ACIT 4880 - Data Analytics - Project 2

US Airline Flight Passenger Satisfaction [Dataset]

Group 2: Mahsa Taer, Maryam Taer
Instructor: Motasem Aldiab

November 2021

Dataset Source and explanation:

Problem Statement

Flight Passenger Satisfaction dataset helps the Airline to analyze their customer satisfaction status based on the level of their satisfaction of the Airline’s offered services as well as some personal information like Age, Gender, etc.

Having knowledge of this has a determining role in the Airline’s future success as they can find the improvement areas and which type of customer each of their services should mostly be targeted to.

Content & Attributes

- **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 (Satisfaction, neutral or dissatisfaction)

Tools Utilized

Most of the data analysis has been done using **Python** and **Jupyter Notebook** which we will demonstrate later in the document. The majority includes Data Cleaning, EDA, Univariate Analysis, Regression, kNN Classification, Random Forest, Decision Tree, and Clustering.

Data Cleaning & Preparation

Data Observation

Before doing any data cleaning, we will do a preliminary observation of the data in the dataset. The first step is to load the CSV file using the *pandas* module into the Jupyter Notebook since we used it for our project. The dataset originally has 103904 rows and 24 columns, and we loaded it as a DataFrame.

Here, there are 2 datasets - test and train - that we downloaded and used. We mostly use the train dataset, but we also clean and prepare the test dataset as well.

Then, for the purpose of observation, we print out the first 14 rows of the dataframe, the types of each column in the dataset including integer, object, and float.

Data Cleaning Preparation

After we realize what kind of dataset and data we are dealing with in the previous step, we start to prepare the data for cleaning.

Missing Values

The first step is to see if there are any missing or null values anywhere in the dataset. After we confirm that there are missing values, we can use `df.isnull().sum()` Command to see which column(s) have missing values and how many missing values there are. In the *Flight Passenger Satisfaction* dataset, there were 310 missing values in the "*ArrivalDelayInMinutes*" column of the train dataset and 83 missing values in the same column in the test dataset.

Now, there are 3 ways to deal with the missing values including removing the rows that have missing values, filling the missing values with either 0 or the mean of the column that has missing values, or imputing the missing values.

Normally, imputation of the data is the best way, but since it is not covered in the course yet, we decided to fill the missing values with the mean of the "*ArrivalDelayInMinutes*" column which seemed the best way.

Rename and change the Type of Some Columns for Better Analysis

Before we start identifying the outliers, we rename the columns to work with the names easier. We just removed the space between the words in the names of the columns. Then, we change the type of some of the columns that could be categorical to use them in other stages of our project like regression.

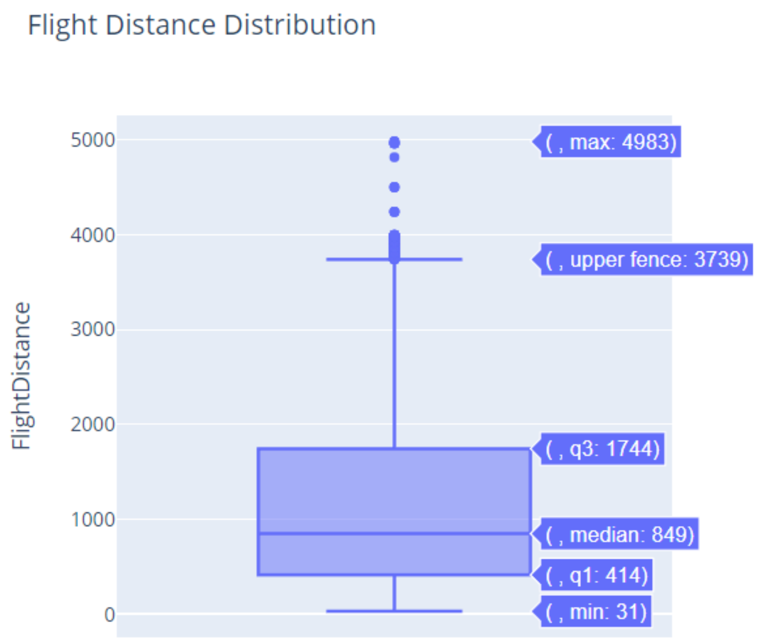
Identifying and Removing Outliers

In this step, we used the `df.describe()` command to get the descriptive statistics of the whole dataset and see if we could spot an outlier.

Then, we got the skewness of the "*FlightDistance*" column and its descriptive statistics, and as we can see in our Jupyter Notebook, the skewness is **1.1** which means the "*FlightDistance*" column is right-skewed and has outliers.

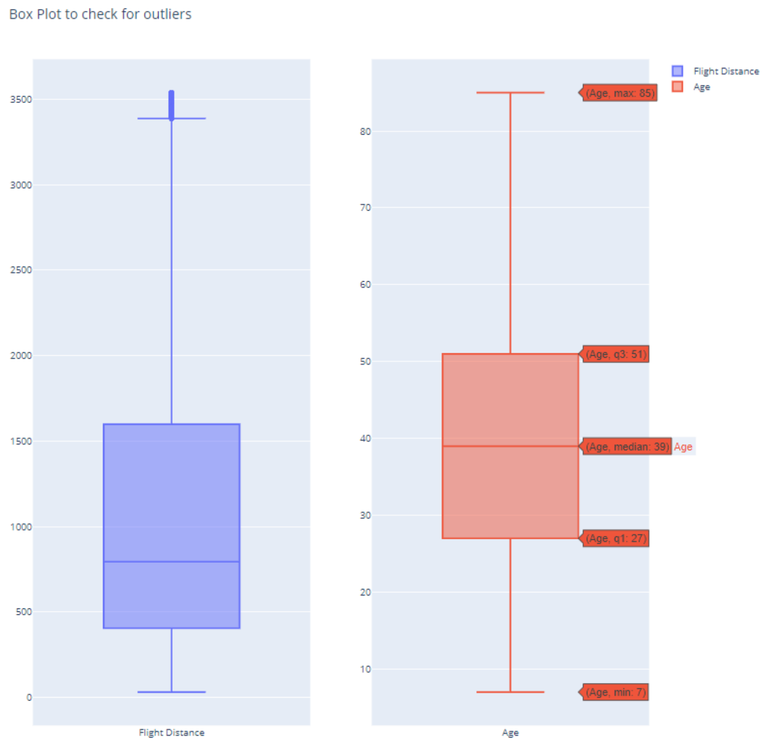
Another way to spot the outliers and get the exact amount of outliers is using *boxplots*. In our project, we used `plotly` and `plotly.express` modules to create the *boxplot* since it represents the outliers and their exact value in a better graphical way. (If plotly module is not installed on your anaconda, it needs to be installed by navigating to the anaconda command prompt and opening it as administrator and run `conda install -c plotly plotly=4.8.1` command)

As you can see in the figure below there are some outliers in the "*FlightDistance*" column.



So, we remove all the rows that are considered outliers and their FlightDistance is more than 3201 Km.

We also got the boxplots for all the continuous variables including "*Age*". As you can see below, "*Age*" does not have outliers.



After removing all the outliers, all the boxplots will be cleaned of outliers, and we will have a cleaned dataset. We exported the cleaned datasets to other CSV files called Airline_Passenger_Satisfaction_Test_Cleaned.csv and Airline_Passenger_Satisfaction_Train_Cleaned.csv so that when we need the clean datasets in the next steps to do the analysis, we would not have to run all the steps of data cleaning from top to bottom.

EDA

Having a cleaned dataset without any anomalies, we started using the new copy of the dataset for visualization and EDA analysis of the data.

Generally, there are a considerable amount of data virtualization tools such as tables, plots, pie charts, stacked bar charts, histograms, scatterplots, and so on. For this project, we used all said virtualization as well as python data analytics modules such as pandas, seaborn, matplotlib, and NumPy.

Using Pie chart, we found that nearly 60% of passengers were neutral or dissatisfied with the services that the target Airline offered them. However, gender doesn’t seem to be a determining factor since the percentage of both females and males is close.

From the start, we can notice that most of the customers are leaning towards using **Eco** flight class, where **Eco Plus** follows second and **Business** class is the least popular among the passengers. However, we conducted -using a line chart- those passengers are most dissatisfied with **Eco Plus** where **Business** class interestingly holds the highest satisfaction rate.

Another factor that we found to be important in measuring satisfaction is **Travel Type**. We observed that passengers who choose **Business travel** with this airline, are more leaned towards **Business class** which we previously conducted to have the highest satisfaction rate. In contrast, customers having **Personal travel**, ideally choose **Eco** class. And **Eco Plus** has the **lowest** utility rate in both categories.

Loyal customers seem to be more interested in traveling in a Business class (for Business travel) and Eco class (for personal travel). Comparatively, Disloyal customers do not even bother using this airline for their personal travels.

We also analyzed the services offered by this airline that have an important role in changing the satisfaction level.

Services that greatly reduces satisfaction:

- low-quality wi-fi
- lack of cleanliness
- low quality or lack of ease of online booking
- poor quality food or drinks or lack of them
- low seating comfort

Services that greatly increases satisfaction:

- quality of food and drinks
- online boarding
- quality entertainment onboard (video, music, etc.)
- high-quality service on board
- quality luggage service

Univariate Statistical Analysis

In this part of the project, we used Univariate Statistical Analysis to analyze our cleaned train dataset by loading the "*Airline_Passenger_Satisfaction_Train_Cleaned.csv*" file.

Results from this section analysis shows that:

- We are 95% confident that the population mean of Flight Distance is between ***1018.5048*** and ***1028.4677***. And our margin error is ***4.9814***.
- We are 95% confident that the population mean Arrival Delay Minutes for those passengers who are both loyal customers with the 100 minutes of Departure Delay and are more than 50 in age is between ***84.5678*** and ***108.0988***. Our margin error is ***11.7654***.
    - Here, our estimate for this subset is less precise than the previous one since the margin error is greater.

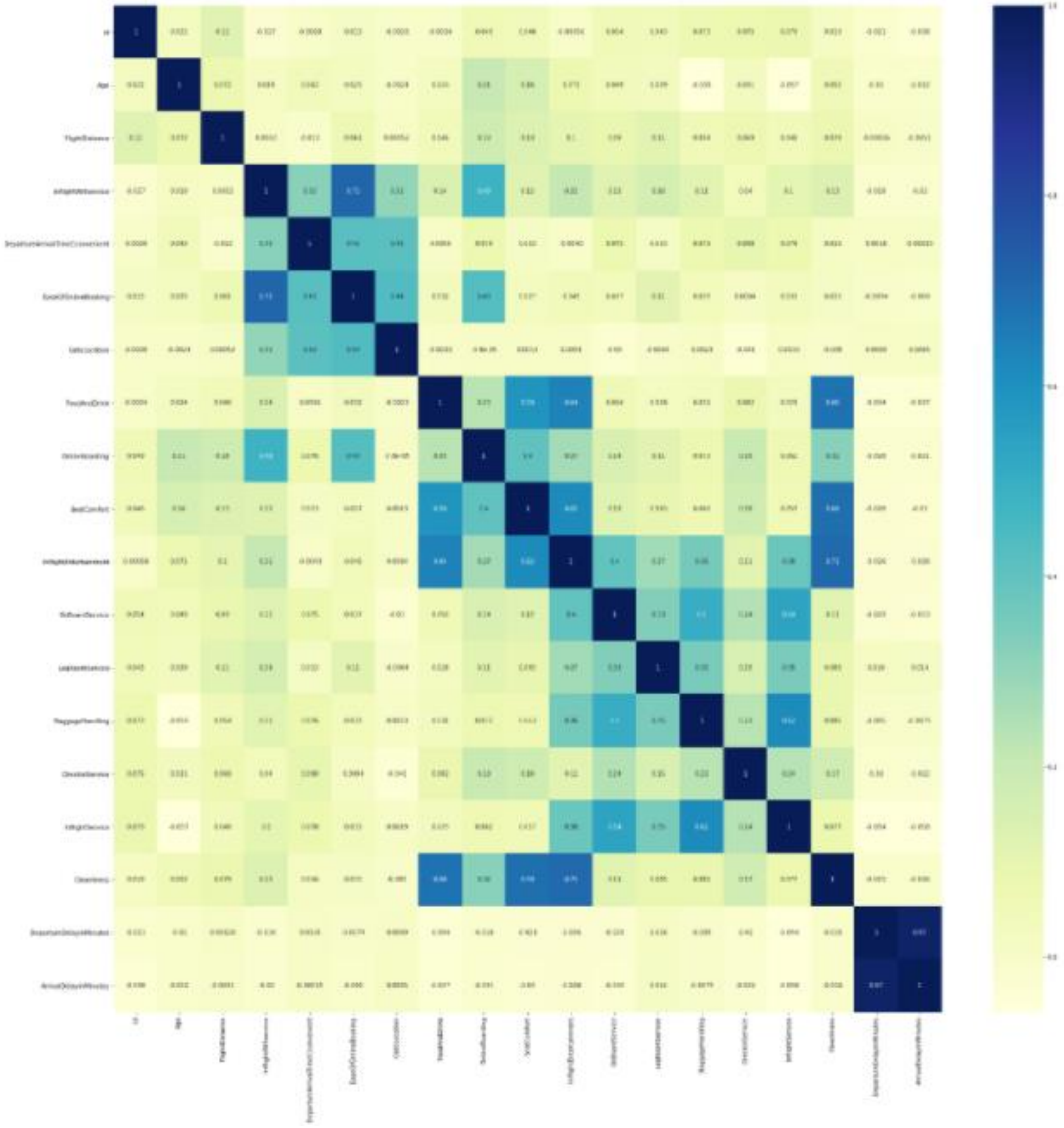
- We are using the FlightDistance column in this step since it is normally distributed and we can see that by using a 95% confidence interval on different sample sizes, as the sample size grows, the margin error decreases. We can gather that the analysis in larger datasets is with fewer errors and more accuracy.
- Here, we are 95% confident that this interval captures the population proportion, and we can estimate the proportion of ***satisfaction*** to within ***0.003*** (or ***0.3%***) with 95% confidence.
- Here, we have a fixed number of total customers (as population) and customers who are satisfied (as a sample) and we see the effect of margin of error on different confidence intervals. As we can see, the more the confidence level, the lower the margin of error will be, and more accurate.
- Here, we tested a null hypothesis to see whether the mean Arrival Delay Minutes for those passengers who are both loyal customers with the 100 minutes of Departure Delay and are more than 50 in age differs from ***2.4*** with a level of significance ***α=0.05***. As a result, the p-value is greater than 0.05 which means that *the null hypothesis is not rejected.*
- Here, we tested a null hypothesis to see with 142 of 97248 customers who were satisfied differed from ***0.15*** with a level of significance ***α=0.10***. As a result, z_data is less than 0 and the p-value is also less than 0.10 which means that we *would reject H0*.

Regression and Prediction

In this part, we use regression which is a supervised method for our data mining analysis on this dataset. Generally, we can use regression and regression models to predict what the target variable would be considering the predictors with values that we choose for them.

Before creating a regression model for prediction, we created a heatmap for the cleaned dataset to observe the correlation relationship between each variable by using the correlation matrix.

As you can see below, the correlation coefficients in the matrix can be in the range of -1 to 1 where -1 is a perfectly negative correlation and 1 is a perfectly positive correlation. This heatmap visualizes how two variables in the dataset move together, either positively or negatively. (Since the dataset is large and has many data, the heatmap has a large scale. The actual map can be found in our Jupyter Notebook - Assignment #2 - Regression & Prediction)



The correlation coefficients can help predict our target variable with linear regression.
For regression purposes, we modified the cleaned dataset and changed the data in the columns that were string to integers so that it would be easier to work with and easier to analyze and predict using regression models.

Here, we started the process of creating a model for linear regression. For linear regression, we wanted to predict customers' Flight Distance considering other variables in the dataset.

First, we chose our predictors and target variables as the whole dataset and Flight Distance as the target.

Then, we split the dataset into 2 sets of train and test variables for X and Y, X being the input or predictors and Y being the target. We basically use the train set to train the variables based on the regression we create so that we can use the test set, being 25% of the dataset, to predict the target based on the predictors.

After we trained our train set, we create a linear regression model to use to predict the test set later. After we created the model, we will fit the train set, and now we can use the model we created to predict the input values of the test set (X for test set).

And we printed the intercept and coefficients at the end. (We included a line of code so that the actual amount (Y for the test) and the predictions would be accessible to compare).

After that, we calculated the MAE, MSE, RMSE, and R2-Square for this test set in linear regression. Then, we predicted the target value for a given set of values to the predictors.

Then, we repeated the same steps to create a regression model using logistic regression for the target of the satisfaction column and predicted some given values at the end.

Classification

Here, we use 3 classification algorithms including Decision Tree, kNN, and Random Forest Classification.

Decision Tree Classification

First, we import the libraries and the cleaned dataset to use, and then we prepare the dataset by changing the data in the columns that have string values to integers to work and analyze better including `‘Gender’`, `‘CustomerType’`, `‘TypeOfTravel’`, `‘Class’`, and `‘satisfaction’`.

Then, we split the dataset to train and test sets to train them, and do the feature scaling. After that, we train the sets using the Decision Tree Classifier module. After creating the model, we can use some data to predict the satisfaction of a customer in a certain scenario.

Then, we can get the predicted set to compare to the actual results and see if they were predicted as expected. Here, we can make the confusion matrix and get the accuracy of the Decision Tree Classifier Algorithm.

And as the last step, we can get the decision tree itself as a visual to see the tree and its depth.

K Nearest Neighbor (kNN) Classification

First, we import the cleaned dataset to use, and then we prepare the dataset by changing the data in the columns that have string values to integers to work and analyze better including `‘Gender’`, `‘CustomerType’`, `‘TypeOfTravel’`, `‘Class’`, and `‘satisfaction’`.

Then, we choose all the columns in the dataset as our predictors except `‘satisfaction’` which is the target that we want to predict. Then, we split the dataset to train and test sets to train them, and after that, we train the sets using the *KNeighborClassifier* module.
Then, we can get the accuracy of the kNN Classification Algorithm. After creating the model, we can use some data to predict the satisfaction of a customer in a certain scenario.

Here, we can make the confusion matrix and get the accuracy of the kNN Classifier Algorithm, and the Area Under Curve (AUC) which seems to be around 78%.

Random Forest Classification

First, we import the libraries and the cleaned dataset to use, and then we prepare the dataset by changing the data in the columns that have string values to integers to work and analyze better including `‘Gender’`, `‘CustomerType’`, `‘TypeOfTravel’`, `‘Class’`, and `‘satisfaction’`.

Then, we choose all the columns in the dataset as our predictors except `‘satisfaction’` which is the target that we want to predict. Then, we split the dataset to train and test sets to train them, and do the feature scaling. After that, we train the sets using the Random Forest Classifier module.

After creating the model, we can use some data to predict the satisfaction of a customer in a certain scenario.

Then, we can get the predicted set to compare to the actual results and see if they were predicted as expected. Here, we can make the confusion matrix and get the accuracy of the Random Forest Classifier Algorithm.

Clustering

Here, we use K-Means Clustering to cluster our train dataset and categorize them for better analysis.

K-Means Clustering

For the purpose of this clustering, we set the objective to find the services offered by the airline that are most likely to decrease/increase the satisfaction level of the passengers.
Early in the clustering for overall satisfaction level, we observed that cluster 2, cluster 1, and cluster 0 respectively have high satisfaction of the airline (most of the **cluster 2** were satisfied, most of the **cluster 0** were dissatisfied, and **cluster 1** is nearly neutral between the two).

Having that in mind, we also analyzed customer types in these clusters and our observation was that cluster 2 again proceeds to have the highest number of loyal customers and cluster 0, highest disloyal customers. We can here conclude that loyal customers are more likely to be satisfied with the airline which makes sense. Hence the airline should increase the loyalty based on the services they offer.

We also analyzed **Travel type** (**Personal** Travel, **Business** Travel), and **Flight Class** (**Eco**, **Eco Plus**, and **Business**). Our observation was that the majority of cluster 2 (most satisfied among clusters) prefer Business travel with this airline and with Business class. Eco and Eco Plus follow next respectively, however, they have received the most dissatisfaction among clusters 1 and 0 which can mean the airline should address them.

The services that we analyzed for the purpose of this clustering are Inflight Entertainment, Inflight Wifi Services, Ease of Online Booking, and Seat Comfort.

Early on, we realized that cluster 0 (the most dissatisfied among clusters) had given Inflight Entertainment and Seat Comfort a very low rate (from 1-3) and the percentage of this rate between the passengers in these clusters are also significantly high which can lead us to the speculation that these two services can have a significant impact on decreasing the overall satisfaction level of these customers and thus, making them disloyal customers.

Additionally, for Inflight Wifi Services and ease of online booking, they received a very high rate (4-5) from cluster 2 which can mean that these two services can have a significant impact on increasing the overall satisfaction.