2022

Flight Delay Prediction

ST2195_COURSEWORK

190526753

Table of Contents

1 Introduction

1.1 Background

Flight delays are one of the issues we face today. One of the common aspects we have on the cause of delays is due to weather. However, there are a lot of other causes to consider. Hence this research tries to examine and demonstrate the cause of delay in several aspects.

1.2 Scope of Study

The research will mainly focus on year 2003 to 2005 flight arrival and departure details for all commercial flights on major carriers within the USA. It aims to scrutinize on the flight delays using a data analysis tool called R and python to explore the data and construct a model that could predict flight delays. For questions 1 and 2, data sets excluding the delay time outliers will be used while the rest will be using data sets including all the delay time. This is due to more proportion that the larger delay value could take when calculating the proportion which then could result to affecting the analysis result. Furthermore, for question 5, although we will be going through different kind of algorithm models, we will mainly focus on Random Forest model for further discussion. The models we will be going through to identify the best are, for python, Logistic Regression, Random Forest, and Decision Tree Model while for R, Linear Discriminant Analysis, Random Forest, Classification and Regression Tree (CART) and K-nearest neighbors (KNN). For feature selection, although we could go through different algorithms such as filter, wrapper, embed and XAI method to see the significance of each variable to the model, we will mainly use embedded method to identify. However, when considering the features for machine learning model, we will base it on the analysis we have gone through.

2 Methodology

2.1 Purpose of research

The purpose of this research is to present an exploratory data analysis (EDA) that could test our hypothesis and assumptions with its summary of statistics and visualization. By answering the five questions we aim to answer, we would be able to construct a predictive model that could accurately predict delays.

2.2 Approach

With the insights generated when exploring the data, we aim to figure out the cause and effect of the delay and influential variables that have an impact on flight delays.

2.3 Data collections

Primary data acquired from the EDA will be mainly used for the research. However, some of the general data that we would like to compare to its primary data would be gathered from the external secondary data source.

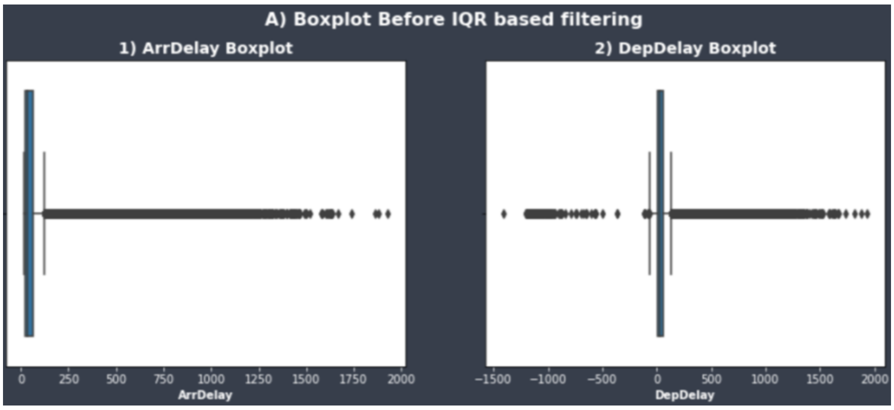
2.4 Data analysis

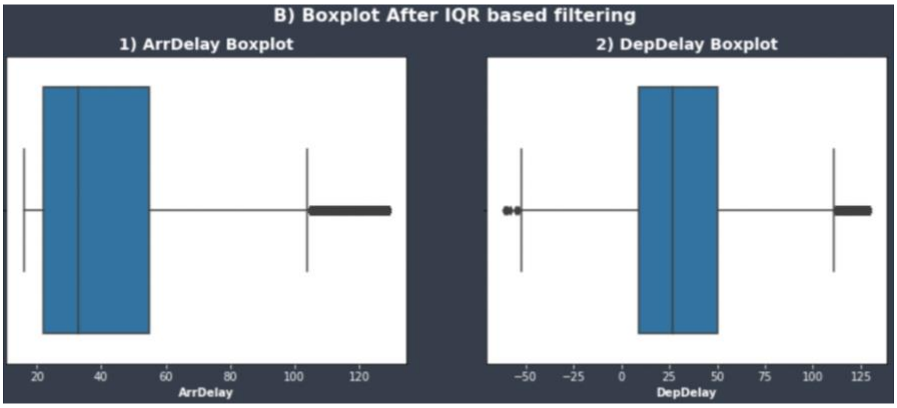
For the data analysis, five questions will be answered to further analyze the flight data. Those five questions are the following: 1. When is the best time of day, day of the week, and time of year to fly to minimize delays? 2. Do older planes suffer more delays? 3. How does the number of people flying between different locations change over time? 4. Can you detect cascading failures as delays in one airport create delays in others? 5) Use the available variables to construct a model that predicts delays. Each question will be a guide to find new insights and come up with an accurate predicting model.

3 Data Analysis

3.1 Data preprocessing

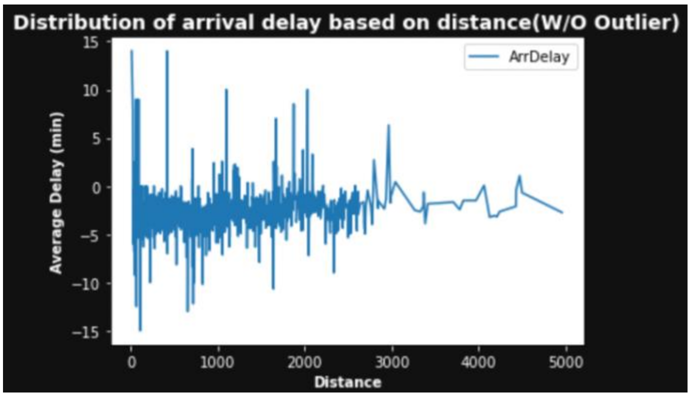
First, we called all the datasets and packages that will be used for the research and concatenated 3 consecutive years flight data which contains specific details of the flight from 2003 to 2005 into one whole data frame. According to Bureau of Transportation statistics, a flight is counted as delayed if the flight operated more than 15 minutes after the scheduled time shown in the carriers' Computerized Reservations Systems (CRS). Hence to answer the questions regarding the flight delays, we classified each of the flights into on-time or delayed.

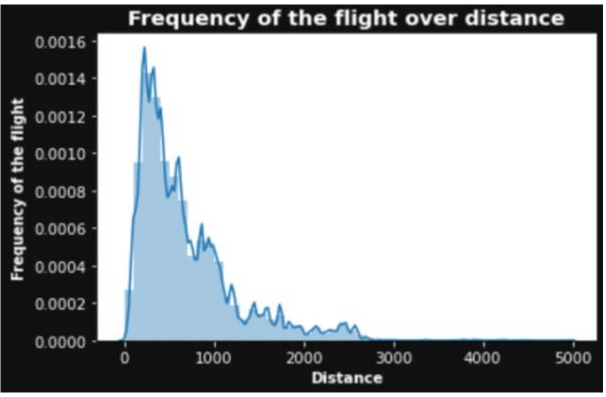




Furthermore, we handled null values and excluded columns that we won’t be needing for the analysis. However, upon exploring the data with delays, we found outliers that may affect the analysis results. To minimize the bias caused by the proportion that the outlier of the delay values may cause, we filtered them out systematically by using IQR based filtering method while keeping the original dataset in place for comparison.

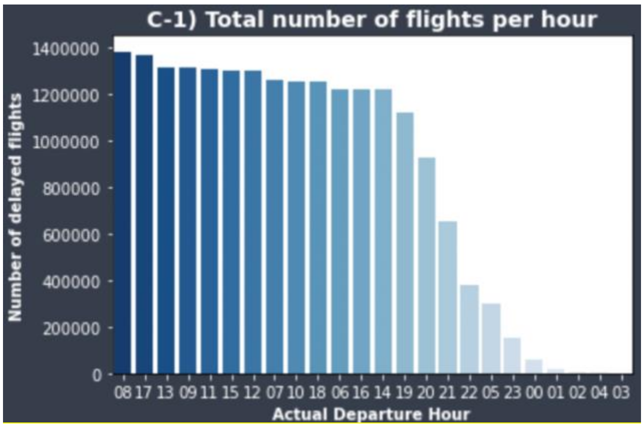
Additionally, we identified that if the flight was cancelled or diverted, there were no delay occurred since the aircraft did not arrive at its destination. Hence, we excluded the rows that were cancelled or diverted. For the time related columns, we converted concatenated hours and minutes into "hh:mm" format and 24-hour notation to make it easily readable and accessible for future use. Moreover, with preprocessed data named “df_copy”, we scrutinized correlations between each variable and deep dived into the dataset to find new insights.

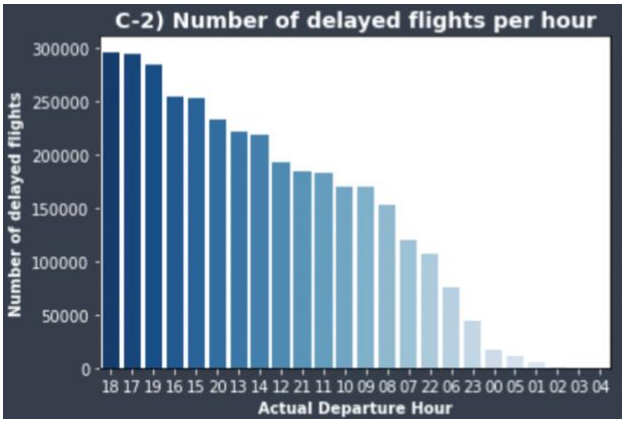




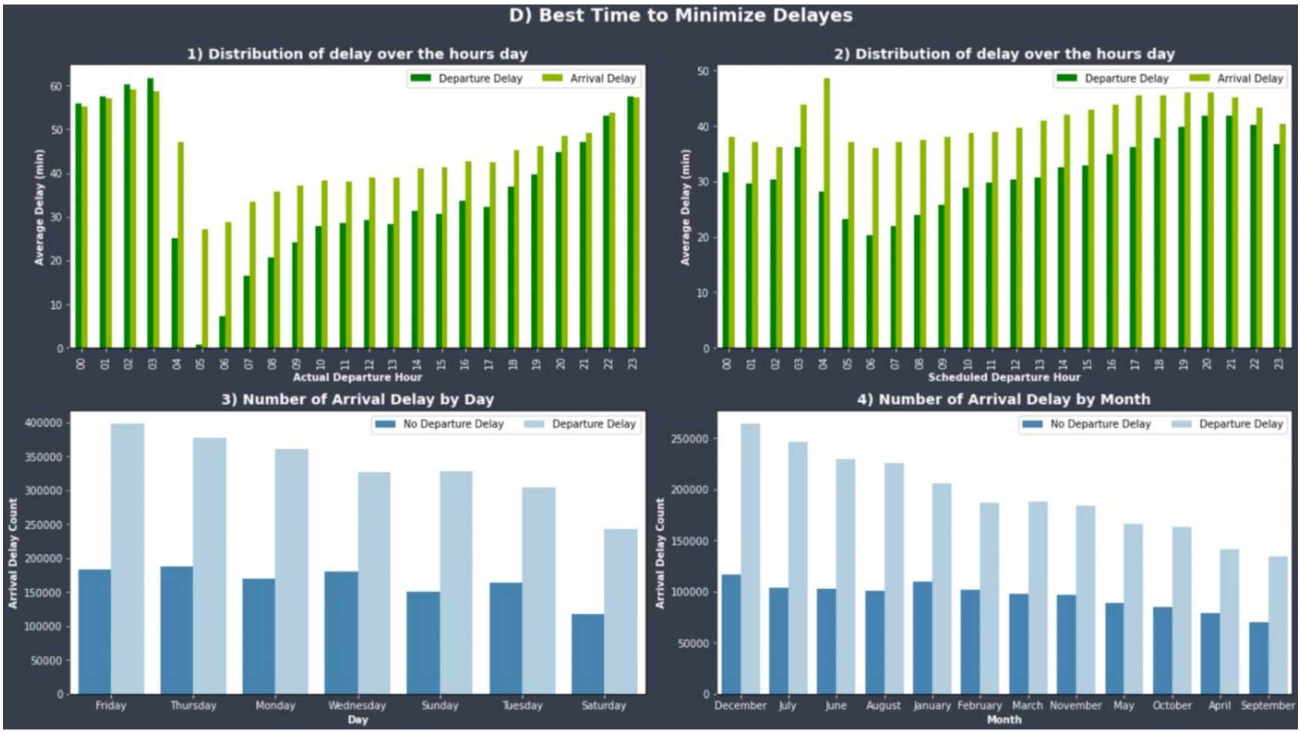
3.2 Question 1 (When is the best time of day, day of the week, and time of year to fly to minimize delays?)

First, we filtered flights with arrival delay and named it as “Delayed_flights”. With 3488440 rows of delayed flights, we made some visualization to clearly see the patterns within the data. For plots D-1 and D-2, we visualized the distribution of actual and scheduled average delay in minutes for each hour of the day. From the plots, we identified that morning flights from 5 to 11 seem to have less delays compared to other departure hour. However, this may be due to lesser flights in the morning compared to other timing. Hence, we visualized the number of flights per hour and found out that even though morning flights and afternoon flights have similar number of flights (see C-1), most of the delayed flights occurred in the afternoon hours (see C-2).



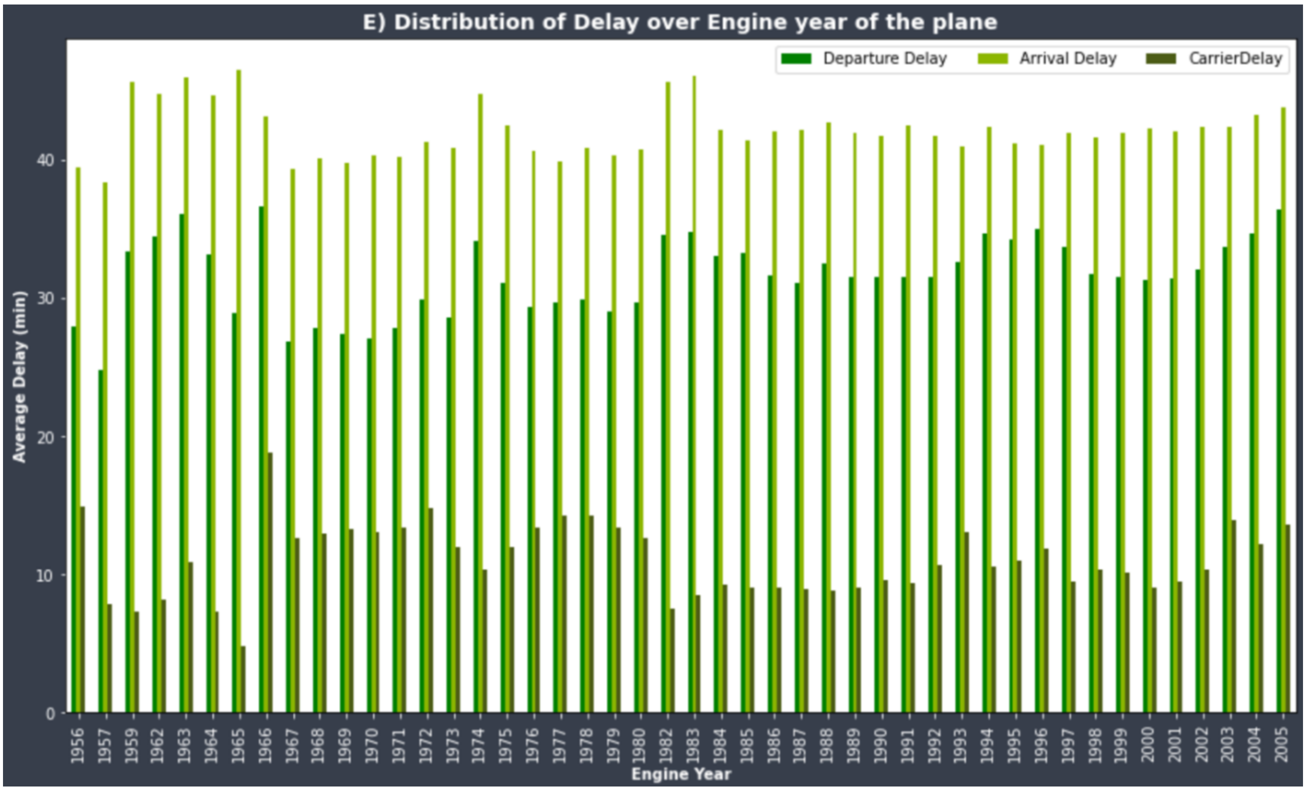


Furthermore, for plots D-3 and D-4, we visualized the number of arrival delayed flights for each day of the week and month with and without departure delay by descending order. From the graph, we identified that weekend seems to have lesser delays compared to weekdays. Also, flights in autumn (September to November) and spring (March to May) tend to have lesser delays compared to months of summer and winter. Hence, as we can see the pattern of delays that are affected by time related variables, we can approximately identify when to book and not to book flights to minimize delays.

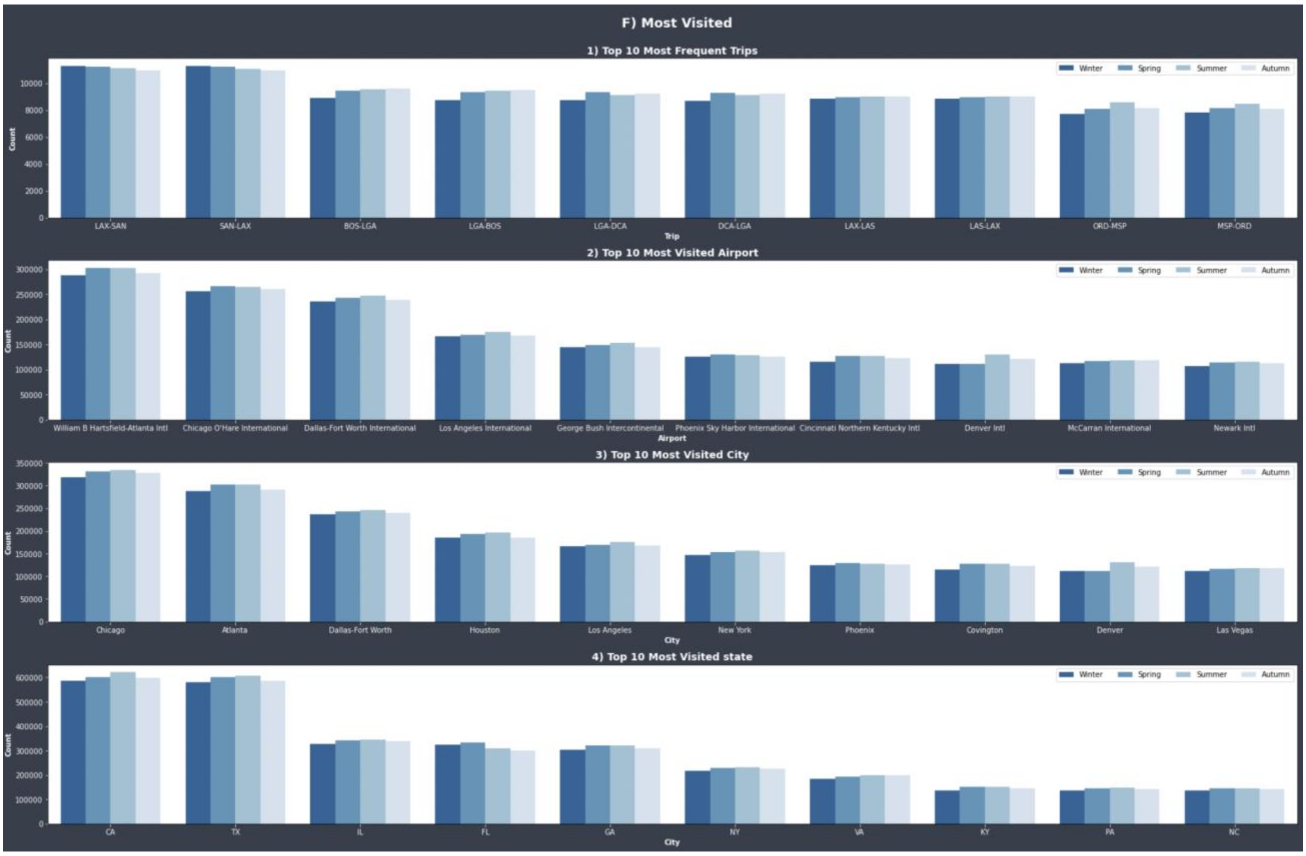


3.3 Question 2 (Do older planes suffer more delays?)

To answer question 2, we merged “plane” dataset which includes details about the aircraft into “Delayed_flights” dataset from question 1 based on the tail number of each aircraft and named it as “plane_characteristic”. Next, from the “plane_characteristic”, we dropped non-indicated engine year values and plotted graph based on the average of departure delay, arrival delay, and carrier delay of each engine year. From plot E, we couldn’t find anything interesting for each engine year as it shows no distinctive patterns of delays. Hence, we can assume that age of the plane does not have great impact on its delays.



3.4 Question 3 (How does the number of people flying between different locations change over time?)

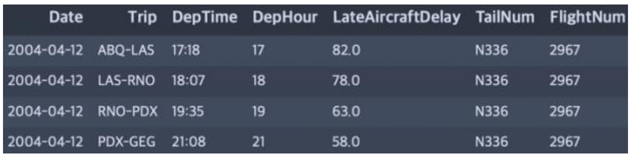


To answer question 3, we merged “airports” dataset which include comprehensive airport details into our preprocessed dataset “df_copy” based on the destination and named it as “Destination”. To see the time trend of the trip, we categorized each flight seasonally based on their month of the flight. Additionally, we concatenated its origin and destination airport to see the route of each plane.

We divided 4 different sections to demonstrate how the number of people fly between different locations change over time. As we can see from F-1 plot, most frequent trip happened in between Los Angeles International Airport (LAX) and San Diego International Airport (SAN) during winter compared to other seasons. This makes sense because temperatures in San Diego are sufficiently warm year-round, hence we can assume that people tend to visit San Diego during winter to get rid of the cold weather and fill the warm air. Similar finding could be applicable for other flight's route too. Moreover, comparing it with the rest of the F plots, we can identify that having the most frequent trip doesn’t necessarily mean it would be parallel to the most visited airport, city, and state since numerous airports are correlated with lots of different city and state. Thus, F-2 plot with William B Hartsfield-Atlanta Intl airport as the most visited airport located in the city of Atlanta is contrasting to F-3 plot showing Chicago as the most visited city. Furthermore, people tend to travel more on spring and summer when the weather is ideal. For summer, it may be due to summer vacation trips. While for spring, considered as off-season compared to summer, could have fewer crowds and the prices of touring may be less compared to summer when prices soar which would then result to encourage people to travel during spring. Hence month and origin of the airport could be an important indicator for prediction model

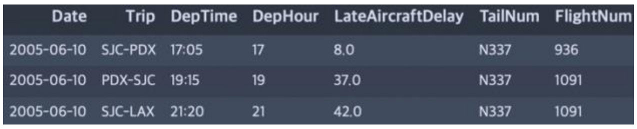
3.5 Question 4 (Can you detect cascading failures as delays in one airport create delays in others?)

First, cascading delay occurs when a delay at previous airport flight causes a ripple effect in the following airport flight. To answer question 4, we used "Destination" dataset from question 3 which includes airport details to sample some of the flights to answer the question. Furthermore, since we are trying to detect cascading failures, we filtered out flights with no late aircraft delay, a delay caused by late incoming of aircraft from the previous airport, and flights with no delayed arrival and departure. To further investigate, we sampled some of the most frequent flights, flights with tail number "N336" and “N337" and demonstrated its route to see if we can detect cascading failures. We sorted departure hour value in ascending order to follow the timeline of the flight for specific date. As we can see in Table 1 “LateAircraftDelay”, the flight departed at 17:18 ended up having 82 minutes of delay. However, as it continues its flight, the delay tends to decrease with the departure time of day. While in table 2, the flight starting with 8 minutes of delay ended up having 42 minutes of late aircraft delay at the end. In fact, the trend seems to be normal and could be also seen with other airports or airline.



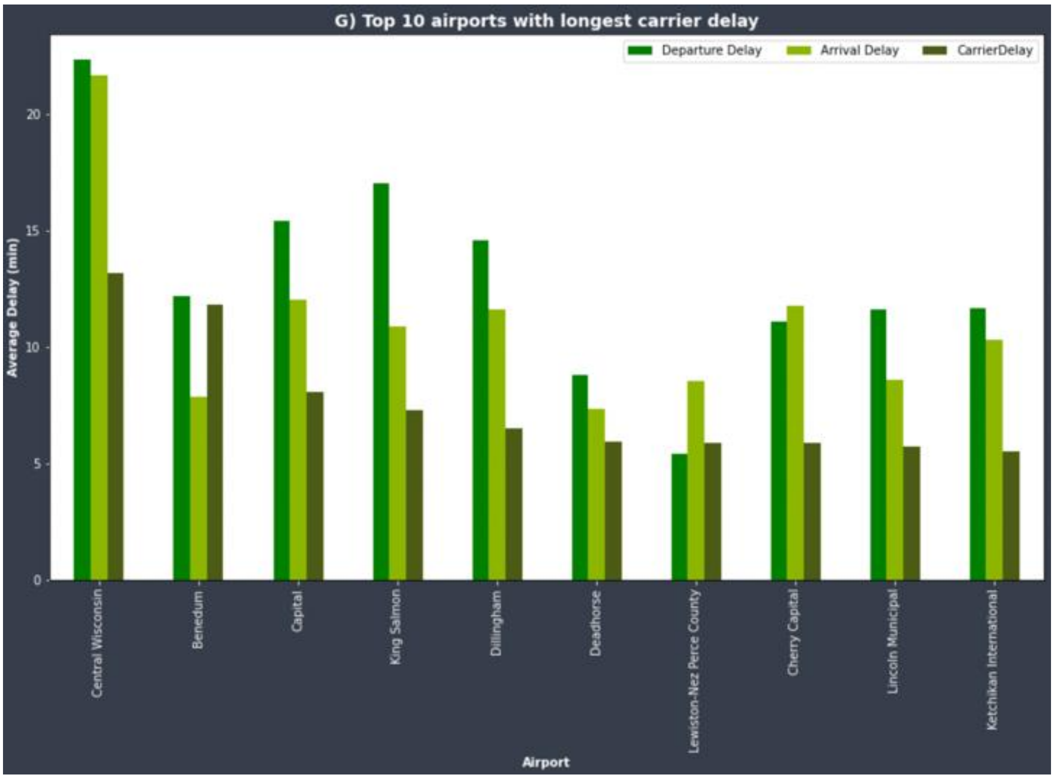
Date	Trip	DepTime	DepHour	LateAircraftDelay	TailNum	FlightNum
2004-04-12	ABQ-LAS	17:18	17	82.0	N336	2967
2004-04-12	LAS-RNO	18:07	18	78.0	N336	2967
2004-04-12	RNO-PDX	19:35	19	63.0	N336	2967
2004-04-12	PDX-GEG	21:08	21	58.0	N336	2967

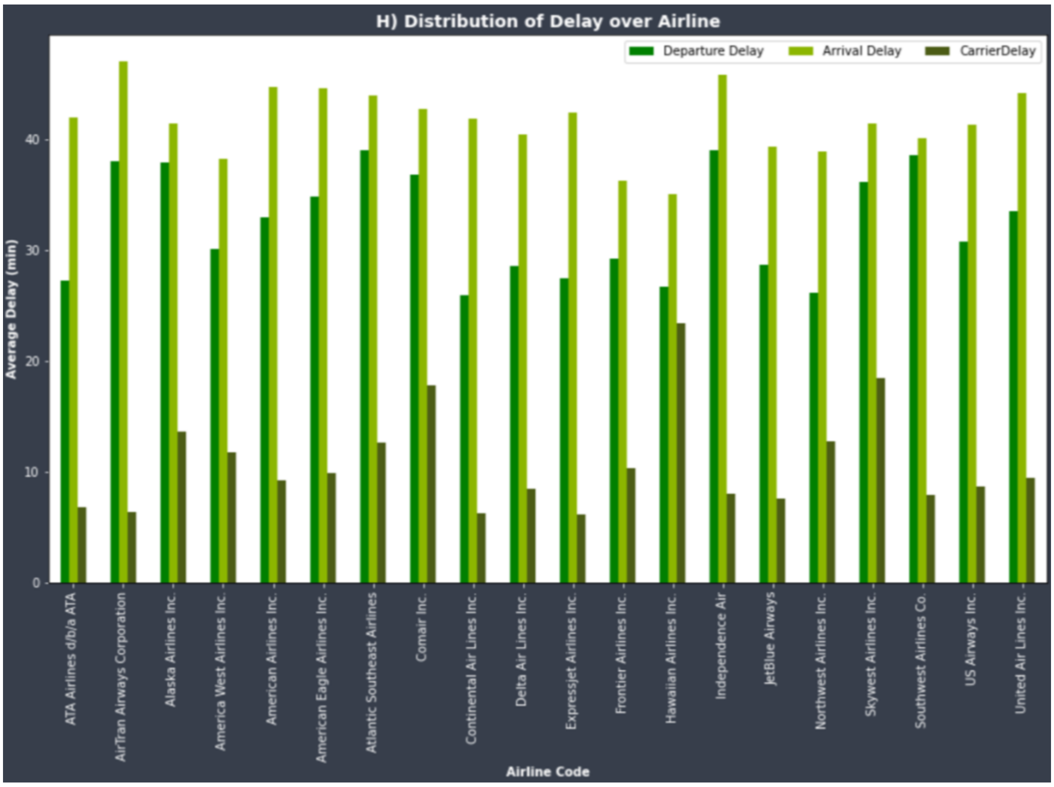
Table 1



Date	Trip	DepTime	DepHour	LateAircraftDelay	TailNum	FlightNum
2005-06-10	SJC-PDX	17:05	17	8.0	N337	936
2005-06-10	PDX-SJC	19:15	19	37.0	N337	1091
2005-06-10	SJC-LAX	21:20	21	42.0	N337	1091

Table 2





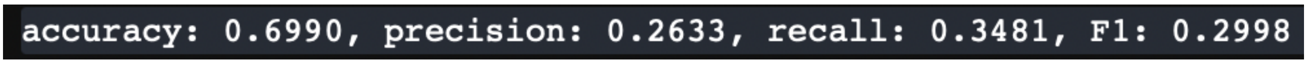
In addition, we checked if different airports and airlines have unusual characteristics on its average delays by plotting graph based on the average of departure delay, arrival delay, and carrier delay of each airline. From plot G and H, we observed some fluctuation of average delays between different airport and airlines. Hence, we concluded that delays could vary based on their origin of the airports and their airlines.

3.6 Question 5 (Use the available variables to construct a model that predicts delays.)

To predict a distinct outcome, there are numerous machine learning models to choose from. Thus, we tested three different models,1) Logistic Regression, 2) Random Forest Classifier, 3) Decision Tree Classifier, to pick the best predicting model. When building these models, we used the data available to find the relationship between features and label using different algorithms each models have. Moreover, to evaluate the prediction of the model, we conducted the following: 1st, confusion matrix, an evaluation metric for the model, 2nd, accuracy, the proportion of the total number of predictions that are correct, 3rd, precision, the total number of correctly classified positive examples, 4th, recall, the measure of positive examples labeled as positive by classifier and lastly, F1, the weighted average of Precision and Recall.

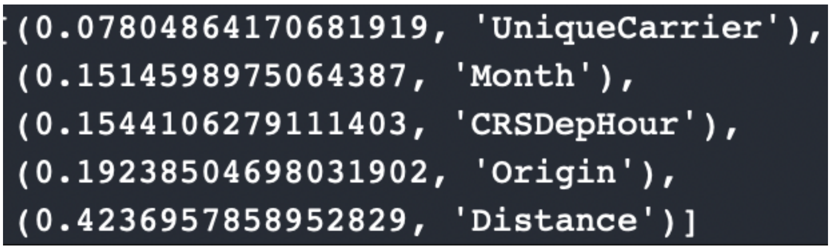
When constructing the model, since the dataset is huge, we used 10% of the preprocessed data which consists of 2035626 rows for python and 0.1% which consists of 20356 rows for R. We encoded categorical features that we will be using into numerical value to allow the model to perform, then selected its independent variable(label) and dependent variables(features) to start the prediction. For the label, since we aimed to predict delay of the flight, we selected "DelayedArr" column and for the features, we selected variables that we assumed to be influential to label and would be available in the future, unique carrier, origin, CRS departure hour, month, and distance. To start predicting, we first divided our train and test data set into half and tested the fit of the three models. Since Logistic Regression depends on a calculation based on 'weights' of the features, having higher number among the features could mislead its understanding. Moreover, decision Tree was also a good model for prediction, however, as it gets more complex, it tends to overfit. On the other hand, random forest makes a classification by aggregating the classifications of many decision trees, hence having overfitted trees in a random forest seemed less affecting compared to decision tree classifier. Hence, as we went through, random forest model seemed to be applicable as it is showed second highest ROC AUC score.

Before predicting the random forest model, as we looked at the proportion of the response dataset, we observed that our positive class(delayed) was much less compared to negative class (not delayed). This may cause class imbalance which may then result to bias. Since our interest was on the flights delay, we needed to revise the model because "not delayed" was dominated over "delayed". Hence, we balanced the class weight before conducting the model. Then, we began to evaluate the model and got the result of the following by train and testing the model.



```
accuracy: 0.6990, precision: 0.2633, recall: 0.3481, F1: 0.2998
```

Furthermore, we identified how influential each feature variable is to the model. From the result, we identified that distance of the flight had the most contribution and their identification code the least to the delay.



```
[(0.07804864170681919, 'UniqueCarrier'),
 (0.1514598975064387, 'Month'),
 (0.1544106279111403, 'CRSDepHour'),
 (0.19238504698031902, 'Origin'),
 (0.4236957858952829, 'Distance')]
```

Next, we demonstrated its ROC curve and demonstrated how hyperparameter tuning, setting of optimal hyperparameters to increase the accuracy and other metrics of the model, could be done using grid search or random search.

4 Future Direction / Recommendations

4.1 Future Research

Some of the ways we could improve our model is by Re-doing the EDA with more than 10 years of data available instead of just using 3 consecutive years data and besides using random forest for this report, we could try using other advanced bagging and boosting algorithms to test the model. To improve the scoring of the model, besides using the given features, we could also try creating new features by grouping existing features to further identify new insights. Furthermore, as we have identified that distance, origin, month, and CRSDepHour have an influence on the delay, we could add related data to enhance the information of the data.

5 References and Appendix

Patil, P. (2021, December 18). *What is exploratory data analysis?* Medium. Retrieved March 21, 2022, from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Ruiz, J. (2021, September 29). *3 things to know when you are facing a prolonged flight delay*. The Points Guy. Retrieved March 21, 2022, from https://thepointsguy.com/news/3-things-to-know-when-you-face-a-lengthy-flight-delay/

How to dealing with imbalanced classes in machine learning. Analytics Vidhya. (2021, January 6). Retrieved March 21, 2022, from https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/

Practical guide to deal with imbalanced classification ... (n.d.). Retrieved March 28, 2022, from https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Finnstats. (2021, May 6). *Class imbalance-handling imbalanced data in R: R-bloggers*. R. Retrieved March 29, 2022, from https://www.r-bloggers.com/2021/05/class-imbalance-handling-imbalanced-data-in-r/