Data Intensive Computing CSE 587

Project Phase 1

TITLE: FLIGHT SATISFACTION CLASSIFICATION

Name	**UB id**
Sriinitha Chinnapatlola	sriinith
Shravani Soma	ssoma

1. Problem Statement:

We aim to consider various factors in predicting the passenger satisfaction for their flight. These factors include flight details like flight distance, inflight wifi services, food and drinks, seat comfort, departure and arrival delays, and many more. We also take customer's details into consideration for evaluating the model. These considerations include gender, age, and their loyalty. This results in a classification model which predicts if a customer would either be 'satisfied' or 'neutral or dissatisfied' based on all the above given factors.

(i) Background:

In a highly competitive business market, airline companies continuously monitor for passengers' feedback in improving their services. They aim to meet their passengers' expectations and in constantly improving their services. Collecting feedback from passengers through various platforms is much needed for in order to make this analysis.

This analysis could result in making satisfactory changes in various aspects like increasing or decreasing the flight ticket prices, infrastructure of the flight, changes in the food and inflight services. Overall, the final analysis would help these companies understand which factors have a major impact on their services and which features do not have much importance to build their services accordingly.

Airline companies must sometimes reward their customers for providing the feedback in order to get a vast range of reviews. These factors that impact airlines' reputation change with time. As thousands of airline companies exist in market today, passengers tend to look for better services each day. They have a wide scope in looking for better services. So, it is very important for companies to retain passengers' loyalty with time and to update their services with time. Feedback pamphlets after the flight, or mails for feedback forms could be sent for immediate and honest responses.

(ii) Potential/ Scope:

In improving the services after analyzing their customers' reviews, companies would build their reputation and build a huge market among various existing companies. This could help them with holding onto their customers' loyalty and building word of mouth. This on whole, would lead airline companies to in building up their game and winning customer's trust.

2. Dataset:

Data is collected from Kaggle, one of the best platforms providing for huge datasets. We have a dataset consisting of 25 columns and 100 thousand plus rows. We perform analysis on this dataset to look for factors impacting the passenger's travel. Following are the attributes in our dataset.

These attributes are:

1. id: id for each passenger
2. satisfaction: two categories – 'satisfied' and 'neutral or dissatisfied'
3. Gender: two categories – 'Male' and 'Female'
4. Arrival Delay in Minutes: delay time during arrival
5. Departure Delay in Minutes: delay time during departure
6. Customer Type: says whether a customer is loyal or disloyal to the airlines
7. Age: age of the passenger.
8. Cleanliness: cleanliness rating on a scale of 1 to 5
9. Type of Travel: says whether it is a personal or business travel
10. Class: Specifies the gender of insured
11. Baggage Handling: Specifies the age of insured
12. Checkin service: checkin service rating on a scale of 1 to 5
13. Flight Distance: travel distance for the flight
14. Inflight wifi service: wifi services rating for the flight on a scale of 0 to 5
15. Inflight service: flight services rating on a scale of 1 to 5
16. Departure/Arrival time convenient: rating for timing convenience with scale 0 to 5
17. Ease of Online booking: rating for ease of booking on a scale of 0 to 5
18. Gate location: satisfactory rating for gate location on a scale of 1 to 5
19. Seat comfort: satisfactory rating for seat comfort on a scale of 1 to 5
20. Online boarding: satisfactory rating for online boarding on a scale of 1 to 5
21. Food and drink: satisfactory rating for food and drinks on a scale of 1 to 5
22. Inflight entertainment: satisfactory rating for entertainment on a scale of 1 to 5
23. On-board service: satisfactory rating for on-board services on a scale of 1 to 5
24. Leg room service: satisfactory rating for online leg room services on a scale of 1 to 5

3. Data Preprocessing/Cleaning:

We preform preprocessing on the data to convert the vague data to a much-structured format. This could be done by identifying errors and rectify them with the apt methods. Cleaning/ Preprocessing needs to be performed before giving the data to the model for the better and much accurate results. A wide range of actions could be performed depending on the data to amplify the results.

Following are the methods involved in cleaning of our data:

1. Identify the problem statement and collecting data
2. Loading data to the environment and importing libraries
3. Describing the statistics of the data
4. Dropping the unnecessary columns
5. Identifying the missing values
6. Filling the missing values in the dataset
7. Removing duplicate rows
8. Initial Choosing of feature set to remove highly correlated features
9. Class imbalance
10. Removing outliers
11. Encoding for categorical columns
12. Normalizing data

1. Identify the problem statement and collecting data:

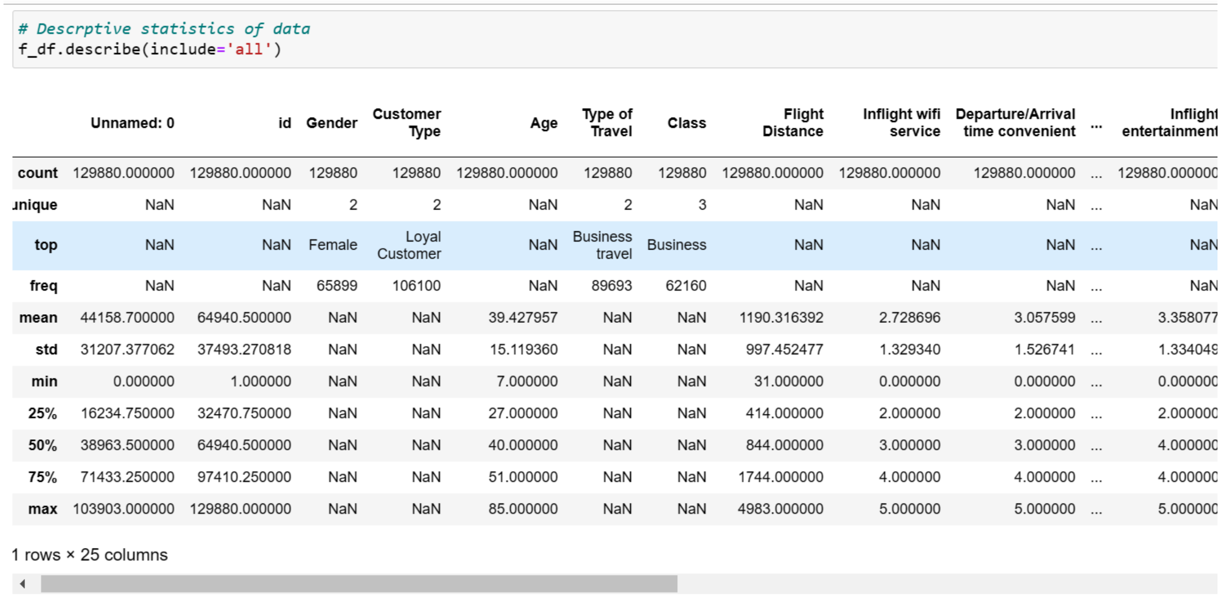
The problem statement should be in relevance to the data collected. Identifying the right dataset for the problem statement is the most important step which gives the accurate results. Better understanding of the problem statement and the apt data with the required features which gives the most significant results for the problem is the vital step. We have collected the dataset from Kaggle consisting of 25 columns and 100 thousand plus rows.

2. The dataset and libraries are imported:

We must import all the required libraries that are used throughout the project. All these libraries can be imported using the keyword ‘import’. We have concatenated the train and test csv files into a single data frame. ‘read_csv’ is the method to import train and test csv files. We have clubbed these data frames into a single data frame.

3. Describing the statistics of the data:

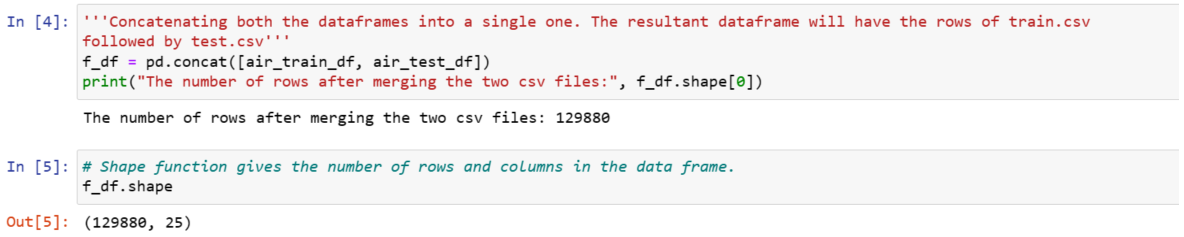
‘describe’ method is used to give the statistics of the data. ‘shape’ is used to give the number of rows and columns.



```
# Descrptive statistics of data
f_df.describe(include='all')
```

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment
count	129880.000000	129880.000000	129880	129880	129880.000000	129880	129880	129880.000000	129880.000000	129880.000000	...	129880.000000
unique	NaN	NaN	2	2	NaN	2	3	NaN	NaN	NaN	...	NaN
top	NaN	NaN	Female	Loyal Customer	NaN	Business travel	Business	NaN	NaN	NaN	...	NaN
freq	NaN	NaN	65899	106100	NaN	89693	62160	NaN	NaN	NaN	...	NaN
mean	44158.700000	64940.500000	NaN	NaN	39.427957	NaN	NaN	1190.316392	2.728696	3.057599	...	3.358077
std	31207.377062	37493.270818	NaN	NaN	15.119360	NaN	NaN	997.452477	1.329340	1.526741	...	1.334049
min	0.000000	1.000000	NaN	NaN	7.000000	NaN	NaN	31.000000	0.000000	0.000000	...	0.000000
25%	16234.750000	32470.750000	NaN	NaN	27.000000	NaN	NaN	414.000000	2.000000	2.000000	...	2.000000
50%	38963.500000	64940.500000	NaN	NaN	40.000000	NaN	NaN	844.000000	3.000000	3.000000	...	4.000000
75%	71433.250000	97410.250000	NaN	NaN	51.000000	NaN	NaN	1744.000000	4.000000	4.000000	...	4.000000
max	103903.000000	129880.000000	NaN	NaN	85.000000	NaN	NaN	4983.000000	5.000000	5.000000	...	5.000000

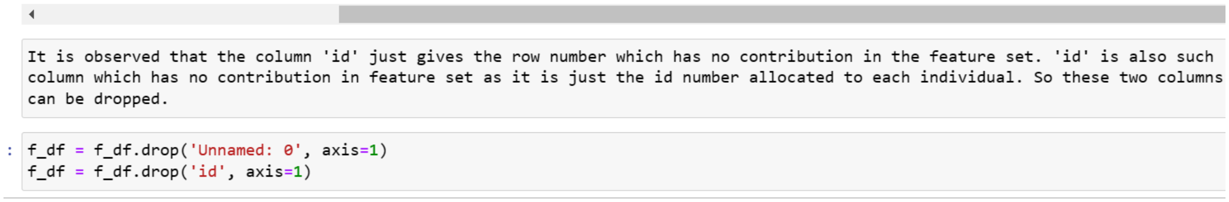
1 rows × 25 columns



```
In [4]: '''Concatenating both the dataframes into a single one. The resultant dataframe will have the rows of train.csv
        followed by test.csv'''
        f_df = pd.concat([air_train_df, air_test_df])
        print("The number of rows after merging the two csv files:", f_df.shape[0])

        The number of rows after merging the two csv files: 129880

In [5]: # Shape function gives the number of rows and columns in the data frame.
        f_df.shape

Out[5]: (129880, 25)
```

4. Dropping the unnecessary columns:

It is observed that the column 'id' just gives the row number which has no contribution in the feature set. 'id' is also such column which has no contribution in feature set as it is just the id number allocated to everyone. So, these two columns can be dropped.

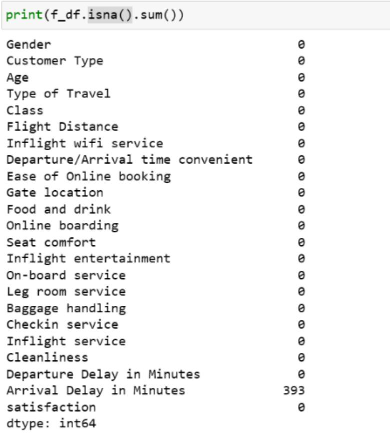


```
It is observed that the column 'id' just gives the row number which has no contribution in the feature set. 'id' is also such
column which has no contribution in feature set as it is just the id number allocated to each individual. So these two columns
can be dropped.

: f_df = f_df.drop('Unnamed: 0', axis=1)
  f_df = f_df.drop('id', axis=1)
```

5. Identifying the missing values:

There are 393 (129880 - 129487) rows that must have even a single null value. We have looked at these columns that are contributing for these null values using ‘isna’ function.

Handling these missing values is important in obtaining good accuracies.

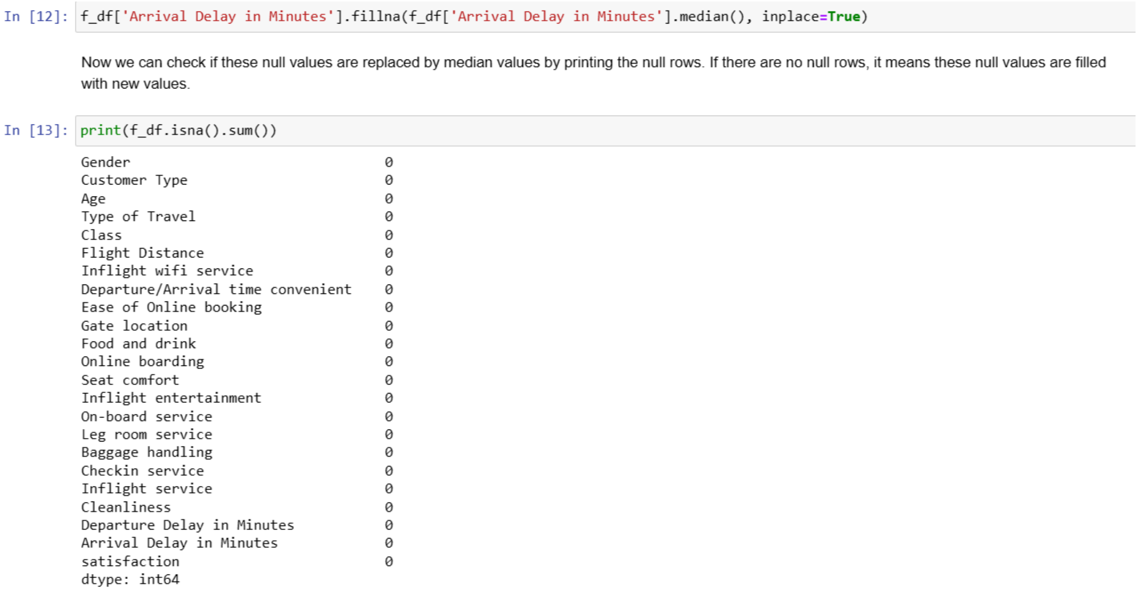


```
print(f_df.isna().sum())
```

```
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             393
satisfaction                           0
dtype: int64
```

6. Filling the missing values in the dataset:

It is observed that "Arrival Delay in Minutes" is the only column that is contributing for these null values. Now we can look for the above statistics in order to know how to handle these values. These values can be filled in with the median values of the column instead of dropping the rows which is usually the opted method. 'fillna' function is used to fill the median values.



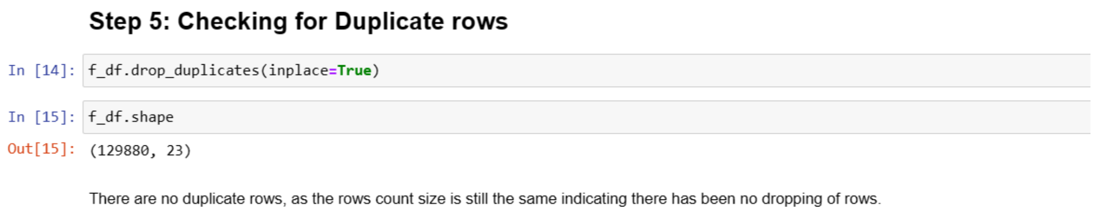
```
In [12]: f_df['Arrival Delay in Minutes'].fillna(f_df['Arrival Delay in Minutes'].median(), inplace=True)
```

Now we can check if these null values are replaced by median values by printing the null rows. If there are no null rows, it means these null values are filled with new values.

```
In [13]: print(f_df.isna().sum())
```

```
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
```

7. Removing Duplicate rows:



Step 5: Checking for Duplicate rows

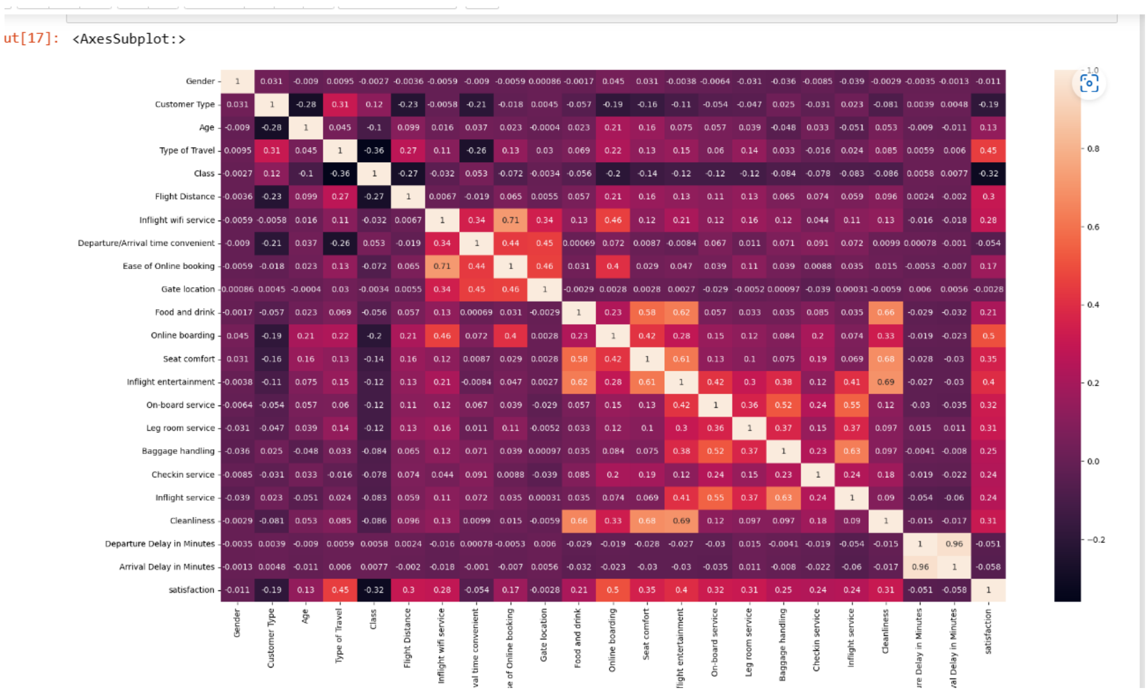
```
In [14]: f_df.drop_duplicates(inplace=True)

In [15]: f_df.shape

Out[15]: (129880, 23)
```

There are no duplicate rows, as the rows count size is still the same indicating there has been no dropping of rows.

‘drop_duplicates’ is the function to remove the duplicate rows from the dataframe. There are no duplicate rows, as the rows count size is still the same indicating there has been no dropping of rows.

8. Initial choosing of feature set to remove highly correlated features:

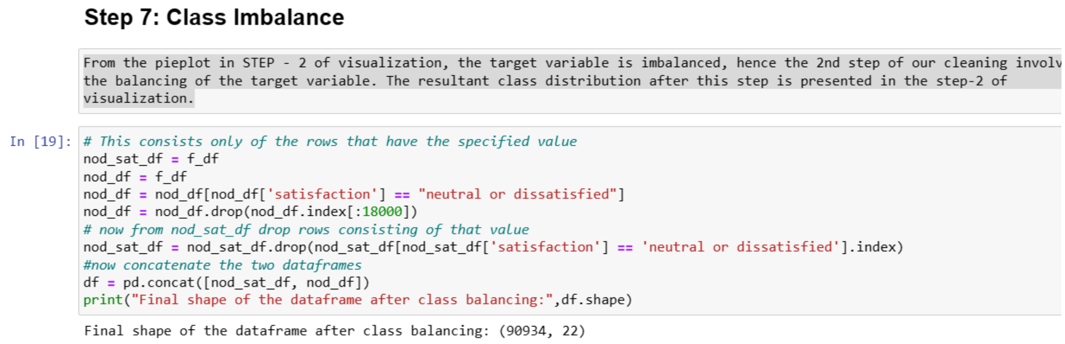
'Departure Delay in Minutes' and 'Arrival Delay in Minutes' are two columns with a very high correlation value of 0.96 (96%). We can drop any of the two columns. 'Departure Delay in Minutes' column is dropped. ‘heatmap’ function from seaborn library is used to plot this graph.



```
ut[17]: <AxesSubplot:>
```

9. Class imbalance:

From the pie chart in step - 2 of visualization, the target variable is imbalanced, hence the 2nd step of our cleaning involves the balancing of the target variable. The resultant class distribution after this step is presented in the step-3 of visualization.



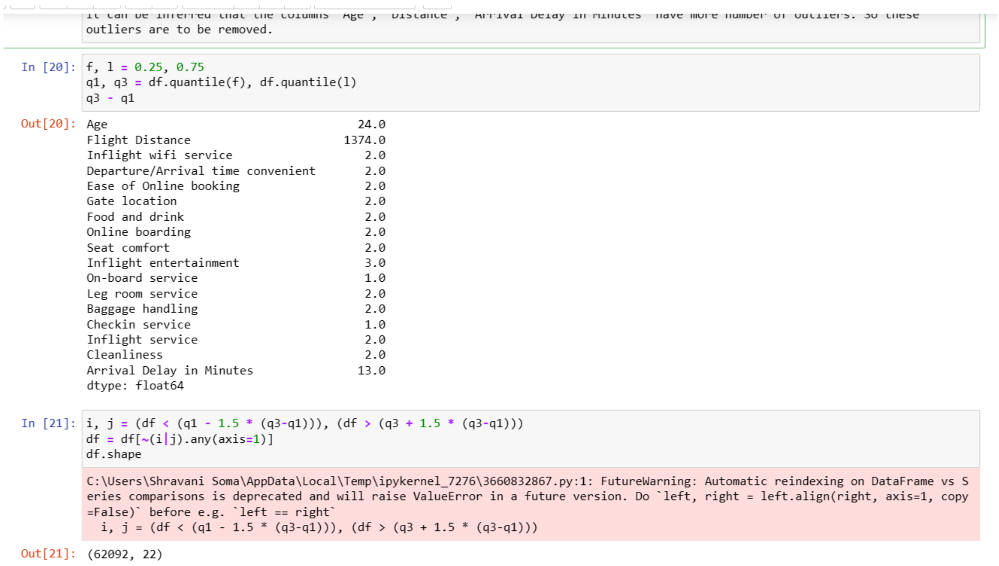
Step 7: Class Imbalance

From the pieplot in STEP - 2 of visualization, the target variable is imbalanced, hence the 2nd step of our cleaning involv
the balancing of the target variable. The resultant class distribution after this step is presented in the step-2 of visualization.

```
In [19]: # This consists only of the rows that have the specified value
         nod_sat_df = f_df
         nod_df = f_df
         nod_df = nod_df[nod_df['satisfaction'] == "neutral or dissatisfied"]
         nod_df = nod_df.drop(nod_df.index[:18000])
         # now from nod_sat_df drop rows consisting of that value
         nod_sat_df = nod_sat_df.drop(nod_sat_df[nod_sat_df['satisfaction'] == 'neutral or dissatisfied'].index)
         #now concatenate the two dataframes
         df = pd.concat([nod_sat_df, nod_df])
         print("Final shape of the dataframe after class balancing:",df.shape)

         Final shape of the dataframe after class balancing: (90934, 22)
```

10. Removing Outliers

We have used IQR method in removing outliers from the data. It can be inferred that the columns 'Age', 'Distance', 'Arrival Delay in Minutes' have a greater number of outliers. So, these outliers are to be removed.

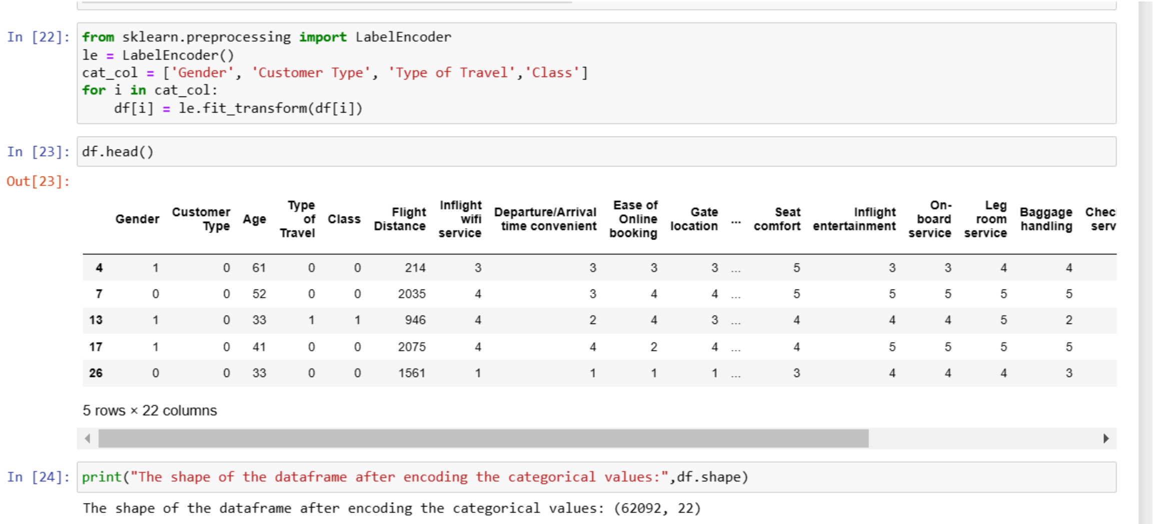


It can be inferred that the columns 'Age', 'Distance', 'Arrival Delay in Minutes' have more number of outliers. So these outliers are to be removed.

```
In [20]: f, l = 0.25, 0.75
         q1, q3 = df.quantile(f), df.quantile(l)
         q3 - q1

Out[20]: Age                                  24.0
         Flight Distance                    1374.0
         Inflight wifi service                 2.0
         Departure/Arrival time convenient     2.0
         Ease of Online booking                2.0
         Gate location                         2.0
         Food and drink                        2.0
         Online boarding                       2.0
         Seat comfort                          2.0
         Inflight entertainment                3.0
         On-board service                      1.0
         Leg room service                      2.0
         Baggage handling                      2.0
         Checkin service                       1.0
         Inflight service                      2.0
         Cleanliness                           2.0
         Arrival Delay in Minutes             13.0
         dtype: float64

In [21]: i, j = (df < (q1 - 1.5 * (q3-q1))), (df > (q3 + 1.5 * (q3-q1)))
         df = df[~(i|j).any(axis=1)]
         df.shape

         C:\Users\Shravani Soma\AppData\Local\Temp\ipykernel_7276\3660832867.py:1: FutureWarning: Automatic reindexing on DataFrame vs S
         eries comparisons is deprecated and will raise ValueError in a future version. Do `left, right = left.align(right, axis=1, copy
         =False)` before e.g. `left == right`
           i, j = (df < (q1 - 1.5 * (q3-q1))), (df > (q3 + 1.5 * (q3-q1)))

Out[21]: (62092, 22)
```

11. Encoding for Categorical Columns

As we have categorical data (text) in our dataset, before we send the data to the model for training, we have to convert the non-numeric data to numeric data. This can be done using label encoding or one-hot encoding. This step of cleaning involves converting non-numerical data to numeric data using encoding.



```
In [22]: from sklearn.preprocessing import LabelEncoder
         le = LabelEncoder()
         cat_col = ['Gender', 'Customer Type', 'Type of Travel','Class']
         for i in cat_col:
             df[i] = le.fit_transform(df[i])

In [23]: df.head()
Out[23]:
```

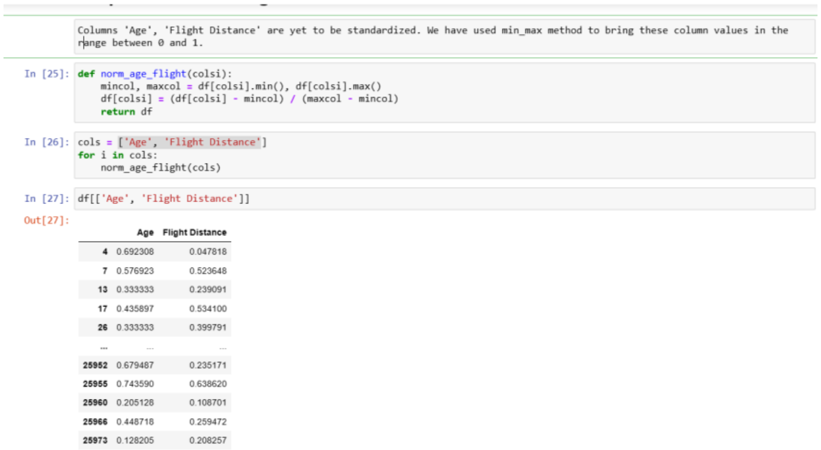
	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	...	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Chec serv
4	1	0	61	0	0	214	3	3	3	3	...	5	3	3	4	4	
7	0	0	52	0	0	2035	4	3	4	4	...	5	5	5	5	5	
13	1	0	33	1	1	946	4	2	4	3	...	4	4	4	5	2	
17	1	0	41	0	0	2075	4	4	2	4	...	4	5	5	5	5	
26	0	0	33	0	0	1561	1	1	1	1	...	3	4	4	4	3	

5 rows × 22 columns

```
In [24]: print("The shape of the dataframe after encoding the categorical values:",df.shape)

         The shape of the dataframe after encoding the categorical values: (62092, 22)
```

12. Normalizing Data:

Columns 'Age', 'Flight Distance' are yet to be standardized. We have used min_max method to bring these column values in the range between 0 and 1.



Columns 'Age', 'Flight Distance' are yet to be standardized. We have used min_max method to bring these column values in the range between 0 and 1.

```
In [25]: def norm_age_flight(colsi):
             mincol, maxcol = df[colsi].min(), df[colsi].max()
             df[colsi] = (df[colsi] - mincol) / (maxcol - mincol)
             return df

In [26]: cols = ['Age', 'Flight Distance']
         for i in cols:
             norm_age_flight(cols)

In [27]: df[['Age', 'Flight Distance']]
Out[27]:
```

	Age	Flight Distance
4	0.692308	0.047818
7	0.576923	0.523648
13	0.333333	0.239091
17	0.435897	0.534100
26	0.333333	0.399791
...	...	...
25952	0.679487	0.235171
25955	0.743590	0.638620
25960	0.205128	0.108701
25966	0.448718	0.259472
25973	0.128205	0.208257

EXPLORATORY DATA ANALYSIS

To get better insights into our data and draw meaningful conclusions from the data, we perform EDA on our data. This EDA can be done either to understand the next cleaning step or to understand the relationships between various categorical and non-categorical features in our dataset.

1. Violin plot: To replace missing values.

We identified that there are missing values in our dataset. Now these values should be replaced either with the mean or median of the column. To decide if we should replace the missing values with mean or median we plot a Violin plot to understand the distribution of data.

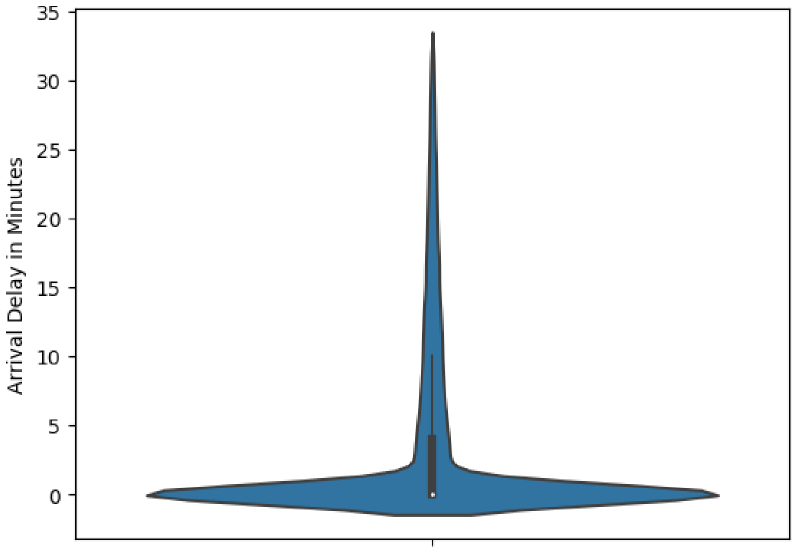


Fig: Violin plot of the column that has missing values.

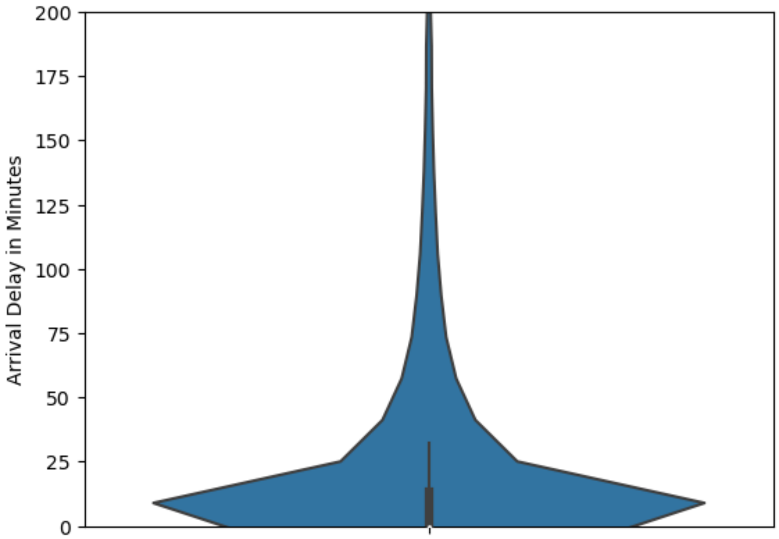


Fig: A closer look at the values distribution

After analyzing the violin plot, since the columns consist of a lot of outliers, it is suitable to replace the missing values with the median rather than the mean. This has been done in the cleaning step of filling the missing values.

2. PIE chart to find if the class is imbalanced:

To see if our data is imbalanced to understand the target value (satisfaction) distribution. 56.6% of the target values belong to the class “neutral or dissatisfied” whereas “ 43.4% of the values belong to the class “satisfied”. To get better training accuracy, we can perform class balancing and the same has been performed as one of the cleaning steps.

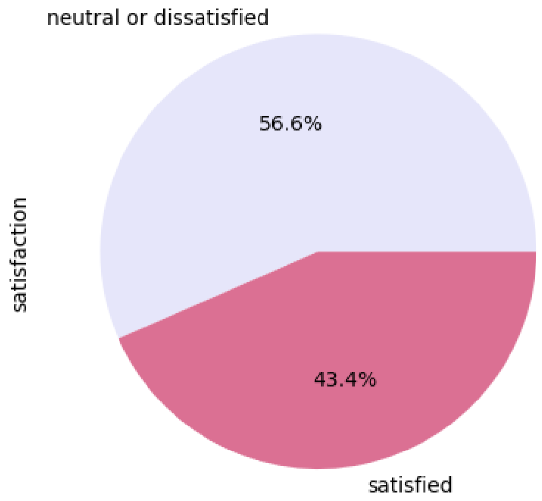


Fig: Target value distribution before class balancing.

3. PIE chart to depict the balanced class: After making the class balanced, we plot a pie chart to see the distribution of the target variable.

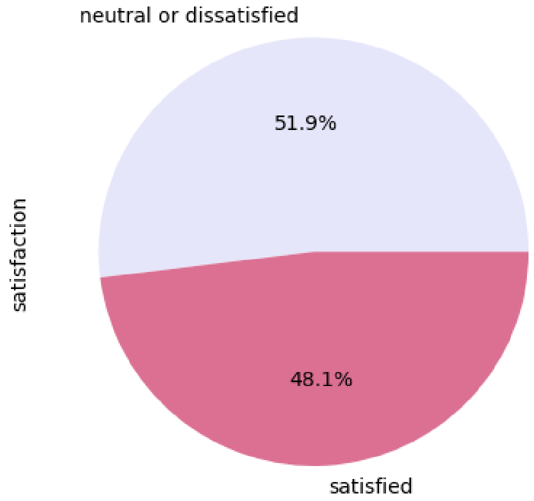
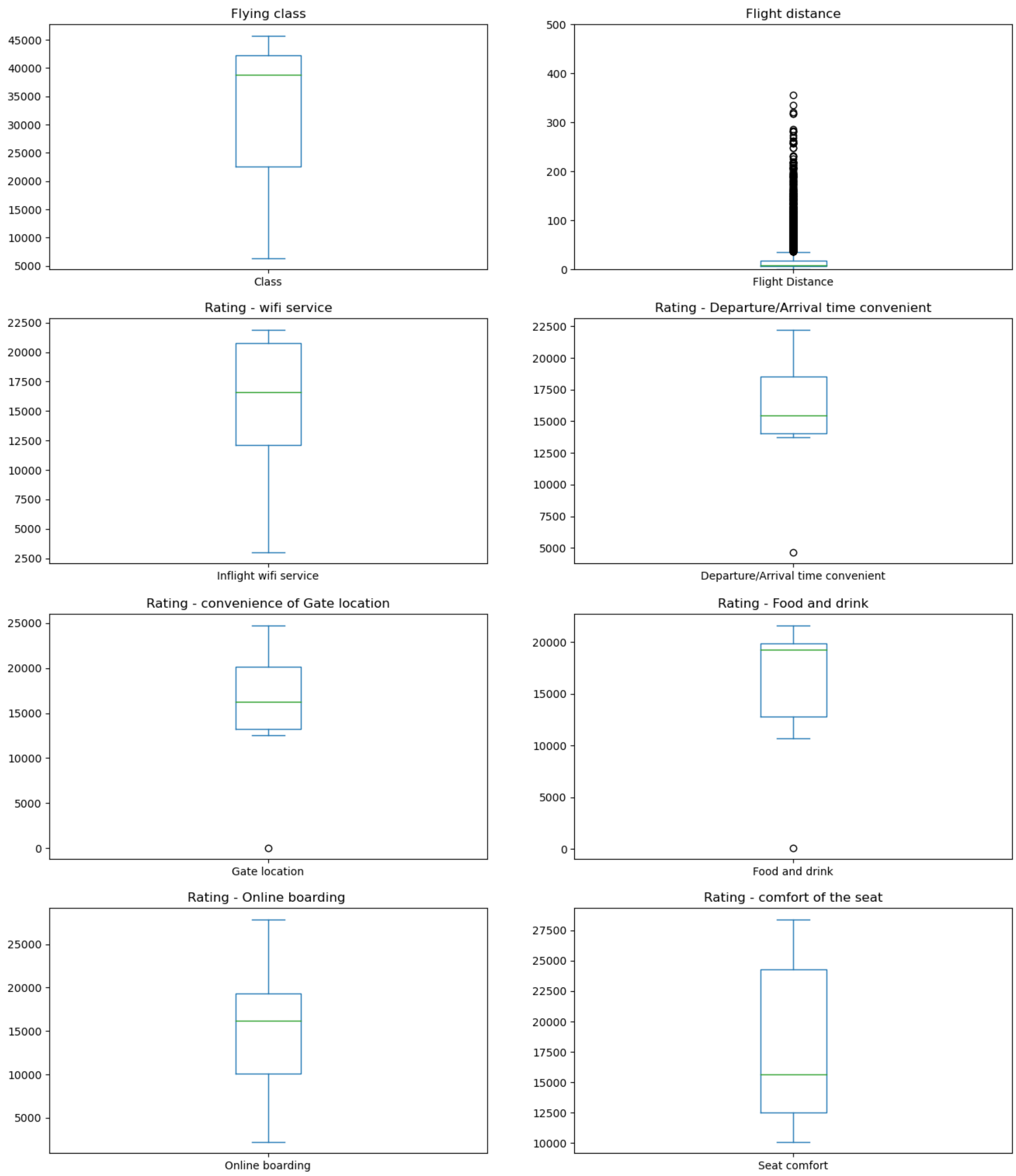
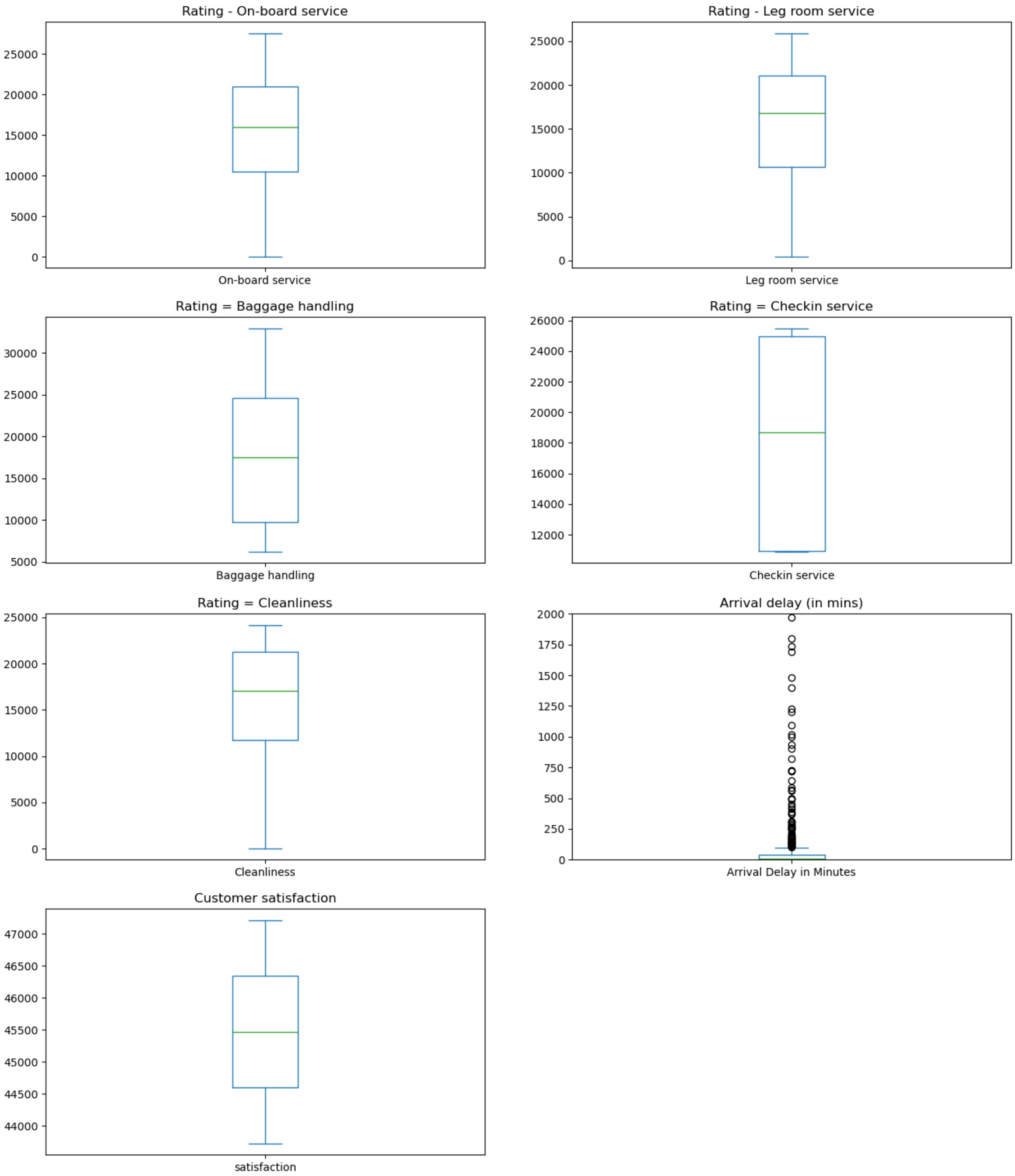


Fig: Target value distribution after class balancing.

4. Boxplots to identify outliers: Before we normalize and scale the data, it is essential to identify the outliers in the data. We plot boxplots for each of the features to identify the outliers and eliminate them. From the below plots, the features “Flight distance”, “Arrival delay in minutes” have outliers. The handling of outliers has been done as one of the cleaning steps.




Rating - On-board service
0
5000
10000
15000
20000
25000
On-board service
Rating - Leg room service
0
5000
10000
15000
20000
25000
Leg room service
Rating = Baggage handling
5000
10000
15000
20000
25000
30000
Baggage handling
Rating = Checkin service
12000
14000
16000
18000
20000
22000
24000
26000
Checkin service
Rating = Cleanliness
0
5000
10000
15000
20000
25000
Cleanliness
Arrival delay (in mins)
0
250
500
750
1000
1250
1500
1750
2000
Arrival Delay in Minutes
Customer satisfaction
44000
44500
45000
45500
46000
46500
47000
satisfaction

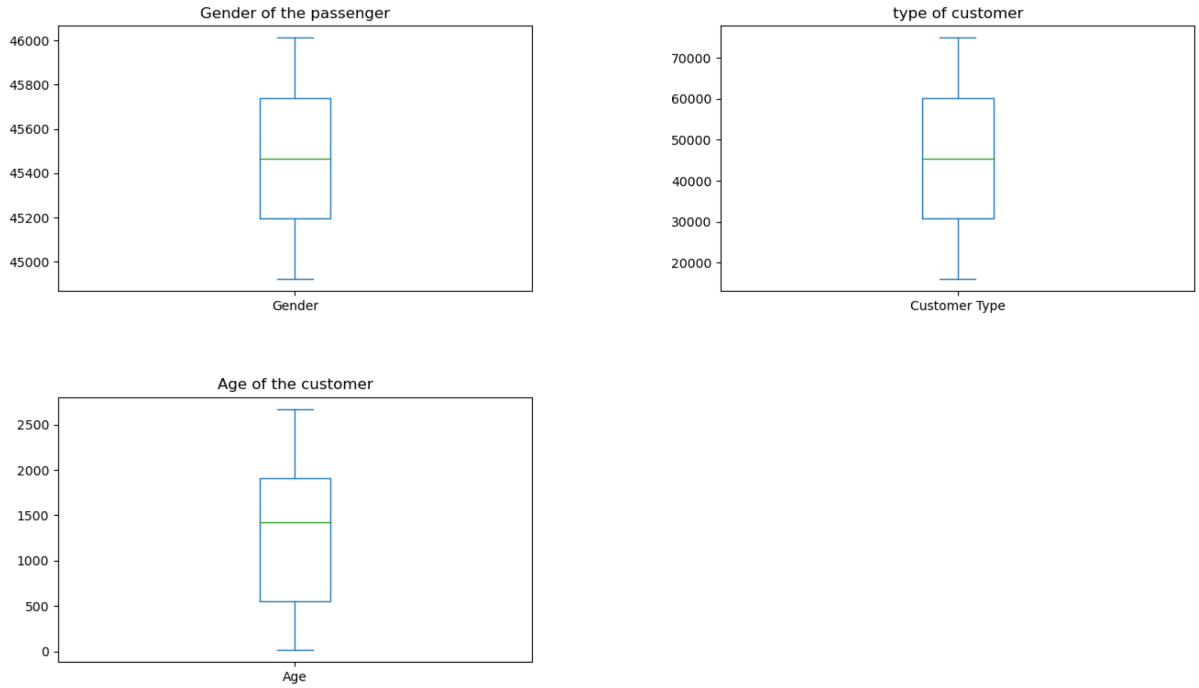


Fig: Boxplots of each of the feature to identify outliers

5. Boxplots after removing outliers: As we identified outliers in two of the columns in our dataset necessary cleaning has been done and the below box plots depict the data distribution of the same columns.

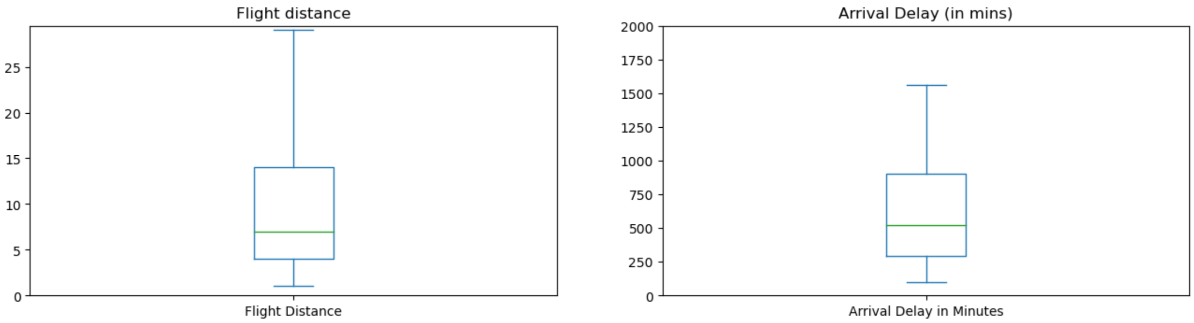


Fig: Box plots after removing outliers from the features that earlier had outliers.

UNIVARIATE ANALYSIS:

1. **Univariate analysis of Gender, customer type, age, type of travel.**

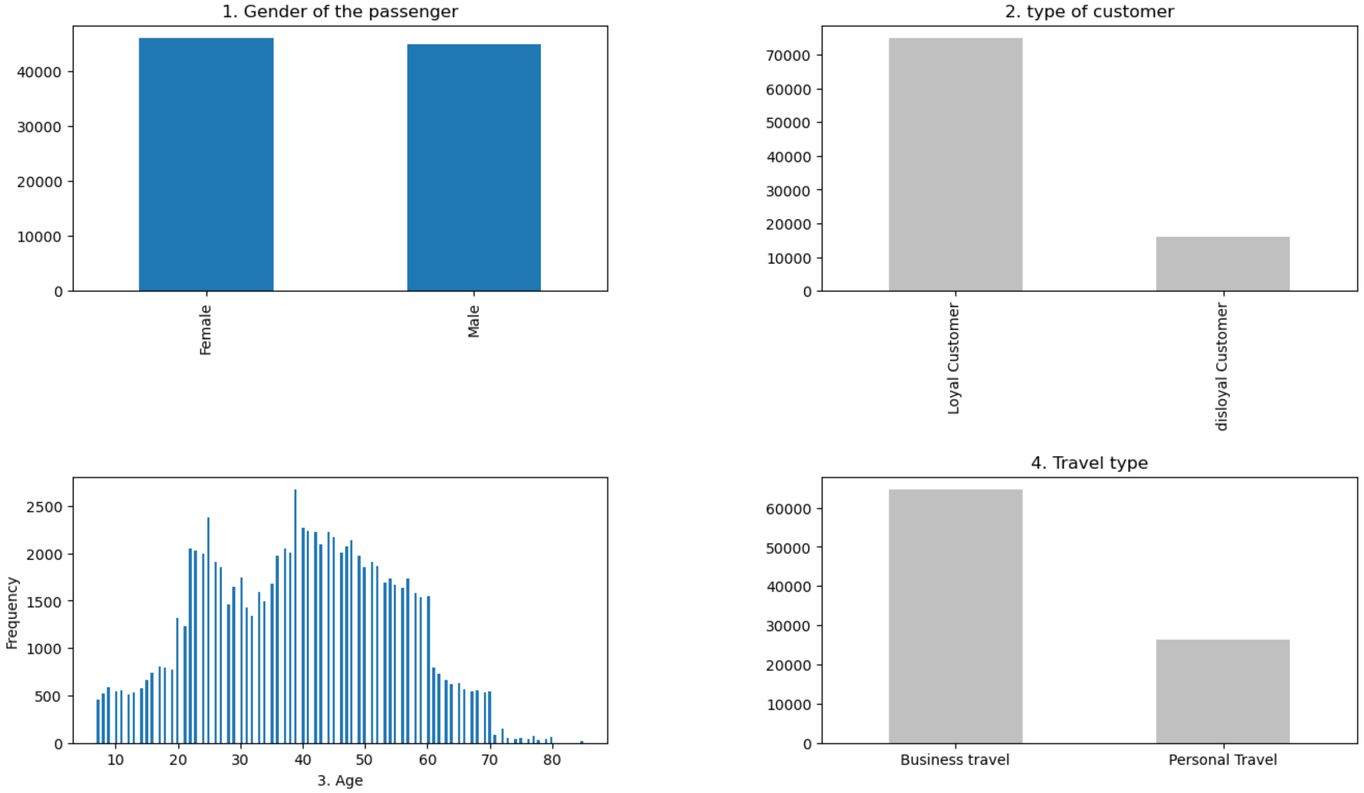


Fig: bar plots of gender, customer type, travel type and histogram of age.

Analysis:

In univariate analysis we try to understand the behavior of each variable separately and derive conclusions about the same.

In our airline passenger satisfaction dataset various EDA and cleaning steps have been performed simultaneously and now we try to analyze the features one at a time. For our dataset, bar plots are the most suitable ones to give an insight about each feature. As some of our columns have values that are more scattered in nature, we choose to plot histogram for the same for a better understanding.

Here is our analysis from the univariate analysis of our dataset:

1. Gender: In the gender feature of our dataset, it can be inferred that the number of female customers are slightly higher than the number of male customers. In our bivariate analysis we can discover some interesting facts by plotting this feature against the target variable.

2. Type of Customers: From this plot it can inferred that the number of loyal customers are significantly higher than the number of disloyal customers and this shows that most of the customers liked the airline's services and travelled with them again.

3. Age: In our dataset, the values in the age category are scattered and it also kind of makes sense as the travelers will be of different age groups. Bar plot would not be an ideal plot for this kind of data hence we plotted a histogram to make some conclusions. From the histogram plot, we concluded that highest number of the airline travelers are aged around 38-40 years.

4. Travel type: From this bar plot if can be inferred that most of the travelers travel for business purposes. This helps the airlines to come up with more business strategies designed for business travelers as majority of their traveler are from this sector.

2. Univariate analysis of flying class, flying distance, wifi-service, departure/arrival time convenience:

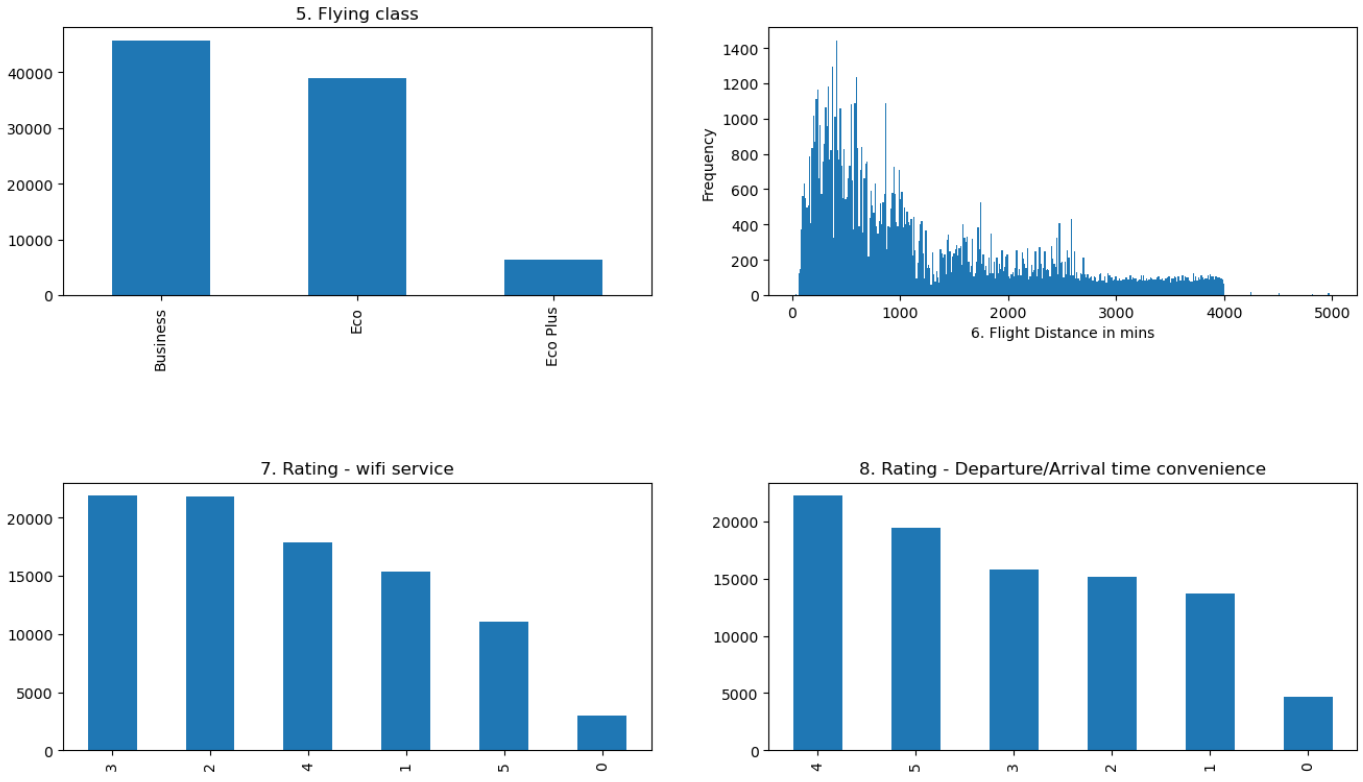


Fig: bar plots of flying class, Wi-Fi-service, departure/arrival time and histogram of flight distance.

Analysis:

5. Flying class: This bar plots gives the number of customers travelling in various flying classes like business, eco and eco plus. A conclusion that can we made from this is very few people are preferring to go for eco plus this might be because there might be no much difference in terms of the services offered in eco plus in comparison to eco and the cost might be considerably higher than Eco. For a traveler in such a case, it might not be feasible to consider "eco plus" over "eco".

6. Flight Distance: For this feature, we have plotted a histogram as the values are scattered. It can be inferred that most of the flights have flying time around 0-1000 mins. Later in the bivariate analysis we can plot this variable against various other features to understand if it's dependent on other features.

7. Wi-Fi Service: This plot is an essential plot as most of us expect a proper wifi connection during our travel time. From the bar plot for this feature, it can be inferred that most of them rated the wifi services a 3/5 which indicates the airplanes for an improvement in the same.

8. Departure/Arrival time convenience: This plot helps the airlines in understanding the flight timings they schedule and chance the time according to the convenience of the customers depending on the rating.

3. Univariate analysis of Gate location, Food and drink, Online boarding, Seat comfort:

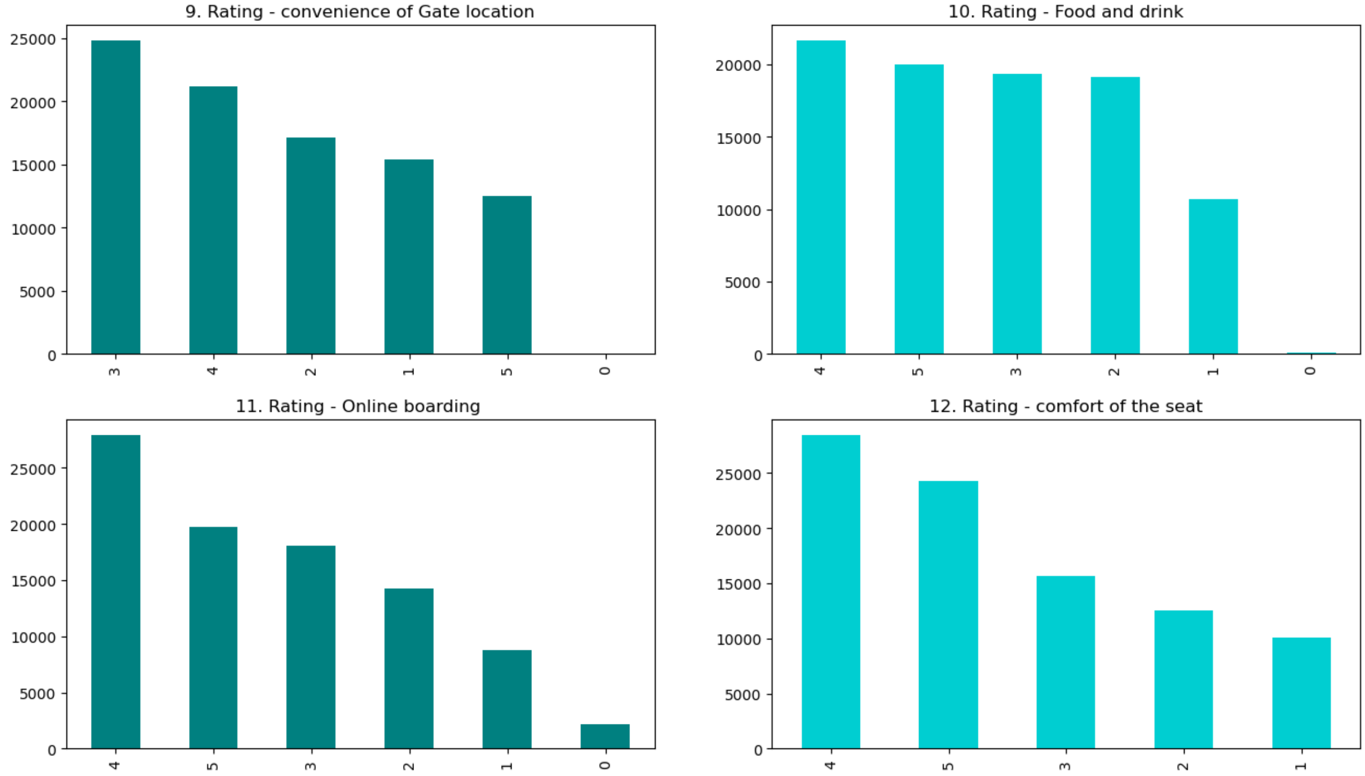


Fig: bar plots of Gate location, Food and drink, Online boarding, Seat comfort

Analysis:

9. Convenience of Gate: How convenient it is for the customer to get to the boarding gate is also important as it's tiresome to find the gate in big airports. Depending on the customer ratings, the airlines can make some changes in the same to increase the number of customers. As most of them rated a 3/5, it means for most of the travelers it's not very convenient to keep searching for the gate.

10. Food and drinks: From the plot the rating is almost similar as the number of customers who gave the rating of 4,5,3,2 are almost equal. This indicates that a significant number of customers are extremely satisfied while a significant amount of them are dissatisfied too. This gives the airlines a chance to improve their food and drink services.

11. Online Boarding: It is very convenient for a traveler to complete the boarding if there's no check in luggage to avoid the long queues at the airport. As most of them rated the online boarding a 4/5, it is understandable that most of the customers use online boarding, and they are satisfied with the service delivered by the airlines too.

12. Comfort of the seat: This can be an essential feature as travelers who are flying long distances would prefer comfortable seats. We can further plot some bivariate plots to see if the comfort of seats, flying class, and flying distance have any relationship.

4. Univariate analysis of On-board service, Leg room service, Baggage handling, Check in service

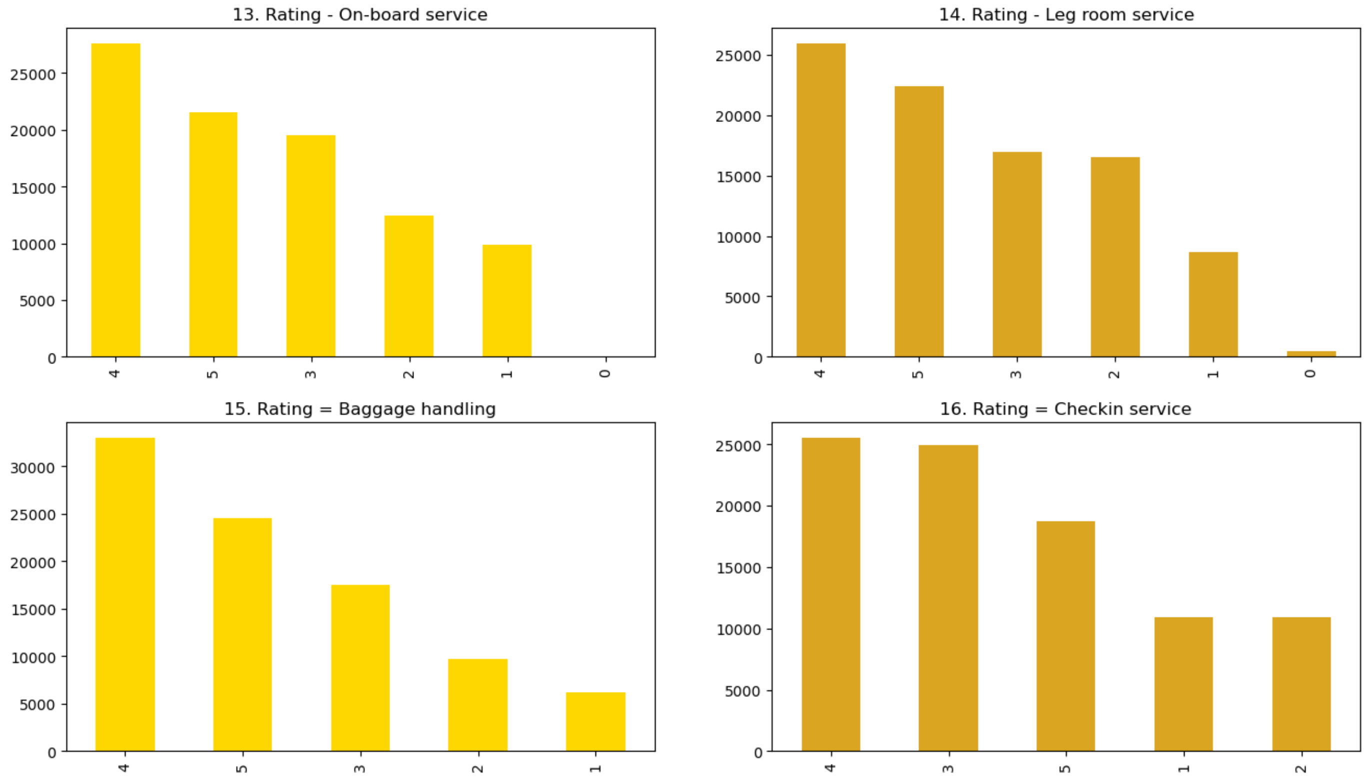


Fig: bar plots of On-board service, Leg room service, Baggage handling, Check in service

Analysis:

13. On- board Service: knowing how efficient the on-boarding service offered by the airlines helps them to improve and train their crew so as to get maximum customer satisfaction. From the bar graph, it can be inferred that most of the travelers are very much satisfied with the services offered as most of them offered 4/5 and 5/5.

14. Leg room service: Most of the customers are satisfied with the legroom offered by the airline. This can be a useful indicator for improving the design of the airplane.

15. Baggage Handling: From the bar plot, it can be inferred that maximum number of customers are satisfied with the baggage handling service offered by the airline.

16. Check-in service: This is an essential indicator. If the check in service offered by the airlines is not smooth and easy the airlines might lose customers. From the bar plot, maximum customers offered a rating of 4/5 and 3/5 which indicates that there are things the airlines can work on to improve the service.

5. Univariate analysis of Inflight entertainment, Inflight service, Cleanliness, satisfaction.

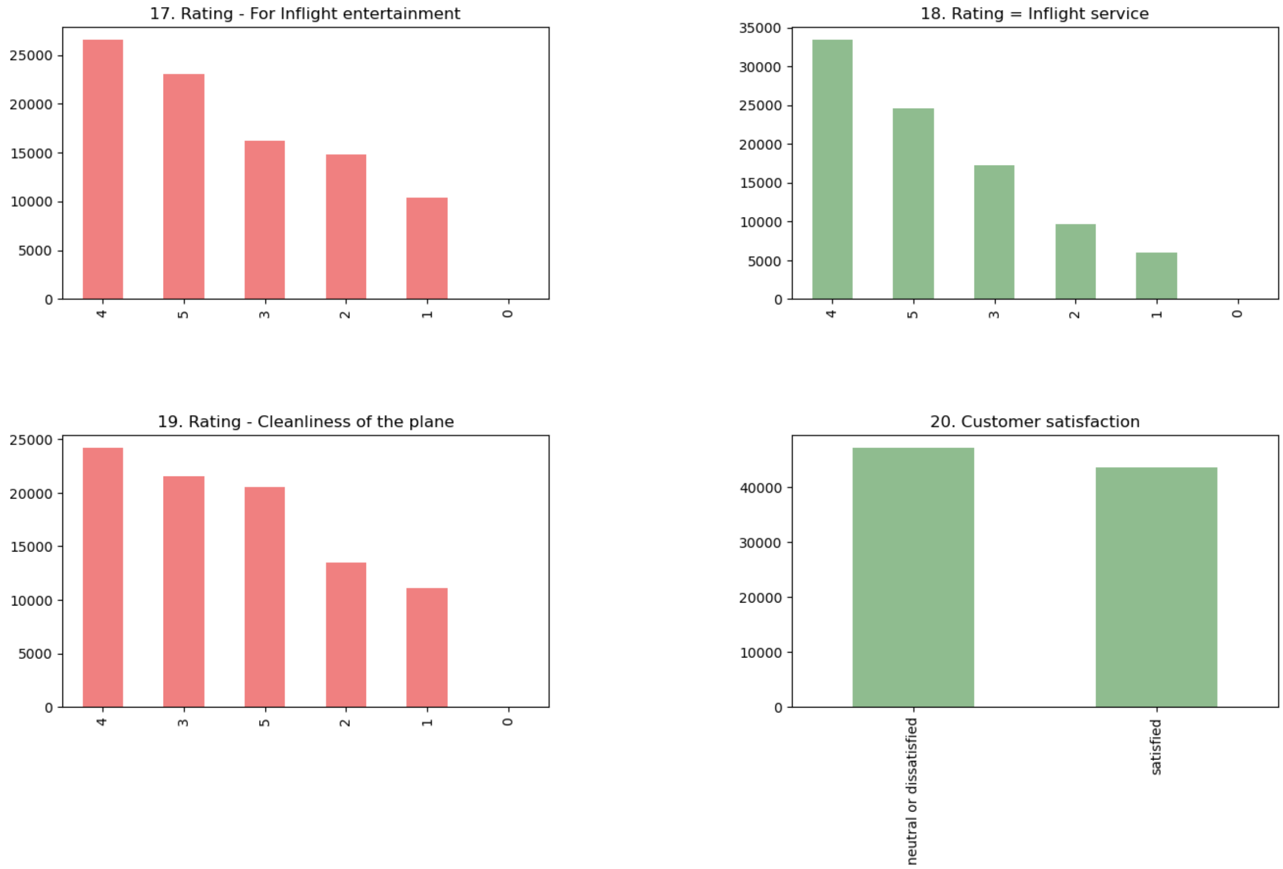


Fig: bar plots of Inflight entertainment, Inflight service, Cleanliness, satisfaction.

Analysis:

17. Inflight entertainment: In - flight entertainment includes various aspects like availability of movies, music albums etc. This is necessary to keep the traveller engaged throughout the journey.

18. Inflight services: From the bar graph, it is observed that the majority of the customers gave a decent rating in terms of the inflight services offered by the airlines. Though there are a considerable number of customers who gave an average rating of 3/5, the number of customers who gave 4/5 and 5/5 are significantly higher.

19. Cleanliness of the plane: Even in this feature, most of the customers rated the service as "above average (4/5). But there are also a considerable number of customers who gave a bad rating (1/5). The airlines must improve in this aspect to reduce the number of customers unhappy with the cleanliness of the plane.

20. Customer Satisfaction: This is the target feature in our dataset, we further plot bivariate plots of other features against this feature to understand the relation between these features and the other independent features.

6. Univariate analysis of Arrival delay in minutes:

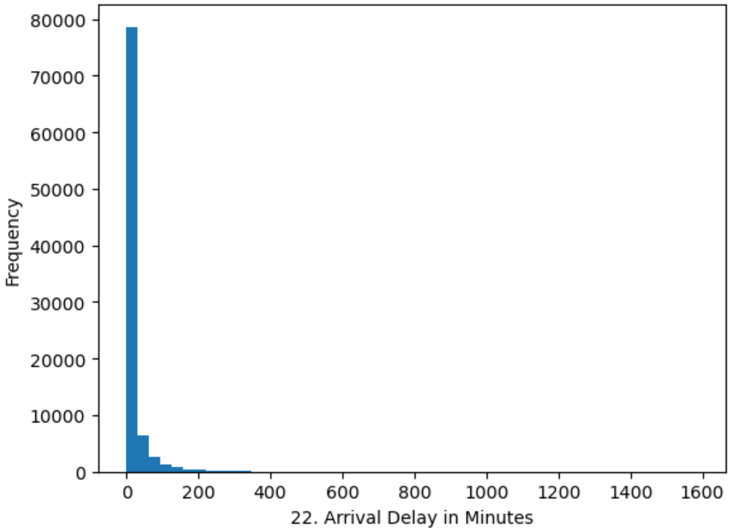


Fig: Histogram of arrival delay (in mins)

Analysis:

21. Arrival delay in minutes: A similar inference can be made from this plot as well, as it is observed that most of the flights are arriving on time as this is essential for customer retainment.

Multivariate analysis of data:

1. Multivariate plot between Age and passenger satisfaction (target variable):

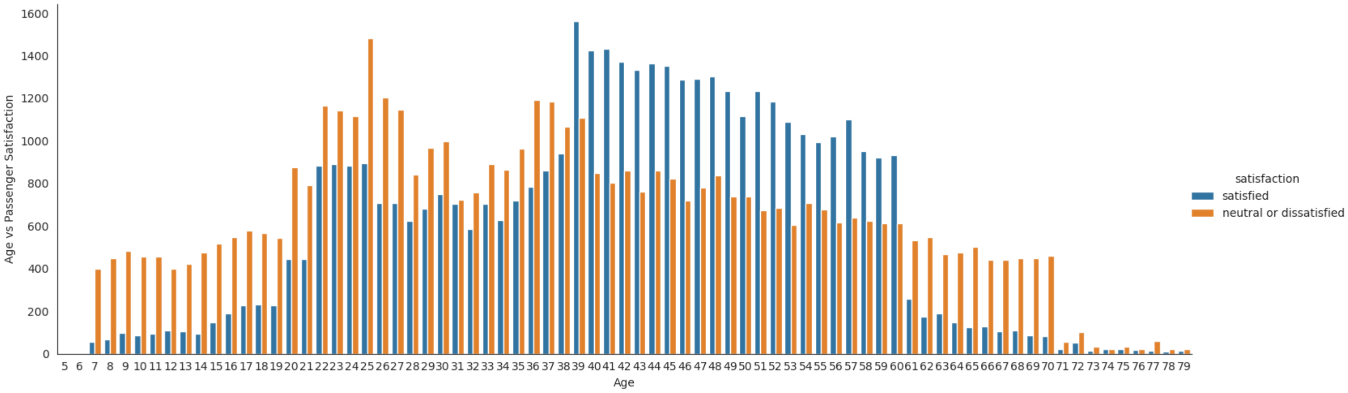


Fig: Catplot of Age vs Passenger satisfaction

Analysis:

As part of the EDA, we have plotted a multivariate plot to understand the relationship between "Age" and "Passenger Satisfaction". From the above count chart, we can conclude that the highest number of satisfied customers belong to the age group of 39 - 60. The highest number of "neutral or dissatisfied" customers are between the age group of 25-27. It can also be inferred that the highest number of customers using the airlines are also spanned between the ages 38 - 60.

2. Multivariate plot of Flight distance, type of travel, satisfaction:

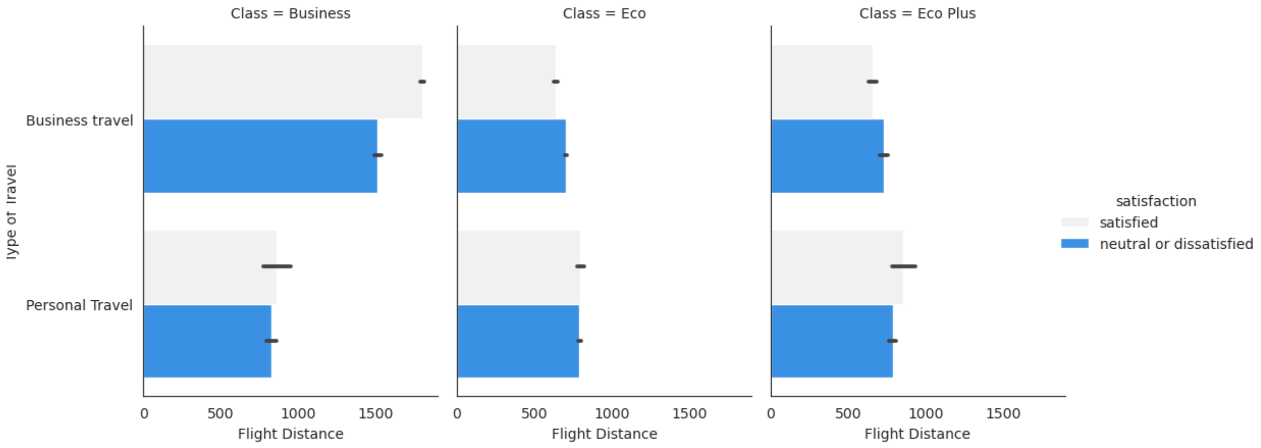


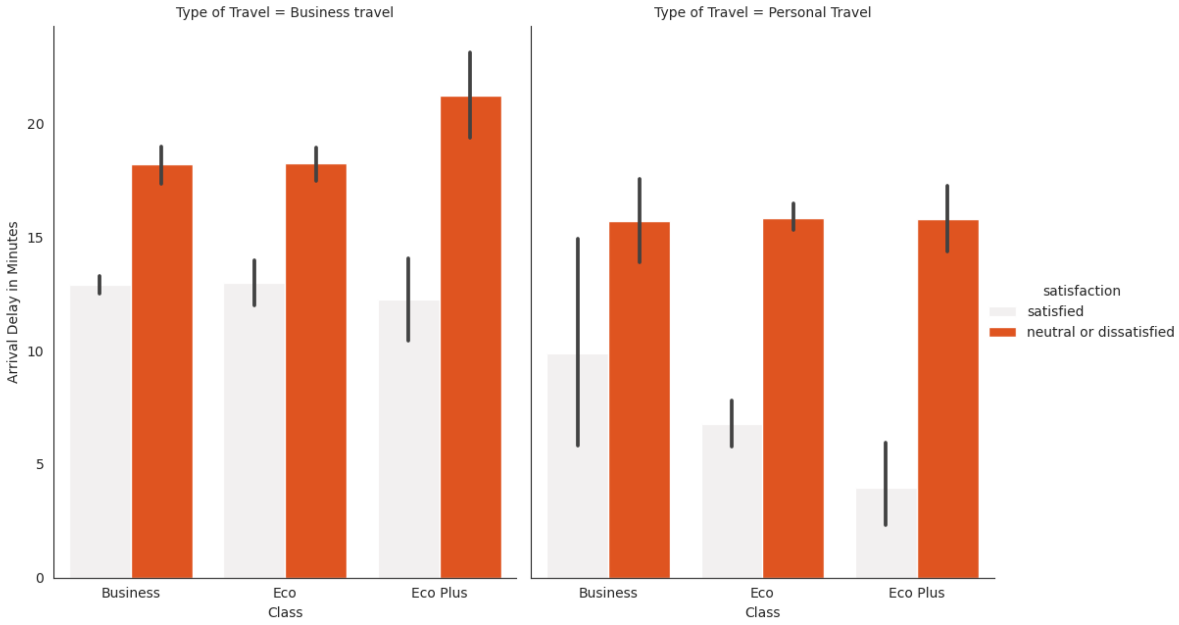
Fig: Catplot of Flight distance, type of travel, target variable.

Analysis:

1. It can be inferred that people who travel long distances usually prefer travelling in "business class" as compared to "Eco" and "Eco Plus".

2. Customers travelling for "Personal" reasons are almost equal in all the three classes i.e Business, Eco and Eco Plus and Customers travelling for "Business" reasons would preferably travel in business class as compared to the other classes.

3. Multivariate Catplot of Arrival Delay in Minutes, Class, satisfaction:

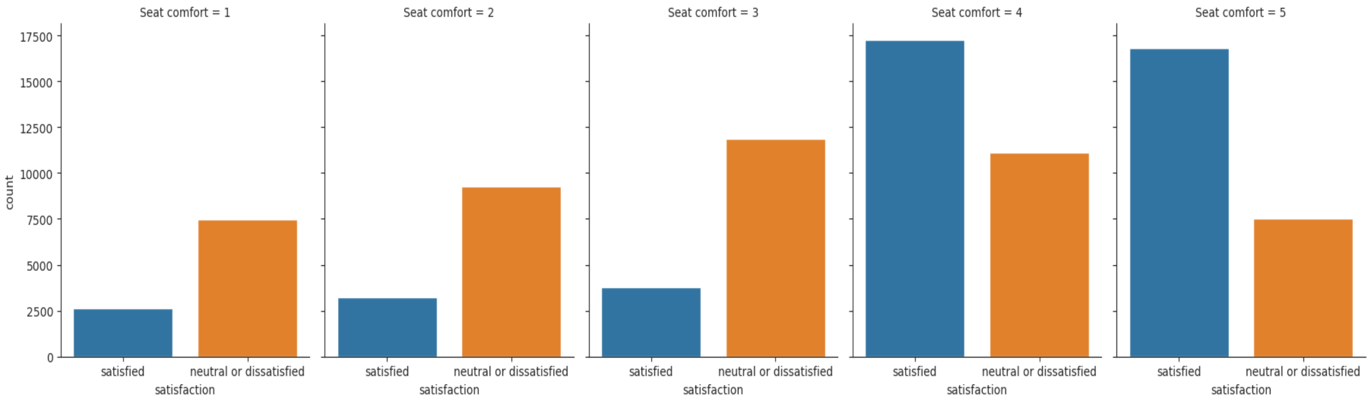


Analysis:

1. In both "Business Travel" and "Personal Travel", the arrival delay is affecting the satisfaction of the customers belonging to the classes “**Eco** " and “**Eco Plus**" as compared to that of "Business class" in both travel types.

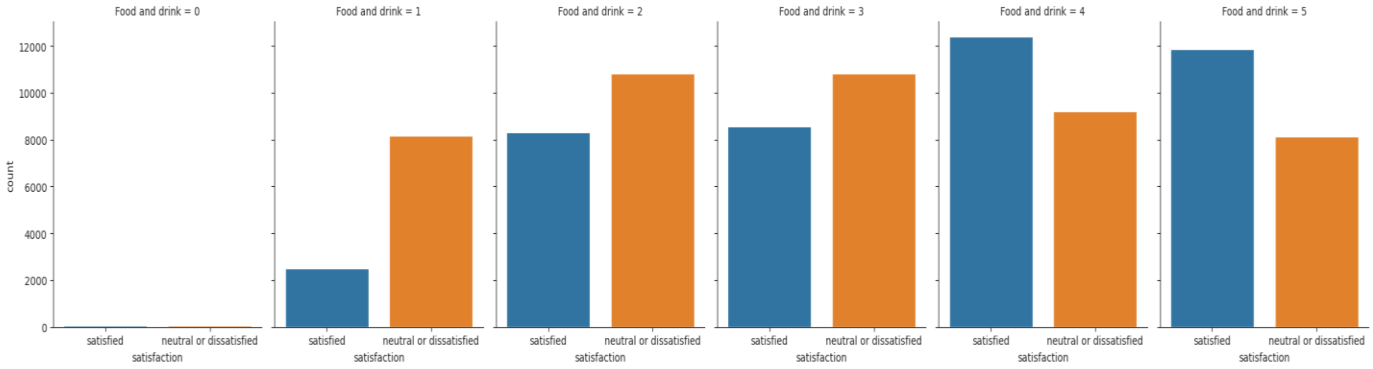
2. Customers travelling for "Business" purposes and travelling in "Eco Plus" class are the most "neutral or dissatisfied" customers out of all the above categories.

4. Bivariate Catplot of seat comfort rating and satisfaction:



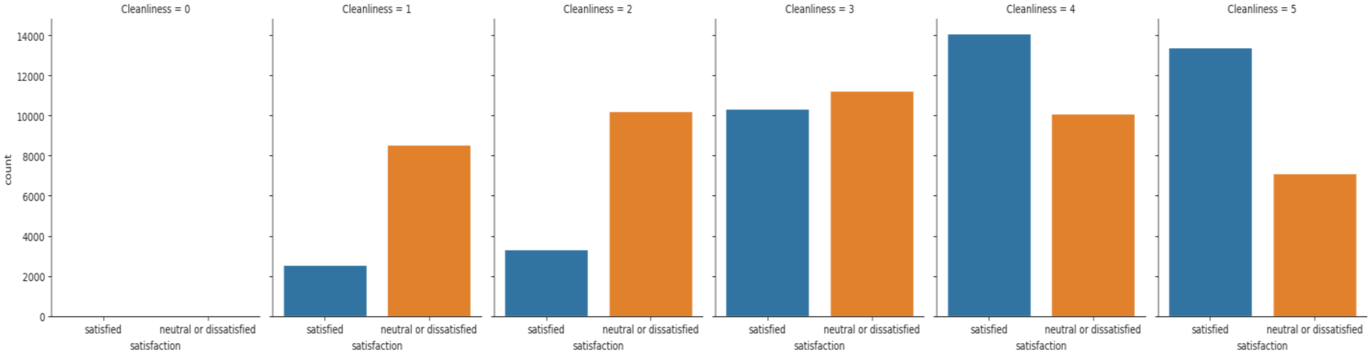
Analysis:
1. As we see, the customers who gave decent (4/5) and great (5/5) ratings for the "Comfort of the seat" are more likely to be "satisfied" with the overall flying experience.
2. Customers who gave an average or below average rating for the "Comfort of the seat" are more likely to be "neutral or dissatisfied" with the overall flying experience.

5. Bivariate Catplot of Food and drink rating and satisfaction:



Analysis:
From the above catplot, we can observe the following things:
1. There are absolutely zero customers to rate the "Food and drinks" service of the airlines a 0/5.
2. Customers who gave decent (4/5) and great (5/5) ratings for the "Food and Drink" service are more likely to be "satisfied" with the overall flying experience as they are considerably large in number as compared with the "neutral or dissatisfied" customers who gave the same rating.
3. Customers who gave the lowest rating (1/5) are more likely to be "neutral or dissatisfied" with the overall experience.

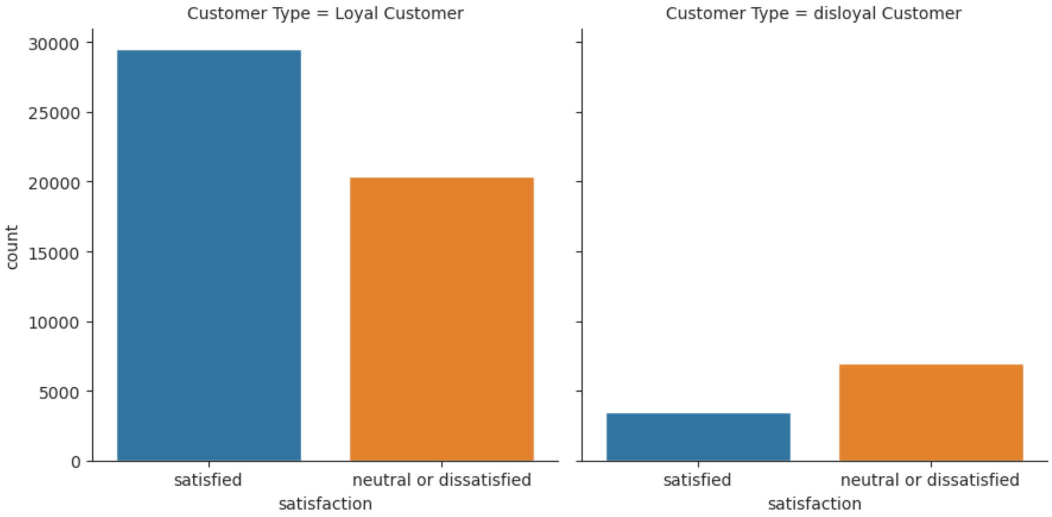
6. Bivariate Catplot of Cleanliness rating and satisfaction:



Analysis:
From the above count plot, we can make the following inferences:

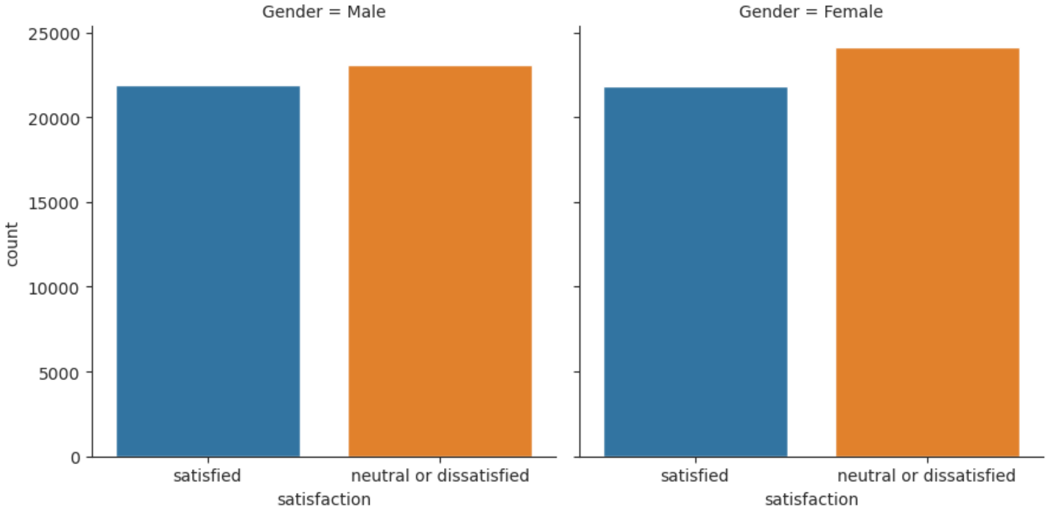
1. Most of the customers gave an average or above average rating (3/5, 4/5, 5/5) for the cleanliness of the airplane.
2. Though most of them have given a rating of above average, there's no significant difference in the number of "neutral or dissatisfied" and "satisfied" customers. This means that "Cleanliness" might not be an important measure for the customer to be satisfied or dissatisfied.

7. Bivariate plot of satisfaction and customer type (Loyal/ disloyal):



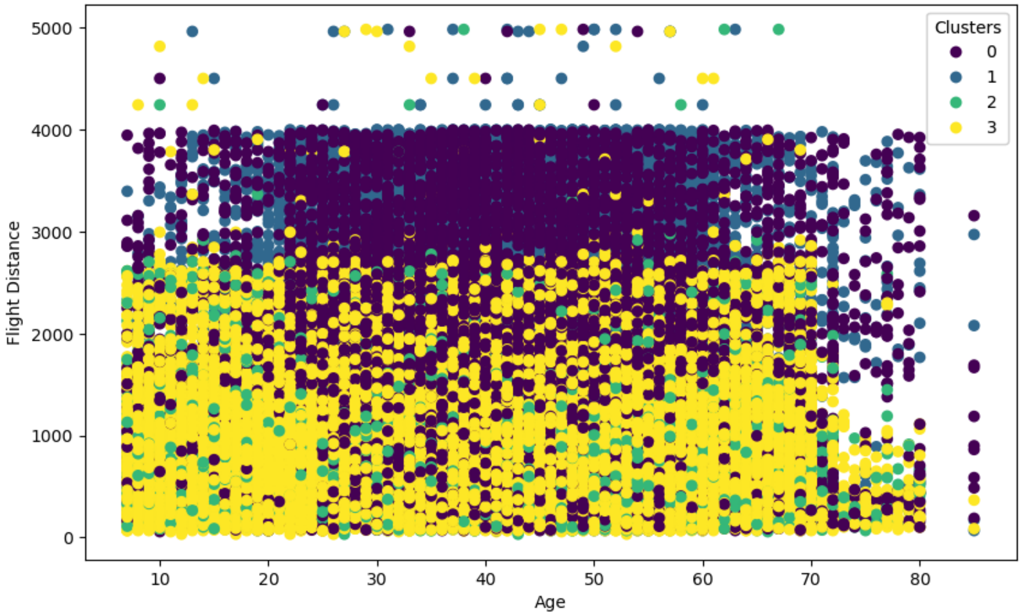
Analysis:
From the above plots, the airlines have a lot of loyal customers which is a positive indicator that the customers are liking the services provided by the airlines. But it is also noticeable that a lot of loyal customers are also "neutral or dissatisfied" with the overall experience. This raises the question that if the customers are not satisfied with the experience, why are they returning to the same airlines?

8. Bivariate plot of satisfaction and gender:



Analysis:
This is a plot to understand the number of "satisfied" or "neutral or dissatisfied" customers based on the gender. We cannot conclude anything based on the gender as in both the genders, the number of "satisfied" and "neutral or dissatisfied" customers are not far apart from each other.

9. Trying to form clustering and draw conclusions:



Analysis:
We tried to perform k-means clustering to understand the data distribution. We tried to form clusters of the following features of our dataset to see if they give any conclusions. We tried to analyse the plots by varying the number of clusters. From the above plot, we can conclude that the range of ages of passengers whose flying time is between 2500 - 4000 is spammed across the entire data range.

10. Feature Engineering:

To identify the correlation between all the independent variables, we have plotted a heatmap which pictorially shows the correlation between each of the features. Based on the correlation values, we choose the features that contribute in predicting our target variable.

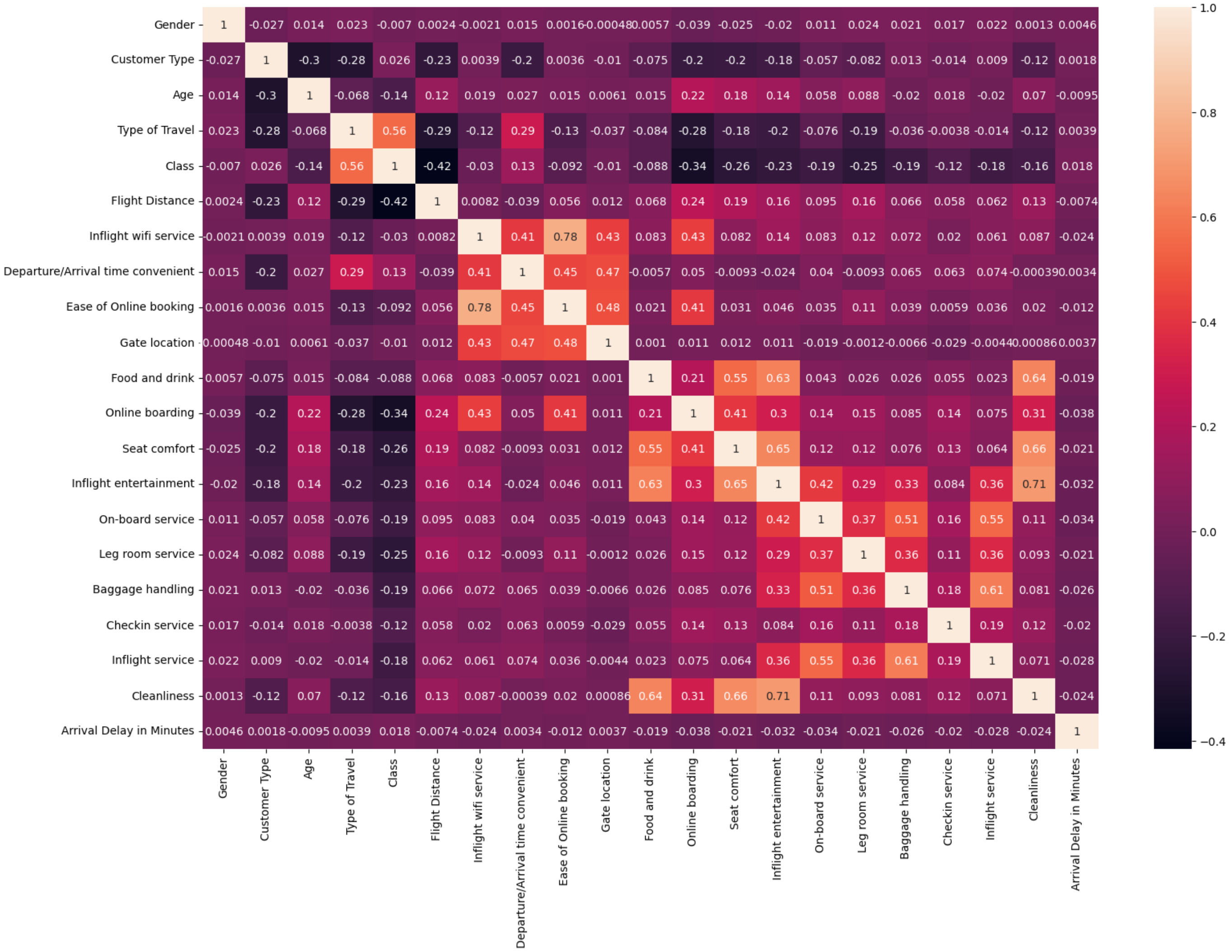


Fig: Heat map

References:

1. **Kaggle dataset:** Airline Passenger Satisfaction | Kaggle

2. What is Univariate, Bivariate & Multivariate Analysis in Data Visualisation? - GeeksforGeeks

3. seaborn heatmap - Python Tutorial (pythonbasics.org)

4. Python Graph Gallery (python-graph-gallery.com)

5. John Tukey (1915 - 2000) - Biography - MacTutor History of Mathematics (st-andrews.ac.uk)