Long answer questions

1. What is data science and what are the benefits of data science?

A. Data science involves using methods to analyze **massive amounts of data** and extract the knowledge it contains. The data in data science is collected from different sources like conducting surveys, social media, internet searches and so on.

While collecting data we should follow 3 C's:

- **Curiosity**
- **Common sense**
- **Communication skills**

Data science is used in many sectors such as:

i. **Self automated cars:** Self-driving cars are vehicles that can drive themselves, without the need for a human driver. They use deep learning algorithms to process data and make decisions about steering, braking, and acceleration.

ii. **Logistics:** Identifying the most efficient delivery routes based on factors like distance, traffic, and weather conditions to minimize transportation costs.

iii. **Airlines:** Airlines use data science to make better decisions about pricing, routes, and maintenance. They also use data science to improve safety, customer service.

iv. **Netflix:** Netflix uses recommendation algorithms powered by data science to analyze user viewing history, ratings, and preferences to suggest content to every individual users as per their interest.

v. **Maps:** This includes traffic prediction, route optimization, and personalized experiences.

vi. **Election voting:** Data science is used to predict and analyze which party is going to rule for the next 5 years basing on their previous work.

Benefits of data science

i. **Gain insights into their customers:** Data science provides benefits to improve business by considering customers interests, identifying buying patterns, etc. Basing on this they improve business by providing discounts, schemes and many more.

ii. **Hire the right players and pit them against the opponents:** Data science provides benefits even in sports by applying statistics to choose right players basing on their opponents to win the match.
Example: In cricket, captains apply statistics to choose bating/bowling after winning the toss.

iii. **Stock Market analysis:** Financial institutions use data science to predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services.

iv. **Online courses:** Recommends courses to the users on their searches, data science is used even to track online course and increase/decrease complexities in the assignments.

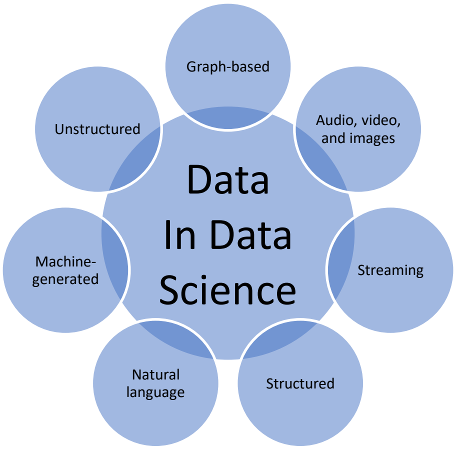
So, data science is mainly needed for

- **Better decision making**-whether to choose A or B?
- **Predictive analysis**-What will happen next?
- **Pattern analysis**-Is there any hidden information in the data?

2. Explain about facets of data with examples.

A. Data can be in different forms. It can be in the forms of text, numbers, image, audio, video and graphs.

The different facets of data are:



i. Structured data

Structured data in generally present in tables within databases or Excel files. Structured Query Language is the preferred way to manage structured data.

Example:

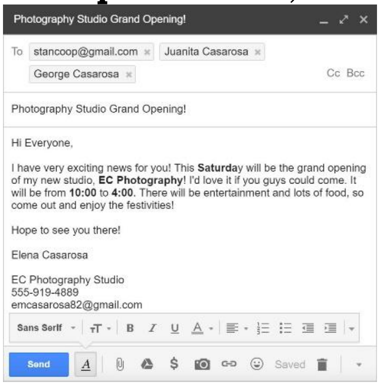
Age	**Income**	**Student**	**Credit_rating**	**Buys_computer**
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31-40	High	No	Fair	Yes

>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
31-40	Low	Yes	Excellent	No
<=30	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
>40	Low	Yes	Fair	Yes
<=30	Medium	Yes	Fair	Yes

ii. Unstructured data

Data is generally in natural Language. Information that doesn't have a predefined structure or model, and is often called raw data. It can be text-heavy, but can also include numbers, dates, and facts. As of now we are able to do text summarization, text completion, and sentiment analysis.

Examples: E-Mail, Wikipedia, Amazon reviews



Photography Studio Grand Opening!

To stancoop@gmail.com, Juanita Casarosa, George Casarosa

Photography Studio Grand Opening!

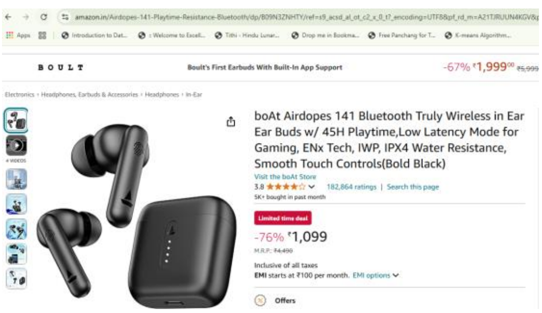
Hi Everyone,

I have very exciting news for you! This **Saturday** will be the grand opening of my new studio, **EC Photography**! I'd love it if you guys could come. It will be from **10:00** to **4:00**. There will be entertainment and lots of food, so come out and enjoy the festivities!

Hope to see you there!

Elena Casarosa

EC Photography Studio
555-919-4889
emcasarosa82@gmail.com



amazon.in/Airdopes-141-Playtime-Resistance-Bluetooth/dp/B09N3ZNHTY/ref=s9_acsd_al_ot_c2_x_0_t?_encoding=UTF8&pf_rd_m=A21TJRUUN4KGV8

BOULT — Boult's First Earbuds With Built-In App Support — -67% ₹1,999⁰⁰ ~~₹5,999~~

Electronics › Headphones, Earbuds & Accessories › Headphones › In-Ear

boAt Airdopes 141 Bluetooth Truly Wireless in Ear Ear Buds w/ 45H Playtime,Low Latency Mode for Gaming, ENx Tech, IWP, IPX4 Water Resistance, Smooth Touch Controls(Bold Black)

Visit the boAt Store

3.8 ★★★★☆ 182,864 ratings | Search this page

5K+ bought in past month

Limited time deal

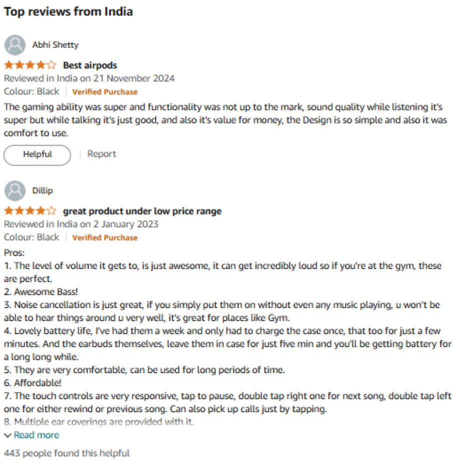
-76% ₹1,099

M.R.P.: ~~₹4,490~~

Inclusive of all taxes

EMI starts at ₹100 per month. EMI options

Offers



Top reviews from India

Abhi Shetty

★★★★☆ **Best airpods**

Reviewed in India on 21 November 2024

Colour: Black | Verified Purchase

The gaming ability was super and functionality was not up to the mark, sound quality while listening it's super but while talking it's just good, and also it's value for money, the Design is so simple and also it was comfort to use.

Dillip

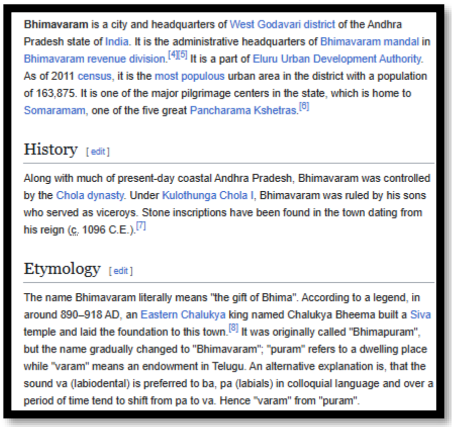
★★★★☆ **great product under low price range**

Reviewed in India on 2 January 2023

Colour: Black | Verified Purchase

Pros:
1. The level of volume it gets to, is just awesome, it can get incredibly loud so if you're at the gym, these are perfect.
2. Awesome Bass!
3. Noise cancellation is just great, if you simply put them on without even any music playing, u won't be able to hear things around u very well, it's great for places like Gym.
4. Lovely battery life, I've had them a week and only had to charge the case once, that too for just a few minutes. And the earbuds themselves, leave them in case for just five min and you'll be getting battery for a long long while.
5. They are very comfortable, can be used for long periods of time.
6. Affordable!
7. The touch controls are very responsive, tap to pause, double tap right one for next song, double tap left one for either rewind or previous song. Can also pick up calls just by tapping.
8. Multiple ear coverings are provided with it.

443 people found this helpful



Bhimavaram is a city and headquarters of West Godavari district of the Andhra Pradesh state of India. It is the administrative headquarters of Bhimavaram mandal in Bhimavaram revenue division.[4][5] It is a part of Eluru Urban Development Authority. As of 2011 census, it is the most populous urban area in the district with a population of 163,875. It is one of the major pilgrimage centers in the state, which is home to Somaramam, one of the five great Pancharama Kshetras.[6]

History [edit]

Along with much of present-day coastal Andhra Pradesh, Bhimavaram was controlled by the Chola dynasty. Under Kulothunga Chola I, Bhimavaram was ruled by his sons who served as viceroys. Stone inscriptions have been found in the town dating from his reign (c. 1096 C.E.).[7]

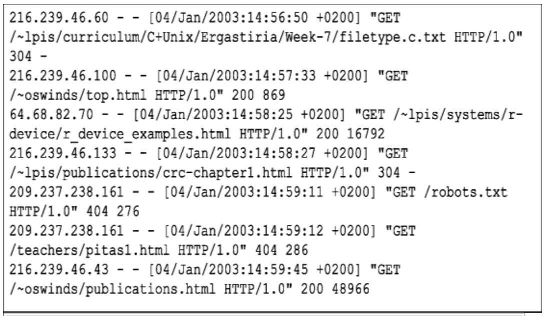
Etymology [edit]

The name Bhimavaram literally means "the gift of Bhima". According to a legend, in around 890–918 AD, an Eastern Chalukya king named Chalukya Bheema built a Siva temple and laid the foundation to this town.[8] It was originally called "Bhimapuram", but the name gradually changed to "Bhimavaram"; "puram" refers to a dwelling place while "varam" means an endowment in Telugu. An alternative explanation is, that the sound va (labiodental) is preferred to ba, pa (labials) in colloquial language and over a period of time tend to shift from pa to va. Hence "varam" from "puram".

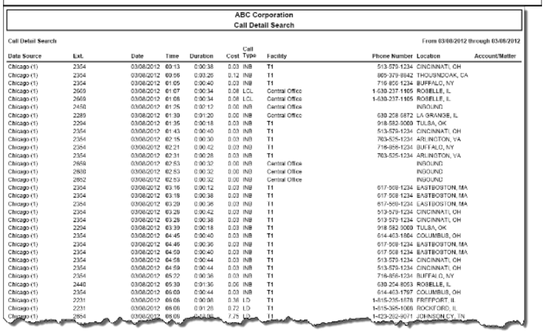
iii. Machine generated data

Automatically created by a computer, process, application, or other machine without human intervention.

Examples: Web server logs, Call detail records



```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET
/~lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0"
304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET
/~oswinds/top.html HTTP/1.0" 200 869
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /~lpis/systems/r-
device/r_device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET
/~lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt
HTTP/1.0" 404 276
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET
/teachers/pitas1.html HTTP/1.0" 404 286
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET
/~oswinds/publications.html HTTP/1.0" 200 48966
```



ABC Corporation
Call Detail Search

Call Detail Search

From 03/08/2012 through 03/08/2012

Data Source	Ext.	Date	Time	Duration	Cost	Call Type	Facility	Phone Number	Location	Account/Matter
Chicago (1)	2354	03/08/2012	00:13	0:00:38	0.03	INB	T1	513-579-1234	CINCINNATI, OH	
Chicago (1)	2354	03/08/2012	00:56	0:03:26	0.12	INB	T1	805-379-8842	THOUSNDOAK, CA	
Chicago (1)	2354	03/08/2012	01:05	0:00:40	0.03	INB	T1	716-856-1234	BUFFALO, NY	
Chicago (1)	2669	03/08/2012	01:07	0:00:34	0.08	LCL	Central Office	1-630-237-1105	ROSELLE, IL	
Chicago (1)	2669	03/08/2012	01:08	0:00:34	0.08	LCL	Central Office	1-630-237-1105	ROSELLE, IL	
Chicago (1)	2450	03/08/2012	01:25	0:02:12	0.00	INB	Central Office		INBOUND	
Chicago (1)	2289	03/08/2012	01:30	0:01:20	0.00	INB	Central Office	630-258-6872	LA GRANGE, IL	
Chicago (1)	2294	03/08/2012	01:35	0:00:18	0.03	INB	T1	918-582-9000	TULSA, OK	
Chicago (1)	2354	03/08/2012	01:43	0:00:40	0.03	INB	T1	513-579-1234	CINCINNATI, OH	
Chicago (1)	2354	03/08/2012	02:15	0:00:30	0.03	INB	T1	703-525-1234	ARLINGTON, VA	
Chicago (1)	2354	03/08/2012	02:21	0:00:42	0.03	INB	T1	716-856-1234	BUFFALO, NY	
Chicago (1)	2354	03/08/2012	02:31	0:00:28	0.03	INB	T1	703-525-1234	ARLINGTON, VA	
Chicago (1)	2659	03/08/2012	02:53	0:00:32	0.00	INB	Central Office		INBOUND	
Chicago (1)	2680	03/08/2012	02:53	0:00:32	0.00	INB	Central Office		INBOUND	
Chicago (1)	2652	03/08/2012	02:53	0:00:32	0.00	INB	Central Office		INBOUND	
Chicago (1)	2354	03/08/2012	03:16	0:00:12	0.03	INB	T1	617-568-1234	EASTBOSTON, MA	
Chicago (1)	2354	03/08/2012	03:18	0:00:38	0.03	INB	T1	617-568-1234	EASTBOSTON, MA	
Chicago (1)	2354	03/08/2012	03:20	0:00:36	0.03	INB	T1	617-568-1234	EASTBOSTON, MA	
Chicago (1)	2354	03/08/2012	03:28	0:00:42	0.03	INB	T1	513-579-1234	CINCINNATI, OH	
Chicago (1)	2354	03/08/2012	03:28	0:00:38	0.03	INB	T1	513-579-1234	CINCINNATI, OH	
Chicago (1)	2294	03/08/2012	03:39	0:00:18	0.03	INB	T1	918-582-9000	TULSA, OK	
Chicago (1)	2354	03/08/2012	04:45	0:00:40	0.03	INB	T1	614-463-1804	COLUMBUS, OH	
Chicago (1)	2354	03/08/2012	04:46	0:00:36	0.03	INB	T1	617-568-1234	EASTBOSTON, MA	
Chicago (1)	2354	03/08/2012	04:50	0:00:40	0.03	INB	T1	617-568-1234	EASTBOSTON, MA	
Chicago (1)	2354	03/08/2012	04:58	0:00:44	0.03	INB	T1	513-579-1234	CINCINNATI, OH	
Chicago (1)	2354	03/08/2012	04:59	0:00:44	0.03	INB	T1	513-579-1234	CINCINNATI, OH	
Chicago (1)	2354	03/08/2012	05:22	0:00:36	0.03	INB	T1	716-856-1234	BUFFALO, NY	
Chicago (1)	2440	03/08/2012	05:30	0:01:36	0.06	INB	T1	630-254-8053	ROSELLE, IL	
Chicago (1)	2354	03/08/2012	06:00	0:00:44	0.03	INB	T1	614-463-1797	COLUMBUS, OH	
Chicago (1)	2231	03/08/2012	06:06	0:00:08	0.36	LD	T1	1-815-235-1876	FREEPORT, IL	
Chicago (1)	2231	03/08/2012	06:06	0:01:28	0.72	LD	T1	1-815-395-1008	ROCKFORD, IL	
Chicago (1)	2654	03/08/2012	06:06	[illegible]	7.75	LD	T1	1-423-282-9071	JOHNSON CY, TN	

iv. Graph based or network data

Data that focuses on the relationship or adjacency of objects. The graph structures use nodes, edges, and properties to represent and store graphical data. Graph-based data is a natural way to represent social networks, and its structure. Allows you to calculate specific metrics such as the influence of a person and the **shortest path between two people.**

Examples: Friends in social media network

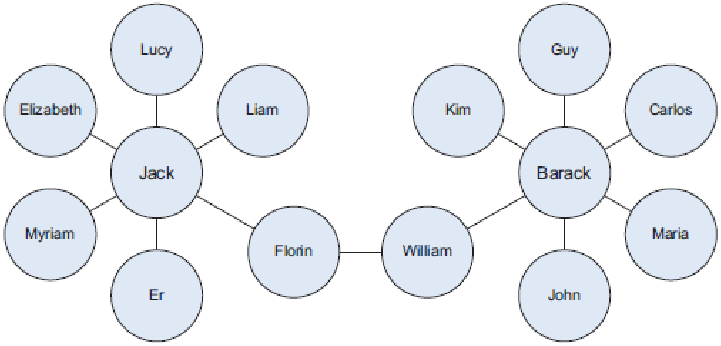


Figure 1.4 Friends in a social network are an example of graph-based data.

v. Audio, Image and Video data

It an be used together to extract insights through a process called audio video analytics.

Examples: Siri/ Ok google**,** Song Search**,** Object detection and analysis

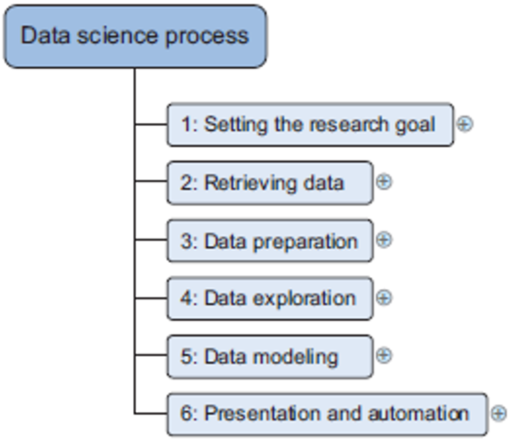
vi. Streaming data

The data flows into the system when an event happens instead of being loaded into a data stored in a batch is known as streaming data.

Examples: “What’s happening” on Twitter, Stock price, Live Sporting even like cricket/ football.

3. A) Write the steps in data science process and explain how to retrieve internal and external data.

A. Data science is a step by step process. It mainly includes 6 steps. They are:



Retrieving data

We can get data either from

- Internal data
- External data

Internal data- The data stored within the company, organization or any institution is known as internal data

i. **Start with data stored within the company**

Most companies have a program for maintaining key data, so much of the cleaning work may already be done”. This data can be stored in official data repositories such as databases, data marts, data warehouses, and data lakes.

Data base is used to store large amount of data in one place which helps organizations to quickly access, manage, modify, update, organize and retrieve their data. Databases are normally controlled using a database management system.

Example: college placement data.

Data mart is a subset of data warehouse. It contains small and selected part of data which is available in a large data warehouse.

Example: Placement data of **Vishnu society** is **Data Warehouse**.

Placement data of **Vishnu Institute of Technology** is **Data Base.**

Placement data of **AI&DS** department is **DataMart**.

A **data warehouse** is a repository of data from multiple data sources. Data can be in structured, unstructured or in semi-structured format.

Examples: placement data of colleges under Vishnu society. A **data lake** can store both structured, semi-structured, and unstructured data in its raw form or unprocessed form.

ii. **Data ownership**

Everyone does not have access to every data. There should be some privacy such that who should have access and who should not.

Example: Just imagine if all your phone numbers or social media accounts are public. It might raise some security issues. So, some data should be given as private.

Chinese walls: Every organization has a strict policy of who can access what part of data.

Example: A student will have access to view his attendance, faculty can have access to post the attendance whereas Ecap coordinator can have permission to edit attendance.

External data- Doing analysis only by collecting data within the organization is not sufficient.

Don't be afraid to shop around

- If data is not sufficient, then look for it outside the organization.
- Many government agencies open the data to the public for analysis purpose.

Example: Let us take placement data set. If we want to increase overall placements in Vishnu society, then considering all the 4 colleges in our society is not sufficient. In such cases we consider data from different universities also for better analysis.

Nowadays, government is also providing some information regarding employment and unemployment in data.gov.in

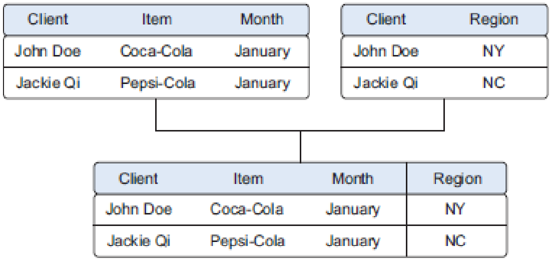
3. B) Explain how we can combine data from different sources.

A. We can combine data from different sources such as:

i. **Joining Tables:** Enriching an observation from one table with information from another table. Combines data from two tables, like purchases with regional demographics, to enrich observations.

Keys: Shared fields (e.g., client name, date) link tables and uniquely identify records.

Type: The number of output rows depends on the type of join, similar to Excel lookups.

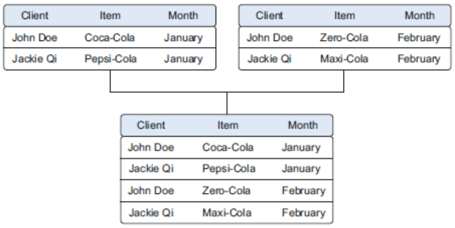


Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Region
John Doe	NY
Jackie Qi	NC

Client	Item	Month	Region
John Doe	Coca-Cola	January	NY
Jackie Qi	Pepsi-Cola	January	NC

ii. **Appending or Stacking:** adding the observation of one table to those of another table. We can append only when all the column names are same.

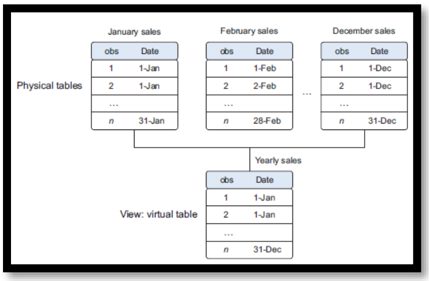


Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

iii. **Views:** While combining we may have a physical table or a virtual table (View).
Views Save Storage: Combine data virtually without duplication, saving storage space.
Dynamic Joins: Joins are recreated with each query, increasing processing demands.
Efficiency Trade-off: Great for large datasets but less efficient for frequent queries.



Physical tables
...
View: virtual table

January sales

obs	Date
1	1-Jan
2	1-Jan
...	
n	31-Jan

February sales

obs	Date
1	1-Feb
2	2-Feb
...	
n	28-Feb

December sales

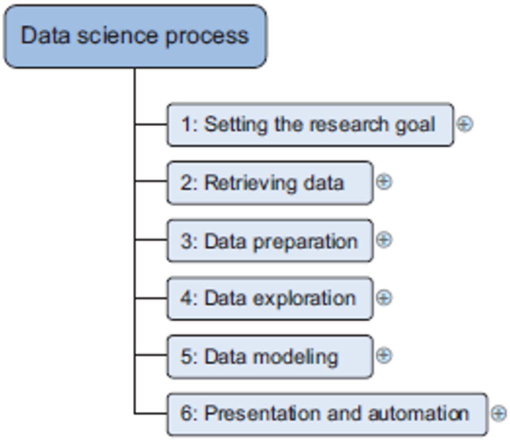
obs	Date
1	1-Dec
2	1-Dec
...	
n	31-Dec

Yearly sales

obs	Date
1	1-Jan
2	1-Jan
...	
n	31-Dec

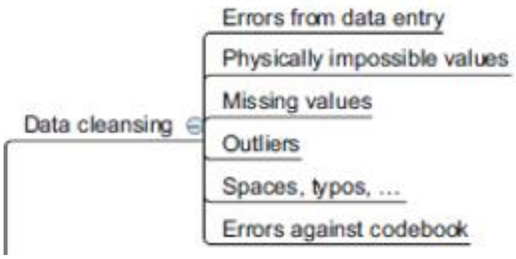
4. How many steps are there in data science process? Explain about data cleansing and how to handle missing values.

A. Data science is a step by step process. It mainly includes 6 steps. They are:



Data science process
1: Setting the research goal
2: Retrieving data
3: Data preparation
4: Data exploration
5: Data modeling
6: Presentation and automation

Data cleansing is one of the major important step in data science process to do better analysis and predictions.
This steps includes:



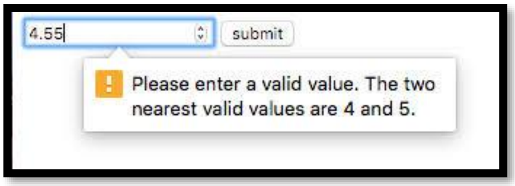
i. **Errors from data entry**

- **Interpretation error:** It occurs when data, information, or a situation is misunderstood.
  Examples: age> 300, percentage>100
- **Inconsistency error:** It occurs when data or information contains conflicting, contradictory, or mismatched values.
  Examples:"Male" in one place and **"M"** in other place, **"F"** and **"f"** are different.

ii. **Redundant white spaces and fixing capital letter mismatches**

A small white space may cause a lot of bugs in the code and can even change the performance of the model.

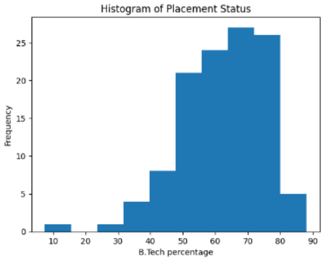
- Capital letter mismatches are common.
  Example: Vishnu Institute of Technology and Vishnu Institute of technology are not same
- Impossible values and sanity checks
  Here you check the value against physically or theoretically impossible values
  Example: check = 0 <= age <= 120



Please enter a valid value. The two nearest valid values are 4 and 5.

iii. **Outliers**

An outlier is an observation that seems to be distant from other observations.



The danger with outliers are:

- Outliers lead to incorrect interpretations or predictions.
- Important information might also miss.

iv. **Handling missing values**

- Omit the missing values.
- Fill the missing values with a string or 0.
- Fill the missing values with mean.
- Fill the missing value with mode.
- Fill the missing value with an estimated value.

But omitting the missing values is not a good practice because in some cases we might miss some important information.

Example:

Gender	**B.tech percentage**	**Communication skills**	**Coding skills**
M	75	AVERAGE	GOOD
F	82	GOOD	GOOD
F	87	EXCELLENT	EXCELLENT
M		GOOD	AVERAGE
M	79		
F	85	GOOD	GOOD

Fillna() is used to replace the missing values. Continuous values are replaced with mean and discrete values are replaced with mode.

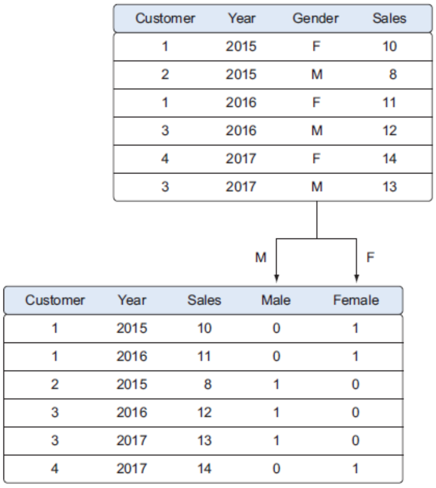
5. A) Explain about data transformation.

A. Data transformation is the process of changing data from one format to another. It's a key step in data management and is used to prepare data for analysis, reporting, and other purposes.

Certain models require their data to be in a certain shape.

- **Reducing the number of variables.**
  Sometimes you have too many variables and need to reduce the number because they don't add new information to the model. Having too many variables in your model makes the model difficult to handle, and certain techniques don't perform well when you overload them with too many input variables. This might result model in overfitting or underfitting.

- **Turning variables into dummies**
  Variables can be turned into dummy variables. Dummy variables can only take two values: true(1) or false(0). The presence of variable is represented with 1 and absence is represented with 0.



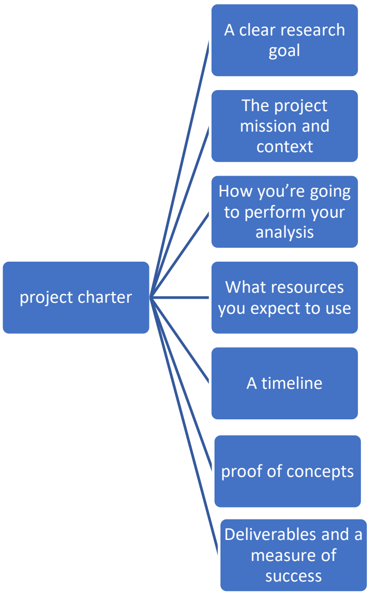
Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13

Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1

5. B) Create a project charter for placement data set.

A. Project charter for placement data set is as follows:

- Get a formal agreement on the deliverables.
- Your client can use this information to make an estimation of the project costs.



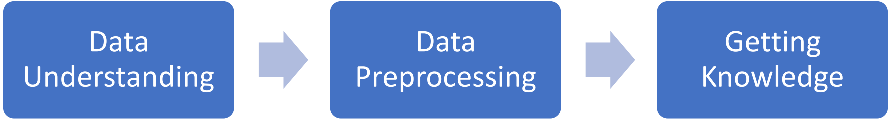
project charter
A clear research goal
The project mission and context
How you're going to perform your analysis
What resources you expect to use
A timeline
proof of concepts
Deliverables and a measure of success

- **A Clear Research Goal-**Improve placement rates.
- **Project Mission and Context-**To empower the college with actionable insights from placement data to enhance student employability and align with industry demands.

The university aims to:

  - Increase its placement percentages
  - Understand employer trends and preferences
  - Use data to showcase the success/failure of their programs for future admissions and partnerships.

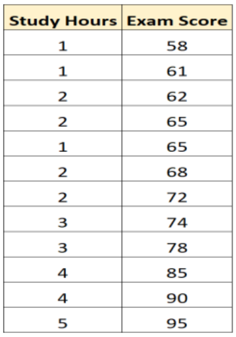
- **Analysis Plan**



- **Resources**
  - Python- Technical Resource
  - Placement information- Data Resource
  - One data analysist and 2 developers – Human Resource
- **Proof of Concept**
  - Develop a small prototype using no-code platforms
- **Deliverables and Measures of Success**
  - **Deliverables**
    - Predictive model for student placement probability.
  - **Measures of Success**
    - The university reports improved placement planning for the next academic year.
- **Timeline**

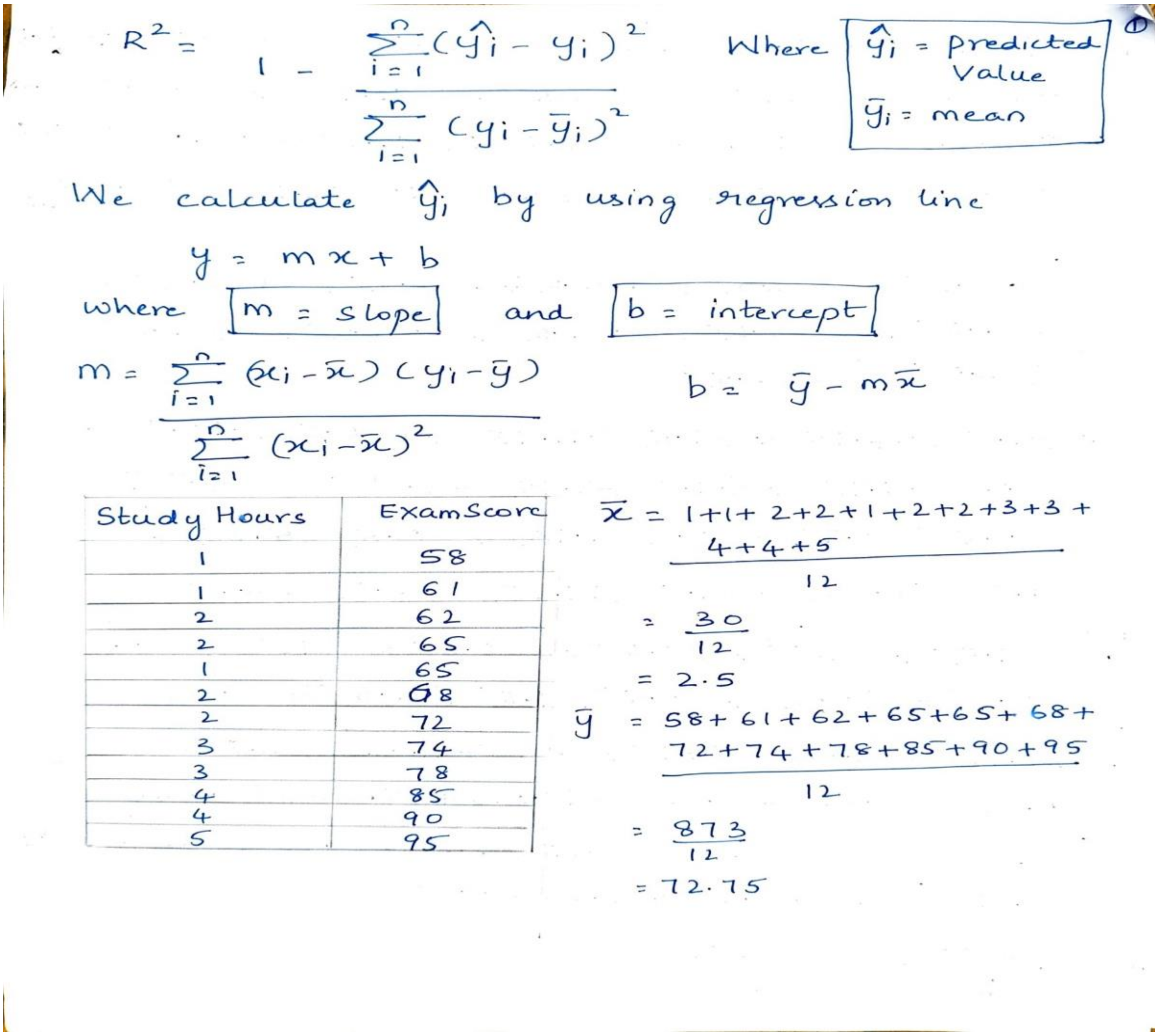
Phase	**Timeframe**
Data Understanding	Week 1
Data Cleaning	Week 2
Exploratory Analysis	Week 3
Advanced Analysis	Week 4
Dashboard Development	Week 5
Report & Recommendations	Week 6
Presentation & Handoff	Week 7

6. Calculate R² score for the below data set using linear regression and find whether the model is good or bad.



Study Hours	Exam Score
1	58
1	61
2	62
2	65
1	65
2	68
2	72
3	74
3	78
4	85
4	90
5	95

A. **R-squared** shows how much of the data is explained by your model. A regression model is said to be a good model if it has R² score more than 85%. R² score can be calculated using the below formula.



$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

Where $\boxed{\hat{y_i} = \text{predicted Value} \quad \bar{y_i} = \text{mean}}$

We calculate $\hat{y_i}$ by using regression line

$$y = mx + b$$

where $\boxed{m = \text{slope}}$ and $\boxed{b = \text{intercept}}$

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b = \bar{y} - m\bar{x}$$

Study Hours	ExamScore
1	58
1	61
2	62
2	65
1	65
2	68
2	72
3	74
3	78
4	85
4	90
5	95

$$\bar{x} = \frac{1+1+2+2+1+2+2+3+3+4+4+5}{12}$$

$$= \frac{30}{12}$$

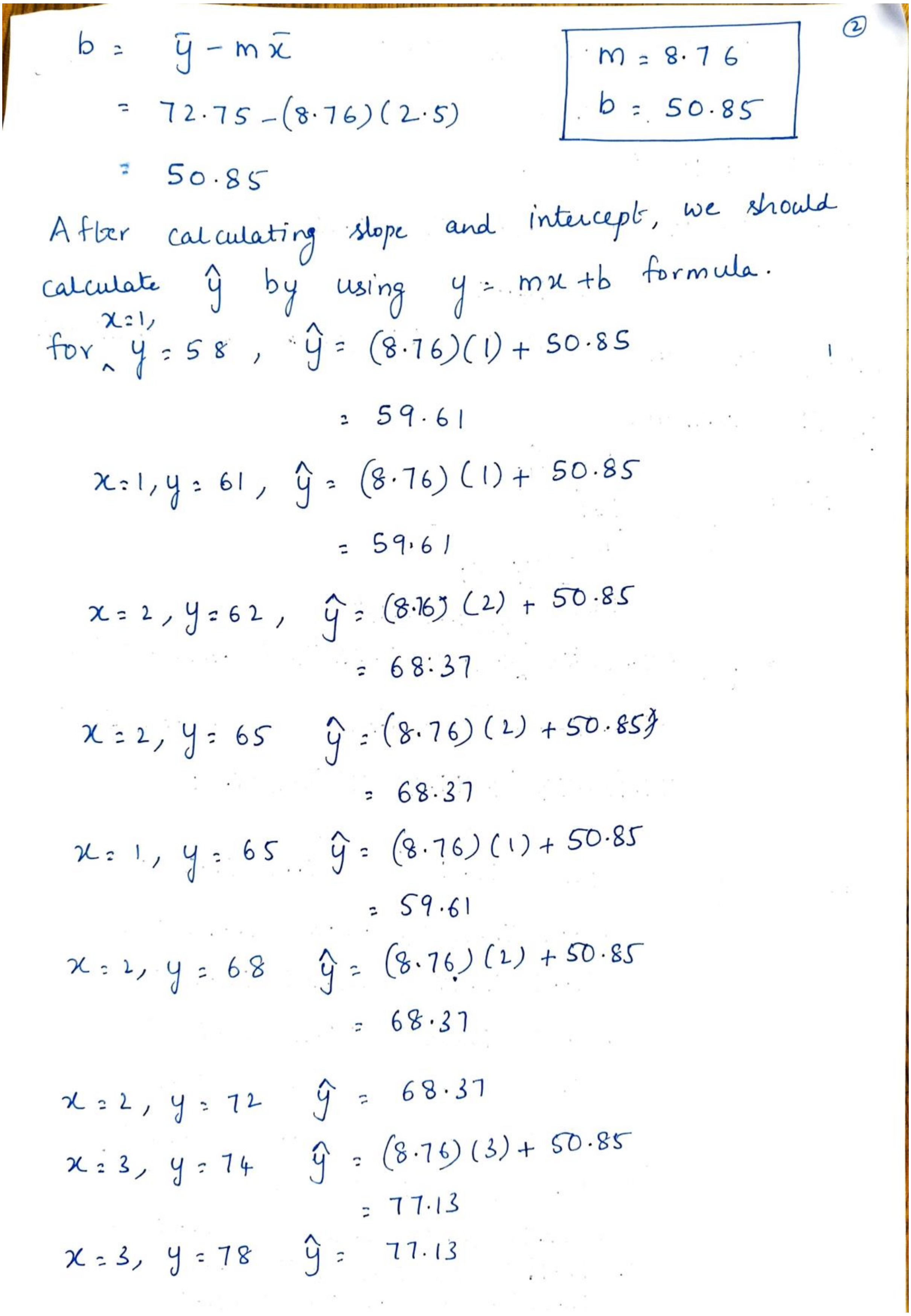
$$= 2.5$$

$$\bar{y} = \frac{58+61+62+65+65+68+72+74+78+85+90+95}{12}$$

$$= \frac{873}{12}$$

$$= 72.75$$

$$m = \frac{166.5}{19} = 8.76$$



2

$$b = \bar{y} - m\bar{x}$$

$$= 72.75 - (8.76)(2.5)$$

$$= 50.85$$

$$\boxed{\begin{aligned} m &= 8.76 \\ b &= 50.85 \end{aligned}}$$

After calculating slope and intercept, we should calculate $\hat{y}$ by using $y = mx + b$ formula.

for $x=1$, $y = 58$, $\hat{y} = (8.76)(1) + 50.85$

$= 59.61$

$x=1, y=61$, $\hat{y} = (8.76)(1) + 50.85$

$= 59.61$

$x=2, y=62$, $\hat{y} = (8.76)(2) + 50.85$

$= 68.37$

$x=2, y=65$ $\hat{y} = (8.76)(2) + 50.85$

$= 68.37$

$x=1, y=65$ $\hat{y} = (8.76)(1) + 50.85$

$= 59.61$

$x=2, y=68$ $\hat{y} = (8.76)(2) + 50.85$

$= 68.37$

$x=2, y=72$ $\hat{y} = 68.37$

$x=3, y=74$ $\hat{y} = (8.76)(3) + 50.85$

$= 77.13$

$x=3, y=78$ $\hat{y} = 77.13$

$x = 4,\ y = 85\ \hat{y} = (8.76)(4) + 50.85$

$= 85.89$

$x = 4,\ y = 90\quad \hat{y} = 85.89$

$x = 5,\ y = 95\quad \hat{y} = (8.76)(5) + 50.85$

$= 94.65$

Now, we can calculate R^2 value by substituting in the formula

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y}_1)^2}$$

$$R^2 = 1 - \frac{172.65 + 172.65 + 19.18 + 19.18 + 172.65 + 19.18 + 19.18 + 23.81 + 23.81 + 186.04 + 186.04 + 501.76}{217.56 + 751.81 + 678.56 + 498.81 + 498.81 + 239.06 + 0.56 + 3.06 + 27.56 + 150.06 + 297.56 + 495.06}$$

$$R^2 = 1 - \frac{1516.13}{3913.91}$$

$= 1 - 0.38$

$= 0.62 \Rightarrow 62\%$

As the percentage is $< 85\%$, the model is not a good model.

7. A) Explain how data science is used in governmental organization.

A. In governmental organizations, data science is used to analyze large datasets from various sources to inform policy decisions, improve public services, enhance efficiency, and increase transparency. Data science is used in many sectors of governmental organizations, including

- **Defense Intelligence Agency**: Data scientists use foreign intelligence to support military operations.
- **Government hospitals**: Data scientists analyze health data to identify major health issues, pandemics, and other health threats.
- **Environmental and weather forecast divisions**: Data scientists monitor climate changes, storms, and cyclones to help people stay safe.
- **Government organizations**: Data scientists use data analysis and machine learning to identify potential cyberattacks and vulnerabilities.

Data science can also help governments improve decision-making, target beneficiaries, and make smart policies.

7. **B)What is MSE? Calculate MSE for the below data set.**

X (Independent Variable)	**Y (Actual Values)**
1	55
2	58
3	65
4	78
5	85

A. Mean square error is a simple measure check for every prediction how far it was from the truth, square this error, and add up the error of every prediction. It is used for regression models.

The formula used to calculate MSE is

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

The Mean Square Error for the above data set is:

- $MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

where $\hat{y}_i$ is the predicted value

y_i is the actual value

To calculate $\hat{y}_i$, we use regression line formula

$$y = mx + b$$

where m is the slope and

b is the intercept

The formulae for m and b is

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$\bar{x} = 3$ $\quad \bar{y} = 68.2$

$$m = \frac{(1-3)(55-68.2) + (2-3)(58-68.2) + (3-3)(65-68.2) + (4-3)(78-68.2) + (5-3)(85-68.2)}{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}$$

$$= \frac{(-2)(-13.2) + (-1)(-10.2) + (0)(-3.2) + (1)(10.2) + (2)(16.8)}{4+1+0+1+4}$$

$$m = \frac{26.4 + 10.2 + 0 + 10.2 + 33.6}{10}$$

$m = 8.04$

$$\boxed{\begin{array}{l} m = 8.04 \\ b = 44.08 \end{array}}$$

$b = \bar{y} - m\bar{x}$

$= 68.2 - (8.04)(3)$

$= 68.2 - 24.12$

$b = 44.08$

By using slope and intercept, we can calculate $\hat{y}$

for $x=1, y=55$ $\quad \hat{y} = (8.04)(1) + 44.08$
$= 52.12$

$x=2, y=58$ $\quad \hat{y} = (8.04)(2) + 44.08$
$= 60.16$

$x=3, y=65$ $\quad \hat{y} = (8.04)(3) + 44.08$
$= 68.2$

$x=4, y=78$ $\quad \hat{y} = (8.04)(4) + 44.08$
$= 76.24$

$x=5, y=85$ $\quad \hat{y} = (8.04)(5) + 44.08$
$= 84.28$

After calculating $\hat{y}$, we can find MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$= \frac{(52.12-55)^2 + (60.16-58)^2 + (68.2-65)^2 + (76.24-78)^2 + (84.28-85)^2}{5}$$

$MSE = 5.358$

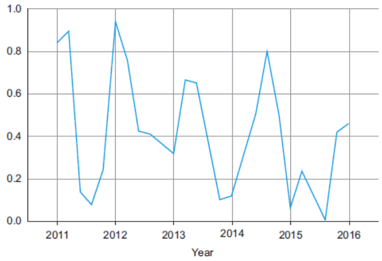
8. Describe about exploratory data analysis and different visualization techniques used.

A. Data exploration step in data science process is used to represent the data in the form of graphs. Information becomes much easier to grasp when shown in a picture, therefore we mainly use **graphical techniques to gain an understanding of your data** and the interactions between variables, identify outliers, identify important variables and to discover patterns. The visualization techniques you use in this phase range from simple line graphs or histograms.

These are some of the graphs:

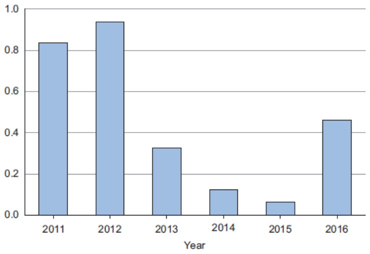
- **Line graph**

  A line graph, also known as a line chart, is a graph that shows how data changes over time.



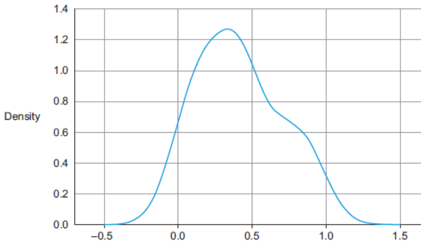
- **Bar graph**

  Bar graphs are the pictorial representation of data in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data.



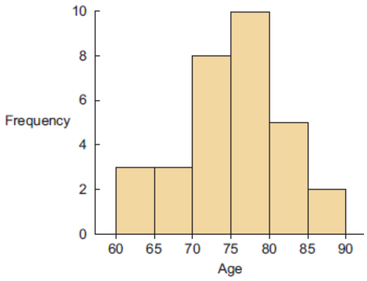
- **Density plot**

  A density plot is a graph that shows how a numerical variable is distributed over time.

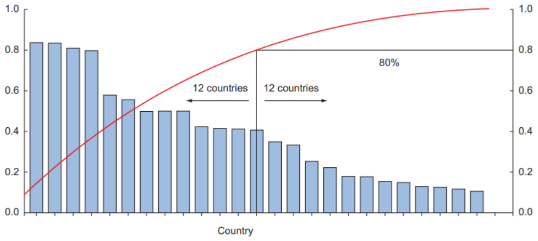


- **Histogram**

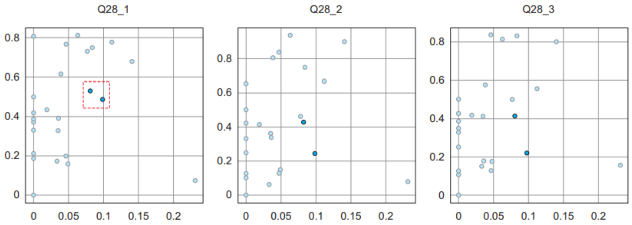
  A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars.



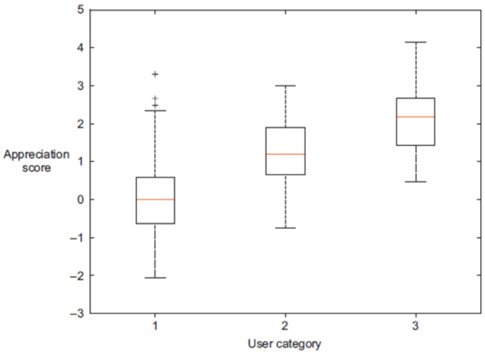
- **Combined graph or pareto diagram**
  A "combined graph" or "Pareto diagram" refers to a type of chart that visually combines both bar graphs and a line graph, where the bars represent individual values in descending order.



- **Link and brush**
  Combines different visualization methods to overcome the shortcomings of single techniques. Interactive changes made in one visualization are automatically reflected in the other visualizations. It is mostly used in google maps.



- **Boxplot**
  A box plot uses boxes and lines to depict the distributions of one or more groups of numeric data. Box limits indicate the range of the central 50% of the data, with a central line marking the median value.



9. Explain in detail about data modelling in data science process.

A. After data exploration, now we build models with the goal of making better predictions, classifying objects, or gaining an understanding of the system that you're modeling.

Most models consist of the following main steps:

- Selection of a modeling technique and variables to enter in the model.
- Execution of the model
- Diagnosis and model comparison

Model and variable selection

- By the time of EDA, we would be knowing which variables are most influential.

  Example: Is the phone number of a student really important for placement analysis?

- Is model easy to implement even in client environment?
- How long is it going to be relevant if untouched?
- Does the model need to be easy to explain? Model and variable selection

Model execution

Model Fit (R-squared & Adjusted R-squared):

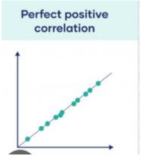
- **R-squared** shows how much of the data is explained by your model (e.g., 0.893 = 89.3%).
- **Adjusted R-squared** adjusts for unnecessary variables. It discourages making models too complex.
- Business: >0.85 is good.

Predictor Coefficients

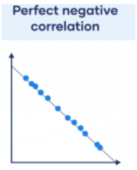
Correlation is a statistical measure that describes the strength and direction of the relationship between two variables.

The different types of correlations are:

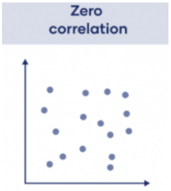
i. **Positive correlation:** Both variables move in the same direction. When one variable increases, the other variable also increases.
Example: communication skills and coding skills



ii. **Negative correlation:** variables move in opposite directions. When one variable increases, the other variable decreases.
Example: B.tech result in CGPA and SSC result in percentage



iii. **Zero correlation:** no relationship between the variables. There is no clear relationship between the two variables.
Example: roll number and marks



Model diagnostics and model comparison

- Holdout Sample for Model Evaluation: A holdout sample is a portion of data set aside during model building, used later to test how well the model performs on unseen data. This ensures the model works on data it hasn't encountered before.
- Error Measures to Compare Models: After using the holdout sample, error measures like Mean Squared Error (MSE) are calculated to evaluate and compare multiple models, helping to select the best one.

10. A) What is k-nearest neighbor technique. Explain it with an example.

A. **K-nearest neighbor algorithm**

Step 1: Selecting the optimal value of K

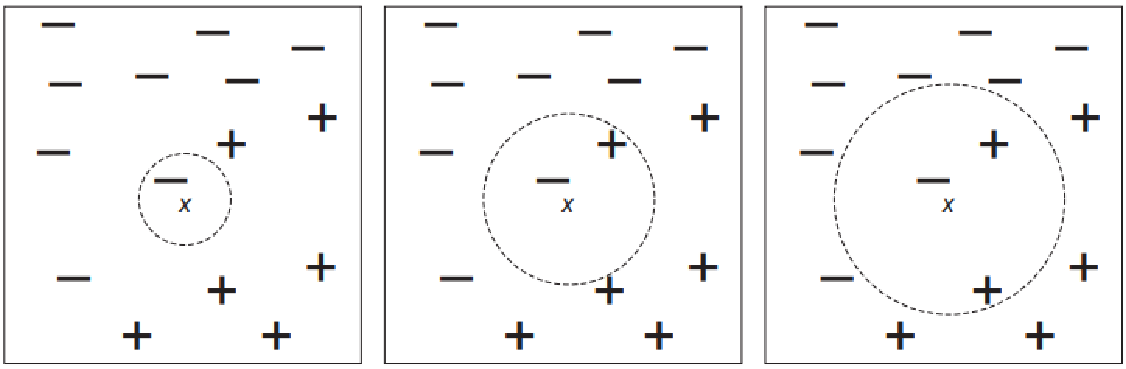
K represents the number of nearest neighbors that needs to be considered while making prediction.

Step 2: Calculating distance

To measure the similarity between target and training data points Euclidean distance is used. Distance is calculated between data points in the dataset and target point.

Step 3: Finding Nearest Neighbors

The k data points with the smallest distances to the target point are nearest neighbors.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

In the above example, when k=1, the value is predicted as **"-"**, when k=2, the value can neither be **"-" nor "+"** , when k=3, as there are maximum samples with “+”, the value is predicted as **“+”.** So, by this observation, we can say that it is better to choose **odd numbers** for **k-value.**

10. B) what is presentation and automation in data science process.

A. This is the final step of your hard work where you present your complete work to the stakeholders. Presenting your results to the stakeholders and industrializing your analysis process for repetitive reuse and integration with other tools.