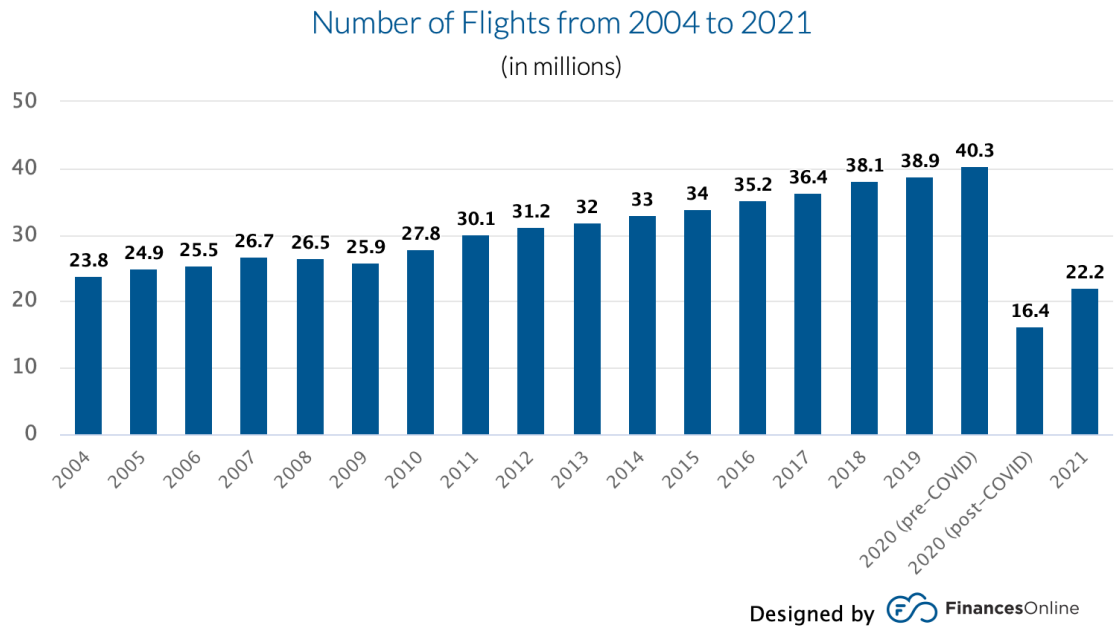
GR STAT5291
Advanced Data Analysis
Final Project

Group Members:

Gexin Chen, Ares Wang, Peiyan Li, Zixuan Liu, Zhihao Ma, Sharon Meng

INTRODUCTION

Airplane, one of the most important transportation facilities in the world, play a remarkable role in shortening the distance between human-beings. People can always arrive at another side of earth in a day by taking an airplane. Furthermore, data also tells us the importance of airplanes as a transportation tool: according to IATA (International Air Transport Association), there are a total of 39 million flights worldwide, which is an average of 10700 airline flights everyday. Although the pandemic of Covid-19 significantly decreases the number of flights everyday, it can be expected with one hundred percent confidence that there will be a huge surge in flight demand in short.



In the process of scientific and technological development and progress, airline services have begun to develop towards standardization. There is almost no longer a gap in the quality of service between airlines. The phenomenon of homogenization has become a trend in the current industry. In order for them to play a prominent role and highlight their advantages, airlines need to improve corresponding management and services. Judging from the current situation that consumers' requirements for services are actually rising, airline companies need to improve and reform their own management structure, do a greater job in controlling operating costs, and

improve customer satisfaction in order to achieve increased market share. To ensure that an airline company can obtain a better development and living environment in the fierce market competition, it is important to investigate the satisfaction level of airline passengers.

From the perspective of consumers, the establishment of the consumer satisfaction model can better grasp the actual needs of the current consumers and the management defects of the airline companies. In this way, airlines can adjust and optimize their own services in combination with the model, and improve customer satisfaction and customer loyalty. That is to say, doing customer satisfaction surveys is the basic task of airlines. The value of feedback from customers is tremendous. Therefore, we are going to utilize a dataset containing airline passengers satisfaction surveys and some basic information about those customers to explore the story behind the satisfaction level of passengers. There are a total of 23 detailed elements that influence passenger flight experience. As data scientists, we intend to use statistical methods including hypothesis testing, data visualization, machine learning etc. to find out which aspects of flight service passengers care the most about. Finally, we can offer recommendations to airline companies based on our findings.

Data Description

The data are contained in two separate files ‘test.csv’ and ‘train.csv’ for the purpose of machine learning implementation.

There are a total of 24 features: 4 of them are numerical variables and the rest 20 variables are all categorical variables. Out of the 20 categorical variables, 14 of them are ordinal variables indicating the satisfaction level of passengers on 14 different aspects of flight service: 1 indicates extremely unsatisfied and 5 indicates very satisfied.

Below are detailed descriptions:

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Flight distance: The flight distance of this journey

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of inflight service

Cleanliness: Satisfaction level of Cleanliness

Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Exploratory Data Analysis

1. Check for Abnormalities & Missing Values

For Exploratory Data Analysis, we combine two files into a single dataframe to increase the sample size. In this way, higher credibility can be obtained.

The first step of Exploratory Data Analysis is to take a brief look at the whole dataset and check if there are any abnormalities.

	count	mean	std	min	25%	50%	75%	max
id	129880.0	64940.500000	37493.270818	1.0	32470.75	64940.5	97410.25	129880.0
Age	129880.0	39.427957	15.119360	7.0	27.00	40.0	51.00	85.0
Flight Distance	129880.0	1190.316392	997.452477	31.0	414.00	844.0	1744.00	4983.0
Inflight wifi service	129880.0	2.728696	1.329340	0.0	2.00	3.0	4.00	5.0
Departure/Arrival time convenient	129880.0	3.057599	1.526741	0.0	2.00	3.0	4.00	5.0
Ease of Online booking	129880.0	2.756876	1.401740	0.0	2.00	3.0	4.00	5.0
Gate location	129880.0	2.976925	1.278520	0.0	2.00	3.0	4.00	5.0
Food and drink	129880.0	3.204774	1.329933	0.0	2.00	3.0	4.00	5.0
Online boarding	129880.0	3.252633	1.350719	0.0	2.00	3.0	4.00	5.0
Seat comfort	129880.0	3.441361	1.319289	0.0	2.00	4.0	5.00	5.0
Inflight entertainment	129880.0	3.358077	1.334049	0.0	2.00	4.0	4.00	5.0
On-board service	129880.0	3.383023	1.287099	0.0	2.00	4.0	4.00	5.0
Leg room service	129880.0	3.350878	1.316252	0.0	2.00	4.0	4.00	5.0
Baggage handling	129880.0	3.632114	1.180025	1.0	3.00	4.0	5.00	5.0
Checkin service	129880.0	3.306267	1.266185	0.0	3.00	3.0	4.00	5.0
Inflight service	129880.0	3.642193	1.176669	0.0	3.00	4.0	5.00	5.0
Cleanliness	129880.0	3.286326	1.313682	0.0	2.00	3.0	4.00	5.0
Departure Delay in Minutes	129880.0	14.713713	38.071126	0.0	0.00	0.0	12.00	1592.0
Arrival Delay in Minutes	129487.0	15.091129	38.465650	0.0	0.00	0.0	13.00	1584.0

It can be detected that the maximum of variables *Departure Delay in Minutes* and *Arrival Delay in Minutes* are unusually large: these numbers indicate that a flight was delayed about a whole day!

We decide to take a further look at samples containing these numbers and nothing abnormal is spotted.

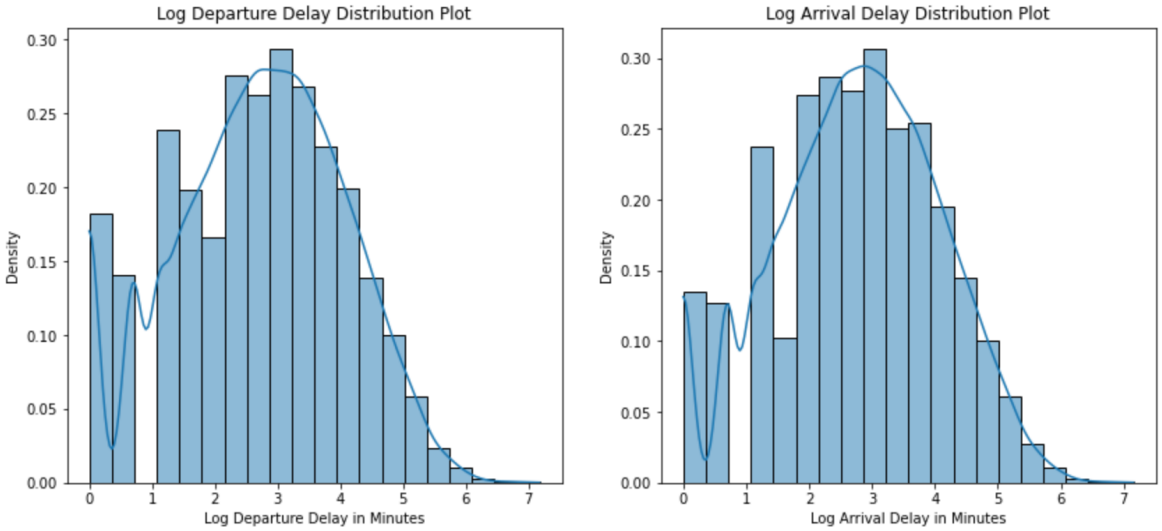
In addition, there are 393 missing values for variable *Arrival Delay in Minutes*. It is probably because some passengers took surveys during the flight and the time of arrival delay could not be

known, which indicates that there is no appropriate estimate to replace these missing values. Since the number of missing values, 393 is small compared to the total number of samples 129880, we decide to simply remove those missing values.

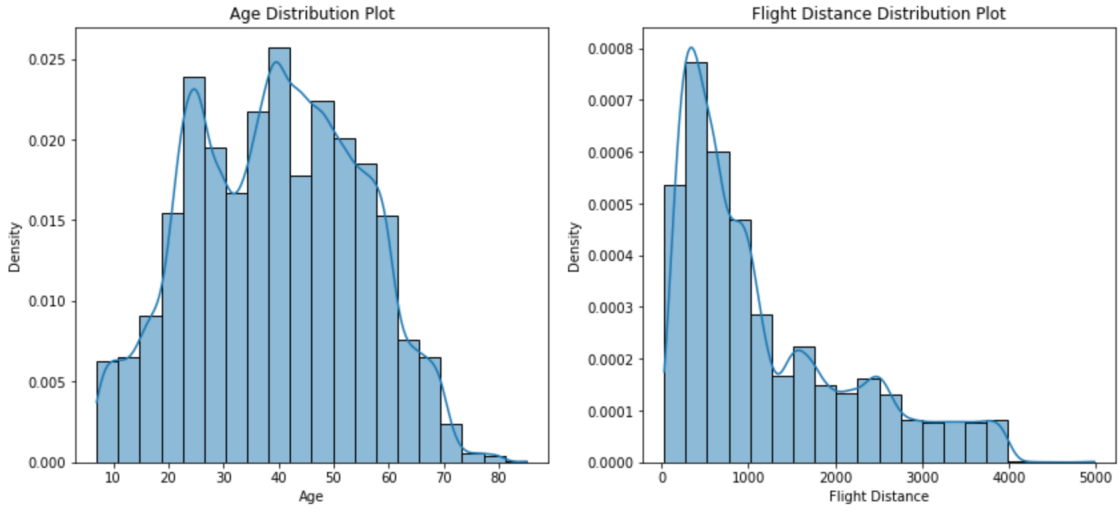
The data types of 24 variables are shown below:
int64(numerical variables): *id, Age, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness, Departure Delay in Minutes*
object(categorical variable): *Gender, Customer Type, Type of Travel, Class, satisfaction*
float64(numerical varaible): *Arrival Delay in Miniutes*

2. Check Underlying Distribution for Numerical Variables

As mentioned in Data Description, there are 4 numerical variables in the dataset: *Age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes.* By checking the underlying distribution of these two variables, we can get a brief sense about whether our samples are generalizable. *Age* should be approximately normally distributed if the sampling method is random. In addition, although we do not have enough information about the overall distribution of *Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes,* we can still get some information from the distribution plots.

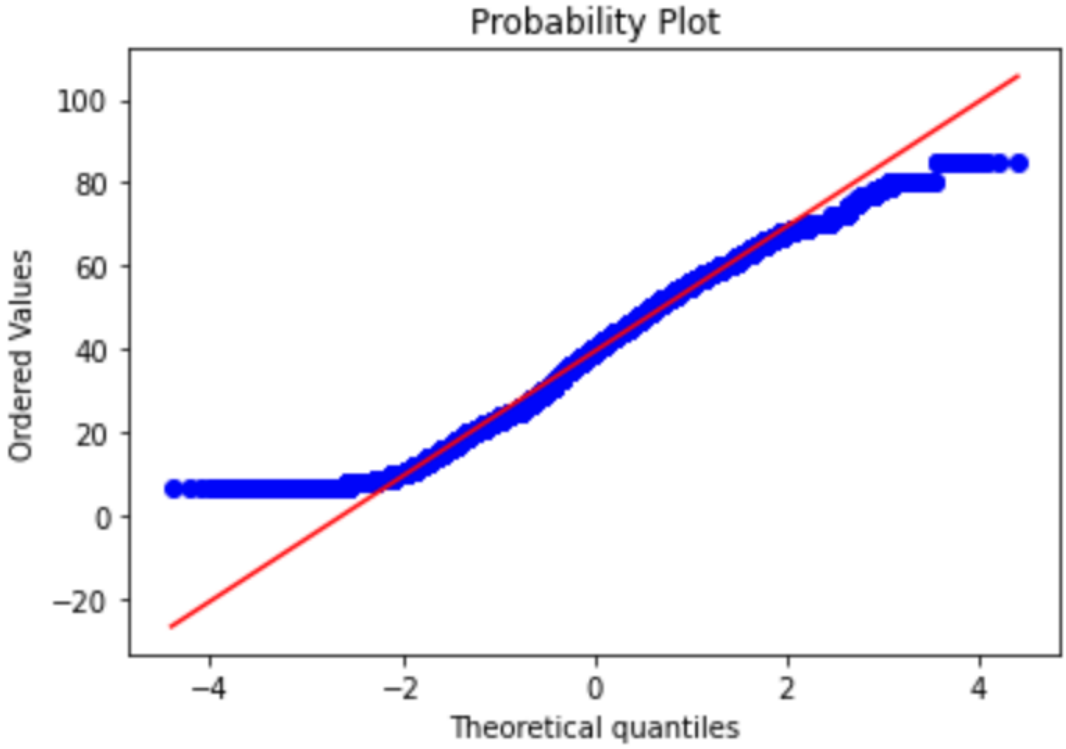


From above, we can see that the distribution of logarithm of *Departure Delay in Minutes & Arrival Delay in Minutes* approximately follow normal distribution. (Log transformation is applied to *Departure Delay in Minutes & Arrival Delay in Minutes* in order to obtain a bell curve.) However, due to the lack of background information, we cannot make a solid conclusion about the sample.



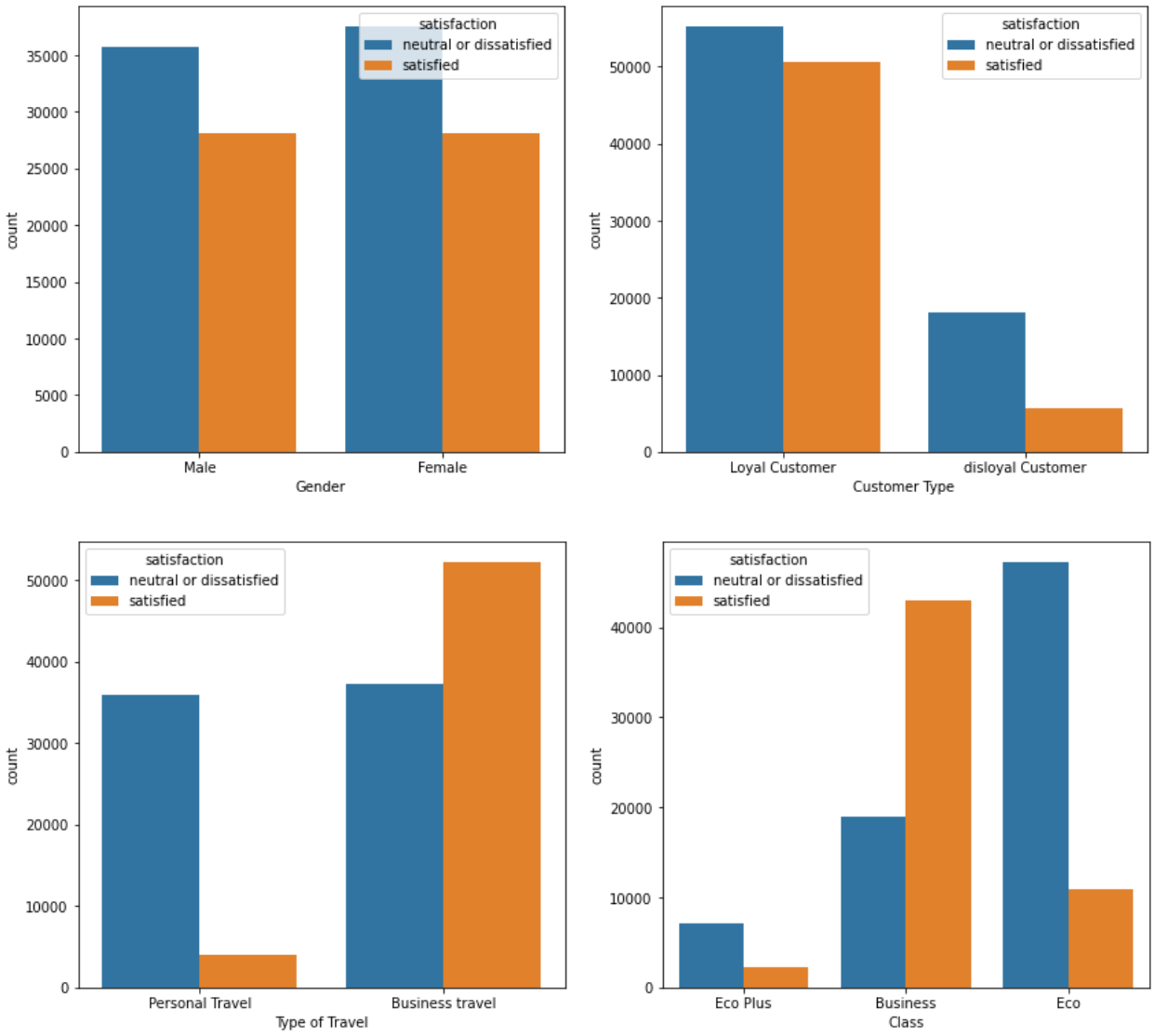
From above, we can see that *Age* approximately follows normal distribution while *Flight*

Distance follows an unknown bell curve. We want to further check the normality of *Age* to ensure the sample is generalizable. Therefore, we utilize a standard visual method, Normal Quantile-Quantile Plot, to further check its normality.

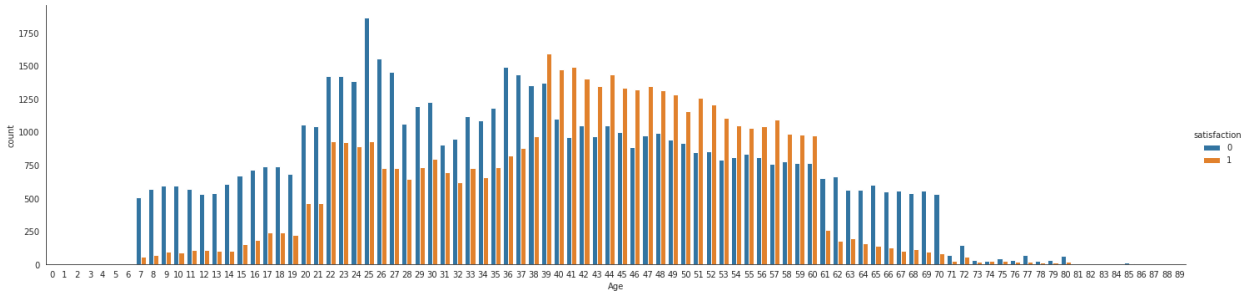


From above, we can see that the plot approximately follows a straight line. We further conduct the Kolmogorow-Smirnov test which is a hypothesis test to check the normality of data and get test statistics equal to 0.9999. Therefore, we can conclude with enough confidence that the distribution of *Age* follows normal distribution and our data is generalizable!

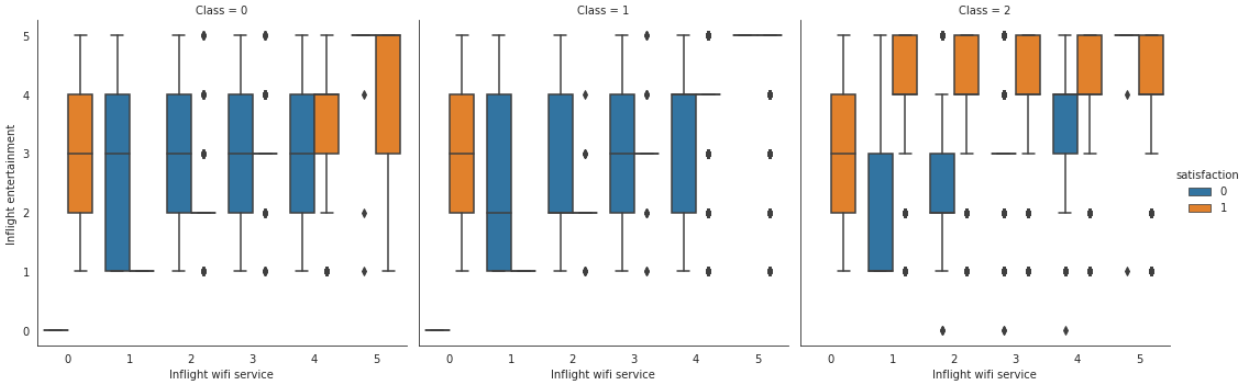
3. Data Visualization for Categorical Variables



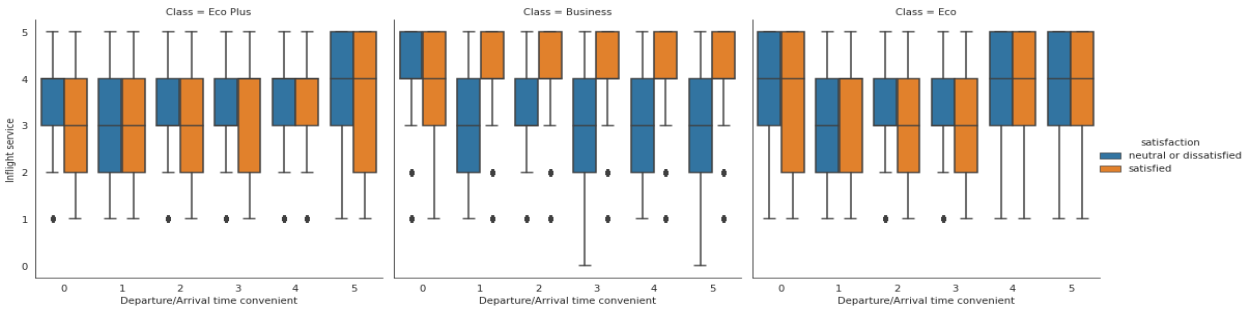
From above, we can see that there is no significant difference between *Gender* types and their satisfaction levels. However, for *Customer Type*, *Type of Travel*, and *Class* types, it can be easily observed that loyal customers and business(business class) travelers are more satisfied with airline services.



From above, we can see that the age of customers has a significant relationship with their satisfaction levels. Among the people aged 7 to 38 and over 61, their dissatisfaction with airlines was significantly higher. Among people aged 39 to 60, they are more satisfied with airlines.



For business class, when the inflight entertainment is 4-5, customers are satisfied with the airline regardless of the number of inflight WiFi services. For economy class and economy class Plus passengers, when the inflight entertainment is 2-4, customers will express dissatisfaction with the airline except that the inflight WiFi service is 4-5.

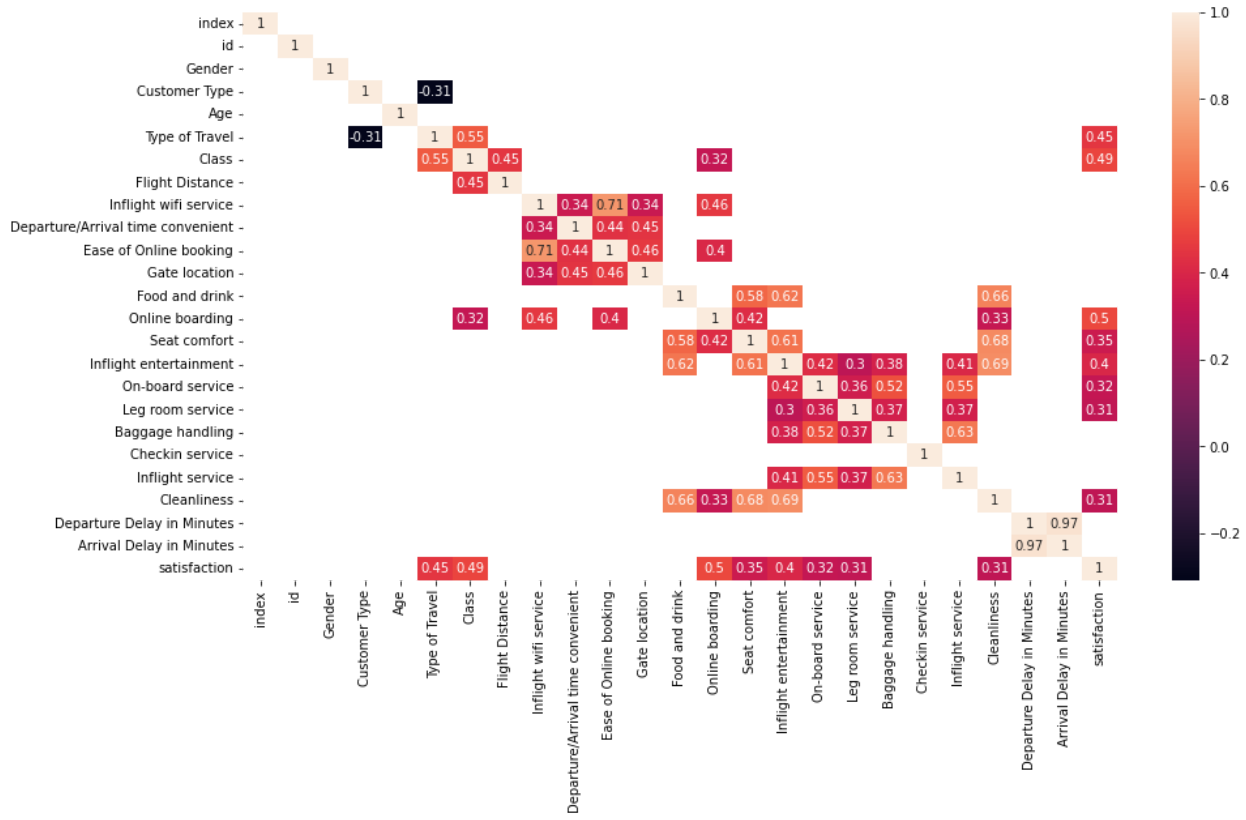


For business class, when the inflight service is 4-5, no matter how many department / arrival time conveners are, customers are satisfied with the airline. For most passengers, when inflight

service is 3-4, customers will express dissatisfaction with the airline.

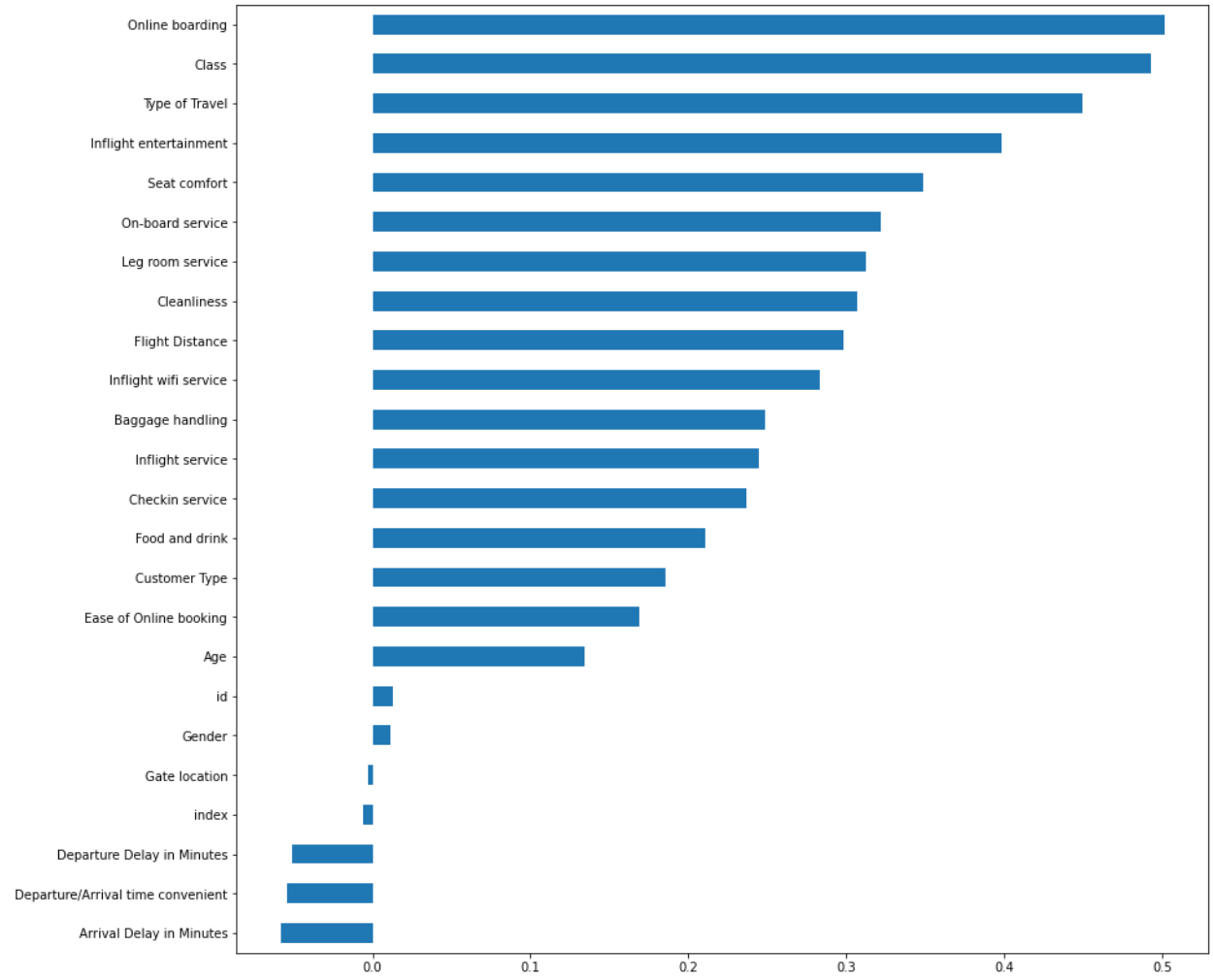
4. Investigate Correlation Between Variables

In order to do Logistic Regression for predicting the target binary variable, it is necessary to check correlation between variables since highly correlated predictor variables may lead our findings not valid. We set the absolute value of correlation coefficient to be higher than 0.3 and get the following graphical representation of correlation matrix output:



On the right side of the heatmap is the sequential palette which shows that lighter colors indicate higher correlation and vice versa.

Then, we move forward to check which variables are correlated with our response variable *satisfaction.* The below horizontal bar plot shows the correlated variables descendingly.



From above, it can be observed that several features are correlated with each other.

We will investigate some correlated features to gain some business insight.

It can be observed that output variable *satisfaction* is highly correlated with *Type of Travel* and *Class*.

Feature Importance Analysis

For feature importance analysis, we divide training data into groups by satisfaction: groups of people who are “satisfied”; “neutral or dissatisfied”.

Then we work on the population of each feature in two groups.

The basic idea is that :

if two populations of one specific feature in two groups have the same distribution, this feature is not important. On the contrary, if these two populations are unlikely to be the same, we could conclude that this feature is important.

We use two methods: Welch’s two-sample t-test (compare the mean) and Wilcoxon signed-rank test (distribution) to test the data.

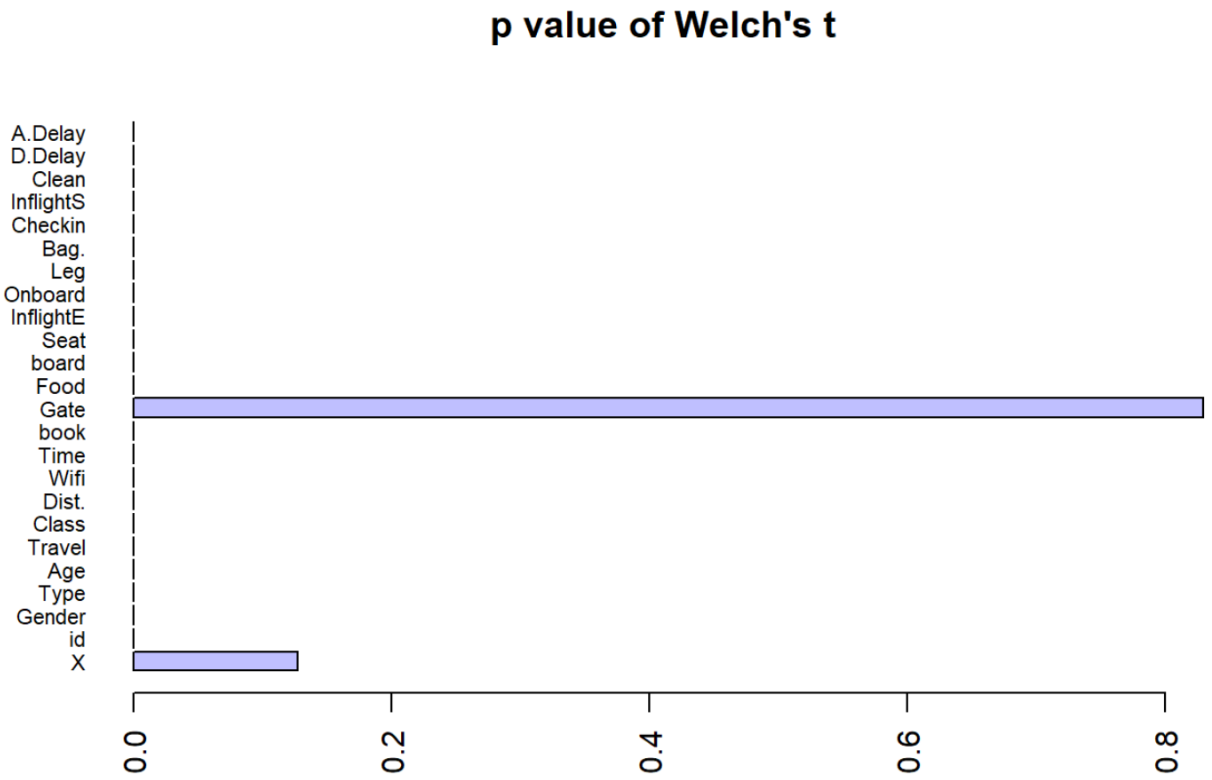
1. Welch’s two sample t-test

Student’s t-test assumes that the sample means being compared for two populations are normally distributed, and that the populations have equal variances. Welch’s t-test is designed for unequal population variances, but the assumption of normality is maintained.

Under the following Null assumption:

The two population means are equal, in which a two-tailed test, we run the test for each feature.

```
##            X           id       Gender         Type          Age       Travel
## 1.273660e-01 9.404291e-06 8.281138e-05 0.000000e+00 0.000000e+00 0.000000e+00
##        Class        Dist.         Wifi         Time         book         Gate
## 0.000000e+00 0.000000e+00 0.000000e+00 1.190888e-61 0.000000e+00 8.290717e-01
##         Food        board         Seat    InflightE      Onboard          Leg
## 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
##         Bag.      Checkin    InflightS        Clean      D.Delay      A.Delay
## 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 1.276708e-61 3.175961e-79
```



We could see that most features are unlikely to have the same mean, which indicates some importance.

However, this test only works under the assumption of normality and could only speak of the mean.

2. Wilcoxon signed-rank test

Considering that the populations are not necessarily normally distributed, we use another test to analyze the importance of features. Wilcoxon test takes numerical data and tests its discrepancy between samples of different classes, namely satisfied and dissatisfied.

The Wilcoxon Rank Sum Test takes numerical data for testing whether samples from two populations have the same distribution. The rank of data $x^{(k)}$ is

$$r^{(k)}(x) = 1 + \sum_{j \neq k} \left(x^{(j)} < x^{(k)} \right)$$

where $x^{(k)}$ denote the k^{th} observation of a feature. Use y to denote the response, and the Wilcoxon rank-sum statistic is:

$$W(x) = \sum_{j=1}^{n} y^{(j)} r^{(j)}(x)$$

The statistics W will be large (small) if the values assumed by X are systematically larger (smaller) in the second population ($Y = 1$). Under the null hypothesis that the distribution of X does not depend on Y, the distribution of W only depends on the number of y equals zero $\left(n_0\right)$ and the number of y equals one $\left(n_1\right)$. For a large sample size n, one can use a Gaussian approximation,

$$W(X) \sim N\left(\frac{n_1(n+1)}{2}, \frac{n_0 n_1(n+1)}{12}\right)$$

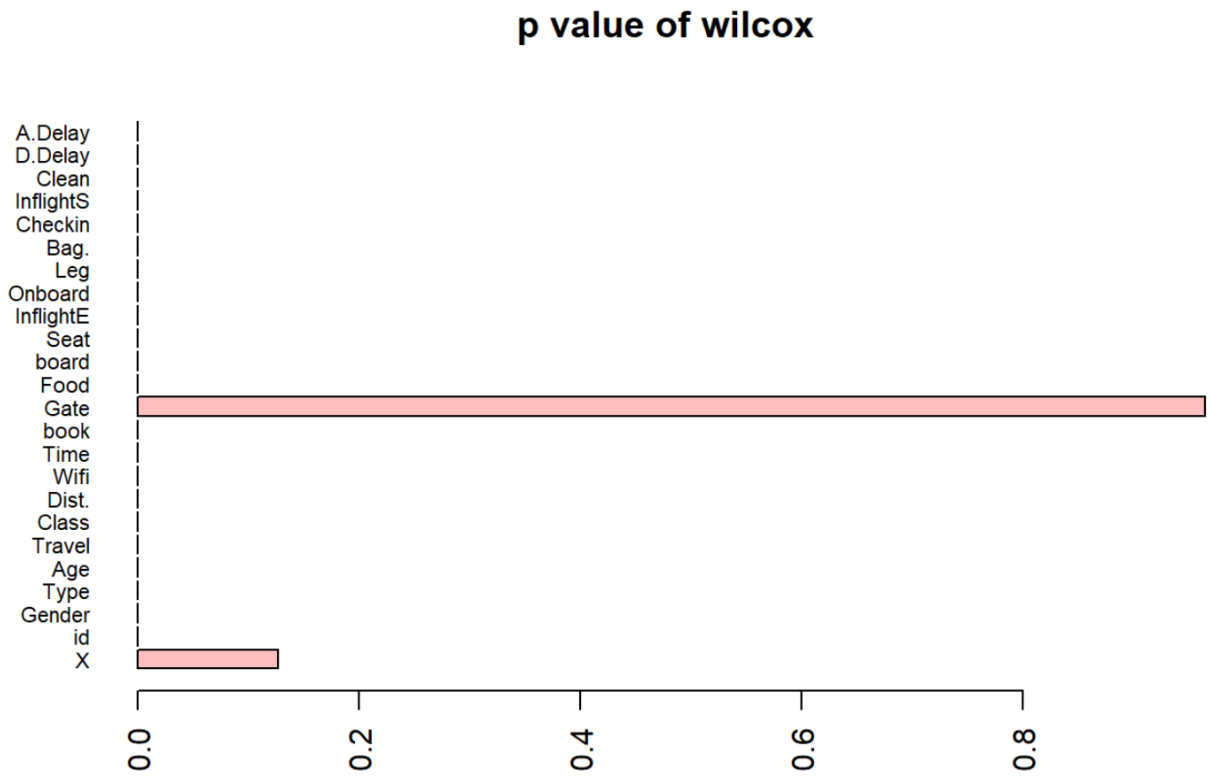
The smaller the p-value is, the more different the distributions of the two datasets are.

An equivalent expression of the Null hypothesis is:

Null hypothesis H_0 The observations $\left(X_i, Y_i\right)$ are exchangeable, meaning that $\left(X_i, Y_i\right)$ and $\left(Y_i, X_i\right)$ have the same distribution. Equivalently, $F(x, y) = F(y, x)$.

Multiple Wilcoxon tests are performed on each feature between satisfied and dissatisfied groups.

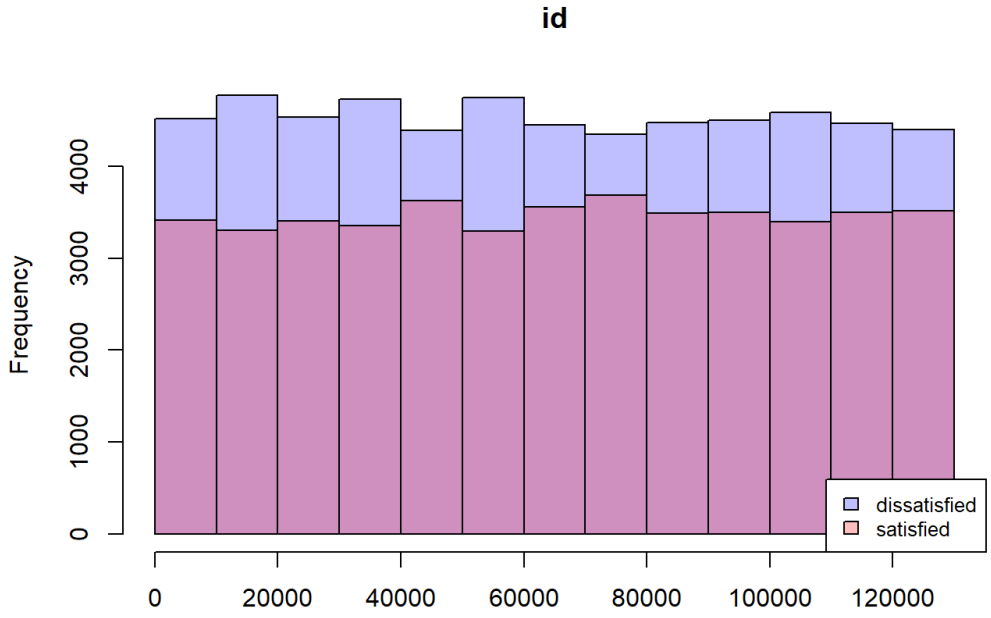
```
##             X            id        Gender          Type           Age
##  1.273000e-01  9.349706e-06  8.278863e-05  0.000000e+00  0.000000e+00
##        Travel         Class         Dist.          Wifi          Time
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  1.754436e-57
##          book          Gate          Food         board          Seat
##  0.000000e+00  9.646642e-01  0.000000e+00  0.000000e+00  0.000000e+00
##     InflightE       Onboard           Leg          Bag.       Checkin
##  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
##     InflightS         Clean       D.Delay       A.Delay
##  0.000000e+00  0.000000e+00 3.153596e-106 1.267778e-229
```



3. Interpret result

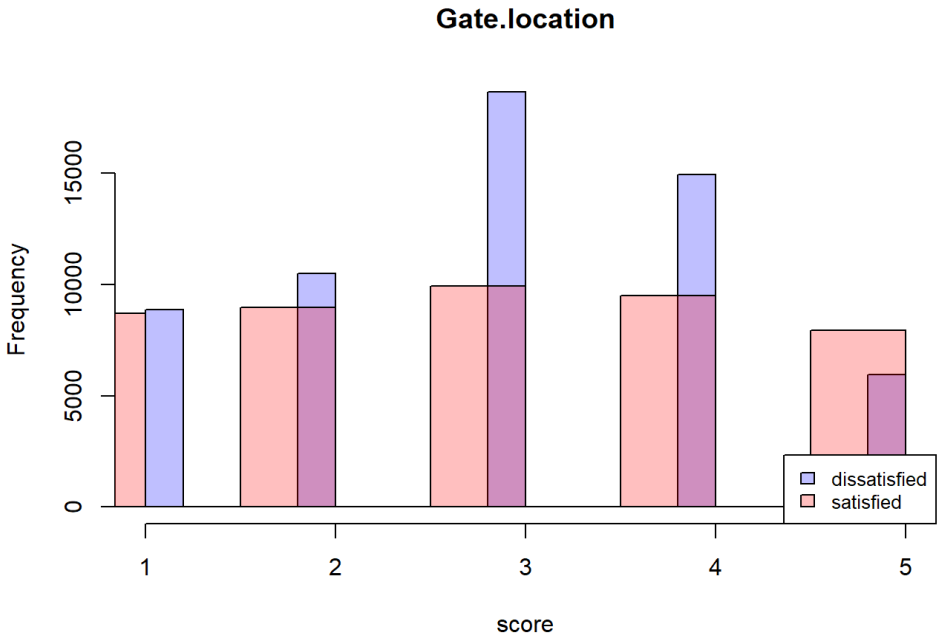
Three observations from the result:

1: Distributions of id in two groups are different, with a very low p-value from the test, suggesting it is an important feature. It is counterintuitive because “id” is usually assigned randomly. However, it could be explained by noticing that “id” may reflect some information of its holder, like “Customer.Type”. We could double examine this observation by picturing it out:

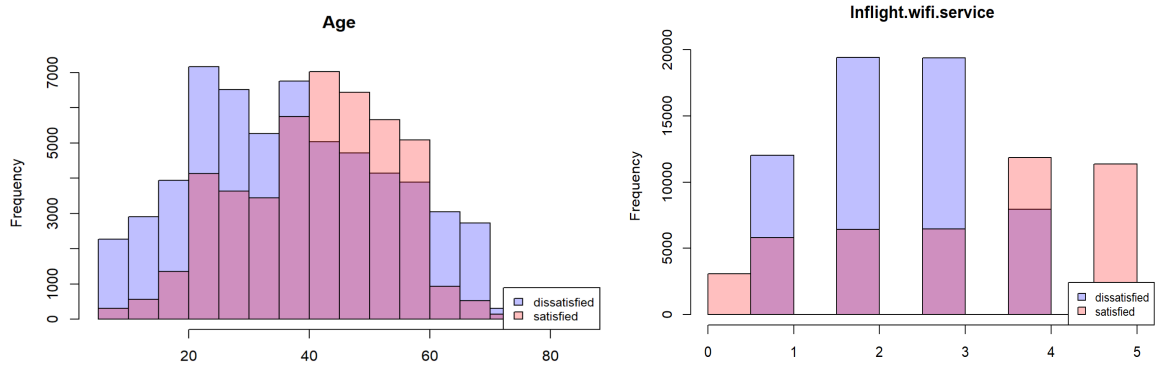


We could see that the distribution of ages in these two groups indeed looks different.

2: Distributions of "Gate.location" in two groups are likely to be the same, with a p-value around 0.9 from the test. So this feature is not important.



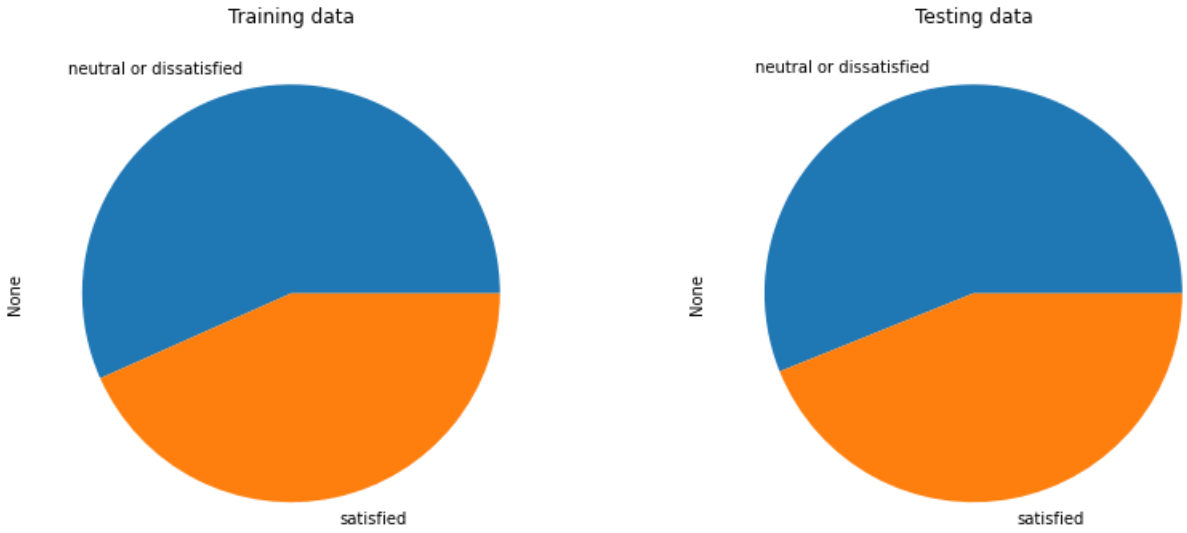
3: Other features are important, unlikely to have the same distribution in two groups. We could take " Age feature" and “Inflight.wifi.service” as examples, double check this result by picturing it out:



We could see that the distribution of ages and wifi service in these two groups indeed looks different, indicating that these features are important.

Machine Learning Implementation

As when we looked into this data, the main variable we want to analyze is the satisfaction variable, which is represented by 0 and 1 indicating yes or no. Thus, based on the dataset, it's intuitive that a classification problem can be raised and analyzed. We first check the distribution of the response variable (satisfaction):



From the data we can see the number of samples with satisfaction or not in the training and testing data are 45025 and 58879, 11403 and 14573 respectively, which implies the ratio of two sets of population is approximately equal to 1 and the data is suitable for machine learning though considering the unfair decision of methods.

1. Traditional Machine Learning Methods

We first applied several traditional machine learning methods on this dataset including Logistic Regression, K Nearest Neighbors, Decision Trees and Support Vector Machine, where the Logistic Regression is the baseline model for our further analysis as it has the simplest structure and this is the dichotomous question.

We recorded the time and the misclassification error of each model and also did the parameter

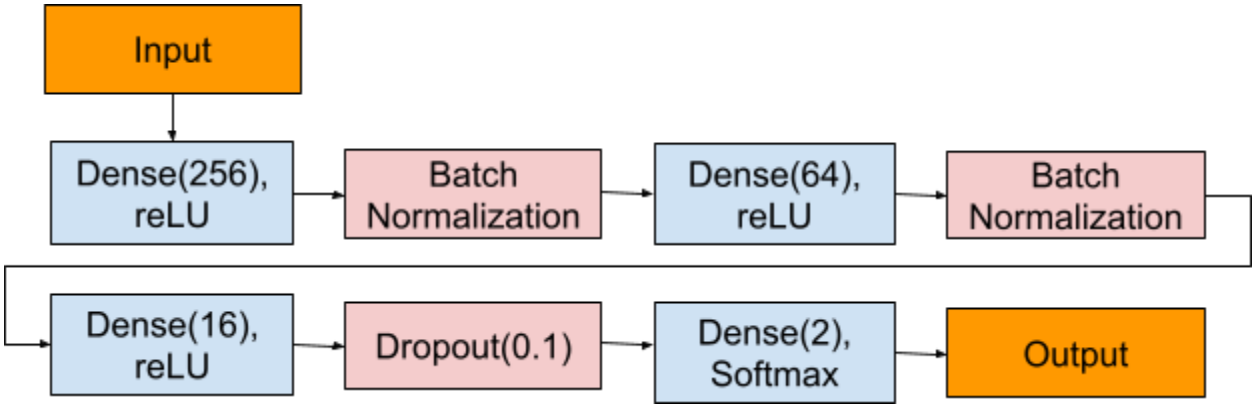
tuning process when applying all 4 machine learning methods in this dataset. From the result of our analysis, we may see our baseline model has a misclassification error of 12.775654% on the test set with only 0.23s training time, and the result of other machine learning methods are shown in the table below.

	Logistic	KNN	Decision Tree	SVM
Error Rate	12.7757%	6.9903%	4.9202%	4.0011%
Training Time	0.2340s	45.8697s	0.4912s	189.2319s
Parameter Set	C=0.01	n_neighbors =9	min_sample_split=10, min_sample_leaf =14, max_depth=20	kernel=gaussian, C=3

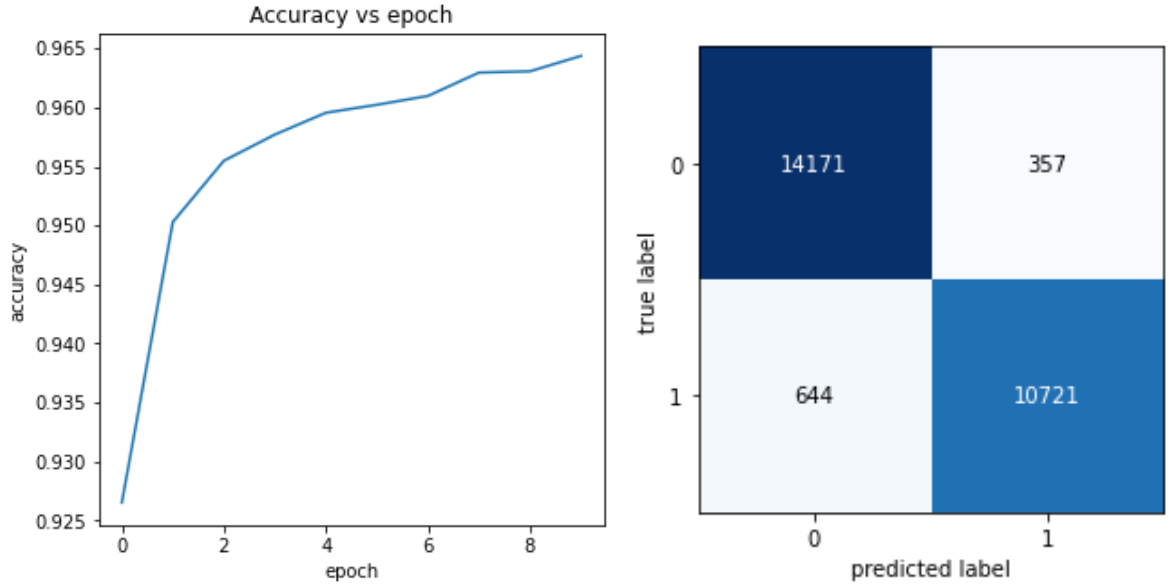
We know the training time and the accuracy have a trade-off relationship. From above results, we may see that KNN and SVM reduced the error rate while the training time goes up, but Decision tree trains better than Logistic while it only increases the time to around 0.5s, which may be due to the dichotomy. But we should also notice that as the Decision Tree has too many attributes to be considered in the hyperparameter tuning process, it actually takes around 30s to get the best parameter set. In all, we may consider more on the Decision tree classifier for our classification problem in the sense of machine learning.

2. Deep Learning Methods

As the question is not a 2D problem, we may use a neural network with dense layers (i.e. the linear layer) for the classification. For each layer of the dense, we use reLU function for activation and softmax as the last, and insert Batch Normalization or Dropout in between dense layers in case of overfitting. The structure of our network is shown below:



We set the epoch times as 10 and split the training data into the final training and validation data by the ratio of 4:1, and finally apply this NN model on the testing data. The accuracy versus the epoch time and the final confusion matrix on the test set are shown below, from which we can see the misclassification error is 3.7385% with the training time of 50.4641s. We may see this is the best performance among all methods, with smallest error and a relatively medium training time.



Recommendation Algorithm

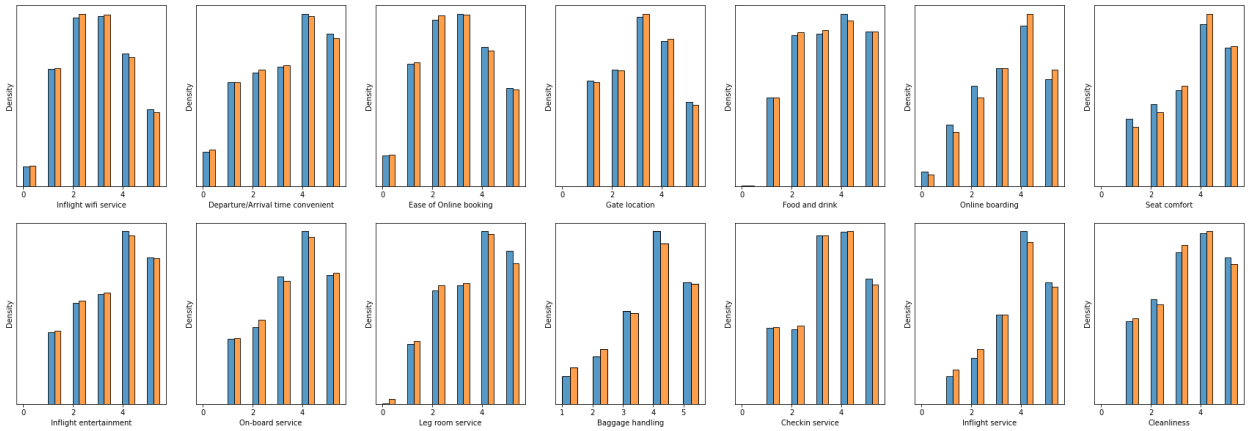
1. Purpose

After predicting the overall satisfaction, we further analyzed the dataset in order to make our project more practical. We observed that the six variables *Gender, Customer Type, Age, Type of Travel, Class and Flight Distance* are the information available to the airline at the time of ticket purchase. By analyzing these six variables, we hope to find out the three areas where different customers are most dissatisfied with the airline's services. The airline can then get this information before the customer gets on the plane, so that it can provide personalized service to the different customer. This will help airlines to improve their overall satisfaction.

2.Analyze Dataset

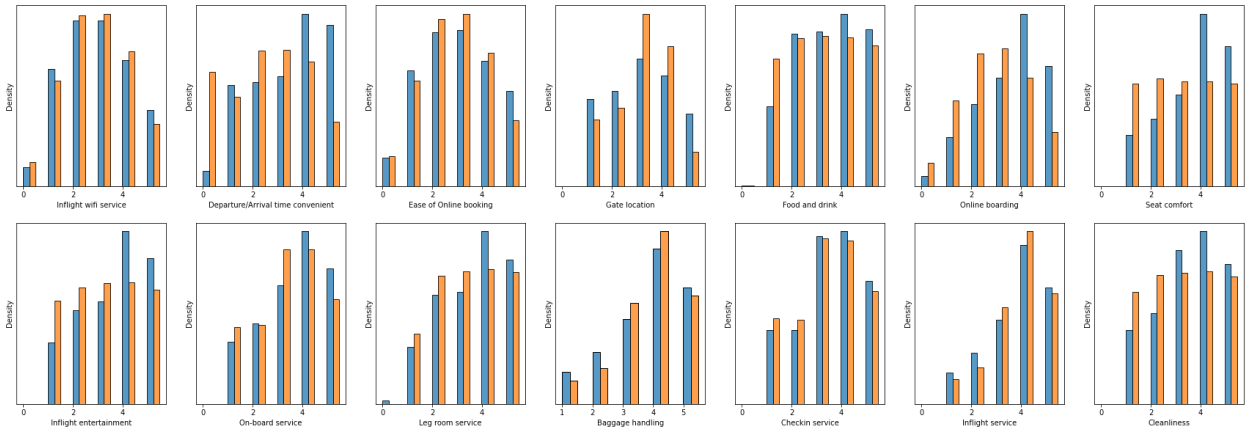
First we need to find the effect of these six variables on the satisfaction of each service during the flying. We plotted the histgrams of these six variables for different satisfactions

· *Gender* (Male: Blue Female: Orange)



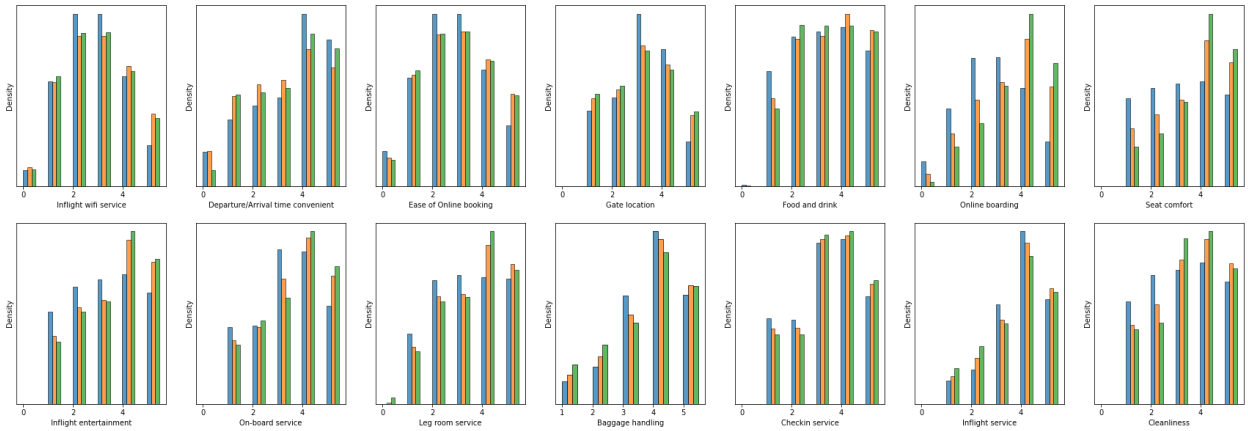
Accoring to the above plots, *Gender* has little effect on these 14 satisfactions.

· *Customer Type* (Loyal Customer: Blue Disloyal Customer: Orange)



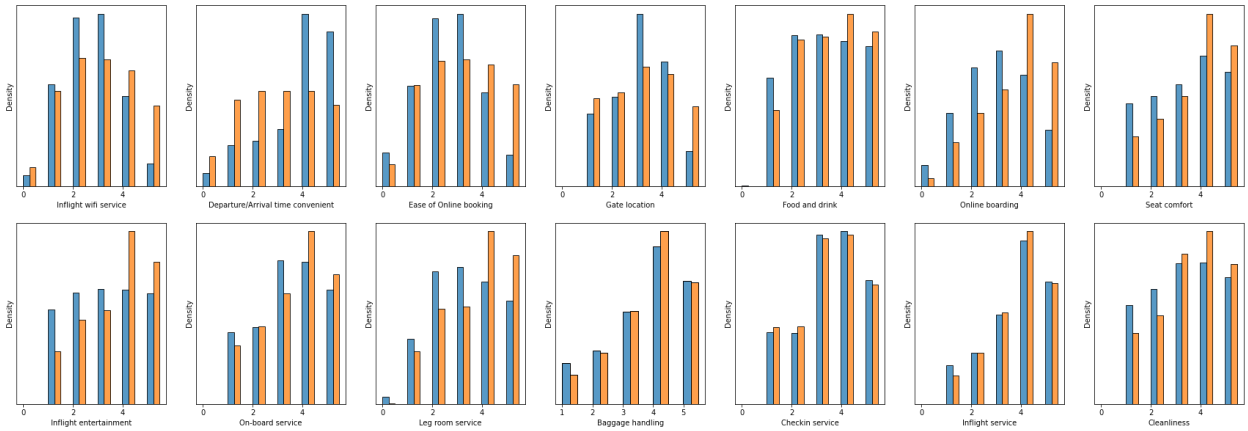
Accoring to the above plots, *Customer Type* has a significant impact on these 14 satisfactions. Especially for *Departure/Arrival time convenient, Gate Location* and *Online boarding,* different Customer type have different satisfactions.

· *Age* (Young (Age<21) : Blue Adult (21<Age<50): Orange Old (Age>50): Green)



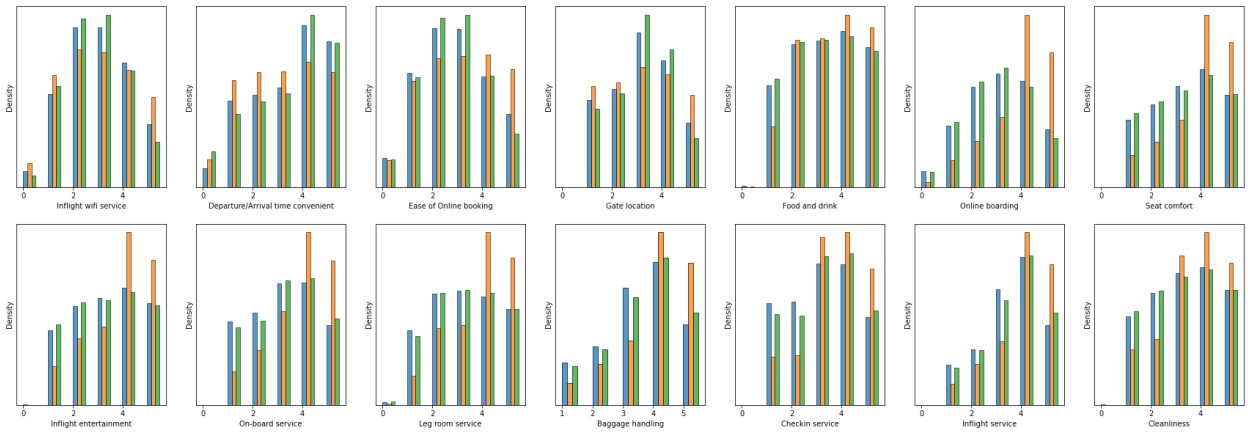
Accoring to the above plots, *Age* only has significant impacts on the *Online boarding* and *Seat comfort*.

· *Type of Travel* (Personal: Blue Business: Orange)



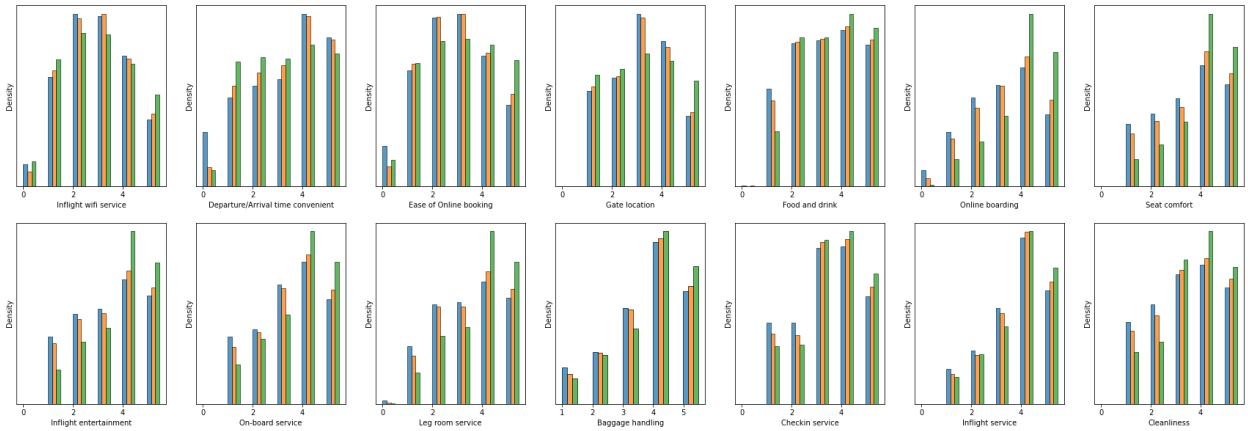
Accoring to the above plots, Type of Travel has a significant impact on these 14 satisfactions. Especially for *Inflight wifi service, Departure/Arrival time convenient, Easy of Oline booking, Seat comfort, Inflight entertainment* and *Leg room service,* different Type of Trave have different satisfactions.

· *Class* (Eco Plus: Blue Business: Orange Eco: Green)



Accoring to the above plots, Class has a significant impact on these 14 satisfactions. Especially for *Online boarding, Seat comfort, Inflight entertainment and Leg room service,* different Class have different satisfactions.

· *Flight Distance* (Short: Blue Medium: Orange Long: Green)



Accoring to the above plots, *Flight Distance* only has significant impacts on the *Online boarding* and *Seat comfort*.

Because the effect of gender on satisfaction is very small, we first exclude this variable. Although Age and flight distance have an effect on Online boarding and seat comfort, they do not have much effect on most other satisfaction. Therefore, we have divided our customers into the following 12 types according to Customer type, Type of Travel and Class.

Loyal Customer|Business travel|Business
Loyal Customer|Business travel|Eco
Loyal Customer|Business travel|Eco Plus
Loyal Customer|Personal Travel|Business
Loyal Customer|Personal Travel|Eco
Loyal Customer|Personal Travel|Eco Plus
Disloyal Customer|Business travel|Business
Disloyal Customer|Business travel|Eco
Disloyal Customer|Business travel|Eco Plus
Disloyal Customer|Personal Travel|Business
Disloyal Customer|Personal Travel|Eco
Disloyal Customer|Personal Travel|Eco Plus

3. Method of implementation

In order to find the most unsatisfactory airplane service for different types of customers, we used the mean comparison method. First we calculated the mean of 14 different service satisfactions of all customers, and then we compared the mean values of the satisfaction of different types of customers with the total. If the average satisfaction of a service for a specific type of customer is lower than the mean satisfaction of all customers, we conclude that this type of customer is more

dissatisfied with this service than other types of customers. The following form is the mean of 14 different services satisfactions of all customers:

Service Type	Mean Satisfaction	Service Type	Mean Satisfaction
Inflight wifi service	2.729753	Inflight entertainment	3.358341
Departure/Arrival time convenient	3.060081	On-board service	3.382609
Ease of Online booking	2.756984	Leg room service	3.351401
Gate location	2.977026	Baggage handling	3.631687
Food and drink	3.202126	Checkin service	3.304323
Online boarding	3.250497	Inflight service	3.640761
Seat comfort	3.439765	Cleanliness	3.286397

4. Conclusion

After applying the above method, we get the following results:

Type of Customers	1st dissatisfaction	2nd dissatisfaction	3rd dissatisfaction
Loyal Customer Business travel Business	Departure/Arrival time convenient	Gate location	Inflight wifi service
Loyal Customer Business travel Eco	Checkin service	Inflight service	Baggage handling
Loyal Customer	Checkin service	On-board service	Inflight service

Business travel Eco Plus			
Loyal Customer Personal Travel Business	Leg room service	Inflight entertainment	Inflight service
Loyal Customer Personal Travel Eco	Online boarding	Inflight entertainment	Ease of Online booking
Loyal Customer Personal Travel Eco Plus	Online boarding	Ease of Online booking	Inflight entertainment
Disloyal Customer Business travel Business	Departure/Arrival time convenient	Seat comfort	Online boarding
Disloyal Customer Business travel Eco	Departure/Arrival time convenient	Online boarding	On-board service
Disloyal Customer Business travel Eco Plus	Online boarding	Departure/Arrival time convenient	On-board service
Disloyal Customer Personal Travel Business	Seat comfort	Cleanliness	Inflight entertainment
Disloyal Customer Personal Travel	Online boarding	Seat comfort	Ease of Online booking

Eco			
Disloyal Customer Personal Travel Eco Plus	Seat comfort	Inflight entertainment	Cleanliness

Reference

1. TJ KLEIN. 2020. *Airline Passenger Satisfaction* Kaggle. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
2. https://harshalvaza.medium.com/invistico-airlines-understanding-customer-satisfaction-6108b500e592