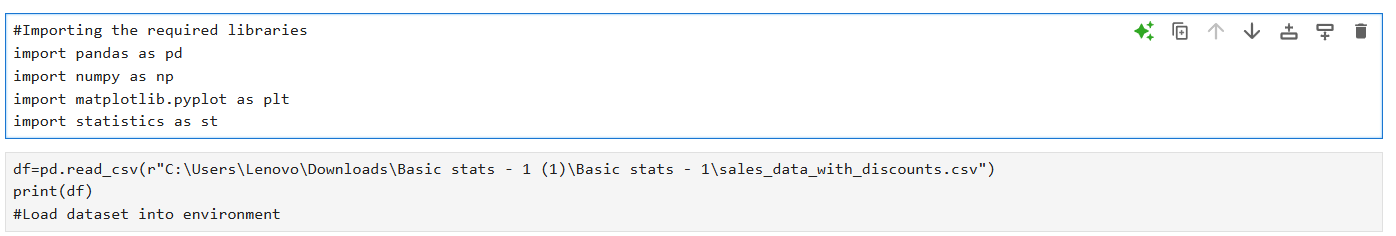
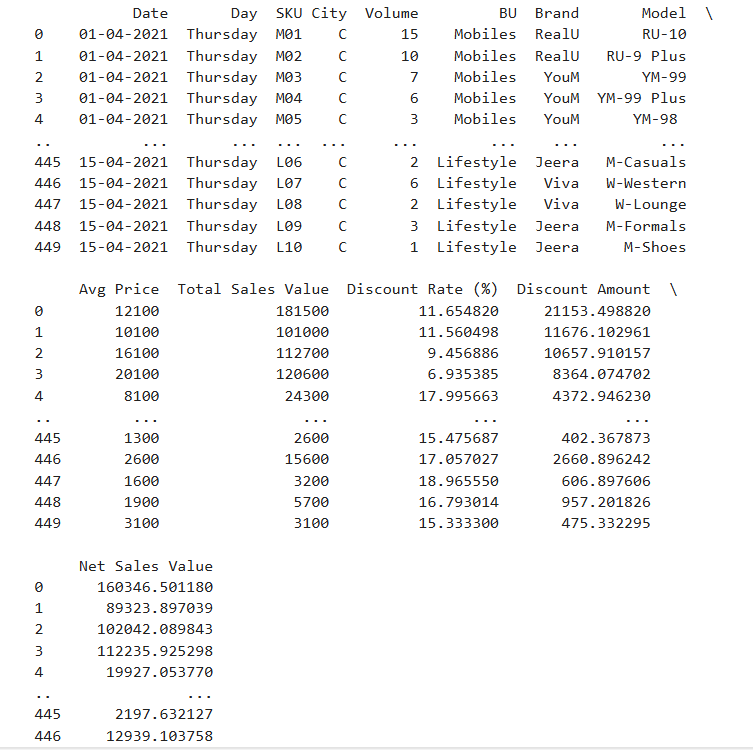
**Basic Statistics**

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

#### Introduction

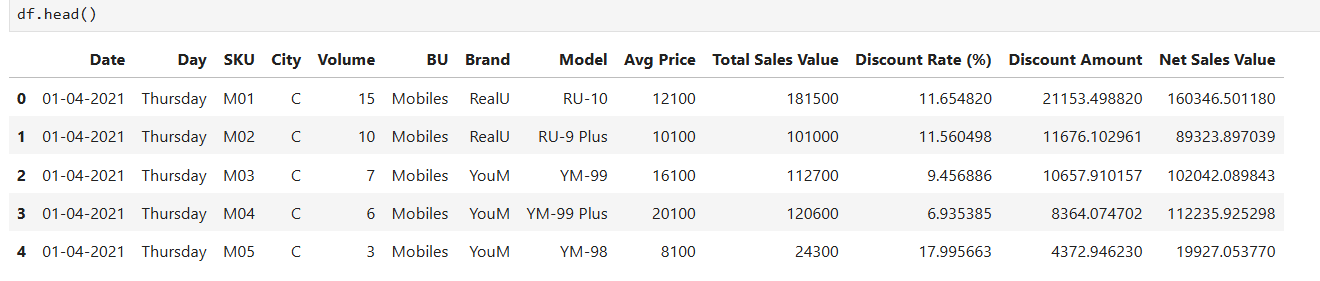
* To perform descriptive analytics, visualize data distributions, and preprocess the dataset for further analysis.
* Importing necessary libraries and loads a CSV file named sales\_data\_with\_discounts.csv into a dataframe for analysis.





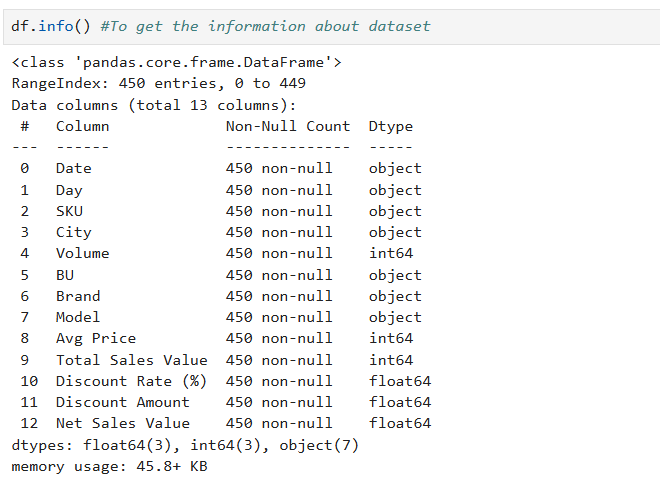
* Analyzing the dataset.





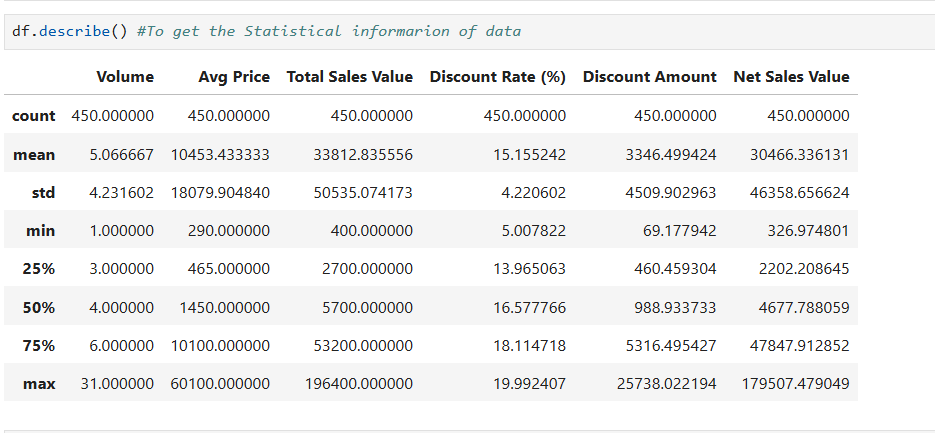
#### Descriptive Analytics for Numerical Columns

* Objective: To compute and analyze basic statistical measures for numerical columns in the dataset.
* Getting the overview of the dataset.

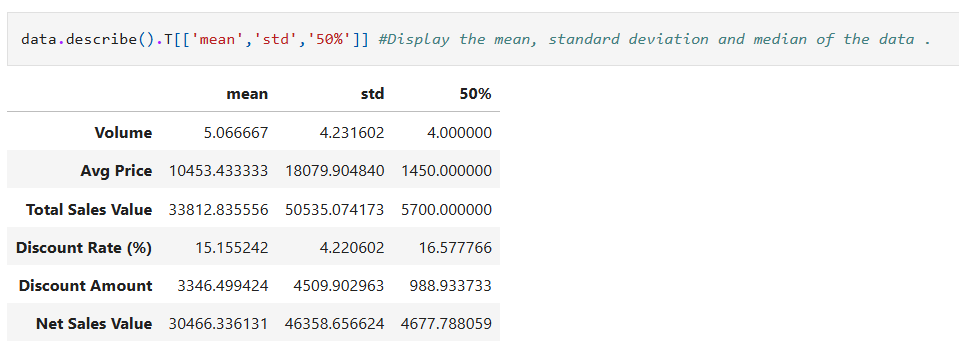


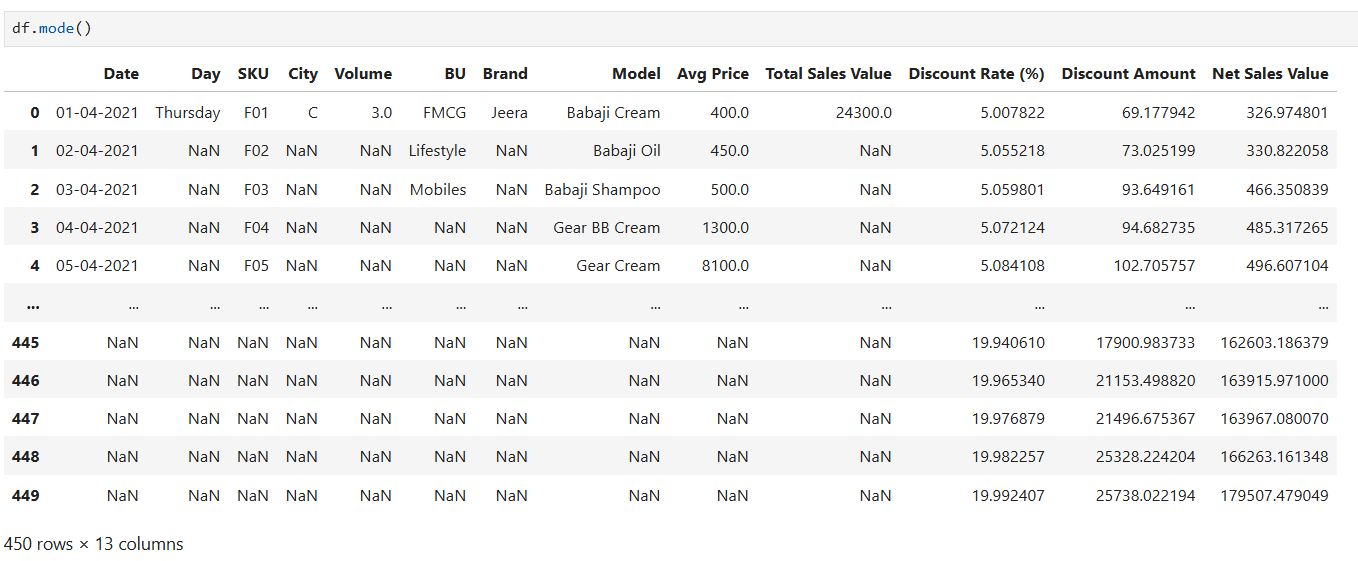
By considering the information of the dataset, I conclude that Volume, Avg Price, Total Sales Value, Discount Rate (%), Discount Amount, Net Sales Value are the numerical Columns .

* To get the Statistical information of the data.



* Calculating the mean, median, mode, and standard deviation for numerical columns.

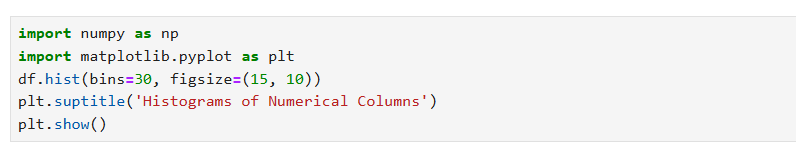


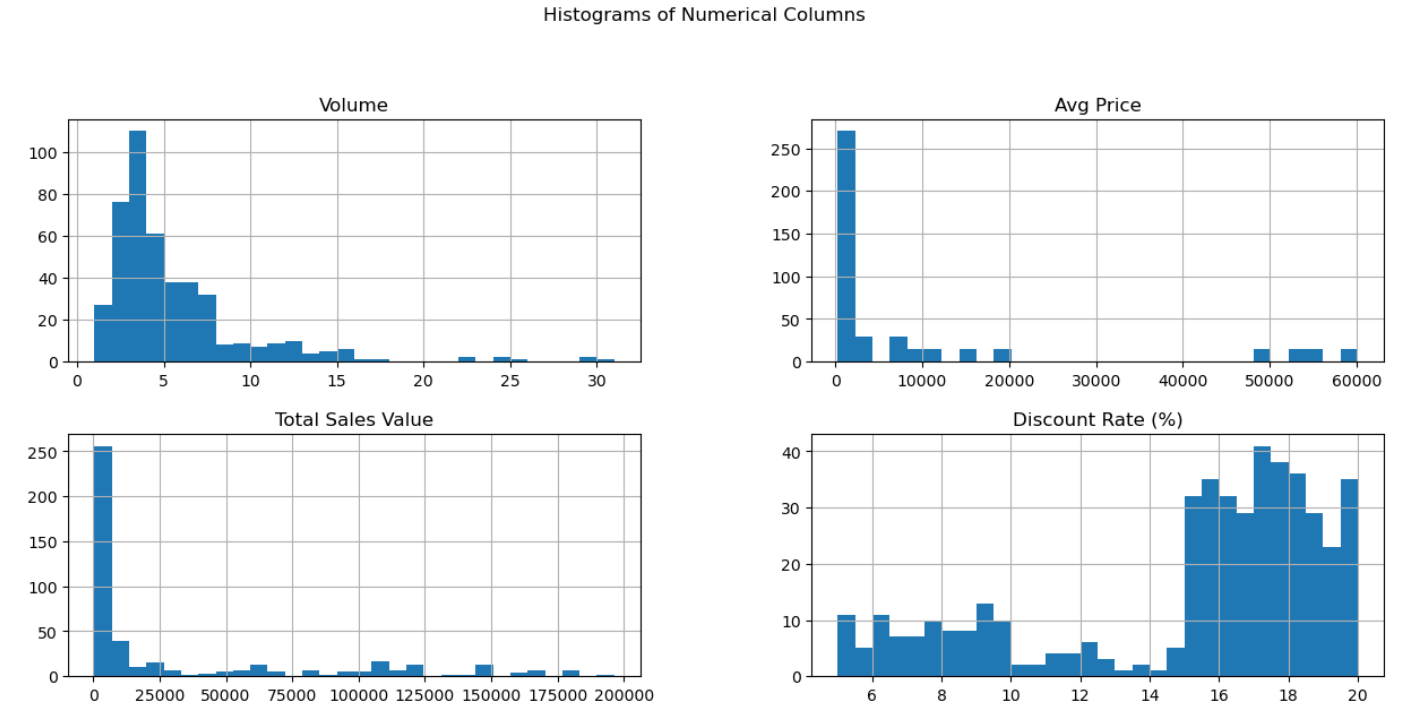


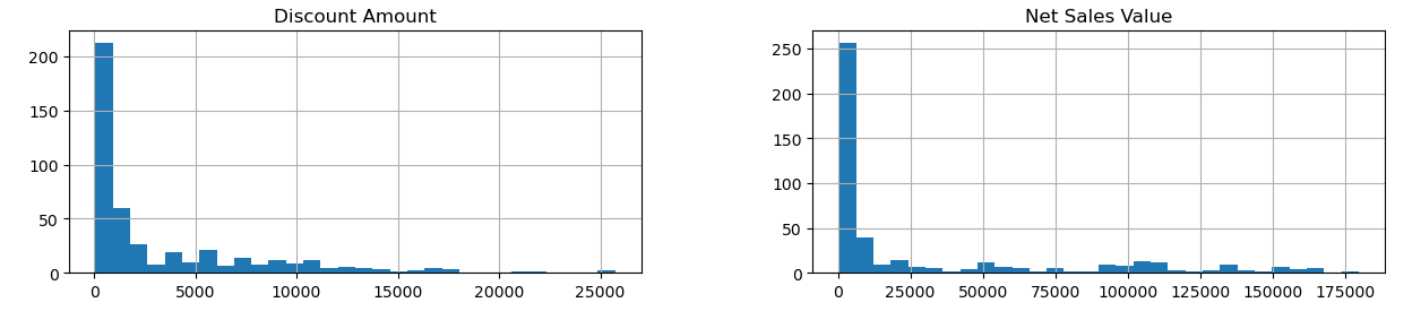
1. **Volume:** The average volume is low at 5.07 with a small standard deviation, indicating consistent low volume sales across the dataset.
2. **Avg Price:** There is a significant difference between the mean and median prices, suggesting a skewed distribution with some very high values inflating the average.
3. **Total Sales Value:** Similar to the average price, the high standard deviation and difference between the mean and median indicate a few very high sales values.
4. **Discount Rate (%):** The average discount rate is around 15%, with a relatively low standard deviation, suggesting that discount rates are fairly consistent across sales.
5. **Discount Amount:** The mean discount amount is significantly higher than the median, indicating that while most discounts are moderate, some large discounts skew the average higher.
6. **Net Sales Value:** Specific statistics are not displayed, but this typically reflects sales after discounts.

#### 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.

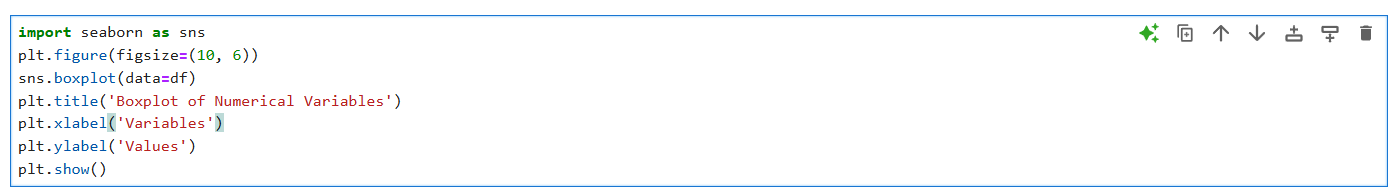


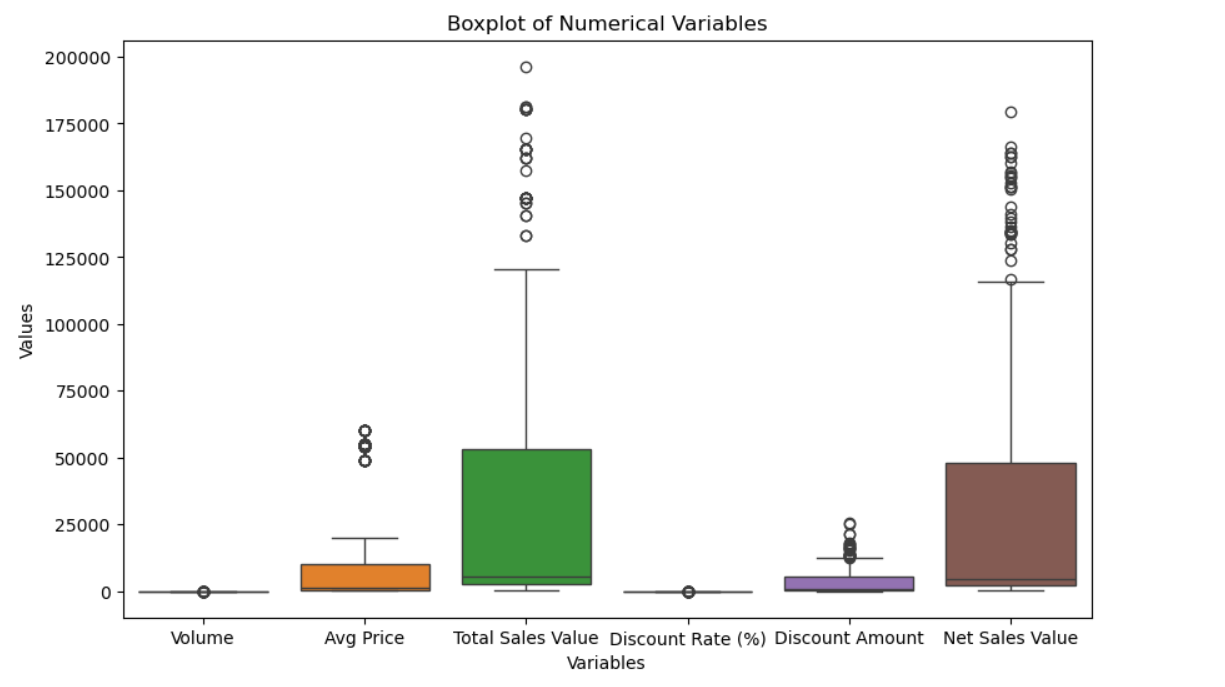




1. Volume: The column's data is right-skewed, following a log-normal distribution with outliers only above the upper bound. The histogram starts with low values, peaks, and then gradually decreases.
2. Average Price: The column follows a Poisson distribution with outliers only above the upper bound. The first bin has most values, while remaining bins have very few or none, indicating an unusual distribution.
3. Total Sales Value: Positively skewed, similar distribution to average price.
4. Discount Rate: The column's data is left-skewed with a log-normal distribution and has outliers only below the lower bound. Low values are found in the initial bins, while higher values are in the remaining bins.
5. Discount Amount: The column's data is left-skewed with a log-normal distribution and has outliers only below the lower bound. Low values are found in the initial bins, while higher values are in the remaining bins.
6. Net Sales Value: The column follows a Poisson distribution and there is no unusual distribution in the coulumn.

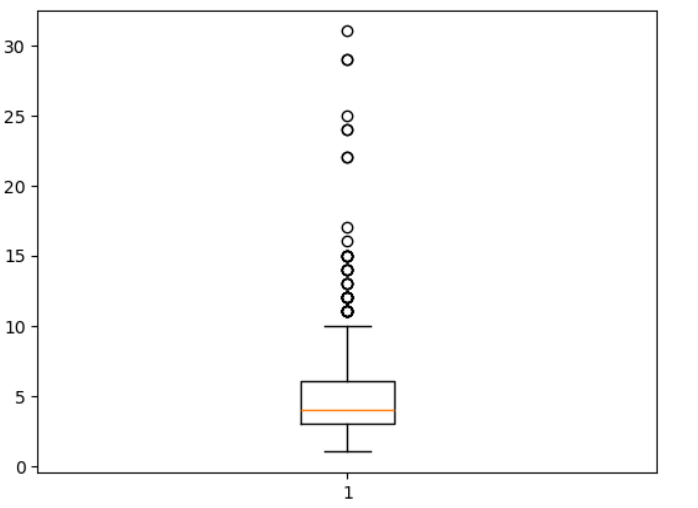
* **Boxplots**: Create boxplots for numerical variables to identify outliers and the interquartile range. Discuss any findings, such as extreme values or unusual distributions.
* Creating boxplot for various numerical variables.





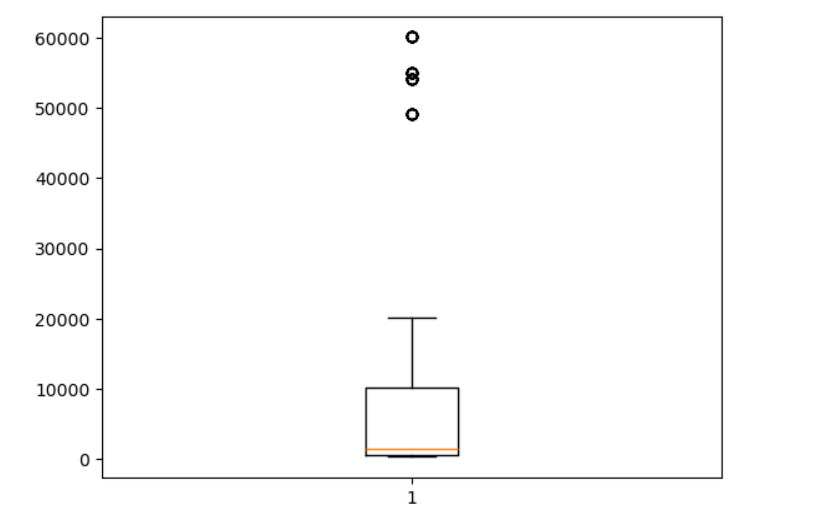
* To identify outliers and the interquartile range.



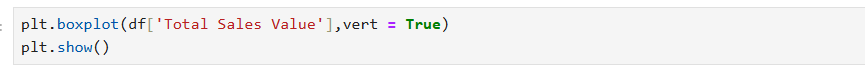


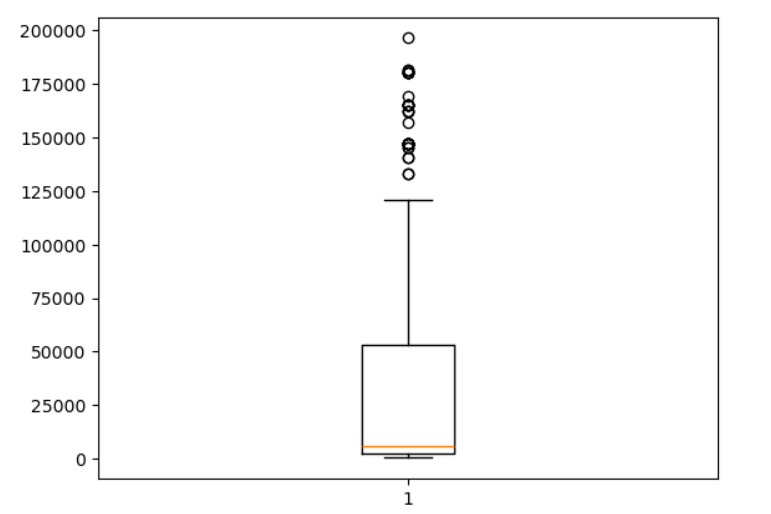
There are 12 outliers in the column, all above the upper bound of 10. More values are distributed above the median, with the lower bound at 1.



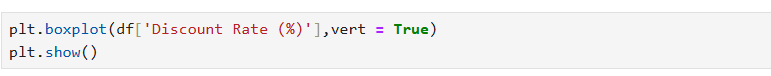


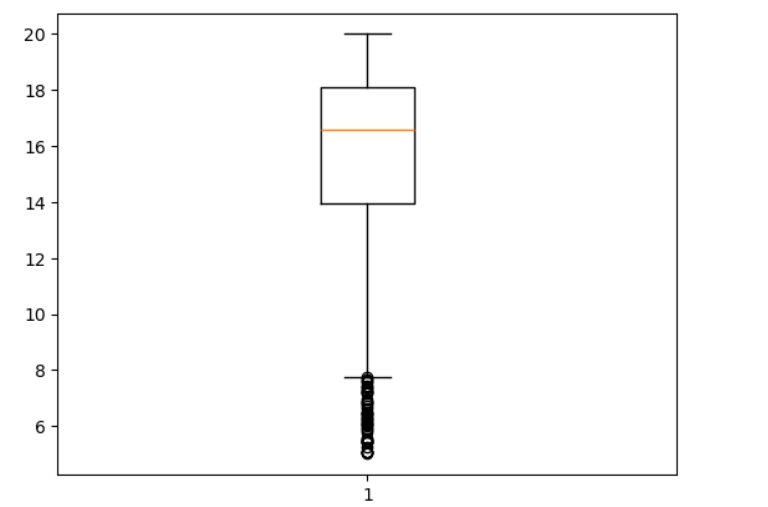
There are 4 outliers in this column, all above the upper bound of 20.1k. The majority of values are distributed above the median, with the lower bound at 290.



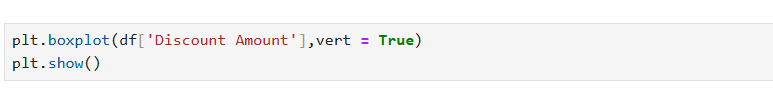


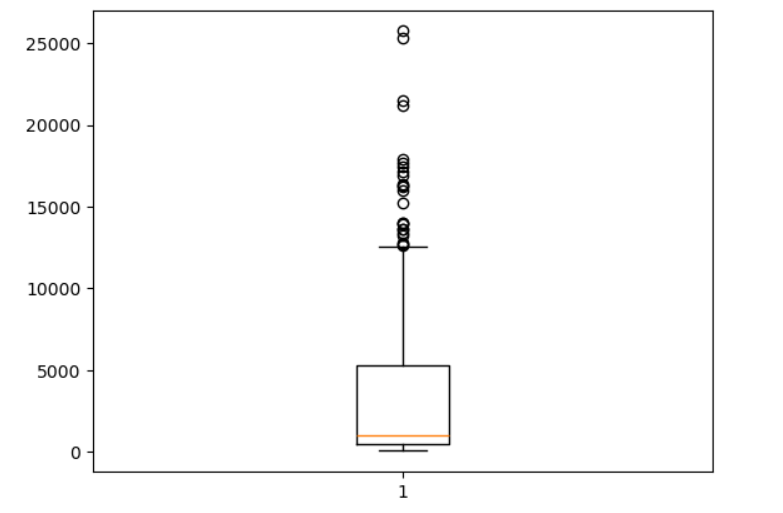
This column has 11 outliers, all above the upper bound of 120.6k. Most values are distributed above the median, with the lower bound at 400.





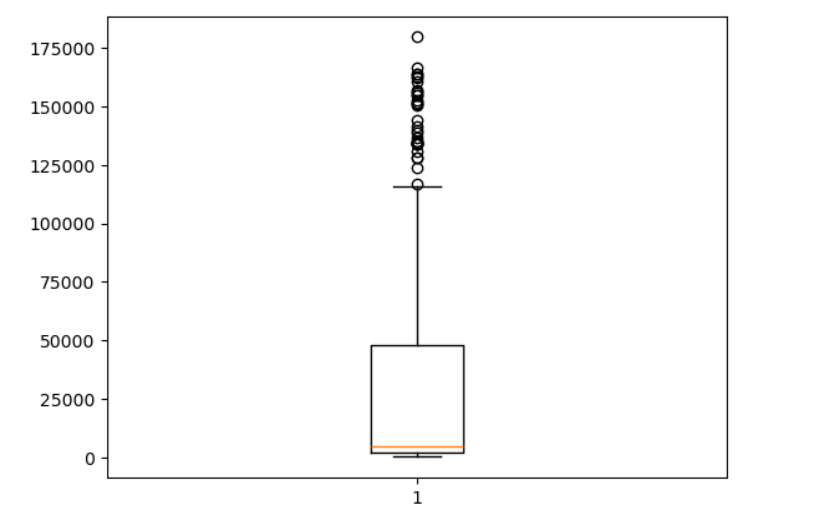
There are several outliers in this column, all at the lower bound of 5.00. The majority of values are distributed above the median, with the upper bound at 19.99.





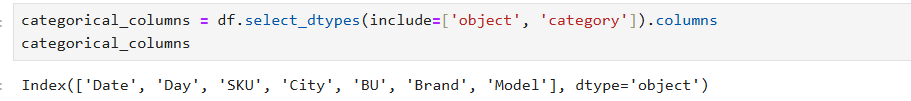
In this column, the majority of values are distributed above the median, with an upper bound of 25.73k and a lower bound of 69.17. There are outliers present at the upper bound.



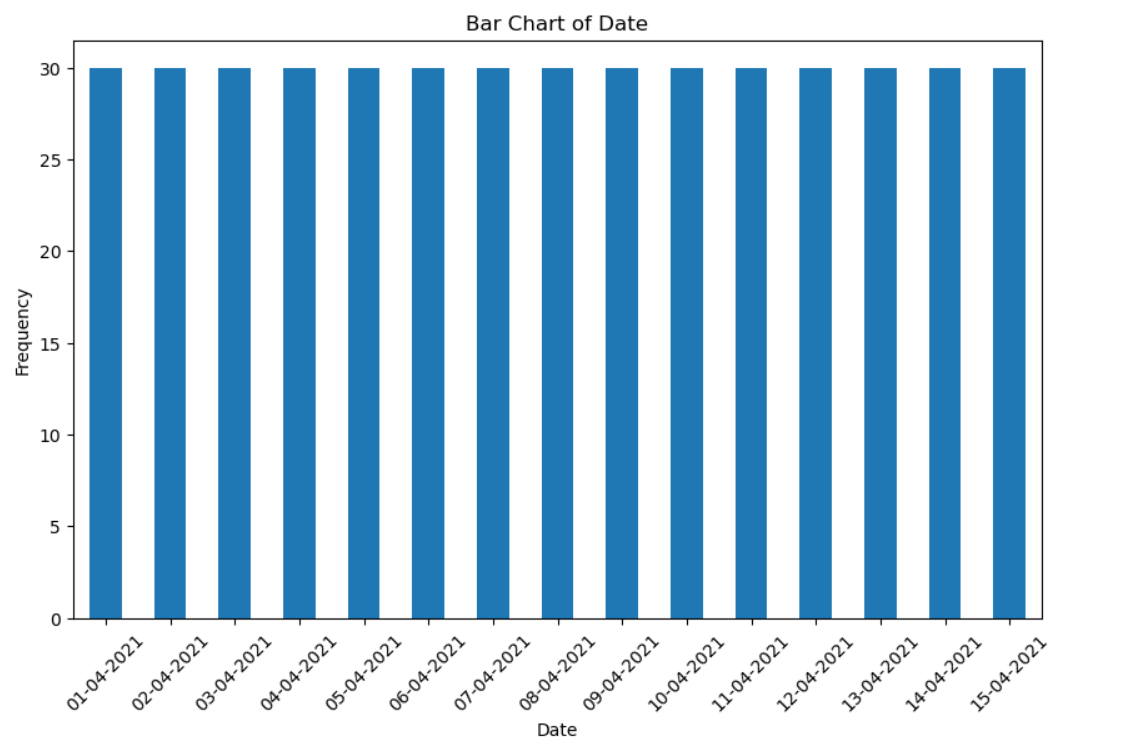


The column has values mostly distributed above the median, with an upper bound of 179.50k and a lower bound of 326.97. Outliers are present at the upper bound.

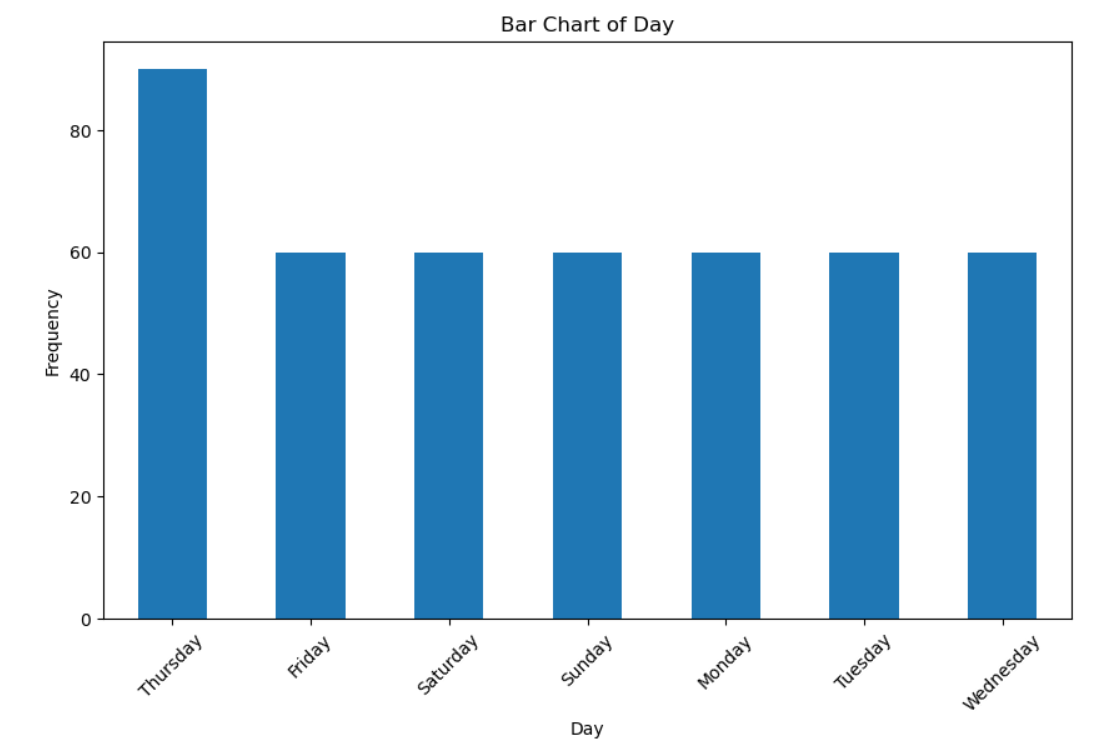
* **Bar Chart Analysis for Categorical Column:** Identify categorical columns in the dataset.Create bar charts to visualize the frequency or count of each category.Analyze the distribution of categories and provide insights.
* To identify categorical columns in the dataset.

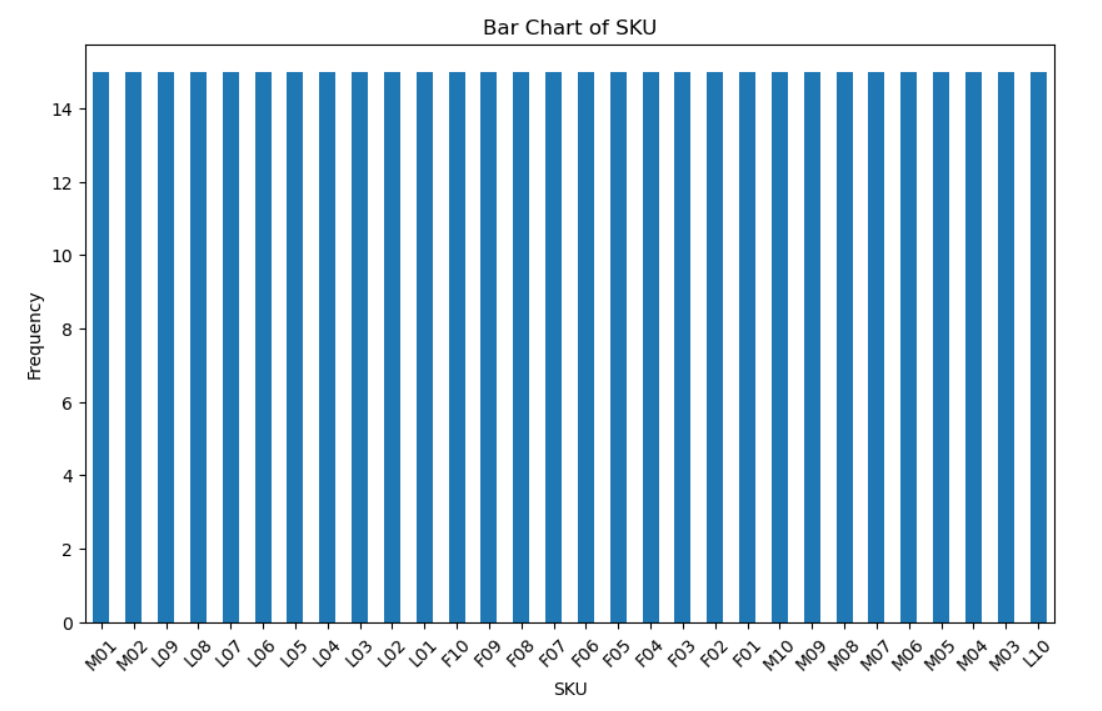


* Creating bar charts to visualize the frequency or count of each category.

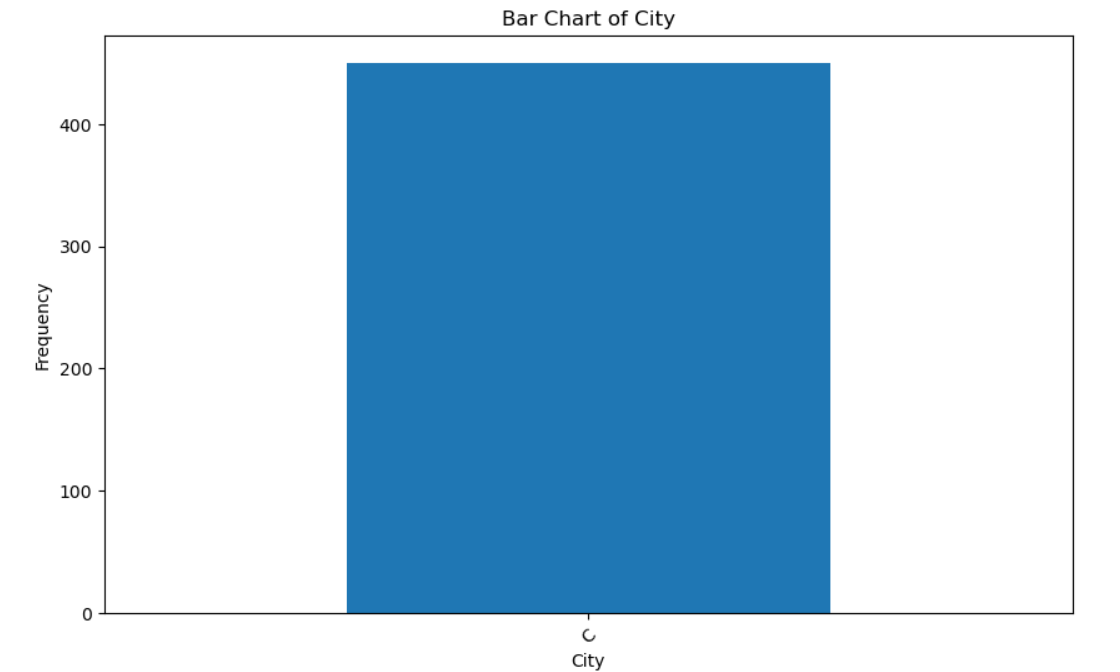


All categories in this column are equally distributed.The count for each unique value is 30.

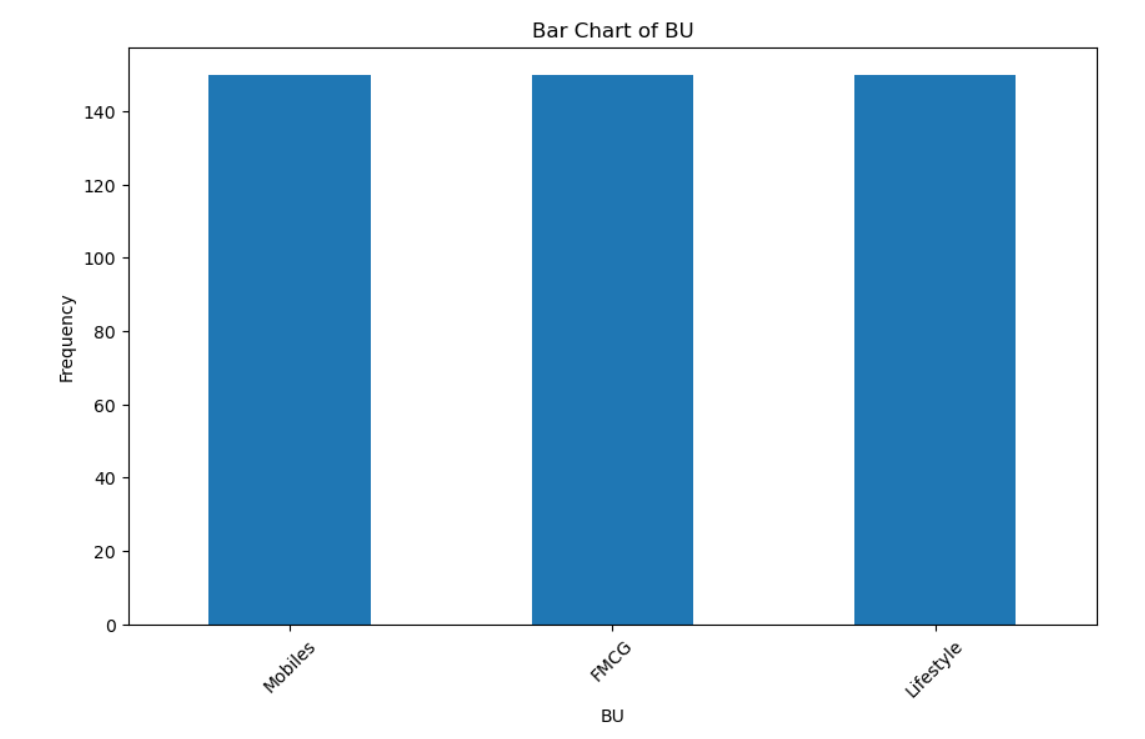


There are a total of 7 unique values in this column, representing the days of the week.All values are distributed equally, except for Thursday, which has a different count.The highest number of sales occur Thursday. 

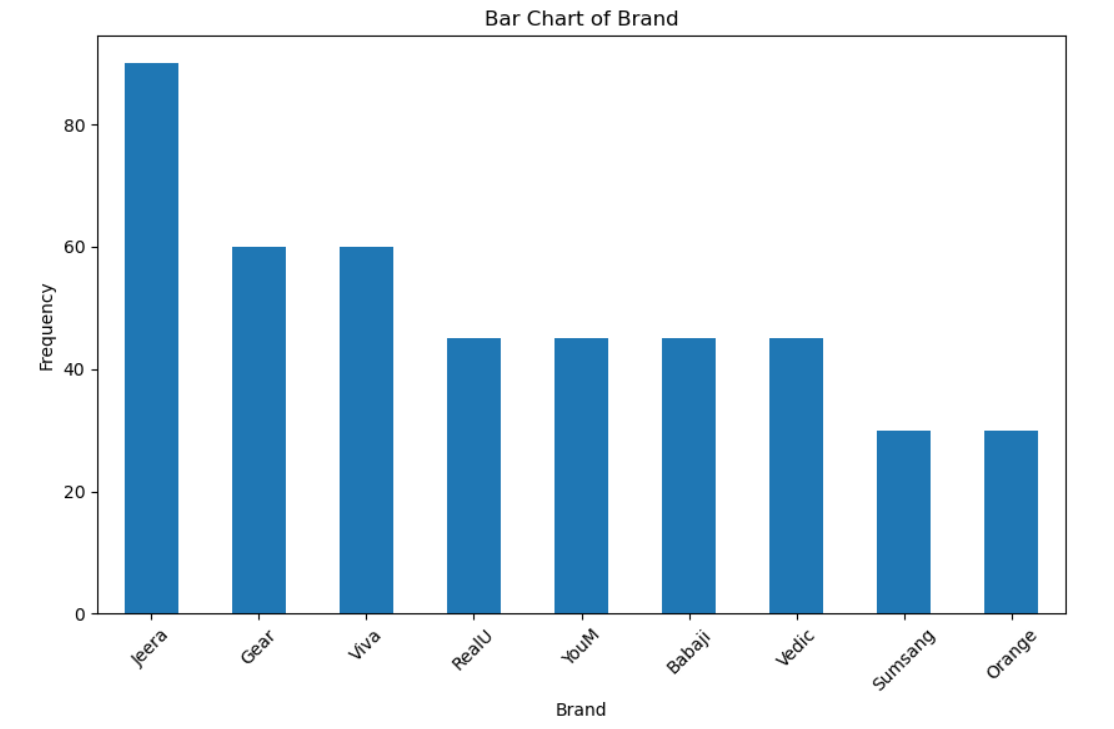
All categories in this column are equally distributed.The count for each unique value is 15.



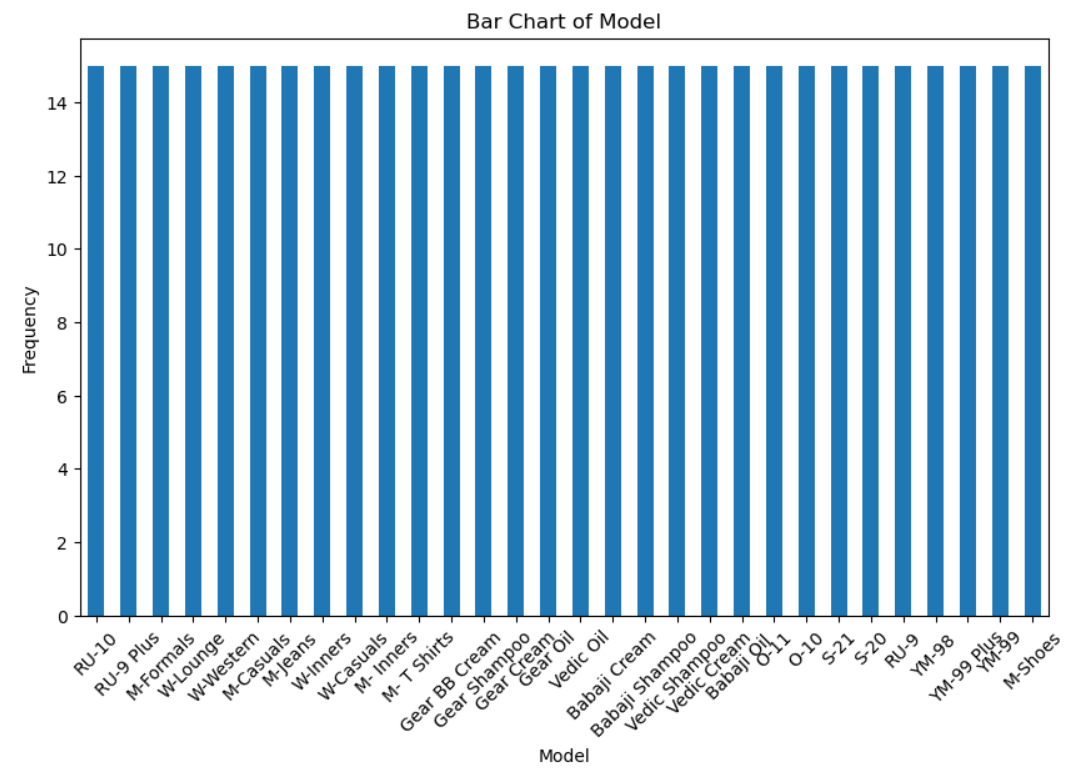
There is only one unique value in this column.The count of this unique value is 450.



All categories in this column are equally distributed. The count for each unique value is 150. There are a total of 3 categories.



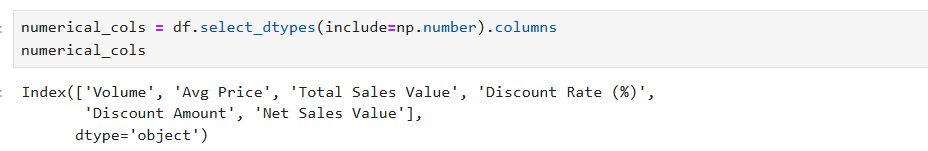
The categories in this column are unequally distributed. The brand Jeera has the highest number of sales.



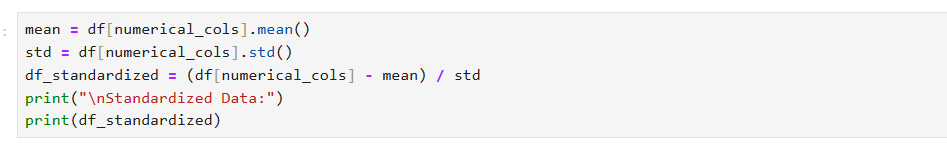
All categories in this column are equally distributed. The count for each unique value is 15.

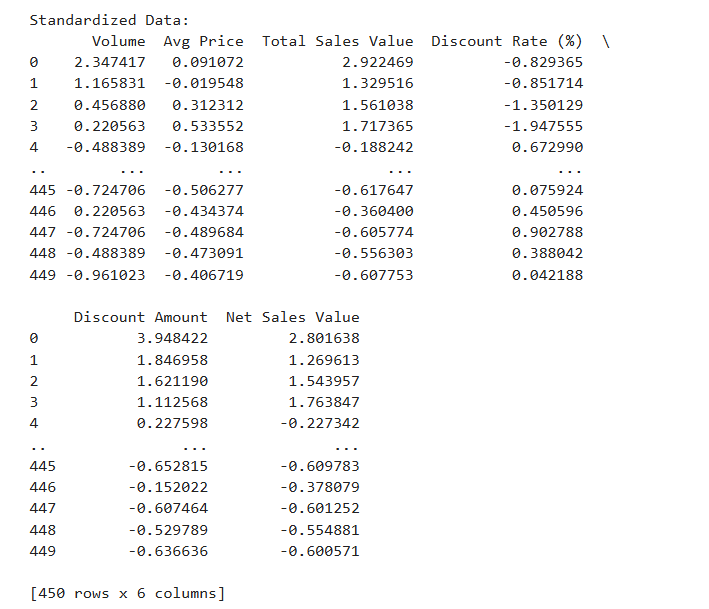
#### Standardization of Numerical Variables

* **Objective**: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Standardization, also known as z-score normalization, is a technique used to transform data so that it has a mean of 0 and a standard deviation of 1. This process adjusts the values of the data based on how far they deviate from the mean, measured in units of standard deviation. The formula for standardization is: z= (x – μ) / σ where: ( x ) is the original value, ( mu ) is the mean of the data, ( sigma ) is the standard deviation of the data, ( z ) is the standardized value
* Selecting the numerical columns.



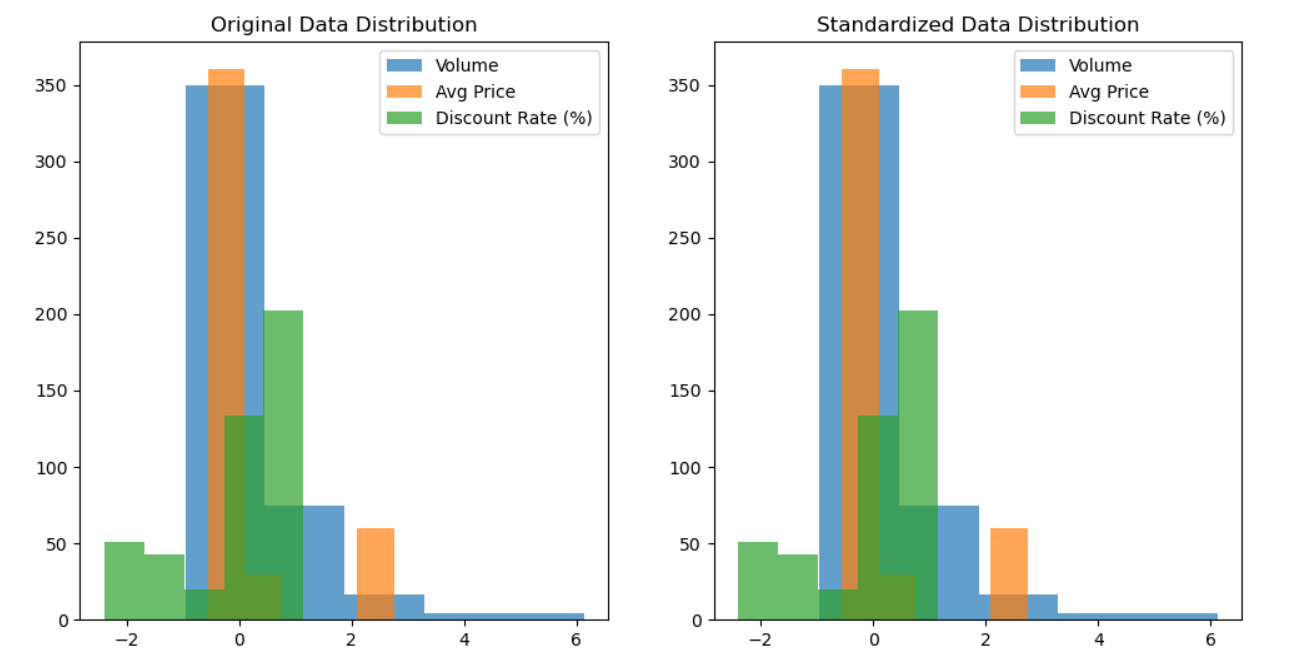
* Standardize the numerical columns using the formula: z=x-mu/sigma.





* Show before and after comparisons of the data distributions.

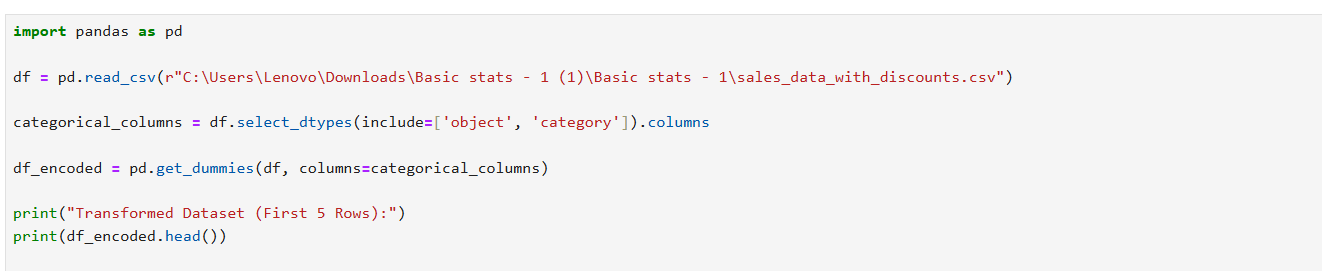


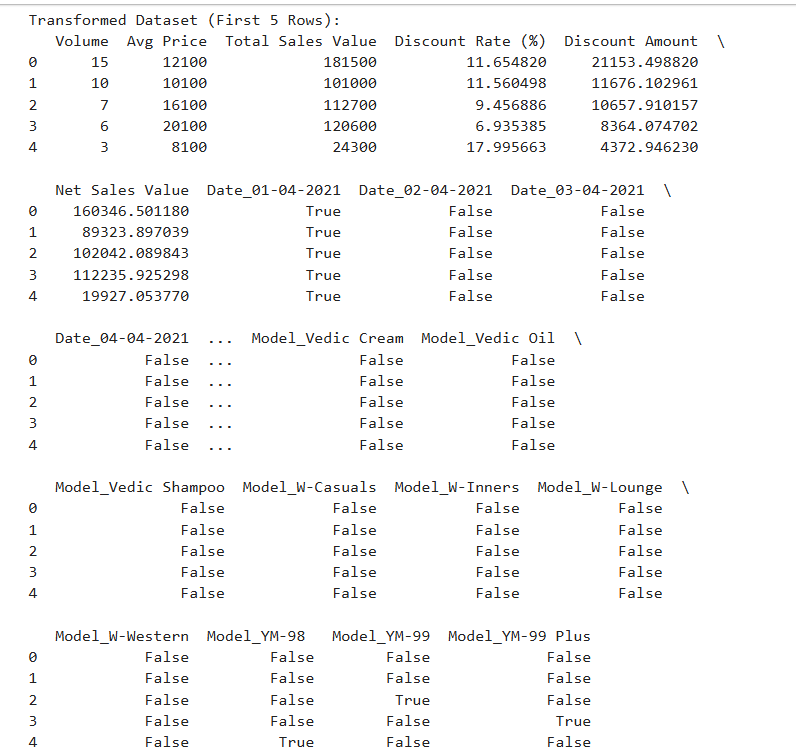


#### Conversion of Categorical Data into Dummy Variables

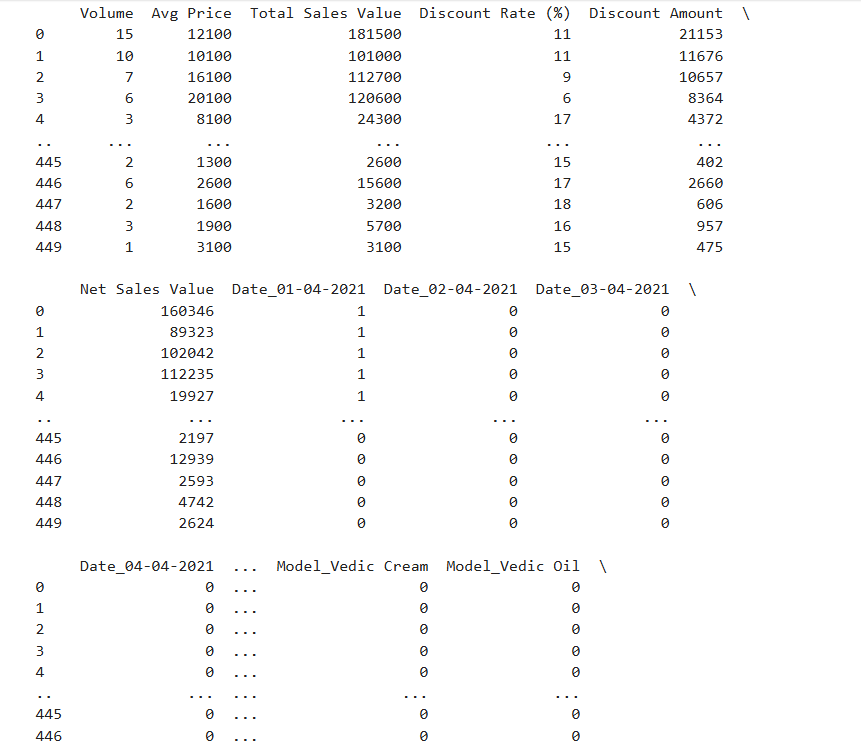
* **Objective**: To transform categorical variables into a format that can be provided to ML algorithms.
* In data analysis and machine learning, converting categorical data into dummy variables also known as one-hot encoding is essential because many algorithms require numerical input. Categorical variables represent data that can be divided into specific categories, but these categories are not inherently ordinal or numerical. To effectively use categorical data in models, we need to transform them into a numerical format. One-hot encoding achieves this by creating binary (0 or 1) columns for each category in the original categorical variable.

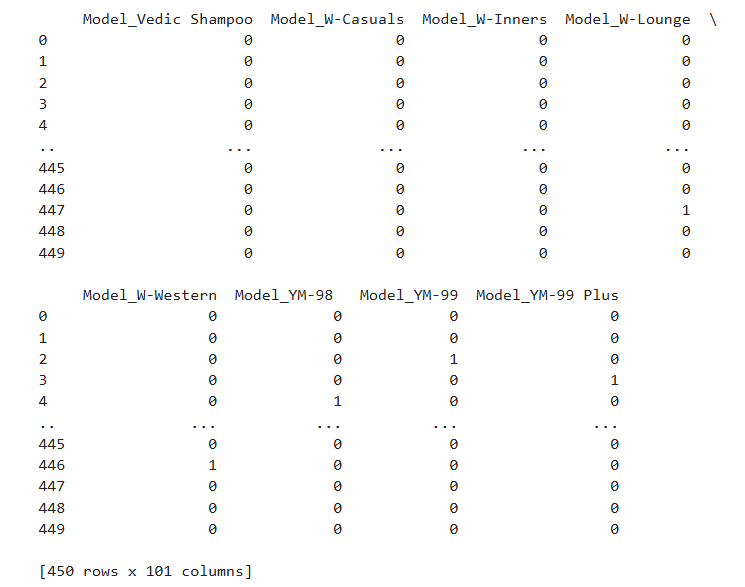
* Applying one-hot encoding to the categorical columns and creating binary (0 or 1) columns for each category.





Screenshot 2025-02-16 163339





Conclusion

* The key findings from the descriptive analytics and data visualizations revealed distinct distribution patterns across numerical variables, including log-normal and Poisson distributions, along with the detection of outliers at various bounds. Categorical data analysis highlighted prevalent categories and potential biases, while standardization effectively scaled numerical variables to a common range, ensuring balanced contributions from all features. One-hot encoding transformed categorical variables into binary vectors, facilitating their inclusion in machine learning algorithms. These preprocessing steps ensured that the data was clean, consistent, and well-prepared for accurate and reliable analysis and modeling.
* Data preprocessing is a pivotal step in data analysis and machine learning, crucial for enabling accurate and reliable modeling. Among the essential preprocessing techniques are standardization and one-hot encoding. Standardization scales data to a common range, preventing features with larger ranges from overshadowing others, thereby ensuring a balanced contribution. One-hot encoding converts categorical variables into binary vectors, making them suitable for processing by models. These techniques help mitigate data leakage, enhance model interpretability, and enable more efficient computation. By applying standardization and one-hot encoding, you can ensure your data is clean, consistent, and primed for effective analysis and modeling