### Descriptive Analytics and Data Pre-processing on Sales & Discounts Dataset

#### Introduction

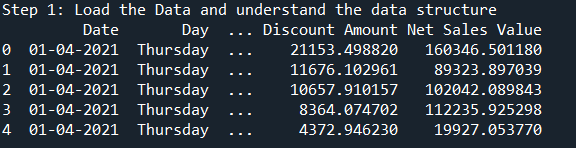
* To perform descriptive analytics, visualize data distributions, and pre-process the dataset for further analysis.

#### Descriptive Analytics for Numerical Columns

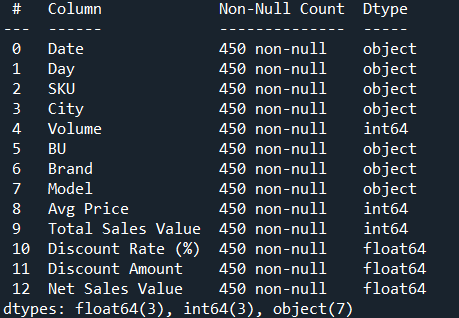
* Objective: To compute and analyse basic statistical measures for numerical columns in the dataset.
* Steps:
  + Load the dataset into a data analysis tool or programming environment (e.g., Python with panda’s library).
  + Identify numerical columns in the dataset.
  + Calculate the mean, median, mode,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

**Observation –**

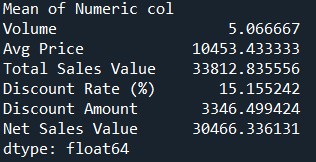
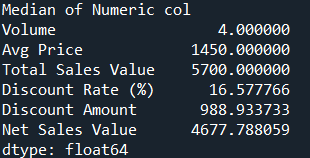
1. Read the dataset with the help of Dataframe of lib pandas and identify the column type with the help of select\_dtypes method.

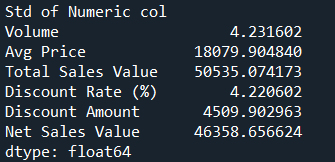


1. Perform data clean-up, check for null and duplicate values



1. Calculated mean, median, mode and S.D of numeric column and below are observations-

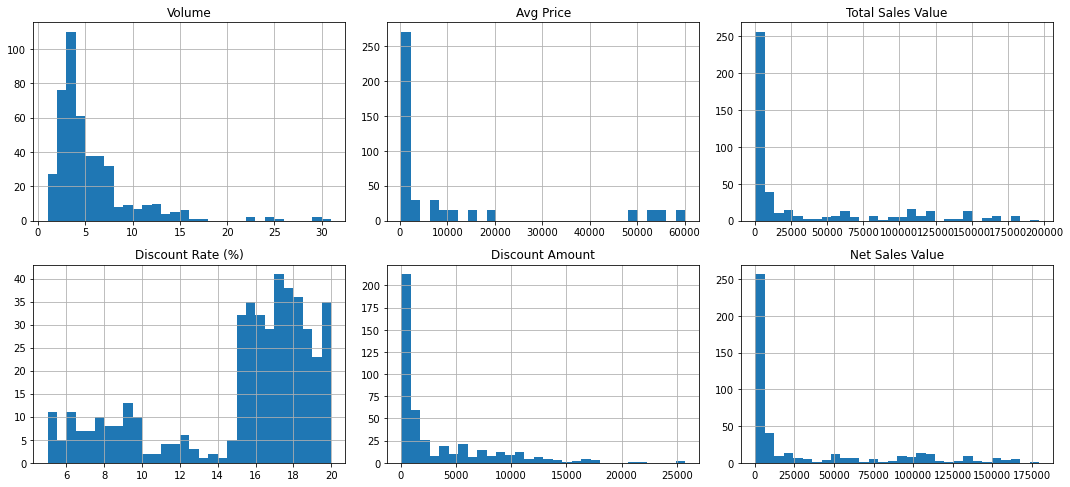


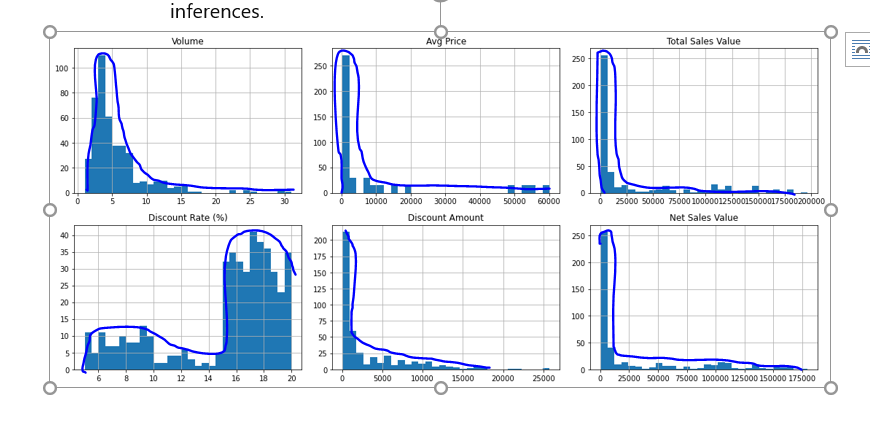
1. **Volume**: Most sales are of lower volumes, with some higher-volume outliers.
2. **Avg Price**: Skewed by a few extremely high prices, leading to a large gap between mean and median.
3. **Total Sales Value**: Dominated by a few high-value transactions, causing significant variability.
4. **Discount Rate**: Relatively consistent, with moderate variation and a symmetric distribution.
5. **Discount Amount**: Influenced by high-value discounts, creating variability and skewness.
6. **Net Sales Value**: A wide range of values due to a few large transactions.

Overall, the data suggests typical sales are smaller, with a few large outliers significantly affecting averages.

**Data Visualizations –**

* **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.





1. **Volume**:
   * The distribution is right-skewed, with most sales having low volumes (around 0 to 5 units).
   * A small number of higher-volume sales exist but are rare, contributing to the long tail.
2. **Avg Price**:
   * Strongly right-skewed, with the majority of sales concentrated at lower price ranges (below 10,000).
   * A few transactions have very high average prices, contributing to the extended tail.
3. **Total Sales Value**:
   * Highly skewed to the right, with most sales values below 25,000.
   * A small number of high-value sales drive the long tail.
4. **Discount Rate (%)**:
   * A more symmetric distribution compared to other variables, with values concentrated between 10% and 20%.
   * The discount rates cluster around the 15% to 17% range, indicating consistency in discount practices.
5. **Discount Amount**:
   * Skewed to the right, with most discount amounts below 5,000.
   * A few transactions with very high discount amounts account for the long tail.
6. **Net Sales Value**:
   * Strongly right-skewed, with most sales below 25,000.
   * A small number of very high net sales values drive the tail, indicating the presence of outliers.

Key Insights:

* Most variables (except for Discount Rate) are highly skewed, with the majority of data concentrated in lower ranges and a few extreme values extending the tails.
* The right-skewness across Avg Price, Total Sales Value, Discount Amount, and Net Sales Value suggests the presence of outliers or specific transactions that are significantly larger than typical ones.
* Discount Rate shows a more stable and consistent pattern compared to other metrics.

**Boxplots**:

* Create boxplots for numerical variables to identify outliers and the interquartile range.
* Discuss any findings, such as extreme values or unusual distributions.

### 1.Volume:

* The majority of the data lies between 0 and 10 (as shown by the interquartile range, IQR).
* Several outliers exist above 15, indicating occasional high-volume sales that are rare compared to typical transactions.
* IQR Range for X - 3.0

### 2. **Avg Price**:

* The IQR is concentrated in the lower range (below approximately 5,000), with the median close to this range.
* Many outliers are present at higher price levels, confirming a few extremely high-priced transactions.
* IQR Range for X - 9635.0

### 3. **Total Sales Value**:

* Most sales values are clustered below 25,000 (IQR), with a median closer to the lower bound.
* Numerous high-value outliers are present, suggesting significant variability caused by infrequent high-value sales.
* IQR Range for X - 50500.0

### 4. **Discount Rate (%)**:

* The distribution is relatively symmetric, with the majority of values between approximately 10% and 18%.
* There are a few outliers below 10%, indicating rare instances of unusually low discount rates.
* IQR Range for X- 4.149655934050962

### 5. **Discount Amount**:

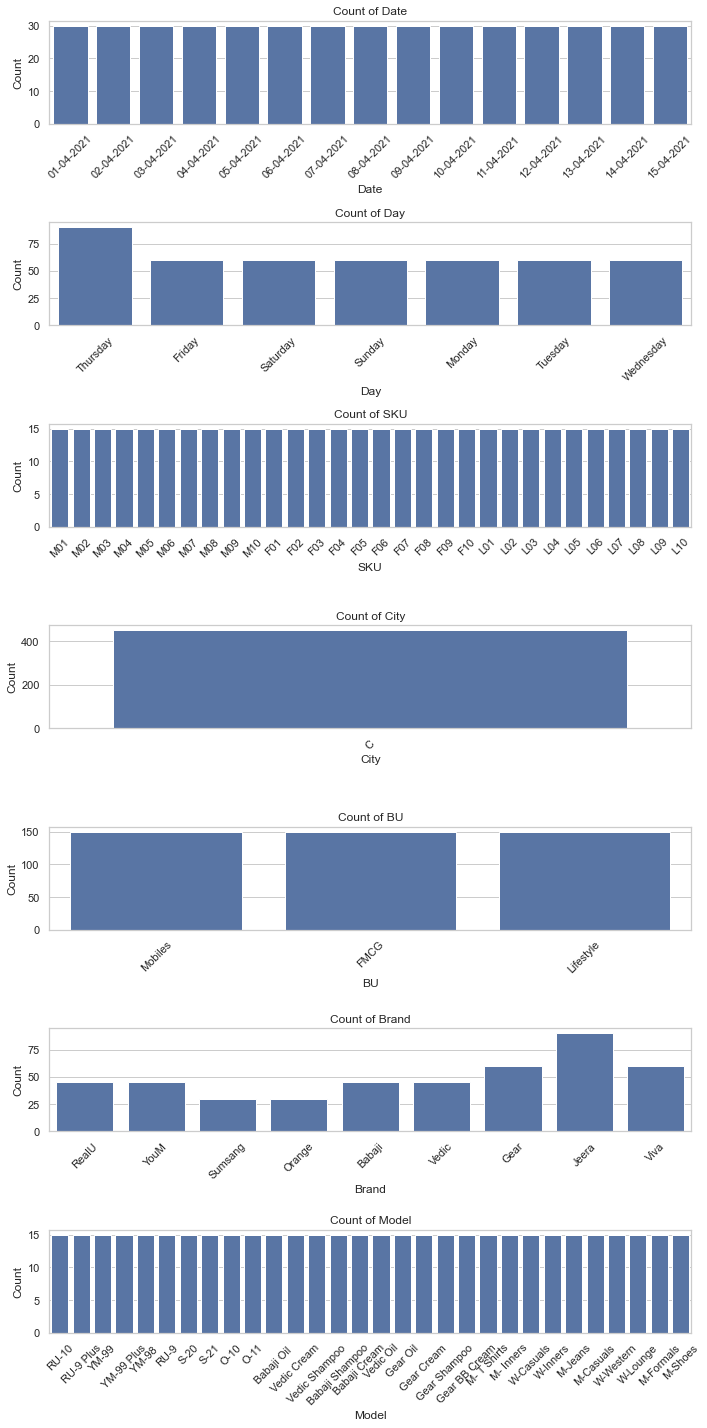
* The IQR lies below 5,000, with the median on the lower end of this range.
* Several high-value outliers suggest occasional large discounts, likely associated with high-value sales.
* IQR Range for X - 4856.03612280112

### 6. **Net Sales Value**:

* The bulk of the data is below 25,000 (IQR), with the median skewed toward the lower end of this range.
* Numerous outliers exist above 50,000, reflecting rare but significantly high-value sales.
* IQR Range for X - 45645.70420627015

### **Overall Observations**:

* **Skewness and Outliers**: Most variables, except Discount Rate, exhibit a strong right-skewed distribution with numerous outliers. These outliers indicate infrequent but impactful high-value transactions or extreme discounts.
* **Stability**: Discount Rate is relatively consistent with fewer outliers, suggesting stable discount policies.
* **Implication**: The data suggests a need to address variability in high-value transactions. These outliers may represent unique customer segments, special deals, or errors, warranting further
* **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.



### 1. **Date**:

* The count of entries across all dates is uniform, indicating data collection is consistent over time.
* There are no evident peaks or drops in counts for specific dates.

### 2. **Day**:

* Thursday has a significantly higher count compared to other days.
* Other days (Friday to Wednesday) have relatively balanced counts, suggesting Thursday might be a peak business day or the data collection is more focused on Thursdays.

### 3. **SKU (Stock Keeping Unit)**:

* Counts are evenly distributed across all SKUs.
* This indicates that all SKUs are equally represented in the dataset without any dominance by specific items.

### 4. **City**:

* One city has a significantly higher count than others, dominating the dataset.
* This suggests that business activity or data collection is concentrated in a single city.

### 5. **BU (Business Unit)**:

* All business units have approximately equal representation in the dataset.
* There’s no clear dominance, which implies a balanced data distribution across business units.

### 6. **Brand**:

* Some brands, such as "Jeera" and "Viva," have noticeably higher counts compared to others like "Orange" or "YouM."
* This suggests these brands are more popular or have higher sales volume in the dataset.

### 7. **Model**:

* The counts for models are uniformly distributed.
* This indicates all models are equally represented, without any particular model standing out in the data.

### **Insights**:

* **Dominance by City**: The data is skewed toward a single city, which might indicate either a large customer base or greater operational focus in that city.
* **Day Effect**: Higher counts on Thursday suggest it might be a critical day for sales or data collection.
* **Balanced Data**: Categories like SKUs, BU, and Model show balanced representation, which is useful for unbiased analysis.
* **Brand Preference**: A few brands stand out, which might indicate higher market share or customer preference.

#### 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

**Explaination-**

Standardization, also known as z-score normalization, is a technique to scale data. It transforms the values in numerical columns to have:

* **Mean (μ) = 0**
* **Standard deviation (σ) = 1**

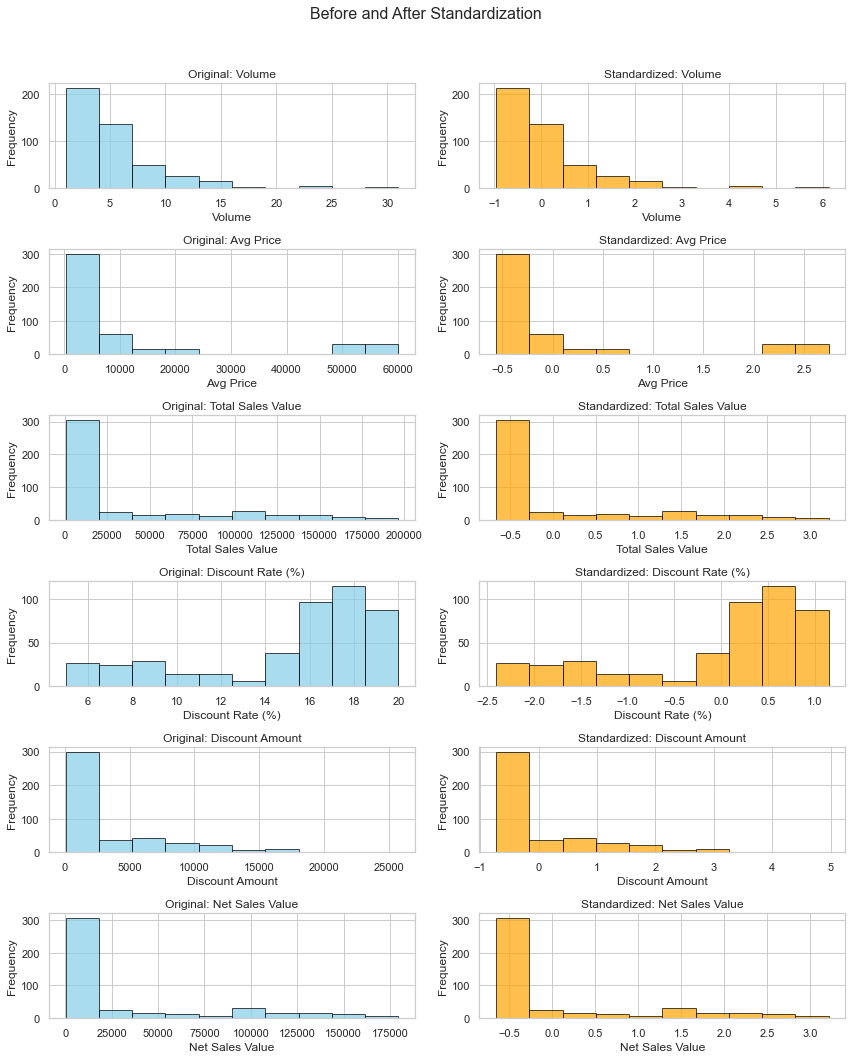
This ensures that all features contribute equally to a model, especially when they have different scales or units.

The formula for standardization is:

z=x-mu/sigma

Where:

* z: Standardized value
* x: Original value
* mμ: Mean of the column
* sigma: Standard deviation of the column
* Essential for algorithms sensitive to feature scaling, such as:
  + Gradient descent-based models (linear regression, neural networks).
  + Distance-based models (k-NN, PCA).
* Reduces bias introduced by differences in scale.



### Observations from Before and After Standardization

1. **Distribution Shape**:
   * The shape of the distributions for all numerical columns remains unchanged after standardization. Standardization does not affect the underlying distribution of the data; it only rescales it.
2. **Scale Transformation**:
   * **Before Standardization**:
     + Each column had its original scale, ranging from small values (e.g., Volume, Discount Rate) to very large values (e.g., Avg Price, Total Sales Value).
     + The scales across columns were inconsistent, with some features spanning a narrow range and others a much wider range.
   * **After Standardization**:
     + The scales are now uniform, centered around 0 with a standard deviation of 1. Each column's data is transformed into z-scores.
3. **Mean-Centering**:
   * After standardization, all columns are centered around 0. The mean of each standardized column is approximately 0.
4. **Spread of Data**:
   * After standardization, the spread (standard deviation) of each column is reduced to 1.
   * The data points are normalized to lie within a consistent range, improving comparability across features.
5. **Outliers**:
   * Outliers in the original data are still present in the standardized data.
   * These outliers are now represented as extreme positive or negative z-scores.

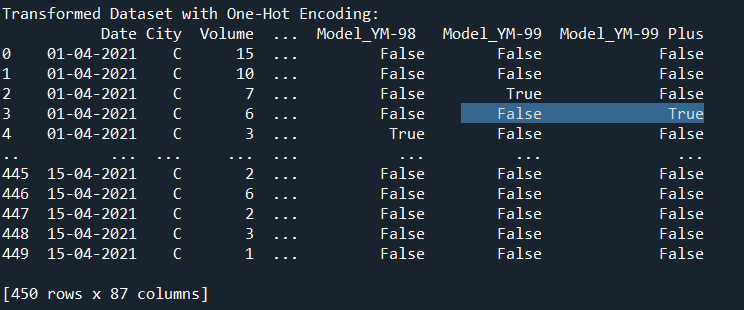
**Note-** Try to normalized data using another method sklearn which follows the same formula

#### Conversion of Categorical Data into Dummy Variables

* + Discuss the need for converting categorical data into dummy variables (one-hot encoding).
  + Categorical data represents qualitative information, such as labels or categories (e.g., "City," "Brand"). Many machine learning algorithms cannot directly work with such non-numeric data because they rely on mathematical operations.
  + Need for Dummy Variables (One-Hot Encoding)
    - **Numerical Representation**: Dummy variables (or one-hot encoding) transform categorical data into a numerical format that machine learning models can understand.
    - **Avoiding Ordinal Assumption**: Assigning numerical labels directly (e.g., "A" = 1, "B" = 2) may mislead the model into assuming an ordinal relationship, which may not exist.
    - **Preserving Information**: Each category is represented as a separate binary column (0 or 1), ensuring no information is lost.

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-**

Descriptive analytics and data visualizations provide an initial understanding of patterns and anomalies in the dataset. Preprocessing steps like **standardization** and **one-hot encoding** are indispensable for ensuring that datasets are clean, consistent, and compatible with machine learning algorithms. These steps ultimately improve model robustness and the reliability of insights drawn from data analysis.