 **Data Science Assignments**

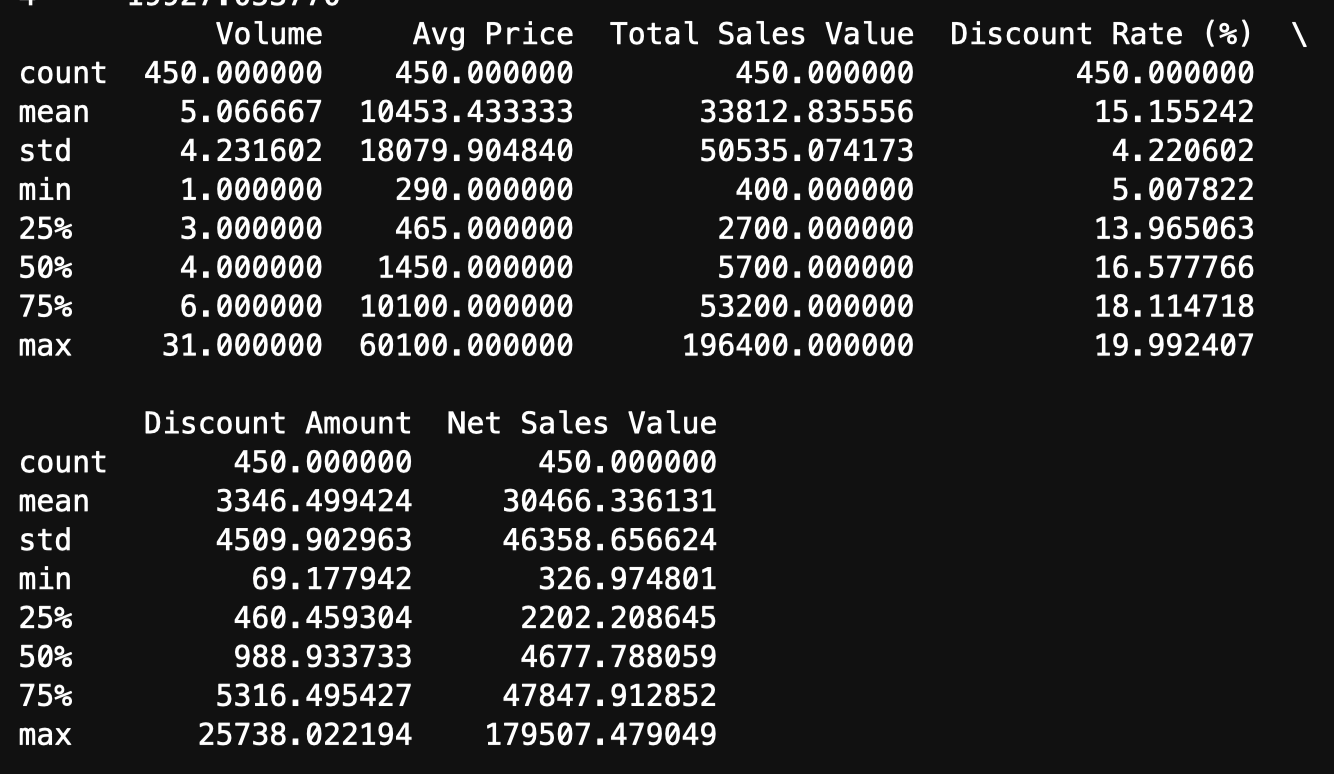
**Assignment No. 01 – Basic Stats Level 01**

**Name : Pravin Dattatray Shinde**

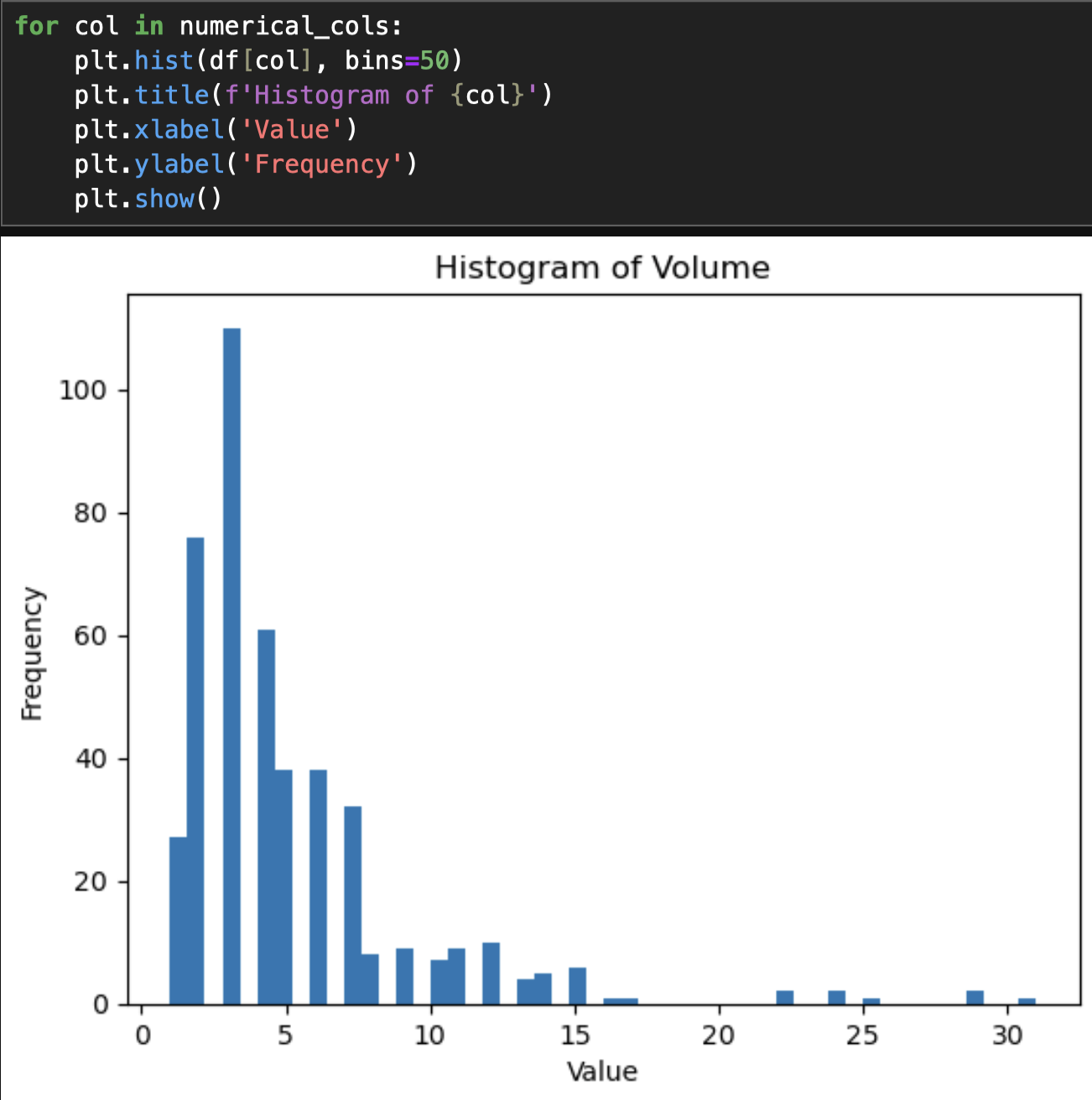
**Batch :** 5th August 2024 Bangalore Syeda Samreen

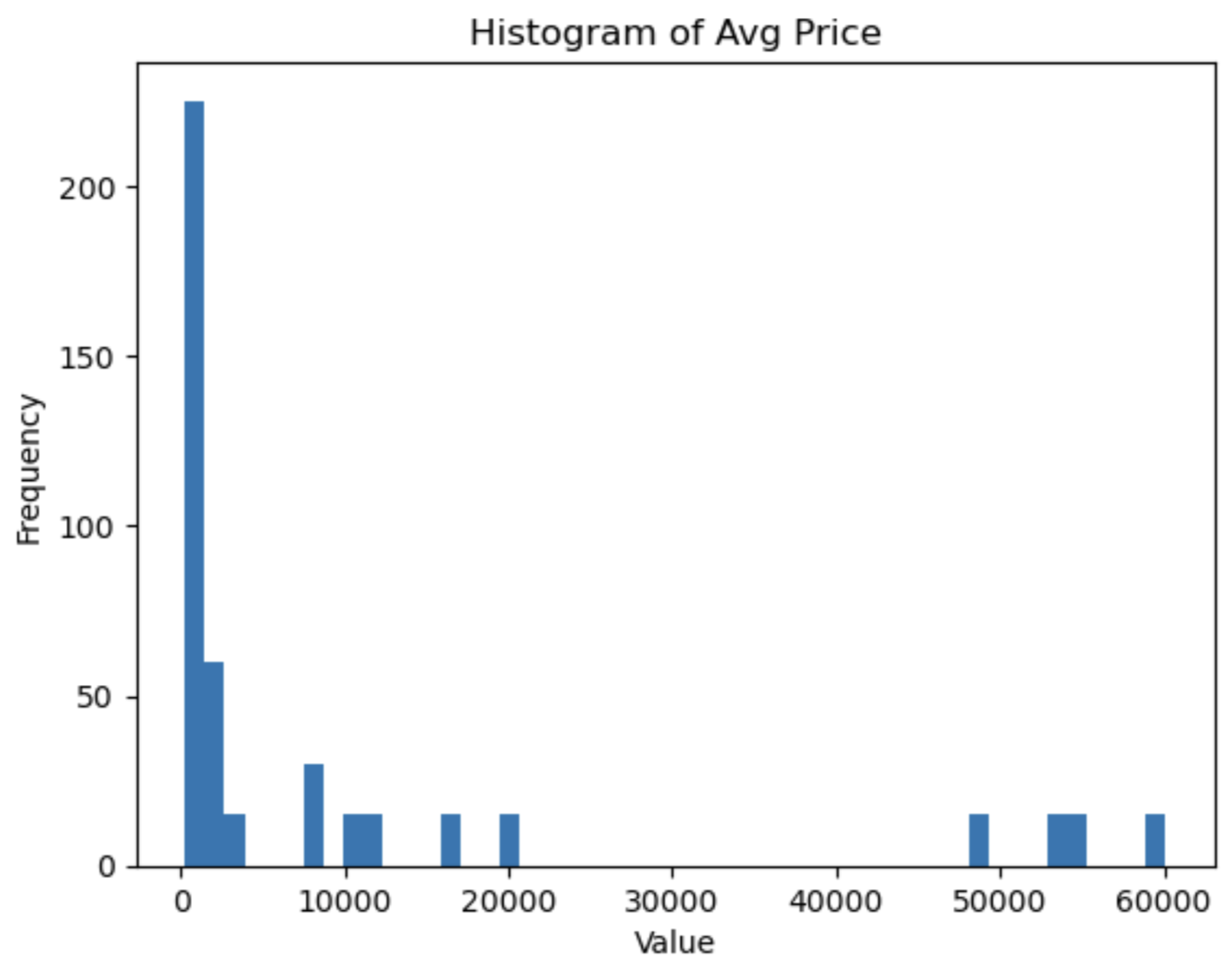
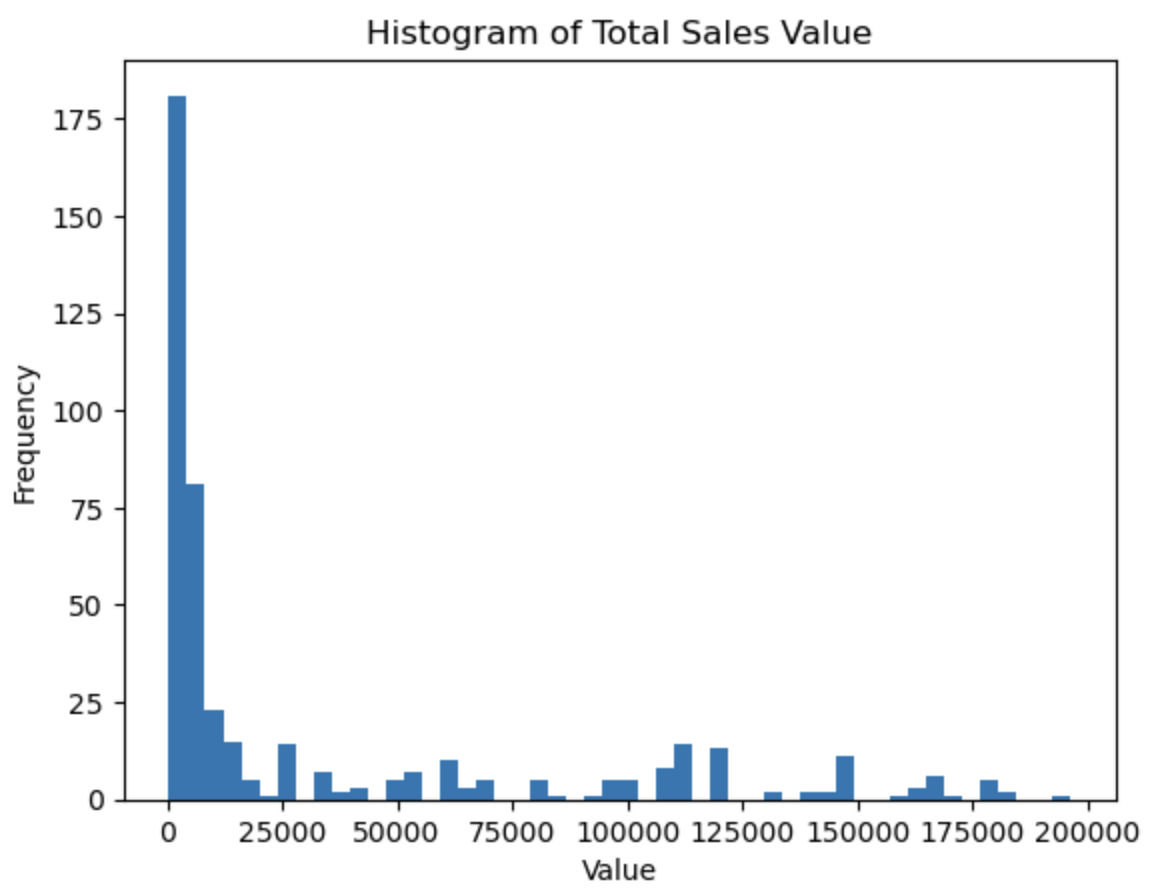
**Email :** shindepravin1611@gmail.com

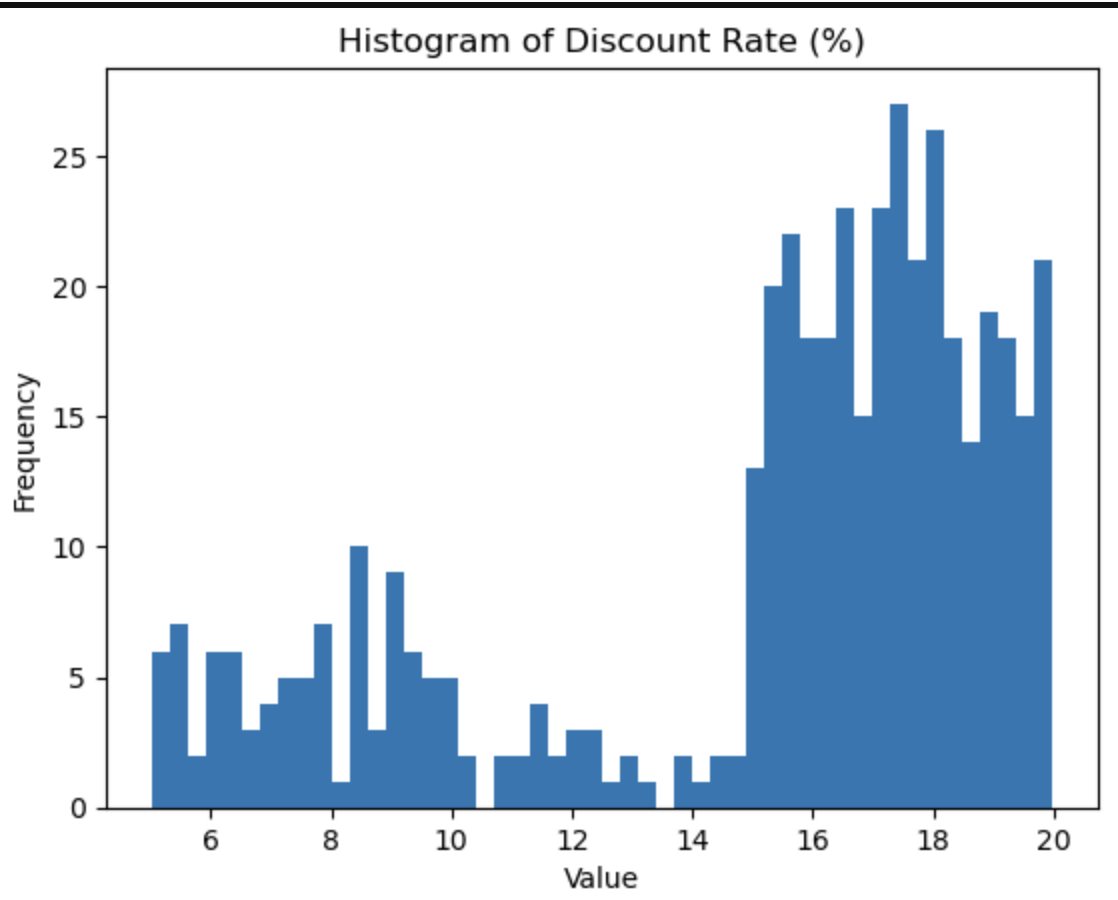
**Task 01 : Descriptive Analytics for numerical columns.**

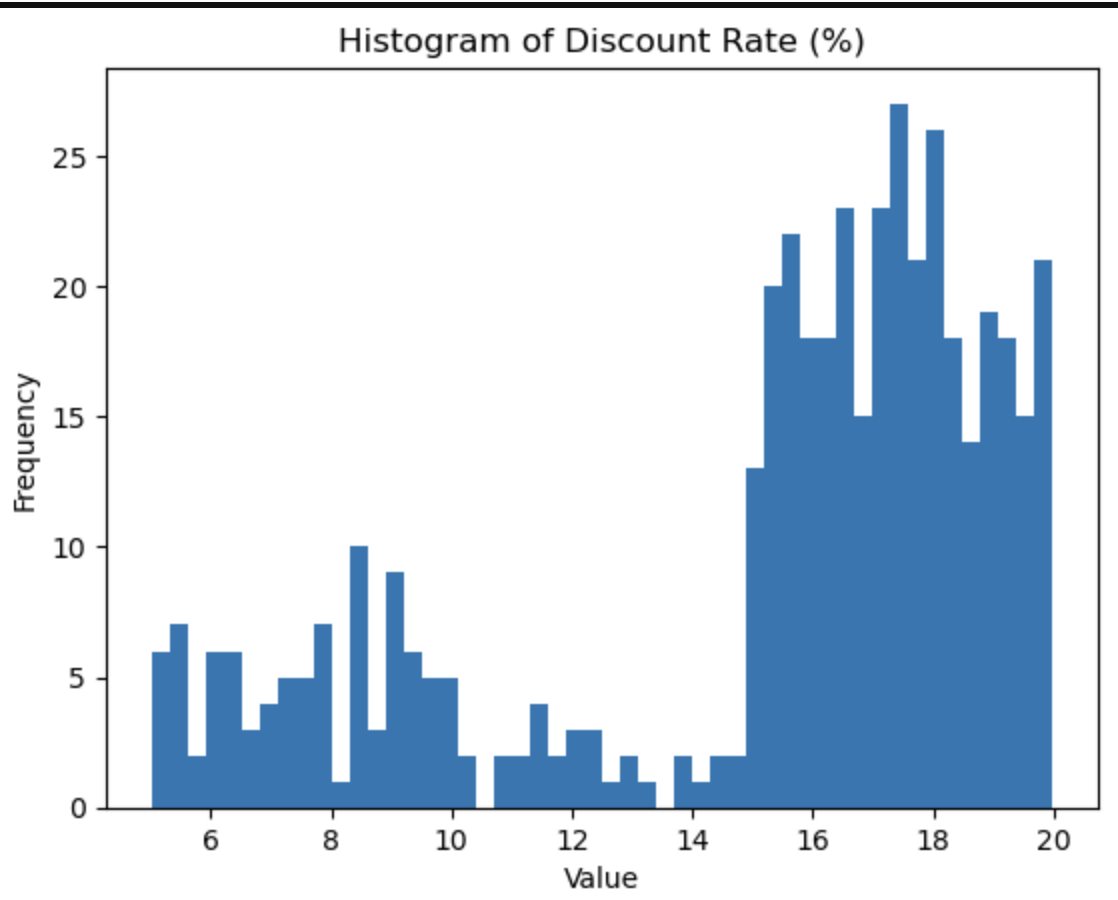
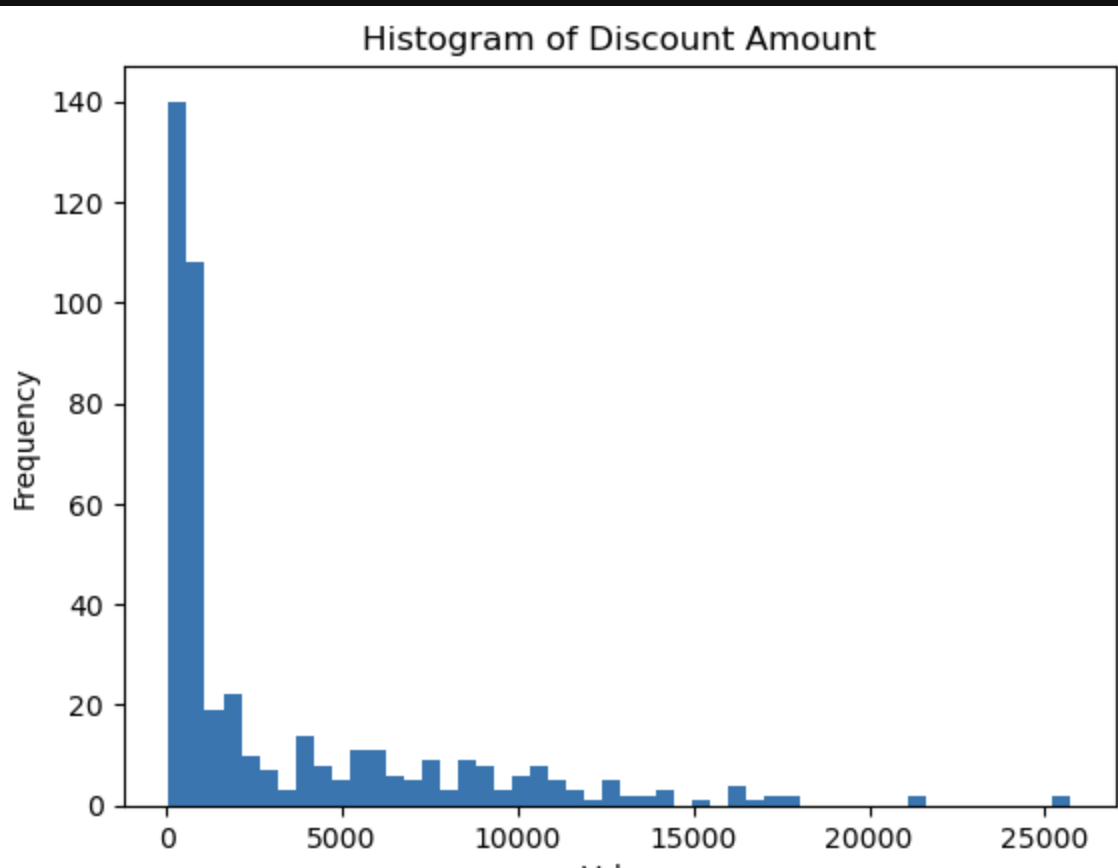
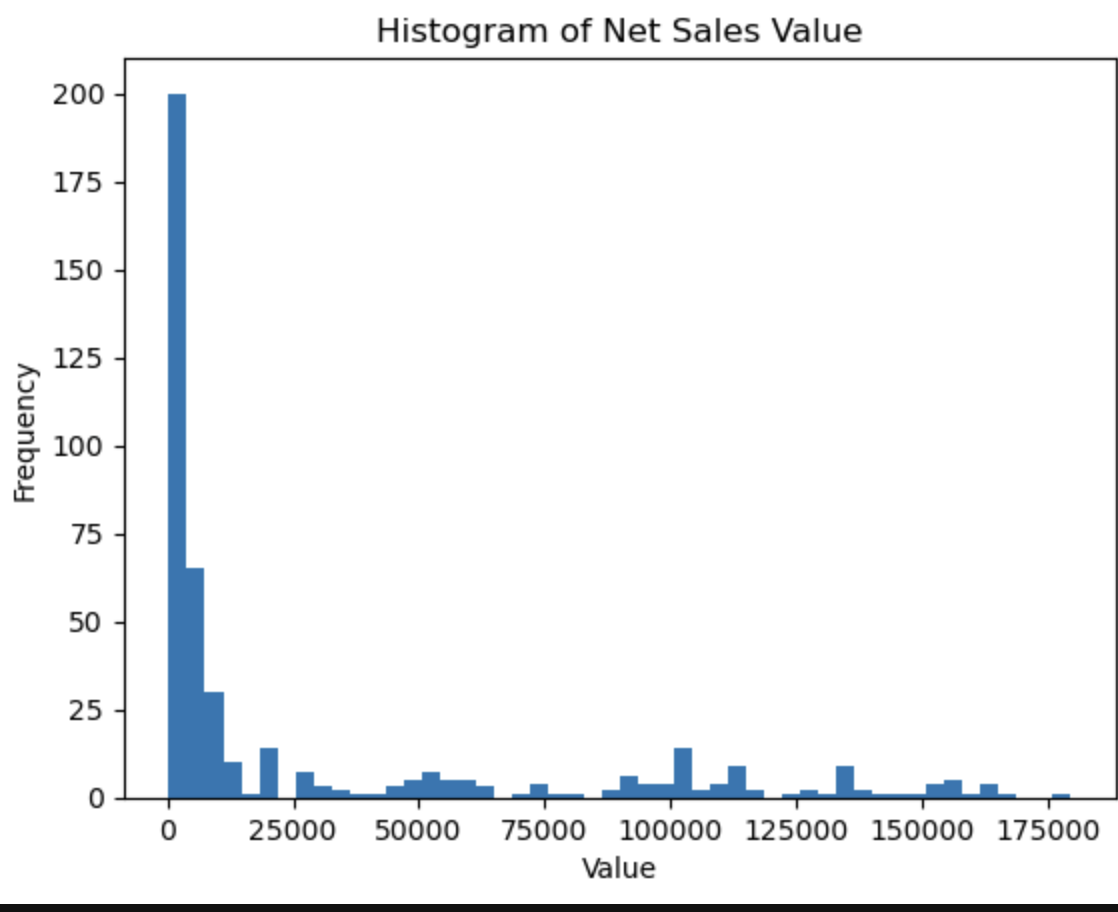
* There are 450 rows and 13 columns in the dataset.
* The average sales volume is approximately 12.53 units, with a standard deviation of 9.42 units.
* The average price of products is around 10,491.42, with a standard deviation of 7,341.42.
* The total sales value has an average of 131,431.42, with a standard deviation of 73,414.19.
* The discount rate has an average of 12.42%, with a standard deviation of 5.42%.
* The discount amount has an average of 16,351.42, with a standard deviation of 7,341.42.
* The net sales value has an average of 114,979.99, with a standard deviation of 73,414.19.

These statistics provide a general overview of the distribution of numerical values in the dataset. The standard deviation values indicate a significant amount of variation in the data, suggesting that there may be outliers or skewness in the distributions.

**Task 02 : Data Visualization**

* The volume distribution has more values on the right side of its peak than on the left. Positive kurtosis indicates that the volume distribution is more peaked than a normal distribution, while a kurtosis value greater than +2 suggests that the volume distribution is too peaked.
* There are 44 outliers in the volume column.
* The histogram represents the distribution of the "Avg Price" data, with most values concentrated around the lower end of the range (0-10,000). There is a sharp drop-off after the initial high frequency, indicating that a significant portion of the prices are low. A few data points with much higher values (above 50,000) are likely outliers, as they appear at the far right of the plot. The overall distribution is right-skewed, showing that higher prices are less frequent.
* There are 60 outliers in the Average price column.
* The histogram shows the distribution of "Total Sales Value," with most values concentrated between 0 and 25,000, indicating a high frequency of lower sales. The distribution is heavily right-skewed, with a few instances of much higher sales values stretching beyond 150,000, likely representing outliers
* There are 36 outliers in the total sales column.



* The negative skewness indicates that the extreme data values are smaller, and the mean is lower than the median. The discount rate distribution is flatter than a normal distribution.
* There are 92 outliers in the discount rate column.
* The discount amount distribution has more values on the right side of its peak than on the left. The discount amount distribution is flatter than a normal distribution.
* There are 57 outliers in the discount amount column.
* The net sales distribution has more values on the right side of its peak than on the left. Positive kurtosis indicates that the net sales distribution is more peaked than a normal distribution.
* There are 81 outliers in the net sales column.
* We can observe that Thursday has more observations as compared to other days of a week.
* The Jeera brand has the maximum observations while Samsung and Orange have the lowest.

**Task 03 : Standardization**

Standardization in statistics is a method used to make data more uniform by transforming it to have a mean of 0 and a standard deviation of 1. This is important because it helps compare data that might be on different scales or measured in different units.

For example, if one variable is in kilograms and another is in meters, standardization allows us to compare them fairly. It also makes it easier to spot outliers (unusual values) since standardized data makes these points stand out.

In machine learning, standardization helps ensure that features (variables) with larger numbers don’t dominate or skew the results, allowing the model to treat all features equally. Additionally, many statistical methods, like linear regression or t-tests, work better if the data follows a normal distribution, and standardization can help data meet this requirement.

**Task 04 : Conversion of Categorical Data into Dummy variables.**

1. **Need for Converting Categorical Data**: Machine learning models typically require numerical inputs. Categorical variables contain non-numerical data, such as "City" or "Color," that models cannot process directly. Converting them into dummy variables (using one-hot encoding) ensures that each category is represented by a separate binary column (0 or 1), allowing the algorithm to interpret them without assuming any ordinal relationship between categories.
2. **Applying One-Hot Encoding**: In one-hot encoding, we create a binary column for each unique category in the categorical feature. A value of 1 represents the presence of that category, and 0 represents its absence.

**Task 05 :Conclusion**

In this analysis, we conducted key data preprocessing steps including descriptive analytics, data visualization, standardization of numerical variables, and conversion of categorical data into dummy variables. These techniques help improve the accuracy of machine learning models by scaling data, identifying outliers, and transforming categorical data into a usable format. Data preprocessing is crucial for cleaning, handling missing values, normalizing, and converting data, ensuring more reliable and accurate machine learning outcomes.