Basic Statistics

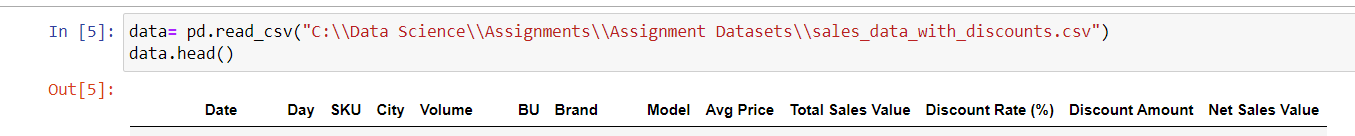
### Descriptive Analytics and Data Preprocessing on Sales & Discounts Dataset

#### 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).

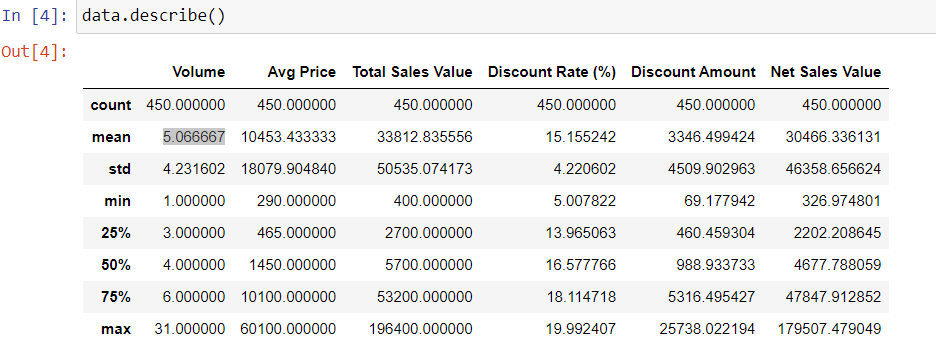


Question: Identify numerical columns in the dataset.

Answer**: ['Volume', 'Avg Price', 'Total Sales Value', 'Discount Rate (%)', 'Discount Amount', 'Net Sales Value']**

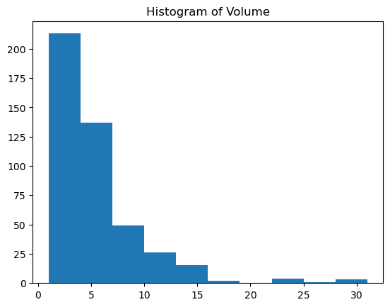
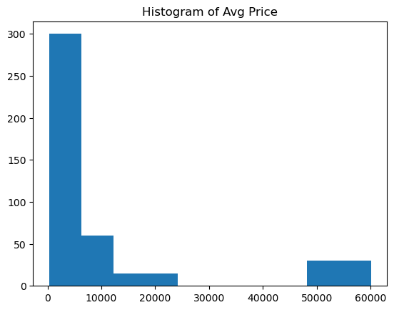
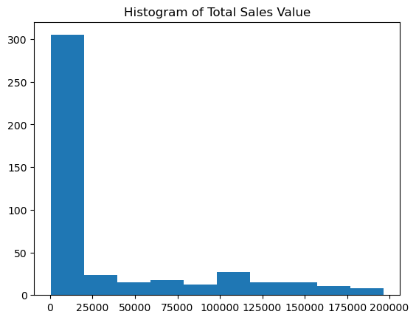
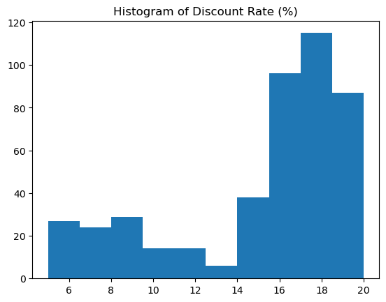
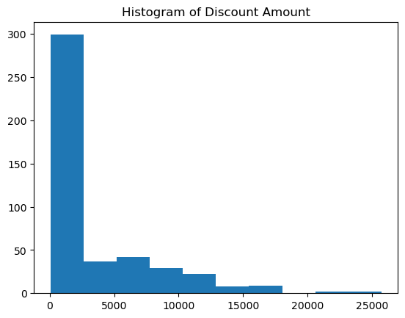
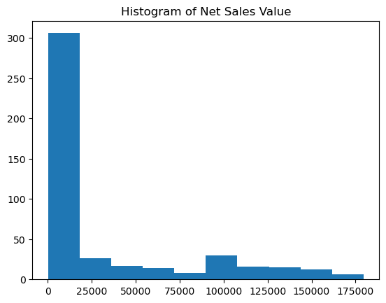
Question: Calculate the mean, median, mode, and standard deviation for these columns.

Answer:



Provide a brief interpretation of these statistics.

#### Data Visualization

* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:
  + Plot histograms for each numerical column.
  + `   
  +  

Question: Analyze the distribution (e.g., skewness, presence of outliers) and provide inferences.

1). Histogram of volume is positive right skewed with outliers from 23 to 25 and 28-30.

2). Histogram of right skewed with outliers with outliers from 5000 to 6000

3). Histogram of Tatal sales is uniformly distributed data

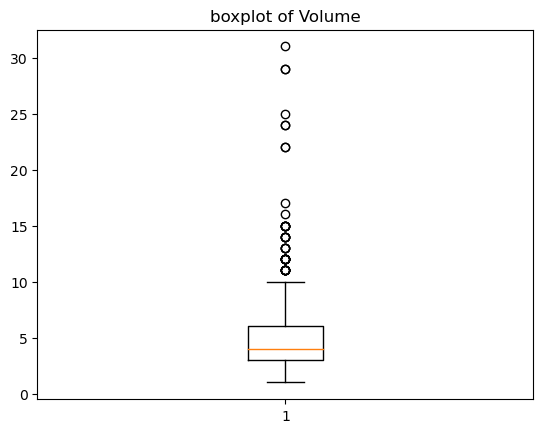
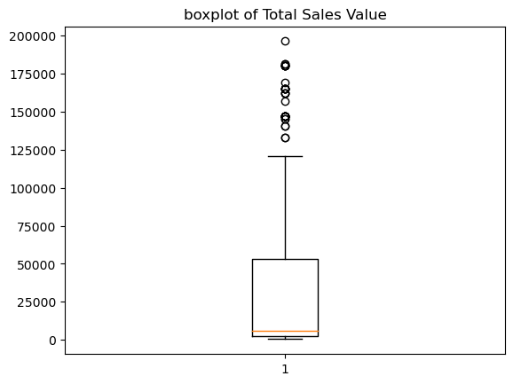
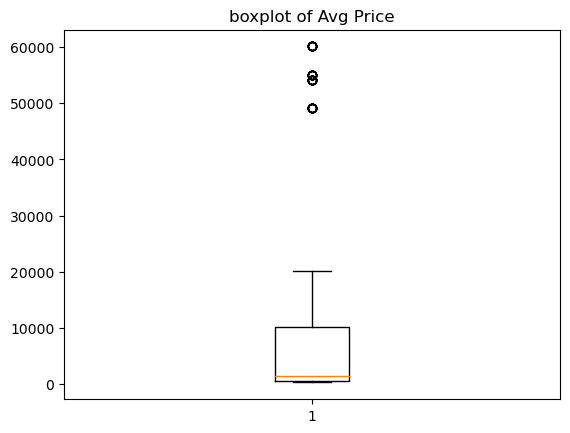
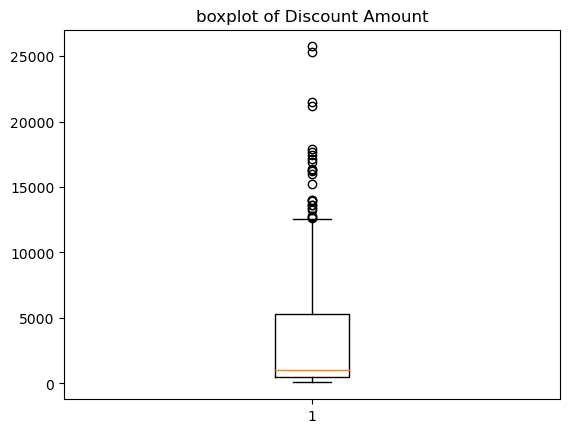
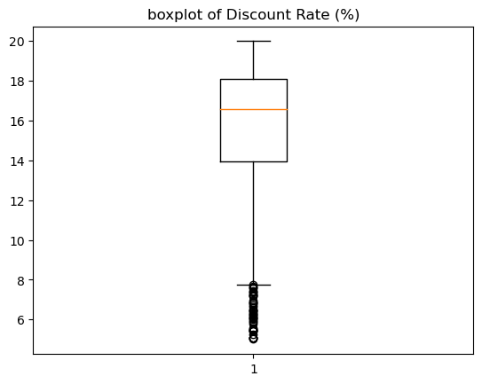
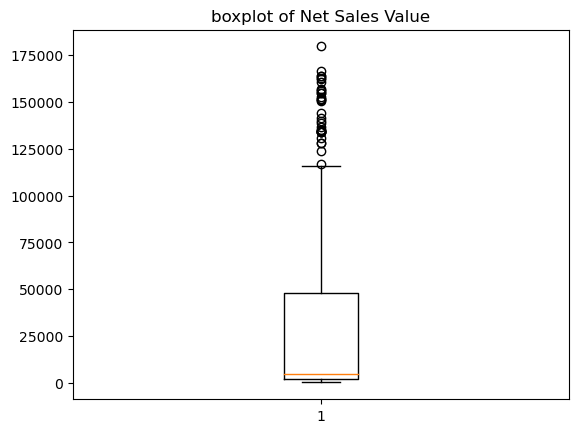
4). Histogram of Discount Rate is left skewed data

5). Histogram of disc amount is left skewed with few outliers

6). Histogram of

* **Boxplots**:

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

Question: Discuss any findings, such as extreme values or unusual distributions.

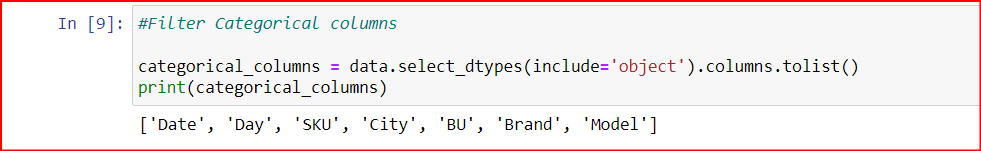
Answer:

These plots shows extreme values for Both discount total and discount range have extreme values

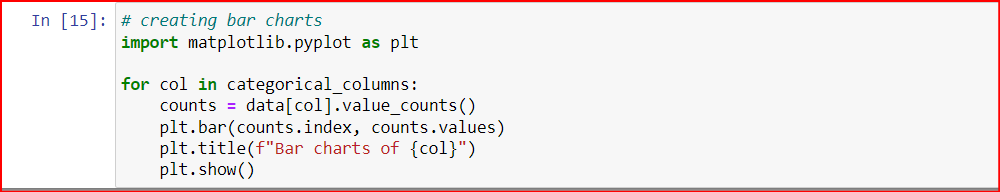
From the above plots we can say all Discount total, discount range and units sold has unusual distribution

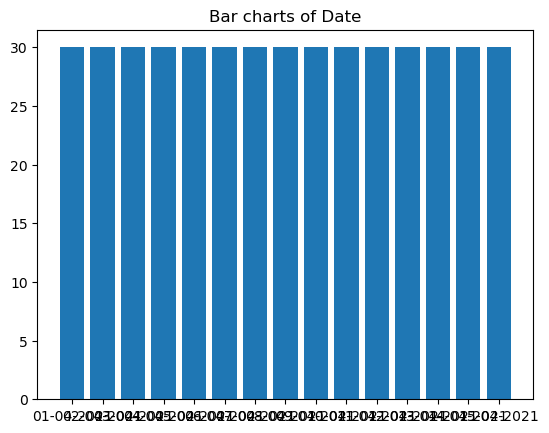
**Bar Chart Analysis for Categorical Column:**

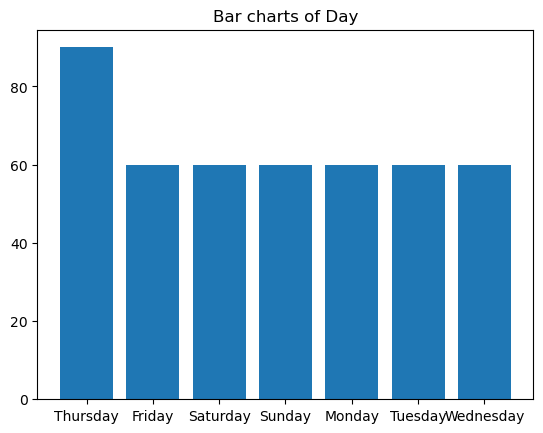
Identify categorical columns in the dataset.

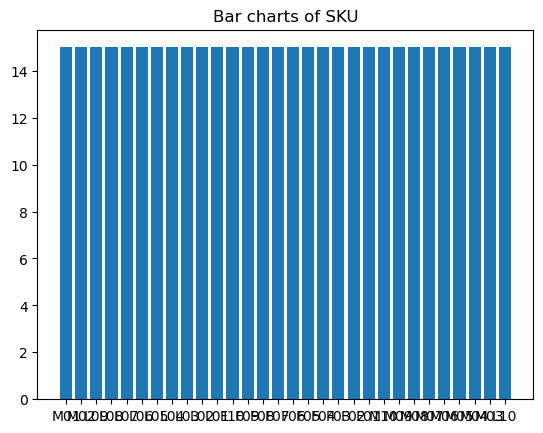


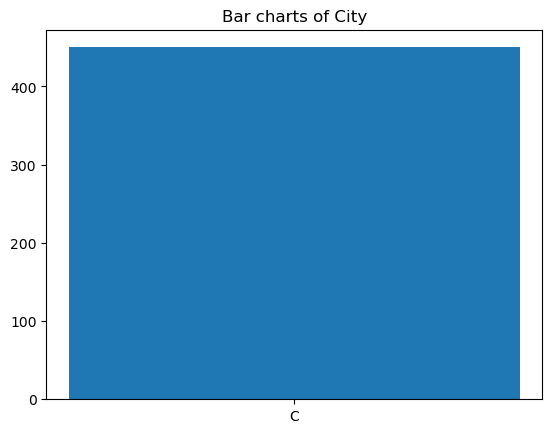
Create bar charts to visualize the frequency or count of each category.

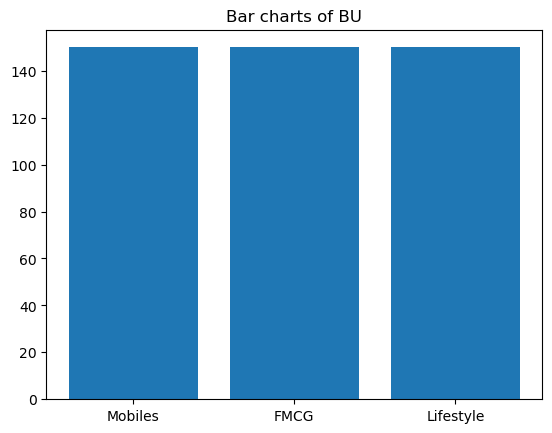


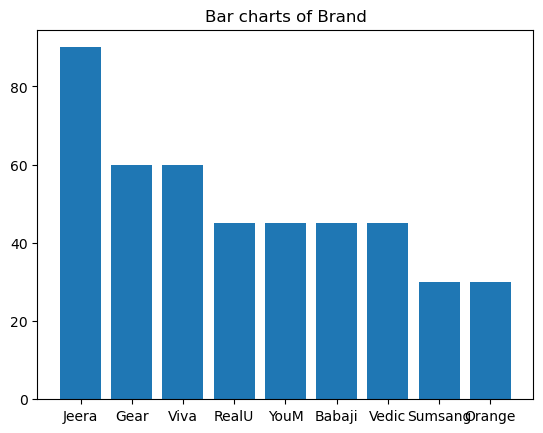


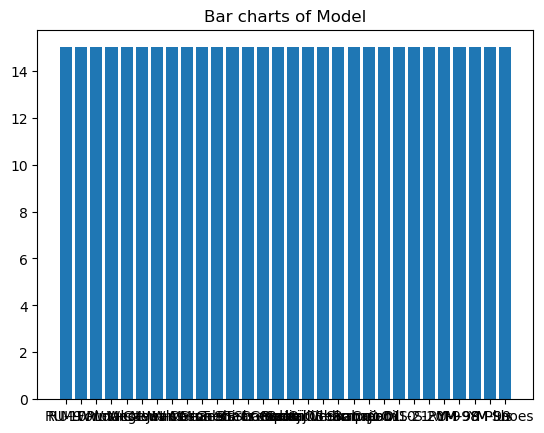












Question: Analyze the distribution of categories and provide insights.

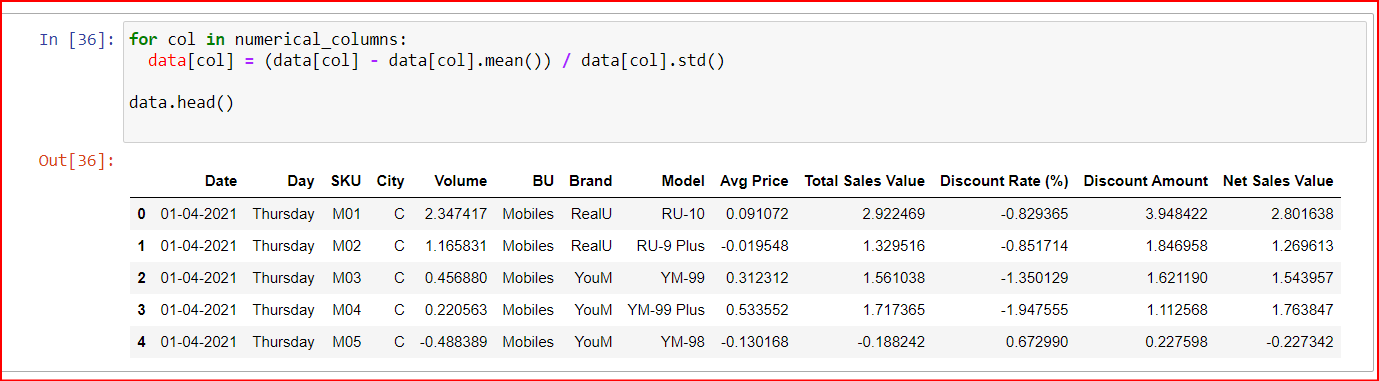
#### Standardization of Numerical Variables

* Objective: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Steps:

Question: Explain the concept of standardization (z-score normalization).

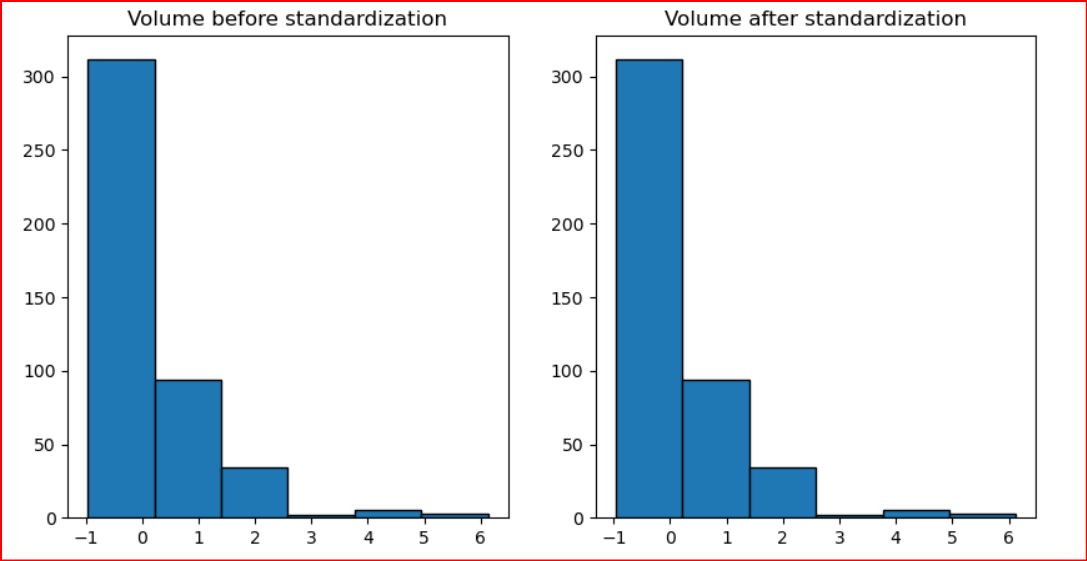
Answer: Standardization is a process to transform the data so that its mean is 0 and its standard deviation is 1 to build the models.

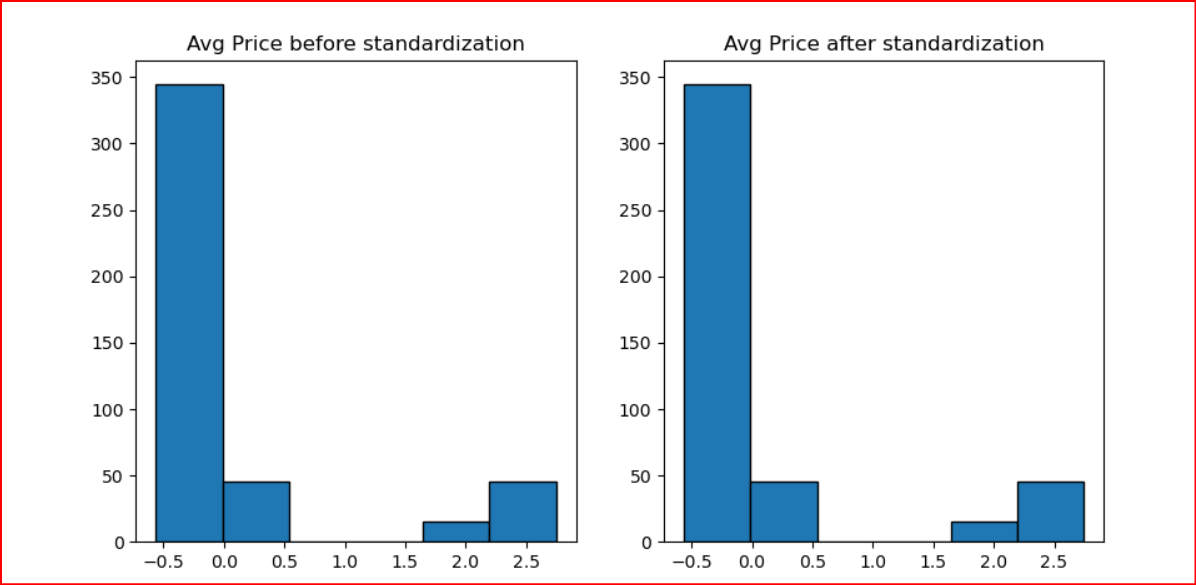
Question: Standardize the numerical columns using the formula: z=x-mu/sigma

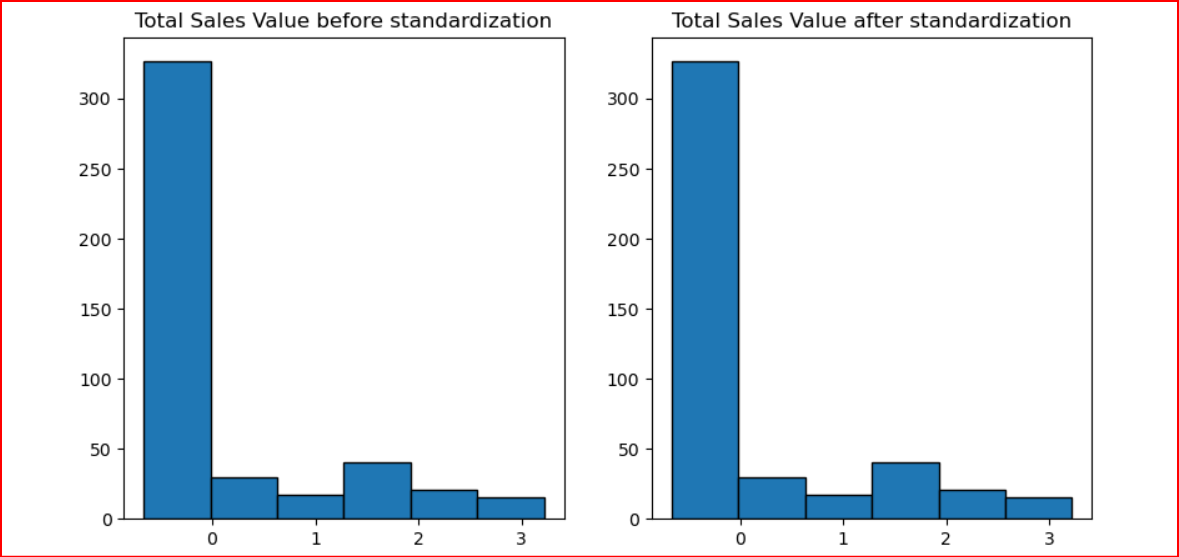


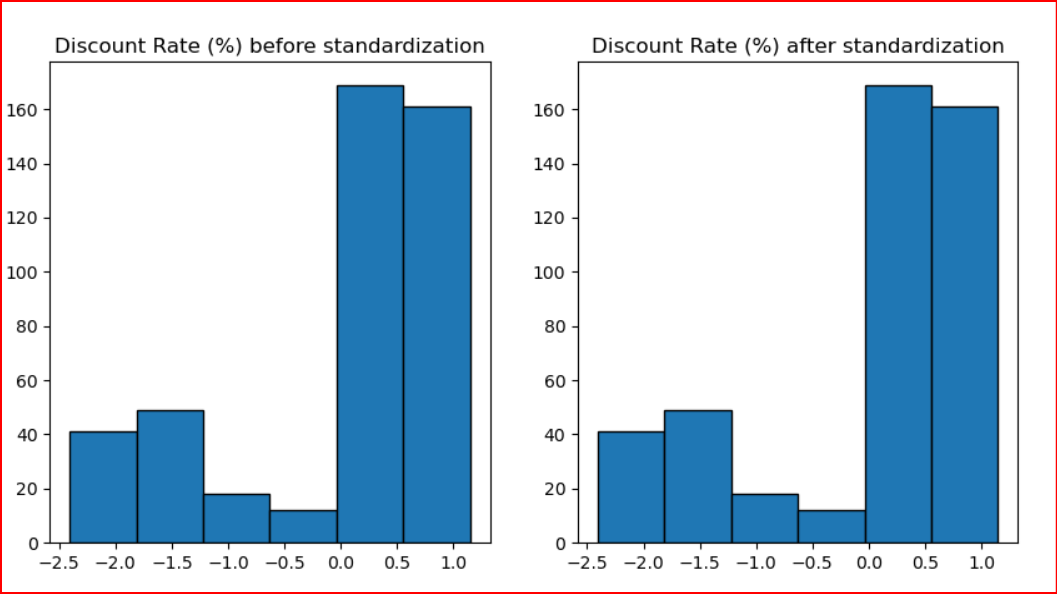
Question: Show before and after comparisons of the data distributions.

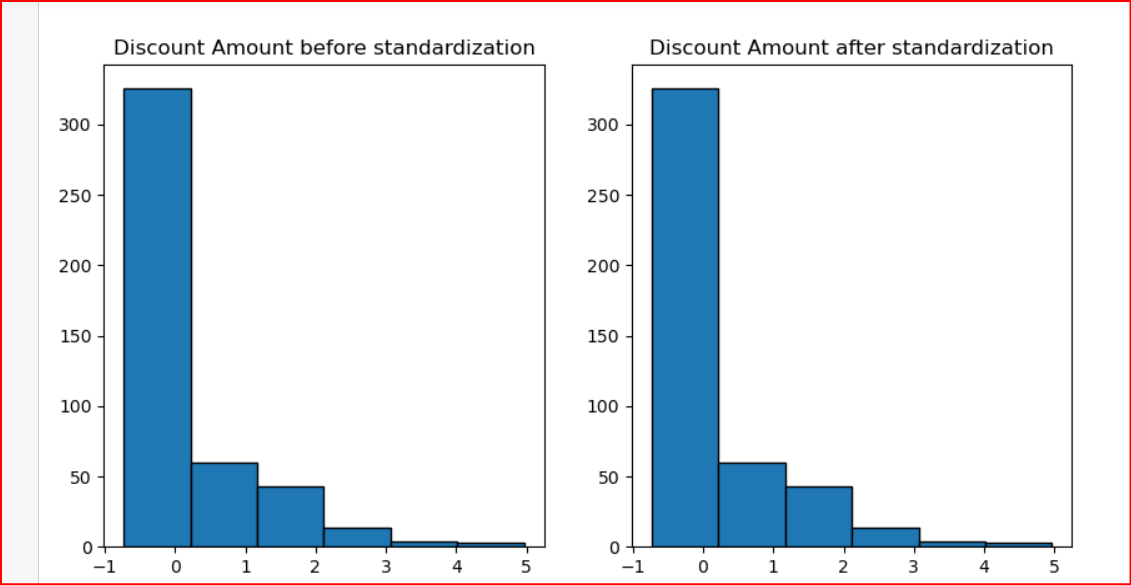
Show before and after comparisons of the data distributions.

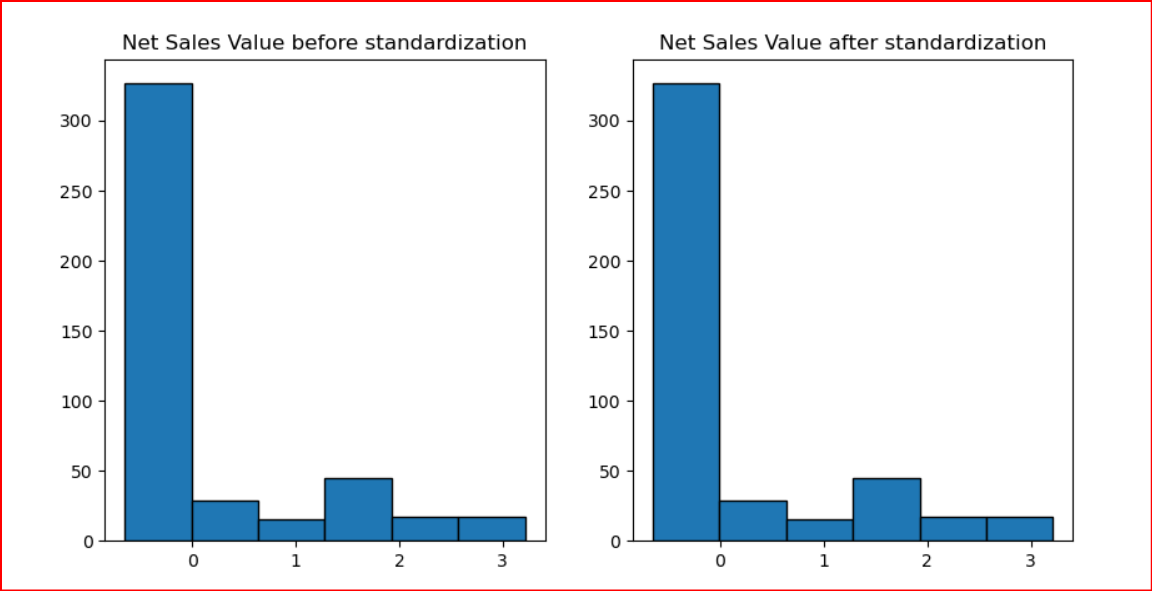












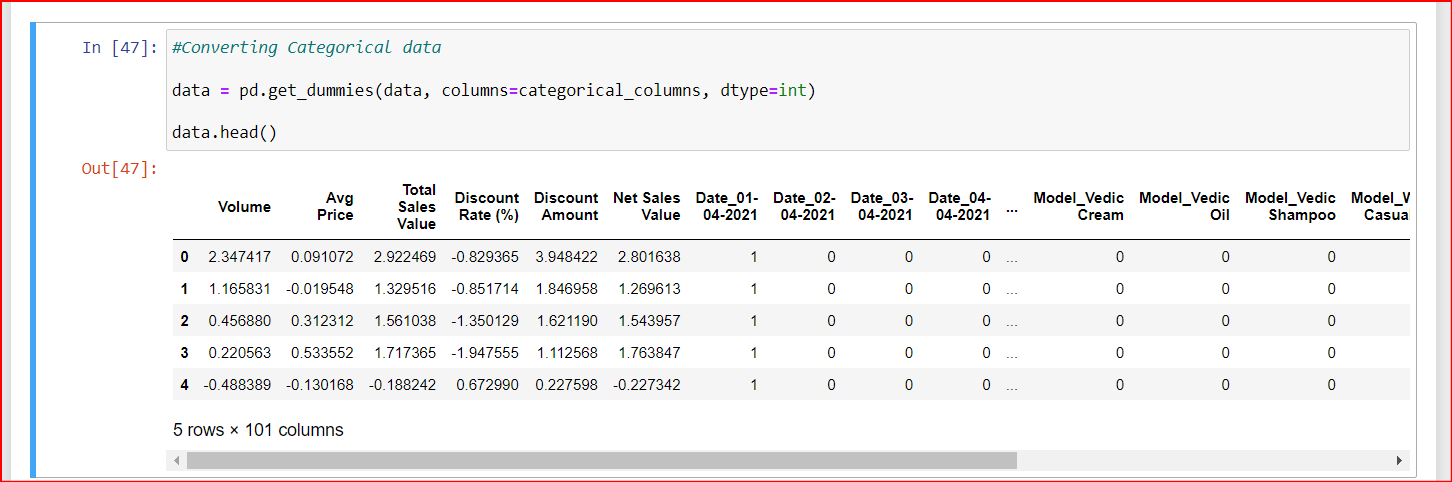
#### Conversion of Categorical Data into Dummy Variables

* Objective: To transform categorical variables into a format that can be provided to ML algorithms.
* Steps:

Question: Discuss the need for converting categorical data into dummy variables (one-hot encoding).

Answer: We can not use the categorical data in Machine Learning models so we have to convert them into binary digits like continuous data to use them in machine learning models.

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



Question: Display a portion of the transformed dataset.

#### Conclusion

* Summarize the key findings from the descriptive analytics and data visualizations.

Key Findings:

In discount total column most products have low discounts and few products having very high discounts.

In discount range column has some extreme values on both the lower and higher ends.

Sales are not distributed across all days evenly and there are certain days showing higher sales.

* Reflect on the importance of data preprocessing steps like standardization and one-hot encoding in data analysis and machine learning.

Standardization will help us improving model performance, using standardized data we can improved model predictions which makes it easier for us to build a better model

One-hot encoding converts categorical data into numerical format and transforms categorical data into a numerical format which we can use in model prediction.