Exploring Passenger Satisfaction

An Explorative Analysis for US Airline Passenger Satisfaction

Nathan Dennis

Introduction

Within this project we will explore the satisfaction of US Airline Passengers based off various factors, many of which stem from data given by the passengers. The dataset can be found here from Kaggle.

The goal of this project is to discover what factors may influence Airline Passenger satisfaction through data analysis with visualizations, statistics, and models. The final objective will be to create a model which accurately predicts customer satisfaction based on the given factors.

We will begin with some exploratory data analysis to help understand the dataset. To do this we may create visualizations or calculate different types of statistics to observe trends or discrepancies in the data. If necessary, the data will have to be cleaned before any further analysis is done if there are data quality issues present.

```
#Load in Dataset
satisfaction <- read_csv('satisfaction.csv', show_col_types = FALSE)
```

Understanding the Dataset

First, we can observe the first 10 rows of the dataset.

```
head(satisfaction,10)
```

```
## # A tibble: 10 x 24
##        id satisfact~1 Gender Custo~2   Age Type ~3 Class Fligh~4 Seat ~5 Depar~6
##     <dbl> <chr>       <chr>  <chr>   <dbl> <chr>   <chr>   <dbl>   <dbl>   <dbl>
##  1  11112 satisfied   Female Loyal ~    65 Person~ Eco       265       0       0
##  2 110278 satisfied   Male   Loyal ~    47 Person~ Busi~    2464       0       0
##  3 103199 satisfied   Female Loyal ~    15 Person~ Eco      2138       0       0
##  4  47462 satisfied   Female Loyal ~    60 Person~ Eco       623       0       0
##  5 120011 satisfied   Female Loyal ~    70 Person~ Eco       354       0       0
##  6 100744 satisfied   Male   Loyal ~    30 Person~ Eco      1894       0       0
##  7  32838 satisfied   Female Loyal ~    66 Person~ Eco       227       0       0
##  8  32864 satisfied   Male   Loyal ~    10 Person~ Eco      1812       0       0
##  9  53786 satisfied   Female Loyal ~    56 Person~ Busi~      73       0       0
## 10   7243 satisfied   Male   Loyal ~    22 Person~ Eco      1556       0       0
## # ... with 14 more variables: `Food and drink` <dbl>, `Gate location` <dbl>,
## #   `Inflight wifi service` <dbl>, `Inflight entertainment` <dbl>,
## #   `Online support` <dbl>, `Ease of Online booking` <dbl>,
## #   `On-board service` <dbl>, `Leg room service` <dbl>,
## #   `Baggage handling` <dbl>, `Checkin service` <dbl>, Cleanliness <dbl>,
## #   `Online boarding` <dbl>, `Departure Delay in Minutes` <dbl>,
## #   `Arrival Delay in Minutes` <dbl>, and abbreviated variable names ...
```

We can see that there are a total of 24 columns and 129,880 rows. We see in the output the column names have spaces in them, which could make them hard to work with. We can rename the columns which have spaces and any other columns which may be incorrectly named. We can observe a full list of the column names now.

```
colnames(satisfaction)
```

```
##  [1] "id"                                "satisfaction_v2"
##  [3] "Gender"                            "Customer Type"
##  [5] "Age"                               "Type of Travel"
##  [7] "Class"                             "Flight Distance"
##  [9] "Seat comfort"                      "Departure/Arrival time convenient"
## [11] "Food and drink"                    "Gate location"
## [13] "Inflight wifi service"             "Inflight entertainment"
## [15] "Online support"                    "Ease of Online booking"
## [17] "On-board service"                  "Leg room service"
## [19] "Baggage handling"                  "Checkin service"
## [21] "Cleanliness"                       "Online boarding"
## [23] "Departure Delay in Minutes"        "Arrival Delay in Minutes"
```

We will now rename any columns with issues, observing the new column names.

```
satisfaction <- satisfaction %>% rename('Satisfaction' = 'satisfaction_v2',
                                  'Customer_type' = 'Customer Type',
                                  'Type_of_travel' = 'Type of Travel',
                                  'Flight_distance' = 'Flight Distance',
                                  'Seat_comfort' = 'Seat comfort',
                                  'Departure_arrival_time_convenient' =
                                    'Departure/Arrival time convenient',
                                  'Food_drink' = 'Food and drink',
                                  'Gate_location' = 'Gate location',
                                  'Inflight_wifi' = 'Inflight wifi service',
                                  'Inflight_entertainment' = 'Inflight entertainment',
                                  'Online_support' = 'Online support',
                                  'Ease_booking' = 'Ease of Online booking',
                                  'On_board_service' = 'On-board service',
                                  'Leg_room_service' = 'Leg room service',
                                  'Baggage_handling' = 'Baggage handling',
                                  'Checkin_service' = 'Checkin service',
                                  'Online_boarding' = 'Online boarding',
                                  'Departure_delay' = 'Departure Delay in Minutes',
                                  'Arrival_delay' = 'Arrival Delay in Minutes')

colnames(satisfaction)
```

```
##  [1] "id"                                "Satisfaction"
##  [3] "Gender"                            "Customer_type"
##  [5] "Age"                               "Type_of_travel"
##  [7] "Class"                             "Flight_distance"
##  [9] "Seat_comfort"                      "Departure_arrival_time_convenient"
## [11] "Food_drink"                        "Gate_location"
## [13] "Inflight_wifi"                     "Inflight_entertainment"
## [15] "Online_support"                    "Ease_booking"
```

```
## [17] "On_board_service"                  "Leg_room_service"
## [19] "Baggage_handling"                  "Checkin_service"
## [21] "Cleanliness"                       "Online_boarding"
## [23] "Departure_delay"                   "Arrival_delay"
```

Now we have the columns renamed and much easier to work with. We can now describe each column, since the column names themselves are not easy to interpret. We can do so with a table where each row will describe the column.

Dataset Column Description

Variable Name	Description
id	Customer id
Satisfaction	Level of Airline passenger satisfaction (satisfied, Neutral/dissastisfied)
Gender	Airline passenger gender (Male/Female in dataset)
Customer_type	Type of customer (Loyal, disloyal)
Age	Airline passenger age
Type_of_travel	Purpose of the flight (Business, Personal)
Class	Travel Class in the plane (Business, Economy, Economy Plus)
Flight_distance	Distance of the flight (miles)
Seat_comfort	Satisfaction Level of Seat Comfort (0-5)
Departure_arrival_time_convenient	Satisfaction Level of Departure/Arrival time (0-5)
Food_drink	Satisfaction level of foods and drinks (0-5)
Gate_location	Satisfaction level of gate location (0-5)
Inflight_wifi	Satisfaction level of in flight wifi (0-5)
Inflight_entertainment	Satisfaction level of in flight entertainment (0-5)
Online_support	Satisfaction level of online support (0-5)
Ease_booking	Satisfaction level of booking ease (0-5)
On_board_service	Satisfaction level of on board service (0-5)
Leg_room_service	Satisfaction level of leg room service (0-5)
Baggage_handling	Satisfaction level of baggage handling (0-5)
Checkin_service	Satisfaction level of check-in service (0-5)
Cleanliness	Satisfaction level of cleanliness (0-5)
Online_boarding	Satisfaction level of online boarding (0-5)
Departure_delay	Departure delay in minutes
Arrival_delay	Arrival delay in minutes

As seen in the dataset columns description, many variables are satisfaction levels based on user input with 5 being highly satisfied and 0 being the least satisfied. We can now begin to do some exploratory data analysis and uncover trends or discrepencies within the dataset.

Exploratory Data Analysis

Starter Exploration

We can start by calculating summary statistics for each numeric column and making sure every column the correct typing.

```r
summary(satisfaction)
```

```
##        id          Satisfaction          Gender           Customer_type     
##  Min.   :     1   Length:129880      Length:129880      Length:129880     
##  1st Qu.: 32471   Class :character   Class :character   Class :character  
##  Median : 64940   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 64940                                                           
##  3rd Qu.: 97410                                                           
##  Max.   :129880                                                           
##                                                                           
##       Age        Type_of_travel        Class           Flight_distance
##  Min.   : 7.00   Length:129880      Length:129880      Min.   :  50   
##  1st Qu.:27.00   Class :character   Class :character   1st Qu.:1359   
##  Median :40.00   Mode  :character   Mode  :character   Median :1925   
##  Mean   :39.43                                         Mean   :1981   
##  3rd Qu.:51.00                                         3rd Qu.:2544   
##  Max.   :85.00                                         Max.   :6951   
##                                                                       
##   Seat_comfort   Departure_arrival_time_convenient   Food_drink   
##  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   :2.839   Mean   :2.991                     Mean   :2.852  
##  3rd Qu.:4.000   3rd Qu.:4.000                     3rd Qu.:4.000  
##  Max.   :5.000   Max.   :5.000                     Max.   :5.000  
##                                                                   
##  Gate_location  Inflight_wifi   Inflight_entertainment Online_support
##  Min.   :0.00   Min.   :0.000   Min.   :0.000          Min.   :0.00  
##  1st Qu.:2.00   1st Qu.:2.000   1st Qu.:2.000          1st Qu.:3.00  
##  Median :3.00   Median :3.000   Median :4.000          Median :4.00  
##  Mean   :2.99   Mean   :3.249   Mean   :3.383          Mean   :3.52  
##  3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.000          3rd Qu.:5.00  
##  Max.   :5.00   Max.   :5.000   Max.   :5.000          Max.   :5.00  
##                                                                      
##   Ease_booking   On_board_service Leg_room_service Baggage_handling
##  Min.   :0.000   Min.   :0.000    Min.   :0.000    Min.   :1.000   
##  1st Qu.:2.000   1st Qu.:3.000    1st Qu.:2.000    1st Qu.:3.000   
##  Median :4.000   Median :4.000    Median :4.000    Median :4.000   
##  Mean   :3.472   Mean   :3.465    Mean   :3.486    Mean   :3.696   
##  3rd Qu.:5.000   3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:5.000   
##  Max.   :5.000   Max.   :5.000    Max.   :5.000    Max.   :5.000   
##                                                                    
##  Checkin_service  Cleanliness    Online_boarding Departure_delay 
##  Min.   :0.000   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 :4.000   Median :   0.00  
##  Mean   :3.341   Mean   :3.706   Mean   :3.353   Mean   :  14.71  
##  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.   :1592.00  
##                                                                  
##  Arrival_delay    
##  Min.   :   0.00  
##  1st Qu.:   0.00  
```

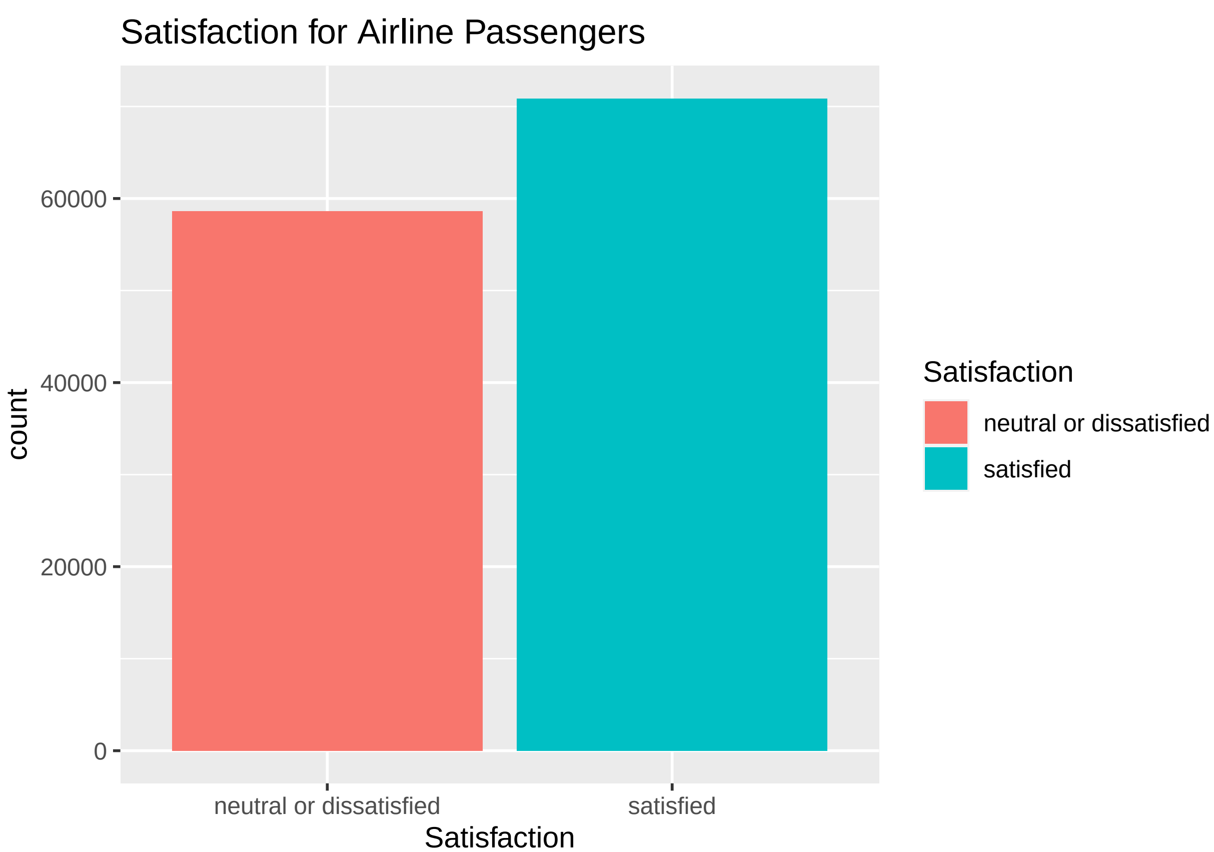
```
##   Median :   0.00
##   Mean   :  15.09
##   3rd Qu.:  13.00
##   Max.   :1584.00
##   NA's   :393
```

We see that every column is of the correct typing. However when one observed the statistics for the "Arrival_delay" column is seems that there are 393 NA values. To replace these missing values, I have decided the best approach would be to remove these values completely. One reason is since we have so much data, losing 393 rows wouldn't hurt analysis dramaticly. Another reason is that filling in this value with some other method, using the median of the column for example, could produce very inaccurate results for the arrival delay and harm our analysis. It is best in this case to remove the data from out data set. Also, it seems like id isn't necessary in analysis since everyone gets a unique id, so we can also remove it.

```r
satisfaction <- na.omit(satisfaction)
satisfaction <- satisfaction %>% dplyr::select(-id)
```

Now we have no NA values in our dataset. Next we should check for some class imbalance between our response variable, Satisfaction, which measures a passengers overall satisfaction. To do this we construct a bar below.

```r
satisfaction_bars <- ggplot(satisfaction, aes(x=Satisfaction, fill=Satisfaction)) +
  geom_bar() +
  labs(title="Satisfaction for Airline Passengers")
satisfaction_bars
```



Looking at the bar chart, there doesn't seem to be any major class imbalance between the two categories.

Outliers

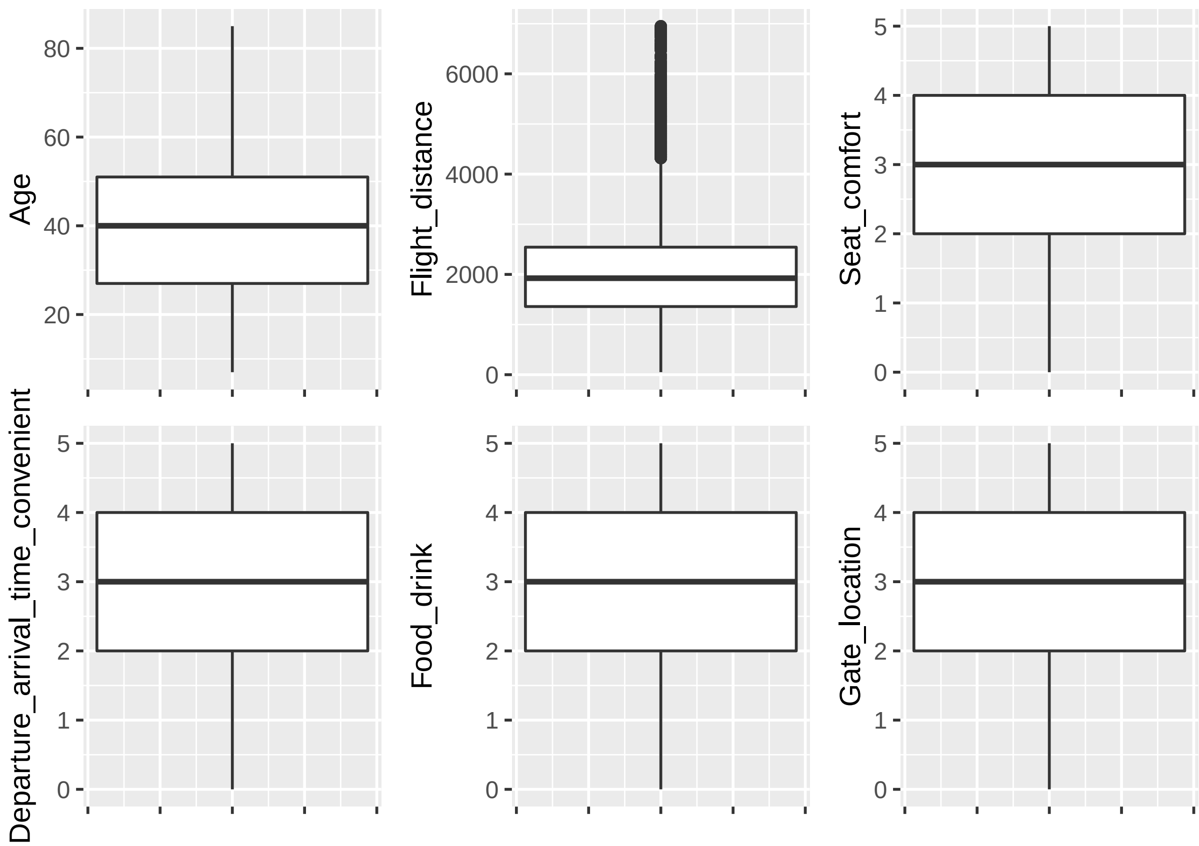
Next, we can check for outliers within our other dependent variables by making a boxplot for each numeric variable. We can organize them into 3 different combinations to improve visability.

```r
p1 <- ggplot(data=satisfaction, mapping=aes(y=Age)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p2 <- ggplot(data=satisfaction, mapping=aes(y=Flight_distance)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p3 <- ggplot(data=satisfaction, mapping=aes(y=Seat_comfort)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p4 <- ggplot(data=satisfaction, mapping=aes(y=Departure_arrival_time_convenient)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p5 <- ggplot(data=satisfaction, mapping=aes(y=Food_drink)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p6 <- ggplot(data=satisfaction, mapping=aes(y=Gate_location)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p7 <- ggplot(data=satisfaction, mapping=aes(y=Inflight_wifi)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p8 <- ggplot(data=satisfaction, mapping=aes(y=Inflight_entertainment)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p9 <- ggplot(data=satisfaction, mapping=aes(y=Online_support)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p10 <- ggplot(data=satisfaction, mapping=aes(y=Ease_booking)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p11 <- ggplot(data=satisfaction, mapping=aes(y=On_board_service)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p12 <- ggplot(data=satisfaction, mapping=aes(y=Leg_room_service)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p13 <- ggplot(data=satisfaction, mapping=aes(y=Baggage_handling)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p14 <- ggplot(data=satisfaction, mapping=aes(y=Checkin_service)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p15 <- ggplot(data=satisfaction, mapping=aes(y=Cleanliness)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p16 <- ggplot(data=satisfaction, mapping=aes(y=Online_boarding)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p17 <- ggplot(data=satisfaction, mapping=aes(y=Departure_delay)) +
```

```
  geom_boxplot() +
  theme(axis.text.x=element_blank())
p18 <- ggplot(data=satisfaction, mapping=aes(y=Arrival_delay)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
```

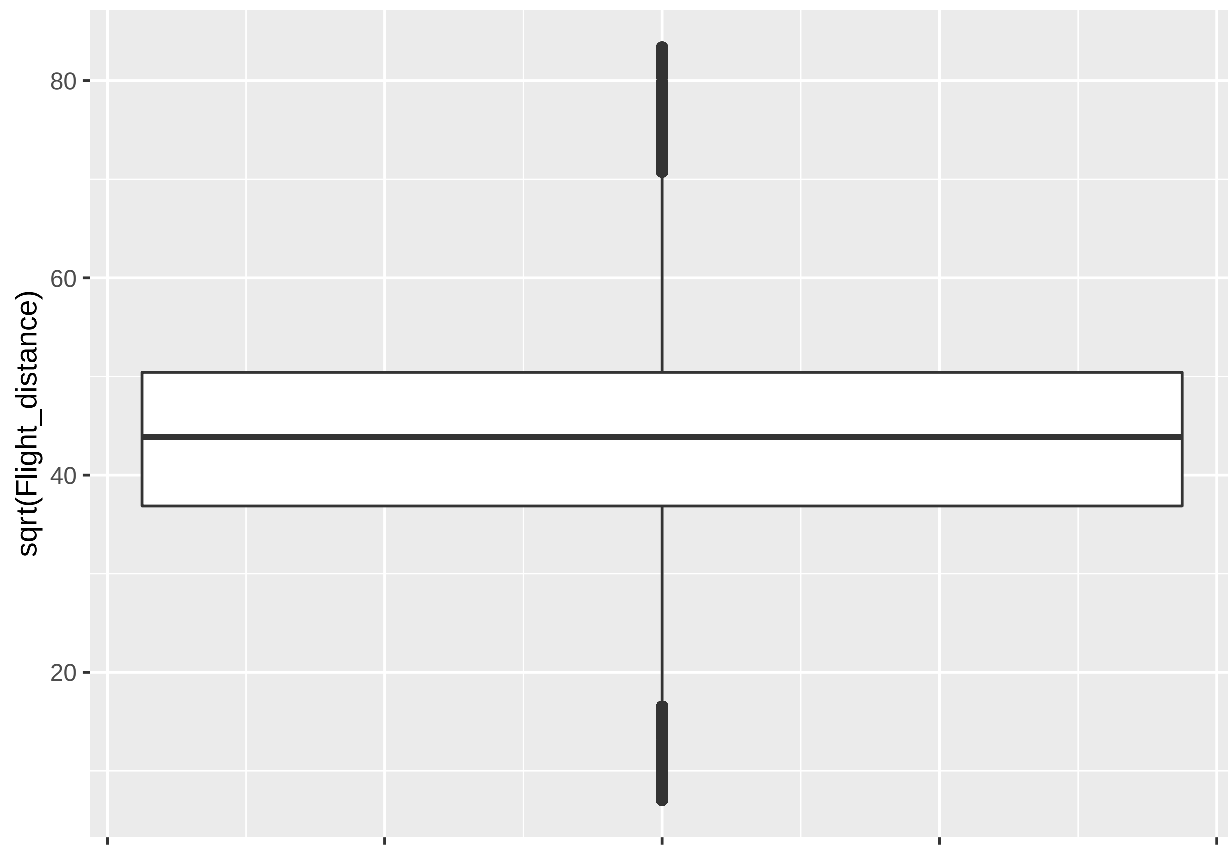
Here we display the first of the plots.

```
p1+p2+p3+p4+p5+p6
```



For this combination of plots, it seems like Flight_distance might have some outliers, as there are many points that are significantly higher than the upper whisker of the boxplot. In order to counter this, we could try a square root transformation to see if this remove the outliers.

```
ggplot(data=satisfaction, mapping=aes(y=sqrt(Flight_distance))) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())
```

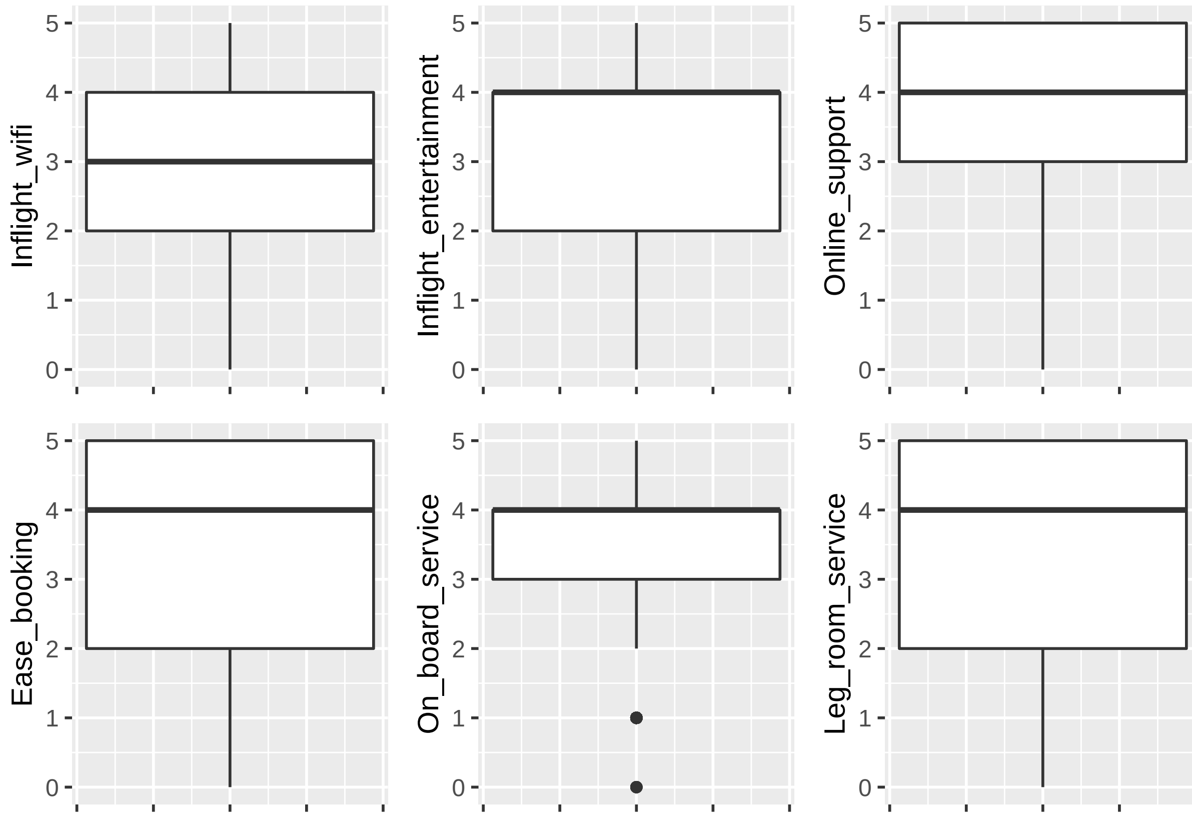


Within this square root transformation, we can see that while there are still some outliers, they are not as harmful as they are relatively close to the whiskers of the plot. We can conclude this variable looks much more normally distributed and has no significant outliers.

```
satisfaction$Flight_distance <- sqrt(satisfaction$Flight_distance)
satisfaction <- satisfaction %>% rename('Flight_distance_sqrt' = 'Flight_distance')
```

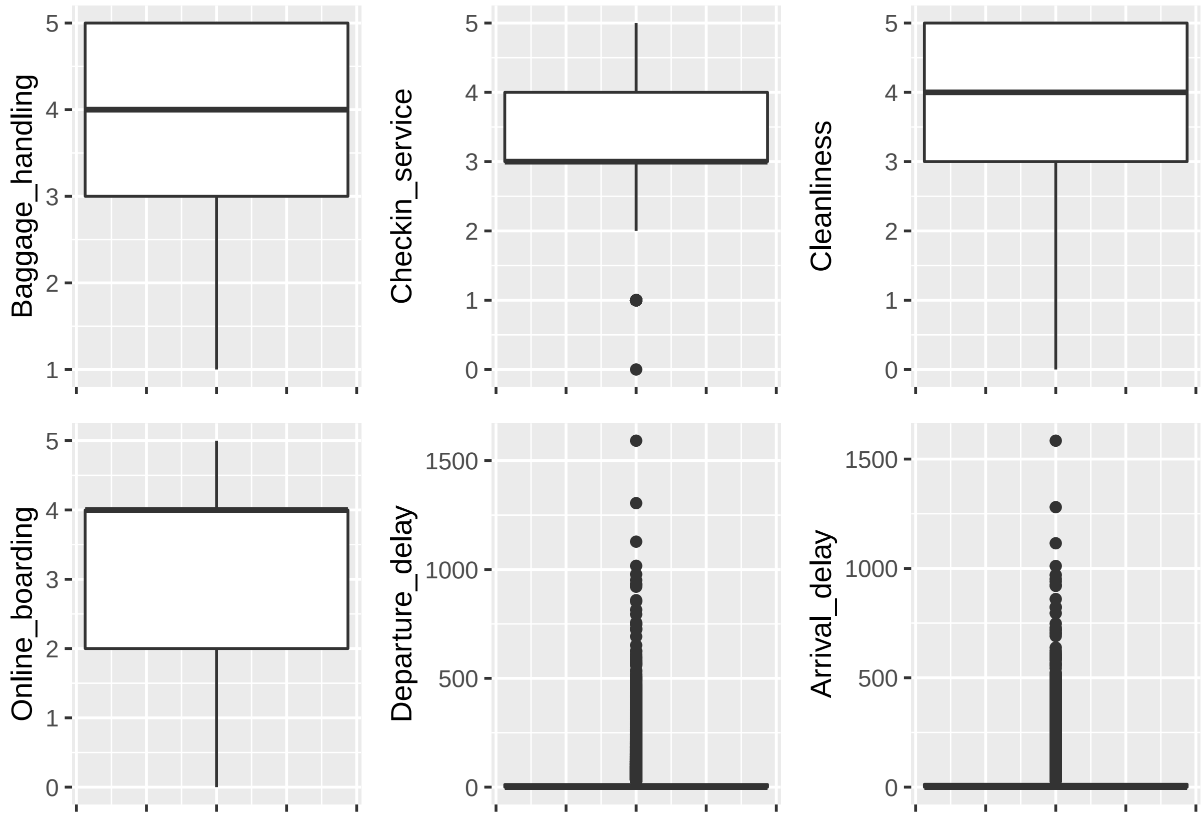
We can now observe the second combination of plots.

```
p7+p8+p9+p10+p11+p12
```



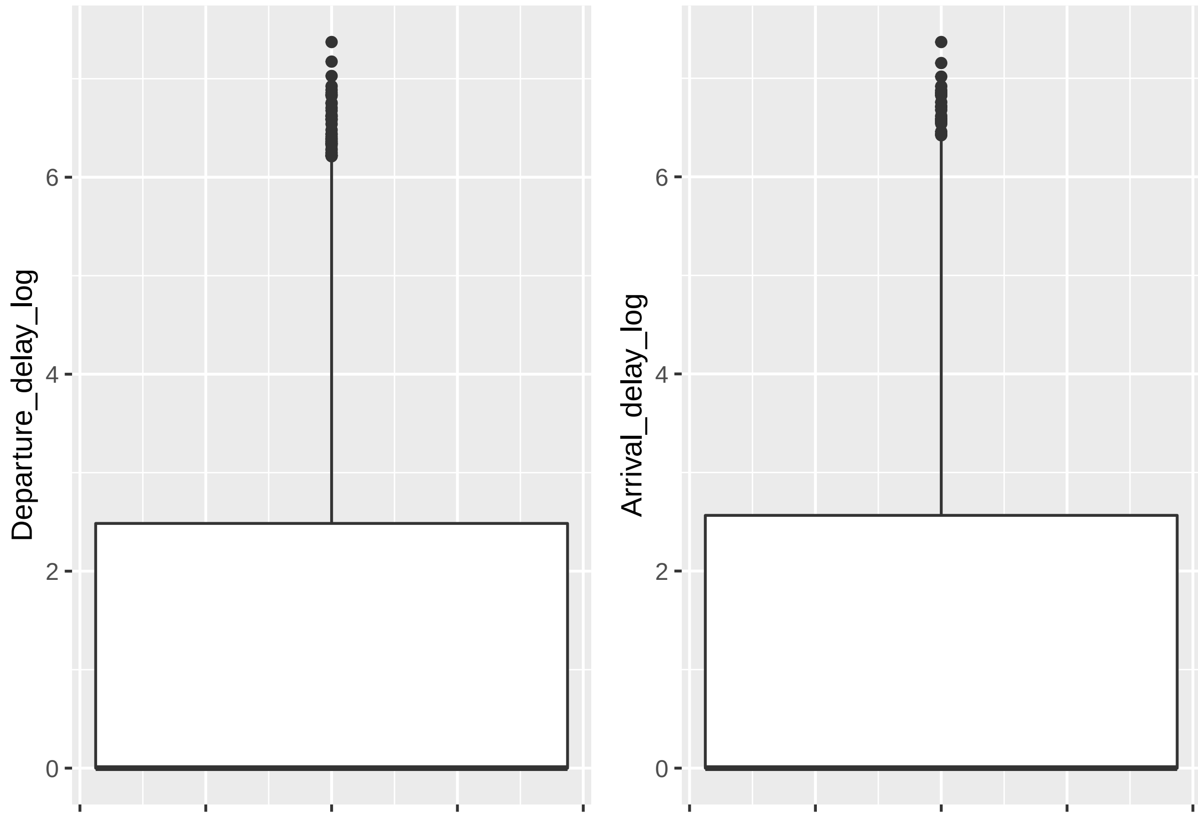
Looking at the next set of boxplots, there don't seem to be any obvious outliers. There is possibly one in On_board_service, but since the range is from 0-5 it doesn't seem too bad. We can conclude that there are no obvious outliers for these variables.

```
p13+p14+p15+p16+p17+p18
```



Within the next combination of plots, there are obvious outliers for both the Departure and Arrival delay plots.

```
satisfaction$Departure_delay <- log(satisfaction$Departure_delay)
satisfaction$Departure_delay[is.infinite(satisfaction$Departure_delay)] <- 0


satisfaction$Arrival_delay <- log(satisfaction$Arrival_delay)
satisfaction$Arrival_delay[is.infinite(satisfaction$Arrival_delay)] <- 0

satisfaction <- satisfaction %>% rename('Departure_delay_log' = 'Departure_delay',
                                        'Arrival_delay_log' = 'Arrival_delay')

p_depdel <- ggplot(data=satisfaction, mapping=aes(y=Departure_delay_log)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())

p_arrdel <- ggplot(data=satisfaction, mapping=aes(y=Arrival_delay_log)) +
  geom_boxplot() +
  theme(axis.text.x=element_blank())

p_depdel + p_arrdel
```



With this log transformation there are significantly less points that trail very far from the upper whisker, and we can conclude we have removed the outliers for both delays. We can conclude that we have removed the outliers from the numeric variables.

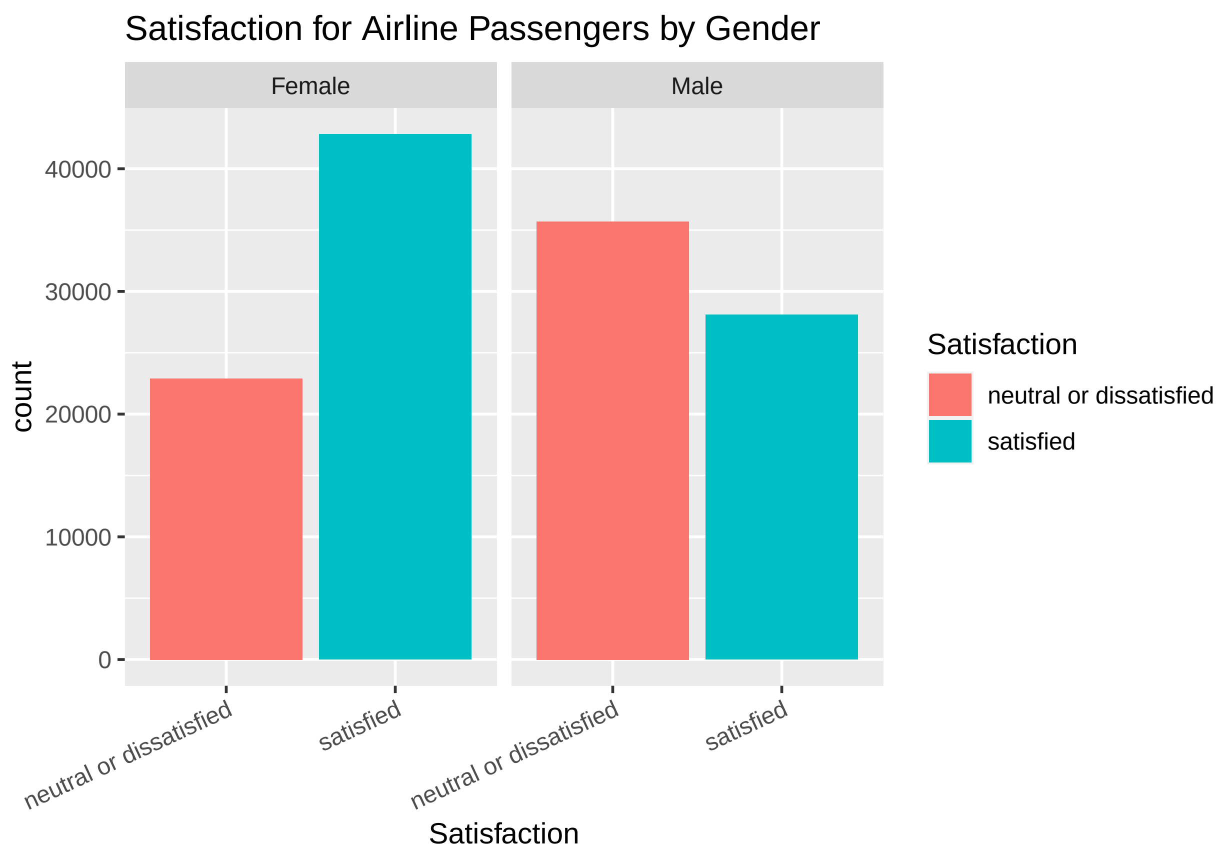
Further Exploring Data

To further explore the data and uncover trends, we can create and calculate different visualizations and statistics. We will explore some of the variables that may have an affect on customer satisfaction using these visualizations and a model created at the end.

Gender

We can observe a side by side bar plot for the gender of airline passengers, comparing the plots to spot any differences.

```
gender_bars <- ggplot(satisfaction, aes(x=Satisfaction, fill=Satisfaction)) +
  geom_bar() +
  labs(title="Satisfaction for Airline Passengers by Gender") +
  facet_wrap(~Gender)
gender_bars + theme(axis.text.x = element_text(angle = 25, vjust = 1, hjust=1))
```



From the bar plot, we can see that is seems like Females are in general more satisfied than they are neutral/dissatisfied, which is not the case for Males. Overall a greater proportion of Female passengers are satisfied with the airline, while a greater proportion of Male passengers are neutral or dissatisfied with the airline. There does seem to be a trend that females are more likely to be satisfied, while males are more likely to be neural or dissatisfied.

However, just observing the plots may not be enough to make a definite conclusion. We can also perform hypothesis testing using a chi-squared test to analyze the relationship between the variables. The null hypothesis is that there is no significant difference in passenger satisfaction between the genders. The alternative hypothesis is that there exists a statistically significant difference in passenger satisfaction between the genders. We will use a significance level of 0.05.

```
gender_table <- table(satisfaction$Satisfaction, satisfaction$Gender)
chi_squared_gender <- chisq.test(gender_table)
chi_squared_gender
```

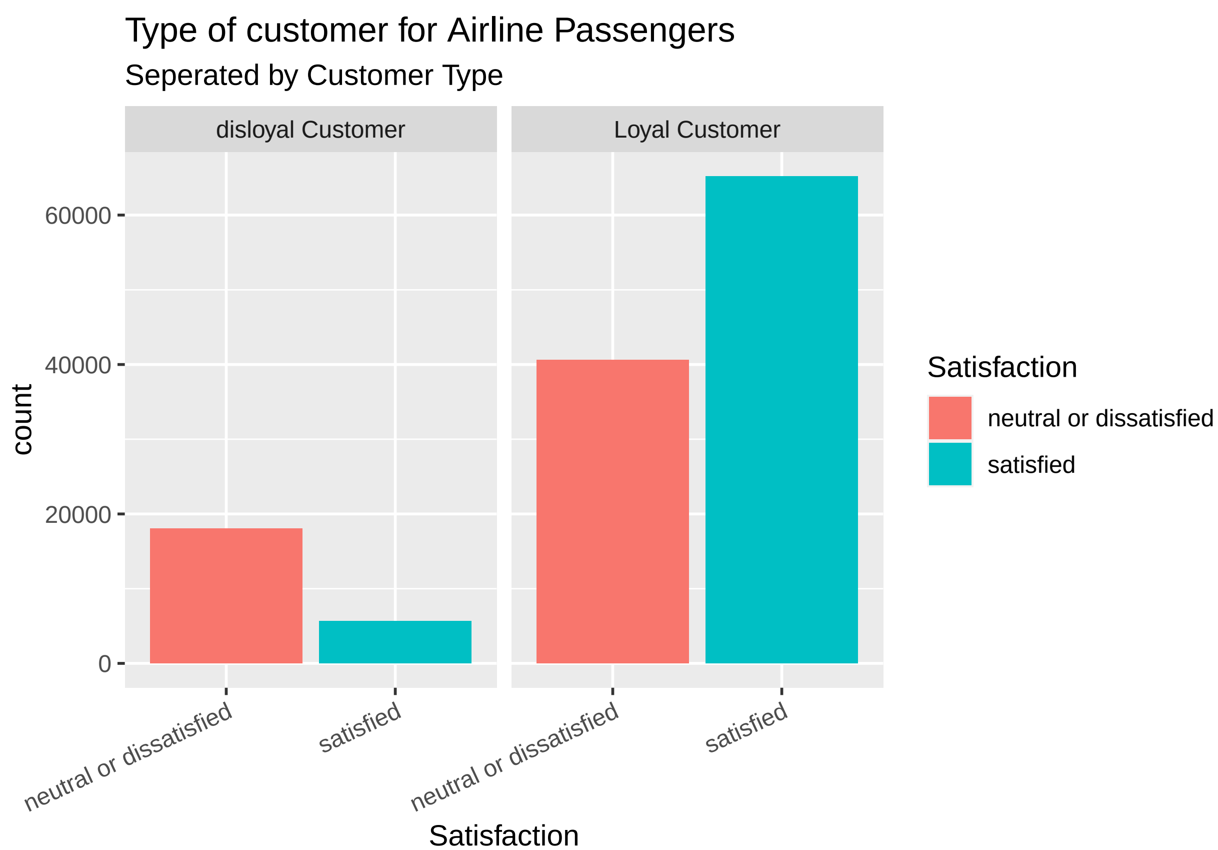
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_table
## X-squared = 5821.6, df = 1, p-value < 2.2e-16
```

Using the chi-squared test we obtain a p-value of 2.2e-16, which is below the significance level of 0.05. So, we reject the null hypothesis in favor of the alternative and conclude there is a statistically significant difference between gender and their corresponding customer satisfaction. To supplement this result with the visualization created above, where we observed that a greater proportion of female customers were satisfied and a greater proportion of male customers were neutral or dissatisfied. We can conclude that female customers are more likely to be satisfied than male customers.

Loyalty

We can now observe another bar plot, this one for the loyalty of the customer.

```
ctype_bars <- ggplot(satisfaction, aes(x=Satisfaction, fill=Satisfaction)) +
  geom_bar() +
  labs(title="Type of customer for Airline Passengers",
       subtitle="Seperated by Customer Type") +
  facet_wrap(~Customer_type)
ctype_bars + theme(axis.text.x = element_text(angle = 25, vjust = 1, hjust=1))
```

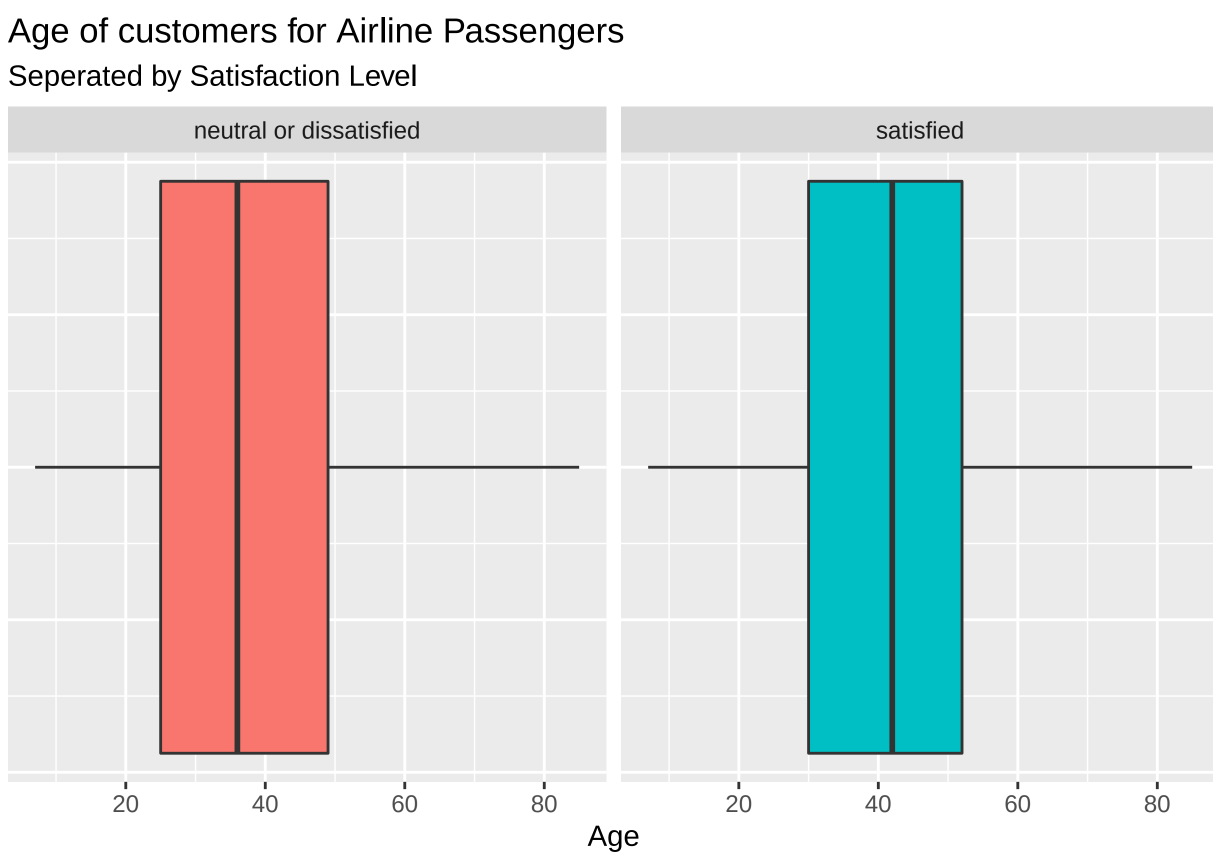


As we can see in the graphs above, the first observation made is that there are much less disloyal customers than loyal customers overall. It is also noticeable that for loyal customers, there are more of them satisfied than neural or dissatisfied. The opposite is true for the disloyal customers, are more of them are neutral or dissatisfied. It seems that overall there are more loyal customers than disloyal customers and loyal customers are more likely to be satisfied.

Age

We can now observe how Age may be related with customer satisfaction by creating a pair of boxplots, observing the differences in age.

```
age_plot <- ggplot(satisfaction, aes(x=Age, fill=Satisfaction)) +
  geom_boxplot(show.legend=FALSE) +
  labs(title="Age of customers for Airline Passengers",
       subtitle="Seperated by Satisfaction Level") +
  facet_wrap(~Satisfaction) +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
age_plot
```

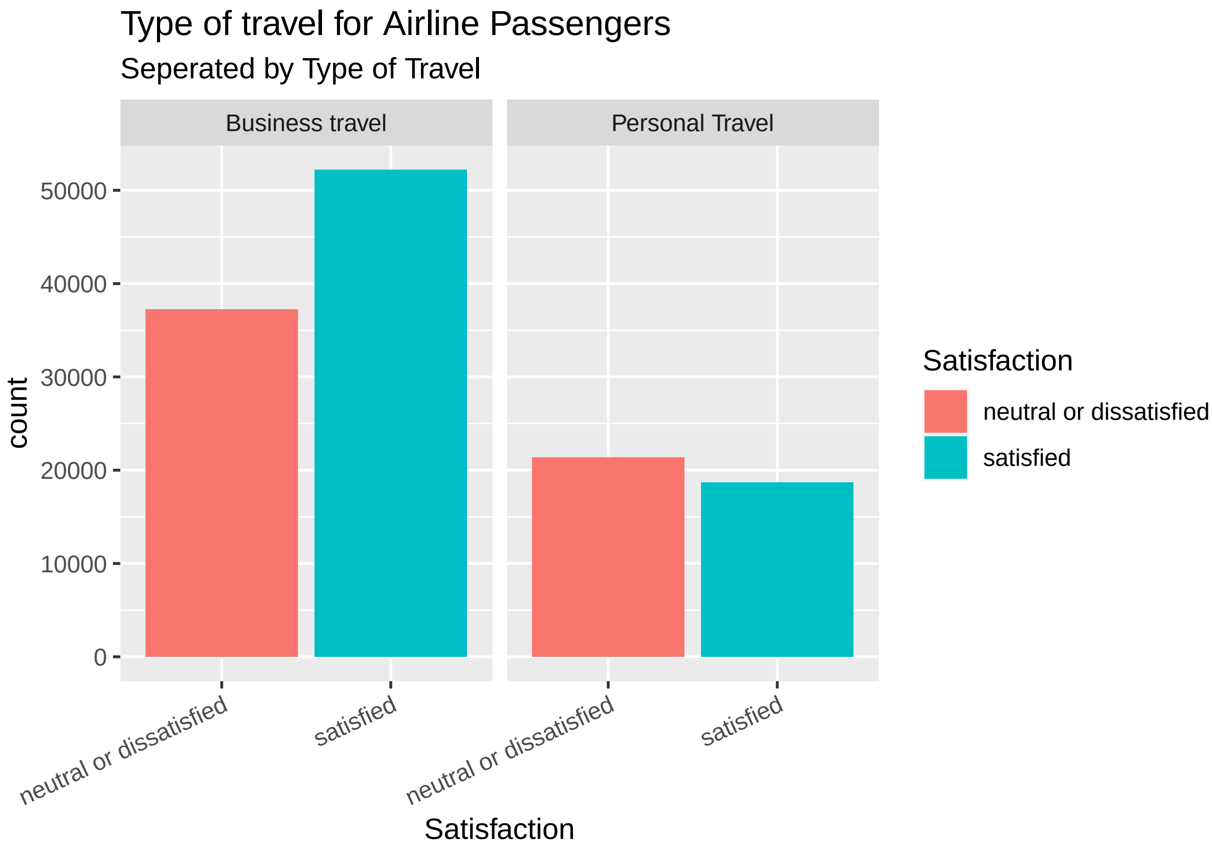


We can see in the boxplots that for the customers who were neutral or dissatisfied, the median age is slightly below the median age for those who were satisfied. Also, the first and third quartiles are higher on the satisfied plot, further proving satisfied customers are usually older age. We can make the conclusion that the more satisfied customers are of older age, while the neutral or dissatisfied customers are of younger age.

Type of Travel

We can now compare the types of travel alongside satisfaction.

```
travelbar <- ggplot(satisfaction, aes(x=Satisfaction, fill=Satisfaction)) +
  geom_bar() +
  facet_wrap(~Type_of_travel) +
  labs(title="Type of travel for Airline Passengers",
       subtitle='Seperated by Type of Travel')

travelbar + theme(axis.text.x = element_text(angle = 25, vjust = 1, hjust=1))
```



We can see in the barplots how satisfaction levels compare based on the travel types. For business travel, more customers were satisfied than not, where the opposite is true for customers who traveled personally. This is noteworthy and based on the visualization it could be inferred that those who travel for business related purposes are more likely to be satisfied with their flight.

It would be interesting to perform chi-squared test to test for statistical significance of this result, with a null hypothesis being there is no significant difference in satisfaction between those to travel for business purposes and those who travel for personal purposes. The alternative hypothesis would be there is a significant difference in satisfaction between the groups. We can perform the chi-squared test and calculate a p-value, observing the result using a 5% level of significance.

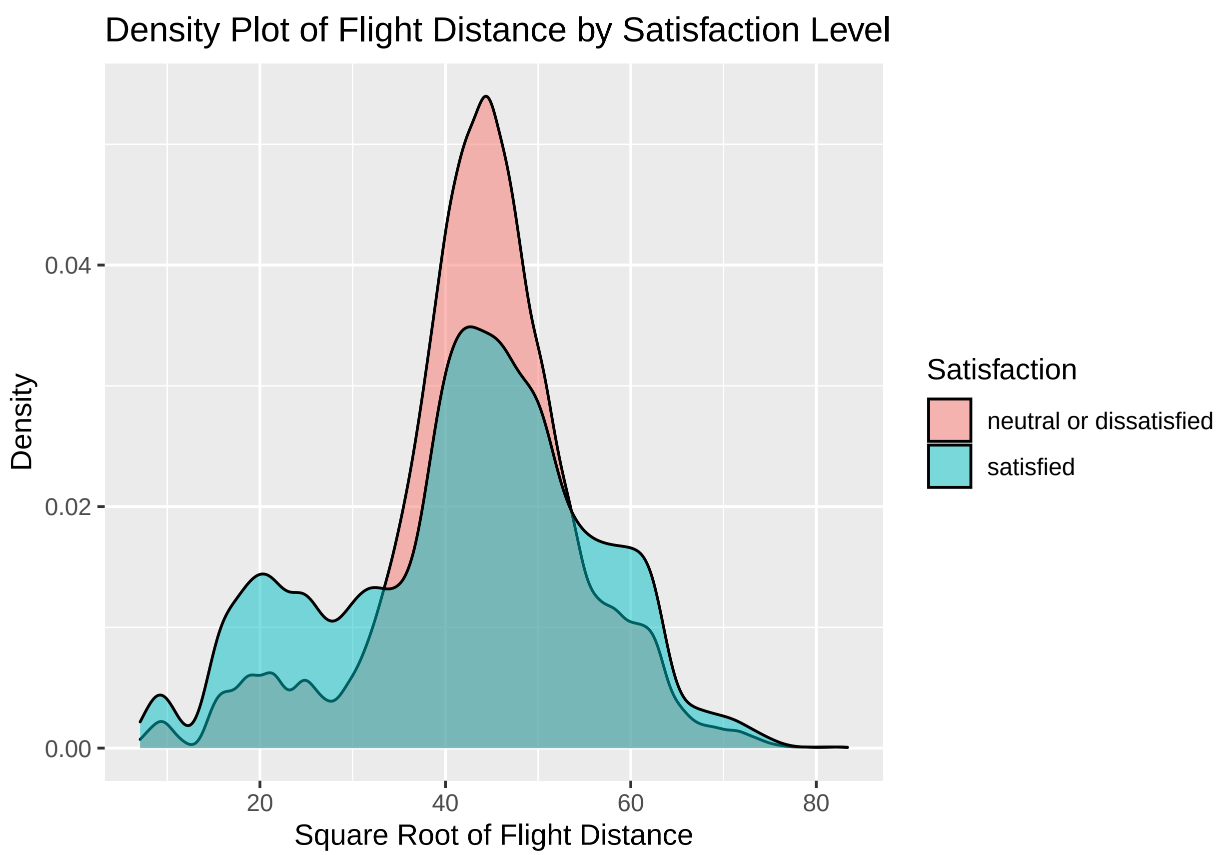
```r
type_of_travel_table <- table(satisfaction$Satisfaction, satisfaction$Type_of_travel)
chi_squared_traveltype <- chisq.test(type_of_travel_table)
chi_squared_traveltype
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  type_of_travel_table
## X-squared = 1535.4, df = 1, p-value < 2.2e-16
```

We obtain a very small p-value of 2.2e-16, which is below the significance level of 0.05. We can safely reject the null hypothesis in favor of the alternative and conclude there exists a statistically significant difference in satisfaction levels between the two groups. By tying this conclusion in with the visualization created, we can conclude that those who travel for business related purposes are more likely to be satisfied with their flight than those who travel for personal purposes.

Flight Distance
Next, we can see how flight distance may affect satisfaction for customers. We will utilize a density plot with the two distributions overlapping each other for comparison.

```
density_flightdist <- ggplot(satisfaction, aes(x = Flight_distance_sqrt, fill = Satisfaction)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Flight Distance by Satisfaction Level",
       x = "Square Root of Flight Distance",
       y = "Density")


density_flightdist
```

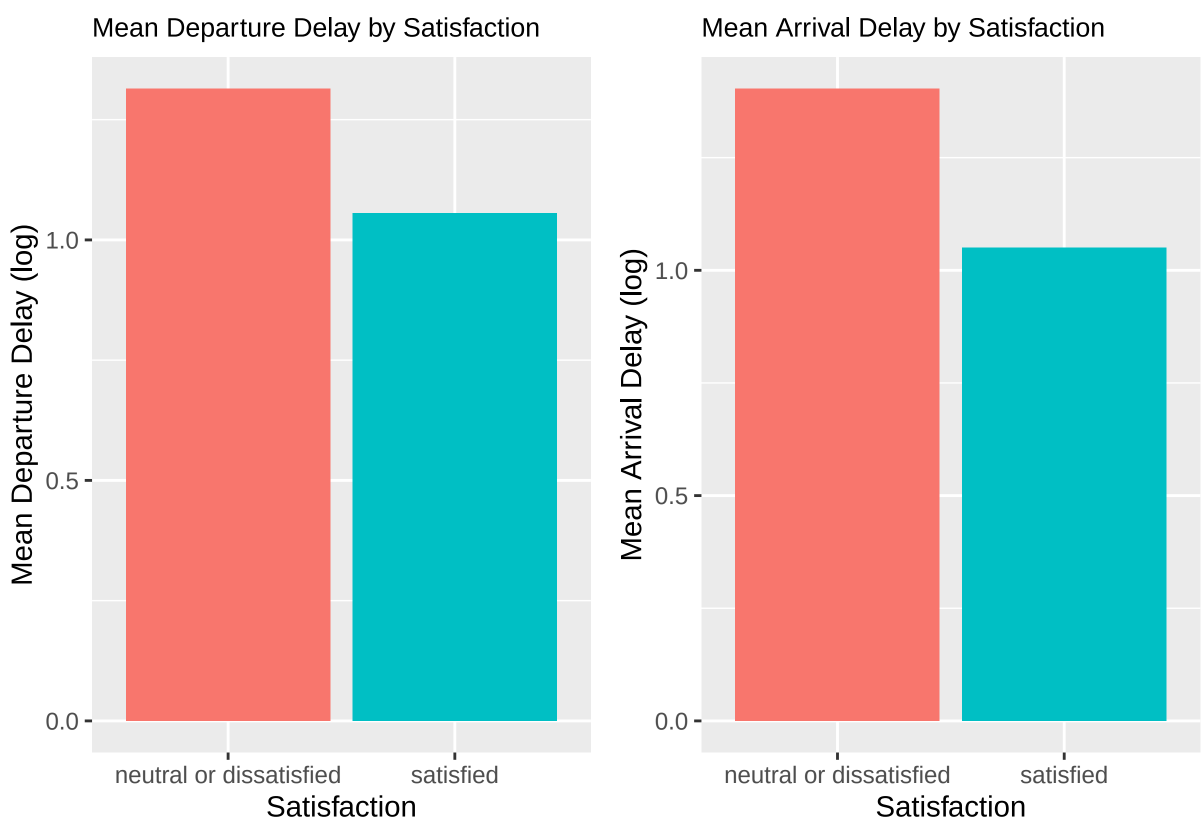


Observing the density plot, there is a noticeable sharp peak for the neutral or dissatisfied group around the 45 square root flight distance mark, where the rest of the graph is flattened out outside of this peak. The satisfied plot is more spread out and resembles more of a normal distribution. One could make the conclusion that customers are satisfied with a variety of different flight distances and it may not necessarily have a direct affect on whether or not they were satisfied the flight.

Delays
Another key variable are the delay times, as delays could significantly impact whether a customer enjoys their flight or not. We can create 2 sets of bar plots for both variables to observe any potential trends.

```
dep_delay_plot <- ggplot(satisfaction, aes(x = Satisfaction, y = Departure_delay_log,
                                           fill=Satisfaction)) +
```

```
  geom_bar(stat = "summary", fun = "mean", show.legend=FALSE) +
  labs(title = "Mean Departure Delay by Satisfaction",
       x = "Satisfaction",
       y = "Mean Departure Delay (log)") +
  theme(plot.title = element_text(size = 10))

arr_delay_plot <- ggplot(satisfaction, aes(x = Satisfaction, y = Arrival_delay_log,
                                          fill=Satisfaction)) +
  geom_bar(stat = "summary", fun = "mean", show.legend=FALSE) +
  labs(title = "Mean Arrival Delay by Satisfaction",
       x = "Satisfaction",
       y = "Mean Arrival Delay (log)") +
  theme(plot.title = element_text(size = 10))
dep_delay_plot+arr_delay_plot
```



We can see that both plots look very similar, with a greater arrival and departure delay for those who were neutral or dissatisfied. This makes sense, as a longer delay would cause customers to feel less satisfied whether that be a departure or arrival delay. For those who were satisfied, their delays were on average much shorter.

Models

We can now proceed to creating models to predict customer satisfaction using a random forest model and logistic regression model. We will utilize a 20/80 test/train split, using the testing dataset to evaluate the accuracy of the model.

```
set.seed(435)
data_split <- initial_split(satisfaction, prop=0.80)
data_train <- training(data_split)
data_test <- testing(data_split)
```

Random Forest

First, we train a random forest regression model. We will use 50 total trees in this random forest model and also have every variable be an option for each tree created, no variables will be removed.

```
data_train$Satisfaction <- as.factor(data_train$Satisfaction)
satis.forest <- randomForest(Satisfaction ~., data=data_train, ntree=50,
                             importance=T, type="classification")
```

After we have created this random forest we can now make predictions and assess its accuracy. Here is a table to observe the accuracy of the predictions, then a calculated error on the testing data.

```
predictions_forest <- predict(satis.forest, newdata = data_test, type = "class")
table_random <- data.frame(True = data_test$Satisfaction, Predicted =
                             as.character(predictions_forest))
table(table_random)
```

```
##                          Predicted
## True                      neutral or dissatisfied satisfied
##   neutral or dissatisfied                   11137       465
##   satisfied                                   663     13633
```

Looking at the table we can observe the errors we make. The model incorrectly classifies 465 customers as satisfied when they were neutral or dissatisfied. The model incorrectly classifies 663 customers as neutral or dissatisfied while they were satisfied. If the model would take into account false positive or negative rates, the model could be adjusted to limit these errors. This model is a good balance between the two.

```
errors <- mean(predictions_forest != data_test$Satisfaction)
errors
```

```
## [1] 0.04355549
```

The error for this model was only 4.36%, a very low error and we can conclude that this model makes fairly accurate predictions.

Logistic Regression

We now can create a logistic regression model to predict customer satisfaction. We make sure to convert categorical variables to factors to create the model. Also, we can use the results from the random forest to determine what features may be of importance in our model and remove features from this logisitic regression model. We can observe below:

```r
importance_scores <- satis.forest$importance
importance_scores
```

```
##                                   neutral or dissatisfied   satisfied
## Gender                                        0.050607045 0.039324866
## Customer_type                                 0.064847188 0.062155743
## Age                                           0.010956395 0.009596063
## Type_of_travel                                0.076862226 0.042006647
## Class                                         0.020503041 0.063062850
## Flight_distance_sqrt                          0.015884126 0.003947463
## Seat_comfort                                  0.160389150 0.143653857
## Departure_arrival_time_convenient             0.014147314 0.020722287
## Food_drink                                    0.043258458 0.016692078
## Gate_location                                 0.005898452 0.048718208
## Inflight_wifi                                 0.026368986 0.018087348
## Inflight_entertainment                        0.133821875 0.042951878
## Online_support                                0.044853987 0.022808653
## Ease_booking                                  0.071887175 0.052167959
## On_board_service                              0.033599612 0.034912537
## Leg_room_service                              0.018507843 0.040133650
## Baggage_handling                              0.050219505 0.021124476
## Checkin_service                               0.043359646 0.009260889
## Cleanliness                                   0.051115282 0.027901409
## Online_boarding                               0.076473135 0.016597639
## Departure_delay_log                           0.004288074 0.002899310
## Arrival_delay_log                             0.005548994 0.003340066
##                                   MeanDecreaseAccuracy MeanDecreaseGini
## Gender                                     0.044439323        1510.7254
## Customer_type                              0.063381633        2215.9262
## Age                                        0.010212143        1529.2726
## Type_of_travel                             0.057803074        1341.7161
## Class                                      0.043761548        1707.5141
## Flight_distance_sqrt                       0.009359115        1705.0158
## Seat_comfort                               0.151254177        6761.5730
## Departure_arrival_time_convenient          0.017743068        1299.2109
## Food_drink                                 0.028739742        2007.1241
## Gate_location                              0.029305873        1051.1580
## Inflight_wifi                              0.021841336         824.9860
## Inflight_entertainment                     0.084159252        9243.0209
## Online_support                             0.032803790        4062.4613
## Ease_booking                               0.061112971        4403.6590
## On_board_service                           0.034313018        2327.5288
## Leg_room_service                           0.030327938        1827.3099
## Baggage_handling                           0.034319504        1213.7034
## Checkin_service                            0.024719447        1299.8672
## Cleanliness                                0.038430267        1364.5371
## Online_boarding                            0.043745699        1796.3410
## Departure_delay_log                        0.003528361         714.3815
## Arrival_delay_log                          0.004340851         757.5926
```

It seems like there are some variables which have very little importance and could be removed from our next model. Features such as Departure_delay_log, Arrival_delay_log, Seat_comfort, Age, and Departure_arrival_time_convenient. These all had "MeanDecreaseAccuracy" scores of about 0.01 or below, which means they may be relatively less important in the overall model. We can safely remove these from our following logistic regression model.

```
log_model <- glm(Satisfaction ~ .-Departure_delay_log-Arrival_delay_log-Seat_comfort-Age
                 -Departure_arrival_time_convenient, data_train, family='binomial')
```

Now that we have created the model, we can observe which features are important in predicting satisfaction. We observe the p-values in the summary of the model.

```
summary(log_model)
```

```
##
## Call:
## glm(formula = Satisfaction ~ . - Departure_delay_log - Arrival_delay_log -
##     Seat_comfort - Age - Departure_arrival_time_convenient, family = "binomial",
##     data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8758  -0.6114   0.1906   0.5364   3.6364
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -6.8870465  0.0723872 -95.142  < 2e-16 ***
## GenderMale                    -0.9600489  0.0181236 -52.972  < 2e-16 ***
## Customer_typeLoyal Customer    1.7308654  0.0260329  66.488  < 2e-16 ***
## Type_of_travelPersonal Travel -0.8542755  0.0251767 -33.931  < 2e-16 ***
## ClassEco                      -0.6057629  0.0226785 -26.711  < 2e-16 ***
## ClassEco Plus                 -0.6729872  0.0351846 -19.127  < 2e-16 ***
## Flight_distance_sqrt          -0.0089694  0.0007387 -12.142  < 2e-16 ***
## Food_drink                    -0.1291245  0.0082102 -15.727  < 2e-16 ***
## Gate_location                  0.0420464  0.0079872   5.264 1.41e-07 ***
## Inflight_wifi                 -0.0672857  0.0097692  -6.888 5.68e-12 ***
## Inflight_entertainment         0.7625693  0.0089005  85.677  < 2e-16 ***
## Online_support                 0.0555097  0.0099504   5.579 2.42e-08 ***
## Ease_booking                   0.3056311  0.0125380  24.376  < 2e-16 ***
## On_board_service               0.2854205  0.0089790  31.788  < 2e-16 ***
## Leg_room_service               0.2302440  0.0076994  29.904  < 2e-16 ***
## Baggage_handling               0.1065393  0.0101325  10.515  < 2e-16 ***
## Checkin_service                0.2644028  0.0075759  34.900  < 2e-16 ***
## Cleanliness                    0.0937973  0.0104504   8.975  < 2e-16 ***
## Online_boarding                0.1473443  0.0109525  13.453  < 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: 142717  on 103588  degrees of freedom
## Residual deviance:  82115  on 103570  degrees of freedom
## AIC: 82153
##
## Number of Fisher Scoring iterations: 5
```

Looking at the summary of the model, every feature has a p-value that is statistically significant and are all important features that highly impact the predictions. We can also calculate the R-squared score, which can help understand the fit of the data.

```r
pR2(log_model)['McFadden']
```

```
## fitting null model for pseudo-r2
```

```
## McFadden
## 0.424628
```

We calculate McFadden's R-squared score to be 0.424628. The indicates that about 42.46% of the variance in the satisfaction of customers is explained by the independent variables. This is moderately high, which indicates the model fits the data well. We can further look into the model by assessing the prediction accuracy on the test data set.

```r
predictions_log <- predict(log_model, newdata = data_test, type = "response")
predictions_log <- ifelse(predictions_log > 0.5, "satisfied", "neutral or dissatisfied")
mean(predictions_log==data_test$Satisfaction)
```

```
## [1] 0.8290602
```

The classification prediction accuracy is around 82.9%, which is moderately high. However, the error rate for the random forest model was slightly lower, so that model may be preferred. Overall both models seem to be accurate with predicting customer satisfaction, but the random forest model takes significantly more time due to the number of trees created and could be avoided for that reason. The random forest model does predict better, but both models predict very well.

Conclusion

In this project, I have performed EDA and machine learning on a passenger satisfaction dataset. Within the EDA, I explore what factors could potentially influence passenger satisfaction through hypothesis testing and data visualization. I then created 2 machine learning models using a random forest and logistic regression to predict passenger satisfaction, testing the models on a testing data set. Overall, this was a beginner personal project to data science techniques.