, and each column represents a latent feature. For example, a row might capture an item's characteristics like genre or popularity.
- **U and V:** The columns of U and V are orthogonal, meaning they capture independent latent features.
- **S:** The singular values in S measure the importance of each latent feature. Larger values indicate more significant features.

Dimensionality Reduction with SVD

To reduce the dimensionality, follow these steps:

1. Compute SVD:
    - Perform SVD on the matrix X to get U, S, and V.
2. Select Top d Singular Values:
    - Identify the top d largest singular values from S. These singular values correspond to the most significant components of the data.
3. Construct Reduced Matrices:
    - Construct reduced matrices U_d, S_d, and V_d
4. Approximate the Original Matrix:

$$X_d = U_d S_d V_d^T$$

- This X_d is the best approximation of X, using only the top d singular values.

In Singular Value Decomposition (SVD), dimensionality reduction is achieved by selecting the most significant singular values and their corresponding singular vectors.

This process leverages the properties of the SVD to capture the most important structures in the data while discarding less important information.

Questions:

1.Explain Real world recommendation engine with neat diagram.

2.What is Dimensionality Problem.

3.Explain SVD in detail

Handouts for Session 8: PCA,Building Recommendation Engine

3.6.2 Principal Component Analysis

- In this approach, we aim to predict preferences by factorizing the **user-item interaction matrix X** into **two lower-dimensional matrices, U and V**, without the need for the singular values matrix S. The goal is to find U and V such that:

$$X \approx U \cdot V^T$$

- The optimization problem is to **minimize the discrepancy between the actual user-item interaction matrix X and its approximation**

$$\tilde{X} U \cdot V^T.$$

- This discrepancy is measured using the squared error:

$$\text{argmin} \sum_{i,j} (x_{ij} - u_i \cdot v_j)^2$$

- x_{ij} is the actual interaction (e.g., rating) between **user i** and **item j**.
- u_i is the i-th row of matrix U, representing the **latent features of user i.**
- v_j is the j-th row of matrix V, representing the **latent features of item j.**
- The **dot product** $u_i \cdot v_j$ is the **predicted preference of user i for item j.**

Latent Features and Matrix Dimensions

- **Number of Latent Features (d)**: This is a parameter that you choose, representing the number of latent features you want to use. It controls the dimensions of matrices U and V.
- **Matrix U**: Has dimensions m×d, where m is the number of users and d is the number of latent features. Each row corresponds to a user.

- **Matrix V**: Has dimensions n×d, where n is the number of items and d is the number of latent features. Each row corresponds to an item.

3.7 Alternating Least Squares

- Here we are not first minimizing the squared error and then minimizing the size of the entries of the matrices U and V. Here we are actually doing both at the same time.
- Alternating Least Squares (ALS) is an algorithm for matrix factorization. ALS is used to decompose a given user-item interaction matrix into two lower-dimensional matrices (U and V) that capture latent features of users and items.

- Here's the algorithm:

  Pick a random V.

  Optimize U while V is fixed.

  Optimize V while U is fixed.
- Keep doing the preceding two steps until you're not changing very much at all. To be precise, you choose an ϵ and if your coefficients are each changing by less than ϵ, then you declare your algorithm "converged."
- Fix V and Update U The way you do this optimization is user by user. So for user i, you want to find:

$$argmin_{u_i} \sum_{j \in P_i} (p_{i,j} - u_i * v_j)^2$$

- where v_j is fixed. In other words, you just care about this user for now. But wait a minute, this is the same as linear least squares, and has a closed form solution! In other words, set:

$$u_i = \left(V_{*,i}^T V_{*,i}\right)^{-1} V_{*,i}^T P_{*i},$$

- where V*,i is the subset of V for which you have preferences coming from user i. Taking the inverse is easy because it's d×d, which is small.And there aren't that many

preferences per user, so solving this many times is really not that hard. Overall you've got a doable update for U.

- When you fix U and optimize V, it's analogous—you only ever have to consider the users that rated that movie, which may be pretty large for popular movies but on average isn't; but even so, you're only ever inverting a d×d matrix.

3.8 Building Recommendation System using Python

The following code is Matt's code to illustrate implementing a recommendation system on a relatively small dataset.

Initialize Matrix V:

- V is initialized with random values.
- This matrix represents the latent features of items.

Initialize Matrix U:

- U is initialized with zeros.
- This matrix will represent the latent features of users.

ALS Algorithm

The ALS algorithm alternates between updating the user latent features (U) and the item latent features (V) to minimize the squared error of the predicted ratings.

For each user:

- Extract the items they have interacted with and their corresponding ratings.
- **Create a matrix vo** containing the latent features of these items.
- **Create a vector pvo** of the ratings.

Solve the regularized least squares problem:

- Update U[i, :] using the formula:

$$U[i,:] = (V_i^T V_i + \lambda I)^{-1} V_i^T X_i$$

- `V_i` is the submatrix of `V` corresponding to the items user `i` has rated.
- `X_i` is the vector of ratings for these items.

3. **Regularized Least Squares Solution**:
    - The equation $U[i, :] = \left(V_i^T V_i + \lambda I\right)^{-1} V_i^T X_i$ is used to update `U[i, :]`.

Explanation of the Equation

- $V_i^T V_i$:
    - This is the product of the transpose of `V_i` and `V_i` itself. It captures the correlations between the latent features of the items that user `i` has interacted with.
- λI:
    - I is the identity matrix of size equal to the number of latent features (`num_features`). Multiplying it by the regularization parameter λ adds a small value to the diagonal elements, which helps in preventing overfitting and ensures numerical stability by making the matrix invertible.
- $\left(V_i^T V_i + \lambda I\right)^{-1}$:
    - This is the inverse of the regularized matrix. It essentially normalizes the correlations captured in $V_i^T V_i$.
- $V_i^T X_i$:
    - This is the product of the transpose of `V_i` and `X_i`. It projects the known interaction values into the latent feature space.
- **Combining these**:
    - The equation $U[i, :] = \left(V_i^T V_i + \lambda I\right)^{-1} V_i^T X_i$ provides the best fit for the latent feature vector of user `i` by solving the regularized least squares problem.

Error Calculation

- Calculate the root mean square error (RMSE):Compute the prediction error for all known user-item interactions.
- Final Predicted Matrix
- Predict the entire user-item interaction matrix:
- Multiply U and V.T to get the predicted ratings.

Code:

```
import math
import numpy as np
# Define the user-item ratings matrix pu
pu = [
    [(0, 0, 1), (0, 1, 22), (0, 2, 1), (0, 3, 1), (0, 5, 0)],
    [(1, 0, 1), (1, 1, 32), (1, 2, 0), (1, 3, 0), (1, 4, 1), (1, 5, 0)],
    [(2, 0, 0), (2, 1, 18), (2, 2, 1), (2, 3, 1), (2, 4, 0), (2, 5, 1)],
    [(3, 0, 1), (3, 1, 40), (3, 2, 1), (3, 3, 0), (3, 4, 0), (3, 5, 1)],
    [(4, 0, 0), (4, 1, 40), (4, 2, 0), (4, 4, 1), (4, 5, 0)],
    [(5, 0, 0), (5, 1, 25), (5, 2, 1), (5, 3, 1), (5, 4, 1)]
]
# Define the item-user ratings matrix pv
pv = [
    [(0, 0, 1), (0, 1, 1), (0, 2, 0), (0, 3, 1), (0, 4, 0), (0, 5, 0)],
    [(1, 0, 22), (1, 1, 32), (1, 2, 18), (1, 3, 40), (1, 4, 40), (1, 5, 25)],
    [(2, 0, 1), (2, 1, 0), (2, 2, 1), (2, 3, 1), (2, 4, 0), (2, 5, 1)],
    [(3, 0, 1), (3, 1, 0), (3, 2, 1), (3, 3, 0), (3, 5, 1)],
    [(4, 1, 1), (4, 2, 0), (4, 3, 0), (4, 4, 1), (4, 5, 1)],
    [(5, 0, 0), (5, 1, 0), (5, 2, 1), (5, 3, 1), (5, 4, 0)]
```

```
]
# Define matrix V
V = np.mat([
    [0.15968384, 0.9441198, 0.83651085],
    [0.73573009, 0.24906915, 0.85338239],
    [0.25605814, 0.6990532, 0.50900407],
    [0.2405843, 0.31848888, 0.60233653],
    [0.24237479, 0.15293281, 0.22240255],
    [0.03943766, 0.19287528, 0.95094265]
])
# Initialize matrix U with zeros
U = np.mat(np.zeros([6, 3]))
# Regularization parameter
L = 0.03


# Perform matrix factorization using alternating least squares
for iter in range(5):
    print("\n----- ITER %s ---- " % (iter + 1))
    print("U")
    urs = []


    # Update U
    for uset in pu:
        vo = []
```

```
        pvo = []
        for i, j, p in uset:
            vor = []
            for k in range(3):
                vor.append(V[j, k])
            vo.append(vor)
            pvo.append(p)

        vo = np.mat(vo)
    ur = np.linalg.inv(vo.T * vo + L * np.mat(np.eye(3))) * vo.T * np.mat(pvo).T
        urs.append(ur.T)

    U = np.vstack(urs)
    print(U)

print("V")
    vrs = []

    # Update V
    for vset in pv:
        uo = []
        puo = []
        for j, i, p in vset:
            uor = []
```

```
            for k in range(3):
                uor.append(U[i, k])
            uo.append(uor)
            puo.append(p)

        uo = np.mat(uo)
        vr = np.linalg.inv(uo.T * uo + L * np.mat(np.eye(3))) * uo.T * np.mat(puo).T
        vrs.append(vr.T)

    V = np.vstack(vrs)
    print(V)
# Calculate RMSE (Root Mean Squared Error)
    err = 0.
    n = 0.
    for uset in pu:
        for i, j, p in uset:
            err += (p - (U[i] * V[j].T)[0, 0]) ** 2
            n += 1

    rmse = math.sqrt(err / n)
    print("RMSE:", rmse)

# Print final U * V.T
print("\nFinal U * V.T")
```

```
print(U * V.T)
```

Output:

```
----- ITER 1 -----
U
[[ 27.48856092  -6.39423002   0.76339065]
 [ 36.40898688 -10.96084483   2.19757677]
 [ 19.79884149  -6.62335704   2.23966925]
 [ 44.36457118 -13.91334026   3.61771668]
 [ 46.25668408 -15.11954225   4.75538378]
 [ 24.08497085 -11.99211139   7.2307236 ]]
V
[[ 0.0545325   0.11205172 -0.0332261 ]
 [ 0.4588277  -1.51294995 -0.61387609]
 [ 0.19340572  0.7878437   0.78829   ]
 [ 0.25513015  1.06866178  1.06871888]
 [-0.16891975 -0.7083464  -0.47721002]
 [-0.01489877 -0.04766114  0.12516857]]
0.458261949227786

----- ITER 2 -----
U
[[ 21.24745657  -9.63073906   5.11174585]
 [ 30.23955333 -13.58612407   5.88615241]
 [ 16.56669002  -8.71313163   5.6304904 ]
 [ 38.39283553 -17.04724985   8.01303322]
 [ 36.73866642 -17.32271228   7.40352948]
 [ 21.68875723 -12.66751199   8.14720285]]
V
[[ 0.28051193  0.75036922  0.39661661]
 [ 0.50058836 -1.50093297 -0.59823823]
 [ 0.19972091  0.76013758  0.77453081]
 [ 0.15257727  0.71636732  0.83603449]
 [-0.28061042 -0.89379554 -0.53418132]
 [ 0.05743832  0.32749297  0.48997755]]
0.23261838473656263

----- ITER 3 -----
U
[[ 20.03935133  -9.53918778   5.11576922]
 [ 28.24667145 -13.626873     5.99360507]
 [ 15.41955317  -8.73439494   5.7077762 ]
 [ 35.92368317 -17.13471274   8.32390394]
 [ 34.53667699 -17.24112282   7.36684371]
 [ 20.7631954  -12.19828589   7.49059224]]
V
[[ 0.33533083  0.82398023  0.39028879]
 [ 0.54163697 -1.47535262 -0.56710963]
 [ 0.17854914  0.68262722  0.74839685]
 [ 0.09531085  0.5462863   0.75891467]
 [-0.33339151 -0.98162282 -0.56601839]
 [-0.00402394  0.17911791  0.45039733]]
0.22364368105565205
```

```
U
[[ 19.11565545  -9.36172458   4.96588607]
 [ 26.88500489 -13.38916067   5.67368401]
 [ 14.72121272  -8.51670509   5.42125844]
 [ 34.07500028 -16.94324729   8.14133866]
 [ 32.96525324 -16.89010525   6.88873846]
 [ 19.91764225 -11.81949685   6.94254806]]
V
[[ 3.77508344e-01  8.80568062e-01  3.91782977e-01]
 [ 5.81916099e-01 -1.45004580e+00 -5.38150854e-01]
 [ 1.53422966e-01  5.97339864e-01  7.18293461e-01]
 [ 6.57958249e-02  4.59551458e-01  7.32442207e-01]
 [-3.78045460e-01 -1.05294790e+00 -5.92358399e-01]
 [-8.14718299e-02 -5.40684446e-04  4.05487216e-01]]
0.21919522056558113

----- ITER 5 -----
U
[[ 18.37609424  -9.14504116   4.75023946]
 [ 25.79448179 -13.10478426   5.28381689]
 [ 14.15580472  -8.27945838   5.08762866]
 [ 32.58674446 -16.65571486   7.77775649]
 [ 31.69560339 -16.47980863   6.29323664]
 [ 19.18869022 -11.48480019   6.45928646]]
V
[[ 0.40861253  0.91893055  0.39447878]
 [ 0.62155035 -1.42528866 -0.51155646]
 [ 0.12949727  0.5172765   0.68830999]
 [ 0.05454709  0.42318187  0.73517698]
 [-0.41861857 -1.11464585 -0.61380526]
 [-0.16925641 -0.19571672  0.3603905 ]]
0.21484614865128004
 [[ 0.9789133  22.02597562  0.91877642  0.62461344 -0.41481413  0.391507  ]
  [ 0.58191546 32.00769898  0.19842198 -0.25415279  0.56590965  0.10318157]
  [ 0.18295347 17.99655423  1.05223443  1.00874867  0.17996792  1.05800087]
  [ 1.07806674 40.01474243  0.95779204  0.44714287  0.14977915  0.54731624]
  [ 0.28996936 39.96955196 -0.08842598 -0.61841076  1.23796031  0.12871294]
  [-0.16494307 24.99160291  0.99005714  0.93524659  0.80399875  1.32782419]]
```

Questions:

1. Explain PCA in detail.

2. Define Alternating Least Sqaure.

3. Write a program for Recommendation system using Python.

Module 4- Data Visualization and Data Exploration

Module 4 Syllabus	**Introduction:** Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization **Comparison Plots:** Line Chart, Bar Chart and Radar Chart**; Relation Plots:** Scatter Plot, Bubble Plot, Correlogram and Heatmap; **Composition Plots:** Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram; **Distribution Plots:** Histogram, Density Plot, Box Plot, Violin Plot**; Geo Plots:** Dot Map, Choropleth Map, Connection Map**; What Makes a Good Visualization?** **Textbook 2: Chapter 1, Chapter 2**

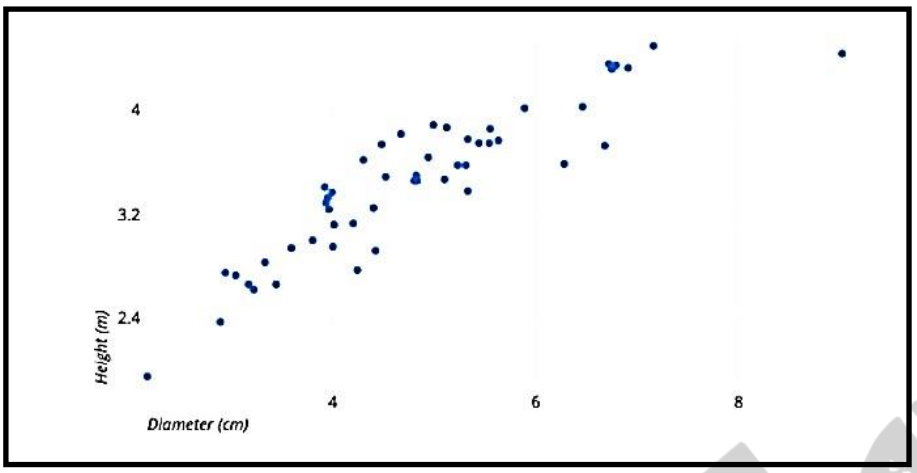
Handouts for Session 1: Introduction: Data Visualization, Importance of Data Visualization

4.1 Introduction to Data Visualization

- Computers and smartphones store names and numbers digitally.
- Data representation involves the forms used to store, process, and transmit data.
- Effective representations convey stories and fundamental discoveries, enhancing the data's value.
- By modeling information properly, we gain clearer, more concise, and understandable insights.
- Representations convert data into useful information, helping to derive meaningful insights.

4.2 The Importance of Data Visualization

- Instead of merely viewing data in Excel columns, visualization helps us better understand our data.
- For instance, visualizing data through charts and graphs can reveal patterns, trends, and insights that are not immediately obvious in raw data form.
- It's easy to see a pattern emerge from the numerical data that's given in the following scatter plot.
- It shows the correlation between diameter and the height of various trees. There is a positive correlation between diameter and height.



Questions

1. Briefly explain Data Visualization.

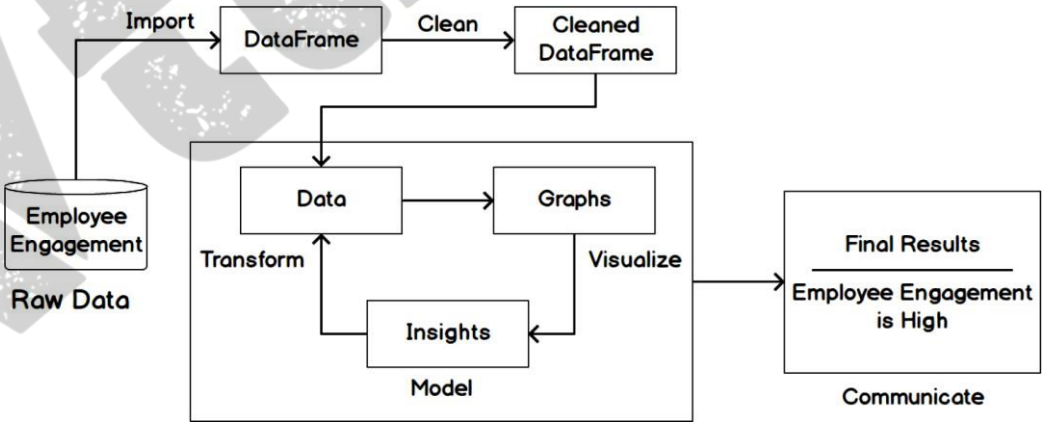
2. Why Data Visualization is Important/Significant?

Handouts for Session 2: Data Wrangling, Tools and Libraries for Visualization

4.3 Data Wrangling

Data wrangling is the process of transforming raw data into a suitable representation for various tasks. It is the discipline of augmenting, cleaning, filtering, standardizing, and enriching data in a way that allows it to be used in a downstream task, which in our case is data visualization.

Examine the following flow diagram of the data wrangling process to understand how precise and actionable data is prepared for business analysts to utilize.



Data wrangling process to measure employee engagement

The following steps explain the flow of the data wrangling process:

1. First, the Employee Engagement data is in its raw form.

2. Then, the data gets imported as a DataFrame and is later cleaned.
3. The cleaned data is then transformed into graphs, from which findings can be derived.
4. Finally, we analyze this data to communicate the final results.

For example, employee engagement can be measured based on raw data gathered from feedback surveys, employee tenure, exit interviews, one-on-one meetings, and so on. This data is cleaned and made into graphs based on parameters such as referrals, faith in leadership, and scope of promotions. The percentages, that is, information derived from the graphs, help us reach our result, which is to determine the measure of employee engagement.

4.4 Tools and Libraries for Visualization

- Several tools are available for creating data visualizations to suit different needs.
- Non-coding tools like Tableau provide an intuitive interface for exploring and understanding data.
- Alongside Python, MATLAB and R are also commonly used in data analytics.
- Python stands out as the industry's preferred language due to its user-friendly nature and efficiency in data manipulation and visualization.
- Its extensive library ecosystem further enhances Python's appeal, making it the optimal choice for robust data visualization tasks.

Questions:

1. What is Data Wrangling?
2. Explain the data wrangling process with an example of employee engagement.
3. With a neat diagram explain the steps involved in the Data Wrangling process.

Handouts for Session 3: Comparison Plots: Line Chart, Bar Chart and Radar Chart

4.5 Comparison Plots

- Comparison plots include charts that are ideal for comparing multiple variables or variables over time.
- **Line charts** are great for visualizing variables over time.

- For comparison among items, **bar charts** (also called column charts) are the best way to go. For a certain time period (say, fewer than 10-time points), vertical bar charts can be used as well.
- **Radar charts** or spider plots are great for visualizing multiple variables for multiple groups.

1. Line Chart

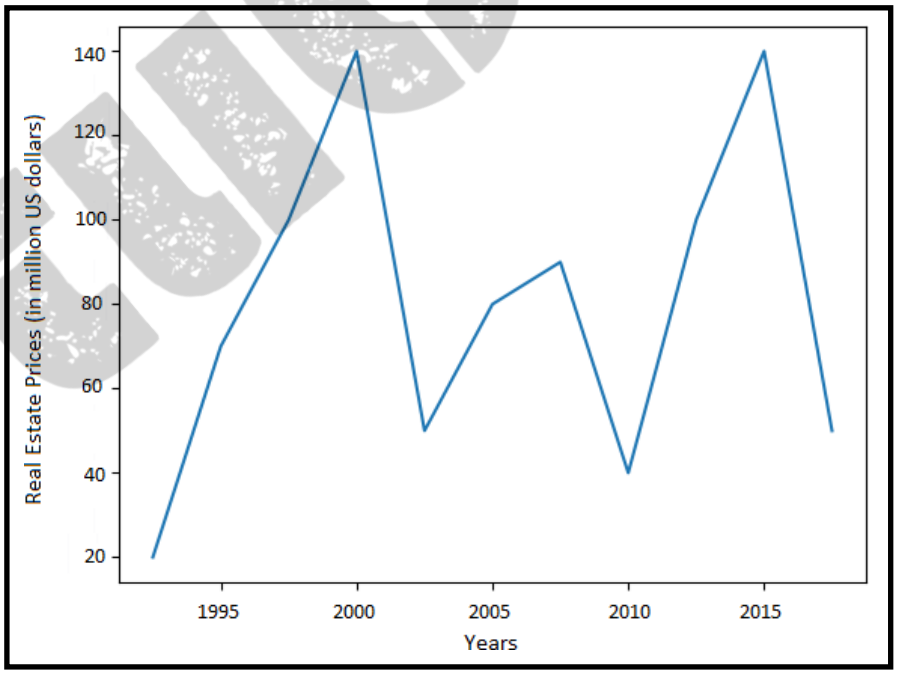
- Line charts are used to display quantitative values over a continuous time period and show information as a series.
- A line chart is ideal for a time series that is connected by straight-line segments.
- The value being measured is placed on the y-axis, while the x-axis is the timescale.

Uses

✓ Line charts are great for comparing multiple variables and visualizing trends for both single as well as multiple variables, especially if your dataset has many time periods (more than 10).

✓ For smaller time periods, vertical bar charts might be the better choice.

Example 1: The following diagram shows a trend of real estate prices (per million US dollars) across two decades. Line charts are ideal for showing data trends:



Line chart for a single variable

Example 2: The following figure is a multiple-variable line chart that compares the stock-closing prices for Google, Facebook, Apple, Amazon, and Microsoft. A line chart is great for comparing values and visualizing the trend of the stock. As we can see, Amazon shows the highest growth:

Line chart showing stock trends for five companies

Design Practices

- ✓ Avoid too many lines per chart.
- ✓ Adjust your scale so that the trend is clearly visible.

2. Bar Charts

- In a bar chart, the bar length encodes the value. There are two variants of bar charts: **vertical bar charts** and **horizontal bar charts.**

Uses

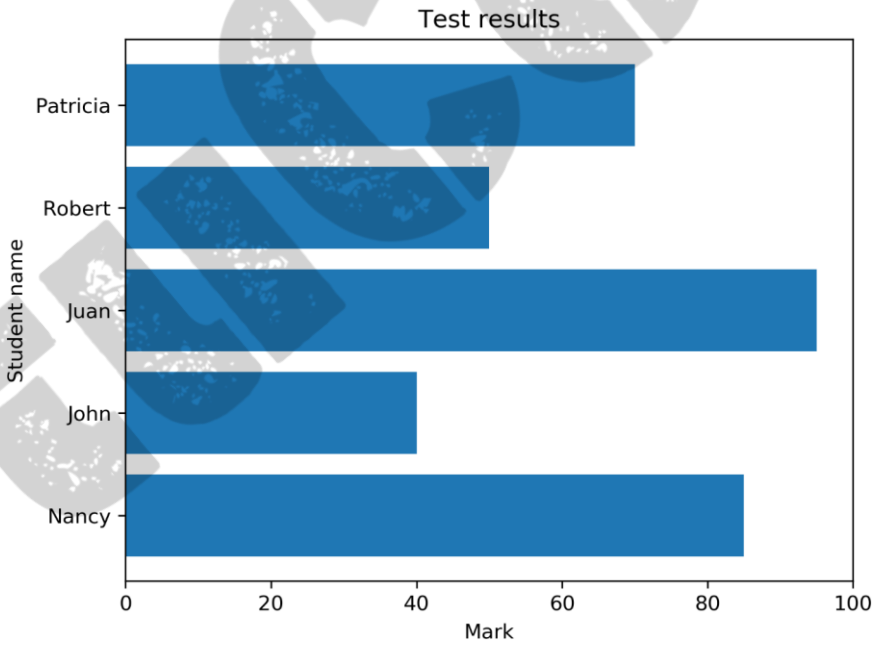
- While they are both used to compare numerical values across categories, vertical bar charts are sometimes used to show a single variable over time.

Example 1:

The following diagram shows a vertical bar chart. Each bar shows the marks out of 100 that 5 students obtained in a test:

Vertical bar chart using student test data

The following diagram shows a horizontal bar chart. Each bar shows the marks out of 100 that 5 students obtained in a test:

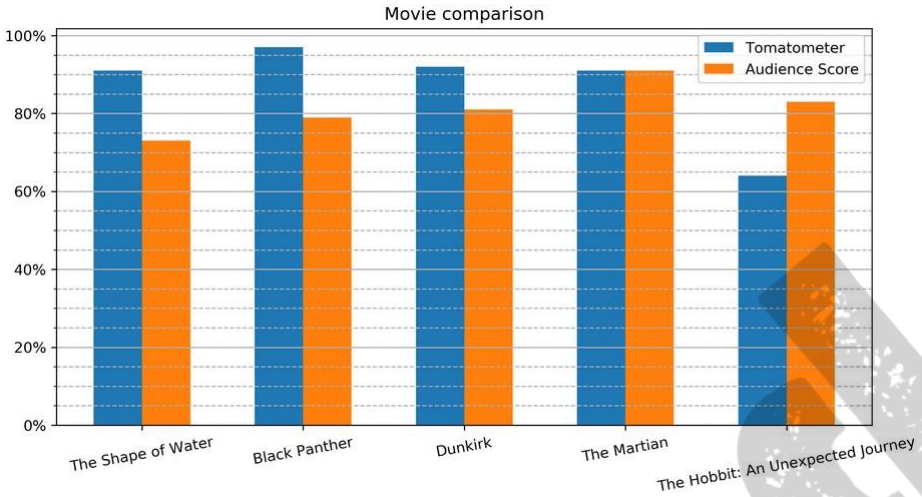


Horizontal bar chart using student test data

Example 2:

The below graph compares movie ratings with two scores: the Tomatometer, representing the percentage of approved critic reviews, and the Audience Score, representing the percentage of users rating 3.5 or higher out of 5. Notably, The

Martian has high scores on both metrics. The Hobbit: An Unexpected Journey has a high Audience Score despite a lower Tomatometer score, likely due to its large fan base.



Comparative bar chart

Design Practices

1. When creating bar charts, ensure the numerical axis starts at zero to avoid misleading representations.
2. Use horizontal labels if the chart isn't too cluttered.
3. If space is limited, rotate the labels at different angles, as seen on the x-axis of the preceding diagram.

3. Radar Charts

- Radar charts (also known as spider or web charts) visualize multiple variables with each variable plotted on its own axis, resulting in a polygon.
- All axes are arranged radially, starting at the center with equal distances between one another, and have the same scale.

Uses

- ✓ Radar charts are great for comparing multiple quantitative variables for a single group or multiple groups.
- ✓ They are also useful for showing which variables score high or low within a dataset, making them ideal for visualizing performance.

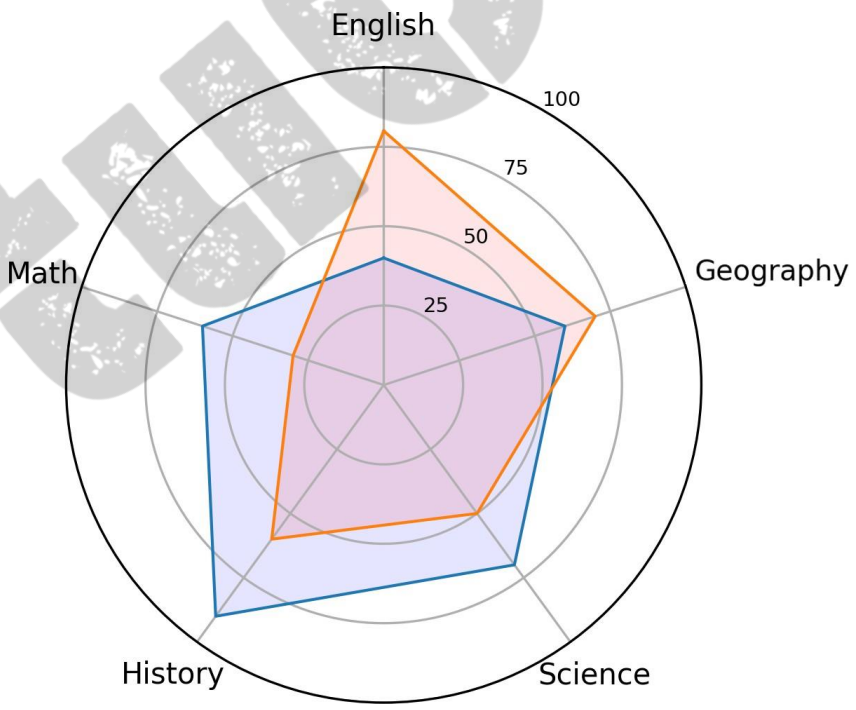
Example 1:

The following diagram shows a radar chart for a single variable. This chart displays data about a student scoring marks in different subjects:

Radar chart for one variable (student)

Example 2:

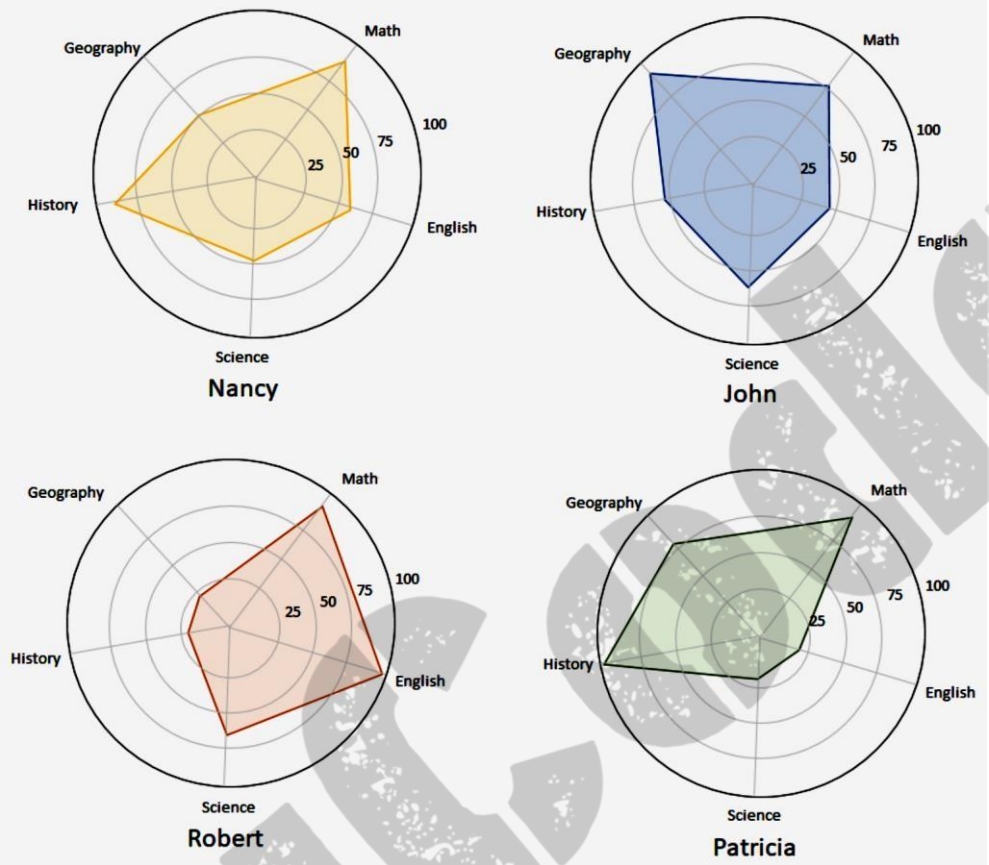
The following diagram shows a radar chart for two variables/groups. Here, the chart explains the marks that were scored by two students in different subjects:



Radar chart for two variables (two students)

Example 3:

The following diagram shows a radar chart for multiple variables/groups. Each chart displays data about a student's performance in different subjects:



Radar chart with faceting for multiple variables (multiple students)

Design Practices

1. Try to display 10 factors or fewer on a single radar chart to make it easier to read.
2. Use faceting (displaying each variable in a separate plot) for multiple variables/ groups, as shown in the preceding diagram, in order to maintain clarity.

Questions:

1. Discuss various comparison plots.

2. Explain what Line, Bar and Radar charts are. Also explain their uses and design practices with examples.

3. Explain the variants of Bar charts with examples.

Handouts for Session 4: Relation Plots: Scatter Plot, Bubble Plot , Correlogram and Heatmap

4.6 Relation Plots

- **Relation plots** are used to show **relationships among variables**.
- A **scatter plot** visualizes the correlation between two variables for one or multiple groups.
- **Bubble plots** can be used to show relationships between three variables. The additional third variable is represented by the dot size.
- **Heatmaps** are great for revealing patterns or correlations between two qualitative variables.
- A **correlogram** is a perfect visualization for showing the correlation among multiple variables.

1. Scatter Plot

- **Scatter plots** show data points for two numerical variables, displaying a variable on both axes.

  Uses

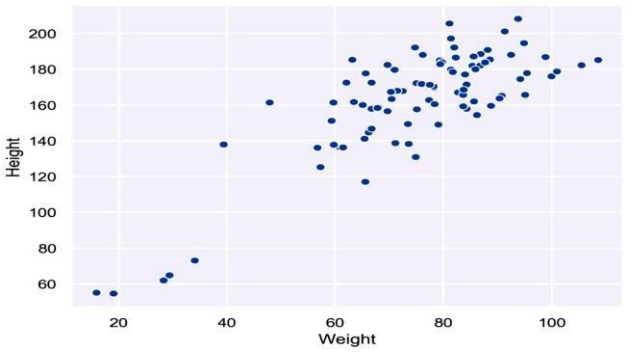
  ✓ Used to determine if a correlation (relationship) exists between two variables.

  ✓ Used to plot the relationship between multiple groups or categories using different colors.

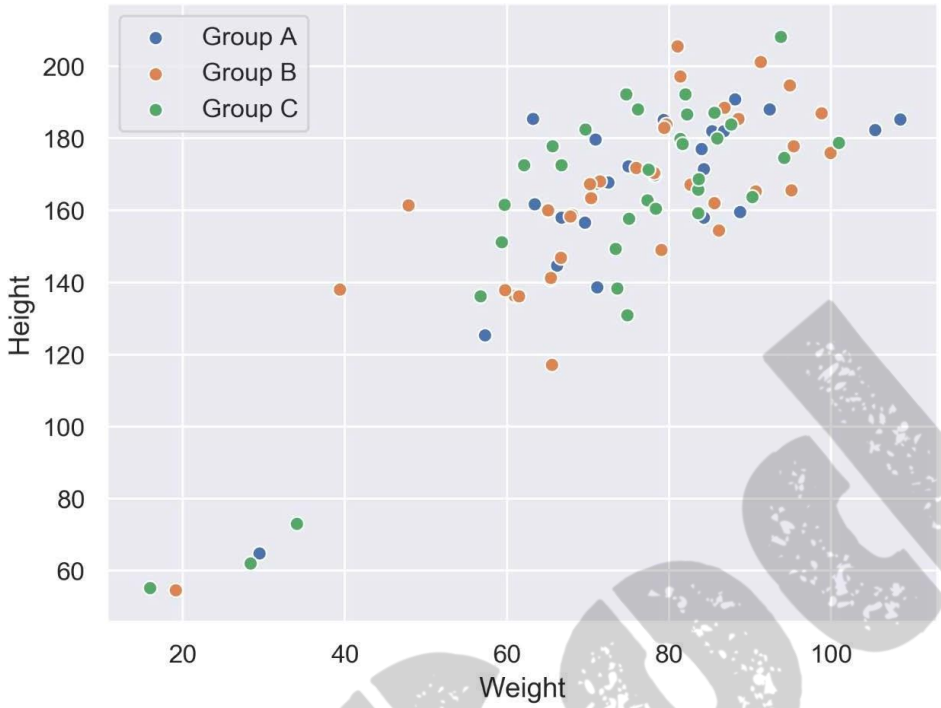
A bubble plot, which is a variation of the scatter plot, is an excellent tool for visualizing the correlation of a third variable.

Examples

The following diagram shows a scatter plot of height and weight of persons belonging to a single group: **Scatter plot with a single group**



The following diagram shows the same data as in the previous plot but differentiates between groups. In this case, we have different groups: A, B, and C:



Scatter plot with multiple groups

Design Practices

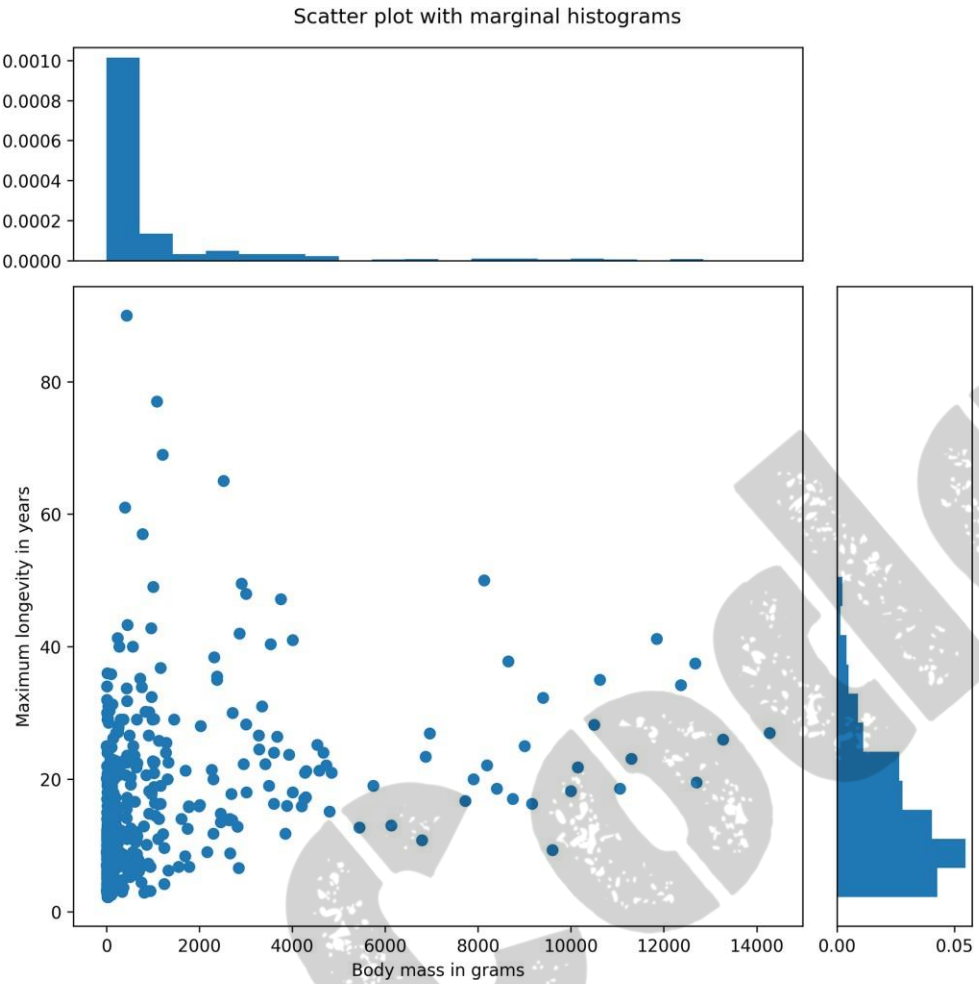
1. Start both axes at zero to represent data accurately.
2. Use contrasting colors for data points and avoid using symbols for scatter plots with multiple groups or categories.

Variants: Scatter Plots with Marginal Histograms

In addition to the scatter plot, which visualizes the correlation between two numerical variables, you can plot the marginal distribution for each variable in the form of histograms to give better insight into how each variable is distributed.

Example

The following diagram shows the correlation between body mass and the maximum longevity for animals in the **Aves** class. The marginal histograms are also shown, which helps to get a better insight into both variables:



Correlation between body mass and maximum longevity of the Aves class with marginal histograms

2. Bubble Plot

- A bubble plot **extends a scatter plot by introducing a third numerical variable**.
- The **value of the variable** is represented by **the size of the dots**.
- The area of the dots is proportional to the value.
- **A legend is used to link the size of the dot to an actual numerical value**.

Uses

- Bubble plots help to show a correlation between three variables.

Example

The following diagram shows a bubble plot that highlights the relationship between heights and age of humans to get the weight of each person, which is represented by the size of the bubble:

Bubble plot showing the relation between height and age of humans

Design Practices

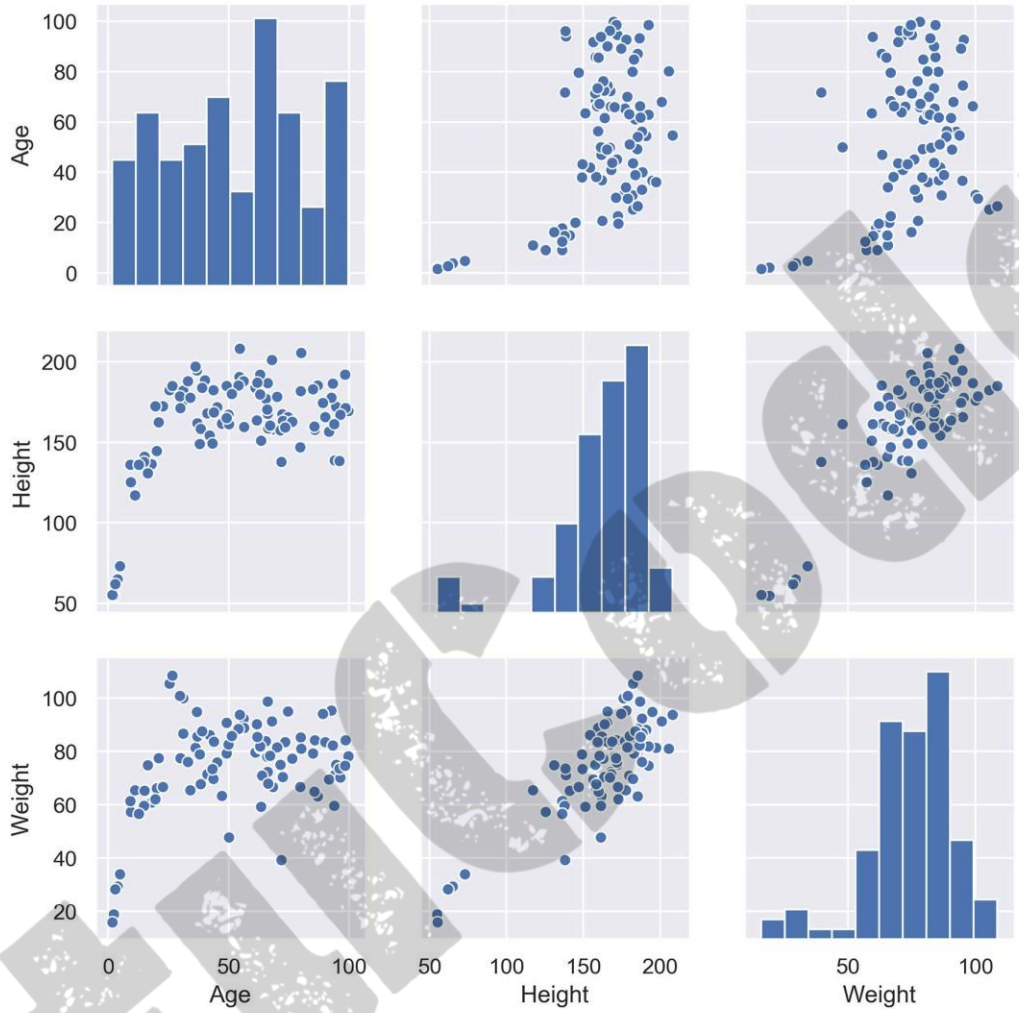
1. The design practices for the scatter plot are also applicable to the bubble plot.
2. Don't use bubble plots for very large amounts of data, since too many bubbles make the chart difficult to read.

3. Correlogram

- A **correlogram** is a combination of scatter plots and histograms.
- A **correlogram** or **correlation matrix** visualizes the **relationship between each pair of numerical variables using a scatter plot.**
- The **diagonals** of the correlation matrix represent the **distribution of each variable in the form of a histogram.**
- Different colors can also be used to plot the relationship between multiple groups or categories.
- A correlogram is **a great chart for exploratory data analysis** to get a feel for your data, especially the correlation between variable pairs.

Examples

The following diagram shows a correlogram for the height, weight, and age of humans. The diagonal plots show a histogram for each variable. The off-diagonal elements show scatter plots between variable pairs:

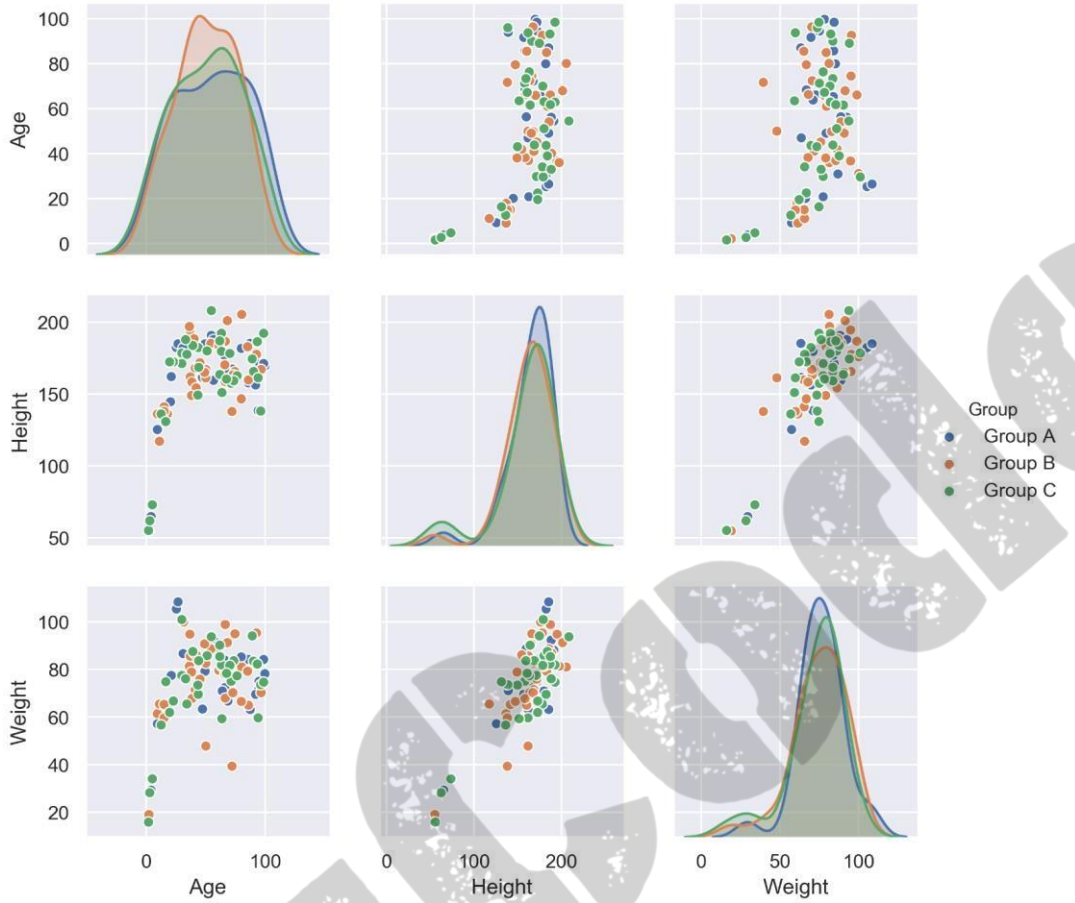


Correlogram with a single category

Design Practices

1. Start both axes at zero to represent data accurately.
2. Use contrasting colors for data points and avoid using symbols for scatter plots with multiple groups or categories.

The following diagram shows the correlogram with data samples separated by color into different groups:



Correlogram with multiple categories

4. Heatmap

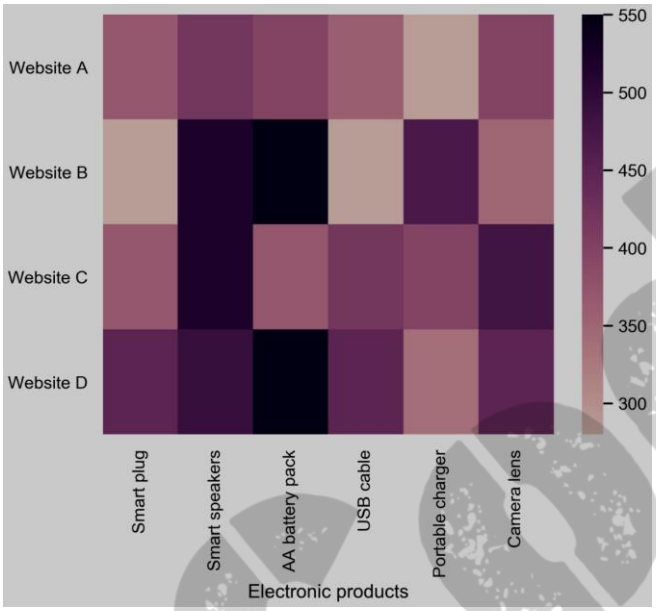
- A **heatmap** is a visualization where **values** contained in a matrix **are represented as colors** or **color saturation**.
- Heatmaps are great for **visualizing multivariate data** (data in which analysis is based on more than two variables per observation).
- In heatmaps **categorical variables are placed in the rows and columns** and **a numerical or categorical variable is represented as colors** or **color saturation**.

Use

- The visualization of multivariate data can be done using heatmaps as they are great for finding patterns in your data.

Example:

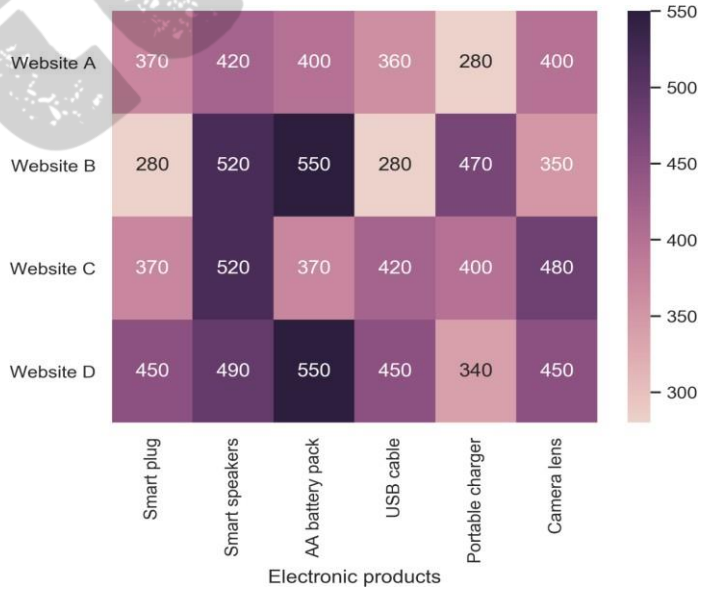
The following diagram shows a heatmap for the most popular products on the electronics category page across various e-commerce websites, where the color shows the number of units sold. In the following diagram, we can analyze that the darker colors represent more units sold, as shown in the key:



Heatmap for popular products in the electronics category

Variants: Annotated Heatmaps

Let's see the same example we saw previously in an annotated heatmap, where the color shows the number of units sold:



Questions:

1. Discuss various Relation plots.

2. Explain what Scatter Plot, Bubble Plot, Correlogram and Heatmap are. Also explain their uses and design practices with examples.

3. Explain the Scatter Plots with Marginal Histograms with an example.

4. Explain Heatmaps and its variant in detail.

Handouts for Session 5: Composition Plots: Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram

4.7 Composition Plots

- **Composition plots** are ideal **to represent some data as a part of a whole.**
- For **static data**, you can use **pie charts**, **stacked bar charts**, or **Venn diagrams**.
- **Pie charts** or **donut charts** help show **proportions and percentages for groups**.
- If you need an additional dimension, **stacked bar charts** are great.
- **Venn diagrams** are the best way **to visualize overlapping groups**, where each group is represented by a circle.
- For **data that changes over time**, you can use either **stacked bar charts** or **stacked area charts.**

1. Pie Chart

- Pie charts illustrate numerical proportions by dividing a circle into slices.
- Each arc length represents a proportion of a category.
- The full circle equates to 100%.
- For humans, it is easier to compare bars than arc lengths; therefore, it is recommended to use bar charts or stacked bar charts the majority of the time.

Use

- To compare items that are part of a whole.

Examples

The following diagram shows household water usage around the world:

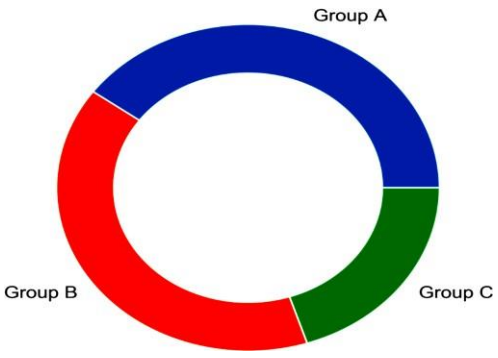
Pie chart for global household water usage

Design Practices

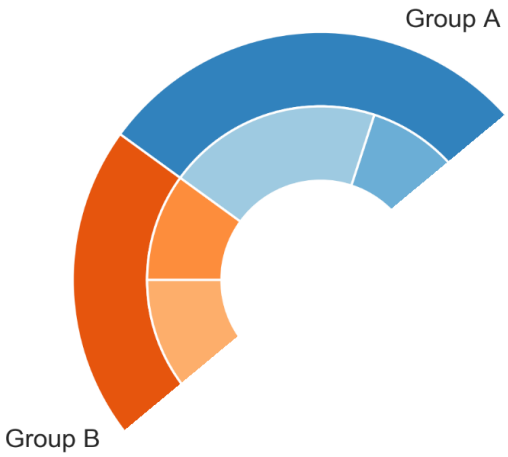
1. Arrange the slices according to their size in increasing/decreasing order, either in a clockwise or counter clockwise manner.
2. Make sure that every slice has a different color.

Variants: Donut Chart

- An alternative to a pie chart is a **donut chart.**
- In **contrast to pie charts**, it is **easier to compare the size of slices**, since the **reader focuses more on reading the length of the arcs** instead of the area.
- Donut charts are also more space-efficient because the center is cut out, so it can be used to display information or further divide groups into subgroups.
- The following diagram shows a basic donut chart:



The following diagram shows a donut chart with subgroups:



Design Practice

1. Use the same color that's used for the category for the subcategories.
2. Use varying brightness levels for the different subcategories.

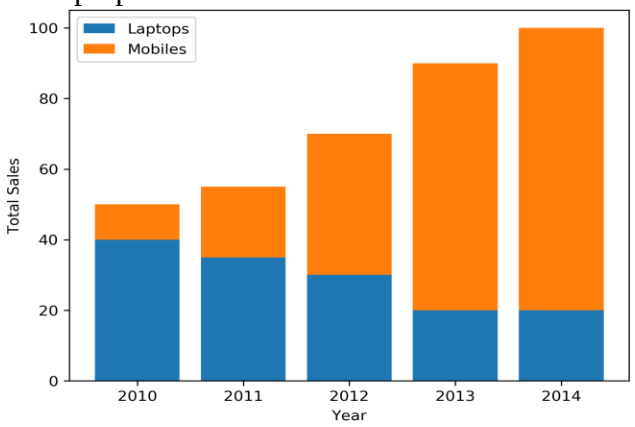
2. Stacked Bar Chart

- **Stacked bar charts** are used to show how a category is divided into subcategories and the proportion of the subcategory in comparison to the overall category.
- Total amounts can be compared across each bar, or the percentage of each group can be displayed.
- The latter is also referred to as a **100% stacked bar chart** and makes it easier to see relative differences between quantities in each group.

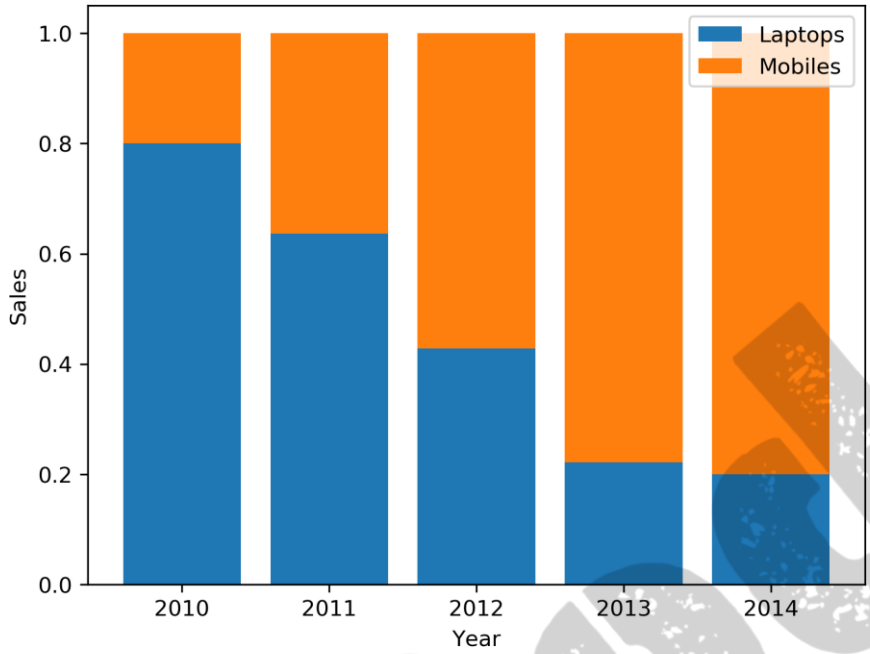
Use

- To compare variables that can be divided into sub-variables.

Example 1:
The following diagram shows a generic stacked bar chart with five groups: Stacked bar chart to show sales of laptops and mobiles



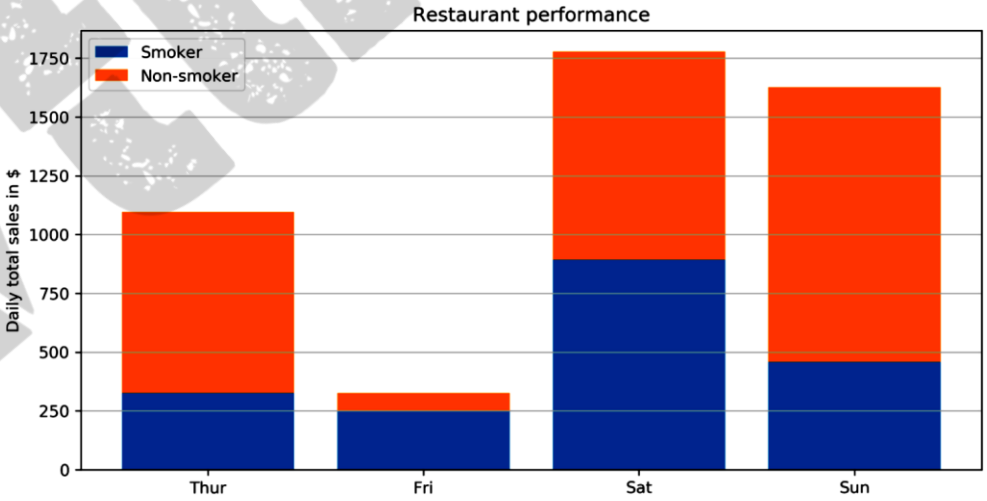
The following diagram shows a 100% stacked bar chart with the same data that was used in the preceding diagram:



100% stacked bar chart to show sales of laptops, PCs, and mobiles

Example 2:

The following diagram illustrates the daily total sales of a restaurant over several days. The daily total sales of non-smokers are stacked on top of the daily total sales of smokers:



Daily total restaurant sales categorized by smokers and non-smokers

Design Practices

- Use contrasting colors for stacked bars.
- Ensure that the bars are adequately spaced to eliminate visual clutter.
- The ideal space guideline between each bar is half the width of a bar.
- Categorize data alphabetically, sequentially, or by value, to uniformly order it and make things easier for your audience.

3. Stacked Area Chart

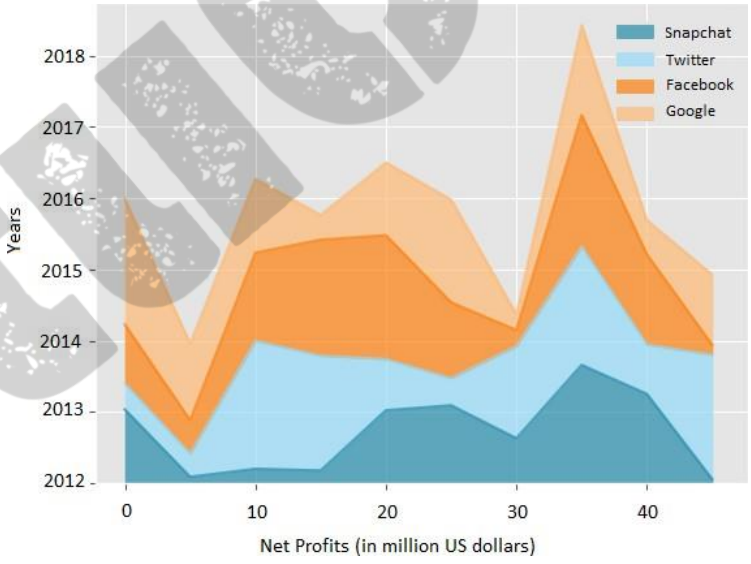
- Stacked area charts show trends for part-of-a-whole relations.
- The values of several groups are illustrated by stacking individual area charts on top of one another.
- It helps to analyze both individual and overall trend information.

Use

- To show trends for time series that are part of a whole.

Examples

The following diagram shows a stacked area chart with the net profits of Google, Facebook, Twitter, and Snapchat over a decade:



Stacked area chart to show net profits of four companies

Design Practice

1. Use transparent colors to improve information visibility.
2. This helps in analyzing overlapping data and makes the grid lines visible.

4. Venn Diagram

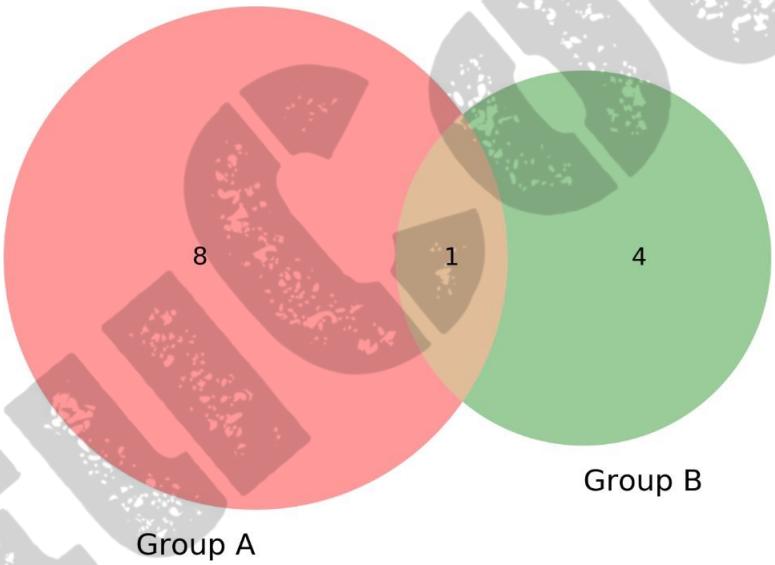
- **Venn diagrams**, also known as **set diagrams**, show all possible logical relations between a finite collection of different sets.
- Each set is represented by a circle.
- The **circle size illustrates the importance of a group**.
- The size of overlap represents the intersection between multiple groups.

Use

- ✓ To show overlaps for different sets.

Example

Visualizing the intersection of the following diagram shows a Venn diagram for students in two groups taking the same class in a semester:



Venn diagram showing students taking the same class

From the preceding diagram, we can note that there are eight students in just group A, four students in just group B, and one student in both groups.

Design Practice

1. It is not recommended to use Venn diagrams if you have more than three groups. It would become difficult to understand.
2. Moving on from composition plots, we will cover distribution plots in the following section.

Questions

1. Discuss various Composition plots.

2. Explain in detail Pie Charts. How Donut Chart can be more convenient than Pie Chart?

3. Explain the Stacked Bar Charts with an example. Also explain the uses and the design practices to ne followed.

4. Compare Stacked Bar Charts and Stacked Area Charts.

5. Discuss Venn Diagram with an example.

Handouts for Session 6: Distribution Plots: Histogram, Density Plot, Box Plot, Violin Plot

4.8 Distribution Plots

- Distribution plots give a deep insight into how your data is distributed.
- For a single variable, a histogram is effective.
- For multiple variables, you can either use a box plot or a violin plot.
- The violin plot visualizes the densities of your variables, whereas the box plot just visualizes the median, the interquartile range, and the range for each variable.

1. Histogram

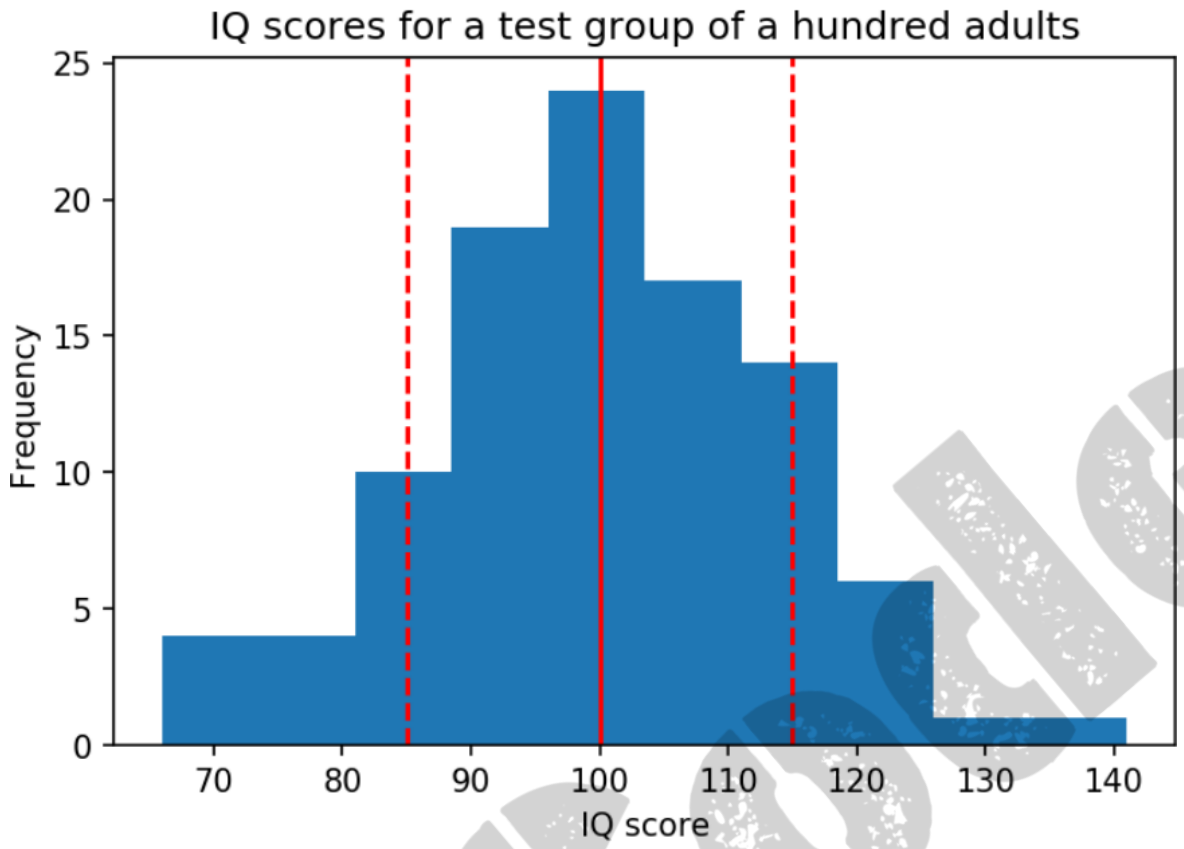
- A histogram visualizes the distribution of a single numerical variable.
- Each bar represents the frequency for a certain interval.
- Histograms provide an estimate of statistical measures, revealing where values are concentrated and making it easy to detect outliers.
- A histogram can be plotted using absolute frequency values or, alternatively, by normalizing the values.
- Different colors for the bars can be used to compare distributions of multiple variables.

Use

- Get insights into the underlying distribution for a dataset.

Example

The following diagram shows the distribution of the **Intelligence Quotient** (**IQ**) for a test group. The dashed lines represent the standard deviation each side of the mean (the solid line):



Distribution of IQ for a test group of a hundred adults

Design Practices

1. Try different numbers of bins (data intervals), since the shape of the histogram can vary significantly.

2. Density Plot

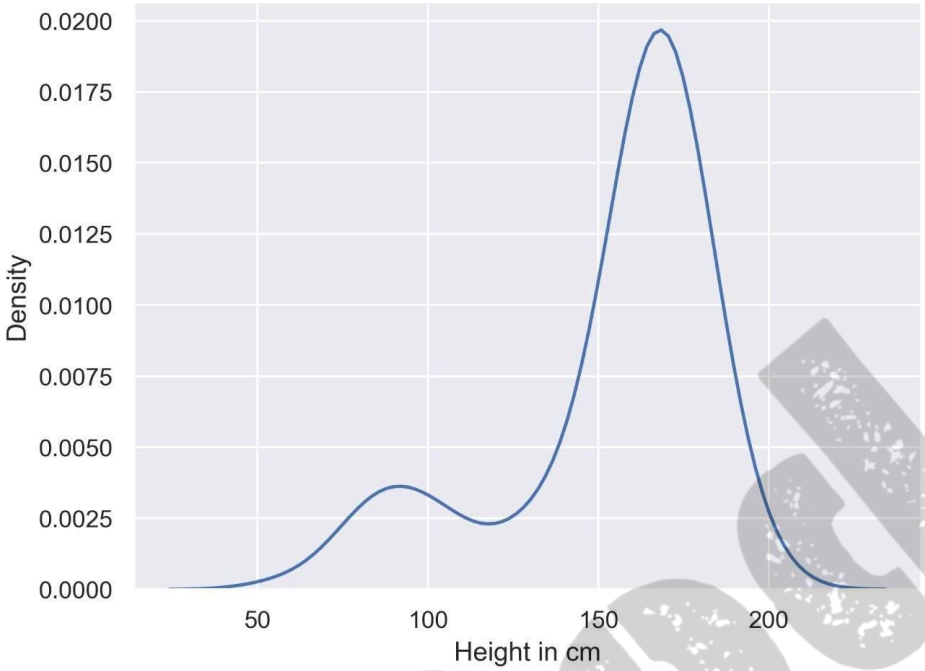
- A density plot shows the distribution of a numerical variable.
- It is a variation of a histogram that uses kernel smoothing, allowing for smoother distributions.
- One advantage these have over histograms is that density plots are better at determining the distribution shape since the distribution shape for histograms heavily depends on the number of bins (data intervals).

Use

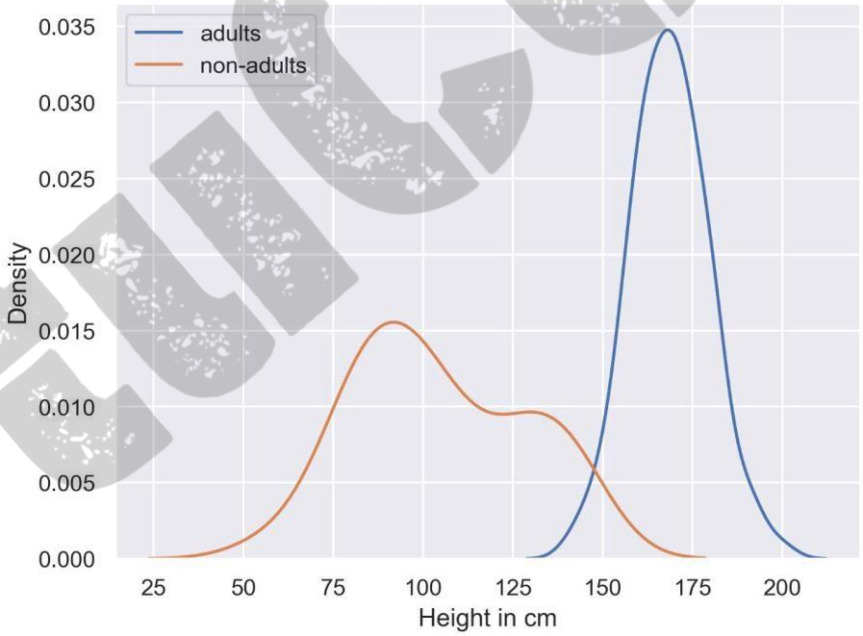
- To compare the distribution of several variables by plotting the density on the same axis and using different colors.

Example

The following diagram shows a basic density plot:



The following diagram shows a basic multi-density plot:



Design Practices

1. Use contrasting colors to plot the density of multiple variables.

3. Box Plot

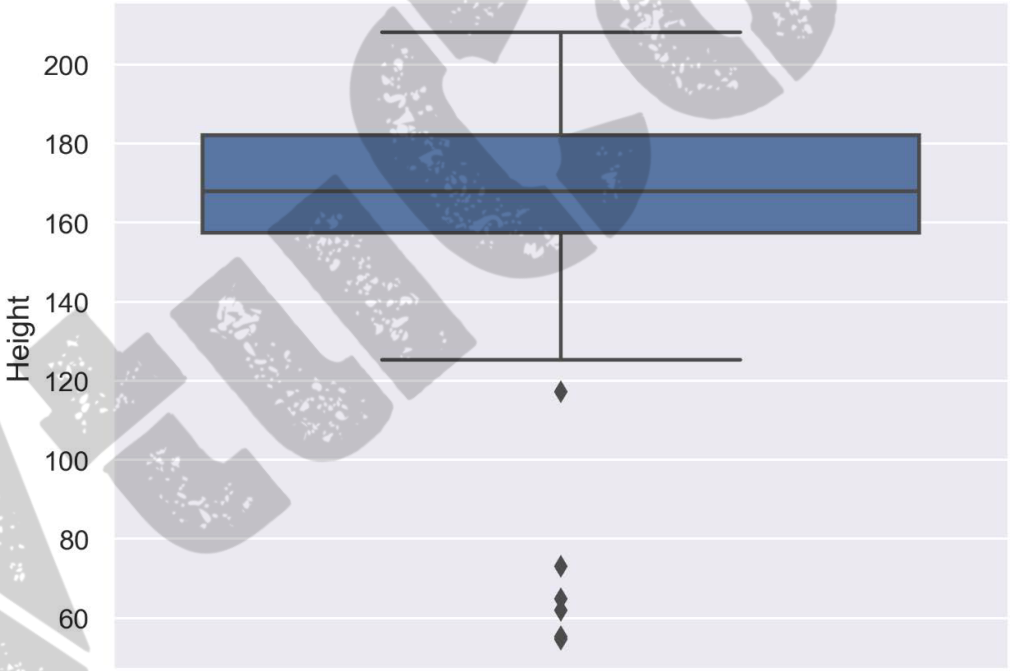
- The box plot shows multiple statistical measurements.
- The box extends from the lower to the upper quartile values of the data, thus allowing us to visualize the interquartile range (IQR).
- The horizontal line within the box denotes the median.
- The parallel extending lines from the boxes are called whiskers; they indicate the variability outside the lower and upper quartiles.
- There is also an option to show data outliers, usually as circles or diamonds, past the end of the whiskers.

Use

✓ Compare statistical measures for multiple variables or groups.

Example

The following diagram shows a basic box plot that shows the height of a group of people:



The following diagram shows a basic box plot for multiple variables. In this case, it shows heights for two different groups – adults and non-adults:

4. Violin Plot

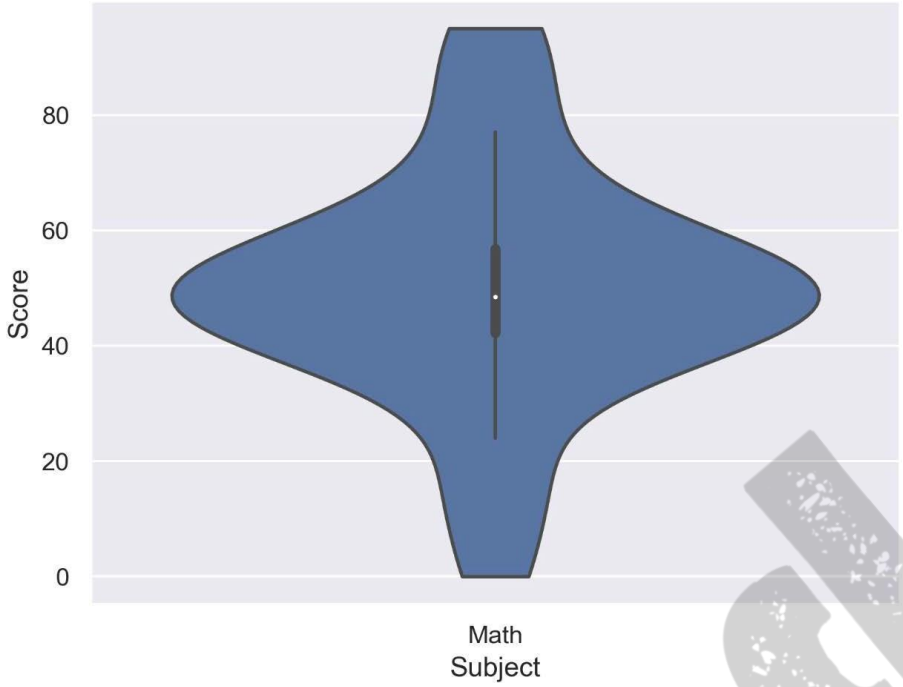
- Violin plots are a combination of box plots and density plots.
- Both the statistical measures and the distribution are visualized.
- The thick black bar in the center represents the interquartile range, while the thin black line corresponds to the whiskers in a box plot.
- The white dot indicates the median.
- On both sides of the centerline, the density is visualized.

Use

✓ Compare statistical measures and density for multiple variables or groups.

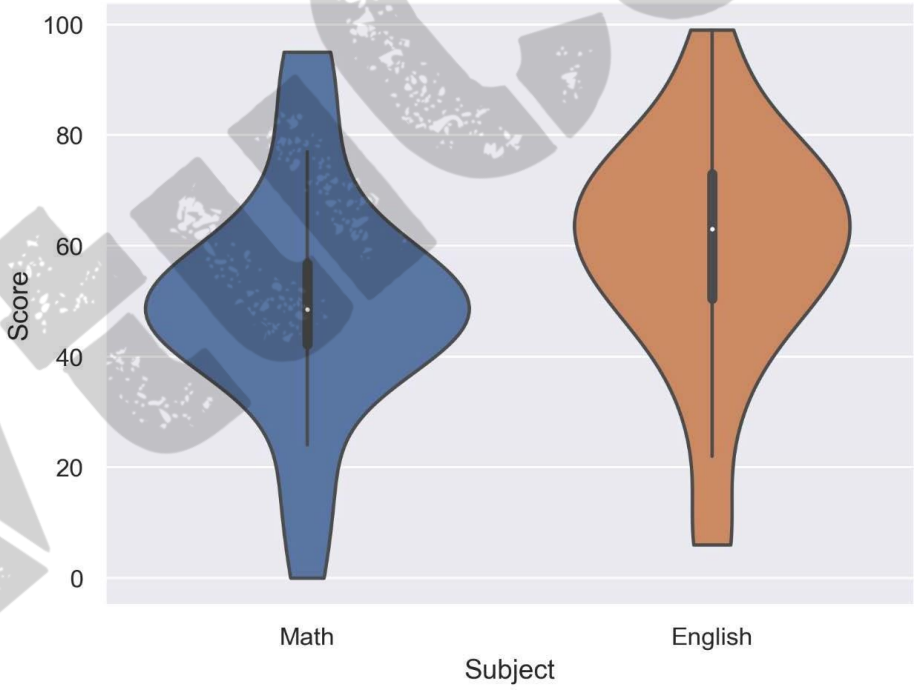
Example

The following diagram shows a violin plot for a single variable and shows how students have performed in Math: From the diagram, we can analyze that most of the students have scored around 40-60 in the Math test.



Violin plot for a single variable (Math)

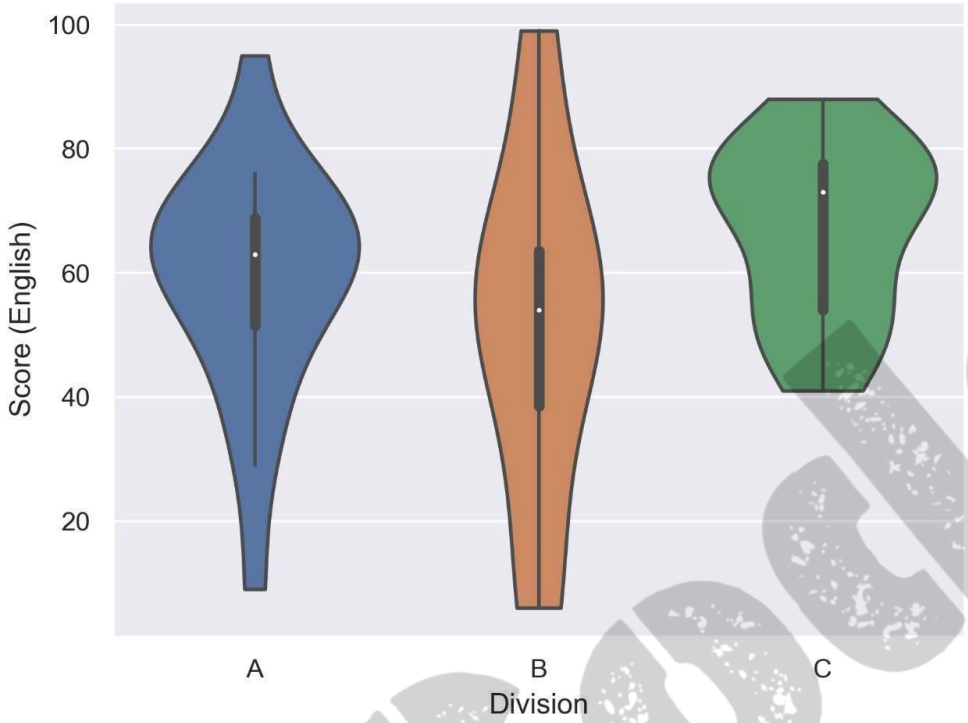
The following diagram shows a violin plot for two variables and shows the performance of students in English and Math:



Violin plot for multiple variables (English and Math)

From the preceding diagram, we can say that on average, the students have scored more in English than in Math.

The following diagram shows a violin plot for a single variable divided into three groups, and shows the performance of three divisions of students in English based on their score:



Violin plot with multiple categories (three groups of students)

From the preceding diagram, we can note that on average, division C has scored the highest, division B has scored the lowest, and division A is, on average, in between divisions B and C.

Design Practices

1. Scale the axes accordingly so that the distribution is clearly visible and not flat.

Questions

1. Discuss various Distribution plots.
2. Explain in detail Histogram Plots in detail with example.
3. Explain the Density Plots how is it different from Histogram Plots.
4. Compare Box Plots and Violin Plots.

Handouts for Session 7: Geo Plots: Dot Map, Choropleth Map, Connection Map

4.9 Geo Plots

- ✓ Geological plots are a great way to visualize geospatial data.
- ✓ Choropleth maps can be used to compare quantitative values for different countries, states, and so on.
- ✓ Connections between different locations can be represented using connection maps.

1. Dot Map

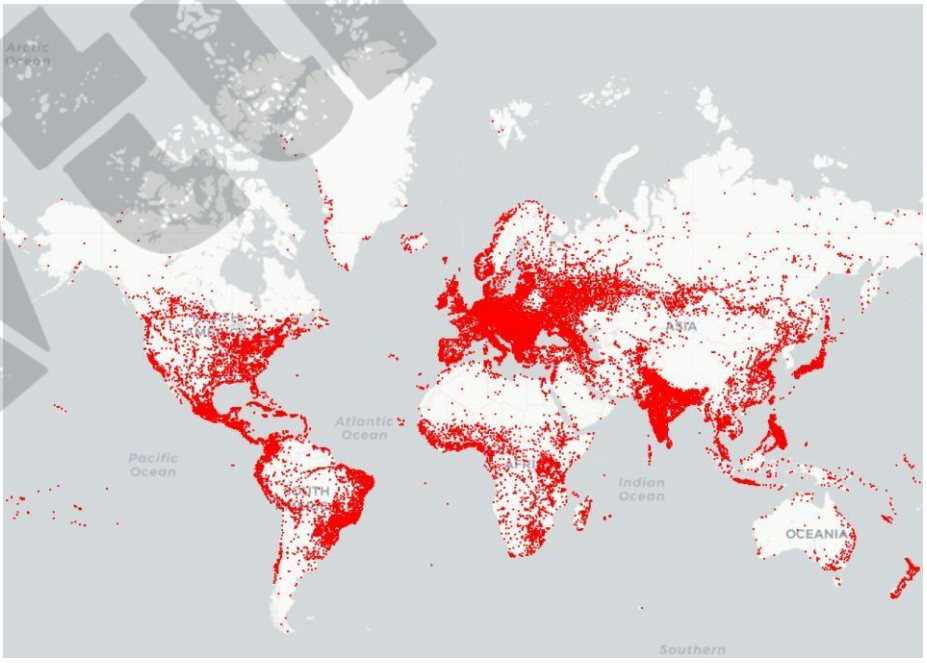
- In a dot map, each dot represents a certain number of observations.
- Each dot has the same size and value (the number of observations each dot represents).
- The dots are not meant to be counted; they are only intended to give an impression of magnitude.
- The size and value are important factors for the effectiveness and impression of the visualization.
- Different colors or symbols can be used for the dots to show multiple categories or groups.

Use

- ✓ To visualize geospatial data.

Example

The following diagram shows a dot map where each dot represents a certain amount of bus stops throughout the world:



Dot map showing bus stops worldwide

Design Practices

1. Avoid displaying too many locations to ensure the map remains clear and the actual locations are discernible.
2. Select an appropriate dot size and value to ensure that in dense areas, the dots blend together, providing a clear impression of the underlying spatial distribution.

2. Choropleth Map

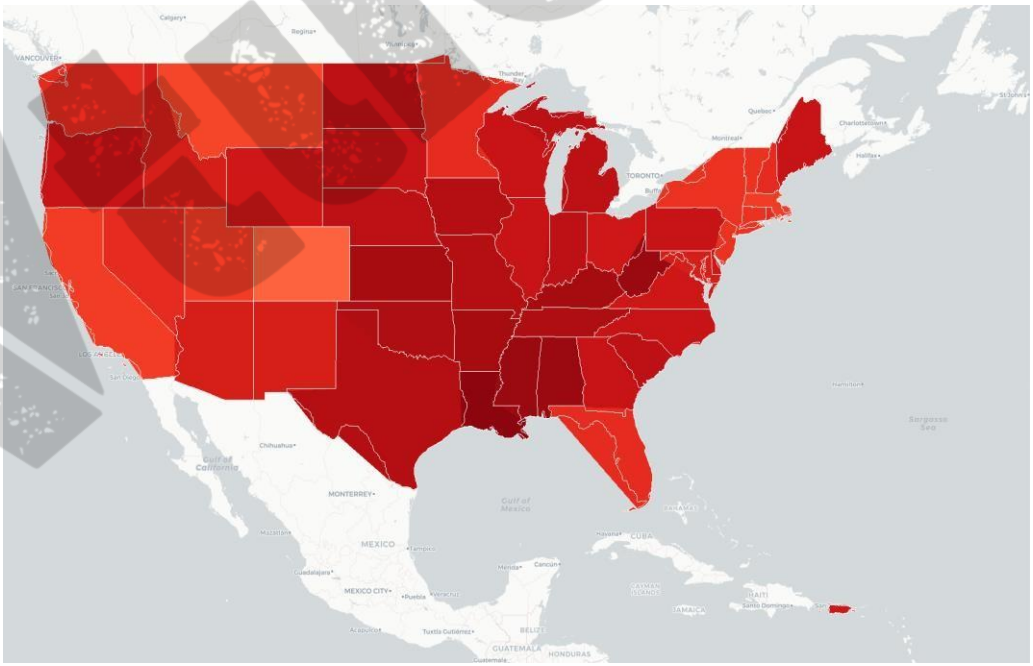
- In a choropleth map, each tile is colored to encode a variable.
- For example, a tile represents a geographic region for counties and countries.
- Choropleth maps provide a good way to show how a variable varies across a geographic area.
- One thing to keep in mind for choropleth maps is that the human eye naturally gives more attention to larger areas, so you might want to normalize your data by dividing the map area-wise.

Use

- ✓ To visualize geospatial data grouped into geological regions—for example, states or countries.

Example

The following diagram shows a choropleth map of a weather forecast in the USA:



Choropleth map showing a weather forecast for the USA

Design Practices

1. Use darker colors for higher values, as they are perceived as being higher in magnitude.
2. Limit the color gradation, since the human eye is limited in how many colors it can easily distinguish between. Seven color gradations should be enough.

3. Connection Map

- In a **connection map**, each line represents a certain number of connections between two locations.
- The link between the locations can be drawn with a straight or rounded line, representing the shortest distance between them.
- Each line has the same thickness and value (the number of connections each line represents).
- The lines are not meant to be counted; they are only intended to give an impression of magnitude.
- The size and value of a connection line are important factors for the effectiveness and impression of the visualization.

Use

- ✓ To visualize connections.

Example

The following diagram shows a connection map of flight connections around the world:

Design Practices

1. Avoid displaying too many connections, as it can make data analysis challenging. Ensure the map remains clear enough to identify the actual locations of the start and end points.
2. Choose a line thickness and value so that the lines start to blend in dense areas. The connection map should give a good impression of the underlying spatial distribution.

Questions:

1. Discuss various Geo plots.

2. Explain in detail Dot Plots in detail with example.

3. Explain the uses of Choropleth Map and what are the design practices to be followed.

Handouts for Session 8: What Makes a Good Visualization?

4.10 What Makes a Good Visualization?

There are multiple aspects to what makes a good visualization:

1. Most importantly, the visualization should be self-explanatory and visually appealing. To make it self-explanatory, use a legend, descriptive labels for your x-axis and y-axis, and titles.
2. A visualization should tell a story and be designed for your audience. Before creating your visualization, think about your target audience; create simple visualizations for a non-specialist audience and more technical detailed visualizations for a specialist audience. Think about a story to tell with your visualization so that your visualization leaves an impression on the audience.

Common Design Practices

- ✓ Use colors to differentiate variables/subjects rather than symbols, as colors are more perceptible.
- ✓ To show additional variables on a 2D plot, use color, shape, and size.
- ✓ Keep it simple and don't overload the visualization with too much information.

6th A 23-24

K.S SCHOOL OF ENGINEERING AND MANAGEMENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
Advanced Learners - 1

YEAR / SEMESTER	**III / VIA**
COURSE TITLE	**Data Science & Visualization**
COURSE CODE	**21CS644**

S.NO	USN	NAME	Signature
1	1KG21CS001	ABBURI PALLAVI	
2	1KG21CS014	B NAYANA	
3	1KG21CS027	DEEKSHITHA R	
4	1KG21CS031	DHANUSH G P	
5	1KG21CS045	HARSHITH K P	
6	1KG21CS048	JAHANAVI S	
7	1KG21CS054	KASTURI POORNIMA CHOWDARY	
8	1KG21CS057	KUSHMITHA T A	
9	1KG21CS059	L LAVANYA	
10	1KG21CS060	LAKSHA SENTHILKUMAR	
11	1KG21CS061	M POOJA	

Faculty Incharge

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

6th A 23-24

K.S SCHOOL OF ENGINEERING AND MANAGEMENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
Advanced Learners - 2

YEAR / SEMESTER	**III / VIA**
COURSE TITLE	**Data Science & Visualization**
COURSE CODE	**21CS644**

S.NO	USN	NAME	Signature
1	1KG21CS003	ABHILASH B R	
2	1KG21CS008	AMITH C SURI	
3	1KG21CS014	B NAYANA	
4	1KG21CS027	DEEKSHITHA R	
5	1KG21CS054	KASTURI POORNIMA CHOWDARY	
6	1KG21CS055	KISHOR KUMAR L	
7	1KG21CS057	KUSHMITHA T A	
8	1KG21CS058	KUSUMA B	
9	1KG21CS061	M POOJA	

Faculty Signature

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K.S SCHOOL OF ENGINEERING AND MANAGEMENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
Attendance for Remedial Class

YEAR / SEMESTER	**III / VIA**
COURSE TITLE	**Data Science & Visualization**
COURSE CODE	**21CS644**

S.NO	USN	NAME	Signature
1	1KG21CS001	ABBURI PALLAVI	
2	1KG21CS002	ABHIJEET DAS	
3	1KG21CS003	ABHILASH B R	
4	1KG21CS004	ABHISHEK V	
5	1KG21CS006	AKSHATHA R GOWDA	
6	1KG21CS007	ALLU CHINNI KRISHNA	
7	1KG21CS009	AMOGH A	
8	1KG21CS016	BEEGALU SRINIVAS AKHIL	
9	1KG21CS017	BHARATH GOWDA J	
10	1KG21CS018	BHAVANA D	
11	1KG21CS019	BHAVANA S	
12	1KG21CS021	C SUSMITHA	
13	1KG21CS022	CHALLA HARI KISHORE	
14	1KG21CS023	CHANDAN TAVANE	
15	1KG21CS026	DARSHAN R	
16	1KG21CS028	DEEPAK ATHRESH R	
17	1KG21CS031	DHANUSH G P	
18	1KG21CS032	DHANUSH U S	
19	1KG21CS035	DIBYAJYOTI SAHU	
20	1KG21CS036	DINESH J L	
21	1KG21CS038	DIVYA P	
22	1KG21CS040	GONGATI RAGHU	
23	1KG21CS042	GURUJALA BHARATH	
24	1KG21CS043	GURURAJ B	
25	1KG21CS044	HANOCH CHRISTIAN R	
26	1KG21CS045	HARSHITH K P	
27	1KG21CS046	HARSHITHA D G	
28	1KG21CS047	HITESH A REDDY	
29	1KG21CS050	K J PRAKRUTHI	
30	1KG21CS051	K NITHISH	

31	1KG21CS052	K ROHITH	
32	1KG21CS053	KARTHIK G	
33	1KG21CS059	L LAVANYA	
34	1KG21CS063	MADINENI BHUVANA	
35	1KG22CS400	AKSHAY U	
36	1KG22CS401	BALAJI N	
37	1KG22CS402	BHAVANA M	
38	1KG22CS403	DHANUSH R	
39	1KG22CS404	DHANUSHREE A	
40	1KG22CS405	KIRAN KUMAR	

Faculty Signature

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K.S SCHOOL OF ENGINEERING AND MANAGEMENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
Attendance for Remedial Class

YEAR / SEMESTER	**III / VIA**
COURSE TITLE	**Data Science & Visualization**
COURSE CODE	**21CS644**

S.NO	USN	NAME	Signature
1	1KG21CS003	ABHILASH B R	
2	1KG21CS004	ABHISHEK V	
3	1KG21CS005	AKHILA A	Akhila A
4	1KG21CS006	AKSHATHA R GOWDA	
5	1KG21CS007	ALLU CHINNI KRISHNA	
6	1KG21CS012	ARPITHA S	
7	1KG21CS017	BHARATH GOWDA J	
8	1KG21CS019	BHAVANA S	
9	1KG21CS026	DARSHAN R	
10	1KG21CS028	DEEPAK ATHRESH R	
11	1KG21CS034	DHEERAJ D RAIKAR	
12	1KG21CS035	DIBYAJYOTI SAHU	
13	1KG21CS039	GEOFFREY SAMUEL	
14	1KG21CS040	GONGATI RAGHU	
15	1KG21CS044	HANOCH CHRISTIAN R	
16	1KG21CS046	HARSHITHA D G	
17	1KG21CS053	KARTHIK G	
18	1KG21CS056	KUNALA GANESH	
19	1KG21CS063	MADINENI BHUVANA	M. Bhuvana
20	1KG22CS400	AKSHAY U	
21	1KG22CS401	BALAJI N	
22	1KG22CS402	BHAVANA M	
23	1KG22CS404	DHANUSHREE A	
24	1KG22CS405	KIRAN KUMAR	

Faculty Signature

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K S SCHOOL OF ENGINEERING AND MANAGEMENT,BENGALURU-560 109

DEPARTMENT OF COMPUTER SCIENCE & ENGG.

2023-24 EVEN SEMESTER VI SEM A SECTION

DSV ASSIGNMENT MARKS

SI.NO.	USN	NAME	ASN1	ASN2	TOTAL	Signature
1	1KG21CS001	ABBURI PALLAVI	10	10	20	
2	1KG21CS002	ABHIJEET DAS	10	10	20	
3	1KG21CS003	ABHILASH B R	10	10	20	
4	1KG21CS004	ABHISHEK V	10	10	20	
5	1KG21CS005	AKHILA A	10	10	20	Akhila A
6	1KG21CS006	AKSHATHA R GOWDA	10	10	20	
7	1KG21CS007	ALLU CHINNI KRISHNA	10	10	20	
8	1KG21CS008	AMITH C SURI	10	10	20	
9	1KG21CS009	AMOGH A	10	10	20	Amogh
10	1KG21CS010	ANKITHA VENKATESH	10	10	20	
11	1KG21CS011	ANKUSH GOWDA K	10	10	20	
12	1KG21CS012	ARPITHA S	10	10	20	
13	1KG21CS013	ASHWINI C	10	10	20	ashwini C
14	1KG21CS014	B NAYANA	10	10	20	
15	1KG21CS015	BATTA PREETHI	08	08	16	
16	1KG21CS016	BEEGALU SRINIVAS AKHIL	10	10	20	
17	1KG21CS017	BHARATH GOWDA J	10	10	20	
18	1KG21CS018	BHAVANA D	10	10	20	
19	1KG21CS019	BHAVANA S	10	10	20	Bhavana
20	1KG21CS021	C SUSMITHA	10	10	20	
21	1KG21CS022	CHALLA HARI KISHORE NAIDU	10	10	20	
22	1KG21CS023	CHANDAN TAVANE	10	10	20	
23	1KG21CS024	CHINMAIY P	10	10	20	
24	1KG21CS026	DARSHAN R	10	10	20	
25	1KG21CS027	DEEKSHITHA R	10	10	20	Deekshitha
26	1KG21CS028	DEEPAK ATHRESH R	10	10	20	Deepak
27	1KG21CS029	DEVDAS	10	10	20	Devdas
28	1KG21CS030	DHAKSHITHA A	10	10	20	Dhakshitha
29	1KG21CS031	DHANUSH G P	10	10	20	
30	1KG21CS032	DHANUSH U S	10	10	20	
31	1KG21CS033	DHARINI	10	10	20	
32	1KG21CS034	DHEERAJ D RAIKAR	10	10	20	
33	1KG21CS035	DIBYAJYOTI SAHU	10	10	20	
34	1KG21CS036	DINESH J L	10	10	20	

SI.NO.	USN	NAME	ASN1	ASN2	TOTAL	Signature
35	1KG21CS037	DIVYA H U	10	10	20	
36	1KG21CS038	DIVYA P	10	10	20	
37	1KG21CS039	GEOFFREY SAMUEL	10	10	20	
38	1KG21CS040	GONGATI RAGHU	10	10	20	
39	1KG21CS041	GORANTLA DIVYA SREE	10	10	20	
40	1KG21CS042	GURUJALA BHARATH	10	0	10	
41	1KG21CS043	GURURAJ B	10	10	20	
42	1KG21CS044	HANOCH CHRISTIAN R	10	10	20	
43	1KG21CS045	HARSHITH K P	10	10	20	
44	1KG21CS046	HARSHITHA D G	10	10	20	
45	1KG21CS047	HITESH A REDDY	10	10	20	
46	1KG21CS048	JAHANAVI S	10	10	20	
47	1KG21CS050	K J PRAKRUTHI	10	10	20	
48	1KG21CS051	K NITHISH	10	10	20	
49	1KG21CS052	K ROHITH	10	10	20	
50	1KG21CS053	KARTHIK G	10	10	20	
51	1KG21CS054	KASTURI POORNIMA CHOWDARY	10	10	20	
52	1KG21CS055	KISHOR KUMAR L	10	10	20	
53	1KG21CS056	KUNALA GANESH	10	10	20	
54	1KG21CS057	KUSHMITHA T A	10	10	20	
55	1KG21CS058	KUSUMA B	10	10	20	
56	1KG21CS059	L LAVANYA	10	10	20	
57	1KG21CS060	LAKSHA SENTHILKUMAR	10	10	20	
58	1KG21CS061	M POOJA	10	10	20	
59	1KG21CS062	M SURABHI	10	10	20	
60	1KG21CS063	MADINENI BHUVANA	10	10	20	
61	1KG22CS400	AKSHAY U	10	10	20	
62	1KG22CS401	BALAJI N	10	10	20	
63	1KG22CS402	BHAVANA M	10	10	20	
64	1KG22CS403	DHANUSH R	10	10	20	
65	1KG22CS404	DHANUSHREE A	10	10	20	
66	1KG22CS405	KIRAN KUMAR	10	10	20	

FACULTY INCHARGE

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K.S SCHOOL OF ENGINEERING AND MANAG IENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
VI - 'A' SEC FINAL AVERAGE MARKS (2023-2024)

YEAR / SEMESTER	VI- 'A'
COURSE TITLE	Data Science & Visualization
COURSE CODE	21CS644
ACADEMIC YEAR	2023-2024

S.NO	USN	NAME	IA1	IA2	IA3	IMP	ASS (20)	QUIZ (20)	TOTAL (100)	SCALE DOWN(50)	SIGNATURE
1	1KG21CS001	ABBURI PALLAVI	16	9	19		20	20	84	42	
2	1KG21CS002	ABHIJEET DAS	0	5	8	14	20	20	67	34	
3	1KG21CS003	ABHILASH B R	10	15	20		20	20	85	43	
4	1KG21CS004	ABHISHEK V	4	8	11		20	20	63	32	
5	1KG21CS005	AKHILA A	9	13	20		20	20	82	41	Akhila A
6	1KG21CS006	AKSHATHA R GOWDA	7	4	10		20	20	61	31	
7	1KG21CS007	ALLU CHINNI KRISHNA	6	6	16		20	20	68	34	
8	1KG21CS008	AMITH C SURI	14	17	20		20	20	91	46	
9	1KG21CS009	AMOGH A	11	9	19		20	20	79	40	
10	1KG21CS010	ANKITHA VENKATESH	14	12	20		20	20	86	43	
11	1KG21CS011	ANKUSH GOWDA K	11	0	15	15	20	20	81	41	
12	1KG21CS012	ARPITHA S	8	0	20	15	20	20	83	42	
13	1KG21CS013	ASHWINI C	14	0	19	16	20	20	89	45	Ashwini C
14	1KG21CS014	B NAYANA	15	17	20		20	20	92	46	B. Nayana
15	1KG21CS015	BATTA PREETHI	0	0	0	12	16	20	48	24	Preethi
16	1KG21CS016	BEEGALU SRINIVAS AKHIL	0	9	15	14	20	20	78	39	
17	1KG21CS017	BHARATH GOWDA J	14	14	20		20	20	88	44	
18	1KG21CS018	BHAVANA D	12	9	17		20	20	78	39	
19	1KG21CS019	BHAVANA S	13	8	19		20	20	80	40	Bhavana S.
20	1KG21CS021	C SUSMITHA	0	10	17	15	20	20	82	41	Susmitha
21	1KG21CS022	CHALLA HARI KISHORE NAIDU	0	10	19	15	20	20	84	42	C. Hari Naidu
22	1KG21CS023	CHANDAN TAVANE	11	6	20		20	20	77	39	
23	1KG21CS024	CHINMAIY P	14	11	20		20	20	85	43	
24	1KG21CS026	DARSHAN R	9	7	15		20	20	71	36	
25	1KG21CS027	DEEKSHITHA R	17	15	20		20	20	92	46	Deekshitha
26	1KG21CS028	DEEPAK ATHRESH R	7	11	18		20	20	76	38	Deepak
27	1KG21CS029	DEVDAS	11	13	19		20	20	83	42	
28	1KG21CS030	DHAKSHITHA A	13	11	20		20	20	84	42	Dhakshitha A
29	1KG21CS031	DHANUSH G P	17	8	20		20	20	85	43	
30	1KG21CS032	DHANUSH U S	14	8	20		20	20	82	41	
31	1KG21CS033	DHARINI	12	0	19	15	20	20	86	43	

K.S SCHOOL OF ENGINEERING AND MANAGEMENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
VI - 'A' SEC FINAL AVERAGE MARKS (2023-2024)

YEAR / SEMESTER	**VI- 'A'**
COURSE TITLE	**Data Science & Visualization**
COURSE CODE	**21CS644**
ACADEMIC YEAR	**2023-2024**

S.NO	USN	NAME	IA1	IA2	IA3	IMP	ASS (20)	QUIZ (20)	TOTAL (100)	SCALE DOWN(50)	SIGNATURE
32	1KG21CS034	DHEERAJ D RAIKAR	7	13	16		20	20	76	38	
33	1KG21CS035	DIBYAJYOTI SAHU	11	7	17		20	20	75	38	
34	1KG21CS036	DINESH J L	13	4	19		20	20	76	38	
35	1KG21CS037	DIVYA H U	14	14	20		20	20	88	44	
36	1KG21CS038	DIVYA P	0	9	15	16	20	20	80	40	
37	1KG21CS039	GEOFFREY SAMUEL	4	0	13	14	20	20	71	36	
38	1KG21CS040	GONGATI RAGHU	10	4	16		20	20	70	35	
39	1KG21CS041	GORANTLA DIVYA SREE	11	13	20		20	20	84	42	
40	1KG21CS042	GURUJALA BHARATH	0	5	15	9	10	20	59	30	
41	1KG21CS043	GURURAJ B	0	9	15	15	20	20	79	40	
42	1KG21CS044	HANOCH CHRISTIAN R	8	12	18		20	20	78	39	
43	1KG21CS045	HARSHITH K P	17	10	20		20	20	87	44	
44	1KG21CS046	HARSHITHA D G	8	10	20		20	20	78	39	
45	1KG21CS047	HITESH A REDDY	0	7	12	5	20	20	64	32	
46	1KG21CS048	JAHANAVI S	15	13	19		20	20	87	44	
47	1KG21CS050	K J PRAKRUTHI	0	0	10	5	20	20	55	28	
48	1KG21CS051	K NITHISH	0	9	15		20	20	64	32	
49	1KG21CS052	K ROHITH	0	6	17	9	20	20	72	36	
50	1KG21CS053	KARTHIK G	9	7	20		20	20	76	38	
51	1KG21CS054	KASTURI POORNIMA CHOWDARY	20	20	20		20	20	100	50	
52	1KG21CS055	KISHOR KUMAR L	14	17	20		20	20	91	46	
53	1KG21CS056	KUNALA GANESH	4	0	7	9	20	20	60	30	
54	1KG21CS057	KUSHMITHA T A	20	17	20		20	20	97	49	
55	1KG21CS058	KUSUMA B	14	16	19		20	20	89	45	
56	1KG21CS059	L LAVANYA	15	12	20		20	20	87	44	
57	1KG21CS060	LAKSHA SENTHILKUMAR	15	13	20		20	20	88	44	
58	1KG21CS061	M POOJA	20	15	20		20	20	95	48	
59	1KG21CS062	M SURABHI	0	12	20	15	20	20	87	44	
60	1KG21CS063	MADINENI BHUVANA	7	9	13		20	20	69	35	
61	1KG22CS400	AKSHAY U	9	5	0	12	20	20	66	33	
62	1KG22CS401	BALAJI N	10	4	9		20	20	63	32	

K.S SCHOOL OF ENGINEERING AND MANAC IENT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
VI - 'A' SEC FINAL AVERAGE MARKS (2023-2024)

YEAR / SEMESTER	VI- 'A'
COURSE TITLE	Data Science & Visualization
COURSE CODE	21CS644
ACADEMIC YEAR	2023-2024

S.NO	USN	NAME	IA1	IA2	IA3	IMP	ASS (20)	QUIZ (20)	TOTAL (100)	SCALE DOWN(50)	SIGNATURE
63	1KG22CS402	BHAVANA M	5	4	15		20	20	64	32	
64	1KG22CS403	DHANUSH R	13	10	20		20	20	83	42	
65	1KG22CS404	DHANUSHREE A	9	4	15		20	20	68	34	
66	1KG22CS405	KIRAN KUMAR	6	6	5		20	20	57	29	

Faculty Incharge

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE

Branch : CS Semester : 6

Sl NO.	USN	21CS644
1	1KG21CS001	42
2	1KG21CS002	34
3	1KG21CS003	43
4	1KG21CS004	32
5	1KG21CS005	41
6	1KG21CS006	31
7	1KG21CS007	34
8	1KG21CS008	46
9	1KG21CS009	40
10	1KG21CS010	43
11	1KG21CS011	41
12	1KG21CS012	42
13	1KG21CS013	45
14	1KG21CS014	46
15	1KG21CS015	24
16	1KG21CS016	39
17	1KG21CS017	44
18	1KG21CS018	39
19	1KG21CS019	40
20	1KG21CS021	41
21	1KG21CS022	42
22	1KG21CS023	39
23	1KG21CS024	43
24	1KG21CS026	36
25	1KG21CS027	46
26	1KG21CS028	38
27	1KG21CS029	42
28	1KG21CS030	42
29	1KG21CS031	43
30	1KG21CS032	41
31	1KG21CS033	43
32	1KG21CS034	38
33	1KG21CS035	38
34	1KG21CS036	38
35	1KG21CS037	44
36	1KG21CS038	40

Sl NO.	USN	21CS644
37	1KG21CS039	36
38	1KG21CS040	35
39	1KG21CS041	42
40	1KG21CS042	30
41	1KG21CS043	40
42	1KG21CS044	39
43	1KG21CS045	44
44	1KG21CS046	39
45	1KG21CS047	32
46	1KG21CS048	44
47	1KG21CS050	28
48	1KG21CS051	32
49	1KG21CS052	36
50	1KG21CS053	38
51	1KG21CS054	50
52	1KG21CS055	46
53	1KG21CS056	30
54	1KG21CS057	49
55	1KG21CS058	45
56	1KG21CS059	44
57	1KG21CS060	44
58	1KG21CS061	48
59	1KG21CS062	44
60	1KG21CS063	35
61	1KG21CS064	48
62	1KG21CS065	44
63	1KG21CS066	30
64	1KG21CS067	46
65	1KG21CS068	45
66	1KG21CS069	28
67	1KG21CS070	44
68	1KG21CS071	28
69	1KG21CS072	48
70	1KG21CS073	40
71	1KG21CS074	46
72	1KG21CS075	46
73	1KG21CS076	45
74	1KG21CS077	48
75	1KG21CS078	39

Sl NO.	USN	21CS644
76	1KG21CS079	22
77	1KG21CS080	36
78	1KG21CS081	42
79	1KG21CS082	32
80	1KG21CS083	45
81	1KG21CS084	43
82	1KG21CS085	47
83	1KG21CS086	44
84	1KG21CS087	39
85	1KG21CS088	39
86	1KG21CS089	42
87	1KG21CS090	41
88	1KG21CS091	41
89	1KG21CS092	40
90	1KG21CS093	45
91	1KG21CS094	25
92	1KG21CS095	47
93	1KG21CS096	45
94	1KG21CS097	21
95	1KG21CS098	32
96	1KG21CS099	45
97	1KG21CS100	42
98	1KG21CS101	42
99	1KG21CS102	43
100	1KG21CS103	44
101	1KG21CS104	43
102	1KG21CS106	38
103	1KG21CS108	47
104	1KG21CS109	43
105	1KG21CS110	24
106	1KG21CS111	38
107	1KG21CS112	35
108	1KG21CS113	40
109	1KG21CS114	43
110	1KG21CS115	49
111	1KG21CS116	38
112	1KG21CS117	36
113	1KG21CS118	43
114	1KG21CS119	41

Sl NO.	USN	21CS644
115	1KG21CS120	47
116	1KG21CS121	44
117	1KG21CS122	36
118	1KG21CS123	36
119	1KG21CS124	40
120	1KG21CS125	46
121	1KG21CS126	35
122	1KG22CS400	33
123	1KG22CS401	32
124	1KG22CS402	32
125	1KG22CS403	42
126	1KG22CS404	34
127	1KG22CS405	29
128	1KG22CS406	36
129	1KG22CS407	40
130	1KG22CS408	36
131	1KG22CS409	45
132	1KG22CS410	34
133	1KG22CS411	38

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024 (EVEN SEMESTER)

ACTIVITY - QUIZ : DATA SCIENCE AND VISUALIZATION(21CS644)

Timestamp	Email Address	Score	Name	USN	Seme-ster	Secti-on	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
01-08-A2:AA222024	abburipallavi123@gmail.com	20 / 20	Abburi pallavi	1KG21CS001	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:46:19	abhijeetdaz56@gmail.com	20 / 20	Abhijeet Das	1kg21cs002	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:31:15	abilashbr182@gmail.com	20 / 20	Abhilash B R	1KG21CS003	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:28:47	sjvabhishek@gmail.com	20 / 20	Abhishek v	1kg21cs004	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:03:17	akhilaamarnath26@gmail.com	20 / 20	Akhila A	1KG21CS005	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-5-2024 14:21:14	akshathagowda2322@gmail.com	20 / 20	Akshatha R Gowda	1KG21CS006	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:45:55	alluchinnikrishna758@gmail.com	20 / 20	Allu.chinni krishna	1KG21CS007	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:57:02	amithsuri818@gmail.com	20 / 20	amith C Suri	1KG21CS008	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:21:25	amogh.official2019@gmail.com	20 / 20	Amogh A	1KG21CS009	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:58:59	ankithavenkatesh04@gmail.com	20 / 20	Ankitha Venkatesh	1KG21CS010	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 18:27:36	ankushsonugowda@gmail.com	20 / 20	Ankush Gowda K	1KG21CS011	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-2-2024 13:04:25	arpithasarpithas1@gmail.com	20 / 20	Arpitha S	1KG21CS012	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 18:55:43	ashwiniashu0458@gmail.com	20 / 20	Ashwini C	1KG21CS013	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 21:13:03	nayanabashyam@gmail.com	20 / 20	B Nayana	1KG21CS014	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:50:18	battapreethi8@gmail.com	20 / 20	B.Preethi	1KG21CS015	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:39:14	akhilyadavbs07@gmail.com	20 / 20	B.S Akhil	1KG21CS016	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:38:56	bharathgowdaj18@gmail.com	20 / 20	Bharath Gowda J	1KG21CS017	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 16:56:47	bhavanad2516@gmail.com	20 / 20	Bhavana D	1kg21cs018	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:38:59	bhavanaseenappa@gmail.com	20 / 20	Bhavana S	1KG21CS019	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:25:34	susmithachintagumpala@gmail.com	20 / 20	Sushmitha C	1KG21CS021	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:42:03	harikishorechalla28@gmail.com	20 / 20	Hari Kishore	1KG21CS022	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:23:10	chandantavane99@gmail.com	20 / 20	Chandan Tavane	1KG21CS023	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:39:25	sudhachinmy09@gmail.com	20 / 20	Chinmaiy P	1KG21CS024	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:38:41	darshur12@gmail.com	20 / 20	Darshan R	1KG21CS026	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:11:34	deekshithar1307@gmail.com	20 / 20	Deekshitha R	1KG21CS027	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b

8-1-2024 14:34:47	deepakr0320@gmail.com	20 / 20	Deepak Athresh R	1KG21CS028	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:36:25	devdaskishanrao@gmail.com	20 / 20	Devdas	1KG21CS029	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:34:44	dhakshithagowda@gmail.com	20 / 20	Dhakshitha A	1KG21CS030	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:58:20	dhanushgp6@gmail.com	20 / 20	Dhanush G P	1KG21CS031	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:21:38	dhanushupallapaate@gmail.com	20 / 20	DHANUSH US	1KG21CS032	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:47:20	dharini.nandhini@gmail.com	20 / 20	Dharini	1KG21CS033	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:38:55	dheerajraikar2003@gmail.com	20 / 20	Dheeraj D Raika	1KG21CS034	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 20:40:37	sahudibyajyoti61@gmail.com	20 / 20	Dibyajyoti sahu	1KG21CS035	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:55:21	dnshmk85jly@gmail.com	20 / 20	Dinesh J L	1KG21CS036	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:52:11	ddiv2154@gmail.com	20 / 20	Divya H U	1KG21CS037	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:42:07	paladivyayadav0528@gmail.com	20 / 20	Divya P	1KG21CS038	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:36:14	geoffreysamuel28@gmail.com	20 / 20	Geoffrey Samue	1KG21CS039	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:06:54	raghusunny394@gmail.com	20 / 20	Raghu	1KG21CS040	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:30:42	sreed3576@gmail.com	20 / 20	G.Divyasree	1kg21cs041	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-5-2024 12:52:44	bharathchowdary1560@gmail.com	20 / 20	G.bharath	1kg21cs042	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 19:32:34	gururangar18@gmail.com	20 / 20	Gururaj B	1kg21cs043	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:26:38	hanochchristian@gmail.com	20 / 20	Hanoch Christian R	1KG21CS044	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:46:35	harshithkpgowda30@gmail.com	20 / 20	Harshith KP	1KG21CS045	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:34:53	harshithadg97@gmail.com	20 / 20	Harshitha D G	1KG21CS046	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:30:57	hiteshareddy3@gmail.com	20 / 20	Hitesh A Reddy	1KG21CS047	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:12:16	jenvinaidu2325@gmail.com	20 / 20	Jahanavi S	1KG21CS048	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:39:32	kjprakruthi@gmail.com	20 / 20	Kj Prakruthi	1KG21CS050	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:34:52	nithishnaidu18@gmail.com	20 / 20	K NITHISH	1KG21CS051	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 18:45:54	rohithnaidu321@gmail.com	20 / 20	K Rohith	1KG21CS052	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:51:09	karthikkg924@gmail.com	20 / 20	Karthik G	1KG21CS053	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 16:55:30	kasturipoornima@gmail.com	20 / 20	K poornima	1KG21CS054	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:40:50	kishorabd2003@gmail.com	20 / 20	Kishor Kumar I	1KG21CS055	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:29:26	kunalaganesh1118@gmail.com	20 / 20	ganesh	1kg21cs056	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:09:17	kushmitha8904@gmail.com	20 / 20	Kushmitha T A	1KG21CS057	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 18:25:31	kusumashiva523@gmail.com	20 / 20	Kusuma.B	1KG21CS058	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:25:24	lekkalapudilavanya@gmail.com	20 / 20	L LAVANYA	1KG21CS059	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 16:36:26	lakmyscbe@gmail.com	20 / 20	Laksha Senthilkumar	1KG21CS060	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:00:19	mpooja2927@gmail.com	20 / 20	M Pooja	1KG21CS061	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:17:07	surabhimkashyap@gmail.com	20 / 20	M Surabhi	1KG21CS062	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:32:38	bhuvanamadineni229@gmail.com	20 / 20	Madineni Bhuvana	1kg21cs063	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b

8-1-2024 14:49:03	akshaykanakapura@gmail.com	20 / 20	Akshay U	1KG22CS400	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-2-2024 12:54:38	balajishetty36@gmail.com	20 / 20	Balaji N	1KG22CS401	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-2-2024 10:33:12	bhavanam791@gmail.com	20 / 20	Bhavana M	1KG22CS402	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 15:06:37	dhanushbhovi@gmail.com	20 / 20	Dhanush R	1KG22CS403	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
8-1-2024 14:53:55	dhanu28shree@gmail.com	20 / 20	Dhanushree A	1kg22cs404	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b
08-02-2024 12:45	kirunayak14@gmail.com	20 / 20	Kiran Kumar	1KG22CS405	6	A	c	c	c	d	b	a	d	b	c	a	d	b	a	a	d	d	b	a	c	b

Course In charge

HOD

HOD
Department of Computer Science E[illegible]g
K.S School of Engineering & Management
Bangalore-560109

K.S. School of Engineering and Management- 560109
DEARTMENT OF COMPUTER SCIENCS & ENGINEERING
TEACHING AND LEARNING

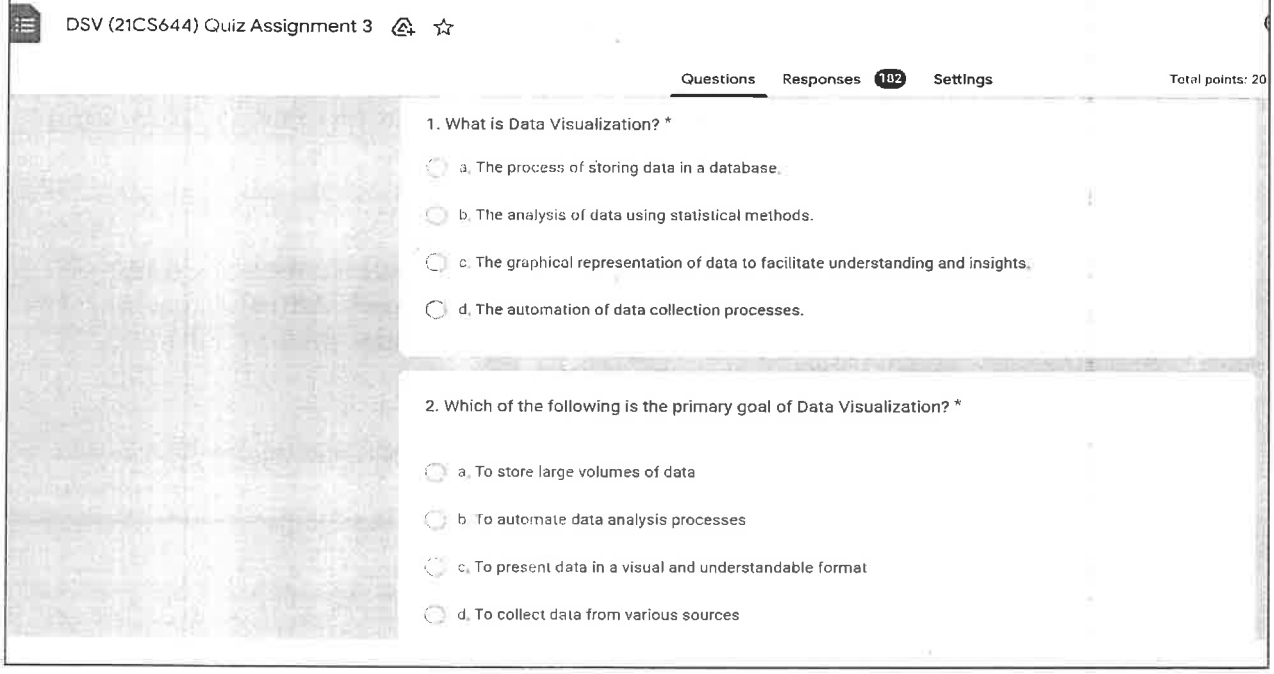
PEDAGOGY REPORT

Academic Year	**2023-24**
Name of the Faculty	**Mrs. KAVITHA K S**
Course Name /Code	**Data Science and Visualization/21CS644**
Semester/Section	**'VI-A' Section**
Activity Name	**Quiz**
Topic Covered	**Data Visualization and Data Exploration**
Date	**08/01/2024**
No. of Participants	**66 Students**
Objectives/Goals	**To analyze the understanding of the students regarding the subject**
ICT Used	**Online**

Appropriate Method/Instructional materials/Exam Questions/CO

CO1: Explore the data in different forms and pre-processing techniques for data science.
CO2: Apply different techniques to Explore Data Analysis and the Data Science Process.
CO3: Interpret feature selection algorithms & design a recommender system.
CO4: Make use of data visualization tools and libraries and plot graphs.
CO5: Develop different charts and include mathematical expressions.

The students were given 20 multiple choice questions from module 4 and 5. It was conducted in online mode.
Questions should be answered:


DSV (21CS644) Quiz Assignment 3

Questions Responses 102 Settings Total points: 20

1. What is Data Visualization? *

a. The process of storing data in a database.

b. The analysis of data using statistical methods.

c. The graphical representation of data to facilitate understanding and insights.

d. The automation of data collection processes.

2. Which of the following is the primary goal of Data Visualization? *

a. To store large volumes of data

b. To automate data analysis processes

c. To present data in a visual and understandable format

d. To collect data from various sources

3. What are some common elements used in Data Visualization? *

- ◯ a. Spreadsheets and word processors
- ◯ b. Tables and paragraphs
- ◯ c. Charts, graphs, and infographics
- ◯ d. Audio and video clips

4. What type of Data Visualization is best suited for showing the distribution of a continuous dataset? *

- ◯ a. Bar chart
- ◯ b. Line chart
- ◯ c. Pie chart
- ◯ d. Histogram

Relevant PO's:	**PO: 1, 2,3,4,5,6,10,11 and 12**
Significance of Results/Outcomes	Students learning data visualization will gain the ability to understand, interpret, and communicate complex data through visual representations, ultimately enhancing their analytical and storytelling skills.
Reflective Critique	Students will learn to identify patterns, trends, and outliers in data that might be missed in raw data form.

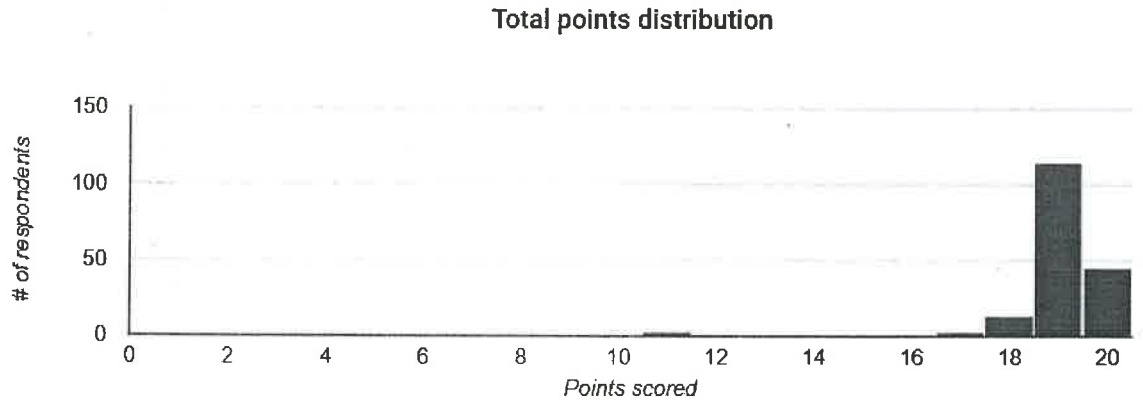
Proofs (Photographs/Videos/Reports/Charts/Models)

DSV (21CS644) Quiz Assignment 3 (Responses)

File Edit View Insert Format Data Tools Extensions Help

A1 8/1/2024 14:27:50

	A	B	C	D	E	F	G	H	I	J	K	L
34	8/1/2024 14:28:44	deepakr0320@gmail.com	18 / 20	Deepak Athresh R	1KG21CS028	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
35	8/1/2024 14:34:47	deepakr0320@gmail.com	20 / 20	Deepak Athresh R	1KG21CS028	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
36	8/1/2024 14:36:25	devdaskishanrao@gmail.com	19 / 20	Devdas	1KG21CS029	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
37	8/1/2024 14:26:36	dhakshithagowda@gmail.com	19 / 20	Dhakshitha A	1KG21CS030	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
38	8/1/2024 14:34:44	dhakshithagowda@gmail.com	20 / 20	Dhakshitha A	1KG21CS030	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
39	8/1/2024 14:58:20	dhanushgp6@gmail.com	20 / 20	Dhanush G P	1KG21CS031	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
40	8/1/2024 21:19:10	dhanushgp6@gmail.com	20 / 20	DHANUSH GP	1KG21CS031	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
41	8/1/2024 14:21:38	dhanushupallapeate@gmail.com	19 / 20	DHANUSH US	1KG21CS032	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
42	8/1/2024 14:40:22	dharini.nandhini@gmail.com	17 / 20	Dharini	1KG21CS033	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
43	8/1/2024 14:47:20	dharini.nandhini@gmail.com	20 / 20	Dharini	1KG21CS033	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
44	8/1/2024 14:52:42	amudanandhini@gmail.com	20 / 20	Dharini	1KG21CS033	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
45	8/1/2024 15:38:55	dheerajraikar2003@gmail.com	20 / 20	Dheeraj D Raikar	1KG21CS034	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
46	8/1/2024 20:35:13	sahudibyajyoti61@gmail.com	19 / 20	Dibyajyoti sahu	1KG21CS035	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
47	8/1/2024 20:38:04	sahudibyajyoti61@gmail.com	19 / 20	Dibyajyoti sahu	1KG21CS035	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
48	8/1/2024 20:40:37	sahudibyajyoti61@gmail.com	20 / 20	Dibyajyoti sahu	1KG21CS035	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
49	8/1/2024 14:50:11	dnshmk85jly@gmail.com	18 / 20	Dinesh J L	1KG21CS036	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
50	8/1/2024 14:52:10	dnshmk85jly@gmail.com	19 / 20	Dinesh J L	1KG21CS036	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
51	8/1/2024 14:55:21	dnshmk85jly@gmail.com	20 / 20	Dinesh J L	1KG21CS036	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
52	8/1/2024 14:52:11	ddlv2154@gmail.com	20 / 20	Divya H U	1KG21CS037	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
53	8/1/2024 14:42:07	paladivyayadav0528@gmail.com	20 / 20	Divya P	1KG21CS038	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
54	8/1/2024 14:36:14	geoffreysamuel28@gmail.com	19 / 20	Geoffrey Samuel	1KG21CS039	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart
55	8/1/2024 15:06:54	raghusunny394@gmail.com	19 / 20	Raghu	1KG21CS040	6	A	c. The graphical represer	c. To present data in a vis	c. Charts, graphs, and inf	d. Histogram	b. Bar chart



Signature of Course In charge

Signature of HOD CSE

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

KSSEM

K.S. SCHOOL OF ENGINEERING AND MANAGEMENT, BANGALORE - 560109

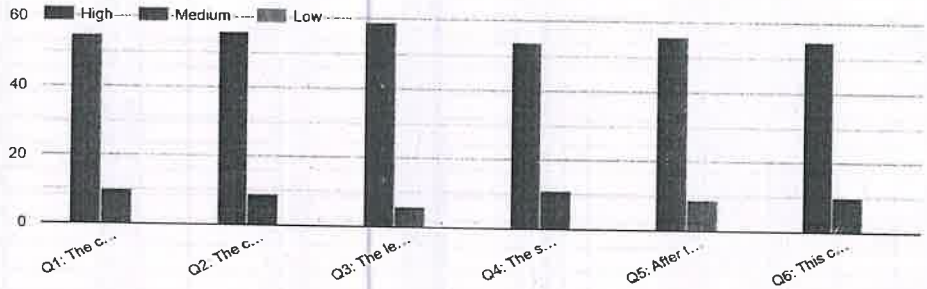
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESSION: 2023-2024 (EVEN SEMESTER)

COURSE END SURVEY: DATA SCIENCE AND VISUALIZATION(21CS644) SEMESTER: VI A SECTION

Timestamp	EMAIL	NAME OF THE STUDENT	USN	[Q1: The course increased your level of interest?]	[Q2: The course content was appropriate and was presented in a structured manner]	[Q3: The learning material, practical sessions were relevant to the course outcomes]	[Q4: The self-study (including reading) required for this course will ensure better achievement of course objectives]	[Q5: After this course, you will be able to solve analyze real life problems related to this course]	[Q6: This course has given you enough understanding to take next level courses]	Signature
7-25-2024 15:19:49	akhilaamarnath26@gmail.com	Akhila A	1KG21CS005	Medium	High	High	Medium	Medium	Medium	
7-25-2024 15:20:35	darshur12@gmail.com	Darshan R	1KG21CS026	High	High	High	High	High	High	
7-25-2024 15:20:38	surabhimkashyap@gmail.com	M Surabhi	1KG21CS062	High	High	High	High	High	High	
7-25-2024 15:20:56	kasturipoornima@gmail.com	K poornima	1KG21CS054	Medium	High	High	High	High	High	
7-25-2024 15:21:45	battapreethi8@gmail.com	B.Preethi	1KG21CS015	Medium	Medium	Medium	Medium	Medium	Medium	
7-25-2024 15:23:18	ashwiniashu0458@gmail.con	Ashwini C	1KG21CS013	High	High	High	High	High	High	
7-25-2024 15:29:55	deepakr0320@gmail.com	Deepak Athresh R	1KG21CS028	Medium	High	High	Medium	Medium	Medium	
7-25-2024 15:36:45	abilashbr182@gmail.com	ABHILASH B R	1KG21CS003	High	High	High	High	High	High	
7-25-2024 15:37:21	divyasreeg61@gmail.com	G.Divyasree	1kg21cs041	Medium	Medium	High	Medium	High	Medium	
7-25-2024 15:40:05	lakmyscbe@gmail.com	Laksha Senthilkumar	1KG21CS060	High	High	High	High	High	High	
7-25-2024 15:44:15	hanochchristian@gmail.com	Hanoch Christian R	1KG21CS044	High	High	High	High	High	High	
7-25-2024 15:45:11	kusumashiva523@gmail.com	Kusuma B	1KG21CS058	High	High	High	High	High	High	
7-25-2024 15:50:52	bhavanam791@gmail.com	Bhavana M	1KG22CS402	High	High	High	Medium	High	High	
7-25-2024 15:53:09	dhakshithagowda@gmail.com	Dhakshitha A	1KG21CS030	High	High	High	High	High	High	
7-25-2024 15:54:44	dhanushgp6@gmail.com	DHANUSH GP	1KG21Cs031	High	High	High	High	High	High	
7-25-2024 16:02:40	bharathgowdaj18@gmail.com	Bharath Gowda J	1KG21CS017	High	High	High	High	High	High	
7-25-2024 16:13:19	lekkalapudilavanya@gmail.com	L LAVANYA	1KG21CS059	High	High	High	High	High	High	
7-25-2024 16:18:13	dharini.nandhini@gmail.com	Dharini	1KG21CS033	High	High	High	High	High	High	
7-25-2024 16:19:36	bhavanad2516@gmail.com	Bhavana D	1KG21CS018	High	High	High	High	High	High	
7-25-2024 16:26:26	abhijeetdaz56@gmail.com	Abhijeet Das	1kg21cs002	Medium	Medium	Medium	Medium	Medium	Medium	
7-25-2024 16:44:08	sudhachinmy09@gmail.com	Chinmaiy P	1KG21CS024	High	High	High	High	High	High	
7-25-2024 16:52:16	kirunayak14@gmail.com	Kiran Kumar	1KG22CS405	High	High	High	High	High	High	
7-25-2024 16:52:33	bhavanaseenappa@gmail.com	Bhavana S	1KG21CS019	High	High	High	High	High	High	
7-25-2024 17:17:11	dhanushbhovi@gmail.com	Dhanush R	1KG22CS403	High	High	High	High	High	High	
7-25-2024 17:30:42	akshaykanakapura@gmail.com	Akshay I	1KG22CS400	High	High	High	High	High	High	
7-25-2024 17:43:03	janvinaidu2325@gmail.com	Jahanavi S	1KG21CS048	High	High	High	High	High	High	
7-25-2024 17:51:07	harshithkpgowda30@gmail.com	Harshith KP	1KG21CS045	High	High	High	High	High	High	
7-25-2024 18:12:09	dheerajraikar2003@gmail.com	Dheeraj D Raikar	1kg21cs034	High	High	High	High	High	High	
7-25-2024 19:04:12	akhilyadavbs11@gmail.com	B.S Akhil	1KG21CS016	High	High	High	High	High	High	
7-25-2024 19:15:00	bhuvanamadineni229@gmail.com	Madineni Bhuvana	1kg21cs063	High	High	High	High	High	High	
7-25-2024 20:09:38	ddiv2154@gmail.com	Divya H U	1KG21CS037	High	High	High	High	High	High	
7-25-2024 20:13:30	susmithachintagumpala@gmail.c	Sushmitha	1KG21cs021	High	Medium	High	High	High	High	

7-25-2024 22:32:58	gurubangar18@gmail.com	Gururaj B	1kg21cs043	High	High	High	High	High	High	
7-26-2024 0:26:35	harshithadg0@gmail.com	Harshitha D.G	1KG21CS046	Medium	Medium	Medium	Medium	Medium	Medium	
7-26-2024 12:02:26	deekshithar1307@gmail.com	Deekshitha R	1KG21CS027	High	High	High	High	High	High	
7-26-2024 12:05:35	kjprakruthi@gmail.com	KJ PRAKRUTHI	1KG21CS050	High	High	High	High	High	High	
7-26-2024 12:50:40	arpithasarpithas1@gmail.com	Arpitha S	1KG21CS012	Medium	Medium	Medium	Medium	Medium	Medium	
7-26-2024 14:18:03	mpooja2927@gmail.com	M Pooja	1KG21CS061	High	High	High	High	High	High	
7-26-2024 17:29:38	nayanabashyam@gmail.com	B NAYAna	1kg21cs014	High	High	High	High	High	High	
7-26-2024 18:26:01	kushmitha8904@gmail.com	Kushmitha T A	1KG21CS057	High	Medium	High	Medium	Medium	Medium	
7-26-2024 19:54:51	karthikkg924@gmail.com	Karthik G	1KG21CS053	High	High	High	High	High	High	
7-26-2024 19:55:57	devdaskishanrao@gmail.com	Devdas	1KG21CS029	High	High	High	High	High	High	
7-26-2024 19:56:39	paladivyayadav0528@gmail.com	Divya P	1KG21CS038	Medium	Medium	Medium	Medium	Medium	Medium	
7-26-2024 20:03:16	ankushsonugowda@gmail.com	Ankush Gowda K	1KG21CS011	High	High	High	High	High	High	
7-26-2024 20:05:03	dhanu28shree@gmail.com	Dhanushree A	1kg22cs404	High	High	High	High	High	High	
7-26-2024 20:06:13	amithsuri818@gmail.com	Amith C Suri	1KG21CS008	High	High	High	High	High	High	
7-26-2024 20:17:47	kishorabd2003@gmail.com	Kishor Kumar l	1KG21CS055	High	High	High	High	High	High	
7-26-2024 20:39:51	harikishorechalla28@gmail.com	Challa HariKishore Naidu	1KG21CS022	High	High	High	High	High	High	
7-26-2024 21:18:37	geoffreysamuel28@gmail.com	Geoffrey Samuel	1KG21CS039	High	High	High	High	High	High	
7-27-2024 8:36:01	alluchinnikrishna758@gmail.com	Allu.chinni krishna	1KG21CS007	High	High	High	High	High	High	
7-27-2024 8:54:24	kunalaganesh1118@gmail.com	ganesh	1kg21cs056	High	High	High	High	High	High	
7-27-2024 9:12:34	nithishnaidu18@gmail.com	K NITHISH	1KG21CS051	High	High	High	High	High	High	
7-27-2024 10:51:24	akshathagowda2322@gmail.com	Akshatha R Gowda	1KG21CS006	High	High	High	High	High	High	
7-27-2024 10:56:34	raghusunny394@gmail.com	Raghu	1KG21CS040	High	High	High	High	High	High	
7-27-2024 10:56:47	dnshmk85jly@gmail.com	Dinesh J L	1KG21CS036	High	High	High	High	High	High	
7-27-2024 12:16:25	hiteshareddy3@gmail.com	Hitesh A Reddy	1KG21CS047	High	High	High	High	High	High	
7-27-2024 13:55:47	amogh.official2019@gmail.com	Amogh A	1KG21CS009	High	High	High	High	High	High	
7-27-2024 14:09:59	ankithavenkatesh04@gmail.com	Ankitha Venkatesh	1KG21CS010	High	High	High	High	High	High	
7-27-2024 14:14:13	chandantavane99@gmail.com	Chandan Tavane	1KG21CS023	High	High	High	High	High	High	
7-27-2024 16:36:51	balajishetty36@gmail.com	Balaji N	1KG22CS401	High	High	High	High	High	High	
7-27-2024 19:18:30	sahudibyajyoti61@gmail.com	Dibyajyoti sahu	1KG21CS035	Medium	Medium	Medium	Medium	Medium	Medium	
7-27-2024 20:25:02	abburipallavi123@gmail.com	Abburi pallavi	1kg21cs001	High	High	High	High	High	High	
7-28-2024 14:59:19	sjvabhishek@gmail.com	Abhishek v	1KG21CS004	High	High	High	High	High	High	
8-1-2024 13:36:37	dhanushupallapaate@gmail.com	DHANUSH US	1KG21CS032	High	High	High	High	High	High	
8-5-2024 12:55:30	Bharathchowdary1560@gmail.co	G.bharath	1kg21cs042	High	High	High	High	High	High	



Faculty Incharge

HOD

HOD
Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

K. S. School of Engineering & Management, Bangalore - 560109

Department of Computer Science Engineering

Staff Feedback (2023-24) Even Sem

Sixth Sem 'A' Section

Faculty Name: Mrs. Kavitha K S

Course Name & Code: Data Science & Visualization/21CS644

Class Strength:66

Sl. No.	1. Effective Planning & organisation	2. Punctuality / Class time Utilization	3. Ability to teach /explain / effective use of board	4. Interaction / Motivating students	5. Subject knowledge	6. Presentation of the subject / communication	7. Linking subject with practical application	8. Syllabus covearage / exam point of view	9. Evaluation of test / counselling	10. Attitude towards students
1	5	5	5	5	5	5	5	5	5	5
2	5	5	5	5	5	5	5	5	5	5
3	5	5	5	5	5	5	5	5	5	5
4	5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5	5
6	5	5	5	5	5	5	5	5	5	5
7	5	5	5	5	5	5	5	5	5	5
8	5	5	5	5	5	5	5	5	5	5
9	5	5	5	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5	5	5	5
11	5	5	5	5	5	5	5	5	5	5
12	5	5	5	5	5	5	5	5	5	5
13	3	3	3	3	3	3	3	3	3	3
14	3	3	3	3	3	3	3	3	3	3
15	4	4	5	5	4	4	5	5	5	5
16	5	5	5	5	5	5	5	5	5	5
17	5	5	5	5	5	5	5	5	5	5
18	5	5	5	5	5	5	5	5	5	5
19	5	5	5	5	5	5	5	5	5	5
20	5	5	5	5	5	5	5	5	5	5
21	5	5	5	5	5	5	5	5	5	5
22	5	5	5	5	5	5	5	5	5	5
23	5	5	5	5	5	5	5	5	5	5
24	5	5	5	5	5	5	5	5	5	5
25	5	5	5	5	5	5	5	5	5	5
26	4	4	4	4	4	4	4	4	4	4
27	5	4	5	4	5	5	5	4	5	5
28	5	5	5	5	5	5	5	5	5	5
29	5	5	5	5	5	5	5	5	5	5
30	5	5	5	5	5	5	5	5	5	5
31	5	5	5	5	5	5	5	5	5	5
32	5	5	5	5	5	5	5	4	5	5
33	5	5	5	5	5	5	5	5	5	5
34	5	5	5	5	5	5	5	5	5	5
35	5	5	5	5	5	5	5	5	5	5
36	5	5	5	5	5	5	5	5	5	5
37	5	5	5	5	5	5	5	5	5	5

38	5	5	5	5	5	5	5	5	5	5
39	5	5	5	5	5	5	5	5	5	5
40	5	5	5	5	5	5	5	5	5	5
41	5	5	5	5	5	5	5	5	5	5
42	5	5	5	5	5	5	5	5	5	5
43	5	5	5	5	5	5	5	5	5	5
44	5	5	5	5	5	5	5	5	5	5
45	5	5	5	5	5	5	5	5	5	5
46	5	5	5	5	5	5	5	5	5	5
47	5	5	5	5	5	5	5	5	5	5
48	5	5	5	5	5	5	5	5	5	5
49	5	5	5	5	5	5	5	5	5	5
50	5	5	5	5	5	5	5	5	5	5
51	5	5	5	5	5	5	5	5	5	5
52	5	5	5	5	5	5	5	5	5	5
53	5	5	5	5	5	5	5	5	5	5
54	4	4	4	3	4	4	4	4	4	4
55	5	5	5	5	5	5	5	5	5	5
56	5	5	5	5	5	5	5	5	5	5
57	5	5	5	5	5	5	5	5	5	5
58	5	5	5	5	5	5	5	5	5	5
59	5	5	5	5	5	5	5	5	5	5
60	5	5	5	5	5	5	5	5	5	5
Col. Total	293	292	294	292	293	293	294	292	294	294
Col. Avg.	4.88	4.87	4.90	4.87	4.88	4.88	4.90	4.87	4.90	4.90
Over all %	97.70									

Head of Department

HOD

Department of Computer Science Engineering
K.S School of Engineering & Management
Bangalore-560109

Principal