Basic Statistic\_Level 1

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

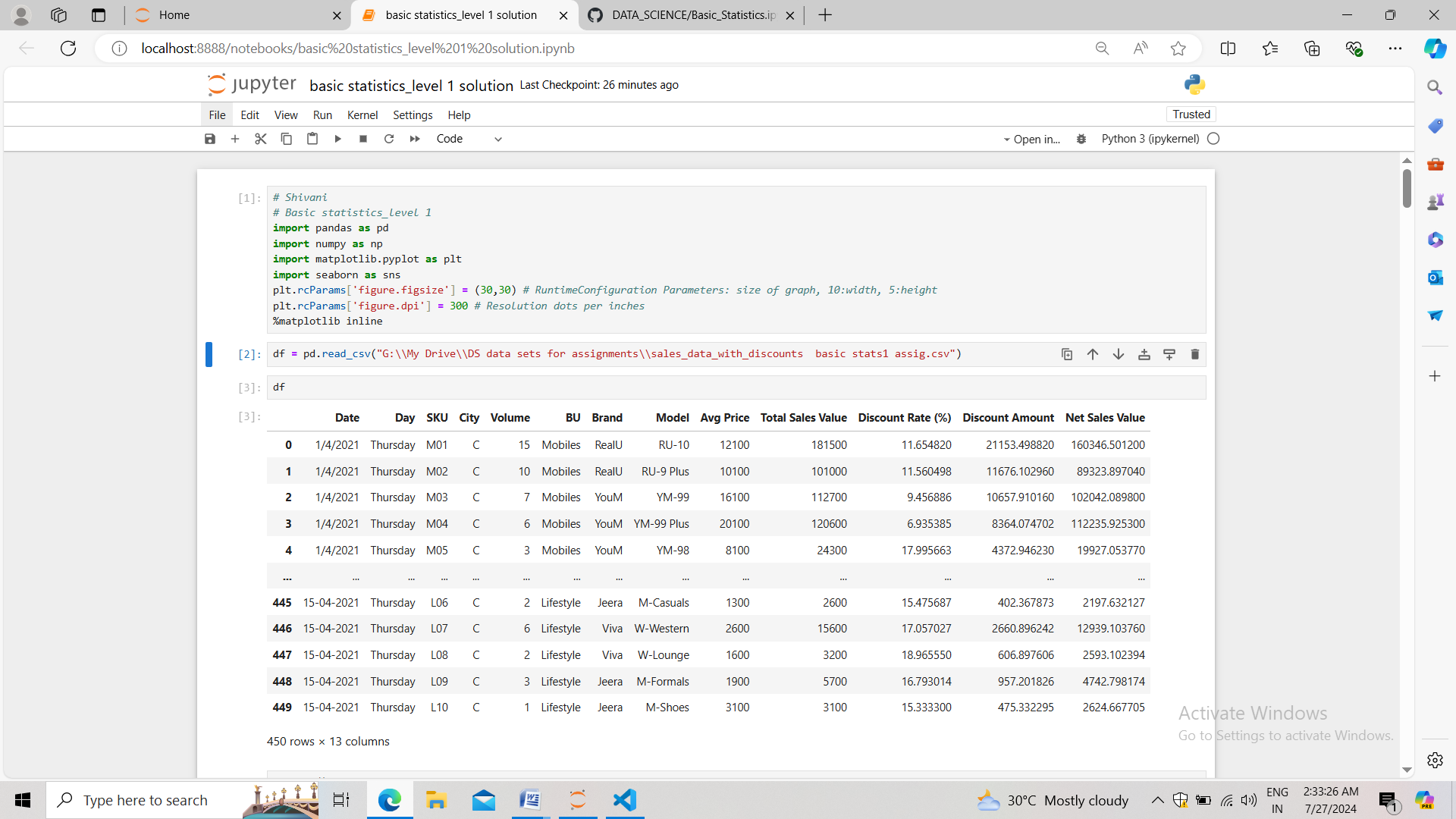
#### Introduction

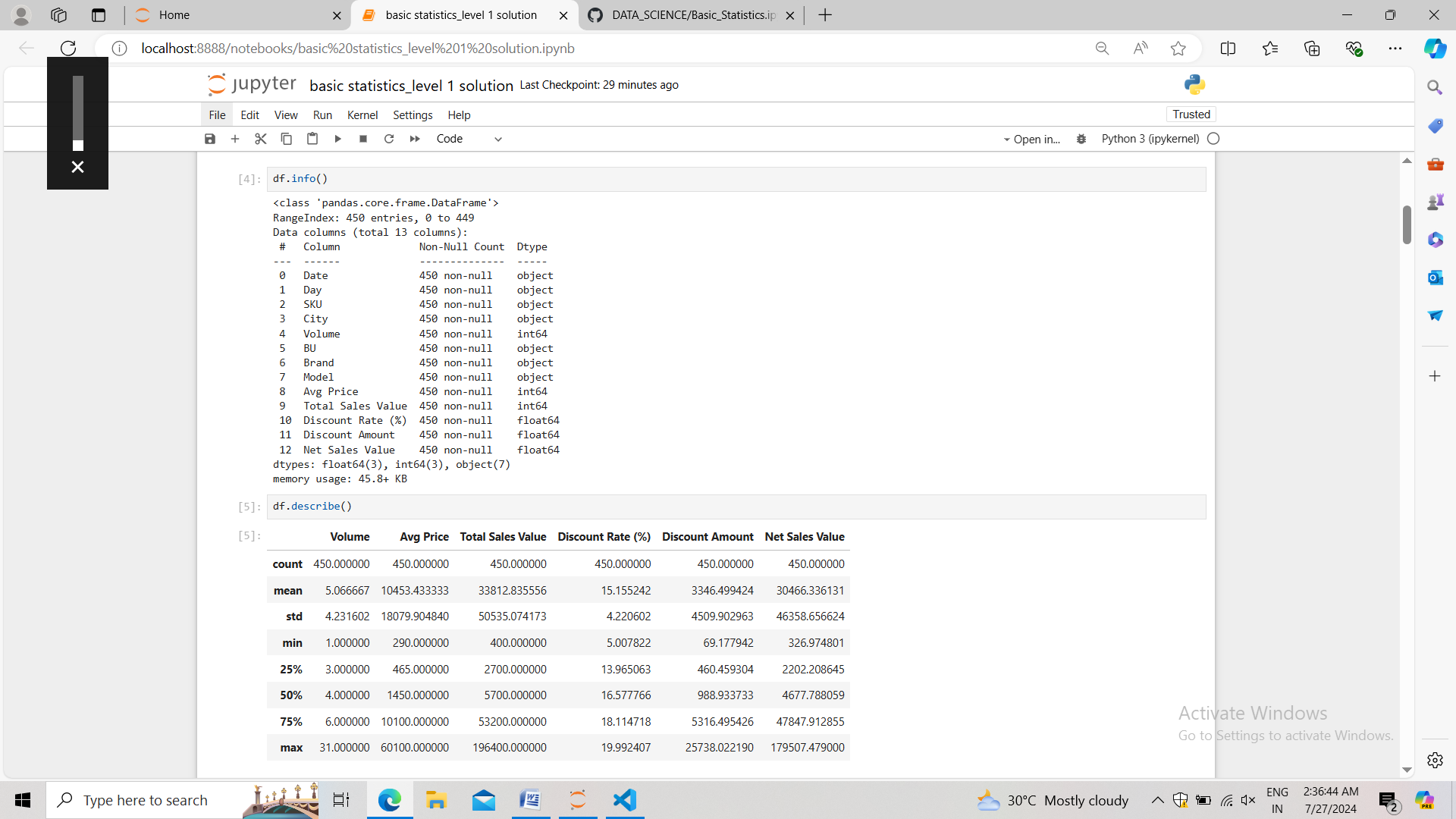
* To perform descriptive analytics, visualize data distributions, and preprocess the dataset for further analysis.

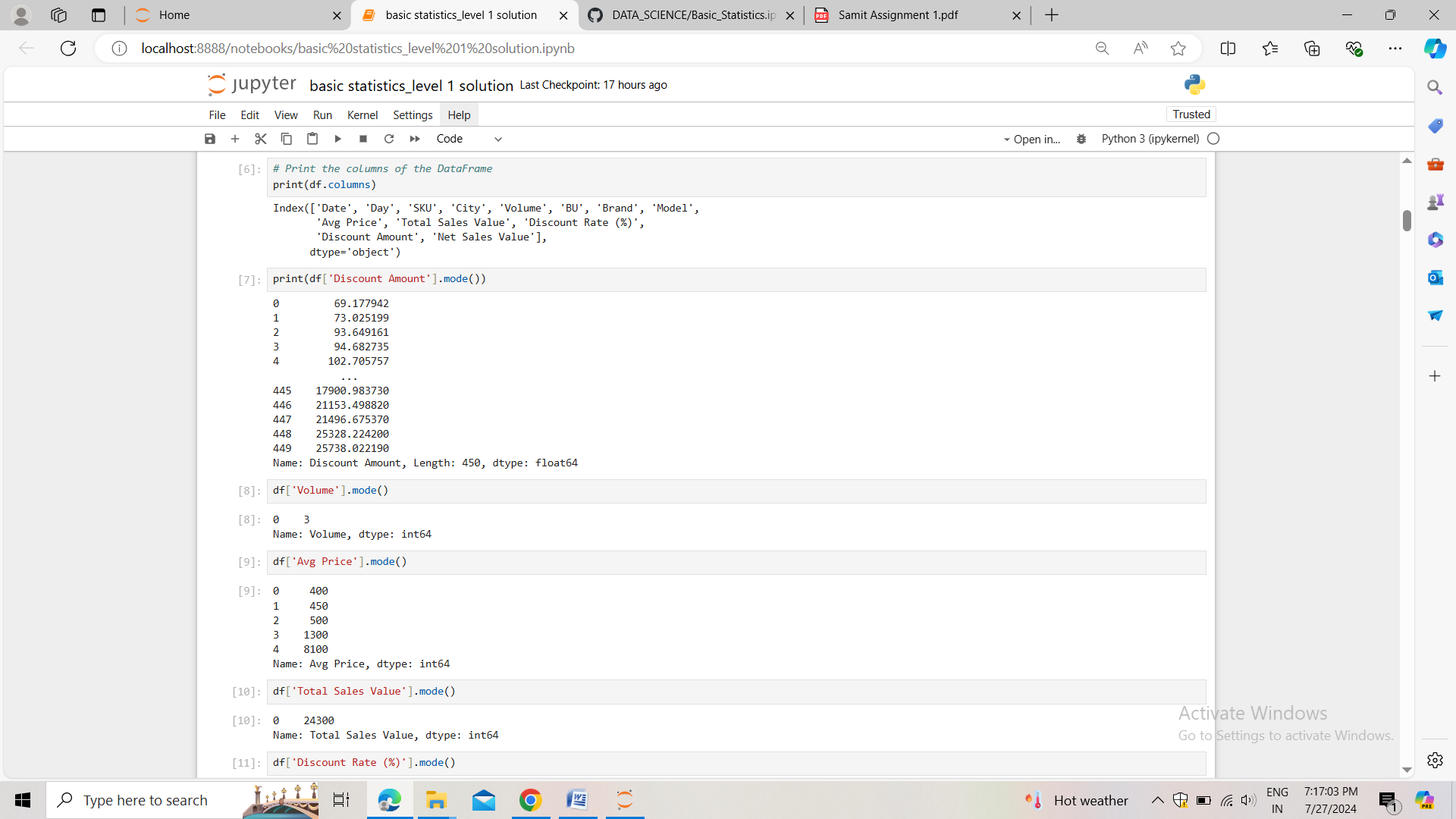
#### **Descriptive Analytics for Numerical Columns**

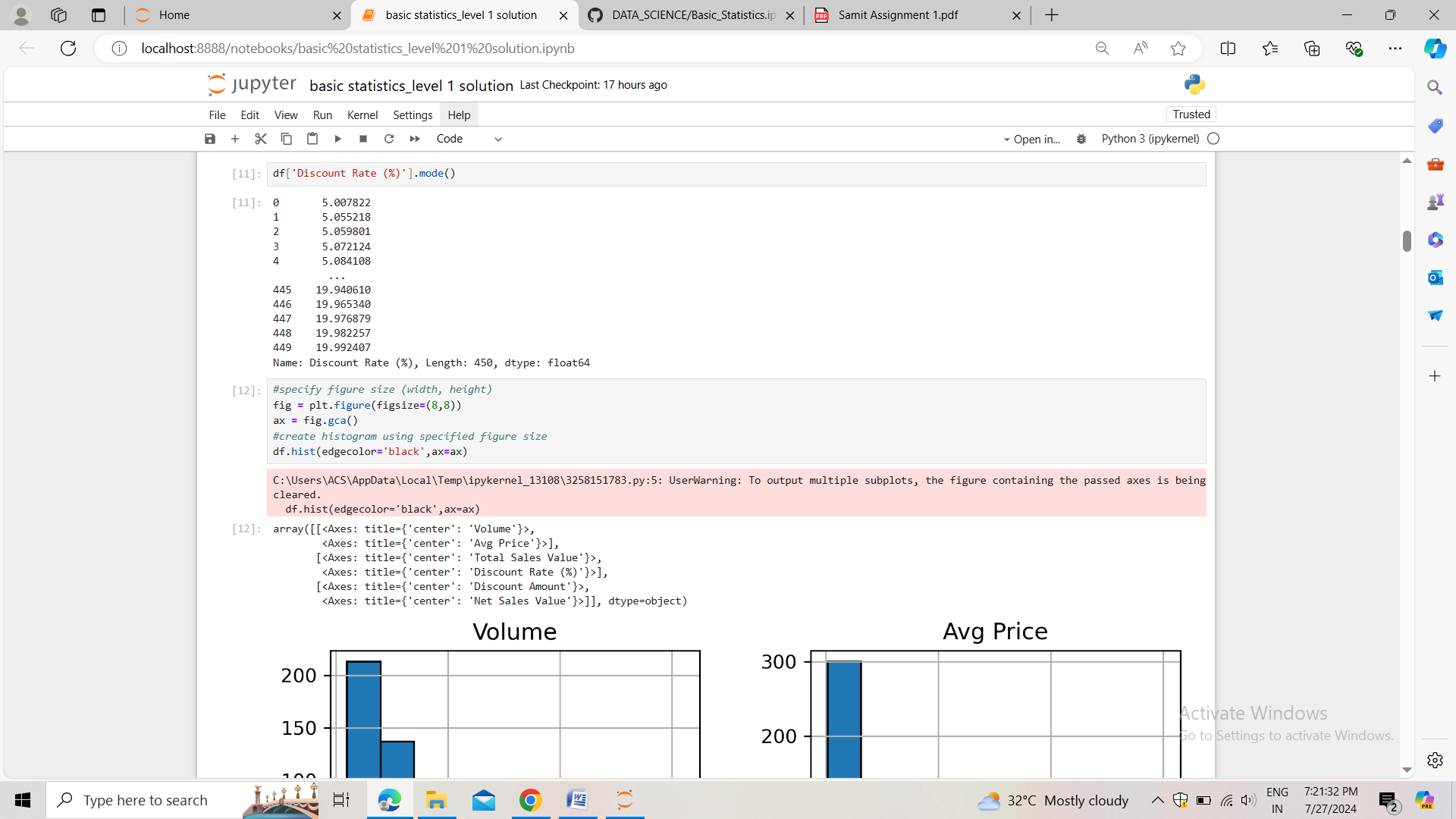
* Objective: To compute and analyze basic statistical measures for numerical columns in the dataset.
* Steps:
  1. Load the dataset into a data analysis tool or programming environment (e.g., Python with pandas library).
  2. Identify numerical columns in the dataset.
  3. Calculate the mean, median,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

**Answer:-**



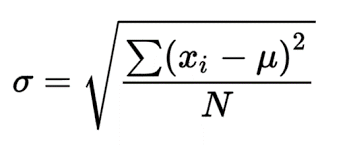




**mean** = (sum of all data points)/(total no. Of data) 

**mode** = more frequent item in the data. It can have more than one mode value.

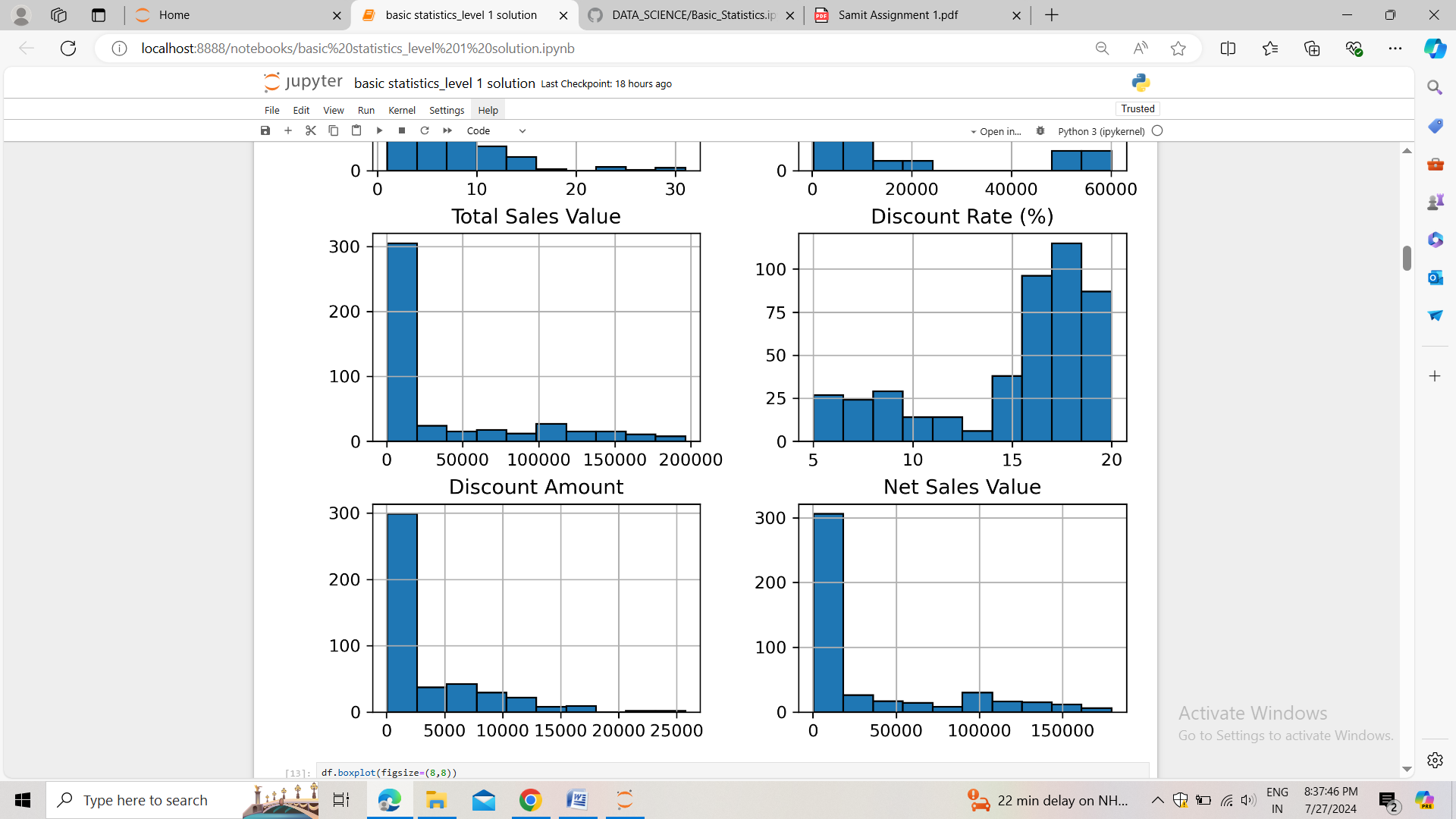
**Median** = (n+1)/2 :-for odd no. Of data items ((n/2)+(n/2+1))/2 :- for even no

**Standard deviation =**

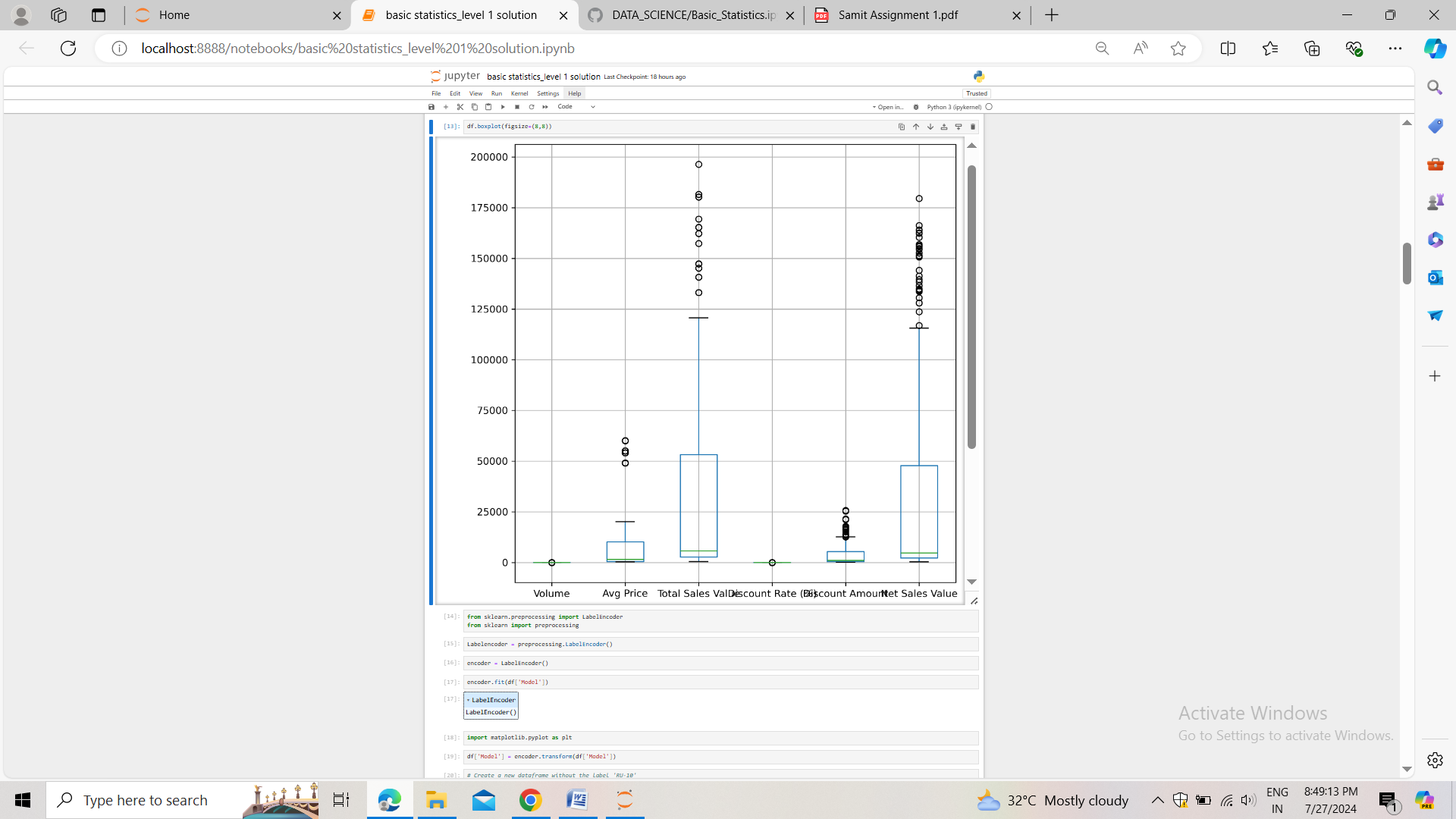
#### **Data Visualization**

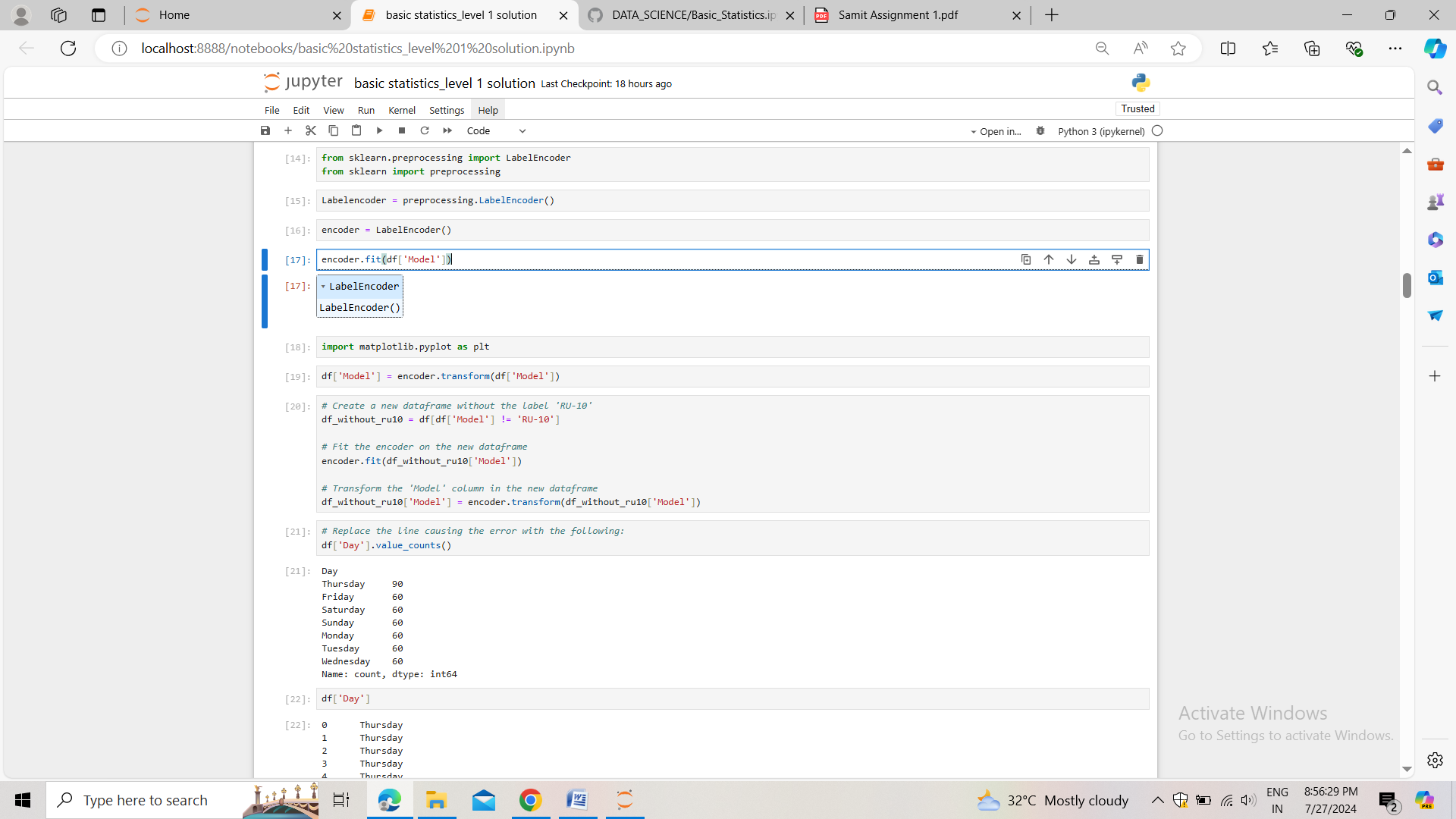
* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:
  1. Plot histograms for each numerical column.
  2. Analyze the distribution (e.g., skewness, presence of outliers) and provide inferences.
* **Boxplots**:
  1. Create boxplots for numerical variables to identify outliers and the interquartile range.
  2. Discuss any findings, such as extreme values or unusual distributions.

Answer:-



* Except the discount rate % all the numerical columns like volume, average price, total sales value, discount amount, net sales value all are positively (+ve) skewed and only the discount rate % is negatively skewed.



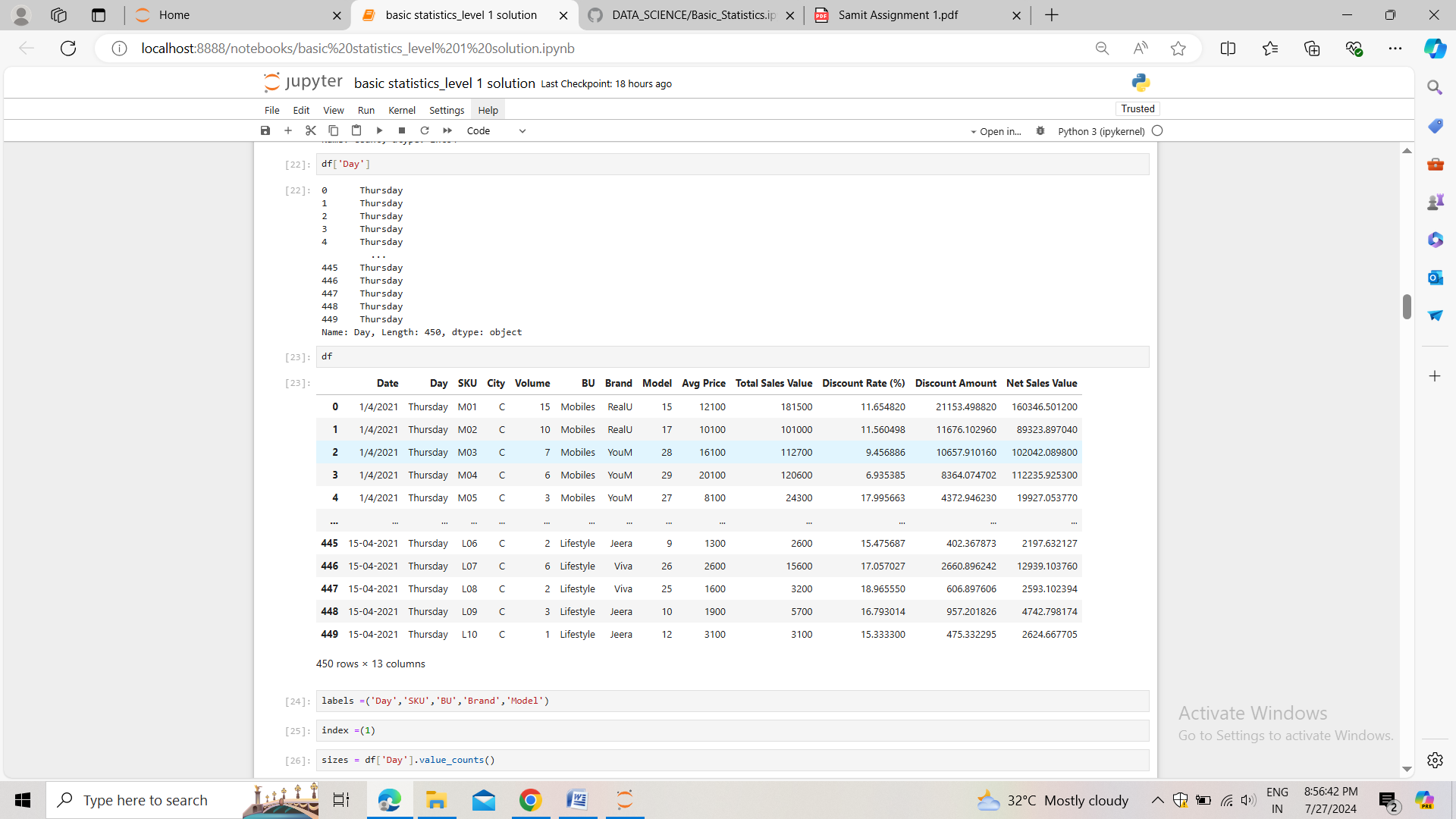
* As we can see avg price, total sales value, discount amount and the net sales value are having more outliers. And also these numerical columns are of large range.
* Volume and discount rate % are of low range.  
  

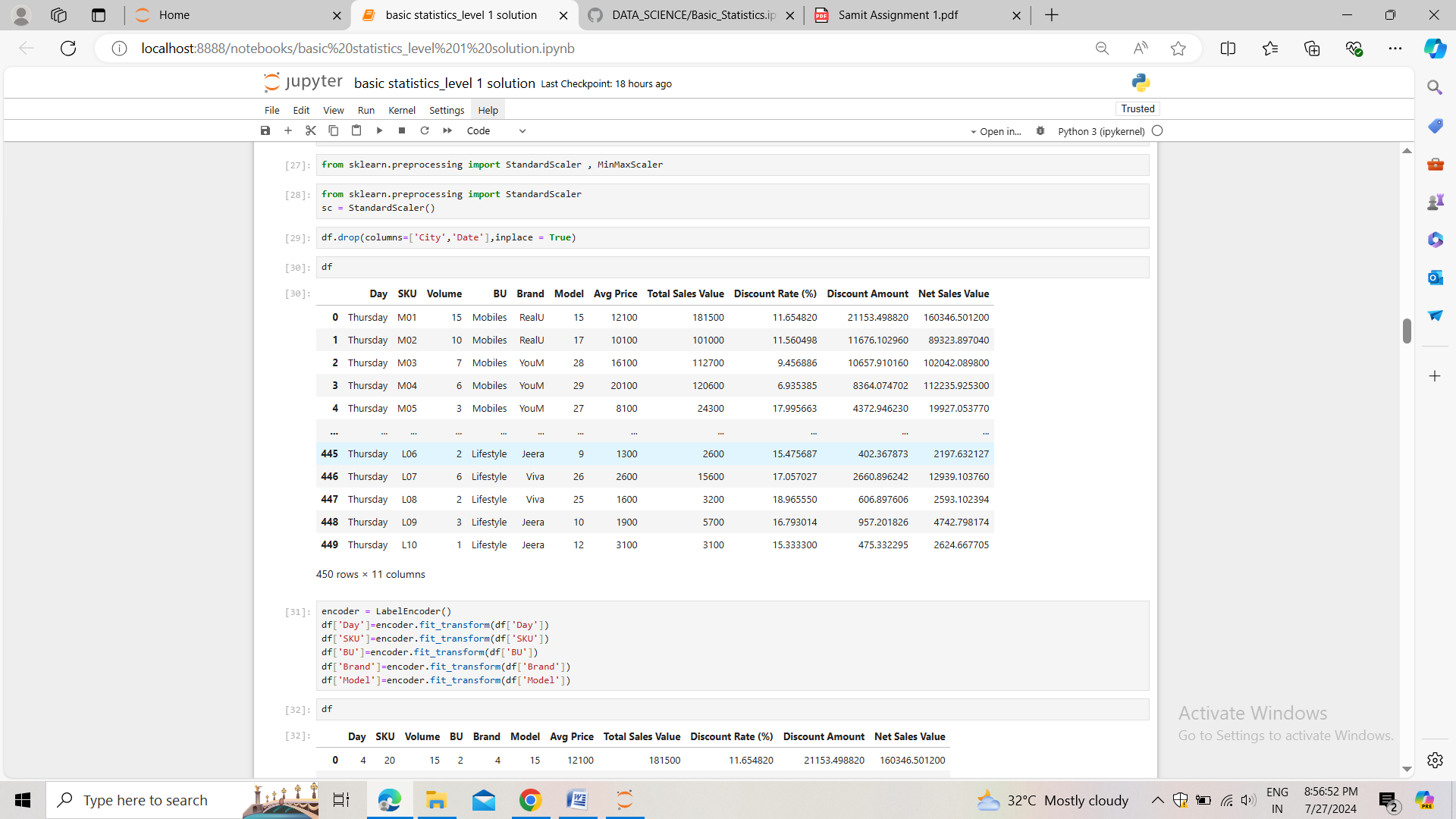
#### **Standardization of Numerical Variables**

* Objective: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Steps:
  1. Explain the concept of standardization (z-score normalization).
  2. Standardize the numerical columns using the formula: z=x-mu/sigma
  3. ​Show before and after comparisons of the data distributions.

Answer:-

**Standardization**:- In geometric intuition of standardization we do mean centering and scaling by the factor of standard deviation. the formula: z=x-mu/sigma transformed our original data to somewhere around zero. It is the process which removes the unit from the variables and put the data around -1 to +1. Or -3 to +3.

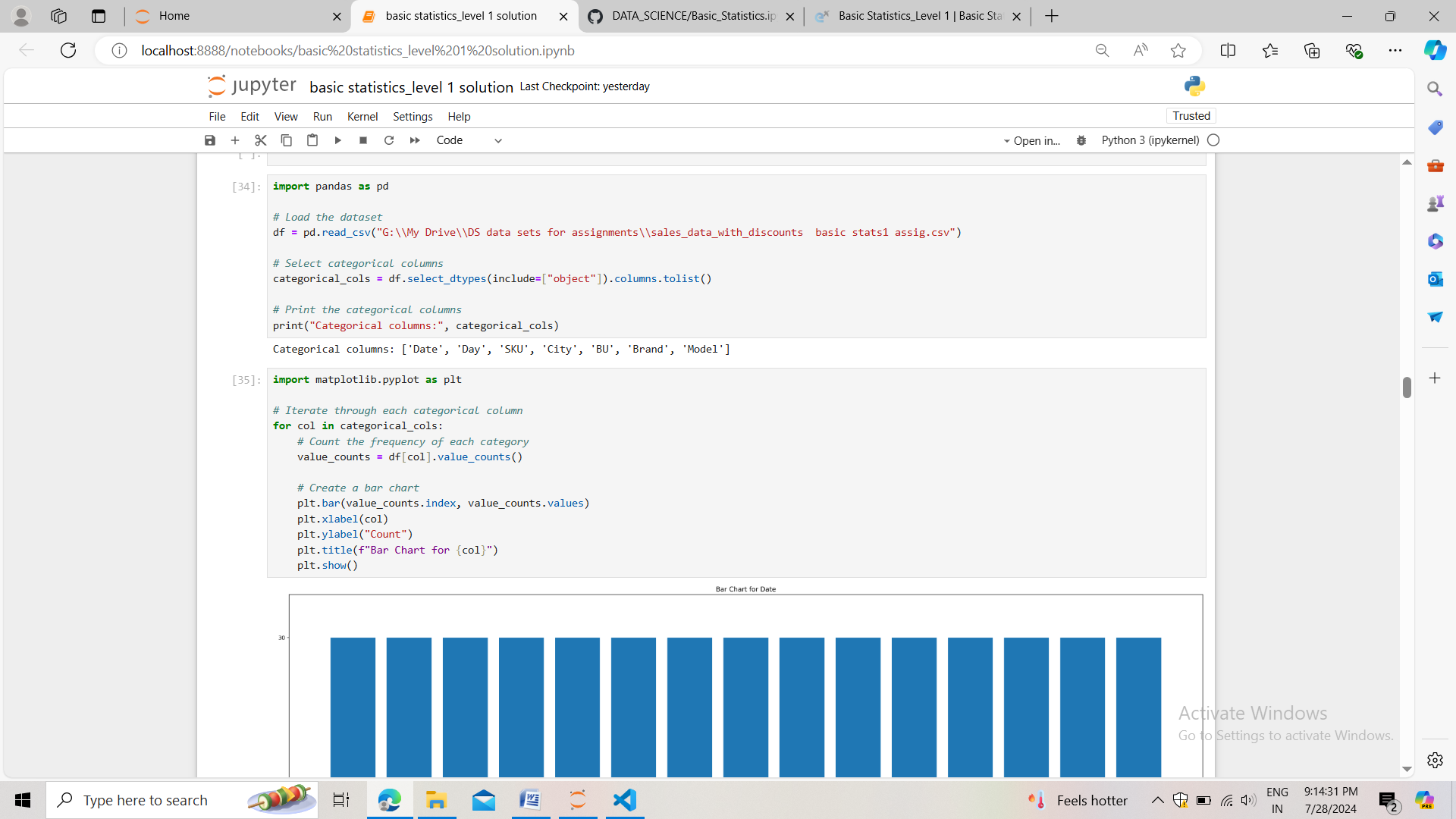




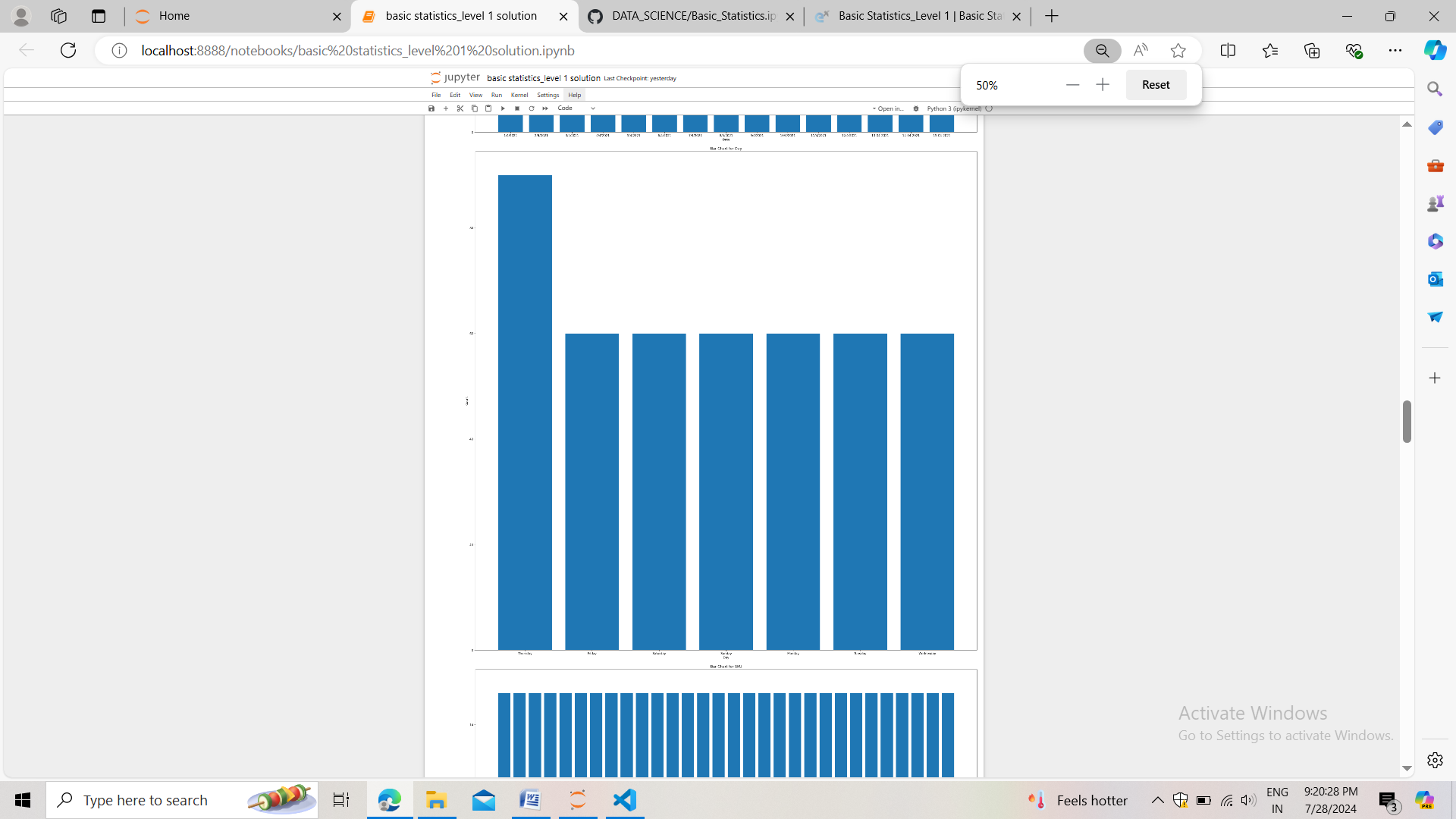


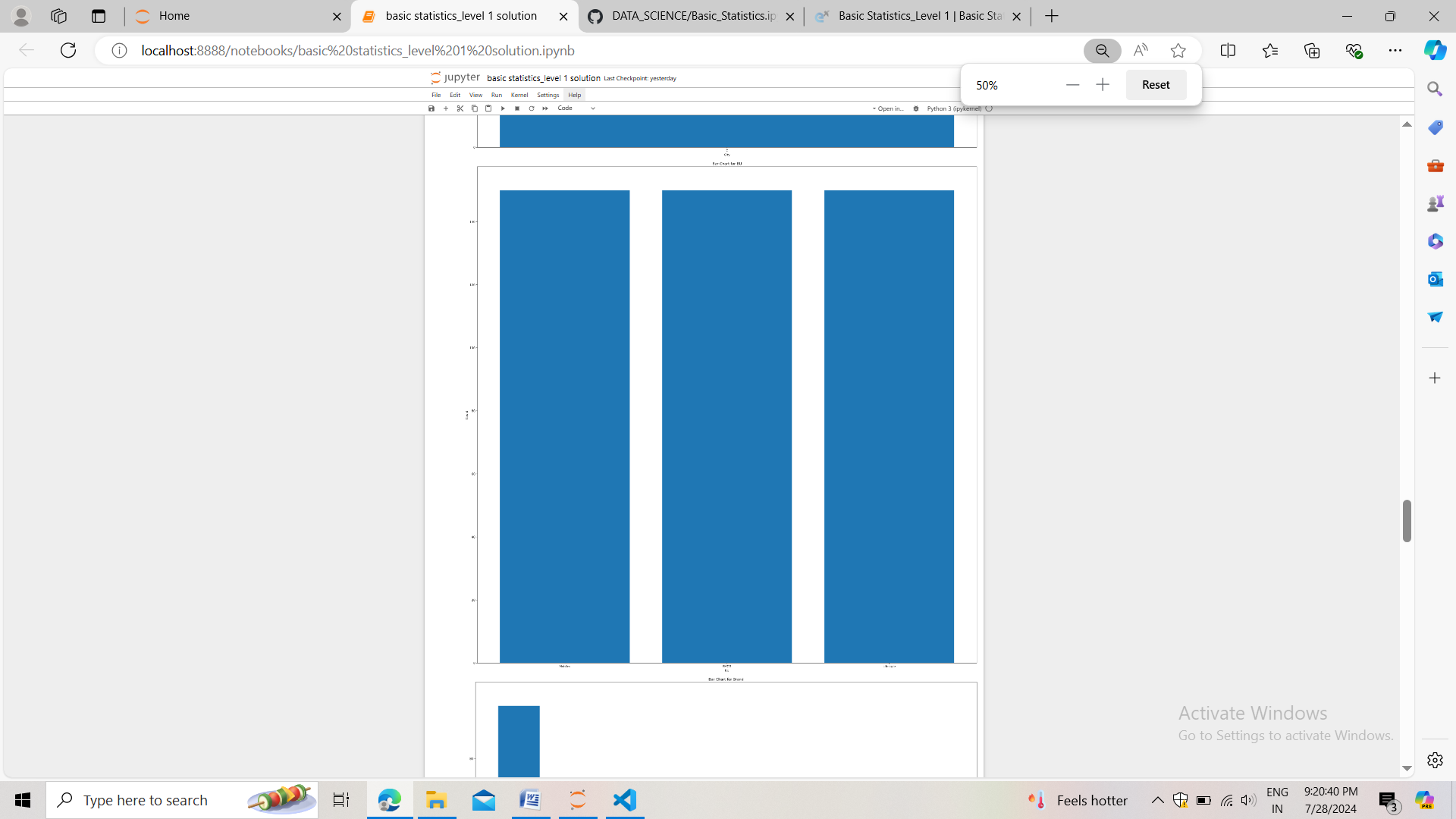
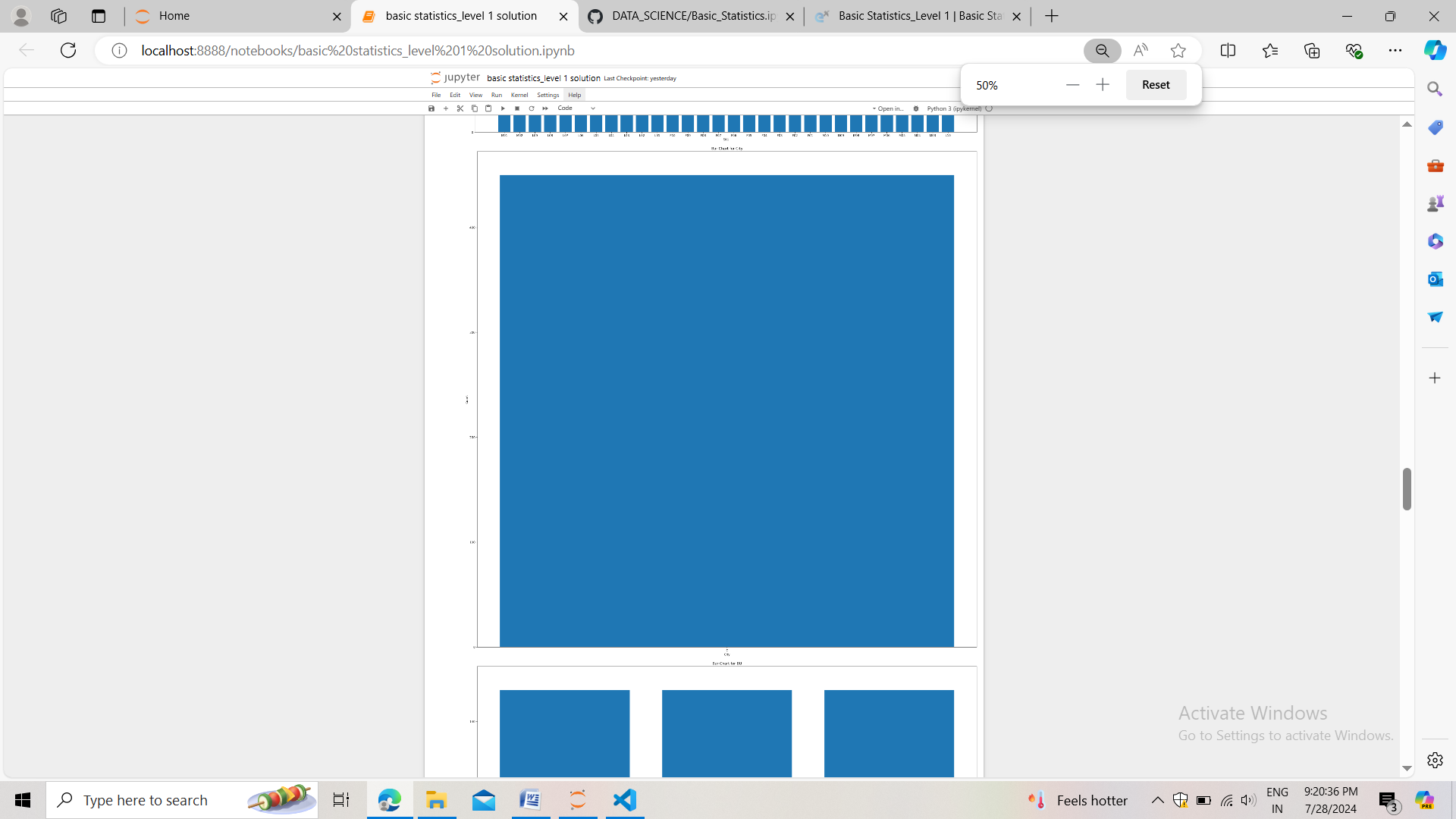
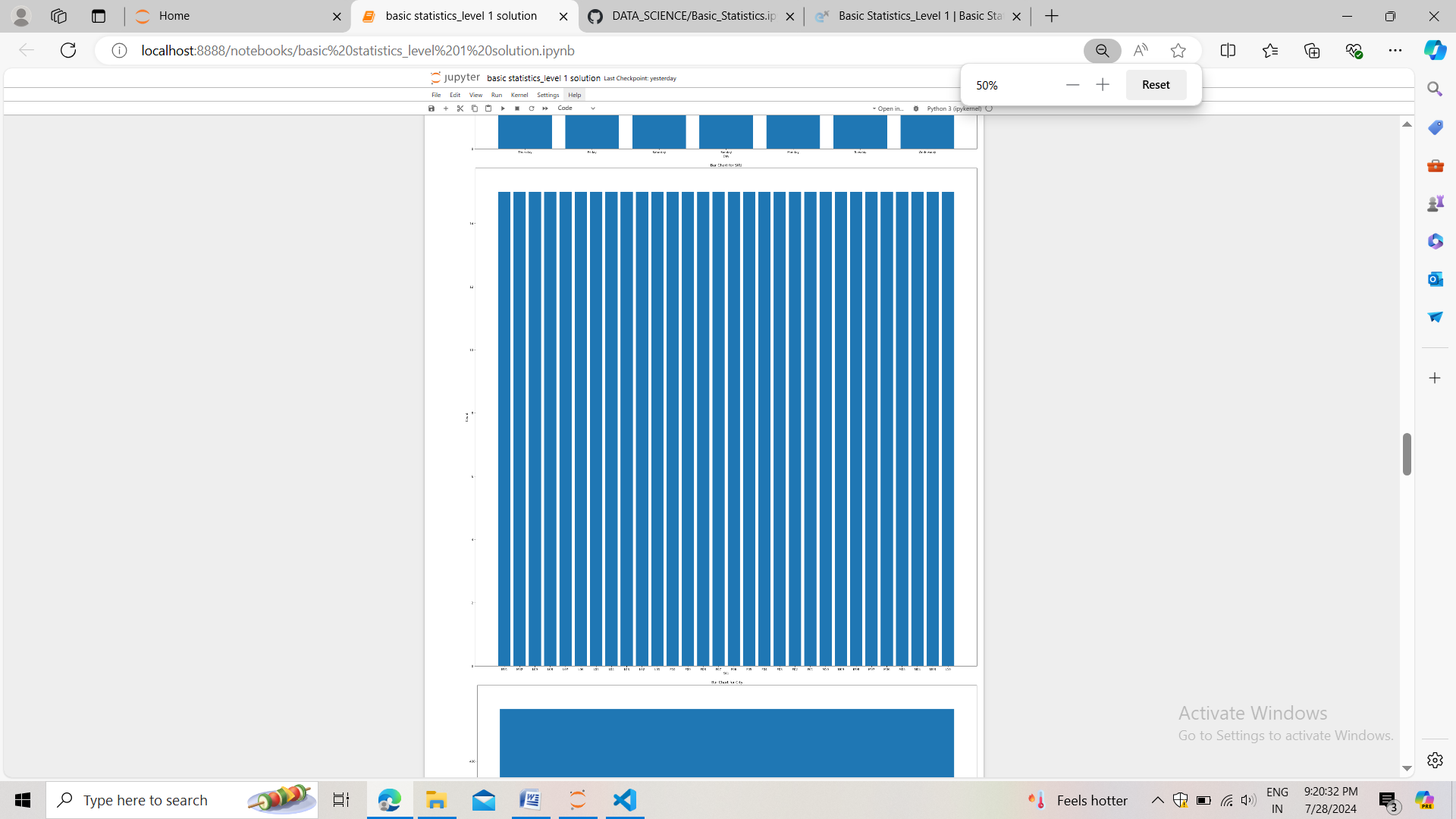
* **Bar Chart Analysis for Categorical Column:**
  1. Identify categorical columns in the dataset.
  2. Create bar charts to visualize the frequency or count of each category.
  3. Analyze the distribution of categories and provide insights.

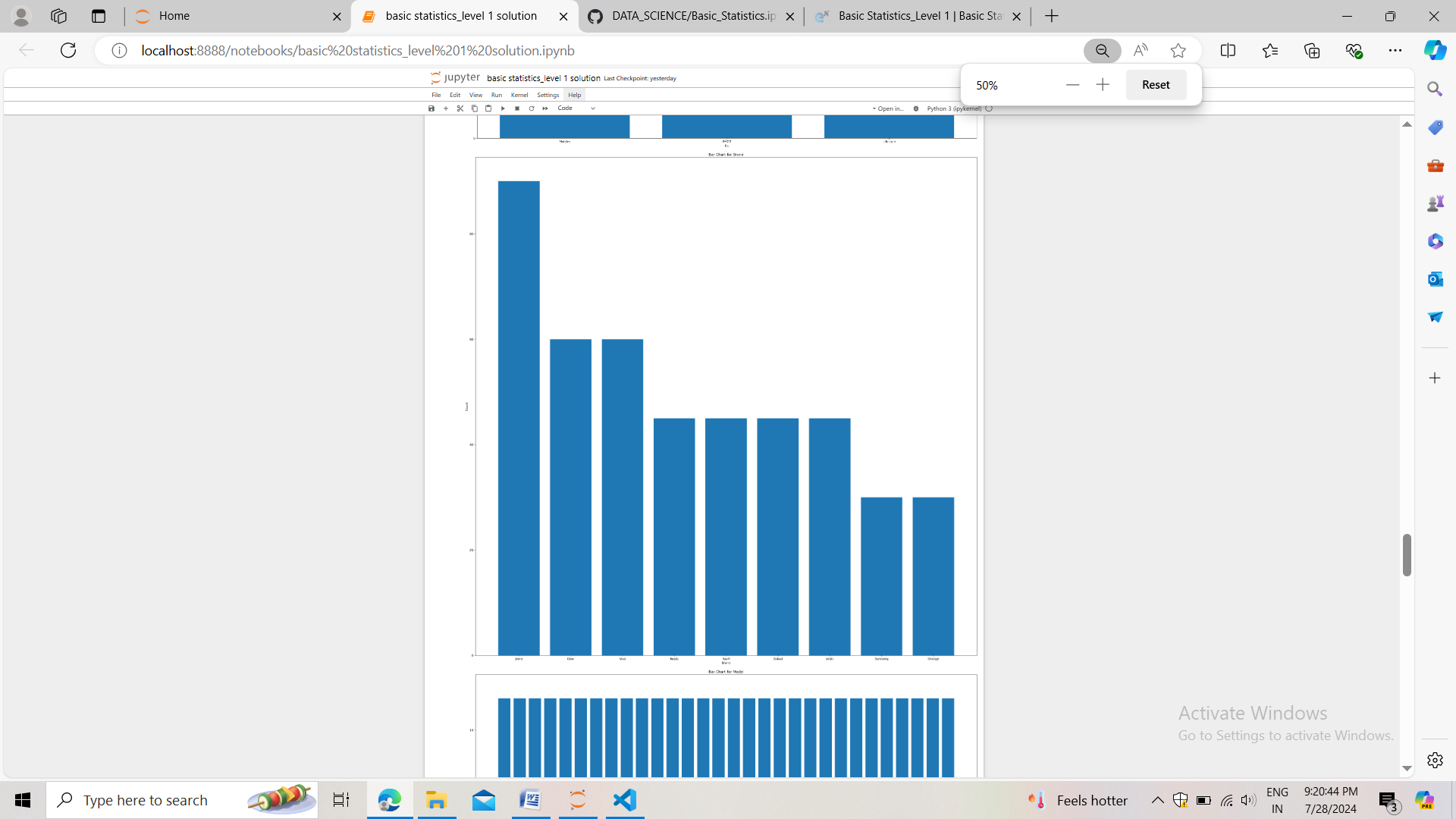
Answer:-

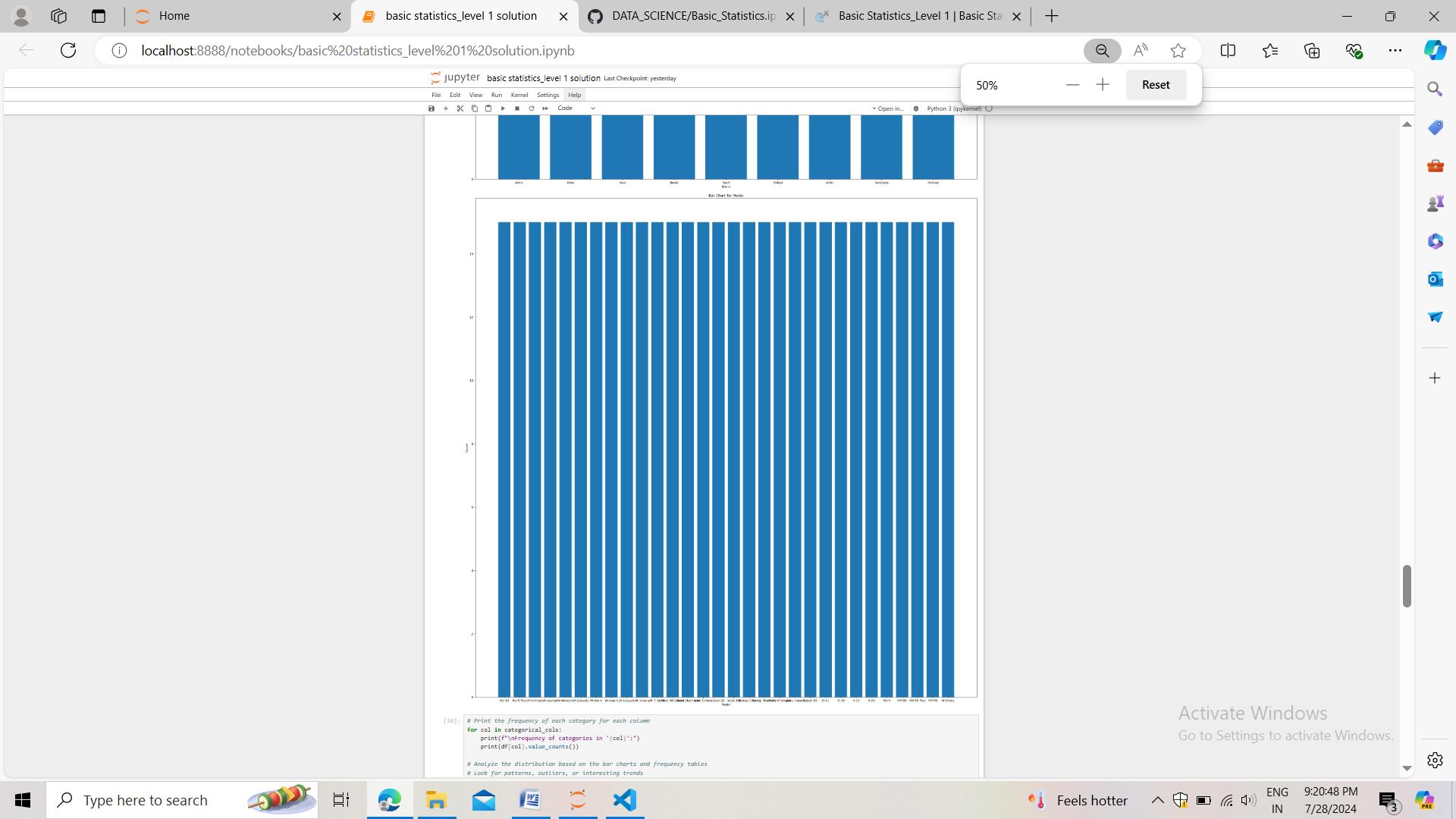


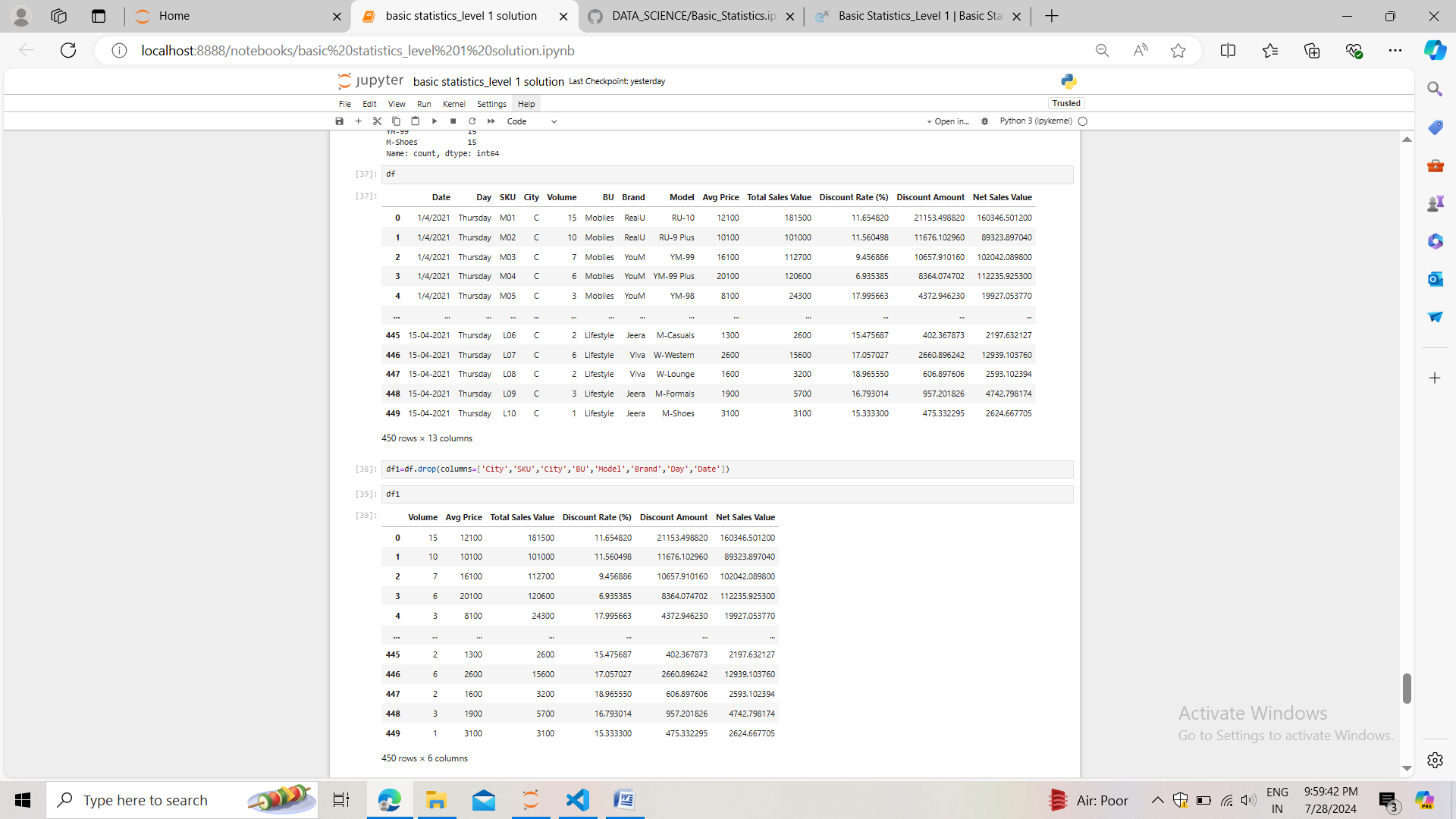
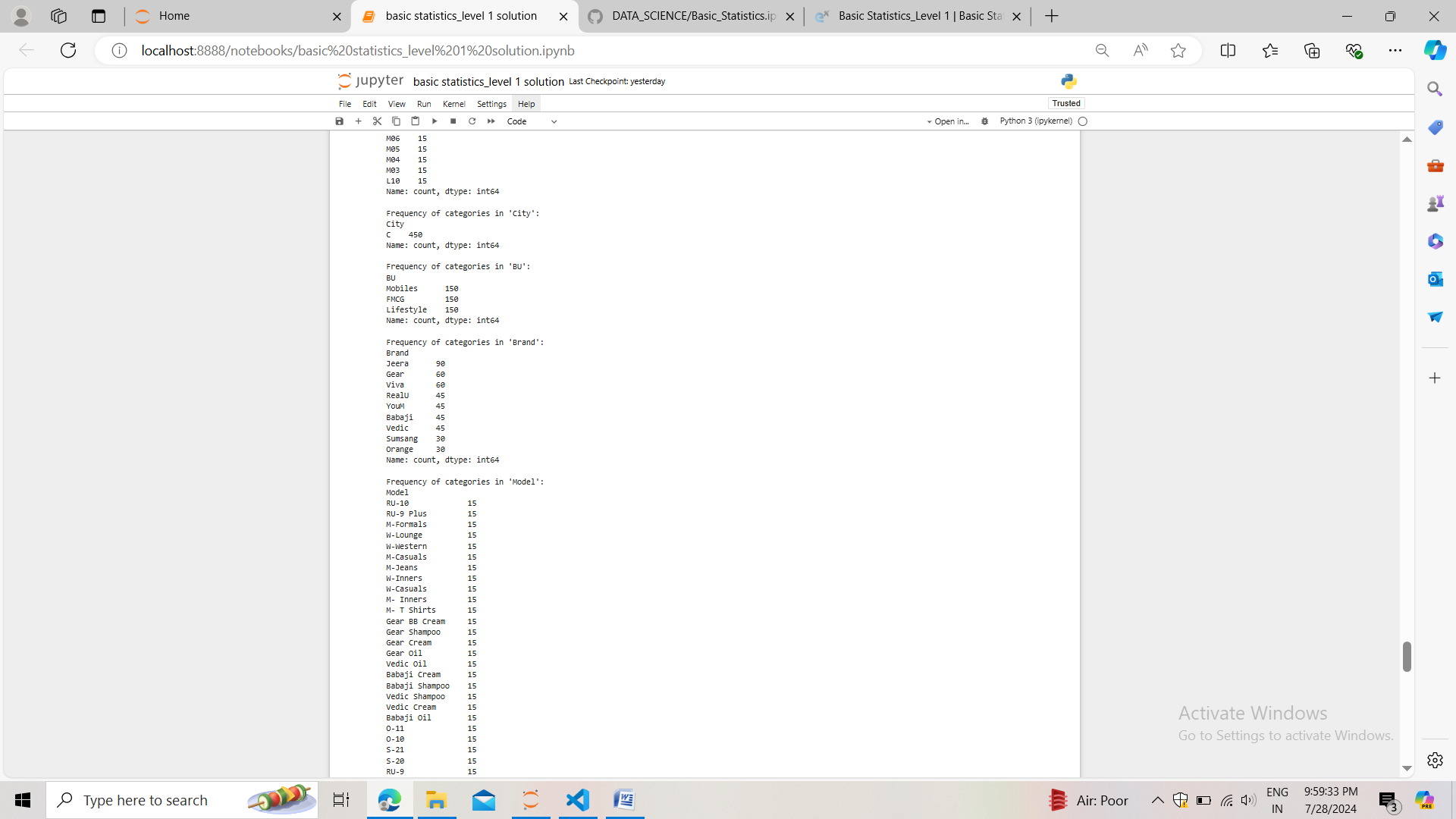
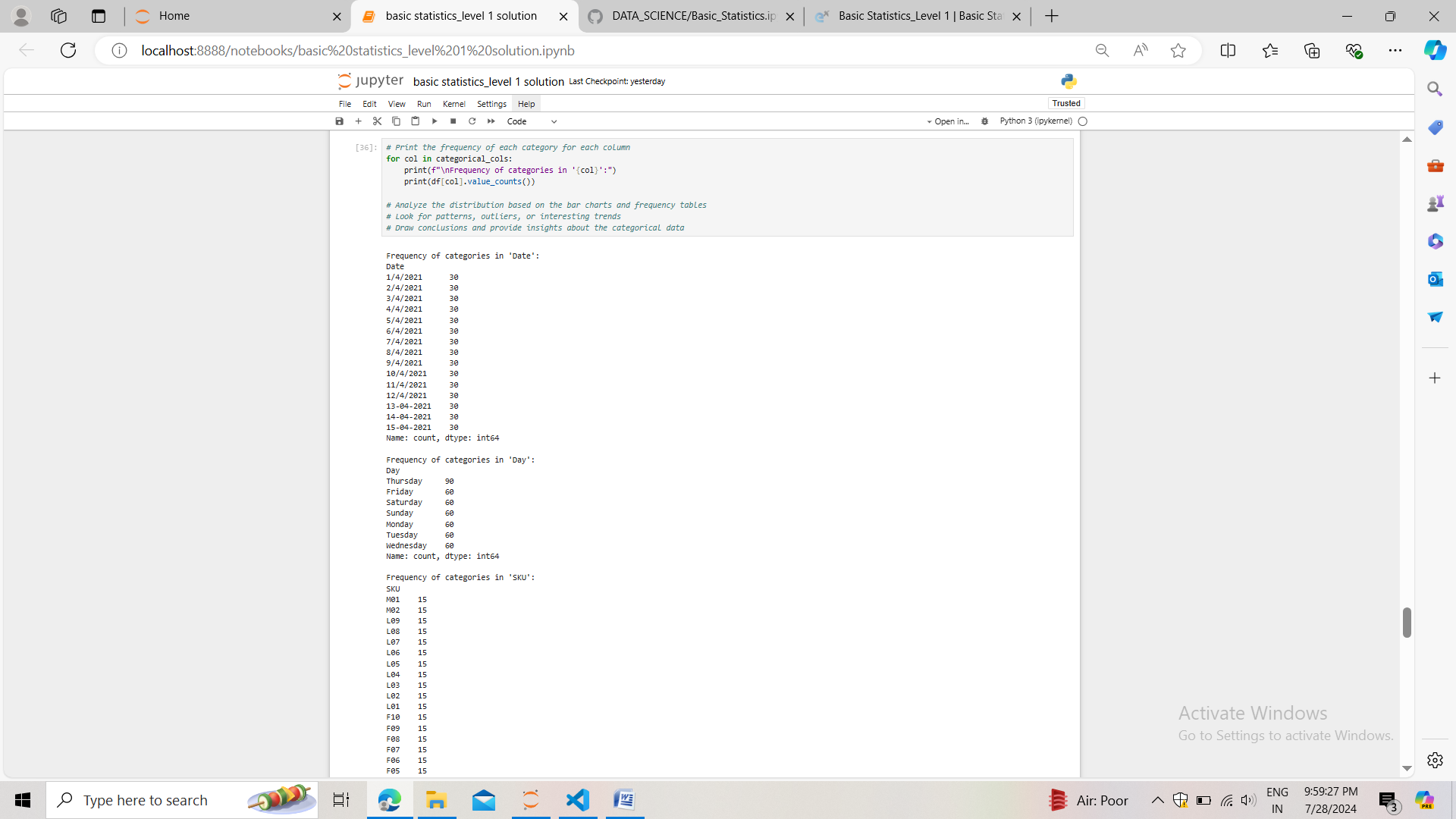










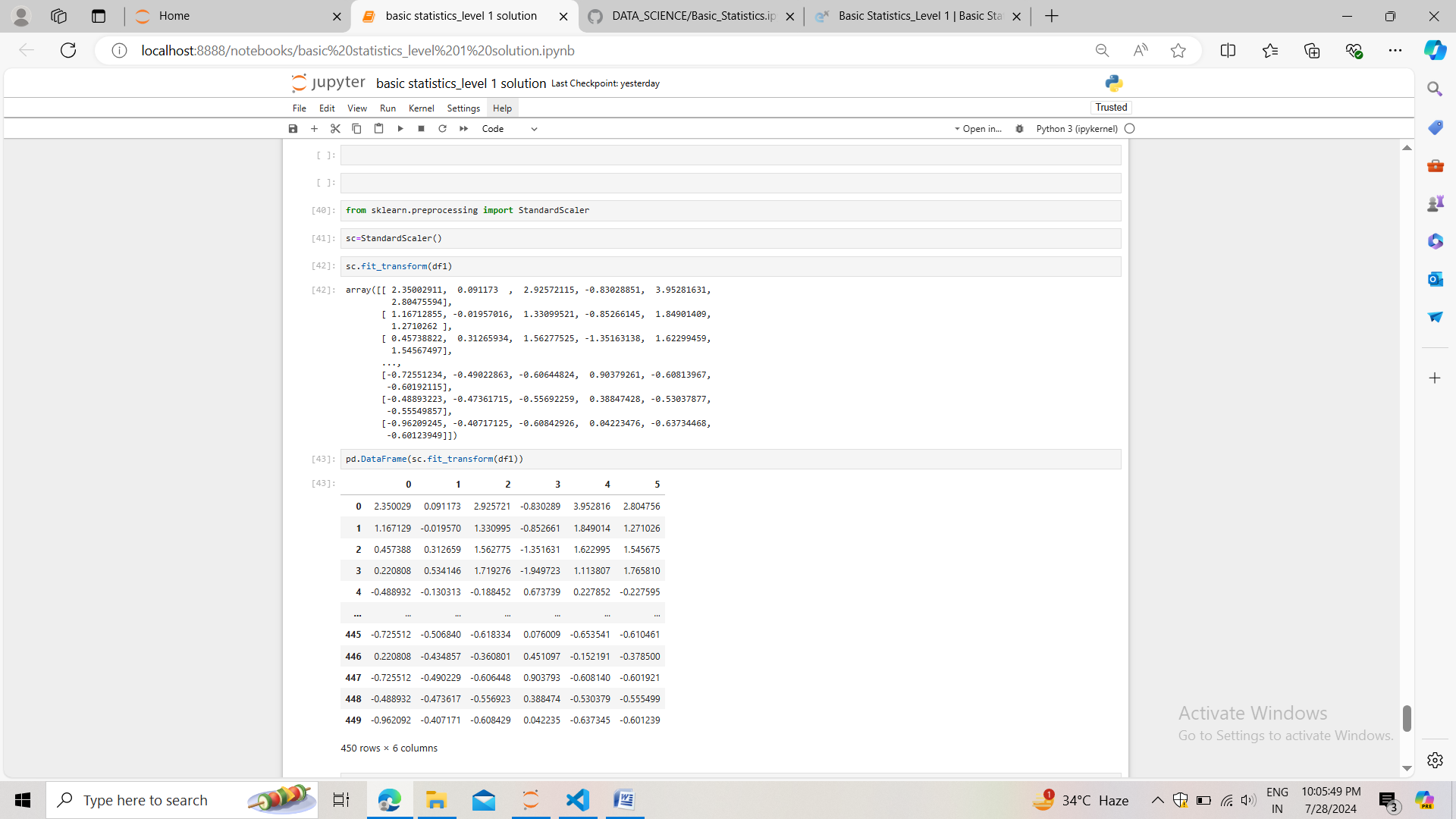


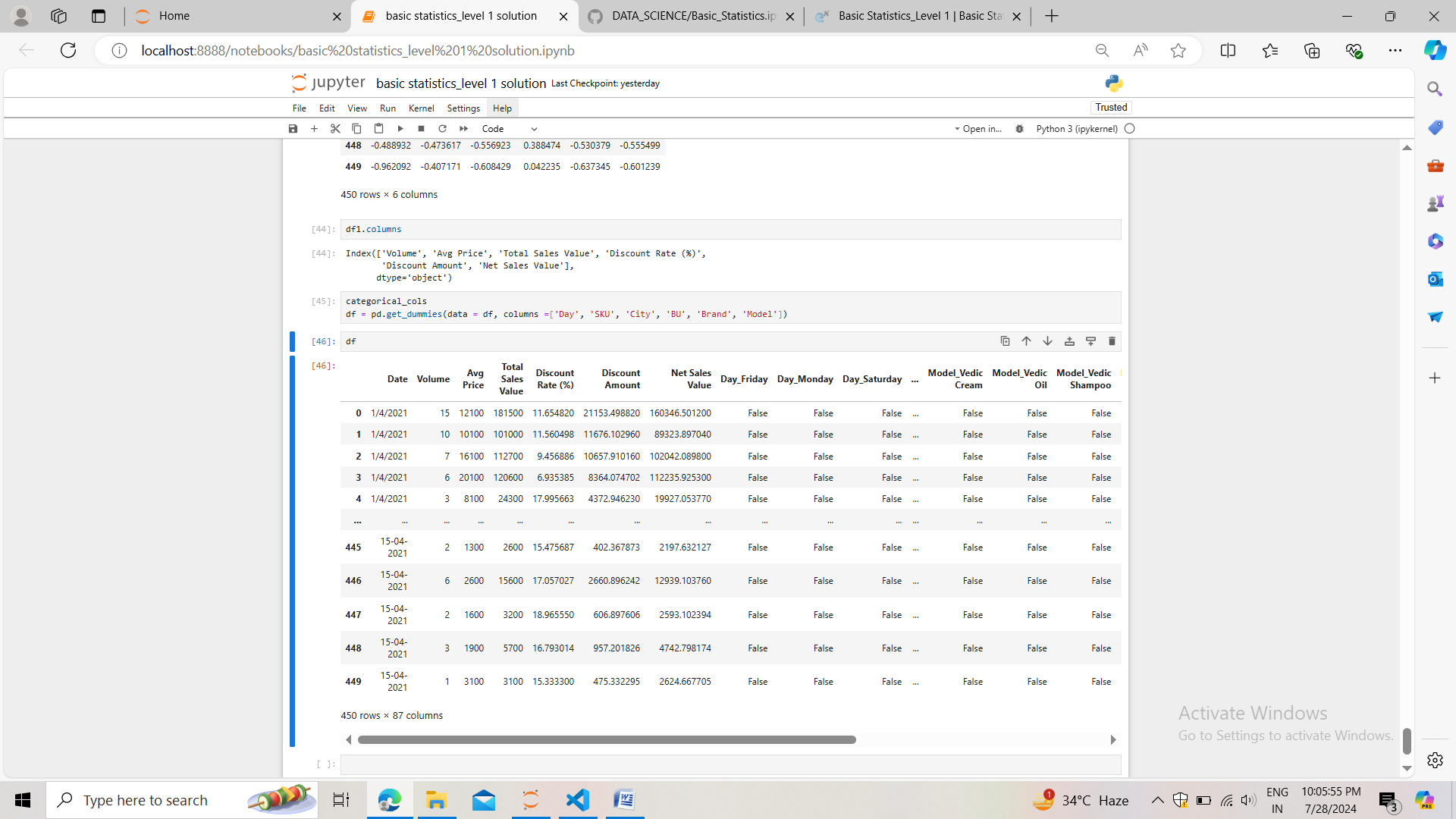
#### **Conversion of Categorical Data into Dummy Variables**

* Objective: To transform categorical variables into a format that can be provided to ML algorithms.
* Steps:
  1. Discuss the need for converting categorical data into dummy variables (one-hot encoding).
  2. Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.
  3. Display a portion of the transformed dataset.

Answer:-

* One hot encoding:- machine learning algos only understand numeric data that’s why we have to transform categorical data into numerical form.
* one hot encoding is used to transform nominal data. When X variables contains nominal data.
* In one hot encoding we create dummy variables because in nominal data we have multiple categories in one column and the categories don’t follow any order also don’t have any relation between categories. If we have 3 categories in one column then we will generate 3 dummy variables for all three categories.
* And then we do dummy variable trap and eliminate one of the dummy variable to avoid the multicolinearity problem.

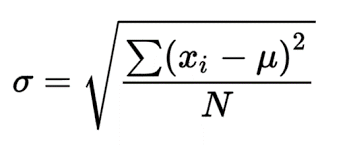




#### Conclusion

* 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.

Answer:-

* As we can see avg price, total sales value, discount amount and the net sales value are having more outliers. And also these numerical columns are of large range. Volume and discount rate % are of low range
* Except the discount rate % all the numerical columns like volume, average price, total sales value, discount amount, net sales value all are positively (+ve) skewed and only the discount rate % is negatively skewed.
* **mean** = (sum of all data points)/(total no. Of data)
* **mode** = more frequent item in the data. It can have more than one mode value.
* **Median** = (n+1)/2 :-for odd no. Of data items ((n/2)+(n/2+1))/2 :- for even no
* **Standard deviation =**
* machine learning algos only understand numeric data that’s why we have to transform categorical data into numerical form.
* **Standardization**:- In geometric intuition of standardization we do mean centering and scaling by the factor of standard deviation. the formula: z=x-mu/sigma transformed our original data to somewhere around zero. It is the process which removes the unit from the variables and put the data around -1 to +1. Or -3 to +3.
* **One hot encoding**:- machine learning algos only understand numeric data that’s why we have to transform categorical data into numerical form.
* one hot encoding is used to transform nominal data. When X variables contains nominal data.
* In one hot encoding we create dummy variables because in nominal data we have multiple categories in one column and the categories don’t follow any order also don’t have any relation between categories. If we have 3 categories in one column then we will generate 3 dummy variables for all three categories.
* And then we do dummy variable trap and eliminate one of the dummy variable to avoid the multicolinearity problem.