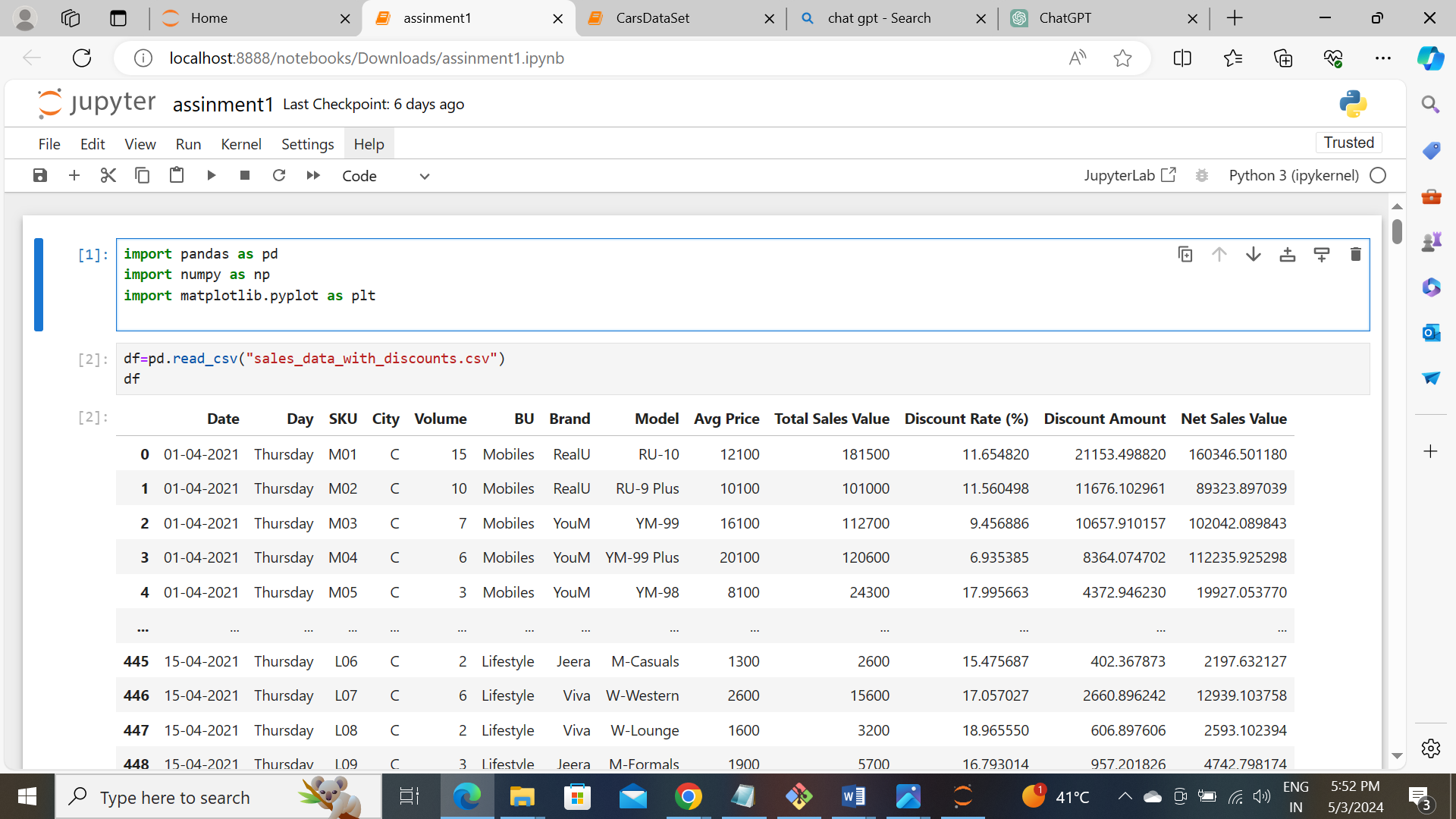
Assainment1

Basic Statistics:-

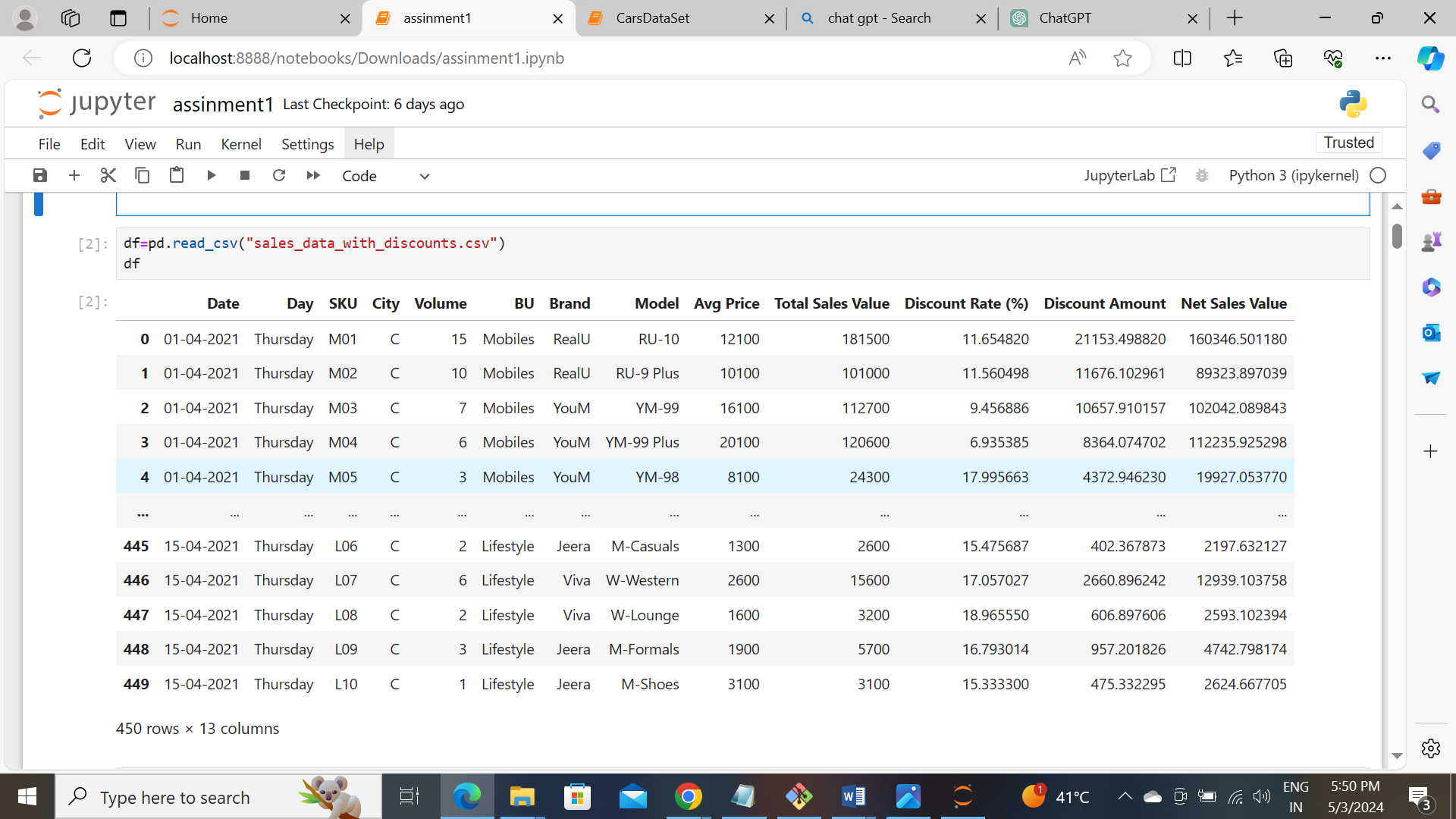
* Objective: To compute and analyze basic statistical measures for numerical columns in the dataset.
* Steps:
  + Load the dataset into a data analysis tool or programming environment (e.g., Python with pandas library).
  + Identify numerical columns in the dataset.
  + Calculate the mean, median, mode, and standard deviation for these columns.
  + Provide a brief interpretation of these statistics.

Practical for the above scenario:

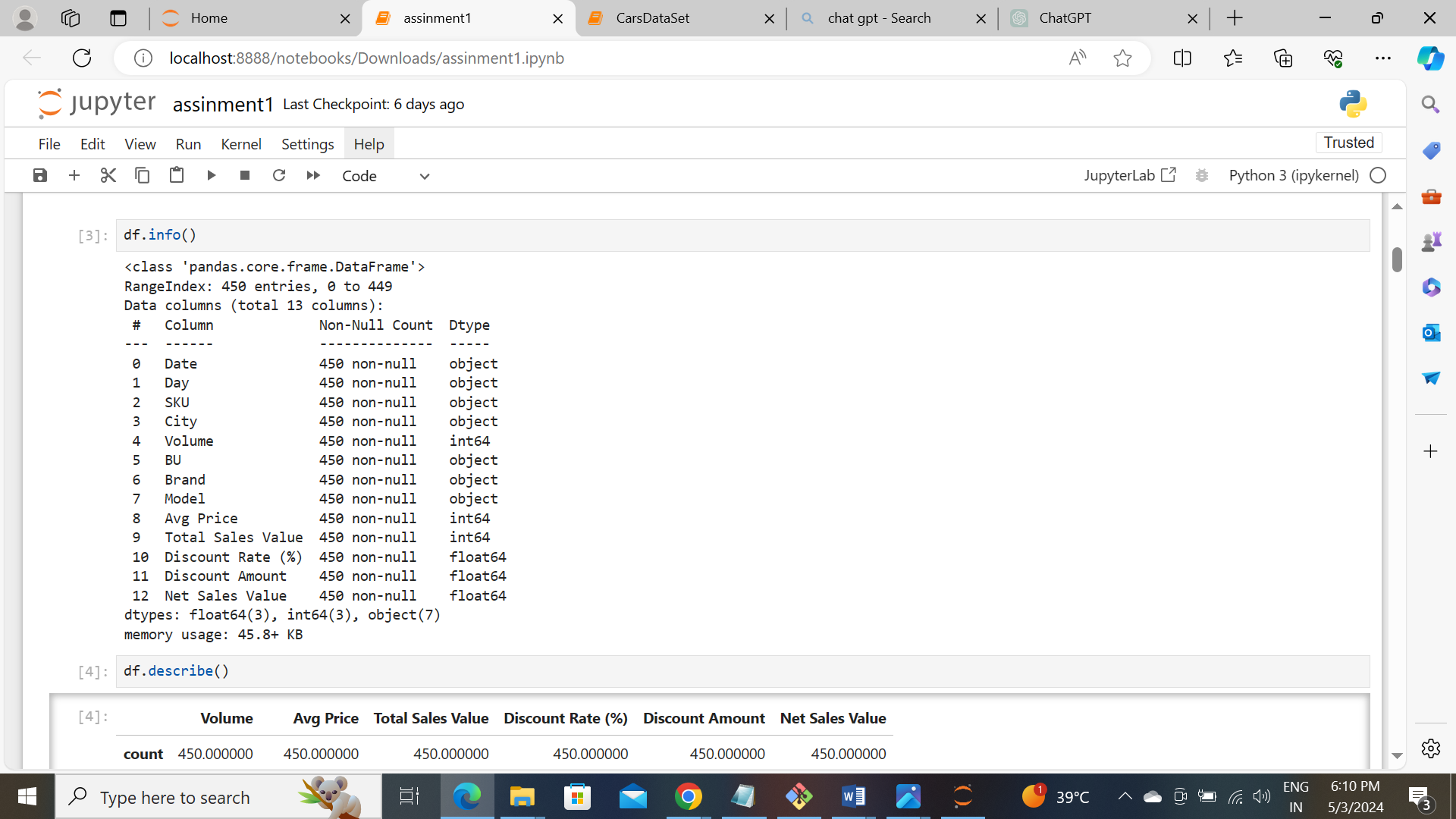
=>Import the Libraries



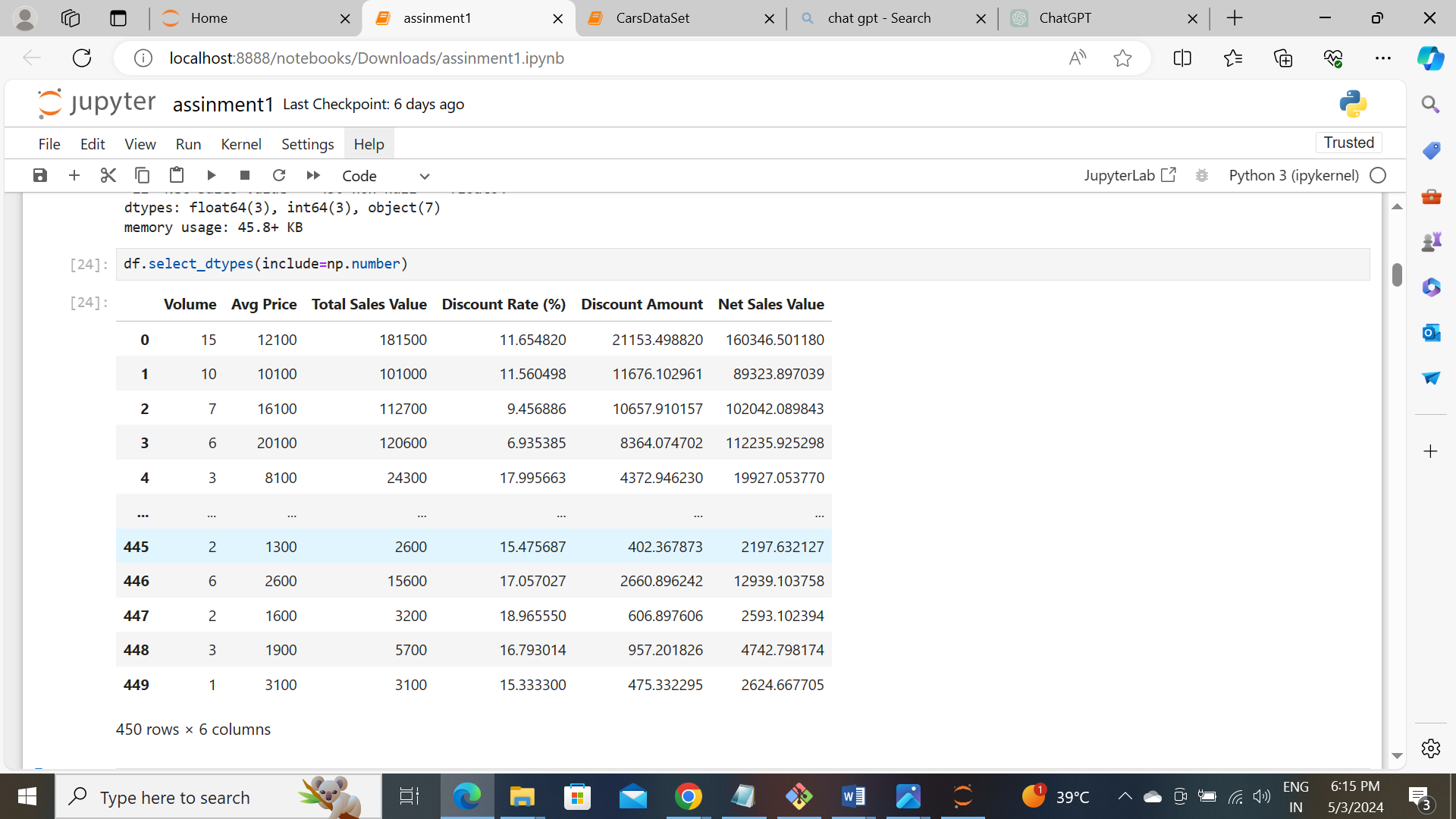
=>Load the Dataset

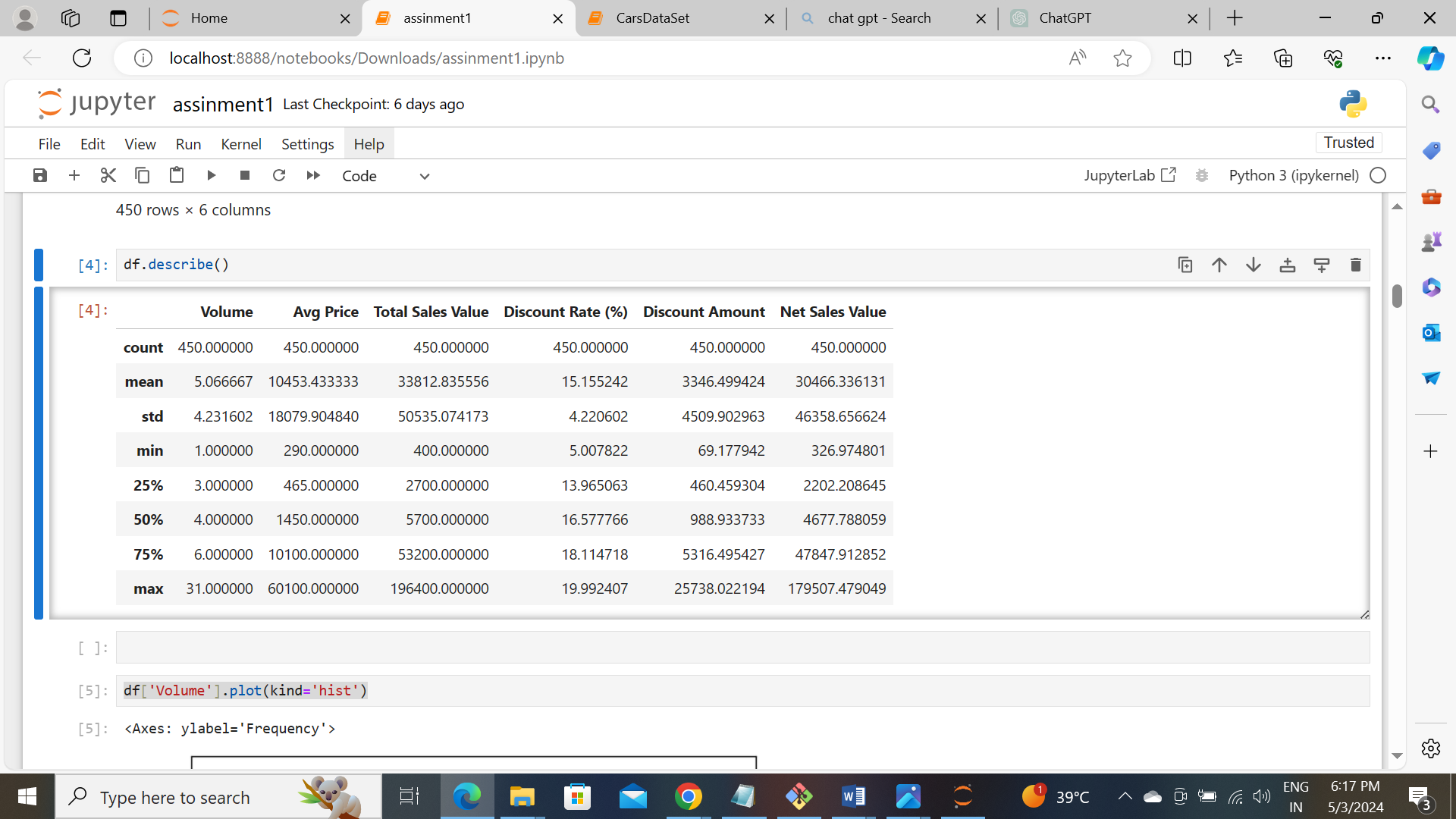


=>Get the information of the Dataset



=>Identifying the numerical columns from dataset



=>Calculating the mean,median,mode and standard deviation of numerical columns

\* It is a small dataset, we have in this numerical and categorical columns. We identified the numerical columns and get its mean,median,mode and standard deviation.

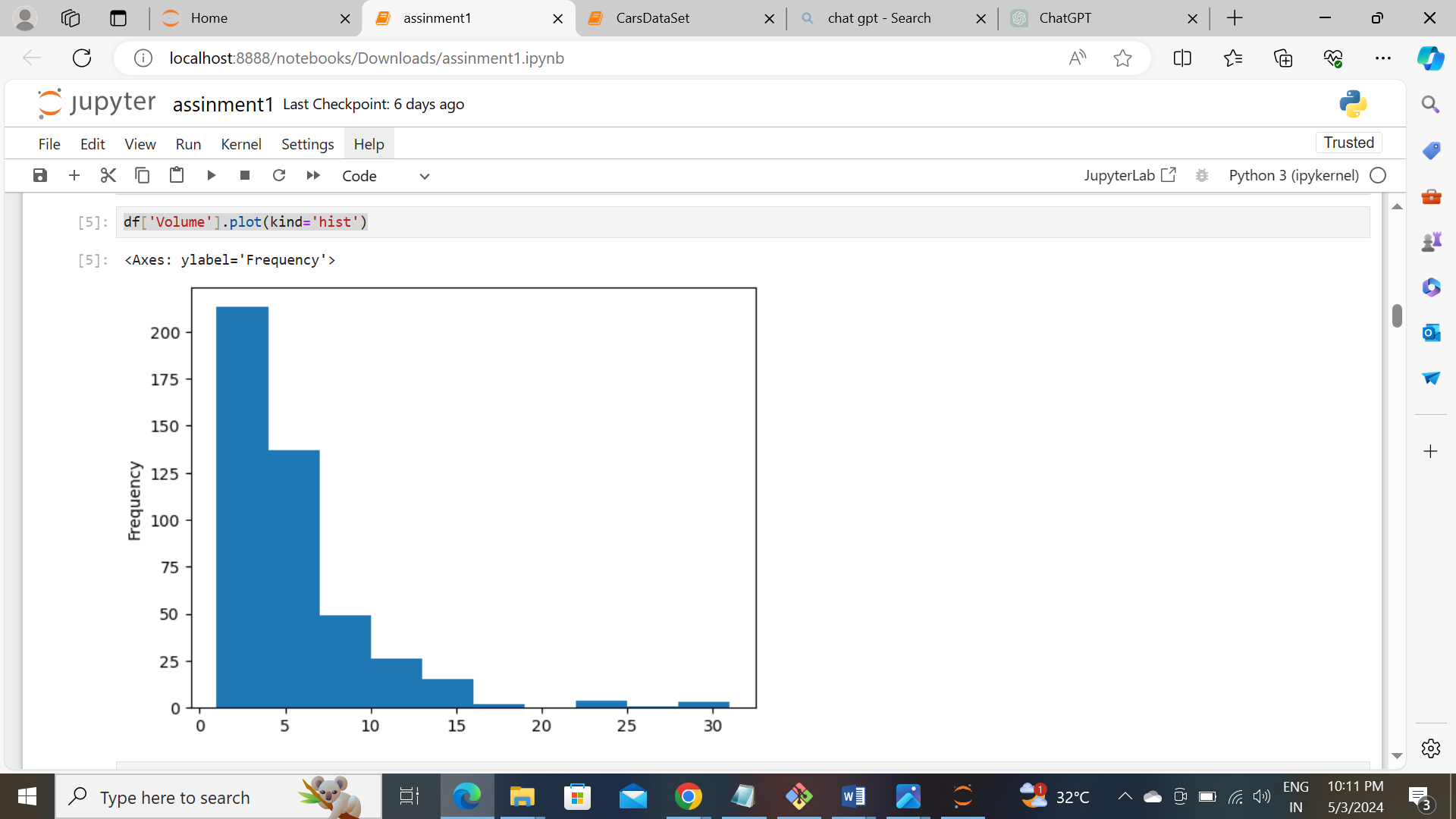
#### Data Visualization

* **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.

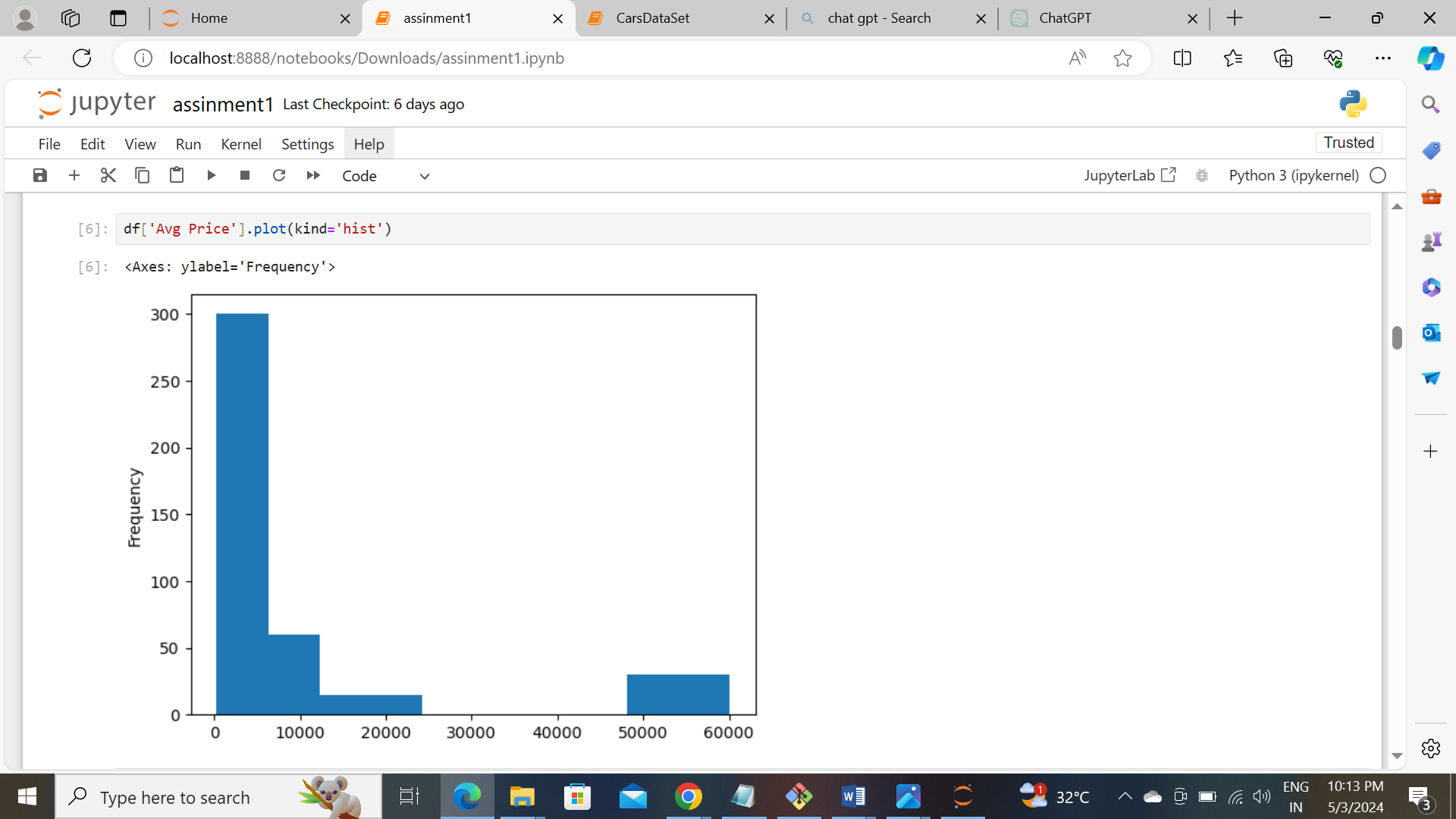
=>We have numerical columns are,

Volume, Avg Price, Total sale value, Discount Rate, Discount Amount, Net sale value.

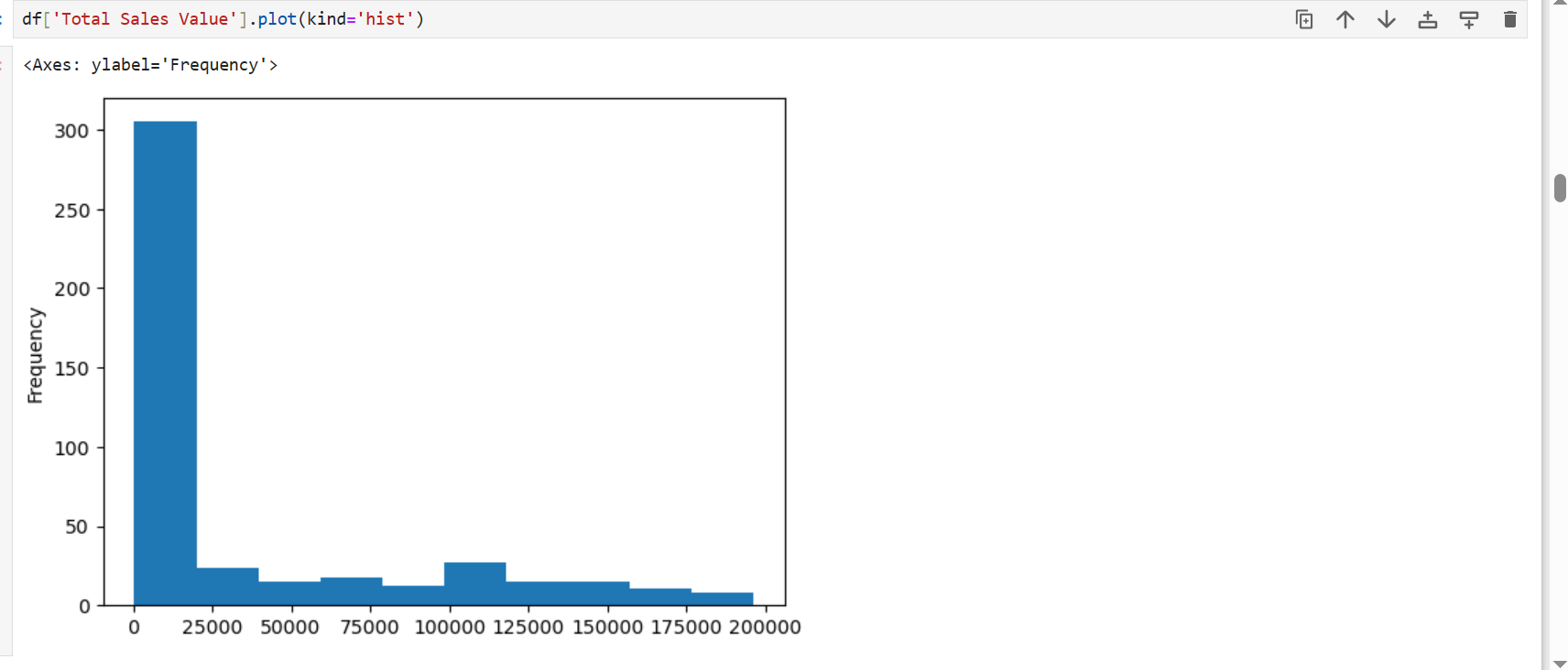
* Histogram of Volume



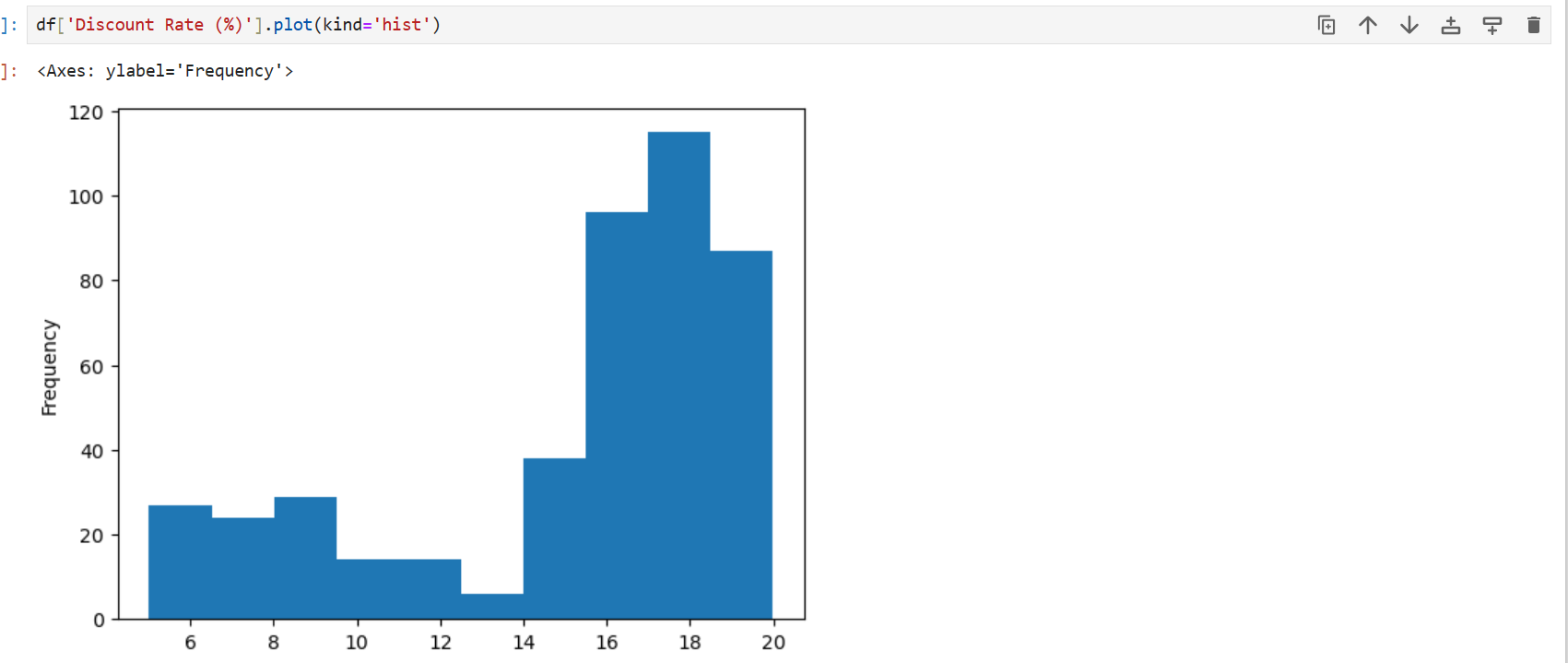
* Histogram of Avg Price



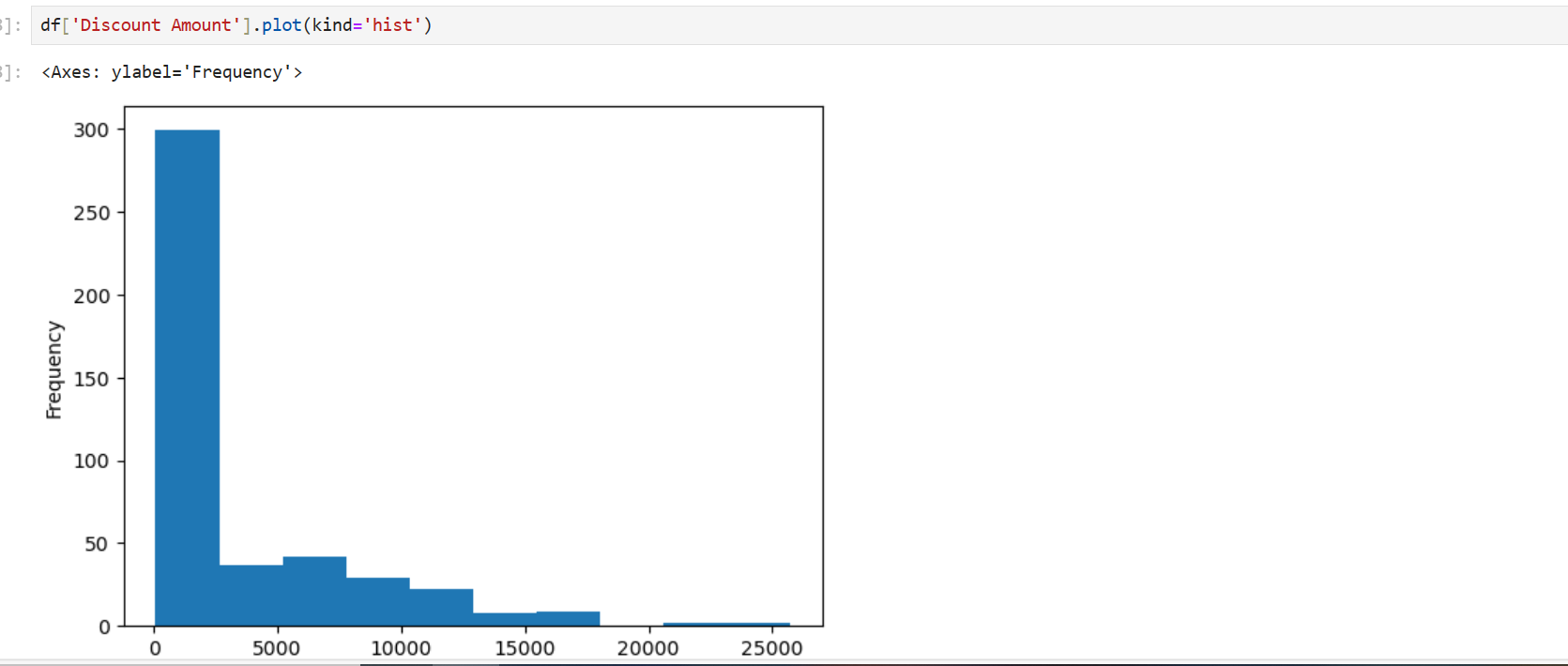
* Histogram of Total sale value



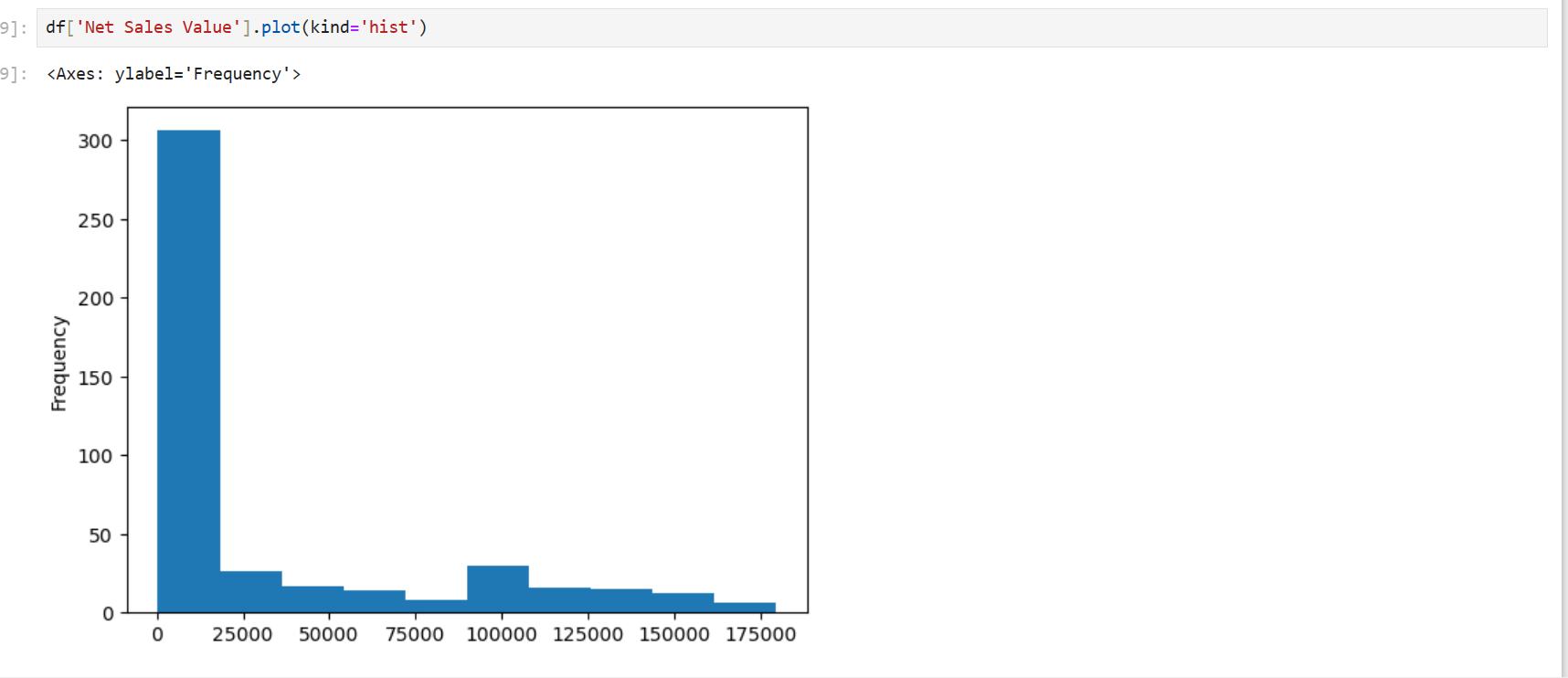
* Histogram of Discount Rate



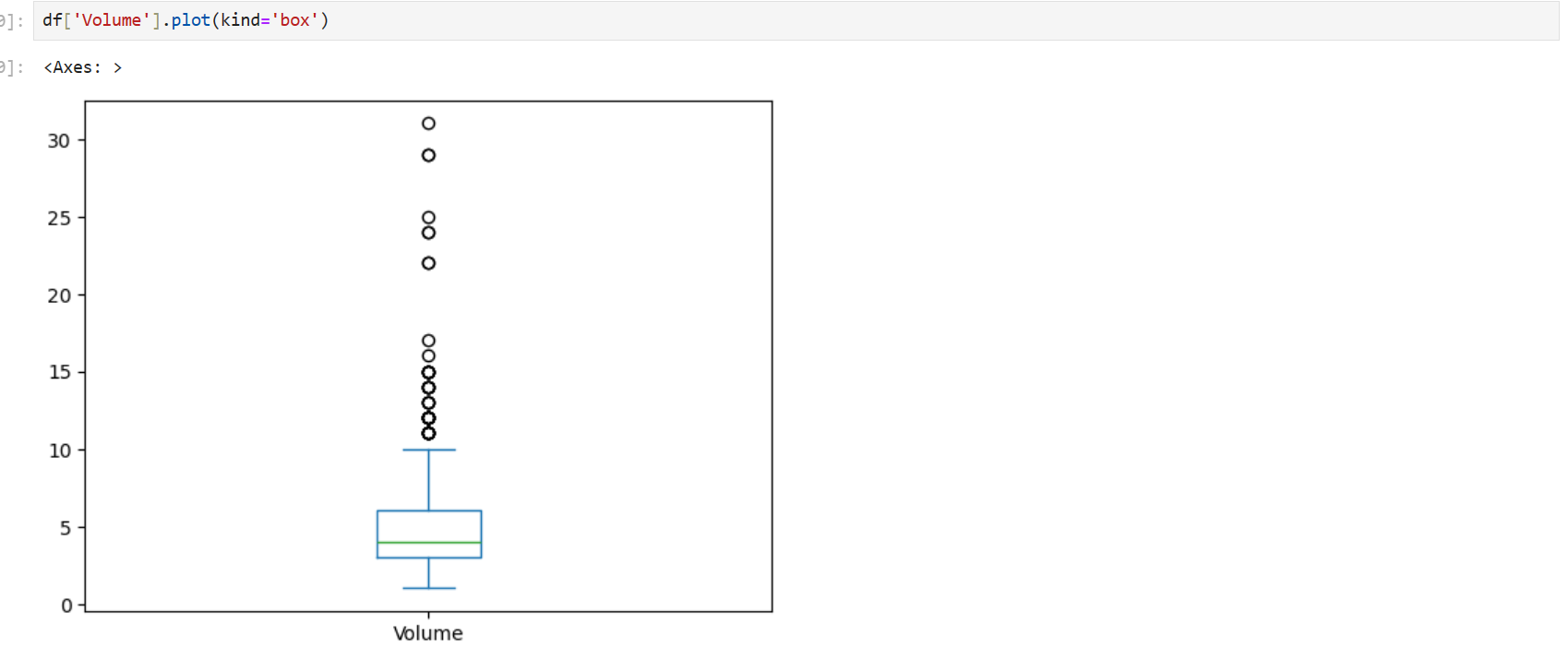
* Histogram of Discount Amount



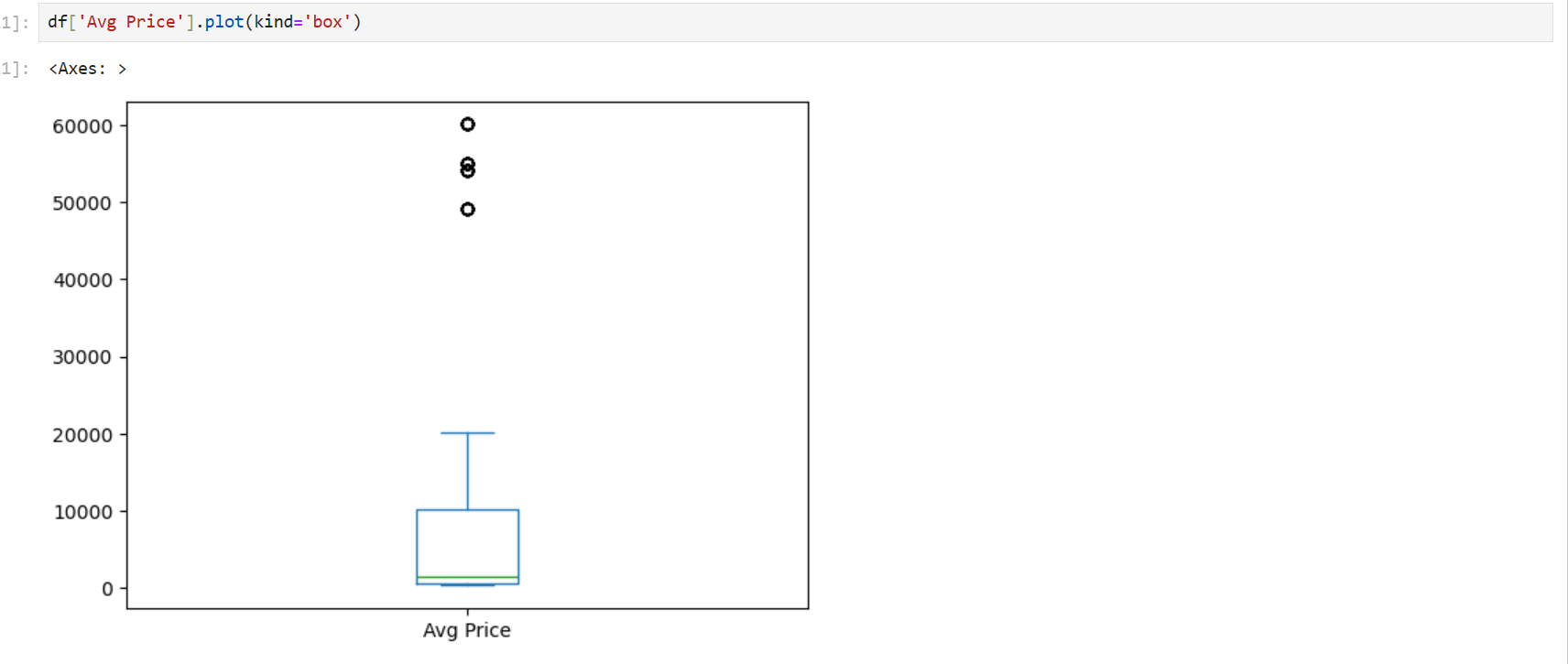
* Histogram of Net sale value



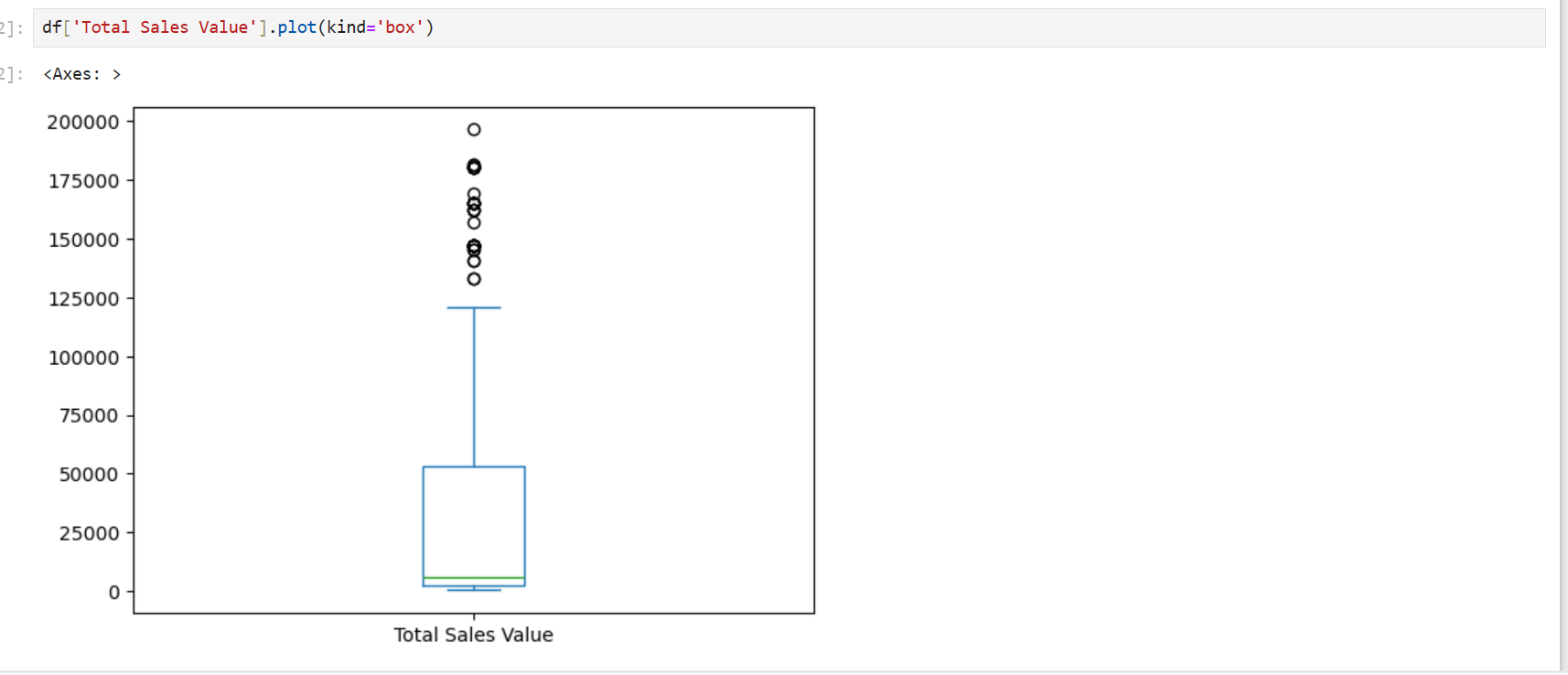
* **Boxplots**:
  + Create boxplots for numerical variables to identify outliers and the interquartile range.
  + Discuss any findings, such as extreme values or unusual distributions.
* Box plot of Volume



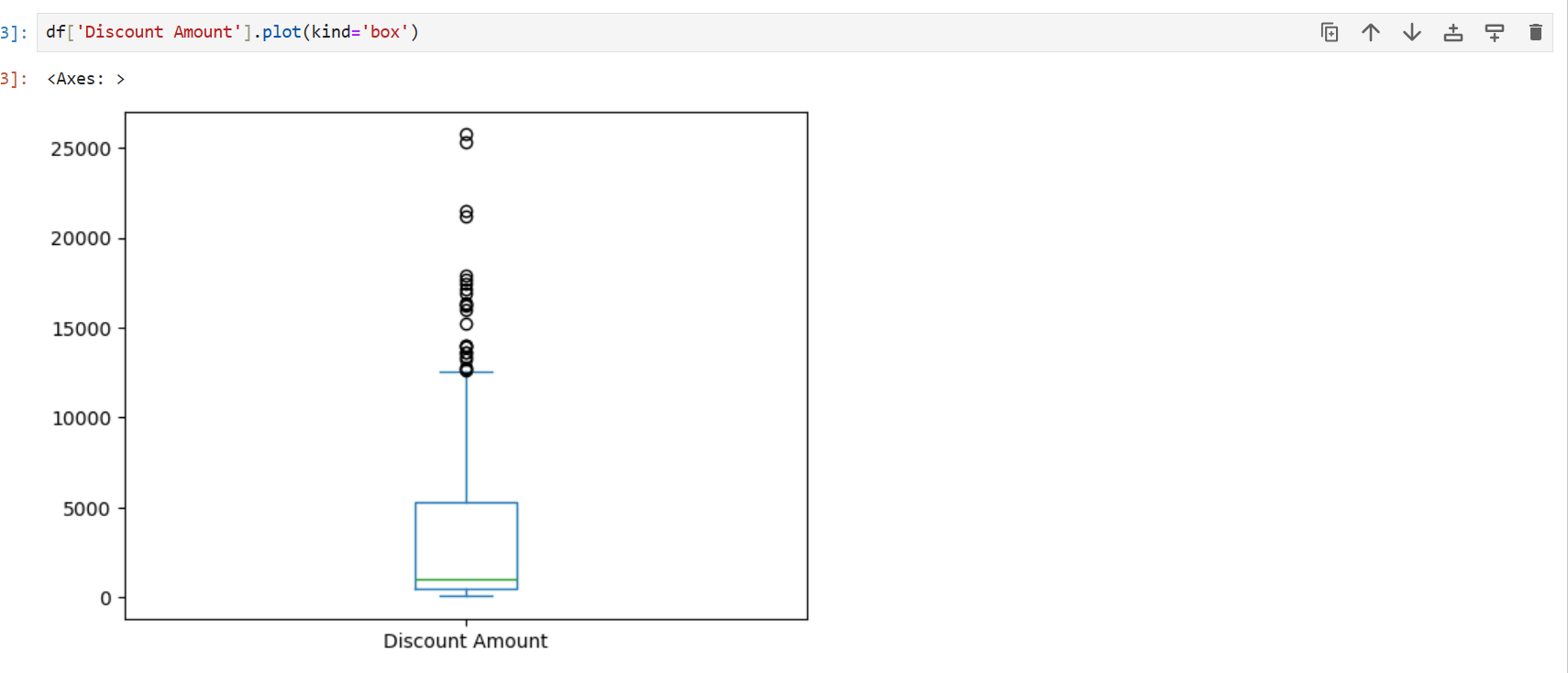
* Box plot of Avg Price



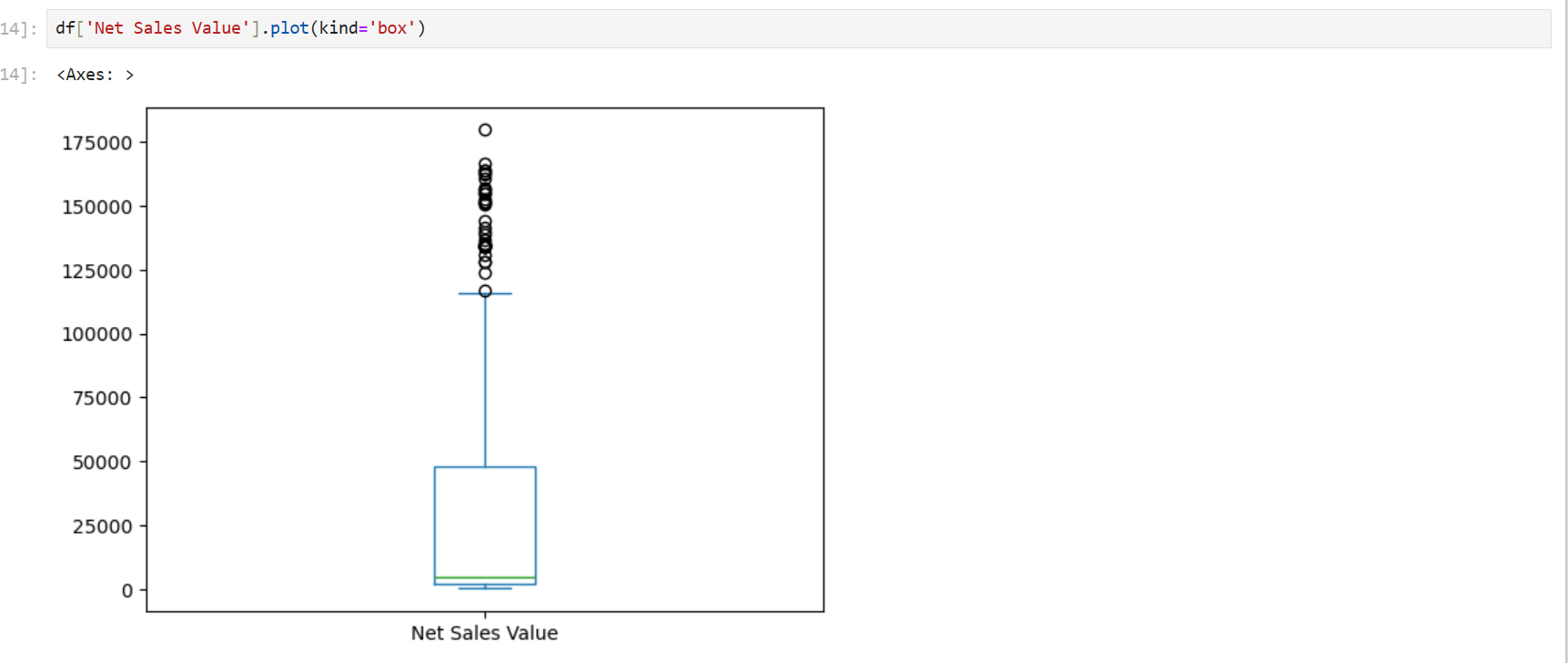
* Box plot of Total sale Value



* Box plot of Discount Amount

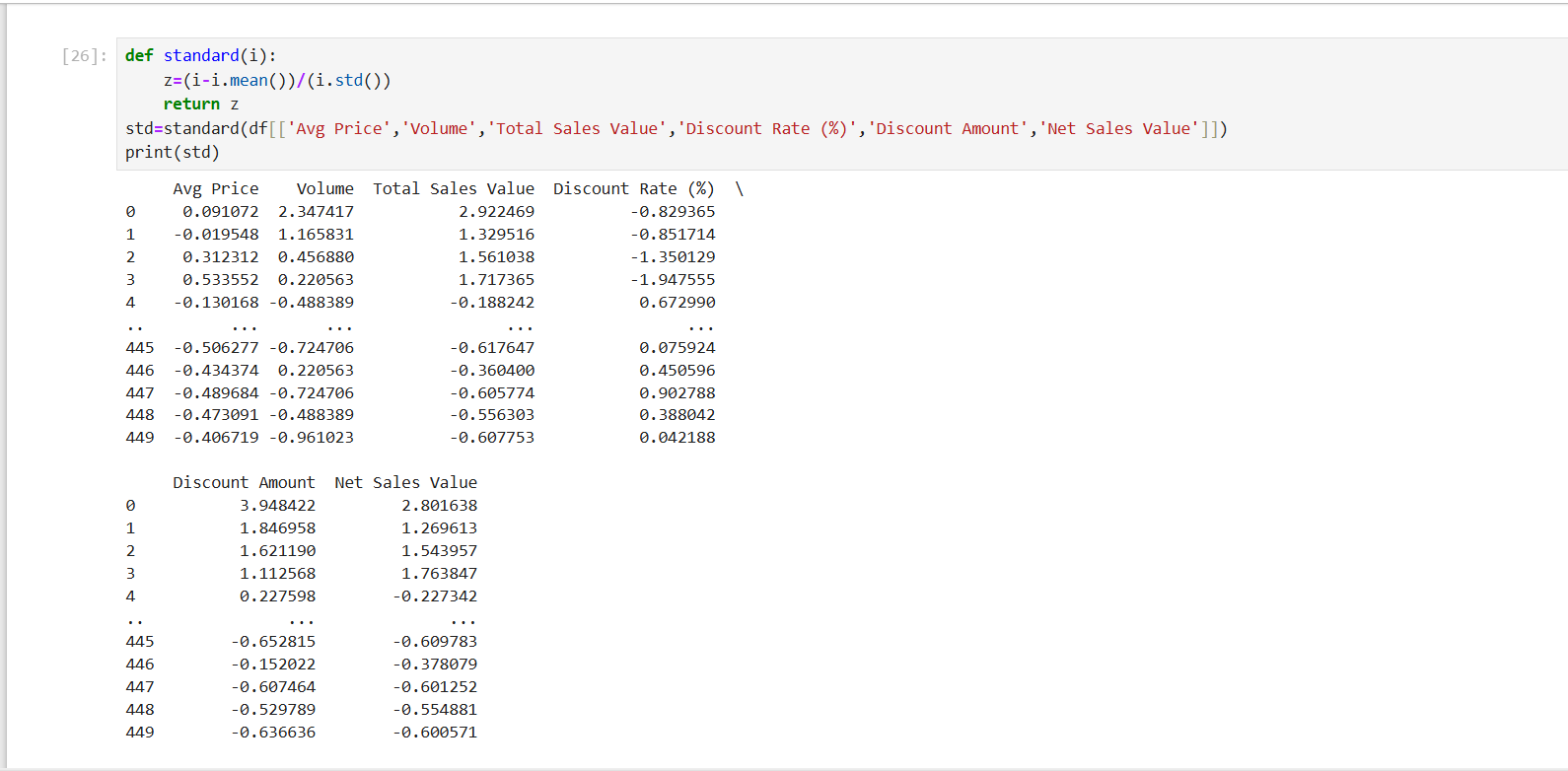


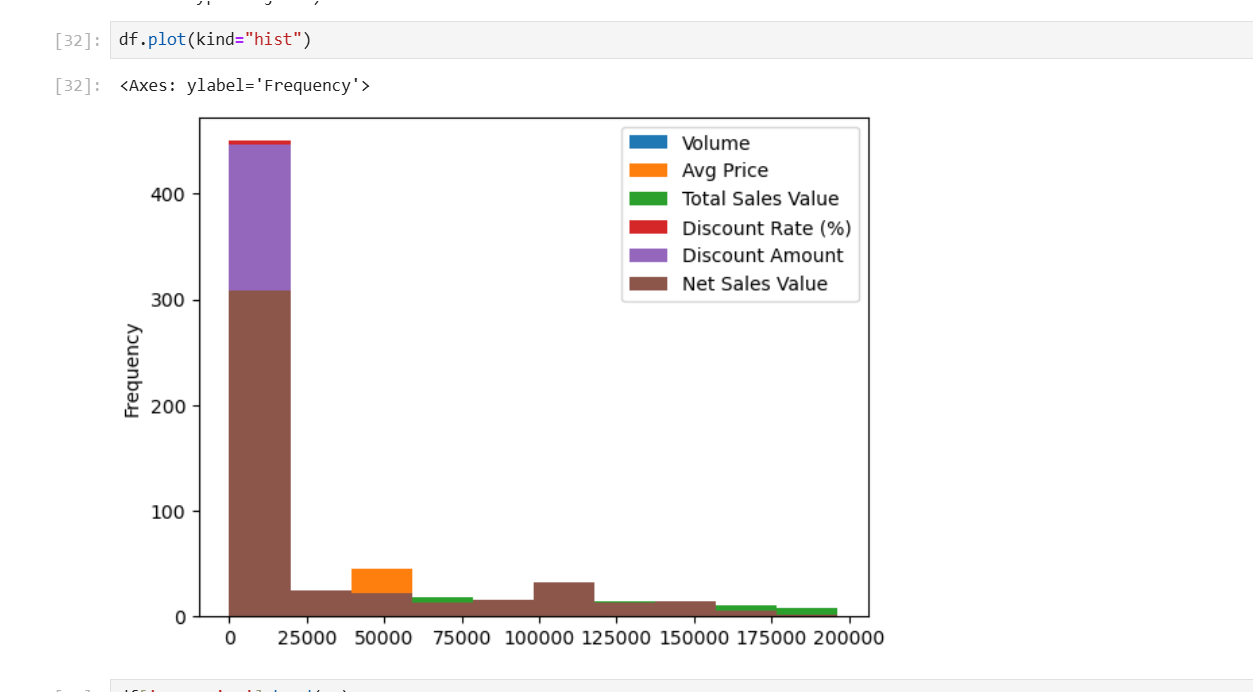
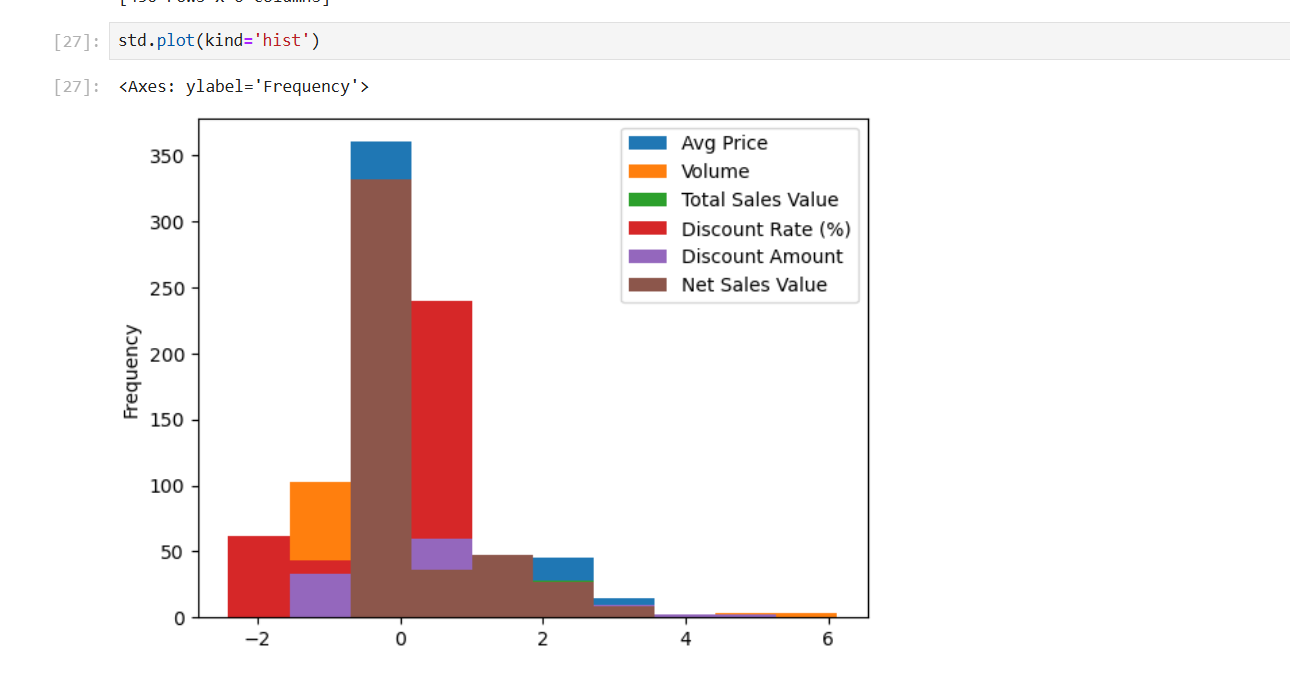
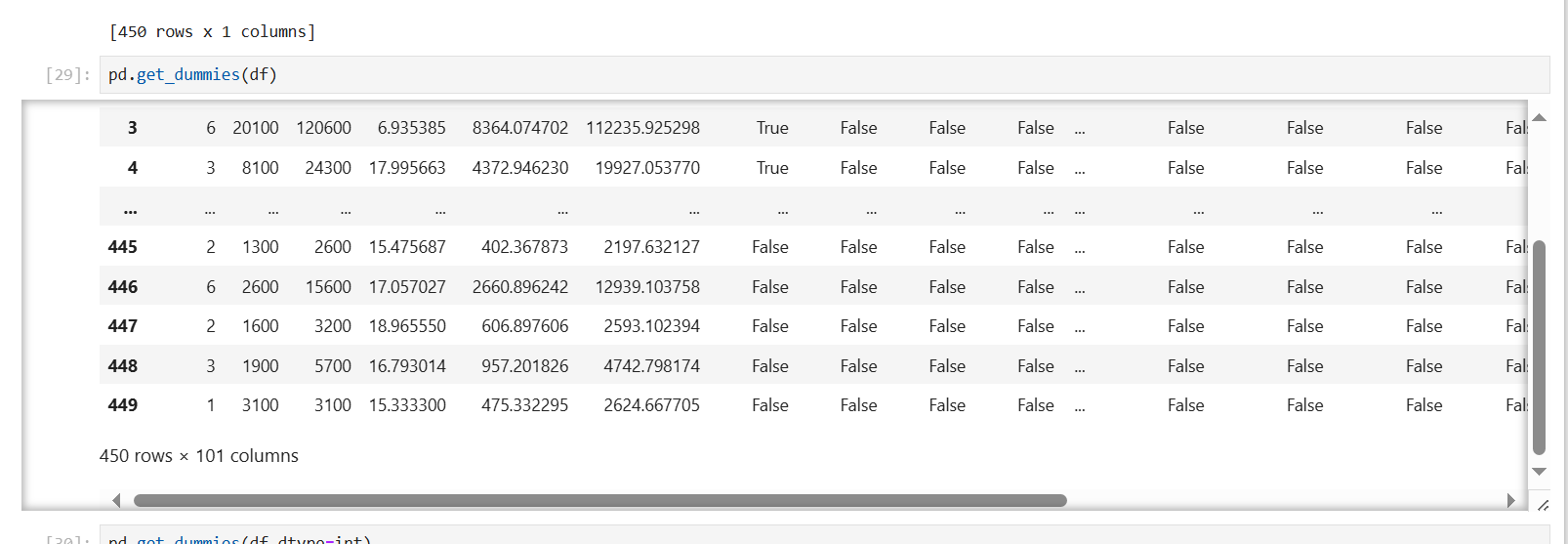
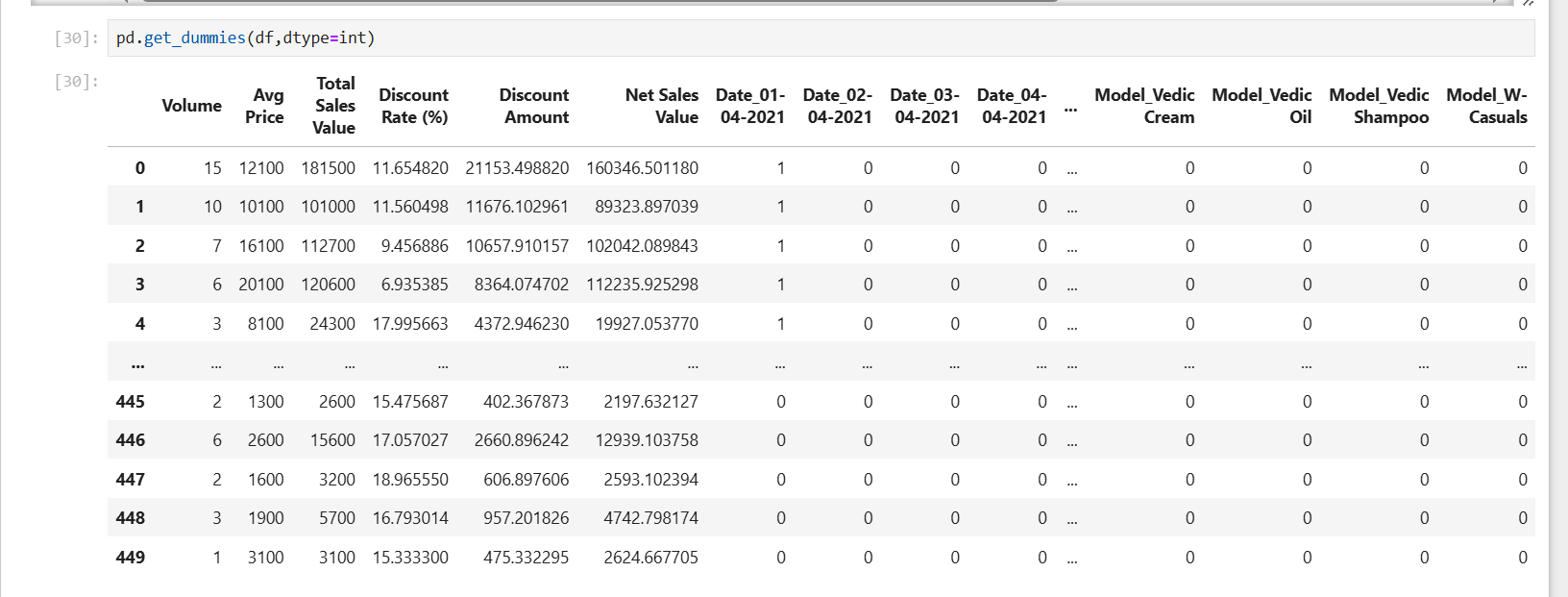
* Box plot of Net sale value



* Standardization

Standardization is a fundamental technique in statistics and data analysis that transforms data to a common scale, making it easier to interpret and compare different datasets or variables.



* Before Standardization of data (in Histogram)
* 
* After Standardization of data (in Histogram)
* 
* Dummy Variables:
* One-hot encoding can increase the dimensionality of the dataset, especially if there are many categories within a variable. This expansion can potentially lead to increased computational requirements and the curse of dimensionality in some cases.
* 
* One-hot encoding
* 
* Conclusion
* Data visualization is more useful to identify the skewness of data and visualize the data flowing. Easily understanding what is the data and outliers.
* Standardization brings all features to a comparable scale, preventing some features from dominating solely due to their larger scale. After easily predict the probability of the numerical data.