**ALY6040: Data Mining**

**Final Project: Final Report.**

**By: Rohit Raman**

**Date – 19-05-2023.**

# Introduction.

# Research Questions.

1. **How accurately can logistic regression predict customer satisfaction and how can these insights from predictive modelling be leveraged to improve customer satisfaction and retention in a business context?**

**The research question is focused on predicting customer satisfaction (a binary variable) using logistic regression. (What we will do:** Computing **logistic regression** for the prediction of customer satisfaction (Binary variable) by splitting the dataset (training, testing) and analyses the result using ROC curve, accuracy, specificity etc)

The problem we hope to solve by predicting customer satisfaction using **logistic regression** is to identify the factors that influence customer **satisfaction levels** and predict how likely a customer is to be satisfied or dissatisfied with the **airline's services**. By identifying these factors, airlines can take necessary measures to improve the areas that are impacting customer satisfaction levels negatively and enhance the areas that are positively impacting satisfaction. This will ultimately lead to higher customer retention, better customer experiences, and increased profitability for the airline.

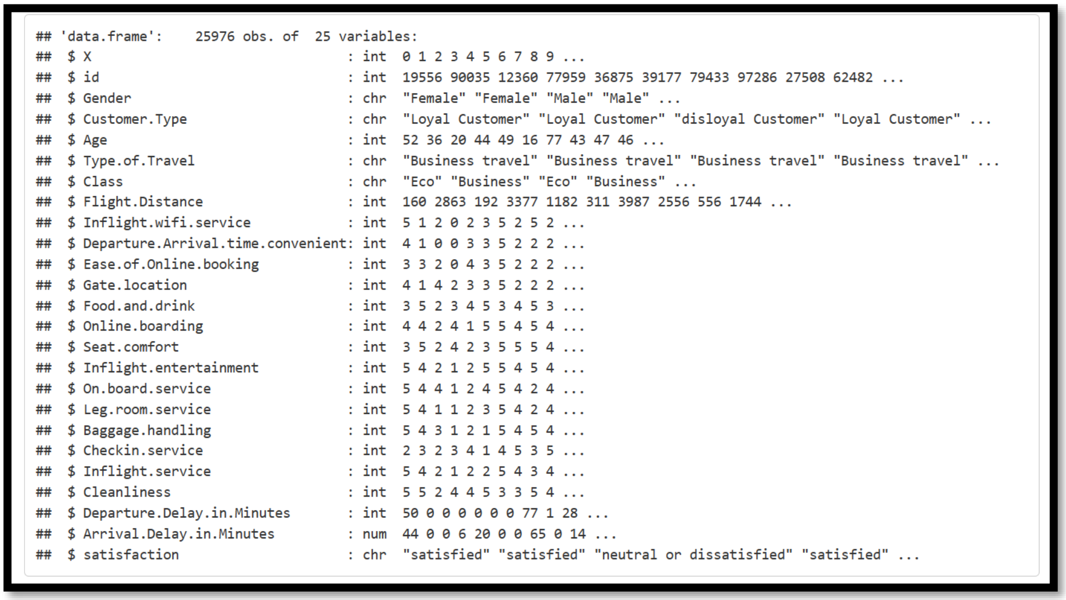
2. **What are the most important factors that impact customer satisfaction for an airline, as identified through a random forest model, and how can this information be leveraged to improve customer satisfaction and loyalty in the airline industry?**

**The next research question is focused on identifying the factors that have the most impact** on **customer satisfaction** for an airline using a **random forest model**. (What we will do): The aim is to display the variable of **importance and provide insights** into how **the airline can improve its services based on these factors**. The problem being addressed is to identify the key drivers of customer satisfaction, which can help the airline prioritize its resources and efforts towards improving these areas and enhance customer experience.

### 3. How can we understand and segment our airline customers based on their satisfaction with various facilities (clustering)?

I will perform clustering analysis on airline customer data to understand and segment customers based on their satisfaction with various facilities. It uses k-means clustering with an optimal number of clusters determined through the elbow method. The resulting clusters are visualized using PCA, providing insights into distinct customer segments and their satisfaction patterns. This information can help airlines tailor their services and strategies to meet the specific needs and preferences of different customer groups.

## Basic overview of the dataset.



```
## 'data.frame':    25976 obs. of  25 variables:
##  $ X                                : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ id                               : int  19556 90035 12360 77959 36875 39177 79433 97286 27508 62482 ...
##  $ Gender                           : chr  "Female" "Female" "Male" "Male" ...
##  $ Customer.Type                    : chr  "Loyal Customer" "Loyal Customer" "disloyal Customer" "Loyal Customer" ...
##  $ Age                              : int  52 36 20 44 49 16 77 43 47 46 ...
##  $ Type.of.Travel                   : chr  "Business travel" "Business travel" "Business travel" "Business travel" ...
##  $ Class                            : chr  "Eco" "Business" "Eco" "Business" ...
##  $ Flight.Distance                  : int  160 2863 192 3377 1182 311 3987 2556 556 1744 ...
##  $ Inflight.wifi.service            : int  5 1 2 0 2 3 5 2 5 2 ...
##  $ Departure.Arrival.time.convenient: int  4 1 0 0 3 3 5 2 2 2 ...
##  $ Ease.of.Online.booking           : int  3 3 2 0 4 3 5 2 2 2 ...
##  $ Gate.location                    : int  4 1 4 2 3 3 5 2 2 2 ...
##  $ Food.and.drink                   : int  3 5 2 3 4 5 3 4 5 3 ...
##  $ Online.boarding                  : int  4 4 2 4 1 5 5 4 5 4 ...
##  $ Seat.comfort                     : int  3 5 2 4 2 3 5 5 5 4 ...
##  $ Inflight.entertainment           : int  5 4 2 1 2 5 5 4 5 4 ...
##  $ On.board.service                 : int  5 4 4 1 2 4 5 4 2 4 ...
##  $ Leg.room.service                 : int  5 4 1 1 2 3 5 4 2 4 ...
##  $ Baggage.handling                 : int  5 4 3 1 2 1 5 4 5 4 ...
##  $ Checkin.service                  : int  2 3 2 3 4 1 4 5 3 5 ...
##  $ Inflight.service                 : int  5 4 2 1 2 2 5 4 3 4 ...
##  $ Cleanliness                      : int  5 5 2 4 4 5 3 3 5 4 ...
##  $ Departure.Delay.in.Minutes       : int  50 0 0 0 0 0 0 77 1 28 ...
##  $ Arrival.Delay.in.Minutes         : num  44 0 0 6 20 0 0 65 0 14 ...
##  $ satisfaction                     : chr  "satisfied" "satisfied" "neutral or dissatisfied" "satisfied" ...
```

Figure 1: Basic overview of the dataset using structure function of the dataset

### Observation.

The data has **various numerical** and **categorical data**. It has **25976 observations** with **25 variables** such as **id, Gender, Customer. Type, Age, Type of travel, Class, Flight Distance, WI-FI services** etc.

## Type of variable in the datasets:

**Gender**: a **character variable** representing the gender of the customer.

**Customer.Type**: a **character variable** representing the type of customer (Loyal or disloyal).

**Age:** an **integer variable** representing the age of the customer.

**Type.of.Travel**: a **character variable** representing the type of travel (Business or Personal).

**Class**: a **character variable** representing the class of travel (Eco, Business or Eco Plus).

**Flight.Distance:** an **integer variable** representing the distance of the flight.

**Inflight.wifi.service:** an **integer variable** representing the rating of inflight WiFi service.

**Departure.Arrival.time.convenient:** an **integer variable** representing the rating of departure and arrival time convenience.

**Ease.of.Online.booking:** an **integer variable** representing the rating of ease of online booking.

**Gate.location:** an **integer variable** representing the rating of gate location.

**Online.boarding:** an **integer variable** representing the rating of online boarding.

**Seat.comfort:** an **integer variable** representing the rating of seat comfort.

**Inflight.entertainment:** an **integer variable** representing the rating of inflight entertainment.

**On.board.service:** an **integer variable** representing the rating of on-board service.

Leg.room.service: an **integer variable** representing the rating of legroom service.

**Baggage.handling**: an **integer variable** representing the rating of baggage handling.

**Checkin.service**: an **integer variable** representing the rating of check-in service.

**Inflight.service:** an **integer variable** representing the rating of inflight service.

**Cleanliness**: an **integer variable** representing the rating of cleanliness.

**Departure.Delay.in.Minutes:** an **integer variable** representing the number of minutes of departure delay.

**Arrival.Delay.in.Minutes:** a **numeric variable** representing the number of minutes of arrival delay.

**satisfaction:** a **character variable** representing the satisfaction level of the customer (satisfied or neutral or dissatisfied).

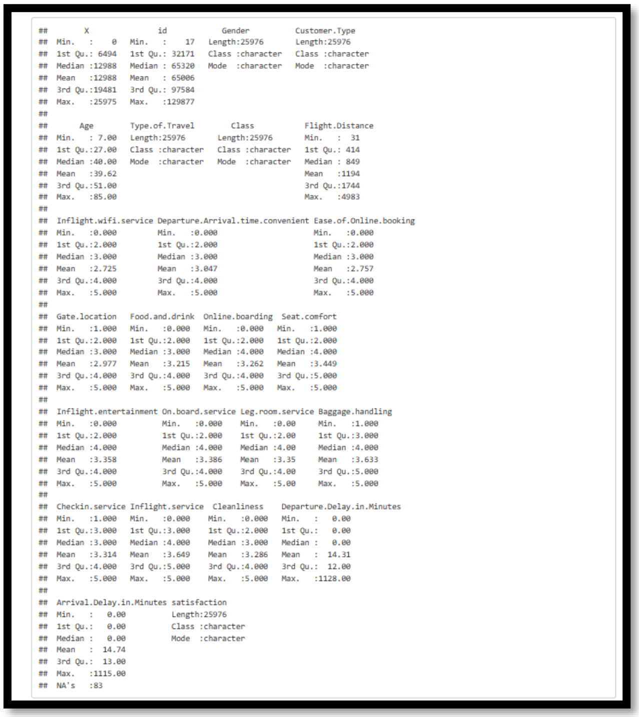
# Write unit of analysis?

**Each row in an airline customer satisfaction dataset conveys** the **feedback or response** from a **single customer regarding their experience with an airline**. This include information **such**

**as their overall rating** of the airline, the **specific aspects** of their flight that they were asked to evaluate (e.g., flight attendants, in-flight entertainment, comfort of seats, etc.), as well as demographic information such as the customer's **age, gender, and location**. The **data in each row** can be used **to analyze theoverall satisfaction levels** of the airline's customers, compare the **performance of different airlines**, and **identify areas for improvement.**

# Methodology: Data preparation and analysis

During the data preparation, we examined the presence of any missing values in the dataset, studied the distribution of variables before handling the missing values, and conducted outlier analysis using box plots. Furthermore, we created a correlation matrix to assess the association between variables.



```
##        X               id              Gender          Customer.Type
##  Min.   :    0   Min.   :    17   Length:25976       Length:25976
##  1st Qu.: 6494   1st Qu.: 32171   Class :character   Class :character
##  Median :12988   Median : 65320   Mode  :character   Mode  :character
##  Mean   :12988   Mean   : 65006
##  3rd Qu.:19481   3rd Qu.: 97584
##  Max.   :25975   Max.   :129877
##
##       Age        Type.of.Travel         Class           Flight.Distance
##  Min.   : 7.00   Length:25976       Length:25976       Min.   :  31
##  1st Qu.:27.00   Class :character   Class :character   1st Qu.: 414
##  Median :40.00   Mode  :character   Mode  :character   Median : 849
##  Mean   :39.62                                         Mean   :1194
##  3rd Qu.:51.00                                         3rd Qu.:1744
##  Max.   :85.00                                         Max.   :4983
##
##  Inflight.wifi.service Departure.Arrival.time.convenient Ease.of.Online.booking
##  Min.   :0.000         Min.   :0.000                     Min.   :0.000
##  1st Qu.:2.000         1st Qu.:2.000                     1st Qu.:2.000
##  Median :3.000         Median :3.000                     Median :3.000
##  Mean   :2.725         Mean   :3.047                     Mean   :2.757
##  3rd Qu.:4.000         3rd Qu.:4.000                     3rd Qu.:4.000
##  Max.   :5.000         Max.   :5.000                     Max.   :5.000
##
##  Gate.location   Food.and.drink  Online.boarding  Seat.comfort
##  Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :4.000   Median :4.000
##  Mean   :2.977   Mean   :3.215   Mean   :3.262   Mean   :3.449
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##
##  Inflight.entertainment On.board.service Leg.room.service Baggage.handling
##  Min.   :0.000          Min.   :0.000    Min.   :0.00     Min.   :1.000
##  1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.00     1st Qu.:3.000
##  Median :4.000          Median :4.000    Median :4.00     Median :4.000
##  Mean   :3.358          Mean   :3.386    Mean   :3.35     Mean   :3.633
##  3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.00     3rd Qu.:5.000
##  Max.   :5.000          Max.   :5.000    Max.   :5.00     Max.   :5.000
##
##  Checkin.service Inflight.service  Cleanliness    Departure.Delay.in.Minutes
##  Min.   :1.000   Min.   :0.000    Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000   1st Qu.:3.000    1st Qu.:2.000   1st Qu.:   0.00
##  Median :3.000   Median :4.000    Median :3.000   Median :   0.00
##  Mean   :3.314   Mean   :3.649    Mean   :3.286   Mean   :  14.31
##  3rd Qu.:4.000   3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000   Max.   :5.000    Max.   :5.000   Max.   :1128.00
##
##  Arrival.Delay.in.Minutes satisfaction
##  Min.   :   0.00          Length:25976
##  1st Qu.:   0.00          Class :character
##  Median :   0.00          Mode  :character
##  Mean   :  14.74
##  3rd Qu.:  13.00
##  Max.   :1115.00
##  NA's   :83
```

Figure 2: Descriptive statistics of the dataset.

### Observations:

**Figure 2 represents** the summary of the **customer satisfaction dataset**. Here, we can see the mean, median, mode of each variable individually, let say if we want to know the average delay in the flight departure, then from the above we figure we can say the average delay in departure is **round 14 min**.

Likewise anyone interested in knowing the 1 quarter value of flight distance then it would be around **414 miles**. The age of customes range from 7 and 85 years. Moreover, Furthermore, while checking the summary statistic, we found the Na values in variable **“Arrival.Delay.in.Minutes”.** There were around 83 Na values in the variable.

## Na values inspections

## Replacing the Na value (Data Cleaning).

**Before removing the Na values, we checked the distribution of variable “Arrival Delay in minutes”**

Distribution of Arrival delay in minutes.

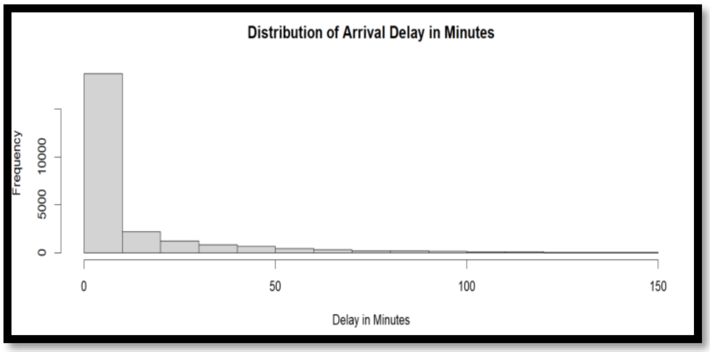
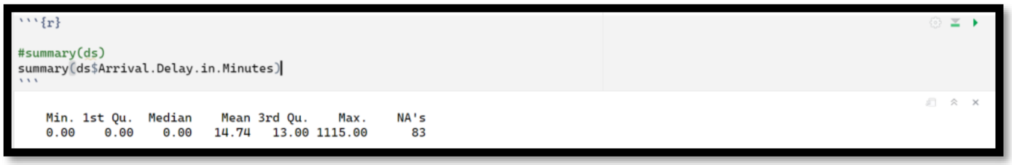


Figure 3: Distribution of Arrival delay in minutes.



~~~
```{r}

#summary(ds)
summary(ds$Arrival.Delay.in.Minutes)
```
~~~

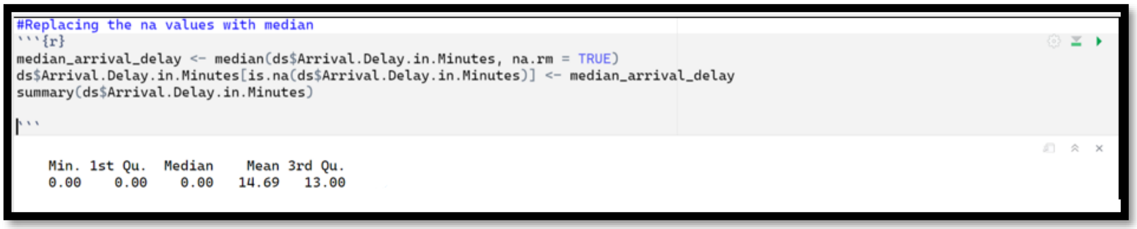
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00    0.00    0.00   14.74   13.00 1115.00      83 
```

Table 4: Descriptive Statistics of Arrival delay

### Observations:

Here from figure 3 and table 4, we can see the distribution arrival delay is right skewed also the summary statistic of Arrival delays says it is very likely that the variable is right-skewed, as the mean (14.74) is greater than the median (0.00) and the maximum value (1115.00) is much larger than the third quartile (13.00). The presence of a large maximum value relative to the rest of the data suggests that there may be a long tail to the right of the distribution, further supporting the idea of a right-skewed distribution so our next step is to remove the NA with median values.

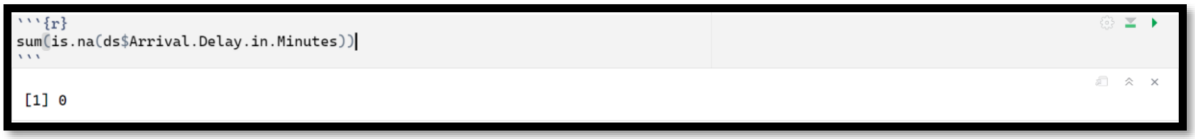
## Replacing the Na values with median



~~~~
#Replacing the na values with median
```{r}
median_arrival_delay <- median(ds$Arrival.Delay.in.Minutes, na.rm = TRUE)
ds$Arrival.Delay.in.Minutes[is.na(ds$Arrival.Delay.in.Minutes)] <- median_arrival_delay
summary(ds$Arrival.Delay.in.Minutes)

```

    Min. 1st Qu.  Median    Mean 3rd Qu.
    0.00    0.00    0.00   14.69   13.00
~~~~

Figure 5: Replacing Na values with median.



~~~~
```{r}
sum(is.na(ds$Arrival.Delay.in.Minutes))
```

 [1] 0
~~~~

Figure 6: Checking if the variable value has been replaced with Na values.

## Observation:

Here from summary analysis **(Figure 2)** of the dataset we can see around **83 observations** in the variable arrival delay in minutes have NA values, considering the presence of the **Na value** we have replaced the **NA value** with its own median value as per the skewness and summary statistics. Now, from figure 6 we can see the dataset does not contain any **Na values**

**Why I removed the outlier from the dataset?**

Outlier have significant impact on the model fitting and prediction in logistic regression, especially when the number of outliers is high or when they influential in determining the relationship between the predictor variable and the outcome. Outlier can lead to biased and inaccurate logistic regression model, resulting in poor prediction and reduced model performance.

## Outlier detection in the dataset.

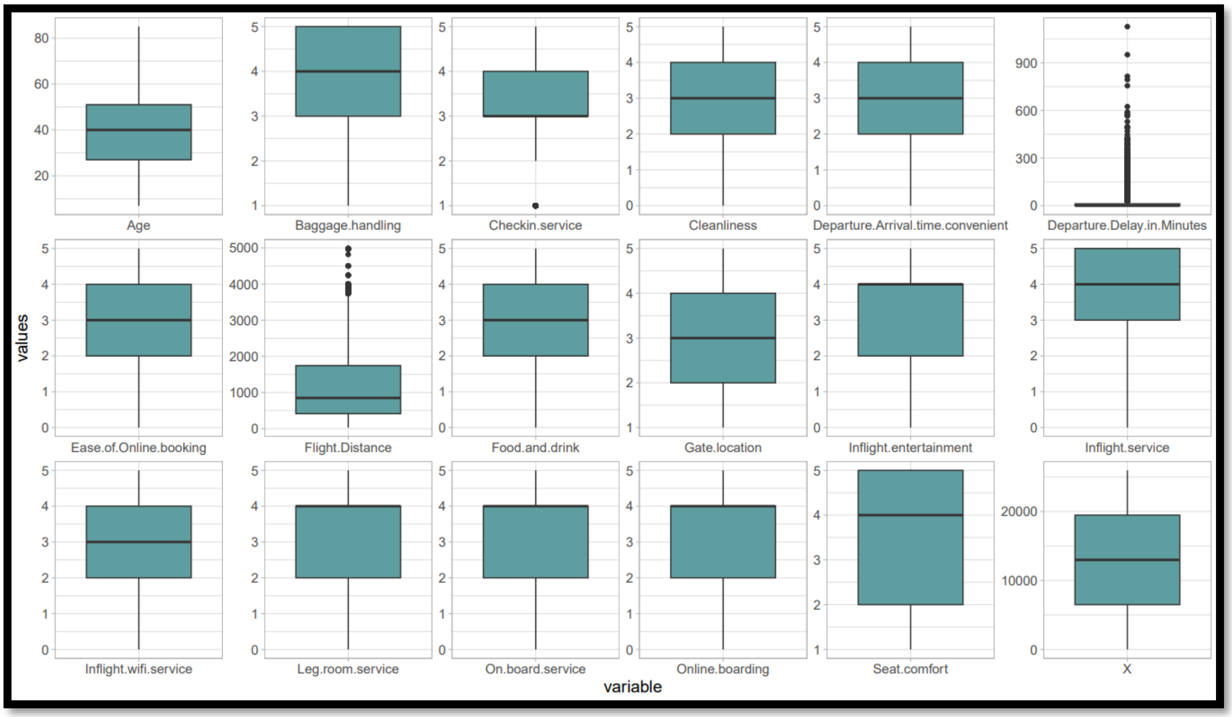
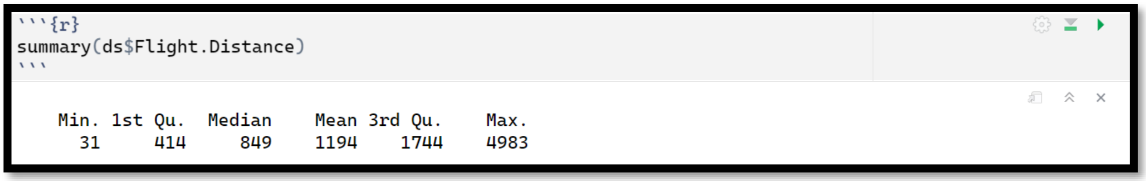
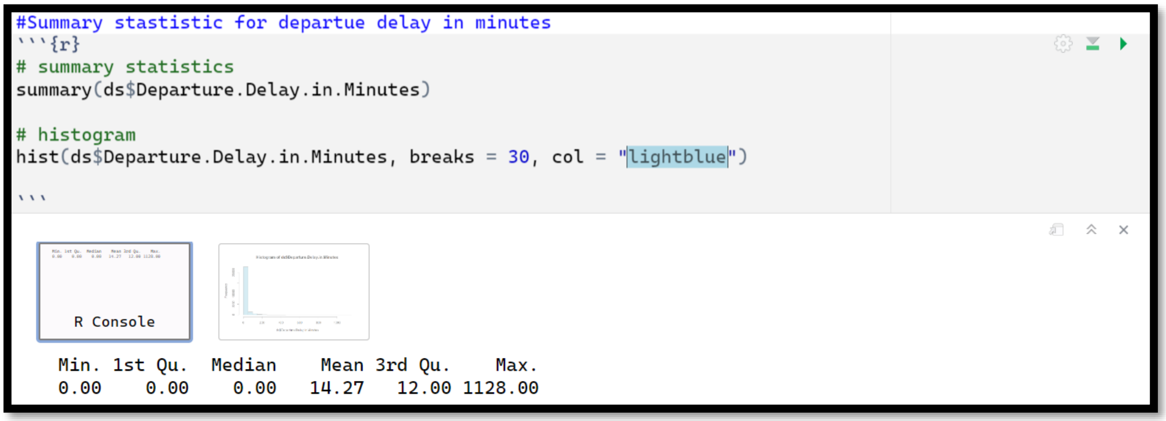


Figure 7: Boxplot (before outlier removal) for outlier detection.



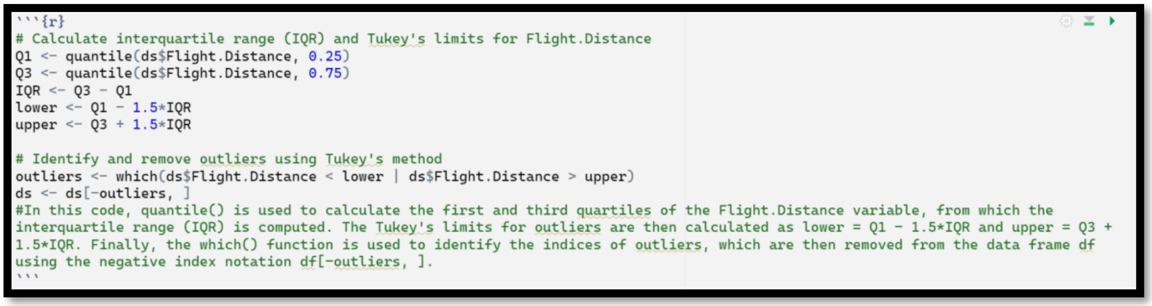
~~~
```{r}
summary(ds$Flight.Distance)
~~~

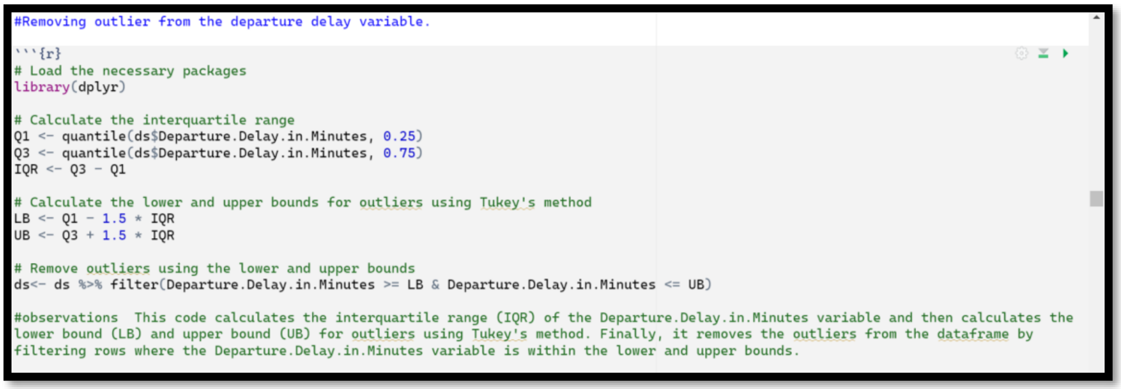
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     31     414     849    1194    1744    4983 
```



```
#Summary stastistic for departue delay in minutes
```{r}
# summary statistics
summary(ds$Departure.Delay.in.Minutes)

# histogram
hist(ds$Departure.Delay.in.Minutes, breaks = 30, col = "lightblue")

```

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.00    0.00   14.27   12.00 1128.00 
```

Figure 7.1 Descriptive statistic of Flight distance and departure delay before removing outlier.

## Observation:
```

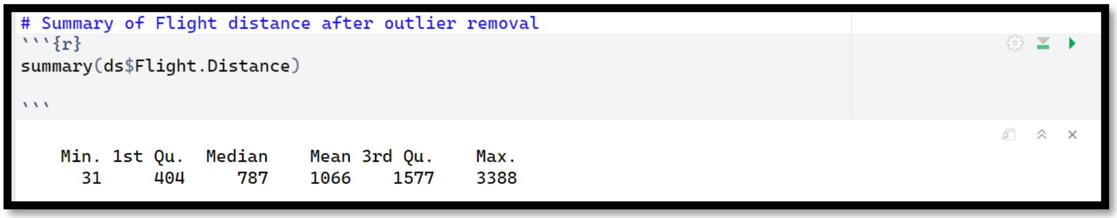
From figure 7, the variables with outliers are departure delay in minutes, arrival, flight distance, and rating on check-in service (not considering it as it is not of my interest)**. I will impute** the outlier from those variables (**Flight.Distance, Departure.Delay.in.Mintues**), as **these variables** are the **interest** of my **further analysis.**

## Outer removal using IQR technique

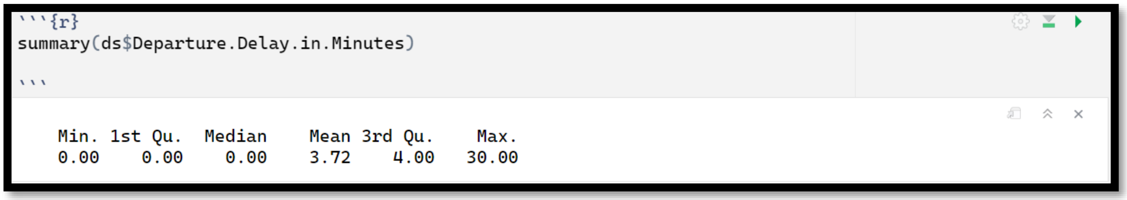


~~~
```{r}
# Calculate interquartile range (IQR) and Tukey's limits for Flight.Distance
Q1 <- quantile(ds$Flight.Distance, 0.25)
Q3 <- quantile(ds$Flight.Distance, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5*IQR
upper <- Q3 + 1.5*IQR

# Identify and remove outliers using Tukey's method
outliers <- which(ds$Flight.Distance < lower | ds$Flight.Distance > upper)
ds <- ds[-outliers, ]
#In this code, quantile() is used to calculate the first and third quartiles of the Flight.Distance variable, from which the
interquartile range (IQR) is computed. The Tukey's limits for outliers are then calculated as lower = Q1 - 1.5*IQR and upper = Q3 +
1.5*IQR. Finally, the which() function is used to identify the indices of outliers, which are then removed from the data frame df
using the negative index notation df[-outliers, ].
~~~
```



```
#Removing outlier from the departure delay variable.

```{r}
# Load the necessary packages
library(dplyr)

# Calculate the interquartile range
Q1 <- quantile(ds$Departure.Delay.in.Minutes, 0.25)
Q3 <- quantile(ds$Departure.Delay.in.Minutes, 0.75)
IQR <- Q3 - Q1

# Calculate the lower and upper bounds for outliers using Tukey's method
LB <- Q1 - 1.5 * IQR
UB <- Q3 + 1.5 * IQR

# Remove outliers using the lower and upper bounds
ds<- ds %>% filter(Departure.Delay.in.Minutes >= LB & Departure.Delay.in.Minutes <= UB)

#observations  This code calculates the interquartile range (IQR) of the Departure.Delay.in.Minutes variable and then calculates the
lower bound (LB) and upper bound (UB) for outliers using Tukey's method. Finally, it removes the outliers from the dataframe by
filtering rows where the Departure.Delay.in.Minutes variable is within the lower and upper bounds.
```



~~~
# Summary of Flight distance after outlier removal
```{r}
summary(ds$Flight.Distance)

```

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     31     404     787    1066    1577    3388 
~~~



~~~
```{r}
summary(ds$Departure.Delay.in.Minutes)

```

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.00    0.00    3.72    4.00   30.00 
~~~

Figure 8: Code for removing outlier from Flight distance and Departure delay variable and its summary statistic after removing outlier.

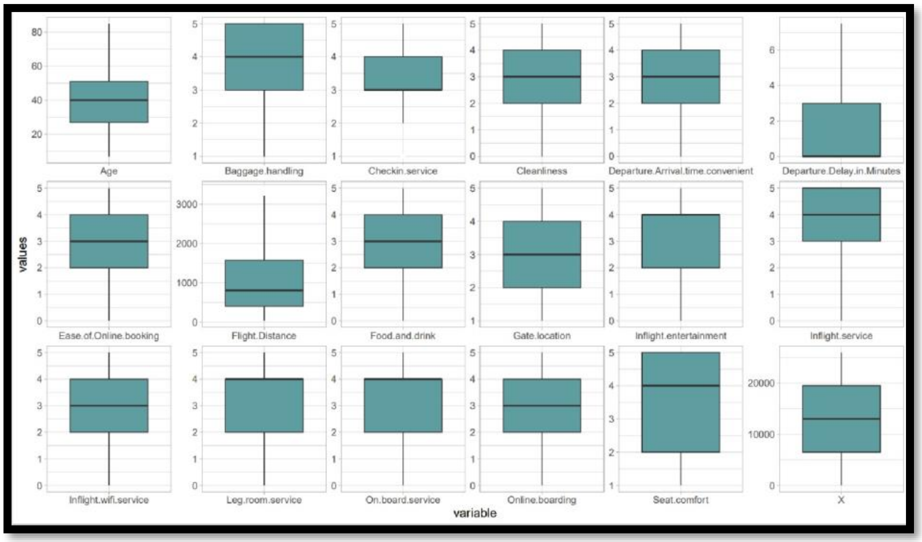


Figure 9: Boxplot (after outlier removal) of the variables.

## Observations:

From figure 8, I have used Tukey's method which is a commonly used technique to identify and remove outliers from a dataset. The method is based on the interquartile range (IQR) of the data, which is the range between the first and third quartiles. Tukey's limits for outliers are calculated as lower = Q1 - 1.5*IQR and upper = Q3 + 1.5*IQR, where Q1 and Q3 are the first and third quartiles, respectively. Any data point outside these limits is considered an outlier and is removed from the dataset. **“After removing the outlier, we observed significant changes in the mean and maximum values of flight distance and departure delay. The mean value of flight distance decreased from 1194 miles to 1106 miles, and the maximum value decreased from 4983 miles to 3388 miles. Similarly, the mean value of departure delay decreased from 14.27 minutes to 3.7 minutes, and the maximum value decreased from 1128 minutes to 30 minutes”.**

**What all variables are correlated with each other.**

## Correlation analysis

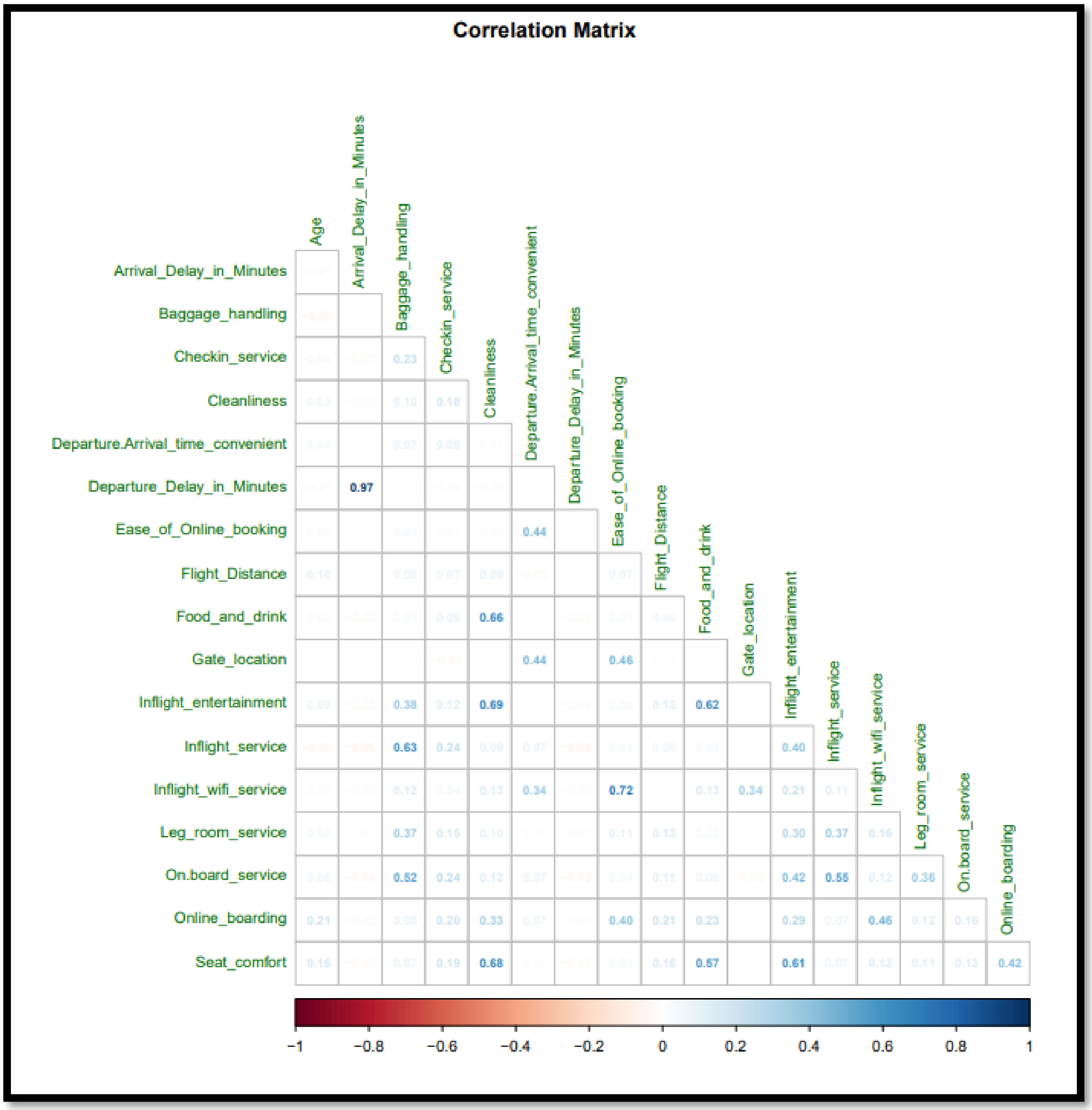


Figure 10: Correlation Matrix analysis.

## Observation.

From the figure 10, here we saw departure delay in minutes and arrival delay minutes are correlated with each other having a value of **0.96 (highly correlated)**. Moreover, **seat comfort** and **cleanliness** have a correlation value of **0.68**, it means that there **is moderate positive correlation** between seat and comfort and cleanliness. This mean that as seat comfort increases, cleanliness is likely to increase as well and vice versa. We found there is positive relationship between inflight wife services and ease of booking with a value of 0.72.

# Methodology: Data analysis with result and findings (discussed as an observations).

In data analysis we will have done the visualization to see the distribution of various features like gender proportion, customer satisfaction, travel history, flight distance etc.

**Distribution of Numerical variable (Discrete and continuous variable).**

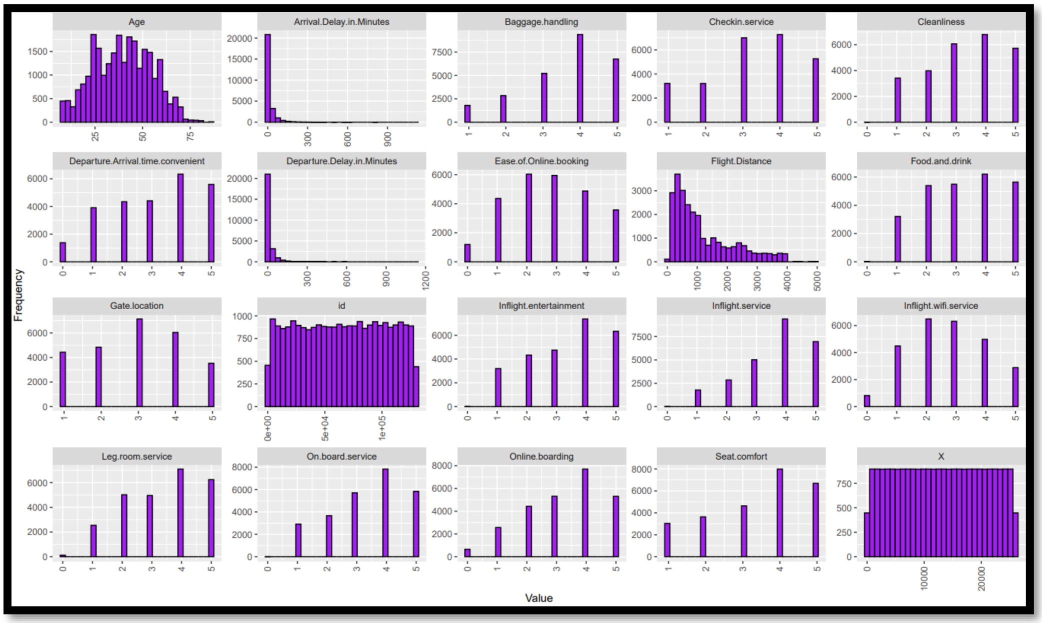


Figure 11: Histogram plot to visualize the distribution of each numeric variable.

## Observation:

Figure 11 displays a histogram that provides a visualization of the distribution of each numerical variable (continuous and discrete variable). Focusing on the Age variable, it appears that the majority of travelers are between the ages of 25 and 50. The visual representations indicate that the **age variable follows a normal distribution**. However, **"Departure delay in min" and "Flight Distance" initially** display **positive skewness**. After removing outliers**, the mean for departure delay is 3.7 min**, and for **flight distance, the mean is 1106 miles**. On the other hand, discrete variables such as **"baggage handling,"** "**check-in service**," **"ease of online booking," "online boarding,"** and **"on-board service"** all appear to follow **a normal distribution** as per their respective distribution plots.

## 1. What is the distribution of Customer based on the gender and what class type the prefer to travel?

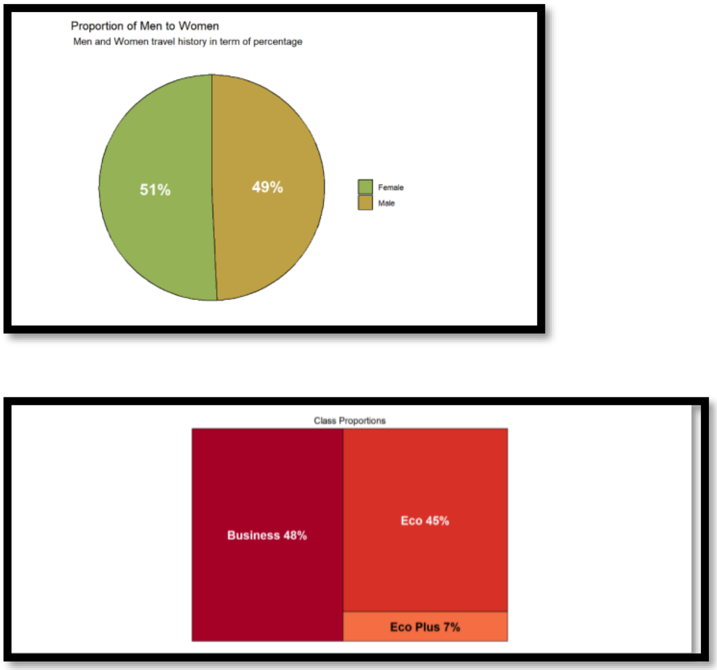


Figure 14: Pie plot for men and women travel history.

### Observation.

From Figure 14 we can say, women made a significant contribution in flight travel with a **percentage of 51** and men were at **around 49 percent**. 48 percent of customer prefer to travel with business and 45 percent travel with economy.

## 2. What is the preference of traveler with the distance of travel?

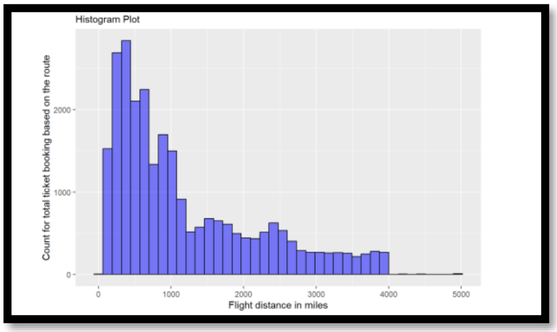


Figure 17. Histogram for flight Distance.

### Observation.

From the above histogram we realized that **majority of travel** has been booked for small **flight distance** between **200 to 1500 miles**, moreover, there is a moderate number of bookings done for the travel between **1000 and 3000 miles,** however **marginal bookings** can be seen after **4000 miles**.

### 3. What is the distribution of customer satisfaction with loyalty?

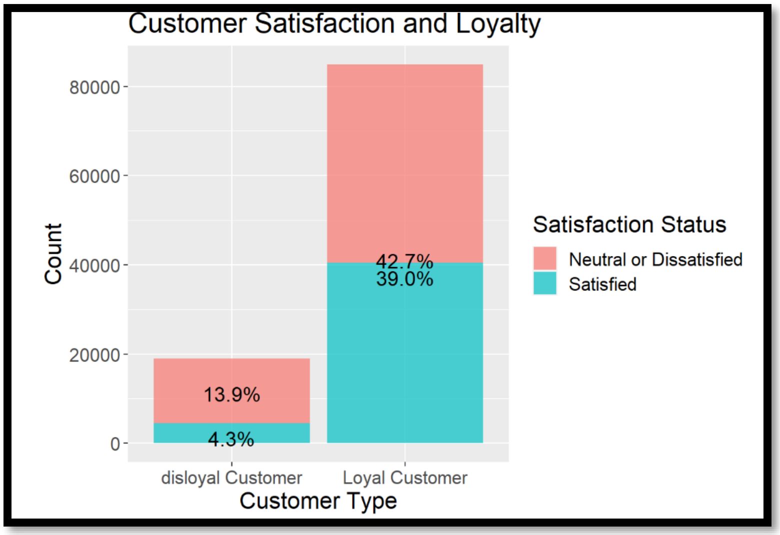


Figure 19: Stacked bar chart to visualize satisfaction over loyalty

### Observations:

The stacked plot above reveals that approximately 4.3% of disloyal customers are content with the airline services, while 13.9% have expressed dissatisfaction. Conversely, a significant contrast is observed among loyal customers, with approximately 42.7% of them expressing dissatisfaction with flight services and the remaining 39% being satisfied.

### 4. What is the Distribution of customer satisfaction over other airline facilities.

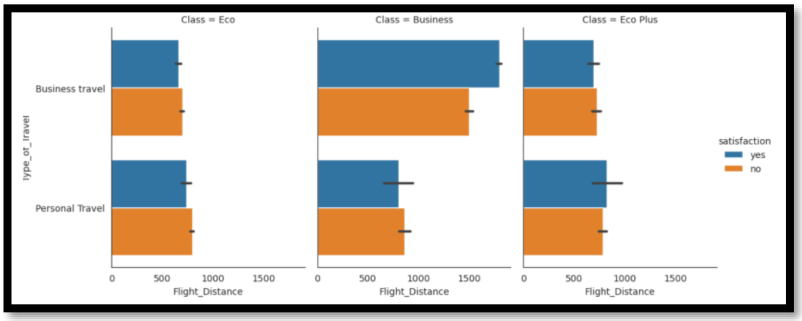


Fig 19.1: Distribution of satisfaction over airline travel class based on the distance and type of travel.

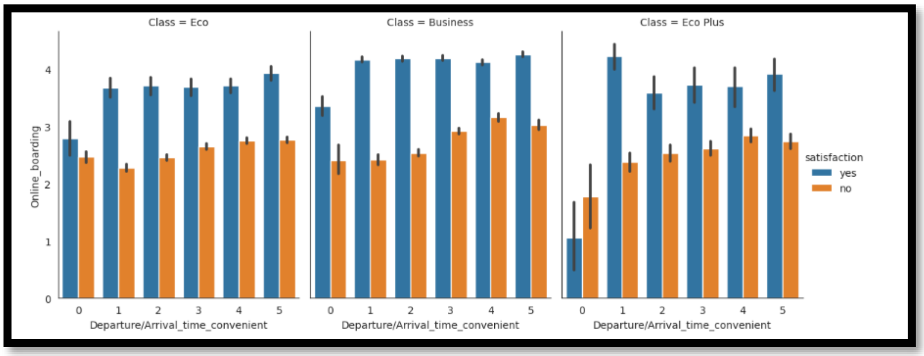


Fig 19.2: Distribution of satisfaction over airline travel class based the online boarding and departure and arrival time convenient.

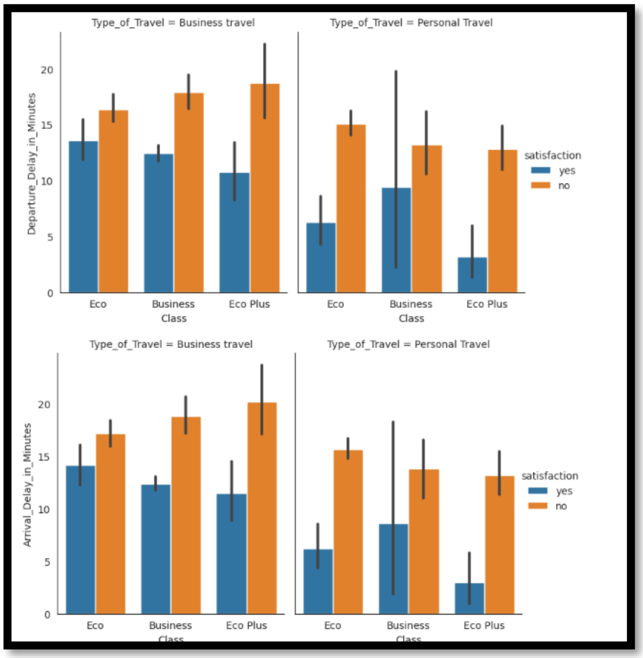


Fig 19.3: Distribution of satisfaction over type travel with respect to type of class and arrival and departure delay.

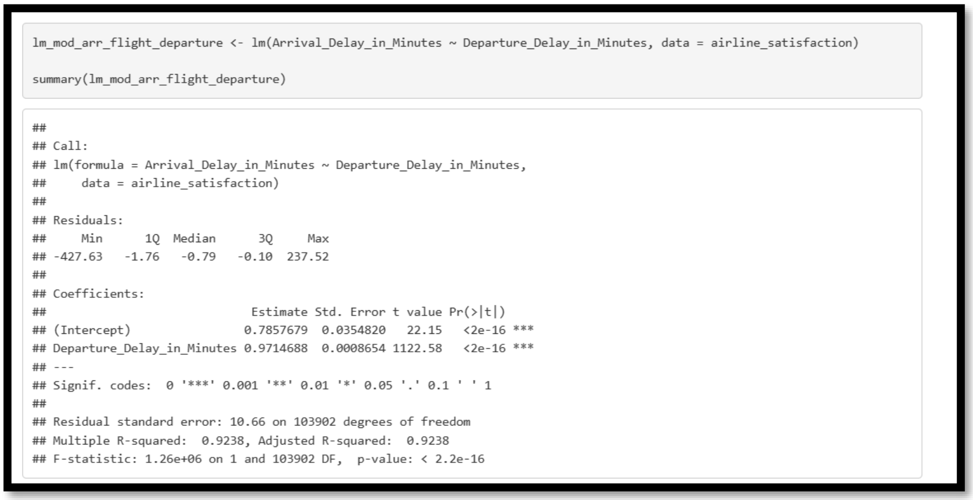
## Observations:

From figure 19.1, The satisfaction levels of airline passengers were studied based on three variables: Type of Travel, Class, and Flight Distance. Passengers in business class who travel longer distances tend to be more satisfied than those in other categories, where satisfaction is distributed equally among satisfied and dissatisfied passengers.

Figure 19.2, Another analysis was performed based on the variables of Online Boarding, Departure/Arrival Time Convenience, and Class. Passengers in the Eco Plus class are more likely to be dissatisfied when Departure/Arrival times are inconvenient, even if the online boarding process is seamless. In contrast, other categories have higher numbers of satisfied passengers than dissatisfied ones.

Figure 19.3, A third analysis was conducted based on the variables of Departure Delay, Arrival Delay, and Type of Travel. Passengers on personal trips, particularly those in the Eco Plus and Eco categories, are more likely to be dissatisfied when there is a high delay in arrival. In general, all combinations have a higher number of dissatisfied passengers than satisfied passengers

## 5. What is the relation between Arrival delay in Minute and Departure delay in minutes. (Linear Regression)



```
lm_mod_arr_flight_departure <- lm(Arrival_Delay_in_Minutes ~ Departure_Delay_in_Minutes, data = airline_satisfaction)

summary(lm_mod_arr_flight_departure)
```

```
##
## Call:
## lm(formula = Arrival_Delay_in_Minutes ~ Departure_Delay_in_Minutes,
##     data = airline_satisfaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -427.63   -1.76   -0.79   -0.10  237.52
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.7857679  0.0354820   22.15   <2e-16 ***
## Departure_Delay_in_Minutes 0.9714688  0.0008654 1122.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.66 on 103902 degrees of freedom
## Multiple R-squared:  0.9238, Adjusted R-squared:  0.9238
## F-statistic: 1.26e+06 on 1 and 103902 DF,  p-value: < 2.2e-16
```

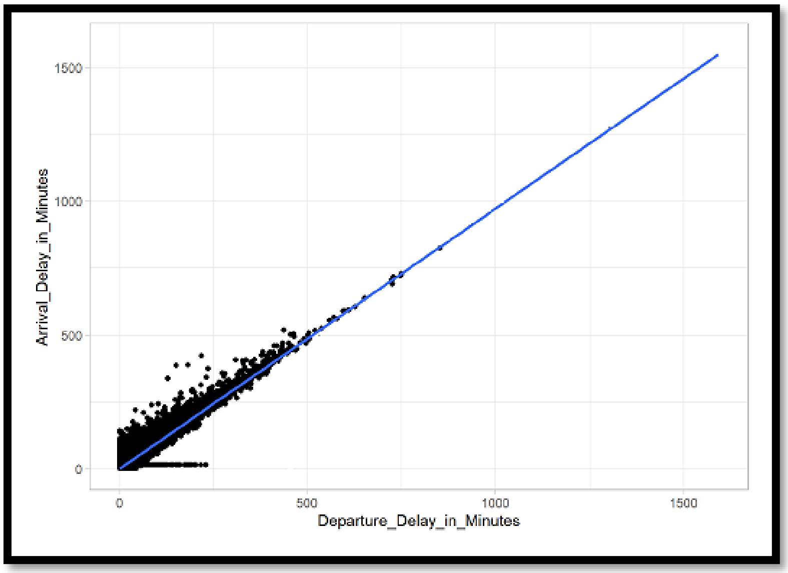
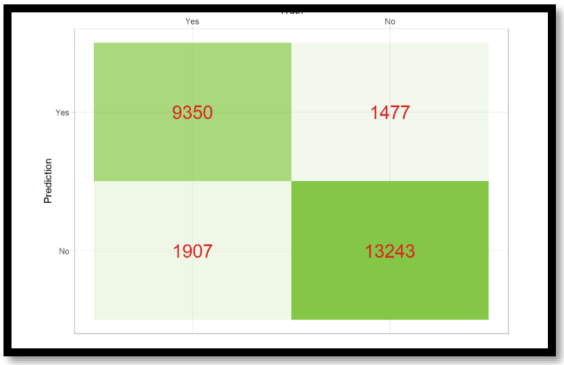


Figure 21: Linear regression for finding the Relation between Arrival Delay in minutes and departure delay in minutes.

### Observation:

To analyze the relation between variable; I have conducted the linear regression. The response variable is "**Arrival_Delay_in_Minutes**" and the predictor variable is "**Departure_Delay_in_Minutes**". The coefficients table shows that the intercept (the expected mean value of the response when the predictor is zero) is estimated to be 0.79, and the slope (the change in the response for each unit change in the predictor) is estimated to be 0.97**. The t-value and p-value** for each coefficient test the hypothesis that the corresponding **coefficient is equal to zero**. In this case, the p-values for both coefficients are very small, indicating that the corresponding coefficients are significantly different from zero and suggesting a strong relationship between **"Departure_Delay_in_Minutes" and "Arrival_Delay_in_Minutes".**

## Logistic Regression and prediction of customer satisfactions (Research Question 1)



| Prediction \ Truth | Yes | No |
|---|---|---|
| Yes | 9350 | 1477 |
| No | 1907 | 13243 |

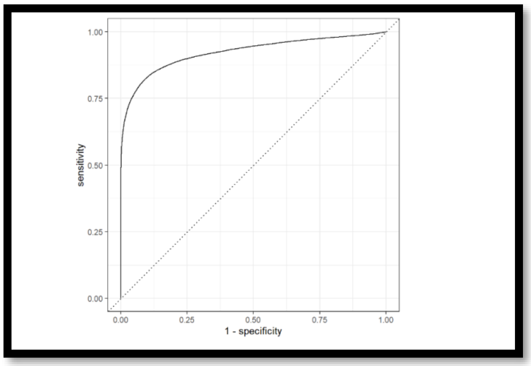
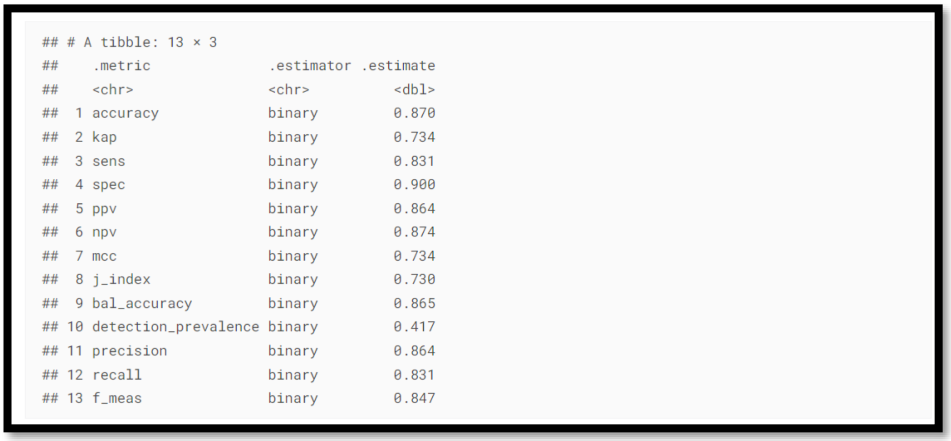
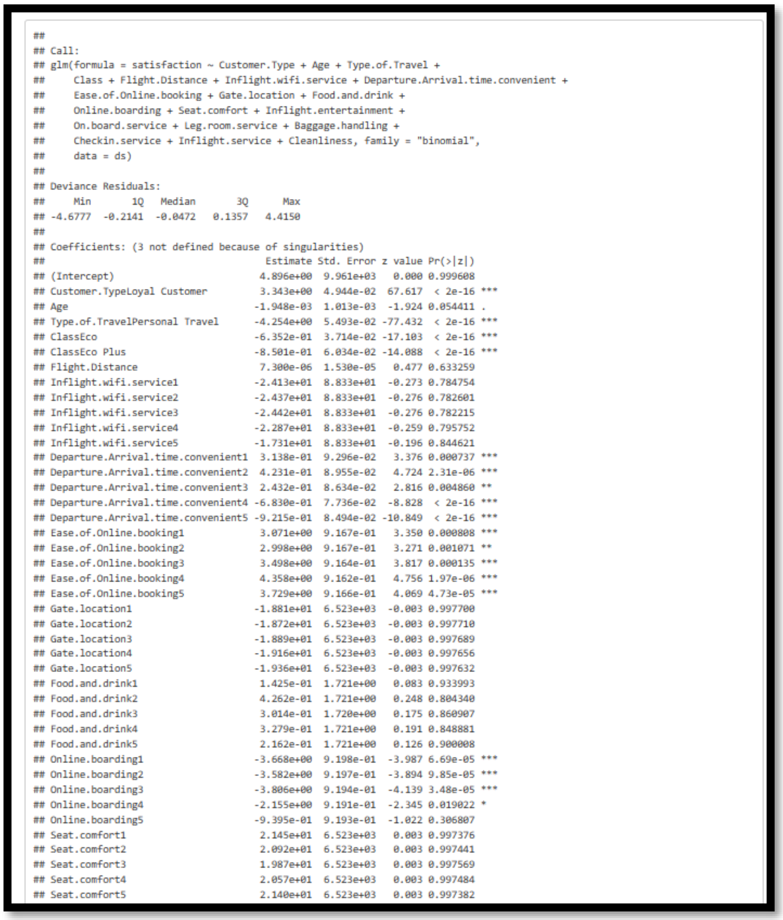


Figure 22: Confusion matrix of Logistic regression      Figure 23: Roc curve of Logistic Regressions.



```
## # A tibble: 13 × 3
##    .metric              .estimator .estimate
##    <chr>                <chr>          <dbl>
##  1 accuracy             binary         0.870
##  2 kap                  binary         0.734
##  3 sens                 binary         0.831
##  4 spec                 binary         0.900
##  5 ppv                  binary         0.864
##  6 npv                  binary         0.874
##  7 mcc                  binary         0.734
##  8 j_index              binary         0.730
##  9 bal_accuracy         binary         0.865
## 10 detection_prevalence binary         0.417
## 11 precision            binary         0.864
## 12 recall               binary         0.831
## 13 f_meas               binary         0.847
```

Figure 24: Logistic regression metrics.



```
##
## Call:
## glm(formula = satisfaction ~ 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 +
##     Checkin.service + Inflight.service + Cleanliness, family = "binomial",
##     data = ds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6777  -0.2141  -0.0472   0.1357   4.4150
##
## Coefficients: (3 not defined because of singularities)
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       4.896e+00  9.961e+03   0.000 0.999608
## Customer.TypeLoyal Customer       3.343e+00  4.944e-02  67.617  < 2e-16 ***
## Age                              -1.948e-03  1.013e-03  -1.924 0.054411 .
## Type.of.TravelPersonal Travel    -4.254e+00  5.493e-02 -77.432  < 2e-16 ***
## ClassEco                         -6.352e-01  3.714e-02 -17.103  < 2e-16 ***
## ClassEco Plus                    -8.501e-01  6.034e-02 -14.088  < 2e-16 ***
## Flight.Distance                   7.300e-06  1.530e-05   0.477 0.633259
## Inflight.wifi.service1           -2.413e+01  8.833e+01  -0.273 0.784754
## Inflight.wifi.service2           -2.437e+01  8.833e+01  -0.276 0.782601
## Inflight.wifi.service3           -2.442e+01  8.833e+01  -0.276 0.782215
## Inflight.wifi.service4           -2.287e+01  8.833e+01  -0.259 0.795752
## Inflight.wifi.service5           -1.731e+01  8.833e+01  -0.196 0.844621
## Departure.Arrival.time.convenient1  3.138e-01  9.296e-02   3.376 0.000737 ***
## Departure.Arrival.time.convenient2  4.231e-01  8.955e-02   4.724 2.31e-06 ***
## Departure.Arrival.time.convenient3  2.432e-01  8.634e-02   2.816 0.004860 **
## Departure.Arrival.time.convenient4 -6.830e-01  7.736e-02  -8.828  < 2e-16 ***
## Departure.Arrival.time.convenient5 -9.215e-01  8.494e-02 -10.849  < 2e-16 ***
## Ease.of.Online.booking1           3.071e+00  9.167e-01   3.350 0.000808 ***
## Ease.of.Online.booking2           2.998e+00  9.167e-01   3.271 0.001071 **
## Ease.of.Online.booking3           3.498e+00  9.164e-01   3.817 0.000135 ***
## Ease.of.Online.booking4           4.358e+00  9.162e-01   4.756 1.97e-06 ***
## Ease.of.Online.booking5           3.729e+00  9.166e-01   4.069 4.73e-05 ***
## Gate.location1                   -1.881e+01  6.523e+03  -0.003 0.997700
## Gate.location2                   -1.872e+01  6.523e+03  -0.003 0.997710
## Gate.location3                   -1.889e+01  6.523e+03  -0.003 0.997689
## Gate.location4                   -1.916e+01  6.523e+03  -0.003 0.997656
## Gate.location5                   -1.936e+01  6.523e+03  -0.003 0.997632
## Food.and.drink1                   1.425e-01  1.721e+00   0.083 0.933993
## Food.and.drink2                   4.262e-01  1.721e+00   0.248 0.804340
## Food.and.drink3                   3.014e-01  1.720e+00   0.175 0.860907
## Food.and.drink4                   3.279e-01  1.721e+00   0.191 0.848881
## Food.and.drink5                   2.162e-01  1.721e+00   0.126 0.900008
## Online.boarding1                 -3.668e+00  9.198e-01  -3.987 6.69e-05 ***
## Online.boarding2                 -3.582e+00  9.197e-01  -3.894 9.85e-05 ***
## Online.boarding3                 -3.806e+00  9.194e-01  -4.139 3.48e-05 ***
## Online.boarding4                 -2.155e+00  9.191e-01  -2.345 0.019022 *
## Online.boarding5                 -9.395e-01  9.193e-01  -1.022 0.306807
## Seat.comfort1                     2.145e+01  6.523e+03   0.003 0.997376
## Seat.comfort2                     2.092e+01  6.523e+03   0.003 0.997441
## Seat.comfort3                     1.987e+01  6.523e+03   0.003 0.997569
## Seat.comfort4                     2.057e+01  6.523e+03   0.003 0.997484
## Seat.comfort5                     2.140e+01  6.523e+03   0.003 0.997382
```

Figure 25: Summary of prediction on customer satisfaction

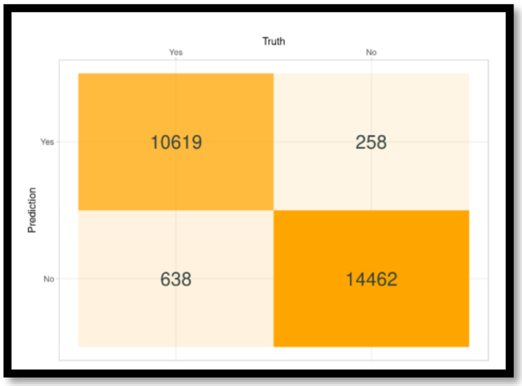
**Observations:**

For logistic regression model building I have used folds in **data validation** because it eventually reduced the **risk of variance** in the **evaluation metric** and consequently **minimizing the risk of overfitting** as the model test on data has not seen before. From figure 23 and 24 we can see the **accuracy of 87 percent which means model is able to correctly predict the target class 87 percent of time.**

The given output is a logistic regression output used for predicting the satisfaction of passengers based on various predictors. **The significant predictors with their respective coefficients and p-values are: Customer.TypeLoyal , Type.of.TravelPersonal Travel, ClassEco , ClassEco Plus, departure arrival time(categories 1-5) as Ease of online booking(categories 1-5) as the p value is less than 0.05 or almost**

**equal to 0.05 and those variable are considered to significant when predicting the customer satisfactions**

# Identifying Key Variable Contributions to Customer Satisfaction Using Random Forests. (Research Question2)



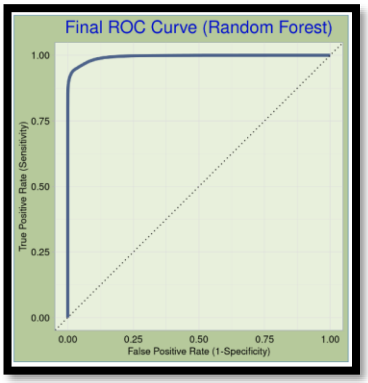
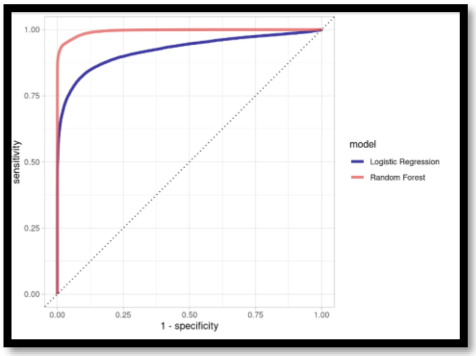


Figure 26: Confusion matrix of Random forests. Figure 27: Random Forest Roc curve



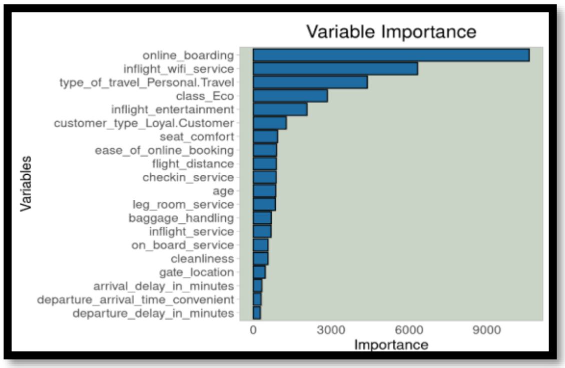
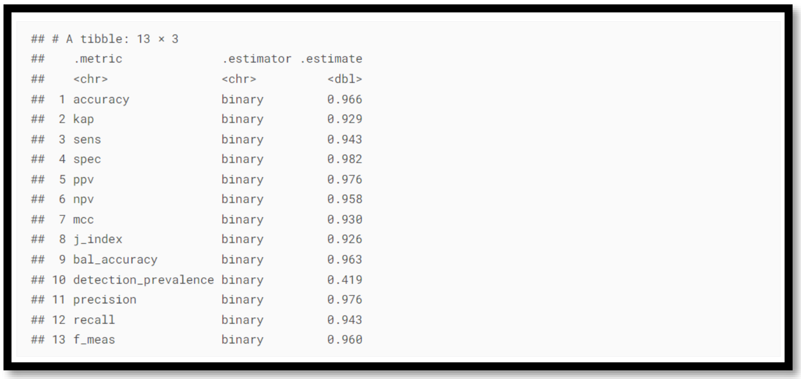


Figure 28: Comparison of Roc Figure 29: Importance variable from Random forests



```
## # A tibble: 13 × 3
##    .metric              .estimator .estimate
##    <chr>                <chr>          <dbl>
##  1 accuracy             binary         0.966
##  2 kap                  binary         0.929
##  3 sens                 binary         0.943
##  4 spec                 binary         0.982
##  5 ppv                  binary         0.976
##  6 npv                  binary         0.958
##  7 mcc                  binary         0.930
##  8 j_index              binary         0.926
##  9 bal_accuracy         binary         0.963
## 10 detection_prevalence binary         0.419
## 11 precision            binary         0.976
## 12 recall               binary         0.943
## 13 f_meas               binary         0.960
```

Figure 30: Random Forest Metrics.

**Observations:**

From figure 26, 27, 28, 29, 30. The accuracy of the random forest model is measured at 0.966, indicating that it correctly predicts customer satisfaction in 96.6% of cases. And in terms of accuracy random forest has performed better than the logistic regressions. This metric provides an overall assessment of the model's performance. The sensitivity (sens) of the model, which represents its ability to correctly identify positive cases of customer satisfaction, is estimated to be 0.943. This means that the model successfully identifies satisfied customers in approximately 94.3% of cases. From **random forests**

When analyzing the satisfaction levels of passengers in the airline industry and determining the factors that **significantly impact their satisfaction**, it becomes evident that the following elements play a crucial role: **Online boarding experience, Availability of in-flight WIFI service, Type of travel, particularly personal travel, Class of service, specifically economy class, In-flight entertainment options, Customer type, particularly loyal customers, Seat comfort during the flight, Ease of online booking process, Quality of leg room service**. These factors have been identified as the key drivers of passenger satisfaction in the airline industry.

# Decision trees (Based on the recommendation of random forest important variables)

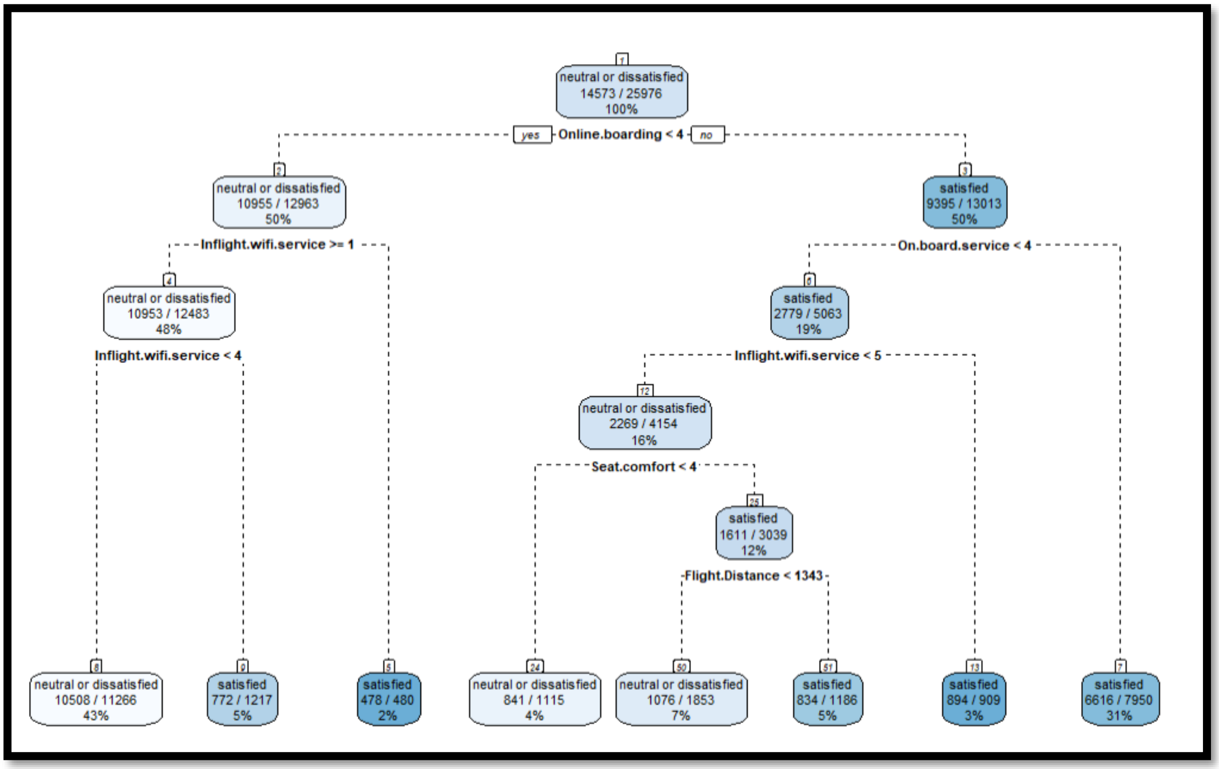


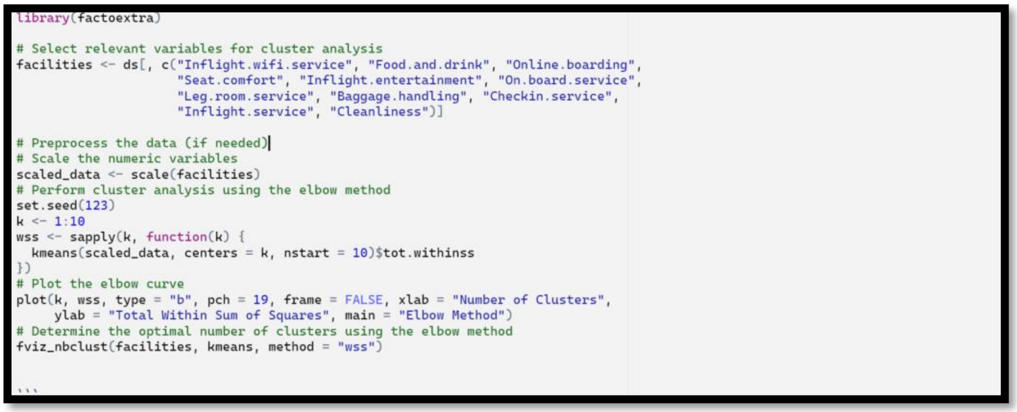
Figure 33: Decision Trees visualizations

**Observations:**

The code selects a set of important variables from a dataset based on the random forest models, including factors such as online boarding, in-flight Wi-Fi service, customer type, flight distance, gate location, food and drink, seat comfort, and on-board service. These variables are then used to create a decision tree model to predict customer satisfaction. The decision tree analysis provides insights into the factors influencing customer satisfaction in the airline industry. With 25,976 observations, the root node reveals that a majority of customers (11,403) are classified as "neutral or dissatisfied" (56.1% probability), while the remaining customers (14,573) are classified as "satisfied" (43.9% probability). The tree splits based on the value of the "Online.boarding" variable. If the value is less than 3.5, it leads to node 4, which predicts a "neutral or dissatisfied" outcome with a higher probability (27.8%). Conversely, if the value is greater than or equal to

3.5, it leads to node 3, which predicts a "satisfied" outcome with a higher probability (72.1%). Node 4 further splits based on the "Inflight.wifi.service" variable. For values less than 0.5, node 5 predicts a "satisfied" outcome with a very high probability (99.6%). For values greater than or equal to 0.5, node 8 predicts a "neutral or dissatisfied" outcome with a higher probability (93.3%). However, if the "Inflight.wifi.service" value is 3.5 or higher, node 9 predicts a "satisfied" outcome with a higher probability (63.4%). These findings highlight the importance of variables such as online boarding and in-flight Wi-Fi service in determining customer satisfaction levels in the airline industry.

## Clustering. (How can we understand and segment our airline customers based on their satisfaction with various facilities?)



~~~
library(factoextra)

# Select relevant variables for cluster analysis
facilities <- ds[, c("Inflight.wifi.service", "Food.and.drink", "Online.boarding",
                     "Seat.comfort", "Inflight.entertainment", "On.board.service",
                     "Leg.room.service", "Baggage.handling", "Checkin.service",
                     "Inflight.service", "Cleanliness")]

# Preprocess the data (if needed)
# Scale the numeric variables
scaled_data <- scale(facilities)
# Perform cluster analysis using the elbow method
set.seed(123)
k <- 1:10
wss <- sapply(k, function(k) {
  kmeans(scaled_data, centers = k, nstart = 10)$tot.withinss
})
# Plot the elbow curve
plot(k, wss, type = "b", pch = 19, frame = FALSE, xlab = "Number of Clusters",
     ylab = "Total Within Sum of Squares", main = "Elbow Method")
# Determine the optimal number of clusters using the elbow method
fviz_nbclust(facilities, kmeans, method = "wss")


```
~~~

Figure 34: Code for cluster analysis

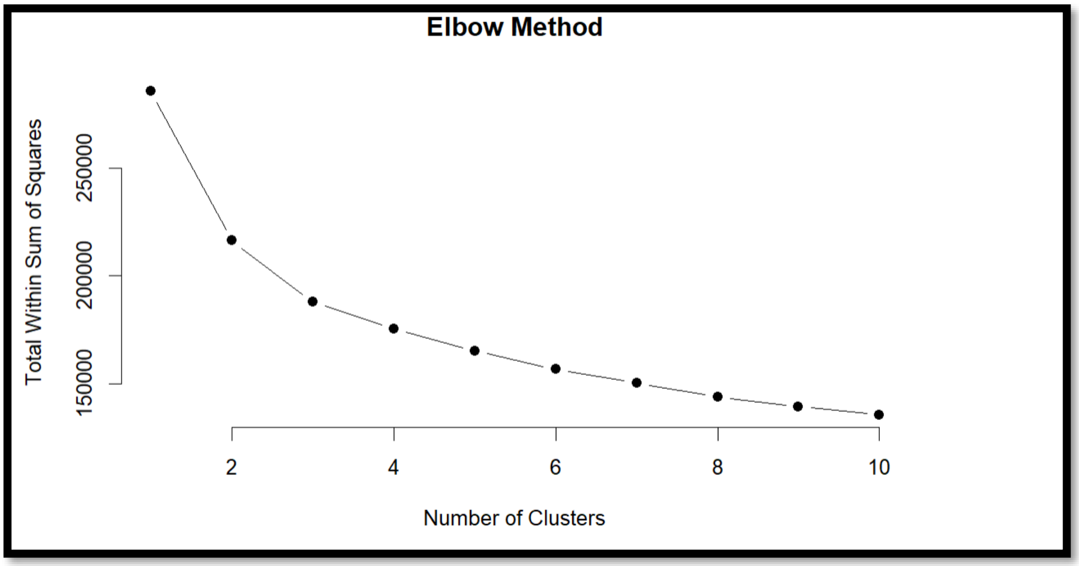


Figure 35: Visualizations for optimal number of clusters.

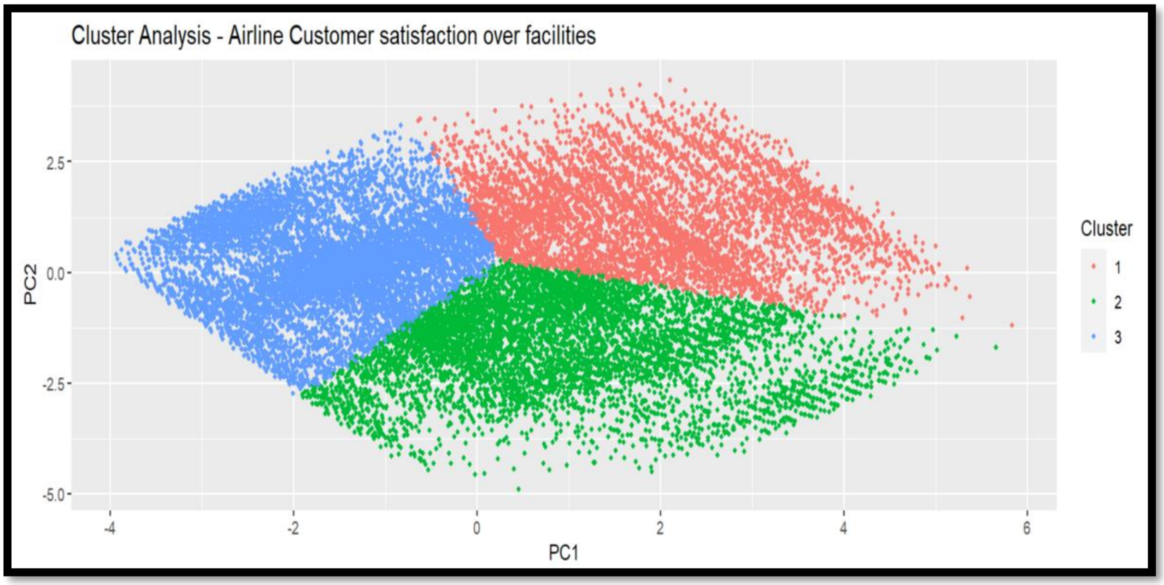


Figure 36: Customer satisfaction over facilities

**Observations:**

We performed k-means clustering with an optimal number of clusters (in this case, 3) on the scaled dataset of facility satisfaction ratings. The k-means algorithm assigns each observation to one of the clusters based on its similarity to the cluster centroids. We added the cluster assignment to the dataset.

To visualize and interpret the clusters, we used Principal Component Analysis (PCA) to reduce the dimensionality of the data. PCA identifies the principal components that explain the most variance in the dataset. In this case, we used the first two principal components, PC1 and PC2, to create a scatter plot. The scatter plot represents each observation as a point in the PC1-PC2 space, and the color of the point represents the assigned cluster. This visualization allows us to observe the separation and grouping of the observations into different clusters based on their facility satisfaction ratings. PC1 and PC2 are used because they capture the most significant variation in the data and provide a concise representation of the clusters' structure.

The plot helps us understand the distinct groups of customers based on their satisfaction with the facilities. By analyzing these clusters, airlines can gain insights into customer preferences, identify target segments, and develop tailored strategies to enhance customer satisfaction and improve their overall service quality.

## Interpretations from overall EDA Analysis and Research questions.

The given analysis provides important insights into the airline industry and highlights key areas that require attention for improving customer satisfaction.

Firstly, for **business class passengers**, it may be **beneficial to invest in longer flights** or offer more **amenities on shorter flights to increase satisfaction levels**. **Secondly, airlines could focus on improving Departure/Arrival Time Convenience for passengers in Eco Plus class to reduce dissatisfaction levels**, **as this was found to be a significant factor impacting their satisfaction levels**. Moreover, the strong positive correlation between departure delay and arrival delay highlights the importance of minimizing departure delays to reduce the likelihood of arrival delays. **Additionally, airlines should consider compensation or other measures for passengers experiencing significant Departure or Arrival Delays, particularly those on personal trips, to address their dissatisfaction**. **The analysis suggests that optimizing flight schedules for short to medium distances could lead to more bookings and increased revenue for airlines**.

Based on the analysis using random forests, the airline industry can enhance passenger satisfaction by focusing on improving the online boarding experience, providing reliable and accessible in-flight WiFi service, tailoring services for passengers traveling for personal reasons, enhancing comfort and amenities in economy class, offering a diverse range of in-flight entertainment options, recognizing and appreciating loyal customers through personalized services and incentives, ensuring comfortable seating arrangements and amenities for passenger comfort, simplifying and streamlining the online booking process, and ensuring adequate legroom space and attentive service. By addressing these factors, the airline industry can significantly improve overall customer satisfaction and deliver a better travel experience to passengers.

In conclusion, the given analysis provides valuable insights into the airline industry and highlights important areas that need attention for improving customer satisfaction. By addressing the issues identified, airlines could improve their services and attract more customers, leading to increased revenue and profitability.

# Summary and Conclusion.

In the initial project report of Aly 6040 we have worked on the flight customer satisfaction dataset where we have cleaned and done the EDA. Following observation can be made.

- Descriptive statistics revealed that the arrival delay in minutes had 83 null values.
- After removing these null values and replacing them with the median, the mean of the arrival delay changed from 14.74 to 14.69.
- Outliers were identified in the departure delay in minutes and flight distance variables using Tukey limits.
- After replacing the outliers, the mean of the departure delay changed from 14.31 to 3.7, and maximum value of departure delay changed from 1128 to 30 minutes and the flight distance mean changed from 1194 to 1106, and maximum value of flight distance decreased from 4983 to 3388 miles.
- A correlation plot revealed a strong positive correlation (0.96) between departure delay and arrival delay.
- 43% of customers were satisfied, and 57% were not satisfied.
- Men contributed 51% to the airline industry, while women contributed 49%.

- 48% of customers preferred to travel in business class, while 45% preferred economy class.
- A histogram analysis showed that most travel was booked for small flight distances between 200 to 1500 miles, with marginal bookings above 4000 miles.
- **Linear regression** revealed a strong positive relationship between arrival delay and departure delay, indicating that when the flight departs late, arrival delay is also impacted.
- From **logistic regression we** have seen the most significant predictor variable for predicting customer satisfaction are Customer.Type, Type.of.Travel, Class, Departure.Arrival.time.convenient, and Ease.of.Online.booking. These predictors have high **coefficient values and very low p-values.**
- From **Radom Forest** we have witnessed the most significant variable contributing to customer satisfaction are: Online boarding, inflight WIFI services, type of travel, class economics, inflight entertainment etc. and based on the random forest important variables recommendation we have created the decision trees.
- We improvised the **K-means clustering using the elbow method** and selected the optimal number of three clusters. This allowed us to understand the customer satisfaction and identify groups of customers with similar interests regarding various flight facilities.

# References and Bibliography.

1. Find duplicated rows (based on 2 columns) in Data Frame in R. (2011, August 8). Stack Overflow. https://stackoverflow.com/questions/6986657/find-duplicated-rows-based-on-2-columns-in-data-frame-in-r
2. Nguyen, C. (2021, September 29). Guide To Data Visualization With ggplot2 - Towards Data Science. Medium. https://towardsdatascience.com/guide-to-data-visualization-with-ggplot2-in-a-hour-634c7e3bc9dd
3. A Grammar of Data Manipulation. (2021). Dplyr. https://dplyr.tidyverse.org/
4. A Grammar of Data Manipulation. (2021). Dplyr. https://dplyr.tidyverse.org/.