**Fall 2023 DATA 230 – 11**

**Data Visualization**

**Homework – 1**

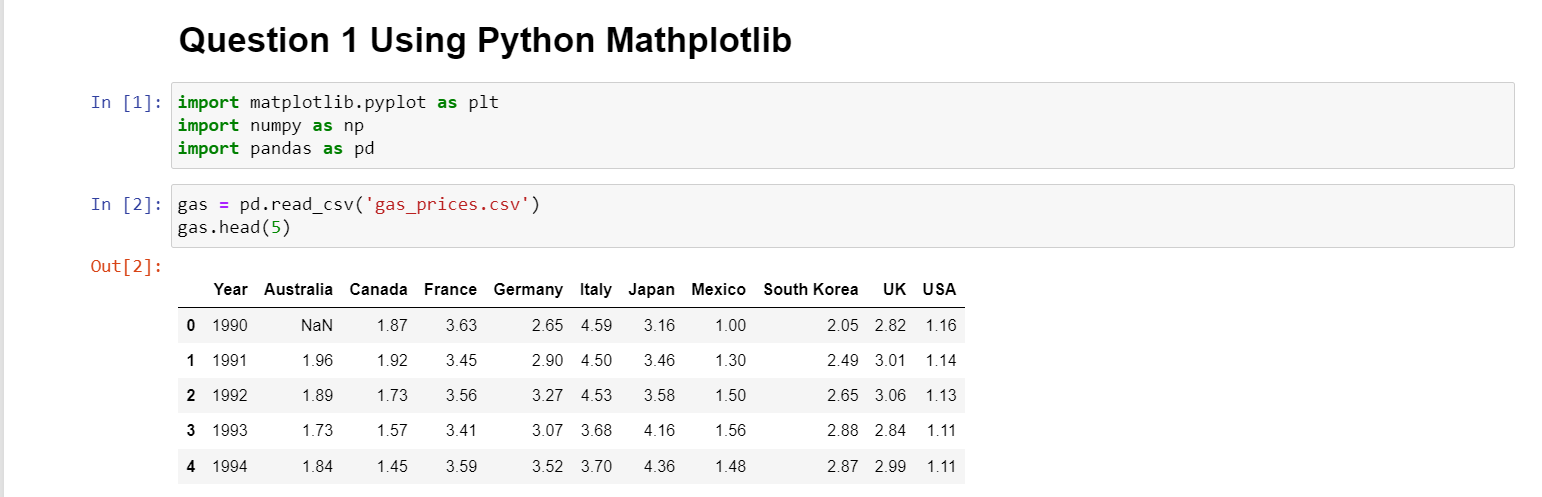
**Name :- Prayag Nikul Purani**

**SJSU Id :- 017416737**

**Question 1:**

**Data importing:**

We will be using pandas for importing the data as data frame

****

Now we will get the knowledge of the data by using the describe function.

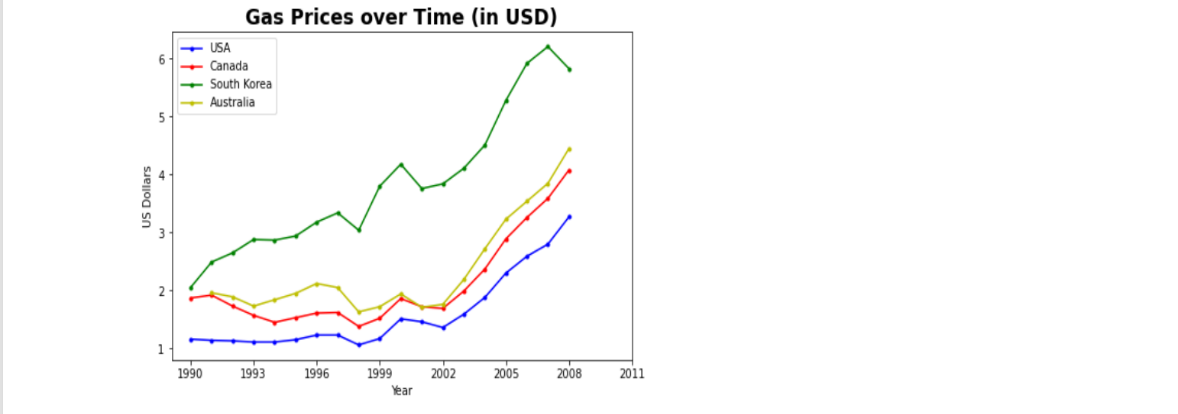
****

from the above image we can get many data about the database, so it will be easy to process for the visualization. So , the data tells us about the gas price of different countries from 1990 to 2008.

**Visualization 1: Line graph:**

So we will be comparing the prices of the some countries like USA, Canada, South Korea, Australia by using line graph of 4 different color. The x-axis will contain the years and the y-axis will contain the prices of the gas

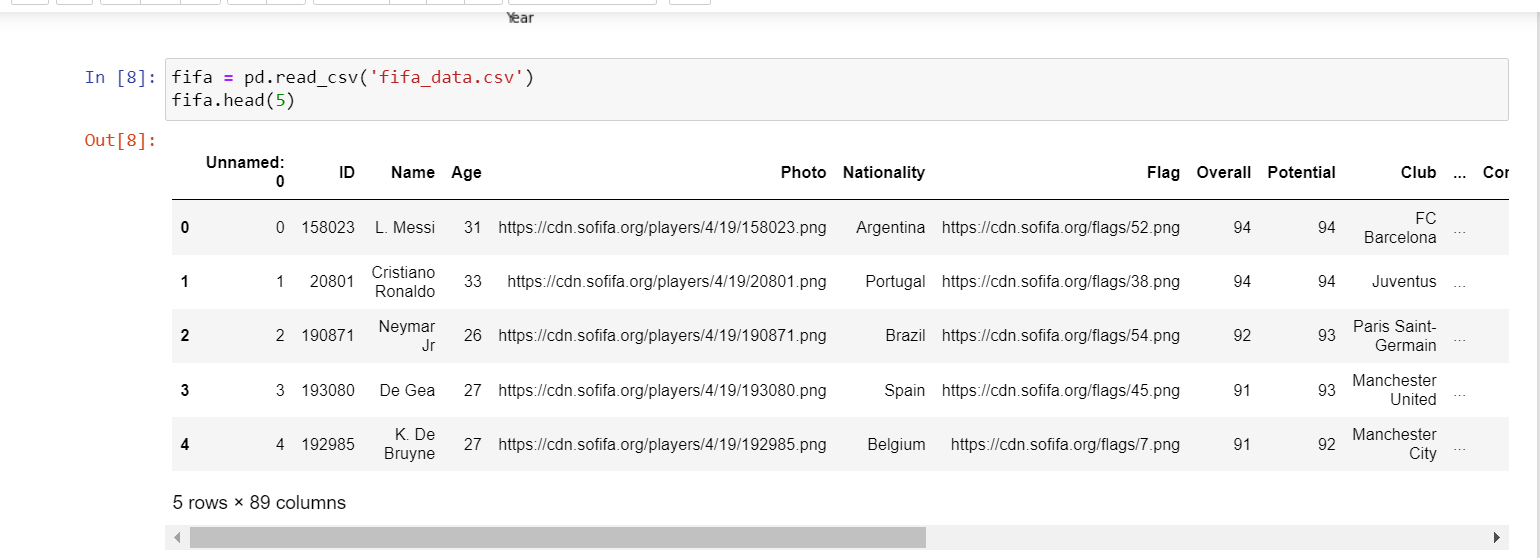
****

****

**Interpretation and Explanation:**

We can see that the South Korea has the highest price in 1990 and USA is at the lowest and at the also the trend remains the same the order of prices of all 4 country remain the same except between 1999 t0 2002 the price in Canada is slightly higher the Australia.

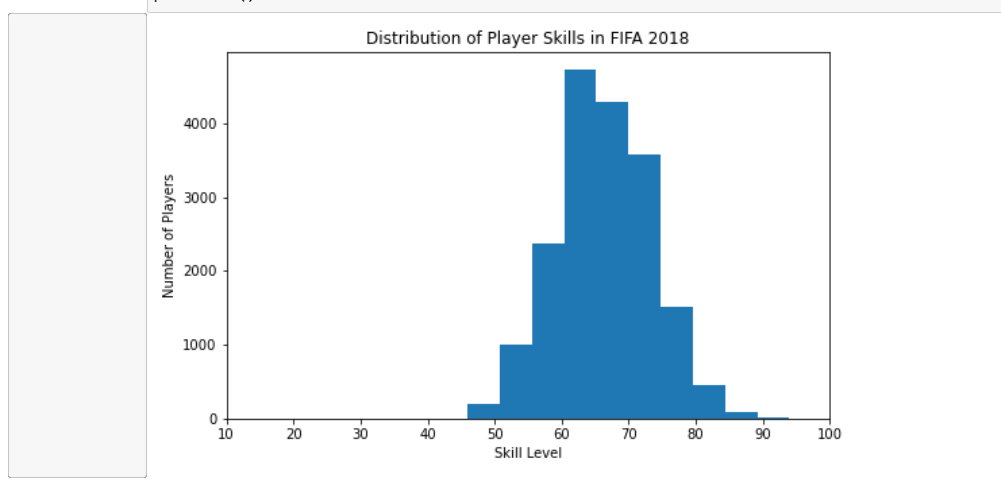
**Visualization 2: Histogram graph:**



The new data which is used is the fifa data which tells about the players who are playing the game and the photos, nationality their flag and so on



So, we will plot the histogram of the overall column which tell us about the skill level of the different player so we can know the density of the skill in which section is more.



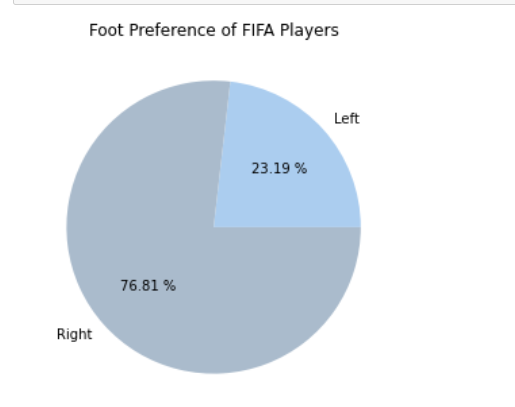
**Interpretation and Explanation :**

We see the bin size in the skill level is 10 and the highest is 100, we can also see that the density or the number of skills between 60-70 is more and the peak is the reason of this observation and the distribution is normal and the number of outliners are also very less.

**Visualization 3: Pie Chart:**

****

So, for the next visualization we will see the percentage of players using left or right foot as there are only two option for leg, it will be easy to understand using pie chart.



**Interpretation and Explanation :**

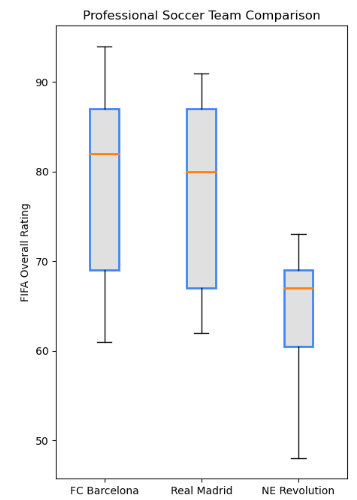
So, from the pie chart we can see that more than 75% of players are right footed and only 23% are left footed. So, from the pie chart we came to know that the ratio is not equal so the promotion or the advertisement could be in that way.

**Visualization 4: Box plot:**

****

We will be plotting the box for the 3 famous football clubs vs their overall rating so can know their mean, max, min, q1 and q2 and median about the different clubs.

1. **Median (Q2):** The line inside the box represents the median, which is the middle value of the dataset when it is ordered from lowest to highest. It indicates the central tendency of the data.
2. **Interquartile Range (IQR):** The box itself spans the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). The IQR contains the middle 50% of the data and provides information about the data's spread.
3. **Whiskers:** The whiskers extend from the box and show the range of the data within a certain limit. Typically, the whiskers extend to the minimum and maximum values within 1.5 times the IQR. Any data points beyond this range are considered potential outliers and are usually plotted individually.
4. **Outliers:** Individual data points that fall outside the whiskers are plotted as individual points and are considered potential outliers. Outliers can be significant because they may indicate data anomalies or interesting features of the dataset.
5. **Skewness:** The box plot's asymmetry can provide insights into the skewness of the data. If one whisker is longer than the other, it may suggest that the data is positively or negatively skewed.
6. **Spread and Variability:** You can visually assess how spread out or concentrated the data is based on the width of the box and the length of the whiskers. A wider box and longer whiskers indicate greater variability, while a narrower box and shorter whiskers indicate less variability.
7. **Data Distribution:** Box plots can help you understand the overall shape of the data distribution. For example, a symmetric box with equally long whiskers on both sides of the median suggests a roughly symmetric distribution, while an asymmetric box may indicate a skewed distribution.
8. **Comparative Analysis:** Box plots are useful for comparing multiple datasets. You can place multiple box plots side by side to compare their distributions and central tendencies easily.

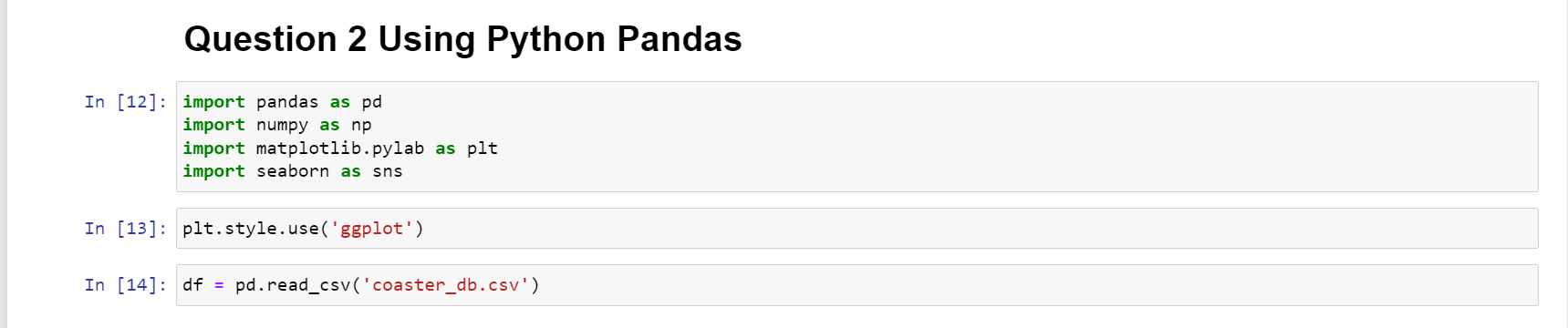


**Interpretation and Explanation :**

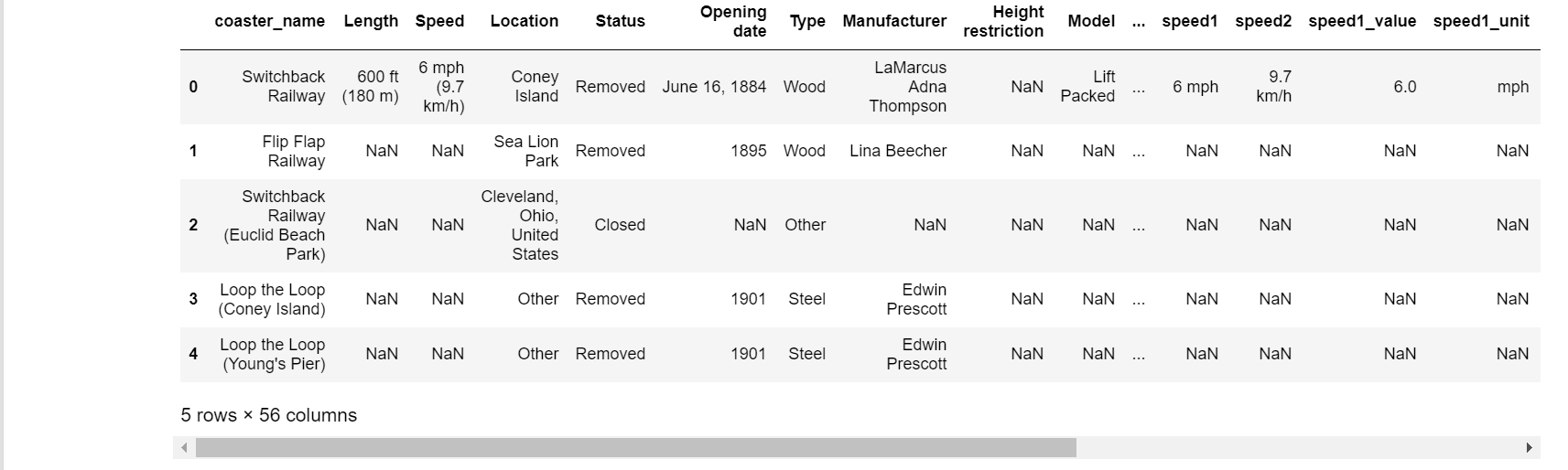
From the box plot we can know that the maximum value is for FC Barrcelona above 90 and the NE Revolution has the minimum value. In summary, when you observe a box plot, you can quickly assess the central tendency, spread, skewness, and presence of outliers within a dataset. This graphical representation is particularly valuable for understanding the characteristics of your data and making informed decisions in data analysis and visualization.

**Question 2:**

**Data importing:**

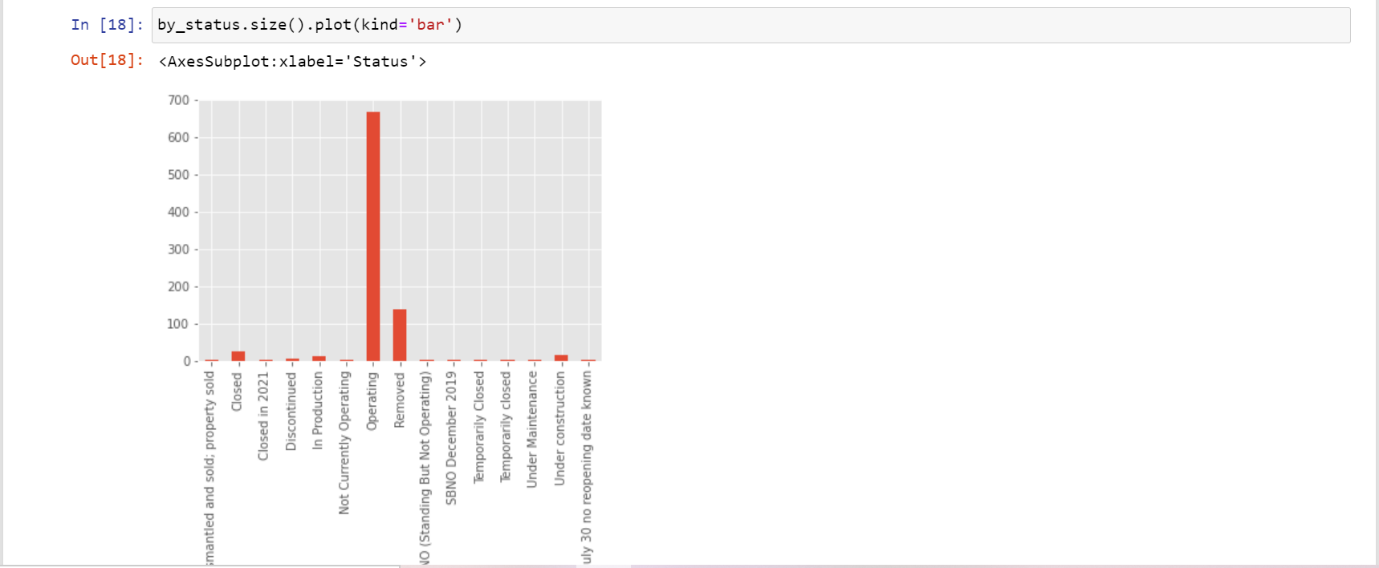
****

We are uploading the roller coaster dataset.



So from the dataset picture we can see that the data cleaning has to be performed and is done in the jupyter file.

**Visualization 1: Bar Chart:**

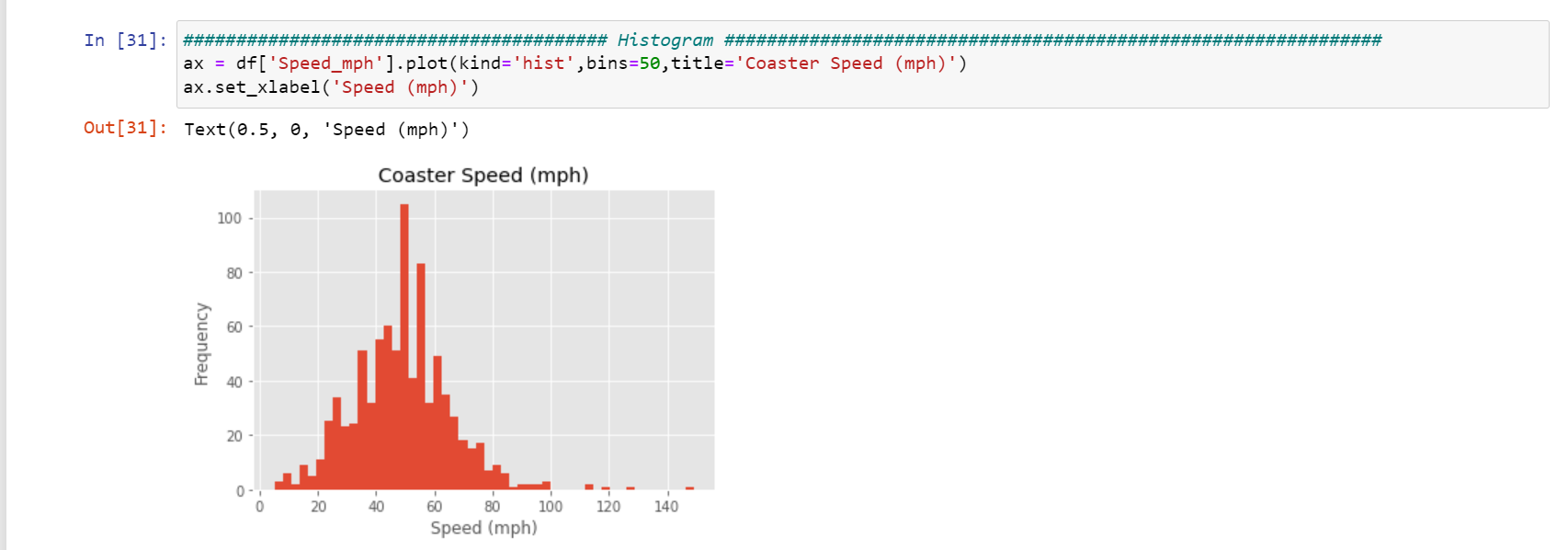


So the main aim of this task is to perform the visualization using pandas. So we will be plotting the bar graph with the x-axis containing the status condition and the y-axis containing the number of roller coaster in that particular condition.

**Interpretation and Explanation :**

We can visualize that nearly 700 number of roller coasters are in working condition and the number outliners are higher, as the categories present in the status is very high so it is difficult to know the count of the status that are less then 10.

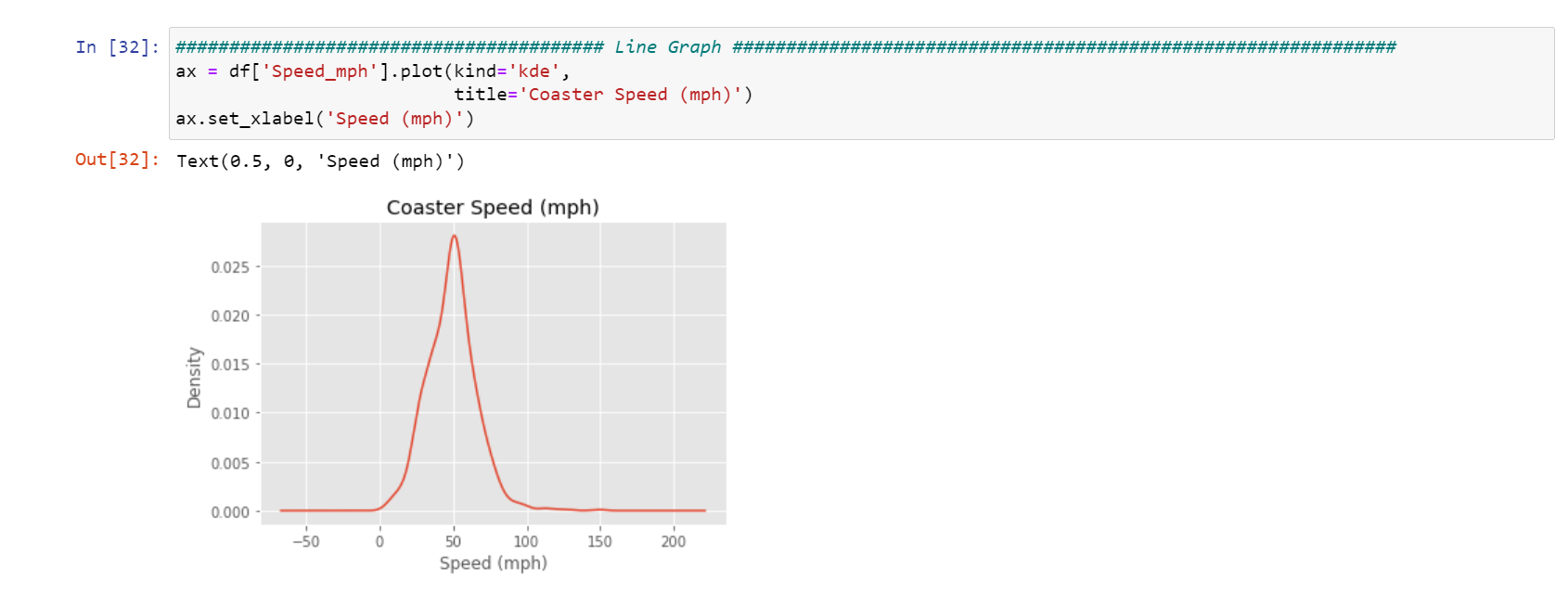
**Visualization 2: Histogram Chart:**



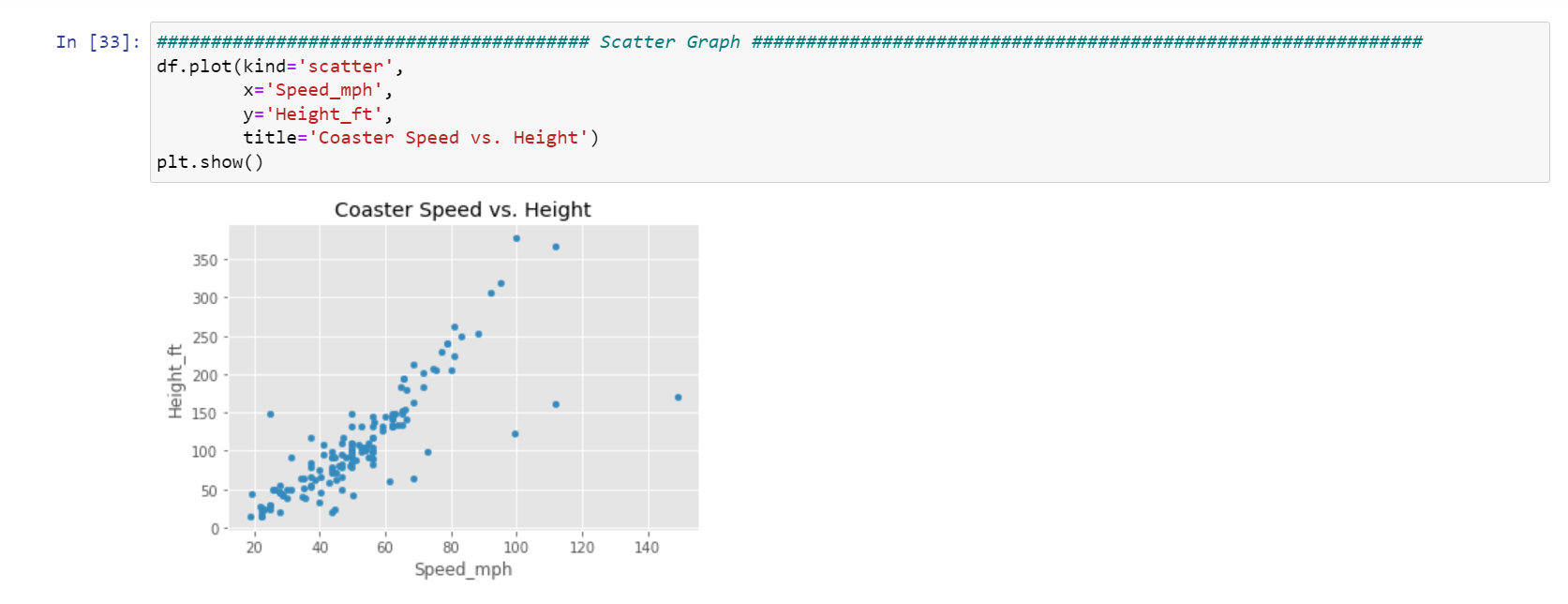
We are plotting the histogram of speed in mph of all the roller coaster present in the dataset to see the maximum number of roller coaster and the spread of the data with will help us to know the further analysis and the conclusion for some events can be made from this type of fields only.

**Interpretation and Explanation :**

We can see that from the plot that the data is spread from 0 to 150 and the the density of the data is high between 40-60, from this graph we can also tell that the median is also nearly 50 and that can be proved from the line graph shown below.



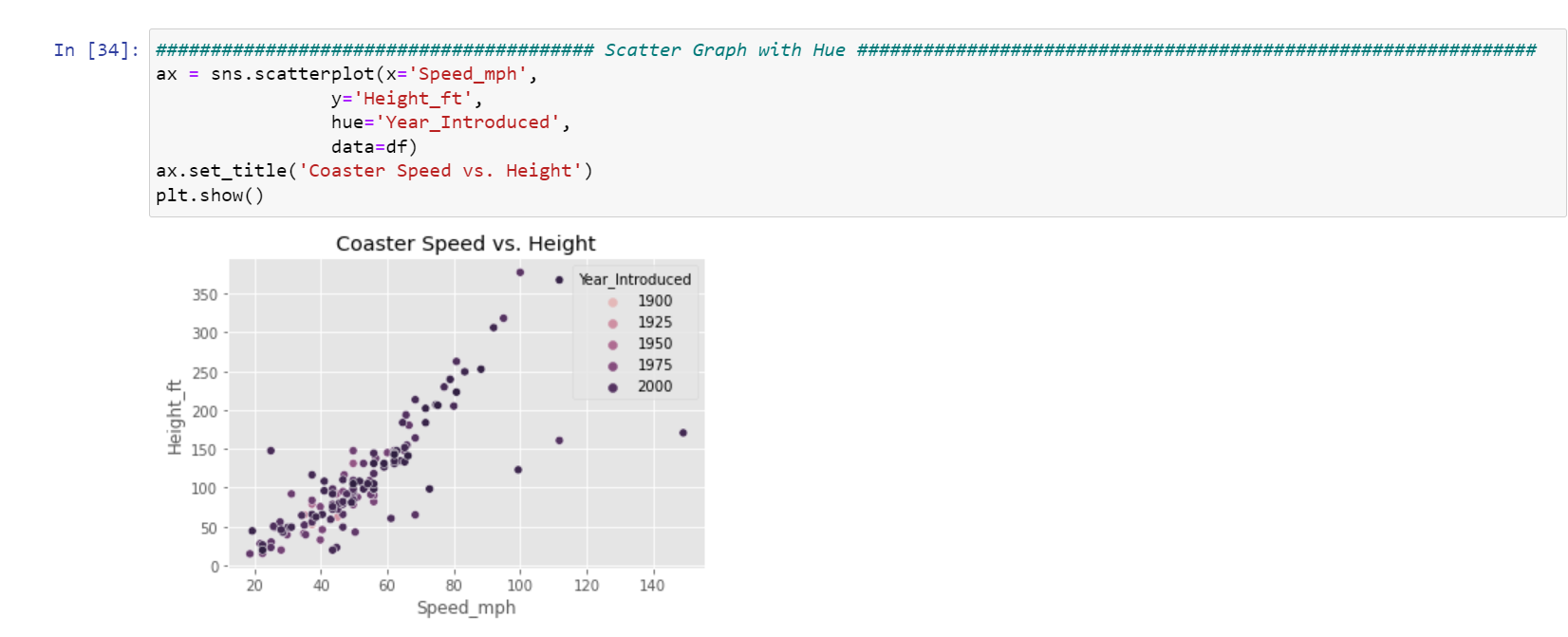
**Visualization 3: Scatter Plot:**

****

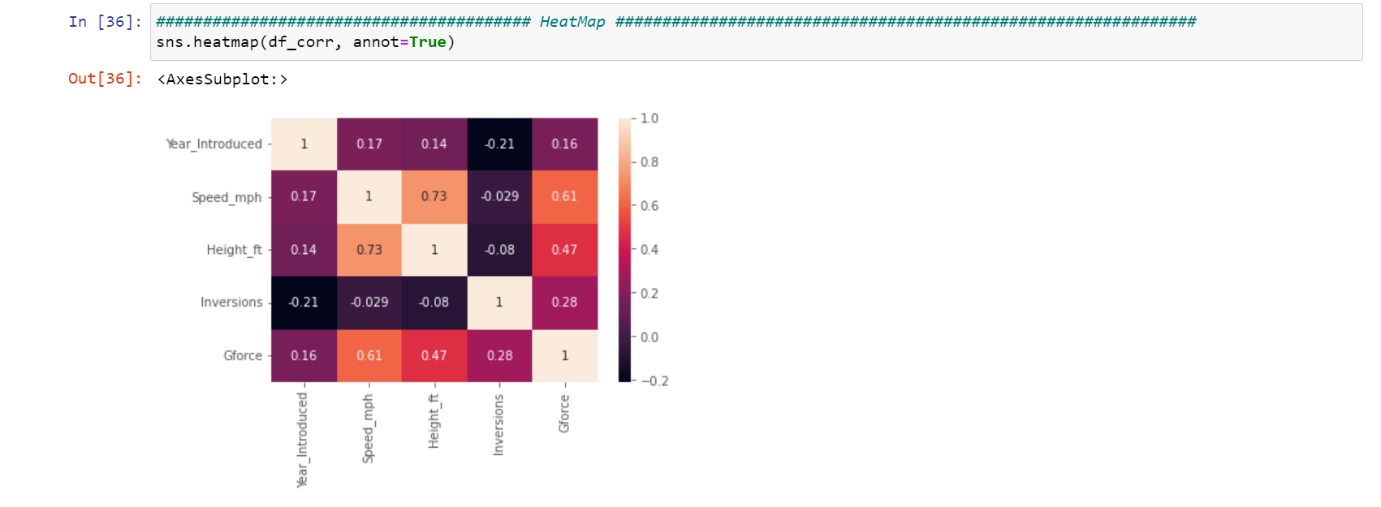
The Scatter plot helps us to tell the relation between the two quantities, so here we have selected the speed vs height of the roller coaster.

**Interpretation and Explanation :**

From the scatter plot we can say that these two quantities are highly related to one another as the graph is increasing for each values apart from some points. The scatter plot has high importance because the it can add many features in it for better understanding for example the below plot has the hue feature combined in it with year the coaster was introduced.



**Visualization 4: Heatmap:**



So, heatmap is used to find the correlation between all the columns present in them and the colors will talk about how strong the relation is between the two entities.

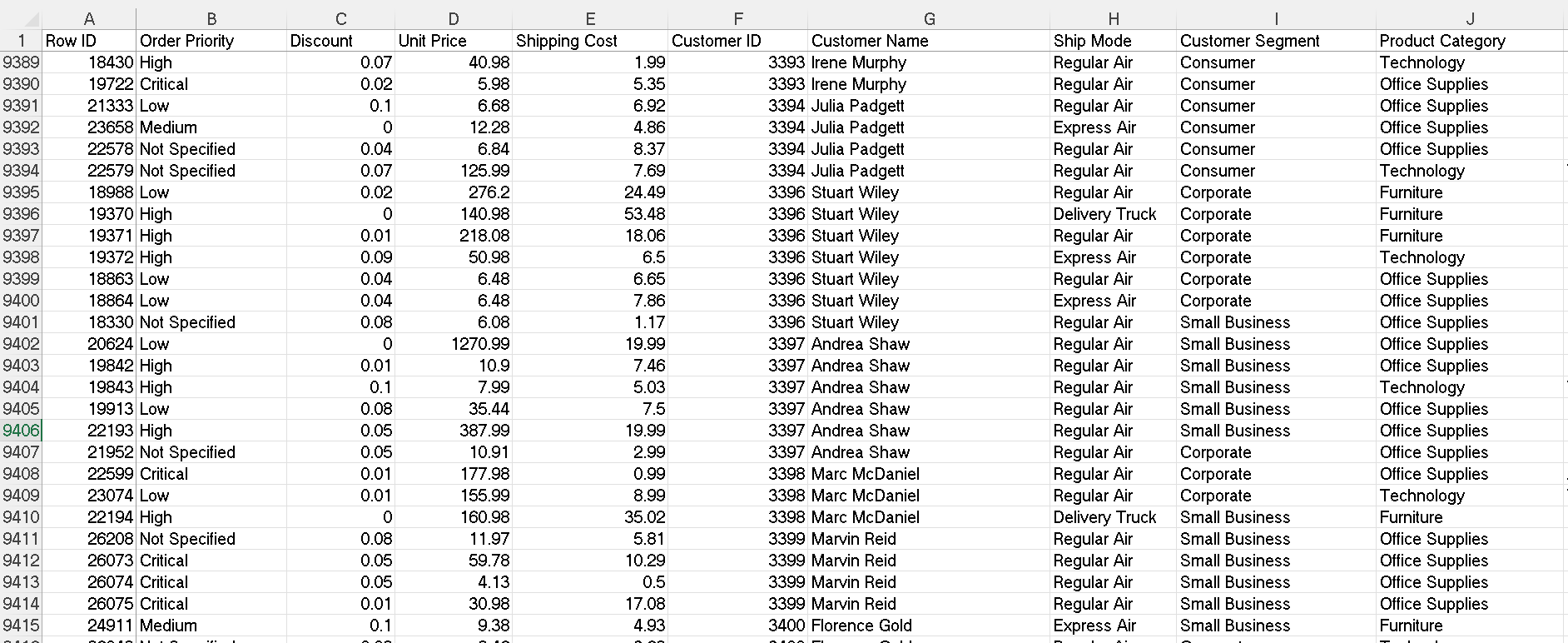
**Interpretation and Explanation:**

Form the heatmap we can prove the height and the speed have a good and strong relation between them and same can be done for rest of the columns.

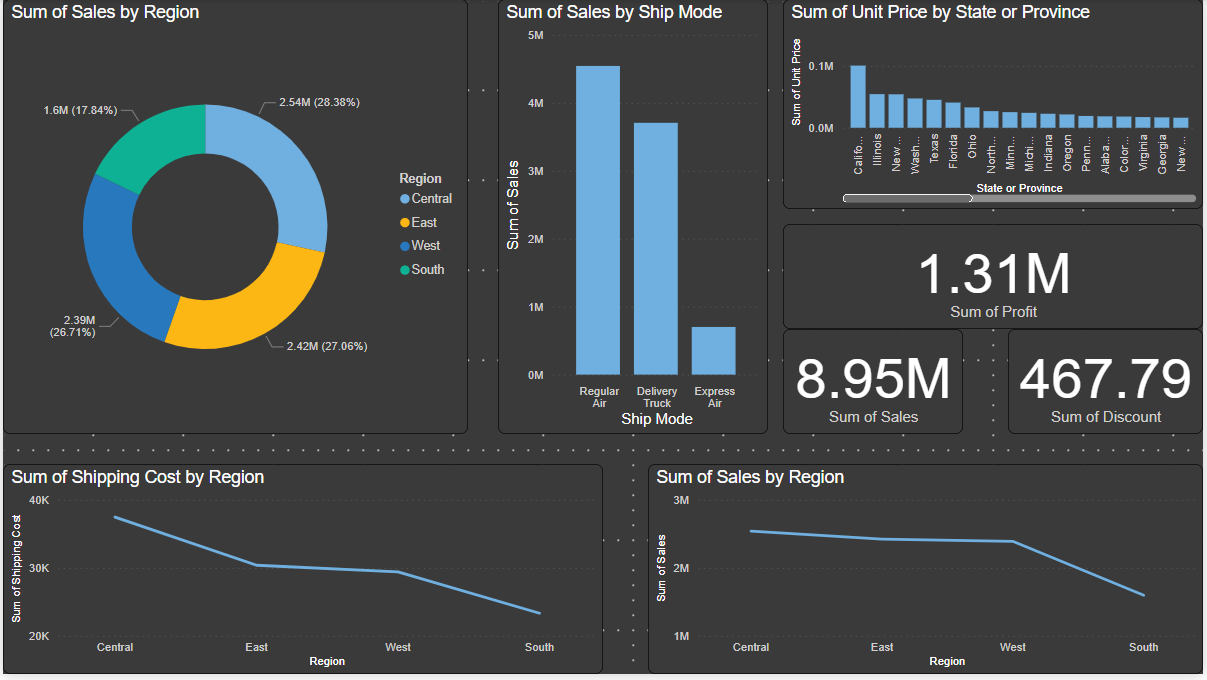
**Question 3:**

**Power BI:-**

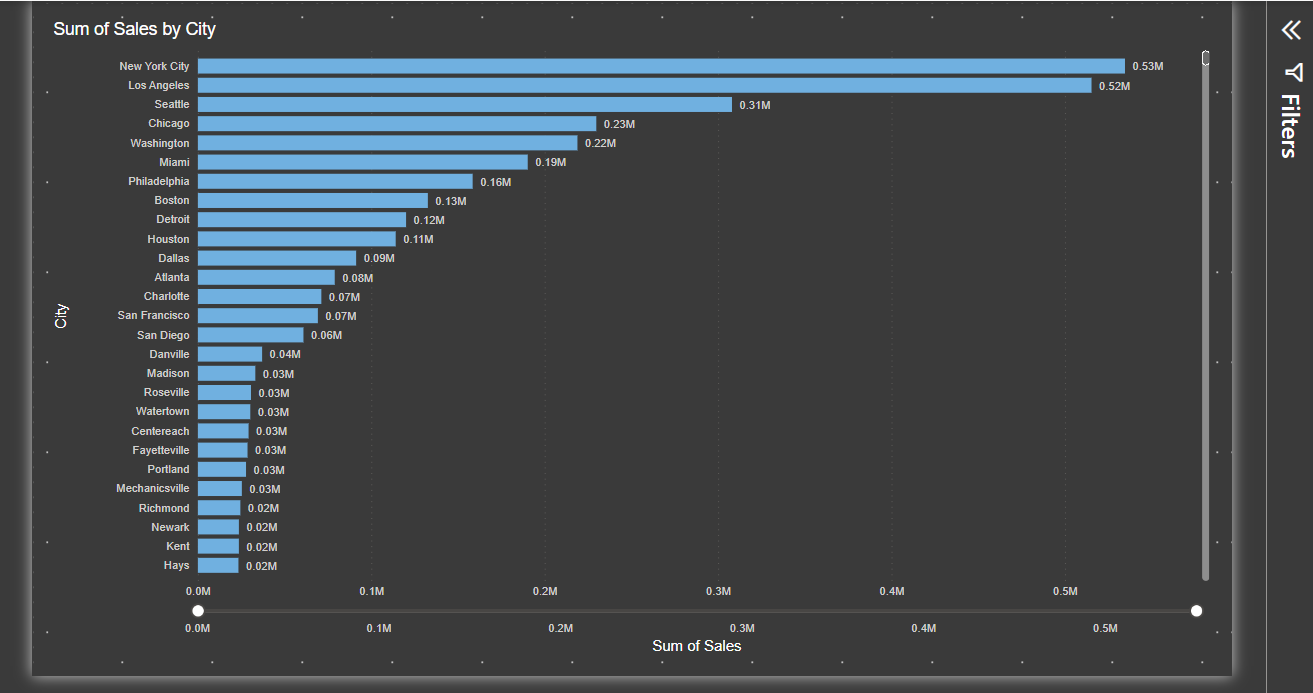
**So, I have used the Superstore dataset for the power bi dashboard creation:**

****

So, this is the small sample of my learning from the guest lecture. The dash board has the 4 types of different plots donut pie chat, line chart, bar graph and the sum box plot. So, from this we can see how powerful is this tool.

****

****

****

Pie chart: Its tells us about the sum of sales in different region.

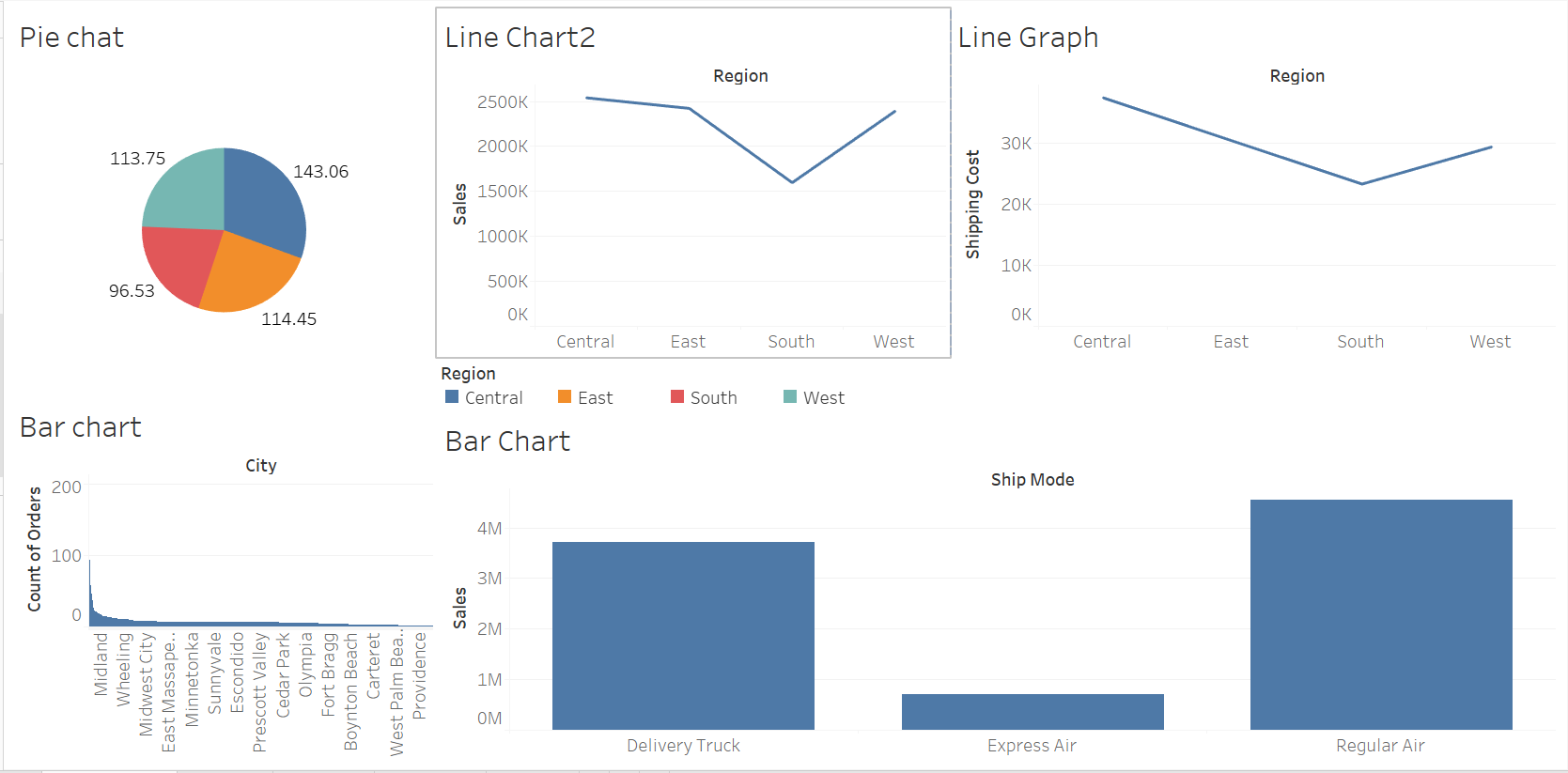
Bar chart : It tells us how the shipping mode effect the purchases.

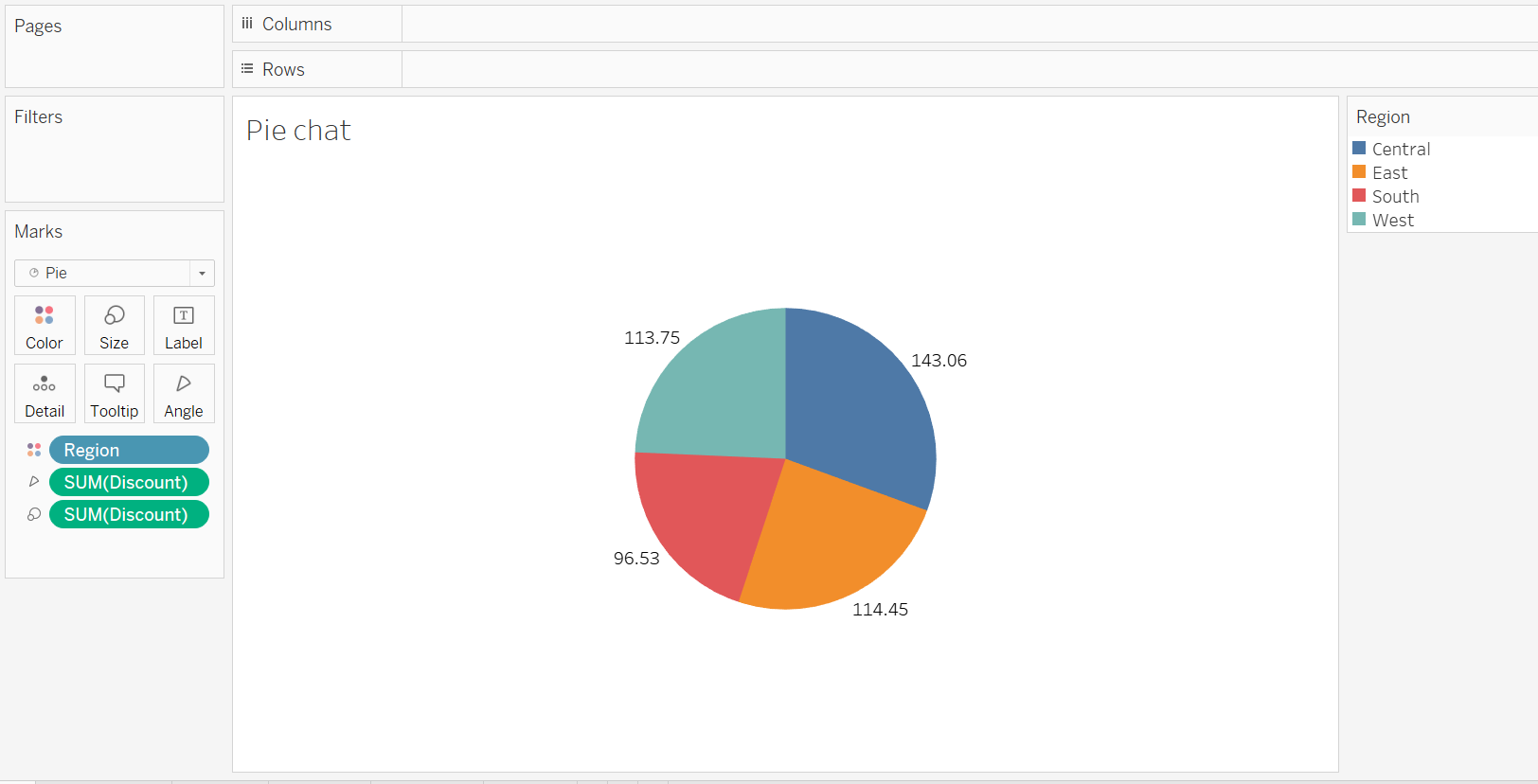
Total Box: It tells us about the total amount of sales, profit and discount

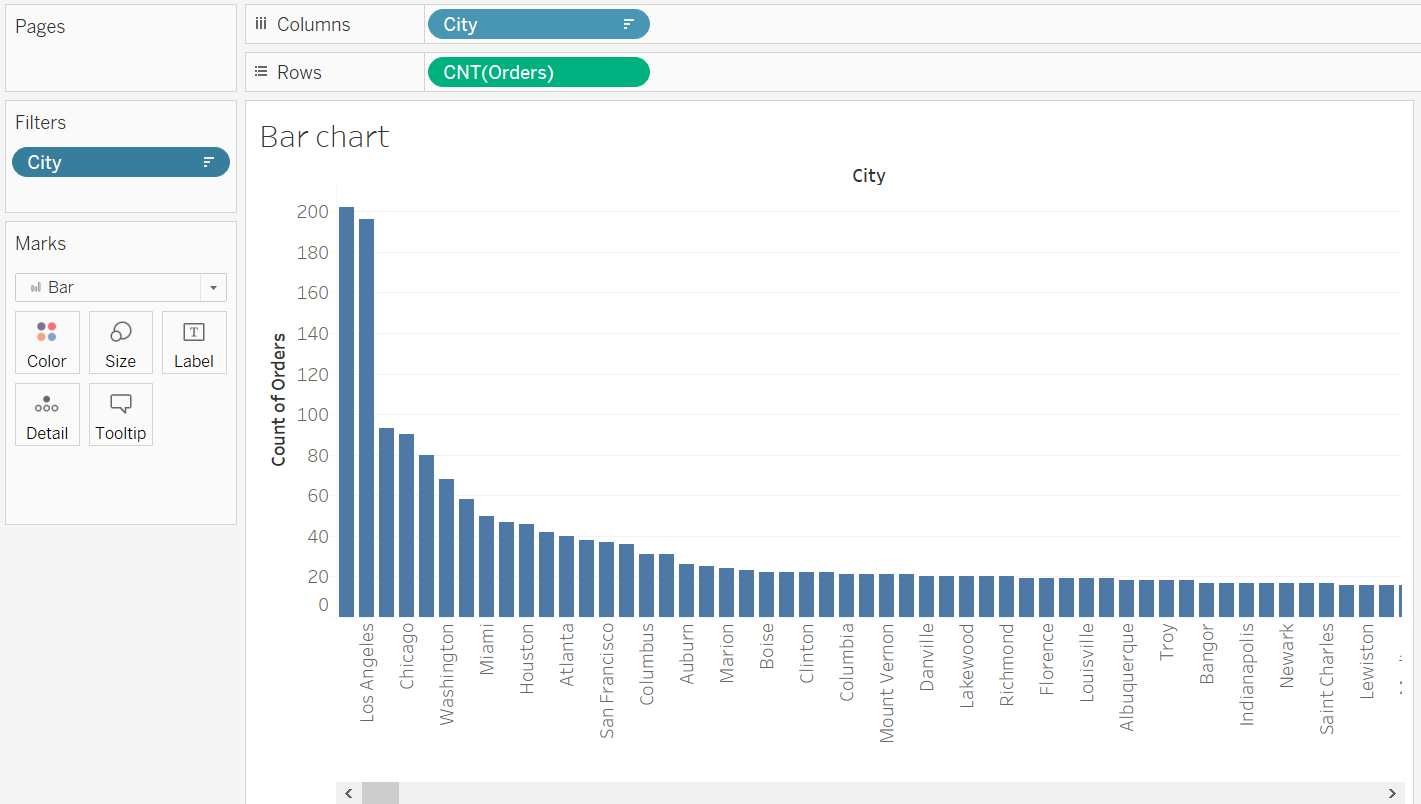
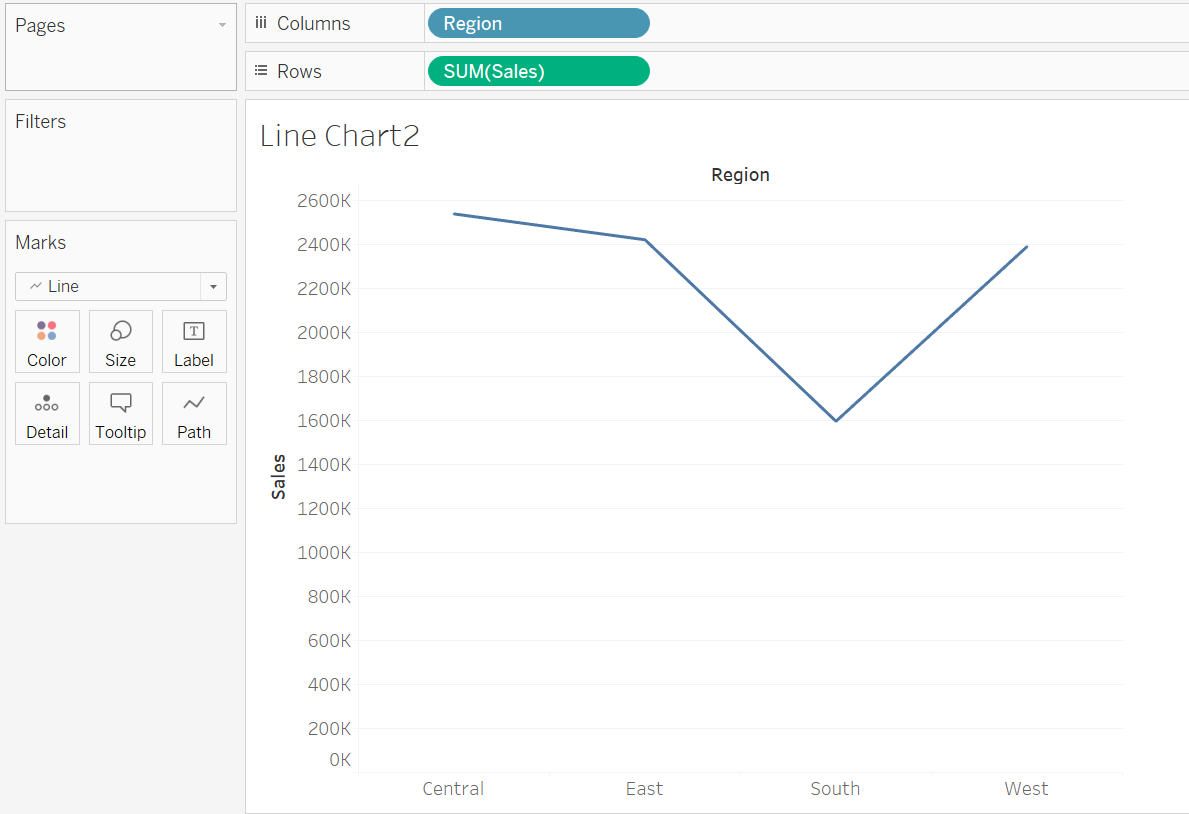
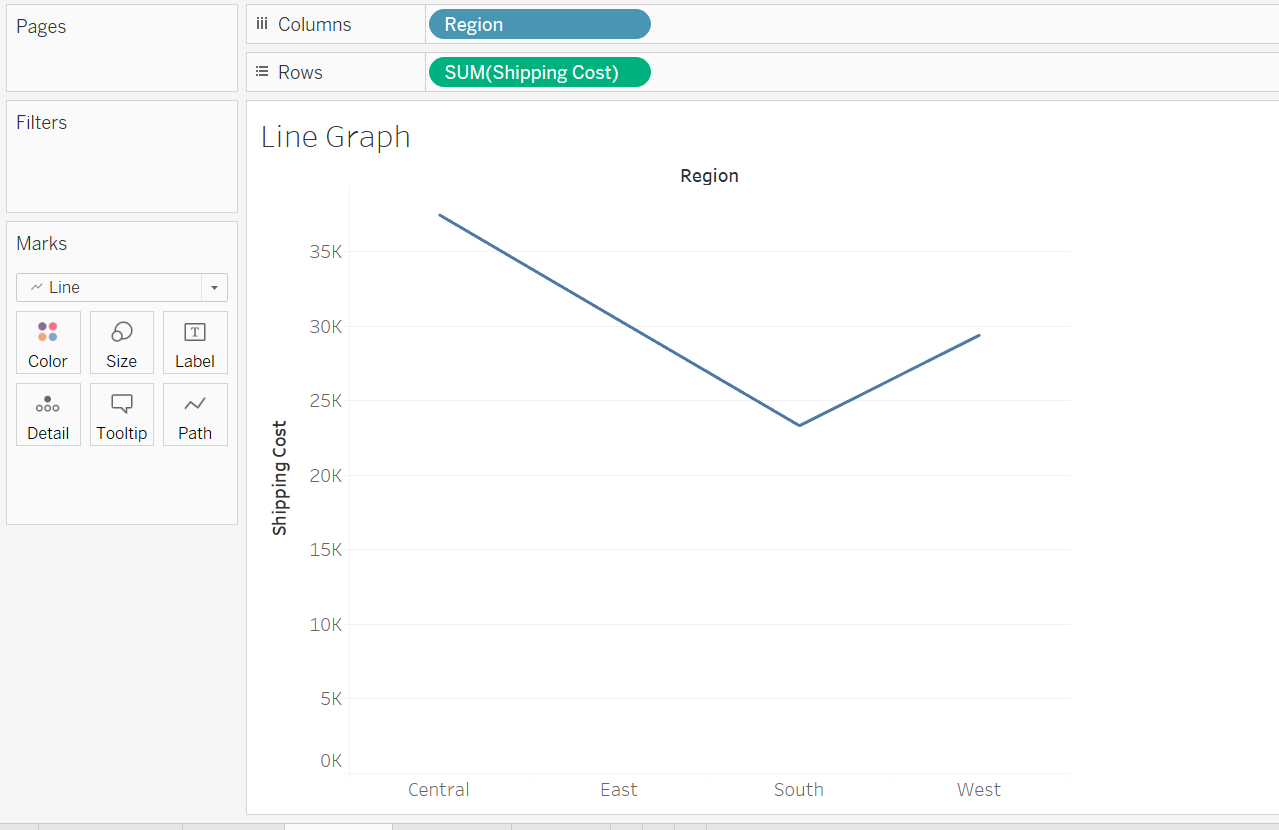
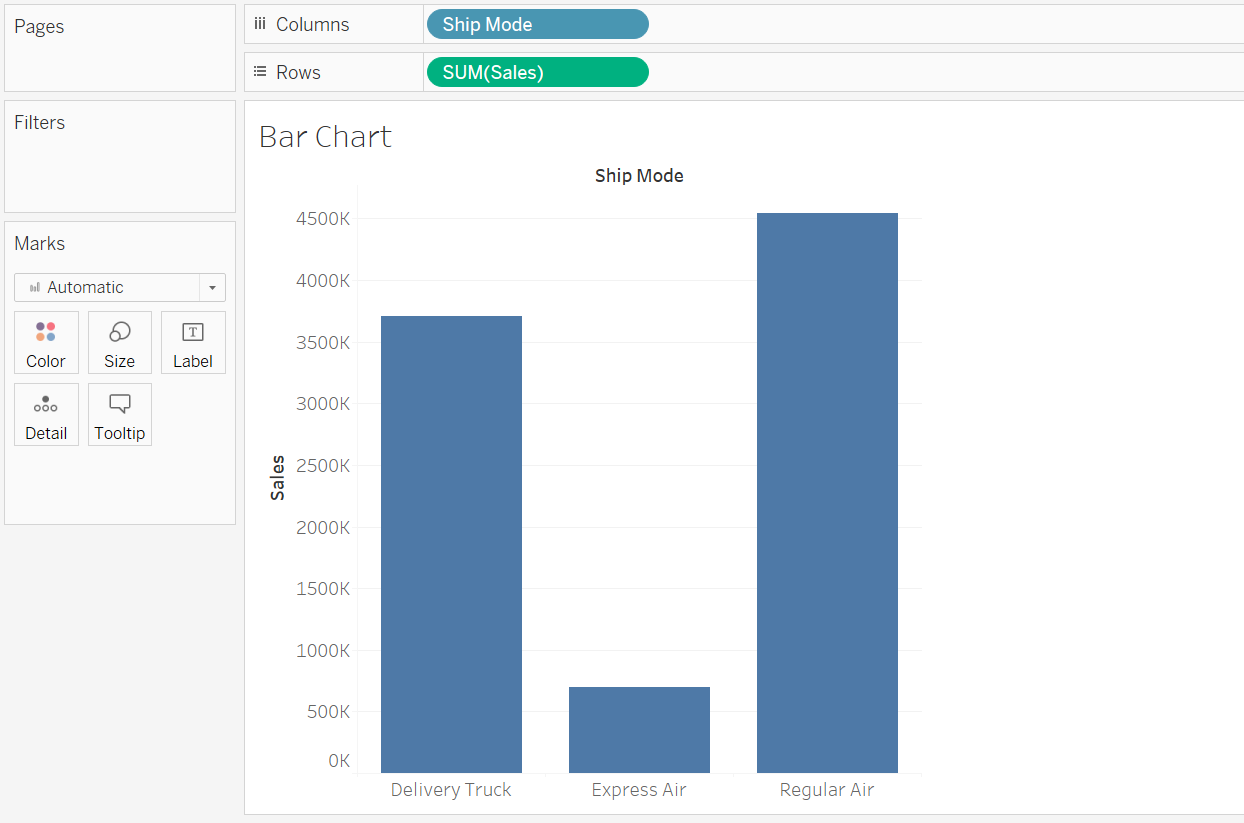
Line plot: Sum of different plots depending on the areas.

**Tableau: -**

**So, for this also I am using the same dataset.**

****

****

****

1. Line Chart: A line chart is a graphical representation of data that uses a series of data points connected by straight lines. It is often used to display trends and changes in data over time, making it useful for showing continuous data, such as stock prices or temperature fluctuations.
2. Bar Chart: A bar chart, also known as a bar graph, uses rectangular bars of varying lengths to represent data values. It is effective for comparing discrete categories or data sets. The height or length of each bar is proportional to the data it represents, making it easy to visualize differences between categories.
3. Pie Chart: A pie chart is a circular graph divided into slices, each representing a portion of a whole. It is ideal for illustrating the composition of a single data set, where each slice represents a percentage or proportion of the total. Pie charts help visualize the distribution of data categories in relation to the whole.