Stats Project Theory and Ans

---

1. Data Handling

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

    - **Imputation:** Replace missing values with statistical measures like mean, median, or mode, depending on the data type. For instance, replace missing numerical values with the mean or median of the column.
    - **Deletion:** Remove rows or columns with missing values if the proportion of missing data is small and does not significantly affect the dataset's integrity.
- **Why is it necessary to convert data types before performing an analysis?**
  Converting data types ensures consistency and compatibility for computations. For example:

    - **Numerical Data:** String or categorical data needs conversion to numerical format for mathematical operations.
    - **Date/Time Data:** Dates stored as strings should be converted to a datetime format to enable operations like sorting, filtering, or calculating time differences.

---

2. Statistical Analysis

- **What is a T-test, and in what scenarios would you use it? Provide an example based on sales data.**
  A T-test is a statistical test used to compare the means of two groups to determine if they are significantly different.
  Example: Compare the average sales in two regions to determine if the sales performance differs significantly.

- **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?**
  The Chi-square test for independence assesses whether two categorical variables are associated.
  Application: Use a contingency table of shipping mode and customer segment, calculate expected frequencies, and compare them with observed frequencies using the Chi-square statistic. A significant result indicates a relationship between shipping mode and

customer segment.

---

3. Univariate and Bivariate Analysis

- **What is univariate analysis, and what are its key purposes?**
  Univariate analysis involves examining a single variable's distribution and characteristics (e.g., mean, median, mode, variance). Its purpose is to summarize data and identify patterns or outliers.

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

    - **Univariate Analysis:** Focuses on one variable (e.g., analyzing the sales distribution using a histogram).
    - **Bivariate Analysis:** Examines the relationship between two variables (e.g., analyzing the correlation between advertising spend and sales using a scatter plot).

---

4. Data Visualization

- **What are the benefits of using a correlation matrix in data analysis? How would you interpret the results?**
  A correlation matrix identifies relationships between numerical variables.

    1. Positive values indicate a direct relationship, negative values indicate an inverse relationship, and values near 0 suggest no relationship.
    2. Example: Use the matrix to identify which variables (e.g., discount and sales) have the strongest correlations.
- **How would you plot sales trends over time using a dataset? Describe the steps and tools you would use.**
  Steps:

    1. Ensure the date column is in datetime format.
    2. Aggregate sales data by time intervals (e.g., monthly or yearly).
    3. Use a line chart to visualize trends.
       Tools: Python libraries like Matplotlib or Seaborn, or visualization software like Tableau or Excel.

---

5. Sales and Profit Analysis

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

    1. Aggregate sales and profit data by product category.
    2. Rank categories by total sales and total profit.
    3. Identify the top-performing categories based on combined rankings or thresholds.
- **Explain how you would analyze seasonal sales trends using historical sales data.**

    1. Group sales data by time units (e.g., months or quarters).
    2. Calculate average sales for each unit across years to identify patterns.
    3. Visualize trends using line or bar charts.

---

6. Grouped Statistics

- **Why is it important to calculate grouped statistics for key variables? Provide an example using regional sales data.**
  Grouped statistics allow for the comparison of metrics across different groups.
  Example: Calculate average sales and profit for each region to identify regional performance differences.
