Basic Statistics

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

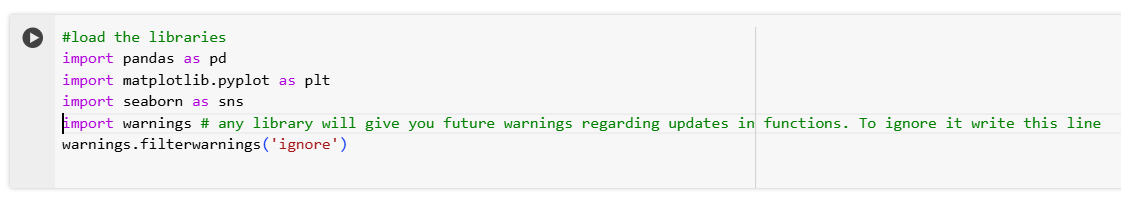
#### Introduction

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

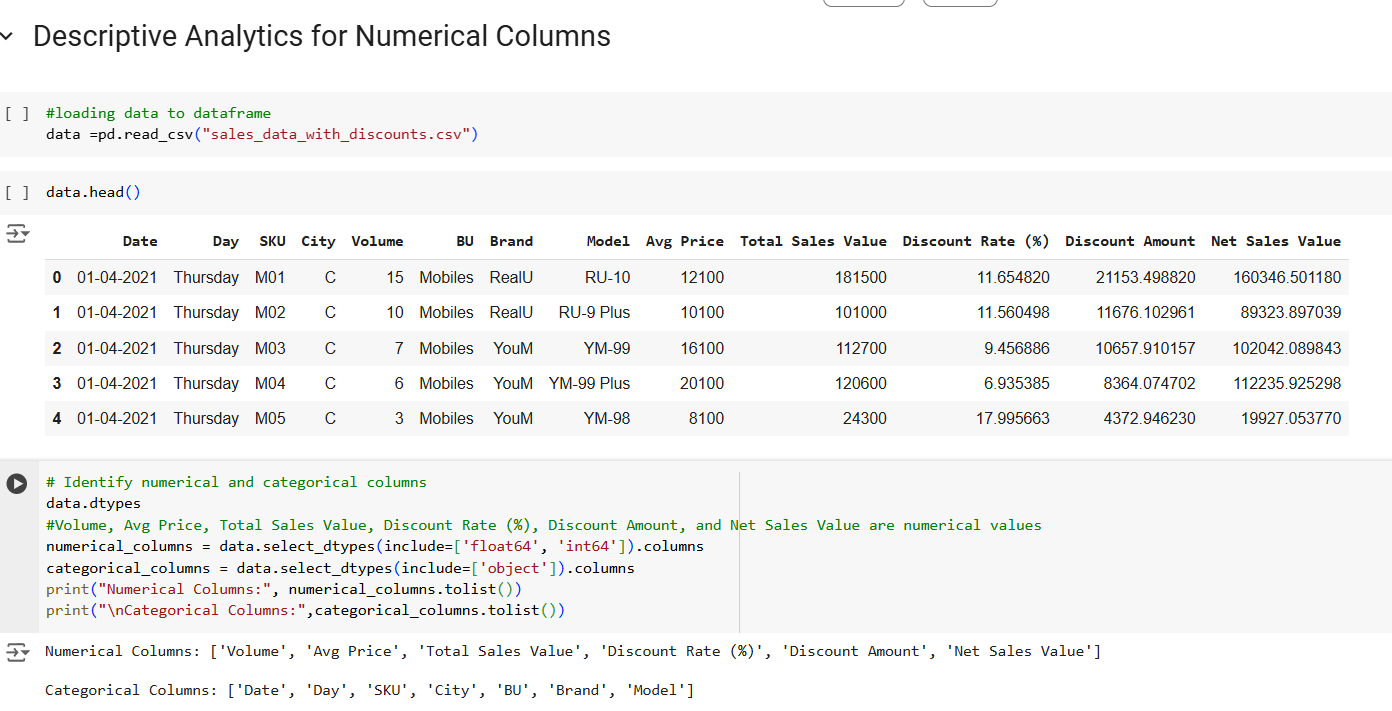
#### Descriptive Analytics for Numerical Columns

* 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.
* Solution:

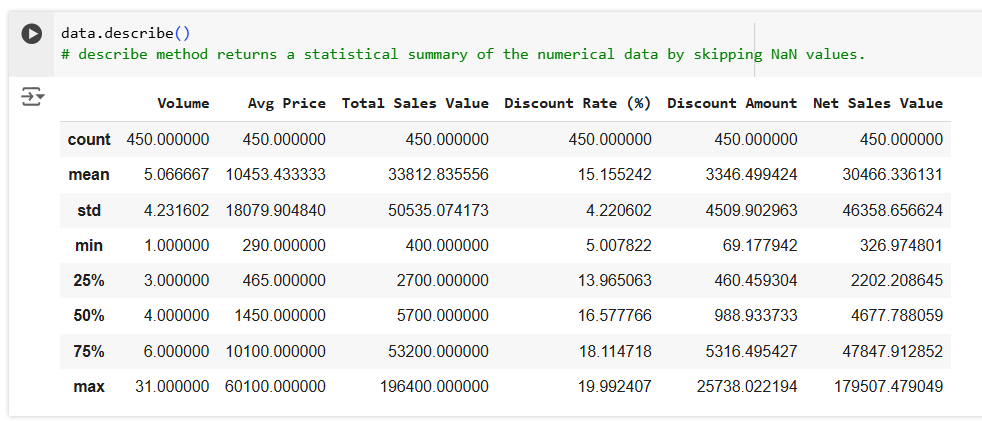
Load the python libraries



Loading data into a data frame and identifying numerical and categorical columns in data set



Calculate the mean, median, mode, and standard deviation using describe() method for all the numeric columns in the dataset.

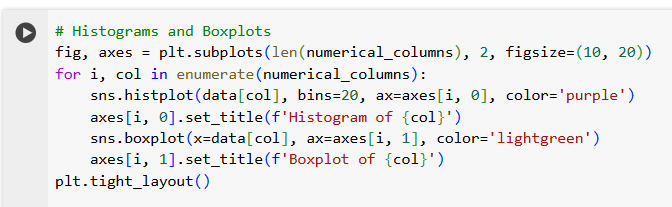


**Interpretation:** All the numerical columns have outliers and a right skewness as Mean greater than Median, where *Avg Price* and *Total Sales Value* has high variance compare to other numerical columns.

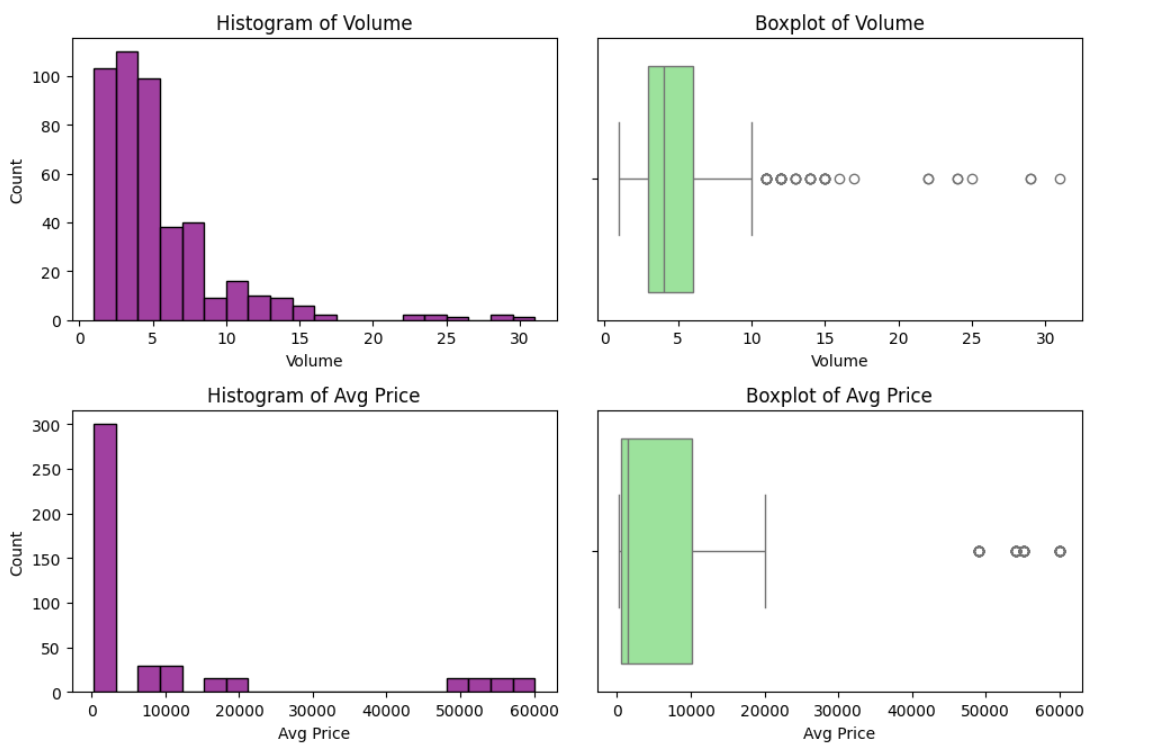
#### Data Visualization

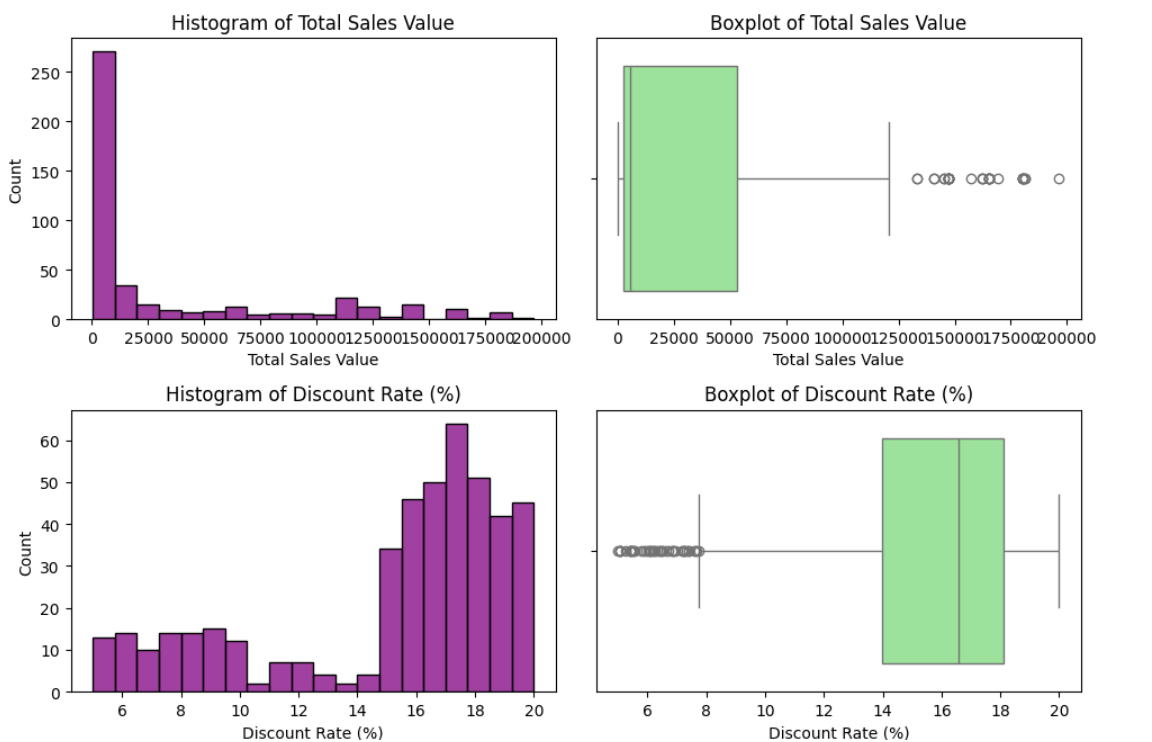
* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:
  + Plot histograms for each numerical column.
  + Analyze the distribution (e.g., skewness, presence of outliers) and provide inferences.
* **Boxplots**:
  + Create 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.
* Solution:

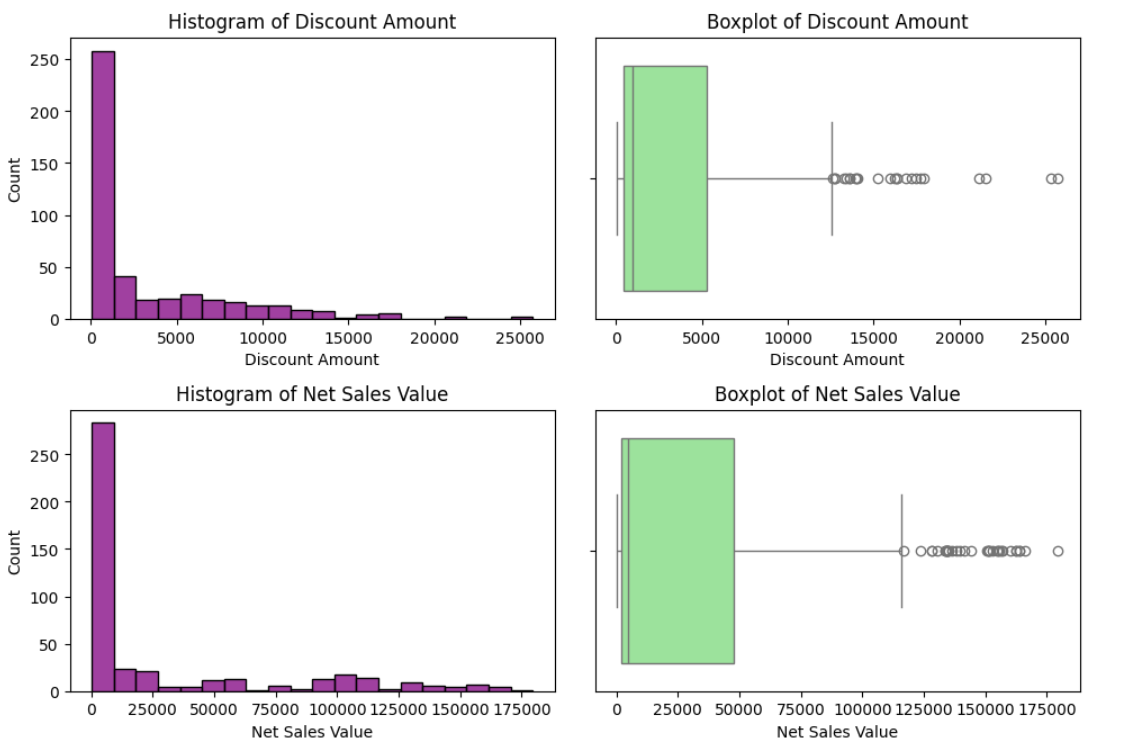
Creating boxplot and histograms for all the numeric values



Output:







**Analysis and Findings from histogram and boxplot:**

*Volume:* There are outliers and a slight positive skew with a small number of high-volume values extending the range.

*Avg Price:* there are outliers and high positive skew where a few high prices significantly influence the dataset.

*Total Sales Value:* There are outliers due to high value transactions and high positive skew

*Discount Rate (%):* There are outliers visible at higher rates with minimal skew

*Discount Amount:* There are outliers and positive skew with a small number of high-volume values.

*Net Sales Value:* There are outliers and positive skew due to high sales values.

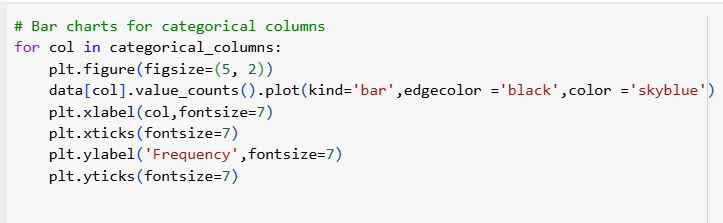
**Inferences:**

Outliers in Avg Price, Total Sales Value, and Net Sales Value could mislead summary statistics. Hence, they should be treated carefully

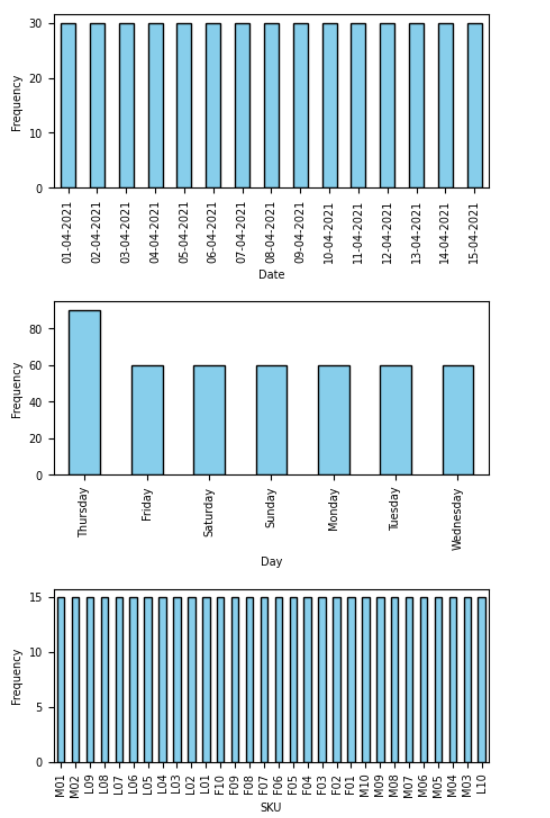
Strong positive skew occurs in columns Avg Price, Total Sales Value. the data can be normalized for analysis.

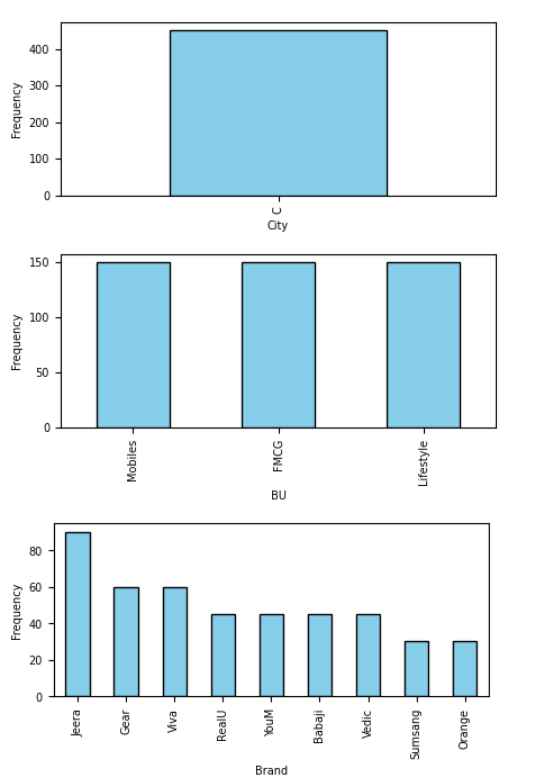
Discount Rate (%) has minimal skewness and fewer outliers.

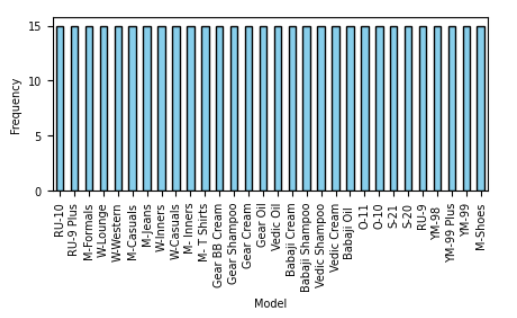
**Bar chart for categorical columns in python:**



Output:







**Analysis from Bar chart:**

Date: The count for each unique value in the date column is 30 and they are equally distributed.

Day: The count for each unique value in the day column is 7 and they are equally distributed, except for Thursday. And sales are high on Thursday.

SKU: The count for each unique value in the SKU column is 15 and they are equally distributed.

City: This column has only one unique value i.e 'c' and the count for each unique value is 450.

BU: This column has 3 categories. The count for each unique value in the BU column is 150 and they are equally distributed.

Brand: The categories in this column are unequally distributed, But the brand Jeera has the highest number of sales.

Model: The count for each unique value in the Model column is 15 and they are equally distributed.

#### 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.
* Solution:

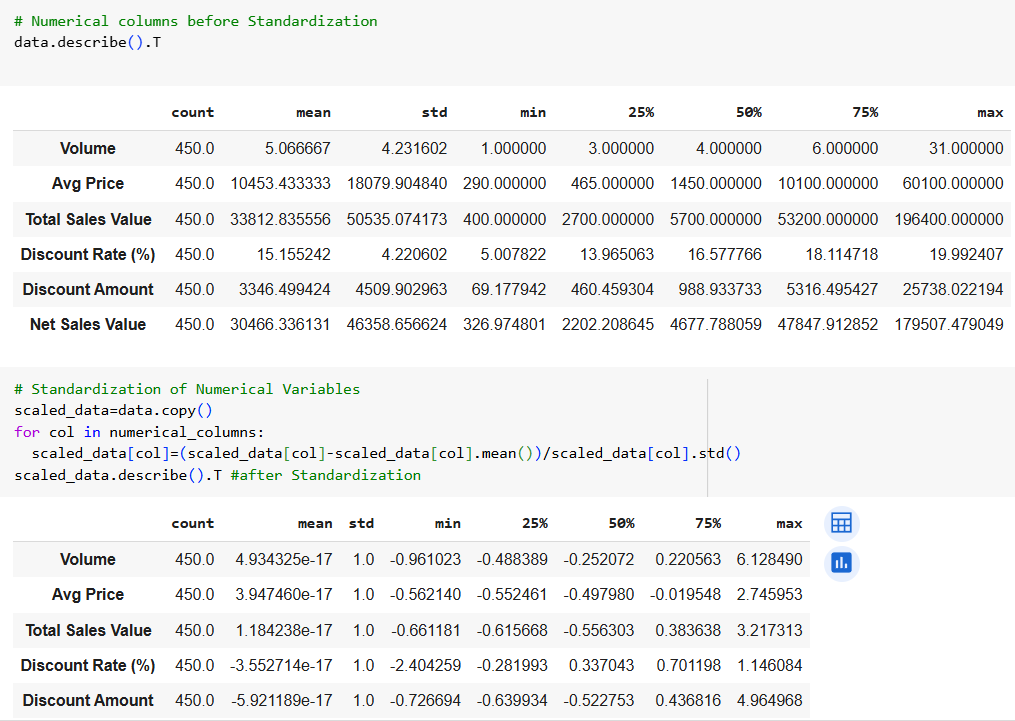
*Concept of standardization (z-score normalization):*

Standardization involves transforming raw data values to a standard scale with a mean of 0 and a standard deviation of 1.

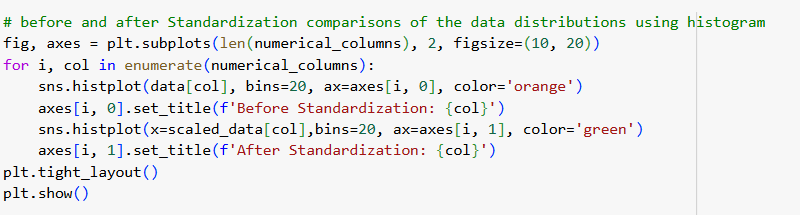
Standardization or Z-Score is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score. Formula: z=x-mu/sigma.

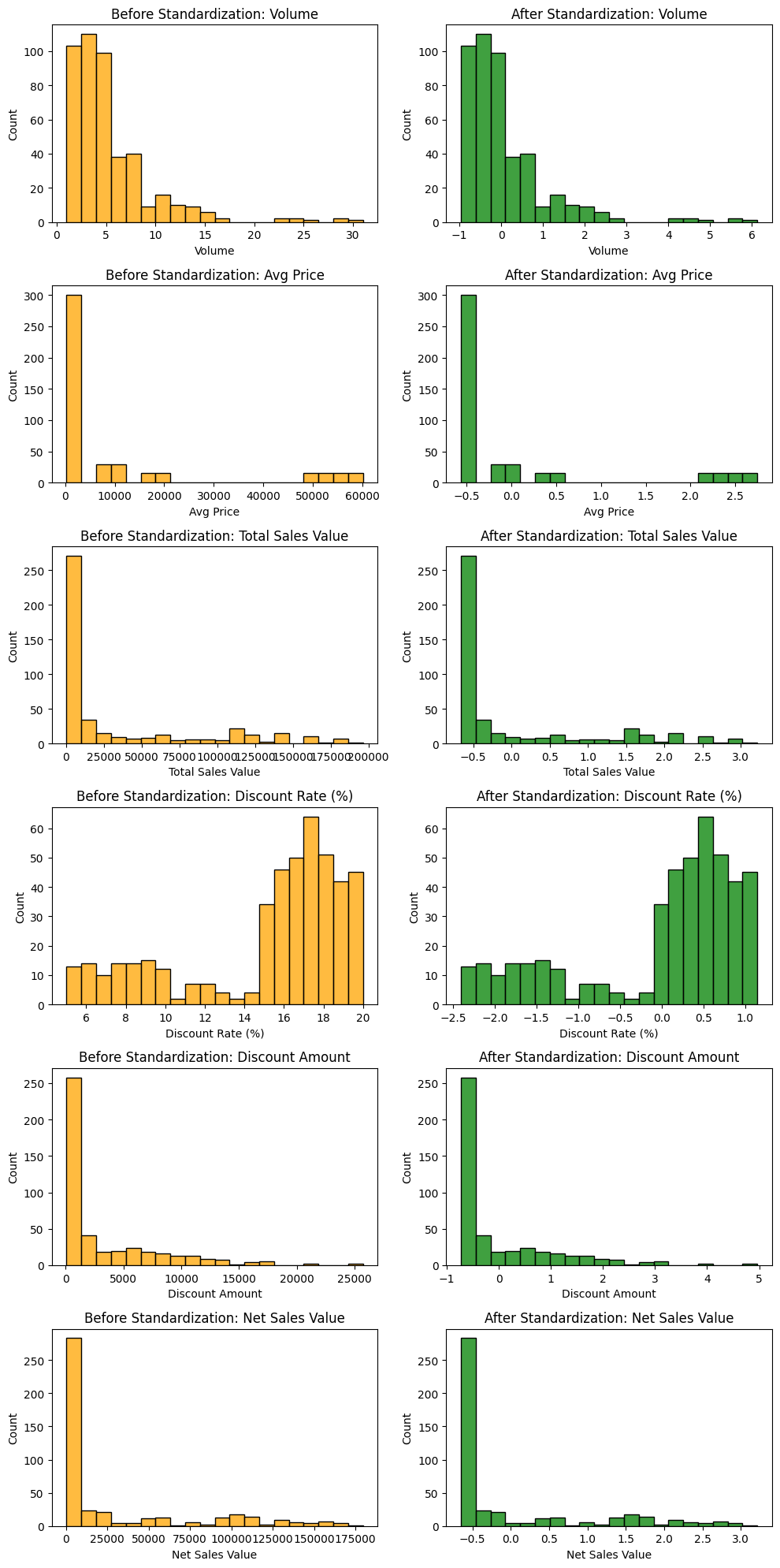
It helps in reduciong the impact of extreme values, improves model performance, and makes data more comparable across different datasets.

Python code and it’s output:



Python code to compare before and after standardization of numeric columns using histogram





#### Conversion of Categorical Data into Dummy Variables

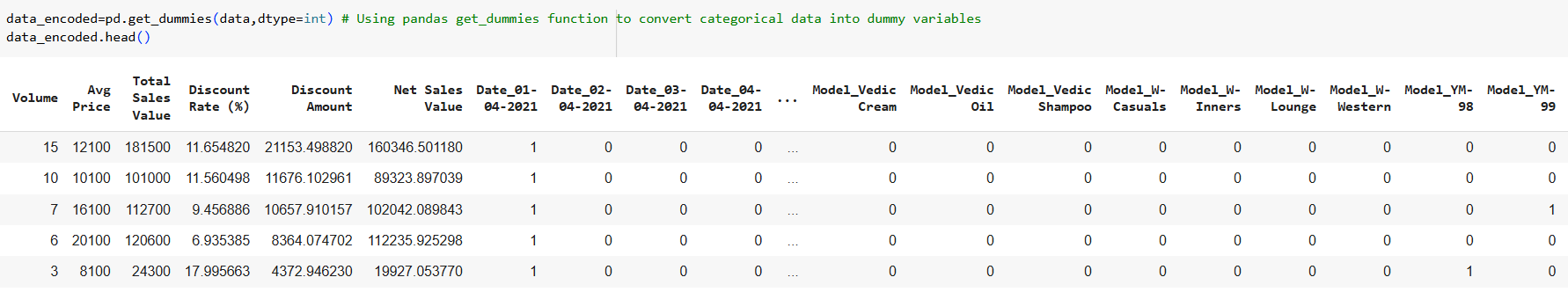
* Objective: To transform categorical variables into a format that can be provided to ML algorithms.
* Steps:
  + Discuss the need for converting categorical data into dummy variables (one-hot encoding).
  + Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each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

**Concept:**

* One Hot Encoding is a method for converting categorical variables into a binary format.
* statistical models require numeric inputs to function correctly. Hence, by transforming categorical data into numeric data using dummy variables, we can include these variables in our models. Also it improves the overall performance of machine learning models.
* python code to apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.

data\_encoded=pd.get\_dummies(data,dtype=int) # Using pandas get\_dummies function to convert categorical data into dummy variables

* 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.

Solution:

* **Descriptive Analytics:**

The process involves using various statistical and visualization techniques to describe and present data meaningfully. It Computes mean, median, mode, and standard deviation.

**Histograms and Boxplots:** these Visualizes the distributions, skewness and outliers for numerical data.

**Bar Charts:** It Shows the frequency of categorical data.

* **Key benefits of data preprocessing:**

Pre-processed data helps machine learning models perform better by ensuring the data is clean, well-organized, and relevant. By removing noise (such as errors or irrelevant features), models can focus on learning meaningful patterns, leading to improve accuracy and generalizability.

Data preprocessing steps:

**Standardization:** technique scales data to a common range to prevent features with large ranges from dominating others. It transforms data to have a mean of 0 and a standard deviation of 1.

**One-Hot Encoding:** This technique converts categorical variables into binary vectors, allowing them to be processed by models.