CSE519: Project Progress Report
Understanding Flight Delays

Objective:

1. Predicting total delay in a particular flight's departure from a particular airport.
2. Finding interesting insights and answering/visualizing interesting questions from the flight data collected.

Dataset:

For our analysis we have selected the **top 10 busiest airports**(*Hartsfield–Jackson Atlanta International Airport (ATL), Los Angeles International Airport (LAX), O'Hare International Airport (ORD), Dallas/Fort Worth International Airport (DFW), Denver International Airport (DEN), John F. Kennedy International Airport (JFK), San Francisco International Airport (SFO), McCarran International Airport (LAS), Seattle-Tacoma International Airport (SEA), Charlotte Douglas International Airport (CLT)*) and **top 5 ranked airlines**(*Delta Air Lines, Southwest Airlines, United Airlines, Alaska Airlines, American Airlines*) for the years **2017** and **2018**.

The major portion of our **flight delay data** is gathered from the **Bureau of Transportation Statistics**'s airline **departures section**. Also, to collect the **weather data** we are using **Weather Underground**'s services. We are collecting **hourly weather data** for a given day. We have started performing web scraping using **Selenium** library. We have collected a fair amount of data and are aiming to collect a complete historical hourly weather dataset of the selected 10 airport locations for the year of **2017** and **2018**. We will further map the collected weather data with the current dataset that we are working on, based on **Date** and **Scheduled Departure Time**.

Preliminary Data Preprocessing:

1. Flight delay data for each airport and each airline is available separately. Therefore, after collecting the individual data we merged all of it into a single csv file to start our analysis.
2. The **Date** column of the merged dataset was further converted to ***panda***'s *datetime* format.
3. In order to facilitate our analysis, we added new columns to our dataset.
   Added columns: **Month**(from Date), **Scheduled Departure Hour(**from Scheduled Departure Time), **Actual Departure Hour**(from Actual Departure Time), **Total Delay** (by the cumulative addition of all the five delay types(***Carrier Delay, Weather Delay, National Aviation System Delay, Security Delay, Late Aircraft Arrival Delay***)).
   Criteria for creating **Scheduled Departure Hour** and **Actual Departure Hour:** We have divided the whole day into 24 hours, starting from 00 hours to 23 hours. Any flight whose departure time is in the range from say A:00 - A:59, we have considered it's corresponding departure hour to A hours(which is from 0-23).

4. Going ahead with the idea of departure hour, we came up with a novel idea of finding **rush hour** for each airport which is calculated by the frequency of departures from an airport for each of the 24 hours. The hour with most departures is the **rush hour.**

Exploratory Data Analysis:

1) **Total Delay**

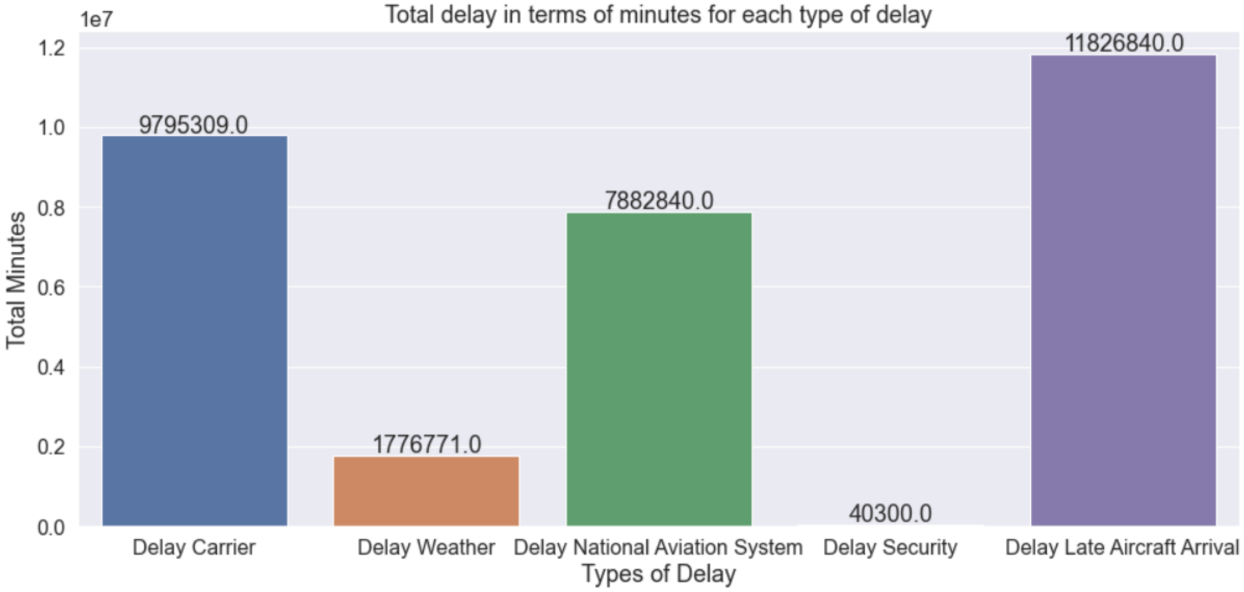


Fig. 1: Total Delays for each delay type.

In 2017 and 2018, the major reason for the total delay was due to **Late Aircraft Arrival** which means there was a delay in the arrival of the aircraft whose current flight got delayed. In Fig. 1, the second reason for the most delay was due to the **Carrier**. It is interesting to note that the aircraft which is arriving late can be late due to one of the other remaining reasons, because from the source where the delay has begun, the reason for its delay should not be due to Late Arrival.

2) **Busy Airports and their corresponding Delays**

In Fig. 2(a), the airports are arranged in decreasing order of the total flights departing and the vertical bars show the average minutes of delay for each airport. Based on this, and Fig. 2(b), it can be commented that during 2017 and 2018 even though **ATL Airport** has the **highest percentage** of departing flights. Rest all the other airports have a similar percentage of departing flights with **Seattle** having the **least percentage** of departing flights (3.5%). Interesting fact to note here is that **Atlanta International Airport** has significantly **lower mean delay**(*9.2 mins*). Consequently, **JFK Airport** has the **least number of departing flights** after Seattle; it is likely to have **delayed departing flights more(*14.2 mins*)** in comparison to ATL Airport. Similarly, every other airport on the chart tends to have less frequent flights departing compared to ATL Airport and almost all of those flights have actually accounted for more mean delay than ATL

Airport. Therefore, it is actually **not very obvious that with an increase in the frequency of departing flights, the average delay would also increase**.

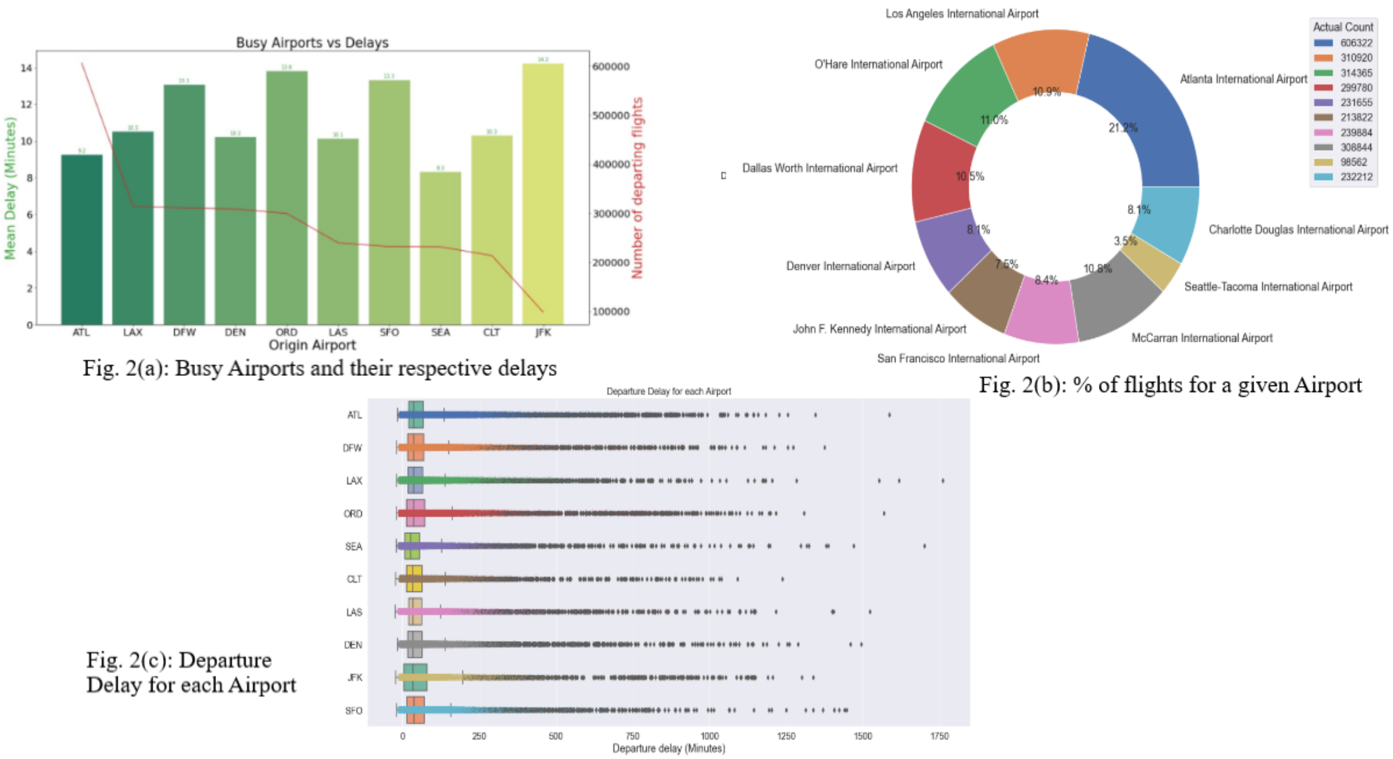


Fig. 2(a): Busy Airports and their respective delays

Fig. 2(b): % of flights for a given Airport

Fig. 2(c): Departure Delay for each Airport

Fig. 2: Busy Airports and their respective delays(minutes).

In Fig. 2(c), using the values where the flight departed before scheduled departure time, it will have a negative value, also, considering the 0 values will make the graph skewed a bit towards left. Thus, considering only the positive values(where delay has occurred), we can see that a passenger can face a **median delay of approximately 36 minutes** for a departing flight. Although the majority of delays are in the time range of 0 to 250 minutes, there are several instances where delay occurred for more than 500 minutes.

3) <u>Delays experienced on holidays</u>

Fig. 3, shows the analysis of how the delays from the selected airlines would show up on holidays for the year 2018. On narrowing down to the selected official holidays(***New Year's Day, Valentine's Day, Presidents' Day, Good Friday, Easter Sunday, Mother's Day, Memorial Day, Independence Day, Labor Day, Columbus Day, Halloween, Veterans Day, Thanksgiving Day, Christmas Eve, Christmas Day***) in the USA, it can be noted that almost all airlines showed **similar** trends till the beginning of **July**(i.e. during *New Year's Day, Valentine's Day, Presidents' Day, Good Friday, Easter Sunday, Mother's Day, Memorial Day).*

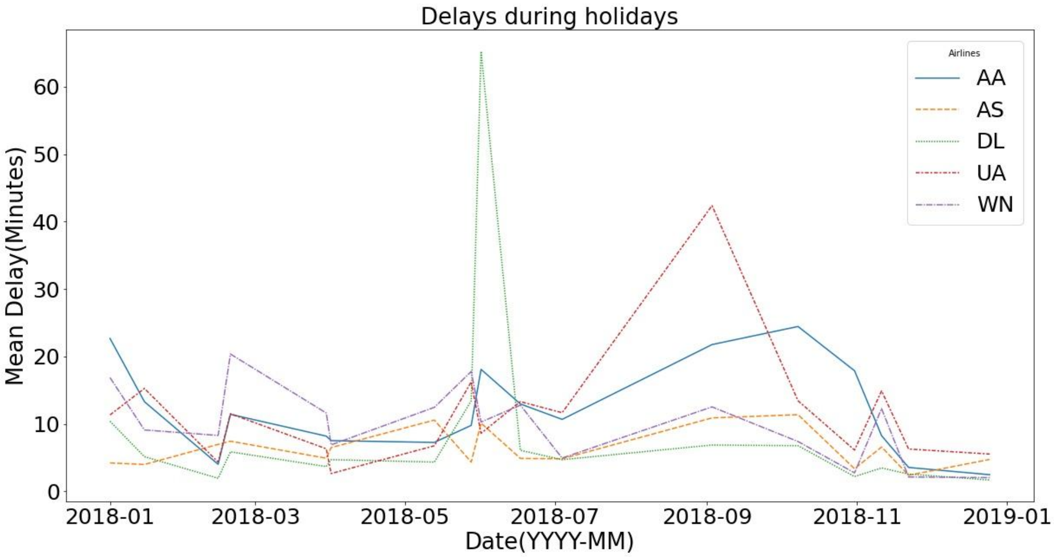
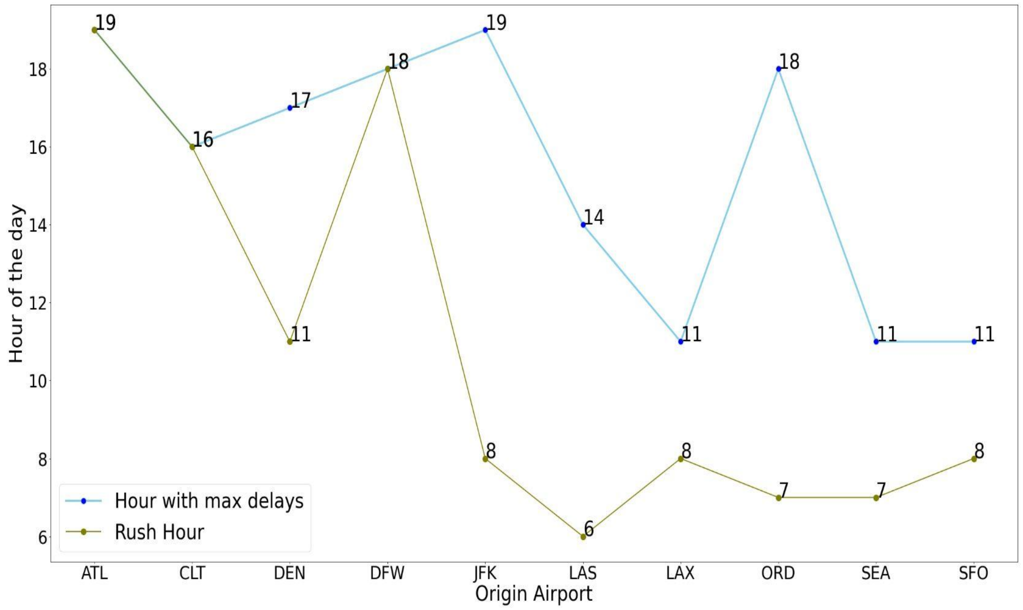


Fig. 3: Delays during holidays

However, during the month of July around the ***Independence Day week***, **Delta Airlines** gave a significant **hike** in delays by increasing the ***mean*** to slightly above *60 mins*, while all the other airlines continued with their similar trends. During ***Labor Day***, mean delays for *United Airlines* and *American Airlines* fairly inclined to roughly *40 mins* and *25 mins* respectively. It is intriguing to note that after the great hike in delays by **Delta Airlines**, the airlines **managed** to keep the **delays** less than *10 mins* for the **rest of the year** with having the **least** delay among all the other flights by slightly less than *5 min* during ***Thanksgiving*** and ***Christmas.***

4) Rush Hour, Hour of the day with maximum delays:

Going forward with our idea of converting the day into 24 periods of an hour each from 0-23, we computed the **Rush Hour and Hour of the day which have maximum delays** for each of the 10 airports we have considered for our analysis. The below graph succinctly shows that for only 2 airports namely ***Hartsfield–Jackson Atlanta International Airport (ATL) and Charlotte Douglas International Airport (CLT)*)** out of the 10 airports in our analysis, the rush hour matches with the hour of the day in which the maximum number of flight delays have occurred. This provides an insight that frequency of flights is not the only major factor which affects flight delays heavily and there are other factors which can cause more delays.



Delay Prediction Modelling:

One of the major aims of our project is to predict the future delay(total, specific delay due to weather conditions) based on the flight data features available to us.

Baseline Model 1

We started by creating a baseline model (**Linear Regression model**) for **predicting total departure delay** in a particular flight with only the numerical features selected(number of features = 5) to set up a standard for comparison for our future models. We split our data into training (80%) and testing (20%) data to test our model. We are using **Mean Squared Error (MSE)** as our evaluation metric. For the Baseline Model 1 we achieved a **MSE** of **97.19** on the testing data.

Baseline Model 2

We started working on our second baseline model with the aim of achieving a better MSE from our first baseline model which has no data preprocessing and only numerical features.

Preprocessing for the Baseline Model 2:

1. Dropped a few categorical columns which couldn’t be efficiently encoded.
2. Included a **total delay** column by adding the individual delays for each row, dropped the individual delay columns.

3. Added the month column by extracting the month from the date column already in our dataset.
4. **Dropped** the date column.
5. Changed the categorical columns to ‘category’ dtype.
6. Converting Scheduled and actual departure time into Scheduled and actual departure **hours columns (0-23)** and dropping scheduled and actual departure time columns.
7. Split the data into training and test data, 80% into training and 20% into test.
8. **One hot encoded** the categorical columns.
9. **Scaled** the dataset features.

After the above preprocessing, we again trained a Linear regression model with our training data and predicted the total delay for our test data. For this model we achieved a **MSE of 92.19**. This MSE is **better** than our first baseline model, which shows our preprocessing has helped improve the performance of the model Also, adding the additional features and including the **categorical features** also helped.

Advanced Predictive Model

Further, we used the same preprocessed data, to train a **Gradient Boosting Regressor** model with our training data and predicted the total delay for our test data. We kept the **hyper parameters** for our model as **default ones.** Gradient Boosting is fairly robust to overfitting so a large number of estimators can result in better performance (Increasing the number of estimators will be a part of our next steps). For this model we achieved a **MSE of 40.38** on testing data.

Next Steps:

- In our next steps, we are aiming to build a robust model that is able to predict the future flight delays. We will achieve this by adding new features and preprocessing the data and apply these changes on different predictive models. This will help us to reduce the errors more i.e. decrease the MSE values.
- Collection of the complete weather data and merging it with our current dataset.
- Exploratory Data Analysis on the merged dataset (current dataset + weather dataset).
- Further to make our model more intuitive, we aim to predict delays by considering the weather conditions separately, in addition to the total delay in the flight in our future work.
- We have also planned to explore the Open-Sky Database and if we find any valuable data through it then we’ll try to merge it with our current data set.