ANALYSIS OF US AIRLINES PASSENGER SATISFACTION SURVEY

Sanchita Ajay Gawand

MS: Information Systems

California State University Los Angeles

CIS 5250:Visual Analytics

Submitted to: Dr. Shilpa Balan

Table of Contents

INTRODUCTION

The airline industry is one that passengers love to hate. A lot of factors are taken into consideration that range from flight schedules, comfort, pricing, in-flight entertainment, food and beverages and customer support. The ACSI has over two decades of experience analyzing passenger satisfaction with the largest airlines operating in the U.S. consumer market. Each year, the ACSI interviews hundreds of passengers about their recent flight experiences, including both business and leisure travelers. The customer survey data serve as inputs to ACSI's proprietary model, which embeds customer satisfaction within a series of cause-and-effect relationships.And yet, overall customer satisfaction hit an all-time high in 2017 and declined only slightly in 2018, according to the latest American Customer Satisfaction Index Travel Report, which was released in late April.

In the report, Southwest Airlines, Alaska Airlines and JetBlue received the highest marks from customers, while budget carriers Spirit Airlines and Frontier Airlines were at the bottom. Legacy carriers, including Delta Air Lines, American Airlines and United Airlines, were in the middle. The report also shows that customer satisfaction in the hotel industry remains stable, while marks for online travel services have slightly declined.

DATASET DESCRIPTION

Dataset URL: https://www.kaggle.com/johnddddd/customer-satisfaction#satisfaction_2015.xlsx

Dataset is about the survey taken about the services offered by the US Airline. It shows the passenger satisfaction. It was taken from kaggle which is an open source website for dataset. Dataset comprises of 23 columns and more than 3000 rows. The responses are on the rate of 1 to 5. It uses the Likert scale to analyze the data.

1 means not at all satisfied

2 means slightly satisfied

3 means neutral

4 means very satisfied

5 means extremely satisfied

Below are the columns in dataset:

1. Year: Data is for three years ie form 2015 to 2017.
2. Month: This column shows survey recorded by month.
3. Satisfaction: Airline satisfaction level denotes the passengers overall satisfaction level ie Satisfied, neutral or dissatisfied.
4. Age: This column shows the actual age of the passengers
5. Gender: This column shows the Gender of the passengers (Female, Male)
6. Type of Travel:This column shows the Purpose of the flight of the passengers (Personal Travel, Business Travel)

7. Class: This column shows the Travel class in the plane of the passengers (Business, Economy , Economy Plus)
8. Customer Type: This column shows the customer type (Loyal customer, disloyal customer)
9. Flight distance:This column shows the The flight distance of the journey
10. Inflight wifi service: This column shows the Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)"
11. Ease of Online booking: This column shows the Satisfaction level of online booking
12. Inflight service: This column shows the Satisfaction level of inflight service
13. Online boarding:This column shows the Satisfaction level of online boarding
14. Inflight entertainment: This column shows the Satisfaction level of inflight entertainment
15. Food and drink: This column shows the Satisfaction level of Food and drink
16. Seat comfort: This column shows the Satisfaction level of Seat comfort
17. On-board service: This column shows the Satisfaction level of On-board service
18. Leg room service: This column shows the Satisfaction level of Leg room service
19. Departure/Arrival time convenient: This column shows the Satisfaction level of Departure/Arrival time convenient
20. Baggage handling: This column shows the Satisfaction level of baggage handling
21. Gate location: This column shows the Satisfaction level of Gate location
22. Cleanliness: This column shows the Satisfaction level of Cleanliness

23. Check-in service: This column shows the satisfaction level of Check-in service
24. Departure Delay in Minutes: This column shows the Minutes delayed when departure
25. Arrival Delay in Minutes: This column shows the minutes delayed when Arrival
26. Flight cancelled: This column shows whether the flight is cancelled or not (Yes, No)
27. Flight time in minutes: This column shows the Minutes flight takes

Datacleaning:

Type 1: Removing NaN rows

Dataset had missing lot of missing values. It was cleaned by using dropna(). Removing the NaN data will give the better quality of data.

Before

```
In [4]: runfile('/Users/sgawand/Desktop/Tes
        Day  Month  Year      Customer Type
0         5      2  2015  disloyal Customer
1         5      2  2015  disloyal Customer
2         5      2  2015  disloyal Customer
3         5      2  2015  disloyal Customer
4         5      2  2015  disloyal Customer
5         5      2  2015  disloyal Customer
6         5      2  2015                NaN
7         5      2  2015                NaN
8         5      2  2015                NaN
9         5      2  2015  disloyal Customer
10       12      1  2015  disloyal Customer
```

After:

```
        Day  Month  Year      Customer Type
0         5      2  2015  disloyal Customer
1         5      2  2015  disloyal Customer
2         5      2  2015  disloyal Customer
3         5      2  2015  disloyal Customer
4         5      2  2015  disloyal Customer
5         5      2  2015  disloyal Customer
9         5      2  2015  disloyal Customer
10       12      1  2015  disloyal Customer
11       12      1  2015  disloyal Customer
```

Code:

```
import pandas as pd

# Airlines=pd.read_csv('Airlines.csv', low_memory=False)

Airlines=pd.read_csv("Airlines.csv")
```

#print(Airlines.head(5))

#print(Airlines[['Day','Month','Year','Customer Type']])

#df_no_missing = Airlines.dropna()

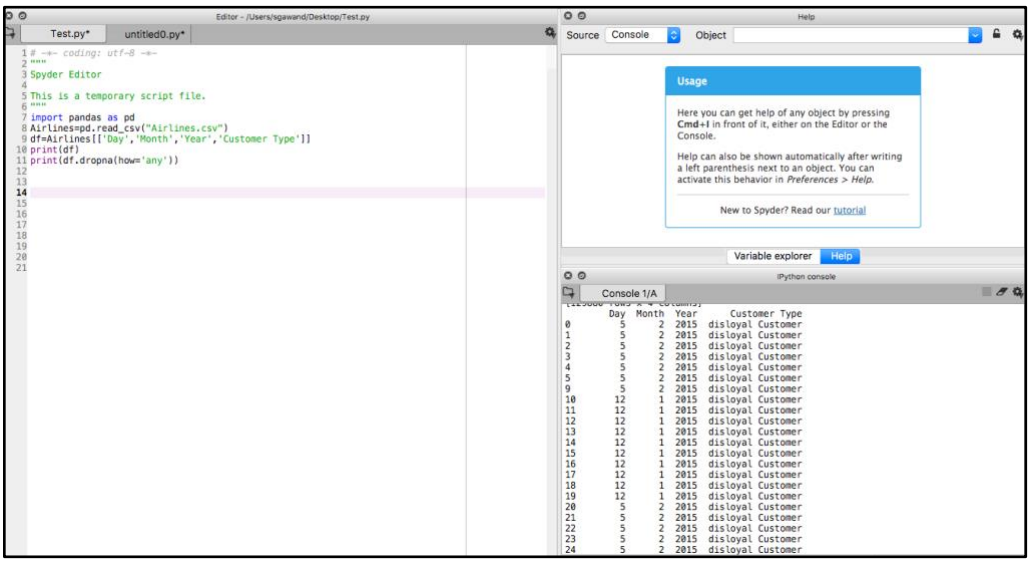
#print(Airlines[['Day','Month','Year','Customer Type']])

df=Airlines[['Day','Month','Year','Customer Type']]

print(df)

print(df.dropna(how='any'))

```
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
df=Airlines[['Day','Month','Year','Customer Type']]
print(df)
print(df.dropna(how='any'))
```



Type 2: Renaming Columns

Dataset had misspelled column names. Gender was misspelled as Gender, Year was misspelled as Yar. This data was cleaned by using .rename()

Before:

Yar	Customer Type	Satisfaction	Gandar
2015	disloyal Customer	satisfied	Male
2015	disloyal Customer	satisfied	Male
2015	disloyal Customer	satisfied	Male
2015	disloyal Customer	satisfied	Female
2015	disloyal Customer	satisfied	Male

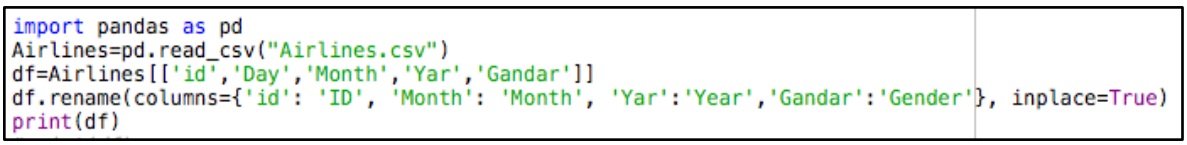
After:

	ID	Day	Month	Year	Gender
0	117135	5	2	2015	Male
1	72091	5	2	2015	Male
2	29663	5	2	2015	Male
3	81849	5	2	2015	Female

Code:

```
import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

df=Airlines[['id','Day','Month','Yar','Gandar']]

df.rename(columns={'id': 'ID', 'Month': 'Month', 'Yar':'Year','Gandar':'Gender'},

inplace=True)

print(df)
```

Code Screenshot:



Type 3: Combining Columns

Category: Concat function

Dataset had different columns of day,month and year. They were merged by using concat function.

Before:

id	Day	Month	Year	Customer Type
117135	5	2	2015	disloyal Customer
72091	5	2	2015	disloyal Customer
29663	5	2	2015	disloyal Customer
81849	5	2	2015	disloyal Customer
83693	5	2	2015	disloyal Customer
129488	5	2	2015	disloyal Customer

After:

	ID	Day	Month	Year	Gender	Month_Year_concat
0	117135	5	2	2015	Male	2 / 2015
1	72091	5	2	2015	Male	2 / 2015
2	29663	5	2	2015	Male	2 / 2015
3	81849	5	2	2015	Female	2 / 2015
4	83693	5	2	2015	Male	2 / 2015
5	129488	5	2	2015	Female	2 / 2015
6	69507	5	2	2015	Male	2 / 2015
7	126309	5	2	2015	Female	2 / 2015

Code:

```
import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

df=Airlines[['id','Day','Month','Yar','Gandar']]

df.rename(columns={'id': 'ID', 'Month': 'Month', 'Yar':'Year','Gandar':'Gender'},
inplace=True)
```

df['Month_Year_concat'] = df["Month"].map(str)+ " / " + df["Year"].map(str)

print(df)

Code screenshot:

```
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
df=Airlines[['id','Day','Month','Yar','Gandar']]
df.rename(columns={'id': 'ID', 'Month': 'Month', 'Yar':'Year','Gandar':'Gender'}, inplace=True)
df['Month_Year_concat'] = df["Month"].map(str)+ " / " + df["Year"].map(str)
print(df)
```

SUMMARY STATISTICS

Show min, max, mean, standard deviation of age of all passengers?

Code:

import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

print('Minimum Value:',Airlines['Age'].min())

print('Maximum Value:',Airlines['Age'].max())

print('Mean Value:',Airlines['Age'].mean())

print('Standard Deviation Value:',Airlines['Age'].std())

Code Screenshot:

```
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
print('Minimum Value:',Airlines['Age'].min())
print('Maximum Value:',Airlines['Age'].max())
print('Mean Value:',Airlines['Age'].mean())
print('Standard Deviation Value:',Airlines['Age'].std())
```

Output:

```
In [129]: runfile('/Users/sgawand/Desktop/Q.5.py', wdir='/Users/sgawand/Desktop')
Minimum Value: 7
Maximum Value: 85
Mean Value: 39.42795657530028
Standard Deviation Value: 15.119359950371692
count    129880.000000
mean         39.427957
std          15.119360
min           7.000000
25%          27.000000
50%          40.000000
75%          51.000000
max          85.000000
Name: Age, dtype: float64
```

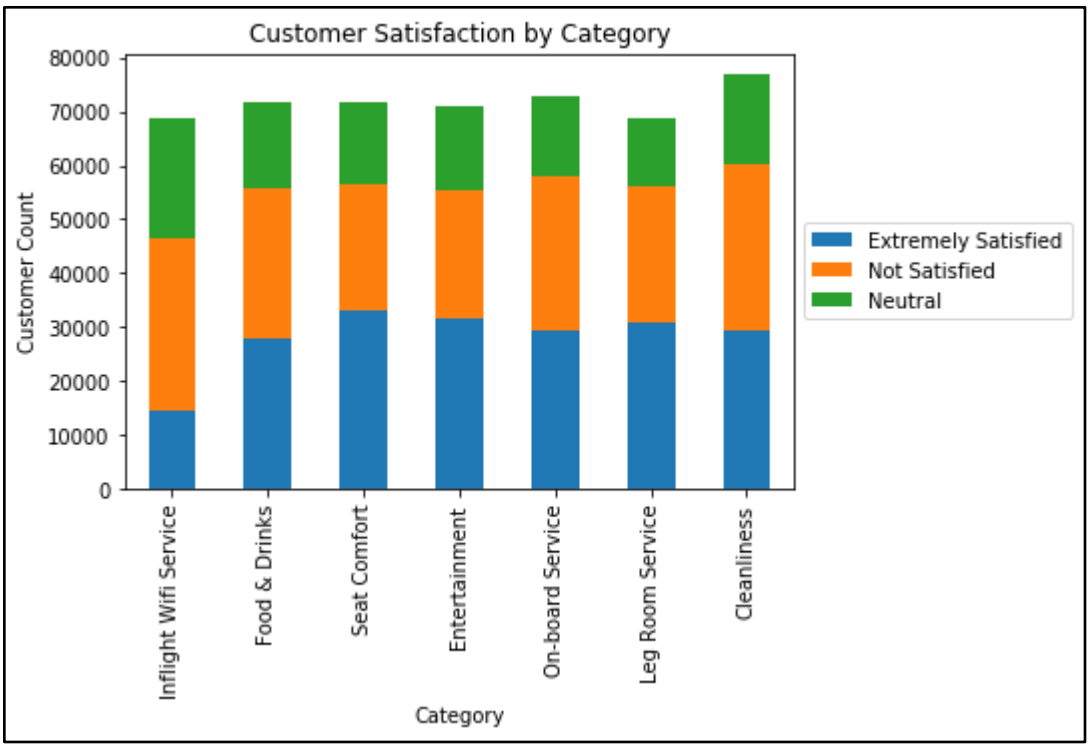
Description:

As shown, Minimum average age of the passengers travelling is 7 years, maximum is 85 years, Mean value of age is 39.42 ie around 40. Standard Deviation is 15.11. Looking at the statistics, we can say that almost passengers with all age groups opt for US Airlines.

Analysis & Visualizations

Q.1 Show the satisfied, neutral and not satisfied customers for inflight wifi service, food and drinks, seat comfort, entertainment, on-board service, leg room services, cleanliness?

Visual:



Category Used: List, Dictionary, Pandas, Dataframe, matplot, value_count() function

Chart Type: Stacked Bar

The above visualization shows Passenger Satisfaction by category. X Axis represent categories such as Inflight Wifi Service, Food & Drink, Seat Comfort, Entertainment, On-Board Service, Leg Room Service and Cleanliness. Y Axis represents the Passenger Count. Colors used represent the satisfaction level for each category. Blue

represents extremely satisfied, orange represents not satisfied and green represents neutral response. As visualized, it can be said that large count of passengers are extremely satisfied by seat comfort , not satisfied with cleanliness and inflight wifi service.

Code:

```
import numpy as np

import matplotlib.pyplot as plt

from matplotlib import rc

import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

#to see wifi counts

Cnt_wifi=Airlines['Inflight wifi service'].value_counts()

print(Cnt_wifi)

#to see food & drink count

Cnt_food=Airlines['Food and drink'].value_counts()

print(Cnt_food)

#to see seat confort count

Cnt_seat=Airlines['Seat comfort'].value_counts()

print(Cnt_seat)
```

```
#to see inflight entertainmanet count

Cnt_entertainment=Airlines['Inflight entertainment'].value_counts()

print(Cnt_entertainment)

#to see onboard service count

Cnt_onbrdservice=Airlines['On-board service'].value_counts()

print(Cnt_onbrdservice)

#to see legroomservice count

Cnt_leg=Airlines['Leg room service'].value_counts()

print(Cnt_leg)

#to see cleaniliness count

Cnt_clean=Airlines['Cleanliness'].value_counts()

print(Cnt_clean)

#to create list

#scale=['Extremely Satisfied','Not Satisfied','Neutral']

category=['Wifi','Food','Seat','Entertainment','On-board','Leg']

Extremely_Satisfied=[14356,27957,33158,31544,29492,30905,29492]

Not_Satisfied=[32185,27794,23328,23884,28542,25056,30639]

neutral=[22328,16051,15108,15675,14787,12895,16729]
```

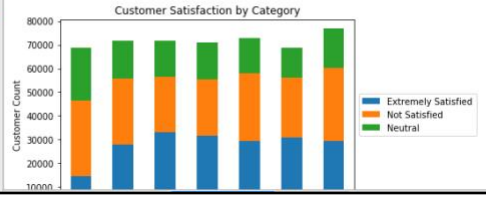
```
#to create dictionary

Sat_dict={'Extremely Satisfied':Extremely_Satisfied,'Not
Satisfied':Not_Satisfied,'Neutral':Neutral}

#to create dataframe

Sat_df=pd.DataFrame(Sat_dict)

#row labels

row_labels=['Inflight Wifi Service','Food & Drinks','Seat Comfort','Entertainment','On-
board Service','Leg Room Service','Cleanliness']

Sat_df.index=row_labels

print(Sat_df)

my_plot = Sat_df.plot(kind='bar',stacked=True,title="Customer Satisfaction by
Category")

my_plot.set_xlabel("Category")

my_plot.set_ylabel("Customer Count")

my_plot.legend(loc='center left', bbox_to_anchor=(1, 0.5))

plt.show()
```

Code Screenshot:

```
@author: sgawand
"""
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rc
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
#to see wifi counts
Cnt_wifi=Airlines['Inflight wifi service'].value_counts()
print(Cnt_wifi)
#to see food & drink count
Cnt_food=Airlines['Food and drink'].value_counts()
print(Cnt_food)
#to see seat confort count
Cnt_seat=Airlines['Seat comfort'].value_counts()
print(Cnt_seat)
#to see inflight entertainmanet count
Cnt_entertainment=Airlines['Inflight entertainment'].value_counts()
print(Cnt_entertainment)
#to see onboard service count
Cnt_onbrdservice=Airlines['On-board service'].value_counts()
print(Cnt_onbrdservice)
#to see legroomservice count
Cnt_leg=Airlines['Leg room service'].value_counts()
print(Cnt_leg)
#to see cleaniliness count
Cnt_clean=Airlines['Cleanliness'].value_counts()
print(Cnt_clean)
#to create list
#scale=['Extremely Satisfied','Not Satisfied','Neutral']
category=['Wifi','Food','Seat','Entertainment','On-board','Leg']
Extremely_Satisfied=[14356,27957,33158,31544,29492,30905,29492]
Not_Satisfied=[32185,27794,23328,23884,28542,25056,30639]
neutral=[22328,16051,15108,15675,14787,12895,16729]
#to create dictionary
Sat_dict={'Extremely Satisfied':Extremely_Satisfied,'Not Satisfied':Not_Satisfied,'Neutral':Neutral}
#to create dataframe
Sat_df=pd.DataFrame(Sat_dict)
#row labels
row_labels=['Inflight Wifi Service','Food & Drinks','Seat Comfort','Entertainment','On-board Service','Leg Roo
Sat_df.index=row_labels
print(Sat_df)
my_plot = Sat_df.plot(kind='bar',stacked=True,title="Customer Satisfaction by Category")
my_plot.set_xlabel("Category")
my_plot.set_ylabel("Customer Count")
my_plot.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```

```python
@author: sgawand
"""
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rc
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
#to see wifi counts
Cnt_wifi=Airlines['Inflight wifi service'].value_counts()
print(Cnt_wifi)
#to see food & drink count
Cnt_food=Airlines['Food and drink'].value_counts()
print(Cnt_food)
#to see seat comfort count
Cnt_seat=Airlines['Seat comfort'].value_counts()
print(Cnt_seat)
#to see inflight entertainmanet count
Cnt_entertainment=Airlines['Inflight entertainment'].value_counts()
print(Cnt_entertainment)
#to see onboard service count
Cnt_onbrdservice=Airlines['On-board service'].value_counts()
print(Cnt_onbrdservice)
#to see legroomservice count
Cnt_leg=Airlines['Leg room service'].value_counts()
print(Cnt_leg)
#to see cleaniliness count
Cnt_clean=Airlines['Cleanliness'].value_counts()
print(Cnt_clean)
#to create list
#scale=['Extremely Satisfied','Not Satisfied','Neutral']
category=['Wifi','Food','Seat','Entertainment','On-board','Leg']
Extremely_Satisfied=[14356,27957,33158,31544,29492,30905,29492]
Not_Satisfied=[32185,27794,23328,23884,28542,25056,30639]
neutral=[22328,16051,15108,15675,14787,12895,16729]
#to create dictionary
Sat_dict={'Extremely Satisfied':Extremely_Satisfied,'Not Satisfied':Not_Satisfied,'Neutral':Neutral}
#to create dataframe
Sat_df=pd.DataFrame(Sat_dict)
#row labels
row_labels=['Inflight Wifi Service','Food & Drinks','Seat Comfort','Entertainment','On-board Service','Leg Room
Sat_df.index=row_labels
print(Sat_df)
my_plot = Sat_df.plot(kind='bar',stacked=True,title="Customer Satisfaction by Category")
my_plot.set_xlabel("Category")
my_plot.set_ylabel("Customer Count")
my_plot.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```

```
3    25056
2    24540
1    12895
0      598
Name: Leg room service, dtype: int64
4    33969
3    30639
5    28416
2    20113
1    16729
0       14
Name: Cleanliness, dtype: int64
                      Extremely Satisfied  Not Satisfied  Neutral
Inflight Wifi Service                14356          32185    22328
Food & Drinks                        27957          27794    16051
Seat Comfort                         33158          23328    15108
Entertainment                        31544          23884    15675
On-board Service                     29492          28542    14787
Leg Room Service                     30905          25056    12895
Cleanliness                          29492          30639    16729
```

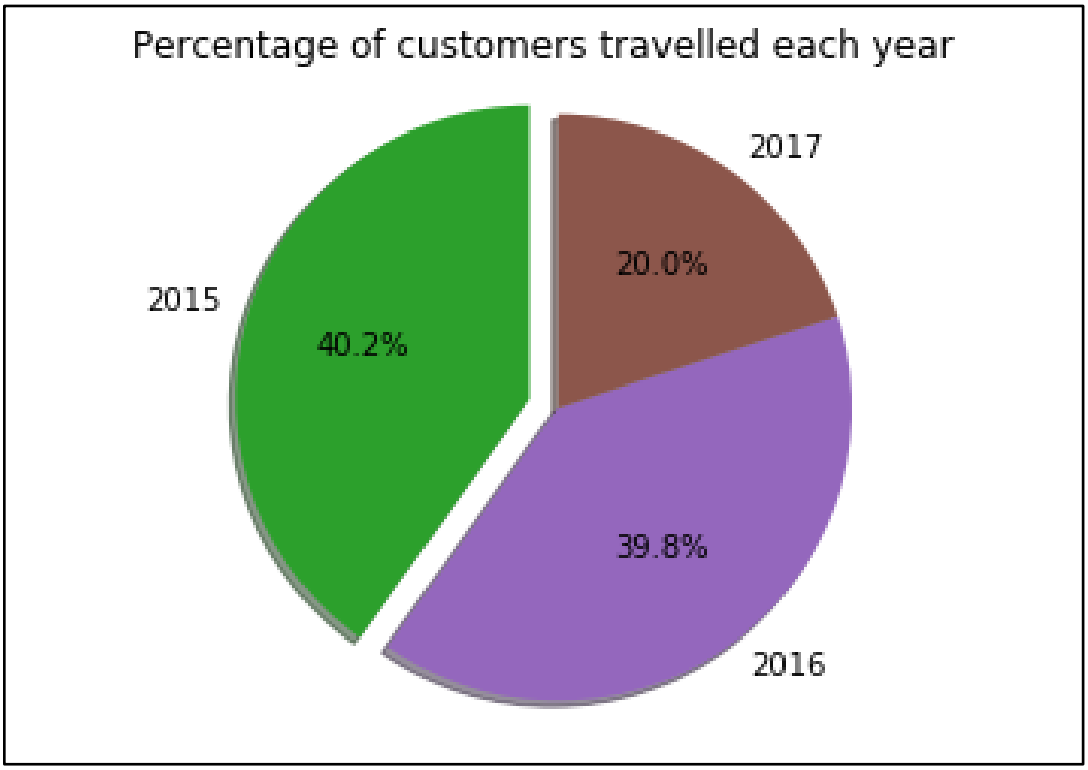


Q.2. Show the percentage of customers travelled each year?

Category Used: Matplot, Tuples

Chart Type: Pie Chart

Visual:



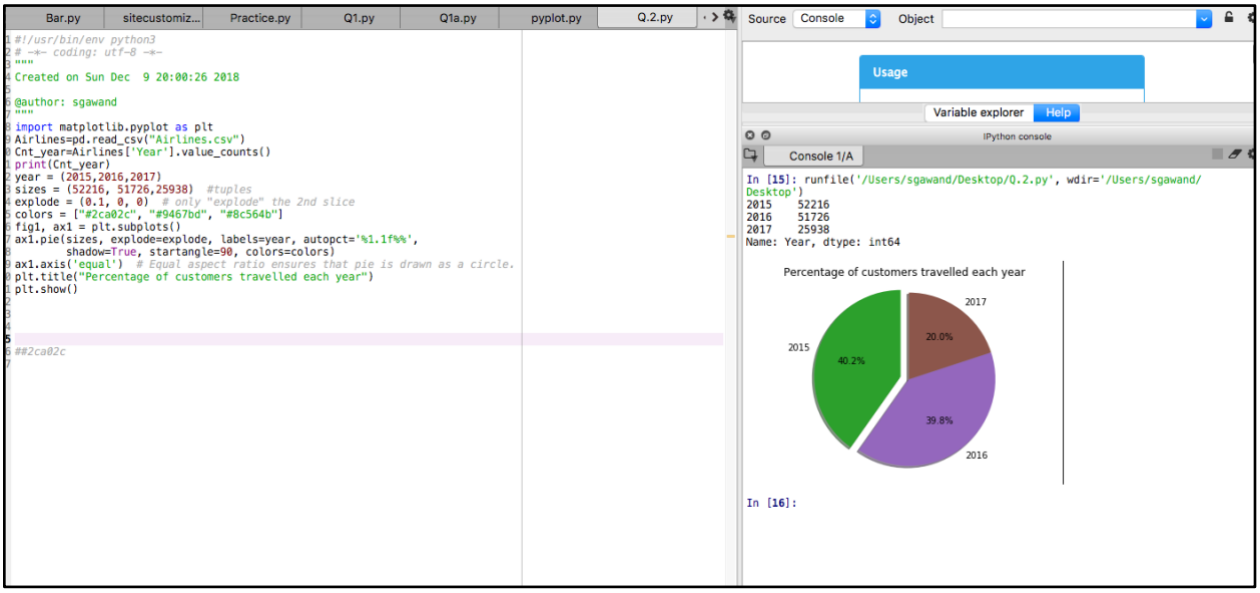
Above visualization shows the percentage of passengers travelled each year ie for year 2015, 2016 and 2017. Maximum percentage of passengers ie 40.2% have travelled in year 2015, followed by 2016. Least are travelled in 2017 ie only 20%. Green color represent year 2015. Purple color represent year 2016 and brown color represents year 2017. It can be interpreted that the percentage of passengers is decreasing each year.

Code:

```
import matplotlib.pyplot as plt

Airlines=pd.read_csv("Airlines.csv")

Cnt_year=Airlines['Year'].value_counts()

print(Cnt_year)

year = (2015,2016,2017)

sizes = (52216, 51726,25938)  #tuples

explode = (0.1, 0, 0)  # only "explode" the 2nd slice

colors = ["#2ca02c", "#9467bd", "#8c564b"]

fig1, ax1 = plt.subplots()

ax1.pie(sizes, explode=explode, labels=year, autopct='%1.1f%%',

        shadow=True, startangle=90, colors=colors)

ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

plt.title("Percentage of customers travelled each year")

plt.show()
```

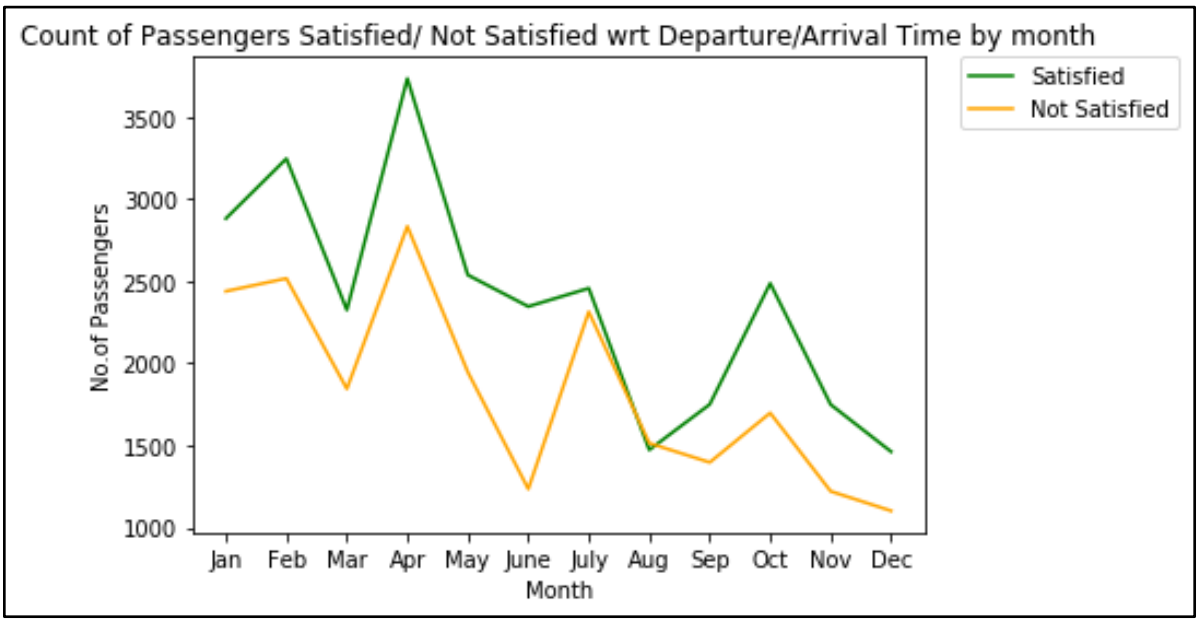
Code Screenshot:

```
@author: sgawand
"""
import matplotlib.pyplot as plt
Airlines=pd.read_csv("Airlines.csv")
Cnt_year=Airlines['Year'].value_counts()
print(Cnt_year)
year = (2015,2016,2017)
sizes = (52216, 51726,25938)  #tuples
explode = (0.1, 0, 0)  # only "explode" the 2nd slice
colors = ["#2ca02c", "#9467bd", "#8c564b"]
fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=year, autopct='%1.1f%%',
        shadow=True, startangle=90, colors=colors)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.title("Percentage of customers travelled each year")
plt.show()
```



Q.3. Show the passengers satisfied/not satisfied wrt Departure/Arrival time for each month?

Visual:



Category Used: List,Dictionary, Dataframe, matplot, numpy, Groupby, Aggregate function

Char Type: Line

Above visualization shows the count of passengers satisfied/not satisfied wrt Departure/Arrival time for each month**.** X axis represent month ie from January to December. Y axis represent no.of passengers who are satisfied and not satisfied with the departure and arrival time of flights. Green line indicates the count of satisfied passengers over the month and the yellow line indicates the count of passengers not satisfied over the months. As shown in visual, majority of passengers travelling in month of April are unsatisfied with the departure/arrival time followed by July. Graph of the not satisfied customers remains low for all months except slight decrease in August.

Code:

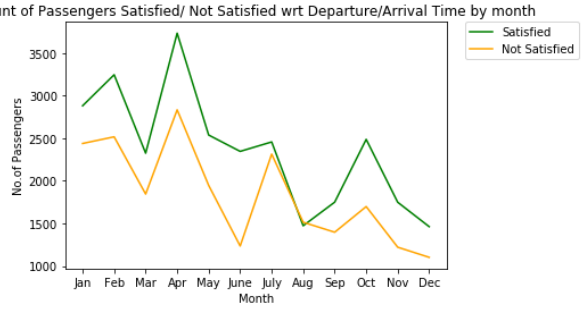
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
groupby_gender_and_answer = Airlines.groupby(['Month','Departure/Arrival time
convenient'])
results = groupby_gender_and_answer['Departure/Arrival time
convenient'].aggregate('count')
print(results.value_counts)
Extremely_Satisfied=[2881,3245,2324,3733,2537,2345,2456,1472,1749,2486,1748,146
1]
Not_Satisfied=[2438,2516,1844,2833,1946,1234,2313,1513,1396,1697,1220,1101]
Dict={'Extremely Satisfied':Extremely_Satisfied,'Not Satisfied':Not_Satisfied}
df=pd.DataFrame(Dict)
#row labels
Month=['Jan','Feb','Mar','Apr','May','June','July','Aug','Sep','Oct','Nov','Dec']
df.index=row_labels
print(df)
plt.plot(Month, Extremely_Satisfied, color='g',label="Satisfied")
plt.plot(Month, Not_Satisfied, color='orange', label="Not Satisfied")
plt.xlabel('Month')
plt.ylabel('No.of Passengers')
```

plt.title('Count of Passengers Satisfied/ Not Satisfied wrt Departure/Arrival Time by month')

plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)

plt.show()

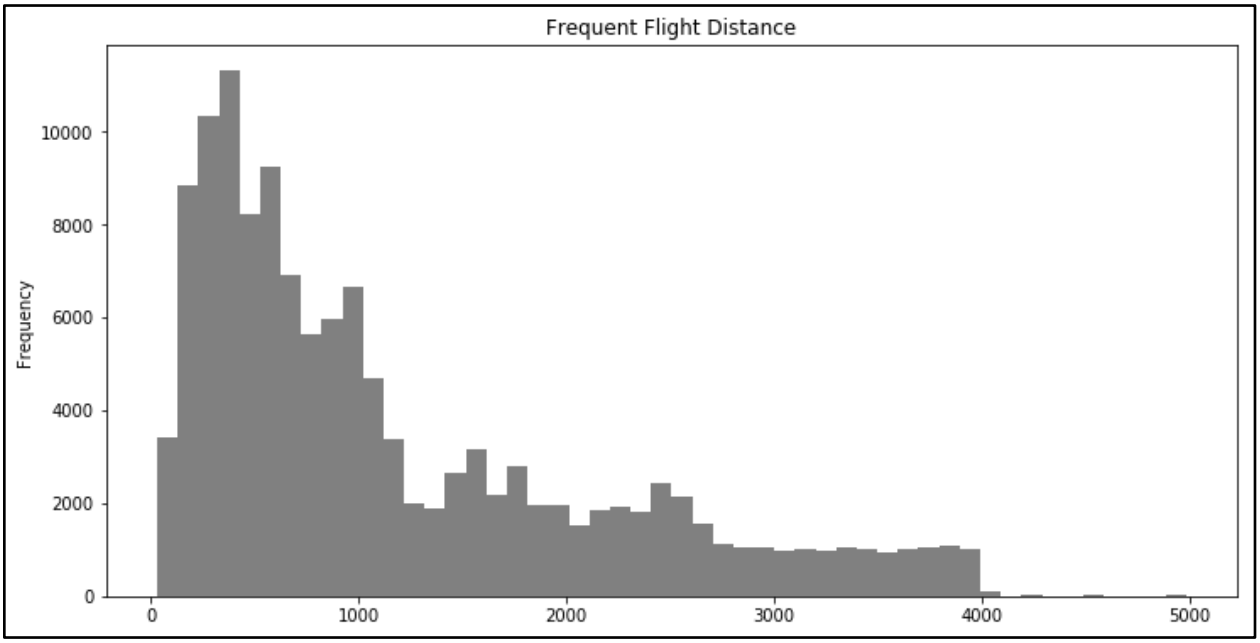
```
      5                                    1461
Name: Departure/Arrival time convenient, Length: 72, dtype: int64>
      Extremely Satisfied  Not Satisfied
Jan                  2881           2438
Feb                  3245           2516
Mar                  2324           1844
Apr                  3733           2833
May                  2537           1946
June                 2345           1234
July                 2456           2313
Aug                  1472           1513
Sep                  1749           1396
Oct                  2486           1697
Nov                  1748           1220
Dec                  1461           1101
```




```
@author: sgawand
"""

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
groupby_gender_and_answer = Airlines.groupby(['Month','Departure/Arrival time convenient'])
results = groupby_gender_and_answer['Departure/Arrival time convenient'].aggregate('count')
print(results.value_counts)
Extremely_Satisfied=[2881,3245,2324,3733,2537,2345,2456,1472,1749,2486,1748,1461]
Not_Satisfied=[2438,2516,1844,2833,1946,1234,2313,1513,1396,1697,1220,1101]
Dict={'Extremely Satisfied':Extremely_Satisfied,'Not Satisfied':Not_Satisfied}
df=pd.DataFrame(Dict)
#row labels
Month=['Jan','Feb','Mar','Apr','May','June','July','Aug','Sep','Oct','Nov','Dec']
df.index=row_labels
print(df)
plt.plot(Month, Extremely_Satisfied, color='g',label="Satisfied")
plt.plot(Month, Not_Satisfied, color='orange', label="Not Satisfied")
plt.xlabel('Month')
plt.ylabel('No.of Passengers')
plt.title('Count of Passengers Satisfied/ Not Satisfied wrt Departure/Arrival Time by month' )
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.show()
```

Q.4. Show the frequent flight distance covered?

Visual:



Category Used: Numpy, pandas

Chart Type: Histogram.

Above visualization shows the frequent flight distance covered. As shown in the figure maximum distance is covered between 0 to 1000 miles. Minimum is between 3000 to 4000 miles. Flight distance is constant between 0 to 1000. Histogram is used for above visualization.

Code:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
```

Airlines['Flight Distance'].plot(kind='hist',bins=50,figsize=(12,6), color='grey',

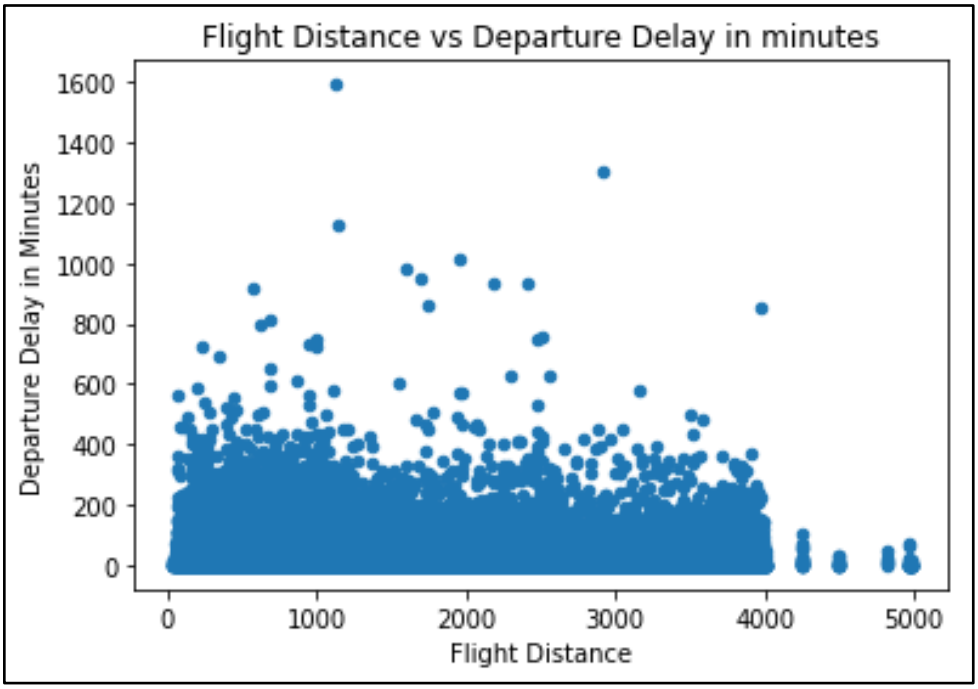
title='Frequent Flight Distance')

plt.show()

Code Screenshot:

Q.5. Show the flight distance vs Departure Delay in minutes?

Visual:



Category Used: Numpy, pandas, matplot

Chart Type: Scatter

Above visualization shows the departure delay in minutes for flight distance. X axis represents the flight distance and y axis represents the departure delay in minutes. As shown, it can be seen that the for the flight distance upto 4000 miles there have been maximum delays.

Code:

```
import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

Airlines=pd.read_csv("Airlines.csv")
```

Airlines.plot(kind='scatter',x='Flight Distance',y='Departure Delay in Minutes', title='Flight Distance vs Departure Delay in minutes')

plt.show()

Code Screenshot:

```
@author: sgawand
"""

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
Airlines.plot(kind='scatter',x='Flight Distance',y='Departure Delay in Minutes')
plt.show()
```

CODES:

Data Cleaning

Type 1: Removing NaN values

```
import pandas as pd

# Airlines=pd.read_csv('Airlines.csv', low_memory=False)

Airlines=pd.read_csv("Airlines.csv")

#print(Airlines.head(5))

#print(Airlines[['Day','Month','Year','Customer Type']])

#df_no_missing = Airlines.dropna()

#print(Airlines[['Day','Month','Year','Customer Type']])

df=Airlines[['Day','Month','Year','Customer Type']]

print(df)

print(df.dropna(how='any'))
```

Type 2: Renaming Columns

```
import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

df=Airlines[['id','Day','Month','Yar','Gandar']]

df.rename(columns={'id': 'ID', 'Month': 'Month', 'Yar':'Year','Gandar':'Gender'},

inplace=True)
```

```
print(df)
```

Type 3: Combining Columns

```
import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

df=Airlines[['id','Day','Month','Yar','Gandar']]

df.rename(columns={'id': 'ID', 'Month': 'Month', 'Yar':'Year','Gandar':'Gender'},
inplace=True)

df['Month_Year_concat'] = df["Month"].map(str)+ " / " + df["Year"].map(str)

print(df)
```

Summary Statistics

```
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
print('Minimum Value:',Airlines['Age'].min())
print('Maximum Value:',Airlines['Age'].max())
print('Mean Value:',Airlines['Age'].mean())
print('Standard Deviation Value:',Airlines['Age'].std())
```

Analysis and Visuals

Q.1 Show the satisfied, neutral and not satisfied customers for inflight wifi service, food and drinks, seat comfort, entertainment, on-board service, leg room services, cleanliness?

Code:

```
import numpy as np

import matplotlib.pyplot as plt

from matplotlib import rc

import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

#to see wifi counts

Cnt_wifi=Airlines['Inflight wifi service'].value_counts()

print(Cnt_wifi)

#to see food & drink count

Cnt_food=Airlines['Food and drink'].value_counts()

print(Cnt_food)

#to see seat confort count

Cnt_seat=Airlines['Seat comfort'].value_counts()

print(Cnt_seat)

#to see inflight entertainmanet count

Cnt_entertainment=Airlines['Inflight entertainment'].value_counts()

print(Cnt_entertainment)
```

```
#to see onboard service count

Cnt_onbrdservice=Airlines['On-board service'].value_counts()

print(Cnt_onbrdservice)

#to see legroomservice count

Cnt_leg=Airlines['Leg room service'].value_counts()

print(Cnt_leg)

#to see cleaniliness count

Cnt_clean=Airlines['Cleanliness'].value_counts()

print(Cnt_clean)

#to create list

#scale=['Extremely Satisfied','Not Satisfied','Neutral']

category=['Wifi','Food','Seat','Entertainment','On-board','Leg']

Extremely_Satisfied=[14356,27957,33158,31544,29492,30905,29492]

Not_Satisfied=[32185,27794,23328,23884,28542,25056,30639]

neutral=[22328,16051,15108,15675,14787,12895,16729]

#to create dictionary

Sat_dict={'Extremely Satisfied':Extremely_Satisfied,'Not
Satisfied':Not_Satisfied,'Neutral':Neutral}

#to create dataframe
```

```
Sat_df=pd.DataFrame(Sat_dict)

#row labels

row_labels=['Inflight Wifi Service','Food & Drinks','Seat Comfort','Entertainment','On-board Service','Leg Room Service','Cleanliness']

Sat_df.index=row_labels

print(Sat_df)

my_plot = Sat_df.plot(kind='bar',stacked=True,title="Customer Satisfaction by Category")

my_plot.set_xlabel("Category")

my_plot.set_ylabel("Customer Count")

my_plot.legend(loc='center left', bbox_to_anchor=(1, 0.5))

plt.show()
```

Q.2. Show the percentage of customers travelled each year?

```
import matplotlib.pyplot as plt

Airlines=pd.read_csv("Airlines.csv")

Cnt_year=Airlines['Year'].value_counts()

print(Cnt_year)

year = (2015,2016,2017)

sizes = (52216, 51726,25938)  #tuples
```

```
explode = (0.1, 0, 0)  # only "explode" the 2nd slice

colors = ["#2ca02c", "#9467bd", "#8c564b"]

fig1, ax1 = plt.subplots()

ax1.pie(sizes, explode=explode, labels=year, autopct='%1.1f%%',

        shadow=True, startangle=90, colors=colors)

ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

plt.title("Percentage of customers travelled each year")

plt.show()
```

Q.3. Show the passengers satisfied/not satisfied wrt Departure/Arrival time for each month?

```
import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

Airlines=pd.read_csv("Airlines.csv")

groupby_gender_and_answer = Airlines.groupby(['Month','Departure/Arrival time

convenient'])

results = groupby_gender_and_answer['Departure/Arrival time

convenient'].aggregate('count')

print(results.value_counts)
```

```
Extremely_Satisfied=[2881,3245,2324,3733,2537,2345,2456,1472,1749,2486,1748,146
1]
Not_Satisfied=[2438,2516,1844,2833,1946,1234,2313,1513,1396,1697,1220,1101]
Dict={'Extremely Satisfied':Extremely_Satisfied,'Not Satisfied':Not_Satisfied}
df=pd.DataFrame(Dict)
#row labels
Month=['Jan','Feb','Mar','Apr','May','June','July','Aug','Sep','Oct','Nov','Dec']
df.index=row_labels
print(df)
plt.plot(Month, Extremely_Satisfied, color='g',label="Satisfied")
plt.plot(Month, Not_Satisfied, color='orange', label="Not Satisfied")
plt.xlabel('Month')
plt.ylabel('No.of Passengers')
plt.title('Count of Passengers Satisfied/ Not Satisfied wrt Departure/Arrival Time by
month' )
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.show()
```

Q.4. Show the frequent flight distance covered?

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
```

```
Airlines['Flight Distance'].plot(kind='hist',bins=50,figsize=(12,6), color='grey',
title='Frequent Flight Distance')
plt.show()
```

Q.5. Show the flight distance vs Departure Delay in minutes?

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
Airlines=pd.read_csv("Airlines.csv")
Airlines.plot(kind='scatter',x='Flight Distance',y='Departure Delay in Minutes', title='Flight
Distance vs Departure Delay in minutes')
plt.show()
```