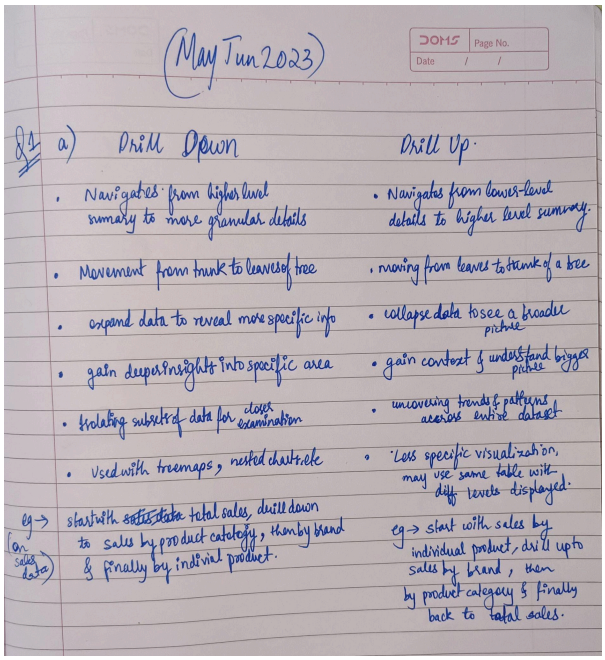
Q1) A)Explain in detail drill up & drill down.



(May Jun 2023)

Q1 a)

Drill Down	Drill Up
• Navigates from higher level summary 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 → start with total sales, drill down to sales by product catology, then by brand & finally by individual product. (on sales data)	eg → start with sales by individual product, drill up to sales by brand, then by product category & finally back to total sales.

Q1) B) Explain multidimensional data model with example

A multidimensional data model is a way of organizing data for analytical purposes, particularly in data warehouses. It allows you to view and analyze data from multiple perspectives, also known as dimensions.

Advantages:

- **Improved Data Analysis:** Multidimensional models allow for faster and more efficient analysis of complex data. Users can easily slice and dice data based on different dimensions (e.g., product category, time period, location) to gain deeper insights.
- **Simplified Queries:** These models enable users to ask complex questions without writing intricate SQL queries. They can simply select the desired dimensions and facts using tools designed for multidimensional analysis.
- **OLAP Support:** Multidimensional models are optimized for Online Analytical Processing (OLAP) tools, which are specifically designed for tasks like drill-down (focusing on details within a dimension), roll-up (summarizing data across a dimension), pivoting (changing the perspective of analysis), and slicing (isolating a specific subset of data).
- **Faster Performance:** Compared to traditional relational databases, multidimensional models can offer faster data retrieval and analysis, particularly for complex queries involving multiple dimensions.
- **Data Integrity:** The structure of multidimensional models enforces data integrity by ensuring consistency between facts and their corresponding dimensions.

Disadvantages:

- **Complexity When Modeling:** Designing and implementing a multidimensional data model can be more complex than a simple relational model. It requires careful planning to define relevant dimensions and facts, ensuring proper relationships between them.
- **Limited Flexibility:** While multidimensional models excel at predefined analysis based on dimensions, they may not be as flexible for exploring entirely new data relationships that weren't anticipated during the model design phase.
- **Scalability Challenges:** Scaling a multidimensional model to accommodate very large datasets can be challenging. Adding new dimensions or facts can require significant restructuring of the model.
- **Data Redundancy:** Dimension tables can sometimes contain redundant data, especially if certain attributes are shared across multiple dimensions. This redundancy can increase storage requirements.
- **Not Ideal for All Data:** Multidimensional models are best suited for analytical data with well-defined dimensions and facts. They may not be the best choice for storing and managing transactional data or data that doesn't naturally fit into a dimensional structure.

Example:

Imagine you run an online store and want to analyze your sales data. Here's how a multidimensional data model could look:

Dimensions:

Time: Year, Quarter, Month, Day
Product: Product Category (e.g., clothing, electronics), Brand, Product ID
Customer: Customer ID, Demographics (age, gender)
Location: Country, City, Store
Facts:

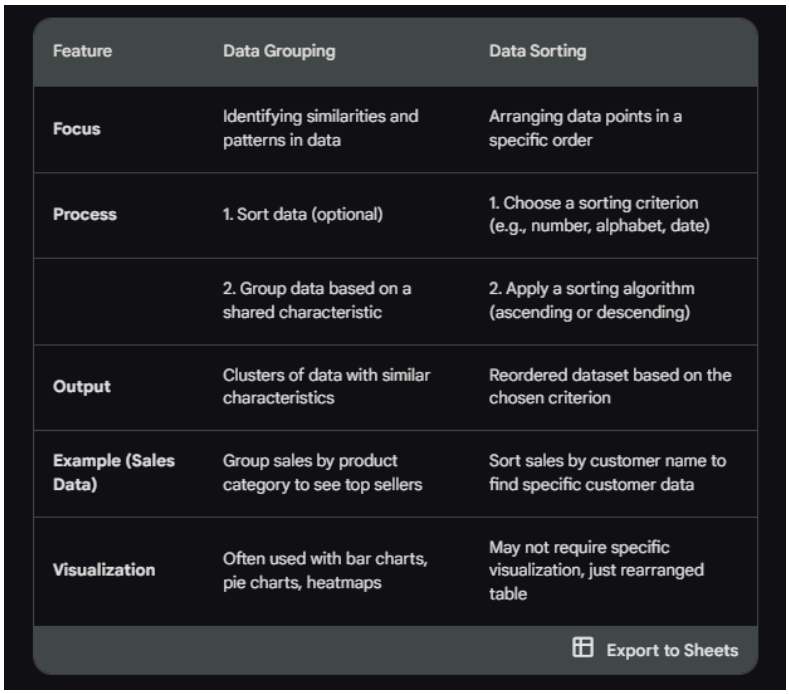
Sales Amount
Number of Units Sold
Average Order Value
Fact Table:

This table would store the sales data for each transaction, including the following columns:

Transaction ID (primary key)
Foreign key to Time dimension table (linking to specific date)
Foreign key to Product dimension table (linking to specific product)
Foreign key to Customer dimension table (linking to specific customer)
Foreign key to Location dimension table (linking to location of purchase)
Sales Amount
Number of Units Sold
Dimension Tables:

Time: This table would hold details about each date, such as year, quarter, month, day of week, etc.
Product: This table would include details about each product, such as category, brand, product name, price, etc.
Customer: This table would hold customer information, such as customer ID, name, demographics (age, gender, location), etc.
Location: This table would include details about each store or location, such as country, city, store ID, etc.

Q1) C) What is data grouping and sorting? Write example of each.



Feature	Data Grouping	Data Sorting
Focus	Identifying similarities and patterns in data	Arranging data points in a specific order
Process	1. Sort data (optional)	1. Choose a sorting criterion (e.g., number, alphabet, date)
	2. Group data based on a shared characteristic	2. Apply a sorting algorithm (ascending or descending)
Output	Clusters of data with similar characteristics	Reordered dataset based on the chosen criterion
Example (Sales Data)	Group sales by product category to see top sellers	Sort sales by customer name to find specific customer data
Visualization	Often used with bar charts, pie charts, heatmaps	May not require specific visualization, just rearranged table

Export to Sheets

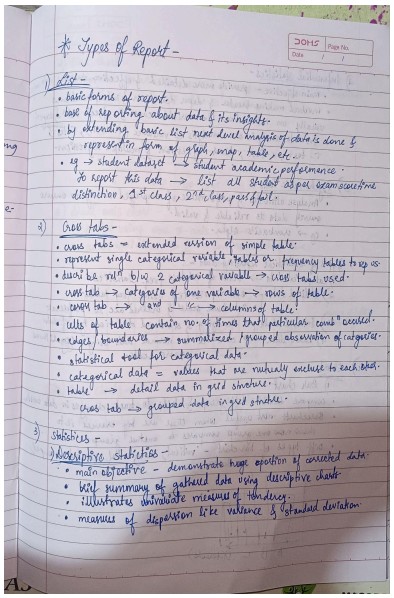
Data Grouping Example:

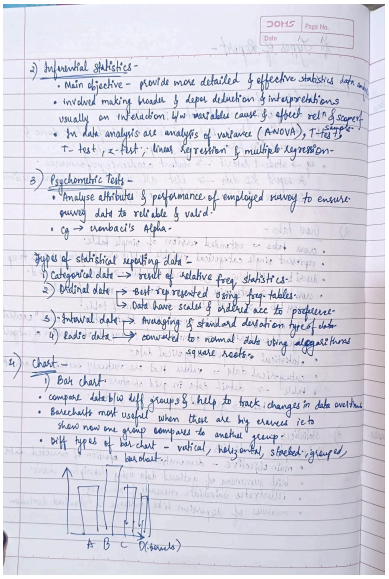
- Goal: Analyze customer purchase behavior by product category.
- Process:
    1. Sort the data by Product Category (optional but can improve efficiency).
    2. Group the data by Product Category. This will create groups of orders for each product category (e.g., Electronics, Clothing, Home Goods).
- Output: You can then calculate various metrics for each group, such as total sales, average order value, and number of orders placed. This analysis might reveal which product categories are the most popular, which categories have higher order values, or if there are any seasonal trends in purchases.

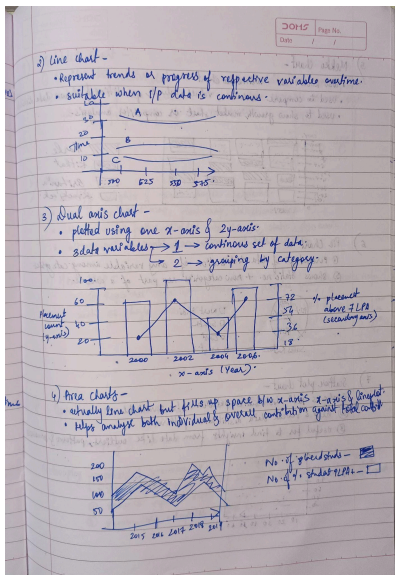
Data Sorting Example:

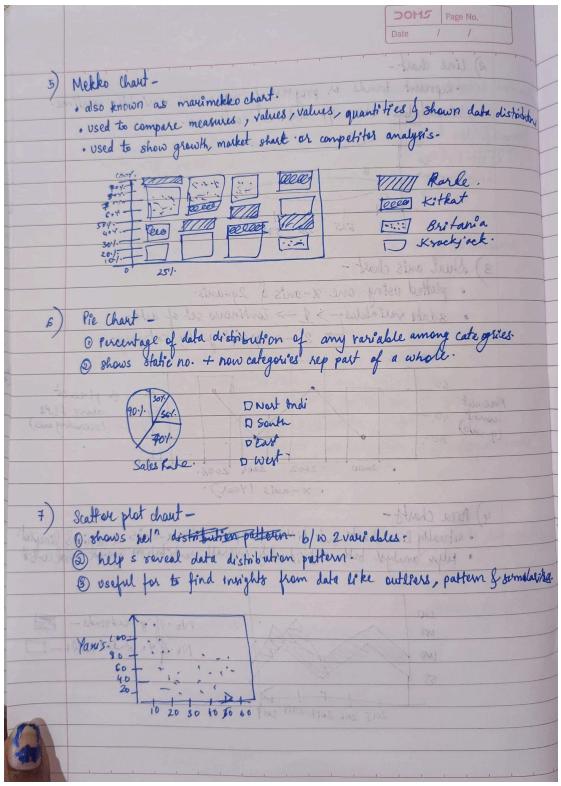
- Goal: Identify the customers who spent the most money in the last month.
- Process:
    1. Filter the data to include orders from the last month (based on Order Date).
    2. Sort the filtered data by Order Amount in descending order (highest to lowest).
- Output: This will show you the orders with the highest order amounts at the top of the table. You can then identify the corresponding customer names to see who your top spenders were in the last month. This information can be valuable for targeted marketing campaigns or loyalty programs.

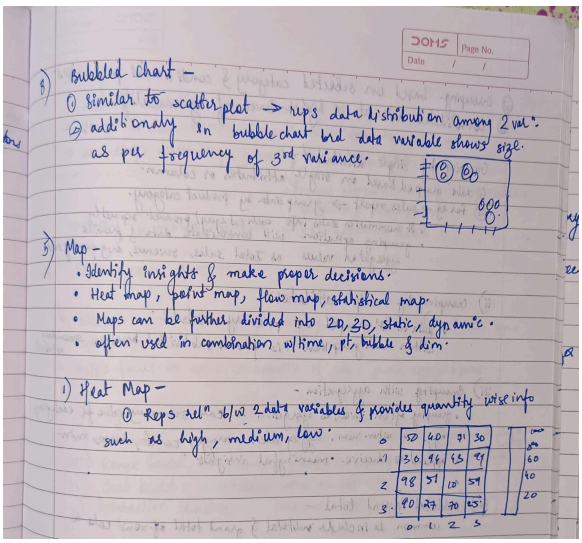
Q2) A) Explain different types of reports in detail?











Q2) B) Explain relational data Model with example.

A relational data model is a way of organizing data in tables (also called relations) with a focus on establishing relationships between them. It's the most widely used data model for storing and managing structured data in databases

Advantages of Relational Data Models:

- **Simplicity:** The basic structure of tables and relationships is easy to understand and manage.
- Flexibility: You can add new tables or modify existing ones to accommodate new data or changing requirements.
- **Data Integrity:** The concept of primary and foreign keys helps ensure data consistency and prevent duplicate entries.
- **Standardization:** SQL (Structured Query Language) is a widely used standard for querying and manipulating data in relational databases.
- **Maturity:** Relational databases are well-established with robust technology and a wide range of tools available.

Disadvantages of Relational Data Models:

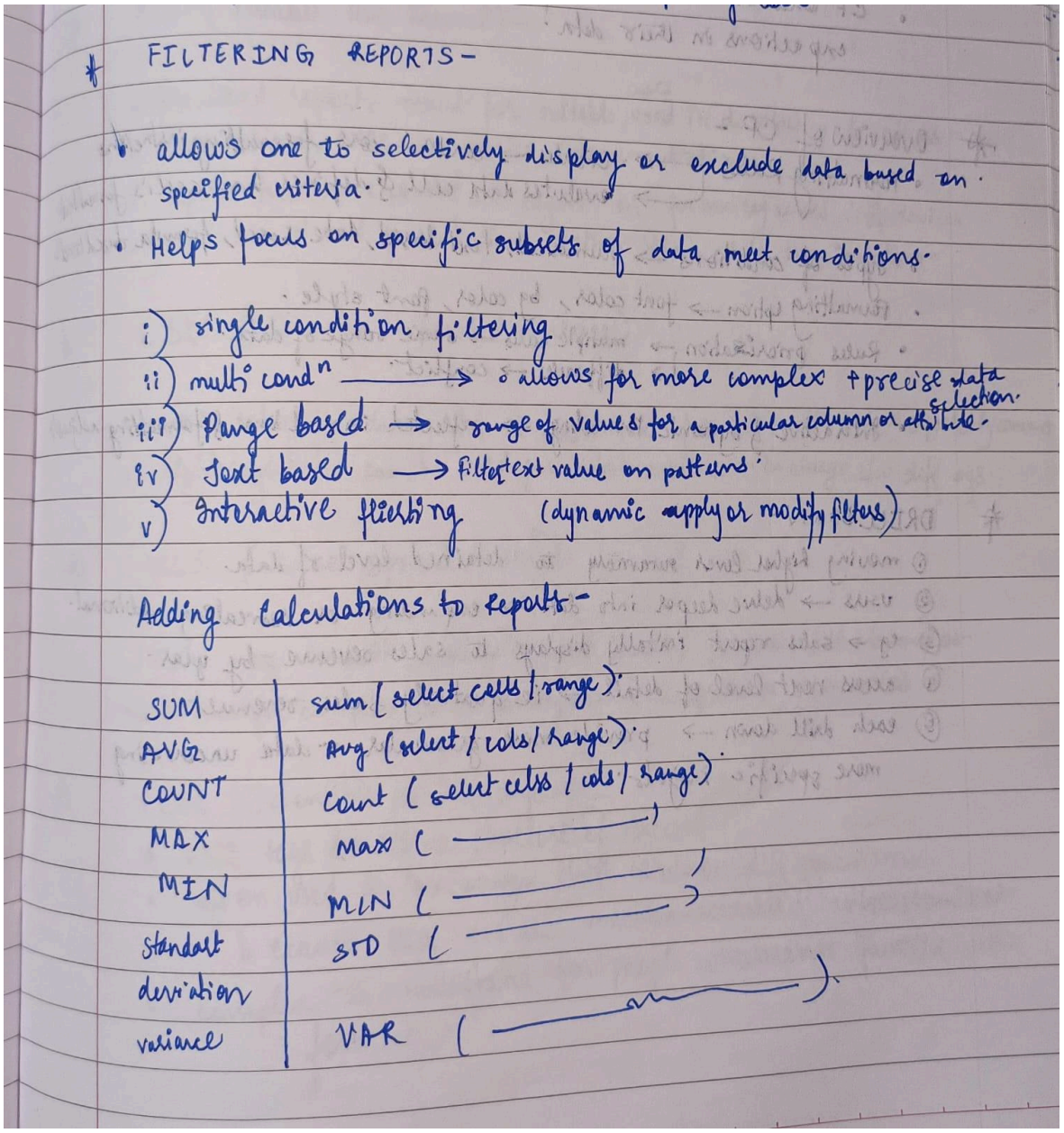
- **Normalization Complexity:** Designing an efficient and well-normalized relational schema can be complex, especially for large and intricate datasets.
- **Scalability Limitations**: Relational databases may face challenges scaling to extremely large datasets, and performance can degrade as data volume increases.
- **Limited Flexibility for Complex Data:** While flexible, relational models might not be ideal for storing and managing highly complex data structures or data with inherently non-tabular relationships.
- **Joins for Complex Queries:** Joining multiple tables to answer complex questions can lead to intricate SQL queries, which can be challenging to write and optimize.

Example:

Consider an online store database. We can have two tables:

Customers: This table might have columns like Customer ID (primary key), Name, Email, Address, etc.
Orders: This table might have columns like Order ID (primary key), Customer ID (foreign key referencing Customers.CustomerID), Order Date, Order Amount, etc.
The Customer ID in the Orders table connects it to the Customers table, allowing you to see which customer placed a specific order.

Q2) C) Write short note on filtering reports.



FILTERING REPORTS -

- allows one to selectively display or exclude data based on specified criteria.
- Helps focus on specific subsets of data meet conditions.

i) single condition filtering
ii) multi cond^n → • allows for more complex & precise data selection.
iii) Range based → range of values for a particular column or attribute.
iv) Text based → Filter text value on patterns.
v) Interactive filtering (dynamic apply or modify filters)

Adding Calculations to Reports -

SUM	sum (select cells / range).
AVG	Avg (select / cols / range).
COUNT	count (select cells / cols / range).
MAX	Max (———)
MIN	MIN (———)
Standard deviation	STD (———)
variance	VAR (———).

Q3) A) Explain data exploration in detail with example

Example: Exploring Customer Purchase Data

Imagine you have a dataset containing customer purchase information for an online store. You might use data exploration techniques like:

- Data Profiling: Analyze data types (e.g., customer ID, product category, purchase amount, date of purchase). Check for missing values and calculate descriptive statistics for numerical variables like purchase amount.
- Visualization: Create histograms to see the distribution of purchase amounts and identify potential outliers (very high or low purchases). Use scatter plots to explore relationships between purchase amount and factors like customer age or location.
- Statistical Analysis: Calculate average purchase amount and correlations between purchase amount and other variables (e.g., customer age, product category).
- Data Subsetting: Focus on specific customer segments (e.g., first-time buyers, high spenders) to understand their purchasing behavior in more detail.

Q3) B) Explain data transformation in detail with example.

Data transformation is the process of converting raw data into a format that's more suitable for analysis.

Advantages:

- **Improved Data Quality:** Transformation helps clean and correct errors, inconsistencies, and missing values, leading to more reliable data for analysis.
- **Enhanced Analysis Efficiency:** Data in a suitable format allows for smoother and faster analysis using specific tools.
- **Standardized Data:** Transformation ensures consistency across datasets, enabling easier comparison, combination, and merging of information.
- **Meaningful Insights Extraction:** By manipulating data strategically, you can reveal hidden patterns, trends, and relationships that might not be evident in raw data.
- **Flexibility for Different Analyses:** Transformed data can be adapted to fit the requirements of various analytical techniques, like machine learning or statistical modeling.

Disadvantages:

- **Increased Processing Time:** Extensive data transformation can be time-consuming, especially for large datasets.
- **Potential for Errors:** Introducing transformations can introduce errors if not done carefully or documented properly.
- **Data Loss or Inaccuracy:** In some cases, transformation steps might lead to accidental data loss or introduce inaccuracies.
- **Complexity for New Users:** Understanding and implementing data transformation techniques can have a learning curve for beginners.

- **Potential Bias:** Transformations can introduce bias if not done objectively or with a clear understanding of the data and analysis goals.

Example: Analyzing Weather Data

Imagine you have a dataset containing weather data for a city, including:

- Date (various formats)
- Time (various formats)
- Temperature (Celsius)
- Humidity (percentage symbol)
- Precipitation (text values like "rainy", "sunny")

Here's how data transformation can be beneficial:

- **Cleaning:** Standardize date and time formats (e.g., YYYY-MM-DD, HH:MM). Identify and handle any missing values.
- **Formatting:** Convert temperatures to a common unit (e.g., Fahrenheit).
- **Derivation:** Calculate a new variable "Feels Like Temperature" based on temperature and humidity.
- **Aggregation:** Group data by month and calculate average temperature and total precipitation.
- **Filtering:** Focus on data for the summer months to analyze seasonal weather patterns.

Q3) C) 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) a) Explain data reduction in detail with example.

Data reduction is a crucial technique in data analysis that aims to condense a large dataset while preserving its essential characteristics. Imagine it like summarizing a long book into its key points – you retain the core information in a more manageable format. Here's a breakdown of data reduction:

Advantages (ADV):

- **Improved Processing Efficiency:** Working with a smaller dataset reduces processing time and computational resources required for analysis.
- **Enhanced Visualization:** Smaller datasets are often easier to visualize effectively, leading to clearer communication of insights.

- **Reduced Storage Requirements:** Less data translates to less storage space needed, which can be a significant benefit for large datasets.
- **Noise Reduction:** Data reduction techniques can help identify and remove irrelevant information or noise, potentially leading to more accurate analysis.
- **Improved Model Performance:** In machine learning, reducing data dimensionality can sometimes improve the performance of models by reducing complexity and overfitting.

Disadvantages (DIS):

- **Information Loss:** The core principle of data reduction involves discarding some data, which can lead to a loss of information if not done carefully.
- **Potential Bias:** The techniques used for reduction can introduce bias if not chosen or implemented appropriately for the specific data and analysis goals.
- **Limited Generalizability:** Insights gained from a reduced dataset might not be generalizable to the entire population represented by the original data.
- **Complexity for New Users:** Understanding and choosing the right data reduction technique can be challenging for beginners in data analysis.

Common Data Reduction Methods:

- **Dimensionality Reduction:** This method focuses on reducing the number of variables in a dataset while retaining the most significant information. Techniques include Principal Component Analysis (PCA) and feature selection.
- **Sampling:** This method involves selecting a smaller subset of data points from the original dataset that is statistically representative of the whole. Sampling techniques include random sampling, stratified sampling, and cluster sampling.
- **Aggregation:** This method involves summarizing data points based on certain criteria, often resulting in a smaller dataset with fewer data points but capturing overall trends. Techniques include averaging, summing, and counting.
- **Data Compression:** This method uses techniques to represent the data in a more compact format without significant loss of information. This is often used for storing or transmitting large datasets.

Example: Analyzing Customer Purchase Data:

Imagine you have a dataset containing detailed purchase information for all customers of an online store, including:

- Customer ID
- Product ID
- Product Name
- Purchase Date
- Purchase Amount
- Customer Demographics (Age, Location, etc.)

This dataset might be very large and complex. Here's how data reduction could be applied:

- **Dimensionality Reduction:** PCA could be used to identify the most important product features that influence purchase behavior, reducing the number of variables needed for analysis.

- **Sampling:** You could randomly sample a representative subset of customers for a specific timeframe, allowing you to analyze recent purchase trends without processing the entire dataset.
- **Aggregation:** You could aggregate purchase data by product category and month, providing insights into overall sales trends across different product categories.

Q4) B) Difference between univariate, Bivariate, Multivariate analysis.

Feature	Univariate Analysis	Bivariate Analysis	Multivariate Analysis
Focus	Single variable	Two variables	More than two variables
Objective	Describe the distribution and characteristics of a single variable	Explore the relationship between two variables	Identify patterns and relationships among multiple variables
Examples	Analyzing average income in a city, distribution of exam scores	Studying the correlation between age and income, price and sales	Exploring the relationship between income, age, education level, and credit score on loan approval
Techniques	Mean, median, mode, standard deviation, frequency distribution tables, histograms	Scatter plots, correlation coefficients	Regression analysis, factor analysis, cluster analysis
Benefits	Simple to understand and implement, provides basic insights into the data	Reveals potential relationships between variables	Provides a more comprehensive understanding of complex relationships among multiple variables
Limitations	Ignores relationships between variables, limited insights	Limited to understanding the connection between two variables	Can be complex to interpret, requires careful consideration of confounding factors

Q4) C) Write a short note on data discretization.

Data discretization, the process of converting continuous data into categories (bins), offers benefits and drawbacks in data analysis. Here's a breakdown:

Advantages (ADV):

- **Improved Analysis Efficiency:** Discretized data can be more efficient to analyze with certain algorithms, especially those designed for categorical data. For instance, decision trees perform well with discrete features.
- **Reduced Data Complexity:** By grouping similar values, discretization simplifies complex datasets and makes them easier to understand and visualize. Imagine a histogram with fewer, broader bars instead of many detailed ones.
- **Prepares for Modeling:** Some machine learning models work better with discrete data. Discretization helps prepare continuous data for such models. For example, a logistic regression model might require categorical inputs.

- **Identifies Patterns:** Discretization can sometimes reveal hidden patterns or trends within the data that might not be readily apparent in its continuous form. Grouping similar values can highlight underlying relationships.

Disadvantages (DIS):

- **Information Loss:** Discretization inevitably leads to some loss of information about the original data. The finer details within each bin are lost.
- **Choosing the Right Method:** Selecting the best discretization method depends on the data and analysis goals. A poor choice can distort the data and lead to misleading results.
- **Impact on Analysis:** Discretized data might require different analysis techniques compared to continuous data. Statistical tests suitable for continuous data might not be applicable anymore.

Discretization Methods:

- **Equal-Width Binning:** This is a simple method that divides the data range into bins of equal width. It's easy to implement but might not capture the underlying distribution of the data well, especially if the data is not uniformly spread.
- **Equal-Frequency Binning:** This method creates bins containing roughly the same number of data points. It can be more representative of the data distribution but might result in bins of unequal width, which can be less intuitive to interpret.
- **K-Means Discretization:** This method uses a clustering algorithm (k-means) to group similar data points into bins. It can be effective but requires specifying the desired number of bins (k). The choice of k can significantly impact the resulting categories.

Example: Analyzing Customer Age

Imagine you have a dataset containing customer purchase information, including customer age (continuous numerical data). You might want to discretize age for analysis:

- **Original Data:** Age (years) - continuous values ranging from 18 to 80.
- **Equal-Width Binning (3 bins):**
    - Bin 1: 18-33 years old (young adults)
    - Bin 2: 34-50 years old (middle-aged adults)
    - Bin 3: 51-80 years old (seniors)
- **Equal-Frequency Binning (3 bins):**
    - Bin 1: 18-28 years old
    - Bin 2: 29-45 years old
    - Bin 3: 46-80 years old

Discretization allows you to analyze purchase behavior based on age groups, potentially revealing trends (e.g., higher spending in a specific age group). However, information about the exact age within each bin is lost.

Q5) A) What is the association rule mining? Explain the terms support, confidence, lift.

Association rule mining is a technique used in data analysis to discover interesting relationships, or associations, between different items or variables within large datasets.

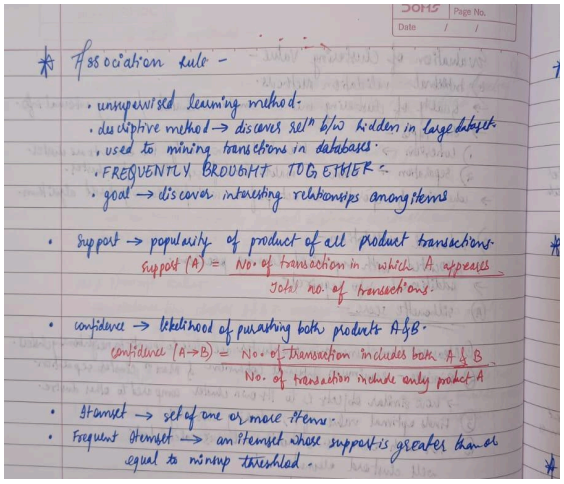
- **Uncovers hidden relationships:** Discovers frequent co-occurrences of items in large datasets.
- **Example:** Identifies products customers often buy together (bread & milk).

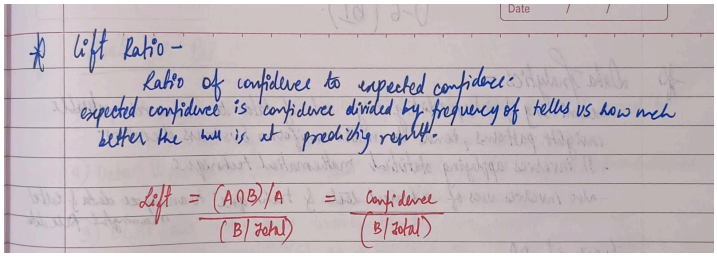
Key Metrics:

- **Support:** How often an itemset (group of items) appears together (e.g., 10% of transactions have bread & milk).
- **Confidence:** Likelihood of finding item B given you already have item A (e.g., if bread is bought, how likely is milk?).
- **Lift:** Strength of the association between items. Considers if the co-occurrence is by chance.
    - Lift > 1: Positive association (finding B with A is more likely than random).
    - Lift = 1: No association (items appear together by chance).
    - Lift < 1: Negative association (finding B with A is less likely than random).

Benefits:

- Understand customer behavior (e.g., recommend products based on past purchases).
- Improve marketing strategies (e.g., promote complementary products together).
- Gain insights from large datasets.



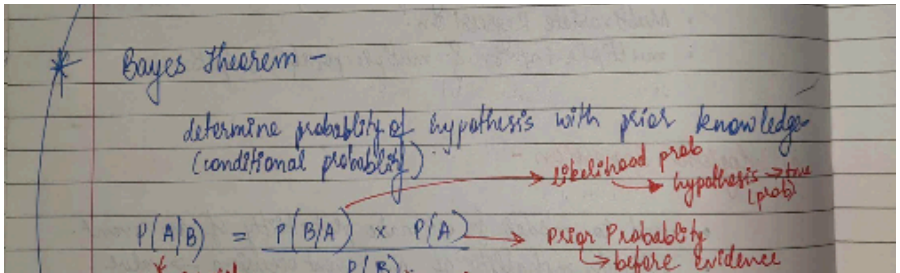


Q5) B) What is the difference between hierarchical clustering and partitioning method?

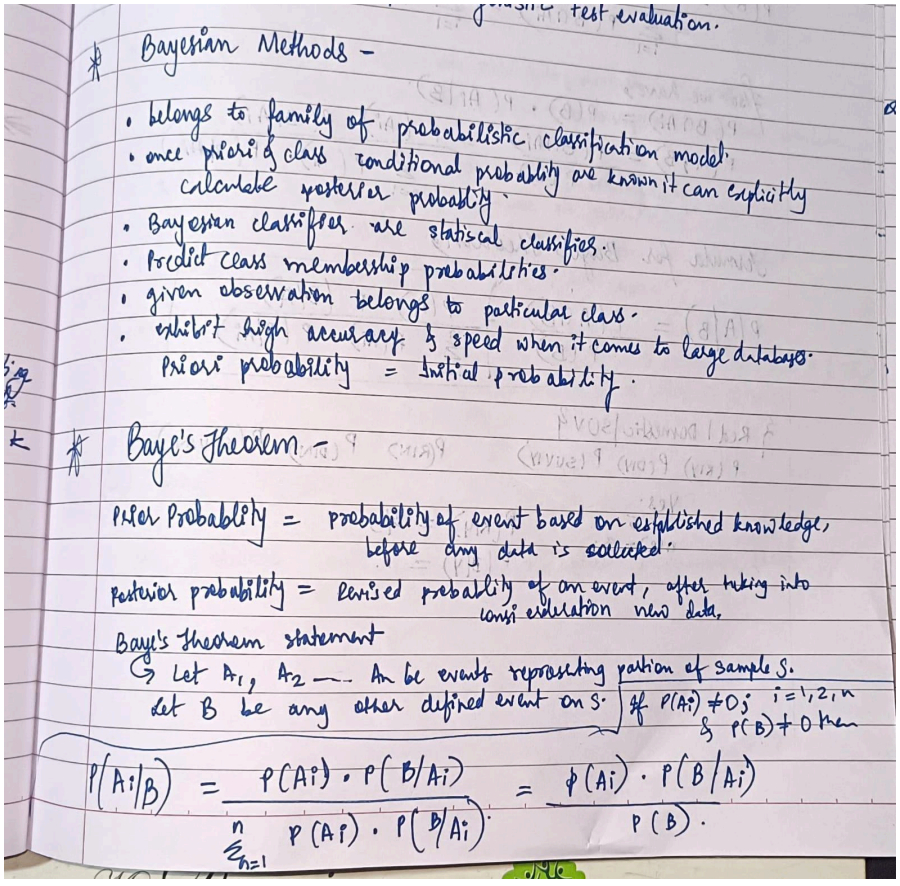
Feature	Hierarchical Clustering	Partitioning Clustering
Type of Clusters	Hierarchical	Partitional
Cluster Structure	Creates a hierarchy of clusters (tree-like structure)	Divides data points into **non-overlapping** clusters
Number of Clusters	**No pre-defined number** needed (explore through stopping criteria)	Requires specifying the **desired number** upfront
Algorithm	Merges or splits clusters based on distances	Assigns data points to clusters based on similarity (one iteration)
Visualization	Dendrogram (tree diagram)	Scatter plots with color-coded clusters
Examples	Agglomerative, Divisive	K-Means, K-Medoids
Advantages	* No need to pre-define cluster count	* Reveals natural groupings
Disadvantages	* Computationally expensive (large datasets)	* Subjective dendrogram interpretation
Advantages	* Faster and more efficient (large datasets)	* Easier to implement and understand
Disadvantages	* Pre-defined cluster count required (challenging)	* Sensitive to initial assignments (local optima)

Q6)

A) Explain Bayes



theorem in detail.



test evaluation.

* Bayesian Methods -

- belongs to family of probabilistic classification model.
- once priori & class conditional probability are known it can explicitly calculate posterior probability.
- Bayesian classifiers are statistical classifies.
- Predict class membership probabilities.
- given observation belongs to particular class.
- exhibit high accuracy & speed when it comes to large databases.

Priori probability = Initial probability.

* Baye's Theorem -

Prior Probability = probability of event based on established knowledge, before any data is collected.

Posterior probability = Revised probability of an event, after taking into consideration new data.

Baye's theorem statement

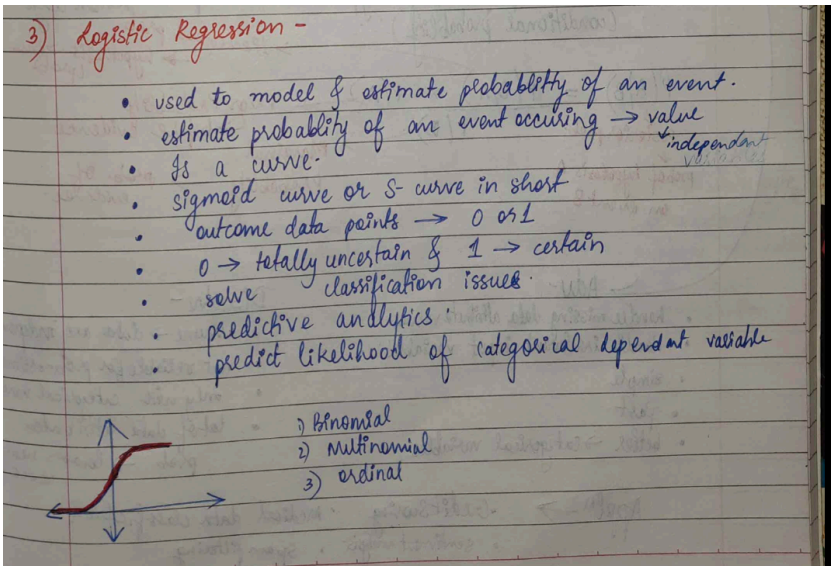
↳ Let $A_1, A_2 \ldots A_n$ be events representing portion of sample S.
Let B be any other defined event on S. If $P(A_i) \neq 0$; $i = 1, 2, n$ & $P(B) \neq 0$ then

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{n=1}^{n} P(A_i) \cdot P(B/A_i)} = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

Q6) B) Different between classification and clustering.

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

Q6) C) Explain logistic regression with example.



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.

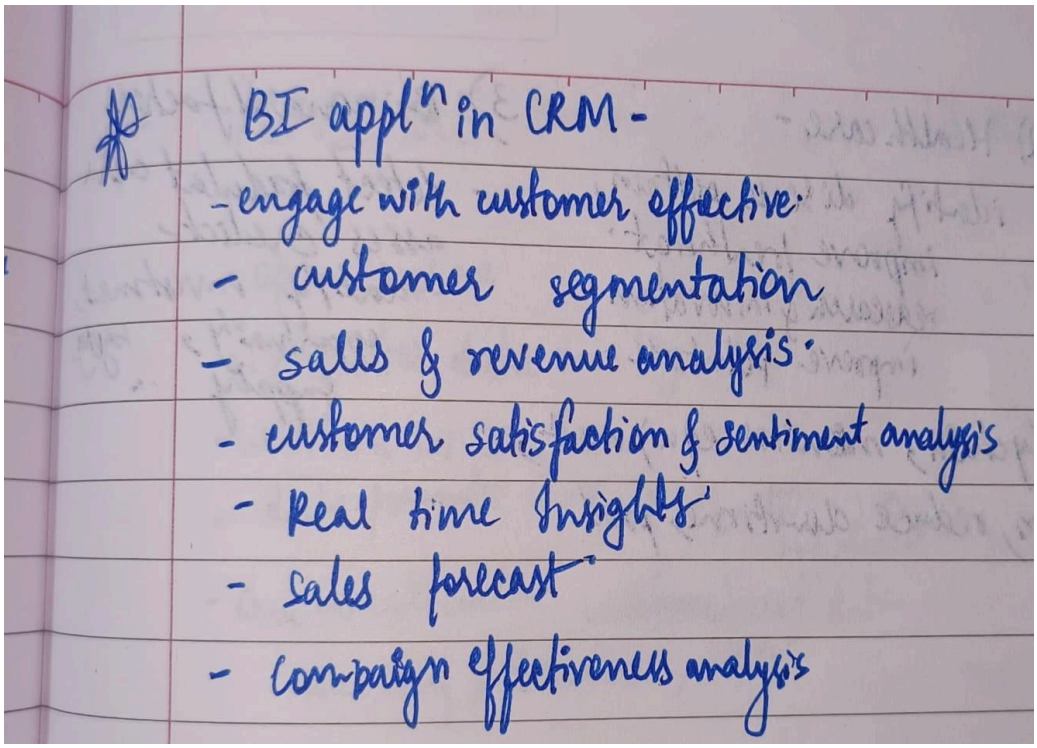
Method:

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.

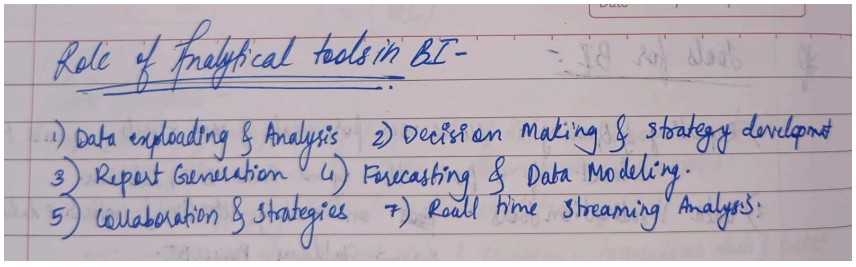
Q7) A) Explain BI application in CRM.



Q7) B) Explain roles of Analytical tools in BI

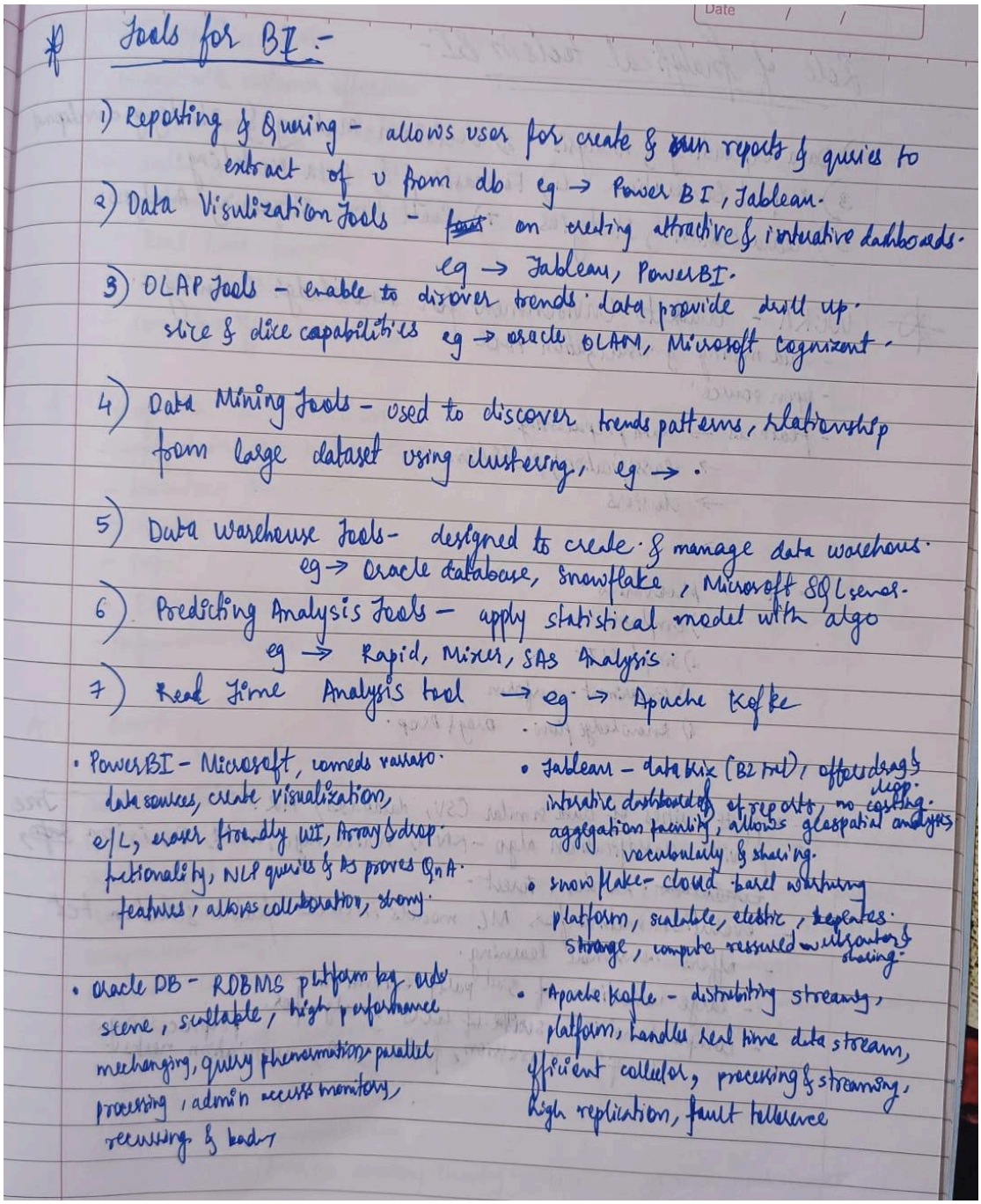
Analytical tools are the workhorses of Business Intelligence (BI). They act as the bridge between raw data and actionable insights, playing several key roles:

- **Data Transformation & Cleaning:** Analytical tools can clean and pre-process messy data, handling inconsistencies and errors to ensure accurate analysis.
- **Data Exploration & Analysis:** These tools allow users to explore data from various angles, identify patterns, trends, and relationships through statistical analysis and calculations.
- **Data Visualization:** Analytical tools create charts, graphs, and dashboards that present complex data in an easy-to-understand visual format, making insights readily accessible.
- **Reporting & Communication:** They help generate reports that summarize findings and communicate insights to stakeholders across the organization.
- **Self-Service Analytics:** Some tools offer self-service functionalities, empowering business users to independently analyze data without relying solely on IT specialists.
- **Predictive Analytics:** Advanced tools can go beyond historical analysis to use data for predictive modeling, forecasting future trends and customer behavior.



Q7) C) Define business intelligence. List and explain any 03 tools for Business intelligence.

Business Intelligence (BI) is a process driven by technology that analyzes an organization's data to provide actionable information for better decision-making. It involves collecting, storing, and analyzing data from various sources to uncover trends, patterns, and insights.

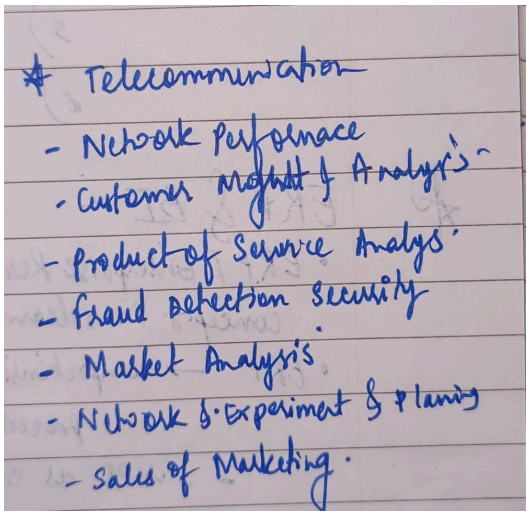


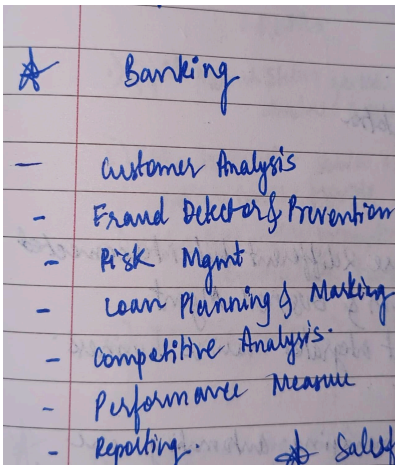
Tools for BI :-

1) Reporting & Quering - allows user for create & run reports & queries to extract of v from db eg → Power BI, Tableau.
2) Data Visulization Tools - ~~fast~~ on creating attractive & intuitive dashboards. eg → Tableau, PowerBI.
3) OLAP Tools - enable to discover 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 warehouse. 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 Kafka

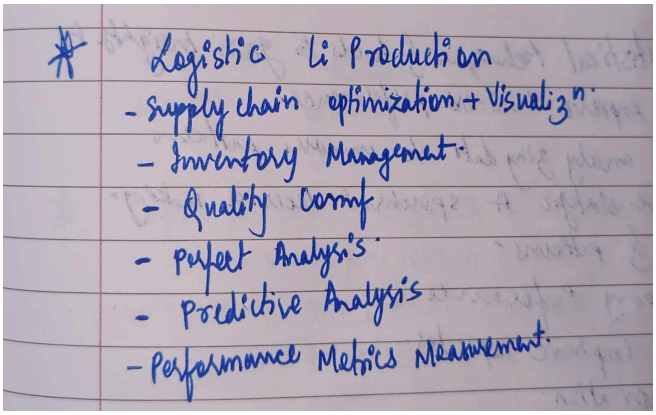
- PowerBI - Microsoft, connects various data sources, create visualization, e/L, easier friendly UI, drag & drop. functionality, NLP queries & AI powers Q&A. features, allows collaboration, strong.
- Oracle DB - RDBMS platform by oracle serve, scalable, high performance mechanism, query optimization parallel processing, admin access monitoring, recovering & backup
- Tableau - data blix (BI tool), offers drag & drop. interactive dashboard & reports, no coding. aggregation facility, allows glaspatial analytics vocabularity & sharing.
- snowflake - cloud based warehouse platform, scalable, elastic, separates storage, compute resources with concurrency sharing.
- Apache Kafka - distributing streaming, platform, handles real time data stream, efficient collector, processing & streaming, high replication, fault tolerance

Q8) A) Explain applications of BI in telecommunication and banking.





Q8) B)Explain BI application in Logistics and production



Q8) C)Explain Role of BI in finance and marketing.