AIRLINE SERVICE SATISFACTION

Abstract:

The airline industry continues to remain as one of the most competitive industries. All airlines provide the same core service: to transport passengers around the world via air transport. To differentiate themselves from the rest, each airline must have their own unique selling point to attract new and existing customers. In our paper, the services provided by the airline, both general services and in-flight services will be focused.

Using a dataset obtained from Kaggle, consisting of 129,880 customer responses, this study aims to analyse and prioritise various observable variables related to in-flight and general services. Then, by performing factor analysis on the variables, airlines can gain insights into which aspects of in-flight services they should prioritize to enhance customer satisfaction. The goal is to provide airlines with valuable information that can help them improve their services, increase customer satisfaction with minimum effort, and ultimately boost profits.

This research seeks to contribute to the airline industry's understanding of the factors that influence customer satisfaction and provide recommendations for airlines to improve their in-flight services. Through this, airlines can establish a competitive edge in this competitive market.

INTRODUCTION

With over 5000 different companies, the airline industry is one of the most competitive markets around. Every single company in this industry provides the same main service, that is picking up a passenger from a source area and sending them to their destination via air transport. However, with many companies that provide the exact same service to compete with, each company has their own unique selling point that encourages a customer to come back and repeatedly use their airlines for their travels. This unique selling point could be low price points, flexible times, loyalty programs, customer service and in-flight services and so on. Each selling point could be a major factor in attracting existing and new customers. In our research, the in-flight services that are provided by these airlines will be focused on.

Researching these services is helpful because it will enable an airline to know how satisfied a customer is. If a customer is satisfied, they are highly likely to use the airline again. Besides that, if the airline focusses on these services, it can be a good selling point to attract new customers. Our airline satisfaction dataset has been obtained from Kaggle. It has 24 observable variables obtained from 129880 customers that will allow us to determine how satisfied the customer is.

The **objective** of this study is to help the airline improve customer's satisfaction with minimum money spent. This is done by **identifying the key features** that significantly affect airline passenger satisfaction. Consequently, all the **features will be ranked based on their importance** so that airlines will know which features to place more focus on. Lastly, the **underlying factors** that contribute to passenger satisfaction **will be determined** by conducting exploratory factor analysis and **will be ranked** by Random Forest. By identifying the factors, some other **important features that do not exist in the dataset may be figured out.**

REVIEW

1. Method involved to preprocess the data

a) Quantile Transform (Brownlee, 2020a)

- Non-Linear transformation technique
- Deals with outliers, and highly skewed data
- Transform the data so that they possess a Gaussian or uniform probability distribution
- Transformation of data is applied on each feature independently

b) Data Scaling (Bhandari, 2020)

- transforming data to have a cleaner scale
- Typically involves standardization or Normalization
- helps so that all features that contribute in a more equal manner

Normalization	Standardization
rescales values to have range [0,1]	centers data around mean and scales to a standard deviation of 1
Sensitive to outliers	Less sensitive to outliers

2. Model to rank the features based on their importance

a) SelectKBest in sklearn

- There are 2 popular feature selection techniques that can be used when **variables/features are numerical** and **target is categorical** which is **(Brownlee, 2020c)** :
    - ANOVA F-value method
    - Mutual Information Statistics
- The main difference between them is that
    - F-test can estimate the degree of linear dependency between two random variables
    - Mutual information methods can capture any kind of nonparametric statistical dependency

b) Permutation Importance using ANN

- Permutation important is a feature selection technique
- It is done by randomly permuting the data in a column/feature and computing the metrics of actual_y and predicted_y
- If the metric remains high after randomly permuting the data, it means that this column of feature is not that important

c) Recursive Feature Elimination (Brownlee, 2020b)

- RFE is a feature selection method that recursively removes small number of features per loop until the specified number of features are met
- It will rank features by importance and refit the model after some features are discarded and repeats, making it recursive
- In sklearn, RFE() will only work for the estimators that have provided the information about feature importance (E.g. feature_importances_, coef_)
- Below is a list of RFE models that used in the project:

Random Forest **(Mbaabu, 2020)**	● Random Forest is a Supervised machine learning model ● It creates a forest of decision trees, while each decision tree is formed based on randomly extracting the observations and features from the dataset. This means that not every tree will see all observations or see all features. ● From a forest of decision trees, we calculate the importance of a feature based on how it removes impurities ● This essentially means that we can observe the change in accuracy of the decision tree when we include or don't include a feature ● Impurity decrease from each feature can be averaged across trees to see the significance of the feature

	● A major advantage of random forest is: ■ robustness ■ able to capture nonlinear relationships between the target and the features
Gradient Boosting Classifier **(Saini, 2021a)**	● Boosting is a type of ensemble learning technique that combines several weak models to become a strong model ● Boosting algorithm will build a sequences of n model (M1, M2, .., Mn) such that Mi+1 will always use to rectify the errors present in Mi for i in [1,n] ● Essentially, the model will learn the mistakes from the previous model to become a better model ● Gradient Boosting Classifier is a **classifier** that uses **boosting** algorithm and it is trained to minimize the loss function of its predecessor using by using the **gradient descent** ● Per iteration, the gradient of the loss function of the current model will be given to a new weak model. The new weak model will then minimize this gradient ● Typically, we use decision trees as our weak model. Each new decision tree focuses on data points that were not well predicted by the previous models.
Logistic Regression	It is just a linear regression followed by a function to make the output become a probability distribution (like softmax). Then, the class with the maximum probability will be the output.
Linear Support Vector	Linear Support Vector Classification is supervised machine

Classification **(Saini, 2021b)**	learning algorithm which perform classification by creating an lines/boundaries/hyperplane that best separates the classes with maximum margin between hyperplane and the data points

3. Model to extract the factors from the features

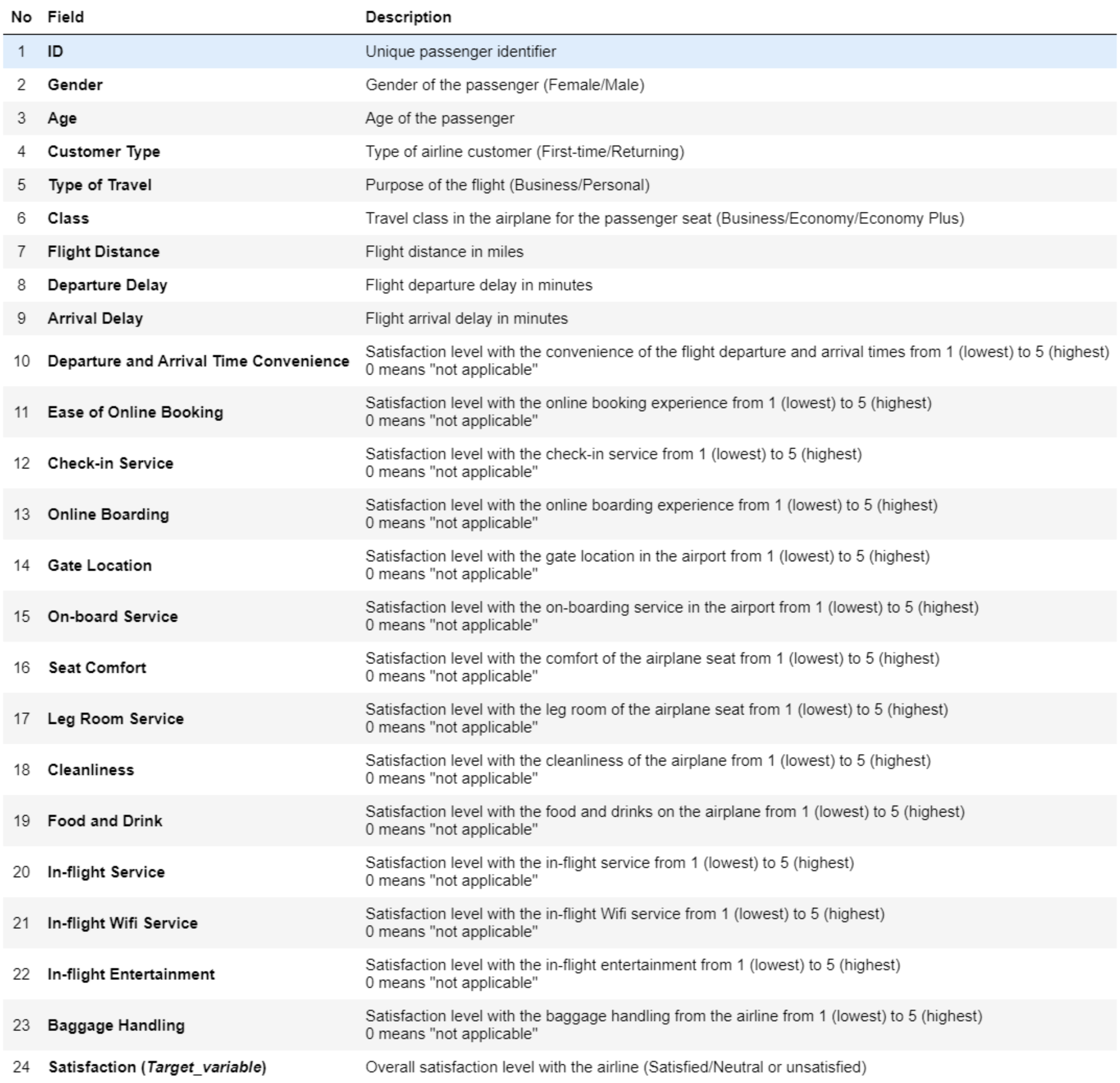
a) EFA

- EFA is a statistical method used to identify underlying factors that explain correlations among a set of observed variables.
- It is able to group data into categories known as factors, which significantly reduces the size of the data.
- It can help to uncover hidden patterns in the dataset as well.

METHODOLOGY

0 Data understanding and analyse logically

- Here is the information given in Kaggle



No	Field	Description
1	ID	Unique passenger identifier
2	Gender	Gender of the passenger (Female/Male)
3	Age	Age of the passenger
4	Customer Type	Type of airline customer (First-time/Returning)
5	Type of Travel	Purpose of the flight (Business/Personal)
6	Class	Travel class in the airplane for the passenger seat (Business/Economy/Economy Plus)
7	Flight Distance	Flight distance in miles
8	Departure Delay	Flight departure delay in minutes
9	Arrival Delay	Flight arrival delay in minutes
10	Departure and Arrival Time Convenience	Satisfaction level with the convenience of the flight departure and arrival times from 1 (lowest) to 5 (highest) 0 means "not applicable"
11	Ease of Online Booking	Satisfaction level with the online booking experience from 1 (lowest) to 5 (highest) 0 means "not applicable"
12	Check-in Service	Satisfaction level with the check-in service from 1 (lowest) to 5 (highest) 0 means "not applicable"
13	Online Boarding	Satisfaction level with the online boarding experience from 1 (lowest) to 5 (highest) 0 means "not applicable"
14	Gate Location	Satisfaction level with the gate location in the airport from 1 (lowest) to 5 (highest) 0 means "not applicable"
15	On-board Service	Satisfaction level with the on-boarding service in the airport from 1 (lowest) to 5 (highest) 0 means "not applicable"
16	Seat Comfort	Satisfaction level with the comfort of the airplane seat from 1 (lowest) to 5 (highest) 0 means "not applicable"
17	Leg Room Service	Satisfaction level with the leg room of the airplane seat from 1 (lowest) to 5 (highest) 0 means "not applicable"
18	Cleanliness	Satisfaction level with the cleanliness of the airplane from 1 (lowest) to 5 (highest) 0 means "not applicable"
19	Food and Drink	Satisfaction level with the food and drinks on the airplane from 1 (lowest) to 5 (highest) 0 means "not applicable"
20	In-flight Service	Satisfaction level with the in-flight service from 1 (lowest) to 5 (highest) 0 means "not applicable"
21	In-flight Wifi Service	Satisfaction level with the in-flight Wifi service from 1 (lowest) to 5 (highest) 0 means "not applicable"
22	In-flight Entertainment	Satisfaction level with the in-flight entertainment from 1 (lowest) to 5 (highest) 0 means "not applicable"
23	Baggage Handling	Satisfaction level with the baggage handling from the airline from 1 (lowest) to 5 (highest) 0 means "not applicable"
24	Satisfaction (*Target_variable*)	Overall satisfaction level with the airline (Satisfied/Neutral or unsatisfied)

- Intuitively, the variables can be divided into 6 different categories:
  - **(4) variables** are About customer's personal information (gender, age, customer type, type of travel)
  - **(2) variables** are About Flight type (class, flight distance)
  - **(3) variables** are About the flight punctuality (departure delay, arrival delay, departure and arrival time convenience)

  - **(7) variables** are About In-flight services (Cleanliness, Food and Drink, Seat Comfort, Leg Room Service, In-flight Service, In-flight Wifi Service, In-flight Entertainment)
  - **(5) variables** are About Other flight services (Check-in Service, Online Boarding, Ease of Online Booking, On-board Service, Baggage Handling)
  - **(1) variable** is miscellaneous (Gate Location)

- **Hypothesis:** all the **(4+2+3+1=10)** variables except variables about **'In-flight services'**, **'Other-flight services'** seem to be useless in improving/predicting the customers' satisfaction (gender, age, customer type, type of travel, class, flight distance, departure delay, arrival delay, departure and arrival time convenience, gate location)

1 Basic data pre-processing

- After the dataset is loaded, the variable ‘id’ is removed as it is useless and just simply states the unique identifier for each passenger, i.e. id numbers from 0 – 129879 in each row.
- Then, the 393 missing values in the variables “Arrival delay” have been filled by the mean of the current available values in the “Arrival delay” column

2 Descriptive Analysis

- Some basic and standard data analysis on the data have been obtained. For example, the count, mean, standard deviation, and so on.
- Then, the data is visualized using boxplots, kernel density estimates, histograms, and bivariate bar plots
  - The boxplots help us to identify outliers in the data
  - Kernel density estimates visualize the symmetry of the data
  - Histograms and bivariate bar plots show the satisfaction of customers depending on the variables
- After that, all values in categorical columns are converted to numerical values.
- Lastly, the correlation heatmap is plotted to display the correlation between variables.

3 Further pre-processing after analysis

- After some basic analysis, some further pre-processing on data is performed.
- This involves
  - ■ removing extreme outliers by quantile transformation,
  - ■ normalizing the data
  - ■ splitting the data into training and testing.

4 Linear Discriminant Analysis

- Linear Discriminant Analysis is performed to determine if there is enough information to predict the customer's satisfaction.
- This is done by observing the precision, recall and f1-score on the test dataset.

5 Feature Selection

- Some feature selection techniques/models are utilized to rank the features (1 being most significant, 22 being the least significant) based on their importance in predicting the customers' satisfaction.
  (1) Lasso Regression
  (2) SelectKBest in sklearn
    - ● Anova F-value method
    - ● Mutual Information
  (3) Permutation Importance using ANN
  (4) Recursive Feature Elimination
    - ○ Random Forest
    - ○ Gradient Boosting Classifier
    - ○ Logistics Regression
    - ○ Linear Support Vector Classification

Then, the average rankings will be obtained from (3) and (4) only (since 1 and 2 is from another category) for all variables and use the average ranking to see the final ranking. Lastly, top k features will be selected to perform Exploratory Factor Analysis.

6 Exploratory Factor Analysis (EFA)

The steps for EFA are as follows:

1. Check adequacy of data by using:
    a. Kaiser-Meyer-Olkin test (KMO score > 0.5 considers dataset to be adequate)
    b. Bartlett's test of Sphericity ($p < 0.01$ considers dataset to be adequate)
2. Calculate the correlation matrix
3. Fit the correlation matrix into the FactorAnalyzer() model
4. Obtain the number of factors using Kaiser-Guttman rule
5. Plot out Scree Plot
6. Fit the training data into the FactorAnalyzer model, and set rotations to difference orthogonal or oblique rotations
7. Tabulate the results of the different rotations and interpret the factors (giving the factors names)

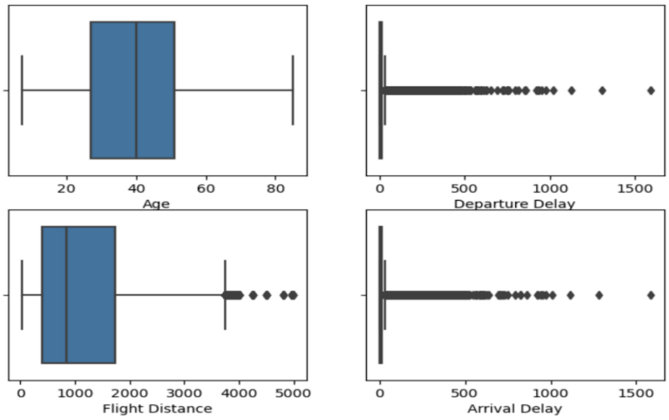
- After performing EFA, a random forest model is used on the transformed (rotated) data to see if all rotation types will yield high test accuracy (high accuracy → the transformed data preserves the original data's information).

- Lastly, the importance of the factors will be found based on features important in Random Forest. It will tell the airline what are the factors that they should focus on and prioritize first.

EXPERIMENTAL ANALYSIS

2 Descriptive Analysis

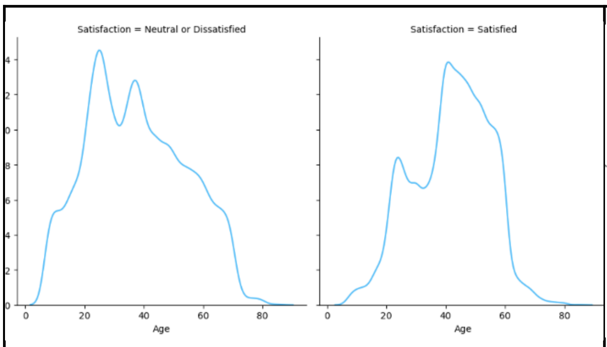
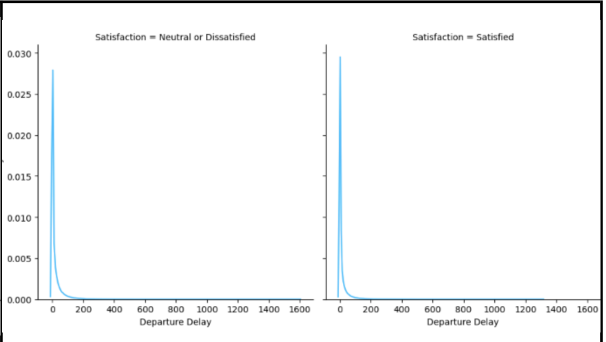
2.1 Boxplot for Numerical Variables

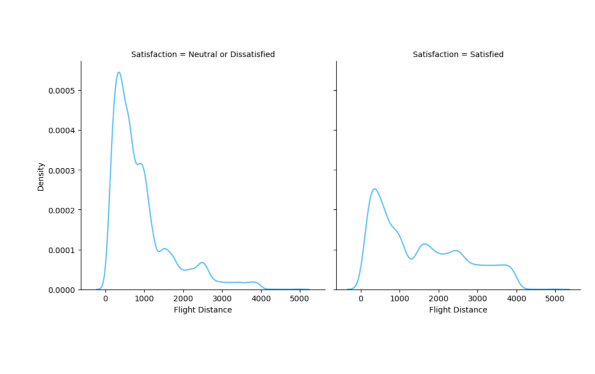
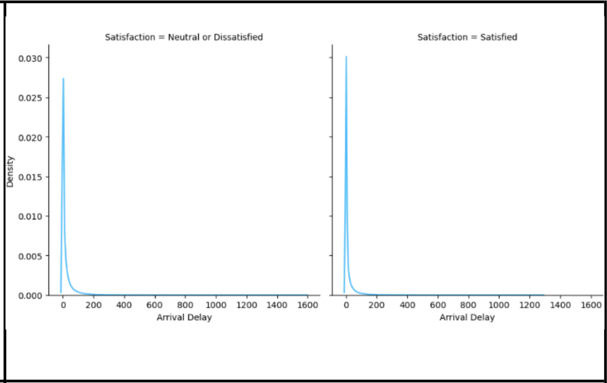


For the numerical variables, there are 4 variables to deal with. The boxplot will indicate and show any outliers in the variables. It also gives a clear understanding of the shape and distribution of the variables. It can be observed that there are many outliers in the variables “Flight distance”, “Departure Delay” and “Arrival Delay”. This could be a problem as the extreme outliers can cause our classification to be inaccurate. From this information, a technique such as quantile transformation is used to remove the outliers and replace it with new values. Besides that, by observing the graph, the overall data distribution of "Flight distance","Departure Delay" and "Arrival Delay" are heavily left skewed. Better visualization on skewness of data will be present in the KDE in the next section

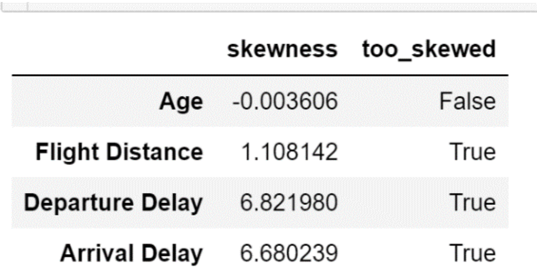
2.2 Kernel Density Estimate for Numerical Variable

We produce the graphs for our 4 numerical variables:

	
Age	Departure Delay

	
Flight Distance	Arrival Delays

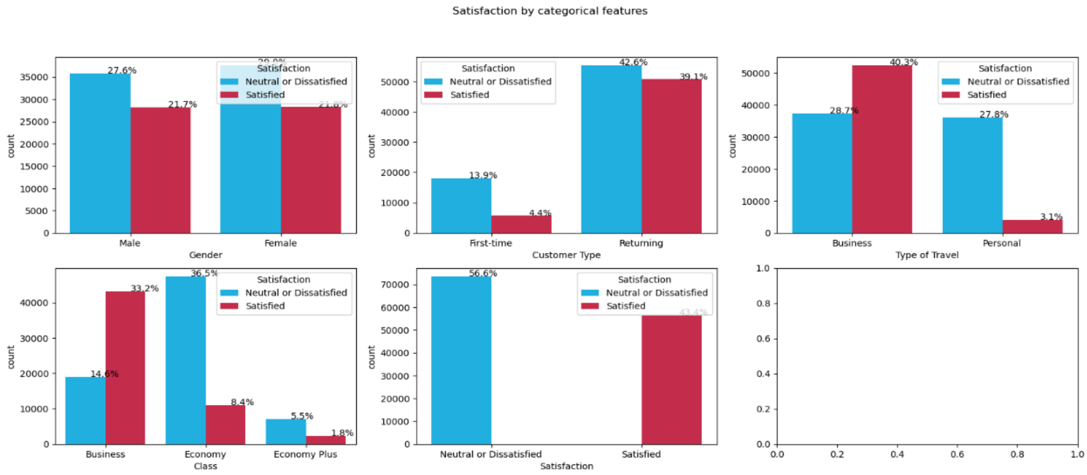
It seems that Age has a relatively symmetric shape compared to the other variables. To confirm this, the skewness is obtained by using df.skew() :

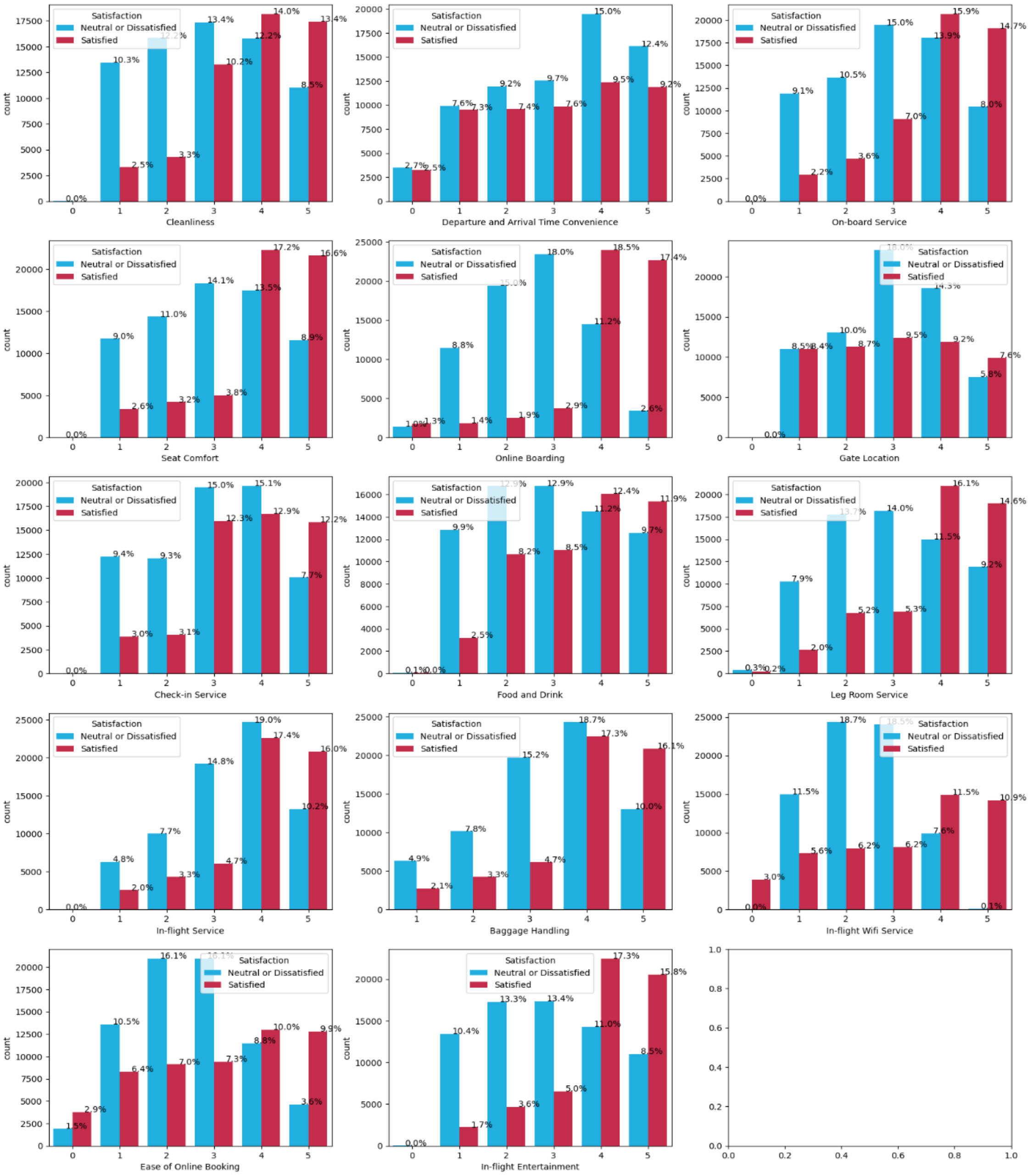


	skewness	too_skewed
Age	-0.003606	False
Flight Distance	1.108142	True
Departure Delay	6.821980	True
Arrival Delay	6.680239	True

It can now be confirmed that Age has very low skewness, while the other 3 variables are considered too skewed (|skewness| > 0.5). Quantile transformation will be applied to transform this data to have a gaussian probability distribution.

2.3 Histogram and Bivariate bar plot



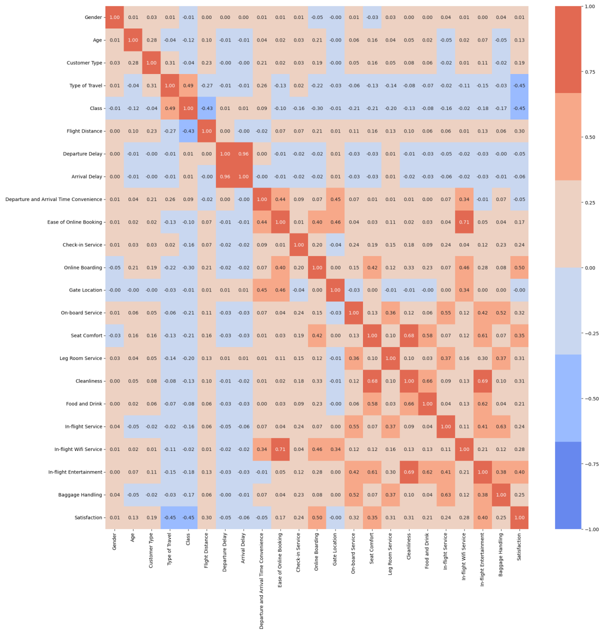


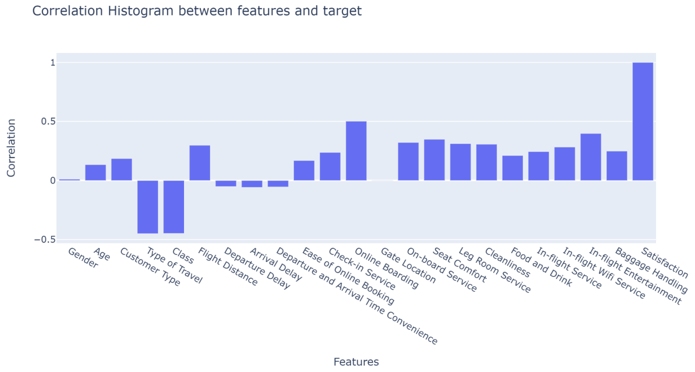
From the histograms above, we make the following conclusions:

Feature	Analysis
Gender	There are almost an equal number of males and females in our dataset. Both males and females have almost equal percentages when it comes to the neutral or dissatisfied and satisfied customers. Therefore, we can conclude that the variable gender most likely does not contribute in predicting satisfaction of a customer
Customer type	More than 80% of the passengers are returning customers, and the difference in satisfaction seems to be relatively small.

	□ Overall, people seemed to be more dissatisfied when they were returning customers.
Type of travel	□ The difference between whether it is a personal or business trip seems to heavily influence the satisfaction. □ When it is a business trip, there seems to be a higher satisfaction as all the expenses is paid by the company
Class	□ People were satisfied when it came to using business class. However, they became heavily biased to being neutral or dissatisfied in economy, and even in economy plus. □ This could show that the treatment and service that they receive in business class is very different from economy and economy plus.
Satisfaction	□ Overall, more people are neutral or dissatisfied rather than satisfied.
All other features (Likert Scale)	□ Overall, we can see that if the customers are dissatisfied or neutral, they usually give ratings between 1-3 □ Customers who are satisfied usually give ratings from 4-5. □ However, this trend doesn't show in "Departure and Arrival Time Convenience", where a majority of the customers are dissatisfied no matter the score being 1-5. But when observing the figures, we can see that the percentages are not that far apart.

2.4 Correlation Heatmap and Histogram





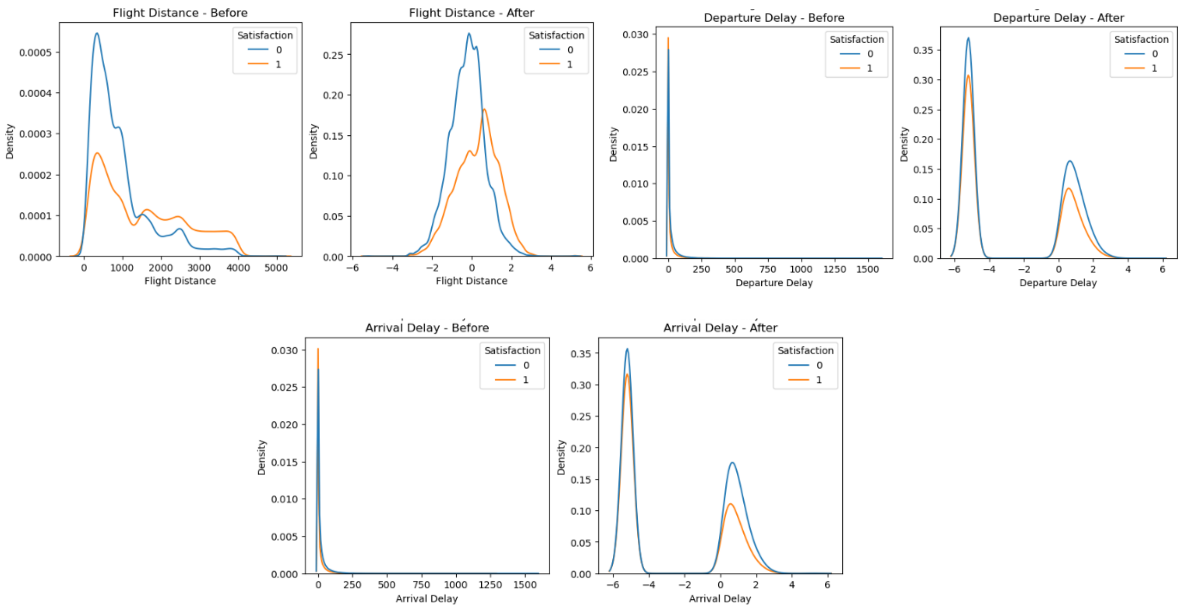
Based on the spectrum, redder means a higher correlation while bluer means a lower correlation. We ignore the diagonal reds and we observe variables with either relatively high or low correlation:

Pair of features with high positive correlation (> 0.6)		Possible reason
Ease of Online Booking	Wifi service	-
In flight Entertainment	Food and Drinks	Since the customer is unable to move around during the flight, in-flight entertainment will be appealing to the customer, and is typically paired along with food and drinks, and seat comfort
	Seat Comfort	
Cleanliness	Food and Drinks	Cleanliness has a high correlation with features regarding services that are provided during the flight. We can conclude that if cleanliness rating is high, the ratings for these 3 features will also be high
	Seat Comfort	
	In-flight Entertainment	
Departure Delay	Arrival Delay	If an airplane departs late, it will also arrive late.

- There are no features that are highly negative correlation with each other (< -0.6)
- There are no features that have high positive or negative correlation with the target variable, satisfaction
- It can be concluded that the relationship between the features and the target variable is likely to be non-linear

3 Further pre-processing after analysis

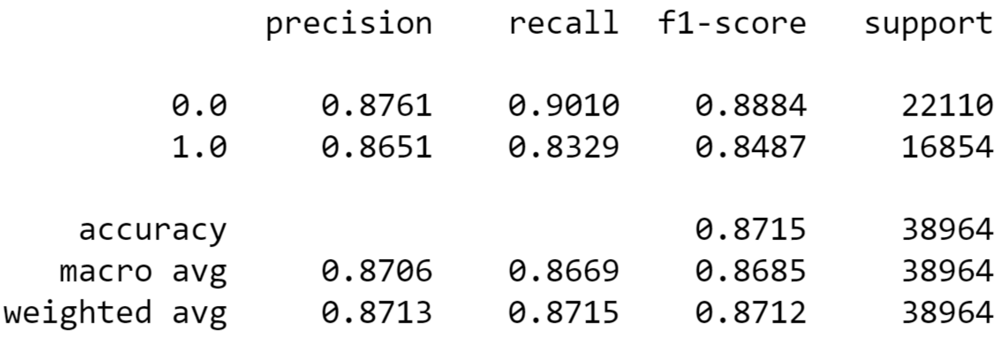
3.1 Quantile Transformed Data



Data after quantile transformation will have a normal distribution

4 Linear Discriminant Analysis

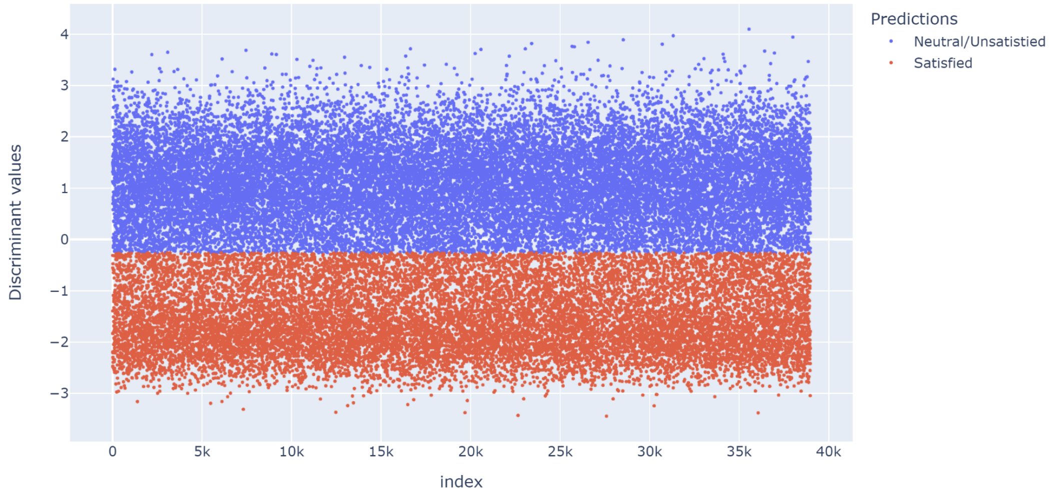
- The performance of LDA in the test dataset:



```
              precision    recall  f1-score   support

         0.0     0.8761    0.9010    0.8884     22110
         1.0     0.8651    0.8329    0.8487     16854

    accuracy                         0.8715     38964
   macro avg     0.8706    0.8669    0.8685     38964
weighted avg     0.8713    0.8715    0.8712     38964
```

The precision and recall is relatively high, giving a good f1 score as well

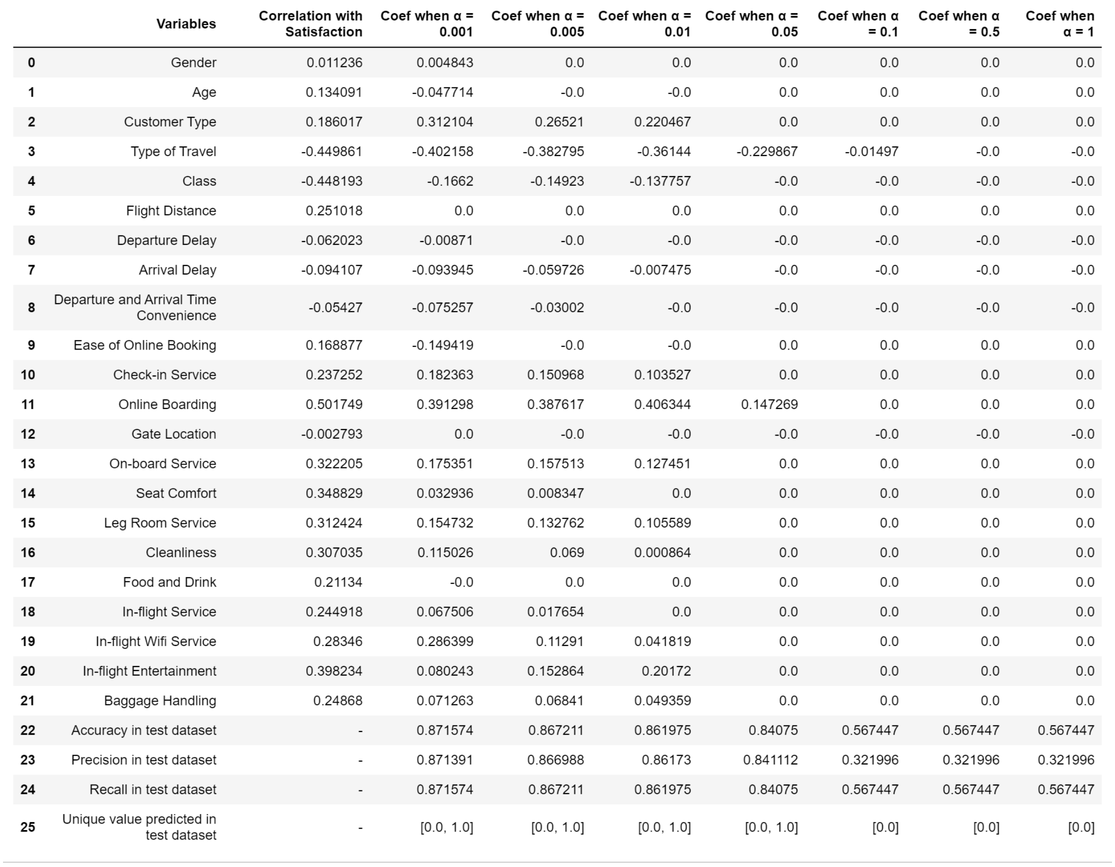
- It can be concluded that there is enough information in the dataset in predicting the customers' satisfaction using the features provided in the dataset.
- Scatter plot on the transformed dataset using LDA



The estimated decision boundary is between (-0.5, 0)

5 Feature Selection

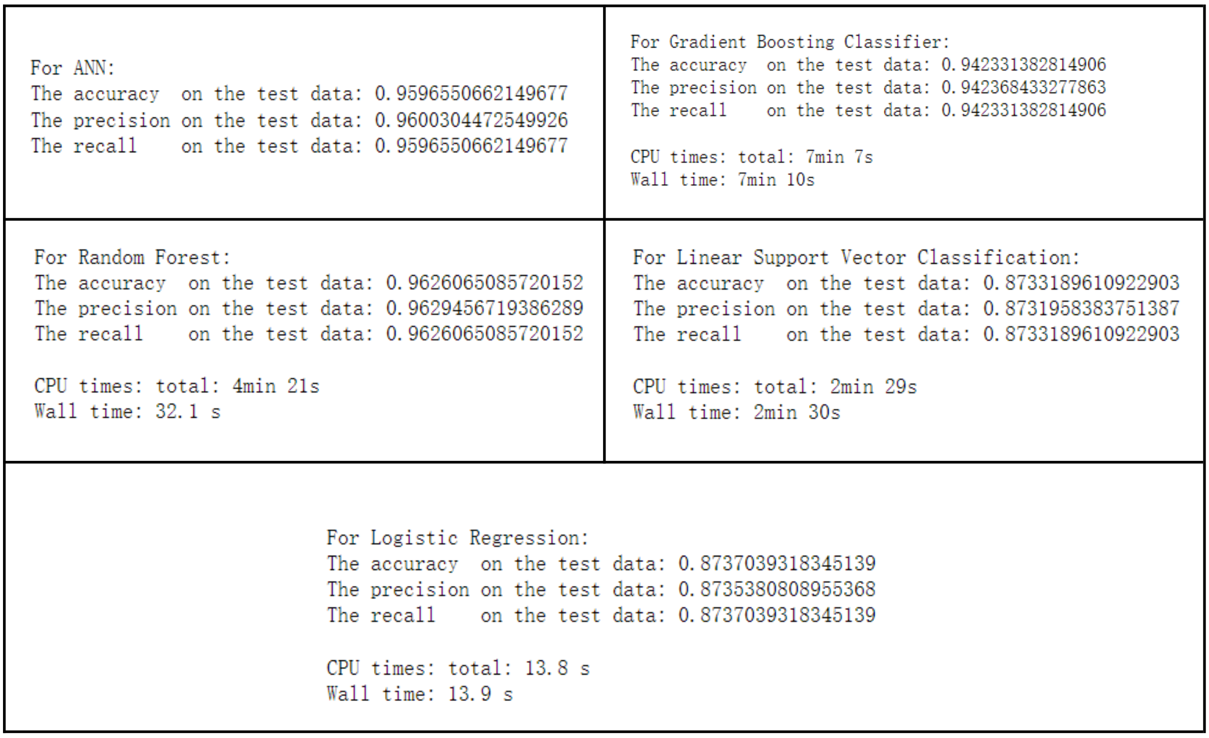
5.1 Lasso Regression



	Variables	Correlation with Satisfaction	Coef when α = 0.001	Coef when α = 0.005	Coef when α = 0.01	Coef when α = 0.05	Coef when α = 0.1	Coef when α = 0.5	Coef when α = 1
0	Gender	0.011236	0.004843	0.0	0.0	0.0	0.0	0.0	0.0
1	Age	0.134091	-0.047714	-0.0	-0.0	0.0	0.0	0.0	0.0
2	Customer Type	0.186017	0.312104	0.26521	0.220467	0.0	0.0	0.0	0.0
3	Type of Travel	-0.449861	-0.402158	-0.382795	-0.36144	-0.229867	-0.01497	-0.0	-0.0
4	Class	-0.448193	-0.1662	-0.14923	-0.137757	-0.0	-0.0	-0.0	-0.0
5	Flight Distance	0.251018	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	Departure Delay	-0.062023	-0.00871	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
7	Arrival Delay	-0.094107	-0.093945	-0.059726	-0.007475	-0.0	-0.0	-0.0	-0.0
8	Departure and Arrival Time Convenience	-0.05427	-0.075257	-0.03002	-0.0	-0.0	-0.0	-0.0	-0.0
9	Ease of Online Booking	0.168877	-0.149419	-0.0	-0.0	0.0	0.0	0.0	0.0
10	Check-in Service	0.237252	0.182363	0.150968	0.103527	0.0	0.0	0.0	0.0
11	Online Boarding	0.501749	0.391298	0.387617	0.406344	0.147269	0.0	0.0	0.0
12	Gate Location	-0.002793	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
13	On-board Service	0.322205	0.175351	0.157513	0.127451	0.0	0.0	0.0	0.0
14	Seat Comfort	0.348829	0.032936	0.008347	0.0	0.0	0.0	0.0	0.0
15	Leg Room Service	0.312424	0.154732	0.132762	0.105589	0.0	0.0	0.0	0.0
16	Cleanliness	0.307035	0.115026	0.069	0.000864	0.0	0.0	0.0	0.0
17	Food and Drink	0.21134	-0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	In-flight Service	0.244918	0.067506	0.017654	0.0	0.0	0.0	0.0	0.0
19	In-flight Wifi Service	0.28346	0.286399	0.11291	0.041819	0.0	0.0	0.0	0.0
20	In-flight Entertainment	0.398234	0.080243	0.152864	0.20172	0.0	0.0	0.0	0.0
21	Baggage Handling	0.24868	0.071263	0.06841	0.049359	0.0	0.0	0.0	0.0
22	Accuracy in test dataset	-	0.871574	0.867211	0.861975	0.84075	0.567447	0.567447	0.567447
23	Precision in test dataset	-	0.871391	0.866988	0.86173	0.841112	0.321996	0.321996	0.321996
24	Recall in test dataset	-	0.871574	0.867211	0.861975	0.84075	0.567447	0.567447	0.567447
25	Unique value predicted in test dataset	-	[0.0, 1.0]	[0.0, 1.0]	[0.0, 1.0]	[0.0, 1.0]	[0.0]	[0.0]	[0.0]

- Variables **"Flight Distance", "Food and Drink" and "Gate Location"** are the least important features among 22 features in predicting the satisfaction in lasso regression.
- This is because the coefficient of these variables is always 0 for different α values.
- Besides that, we can observe a slight decrease of accuracy in the test dataset when some of the other features' coefficient is set to be near 0.
  - For example: When "Gender, Age, Departure Delay, Ease of Online Booking" is set to near 0 (from α = 0.001 to α = 0.005), the test accuracy decrease from 0.871574 to 0.867211
- However, this **does not** imply that the variables "Flight Distance", "Food and Drink" and "Gate Location" can be dropped. There may exist a non-linear relationship between these variables and satisfaction.
- Therefore, more feature selection techniques will be done to drop the useless and unnecessary features and retain the most important features in predicting the customer's satisfaction

5.2 Performance for each model in test dataset



```
For ANN:
The accuracy  on the test data: 0.9596550662149677
The precision on the test data: 0.9600304472549926
The recall    on the test data: 0.9596550662149677
```

```
For Gradient Boosting Classifier:
The accuracy  on the test data: 0.942331382814906
The precision on the test data: 0.942368433277863
The recall    on the test data: 0.942331382814906

CPU times: total: 7min 7s
Wall time: 7min 10s
```

```
For Random Forest:
The accuracy  on the test data: 0.9626065085720152
The precision on the test data: 0.9629456719386289
The recall    on the test data: 0.9626065085720152

CPU times: total: 4min 21s
Wall time: 32.1 s
```

```
For Linear Support Vector Classification:
The accuracy  on the test data: 0.8733189610922903
The precision on the test data: 0.8731958383751387
The recall    on the test data: 0.8733189610922903

CPU times: total: 2min 29s
Wall time: 2min 30s
```

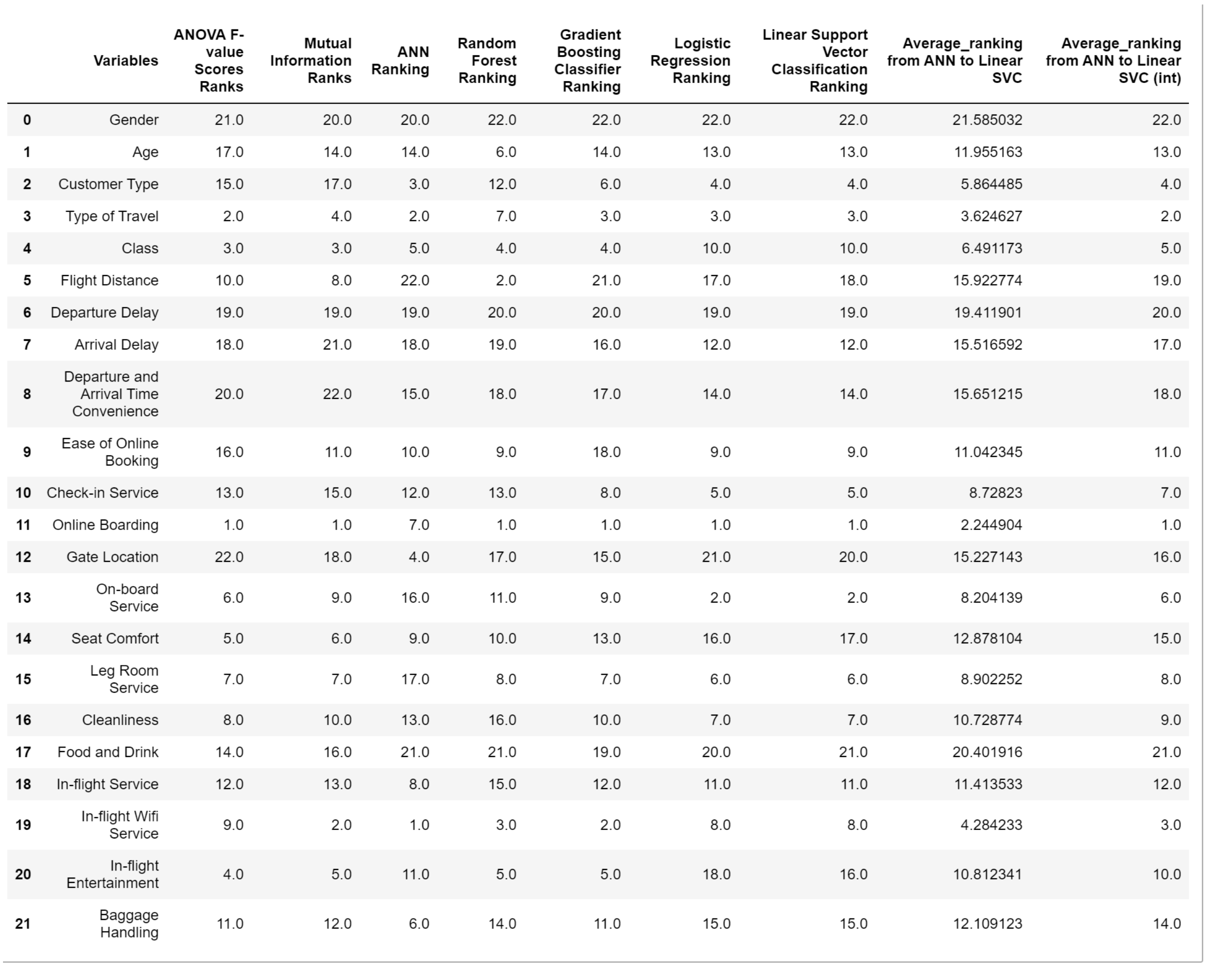
```
For Logistic Regression:
The accuracy  on the test data: 0.8737039318345139
The precision on the test data: 0.8735380808955368
The recall    on the test data: 0.8737039318345139

CPU times: total: 13.8 s
Wall time: 13.9 s
```

Overall, the test accuracy, recall, and precision of the models are as follows (in descending order):

1. Random Forest
2. ANN
3. Gradient Boosting Classifier
4. Logistic Regression
5. Linear Support Vector Classification

Therefore, there will be more trust placed in the model's ranking with better performance on the test dataset. This is done by using the exponential function to place more weights on the models with higher accuracy when computing the average ranking.

5.3 Overall ranking on different feature selection models



	Variables	ANOVA F-value Scores Ranks	Mutual Information Ranks	ANN Ranking	Random Forest Ranking	Gradient Boosting Classifier Ranking	Logistic Regression Ranking	Linear Support Vector Classification Ranking	Average_ranking from ANN to Linear SVC	Average_ranking from ANN to Linear SVC (int)
0	Gender	21.0	20.0	20.0	22.0	22.0	22.0	22.0	21.585032	22.0
1	Age	17.0	14.0	14.0	6.0	14.0	13.0	13.0	11.955163	13.0
2	Customer Type	15.0	17.0	3.0	12.0	6.0	4.0	4.0	5.864485	4.0
3	Type of Travel	2.0	4.0	2.0	7.0	3.0	3.0	3.0	3.624627	2.0
4	Class	3.0	3.0	5.0	4.0	4.0	10.0	10.0	6.491173	5.0
5	Flight Distance	10.0	8.0	22.0	2.0	21.0	17.0	18.0	15.922774	19.0
6	Departure Delay	19.0	19.0	19.0	20.0	20.0	19.0	19.0	19.411901	20.0
7	Arrival Delay	18.0	21.0	18.0	19.0	16.0	12.0	12.0	15.516592	17.0
8	Departure and Arrival Time Convenience	20.0	22.0	15.0	18.0	17.0	14.0	14.0	15.651215	18.0
9	Ease of Online Booking	16.0	11.0	10.0	9.0	18.0	9.0	9.0	11.042345	11.0
10	Check-in Service	13.0	15.0	12.0	13.0	8.0	5.0	5.0	8.72823	7.0
11	Online Boarding	1.0	1.0	7.0	1.0	1.0	1.0	1.0	2.244904	1.0
12	Gate Location	22.0	18.0	4.0	17.0	15.0	21.0	20.0	15.227143	16.0
13	On-board Service	6.0	9.0	16.0	11.0	9.0	2.0	2.0	8.204139	6.0
14	Seat Comfort	5.0	6.0	9.0	10.0	13.0	16.0	17.0	12.878104	15.0
15	Leg Room Service	7.0	7.0	17.0	8.0	7.0	6.0	6.0	8.902252	8.0
16	Cleanliness	8.0	10.0	13.0	16.0	10.0	7.0	7.0	10.728774	9.0
17	Food and Drink	14.0	16.0	21.0	21.0	19.0	20.0	21.0	20.401916	21.0
18	In-flight Service	12.0	13.0	8.0	15.0	12.0	11.0	11.0	11.413533	12.0
19	In-flight Wifi Service	9.0	2.0	1.0	3.0	2.0	8.0	8.0	4.284233	3.0
20	In-flight Entertainment	4.0	5.0	11.0	5.0	5.0	18.0	16.0	10.812341	10.0
21	Baggage Handling	11.0	12.0	6.0	14.0	11.0	15.0	15.0	12.109123	14.0

1 implies most important, 22 implies least important

- From the table above, for the column “Average_ranking from ANN to Linear SVC”
  - In terms of flight services, the variable that contributes the most in predicting the satisfaction is **Online Boarding** (rank 1*st*), followed by Type of Travel, In-flight Wifi Service, ...
  - The variable in ’flight services‘ that contributes the least in predicting satisfaction is **Food and Drink** (rank 20*th*)
- So, some decisions have been made:
  - 7 features that have **average_ranking > threshold(15**) will be dropped before performing factor analysis to easier the analysis process i.e. ['Gender' 'Flight Distance' 'Departure Delay' 'Arrival Delay' 'Departure and Arrival Time Convenience' 'Gate Location' 'Food and Drink']
  - Among them,

        - **(1)** variables are from customer's personal information (Gender)
        - **(1)** variables is from Flight type (Flight distance)
        - **(3)** variables is from whether the flight punctual or not (departure delay, arrival delay, departure and arrival time convenience)
        - **(1)** variables is from In-flight services (Food and Drink)
        - **(1)** variables is from miscellaneous (Gate Location)

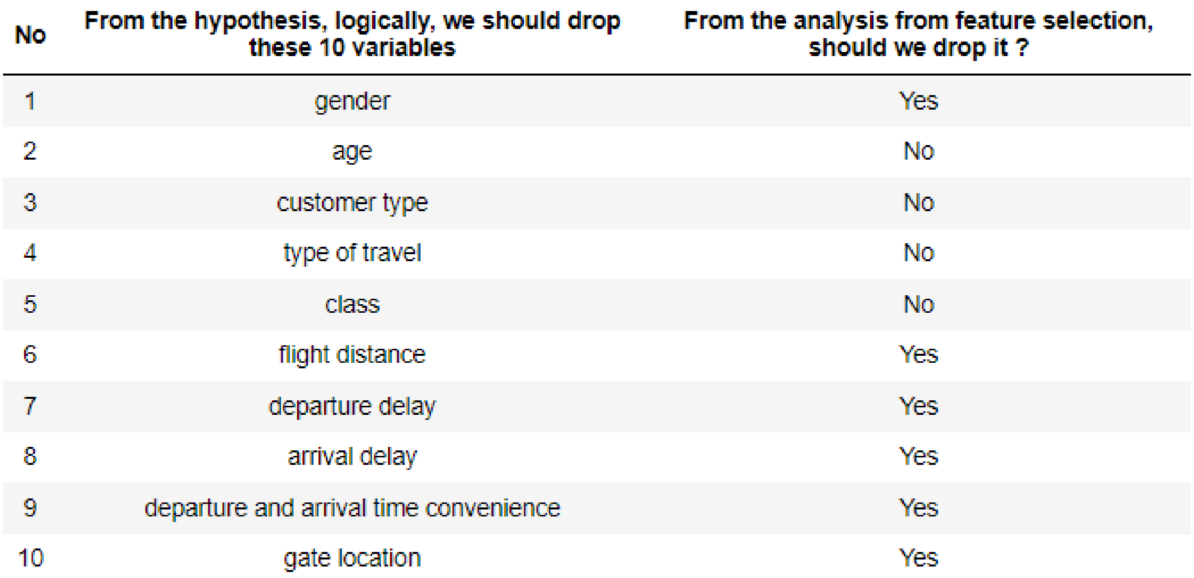
- The reason of why threshold is selected as 15:
    - Variables with 10 <= ranking <=15 are ['Age', 'Ease of Online Booking', 'Seat Comfort', 'Cleanliness', 'In-flight Service', 'Baggage Handling'].
        - They are all variables **related to flight services** provided except Age (ranking 13*th*).
        - It seems that they will **likely** contribute in forming the factors that relate to flight-services.
    - Variables with ranking > 15 are ['Gender', 'Flight Distance', 'Departure Delay', 'Arrival Delay', 'Departure and Arrival Time Convenience', 'Gate Location', 'Food and Drink'].
        - There are all variables **not related to flight services** except Food and Drink (ranking 20*th*).
        - It seems that they will **not likely** contribute in forming the factors that related to flight-services

- The possible reasons as to why these 7 features ['Gender' 'Flight Distance' 'Departure Delay' 'Arrival Delay' 'Departure and Arrival Time Convenience' 'Gate Location' 'Food and Drink'] does not contribute much in predicting the satisfaction of customers are as below:

Variables	Possible Reason
Gender	Typically, the gender will not affect the results since it's not related to the features of the flight or the service provided. This means that the satisfaction will not be affected
Flight Distance	Most people booking the flight should already know the distance and travel

	time of their journey, so this shouldn't affect the satisfaction of the passengers. This feature is also not really related to the service provided by the airline
Departure Delay	The departure delay doesn't really affect the satisfaction because quite often, it is not the airlines' fault that a flight has been delayed. This issue is usually due to high traffic in the airport, with many planes wanting to take off from the same airport and using the same route. When it's related to the airline, like a component of the airplane malfunctioning, it is very rare and often doesn't happen.
Arrival Delay	This is related to departure delay. If there is a departure delay, there will definitely be an arrival delay. However, as stated above, it is typically not the airlines' fault, so it doesn't affect the customer's satisfaction with the airline
Departure and Arrival Time Convenience	The passenger is the one who books the flights, so they are the ones selecting their departure and arrival time. The airline does give options for different times with different prices, but is ultimately up to the customer as to whether they want a more convenient time for a higher price or a less convenient time with a lower price
Gate Location	Whether the customer enter at which gate location, it will not affect the customer's satisfaction on the flight-services unless there is something wrong at the gate location
Food and drinks	There are 2 aspects when it comes to the food and drinks, which are quality and the price of the food. In terms of quality, customers usually don't expect much from airline food as it is all prepackaged and typically there is nothing unique about the food being served. In terms of price, there is food often served for free, especially packaged snacks. On long flights, airlines will usually provide a meal for every customer inclusive of their ticket price, so customers' satisfaction of food in terms of price should not be affected. Essentially, customers typically do not expect very much from the food and is not considered an important feature in terms of satisfaction as they might be more focussed on the other services provided

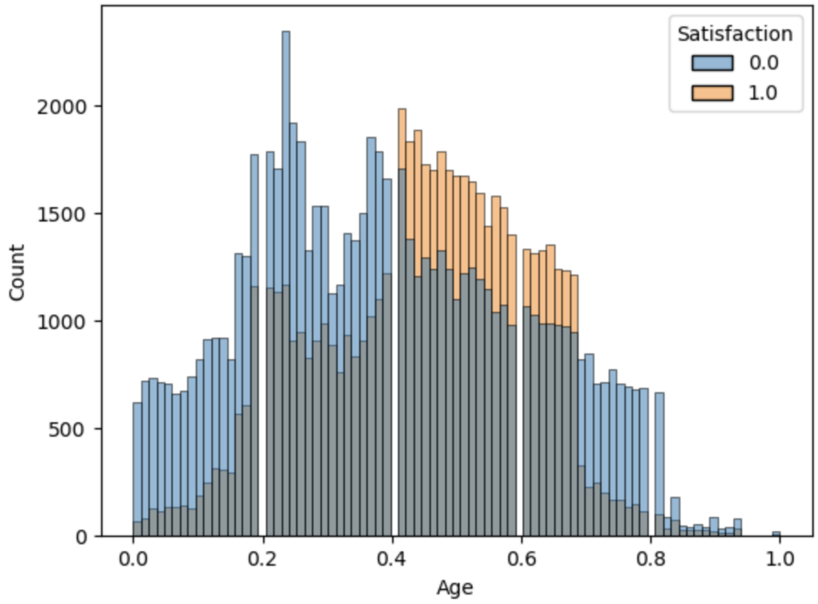
Based on this analysis, there are some contradictions with the **hypothesis**:



No	From the hypothesis, logically, we should drop these 10 variables	From the analysis from feature selection, should we drop it ?
1	gender	Yes
2	age	No
3	customer type	No
4	type of travel	No
5	class	No
6	flight distance	Yes
7	departure delay	Yes
8	arrival delay	Yes
9	departure and arrival time convenience	Yes
10	gate location	Yes

The possible reason of why these 4 features ['Age','Customer Type' 'Type of Travel' 'Class' 'Arrival Delay'] contribute much in predicting satisfaction of customer is as below:

<table>
<tr><td>Variables</td><td>Possible reasons</td></tr>
<tr><td>Age</td><td>It ranges from 7 to 85 years old. Different age groups have different percentages of satisfaction. For example, when the customer is about 40 years old, most of them will be satisfied with the airline services. (This can be shown in the histogram plot at below)

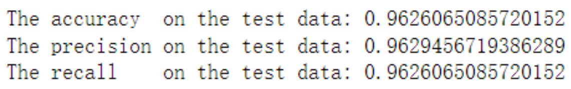
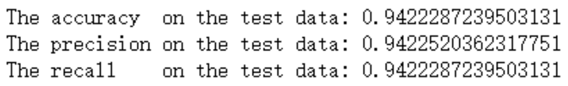

</td></tr>
</table>

Customer Type	It can be (First-time/Returning). If the customer is returning to the same airline for their travels, it means that the airline is doing a good job in attracting the customers to use their airlines instead of competitors
Type of Travel	It can be (Business/Personal) As we have mentioned above, business trips are typically paid for. Therefore, this means that customers on business trips get to enjoy the benefits for free. On the contrary, personal trip customers have high expectations as they have paid for the service and want to get their money's worth
Class	It can be (Business/Economy/Economy Plus) For each class, there is a different level of service and experience. The service in Economy tends to be the worst as airlines place no priority on these passengers. On the contrary, passengers in business will get the best service as they have paid a high amount of money. Therefore, the better the service, the more satisfied a customer will be, and the service is dependent on which class the passenger is in

- However, we will drop the 'age','customer type' and 'type of travel' as it is not helping in improving the airline services. For variables 'class', we will not drop it as it seems to be related to the flight services (as different classes will have different distributions of satisfaction). For example, when the class="Business", the people will be satisfied overall (This can be shown at the bivariate bar plot at Experimental Analysis - Section 2.3).

5.4 Remove some features before factor Analysis

- From the previous section, the following **12 variables** will be selected to perform **EFA**:
  - ['Class', 'Ease of Online Booking', 'Check-in Service', 'Online Boarding', 'On-board Service', 'Seat Comfort', 'Leg Room Service', 'Cleanliness', 'In-flight Service,' 'In-flight Wifi Service', 'In-flight Entertainment', 'Baggage Handling']
- Below shows that even after removing the variables, it is still very accurate and that the features that were removed have little contribution in predicting the satisfaction of the customer.

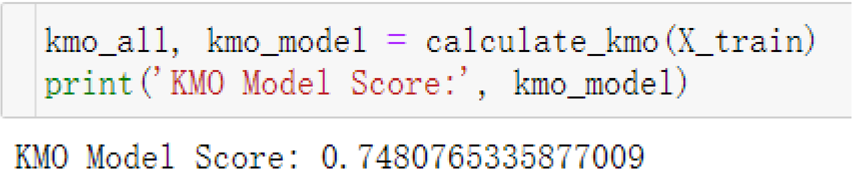
 The accuracy on the test data: 0.9626065085720152 The precision on the test data: 0.9629456719386289 The recall on the test data: 0.9626065085720152	 The accuracy on the test data: 0.9422287239503131 The precision on the test data: 0.9422520362317751 The recall on the test data: 0.9422287239503131
Before dropping the variables, result of Random Forest on test dataset	After dropping some variables, result of Random Forest on test dataset

Accuracy, precision and recall on test dataset before and after dropping the variables

6 Exploratory Factor Analysis (EFA)

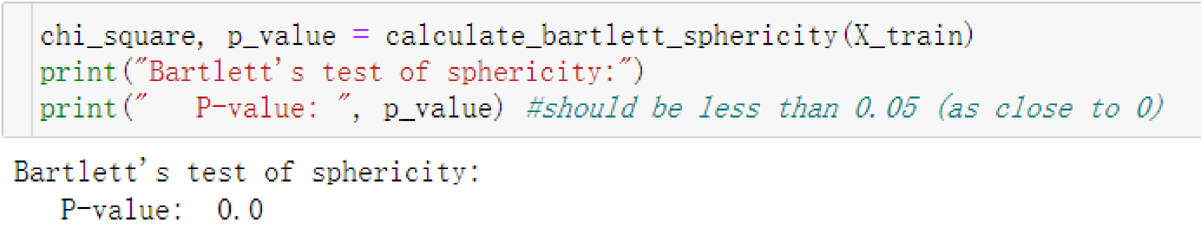
6.1 Perform EFA

- To begin with, the sample size needs to be adequate in relation to the number of variables that we have. Two tests (Kaiser-Meyer-Olkin Bartlett's test of sphericity) and are performed, and the results are:



```
kmo_all, kmo_model = calculate_kmo(X_train)
print('KMO Model Score:', kmo_model)
```

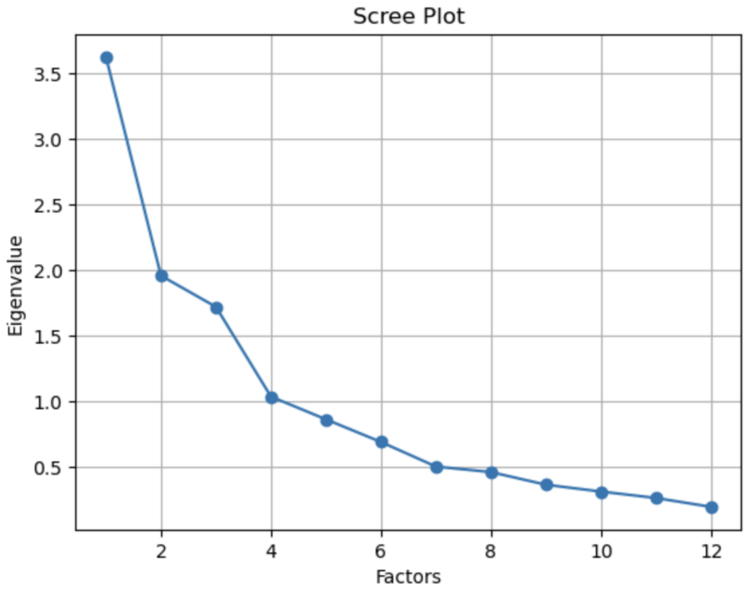
```
KMO Model Score: 0.7480765335877009
```



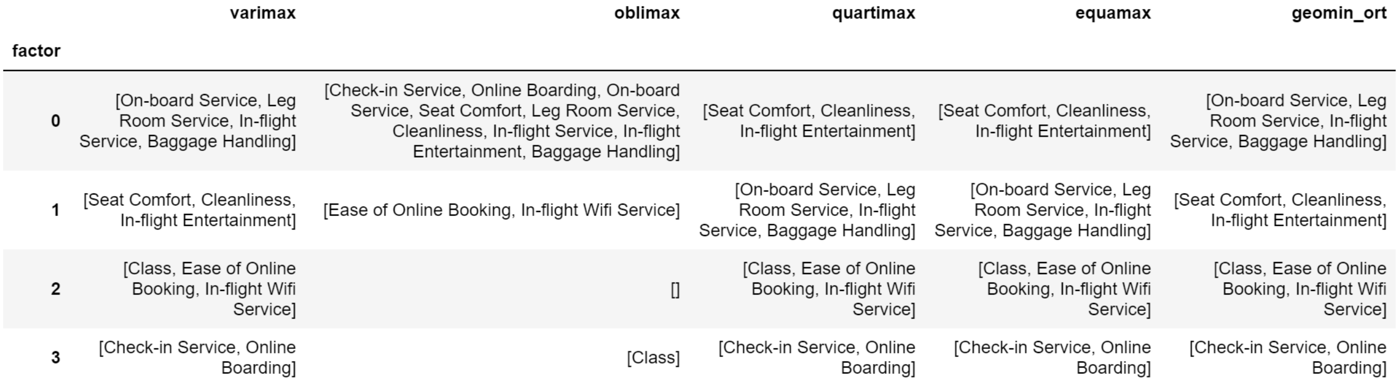
```
chi_square, p_value = calculate_bartlett_sphericity(X_train)
print("Bartlett's test of sphericity:")
print("   P-value: ", p_value) #should be less than 0.05 (as close to 0)
```

```
Bartlett's test of sphericity:
   P-value:  0.0
```

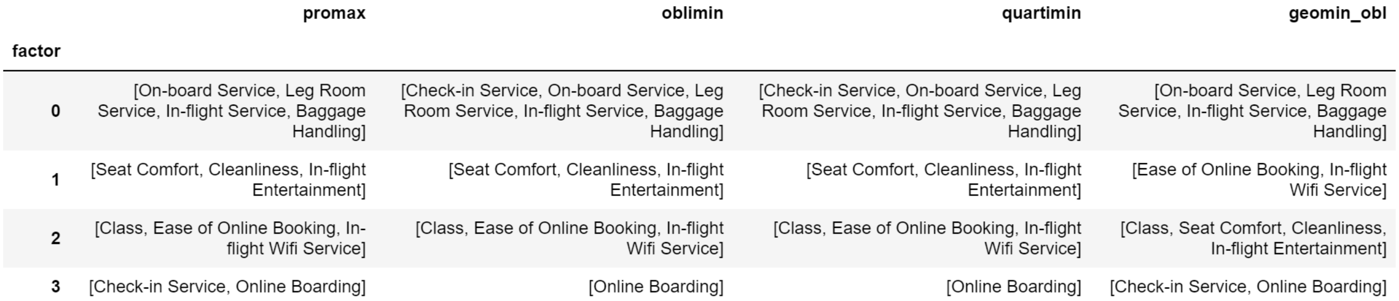
- It can be concluded that there is an adequate amount of data, since the KMO model score is > 0.5 and the p-value is < 0.01 (significant value).
- The Kaiser-Guttman rule and scree plot shows the number of factors generated, which is 4.



- There are a total of 9 rotations that can be done in the FactorAnalyzer model. Since we are not sure which will be the optimal rotation, we use all 9 rotations and decide later on which is the best to use.
  - The orthogonal rotations are ['varimax', 'oblimax', 'quartimax', 'equamax', 'geomin_ort']
  - The oblique rotations are ['promax', 'oblimin', 'quartimin', 'geomin_obl']
- Below is the results of all 9 rotations:



factor	varimax	oblimax	quartimax	equamax	geomin_ort
0	[On-board Service, Leg Room Service, In-flight Service, Baggage Handling]	[Check-in Service, Online Boarding, On-board Service, Seat Comfort, Leg Room Service, Cleanliness, In-flight Service, In-flight Entertainment, Baggage Handling]	[Seat Comfort, Cleanliness, In-flight Entertainment]	[Seat Comfort, Cleanliness, In-flight Entertainment]	[On-board Service, Leg Room Service, In-flight Service, Baggage Handling]
1	[Seat Comfort, Cleanliness, In-flight Entertainment]	[Ease of Online Booking, In-flight Wifi Service]	[On-board Service, Leg Room Service, In-flight Service, Baggage Handling]	[On-board Service, Leg Room Service, In-flight Service, Baggage Handling]	[Seat Comfort, Cleanliness, In-flight Entertainment]
2	[Class, Ease of Online Booking, In-flight Wifi Service]	[]	[Class, Ease of Online Booking, In-flight Wifi Service]	[Class, Ease of Online Booking, In-flight Wifi Service]	[Class, Ease of Online Booking, In-flight Wifi Service]
3	[Check-in Service, Online Boarding]	[Class]	[Check-in Service, Online Boarding]	[Check-in Service, Online Boarding]	[Check-in Service, Online Boarding]



factor	promax	oblimin	quartimin	geomin_obl
0	[On-board Service, Leg Room Service, In-flight Service, Baggage Handling]	[Check-in Service, On-board Service, Leg Room Service, In-flight Service, Baggage Handling]	[Check-in Service, On-board Service, Leg Room Service, In-flight Service, Baggage Handling]	[On-board Service, Leg Room Service, In-flight Service, Baggage Handling]
1	[Seat Comfort, Cleanliness, In-flight Entertainment]	[Seat Comfort, Cleanliness, In-flight Entertainment]	[Seat Comfort, Cleanliness, In-flight Entertainment]	[Ease of Online Booking, In-flight Wifi Service]
2	[Class, Ease of Online Booking, In-flight Wifi Service]	[Class, Ease of Online Booking, In-flight Wifi Service]	[Class, Ease of Online Booking, In-flight Wifi Service]	[Class, Seat Comfort, Cleanliness, In-flight Entertainment]
3	[Check-in Service, Online Boarding]	[Online Boarding]	[Online Boarding]	[Check-in Service, Online Boarding]

6.2 Interpretation of EFA

- Based on the final results, it is evident that different rotation techniques yield varying factor loadings.
- The **varimax, quartimax, equamax, geomin_ort** (orthogonal rotation), and **promax rotation** (oblique rotation) techniques **produce similar factors**. When the factors produced by oblique rotation resemble those produced by orthogonal rotation, it indicates that the factors are nearly orthogonal/uncorrelated.
- Similarly, **the oblimin and quartimin** techniques yield **similar factors**.
- Moreover, **the oblimax and geomin_obl** techniques yield **distinct factors**.
- The table below displays the factor naming conventions by us for each rotation technique.

Factor	varimax / promax / geomin_ort	quartimax / equamax	oblimax	oblimin / quartimin	geomin_obl
0	Overall Service Quality	Comfort and Entertainment	Overall Service Quality	Overall Service Quality	Overall Service Quality
1	Comfort and Entertainment	Overall Service Quality	Technology services	Comfort and Entertainment	Connectivity
2	Travel Class and Connectivity	Travel Class and Connectivity	-	Travel Class and Connectivity	Comfort and Entertainment
3	Check-in and boarding services	Check-in and boarding services	Class	Online Boarding	Check-in and boarding services

- From the table above, all possible naming of the factors are: {Overall service quality, Comfort and Entertainment, Travel Class and Connectivity, Check-in and boarding services, Technology services, Connectivity} ∪ {Class, Online Boarding} - Original variable name
- Based on these factors, we interpret the possible features in these factors:

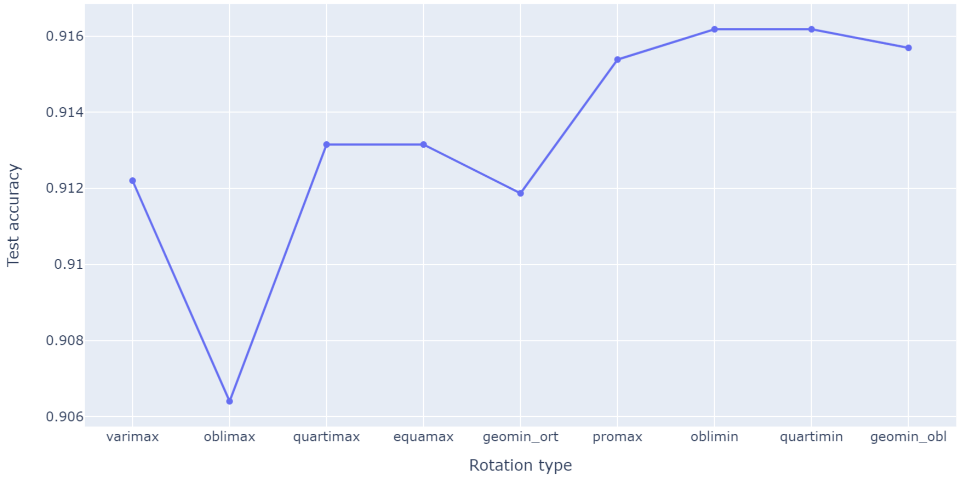
Factor	**Possible Features in the Factor that are not in the dataset**
Overall service quality	Power Outlets and USB Ports, ...
Comfort and Entertainment	Cabin Air Quality, Cabin Layout and Design, Audio Options, ...

Travel Class and Connectivity	Mood Lighting, Real-Time Flight Tracking, ...
Check-in and boarding services	Online Check-in services, Offline check-in services, Automated Boarding Gates services, ...
Technology services	Seat Controls, Electronic Amenities, ...
Connectivity	Real-Time Flight Tracking, ...

- By identifying the features associated with each factor, airlines can enhance customer satisfaction by improving those features.

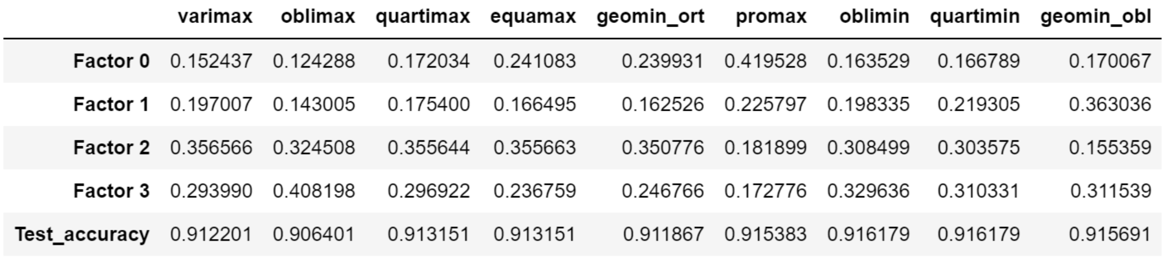
6.3 Random Forest on the transformed dataset

- This section focuses on building a model using the transformed dataset (i.e., data after rotation) and evaluating its performance on a test dataset. This analysis helps us assess how well each rotation technique captures the variances in the features.
- Since oblique rotation maximizes the factor loadings without assuming orthogonality between factors, data transformed using oblique rotation is expected to achieve higher accuracy in the model and capture more variance in the features.
- Moreover, as Random Forest demonstrated the highest performance in the feature selection section, we will use Random Forest classification to evaluate the model's performance on the test dataset.



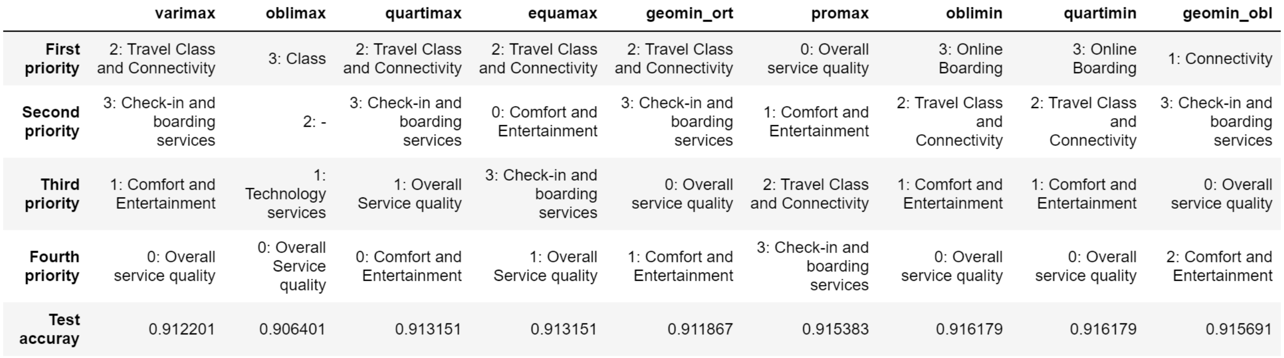
- From the evaluation graph shown, we can observe that the results from each rotation has roughly 91% accuracy on test dataset (Originally, before transformed, the data performance at test dataset is about 94%)
- This implies that each rotation managed to perform well in capturing the variance in each feature.

6.4 Feature Importance in each factor



	varimax	oblimax	quartimax	equamax	geomin_ort	promax	oblimin	quartimin	geomin_obl
Factor 0	0.152437	0.124288	0.172034	0.241083	0.239931	0.419528	0.163529	0.166789	0.170067
Factor 1	0.197007	0.143005	0.175400	0.166495	0.162526	0.225797	0.198335	0.219305	0.363036
Factor 2	0.356566	0.324508	0.355644	0.355663	0.350776	0.181899	0.308499	0.303575	0.155359
Factor 3	0.293990	0.408198	0.296922	0.236759	0.246766	0.172776	0.329636	0.310331	0.311539
Test_accuracy	0.912201	0.906401	0.913151	0.913151	0.911867	0.915383	0.916179	0.916179	0.915691

- The table above displays the distribution of feature (factor) scores for each rotation. By sorting the scores for each rotation, we can determine the order in which to prioritize improvements for each factor.



	varimax	oblimax	quartimax	equamax	geomin_ort	promax	oblimin	quartimin	geomin_obl
First priority	2: Travel Class and Connectivity	3: Class	2: Travel Class and Connectivity	2: Travel Class and Connectivity	2: Travel Class and Connectivity	0: Overall service quality	3: Online Boarding	3: Online Boarding	1: Connectivity
Second priority	3: Check-in and boarding services	2: -	3: Check-in and boarding services	0: Comfort and Entertainment	3: Check-in and boarding services	1: Comfort and Entertainment	2: Travel Class and Connectivity	2: Travel Class and Connectivity	3: Check-in and boarding services
Third priority	1: Comfort and Entertainment	1: Technology services	1: Overall Service quality	3: Check-in and boarding services	0: Overall service quality	2: Travel Class and Connectivity	1: Comfort and Entertainment	1: Comfort and Entertainment	0: Overall service quality
Fourth priority	0: Overall service quality	0: Overall Service quality	0: Comfort and Entertainment	1: Overall Service quality	1: Comfort and Entertainment	3: Check-in and boarding services	0: Overall service quality	0: Overall service quality	2: Comfort and Entertainment
Test accuray	0.912201	0.906401	0.913151	0.913151	0.911867	0.915383	0.916179	0.916179	0.915691

- After analyzing the table above, it is recommended that the airline service team concentrates on factor 2 (Travel Class and Connectivity). It should be given the highest priority when determining which factor to focus on for improvement of customers' satisfaction
- This is due to the fact that factor 2 has the highest score and highest majority vote in the first priority

- Consequently, the factor that has the 2nd priority for improvement is the Check-in and boarding services and so on.

Conclusion

We have managed to perform factor analysis on our dataset and have extracted 4 factors. These 4 factors are able to group new features beyond this dataset and give the airline an idea of how much importance they should place on it. Besides that, airlines also have the ranking of the factors, with the highest ranked factor being Travel Class and Connectivity. The airline should place high priority on this factor since it seems to heavily contribute to the satisfaction of the customer.

We have also removed some variables and proved that though the tests say that these variables are significant in relation to the satisfaction, removing them can still produce satisfactory results.

For future works, we aim to perform CFA on this dataset. This way, we can confirm that the analysis that we have performed is optimal and validated.

Reference

Bhandari, A. (2020, April 3). *Feature Engineering: Scaling, Normalization and Standardization*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Brownlee, J. (2020a, May 20). *How to Use Quantile Transforms for Machine Learning - MachineLearningMastery.com*. Machine Learning Mastery. https://machinelearningmastery.com/quantile-transforms-for-machine-learning/

Brownlee, J. (2020b, May 25). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. MachineLearningMastery. https://machinelearningmastery.com/rfe-feature-selection-in-python/

Brownlee, J. (2020c, June 5). *How to Perform Feature Selection With Numerical Input Data - MachineLearningMastery.com*. https://machinelearningmastery.com/feature-selection-with-numerical-input-data/

Mbaabu, O. (2020, December 11). *Introduction to Random Forest in Machine Learning | Engineering Education (EngEd) Program | Section*. Section. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Saini, A. (2021a, September 20). *Gradient Boosting Algorithm: A Complete Guide for Beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

Saini, A. (2021b, October 12). *Support Vector Machine(SVM): A Complete guide for beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/