Basic Statistics

MULAGAPATI PADMAVATHI

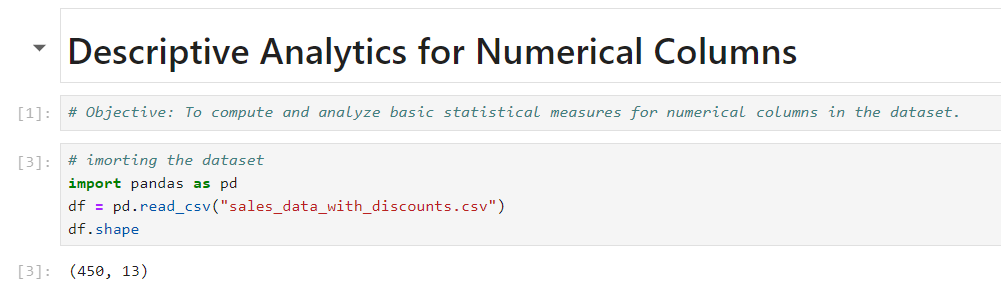
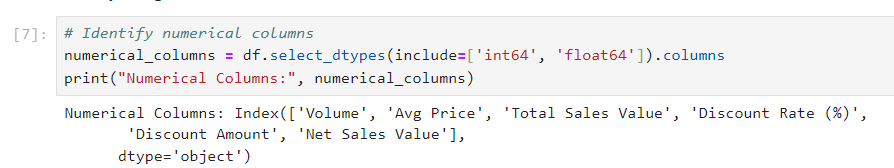
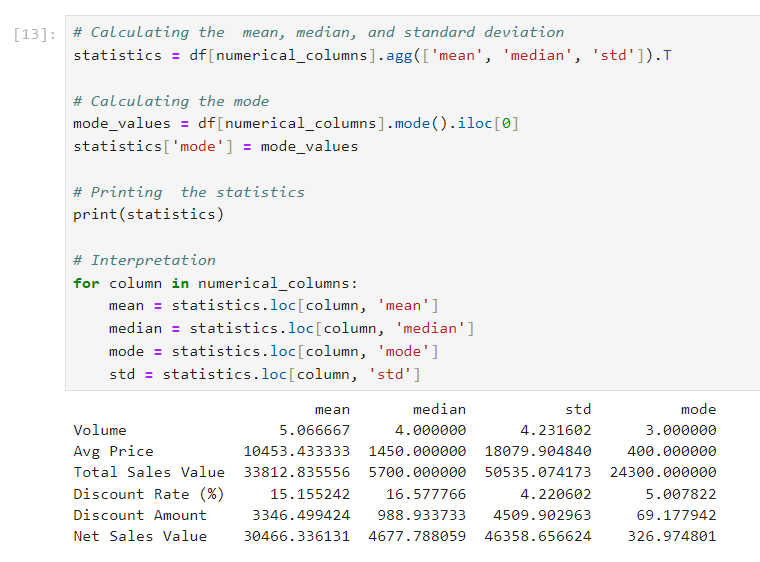
DATA SCIENCE WEEKDAY BATCH

### 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 analyse basic statistical measures for numerical columns in the dataset.
* **Steps:**
  + **1) Load the dataset into a data analysis tool or programming environment (e.g., Python with pandas library).**
  + **2) Identify numerical columns in the dataset.**
  + **3) Calculate the mean, median, mode, and standard deviation for these columns.**

**4) Provide a brief interpretation of these statistics.**

**1. Mean:** The mean represents the average value of each numerical column in the dataset. For example, if the mean of "Net Sales Value" is high, it indicates that on average, sales generate a substantial amount of revenue. However, the mean can be sensitive to extreme values or outliers, which can skew this average.

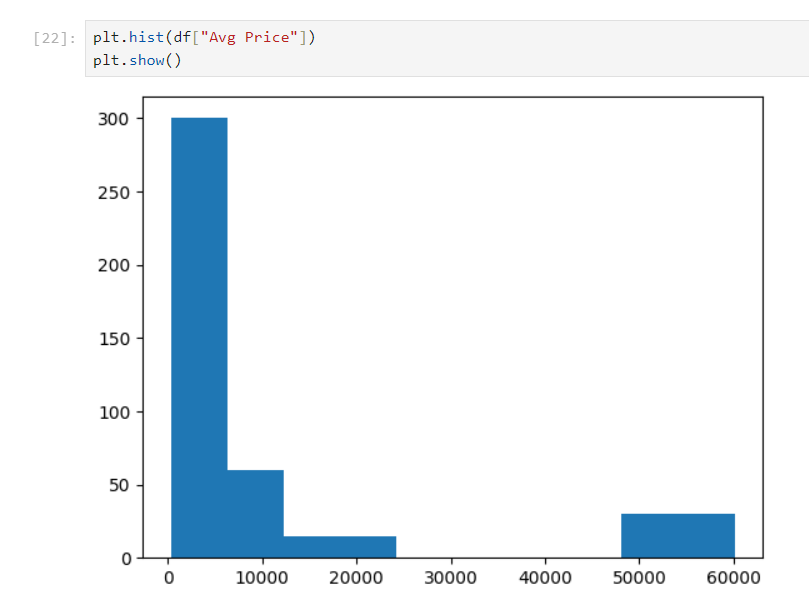
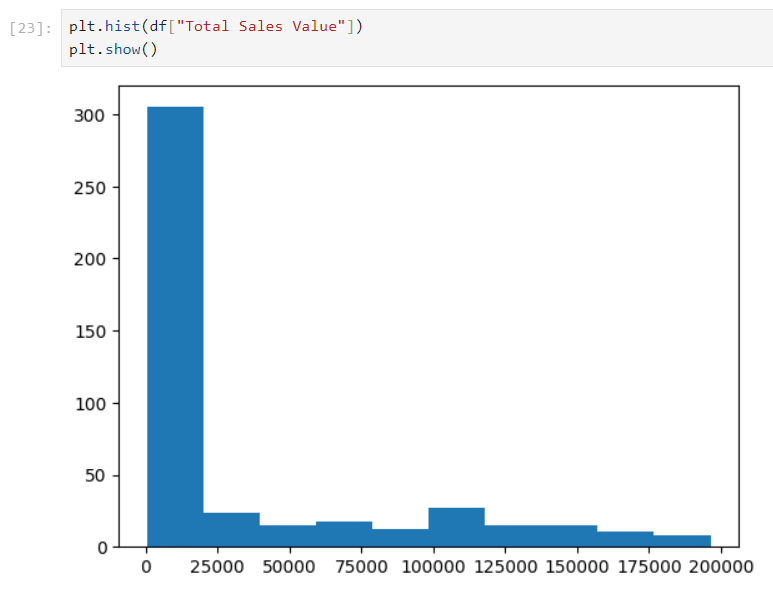
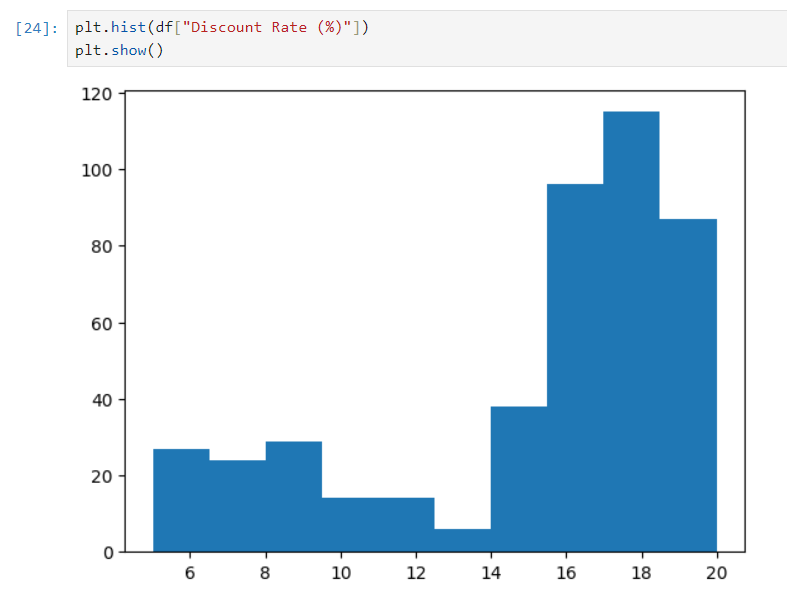
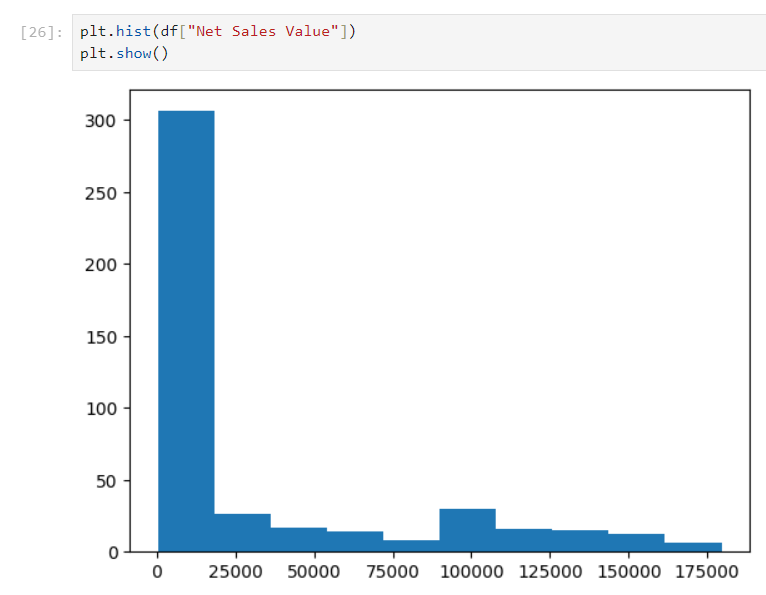
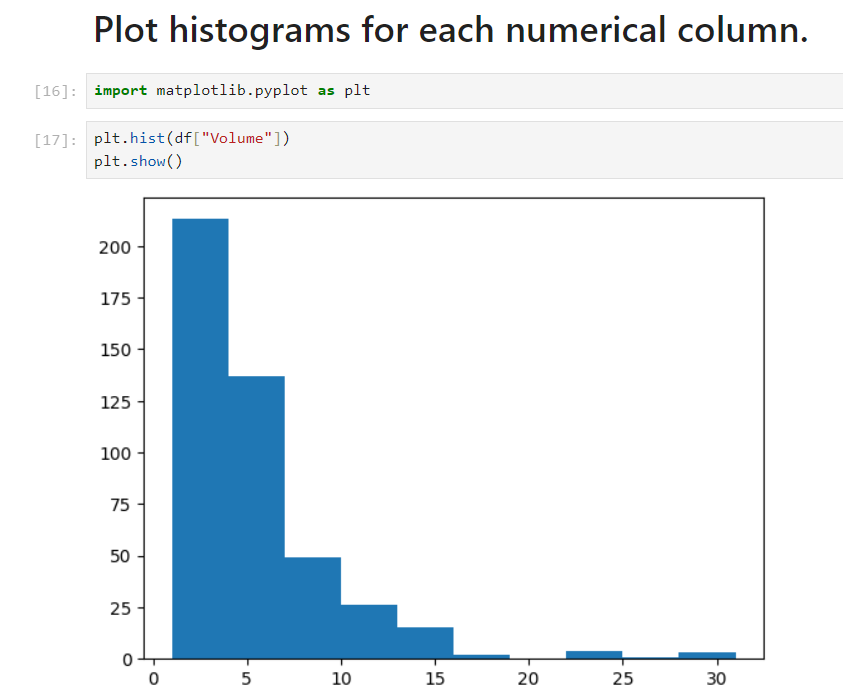
**2. Median**: The median is the middle value of each numerical column when the data is ordered. It provides a better measure of central tendency in the presence of outliers or skewed data. For instance, if the "Discount Rate" has a median lower than the mean, it suggests that most transactions have a discount rate lower than the average, possibly due to a few high-discount outliers.

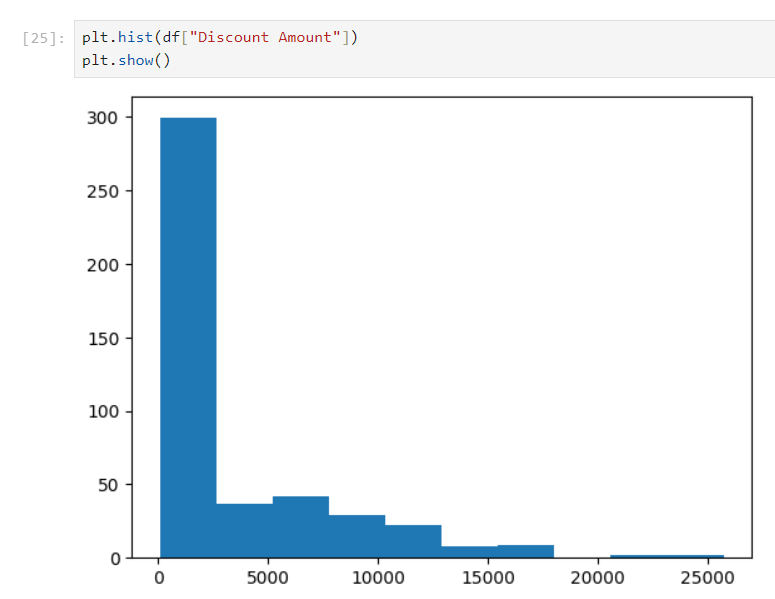
**3. Mode**: The mode is the most frequently occurring value in each numerical column. It’s particularly useful in understanding the most common data points. For instance, if the mode of "Volume" is 1, it indicates that a single unit is the most frequently sold quantity. This can help in inventory planning and sales strategy.

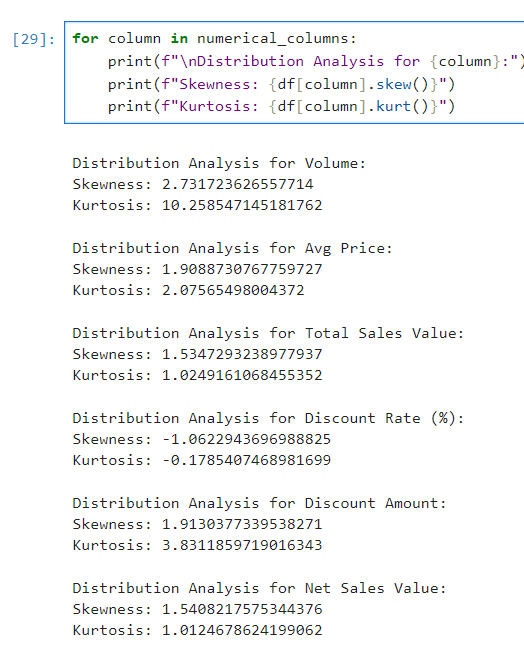
**4. Standard Deviation**: The standard deviation measures the spread of the data around the mean. A high standard deviation in columns like "Total Sales Value" or "Net Sales Value" suggests significant variability in the sales amounts, which might indicate that some transactions are much larger or smaller than others. This could point to a diverse product range or varied customer purchasing behaviour.

#### Data Visualization

* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:

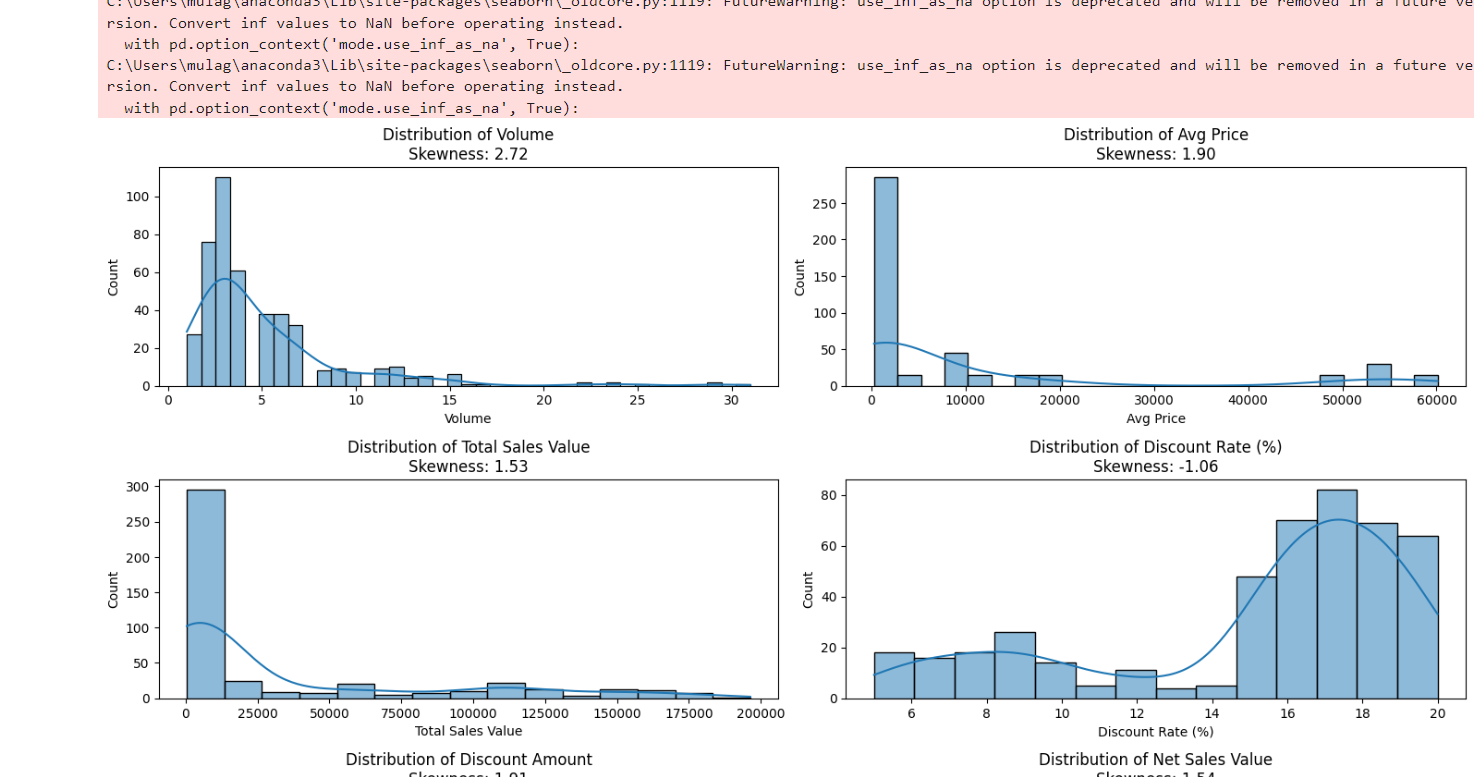
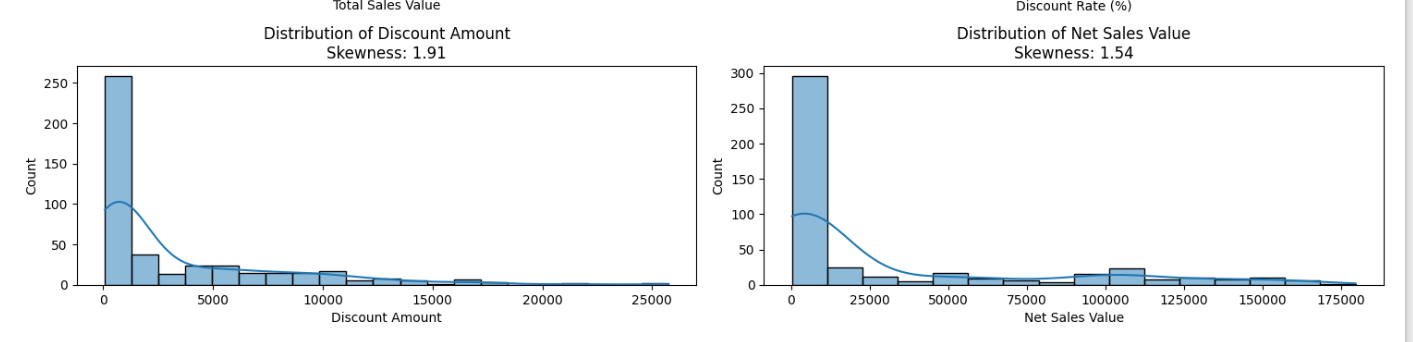
**1) Plot histograms for each numerical column.******

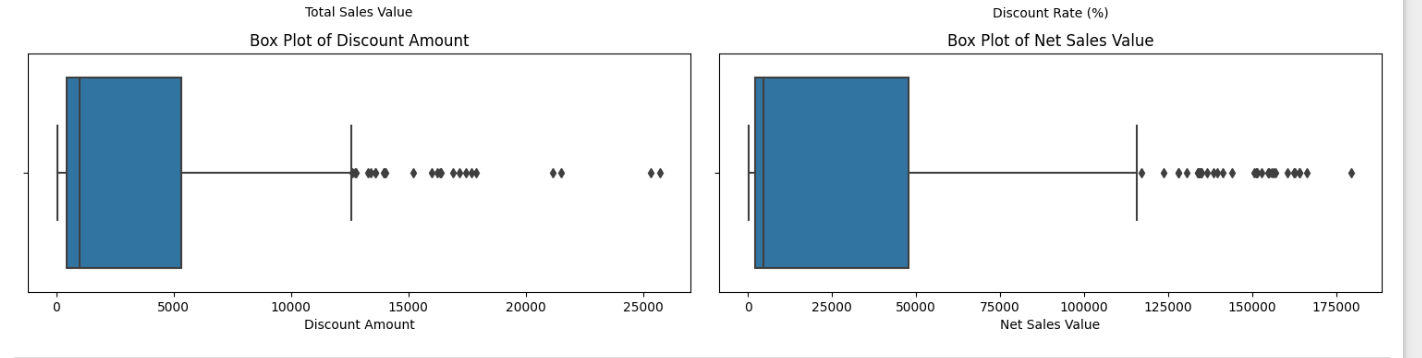
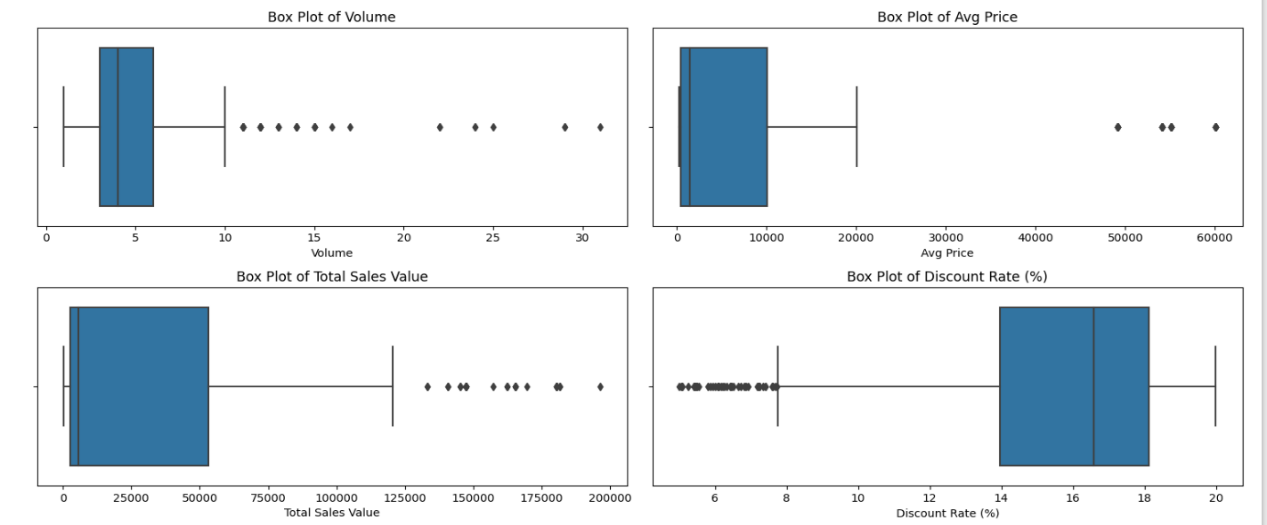


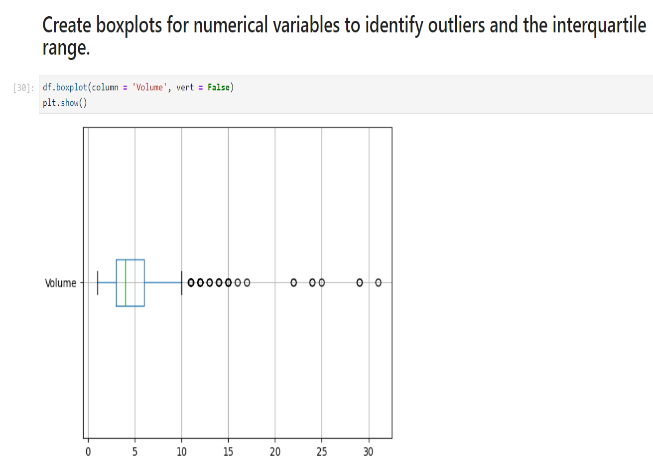
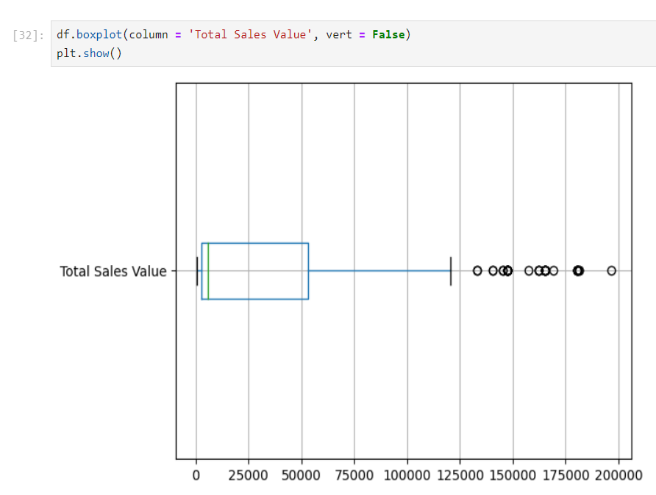
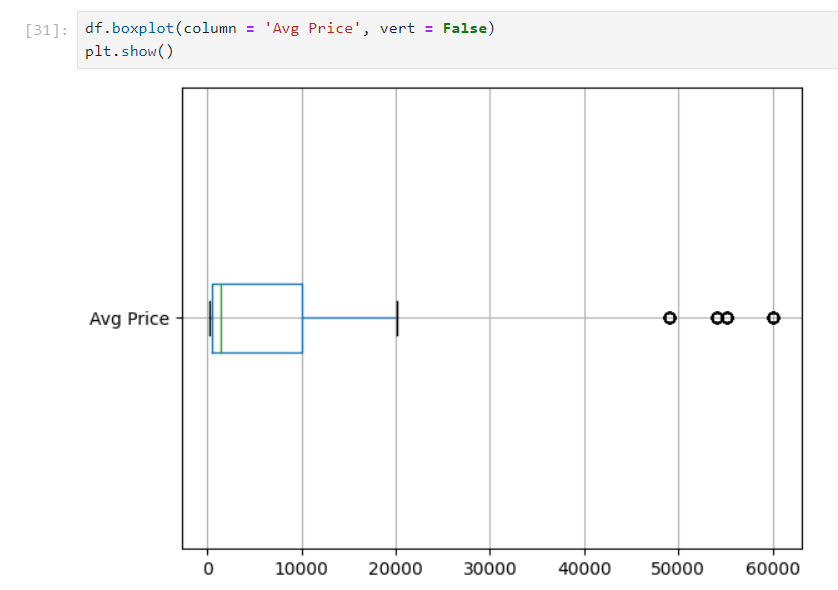
******2) Analyse the distribution (e.g., skewness, presence of outliers) and provide inferences.**

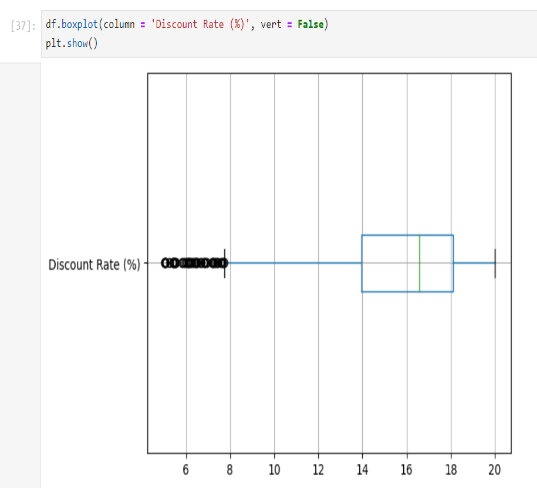
1) **Positive Skewness**: If the skewness value is greater than 0, the data is right-skewed (long tail on the right).

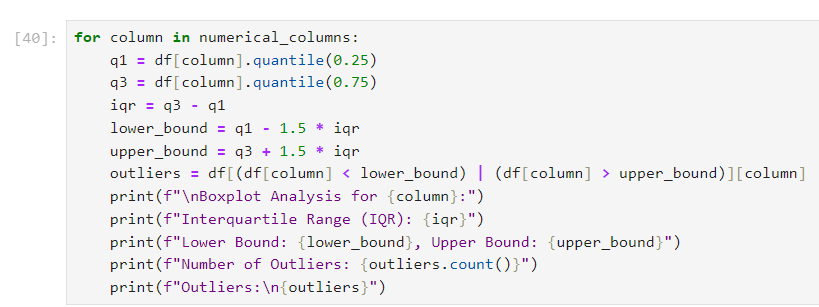
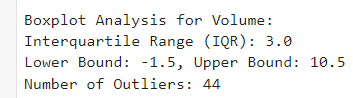
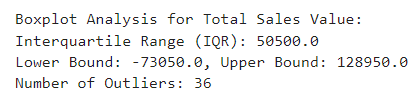
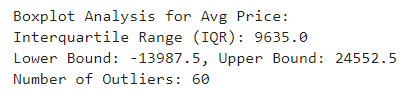
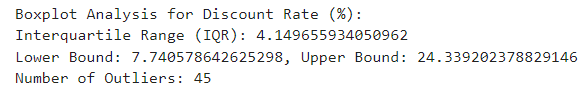
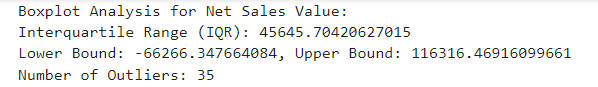
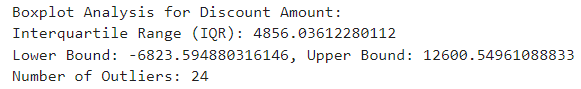
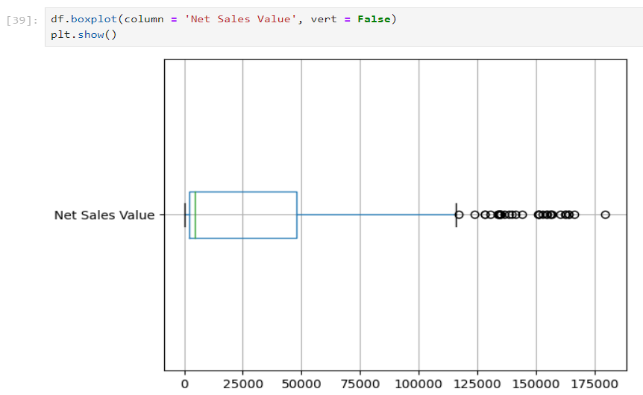
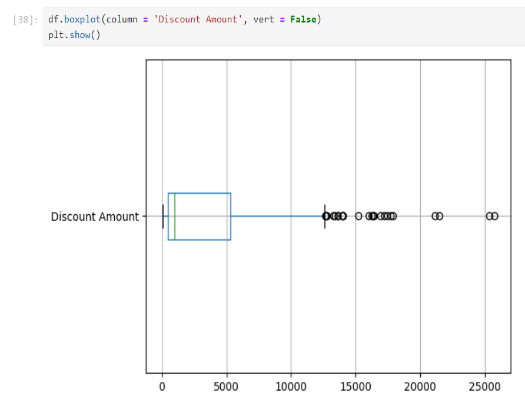
2) **Negative Skewness**: If the skewness value is less than 0, the data is left-skewed (long tail on the left).

3) **Symmetric**: If the skewness is around 0, the data is approximately symmetric.

* **Boxplots**

1. **Create boxplots for numerical variables to identify outliers and the interquartile range.**

****

****

1. **Discuss any findings, such as extreme values or unusual distributions.**

**1. Volume:**

- Distribution: The Volume distribution is highly right-skewed, indicating that most transactions involve lower volumes, with fewer instances of high-volume sales.

- Outliers: Significant outliers are present, suggesting that certain transactions had unusually high sales volumes compared to the rest of the dataset. These could be bulk purchases or special promotions.

**2. Average Price:**

- Distribution: The average price shows a slight right skew, indicating that lower prices are more common, but there's a tail towards higher prices.

- Outliers: Some outliers exist, particularly on the higher end of the price spectrum. This could indicate premium products or errors in pricing.

**3. Total Sales Value:**

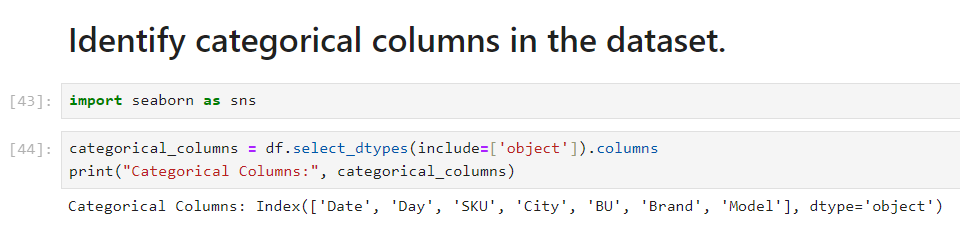
- Distribution: The Total Sales Value is also right-skewed, consistent with the fact that higher total sales values are less frequent.

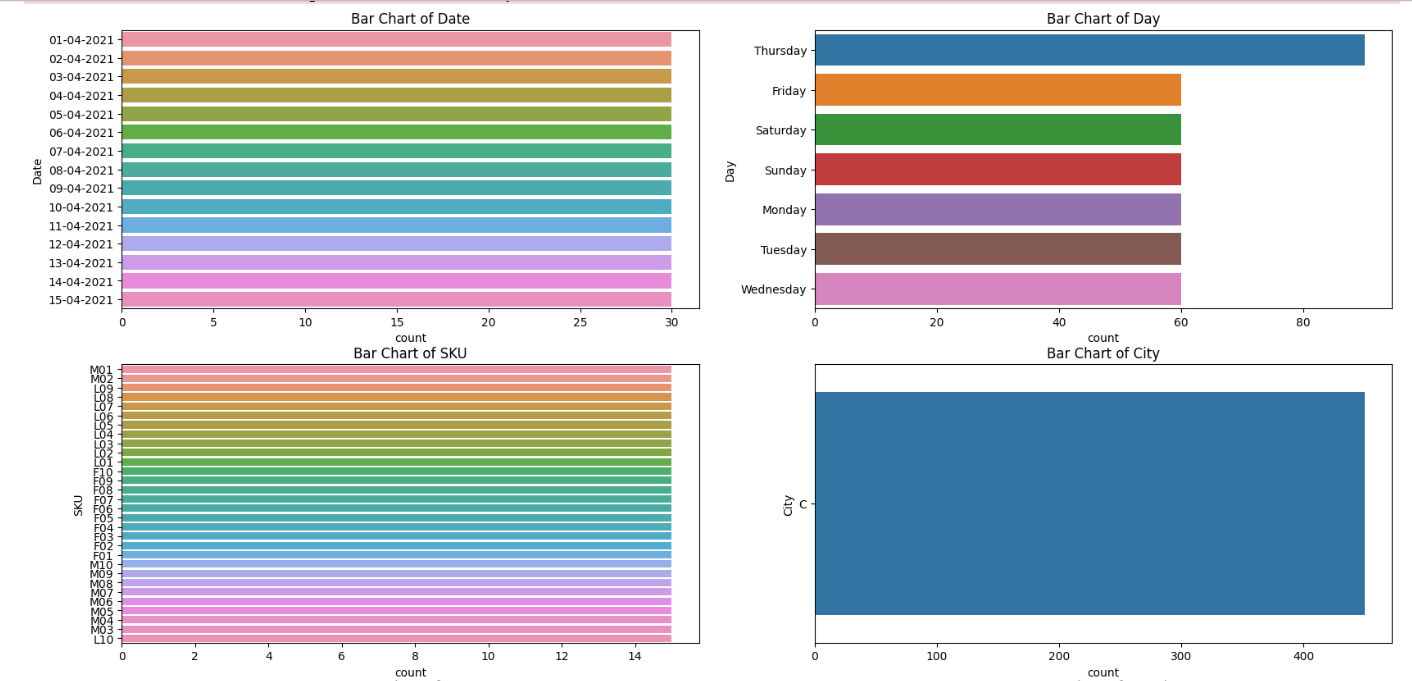
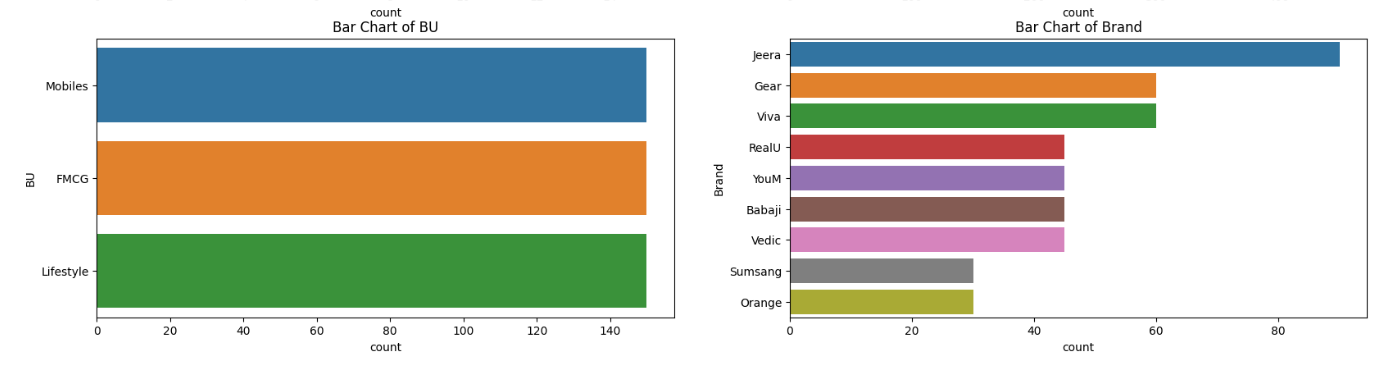
- Outliers: The presence of outliers here suggests transactions with exceptionally high total sales values, likely linked to high volume and/or high-priced products.

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

- Distribution: The discount rate distribution is left-skewed, meaning that higher discount rates are more common. This could be due to frequent promotions or sales.

- Outliers: The outliers are mainly at the lower end, indicating some transactions with very low or no discounts, which might be for high-demand.

* **Bar Chart Analysis for Categorical Column:**
  + **1) Identify categorical columns in the dataset.**
  + **2) Create bar charts to visualize the frequency or count of each category. Analyse the distribution of categories and provide insights.**



**1. City:**

- Certain cities may dominate sales, contributing significantly more compared to others.

- Cities with fewer sales could either be less populous or less targeted in promotions.

**2. Day:**

- Sales distribution across days can reveal patterns, like higher sales on weekends or specific weekdays. This may highlight peak sales periods.

**3. Brand/Model:**

- Some brands and models likely have much higher sales frequencies, indicating popular or flagship products.

- Lesser-known brands or models may show fewer sales, indicating niche or lower-demand items.

**4. BU (Business Unit):**

- One business unit (e.g., Mobiles) could dominate the dataset, suggesting a focus on a particular product category.

#### **Standardization of Numerical Variables**

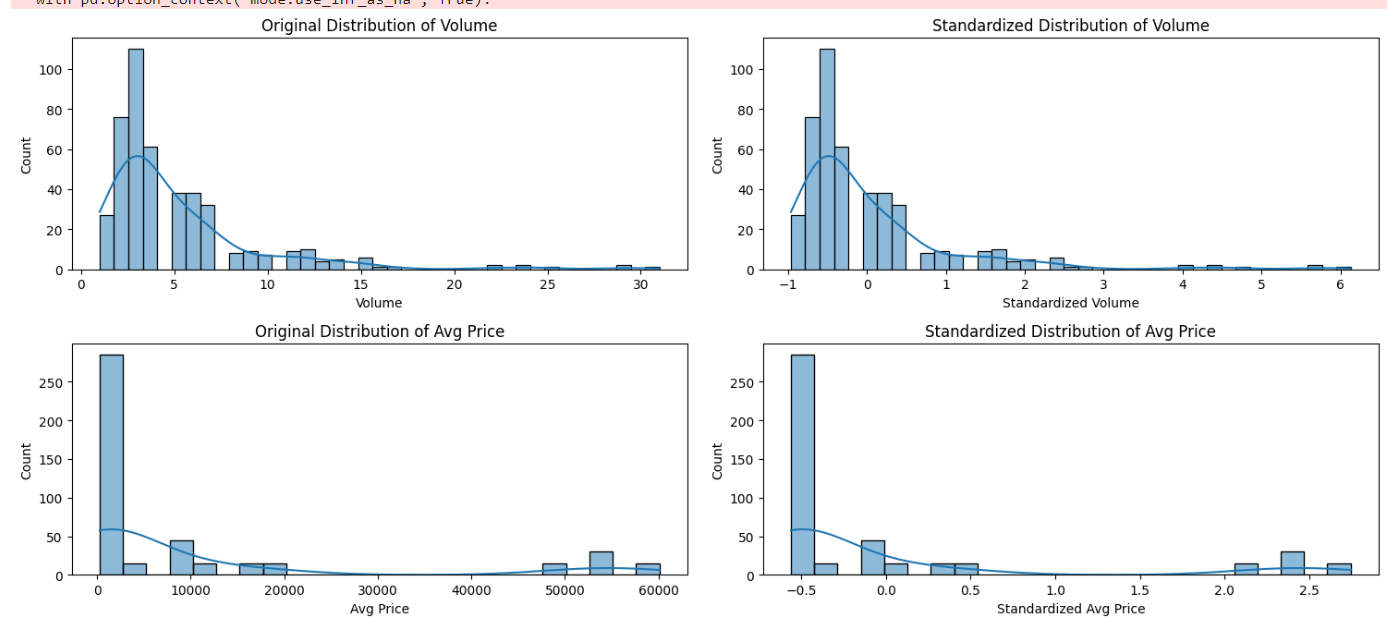
* **Objective:** To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* **Steps:**
  + **1) Explain the concept of standardization (z-score normalization).**
  + **Standardize the numerical columns using the formula: z=x-mu/sigma**

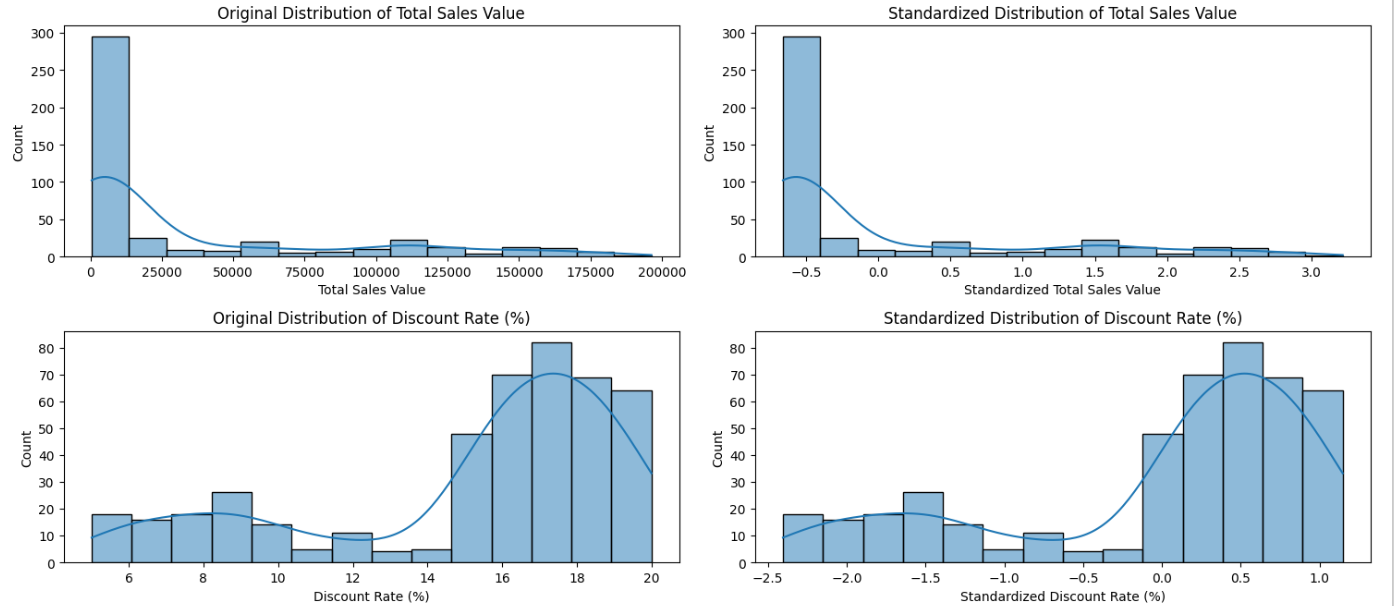
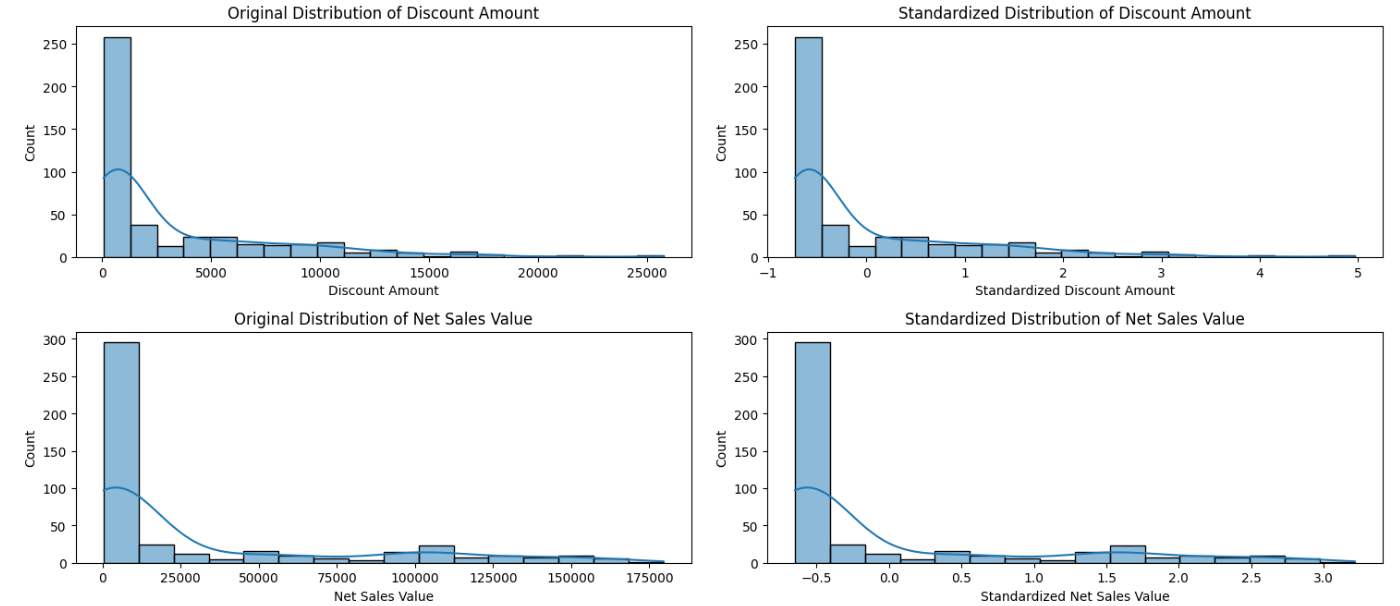
**Standardization** (also called Z-score normalization) transforms numerical data to have a mean of 0 and a standard deviation of 1.

This ensures that the numerical variables are on the same scale, making them more suitable for analytical models (especially distance-based models like k-NN, clustering, etc.).

**Formula**: The Z-score formula is: z=x−μ/σ

Where:

* x is the original data point,
* μ is the mean of the variable,
* σ is the standard deviation of the variable,
* z is the standardized value.



#### **Conversion of Categorical Data into Dummy Variables**

* **Objective:** To transform categorical variables into a format that can be provided to ML algorithms.
* **Steps:**
  + **1) Discuss the need for converting categorical data into dummy variables (one-hot encoding).**

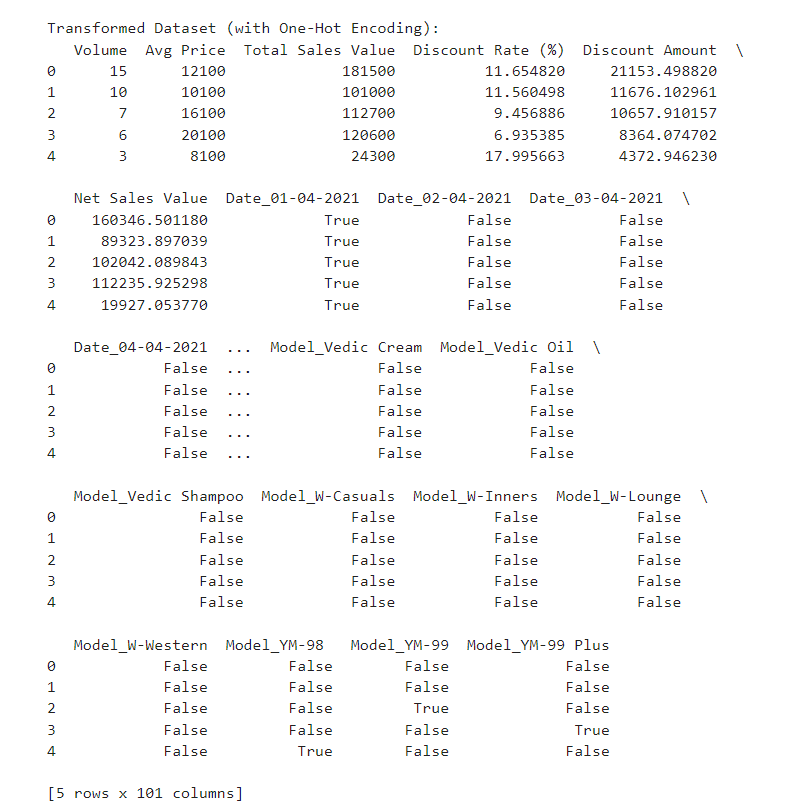
Machine learning algorithms generally require numerical inputs. Categorical data, especially non-ordinal categories (e.g., city names, product brands), need to be transformed into a numerical format.

* **One-hot encoding** is a technique used to convert categorical variables into a series of binary columns (0 or 1). Each category of the original categorical variable is turned into a separate column where:
  + **1** indicates the presence of that category.
  + **0** indicates its absence.

This process ensures that the ML algorithm can interpret the categorical data without making incorrect assumptions about the order or relationship between categories.

**2) Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.**

**Display a portion of the transformed dataset.**



#### Conclusions

* 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.
* **Key Findings from Descriptive Analytics and Data Visualizations:**

**1. Numerical Variables:**

- Most of the numerical variables, such as Volume, Total Sales Value, and Discount Amount, were right-skewed, indicating that the majority of sales and discounts were smaller, with only a few large transactions.

- Outliers were present in almost every numerical variable, especially in Volume and Total Sales Value, pointing to some unusually large transactions.

- After standardization, the numerical variables were scaled to have a mean of 0 and a standard deviation of 1, making them more suitable for machine learning models.

**2. Categorical Variables:**

- City: Some cities had significantly more sales than others, likely major markets or focus areas.

- Day: There may be trends related to higher sales on certain days of the week (e.g., weekends).

- Brand/Model: Certain brands or models were much more frequently sold, suggesting these are flagship products.

- After one-hot encoding, categorical columns were converted into binary features, allowing them to be utilized in machine learning models.

* **Importance of Data Preprocessing in Data Analysis and Machine Learning:**

**1. Standardization:** Ensures that numerical variables are on the same scale, which is crucial for algorithms like k-NN, SVMs, and gradient descent-based models that are sensitive to the magnitude of the input data.

- Impact: Helps the model converge faster and perform better since all features contribute equally to the prediction, rather than features with larger values dominating.

**2. One-Hot Encoding:** Transforms categorical variables into a format that can be used by machine learning models. Since most algorithms can't handle non-numeric data, this step is essential.

- Impact: Converts categories into binary columns, allowing the model to recognize distinct classes without assuming any ordinal relationship between them. This ensures that models treat categorical variables as separate and unrelated.