Answers to Stats Project Questions

1. Data Handling:

Handling Missing Values:

1. Imputation:

- Replace missing values with the mean, median, or mode of the column, depending on the type of data.

- Example: For numerical sales data, replace missing values with the mean.

2. Deletion:

- Remove rows or columns containing missing data if the percentage of missing values is minimal and does not affect the analysis significantly.

Converting Data Types:

- Necessity:

  - Ensures compatibility with statistical or machine-learning models.

  - Avoids errors when performing mathematical operations (e.g., converting dates to datetime format or sales to numeric format for aggregation).

  - Example: A "price" column stored as text must be converted to a numeric format for analysis.

2. Statistical Analysis:

T-Test:

- Definition:

- A statistical test used to compare the means of two groups.

- Scenario:

  - To determine if the average sales differ between two regions.

- Example:

  - Comparing average sales in Region A vs. Region B during a quarter.

Chi-square Test for Independence:

- Definition:

  - Tests if two categorical variables are independent of each other.

- Scenario:

  - Assessing the relationship between shipping mode (e.g., Standard, Express) and customer segment (e.g., Corporate, Consumer).

- Application:

  - Create a contingency table with frequencies, calculate the Chi-square statistic, and compare it with the critical value to infer independence.

3. Univariate and Bivariate Analysis:

Univariate Analysis:

- Definition:

  - Analysis of a single variable to understand its distribution, central tendency, and spread.

- Purpose:

  - Identify outliers, summarize data, and visualize distributions.

- Example:

  - Analyzing sales data to calculate average sales and plot a histogram.

Bivariate Analysis:

- Definition:
  - Analysis of the relationship between two variables.
- Example:
  - Examining the correlation between marketing spend and sales using a scatter plot.

4. Data Visualization:

Correlation Matrix:

- Benefits:
  - Identifies relationships between multiple variables simultaneously.
  - Highlights positive or negative correlations.
- Interpretation:
  - Values range from -1 (strong negative) to +1 (strong positive); zero indicates no correlation.

Plotting Sales Trends Over Time:

1. Convert the date column to a datetime format.
2. Group sales data by time intervals (e.g., monthly).
3. Use line plots (e.g., via Python's Matplotlib or Excel).

5. Sales and Profit Analysis:

Identifying Top-performing Product Categories:

1. Group data by product categories.

2. Sum sales and profit for each category.

3. Rank categories based on totals.

Analyzing Seasonal Sales Trends:

- Group sales data by seasons or months.

- Compare year-over-year or quarter-over-quarter trends using line or bar charts.

6. Grouped Statistics:

Importance of Grouped Statistics:

- Helps in segment-specific insights.

- Reveals trends, patterns, and anomalies for targeted decisions.

Example:

- Regional sales analysis:

  - Calculate mean, median, and variance of sales for each region to identify high-performing areas.