Basic Statistics

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

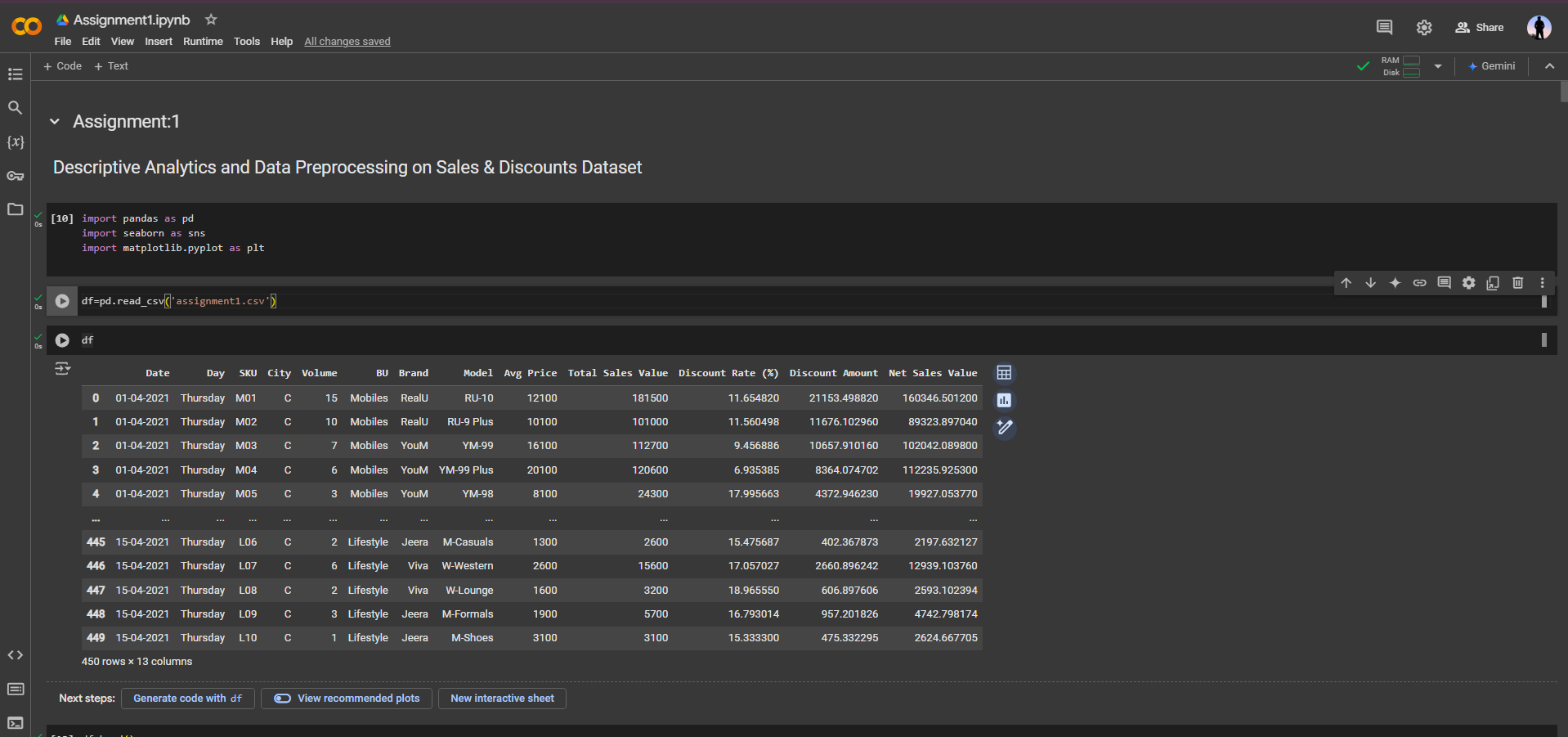
#### **Introduction**

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

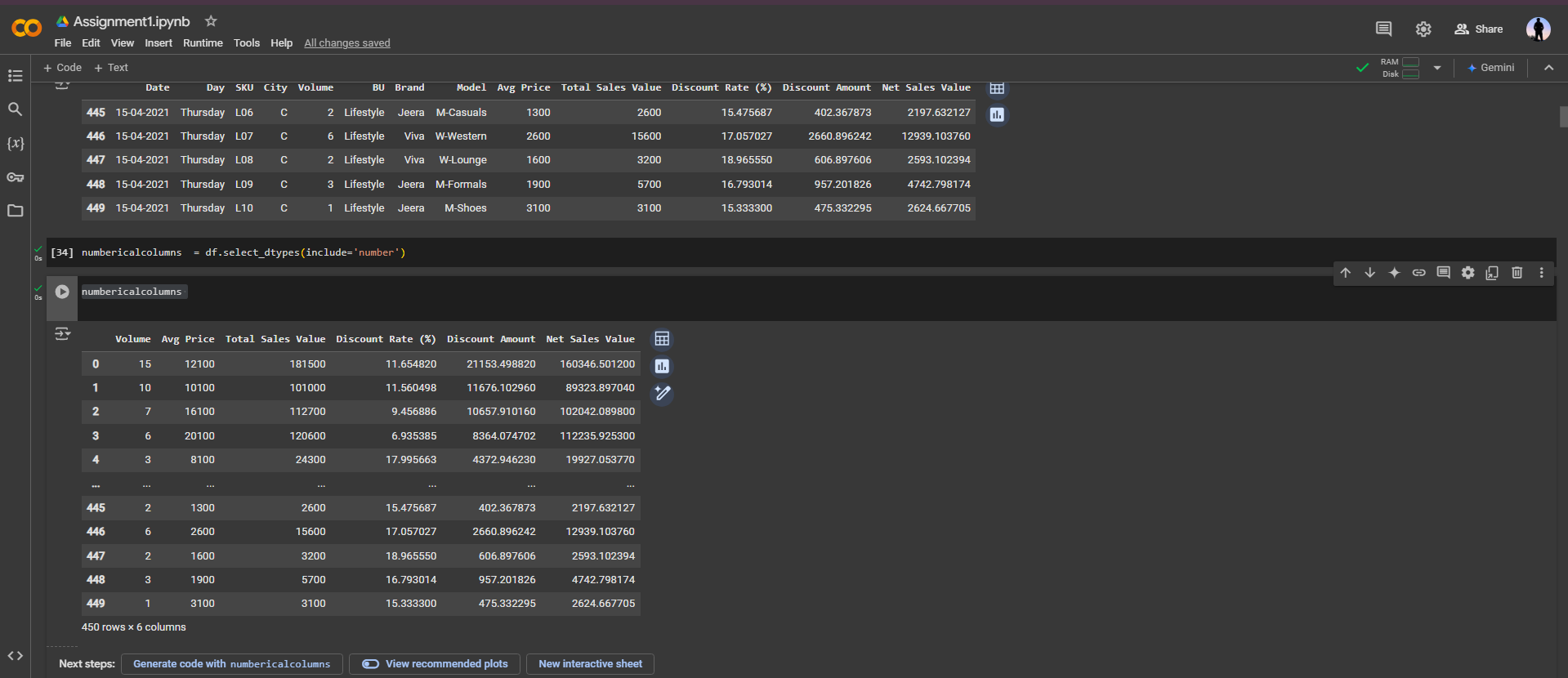
#### **Descriptive Analytics for Numerical Columns**

**Objective: To compute and analyse basic statistical measures for numerical columns in the dataset.**

1. ***Load the dataset into a data analysis tool or programming environment (e.g., Python with pandas library).***

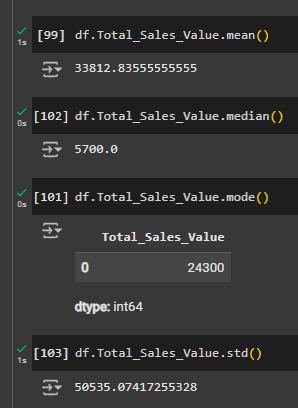
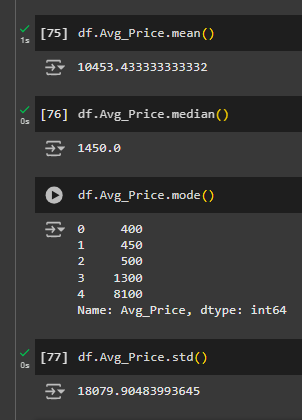
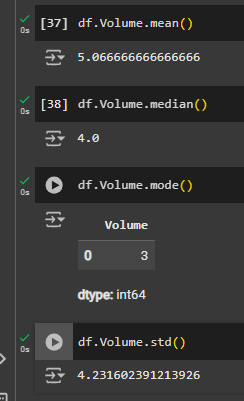


1. ***Identify numerical columns in the dataset.***

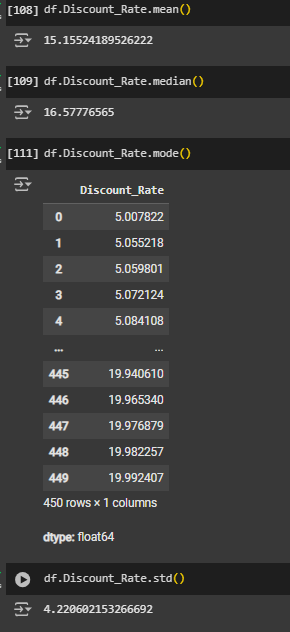
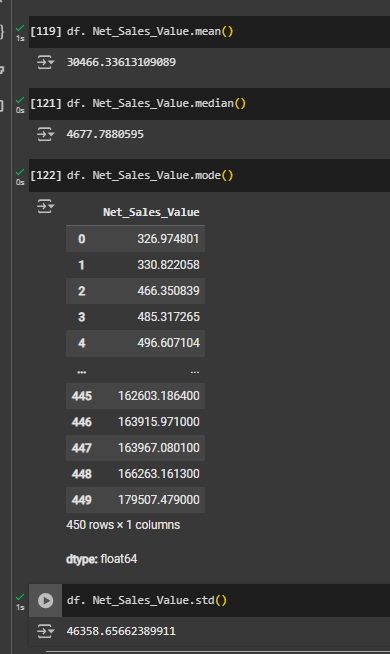
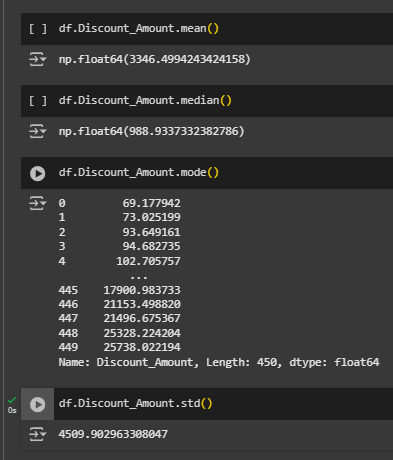


1. ***Calculate the mean, median, mode, and standard deviation for these columns.***

**Volume**  **Avg Price Total Sales Value**



**Discount Rate Discount Amount Net Sales Value**

****

1. ***Provide a brief interpretation of these statistics.***

The total number of rows are 450

There are no Null Value or NA Data in the Entries

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Relation** | **Outliers and Skewness Behaviour** |
| **Volume** | 5 | 4 | Mean<Median | Right Skewed, Outlier Present at Right hand Side |
| **Avg Price** | 10453 | 1450 | Mean>Median | Left Skewed, Outlier Present at Left hand Side |
| **Total Sales Value** | 33812 | 5700 | Mean>Median | Left Skewed, Outlier Present at Left hand Side |
| **Discount Rate** | 15 | 16 | Mean<Median | Right Skewed, Outlier Present at Right hand Side |
| **Discount Amount** | 3346 | 988 | Mean>Median | Left Skewed, Outlier Present at Left hand Side |
| **Net Sales Value** | 30466 | 4677 | Mean>Median | Left Skewed, Outlier Present at Left hand Side |

#### **Data Visualization**

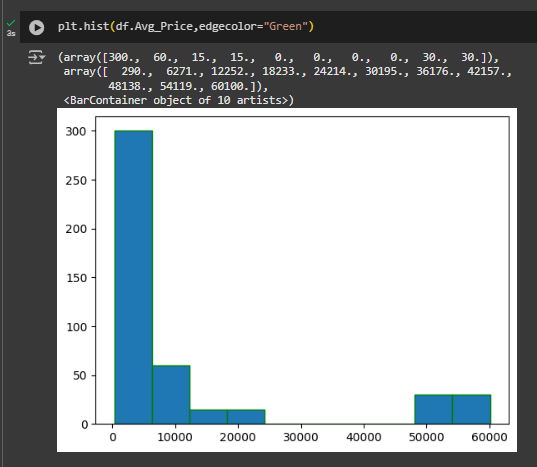
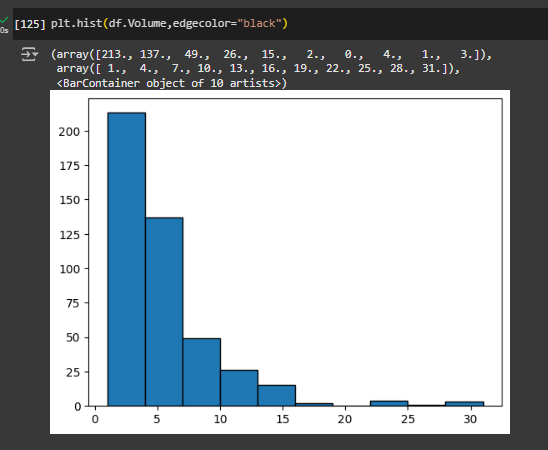
**Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.

1. ***Histograms****:*

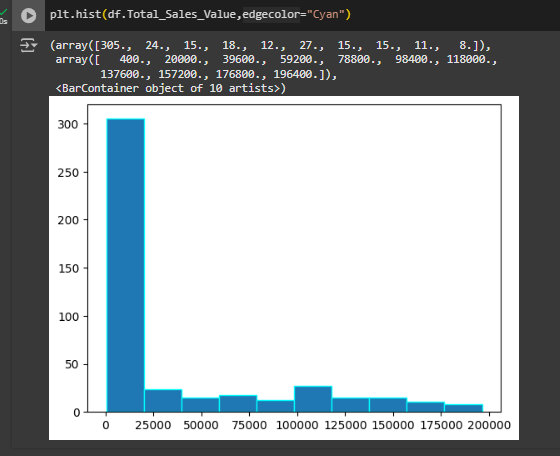
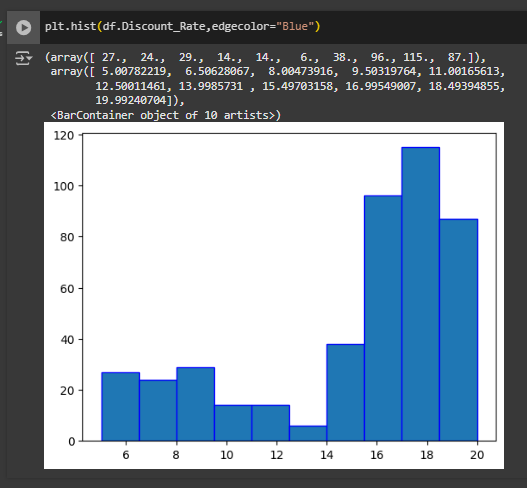
Plot histograms for each numerical column.

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

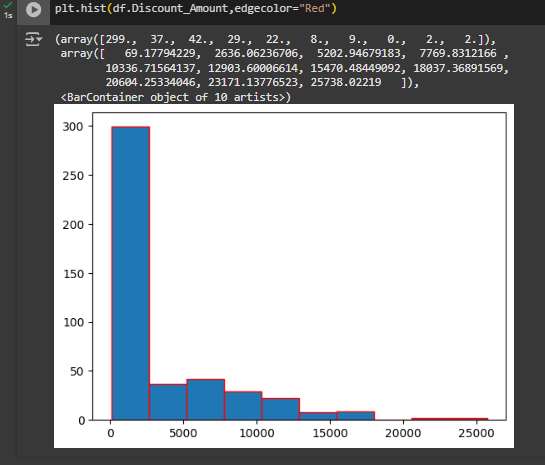
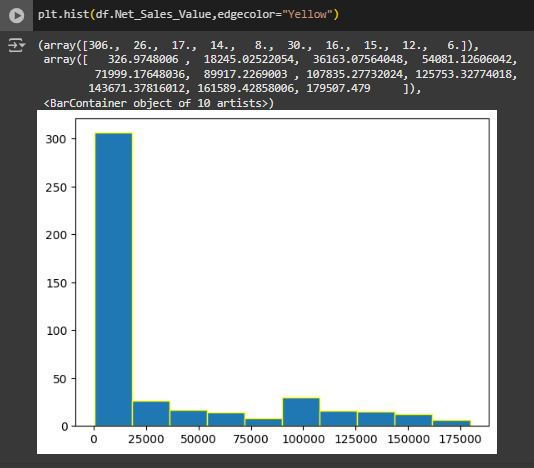
Volume Avg Price



Total Sales Value Discount Rate

Discount Amout Value Net Sales Value

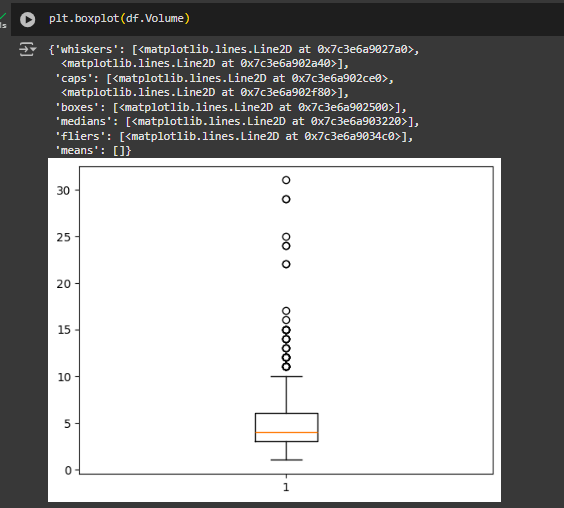
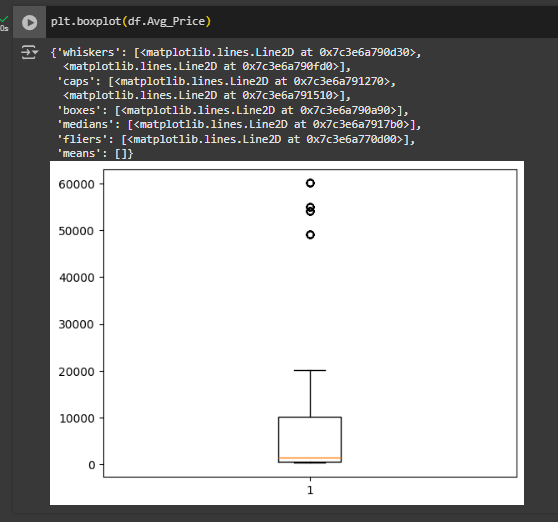
 

|  |  |
| --- | --- |
|  | **Outliers and Skewness Behaviour** |
| **Volume** | Right Skewed, Outlier Present at Right hand Side |
| **Avg Price** | Left Skewed, Outlier Present at Left hand Side |
| **Total Sales Value** | Left Skewed, Outlier Present at Left hand Side |
| **Discount Rate** | Right Skewed, Outlier Present at Right hand Side |
| **Discount Amount** | Left Skewed, Outlier Present at Left hand Side |
| **Net Sales Value** | Left Skewed, Outlier Present at Left hand Side |

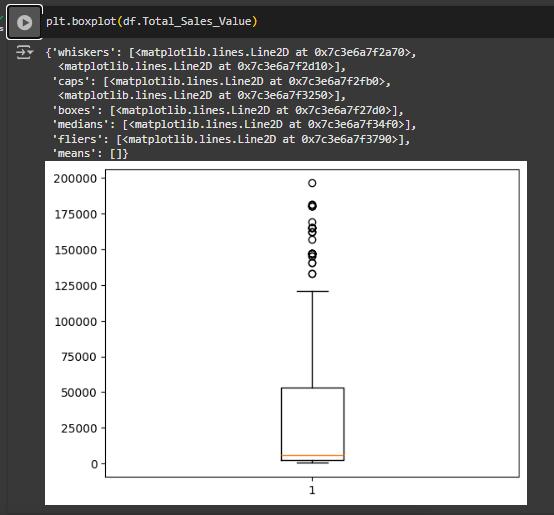
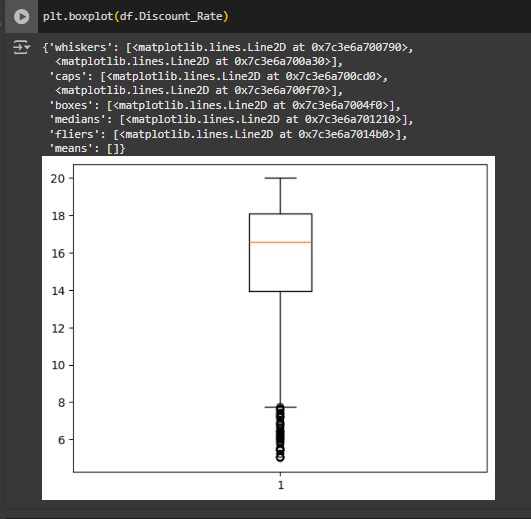
1. ***Boxplots****:*

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

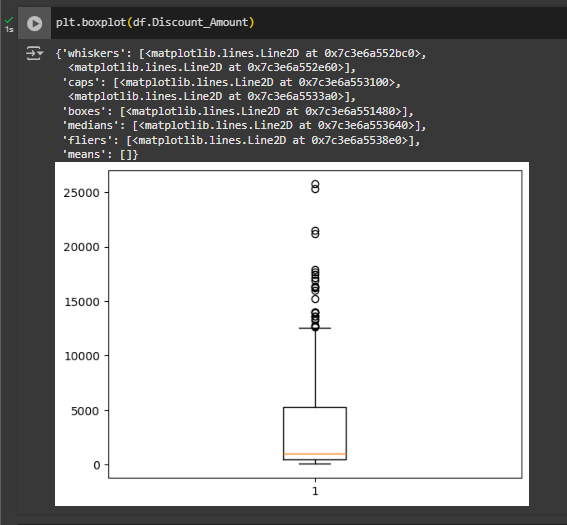
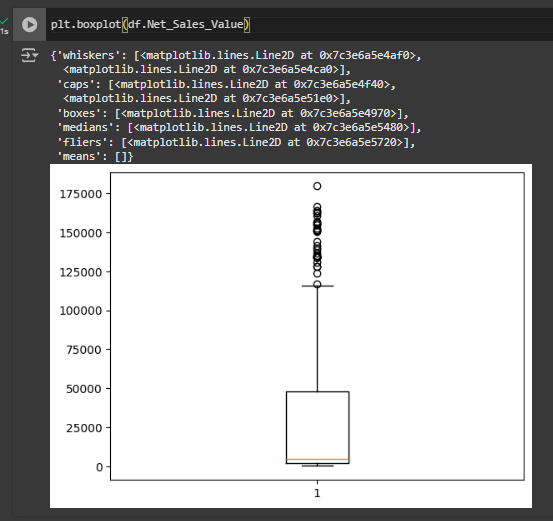
Volume Avg Price

Total Sales Value Discount Rate

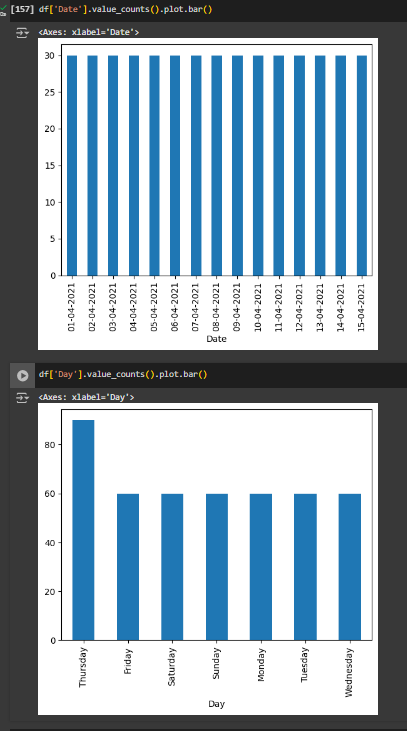
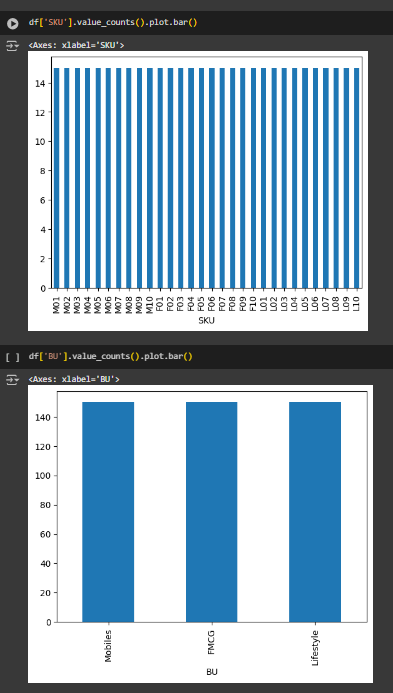
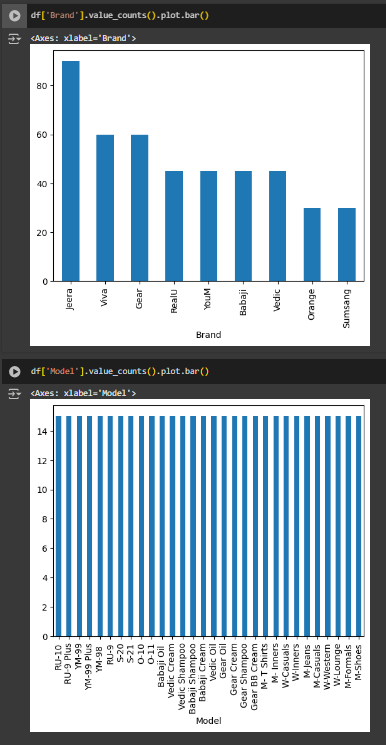
Discount Amount Net Sales Value

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

|  |  |
| --- | --- |
|  | **Outliers Behaviour** |
| **Volume** | Outliers are present on the upper extreme part |
| **Avg Price** | Outliers are present on the upper extreme part |
| **Total Sales Value** | Outliers are present on the upper extreme part |
| **Discount Rate** | Outliers are present on the Lower extreme part |
| **Discount Amount** | Outliers are present on the upper extreme part |
| **Net Sales Value** | Outliers are present on the upper extreme part |

1. **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.

#### 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.

#### 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.

#### 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.