Times asked: 6 times
5 times
4 times
3 times
2 times
1 time

# indicates 5-mark question

% indicates questions from same topic merged

View complete PYQ analysis & Previous year papers

Applied Data Science Question bank

1. Introduction to Data Science

1. Explain the data science process in detail.
2. Explain data science tasks with suitable examples.
3. Explain the significance of Volume, Dimensionality, and Complexity of data in Data Science techniques.
4. Differentiate between Data Science and Data Analytics.

2. Data Exploration

5. Calculate Bowley's coefficient of skewness for a given data set:

i.

Size	4	4.5	5	5.5	6	6.5	7	7.5	8
F	10	18	22	25	40	15	10	8	7

ii.

Salary	4-8	8-12	12-16	16-20	20-24
No of Candidates	4	10	15	8	3

6. Calculate Karl Pearson's coefficient of correlation for a given data set:

i.

X	15	18	20	28	34
Y	40	42	46	50	52

ii.

X	10	20	30	40	50	60	70	80	90	100
Y	2	4	8	5	10	15	14	20	22	50

7. Explain Data Exploration and its objectives and types. %
8. Explain Type-I and Type-II errors with suitable examples. %
9. Explain Hypothesis Testing. Describe the steps involved with an example, and the types of hypothesis testing. %

10. Explain ANOVA, its advantages, types, and how it differs from a t-test. %
11. Perform hypothesis testing using appropriate statistical tests (Z-test or t-test) for a given problem:
    i. In certain food experiment to compare two types of baby foods A and B, the following results of the increase in weight (lbs) observed in 8 children as follows:

Food A	49	53	51	52	47	50	52	53
Food B	52	55	52	53	50	54	54	53

    Examine the significance of the increase in weight of children due to food B. (Given t-value at alpha = 0.05 is 2.365)
    ii. A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trails in which she demonstrates a mean of 116 words with a standard deviation of 15 words? Use 5% level of significance. $Z_\alpha = 1.96$.

3. Methodology and Data Visualization

12. Explain Data Visualization, its importance, and its types (Univariate and Multivariate). Explain the purpose of Histogram, Quartile plot, Scatter plot, Bubble chart, and Density chart with suitable examples. %
13. Explain model validation techniques: Cross-validation, K-fold cross-validation, Leave-one-out cross-validation, and Bootstrapping.

4. Anomaly Detection

14. What are outliers? Explain the causes of outliers and different outlier detection methods. %
15. Explain SMOTE in detail.
16. Explain the DBSCAN algorithm to detect outliers along with its advantages and disadvantages.

5. Time Series Forecasting

17. Explain the Auto Regressive Integrated Moving Average (ARIMA) model. Describe its working, advantages, limitations, and applications. %
18. Explain the taxonomy of time series forecasting techniques. #
19. What is Time Series Decomposition? Explain its components and the classical decomposition technique.
20. Explain smoothing methods used in time series forecasting.
21. Explain how the time series approach is used to forecast the demand for a product.

6. Applications of Data Science

22. Explain how predictive modelling can be applied for house price prediction.
23. Write a note on Applications of Data Science. #

Asked once:

1. Introduction to Data Science

-

2. Data Exploration

1. Compare and contrast descriptive and inferential statistics. #
2. Write a note on measure of spread. #

3. Methodology and Data Visualization

3. Explain the roadmap for data exploration.

4. Anomaly Detection

4. Explain the Distance-based approach to outlier detection.
5. What is anomaly detection? Explain the process of anomaly detection. #
6. Can statistics be used to detect outliers if yes, Explain. #

5. Time Series Forecasting

7. Explain performance evaluation with respect to Time series forecasting. #
8. Explain Time series analysis using linear regression.

6. Applications of Data Science

9. Explain how predictive modelling can be applied for fraud detection.
10. Explain the steps to build a product recommendation model in detail.
11. What are Recommendation engines? Explain.

Module-wise Marks Weightage and Question Count

	1	2	3	4	5	6
2025 Aug	15 (2)	25 (3)	30 (4)	20 (2)	35 (4)	0
2025 May	20 (2)	55 (6)	15 (2)	5 (1)	15 (2)	15 (2)
2024 Dec	20 (2)	25 (3)	10 (1)	30 (4)	25 (3)	10 (1)
2024 May	15 (2)	55 (6)	20 (2)	15 (2)	15 (2)	10 (1)
2023 Dec	15 (2)	25 (3)	20 (2)	20 (2)	25 (3)	15 (2)
2023 May	10 (1)	35 (4)	10 (2)	25 (3)	35 (4)	10 (1)
Estimate	15 (2)	35+ (4+)	20 (2)	20 (2)	25-35 (3-4)	10 (1)
Total	95	220	105	115	150	60

Applied Data Science Answer bank

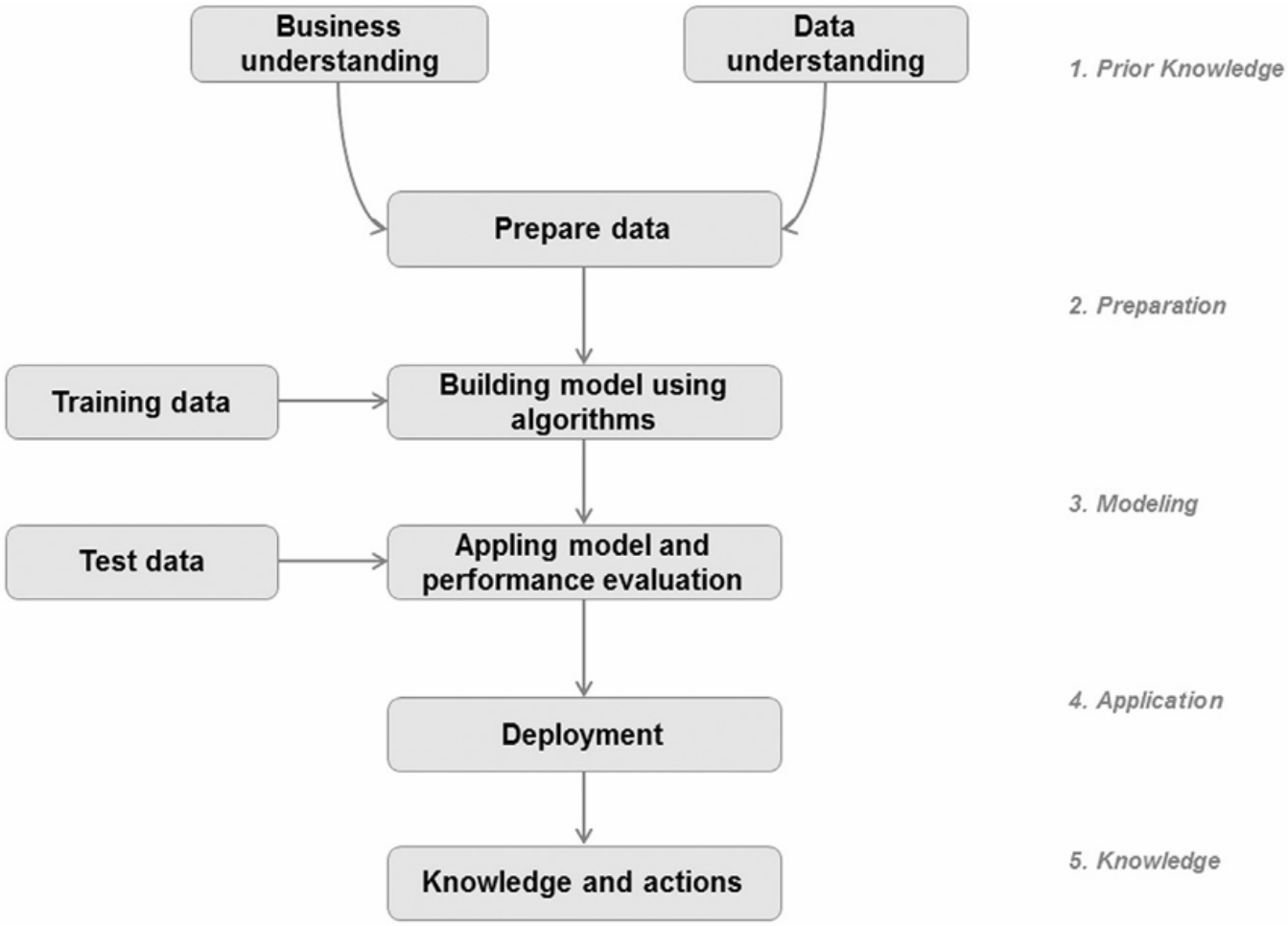
# indicates 5-mark question

% indicates questions from same topic merged

1. Introduction to Data Science

1. Explain the data science process in detail.

The Data Science Process is a systematic approach used to extract meaningful insights from data and support decision-making.
The process can be divided into five major stages: Prior Knowledge, Preparation, Modeling, Application, and Knowledge generation.



1. Prior Knowledge

This stage focuses on gaining a clear understanding of the problem and the data before any analysis is performed.

(a) Business Understanding

- Defines the objective of the problem and what the business wants to achieve, along with constraints and success criteria.

(b) Data Understanding

- Identifies and explores available data to understand its structure, type, and quality before further processing.

Example: Looking at students' marks and attendance to understand what data is available.

2. Data Preparation

- Raw data is cleaned and transformed into a usable format.
- Includes handling missing values, removing duplicates, and correcting inconsistencies.
- Relevant features are selected and data is structured for modeling.

Example: Cleaning a marks dataset by removing blanks, fixing wrong entries, and selecting subjects for prediction.

3. Modeling

- Appropriate algorithms are selected based on the problem (classification/regression)
- The model is trained using the training dataset.
- Multiple models may be built and compared.

Example: Using Linear Regression to predict student marks.

4. Application

- The model is evaluated using test data to check its performance.
- Metrics like accuracy or error are calculated.
- The model is deployed if performance is satisfactory.

Example: Checking how accurately the model predicts marks and using it for new students.

5. Knowledge Generation

- Model results are understood to find useful insights.
- These insights are used to take decisions or actions.
- Feedback can be used to improve the process further.

Example: Finding that students who study more hours score higher marks, and using this to guide study plans.

2. Explain data science tasks with suitable examples.

1. Classification

- Classification is used to assign data into predefined categories or classes.
- Output is discrete (labels like yes/no, spam/not spam).

Example: Classifying emails as spam or not spam

2. Regression

- Regression is used to predict continuous numerical values.
- Shows relationship between variables.

Example: Predicting student marks based on study hours.

3. Clustering

- Clustering is used to group similar data points without predefined labels.
- Helps in finding hidden patterns in data.

Example: Grouping customers based on buying behaviour.

4. Association Rule Mining

- Association Rule Mining is used to find relationships between variables in datasets.
- Often used in market basket analysis.

Example: Customers who buy bread also buy butter.

5. Anomaly Detection

- Anomaly Detection is used to identify unusual or abnormal data points.
- Useful in detecting fraud or errors.

Example: Detecting fraudulent bank transactions.

6. Recommendation Systems

- Recommendation Systems are used to suggest items based on user preferences or behaviour.
- Improves user experience through personalization.

Example: Recommending movies based on watch history.

7. Feature Selection

- Feature Selection is used to select important features (variables) for modeling.
- Reduces complexity and improves model performance.

Example: Choosing only study hours and attendance to predict marks, ignoring irrelevant data.

3. Explain the significance of Volume, Dimensionality, and Complexity of data in Data Science techniques.

1. Volume of Data

- Volume refers to the amount of data generated and stored.
- Large volumes of data require more storage and faster processing techniques.

Significance

- Traditional tools may fail to handle very large datasets.
- Requires use of big data technologies like distributed computing.
- More data can improve model accuracy if handled properly.

Example: Large amounts of data generated by social media platforms.

2. Dimensionality of Data

- Dimensionality refers to the number of features (attributes/variables) in a dataset
- High dimensional data means a large number of input variables

Significance

- Leads to the curse of dimensionality, making models complex and less efficient.
- Increases computation time and may reduce accuracy (overfitting).
- Requires techniques like feature selection or dimensionality reduction.

Example: A dataset with hundreds of features like age, income, location, etc.

3. Complexity of Data

- Complexity refers to the structure and variety of data.
- Data can be structured, semi-structured, or unstructured.

Significance

- Complex data is harder to process and analyze.
- Complex data requires advanced techniques like machine learning and deep learning.
- May include text, images, videos, or interconnected data.

Example: Analyzing social media data containing text, images, and videos.

4. Differentiate between Data Science and Data Analytics.

Parameter	Data Science	Data Analytics
Meaning	Study of extracting knowledge using advanced methods.	Study of analyzing data to find useful insights.
Scope	Broad field covering full data process.	Limited mainly to data analysis.
Main Goal	Predict future outcomes.	Understand past data.
Focus	Model building and prediction.	Reporting and interpretation.
Skill Set	Requires knowledge of ML, statistics, and programming.	Requires knowledge of data handling, basic statistics, and visualization.
Programming	Requires strong coding skills.	Requires basic tools like Excel/Power BI.
Data Type	Works with structured and unstructured data.	Mostly works with structured data.
Output	Predictive models.	Reports and dashboards.
Complexity	More complex and harder to implement.	Less complex and easier to implement.
Tools Used	Python, R, TensorFlow.	Excel, SQL, Power BI.
Example Use	Predicting student marks based on study data.	Analyzing student marks to find top performers

2. Data Exploration

5. Calculate Bowley's coefficient of skewness for a given data set:

i.

Size	4	4.5	5	5.5	6	6.5	7	7.5	8
F	10	18	22	25	40	15	10	8	7

Step 1: Create a Cumulative Frequency Table

Size (x)	Frequency (f)	Cumulative Frequency (cf)
4	10	10
4.5	18	28
5	22	50
5.5	25	75
6	40	115
6.5	15	130
7	10	140
7.5	8	148
8	7	155

The total frequency, N, is 155.

Step 2: Calculate the Quartiles

Now we'll find the first quartile (Q_1), the median/second quartile (Q_2), and the third quartile (Q_3).

1. First Quartile (Q_1)

- Position = $\frac{N+1}{4}$th item
- Position = $\frac{155+1}{4} = \frac{156}{4} = 39$th item
- Looking at the cf column, the 39th item falls into the cumulative frequency of 50, which corresponds to the size 5.
- $Q_1 = 5$

2. Second Quartile / Median (Q_2)

- Position = $\frac{N+1}{2}$th item
- Position = $\frac{155+1}{2} = \frac{156}{2} = 78$th item
- The 78th item falls into the cumulative frequency of 115, which corresponds to the size 6.
- $Q_2 = 6$

3. Third Quartile (Q_3)

- Position = $3 \times \frac{N+1}{4}\text{th item}$
- Position = $3 \times 39 = 117\text{th item}$
- The 117th item falls into the cumulative frequency of 130, which corresponds to the size 6.5.
- $Q_3 = 6.5$

Step 3: Apply Bowley's Formula

Bowley's coefficient of skewness (S_k) is calculated using the following formula:

$$S_k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Plug in our quartile values:

$$S_k = \frac{6.5 + 5 - 2(6)}{6.5 - 5}$$

$$S_k = \frac{11.5 - 12}{1.5}$$

$$S_k = \frac{-0.5}{1.5}$$

$$S_k = -0.333...$$

Final Answer: Bowley's coefficient of skewness for this series is approximately **-0.33**.

ii.

Salary	4-8	8-12	12-16	16-20	20-24
No of Candidates	4	10	15	8	3

Step 1: Create a Cumulative Frequency Table

Salary (Class)	No of Candidates (f)	Cumulative Frequency (cf)
4–8	4	4
8–12	10	14
12–16	15	29
16–20	8	37
20–24	3	40

The total frequency, N, is 40.

Step 2: Calculate the Quartiles

For continuous grouped data, the formula for a quartile is:

$$Q_i = L + \frac{\frac{iN}{4} - cf}{f} \times h$$

(Where L is the lower limit of the quartile class, cf is the cumulative frequency of the preceding class, f is the frequency of the quartile class, and h is the class width)

1. First Quartile (Q_1)

- Position = $\frac{N}{4}$th item
- Position = $\frac{40}{4} = 10$th item
- Looking at the cf column, the value just greater than 10 is 14. This means the Q_1 class is **8–12**.
- $L = 8, cf = 4, f = 10, h = 4$
- $$Q_1 = 8 + \frac{10 - 4}{10} \times 4$$
- $$Q_1 = 8 + \frac{6}{10} \times 4$$
- $$Q_1 = 8 + 2.4$$
- $Q_1 = 10.4$

2. Second Quartile / Median (Q_2)

- Position = $\frac{N}{2}$th item
- Position = $\frac{40}{2} = 20$th item
- Looking at the cf column, the value just greater than 20 is 29. This means the Q_2 class is **12–16**.

- $L = 12, cf = 14, f = 15, h = 4$
- $$Q_2 = 12 + \frac{20 - 14}{15} \times 4$$
- $$Q_2 = 12 + \frac{6}{15} \times 4$$
- $$Q_2 = 12 + \frac{24}{15}$$
- $$Q_2 = 12 + 1.6$$
- $Q_2 = 13.6$

3. Third Quartile (Q_3)

- Position = $\frac{3N}{4}\text{th item}$
- Position = $3 \times 10 = 30\text{th item}$
- Looking at the cf column, the value just greater than 30 is 37. This means the Q_3 class is **16–20**.
- $L = 16, cf = 29, f = 8, h = 4$
- $$Q_3 = 16 + \frac{30 - 29}{8} \times 4$$
- $$Q_3 = 16 + \frac{1}{8} \times 4$$
- $$Q_3 = 16 + 0.5$$
- $Q_3 = 16.5$

Step 3: Apply Bowley's Formula

Now apply the values to Bowley's coefficient of skewness (S_k) formula:

$$S_k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Plug in our calculated quartile values:

$$S_k = \frac{16.5 + 10.4 - 2(13.6)}{16.5 - 10.4}$$

$$S_k = \frac{26.9 - 27.2}{6.1}$$

$$S_k = \frac{-0.3}{6.1}$$

$$S_k = -0.04918...$$

Final Answer: Bowley's coefficient of skewness for this data is approximately **-0.049**.

6 Calculate Karl Pearson's coefficient of correlation for a given data set:

i.

X	15	18	20	28	34
Y	40	42	46	50	52

Step 1: Formula

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Step 2: Prepare the Table

X	Y	X^2	Y^2	XY
15	40	225	1600	600
18	42	324	1764	756
20	46	400	2116	920
28	50	784	2500	1400
34	52	1156	2704	1768
∑X = 115	**∑Y = 230**	**$\sum X^2$= 2889**	**$\sum Y^2$ = 10684**	**∑XY = 5444**

n = 5

Step 3: Substitute in Formula

Numerator:

$$n\sum XY - (\sum X)(\sum Y)$$

$$5(5444) - (115)(230)$$

$$27220 - 26450 = 770$$

Denominator:

$$\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}$$

$$\sqrt{[5(2889) - (115)^2][5(10684) - (230)^2]}$$

First bracket:

$$14445 - 13225 = 1220$$

Second bracket:

$$53420 - 52900 = 520$$

$$\sqrt{1220 \times 520}$$

$$\sqrt{634400} = 796.5$$

Step 4: Final Calculation

Now, divide the numerator by the denominator:

$$r = \frac{770}{796.5}$$

$$r = 0.97$$

Final Answer: Karl Pearson's coefficient of correlation is 0.97. Since r is close to +1, there is a very strong positive correlation between X and Y.

ii.

X	10	20	30	40	50	60	70	80	90	100
Y	2	4	8	5	10	15	14	20	22	50

Step 1: Formula

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Step 2: Prepare the Table

X	Y	X^2	Y^2	XY
10	2	100	4	20
20	4	400	16	80
30	8	900	64	240
40	5	1600	25	200
50	10	2500	100	500
60	15	3600	225	900
70	14	4900	196	980
80	20	6400	400	1600
90	22	8100	484	1980
100	50	10000	2500	5000
$\sum X = 550$	**$\sum Y = 150$**	**$\sum X^2 = 38500$**	**$\sum Y^2 = 4014$**	**$\sum XY = 11500$**

n = 10

Step 3: Substitute in Formula:

Numerator:

$$n\sum XY - (\sum X)(\sum Y)$$

$$10(11500) - (550)(150)$$

$$115000 - 82500 = 32500$$

Denominator:

$$\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}$$

$$\sqrt{[10(38500) - (550)^2][10(4014) - (150)^2]}$$

First bracket:

$$385000 - 302500 = 82500$$

Second bracket:

$$40140 - 22500 = 17640$$

$$\sqrt{82500 \times 17640}$$

$$\sqrt{1455300000} = 38148.4$$

Step 4: Final Calculation

$$r = \frac{32500}{38148.4}$$

$$r = 0.85$$

Final Answer: Karl Pearson's coefficient of correlation is approximately 0.85. This indicates a strong positive correlation between X and Y.

7. Explain Data Exploration and its objectives and types. %

Data Exploration (also called Exploratory Data Analysis - EDA) is the process of analyzing and summarizing a dataset using statistical methods and visualization techniques to understand its structure, patterns, and relationships before building models.
It helps in understanding the data better and preparing it for further analysis.

Objectives of Data Exploration

- To understand the structure and nature of data (size, variables, types)
- To identify missing values, errors, duplicates, and outliers
- To find patterns, trends, and relationships between variables
- To study the distribution of data (normal, skewed, etc.)
- To select important variables and prepare data for model building

Types of Data Exploration

1. Univariate Exploration

Univariate exploration studies one variable at a time to understand its characteristics.

- Helps in analyzing distribution, central tendency (mean, median), and spread (range, standard deviation)
- Common tools: Histogram, Bar Chart, Pie Chart, Mean, Median, Mode.

Example: Studying students' marks to find average and distribution.

2. Bivariate Exploration

Bivariate exploration studies two variables together to understand the relationship between them.

- Helps in identifying correlation or association between variables.
- Common tools: Scatter plot, Correlation coefficient.

Example: Studying the relationship between study hours and marks.

3. Multivariate Exploration

Multivariate exploration studies more than two variables together to understand complex relationships.

- Helps in analyzing how multiple variables affect each other
- Common tools: Correlation matrix, Heatmaps, ANOVA.

Example: Analyzing marks based on study hours, attendance, and sleep.

8. Explain Type-I and Type-II errors with suitable examples. %

In hypothesis testing, errors can occur when making decisions about the null hypothesis (H_0). These errors are classified as Type I error and Type II error.

1. Type I Error (α Error)

- Type I error occurs when we reject the null hypothesis even though it is true.
- It means we conclude that there is an effect or difference when actually there is none.
- Also called a False Positive.
- It is controlled by the level of significance (α), which is usually set before the test (e.g., 0.05).

Example:
Suppose we test whether a new teaching method improves marks.
If we conclude that the method is effective when actually it is not, this is a Type I error.

2. Type II Error (β Error)

- Type II error occurs when we fail to reject the null hypothesis even though it is false.
- It means we conclude that there is no effect when actually there is one.
- Also called a False Negative.
- It is related to the power of the test, where lower β means higher ability to detect real effects.

Example:
If we conclude that the new teaching method does not improve marks when actually it does, this is a Type II error.

The left curve represents H_0 and the right curve represents H_1. The shaded region α shows the probability of Type I error, while β shows Type II error. The area (1 − α) represents correct acceptance of H_0, and (1 − β) represents the power of the test.

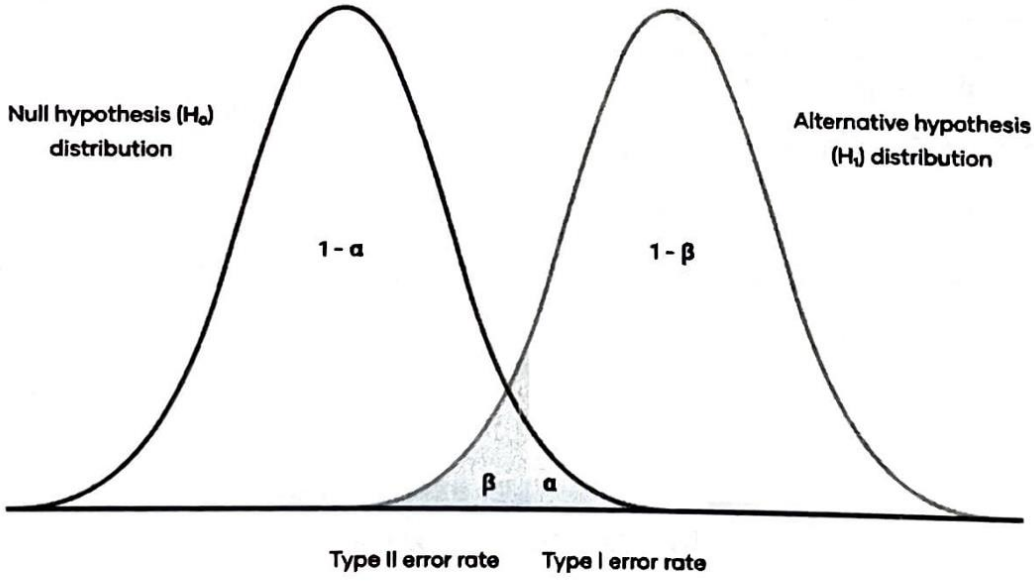


Figure 2.15: Probability of making a Type I & Type II Error.

9. Explain Hypothesis Testing. Describe the steps involved with an example, and the types of hypothesis testing. %

Hypothesis testing is a statistical method used to make a decision about a population based on sample data.
It helps us determine whether a claim or assumption is statistically significant or occurred by chance.

Steps in Hypothesis Testing

Step 1: State the Hypotheses

- **Null Hypothesis (H_0):** No effect or no difference.
- **Alternative Hypothesis (H_1):** There is an effect or difference.

Example:
H_0 : The new teaching method does not improve marks.
H_1 : The new teaching method improves marks.

Step 2: Choose the Significance Level (α)

- Select the probability of making a Type I error (usually 0.05).

Step 3: Select the Test Statistic

- Choose appropriate test (t-test, ANOVA, etc.) depending on data.

Step 4: Calculate the Test Statistic

- Compute the value using sample data.

Step 5: Make a Decision

- Compare calculated value with critical value (or p-value with α).
- If p-value < α → Reject H_0
- If p-value ≥ α → Do not reject H_0

Step 6: Draw Conclusion

- State the final result in simple words.

Example Conclusion:
If p-value < 0.05, we conclude that the new teaching method significantly improves marks.

Types of Hypothesis Testing

(a) One-Tailed Test

- Tests for effect in one direction only.
- Example: Checking if marks have increased.

(b) Two-Tailed Test

- Tests for effect in both directions.
- Example: Checking if marks have changed (increase or decrease).

(c) Parametric Tests

- Assume data follows a normal distribution.
- Example: Z-test, t-test.

(d) Non-Parametric Tests

- Do not assume any specific distribution.
- Example: Chi-square test.

10. Explain ANOVA, its advantages, types, and how it differs from a t-test. %

ANOVA (Analysis of Variance) is a statistical technique used to compare the means of three or more groups to determine whether there is a significant difference between them. It checks whether the variation between groups is real or due to random chance.
Example: Comparing the average marks of students from three different classes.

Types of ANOVA:

1. One-Way ANOVA

- Used when there is one independent variable (one factor).
- Compares means of multiple groups under one condition.
- Example: Comparing marks of students taught using different teaching methods.

2. Two-Way ANOVA

- Used when there are two independent variables (two factors).
- Studies the individual and combined effect of two factors.
- Example: Studying the effect of study hours and teaching method on marks.

Benefits of ANOVA:

- It allows comparison of more than two groups at the same time.
- It reduces the need for multiple t-tests, which increases error.
- It helps determine whether group differences are statistically significant.
- It saves time and provides more reliable results.
- It helps in studying the effect of one or more factors on a dependent variable.

Difference Between ANOVA and t-Test:

Parameter	**ANOVA**	**t-Test**
Number of Groups	Compares 3 or more groups	Compares only 2 groups
Factors	Can analyze one or two factors	Usually analyzes one factor
Error Rate	Controls Type I error when comparing many groups	Error increases if used multiple times
Purpose	Checks overall difference among group means	Checks difference between two means
Output	Gives F-value	Gives t-value

11. Perform hypothesis testing using appropriate statistical tests (Z-test or t-test) for a given problem:

i. In certain food experiment to compare two types of baby foods A and B, the following results of the increase in weight (lbs) observed in 8 children as follows:

Food A	49	53	51	52	47	50	52	53
Food B	52	55	52	53	50	54	54	53

Examine the significance of the increase in weight of children due to food B. (Given t-value at alpha = 0.05 is 2.365)

Since we are comparing two related samples (the same 8 children trying two different foods), we will use a Paired t-test.

Step 1: Prepare the Table

Let d be the difference in weight (Food B - Food A). Here, the number of pairs n = 8.

Child	**Food A**	**Food B**	**Difference (d = B – A)**	**d^2**
1	49	52	3	9
2	53	55	2	4
3	51	52	1	1
4	52	53	1	1
5	47	50	3	9
6	50	54	4	16
7	52	54	2	4
8	53	53	0	0
Totals			**$\sum d = 16$**	**$\sum d^2 = 44$**

Step 2: Hypothesis and Formulas

- **Null Hypothesis (H_0):** There is no significant increase in weight ($\mu_d = 0$).
- **Alternative Hypothesis (H_a):** There is a significant increase in weight due to Food B ($\mu_d > 0$).

We will need these formulas:

- **Mean of differences:** $\bar{d} = \frac{\Sigma d}{n}$
- **Standard Deviation:** $s = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}}$
- **Paired t-test:** $t = \frac{\bar{d}}{s/\sqrt{n}}$

Step 3: Substitute in Formula:

Mean of difference ($\bar{d}$):

$$\bar{d} = \frac{16}{8} = 2$$

Standard Deviation (s):

$$s = \sqrt{\frac{44 - \frac{(16)^2}{8}}{8-1}}$$

$$s = \sqrt{\frac{44 - \frac{256}{8}}{7}}$$

$$s = \sqrt{\frac{44 - 32}{7}}$$

$$s = \sqrt{\frac{12}{7}}$$

$$s = \sqrt{1.714} = 1.309$$

Calculate t-value (t):

$$t = \frac{2}{1.309/\sqrt{8}}$$

$$t = \frac{2}{1.309/2.828}$$

$$t = \frac{2}{0.463}$$

$$t = 4.32$$

Step 4: Compare with Given t-value

Calculated t = **4.32**
Given t (α = 0.05) = **2.365**

4.32 > 2.365

Because the calculated value (4.32) is greater than the given value (2.365), **we reject the null hypothesis**.
Therefore, there is a **highly significant increase in weight** of the children due to Food B.

ii. A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trails in which she demonstrates a mean of 116 words with a standard deviation of 15 words? Use 5% level of significance. $Z_\alpha = 1.96$.

To determine whether we can reject the stenographer's claim, we will perform a **One-Sample Z-Test** (since the sample size is large, $n = 100$).

1. State the Hypotheses

- **Null Hypothesis (H_0):** $\mu = 120$ (The claim is true; her mean speed is 120 wpm).
- **Alternative Hypothesis (H_1):** $\mu \neq 120$ (The claim is false; her mean speed is not 120 wpm).

2. Given Data

- Claimed Mean (μ): **120**
- Sample Mean ($\bar{x}$): **116**
- Standard Deviation (s): **15**
- Sample Size (n): **100**
- Significance Level (α): **5% (0.05)**
- Critical Z-value (Z_α): **1.96**

3. Calculate the Test Statistic (Z)

The formula for the Z-test is:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{116 - 120}{15/\sqrt{100}} = \frac{-4}{15/10} = \frac{-4}{1.5} = -2.67$$

The absolute value of our calculated Z is **2.67**.

4. Decision Rule

We compare the calculated $|Z|$ value with the critical Z value:

- If $|Z|_{cal} > Z_{crit}$, we **Reject** H_0.
- If $|Z|_{cal} \leq Z_{crit}$, we **Fail to Reject** H_0.

Comparison:

$$2.67 > 1.96$$

5. Conclusion

Since the calculated value (2.67) is greater than the critical value (1.96), it falls in the **rejection region**.

Final Answer: Yes, we **reject the stenographer's claim** at the 5% level of significance.

3. Methodology and Data Visualization

12 Explain Data Visualization, its importance, and its types (Univariate and Multivariate). Explain the purpose of Histogram, Quartile plot, Scatter plot, Bubble chart, and Density chart with suitable examples. %

Data Visualization refers to the representation of data in graphical or visual formats such as charts, graphs, and plots. Instead of analyzing raw numerical data, visualization helps in presenting information in a way that is easier to understand and interpret. It plays an important role in data analysis by making complex data more meaningful.

Importance of Data Visualization

- Makes large and complex data easier to understand and interpret.
- Helps in finding patterns and trends.
- Supports better decision-making.
- Helps in detecting outliers and anomalies.
- Improves presentation and communication of results.

Types of Data Visualization:

1. Univariate Visualization

- Involves visualization of a single variable.
- Used to understand distribution and spread of data.
- Helps in identifying central tendency and variability.

Examples: Histogram, Density plot, Box plot

2. Multivariate Visualization

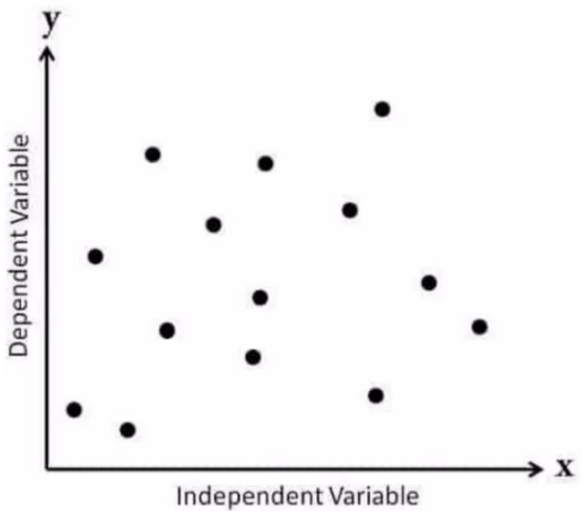
- Involves two or more variables.
- Used to understand relationships and interactions.
- Helps in identifying correlation and combined effects of variables

Examples: Scatter plot, Bubble chart

Data Visualization Techniques

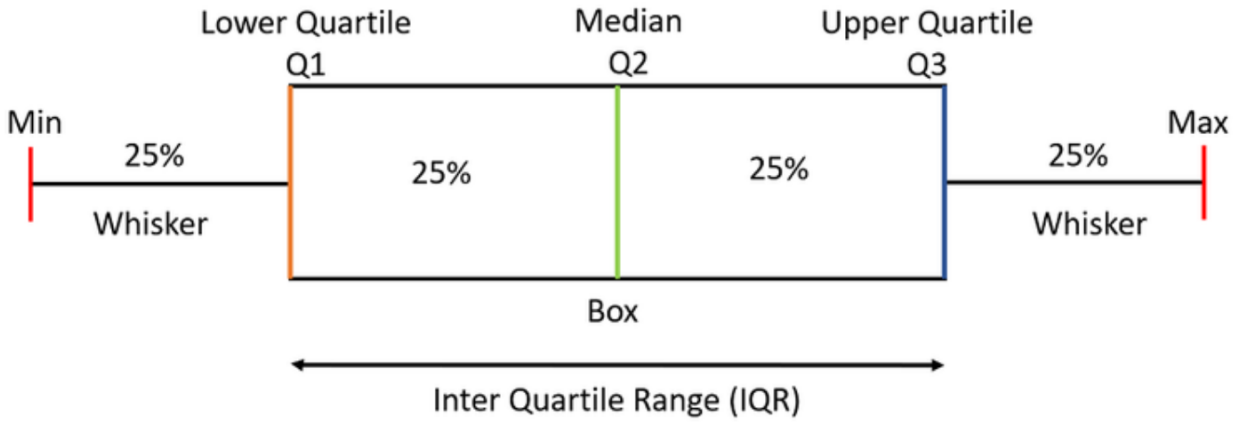
1. Scatter Plot: A scatter plot is used to represent the relationship between two variables. Each data point is plotted as a dot on a graph based on its values.

- Useful for identifying correlation (positive, negative, or no correlation)
- Helps in detecting outliers.



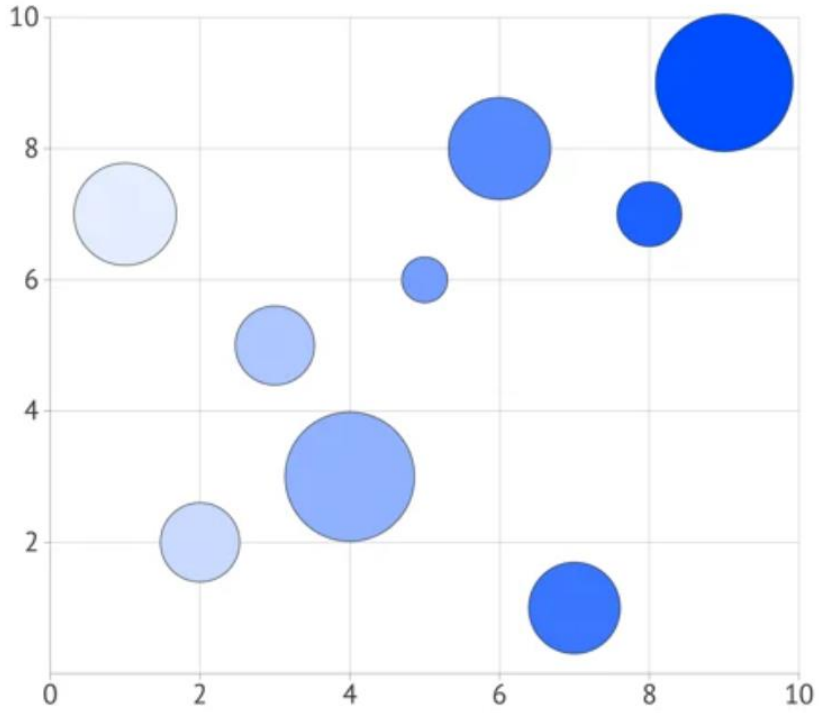
2. Quartile Plot (Box Plot): A quartile plot, also known as a box plot, represents the distribution of data using quartiles.

- Shows median, first quartile (Q1), third quartile (Q3), minimum and maximum values.
- Helps in identifying outliers clearly.



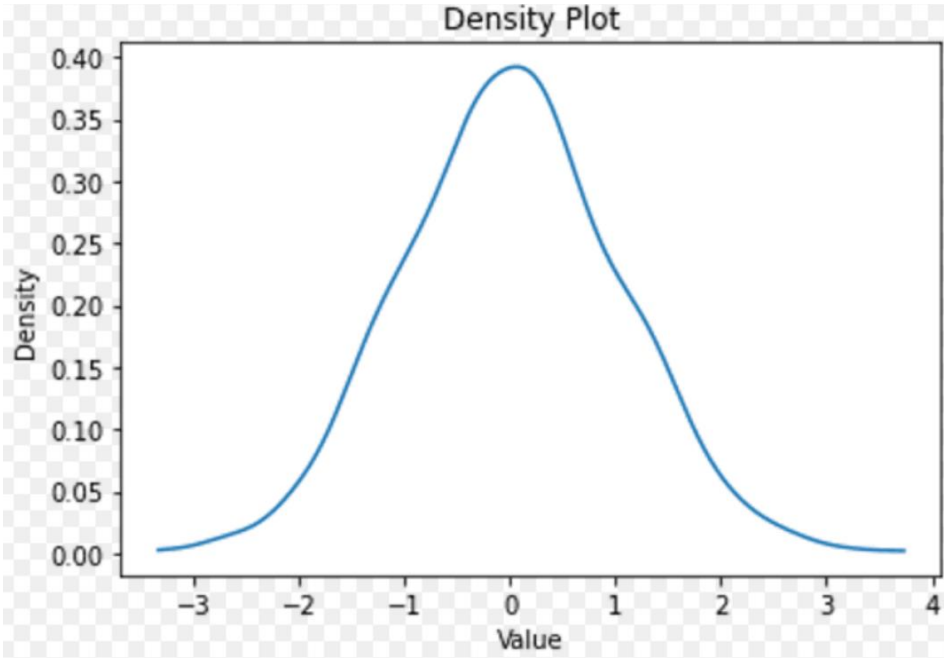
3. Bubble chart: A bubble chart is an extension of a scatter plot where a third variable is represented using the size of the bubbles.

- Allows visualization of three variables at once.
- Larger bubbles indicate higher values.



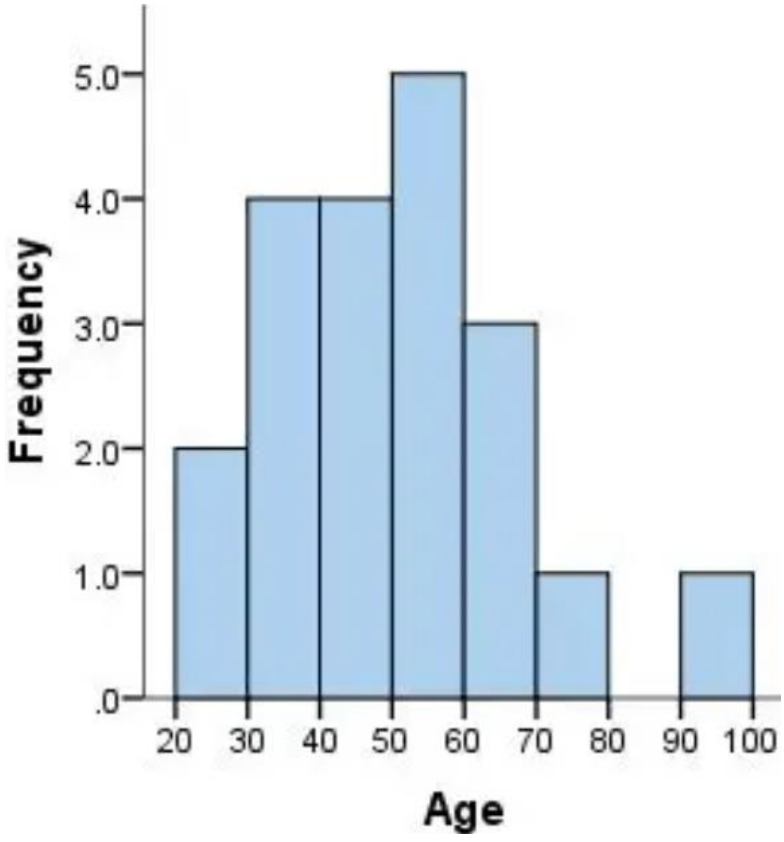
4. Density plot: A density plot represents the distribution of continuous data using a smooth curve.

- Smoother alternative to histograms.
- Helps in identifying peaks and spread of data.



5. Histogram: A histogram represents the distribution of continuous data using bars, where the height of each bar shows the frequency of values in each interval.

- Helps in understanding the shape of data (normal, skewed, etc.).
- Shows how frequently values occur within ranges.



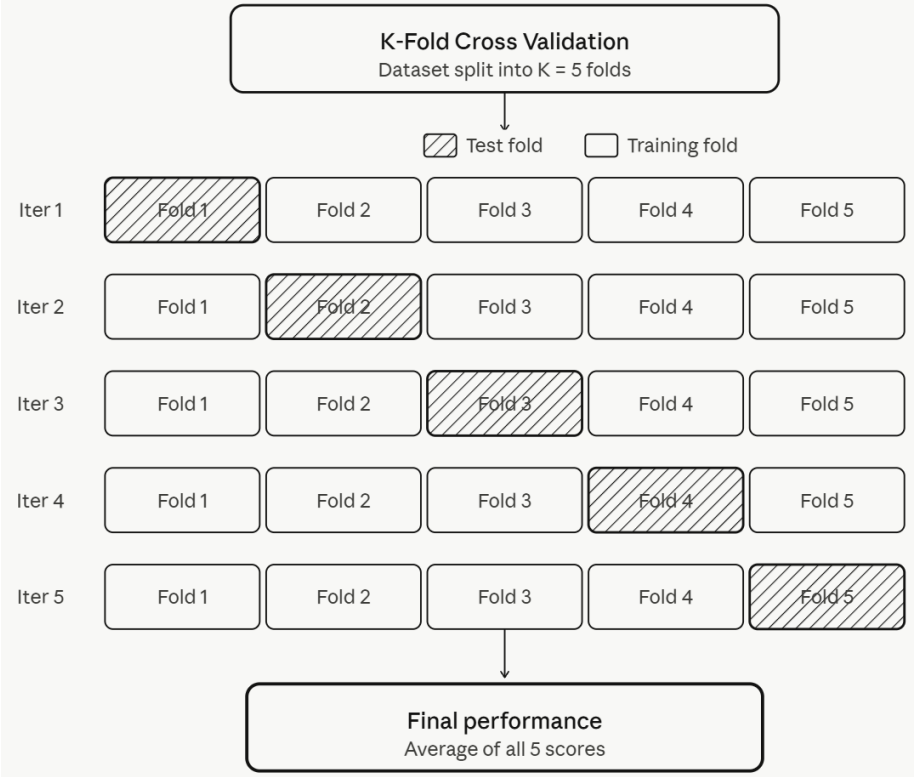
13. Explain model validation techniques: Cross-validation, K-fold cross-validation, Leave-one-out cross-validation, and Bootstrapping.

1. Cross Validation

- Cross validation is a resampling technique used to evaluate the performance of a machine learning model.
- Instead of splitting the data only once, the dataset is divided multiple times so that the model is tested on different subsets of data.
- This helps in reducing overfitting and gives a more accurate result.

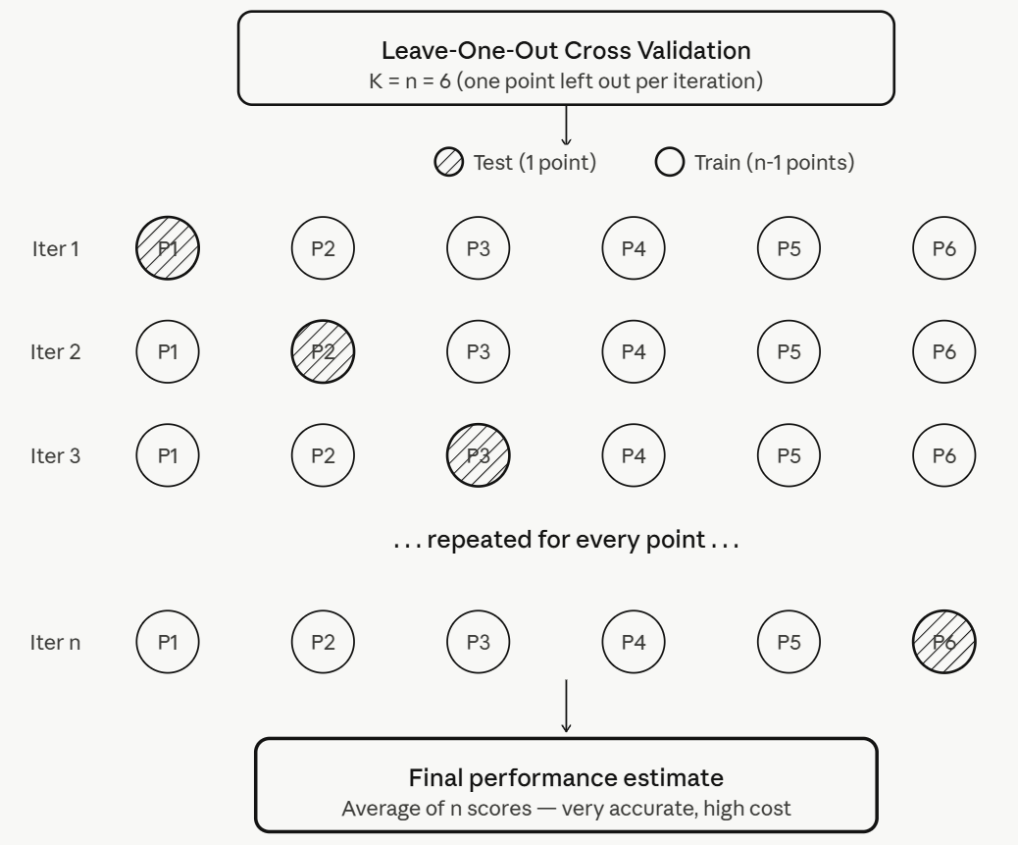
2. K-Fold Cross Validation

- In K-Fold Cross Validation, the dataset is divided into K equal parts (folds).
- The model is trained on K−1 folds and tested on the remaining fold.
- This process is repeated K times, with each fold used once as the test set.
- The final performance is the average of all K results.
- It provides a better evaluation compared to a simple train-test split.



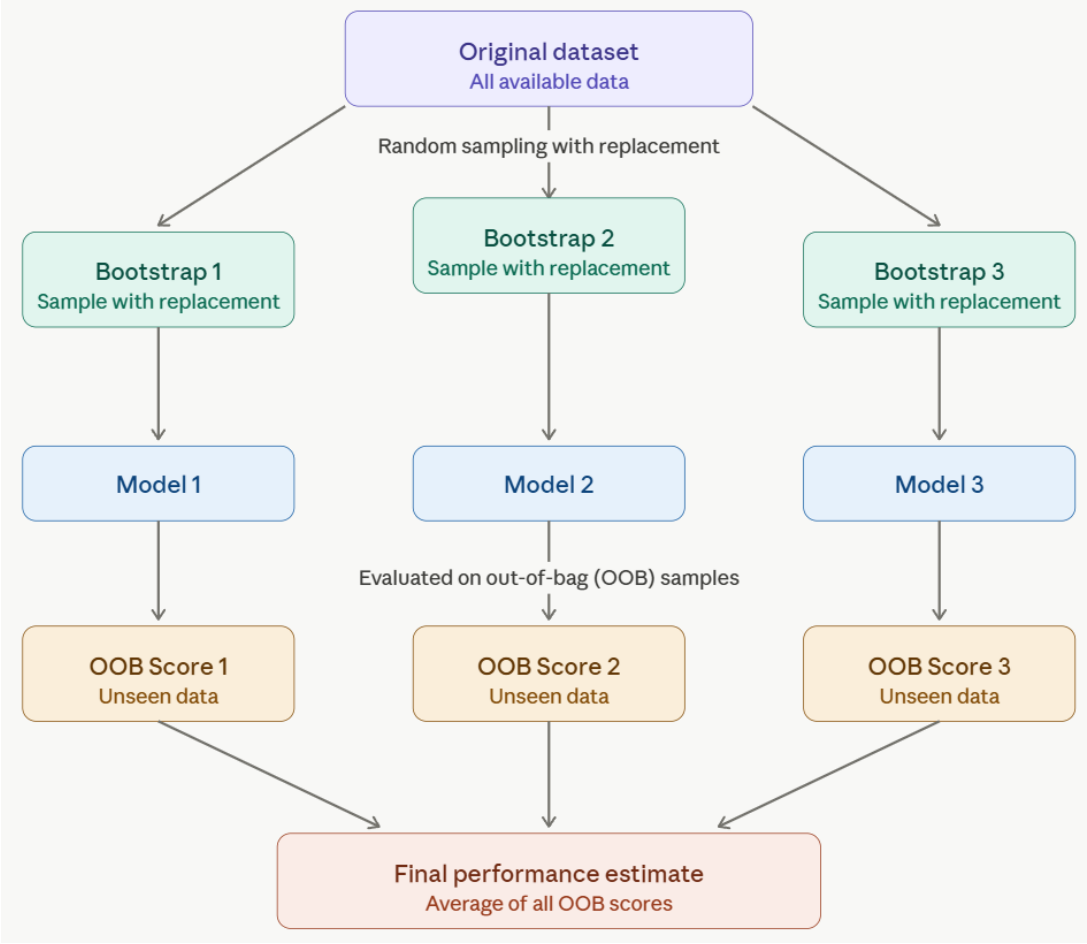
3. Leave-One-Out Cross Validation (LOOCV)

- LOOCV is a special case of K-Fold Cross Validation where K equals the number of data points (n).
- In this method, one data point is used as the test set and the remaining n−1 points are used for training.
- This process is repeated for every data point.
- It gives very accurate results but takes more time.



4. Bootstrapping

- Bootstrapping is a resampling technique where multiple datasets are created by randomly sampling the original dataset with replacement.
- Each sampled dataset is used to train the model.
- Performance is evaluated on the remaining data (out-of-bag samples).
- It is useful for estimating model performance.



4. Anomaly Detection

14. What are outliers? Explain the causes of outliers and different outlier detection methods. %

Outliers

- An outlier is a data point that significantly differs from other values in a dataset.
- It lies far away from the general pattern or distribution of the data.
- Outliers may indicate errors, rare events, or important variations.

Example: A student scoring 7 marks in a class where most score between 50-60.

Causes/Reasons of Outliers

1. Data Entry Errors: Incorrect values entered during data recording (e.g., typing 500 instead of 50).

2. Measurement Errors: Inaccurate readings caused by faulty instruments or improper measurement techniques.

3. Normal Variations in Data: Genuine extreme values that naturally occur in real-world data, representing rare but valid observations, such as unusually tall individuals.

4. Data from Different Distribution Classes: Values belonging to a different group or category (e.g., bot traffic vs human users), showing different behaviour.

5. Sampling Errors: Occur when the sample does not properly represent the population, leading to unusual or biased observations.

6. Data Processing Errors: Mistakes during data cleaning or transformation that result in abnormal values.

Outlier detection methods

Outlier detection methods are used to identify data points that significantly differ from the rest of the dataset. These methods help in detecting anomalies, errors, or rare events.

1. Statistical Methods

- These methods assume that the data follows a statistical distribution. Outliers are identified based on measures like mean and standard deviation.
- **Example:** Z-score method, where values far from the mean are considered outliers.

2. Distance-Based Methods

- In this method, the distance between data points is calculated. Points that are far away from most of the data are treated as outliers.
- **Example:** Uses distance measures like Euclidean distance, where points with large distances from others are identified as outliers

3. Density-Based Methods

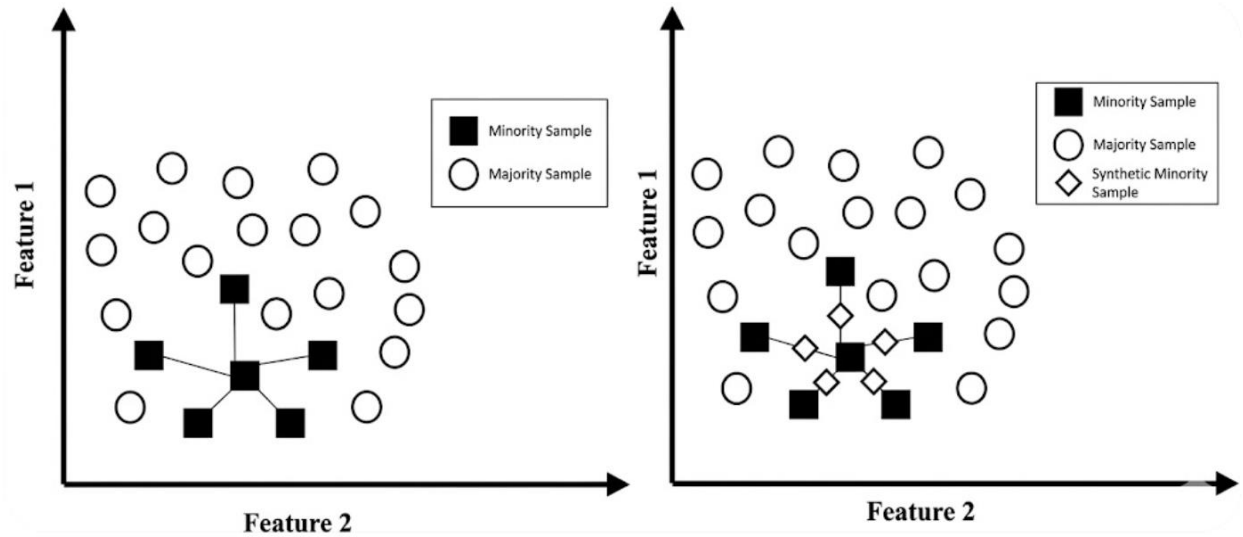
- These methods compare the density of a data point with its neighbouring points. Points in low-density regions are considered outliers, while points in high-density regions are normal.
- **Example:** DBSCAN, where points not belonging to any dense cluster are marked as noise (outliers)

4. Clustering-Based Methods

- Data is grouped into clusters, and points that do not belong to any cluster or belong to very small clusters are treated as outliers.
- **Example:** K-means, where points far from cluster centroids or in small clusters are considered outliers.

15. Explain SMOTE in detail.

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to handle imbalanced datasets, where one class (minority class) has much fewer samples than the majority class. Instead of duplicating minority samples, SMOTE generates new synthetic (artificial) samples to balance the dataset and improve model performance.



Working of SMOTE

1. A minority class data point is selected from the dataset.
2. The algorithm finds its k-nearest neighbours from the minority class.
3. A synthetic data point is created between the selected point and its neighbours.
4. The new synthetic sample is added to the dataset.
5. The process is repeated for other minority points until the dataset becomes balanced.

Advantages

- Reduces class imbalance in datasets.
- Improves performance for minority class.
- Avoids overfitting caused by simple duplication of samples.

Disadvantages

- May create noisy or overlapping samples.
- Increases computational cost for large datasets.

Applications

SMOTE is widely used in fraud detection, medical diagnosis, spam detection, and anomaly detection where minority class prediction is important.

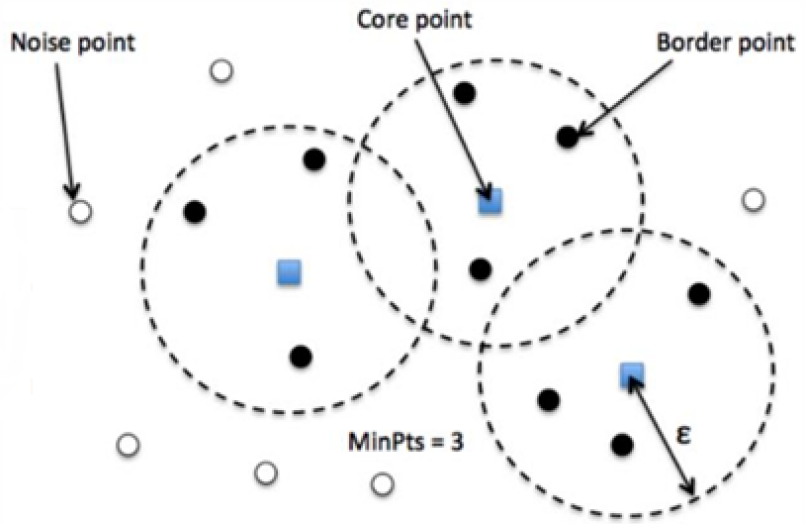
10. Explain the DBSCAN algorithm to detect outliers along with its advantages and disadvantages.

Density-based outlier detection identifies outliers based on the density of data points in a region.

- Data points in high-density regions (close to many points) are considered normal.
- Data points in low-density regions (far from others) are considered outliers.

A commonly used method based on this concept is DBSCAN,

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm that groups together data points that are closely packed (high-density regions) and marks points in low-density regions as noise or outliers.



Key Concepts

1. **ε (epsilon):** The neighbourhood radius, how far we look around a point.
2. **MinPts:** Minimum number of points required to form a dense region.
3. **Core Point:** A point having at least MinPts neighbours within distance ε.
4. **Border Point:** A point that is not a core point but falls within the ε-neighbourhood of a core point.
5. **Noise (Outlier):** A point that is neither a core point nor a border point.

Algorithm Steps

1. Pick an unvisited point.
2. If it has at least MinPts neighbours within ε → mark it as a core point and form a new cluster.
3. Expand the cluster by recursively including all points that are density-reachable from the core point.
4. If a point is not density-reachable from any core point → mark it as noise.
5. Repeat until all points are visited.

Example: In GPS data of trees, dense regions form clusters (forest areas), while isolated trees are marked as outliers.

Advantages of DBSCAN

- Can detect outliers directly as noise points.
- Does not require specifying number of clusters in advance.
- Can identify clusters of arbitrary shapes.

Disadvantages of DBSCAN

- Choice of ε and MinPts is critical and affects results.
- Not suitable for datasets with varying densities.
- Performance decreases for high-dimensional data.

5. Time Series Forecasting

17. Explain the Auto Regressive Integrated Moving Average (ARIMA) model. Describe its working, advantages, limitations, and applications. %

ARIMA (Auto Regressive Integrated Moving Average)is a time series forecasting model that combines Auto Regression (AR), Integration (I), and Moving Average (MA) to predict future values using past data and error terms.

Components of ARIMA

- **Auto Regression (AR):**
  Uses previous values of the time series to predict the current value.
- **Integration (I):**
  Applies differencing to make the data stationary by removing trend.
- **Moving Average (MA):**
  Uses past forecast errors to improve prediction accuracy.

Working of ARIMA Model

- The time series data is first analyzed to check stationarity, since ARIMA requires constant mean and variance.
- If the data is non-stationary, differencing is applied one or more times to remove trend and stabilize the series.
- The number of differencing steps is denoted by d.
- Next, the Auto Regression (AR) part is applied, where the current value is expressed as a function of its past values.
- The number of past values used is denoted by p.
- Then, the Moving Average (MA) part is applied, which models the current value based on past error terms.
- The number of error terms used is denoted by q.
- The model is represented as ARIMA(p, d, q).
- Appropriate values of (p, d, q) are selected using methods like ACF and PACF plots.
- The model is then trained on historical data and used to forecast future values.

Advantages of ARIMA

- Effective for short-term forecasting
- Works well with time series data having trends
- Flexible model combining multiple techniques
- Does not require large datasets

Limitations of ARIMA

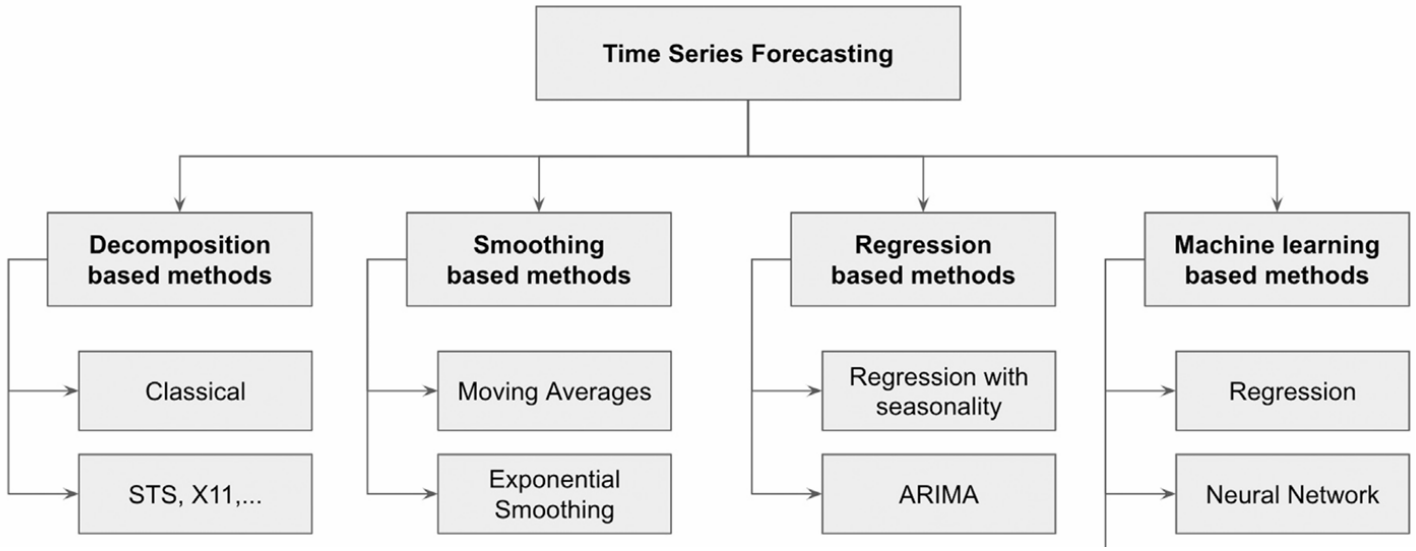
- Requires data to be stationary.
- Difficult to select correct (p, d, q) values.
- Not suitable for highly complex or nonlinear data.
- Cannot handle seasonality directly.

Applications of ARIMA

- Stock price and financial forecasting.
- Sales and demand forecasting.
- Weather prediction.
- Economic and business trend analysis.

18. Explain the taxonomy of time series forecasting techniques. #

Taxonomy of time series forecasting refers to the classification of forecasting techniques based on how future values are predicted from past data. Time series forecasting methods can be broadly divided into four categories.



1. Decomposition-Based Methods

- Time series is divided into trend, seasonality, and noise.
- Trend and seasonality are predictable, noise is random.
- Components are forecasted separately and then combined.

  Example: Classical decomposition

2. Smoothing-Based Methods

- Forecasts are made using weighted averages of past values.
- Helps in reducing noise and highlighting patterns.

  Example: Moving Average, Exponential Smoothing

3. Regression-Based Methods

- Models relationship between time and data values.
- Uses functions like linear or polynomial.
- Considers relation between past and present values.

  Example: Linear Regression, ARIMA

4. Machine Learning-Based Methods

- Uses ML models to learn patterns from data.
- Converts time series into input-output format.
- Can handle complex patterns.

  Example: Neural Networks

19. What is Time Series Decomposition? Explain its components and the classical decomposition technique.

Time Series Decomposition is the process of breaking down a time series into its individual components to understand underlying patterns and improve forecasting. It helps in identifying how different factors such as trend, seasonality, and randomness influence the data over time.

Components of Time Series Decomposition

1. **Trend (T)**
   The long-term movement of data over a long period of time. It shows the general direction in which the data is moving (increasing or decreasing).
2. **Seasonal Component (S)**
   Regular patterns that repeat over a fixed time period such as months, quarters, or seasons.
3. **Cyclical Component (C)**
   Long-term fluctuations occurring due to economic or business cycles. These cycles usually last for several years.
4. **Irregular or Random Component (I)**
   Unpredictable variations caused by random events such as natural disasters or unexpected changes.

Classical Decomposition Technique
The classical decomposition technique is used to separate a time series into its individual components (Trend, Seasonal, Cyclical, and Irregular) to analyze patterns and improve forecasting. Two common models are used:

1. **Additive Model**

Used when seasonal variations remain constant over time.

Time Series = Trend + Seasonal + Cyclical + Irregular

2. **Multiplicative Model**

Used when seasonal variations change proportionally with the trend.

Time Series = Trend × Seasonal × Cyclical × Irregular

Steps in Classical Decomposition

1. Estimate Trend (T)

- Use moving averages or smoothing techniques.

2. Estimate Seasonal Component (S)

- Remove trend and calculate seasonal indices.

3. Estimate Cyclical Component (C)

- Analyze long-term fluctuations after removing trend and seasonality.

4. Estimate Irregular Component (I)

- Remaining random variation after removing all other components.

20. Explain smoothing methods used in time series forecasting.

Smoothing methods are techniques used to reduce noise in time series data and highlight important patterns such as trend and seasonality, making forecasting more reliable.

Types of Smoothing Methods

1. Moving Average Method

- Forecast is calculated as the average of a fixed number of past observations.
- Helps remove short-term fluctuations and show the overall trend.
- Larger window gives smoother results but reacts slower to changes.

Example: A 3-period moving average uses the last 3 values to predict the next value

2. Weighted Moving Average

- Assigns different weights to past observations, giving more importance to recent data
- More accurate than simple moving average as it reflects recent changes better.
- Weights are chosen based on importance of observations.

Example: Recent sales values given higher weight than older values.

3. Simple Exponential Smoothing

- Applies a smoothing factor (α) to give more weight to recent observations.
- Suitable for data with no clear trend or seasonality.

4. Holt's Linear Method

- Extends exponential smoothing by including a trend component.
- Useful when data shows a consistent upward or downward trend.

5. Holt-Winters Method

- Extends Holt's method by including seasonality.
- Suitable for data with both trend and repeating seasonal patterns.

21. Explain how the time series approach is used to forecast the demand for a product.

The time-series approach is used to forecast future demand by analyzing past data collected over time. It assumes that past patterns will continue in the future.

Steps in Time-Series Forecasting with an example

1. Collection of Historical Data

- Past demand data is collected over time (daily, monthly, yearly).
- **Example:** Monthly sales of air conditioners (ACs) for the last 3 years.

2. Identification of Components

- The time series is analyzed into components:
  - **Trend (T):** Overall increase in AC sales over years.
  - **Seasonal (S):** Higher demand in summer months.
  - **Cyclical (C):** Changes due to economic conditions.
  - **Irregular (I):** Sudden changes due to unexpected events.

3. Model Selection

- A suitable forecasting model is selected based on data pattern.
- **Example:** Use Moving Average or ARIMA to model AC sales data.

4. Data Smoothing / Transformation

- Irregular fluctuations are reduced to make patterns clearer.
- **Example:** Apply moving average to smooth monthly AC sales.

5. Model Fitting

- The selected model is applied to historical data.
- **Example:** Fit ARIMA model using past AC sales data.

6. Forecasting Future Demand

- The model is used to predict future demand values.
- **Example:** Predict AC sales for upcoming summer season.

7. Evaluation of Forecast

- Predicted values are compared with actual values.
- **Example:** Compare forecasted vs actual AC sales using MAE/MSE.

6. Applications of Data Science

Q2. Explain how predictive modelling can be applied for house price prediction.

Predictive modelling uses historical data and machine learning techniques to predict future or unknown values. In house price prediction, it helps estimate the price of a house based on its features.

Steps in Predictive Modelling with an example:

1. Data Collection

- Collect historical data of houses with features such as location, area, number of rooms, and price.

Example dataset:

Area (sq. ft)	Bedrooms	Price (₹ Lakhs)
800	2	40
1000	2	50
1200	3	65
1500	3	80

2. Data Preprocessing

- Handle missing values, remove errors, and normalize data.
- Convert categorical data (e.g., location) into numerical form.

3. Feature Selection

- Select important features such as area, location, and number of bedrooms.
- Remove irrelevant or redundant features.

4. Model Selection

- Choose suitable algorithm like Linear Regression, Decision Tree, or Random Forest.
- Linear regression is commonly used for price prediction.

5. Model Training

- Train the model using historical data.
- The model learns relationships such as: “Larger area and more bedrooms → higher price”

6. Model Evaluation

- Test the model using unseen data.
- Use metrics like MAE, MSE, RMSE to measure accuracy.

7. Prediction

Using a simple regression model:

$$Price = 5 + 0.04(Area) + 5(Bedrooms)$$

For a new house:

- Area = 1100 sq. ft
- Bedrooms = 2

$$Price = 5 + 0.04(1100) + 5(2) = 59 \text{ Lakhs}$$

Predicted price = ₹59 Lakhs

Conclusion

Predictive modelling uses past data to learn relationships between features and price, enabling accurate and practical house price prediction.

Advantages

- Helps buyers and sellers make informed decisions.
- Automates price estimation.
- Improves accuracy compared to manual methods.

23. Write a note on Applications of Data Science. #

1. Finance

- Data science is used to analyze financial data for fraud detection, risk assessment, and investment decision-making.
  Example: Detecting fraudulent credit card transactions.

2. Healthcare

- Helps in predicting diseases, improving diagnosis, and analyzing patient data for better treatment decisions.
  Example: Predicting diabetes using patient health records.

3. Social Media

- Used to analyze user behaviour and public opinion from large volumes of user-generated data.
  Example: Sentiment analysis of tweets to understand public opinion.

4. Marketing

- Helps in customer segmentation, targeted advertising, and improving marketing strategies.
  Example: Showing personalized ads based on user browsing history.

5. Education

- Used to monitor student performance and improve learning outcomes.
  Example: Identifying students who need extra academic support.

6. E-commerce

- Used to recommend products and understand customer preferences.
  Example: Suggesting products based on past purchases.

7. Sports Analytics

- Helps in analyzing player performance and developing game strategies using data.
  Example: Evaluating player statistics to improve team performance.

Asked once:

2. Data Exploration

1. Compare and contrast descriptive and inferential statistics. #

Parameter	Descriptive Statistics	Inferential Statistics
Meaning	Summarizes and describes data.	Draws conclusions from data.
Purpose	To present data in a simple form.	To make predictions or decisions.
Data Used	Uses complete data (dataset).	Uses sample data.
Techniques	Mean, median, charts, graphs.	Hypothesis testing, regression.
Output	Tables, graphs or summaries.	Conclusions & predictions.
Complexity	Simple and easy to understand.	More complex and mathematical.
Accuracy	Exact for given data.	Based on probability (may involve error).
Example	Finding average marks of a class.	Predicting marks of all students from a sample.

2. Write a note on measure of spread. #

A measure of spread tells us how much the data values are spread out or vary from each other or from the central value. It helps us understand whether the data is consistent or varies a lot.

Common Measures of Spread

1. Range

- Difference between the maximum and minimum values, showing the total spread of data.

$$\text{Range} = \text{Max} - \text{Min}$$

2. Interquartile Range (IQR)

- Difference between Q_3 and Q_1, representing the spread of the middle 50% of data.

$$\text{IQR} = Q_3 - Q_1$$

3. Mean Absolute Deviation

- Average of absolute differences from the mean, indicating overall variability in data.

4. Variance

- Average of squared deviations from the mean, measuring how much values differ from the average.

5. Standard Deviation

- Square root of variance, showing the typical distance of values from the mean.
- Most commonly used measure of spread.

3. Methodology and Data Visualization

3. Explain the roadmap for data exploration.

The roadmap for data exploration is a structured approach to understand and analyze a dataset using statistical measures and visualization techniques before modeling.

1. Organizing the Dataset

- Arrange data in rows (observations) and columns (attributes) for proper analysis
- Identify the target variable (class label) if present

2. Determining Central Tendency

- Calculate mean, median, and mode for each attribute
- Differences may indicate skewness or presence of outliers

3. Analyzing Data Dispersion

- Use range and standard deviation to understand variability
- Helps in knowing how spread out the data is

4. Studying Data Distribution

- Use histograms to observe distribution of values
- Helps identify normal or skewed data

5. Pivoting and Slicing Data

- Rearrange data to analyze different attribute values
- Useful for summarizing and comparing data

6. Identifying Outliers

- Detect unusual values using quartiles or plots
- Outliers can affect mean and variance

7. Examining Relationships Between Attributes

- Measure correlation between variables
- Helps identify dependent attributes.

8. Visualizing Relationships

- Use scatter plots to study relationships.
- Helps identify trends and patterns.

9. Exploring High-Dimensional Data

- Use advanced charts like parallel coordinates.
- Helps analyze multiple attributes together.

4. Anomaly Detection

4. Explain the Distance-based approach to outlier detection.

The distance-based approach identifies outliers based on the distance between data points. A data point is considered an outlier if it lies far away from most other points in the dataset.

Concept

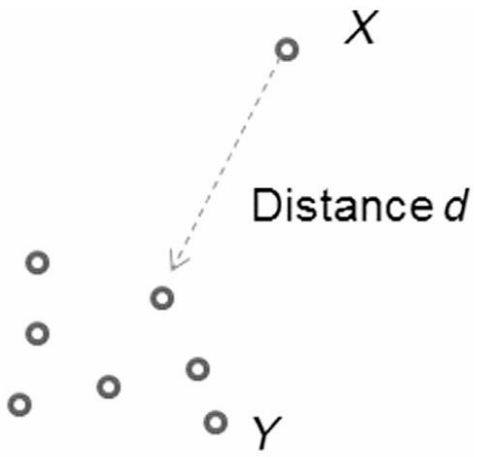
- Normal data points are located in dense regions (close to each other).
- Outliers are points that are far from the majority of data.
- Distance is measured using metrics like Euclidean distance.

Working of Distance-Based Approach

1. For each data point, calculate the distance to other points.
2. Define a threshold distance (d) or number of neighbours (k).
3. Count how many points lie within distance d of a given point.
4. If a point has very few neighbours within this distance, it is marked as an outlier.

Example:

- In the diagram, most points form a cluster (dense region).
- Point X lies far away from other points.
- The distance d between X and nearest points is large.



Hence, X is identified as an outlier, while points like Y are normal.

Advantages

- Simple and easy to understand.
- Does not assume any data distribution.
- Effective for small and low-dimensional datasets.

Disadvantages

- Computationally expensive for large datasets.
- Not suitable for high-dimensional data.
- Sensitive to choice of distance threshold.
- Cannot handle varying density well.

5. What is anomaly detection? Explain the process of anomaly detection.

Anomaly Detection is the process of identifying data points that significantly differ from normal patterns in a dataset. These unusual points are called anomalies or outliers and may indicate errors, fraud, or rare events.

Process of Anomaly Detection

1. Data Collection

- Gather relevant data from sources such as databases, sensors, or logs.
- Ensure data is suitable for analysis.

2. Data Preprocessing

- Clean the data by handling missing values and noise.
- Normalize or transform data if required.

3. Feature Selection

- Select important attributes that help in identifying anomalies.
- Remove irrelevant or redundant features.

4. Model Selection

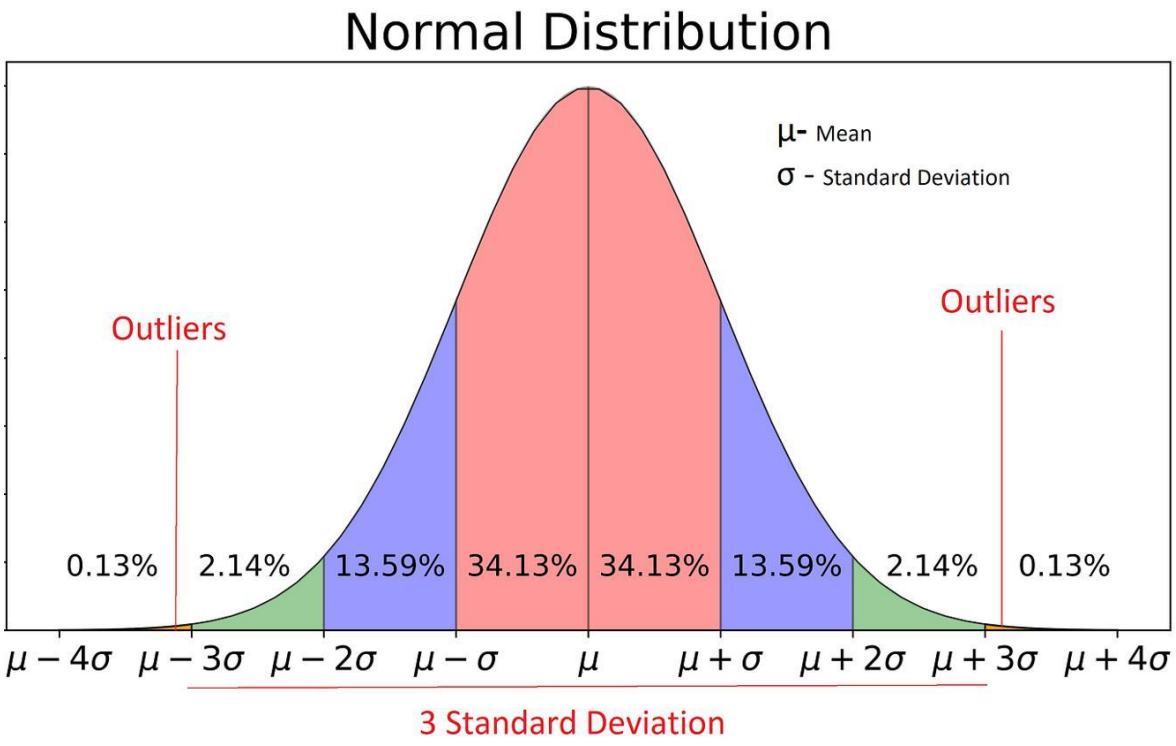
- Choose an appropriate method (statistical, distance-based, or machine learning).
- Depends on type and nature of data.

5. Detection of Anomalies

- Apply the chosen method to identify unusual data points.
- Points that deviate significantly are marked as anomalies.

6. Can statistics be used to detect outliers if yes, Explain. #

Yes, statistical methods can be used to detect outliers by modeling the data using a probability distribution and identifying values that lie far from the normal range.



- Many real-world datasets follow a normal (Gaussian) distribution.
- The distribution is defined using mean and standard deviation.
- A normal distribution curve is constructed using these parameters.
- Data points that lie at the extreme ends (tails) of the distribution do not fit the model.
- Such points are considered outliers, as they are far from the majority of data.
- Typically, values beyond ±3 standard deviations from the mean are treated as outliers.

5. Time Series Forecasting

7. Explain performance evaluation with respect to Time series forecasting.

Performance evaluation in time series forecasting measures how accurately predicted values match actual values. It helps in selecting the most suitable forecasting model.

Let error be e_t, actual value be y_t and predicted(forecasted) value be $\hat{y}_t$.

$$e_t = y_t - \hat{y}_t$$

1. Mean Absolute Error (MAE)

- Measures average magnitude of errors without considering direction.
- Easy to interpret and less sensitive to large errors.

$$MAE = \frac{1}{n}\sum | e_t |$$

2. Mean Squared Error (MSE)

- Measures average squared errors, giving more weight to large errors.
- Gives more importance to larger errors.

$$MSE = \frac{1}{n}\sum e_t^2$$

3. Root Mean Squared Error (RMSE)

- Square root of MSE, expressed in same units as data.
- Easier to interpret than MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

4. Mean Absolute Percentage Error (MAPE)

- Measures error as a percentage of actual values.
- Useful for comparing performance across datasets.

$$MAPE = \frac{100}{n}\sum \left| \frac{e_t}{y_t} \right|$$

8. Explain Time series analysis using linear regression.

Time series analysis using linear regression models the relationship between time and the observed values to identify trends and forecast future values.

Linear Regression Model

- The basic form of the model is:

$$y_t = a + bt$$

Where:

- y_t= value at time t
- a= intercept (starting value)
- b= slope (rate of change over time)

Working of the Method

- Time is treated as the independent variable, and data values as the dependent variable.
- A regression line is fitted to the historical data.
- The slope b shows whether the data is increasing or decreasing.
- The fitted line is used to predict future values.

Example

Suppose monthly sales data is:

Month (t)	Sales (y)
1	100
2	120
3	140
4	160

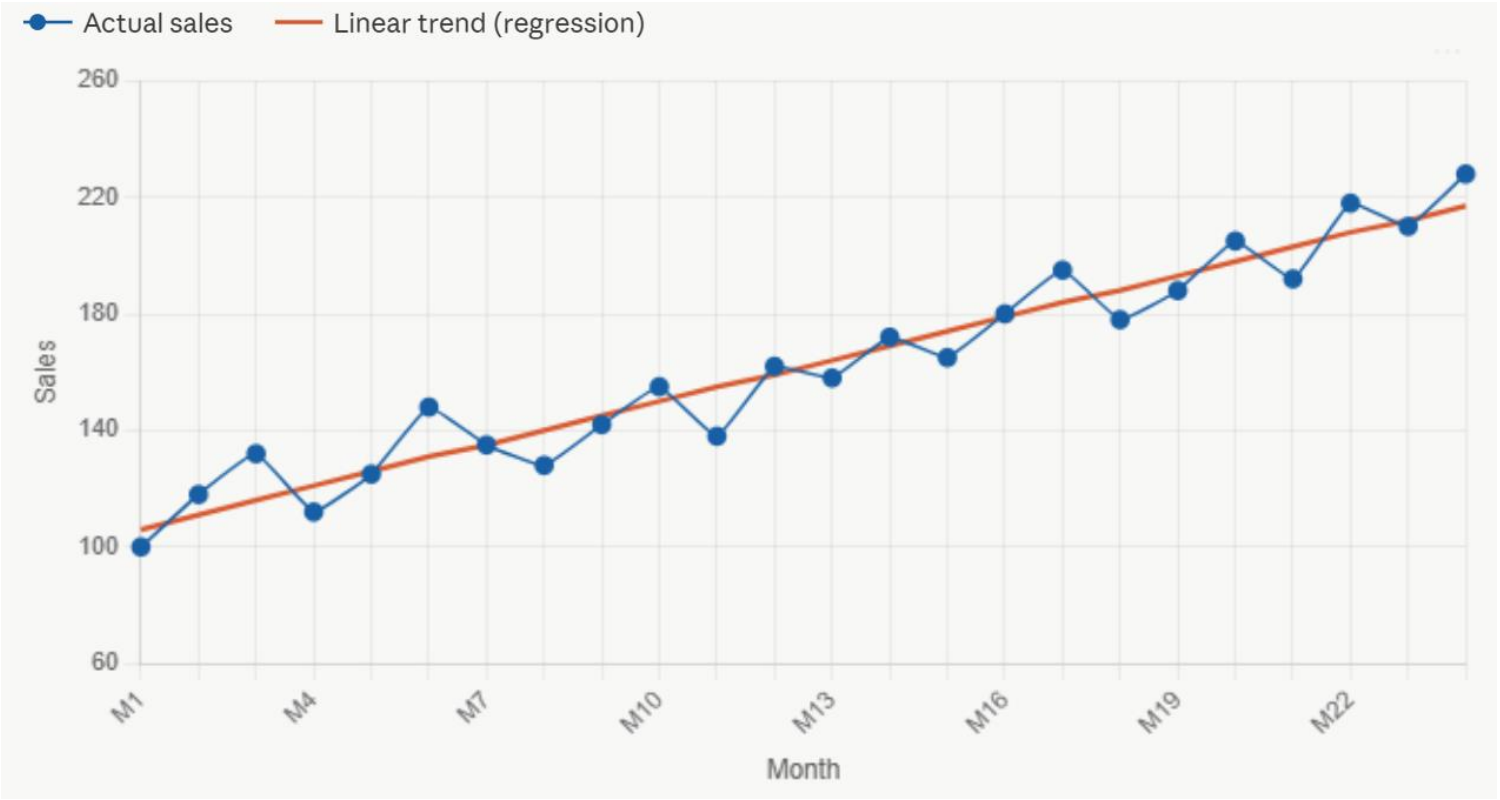
A regression line is fitted:

$$y_t = 80 + 20t$$

To predict sales for month 5:

$$y_5 = 80 + 20(5) = 180$$

Predicted sales = 180



The straight line represents the regression trend used to forecast future values from time series data.

Advantages

- Simple and easy to implement.
- Works well for data with a linear trend.
- Easy to interpret.

Limitations

- Cannot handle seasonality or non-linear patterns.
- Assumes a constant rate of change.
- Not suitable for complex time series data.

6. Applications of Data Science

9. Explain how predictive modelling can be applied for fraud detection.

Predictive modelling is used to detect fraud by analyzing past transaction data and identifying patterns that distinguish fraudulent and normal behaviour. It is mainly a classification problem.

Example Dataset:

Amount	Location	Time	Transactions/Hour	Fraud
500	Mumbai	Day	2	No
20000	Delhi	Night	10	Yes
700	Mumbai	Day	1	No
15000	Unknown	Night	8	Yes

Model learns patterns such as high transaction amount, unusual location, night transactions, and high transaction frequency, which indicate higher probability of fraud

Predictive Model (Classification)

- Target variable: Fraud (Yes/No)
- Model learns:

$$P(\text{Fraud}) = f(\text{Amount, Location, Time, Frequency})$$

- Common algorithms include Logistic Regression, Decision Tree, and Random Forest.

Working

1. Feature Identification

- Important features include transaction amount, location, time, and transaction frequency.

2. Model Training

- Train model using historical labelled data.
- Model learns relationship between features and fraud.

3. Pattern Learning

- Learns normal vs abnormal behaviour.
- Example:
    - ₹500 → normal
    - ₹20,000 at midnight → suspicious

4. Prediction (Real-Time Use)

New transaction:

- Amount = ₹18,000
- Location = Unknown
- Time = Night
- Frequency = High

Model output:

$$P(\text{Fraud}) = 0.92$$

Classified as: Fraud

Advantages

- Detects fraud in real time.
- Reduces financial losses
- Improves security and monitoring.

Conclusion

Predictive modelling helps detect fraud by identifying unusual patterns in transaction data, making financial systems more secure.

10. Explain the steps to build a product recommendation model in detail.

Steps to Build a Recommendation Model

1. Data Collection

- Collect user data such as purchase history, ratings, clicks, and browsing behaviour.
- **Example:** Amazon collects products viewed, bought, and rated by users.

2. Data Preprocessing

- Clean data by removing missing or incorrect values.
- Convert user-item interactions into structured format (user–product matrix).

3. Feature Engineering

- Extract useful features such as user preferences, product categories, and frequency of purchases.
- Represent users and items numerically.

4. Model Selection

- Choose recommendation technique:
  - **Content-Based Filtering** (based on product features)
  - **Collaborative Filtering** (based on similar users/items)
  - **Hybrid Methods** (combination of both)

5. Model Training

- Train the model using historical user-item interaction data.
- Learn patterns such as: users with similar behaviour like similar products.

6. Recommendation Generation

- Generate product suggestions for users.
- **Example:** "Users who bought this also bought..."

7. Model Evaluation

- Evaluate using metrics like precision, recall, and accuracy.
- Check how relevant the recommendations are.

8. Deployment and Update

- Deploy model in real system.
- Continuously update using new user data.

Example

- User buys a mobile phone
- Model recommends phone case, charger, earphones.

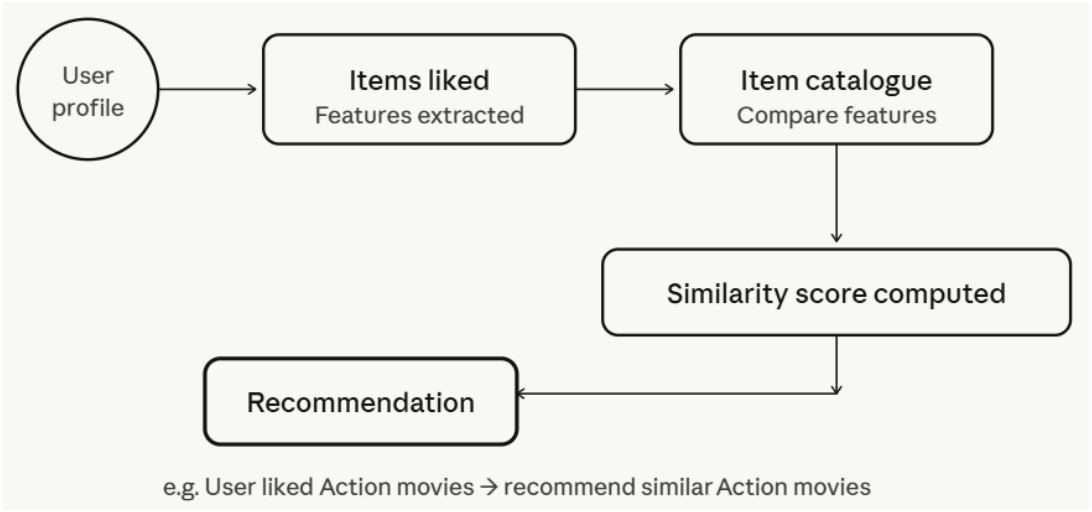
11. What are Recommendation engines? Explain.

Recommendation engines are systems that suggest relevant items to users based on their preferences, behaviour, or past interactions. They are widely used in platforms like e-commerce, streaming services, and social media.

Types of Recommendation Engines

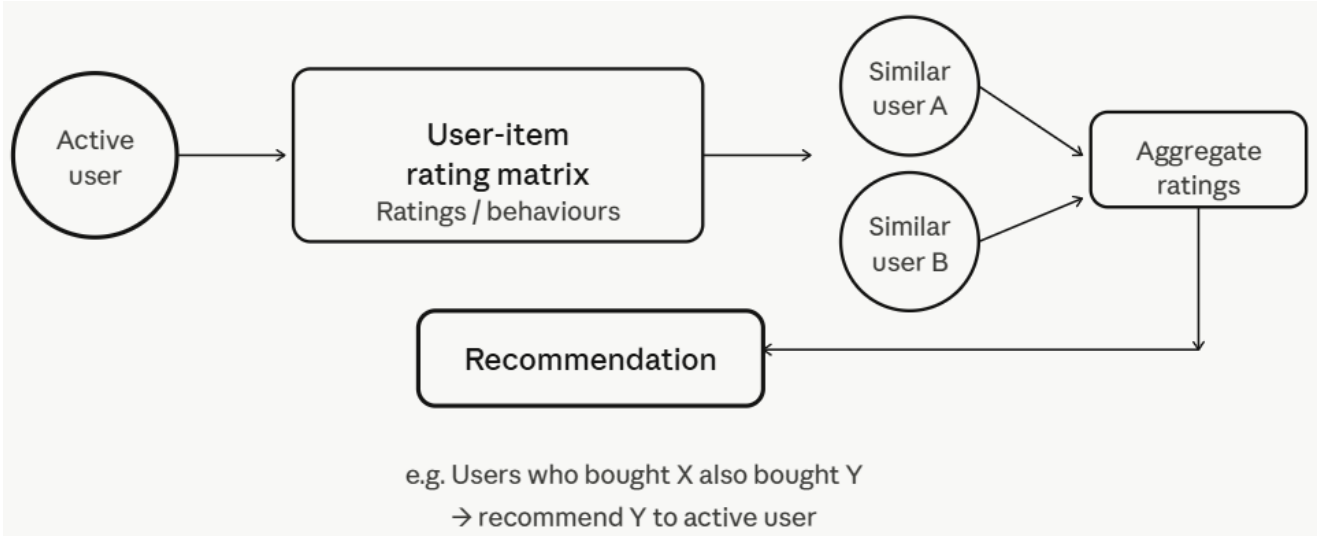
1. Content-Based Filtering

- Recommends items based on similarity between product features.
- Uses user's past preferences to suggest similar items.
- Example: Recommending movies of same genre.



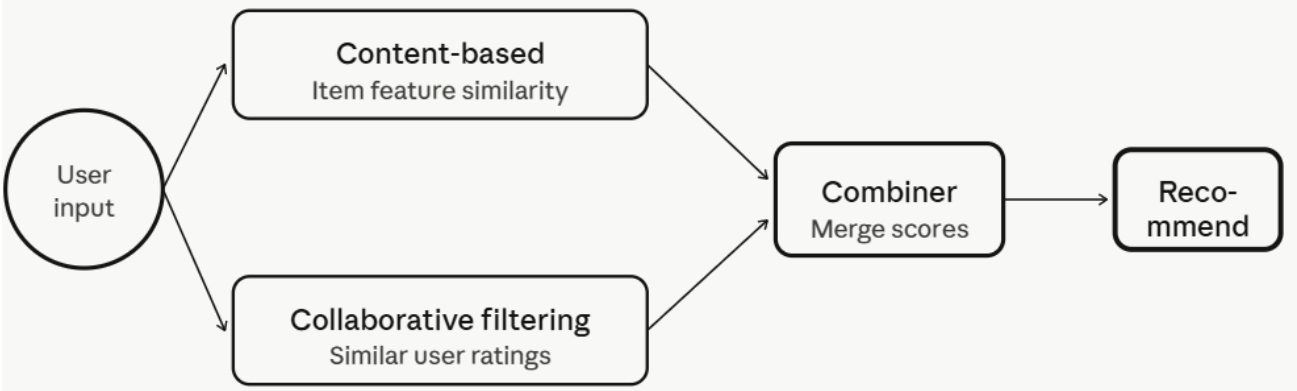
2. Collaborative Filtering

- Recommends items based on similar users or similar user behaviour
- Assumes users with similar interests will like similar items.
- Example: "Users who bought this also bought..."



3. Hybrid Recommendation

- Combines collaborative and content-based methods.
- Improves accuracy and overcomes limitations of individual methods.



~ AJ