IDS 494
PYTHON FOR DATA SCIENCE
FINAL PROJECT REPORT

DEVYANI SINGH-654766729
TARUN KHURANA-667683595
RUITING CHEN-669199419

EXECUTIVE SUMMARY

This report summarizes the statistical, descriptive analysis and implementation of models for the purpose of predicting flight delays by understanding python.
For this project we acquired the dataset from US Transport website which comprised of three excel files representing airports, airlines and flights information.
Our aim was to not only to study the datasets in detail by analysing flight delays but also to predict them to the best of our ability ,hence, we started with cleaning the dataset and merging all three excel files into a single dataset. After cleaning and merging, we had close to 5.7 million rows and 31 columns and handling this amount data really proved to be challenging.
After our data was prepared for use, we started out analysis as we wanted to know how each numerical and categorical variable looked like. We did a thorough analysis on each variable and explained those using statistical figures and plots. Our main numerical variable was "Delay" which was further subdivided into Departure delay, Arrival delay, Weather delay, Late aircraft delay and Airline delay. We analysed which delay was the maximum, minimum and also plotted the total average delay. For our categorical variables, Airline and Airport were the most important variables. We analysed which airline and which airport attributed to the most delay.
After doing an extensive statistical and descriptive research on our variables, we proceeded towards building models for predicting flight delays, starting with linear regression model. After linear regression we tried clustering, followed by Time series, Fixed effect and in the end we also tried implementing Random Forest model.
We faced many problems while implementing our models and would consider the amount of data volume (close to 1GB) as the main cause.

INTRODUCTION

About the datasets:

The datasets were acquired from US Transport website and contained 3 excels files namely, Airport dataset, Airline dataset and Flight delay dataset.

The Flight delays dataset is for the year 2015 for the airlines flying in United States. This is our main dataset which has 5819079 rows and 31 columns in raw form.

Here is the list of few of the important variables in the main dataset when in raw form-

- YEAR, MONTH, DAY, DAY_OF_WEEK: dates of the flight
- AIRLINE: an identification number assigned by us dot to identify a unique airline
- ORIGIN_AIRPORT,DESTINATION_AIRPORT: code attributed by iata to identify the airports
- SCHEDULED_DEPARTURE,SCHEDULED_ARRIVAL : scheduled times of take-off and landing
- DEPARTURE_TIME,ARRIVAL_TIME: real times at which take-off and landing took place
- DEPARTURE_DELAY,ARRIVAL_DELAY: difference (in minutes) between planned and real times
- DISTANCE: distance (in miles)

We would be considering these variables in our analysis and prediction.

The Airline Dataset represents a list of all the airlines operating in the United States whereas, the Airport Dataset represents a list of all the airports currently in the United States. The Airline dataset consists of IATA code and name of each airline. There are 14 airlines in that data. The Airport dataset consists of IATA code, name of each airport and city, state and country that airport is situated in. It also contains the latitude and longitude of each airport in the dataset.

The first roadblock we faced while trying to load our main dataset, a file with more than 58 million rows, in our 32 bit PYTHON IDE was that of a low memory timeout which not only resulted in extensive load on the CPU's processor, resulting in pausing the windows for quite a few number of times but also absorbed 5 days of our project (before even starting!!).After struggling for a few days and googling all over the internet, we came to a conclusion that a 32 bit PYTHON IDLE can maximum be allowed to have 2 GB RAM at once which was quite less than what our dataset wanted. We uninstalled the 32 bit PYTHON IDLE version only to install the 64 bit PYTHON IDLE version which resulted in the smooth execution of loading the dataset as the 64 bit PYTHON had no upper limit for the memory.

THE PROBLEM AND ITS IMPORTANCE

Now a days ,with quite a lot of airlines operating, it has become really necessary to have some kind of predictive analysis in hand so has to select the best one possible at an economical price. Hence, the problem we considered in our final study report was that how much a flight will probably be delayed and what may be the possible reasons for the same?

PREPARING THE DATASETS

CLEANING THE DATASETS

After the successful loading the flights dataset we started working on cleaning the dataset and the following steps were followed to accomplish the same:

- Step1: Combined columns like YEAR, MONTH, and DAY etc. to create a DATE column.
- Step2: Formatted various columns to time format with the help of a function which changed the formats of many delay type columns like SCHEDULED_DEPARTURE and DEPARTURE_TIME to have an exact time of a day in 24 hr. format.
- Step3: Removing unwanted columns like YEAR, MONTH, and DAY etc.
- Step4: Deleting/Replacing blank value rows, that is, handling NA's and Cancelled/Diverted flights as they won't affect delays.
- Step5: The last was rearranging the columns in order to better understand the dataset

MERGING THE DATASETS

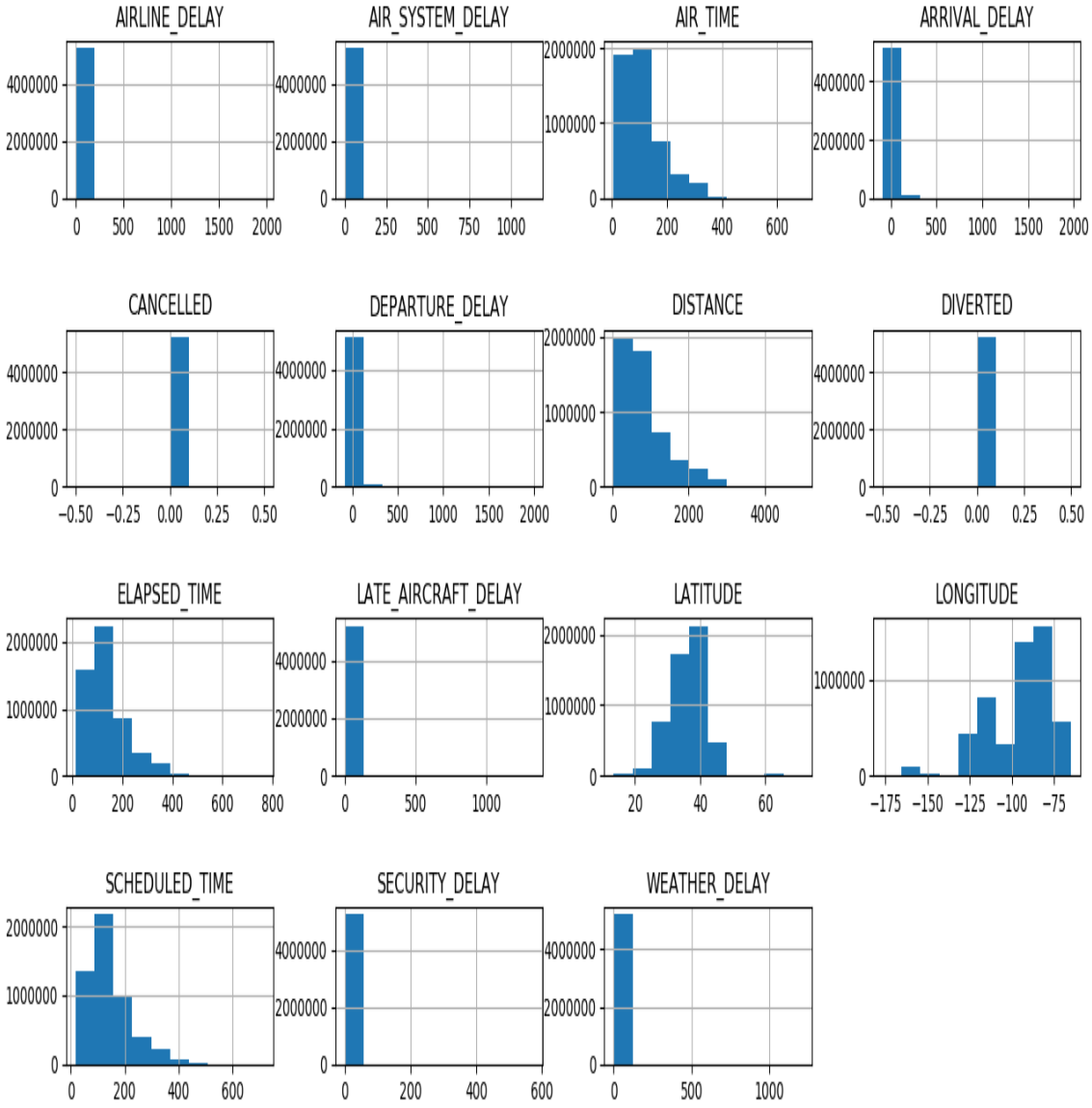
The main dataset had just the code instead of the airline name which made us merge the two datasets, that is, airlines and flights for the easy understanding of the datasets.

After performing the preparing the data process, the datasets were ready to use for further analysis.

DESCRIPTIVE ANALYSIS

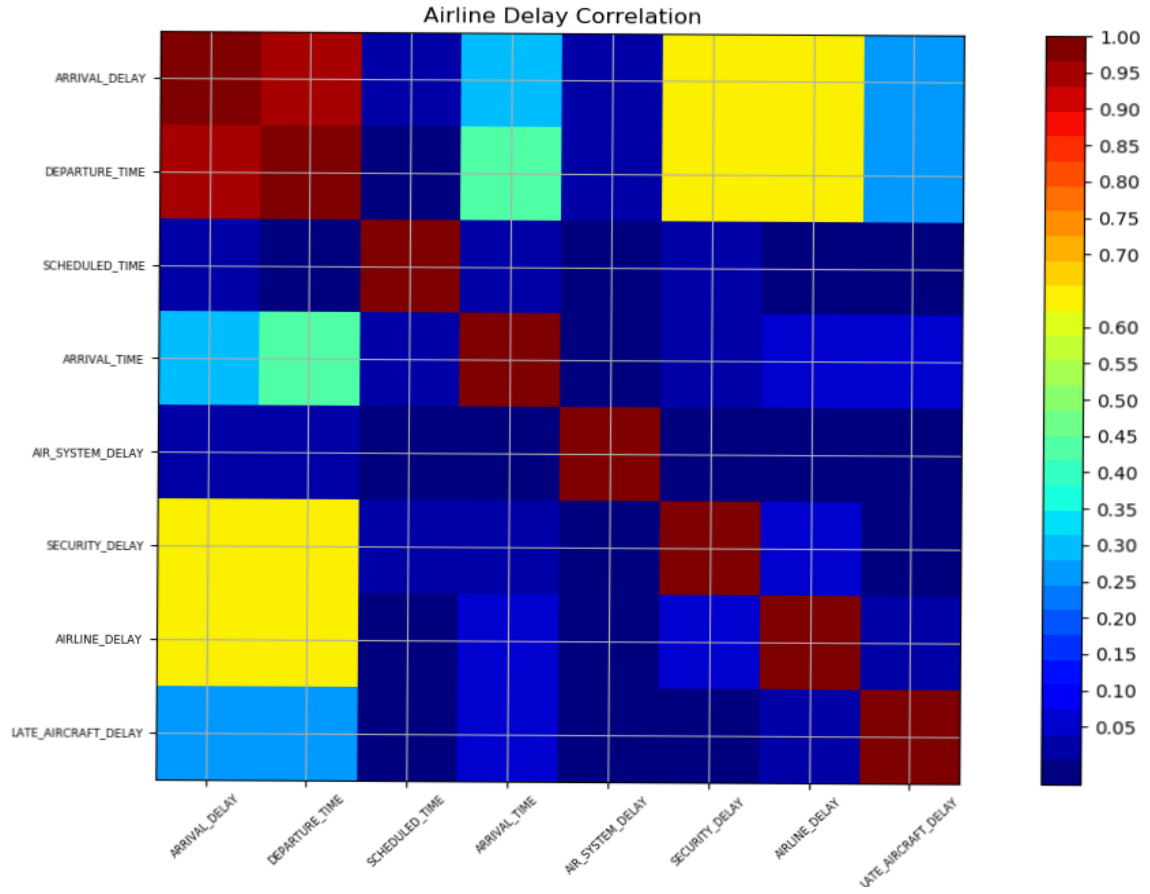
NUMERICAL DATA ANALYSIS

In numerical data analysis, we analysed each variable individually, hence, plotted the distribution of each individual variable below. From this figure, we can see that Time, latitude and longitude is more distributed than Delay as the Time variables in minutes show the distribution in a day whereas, Delay variables also in minutes show a smaller distribution because the delay may range from a few minutes to a few hours but very rarely will the distribution be of a day or more. Same case applies to the variables Longitude and Latitude, as they represent the location of airports in whole of USA they have a wider distribution



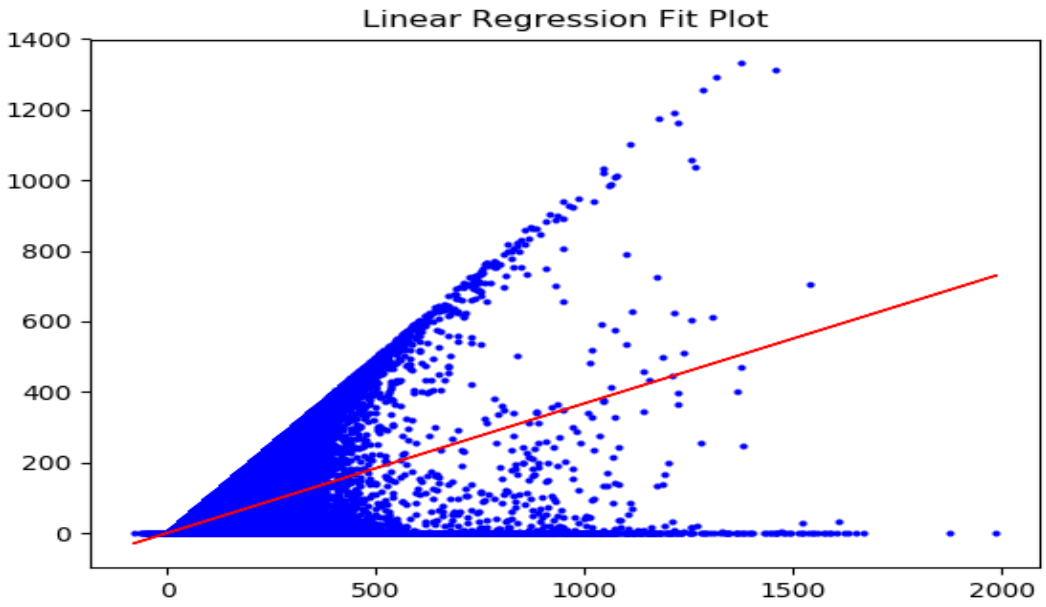
After individually visually analysing each variable, we plotted the correlation matrix to analyse the relationship of all numerical data with each other. From Correlation Matrix, we see that a strong correlation exists between departure delay, arrival delay and late aircraft delay. We also see that a strong correlation exists between departure delay, arrival delay and airline delay. Some direct correlation like departure delay to arrival delay also exists.

From the analysis of correlation matrix, we identified two relationships of departure delay with arrival delay and late aircraft delay.



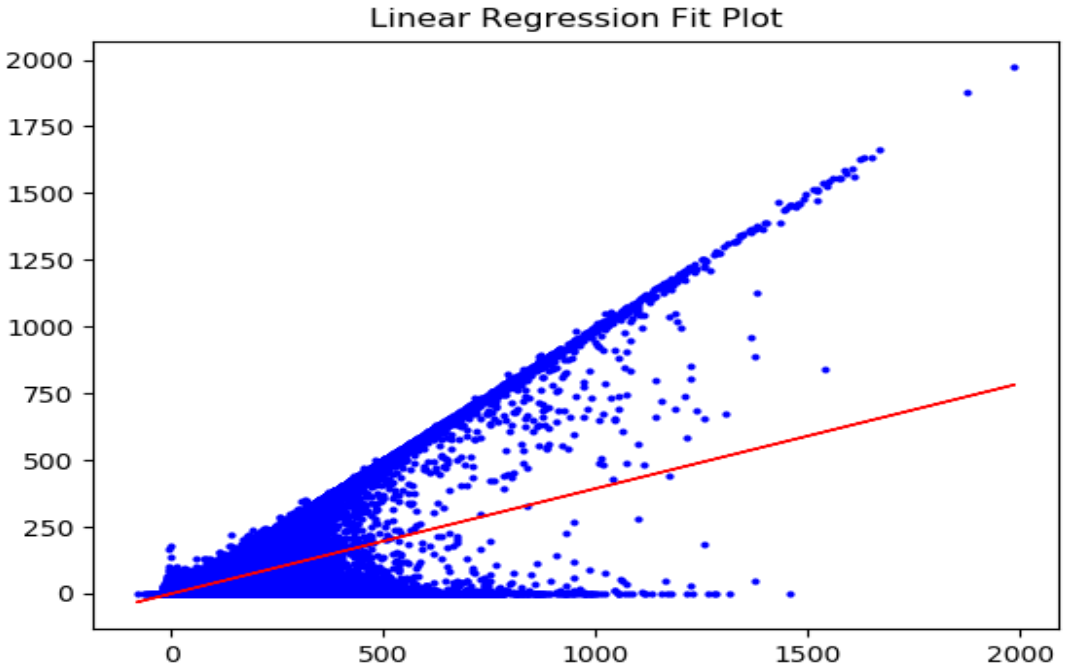
	DEPARTURE_DELAY	ARRIVAL_DELAY	DEPARTURE_TIME	SCHEDULED_TIME	ARRIVAL_TIME	AIR_SYSTEM_DELAY	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
DEPARTURE_DELAY	1.000000	0.944672	0.171840	0.027582	0.049652	0.304895	0.028290	0.658352	0.652094	0.265420
ARRIVAL_DELAY	0.944672	1.000000	0.159787	-0.030029	0.049876	0.424414	0.028074	0.627489	0.622980	0.270098
DEPARTURE_TIME	0.171840	0.159787	1.000000	-0.020817	0.650051	0.055973	0.003888	0.051761	0.150353	0.021641
SCHEDULED_TIME	0.027582	-0.030029	-0.020817	1.000000	0.020561	0.014741	0.003270	0.015696	-0.014251	0.000173
ARRIVAL_TIME	0.049652	0.049876	0.650051	0.020561	1.000000	0.038267	-0.000826	0.001293	0.034496	-0.001344
AIR_SYSTEM_DELAY	0.304895	0.424414	0.055973	0.014741	0.038267	1.000000	0.002863	0.029539	0.055270	0.050190
SECURITY_DELAY	0.028290	0.028074	0.003888	0.003270	-0.000826	0.002863	1.000000	-0.001400	0.003678	-0.000709
AIRLINE_DELAY	0.658352	0.627489	0.051761	0.015696	0.001293	0.029539	-0.001400	1.000000	0.057689	-0.004793
LATE_AIRCRAFT_DELAY	0.652094	0.622980	0.150353	-0.014251	0.034496	0.055270	0.003678	0.057689	1.000000	0.036826
WEATHER_DELAY	0.265420	0.270098	0.021641	0.000173	-0.001344	0.050190	-0.000709	-0.004793	0.036826	1.000000

We now plot linear regression between late aircraft delay and departure delay.



Analysis- From the linear regression line we can see that a Positive relationship exists. Whenever the delay is because of the aircraft, the flight departure will be delayed.

We now plot linear regression between airline delay and departure delay



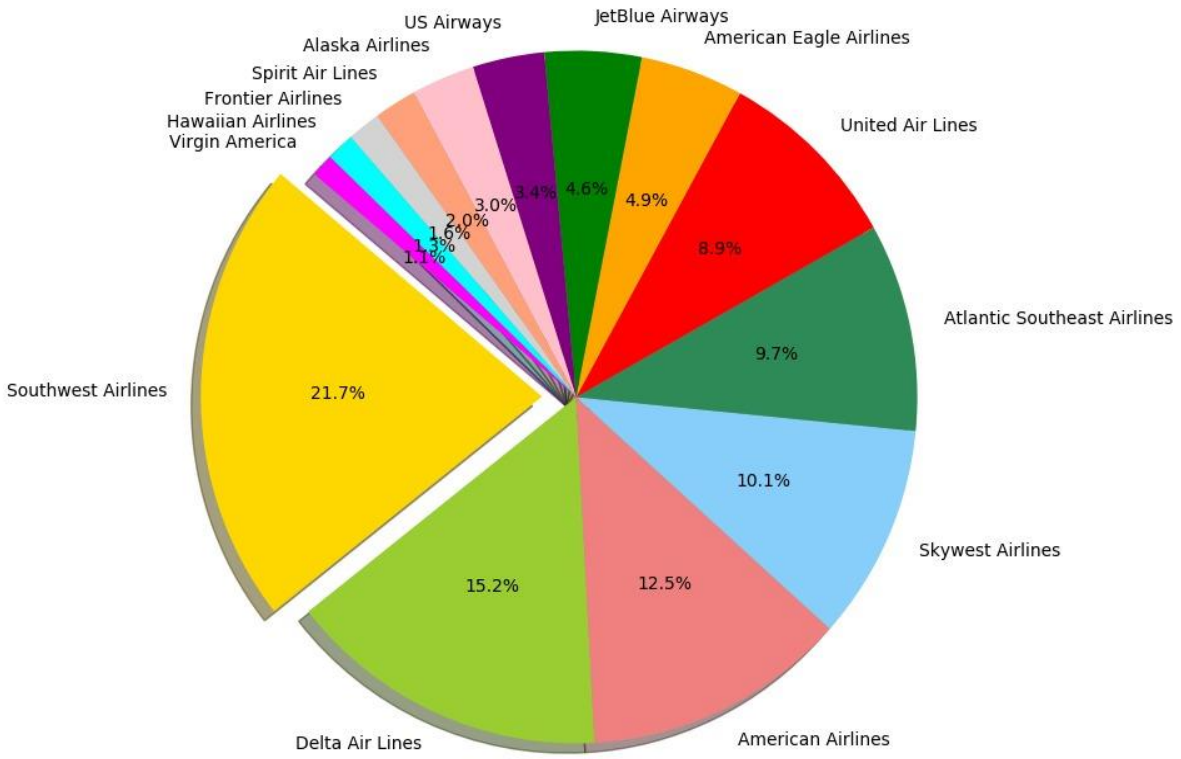
Analysis- From the linear regression line we can see that a Positive relationship exists. Whenever the delay is because of the airline, the flight departure will be delayed.

CATEGORICAL DATA ANALYSIS

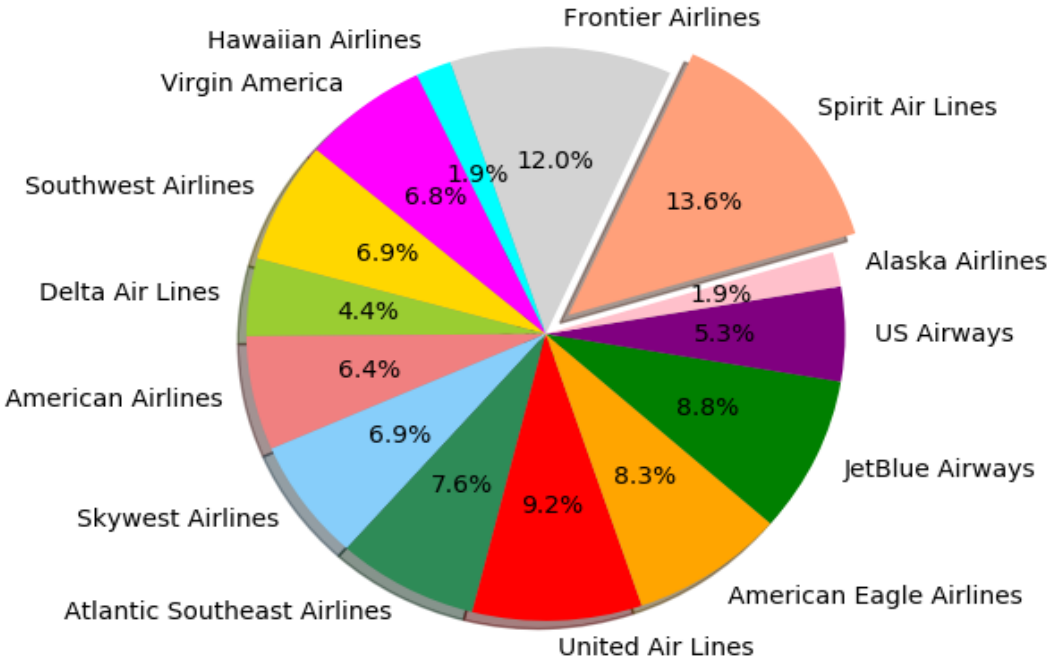
From the set of categorical data, we have chosen Airline and Airport variables to analyse.

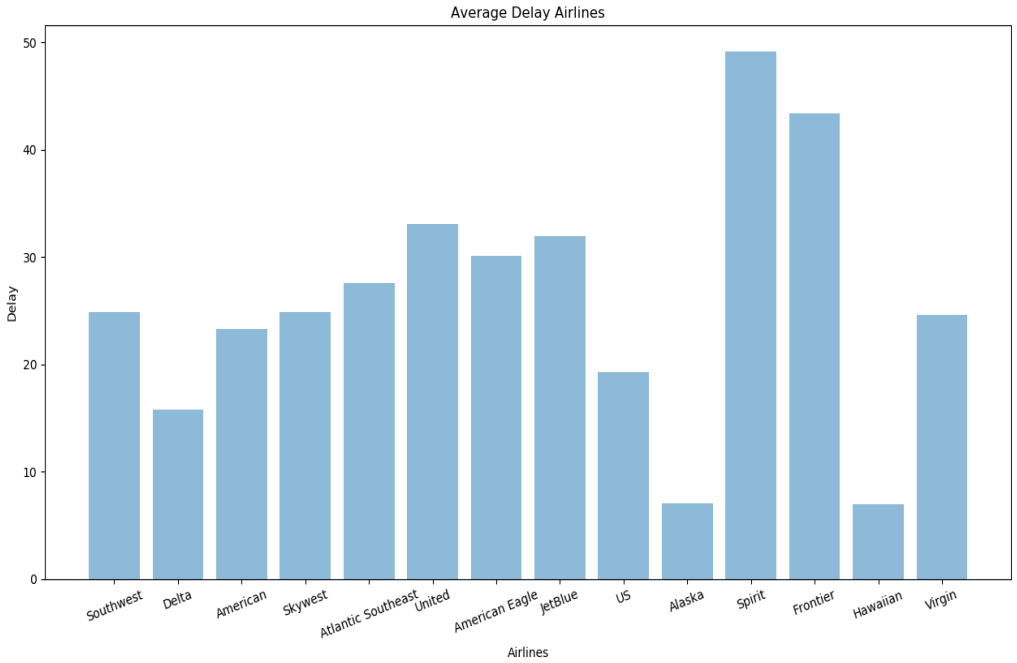
AIRLINES:

After looking at the data we believed that higher the flights count of the airline more should be the delay because of which we tried finding the airlines that flew the most in United States for the year 2015 in order to see if the count of flights has any role to play in it being delayed.
After statistically analysing, we found out that Airline with the maximum number of flights ('WN', 1242403) where WN is Southwest Airlines.
Here is the pie chart depicting the count of flights each airlines operated in the year 2015.



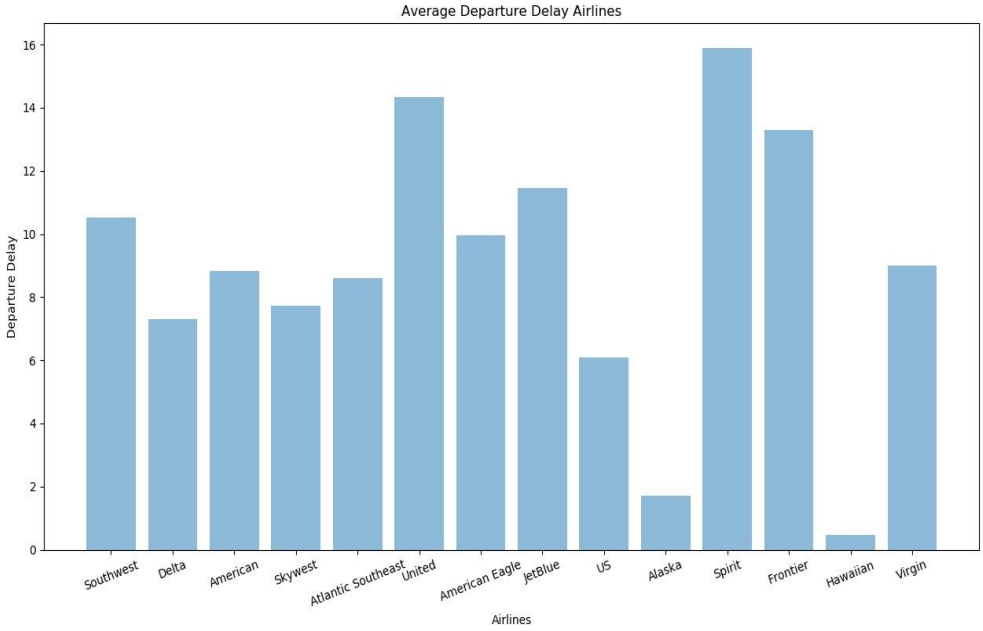
From the graph we can see that the Southwest airlines has the highest percentage of flights operating in the United States for the year 2015 followed by Delta airlines and American airlines with 15.2% and 12.5% respectively.
Analysis: As Southwest airlines co. has the maximum flights count in the United States, we tried finding if the Southwest airlines co. also has the maximum average delay.
Calculating the total average delay of each airline we find that the airline codes as ‘NK’ has the highest delay, which is not Southwest Airlines.
Now we will plot the average delays of each airline to see which airline has the maximum delay.

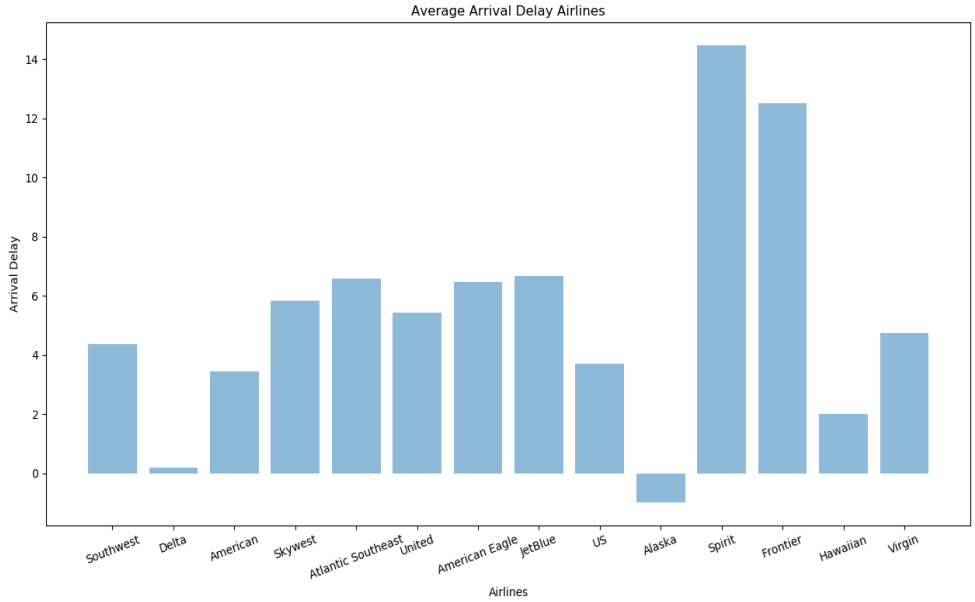




After plotting we find out that Southwest airlines co. may have the maximum number of flights running in the United States but it does not account for the maximum average delay. On the other hand, we noticed that the frontier airlines has the maximum average delay.

Breaking the delay in two parts, departure delay and arrival delay to analyzing them separately.





Analysis: After looking at the total flights number pie chart, we conclude that southwest airlines co. has the maximum number of flights flying in the year 2015 but does not play much of a role in the delays

Also, we conclude that although Spirit (2%), Frontier (1.6%) and United (8.8%) airlines hold very less share in the same plot, they are the main contributors in the delays for the year 2015.

Here is some information about the airlines causing the most delays

- The maximum average delay are spirit, frontier and united airlines respectively
- The maximum departure delay are spirit, united and frontier airlines respectively
- The maximum arrival delay are spirit, frontier and JetBlue airlines respectively.

AIRPORTS:

Now we want to the analyze airports to find the busiest origin airport and destination airport.

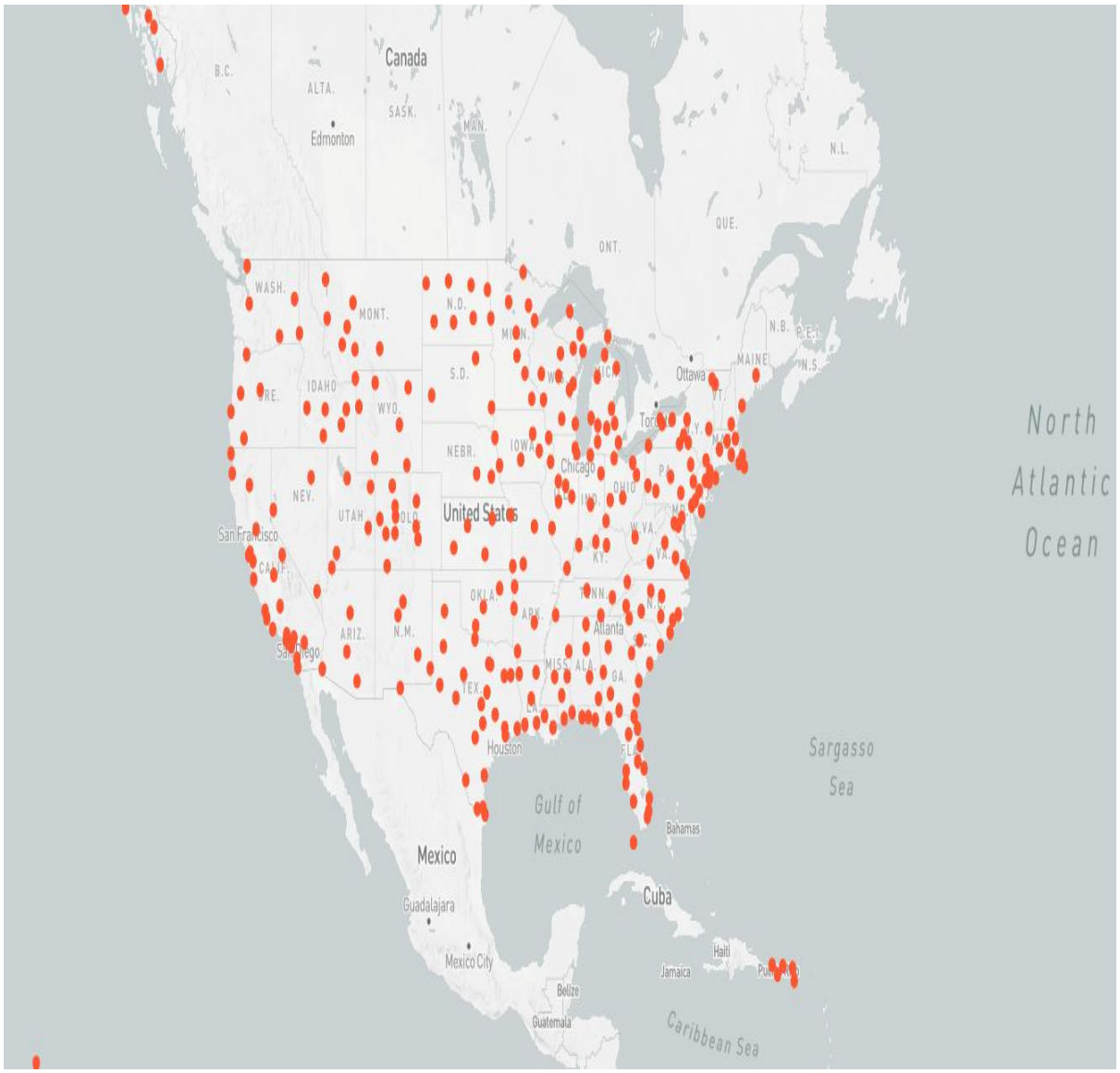
RESULTS

- Busiest Origin Airport ('ATL', 343506)
- Busiest Destination Airport ('ATL', 343076)

where 'ATL' is the Atlanta airport

Now, we wanted to see if the busiest airport is the airport with the most delays. So we calculated the airport with the maximum average delay and we find that the airport called Wilmington Airport has highest delay, which is not Atlanta airport (the busiest airport).

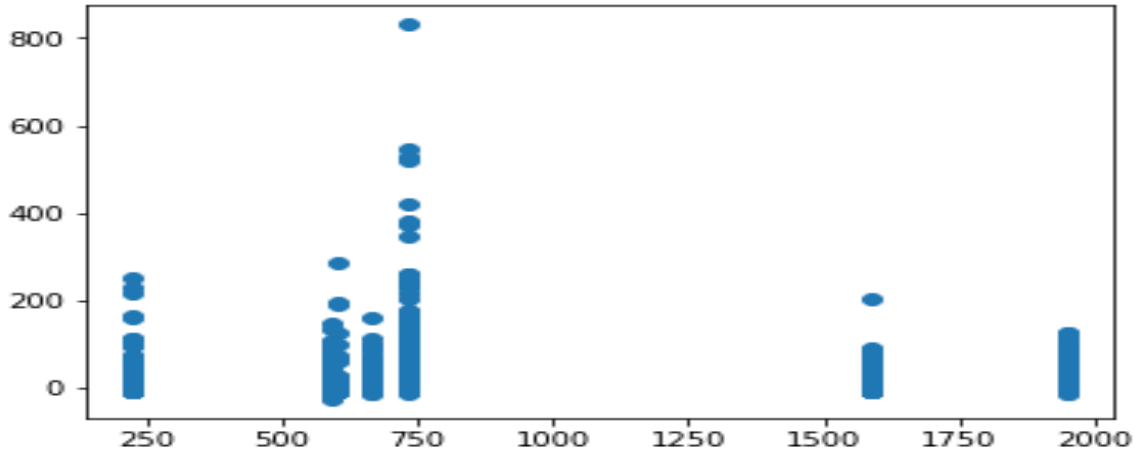
Note: as the count of number of airports is more than 300, showing a plot would make no sense, hence we tried creating plot of all the airports on the map on United States using plotly package which also had the functionality of displaying the name of the airport with average delays when you hovered over the dots in the map.



MODELING

MODEL 1: CLUSTERING

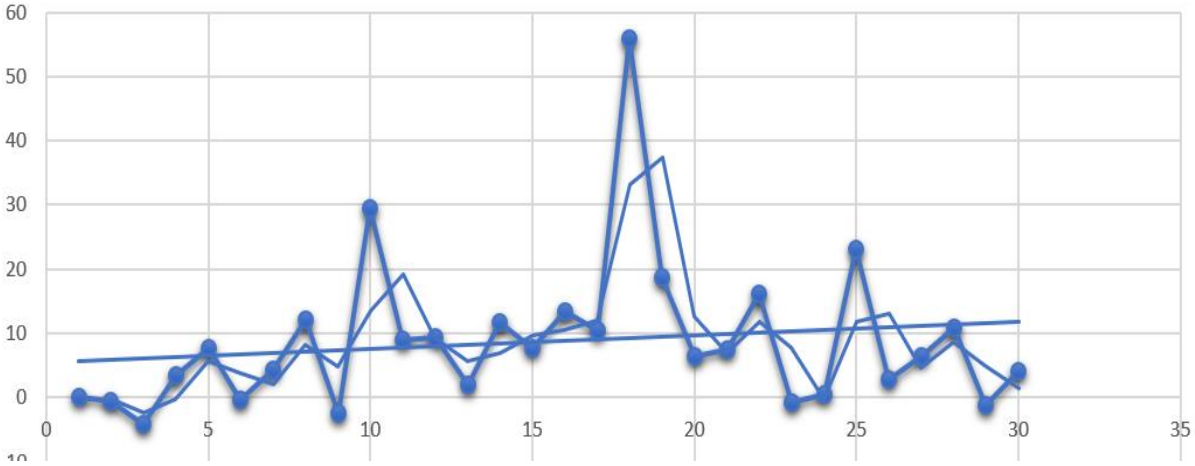
We considered using clustering model to group a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.



We expected to use this model to find a relationship between distance and delay time. After cleaning data and simple analysis, we select the airport-ATL and airline-AA to analyze. In the figure shown above, points are mainly clustered between 500 and 750 miles under 300 minutes. The distance between 750 to 1500 doesn't have any delay. This model doesn't make any sense, because points shown in this figure 1-1 aren't clustered, they are scattered as a array, which means the flight just delays at some certain random distance, which is wrong.
Hence we moved to our next model, called the Time series.

MODEL 2:TIME SERIES

A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.



We introduced this model after considering various factors which leads the flight delays such as, the trend we see when more people take flight during the weekend or business travel on

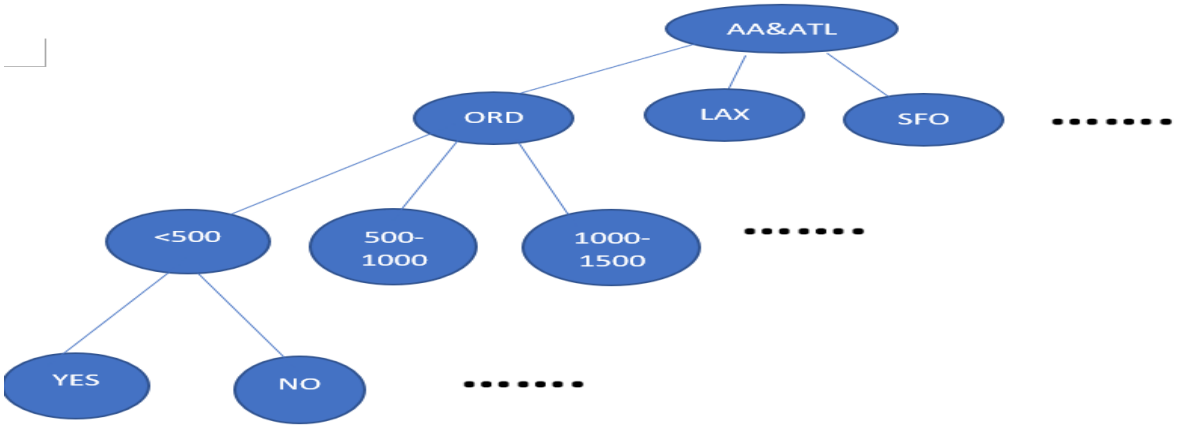
Monday, which may have a huge impact on flight delays directly. For time series model, after cleaning and filtering records, we used data from Jun.01 to Jun.30 , ATL (Atlanta) as departure airport and AA (American airlines) as airline in this model. Y axis as average delayed time and X axis as timeline. The points the figure 1-2 above moves cyclically and periodically. The reason we cannot make this model in consideration for predictiong is that there is no time series package or model in Python and it seems impossible to write one.
For next model, we tried Fixed effect.

MODEL 3:RANDOM FOREST

Random forests model is a model for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of classification or regression of the individual trees. It can overcome and correct the shortage of decision trees, which is overfitting to their training set.

```
('Number of observations in the training data:', 4653430)
('Number of observations in the test data:', 1165649)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None,
criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=5, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
{'n_estimators': 5, 'criterion': 'gini'}
[mean: 1.00000, std: 0.00000, params: {'n_estimators': 5, 'criterion':
'gini'}]
Mean cross validation score is: 1.0
('Random forest training and testing with with non-redundant variables
took [', datetime.timedelta(0, 578, 356000), '] seconds.')
Accuracy: 100.0%
Recall: 100.0%
Confusion matrix:
[[443170      0]
 [     0 442074]]
Area under the ROC curve: 0.0
```



1-7

In Random forest model, because the dataset we use is medium sized, analysing the categorical nature of many variables we find Non-linearity of the dataset (Linear Regression was attempted but had high MSE; Linear relation doesn't exist for many variables) and underlying complex dependencies exisit between variables and Non-parameteric. We think the random forest should work and perdict the flight delays.
When training the model, we selected the data of the month of Decmber. The training set info: test set = 8:2, airline is AA (American airlines), departure airport is ATL (Atlanta), every 500 miles per distance unit, etc in total 27 dimensions, also adding a binary element for delay, if delay exisits, the answer is yes else no replaced by 1 and 0 while coding. In ROC, from fig 1-6, we get a straight line, meaning we get a 100% accurate model, which is impossible when the dataset includes 40 thousand irregular records .

MODEL 4:FIXED EFFECT

A fixed effects model refers to a regression model in which the group means are fixed (non-random). Generally, data can be grouped according to several observed factors. The group means could be modelled as fixed or random effects for each grouping. In a fixed effects model each group mean is a group-specific fixed quantity.

Formula $$y_{it} = X_{it}\beta + \alpha_i + u_{it}$$

Fixed effect fit for the model parameters are fixed or non-random quantities. This is in contrast to random effects models and mixed models in which all or some of the model parameters are considered as random variables.
In this model, Y_{it} is the dependent variable observed for individual *i* at time *t*. We separate the formula 1-3 into two parts, the first part is $Y_1 = X_{it}$, where X_{it} is a matrix, where we set *i* as time frame concluded by 96 parts (every 15 minutes for a 24 hoursday) and set *t* as dates in the month of June. The average delay per 15 minutes is recorded. The second part is $\alpha_i + U_{it}$ as erro. α_i Is the unobserved time-invariant individual effect and U_{it} is a matrix. $N^k = 0$, the difference between the average delay and schedule of every element in U_{it}, so U_{it} matrix has the same scale as X_{it}. This model is supposed to work and we are still building this model.

Although we have finished this course, we will revise and adjust this random forest model as well as fixed effect model.

CONCLUSION

The most substantial take away we received from this project was that we learned how it is to thoroughly understand your data and how carefully you must choose a dataset to accurately analyse the data. This became abundantly clear when we had already faced our first roadblock. After finding a solution for the same we worked further to expand the boundaries of the project as we began to see how much difference having a large amount of data can have. It allows you to have more insight into how the variables interact with each other and try to form relationships between them. Working on this project has overall helped us in developing hands on experience in dealing with data.

We analysed all the variables, numerical as well as categorical, univariately and then bivariately. We formed relations between them and plotted several plots and figures to visually analyze them. We tried to model these formed relationships by implementing several models. We eliminated many models after implementation because they couldn't successfully predict flight delays. The two main models we focussed on were Fixed effect model and Random Forest model.