Stats Project-

1. **Data Handling**:
   - How would you handle missing values in a dataset? Describe at least two methods.
     - **Removing Missing Data**: Sometimes, if the missing values are sparse, you can simply drop rows or columns with missing data using dropna(). However, this can result in data loss, so it's used cautiously when the missing data is not substantial.
     - **Imputation**: This method fills missing values with a calculated estimate based on existing data. Common imputation techniques include: **Mean/Median/Mode Imputation**: Replacing missing values in a numeric column with the mean (for continuous variables) or median/mode (for categorical data).

   - Explain why it might be necessary to convert data types before performing an analysis.
     - Converting data types ensures that the data is in an appropriate format for analysis and helps prevent errors.

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 used to compare the means of two groups to determine if they are statistically significantly different from each other. You would use it when you have two independent groups and want to compare a metric across those groups. For eg: Sales between two cities.
   - 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 independent of each other. It compares the observed frequencies of categories with the expected frequencies (if there were no relationship).
     - When to use: The Chi-square test is used when both variables are categorical, and you want to determine if there is a significant association between them.

3. **Univariate and Bivariate Analysis**:
   - What is univariate analysis, and what are its key purposes?

Univariate analysis involves analyzing one variable at a time to summarize and find patterns in the data.

The key purposes include:

- Understanding the central tendency (mean, median, mode).
- Understanding the spread (range, variance, standard deviation).
- Identifying outliers and anomalies.
- Determining the distribution of a variable.

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

Univariate analysis: Deals with a single variable.

Example: Calculating the average sales (mean) of a product.

Bivariate analysis: Examines the relationship between two variables.

Example: Analyzing how Price Quantity relate to City.

4. **Data Visualization**:
   - What are the benefits of using a correlation matrix in data analysis? How would you interpret the results?
     - A **correlation matrix** visually represents the relationships between numeric variables.
     - The benefits include:
       - Identifying patterns of strong positive or negative correlations.
       - Spotting multicollinearity issues (when two variables are highly correlated).
       - Quickly identifying variables to focus on in further analyses or modeling.
     - **Interpretation**:
     - A correlation close to +1 indicates a strong positive relationship.
     - A correlation close to -1 indicates a strong negative relationship.
     - A correlation near 0 suggests no linear relationship.

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

   To plot sales trends over time:
     - **Convert the date column** to a datetime format (pd.to_datetime()).
     - **Group by time intervals** (e.g., daily, weekly, monthly).
     - **Aggregate sales** for each period using sum().
     - **Plot the data** using matplotlib or seaborn.

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

- **Group data** by product category.

- **Calculate total sales** using .sum() for each category.
- **Sort categories** by sales in descending order to identify the top performers.

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

To analyze seasonal sales trends:

- **Extract the month or season** from the date
- **Group data** by month or season and calculate total sales or average sales.
- **Plot trends** to see if certain months or seasons have consistently higher sales.

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

   Calculating grouped statistics allows you to:

- Compare key variables across different subgroups (e.g., City, time periods).
- Identify regional differences in performance (e.g., higher sales in one City).
- Understand variability across groups.