Final Project Report

Team Members ¶

Phu Gia Pham

Tejas Mirashi

1. Introduction

I. The Purposes

a. Motivation of the Project

Recently, due to the COVID-19 pandemic, lots of companies have been struggling to maintain their businesses, and some of them even had to file for bankruptcy. One of the businesses that have been affected the most is the Airline industry as it is based on the hospitality industry and most people in the world are now afraid of travelling with the COVID-19 and government restriction. However, once the effect of pandemic subsides, air travel will again pick up pace. Therefore, all the Airline companies need to focus on the customer's experience and satisfaction to create their competitive advantage.

b. The Purpose

In this project, we focus on analyzing what factors affect the customer satisfaction when experiencing their flight in order to help the Airlines companies have a deep look into their services.

II. Why is this question important?

It is important for Airlines companies, especially their sales and operating managers, to understand their customers behavior in order to not only emphasize their strengths but also find out solutions to improve their weakness in customer services. From here, the quality of services provided by Airline companies will be controlled and adjusted in accordance with customers' needs and promote the company revenue by bringing in new customers and returning customers.

III. The brief summary of findings

Post

2. Main

I. Data Processing/Cleaning

At the beginning, we were about to use both datasets from Kaggle, but we decided to analyze the first dataset which includes more than 25,800 observations. Even though using all dataset will provide enough data to improve accuracy, but at the same time, this will affect our conclusions by outliers such as departure delay in minutes and arrival delay in minutes.

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
import statsmodels.api as sm
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
```

In [2]:
```
raw = r'test.csv'
df_1=pd.read_csv(raw)
```

In [3]:
```
#Number of Observations before revoming NaN value
df_1.shape
```

Out[3]: (25976, 25)

In [4]:
```
#Revoming NaN values in column Arrival Delay in Minutes
df_1.dropna(subset = ["Arrival Delay in Minutes"], inplace=True)
```

In [5]:
```
df_1.shape
```

Out[5]: (25893, 25)

II. Description of the Dataset including the List of Fields

a. Source of Dataset

The data is taken from Kaggle dataset which is a web service platform for data.

https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction (https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction)

b. Description of the Dataset

c. Data Type of Fields

In [6]:
```
#Type of the Dataset
type(df_1)
```

Out[6]:
```
pandas.core.frame.DataFrame
```

In [7]:
```
# Series with the data type of each column
df_1.dtypes
```

Out[7]:
```
Unnamed: 0                             int64
id                                     int64
Gender                                object
Customer Type                         object
Age                                    int64
Type of Travel                        object
Class                                 object
Flight Distance                        int64
Inflight wifi service                  int64
Departure/Arrival time convenient      int64
Ease of Online booking                 int64
Gate location                          int64
Food and drink                         int64
Online boarding                        int64
Seat comfort                           int64
Inflight entertainment                 int64
On-board service                       int64
Leg room service                       int64
Baggage handling                       int64
Checkin service                        int64
Inflight service                       int64
Cleanliness                            int64
Departure Delay in Minutes             int64
Arrival Delay in Minutes             float64
satisfaction                          object
dtype: object
```

In [8]:
```
#Columns labels of the Dataset
df_1.columns
```

Out[8]:
```
Index(['Unnamed: 0', 'id', 'Gender', 'Customer Type', 'Age', 'Type of T
ravel',
       'Class', 'Flight Distance', 'Inflight wifi service',
       'Departure/Arrival time convenient', 'Ease of Online booking',
       'Gate location', 'Food and drink', 'Online boarding', 'Seat comf
ort',
       'Inflight entertainment', 'On-board service', 'Leg room servic
e',
       'Baggage handling', 'Checkin service', 'Inflight service',
       'Cleanliness', 'Departure Delay in Minutes', 'Arrival Delay in M
inutes',
       'satisfaction'],
      dtype='object')
```

In [9]:
```
#An array of values of the Dataset
df_1.values
```

Out[9]:
```
array([[0, 19556, 'Female', ..., 50, 44.0, 'satisfied'],
       [1, 90035, 'Female', ..., 0, 0.0, 'satisfied'],
       [2, 12360, 'Male', ..., 0, 0.0, 'neutral or dissatisfied'],
       ...,
       [25973, 37675, 'Female', ..., 0, 0.0, 'neutral or dissatisfie
d'],
       [25974, 90086, 'Male', ..., 0, 0.0, 'satisfied'],
       [25975, 34799, 'Female', ..., 0, 0.0, 'neutral or dissatisfie
d']],
      dtype=object)
```

III. Basic descriptive features of the data

a. Number of Observations

83 observations were removed from the dataset as they had missing values in column Arrival Delay in Minutes. The reason why we decided to remove these observations becaues the data is already big enough for us to draw conclusion from the dataset.

In [11]:
```
# The dimensiionality of the Dataset
df_1.shape
```

Out[11]:
```
(25893, 25)
```

25893 observations will be observed in this project

b. Number of Observations by Variable/Field

In [10]:
```
#Summary of the Dataset
df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 25893 entries, 0 to 25975
Data columns (total 25 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   Unnamed: 0                        25893 non-null  int64  
 1   id                                25893 non-null  int64  
 2   Gender                            25893 non-null  object 
 3   Customer Type                     25893 non-null  object 
 4   Age                               25893 non-null  int64  
 5   Type of Travel                    25893 non-null  object 
 6   Class                             25893 non-null  object 
 7   Flight Distance                   25893 non-null  int64  
 8   Inflight wifi service             25893 non-null  int64  
 9   Departure/Arrival time convenient 25893 non-null  int64  
 10  Ease of Online booking            25893 non-null  int64  
 11  Gate location                     25893 non-null  int64  
 12  Food and drink                    25893 non-null  int64  
 13  Online boarding                   25893 non-null  int64  
 14  Seat comfort                      25893 non-null  int64  
 15  Inflight entertainment            25893 non-null  int64  
 16  On-board service                  25893 non-null  int64  
 17  Leg room service                  25893 non-null  int64  
 18  Baggage handling                  25893 non-null  int64  
 19  Checkin service                   25893 non-null  int64  
 20  Inflight service                  25893 non-null  int64  
 21  Cleanliness                       25893 non-null  int64  
 22  Departure Delay in Minutes        25893 non-null  int64  
 23  Arrival Delay in Minutes          25893 non-null  float64
 24  satisfaction                      25893 non-null  object 
dtypes: float64(1), int64(19), object(5)
memory usage: 5.1+ MB
```

c. Number of Missing Values by Variable/Field

In [12]:
```
#The number of missing values
df_1.isnull().sum(axis=0)
```

Out[12]:
```
Unnamed: 0                            0
id                                    0
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Inflight wifi service                 0
Departure/Arrival time convenient     0
Ease of Online booking                0
Gate location                         0
Food and drink                        0
Online boarding                       0
Seat comfort                          0
Inflight entertainment                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Inflight service                      0
Cleanliness                           0
Departure Delay in Minutes            0
Arrival Delay in Minutes              0
satisfaction                          0
dtype: int64
```

According to the code above, there are 83 values missing in the Arrival Delay in Minutes. The rest is fine and has no missing data.

d. Describe the Data

In [13]:
```
round(df_1.describe(),2)
```

Out[13]:

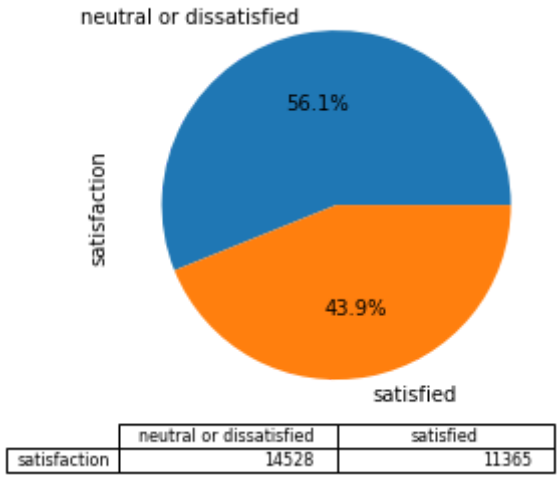
	Unnamed: 0	id	Age	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking
count	25893.00	25893.00	25893.00	25893.00	25893.00	25893.00	25893.00
mean	12987.84	65021.97	39.62	1193.75	2.72	3.05	2.76
std	7499.18	37606.10	15.13	998.63	1.33	1.53	1.41
min	0.00	17.00	7.00	31.00	0.00	0.00	0.00
25%	6496.00	32209.00	27.00	414.00	2.00	2.00	2.00
50%	12984.00	65344.00	40.00	849.00	3.00	3.00	3.00
75%	19482.00	97623.00	51.00	1744.00	4.00	4.00	4.00
max	25975.00	129877.00	85.00	4983.00	5.00	5.00	5.00

iv. Analysis and Explanations

How many percentage of customers are satisfied or dissatisfied about the overall airline experience?

In [14]:
```
df_1['satisfaction'].value_counts().plot(kind='pie',table=True, autopct='%1.1f%%')
```

Out[14]: <AxesSubplot:ylabel='satisfaction'>



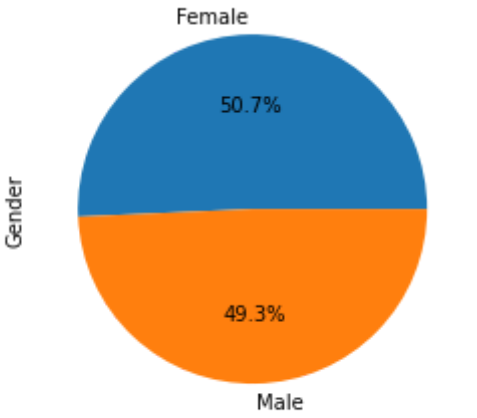
	neutral or dissatisfied	satisfied
satisfaction	14528	11365

As we can see, there are more dissatisfied / neutral passengers than satisfied ones. The airline needs to dig deeper in to the reasons and implement measures to improve this scenario.

What is the percentage of satisfied and dissatisfied customers amongst genders?

In [15]:
```
df_1['Gender'].value_counts().plot(kind='pie',table=True, autopct='%1.1
f%%')
```

Out[15]: <AxesSubplot:ylabel='Gender'>



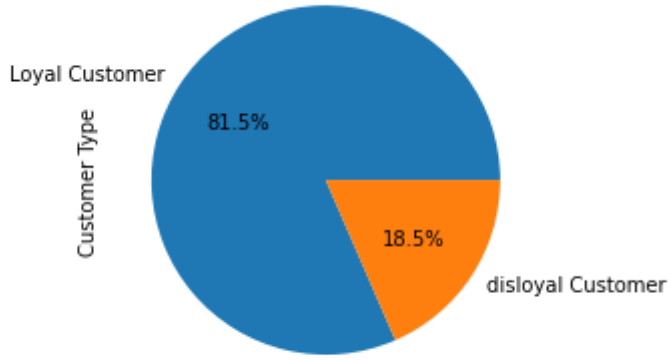
	Female	Male
Gender	13127	12766

Continuing the earlier scenario, the distribution based on gender is almost equal. As a remedy, the airline may analyze the genderwise preferences to make the flight experience a better one.

Distribution of passenger types

In [16]:
```
df_1['Customer Type'].value_counts().plot(kind='pie',table=True, autopct
='%1.1f%%')
```

Out[16]: <AxesSubplot:ylabel='Customer Type'>



	Loyal Customer	disloyal Customer
Customer Type	21111	4782

As a percentage, the distribution is highly tilted towards the passengers tagged as loyal. However, we need to analyze if the criteria of tagging if the distribution as per the above pie chart, is similar to the distribution between satisfied and dissatisfied / neutral.

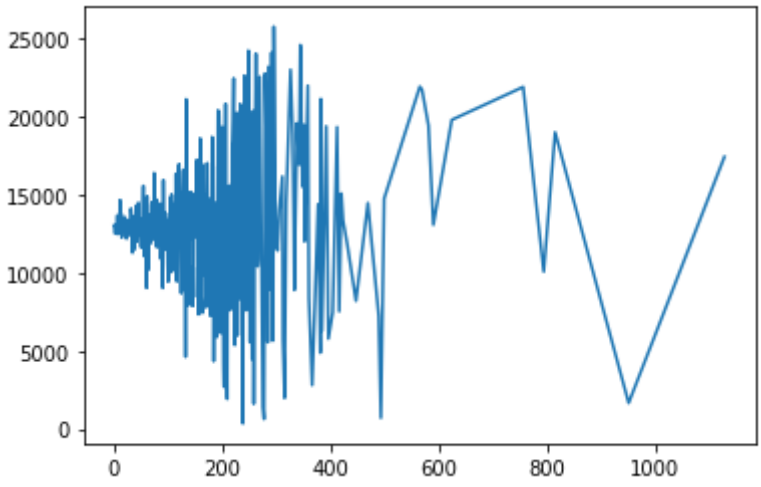
1. Is there a regression between the delay in departure and arrival times?

In [17]:
```
df_delay=df_1.groupby(['Departure Delay in Minutes']).mean()
df_delay.head()
```

Out[17]:

Departure Delay in Minutes	Unnamed: 0	id	Age	Flight Distance	Inflight wifi service	Departure/Arr time conveni
0	13021.841646	62479.982943	39.644880	1183.155352	2.748243	3.054991
1	13032.566940	74755.225410	38.926230	1205.103825	2.693989	2.995902
2	12503.811744	73226.663212	38.974093	1288.132988	2.747841	3.012090
3	13046.286260	69788.326336	39.328244	1169.318702	2.664122	2.975191
4	13096.794702	66793.441501	39.456954	1262.342163	2.704194	3.050773

In [18]:
```
plt.plot(df_delay['Unnamed: 0'])
plt.show()
```

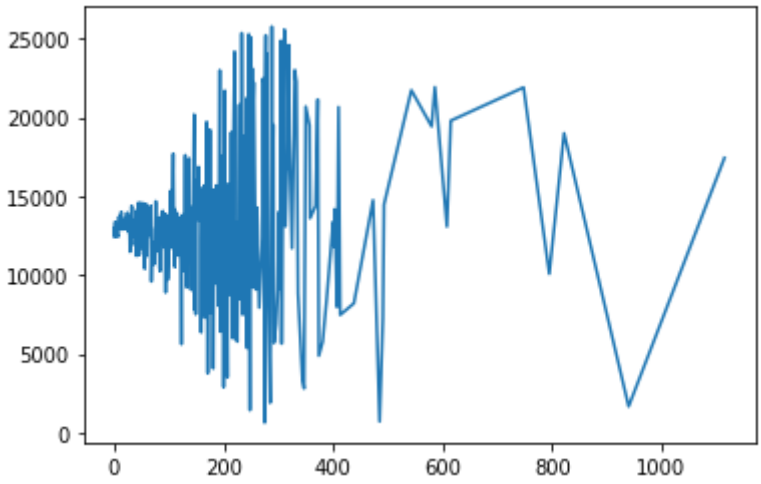


In [19]:
```
df_arrival=df_1.groupby(['Arrival Delay in Minutes']).mean()
df_arrival.head()
```

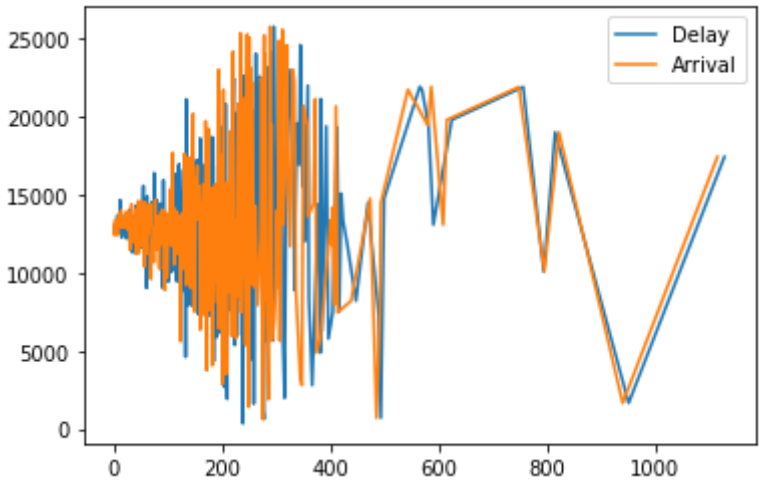
Out[19]:

	Unnamed: 0	id	Age	Flight Distance	Inflight wifi service	Departure/Arriv time convenie
Arrival Delay in Minutes						
0.0	13001.656708	64627.640263	39.653282	1200.067905	2.761066	3.048170
1.0	12398.304104	67888.552239	39.927239	1301.912313	2.742537	3.057836
2.0	13379.617591	65531.684512	39.476099	1169.504780	2.711281	2.971319
3.0	12917.681633	69212.932653	39.906122	1192.373469	2.706122	3.112245
4.0	12974.399142	67539.060086	39.875536	1186.879828	2.776824	3.152361

In [20]:
```
plt.plot(df_arrival['Unnamed: 0'])
plt.show()
```

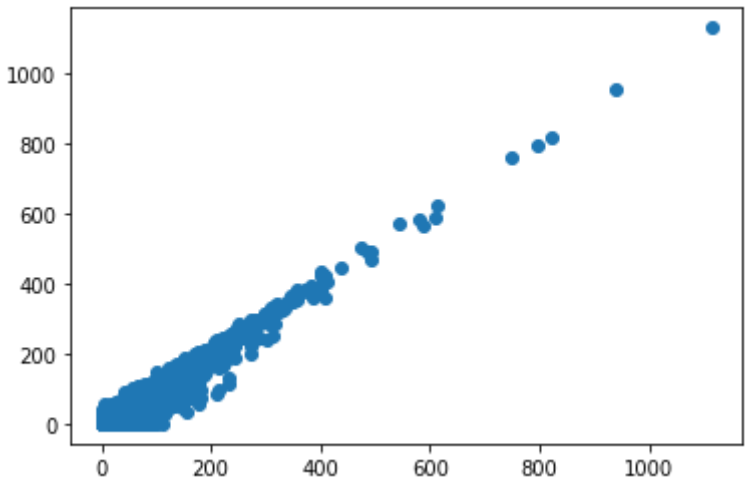


In [21]:
```
plt.plot(df_delay['Unnamed: 0'], label ="Delay")
plt.plot(df_arrival['Unnamed: 0'], label ="Arrival")
plt.legend()
plt.show()
```



In [22]:
```
plt.scatter(df_1['Arrival Delay in Minutes'],df_1['Departure Delay in Minutes'])
```

Out[22]: <matplotlib.collections.PathCollection at 0x7f17db6547d0>



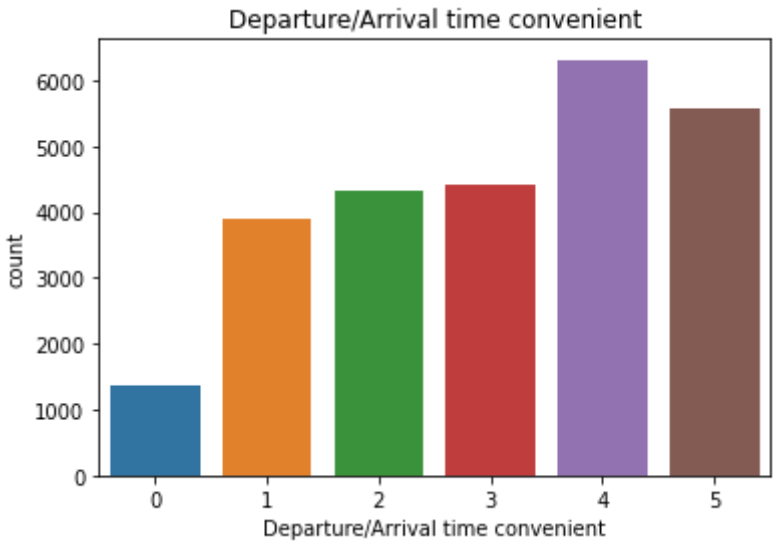
On analyzing the arrival and departure delay departure times, we understand that there is a very close relation between both the variables. The outcome is quite logical as once there is a delay in arrival, there ought to be a delay in departure.

Predictive suggestion – The airline can predict the estimated delay in departure once they know the delay in arrival.

Prescriptive suggestion – Once the delay in departure is predicted, the airline can arrive at a cut-off delay in time which can be covered up by better ground management and increase in cruising speed. Say, the airline arrives at a conclusion that up to a 20-minute delay, the same can be covered up by better ground service management and an increase in cruising speed.

2. What is the level of satisfaction of passengers based on delay in departure and arrival?

In [23]:
```
#Distribution of satisfaction level in Departure/Arrival time convenient
sns.countplot(x ='Departure/Arrival time convenient', data = df_1)
plt.title("Departure/Arrival time convenient")
plt.show()
```

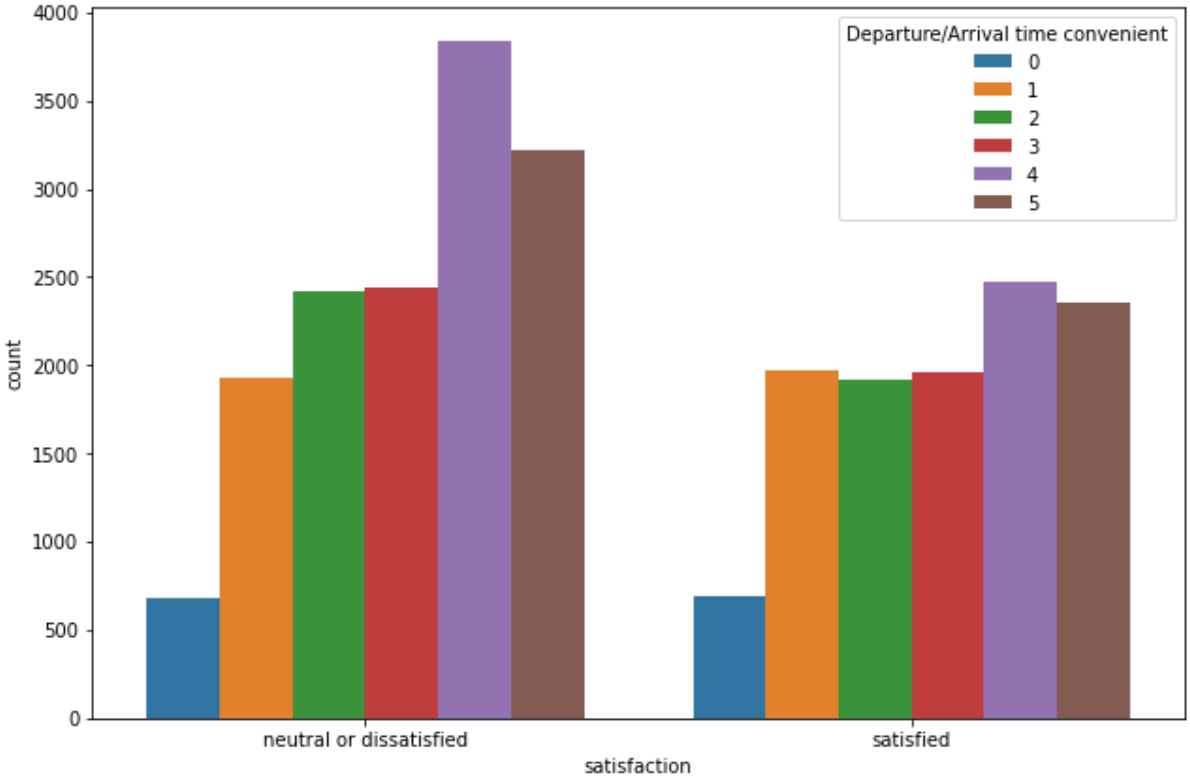


In [24]:
```
print(round((df_1['Departure/Arrival time convenient'].value_counts()/le
n(df_1['Departure/Arrival time convenient']))*100),2)
```

```
4    24.0
5    22.0
3    17.0
2    17.0
1    15.0
0     5.0
Name: Departure/Arrival time convenient, dtype: float64 2
```

In [25]:
```
def compare_2_vars(var1, var2):
    fig, ax = plt.subplots(figsize = (9, 6))
    plt.subplot(111)
    sns.countplot(data = df_1, x = var1, hue = var2, order = ["neutral o
r dissatisfied", "satisfied"])
    plt.tight_layout()
    plt.show()
```

In [26]:
```
compare_2_vars('satisfaction', 'Departure/Arrival time convenient')
```



The results are quite surprising. As we can see from the pie chart above, the number of satisfied and dissatisfied / neutral passengers is almost equal. Out of the individual cohort, almost more than half of the dissatisfied passengers have given a rating of 4 and above for the convenience of arrival and departure time. Hence, a high rating for convenience of arrival and departure time does not result in the passenger being satisfied.

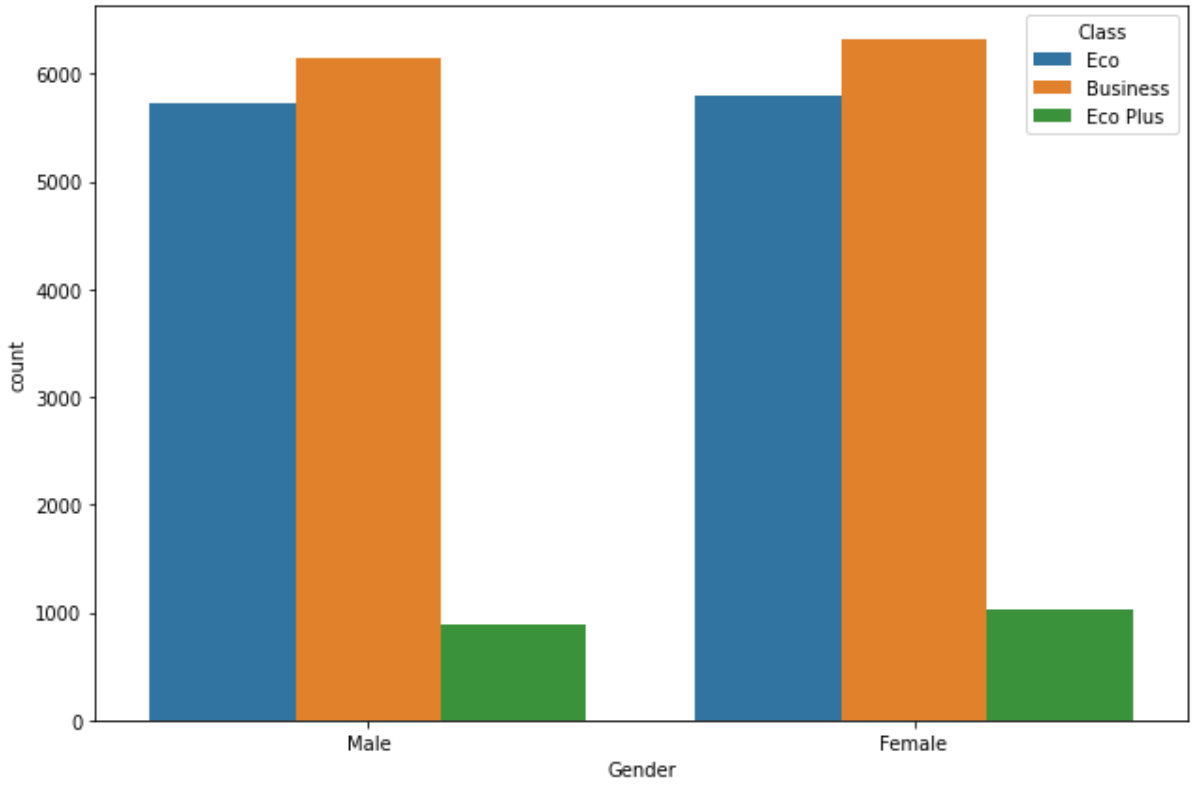
Predictive suggestion – The airline can further analyze the cohort of the dissatisfied passengers despite rating the convenience of arrival and departure time based on other factors such as age, gender, class of travel and type of travel.

Prescriptive suggestion – The airline can contact such passengers (dissatisfied passengers who have given a rating 3 and below) via mailers about various options of day and time combinations on which the routes are operational. A similar exercise can also be done with satisfied passengers with a rating of 3 and below.

3. Does the age, gender and type of travel affect the choice of class?

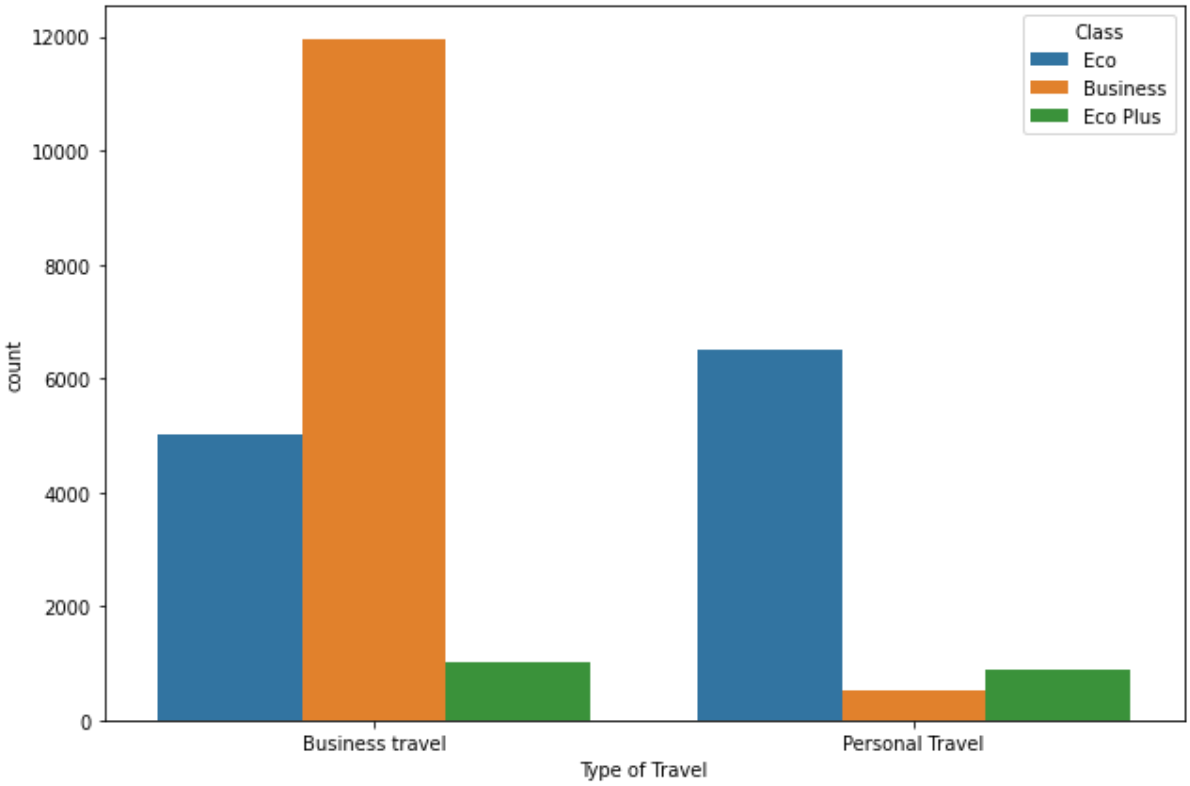
In [27]:
```
def gender(var1, var2):
    fig, ax = plt.subplots(figsize = (9, 6))
    plt.subplot(111)
    sns.countplot(data = df_1, x = var1, hue = var2, order = ["Male", "F
emale"])
    plt.tight_layout()
    plt.show()
```

In [28]:
```
gender('Gender', 'Class')
```



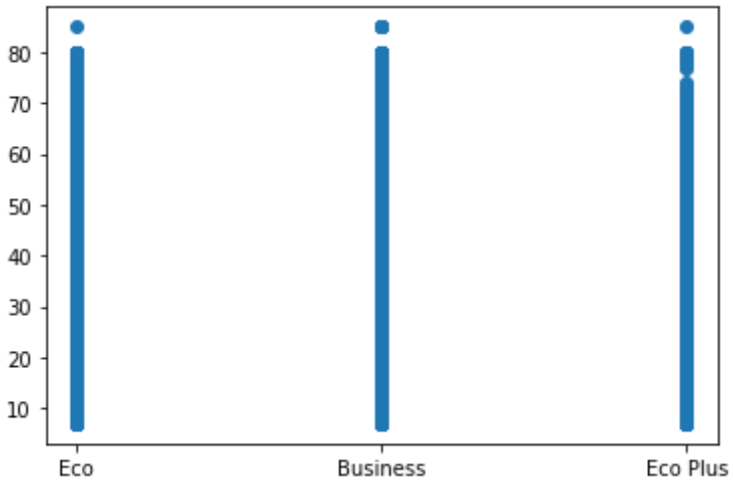
In [29]:
```
def travel_type(var1, var2):
    fig, ax = plt.subplots(figsize = (9, 6))
    plt.subplot(111)
    sns.countplot(data = df_1, x = var1, hue = var2, order = ["Business
 travel", "Personal Travel"])
    plt.tight_layout()
    plt.show()
```

In [30]:
```
travel_type("Type of Travel", "Class")
```



In [31]:
```
plt.scatter(df_1['Class'],df_1['Age'])
```

Out[31]: <matplotlib.collections.PathCollection at 0x7f17db7157d0>



The distribution of choice of class in context to gender is almost identical.

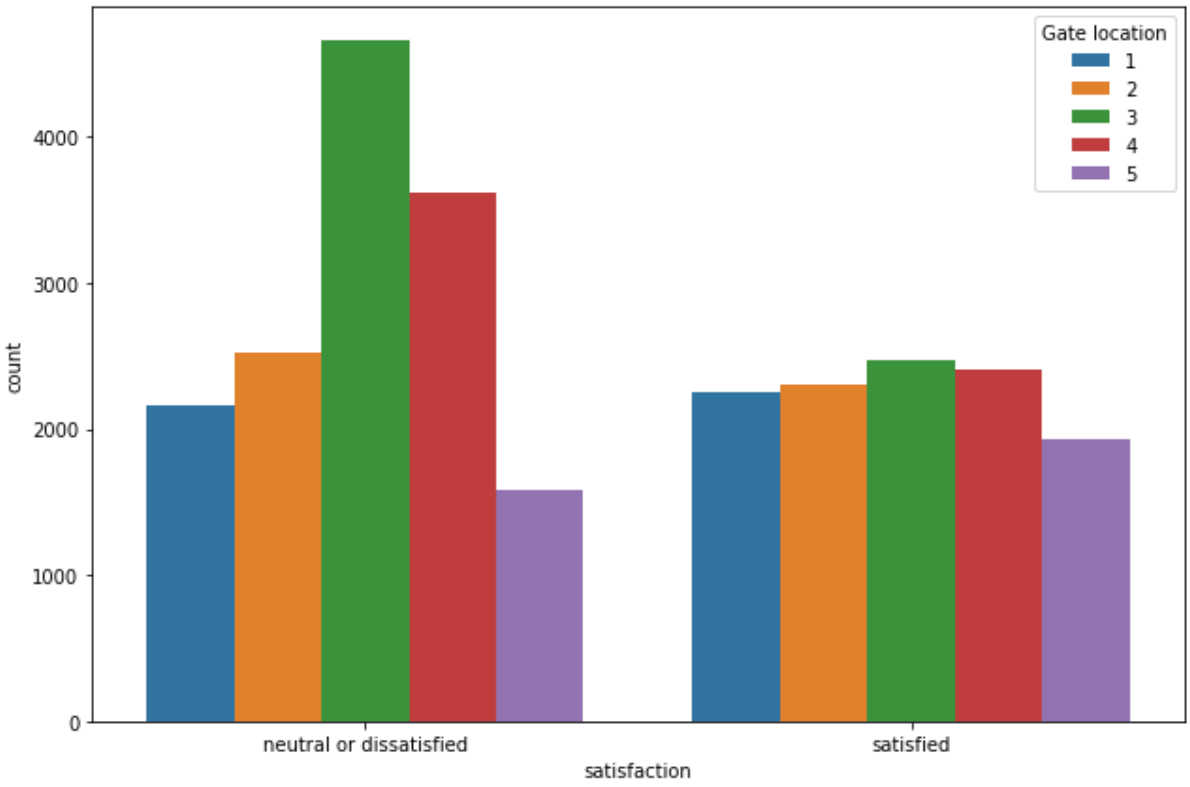
We see that extremely high number of passengers traveling for business purpose prefer to travel via business class. There is a possibility that such a scenario has risen as business travel is mostly paid by the company and not the individual.

Predictive suggestion – The airline can use data such as name of the company, different companies among same sector and time of travel during the calendar year to get information about corporate passengers using 'Eco' and 'Eco Plus' class.

Prescriptive suggestion – The airline can contact such companies and offer them deals so that the economy class passengers choose business class for business purpose travel. They can even tie up with local cab operators and hotels at the frequent destinations as an added service for business class passengers. This way the revenue can be increased as business class tickets would be sold at a higher rate and also, the local cab operators and hotels can be roped in on a revenue sharing model.

4. Does the gate location affect the level of satisfaction?

```
In [32]: compare_2_vars('satisfaction', 'Gate location')
```



In [33]:
```
print(round((df_1['Gate location'].value_counts()/len(df_1['Gate locatio
n']))*100),2)
```

```
3    28.0
4    23.0
2    19.0
1    17.0
5    14.0
Name: Gate location, dtype: float64 2
```

We can see that a high number of passengers in both cohorts, satisfied as well as dissatisfied have rated the gate location 3 or below. Hence, we can conclude that the gate location plays an important role on the level of satisfaction.

Predictive suggestion – The airline can predict the most convenient arrival and departure time using historic data where they have the maximum passenger traffic and accordingly try to allot a convenient gate location.

Prescriptive suggestion – If a convenient gate location cannot be allotted, the airline can provide a buggy service to transfer the passengers from the security check-in to the allotted gate.

5. What is the trend of ratings for in-flight wifi service and in-flight entertainment in context to the flight distance?

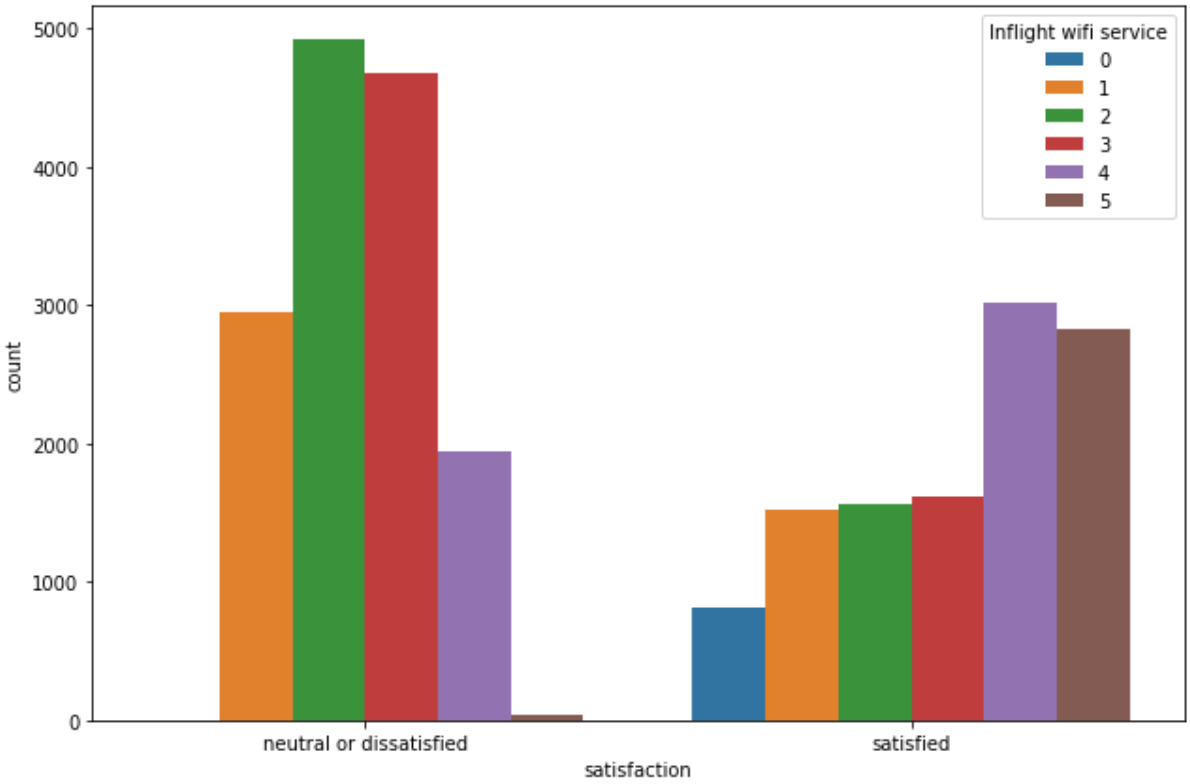
In [34]:
```
long_distance_flight = df_1[(df_1['Flight Distance'] > 1000)]
long_distance_flight.head()
```

Out[34]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depa time
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3
6	6	79433	Female	Loyal Customer	77	Business travel	Business	3987	5	5
7	7	97286	Female	Loyal Customer	43	Business travel	Business	2556	2	2

5 rows × 25 columns

In [35]:
```
compare_2_vars('satisfaction', 'Inflight wifi service')
```



In [36]:
```
df_flight_distance=df_1.groupby(['Flight Distance']).mean()
df_flight_distance.head()
```

Out[36]:

Flight Distance	Unnamed: 0	id	Age	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking
31	4922.333333	30053.666667	45.000000	3.000000	4.000000	1.666667
56	15089.666667	49790.000000	26.000000	1.666667	3.000000	2.000000
67	14114.750000	38517.562500	39.250000	2.718750	3.312500	2.437500
73	13607.555556	41493.277778	41.166667	2.888889	2.111111	2.500000
74	10577.000000	16284.916667	38.083333	2.416667	3.000000	2.750000

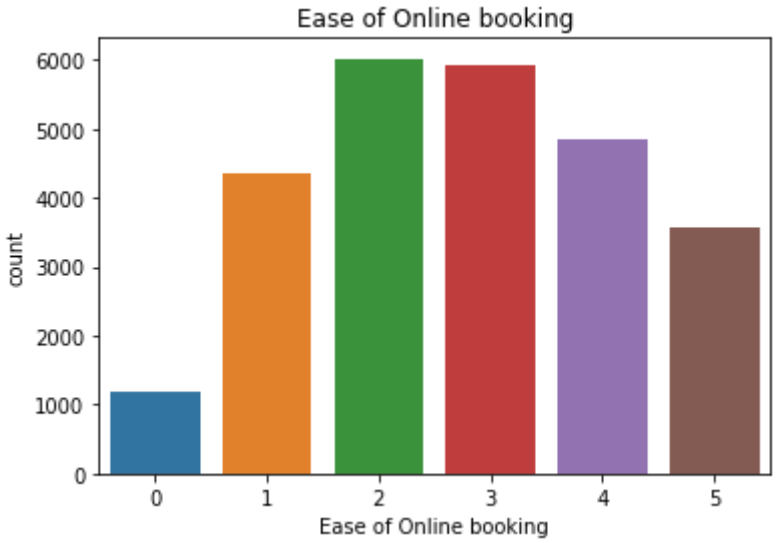
We have considered a long duration flight as the one which travels more than 1,000 miles. One limitation is that no data is available if there is a stop-over in such flights. From the analysis above, in-flight wifi service is an important factor to determine the level of satisfaction. Majority of the dissatisfied passengers have rated the in-flight wifi service 3 or below. On the contrary, majority of the satisfied passengers have rated the in-flight wifi 4 or above.

Predictive suggestion – The longer the flight distance, the more difficult it is to keep the passengers satisfied.

Prescriptive suggestion – For long duration flights, better in-flight entertainment such as playing cards or small board games can be provided. Also, free wifi service can be made available for certain slots depending on the time-zone so that the passengers are able to communicate with their family and friends. It would also be helpful to appease the business purpose passengers as they would have access to e-mails.

6. Do the age criteria have any effect on ratings for ease of online booking and online boarding?

In [35]:
```
sns.countplot(x ='Ease of Online booking', data = df_1)
plt.title("Ease of Online booking")
plt.show()
```

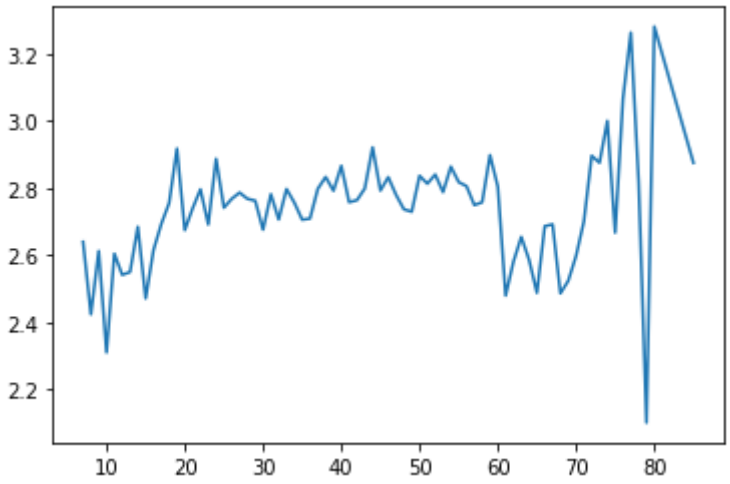


In [36]:
```
df_age=df_1.groupby(['Age']).mean()
df_age.head()
```

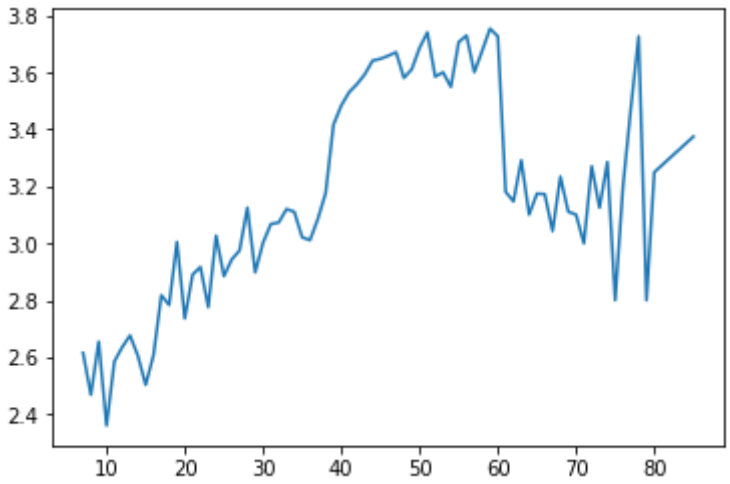
Out[36]:

	Unnamed: 0	id	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	loc
Age							
7	12231.688525	70043.147541	908.663934	2.614754	3.786885	2.639344	2.77
8	12647.987179	60108.070513	862.500000	2.378205	3.474359	2.423077	2.80
9	13416.400000	64834.909091	847.472727	2.606061	3.309091	2.612121	2.95
10	13287.611511	73537.050360	876.949640	2.410072	3.244604	2.309353	3.02
11	13355.685535	64889.805031	897.761006	2.490566	3.383648	2.603774	2.88

In [37]:
```
plt.plot(df_age['Ease of Online booking'])
plt.show()
```



In [38]:
```
plt.plot(df_age['Online boarding'])
plt.show()
```



We can see that most passengers have rated the ease of online booking a 3 or lower.

For the first 3 quartiles, i.e. for the age group from 7 to 51, we see that the rating is between 2.3 and 2.9. Here, we are referring to the age group that we can say to be conversant with technology. So we can infer that the airline has to work more on smoothening the online booking process.

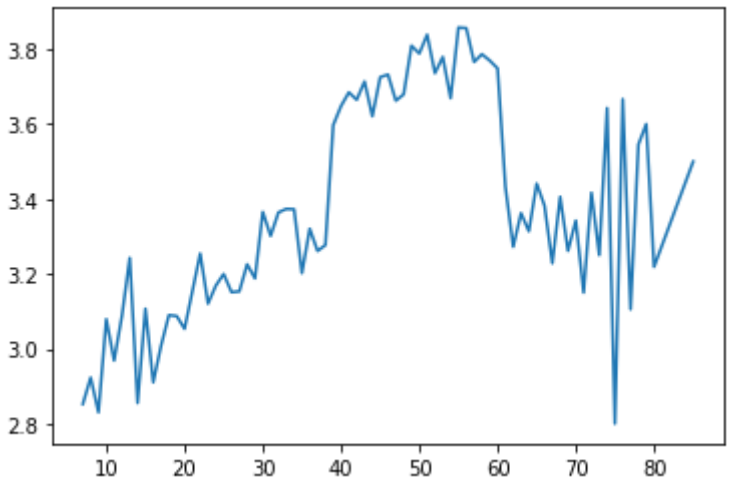
For the online boarding process, we can see that as the age factor increases, the online boarding is rated higher. There are certain outliers but overall we can see that the rating for ease of online boarding is on the higher side.

Predictive suggestion – Using historic data the airline can predict the age group that prefers to travel during a particular time of the year.

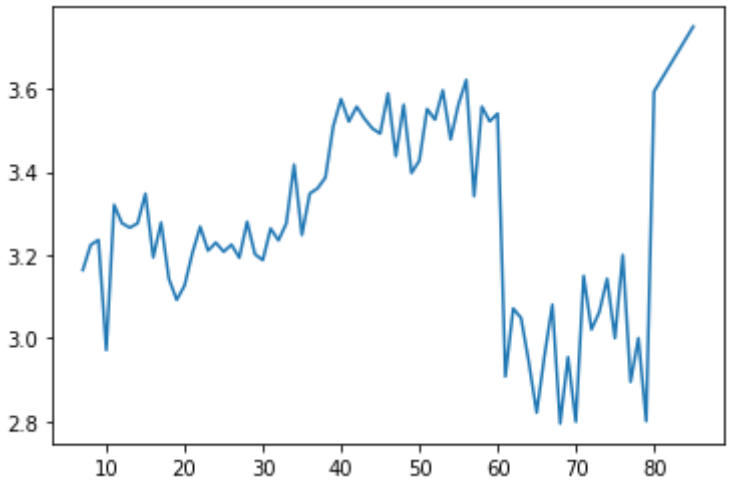
Prescriptive suggestion – Depending on the age group of passengers analyzed using predictive suggestion, the airline can concentrate if online booking and online boarding can be worked upon. For example, if for a particular flight, the passenger cohort is young, the airline can suggest the passengers to do online boarding and keep a very few counters open for physical check-in.

7. Do age and type of travel have any effect on the ratings for seat comfort and leg room service?

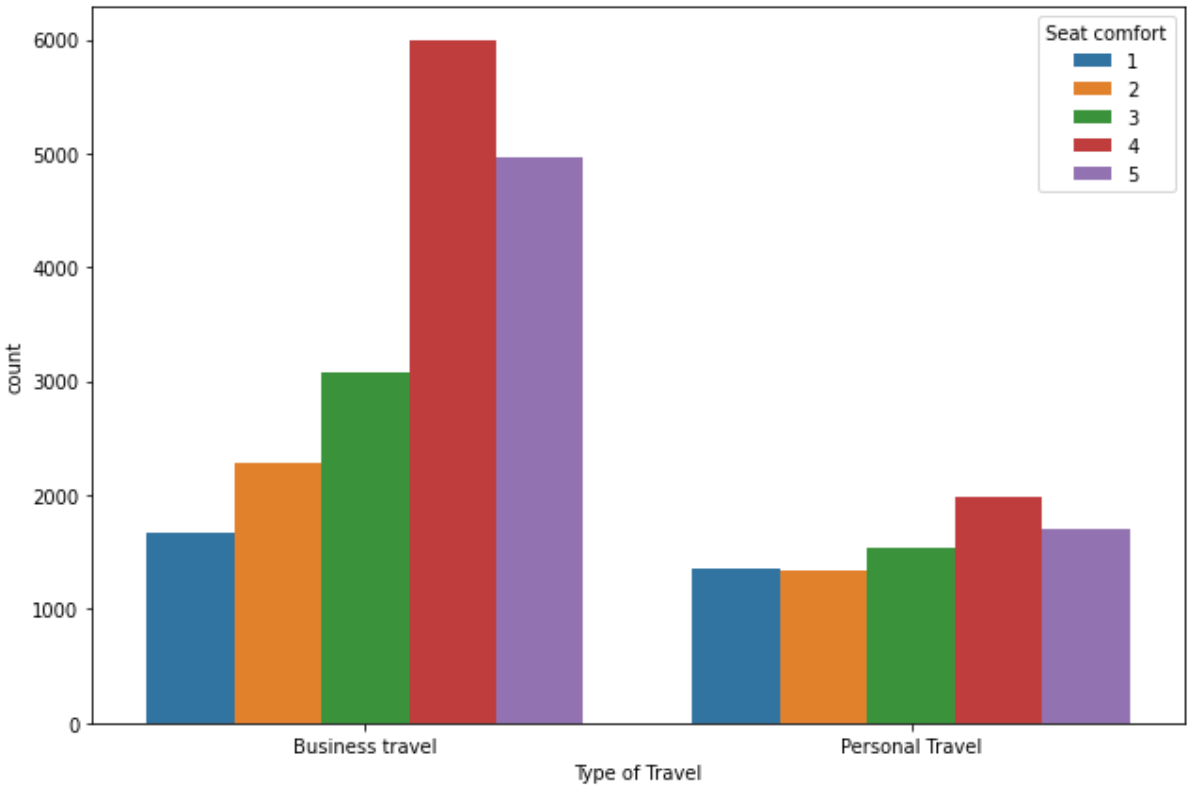
In [39]:
```
plt.plot(df_age['Seat comfort'])
plt.show()
```



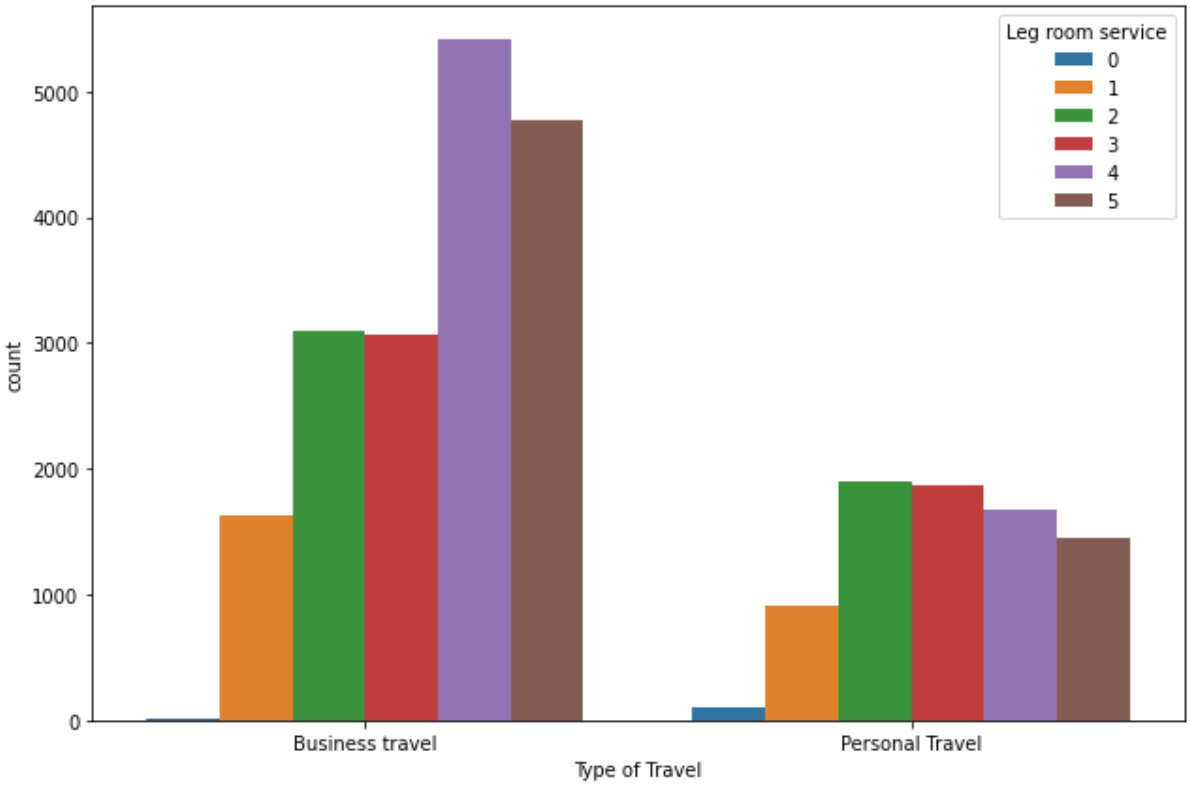
In [40]:
```
plt.plot(df_age['Leg room service'])
plt.show()
```



In [42]:
```
travel_type("Type of Travel", "Seat comfort")
```



In [43]:
```
travel_type("Type of Travel", "Leg room service")
```



We can see that most passengers have rated the ease of online booking a 3 or lower.

For the first 3 quartiles, i.e. for the age group from 7 to 51, we see that the rating is between 2.3 and 2.9. Here, we are referring to the age group that we can say to be conversant with technology. So we can infer that the airline has to work more on smoothening the online booking process.

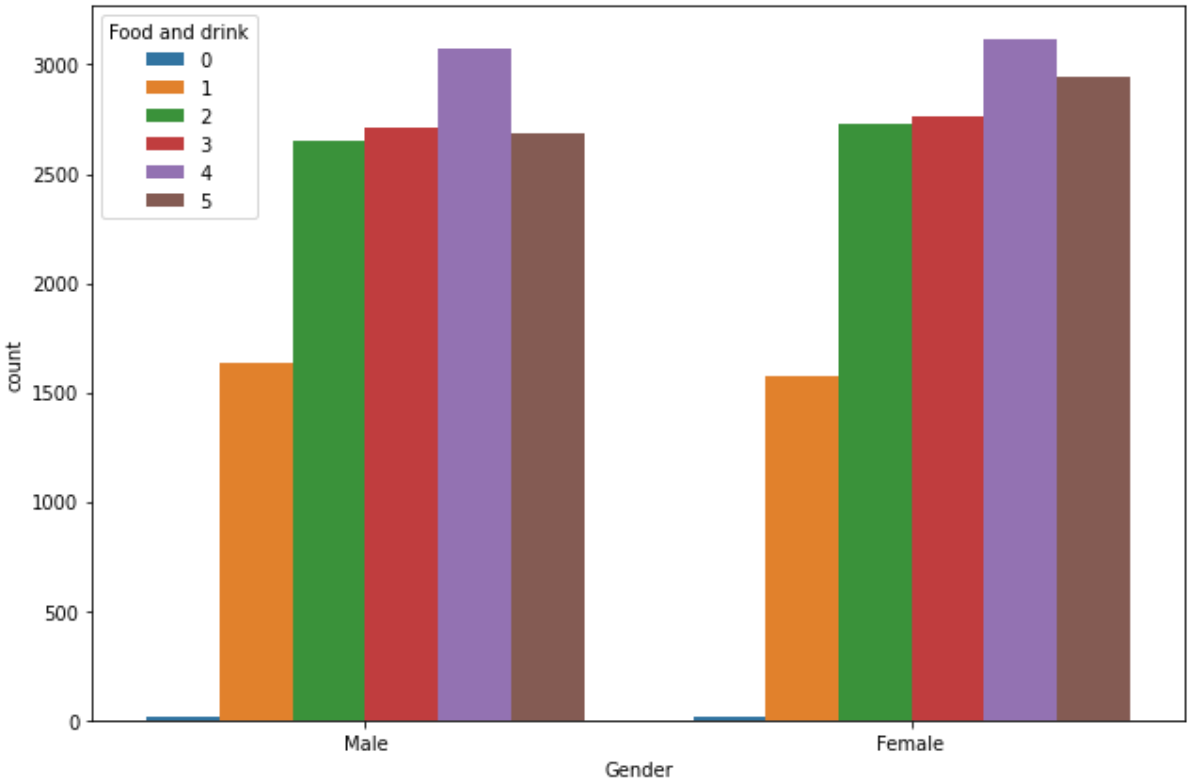
For the online boarding process, we can see that as the age factor increases, the online boarding is rated higher. There are certain outliers but overall we can see that the rating for ease of online boarding is on the higher side.

Predictive suggestion – Using historic data the airline can predict the age group that prefers to travel during a particular time of the year.

Prescriptive suggestion – Depending on the age group of passengers analyzed using predictive suggestion, the airline can concentrate if online booking and online boarding can be worked upon. For example, if for a particular flight, the passenger cohort is young, the airline can suggest the passengers to do online boarding and keep a very few counters open for physical check-in.

8. What is the trend of ratings for food and drink in context of the gender of the passenger?

In [44]:
```
gender('Gender', 'Food and drink')
```



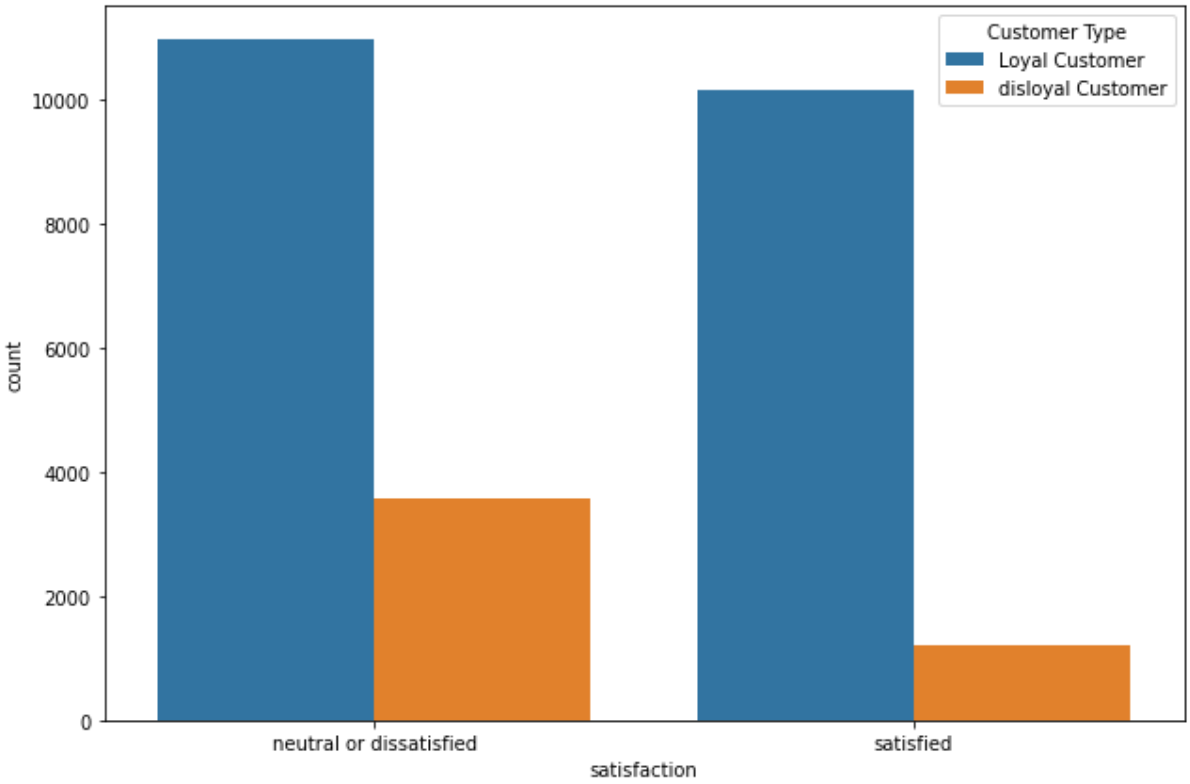
Analyzing both the histograms, we can see that the ratings for food and drink irrespective of gender are almost equally distributed in terms of count. It is because of such a distribution, it is difficult to prescribe a suggestion to the airline for areas of improvement.

Predictive suggestion – Since there are only two variables, the airline can predict the range of ratings for food and drink given the gender.

Prescriptive suggestion – If the trip consists of one gender majorly, the menu can be tweaked according to the preferences.

9. What is the correlation between the type of customer and the level of satisfaction? Is it logical to assume that a loyal customer would be satisfied irrespective of the ratings?

In [45]:
```
compare_2_vars('satisfaction', 'Customer Type')
```



It is quite surprising to see that almost equal number of loyal as well as disloyal passengers are satisfied and dissatisfied / neutral.

Predictive suggestion – The airline can predict if the type of customer is related to the level of satisfaction.

Prescriptive suggestion – The airline may have to deep dive in the criteria which tags a passenger as loyal or disloyal.

In [61]:
```
def satisfaction_scale(df_1):
    if df_1['satisfaction'] == 'satisfied':
        return 1
    else:
        return 0
df_1['satisfaction_scale'] = df_1.apply(satisfaction_scale, axis=1)
df_1.head()
```

Out[61]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depa tim
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3

5 rows × 28 columns

In [62]:
```
def customer_type_scale(df_1):
    if df_1['Customer Type'] == "Loyal Customer":
        return 1
    else:
        return 0
df_1['customer_type_scale'] = df_1.apply(customer_type_scale, axis=1)
df_1.head()
```

Out[62]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depa time
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3

5 rows × 29 columns

In [63]:
```
correlation = df_1['satisfaction_scale'].corr(df_1['customer_type_scale'
])
correlation
```

Out[63]: 0.1794822606093637

We can see that the correlation between the satisfaction and type of customer though positive, is not strong. We cannot say that a customer though classified as loyal, cannot be said to be satisfied with the overall flight experience. The criteria using which the airline decides the loyalty of a passenger needs to be re-evaluated.

10. Given the type of travel and the delay in arrival and departure being a non-null value, does it affect the satisfaction?

We have combined two columns Departure Delay in Minutes and Arrival Delay in Minutes and treated the same as one single variable for addressing this question.

In [47]:
```
df_1['Departure/Arrival Delay in Minutes'] = df_1['Departure Delay in Mi
nutes'] + df_1['Arrival Delay in Minutes']
```

In [48]:
```
df_1['Departure/Arrival Delay in Minutes'].head()
```

Out[48]:
```
0    94.0
1     0.0
2     0.0
3     6.0
4    20.0
Name: Departure/Arrival Delay in Minutes, dtype: float64
```

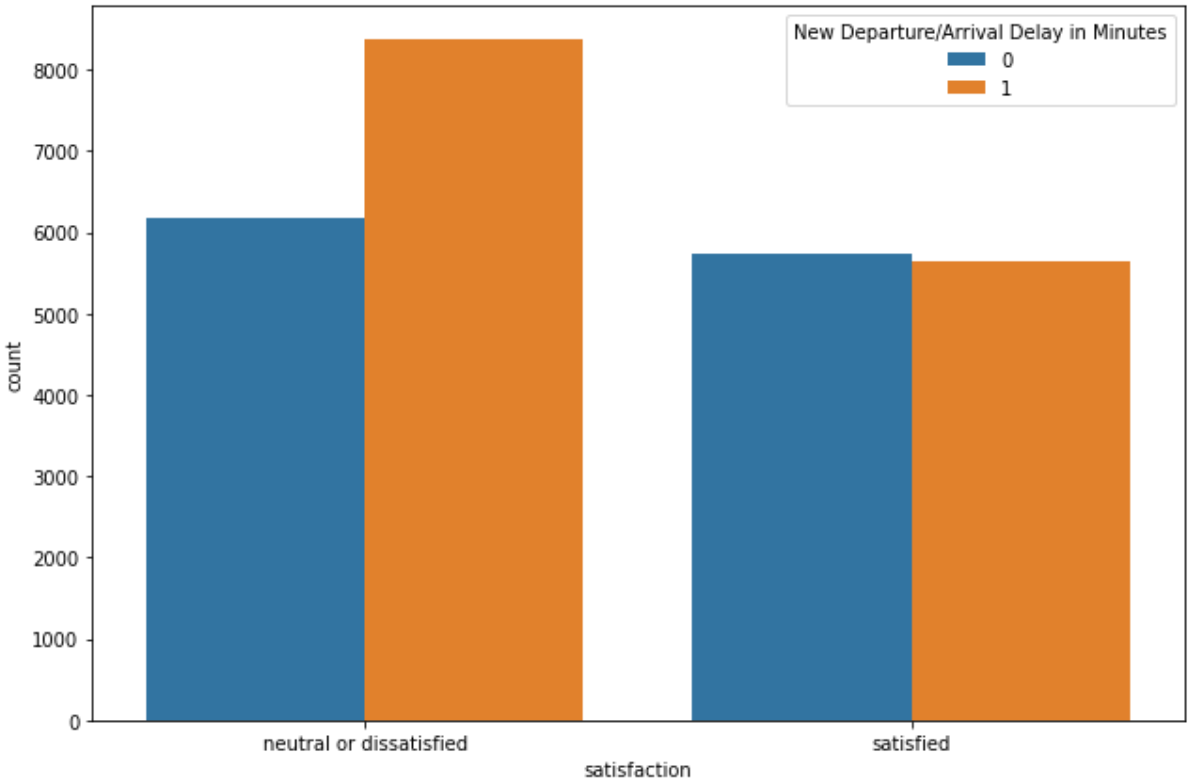
In [54]:
```
def change(df_1):
    if df_1['Departure/Arrival Delay in Minutes'] == 0.0:
        return 0
    else:
        return 1
df_1['New Departure/Arrival Delay in Minutes'] = df_1.apply(change, axis
= 1)
df_1.head()
```

Out[54]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depa time
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3

5 rows × 27 columns

```
In [55]: compare_2_vars('satisfaction', 'New Departure/Arrival Delay in Minutes')
```



For this question, we have combined the arrival delay and departure delay columns. Then if both the columns have '0' i.e. no delay in either arrival or departure, the column is tagged as 0, else 1.

The analysis for dissatisfied/neutral passengers is logical as more number of passengers are dissatisfied. However, for the satisfied passengers cohort, we can see that almost equal number of flights were delayed or not.

The airline can analyse these flights for the trend of other ratings to derive insights for the passengers being satisfied inspite of the flight being delayed. Those insights can be useful to improve the overall flight experience.

Hypothesis

Hypothesis 1

H0: A female passenger travelling business class is satisfied with the flight experience.
HA: A female passenger travelling business class is dissatisfied with the flight experience.

In [67]:
```
female_number = df_1[(df_1['Gender'] == 'Female')]
female_number.head()
```

Out[67]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depa time
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3
6	6	79433	Female	Loyal Customer	77	Business travel	Business	3987	5	5
7	7	97286	Female	Loyal Customer	43	Business travel	Business	2556	2	2

5 rows × 29 columns

In [79]:
```
female_business_class = female_number[(female_number.Class == 'Business'
)]
female_business_class.head()
```

Out[79]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	De ti
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
6	6	79433	Female	Loyal Customer	77	Business travel	Business	3987	5	5
7	7	97286	Female	Loyal Customer	43	Business travel	Business	2556	2	2
9	9	62482	Female	Loyal Customer	46	Business travel	Business	1744	2	2
11	11	115550	Female	Loyal Customer	33	Business travel	Business	325	2	5

5 rows × 29 columns

In [80]:
```
female_business_class['satisfaction'].value_counts(normalize=True)
```

Out[80]:
```
satisfied                 0.691808
neutral or dissatisfied   0.308192
Name: satisfaction, dtype: float64
```

In [82]:
```
def female_business_class(df_1):
    if df_1['Gender'] == 'Female' and df_1['Class'] == 'Business':
        return 1
    else:
        return 0
df_1['Female_Business_Class'] = df_1.apply(female_business_class, axis =
1)
```

In [91]:
```
model=smf.ols(formula='satisfaction_scale~Female_Business_Class',data=df
_1)
results=model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:     satisfaction_scale   R-squared:
0.084
Model:                            OLS   Adj. R-squared:
0.084
Method:                 Least Squares   F-statistic:
2365.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):
0.00
Time:                        04:19:26   Log-Likelihood:
-17467.
No. Observations:               25893   AIC:                            3.
494e+04
Df Residuals:                   25891   BIC:                            3.
495e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==================
                             coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
------------------
Intercept                  0.3574      0.003    105.285      0.000
0.351       0.364
Female_Business_Class      0.3344      0.007     48.629      0.000
0.321       0.348
==============================================================================
=======
Omnibus:                   134057.103   Durbin-Watson:
2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              2
993.248
Skew:                           0.276   Prob(JB):
0.00
Kurtosis:                       1.429   Cond. No.
2.50
==============================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

Here, from summary table we can see that t-statistic for satisfaction intercept is 105.285 and that for a female passenger traveling business class is 48.629. We can see that the p-value is 0 which in fact is not exactly 0 but very close to 0 as we have very large t-statistics.

R-squared value explains us about how much percentage change can in 'x' (female passenger traveling business class in our case) can be explained by 'y' (level of satisfaction in our case). In our case this variable is low at 8.40%. It means the regression is very skewed in our case.

F statistic tells us about the goodness of test for regression. F value is 2365. Hence, from the t-distribution table, we can see that the probability at df = 25,891 will be very close to 0. Hence, we can reject the null hypothesis stated above.

Hypothesis 2

H0: A male passenger travelling business class is satisfied with the flight experience.
HA: A male passenger travelling business class is dissatisfied with the flight experience.

In [85]:
```
male_number = df_1[(df_1['Gender'] == 'Male')]
male_number.head()
```

Out[85]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Dep tin
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
5	5	39177	Male	Loyal Customer	16	Business travel	Eco	311	3	3
8	8	27508	Male	Loyal Customer	47	Business travel	Eco	556	5	2
15	15	22470	Male	Loyal Customer	50	Personal Travel	Eco	83	3	4

5 rows × 30 columns

In [86]:
```
male_business_class = male_number[(male_number.Class == 'Business')]
male_business_class.head()
```

Out[86]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	De ti
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
20	20	63995	Male	Loyal Customer	60	Business travel	Business	612	4	4
23	23	44304	Male	Loyal Customer	25	Business travel	Business	1428	4	4
26	26	127781	Male	Loyal Customer	24	Business travel	Business	3680	4	1
28	28	121658	Male	Loyal Customer	44	Business travel	Business	1543	3	5

5 rows × 30 columns

In [87]:
```
male_business_class['satisfaction'].value_counts(normalize=True)
```

Out[87]:
```
satisfied                 0.698666
neutral or dissatisfied   0.301334
Name: satisfaction, dtype: float64
```

In [88]:
```
def male_business_class(df_1):
    if df_1['Gender'] == 'Male' and df_1['Class'] == 'Business':
        return 1
    else:
        return 0
df_1['Male_Business_Class'] = df_1.apply(male_business_class, axis = 1)
```

In [92]:
```
model=smf.ols(formula='satisfaction_scale~Male_Business_Class',data=df_1
)
results=model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:     satisfaction_scale   R-squared:
0.085
Model:                            OLS   Adj. R-squared:
0.085
Method:                 Least Squares   F-statistic:
2413.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):
0.00
Time:                        04:20:07   Log-Likelihood:
-17444.
No. Observations:               25893   AIC:                            3.
489e+04
Df Residuals:                   25891   BIC:                            3.
491e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
================
                           coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
----------------
Intercept               0.3581      0.003    106.013      0.000
0.351       0.365
Male_Business_Class     0.3406      0.007     49.126      0.000
0.327       0.354
==============================================================================
=======
Omnibus:                   135683.150   Durbin-Watson:
2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2
979.602
Skew:                           0.280   Prob(JB):
0.00
Kurtosis:                       1.435   Cond. No.
2.51
==============================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

Here, from summary table we can see that t-statistic for satisfaction intercept is 106.013 and that for a male passenger traveling business class is 49.126. We can see that the p-value is 0 which in fact is not exactly 0 but very close to 0 as we have very large t-statistics.

R-squared value explains us about how much percentage change can in 'x' (male passenger traveling business class in our case) can be explained by 'y' (level of satisfaction in our case). In our case this variable is low at 8.50%. It means the regression is very skewed in our case.

F statistic tells us about the goodness of test for regression. F value is 2413. Hence, from the t-distribution table, we can see that the probability at df = 25,891 will be very close to 0. Hence, we can reject the null hypothesis stated above.

Hypothesis 3

H0: A female passenger travelling economy class is satisfied with the flight experience.
HA: A female passenger travelling economy class is dissatisfied with the flight experience.

In [94]:
```
female_eco_class = female_number[(female_number.Class == 'Eco')]
female_eco_class.head()
```

Out[94]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depa time
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3
10	10	47583	Female	Loyal Customer	47	Business travel	Eco	1235	4	1
16	16	124915	Female	Loyal Customer	31	Business travel	Eco	728	2	5
18	18	76872	Female	Loyal Customer	43	Personal Travel	Eco	1927	3	4

5 rows × 29 columns

In [95]:
```
female_eco_class['satisfaction'].value_counts(normalize=True)
```

Out[95]:
```
neutral or dissatisfied    0.809318
satisfied                  0.190682
Name: satisfaction, dtype: float64
```

In [96]:
```
def female_eco_class(df_1):
    if df_1['Gender'] == 'Female' and df_1['Class'] == 'Eco':
        return 1
    else:
        return 0
df_1['Female_Eco_Class'] = df_1.apply(female_business_class, axis = 1)
```

In [98]:
```
model=smf.ols(formula='satisfaction_scale~Female_Eco_Class',data=df_1)
results=model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:     satisfaction_scale   R-squared:
0.084
Model:                            OLS   Adj. R-squared:
0.084
Method:                 Least Squares   F-statistic:
2365.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):
0.00
Time:                        04:28:00   Log-Likelihood:
-17467.
No. Observations:               25893   AIC:                            3.
494e+04
Df Residuals:                   25891   BIC:                            3.
495e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=============
                      coef    std err          t      P>|t|      [0.02
5      0.975]
------------------------------------------------------------------------------
-------------
Intercept            0.3574      0.003    105.285      0.000       0.35
1       0.364
Female_Eco_Class     0.3344      0.007     48.629      0.000       0.32
1       0.348
==============================================================================
=======
Omnibus:                   134057.103   Durbin-Watson:
2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2
993.248
Skew:                           0.276   Prob(JB):
0.00
Kurtosis:                       1.429   Cond. No.
2.50
==============================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

Here, from summary table we can see that t-statistic for satisfaction intercept is 105.285 and that for a female passenger traveling economy class is 48.629. We can see that the p-value is 0 which in fact is not exactly 0 but very close to 0 as we have very large t-statistics.

R-squared value explains us about how much percentage change can in 'x' (female passenger traveling economy class in our case) can be explained by 'y' (level of satisfaction in our case). In our case this variable is low at 8.40%. It means the regression is very skewed in our case.

F statistic tells us about the goodness of test for regression. F value is 2365. Hence, from the t-distribution table, we can see that the probability at df = 25,891 will be very close to 0. Hence, we can reject the null hypothesis stated above.

Hypothesis 4

H0: A male passenger travelling economy class is dissatisfied with the flight experience.
HA: A male passenger travelling economy class is satisfied with the flight experience.

In [99]:
```
male_eco_class = male_number[(male_number.Class == 'Eco')]
male_eco_class.head()
```

Out[99]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Depart time
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0
5	5	39177	Male	Loyal Customer	16	Business travel	Eco	311	3	3
8	8	27508	Male	Loyal Customer	47	Business travel	Eco	556	5	2
15	15	22470	Male	Loyal Customer	50	Personal Travel	Eco	83	3	4
21	21	75855	Male	Loyal Customer	43	Personal Travel	Eco	1437	3	4

5 rows × 30 columns

In [100]:
```
male_eco_class['satisfaction'].value_counts(normalize=True)
```

Out[100]:
```
neutral or dissatisfied    0.803456
satisfied                  0.196544
Name: satisfaction, dtype: float64
```

In [101]:
```
def male_eco_class(df_1):
    if df_1['Gender'] == 'Male' and df_1['Class'] == 'Eco':
        return 1
    else:
        return 0
df_1['Male_Eco_Class'] = df_1.apply(male_business_class, axis = 1)
```

In [102]:
```
model=smf.ols(formula='satisfaction_scale~Male_Eco_Class',data=df_1)
results=model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:     satisfaction_scale   R-squared:
0.085
Model:                            OLS   Adj. R-squared:
0.085
Method:                 Least Squares   F-statistic:
2413.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):
0.00
Time:                        04:33:10   Log-Likelihood:
-17444.
No. Observations:               25893   AIC:                             3.
489e+04
Df Residuals:                   25891   BIC:                             3.
491e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===========
                     coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
Intercept          0.3581      0.003    106.013      0.000       0.351
0.365
Male_Eco_Class     0.3406      0.007     49.126      0.000       0.327
0.354
==============================================================================
=======
Omnibus:                   135683.150   Durbin-Watson:
2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2
979.602
Skew:                           0.280   Prob(JB):
0.00
Kurtosis:                       1.435   Cond. No.
2.51
==============================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

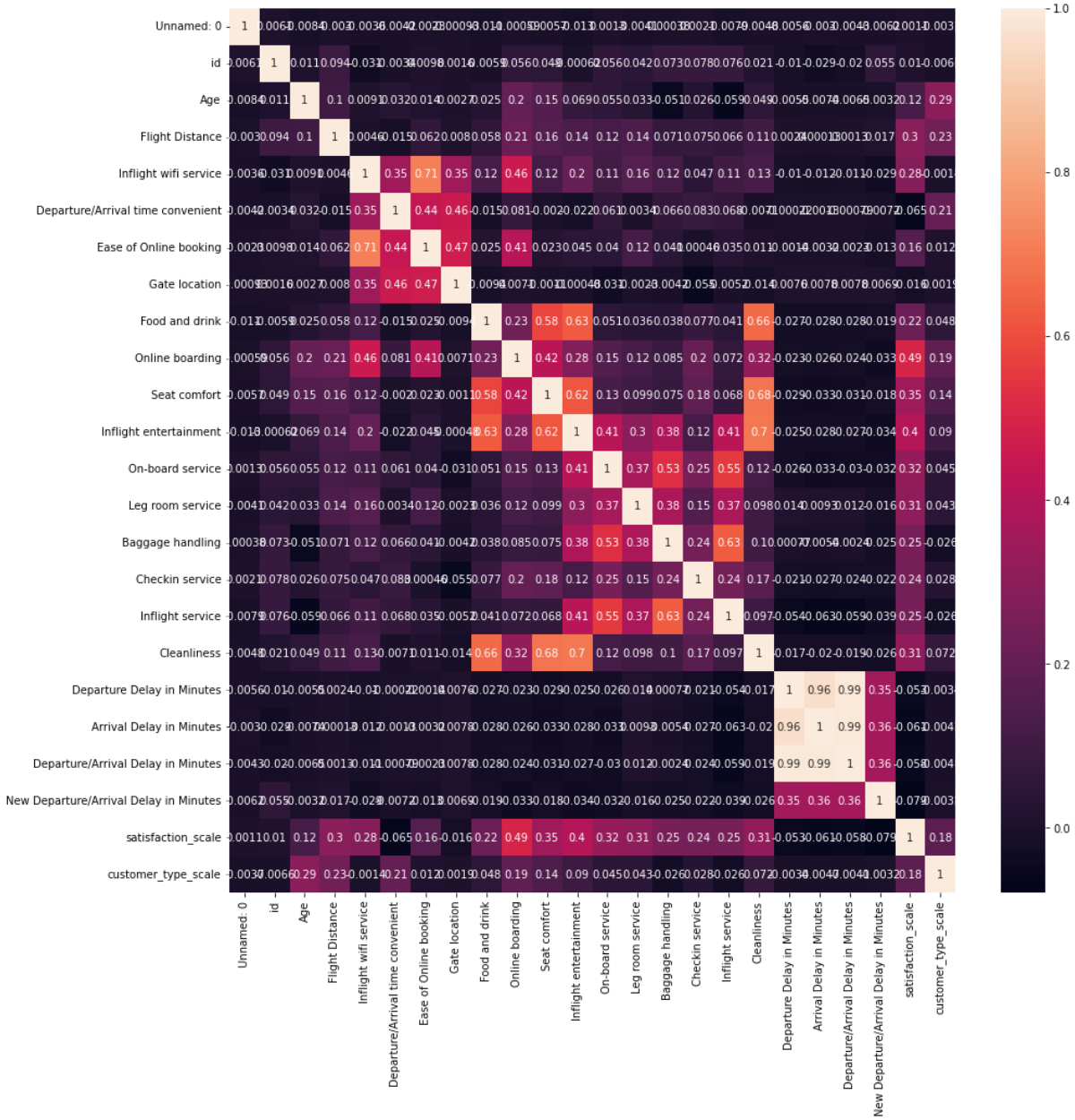
Here, from summary table we can see that t-statistic for satisfaction intercept is 106.013 and that for a male passenger traveling economy class is 49.126. We can see that the p-value is 0 which in fact is not exactly 0 but very close to 0 as we have very large t-statistics.

R-squared value explains us about how much percentage change can in 'x' (male passenger traveling economy class in our case) can be explained by 'y' (level of satisfaction in our case). In our case this variable is low at 8.50%. It means the regression is very skewed in our case.

F statistic tells us about the goodness of test for regression. F value is 2413. Hence, from the t-distribution table, we can see that the probability at df = 25,891 will be very close to 0. Hence, we can reject the null hypothesis stated above.

Final Analysis

In [65]:
```
flight_cor=df_1.corr()
flight_cor
fig, ax = plt.subplots(figsize=(15,15))
sns.heatmap(flight_cor, annot=True)
plt.show()
```



We can see that there is significantly strong correlation between cleanliness and food and drink; cleanliness and inflight entertainment; leg room and inflight service. Hence, we can conclude that the airline has to concentrate more on cleanliness, inflight service and food and drink service majorly to increase the number of satisfied passengers.

3. Conclusion

The summary of findings:

We saw that we rejected the null hypothesis in all the scenarios. Overall we can conclude that the passengers are not statified with the overall flight experience.

Especially keeping the COVID situation in mind, following all the protocols, the airline needs to work hard on improving the flight experience.

Business implications for audience:

The individual business implications have been addressed with the particular business and analysis questions raised. For an overall analysis, we feel that the criteria for classifying a passenger as loyal / disloyal needs to be revisited. Also, the individual suggestions given need to be considered by the airline.

Limitations of this project

Some of the variables are left unrated by the passengers. It cannot be established whether this act was arbitrary or if the customers did not have any opinion.

We cannot establish if they type of customer as rated by the airline has an impact on the ratings given by the passenger.

The data set has three types of satisfaction criteria: satisfied, neutral or dissatisfied. However, dissatisfied, and neutral are combined. As per our logic, the possibility of a neutral passenger using the airline again for travel is more than a dissatisfied passenger. The data set should have been made keeping neutral and dissatisfied criteria separate.

Outliers; especially in the 'Departure Delay in Minutes' and 'Arrival Delay in Minutes' columns are mainly responsible for the skewness for the normal distribution. Also, we believe that since both these columns have a significant impact on the satisfaction rating, we cannot rule out the possibility of an element of biasness in the final satisfaction rating

The location of gate is not always decided by the airline. Other than the long term leased gates (the airline has the autonomy to use these gates), the airline sometimes must use common gates (the airport authority decides the allocation). Hence, we cannot suggest any prescriptive analytics for a lower rating for convenience of 'Gate location' as it may not be implementable.(https://www.quora.com/How-do-airports-assign-flights-to-gates (https://www.quora.com/How-do-airports-assign-flights-to-gates))

Potential project by adopting advanced analytics techniques such as predictive and prescriptive analytics

The same is suggested along with the individual analysis questions in form of suggestions. The respective managers may work on the individual suggestions based on statistical analysis and try to implement the prescriptive measures suggested.