UEL-university of east London

Data Science Dissertation

Name: OKUNBOR PROMISE AISOSA

Student ID: 2218552

Supervisor: Mrs Romina Novaku

Submission Date: 1st September 2023

PREDICTING FACTORS INFLUENCING CUSTOMER SATISFACTION

IN THE AIRLINE INDUSTRY USING MACHINE LEARNING: A CASE STUDY

Table of Contents

Abstract

In the competitive airline industry, understanding factors that make passengers satisfied is as important as knowing what makes them dissatisfied. This study used machine learning techniques to analyse a dataset containing factors that might influence their decision to reuse a particular airline based on the quality of their service. Our findings show that inflight entertainment, seat comfort, online booking, and support influence customers satisfaction, while arrival delays influence their dissatisfaction. Random forests and K-nearest neighbours gave accurate results. The Random Forest was incredibly accurate, with a 96% success rate and K-Nearest Neighbours 89% success rate, respectively. These findings can help airlines improve and excel in a competitive industry by focusing on what matters most to their passengers.

1. Introduction

In today's highly competitive industry, customer satisfaction plays a fundamental role in determining the success or failure of a company (Vu, 2021). This holds especially true for the airline industry, where passengers' experiences and perceptions significantly impact their loyalty, word of mouth recommendation, and overall brand image. According to Zaharias (2016), many people travel by air for both work and fun. Some need to move between places quickly due to their jobs, while others want a break from their routine and are ready to spend on enjoyable travel experiences. They are also willing to set aside a separate budget for leisure travel and want to make the most of it (Zaharias, 2016). As a result, airlines are constantly seeking ways to better understand and predict factors that influence customer satisfaction. With the rapid advancements in technology, particularly in the field of machine learning, airlines have an opportunity to harness vast amounts of data to gain insights into customer preferences and behaviour.

Bougie, Pieters, and Zeelenberg (2003) did a study on anger and satisfaction in the context of customer responses to failed service encounters across different industries. It shows that anger and dissatisfaction have distinct effects on how customers react to service failures, with anger serving as a mediator between dissatisfaction and subsequent behaviour.

The airline industry is characterised by a complex web of services, encompassing ticket booking, check-in procedures, in-flight services, baggage handling, online booking, customer support, and many more. Each of these touchpoints contributes to the overall customer journey, and any negative experience at any stage can lead to dissatisfaction. Equally, a positive experience can foster loyalty and encourage word-of-mouth recommendations. For example, the Yuliani Dwi and Eko Agus (2020) study focuses on customer satisfaction within the commercial airline industry, emphasising its importance due to heightened competition and its impact on customer decisions and recommendations. The author's primary objective was to identify the key factors that influence customer satisfaction and make recommendations for the top 10 airlines listed on the Skytrax ranking. They employed a big data approach, collecting and analysing data from 10,189 customer reviews spanning the years 2012 to 2019. The researchers used binary logistic regression for their prediction. And they discovered that customer recommendations for airlines are significantly influenced by factors such as airline rating, seat comfort, cabin staff service, food and beverages, and value for money. However, in-flight entertainment and ground services did not show a significant impact on customer recommendations.

Given the complex nature of the industry, traditional methods of analysis often fall short of capturing the complex relationships between various factors and overall customer satisfaction. Machine learning, a subset of artificial intelligence, offers a data-driven approach to tackling this challenge. By leveraging algorithms that can learn from data, machine learning models can identify patterns, correlations, and predictive factors that might avoid manual analysis. This ability to handle large and complex datasets, coupled with the capacity to uncover hidden insights, has made machine learning an attractive tool for the airline industry to enhance customer satisfaction and gain insight on what makes them dissatisfied.

This literature review dive into the application of machine learning techniques to predict the factors that influence customer satisfaction and dissatisfaction in the airline industry. The focus is on how well the machine learning model performs in predicting the factors that influence customer satisfaction and dissatisfaction in the airline business. By examining existing research and outcomes, this review aims to understand customer needs that can lead to increased customer loyalty, enable a competitive edge, attract more customers, and foster overall airline growth.

2. Literature review

2.1 Airline quality service on customer satisfaction

Park et al., 2019; Tam, 2004; Gardial et al., 1994; Rust and Oliver, 1993, defined customer satisfaction "as an emotional response that results from a cognitive process of evaluating the service received against the costs of obtaining the service". They explained that most of the tourism research focused on customer satisfaction as a key determinant of industry success. Which means that customer satisfaction is a determinant factor that tells if a customer is happy with a service or product delivered in the airline industry. Service quality and passenger satisfaction are integral components of airline research, each contributing uniquely to the understanding of passengers' attitudes and expectations (Park et al., 2019).

In recent years, several scholars have presented their studies focusing on algorithms to predict how airline service quality can influence customer satisfaction. This case study closely examines the prediction of some of these influential factors, aligning with the broader trend of utilising machine learning for predictive analyses. For example, Suhartanto and Ariani Noor (2012) used regression analysis to determine "how service quality and price affect customer satisfaction on full-service and low-cost airlines". According to the findings of their analysis, airline service quality, particularly the attitudes and prices of service employees, are factors that should be given more attention to develop customer satisfaction in both types of airlines. Bellizzi, Eboli, and Mazzulla (2020) did an analysis of airline service quality under aspects such as on-time performance, baggage handling, food quality, seat comfort, check-in processes, and in-flight service. Their findings show that food quality, seat comfort, and service need to be improved to attract loyal customers.

Ridwan's (2022) recent study used machine learning algorithms like support vector, random forest, logistic regression, etc. to predict factors such as "inflight entertainment, online support, arrival/departure delay in minutes, food and drink, online support and bookings, seat comfort, cleanliness, and bag handling" has an impact on customer satisfaction in the airline industry. His result showed that Randon Forest gave an accuracy of 92%, with "inflight entertainment and seat comfort" having the most significant impact on customer satisfaction and passengers being dissatisfied with delay in their arrival.

Lestari and Murjito (2020) examined seven (7) determinant factors that can influence customer satisfaction and recommendation for ten (10) airlines on the "Skytrax list" using the logistic regression algorithm. According to their findings, "seat comfort, cabin staff service, food and beverages, and value for money" appear to have a significant influence on customer satisfaction and recommendation, with an R-square value of 76.53%, while ground services and in-flight entertainment were insignificant in their results. The aim of the authors findings is to help international airlines improve their service quality.

Furthermore, An and Noh (2009) investigated the "impact of in-flight service quality on customer satisfaction and loyalty in the airline industry" by analysing data collected from passengers that use business and economy classes. The result indicated that for passengers in the business class, six service quality factors were identified as important, such as "alcoholic and non-alcoholic beverage offerings, responsiveness and empathy from staff, reliability, assurance, presentation style of food, food quality, and entertainment". While for economy class, five important service quality factors were highlighted, such as "responsiveness and empathy from staff, food quality, availability of alcoholic beverages, availability of non-alcoholic beverages, and reliability". In essence, the authors research highlighted the importance for airlines to understand the different needs and expectations of different customer segments when designing and delivering in-flight services, which can ultimately contribute to higher levels of customer satisfaction and loyalty. (An and Noh, 2009).

Also, Noviantoro and Huang (2021) used data mining and machine learning algorithms to predict key service attributes that have the most impact on passenger satisfaction in the airline industry. The authors results showed that with "online boarding, inflight Wi-Fi, baggage handling, and inflight entertainment", airlines can enhance their competitive position and attract more passengers. The aim of their study is to highlight the importance of service quality as a determinant in airline competition, urging airlines to prioritise strategies that align with customer preferences (Noviantoro and Huang, 2021).

Wang’ondu (2009) also investigated factors affecting customer satisfaction in the airline industry, with a focus on Kenya Airways (KQ). According to his results, there is a big impact on customer satisfaction in the airline industry from things like reservations and ticketing, check-in procedures, in-flight services, and baggage handling and collection.

Enhancing customer loyalty is a highly successful strategy for any airline company hoping to succeed, as it increases market share, improves market positioning, increases profitability, and increases the number of devoted customers. This statement is supported by the investigation conducted by Wang and Chaipoopirutana (2015) as regards factors shaping customer loyalty within the airline industry, particularly in "Thai Airways". They underlined the essential role of customer satisfaction and its influence on fostering customer loyalty by demonstrating the relationships among service quality, "complaint handling, corporate image, customer satisfaction, and customer loyalty". The authors research offers valuable guidance for airlines seeking to enhance customer loyalty and overall business success.

Suki (2014) Used a convenience sampling method of 300 respondents who frequently used Malaysia Airlines, they examined the effects of airline service quality such as in-flight entertainment, flight punctuality, and cleanliness on customer recommendations and satisfaction. The author used "structural equation modelling" (SEM) to analyse the data, and the result showed that when customers are satisfied with the airline service, word-of-mouth is more likely to be recommended to others.

Fazidah and Mohamed (2015) used the SERVQUAL model to identify five service quality factors that can impact customer satisfaction in the airline industry. Factors like tangibility, information quality, responsiveness, trust, and personalization. According to the authors findings, the main service quality factors that had a significant impact on customer satisfaction were tangibleness, trust, and personalization. The authors went on to state that their findings implied that maintaining users' personal information security is important and that trust is acknowledged as a crucial component in online transactions in the airline industry. Additionally, Han et al.'s (2012) study involved an empirical examination of how passengers view airline lounges. His findings revealed that the most significant factor shaping customer satisfaction and the likelihood of passengers revisiting the lounge is the quality of food and beverage service.

Customer satisfaction is important because it creates customer loyalty, leads to customer recommendations or referrals, allows for a competitive edge, and contributes to growth and profitability in the airline industry, as explained by (Joseph and Ismail 2012; Baker 2013).

2.2 Customer dissatisfaction

Dissatisfaction occurs when customers are unhappy with a product or service that has been delivered. This statement is supported by Bougie, Pieters, and Zeelenberg (2003), who said that "customer dissatisfaction can result from product defects and poor service". Just as customer satisfaction can lead to happy customers, dissatisfied customers can also wreak havoc on any aspect of any business (Fuchs, 2022).

Dissatisfied customers in the airline industry can have several negative effects on the company's bottom line, including decreased revenue, decreased customer loyalty, a damaged reputation, and even the airline industry's automatic shutdown. Uslu and Sarper Karakadilar (2022) investigated the effects of revenue management complaints on customer perceptions and loyalty. Complaints about revenue management practices, according to the author's findings, have a positive effect on the travel experience. Matusitz and Breen (2009) also investigated customer service complaints. According to the research, poor service delivery can result in the loss of both loyal and potential customers.

Moreover, Park, Lee, and Nicolau (2020) examined the relationship between “airline service attribute quality” and “overall satisfaction", considering potential mixed effects. Their study analyses 157,035 consumer reviews to explore the impact of various service attributes on positive and negative ratings. The authors findings showed that positive ratings (satisfaction) and certain service attributes like "cleanliness, food and beverages, and in-flight entertainment" significantly influence positive ratings, contributing to customer satisfaction. While attributes like customer service, check-in, and boarding significantly impact negative ratings (dissatisfaction).

Another author who wrote about customer satisfaction and dissatisfaction in the airline industry is Anand and Bansal (2016). They created a predictive model for identifying customer satisfaction and dissatisfaction with airline service, using mobile phones as the product of interest and 11 decision-making variables as independent factors. The author employed logistic regression, a powerful multivariate technique, to validate the predictive model for the opposite dependent variable. Their findings showed that logistic regression is effective in predicting customer dissatisfaction with service, such as departure delays and arrival delays.

Insights into the aspects of airport services that are most likely to give rise to complaints and dissatisfaction can help to better understand how poor service experiences can affect the behaviour of passengers. For instance, Chang et al. (2008) examined how dissatisfaction with service quality might lead to future complaints, particularly at Taoyuan International Airport in Taiwan. His study involved surveying passengers who had experienced unpleasant or failed service encounters at the airport over the past six months. The survey gathered 323 valid responses from passengers, primarily females aged 17 to 26, with university degrees, often travelling as part of a group. He observed that most of the respondents had used the airport frequently, and many had experienced negative incidents primarily in Terminal 1. He also talked about how airport direction signs and layout, airport service facilities, and airline company schedules are related to the issues passengers are dissatisfied with. The author stated that "poor airport services experienced in failed encounters directly influence passengers' intentions to make complaints in the future".

2.3 Aftermath effect of covid19 as relate to customer satisfaction in Airline.

The era of the COVID-19 pandemic's effects on the airline industry were specifically examined in Zahraee et al.'s (2022) study, which investigated customer satisfaction and airline response times. He pointed out that not much research has systematically examined how airlines handled the pandemic and its impact on client satisfaction. His research sought to close this gap by evaluating consumer satisfaction with the aviation sector during COVID-19 through a questionnaire survey in China and investigating policies that could help airlines by examining feedback from 49 major airlines around the world, taking operational costs and passenger safety into account. The authors used a questionnaire survey in China to gauge customer satisfaction using 22 constructs in four phases: "pre-flight, in-flight, after-arrival," and additional measures (like face mask regulations and HEPA filters). The authors results showed that the measures related to "hygiene products, thermal scanners, and cabin cleansing" ranked highest in terms of passenger satisfaction and that passengers' satisfaction varied across different COVID-19 measures, highlighting areas for improvement in in-flight stage measures. The analysis provided by the author points to potential directions for boosting passenger satisfaction, enhancing in-flight stage measures, and ultimately accelerating the industry's post-COVID-19 recovery.

In addition, Istijanto (2021) also investigated the impact of customer participation and service failure on customer recovery satisfaction within the airline industry, particularly in the context of the COVID-19 pandemic, The author study used a scenario-based experiment involving 180 respondents to analyse his findings. His research employs exploratory factor analysis to validate measurements and a general linear model to examine how customer participation and service failure influence customer recovery satisfaction. The authors findings indicated that higher customer satisfaction is observed when customers actively participate in service recovery during failures caused by COVID-19. However, increased customer participation during service failures due to reasons like pilots being on strike leads to reduced customer recovery satisfaction.

2.4 Using of machine learning algorithms in airline

Machine learning models can aid airlines in the analysis of massive amounts of data, the discovery of untapped knowledge, and the highly accurate forecasting of customer satisfaction. By implementing these insights, airlines can enhance their offerings to customers and how they interact with them, fostering greater customer loyalty and boosting retention rates as well as their overall success. For example, Sugara and Purwitasari's (2022) developed a machine learning model to predict flight delays using various algorithms. The author’s outcome showed how flight delays negatively impact customer satisfaction and airport revenue. They also incorporated weather data and employed techniques to handle imbalanced data. After evaluating their model's performance, they found that a combination of the Random Forest classifier with ROS and RUS techniques yielded the best results. Their model achieved an accuracy of 82.58%, an error rate of 17.42%, and an AUC value of 81.1%. The authors believe that their observation can help airports make strategic decisions to optimise revenue and enhance operations by better managing flight delays. Yazdi et al.'s (2020) article supports this claim, as the author used deep learning (DL) to predict how flight departure delays can lead to passenger dissatisfaction. Yazdi et al.’s (2020) and Sugara and Purwitasari’s (2022) research is valuable for airline insight because it enables airlines to provide more precise delay predictions and to keep passengers updated in real time. By lowering uncertainty and empowering users to make more knowledgeable decisions like changing connecting flights or travel schedules, this transparency can improve customer satisfaction.

Another related article that talked about the use of machine learning techniques in the aspect of differentiating between travellers, such as business and causal, is Samunderu and Farrugia (2022). The authors concentrated on locating “critical traits, heuristics, and models” that can precisely forecast a passenger's intended travel destination at the time of purchase. They emphasised the usefulness of this prediction in real-time dynamic product customization for customers. The authors also discussed the need for machine learning algorithms that are easy for business stakeholders to understand without requiring deep technical knowledge. Their study showed that combining techniques can produce insights into consumer behaviour with useful applications that generate revenue. In general, the article discussed the significance of comprehending consumer behaviour and preferences to improve the services provided by the airline industry. Another author who wrote about the use of machine learning to analyse customer satisfaction within Korean Airlines is Franklin (2023).

With various studies conducted by different authors, there is no doubt that service quality plays a pivotal role in shaping customer perceptions. Elements such as cabin cleanliness, food quality, staff responsiveness, and service reliability contribute to passengers' satisfaction. A comfortable and enjoyable in-flight experience, characterised by factors like seat comfort, entertainment options, and overall ambiance, significantly influences customers' overall contentment. Also, the importance of machine learning techniques in airline organisations for understanding customer satisfaction and dissatisfaction cannot be overemphasised, as it opens the door for airline development and can also cause airline failure. The relevancy of an airline can only be achieved when it understands its customers’ needs and acknowledges their feedback. Hence the need to use machine learning techniques to predict what services will satisfy their needs.

3 Methodology

This methodology section presents the systematic approach employed to predict customer satisfaction and dissatisfaction within the airline industry. It encompasses data description, exploration, analysis, and predictive modelling, ensuring the establishment of a comprehensive framework yielding reliable insights. A comprehensive framework is presented, detailing the steps to be undertaken to attain accurate predictions and actionable insights.

3.1 Data Collection and Description:

The first step entails describing the dataset obtained from the reputable online resource Kaggle. This dataset effectively documents a series of customer comments regarding various aspects of their travel experience. The dataset is thoroughly explained, including information about the number of observations, the names of the variables, and the key characteristics of the data being represented. The transparency makes sure that subsequent analysis and modelling are built on a strong understanding of the data's composition and characteristics.

3.2 Data Exploration and Analysis:

The following phase involves conducting an examination of the data. The journey towards predicting levels of customer satisfaction and dissatisfaction starts with the process of data exploration and analysis, often referred to as preprocessing. The unprocessed dataset will undergo a comprehensive preprocessing procedure to effectively address potential instances of data quality concerns. The methods employed to manage missing data, transform, normalise, handle imbalanced data where necessary, and reduce the number of columns to ensure the dataset maintains its integrity and relevance will be discussed.

Employing exploratory data analysis (EDA) serves as the initial step to comprehend the latent patterns and interconnections within the dataset. This stage, facilitated through visual techniques like histograms, scatterplots, correlation matrices, and statistical testing, uncovers insights that significantly influence the process of selecting relevant features and constructing predictive models. To quantitatively determine the associations between variables and instances of customer satisfaction and dissatisfaction. This approach enriches our comprehension of pivotal predictors and enhances our ability to comprehend the dynamics at play.

3.3 Predictive Modelling:

After proper cleaning and visualisation, we move forward with machine learning modelling. The foundation of this method is predictive modelling, which uses machine learning algorithms to predict customer satisfaction outcomes and dissatisfied customers. Decision trees, random forests, logistic regression, K-nearest neighbours (KNN), and naive bayes are some of the algorithms that will be used. Furthermore, the test result or outcome for each algorithm will be discussed. The model will be evaluated and compared for accuracy, sensitivity, specificity, precision, and F1 score. The final predictive model is chosen based on how well it can produce precise and trustworthy predictions of customer satisfaction, increasing its practical applicability.

In summation, this methodology summarises an organised and transparent approach to predicting customer satisfaction and dissatisfaction in the airline industry. The workings of data collection, preparation, exploration, and predictive modelling conclude in a strong framework that generates valuable insights. By adopting this methodology, this research aims to empower the airline industry with actionable information to enhance customer experiences. The transparent and thorough approach ensures the credibility of the findings, positioning the study as a reliable resource for both academia and industry practitioners.

Through the identification of critical predictors, this research illuminates' avenues for airlines to tailor services and strategies that foster customer satisfaction and loyalty. As a result, this study contributes to the broader field of predictive analytics, guiding industries beyond aviation in their pursuit of elevating customer contentment within competitive markets. Overall, this study provides valuable insights into customer behaviour and preferences, enabling the airline industry to thrive in a competitive market by effectively catering to customer needs.

Process flow

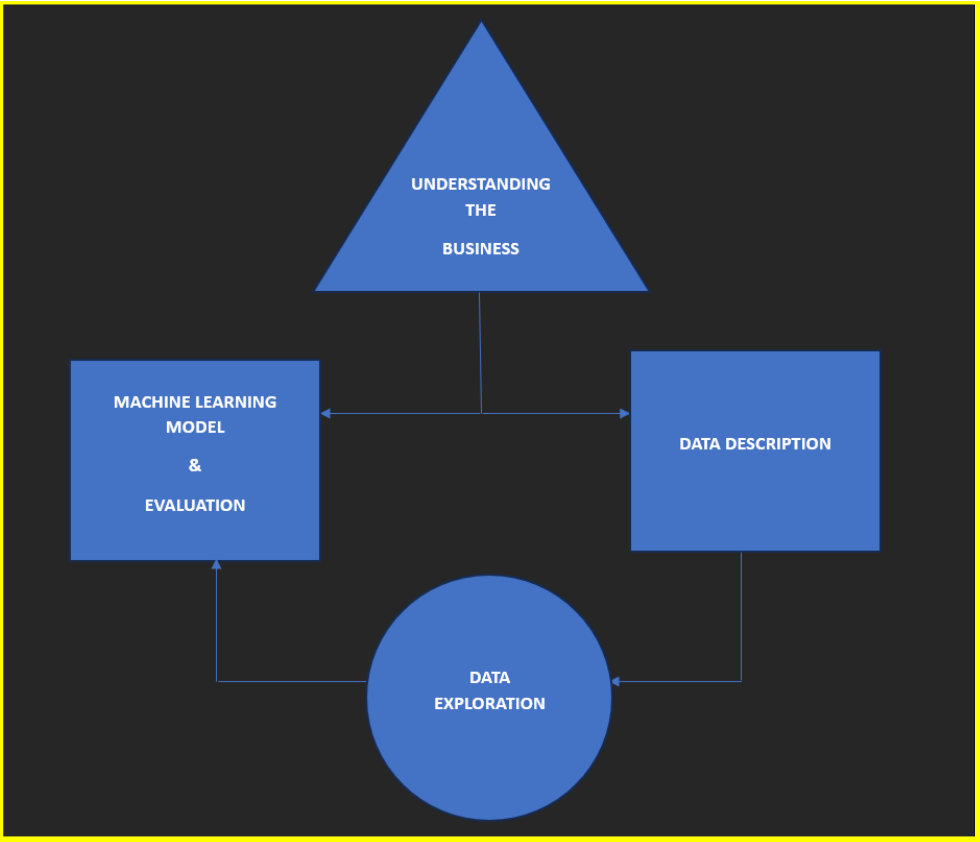


Figure 1

4 Data Description

4.1 Overview

Data description is the initial process of inspecting and providing an overview of the dataset that is used. It involves understanding the basic characteristics of the data, such as its structure, variable types, contents, and general statistics. Data description is often the first essential step in any data analysis and exploration process, as it helps us gain a preliminary understanding of the dataset and its properties before diving into more in-depth analysis or modelling.

According to Tukey (1970), data description involves understanding complex datasets, which is the crucial initial step in the data analysis process before formal statistical modelling is carried out.

Hence, it is essential to give a detailed description of the dataset that will be used in this research. Understanding the factors that make customers satisfied plays a critical role in shaping an airline's reputation and fostering customer loyalty. As airlines strive to provide exceptional experiences to their passengers, data-driven approaches and predictive analytics have become invaluable tools for understanding customer feedback and improving overall service quality.

The dataset, as earlier mentioned, was obtained from the Kaggle website, Airlines Customer Satisfaction | Kaggle, a well-known online platform for data science competitions and machine learning practitioners. The dataset includes a compilation of customer feedback and flight details from customers who have flown with the airlines. The dataset consists of 23 columns and 129,800 rows, where each row represents the details and experiences of the customers. Below is a brief explanation of what each variable represents:

4.2 Variable description

The independent variables in the dataset represent various factors that could potentially influence customer satisfaction. These factors such as “seat comfort, departure convenience, food and drink, gate location, inflight entertainment," etc. These factors are assessed using a scoring system ranging from 0 to 5.

The dependent variable in the dataset is "customer satisfaction," which serves as the target variable for prediction. Customer satisfaction is measured as a binary of 0 and 1, indicating whether the customer is satisfied or dissatisfied with the travel experience.

Additionally, the dataset includes passenger details like gender, age, customer type, type of travel, class of travel, and flight distance. This helps account for differences in customer satisfaction based on different customer segments.

Table 1 shows a more detailed description of the variables and what they represent.

Table 1: Detailed description of all variables

S/N	Numerical Variable	Description
1	Age	Passenger age- (Min=7, Median=40, Mean=39 and Max 85)
2	Flight distance	Travel distance in kilometres- (Min=50, Median=1924, Mean=1981 and Max= 6951)
3	Departure delay in minute	Minute flight departures (Min=0, Median=0, Mean=14 and Max= 1592)
4	Arrival delay in minute	Minute flight arrives (Min=0, Median=0, Mean=15 and Max= 1584)
	Categorical variables	**Description**
5	Gender	Female (1) and Male (0)
6	Customer type	Loyal (1) and disloyal (0)
7	Type of travel	Business (1) and Personal (0)
8	Class of travel	Business (0), Eco (1) and Eco plus (2)
9	Satisfaction	Satisfied (1) and dissatisfied (0)
10	Seat comfort	Rating 0 to 5
11	Departure Convenient	
12	Food and Drink	
13	Gate Location	
14	Inflight WiFi Service	
15	Inflight Entertainment	
16	Online Support	
17	Ease of Online Booking	
18	On-board Service	
19	Leg Room Service	
20	Baggage Handling	
21	Check-in Service	
22	Cleanliness	
23	Online Boarding	

The statistical summary result of the dataset can be found in figure 2.

5 Data Exploration

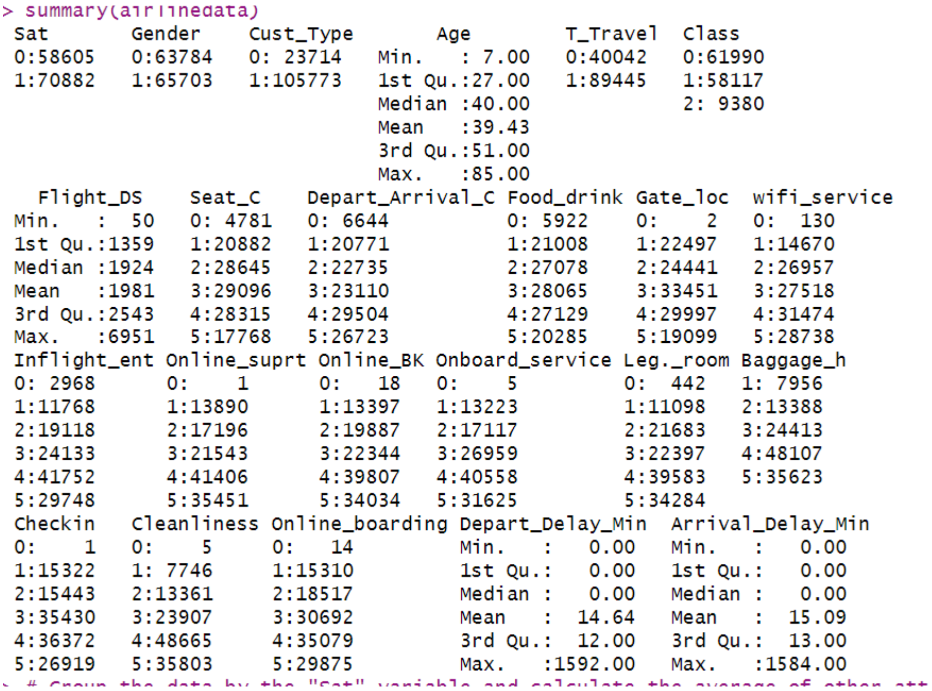
5.1 Data Preparation

Data exploration, also known as exploratory data analysis (EDA), is the second critical step in any data analysis process. It involves data cleaning, transformation, handling missing values, checking for imbalanced data, and identifying outliers in a dataset. Hadley Wickham's article talked about the concept of tidy data, which emphasises structuring data in a consistent and organised manner to facilitate data analysis and visualisation (Hadley, 2021). Tidy data principles play a fundamental role in effective data exploration and analysis (Hadley, 2021).

For this reason, the main goal of carrying out data exploration in this research is to gain insights and a deeper understanding of the variables before performing more advanced modelling. Before commencing the exploration of the variables in RStudio, Excel was used to remove 393 missing variables, and we were left with 12,9487 values. Column names were reduced. After that, the categorical variables were converted to factor-based variables to create plots and perform correlation matrices.

5.2 Data visualization

Next, charts, graphs, maps, and tables are presented below to visualise the distribution, normality, and correlation of dependent and independent variables.



```
> summary(airlinedata)
 Sat       Gender    Cust_Type       Age        T_Travel  Class
 0:58605   0:63784   0: 23714   Min.   : 7.00   0:40042   0:61990
 1:70882   1:65703   1:105773   1st Qu.:27.00   1:89445   1:58117
                                Median :40.00             2: 9380
                                Mean   :39.43
                                3rd Qu.:51.00
                                Max.   :85.00
   Flight_DS    Seat_C    Depart_Arrival_C Food_drink Gate_loc   wifi_service
 Min.   :  50   0: 4781   0: 6644          0: 5922    0:     2   0:  130
 1st Qu.:1359   1:20882   1:20771          1:21008    1:22497   1:14670
 Median :1924   2:28645   2:22735          2:27078    2:24441   2:26957
 Mean   :1981   3:29096   3:23110          3:28065    3:33451   3:27518
 3rd Qu.:2543   4:28315   4:29504          4:27129    4:29997   4:31474
 Max.   :6951   5:17768   5:26723          5:20285    5:19099   5:28738
 Inflight_ent Online_suprt Online_BK Onboard_service Leg._room Baggage_h
 0: 2968      0:     1     0:    18  0:     5        0:  442   1: 7956
 1:11768      1:13890      1:13397   1:13223         1:11098   2:13388
 2:19118      2:17196      2:19887   2:17117         2:21683   3:24413
 3:24133      3:21543      3:22344   3:26959         3:22397   4:48107
 4:41752      4:41406      4:39807   4:40558         4:39583   5:35623
 5:29748      5:35451      5:34034   5:31625         5:34284
 Checkin   Cleanliness Online_boarding Depart_Delay_Min  Arrival_Delay_Min
 0:     1  0:     5    0:    14        Min.   :   0.00   Min.   :   0.00
 1:15322   1: 7746     1:15310         1st Qu.:   0.00   1st Qu.:   0.00
 2:15443   2:13361     2:18517         Median :   0.00   Median :   0.00
 3:35430   3:23907     3:30692         Mean   :  14.64   Mean   :  15.09
 4:36372   4:48665     4:35079         3rd Qu.:  12.00   3rd Qu.:  13.00
 5:26919   5:35803     5:29875         Max.   :1592.00   Max.   :1584.00
> # Group the data by the "Sat" variable and calculate the average of other att
```

Figure 2: Summary statistic

Display the summary statistic of the dataset, such as mean, media and max.

- **Visualizing the distribution of the dependent variable (satisfaction)**

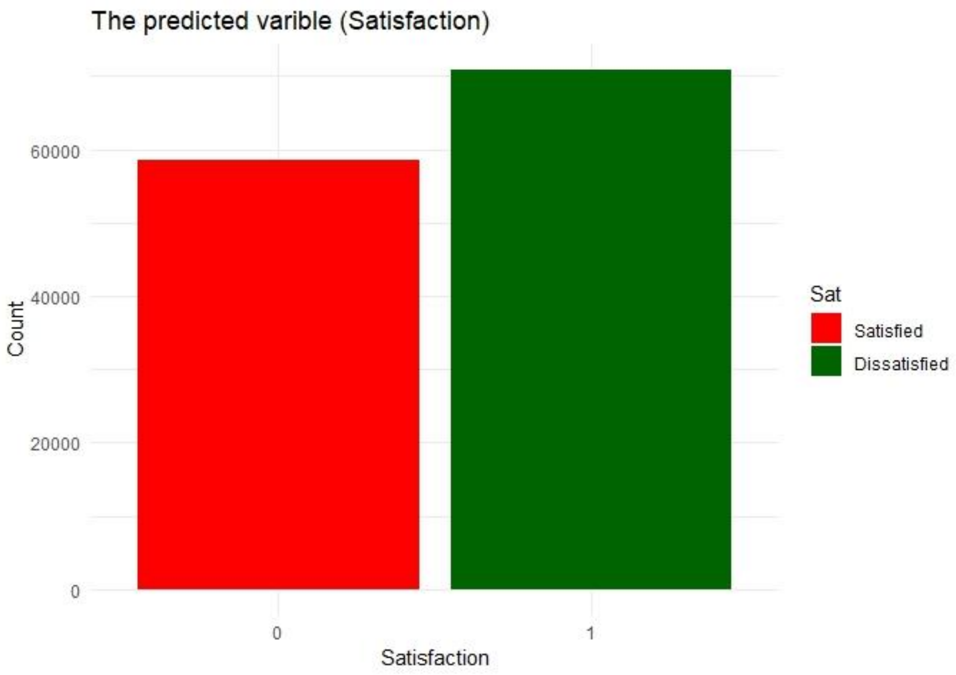


Figure 3: Distribution of the dependent variable before balancing.

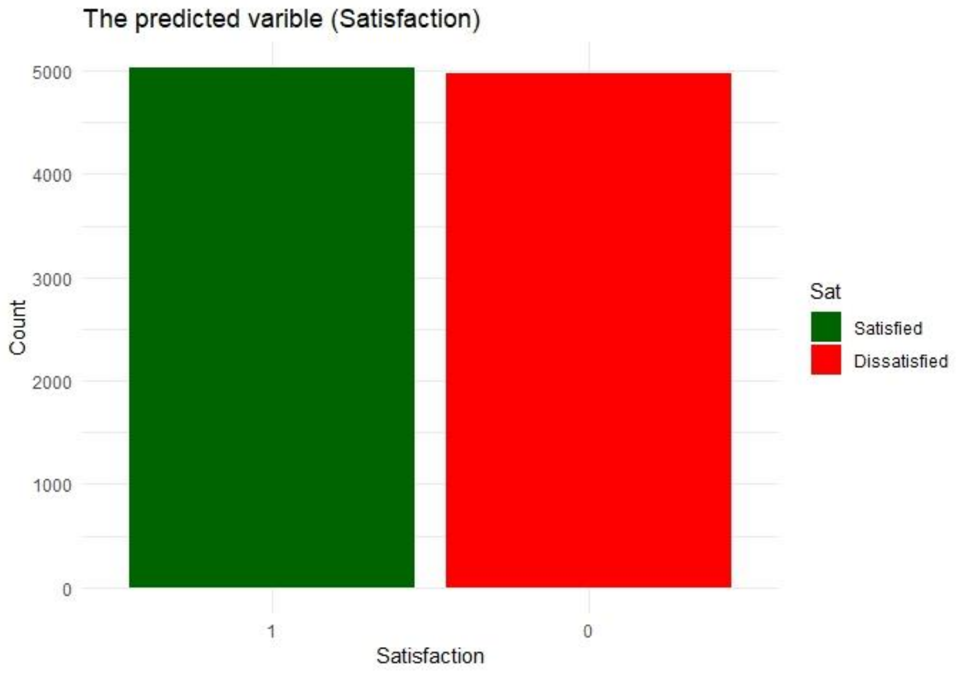


Figure 4: Distribution of the dependent variable after balancing

The initial code executed revealed an imbalance within the satisfaction variable, which is our predictive target, as seen in figure 3. This can cause biases during predictions when using our model. To address this potential bias, we adopted an oversampling approach that led to the balancing of the dataset in figure 4. After visualising the dependent variable, we proceeded to the distribution of our independent variables and correlation.

- **Visualizing the distribution of the independent variables**

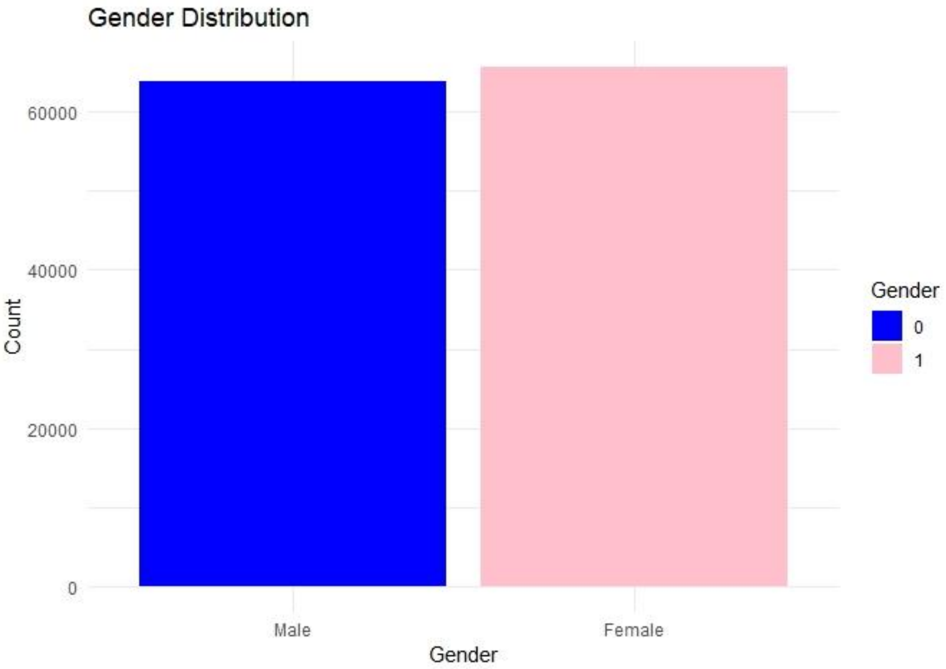


Figure 5: Visualization of gender distribution

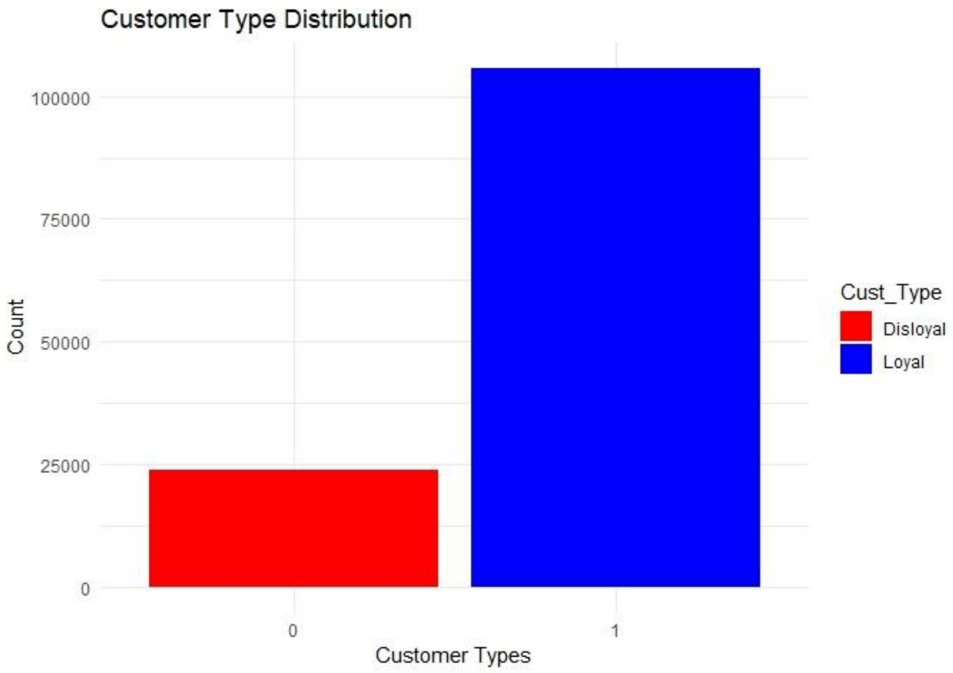


Figure 6: Customer type distribution

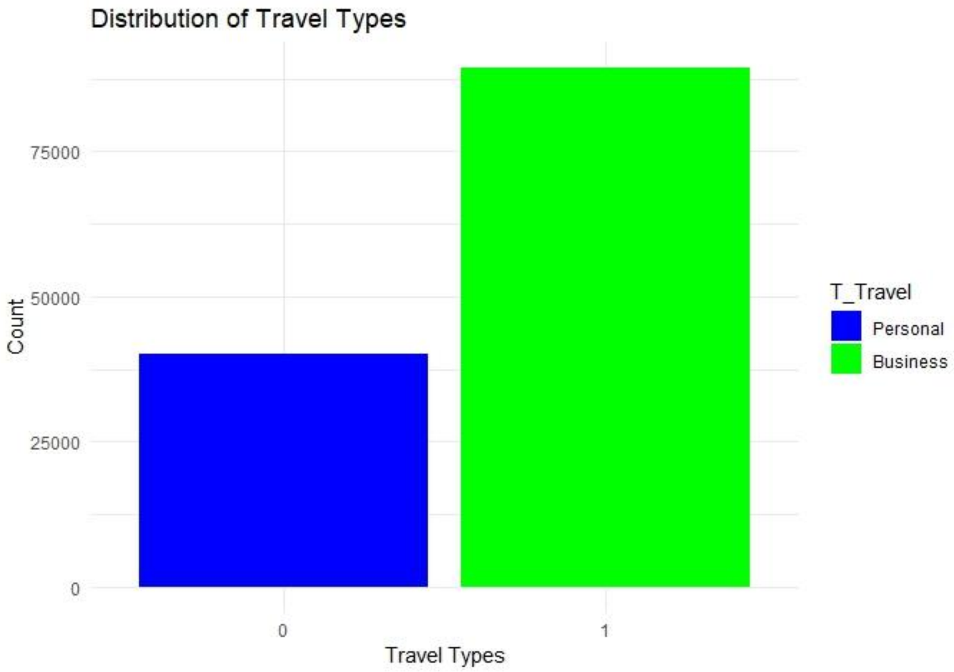


Figure 7: Distribution of types of passenger travel

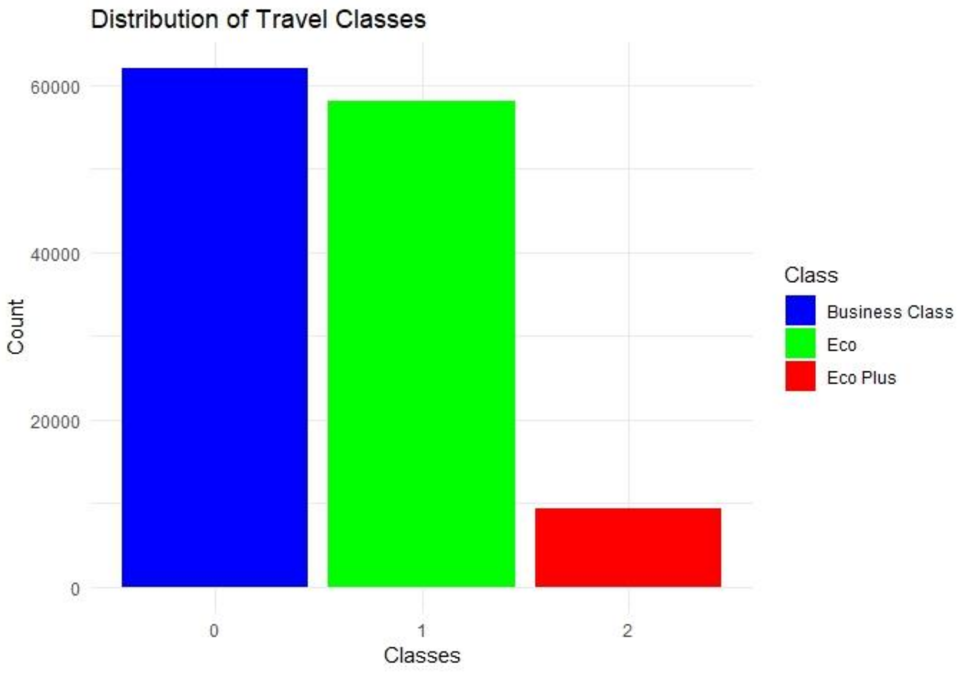


Figure 8: Distribution of travel classes

From figure 5, we can observe the relative distribution of male and female passengers in the dataset. This suggests that there are more female passengers than male passengers in the dataset. The gender variable shows a slight distribution but is not perfect. Figure 6 visually shares the distribution of customer types, revealing that the number of "loyal" customers exceeds the number of "disloyal" customers. This observation allows us to view the imbalance variable as the rate at which one class notably outweighs the other. Additionally, in the visual representation of the distribution of travel types in figure 7, it is evident that there exists a notable imbalance between business and personal travel categories.

The graphical interpretation highlights that the count of business travel instances significantly outweighs that of personal travel instances. This visualisation underlines the imbalanced nature of the variable with respect to travel types, emphasising the need for careful consideration when interpreting results and making inferences from the data. We can also see that the class distribution in figure 8 is not perfectly distributed, as both business and the economy classes outweigh the economy plus.

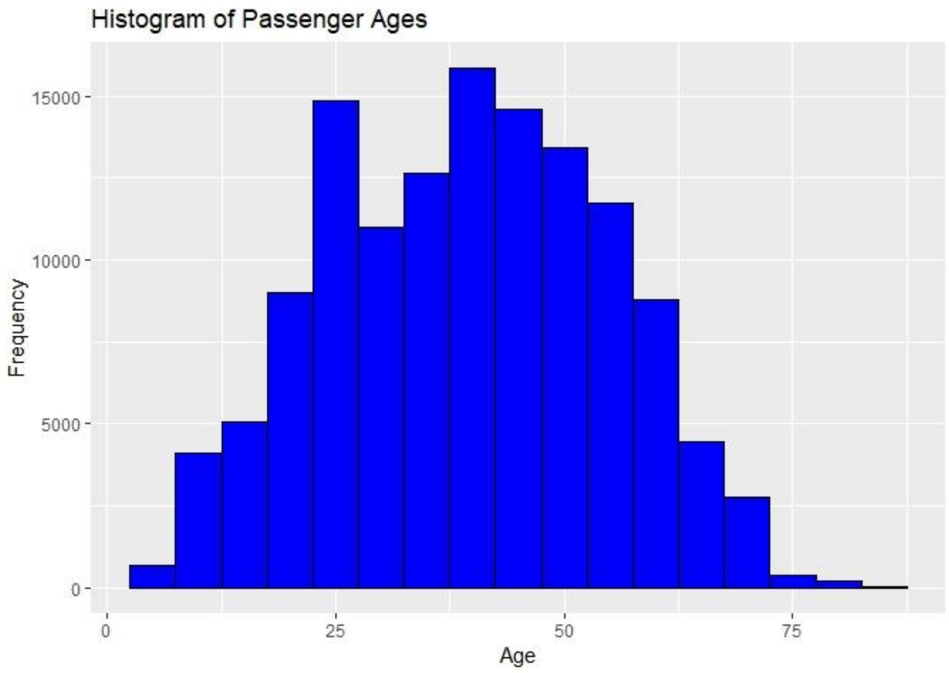


Figure 9: Distribution of passenger's age

The histogram of Passenger ages provides a visual representation of the age distribution among airline customers. The histogram shows the frequency of age groups, with the x-axis representing age and the y-axis representing the frequency count. The distribution appears to be slightly right skewed, with most customers falling within the middle age range. The mean age, which is around 39.43, indicates the central tendency of the age distribution. Similarly, the median age of 40.00 provides a robust measure of the centre, confirming the presence of right-skewness as it is slightly less than the mean. This observation suggests that the customer base tends to be concentrated around a specific age range, potentially influencing marketing strategies and service offerings to cater to the preferences of this age group. This insight into age distribution can inform targeted advertising and service development efforts.

- **Visualizing the distribution of the satisfaction score based on the airline service delivery.**

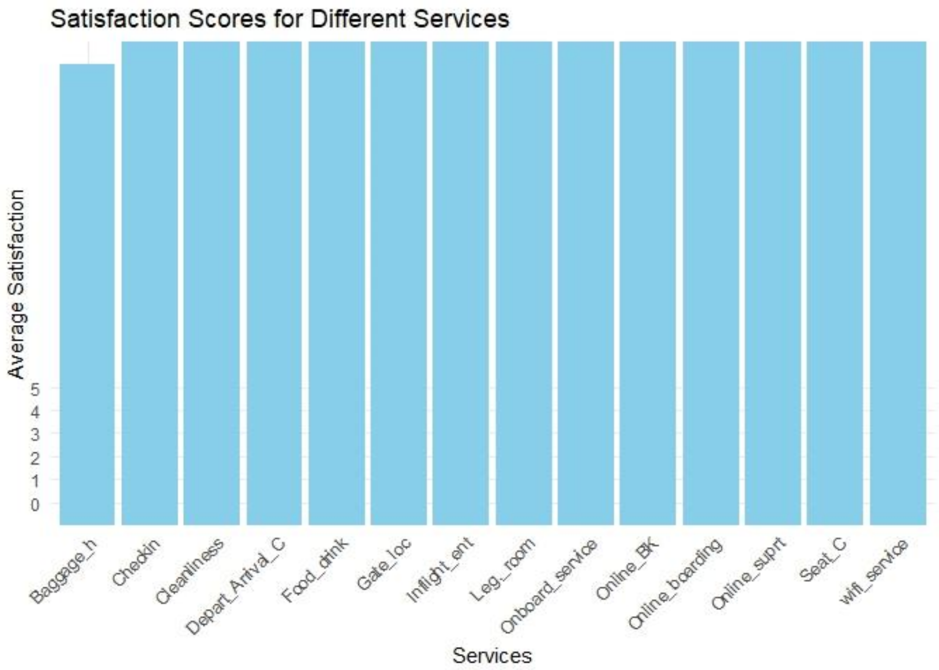


Figure 10: Distribution of airline service delivery

The visualised result illustrates the average satisfaction scores for various airline services such as seat comfort, food and drinks, online support services, inflight entertainment, cleanliness, etc. Among these services, all show a relatively consistent level of average satisfaction, except for "Baggage handling" which exhibits a slightly lower average satisfaction score. This variation in satisfaction scores across services indicates that customers generally have similar levels of satisfaction with most aspects of the airline experience. However, the lower satisfaction score for "Baggage handling" suggests that customers might have comparatively fewer positive perceptions or experiences related to baggage handling. Further analysis and investigation would be carried out to understand the specific reasons behind this observed difference in satisfaction levels.

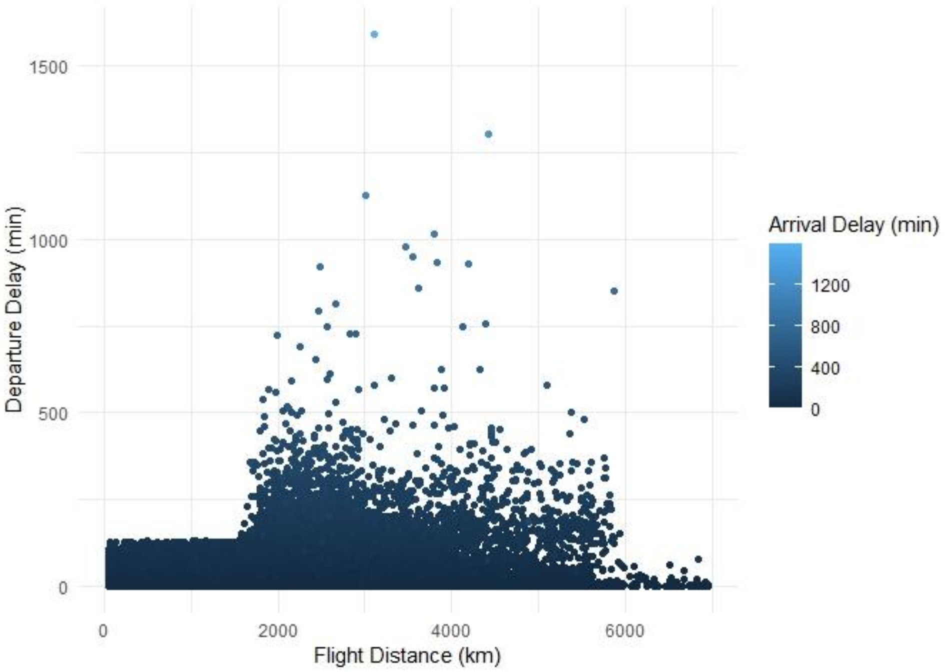


Figure 10: Distribution of departure and arrival delay in minute vs. flight distance

The scatter plot allows us to visualise the distribution of flight distance, departure delay, and arrival delay. The concentration of data points towards lower values of departure delay in minutes and flight distance, suggesting that most flights have minimal delays and shorter distances. However, the flight distance doesn't affect the departure delay in terms of when the flight will arrive at its destination.

Furthermore, we move forward to checking the correlation between the independent variables. To understand how each variable correlates with each other, this process will help us determine which variables to drop before proceeding with the machine learning modelling.

5.3 Correlation analysis

Correlation is a method that assesses the relationship or association between variables. It allows us to have an understanding or insight into how trends and patterns between two or more variables. With correlation analysis, we can comprehend how changes in one variable impact changes in another. It does not always imply causation; it can show whether changes in one variable tend to be correlated with those in another.

- **Visualising the independent variables that are moderately or strongly correlated**

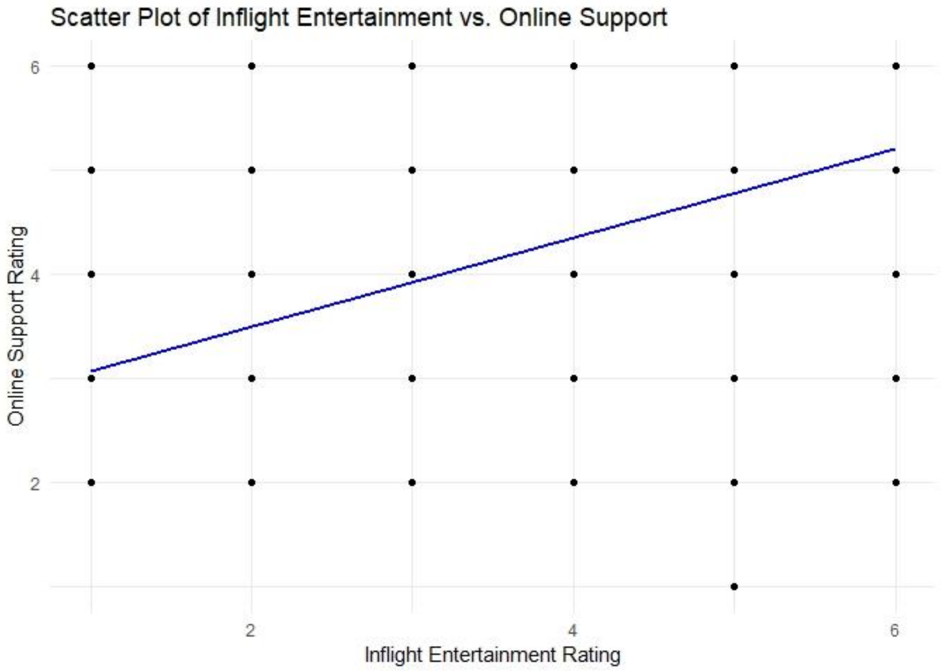


Figure 12: Correlation between Inflight entertainment vs Online support

The plot shows that the correlation coefficient is close to 1, indicating a strong positive relationship. This is because inflight entertainment and online support exhibit a moderately strong positive correlation (r ≈ 0.67). This suggests that higher ratings for in-flight entertainment are often associated with higher ratings for online support.

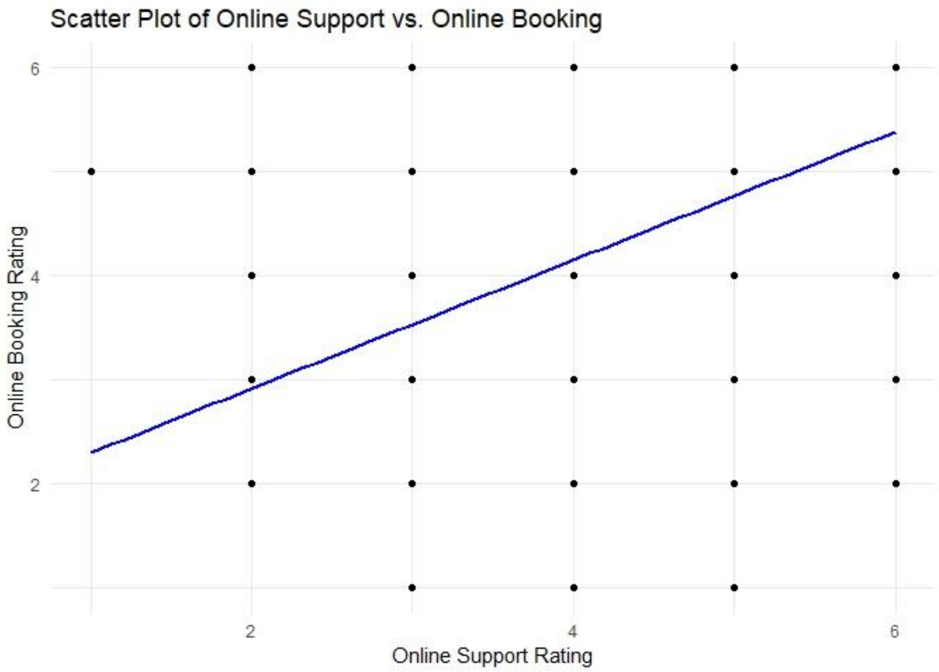


Figure 13: Correlation between online support vs online booking service

This scatter plot shows us how "Online Support Rating" and "Online Booking Rating" variables provides valuable insights into their relationship. The calculated correlation coefficient of 0.6176415871 signifies a positive correlation, indicating a constructive linear relationship between these two variables.

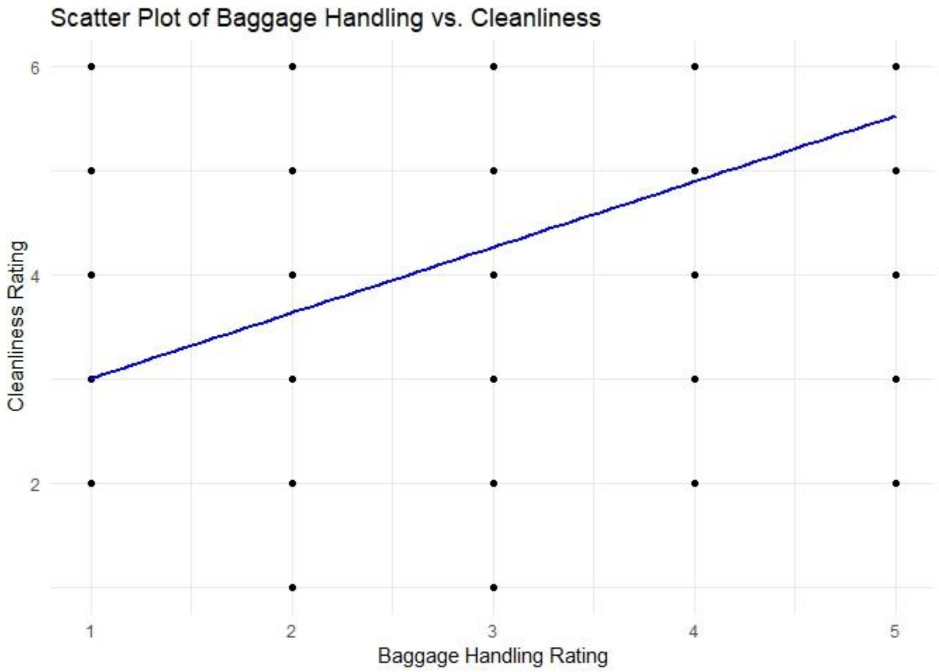


Figure 14: Correlation between baggage handling vs cleanliness

The scatter plot between "baggage handling rating" and "cleanliness rating" shows a positive correlation (correlation coefficient: 0.6320466622). Passengers who give higher ratings for baggage handling also tend to rate cleanliness more positively. The linear regression line adds weight to this observation.

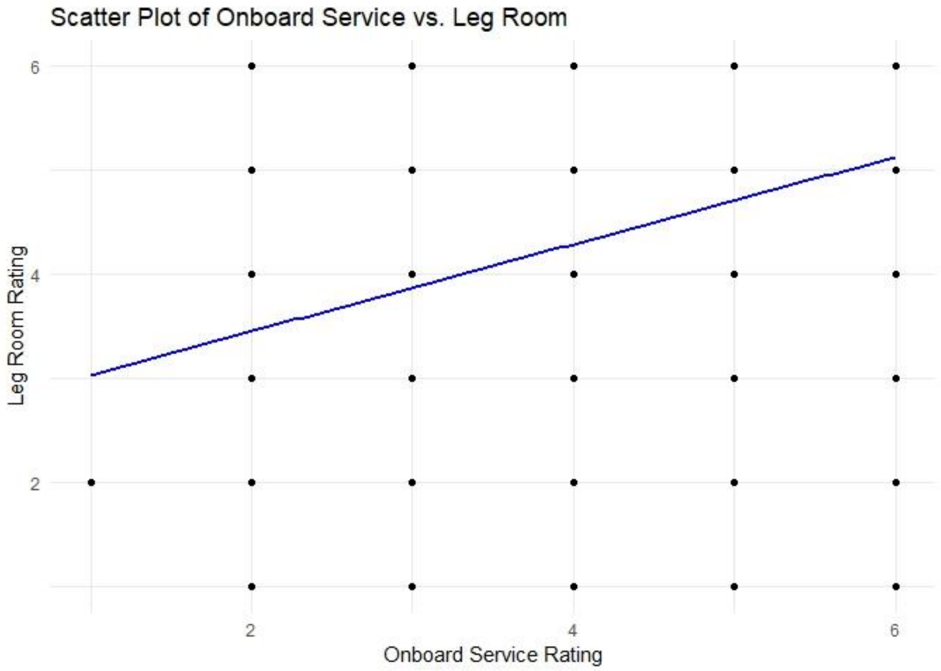


Figure 15: Correlation between onboarding vs leg room service

The scatter plot between "baggage handling rating" and "cleanliness rating" shows a positive correlation (correlation coefficient: 0.6320466622). Passengers who give higher ratings for baggage handling might tend to rate cleanliness more positively. The linear regression line adds weight to this observation.

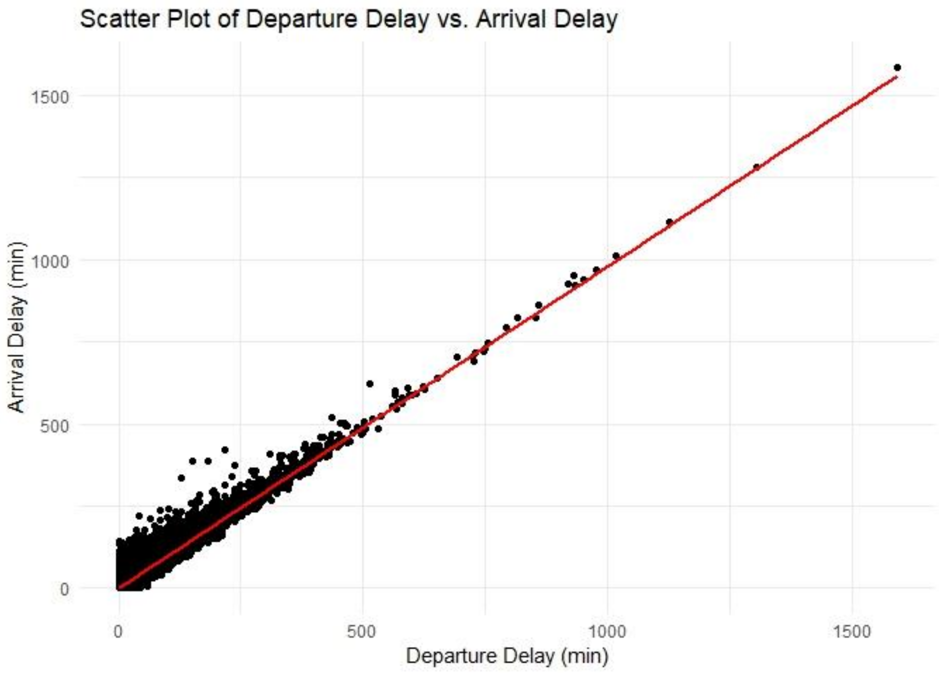


Figure 16: Correlation between arrival and departure delay in min

The plot between "Departure Delay (min)" and "Arrival Delay (min)" reveals a strong positive correlation (correlation coefficient: 0.9652911835). Points are strongly clustered along a rising trend line, indicating that flights with longer departure delays also experience longer arrival delays. The linear regression line confirms this relationship.

- **Heatmap for independent variables that are correlated.**

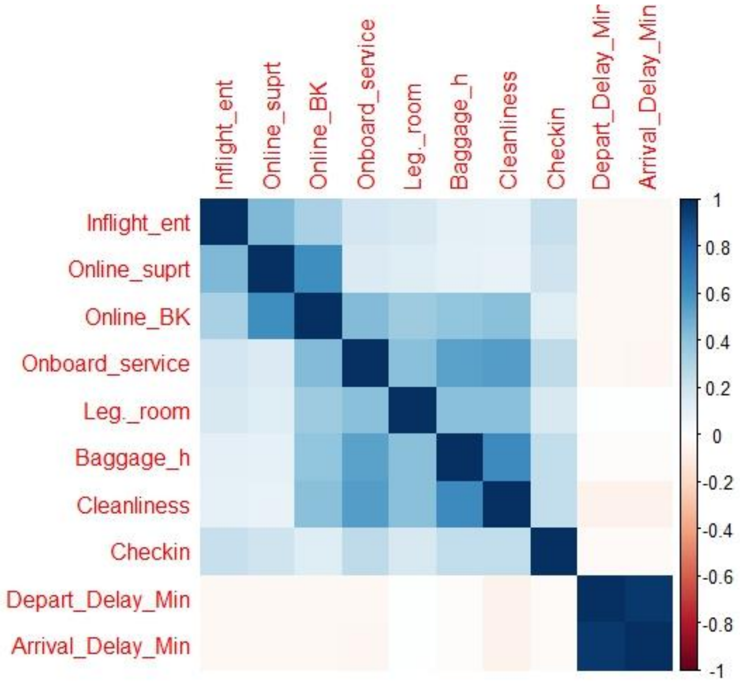


Figure 17: *The heatmap provided offers a graphical representation of strongly and moderately high correlation coefficients among different pairs of independent variables in the dataset.*

As explained in figure 10, a significant positive correlation exists between passengers' evaluations of in-flight entertainment and online support. This coefficient indicates that as passengers rate in-flight entertainment higher, they are inclined to rate online support higher as well. while for in-flight entertainment and online booking ratings. This suggests that passengers with favourable perceptions of in-flight entertainment are likely to hold favourable opinions of online booking services. Figure 11 shows that as passengers rate online support more positively, they tend to exhibit similar positive ratings for online booking. Also, a moderate positive correlation is observed between passengers' assessments of onboard service quality and available legroom (Figure 13). This implies that passengers who express higher satisfaction with onboard service also tend to hold favourable views regarding legroom. A relatively weaker positive correlation is present between evaluations of check-in service and cleanliness, as shown oppositely in the map. This suggests a mild association between these factors; passengers who rate check-in service positively may also have a modest tendency to rate cleanliness positively.

Additionally, a strong positive correlation describes the connection between passengers' perceptions of baggage handling and overall cleanliness. Meaning that passengers who provide higher ratings for baggage handling also tend to provide correspondingly higher ratings for cleanliness. We can also see that Figure 14 shows a highly significant positive correlation between departure delay and arrival delay. This implies that instances of long departure delays are closely associated with long arrival delays.

The main goal of locating and analysing the strongest correlations between independent variables is to reduce the problem of multicollinearity. According to Voss (2005) and Frgestad et al. (2009), multicollinearity is a problem with models that makes it difficult to separate the individual effects of each variable. According to Voss (2005), this can result in unreliable coefficient estimates and a model that is harder to interpret. Therefore, it is a good practice to consider eliminating one of the correlated variables or using techniques to address multicollinearity before processing your data for modelling (Voss 2005).

- **Visualising the dependent variables that are correlated with the independent variable.**

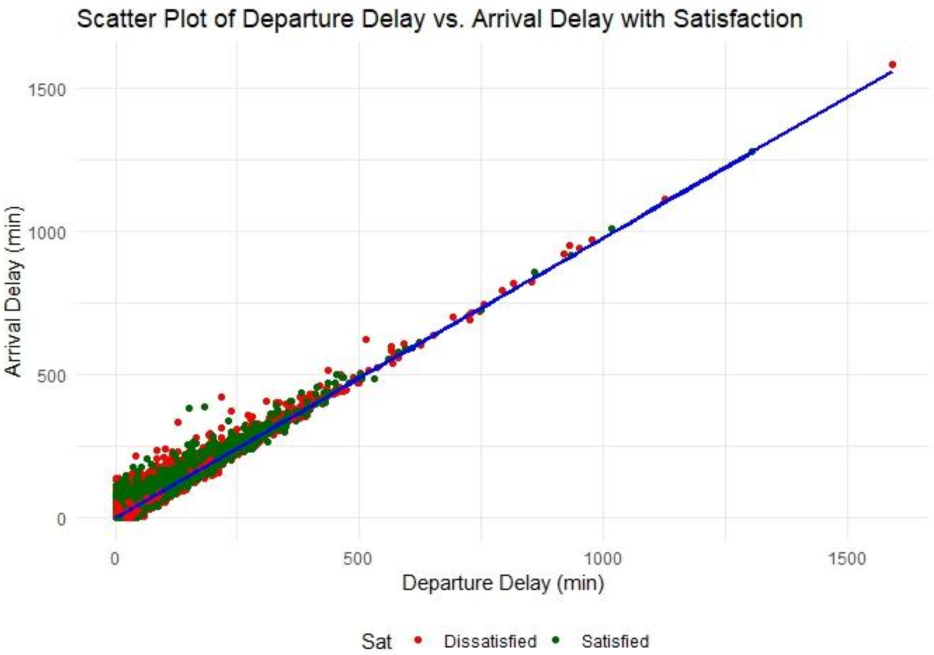


Figure 18: Departure and arrival delay in min vs Satisfaction

The scatter plot illustrates a strong positive linear relationship, and the high correlation coefficient indicates a nearly one-to-one correspondence between departure delay and arrival delay. The high correlation indicates that the two delays are closely linked, with one often causing the other. This implies that longer departure delays are likely to result in longer arrival delays.

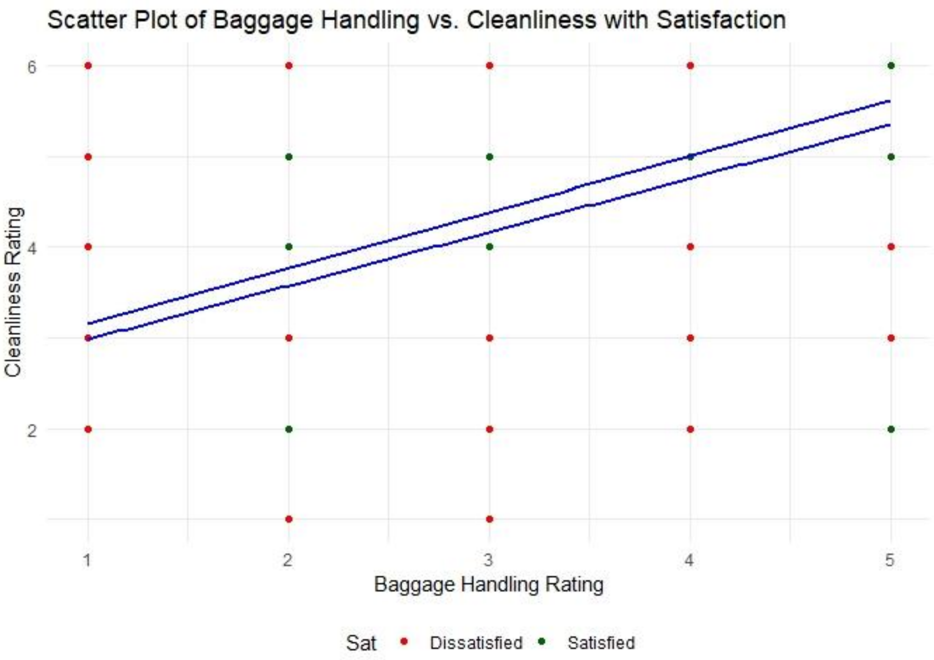


Figure 19: Baggage handling and cleanliness vs Satisfaction

The scatter plot shows a positive linear trend. Passengers who rate baggage handling positively also tend to give higher ratings for cleanliness. This suggests that positive experiences with baggage handling might correspond to positive perceptions of cleanliness on the flight.

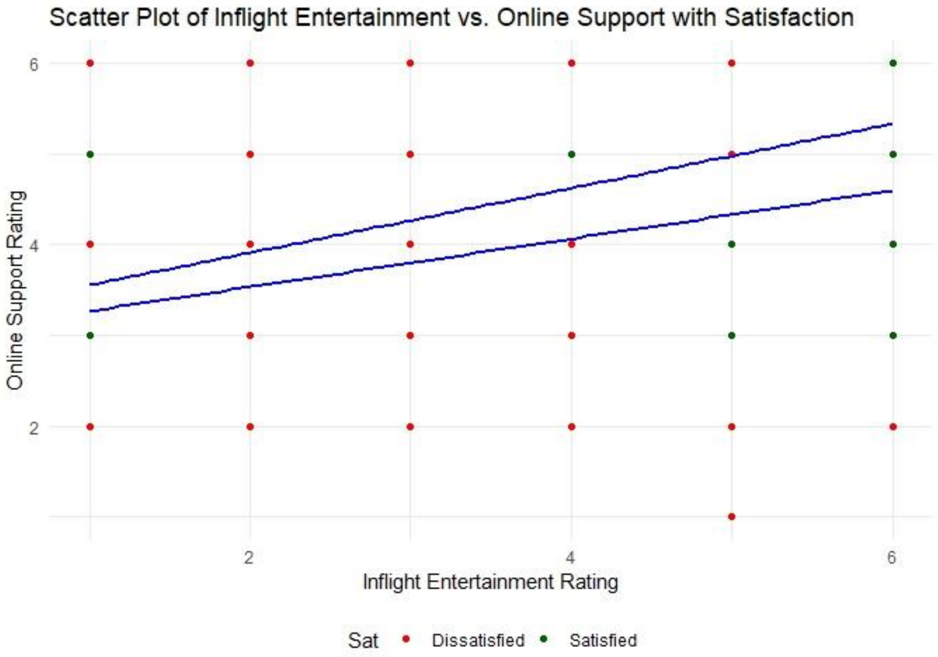


Figure 20: Inflight entertainment and online support vs Satisfaction

Similarly, the scatter plot demonstrates a positive but nor perfect linear relationship. Passengers who rate the in-flight entertainment most likely tend to give higher ratings for online booking services. This suggests a potential connection between a positive in-flight entertainment experience and favourable opinions about the online booking process.

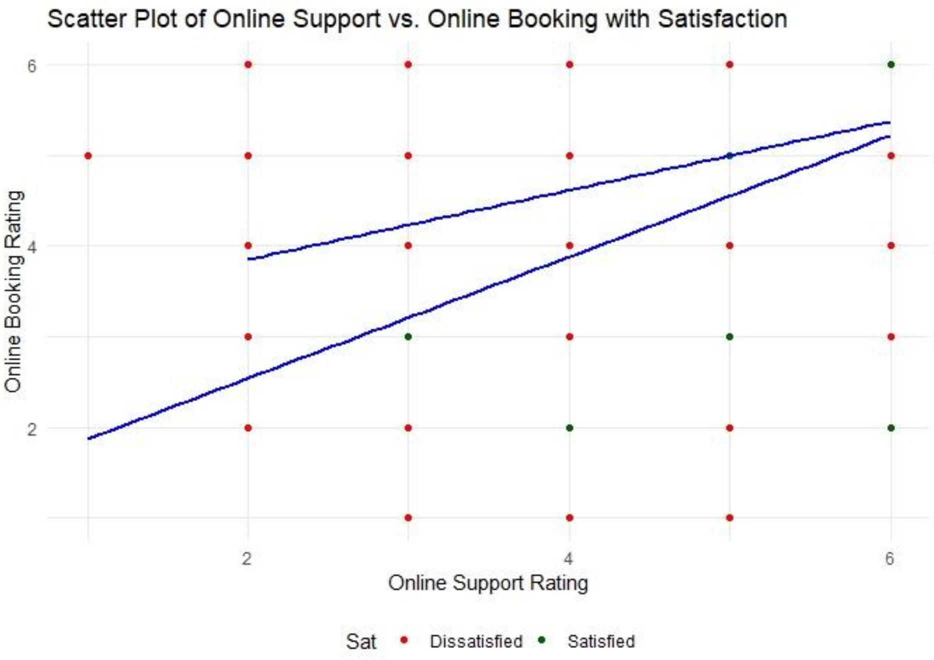


Figure 21: Online support and online booking vs Satisfaction

Figure 21 also illustrates a not so positive linear trend. Which means that passengers who rate online support mighty likely provide higher ratings for online booking services. This indicates that positive experiences with online support may lead to favourable perceptions of the online booking system.

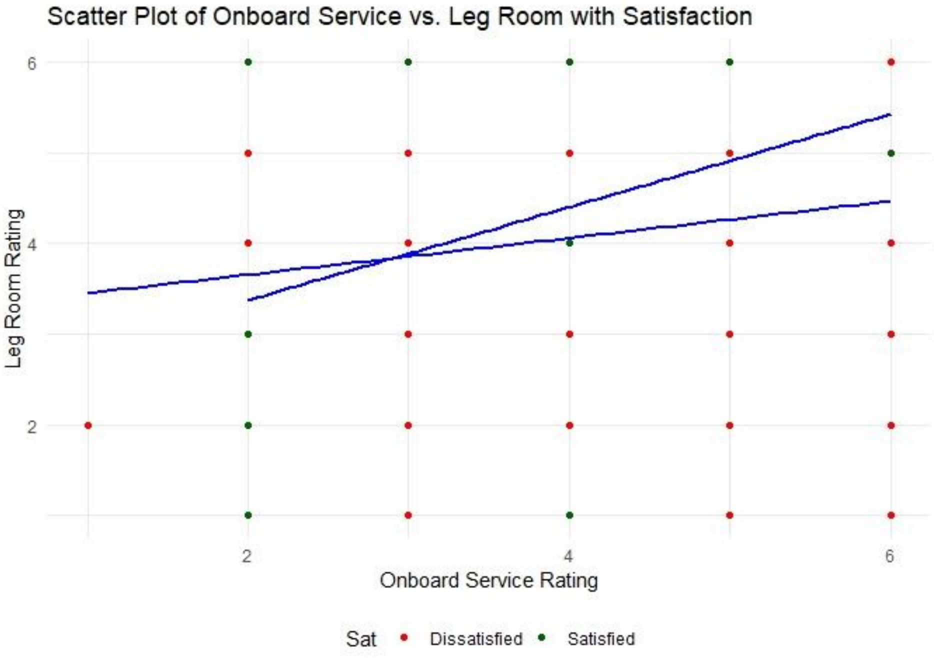


Figure 22: Online boarding and Leg room service vs Satisfaction

The scatter plot indicates a positive linear relationship, but the correlation coefficient is relatively low. Passengers who rate check-in processes positively also tend to rate cleanliness positively. However, the weaker correlation suggests that the relationship is not as pronounced.

- **Correlation heatmap for all the independent variables vs the independent variables**

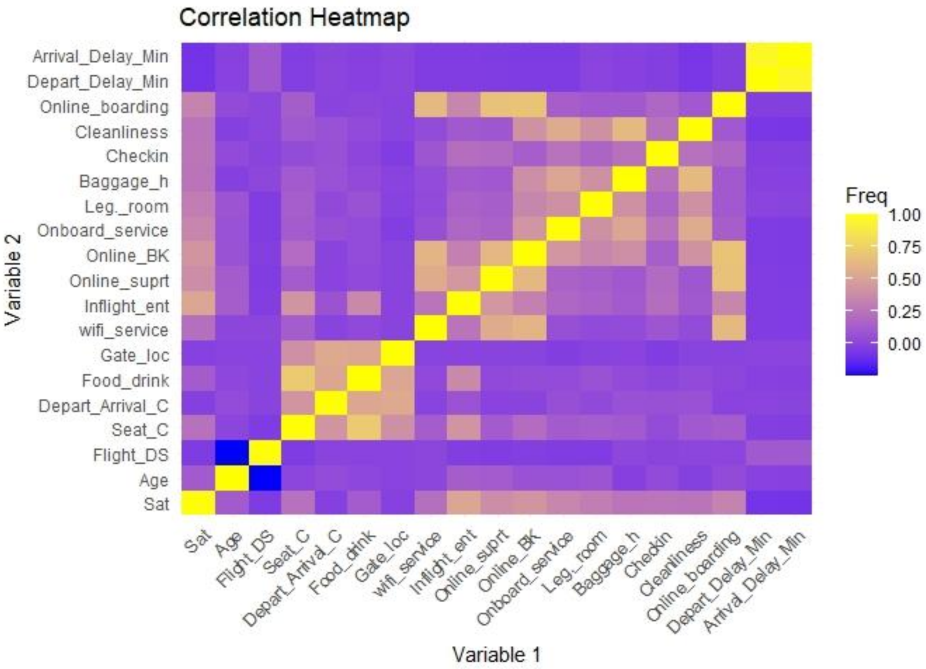


Figure 23: Correlation heatmap

In the context of our research endeavour, the process of converting the correlation matrix into a heatmap offers us a valuable visual tool for comprehending the intricate relationships among our selected variables.

This visualisation technique employs colours to encapsulate the correlation coefficients within the matrix, thereby facilitating a more intuitive grasp of the interconnections within our dataset. The blue squares in the heatmap indicate negative correlations between flight distance and age. This coloration clearly shows that when the value of one variable rises, the corresponding variable usually shows a decline. In this case, there is a tendency for "flights distance" to decrease as "age" increases. On the other hand, the heatmap's yellow hues correspond to the positive correlations represented by yellow squares in the matrix. This means that there is a propensity for another variable to increase when the value of one increase. For example, food and drink and seat comfort suggest that as passengers' satisfaction with food and drink rises, so too does their satisfaction with the comfort of their seats.

Furthermore, when the heatmap contains white squares, which indicate weak or negligible correlations, these are depicted as lighter or neutral colours. This is an indication that changes in one variable's value don't significantly influence the other variable's value. An example will be the near-zero correlation value between "gate location" and age," which is reflected by the lack of a pronounced colour in the heatmap, reaffirming the absence of a substantial relationship between these variables.

5.4 Statistical testing

In addition to observing the correlation coefficient of the independent variable to the dependent variable variables, we went further to carry out a K-s statistical test. To see statistical significance. According to Boyerinas (2016), the Kolmogorov-Smirnov (KS) test is used to compare the distribution of a sample with a reference distribution or to compare the distributions of two samples. In this case, we are comparing the distribution of the dependent variable ("Sat") against the distributions of some independent factors to see if they are significantly different.

Usually, when conducting a statistical test, if the p-value associated with the test statistic is very small (typically less than the chosen significance level, often 0.05), we reject the null hypothesis. This implies there is strong evidence for a significant effect, difference, or relationship in the population beyond what can be attributed to random chance.

If the p-value is not very small (greater than or equal to the chosen significance level), we fail to reject the null hypothesis. This indicates that the observed results are reasonably consistent with what could happen by random chance. However, it doesn't prove the null hypothesis is true; rather, it means we lack strong evidence for a significant effect, difference, or relationship.

Table 2: Statistical test

	Independent Variable	**Kolmogorov-Smirnov test**	**P-Value**	**Null Hypothesis**
1	Seat Comfort	0.33845	< 2.2e-16	Reject the Hypothesis
2	Departure/Arrival convivence	0.02406		
3	Food and drinks	0.18336		
4	Gate location	0.069941		
5	Inflight entertainment	0.59969		
6	Online support	0.42122		
7	Online booking	0.44282		
8	Onboarding service	0.34159		
9	Leg room	0.32757		
10	Baggage handling	0.27081		
11	Cheak-in service	0.21379		
12	Cleanliness	0.26538		
13	Inflight Wi-Fi	0.19827		
14	Online boarding	0.28864		
15	Age	0.1694		
16	Gender	0.21299		

17	Customer type	0.22734		
18	Type of travel	0.10113		
19	Class	0.31306		
20	Flight distance	0.14056		
21	Departure	0.065256		
22	Arrival in minute	0.10572		

Based on the results of the Kolmogorov-Smirnov tests, the p-values for all the independent variables tested against the dependent variable "Sat" (satisfaction) are significantly less than the significance level of 0.05 (p-value < 0.05). Hence, we reject the hypothesis test, seeing that all the independent variables have statistically significant differences in their distributions between the "dissatisfied" and "satisfied" groups in relation to passenger satisfaction, as seen in Table 2.

After understanding the distribution and relationship with the through visualisation and correlation, we move on to training and testing our machine learning model to make predictions. To obtain an accurate result before running our model. We did future selection by removing the departure delay in min in logistic regression, KNN and Naïve bayes.

6 Machine learning modelling and evaluation

Machine learning models can be defined as algorithms that enable computers to learn from a group of datasets and make predictions or decisions based on the dataset. It is designed to understand relationships, trends, and patterns in a dataset and then use these insights to make predictions for unseen data. There are three major types of models that are suited for different types of datasets: supervised, unsupervised, and reinforcement learning.

The supervised models are trained on labelled data, where the input data is paired with the correct output. They include logistic and linear regression, naive bayes, K-nearest neighbour, Radnom forest, decision tree, support vector machine, and neutral techniques. The goal is for the model to learn the relationship between input and output so that it can accurately predict the output for new inputs.

The unsupervised model has to do with unlabeled data, where its task is to find patterns, clusters, or structures within the data, while the reinforcement model learns through interaction with an environment. They receive feedback in the form of rewards or penalties for the actions they take, allowing them to learn optimal strategies over time. For the purposes of this research, we used logistic regression, Naive Bayes, K-nearest neighbour, Radnom forest, and decision trees to train and test our dataset. The performance of these models was evaluated using sensitivity (recall), accuracy, specificity, and F1 score. The highest score will determine the best model that predicts customer satisfaction. Before proceeding, the data was partitioned into training sets of 0.7 and testing sets of 0.3, respectively.

6.1 Logistic regression model and result evaluation

As a statistical technique for binary classification, logistic regression is primarily used to estimate the likelihood that an observation will fall into one of two possible classes or categories. It is a kind of generalised linear model (GLM) and is especially helpful when the dependent variable is categorical and the relationship between the independent variables and the result is not linear. Fritz and Berger (2015). In this analysis, the logistic regression model was built to predict what services in the airline industry can cause customer satisfaction and identify what services can cause dissatisfaction.

The logistic regression model, as seen below, was well-fitted to the training data, and the coefficient estimates provide insights into all predicted variables, excluding the departure arrival time, that are statistically significant in predicting customer satisfaction within the airline industry. This means that the model has the potential to provide valuable insights and predictions once it moves on to the prediction phase. Future selection was carried out to obtain an accurate picture of our model's performance.

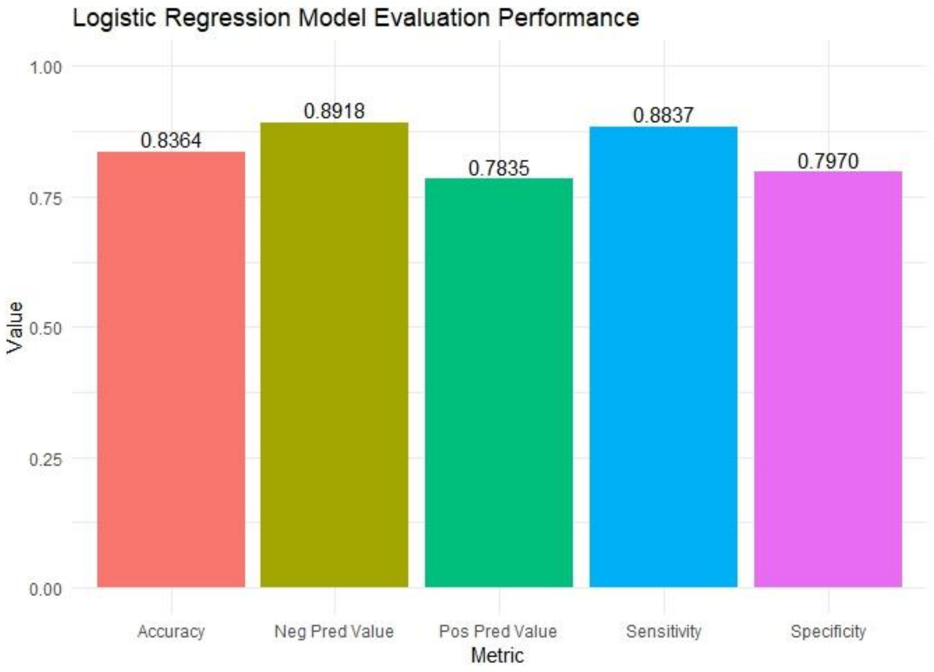


Figure 24: Logistic regression result

Precision (Pos Pred Value): Precision is a critical metric that evaluates the accuracy of positive predictions made by our classification model, specifically in the context of predicting passenger satisfaction, where "Satisfied" is represented as 1.

In this analysis, the logistic regression precision for passenger satisfaction (class 1) is approximately 0.8835. This signifies that when the model predicts passenger satisfaction, it is correct approximately 88.35% of the time. In other words, it demonstrates a high level of accuracy in identifying satisfied passengers.

Recall (Sensitivity): Recall, also referred to as sensitivity, is a fundamental measure assessing our model's ability to correctly capture all actual instances of passenger satisfaction, where "Satisfied" is represented as 1.

The logistic regression recall for passenger satisfaction (class 1) is approximately 0.7833. Which implies that the model effectively identifies about 78.33% of all genuinely satisfied passengers. In practical terms, it indicates that the model is successful in recognizing a substantial proportion of passengers who are truly satisfied with the airline service.

F1 Score: The F1 score serves as a comprehensive metric that strikes a balance between precision and recall. It is particularly useful when we need to assess the model's performance while considering both the accuracy of positive predictions and its ability to identify actual positive instances.

For the logistic regression the F1 score for passenger satisfaction (class 1) is approximately 0.8291. This metric summarizes the balance between making precise positive predictions and correctly capturing actual instances of passenger satisfaction. An F1 score of 0.8291 indicates that our model demonstrates a balanced performance in identifying satisfied passengers with an appropriate blend of accuracy and coverage.

Precision (Pos Pred Value) and Recall (Sensitivity): Calculation:

Precision = True Positives / (True Positives + False Positives) Precision = 15580 / (15580 + 2050) ≈ 0.8835

Recall = True Positives / (True Positives + False Negatives) Recall = 15580 / (15580 + 4305) ≈ 0.7833

F1 Score: The F1 score balances precision and recall and is calculated as:

*F1 Score = 2 * (Precision * Recall) / (Precision + Recall) F1 Score = 2 * (0.8835 * 0.7833) / (0.8835 + 0.7833) ≈ 0.8291*

6.2 K-nearest neighbour model and result

According to Kumar (2020), the KNN non-parametric algorithm classifies data points based on their proximity to labelled data points. It assigns a new data point to the majority class of its k nearest neighbours. K-nearest neighbours is a simple and intuitive classification algorithm that is often used for pattern recognition and classification tasks. In KNN, the class of a data point is determined by the class of its k-nearest neighbours in the feature space.

Before performing the KNN modelling, the dataset was scaled since unscaled numerical variables can cause the model to be ineffective based on its sensitivity. And the only independent variable that was taken out is the departure delay in minutes.

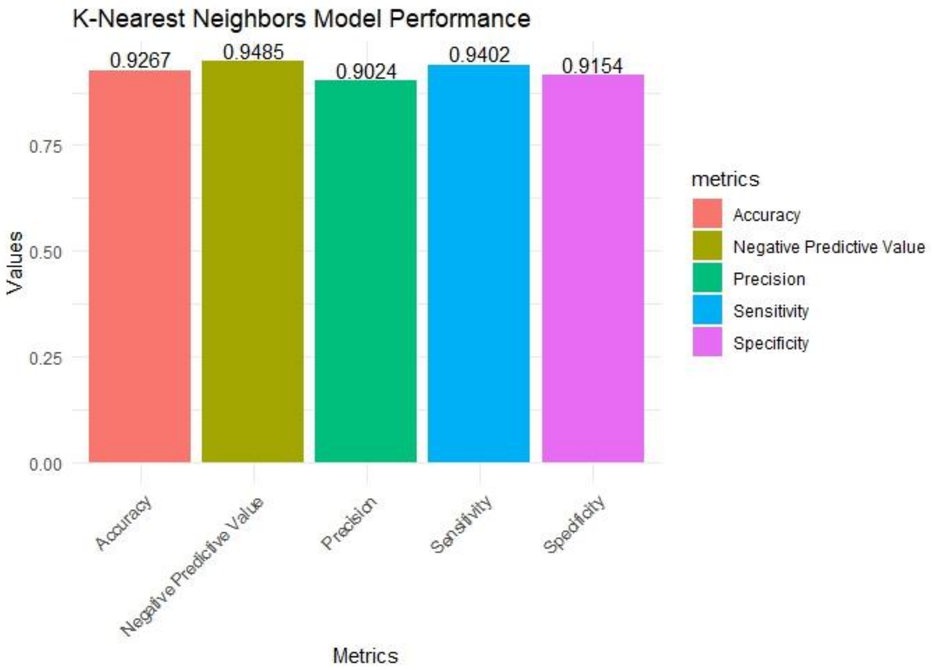


Figure 25: K-nearest neighbour result

The K-Nearest Neighbours (KNN) classification model result shows an F1 score of approximately 0.8917. This metric captures the model's ability to make accurate positive predictions (precision) while effectively capturing actual positive instances (recall).

The score signifies that our KNN model strikes a balanced performance, demonstrating ability in accurately identifying passenger satisfaction (class 0) while appropriately capturing genuinely satisfied passengers. This score indicates the model's effectiveness in maintaining an appropriate balance between precision and recall, making it a valuable tool for the task of passenger feedback analysis in the airline industry.

6.3 Naive Bayes model

According to Berrar (2018), The Nave Bayes theorem is a crucial component of science because it uses logical reasoning to update the likelihood of hypotheses considering new evidence. A probabilistic machine learning algorithm called the Naive Bayes model is employed for classification and occasionally regression tasks. It is based on the Bayes theorem, which estimates the likelihood of an event occurring based on knowledge of the circumstances surrounding the event in the past. The "naive" in Naive Bayes refers to the presumption of feature independence, which simplifies calculations and increases the computational efficiency of the algorithm. The analysis's Bayes performance is shown in the table below. Berrar (2018),

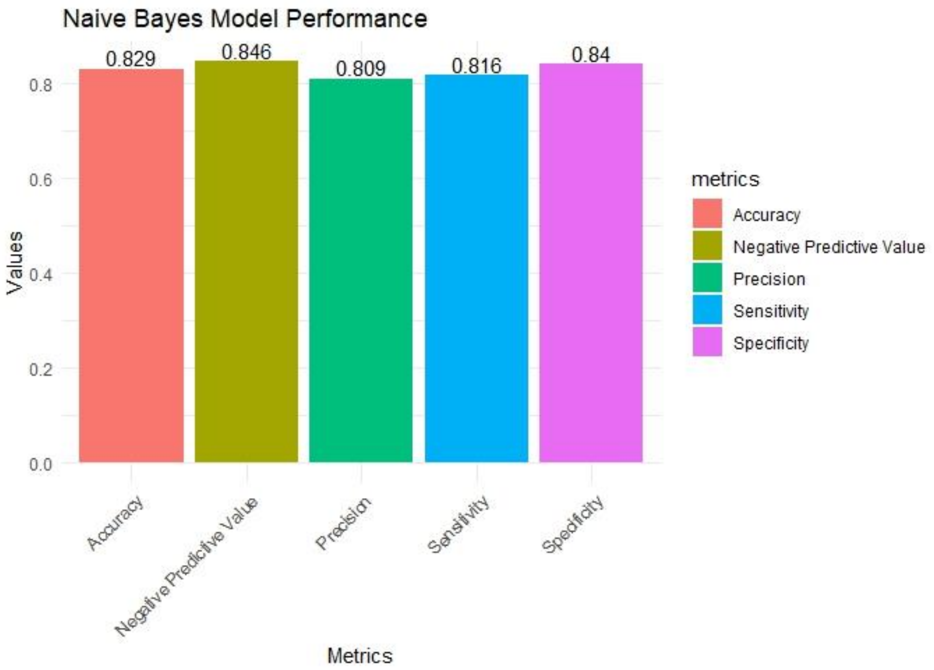


Figure 26: Naïve Bayes model result and evaluation

The F1 score of 0.8092 signifies our Navie bayes model's proficiency in maintaining a balanced between precision and recall. It demonstrates the model's capability to make accurate predictions of passenger satisfaction (class 1) while effectively capturing actual instances of satisfaction. This score underscores the model's robustness and suitability for the task of classifying passenger feedback in the airline industry, where achieving the right balance between precision and recall is of paramount importance.

"In the context of assessing the performance of our Naive Bayes classifier for predicting passenger satisfaction, we employ the F1 Score as a crucial evaluation metric. The F1 Score is particularly valuable in scenarios where there is an imbalance between the classes, such as in this study where we distinguish between 'satisfied' and 'dissatisfied' passengers.

6.4 Decision Tree

This is a type of machine learning technique create a tree-like model of decisions and their possible consequences. They are effective in capturing complex interactions between features and can handle both categorical and numerical data (de Ville, 2013).

The algorithm breaks down a dataset into smaller subsets by making decisions based on feature values, ultimately leading to the prediction of a target variable. Below is the performance of the decision tree in our analysis.

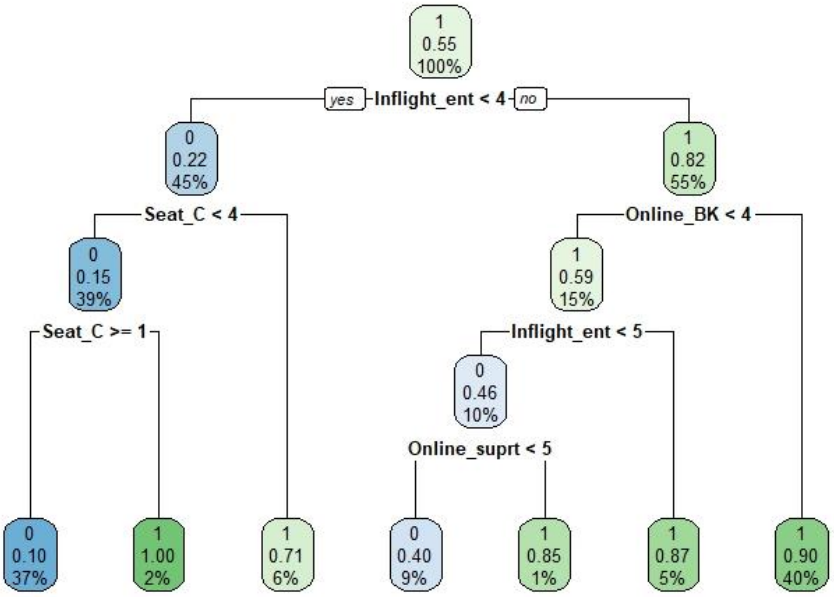


Figure 27: Decision Tree result and evaluation

Below, the decision tree model's performance is evaluated through the provided confusion matrix and associated statistics.

Model Performance: The Decision Tree successfully classified passenger feedback into two classes: satisfied (1) and dissatisfied (0). The confusion matrix reveals the following:

- True positives (predicted as satisfied and actually satisfied): 14,377
- False positives (predicted as satisfied but actually dissatisfied): 3,391
- True Negatives (predicted as dissatisfied and actually dissatisfied): 17,814
- False negatives (predicted as dissatisfied but actually satisfied): 3,253

These metrics reveal that the model excels at correctly identifying actual satisfied passengers while maintaining a strong ability to identify actual dissatisfied passengers.

Positive Predictive Value: The positive predictive value (PPV) is 80.92%, indicating that when the model predicts satisfaction, it is accurate approximately 80.92% of the time.

Negative Predictive Value: The negative predictive value (NPV) is 84.56%, showcasing the model's accuracy when predicting dissatisfaction.

Based on our analysis, we can see that the decision tree model performs admirably in predicting passenger satisfaction from airline service feedback. It effectively maintains a balance between precision and recall, making it a valuable tool for the airline industry. With an accuracy of 82.89%, our model provides a robust framework for classifying passenger feedback and offers valuable insights into four main service quality areas such as inflight entertainment, online booking, online support, and seat comfort that can lead to passenger satisfaction. Table 3 shows the decision tree result. We go ahead and use the random forest to compare this result.

6.5 Random Forest Model

Random forests are an ensemble of decision trees. They combine multiple decision trees to make predictions and reduce the risk of overfitting. Random forests are known for their high accuracy and robustness. (E R, 2021). It's an extension of the decision tree algorithm that aims to improve predictive accuracy and control overfitting. Let’s look at how the random forest model performed in our analysis.

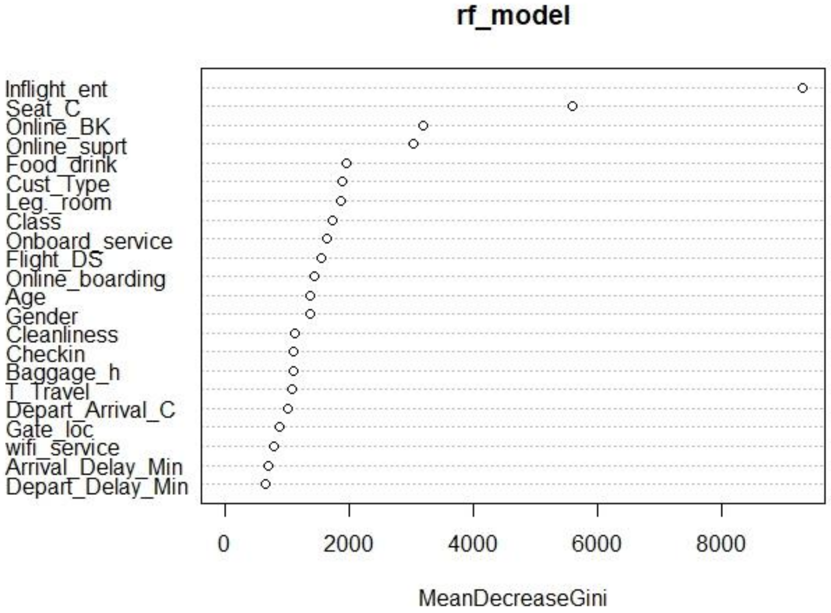


Figure 28: Random Forest model

We evaluated the performance of the Random Forest classification model for predicting airline passenger satisfaction based on feedback. The result in table 3 shows that the model can predict 96% accurately the four major factors that can influence customer satisfaction in the airline industry, which are Inflight entertainment, Seat comfort, Online booking, and Online support. While arrival delay, Wi-Fi service and gate location how dissatisfied customers are with such services. This result

These results collectively indicate the Random Forest model's robust performance in predicting airline passenger satisfaction. With high accuracy, sensitivity, specificity, and strong agreement with the actual data, the model effectively discriminates between satisfied and dissatisfied passengers, making it a valuable tool for analysing and improving airline services.

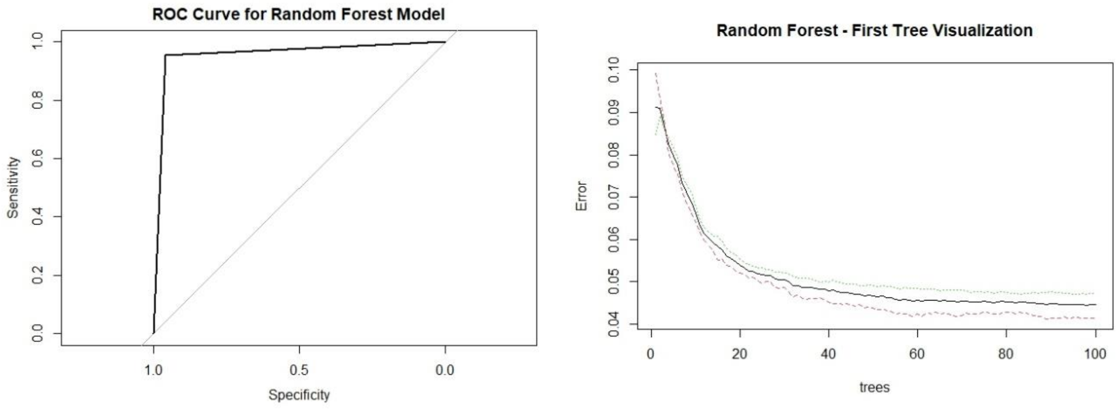


Figure 29: Random Forest ROC Curve and Tree visualization

The ROC curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold values. Each point on the curve corresponds to a different threshold setting. As we can see, the ROC indicate high sensitivity and specificity across various thresholds and a random classifier of the ROC curve close to the diagonal line (45-degree line).

The under-ROC curve (AUC-ROC) summarizes the overall performance of the model. A higher AUC-ROC value (closer to 1) indicates better discrimination between satisfied and dissatisfied airline passengers based on their feedback. The result of the random forest seems to give the height and best accuracy among the other model 96% and a high F1 score 0.95%.

6.6 Comparing result with other authors findings

Based on our model analysis, several key factors in airline services significantly influence customer satisfaction. These factors include inflight entertainment, seat comfort, online booking, and online support. Conversely, factors such as departure and arrival delays are identified as major contributors to customer dissatisfaction. This finding is consistent with Ridwan's (2022) study, which investigated the impact of various factors on customer satisfaction in the airline industry using supervised learning techniques. Ridwan's research revealed that both inflight entertainment and seat comfort play crucial roles in influencing customer loyalty, with a predictive accuracy of 96% achieved using a random forest model.

In a similar vein, Noviantoro and Huang (2021) employed supervised machine learning algorithms, including deep learning and random forest, to analyze customer feedback collected from Kaggle. Their study highlighted areas that require improvement to enhance customer satisfaction within carrier services. Specifically, their findings emphasized inflight services, online boarding, baggage handling, and Wi-Fi connectivity as critical areas for airline carriers to focus on.

Furthermore, Park et al. (2019) conducted research emphasizing the importance of enhancing online airline services and the overall in-flight customer experience to bolster both customer satisfaction and repeat business.

Sugara and Purwitasari (2022), along with Yazdi et al. (2020), examined the adverse effects of flight departure delays on customer satisfaction and proposed strategies to mitigate such occurrences. This literature review reveals a plethora of authors who have explored factors influencing customer satisfaction within the airline industry."

Model comparison table

Table 3

Metric	Logistic Regression	K-nearest neighbour	Naive Bayes	Decision Tree	Random Forest
Sensitivity (Recall)	0.8837%	0.9402%	0.816%	0.8574%	0.9589%
Precision (Positive Predictive Value)	0.7835%	0.9024%	0.809%	0.8460%	0.9466%
Accuracy	0.8364%	0.9267%	0.829%	0.8644%	0.9568%
Negative Predictive Value (NPV)	0.8918%	0.9485%	0.846%	0.8801%	0.9655%
Specificity	0.7970%	0.9154%	0.84%	0. 8703%	0.9551%

7 Conclusion

The comprehensive analysis conducted in this study has investigated aspects of customer satisfaction and dissatisfaction within the airline industry. The dataset encompasses various displays of independent variables ranging from seat comfort and in-flight entertainment services to online booking support and departure and arrival delays. Throughout the EDA phase, various visualisation techniques were employed to furnish valuable insights into variable distributions, interrelationships, and correlations with customer satisfaction. These visualisations have not only enhanced the comprehension of the dataset but have also laid the groundwork for subsequent modelling endeavours.

We used machine learning algorithms, including logistic regression, K-nearest neighbours, naive bayes, decision trees, and random forests, to predict customer satisfaction. Each of these models was evaluated using performance metrics such as sensitivity, precision, accuracy, F1 score, and specificity, all of which are integral for the assessment of classification tasks. The outcome of the model comparison table reveals the exceptional performance of both the Random Forest and K-nearest Neighbours models in forecasting the predominant services conducive to customer satisfaction while demonstrating robust performance in predicting customer dissatisfaction.

The research findings have yielded insightful revelations into the factors underpinning customer satisfaction and dissatisfaction in the airline industry. Variables such as seat comfort, in-flight services, online support, and ease of online booking have exhibited noteworthy positive correlations with customer satisfaction. Conversely, departure and arrival delays have emerged as contributory factors to customer dissatisfaction.

Regarding departure and arrival delays, it is imperative to acknowledge the multitude of factors within the airline industry that can engender these delays. Weather-related disruptions, technical anomalies, baggage handling inefficiencies, security checks, immigration and customs procedures, and staffing constraints have all been recognised as potential culprits in delaying flight schedules.

To address and mitigate the impact of departure and arrival delays, it is strongly recommended that the airline industry adopt proactive measures. These measures encompass the deployment of advanced weather monitoring and management systems, enhancements in aircraft maintenance protocols, streamlining of baggage handling procedures, and optimisation of security processes. Furthermore, augmenting staffing levels during peak travel periods and enhancing communication with passengers regarding delays can significantly improve customer dissatisfaction. These strategic initiatives are paramount for strengthening operational efficiency, elevating the overall passenger experience, and cultivating a reputation for dependable service within a fiercely competitive airline market.

Furthermore, to enhance major services that gain higher levels of customer satisfaction—such as seat comfort, in-flight services, online support, and ease of online booking—airlines can implement or improve these strategies. These strategies encompass expanding seating options to cater to varying passenger preferences, expanding in-flight entertainment offerings, ensuring the proficiency and affability of the cabin crew, provisioning additional amenities to enhance passenger comfort, and establishing round-the-clock online customer service support channels.

Moreover, the ease of the online booking process can be enhanced through the development of user-friendly websites and mobile applications, transparent pricing structures, and the implementation of systems that retain passenger preferences for future bookings. Encouraging passenger feedback and perpetually refining staff training protocols also constitute integral components of the endeavour to augment service quality.

This dissertation has helped us better understand what makes airline passengers happy or unhappy. We used data to figure this out. The important insights we gained can be like a guide for airlines to make their customers cheerful and keep them coming back. In a tough competition, knowing how to make customers delighted is super important for both big airlines and new ones as they try to do well in the industry.

Reference list

An, M. and Noh, Y. (2009). Airline customer satisfaction and loyalty: impact of in-flight service quality. *Service Business*, [online] 3(3), pp.293–307. doi:https://doi.org/10.1007/s11628-009-0068-4.

Anand, A. and Bansal, G. (2016). Predicting Customer's Satisfaction (Dissatisfaction) Using Logistic Regression. *International Journal of Mathematical, Engineering and Management Sciences*, 1(2), pp.77–88. doi:https://doi.org/10.33889/ijmems.2016.1.2-009.

Baker, D.M.A. (2013). Service Quality and Customer Satisfaction in the Airline Industry: A Comparison between Legacy Airlines and Low-Cost Airlines. *American Journal of Tourism Research*, [online] 2(1), pp.67–77. doi:https://doi.org/10.11634/216837861403317.

Bellizzi, M.G., Eboli, L. and Mazzulla, G. (2020). An online survey for the quality assessment of airlines' services. *Research in Transportation Business & Management*, [online] 37(100515), p.100515. doi:https://doi.org/10.1016/j.rtbm.2020.100515.

Berrar, D. (2018). Bayes' Theorem and Naive Bayes Classifier Call for Papers for Machine Learning journal: Machine Learning for Soccer View project Bayes' Theorem and Naive Bayes Classifier. *Data Science Laboratory, Tokyo Institute of Technology*. doi:https://doi.org/10.1016/B978-0-12-809633-8.20473-1.

Bougie, R., Pieters, R. and Zeelenberg, M. (2003). Angry Customers don't Come Back, They Get Back: The Experience and Behavioral Implications of Anger and Dissatisfaction in Services. *Journal of the Academy of Marketing Science*, 31(4), pp.377–393. doi:https://doi.org/10.1177/0092070303254412.

Boyerinas, B. (2016). *Determining the Statistical Power of the Kolmogorov-Smirnov and Anderson-Darling Goodness-of-Fit Tests via Monte Carlo Simulation*. [online] Available at: https://www.cna.org/archive/CNA_Files/pdf/dop-2016-u-014638-final.pdf [Accessed 31 Aug. 2023].

Chang, W.-L., Liu, H.-T., Wen, Y.-S. and Lin, T.-A. (2008). Building an integrated model of future complaint intentions: The case of Taoyuan International Airport. *Journal of Air Transport Management*, 14(2), pp.70–74. doi:https://doi.org/10.1016/j.jairtraman.2007.11.004.

Conrad, E., Misenar, S. and Feldman, J. (2016). *Chapter 4 - Domain 3: Security Engineering (Engineering and Management of Security)*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780128024379000047 [Accessed 31 Aug. 2023].

de Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6), pp.448–455. doi:https://doi.org/10.1002/wics.1278.

E R, S. (2021). *Random Forest | Introduction to Random Forest Algorithm*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

Færgestad, E.M., Langsrud, Ø., Høy, M., Hollung, K., Sæbø, S., Liland, K.H., Kohler, A., Gidskehaug, L., Almergren, J., Anderssen, E. and Martens, H. (2009). *4.08 - Analysis of Megavariate Data in Functional Genomics*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780444527011000119.

Fazidah, N. and Mohamed, H. (2015). Digital Records Managemnt for Public Organizations View project Model for sustainable use of e-learning View project. *Article in International Journal of Business Information Systems*. doi:https://doi.org/10.1504/IJBIS.2015.072249.

Franklin, A. (2023). *3 steps to achieving customer satisfaction and loyalty - Zendesk*. [online] Zendesk. Available at: https://www.zendesk.com/blog/3-steps-achieving-customer-satisfaction-loyalty/.

Fuchs, J. (2022). *How to Handle and Remedy Customer Dissatisfaction*. [online] blog.hubspot.com. Available at: https://blog.hubspot.com/service/customer-dissatisfaction.

Gardial, S.F., Clemons, D.S., Woodruff, R.B., Schumann, D.W. and Burns, M.J. (1994). Comparing Consumers' Recall of Prepurchase and Postpurchase Product Evaluation Experiences. *Journal of Consumer Research*, [online] 20(4), pp.548–560. Available at: https://www.jstor.org/stable/2489758 [Accessed 9 Mar. 2022].

Hadley , W., Volume, V. and Issue, I. (2021). *Journal of Statistical Software Tidy Data*. [online] Available at: https://vita.had.co.nz/papers/tidy-data.pdf.

Han, S., Ham, S. (Sunny), Yang, I. and Baek, S. (2012). Passengers' perceptions of airline lounges: Importance of attributes that determine usage and service quality measurement. *Tourism Management*, [online] 33(5), pp.1103–1111. doi:https://doi.org/10.1016/j.tourman.2011.11.023.

Istijanto (2021). Impacts of the COVID-19 pandemic on airline passengers' recovery satisfaction: An experimental study. *Transportation Research Interdisciplinary Perspectives*, 12(PMC8669095), p.100487. doi:https://doi.org/10.1016/j.trip.2021.100487.

Josephat, P. and Ismail, A. (2012). A Logistic Regression Model of Customer Satisfaction of Airline. *International Journal of Human Resource Studies*, 2(4), p.255. doi:https://doi.org/10.5296/ijhrs.v2i4.2868.

Kaggle (2022). *Kaggle: Your Home for Data Science*. [online] Kaggle.com. Available at: https://www.kaggle.com/.

Khosrow-Pour, D.B.A. (2021). *Encyclopedia of Information Science and Technology, Fourth Edition (10 Volumes)*. [online] *www.igi-global.com*. IGI Global. Available at: https://www.igi-global.com/book/encyclopedia-information-science-technology-fourth/173015.

Kumar, A. (2020). *KNN Algorithm: What?When?Why?How?* [online] Medium. Available at: https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f.

Kumar, N. (2019). *Naive Bayes Classifiers - GeeksforGeeks*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/naive-bayes-classifiers/.

Lestari, Y.D. and Murjito, E.A. (2020). *View of Factor Determinants of Customer Satisfaction with Airline Services Using Big Data Approaches*. [online] journal.unj.ac.id. Available at: https://journal.unj.ac.id/unj/index.php/jpeb/article/view/14032/8290.

Matusitz, J. and Breen, G.-M. (2009). Consumer Dissatisfaction, Complaints, and the Involvement of Human Resource Personnel in the Hospitality and Tourism Industry. *Journal of Human Resources in Hospitality & Tourism*, 8(2), pp.234–246. doi:https://doi.org/10.1080/15332840802269866.

Noviantoro, T. and Huang, J.-P. (2021). Investigating airline passenger satisfaction: Data mining method. *Research in Transportation Business & Management*, 43, p.100726. doi:https://doi.org/10.1016/j.rtbm.2021.100726.

Park, E., Jang, Y., Kim, J., Jeong, N.J., Bae, K. and del Pobil, A.P. (2019). Determinants of customer satisfaction with airline services: An analysis of customer feedback big data. *Journal of Retailing and Consumer Services*, 51(186-190), pp.186–190. doi:https://doi.org/10.1016/j.jretconser.2019.06.009.

Park, S., Lee, J.-S. and Nicolau, J.L. (2020). Understanding the dynamics of the quality of airline service attributes: Satisfiers and dissatisfiers. *Tourism Management*, 81, p.104163. doi:https://doi.org/10.1016/j.tourman.2020.104163.

Ridwan, M. (2022). *Predicting & Optimizing Airlines Customer Satisfaction Using Predicting & Optimizing Airlines Customer Satisfaction Using Classification Classification*. [online] Available at: https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12513&context=theses.

Rust, R.T. and Oliver, R.L. (1993). *Service Quality: New Directions in Theory and Practice*. [online] *Google Books*. SAGE Publications. Available at: https://books.google.co.uk/books?hl=en&lr=&id=c3woDAAAQBAJ&oi=fnd&pg=PT7&ots=zq39wcwUJB&sig=EtQxhy9fikPw_np0JlgI8o7pX6o&redir_esc=y [Accessed 26 Aug. 2023].

Samunderu, E. and Farrugia, M. (2022). Predicting customer purpose of travel in a low-cost travel environment—A Machine Learning Approach. *Machine Learning with Applications*, 9(100379), p.100379. doi:https://doi.org/10.1016/j.mlwa.2022.100379.

Sugara, R.A. and Purwitasari, D. (2022). *Flight Delay Prediction for Mitigation of Airport Commercial Revenue Losses Using Machine Learning on Imbalanced Dataset*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CENIM56801.2022.10037369.

Suhartanto, D. and Ariani Noor, A. (2012). *CUSTOMER SATISFACTION IN THE AIRLINE INDUSTRY: THE ROLE OF SERVICE QUALITY AND PRICE Analyzing the Complex and Dynamic Nature of Brand Loyalty in the Hotel Industry View project Increase the quality of travelling through sustainable holiday program View project Dwi Suhartanto Politeknik Negeri Bandung*.

Suki, N.M. (2014). Passenger satisfaction with airline service quality in Malaysia: A structural equation modeling approach. *Research in Transportation Business & Management*, [online] 10, p.26. Available at: https://www.academia.edu/95710225/Passenger_satisfaction_with_airline_service_quality_in_Malaysia_A_structural_equation_modeling_approach [Accessed 11 Aug. 2023].

Tam, J.L.M. (2004). Customer Satisfaction, Service Quality and Perceived Value: An Integrative Model. *Journal of Marketing Management*, 20(7-8), pp.897–917. doi:https://doi.org/10.1362/0267257041838719.

Tukey, J.W. (1970). *EXPLORATORY DATA ANALYSIS*. [online] Available at: http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf.

Uslu, T. and Sarper Karakadilar, I. (2022). *The Impact of CustomerDissatisfaction RegardingRevenue Managementon Perceptions of AirlineExperience and Loyalty*. [online] Researchgate. Available at: https://www.researchgate.net/publication/361431056_The_Impact_of_Customer_Dissatisfaction_Regarding_Revenue_Management_on_Perceptions_of_Airline_Experience_and_Loyalty#fullTextFileContent.

Voss, D.S. (2005). *Multicollinearity*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/abs/pii/B012369398500428X [Accessed 31 Aug. 2023].

Vu, T. (2021). Service Quality And Its Impact On Customer Satisfaction. *Researchgate*. doi:https://doi.org/10.6084/m9.figshare.17089454.

Wang, T. and Chaipoopirutana, S. (2015). *A Study of the Factors Influencing Customer Loyalty: A Case Study of Thai Airways*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3041792.

Wang'ondu, R.W. (2009). *Factors affecting customer satisfaction in airline industry*. [online] erepository.uonbi.ac.ke. Available at: http://erepository.uonbi.ac.ke/handle/11295/13169.

Yazdi, M.F., Kamel, S.R., Chabok, S.J.M. and Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, 7(1). doi:https://doi.org/10.1186/s40537-020-00380-z.

Zach (2020). *Introduction to Logistic Regression*. [online] Statology. Available at: https://www.statology.org/logistic-regression/.

Zaharias, B. (2016). *ANALYSING CUSTOMER SATISFACTION IN THE AIRLINE INDUSTRY*.

Zahraee, S.M., Shiwakoti, N., Jiang, H., Qi, Z., He, Y., Guo, T. and Li, Y. (2022). A study on airlines' responses and customer satisfaction during the COVID-19 pandemic. *International Journal of Transportation Science and Technology*. doi:https://doi.org/10.1016/j.ijtst.2022.11.004.

https://docs.google.com/document/d/1Ju5hvbqAHmitxCH64HilrNhUGxZAkengFMHNaun8Ozg/edit?usp=sharing- (Analysis code)

Fritz, M. and Berger, P.D. (2015). Chapter 11 - Will anybody buy? Logistic regression. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780128006351000112.