**Stats Project- Answer**

**1. Data Handling:**

**1.Mean/Median Imputation:** Replace missing values with the mean or median of the column. This is simple and maintains the overall distribution.

**2. Deletion**: Remove rows or columns with missing values if the dataset is large and the proportion of missing data is small.

**\*Explain why it might be necessary to convert data types before performing an analysis:**

Data type conversion ensures compatibility and accuracy in computations. For example, converting string representations of numbers to numeric types allows for mathematical operations, and converting dates to datetime types enables time-series analysis.

**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 compares the means of two groups to see if they are statistically different. For example, you could compare the average sales between two regions to determine if the difference in sales performance is significant.

**\*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 assesses if two categorical variables are related. For instance, you could use it to test whether the shipping mode (standard, express, etc.) is related to the customer segment (retail, corporate, etc.).

**3. Univariate and Bivariate Analysis:**

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

Univariate analysis examines a single variable to summarize and understand its distribution, central tendency, and variability. It is useful for detecting patterns and anomalies

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

**~Univariate Analysis:** Analyzes one variable (e.g., examining the distribution of sales).

**~Bivariate Analysis**: Analyzes the relationship between two variables (e.g., comparing sales and profit to see if they are correlated).

**4. Data Visualization:**

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

A correlation matrix shows the strength and direction of relationships between multiple variables. High correlation coefficients (close to 1 or -1) indicate strong relationships, which can guide further analysis.

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

**1.Data Preparation**: Ensure the sales data is time-stamped.

**2.Tool Selection**: Use tools like Excel, Python (matplotlib, seaborn).

**3.Plotting**: Create line charts to visualize trends over time.

**5. Sales and Profit Analysis:**

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

**1.Data Aggregation:** Sum sales and profit by product category.

**2.Ranking:** Rank categories based on total sales and profit.

**3.Analysis:** Identify and focus on the top-ranked categories.

**\*Explain how you would analyze seasonal sales trends using historical sales data:**

**1.Data Segmentation:** Divide data into time periods (e.g., months, quarters).

**2.Trend Analysis**: Use line charts or seasonal decomposition methods to identify patterns.

**3.Tool Usage:** Apply tools like Excel, Python (pandas, statsmodels), or R.

**6. Grouped Statistics:**

**\*Why is it important to calculate grouped statistics for key variables? Provide an example using regional sales data:**

Grouped statistics provide insights into specific segments, enabling tailored strategies. For example, calculating average sales by region helps identify high-performing areas and regions needing improveme