Northeastern University

ALY6015: Intermediate Analytics

Final Project Report for Predictive modeling on airline customer satisfaction

By: Rohit Raman

Date – 17-02-2023.

Final Project Report.

Motivation for choosing dataset.

I have chosen to work on the Flight Customer satisfaction dataset, The satisfaction of airline customers is a crucial element in determining the success of the airline industry. In an increasingly competitive market, it is necessary for airlines to enhance customer satisfaction levels to retain their existing customers and attract new ones. As a result, a thorough analysis of the airline customer satisfaction dataset can provide valuable insights into the factors that impact customer satisfaction, leading to potential improvements in the customer experience. The dataset is widely accessible and easily available from various sources, including public data repositories, making it a suitable option for conducting logistic regression and prediction analysis. The conclusions derived from the examination of the airline customer satisfaction dataset can be put into practical use by the airline industry. For instance, airlines can leverage the insights to improve their operations and services, thereby increasing customer satisfaction and generating greater revenue.

Credibility of the dataset.

Final Project Report.

I picked the flight customer satisfaction dataset from the Open source from northeastern database repository. This dataset belongs to ***[TJ Klein, December 2019, Airline Customer Satisfaction]***. The link of the dataset is attached in the section of references.

Introduction.

The goal of the project to know what all facilities Flight companies have to improve to operate well in the future, some of the question that we are going to answer in the analysis are.

1. Briefing the overview of the dataset.
2. What all variable has Na value, methods to deals with Na values.
3. What type of variable is satisfaction? If it's not binary then how to change it.
4. What all variable have outlier? How to detect outlier in the dataset.
5. What is the distribution of customer satisfaction
6. What all variable are correlated with each other
7. What is the gender proportion in flight travel?
8. What class type is preferred by the customer during flight travel.
9. What percentage of customer are loyal with the flight travel.
10. Does the customer prefer long distance travel?
11. Generalize the different age group with flight travel using bar plot?
12. What is the relation between Arrival delay in Minute and Departure delay in minutes.
13. What is relation between arrival delay in minutes and flight distance.
14. What is the customer satisfaction on Gate location?
15. What is the customer satisfaction on seat comfort?

16. Build two different model for the customer satisfaction as response variable with all other different variable (including correlated variable) and compare it with other model which do not have correlated variable using Anova and chi square testing and summarize the result which model is good and why? (**GLM, Anova, Chi Square Testing**).
17. **Computing logistic regression** for the **prediction of customer satisfaction (Binary variable)** by splitting the dataset (training, testing) and analyses the result using ROC curve, accuracy, specificity etc.
18. What all **areas** does **airline** need to **improve** for better **customer satisfaction?**

EDA (Exploratory data analysis)

1. Basic overview of the dataset.

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

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

Observation.

The data has various numerical and categorical data. It has 25976 observations with 25 variables such as id, Gender, Customer. Type, Age, Type of travel, Class, Flight Distance, WI-FI services etc. Each observation in the data set referees to the details about the flight and its services.

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

Figure 2: Summary of the dataset.

Observations:

Figure 2 represents the summary of the customer satisfaction dataset. Here, the we can see the mean median mode of each variable individually, let say if we want to know the average delay in the flight departure, then from the above we figure we cay the average delay in departure is round 14 min. Likewise anyone interested in knowing the 1 quarter value of flight distance then it would be around 414 miles. The age range 7 and 85 years.

Moreover, we will convert the satisfaction with “yes” and “no” (**binary variable**) for Satisfied and Neutral or dissatisfied.

2. Replacing the Na value with the median value and sub setting the data (Data Cleaning).

```
airline_satisfaction <- subset(airline_satisfaction,select = -X)
airline_satisfaction$Departure.Delay.in.Minutes = as.numeric(airline_satisfaction$Departure.Delay.in.Minutes)
airline_satisfaction$Arrival.Delay.in.Minutes[is.na(airline_satisfaction$Arrival.Delay.in.Minutes)] <-0
any(is.na(airline_satisfaction))
```

```
## [1] FALSE
```

Figure 3: Removal of Na value with the median values.

Observation:

Here from summary analysis (Figure 2) of the dataset we can see around 83 observations in the variable arrival delay in minutes have NA values, considering the presence of the Na value we have replaced the NA value with its own median value which is 0. Now, from figure 3 we can the dataset does not contain any Na values. Also, we have subset the dataset as variable X does not give any meaningful information since it is used for the row identification.

3. Conversion of Satisfaction variable into binary variable.

```
airline_satisfaction <- airline_satisfaction %>%
       mutate(across(where(is.character), ~ as.factor(str_squish(str_to_title(.))))) %>%
       mutate(
              satisfaction = str_replace_all(satisfaction, "Neutral Or Dissatisfied", replacement = "No"),
              satisfaction = str_replace_all(satisfaction, "Satisfied", replacement = "Yes"),
              satisfaction = factor(satisfaction, levels = c("Yes", "No")),
              Arrival_Delay_in_Minutes = as.numeric(str_replace_na(
                     Arrival_Delay_in_Minutes,
                     mean(Arrival_Delay_in_Minutes,
                          na.rm = TRUE)
              ))
       )
```

Figure 4: Code for converting satisfaction variable into binary variable.

```
## Columns: 25
## $ SR                                <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11~
## $ id                                <int> 70172, 5047, 110028, 24026, 119299, ~
## $ Gender                            <fct> Male, Male, Female, Female, Male, Fe~
## $ Customer_Type                     <fct> Loyal Customer, Disloyal Customer, L~
## $ Age                               <int> 13, 25, 26, 25, 61, 26, 47, 52, 41, ~
## $ Type_of_Travel                    <fct> Personal Travel, Business Travel, Bu~
## $ Class                             <fct> Eco Plus, Business, Business, Busine~
## $ Flight_Distance                   <int> 460, 235, 1142, 562, 214, 1180, 1276~
## $ Inflight_wifi_service             <int> 3, 3, 2, 2, 3, 3, 2, 4, 1, 3, 4, 2, ~
## $ Departure.Arrival_time_convenient <int> 4, 2, 2, 5, 3, 4, 4, 3, 2, 3, 5, 4, ~
## $ Ease_of_Online_booking            <int> 3, 3, 2, 5, 3, 2, 2, 4, 2, 3, 5, 2, ~
## $ Gate_location                     <int> 1, 3, 2, 5, 3, 1, 3, 4, 2, 4, 4, 2, ~
## $ Food_and_drink                    <int> 5, 1, 5, 2, 4, 1, 2, 5, 4, 2, 2, 1, ~
## $ Online_boarding                   <int> 3, 3, 5, 2, 5, 2, 2, 5, 3, 3, 5, 2, ~
## $ Seat_comfort                      <int> 5, 1, 5, 2, 5, 1, 2, 5, 3, 3, 2, 1, ~
## $ Inflight_entertainment            <int> 5, 1, 5, 2, 3, 1, 2, 5, 1, 2, 2, 1, ~
## $ On.board_service                  <int> 4, 1, 4, 2, 3, 3, 3, 5, 1, 2, 3, 1, ~
## $ Leg_room_service                  <int> 3, 5, 3, 5, 4, 4, 3, 5, 2, 3, 3, 2, ~
## $ Baggage_handling                  <int> 4, 3, 4, 3, 4, 4, 4, 5, 1, 4, 5, 5, ~
## $ Checkin_service                   <int> 4, 1, 4, 1, 3, 4, 3, 4, 4, 4, 3, 5, ~
## $ Inflight_service                  <int> 5, 4, 4, 4, 3, 4, 5, 5, 1, 3, 5, 5, ~
## $ Cleanliness                       <int> 5, 1, 5, 2, 3, 1, 2, 4, 2, 2, 2, 1, ~
## $ Departure_Delay_in_Minutes        <int> 25, 1, 0, 11, 0, 0, 9, 4, 0, 0, 0, 0~
## $ Arrival_Delay_in_Minutes          <dbl> 18, 6, 0, 9, 0, 0, 23, 0, 0, 0, 0, 0~
## $ satisfaction                      <fct> No, No, Yes, No, Yes, No, No, Yes, N~
```

Figure 5: Conversion of satisfaction variable into binary variable.

Observation:

From figure 4 and 5, represent the process of converting the satisfaction variable into satisfaction variable, here satisfaction variable has been converted into factor variable with two level with Yes or No.

4. Outlier detection in the dataset.

```
as_num<-select_if(airline_satisfaction,is.numeric)%>%select(-id)
as_num_p<-as_num %>% gather(variable,values,1:18 )
options(repr.plot.width = 14, repr.plot.height = 8)
ggplot(as_num_p)+
  geom_boxplot(aes(x=variable,y=values),fill="cadetblue") +
  facet_wrap(~variable,ncol=6,scales="free") +
  theme(strip.text.x = element_blank(),
        text = element_text(size=14))
```

Figure 6: Code for boxplot for outlier analysis.

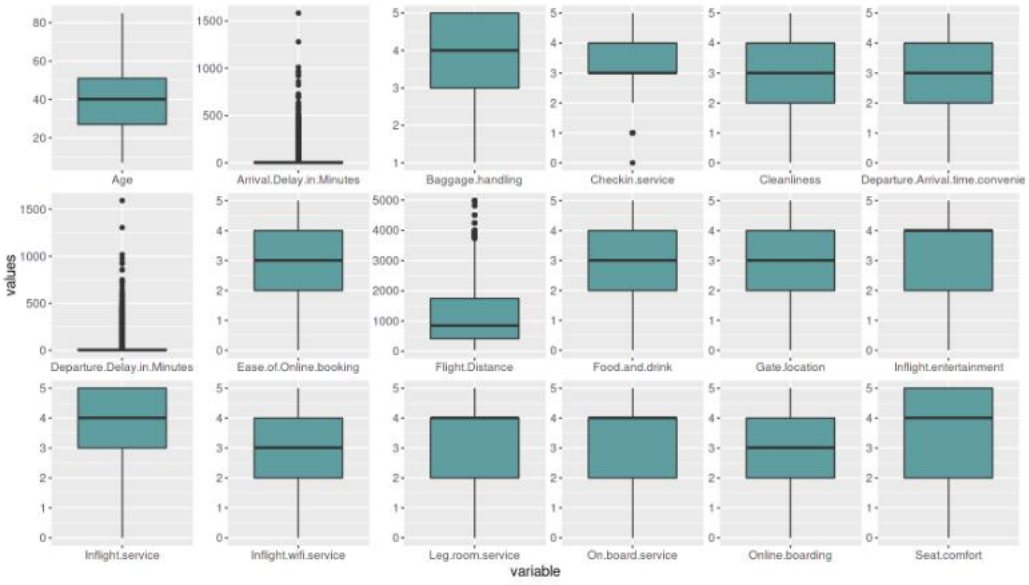


Figure 7: Boxplot for outlier detection.

Observation:

From figure 7, the variables with outliers are departure, arrival, flight distance, and rating on check-in service. I will not impute the outlier from those variables, as the variables are the interest of my further analysis.

5. What is the distribution of Customer Satisfaction.

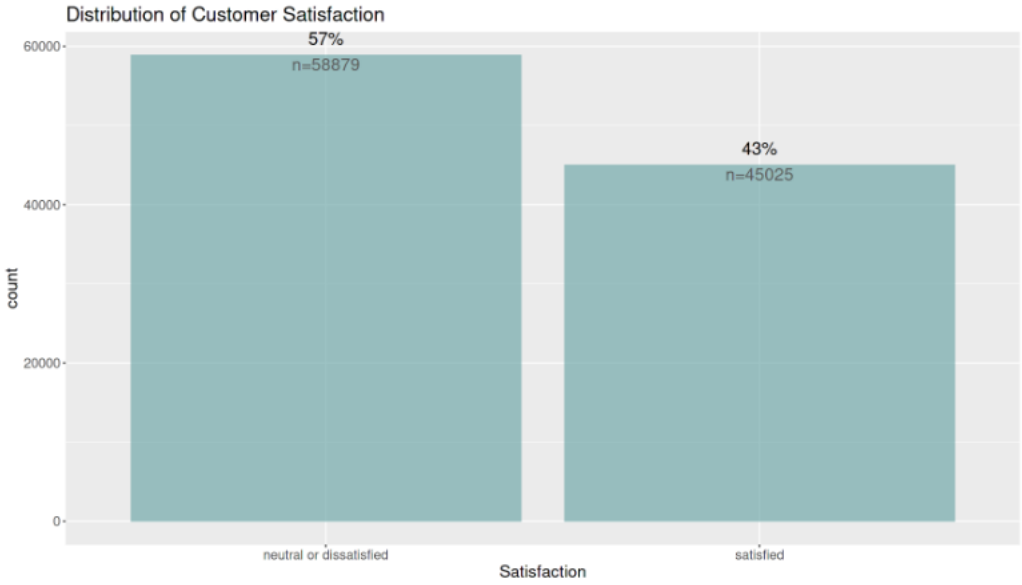


Figure 8: Distribution of Customer Satisfaction.

Observation.

From the figure 8 we have used bar plot to know the proportion of satisfied and unsatisfied customer, around 43 percent of customer are satisfied and around 57 percent of them are unsatisfied.

6. What all variables are correlated with each other.

```r
cor_mat <-
  airline_satisfaction %>%
  select(where(is.numeric), -c(id, SR)) %>%
  cor(use = "pairwise.complete.obs")

corrplot(
  title = "\n\nCorrelation Matrix",
  cor_mat,
  method = "number",
  order = "alphabet",
  type = "lower",
  diag = FALSE,
  number.cex = 0.7,
  tl.cex = 0.8,
  tl.col = "darkgreen",
  addgrid.col = "gray"
)
```

Figure 9: Code for correlation matrix.

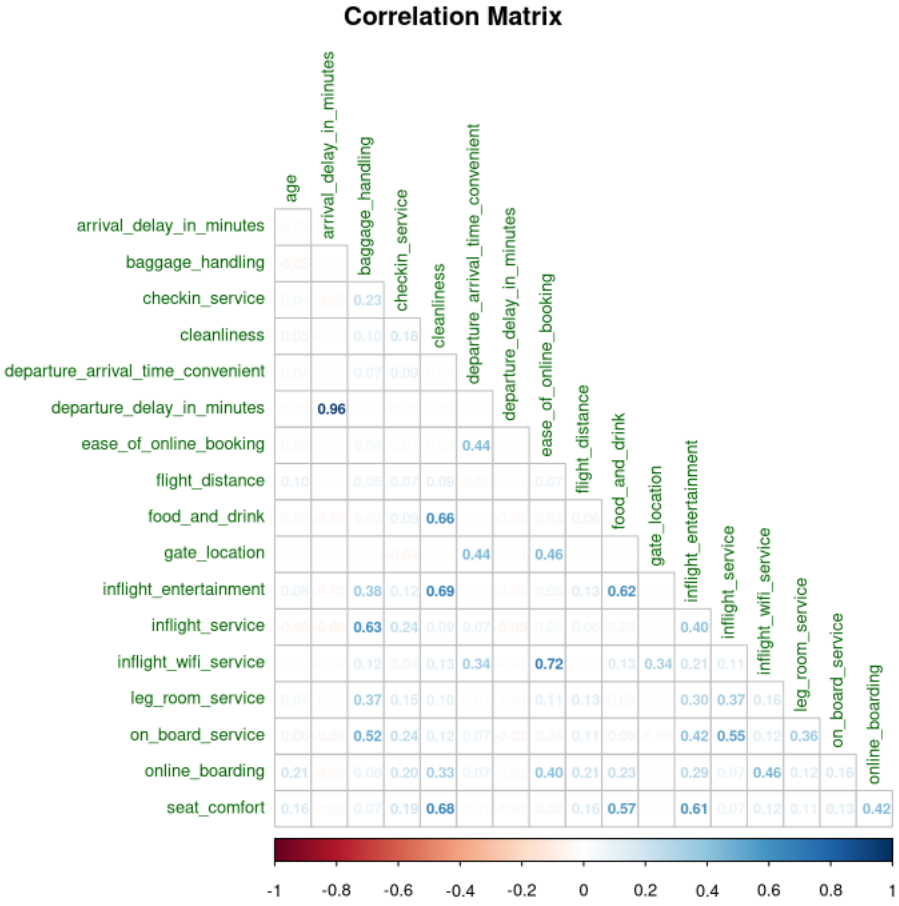


Figure 10: Correlation Matrix analysis.

Observation.

From the figure 10, here we saw departure delay in minutes and arrival delay minutes are correlated with each other, so for regression analysis (logistic) we would consider either one of them.

7. What is the Gender Proportion in flight travel.

```
airline_satisfaction %>%
  group_by(Gender) %>%
  summarize(counts = n()) %>%
  mutate(perc = (counts / sum(counts)) * 100) %>%
  arrange(desc(perc)) %>%
ggplot(aes("", counts)) +
  geom_col(
    position = "fill",
    color = "black",
    width = 1,
    aes(fill = factor(Gender))
  ) +
  geom_text(
    aes(label = str_c(round(perc), "%"), group = factor(Gender)),
    position = position_fill(vjust = 0.5),
    color = "white",
    size = 6,
    show.legend = FALSE,
    fontface = "bold"
  ) +
  coord_polar(theta = "y") +
  scale_fill_manual (values = c("#95b356", "#bda144")) +
  theme_void() +
  labs(
    title = "Proportion of Men to Women",
    subtitle = " Men and Women travel history in term of percentage",
    caption = "Data Source: Airline Passenger Satisfaction",
    fill = ""
  )
```

Figure 10: Code for the pie plot for Gender proportion.

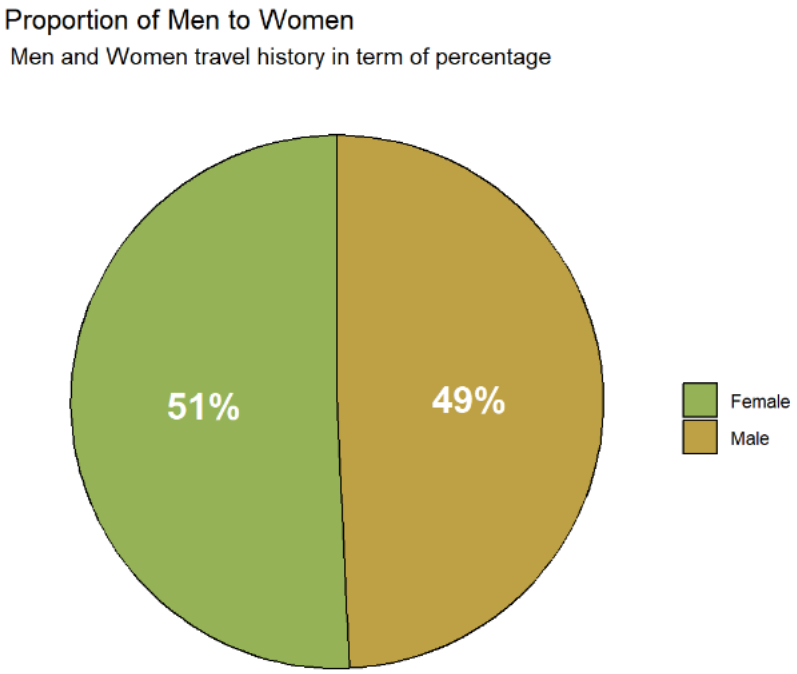


Figure 11: Pie plot for men and women travel history.

Observation.

From Figure 11 we can say, women made a significant contribution in flight travel with a percentage of 51 and men were at around 49 percent.

8. What class type the customer prefer to have during the travel?

```
tree_plot(Class, pal = "RdYlBu")
```

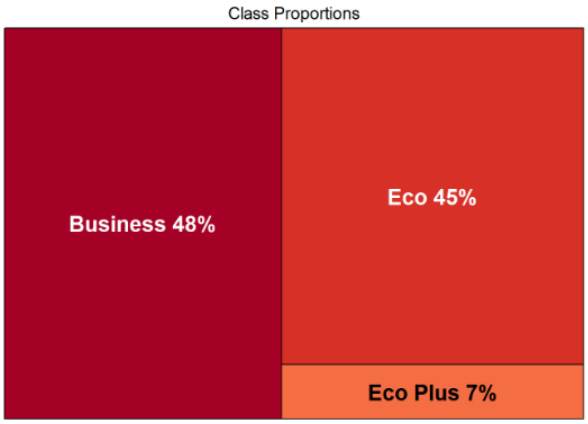


Figure 12: Tree plot for proportion of class.

Observation.

For the proportion of class, we have used the tree plot function and realized that the majority of traveler prefer to travel with business class with a value of 48 percent while 45 travelers prefer to travel with economic class.

9. How many customers are loyal and how many were disloyal with the travel?

```
#Customer type proportion

tree_plot(Customer_Type, pal = c("#F7495D", "#5FB9E7"))
```

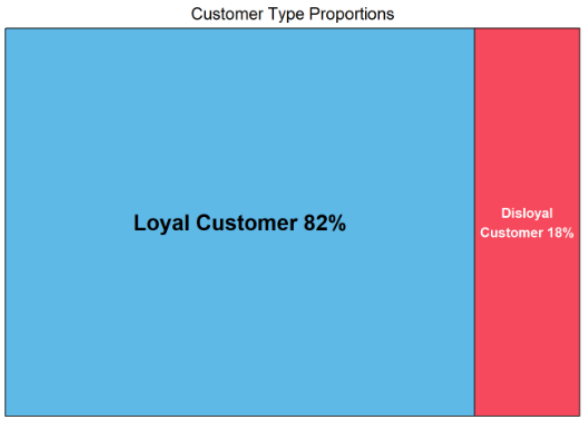


Figure 13: Tree plot for analysis of customer type proportion.

Observation.

Here, we observed that around 82 percent of customer are loyal with the flight travel how 18 percent were disloyal with the travel history.

10. What is the preference of traveler with the distance of travel?

```
airline_satisfaction %>%
  select(Flight.Distance) %>%
  ggplot(aes(Flight.Distance)) +
  geom_histogram(color = "black", fill = "blue", bins = 40, alpha = 0.5) +

  labs(
    title = "Flight Distance Distributions",
    subtitle = "Histogram Plot",
    x = "Flight Distance in miles",
    y = "Count for total ticket booking based on the route"
  )
```

Figure 14: Code for plotting the histogram for the flight distance.

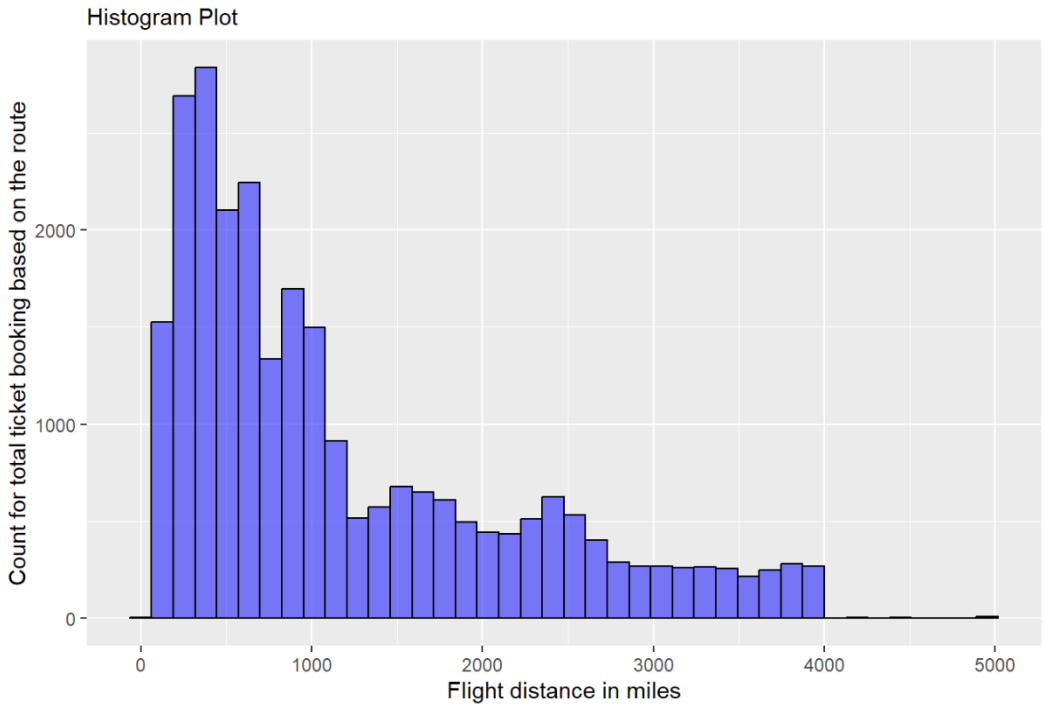


Figure 15. Histogram for flight Distance.

Observation.

From the above histogram we realized that majority of travel has been booked for small flight distance between 200 to 1500 miles, however there are very few numbers of bookings seen after 4000 miles.

Distribution of different age group who travelled with flight

```
airline_satisfaction %>%
  select(Age, Gender) %>%
  arrange(desc(Age)) %>%
  mutate(age_group = case_when(
                               Age >= min(Age) & Age <= 10 ~ glue(min(Age)," -10"),
                               Age > 10 & Age <= 18 ~ "11-18",
                               Age > 18 & Age <= 25 ~ "18-25",
                               Age > 25 & Age <= 35 ~ "26-35",
                               Age > 35 & Age <= 45 ~ "36-45",
                               Age > 45 & Age <= 50 ~ "46-50",
                               Age > 50 & Age <= 64 ~ "51-64",
                               Age > 64 ~ "> 64"
                               ),
    age_group = factor(
      age_group,
      level = c(glue(min(Age)," -10"), "11-18","18-25", "26-35", "36-45", "46-50", "51-64", "> 64")
    )
  ) %>%
  count(age_group, Gender, name = "counts") %>%
  mutate(perct = round(counts/sum(counts),2)) %>%
  ggplot(aes(x = age_group, y = counts)) +
  geom_col(aes(fill = Gender), size = 1) +
  scale_fill_manual(values = c( "Female" = "#40615d", "Male" = "#ebcf34"))+
  labs(
    title = "Satisfaction by Flight Distance and Age",x = "Age",
    y = "Count",
    fill = NULL
  )
```

Figure 16: Code for Bar plot for different age group who preferred to travel with air.

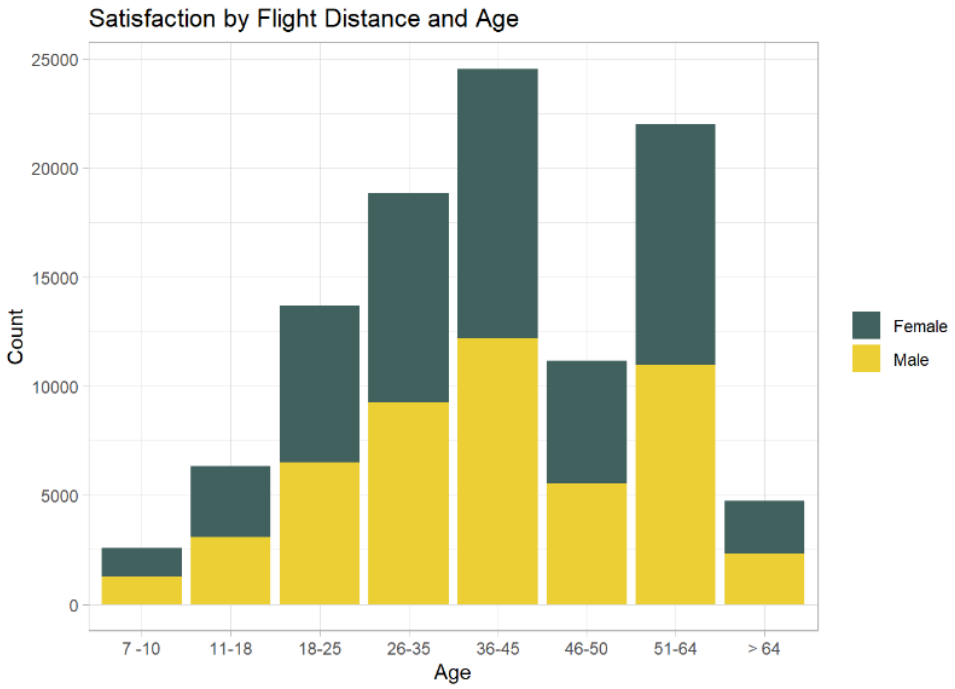


Figure 17: Bar plot for different age group who preferred to travel with air.

Observation:

From the figure 17 the proportion of Female and Male by age is around 50 percent in all the age group eventually.

11. What is the relation between Arrival delay in Minute and Departure delay in minutes.

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

summary(lm_mod_arr_flight_departure)
```

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

Figure 18: Linear regression code for finding the Relation between Arrival Delay in minutes and departure delay in minutes.

Observation:

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

The residual standard error of 10.66 on 103902 degrees of freedom is an estimate of **the standard deviation of the residuals** (the difference between the observed and predicted values of the response). The multiple **R-squared of 0.9238** is the proportion of variation in the response that is explained by **the predictor**, and the adjusted **R-squared** is a **corrected version of R-squared** that adjusts for the number of predictors in the model. **The F-statistic** and its **p-value** test the hypothesis that all **coefficients are equal to zero**, providing a measure of the overall significance of the model. In this case, the **F-statistic** and **p-value** indicate that the model is **very significant**.

12. What is relation between arrival delay in minutes and flight distance.

```
lm_mod_arr_flight_distance <- lm(Arrival_Delay_in_Minutes ~ Flight_Distance, data = airline_satisfaction)

summary(lm_mod_arr_flight_distance)
```

```
##
## Call:
## lm(formula = Arrival_Delay_in_Minutes ~ Flight_Distance, data = airline_satisfaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -15.29  -15.22  -15.04   -2.11 1568.81
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.529e+01  1.866e-01  81.944   <2e-16 ***
## Flight_Distance -9.388e-05  1.202e-04  -0.781    0.435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.64 on 103902 degrees of freedom
## Multiple R-squared:  5.869e-06,  Adjusted R-squared:  -3.756e-06
## F-statistic: 0.6098 on 1 and 103902 DF,  p-value: 0.4349
```

Figure 19: Linear regression code for finding the relation between Arrival delay in minutes and flight distance.

Observation:

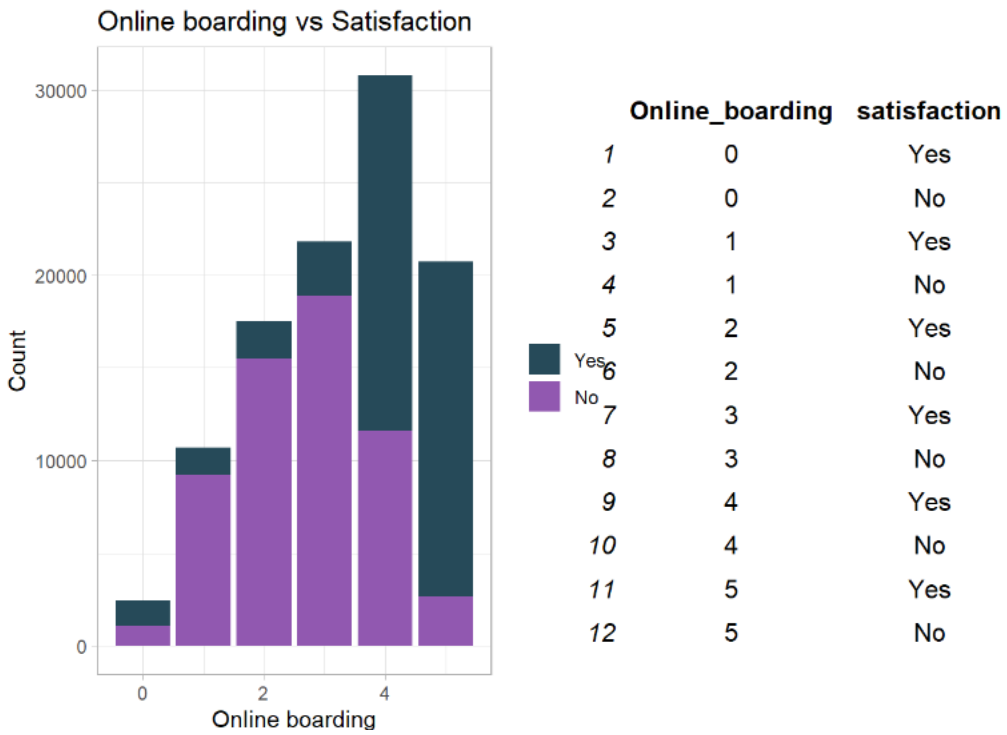
From figure 19, the results show the results of a linear regression analysis where the dependent variable (Arrival_Delay_in_Minutes) is being modeled based on the independent variable (Flight_Distance).The **residuals section** shows **the distribution** of the **difference between the predicted values and the actual values, with the minimum, 1st quartile, median, 3rd quartile, and maximum values**. The coefficients section shows the regression coefficients, including the intercept and the estimated coefficient for the Flight_Distance variable. The **estimate of the intercept** is **15.29** and the **estimate of the coefficient for Flight_Distance is -0.00009388**. **The standard error** for the intercept and Flight_Distance **is 0.1866 and 0.0001202** respectively. **The t-value** and **p-value** show the results of the hypothesis test of whether each coefficient is significantly different from zero. **The p-value for the Flight_Distance coefficient is 0.435, which is not significant at the 0.05 level and suggests that the relationship between Flight_Distance and Arrival_Delay_in_Minutes is not statistically significant.** The residual standard error of 38.64 on 103902 degrees of freedom, and an adjusted R-squared value of -3.756e-06, suggest that the model **explains a very small amount of the variability in Arrival_Delay_in_Minutes**.

13. What is the customer satisfaction on Gate location?

```
table_boarding <- tableGrob(airline_satisfaction %>%
  select(Online_boarding, satisfaction) %>%
  count(Online_boarding, satisfaction), theme = ttheme_minimal())


  airline_satisfaction %>%
  select(Online_boarding, satisfaction) %>%
  count(Online_boarding, satisfaction) %>%
ggplot(aes(x = Online_boarding, y = n, fill = satisfaction)) +
  geom_col(size = 1) +
  scale_fill_manual(values = c("#264a59", "#9158b0"))+
  theme(strip.background = element_rect(fill = "#673B38"),
        strip.text = element_text(face = "bold")) +
  labs(
    title = "Online boarding vs Satisfaction",
    x = "Online boarding" ,
    y = "Count",
    fill = ""
  ) + table_boarding
```

Figure 20: Code for bar plot for online boarding vs Satisfaction.



	Online_boarding	satisfaction
1	0	Yes
2	0	No
3	1	Yes
4	1	No
5	2	Yes
6	2	No
7	3	Yes
8	3	No
9	4	Yes
10	4	No
11	5	Yes
12	5	No

Figure 21: Gate Location and satisfaction bar plot.

Observation:

Looking at the gate location satisfaction, we can say majority of traveler are not satisfied with the gate location offered by the airport.

14. What is the customer satisfaction on seat comfort?

```
airline_satisfaction %>%
  select(Seat_comfort, satisfaction, Class) %>%
  count(Seat_comfort, satisfaction, Class) %>%
ggplot(aes(x = Seat_comfort, y = n, fill = satisfaction)) +
  geom_col(size = 1) +
  scale_fill_manual(values = c("#7d9669", "#453d2a"))+
  facet_wrap(vars(Class))+
  theme(strip.background = element_rect(fill = "#673B38"),
        strip.text = element_text(face = "bold")) +
  labs(
    title = "Satisfaction by Seat Comfort and Class",
    x = "Seat_comfort" ,
    y = "Count",
    fill = ""
  )
```

Figure 22: Code for seating comfort and satisfaction analysis using bar plot.

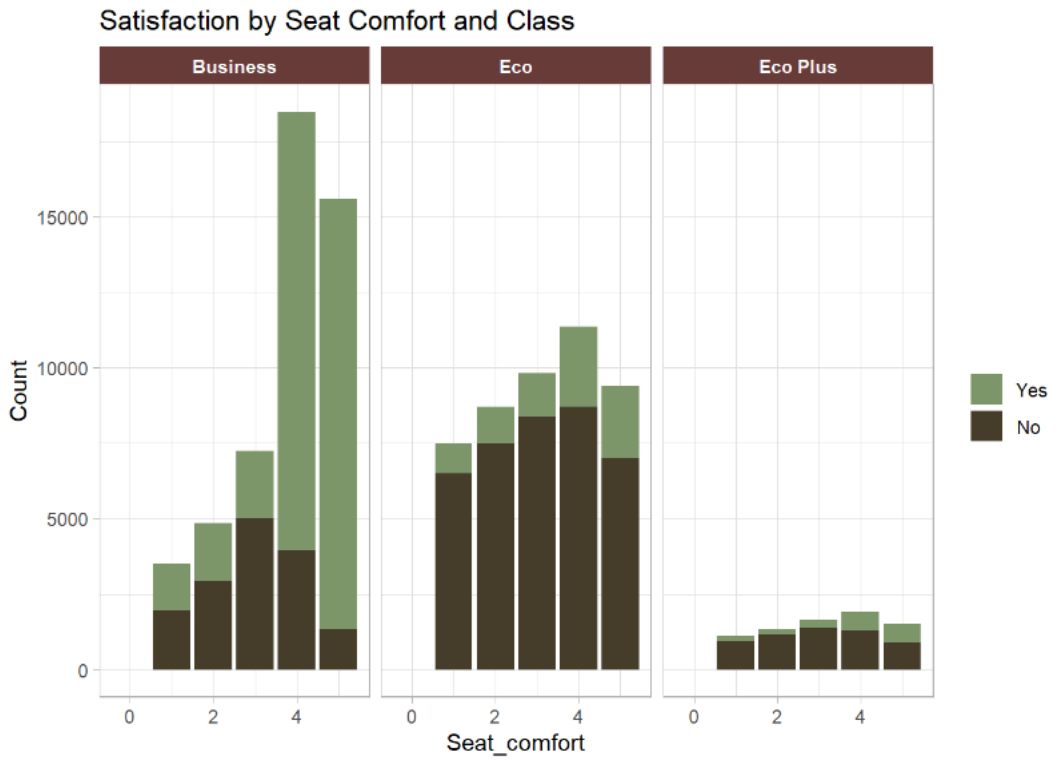


Figure 23: Visualization for seat comfort and satisfaction analysis.

Observation:

Here we can say the Majority of customer felt discomfort while travelling as the count for satisfaction as “No” is significant high.

15. Generalized linear Model and Chi Square Testing.

For Generalized linear model I have created two model first and one by one we have applied Glm function to assess the model performance.

```
set.seed(123)
model1 <- glm(satisfaction ~  Customer.Type + Age +

                Type.of.Travel + Class + Flight.Distance + Inflight.wifi.service +

                Departure.Arrival.time.convenient + Ease.of.Online.booking +
                Gate.location + Food.and.drink + Online.boarding + Seat.comfort +
                Inflight.entertainment + On.board.service + Leg.room.service +
                Baggage.handling + Checkin.service + Inflight.service +
                Cleanliness
              ,
               data = ds, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model1)
```

Figure 24: Model1 building using Glm function.

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142189  on 103903  degrees of freedom
## Residual deviance:  37171  on 103831  degrees of freedom
## AIC: 37317
##
## Number of Fisher Scoring iterations: 17
```

Figure 25: Model 1 statistics after applying glm function

Observation:

The output is from a logistic regression model where the response variable is "satisfaction" and the predictor variables are "Customer.Type," "Age," "Type.of.Travel," "Class," "Flight.Distance," "Inflight.wifi.service," "Departure.Arrival.time.convenient," "Ease.of.Online.booking," "Gate.location," "Food.and.drink," "Online.boarding," "Seat.comfort," "Inflight.entertainment," "On.board.service," "Leg.room.service," "Baggage.handling," "Checkin.service," and "Cleanliness."

The "Deviance Residuals" are a measure of the model's goodness of fit, with smaller residuals indicating a better fit.

The "Coefficients" table shows the estimated coefficient for each predictor variable and the associated standard error, z-value, and p-value. The p-values are used to test the null hypothesis that the corresponding coefficient is zero, and if the p-value is less than a significance level (such as 0.05), it suggests that the corresponding predictor variable is significantly related to the response variable. A small p-value indicates strong evidence against the null hypothesis and in favor of the alternative hypothesis that the predictor variable has a non-zero effect on the response variable. The z-value is the ratio of the estimated coefficient to its standard error and provides a measure of how many standard deviations the coefficient is away from zero.

GLM for Model 2

```
set.seed(123)
model2 <- glm(satisfaction ~ Gender+ Customer.Type + Age +

                Type.of.Travel + Class +
                Departure.Arrival.time.convenient + Ease.of.Online.booking +
                Online.boarding + Seat.comfort +
                Inflight.entertainment + On.board.service + Leg.room.service +
                Baggage.handling + Checkin.service + Inflight.service +Cleanliness+
                    Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes,
              data = ds, family = "binomial")
summary(model2)
```

Figure 24: Model1 building using Glm function.

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142189  on 103903  degrees of freedom
## Residual deviance:  50284  on 103844  degrees of freedom
## AIC: 50404
##
## Number of Fisher Scoring iterations: 11
```

Figure 25: Statistics analysis after applying Glm on model2.

Observation:

This output is from a logistic regression model in R. The model was fit using the "glm" function, with the formula "satisfaction ~ Gender + Customer.Type + Age + Type.of.Travel + Class + Departure.Arrival.time.convenient + Ease.of.Online.booking + Online.boarding + Seat.comfort + Inflight.entertainment + On.board.service + Leg.room.service + Baggage.handling + Checkin.service + Inflight.service + Cleanliness + Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes". The family argument is set to "binomial", indicating that this is a logistic regression model. The data used to fit the model is stored in the "ds" object.

The output provides a summary of the model coefficients and their associated statistics, such as standard errors, z-values, and p-values. These statistics can be used to assess the significance of each predictor in explaining the response (satisfaction). For example, a predictor with a low p-value (e.g., < 0.05) indicates that it has a significant effect on the response.

16. Analysis of variance using two Glm models (model1 and model 2) using chi-square test to determine which model provide better fit. But why we used Anova function for two Glm model using chi-square test to determine best fit model.

The **ANOVA function** in R to compare **two or more generalized linear models** (GLMs) using the **chi-square test** to determine which model is the best fit.

Here, in this example, the output shows the results of an Analysis of Deviance Table, which compares two generalized linear models (GLMs): Model 1 and Model 2. The hypothesis being tested is whether adding the predictors in Model 2 significantly improves the fit of the model compared to Model 1, which includes fewer predictors.

The null hypothesis(H0): Difference in deviance between the two models is not significant.

The alternative hypothesis (H1): the Difference in deviance is significant

Anova Test.

```
anova(model1, model2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: satisfaction ~ Customer.Type + Age + Type.of.Travel + Class +
##     Flight.Distance + Inflight.wifi.service + Departure.Arrival.time.convenient +
##     Ease.of.Online.booking + Gate.location + Food.and.drink +
##     Online.boarding + Seat.comfort + Inflight.entertainment +
##     On.board.service + Leg.room.service + Baggage.handling +
##     Checkin.service + Inflight.service + Cleanliness
## Model 2: satisfaction ~ Gender + Customer.Type + Age + Type.of.Travel +
##     Class + Departure.Arrival.time.convenient + Ease.of.Online.booking +
##     Online.boarding + Seat.comfort + Inflight.entertainment +
##     On.board.service + Leg.room.service + Baggage.handling +
##     Checkin.service + Inflight.service + Cleanliness + Departure.Delay.in.Minutes +
##     Arrival.Delay.in.Minutes
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1    103831      37171
## 2    103844      50284 -13   -13113 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 26: Analysis of Variance on model1 and model2

Observations:

The **p-value** of the **Chi-squared test is less than 0.001**, which indicates strong evidence against the null hypothesis. Therefore, **we reject the null hypothesis** and **conclude that difference in deviance is significant.**

The residual deviances of **Model 1** and **Model 2** are **37317 and 50404**, respectively. The **AIC value of Model 1 is 37316.9**, **while that of Model 2 is 50403.** The AIC value is an **information criterion that estimates the quality of a statistical model**. The model with **the lower AIC value is considered to be a better fit for the data.** In this case, Model 1 **has a lower AIC value**, which suggests that it is a **better fit for the data. However,** since the Chi-squared test indicates that **Model 2 has a significantly better fit**, **we should choose Model 2 over Model 1**.

In summary, the results suggest that **Model 2 is a better fit for the data**, even though it has a higher **AIC value than Model 1.** The Chi-squared test indicates that **the difference in deviance between the two models is significant**, and **Model 2 includes more predictors than Model 1, which is likely the reason for the better fit**.

Logistic Regression and prediction of customer satisfaction.

17. Model prediction and data split.

Before making the prediction, I will split the dataset in training and testing data in the proportion of ¾.

Remember on the correlation analysis we realized that the **Arrival delay** and **departure delay** are **highly correlated** so before making a prediction I will eliminate one of those.

```
#Model
#Create a new data frame for building model

airline_satisfaction_for_model <- airline_satisfaction %>%
  select(-c(
    id,
    SR
  ))
```

Figure 26: Code for Eliminating id and Sr before making prediction.

```
set.seed(31967)

airline_satisfaction_split <- initial_split( airline_satisfaction_for_model, prop = 3/4, strata = satisfaction)

train_data <- training(airline_satisfaction_split)
test_data  <- testing(airline_satisfaction_split)
```

Figure 27: Splitting the data in training and testing.

```
set.seed(31967)
fold_cv <- vfold_cv(train_data, times = 10, apparent = TRUE)
```

Figure 28: Code for generating the 10 folds

Now **model building** will be done **our aim is to penalize the coefficient**. logistic regression, **the coefficients represent** the change in **the log odds** of the **response variable** for a **one-unit change** in the **predictor**. However, large **coefficients** can lead to **overfitting and unstable models**. To prevent this, **penalization techniques** are used to **shrink the coefficients towards zero**. These techniques **balance model fit** and complexity by adding a penalty term to the log-likelihood function that discourages large coefficients. **The goal is to find a balance between a good fit to the data and stability, which can lead to improved predictive performance and reduced variance**.

```
# Building  the model using glmnet
logis_mod <-
  logistic_reg(penalty = tune(), mixture = 1) %>%
  set_engine("glmnet")
```

```
# Create the logistic_ recipe  variable and workflow. Workflow will help in bundling with modelling and post processin reque
st.

logis_recipe <-
  recipe(satisfaction ~ ., data = train_data) %>%
  step_dummy(all_nominal_predictors(), -all_outcomes()) %>%
  step_zv(all_numeric()) %>%
  step_normalize(all_numeric()) %>%
  step_corr(all_numeric_predictors(), threshold = .7)

#step_corr(): It will remove variables that have large absolute correlations with other variables that is only for glmnet en
gine.

#Create the workflow (glmnet)

logis_workflow <-
  workflow() %>%
  add_model(logis_mod) %>%
  add_recipe(logis_recipe)

logis_workflow
```

Figure 29: Code for building the model and creating the recipe variable.

```
set.seed(31967)
logis_res <-
  logis_workflow %>%
  tune_grid(fold_cv,
            grid = logis_reg_grid,
            control = control_grid(save_pred = TRUE),
            metrics = metric_set(roc_auc))
```

Figure 30: Code for training and tuning the model.

```
#Set the best model (glmnet)
logis_best <-
  logis_res %>%
  show_best() %>%
  arrange(desc(mean)) %>%
  dplyr::slice(2)

logis_best
```

```
## # A tibble: 1 x 7
##   penalty .metric .estimator  mean     n std_err .config
##     <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00108 roc_auc binary     0.923    10 0.00124 Preprocessor1_Model02
```

```
logis_auc <-
  logis_res %>%
  collect_predictions(parameters = logis_best) %>%
  roc_curve(satisfaction, .pred_Yes) %>%
  mutate(model = "Logistic Regression")

autoplot(logis_auc)
```

Figure 31: Code for setting the best model.

```
set.seed(31967)
final_logis_res <-
  logis_workflow %>%
  finalize_workflow(logis_best) %>%
  last_fit(airline_satisfaction_split)

final_logis_res
```

```
## # Resampling results
## # Manual resampling
## # A tibble: 1 x 6
##   splits               id              .metrics .notes   .predict~1 .workflow
##   <list>               <chr>           <list>   <list>   <list>     <list>
## 1 <split [77927/25977]> train/test split <tibble> <tibble> <tibble>   <workflow>
## # ... with abbreviated variable name 1: .predictions
```

```
collect_metrics(final_logis_res)
```

```
## # A tibble: 2 x 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.870 Preprocessor1_Model1
## 2 roc_auc  binary         0.923 Preprocessor1_Model1
```

Figure 32: Code for final regression fit.

```
collect_predictions(final_logis_res) %>%
  conf_mat(satisfaction, .pred_class) %>%
pluck(1) %>%
  as_tibble() %>%
  ggplot(aes(Truth, Prediction, alpha = n)) +
  geom_tile(show.legend = FALSE, fill = "#84c746") +
  geom_text(aes(label = n), colour = "#db2525", alpha = 1, size = 7) +
  scale_x_discrete(position = "top", limits = c("Yes","No"))
```

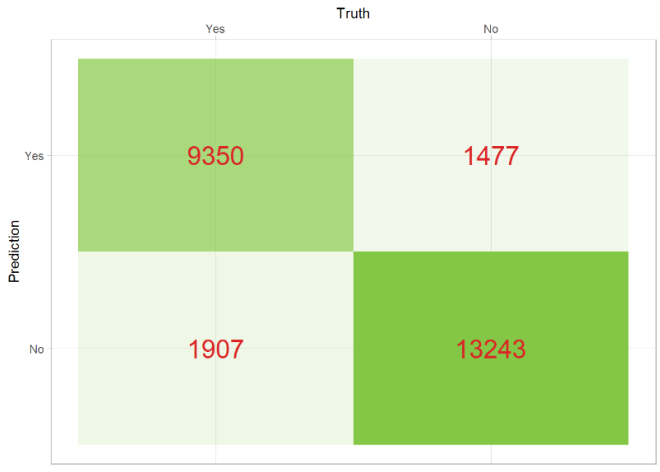


Figure 33: Code for confusion matrix.

```
collect_predictions(final_logis_res) %>%
  metrics(satisfaction, .pred_class) %>%
  select(-.estimator) %>%
  filter(.metric == "accuracy")
```

```
## # A tibble: 1 x 2
##   .metric  .estimate
##   <chr>        <dbl>
## 1 accuracy     0.870
```

Figure 34: Model accuracy.

```
logis_auc <-
  logis_res %>%
  collect_predictions(parameters = logis_best) %>%
  roc_curve(satisfaction, .pred_Yes) %>%
  mutate(model = "Logistic Regression")

autoplot(logis_auc)
```

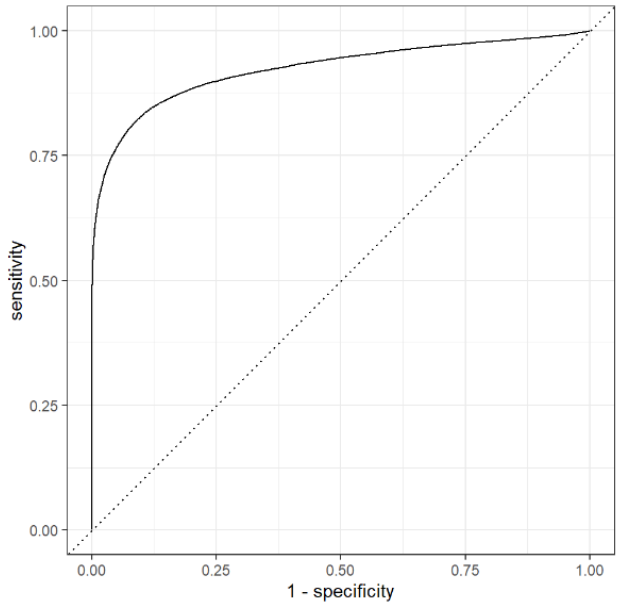


Figure 36: Roc Curve.

```
summary(confusion_matrix_logis)
```

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

Figure 37: Summary of confusion matrix.

Observation:

Figure 27 to 37 represent the steps for logistic regression, here I have created the training and testing in a proportion of 3:4 and created a fold of 10. The **idea behind using folds** in data validation is to get a better estimate of the model's performance by reducing the variance in the evaluation metric. By dividing the data into "**folds**" and training the model on k-1 folds and testing it on the remaining fold, we can get k different performance measurements. This can help reduce the risk of **overfitting**, as the model is tested on data it hasn't seen before. Additionally, by averaging the performance measurements from each fold, we can get a more robust estimate of the **model's performance** on unseen data. In summary, using folds in data validation provides a more robust and accurate estimate of the model's performance, reducing the risk of overfitting and increasing the model's generalization ability. Moreover, I have taken the prediction of model using the

confusion matrix, The. metric column contains the name of the evaluation metric and in this case, it is "accuracy". The .estimate column contains the value of the metric and in this case, it is 0.870, which represents an **accuracy of 87%.** This data can be interpreted as the accuracy of a model on a classification problem, where accuracy is the proportion of correct predictions made by the model. A value of **0.870 indicates** that the model is able to **correctly predict the target** class **87% of the time**. From the confusion matrix summary (figure 37) I can say

This data can be interpreted as the accuracy of a model on a classification problem, where accuracy is the proportion of correct predictions made by the model. A value of 0.870 indicates that the model is able to correctly predict the target class 87% of the time.
The metrics included in this table are commonly used to evaluate binary classification models. Here are some brief descriptions of **each metric**:

accuracy: proportion of correct predictions made by the model.

sens: sensitivity or true positive rate, the proportion of positive cases that are correctly identified.

spec: specificity or true negative rate, the proportion of negative cases that are correctly identified.

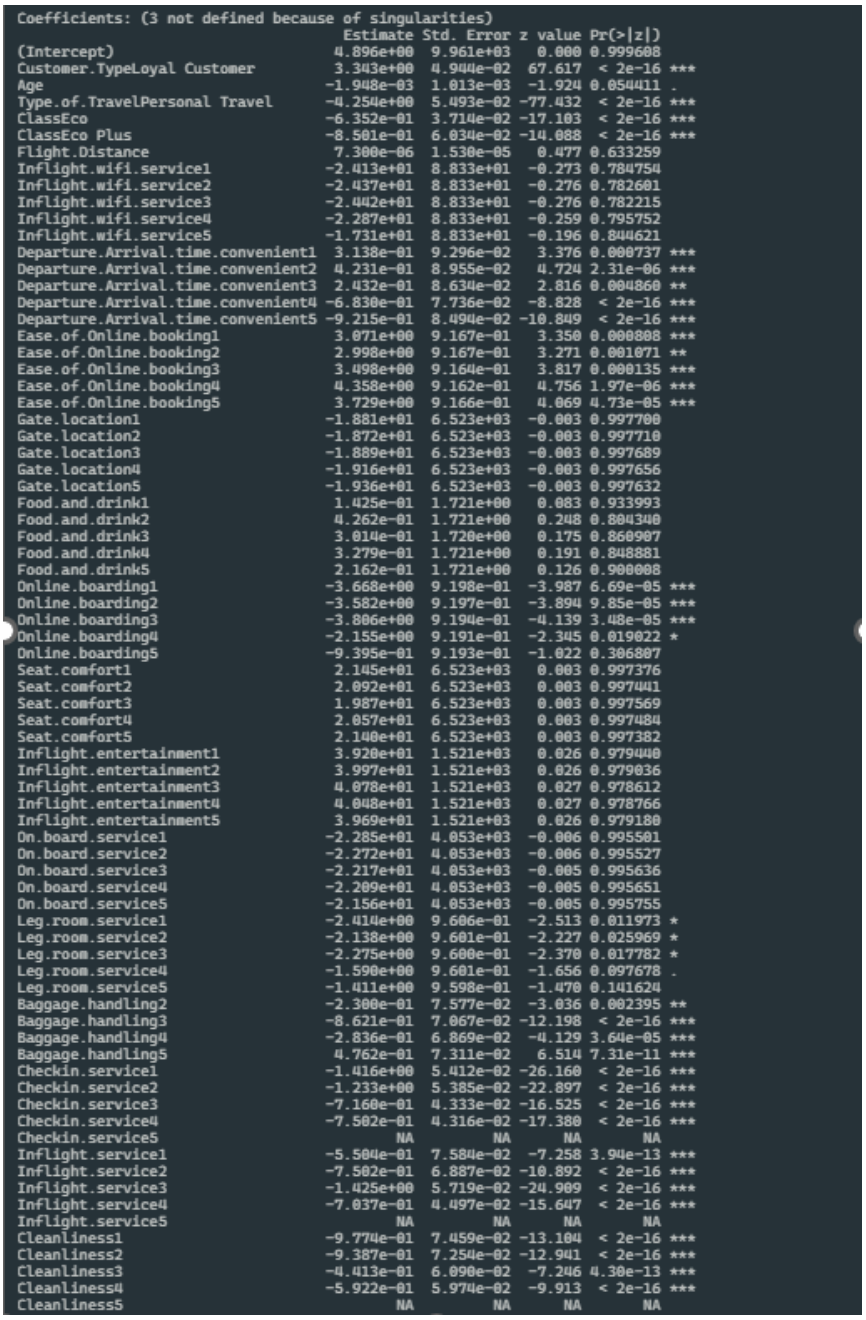
npv: negative predictive value, the proportion of true negative predictions among all negative predictions.

bal_accuracy: balanced accuracy, an average of sensitivity and specificity, taking into account both false positive and false negative predictions.

precision: the proportion of true positive predictions among all positive predictions.

By using multiple evaluation metrics, it's possible to get a comprehensive understanding of the model's performance and identify its strengths and weaknesses.

18. What all areas does airline need to improve for better customer satisfaction?



```
Coefficients: (3 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    4.896e+00  9.961e+03   0.000 0.999608
Customer.TypeLoyal Customer    3.343e+00  4.944e-02  67.617  < 2e-16 ***
Age                           -1.948e-03  1.013e-03  -1.924 0.054411 .
Type.of.TravelPersonal Travel -4.254e+00  5.493e-02 -77.432  < 2e-16 ***
ClassEco                      -6.352e-01  3.714e-02 -17.103  < 2e-16 ***
ClassEco Plus                 -8.501e-01  6.034e-02 -14.088  < 2e-16 ***
Flight.Distance                7.300e-06  1.530e-05   0.477 0.633259
Inflight.wifi.service1        -2.413e+01  8.833e+01  -0.273 0.784754
Inflight.wifi.service2        -2.437e+01  8.833e+01  -0.276 0.782601
Inflight.wifi.service3        -2.442e+01  8.833e+01  -0.276 0.782215
Inflight.wifi.service4        -2.287e+01  8.833e+01  -0.259 0.795752
Inflight.wifi.service5        -1.731e+01  8.833e+01  -0.196 0.844621
Departure.Arrival.time.convenient1  3.138e-01  9.296e-02   3.376 0.000737 ***
Departure.Arrival.time.convenient2  4.231e-01  8.955e-02   4.724 2.31e-06 ***
Departure.Arrival.time.convenient3  2.432e-01  8.634e-02   2.816 0.004860 **
Departure.Arrival.time.convenient4 -6.830e-01  7.736e-02  -8.828  < 2e-16 ***
Departure.Arrival.time.convenient5 -9.215e-01  8.494e-02 -10.849  < 2e-16 ***
Ease.of.Online.booking1        3.071e+00  9.167e-01   3.350 0.000808 ***
Ease.of.Online.booking2        2.998e+00  9.167e-01   3.271 0.001071 **
Ease.of.Online.booking3        3.498e+00  9.164e-01   3.817 0.000135 ***
Ease.of.Online.booking4        4.358e+00  9.162e-01   4.756 1.97e-06 ***
Ease.of.Online.booking5        3.729e+00  9.166e-01   4.069 4.73e-05 ***
Gate.location1                -1.881e+01  6.523e+03  -0.003 0.997700
Gate.location2                -1.872e+01  6.523e+03  -0.003 0.997710
Gate.location3                -1.889e+01  6.523e+03  -0.003 0.997689
Gate.location4                -1.916e+01  6.523e+03  -0.003 0.997656
Gate.location5                -1.936e+01  6.523e+03  -0.003 0.997632
Food.and.drink1                1.425e-01  1.721e+00   0.083 0.933993
Food.and.drink2                4.262e-01  1.721e+00   0.248 0.804340
Food.and.drink3                3.014e-01  1.720e+00   0.175 0.860907
Food.and.drink4                3.279e-01  1.721e+00   0.191 0.848881
Food.and.drink5                2.162e-01  1.721e+00   0.126 0.900008
Online.boarding1              -3.668e+00  9.198e-01  -3.987 6.69e-05 ***
Online.boarding2              -3.582e+00  9.197e-01  -3.894 9.85e-05 ***
Online.boarding3              -3.806e+00  9.194e-01  -4.139 3.48e-05 ***
Online.boarding4              -2.155e+00  9.191e-01  -2.345 0.019022 *
Online.boarding5              -9.395e-01  9.193e-01  -1.022 0.306807
Seat.comfort1                  2.145e+01  6.523e+03   0.003 0.997376
Seat.comfort2                  2.092e+01  6.523e+03   0.003 0.997441
Seat.comfort3                  1.987e+01  6.523e+03   0.003 0.997569
Seat.comfort4                  2.057e+01  6.523e+03   0.003 0.997484
Seat.comfort5                  2.140e+01  6.523e+03   0.003 0.997382
Inflight.entertainment1        3.920e+01  1.521e+03   0.026 0.979440
Inflight.entertainment2        3.997e+01  1.521e+03   0.026 0.979036
Inflight.entertainment3        4.078e+01  1.521e+03   0.027 0.978612
Inflight.entertainment4        4.048e+01  1.521e+03   0.027 0.978766
Inflight.entertainment5        3.969e+01  1.521e+03   0.026 0.979180
On.board.service1             -2.285e+01  4.053e+03  -0.006 0.995501
On.board.service2             -2.272e+01  4.053e+03  -0.006 0.995527
On.board.service3             -2.217e+01  4.053e+03  -0.005 0.995636
On.board.service4             -2.209e+01  4.053e+03  -0.005 0.995651
On.board.service5             -2.156e+01  4.053e+03  -0.005 0.995755
Leg.room.service1             -2.414e+00  9.606e-01  -2.513 0.011973 *
Leg.room.service2             -2.138e+00  9.601e-01  -2.227 0.025969 *
Leg.room.service3             -2.275e+00  9.600e-01  -2.370 0.017782 *
Leg.room.service4             -1.590e+00  9.601e-01  -1.656 0.097678 .
Leg.room.service5             -1.411e+00  9.598e-01  -1.470 0.141624
Baggage.handling2             -2.300e-01  7.577e-02  -3.036 0.002395 **
Baggage.handling3             -8.621e-01  7.067e-02 -12.198  < 2e-16 ***
Baggage.handling4             -2.836e-01  6.869e-02  -4.129 3.64e-05 ***
Baggage.handling5              4.762e-01  7.311e-02   6.514 7.31e-11 ***
Checkin.service1              -1.416e+00  5.412e-02 -26.160  < 2e-16 ***
Checkin.service2              -1.233e+00  5.385e-02 -22.897  < 2e-16 ***
Checkin.service3              -7.160e-01  4.333e-02 -16.525  < 2e-16 ***
Checkin.service4              -7.502e-01  4.316e-02 -17.380  < 2e-16 ***
Checkin.service5                      NA         NA      NA       NA
Inflight.service1             -5.504e-01  7.584e-02  -7.258 3.94e-13 ***
Inflight.service2             -7.502e-01  6.887e-02 -10.892  < 2e-16 ***
Inflight.service3             -1.425e+00  5.719e-02 -24.909  < 2e-16 ***
Inflight.service4             -7.037e-01  4.497e-02 -15.647  < 2e-16 ***
Inflight.service5                     NA         NA      NA       NA
Cleanliness1                  -9.774e-01  7.459e-02 -13.104  < 2e-16 ***
Cleanliness2                  -9.387e-01  7.254e-02 -12.941  < 2e-16 ***
Cleanliness3                  -4.413e-01  6.090e-02  -7.246 4.30e-13 ***
Cleanliness4                  -5.922e-01  5.974e-02  -9.913  < 2e-16 ***
Cleanliness5                          NA         NA      NA       NA
```

Figure 38: Summary of prediction on customer satisfaction.

Observation:

The significant factors for prediction are those with **p-values less than 0.05**, indicating a **statistically significant relationship** with the **outcome variable**.

Based on the output, the **significant factors** for **predicting satisfaction include**: **Customer type (Loyal Customer), Type of travel (Personal Travel), Class (Eco, Eco Plus), Departure/arrival time convenience (categories 1-5), Ease of online booking (categories 1-5), Inflight Wi-Fi service (categories 1-5), Age and flight distance** have p-values slightly **above 0.05**, so they may or may **not be significant** depending on the chosen significance level. Note that some of the variables have multiple categories, with one category (the reference category) omitted from the output. The **coefficients** for the other categories indicate how they differ from the reference category in their relationship with the **outcome variable**. the factors with **p-values** greater than **0.05** are considered statistically insignificant. The following variables have a **p-value** greater than **0.05** and hence are **considered statistically insignificant**: **Age, Flight. Distance, Inflight WIFI service, Gate location, Food and drink**.

Note that these **variables** may still **have some impact** on the **response variable**, but their impact is **not statistically significant** based on the **given model and data**.

Summary.

In the final project of Aly 6015 I have worked on the flight customer satisfaction dataset, to predict the customer satisfaction on various predictor variable. Starting with the dataset I have done exploratory data analysis to know the nature of dataset.

Final Project Report.

Here I have done **EDA** and summary statistic of the dataset. After the initial Eda I move forward for the inspection of **outlier** in the variable and found few variables possess the outlier but those are not significant to the analysis using **boxplot**. Furthermore, I realized one possess 83 Na value and we removed Na value with the mean of it. **Correlation analysis** is also done and I realized arrival delay and departure delay are correlated with each other. For the visualization I have created several plots to know the **customer satisfaction proportion**, **distribution of gender**, **what proportion of customer is happy with the gate location, seating area** and many more. Furthermore, I have built the **linear regression model** to know the relationship between arrival delay departure delay, both of these variables hold **positive relation** with each other**. Anova** and **chisquare** testing was done to compare the 2 Generalized **linear model** with each other and we got to know the model with **predictor variable** arrival delay and departure significantly performed well with the performance based on the low **AIC value**. Furthermore, using **logistic regression** and **data validation techniques**, I have conducted an analysis of **the airline customer satisfaction dataset**, yielding an **87% accuracy and 86% precision**. The results reveal that the most **influential positive factors** contributing to customer satisfaction are customer **loyalty**, an **efficient online booking process, and convenient departure/arrival times**. In contrast, the most significant negative factors affecting satisfaction are traveling for personal reasons, being in **Economy or Economy Plus class**, and having lower ratings for gate location or departure/arrival time convenience. Notably, other factors such as **age** and **flight distance** show no clear association with customer satisfaction.

As a result, **airlines** should **prioritize improvements** to the **booking process** and **customer loyalty** programs, while also enhancing the overall gate and departure/arrival experience for passengers.

References and Bibliography.

1. Airline customer satisfaction dataset. (2021, September 29).https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.
2. Find duplicated rows (based on 2 columns) in Data Frame in R. (2011, August 8). Stack Overflow. https://stackoverflow.com/questions/6986657/find-duplicated-rows-based-on-2-columns-in-data-frame-in-r
3. Nguyen, C. (2021, September 29). Guide To Data Visualization With ggplot2 - Towards Data Science. Medium. https://towardsdatascience.com/guide-to-data-visualization-with-ggplot2-in-a-hour-634c7e3bc9dd
4. A Grammar of Data Manipulation. (2021). Dplyr. https://dplyr.tidyverse.org/
5. A Grammar of Data Manipulation. (2021). Dplyr. https://dplyr.tidyverse.org/.