DATA MINING FOR AIRLINE SATISFACTION

CONTENT

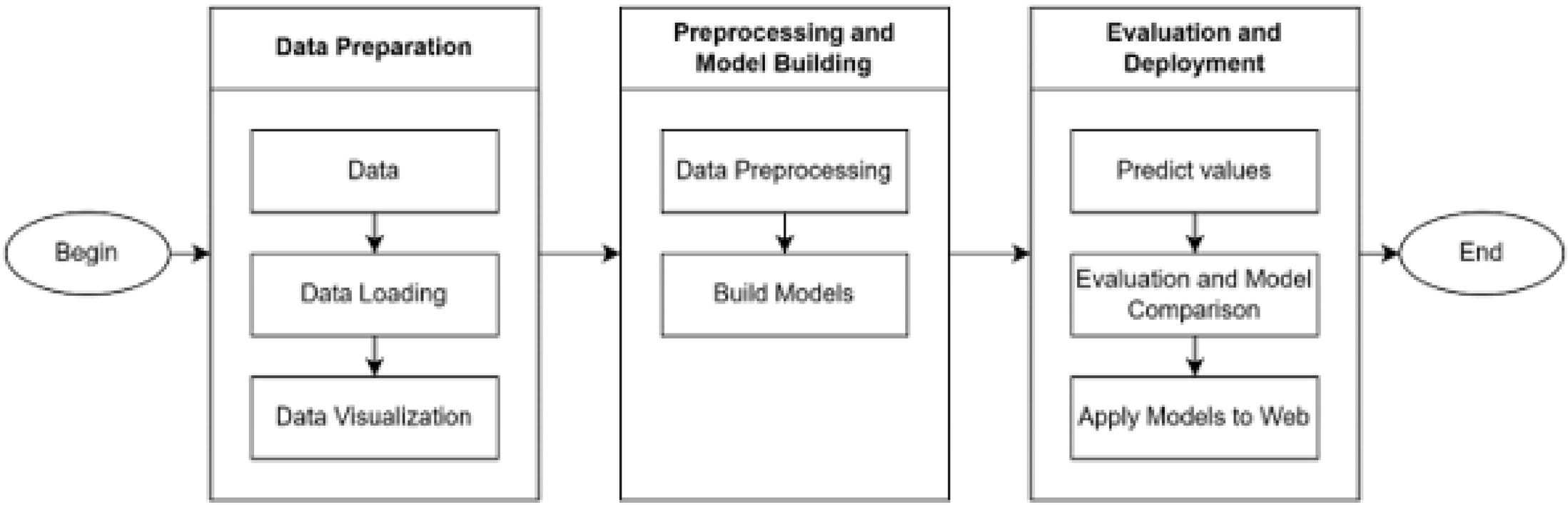
1. Overview
2. Data visualization
3. Preprocessing
4. Experimental
5. Evaluation
6. Conclusion

1. OVERVIEW

REASON

The rapid development of the aviation industry leads to great competition in this industry. Understanding customers, knowing whether they are satisfied or not, thereby improving quality is very necessary.

PROCESS


Begin
Data Preparation
Data
Data Loading
Data Visualization
Preprocessing and Model Building
Data Preprocessing
Build Models
Evaluation and Deployment
Predict values
Evaluation and Model Comparison
Apply Models to Web
End

DATA OVERVIEW

Column	Explain
#	Numerical order
id	Flight ID code
Gender	Customer's gender (Male, Female)
Customer Type	Customer type (Loyal customer, disloyal customer)
Age	Customer's age
Type of Travel	Purpose of the customer's flight (Personal Travel, Business Travel)
Class	Customer's ticket class (Business, Eco, Eco Plus)
Flight distance	Distance of the flight journey
Inflight wifi service	Satisfaction level with in-flight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient	Level of satisfaction with convenient departure/arrival time

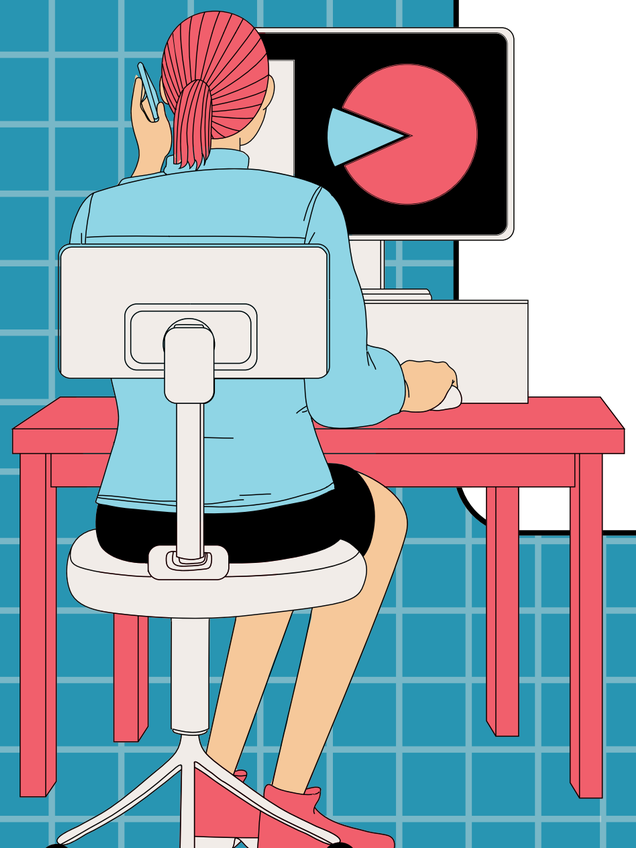
DATA OVERVIEW

Ease of Online booking	Level of satisfaction when booking tickets online
Gate location	Level of satisfaction with Gate location
Food and drink	Level of satisfaction with food and drinks
Online boarding	Satisfaction level with online check-in
Seat comfort	Level of satisfaction with seat comfort
Inflight entertainment	Level of satisfaction with in-flight entertainment
On-board service	Level of satisfaction with on-board service
Leg room service	Level of satisfaction with seats with wide legroom
Baggage handling	Level of satisfaction with baggage handling
Check-in service	Level of satisfaction with Check-in service
Inflight service	Level of satisfaction with in-flight service
Cleanliness	Level of satisfaction with cleanliness

DATA OVERVIEW

Departure Delay in Minutes	Departure minutes
Arrival Delay in Minutes	Number of minutes delayed upon arrival
Satisfaction	Customer satisfaction level with the airline (Satisfaction, neutral or dissatisfaction)

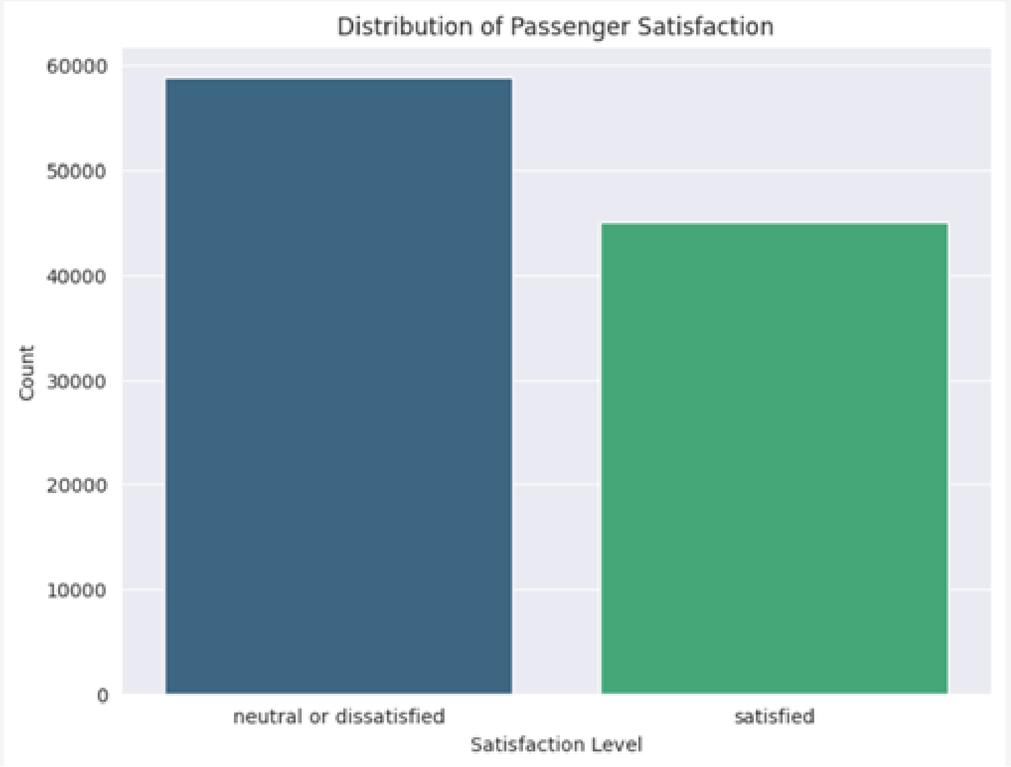
2. DATA VISUALIZATION

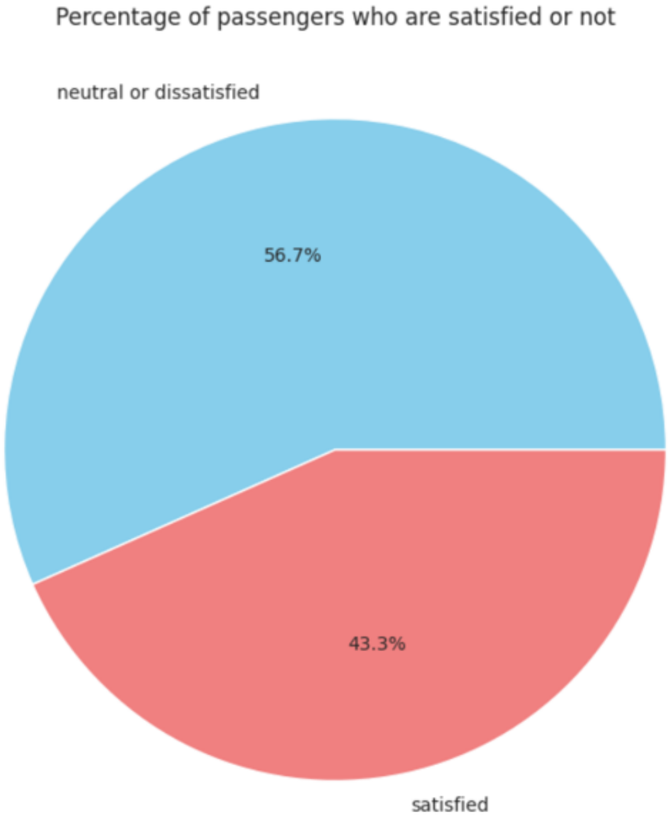


DATA VISUALIZATION

- Distribution of Customer Satisfaction
- Distribution of 'Gender' and its influence on satisfaction levels
- Distribution of 'Age' and its influence on satisfaction levels
- Sastifaction by Age and Gender 10
- Distribution of ‘Customer Type’ and its influence on satisfaction levels.
- Distribution of ‘Type of Travel’ and ‘Travel Class ’
- Distribution of Flight Distance
- Service quality analysis
- Correlation between delays (departure and arrival) and satisfaction levels

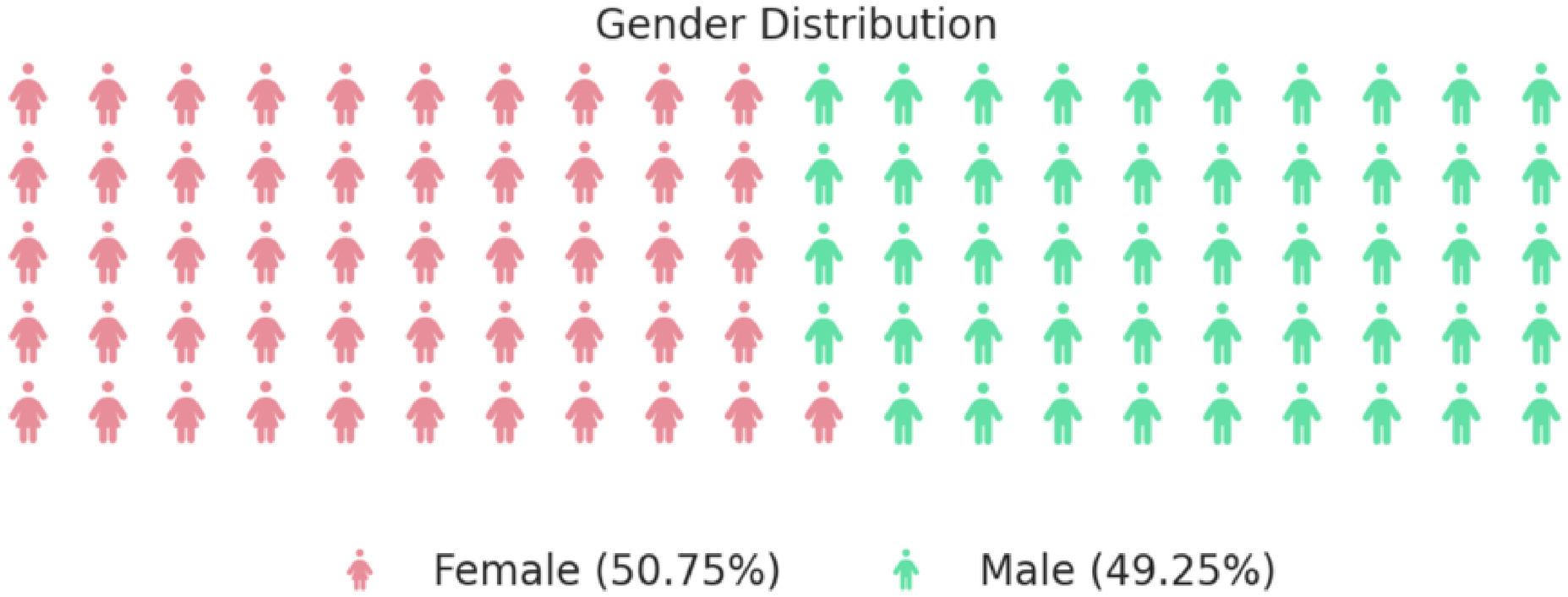
DISTRIBUTION OF CUSTOMER SATISFACTION





This analysis reveals that the number of dissatisfied passengers exceeds that of satisfied passengers by a noticeable margin, indicating an imbalance in our train data.

DISTRIBUTION OF 'GENDER' AND ITS INFLUENCE ON SATISFACTION LEVELS



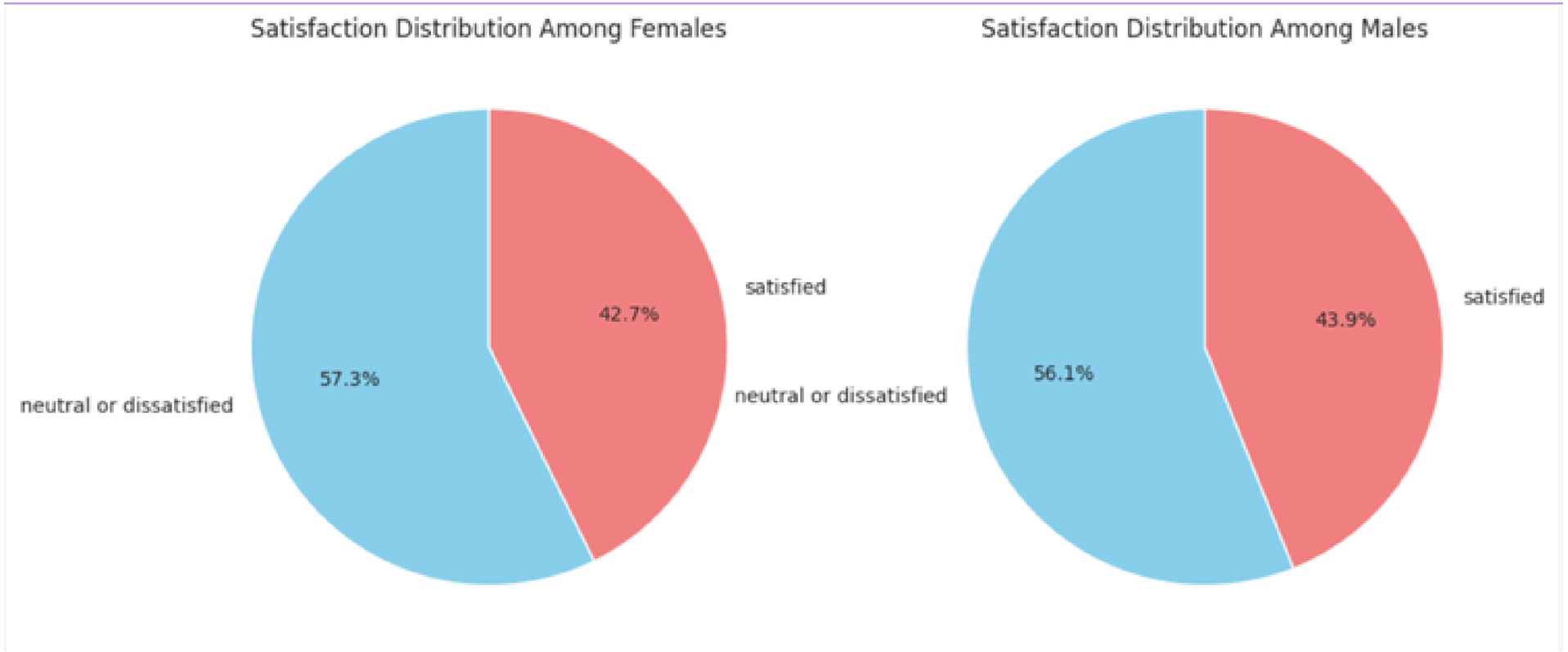
The distribution of gender among passengers is nearly balanced, with a relatively equal representation of male and female travelers.

DISTRIBUTION OF 'GENDER' AND ITS INFLUENCE ON SATISFACTION LEVELS

satisfaction	neutral or dissatisfied	satisfied
Gender		
Female	30193	22534
Male	28686	22491

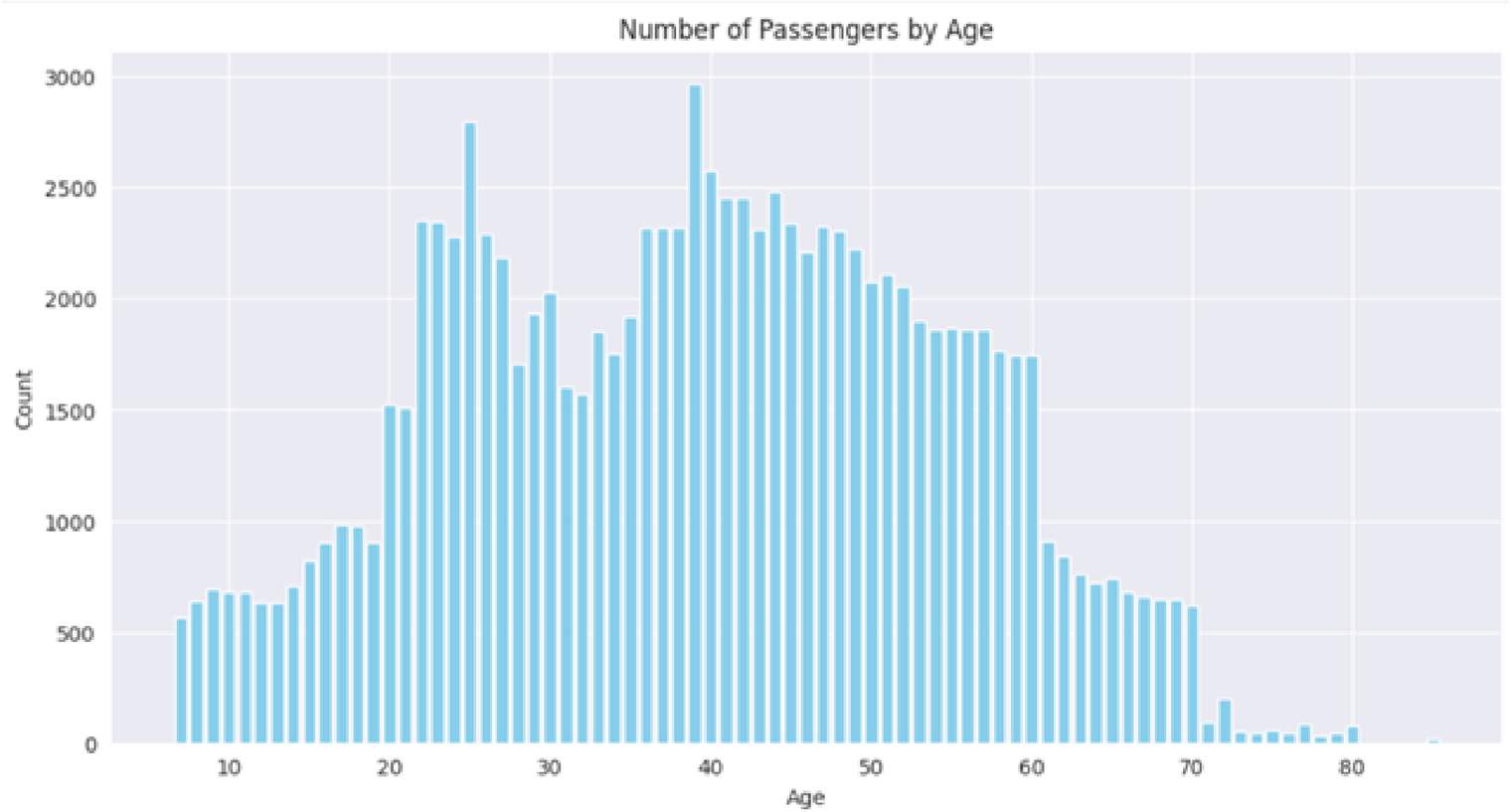
There are 22534 satisfied and 30193 dissatisfied female customer.
There are 22491 satisfied and 28868 dissatisfied male customer.

DISTRIBUTION OF 'GENDER' AND ITS INFLUENCE ON SATISFACTION LEVELS



In general, the data indicates an even distribution of satisfaction between genders, with a slightly higher percentage of satisfied individuals in both male and female groups.

DISTRIBUTION OF 'AGE' AND ITS INFLUENCE ON SATISFACTION LEVELS

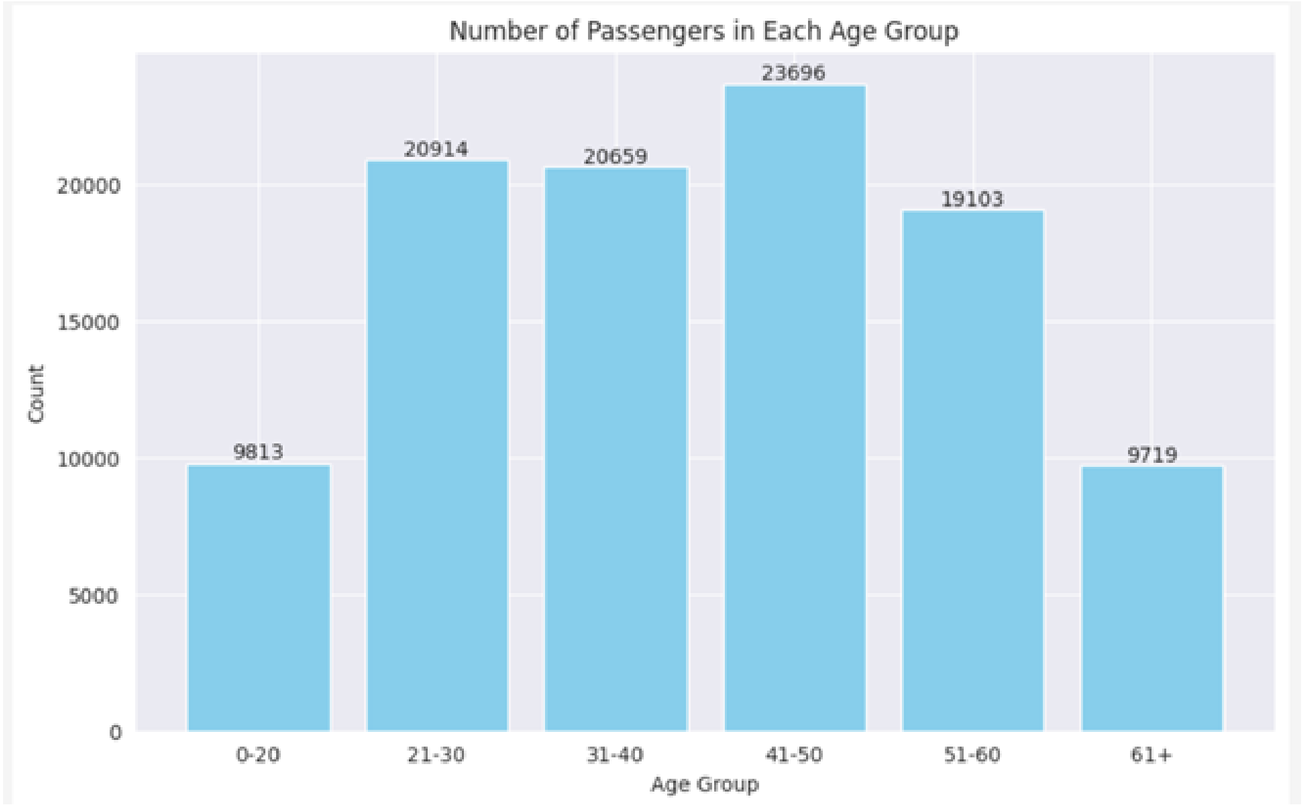


The individuals in the dataset span a wide range of ages, from 7 to 85 years old. This broad age range reflects the diversity of generations.

DISTRIBUTION OF 'AGE' AND ITS INFLUENCE ON SATISFACTION LEVELS

PREDOMINANT AGE GROUP

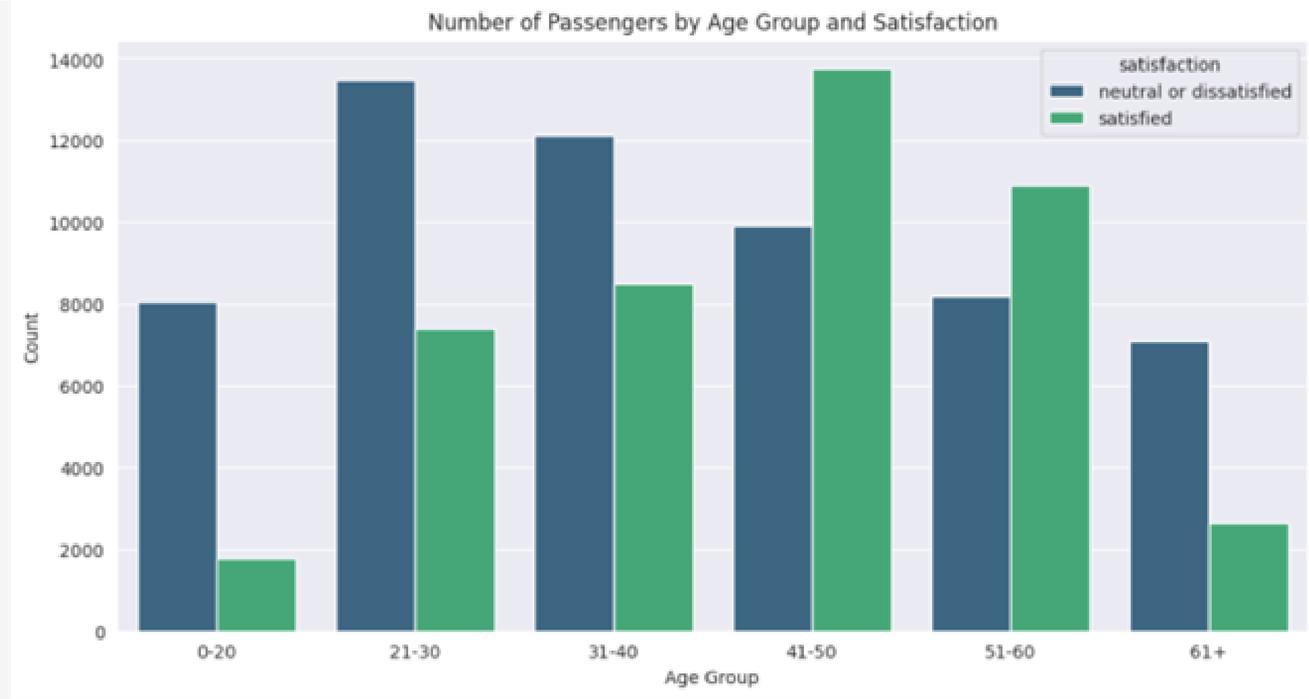
	Age Group	Count
0	41-50	23696
1	21-30	20914
2	31-40	20659
3	51-60	19103
4	0-20	9813
5	61+	9719



Overall, the data exhibits a diverse distribution across various age groups, with a notable concentration in the '41-50,' '21-30,' and '31-40' categories. Meanwhile, the '0-20' group. Although they are the smallest, still represents a significant portion of the dataset.

DISTRIBUTION OF 'AGE' AND ITS INFLUENCE ON SATISFACTION LEVELS

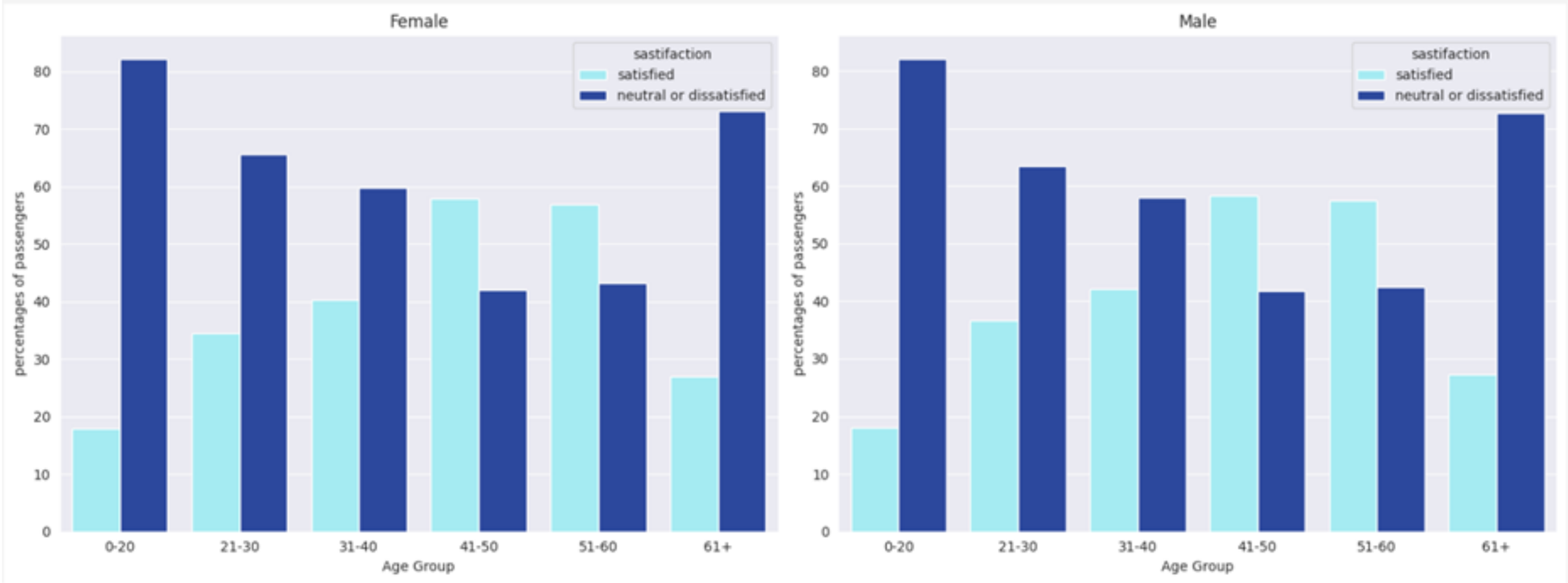
PREDOMINANT AGE GROUP



To sum up, the data uncovers intriguing satisfaction patterns among various age groups. Higher satisfaction levels are observed in the 41-50 and 51-60 age brackets, whereas the 0-20 and 61+ age groups demonstrate larger proportions of individuals expressing neutrality or dissatisfaction. The 31-40 age group falls in the middle, suggesting a relatively balanced distribution of sentiments.

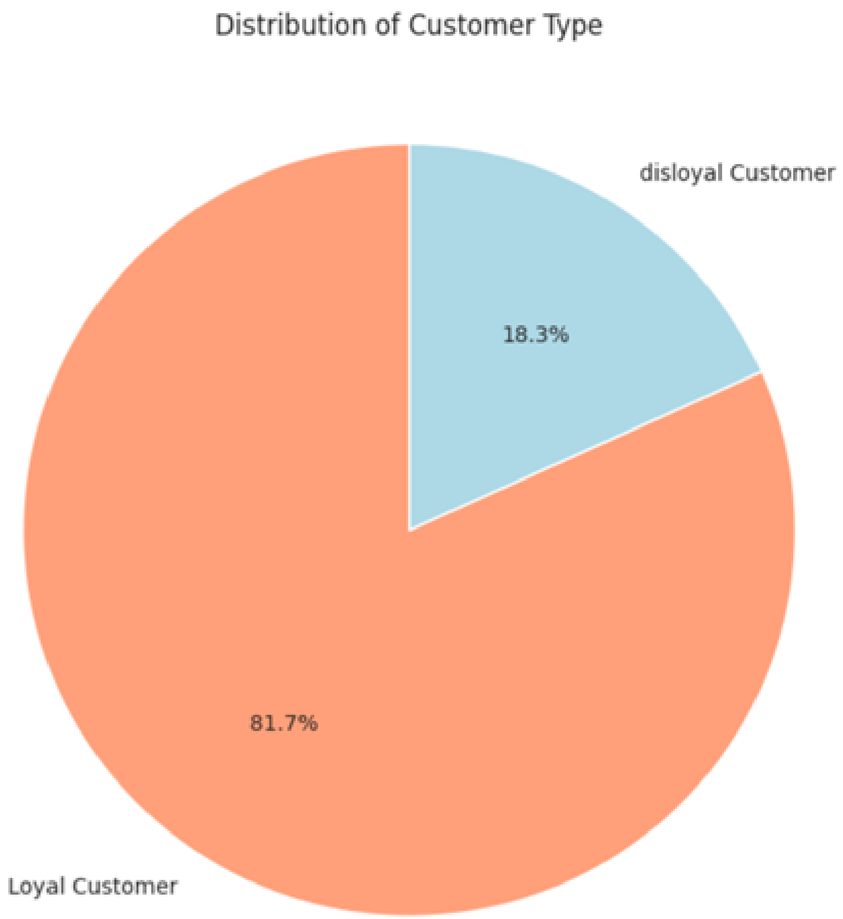
SASTIFACTION BY AGE AND GENDER

Gender	Age Group	neutral or dissatisfied	satisfied
Female	0-20	0.821120	0.178880
	21-30	0.655131	0.344869
	31-40	0.596979	0.403021
	41-50	0.420732	0.579268
	51-60	0.431317	0.568683
	61+	0.730800	0.269200
Male	0-20	0.819950	0.180050
	21-30	0.634110	0.365890
	31-40	0.579075	0.420925
	41-50	0.416923	0.583077
	51-60	0.424974	0.575026
	61+	0.726704	0.273296



In a broader sense, the data suggests consistent satisfaction patterns between genders throughout various age groups, with only slight variations. Generally, younger age groups tend to show higher proportions of neutrality or dissatisfaction, whereas older age groups demonstrate elevated levels of satisfaction.

DISTRIBUTION OF ‘CUSTOMER TYPE’ AND ITS INFLUENCE ON SATISFACTION LEVELS



Most customers (over 80%) belong to the loyal customer category, underscoring a robust rate of customer retention.

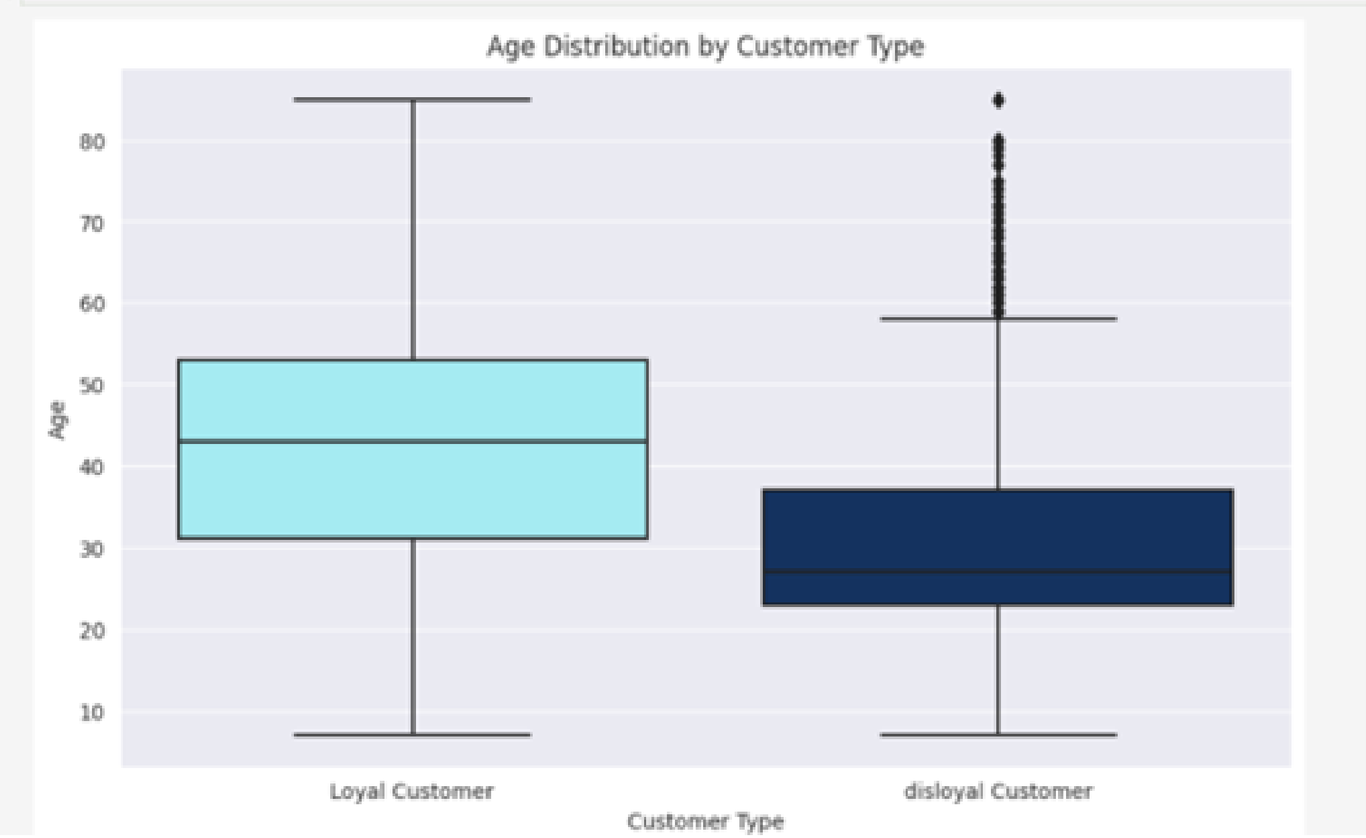
DISTRIBUTION OF ‘CUSTOMER TYPE’ AND ITS INFLUENCE ON SATISFACTION LEVELS

RELATIONSHIP BETWEEN SATISFACTION, AGE AND LOYALTY

satisfaction	neutral or dissatisfied	satisfied
Customer Type		
Loyal Customer	52.270881	47.729119
disloyal Customer	76.334229	23.665771

Loyal customers typically exhibit a well-balanced distribution between neutral/dissatisfied and satisfied sentiments, while a significant majority of disloyal customers express either neutrality or dissatisfaction.

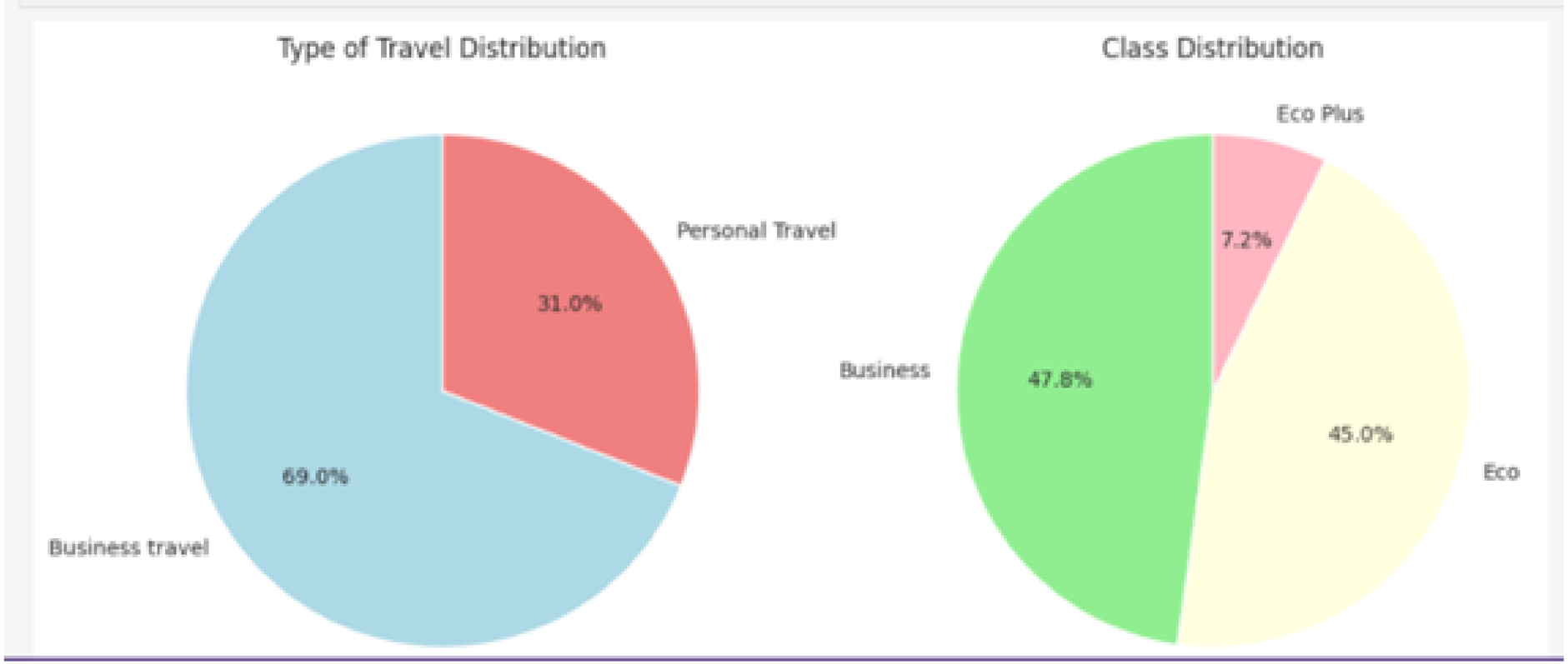
DISTRIBUTION OF ‘CUSTOMER TYPE’ AND ITS INFLUENCE ON SATISFACTION LEVELS



RELATIONSHIP BETWEEN SATISFACTION, AGE AND LOYALTY

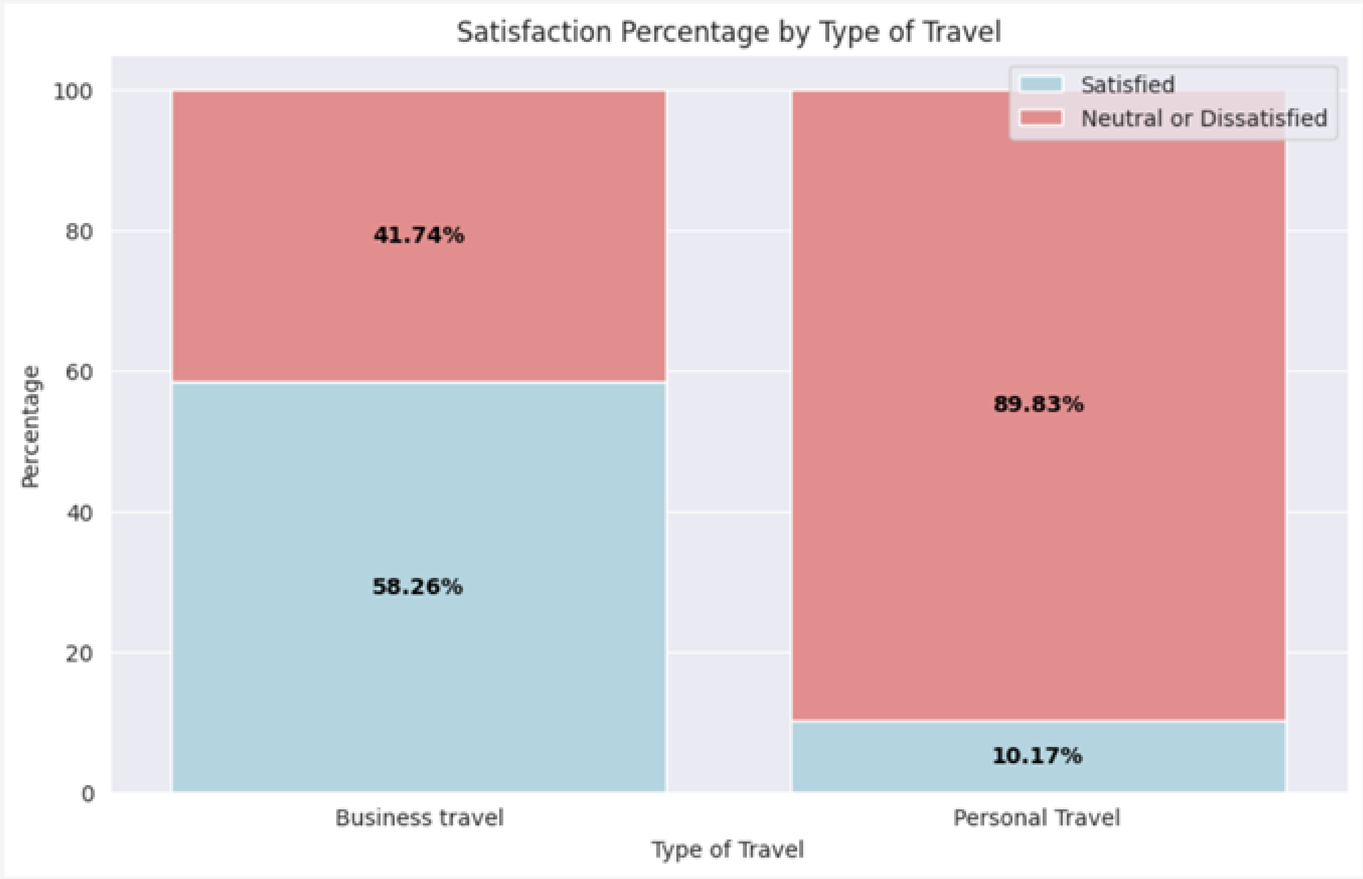
On average, loyal customers are older in comparison to their disloyal counterparts. Age appears to be a potential contributing factor to customer loyalty, as older individuals demonstrate a higher likelihood of being classified as loyal customers.

DISTRIBUTION OF ‘TYPE OF TRAVEL’ AND ‘TRAVEL CLASS’



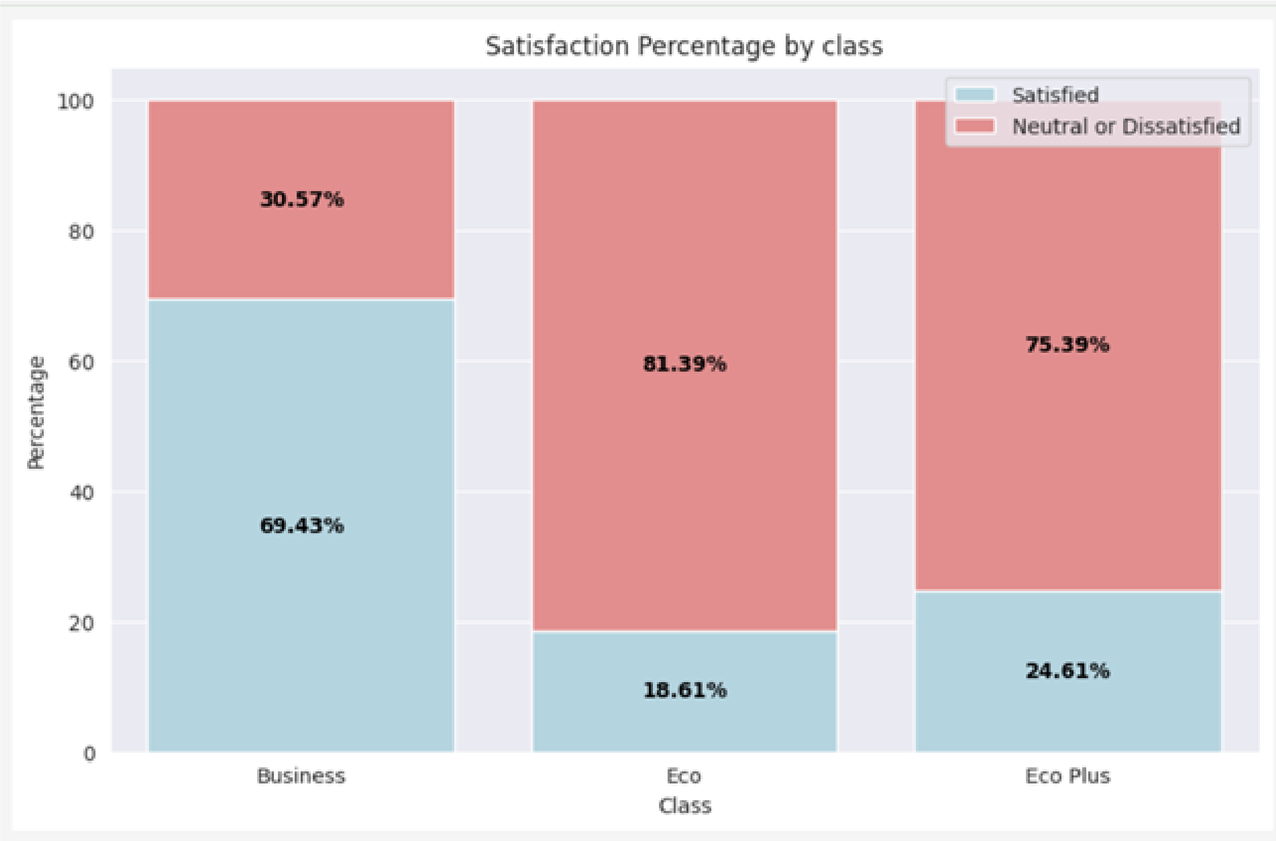
In general, ‘Business Travel’ take a higher count compared to ‘Personal Travel’. Furthermore, the Economy class exhibits the highest ticket count, followed by Business class and Economy Plus class in descending order.

DISTRIBUTION OF ‘TYPE OF TRAVEL’ AND ‘TRAVEL CLASS’



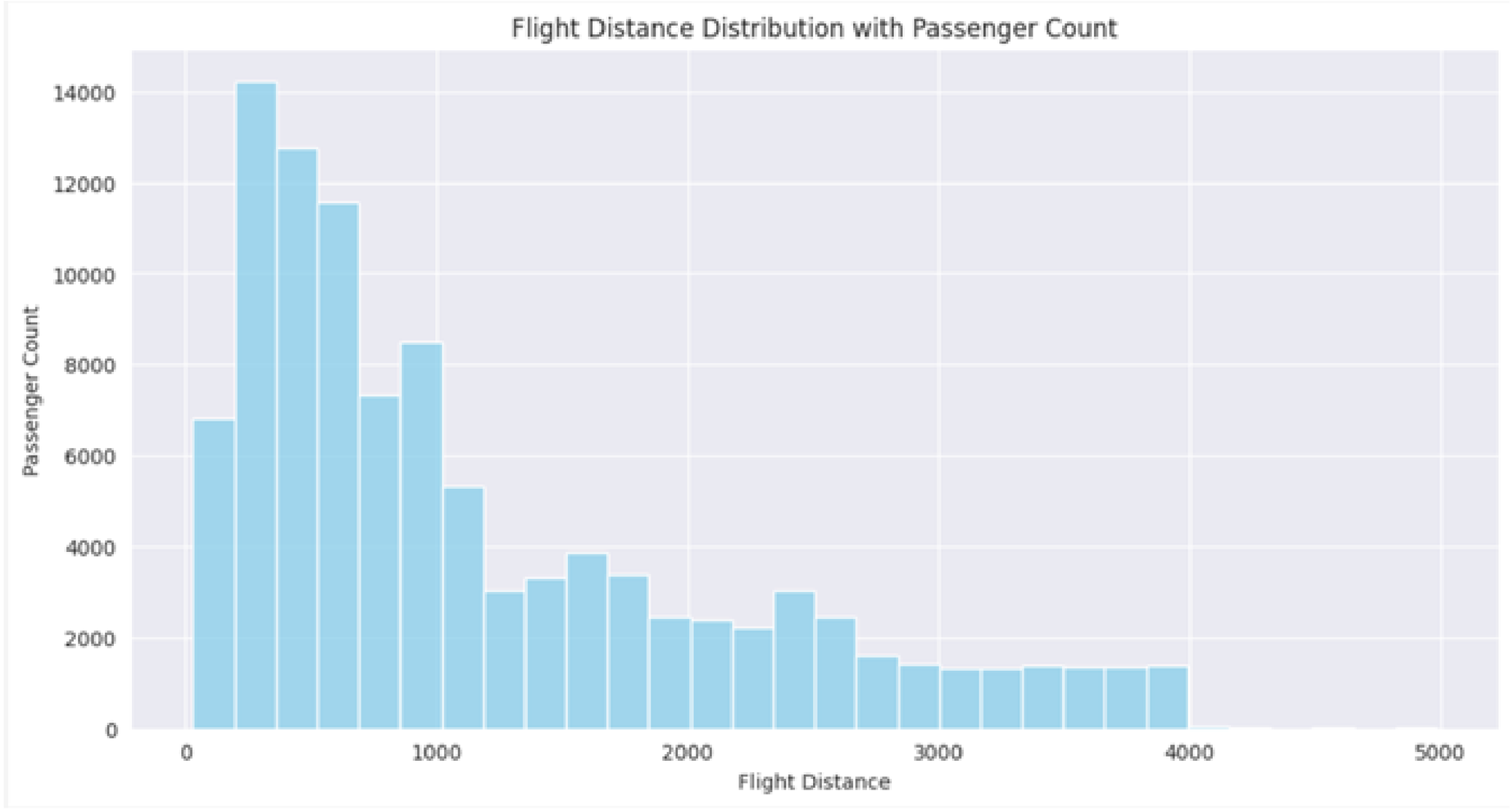
"Business Travelers show a significantly higher satisfaction rate compared to Personal Travelers, with the majority of Business Travelers expressing satisfaction, whereas the majority of Personal Travelers tend to express neutrality or dissatisfaction."

DISTRIBUTION OF ‘TYPE OF TRAVEL’ AND ‘TRAVEL CLASS’

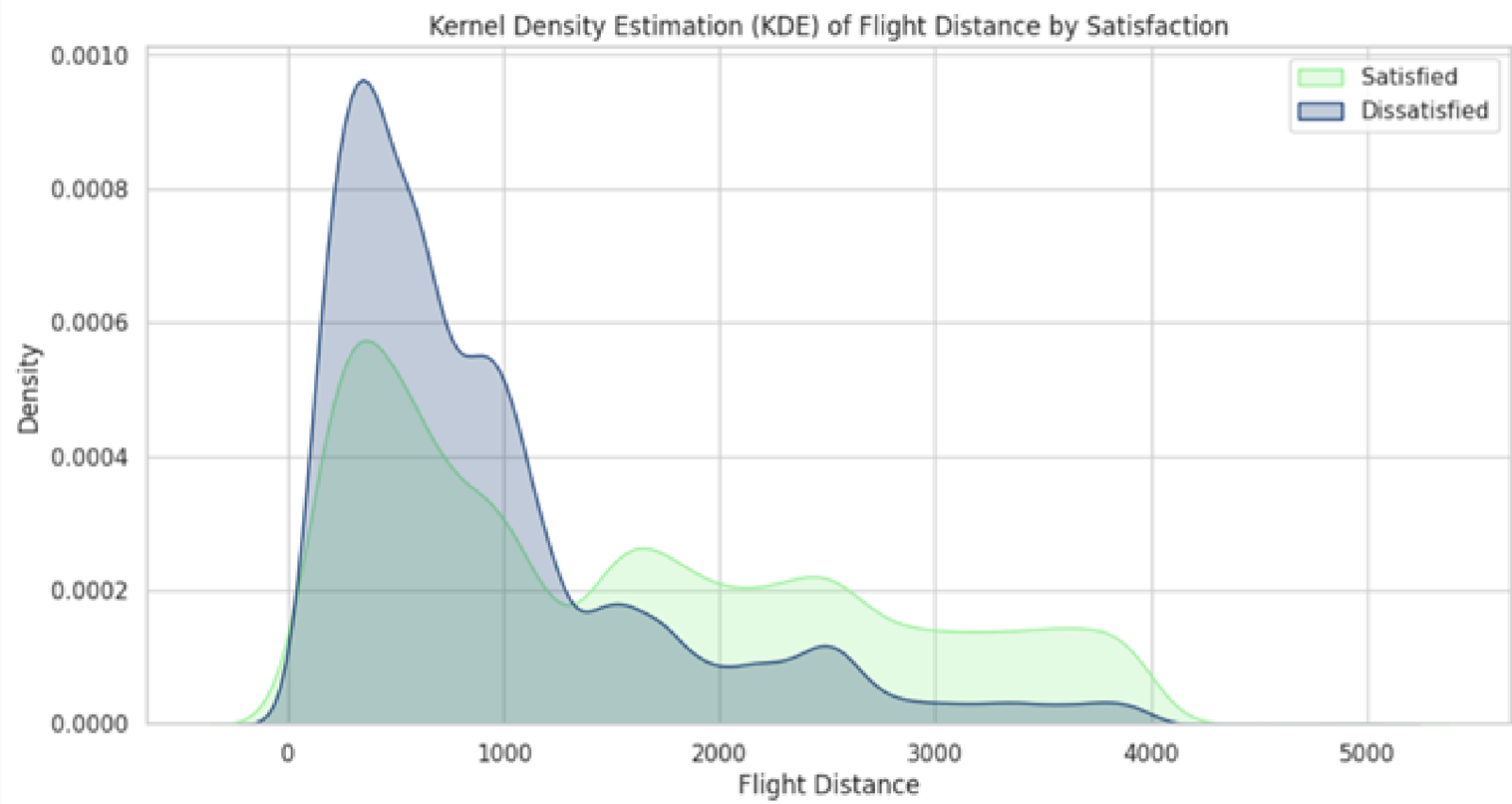


The data reveals that 'Business Class' has the highest proportion of satisfied customers, accounting for 69.43% of total satisfaction across all classes. Conversely, 'Eco Plus Class' exhibits the highest percentage of dissatisfied or neutral customers, comprising 75.39% of such sentiments across all classes.

DISTRIBUTION OF FLIGHT DISTANCE

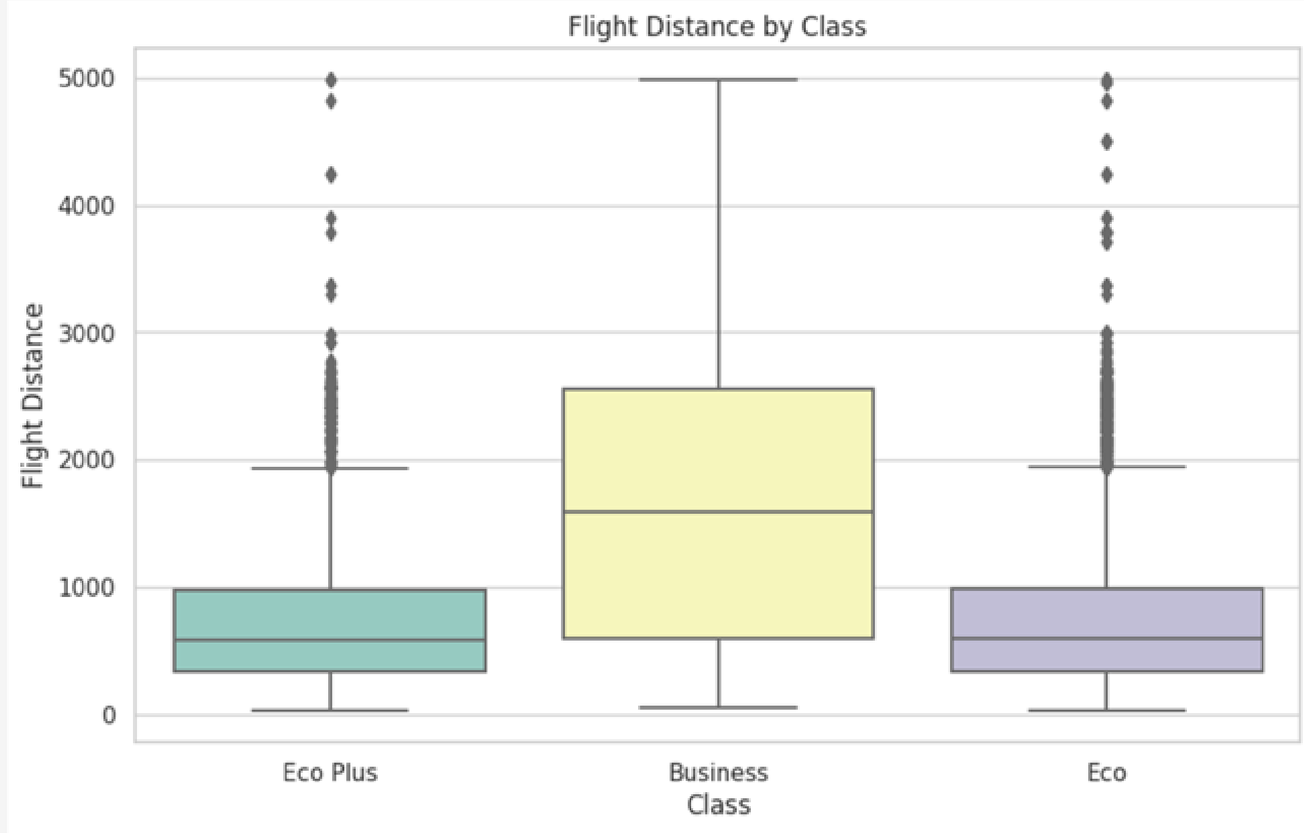


DISTRIBUTION OF FLIGHT DISTANCE



There appears to be a positive correlation between travel distance and satisfaction, suggesting that longer journeys are more likely to result in higher satisfaction. This correlation may be attributed to the observation that services on longer flights tend to be of higher quality compared to shorter ones.

DISTRIBUTION OF FLIGHT DISTANCE

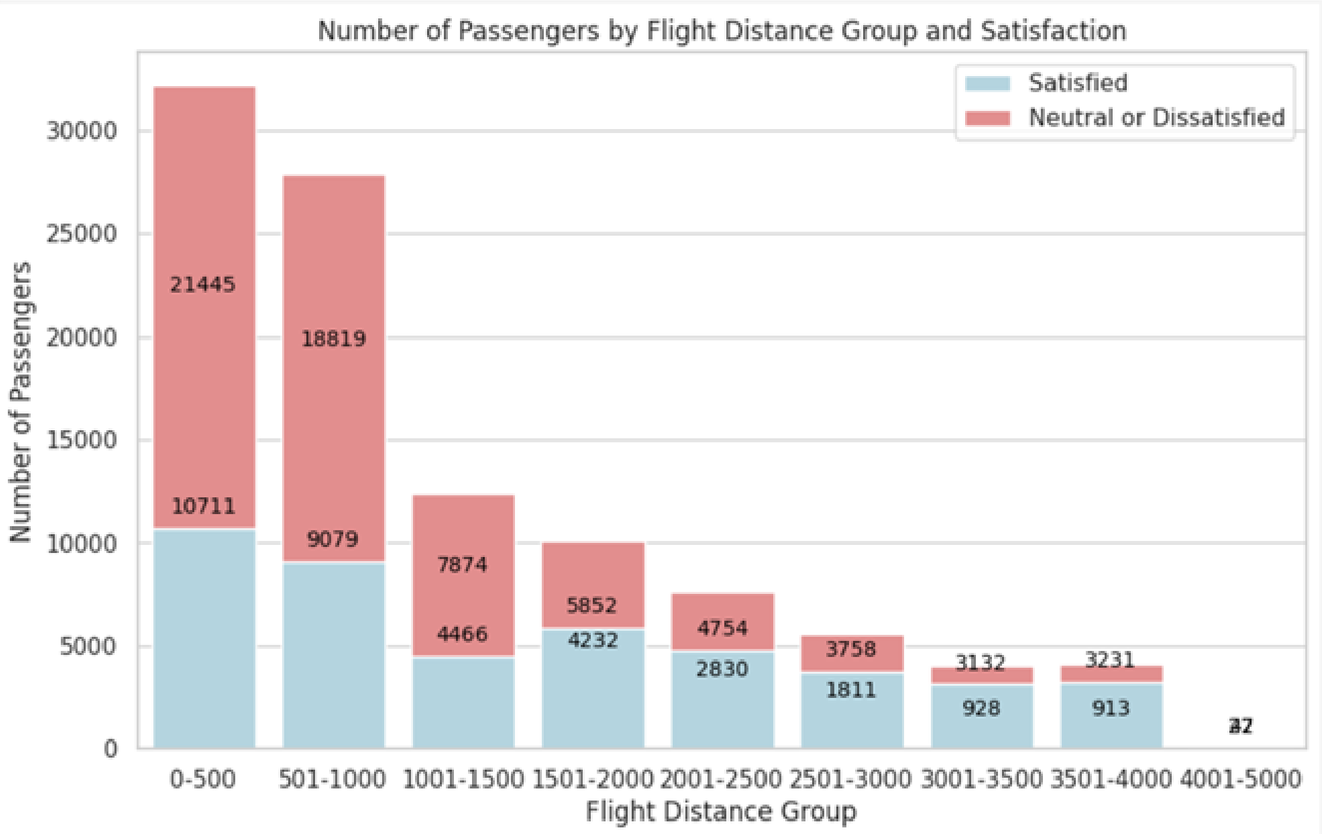


AVERAGE FLIGHT DISTANCE FOR EACH CLASS OF TRAVEL

	Class	Flight Distance
0	Business	1675.976925
1	Eco	743.439748
2	Eco Plus	747.125567

Business Class travelers, on average, cover longer flight distances compared to passengers in Economy Class and Economy Plus Class. In contrast, the flight distances for Economy Class and Economy Plus Class are notably shorter and exhibit a similar range.

DISTRIBUTION OF FLIGHT DISTANCE

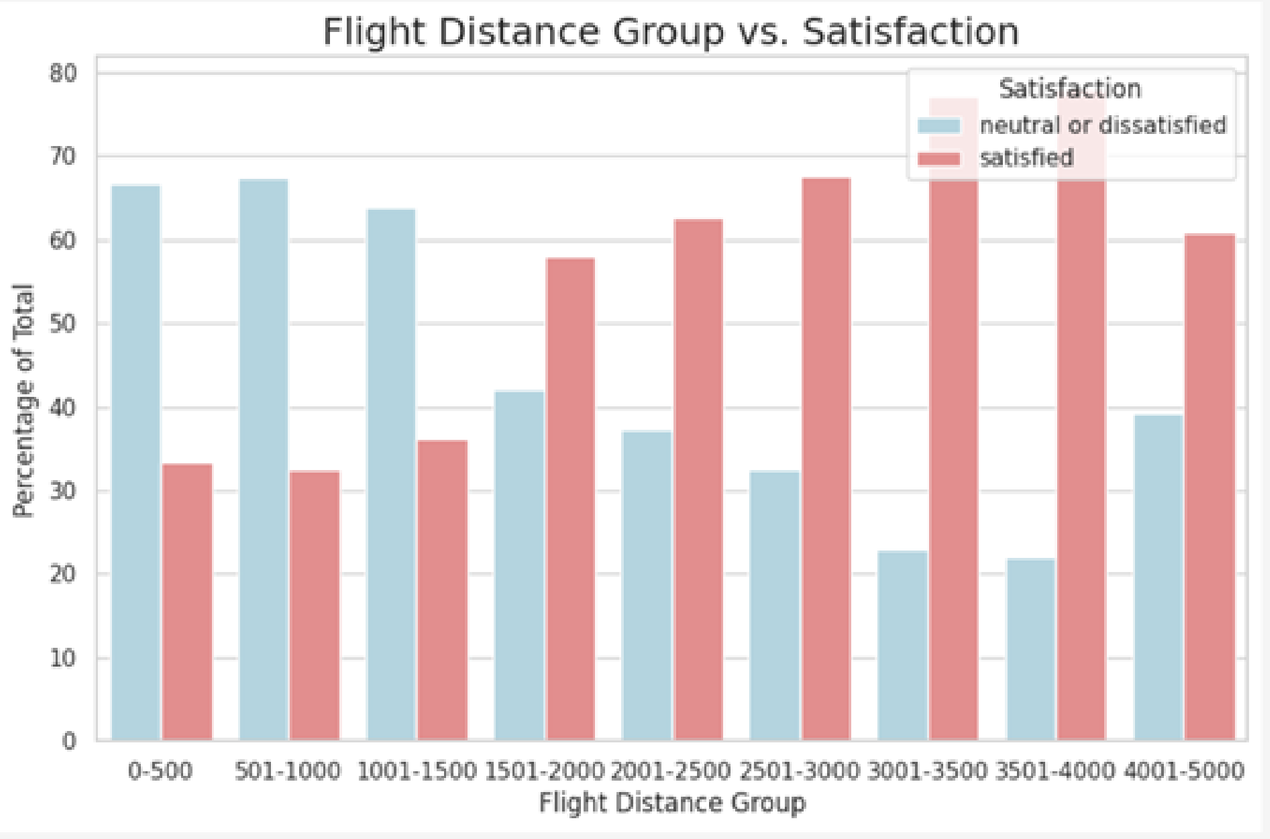


AVERAGE FLIGHT DISTANCE FOR EACH CLASS OF TRAVEL

	Flight Distance Group	Count
0	0-500	32156
1	501-1000	27898
2	1001-1500	12340
3	1501-2000	10084
4	2001-2500	7584
5	2501-3000	5569
6	3501-4000	4144
7	3001-3500	4060
8	4001-5000	69

The majority of flights are concentrated within shorter distance ranges (0-500 miles and 501-1000 miles), with a gradual decrease in count as the flight distance range expands. The data offers a comprehensive overview of flight distribution across different distance groups.

DISTRIBUTION OF FLIGHT DISTANCE

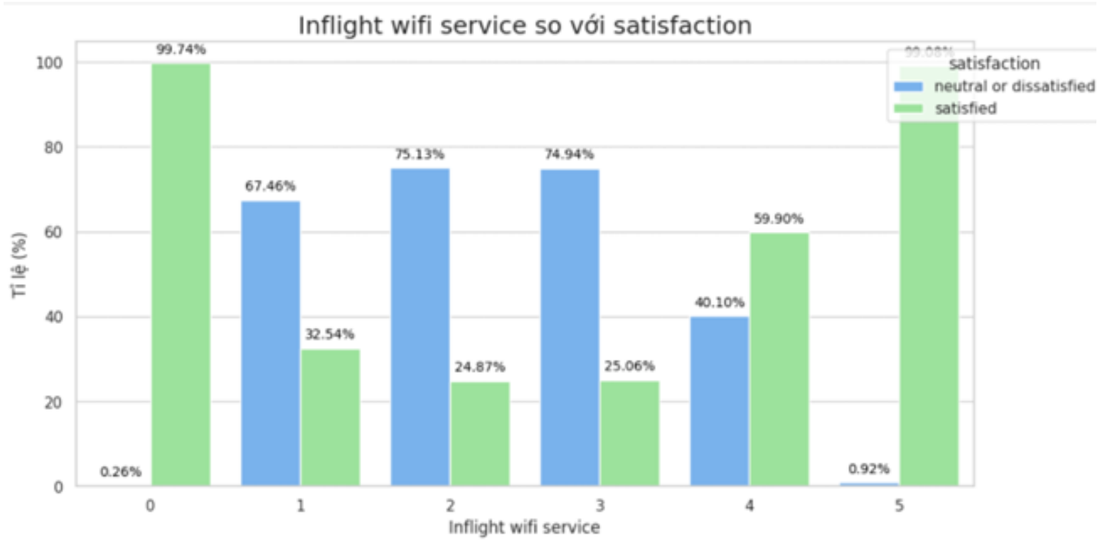


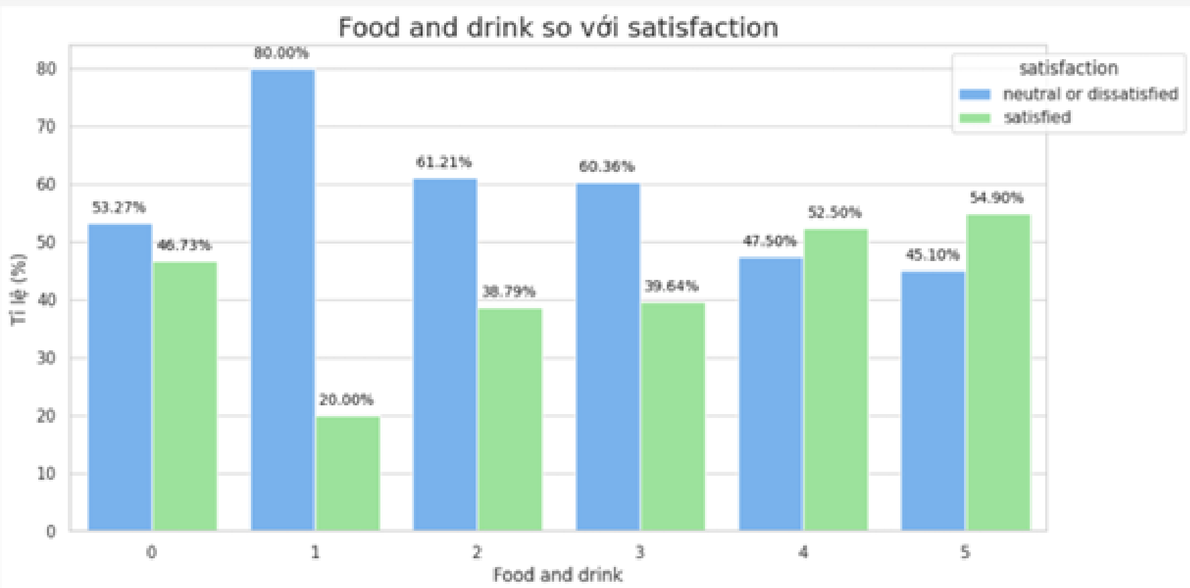
CORRELATION BETWEEN TRAVEL CLASS AND SATISFACTION LEVELS

Generally, shorter-distance flights tend to have lower satisfaction levels, whereas long-distance flights exhibit a mix of satisfaction and neutral/dissatisfied sentiments. Some distance groups show a majority of passengers expressing satisfaction.

SERVICE QUALITY ANALYSIS

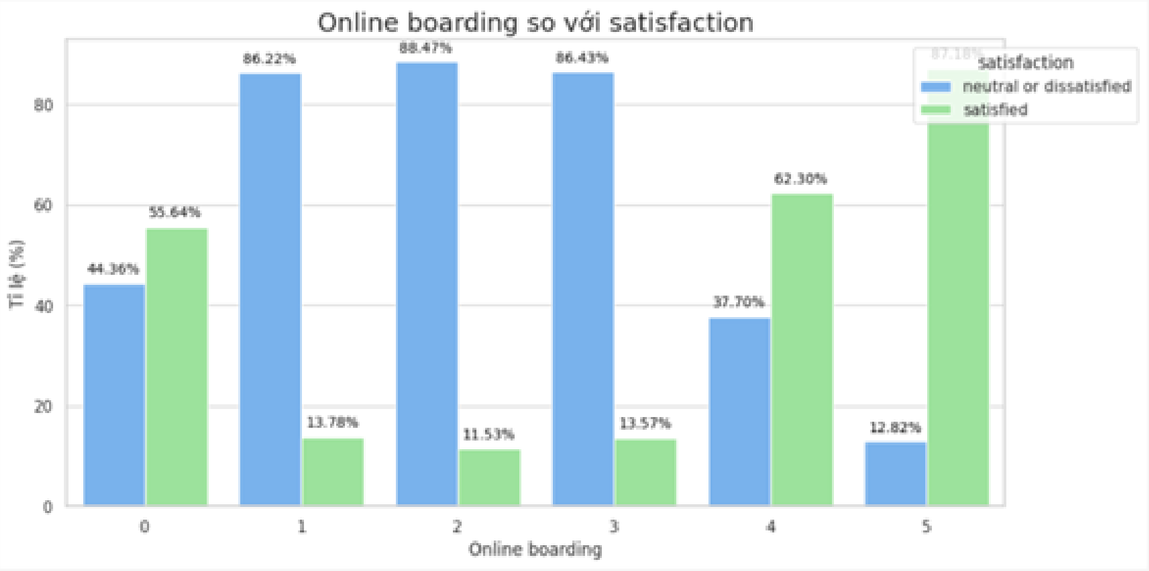
Let see what service factors have the highest and lowest satisfaction levels:

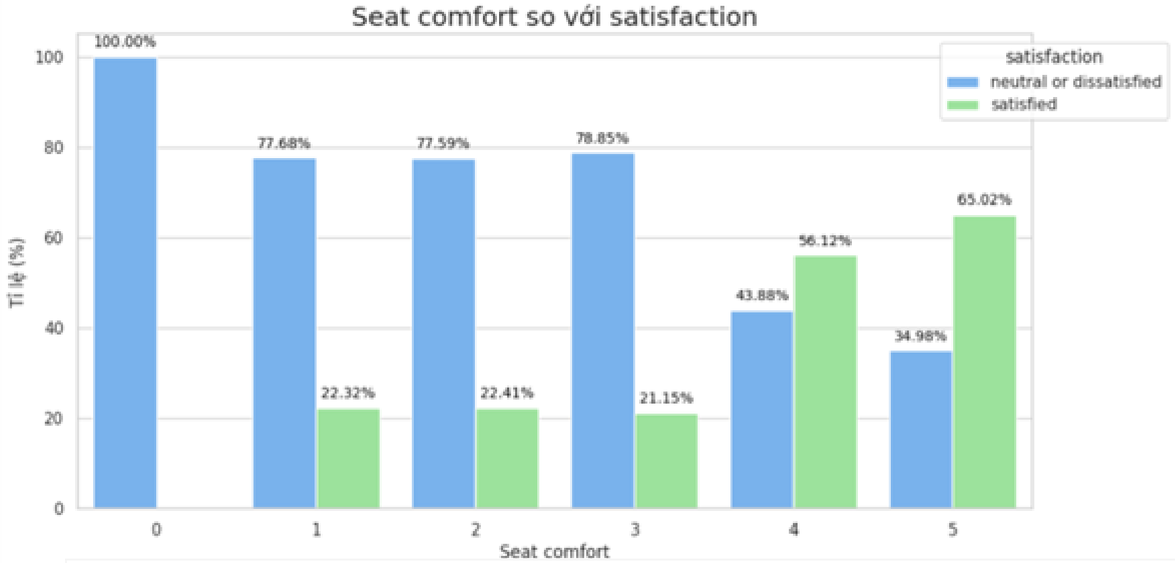


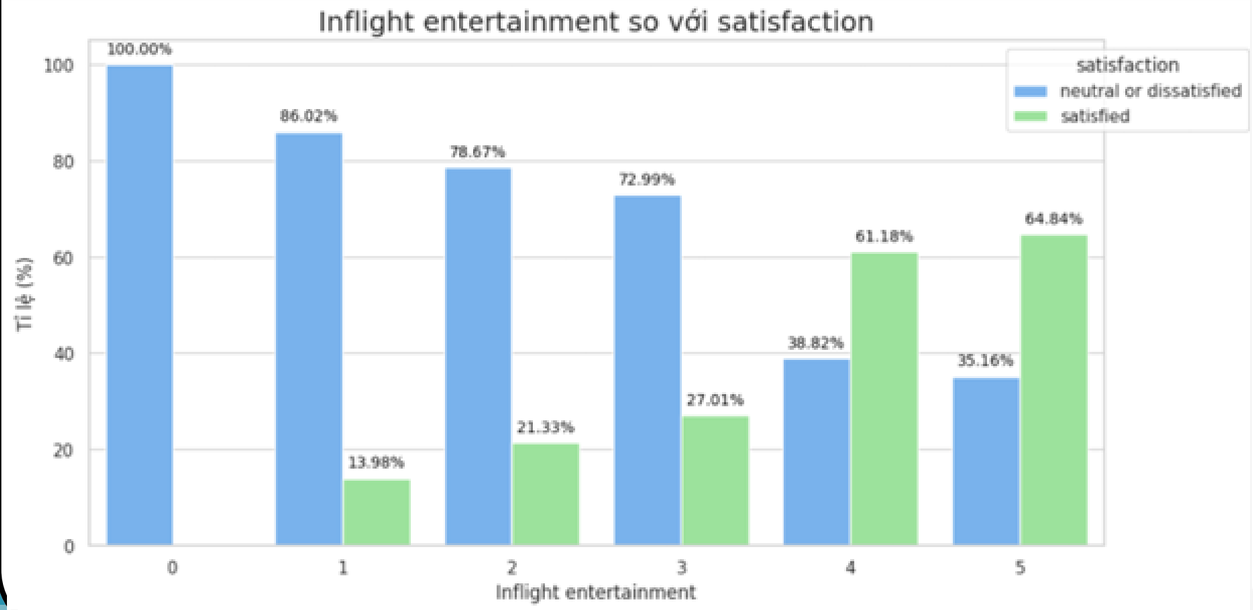


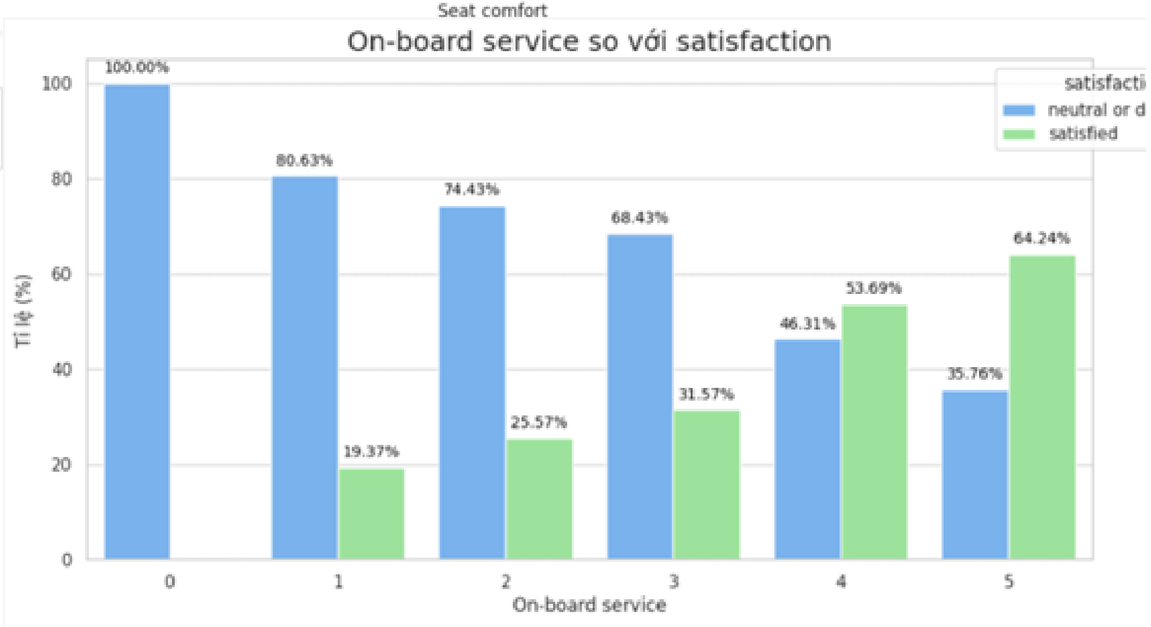
SERVICE QUALITY ANALYSIS

Let see what service factors have the highest and lowest satisfaction levels:



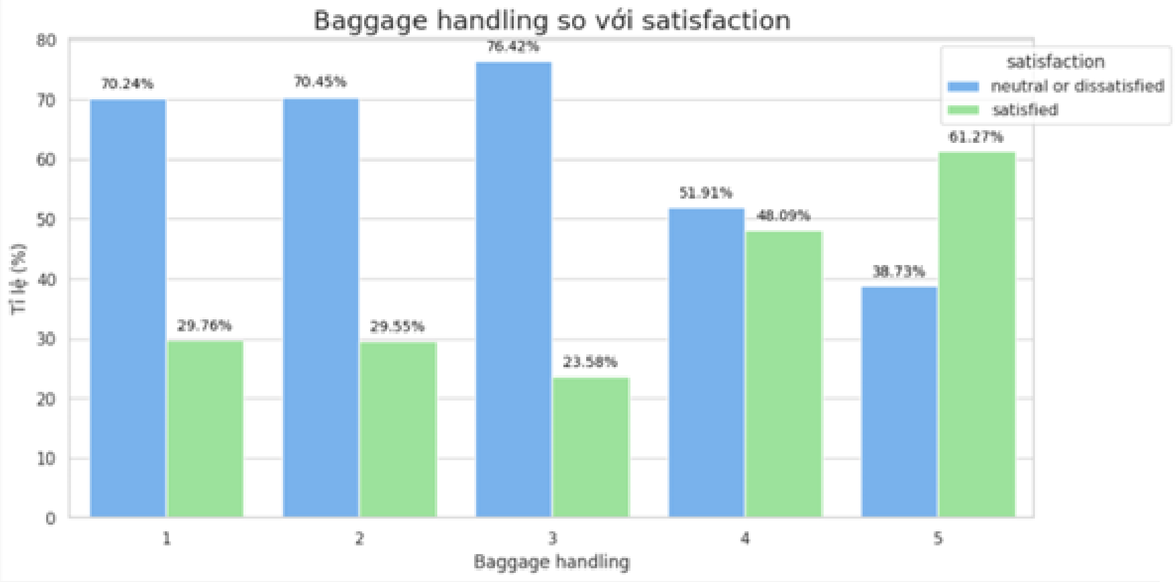


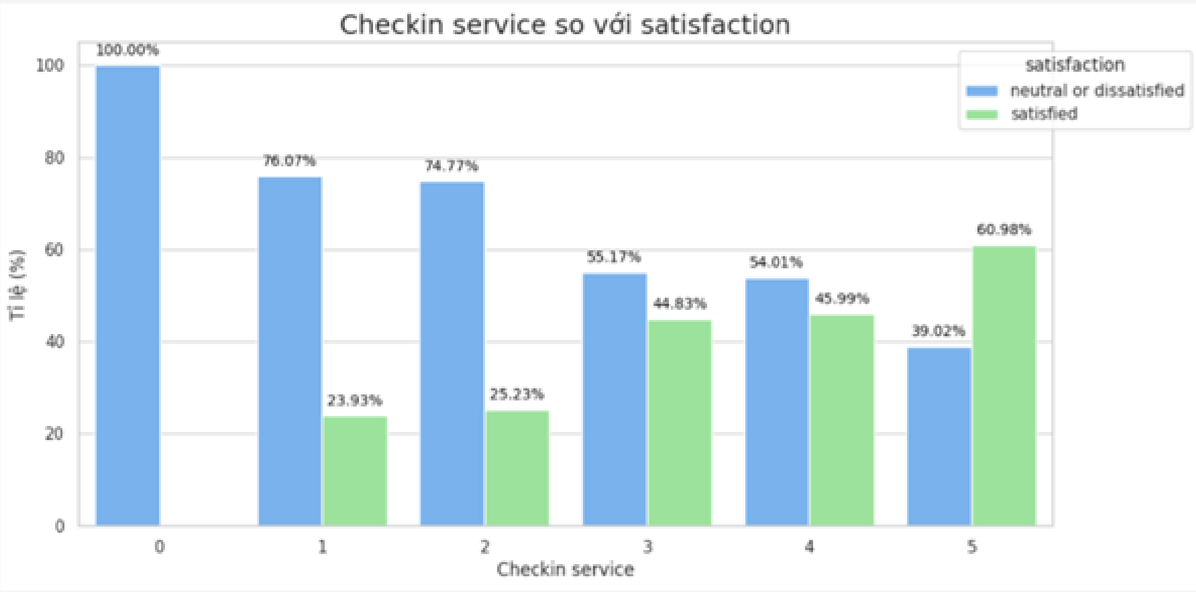


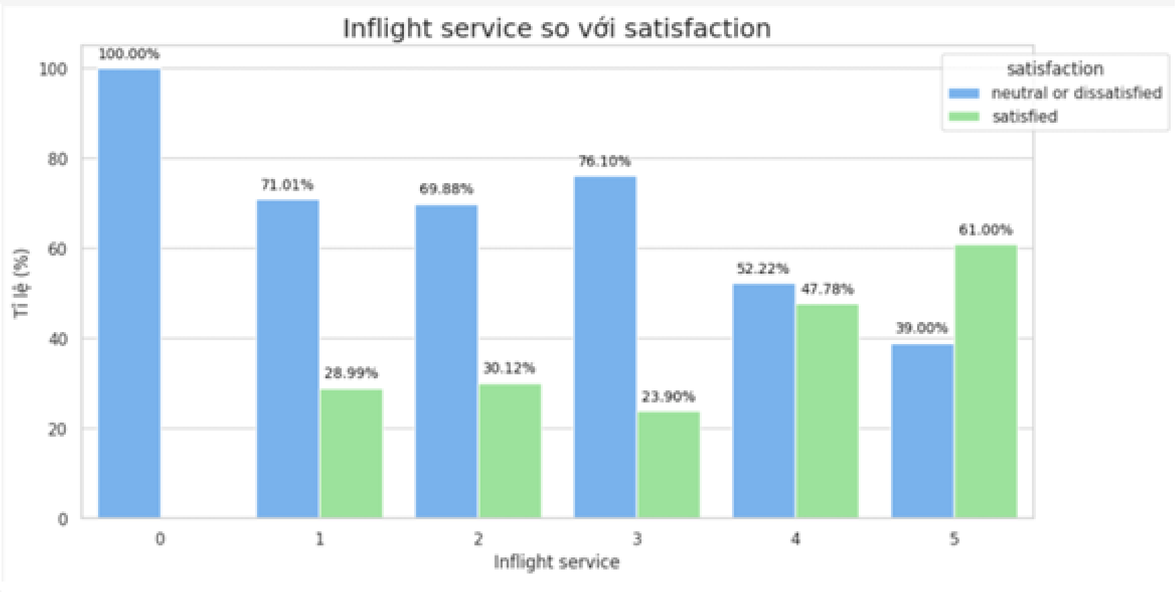


SERVICE QUALITY ANALYSIS

Let see what service factors have the highest and lowest satisfaction levels:

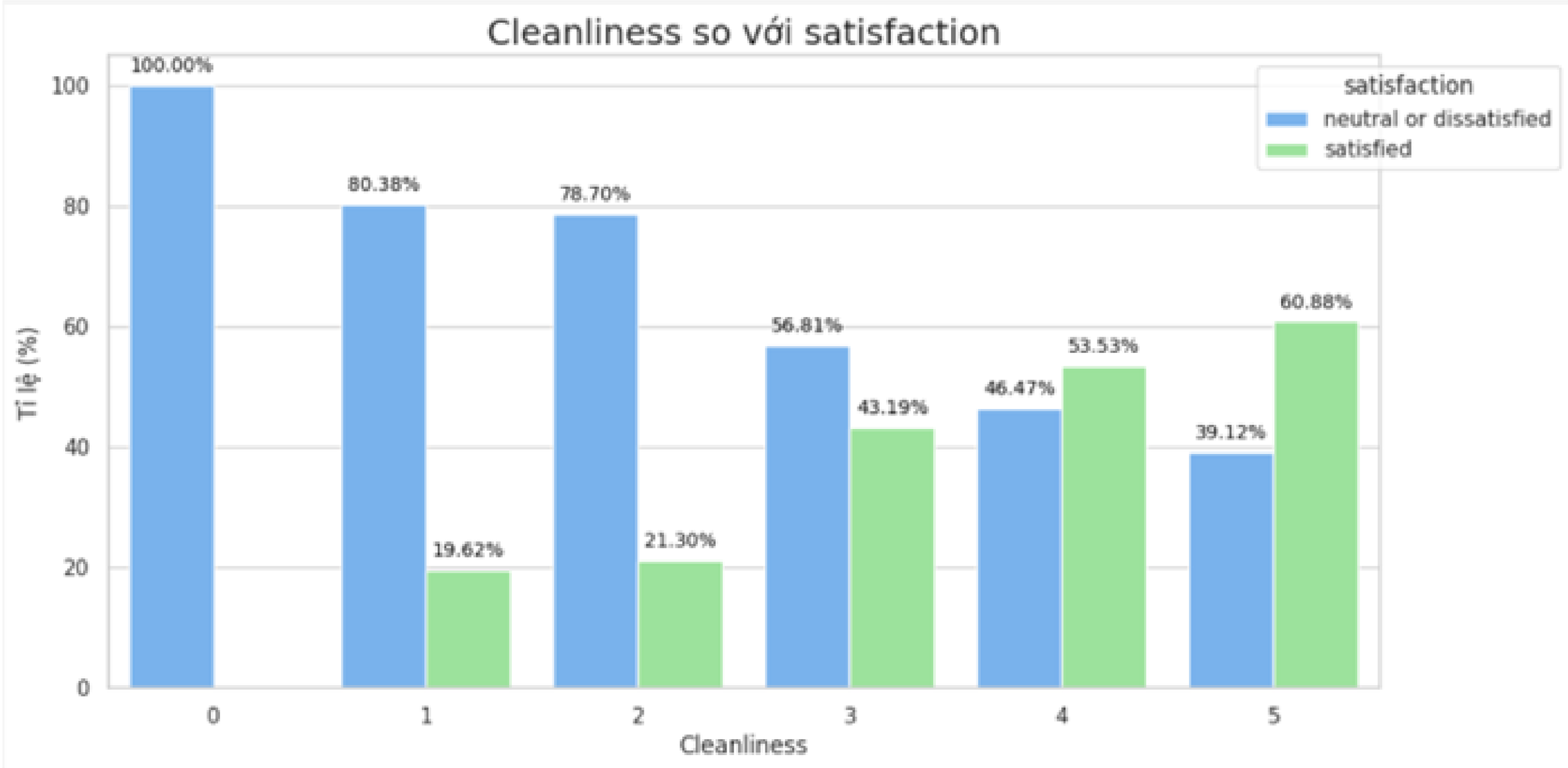




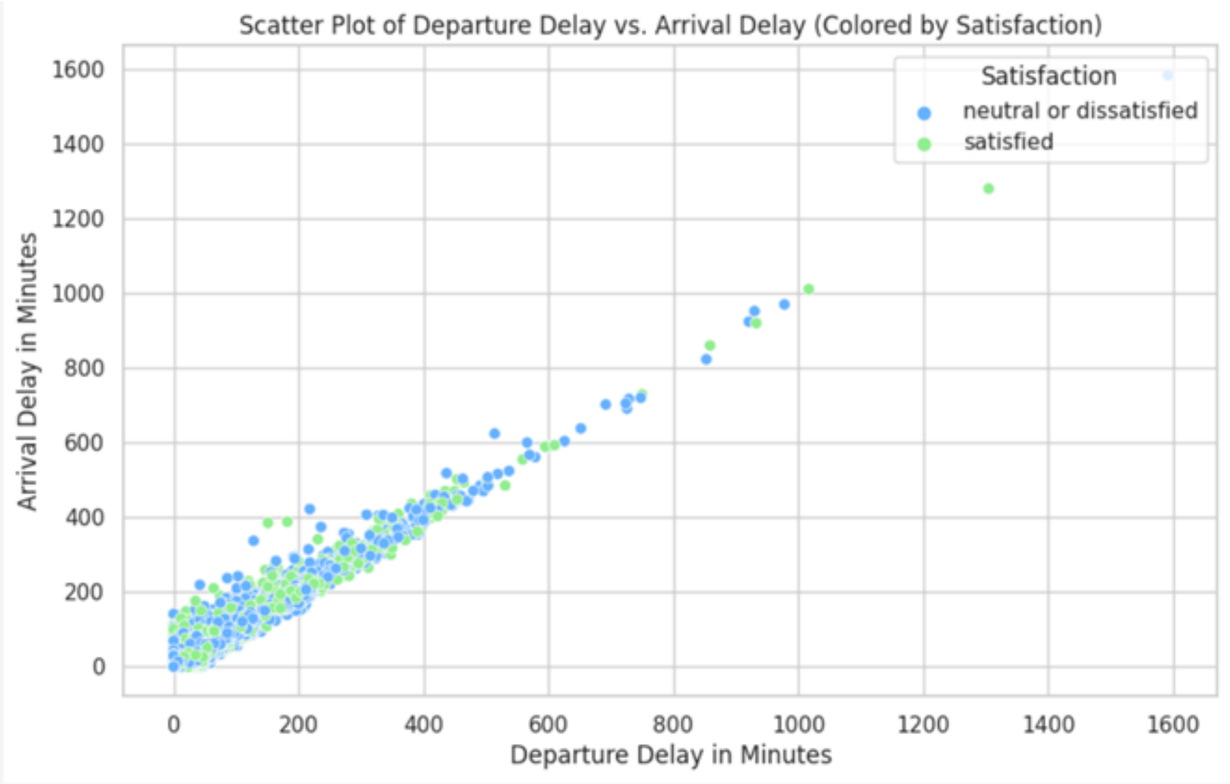


SERVICE QUALITY ANALYSIS

Let see what service factors have the highest and lowest satisfaction levels:

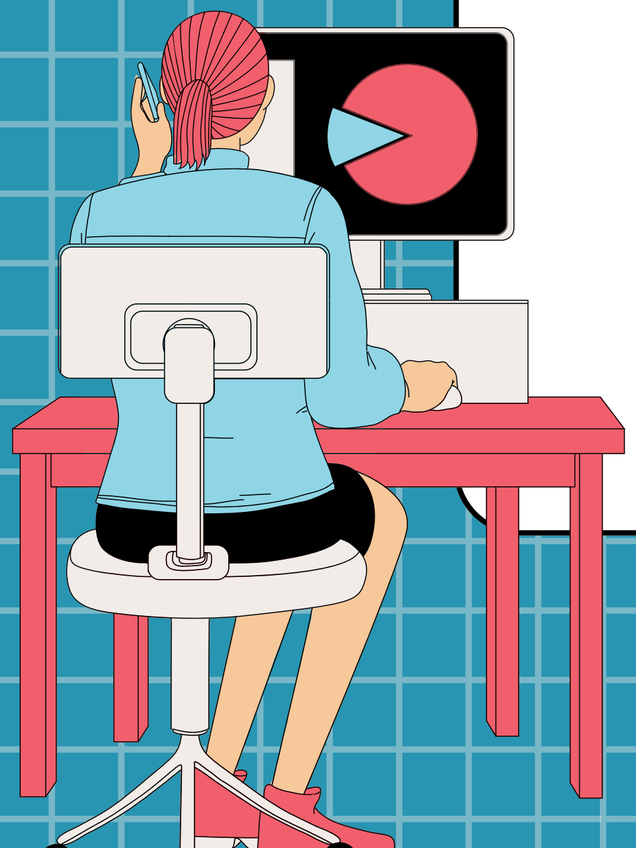


CORRELATION BETWEEN DELAYS (DEPARTURE AND ARRIVAL) AND SATISFACTION LEVELS

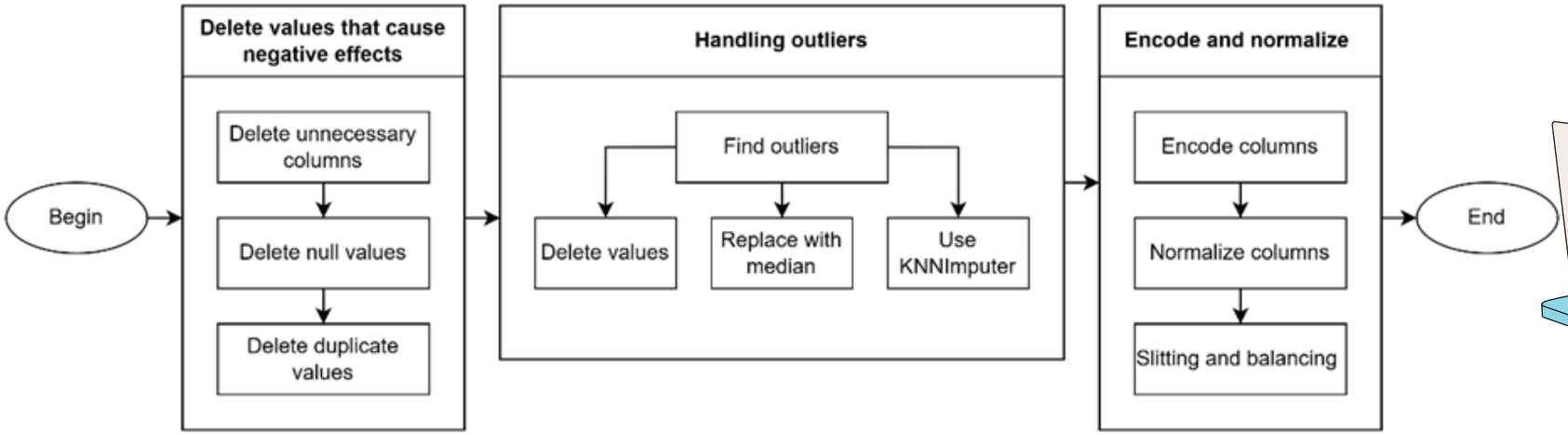


Individuals expressing neutral or dissatisfied sentiments frequently encounter more significant delays in their journeys compared to satisfied passengers. This underscores the pivotal role that punctuality and minimal delays play in shaping overall customer satisfaction.

3. DATA PREPROCESSING



DATA PREPROCESSING



DATA PREPROCESSING

- delete unnecessary columns

#	Numerical order
id	Flight ID code

- id and # columns dont use in any model so we can delete these columns

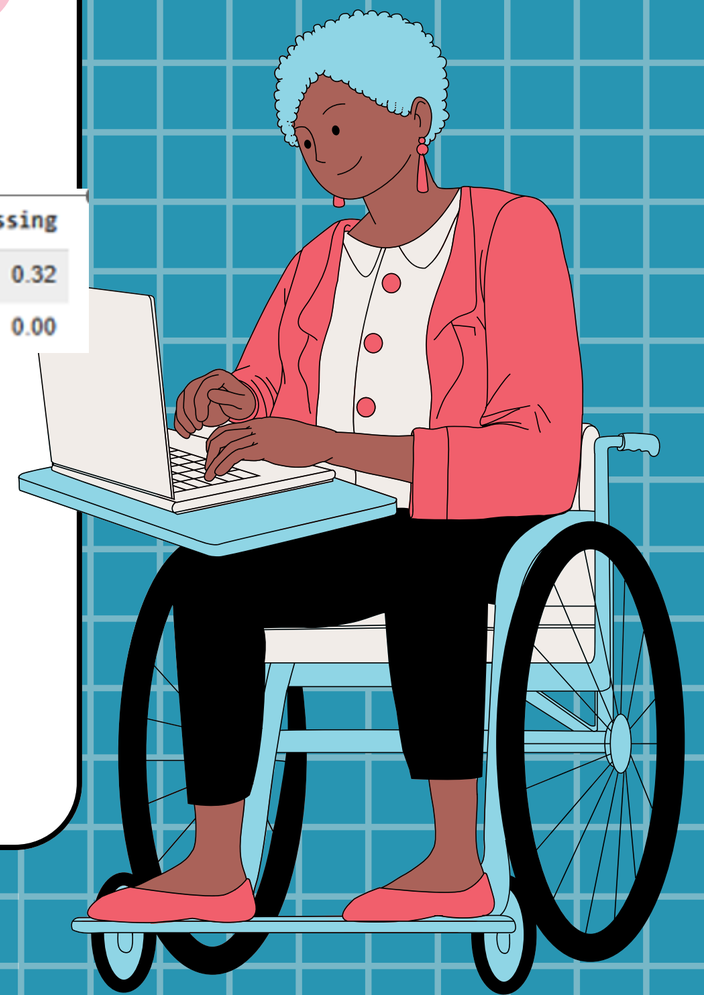
DATA PREPROCESSING

- delete null value

	Total Missing	Percent Missing
Arrival Delay in Minutes	310	0.298
Gender	0	0.000

	Total Missing	Percent Missing
Arrival Delay in Minutes	83	0.32
Gender	0	0.00

- We can see that in this column has ~0,2 null value, so we can delete them in train and test dataset



DATA PREPROCESSING

- Determine duplicated value

```
There is 0 duplicated data in training data
There is 0 duplicated data in testing data
```

Vậy không có giá trị nào bị trùng lấp

DATA PREPROCESSING

- Handing outliers

- Using quantile ranges to divide values into 4 parts, each part representing 25%.
- Find the interquartile range.
- Determine the lower bound and upper bound beyond which any values are considered outliers.

DATA PREPROCESSING

	Column Name	% Outliers
0	Age	0.0000
1	Arrival Delay in Minutes	13.4699
2	Baggage handling	0.0000
3	Checkin service	12.4071
4	Class	0.0000
5	Cleanliness	0.0000
6	Customer Type	0.0000
7	Departure Delay in Minutes	13.9274
8	Departure/Arrival time convenient	0.0000
9	Ease of Online booking	0.0000
10	Flight Distance	2.2077

	Column Name	% Outliers
0	Age	0.0000
1	Arrival Delay in Minutes	13.6639
2	Baggage handling	0.0000
3	Checkin service	12.3817
4	Class	0.0000
5	Cleanliness	0.0000
6	Customer Type	0.0000
7	Departure Delay in Minutes	13.6794
8	Departure/Arrival time convenient	0.0000
9	Ease of Online booking	0.0000
10	Flight Distance	2.2400

DATA PREPROCESSING

- Handling outliers

- In the 'Flight Distance' column, since the percentage of outliers is quite small, approximately 2%, we can consider removing these values.

DATA PREPROCESSING

- Handling outliers
- In the 'Checkin service' column, as the values range from 0 to 5 and outliers constitute approximately 12%, we can use the median to replace these outliers.

DATA PREPROCESSING

- Handling outliers

In the 'Departure Delay in Minutes' column, we can use the KNNimputer algorithm to replace outliers with a specified value of n_neighbors = 5. This means utilizing the values of the five nearest neighbors to predict the outlier values.

DATA PREPROCESSING

- Encode

Encode the rating values to a range from 0 to 1 to facilitate faster processing by algorithms.

DATA PREPROCESSING

Data correlations

	Age	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	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
Age	1.00	0.09	0.02	0.04	0.02	-0.00	0.02	0.21	0.16	0.07	0.05	0.04	-0.05	0.02	-0.05	0.05	-0.01	-0.01	0.13
Flight Distance	0.09	1.00	0.01	-0.02	0.07	0.00	0.05	0.21	0.15	0.12	0.11	0.13	0.06	0.04	0.05	0.09	0.00	-0.00	0.29
Inflight wifi service	0.02	0.01	1.00	0.34	0.72	0.33	0.14	0.47	0.12	0.21	0.12	0.16	0.12	0.01	0.11	0.13	-0.02	-0.02	0.29
Departure/Arrival time convenient	0.04	-0.02	0.34	1.00	0.43	0.43	0.01	0.07	0.01	-0.00	0.07	0.01	0.07	0.04	0.07	0.02	0.00	-0.00	-0.05
Ease of Online booking	0.02	0.07	0.72	0.43	1.00	0.45	0.03	0.41	0.03	0.05	0.04	0.11	0.04	0.01	0.03	0.02	-0.01	-0.01	0.17
Gate location	-0.00	0.00	0.33	0.43	0.45	1.00	-0.00	0.00	0.00	0.00	-0.03	-0.01	0.00	-0.01	0.00	-0.00	0.01	0.01	0.00
Food and drink	0.02	0.05	0.14	0.01	0.03	-0.00	1.00	0.23	0.58	0.63	0.06	0.03	0.03	0.04	0.03	0.66	-0.03	-0.03	0.21
Online boarding	0.21	0.21	0.47	0.07	0.41	0.00	0.23	1.00	0.42	0.28	0.15	0.12	0.08	0.09	0.07	0.33	-0.02	-0.02	0.50
Seat comfort	0.16	0.15	0.12	0.01	0.03	0.00	0.58	0.42	1.00	0.61	0.13	0.10	0.07	0.09	0.07	0.68	-0.03	-0.03	0.34
Inflight entertainment	0.07	0.12	0.21	-0.00	0.05	0.00	0.63	0.28	0.61	1.00	0.41	0.29	0.37	0.04	0.40	0.70	-0.03	-0.03	0.39
On-board service	0.05	0.11	0.12	0.07	0.04	-0.03	0.06	0.15	0.13	0.41	1.00	0.35	0.51	0.11	0.55	0.12	-0.03	-0.03	0.32
Leg room service	0.04	0.13	0.16	0.01	0.11	-0.01	0.03	0.12	0.10	0.29	0.35	1.00	0.36	0.07	0.36	0.09	0.01	0.01	0.31
Baggage handling	-0.05	0.06	0.12	0.07	0.04	0.00	0.03	0.08	0.07	0.37	0.51	0.36	1.00	0.11	0.63	0.09	-0.00	-0.01	0.24
Checkin service	0.02	0.04	0.01	0.04	0.01	-0.01	0.04	0.09	0.09	0.04	0.11	0.07	0.11	1.00	0.12	0.08	-0.02	-0.02	0.09
Inflight service	-0.05	0.05	0.11	0.07	0.03	0.00	0.03	0.07	0.07	0.40	0.55	0.36	0.63	0.12	1.00	0.09	-0.05	-0.06	0.24
Cleanliness	0.05	0.09	0.13	0.02	0.02	-0.00	0.66	0.33	0.68	0.70	0.12	0.09	0.09	0.08	0.09	1.00	-0.01	-0.02	0.30
Departure Delay in Minutes	-0.01	0.00	-0.02	0.00	-0.01	0.01	-0.03	-0.02	-0.03	-0.03	-0.03	0.01	-0.00	-0.02	-0.05	-0.01	1.00	0.97	-0.05
Arrival Delay in Minutes	-0.01	-0.00	-0.02	-0.00	-0.01	0.01	-0.03	-0.02	-0.03	-0.03	-0.03	0.01	-0.01	-0.02	-0.06	-0.02	0.97	1.00	-0.06
satisfaction	0.13	0.29	0.29	-0.05	0.17	0.00	0.21	0.50	0.34	0.39	0.32	0.31	0.24	0.09	0.24	0.30	-0.05	-0.06	1.00

DATA PREPROCESSING

- Unnecessary Features

It can be observed that columns such as 'Gender,' 'Arrival Delay in Minutes,' 'Gate location,' 'Departure/Arrival time convenient' exhibit low data correlation, making them unnecessary for the algorithm.

DATA PREPROCESSING

- One hot encoding

Utilize the one-hot encoding algorithm for the columns ['Customer Type', 'Type of Travel', 'Class']. Since these columns do not have ordinal distinctions (no inherent order), employing this algorithm is appropriate.

DATA PREPROCESSING

- One hot encoding

Customer Type_Loyal Customer	Customer Type_disloyal Customer	Type of Travel_Business travel	Type of Travel_Personal Travel	Class_Business	Class_Eco	Class_Eco Plus
1.0	0.0	0.0	1.0	0.0	0.0	1.0
0.0	1.0	1.0	0.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0	1.0	0.0	0.0

DATA PREPROCESSING

- Standardization
- Standardization scales features by subtracting the mean and then dividing by the standard deviation. This results in features that have a mean of 0 and a standard deviation of 1.

DATA PREPROCESSING

- Balance dataset

There is an imbalance in the outcome variable, with more occurrences of 0 indicating "Neutral or dissatisfied" compared to 1 indicating "Satisfied."

DATA PREPROCESSING

- Balance data

Utilize the Random Over-sampling method with a seed value of 42. This method involves increasing the number of samples in the minority class by randomly duplicating existing samples.

4. EXPERIMENTAL

SVM ALGORITHM

Support Vector Machine (SVM) model works by finding a hyperplane in a high-dimensional space (corresponding to the number of features in the data) in such a way that it optimally classifies data points into different classes

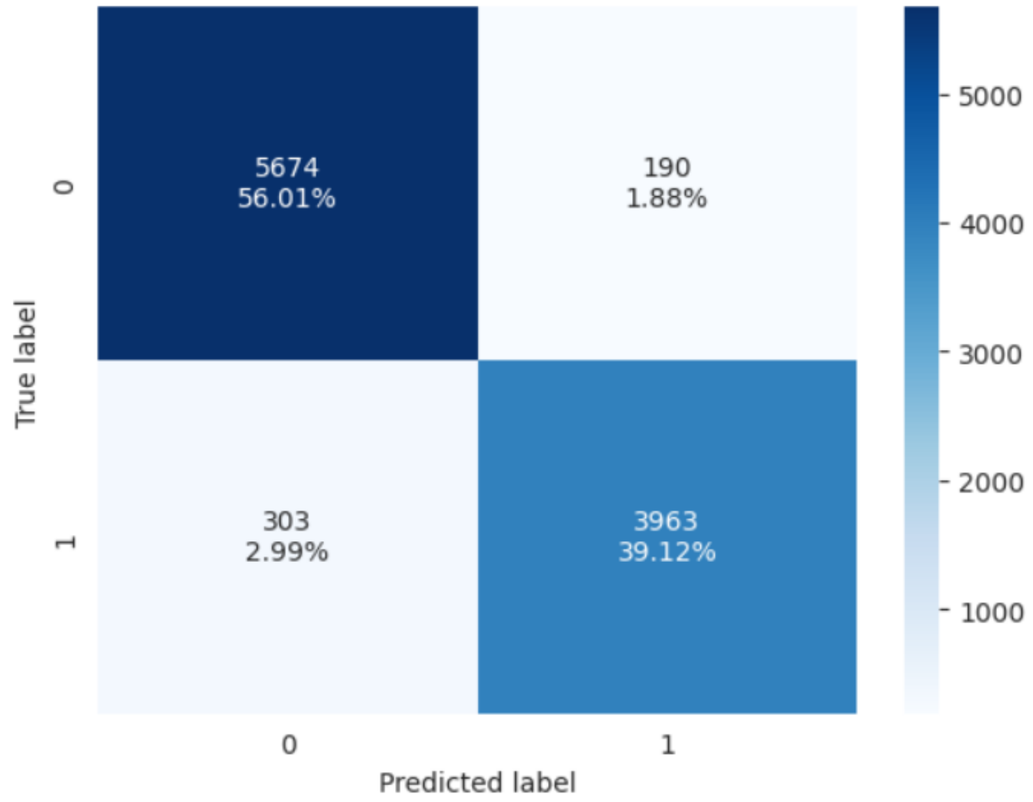
SVM ALGORITHM

Hyperplane is chosen so that the distance from the hyperplane to the nearest data points (known as support vectors) is maximized. The goal of SVM is to maximize the margin between classes, which is synonymous with optimizing classification performance.

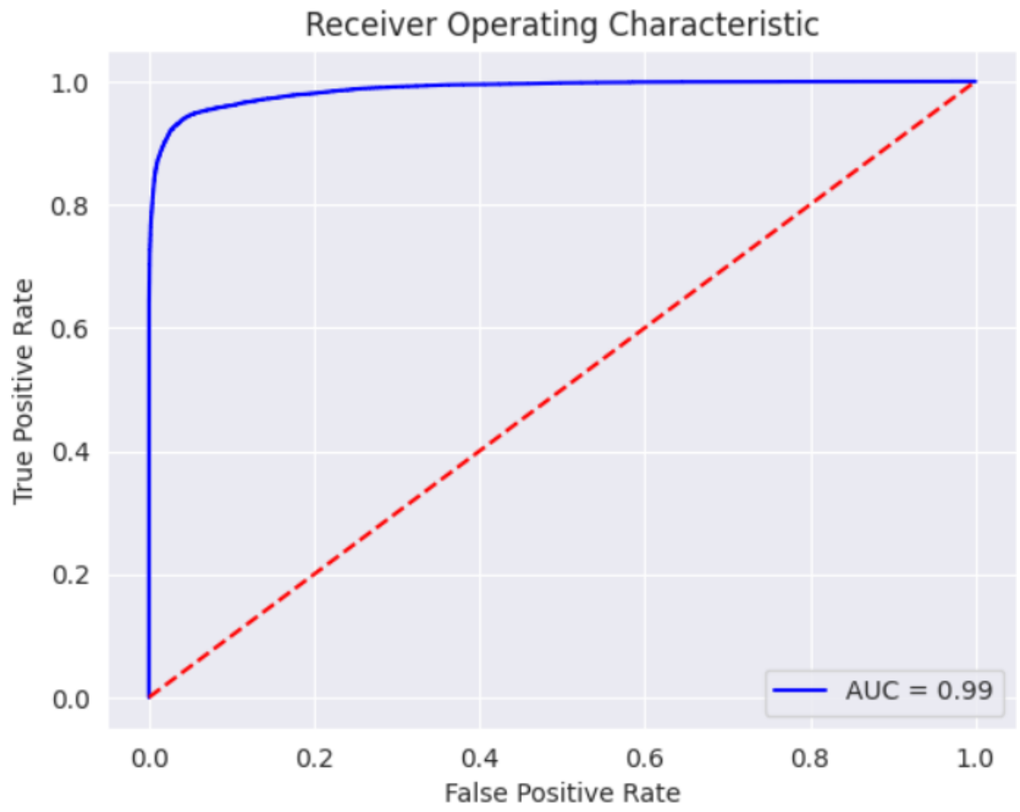
SVM ALGORITHM

Our team utilizes the default kernel, which is the Radial Basis Function (RBF), and applies a value of C equal to 1.0.

SVM ALGORITHM



SVM ALGORITHM



KNN ALGORITHM

Based on the assumption that data points with similar labels tend to have similar features.

To predict the label of a new data point, K-Nearest Neighbors (KNN) identifies its nearest neighbors in the feature space and predicts the label based on the majority label among those neighbors.

KNN ALGORITHM

Choose the number of neighbors (K): This is the number of nearest data points the algorithm will consider for making predictions.

Define a distance measure: Use a distance metric (usually Euclidean distance) to determine the nearest neighbors.

KNN ALGORITHM

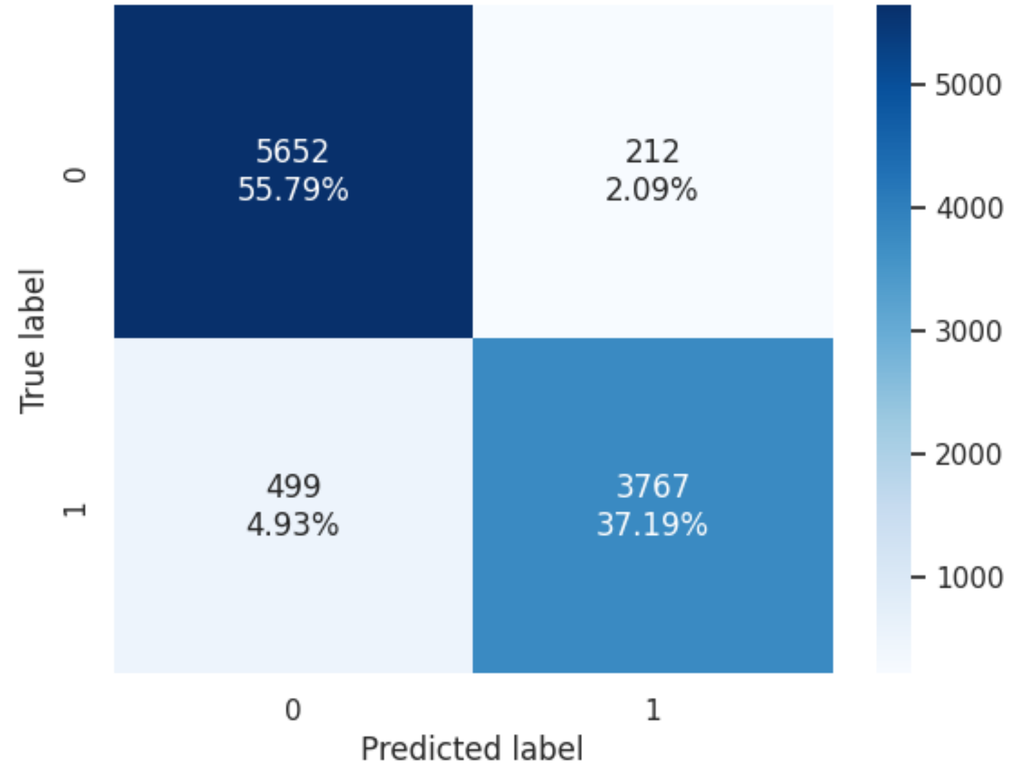
Pros:

- Easy to understand and implement.
- Doesn't make assumptions about the distribution of data.

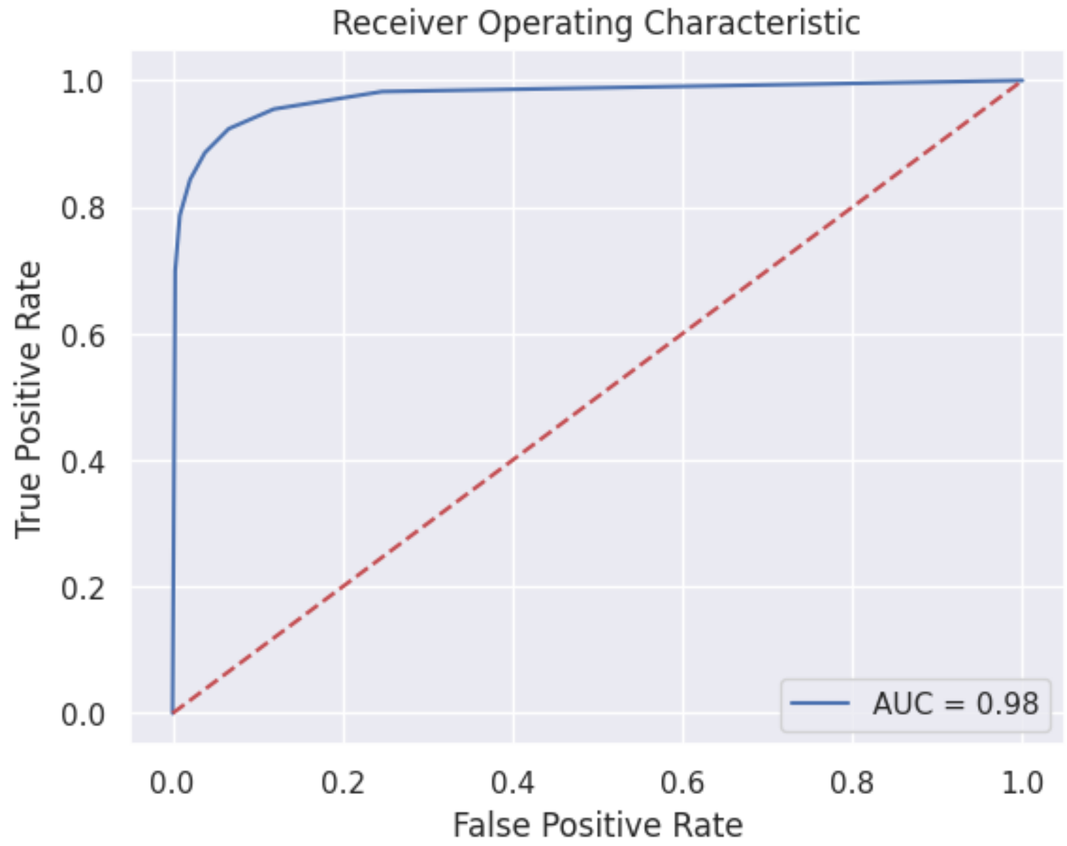
Cons:

- Sensitive to noise and outliers.
- Requires careful selection of the K value, and there's no specific rule for choosing K.

KNN ALGORITHM



KNN ALGORITHM

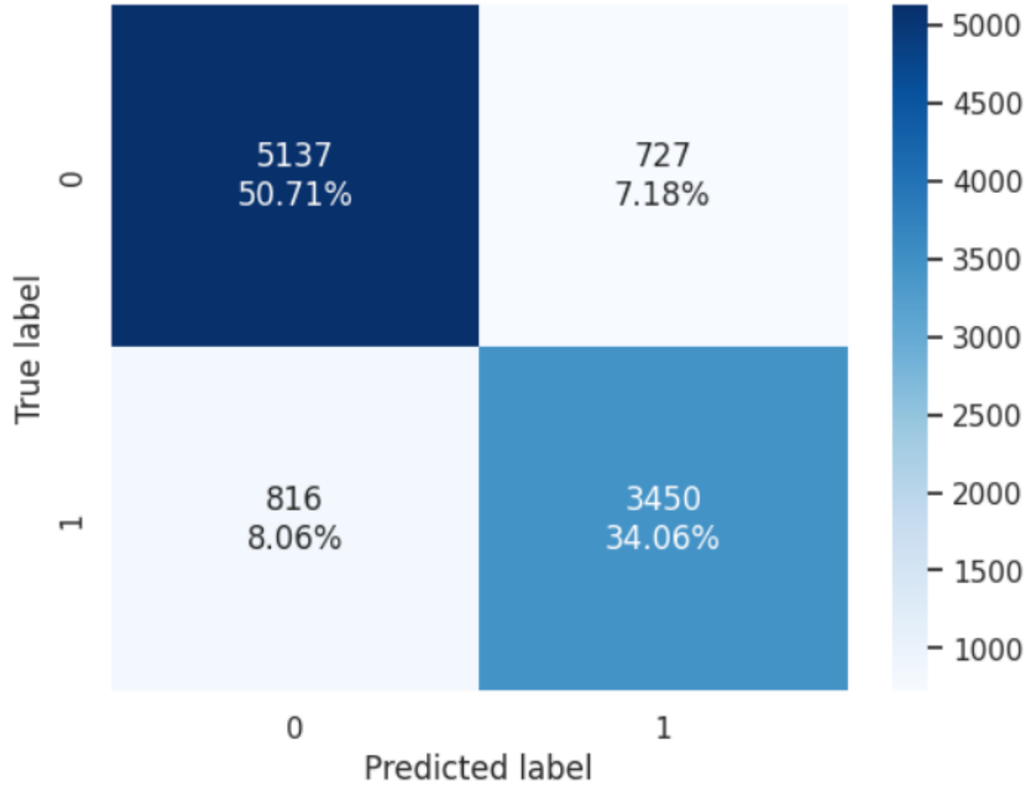


NAIVE BAYES ALGORITHM

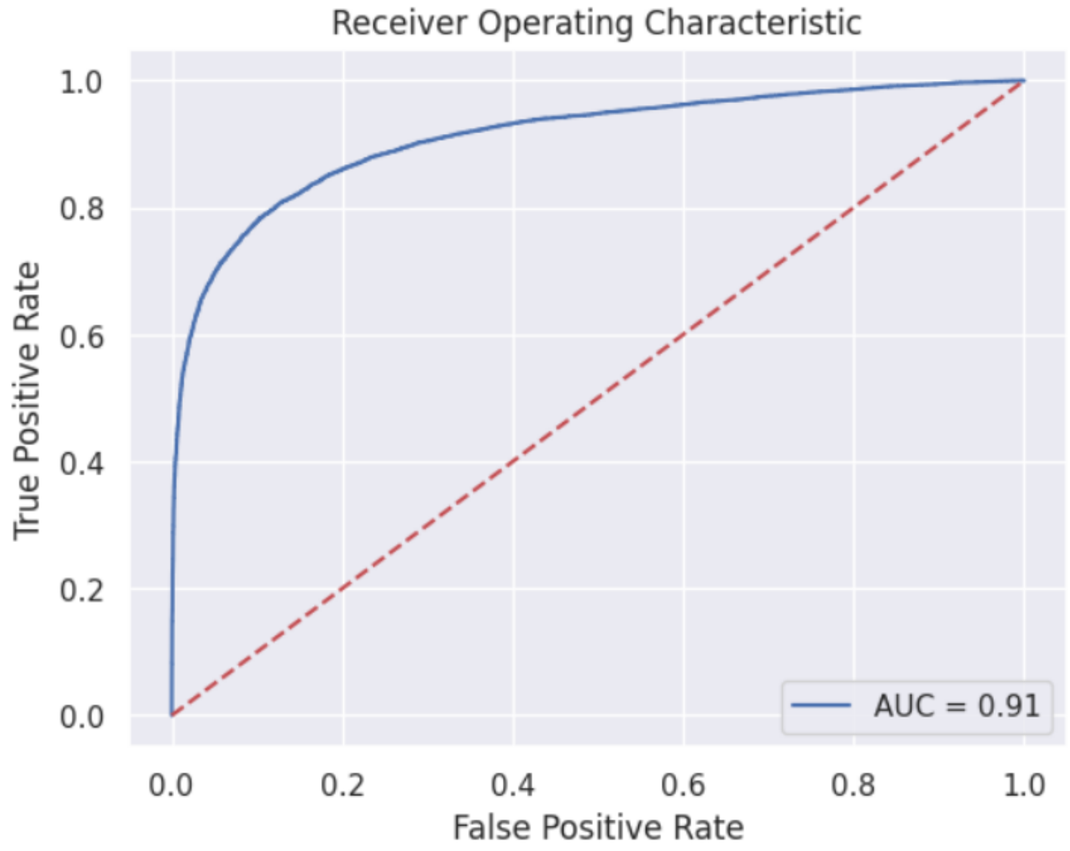
Naïve Bayes is a class of algorithmic analysis modeled on Bayesian reasoning in validation lists, used in many types of task assignment.

When independence is assumed, Naive Bayes performs better than other models such as logistic regression.

NAIVE BAYES ALGORITHM



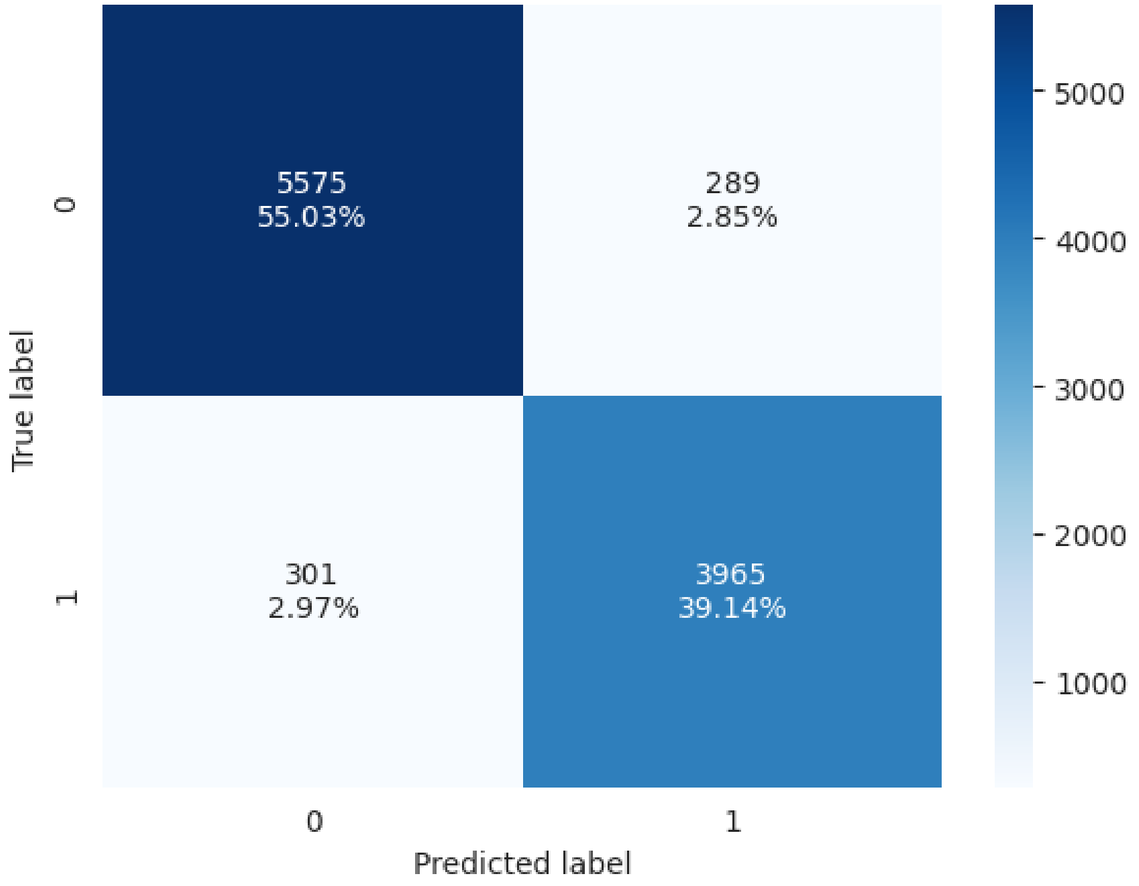
NAIVE BAYES ALGORITHM



DECISION TREE ALGORITHM

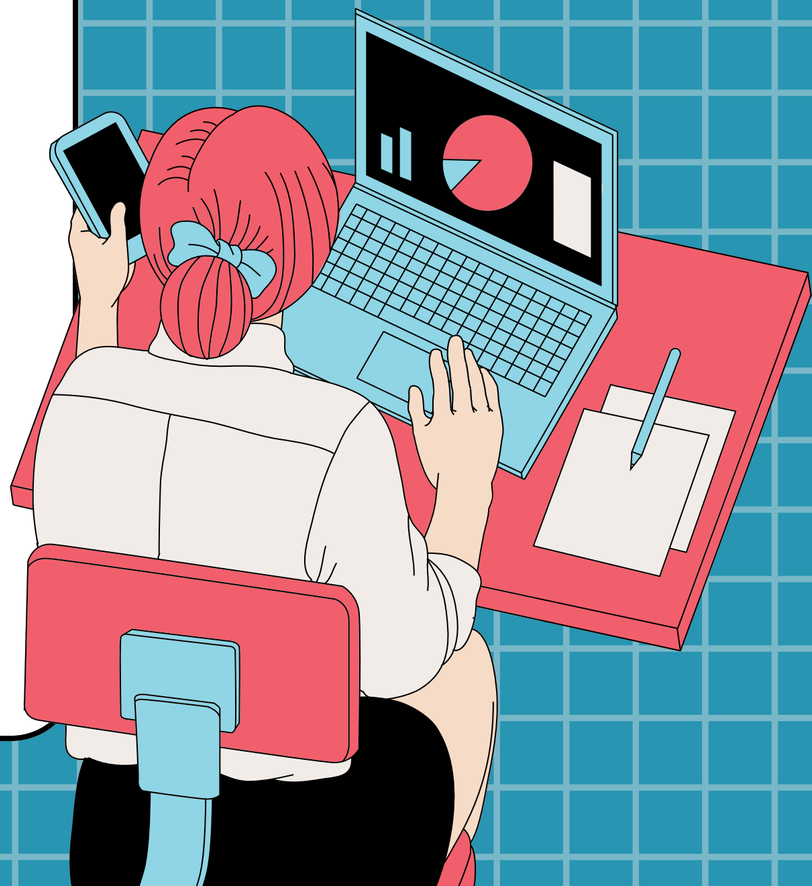
Decision tree, one of the most widely used algorithms in machine learning, is an explainable and white box algorithm that shows classification results using an if-then rule format [15]. It is used in both classification and regression problems. In a decision tree, every node symbolizes a feature, each branch indicates a rule, and every leaf represents a result, either a specific value or the continuation of further branching.

DECISION TREE ALGORITHM

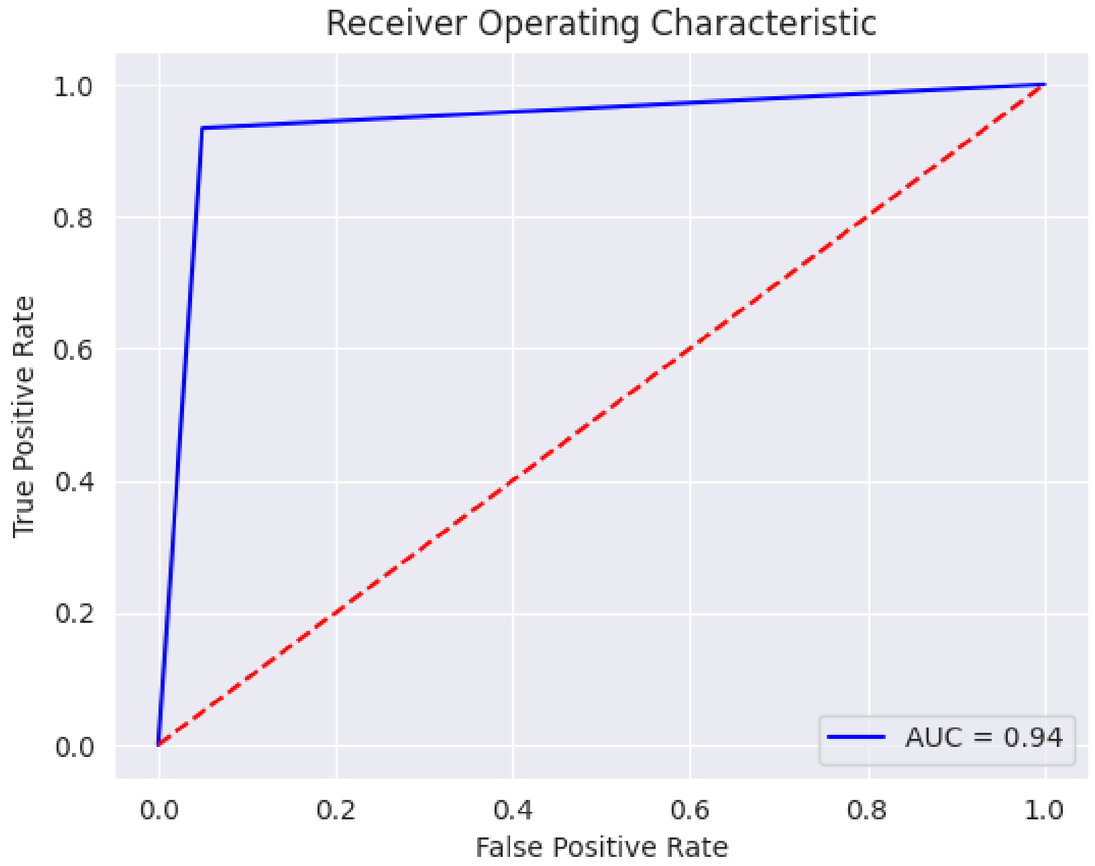


Predicted label

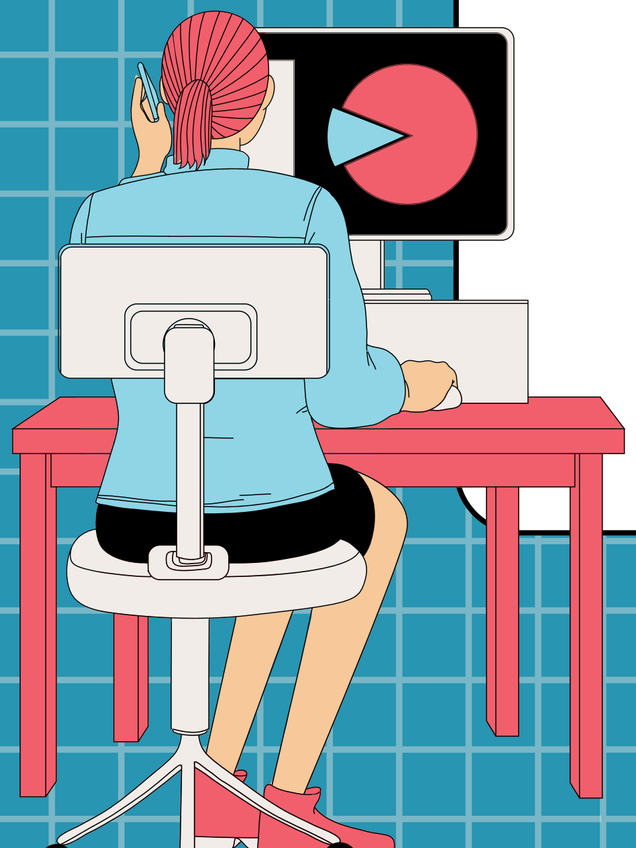
Accuracy=0.942
Precision=0.932
Recall=0.929
F1 Score=0.931



DECISION TREE ALGORITHM



5. EVALUATION



DISCUSSION

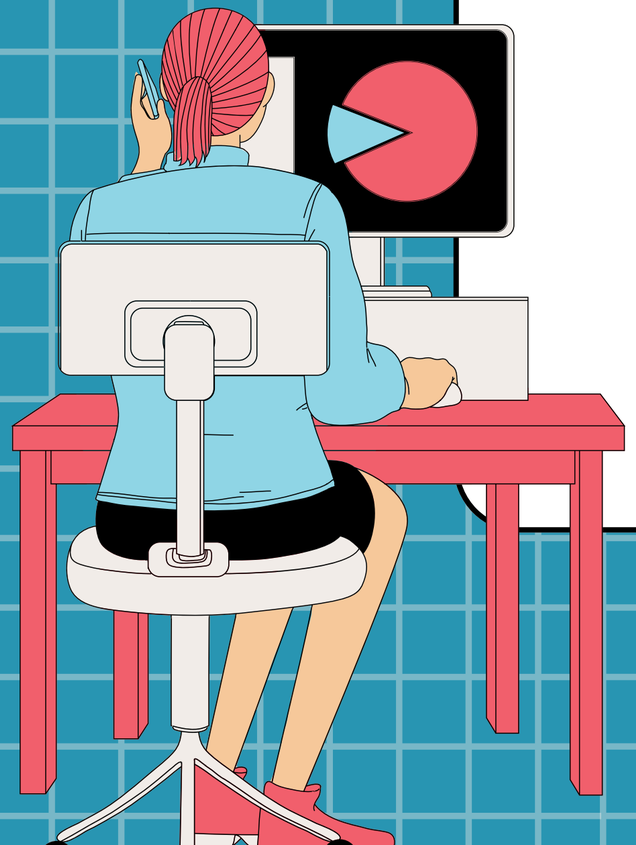
Results above 0.8 across KNN, SVM, Naïve Bayes, and Decision Tree indicate excellent performance, attributed to quality data and effective processing.

SVM emerges as the most promising model, excelling in high-dimensional datasets typical of air travel satisfaction analysis.

Naïve Bayes shows less favorable performance due to oversimplified assumptions in handling intricately interrelated factors.

Our machine learning exploration concludes with SVM as the optimal choice for predicting customer satisfaction in air travel, meeting our objectives.

6. CONCLUSION



CONCLUSION

Future research could explore advanced machine learning models and real-time data streams for enhanced accuracy. Additionally, integrating customer feedback loops and sentiment analysis tools would offer actionable insights for immediate service improvements.

Our ongoing commitment to leveraging cutting-edge technologies is crucial in meeting evolving passenger expectations and ensuring sustained customer loyalty.

THANK YOU