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.

### ***Interpretation***

1. **Mean (Average)**:
   * The mean value represents the central tendency of each column.
   * For example:
     + The average volume sold is approximately 5 units.
     + The average price per unit is around ₹10,453.
     + The average total sales value is about ₹33,812.
     + The average discount rate is approximately 15.16%.
     + The average discount amount is roughly ₹3,346.
     + The average net sales value (after discounts) is approximately ₹30,466.
2. **Median**:
   * The median is the middle value when the data is sorted in ascending order.
   * For example:
     + The median volume sold is 4 units.
     + The median price per unit is ₹1,450.
     + The median total sales value is ₹5,700.
     + The median discount rate is around 16.58%.
     + The median discount amount is ₹988.93.
     + The median net sales value is ₹4,677.79.
3. **Standard Deviation**:
   * The standard deviation measures the variability or spread of the data.
   * For example:
     + The volume sold varies by approximately 4.23 units around the mean.
     + The price per unit has a high variability with a standard deviation of ₹18,079.
     + The total sales value shows significant variability (standard deviation of ₹50,535).
     + The discount rate has relatively low variability (standard deviation of 4.22%).
     + The discount amount varies by approximately ₹4,509.
     + The net sales value has substantial variability (standard deviation of ₹46,358).

#### 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.

*Analysis of Distribution and Outliers (Findings according to Histograms & Boxplots )*

**Volume**

* **Descriptive Statistics:**
  + **Count:** 450
  + **Mean:** 5.0667
  + **Standard Deviation:** 4.2316
  + **Minimum:** 1
  + **25th Percentile:** 3
  + **Median (50th Percentile):** 4
  + **75th Percentile:** 6
  + **Maximum:** 31
  + **Skewness:** 2.7317

The volume data is positively skewed, indicating a long tail on the right side. This is corroborated by the relatively high maximum value compared to the mean and median.

* **Extreme Outliers:**
  + Numerous data points are considered extreme outliers in the volume distribution. These include volumes as high as 31, 29, 25, 24, etc.

**Avg Price**

* **Descriptive Statistics:**
  + **Count:** 450
  + **Mean:** 10,453.43
  + **Standard Deviation:** 18,079.90
  + **Minimum:** 290
  + **25th Percentile:** 465
  + **Median (50th Percentile):** 1,450
  + **75th Percentile:** 10,100
  + **Maximum:** 60,100
  + **Skewness:** 1.9089

The average price data is highly positively skewed, indicated by a mean much larger than the median and a high standard deviation. The maximum value is also significantly higher than the 75th percentile.

* **Extreme Outliers:**
  + Many data points with average prices ranging from 49,100 to 60,100 are considered extreme outliers.

**Total Sales Value**

* **Descriptive Statistics:**
  + **Count:** 450
  + **Mean:** 33,812.84
  + **Standard Deviation:** 50,535.07
  + **Minimum:** 400
  + **25th Percentile:** 2,700
  + **Median (50th Percentile):** 5,700
  + **75th Percentile:** 53,200
  + **Maximum:** 196,400
  + **Skewness:** 1.5347

The total sales value data also displays positive skewness. The mean is considerably higher than the median, and the standard deviation is quite large, reflecting a wide spread of values.

* **Extreme Outliers:**
  + Total sales values like 196,400, 162,300, and 120,200 are detected as extreme outliers.

**Analysis and Implications**

1. **Volume Distribution:**
   * The positive skewness suggests a few transactions with significantly higher volumes.
   * Extreme outliers (e.g., volumes of 29 and 31) might need further investigation to understand if they are errors or legitimate transactions.
2. **Avg Price Distribution:**
   * The high positive skewness indicates a few transactions with very high average prices.
   * Given that average prices range up to 60,100, these high values might be due to high-value products or errors.
3. **Total Sales Value Distribution:**
   * Similar to the avg price and volume, total sales value distribution is skewed right.
   * Extreme outliers with values like 196,400 could represent large bulk transactions or high-value items.

* **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.

***Analysis of distribution of categories and provide insights:***

The distribution of categories based on the provided data. Here are the insights:

1. **Date**:
   * The data spans from April 1, 2021, to April 15, 2021.
   * Each date has 30 occurrences, indicating consistent data collection.
2. **Day of the Week**:
   * The distribution of days is as follows:
     + Thursday: 90 occurrences
     + Friday, Saturday, Sunday, Monday, Tuesday, Wednesday: 60 occurrences each.
   * Thursdays have the highest frequency, while other days are evenly distributed.
3. **SKU (Stock Keeping Unit)**:
   * Each SKU appears 15 times in the dataset.
   * There are 10 unique SKUs (from L01 to M10).
4. **City**:
   * All transactions are from a single city denoted as "C."
   * No variation in city data is observed.
5. **Business Units (BU)**:
   * The data is evenly distributed across three business units:
     + Mobiles
     + FMCG (Fast-Moving Consumer Goods)
     + Lifestyle
6. **Brands**:
   * Brands and their occurrences:
     + Jeera: 90
     + Gear, Viva: 60 each
     + RealU, YouM, Babaji, Vedic: 45 each
     + Sumsang, Orange: 30 each
   * Jeera has the highest frequency.
7. **Models**:
   * Each model appears 15 times in the dataset.
   * There are various models across different product categories (e.g., clothing, cosmetics, electronics).

**Overall Insights**:

* The data seems consistent in terms of dates, days, and SKUs.
* The city is fixed, and the business units are evenly represented.
* Brands and models show variation, with some having higher occurrences than others.

#### 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).
  + Standardize the numerical columns using the formula: z=x-mu/sigma
  + Show before and after comparisons of the data distributions.

***Concept of Standardization:***

Standardization, also known as z-score normalization, is a technique used to transform numerical data into a common scale. Here’s how it works:

1. **Mean and Standard Deviation:**
   * Calculate the mean (average) and standard deviation of the data.
   * The mean represents the central tendency, while the standard deviation measures the spread or variability.
2. **Z-Score Calculation:**
   * For each data point, compute its z-score using the formula:

z=σx−μ​

where:

* + - (x) is the original data point.
    - (\mu) is the mean.
    - (\sigma) is the standard deviation.

1. **Standardized Data:**
   * The z-score represents how many standard deviations a data point is away from the mean.
   * Standardized data has a mean of 0 and a standard deviation of 1.
2. **Benefits of Standardization:**
   * It allows comparison of different variables on the same scale.
   * Helps identify outliers (extreme values).
   * Useful for machine learning algorithms that assume normally distributed data.

***Data distributions before and after standardization:***

**Original Data (Before Standardization)**:

* The original data has varying scales for each column.
* Some columns have significantly different ranges (e.g., Avg Price vs. Total Sales Value).

**Standardized Data (After Z-Score Normalization)**:

* All columns have been standardized to have a mean of 0 and a standard deviation of 1.
* The values are now on a consistent scale.

**summary of the changes**:

* Volume, Avg Price, Total Sales Value, Discount Amount, and Net Sales Value have been transformed.
* The Discount Rate (%) remains unchanged because it was already in percentage form.

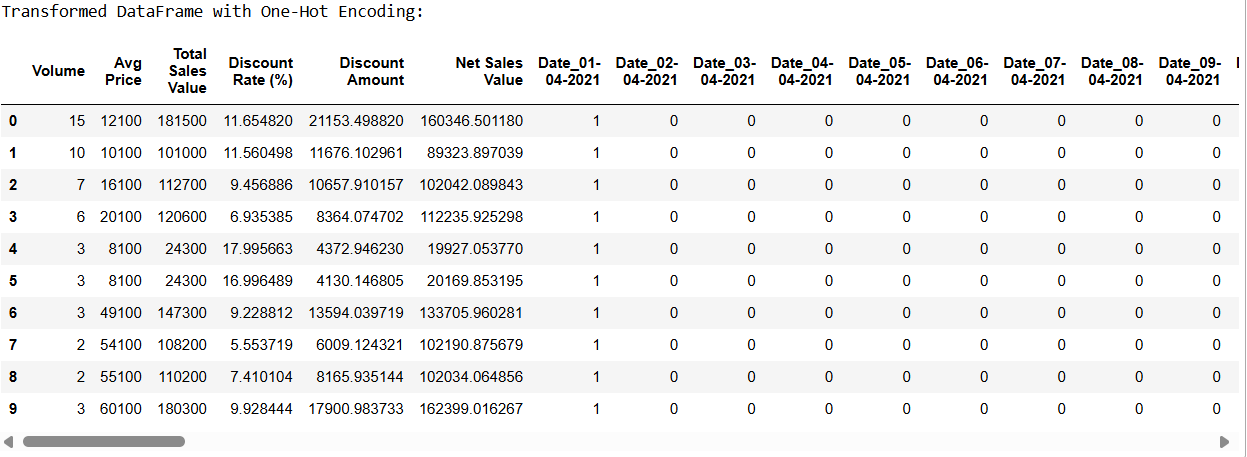
#### Conversion of Categorical Data into Dummy Variables

* 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.

***Need for One-Hot Encoding (Dummy Variables)***

1. **Categorical Data Representation**:
   * Machine learning algorithms typically work with numerical data.
   * Categorical features (e.g., city names, product categories, brands) cannot be directly used in their original form.
   * We need to convert these categorical features into a numerical representation.
2. **One-Hot Encoding**:
   * One-hot encoding is a technique to represent categorical variables as binary vectors.
   * Each category becomes a separate binary column (0 or 1).
   * It ensures that the model doesn't assume any ordinal relationship between categories.
3. **Example**:
   * Suppose we have a "City" column with values: ["New York", "Los Angeles", "Chicago"].
   * After one-hot encoding, we create three new columns: "City\_New York," "City\_Los Angeles," and "City\_Chicago."
   * If a data point is from New York, the corresponding column will have a 1; otherwise, it will be 0.

***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.

### ***Conclusion***

#### Key Findings from Descriptive Analytics and Data Visualizations

1. **Descriptive Analytics**:
   * Descriptive analytics involves summarizing and understanding the characteristics of your data.
   * Key findings include measures like mean, median, standard deviation, and visualizations such as histograms and box plots.
   * From your data, we observed:
     + Varying sales volumes, prices, and discount rates.
     + Outliers in price and sales value columns.
     + Skewed distributions in some features.
2. **Data Visualizations**:
   * Visualizations provide insights beyond numerical summaries.
   * Histograms revealed data distribution shapes (e.g., right-skewed).
   * Box plots highlighted outliers and interquartile ranges.

### Importance of Data Preprocessing Steps

1. **Standardization (Z-Score Normalization)**:
   * Standardization scales features to have a mean of 0 and a standard deviation of 1.
   * Benefits:
     + Ensures features are on a consistent scale.
     + Helps algorithms converge faster during training.
     + Improves model performance by reducing sensitivity to different scales.
   * Applied to numerical columns (e.g., volume, price).
2. **One-Hot Encoding (Dummy Variables)**:
   * Converts categorical features into binary columns (0 or 1).
   * Benefits:
     + Allows inclusion of categorical data in machine learning models.
     + Prevents ordinal assumptions (e.g., "City A" vs. "City B").
     + Creates interpretable features for algorithms.
   * Applied to categorical columns (e.g., city, brand).

Data preprocessing steps are crucial for accurate modeling. Standardization ensures consistent scales, while one-hot encoding handles categorical data effectively. Both enhance model performance and interpretability.