Basic Statistics

### Descriptive Analytics and Data Preprocessing on Sales & Discounts Dataset

#### Introduction

* To perform descriptive analytics, visualize data distributions, and preprocess the dataset for further analysis.

#### Descriptive Analytics for Numerical Columns

* Objective: To compute and analyze basic statistical measures for numerical columns in the dataset.
* Steps:
  + Load the dataset into a data analysis tool or programming environment (e.g., Python with pandas library).
  + Identify numerical columns in the dataset.
  + Calculate the mean, median, mode, and standard deviation for these columns.
  + Provide a brief interpretation of these statistics.

#### Data Visualization

* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:
  + Plot histograms for each numerical column.
  + Analyze the distribution (e.g., skewness, presence of outliers) and provide inferences.
* **Boxplots**:
  + Create boxplots for numerical variables to identify outliers and the interquartile range.
  + Discuss any findings, such as extreme values or unusual distributions.
* **Bar Chart Analysis for Categorical Column:**
  + Identify categorical columns in the dataset.
  + Create bar charts to visualize the frequency or count of each category.
  + Analyze the distribution of categories and provide insights.

#### Standardization of Numerical Variables

* Objective: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Steps:
  + Explain the concept of standardization (z-score normalization).
  + Standardization, also known as z-score normalization, is a data preprocessing technique used to scale numerical variables. It transforms the data so that it has a mean (μ\muμ) of 0 and a standard deviation (σ\sigma σ) of 1. The formula for standardization is:
  + **z=(x−μ)/σ**
  + Here:
* x is the original value,
* μ is the mean of the column,
* σ is the standard deviation of the column.
  + This process ensures uniformity in the data, which is particularly beneficial for machine learning models that are sensitive to the scale of input data.
  + Standardize the numerical columns using the formula: z=x-mu/sigma
  + ​Show before and after comparisons of the data distributions.

#### Conversion of Categorical Data into Dummy Variables

* Objective: To transform categorical variables into a format that can be provided to ML algorithms.
* Steps:
  + Discuss the need for converting categorical data into dummy variables (one-hot encoding).
  + Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.
  + Display a portion of the transformed dataset.

#### Conclusion

* Summarize the key findings from the descriptive analytics and data visualizations.
* Reflect on the importance of data preprocessing steps like standardization and one-hot encoding in data analysis and machine learning.