1. Data Handling

- Handling Missing Values:

Deletion Method: One simple method is to remove rows or columns with missing values. This can be done if the missing data is not a large proportion of the dataset, and deletion won't significantly impact the analysis.

Imputation Method: Another approach is imputing missing values based on other available data. This can be done through various techniques such as:

Mean/Median Imputation: Replacing missing values with the mean or median of the column.

Regression Imputation: Predicting missing values using other variables in the dataset via regression models.

-Converting Data Types:

Before analysis, converting data types ensures that the variables are treated correctly by statistical models and algorithms. For example, if a categorical variable is mistakenly encoded as a numeric variable (e.g., "male" and "female" as 1 and 0), it could lead to incorrect conclusions in analyses like regression or correlation.

2. Statistical Analysis

-T-test: A T-test is a statistical test used to compare the means of two groups to determine if there is a significant difference between them. It’s particularly useful when the sample size is small and the population variance is unknown.

Example using sales data: You could use a T-test to compare the average sales performance between two stores, Store A and Store B, to see if there's a significant difference in their average monthly sales.

-Chi-square Test for Independence:

The Chi-square test is used to determine if there is an association between two categorical variables. It is based on comparing the observed frequencies with the expected frequencies under the assumption that the variables are independent.

Example (Shipping Mode and Customer Segment): To test whether shipping mode (e.g., standard vs. expedited) is related to customer segment (e.g., retail vs. wholesale), you would collect data on both variables, create a contingency table, and apply the Chi-square test to check for independence.

3. Univariate and Bivariate Analysis

-Univariate Analysis:

Univariate analysis involves examining a single variable to summarize and find patterns. Key purposes include understanding the distribution of the data and identifying potential outliers.

Example: Analyzing the distribution of sales in a dataset, including measures like mean, median, mode, and standard deviation.

-Bivariate Analysis:

Bivariate analysis involves examining the relationship between two variables. It helps in understanding how two variables are related or whether changes in one variable affect another.

Example: Analyzing the relationship between advertising spending and sales revenue.

4. Data Visualization

-Correlation Matrix:

A correlation matrix displays the pairwise correlation coefficients between multiple variables. It helps identify linear relationships, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Interpretation: Variables with correlation coefficients near +1 or -1 indicate a strong relationship, while those near 0 suggest little or no linear relationship.

-Plotting Sales Trends Over Time:

To plot sales trends over time, you would:

Prepare the data by ensuring time (e.g., dates or months) is in a usable format.

Use a line chart, where the x-axis represents time (e.g., months or days), and the y-axis represents sales.

Tools like Excel, Google Sheets, or Python's Matplotlib/Seaborn can be used to create the plot.

5. Sales and Profit Analysis

-Identifying Top-Performing Product Categories:

You can identify top-performing product categories by aggregating total sales and profit for each category.

Process: Sum the sales and profit for each category and rank them to identify the highest values.

Example: Group sales data by category, calculate the total sales and profit, and then sort the categories based on these metrics.

-Analyzing Seasonal Sales Trends:

Seasonal trends can be analyzed by grouping sales data by time periods, such as months or quarters.

Process: Plot the sales data over time and look for recurring patterns that correspond to certain seasons (e.g., higher sales in the winter or around holidays).

Tools like Excel, Google Sheets, or R can be useful to visualize trends.

6. Grouped Statistics

Importance of Grouped Statistics:

Grouped statistics help to understand variations within different segments or categories of data. They provide insights into patterns that may not be apparent in overall statistics.

Example (Regional Sales Data): If you have sales data from different regions, you can calculate the average sales, profit margins, and other statistics for each region. This helps identify regions that are performing well or underperforming.