Table of Contents

[Exploratory Data Analysis 2](#_Toc139652028)

[1. Data Overview and exploration 2](#_Toc139652029)

[2. Correlation Analysis 3](#_Toc139652030)

[3. Data Visualization 4](#_Toc139652031)

[4. Outliers 4](#_Toc139652032)

[Modelling 4](#_Toc139652033)

[1. Variable Selection using Forward Selection 4](#_Toc139652034)

# **Exploratory Data Analysis**

## Data Overview and exploration

We chose datasets from [kaggle](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction), which was already split into train and test sets. Our train dataset contains 103,904 rows and 24 features, of which we decided to exclude id column. The data description is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Initial Data Type** | **Feature** | **Initial Data Type** |
| Gender | Character | Seat comfort | Integer |
| Customer Type | Character | Inflight entertainment | Integer |
| Age | Integer | On-board service | Integer |
| Type of Travel | Character | Leg room service | Integer |
| Class | Character | Baggage handling | Integer |
| Flight Distance | Integer | Checkin service | Integer |
| Inflight wifi service | Integer | Inflight service | Integer |
| Departure/ Arrival time convenient | Integer | Cleanliness | Integer |
| Ease of Online booking | Integer | Departure Delay in Minutes | Integer |
| Gate location | Integer | Arrival Delay in Minutes | Integer |
| Food and drink | Integer | Satisfaction (response variable) | Integer |
| Online boarding | Integer |  |  |

***Data types:***

Of these, the variables from Inflight wifi service to Cleanliness take values of Likert scale (satisfaction level from 1 to 5 and 0 as “not applicable”). So, we considered those as categorical variables (ordinal).

***Missing values:***

In the ordinal categorical variables, there are few 0s which mean “not applicable” as passengers haven’t rated the respective variables. And the missing data of each variable is less than 5% of the total data and hence we imputed those values with mode.

Besides this, arrival delay also has 310 missing values. Our initial hypothesis was arrival delay and departure delay might be highly correlated as the formal depends on the latter because flights rarely make-up for the delayed time in air. And hence we removed the 310 rows to check the correlations first before making the decision of dropping the feature.

## Correlation Analysis

*Heatmap for numeric variables:*

A red and grey squares with white text

Description automatically generated

This shows that arrival and departure delays are highly correlated. So, instead of removing the missing values from arrival delay, we decided to drop the entire column from the data. In a similar manner, we generated a heat map for the categorical variables as well.

*Heatmap for categorical variables:*

A screen shot of a data

Description automatically generated

All the values are less than 0.7 which suggests that there are no strong correlations between any of the variables.

## Data Visualization

Among the numeric variables, flight distance has values in the range of 31-4983 and departure delay in the range of 0-1592. We visualized the data using QQ plots, histograms, box plots after applying various transformations like log, square root, box-cox. All the visualizations are listed in the appendix (Figure 1 and 2). Data of both variables doesn’t assume normal distribution even after transformations and box plot shows lots of data as outliers except for the log transformed data. Hence, based on our group’s judgement, we decided to apply log transformation. Removal of outliers is explained in the next section.

We also looked at the visualizations of categorical variables w.r.t the response variable (Figure 3). Few key observations from the plots:

* Gender doesn’t seem to play a major role as both have almost the same satisfaction levels.
* People in the age group of approximately 20-35 seem more dissatisfied than any other age group. And interestingly, people in this age group are more likely to be disloyal (Figure 4).
* Business class passengers have high satisfaction levels as expected. They rated high for check-in service, seat comfort and food compared to the Eco plus and Eco class (Figure 4).
* Surprisingly, cleanliness with even 4 and 5 ratings has passengers with satisfaction level as 0 (figure 3).
* Passengers have expressed more dissatisfaction with other services like inflight wifi, inflight entertainment, baggage handling, on-board service, inflight service. Only when the ratings of these services are above 3, there are more passengers who are satisfied but then this number doesn’t seem to be significantly large.

## Outliers

We examined the box plots after applying log transformations to flight distance and departure delay and there are outliers only in departure delay. We explored two methods to remove the outliers – one is by interquartile range (IQR) and the other by z-score (values<3). But z-score is an effective method if the data is normally distributed, which is not the case with our data. Hence, we chose the IQR method to remove outliers. After removing the outliers, the data has reduced to 103,886 rows from 103,904.

# **Modelling**

## Variable Selection using Forward Selection

To understand the important features to be included in the modelling, we used forward selection to choose the factors (Table 1 in appendix). The results are compelling:

* Our initial hypothesis was that flight distance was an important feature but as per forward selection, it isn’t.
* We didn’t anticipate that online boarding, type of travel, wifi service, customer type and inflight service are the top 5 important features.
* We expected departure delay to be among the top important features but as per forward selection, it is the 16th important feature.

Except flight distance, forward selection has chosen all the features. Using these results, we would like to build a logistic regression model and check p-values if there are still any insignificant variables.

**Appendix**

|  |  |  |
| --- | --- | --- |
| A picture containing text, diagram, plot, line  Description automatically generated | A graph of a graph  Description automatically generated | A graph with a line and a square  Description automatically generated |
| A picture containing text, diagram, line, plot  Description automatically generated | A graph of a building  Description automatically generated | A graph with a line and a square  Description automatically generated |
| A picture containing text, diagram, plot, line  Description automatically generated | A graph of a tall building  Description automatically generated | A graph with a line and a square  Description automatically generated |
| A graph of a line  Description automatically generated | A graph of a tall tower with Willis Tower in the background  Description automatically generated | A graph with a line and a square  Description automatically generated |

**Figure 1: Plots of flight distance**

|  |  |  |
| --- | --- | --- |
| A graph with a line  Description automatically generated | A graph with numbers and a bar  Description automatically generated | A graph of a line  Description automatically generated |
| A graph with a line  Description automatically generated | A graph with numbers and a bar  Description automatically generated | A graph with a line and a square  Description automatically generated |
| A graph with a line  Description automatically generated | A graph with a graph  Description automatically generated | A graph with a rectangle and a line  Description automatically generated |

**Figure 2: Plots of Departure delay in minutes**

**A screenshot of a computer screen

Description automatically generated**

**Figure 3: Categorical Variables (except Age) against response variable**

A group of graphs showing different types of data

Description automatically generated

**Figure 4: Categorical variables (contd.)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Step** | **Df** | **Deviance** | **Resid. Df** | **Resid. Dev** | **AIC** |
| 1 |  | NA | NA | 103885 | 142165.13 | 142167.13 |
| 2 | + Online.boarding | -4 | 43771.50806 | 103881 | 98393.62 | 98403.62 |
| 3 | + Type.of.Travel | -1 | 19766.82322 | 103880 | 78626.8 | 78638.8 |
| 4 | + Inflight.wifi.service | -4 | 10164.54812 | 103876 | 68462.25 | 68482.25 |
| 5 | + Customer.Type | -1 | 6546.868551 | 103875 | 61915.38 | 61937.38 |
| 6 | + Inflight.service | -4 | 7182.501528 | 103871 | 54732.88 | 54762.88 |
| 7 | + Leg.room.service | -4 | 2211.843474 | 103867 | 52521.04 | 52559.04 |
| 8 | + Checkin.service | -4 | 1750.071613 | 103863 | 50770.97 | 50816.97 |
| 9 | + Seat.comfort | -4 | 1186.632167 | 103859 | 49584.34 | 49638.34 |
| 10 | + Baggage.handling | -4 | 980.203787 | 103855 | 48604.13 | 48666.13 |
| 11 | + Inflight.entertainment | -4 | 864.653205 | 103851 | 47739.48 | 47809.48 |
| 12 | + Class | -2 | 874.104126 | 103849 | 46865.37 | 46939.37 |
| 13 | + Departure..Arrival.time.convenient | -4 | 652.479314 | 103845 | 46212.9 | 46294.9 |
| 14 | + Ease.of.Online.booking | -4 | 539.367682 | 103841 | 45673.53 | 45763.53 |
| 15 | + On.board.service | -4 | 409.296861 | 103837 | 45264.23 | 45362.23 |
| 16 | + Departure.Delay.in.Minutes | -1 | 341.223488 | 103836 | 44923.01 | 45023.01 |
| 17 | + Cleanliness | -4 | 219.708532 | 103832 | 44703.3 | 44811.3 |
| 18 | + Gate.location | -4 | 135.231073 | 103828 | 44568.07 | 44684.07 |
| 19 | + Age | -1 | 50.884056 | 103827 | 44517.18 | 44635.18 |
| 20 | + Food.and.drink | -4 | 29.758659 | 103823 | 44487.42 | 44613.42 |
| 21 | + Gender | -1 | 4.628407 | 103822 | 44482.8 | 44610.8 |

**Table 1: Forward Selection to choose features**