

WQD7001 Principles of Data Science

Group 05

Group Project Part 1: Quantifying the Air Travel Experience: A Comprehensive Analysis of Influential Factors in Passenger Satisfaction

|  |  |
| --- | --- |
| Name | Matric Number |
| MUHAMMAD ZAHRIEL BIN ISMAIL | 22085509 |
| HAN XIANG | 22093085 |
| IZZUL ILHAM BIN YUSOF | 22107573 |
| TASSLIM BIN MANSOOR ALI | 23056322 |
| ZHANG LEPING | 22098083 |

Table of Contents

[1.0 Team Member & Roles 3](#_Toc152551470)

[2.0 Abstract 4](#_Toc152551471)

[3.0 Introduction 5](#_Toc152551472)

[3.1 Problem Statement 5](#_Toc152551473)

[3.2 Project Objective 6](#_Toc152551474)

[3.3 Scope & Domain 6](#_Toc152551475)

[4.0 Literature Review 7](#_Toc152551476)

[5.0 Data Science Pipeline 8](#_Toc152551477)

[5.1 Data Structure 8](#_Toc152551478)

[5.2 Data Quality Assessment 10](#_Toc152551479)

[5.2.1 Missing Data 10](#_Toc152551480)

[5.2.2 Outliers 11](#_Toc152551481)

[5.2.3 Grouping of Data 13](#_Toc152551482)

[5.3 Exploratory Data Analysis 14](#_Toc152551483)

[5.3.1 Distribution of Data 14](#_Toc152551484)

[5.3.2 Correlation Analysis 19](#_Toc152551485)

[5.3.3 Principal Component Analysis 20](#_Toc152551486)

[6.0 Ethical Considerations 21](#_Toc152551487)

[7.0 Project Contribution 22](#_Toc152551488)

[8.0 Conclusion 23](#_Toc152551489)

[9.0 References 24](#_Toc152551490)

# 1.0 Team Member & Roles

Team members of Group 07 and their respective roles are listed in the table below:

|  |  |
| --- | --- |
| Name | Role |
| MUHAMMAD ZAHRIEL BIN ISMAIL | Leader |
| HAN XIANG | Maker |
| IZZUL ILHAM BIN YUSOF | Detective |
| TASSLIM BIN MANSOOR ALI | Oracle |
| ZHANG LEPING | Presenter |

Role Responsibilities:

1. Leader: Guiding the project to ensure it stays on track as well as keeping track of the project progress.
2. Maker: Designing and writing code to conduct analysis on the data, more focused on the technical aspects of the project
3. Detective: Analyses the data to understand and answer problems. Proposing solutions based on further investigating the issue or data.
4. Oracle: Project Subject Matter Expert (SME) provides guidance for regarding background of the data.
5. Presenter: Communicates project progress and status to key stakeholders, creates documentation and reports.

# 2.0 Abstract

As the world economy and aviation sector grow, a growing number of individuals are opting to travel by air. The purpose of this study is to investigate the factors that influence passenger satisfaction in the aviation business to meet passengers' growing demands for aviation services. We examined the "Airline Passenger Satisfaction" data set to increase passenger satisfaction during the flight. We then filled the gaps in the data using the Multiple Imputation by Chained Equations (MICE) method and employed Exploratory Data Analysis (EDA) to identify and resolve the main issues influencing the passenger experience. Our main goal is to analyse which factors influence air traveller satisfaction and to uncover the relative weights of these factors and their impact on the overall experience. This research will provide airlines with trusted insights to help them optimize the passenger experience, improve passenger satisfaction and loyalty, and thereby gain more profits.

# 3.0 Introduction

The aviation industry is a large industry encompassing and assisting in several different economic sectors (Transportation, tourism, etc.) It is one of the largest driving factors in tourism which contributes to approximately 7.6% of the global GDP in 2022 (WTTC, 2023) and has been steadily growing over the years. In an era where travellers are growing more discerning and have higher expectations of travelling services, understanding the key factors that can contribute to a traveller’s satisfaction has become essential for all stakeholders in the aviation industry.

This study aims to analyse and understand the key factors in providing passenger satisfaction within the aviation industry by reviewing a dataset obtained from a Passenger Satisfaction Survey. The survey contains thousands of rows of data in which the passengers rate the different aspects of their flight experience and capturing these sentiments within the dataset. By analysing this dataset, we seek to gain a valuable insight into understanding passenger expectations and identifying which factors need to be acted upon for airlines to enhance their customer’s experience.

## 3.1 Problem Statement

The aviation industry worldwide is extremely competitive, with passenger satisfaction playing a vital role in determining the airline performance. The airline industry recognizes the importance of passenger satisfaction in customer retention as well as maintaining a competitive advantage over other companies in the same industry. Understanding passengers’ expectations and the factors that influences their satisfaction is important for airline companies aiming to be the best in the industry. Airlines can make informed decisions to address challenges, improve services and ultimately boost customer happiness by exploring various aspects of the passenger experience.

This study aims to investigate the passengers experience from their pre-flight, in-flight, and post-flight experiences and to identify key elements that impact their overall satisfaction. With thousands of passenger’s responses recording their views of various aspects of their flying experience, the goal is to extract valuable insights that illustrate passengers’ expectation.

By conducting this study, we aim to answer two primary questions:

1) What are the key contributing factors that impact the passengers overall experience while flying?

2) What is the impact of flight delays and arrival delays on overall passenger’s experience?

## 3.2 Project Objective

This Section focuses on outlining the objectives of the study, below are two of the primary project objectives:

1) To identify and analyse the key contributing factors that impact the overall customer experience in the context of airline travel. This objective includes:

- Analysing the collected data to determine the relative importance and impact of various factors on overall customer satisfaction.

- Identifying the weightage of each factor as compared to one another in deciding the overall customer experience

2) To investigate the specific impacts of departure and arrival delays on overall customer experience. This objective includes:

- Examining the correlation between the duration and frequency of delays and customer satisfaction.

- Identifying which delay has a higher effect on customer satisfaction

## 3.3 Scope & Domain

The scope of the project will encompass the several aspects of that a customer may experience during their flight such as check-in services, online boarding, on-board service, cleanliness, food & beverages and several other pre to post flight services.

The research domain will be limited to a customer airline satisfaction (N > 120,000) and focuses on the data provided from customer feedback as some customer data such as Age and Gender.

# 4.0 Literature Review

Airline passenger satisfaction is a very important variable that is used in many previous studies because it is linked to the attitude of the passenger regarding service expectations and perceptions towards an airline service. In the airline sector, passenger satisfaction is a complicated customer knowledge and experience that can be defined as an evaluation that passengers have faced (Tahanisaz & Shokuhyar, 2020)

The paper authored by Xuchu Jiang, Ying Zhang, Ying Li and Biao Zhang presents a study on the factors affecting airline passenger satisfaction using a RF-RFE-LR model to extract important variables affecting passenger satisfaction. According to the study, to increase service quality and determine which service tactics work best for attracting new customers, airlines should implement a customer-oriented service evaluation approach. The research also emphasizes how the COVID-19 pandemic affected the aviation sector and how crucial providing high-quality services is to obtain a competitive edge. The random forest after RF-RFE feature selection performs the best in the classification prediction model (Jiang et al., 2022).

Another study by Sachin Kumar and Mikhail Zymbler discusses on analysing customer satisfaction from airline tweets using machine learning. The paper explores numerous approaches to sentiment analysis, including document-level, phrase-level, and entity-level sentiment analysis, it also explains how to extract features from tweets using word embedding with the Glove dictionary approach and the n-gram approach, and then use those features to build a classification model that divides tweets into positive and negative categories. The study concludes that social media sentiment analysis of customers can assist airlines in identifying problems that impact customer satisfaction (Kumar & Zymbler, 2019).

The study authored by Tri Noviantoro and Jen-Peng Huang focuses on the impact on the full-service airline business model from the airline passenger satisfaction data. The research highlights on the significance of providing excellent services to travellers to maintain a competitive edge within the airline industry. According to their findings, deep learning is the best machine learning model to predict airline passenger satisfaction. The study suggests that airlines must focus on enhancing their services to suit the passenger’s expectations in order to ensure their sustainability in the business (Noviantoro & Huang, 2022).

# 5.0 Data Science Pipeline

This section focuses on the study of the basic structure of the dataset and initial EDA steps taken by the team.

## 5.1 Data Structure

Understanding the dataset is an important step in providing a deeper insight when analysing the dataset, the following section focuses on the basic structure of the Airline Passenger Satisfaction dataset. The basic information of the dataset that is being analysed is provided in the table below:

Table 1.0: Dataset Information

|  |  |
| --- | --- |
| Data Structure | Property |
| Data Source | <https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction> |
| Data Name | Airline Passenger Satisfaction |
| File Type | .csv |
| File Size | 12.5 MB |
| Year | 2022 |
| Data Dimensions | 24 Columns, 129881 Rows |

|  |  |  |
| --- | --- | --- |
| Attribute | Data Type | Description |
| ID | Integer | Unique Integer to identify unique entries |
| Gender | String (Categorical) | Customer Gender |
| Age | Integer (Continuous) | Customer Age |
| Customer Type | String (Categorical) | Identifies if the Customer is a first-time or returning customer |
| Type of Travel | String (Categorical) | Identifies if the customer’s flight is personal or for business |
| Class | String (Categorical) | Flight Class (Business, Economy, Economy plus) |
| Flight Distance | Integer (Continuous) | Distance of the customers flight |
| Departure Delay | Integer (Continuous) | Delay from actual departure time from point of origin |
| Arrival Delay | Integer (Continuous) | Delay from actual arrival time to the destination |
| Departure and Arrival Time Convenience | Integer (Ordinal) | Rating from 1-5 on Convenience of Dept/Arrival |
| Ease of Online Booking | Integer (Ordinal) | Rating from 1-5 on online booking service |
| Check-in Service | Integer (Ordinal) | Rating from 1-5 on check-in service |
| Online Boarding | Integer (Ordinal) | Rating from 1-5 on online boarding service |
| Gate Location | Integer (Ordinal) | Rating from 1-5 on gate location convenience |
| On-board Service | Integer (Ordinal) | Rating from 1-5 on on-board service |
| Seat Comfort | Integer (Ordinal) | Rating from 1-5 on seat comfort |
| Leg Room Service | Integer (Ordinal) | Rating from 1-5 on leg room service |
| Cleanliness | Integer (Ordinal) | Rating from 1-5 on overall cleanliness |
| Food and Drink | Integer (Ordinal) | Rating from 1-5 on food and drinks provided |
| In-flight service | Integer (Ordinal) | Rating from 1-5 on In-flight service |
| In-flight Wi-Fi Service | Integer (Ordinal) | Rating from 1-5 on In-flight Wi-Fi Service |
| In-flight entertainment | Integer (Ordinal) | Rating from 1-5 on In-flight entertainment |
| Baggage Handling | Integer (Ordinal) | Rating from 1-5 on Baggage Handling |
| Satisfaction | String (Categorical) | Overall customer satisfaction Rated between Satisfied and Neutral or Dissatisfied |

## 5.2 Data Quality Assessment

This section focuses on analysing the data for any outliers and missing data in the several columns. The Data Quality Assessment will be done primarily using tools such as RStudio and Python.

### 5.2.1 Missing Data

Identifying if any of the rows have missing data is important in reducing potential noise in the results of the study. Running a script as per the figure below in RStudio presents us with any missing data by column.

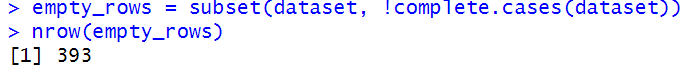
Figure 1.0: Summary of Missing Data

A computer screen shot of a computer screen

Description automatically generated

The result of the code identifies only one column that has missing data, “Arrival.Delay”. Further analysis into the column identifies 393 instances of missing data as per the following figure.

Figure 1.1: Number of Missing Rows



Despite the missing data making a very small percentage of the data (N < 0.302%), the team has decided to impute the data given that the 393 rows are only missing data in one column. The imputation method used is Multiple Imputation by Chained Equations (MICE), which is a widely used imputation method used in statistics and data analysis.

### 5.2.2 Outliers

Further investigation revealed that their exist outliers in two different columns (Departure.Delay and Arrival.Delay). Below are boxplots of the two columns:

Figure 2.0 Departure Delay Boxplot

A graph of a pointy point

Description automatically generated with medium confidence

Figure 2.1 Arrival Delay Boxplot

A graph of a graph

Description automatically generated with medium confidence

Added to the boxplots above, the following is a table of the characteristics of the mentioned columns.

|  |  |  |
| --- | --- | --- |
| Characteristics | Departure Delay | Arrival Delay |
| Number of Entries | 129880 | 129880 |
| Mean | 14.71 | 15.16 |
| Median | 0 | 0 |
| Mode | 0 | 0 |

### 5.2.3 Grouping of Data

Data Grouping/Binning is conducted to ensure that the data is easier to process or present in later stages, the primary columns that have been identified for binning is “Age” and “Flight Distance”. The proposed binning for the two said columns is to bin “Age” by Age Groups and “Flight Distance” by Flight Distance Categories. Below is a table on the binning criteria of “Age” and “Flight Distance”.

|  |  |
| --- | --- |
| Age | Age Group |
| 0-12 | Child |
| 13-17 | Adolescents |
| 18-65 | Adult |
| 66+ | Elderly |

|  |  |
| --- | --- |
| Flight Distance (Miles) | Flight Distance Category |
| 0 – 999 | Short Distance |
| 1000 – 3000 | Medium Distance |
| 3001+ | Long Distance |

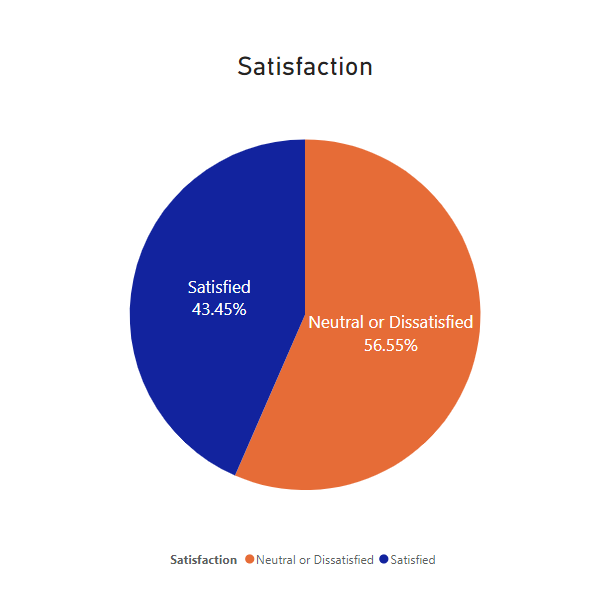
## 5.3 Exploratory Data Analysis

This section focuses primarily on understanding the data by employing variety of techniques to uncover underlying patterns and extracts meaningful insights from the data before further investing into advanced modelling. It is important to perform EDA to guide data scientist to generate the right questions and utilise the data into answering them to meet the business objectives.

### 5.3.1 Distribution of Data

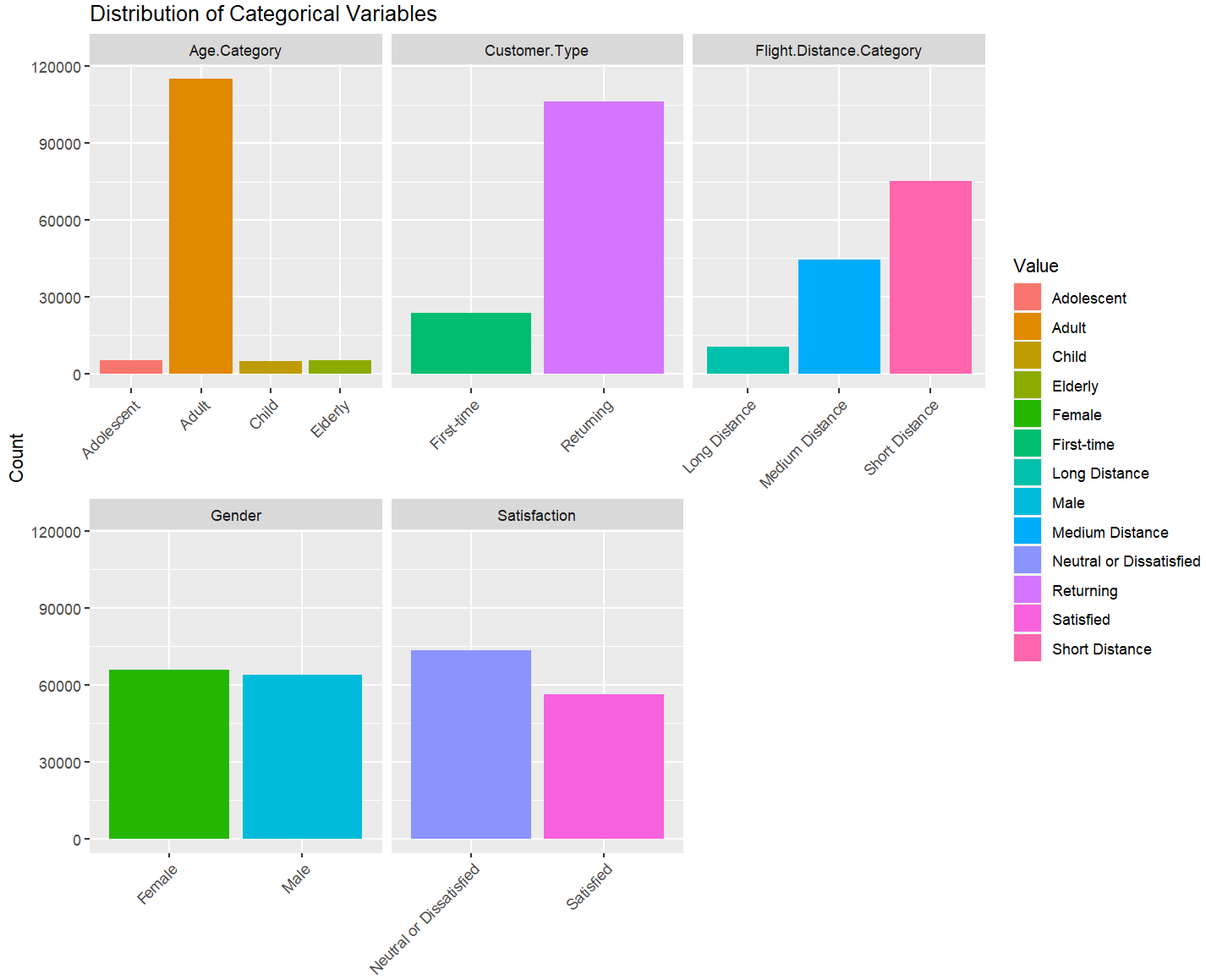
Firstly, we perform some simple distribution analysis on the responses by doing a simple demographic background analysis. Firstly, there is a higher number of “Neutral or Dissatisfied” entries as compared to “Satisfied” entries (6.55% Difference).

Figure 3.0 Distribution of Satisfaction



Furthermore, analysis of the distribution of each demographical value was conducted and shown below figure:

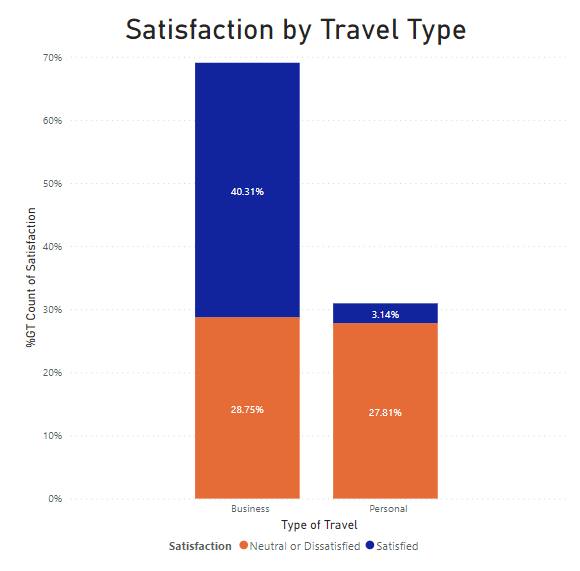
Figure 3.1 Distribution of Demographics



The above figure shows that most of the data consists of “Adults”, “Returning” Customers and “Short Distance” Flights. The distribution between Gender and Satisfaction is almost evenly distributed.

Further investigation into the satisfaction of customers by their demographic data to show a surface level understanding of the relationship between the demographic data is important to outline before starting further analysis into the data. As per the figure below, a plot showing the relationship between Satisfaction and Travel Types is shown.

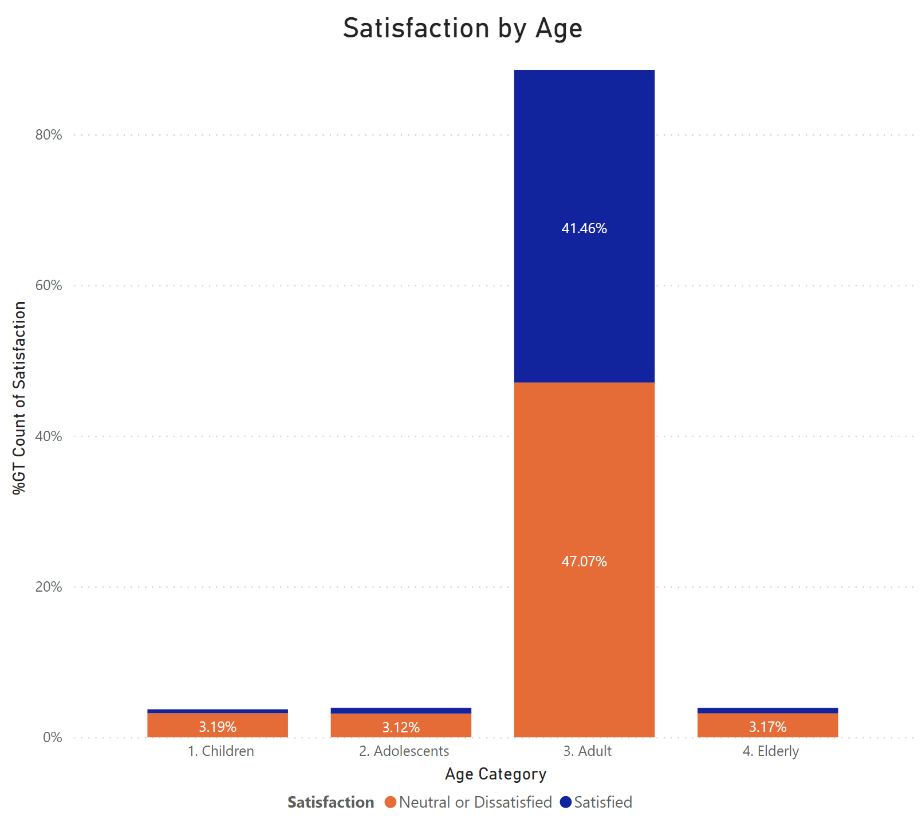
Figure 3.2 Satisfaction by Travel Type



Based on the figure Above, we can deduce that many of the satisfied customers consists of customers travelling for business purposes and a small minority (~3.14%) consist of Personal type travellers

Alongside the Travel Type, a plot to discover any relations between ages is conducted and presented in the figure below.

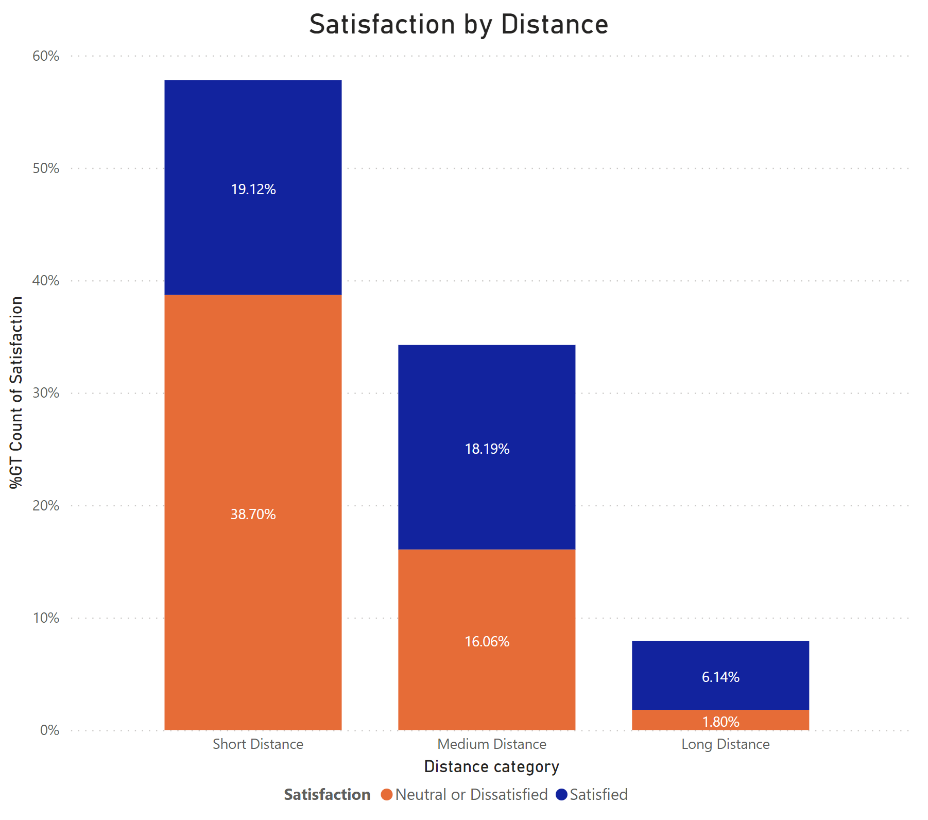
Figure 3.2 Satisfaction by Age



Based on the figure we can see that the highest contribution to “Satisfied” and “Neutral or Unsatisfied” customers is Adults, this is because “Adults” are the largest contributing age group for the dataset. We also can see that across all age groups, neutral and dissatisfied customers are the predominant side. Even adults which are the huge contributors of customers headcount are more prone to negative view on their experiences, a slight change in their tendencies would impact significantly to airline profitability.

Furthermore, a relationship between “Satisfaction” and “Distance Categories” is conducted and represented in the table below.

Figure 3.3 Satisfaction by Distance

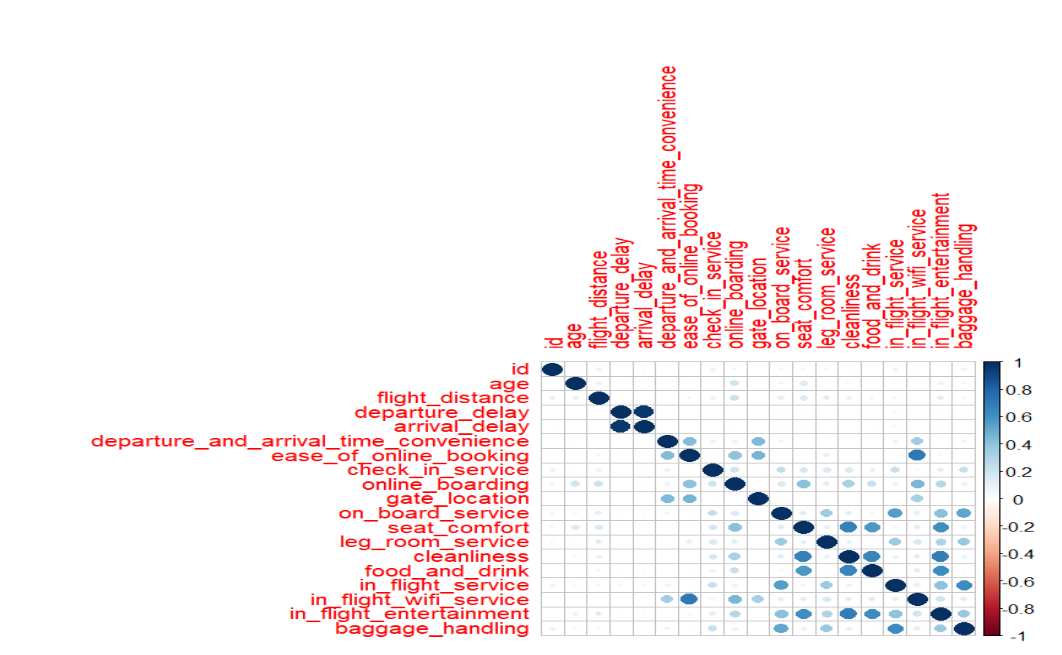


Based on the chart, it was shown that short distance flight customers are more likely to show neutral or dissatisfaction toward their flight experiences. Comparing to medium and long-distance flight, the distribution of customers is more to positive view towards their experience. Although this is not confirmed but likely the efforts on customer service have been focusing on long distance flight but need to be taken note that more than 50% of the customers are coming from short distance flight.

### 5.3.2 Correlation Analysis

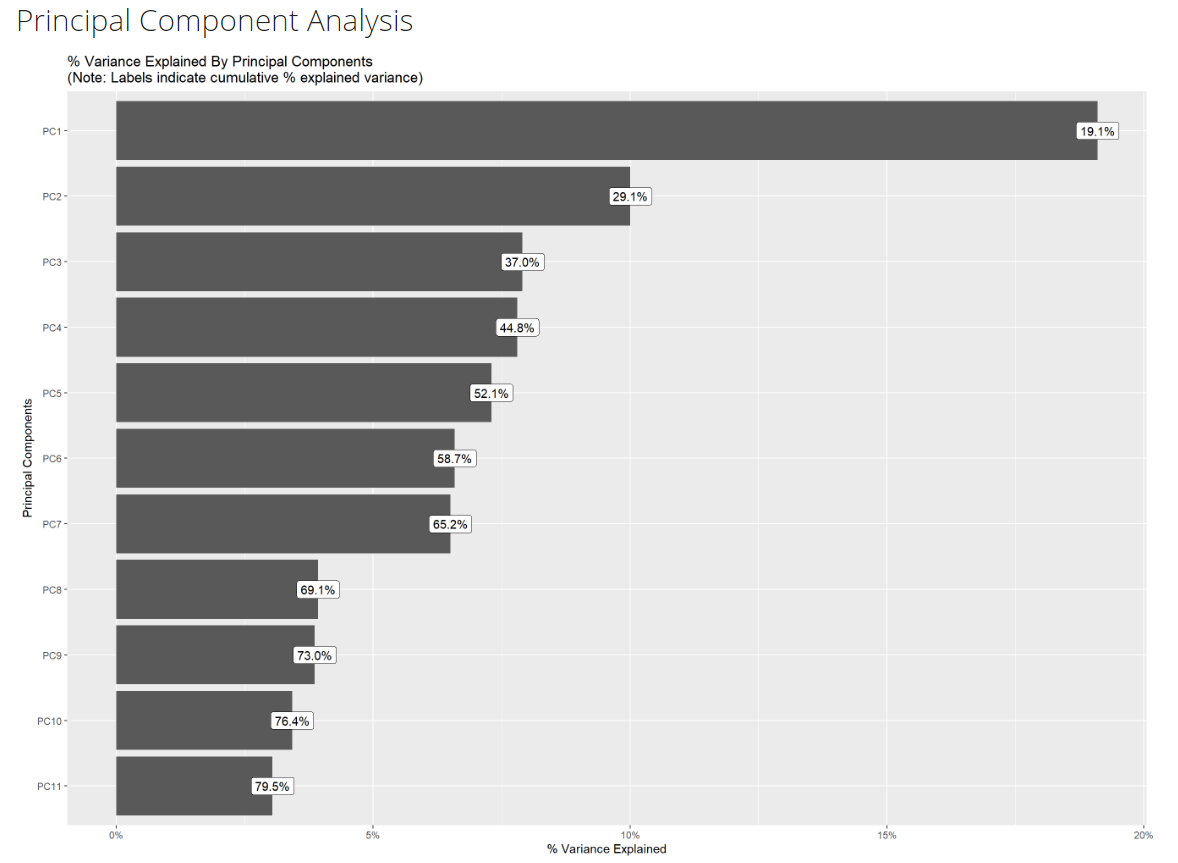
Secondly, going beyond simple distribution, we investigate the correlation relationship between each numerical response type. Some interesting findings is that cleanliness, seat comfort, food & drink, inflight entertainment are highly correlated pairs. These factors correlations may not explain causality; however, these may show that these items go hand in hand in determining customer satisfaction and should be considered when performing advanced modelling. Another interesting view to look at is that all factors' correlations are neutral or positively correlated. This means that availability of certain aspects of the factors would only contribute to the customer satisfaction rather than the other way around.

Figure 4.0 Correlation Analysis



### 5.3.3 Principal Component Analysis

Thirdly, before we can fit the features into developing a model, we need to know how much the features can be used to explain the output. If there is low level of component importance, we need to enhance the dataset with more data or perform features engineering to improve the performance. One such method is to perform Principal Component Analysis (PCA). In PCA, we develop principal components, which are new features developed by combining multiple uncorrelated original features in the dataset. From there, we can use Explained Variance Ratio, which shows how the usefulness of the components in explaining the output required. From our dataset, we found out that our dataset can be explained in terms of 11 principal components, which attributable to around 80%. What this means is that the using the 11 principal components would suffice to explain around 80% of the output. Generally, it is recommended to use the ratio up to 80% as we would not want the model to incorporate too much noise and cause overfitting which then make the model doesn’t perform well in real world data.



In summary, the factors in the dataset are important and does provide insight of the responses whether the customer is satisfied. Thus, we can rely on the dataset to perform prediction or any other purpose in relation to customer satisfaction in airline. However, choosing the right algorithm also plays an important role in the performance.

# 6.0 Ethical Considerations

When conducting customer satisfaction surveys, airlines should take the following ethical considerations into account:

1) Obtain the consent of the participants and ensure that the data processing and analysis of the project are legal and compliant.

2)To protect the identity of respondents, the only private information collected in this dataset is age and gender.

3) Treat participants equally and ensure there is no bias in analysis and reporting.

4) Social responsibility is positive, and the goal of the analysis is to promote airline services while improving passengers’ aviation experience.

5) Adhere to the principles of honesty and objectivity during the investigation process to ensure that the results of data analysis are accurate and will not exaggerate or dilute the actual situation.

# 7.0 Project Contribution

Airlines can optimize flight services by using airline satisfaction surveys, which not only assess passengers' contentment with flight services but also identify crucial aspects that impact satisfaction. The specific contributions are as follows:

1) Established a model that can be used to analyse aviation satisfaction

2) Make improvement suggestions to airlines based on the analysis results to improve aviation service satisfaction

3) Can identify the needs and preferences of specific groups of people, provide personalized service recommendations, and meet the expectations of different groups.

4) Reduce complaints and disputes and help create a more harmonious travel environment.

5) Promote technological innovation and allow the aviation industry to use more advanced technologies to improve efficiency and service quality.

6) Make people prefer aviation as a mode of travel and improve their comfort during the journey

# 8.0 Conclusion

This study aims to explore the data in depth, discovering that 11 principal components are sufficient to explain 80% of the output through principal component analysis (PCA) before developing the model, and mining the data through exploratory data analysis (EDA) to gain meaningful insights.

According to research, in terms of business travel and personal travel, personal travel has higher requirements for flight services, and the satisfaction rate only accounts for 3%. Across all age groups, the customers tend to be neutral or dissatisfied, while adults are slightly more tolerant of imperfections in airline service with near balance proportion, the proportion in other age groups are predominant in neutral and dissatisfaction. Medium and Long-distance passengers tend to be more satisfied than short-distance passengers. In addition, factors such as aircraft cleanliness, seat comfort, and food were also found to affect customer satisfaction. The analysis results show that in-flight service is very important because these service details will affect overall passenger satisfaction and thus reduce passenger loyalty. Airlines should take the lead in focusing on and optimizing the factors mentioned above to improve services and competitiveness, thereby attracting more passengers.

# 

# 

# 

# 

# 

# 

# 

# 

# 9.0 References

WTTC. (2023). *Economic Impact Research.* World Travel & Tourism Council.

Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022). Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. *Scientific Reports*, *12*(1), 11174. <https://doi.org/10.1038/s41598-022-14566-3>

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets [Article]. *Journal of Big Data*, *6*(1), Article ARTN 62. <https://doi.org/10.1186/s40537-019-0224-1>

Noviantoro, T., & Huang, J. (2022). Investigating airline passenger satisfaction: Data mining method [Article]. *Research in Transportation Business and Management*, *43*, Article ARTN 100726. <https://doi.org/10.1016/j.rtbm.2021.100726>

Tahanisaz, S., & Shokuhyar, S. (2020). Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry [Article]. *Journal of Air Transport Management*, *83*, Article ARTN 101764. <https://doi.org/10.1016/j.jairtraman.2020.101764>