1. Data Handling:

- How would you handle missing values in a dataset? Describe at least two methods.

Handling Missing Values

i. **Imputation**: Replace missing values with mean, median, mode, or predicted values using models.

ii. **Deletion**: Remove rows or columns with missing data if the proportion of missing values is small or they are irrelevant.

- Explain why it might be necessary to convert data types before performing an analysis.

Necessary Of Data Type Conversion

i. **Accuracy:** Ensures correct mathematical and logical operations (e.g., treating numerical data as numbers, not strings).

ii. **Efficiency**: Reduces memory usage and speeds up computation by using optimized data types.

2. Statistical Analysis:

- What is a T-test, and in what scenarios would you use it? Provide an example based on sales data.

T-Test :-

A **T-test** is a statistical method used to determine if there is a significant difference between the means of two groups. It assumes the data is normally distributed and that variances are roughly equal.

Scenarios for Use :-

i. Comparing the performance of two groups.
ii. Evaluating the impact of a change or intervention.

Example in Sales Data:-

i. Sales before implementing the strategy (Group A).
ii. Sales after implementing the strategy (Group B).

By conducting a **two-sample T-test**, you can test if the mean sales in Group B are significantly higher than in Group A.

- Describe the Chi-square test for independence and explain when it should be used. How would you apply it to test the relationship between shipping mode and customer segment?

Chi-Square Test for Independence:-

The **Chi-square test for independence** is used to determine whether two categorical variables are related. It compares the observed frequencies of combinations of categories to the expected frequencies under the assumption of independence.

Scenarios for Use:-

i. Testing relationships between categorical variables.
ii. Understanding associations between different groups or segments.

* To test the relationship between shipping mode and customer segment.

Create a contingency table of shipping modes vs. customer segments and test if the observed distribution deviates from what is expected by chance

3. Univariate and Bivariate Analysis:

- What is univariate analysis, and what are its key purposes?

Univariate Analysis:-

Univariate analysis examines a single variable to understand its distribution, central tendency, and spread.

Key Purposes:-

iii. **Summarization:-** Provides insights into the data through measures like mean, median, mode, variance, and standard deviation.

iv. **Visualization**: Helps identify patterns or outliers using histograms, bar charts, or box plots.

Example:- Analyzing the distribution of monthly sales revenue in a dataset.

- Explain the difference between univariate and bivariate analysis. Provide an example of each.

Difference Between Univariate and Bivariate Analysis

Aspect	Univariate Analysis	Bivariate Analysis
Number of Variables	Focuses on one variable at a time.	Examines the relationship between two variables.
Objective	Describes data distribution and variation.	Identifies correlations, patterns, or trends.
Techniques	Mean, median, histograms, box plots.	Scatter plots, correlation coefficients, cross-tabulations.

Example of Univariate Analysis: Examining customer age distribution using a histogram.

Example of Univariate Analysis: Analyzing the relationship between advertising spend and sales revenue using a scatter plot and correlation analysis.

4. Data Visualization:

- What are the benefits of using a correlation matrix in data analysis? How would you interpret the results?

Benefits of Using a Correlation Matrix:

i. **Identifying Relationships**: Shows how strongly variables are related to each other.
ii. **Data Reduction**: Helps identify highly correlated variables, which might be redundant for predictive modeling.
iii. **Trend Discovery**: Highlights positive or negative relationships that can guide further analysis.

Interpretation Of Result:

Correlation values range from -1 to 1:

i. **1**: Strong positive correlation (variables move in the same direction).
ii. **-1**: Strong negative correlation (variables move in opposite directions).
iii. **0**: No correlation (no linear relationship).

For example, if sales and advertising have a correlation of 0.91, it indicates a strong positive relationship.

- How would you plot sales trends over time using a dataset? Describe the steps and tools you would use.

Plotting Sales Trends Over Time

i. **Prepare the Dataset**: Ensure the dataset has a time column (e.g., dates) and a sales column. Aggregate data if needed (e.g., monthly sales).

ii. **Select Tools**: Use tools like Python (Matplotlib, Seaborn, Plotly), Excel, or Power BI for visualization.

Create Plot (Line Chart):

iii. **In Python:** Use Matplotlib for a line chart.

```
import matplotlib.pyplot as plt
sales_data.groupby('Date')['Sales'].sum().plot(kind='line')
plt.title('Sales Trends Over Time')
plt.xlabel('Date')
plt.ylabel('Sales')
plt.show()
```

In Excel: Create a line chart by plotting time on the X-axis and sales on the Y-axis.

In Power BI: Drag date fields to columns and sales fields to rows, and choose the line chart type.

Output: The plot will show fluctuations and trends, such as seasonality or growth over time.

5. Sales and Profit Analysis:

- How can you identify top-performing product categories based on total sales and profit? Describe the process.

Identifying Top-Performing Product Categories Based on Sales and Profit

To identify top-performing product categories, follow these steps:

(a) **Aggregate Data**:

i. **Sales**: Sum the total sales for each product category

ii. **Profit**: Sum the total profit for each product category (profit = sales - cost).

(b) **Rank Categories:**

i. Sort the product categories by total sales and profit in descending order. Identify the categories with the highest total sales and profit.

(c) Visualozation:

i. Use bar charts or pie charts to visualize the distribution of sales and profit across categories.

- Explain how you would analyze seasonal sales trends using historical sales data.

Analyzing Seasonal Sales Trends Using Historical Sales Data

i. **Aggregate Data by Time Period**: Group sales data by time units such as months or quarters.

ii. **Identify Seasonal Patterns**: Look for recurring spikes or drops in sales over specific months or seasons.

iii. **Use Visualization**: Plot time-series charts (e.g., line charts) to visualize trends and identify seasonality.

iv. **Decompose the Time Series**: Apply time series analysis techniques to isolate seasonal, trend, and residual components.

6. Grouped Statistics:

- Why is it important to calculate grouped statistics for key variables? Provide an example using regional sales data.

Important To Calculate Of Grouped Statistics

Grouped statistics help summarize and analyze data by specific categories or segments, revealing patterns or differences within groups.

Example:

For **regional sales data**, calculating grouped statistics like mean, median, or total sales per region can reveal performance differences. For instance, if you calculate the total sales for each region (North, South, East, West), it helps identify which regions are outperforming or underperforming, guiding targeted strategies for improvement.