BASIC STATISTICS ASSIGNMENT

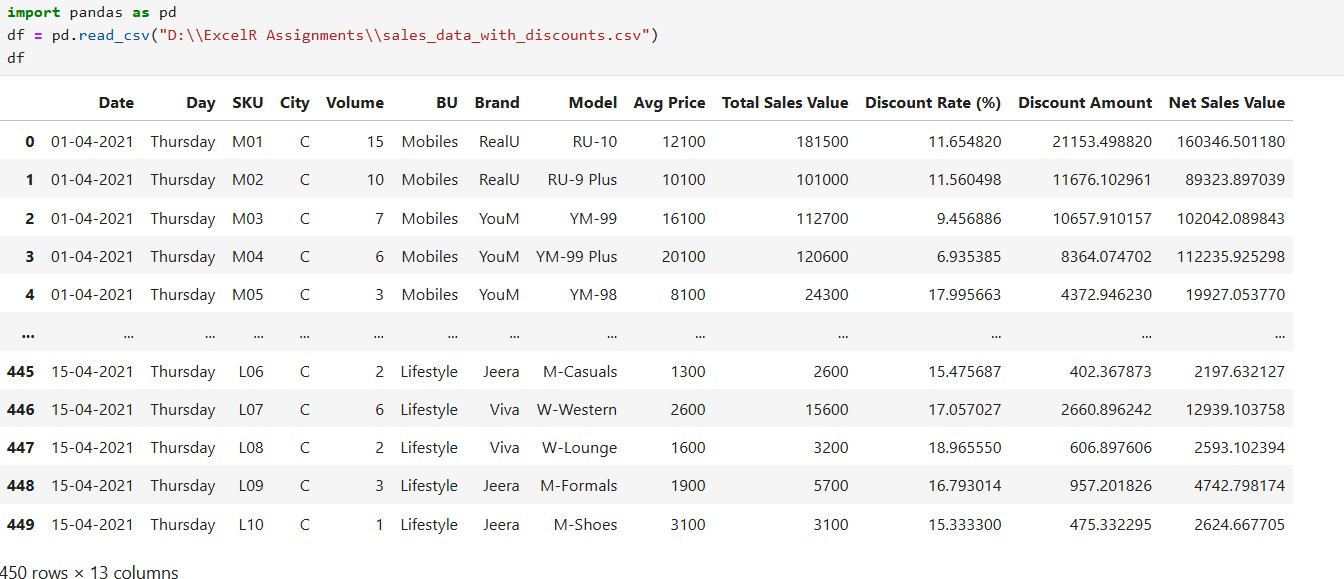
# INTRODUCTION

This analysis aims to perform descriptive analytics and data preprocessing on a given dataset related to sales and discounts. We will compute basic statistical measures for numerical columns, visualize data distributions, and standardize numerical variables and convert categorical data into dummy variables.

# DESCRIPTIVE ANALYTICS FOR NUMERICAL COLUMNS

**Step 1:**

Loading the dataset into a data analysis tool or programming environment. The dataset is in a CSV format, we can load it into a Python environment using pandas.



# Step 2:

The numerical columns in the dataset are:

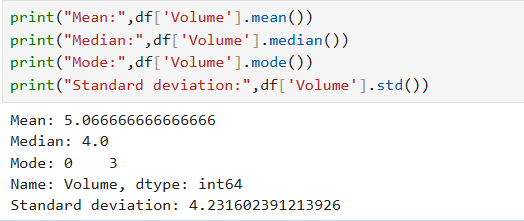
* Volume
* Avg Price
* Total Sales
* Discount Rate
* Discount Amount
* Net Sales Value

# Step 3:

Calculating Statistical Measures.

# Volume:

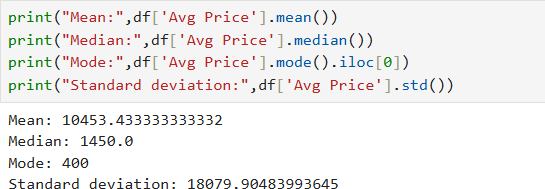
Calculating the mean, median, mode and standard deviation for the volume.



The average number of items sold is around 5, with a median of 4. The standard deviation is relatively high, indicating a spread in the number of items sold.

# Avg Price:

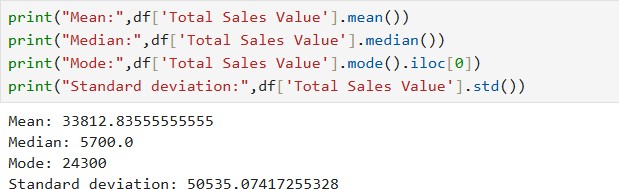
Calculating the mean, median, mode and standard deviation for the Avg Price.



The average price is around 10453 with a median of 1450 and with mode of 400. The standard deviation is again high, suggesting variation in prices.

# Total Sales Value:

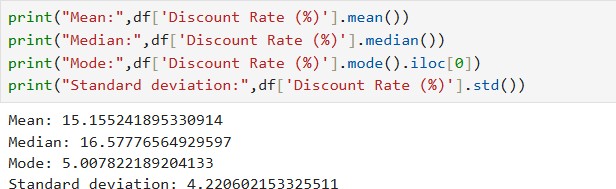
Calculating the mean, median, mode and standard deviation for the Total Sales Value.



The average total sales value are around 33,812 with a median of 5700. The standard deviation is high, indicating a wide range of total sales values.

# Discount Rate:

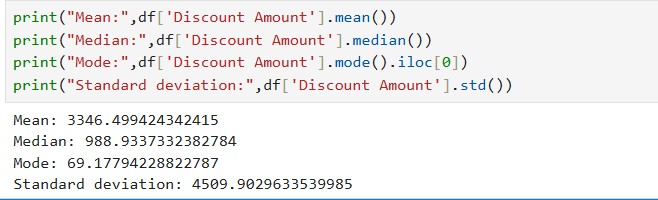
Calculating the mean, median, mode and standard deviation for the Dicount Rate.



The average and median discounts are around 15% and 16% respectively, with moderate standard deviations.

# Discount Amount:

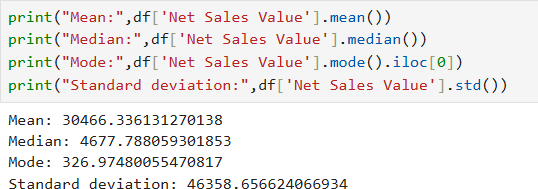
Calculating the mean, median, mode and standard deviation for the Dicount Amount.



The average Discount Amount is around 3346 with a median of 988. The standard deviation is moderate, indicating a medium range of Discount Amount.

# Net Sales Value:

Calculating the mean, median, mode and standard deviation for the Net Sales Value.

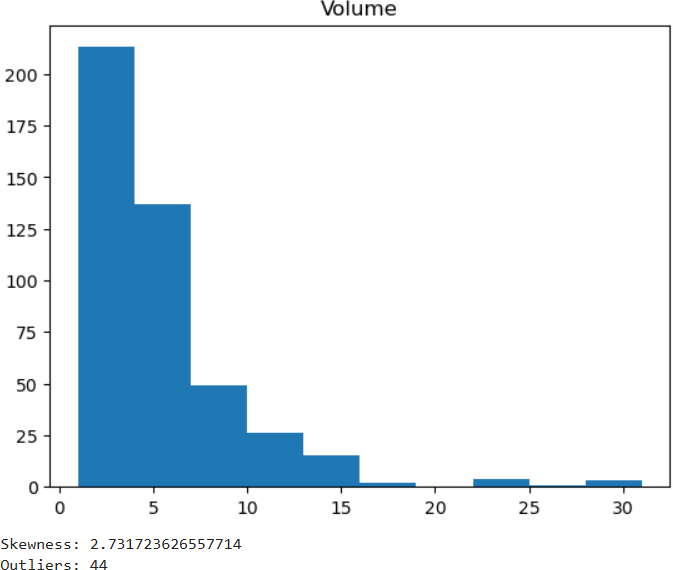
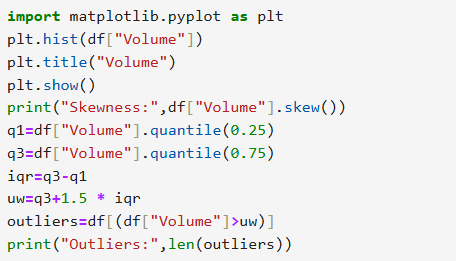


The average net sales value is around 30466 with a median of 4677. The standard deviation is high, reflecting variability in net sales values.

# DATA VISUALIZATION

**Histograms 1.Volume:**

Showing the histogram for volume.

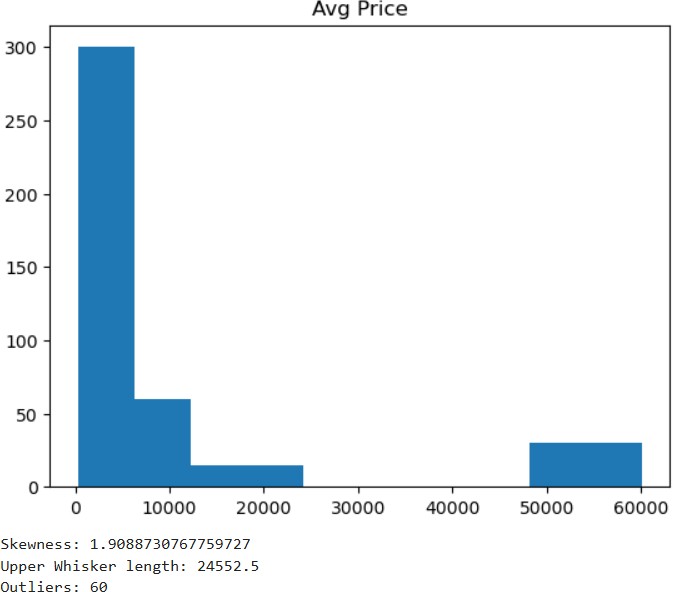


The histogram for Volume of Sales and Discounts dataset is shown above. The Volume dataset contains the skewness around 2.73 which is positively skewed. So the outliers lies on the right side of the upper whisker length. The calculated upper whisker length for the above histogram is 10.5.There are totally 44 outliers in the Volume of Sales and Discounts dataset.

# Avg Price:

Showing the histogram for Avg Price.

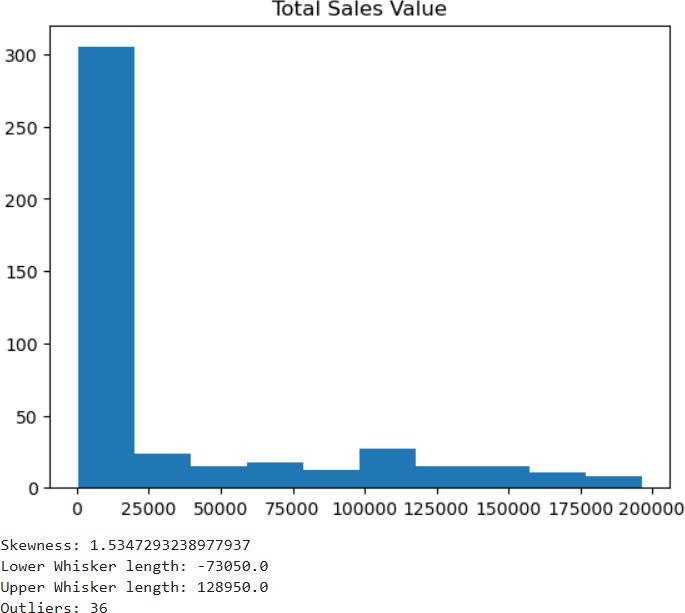
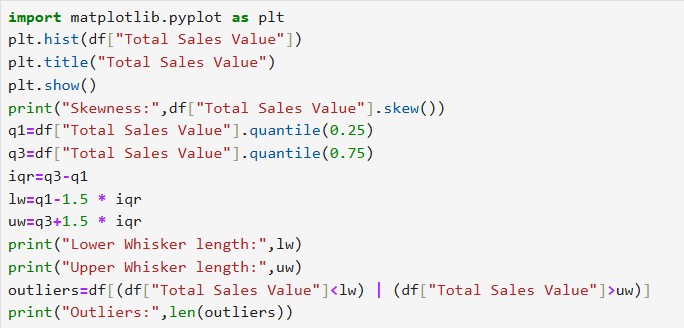




The histogram for Avg Price of Sales and Discounts dataset is shown above. The Avg Price dataset contains the skewness around 1.90 which is positively skewed. So the outliers lies on the right side of the upper whisker length. The calculated upper whisker length for the above histogram is 24552.5 .So the avg price which exceeds 24552.5 are the outliers. There are totally 60 outliers in the Avg Price of Sales and Discounts dataset.

# Total Sales Value:

Showing the histogram for Total Sales Value.

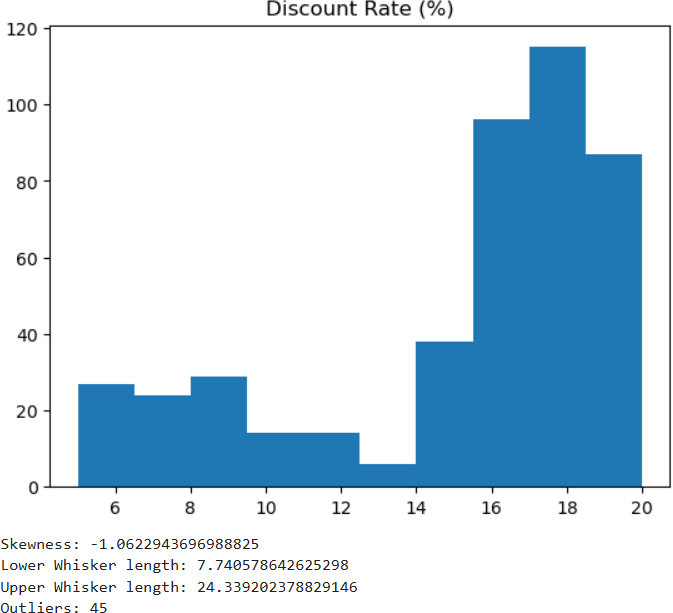


The histogram for Total Sales Value of Sales and Discounts dataset is shown above. The Total Sales Value dataset contains the skewness around 1.53 which is positively skewed. So the outliers lies on the right side of the upper whisker length. The calculated upper whisker length for the above histogram is 128950.So the Total Sales Value which exceeds 128950 are the outliers. There are totally 36 outliers in the Total Sales Value of Sales and Discounts dataset.

# Discount Rate:

Showing the histogram for Discount Rate.

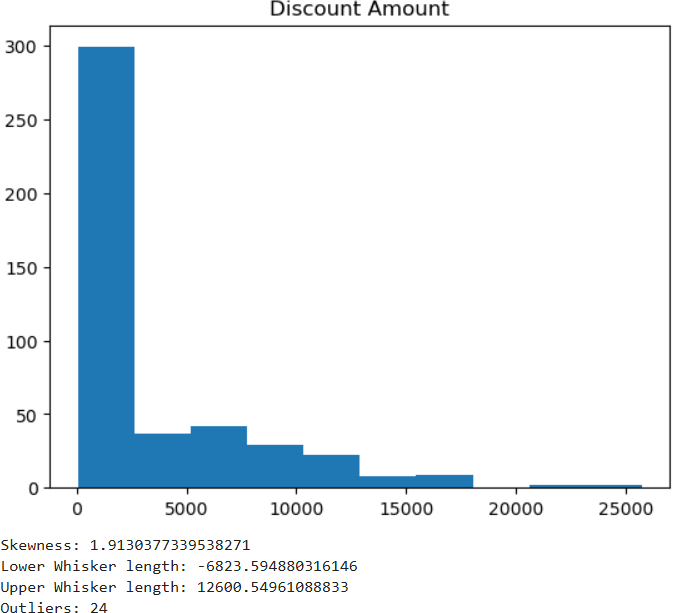
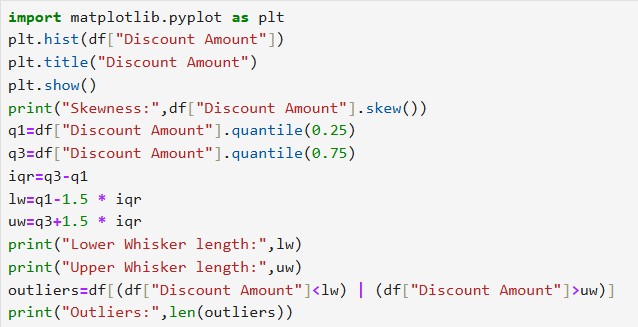




The histogram for Discount Rate of Sales and Discounts dataset is shown above. The Discount Rate dataset contains the skewness around -1.06 which is negatively skewed. So the outliers lies on the left side of the Lower whisker length. The calculated lower whisker length for the above histogram is 7.7405. So the Discount Rate which goes below 7.7405 are the outliers. There are totally 45 outliers in the Discount Rate of Sales and Discounts dataset.

# Discount Amount:

Showing the histogram for Discount Amount.

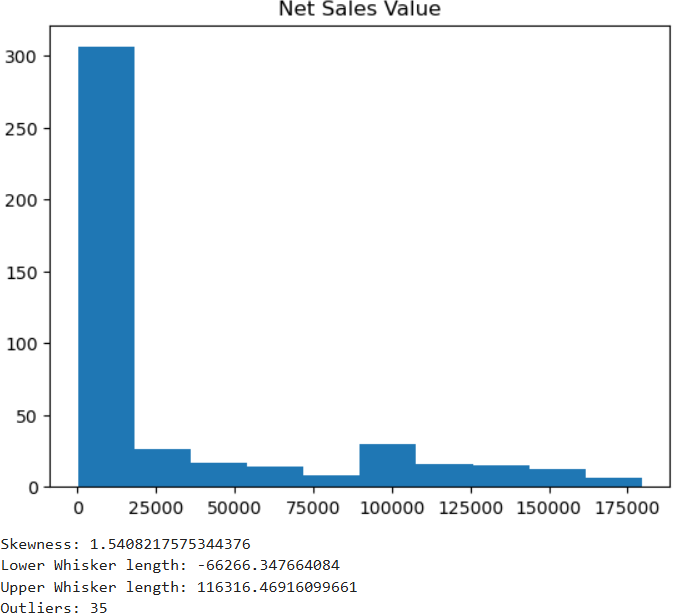


The histogram for Discount Amount of Sales and Discounts dataset is shown above. The Discount Amount dataset contains the skewness around 1.91 which is positively skewed. So the outliers lies on the right side of the upper whisker length. The calculated upper whisker length for the above histogram is 12600.54

.So the Total Sales Value which exceeds 12600.54 are the outliers. There are totally 24 outliers in the Discount Amount of Sales and Discounts dataset.

# Net Sales Value:

Showing the histogram for Net Sales Value.

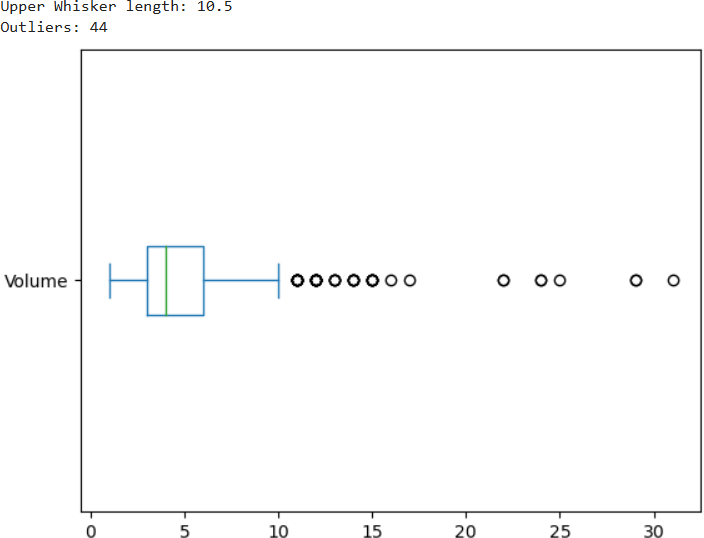
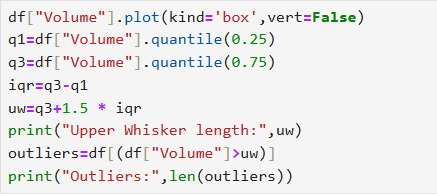


The histogram for Net Sales Value of Sales and Discounts dataset is shown above. The Net Sales Value dataset contains the skewness around 1.54 which is positively skewed. So the outliers lies on the right side of the upper whisker length. The calculated upper whisker length for the above histogram is 116316.So the Net Sales Value which exceeds 116316 are the outliers. There are totally 35 outliers in the Net Sales Value of Sales and Discounts dataset.

# Boxplots:

1. **Volume:**

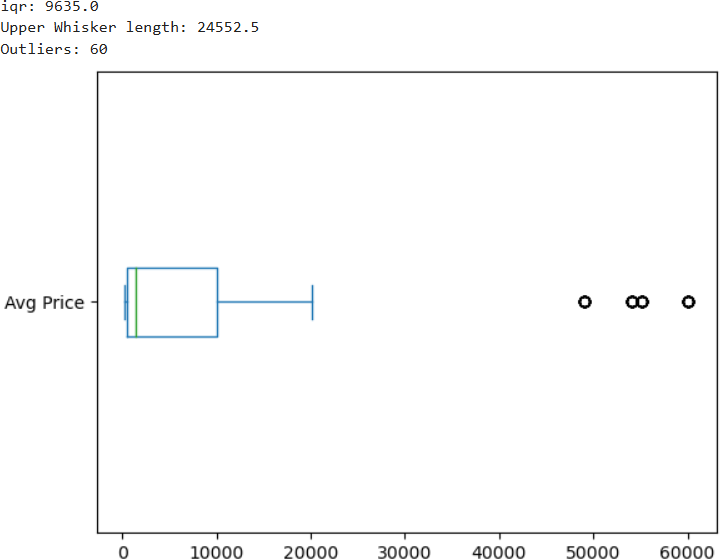
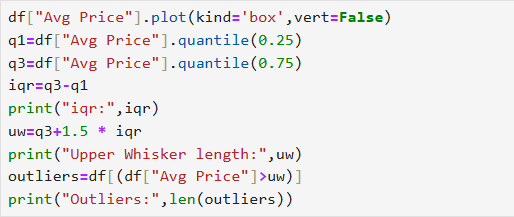
Showing the boxplot for volume.



The boxplot for Volume of Sales and Discounts dataset is shown above. The outliers are on the right side of the upper whisker length. The calculated upper whisker length for the above boxplot is 10.5.There are totally 44 outliers in the Volume of Sales and Discounts dataset.

# Avg Price:

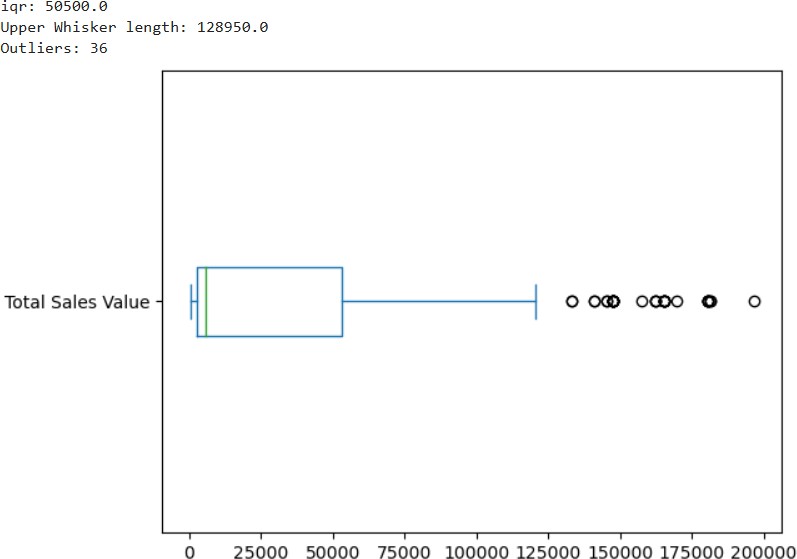
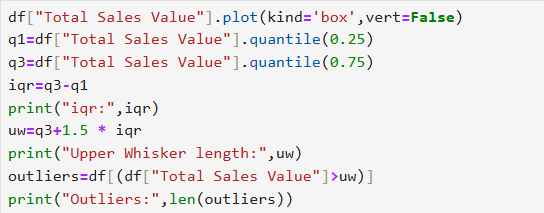
Showing the boxplot for Avg Price.



The boxplot for Avg price of Sales and Discounts dataset is shown above. The outliers are on the right side of the upper whisker length. The calculated upper whisker length for the above boxplot is 24552.5.There are totally 60 outliers in the Avg price of Sales and Discounts dataset.

# Total Sales Value:

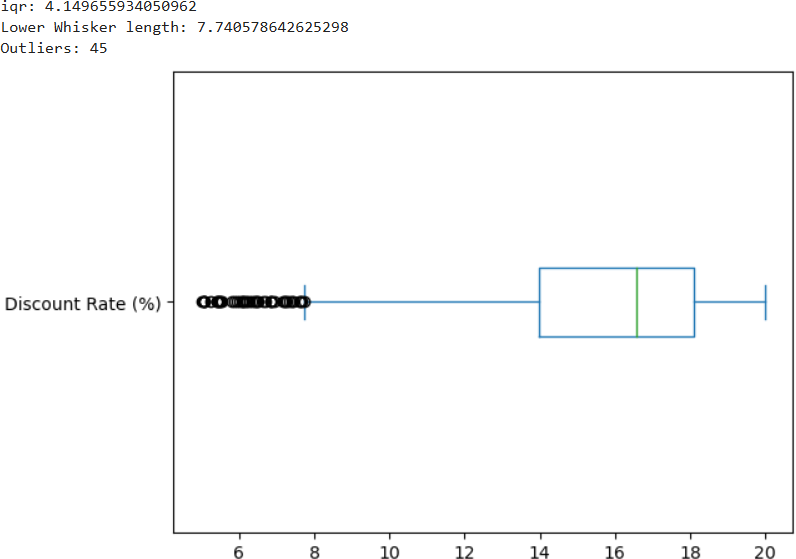
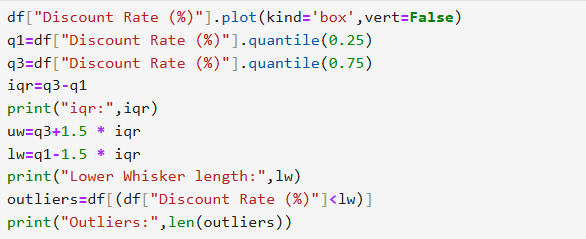
Showing the boxplot for Total Sales Value.



The boxplot for Total Sales Value of Sales and Discounts dataset is shown above. The outliers are on the right side of the upper whisker length. The calculated upper whisker length for the above boxplot is 128950.There are totally 36 outliers in the Total Sales Value of Sales and Discounts dataset.

# Discount Rate:

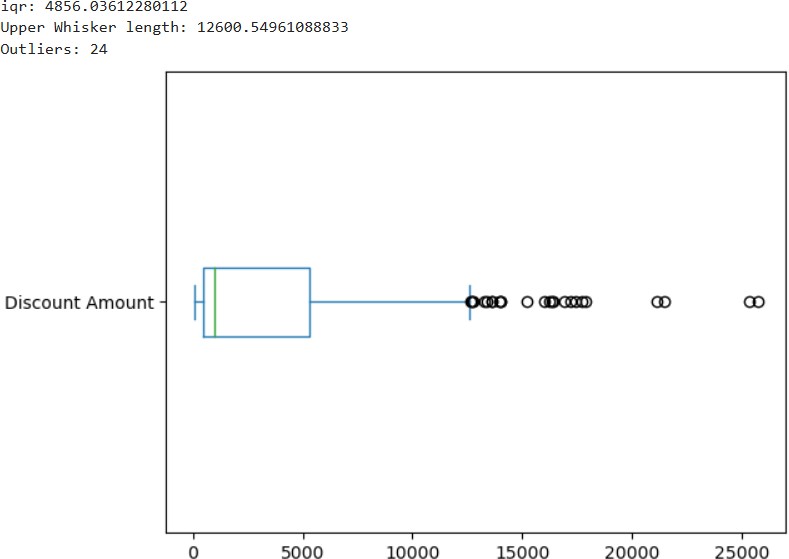
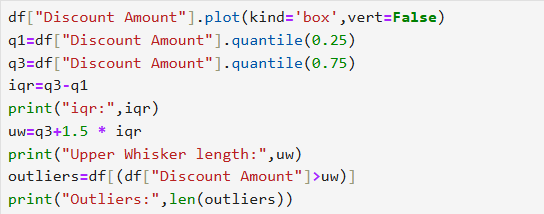
Showing the boxplot for Discount Rate.



The boxplot for Discount Rate of Sales and Discounts dataset is shown above. The outliers are on the left side of the lower whisker length. The calculated lower whisker length for the above boxplot is 7.7405. There are totally 45 outliers in the Discount Rate of Sales and Discounts dataset.

# Discount Amount:

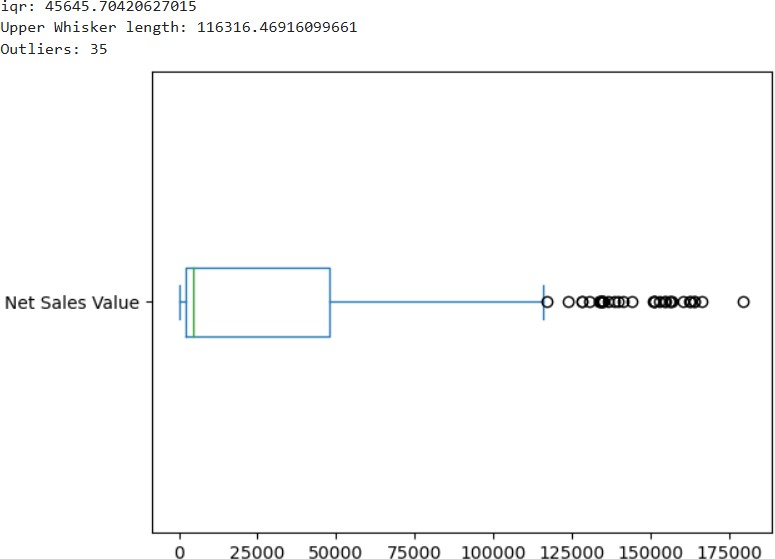
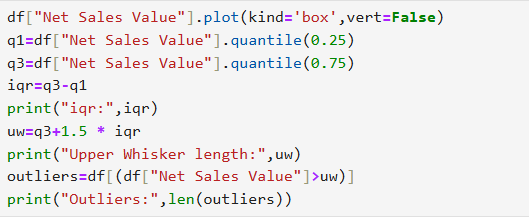
Showing the boxplot for Discount Amount.



The boxplot for Discount Amount of Sales and Discounts dataset is shown above. The outliers are on the right side of the upper whisker length. The calculated upper whisker length for the above boxplot is 12600.5. There are totally 24 outliers in the Discount Amount of Sales and Discounts dataset.

# Net Sales Value:

Showing the boxplot for Net Sales Value.



The boxplot for Net Sales Value of Sales and Discounts dataset is shown above. The outliers are on the right side of the upper whisker length. The calculated upper whisker length for the above boxplot is 116316.4. There are totally 35 outliers in the Net Sales Value of Sales and Discounts dataset.

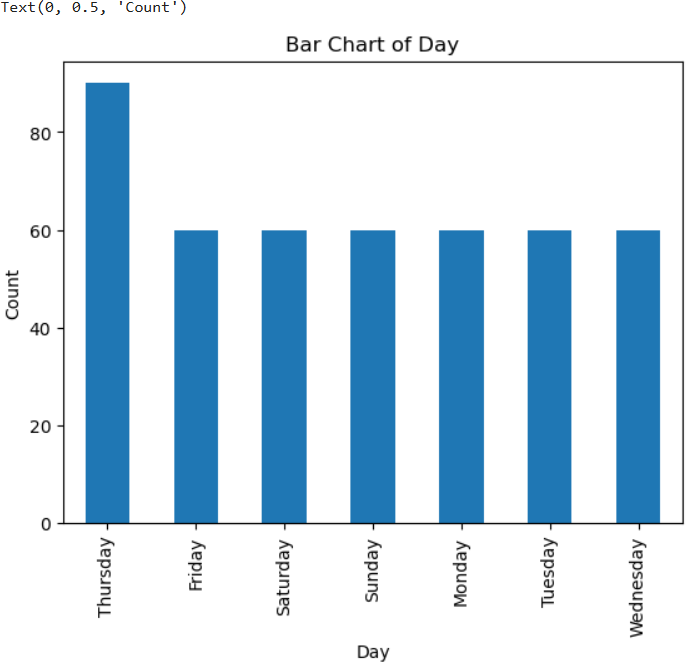
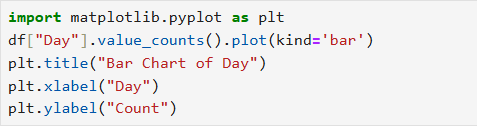
# Bargraphs:

Based on the given dataset, the following columns appear to be categorical:

* + **Day:** Contains the day of the week (Thursday in this case).
  + **SKU:** Unique identifier for each product.
  + **City:** City where the sale occurred.
  + **BU:** Business Unit.
  + **Brand:** Brand of the product.
  + **Model:** Model of the product.

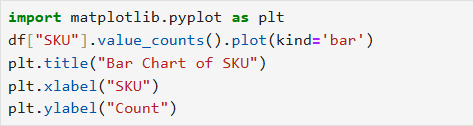
# Day:

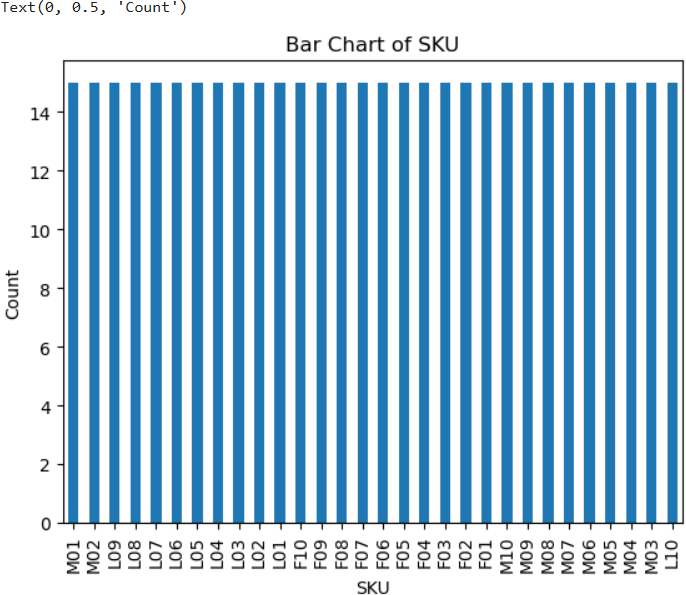
Showing the bargraph for Day.



# SKU:

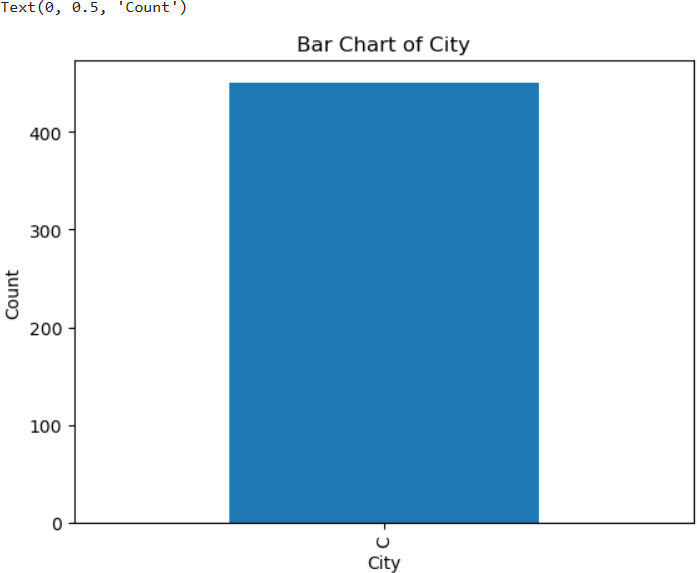
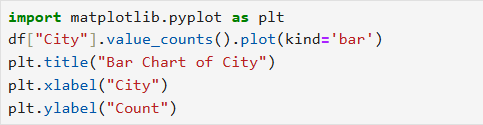
Showing the bargraph for SKU.





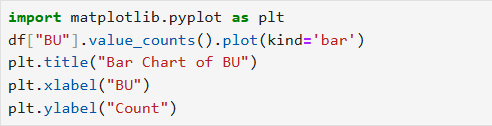
# City:

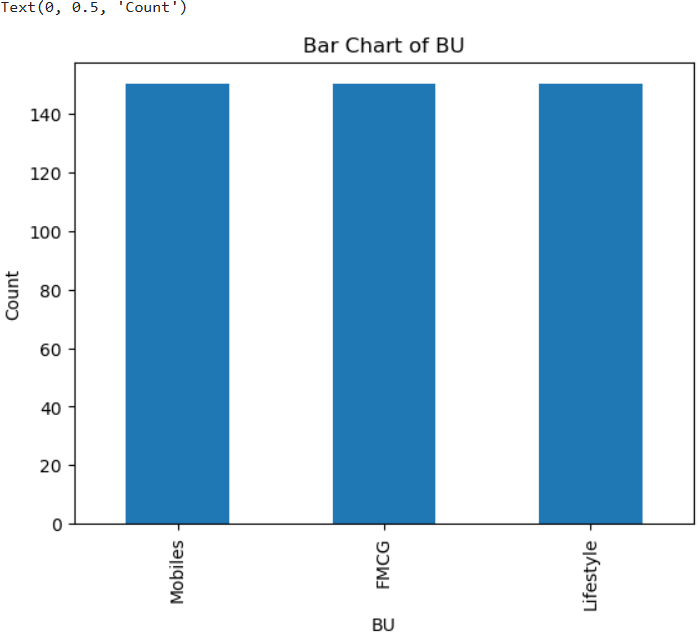
Showing the bargraph for City.



# BU:

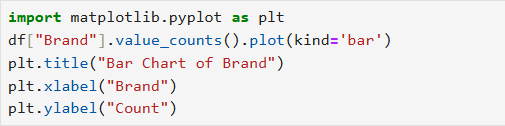
Showing the bargraph for BU.

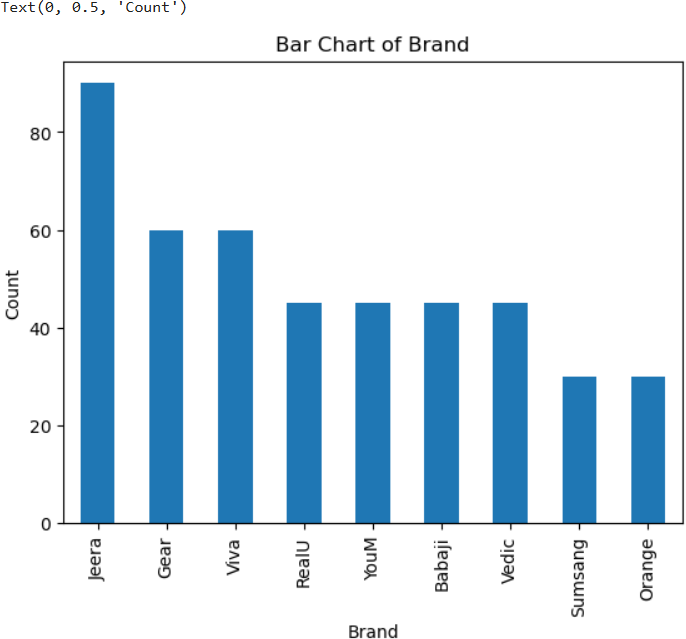




# Brand:

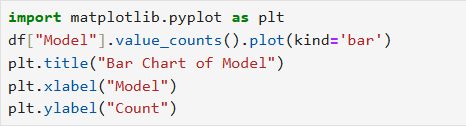
Showing the bargraph for Brand.

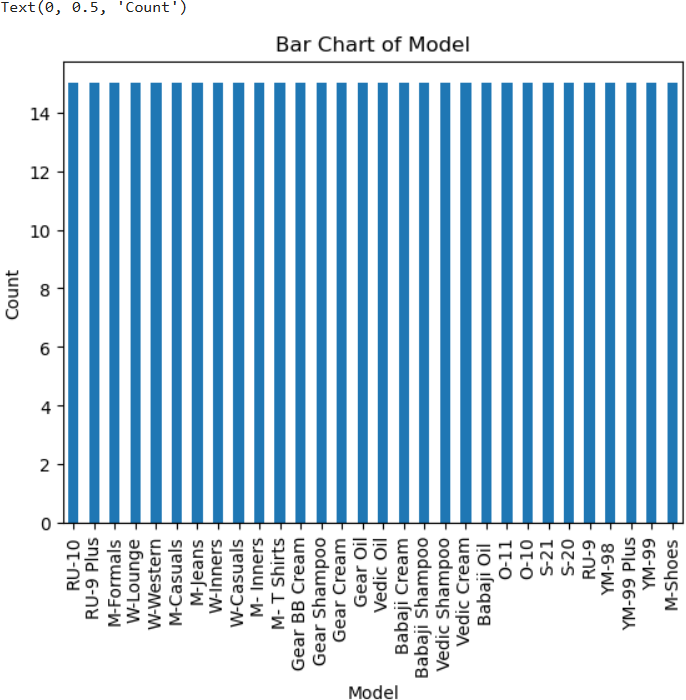




# Model:

Showing the bargraph for Model.





By examining the bar charts, we can observe the following:

* + **Day:** All sales occurred on Thursday.
  + **City:** All sales happened in the same city (C).
  + **BU:** There are three distinct Business Units (Mobiles, FMCG, Lifestyle).
  + **Brand:** There are multiple brands involved (RealU, YouM, Sumsang, Orange, Babaji, Vedic, Gear, Jeera, Viva).
  + **Model:** There are various models within each brand.

# STANDARDIZATION OF NUMERICAL VARIABLES:

Understanding Standardization (Z-score Normalization):

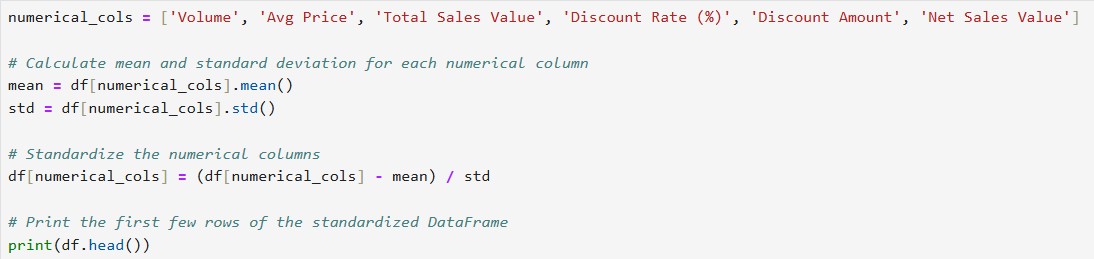
Standardization is a technique used to transform numerical variables into a common scale, often with a mean of 0 and a standard deviation of 1. This process is essential for many machine learning algorithms, as it helps to ensure that features with different scales contribute equally to the model.

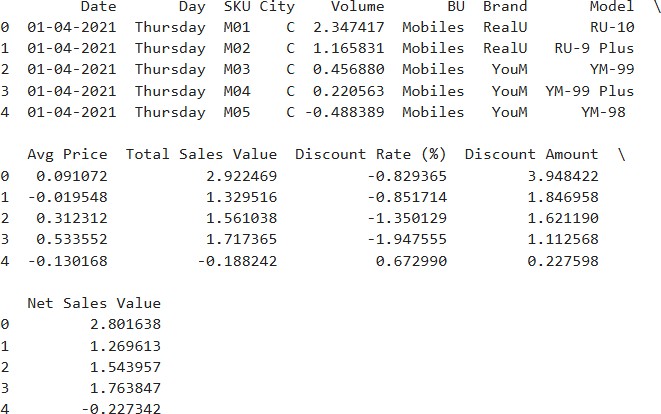
# Steps:

1. Calculate the mean (μ) and standard deviation (σ) for each numerical column.
2. Standardize each data point using the formula:

z = (x - μ) / σ

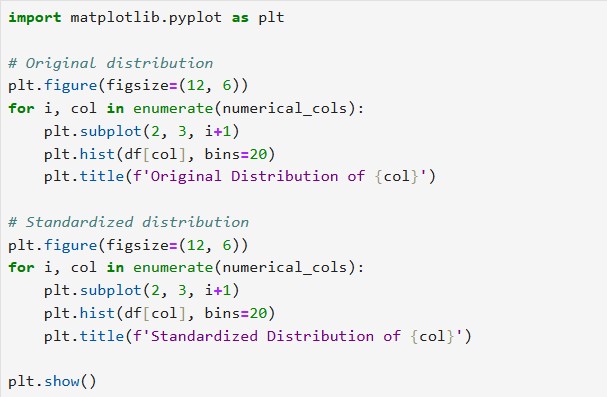
# Implementation in Python:

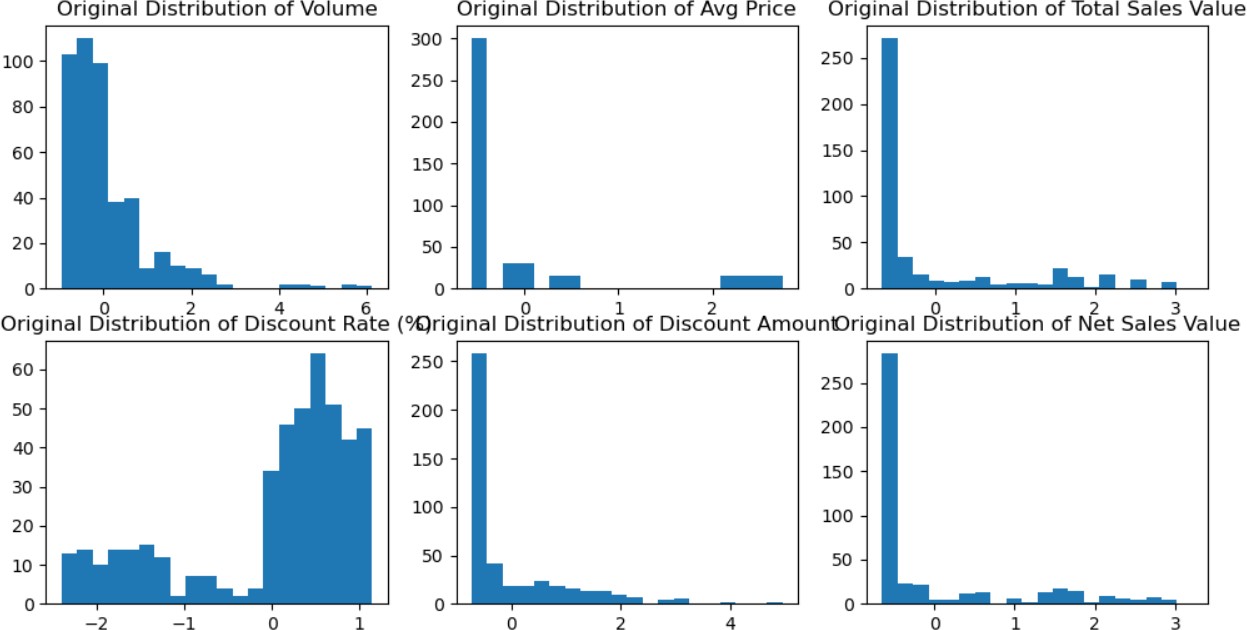


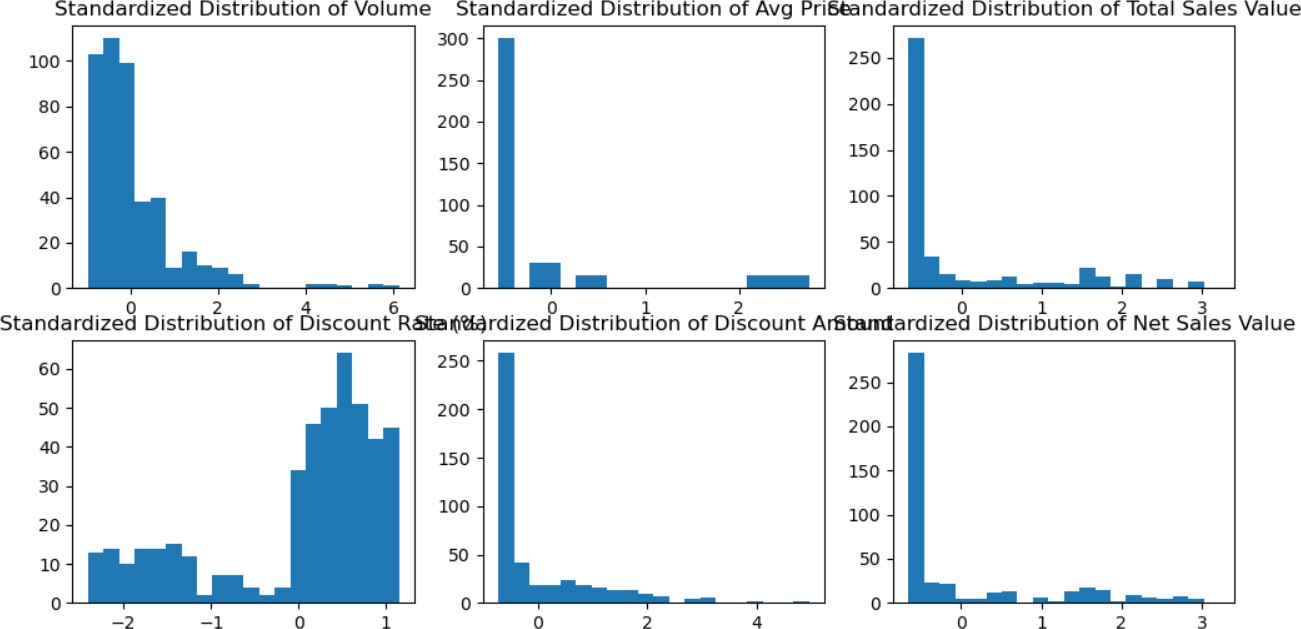


**Comparing Before and After Distributions:**

To visualize the impact of standardization, we can create histograms or box plots for the numerical columns before and after the transformation.







# Key Observations from the Standardized Data:

* + The mean of each standardized column will be close to 0.
  + The standard deviation of each standardized column will be close to 1.
  + The shape of the distribution (e.g., normal, skewed) remains the same after standardization.

By standardizing numerical variables, you can improve the performance of many machine learning algorithms, especially those that rely on distance-based calculations or gradient descent optimization.

# CONVERSION OF CATEGORICAL DATA INTO DUMMY VARIABLES:

To transform categorical variables into a format that can be understood by machine learning algorithms, we need to convert them into numerical representations. One-hot encoding is a common technique for this purpose.

# Steps:

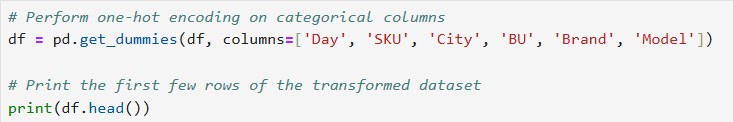
1. Identify Categorical Columns:

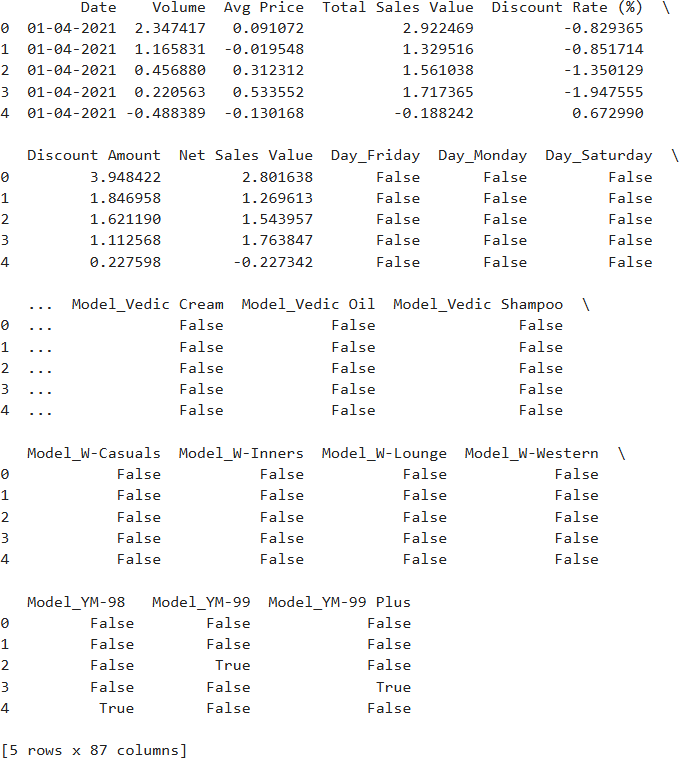
In the given dataset, the following columns are categorical:

* + Day
  + SKU
  + City
  + BU
  + Brand
  + Model

1. Apply One-Hot Encoding:

We can use the get\_dummies() function from pandas to create dummy variables for each category:





* For each categorical column, the get\_dummies() function creates a new binary column for each unique category.
* The value in the new column is True if the original category matches, and False otherwise.
* This process effectively converts categorical data into a numerical format that can be used by machine learning algorithms**.**

# CONCLUSION:

**Key Findings from Descriptive Analytics and Data Visualization: Numerical Variables:**

* The dataset exhibits variability in numerical variables like Volume, Avg Price, and Total Sales.
* Some variables, such as Discount R and Discount A, seem to have a relatively narrow range.

# Categorical Variables:

* The majority of sales occurred on a single day, Thursday.
* Sales are concentrated in a specific city (C).
* The dataset includes a variety of brands and models, particularly in the mobile category.
* The BU column suggests different business units or divisions within the company.

# Importance of Data Preprocessing:

Data preprocessing is a crucial step in data analysis and machine learning pipelines. It ensures that the data is clean, consistent, and suitable for modeling.

# Standardization:

Standardizing numerical variables helps in:

* Scaling features to a common range, preventing features with larger scales from dominating the model.
* Improving the convergence of optimization algorithms.
* Making feature comparisons more meaningful.

# One-Hot Encoding:

* Converting categorical variables into numerical representations is essential for machine learning algorithms.
* One-hot encoding creates binary features for each category, allowing the model to learn the importance of each category.

By performing these preprocessing steps, we can enhance the quality of our analysis and improve the performance of machine learning models.