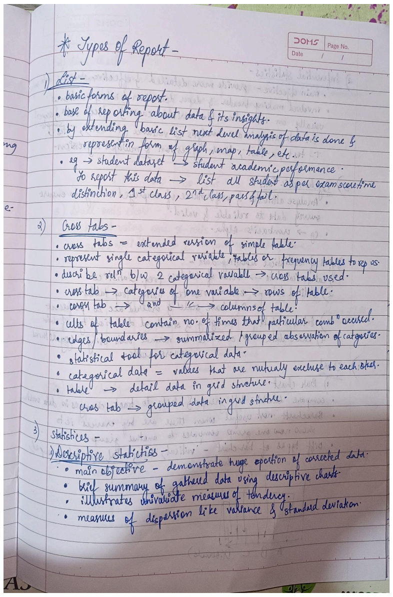
Q1) a) State different types of reports with their application.



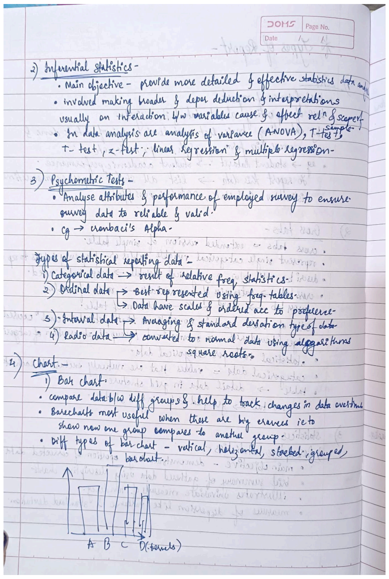
* Types of Report -

1) List -
- basic forms of report.
- base of reporting about data & its insights.
- by extending basic list next level analysis of data is done & represent in form of graph, map, table, etc.
- eg → student dataset → student academic performance. To report this data → list all student as per exam score time distinction, 1st class, 2nd class, pass & fail.

2) Cross tabs -
- cross tabs = extended version of simple table.
- represent single categorical variable, tables or frequency tables to rep us.
- describe reln b/w 2 categorical variable → cross tabs used.
- cross tab → categories of one variable → rows of table.
- cross tab → 2nd " → columns of table.
- cells of table contain no. of times that particular combn occured.
- edges / boundaries → summarized / grouped observation of catgories.
- statistical tool for categorical data.
- categorical data = values that are nutually excluse to each other.
- table → detail data in grid structure.
  cros tab → grouped data in grid structure.

3) Statistices -

1) Descriptive statictiss -
- main objective - demonstrate huge portion of corrected data.
- brief summary of gathered data using descriptive charts.
- illustrates univariate measures of tendency.
- measures of dispersion like variance & standard deviation.



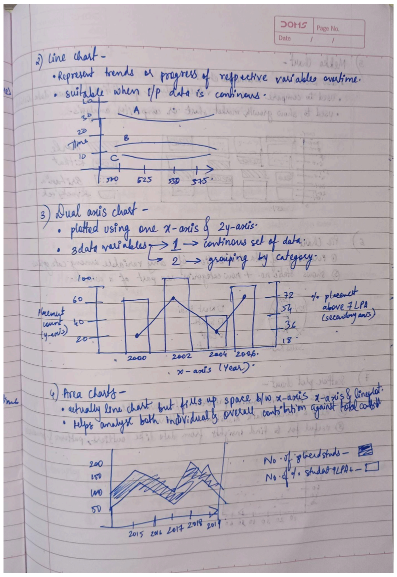
2) Inferential statistics -
- Main objective - provide more detailed & effective statistics data analysis
- involved making broader & deeper deduction & interpretations usually on interaction b/w variables cause & effect reln & scope of
- In data analysis are analysis of variance (ANOVA), T-test, sample T-test, z-test, linear regression & multiple regression.

3) Psychometric Tests -
- Analyse attributes & performance of employed survey to ensure survey data to reliable & valid.
- eg → cronbaci's Alpha.

Types of statistical reporting data :-
1) Categorical data → result of relative freq statistics.
2) Ordinal data → Best represented using freq. tables.
   → Data have scales & ordered acc to preference.
3) Interval data → Averaging & standard deviation type of data
4) Radio data → converted to normal data using algorithms square roots.

4) Chart. -
1) Bar chart:
- compare data b/w diff groups & help to track changes in data overtime
- Barecharts most useful when there are big changes ie to show now one group compares to another group.
- Diff types of bar chart - vertical, horizontal, stacked, grouped, bar chart.

A B C D(brands)



2) Line chart -
- Represent trends or progress of respective variables overtime.
- suitable when i/p data is continous.

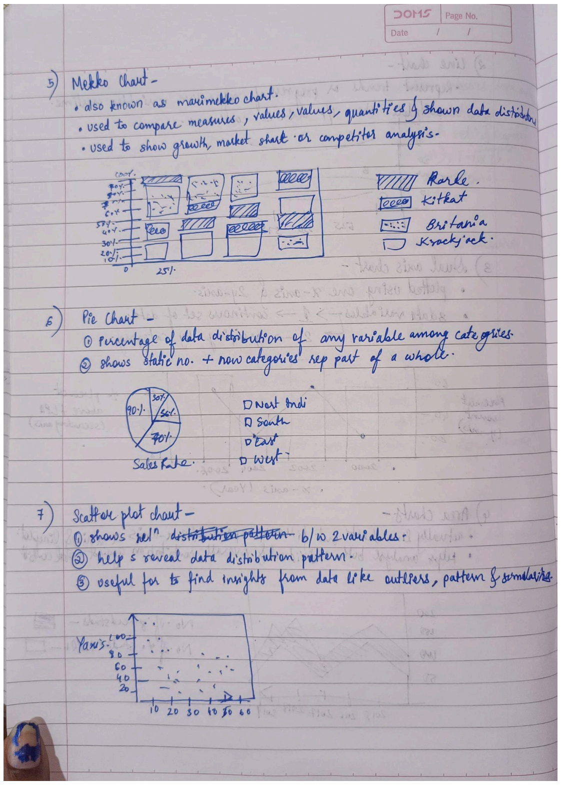
40
30
20
10
Time
A
B
C
500
525
550
575

3) Dual axis chart -
- plotted using one x-axis & 2y-axis.
- 3data variables → 1 → continous set of data.
  → 2 → grouping by category.

100
60
40
20
Placement count (y-axis)
2000
2002
2004
2006
x - axis (Year)
72
54
36
18
% placement above 7 LPA (secondary axis)

4) Area charts -
- actually line chart but fills up space b/w x-axis x-axis & lineplot.
- helps analyse both individual & overall contribution against total contribution

200
150
100
50
2015
2016
2017
2018
2019
No. of placed studs - ▨
No. of % studat 9LPA+ - □



5) Mekko Chart -
- also known as marimekko chart.
- used to compare measures, values, values, quantities & shown data distribution
- used to show growth, market share or competitor analysis.

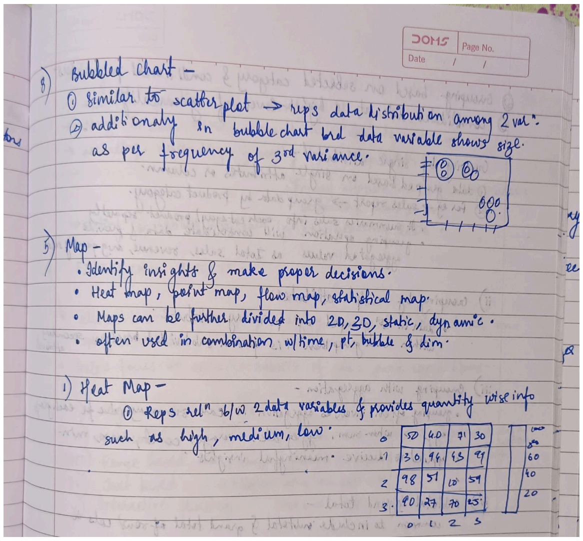
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0
25%
▨ Parle.
Kitkat
Britania
Krackjack.

6) Pie Chart -
1. Percentage of data distribution of any variable among categories.
2. shows static no. + now categories rep part of a whole.

30%
50%
90%
70%
Sales Rate.
□ Nort Indi
□ South
□ East
□ West

7) Scatter plot chart -
1. shows reln ~~distribution pattern~~ b/w 2 variables.
2. helps reveal data distribution pattern.
3. useful for to find insights from data like outliers, pattern & similarities.

Y axis.
100
80
60
40
20
10
20
30
40
50
60



8) Bubbled chart -
1. similar to scatterplot → reps data distribution among 2 varn.
2. additionaly in bubble chart 3rd data variable shows size as per frequency of 3rd variance.

5) Map -
- Identify insights & make proper decisions.
- Heat map, point map, flow map, statistical map.
- Maps can be further divided into 2D, 3D, static, dynamic.
- often used in combination w/time, pt, bubble & dim.

1) Heat Map -
1. Reps reln b/w 2 data variables & provides quantity wise info such as high, medium, low.

	0	1	2	3
0	50	40	71	30
1	30	94	43	97
2	98	51	10	59
3	90	27	70	25

Q1) B) What are the best practices in dashboard design?

Know your audience: Cater to their goals and information needs.
Start with a purpose: Clearly define what insights the dashboard should deliver.
Simplicity is key: Avoid overwhelming users with excessive data or visuals.
Clarity matters: Use clear labels and prioritize visual hierarchy for easy understanding.
Visualizations that work: Choose charts that effectively represent the data.
Strategic color use: Colors can highlight info but avoid overuse or clashing.
Consistency is king: Maintain consistent formatting and styling throughout.
Interactivity is key: Allow users to filter, drill down, or explore different timeframes.
Context is king: Provide explanations or tooltips for deeper understanding.
Right tool for the job: Pick a BI tool with features that match your needs.
Whitespace is your friend: Use it to improve readability and avoid clutter.
Test and improve: Gather user feedback and iterate on your design.
Data security matters: Restrict access and keep data secure.

Business intelligence (BI) dashboards are all about making data understandable and actionable. Here are some key best practices to follow for effective BI dashboard design:

Know your audience and goals:

- Who will be using the dashboard? Executives? Sales teams? Tailor the information and complexity to their needs.
- What do you want users to achieve with the dashboard? Make informed decisions? Track progress towards a goal? Design the layout and metrics to support those goals.

Focus on clarity and conciseness:

- Avoid cramming too much information onto a single screen. Prioritize the most important KPIs (Key Performance Indicators).
- Use clear labels, consistent formatting, and uncluttered visuals. Let the data itself be the star.

Choose the right visuals:

- Don't use fancy charts if a simple bar graph tells the story better. Match the chart type to the data you're presenting (e.g., pie charts for proportions, line charts for trends).

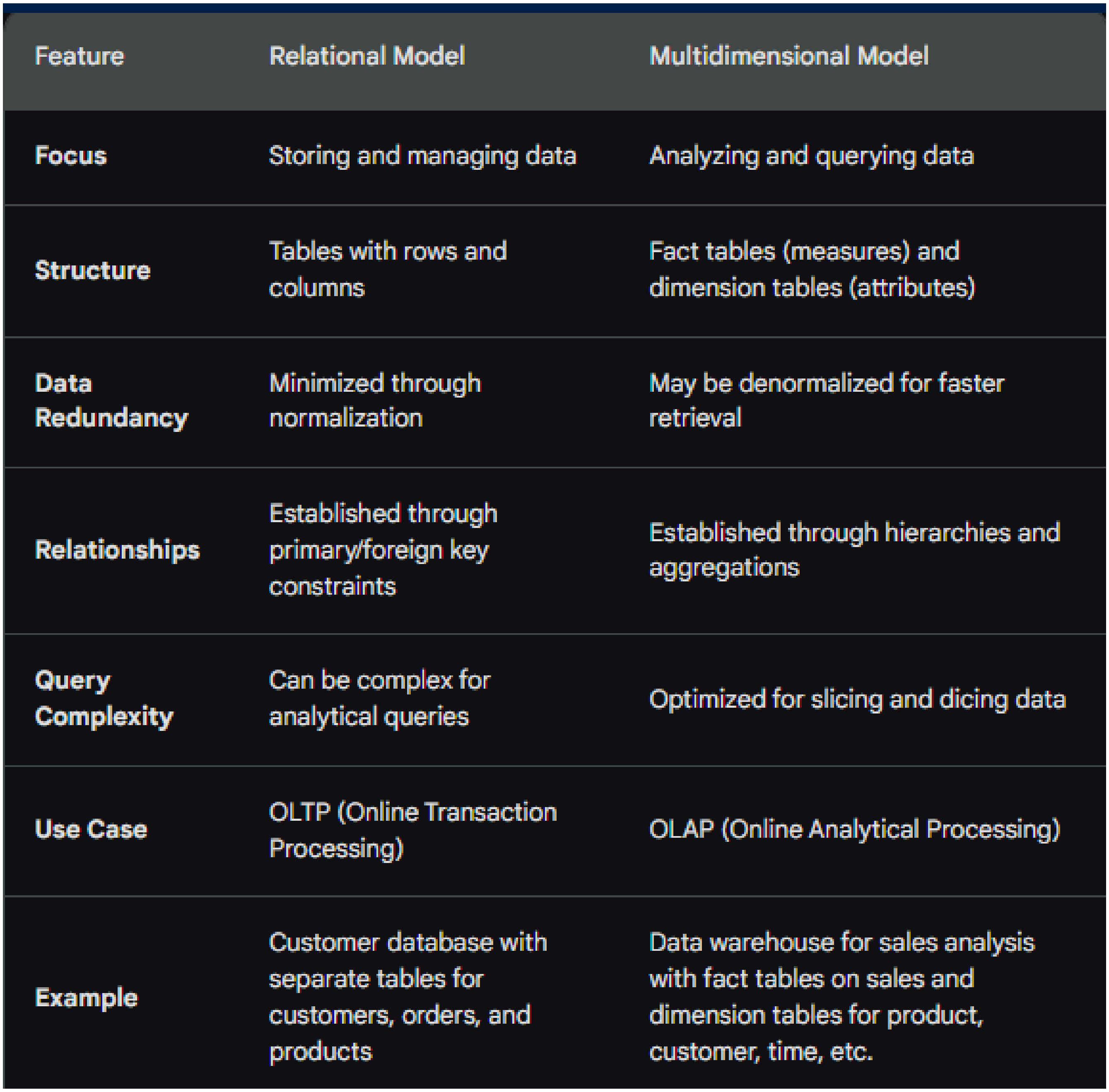
Let users interact with the data:

- Allow users to filter data by time period, department, etc. This empowers them to find the information they need quickly.
- Consider adding drill-down capabilities, where users can click on a data point to see more details.

Design for usability:

- Ensure the dashboard is easy to navigate and understand.
- Optimize for different devices (desktop, mobile) so users can access information on the go.

Q1) C) State the difference between relational and multidimensional data model.



Feature	Relational Model	Multidimensional Model
Focus	Storing and managing data	Analyzing and querying data
Structure	Tables with rows and columns	Fact tables (measures) and dimension tables (attributes)
Data Redundancy	Minimized through normalization	May be denormalized for faster retrieval
Relationships	Established through primary/foreign key constraints	Established through hierarchies and aggregations
Query Complexity	Can be complex for analytical queries	Optimized for slicing and dicing data
Use Case	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
Example	Customer database with separate tables for customers, orders, and products	Data warehouse for sales analysis with fact tables on sales and dimension tables for product, customer, time, etc.

Q2) A) Suggest the use of Data Grouping & Sorting, Filtering Reports.

Data Grouping, Sorting, and Filtering are all powerful tools for making sense of large datasets in reports. Here's a breakdown of their uses:

Data Grouping

- **Purpose:** Organize data into categories for easier analysis.
- **Use Case:** Imagine a sales report. You can group data by product category to see which categories are performing well.
- **Benefit:** Identify trends and patterns within specific groups.

Data Sorting

- **Purpose:** Arrange data in a specific order (ascending or descending) based on a chosen column.
- **Use Case:** In a customer list report, sort by purchase history (highest to lowest) to identify your most valuable customers.
- **Benefit:** Quickly find the most important information or spot outliers.

Filtering Reports

- **Purpose:** Focus on a specific subset of data that meets certain criteria.
- **Use Case:** In a website traffic report, filter by a specific time period (e.g., Black Friday week) to analyze sales trends during that period.
- **Benefit:** Simplify complex data sets and focus on the information most relevant to your current task.

Working Together:

These techniques are often used together for even more powerful analysis. For instance, you can group data, then sort within each group for a more granular view. Additionally, you can filter the data before grouping or sorting to focus on a specific subset from the start.

By effectively using these features, you can transform raw data into clear and actionable insights.

Q2) B) What is a File Extension? Explain the structure of CSV file

A file extension is a short identifier typically added to the end of a filename to indicate the type of file it is. It acts like a label that tells your computer what program to use to open the file.

CSV (Comma-Separated Values) files are a popular format for transferring data between different applications because of their simplicity and universality.

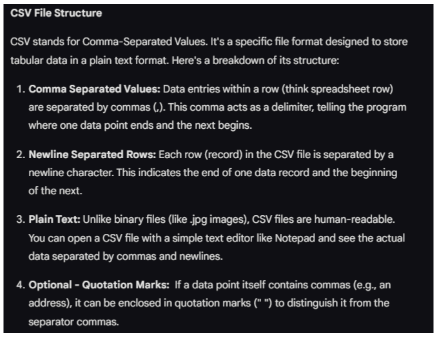
Here are some of the common uses of CSV files:

Data Exchange:

- **Spreadsheets:** The most common use of CSV files is to transfer data between spreadsheets like Microsoft Excel, Google Sheets, etc. CSV provides a common format that different spreadsheet programs can understand and import/export data from.
- **Databases:** Databases can also use CSV files to import or export data. This allows for easy migration of data between different database systems or transferring data from a database to a spreadsheet for analysis.

Data Analysis:

- **Data Cleaning & Manipulation:** CSV files are often used as an intermediate step in data analysis workflows. Data can be extracted from various sources into CSV format, then cleaned, filtered, or manipulated using programming languages like Python before being loaded into a data analysis tool.
- **Sharing Data Sets:** Researchers and analysts often share data sets in CSV format for reproducibility and ease of use by others. The simple structure allows researchers to import the data into their preferred analysis tools.
- **System Backups:** Some basic applications might use CSV files for simple data backups or configuration settings. The easy human-readable format allows for straightforward backups and restores.
- **Mailing Lists:** Email marketing tools may allow uploading CSV files containing email addresses and other subscriber information for creating mailing lists.



CSV File Structure

CSV stands for Comma-Separated Values. It's a specific file format designed to store tabular data in a plain text format. Here's a breakdown of its structure:

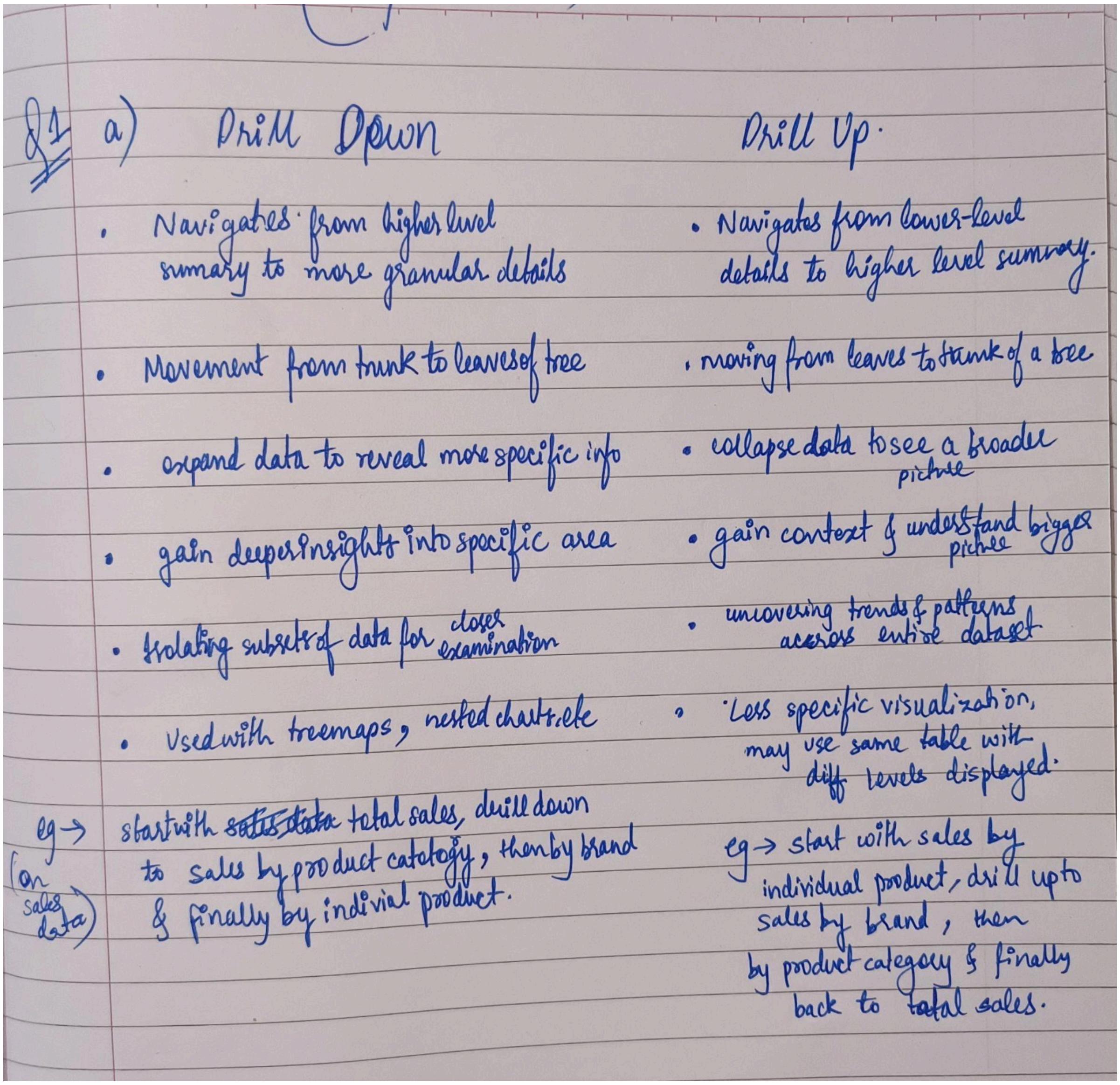
1. **Comma Separated Values:** Data entries within a row (think spreadsheet row) are separated by commas (,). This comma acts as a delimiter, telling the program where one data point ends and the next begins.

2. **Newline Separated Rows:** Each row (record) in the CSV file is separated by a newline character. This indicates the end of one data record and the beginning of the next.

3. **Plain Text:** Unlike binary files (like .jpg images), CSV files are human-readable. You can open a CSV file with a simple text editor like Notepad and see the actual data separated by commas and newlines.

4. **Optional - Quotation Marks:** If a data point itself contains commas (e.g., an address), it can be enclosed in quotation marks (" ") to distinguish it from the separator commas.

Q2) C) Explain in detail Drill up and Drill Down.



Q1 a)

	Drill Down	Drill Up.
	Navigates from higher level sumary to more granular details	Navigates from lower-level details to higher level summary.
	Movement from trunk to leaves of tree	moving from leaves to trunk of a tree
	expand data to reveal more specific info	collapse data to see a broader picture
	gain deeper insights into specific area	gain context & understand bigger picture
	Isolating subsets of data for closer examination	uncovering trends & patterns across entire dataset
	Used with treemaps, nested charts, etc	Less specific visualization, may use same table with diff levels displayed.
eg→ (on sales data)	start with ~~sales data~~ total sales, drill down to sales by product catatogy, then by brand & finally by indivial product.	eg→ start with sales by individual product, drill upto sales by brand, then by product category & finally back to total sales.

Q3) A) mean , median mode (numerical skipped)
Q3) b) What is data Transformation? Explain Data Transformation Process in Detail

Data transformation is the process of converting raw data into a usable format that's suitable for analysis and further processing. Imagine raw data as a pile of unorganized rocks. Data transformation takes those rocks and crushes, sorts, and cleans them into usable gravel or building blocks.

Here's a detailed breakdown of the data transformation process:

1. Data Understanding:

- This initial step involves getting to know your data. You'll analyze the data source, identify its format (CSV, JSON, etc.), and understand the meaning and structure of the data points (columns).

2. Data Cleaning:

- Raw data is rarely perfect. This stage focuses on fixing errors and inconsistencies. Common cleaning tasks include:
  - Handling missing values: Filling in missing data points with estimated values or removing them entirely depending on the situation.
  - Identifying and correcting inconsistencies: Fixing typos, standardizing formats (e.g., converting dates to a consistent format), and addressing any outliers that might skew results.

3. Data Integration:

- Often, data comes from multiple sources (databases, spreadsheets, etc.). Integration involves combining this data into a single, unified format. This may involve resolving data conflicts (e.g., duplicate entries) and ensuring all data adheres to a consistent structure.

4. Data Transformation:

- This is where the real transformation happens. You'll manipulate the data to suit your analysis needs. Common transformations include:
  - Deriving new attributes: Creating new data points based on existing ones (e.g., calculating total sales from product quantity and price).
  - Data normalization: Scaling numerical data to a common range for better analysis in some statistical methods.
  - Data aggregation: Summarizing data by grouping similar records together (e.g., calculating total sales per product category).

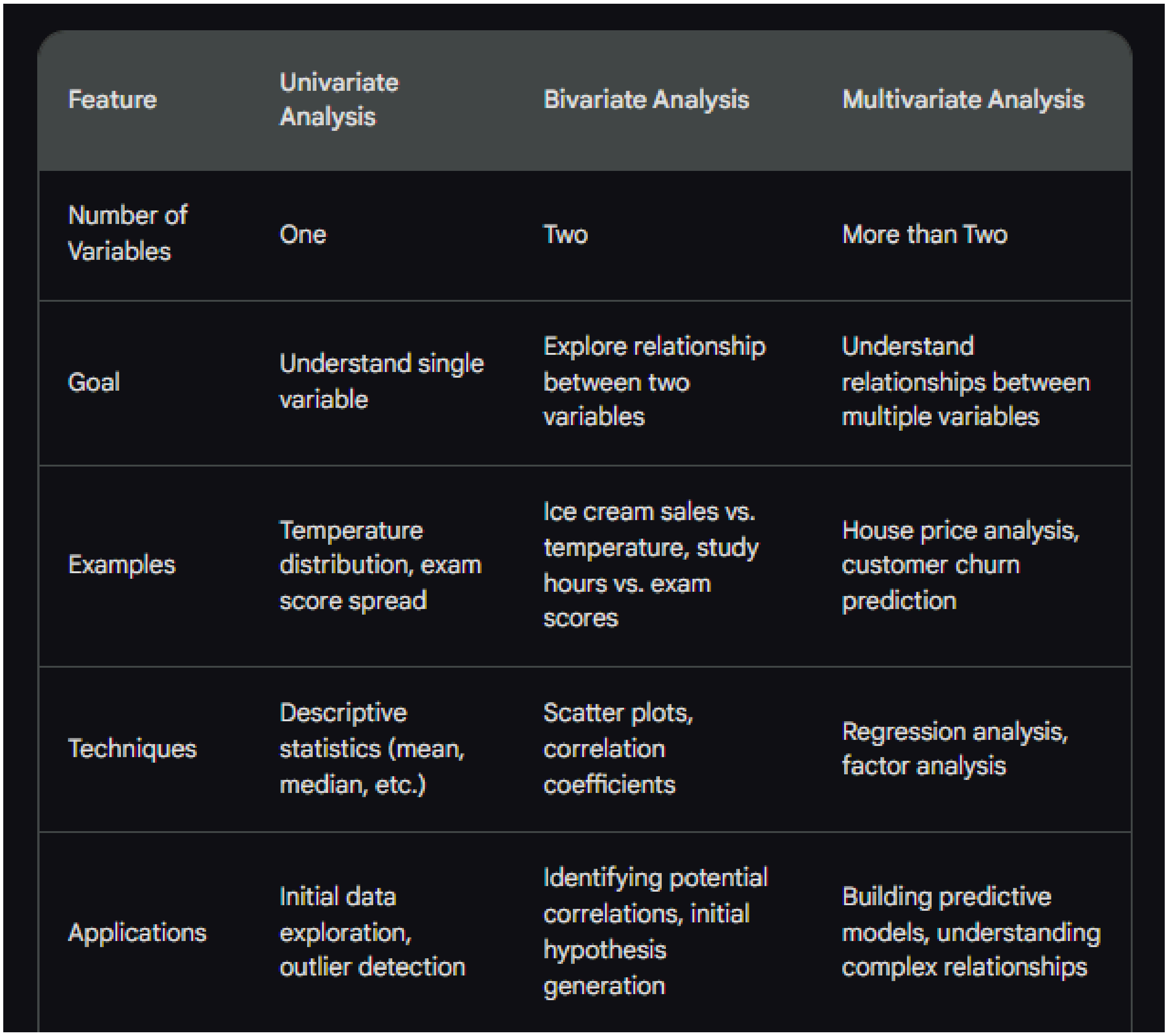
5. Data Validation:

- After transformation, it's crucial to ensure the data is accurate and reflects the intended changes. Data validation involves checking for errors introduced during the transformation process and making sure the transformed data aligns with the business goals.

6. Data Storage:

- The final step involves storing the transformed data in a usable format for further analysis or use in other applications. This could be a data warehouse, data lake, or another data storage system depending on your needs.

Q3) C) Explain univariate, bi variate and multivariate analysis with example and applications.



Feature	Univariate Analysis	Bivariate Analysis	Multivariate Analysis
Number of Variables	One	Two	More than Two
Goal	Understand single variable	Explore relationship between two variables	Understand relationships between multiple variables
Examples	Temperature distribution, exam score spread	Ice cream sales vs. temperature, study hours vs. exam scores	House price analysis, customer churn prediction
Techniques	Descriptive statistics (mean, median, etc.)	Scatter plots, correlation coefficients	Regression analysis, factor analysis
Applications	Initial data exploration, outlier detection	Identifying potential correlations, initial hypothesis generation	Building predictive models, understanding complex relationships

Q4) A) What is a Contingency Table? What is Marginal Distribution? Justify with suitable example.

Contingency Table:

1. Tabular representation of the joint distribution of two or more categorical variables.
2. Shows frequencies or counts of observations for each combination of categories.
3. Useful for examining relationships between categorical variables and performing hypothesis tests.

Marginal Distribution:

1. Distribution of frequencies or proportions of one categorical variable in a contingency table.
2. Shows the overall distribution of one variable across its categories.
3. Helps understand characteristics of the sample population without considering relationships between variables.

Example:

- **Contingency Table:**

```
             | Liberal | Conservative | Total
------------------------------------------------
Male          | 150     | 100          | 250
Female        | 200     | 120          | 320
Total         | 350     | 220          | 570
```

- **Marginal Distribution:**
    - Gender: Male - 43.86%, Female - 56.14%
    - Voting Preference: Liberal - 61.40%, Conservative - 38.60%

These concepts provide insights into categorical data relationships and overall distributions, facilitating further analysis and interpretation.

Q4) B) Explain data validation, Incompleteness, noise, inconsistency of quality of input data.

1. Data Validation:

- **Concept:** Data validation refers to the process of ensuring that the data entered into a system meets predefined criteria. This helps prevent errors and inconsistencies from entering the data pool in the first place.
- **Example:** Imagine an online form where users enter their age. Data validation can ensure the entered value is a number within a reasonable age range (e.g., 18-100).
- **Solutions:** Techniques like data type checks (numbers only for age), format checks (valid email format), and range checks (age within limits) can be implemented during data entry.

2. Data Incompleteness:

- **Concept:** Data incompleteness refers to missing values within a dataset. This can occur due to human error during data entry, system failures, or limitations in data collection methods.
- **Example:** A customer survey might have missing responses for some questions if participants skipped them.
- **Solutions:** Depending on the situation, missing values can be imputed (estimated based on other data points), ignored if a small percentage of data is missing, or the data collection process can be improved to minimize missing entries.

3. Noise:

- **Concept:** Data noise refers to errors or random fluctuations within the data that can distort the true signal or pattern. This can arise from faulty sensors, data transmission errors, or human error during data entry.
- **Example:** Temperature readings from a sensor might be slightly inaccurate due to sensor malfunction.
- **Solutions:** Data cleaning techniques like outlier detection (identifying and removing extreme values) and data smoothing (averaging multiple data points) can help mitigate the impact of noise.

4. Inconsistency of data quality,-
also referred to as data heterogeneity, refers to the lack of uniformity within a dataset. This means the data exhibits variations in its characteristics that can hinder analysis and lead to inaccurate results.

Causes of Inconsistency:

- **Multiple Data Sources:** When data is integrated from various sources, inconsistencies can arise due to differences in data collection methods, storage formats, or internal coding schemes used by each source.
- **Manual Data Entry:** Human error during data entry can lead to inconsistencies in formats, spellings, or how information is captured.
- **Changes over Time:** Data collection practices or coding schemes might evolve over time, leading to inconsistencies between older and newer data points within a dataset.

Impacts of Inconsistency:

- **Hindering Data Analysis:** Inconsistent data can make it difficult to merge or compare data points from different parts of the dataset, hindering analysis and potentially leading to skewed results.
- **Misleading Insights:** Inconsistent data can lead to inaccurate or misleading conclusions if the variations are not properly accounted for during analysis.
- **Inefficient Data Processing:** Inconsistent data formats can complicate data cleaning, transformation, and processing steps, requiring additional effort to achieve consistency.

Mitigating Inconsistency:

- **Data Standardization:** Define and enforce consistent data formats, units of measurement, and coding schemes across all data sources and throughout the data lifecycle.
- **Data Cleaning:** Implement processes to identify and rectify inconsistencies in existing data, potentially involving data transformation or manual correction.
- **Data Profiling:** Regularly analyze your data to identify and address potential inconsistencies before they impact analysis.
- **Data Governance:** Establish clear data governance policies that outline data quality standards and procedures for maintaining consistency.

Q4) C) Explain following Data reduction technique: Sampling, Feature selection, Principal component analysis

Sampling, Feature Selection, and Principal Component Analysis (PCA) - address the challenge of high dimensionality in data analysis. High dimensionality refers to datasets with a large number of variables. While more variables can provide a richer picture, it can also lead to issues like:

- **Increased computational cost:** Processing and analyzing massive datasets with many variables can be time-consuming and resource-intensive.
- **The Curse of Dimensionality:** With more variables, the space needed to represent the data grows exponentially. This can make it harder to identify meaningful patterns and relationships.
- **Overfitting:** Models trained on high-dimensional data can become overly complex and fit the training data too closely, leading to poor performance on unseen data (generalizability).

Here's a breakdown of how each technique tackles dimensionality reduction:

1. Sampling:

- **Concept:** Instead of analyzing the entire dataset, sampling involves selecting a representative subset of data points. There are different sampling methods (random, stratified, etc.) to ensure the chosen subset reflects the characteristics of the whole dataset.
- **Benefits:**
  - Reduced processing time and computational cost.
  - Can still provide accurate insights if the sample is chosen carefully.
- **Drawbacks:**
  - The representativeness of the sample is crucial. A poorly chosen sample can lead to misleading results.
  - Not ideal for situations where every data point is important (e.g., financial transactions).

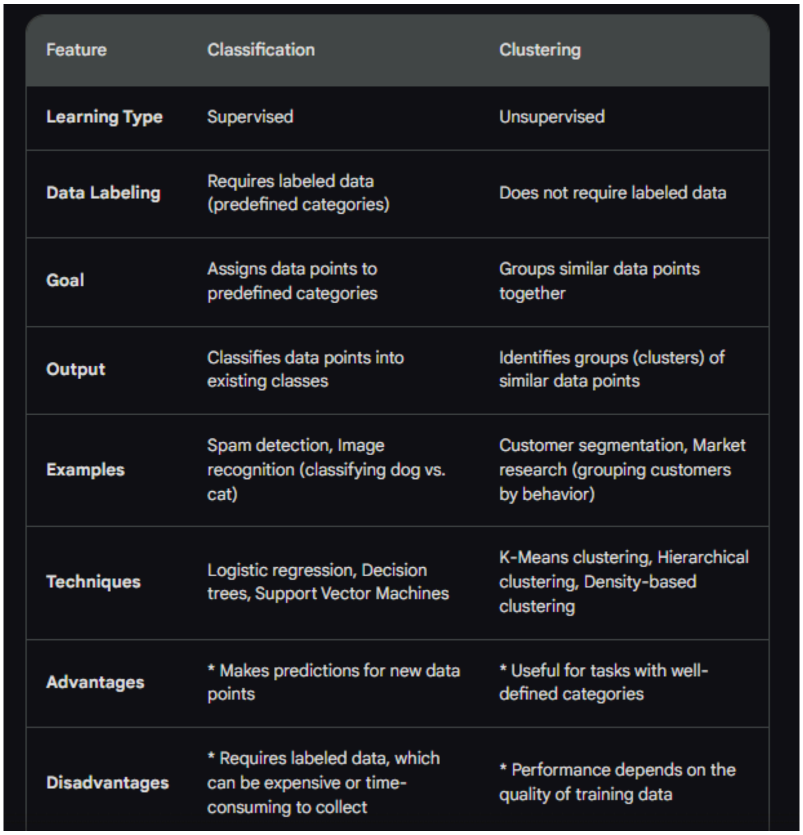
2. Feature Selection:

- **Concept:** This technique focuses on identifying and keeping only the most relevant features (variables) that contribute significantly to the analysis.
- **Benefits:**
  - Improves model performance by reducing noise and irrelevant information.
  - Simplifies models, making them easier to interpret and understand.
- **Drawbacks:**
  - Choosing the right features requires domain knowledge and can be subjective.
  - Excluding important features can lead to inaccurate results.

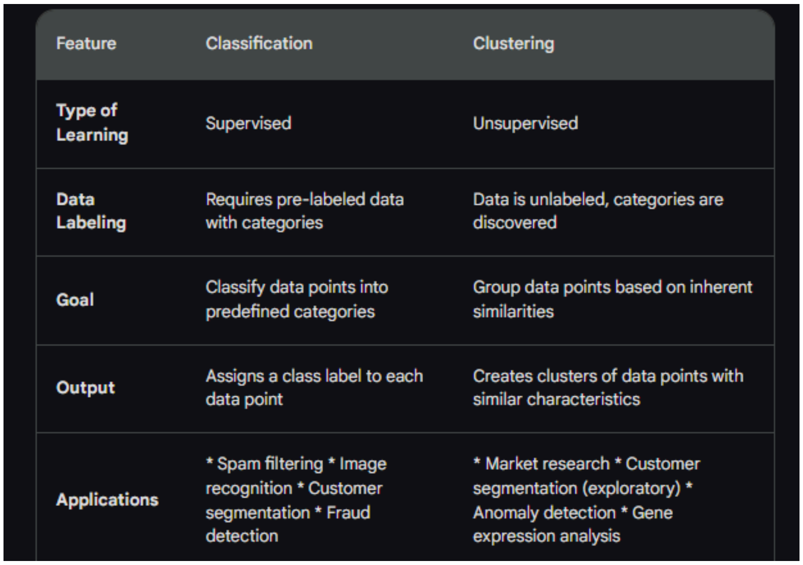
3. Principal Component Analysis (PCA):

- **Concept:** PCA is a transformation technique that creates a new set of features (called principal components) from the existing ones. These new features capture the most significant variation in the data while minimizing redundancy.
- **Benefits:**
  - Reduces dimensionality while preserving most of the important information in the data.
  - Improves model performance and reduces overfitting.
- **Drawbacks:**
  - PCA assumes a linear relationship between features, which may not always be the case.
  - Interpreting the new principal components can be challenging compared to original features.

Q5) A) Write a difference between classification and clustering with applications.

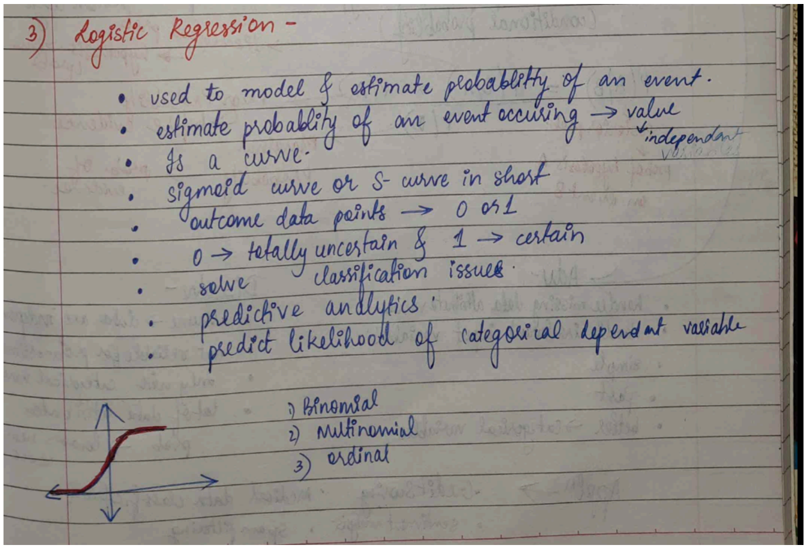


Feature	Classification	Clustering
Learning Type	Supervised	Unsupervised
Data Labeling	Requires labeled data (predefined categories)	Does not require labeled data
Goal	Assigns data points to predefined categories	Groups similar data points together
Output	Classifies data points into existing classes	Identifies groups (clusters) of similar data points
Examples	Spam detection, Image recognition (classifying dog vs. cat)	Customer segmentation, Market research (grouping customers by behavior)
Techniques	Logistic regression, Decision trees, Support Vector Machines	K-Means clustering, Hierarchical clustering, Density-based clustering
Advantages	* Makes predictions for new data points	* Useful for tasks with well-defined categories
Disadvantages	* Requires labeled data, which can be expensive or time-consuming to collect	* Performance depends on the quality of training data



Feature	Classification	Clustering
Type of Learning	Supervised	Unsupervised
Data Labeling	Requires pre-labeled data with categories	Data is unlabeled, categories are discovered
Goal	Classify data points into predefined categories	Group data points based on inherent similarities
Output	Assigns a class label to each data point	Creates clusters of data points with similar characteristics
Applications	* Spam filtering * Image recognition * Customer segmentation * Fraud detection	* Market research * Customer segmentation (exploratory) * Anomaly detection * Gene expression analysis

Q5) B) Write a short note on Logistic Regression.



3) Logistic Regression -

- used to model & estimate probablity of an event.
- estimate probablity of an event occuring → value
  independant variables
- Is a curve.
- sigmoid curve or S- curve in short
- outcome data points → 0 or 1
- 0 → totally uncertain & 1 → certain
- solve classification issues.
- predictive analytics :
- predict likelihood of categorical dependent variable

1) Binomial
2) Multinomial
3) ordinal

A statistical method for **classification** tasks in machine learning. It predicts the probability of an event belonging to one of two categories (e.g., spam or not spam).

Advantages (ADV):

- **Simple to understand and interpret:** Coefficients reveal which factors most influence the prediction.
- **Good for binary classification:** Works well for problems with two outcome categories.
- **Efficient for large datasets:** Can handle large amounts of data efficiently.

Disadvantages (DIS):

- **Assumes linear relationships:** May not be suitable for complex relationships between features.
- **Limited to two categories:** Can't handle problems with more than two outcome categories (multiclass).
- **Data quality sensitive:** Relies on good quality data for accurate predictions.

Method:

1. Uses features (data points) to estimate the probability of belonging to a class (0 to 1).
2. Employs the sigmoid function to transform a linear combination of features into a probability.
3. Sets a decision boundary (often 0.5) to classify data points based on the predicted probability.

Example:

- **Scenario:** Predicting loan approval (Yes/No) based on income, credit score, and loan amount.
- **Model:** Logistic regression analyzes historical loan data to learn the relationships between these factors and approval.
- **Prediction:** For a new loan application, the model calculates the probability of approval based on the applicant's data.
- **Decision:** The bank uses the probability (and a threshold) to decide whether to approve the loan.

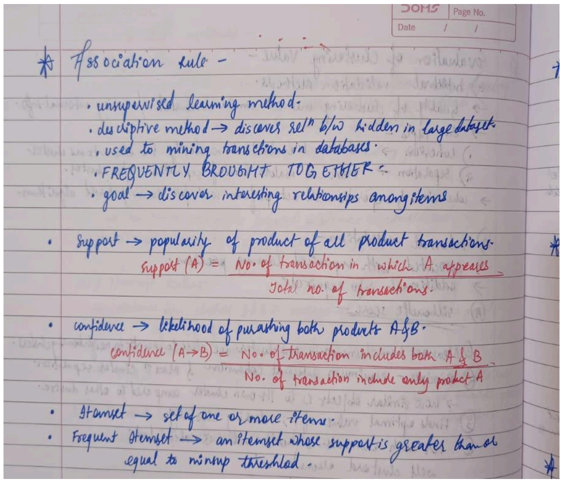
Q5) C) skipped(apriori algo numerical)

Q6) A) What are association rules? How to evaluate them using Support and Confidence? Explain with Example.

Association rules are a fundamental concept in data mining, used primarily to discover interesting relationships, patterns, or associations among a set of items in large databases. These rules are often applied in market basket analysis to identify products that frequently co-occur in transactions.

Key Concepts

1. **Itemset**: A collection of one or more items.
2. **Support**: The frequency or proportion of transactions in the dataset that contain a particular itemset.
3. **Confidence**: The likelihood that a transaction containing a certain itemset also contains another itemset.



* Association rule -

- unsupervised learning method.
- descriptive method → discovers rel^n b/w hidden in large dataset.
- used to mining transctions in databases.
- FREQUENTLY BROUGHT TOGETHER.
- goal → discover interesting relationships among items

- Support → popularity of product of all product transactions.

$$\text{support}(A) = \frac{\text{No. of transaction in which A appeares}}{\text{Total no. of transactions.}}$$

- confidence → likelihood of purashing both products A&B.

$$\text{confidence}(A \rightarrow B) = \frac{\text{No. of transaction includes both A \& B}}{\text{No. of transaction include only product A}}$$

- Itemset → set of one or more items.
- Frequent Itemset → an itemset whose support is greater than or equal to minsup threshlad.

Example

Consider a small database of transactions as follows:

Transaction ID	Items Purchased
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

Let's derive association rules and evaluate them using support and confidence.

Step 1: Calculate Support

First, find the support for each itemset:

- Support(Bread) = 4/5 = 0.8
- Support(Milk) = 4/5 = 0.8
- Support(Diapers) = 4/5 = 0.8
- Support(Beer) = 3/5 = 0.6
- Support(Bread, Milk) = 3/5 = 0.6
- Support(Bread, Diapers) = 3/5 = 0.6
- Support(Milk, Diapers) = 3/5 = 0.6
- Support(Bread, Milk, Diapers) = 2/5 = 0.4

Step 2: Generate Rules and Calculate Confidence

Now, generate possible rules and calculate their confidence:

1. **Rule: Bread → Milk**
   - Support(Bread ∩ Milk) = Support(Bread, Milk) = 0.6
   - Confidence(Bread → Milk) = Support(Bread ∩ Milk) / Support(Bread) = 0.6 / 0.8 = 0.75
2. **Rule: Milk → Bread**
   - Support(Milk ∩ Bread) = Support(Bread, Milk) = 0.6
   - Confidence(Milk → Bread) = Support(Milk ∩ Bread) / Support(Milk) = 0.6 / 0.8 = 0.75
3. **Rule: Diapers → Beer**
   - Support(Diapers ∩ Beer) = 3/5 = 0.6
   - Confidence(Diapers → Beer) = Support(Diapers ∩ Beer) / Support(Diapers) = 0.6 / 0.8 = 0.75
4. **Rule: Bread, Milk → Diapers**
   - Support(Bread ∩ Milk ∩ Diapers) = 0.4
   - Confidence(Bread, Milk → Diapers) = Support(Bread ∩ Milk ∩ Diapers) / Support(Bread ∩ Milk) = 0.4 / 0.6 = 0.67

Interpretation

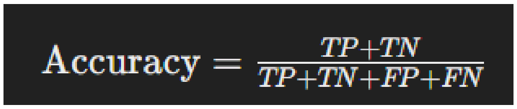
- The rule **Bread → Milk** has a support of 0.6 and a confidence of 0.75. This means 60% of all transactions contain both Bread and Milk, and if a transaction contains Bread, there is a 75% chance it also contains Milk.
- The rule **Diapers → Beer** has a confidence of 75%, indicating that in 75% of the cases where Diapers are purchased, Beer is also purchased.

Q6) B)State different formulae for Evaluation of classification models.

Evaluating classification models is crucial to determine their effectiveness and performance. There are several metrics and formulae commonly used for this purpose. Here are some of the key evaluation metrics:

1. Accuracy

Accuracy is the proportion of correctly classified instances (both true positives and true negatives) among the total number of instances.

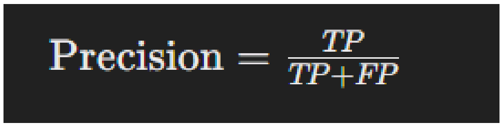


$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **TP**: True Positives
- **TN**: True Negatives
- **FP**: False Positives
- **FN**: False Negatives

2. Precision

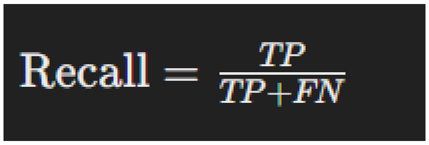
Precision, also known as Positive Predictive Value, measures the proportion of true positives among the instances classified as positive.



$$\text{Precision} = \frac{TP}{TP+FP}$$

3. Recall (Sensitivity or True Positive Rate)

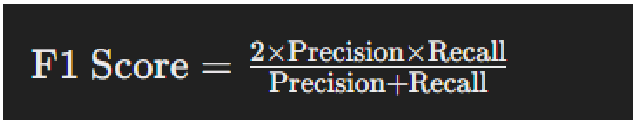
Recall measures the proportion of true positives among the actual positives.



$$\text{Recall} = \frac{TP}{TP+FN}$$

4. F1 Score

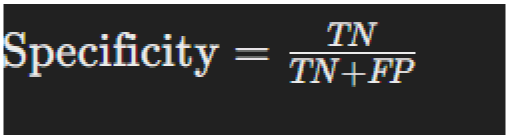
The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics.



$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Specificity (True Negative Rate)

Specificity measures the proportion of true negatives among the actual negatives.



$$\text{Specificity} = \frac{TN}{TN + FP}$$

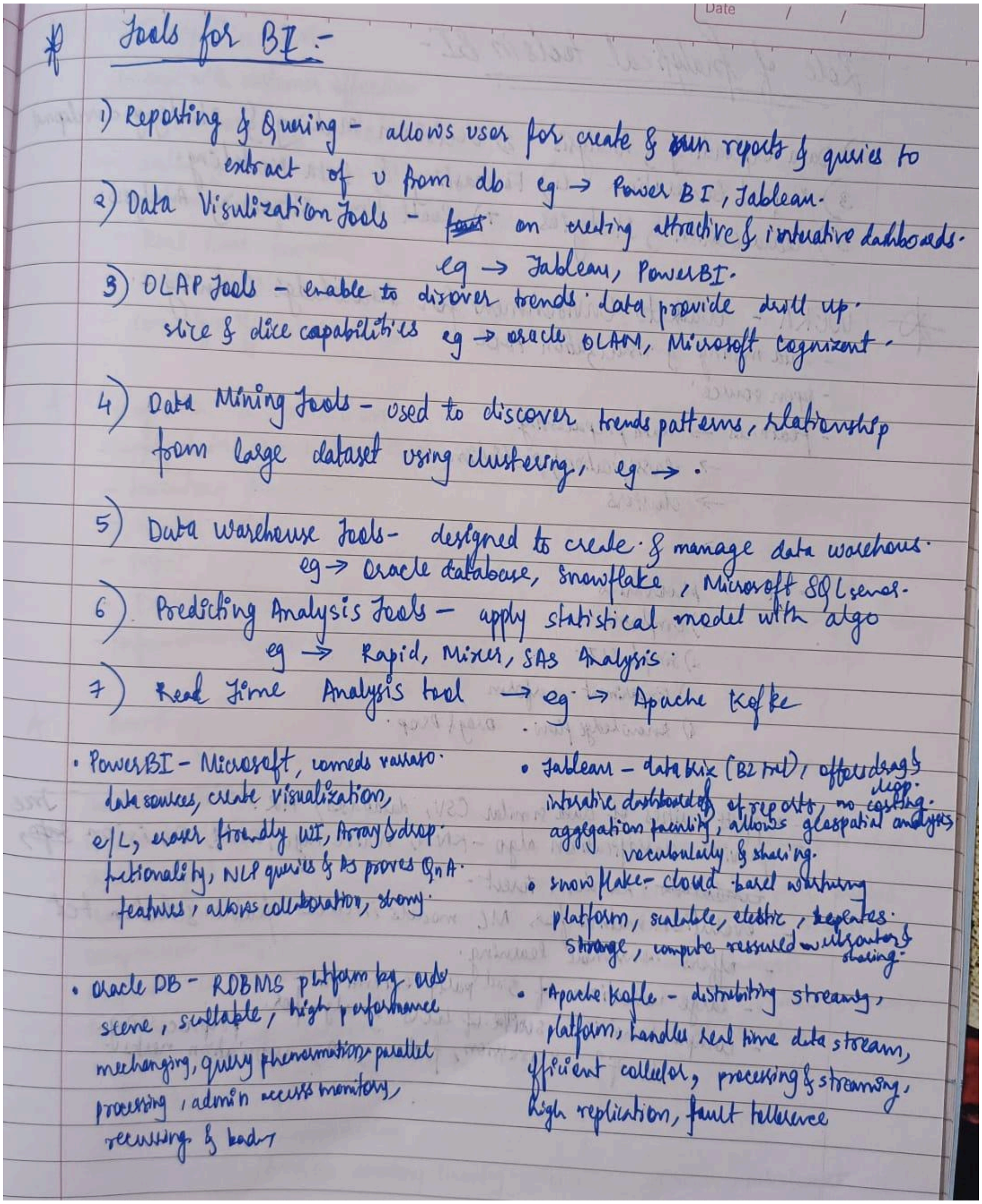
6. Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

Concept: ROC AUC is a metric that considers all possible classification thresholds. It plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis. The AUC represents the area under the ROC curve, with a higher AUC indicating better model performance.

Interpretation: ROC AUC is useful for comparing models, especially when dealing with imbalanced datasets or when the class distribution might change over time. A perfect model would have an AUC of 1, while a random model would have an AUC of 0.5.

Q6) C) CLUSTERING (NUMERICAL SKIPPED)

Q7) A) State and explain different Tools for Business Intelligence



Date / /

* Tools for BI :-

1) Reporting & Quering - allows user for create & ~~run~~ reports & quries to extract of v from db eg → Power BI, Jablean.

2) Data Visulization Tools - ~~focus~~ on creating attractive & intuative dashboards. eg → Jablean, PowerBI.

3) OLAP Tools - enable to disover trends data provide drill up. slice & dice capabilities eg → oracle OLAM, Microsoft cognizent.

4) Data Mining Tools - used to discover trends patterns, relationship from large dataset using clustering, eg → .

5) Data Warehouse Tools - designed to create & manage data warehous. eg → Oracle database, Snowflake, Microsoft SQL server.

6) Predicting Analysis Tools - apply statistical model with algo eg → Rapid, Mixer, SAS Analysis.

7) Real Time Analysis tool → eg → Apache Kafke

- PowerBI - Microsoft, conneds various data sources, create visualization, e/L, easier friendly UI, Array & drop fuctionality, NLP quries & AI proves QnA features, allows collaboration, strong.

- Oracle DB - RDBMS platform by orde scene, scalable, high performance mechanging, query phenomation parallel processing, admin access monitoring, recovering & backy

- Jablean - data blix (B2 tool), offers drag & drop. interative dashboard of reports, no costing. aggegation facility, allows geospatial analysis vocubularity & sharing.

- snowflake - cloud based warehousing platform, scalable, elastic, seperates storage, compute ressurced with security & sharing.

- Apache Kafle - distributing streaming, platform, handles real time data stream, efficient collector, processing & streaming, high replication, fault tollerence

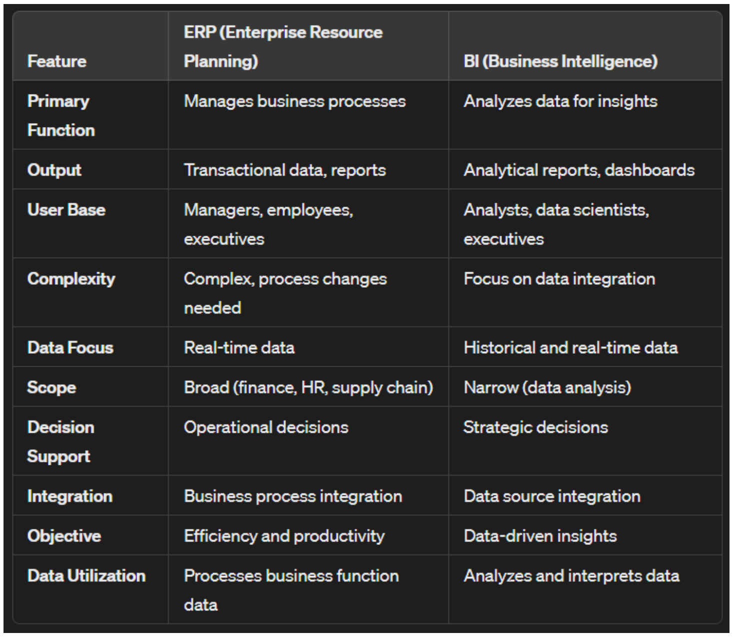
Q7) B) State and Elaborate similarities & differences in ERP and Business Intelligence.

Similarities

1. **Data Utilization**:
   - Both systems rely on data collection, integration, and analysis.
2. **Integration**:
   - Both integrate data from multiple sources to provide a comprehensive view of the business.
3. **Decision Support**:
   - Both aid in decision-making processes (ERP for operational, BI for strategic).

Differences

1. **Primary Function**:
   - **ERP**: Manages and streamlines core business processes.
   - **BI**: Analyzes data to generate insights and support strategic decisions.
2. **Output**:
   - **ERP**: Produces transactional data and operational reports.
   - **BI**: Generates analytical reports, dashboards, and data visualizations.
3. **User Base**:
   - **ERP**: Used by operational managers, employees, and executives.
   - **BI**: Used by analysts, data scientists, and top executives.
4. **Complexity and Implementation**:
   - **ERP**: More complex, involving significant changes to business processes.
   - **BI**: Focuses on data integration and analytics.
5. **Real-Time vs. Historical Data**:
   - **ERP**: Emphasizes real-time data.
   - **BI**: Primarily deals with historical data but can include real-time analysis.
6. **Scope**:
   - **ERP**: Broad, covering various business functions.
   - **BI**: Narrower, focusing on data analysis and insights.



Feature	ERP (Enterprise Resource Planning)	BI (Business Intelligence)
Primary Function	Manages business processes	Analyzes data for insights
Output	Transactional data, reports	Analytical reports, dashboards
User Base	Managers, employees, executives	Analysts, data scientists, executives
Complexity	Complex, process changes needed	Focus on data integration
Data Focus	Real-time data	Historical and real-time data
Scope	Broad (finance, HR, supply chain)	Narrow (data analysis)
Decision Support	Operational decisions	Strategic decisions
Integration	Business process integration	Data source integration
Objective	Efficiency and productivity	Data-driven insights
Data Utilization	Processes business function data	Analyzes and interprets data

Q7) C) Write a note on: BI Applications in CRM.

Customer Segmentation: BI tools divide customers based on demographics and behavior for targeted marketing.

Sales and Marketing Analytics: Analyze sales data to refine marketing strategies and optimize sales processes.

Customer Retention: Predict and prevent churn through analysis of customer interactions.

Personalization: Customize product recommendations and promotions based on customer preferences.

Customer Lifetime Value: Calculate CLV to prioritize high-value customers and allocate resources effectively.

Feedback Analysis: Gauge customer sentiment from various channels to address concerns promptly.

Cross-Selling and Upselling: Identify opportunities for additional sales based on purchase patterns.

Operational Efficiency: Automate data handling to focus on strategic initiatives and customer engagement.

Predictive Analytics: Forecast customer behavior and market trends to prepare for future demands.

Performance Monitoring: Real-time dashboards track CRM KPIs for quick decision-making and issue resolution.

Q8) A) State the role of Data Analytics in any business with example

The role of data analytics in business is to transform raw data into actionable insights, aiding decision-making processes, optimizing operations, and gaining competitive advantages

Customer Insights:

- Analyzing customer data to understand preferences and behavior for targeted marketing.

Inventory Optimization:

- Predicting demand and optimizing inventory levels to minimize stockouts and excess inventory.

Price Optimization:

- Setting optimal pricing strategies based on competitor pricing and demand elasticity.

Customer Experience Enhancement:

- Improving the shopping experience by analyzing feedback and sentiment analysis from various channels.

Fraud Detection:

- Identifying and preventing fraudulent transactions by analyzing transactional data for anomalies.

Q8) B) Comment “How might you implement business intelligence findings within an organization?”

Implementing business intelligence (BI) findings within an organization involves several key steps to ensure that insights are effectively translated into action.

1. **Strategic Alignment**: Ensure that BI findings align with the organization's strategic goals and objectives. Identify areas where BI insights can contribute to improving performance, reducing costs, or driving innovation.
2. **Stakeholder Engagement**: Engage key stakeholders across different departments, including executives, managers, and frontline employees. Communicate the value of BI findings and involve stakeholders in the decision-making process.
3. **Actionable Insights**: Translate BI findings into actionable insights that can be easily understood and implemented by relevant teams. Focus on providing specific recommendations and solutions rather than presenting raw data.
4. **Training and Education**: Provide training and education to employees on how to interpret and use BI insights effectively. Equip teams with the necessary skills and tools to leverage BI findings in their day-to-day activities.
5. **Integration with Business Processes**: Integrate BI findings into existing business processes and workflows to ensure seamless implementation. Embed BI dashboards and reports into relevant systems and applications used by employees.
6. **Continuous Monitoring and Improvement**: Establish mechanisms for continuous monitoring and improvement of BI initiatives. Track key performance indicators (KPIs) to measure the impact of BI findings and identify areas for optimization.
7. **Cultural Change**: Foster a data-driven culture within the organization by promoting transparency, accountability, and collaboration around BI initiatives. Encourage employees to embrace data-driven decision-making and incorporate BI insights into their decision-making processes.
8. **Feedback Loop**: Establish a feedback loop to gather input from users and stakeholders on the effectiveness of BI implementations. Use feedback to iterate and refine BI strategies over time, ensuring ongoing relevance and value.

Q8) C) Write a note on: BI Applications in Logistics.

Business Intelligence (BI) applications play a crucial role in optimizing logistics operations by providing valuable insights into various aspects of the supply chain. From inventory management to transportation optimization, BI empowers logistics companies to make data-driven decisions, enhance efficiency, and improve customer satisfaction.

Demand Forecasting:

- Analyze historical sales data and market trends for accurate demand prediction.

Inventory Management:

- Monitor inventory levels and turnover rates in real-time for optimization.

Route Optimization:

- Analyze transportation data to optimize routes and minimize costs.

Warehouse Management:

- Track warehouse operations and KPIs to identify bottlenecks and improve productivity.

Supplier Performance Monitoring:

- Assess supplier performance for on-time delivery and quality assurance.

Customer Service Optimization:

- Analyze customer data to enhance service levels and satisfaction.

Risk Management:

- Identify and mitigate supply chain risks through data analysis.

Performance Monitoring and Reporting:

- Provide real-time dashboards and reports for performance tracking and decision-making.