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

Question 1 Using Python Mathplotlib

```
In [1]: import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd

In [2]: gas = pd.read_csv('gas_prices.csv')
        gas.head(5)
```

Out[2]:

	Year	Australia	Canada	France	Germany	Italy	Japan	Mexico	South Korea	UK	USA
0	1990	NaN	1.87	3.63	2.65	4.59	3.16	1.00	2.05	2.82	1.16
1	1991	1.96	1.92	3.45	2.90	4.50	3.46	1.30	2.49	3.01	1.14
2	1992	1.89	1.73	3.56	3.27	4.53	3.58	1.50	2.65	3.06	1.13
3	1993	1.73	1.57	3.41	3.07	3.68	4.16	1.56	2.88	2.84	1.11
4	1994	1.84	1.45	3.59	3.52	3.70	4.36	1.48	2.67	2.99	1.11

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

```
In [4]: gas.describe()
```

Out[4]:

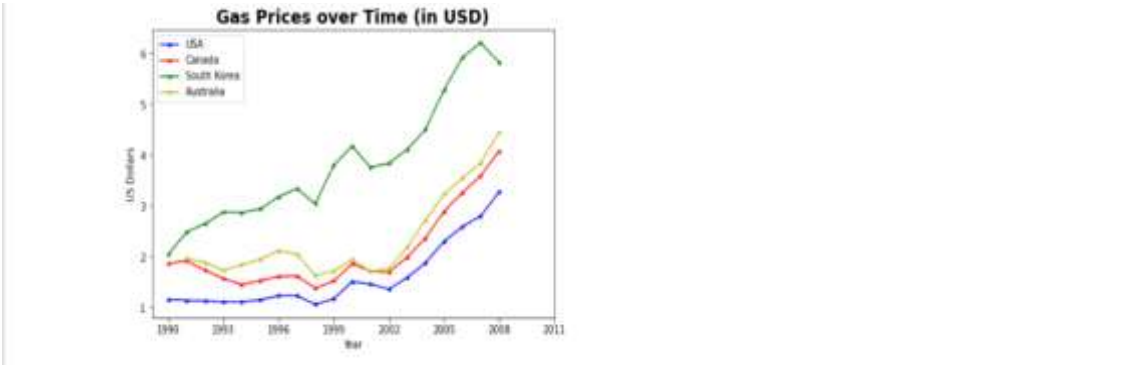
	Year	Australia	Canada	France	Germany	Italy	Japan	Mexico	South Korea	UK	USA
count	19.000000	18.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000
mean	1999.000000	2.348889	2.086842	4.407895	4.224737	4.645789	3.820526	1.781579	3.835789	4.382632	1.592105
std	5.627314	0.845931	0.786818	1.167531	1.425749	1.149610	0.696815	0.462148	1.226170	1.455170	0.683355
min	1990.000000	1.630000	1.380000	3.410000	2.650000	3.570000	2.820000	1.000000	2.050000	2.820000	1.060000
25%	1994.500000	1.780000	1.590000	3.605000	3.370000	3.805000	3.270000	1.475000	2.910000	3.135000	1.145000
50%	1999.000000	1.955000	1.730000	3.870000	3.530000	4.390000	3.640000	1.790000	3.760000	4.130000	1.230000
75%	2003.500000	2.587500	2.180000	4.700000	4.915000	4.940000	4.320000	2.210000	4.345000	5.130000	1.735000
max	2008.000000	4.450000	4.080000	7.510000	7.750000	7.630000	5.740000	2.450000	6.210000	7.420000	3.270000

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

```
In [5]: ########################################## Line Graph ##########################################################################
plt.figure(figsize=(8,5))

plt.title('Gas Prices over Time (in USD)', fontdict={'fontweight':'bold', 'fontsize': 18})

plt.plot(gas.Year, gas.USA, 'b.-', label='United States')
plt.plot(gas.Year, gas.Canada, 'r.-')
plt.plot(gas.Year, gas['South Korea'], 'g.-')
plt.plot(gas.Year, gas.Australia, 'y.-')

# Another way to plot many values!
countries_to_look_at = ['USA','Canada', 'South Korea', 'Australia']

plt.xticks(gas.Year[::3].tolist()+[2011])

plt.xlabel('Year')
plt.ylabel('US Dollars')

plt.legend(countries_to_look_at)

plt.savefig('Gas_price_figure.png', dpi=300)

plt.show()
```



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:

```
In [8]: fifa = pd.read_csv('fifa_data.csv')
fifa.head(5)
```

Out[8]:

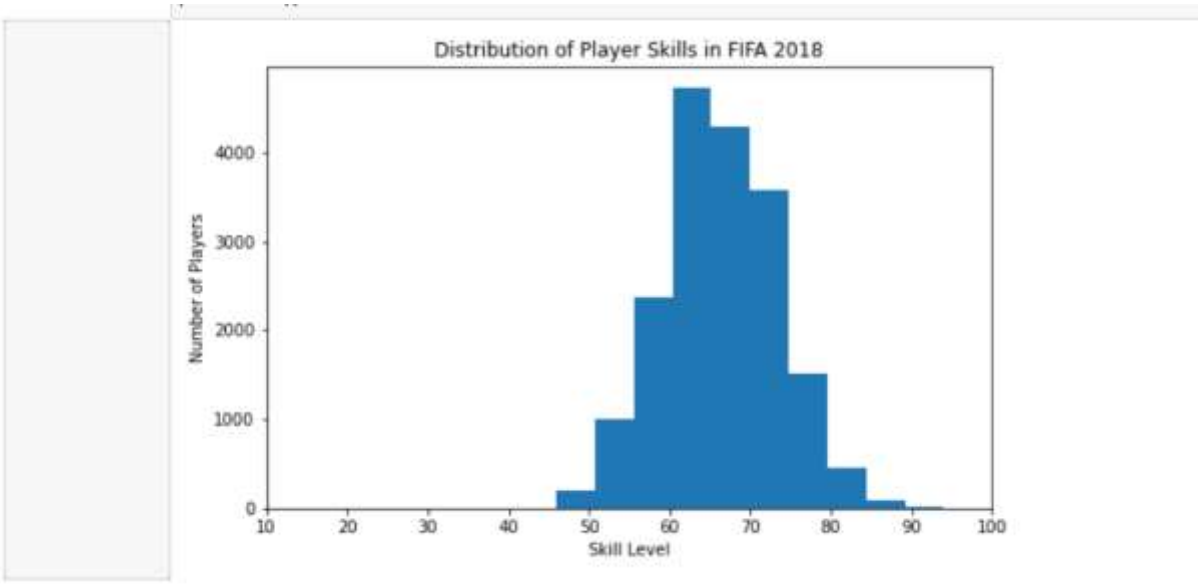
	Unnamed: 0	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	...	Cor
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona	...	
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus	...	
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-German	...	
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United	...	
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City	...	

5 rows × 89 columns

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

```
In [10]: ######################################## Histogram Graph ########################################################################
bins = [10,20,30,40,50,60,70,80,90,100]

plt.figure(figsize=(8,5))

plt.hist(fifa.Overall)

plt.xticks(bins)

plt.ylabel('Number of Players')
plt.xlabel('Skill Level')
plt.title('Distribution of Player Skills in FIFA 2018')

plt.show()
```

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.



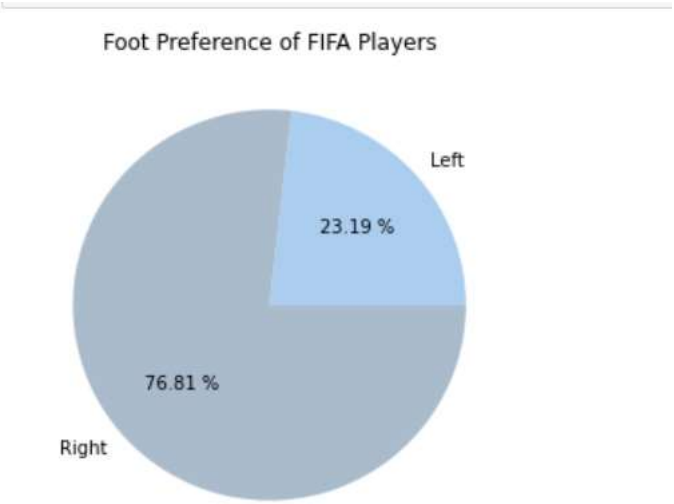
Interpretation and Explanation :

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:

```
In [11]: ######################################## Pie Chat ########################################################################
left = fifa.loc[fifa['Preferred Foot'] == 'Left'].count()[0]
right = fifa.loc[fifa['Preferred Foot'] == 'Right'].count()[0]

plt.figure(figsize=(8,5))

labels = ['Left', 'Right']
colors = ['#abcdef', '#aabbcc']

plt.pie([left, right], labels = labels, colors=colors, autopct='%.2f %%')

plt.title('Foot Preference of FIFA Players')

plt.show()
```

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:

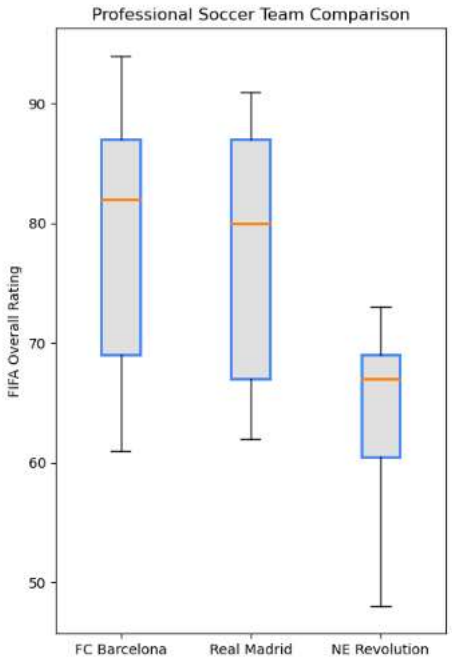
```
In [17]: ######################################## Box Graph ##########################################################################
         plt.figure(figsize=(5,8), dpi=100)

         plt.style.use('default')

         barcelona = fifa.loc[fifa.Club == "FC Barcelona"]['Overall']
         madrid = fifa.loc[fifa.Club == "Real Madrid"]['Overall']
         revs = fifa.loc[fifa.Club == "New England Revolution"]['Overall']

         #bp = plt.boxplot([barcelona, madrid, revs], labels=['a','b','c'], boxprops=dict(facecolor='red'))
         bp = plt.boxplot([barcelona, madrid, revs], labels=['FC Barcelona','Real Madrid','NE Revolution'], patch_artist=True, medianprops

         plt.title('Professional Soccer Team Comparison')
         plt.ylabel('FIFA Overall Rating')

         for box in bp['boxes']:
             # change outline color
             box.set(color='#4286f4', linewidth=2)
             # change fill color
             box.set(facecolor = '#e0e0e0' )
             # change hatch
             #box.set(hatch = '/')

         plt.show()
```

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:

Question 2 Using Python Pandas

```
In [12]: import pandas as pd
         import numpy as np
         import matplotlib.pylab as plt
         import seaborn as sns

In [13]: plt.style.use('ggplot')

In [14]: df = pd.read_csv('coaster_db.csv')
```

We are uploading the roller coaster dataset.

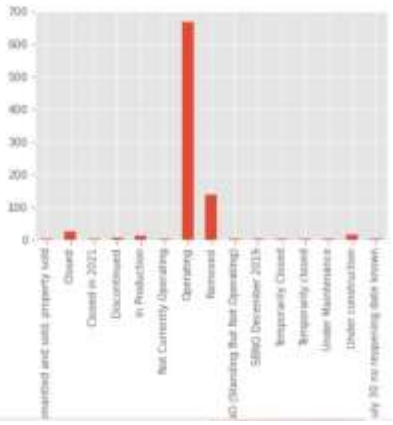
	coaster_name	Length	Speed	Location	Status	Opening date	Type	Manufacturer	Height restriction	Model	...	speed1	speed2	speed1_value	speed1_unit
0	Switchback Railway	600 ft (180 m)	6 mph (9.7 km/h)	Coney Island	Removed	June 16, 1884	Wood	LaMarcus Adna Thompson	NaN	Lift Packed	...	6 mph	9.7 km/h	6.0	mph
1	Flip Flap Railway	NaN	NaN	Sea Lion Park	Removed	1895	Wood	Lina Beecher	NaN	NaN	...	NaN	NaN	NaN	NaN
2	Switchback Railway (Euclid Beach Park)	NaN	NaN	Cleveland, Ohio, United States	Closed	NaN	Other	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
3	Loop the Loop (Coney Island)	NaN	NaN	Other	Removed	1901	Steel	Edwin Prescott	NaN	NaN	...	NaN	NaN	NaN	NaN
4	Loop the Loop (Young's Pier)	NaN	NaN	Other	Removed	1901	Steel	Edwin Prescott	NaN	NaN	...	NaN	NaN	NaN	NaN

5 rows × 56 columns

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:

```
In [18]: by_status.size().plot(kind='bar')
Out[18]: <AxesSubplot:xlabel='Status'>
```



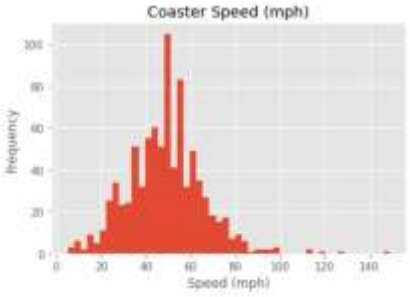
So the main aim of this task is to perform the visualization using pandas. So we will be plotting the bar graph with the x-axis containing the status condition and the y-axis containing the number of roller coaster in that particular condition.

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:

```
In [31]: ######################################## Histogram ##########################################################
         ax = df['Speed_mph'].plot(kind='hist',bins=50,title='Coaster Speed (mph)')
         ax.set_xlabel('Speed (mph)')
Out[31]: Text(0.5, 0, 'Speed (mph)')
```

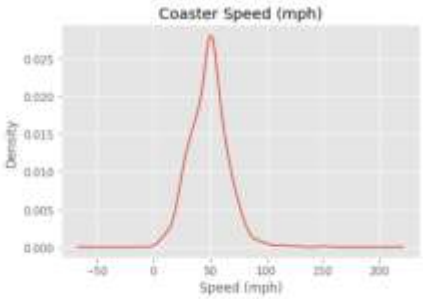


We are plotting the histogram of speed in mph of all the roller coaster present in the dataset to see the maximum number of roller coaster and the spread of the data with will help us to know the further analysis and the conclusion for some events can be made from this type of fields only.

Interpretation and Explanation :

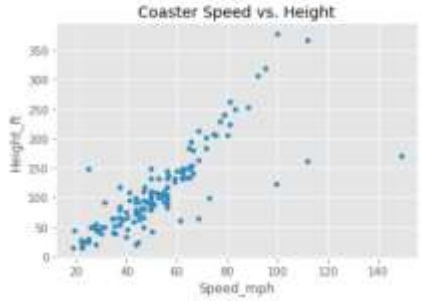
We can see that from the plot that the data is spread from 0 to 150 and the the density of the data is high between 40-60, from this graph we can also tell that the median is also nearly 50 and that can be proved from the line graph shown below.

```
In [32]: ######################################## Line Graph ##########################################################
         ax = df['Speed_mph'].plot(kind='kde',
                                   title='Coaster Speed (mph)')
         ax.set_xlabel('Speed (mph)')
Out[32]: Text(0.5, 0, 'Speed (mph)')
```



Visualization 3: Scatter Plot:

```
In [33]: ######################################## Scatter Graph ########################################################################
df.plot(kind='scatter',
        x='Speed_mph',
        y='Height_ft',
        title='Coaster Speed vs. Height')
plt.show()
```

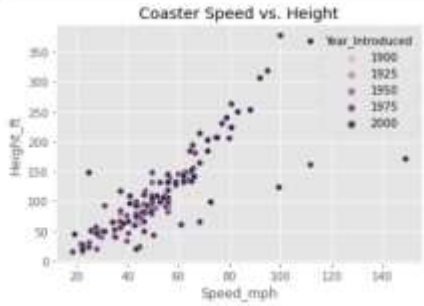


The Scatter plot helps us to tell the relation between the two quantities, so here we have selected the speed vs height of the roller coaster.

Interpretation and Explanation :

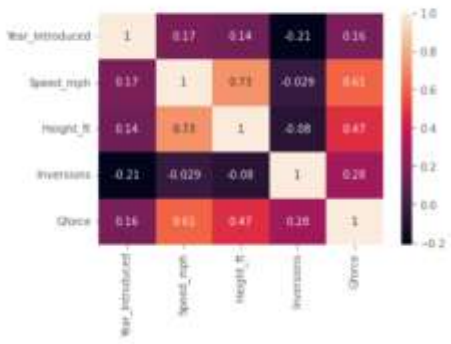
From the scatter plot we can say that these two quantities are highly related to one another as the graph is increasing for each values apart from some points. The scatter plot has high importance because the it can add many features in it for better understanding for example the below plot has the hue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

```
In [34]: ######################################## Scatter Graph with Hue ###############################################################
ax = sns.scatterplot(x='Speed_mph',
                     y='Height_ft',
                     hue='Year_Introduced',
                     data=df)
ax.set_title('Coaster Speed vs. Height')
plt.show()
```



Visualization 4: Heatmap:

```
In [36]: ######################################## HeatMap ##############################################################################
sns.heatmap(df_corr, annot=True)
Out[36]: <AxesSubplot:>
```



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.

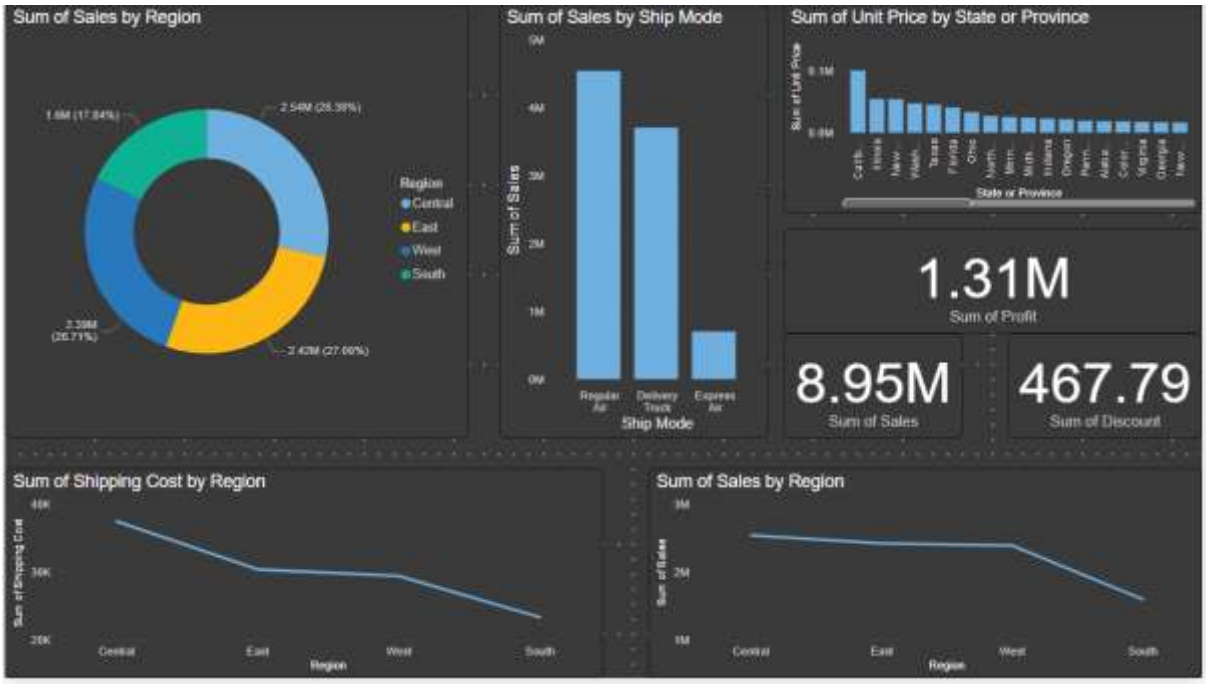
<u>Question 3:</u>

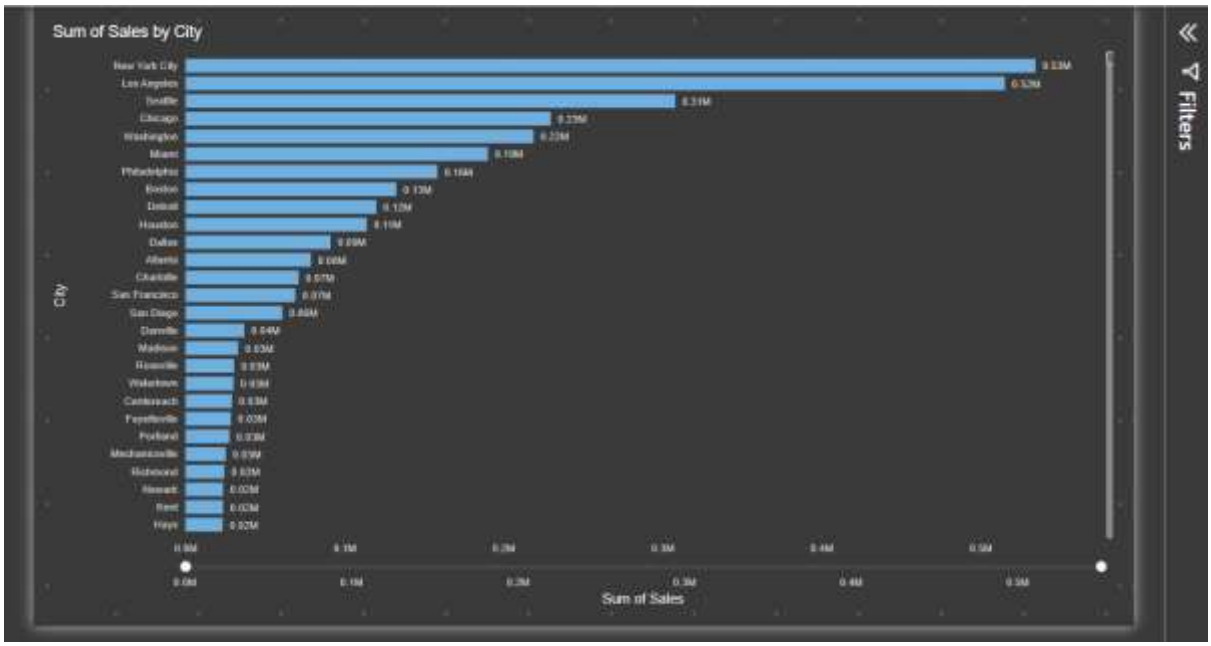
Power BI:-

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

	A	B	C	D	E	F	G	H	I	J
1	Row ID	Order Priority	Discount	Unit Price	Shipping Cost	Customer ID	Customer Name	Ship Mode	Customer Segment	Product Category
9389	18430	High	0.07	40.98	1.99	3393	Irene Murphy	Regular Air	Consumer	Technology
9390	19722	Critical	0.02	5.98	5.35	3393	Irene Murphy	Regular Air	Consumer	Office Supplies
9391	21333	Low	0.1	6.68	6.92	3394	Julia Padgett	Regular Air	Consumer	Office Supplies
9392	23658	Medium	0	12.28	4.86	3394	Julia Padgett	Express Air	Consumer	Office Supplies
9393	22578	Not Specified	0.04	6.84	8.37	3394	Julia Padgett	Regular Air	Consumer	Office Supplies
9394	22579	Not Specified	0.07	125.98	7.69	3394	Julia Padgett	Regular Air	Consumer	Technology
9395	18988	Low	0.02	276.2	24.49	3396	Stuart Wiley	Regular Air	Corporate	Furniture
9396	19370	High	0	140.98	53.48	3396	Stuart Wiley	Delivery Truck	Corporate	Furniture
9397	19371	High	0.01	218.08	18.06	3396	Stuart Wiley	Regular Air	Corporate	Furniture
9398	19372	High	0.09	50.98	6.5	3396	Stuart Wiley	Express Air	Corporate	Technology
9399	18863	Low	0.04	6.48	6.65	3396	Stuart Wiley	Regular Air	Corporate	Office Supplies
9400	18864	Low	0.04	6.48	7.86	3396	Stuart Wiley	Express Air	Corporate	Office Supplies
9401	18330	Not Specified	0.08	6.08	1.17	3396	Stuart Wiley	Regular Air	Small Business	Office Supplies
9402	20624	Low	0	1270.99	19.99	3397	Andrea Shaw	Regular Air	Small Business	Office Supplies
9403	19842	High	0.01	10.9	7.48	3397	Andrea Shaw	Regular Air	Small Business	Office Supplies
9404	19843	High	0.1	7.98	5.03	3397	Andrea Shaw	Regular Air	Small Business	Technology
9405	19813	Low	0.08	35.44	7.5	3397	Andrea Shaw	Regular Air	Small Business	Office Supplies
9406	22193	High	0.05	387.99	19.99	3397	Andrea Shaw	Regular Air	Small Business	Office Supplies
9407	21952	Not Specified	0.05	10.91	2.99	3397	Andrea Shaw	Regular Air	Corporate	Office Supplies
9408	22599	Critical	0.01	177.98	0.99	3398	Marc McDaniel	Regular Air	Corporate	Office Supplies
9409	23074	Low	0.01	155.99	8.99	3398	Marc McDaniel	Regular Air	Corporate	Technology
9410	22194	High	0	180.98	35.02	3398	Marc McDaniel	Delivery Truck	Small Business	Furniture
9411	26208	Not Specified	0.08	11.97	5.81	3399	Marvin Reid	Regular Air	Small Business	Office Supplies
9412	26073	Critical	0.05	59.76	10.29	3399	Marvin Reid	Regular Air	Small Business	Office Supplies
9413	26074	Critical	0.05	4.13	0.5	3399	Marvin Reid	Regular Air	Small Business	Office Supplies
9414	26075	Critical	0.01	30.98	17.08	3399	Marvin Reid	Regular Air	Small Business	Office Supplies
9415	24911	Medium	0.1	9.38	4.93	3400	Florence Gold	Express Air	Small Business	Furniture

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.

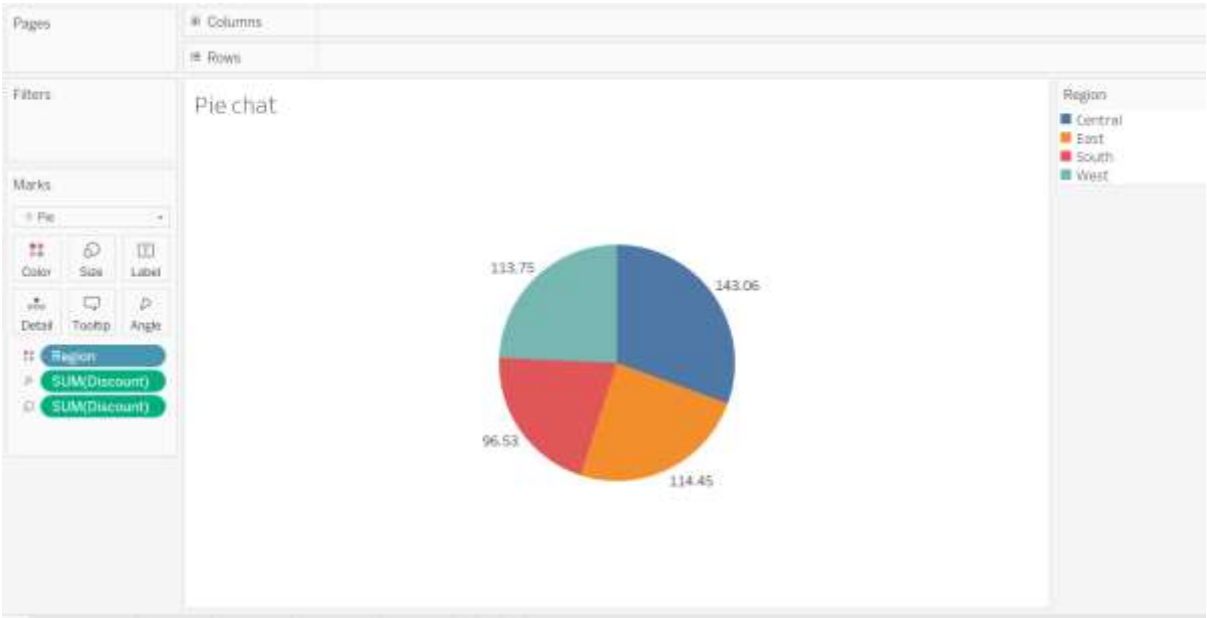
Bar chart : It tells us how the shipping mode effect the purchases.

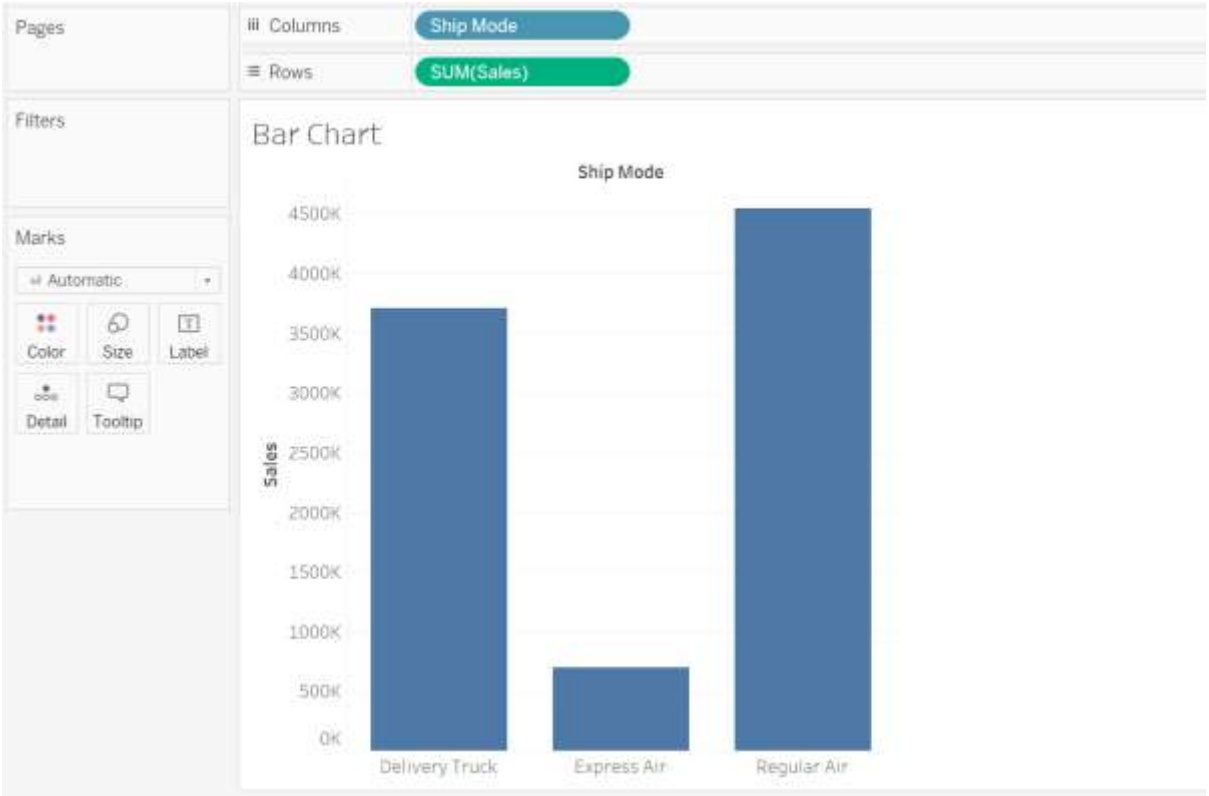
Total Box: It tells us about the total amount of sales, profit and discount

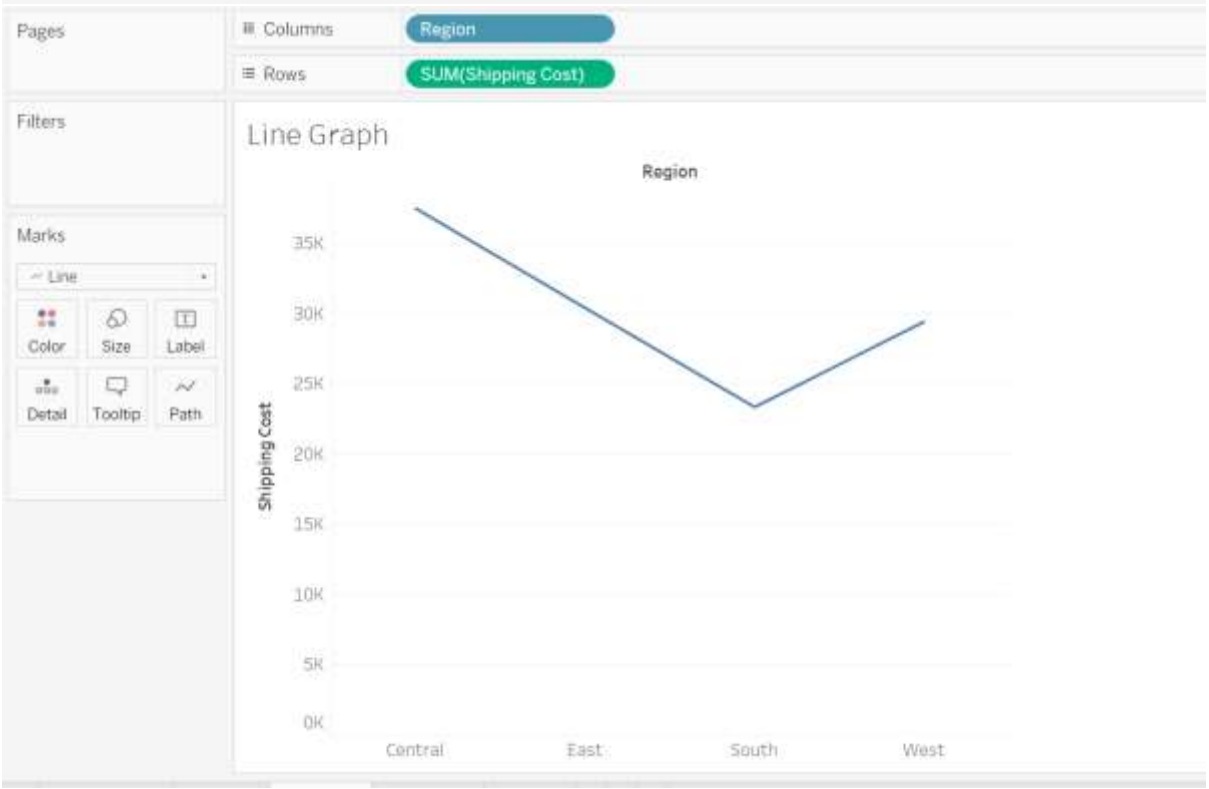
Line plot: Sum of different plots depending on the areas.

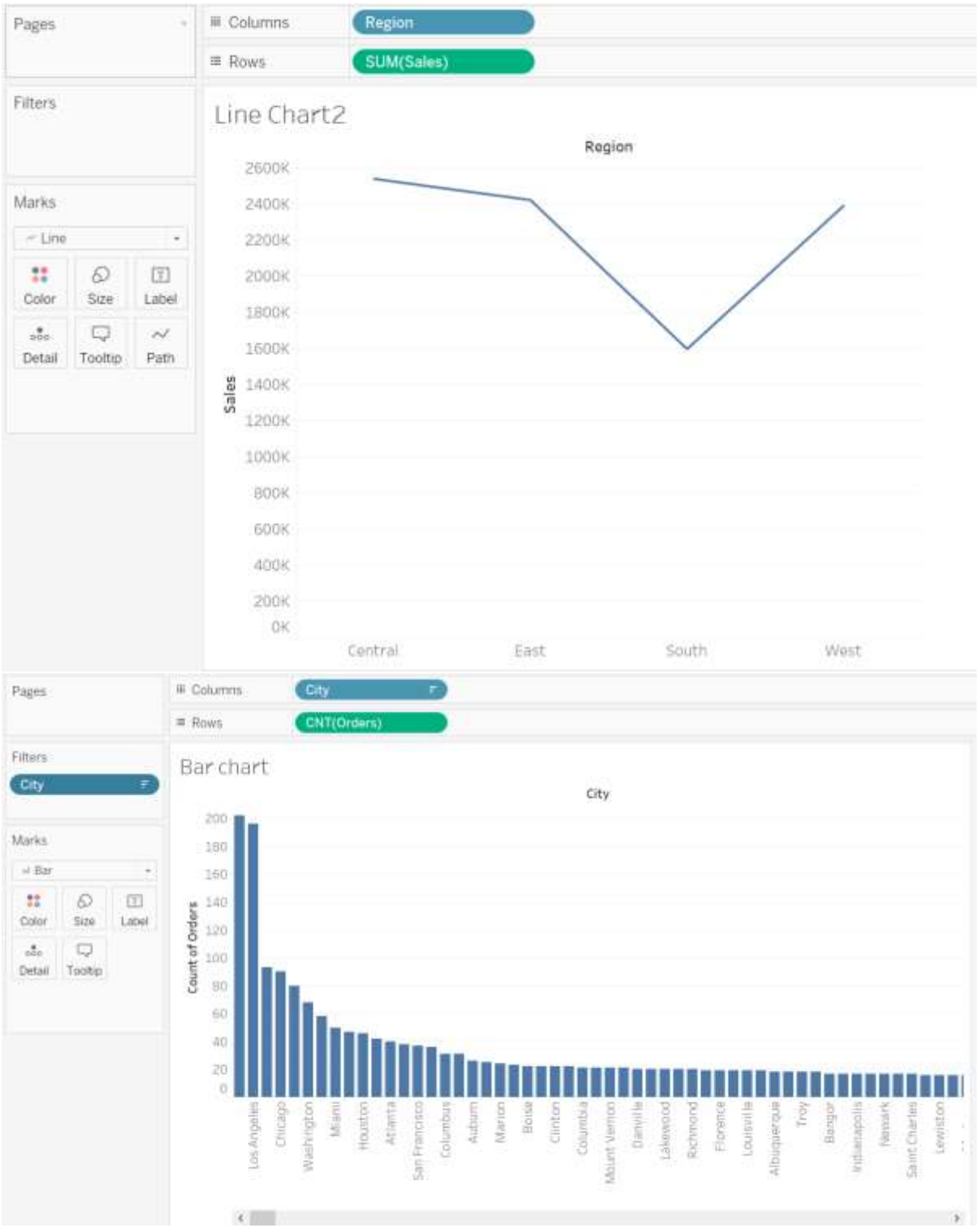
Tableau: -

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




Pages
Columns
Ship Mode
Rows
SUM(Sales)
Filters
Bar Chart
Ship Mode
Marks
Automatic
Color
Size
Label
Detail
Tooltip
4500K
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
Sales
Delivery Truck
Express Air
Regular Air


Pages
Columns
Region
Rows
SUM(Shipping Cost)
Filters
Line Graph
Region
Marks
Line
Color
Size
Label
Detail
Tooltip
Path
35K
30K
25K
20K
15K
10K
5K
0K
Shipping Cost
Central
East
South
West



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.