# Statement of Work SOW(V2)

## AI Algorithms

# Airline Passenger Satisfaction Analysis

# Submitted To:

# Professor Marcos Bittencourt

# Submitted By:

# Sonal Raghuvanshi

# 100805082

CONTENTS

[ABSTRACT 3](#_Toc57754899)

[INTRODUCTION 3](#_Toc57754900)

[SCOPE OF WORK 3](#_Toc57754901)

[DELIVERABLES 3](#_Toc57754902)

[MILESTONES 3](#_Toc57754903)

[DATASET SOURCE 4](#_Toc57754904)

[DATA ASSUMPTIONS 5](#_Toc57754905)

[DATA CONSTRAINTS 5](#_Toc57754906)

[TEST PROCESS 5](#_Toc57754907)

[EXPLORATORY DATA ANALYSIS 6](#_Toc57754908)

[PRELIMINARY DATA MANIPULATIONS 11](#_Toc57754909)

[STATISTICAL ANALYSIS 13](#_Toc57754910)

[FEATURE ENGINEERING 25](#_Toc57754911)

[ACCEPTANCE 28](#_Toc57754912)

# ABSTRACT

Over the last decade, there has been immense growth in the number of airline industries around the globe which has provided the passengers a plethora of options to choose from. Passengers have the ability to access the features, convenience, prices, services, and a lot of other key factors along with providing their rating to the airline services they have experienced. To keep themselves abreast of the competition, airlines need to know how to maximize their potential. Passenger satisfaction plays a major role when it comes to identifying the important factors which affect productivity and overall turnover. In this report, we have analyzed the features pertaining to more satisfied airline passengers and compared the results obtained by various machine learning algorithms to achieve the best performance.

# INTRODUCTION

The selected topic for this analysis is ‘Trees’. This is a tree analogy in machine learning that can be used for classification and regression. It is a flowchart like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test and each leaf represents the class label which is the decision taken after computing all the attributes. A classification tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting data into partitions and then splitting it up further on each of the branches. There are different decision rules which are defined for each layer and as the depth increases, the complexity of the rules for partition also increases. A classification tree labels, records, and provide a level of confidence that the classification is correct. There are various ensemble methods such as bagging and boosting which combine weak classification tree models and form a more accurate classification model.

# SCOPE OF WORK

The outcome of this project is to analyze the features and find out those features that are affecting the customer more and making the customer more satisfied and happier. It is achieved by using various machine learning algorithms and then comparing the results obtained to get the best performing algorithm.

## DELIVERABLES

There will be the whole project at the end with a python file and a detailed document used for analyzing the airline passenger satisfaction data. It helps to identify the best features that are on more satisfied airline passengers and thus, increasing the airline productivity and overall turnover.

## MILESTONES

|  |  |
| --- | --- |
| Milestone | Estimated Delivery Date |
| Statement of Work | 06-Nov-2020 |
| Data Acquisition and Understanding | 01-Dec-2020 |
| Modelling | 18-Dec-2020 |
| Prototyping | 18-Dec-2020 |
| Deployment | 18-Dec-2020 |

# DATASET SOURCE

Passenger satisfaction is the key element of modern businesses as it can significantly affect service quality improvement. The data used for this analysis is obtained from the airline passenger satisfaction survey that consisted of 24 variables and 129,880 records that are obtained from 2 CSV files. These records describe the various ratings that the passengers have given to the airline services after their journeys. The response variable in the data is the ‘Satisfaction’ variable which describes if the passenger feels satisfied or unsatisfied/neutral from their journey based on their seat location, cleanliness, food quality, etc. These features are useful to measure how the products and services provided by the airline surpass the passenger’s satisfaction level.

Data Source : Klein, T.J. (2020, Feb 20). “Airline Passenger Satisfaction”. Retrieved from <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables Information** | | | | | | | |
| ***Variable Name*** | | ***Description*** | | | | | |
| *ID* | | *ID of the passenger* | | | | | |
| *Gender* | |  | *Gender of the passengers (Female, Male)* | | | |  |
| *Customer Type* | | *The customer type (Loyal customer, disloyal customer)* | | | | | |
| *Age* | |  | *The actual age of the passengers* |  | | | |
| *Type of Travel* | | *Purpose of the flight of the passengers (Personal Travel, Business Travel)* | | | | | |
| *Class* | | *Travel class in the plane of the passengers (Business, Eco, Eco Plus)* | | | | | |
| *Flight Distance* | | *The flight distance of this journey* | | | | | |
| *Inflight Wifi Service* | | *Satisfaction level of the inflight wifi service (0: Not Applicable;1-5)* | | | | | |
| *Departure/Arrival time convenient* | | *Satisfaction level of Departure/Arrival time convenient* | | | | | |
| *Ease of Online Booking* | |  | *Satisfaction level of online booking* | |  | | |
| *Gate Location* | | *Satisfaction level of Gate location* | | | | | |
| *Food and Drink* | |  | *Satisfaction level of Food and drink* | |  | | |
| *Online Boarding* | | *Satisfaction level of online boarding* | | | | | |
| *Seat Comfort* | |  | *Satisfaction level of Seat comfort* |  | | | |
| *Inflight Entertainment* | | *Satisfaction level of inflight entertainment* | | | | | |
| *On-Board Service* | |  | *Satisfaction level of On-board service* | | |  | |
| *Leg Room Service* | | *Satisfaction level of Leg room service* | | | | | |
| *Baggage Handling* | |  | *Satisfaction level of baggage handling* | | |  | |
| *Check-in Service* | | *Satisfaction level of Check-in service* | | | | | |
| *Inflight Service* | |  | *Satisfaction level of inflight service* | |  | | |
| *Cleanliness* | | *Satisfaction level of Cleanliness* | | | | | |
| *Departure Delay in Minutes* |  |  | *Minutes delayed when departure* |  | | | |
| *Arrival Delay in Minutes* | | *Minutes delayed when Arrival* | | | | | |
| *Satisfaction* | | *Airline satisfaction level (1 = Satisfaction, 0= neutral or dissatisfaction)* | | | | | |

# DATA ASSUMPTIONS

The first column of the data obtained was useless as it consisted of the row number and the second column consisted of the ID of the airline passenger which again doesn’t contribute to our analysis, hence we have dropped it in the beginning.

# DATA CONSTRAINTS

The dataset consisted of 18 categorical predictors and 4 numerical variables apart from our variable of interest. A lot of categorical variables were encoded in numbers but the rest of them such as Gender, Class, Type of Travel, etc. were encoded using dummy encoding technique. A few values were missing from the data which were removed and while checking for the outliers using boxplot, we could see that 3 or 4 values were out of the box for Departure Delay and Flight Distance but they could not be considered as an outlier as the airline departure delay can easily range from 0 to a few hours. Also, for some of the domestic flights, the distance is small whereas, for international flights, the distance can be large so after checking the out of the box values, we found that they cannot be considered as an outlier.

# TEST PROCESS

As we can see by the dataset, there is not much class imbalance but still, to cater to some of the count differences, we have used f-measure. If we will use only the accuracy for example, then our model will be biased towards the dissatisfied passengers as their number of counts is more. If we use precision, the it tends to minimize false positives, it will lead our classifier to declare more cases as dissatisfied and we might end up not recognizing the important factors that have led to customer satisfaction. If we were to use recall, as it tends to minimize false negatives, we may end up giving too much importance to the services which might not be affecting our passenger’s satisfaction level. So, to balance this, we need to use f-measure which is a trade-off between recall and precision.







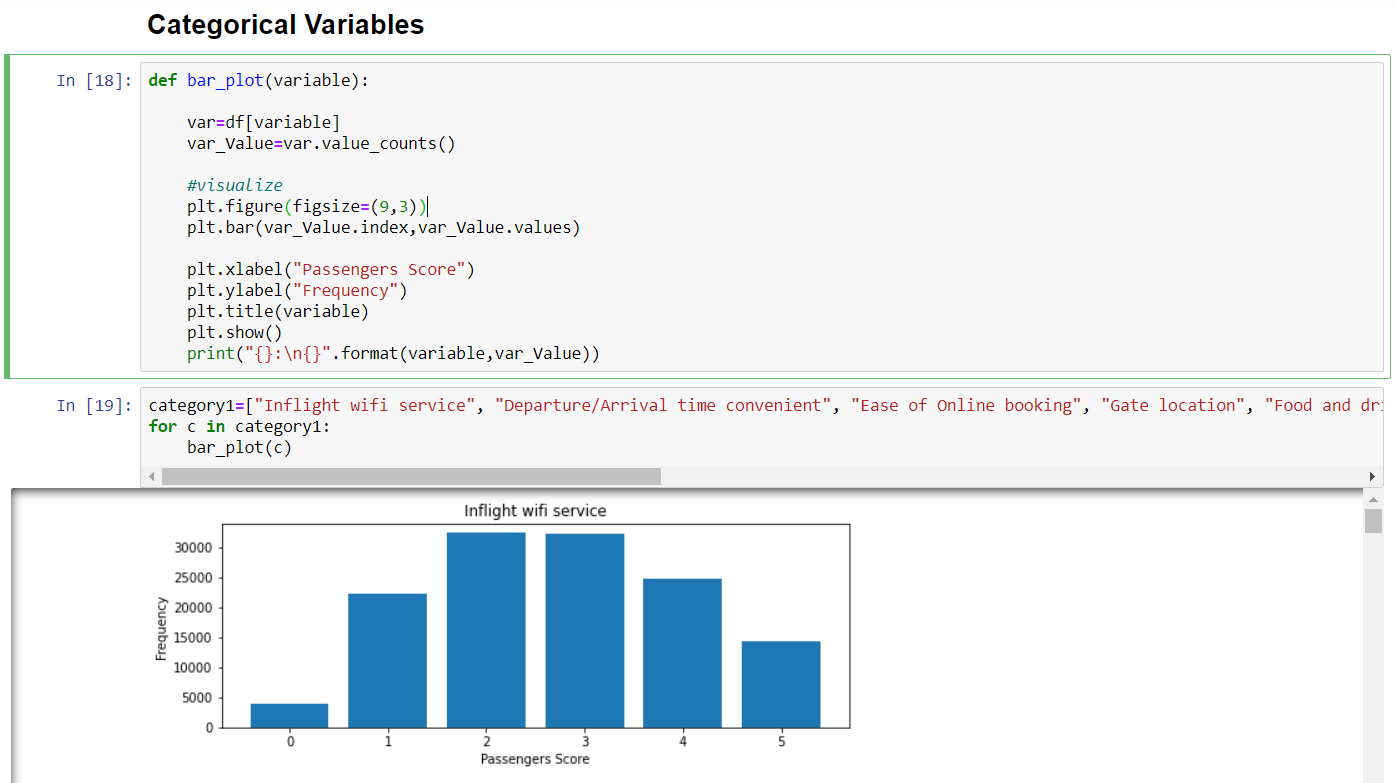
Further, we have split our data into training and test set in the ratio of 80:20 taking the larger section for training purposes. Since most of the machine learning algorithms work better when the data is on a similar scale and normalized, hence we have scaled the feature set using the standard scaler to bring the mean to 0 and a unit standard deviation.

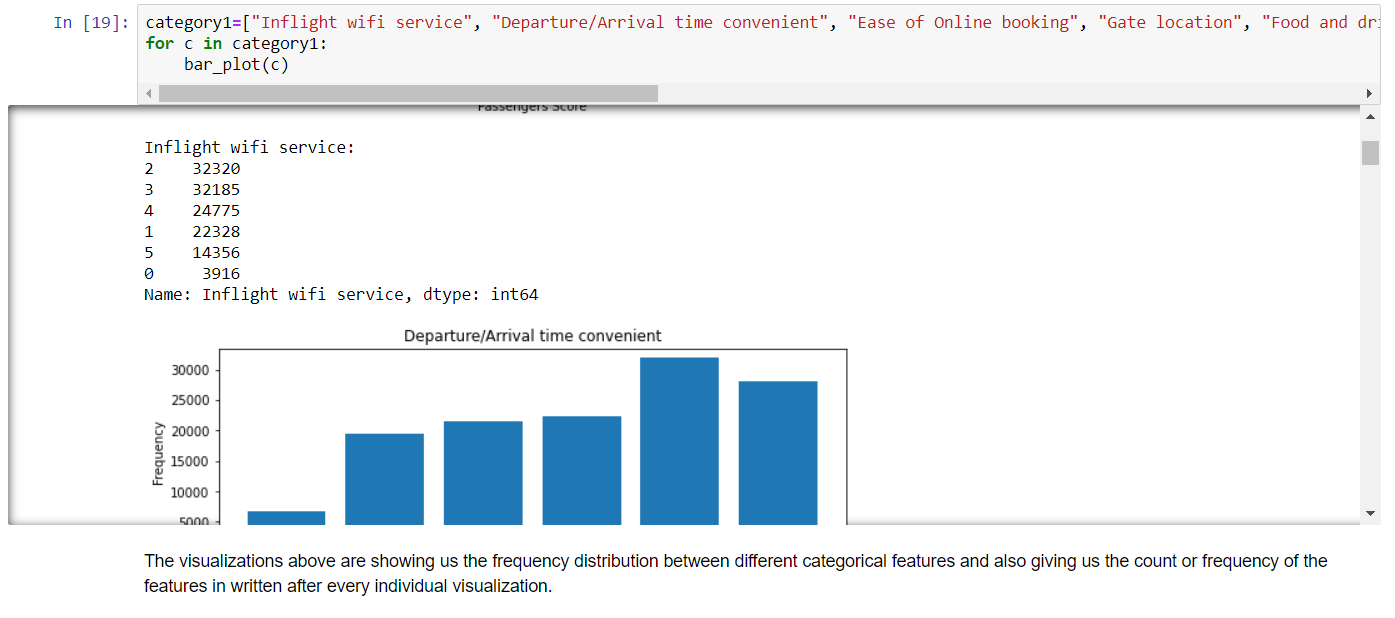
Some so many models can give the airline passenger satisfaction analysis like Decision Tree, Random Forest, Logistic Regression, KNN, etc. but in the end, we are going to use cross-validation to find the best out of the available model options.

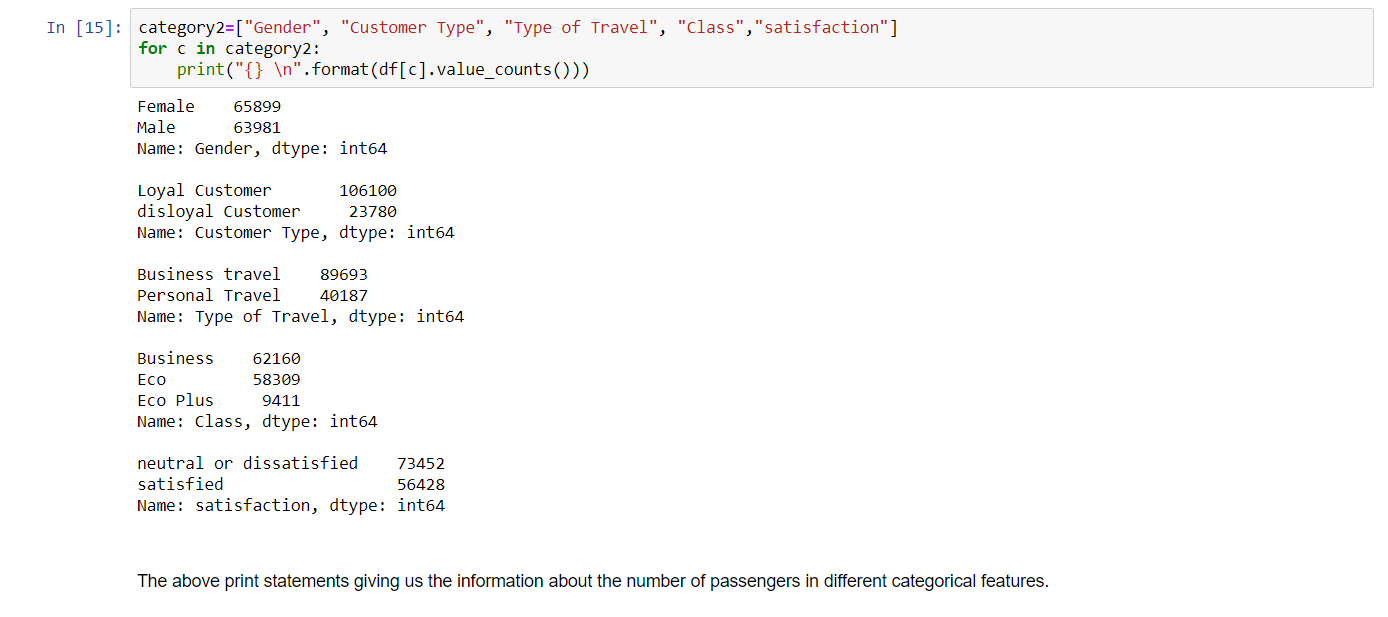
# EXPLORATORY DATA ANALYSIS

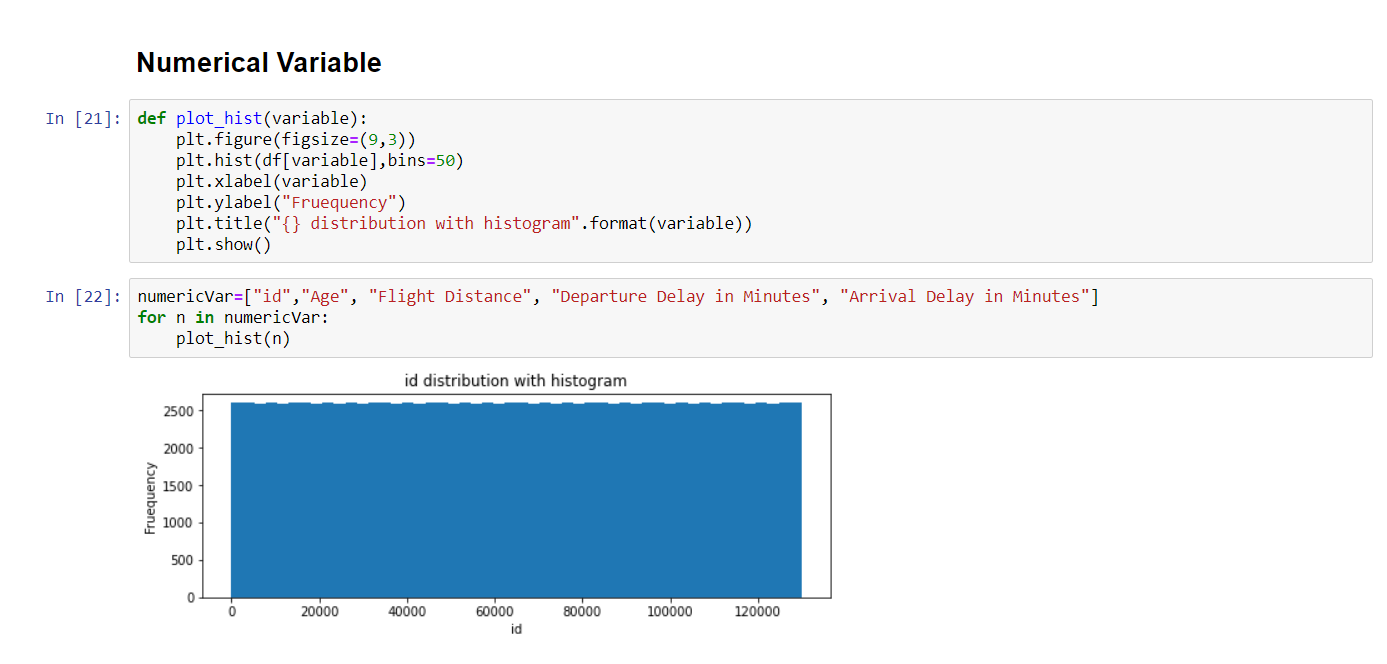
Preliminary analysis of our data gives some insights about the data, how the features are related to each other and what all are the features that are responsible for satisfied customer analysis.

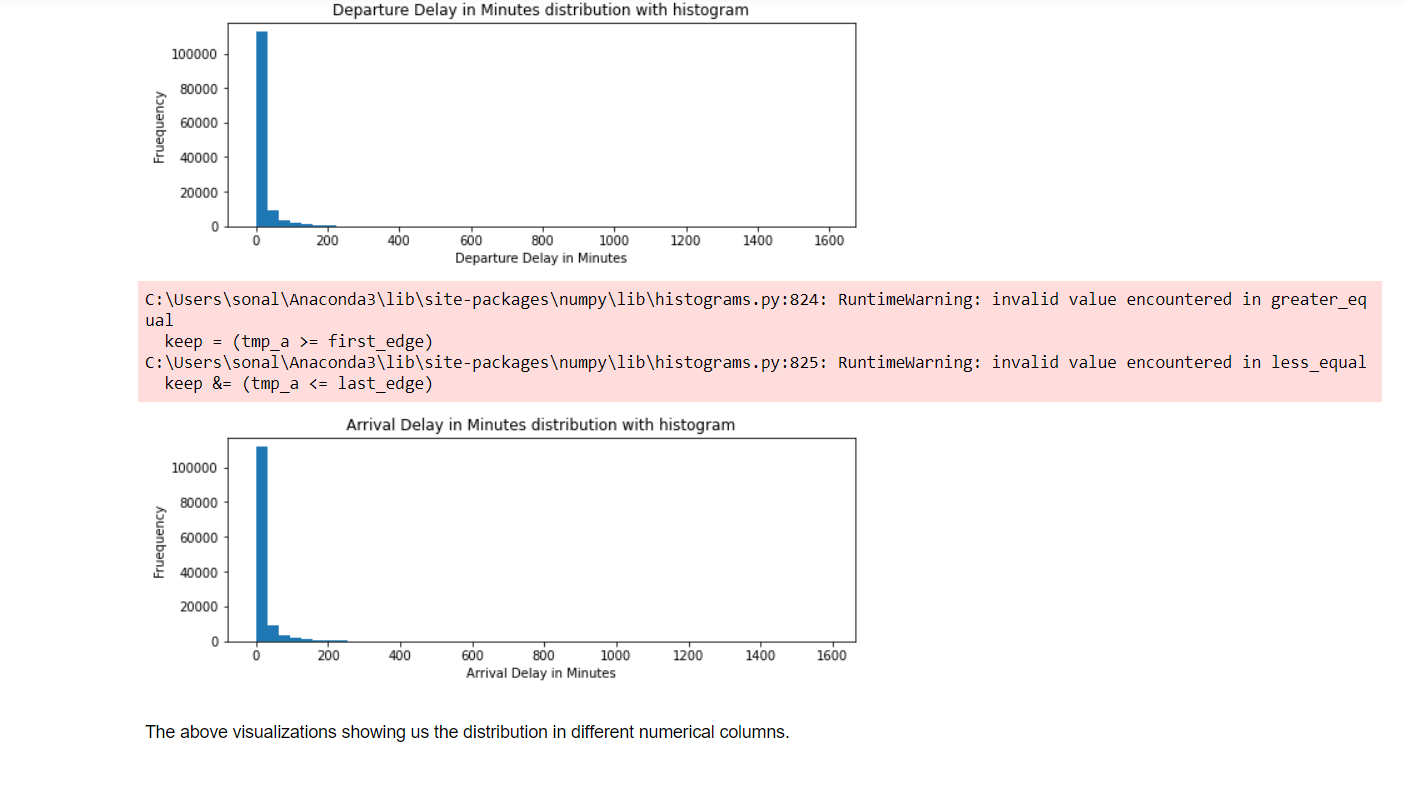
Please find below some of the basic analysis with information:

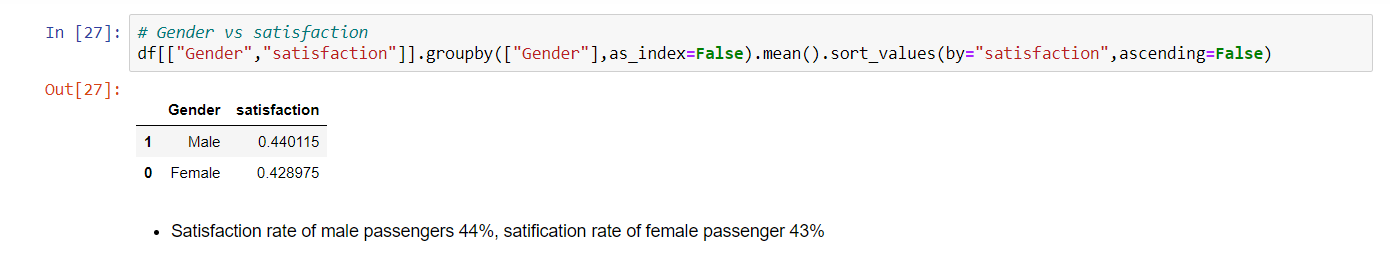






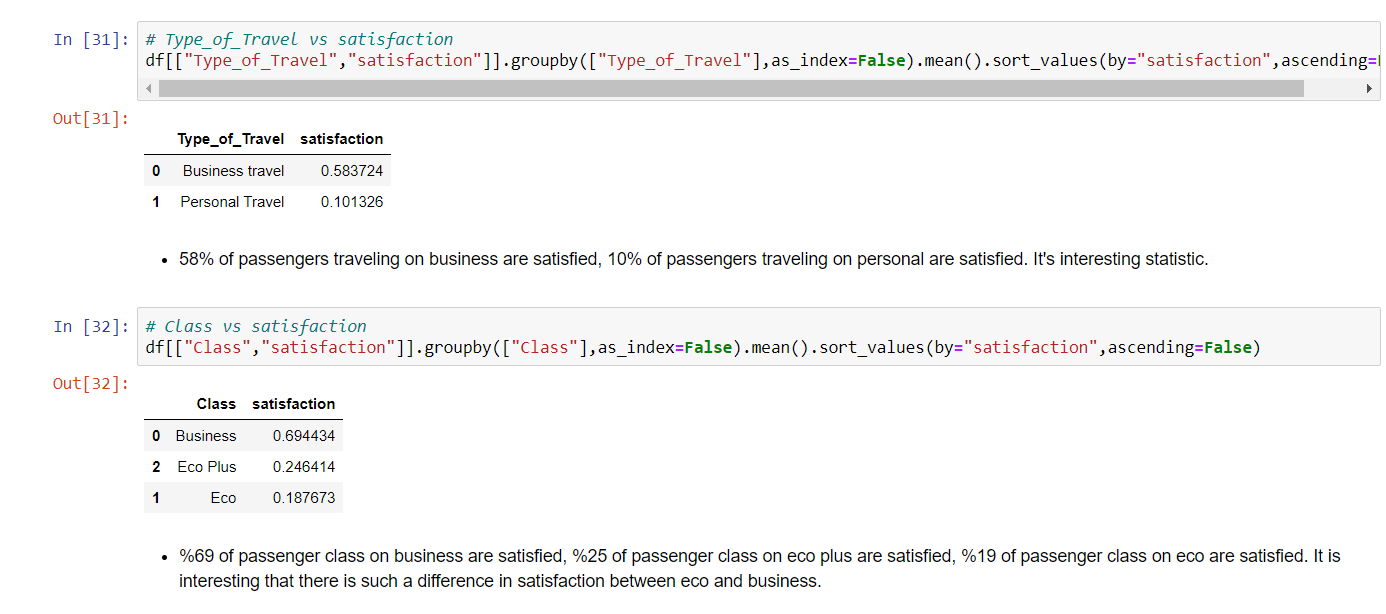


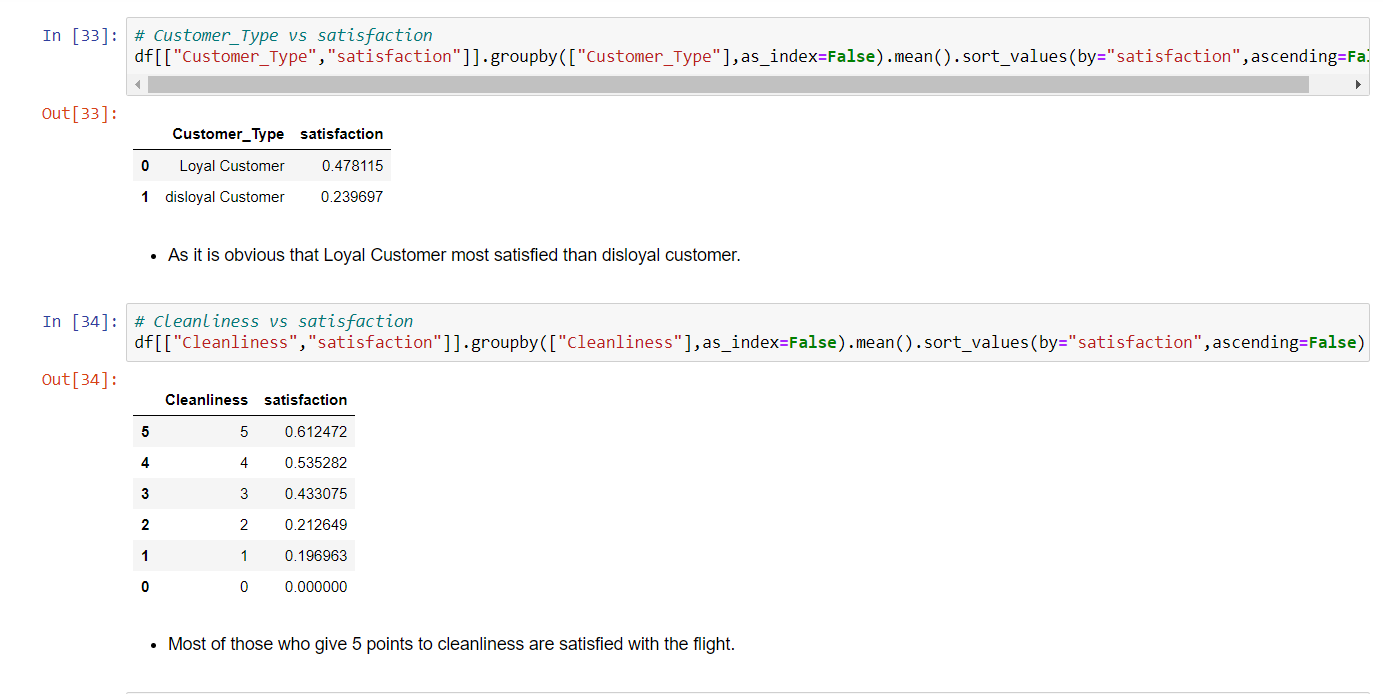




The visualization above shows there is the percentage of satisfied men and women passengers in our survey with 44% as male and 43% as female. The number of satisfied passengers is less than the number of dissatisfied or neutral passengers as we can see in the above visualizations too.











# PRELIMINARY DATA MANIPULATIONS

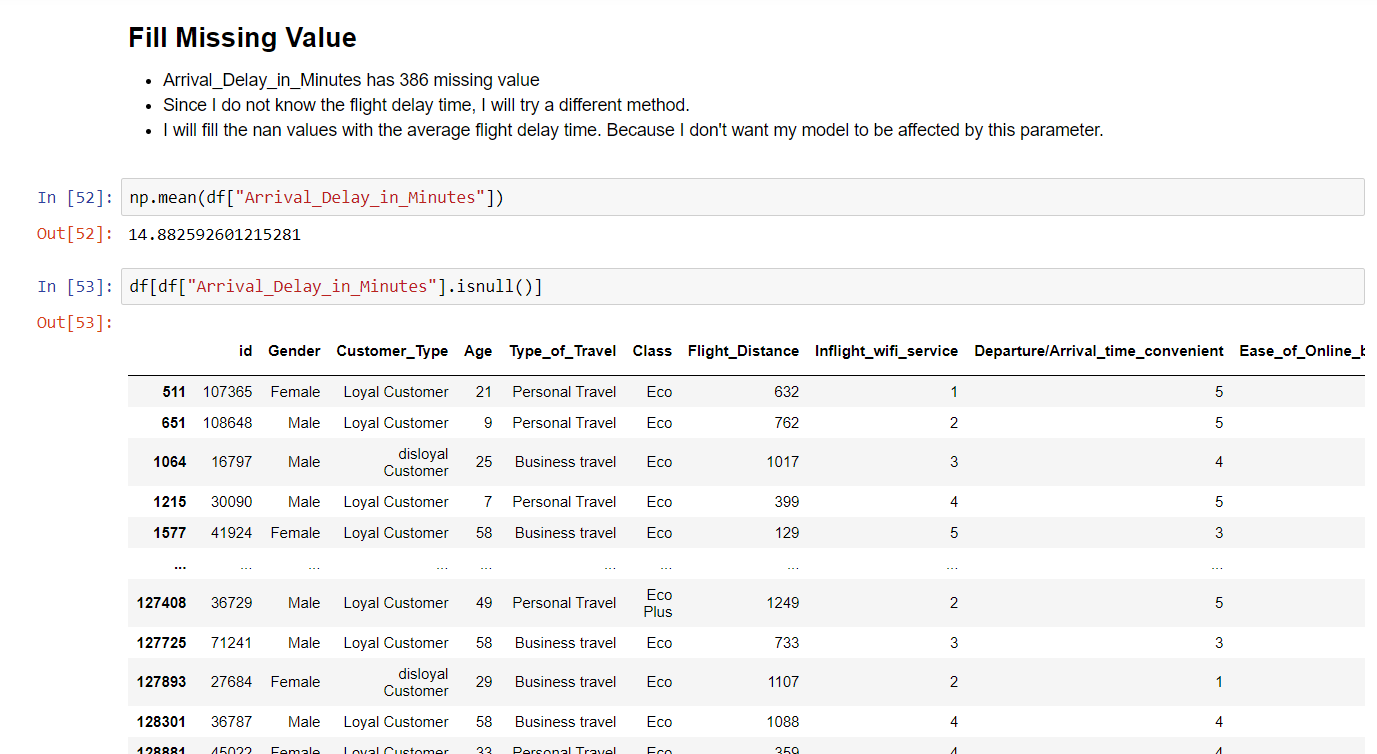
As we can see in the below code snippets that there are some columns that have been dropped off and some columns that have been rescaled and cleaned.

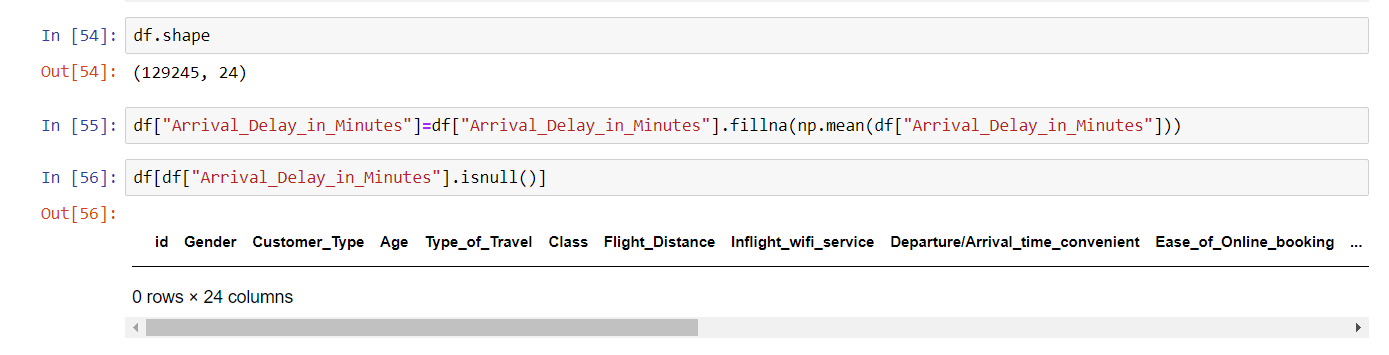
There is one column named Arrival\_Delay\_in\_Minutes in our data which is having 386 missing

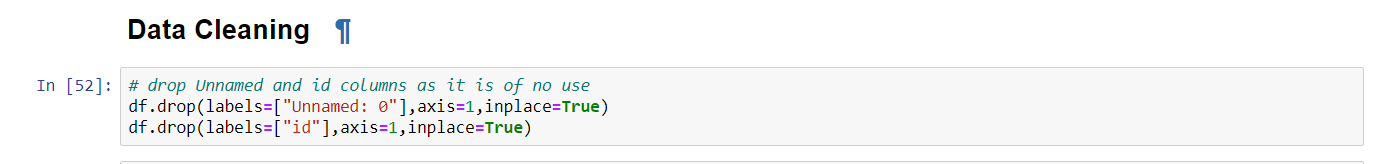
records, since I do not know the flight delay time, I have tried a different method as I have filled the nan values with the average flight delay time because I do not want my model to be affected by this parameter.







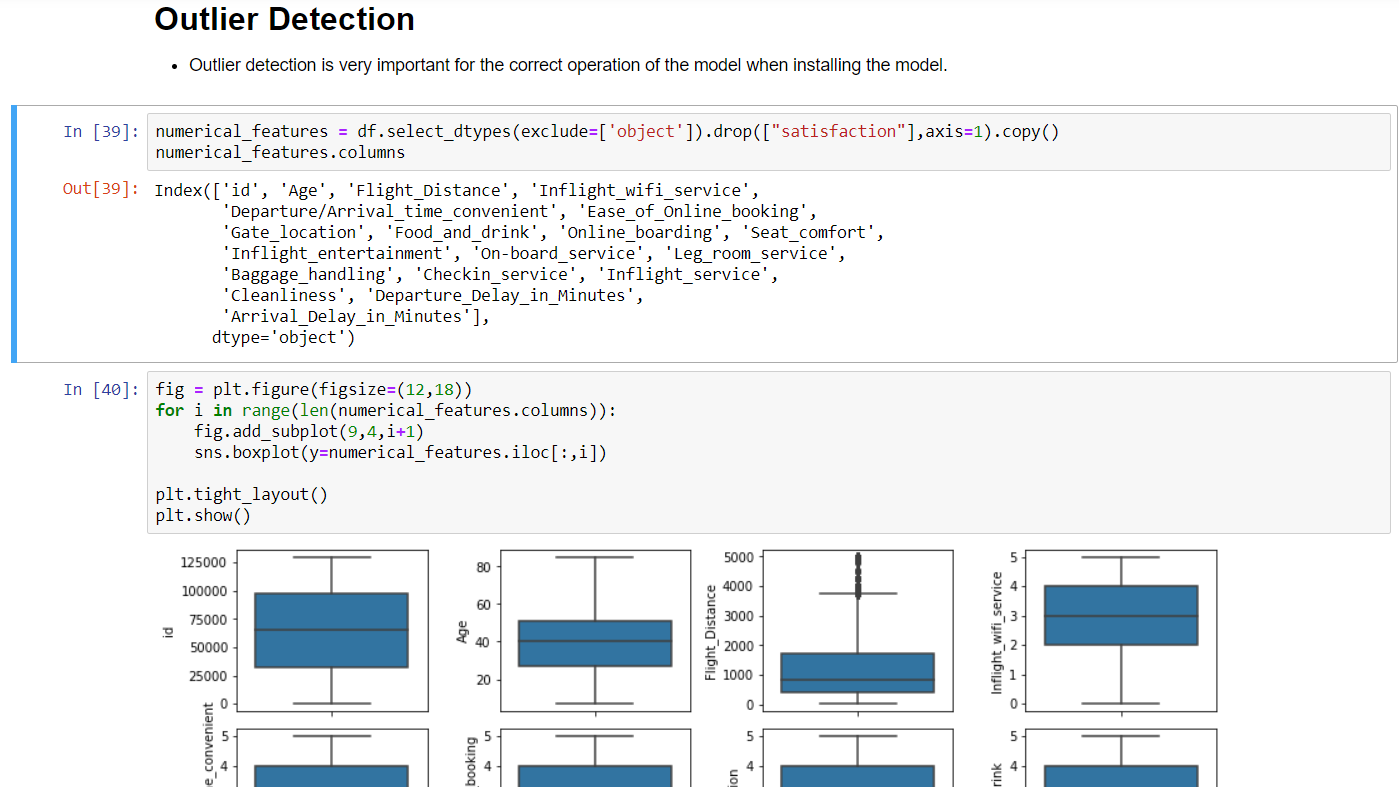




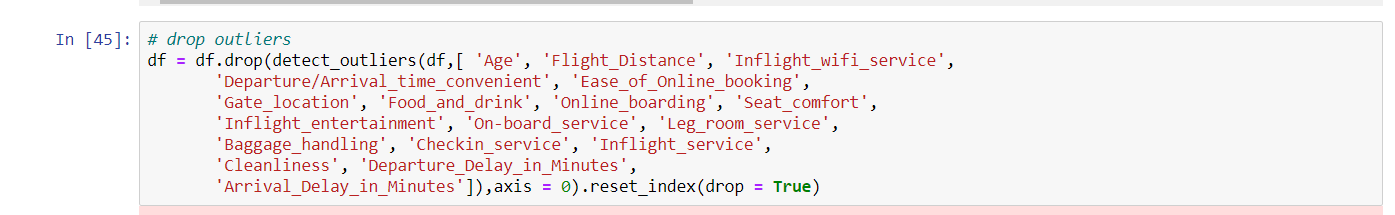
# STATISTICAL ANALYSIS

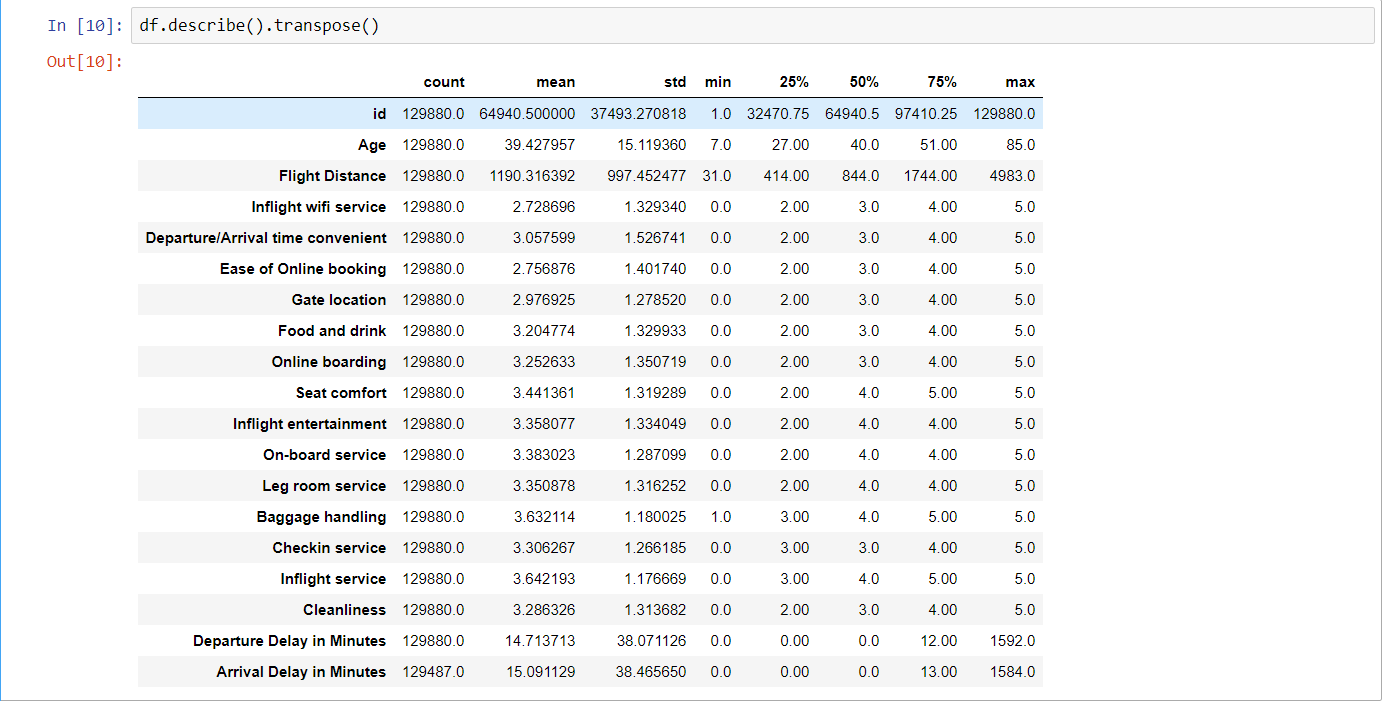
Statistical analysis gives you the ability to assess, understand and make predictions about data, it is at the very bottom of inferential statistics and can be considered of the important thing in python.

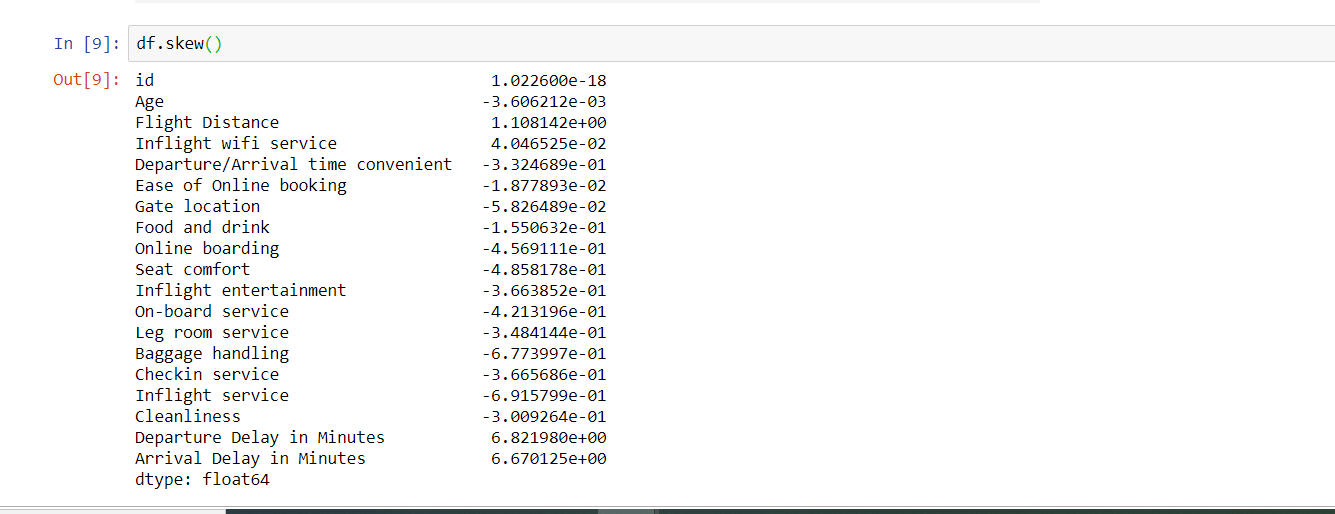
In the python file I have created outlier’s detection code where I have detected the outliers with the help of the method created and then dropped the outliers detected.

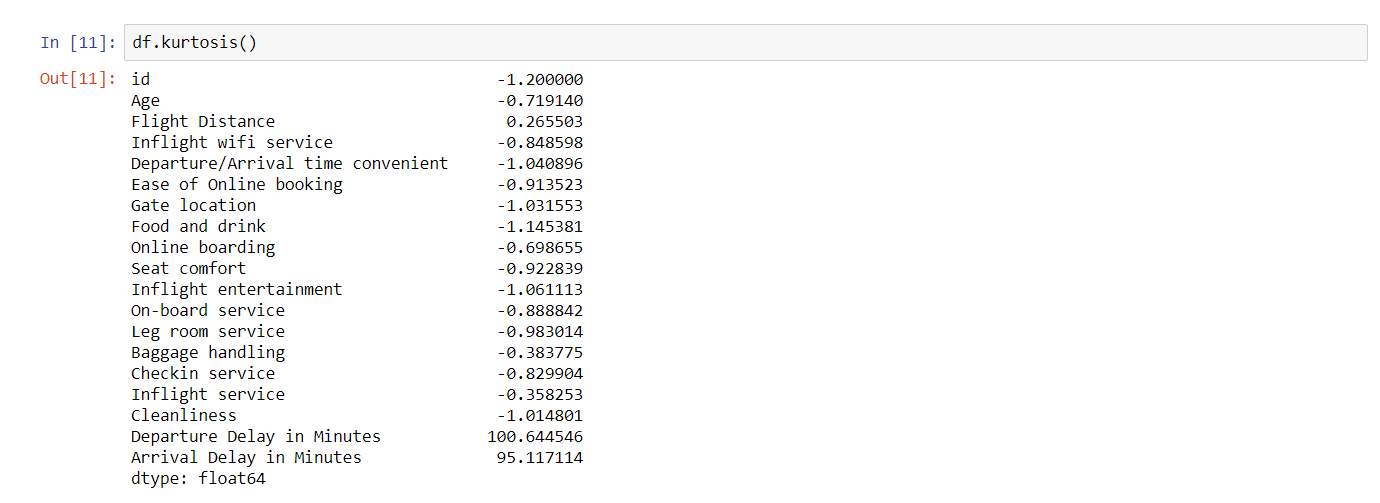






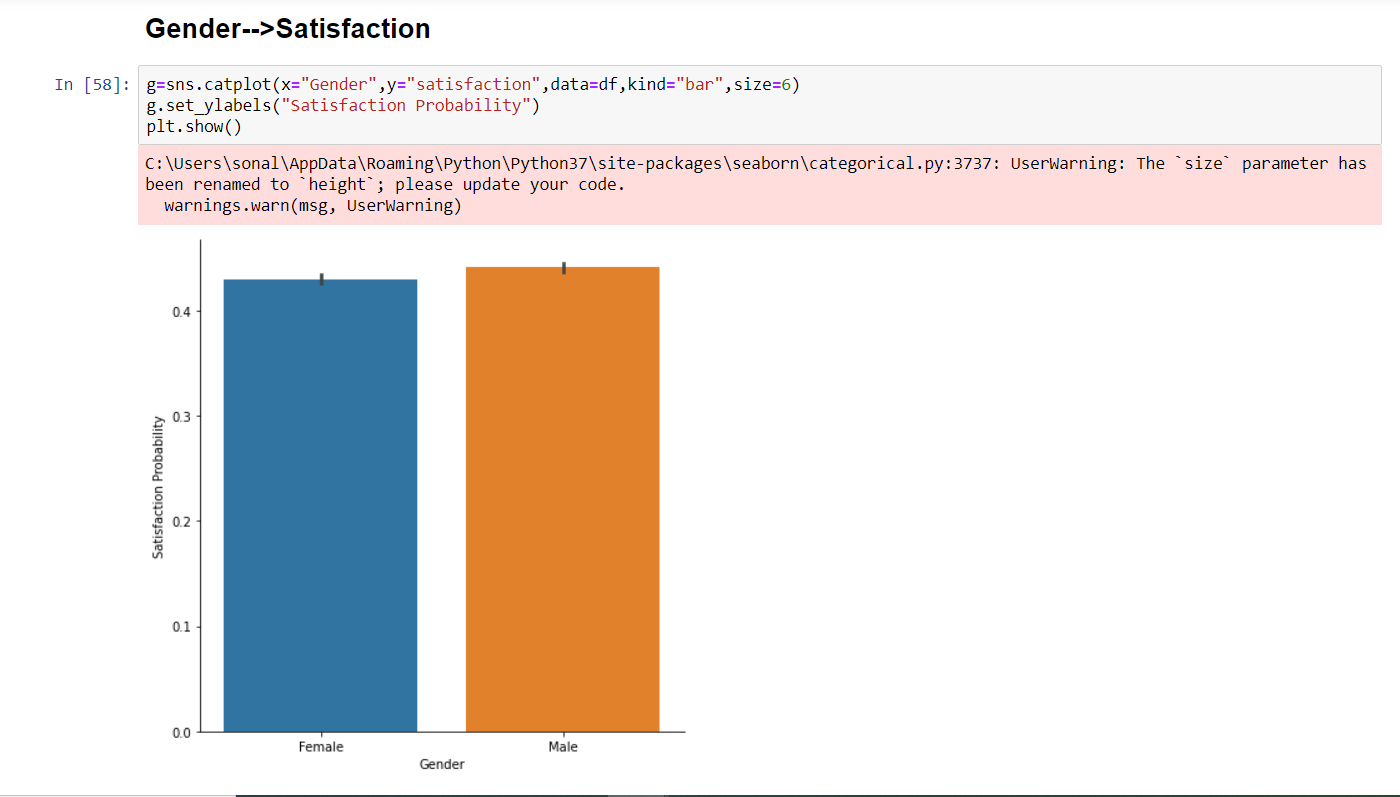


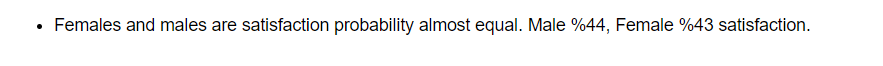


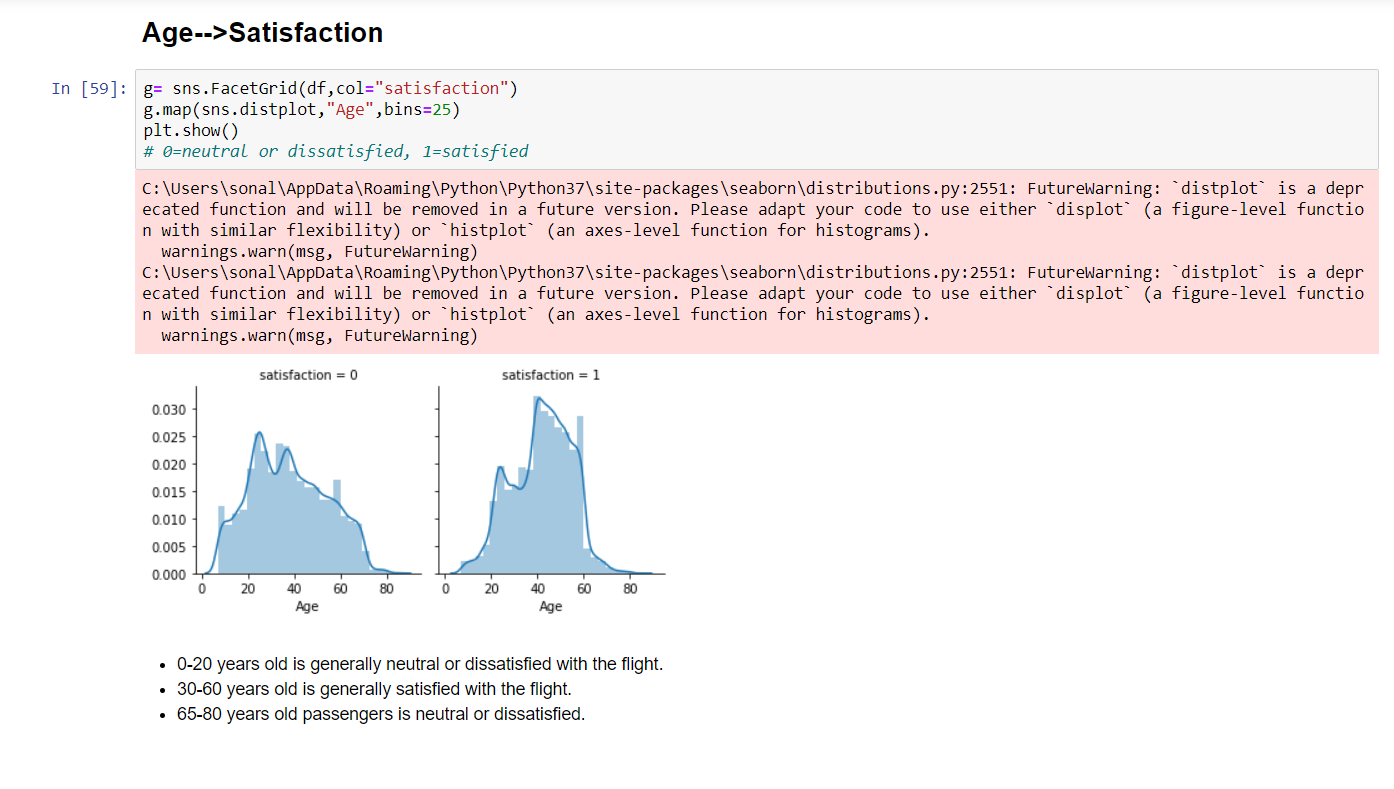




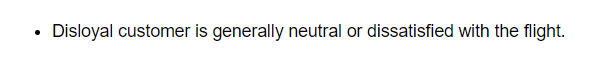
Departure delay and arrival delay are highly correlated and showed linear relationship. Some of the features which looked highly correlated were ‘Food and Drinks’, ‘Cleanliness’, ‘Seat Comfort’ and ‘Inflight Entertainment’ amongst themselves.

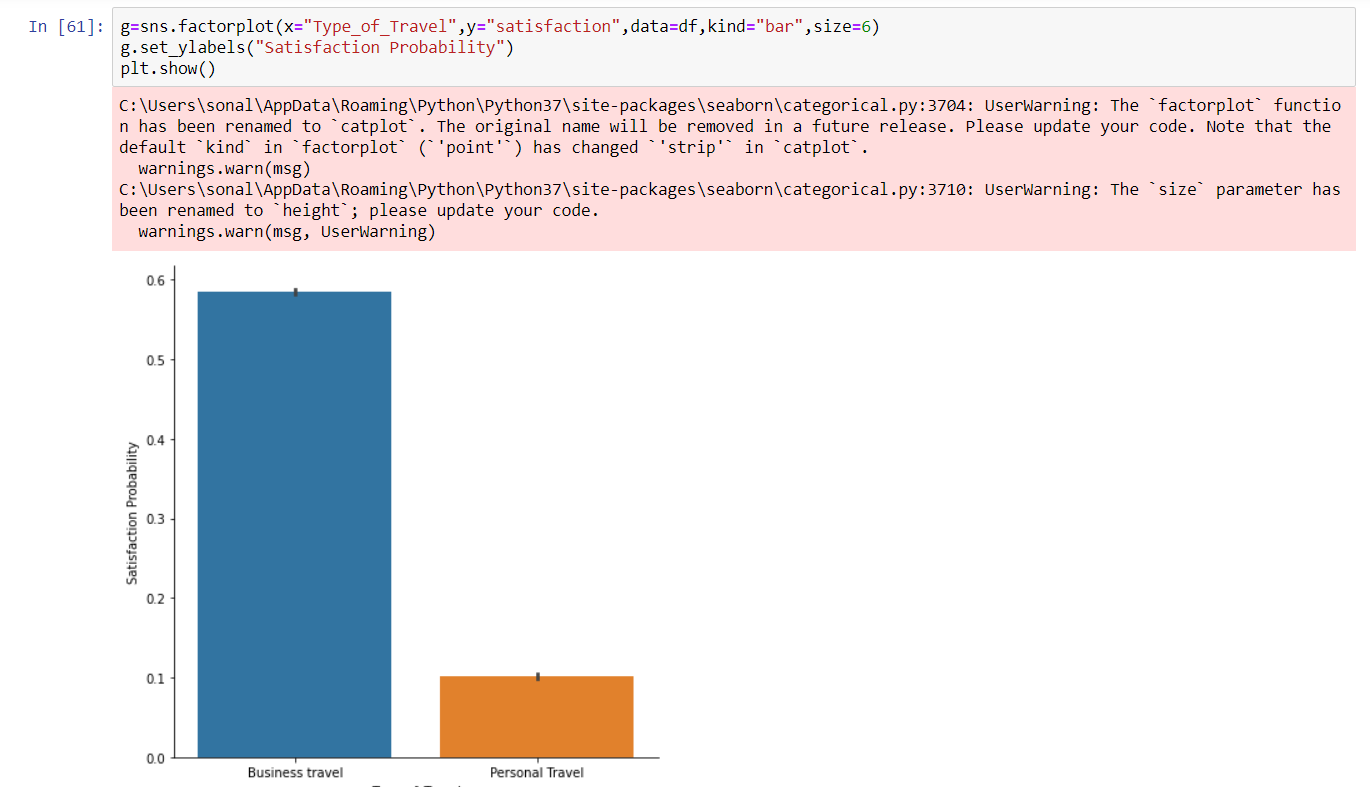


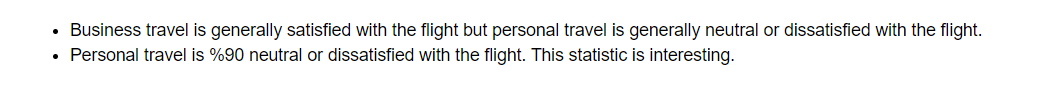


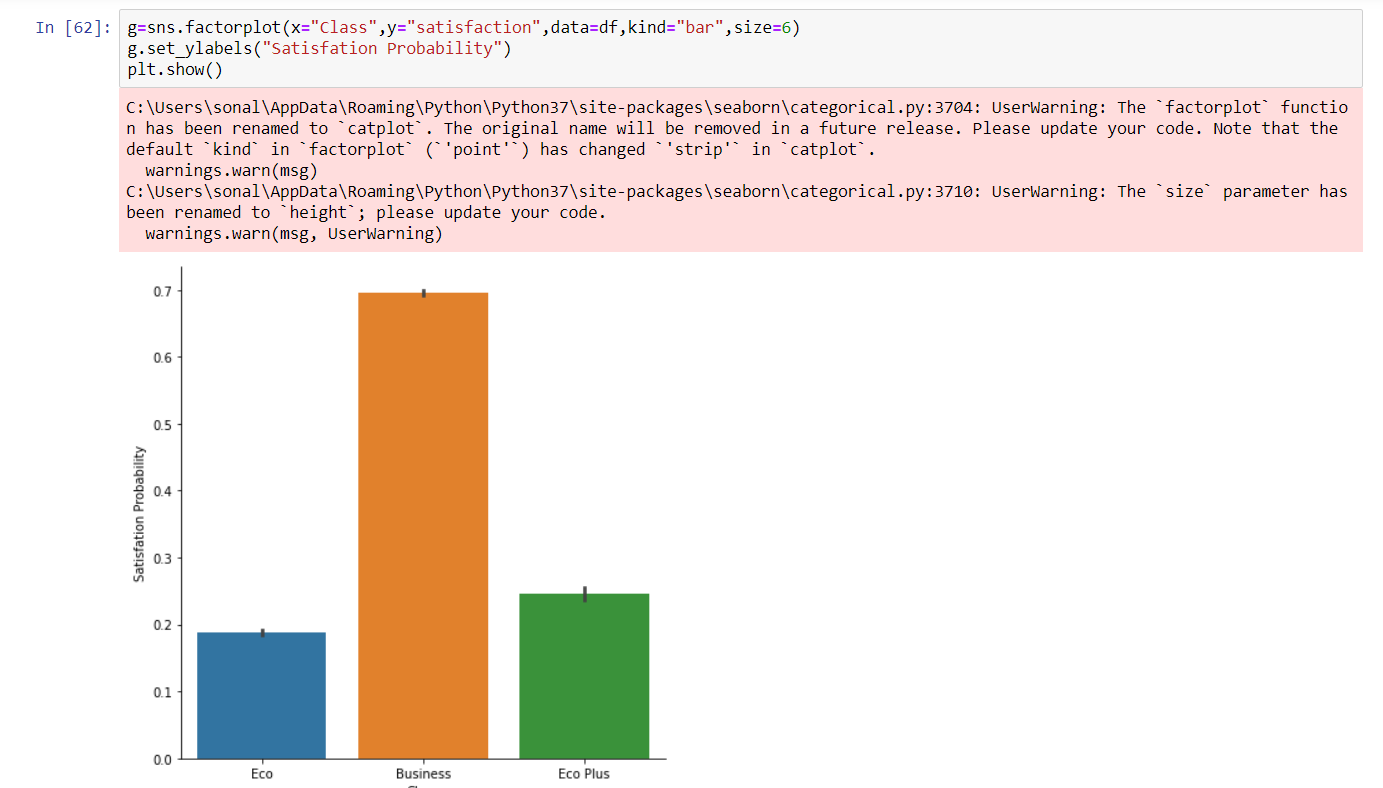


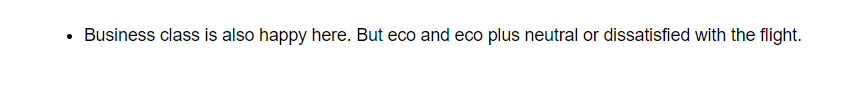




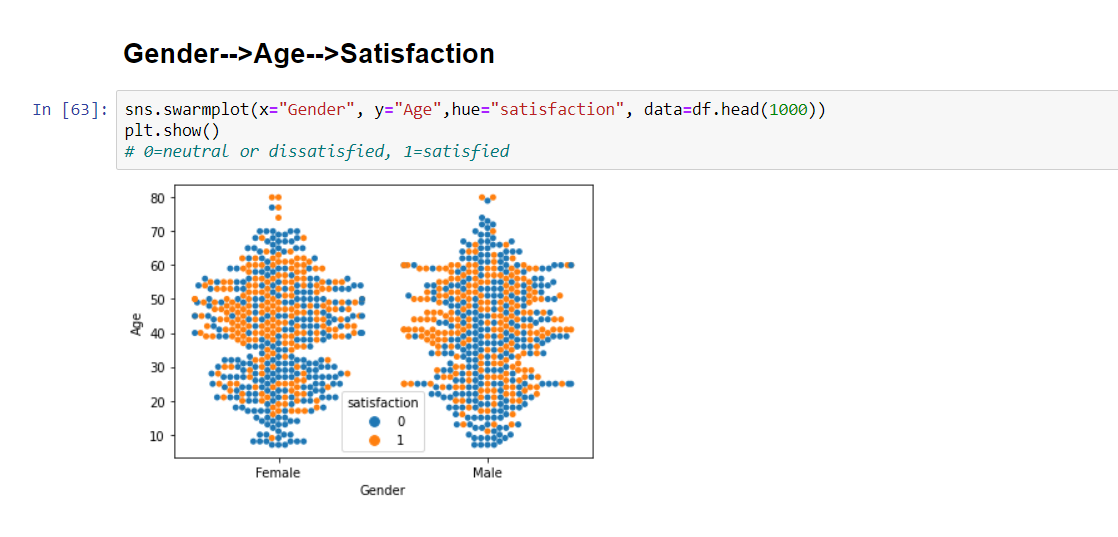


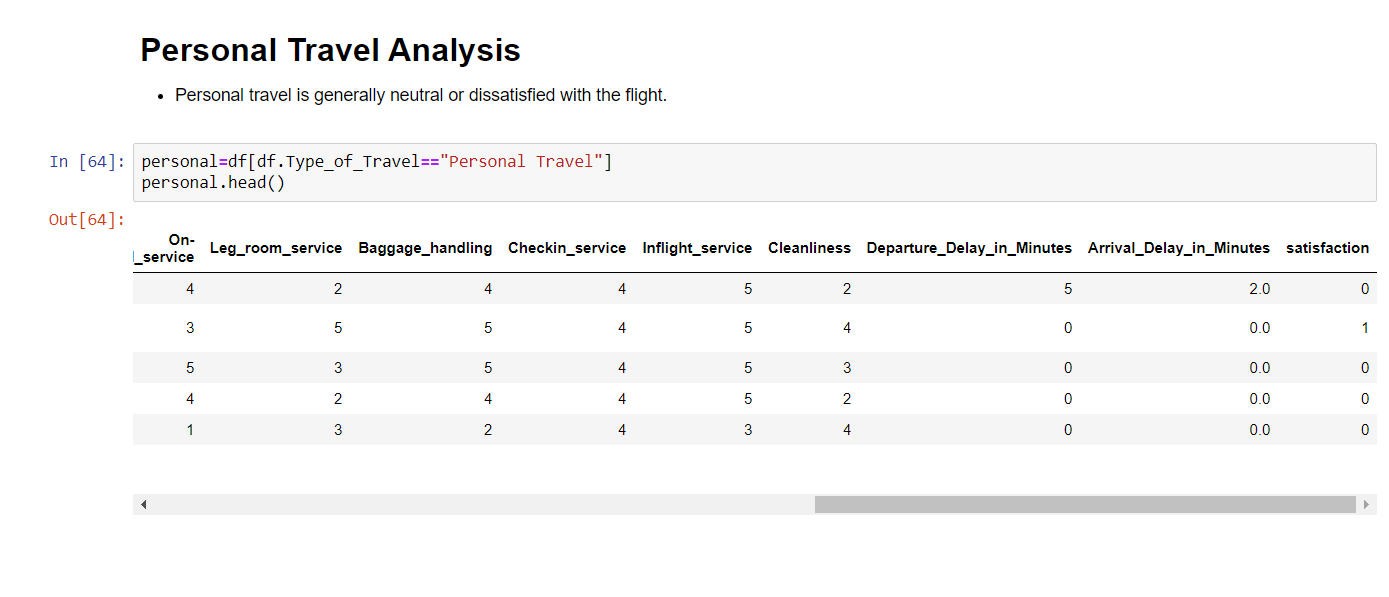


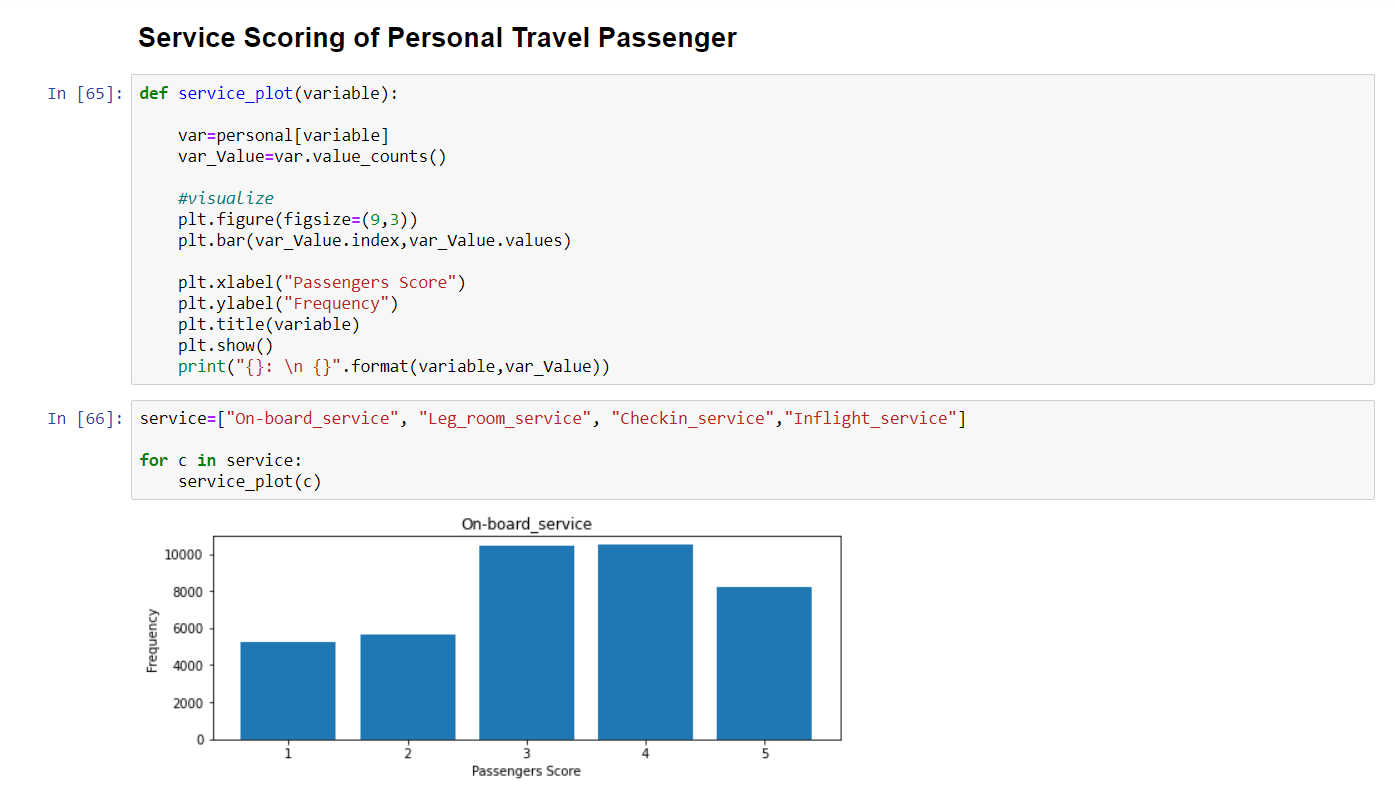


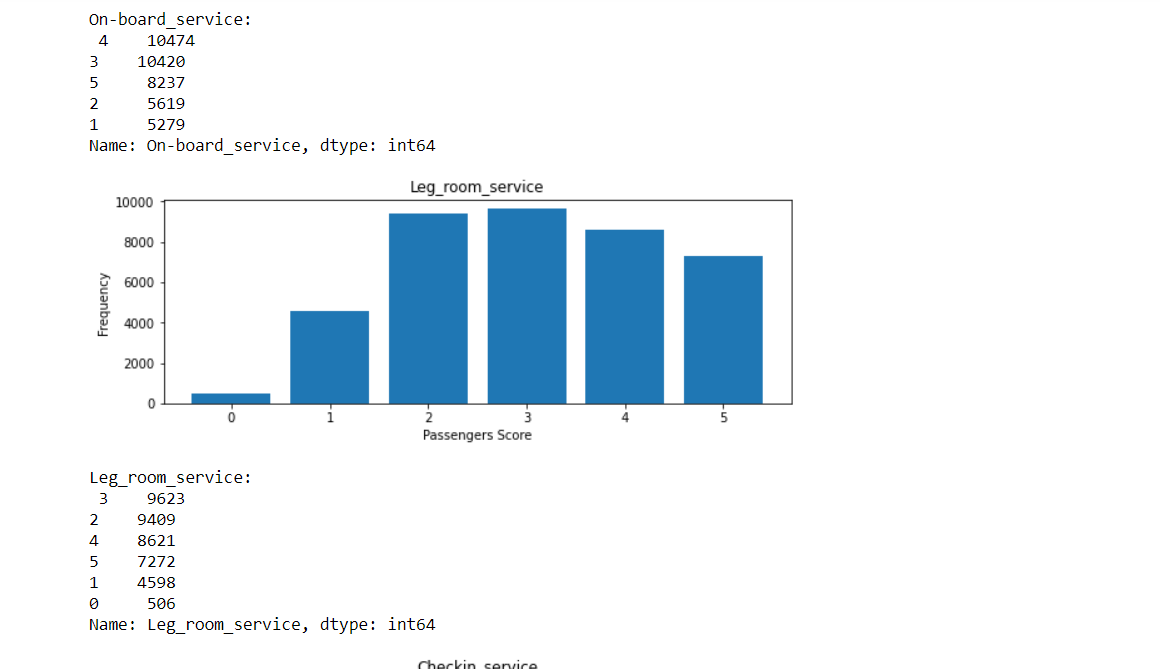


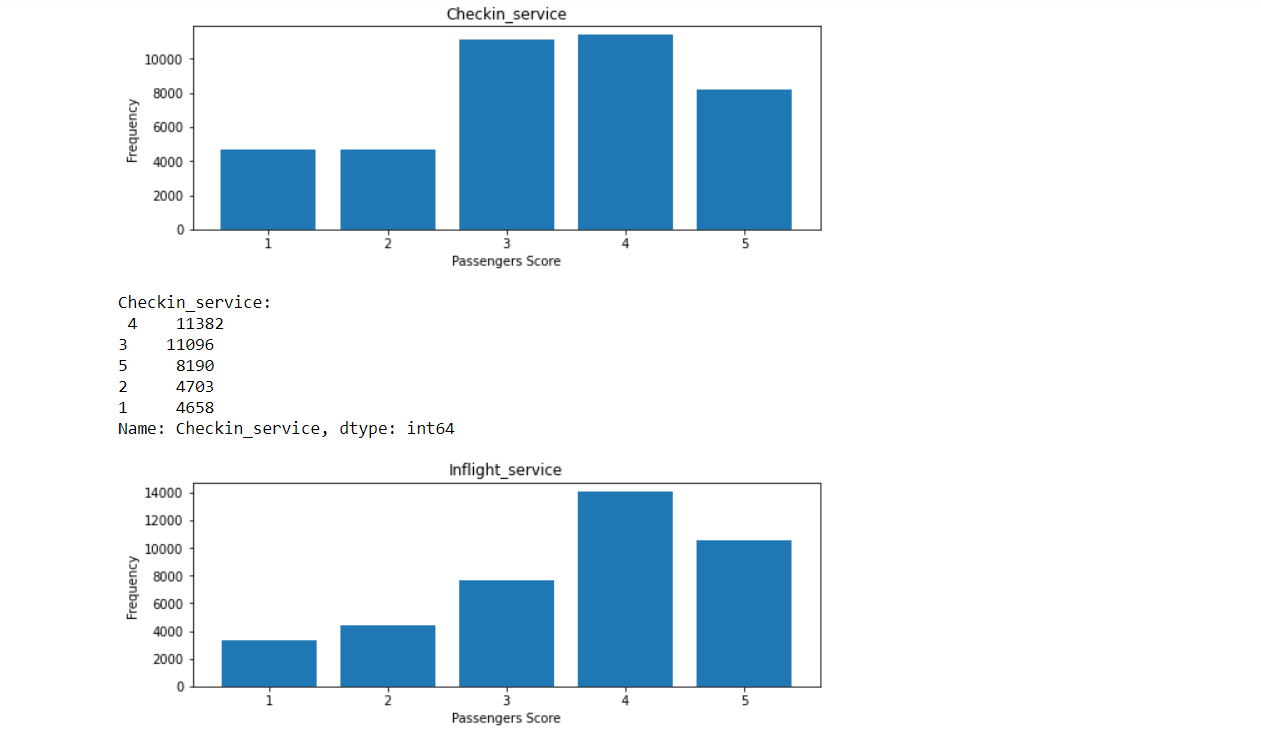
One of the major aspects noted was the satisfaction of the passengers based on the class they were flying in along with their satisfaction with the online boarding. As shown in Figure, the people traveling in the business class who were happy with the online boarding were much more satisfied overall compared to the economy and economy plus classes even if they were happy with the online boarding process. Contrary to this, the economy class people who were not so happy with the online boarding either seems to be highly dissatisfied overall. For economy plus class, online boarding doesn’t seem to be affecting their satisfaction level which means their ratings depend on the factors other than the class and online boarding.

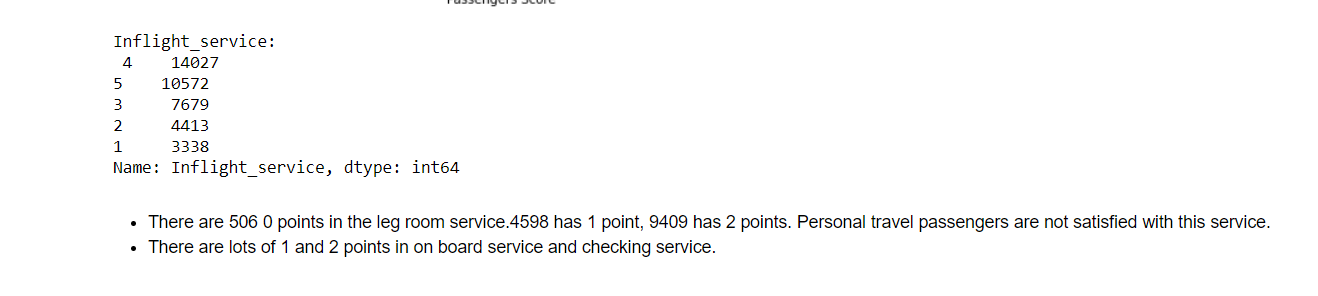




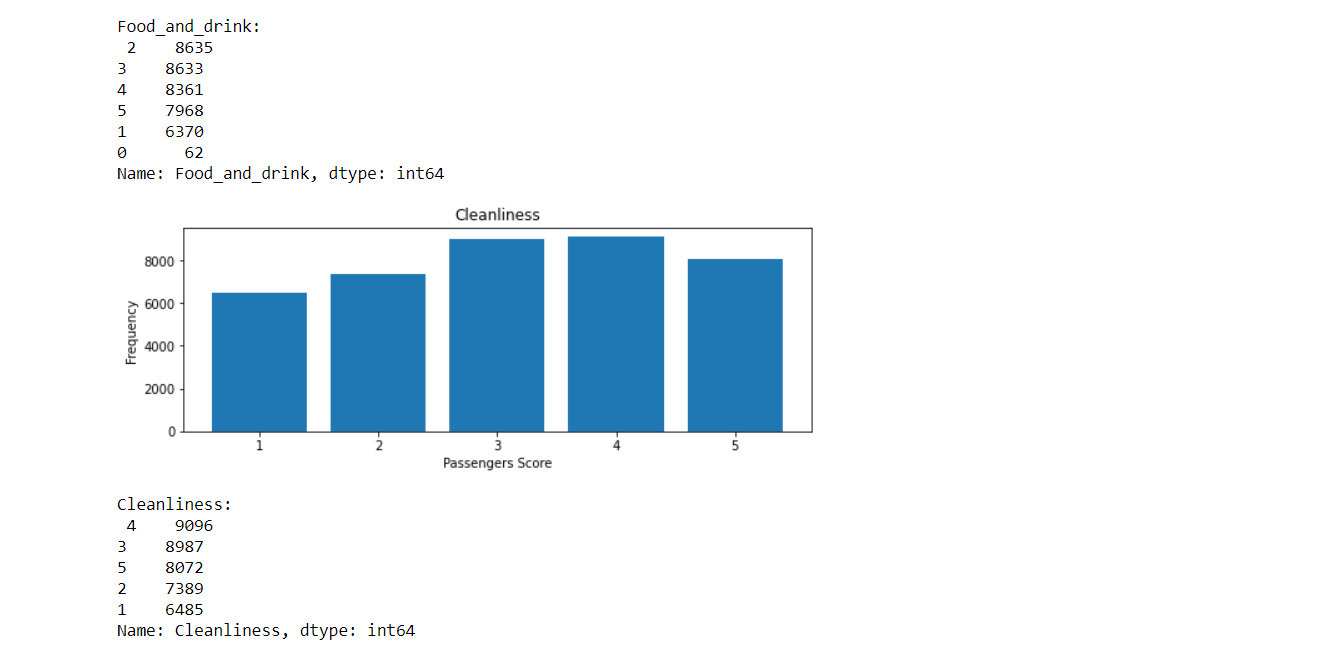


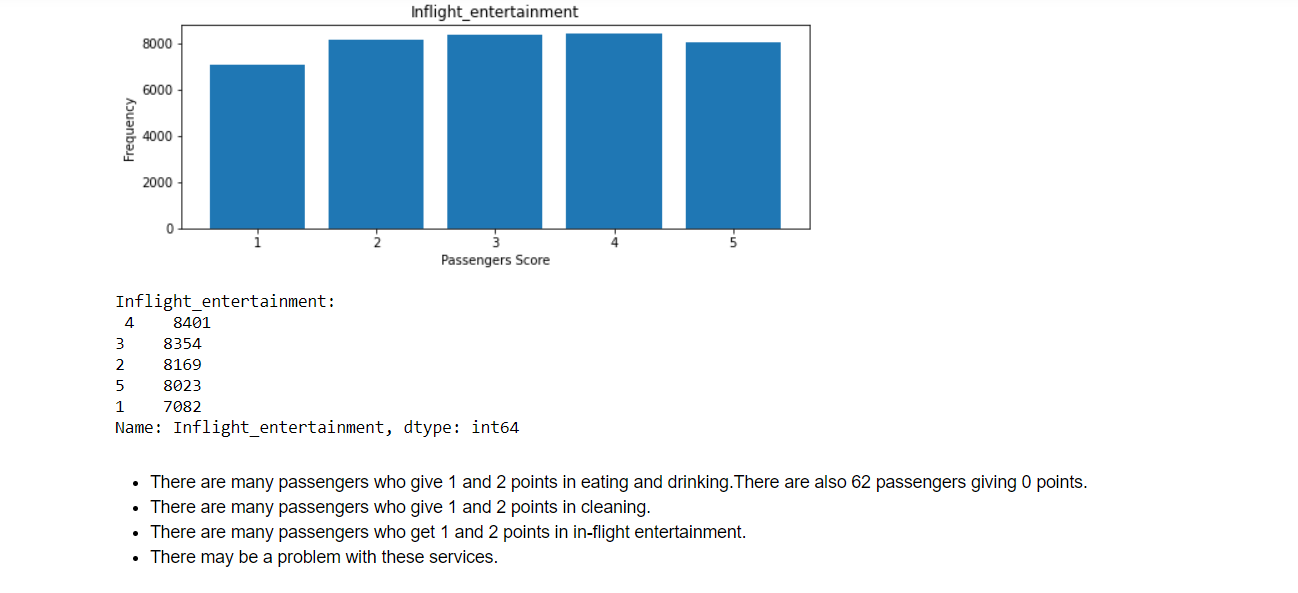




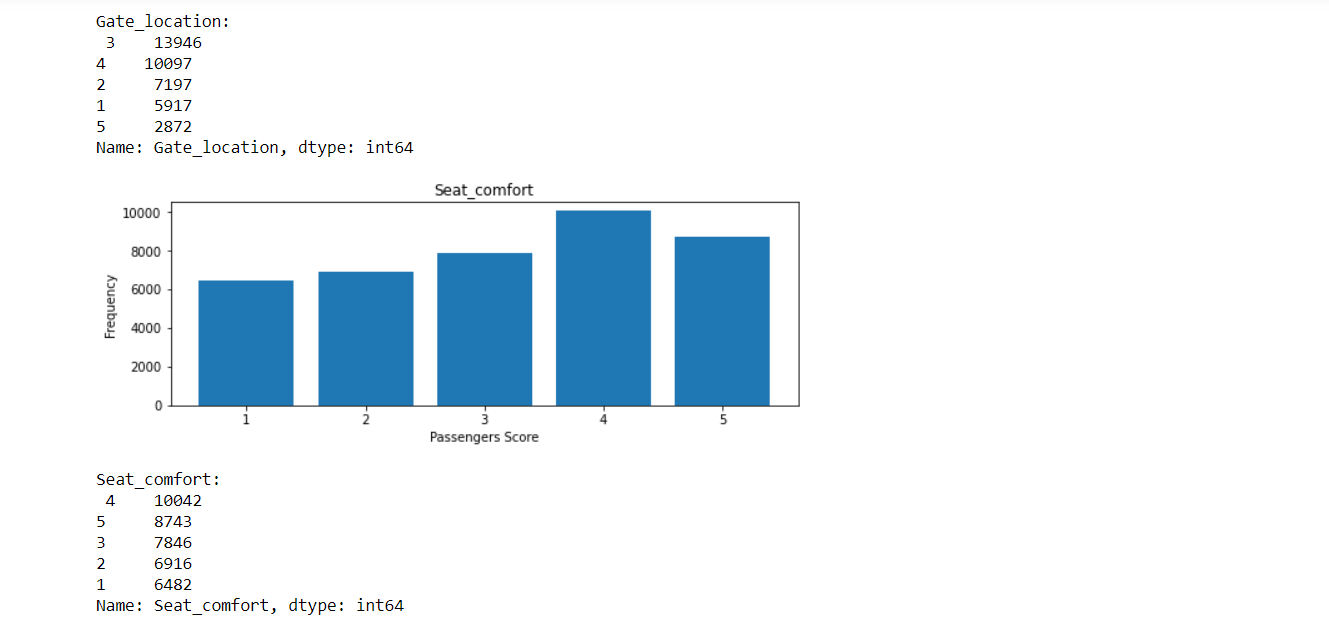


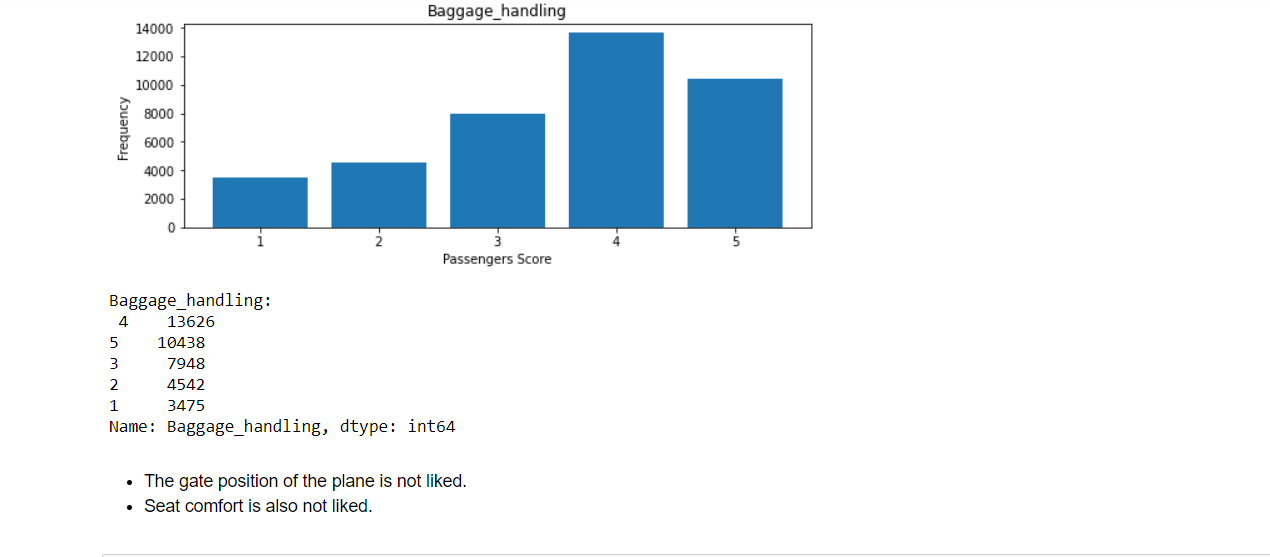




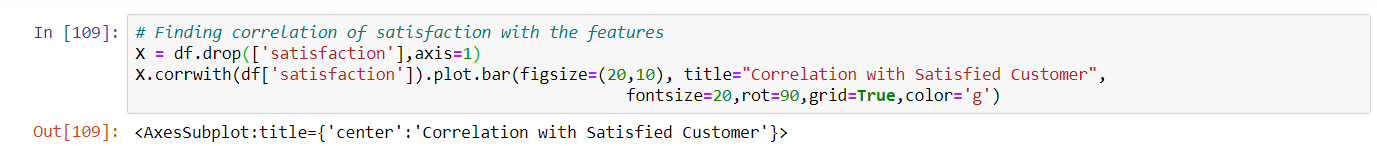


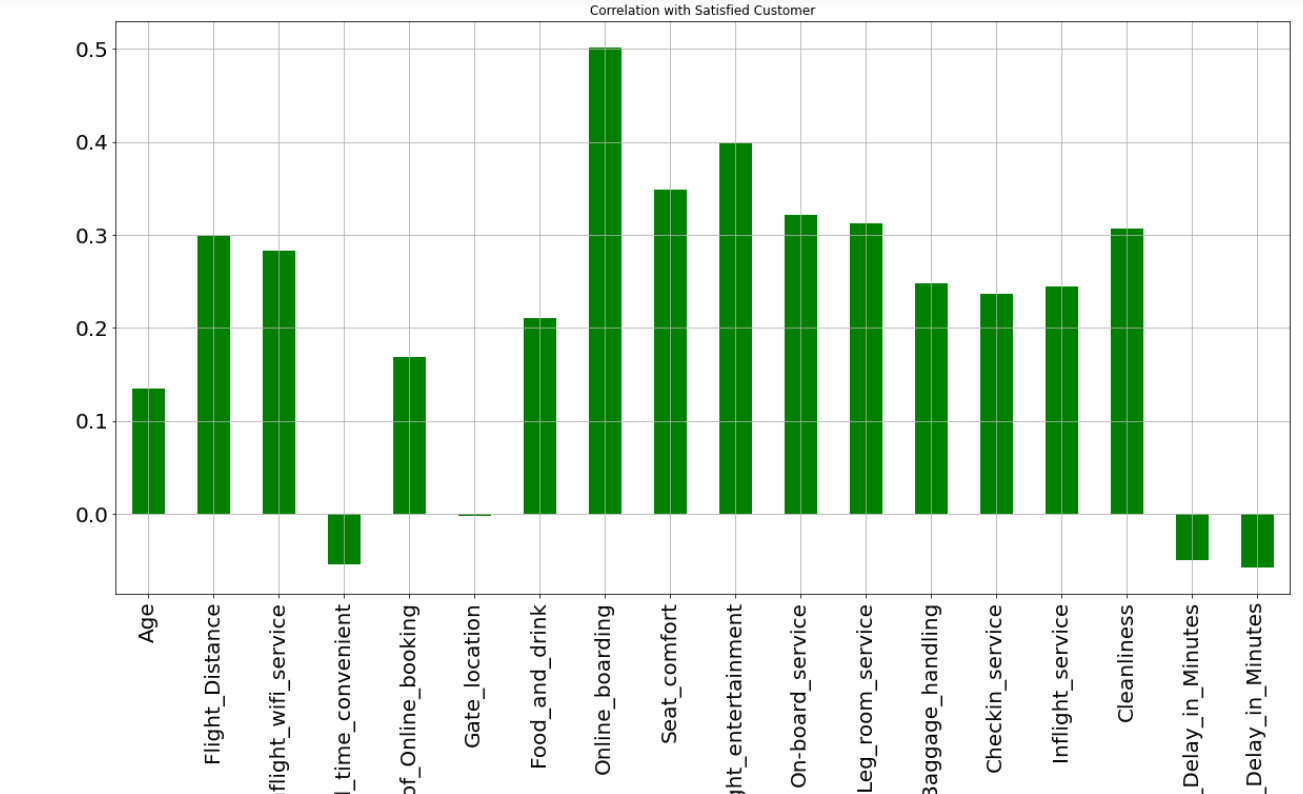






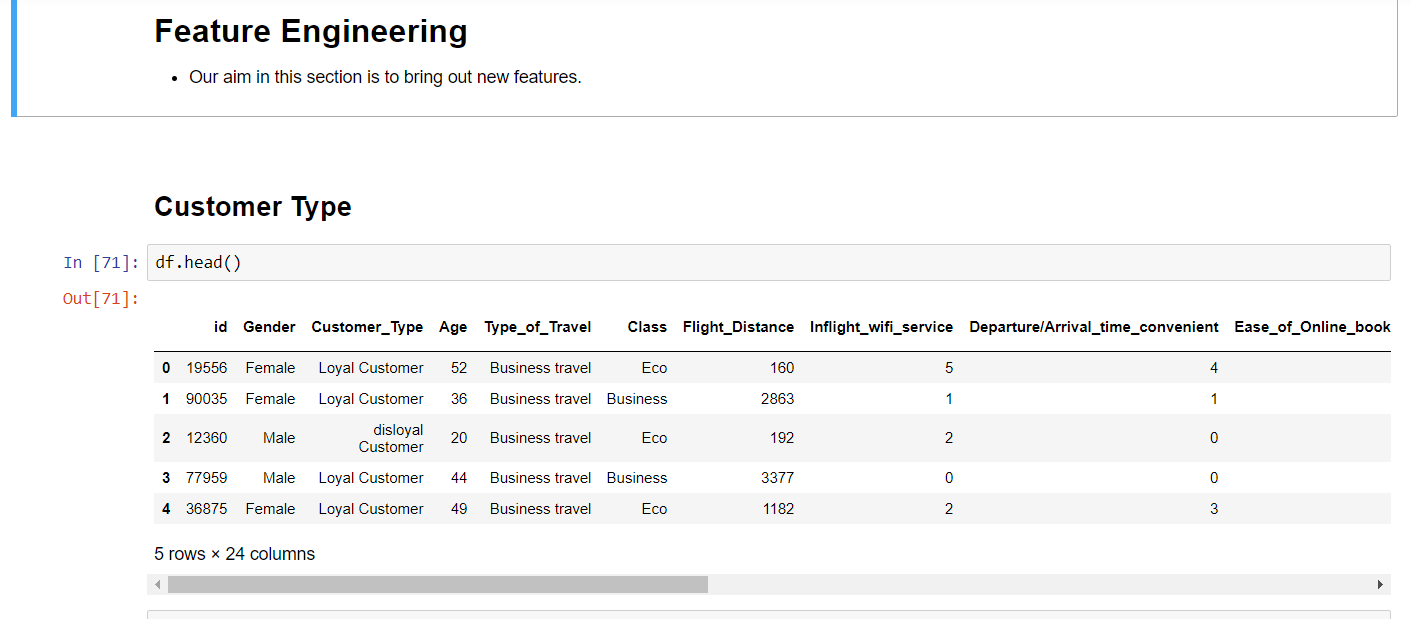
In terms of correlation with our response variable ‘Satisfaction’, some of the features such as ‘Age’, ‘Departure Delay’, ‘Gate Location’ seem to have a negligible influence on the satisfaction level of a passenger as can be seen from the visualizations. It seems that the gate location does not impact the flight experience of a passenger but departure delay’s less impact on the satisfaction level was opposite to our expectations. One of the possible reasons can be that the passengers understand that the delays are mostly never intentional and hence do not base their satisfaction judgment on that. Most of the flights in the survey are for less than 1000 km. The percentage of dissatisfaction amongst disloyal passengers is more than the same amongst the loyal passengers. Another key aspect is the type of traveling where the people going for a business trip are highly satisfied compared to the people traveling personally which seems likely because most of the business trips are funded by the businesses and hence passengers are happy there. Also, the travelers who liked inflight entertainment tend to be more satisfied overall.

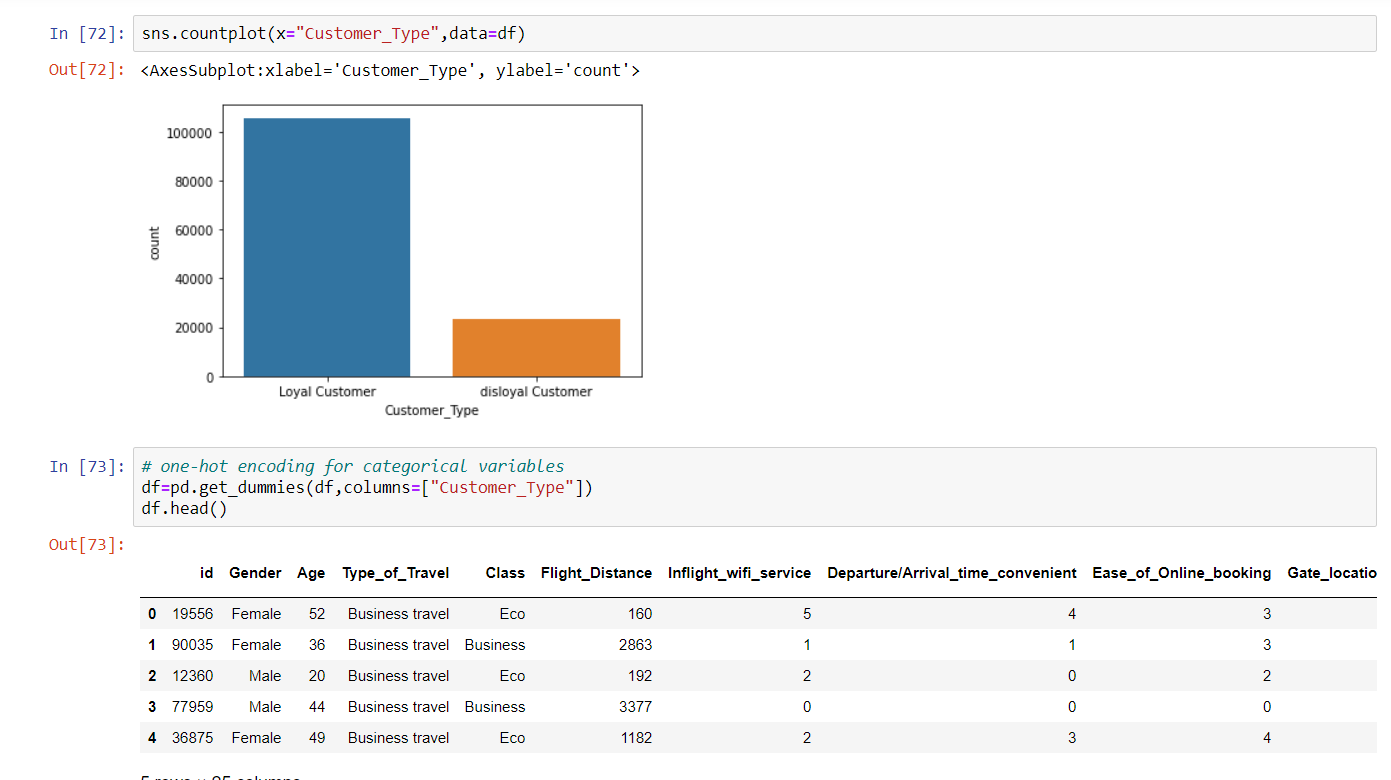


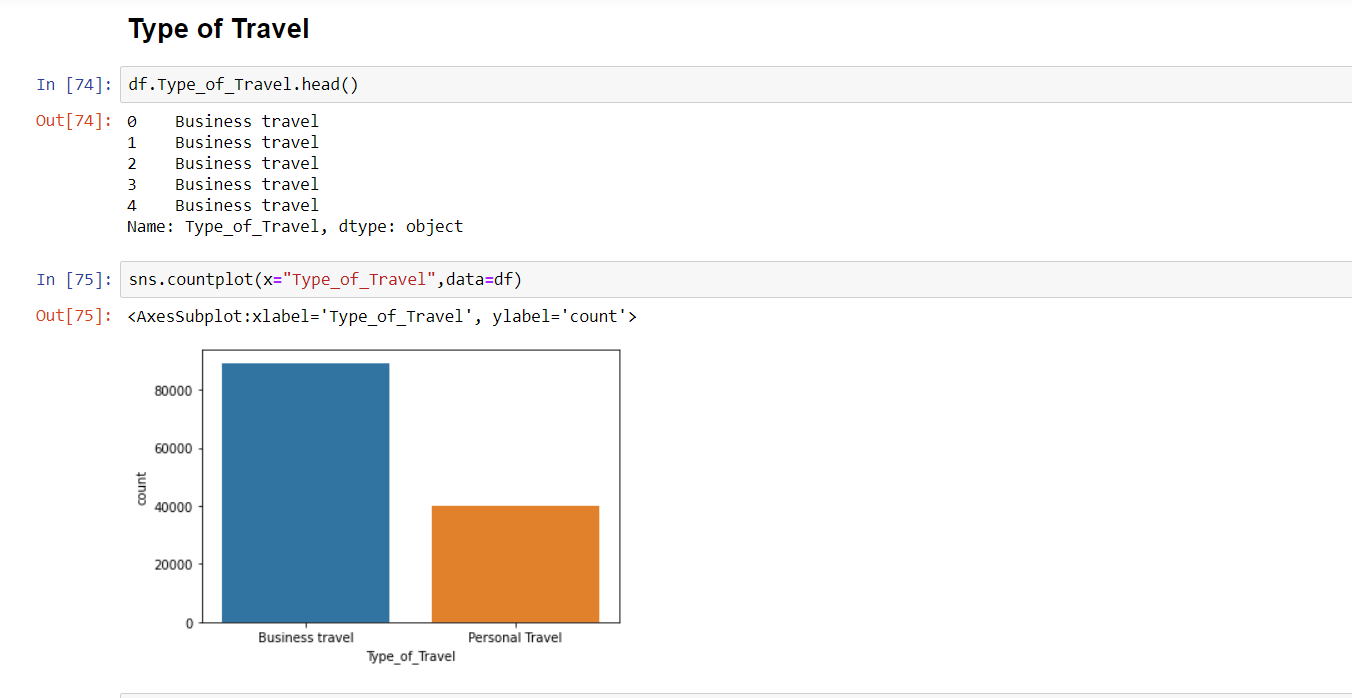


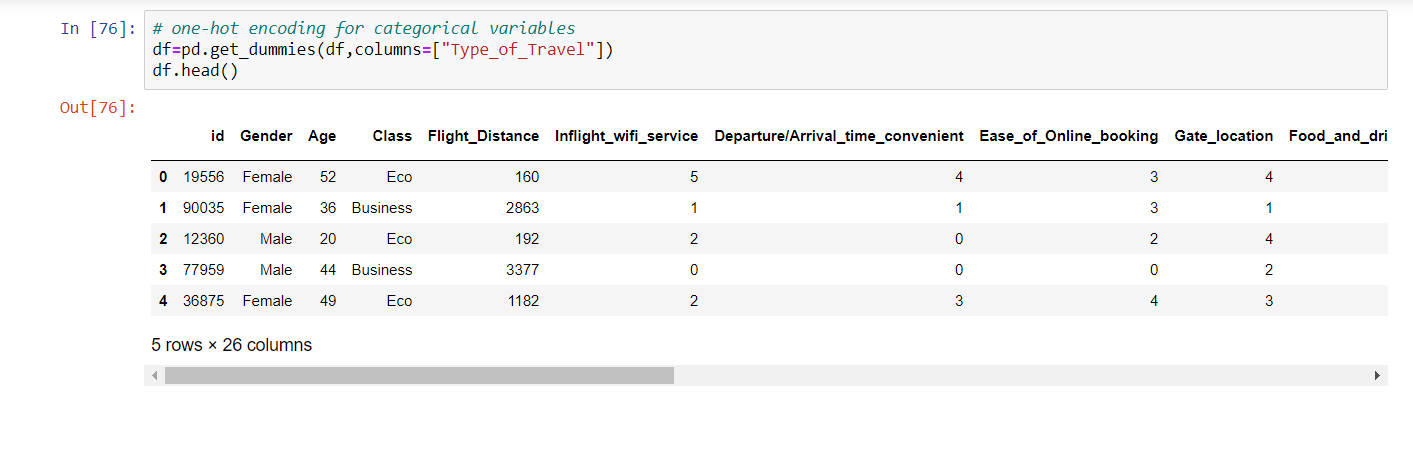
# FEATURE ENGINEERING

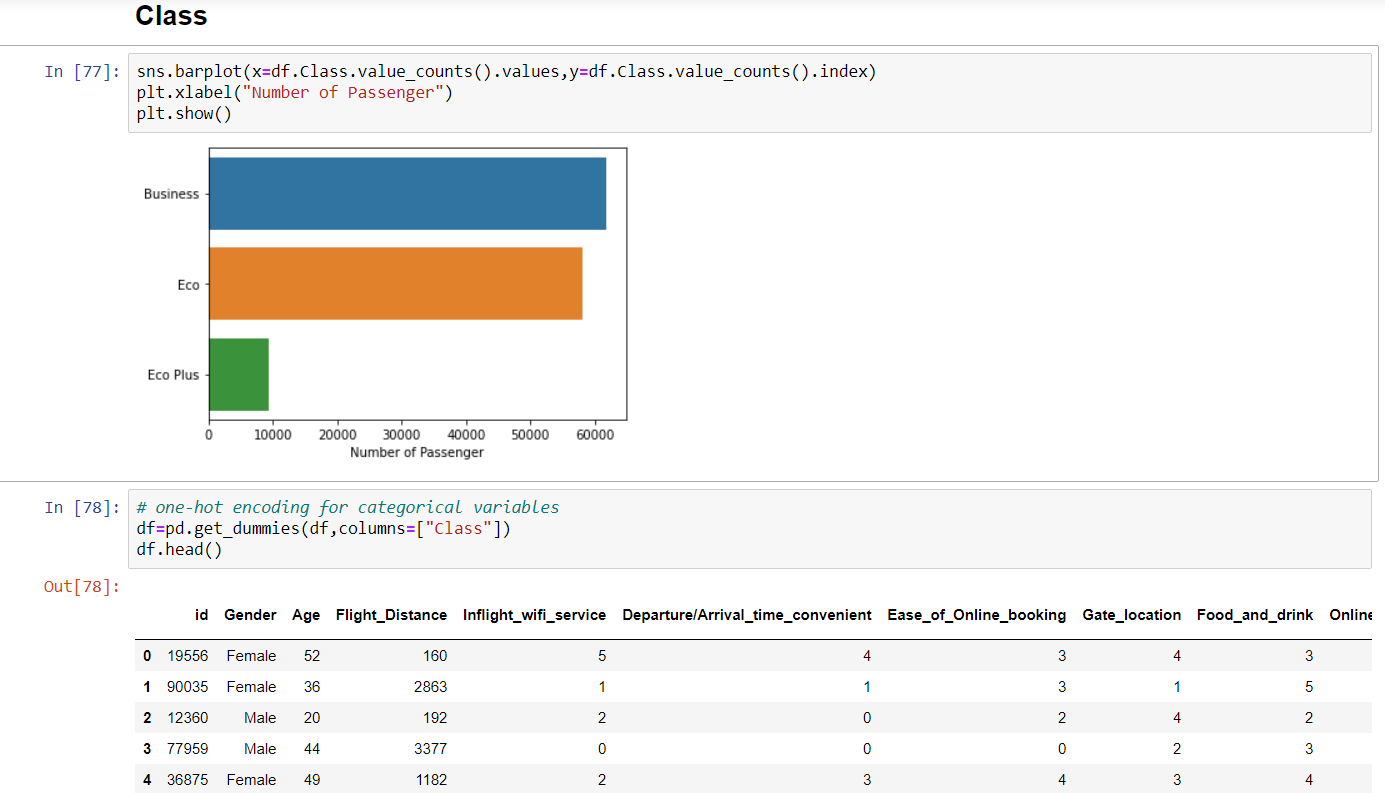
Even the raw dataset has features. Most of the time, the data will be in the form of a table. Each column is a feature. But these features may not produce the best results from the algorithm. Modifying, removing, and merging these features results in a new set that is more suitable to the algorithm training. Feature engineering in machine learning is more than just choosing and transforming the required features. Not only does feature engineering prepare the dataset to be consistent with an algorithm, but it also improves the efficiency of machine learning models. Feature engineering is a critical part of this process. Without this step, the accuracy of your machine learning algorithm will be drastically reduced. A typical machine learning begins with data collection and exploratory analysis. Data cleaning comes next. This phase eliminates redundant values and corrects mislabelled classes and features. Feature engineering is the next step. The performance of the feature engineering is fed to the predictive models, and the results are cross-validated. An algorithm that feeds raw data is unaware of the importance of the features. It is making predictions in the dark. You can think of feature engineering as the guiding light in this scenario. When you have relevant features, the complexity of the algorithms reduces. Even if you use an algorithm that is not optimal for the situation, the results will still be accurate. Simpler models are often easier to understand, code, and maintain.

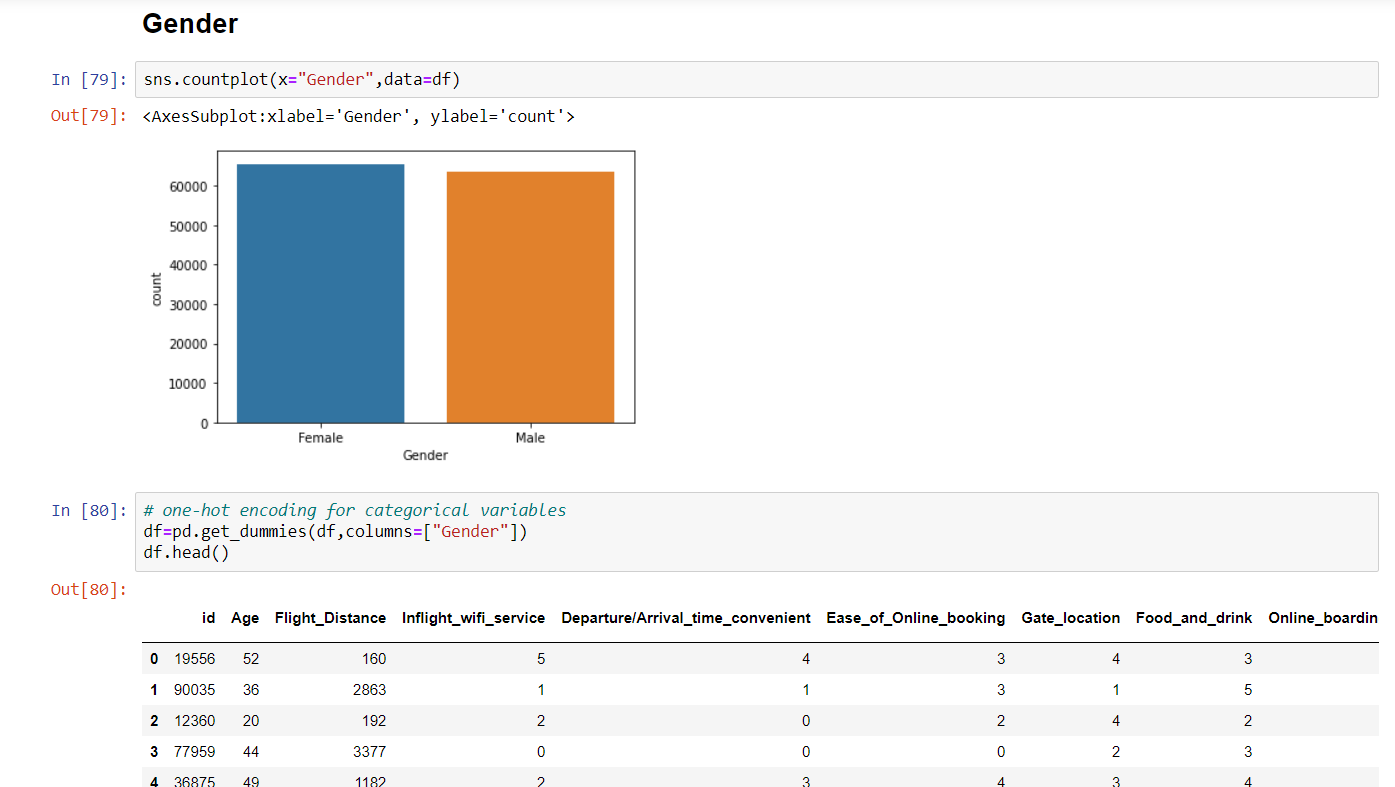












# ACCEPTANCE

Date:01-12-2020

By initialing and signing each page below, I, Sonal Raghuvanshi, as a student at Durham College, agree to and accept the terms outlined in this Statement of Work.

*(Durham College)*

By: Sonal Raghuvanshi