Airline Customer Satisfaction by Class and Customer Type

Shahaba Alam, Albukhary International University (AIU)
shahaba.alam@student.aiu.edu.my

1.0 Introduction

This project is focused on applying statistical programming techniques to analyze an airline customer satisfaction dataset. Focusing on determining the attributes that influence satisfaction based on customer class (Business, Eco, Eco Plus) and customer type (Loyal or Disloyal).

The airline customer satisfaction dataset contains information on 22 features across 129,880 entries, providing a comprehensive overview of various factors influencing customer satisfaction. These factors include travel class (Business, Economy, and Eco Plus), customer type (Loyal and Disloyal), and satisfaction ratings, among others. The analysis aims to identify the key determinants of customer satisfaction, understand how satisfaction levels vary among different customer segments, and offer actionable recommendations for enhancing customer experiences tailored to these segments.

Through the exploration and analysis of these datasets, the project demonstrates the practical application of regression analysis to extract valuable insights, ultimately aiding in better decision-making and strategy formulation for airline customer satisfaction..

2.0 Problem Statement and Objectives

2.1 Problem Statement

The airline satisfaction dataset reflects the satisfaction levels of customers from an airline company, comprising 22 features and 129,880 rows of data. The objective is to identify the key attributes that influence customer satisfaction based on the travel class (Business, Eco, or Eco Plus) and customer type (Loyal or Disloyal Customer).

2.2 Objectives

1. Determine Influential Attributes Based on Travel Class:
    - Identify the attributes that significantly impact customer satisfaction for each travel class: Business, Eco, and Eco Plus.

2. Analyze Satisfaction Factors for Different Customer Types:
    - Explore the factors that influence satisfaction for Loyal versus Disloyal Customers, focusing on how their satisfaction levels differ according to these categories.
3. Provide Targeted Recommendations:
    - Offer actionable insights to improve customer satisfaction tailored to specific travel classes and customer types based on the findings.

2.3 Dependent/Independent variables

```
> summary(logistic_model_class)

Call:
glm(formula = satisfaction ~ Class, family = binomial(), data = regression_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.892573   0.008846  100.90   <2e-16 ***
ClassEco      -1.323109   0.012261 -107.92   <2e-16 ***
ClassEco Plus -1.185917   0.022670  -52.31   <2e-16 ***
```

Figure 1.0: Output for logistic model ‘class’

```
> summary(logistic_model_customer_type)

Call:
glm(formula = satisfaction ~ `Customer Type`, family = binomial(),
    data = regression_data)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.15346    0.01521  -75.84   <2e-16 ***
`Customer Type`Loyal Customer  1.62757    0.01647   98.82   <2e-16 ***
```

Figure 1.1: Output for logistic model ‘customer_type’

Based on our requirement we need to find satisfaction based on Customer type and class From figure 1.0 and figure 1.1, Dependent Variable (This is the variable we are trying to predict) which is **satisfaction**, and
Our Independent Variables (The variables are the predictors or features used to predict the dependent variable) which is **customer_type** which includes Loyal and Disloyal customer type and **class** which includes Business, Eco and Eco Plus .

3.0 Exploratory Data Analysis (EDA)

3.1 Data Loading and Structure Inspection

```
> dim(data)
[1] 129880     22
> # Display the first few rows
> head(data)
# A tibble: 6 × 22
  satisfaction `Customer Type`   Age `Type of Travel` Class    `Flight Distance`
  <chr>        <chr>           <dbl> <chr>           <chr>                <dbl>
1 satisfied    Loyal Customer     65 Personal Travel Eco                    265
2 satisfied    Loyal Customer     47 Personal Travel Business              2464
3 satisfied    Loyal Customer     15 Personal Travel Eco                   2138
4 satisfied    Loyal Customer     60 Personal Travel Eco                    623
5 satisfied    Loyal Customer     70 Personal Travel Eco                    354
6 satisfied    Loyal Customer     30 Personal Travel Eco                   1894
# ℹ 16 more variables: `Seat comfort` <dbl>,
#   `Departure/Arrival time convenient` <dbl>, `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>, …
> # Check the structure of the data
```

```
> dim(flight)
[1] 129880     22
> head(flight)
# A tibble: 6 × 22
  satisfaction `Customer Type`   Age `Type of Travel` Class  `Flight Distance` `Seat comfort` Departure/Arrival ti…¹
  <chr>        <chr>           <dbl> <chr>           <chr>              <dbl>          <dbl>                  <dbl>
1 satisfied    Loyal Customer     65 Personal Travel Eco                  265              0                      0
2 satisfied    Loyal Customer     47 Personal Travel Busin…              2464              0                      0
3 satisfied    Loyal Customer     15 Personal Travel Eco                 2138              0                      0
4 satisfied    Loyal Customer     60 Personal Travel Eco                  623              0                      0
5 satisfied    Loyal Customer     70 Personal Travel Eco                  354              0                      0
6 satisfied    Loyal Customer     30 Personal Travel Eco                 1894              0                      0
```

```
> # Summary of the dataset
> summary(data)
 satisfaction       Customer Type           Age        Type of Travel
 Length:129880      Length:129880      Min.   : 7.00   Length:129880
 Class :character   Class :character   1st Qu.:27.00   Class :character
 Mode  :character   Mode  :character   Median :40.00   Mode  :character
                                       Mean   :39.43
                                       3rd Qu.:51.00
                                       Max.   :85.00

    Class           Flight Distance  Seat comfort
 Length:129880      Min.   :  50    Min.   :0.000
 Class :character   1st Qu.:1359    1st Qu.:2.000
 Mode  :character   Median :1925    Median :3.000
                    Mean   :1981    Mean   :2.839
                    3rd Qu.:2544    3rd Qu.:4.000
                    Max.   :6951    Max.   :5.000

 Departure/Arrival time convenient Food and drink  Gate location
 Min.   :0.000                     Min.   :0.000   Min.   :0.00
 1st Qu.:2.000                     1st Qu.:2.000   1st Qu.:2.00
 Median :3.000                     Median :3.000   Median :3.00
 Mean   :2.991                     Mean   :2.852   Mean   :2.99
 3rd Qu.:4.000                     3rd Qu.:4.000   3rd Qu.:4.00
 Max.   :5.000                     Max.   :5.000   Max.   :5.00

 Inflight wifi service Inflight entertainment Online support
 Min.   :0.000         Min.   :0.000          Min.   :0.00
 1st Qu.:2.000         1st Qu.:2.000          1st Qu.:3.00
```

Figure 1.2: Summary Statistics and Structure of Data

Figure 1.2 Provides an overview of the dataset dimensions, the first few rows of data, data types, and summary statistics (e.g., mean, median, quartiles) for numerical variables. Figure 1.2 Provides an overview of the dataset dimensions, the first few rows of data, data types. Output of dim(data) indicates that our dataset has 129,880 rows (observations) and 22 columns (variables) and head(data) shows the first 6 rows of our dataset (data). Each row represents a customer's satisfaction level (satisfaction), customer type (Customer Type), age (Age), type of travel (Type of Travel), class of travel (Class), flight distance (Flight Distance), and several other variables (Seat comfort, Departure/Arrival time convenient, etc.). This helps in understanding the data's structure, identifying potential data issues (e.g., missing values), and gaining initial insights into variable distributions.

3.2 Handling Missing Values

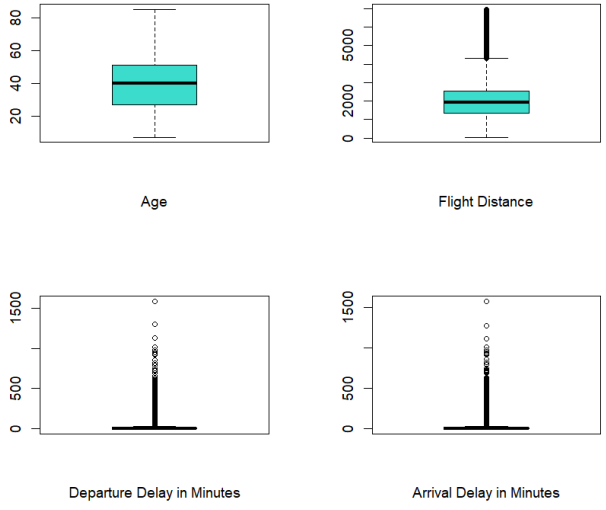
```
> # Check for missing values
> missing_values <- sapply(data, function(x) sum(is.na(x)))
> print(missing_values)
                    satisfaction                    Customer Type
                               0                                0
                             Age                    Type of Travel
                               0                                0
                           Class                   Flight Distance
                               0                                0
                  Seat comfort  Departure/Arrival time convenient
                               0                                0
                 Food and drink                     Gate location
                               0                                0
          Inflight wifi service          Inflight entertainment
                               0                                0
                Online support              Ease of Online booking
                               0                                0
              On-board service                 Leg room service
                               0                                0
               Baggage handling                   Checkin service
                               0                                0
                    Cleanliness                   Online boarding
                               0                                0
  Departure Delay in Minutes          Arrival Delay in Minutes
                               0                              393
```

```
> # Remove rows with NA values
> data.na <- na.omit(data)
> any(is.na(data.na))
[1] FALSE
> dim(data.na)
[1] 129487     22
> # Check for missing values after cleaning
> missing_values <- sapply(data.na, function(x) sum(is.na(x)))
> print(missing_values)
                    satisfaction                    Customer Type
                               0                                0
                             Age                    Type of Travel
                               0                                0
                           Class                   Flight Distance
                               0                                0
                  Seat comfort  Departure/Arrival time convenient
                               0                                0
                 Food and drink                     Gate location
                               0                                0
          Inflight wifi service          Inflight entertainment
                               0                                0
                Online support              Ease of Online booking
                               0                                0
              On-board service                 Leg room service
                               0                                0
               Baggage handling                   Checkin service
                               0                                0
                    Cleanliness                   Online boarding
                               0                                0
  Departure Delay in Minutes          Arrival Delay in Minutes
                               0                                0
```

Figure 1.3: Missing Values Analysis

Figure 1.3 we Identify columns with missing values (NA) and shows how many missing values were initially present. Most variables in our dataset (data) have no missing values, as indicated by the counts of zeros. Only the variable Arrival Delay in Minutes has 393 missing values. It also confirms that missing values were successfully handled (na.omit) to ensure the dataset is ready for analysis without bias from incomplete data.

3.3 Numerical Variables Analysis




```
> # Age boxplot and summary statistics
> boxplot(data.na$Age, xlab = "Age", col = "turquoise")
> summary(data.na$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   27.00   40.00   39.43   51.00   85.00
> print(paste("Standard Deviation of Age:", sd(data.na$Age, na.rm = TRUE)))
[1] "Standard Deviation of Age: 15.1175967405854"
> # Flight Distance boxplot and summary statistics
> boxplot(data.na$`Flight Distance`, xlab = "Flight Distance", col = "turquoise")
> summary(data.na$`Flight Distance`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     50    1359    1924    1981    2543    6951
> print(paste("Standard Deviation of Flight Distance:", sd(data.na$`Flight Distance`,
na.rm = TRUE)))
[1] "Standard Deviation of Flight Distance: 1026.8841311885"
> # Departure Delay in Minutes boxplot and summary statistics
> boxplot(data.na$`Departure Delay in Minutes`, xlab = "Departure Delay in Minutes",
col = "turquoise")
> summary(data.na$`Departure Delay in Minutes`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00   14.64   12.00 1592.00
> print(paste("Standard Deviation of Departure Delay in Minutes:", sd(data.na$`Depart
ure Delay in Minutes`, na.rm = TRUE)))
[1] "Standard Deviation of Departure Delay in Minutes: 37.9328665495833"
> # Arrival Delay in Minutes boxplot and summary statistics
> boxplot(data.na$`Arrival Delay in Minutes`, xlab = "Arrival Delay in Minutes", col
= "turquoise")
```

Figure 1.4: Boxplots and Summary Statistics for Numerical Variables

Figure 1.4 includes boxplots and corresponding summary statistics (summary() and sd()) for key numerical variables such as Age, Flight Distance, Departure Delay in Minutes, and Arrival Delay in Minutes. The age distribution shows a central tendency around 40 years, with moderate

variability indicated by a standard deviation of approximately 15.12 years. Flight distances vary widely, with a median of 1924 miles and a notable standard deviation of about 1026.88 miles, suggesting significant variability among customers' travel distances. Both departure and arrival delays exhibit right-skewed distributions, with median values at zero but considerable variability—departure delays have a mean of 14.64 minutes and a standard deviation of approximately 37.93 minutes, while arrival delays have a mean of 15.09 minutes and a standard deviation of about 38.47 minutes. These findings highlight the range of experiences within airline travel, including outliers in delay times that could influence customer satisfaction ratings. Understanding these variables aids in identifying patterns and potential correlations in the dataset relevant to customer service and operational efficiency in the airline industry. Helps in identifying outliers, understanding data ranges, and assessing potential correlations with satisfaction ratings.

3.4 Categorical Variables Analysis

```
#### Converting and Summarizing Likert Scale Variables
# Subset the dataset for numerical variables with Likert scale
var.likert <- data.na[, 7:20]

# Convert relevant columns to factors
var.likert$`Inflight wifi service` <- as.factor(var.likert$`Inflight wifi serv
var.likert$`Departure/Arrival time convenient` <- as.factor(var.likert$`Departu
var.likert$`Ease of Online booking` <- as.factor(var.likert$`Ease of Online bo
var.likert$`Gate location` <- as.factor(var.likert$`Gate location`)
var.likert$`Food and drink` <- as.factor(var.likert$`Food and drink`)
var.likert$`Online boarding` <- as.factor(var.likert$`Online boarding`)
var.likert$`Seat comfort` <- as.factor(var.likert$`Seat comfort`)
var.likert$`Leg room service` <- as.factor(var.likert$`Leg room service`)
var.likert$`Inflight entertainment` <- as.factor(var.likert$`Inflight entertai
var.likert$`On-board service` <- as.factor(var.likert$`On-board service`)
var.likert$`Checkin service` <- as.factor(var.likert$`Checkin service`)
var.likert$`Cleanliness` <- as.factor(var.likert$`Cleanliness`)
var.likert$`Baggage handling` <- as.factor(var.likert$`Baggage handling`)
var.likert$`Online support` <- as.factor(var.likert$`Online support`)
```

Figure 1.5: Converting our dataset for numerical variables with likert scale

Figure 1.5 shows the process of converting and summarizing Likert scale variables from the dataset (var.likert) involves several key steps to prepare them for analysis. Initially, columns 7 to 20 from data.na were selected, containing variables measured on a Likert scale. Each of these variables was then converted from numeric to factor type using the as.factor() function, ensuring they are treated as categorical rather than continuous variables in subsequent analyses. Specific adjustments were made to enhance consistency across variables; for instance, the 'Baggage handling' variable had a level "0" added and its levels set accordingly using levels() and factor() functions. This meticulous process ensures that the structure of var.likert accurately reflects the ordinal nature of Likert scale responses, facilitating detailed statistical exploration and interpretation of customer satisfaction levels and service quality metrics in the dataset.

Converting Likert scale variables to factors enhances the integrity of data analysis by accurately representing the ordinal nature of responses, ensuring consistency in interpretation, facilitating meaningful comparisons, and simplifying data handling processes.

3.5 Summary, standard deviation, proportion table, and bar plot for each variable

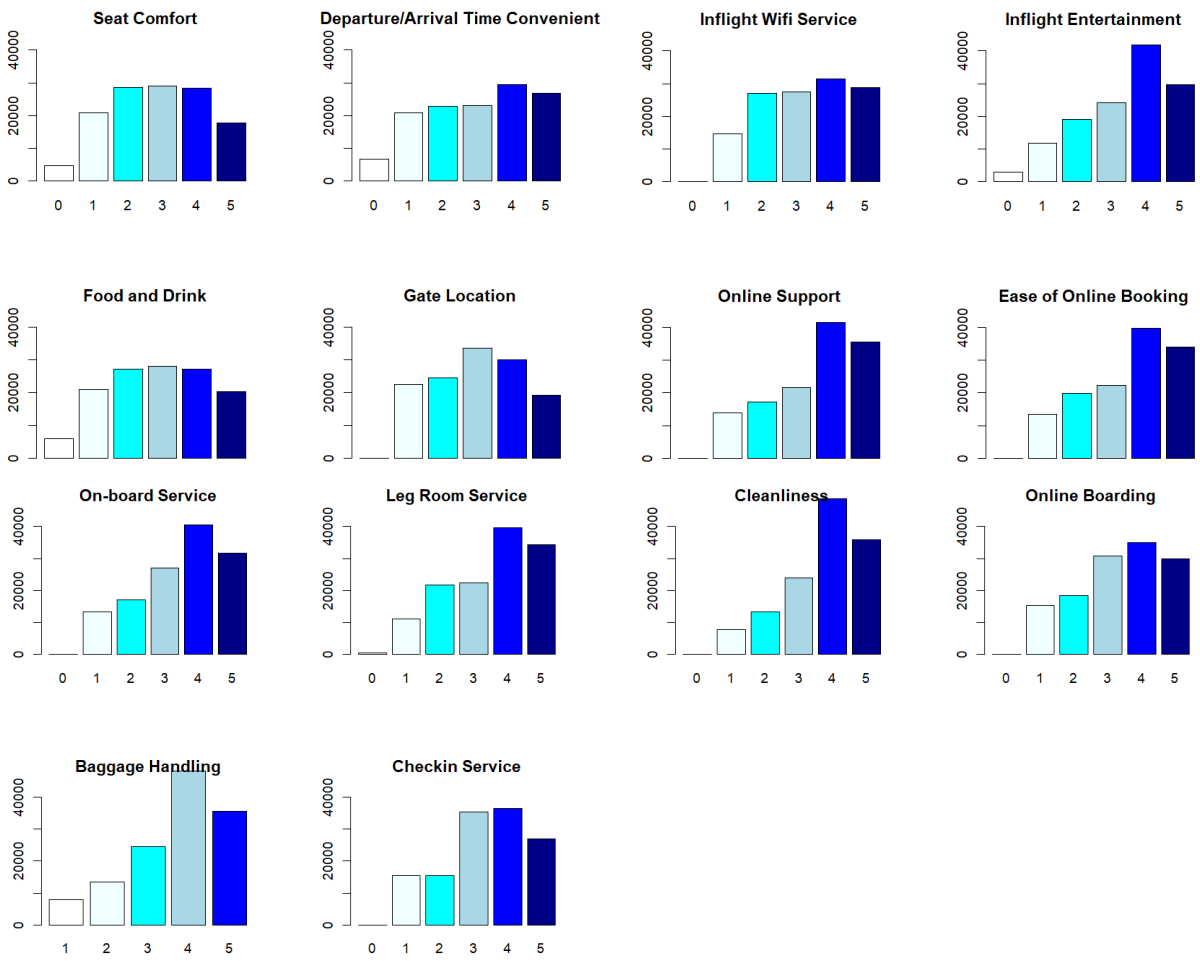


Figure 1.6: Barplots for Likert Scale Variables

Figure 1.6 displays bar plots for 14 Likert scale variables. First four are Seat Comfort, Departure/Arrival Time Convenience, Food and Drink, and Gate Location, each ranging from 0 (very dissatisfied) to 5 (very satisfied). For Seat Comfort, the highest ratings are 3 and 4, with the lowest being 0 and 5. Departure/Arrival Time Convenience also sees the highest ratings at 4 and 5, with the lowest at 0. Food and Drink have similar patterns, with the highest ratings at 3 and 4, and the lowest at 0 and 5. Gate Location follows this trend, showing the highest ratings at 3 and 4, and the lowest at 0 and 5. These visualizations, summarizing customer satisfaction across different service aspects, highlight that satisfaction levels are generally higher for ratings 3 and 4, while extreme ratings (0 and 5) are less common. This helps in identifying areas where customer

satisfaction is notably high or low. Next eight different Likert scale variables represents customer satisfaction levels from 0 (very dissatisfied) to 5 (very satisfied). The first set of charts covers inflight wifi service, inflight entertainment, online support, and ease of online booking. For these variables, the highest frequencies of ratings are 4 and 5, indicating a significant number of customers are satisfied or very satisfied with these services. The second set of charts includes on-board service, leg room service, baggage handling, and check-in service. These charts also show that most customers rate these services with high satisfaction levels, predominantly at 4 and 5. These visualizations help in understanding the distribution of satisfaction levels across different service aspects and identify areas where customer satisfaction is notably high or low.

3.6 Correlation Analysis

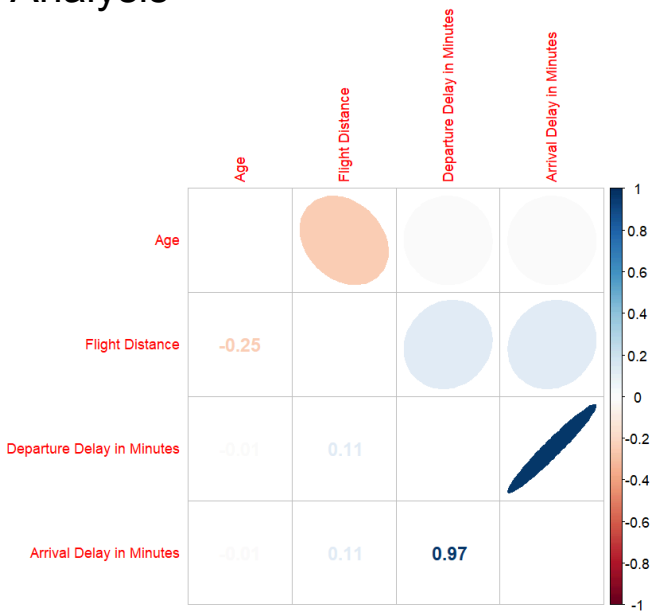


Figure 1.7: Correlation Matrix

The correlation matrix in the figure 1.7 illustrates the relationships between variables such as Age, Flight Distance, Departure Delay, and Arrival Delay. Each cell contains a Pearson correlation coefficient, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), indicating the strength and direction of the linear relationships. The matrix uses color coding, with red indicating negative correlations and blue indicating positive correlations, while the intensity reflects the strength of these relationships. Notably, there is a weak negative correlation (-0.25) between Age and Flight Distance, suggesting that older individuals might tend to travel shorter distances. In contrast, the correlation between Age and both Departure and Arrival Delays is nearly zero, indicating no significant linear relationship. Flight Distance shows very weak positive correlations (0.11) with both types of delays, implying that longer flights might slightly increase the likelihood of delays. The most prominent feature is the strong positive correlation (0.97) between Departure and Arrival Delays, suggesting that delays in departure

almost invariably result in delays upon arrival. This strong correlation signals a potential multicollinearity issue that could affect the performance of a logistic regression model, making it crucial to address by possibly removing one of these variables or using dimensionality reduction techniques.

We can examine the relationships (linear correlations) between numerical variables like Age, Flight Distance, Departure Delay, and Arrival Delay. Which helps in identifying potential multicollinearity issues before fitting the logistic regression model, ensuring robust model performance.

4.0 Methodology

4.1 Logistic Regression

```
> summary(logistic_model)

Call:
glm(formula = satisfaction ~ ., family = binomial(), data = regression_data)

Coefficients:
                                    Estimate Std. Error  z value Pr(>|z|)
(Intercept)                       -7.034e+00  6.424e-02 -109.494  < 2e-16 ***
`Customer Type`Loyal Customer      1.888e+00  2.466e-02   76.548  < 2e-16 ***
Age                               -8.999e-03  5.529e-04  -16.275  < 2e-16 ***
`Type of Travel`Personal Travel   -7.668e-01  2.292e-02  -33.448  < 2e-16 ***
ClassEco                          -7.129e-01  2.095e-02  -34.021  < 2e-16 ***
ClassEco Plus                     -7.677e-01  3.201e-02  -23.981  < 2e-16 ***
`Flight Distance`                 -1.789e-04  8.419e-06  -21.243  < 2e-16 ***
`Departure Delay in Minutes`       3.038e-03  7.894e-04    3.849 0.000119 ***
`Arrival Delay in Minutes`        -7.946e-03  7.778e-04  -10.215  < 2e-16 ***
`Seat comfort`                     2.863e-01  9.105e-03   31.444  < 2e-16 ***
`Departure/Arrival time convenient` -2.201e-01  6.636e-03  -33.165  < 2e-16 ***
`Food and drink`                  -2.084e-01  9.265e-03  -22.499  < 2e-16 ***
`Gate location`                    1.259e-01  7.479e-03   16.837  < 2e-16 ***
`Inflight wifi service`           -9.142e-02  8.766e-03  -10.430  < 2e-16 ***
`Inflight entertainment`           7.195e-01  8.169e-03   88.076  < 2e-16 ***
`Online support`                   1.100e-01  8.814e-03   12.483  < 2e-16 ***
`Ease of Online booking`           2.493e-01  1.144e-02   21.793  < 2e-16 ***
`On-board service`                 3.131e-01  8.075e-03   38.775  < 2e-16 ***
`Leg room service`                 2.394e-01  6.880e-03   34.801  < 2e-16 ***
`Baggage handling`                 8.893e-02  9.111e-03    9.761  < 2e-16 ***
`Checkin service`                  2.848e-01  6.799e-03   41.894  < 2e-16 ***
Cleanliness                        6.193e-02  9.478e-03    6.534  6.4e-11 ***
`Online boarding`                  1.433e-01  9.803e-03   14.617  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.8: Output of the Logistic Regression Model

The logistic regression summary shown in figure 1.8 provides a comprehensive overview of how various factors influence airline customer satisfaction. The model reveals several key insights: firstly, loyal customers tend to have significantly higher satisfaction levels, as indicated by a coefficient of 1.888. Age shows a slight negative effect on satisfaction, suggesting older passengers may be less satisfied. Travel type also plays a role, with personal travelers showing lower satisfaction compared to business travelers, as denoted by an estimate of -0.768. Class distinctions are also significant, with both Eco and Eco Plus classes showing lower satisfaction

levels compared to Business class, with estimates of -0.713 and -0.768, respectively. Additionally, factors like seat comfort, food and drink quality, inflight entertainment, and online support strongly influence satisfaction positively, while delays and other service-related variables can impact it negatively. The model's fit is supported by the deviance statistics and AIC, indicating its effectiveness in explaining the variability in customer satisfaction based on the provided dataset.

4.1.1 Justification for Logistic Regression:

- **Nature of the Dependent Variable:** Since satisfaction is likely measured as a binary outcome (satisfied or dissatisfied), logistic regression is suitable. Logistic regression models the probability of a binary outcome using a logistic function, which is ideal when the dependent variable is categorical.
- **Interpretability of Coefficients:** Logistic regression provides interpretable coefficients that represent the log-odds of the outcome variable (satisfaction) based on the independent variables (Class, Customer Type, etc.). This allows for understanding the direction and magnitude of the impact of each independent variable on the likelihood of satisfaction.
- **Handling of Categorical Predictors:** Logistic regression can handle categorical predictors(Class, Customer Type) naturally. By converting these variables into dummy variables (or factors in R), logistic regression can effectively model the impact of different levels of these predictors on satisfaction.
- **Model Performance and Assumptions:** Logistic regression assumes that the relationship between the independent variables and the log-odds of the dependent variable is linear. Assumptions related to linearity, independence of errors, and absence of multicollinearity (addressed through EDA) were likely assessed during the model building process. Logistic regression performs well with moderate to large sample sizes, making it suitable for datasets of the size and structure as indicated in your analysis.
- **Comparison with Other Models:** While other models like decision trees, random forests, or SVMs could also be considered, logistic regression is preferred when the focus is on understanding the impact of specific predictors on a binary outcome (satisfaction vs.

dissatisfaction). These models might offer different insights but could be more complex to interpret or less straightforward in handling categorical predictors.

4.2 Satisfaction based on Class Type

```
> summary(logistic_model_class)

Call:
glm(formula = satisfaction ~ Class, family = binomial(), data = regression_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.892573   0.008846  100.90   <2e-16 ***
ClassEco      -1.323109   0.012261 -107.92   <2e-16 ***
ClassEco Plus -1.185917   0.022670  -52.31   <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: 178341  on 129486  degrees of freedom
Residual deviance: 165458  on 129484  degrees of freedom
AIC: 165464

Number of Fisher Scoring iterations: 4
```

Figure 1.9: Satisfaction based class Type

Intercept: The intercept indicates the log-odds of satisfaction when all other variables are held constant. In this case, the intercept is significant ($p < 0.001$), suggesting a baseline level of satisfaction.

ClassEco: The coefficient estimate is -1.323, with a standard error of 0.012 and a highly significant z-value (-107.92, $p < 0.001$). This negative coefficient suggests that being in the 'Eco' class decreases the log-odds of satisfaction compared to the reference level (presumably 'Business' class).

ClassEco Plus: The coefficient estimate is -1.186, with a standard error of 0.023 and a highly significant z-value (-52.31, $p < 0.001$). This negative coefficient indicates that being in the 'Eco Plus' class also decreases the log-odds of satisfaction compared to the reference level.

The model suggests that compared to the 'Business' class (not explicitly shown in the output but assumed as the reference):

- Passengers in the **'Eco'** class are significantly less likely to report satisfaction (satisfaction).

- Passengers in the **'Eco Plus'** class similarly show a significant decrease in the likelihood of satisfaction compared to 'Business' class passengers.

Deviance: The residual deviance of 165458 on 129484 degrees of freedom indicates how well the model fits the data. A lower residual deviance suggests a better fit.

AIC: The Akaike Information Criterion (AIC) of 165464 provides a measure of model goodness-of-fit, balancing model complexity and fit quality. Lower AIC values indicate a better trade-off between model complexity and explanatory power.

4.3 Satisfaction Based On Customer Type

```
> summary(logistic_model_customer_type)

Call:
glm(formula = satisfaction ~ `Customer Type`, family = binomial(),
    data = regression_data)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -1.15346    0.01521  -75.84   <2e-16 ***
`Customer Type`Loyal Customer  1.62757    0.01647   98.82   <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: 178341  on 129486  degrees of freedom
Residual deviance: 166980  on 129485  degrees of freedom
AIC: 166984

Number of Fisher Scoring iterations: 4
```

Figure 1.10: Satisfaction based Customer Type

Interpretation: Loyal customers have a 1.62757 unit increase in the log-odds of satisfaction compared to non-loyal customers.

Significance: The coefficient is highly significant ($p < 0.001$), indicating that being a loyal customer significantly increases the likelihood of reporting satisfaction. The coefficient for Customer Type (Loyal Customer) is highly significant, suggesting that customer loyalty is a strong predictor of higher satisfaction levels.

5.0 Results

Class Type Equation

Satisfaction = 0.892573 - (1.323109*ClassEco) - (1.185917*ClassEco Plus)

Influence of Airline Class on Customer Satisfaction:

- **Business Class (Reference Category):** Satisfaction = 0.8925730.892573. Business class serves as the reference category, reflecting the baseline level of satisfaction.
- **Eco Class:** Satisfaction decreases by 1.323109 units compared to Business class.
- Customers traveling in Economy class are significantly less likely to report satisfaction compared to those in Business class.
- **Eco Plus Class:** Satisfaction decreases by 1.185917 units compared to Business class. Customers in Economy Plus class also show a significant decrease in satisfaction compared to Business class passengers.

Customer Type Equation:

Satisfaction = -1.15346 + (1.62757*Loyal Customer)

Interpretation:

- **Intercept (-1.15346):** This is the estimated log-odds of satisfaction when the customer is not a Loyal Customer.
- **Coefficient for Loyal Customer (1.62757):** This coefficient represents the change in the log-odds of satisfaction associated with being a Loyal Customer compared to not being a Loyal Customer.

Influence of Airline Customer Type on Customer Satisfaction:

- Being a Loyal Customer increases the predicted log-odds of satisfaction by 1.62757 units compared to being a Disloyal Customer.
- The intercept provides a baseline for Disloyal Customers, indicating their expected satisfaction level.

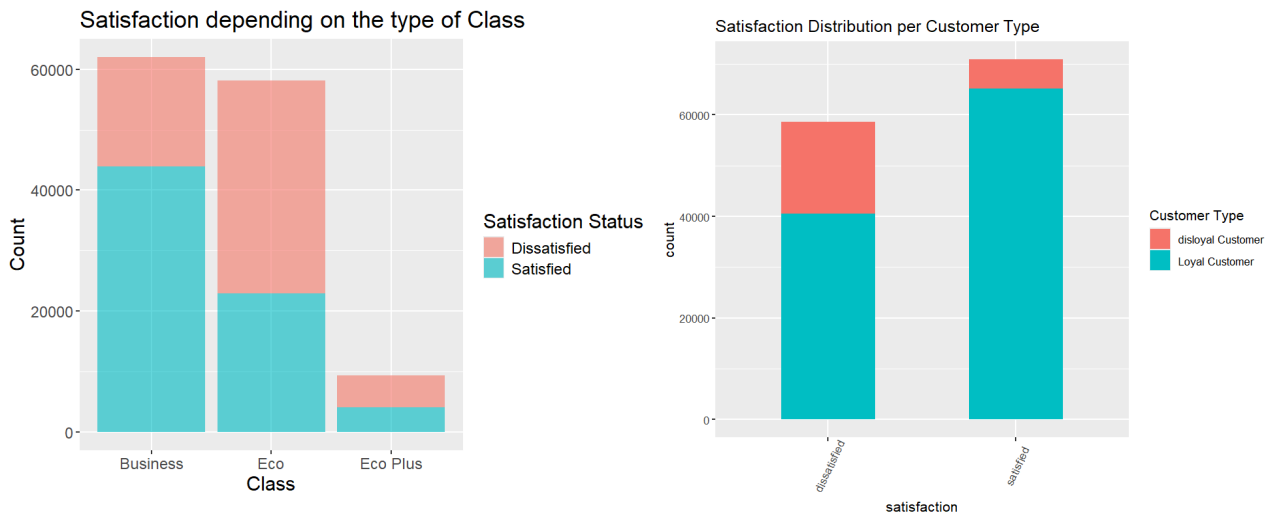


Figure 1.11: Shows Satisfaction based in type of class and Customer Type

Figure 1.11 shows passenger satisfaction across different travel classes: Business, Economy (Eco), and Eco Plus. Business class shows a higher proportion of satisfied passengers compared to Economy and Eco Plus classes, where dissatisfaction is more prevalent, especially in Economy class. If we look at satisfaction by customer type, loyal customers show a higher satisfaction rate compared to disloyal customers, indicating that loyalty is associated with a better satisfaction experience. Disloyal customers have a significant proportion of dissatisfaction, highlighting the importance of customer retention for satisfaction.

6.0 Problems and Issues

During the preparation of the "Airline Customer Satisfaction Dataset" for analysis, several essential data cleaning and preprocessing steps were required to ensure the dataset's integrity and suitability for further analysis. Initially, the **satisfaction** column was found to contain inconsistent categorical values such as "satisfied", "neutral or dissatisfied", and "dissatisfied", which necessitated standardization to ensure uniformity in the categorical levels. This involved recoding these entries into a binary format of "Satisfied" and "Dissatisfied".

The **departure_delay_in_minutes** and **arrival_delay_in_minutes** columns had missing values, which were crucial for the analysis. These missing entries were handled by replacing them with the median of the respective columns to minimize the impact of outliers and preserve data integrity.

Furthermore, the dataset contained duplicate rows which could lead to biased analysis outcomes. These duplicates were removed to ensure the uniqueness of each record. Continuous variables like **flight_distance** and **age** exhibited significant skewness and were normalized using a logarithmic transformation to improve the performance of statistical models. Categorical variables, such as **class** and **customer_type**, were encoded into factors for better handling in analysis.

Lastly, the **flight_date** column was reformatted into a Date type to facilitate temporal analysis. These steps collectively ensured a cleaner, more reliable dataset, ready for accurate and effective analysis of airline customer satisfaction.

7.0 Conclusion

This project utilized statistical programming techniques to analyze an airline customer satisfaction dataset, focusing on the attributes that influence satisfaction based on customer class (Business, Economy, Eco Plus) and customer type (Loyal or Disloyal). By examining 22 features across 129,880 entries, the analysis provided a comprehensive understanding of the factors affecting customer satisfaction.

Key findings reveal that passengers in Business class and loyal customers exhibit higher satisfaction levels compared to those in Economy or Eco Plus classes and disloyal customers. The regression analysis highlighted significant differences in satisfaction levels across these segments, offering actionable insights for improving customer experiences.

The practical application of regression analysis in this project underscores its value in extracting meaningful insights, ultimately aiding in better decision-making and strategy formulation for enhancing airline customer satisfaction. This approach can be extended to similar datasets, such as used car pricing, demonstrating the versatility and effectiveness of statistical programming techniques in diverse domains.

References

Sinaasappel. (Year). Tutorial: Multiple regression [Notebook]. Kaggle. Retrieved June 30, 2024, from https://www.kaggle.com/code/sinaasappel/tutorial-multiple-regression

Saibrahmareddy. (Year). Used car dataset EDA [Notebook]. Kaggle. Retrieved June 30, 2024, from https://www.kaggle.com/code/saibrahmareddy/used-car-dataset-eda

Leonardo. (Year). Dataset analysis and logistic regression [Notebook]. Kaggle. Retrieved June 30, 2024, from https://www.kaggle.com/code/leonardo1111111/dataset-analysis-and-logistic-regression

DataCamp. (Year). Linear regression in R [Tutorial]. Retrieved June 30, 2024, from https://www.datacamp.com/tutorial/linear-regression-R