**Airline Traffic Data Analysis Report**

1. **Introduction**

This report outlines the findings of 2014 U.S airline traffic data analysis, and perform the supervised machine learning algorithms on the prediction of airline delays in 2015. The remainder of the report is structured as follows: Section II provides a general exploration of 2014 U.S Airline Traffic Data as well as Airline Traffic Data of JFK airport. Section III present the process of how we build a model to predict airline delays. Future work are shown in the Section VI.

1. **Data Analysis**
2. **Data Description**

The dataset is available on the United States Department of Transportation [website](http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time). I also downloaded the [Airport Codes mapped to Latitude/Longitude in the United States](https://opendata.socrata.com/dataset/Airport-Codes-mapped-to-Latitude-Longitude-in-the-/rxrh-4cxm) dataset. I first read the raw csv data file into the DataFrame via python, drop the empty entries, and I found that there are over 5.3 million flights in 2014. The dataset I downloaded is over 1.4 GB. The following table describes the details of the important portion of my downloaded dataset

|  |  |
| --- | --- |
| **Name** | **Description** |
| YEAR | 2014 , 2015 |
| MONTH | 1-12 |
| DAY\_OF\_MONTH | 1-31 |
| DAY\_OF\_WEEK | 1 (Monday) - 7 (Sunday) |
| FL\_DATE | Flight date |
| UNIQUE\_CARRIER | Unique carrier code |
| AIRLINE\_ID | Airline identification number assigned by US DOT |
| CARRIER | Code assigned by IATA |
| TAIL\_NUM | Tail number |
| FL\_NUM | Flight number |
| ORIGIN | Origin airport |
| DEST | Destination airport |
| CRS\_DEP\_TIME | Scheduled departure time (local: hhmm) |
| DEP\_TIME | Actual Departure Time (local: hhmm) |
| DEP\_DELAY | Departure delayed minutes |
| DEP\_DEL15 | Departure Delay Indicator, 15 min or more (1=Yes) |
| CRS\_ARR\_TIME | Scheduled arrival time(local: hhmm) |
| ARR\_TIME | Actual Arrival Time |
| ARR\_DELAY | Arrival delayed minutes |
| CANCELLED | Cancelled Flight Indicator (1=Yes) |
| CANCELLATION\_CODE | Specifies the reason for cancellation(A=carrier,B=weather,C=NAS, D=security) |
| CRS\_ELAPSED\_TIME | Scheduled elapsed time of flight |
| ACTUAL\_ELAPSED\_TIME | Actual elapsed time of flight |
| AIR\_TIME | Flight time in minutes |
| DISTANCE | Distance between airport |
| CARRIER\_DELAY | in minutes |
| WEATHER\_DELAY | in minutes |
| NAS\_DELAY | in minutes |
| SECURITY\_DELAY | in minutes |
| LATE\_AIRCRAFT\_DELAY | in minutes |

Table 1. Dataset Description

1. **Exploration of 2014 U.S Airline Traffic Data**

In this part, I’ll present my exploration results in a question-solution form.

***Question 1: Find the top 10 busiest US airport in 2014.***

To find the top busiest airport in the US. We just need to look the annual number of departure and arrival flights for each airport in 2014.

|  | **Total Arrival Flights** | **Total Departure Flights** | **Latitude** | **Longitude** |
| --- | --- | --- | --- | --- |
| **ATL** | 369779 | 369842 | 33.6367 | 84.4281 |
| **ORD** | 287036 | 287377 | 41.9808 | 87.9067 |
| **DFW** | 278309 | 278310 | 32.8969 | 97.0381 |
| **DEN** | 224205 | 224067 | 39.8617 | 104.6731 |
| **LAX** | 221932 | 221941 | 33.9425 | 118.4072 |
| **IAH** | 174210 | 174235 | 29.9844 | 95.3414 |
| **SFO** | 166893 | 166883 | 37.6189 | 122.3750 |
| **PHX** | 160475 | 160466 | 33.4342 | 112.0117 |
| **LAS** | 137055 | 137058 | 36.0800 | 115.1522 |
| **CLT** | 112243 | 112256 | 35.2139 | 80.9431 |

Table 2. Top 10 busiest US airport in 2014

***Question 2: Find the delay situation by airport.***

Here, we’ll look at the delayed number of departure flights and its percentage for each airport.

|  | **Airport** | **Total Departure Flights** | **DELAY\_DEP\_FLT** | **Percentage** |
| --- | --- | --- | --- | --- |
| **1** | GST | 77 | 27 | 35.0649 |
| **2** | OTH | 376 | 127 | 33.7766 |
| **3** | CEC | 963 | 303 | 31.4642 |
| **4** | MDW | 88498 | 27836 | 31.4538 |
| **5** | HOU | 59503 | 17975 | 30.2086 |
| **6** | ADK | 104 | 31 | 29.8077 |
| **7** | DAL | 48294 | 14038 | 29.0678 |
| **8** | ORD | 287377 | 80608 | 28.0496 |
| **9** | GUM | 366 | 101 | 27.5956 |
| **10** | BWI | 90845 | 24822 | 27.3235 |

Table 3. Top 10 US airport suffering severe delaying situations

Now, we visualize the above two questions in the following geographical map, In the following figure, the color bar represent the percentage of delayed flights, and the size of the circles represents the capacity of the airport.

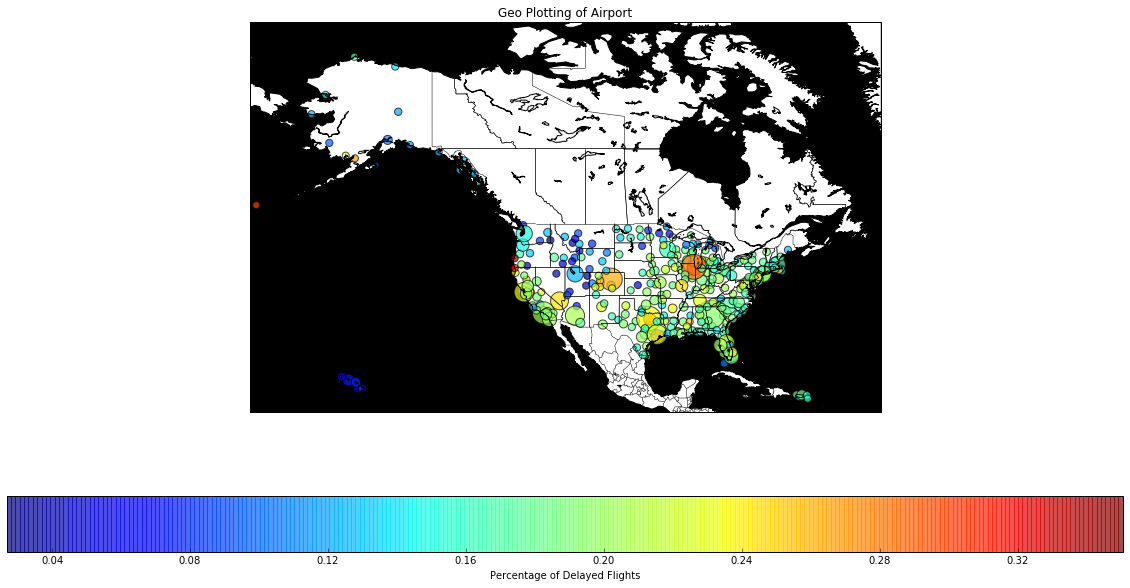


Fig 1. Capacity and severity of delaying situations of US airport 2014

***Question 3: Find the top 10 most connected pairs of airport.***

To solve this problem, we can directly count the number of pairs of airports appearing in the constructed data frame. But there is another way to do this. The solution is to build an airport network. This solution might complicate this problem but it allows people to explore other network searching problems.

We return the top10 pairs of airport with most connections.

The number of air routes is: 2412

[('SFO', 'LAX'), ('JFK', 'LAX'), ('LAX', 'LAS'), ('HNL', 'OGG'), ('ORD', 'LGA'), ('JFK', 'SFO'), ('PHX', 'LAX'), ('ATL', 'LGA'), ('ATL', 'MCO'), ('SAN', 'LAX')]

Let’s also try to visualize the airport network on a geographical map. Note that since there are over 2,000 air routes, we randomly sampling 500 routes for the easy visualization.

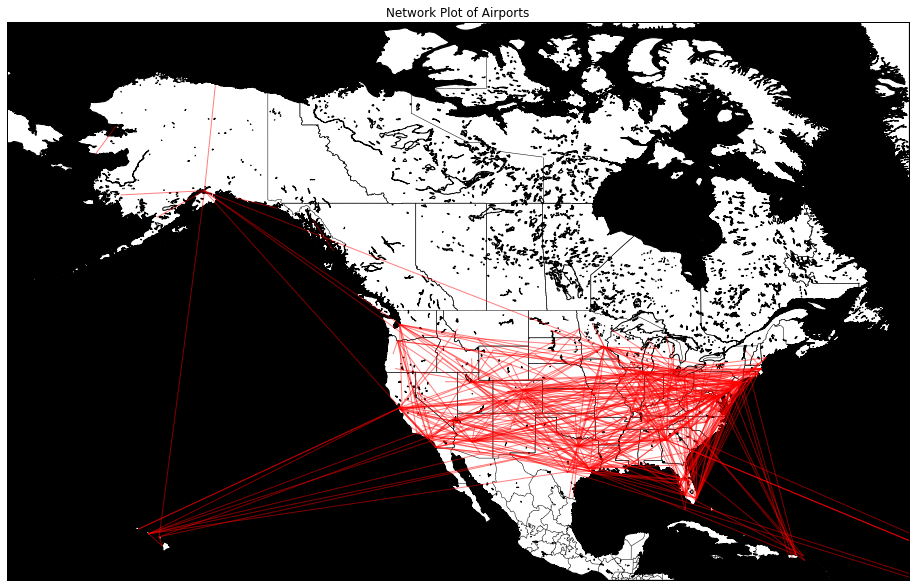


Fig2. Airport network plot 2014

***Question 4: Find the delay situation by carrier.***

Despite we can analyze the delay situation by airport, we can also analyze by carrier.

|  | **CARRIER** | **DEL\_COUNT** | **FL\_COUNT** | **PERCENTAGE** |
| --- | --- | --- | --- | --- |
| **0** | AA | 109406 | 529486 | 20.662680 |
| **1** | AS | 16607 | 159568 | 10.407475 |
| **2** | B6 | 49843 | 243735 | 20.449669 |
| **3** | DL | 115991 | 794033 | 14.607831 |
| **4** | EV | 155334 | 652724 | 23.797807 |
| **5** | F9 | 18320 | 85135 | 21.518764 |
| **6** | FL | 14299 | 78014 | 18.328762 |
| **7** | HA | 3782 | 74585 | 5.070725 |
| **8** | MQ | 83574 | 372401 | 22.441938 |
| **9** | OO | 111354 | 595643 | 18.694755 |
| **10** | UA | 115972 | 486527 | 23.836704 |
| **11** | US | 56928 | 408373 | 13.940197 |
| **12** | VX | 9484 | 57184 | 16.585059 |
| **13** | WN | 334478 | 1159661 | 28.842739 |

Table 4. The percentage of delayed departure flights by carrier

We also made a few plots to visualize the results.

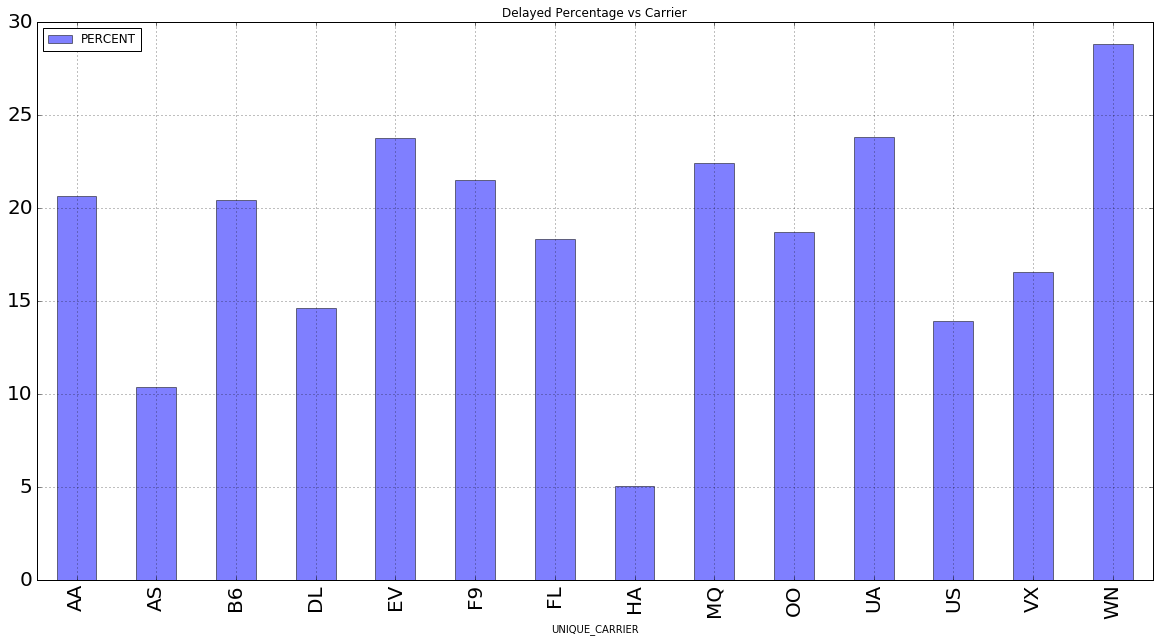


Fig 3. Departure delaying percentage vs Carrier

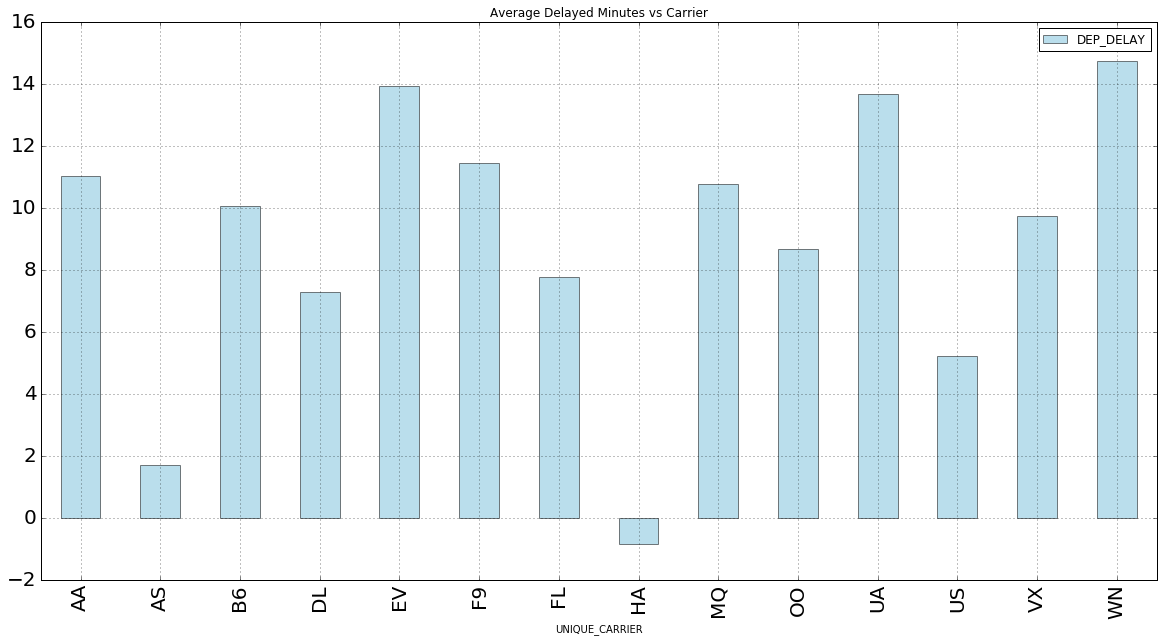


Fig 4. Average Departure delaying minutes vs Carrier

Note that the negative delaying minutes represent the early departures. And we can easily observe that the Southwest Airlines(WN) performs worst in terms of both the percentage of delayed departure flights and the average delaying minutes. More detailed information regarding the delaying situations are available on

1. **Exploration of 2014 JFK Airport Data**

Now let’s look at the specific airline traffic data for JFK airport. Let’s see how delayed flights are distributed by month:

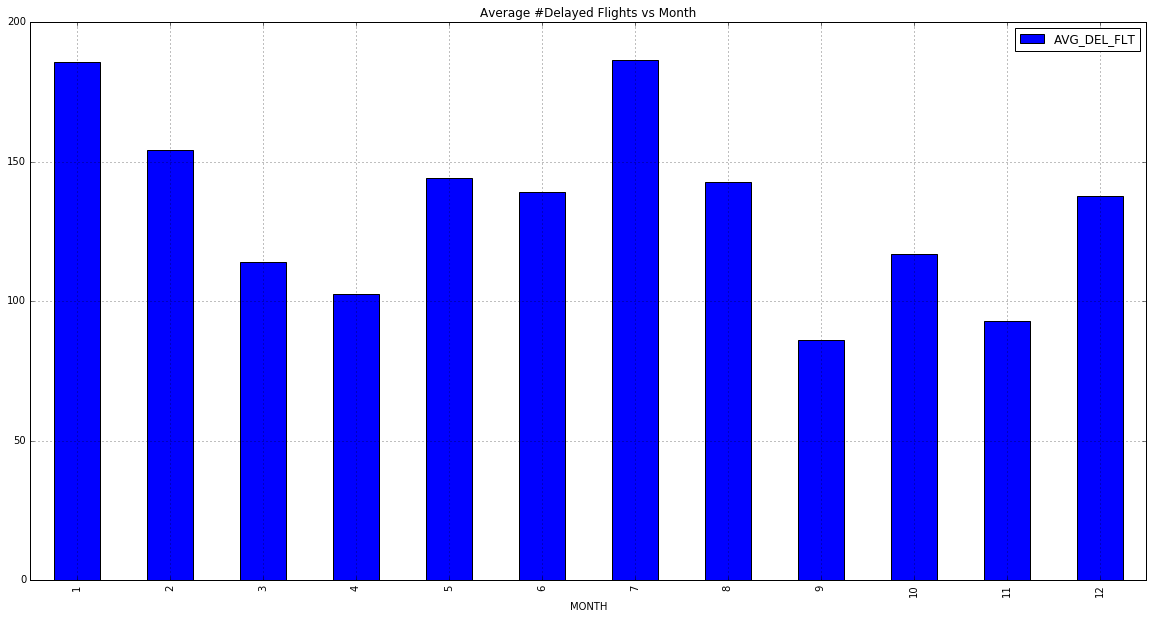


Fig 5. Average number of departure delayed flights vs Month

From the above figure, we find that there are highest number of delayed flights in Jan, Feb, and Jul. This might related to the severe weather conditions in Jan and Feb, and Jul is the tourist season.

Now let's look at the hour-of-day:

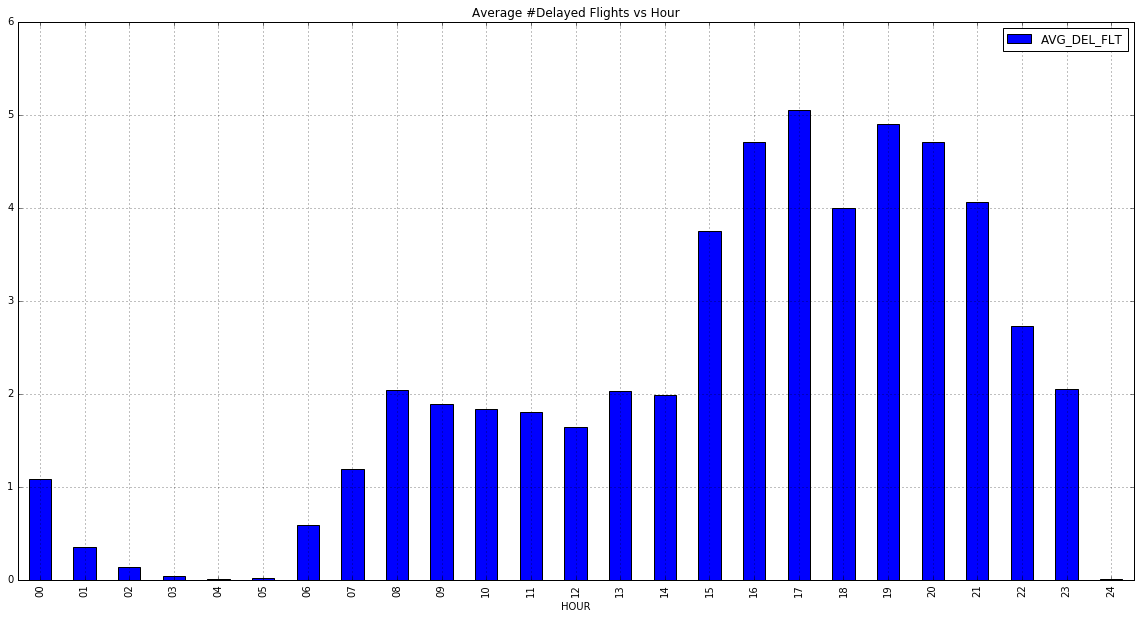


Fig 6. Average number of departure delayed flights vs Hour of the day

This is exactly what we expected, the number of delayed departure flights tend to pile up during the daylight, and start to decline when the night falls.

1. **Airline Delay Prediction**

In this Section, we’ll focus on how to build a supervised machine learning model to predict airline delays in JFK airport. We’ll use the 2014 airline traffic data to predict delays in 2015. Note that the latest data is up to Oct, 2015. Random Forest Classifier is implemented to perform the prediction. The performance is evaluated by the confusion table and prediction accuracy.

The target variable is a binary value (1 for delays of more than 15 minutes, 0 otherwise). Therefore, there are a lot more 0s than 1s in the target vector Y. I constructed the feature vector ***x*** by using the following predictors

1. Month (winter month tend to have more delays)

2. Day of the month (not so obvious, but let’s keep it)

3. Day of the week (weekdays vs weekends)

4. Carrier (delay situations varies from carriers to carries as shown above.)

5. Destination airport

6. Distance (not so obvious, but it’s interesting to see if it is a good predictor.)

7. The number of days from the nearest holiday (this is based on the assumption that the more close to the holiday, more delays will be.)

Model 1.

In the first model, I just used variable 1, 2, 3, 6 and 7 to construct our feature vector for each flight. Then random forest classifier with 25 trees is applied. The confusion table is shown as below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted Label | |
| 0 | 1 |
| True  Label | 0 | 127106 | 6756 |
| 1 | 32404 | 2360 |

Table 5. Confusion table 1.

According to the confusion matrix, we can compute the following scores: Precision = 0.26, Recall = 0.07, F1 = 0.11, Accuracy = 0.77.

Model 2.

In the second model, I used the OneHotEncoding function in Scikit-learn to encode all the categorical features. And 7 features are all used in this model. Then applied the random forest classifier with the same parameters.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted Label | |
| 0 | 1 |
| True  Label | 0 | 131354 | 2508 |
| 1 | 33654 | 1110 |

Table 5. Confusion table 2

Precision = 0.31, Recall = 0.03, F1 = 0.06, Accuracy = 0.79

From the comparison between the two models, we can see that the accuracy is improved by using the Model 2.

1. **Possible Future Work**

We can continue to enrich the model by create more features like weather conditions (wind velocity, visibility, humidity, etc), security conditions of airport and so forth.