Aviation Project

Capstone project

Author: Smitayan Nandy
Date – 21/01/2020

Table of Contents

1. Problem Statement/Objective

This is the dilemma of a reputed US airline carrier 'Falcon airlines'. They aim to determine the relative importance of each parameter with regards to their contribution to passenger satisfaction. Provided is a random sample of 90917 individuals who travelled using their flights. The on-time performance of the flights along with the passengers' information is published in the file named 'Flight data'. These passengers were asked to provide their feedback at the end of their flights on various parameters along with their overall experience. These collected details are made available in the survey report labelled 'Survey data'.

In the survey, the passengers were explicitly asked whether they were satisfied with their overall flight experience and that is captured in the data of survey report under the variable labelled 'Satisfaction'.

The two objective of this project are-

1. To understand which parameters play an important role in swaying a passenger feedback towards 'satisfied'

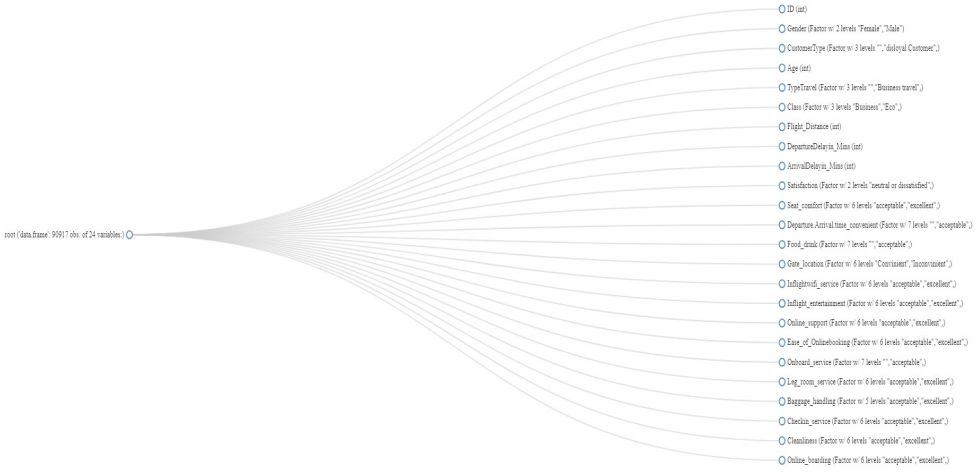
2. To predict whether a passenger will be satisfied or not given the rest of the details are provided.

The findings of this study will hopefully help Falcon airlines to channelize their resources on specific parameters in order to give customers a better flight experience; which in turn can boost their customer base and increase profit.

2. Data

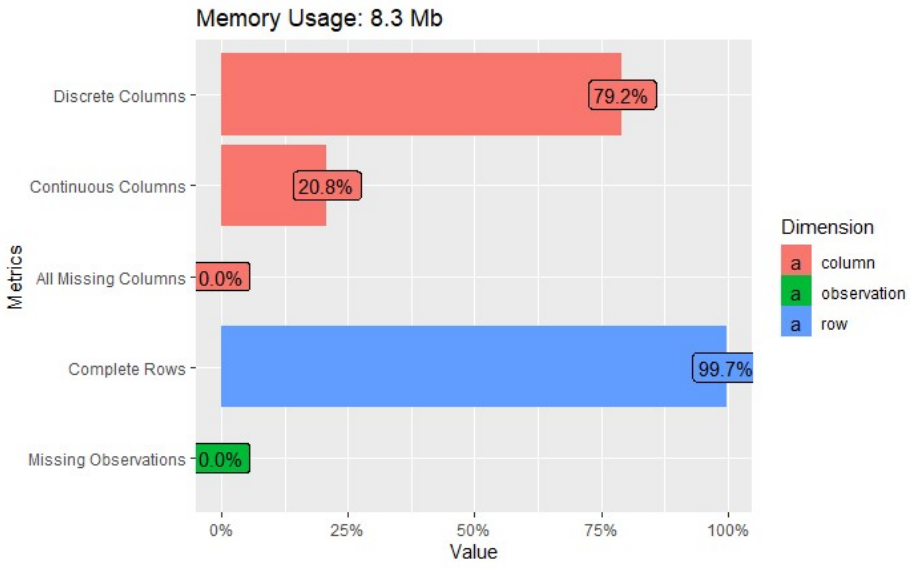
Both the survey data and flight information data have been joined on customer ID field to create one single data set called Aviation. The combined dataset has 90917 observations and 24 variables.

a. Structure

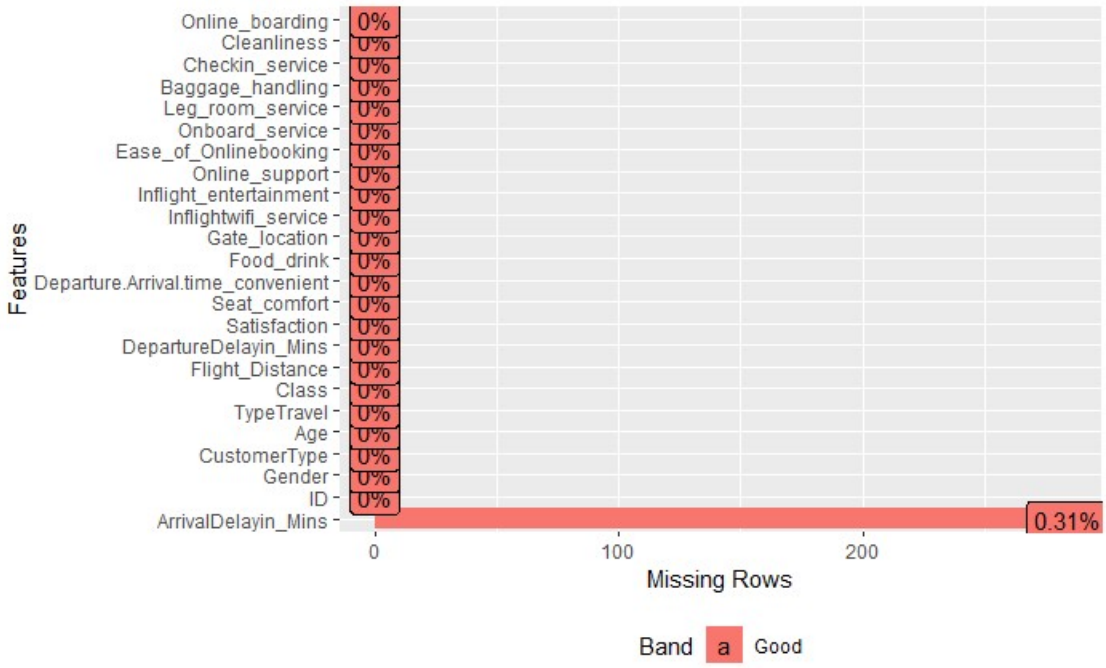


ID	Int
Gender	Factor with 2 levels
CustomerType	Factor with 3 levels
Age	Int
TypeTravel	Factor with 3 levels
Class	Factor with 3 levels
Flight_Distance	Int
DepartureDelayin_Mins	Int
ArrivalDelayin_Mins	Int
Satisfaction	Factor with 2 levels
Seat_Comfort	Factor with 6 levels
Departure.Arrival.time_convenient	Factor with 7 levels
Food_drink	Factor with 7 levels
Gate_location	Factor with 6 levels
Inflightwifi_service	Factor with 6 levels
Inflight_entertainment	Factor with 6 levels
Online_support	Factor with 6 levels
Ease_of_Onlinebooking	Factor with 6 levels
Onboard_service	Factor with 7 levels
Leg_room_service	Factor with 6 levels
Baggage_handling	Factor with 5 levels
Checkin_service	Factor with 6 levels
Cleanliness	Factor with 6 levels
Online_boarding	Factor with 6 levels

b. Distribution between Discrete and Continuous columns



c. Null Check



ArrivalDelayin_Mins have 284 observations with null value.

d. Summary of Data

Flight data:

```
   Gender                CustomerType       Age                TypeTravel        Class       Flight_Distance DepartureDelayin_Mins ArrivalDelayin_Mins
Female:46186                   : 9099   Min.   : 7.00                 : 9088   Business:43535   Min.   :  50    Min.   :   0.00       Min.   :   0.00
Male  :44731   disloyal Customer:14921   1st Qu.:27.00   Business travel:56481   Eco     :40758   1st Qu.:1360    1st Qu.:   0.00       1st Qu.:   0.00
               Loyal Customer   :66897   Median :40.00   Personal Travel:25348   Eco Plus: 6624   Median :1927    Median :   0.00       Median :   0.00
                                         Mean   :39.45                                            Mean   :1982    Mean   :  14.69       Mean   :  15.06
                                         3rd Qu.:51.00                                            3rd Qu.:2542    3rd Qu.:  12.00       3rd Qu.:  13.00
                                         Max.   :85.00                                            Max.   :6950    Max.   :1592.00       Max.   :1584.00
                                                                                                                                        NA's   :284
```

Survey Data:

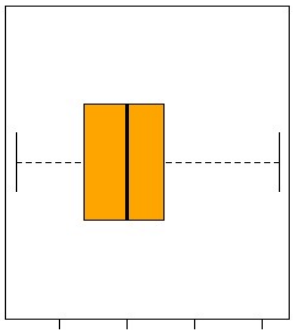
```
              Satisfaction            Seat_comfort   Departure.Arrival.time_convenient            Food_drink             Gate_location        Inflightwifi_service
neutral or dissatisfied:41156   acceptable     :20552                    : 8244                          : 8181   Convinient      :21088   acceptable      :19199
satisfied               :49761   excellent      :12519   acceptable      :14806           acceptable     :17991   Inconvinient    :15876   excellent       :20258
                                 extremely poor : 3368   excellent       :17079           excellent      :12947   manageable      :23385   extremely poor :   96
                                 good           :19789   extremely poor  : 4199           extremely poor : 3794   need improvement :17113   good            :22159
                                 need improvement:20002   good            :18840           good           :17245   very convinient  :13454   need improvement:18894
                                 poor           :14687   need improvement:14539           need improvement:17359   very inconvinient:    1   poor            :10311
                                                         poor            :13210           poor           :13400
     Inflight_entertainment          Online_support      Ease_of_Onlinebooking        Onboard_service          Leg_room_service         Baggage_handling
acceptable      :16995      acceptable      :15090   acceptable      :15686                  : 7179   acceptable      :15775   acceptable      :17233
excellent       :20786      excellent       :24916   excellent       :23960   acceptable     :17411   excellent       :24071   excellent       :25002
extremely poor  : 2038      extremely poor  :    1   extremely poor  :   12   excellent      :20396   extremely poor  :  322   good            :33822
good            :29373      good            :29042   good            :27993   extremely poor :    3   good            :27814   need improvement: 9301
need improvement:13527      need improvement:12063   need improvement:13896   good           :26373   need improvement:15156   poor            : 5559
poor            : 8198      poor            : 9805   poor            : 9370   need improvement:11018   poor            : 7779
                                                                              poor           : 8537
       Checkin_service          Cleanliness       Online_boarding
acceptable      :24941   acceptable      :16930   acceptable      :21427
excellent       :18918   excellent       :25079   excellent       :20993
extremely poor  :    1   extremely poor  :    4   extremely poor  :    9
good            :25483   good            :34246   good            :24676
need improvement:10813   need improvement: 9283   need improvement:13035
poor            :10761   poor            : 5375   poor            :10777
```

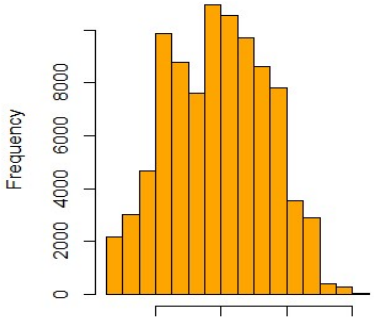
There are some blank values in variables: CustomerType,TypeTravel,Departure.Arrival.time_convenient and Food_drink.

3. Exploratory Data Analysis

We will start with trying to understand the data attributes, their spread and how they are related to each other. Since our objective is to find out which parameters influence the satisfaction level of flight passengers, we will assign satisfaction as the dependent variable.

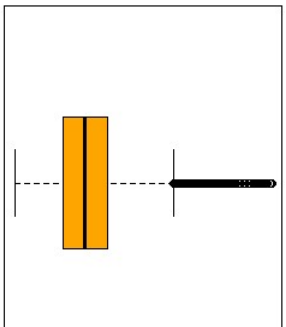
a. Univariate Analysis of continuous variables

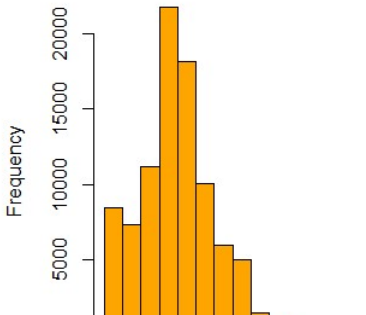




Age:

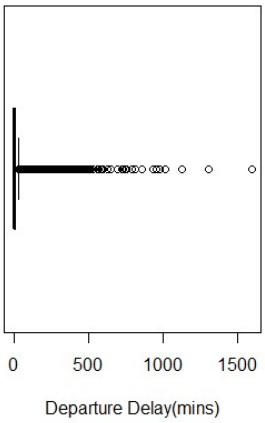
- There are no outliers.
- Distribution is slightly right skewed.
- Majority of the population are aged between 25 and 60.

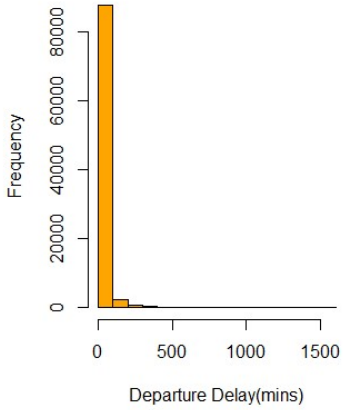




Flight Distance:

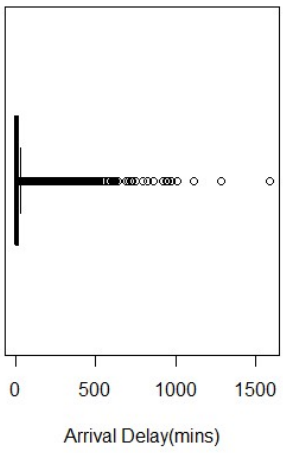
- There are lot of outliers, beyond 4000.
- Distribution is right skewed.
- Majority of the data seems to be of flight distances < 3000 kms, which probably are not long distance international flights.

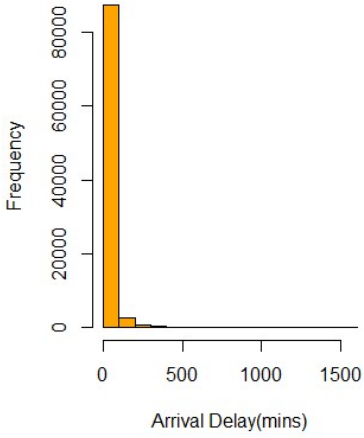




Departure Delay:

- Distribution has lot of outliers.
- Most of the flights have delays in Departure less than 30 mins.

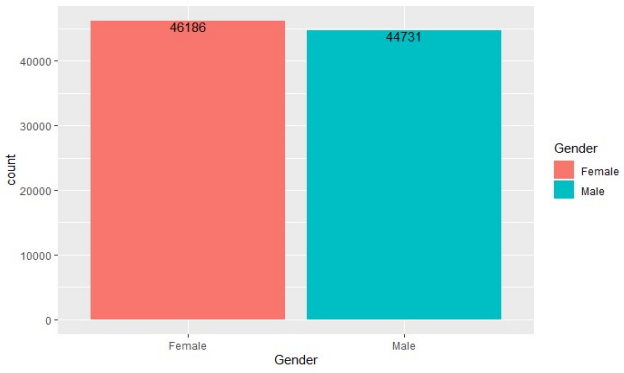




Arrival Delay:

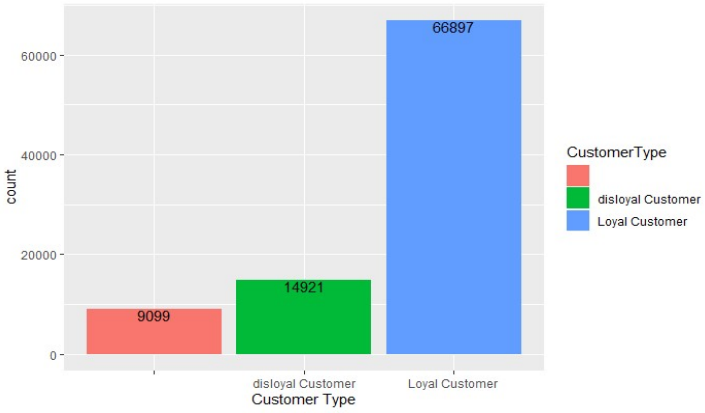
- Distribution has lot of outliers.
- Most of the flights have delays in Arrival less than 30 mins.

b. Univariate Analysis of Discrete variables



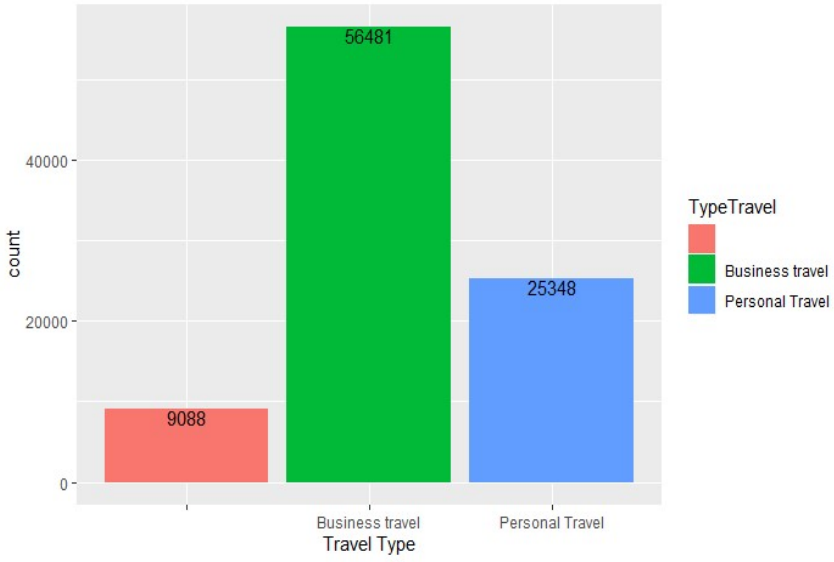
Gender:

Pretty balanced distribution with 46186 Female and 44731 Male.



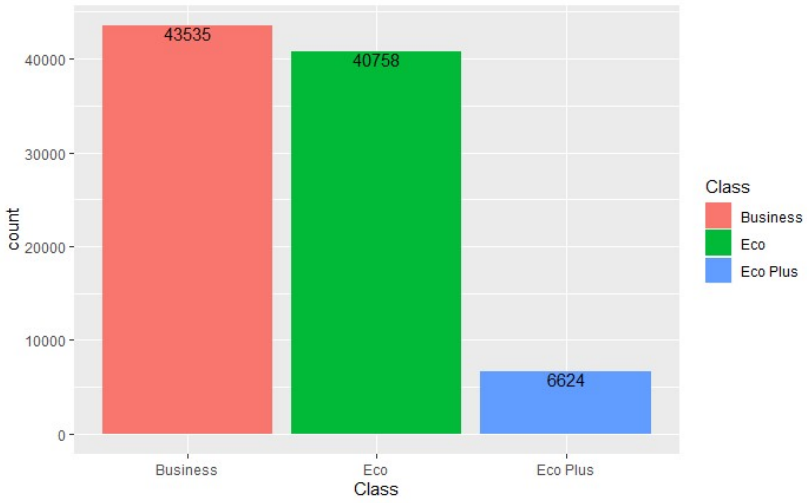
Customer Type:

Majority of the data comprises loyal customers. There are 9099 observations which doesn't have any relevant value. These should be classified as disloyal or loyal customers before model building.



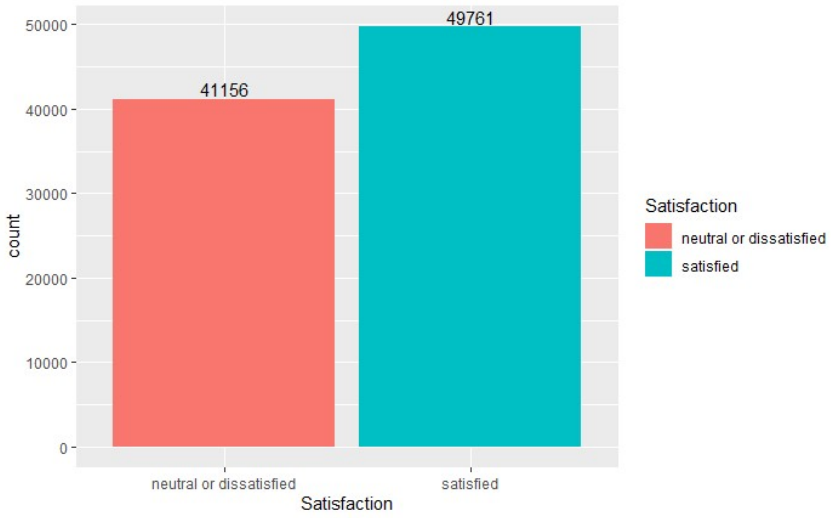
Travel Type:

Majority of the customers travel for business reasons. There are 9088 observations which doesn't have any relevant value. These should be classified as business travel or personal travel before model building.



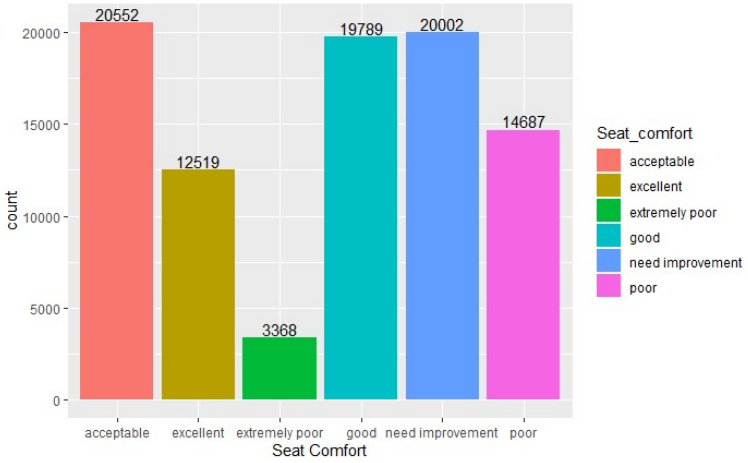
Class:

Most of the customers are business class passengers. But overall it's a balanced distribution between business class and economy class.



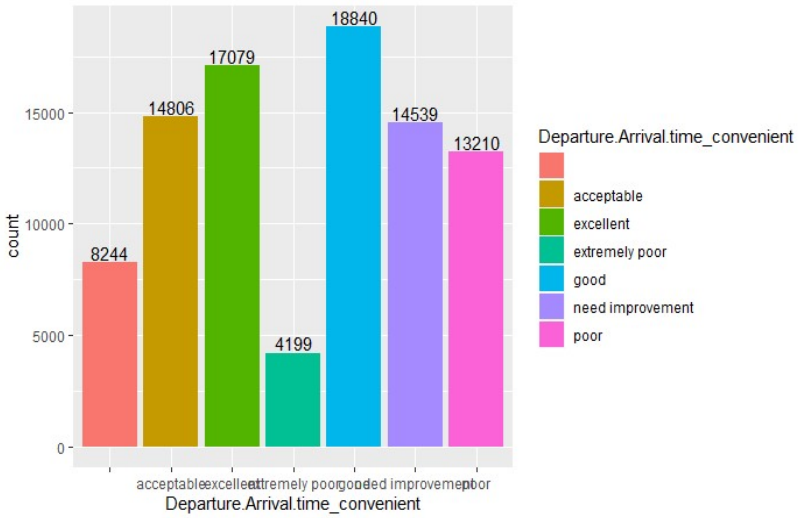
Satisfaction:

There is no imbalance in the dataset with respect to the dependent variable. Majority of the customers are satisfied customers.



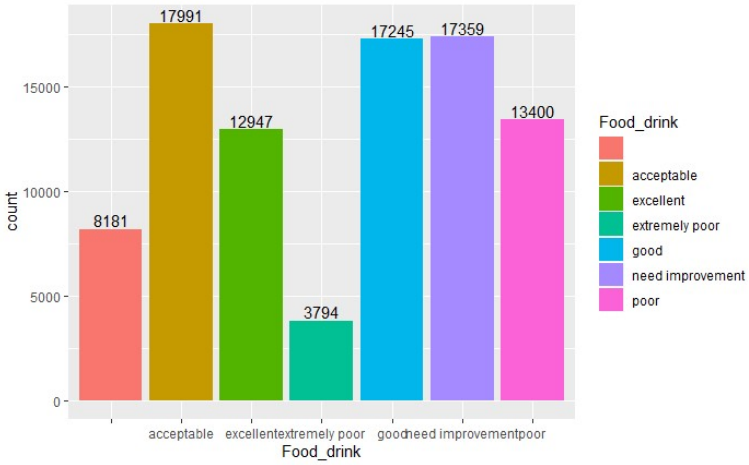
Seat Comfort:

There are no missing values and majority of the customers have given an average rating for seat comfort.



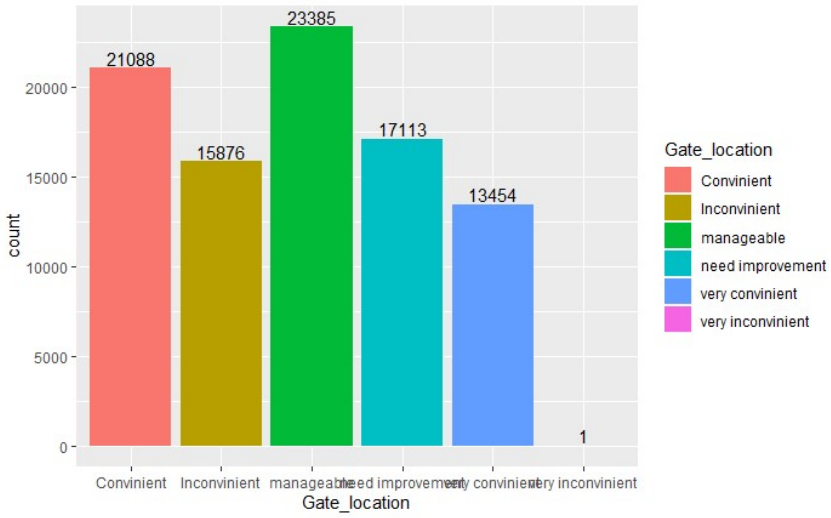
Departure Arrival Time Convenient:

There are 8244 observations which have no relevant values, these will need to be labelled in any of the other categories before model building.



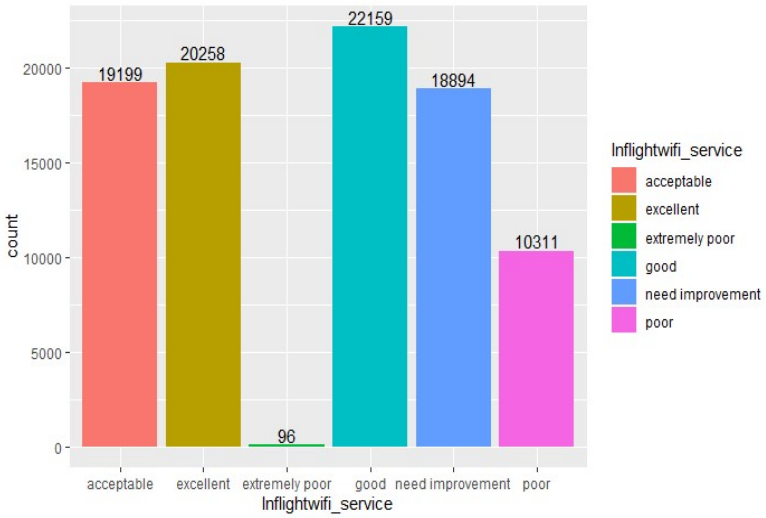
Food_Drink:

There are 8181 observations which have no relevant values, these will need to be labelled in any of the other categories before model building.



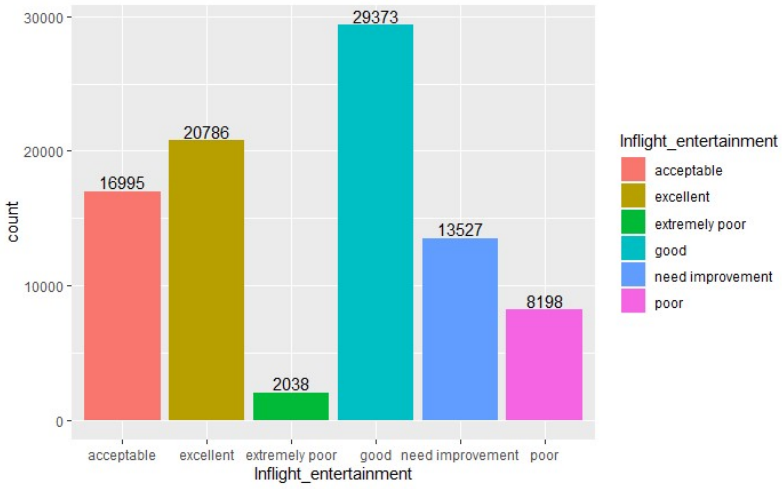
Gate Location:

There is 1 observation which is labelled as ‘very inconvenient’. This can be clubbed under ‘inconvenient’.



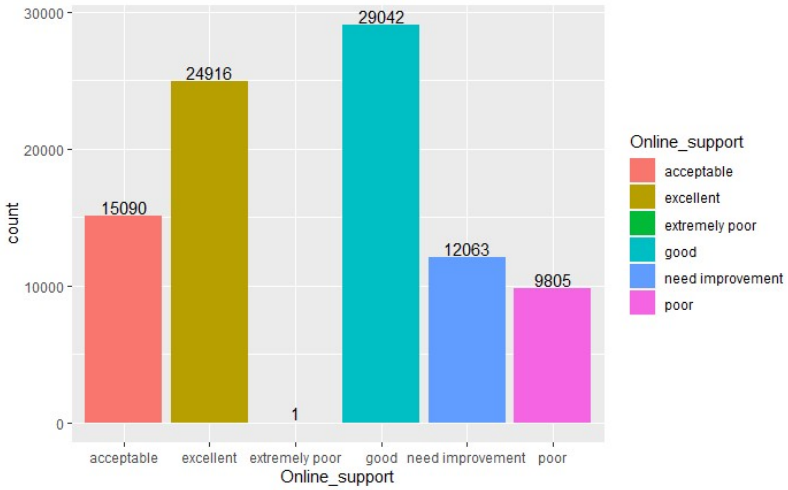
Inflight Wi-Fi Service:

There are 96 observations under the category ‘extremely poor’, these may also be clubbed under ‘poor’.



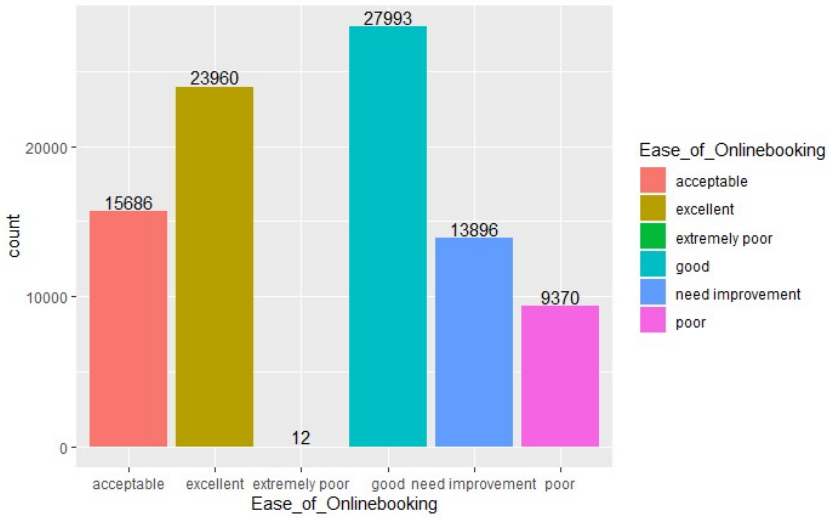
Inflight Entertainment

There are no missing values and majority of the customers have given a good rating towards inflight entertainment.



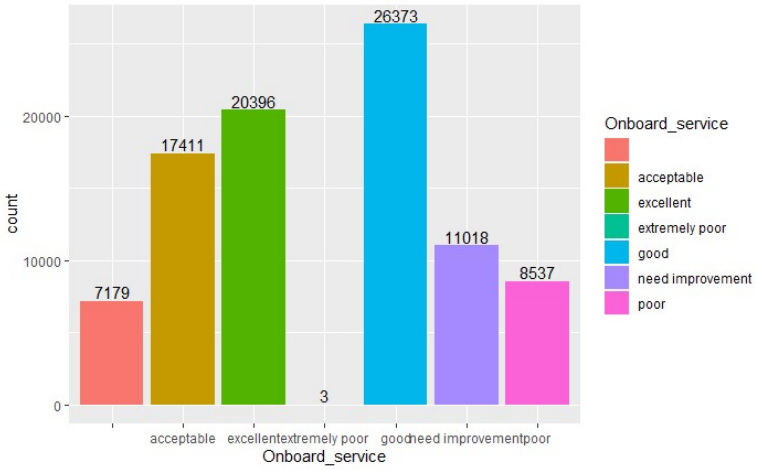
Online Support:

There is 1 observation under the category ‘extremely poor’, this may be clubbed under ‘poor’.



Ease of Online Booking:

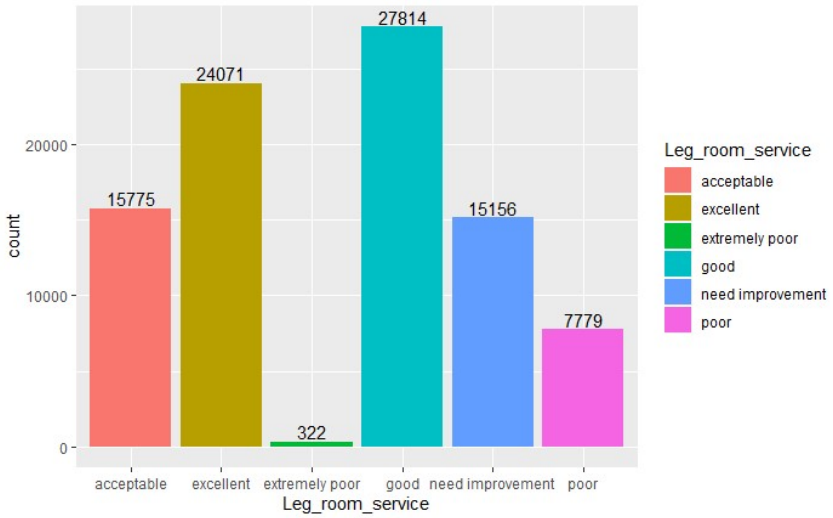
There are 12 observations under the category ‘extremely poor’, these may be clubbed under ‘poor’.



Onboard Service:

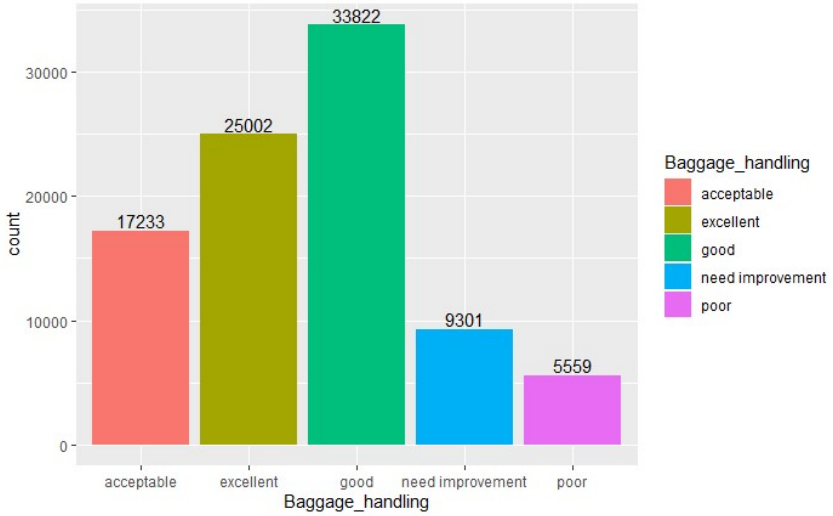
There are 3 observations under the category ‘extremely poor’, these may be clubbed under ‘poor’.

There are 7179 observations without any value, these have to be labelled before model building.



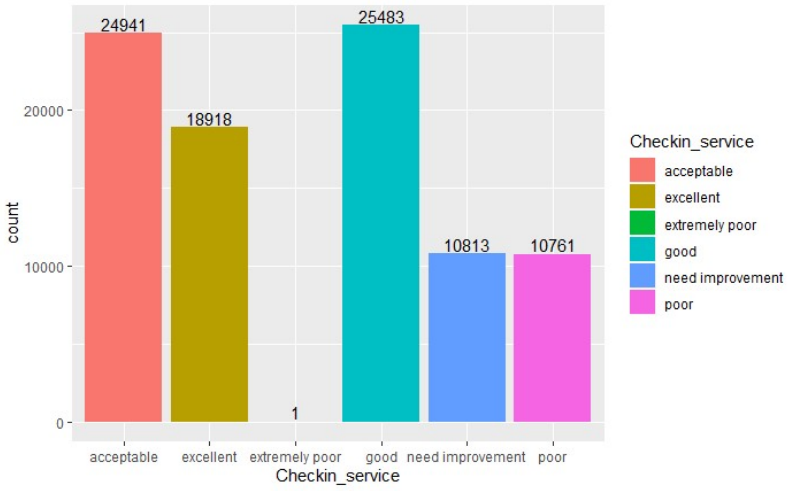
Leg Room Service:

There are 322 observations under the category ‘extremely poor’, these may be clubbed under ‘poor’.



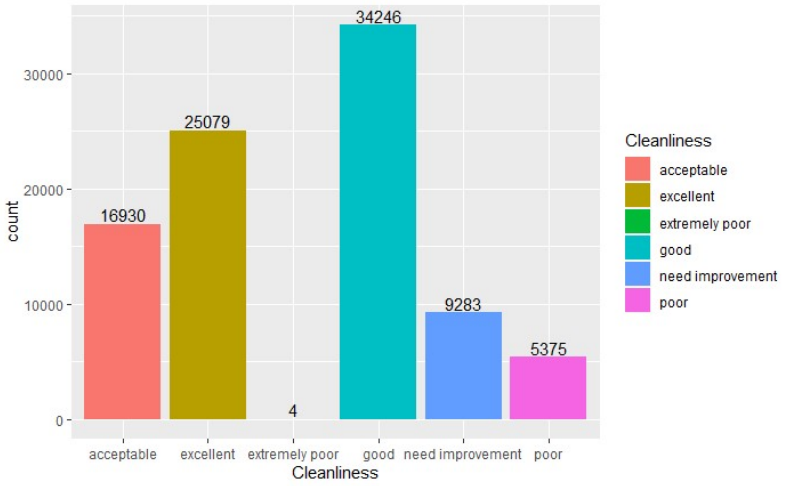
Baggage Handling:

There are no missing values and majority of the customers have given a good rating towards baggage handling.



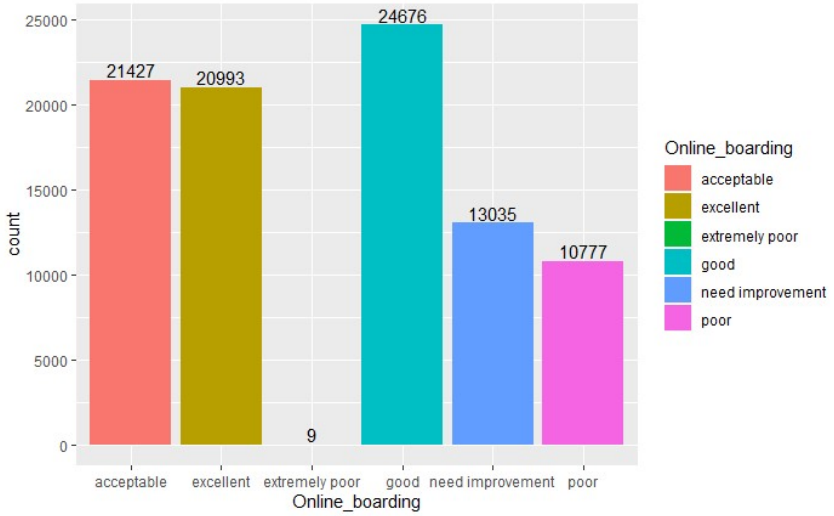
Check-in Service:

There is 1 observation under the category ‘extremely poor’, this may be clubbed under ‘poor’.



Cleanliness:

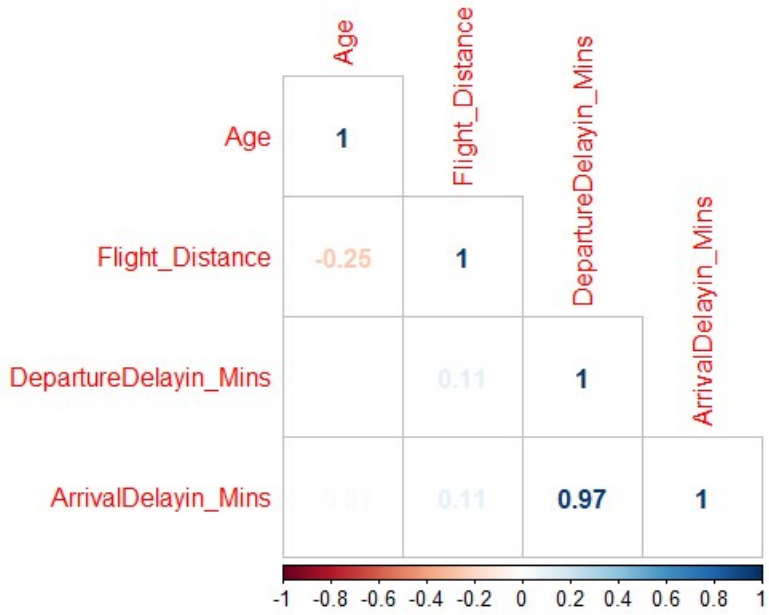
There are 4 observations under the category ‘extremely poor’, these may be clubbed under ‘poor’.



Online Boarding:

There are 9 observations under the category ‘extremely poor’, these may be clubbed under ‘poor’.

c. Bi-variate analysis between continuous variables



There is a strong positive correlation between Arrival Delays and Departure Delays, which signifies that flights which have arrived late also depart late for the onward journey.

There is a weak negative correlation between Age and Flight distance, which signifies that people who are older prefer shorter distance flights.

d. Chi-square tests between factor variables

Since Satisfaction is our dependent variable, we will assess the dependency of other factor variables on Satisfaction.

Our null hypothesis is (Ho) = Variables are independent

Alternative hypothesis is (Ha) = Variables are not independent

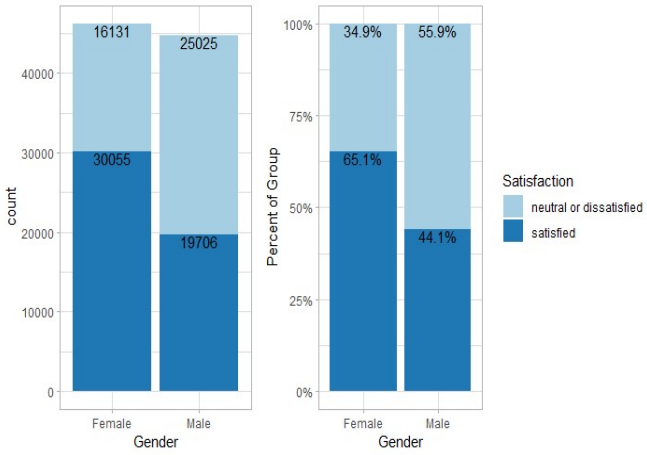
The below table illustrates the p-values from chi-square test run on each of the independent variables:

chisq.test(Aviation$Gender,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$CustomerType,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$TypeTravel,Aviation$Satisfaction)$p.value	7.140985e-220
chisq.test(Aviation$Class,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Seat_comfort,Aviation$Satisfaction)$p.value	0

chisq.test(Aviation$Departure.Arrival.time_convenient,Aviation$Satisfaction)$p.value	1.001355e-35
chisq.test(Aviation$Food_drink,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Gate_location,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Inflightwifi_service,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Inflight_entertainment,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Online_support,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Ease_of_Onlinebooking,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Onboard_service,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Leg_room_service,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Baggage_handling,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Checkin_service,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Cleanliness,Aviation$Satisfaction)$p.value	0
chisq.test(Aviation$Online_boarding,Aviation$Satisfaction)$p.value	0

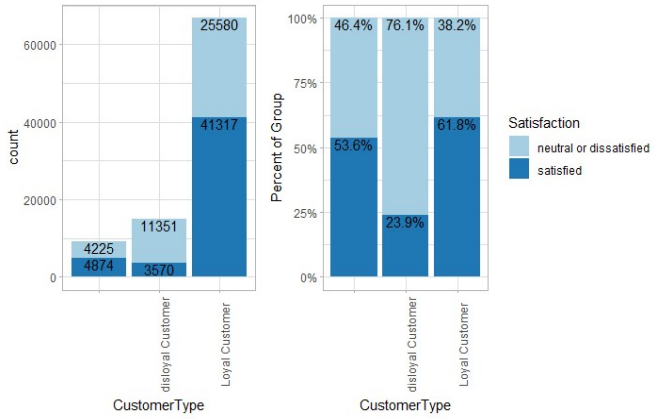
Results of chi-square test shows that all the factor variables are significant. We will plot the data and try get some more insights.

Corresponding bar plots are below:



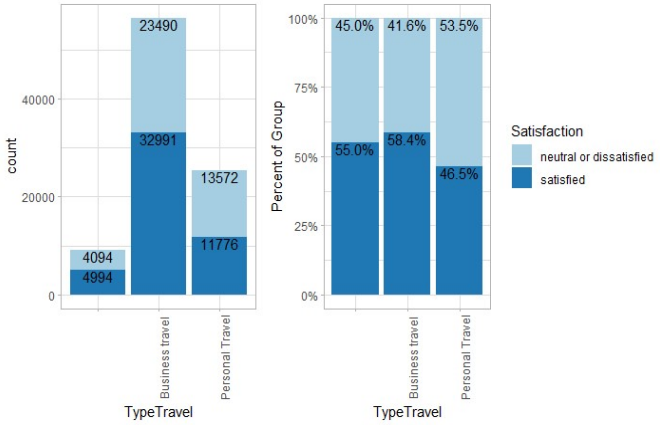
Gender:

Number of Female customers are slightly higher. More than half of the Male customers are showing to be dissatisfied compared to Female customers.



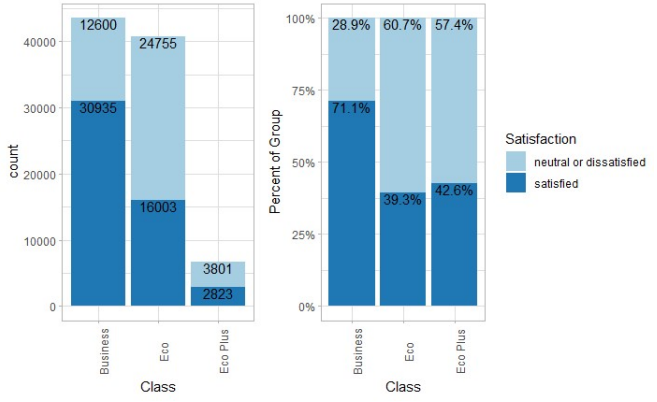
Customer Type:

Loyal customers are higher in numbers with ~40% of them being dissatisfied. ~76% Disloyal customers are dissatisfied. A lot of records are missing values.



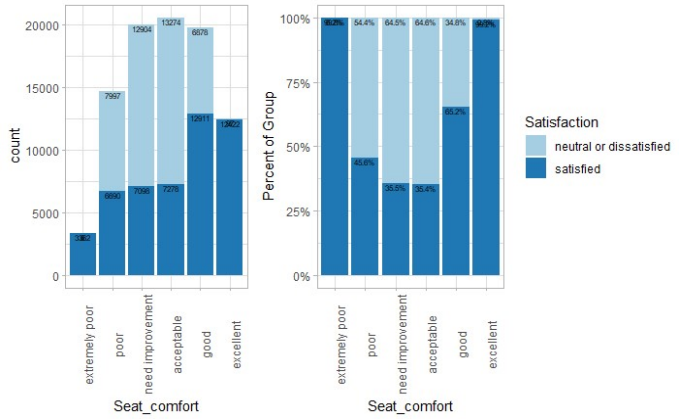
Type Travel:

Most of the customers travel for business. There are lot of missing values as well. The number of dissatisfied customers are almost equal for each type.



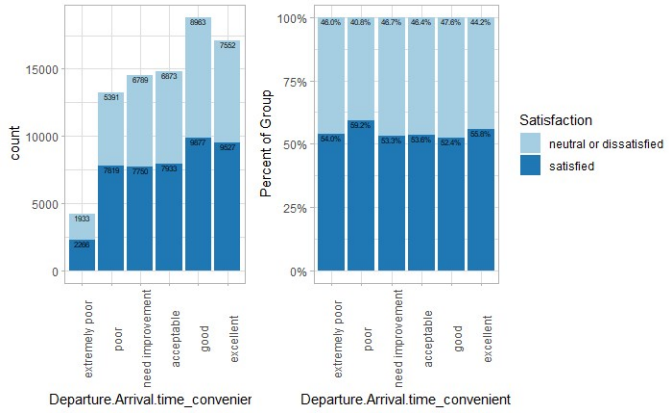
Class:

Most of the customers fly business class. Majority of the dissatisfied customers fly Economy and Economy plus.



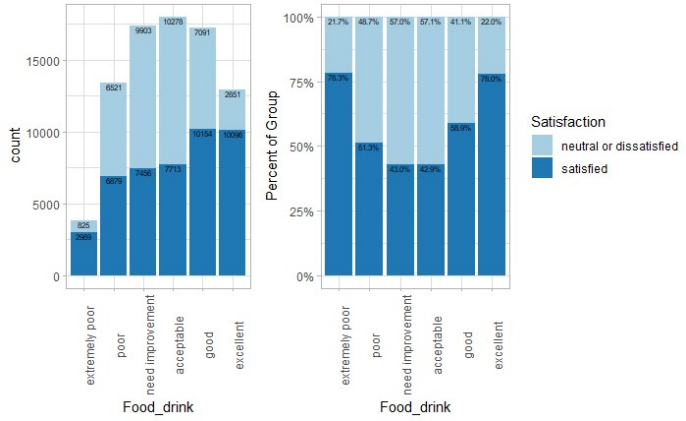
Seat Comfort:

Most customers have rated seat comfort as average. Customers who have rated it excellent are overall satisfied. There are exceptions as well where customers have rated extremely poor but are overall satisfied.



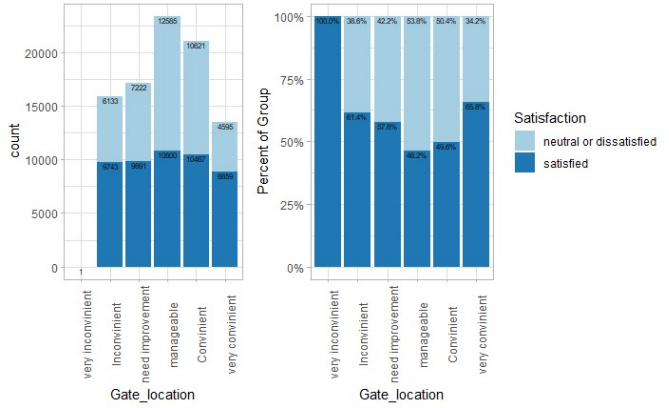
Departure and Arrival Time Convenience:

From the plot seems like Departure Arrival time convenience doesn't have much say on satisfaction.



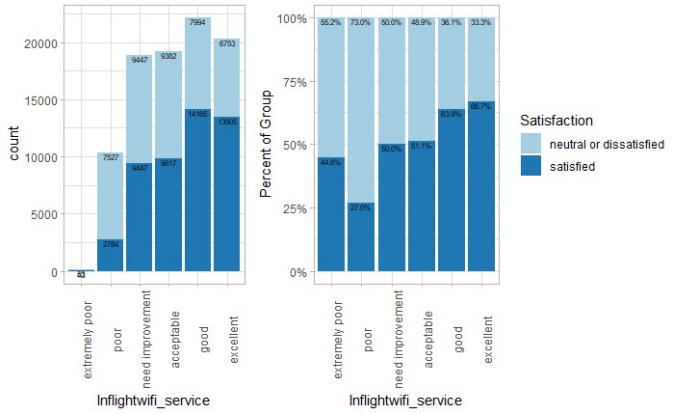
Food and Drink:

Most customers have rated food and drink as average and ones who have rated it as poor are dissatisfied. There are exceptions as well where customers have rated extremely poor but are overall satisfied.



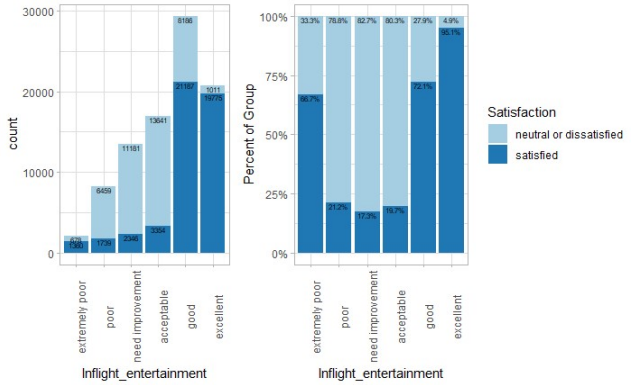
Gate Location:

Gate location largely doesn't show a clear trend towards satisfaction, except for one case where the gate location was extremely inconvenient and the customer was dissatisfied.



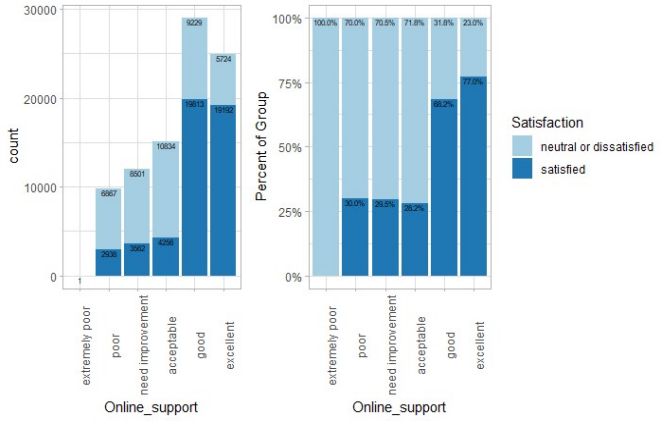
Inflight Wifi Service:

Majority have rated wifi service as good, very few have rated it as extremely poor. Customers who have rated wifi service as good are overall satisfied as well. Exceptions are those who have rated it extremely poor but are still overall satisfied.



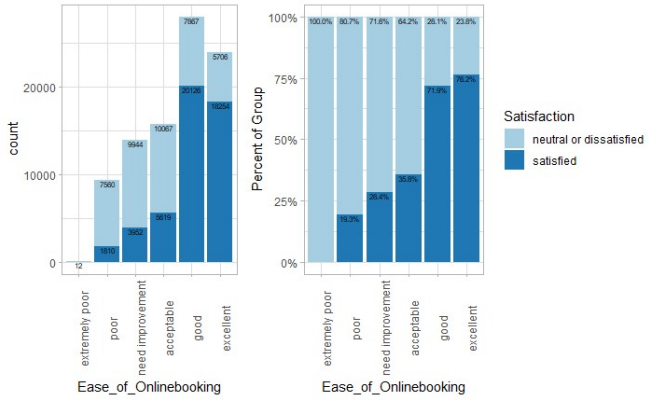
Inflight Entertainment:

Majority of the customers have rated inflight entertainment as good. Customers who have rated entertainment services as good are overall satisfied as well. Exceptions are those who have rated it extremely poor but are still overall satisfied.



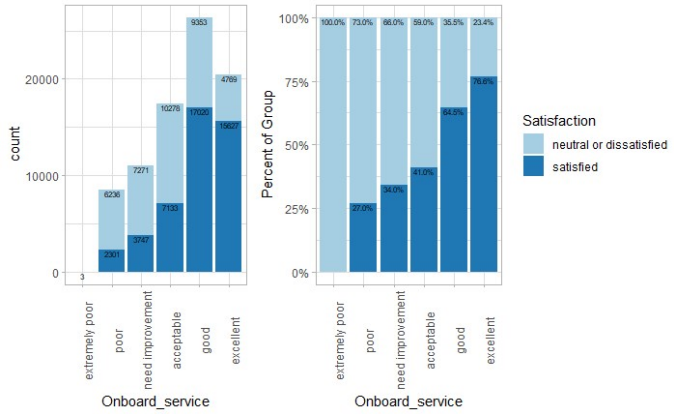
Online Support:

Most of the customers have rated online support as good/excellent. Customers are overall satisfied when customers have rated higher on this feature.



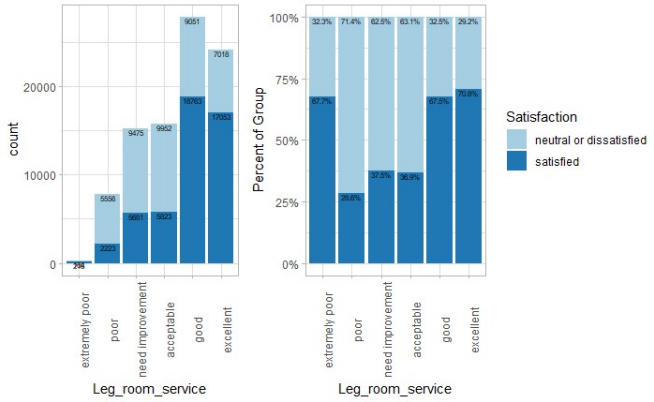
Ease of Online Booking:

Most of the customers have rated online booking as good/excellent. Customers are overall satisfied when customers have rated higher on this feature.



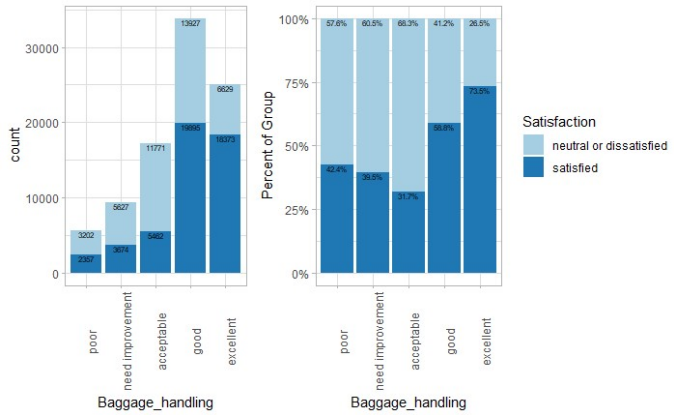
Onboard Service:

Majority of the customers have given a good rating towards on-board service and level of dissatisfaction decreases with an increase in the rating for on-board service.



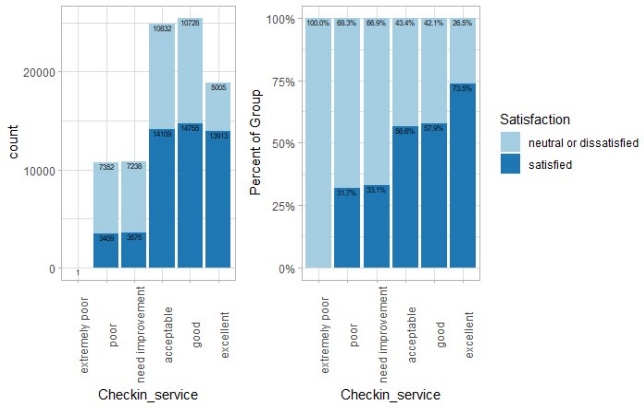
Leg Room Service:

Majority of the customers have given a good rating towards leg-room service and level of dissatisfaction decreases with an increase in the rating for leg-room service. There are very few exceptions where customers have rated extremely poor but are overall satisfied.



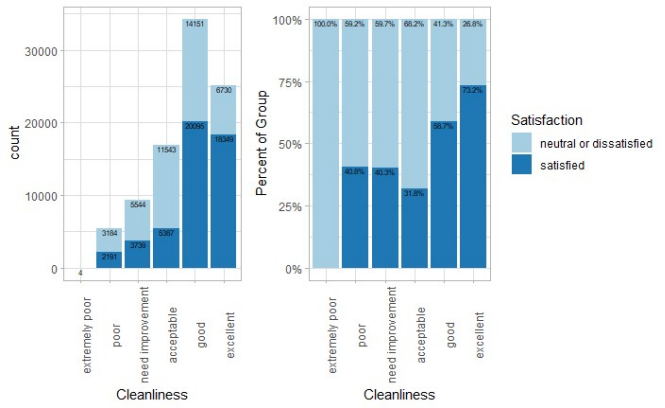
Baggage Handling:

Majority of the customers have given a good rating towards baggage handling. Customers who are not satisfied with baggage handling are dissatisfied customers.



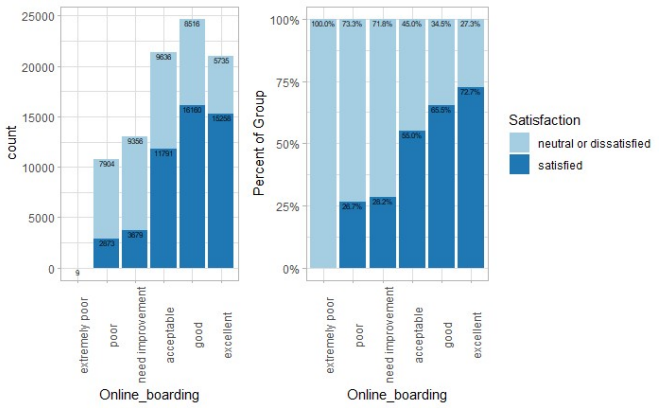
Check-in Service:

Majority of the customers have given a good rating towards check-in service and level of dissatisfaction decreases with an increase in the rating for check-in service.



Cleanliness:

Most of the customers have given a good rating towards cleanliness. However, dissatisfied customers have rated cleanliness poorly.



Online Boarding:

Majority of the customers have rated online boarding as higher than average standard. However dissatisfied customers have rated online boarding poorly.

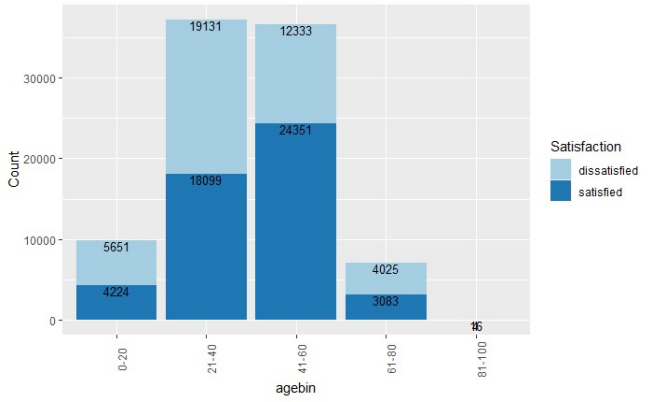
The above bar plots show that almost all the factor variables except Gender, Arrival Departure Time Convenience and Gate Location influence Satisfaction.

e. Relationship between dependent variable and continuous variable

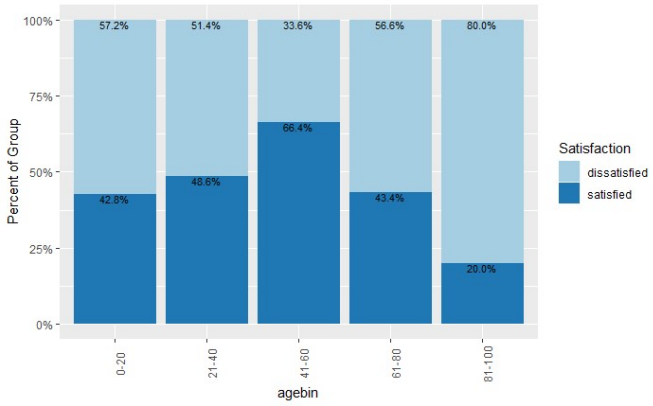
Logistic models were created for each continuous variable to determine whether the continuous variables are significant in determining Satisfaction.

Age:

```
Call:
glm(formula = Satisfaction ~ Age, family = "binomial", data = Aviation,
    na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5803  -1.2230   0.9735   1.0868   1.3155

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4292460  0.0187421  -22.90   <2e-16 ***
Age          0.0157623  0.0004474   35.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125222  on 90916  degrees of freedom
Residual deviance: 123961  on 90915  degrees of freedom
AIC: 123965

Number of Fisher Scoring iterations: 4
```



Majority of the customers are between the age bracket of 21-60. Highest number of dissatisfied customers are in the 21-40 age bracket.



Percentage of dissatisfaction is higher than satisfaction in all age brackets except 41-60.

The p-values of the following variables show that they are significant, but the pseudo R2 values show that each of them individually don't explain the variance too much.

Flight_Distance:

```
Call:
glm(formula = Satisfaction ~ Flight_Distance, family = "binomial",
    data = Aviation, na.action = na.omit)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.325  -1.257   1.047   1.098   1.262

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.448e-01  1.454e-02   23.72   <2e-16 ***
Flight_Distance -7.802e-05  6.495e-06  -12.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125222  on 90916  degrees of freedom
Residual deviance: 125078  on 90915  degrees of freedom
AIC: 125082

Number of Fisher Scoring iterations: 3
```

Departure Delay in Mins:

```
Call:
glm(formula = Satisfaction ~ DepartureDelayin_Mins, family = "binomial",
    data = Aviation, na.action = na.omit)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.285  -1.284   1.073   1.073   3.217

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.2502336  0.0072217   34.65   <2e-16 ***
DepartureDelayin_Mins -0.0041538  0.0001943  -21.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125222  on 90916  degrees of freedom
Residual deviance: 124714  on 90915  degrees of freedom
AIC: 124718

Number of Fisher Scoring iterations: 3
```

Arrival Delay in Mins:

```
Call:
glm(formula = Satisfaction ~ ArrivalDelayin_Mins, family = "binomial",
    data = Aviation, na.action = na.omit)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.288  -1.282   1.071   1.071   3.311

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.2567529  0.0072574   35.38   <2e-16 ***
ArrivalDelayin_Mins -0.0044789  0.0001948  -22.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124828  on 90632  degrees of freedom
Residual deviance: 124236  on 90631  degrees of freedom
  (284 observations deleted due to missingness)
AIC: 124240

Number of Fisher Scoring iterations: 3
```

The McFadden R2 values for each of the models are as below:

pR2(logistic.model.Age)["McFadden"]	.0100687

pR2(logistic.model.FlightDistance)["McFadden"]	.001154021
pR2(logistic.model.ArrivalDelayin_Mins)["McFadden"]	.007871765
pR2(logistic.model.DepartureDelayin_Mins)["McFadden"]	.004054502

4. Data Pre-processing

a. Data Imputation:

In the original data there were lot of observations with missing values. The findings from initial data exploration showed the following attributes had missing data:

CustomerType: 9099 observations

TypeTravel: 9088 observations

Departure.Arrival.time_convenient: 8244 observations

Food_drink: 8181 observations

ArrivalDelayin_Mins: 284 observations

Onboard_service: 7179 observations

Through further analysis it was found that these missing values don't belong to a specific set of customers. If we had to get rid of all these observations with missing values we will end up losing lot of data. Hence Mice package was used for data imputation.

For imputing categorical variables, 'polyreg' method was used and for numeric variable, 'pmm' method was used.

b. Variable Transformation:

Levels for the variable Gate_Location were changed to make it uniform with other factor variables. Following mapping was used: Convinient → good; Inconvinient → poor; manageable → acceptable; very convenient → excellent; very inconvenient → extremely poor

For the variable Satisfaction level name of neutral or dissatisfied was changed to dissatisfied.

c. Variable Creation:

New variable agebin was created with the break up as 0-20, 21-40, 41-60, 61-80 and 81-100.

New variable cluster id was created after customer segmentation with values of '1','2','3','4' and 'others'.

d. Summary of Cleansed data:

```
       ID          ArrivalDelayin_Mins    Gender                CustomerType        Age                TypeTravel         Class       Flight_Distance
 Min.   :149965   Min.   :   0.00     Female:46186   disloyal Customer:16650   Min.   : 7.00   Business travel:62766   Business:43535   Min.   :  50
 1st Qu.:172694   1st Qu.:   0.00     Male  :44731   Loyal Customer   :74267   1st Qu.:27.00   Personal Travel:28151   Eco     :40758   1st Qu.:1360
 Median :195423   Median :   0.00                                             Median :40.00                           Eco Plus: 6624   Median :1927
 Mean   :195423   Mean   :  15.13                                             Mean   :39.45                                            Mean   :1982
 3rd Qu.:218152   3rd Qu.:  13.00                                             3rd Qu.:51.00                                            3rd Qu.:2542
 Max.   :240881   Max.   :1584.00                                             Max.   :85.00                                            Max.   :6950

 DepartureDelayin_Mins       Satisfaction            Seat_comfort   Departure.Arrival.time_convenient            Food_drink             Gate_location
 Min.   :   0.00       dissatisfied:41156   acceptable      :20552   acceptable      :16229            acceptable      :19734   good            :21088
 1st Qu.:   0.00       satisfied   :49761   excellent       :12519   excellent       :18839            excellent       :14223   poor            :15876
 Median :   0.00                            extremely poor  : 3368   extremely poor  : 4624            extremely poor  : 4173   acceptable      :23385
 Mean   :  14.69                            good            :19789   good            :20591            good            :18945   need improvement:17113
 3rd Qu.:  12.00                            need improvement:20002   need improvement:16048            need improvement:18990   excellent       :13454
 Max.   :1592.00                            poor            :14687   poor            :14586            poor            :14852   extremely poor  :    1

 Inflightwifi_service      Inflight_entertainment         Online_support      Ease_of_Onlinebooking        Onboard_service
 acceptable      :19199    acceptable      :16995    acceptable      :15090   acceptable      :15686    acceptable      :18865
 excellent       :20258    excellent       :20786    excellent       :24916   excellent       :23960    excellent       :22143
 extremely poor  :   96    extremely poor  : 2038    extremely poor  :    1   extremely poor  :   12    extremely poor  :   17
 good            :22159    good            :29373    good            :29042   good            :27993    good            :28601
 need improvement:18894    need improvement:13527    need improvement:12063   need improvement:13896    need improvement:11939
 poor            :10311    poor            : 8198    poor            : 9805   poor            : 9370    poor            : 9352

 Leg_room_service          Baggage_handling         Checkin_service            Cleanliness          Online_boarding     agebin
 acceptable      :15775    acceptable      :17233   acceptable      :24941   acceptable      :16930   acceptable      :21427   0-20  : 9875
 excellent       :24071    excellent       :25002   excellent       :18918   excellent       :25079   excellent       :20993   21-40 :37230
 extremely poor  :  322    good            :33822   extremely poor  :    1   extremely poor  :    4   extremely poor  :    9   41-60 :36684
 good            :27814    need improvement: 9301   good            :25483   good            :34246   good            :24676   61-80 : 7108
 need improvement:15156    poor            : 5559   need improvement:10813   need improvement: 9283   need improvement:13035   81-100:   20
 poor            : 7779                             poor            :10761   poor            : 5375   poor            :10777
```

Summary above confirms that there are no more missing values.

e. Training and Test data:

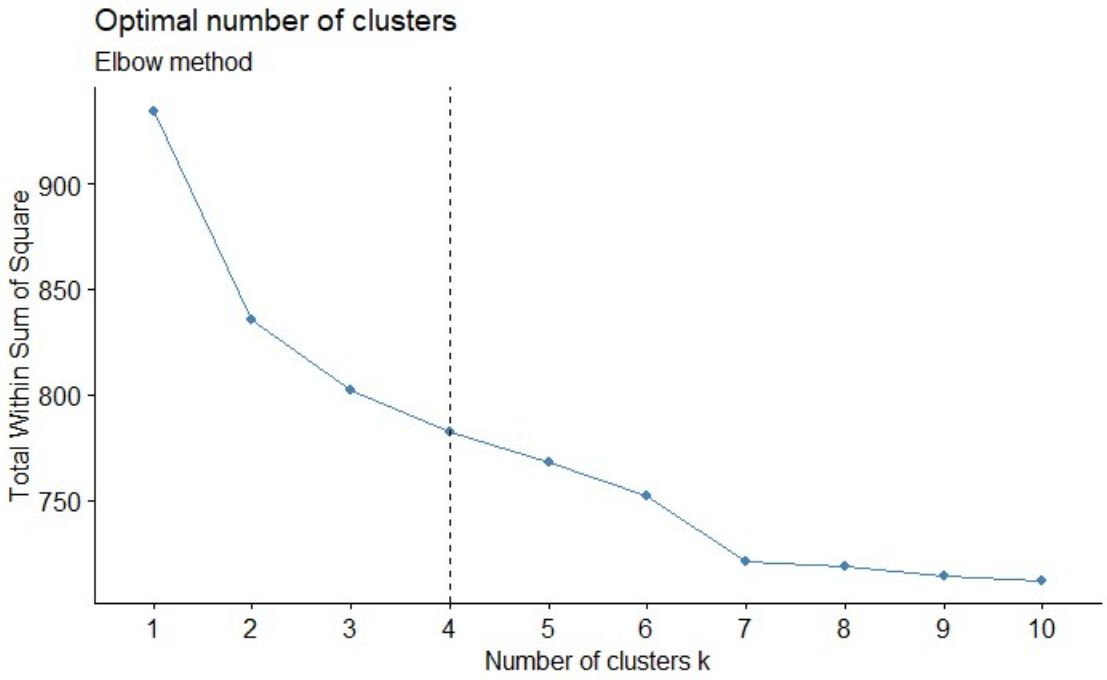
Training and Test data was created by doing a 70-30 split and that will be used for model building.

5. Customer Segmentation:

Clustering was done on passenger attributes like Age, Gender, Customer Type, Travel Type and Class. Segmentation might help us understand some pattern of the type of passengers in the survey. Using the segmentation data we can further drill down on how each segment of passenger has rated the amenities and the overall satisfaction levels.

Since the data has both numeric and categorical features, Gower distance was used to create distance matrix and then clustering.

Number of clusters were determined to be 4.



Cluster profiles created:

Cluster	Gender	Customer Type	Age	Type Travel	Class
1	Female: 0 Male :2485	disloyal Customer: 303 Loyal Customer :2182	Min. : 7.00 1st Qu.:34.00 Median :44.00 Mean :42.95 3rd Qu.:52.00 Max. :85.00	Business travel:2431 Personal Travel: 54	Business:2101 Eco : 241 Eco Plus: 143
2	Female:1249 Male :1509	disloyal Customer: 10 Loyal Customer :2748	Min. : 7.00 1st Qu.:22.00 Median :37.00 Mean :37.43 3rd Qu.:52.00 Max. :74.00	Business travel: 143 Personal Travel:2615	Business: 52 Eco :2464 Eco Plus: 242

3	Female:2547 Male : 0	disloyal Customer: 140 Loyal Customer :2407	Min. : 7.00 1st Qu.:36.00 Median :44.00 Mean :43.96 3rd Qu.:53.00 Max. :80.00	Business travel:2375 Personal Travel: 172	Business:2009 Eco : 324 Eco Plus: 214
4	Female:852 Male :449	disloyal Customer:1209 Loyal Customer : 92	Min. : 7.00 1st Qu.:22.00 Median :25.00 Mean :27.91 3rd Qu.:33.00 Max. :79.00	Business travel:1296 Personal Travel: 5	Business: 180 Eco :1061 Eco Plus: 60

Description of the above Clusters:

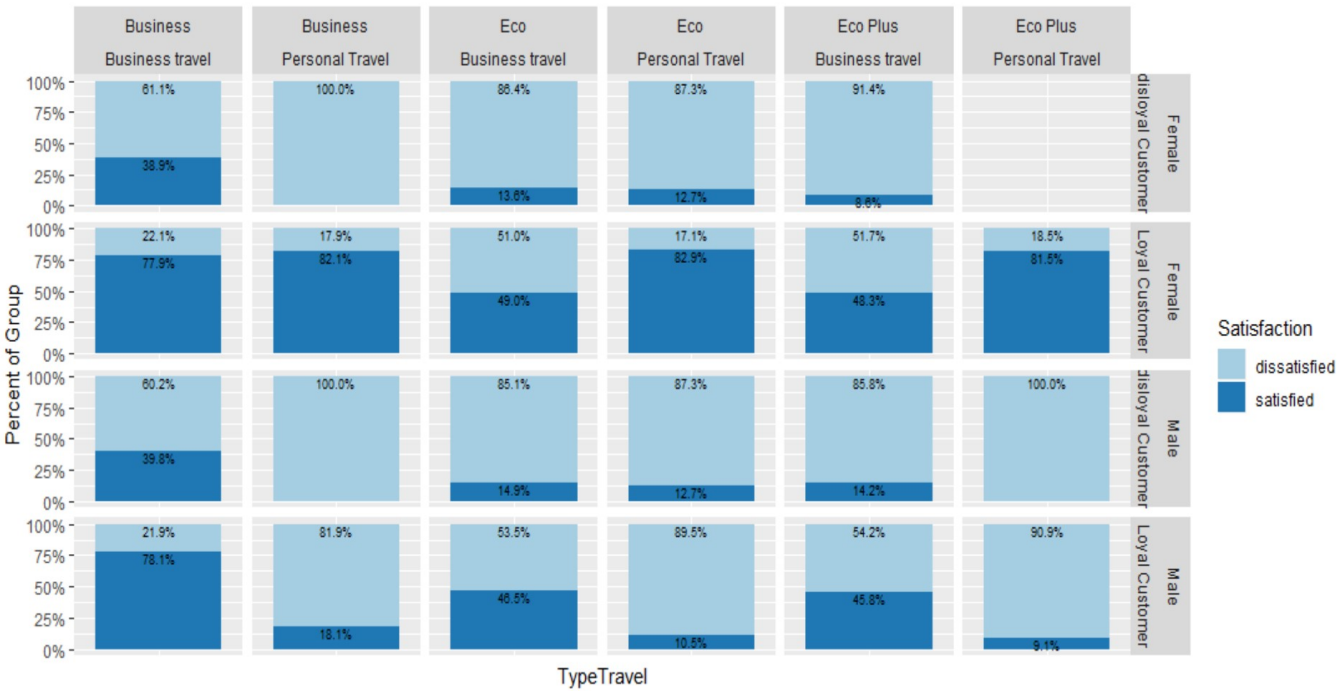
Cluster 1: Male, Loyal customers flying Business class for Business trips.

Cluster 2: Loyal customers flying Economy class for Personal trips.

Cluster 3: Female, Loyal customers flying Business class for Business trips.

Cluster 4: Disloyal customers flying Economy class for Business trips.

The figure below will help us understand how each section of the segments derived has rated against Satisfaction.



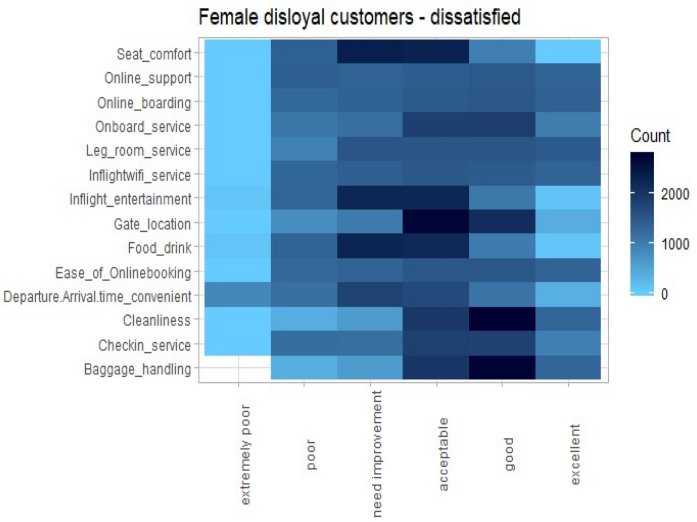
Key Observations:

1. Female Loyal customers are more Satisfied passengers except when they are flying Economy for business. Male Loyal customers are comparatively more dissatisfied.

2. Male/Female disloyal customers flying for personal reasons are always dissatisfied. The numbers of customers in these categories are very less.

3. Loyal customers are Satisfied when they are flying business class for business trips. Disloyal customers still show dissatisfaction in Business class and business trips.

In an effort to understand the trend of ratings received across all flight features by various segment of customers some heatmaps were plotted.

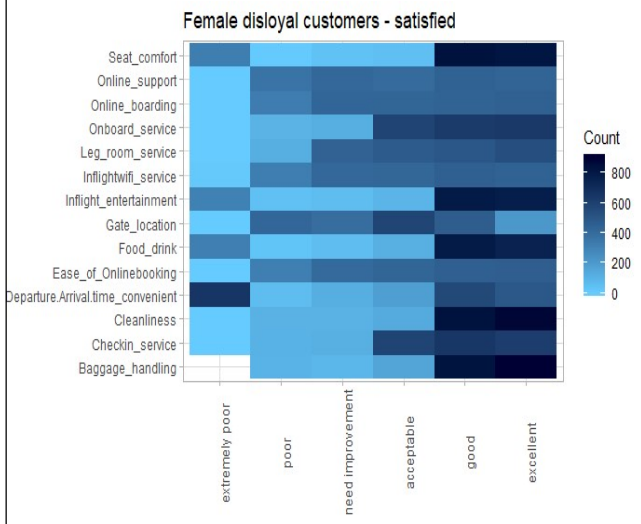


Female Disloyal customers – dissatisfied:

Female disloyal customers have found Seat Comfort, Food_Drink and Inflight Entertainment average.

Onboard Service, Online Boarding, Cleanliness and Baggage Handling are also rated average.

Lot of disloyal customers have found the flight timings to be inconvenient and have rated extremely poor.



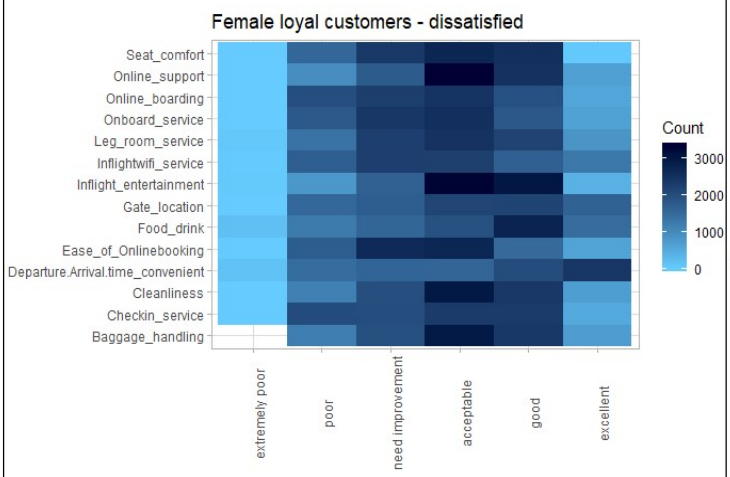
Female Disloyal customers – satisfied:

Although the numbers are much less but customers in this segment have found Seat Comfort, Inflight Entertainment, Food/Drink, Cleanliness and Baggage Handling to be good or excellent. Hence they are overall satisfied.

Lot of disloyal customers have found the flight timings to be inconvenient and have rated extremely poor.

Key differentiators:

Seat Comfort, Inflight Entertainment, Food_Drink, Cleanliness and Baggage Handling



Female loyal customers – dissatisfied:

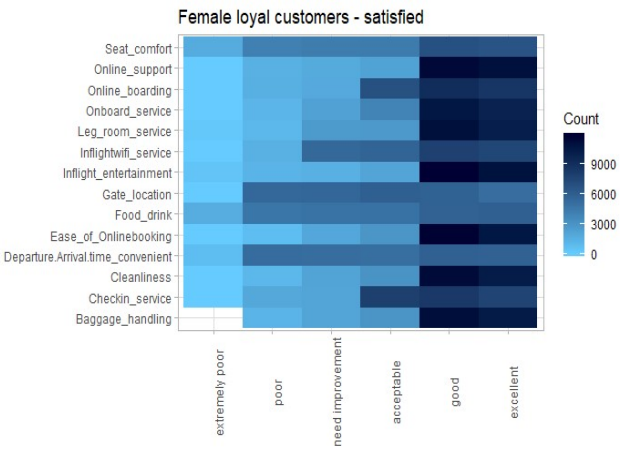
Female loyal customers have found food_drink and inflight entertainment good.

Onboard service, online support, seat comfort, cleanliness and baggage handling are also rated acceptable.

Online boarding and check-in service have been rated poor.

Loyal customers have found the flight timings to be convenient and have rated excellent.

However in this segment of customers lot have found ease of online booking to be wanting and feel there is scope of improvement.

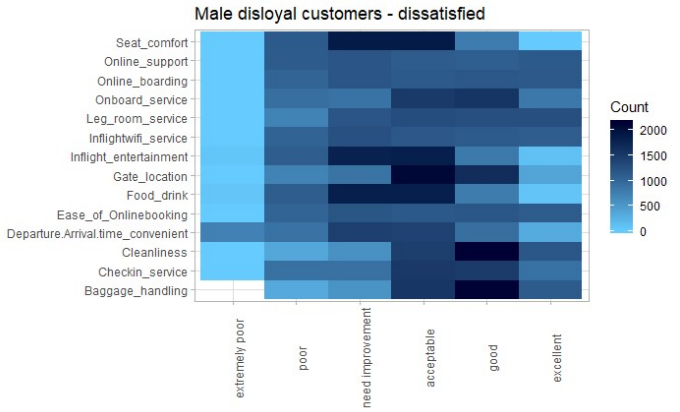


Female loyal customers – satisfied:

Number of female loyal satisfied customers are high. They have rated online support, leg room service, onboard service,inflight entertainment, ease of online booking, cleanliness and baggage handling as good/excellent. Hence they are overall satisfied.

Key differentiators:

Online Support, Leg room service, Onboard Service, Ease of Online Booking, Inflight Entertainment, Cleanliness and Baggage Handling



Male disloyal customers – dissatisfied:

Male disloyal customers have found Seat Comfort, Inflight Entertainment and Food/Drink to be average. Time convenience has also been rated poor.

Gate Location has been rated acceptable by most.

They have rated Cleanliness and Baggage Handling good as well but overall shown dissatisfaction.

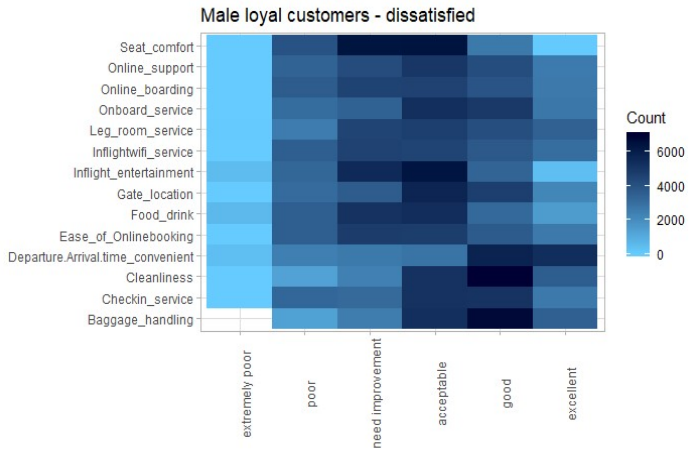
Male disloyal customers – satisfied:

Numbers are low in this segment. But the satisfied customers have rated Seat Comfort, Inflight Entertainment, Food/Drink, Cleanliness, Baggage Handling as good/excellent.

Departure Arrival Time Convenience has been rated extremely poor.

Key differentiators:

Seat Comfort, Inflight Entertainment, Food and Drink, Cleanliness and Baggage Handling.



Male loyal customers – dissatisfied:

Male loyal customers have found Seat comfort, Onboard service, Legroom service,Inflight entertainment and food/drink to be average. Gate location has been rated acceptable by most.

They have rated Cleanliness and Baggage handling good but overall shown dissatisfaction.

Male loyal customers – satisfied:

The satisfied customers have rated Online Support, Inflight entertainment, Ease of online booking, food/drink, Cleanliness, Baggage handling as good/excellent.

Onboard service and Legroom service has been rated good by most of the customers.

Key differentiators:

Seat Comfort, Online Support, Onboard service, Legroom service, Inflight Entertainment, Ease of Online Booking, Cleanliness and Baggage Handling.

Key Observations:

- Heatmap analysis of the ratings given by disloyal customers reveal that Seat comfort, Inflight Entertainment, Food and Drink, Cleanliness and Baggage Handling have been the primary factors that caused dissatisfaction for both male and female passengers. For many the time of flight departure and arrival was also very inconvenient.

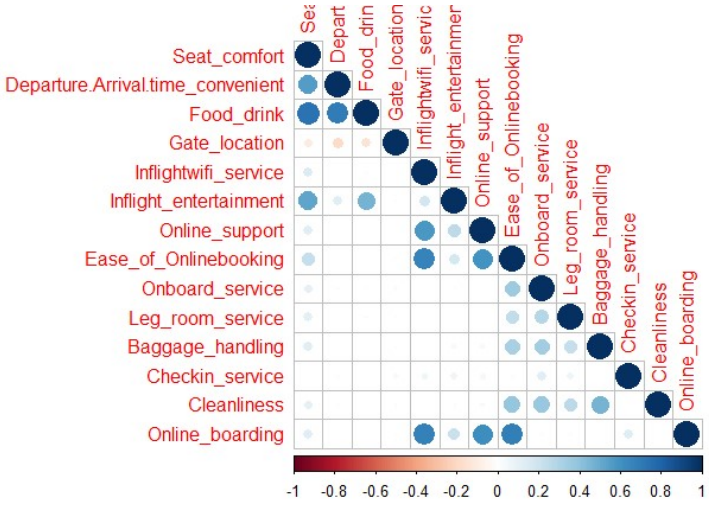
- For loyal customers Online Support, Inflight Entertainment, Onboard service, Legroom service ,Ease of Online Booking, Cleanliness and Baggage Handling have been the primary factors that contributed towards dissatisfaction for both male and female passengers.

Based on the heatmap analysis, the immediate recommendation for Falcon Airlines will be to upgrade their service towards Seat Comfort, Inflight Entertainment, Ease of Online Booking, Online Support,Cleanliness and Baggage Handling across all customer segments so as to deliver a better flight experience.

6. Factor Analysis:

An alternate way of analyzing the features and their impact on satisfaction is Factor analysis. We can reduce the number of features or condense them into categories to help Falcon airlines assess them as a whole. The feature survey data is ordinal data and are factor variables. Usual Correlation functions work best on continuous or numeric variables. Here we will apply Polychoric correlation as the data is ordinal.

Polychoric correlation of factor variables below shows some collinearity in data. Sphericity check and KMO test indicates that dimension reduction is possible for the data.



Bartlett sphericity check:

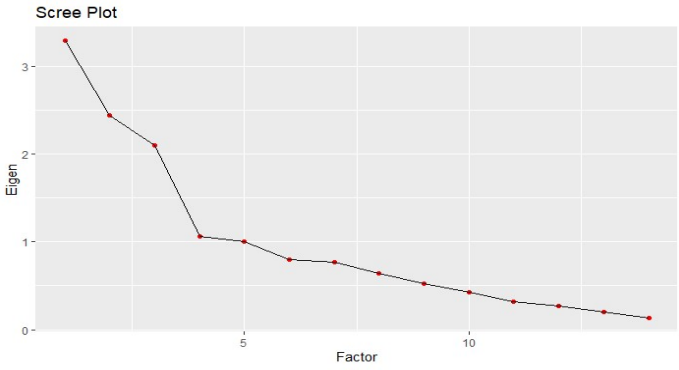
cortest.bartlett(items_polychoric)$p.value: 2.444919e-61

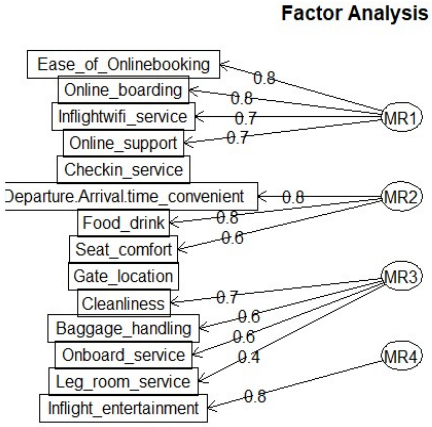
A very low p- value signifies that dimension reduction is possible.

KMO Test:

Kaiser-Meyer-Olkin factor adequacy: 0.71

Scree plot shows that 4 factors could be derived from the data:





Factors derived through varimax rotation:

MR1: Passenger digital amenities

MR2: Passenger convenience

MR3: Passenger onboard services

MR4: Inflight entertainment

7. Model Building:

Model Selection:

Since this is a classification problem, three algorithms were tried – Logistic Regression, Random Forest and XGBoost.

KNN was excluded because dataset chiefly consists of categorical variables. Naïve Bayes was excluded because there is collinearity between some of the factor variables, which violates the primary assumption of Naïve Bayes algorithm.

Data Preparation:

Clustering done earlier using pam method, was executed on only 10% of the dataset because of computational limitations. The clusters/patterns identified was generalized and 80% observations of the dataset was then assigned to the identified 4 clusters. The remaining 20% records were tagged as others or the 5th cluster.

Apart from a general split of 70-30 ratio between Training and Test data, smaller splits of Training and Test datasets were created for each cluster. The models were executed on the same split of Training and Test data. The smaller splits of data for each cluster was used to identify the variable importance for each segment of customer.

a. Logistic Regression

aviation.logistic=glm(Satisfaction~.,data = TrainingData[-c(1)],family = 'binomial')

ID field was excluded for model building. The model summary showed all parameters to be significant.

Confusion matrix on training data:

	Predicted dissatisfied	Predicted satisfied
Actual dissatisfied	26130	2677
Actual satisfied	3021	31813

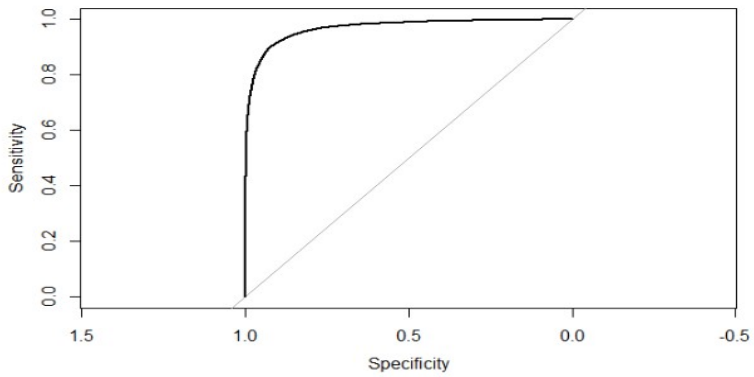
Confusion matrix on test data:

	Predicted dissatisfied	Predicted satisfied
Actual dissatisfied	11207	1142
Actual satisfied	1316	13611

Stats derived from confusion matrix

	Training Data	Test Data
TPR(Sensitivity)	.913	.911
TNR(Specificity)	.907	.907
Accuracy	.910	.909

ROC Curve and other stats derived from Test data:



Area under the curve: 0.968

Variable importance:

Variable Importance study of the Logistic regression model showed the following order of features towards determining satisfaction:

Inflight entertainment, Seat Comfort,Gender,Departure Arrival Time convenience, Online Support, Ease of Online booking, Cleanliness, Baggage Handling, Legroom service, Checkin service, Onboard service.

b. Random Forest

Parameters used:

finalfor = randomForest(Satisfaction~. Data = TrainingData[-c(1)], ntree=501,mtry=6,nodesize=10,importance=T)

Model tuning methods determined the ntree value to be 501 and mtry value to be 6. ID field was disregarded for model building.

```
Call:
 randomForest(formula = Satisfaction ~ ., data = TrainingData[-c(1)],      ntree = 501,
mtry = 6, nodesize = 10, importance = T)
               Type of random forest: classification
                     Number of trees: 501
No. of variables tried at each split: 6

        OOB estimate of  error rate: 4.66%
Confusion matrix:
             dissatisfied satisfied class.error
dissatisfied        27641      1166  0.04047627
satisfied            1800     33034  0.05167365
```

Confusion matrix on training data:

	Predicted dissatisfied	Predicted satisfied
Actual dissatisfied	28604	203
Actual satisfied	516	34318

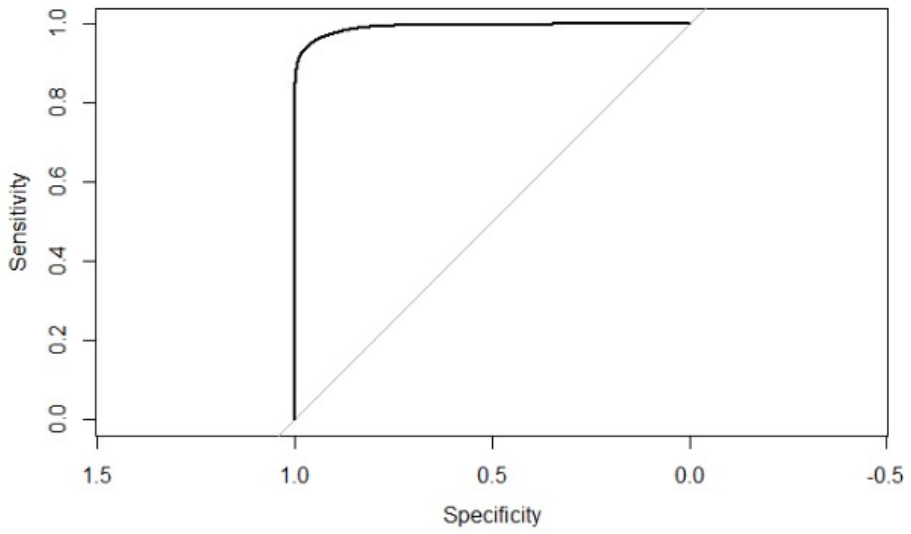
Confusion matrix on test data:

	Predicted dissatisfied	Predicted satisfied
Actual dissatisfied	11835	514
Actual satisfied	740	14187

Stats derived from confusion matrix

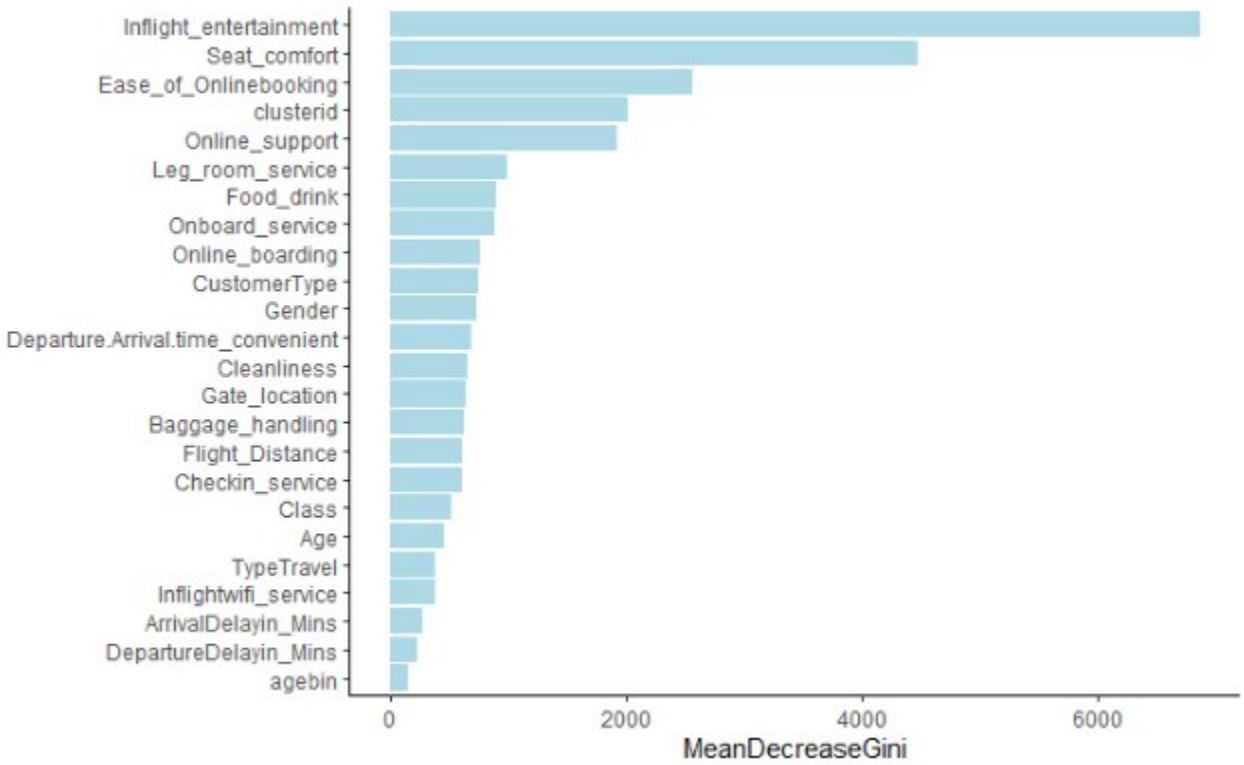
	Training Data	Test Data
TPR(Sensitivity)	.985	.950
TNR(Specificity)	.993	.958
Accuracy	.988	.954

ROC Curve and other stats derived from Test data:



Area under the curve: 0.99

Variable Importance:



The plot above shows that Inflight entertainment, Seat comfort, Ease of online booking, Online support, Legroom service, Food drink are the key drivers of satisfaction.

Inflight WiFi service, Arrival delay and Departure delay are least contributors towards Satisfaction.

Clusterid is also identified as one of the key factor, this field signifies the customer segment on the basis of Customer Type, Class, Type Travel and Gender. Hence we will try study the variable importance levels for each segment by running models on each segment.

c. Xgboost

Various values of parameters were tried for learning rate (eta), max depth and nrounds using for loops and accuracy was checked for each combination. Based on the combination where accuracy was highest, the final parameters for model was decided.

```
aviation.xgb.fit = xgboost(
  data = as.matrix(data.matrix(xgbTrain[,-c(1,10,25)])),
  label = as.matrix(data.matrix(xgbTrain[,c(10)])),
  eta = 0.1,
  max_depth = 6,
  min_child_weight = 5,
  nrounds = 399,
  nfold = 5,
  objective = "binary:logistic",
  verbose = 0,
  early_stopping_rounds = 10
)
```

Confusion matrix on training data:

	Predicted dissatisfied	Predicted satisfied
Actual dissatisfied	28184	623
Actual satisfied	888	33946

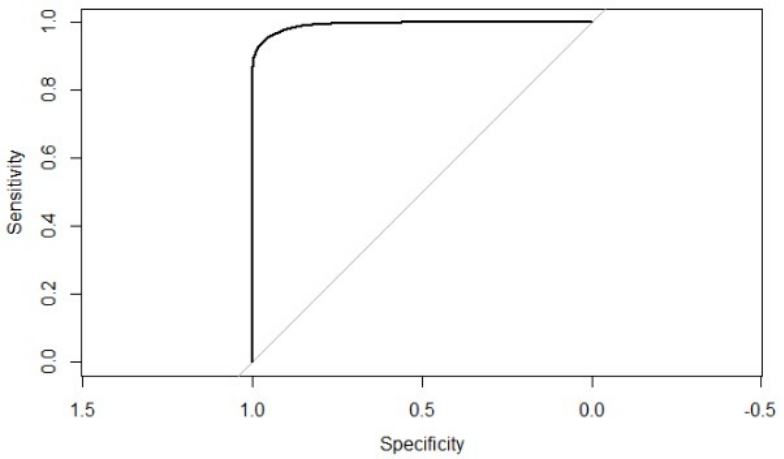
Confusion matrix on test data:

	Predicted dissatisfied	Predicted satisfied
Actual dissatisfied	11779	570
Actual satisfied	683	14244

Stats derived from confusion matrix

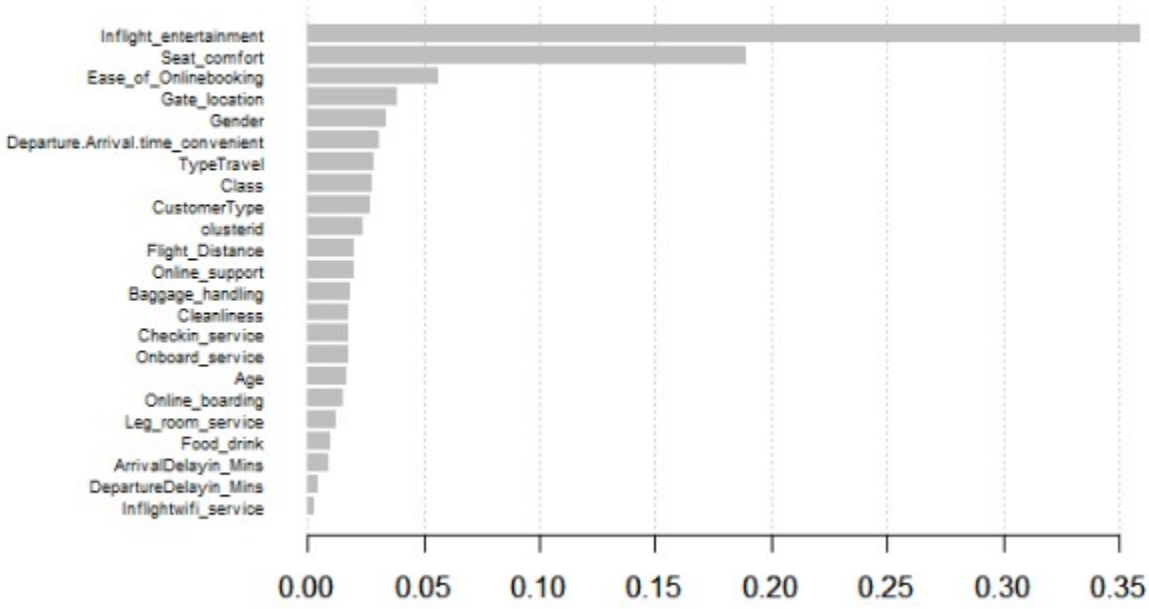
	Training Data	Test Data
TPR(Sensitivity)	.974	.954
TNR(Specificity)	.978	.954
Accuracy	.976	.954

ROC Curve and other stats derived from Test data:



Area under the curve: 0.99

Variable Importance:



As per the variable importance chart, Inflight entertainment, Seat Comfort and Ease of Onlinebooking are the major contributors for Satisfaction, which is the same trend seen in Random Forest. Similarly Inflight WiFi service, Arrival delay and Departure delay are least contributors towards Satisfaction.

However the variable importance order varies for some other factors between XGBoost and Random Forest.

d. Model comparison

Table with performance stats:

Algorithm	TPR	TNR	Accuracy	AUC
Logistic Regression	**.911**	**.907**	**.909**	**.968**
Random Forest	**.950**	**.958**	**.954**	**.99**
XGBoost	**.954**	**.954**	**.954**	**.99**

Random Forest and XGBoost have got similar statistics with respect to Accuracy and AUC. Both Random forest and XGBoost was used to study further trends for each cluster. Separate Xgboost models were created for each cluster for predicting the outcome whereas Random forest models for individual clusters were used to identify the feature importance.

Following are the stats derived from the Test data of each cluster:

Cluster	TPR	TNR	Accuracy	AUC
Cluster 1	**.991**	**.961**	**.985**	**.9975**
Cluster 2	**.923**	**.945**	**.937**	**.988**
Cluster 3	**.992**	**.969**	**.987**	**.998**
Cluster 4	**.929**	**.908**	**.911**	**.968**
Others	**.948**	**.880**	**.911**	**.976**

All the models created exhibited more than 91% accuracy on the respective cluster data.

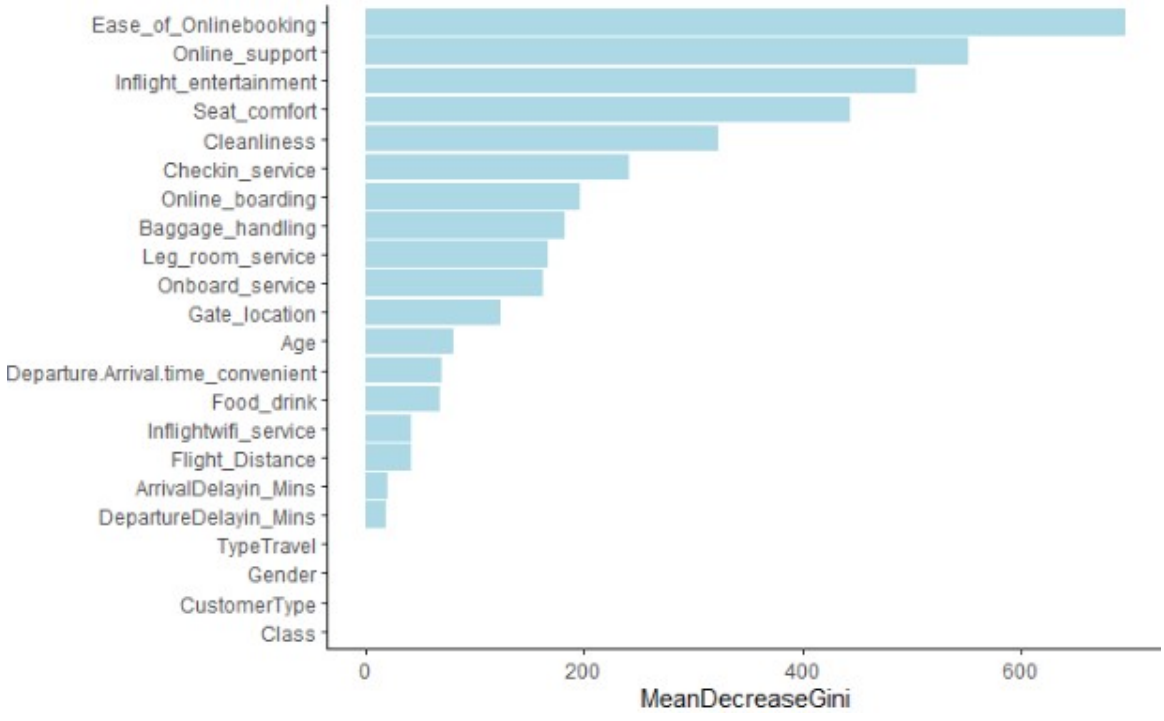
e. Segment insights

Cluster 1: Male, Loyal customers flying Business class for Business trips.

Overall 78.1% of this segment of customers are satisfied.

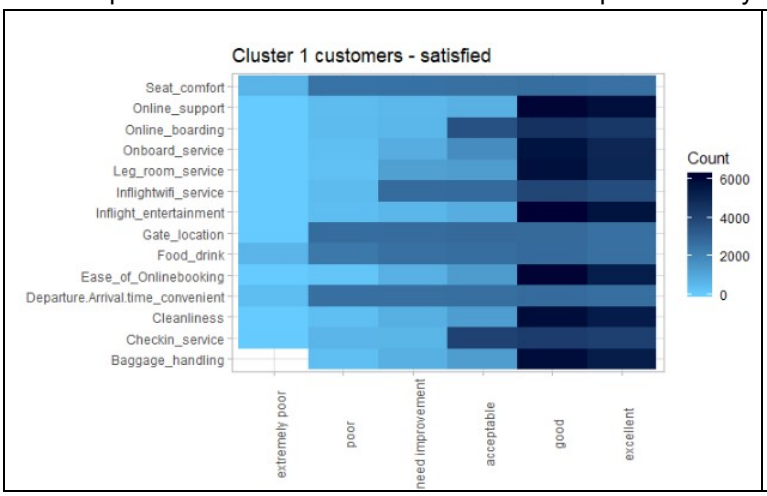
```
dissatisfied    satisfied
        3830        13681
```

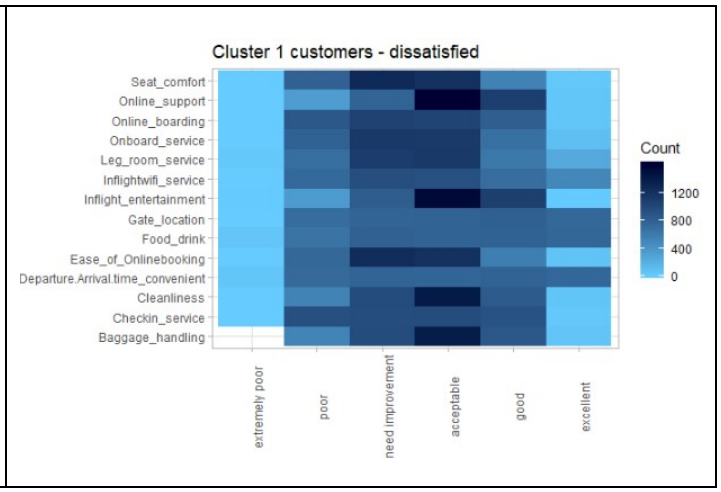
Variable importance:



Variable importance plot show Ease of online booking, Online support, Inflight entertainment, Seat comfort, Cleanliness, Checkin service, Online boarding, Baggage handling, Legroom service and Onboard service to be the key drivers for satisfaction.

Similar rating trend is also shown by the heat maps below. Satisfied customers have rated mostly ‘good’ or ‘excellent’ with respect to the above mentioned factors as represented by darker blocks.





Clearly amenities provided by Falcon airlines in business class is good, hence most of these customers are satisfied. Customers who were not satisfied didn’t find Ease of online booking, Seat comfort, Online boarding, Onboard service, Cleanliness, Baggage handling, Checkin service, Legroom service to be good enough and we see a lot of ratings as ‘poor’ or ‘need improvement’ for these factors.

Food drink, Gate location and Departure Arrival time convenience can be seen to have a similar rating trend between satisfied and dissatisfied customer, which means lot of satisfied customers have rated ‘poor’ for these services.

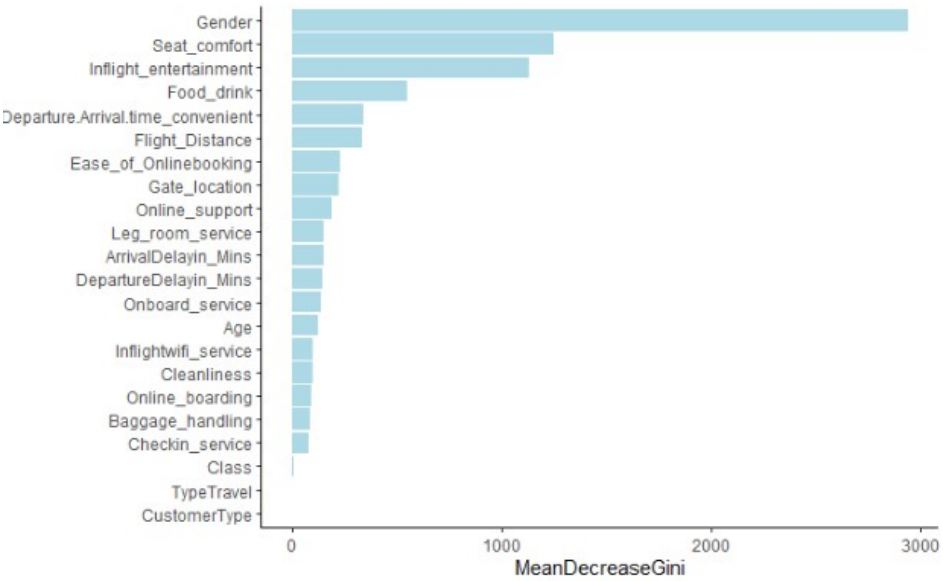
Cluster 2: Loyal customers flying Economy class for Personal trips.

This segment of customers have higher number of dissatisfactions and number of male dissatisfied customers are considerably higher. Female loyal customers have been more satisfied, but the number of male dissatisfied customers have resulted in higher number of dissatisfied customers.

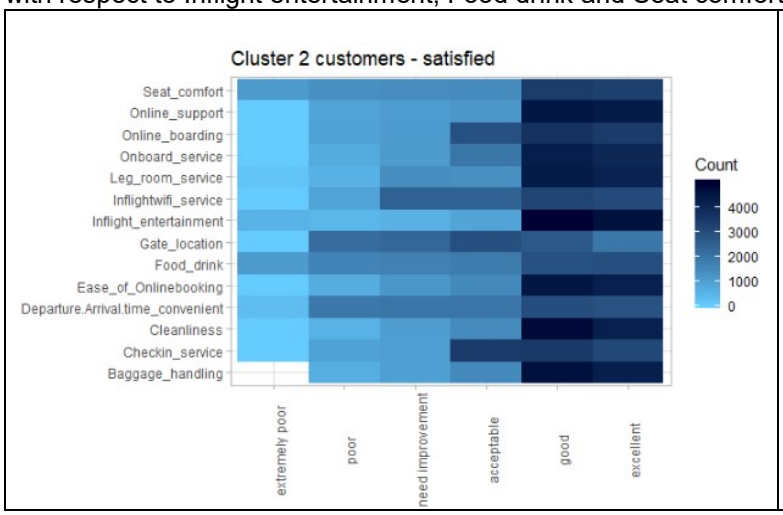
```
              Female  Male
dissatisfied    2253 11720
satisfied      10768  1356
```

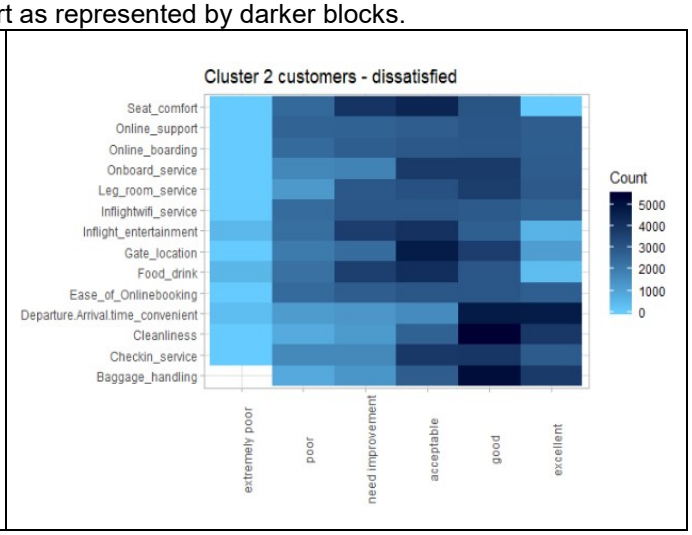
Variable importance:

Because of the imbalance in the satisfaction numbers based on Gender, it is the primary factor identified by the model. As shown by variable importance below apart from Gender, Seat comfort, Inflight Entertainment and Food drink are the most notable factors driving satisfaction.



Similar rating trend is also shown by the heat maps below. Satisfied customers have rated mostly 'good' or 'excellent' with respect to Inflight entertainment, Food drink and Seat comfort as represented by darker blocks.





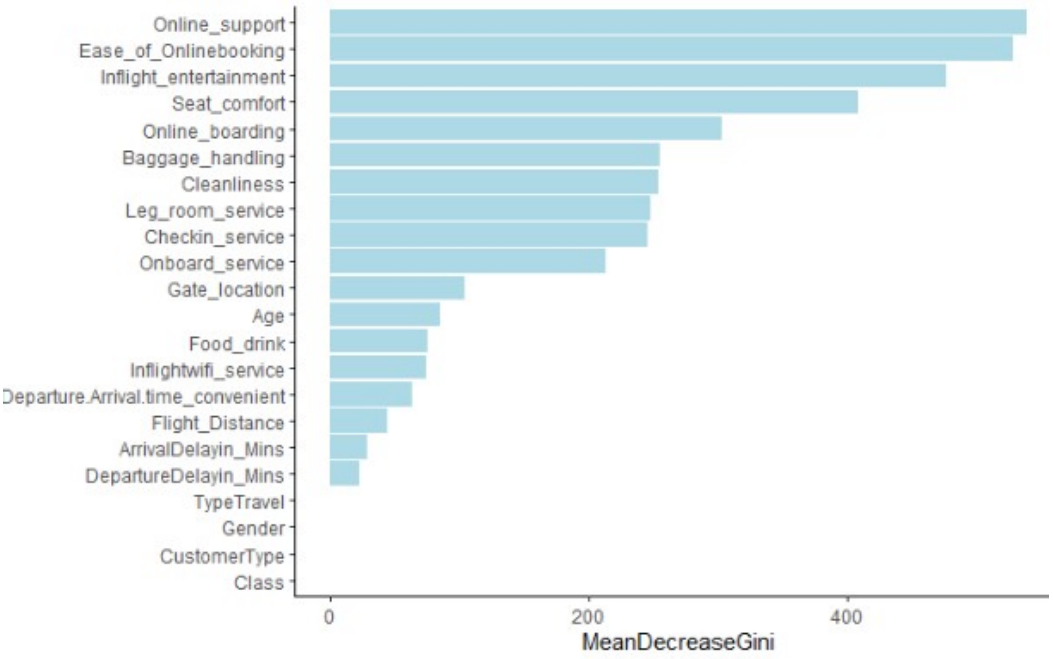
Amenities provided in Economy class is not as good as Business class, rating trend of dissatisfied customers is a mixed bag. Predominantly male who are dissatisfied didn't find Ease of online booking, Seat comfort, Online boarding, Inflight entertainment and Food drink to be good enough and we see a lot of ratings as 'poor' or 'need improvement' for these factors.

Cluster 3: Female, Loyal customers flying Business class for Business trips.

Just like Male loyal customers flying business class, this segment also comprises mostly of satisfied customers. Overall 77.9% of this segment of customers are satisfied.

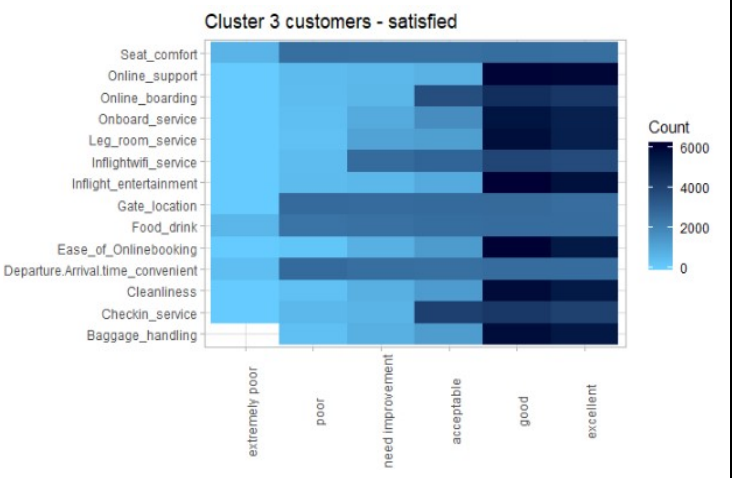
```
dissatisfied    satisfied
        3889        13732
```

Variable importance:



Variable importance plot show Online support, Ease of online booking, Inflight entertainment, Seat comfort, Online boarding, Baggage handling, Cleanliness, Legroom service, Checkin service and Onboard service to be the key drivers for satisfaction.

Similar rating trend is also shown by the heat maps below. Satisfied customers have rated mostly ‘good’ or ‘excellent’ with respect to the above mentioned factors as represented by darker blocks.



Clearly amenities provided by Falcon airlines in business class is good, hence most of these customers are satisfied. Customers who were not satisfied didn’t find Ease of online booking, Seat comfort, Online boarding, Legroom service, Inflight entertainment, Online support, Cleanliness, Baggage handling to be good enough and we see a lot of ratings as ‘poor’ or ‘need improvement’ for these factors.

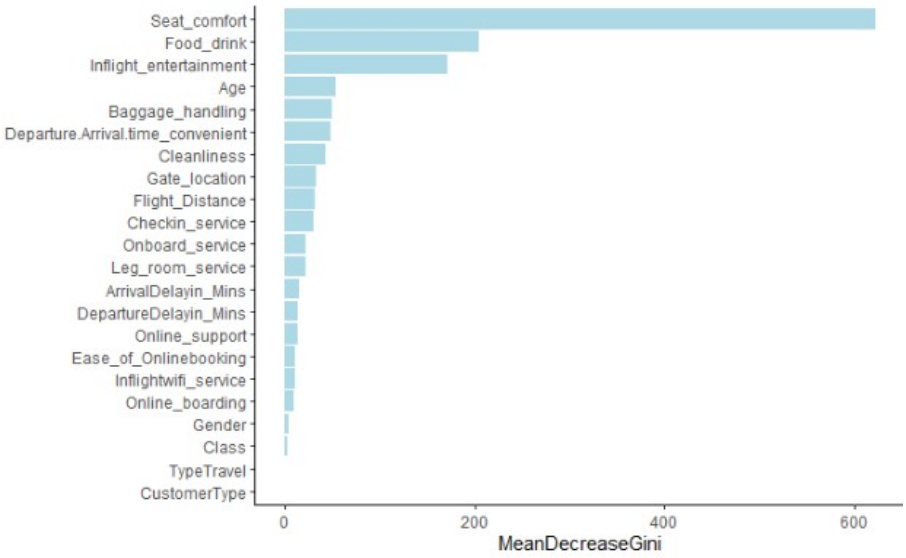
Food drink, Gate location and Departure Arrival time convenience can be seen to have a similar rating trend between satisfied and dissatisfied customer, which means lot of satisfied customers have rated ‘poor’ for these services.

Cluster 4: Disloyal customers flying Economy class for Business trips.

This is a small segment of customers, consisting of around 10K customers, but overall this segment has higher number of dissatisfaction.

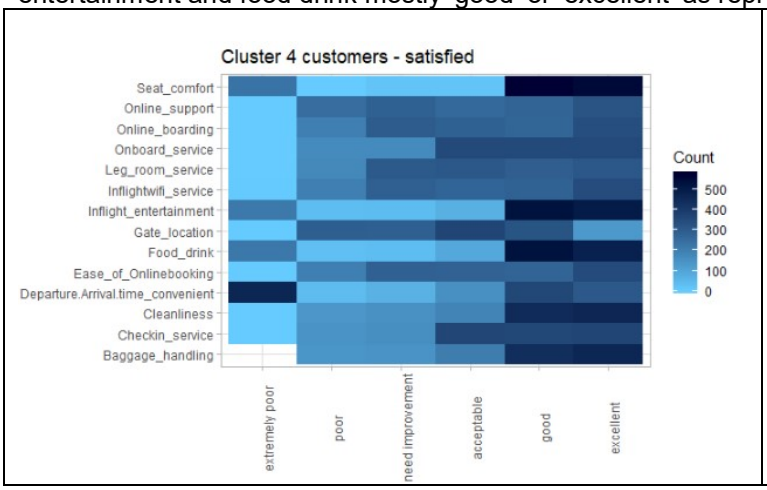
```
dissatisfied    satisfied
        8626         1394
```

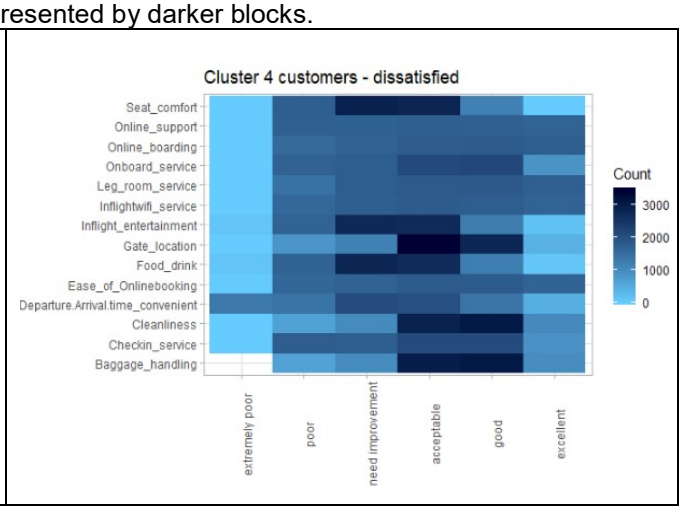
Variable importance:



Variable importance plot shows Seat comfort, Food drink and Inflight entertainment are the major factors of dissatisfaction.

Similar rating trend is also shown by the heat maps below. Satisfied customers have rated Seat comfort, Inflight entertainment and food drink mostly ‘good’ or ‘excellent’ as represented by darker blocks.





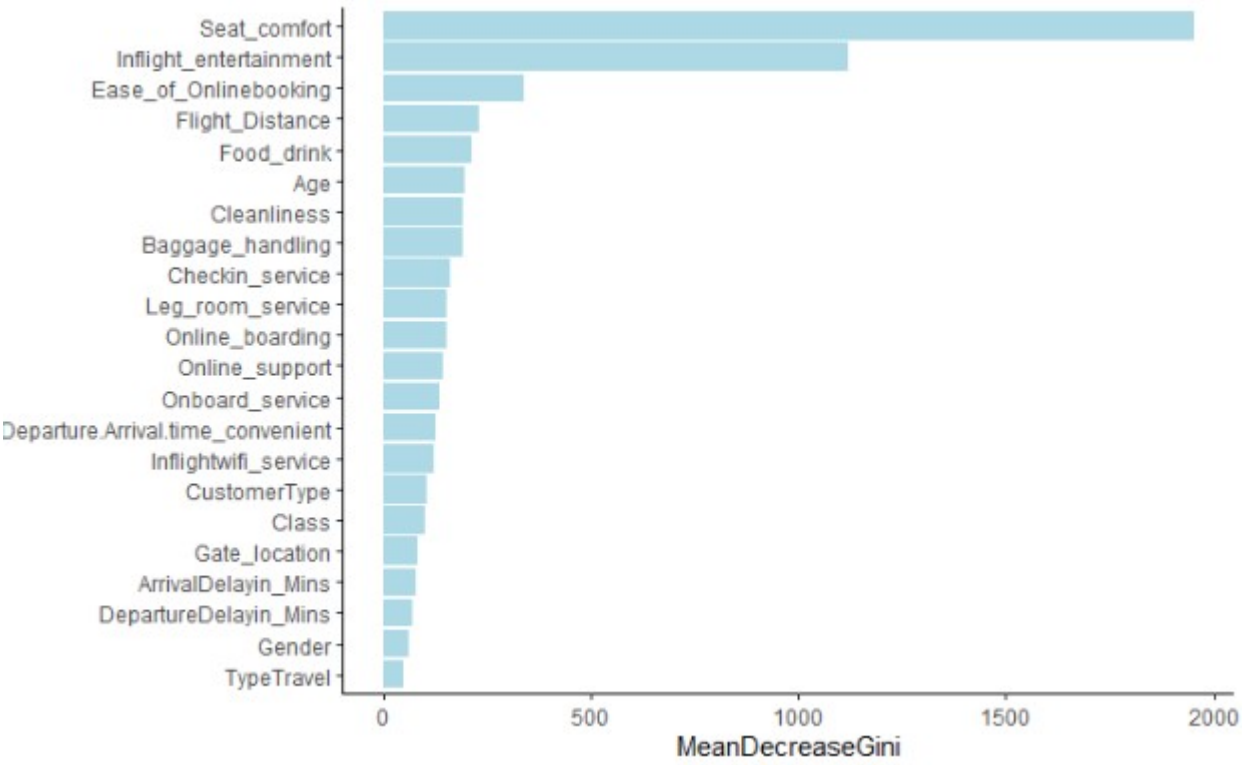
Dissatisfied customers have rated Seat comfort, Inflight entertainment, food drink as ‘need improvement’ or ‘acceptable’.

Others:

Customers who were not tagged in any other cluster comprises of around 19K records. Majority of these customers are also dissatisfied.

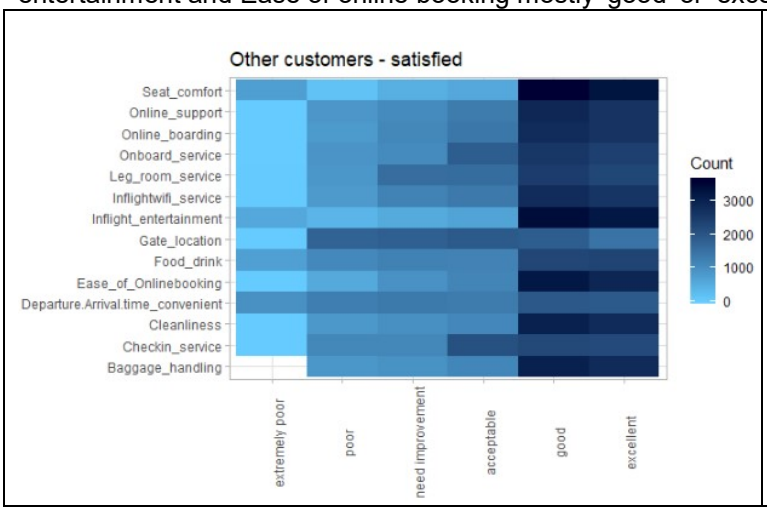
```
dissatisfied    satisfied
       10838         8830
```

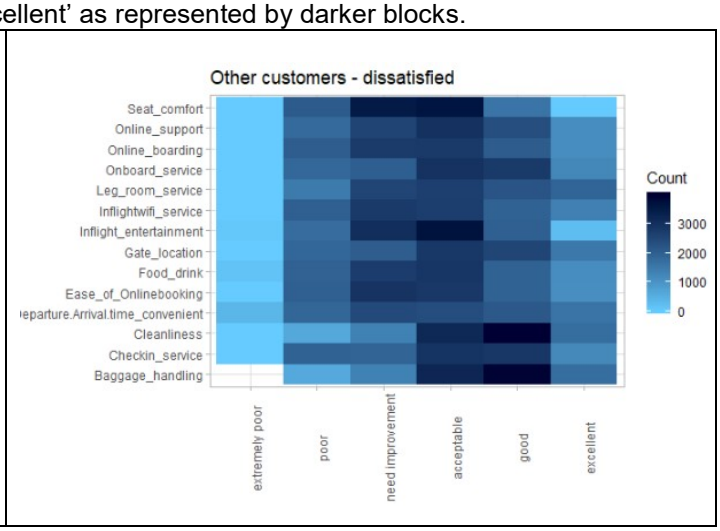
Variable Importance:



Variable importance plot shows Seat comfort, Inflight entertainment and Ease of online booking are the major factors of dissatisfaction.

Similar rating trend is also shown by the heat maps below. Satisfied customers have rated Seat comfort, Inflight entertainment and Ease of online booking mostly ‘good’ or ‘excellent’ as represented by darker blocks.





Dissatisfied customers have rated almost all features as ‘poor’, ‘need improvement’ or ‘acceptable’ as we see all dark blocks concentrated towards the center of the plot. Out of them Seat comfort and Inflight entertainment are the most prominent blocks.

8. Summary:

Observations from exploratory data analysis and model interpretations:

- Based on the exploratory data analysis, we found that the dataset provided has higher number of loyal customers compared to disloyal customers and number of satisfied customers are slightly higher than dissatisfied customers.
- Highest number of customers fly business class for business trips; second highest numbers of customers fly economy class for personal trips. Least number of customers fly Eco plus.
- Majority of the dissatisfied customers are male. Female customers have been less critical of the airline amenities.
- Loyal customers are more satisfied than disloyal customers.
- Inflight entertainment, Seat comfort, Ease of online booking, Online support are the key drivers of Satisfaction.
- Inflight WiFi service, Arrival delay, Departure delay are least contributors towards Satisfaction.

Recommendations for Falcon Airlines:

Business Class:

- Ratings given by customers are always subjective; we have seen that Ease of online booking, Online support, Inflight entertainment, Cleanliness and Baggage handling to be the key drivers for satisfaction. People who are satisfied by these services are most likely to be overall satisfied. Majority business class passengers have found these amenities to be good, but there are few who have rated these factors as poor. So Falcon Airlines might need to reach out to these few customers who have rated these amenities poorly to understand their needs.

- Seat comfort has to be improved as a lot of customers (both satisfied and dissatisfied) have rated it poor or need improvement.
- Food drink has to be improved as a lot of customers (both satisfied and dissatisfied) have rated it poor or need improvement.
- Inflight WiFi service should also be improved in Business class.

Economy Class:

- Overall Falcon Airlines should work towards making amenities in Economy class to be more male friendly. Male customers have been highly critical and dissatisfied with the amenities of Economy class.
- Seat comfort, Food drink and Inflight entertainment are the major factors of dissatisfaction in Economy class. Most of the customers are dissatisfied who have rated these amenities low. These amenities have to be improved in Economy class.
- Online Support, Legroom service, online boarding, Ease of Online booking also needs to be improved for Economy class.

Falcon airlines should have a look at the flight timings as lot of customers have rated Departure arrival timings as poor or need improvement.