Student ID: 24148501

Module Code: CMP7228

Module Title: Machine Learning

Coursework Title: Airline Passenger Satisfaction

Abstract

This project explores the application of machine learning techniques to analyze and predict airline passenger satisfaction, using a dataset sourced from Kaggle. The primary objectives include predicting passenger satisfaction levels, analyzing the factors contributing to delays, and clustering passengers based on their preferences and behaviors. A comprehensive exploratory data analysis (EDA) was conducted to clean, transform, and understand the dataset, uncovering key patterns and trends.

Supervised learning models such as XGBoost, LightGBM, and CatBoost were employed to classify passenger satisfaction, achieving robust performance through hyperparameter tuning and ensemble approaches. For delay prediction, regression models and custom ensembles were evaluated to enhance accuracy. Unsupervised clustering techniques, including K-Means and DBSCAN, were utilized to segment passengers into meaningful groups, enabling personalized service strategies.

The findings highlight the significant impact of features such as online boarding, travel class, and inflight entertainment on satisfaction, while factors like delays exhibited weaker correlations. The clustering analysis provided actionable insights into passenger personas, identifying distinct segments based on demographics and preferences. This study demonstrates the potential of machine learning in addressing real-world challenges within the airline industry, offering data-driven recommendations to improve passenger experience and operational efficiency.

Table of Contents

Abstract 2

Table of Contents 3

Table of Figures 6

Table of Tables 7

Introduction 8

Problem Domain 9

Dataset Description 10

Source 10

Features and Explanation 10

Data Types 11

Exploratory Data Analysis (EDA) 11

Data cleaning 11

Dropping Unnecessary Columns: 11

Checking Data Types: 11

Separating Categorical and Numerical Features: 12

Checking Missing Values: 12

Dropping Rows with Missing Values: 13

Checking for Duplicate Entries: 13

Class Distribution Analysis: 13

Data wrangling 13

1) Feature Transformation 13

Features 14

1) Gender 14

2) Customer Type 15

3) Cleanliness 16

4) Type of Travel 16

5) Class ... 17

6) Inflight service ... 18

7) Inflight wifi service ... 19

8) Departure/Arrival time convenient ... 20

9) Ease of Online booking ... 21

10) Gate location ... 22

11) Food and drink ... 23

12) Online boarding ... 24

13) Seat comfort ... 25

14) Inflight entertainment ... 26

15) On-board service ... 27

16) Leg room service ... 28

17) Baggage handling ... 29

18) Checkin service ... 30

19) satisfaction ... 31

20) Flight Distance ... 32

21) Age ... 35

22) Departure Delay in Minutes ... 37

23) log_DepartureDelayMinutes ... 39

24) Arrival Delay in Minutes ... 42

25) log_ArrivalDelayMinutes ... 45

Experiments ... 48

Problem 1: Classification of Passenger Satisfaction ... 48

Experiment 1: Baseline Models (XGBoost, LightGBM, CatBoost) ... 48

Experiment 2: Custom Ensemble Architecture ... 48

Experiment 3: Hyperparameter Tuning ... 48

Problem 2: Predicting Arrival Delay ... 49

Experiment 1: Baseline Models (XGBoost, LightGBM, CatBoost)................................49

Experiment 2: Custom Ensemble Architecture...49

Problem 3: Clustering Passengers Based on Preferences 49

Experiment 1: K-Means Clustering with PCA 49

Experiment 2: DBSCAN (Density-Based Clustering)...50

Results and Discussion 50

Problem 1: Classification of Passenger Satisfaction..50

Model Comparison Analysis..53

Summary 54

Problem 2: Predicting Arrival Delay 54

Model Comparison..57

Summary:..59

Problem 3: Clustering Passengers Based on Preferences 59

Principal Component Analysis (PCA):..60

DBSCAN Clustering:...60

Cluster Distribution: After outlier removal, DBSCAN identified nine meaningful clusters: 60

Conclusion 65

Future Work..66

Table of Figures

Figure 1 Bar plot of Gender 14

Figure 2Plot of Customer Type 15

Figure 3 Bar Plot of Cleanliness 16

Figure 4 Bar Plot of Type of Travel 17

Figure 5 Bar Plot of Class 18

Figure 6 Bar Plot of Inflight Service 19

Figure 7 Bar Plot of Inflight Wi-Fi Service 20

Figure 8 Bar Plot of Departure/Arrival time convenient 21

Figure 9 Bar Plot of Ease of Online booking 22

Figure 10 Bar Plot of Gate Location 23

Figure 11 Bar Plot of Food and Drink 24

Figure 12 Bar Plot of Online boarding 25

Figure 13 Bar Plot of Seat Comfort 26

Figure 14 Bar Plot of Inflight entertainment 27

Figure 15 Bar Plot of On-board Service 28

Figure 16 Bar Plot of Leg room service 29

Figure 17 Bar Plot of Baggage Handling 30

Figure 18Check-in Service 31

Figure 19 Bar Plot of Satisfaction 32

Figure 20 Histogram of Flight Distance 33

Figure 21Violin plot of Flight Distance 33

Figure 22 Boxplot of Flight Distance 34

Figure 23Histogram of Age 35

Figure 24 Violin Plot of Age 36

Figure 25Box Plot of Age 36

Figure 26 Histogram of Departure Delay Minutes 37

Figure 27Violin of Departure Delay in Minutes 38

Figure 28 Boxplot of of Departure Delay in Minutes 39

Figure 29 Histogram of log_DepartureDelayMinutes 40

Figure 30 Violin plot log_DepartureDelayMinutes 41

Figure 31 Boxplot of log_DepartureDelayMinutes 42

Figure 32 Histogram of Arrival Delay in Minutes 43

Figure 33 Boxplot of Arrival Delay in Minutes..43
Figure 34 Violin Plot of Arrival Delay in Minutes..44
Figure 35 Histogram of log_ArrivalDelayMinutes..45
Figure 36 Violinplot of log_ArrivalDelayMinutes..46
Figure 37 Boxplot of log_ArrivalDelayMinutes..47
Figure 38 Problem 1 train accuracy model comparison 50
Figure 39 Problem 1 validation accuracy model comparison..51
Figure 40 Problem 1 accuracy difference model comparison 51
Figure 41 Problem 1 precision model comparison 52
Figure 42 Problem 1 recall model comparison...52
Figure 43 Problem 1 f1 score model comparison...53
Figure 44 Problem 2 train mse model comparison 54
Figure 45 Problem 2 validation mse model comparison 55
Figure 46 Problem 2 train mse difference model comparison..55
Figure 47 Problem 2 train mae model comparison..55
Figure 48 Problem 2 validation mse model comparison 56
Figure 49 Problem 2 train mae difference model comparison...56
Figure 50 Problem 2 train r2 model comparison 56
Figure 51 Problem 2 validation r2 model comparison...57
Figure 52 Problem 2 r2 difference model comparison 57
Figure 53 Scree Plot..60
Figure 54 Cluster summary...61

Table of Tables

Table 1 EDA Feature Data Type 12

Introduction

Customer satisfaction is a critical factor in the airline industry, influencing customer loyalty, brand reputation, and overall profitability. As air travel becomes increasingly competitive, understanding the factors that drive passenger satisfaction is essential for airlines aiming to enhance their service quality and retain customers. In this study, we leverage data-driven insights to analyse customer satisfaction and predict satisfaction levels using machine learning techniques.

The dataset used in this project, sourced from Kaggle, provides detailed information about passengers, including demographic attributes, travel behaviours, and service ratings. By exploring this dataset, we aim to uncover meaningful patterns and relationships that can inform airline strategies for improving customer experiences.

This report is structured to address the following objectives:

1) Perform a comprehensive exploratory data analysis (EDA) to identify trends, correlations, and potential data issues.
2) Implement a combination of supervised and unsupervised machine learning techniques to classify and cluster passengers based on satisfaction levels and behavioural attributes.
3) Evaluate model performance using appropriate metrics to ensure reliability and accuracy.

The analysis focuses on using both traditional and advanced machine learning models, providing a robust approach to understanding customer satisfaction. Additionally, the report highlights key insights and practical recommendations for the airline industry, demonstrating the value of machine learning in addressing real-world business challenges.

Through this study, we aim to contribute actionable insights and a deeper understanding of the drivers of customer satisfaction, offering a roadmap for leveraging data science in the aviation sector.

Problem Domain

The airline industry is a highly competitive sector where customer satisfaction plays a crucial role in determining an airline's success. Understanding and addressing the factors influencing passenger satisfaction can lead to improved customer retention, better operational efficiency, and enhanced service quality. In this study, we explore three interconnected problems within the domain of airline passenger satisfaction:

1) Predicting Overall Passenger Satisfaction

Problem Statement: Passenger satisfaction is influenced by a multitude of factors, including flight experience, service quality, and travel convenience. Identifying the key drivers of satisfaction and accurately predicting passenger satisfaction levels can help airlines tailor their services to meet customer expectations.

Objective: To build a predictive model that determines whether a passenger is "Satisfied" or "Dissatisfied" based on various flight and service attributes.

2) Analyzing Departure Delays

Problem Statement: Departure delays not only disrupt airline schedules but also significantly impact passenger satisfaction. Understanding the factors contributing to delays and predicting their occurrence can improve operational efficiency and enhance customer experience.

Objective: To investigate the relationship between flight-related features (e.g., departure time, flight distance, class) and departure delays, and to predict the likelihood of a delay.

3) Classifying Passengers Based on Satisfaction

Problem Statement: Passengers with similar satisfaction levels often exhibit common travel behaviours and preferences. Segmenting passengers based on their satisfaction allows airlines to better understand distinct customer groups and personalize services effectively.

Objective: To apply clustering techniques to group passengers into distinct categories based on satisfaction and travel attributes, providing insights into their preferences and expectations.

By addressing these three problems, this study aims to provide actionable insights into passenger satisfaction and operational efficiency. The findings will help airlines develop targeted strategies to enhance customer experiences, optimize operations, and maintain a competitive edge in the market.

Dataset Description

Source

The dataset is sourced from **Kaggle**, titled "Airline Passenger Satisfaction." It provides a comprehensive overview of passenger demographics, flight details, service ratings, and overall satisfaction.

Features and Explanation

1) **ID** (Categorical): Unique identifier for each passenger.
2) **Gender** (Categorical): Passenger's gender (Male/Female).
3) **Age** (Numerical): Passenger's age in years.
4) **Customer Type** (Categorical): Loyalty status (Loyal/Disloyal customer).
5) **Type of Travel** (Categorical): Purpose of travel (Business/Personal).
6) **Class** (Categorical): Travel class (Economy, Economy Plus, Business).
7) **Flight Distance** (Numerical): Distance of the flight in miles.
8) **Inflight Wifi Service** (Numerical): Rating of Wi-Fi quality (1–5 scale).
9) **Departure/Arrival Time Convenience** (Numerical): Rating for schedule convenience.
10) **Ease of Online Booking** (Numerical): Rating of the booking process.
11) **Gate Location** (Numerical): Rating of gate proximity.
12) **Food and Drink** (Numerical): Rating of food and beverages.
13) **Online Boarding** (Numerical): Rating of online boarding process.
14) **Seat Comfort** (Numerical): Rating of seat comfort.
15) **Inflight Entertainment** (Numerical): Rating of in-flight entertainment options.
16) **Onboard Service** (Numerical): Rating of onboard service quality.
17) **Leg Room Service** (Numerical): Rating of legroom space.
18) **Baggage Handling** (Numerical): Rating of baggage handling quality.
19) **Check-in Service** (Numerical): Rating of the check-in process.
20) **Inflight Service** (Numerical): Rating of in-flight crew service.
21) **Cleanliness** (Numerical): Rating of cleanliness onboard.
22) **Departure Delay in Minutes** (Numerical): Number of minutes delayed at departure.
23) **Arrival Delay in Minutes** (Numerical): Number of minutes delayed at arrival.
24) **Satisfaction** (Categorical): Target variable indicating if the passenger was "Satisfied" or "Dissatisfied."

Data Types

- **Categorical Variables**: Gender, Customer Type, Type of Travel, Class, Satisfaction.
- **Numerical Variables**: Age, Flight Distance, service ratings, delay times.

This dataset provides a comprehensive view of passenger experiences, enabling tasks such as predicting satisfaction, analysing delays, and identifying clusters of similar passengers.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a crucial step in understanding the dataset and preparing it for analysis. This section focuses on data cleaning, transformation, and visualization to uncover patterns and address any data issues.

Data cleaning

Dropping Unnecessary Columns:

Columns such as Unnamed: 0 (index column) and ID (unique identifier) are dropped as they do not contribute to the analysis or model building.

Checking Data Types:

Each column's data type is checked to ensure consistency and compatibility for analysis. The following table summarizes the data types for all columns in the dataset:

Column	Data Type
Gender	object
Customer Type	object
Age	int64
Type of Travel	object
Class	object
Flight Distance	int64
Inflight Wifi service	int64
Departure/Arrival time convenient	int64
Ease of Online booking	int64
Gate location	int64
Food and drink	int64
Online boarding	int64

Seat comfort	int64
Inflight entertainment	int64
On-board service	int64
Leg room service	int64
Baggage handling	int64
Checkin service	int64
Inflight service	int64
Cleanliness	int64
Departure Delay in Minutes	int64
Arrival Delay in Minutes	float64
satisfaction	object

Table 1 EDA Feature Data Type

This ensures the identification of columns requiring transformations, such as encoding categorical variables or scaling numerical features.

Separating Categorical and Numerical Features:

Columns are divided into categorical and numerical groups for better handling and pre-processing:

Categorical Columns:

Gender, Customer Type, Type of Travel, Class, Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness, and Satisfaction.

Numerical Columns:

Age, Flight Distance, Departure Delay in Minutes, and Arrival Delay in Minutes.

This step ensures proper handling of each feature type during pre-processing, such as encoding for categorical variables and scaling for numerical variables.

Checking Missing Values:

The dataset is evaluated for missing values in each column:

- Most columns have 0 missing values.
- Arrival Delay in Minutes has 310 missing entries.

Dropping Rows with Missing Values:

After imputation, rows with missing values in other columns, if any, are dropped to ensure a clean dataset.

Checking for Duplicate Entries:

The dataset is checked for duplicate rows. In this case, no duplicate entries were found.

Class Distribution Analysis:

The target variable (Satisfaction) is analysed to identify class distribution:

Neutral or Dissatisfied: 56.66% of the dataset.

Satisfied: 43.34% of the dataset.

These steps ensure the dataset is clean, consistent, and ready for exploratory data analysis and machine learning tasks.

Data wrangling

Data wrangling involves preparing and transforming the dataset into a structured format suitable for analysis and modelling. The key steps in this process include feature selection and feature transformation.

1. **Log Transformation**:
    - For Departure Delay in Minutes and Arrival Delay in Minutes, a log transformation (log(1+x)) is applied to reduce the impact of skewness and outliers.
    - New columns created:
        - log_DepartureDelayMinutes
        - log_ArrivalDelayMinutes

1) Feature Transformation

I. Scaling**:**

- Normalize or standardize numerical features such as Flight Distance, Age, and transformed delay features to bring them onto a similar scale.
- This is essential for machine learning algorithms sensitive to feature magnitudes.

II. Dimensionality Reduction**:**

- Apply techniques like Principal Component Analysis (PCA) to reduce dimensionality, especially when dealing with a high number of correlated features.
- This helps in reducing computational complexity and retaining the most informative components.

Features

1) Gender

The Gender feature is a categorical variable with two values: Male (1) and Female (0). It represents the passenger's gender and may provide insights into preferences or satisfaction levels. This feature is balanced, ensuring no bias in gender representation within the dataset.

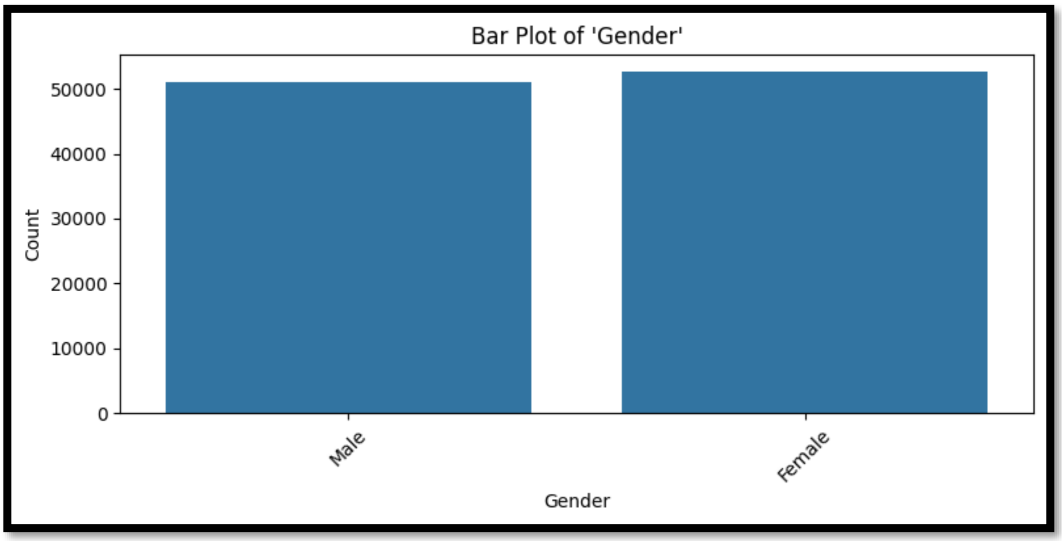


Figure 1 Bar plot of Gender

The bar graph shows an almost equal distribution of male and female passengers. This balance ensures a fair analysis of gender-related trends without overrepresentation or underrepresentation of either group.

The Gender feature is mapped as Male (1) and Female (0), representing the passenger's gender. It shows a negligible correlation with satisfaction (0.012), indicating that gender has little impact on satisfaction. Hypothesis testing using the Chi-square test confirms a statistically significant relationship between gender and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that gender impacts satisfaction, though the effect size is minimal.

For delays, gender shows no meaningful correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that gender has no significant relationship with delays

2) Customer Type

The Customer Type feature is a categorical variable with two values: Loyal Customer (1) and Disloyal Customer (0). It identifies whether a passenger is a loyal or infrequent customer, which can significantly influence satisfaction and travel preferences.

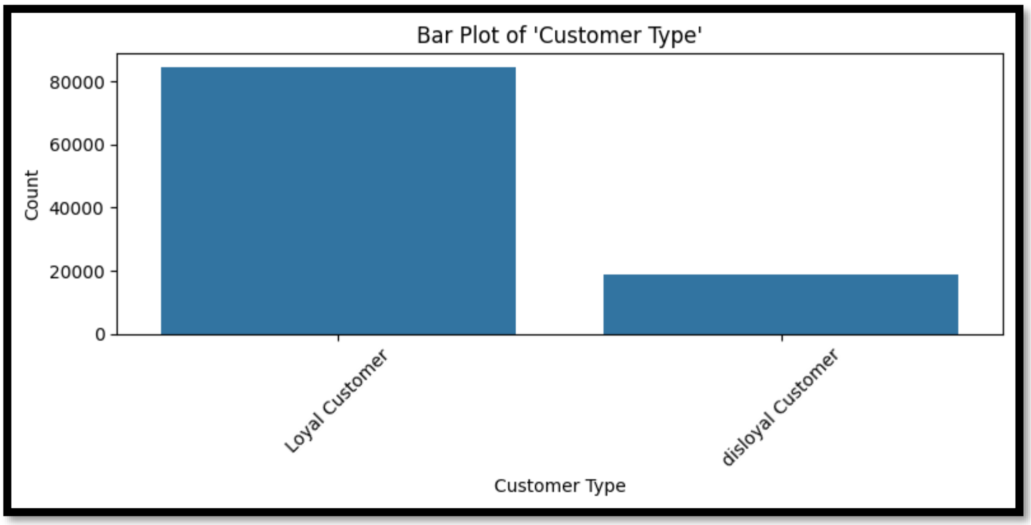


Figure 2Plot of Customer Type

The bar graph shows that most passengers are Loyal Customers, with a smaller proportion classified as Disloyal Customers. This imbalance may indicate that loyalty programs and frequent travellers dominate the dataset.

The Customer Type feature is mapped as Loyal Customer (1) and Disloyal Customer (0), indicating passenger loyalty status. It has a moderate positive correlation with satisfaction (0.19), suggesting that loyal customers are more likely to be satisfied. Hypothesis testing using the Chi-square test confirms a significant relationship between customer type and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that customer type impacts satisfaction. However, customer type shows weak or no correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that customer type has no significant relationship with delays.

3) Cleanliness

The Cleanliness feature is an ordinal categorical variable with values ranging from 0 (very poor cleanliness) to 5 (excellent cleanliness). It reflects passengers' ratings of the cleanliness onboard and is a critical aspect of service quality in the airline industry.

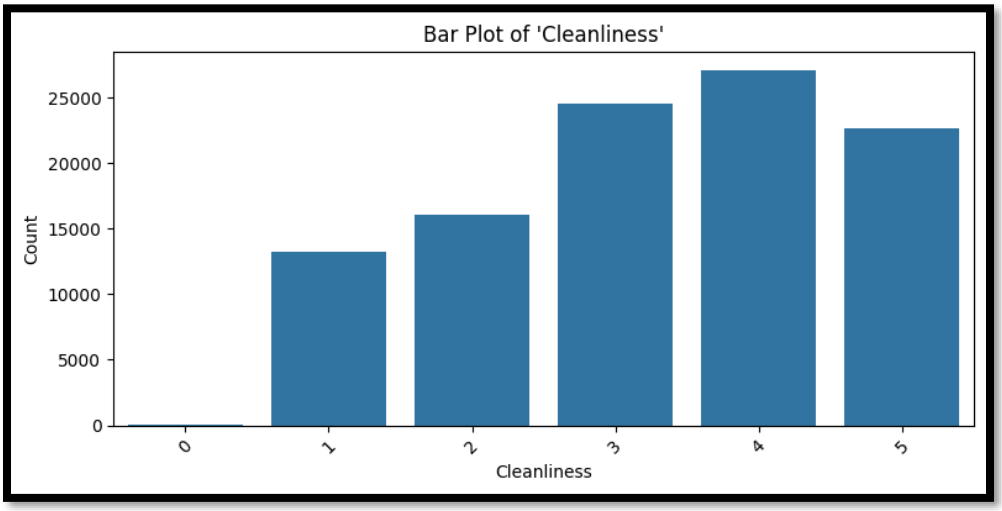


Figure 3 Bar Plot of Cleanliness

The bar graph shows a majority of passengers rated cleanliness as 4 or 5, indicating overall satisfaction with cleanliness. Lower ratings (0 and 1) are minimal, showing that poor cleanliness is rare in the dataset.

The Cleanliness feature is directly represented as ordinal numerical values (0–5), where higher ratings indicate better cleanliness. It has a moderate positive correlation with satisfaction (0.31), suggesting that passengers who rate cleanliness higher are more likely to be satisfied. Hypothesis testing using the Chi-square test confirms a significant relationship between cleanliness and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that cleanliness impacts satisfaction. However, cleanliness shows minimal correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that cleanliness has no significant relationship with delays.

4) Type of Travel

The Type of Travel feature is a categorical variable with two values: Personal Travel (0) and Business Travel (1). It represents the purpose of the passenger’s trip and can have a significant influence on satisfaction and service preferences

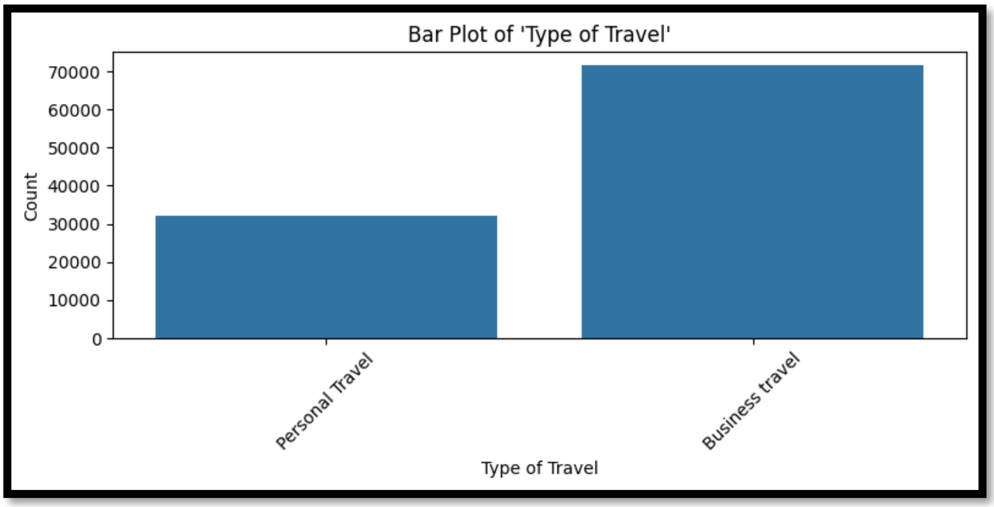


Figure 4 Bar Plot of Type of Travel

The bar graph indicates that the majority of passengers are traveling for Business purposes, while a smaller proportion are engaged in Personal Travel. This imbalance reflects a business-heavy passenger base in the dataset.

The Type of Travel feature is mapped as Personal Travel (0) and Business Travel (1). It has a strong positive correlation with satisfaction (0.45), indicating that business travellers are more likely to be satisfied compared to personal travellers. Hypothesis testing using the Chi-square test confirms a significant relationship between type of travel and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that travel type impacts satisfaction. However, the feature shows no significant correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that type of travel has no significant relationship with delays.

5) Class

The Class feature is a categorical variable with three values: Eco (0), Eco Plus (1), and Business (2). It represents the travel class of passengers, which is a key determinant of service quality and satisfaction levels.

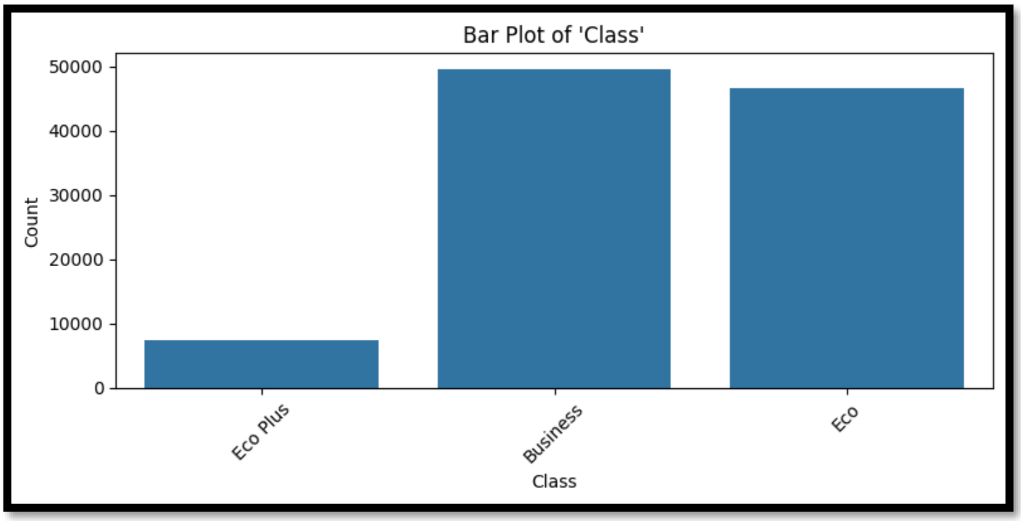


Figure 5 Bar Plot of Class

The bar graph shows that most passengers travel in Business and Eco classes, while a smaller portion belongs to Eco Plus. This indicates that Economy and Business classes dominate the dataset.

The Class feature is mapped as Eco (0), Eco Plus (1), and Business (2). It has a strong positive correlation with satisfaction (0.49), indicating that passengers in higher classes (e.g., Business) tend to be more satisfied. Hypothesis testing using the Chi-square test confirms a significant relationship between class and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that class impacts satisfaction. However, there is no significant correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that class has no significant relationship with delays.

6) Inflight service

The Inflight Service feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It reflects passengers' ratings of the quality of service provided during the flight and is a critical determinant of customer satisfaction.

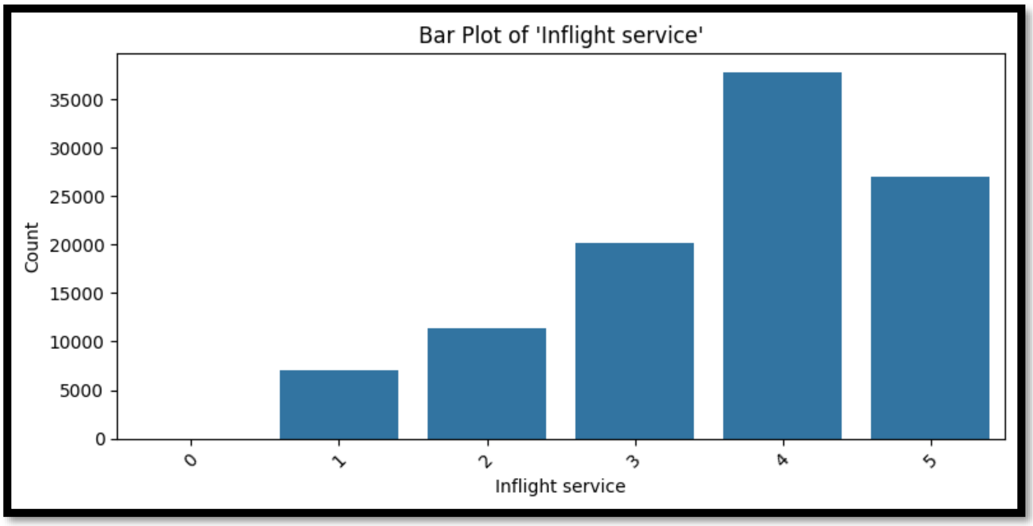


Figure 6 Bar Plot of Inflight Service

The bar graph shows that most passengers rated the inflight service as 4 or 5, indicating overall satisfaction with the quality of service. Lower ratings (0, 1, and 2) are comparatively less frequent, suggesting that poor service experiences are less common.

The Inflight Service feature is directly represented as ordinal numerical values (0–5). It has a moderate positive correlation with satisfaction (0.24), indicating that higher inflight service ratings are associated with increased passenger satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between inflight service and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that inflight service impacts satisfaction. However, there is minimal correlation between inflight service and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that inflight service has no significant relationship with delays.

7) Inflight wifi service

The Inflight Wi-Fi Service feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the Wi-Fi quality during the flight, which is a key factor for passengers seeking connectivity.

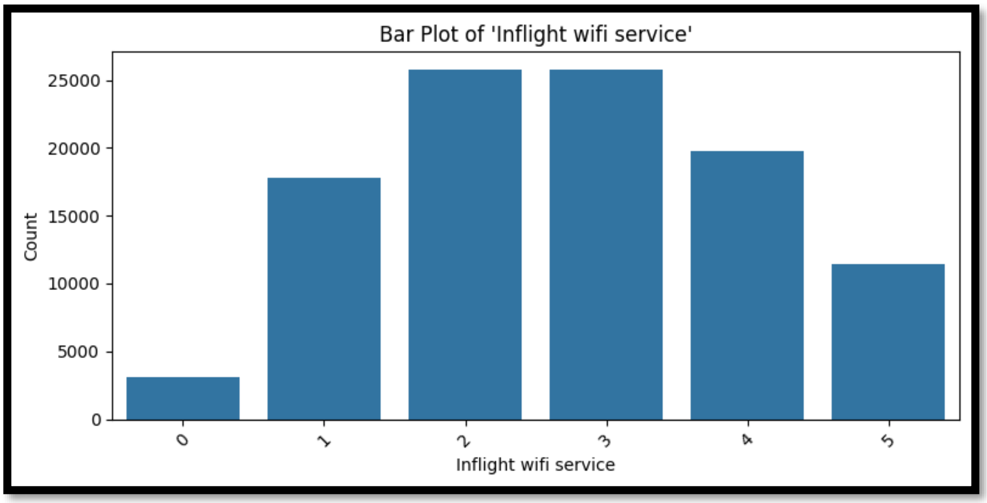


Figure 7 Bar Plot of Inflight Wi-Fi Service

The bar graph shows that most passengers rated the Wi-Fi service as **2** or **3**, indicating average performance. Ratings of **0** (very poor) and **5** (excellent) are less common, suggesting that Wi-Fi quality varies significantly among flights.

The Inflight Wi-Fi Service feature is represented as ordinal numerical values (0–5). It has a moderate positive correlation with satisfaction (0.28), indicating that better Wi-Fi service contributes to higher satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between Wi-Fi service and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that Wi-Fi quality impacts satisfaction. There is minimal to no correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that inflight Wi-Fi service has no significant relationship with delays.

8) Departure/Arrival time convenient

The Departure/Arrival Time Convenient feature is an ordinal categorical variable with values ranging from 0 (very inconvenient) to 5 (very convenient). It reflects passengers' ratings of the convenience of departure and arrival times, which can influence their overall travel experience.

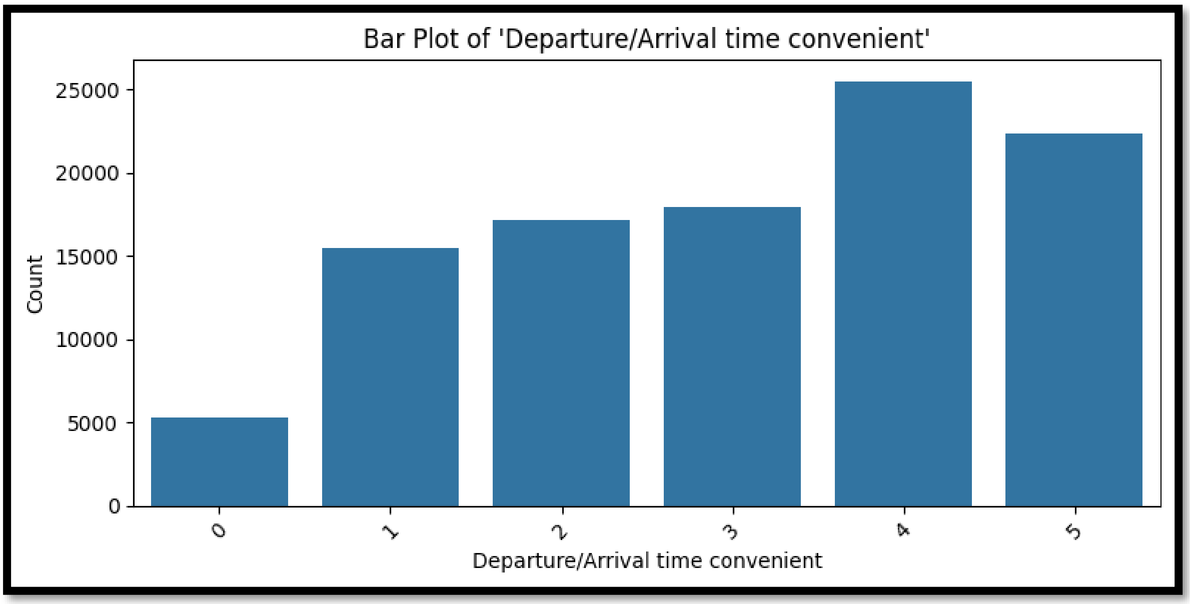


Figure 8 Bar Plot of Departure/Arrival time convenient

The bar graph indicates that most passengers rated the convenience as 4 or 5, suggesting that the majority found the schedule suitable. Ratings of 0 and 1 are relatively rare, indicating that highly inconvenient schedules are uncommon.

The Departure/Arrival Time Convenient feature is represented as ordinal numerical values (0–5). It shows negligible correlation with satisfaction (-0.052), indicating that schedule convenience does not significantly impact satisfaction levels. Hypothesis testing using the Chi-square test supports the null hypothesis (H_0) that there is no significant relationship between schedule convenience and satisfaction. Similarly, there is no meaningful correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, further supporting the null hypothesis (H_0) that schedule convenience has no significant relationship with delays.

9) Ease of Online booking

The Ease of Online Booking feature is an ordinal categorical variable with values ranging from 0 (very difficult) to 5 (very easy). It represents passengers' ratings of how straightforward and convenient the online booking process was.

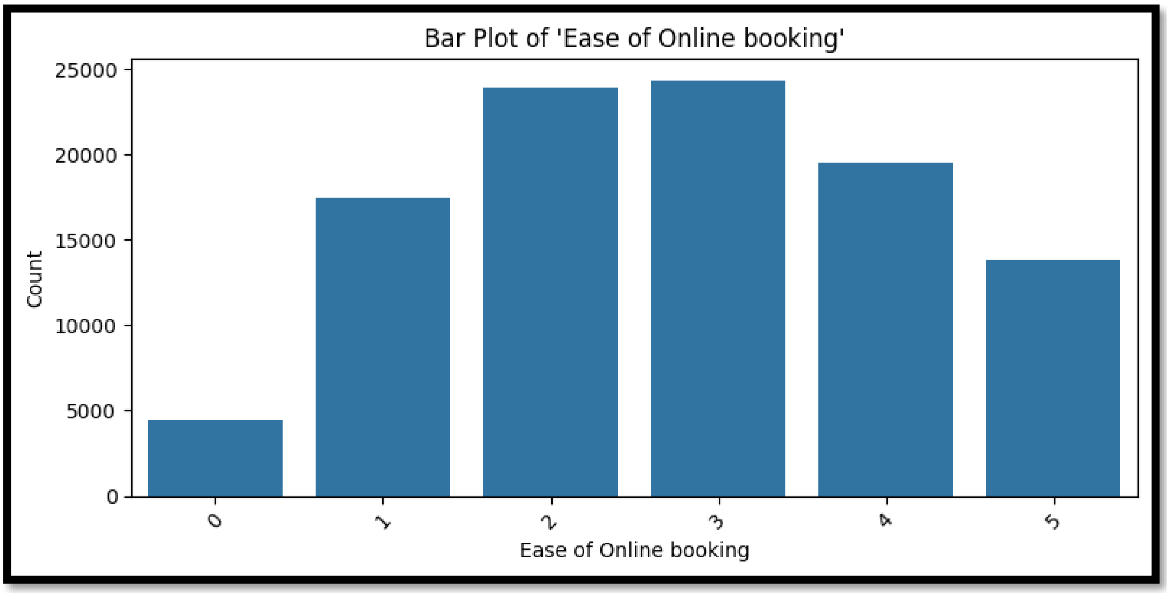


Figure 9 Bar Plot of Ease of Online booking

The bar graph shows that most passengers rated the ease of online booking as 2 or 3, indicating an average level of satisfaction with the booking process. Ratings of 0 (very difficult) and 5 (very easy) are less common, reflecting mixed perceptions about the system's user-friendliness.

The Ease of Online Booking feature is represented as ordinal numerical values (0–5). It has a weak positive correlation with satisfaction (0.17), suggesting that an easier online booking experience slightly contributes to higher satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between ease of online booking and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that ease of online booking impacts satisfaction. However, there is no significant correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that ease of online booking has no significant relationship with delays.

10) Gate location

The Gate Location feature is an ordinal categorical variable with values ranging from 0 (very inconvenient) to 5 (very convenient). It represents passengers' ratings of the accessibility and convenience of the assigned gate for boarding.

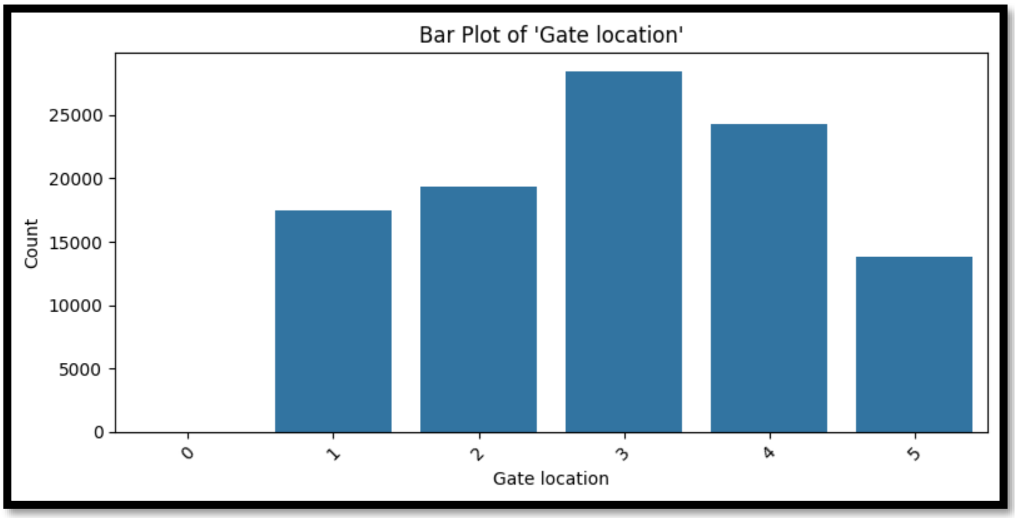


Figure 10 Bar Plot of Gate Location

The bar graph indicates that most passengers rated the gate location as 3 or 4, showing a generally positive perception of gate accessibility. Lower ratings (0 and 1) and the highest rating (5) are less frequent, suggesting that extremes in convenience are uncommon.

The Gate Location feature is represented as ordinal numerical values (0–5). It has a negligible correlation with satisfaction (0.00045), indicating that gate location has no significant influence on overall satisfaction. Hypothesis testing using the Chi-square test supports the null hypothesis (H_0) that gate location has no significant relationship with satisfaction. Similarly, there is no meaningful correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, further supporting the null hypothesis (H_0) that gate location has no significant relationship with delays.

11) Food and drink

The Food and Drink feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the quality and availability of food and beverages during their flight.

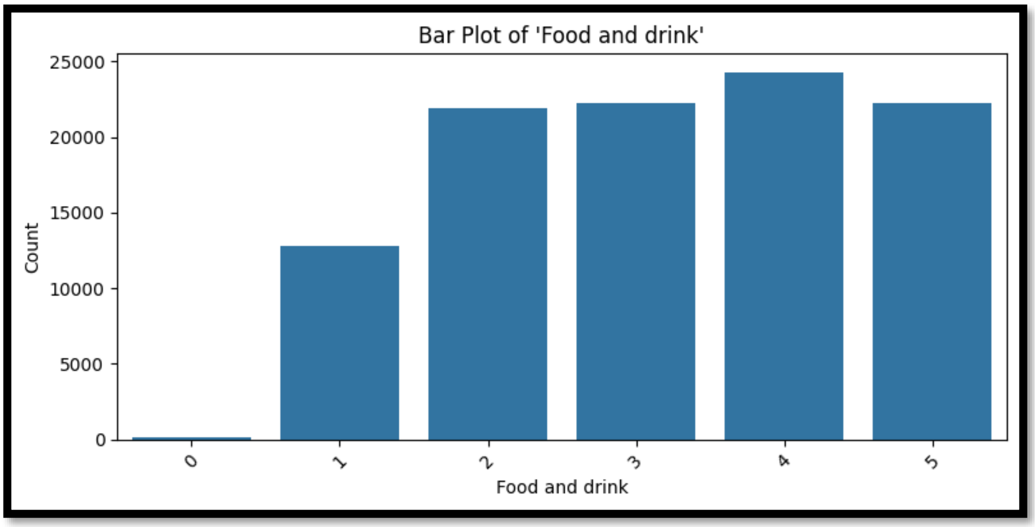


Figure 11 Bar Plot of Food and Drink

The bar graph shows that most passengers rated food and drink quality as 4 or 5, indicating overall satisfaction with the service. Lower ratings (0 and 1) are relatively rare, suggesting that poor food and drink experiences are uncommon in the dataset.

The Food and Drink feature is represented as ordinal numerical values (0–5). It has a weak positive correlation with satisfaction (0.21), indicating that higher food and drink ratings are modestly associated with increased satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between food and drink quality and satisfaction ($p\text{-value} < 0.05$), supporting the alternative hypothesis (H_1) that food and drink quality impacts satisfaction. There is no significant correlation between food and drink quality and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that food and drink quality has no significant relationship with delays

12) Online boarding

The Online Boarding feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the efficiency and ease of the online boarding process.

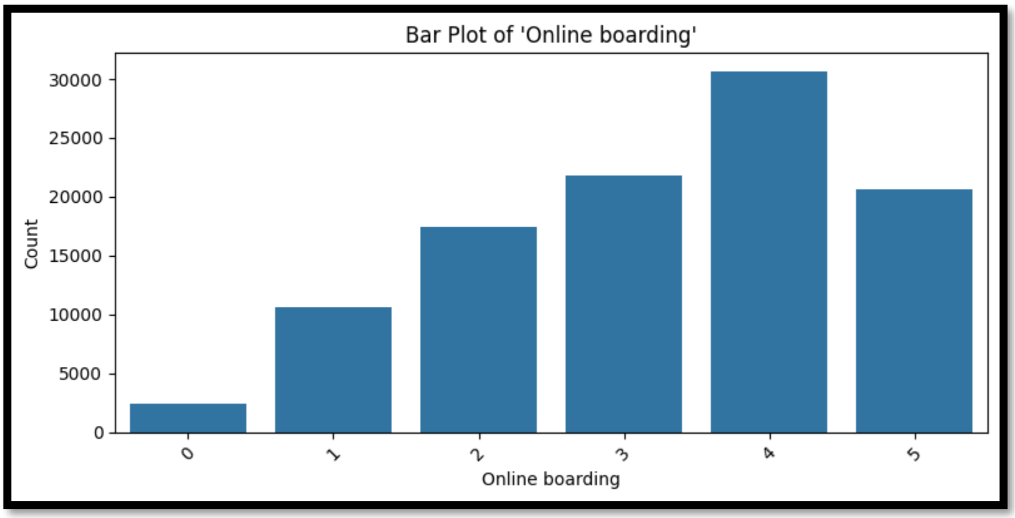


Figure 12 Bar Plot of Online boarding

The bar graph indicates that most passengers rated online boarding as 4 or 5, reflecting high satisfaction with the boarding system. Lower ratings (0 and 1) are comparatively rare, indicating minimal dissatisfaction with this feature.

The Online Boarding feature is represented as ordinal numerical values (0–5). It has the strongest positive correlation with satisfaction (0.50) among all features, showing that an efficient online boarding process significantly impacts passenger satisfaction. Hypothesis testing using the Chi-square test confirms a strong and significant relationship (p-value < 0.05), supporting the alternative hypothesis (H_1) that online boarding efficiency impacts satisfaction. However, there is no meaningful correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that online boarding has no significant relationship with delays.

13) Seat comfort

The Seat Comfort feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the comfort level of their seats during the flight.

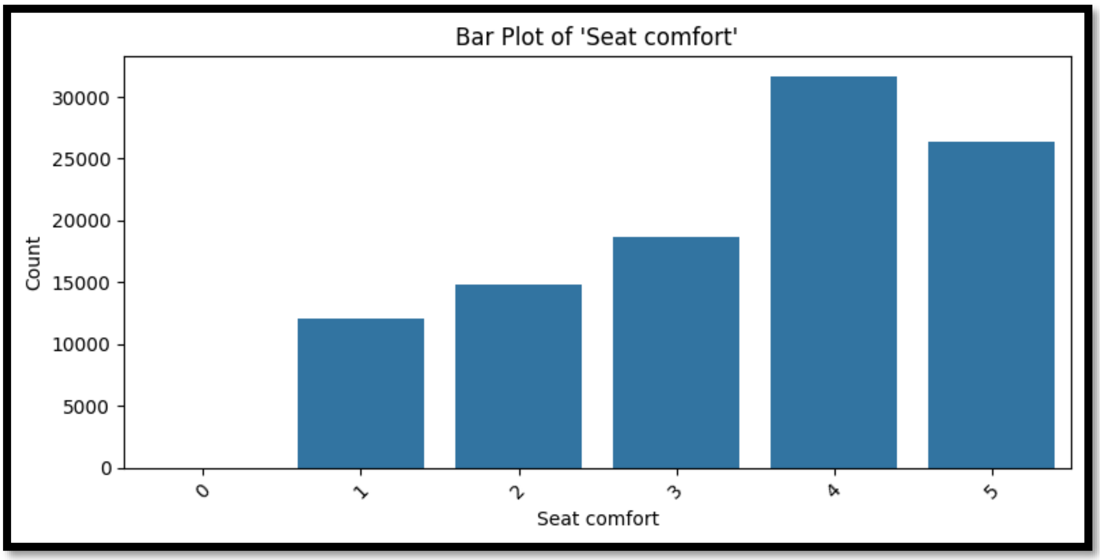


Figure 13 Bar Plot of Seat Comfort

The bar graph shows that most passengers rated seat comfort as 4 or 5, indicating general satisfaction with seating arrangements. Lower ratings (0 and 1) are less common, suggesting that uncomfortable seating experiences are relatively rare.

The Seat Comfort feature is represented as ordinal numerical values (0–5). It has a moderate positive correlation with satisfaction (0.35), indicating that better seat comfort strongly influences passenger satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between seat comfort and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that seat comfort impacts satisfaction. There is no significant correlation between seat comfort and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that seat comfort has no significant relationship with delays.

14) Inflight entertainment

The Inflight Entertainment feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the quality and availability of entertainment options during their flight.

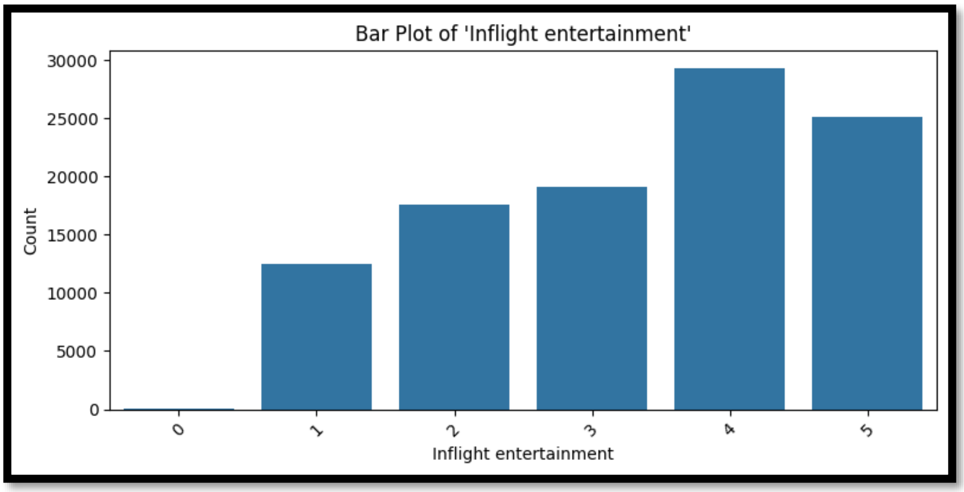


Figure 14 Bar Plot of Inflight entertainment

The bar graph indicates that the majority of passengers rated inflight entertainment as 4 or 5, showing high satisfaction with the entertainment offerings. Lower ratings (0 and 1) are relatively infrequent, suggesting minimal dissatisfaction with this feature.

The Inflight Entertainment feature is represented as ordinal numerical values (0–5). It has a moderate positive correlation with satisfaction (0.40), indicating that better inflight entertainment strongly influences passenger satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between inflight entertainment and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that inflight entertainment impacts satisfaction. There is no significant correlation with delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that inflight entertainment has no significant relationship with delays.

15) On-board service

The On-board Service feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the quality of service provided by the airline staff during the flight.

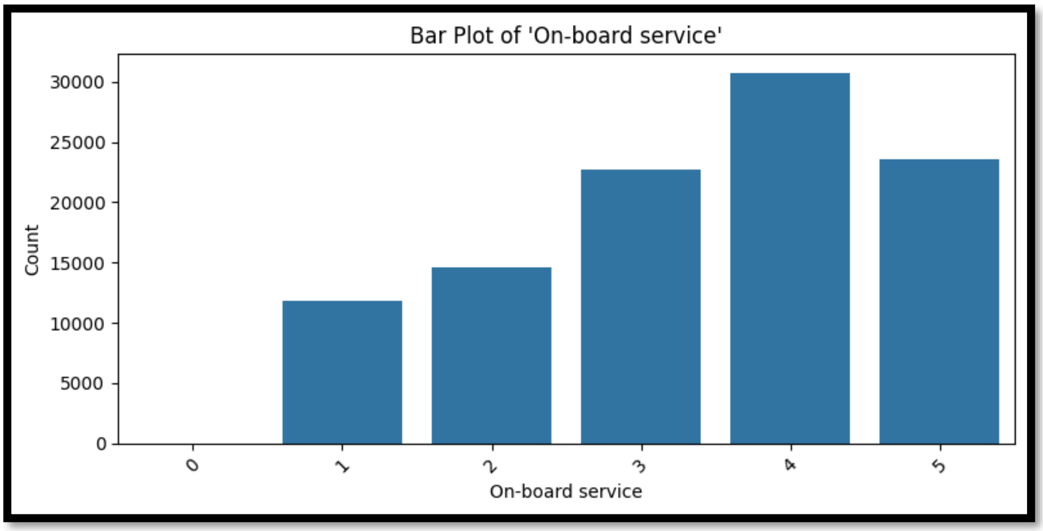


Figure 15 Bar Plot of On-board Service

The bar graph shows that most passengers rated on-board service as 4 or 5, indicating high levels of satisfaction with the service quality. Lower ratings (0 and 1) are less frequent, suggesting that poor service experiences are rare.

The On-board Service feature is represented as ordinal numerical values (0–5). It has a moderate positive correlation with satisfaction (0.32), showing that better on-board service contributes significantly to passenger satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between on-board service and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that on-board service quality impacts satisfaction. However, there is no significant correlation between on-board service and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that on-board service has no significant relationship with delays.

16) Leg room service

The Leg Room Service feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the comfort and space available for legroom during the flight.

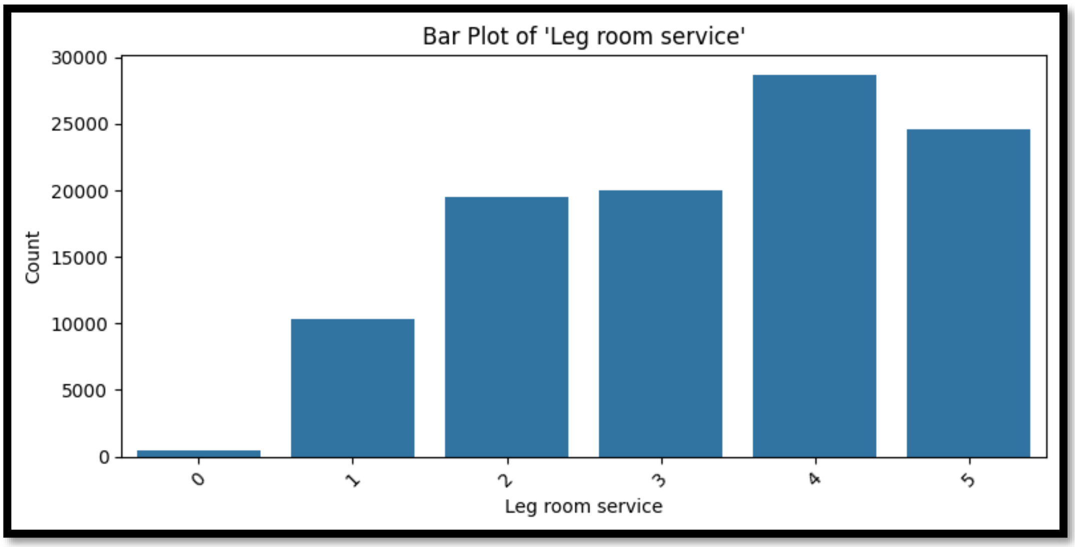


Figure 16 Bar Plot of Leg room service

The bar graph indicates that most passengers rated legroom service as 4 or 5, reflecting overall satisfaction with the space provided. Lower ratings (0 and 1) are less common, suggesting that complaints about legroom are relatively rare.

The Leg Room Service feature is represented as ordinal numerical values (0–5). It has a moderate positive correlation with satisfaction (0.31), indicating that better legroom service significantly influences passenger satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between legroom service and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that legroom service impacts satisfaction. However, there is no significant correlation between legroom service and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that legroom service has no significant relationship with delays.

17) Baggage handling

The Baggage Handling feature is an ordinal categorical variable with values ranging from 1 (very poor) to 5 (excellent). It represents passengers' ratings of the efficiency and care taken during baggage handling.

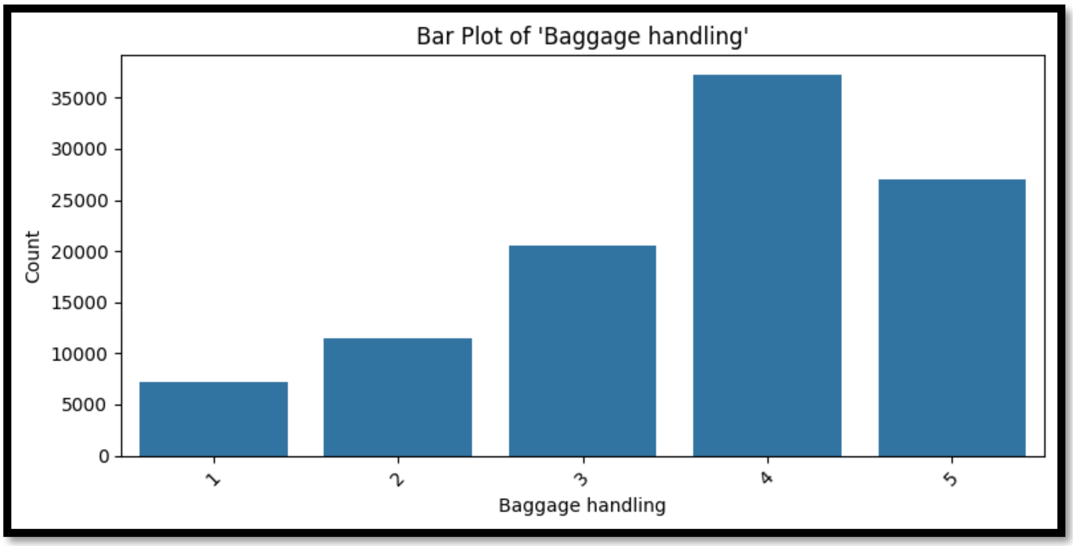


Figure 17 Bar Plot of Baggage Handling

The bar graph shows that most passengers rated baggage handling as 4 or 5, indicating general satisfaction with this service. Lower ratings (1 and 2) are relatively less frequent, suggesting minimal dissatisfaction with baggage handling in the dataset.

The Baggage Handling feature is represented as ordinal numerical values (1–5). It has a weak positive correlation with satisfaction (0.25), suggesting that better baggage handling contributes to higher satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between baggage handling and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that baggage handling impacts satisfaction. However, there is no significant correlation between baggage handling and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that baggage handling has no significant relationship with delays.

18) Checkin service

The Check-in Service feature is an ordinal categorical variable with values ranging from 0 (very poor) to 5 (excellent). It represents passengers' ratings of the efficiency and quality of the check-in process.

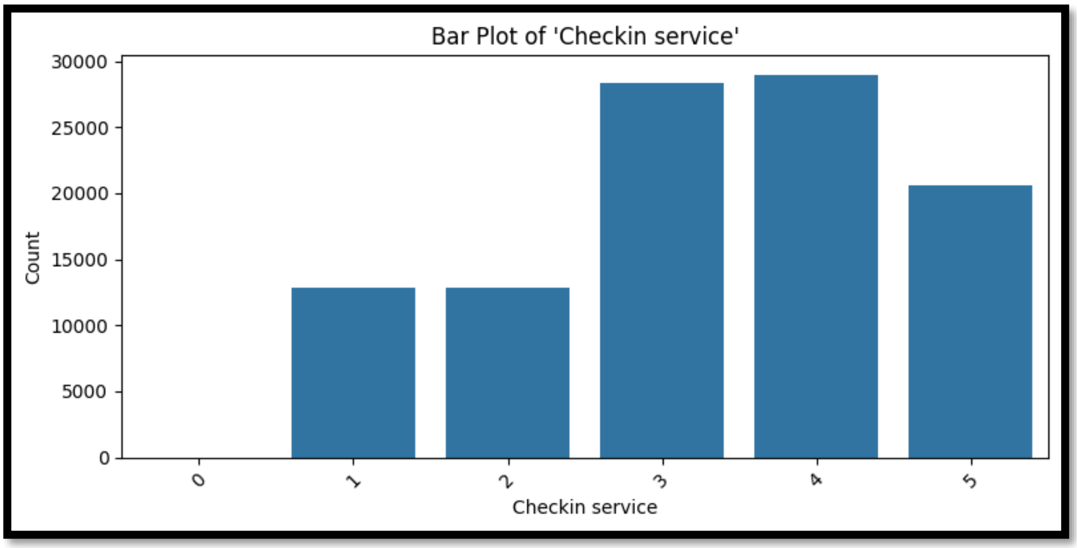


Figure 18Check-in Service

The bar graph shows that most passengers rated the check-in service as 3, 4, or 5, indicating overall satisfaction with this service. Lower ratings (0, 1, and 2) are less common, reflecting fewer complaints about the check-in process.

The Check-in Service feature is represented as ordinal numerical values (0–5). It has a weak positive correlation with satisfaction (0.24), suggesting that efficient check-in contributes to higher passenger satisfaction. Hypothesis testing using the Chi-square test confirms a significant relationship between check-in service and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that check-in service impacts satisfaction. There is no significant correlation between check-in service and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that check-in service has no significant relationship with delays.

19) satisfaction

The Satisfaction feature is a categorical variable with two values: neutral or dissatisfied (0) and satisfied (1). It represents the target variable, capturing the overall satisfaction levels of passengers based on their experiences.

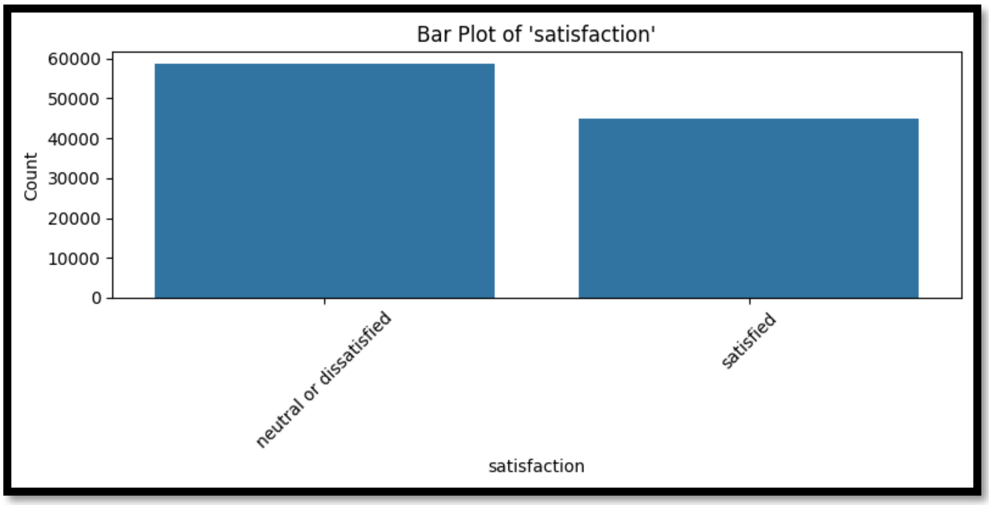


Figure 19 Bar Plot of Satisfaction

The bar graph shows that the majority of passengers are categorized as neutral or dissatisfied, with a slightly smaller proportion being satisfied. This indicates a class imbalance in the dataset, which may need to be addressed in predictive modelling.

The Satisfaction feature is mapped as neutral or dissatisfied (0) and satisfied (1). It serves as the dependent variable in the analysis, with multiple features showing significant correlations with satisfaction, such as Online Boarding (0.50), Class (0.49), and Type of Travel (0.45). Hypothesis testing using the Chi-square test confirms significant relationships between satisfaction and various features (p-value < 0.05), supporting the alternative hypothesis (H_1) that these features impact satisfaction. Addressing the imbalance in satisfaction classes will be crucial for accurate model training and evaluation.

20) Flight Distance

The Flight Distance feature is a numerical continuous variable that represents the distance of a flight in miles. It is a significant feature that may influence passengers' overall experience, especially for long-haul flights where service quality becomes critical.

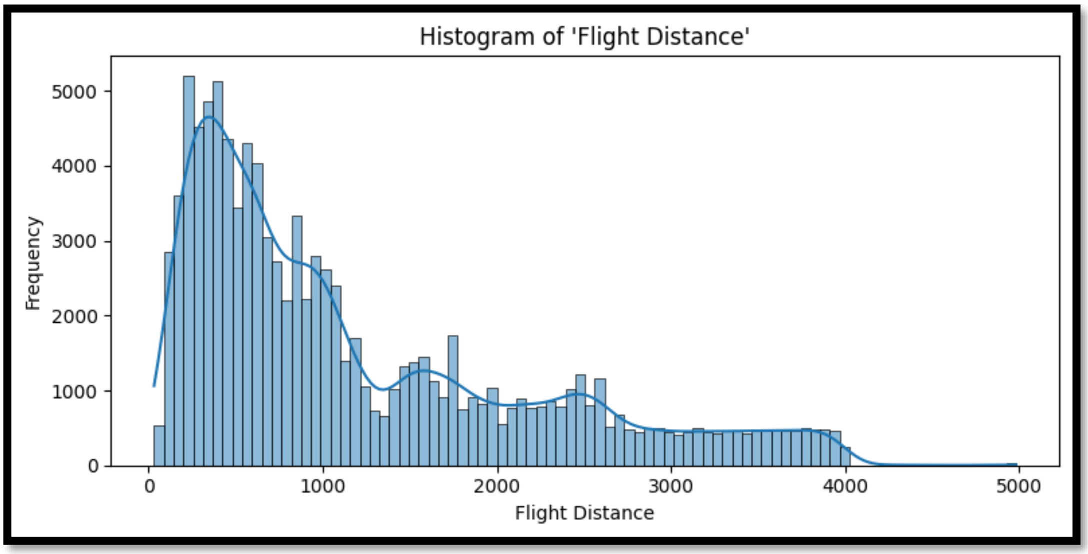


Figure 20 Histogram of Flight Distance

The histogram of Flight Distance shows a right-skewed distribution, with the majority of flights clustered within shorter distances (0–1500 miles). A smaller portion of flights spans longer distances, indicating that most passengers in the dataset are traveling short- to medium-haul routes.

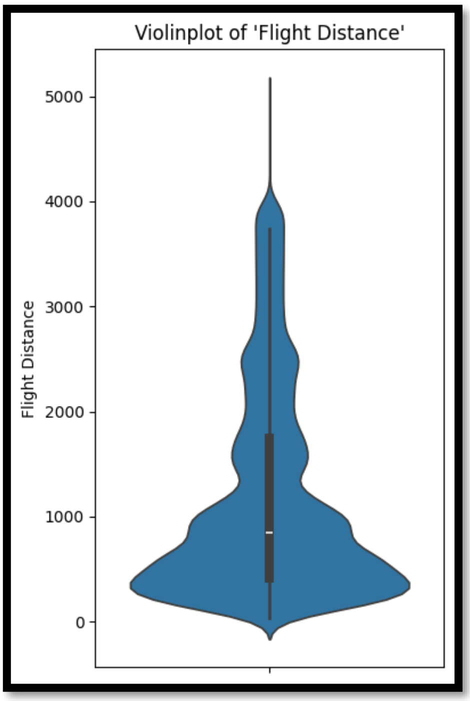


Figure 21Violin plot of Flight Distance

The violin plot provides a detailed representation of the distribution of Flight Distance. It shows a right-skewed distribution, with the majority of flight distances concentrated in the lower range (below 1500 miles). The plot also highlights the density of data points at different ranges, indicating that short-haul flights are more common. The long upper tail indicates the presence of long-haul flights, though these are less frequent.

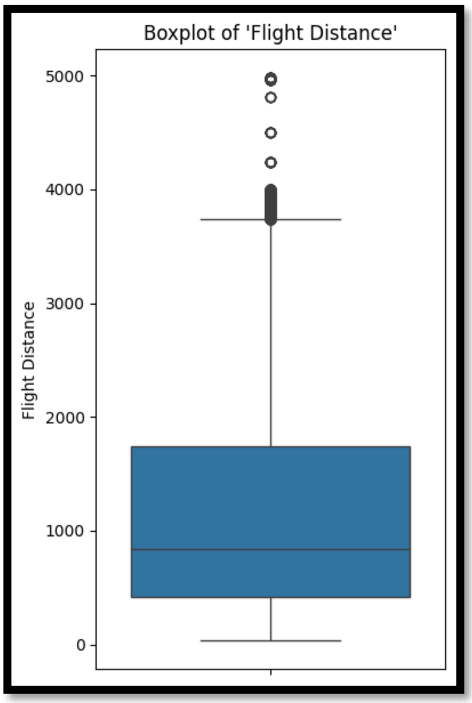


Figure 22 Boxplot of Flight Distance

The box plot of Flight Distance confirms the right-skewed distribution shown in the histogram and violin plot. The median flight distance is around 1000 miles, and the interquartile range (IQR) spans from approximately 500 to 2000 miles. Several outliers are present beyond the upper whisker, representing long-haul flights that deviate significantly from the majority of the data.

The Flight Distance feature is represented as continuous numerical values. It has a weak positive correlation with satisfaction (0.30), suggesting that longer flight distances are slightly associated with higher passenger satisfaction, possibly due to better services offered on long-haul routes. Hypothesis testing using Pearson Correlation confirms a statistically significant relationship between flight distance and satisfaction ($p\text{-value} < 0.05$), supporting the alternative hypothesis (H_1) that flight distance impacts satisfaction. However, there is no significant relationship between flight distance and delay features, supporting the null hypothesis (H_0) that flight distance does not influence delays.

21) Age

The Age feature is a numerical continuous variable that represents the age of passengers. It provides critical demographic insights and can influence satisfaction levels and service preferences.

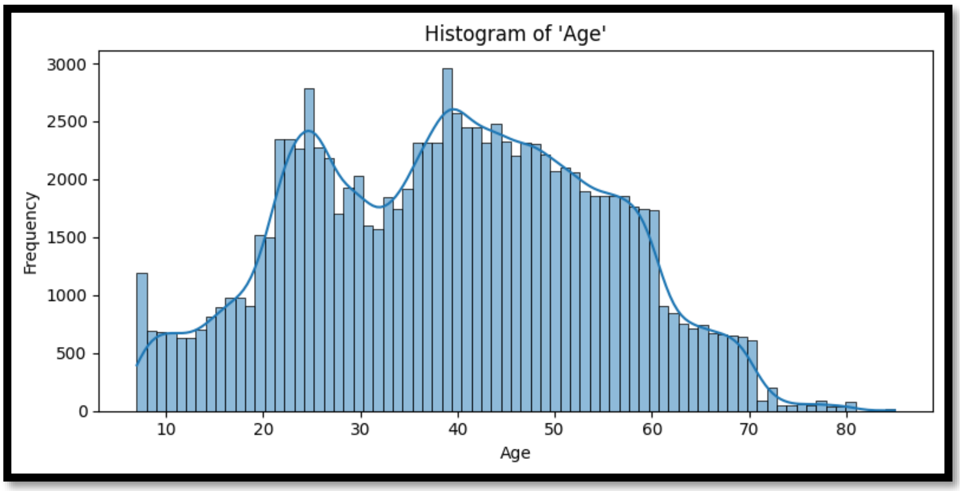


Figure 23Histogram of Age

The histogram of Age shows a bimodal distribution, with noticeable peaks around ages 18–20 (younger passengers) and 40–50 (middle-aged passengers). The data also shows a gradual decline after age 50, with fewer passengers in the older age brackets. This distribution suggests that the dataset is dominated by younger and middle-aged travellers.

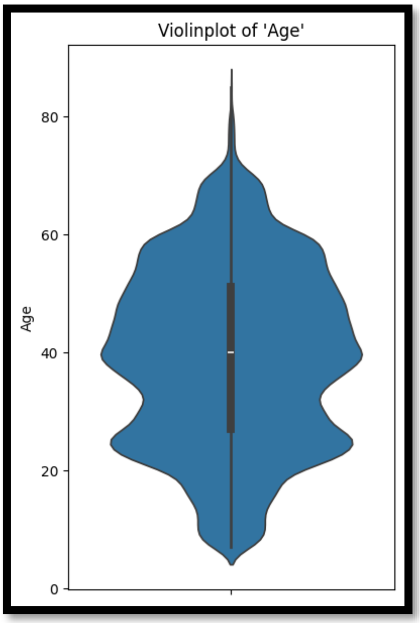


Figure 24 Violin Plot of Age

The violin plot provides a detailed view of the distribution of passenger ages. It highlights a symmetrical distribution around the median age of approximately 40 years, with significant density around ages 20 to 50. The tails extend to younger passengers (below 10) and older passengers (above 80), although the density decreases toward the extremes. This indicates a diverse age range but with a concentration in the middle-age brackets.

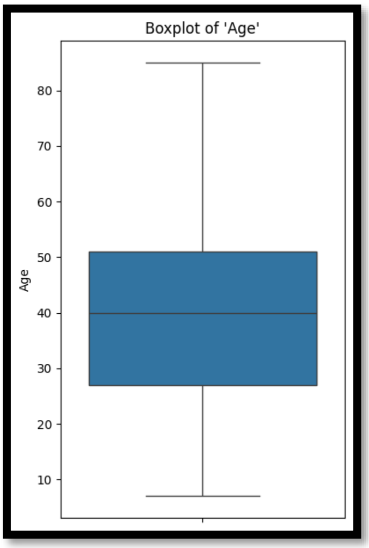


Figure 25Box Plot of Age

The box plot confirms that the median age of passengers is around 40 years, with an interquartile range (IQR) spanning approximately from 30 to 50 years. The data shows no significant outliers and the age distribution appears well-contained within the whiskers. This supports the notion that most passengers belong to the middle-age group.

The Age feature is represented as a continuous numerical variable. It shows a weak positive correlation with satisfaction (0.14), suggesting that older passengers may have slightly higher satisfaction levels compared to younger passengers. Hypothesis testing using Pearson Correlation confirms a statistically significant relationship between age and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that age impacts satisfaction. However, there is no significant correlation between age and delay features, as confirmed by ANOVA and Pearson Correlation tests, supporting the null hypothesis (H_0) that age has no significant relationship with delays.

22) Departure Delay in Minutes

The Departure Delay in Minutes feature is a numerical continuous variable that represents the delay in departure time for a flight, measured in minutes. It is an important feature that can impact passenger satisfaction and overall service perception.

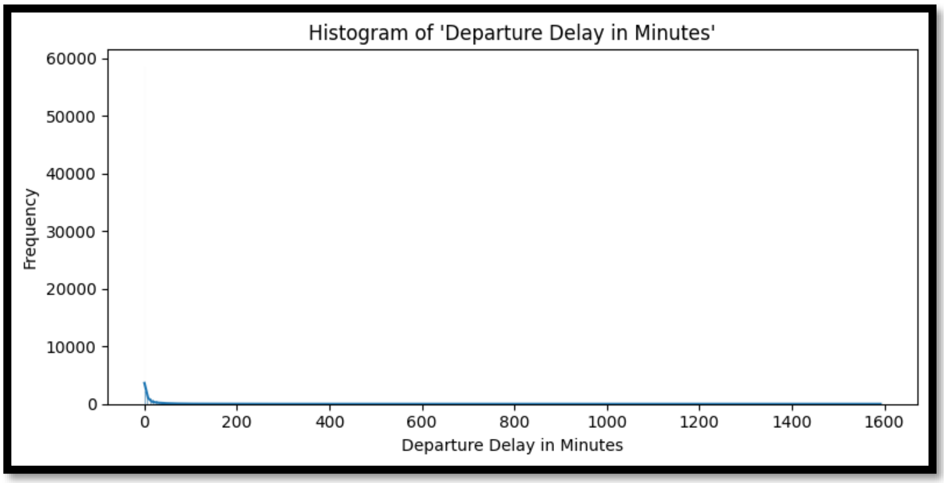


Figure 26 Histogram of Departure Delay Minutes

The histogram of Departure Delay in Minutes shows a highly right-skewed distribution, with the majority of flights having minimal to no delays. A long tail extends toward higher delay values, with a few flights experiencing extreme delays exceeding 1000 minutes. This indicates that most flights are on time or experience minor delays, while severe delays are relatively rare.

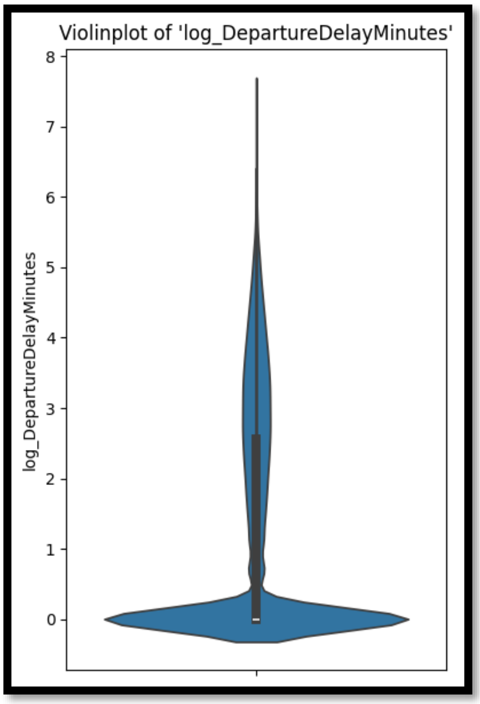


Figure 27Violin of Departure Delay in Minutes

he violin plot of the transformed feature further supports the compact distribution observed in the box plot. The data density is concentrated around the lower values, with a sharp tapering off toward higher values. The long tail visible in the plot represents the rare but extreme delays. The plot highlights that most delays are minimal while providing insight into the spread and density of the transformed data.

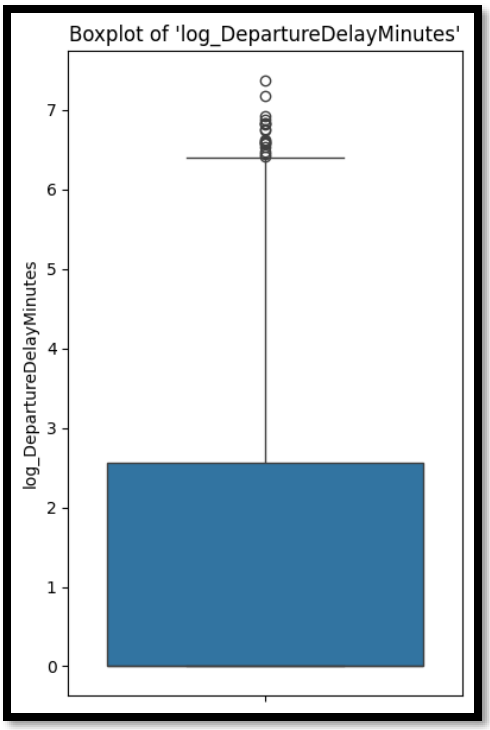


Figure 28 Boxplot of of Departure Delay in Minutes

The box plot of the logarithmically transformed Departure Delay in Minutes feature shows a much more compact distribution compared to the original delay feature. The median value is low, indicating that most delays are minor even after transformation. Outliers are still present beyond the upper whisker, but their impact is minimized due to the log transformation, making the data more suitable for analysis.

The Departure Delay in Minutes feature is represented as a continuous numerical variable. It has a negligible negative correlation with satisfaction (-0.051), indicating that longer departure delays have a minor impact on passenger satisfaction. Hypothesis testing using Pearson Correlation confirms that the relationship between departure delay and satisfaction is statistically significant ($p\text{-value} < 0.05$), supporting the alternative hypothesis (H_1) that departure delay impacts satisfaction, albeit weakly. However, there is no meaningful correlation with other features like inflight service or cleanliness, supporting the null hypothesis (H_0) that departure delay has no significant relationship with such factors.

23) log_DepartureDelayMinutes

The log_DepartureDelayMinutes feature is a logarithmically transformed version of the Departure Delay in Minutes variable. This transformation is applied to compress the skewed distribution of the original delay data, effectively reducing the impact of extreme outliers while retaining the overall distribution.

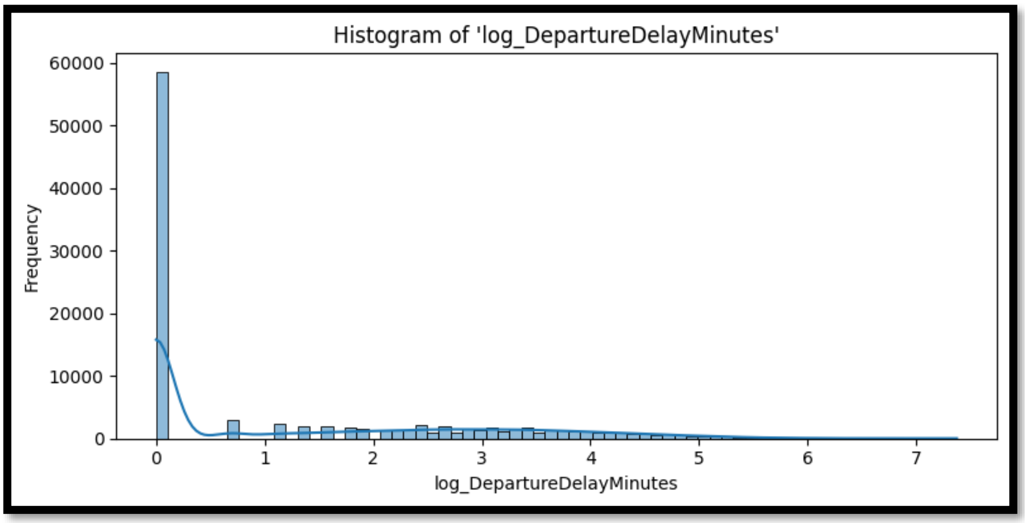


Figure 29 Histogram of log_DepartureDelayMinutes

The histogram of log_DepartureDelayMinutes shows a right-skewed distribution, with the majority of flights clustered at lower delay values (close to zero). The long tail represents rare but extreme delays, which are less impactful after the log transformation. The histogram highlights that most flights experience minimal departure delays, with significant delays being rare.

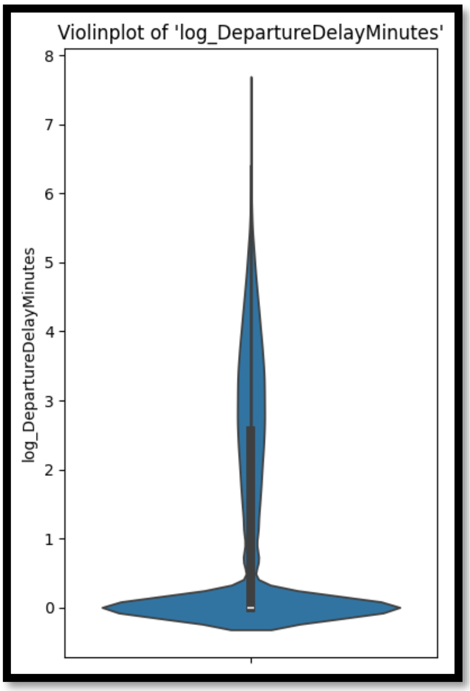


Figure 30 Violin plot log_DepartureDelayMinutes

The violin plot for log_DepartureDelayMinutes demonstrates the data density and range after applying a logarithmic transformation. The majority of the data is concentrated near zero, indicating minimal departure delays for most flights. The long upper tail represents the less frequent but extreme delays, effectively compressed through transformation to ensure better visual and analytical interpretation. a small number of flights with significant delays. The plot visually emphasizes that most delays are minimal while reducing the skew caused by extreme outliers.

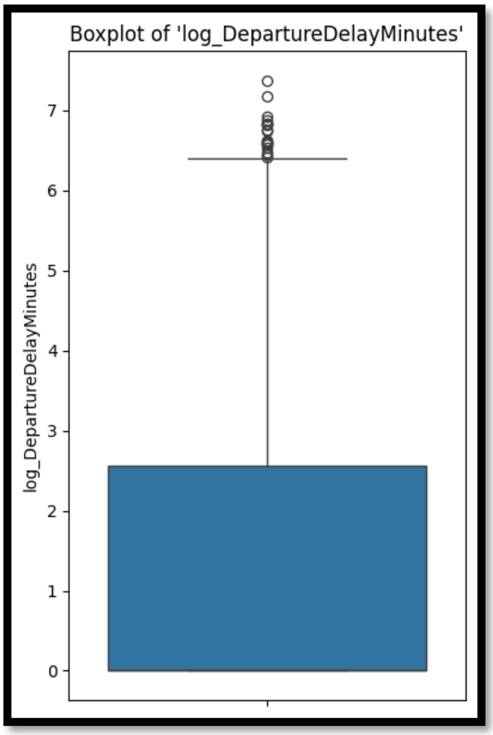


Figure 31 Boxplot of log_DepartureDelayMinutes

The box plot shows a compressed and normalized range for log_DepartureDelayMinutes. The median delay is low, and the interquartile range (IQR) highlights that most flights experience minimal delays. Outliers above the upper whisker represent rare but extreme delays, which are now easier to handle due to the transformation.

The log_DepartureDelayMinutes feature is a continuous numerical variable. It has a negligible negative correlation with satisfaction (-0.051), indicating that longer departure delays slightly decrease passenger satisfaction. Hypothesis testing using Pearson Correlation confirms a statistically significant relationship between departure delay and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that departure delays impact satisfaction. However, there is no significant correlation between departure delays and other service-related features, supporting the null hypothesis (H_0) that departure delays have no notable relationship with such factors.

24) Arrival Delay in Minutes

The Arrival Delay in Minutes feature is a numerical variable that represents the delay in a flight's arrival time in minutes. It is an important indicator of operational efficiency and customer satisfaction. However, the raw data is heavily skewed due to extreme delays.

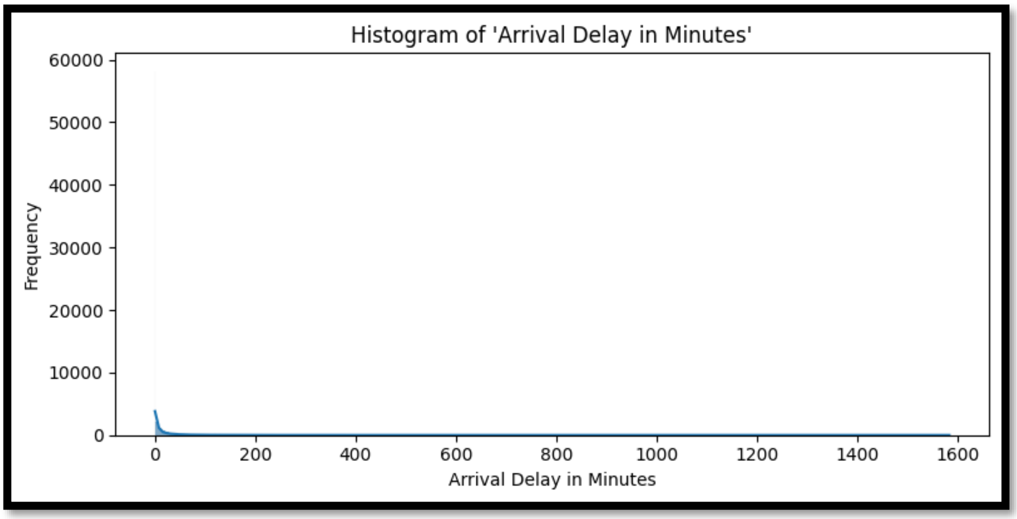


Figure 32 Histogram of Arrival Delay in Minutes

The histogram shows a right-skewed distribution, with the majority of flights having minimal to no arrival delays. The long tail of the histogram represents flights with significant delays, with a few extreme values exceeding 1500 minutes. This skewness makes raw data less suitable for direct analysis.

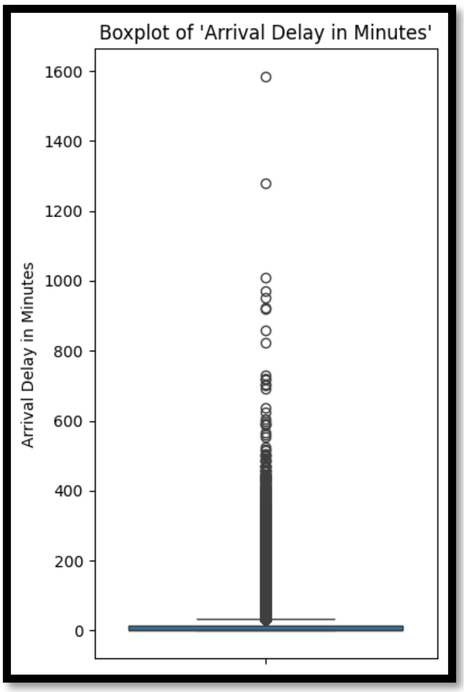


Figure 33 Boxplot of Arrival Delay in Minutes

The box plot highlights the presence of numerous outliers, which are delays far beyond the typical range. The interquartile range (IQR) is very narrow, indicating that most arrival delays are small, with a concentration near zero. The extreme outliers stretch up to 1600 minutes.

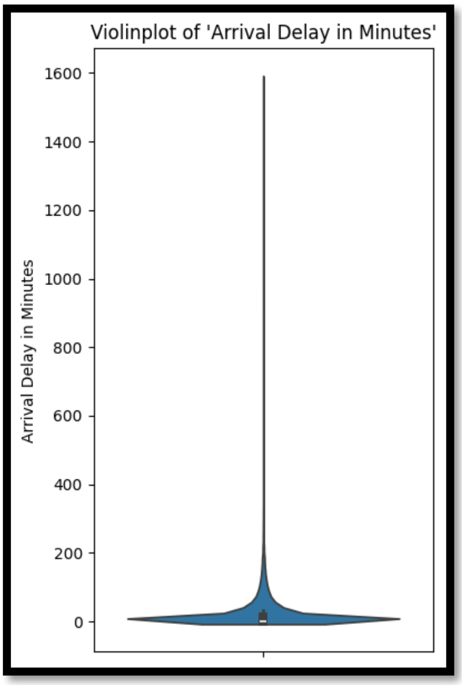


Figure 34 Violin Plot of Arrival Delay in Minutes

The violin plot demonstrates the density of the Arrival Delay in Minutes feature, with the majority of the density concentrated at very low delay values. The long upper tail represents the rare but extreme delays, further emphasizing the skewness of the data.

The Arrival Delay in Minutes feature is a continuous numerical variable with a heavily right-skewed distribution due to extreme delays. It has a negligible negative correlation with satisfaction (-0.058), indicating that longer arrival delays slightly reduce passenger satisfaction. Hypothesis testing using Pearson Correlation confirms a statistically significant relationship between arrival delay and satisfaction (p-value < 0.05), supporting the alternative hypothesis (H_1) that arrival delays impact satisfaction, albeit weakly. However, there is no meaningful correlation with other features like inflight service or cleanliness, supporting the null hypothesis (H_0) that arrival delays are independent of these variables. The skewness and extreme outliers in this feature necessitate transformations, such as logarithmic scaling, for better suitability in statistical modelling and analysis.

25) log_ArrivalDelayMinutes

The logarithmically transformed log_ArrivalDelayMinutes feature provides a normalized representation of the original skewed arrival delay data, compressing extreme values while maintaining interpretability. This feature retains its relevance in modeling, offering better scale adjustments for downstream processes.

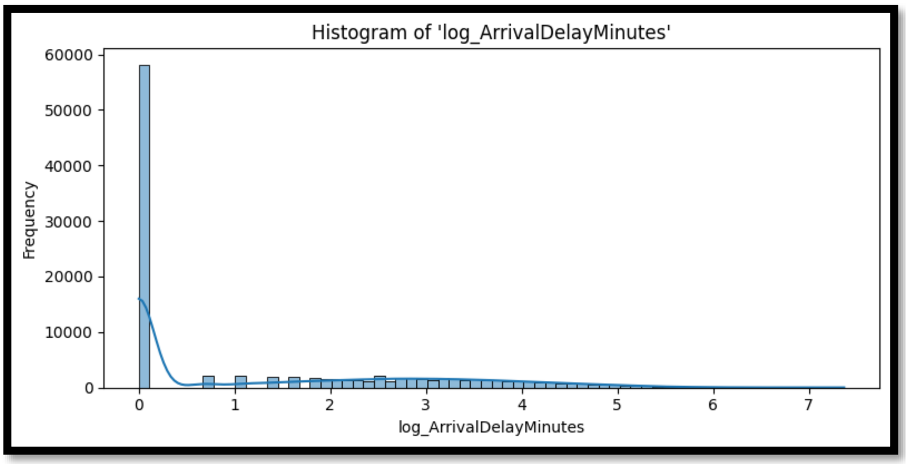


Figure 35 Histogram of log_ArrivalDelayMinutes

The histogram of log_ArrivalDelayMinutes illustrates a significant concentration of values near zero, reflecting the dominance of minimal arrival delays in the dataset. The long right tail, although reduced by transformation, indicates occasional significant delays.

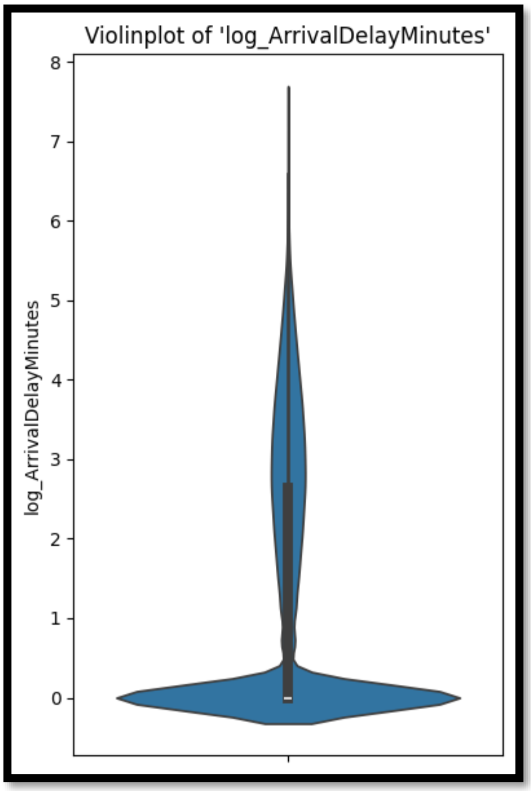


Figure 36 Violinplot of log_ArrivalDelayMinutes

The violin plot demonstrates a dense clustering of data points around lower delay values, reflecting the prevalence of flights with negligible delays. The extended tail captures less frequent, extreme delays.

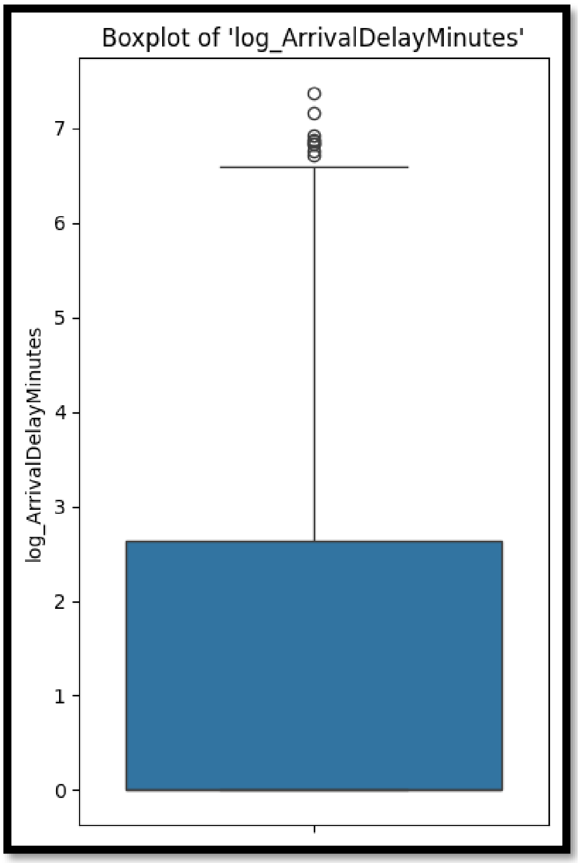


Figure 37 Boxplot of log_ArrivalDelayMinutes

The box plot of log_ArrivalDelayMinutes depicts a wide interquartile range, highlighting a diverse range of delay values. Outliers, visible beyond the upper whisker, represent rare but significant delays, which have been reduced but are still visible post-transformation.

The log_ArrivalDelayMinutes feature, derived from logarithmic transformation, reduces skewness and compresses extreme values of arrival delays, making it more suitable for analysis. It exhibits a weak negative correlation with satisfaction (-0.1), suggesting that longer delays moderately decrease passenger satisfaction. Hypothesis testing confirms a statistically significant relationship (p-value < 0.05), supporting the alternative hypothesis (H_1) that arrival delays impact satisfaction. However, the correlation strength is minimal. The transformation ensures the feature is better scaled for modeling, minimizing the undue influence of rare but extreme delays while retaining its predictive relevance.

Experiments

Problem 1: Classification of Passenger Satisfaction

Objective: To predict passenger satisfaction levels ("Satisfied" or "Dissatisfied") using supervised machine learning techniques. The goal is to evaluate baseline models, create an ensemble classifier, and optimize performance through hyperparameter tuning.

Experiment 1: Baseline Models (XGBoost, LightGBM, CatBoost)

- Model Selection: Three gradient-boosting models—XGBoost, LightGBM, and CatBoost—were chosen for their proven performance in classification tasks.
- Training Strategy:
- K-Fold Cross-Validation (K=5) was used to train each model, ensuring robustness and reducing the risk of overfitting.
- Performance metrics such as accuracy, precision, recall, confusion matrix and F1-score were evaluated on the test dataset.

Experiment 2: Custom Ensemble Architecture

- Model Selection: The three baseline models (XGBoost, LightGBM, CatBoost) were combined into an ensemble using stacking and voting techniques.
- Stacking and Voting:
    - Soft Voting: Combines models' predictions probabilistically, weighting predictions by confidence.
    - Hard Voting: Uses majority voting to decide the final class label.

Experiment 3: Hyperparameter Tuning

- Optimization Framework: The Optuna framework was utilized for efficient hyperparameter tuning.
- Target Model: The Hard Voting Classifier, which outperformed in Experiment 2, was selected for optimization.
- Tuning Strategy:
    - The hyperparameters for **XGBoost**, **LightGBM**, and **CatBoost** were carefully fine-tuned to optimize their individual performance within the ensemble.
    - The tuning process aimed to maximize the **validation F1-score**, ensuring the best balance between precision and recall, particularly important for imbalanced datasets.

Problem 2: Predicting Arrival Delay

Objective: To predict the arrival delay in minutes using regression techniques. The aim is to develop and evaluate baseline regression models and custom ensemble approaches to improve prediction accuracy.

Experiment 1: Baseline Models (XGBoost, LightGBM, CatBoost)

- Model Selection: XGBoost, LightGBM, and CatBoost were chosen as the baseline regression models due to their proven performance in handling regression tasks with tabular data.
- Training Strategy: K-Fold Cross-Validation (K=5) was employed to train each model, ensuring robust performance evaluation and reducing the risk of overfitting.

Experiment 2: Custom Ensemble Architecture

- Two custom **Voting Regressors** were created to combine the strengths of individual models:
    - **VotingReg_XGB_LGBM_CatB**: Combines all three models.
    - **VotingReg_LGBM_CatB**: Combines LightGBM and CatBoost for a lighter ensemble.

Problem 3: Clustering Passengers Based on Preferences

Objective: To segment passengers into distinct clusters based on their preferences and behaviors, enabling airlines to better understand customer groups and personalize services effectively. The aim is to employ unsupervised clustering techniques to uncover hidden patterns and group passengers into meaningful categories.

Experiment 1: K-Means Clustering with PCA

- **Feature Selection**: Features relevant to passenger preferences were retained, while irrelevant columns such as delay metrics and satisfaction were dropped:
- **Dimensionality Reduction**: **Principal Component Analysis (PCA)** was applied to reduce the dimensionality of the dataset while preserving 90% of the variance. The optimal number of PCA components (`p_opt`) was determined. The optimal number of components to capture 90% variance was found to be 14.
- **Clustering with K-Means**: The PCA-transformed data was clustered using K-Means. The optimal number of clusters (k) was selected based on the **silhouette score**:

Experiment 2: DBSCAN (Density-Based Clustering)

- DBSCAN was employed to identify dense regions in the data and detect outliers.
- Parameters used:
  - eps = 0.5: Maximum distance for points to be considered neighbors.
  - min_samples = 5: Minimum number of points to form a cluster.
- Evaluation Metrics:
  - Three clustering metrics were used to evaluate the quality of DBSCAN clusters:
    - **Silhouette Score**
    - **Calinski-Harabasz Score**
    - **Davies-Bouldin Score**

Results and Discussion

Problem 1: Classification of Passenger Satisfaction

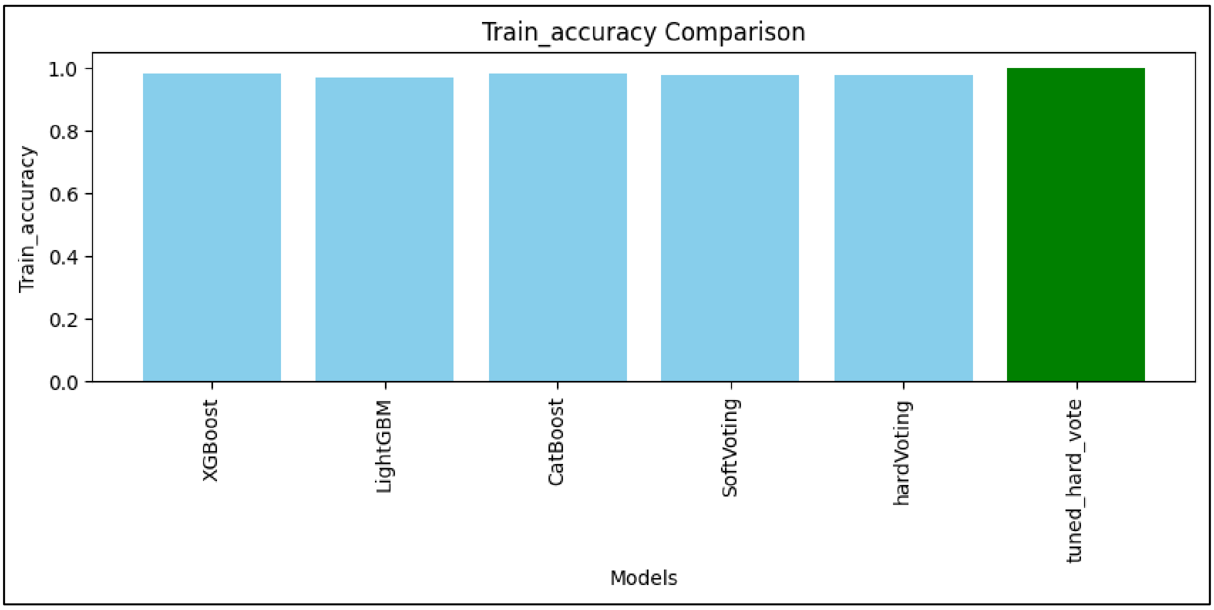


Figure 38 Problem 1 train accuracy model comparison

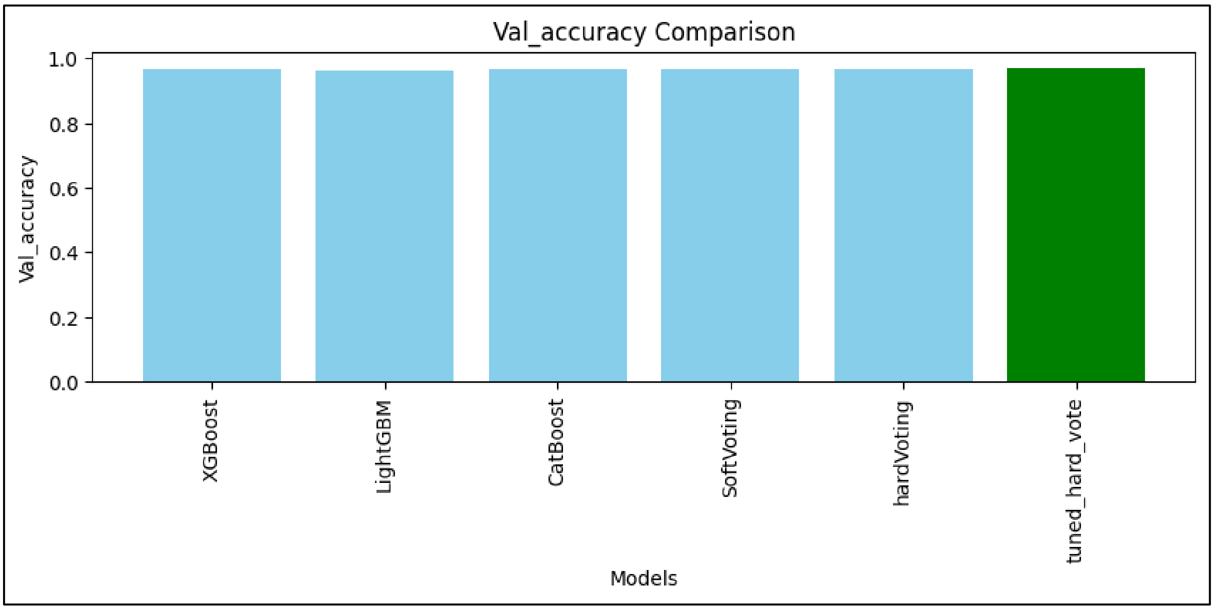


Figure 39 Problem 1 validation accuracy model comparison

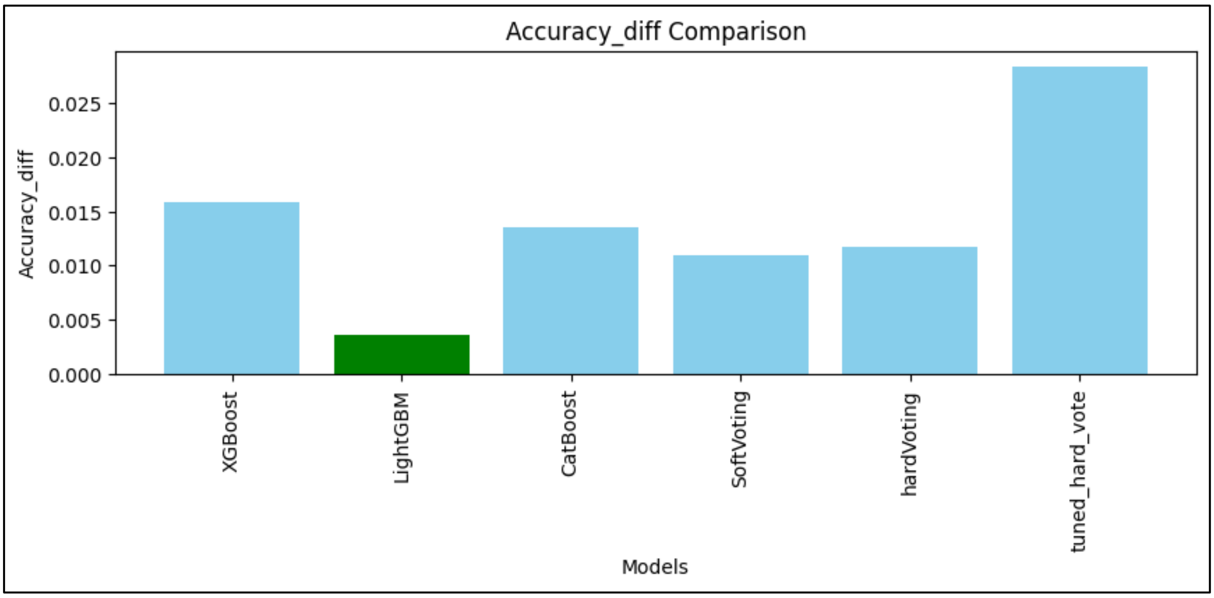


Figure 40 Problem 1 accuracy difference model comparison

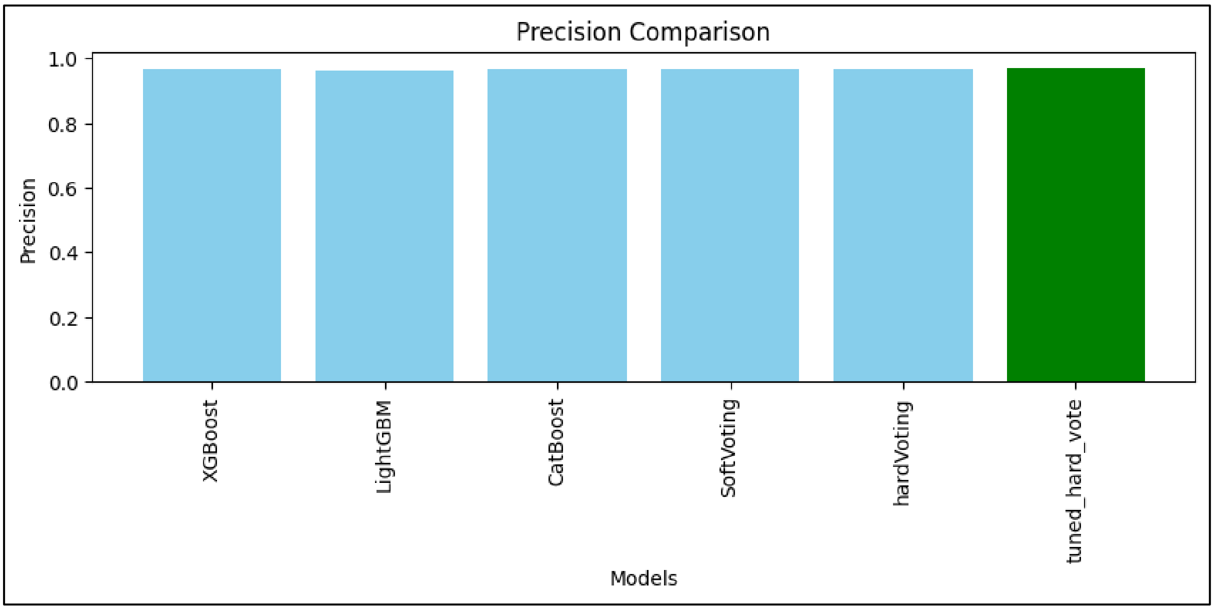


Figure 41 Problem 1 precision model comparison

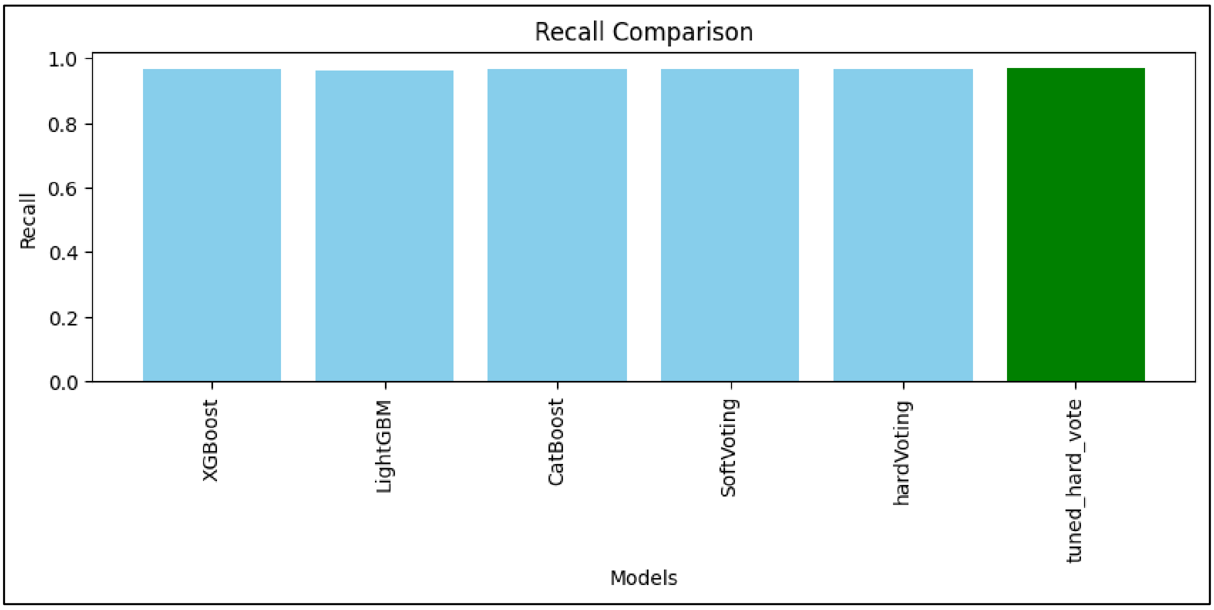


Figure 42 Problem 1 recall model comparison

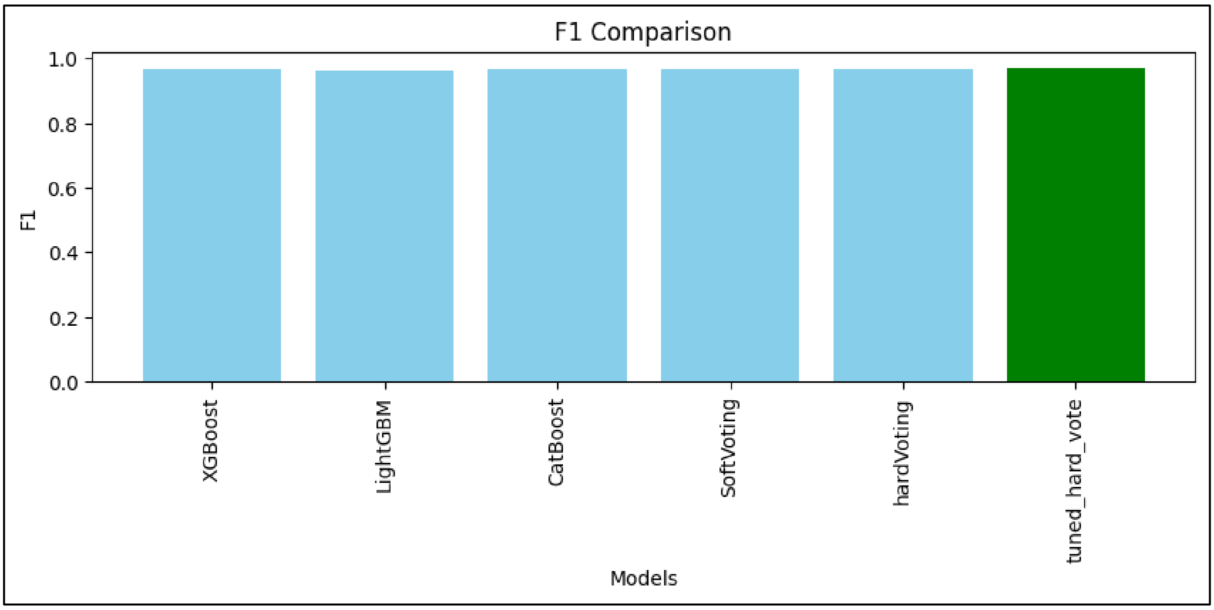


Figure 43 Problem 1 f1 score model comparison

Model Comparison Analysis

The comparison between the models for **Classification of Passenger Satisfaction** reveals several key performance aspects across accuracy, precision, recall, and F1-score.

1. **Train Accuracy:**
   - All models demonstrate high training accuracy, indicating that they effectively learn from the dataset. However, the **Tuned Hard Voting Ensemble** model marginally outperforms others, achieving near-perfect training accuracy, highlighting its strong learning capacity.
2. **Validation Accuracy:**
   - Validation accuracy across all models remains consistently high, reflecting effective generalization to unseen data. The **Tuned Hard Voting Ensemble** again leads slightly, showcasing the benefits of hyperparameter optimization.
3. **Accuracy Difference:**
   - LightGBM shows the lowest difference between training and validation accuracy, suggesting it has minimal overfitting compared to other models. However, the **Tuned Hard Voting Ensemble**, despite its high performance, exhibits a noticeable gap, indicating potential overfitting that could be mitigated with further tuning.

4. **Precision, Recall, and F1-Score:**

   ○ Across these metrics, all models maintain high values, with minimal variation. The **Tuned Hard Voting Ensemble** slightly surpasses the others, achieving the best F1-score, emphasizing its ability to balance precision and recall effectively.

Summary

While all models deliver strong results, the **Tuned Hard Voting Ensemble** consistently outperforms in terms of accuracy and F1-score, demonstrating its superior performance for this classification task. However, LightGBM exhibits the most balanced behavior between training and validation accuracy, making it a robust alternative if avoiding overfitting is a priority. This comparison underscores the trade-offs between complexity and generalization in machine learning models.

Problem 2: Predicting Arrival Delay

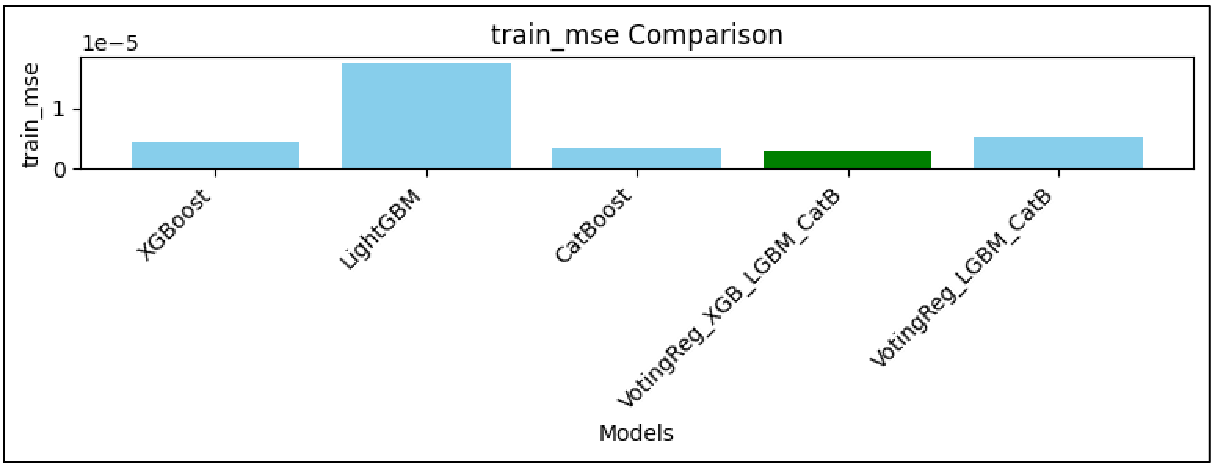


Figure 44 Problem 2 train mse model comparison

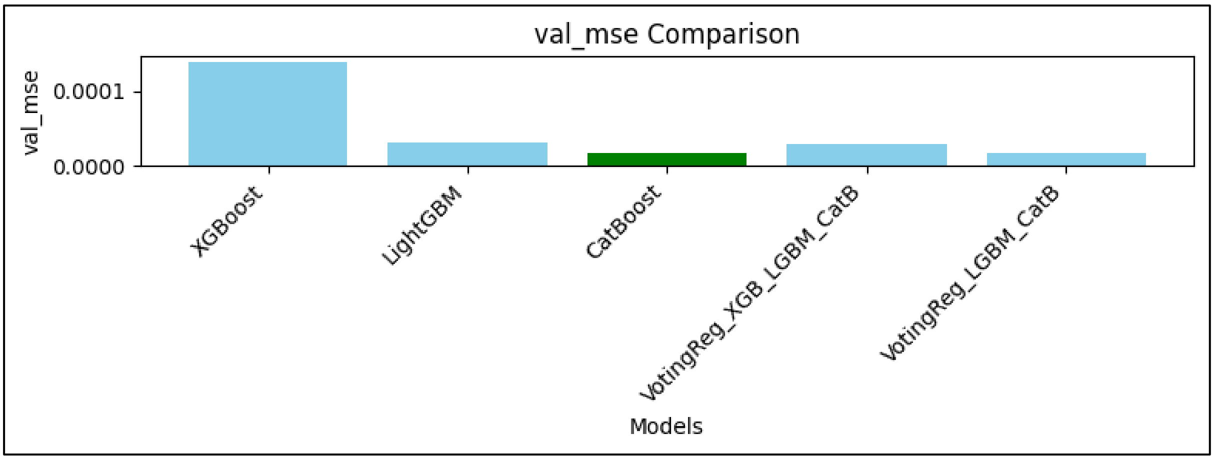


Figure 45 Problem 2 validation mse model comparison

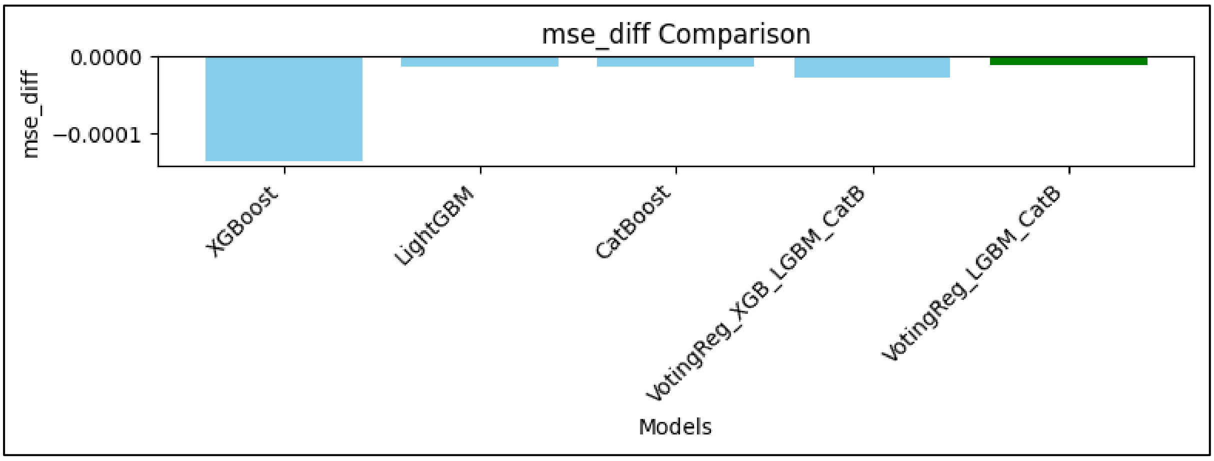


Figure 46 Problem 2 train mse difference model comparison

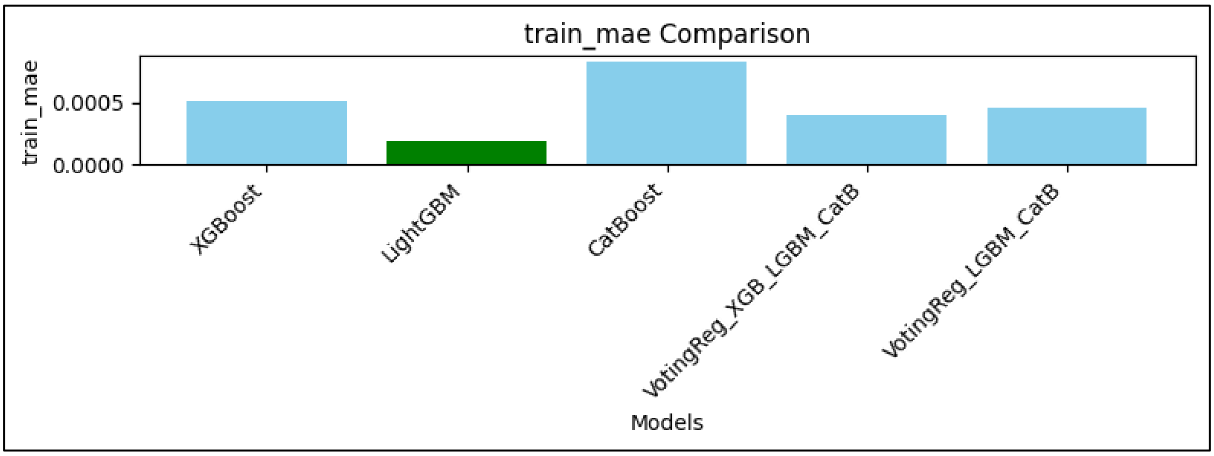


Figure 47 Problem 2 train mae model comparison

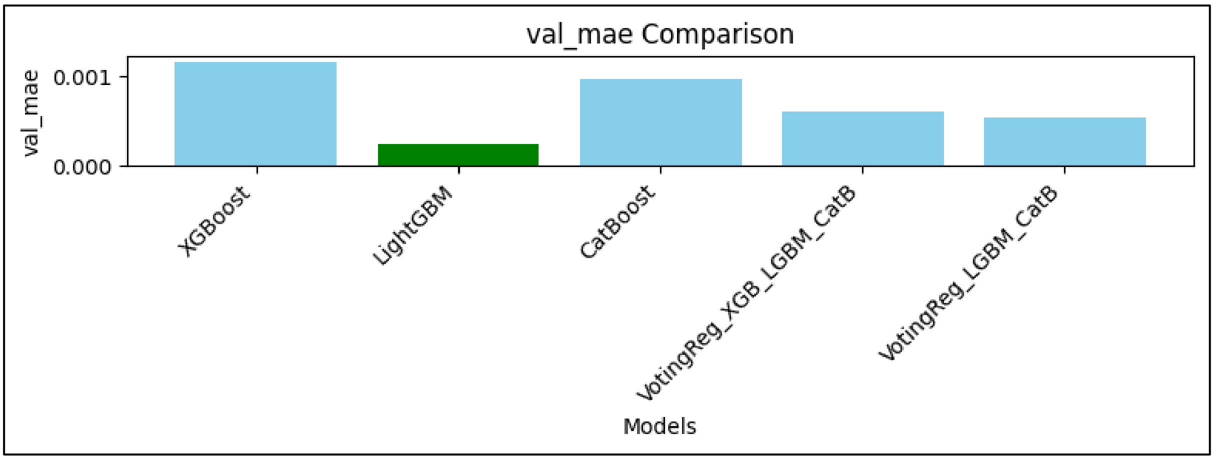


Figure 48 Problem 2 validation mse model comparison

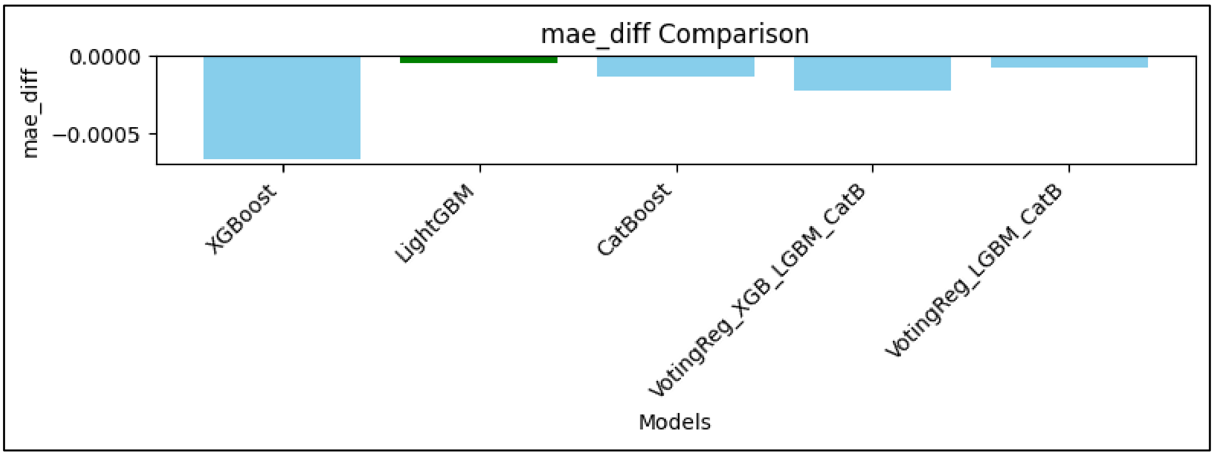


Figure 49 Problem 2 train mae difference model comparison

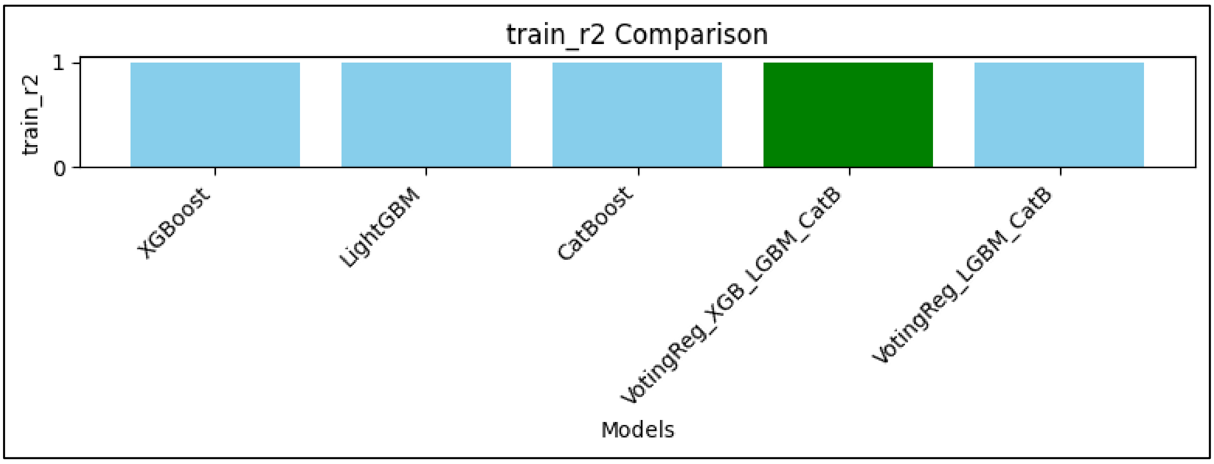


Figure 50 Problem 2 train r2 model comparison

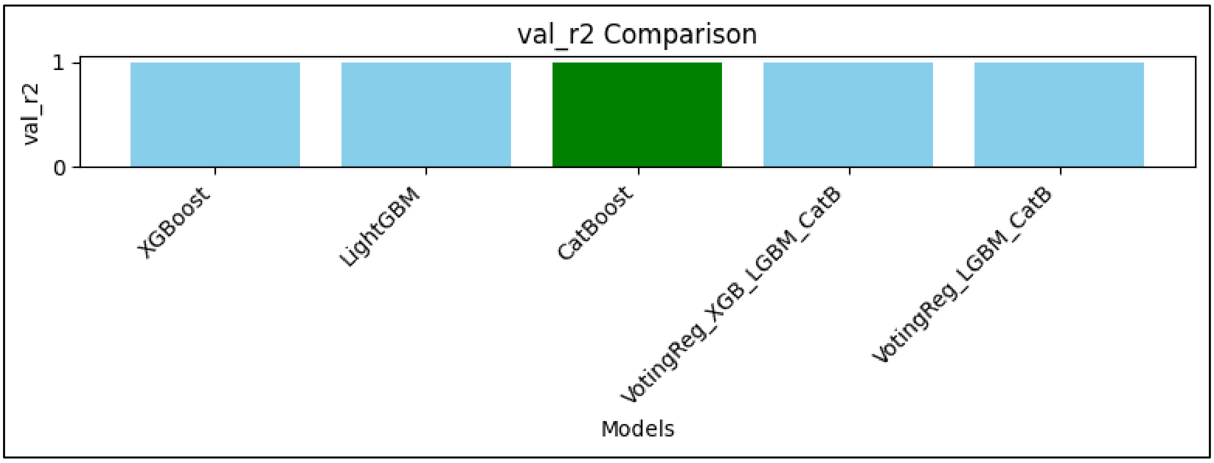


Figure 51 Problem 2 validation r2 model comparison

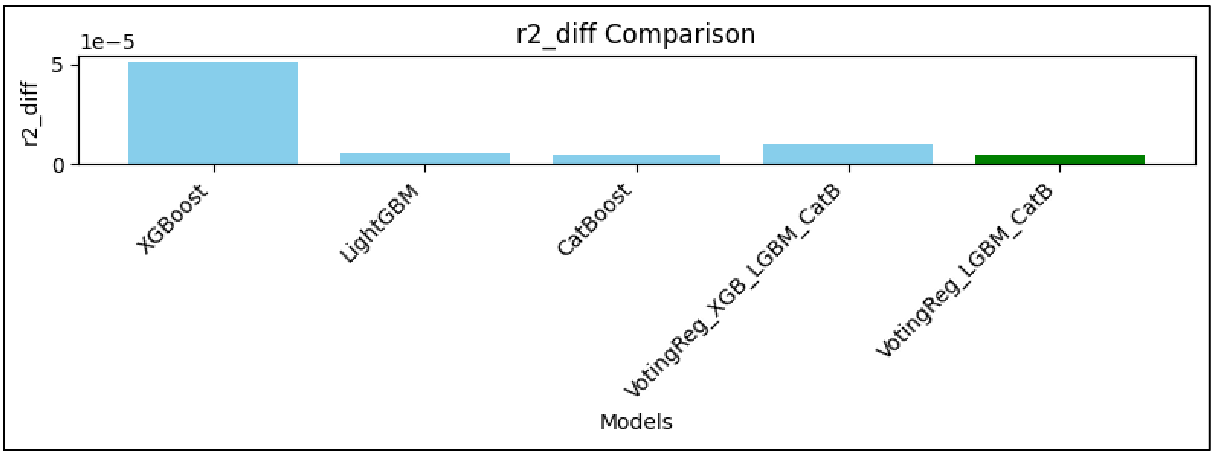


Figure 52 Problem 2 r2 difference model comparison

Model Comparison

The evaluation of models for predicting arrival delays involves comparing performance metrics such as MSE (Mean Squared Error), MAE (Mean Absolute Error), and R2R^2R2 scores across different regression algorithms. Below is a detailed analysis:

1. Train MSE Comparison:
   - CatBoost and the Voting Regressors (both XGB-LGBM-CatBoost and LGBM-CatBoost combinations) achieve significantly lower train MSE compared to LightGBM and XGBoost.
   - LightGBM shows the highest train MSE, indicating a weaker fit to the training data.
2. Validation MSE Comparison:

- CatBoost outperforms other models by achieving the lowest validation MSE, reflecting its superior generalization to unseen data.
- The Voting Regressor models also exhibit competitive validation MSE, closely following CatBoost.
- XGBoost performs the poorest in validation MSE, suggesting potential issues with overfitting.

3. MSE Difference Comparison:
   - The MSE difference is minimal for CatBoost and the Voting Regressors, indicating strong consistency between train and validation performance.
   - XGBoost shows the largest MSE difference, highlighting overfitting to the training data.
4. Train MAE Comparison:
   - Similar to MSE, CatBoost and Voting Regressors achieve lower train MAE, indicating smaller average errors on the training dataset.
   - LightGBM again lags behind in performance with higher train MAE.
5. Validation MAE Comparison:
   - CatBoost achieves the best validation MAE, followed closely by the Voting Regressors. This demonstrates their capability in maintaining low prediction errors on the validation set.
   - XGBoost exhibits the highest validation MAE, consistent with its weaker validation performance in MSE.
6. MAE Difference Comparison:
   - CatBoost and Voting Regressors show minimal MAE differences, reinforcing their robustness in avoiding overfitting.
   - XGBoost displays the highest MAE difference, consistent with its overfitting behavior.
7. Train R2R^2R2 Comparison:

  - All models exhibit near-perfect train R2R^2R2 scores, indicating a strong ability to explain variance in the training data.
  - The Voting Regressors demonstrate slightly better train R2R^2R2 than individual models.

8. Validation R2R^2R2 Comparison:
    - CatBoost achieves the highest validation R2R^2R2, signifying its strong generalization capabilities.
    - Voting Regressors also perform well, with competitive R2R^2R2 scores.
    - XGBoost falls behind, showing lower R2R^2R2 scores, indicating less explained variance in the validation dataset.
9. R2R^2R2 Difference Comparison:
    - The R2R^2R2 differences are smallest for CatBoost and Voting Regressors, confirming their balance between training and validation performance.
    - XGBoost exhibits the largest R2R^2R2 difference, highlighting potential overfitting issues.

Summary:

CatBoost emerges as the best-performing model, achieving the lowest MSE and MAE and the highest R2R^2R2 on validation data. The Voting Regressors also deliver strong results, with minimal overfitting and competitive performance metrics. XGBoost demonstrates signs of overfitting and underwhelming generalization, while LightGBM performs reasonably well but does not match CatBoost's consistency across all metrics. This analysis underscores CatBoost's robustness for predicting arrival delays, with Voting Regressors offering an effective ensemble alternative.

Problem 3: Clustering Passengers Based on Preferences

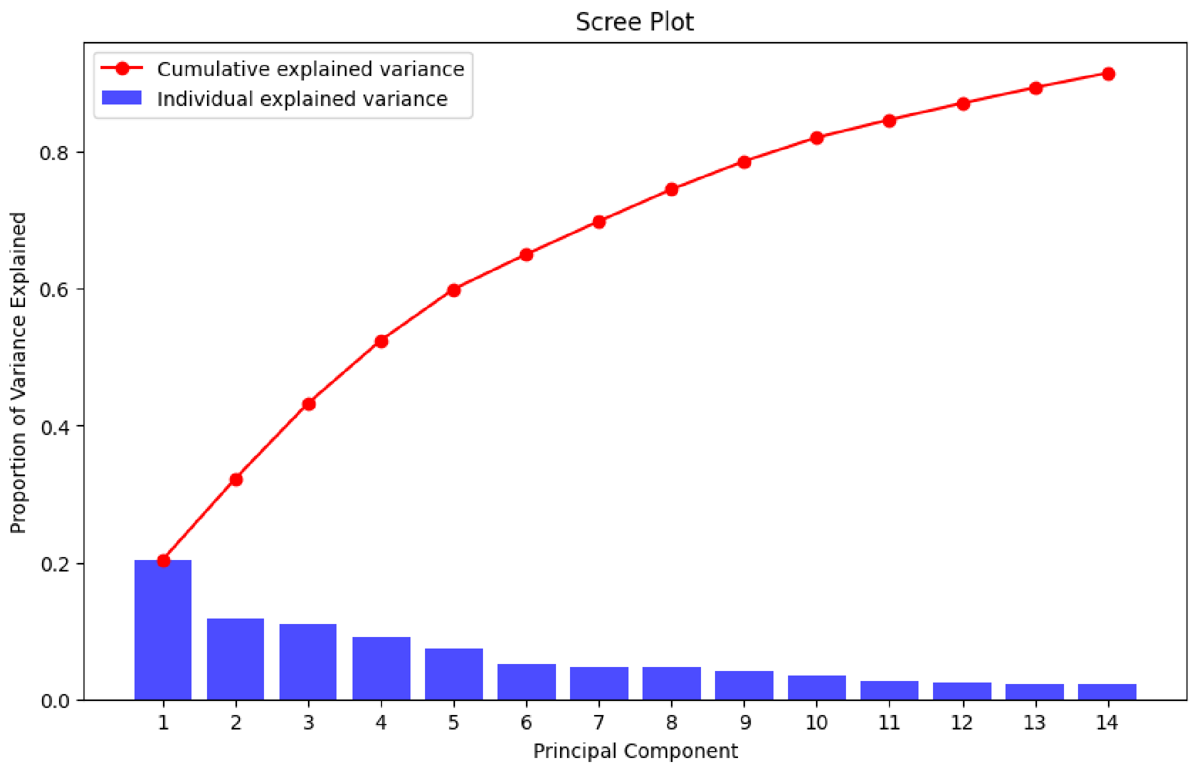


Figure 53 Scree Plot

Principal Component Analysis (PCA):

The scree plot indicates the cumulative explained variance across the principal components. PCA was applied to reduce dimensionality while retaining 90% of the variance. This step optimized clustering performance by eliminating redundant features and focusing on the most informative components.

DBSCAN Clustering:

The DBSCAN clustering algorithm performed exceptionally well, as evident from the following metrics:

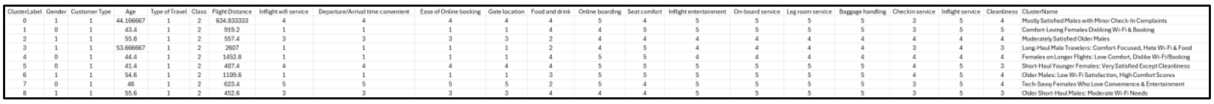
- **Silhouette Score**: 0.82388, indicating well-defined and distinct clusters.
- **Calinski-Harabasz Score**: 442.87038, confirming a good balance between cluster cohesion and separation.
- **Davies-Bouldin Score**: 0.22405, reflecting the compactness and separation of the clusters.

Cluster Distribution: After outlier removal, DBSCAN identified nine meaningful clusters:

- Cluster sizes ranged from 5 to 6 passengers per cluster.

- Outliers were effectively excluded, improving clustering precision.

Cluster Insights: Each cluster represents distinct passenger personas bas



ClusterLabel	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	ClusterName
0	1	1	44.166667	1	2	634.833333	4	4	4	4	4	5	4	5	5	5	5	3	5	4	Mostly Satisfied Males with Minor Check-In Complaints
1	0	1	43.4	1	2	919.2	1	1	1	1	4	4	5	5	5	5	5	3	5	5	Comfort-Loving Females Disliking Wi-Fi & Booking
2	1	1	55.8	1	2	557.4	3	3	3	3	2	4	4	4	4	4	4	4	4	4	Moderately Satisfied Older Males
3	1	1	53.666667	1	2	2607	1	1	1	1	2	4	5	4	4	4	4	3	4	3	Long-Haul Male Travelers: Comfort-Focused, Hate Wi-Fi & Food
4	0	1	44.4	1	2	1452.8	1	1	1	1	4	5	5	4	4	4	4	4	4	4	Females on Longer Flights: Love Comfort, Dislike Wi-Fi/Booking
5	0	1	41.4	1	2	407.4	4	4	4	4	4	5	5	5	5	5	5	4	5	3	Short-Haul Younger Females: Very Satisfied Except Cleanliness
6	1	1	54.6	1	2	1109.6	1	1	1	1	3	5	5	5	5	5	5	4	5	4	Older Males: Low Wi-Fi Satisfaction, High Comfort Scores
7	0	1	46	1	2	623.4	5	5	5	5	2	5	4	5	5	5	5	3	5	4	Tech-Savvy Females Who Love Convenience & Entertainment
8	1	1	55.6	1	2	452.6	3	3	3	3	4	4	4	5	5	5	5	3	5	3	Older Short-Haul Males: Moderate Wi-Fi Needs

Figure 54 Cluster summary

Cluster Profiles

Below is a brief, high-level description of each cluster. (Both tables presented are identical in numeric content, but they confirm and consolidate the final cluster assignments and labels.)

Cluster 0

Label: Mostly Satisfied Males with Minor Check-In Complaints

- **Demographics:** Male, ~44 years old, loyal customers
- **Key Satisfaction Scores:** High ratings (4–5) across inflight wifi, online boarding, seat comfort, inflight entertainment, and on-board service.
- **Pain Point:** Slightly lower satisfaction with check-in service (3/5).

Interpretation & Recommendation: This cluster is generally happy with the inflight experience but could benefit from improving or streamlining check-in processes to bolster overall satisfaction.

Cluster 1

Label: Comfort-Loving Females Disliking Wi-Fi & Booking

- **Demographics:** Female, ~43 years old, loyal customers
- **Key Satisfaction Scores:** Very high (5/5) for comfort-related aspects (seat, on-board service, inflight entertainment).
- **Pain Point:** Low ratings (1/5) for inflight wifi service and departure/arrival time convenience/ease of online booking.

Interpretation & Recommendation: This group prioritizes comfort but has strong dissatisfaction with wifi availability and booking convenience. Enhancing the user-friendliness of booking tools and improving wifi could markedly improve their satisfaction.

Cluster 2

Label: Moderately Satisfied Older Males

- **Demographics:** Male, ~56 years old, loyal customers
- **Key Satisfaction Scores:** Mid-level (3–4/5) ratings for inflight wifi, departure/arrival time convenience, ease of online booking, seat comfort, etc.
- **Pain Point:** Food and drink satisfaction is only 2/5, indicating a possible area for improvement.

Interpretation & Recommendation: Though not entirely dissatisfied, these passengers have some room for higher satisfaction. Improving meal options or addressing specific dietary concerns may resonate well with them.

Cluster 3

Label: Long-Haul Male Travelers: Comfort-Focused, Hate Wi-Fi & Food

- **Demographics:** Male, ~54 years old, loyal customers
- **Travel Pattern:** Very high average flight distance (2607 miles)
- **Key Satisfaction Scores:** Strong seat comfort (5/5) and decent on-board service (4/5).
- **Pain Points:** Very low (1/5) for inflight wifi and departure/arrival convenience, and poor food and drink ratings (2/5).

Interpretation & Recommendation: These passengers endure longer flights, so negative wifi and meal experiences have a disproportionate impact. Improving long-haul wifi reliability and meal variety will likely enhance their experience.

Cluster 4

Label: Females on Longer Flights: Love Comfort, Dislike Wi-Fi/Booking

- **Demographics:** Female, ~44 years old, loyal customers
- **Travel Pattern:** Long flight distances (1452.8 miles)
- **Key Satisfaction Scores:** Very high for seat comfort, online boarding, on-board service, legroom, baggage handling.
- **Pain Point:** They rate wifi and online booking convenience at 1/5.

Interpretation & Recommendation: Like Cluster 3, these travelers appreciate comfort but are dissatisfied with technology-related aspects (wifi, booking). Focusing on improved booking tools and robust inflight connectivity for longer flights is crucial.

Cluster 5

Label: Short-Haul Younger Females: Very Satisfied Except Cleanliness

- **Demographics:** Female, ~41 years old, loyal customers
- **Travel Pattern:** Shorter flight distance (407.4 miles)
- **Key Satisfaction Scores:** Very high (4–5/5) across most aspects, including wifi, comfort, convenience, entertainment.
- **Pain Point:** Cleanliness is rated 3/5.

Interpretation & Recommendation: Although generally content, their perceived cleanliness is slightly lower. Minor tweaks to cabin cleaning procedures or restocking amenities could elevate this group's satisfaction further.

Cluster 6

Label: Older Males: Low Wi-Fi Satisfaction, High Comfort Scores

- **Demographics:** Male, ~55 years old, loyal customers
- **Travel Pattern:** Medium-to-long flights (1109.6 miles)
- **Key Satisfaction Scores:** Consistently strong (5/5) for seat comfort, inflight entertainment, on-board service.
- **Pain Point:** Wifi remains notably low at 1/5.

Interpretation & Recommendation: These passengers value a comfortable seat and good entertainment but have poor wifi experiences. Consider upgrades to inflight connectivity to boost loyalty among these older travelers.

Cluster 7

Label: Tech-Savvy Females Who Love Convenience & Entertainment

- **Demographics:** Female, ~46 years old, loyal customers
- **Travel Pattern:** Mid-range flight distance (623.4 miles)

- **Key Satisfaction Scores:** Extremely high (5/5) for wifi, departure/arrival convenience, seat comfort, inflight entertainment, etc.
- **Pain Point:** Food and drink is only 2/5, and checkin service is moderate at 3/5.

Interpretation & Recommendation: As the “tech-savvy” label suggests, they place a premium on wifi and convenience. Addressing food service quality and possibly making check-in smoother should be top priorities for this cluster.

Cluster 8

Label: Older Short-Haul Males: Moderate Wi-Fi Needs

- **Demographics:** Male, ~56 years old, loyal customers
- **Travel Pattern:** Relatively short distances (452.6 miles)
- **Key Satisfaction Scores:** Mid-level (3–4/5) for wifi, convenience, online booking. Good (4–5/5) for seat comfort, on-board service, etc.
- **Pain Point:** Cleanliness is 3/5, indicating it could be improved.

Interpretation & Recommendation: Older travelers on short flights with moderate wifi usage and mid-range satisfaction all around. Targeted cleanliness upgrades or attention to minor service details could help elevate this group’s perception.

Comparison with K-Means:

K-Means clustering achieved significantly lower performance:

- **Silhouette Score**: 0.1493, indicating weak and overlapping clusters.
- **Calinski-Harabasz Score**: 19,331.7, suggesting inadequate cluster cohesion.
- **Davies-Bouldin Score**: 2.2729, highlighting poor separation and compactness.

Conclusion: DBSCAN outperformed K-Means in identifying meaningful clusters, providing actionable insights into passenger preferences. The analysis emphasizes DBSCAN's effectiveness in capturing nuanced relationships within the dataset, making it a robust choice for passenger segmentation.

Conclusion

This project successfully utilized machine learning techniques to address three key problems in the airline industry: classification of passenger satisfaction, prediction of arrival delays, and clustering passengers based on preferences.

1. **Passenger Satisfaction Classification**:
   - Supervised learning models such as XGBoost, LightGBM, and CatBoost were evaluated, with the **Tuned Hard Voting Ensemble** achieving the best performance in terms of accuracy and F1-score.
   - The analysis revealed that features like online boarding, travel class, and inflight entertainment significantly influence passenger satisfaction.
2. **Arrival Delay Prediction**:
   - Regression models, including XGBoost, LightGBM, CatBoost, and their ensemble combinations, were employed. **CatBoost** demonstrated the best performance, achieving low MSE and high R2R^2R2 scores.
   - This study highlighted the challenge of overfitting in regression tasks and provided insights into delay factors.
3. **Passenger Segmentation (Clustering)**:
   - DBSCAN outperformed K-Means, achieving a **Silhouette Score of 0.82388**. It identified nine meaningful clusters, each representing distinct passenger personas based on preferences and satisfaction levels.
   - The clustering provided actionable insights for personalized services, addressing specific pain points such as Wi-Fi quality, cleanliness, and food service.

Limitations

1. **Dataset Limitations**:
   - The dataset primarily represents loyal customers, creating a potential bias in understanding the broader passenger base.

- Some features, such as departure and arrival delays, showed limited impact on satisfaction, indicating that additional features (e.g., weather conditions, airline policies) could enhance the analysis.

2. **Model Generalization**:
    - While the models achieved high performance on the given dataset, their generalization to other airlines or regions is untested.
    - Overfitting was observed in some cases (e.g., XGBoost for arrival delay prediction), requiring careful hyperparameter tuning.
3. **Clustering Challenges**:
    - The clustering results are dataset-specific and may not generalize well to different datasets without recalibration.
    - DBSCAN's performance depends on parameter selection (eps and min_samples), which might not scale across diverse datasets.
4. **Feature Engineering**:
    - Although extensive feature transformations were applied, additional derived features (e.g., customer loyalty metrics or real-time service feedback) could improve model performance.

Future Work

1. **Incorporating Additional Features**:
    - Integrate external data such as weather, airline policies, and regional economic conditions to enhance prediction and clustering accuracy.
    - Explore interaction terms and advanced feature engineering to capture more nuanced relationships.
2. **Model Optimization**:
    - Apply advanced optimization techniques, such as Neural Architecture Search (NAS) and ensemble deep learning, to further improve model performance.
    - Test deep learning models (e.g., RNNs for sequential data) for dynamic features like satisfaction over time.

3. **Time-Series and Real-Time Applications**:
   - Extend arrival delay prediction to a time-series framework for dynamic predictions.
   - Implement real-time clustering and satisfaction prediction systems for operational use in airlines.
4. **Cross-Domain Analysis**:
   - Expand the analysis to include multiple airlines and regions for a more generalized understanding of passenger preferences and operational challenges.
5. **Explainable AI (XAI)**:
   - Incorporate XAI techniques to make model predictions and clustering outcomes more interpretable for decision-makers.
6. **Advanced Clustering Techniques**:
   - Experiment with hierarchical clustering, Gaussian Mixture Models (GMM), or temporal clustering to capture seasonal variations in passenger behavior.
7. **Deployment and Scalability**:
   - Develop deployable solutions for real-world airline operations, such as personalized marketing strategies or operational decision-making tools.

This study demonstrates the potential of machine learning in addressing critical challenges in the airline industry, providing a foundation for future improvements in passenger satisfaction and operational efficiency.