Predicting customer satisfaction

Introduction

An airline would like to build a model to predict whether customers will be satisfied with its services, given previous customer feedback about their flight experience. The marketing department would also like to know which features are the most important to customer satisfaction and clearly understand how the model makes decisions.

For these purposes, I suggest using the decision tree model, because it provides a set of rules that can be effectively communicated.

In this study, the following steps will be covered:

- Conduction of the basic EDA
- Preparing the data for modeling
- Building and evaluation of the decision tree model
- Hyperparameter tuning

Exploratory data analysis

To complete the previously stated task, airline passenger satisfaction survey data has been used. The data has already been divided into training and testing datasets. The training dataset consists of 25 columns and 103904 rows. The testing dataset consists of 25 columns and 25976 rows.

Columns represent the following data:

Gender: male or female

Customer type: regular or non-regular airline customer

Age: the actual age of the passenger

Type of travel: the purpose of the passenger's flight (personal or business travel)

Class: business, economy, economy plus

Flight distance

Inflight wifi service: satisfaction level with Wi-Fi service on board (0: not rated; 1-5)

Departure/Arrival time convenient: departure/arrival time satisfaction level (0: not rated; 1-5)

Ease of Online booking: online booking satisfaction rate (0: not rated; 1-5)

Gate location: level of satisfaction with the gate location (0: not rated; 1-5)

Food and drink: food and drink satisfaction level (0: not rated; 1-5)

Online boarding: satisfaction level with online boarding (0: not rated; 1-5)

Seat comfort: seat satisfaction level (0: not rated; 1-5)

Inflight entertainment: satisfaction with inflight entertainment (0: not rated; 1-5)

On-board service: level of satisfaction with on-board service (0: not rated; 1-5)

Leg room service: level of satisfaction with leg room service (0: not rated; 1-5)

Baggage handling: level of satisfaction with baggage handling (0: not rated; 1-5)

Checkin service: level of satisfaction with checkin service (0: not rated; 1-5)

Inflight service: level of satisfaction with inflight service (0: not rated; 1-5)

Cleanliness: level of satisfaction with cleanliness (0: not rated; 1-5)

Departure delay in minutes

Arrival delay in minutes

Let's check the data types.

```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 25 columns):
 #   Column                            Non-Null Count   Dtype  
---  ------                            --------------   -----  
 0   Unnamed: 0                        103904 non-null  int64  
 1   id                                103904 non-null  int64  
 2   Gender                            103904 non-null  object 
 3   Customer Type                     103904 non-null  object 
 4   Age                               103904 non-null  int64  
 5   Type of Travel                    103904 non-null  object 
 6   Class                             103904 non-null  object 
 7   Flight Distance                   103904 non-null  int64  
 8   Inflight wifi service             103904 non-null  int64  
 9   Departure/Arrival time convenient  103904 non-null  int64  
 10  Ease of Online booking            103904 non-null  int64  
 11  Gate location                     103904 non-null  int64  
 12  Food and drink                    103904 non-null  int64  
 13  Online boarding                   103904 non-null  int64  
 14  Seat comfort                      103904 non-null  int64  
 15  Inflight entertainment            103904 non-null  int64  
 16  On-board service                  103904 non-null  int64  
 17  Leg room service                  103904 non-null  int64  
 18  Baggage handling                  103904 non-null  int64  
 19  Checkin service                   103904 non-null  int64  
 20  Inflight service                  103904 non-null  int64  
 21  Cleanliness                       103904 non-null  int64  
 22  Departure Delay in Minutes        103904 non-null  int64  
 23  Arrival Delay in Minutes          103594 non-null  float64
 24  satisfaction                      103904 non-null  object 
dtypes: float64(1), int64(19), object(5)
memory usage: 19.8+ MB
```

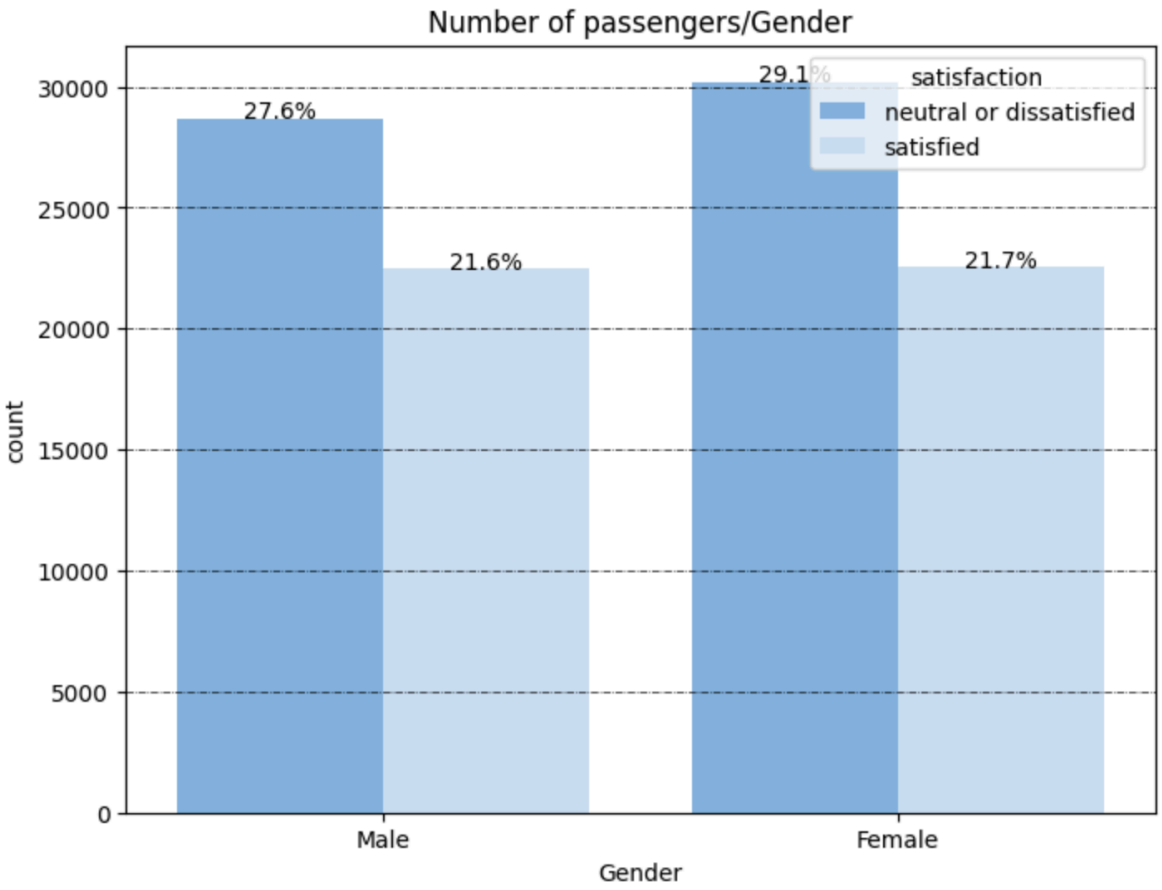
```
test_data.info()
```

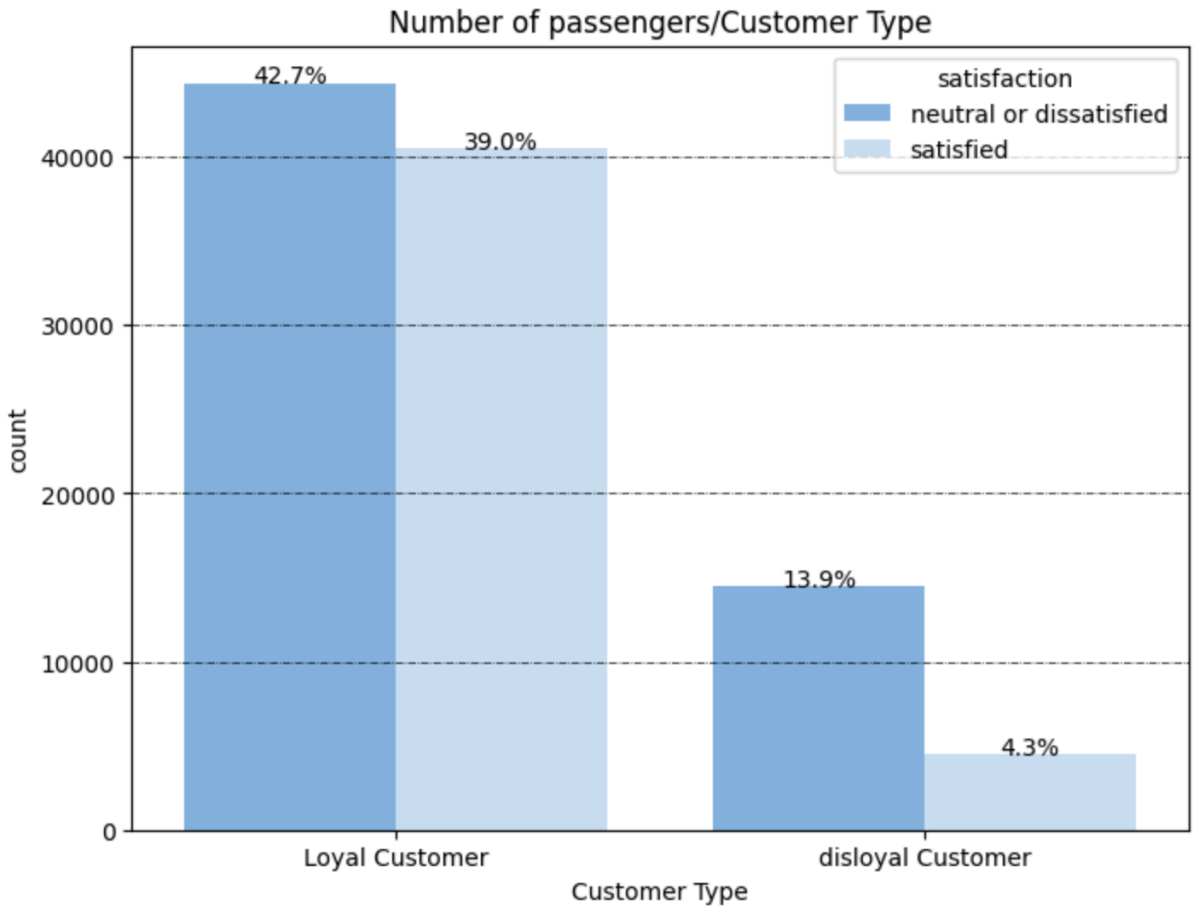
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25976 entries, 0 to 25975
Data columns (total 25 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   Unnamed: 0                        25976 non-null  int64  
 1   id                                25976 non-null  int64  
 2   Gender                            25976 non-null  object 
 3   Customer Type                     25976 non-null  object 
 4   Age                               25976 non-null  int64  
 5   Type of Travel                    25976 non-null  object 
 6   Class                             25976 non-null  object 
 7   Flight Distance                   25976 non-null  int64  
 8   Inflight wifi service             25976 non-null  int64  
 9   Departure/Arrival time convenient  25976 non-null  int64  
 10  Ease of Online booking            25976 non-null  int64  
 11  Gate location                     25976 non-null  int64  
 12  Food and drink                    25976 non-null  int64  
 13  Online boarding                   25976 non-null  int64  
 14  Seat comfort                      25976 non-null  int64  
 15  Inflight entertainment            25976 non-null  int64  
 16  On-board service                  25976 non-null  int64  
 17  Leg room service                  25976 non-null  int64  
 18  Baggage handling                  25976 non-null  int64  
 19  Checkin service                   25976 non-null  int64  
 20  Inflight service                  25976 non-null  int64  
 21  Cleanliness                       25976 non-null  int64  
 22  Departure Delay in Minutes        25976 non-null  int64  
 23  Arrival Delay in Minutes          25893 non-null  float64
 24  satisfaction                      25976 non-null  object 
dtypes: float64(1), int64(19), object(5)
memory usage: 5.0+ MB
```

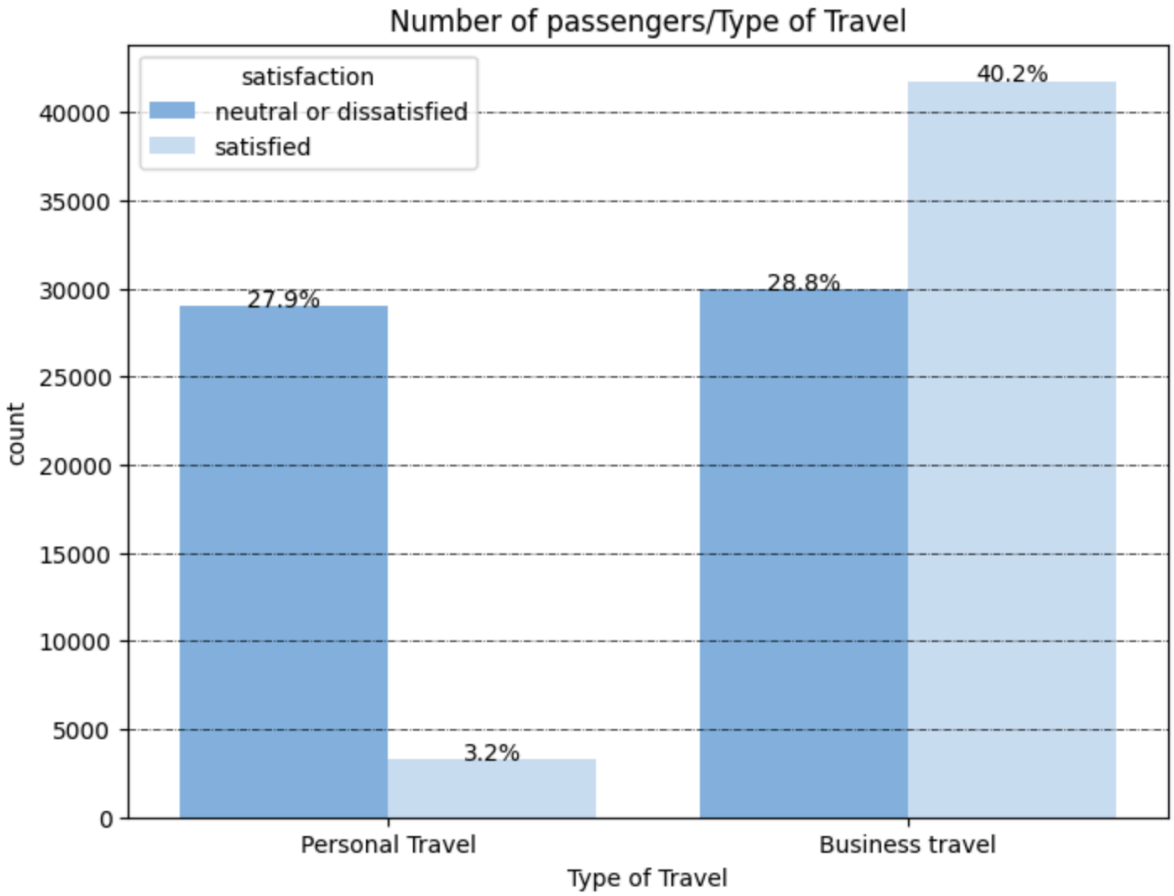
Most of the data is numerical, but there are five string-type columns ('Gender', 'Customer Type', 'Type of Travel', 'Class', 'Satisfaction').

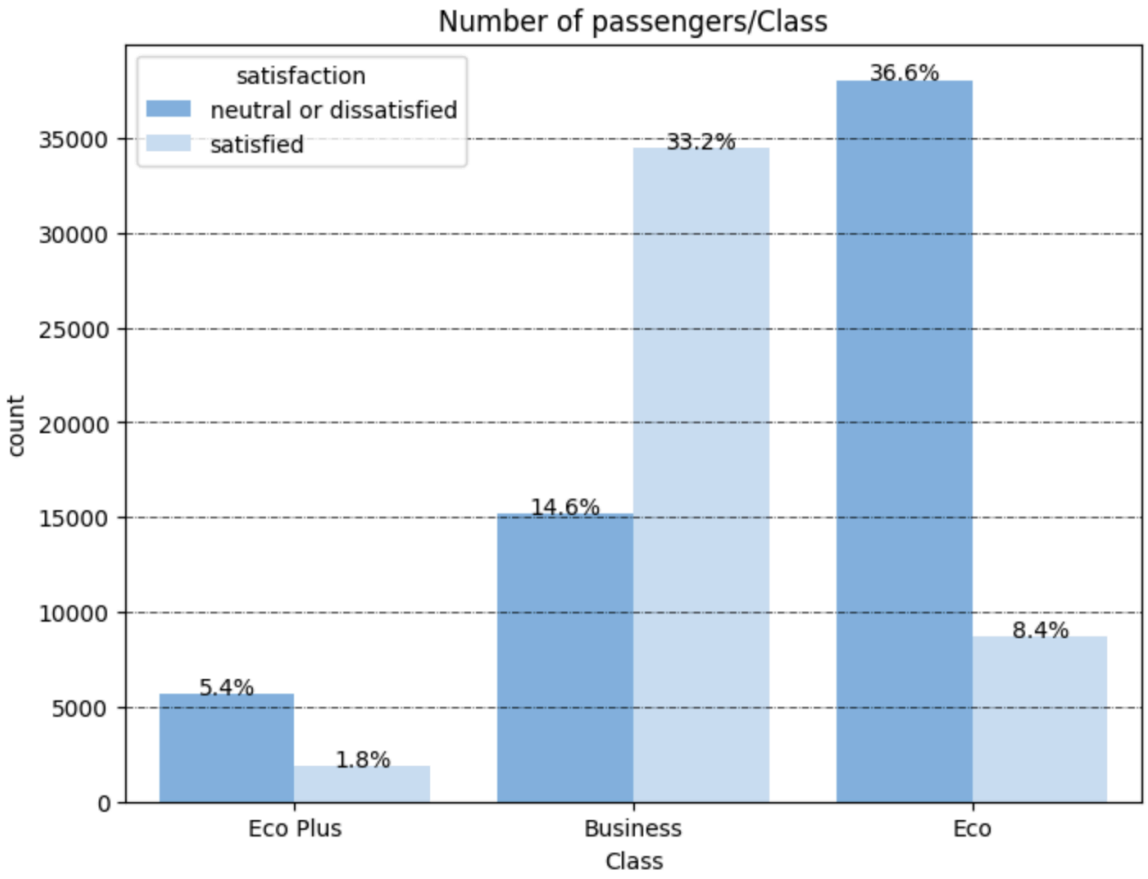
Also, we can observe missing values in the column 'Arrival Delay in Minutes' in both train and test datasets.

Let's examine the satisfaction levels for different genders, loyal and disloyal customers, personal and business travel, and classes.


Number of passengers/Gender
satisfaction
neutral or dissatisfied
satisfied
27.6%
21.6%
29.1%
21.7%
30000
25000
20000
15000
10000
5000
0
count
Male
Female
Gender


Number of passengers/Customer Type
satisfaction
neutral or dissatisfied
satisfied
42.7%
39.0%
13.9%
4.3%
40000
30000
20000
10000
0
count
Loyal Customer
disloyal Customer
Customer Type

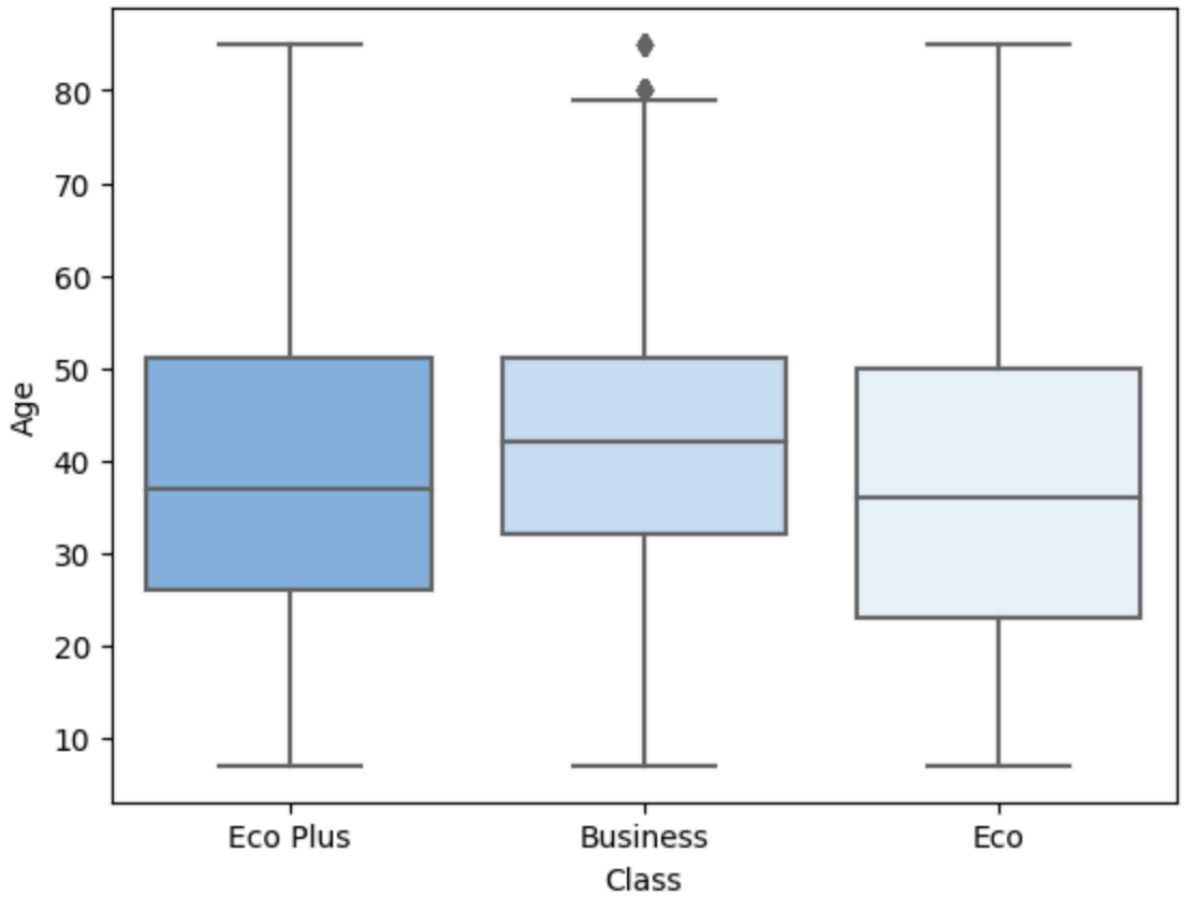

Number of passengers/Type of Travel
satisfaction
neutral or dissatisfied
satisfied
40000
35000
30000
25000
20000
15000
10000
5000
0
count
27.9%
3.2%
28.8%
40.2%
Personal Travel
Business travel
Type of Travel

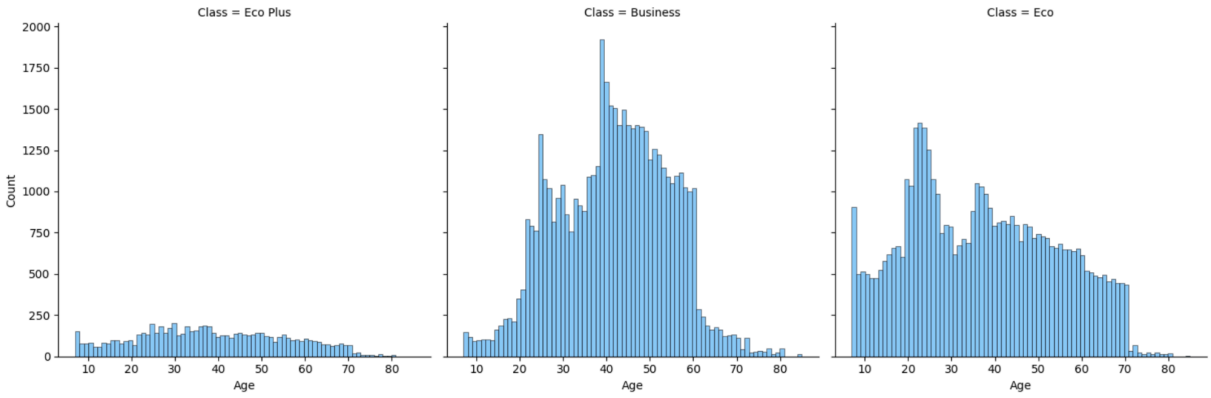


From the bar graphs, we can see the following :

- The ratio of satisfied and dissatisfied customers among male and female genders is similar.
- Among disloyal customers, there are more neutral or dissatisfied customers.
- Most customers who use airlines for personal travel are neutral or dissatisfied. At the same time, there are more satisfied customers among those who travel for business purposes.
- The vast majority of customers who travel Eco class are dissatisfied, and most customers who travel Business class are satisfied.

Let's take a look at the distributions of Age and Flight Distance.

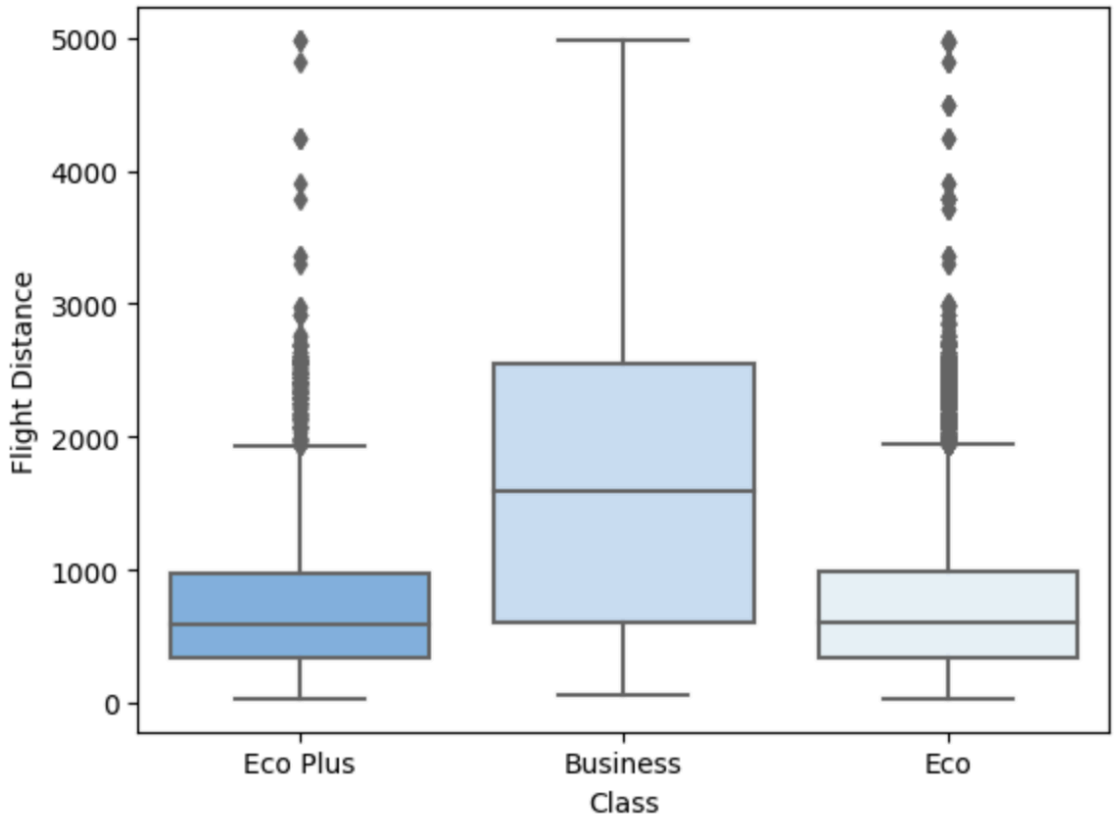


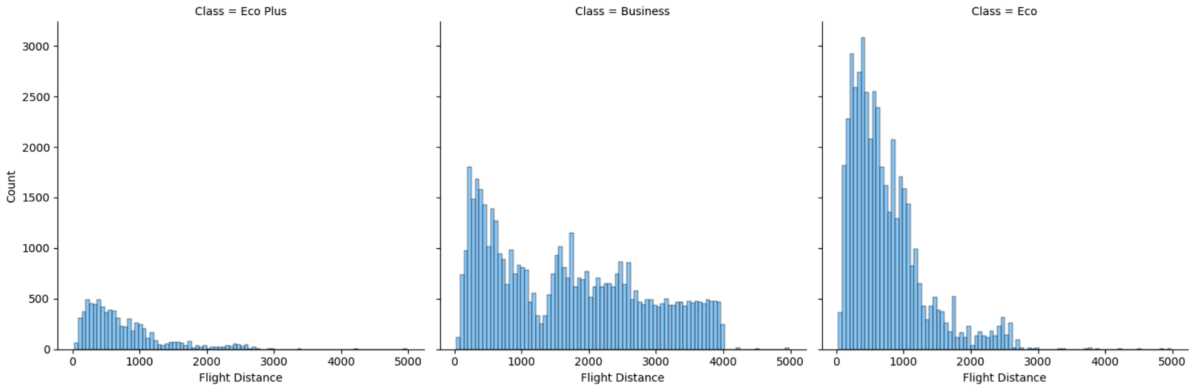


Median age of passengers in Eco Plus class: 37.0

Median age of passengers in Eco class: 36.0

Median age of passengers in Business class: 42.0



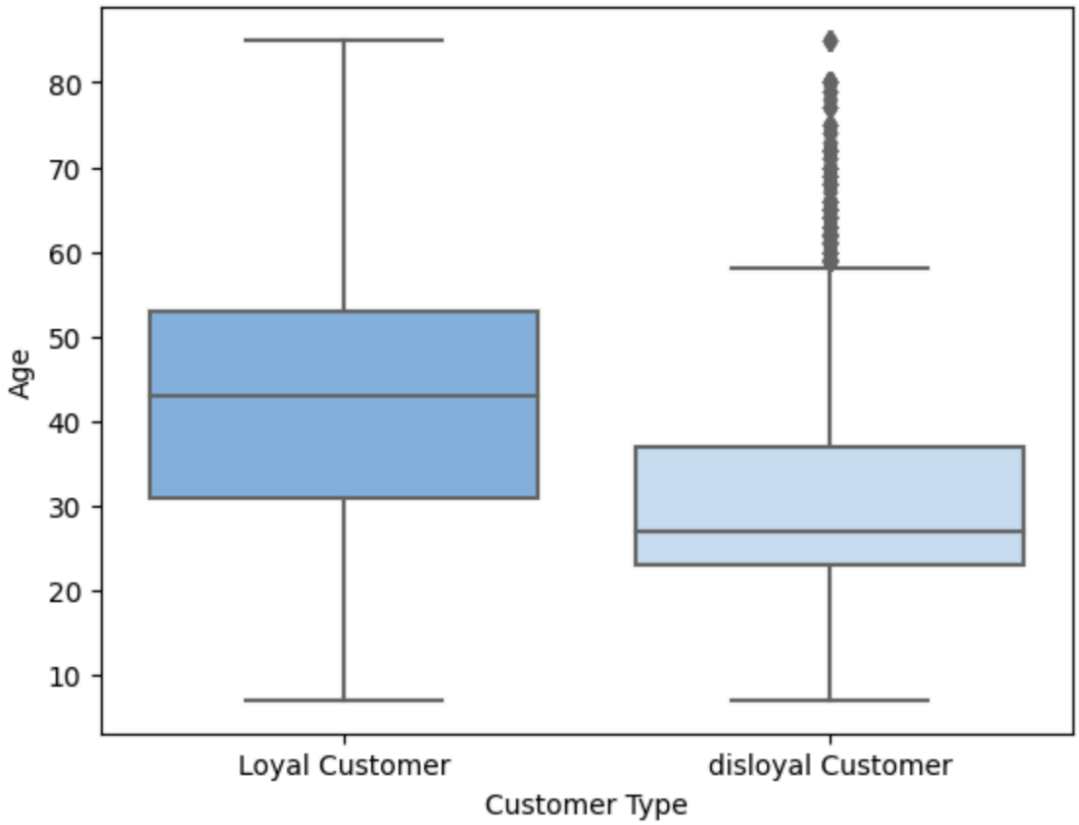


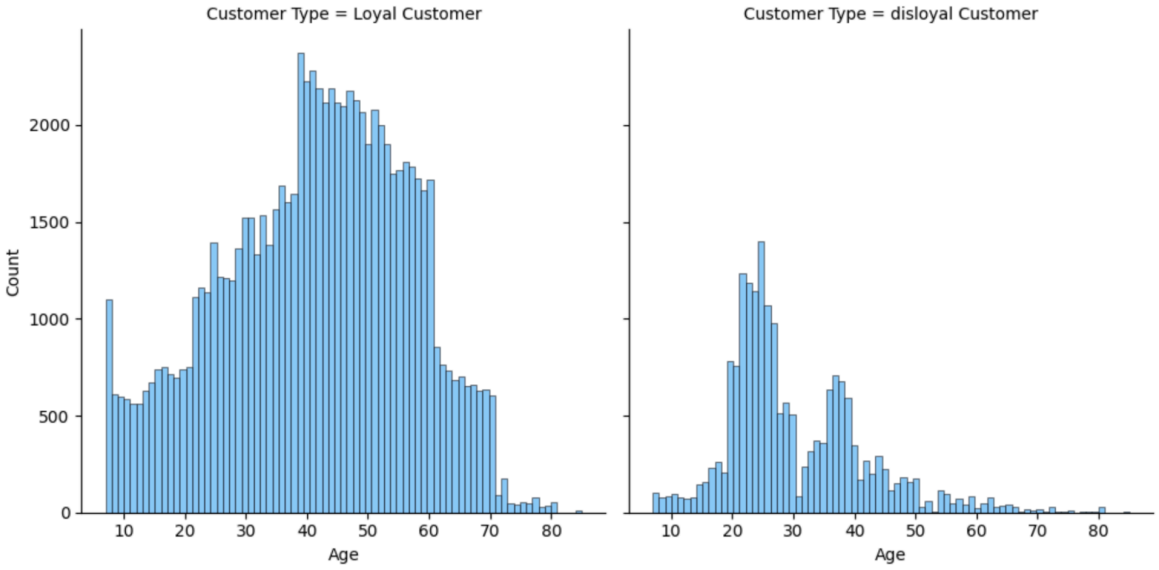
Median flight distance for Eco Plus class: 589.0

Median flight distance for Eco class: 599.0

Median flight distance for Business class: 1589.0

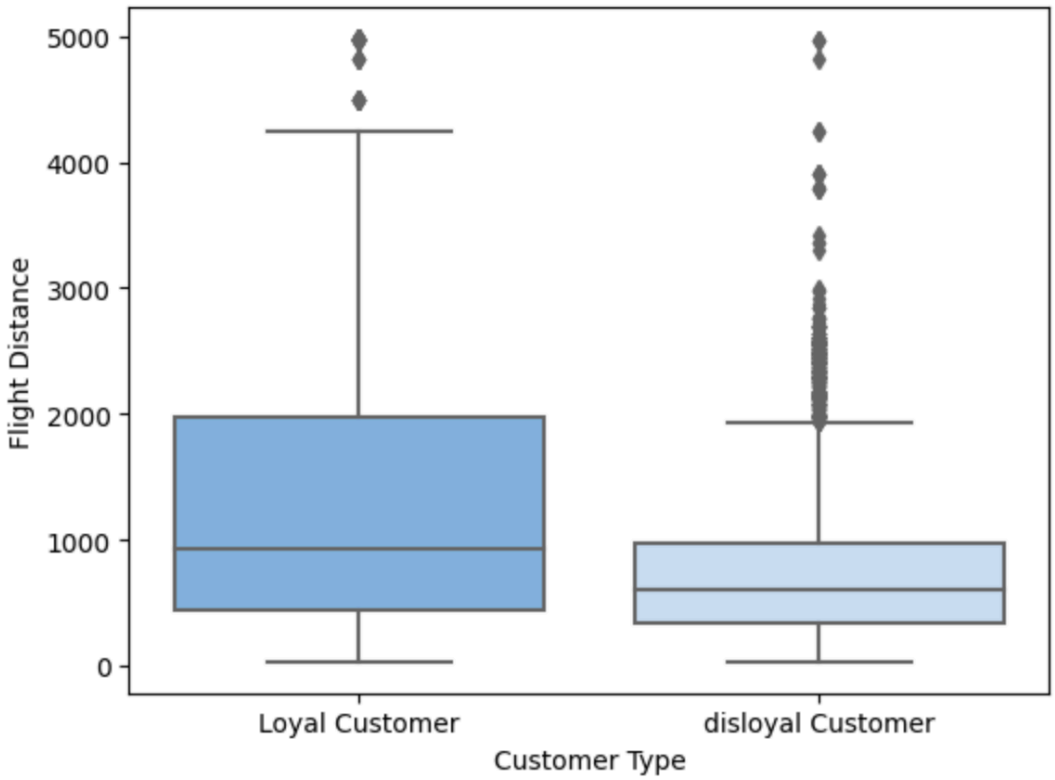
As we can see, the median flight distance of customers in Business class is much longer.

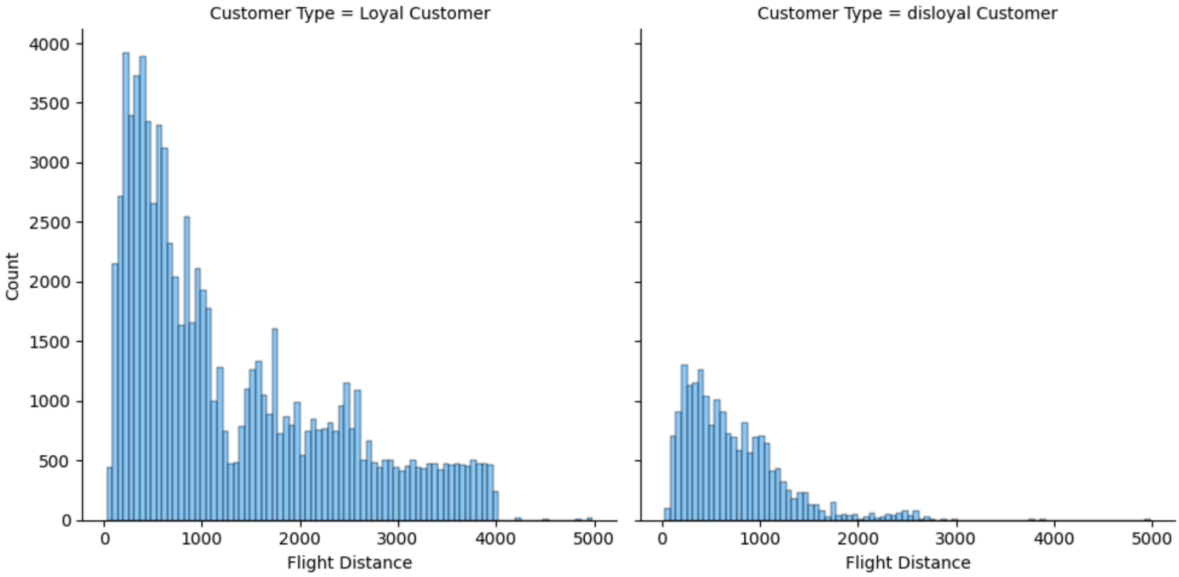




Median age of loyal customers: 43.0

Median age of disloyal customers: 27.0





Median flight distance for loyal customers: 925.0

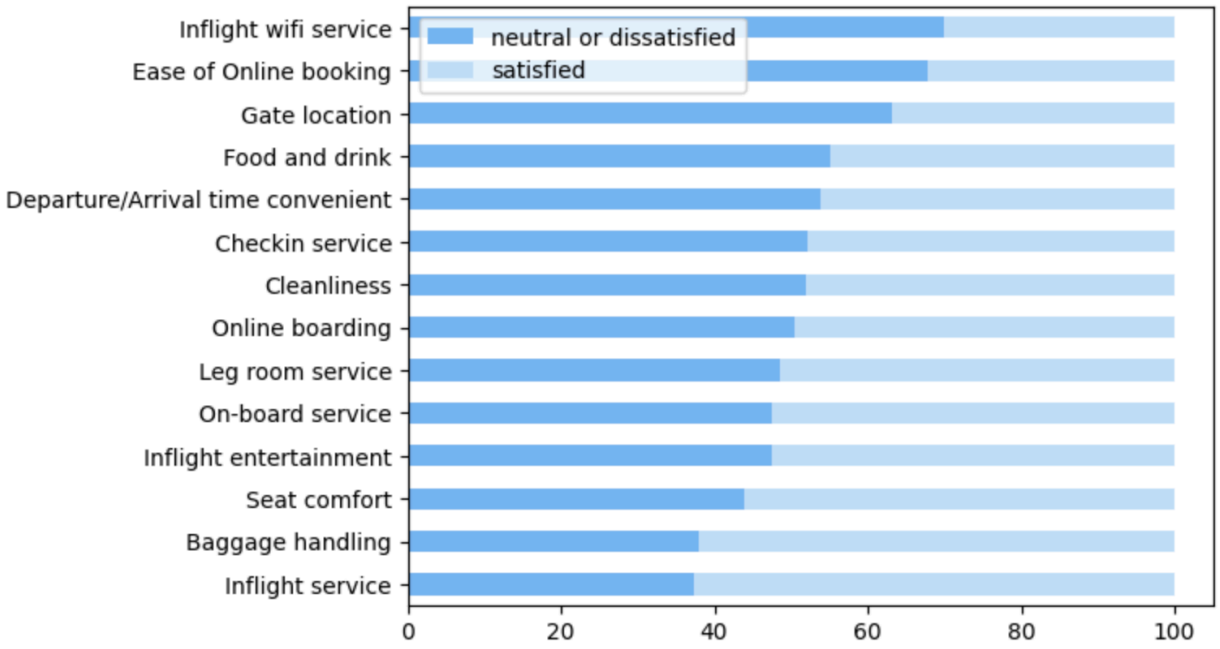
Median flight distance for disloyal customers: 598.0

Let's take a look at the satisfaction levels of different services.

To aggregate this information, let's label the level of satisfaction in the following way:

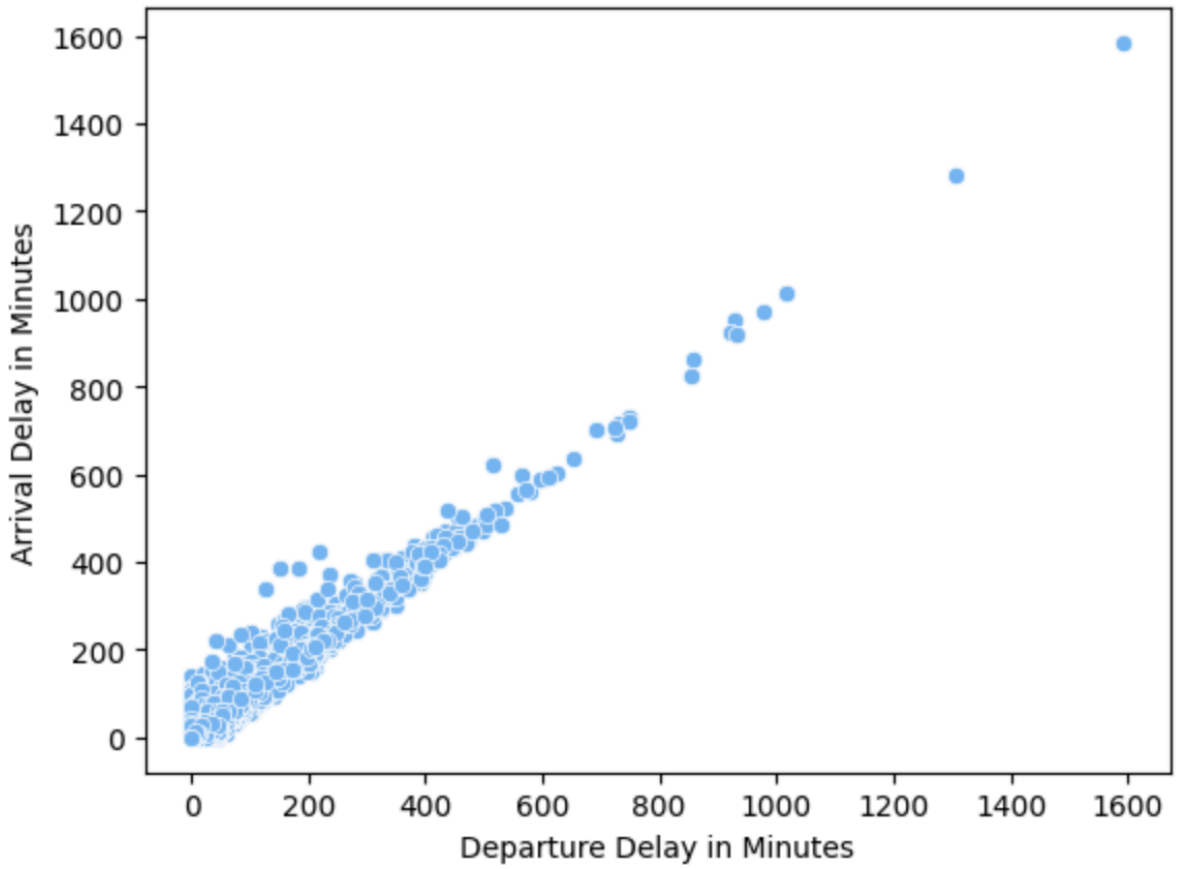
- less than 4 - neutral or dissatisfied;
- 4 and 5 - satisfied.

In the following graph, we can observe the ratio of satisfied and dissatisfied customers for different services:



Customers are more satisfied with baggage handling and inflight services and less satisfied with inflight wifi service, ease of online booking, and gate location.

Let's take a closer look at delays. The column 'Arrival Delay in Minutes' has null values. If there is a departure delay, there is also likely to be an arrival delay. Let's check this statement.



As we can observe from the scatterplot, arrival delay is linearly related to departure delay.

The correlation coefficient between these two variables is 0.97, which means there is a strong positive correlation between departure and arrival delays. Considering the linear relationship and strong positive correlation, I suggest imputing missing arrival delay values with departure delays.

Preparing the data for modeling

The first two columns (Unnamed: 0, id) are useless for prediction, so we must drop them.

Let's check how many unique values we have in categorical variables.

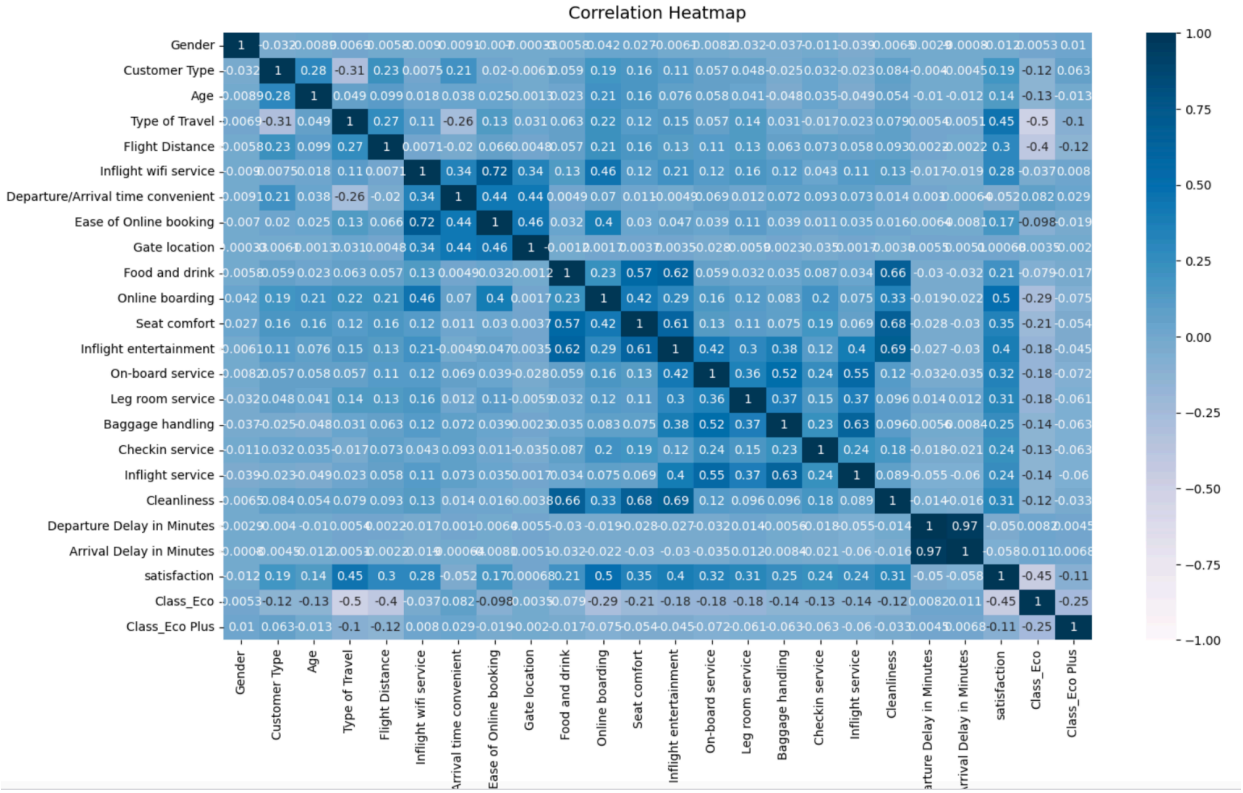
	Number of unique
Gender	2
Customer Type	2
Type of Travel	2
Class	3
satisfaction	2

Label encoding has been provided for the features 'Gender,' 'Customer Type,' 'Type of Travel', and the target variable 'satisfaction.' Since the feature 'Class' has more than two unique values, one-hot encoding has been provided in this case.

Let's check data types one more time before modeling.

```
class 'pandas.core.frame.DataFrame'>
angeIndex: 103904 entries, 0 to 103903
ata columns (total 24 columns):
#   Column                            Non-Null Count   Dtype
--  ------                            --------------   -----
0   Gender                            103904 non-null  int64
1   Customer Type                     103904 non-null  int64
2   Age                               103904 non-null  int64
3   Type of Travel                    103904 non-null  int64
4   Flight Distance                   103904 non-null  int64
5   Inflight wifi service             103904 non-null  int64
6   Departure/Arrival time convenient  103904 non-null  int64
7   Ease of Online booking            103904 non-null  int64
8   Gate location                     103904 non-null  int64
9   Food and drink                    103904 non-null  int64
10  Online boarding                   103904 non-null  int64
11  Seat comfort                      103904 non-null  int64
12  Inflight entertainment            103904 non-null  int64
13  On-board service                  103904 non-null  int64
14  Leg room service                  103904 non-null  int64
15  Baggage handling                  103904 non-null  int64
16  Checkin service                   103904 non-null  int64
17  Inflight service                  103904 non-null  int64
18  Cleanliness                       103904 non-null  int64
19  Departure Delay in Minutes        103904 non-null  int64
20  Arrival Delay in Minutes          103904 non-null  float6
21  satisfaction                      103904 non-null  int64
22  Class_Eco                         103904 non-null  int64
23  Class_Eco Plus                    103904 non-null  int64
types: float64(1), int64(23)
emory usage: 19.0 MB
```

Now, I would like to check the correlation matrix to see which variables strongly correlate with the target variable, 'satisfaction'.



The top five highest correlation coefficients have the following features:

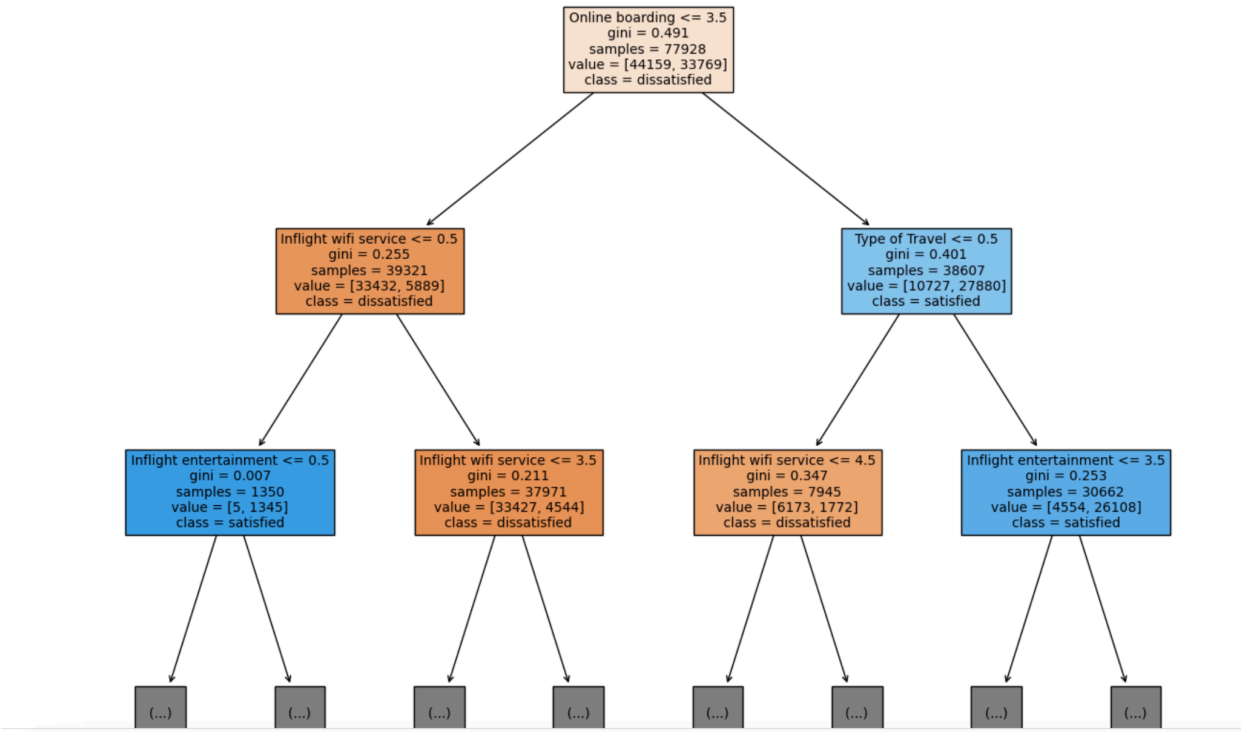
- Online boarding (0.5)
- Type of Travel (0.45)
- Inflight entertainment (0.4)
- Seat comfort (0.35)
- On-board service (0.32)

Interestingly, the correlation coefficient between the target variable and 'Class Eco' is -0.45, meaning there is a negative correlation between these two variables.

Building and evaluating the model

Data was divided into features, target variables, and train and validation sets.

Let's begin with building the decision tree with default parameters using the scikit-learn library. Here, we can take a look at the first splits of the tree:



The following table represents the decision tree model's accuracy, precision, recall, and F1 score.

	Accuracy	Precision	Recall	F1 score
Decision tree default	0.942	0.930	0.938	0.934

Let's proceed with hyperparameter tuning to prevent our model from overfitting the training data.

The following parameters were used to conduct grid search:

```
tree_parameters = {'max_depth':[2,3,4,5,6,7,8,9,10,12,14,16,18,20,30,40,50],
                   'min_samples_leaf': [2,3,4,5,6,7,8,9,10,15]}
```

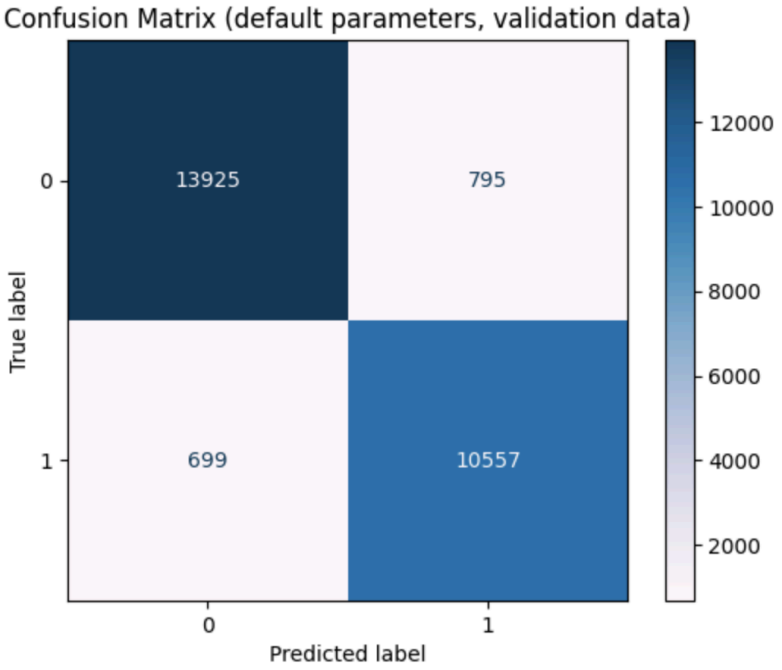
In the following image we can see the best parameters.

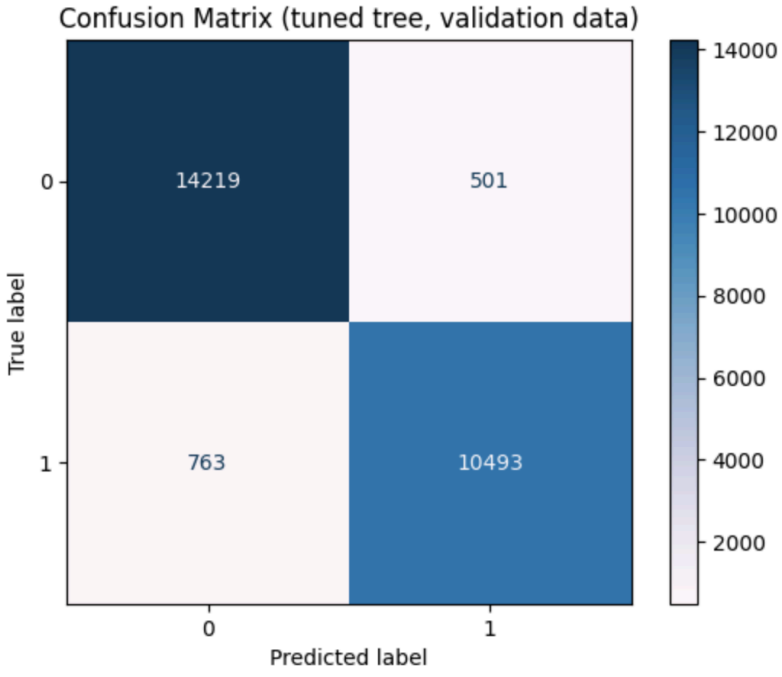
```
▾ DecisionTreeClassifier
DecisionTreeClassifier(max_depth=16, min_samples_leaf=5, random_state=0)
```

We can compare the results in the following table:

	Accuracy	Precision	Recall	F1 score
Decision tree default	0.942	0.930	0.938	0.934
Decision tree tuned	0.951	0.954	0.932	0.943

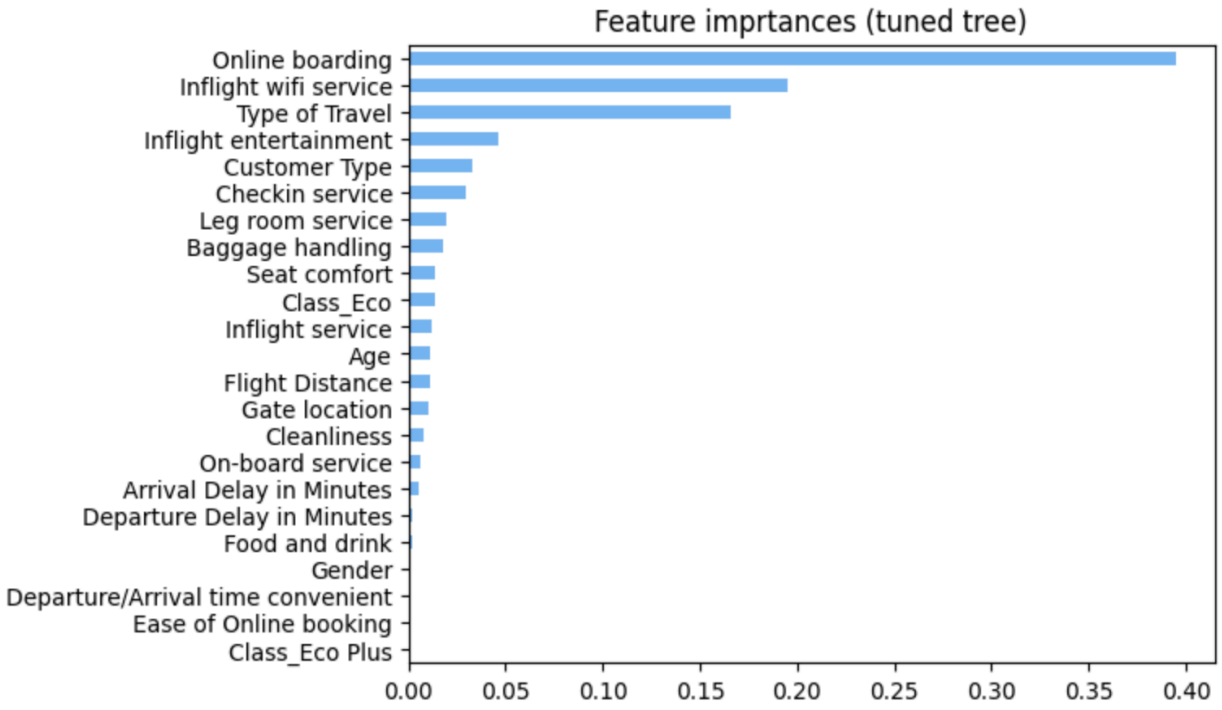
Here are the confusion matrices of the two decision trees:





As we can see, the overall performance of the tuned model is slightly better.

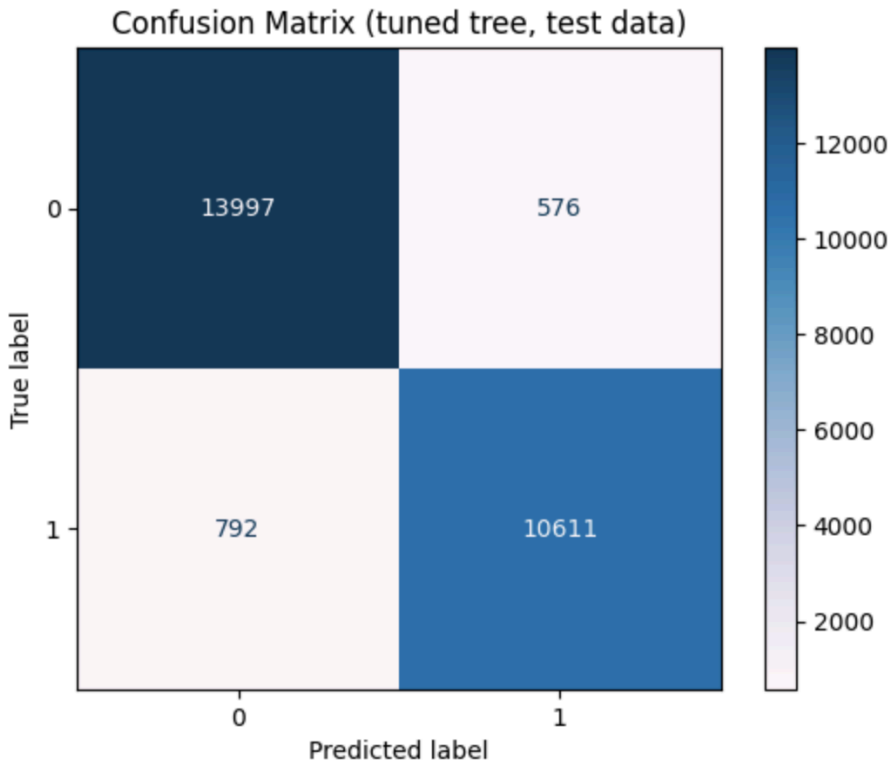
Let's take a look at the feature importances of the tuned tree:



Interestingly, the top five important features are not exactly the same as those with the highest correlation coefficients. This fact underlines the decision trees' ability to discover hidden patterns corresponding to complex interactions in the data.

Let's check the tuned decision tree model on the testing dataset.

	Accuracy	Precision	Recall	F1 score
Decision tree default	0.942	0.930	0.938	0.934
Decision tree tuned	0.951	0.954	0.932	0.943
Tuned tree test	0.947	0.949	0.931	0.939



The results for the testing data are only slightly worse than those for the validation data.

Conclusion

The decision tree model has been built to predict customer satisfaction. For the testing data, there are the following evaluation metrics:

- accuracy: 94.7% of data points were correctly classified;
- precision: 94.9% of positive predictions are true positives;
- recall: 93.1% of actual positives were correctly classified;
- f1 score combines precision and recall into a single expression, giving each equal importance: 93.9%.

All these metrics are helpful. Accuracy is an overall representation of model performance. Precision will be a good metric for stakeholders to avoid falsely claiming a customer is satisfied. Assuming a customer is happy when they are not might lead to customer churn. The airline also might want to limit false negatives. For this purpose, recall is useful. Assuming that actually satisfied people are unsatisfied can lead to the airline wasting resources trying to improve the customer experience of an already happy customer.

The most critical features represented in the given dataset are ‘Online boarding’, ’Inflight wifi service’, and ‘Type of Travel’. We can observe this from the decision tree visualization and the feature importance graph.

Passenger satisfaction survey data showed that customers are mostly unsatisfied with inflight wifi service, ease of online booking, and gate location.

The company should take steps to improve the experience for people who travel for personal purposes because data shows a shallow level of satisfaction. The same applies to the Eco class passengers, who are mostly unsatisfied with airline services.