

GROUP 3

Capstone Project – Interim Report

Airline Passenger Satisfaction Prediction

Mentor: Srikar Muppidi

* **Akshaya ES**
* **Nikita Ann George**
* **Pala Sheshu**
* **Ujjwal Kumar Sharma**
* **Venkat Paritala**
* **Viresh Raj Sah**

# PROBLEM STATEMENT

Predicting customer satisfaction based on their holistic experience while traveling by air.

# PROJECT OUTCOME

For this project, we intend to explore survey results provided by passengers (anonymized), and understand what causes them to be satisfied or dissatisfied. Ultimately, we hope our findings enable us to make recommendations on how to address the factors that cause dissatisfaction.

# BUSINESS IMPACT

# Airline businesses around the world are decimated by Covid-19 as most international air travel has been grounded. In fact, some airlines such as Thai Airways have already filed for bankruptcy. Nonetheless, once the storm is over, demand for air travel is expected to surge as people rush back for overseas holidays. What can airlines prepare to give themselves a competitive edge when the crowd finally arrives?

Customer satisfaction is always top of mind for airlines. Unhappy or disengaged customers naturally mean fewer passengers and less revenue. It’s important that customers have an excellent experience every time they travel. On-time flights, good in-flight entertainment, more (and better) snacks, and more legroom might be the obvious contributors to a good experience and more loyalty.

While we might hear about those aspects the most, the customer experience is not about just the flight itself. It’s everything from purchasing the ticket on the company’s website or mobile app to checking bags in at the airport or via a mobile app to waiting in the terminal. This mindset has been, and continues to be, adapted to the post-security, onboard, and post-flight experience. So how can we determine which of these factors contribute to the satisfaction of the customer?

# To answer this, we intend on building a classification problem to predict the customer satisfaction

# DATA SET AND DOMAIN

* A dataset is a collection of data, and it can be structured or unstructured.
* A structured data is represented in a tabular format, where every column of the table represents a particular variable, and each row corresponds to a given record of the dataset in question.
* Unsupervised data is not represented in a tabular form, data that we fetch from Facebook, Twitter, and Netflix, etc. with the help of recommendation systems are all our unsupervised data.

# DATA DESCRIPTION:

The data set consists of 129880 observations and 25 features before the cleaning and contains information regarding the customers as well as their experience in the flight based off multiple factors such as their reviews for the food/drinks,seat comfort,inflight entertainment etc. The description for each variable along with the datatype as given in the dataset is as follows:

|  |  |  |
| --- | --- | --- |
| VARIABLE | DATATYPE | DESCRIPTION |
| Unnamed:0 | numeric | Index value |
| 1. Id | numeric | Unique identifier |
| 2. Gender | object | Gender of the passengers (Female, Male) |
| 3. Customer Type | object | The customer type (Loyal customer, disloyal customer) |
| 4. Age | numerical | The actual age of the passengers |
| 5. Type of Travel | object | Purpose of the flight of the passengers (Personal Travel, Business Travel) |
| 6. Class | object | Travel class in the plane of the passengers (Business, Eco, Eco Plus) |
| 7. Flight distance | numeric | The flight distance of this journey |
| 8. Inflight wifi service | numeric | Satisfaction level of the inflight wifi service (0:Not Applicable;1-5) |
| 9.Departure/Arrival time convenient | numeric | Satisfaction level of Departure/Arrival time convenient |
| 10. Ease of Online booking | numeric | Satisfaction level of online booking |
| 11. Gate location | numeric | Satisfaction level of Gate location |
| 12. Food and drink | numeric | Satisfaction level of Food and drink |
| 14. Online boarding | numeric | Satisfaction level of online boarding |
| 15. Seat comfort | numeric | Satisfaction level of Seat comfort |
| 16. Inflight entertainment | numeric | Satisfaction level of inflight entertainment |
| 17. On-board service | numeric | Satisfaction level of On-board service |
| 18. Leg room service | numeric | Satisfaction level of Leg room service |
| 19. Baggage handling | numeric | Satisfaction level of baggage handling |
| 20. Check-in service | numeric | Satisfaction level of Check-in service |
| 21. Inflight service | numeric | Satisfaction level of inflight service |
| 22. Cleanliness | numeric | Satisfaction level of Cleanliness |
| 23. Departure Delay in Minutes | numeric | Minutes delayed when departure |
| 24. Arrival Delay in Minutes | numeric | Minutes delayed when Arrival |
| 25. Satisfaction | object | Airline satisfaction level(Satisfaction, neutral or dissatisfaction) |

# Data Types-independent variables:

# The satisfaction levels across columns like'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' work on a rating system and hence can be classified as categorical data.

# The variables: Gender, Type of Travel ,Customer Type and Class are also part of categorical data

# The variables :Age, Distance, Departure Delay in Minutes and Arrival Delay in Minutes are numerical

**Target Variable: Satisfaction-Categorical**

There is a slight imbalance in the distribution of classes in the Target variable.



# PROJECT METHODOLOGY



* Identify the data source available
* Understanding the data
* Data cleansing and Data preparation
* Data Analysis
* Feature Engineering
* Model Building
* Model Validation
* Recommendations

**PRE-PROCESSING DATA ANALYSIS**

# Data Preparation:

Data preprocessing is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

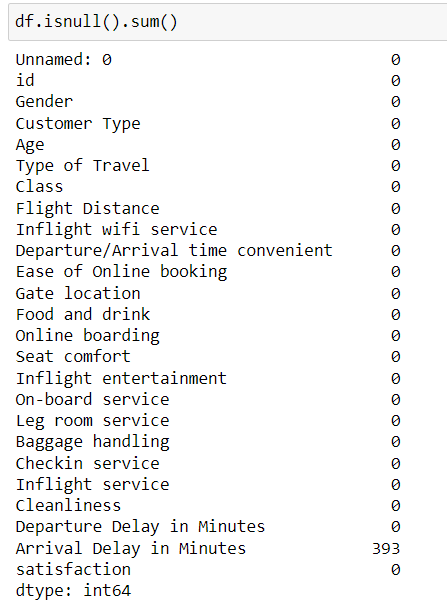
Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis.

Acquire the dataset:

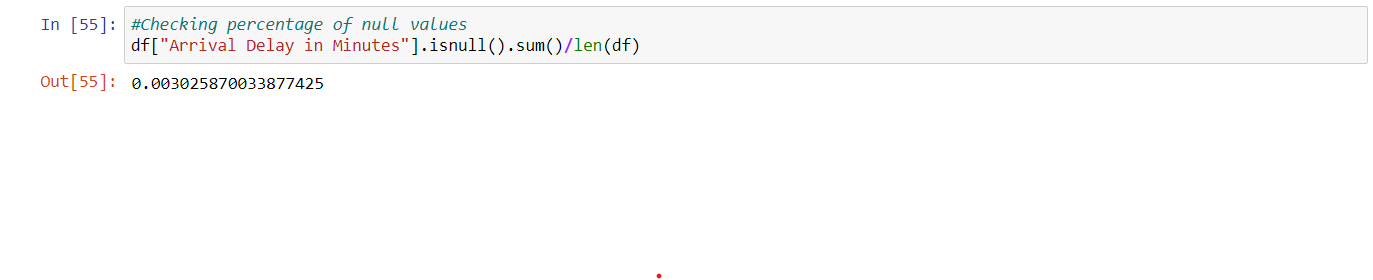
<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>

# Missing/Null Values:

Impute or drop features with missing values based on the percentage of missing values and relevance for model building.

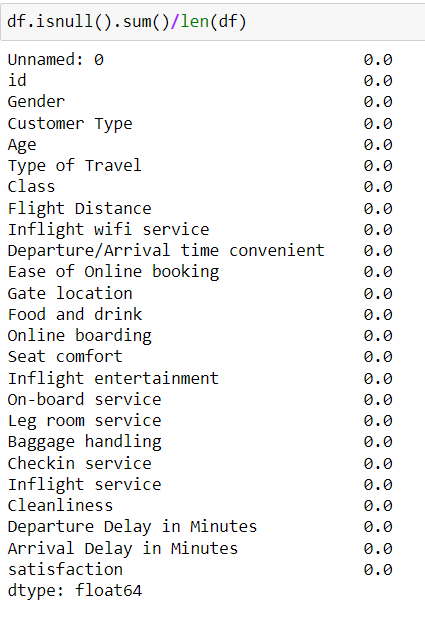


In the dataset we have Null values in one column("Arrival Delay in Minutes"):We calculated the percentage of null values present in the variable to understand their impact on the data

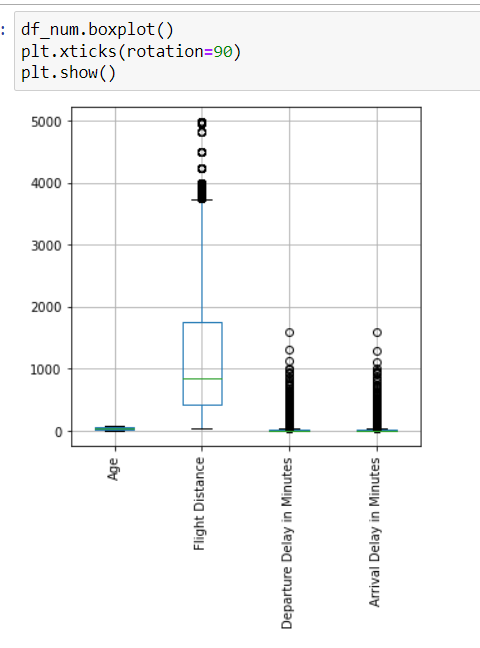


As the percentage of null values was found to be negligible, we dropped the observations containing these null values.

There are no null values in the data now



# OUTLIERS

****

There are outliers present in the Flight Distance, Departure Delay in Minutes and Arrival Delay in Minutes column. As these variables play a role in customer satisfaction (more the delay,we can assume less satisfied the customer), we are not excluding the outliers . As we proceed with model building we can transform the extreme outliers. But for the base model we choose to leave the outliers as it is

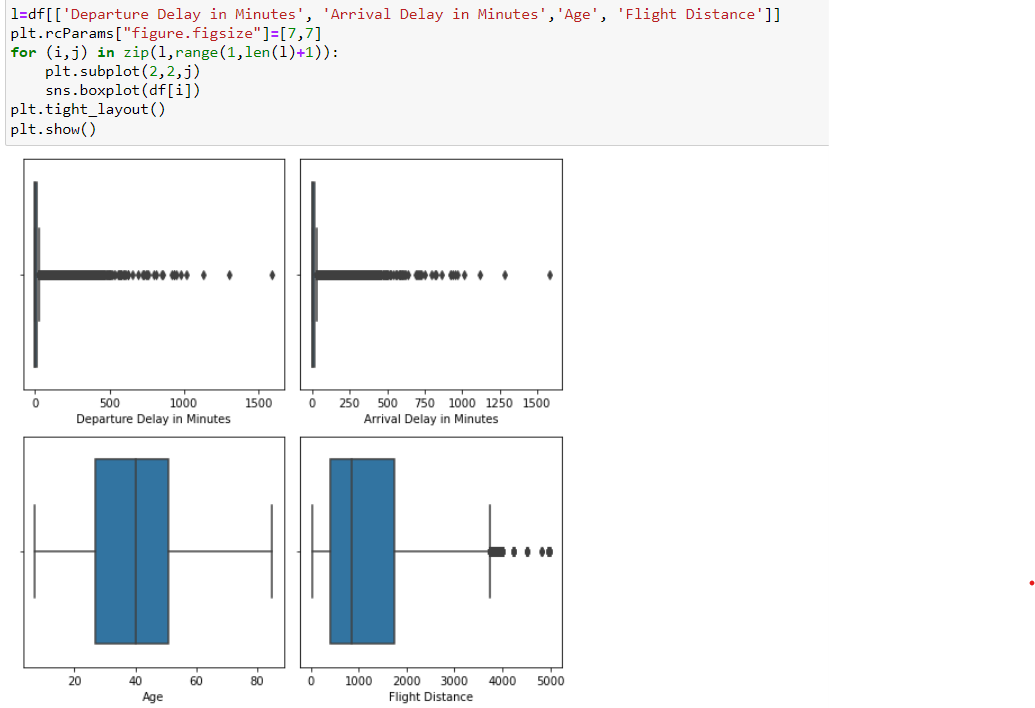
# INSIGNIFICANT COLUMNS

The below are the redundant features that are dropped from the dataset.

* Unnamed: 0- Index value
* Id- Unique identifiers of each row, has no impact on the model

# EXPLORATORY DATA ANALYSIS & BUSINESS INSIGHTS

UNIVARIATE ANALYSIS

1.Overview of the Numerical variables

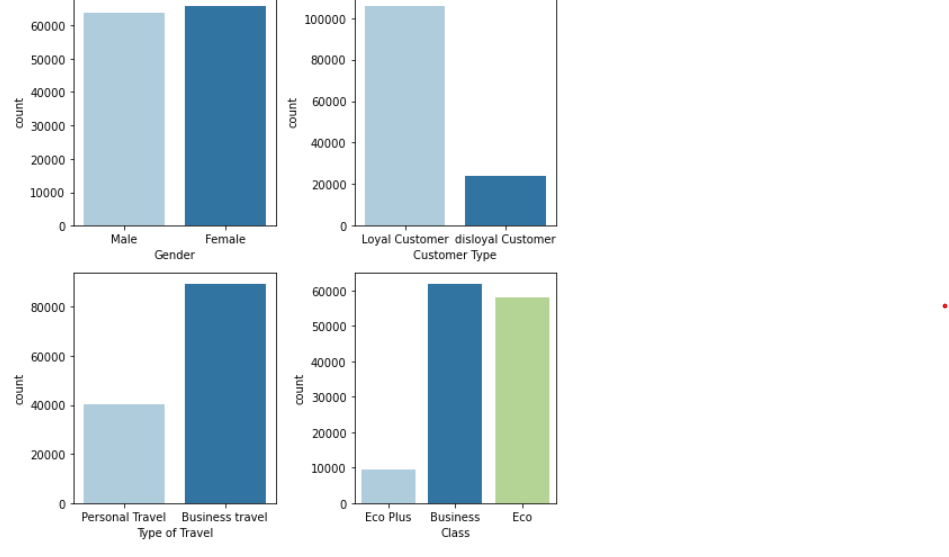
* ***Heavy presence of outliers in variables: 'Departure Delay in Minutes', 'Arrival Delay in Minutes'***
* ***The average age of passengers was found to be ~40***
* ***The average distance was found to be 1190 miles***

2.Overview of Rating columns



* ***A large variety of passengers gave an average rating of 4 out of 5 for various factors like ease of online booking, on-board service, checkin service etc.***
* ***The factor with the largest amount of dissatisfaction (rating 1) was in-flight service***
* ***There were some factors that were voted zero ,which we can assume to be for the cases where the customer refused to provide a rating***

3.Overview of Categorical columns

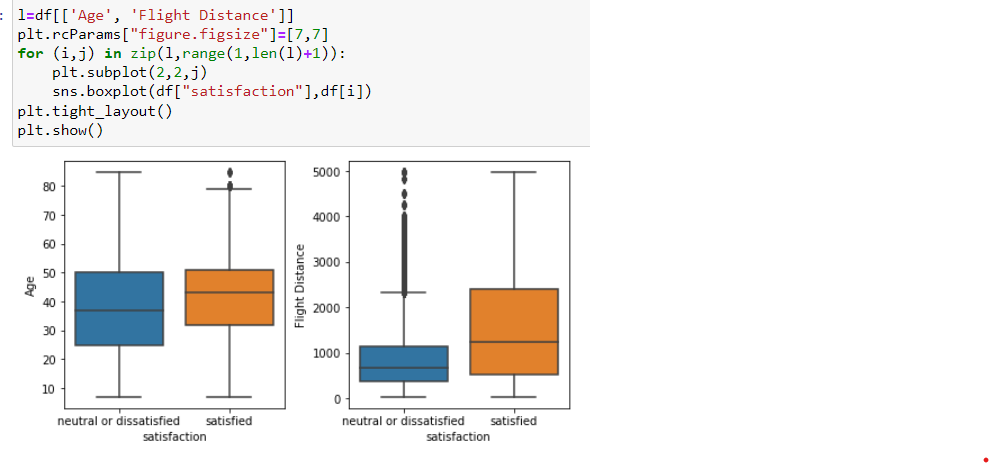
****

* ***Majority of passengers chose Business class.***
* ***The data contains information predominantly of loyal customers***
* ***Both genders were more or less equally captured for this project***

BIVARIATE ANALYSIS:

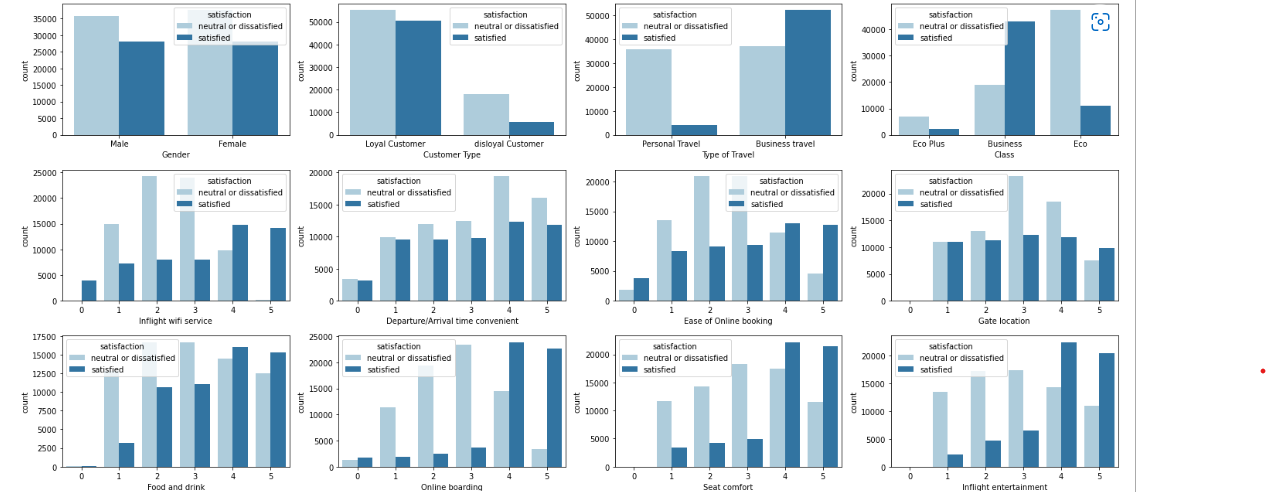
Relationship with the target variable:

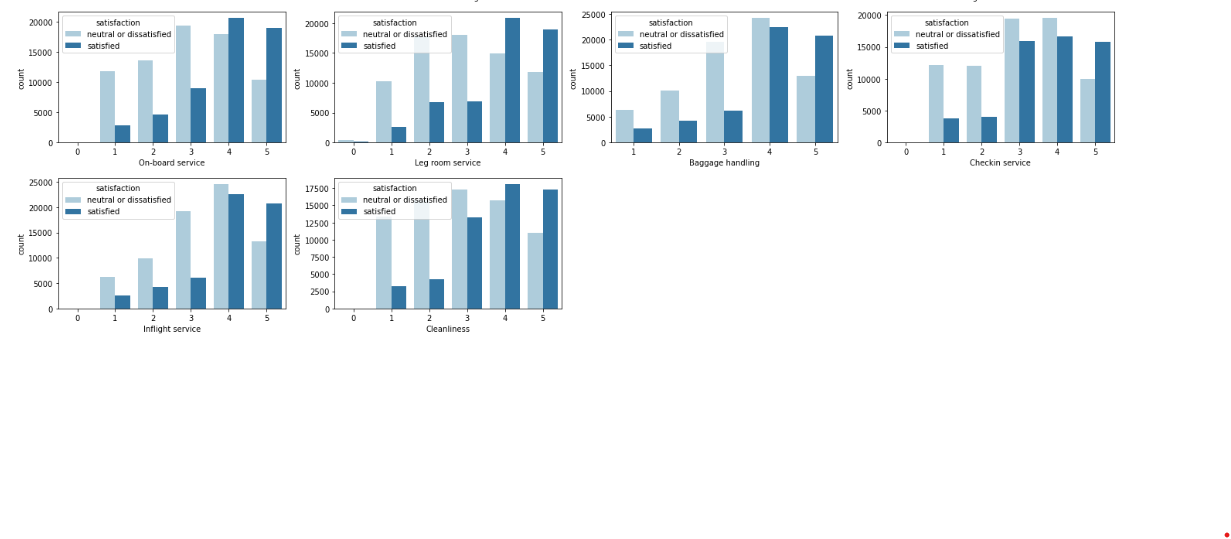
1. Numerical independent variables vs target variable



* ***The average age for satisfied and dissatisfied were found to be 37 and 41 respectively***
* ***The average distance travelled by satisfied customers was higher than that of neutral/dissatisfied***

1. Categorical independent variables vs target variable

****



* ***Loyal customers were found to be satisfied on comparison with disloyal ones***
* ***Business travelers’ are highly satisfied***
* ***Moreover passengers business class get high satisfied count, while passengers in the eco and eco plus class were found to be more dissatisfied***
* ***The Main Causes of dissatisfaction are: In-Flight Wifi, Online Booking and Online boarding***
* ***Less But Observable Dissatisfaction is caused by Leg-Room Services and Cleanliness***
* ***Some Dissatisfaction is also caused by Gate Location, Food and Drinks, Seat Comfort, Inflight Entertainment, On Board Services***
* ***Online Boarding Shows generally satisfactory ratings, but is one of the main cause of dissatisfaction among the dissatisfied customers***

**Statistical Tests:**

So far we have observed that the satisfaction of the customer has a significant relationship with other variables such as the type of travel, their rating of factors like inflight wifi/seat comfort/on-board service etc. We have made these observations by studying the graphs above.

We can further validate these observations using statistical tests like Chi-squared contingency(Test of independence).

**Chi-Square test of independence**

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables.

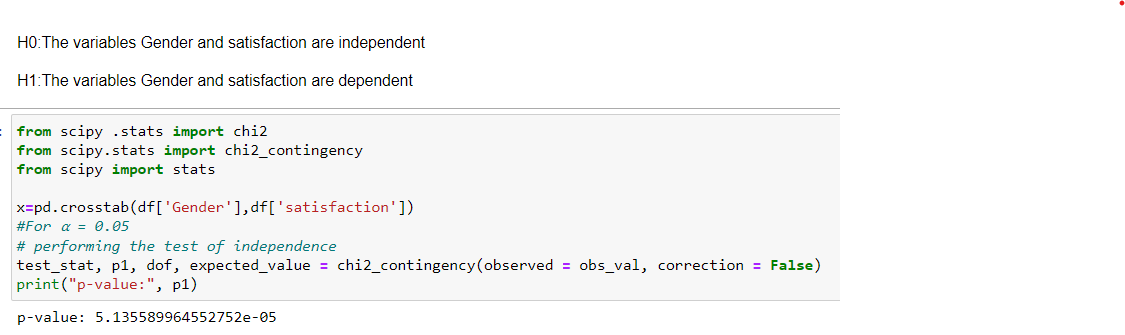
The hypothesis to test the independence of attributes

H0**:** The attributes are independent**.**

H1**:** The attributes are dependent

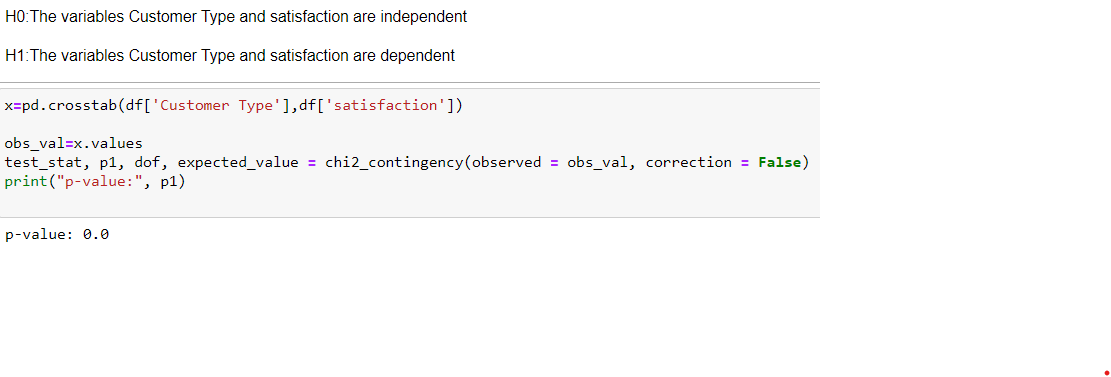
**1.Gender vs Satisfaction:**

To understand the impact of the variable “gender” on satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the customer’s gender on their satisfaction

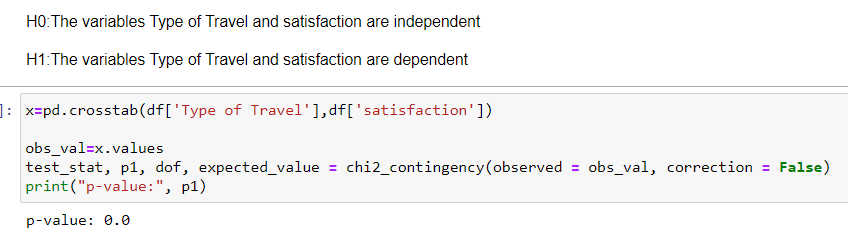
****

**2.Customer Type vs Satisfaction :**

To understand the impact of the variable “customer type” on satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the customer’s loyalty on their satisfaction

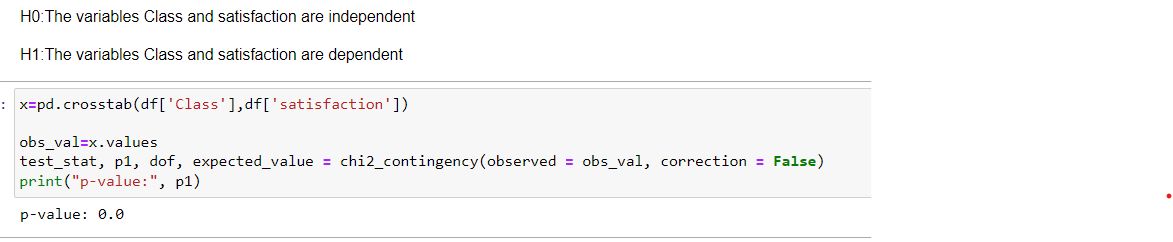
**3.Type of Travel vs Satisfaction**

To understand the impact of the variable “Type of Travel” on satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the customer’s reason for travel on their satisfaction

****

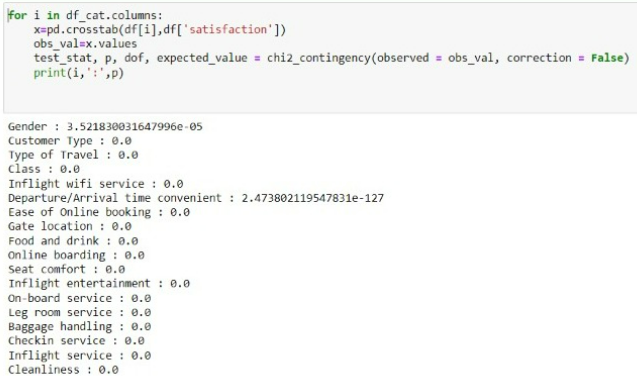
**4.Class vs Satisfaction:**

To understand the impact of the variable “Class” on satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the Class the customer is using (Business/Eco/Ecoplus) on their satisfaction

****

**5. Rating columns vs Satisfaction:**

To understand the impact of the rating variables (example: seat comfort, inflight service, food and drink quality)on satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the way these factors were rated by the customer on their satisfaction.



**Kruskal Wallis test**

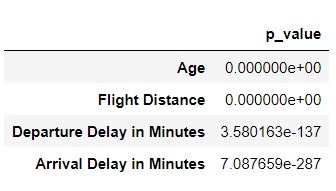
It is used to check the equality of population medians for more than two independent samples. It is a non-parametric test that works similar to ANOVA, but doesn’t require the tests of normality and equal variances to be satisfied.

The null and alternative hypothesis is given as:

Ho:The median of all treatments are the same.

H1: At least one treatment has a different median.

To understand the impact of the numerical variables (example: age,flight distance, departure delay in minutes,arrival delay in minutes )on satisfaction we conducted a Kruskal Wallis test. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the way these factors were rated by the customer on their satisfaction.



**INFERENCE FROM STATISTICAL TESTS:**

Gender and satisfaction

For the variable Gender the p\_value < α ,so we reject the null hypothesis. This means that the Gender and the target variable satisfied are dependent. Gender plays a role in customer satisfaction.

Customer type and satisfaction

For the variable Customer Type the p\_value < α ,so we reject the null hypothesis. The variable customer type also makes a difference in the customer satisfaction.

Type of Travel and satisfaction

For the variable Type of travel the p\_value < α ,so we reject the null hypothesis. The business travelers travel more frequently than people who travel for personal affairs. So they are more familiar with the flight experiences than the other group. This can create an effect in the customer satisfaction.

Class and satisfaction.

For the variable Class the p\_value < α , so we reject the null hypothesis. The variable satisfaction is dependent on the variable class.

Rating variables and satisfaction.

For all the variables the p\_value < α , so we reject the null hypothesis. The variable satisfaction is dependent on the rating variables.

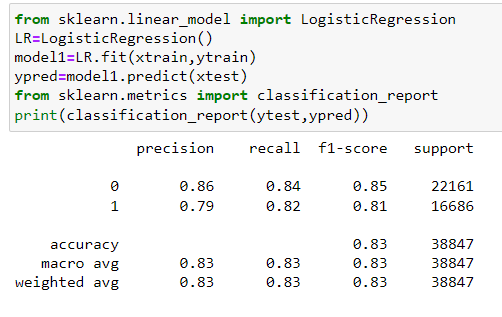
Numerical variables and satisfaction.

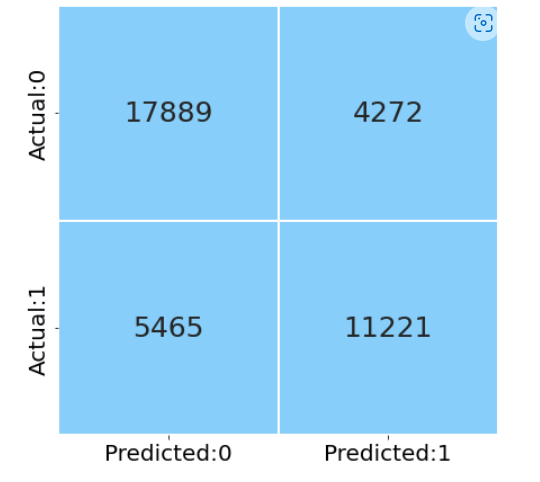
For all the numerical variables the p\_value < α , so we reject the null hypothesis. The variable satisfaction is dependent on the numerical variables.

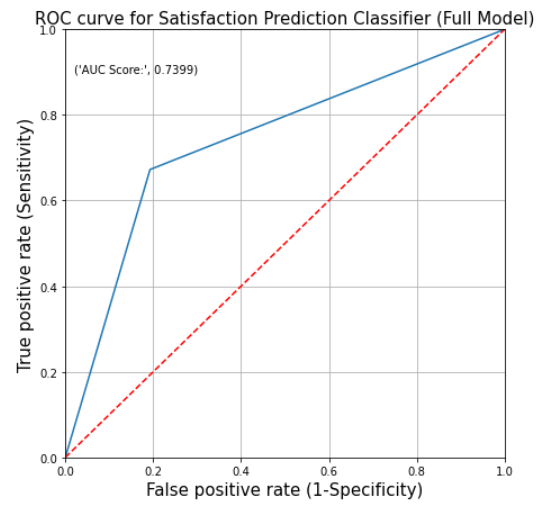
**BASE MODEL**

We used Logistic Regression for our base model:

* Logistic Regression is a binary classification algorithm. It predicts the probability of occurrence of a label class.
* Consider that logistic regression is used to identify whether the consumer satisfied or not.

****

****

****

**Understanding The Metrics:**

**PERFORMANCE METRICS:**

* **Confusion Matrix:**

It is the performance measure for the classification problem. It is a table used to compare predicted and actual values of the target variable.

* **ROC:**

ROC curve is the plot of TPR against the FPR values obtained at all possible threshold values.

**PERFORMANCE EVALUATION METRICS:**

* **Accuracy:**

Accuracy is the fraction of predictions that our model got correct. Higher the accuracy of the model better is the model. The accuracy of the base model was found to be 0.83

* **Precision:**

Precision is the proportion of positive cases that were correctly predicted.

* **Recall:**

A recall is the proportion of actual positive cases that were correctly predicted.

* **F1 score:**

F1score is the harmonic mean of precision and recall values for a classification model.

**BUSINESS JUSTIFICATION:**

* The main motive is to improve the performance of the model. As per the business scenario, the model is predicting the customers is satisfied but in reality they aren’t. In this scenario, we won’t be able to establish satisfaction of customer and this might result in us losing a potentially loyal customer
* In the second scenario if the model is predicted customer is not satisfied but in reality they are satisfied.

### **Future Work**

* Further after building the base model, we will proceed to scale the numerical variables to provide equal weightage to all the numerical variables. We can apply transformation techniques to treat the skewness.
* We will also move forward with feature selection techniques to study what variables have an impact on the target.
* We will proceed with building new and more efficient models made with bagging and boosting techniques followed by hyperparameter tuning to get best results.
* Finally we will proceed with model validation using KFold/cross val techniques.