Common Pre-Test for Technical Domains

SECTION A: DATA ANALYSIS & SCIENCE

1. **Imagine you're given a CSV file with 10,000 customer records, but 15% of the "Age" column values are missing. How would you handle this missing data before analysing customer demographics?**

Ans: Handling Missing Age Data :-

If there are missing values in 15% of the 'Age' values in the CSV file, I would tackle this issue prior to analysis by performing the following steps:

1. Identify missing values:

First, I would explore how many missing Age values there are using Python.

2. Choose how you will handle them:

Since Age is a numeric value:

1. Missing ages would be filled with the median age of the customers.

2. Median is preferred as it is not influenced by the presence of very young and very old variables.

3. Impute :

Handling Missing Values: Use Python (fillna(median)).

4. Proceed with Analysis:

Secondly, after handling the missing values, I would be able to process the customers' demographic data without bias and/or errors.

2. **You need to identify which products are frequently purchased together. What kind of analysis would you perform, and what specific techniques or tools would you use?**

Ans: Identifying Products Purchased Together

I would carry out Market Basket Analysis in order to identify the products that are usually purchased together.

What kind of analysis is done?

Market Basket Analysis

It finds the relationships between the products in customer transactions.

Used techniques:

1.Association Rule Mining

2.Common Algorithms:-

  1.Apriori

  2.FP-Growth

These techniques find rules like:

When a customer buys A, he's likely to buy B.

Utilities employed:

1.Python

2.Libreries:

3.pandas: data handling

4.mlxtend Apriori and association rules

Why it's useful:

1.Helps in bundling the product

2.Improves recommendations

3.Improve sales

Example :

from mlxtend.frequent_patterns import apriori, association_rules

3. **Explain the difference between a bar chart and a histogram. When would you use each, and what does this tell you about the nature of the data they represent?**

Ans:

Difference Between Bar Chart and Histogram

Bar Chart	**Histogram**
Used for categorical data	**Used for numerical (continuous) data**
Bars have gaps between them	**Bars touch each other**
Shows comparison between categories	**Shows distribution of data**
Order of bars can be changed	**Order is fixed (numeric order)**

When to use each?

- Bar Chart
  Used when comparing different categories, such as product types, departments, or genders.
- Histogram
  Used when analyzing how numerical data is distributed, such as age, salary, or marks.

What this tells about the data?

- Bar chart shows count or value per category
- Histogram shows data spread, shape, and frequency

Example:

- **Bar Chart → Sales by product category**
- **Histogram → Age distribution of customers**

4. **Write a simple SQL query to find all customers from "New York" who spent more than $500 in January, ordered by their total spending (highest to lowest).**

Ans: SQL query is as follows:-

```
SELECT customer_id, customer_name, SUM(amount) AS total_spent
FROM orders
WHERE city = 'New York'
  AND order_date BETWEEN '2024-01-01' AND '2024-01-31'
GROUP BY customer_id, customer_name
HAVING SUM(amount) > 500
ORDER BY total_spent DESC;
```

5. **You analyse website traffic data and find that page load times increase dramatically when more than 100 users are active. What are at least three possible reasons for this correlation, and how would you investigate further?**

Ans: Possible Reasons for Increased Page Load Time :

1. Server overload

The server may not support more than 100 users at one time because of limited CPU, memory, or bandwidth.

2. Database bottleneck

The number of users running queries on it might be too many, hence slowing down the speed at which the database responds to requests.

3. Network or bandwidth limitation

Internet bandwidth may be too low to accommodate such a high volume of concurrent traffic.

How I would go about investigating further:

1. Monitoring of server performance:

CPU, Memory, and Disk usage during high traffic.

2. Analyze database queries:

Look out for slow or unoptimized queries.

3. Check load and traffic patterns:

Use logs and monitoring tools to clearly highlight when and for what reason delays occur.