

**AIRLINE CUSTOMER SATISFACTION**

**INTERIM REPORT**

**Submitted to:**

Concerned Faculty

At

Great Learning

**Submitted by:**

Arun .S

Aishwarya .K

Backiyalakshmi Anantharaj

Mudisetty Jyothi

Tejeswaran .V

PGPDSE-FT Bangalore October 2023

Post Graduate Program in Data Science and Engineering

**CONTENTS**

Page no.

|  |  |
| --- | --- |
| 1.Introduction to the business problem | 3 |
| 2. Exploratory Data Analysis and business insights | 4 – 18 |
| 3. Encoding | 19 |
| 4. Base Model | 19-21 |
| 5. Decision Tree | 21-24 |
| 6. Random Forest Classification | 25-26 |
| 7. AdaBoost Classification | 27 |
| 8. Gradient Boosting Classification | 28 |
| 9. XG Boosting classification | 29 |
| 10. Conclusion | 32-33 |

**Introduction:**

Our aim is to showcase the basics of data science (data cleaning, encoding, feature engineering, and model training), all while attempting to solve a problem that is common among businesses to satisfy the customer’s expectation.

We will be treating this as a binary classification problem, where we will attempt to create a model that predicts whether a customer was **satisfied** or **unsatisfied** with the experience and/or service which an airline provided.

The below points include the necessary things from a buyer perspective:

**The features of the Dataset:**

Gender, Customer Type, Age, Type of Travel, Class, 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, Check in Service, Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes,

Satisfaction

**Problem Statement:**

The main objective of the project is to build and train a machine learning algorithm that classifies airline passenger’s satisfaction by entering the data of a specific passenger. The classifier then determines whether this passenger is satisfied or dissatisfied.

**Objective:**

Take advantage of all of the feature variables available below, use it to Analyze and predict passenger is satisfied or dissatisfied.

This dataset contains an airline passenger satisfaction survey.

**Data Dictionary:**

**1.Gender :** Gender of the passengers(Female, Male)

**2.Customer Type:** The customer type (Loyal customer, disloyal customer)

**3. Age:** The actual age of the passengers

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

**5.Class:** Travel class in the plane of the passengers(Business, Eco, Eco plus)

**6.Flight distance:** The flight distance of this journey

**7.Inflight Wi-Fi service :** Satisfaction level of the inflight Wi-Fi service(0:Not,Applicable:1-5)

**8.Departure/Arrival time convenient:** Satisfaction level of Departure/Arrival timeconvenient

**9.Ease of Online booking:** Satisfaction level of online booking

**10. Gate location:** Satisfaction level of Gate location

**11.Food and drink:** Satisfaction level of Food and drink

**12.Online boarding:** Satisfaction level of online boarding

**13.Seat comfort:** Satisfaction level of Sear comfort

**14.Inflight entertainment:** Satisfaction level of inflight entertainment

**15.On-board service:** Satisfaction level of On-board service

**16.Leg room service:** Satisfaction level of Leg room service

**17.Baggage handling:** Satisfaction level of baggage handling

**18.Check-in service :** Satisfaction level of Check-in service

**19.Inflight service:** Satisfaction level of Inflight service

**20.Cleanliness:** Satisfaction level of Cleanliness

**21.Departure Delay in Minutes:** Minutes delayed when departure

**22.Arrival Delay in Minutes:** Minutes delayed when Arrival

**23. Satisfaction:** Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

# Exploratory Data Analysis:

**Solution:**

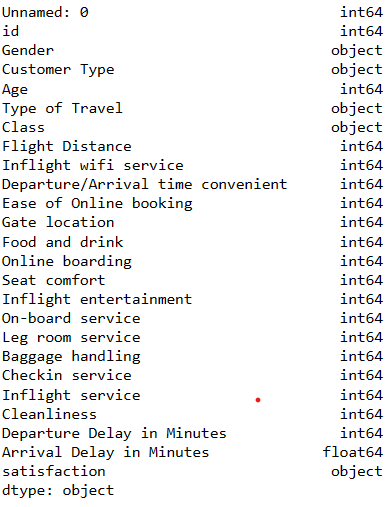


Firstly, after importing all the relevant libraries on Jupyter notebook, we load the data set. Then, we perform EDA to extract and see patterns in the given data set.

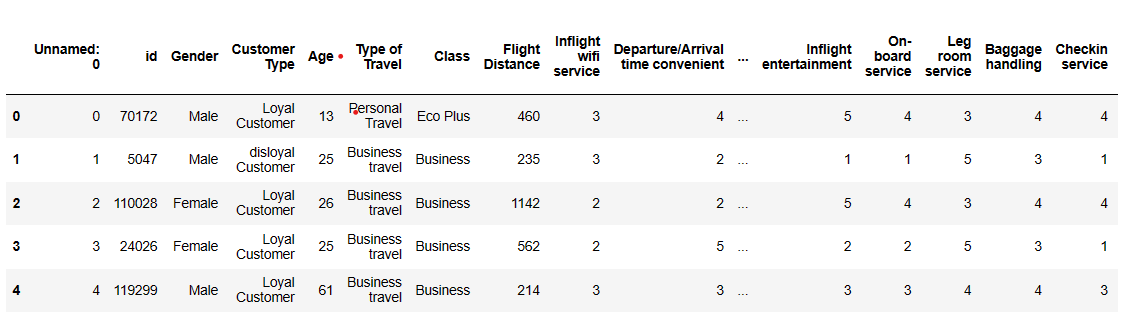
**Shape of the data:**

The shape of the data is (103904,25) that is the data consists of 103904 rows and 25 columns.

**Data types of the Data:**

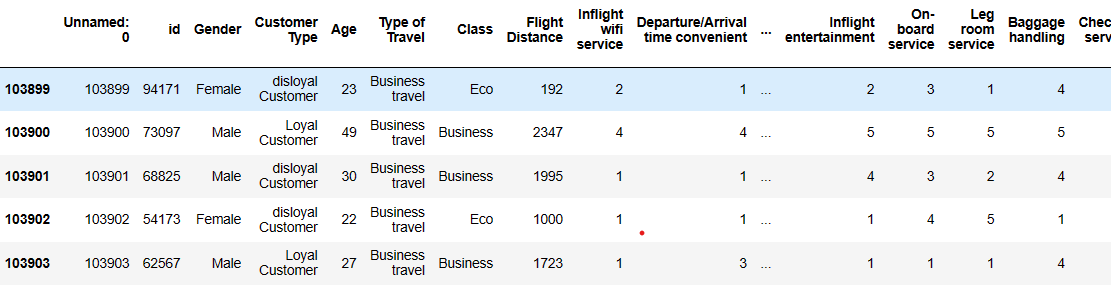
****

**First 5 records in the data:**

****



**Bottom 5 records in the data:**

****

# Information of the data:

# 



**Null values:**



The Null values are present in the dataset:

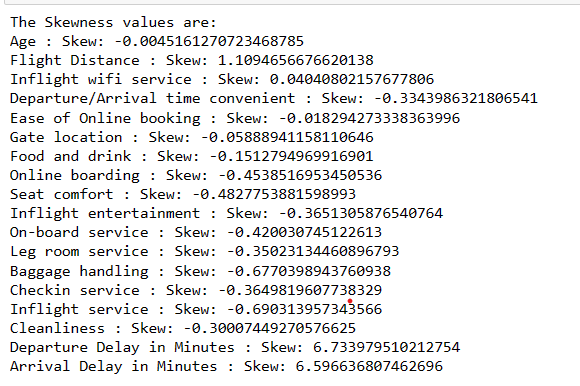
# 

# Duplicates:

There are no duplicates present in the dataset

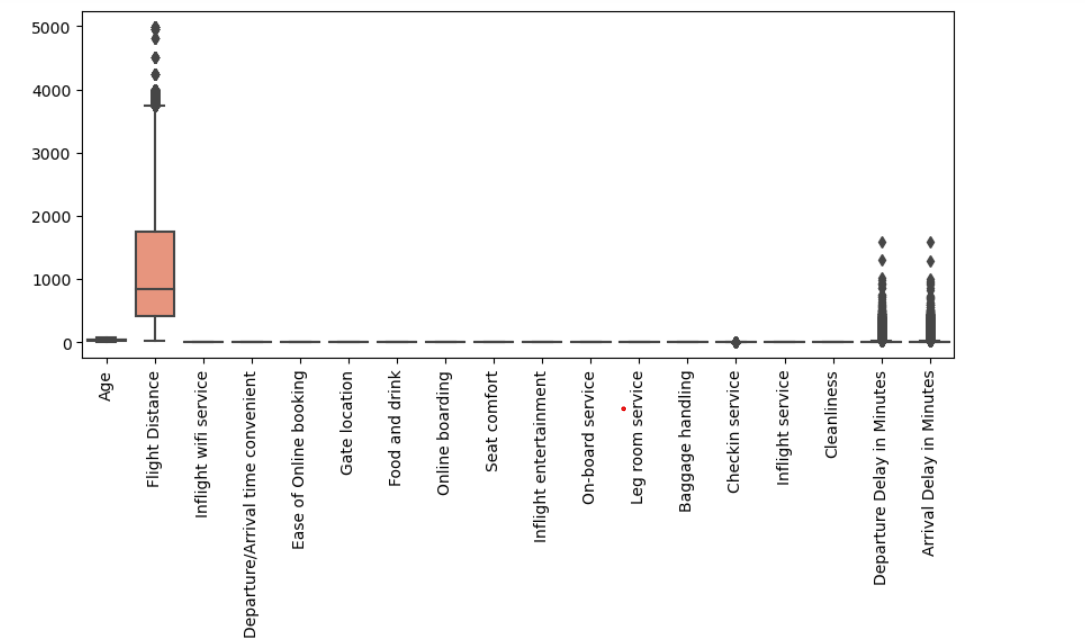
C:\Users\Admin\Pictures\Saved Pictures\duplicate.png

**Skewness of the dataset:**

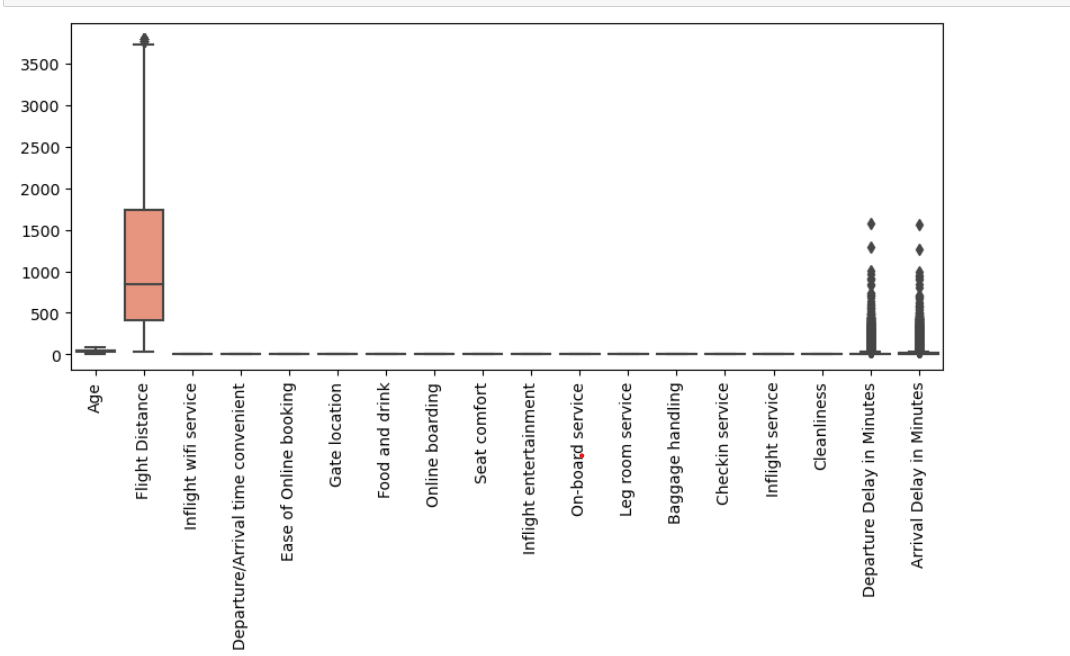
****

**Boxplot:**

**Outliers present in the dataset:**

****

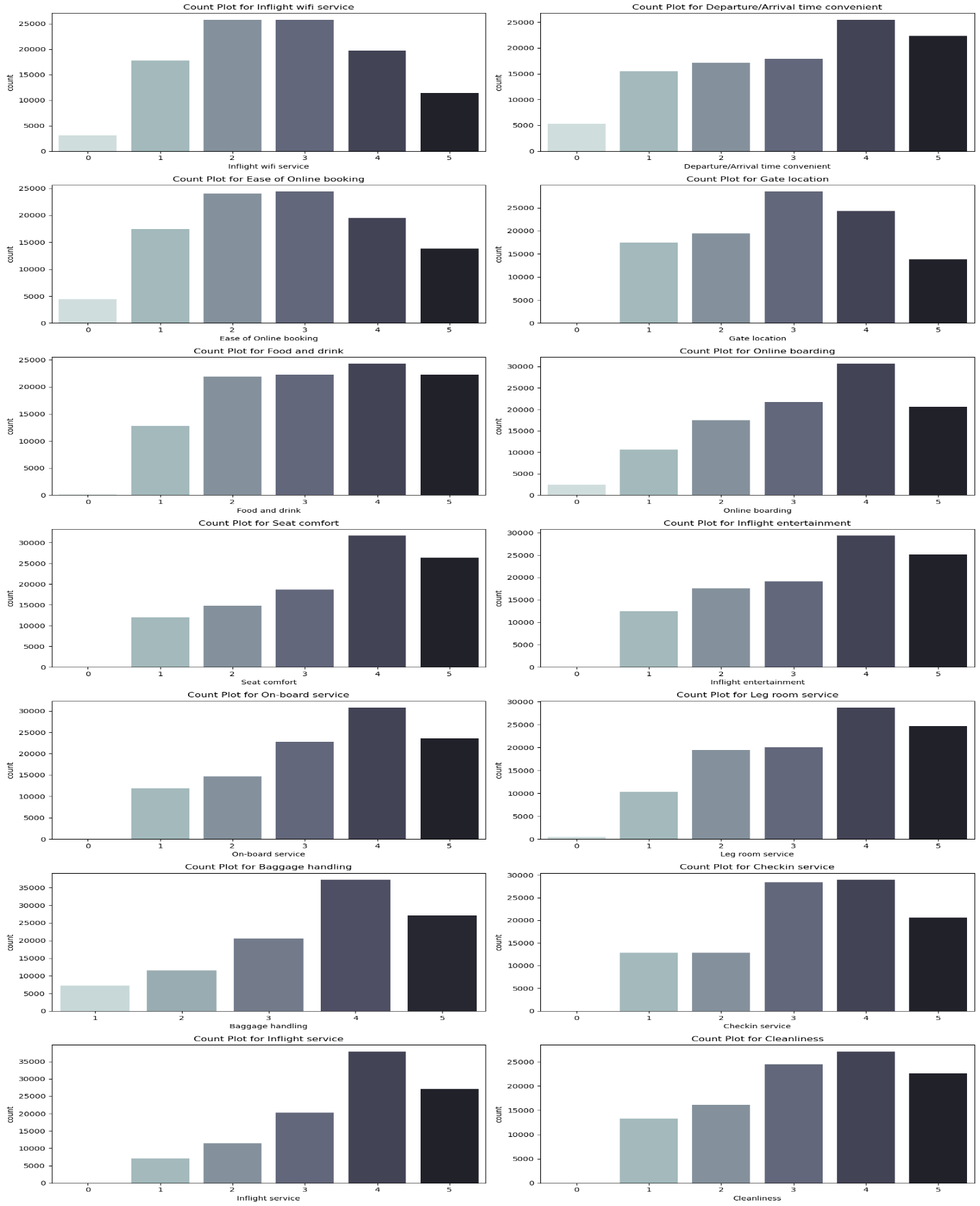
# After Outlier treatment:

****

**Uni-Variate Analysis:**

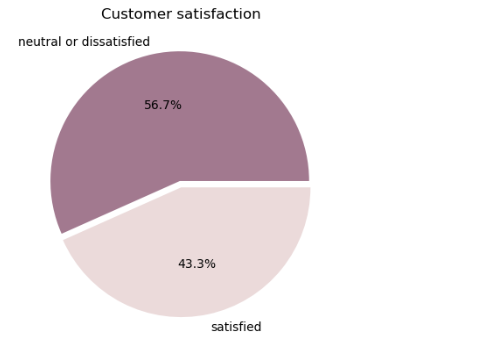
To perform univariate analysis, we have grouped all the services provided (which is numeric) and categorical columns separately.

1. Services Provided:



* The above count plots represents the ratings ( 0 to 5 ) for each service provided by the airline.
* From the analysis performed on the services provided by the airline, we observe that most passengers are satisfied with all the services, often giving a rating of 4 out of 5.

1. **Pie chart to represent the percentage distribution of customer satisfaction** ( Satisfied or neutral or dissatisfied ):



* 56.7% of passengers are neutral of dissatisfied whereas 43.3% of the passengers are satisfied.

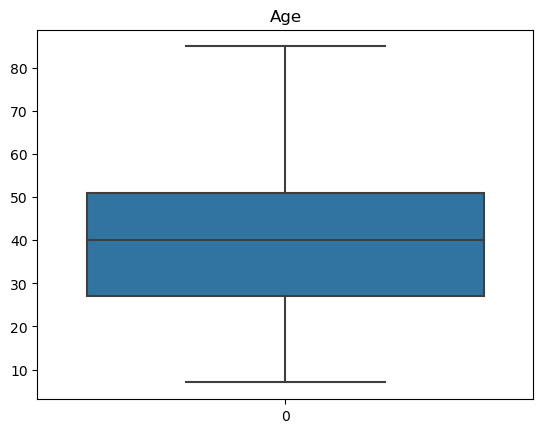
1. **Categorical features:**

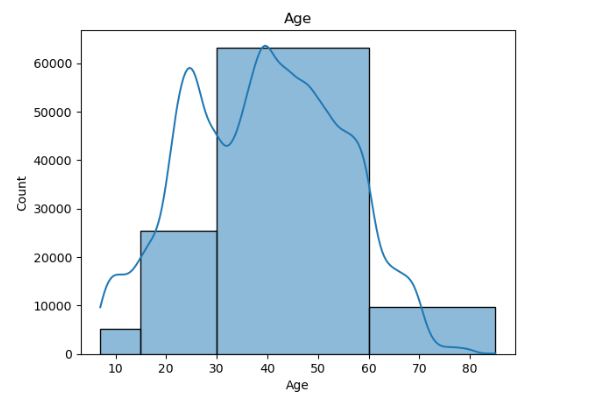
Plotting a pie chart to analyse the distribution of all the categorical features.



* Gender : The proportion of men and women traveling is almost the same
* Customer Type : 81.7 % of customers are loyal customers whereas 18.3% of the customers are disloyal.
* Class : 47.8% , 45% , 7.2% of the people travel in Business class, Eco class and Eco Plus class respectively.
* Type of Travel : Most of people are likely to travel for business purpose.

1. **Distribution of Age:**

****



* From the box plot and the histogram, we observe that most of the passengers age ranges between 30 and 50.

# Bivariate Analysis:

# Distribution of each service based on gender:

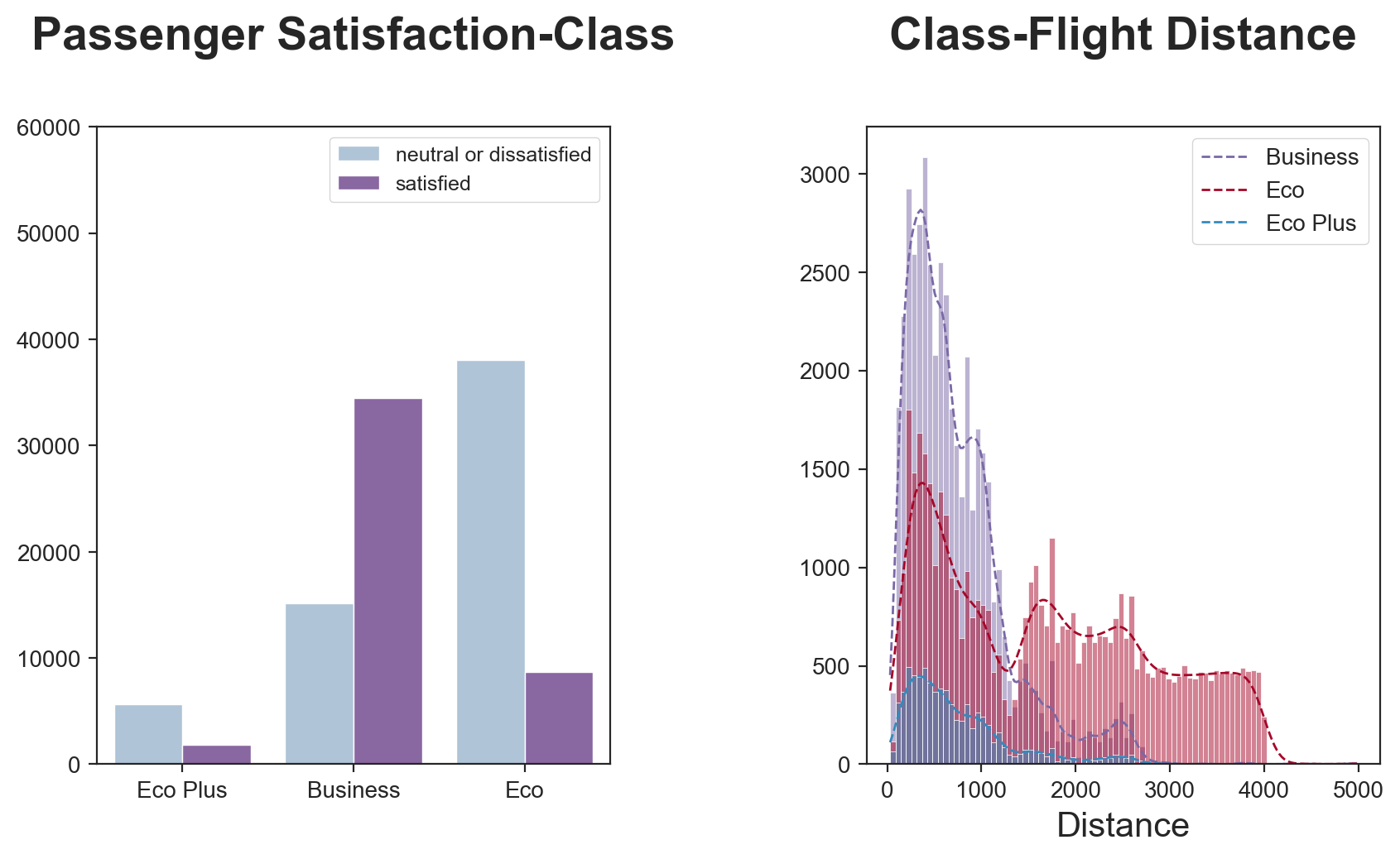
# 

# Considering all the service satisfaction females are more satisfied than males with majority of the ratings ranging between 3 and 5.

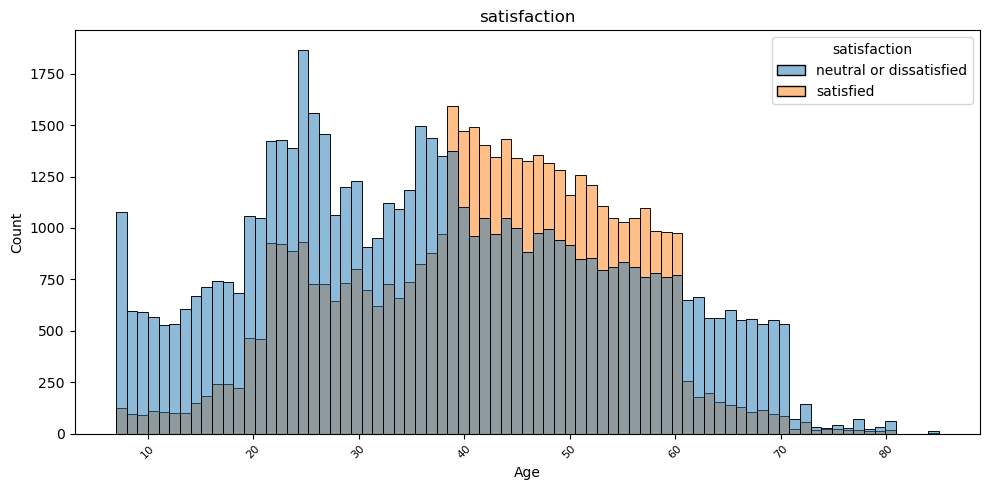
# Distribution of each service based on satisfaction:

# 

* The majority of the reviews are positive or neutral (3-5).
* When it comes to the overall satisfaction and experience of the passengers, 60% of them are neutral or dissatisfied and 40% are satisfied.
* There aren't any categories in which the airline has received overall negative feedback, with most of the reviews being in the range of 3 to 5.
* With that in mind the airline should focus on improving the services that have the biggest impact on the experience of the passenger and therefore the overall satisfaction and some of those categories are Seat Comfort, Leg Room Service, Cleanliness, Online Boarding and Baggage Handling.

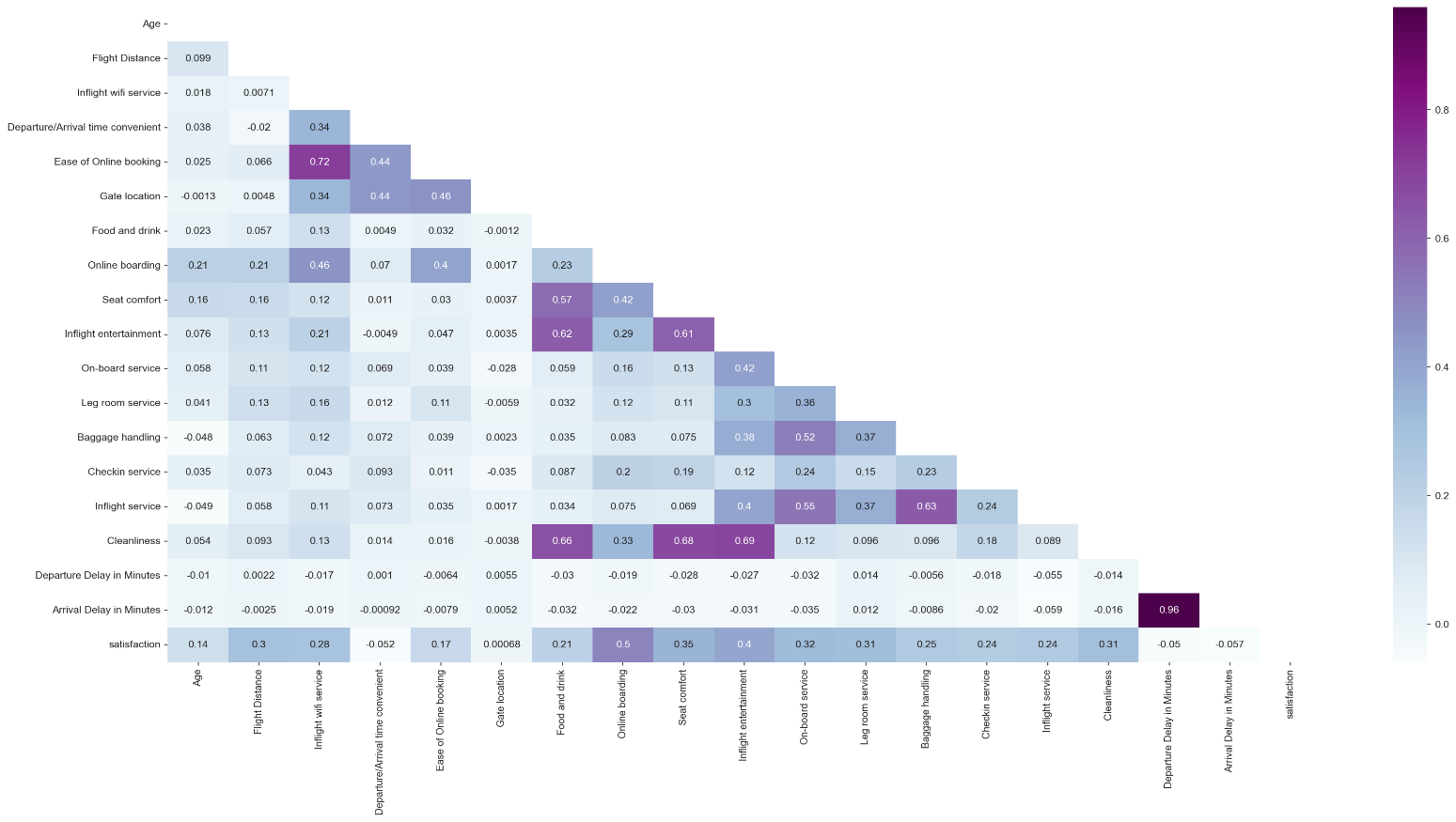


* Very few people fly in the economy plus class. They usually prefer Economy or Business.
* More than 80% of passengers flying in economy are either Neutral or Dissatisfied.
* That shows us that it needs some improvement.
* Passengers that fly for long distances almost always prefer business class. Probably because it is more comfortable.



* The people who gave 'neutral or dissatisfied' having the count more than 1750 are more than those who are satisfied
* People between the age of 40 to 60 have rated it as Satisfied

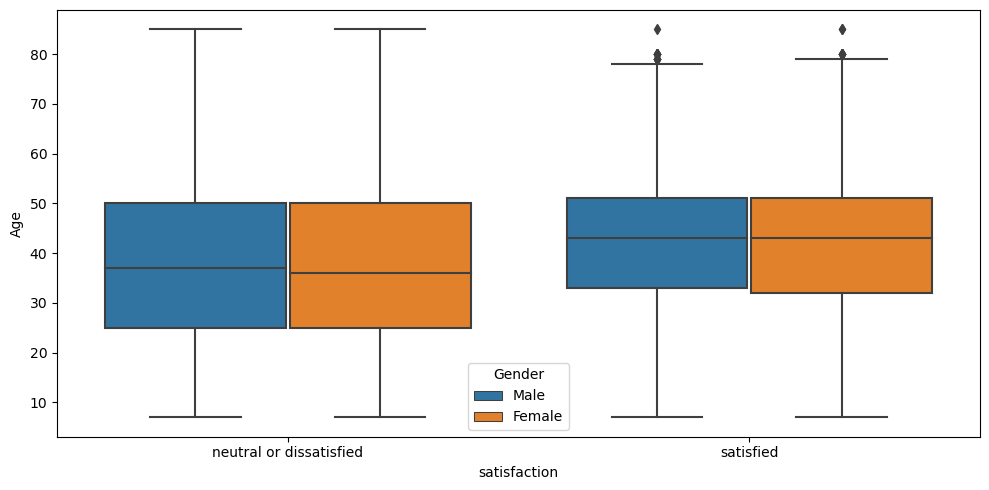
**Heat Map:**



* We can see that there is a match between arrival delay and departure delay which is expected.
* Food and Drink is associated with In-Flight Entertainment.
* There is a correlation between Seat Comfort and clean less.
* Passenger satisfaction is dependent on all of the features of the dataset (Online Boarding, Seat Comfort etc.) almost equally.

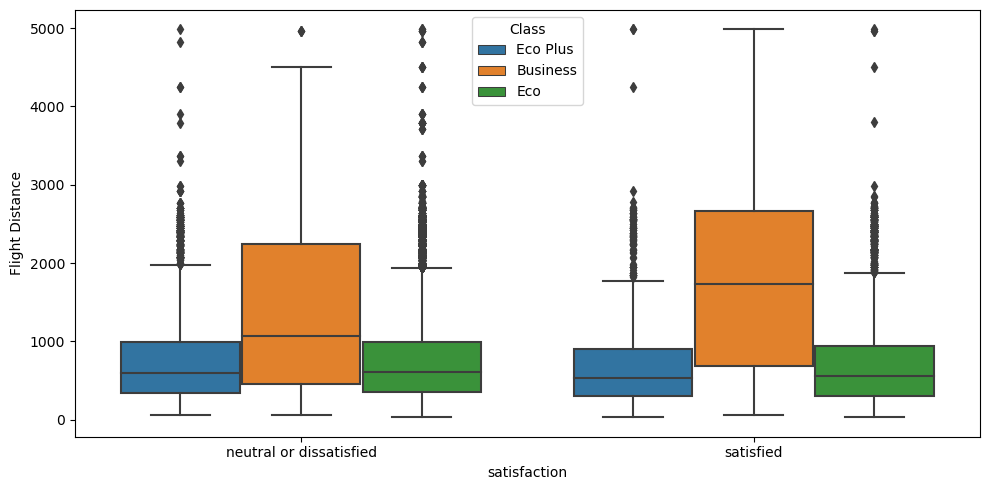
**Multivariate Analysis:**

1. **Distribution of satisfaction ( target feature ) based on Age and Gender:**

****

* From the box plot we observe that, most number of passengers who are neutral or dissatisfied, their age lies between 25 and 50. Whereas the most number of passengers who are satisfied, their age lies between 35 and 50. In both the cases, the male to female ratio is almost similar.

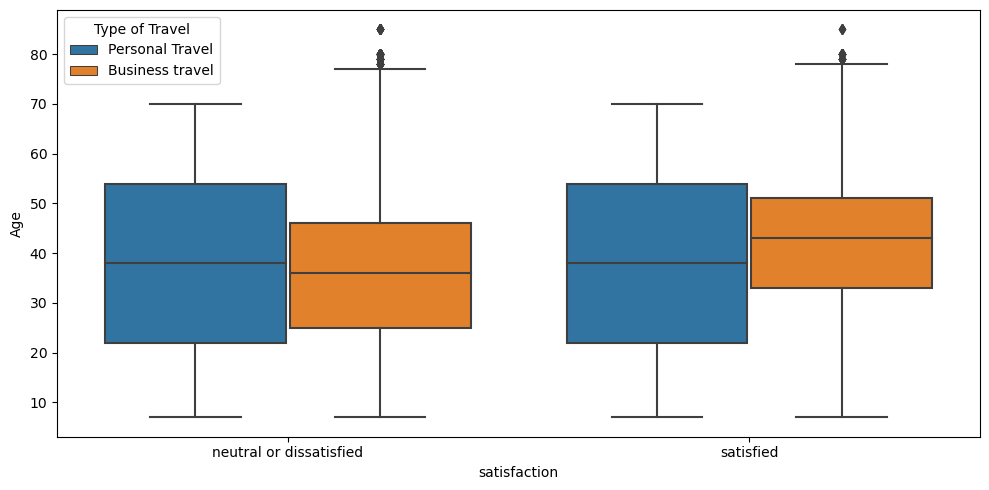
1. **Distribution of satisfaction ( target feature ) based on Flight distance and Class:**



* From the box plot we observe that, most number of passengers who are neutral or dissatisfied, for business class the travel distance is higher when compared to Eco plus or Eco class. Whereas the most number of passengers who are satisfied, for business class the travel distance is higher when compared to Eco plus or Eco class.

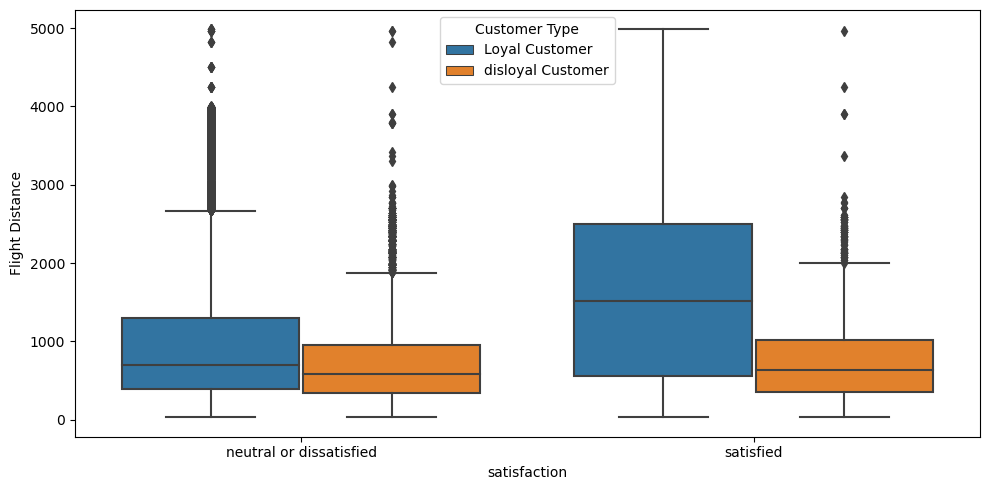
Therefore passengers who are satisfied and travel in business class, their travel distance is the highest.

1. **Distribution of satisfaction ( target feature ) based on Age and Type of Travel:**



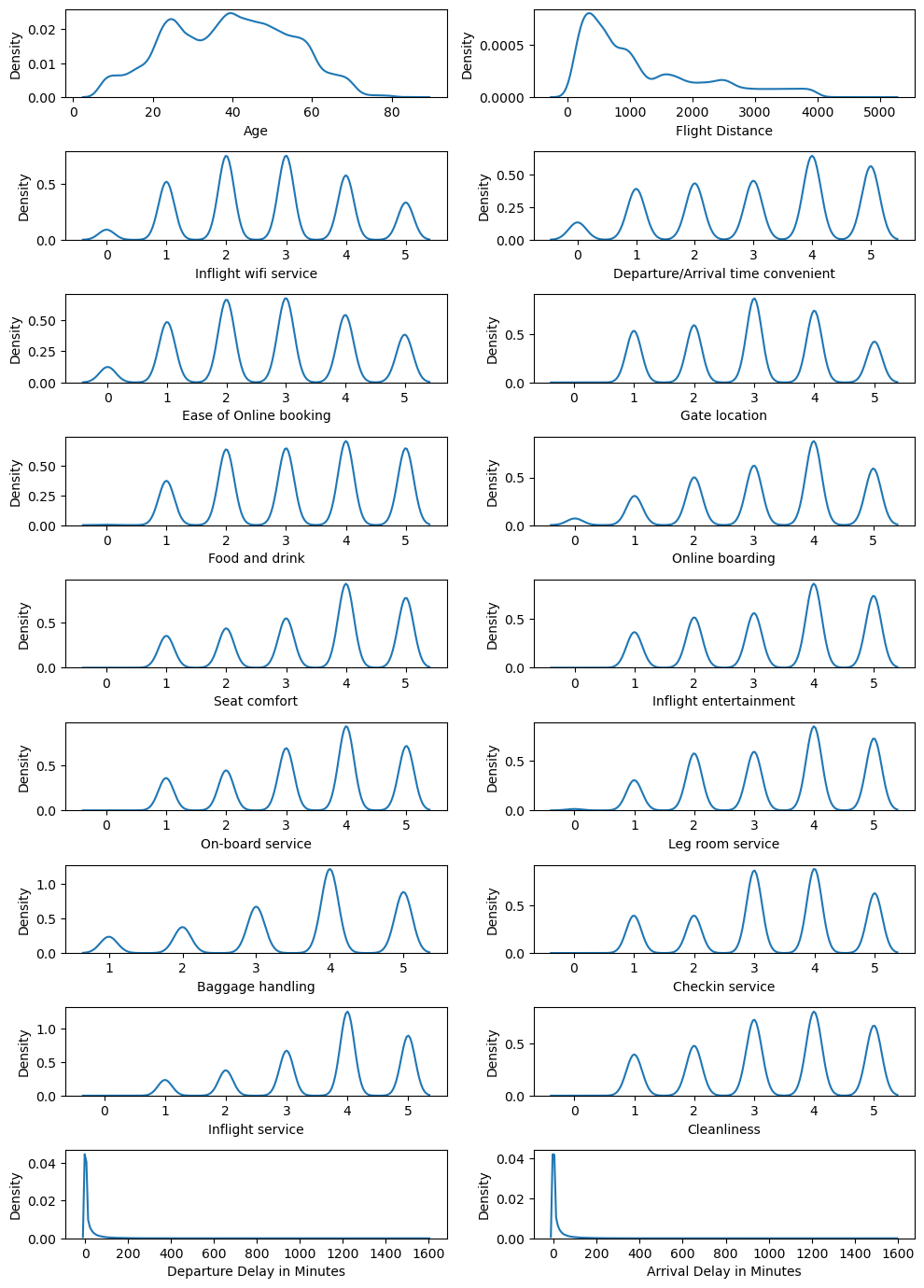
* From the box plot, we observe that, most number of passengers who are neutral or dissatisfied, passengers between age of 20 and 55 are on personal travel and passengers between age of 25 and 45 are on business travel.
* For the passengers who are satisfied, passengers between age of 20 and 55 are on personal travel and passengers between age of 35 and 50 are on business travel .

1. **Distribution of satisfaction ( target feature ) based on Flight distance and Customer Type:**

****

* From the box plot, we can observe that, for neutral or dissatisfied customers, loyal customers travel more than disloyal customers.
* For the passengers who are satisfied, the loyal customers travel distance is the highest when compared to disloyal customers.

**Distribution of Numerical features to identify skewness:**

****

* From the above KDE plot, we observe that all of the numerical features have skewness.
* The feature “Flight Distance”, “Departure Delay” and “Arrival Delay” are Leptokurtic, whereas all the other numeric features are Platykurtic.
* To treat the skewness present in these features, we use a transformation method. Since most of the features are negatively skewed, log or Yeo-Johnson ( for both positive and negative skewness ) transformation techniques are more viable.

# Encoding:

# To encode the categorical columns that are present in the dataset, we use One-Hot encoding. Since the categorical columns present have only 2 or 3 sub-categories / sub classes and has no inherent order or hierarchy associated to it, One-Hot encoding is best suited for these features.

# 

# Scaling and Transformation:

# By checking for the skewness of the data, we observed that majority of the features were left skewed followed by a few features that are positively/right skewed. To handle this, Yeo-Johnson or logarithmic transformation are best suited.

# Using Standard Scaling, we bring all the encoded and the numerical columns to a common scale such that the features have a standard deviation of 1 and mean 0.

# Splitting the data:

# Splitting the dataset into train and test.

# Extracting the target column into separate vectors for training set and test set. For training and testing purpose we are splitting the dataset into train and test data in the ratio 70:30.

# We have bifurcated the dataset into train and test. We have also taken out the target column out of train and test data into separate vector for evaluation purposes.

# Checking the relationship between feature and target:

# 

# From the plot every independent variables (except Gate Location) depend upon the target.

# Evaluation metrics:

# Accuracy and ROC-AUC are common metrics used to evaluate the performance of classification models. They each provide a different perspective on the model's effectiveness, and understanding them can guide us in building better models.

# Accuracy: Accuracy is the ratio of correctly predicted instances to the total number of instances. It's a simple metric that gives you an idea of how often your model is correct. It is a common choice for measuring model performance, especially when dealing with balanced datasets where the classes are roughly equal in frequency.

# ROC-AUC: ROC-AUC stands for "Receiver Operating Characteristic - Area under the Curve." It measures the area under the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate at various threshold settings.

# ROC-AUC is ideal for binary classification problems, especially when dealing with imbalanced datasets. It provides a nuanced view of the model's performance by examining how the true positive rate changes as the threshold for classifying a positive instance varies.

# ROC-AUC is threshold-independent, meaning it evaluates a model's performance over a range of thresholds. This characteristic makes it valuable for understanding a model's robustness and its ability to discriminate between classes.

# Model 1:(Base Model)

# Logistic Regression:

# For Binary Classification, Logistic Regression is the base model.

# Logistic regression offers simplicity, interpretability, and efficiency, making it a suitable choice as a base model for binary classification tasks. Its straightforward nature and ability to handle large datasets make it an essential component in the machine learning toolkit.

# 

# From the base logistic model we obtain an accuracy of 0.88 on both the test dataset and training dataset, meaning that the logistic model is able to classify our target correctly with an accuracy of 88%.

# 

# Confusion Matrix ROC-AUC Curve

# Model 2:

# Decision Tree:

# Building a decision tree model to check how well the model performs on our train and test datasets.

# 

# 

# Confusion Matrix ROC-AUC Curve

# From the decision tree model, we observe that the model obtained is over fitted.

# To rectify this over fitting we hyper tune the model to obtain the best hyperparameters.

# This is done by using Grid Search CV method.

# Model performance after fine tuning:

# 

# 

# Confusion Matrix ROC-AUC Curve

# After tuning the hyper parameters, there is slight difference in precision values, but we can see a clear difference in the ROC-AUC Curve plot.

# Feature Selection:

# Using sequential feature selector to obtain the features that most significant.

# The features that are most significant are:

# Inflight Wi-Fi service

# Gate location

# Online boarding

# Baggage handling

# Inflight service

# Customer Type Loyal Customer

# Type of Travel Business travel

# Class Business

# We now use these significant features in our decision tree model.

# Model Performance after feature selection:

# 

# 

# Confusion Matrix ROC-AUC Curve

# The final decision tree model is able to classify 94% of the data correctly.

# Model 3:

# Random Forest:

# Using a bagging technique, random forest classification to check how well the model is able to classify our data.

# 

# 

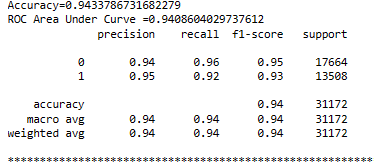
# Confusion Matrix ROC-AUC Curve

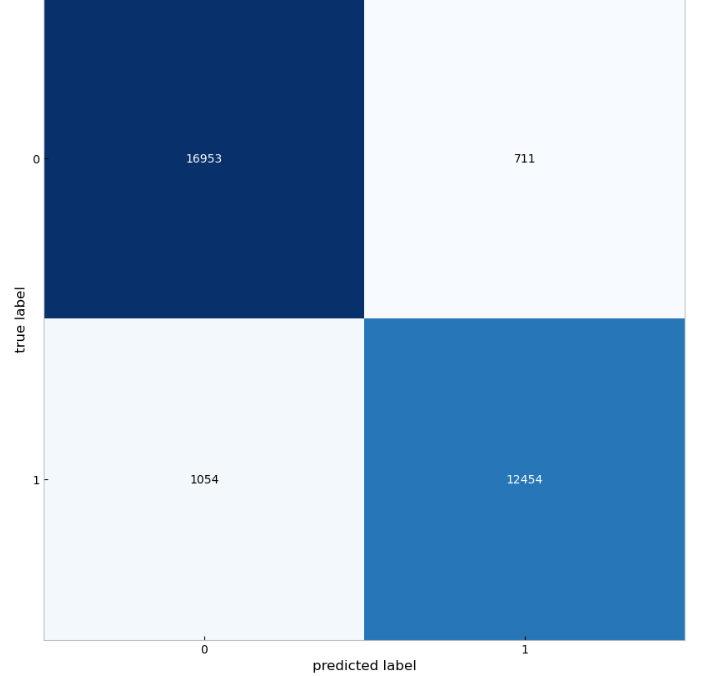
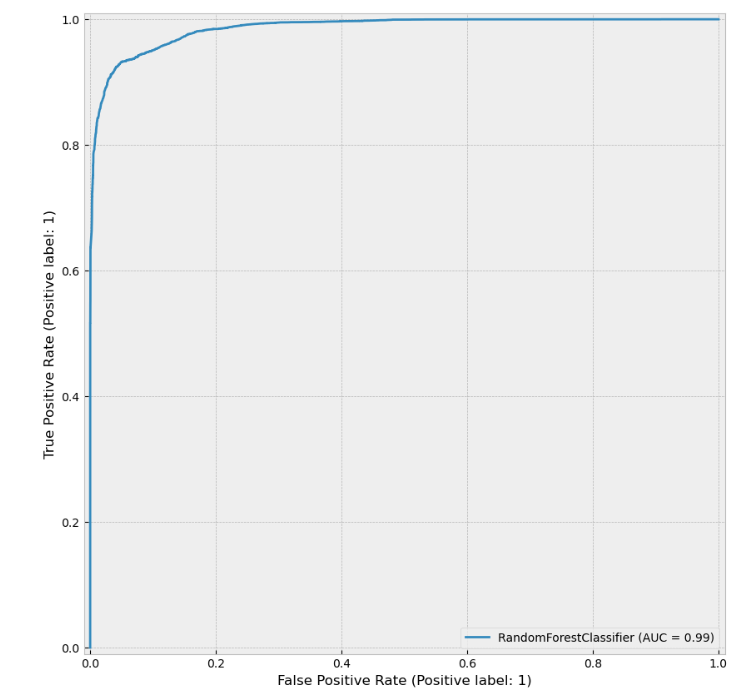
# From the Random Forest model, we observe that the model obtained is over fitted.

# To rectify this over fitting, we hyper tune the model to obtain the best hyperparameters.

# This is done by using Grid Search CV method.

**Model Performance after fine tuning:**



# Confusion Matrix ROC-AUC Curve

# After tuning the hyper parameters, there is slight difference in precision values, but we can see a clear difference in the ROC-AUC Curve plot. Therefore, the final model is able to classify 94% of the data correctly.

**Model 4:**

**Ada Boost Classifier:**

# Using a boosting technique, Ada Boost classification to check how well the model is able to classify our data.

# 

# Model performance:

# 

# Confusion Matrix ROC-AUC Curve

# From the training and test scores, we can conclude that there is no overfitting present.

# Thus the model is able to classify 93% of the data correctly.

# Model 5:

# Gradient Boosting:

# Using a boosting technique, gradient boost classification to check how well the model is able to classify our data.

# 

# Model Performance:

# 

# Confusion Matrix ROC-AUC Curve

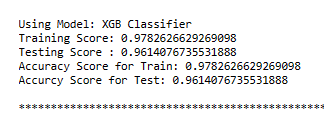
# From the training and test scores, we can conclude that there is no overfitting present.

# Thus the model is able to classify 93% of the data correctly.

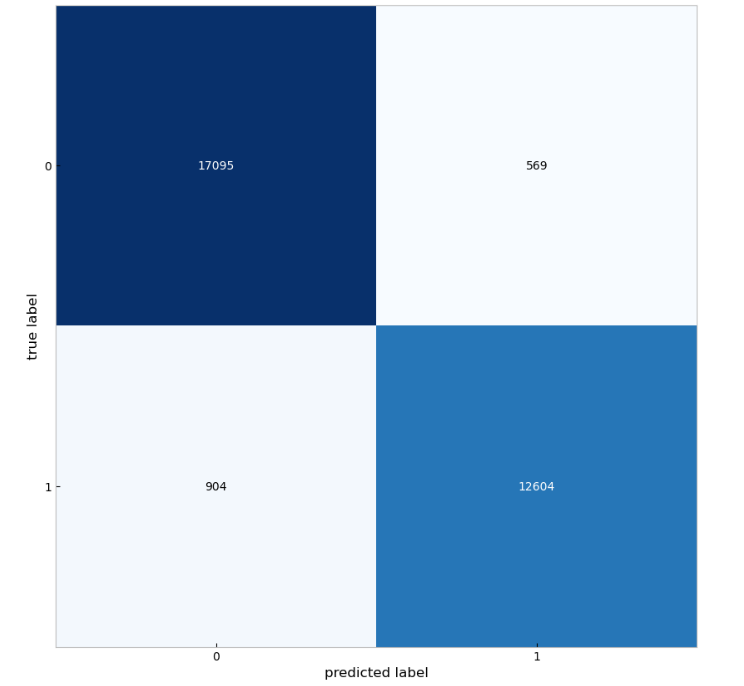
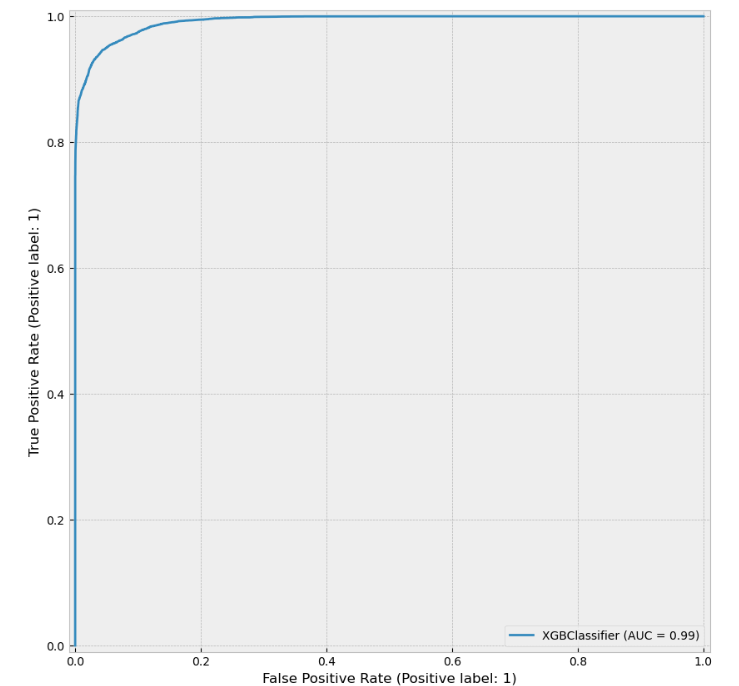
**Model 6:**

**XG Boosting:**

# Using a boosting technique, extreme gradient boost classification to check how well the model is able to classify our data.



**Model Performance:**

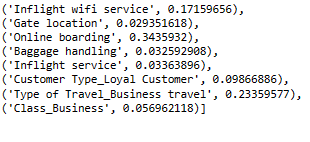
# Confusion Matrix ROC-AUC Curve

# From the training and test scores, we can conclude that there is no overfitting present.

# Thus the model is able to classify 96% of the data correctly.

**Important Features:**

By performing sequential feature selection on our Xg Boost model, the significant features obtained were:

****

From the XG Boost Model performance using Feature selection method, we can able to get three important features,

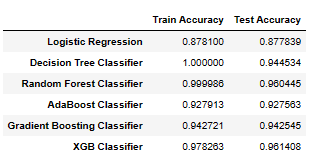
* 1. Inflight Wi-Fi Service
  2. Online Boarding
  3. Types of Business Travel

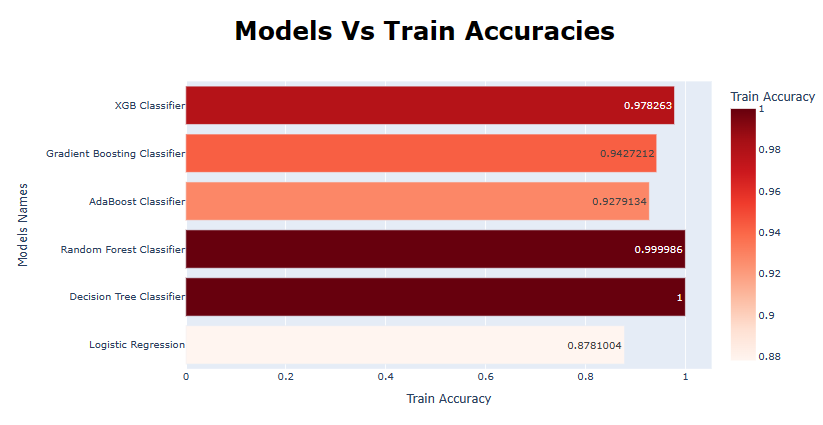
And the features that need to be improve their performance to satisfy the customer are,

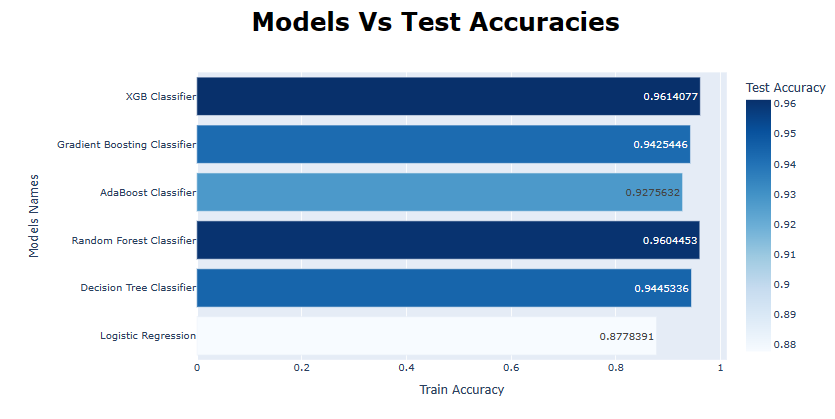
1. Gate Location
2. Baggage Handling
3. Inflight service
4. Customer Type Loyal Customer
5. Class Business

To conclude which classification algorithm would best suit our problem statement, we analyze different metrics or scores of the classification models:

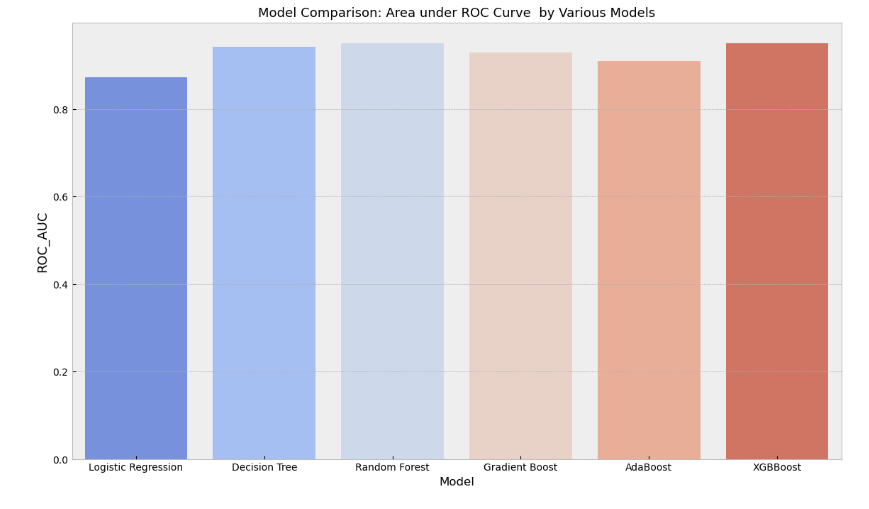
**Model Performance by Accuracy score:**

****



****

**Model Performance by ROC-AUC Score:**

****

**Conclusion:**

**Business Interpretation:**

* Passengers also appear to be sensitive to certain aspects of their services and as such the airline should tackle such issues as a priority to improve services.
* This includes inflight Wi-Fi and departure time. That said, if the airline wishes to continue to focus on business travellers aspects such as online booking and seat comfort should be prioritised.
* There are also aspects such as ticket price, flight location also missing from this dataset.

**Recommendation:**

About business, my advice to the company to improve the services those have low

Score for the Eco class:

1. Food and drink
2. Online boarding
3. Seat comfort
4. Inflight entertainment
5. On board service
6. Leg room service

* Really Important Features: Type of Travel, Inflight Wi-Fi service, Online boarding, Seat comfort
* I did some hyper parameter tuning on some of the models, but not all.
* Thus the Company should look into improving services like “food and drinks”, ”Seat comfort”, “Inflight entertainment”, ”Baggage handling” and services provided for Eco and Eco plus class passengers.

**Overall Performance:**

* Since Decision Tree is a white-box (explainable) model, we can deep-dive into its visualization to get more valuable insight.
* We see that Random Forest and XG Boost have performed very well on both Accuracy and area under ROC curve. So, we are now interested to see how many decision trees are minimally required make the Accuracy consistent (recalling the fact that Random Forest is actually a bagged ensemble of decision trees).
* We observe, Random Forest and XG Boost have performed equally well on producing high ROC\_AUC score (96%). But Random Forest has taken lesser amount of time compared to time taken by XG Boost. So, we will stick to Random Forest as the best model.
* Hence, to satisfy the customer the airline company should give more important to the features that are giving less significant with the target (Customer’s satisfaction).
* By improving this features performance will improve their ratings.

**Future Work:**

We have some questions that we thought it will help us to get to know our data more like, was the departure delay caused by bad weather? Or like was customer’s seatmate is rude?

By Analysing these deep knowledge of data we can able to improve the performance of model by selecting the best features.

**---x---**