Airline Passenger Satisfaction

Sri Krishna Tirthala | Anik Alvi | Rahul Garigipati

This project deals with the Airline Passenger reviews data. When a passenger takes a particular flight, every airline company prefers to take the review of inflight services from the passenger, which helps the company to improve its services if the reviews provided by the passenger are bad. There is a high chance of losing their customers if they don't act upon their bad services. Also, few airline companies use the data from their passengers and find which particular service is better among all the services they provide and might want to advertise to public about it.

Motivation:

The motivation for this project is to apply the Data mining algorithms on the available raw data of airline passenger reviews and use it for different purposes. The airline companies might want to compensate for the passengers who faced bad experience by giving them discounts for their next trip and also thank passengers who gave good reviews for the services by offering them discount on their stay on their holiday trip.

Problem Definition:

What's the problem here?

To compensate passengers for their bad experiences or to thank them for giving good reviews, the company has to know which passenger belongs to which group. If we divide the passengers into two different groups, i.e., satisfied customers and dissatisfied customers. So, we need to build something that classifies the passengers into two different groups.

What's the solution to the problem?

The classification algorithms available in the data science will be a solution by helping the companies to classify their passengers based on their reviews. There are various classification algorithms available that can be used but different algorithms perform differently based on the kind of dataset that they are being implemented on.

So, few of the algorithms can be implemented and one of them can be picked from those which gives the best accuracy rate.

Possible Data Mining Techniques:

Classification techniques can be used as main purpose of classifying the users and act accordingly. Also, clustering techniques can be implemented and plot the data based on the class (Economy, Business, Economy Plus) and show the review of services provided among various classes.

Dataset:

We took this dataset from the Kaggle website and the link is provided below:

https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction?select=test.csv

There are two datasets available (Train.CSV and test.CSV) train.CSV to train the classifying model and test.CSV to test the model. The dataset has 25 different columns including the class label that has to be predicted. Training dataset has 1,03,905 rows of data and the testing dataset has 25,977 rows of data.

Data Analysis:

Using the data, the analysis has been done to find out the relation between various columns and compared how the columns are related.
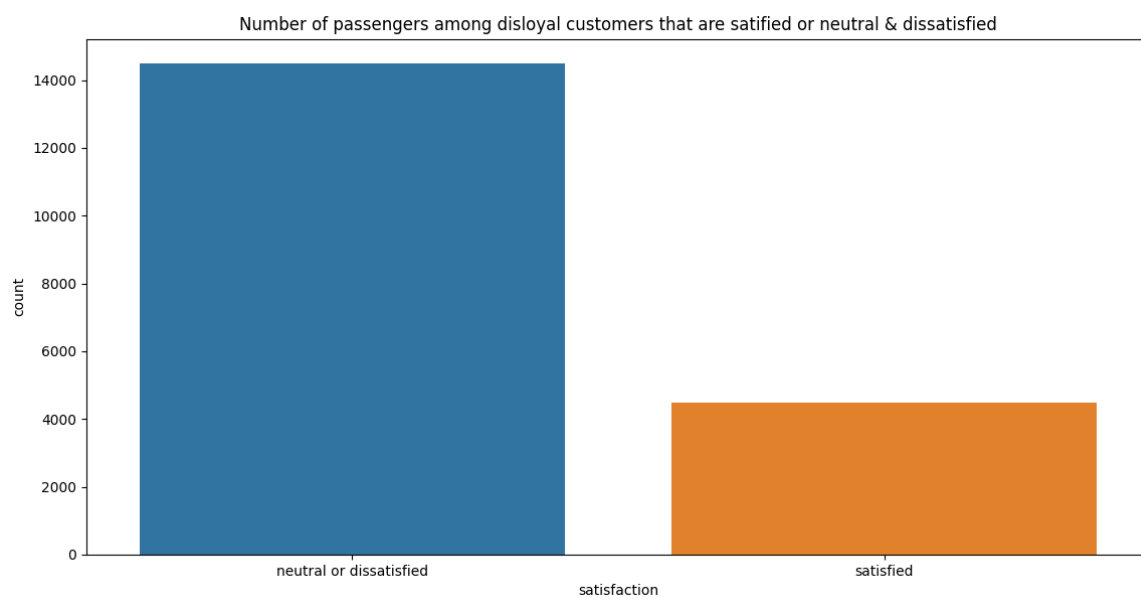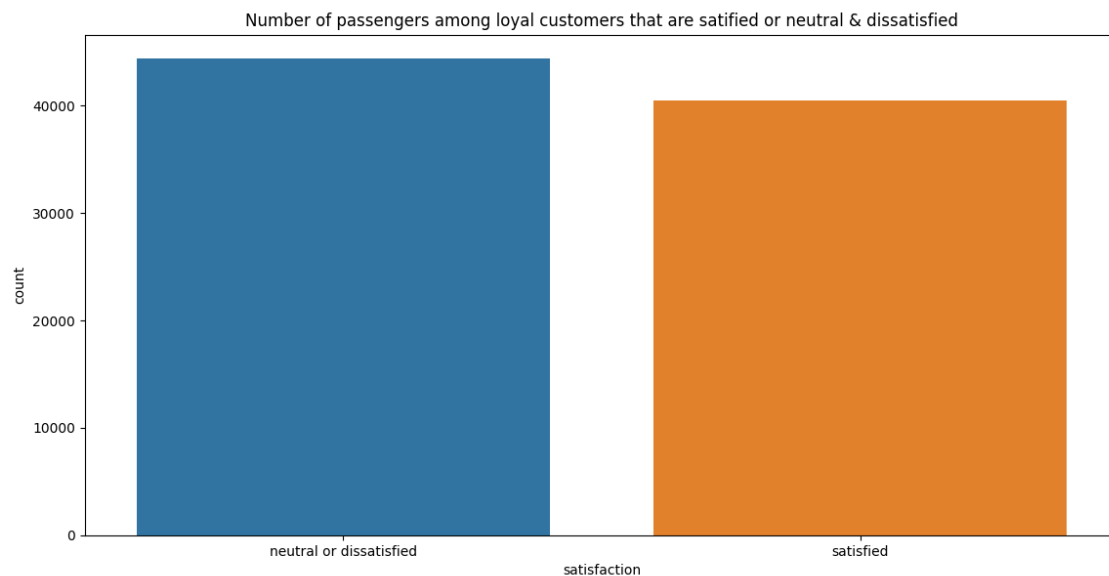
The below is the plot to show the count of passengers that are classified into 'satisfied' and 'neutral or dissatisfied'.

Number of passengers that are satified or neutral & dissatisfied
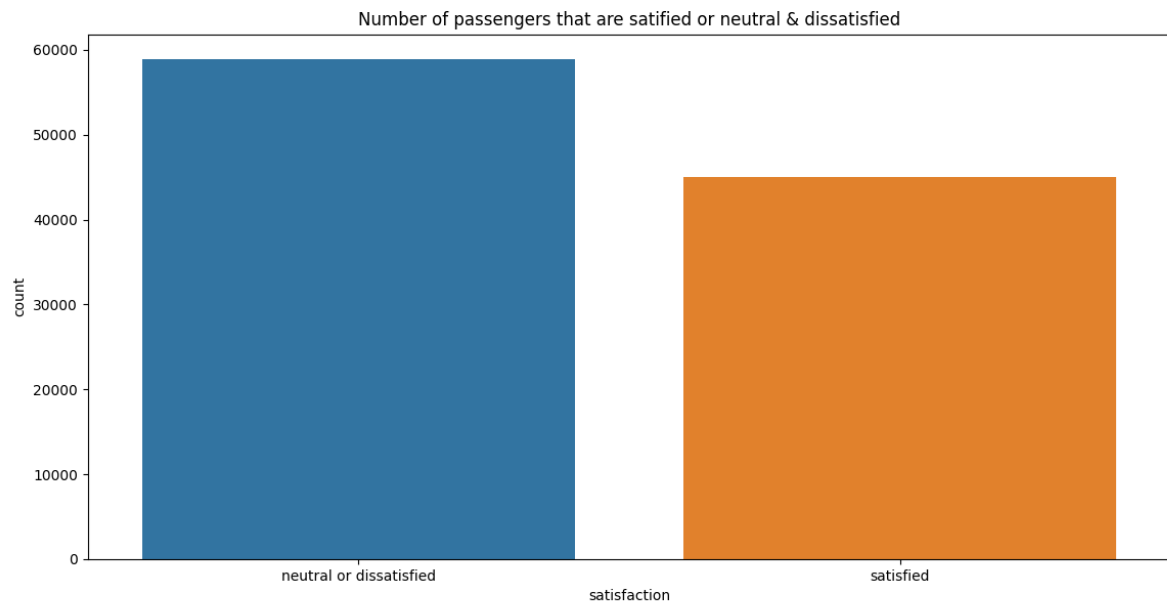
The below is the plot to show the count of passengers that preferred to travel in three available classes of the flight that are Business, Eco Plus and Economy.
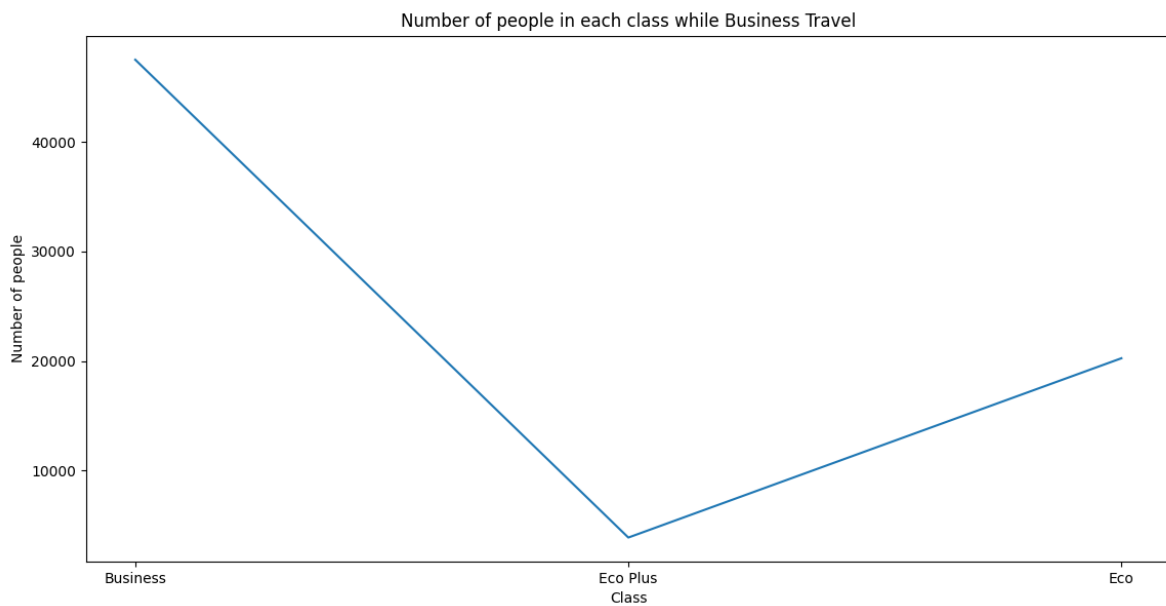

Number of passengers in each Class
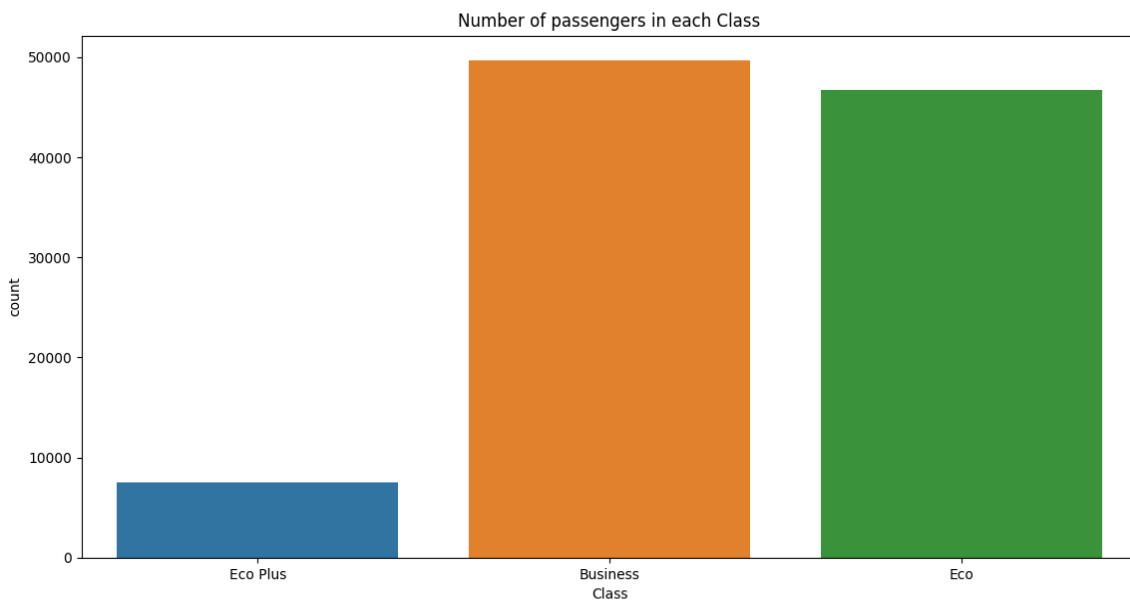
For the below plots, we divided the data into two different dataframes of loyal customers data and disloyal customers data and checked how many of them are into satisfied label and how many of them are into neutral and dissatisfied label.

Number of passengers among loyal customers that are satified or neutral & dissatisfied

Number of passengers among disloyal customers that are satified or neutral & dissatisfied
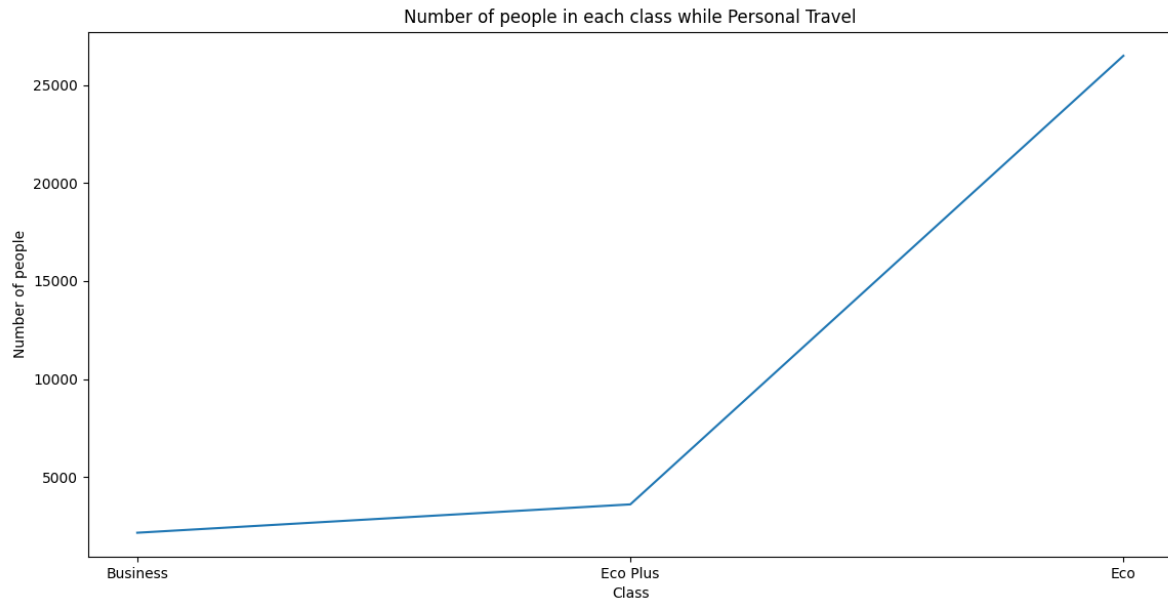

The below graph is to show the overall rating given by each passenger from three different classes. Its through taking the mean of all services ratings and again calculating the mean of all means for each class and seems like business class have provided better services when compared among all three classes.

Comparison of class-vise overall rating
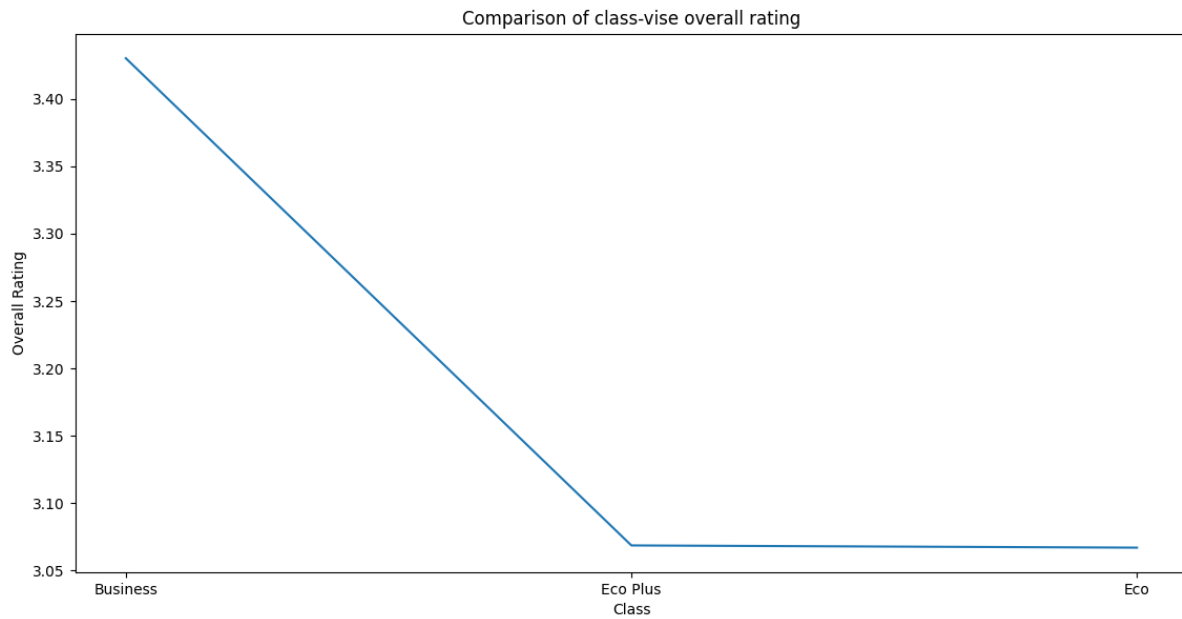
The next two plots are to show how many passengers preferred to travel in what classes based on the purpose of their travel. From the below plots we can tell that, while travelling on business purpose, most of the passengers preferred to travel in business class and while travelling on person reasons, most of them chose to travel in Economy class.

Number of people in each class while Business Travel

Number of people in each class while Personal Travel

Data Pre-processing:

One of the most important tasks in every machine learning task is the data pre-processing section. In our project, few attributes had missing values. To resolve this problem, we replaced these missing values by the average value of that attribute since the real value is not known.

Classification Approaches:

Our project is a classification problem where the model needs to predict whether the customer is satisfied with the airline service or not. Looking into classification algorithms, we implemented logistic regression and k-nearest neighbour classifier algorithms to find how well the algorithms perform. Table below shows the accuracy, precision and recall values obtained from each of these algorithms.

	Accuracy	Precision	Recall
Logistic Regression	0.8269	0.8266	0.8269
KNN	0.915	0.915	0.915

Based on the table above, KNN performs better than Logistic regression for accuracy, precision as well as recall value. At this stage of the project, we have these results but we want to work on few other algorithms and know which gives the better accuracy rate and know what best suits this statistical data.