1 Introduction

The U.S. Department of Transportation (DOT) releases Air Travel Consumer Report (ATCR) on reporting marketing and operating air carrier data compiled monthly. According to DOT's ATCR reports average On-Time performance of the carriers about 75% and cancellation rate is about 1-2% . Consumers, air carriers, and airport personals are all affected by the On-Time performance and spend thousand of dollars.

There are so many reasons behind cancellations and delays for instance airlines' chronicle delay and cancellation nature, heavy or light airport traffic at both origin and destination, severe weather conditions, holidays etc..

Data is acquired from several sources that will be discussed in the next section, used for both supervised machine learning and unsupervised machine learning to develop a model to predict flight cancellation for the US domestic flights between Jan 1,1 2015 00:00 am and Dec 31, 2015 11:59 pm. In addition, the goal of this project is to build an optimization model that helps reduce the amount of cancelation and delay for both airports and airlines. This study include on-time performance of domestic flights operated by 14 large air carriers and top ten most common airports.

2. Data Collection and Upload

Data is acquired from several resources. First flight is acquired from Kaggle 2015 Flight Delays and Cancellations which has detailed information about flight's historical performance for the year of 2015.

Each row contains a unique flight details including flight date, carrier name, origin airport, destination airport, departure time, arrival time, distance, departure delay, arrival

delay, cancellation status, taxi times, and many other on-performance data. This data can be found at Capstone 1 - Preparing Data.ipynb.

Same data set can be acquired from The flight delay and cancellation data was collected and published by the DOT's Bureau of Transportation Statistics.The same data can be acquired from the The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics that tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report. DOT Bureau of Transportation Statistics is providing options to download customized data about flights. The data can be contained in downloadable excel ,csv, or other formats. Airport dataset included name of the airport, latitude and longitude. This information is used to find the location of the airport on the map, used for time zones and merged with original dataset to get additional information. You can see the detail of airport data at Airports.ipynb.

Second, daily DarkSky.net The Dark Sky API allows you to look up the weather anywhere on the globe, returning (where available):

- Current weather conditions
- Minute-by-minute forecasts out to one hour
- Hour-by-hour and day-by-day forecasts out to seven days
- Hour-by-hour and day-by-day observations going back decades

Dark Sky provides the data in JSON format that can be easily converted to pandas DataFrame. This daily data has included as temperature, humidity, visibility, wind direction, weather condition etc.. There are 1000 API call limits per API key. There were 365 days call per airports, so total there were 365 * number of airports call to get weather data. Details of this process can be found at Capstone 1 Weather API.ipynb.

Third, United States of America holidays for year 2015 has been added to set that can be found at Capstone 1 - Preparing Data.ipynb as well. This data will be used to see if holidays have any effect on flight cancellation or delays.

At last, flight details, origin and destination airports, each airline names and codes, weather data for each origin and destination airport are merged. Data is cleaned and prepared for the data exploration and modelling. You can find the merging progress here at Capstone 1 Merging Airports-Airlines-Flights- Weather.ipynb.

3 Data Exploration

3.1 Data Wrangling

Data set contained 5819079 rows 31 dimensions flights in year of 2015. To ease the computation, the top 10 destination airports (in terms of most traffic) were selected. Data was reduced to top 10 destination airports to reduce the amount of computation. Top 10 airports made up about 34% of the data.

Data set has scheduled departure time, departure time, wheel off, wheel on, scheduled arrival time, and arrival time. Those are given in only four letter formats and wont

specify if the day changes because of the delays, or the time zone difference. Data was corrected and all time and dates are accurate.

Each column has a relationship with cancellation and both origin and destination delay. Exploratory data analysis will communicate how these each relationship affect each other and what is the correlation between each feature.

3.2 Introduction to Data Cleaning Step by Step

3.2.1 Getting started with datetime

Data set is provided as below Table 1A. This provided some advantages and also some disadvantages. Getting each day , month, and year is easy however time and date are changing due to delay and timezone and cancellations. This data was kept as is and also converted to datetime object time that makes calculation more accurate when delay or timezone involved.

	YEAR	MONTH	DAY
0	2015	1	1
1	2015	1	1
2	2015	1	1
3	2015	1	1

Table 1.A Original Table

	YEAR/MONTH/DAY
0	2015-01-01
1	2015-01-01
2	2015-01-01
3	2015-01-01

Table 1.B Original Table with Datetime

For the origin airport, following data is provided at table 2. This is data provided quick calculation about summary statistics in terms of mean, median, and standard deviation of delay time and how it various day to day or month to month.

	YEAR	MONTH	DAY	SCHEDULED DEPARTURE	DEPARTURE TIME	DEPARTURE DELAY	TAXI OUT	WHEELS OFF
0	2015	1	1	0005	2354	-11	21	0015
1	2015	1	1	0010	0002	-8	12	0014
2	2015	1	1	0020	0018	-2	16	003

Table 2 Flight time at Origin Airport

Using the timedelta function delay times are added to schedule time, see Table 3 so that if delay causing a change in days, this changes are reflected to dataset. Also taxi out time is also has potential to change days. Instead concatenating date (YEAR/MONTH/DAY), timedelta used to make sure data is accurate.

	YEAR/MONTH/DAY	SCHEDULED DEPARTURE	DEPARTURE TIME	DEPARTURE DELAY	TAXI OUT	WHEELS OFF
0	2015-01-01	2015-01-01 00:05	2015-01-01 23:54	-11	21	2015-01-01 00:15
1	2015-01-01	2015-01-01 00:10	2015-01-01 00:02	-8	12	2015-01-01 00:14
2	2015-01-01	2015-01-01 00:20	2015-01-01 00:18	-2	16	2015-01-01 00:34

Table 3 Considering Delays and Time changes

Original data set is only providing times at the destination airport. And it does not specify if it is the same day or time zone changes etc Table

....	WHEELS ON	TAXI_IN	SCHEDULED ARRIVAL	ARRIVAL TIME	ARRIVAL DELAY
....	0404	4	0430	0408	-22
....	0737	4	0750	0741	-9
....	0800	11	0806	0811	5

Table 4 Flight time at Destination Airport

By considering the airtime which how long airline carrier takes to get to the destination airport and the time zone changes, the most accurate date and time at the destination airport is reflected at the Table 5.Data cleaning and processing discussed in detail in this IPython notebook.

....	WHEELS ON	TAXI_IN	SCHEDULED ARRIVAL	ARRIVAL TIME	ARRIVAL DELAY
....	2015-01-01 *04:04*	4	2015-01-01 *04:30*	2015-01-01 *04:08*	-22
....	2015-01-01 *07:37*	4	2015-01-01 *07:50*	2015-01-01 *07:41*	-9
....	2015-01-02 *08:00*	11	2015-01-02 *08:06*	2015-01-02 *08:11*	5

Table 5 Considering Airtime and Time zone Differences

3.2.2 US Holidays

One of the null hypothesis states that US holidays are affecting flight cancellation, the traffic volume, or delays. Data is inserted as a new column called IsHoliday which is labeled as 0 (False = Not Holiday) and 1 (True = Holiday). Inserting Holiday data is discussed in detail in this IPython notebook.

....	Is Holiday
....	0
....	1
....	1

Table 6 US Holidays

3.2.2 Weather Data

It is hypothesized that weather has a significant effect on delay cancellation and delays. Weather data does not come with the data set. The Dark Sky Company specializes in weather forecasting and visualization and providing minute to minute weather forecasting. Dark Sky API provides the weather conditions, in a convenient JSON format.

Whether data acquired and merged with the data set. Weather data API requests are discussed in detail in this IPython notebook.

Example API Request

https://api.darksky.net/forecast/[key]/[latitude],[longitude],[timestamp]

Here are the definition of each parameter for the API request at Table 7.

API Key	Latitude	Longitude	Timestamp
Unique Key	Airports' Latitude	Airports' Longitude	Airports' Timestamp at 2015-01-01 00:00

Table 7 API Requests

Time Zone differences are very important for the correct API request. Because I get the data based on local timezone. In order to get accurate data, timestamp in target time zone should be adjusted. This process is discussed in detail in this IPython notebook.

"For example, the UNIX time 1466892000 corresponds to 6PM (18:00) in New York, 10PM (22:00) GMT, and midnight (00:00) of the following day in Los Angeles. When converting UNIX timestamps to local times, always use the timezone property of the API response, so that your software knows which timezone is the right one." (Dark Sky Q&A)

4. Exploratory Data Analysis

4.1 Airports

Data set has 5819079 number of domestic flights in 2015. In order to reduce the computation time and the complexity of the dataset, only 10 busiest airports are selected from origin and destination airport. Number of flights have been reduced 533183 which about 9.2% of the dataset. This is a good rule of thumb that sample data is less than 10% and number of samples are more than 35. We can definitely apply Central Limit Theorem for the summary of statistics.

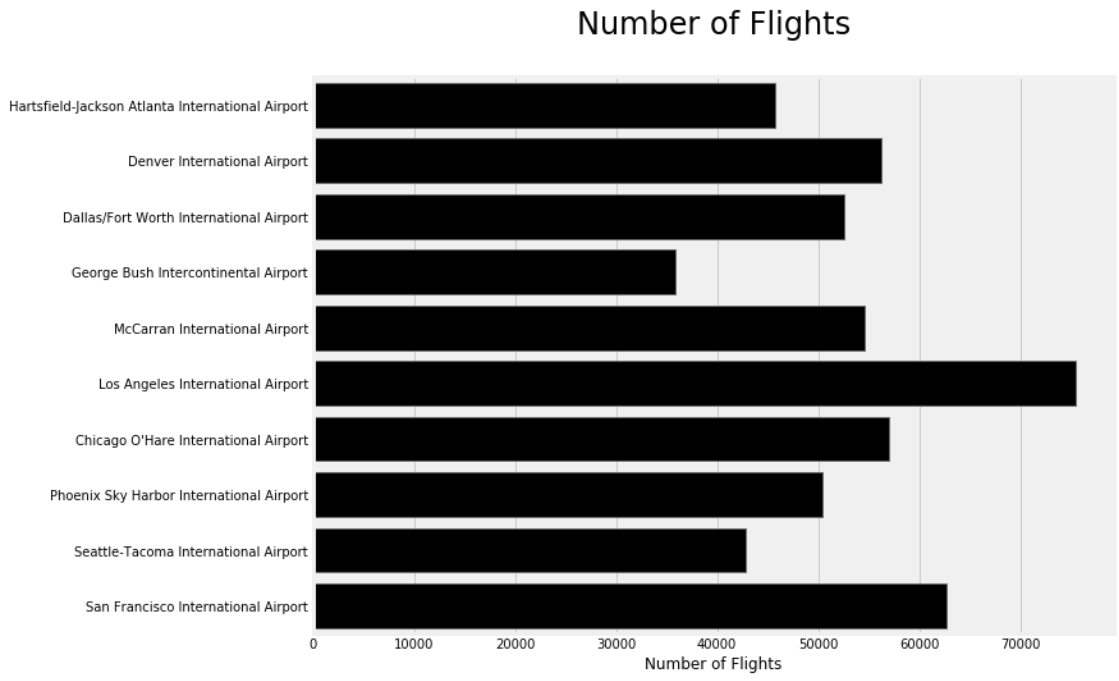

Figure 1. Airport Traffic

Most common 10 airports locations and size of traffic is illustrated in Figure 2 below. Darker the color is more busier the airport is, also larger the size of the circle is more traffic the airport has.

Selection of these 10 most common airport are discussed in detail at this IPython notebook. Data visualization of the figures can be found at the IPython notebook here.

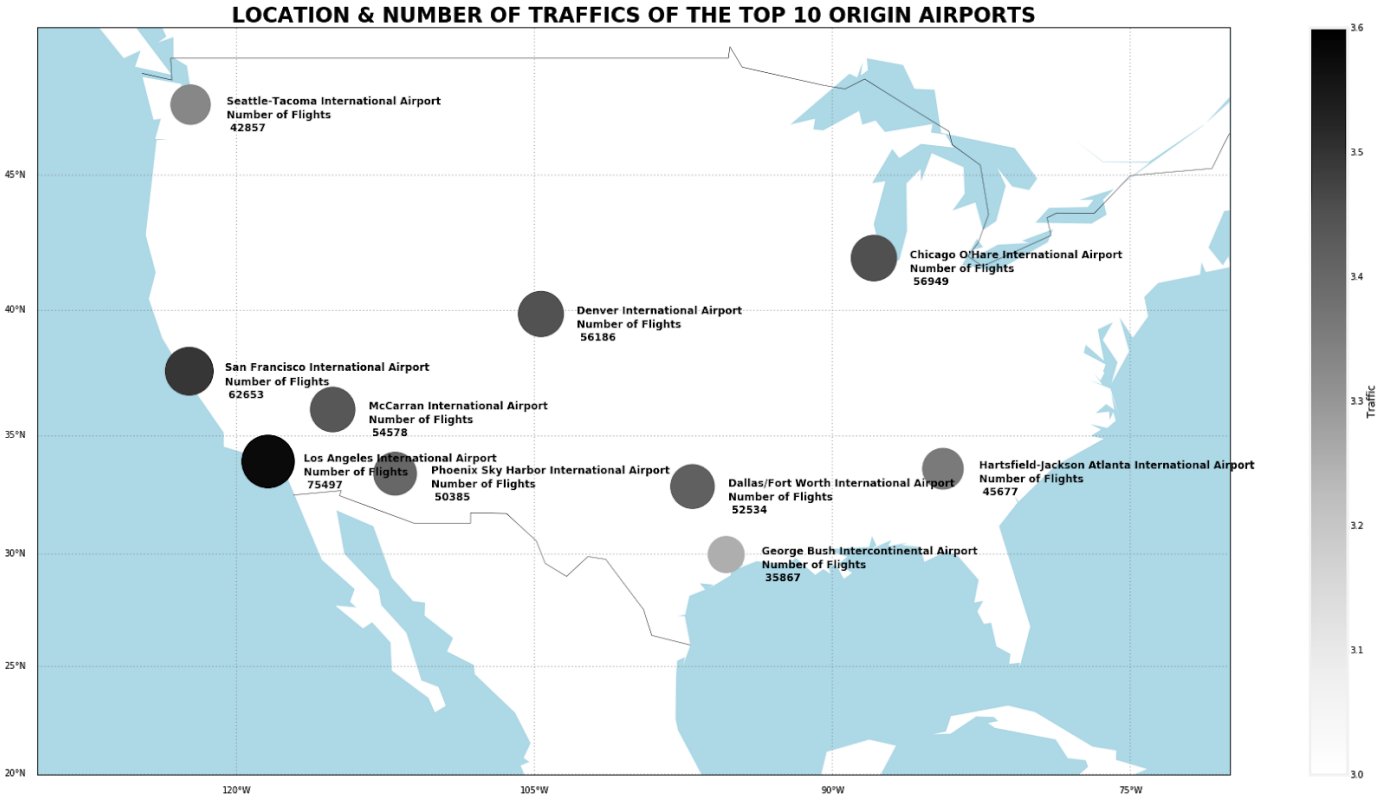

Figure 2. Airport Locations and Traffic Sizes

4.2 Flight Cancellation and Trends

There are about .5M flights between flight from busiest domestic airports in the US. The routes are illustrated in Figure 3.

There are 5398 flights cancelled for these busiest 10 airports. This makes only 0.96% of the data. This cancellation rate looks very insignificant, however 5398 affect thousands of people as passengers, crews and airport personal as well as airline companies.

Cancellations also cost millions of dollars to the US economy and greatly affects airline companies.

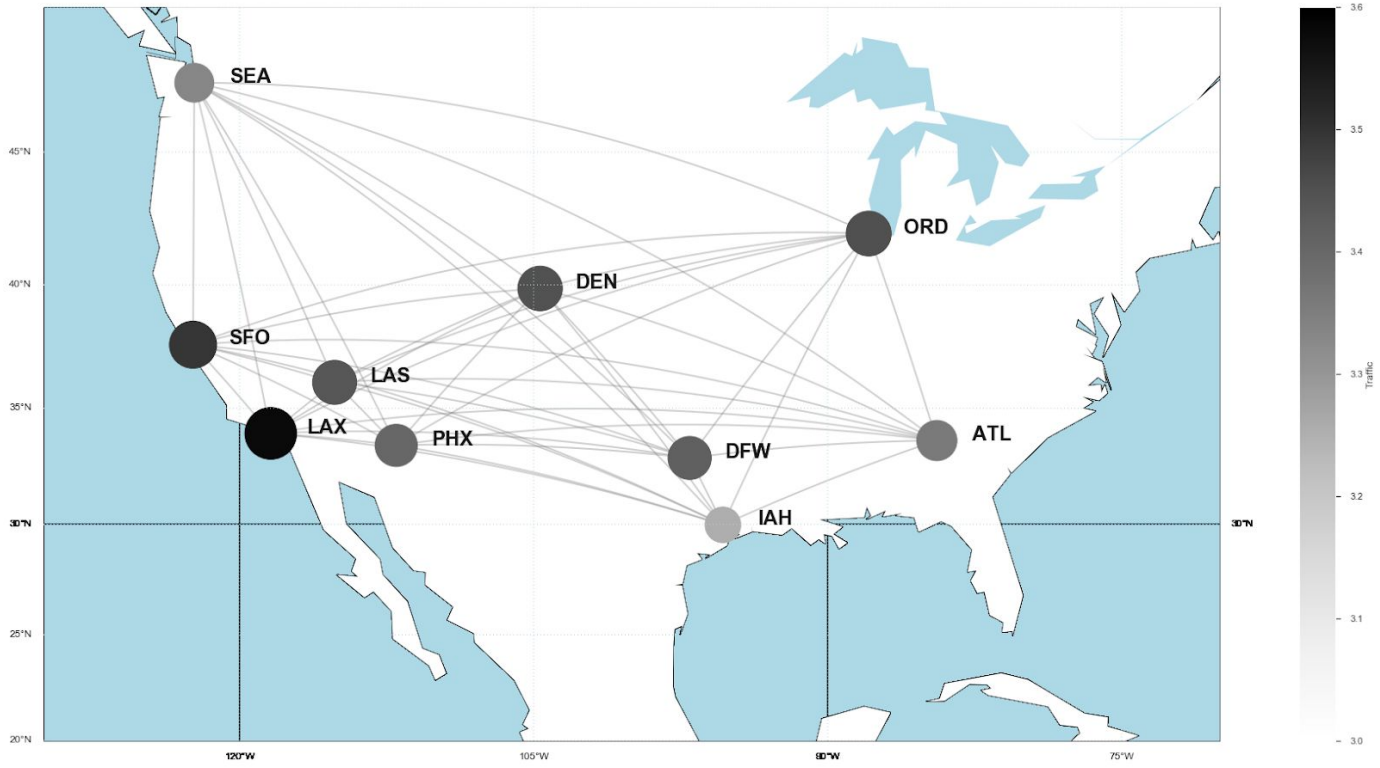

Figure 3. Flight Network Between 10 Most Busiest Airports

It is important to understand how cancellation happens, when it happens, is there a predictable trend?

By plotting the cancellation data per days, months, quarters, and annually there are some clear trends. It can be concluded that there are fewer and fewer cancellation during summer season, however there more cancellation during winter seasons or toward the beginning and the end of the winter. Therefore by looking at the trend, I can be said that weather has a significant effect on flight cancellations.

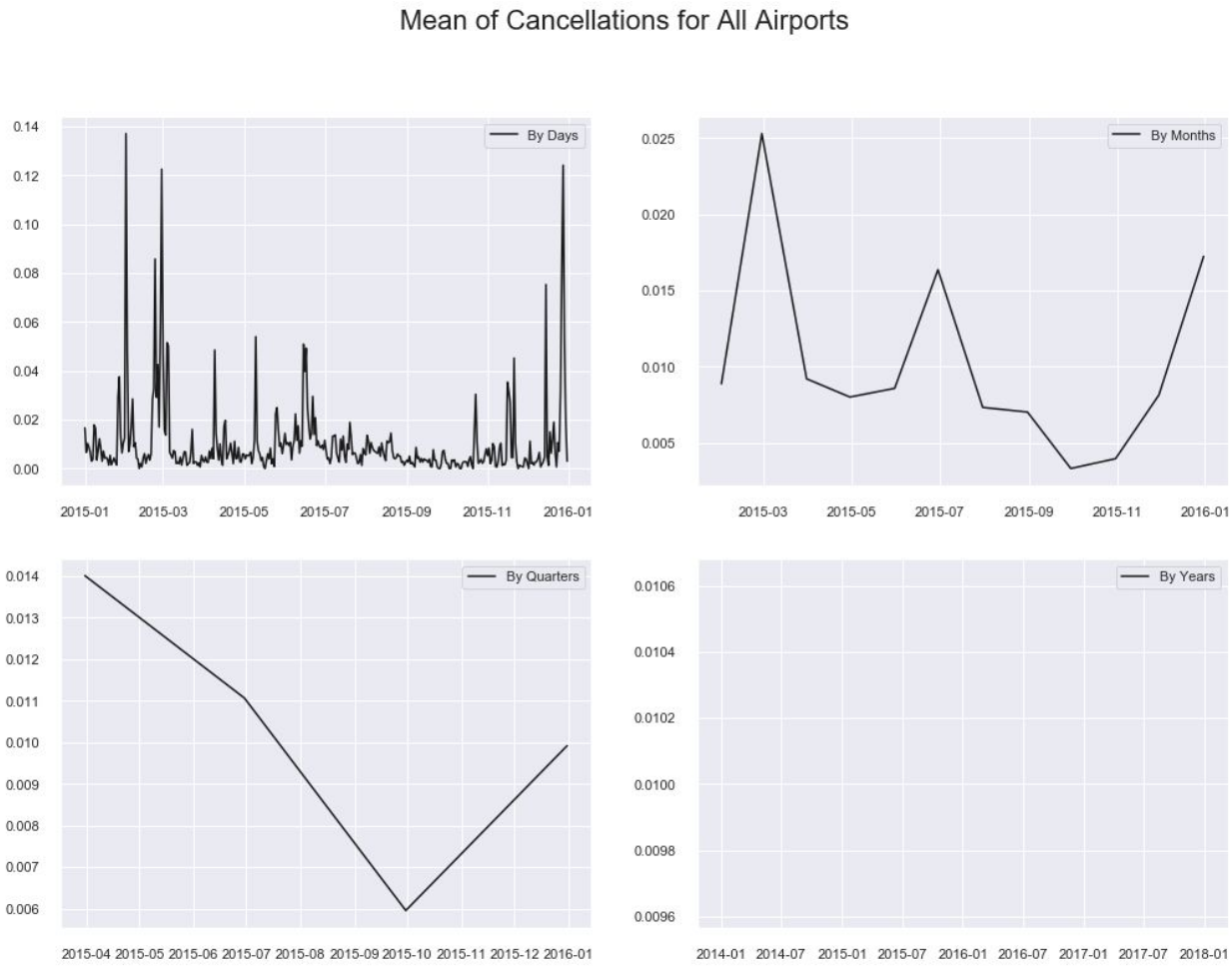

Figure 4. Flight Cancellation Trends Seasonally

To see the cancellation ratio per day, the number of flights per day and number of cancellation per day are plotted on the same graph, see Figure 5. It can be said that the number of flights is pretty much stable throughout the year. There is no clear relations between volume of air traffic and the cancellation. But still, Cancellation ratio for each day can be calculated and plotted, see Figure 6.(Basic numpy function is applied for the ratio)

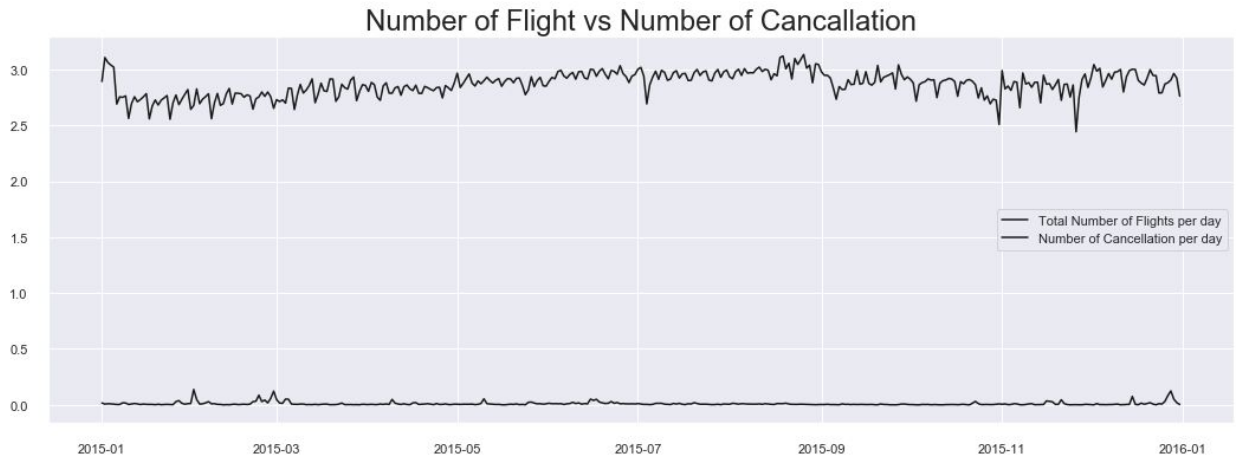

Figure 5. Number of Flight per Day vs Number of Cancellation per Day

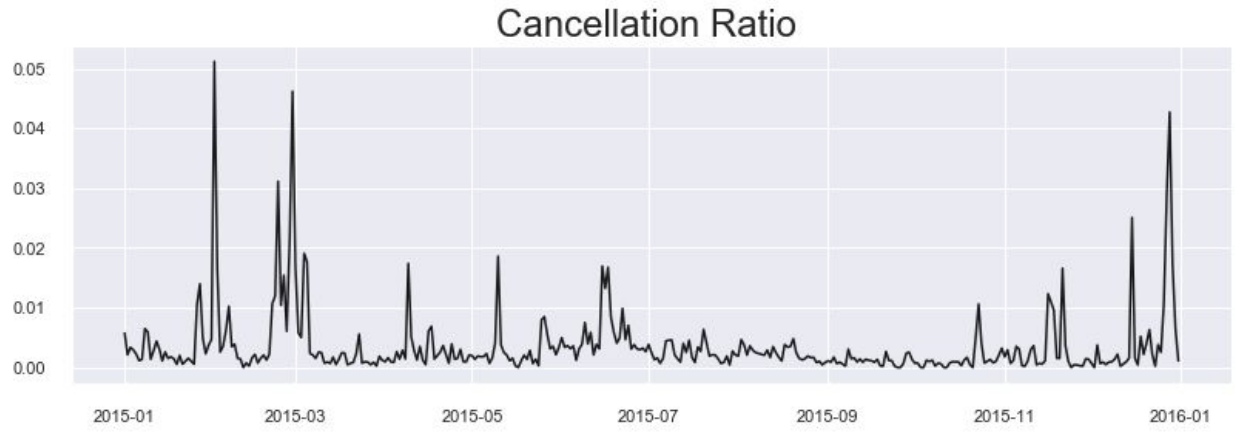

Figure 6. Cancellation Ratio (Cancellation/Number of Flights)

4.3 Other Cancellation Variables

There are 73 dimensions (columns) of the data set. Some of them can be categorized in one larger category such as whether including (wind speed, temperature, etc..). Main larger categories will be investigated on correlation to cancellations. These categories are holidays, days of the week, airports, airlines, flight distance, weather etc..

4.3.1 Holidays

Holidays are plotted along cancellation per day in 2015. At first sight When it is a holiday there is no high jump in delay numbers or otherwise. It can be said that there is no strong relation between holidays and flight cancellations. There is a high peak right after Christmas Holiday, that might be because of the holiday or its is just because of bad weather, see Figure 5.To see the justification see IPython notebook.

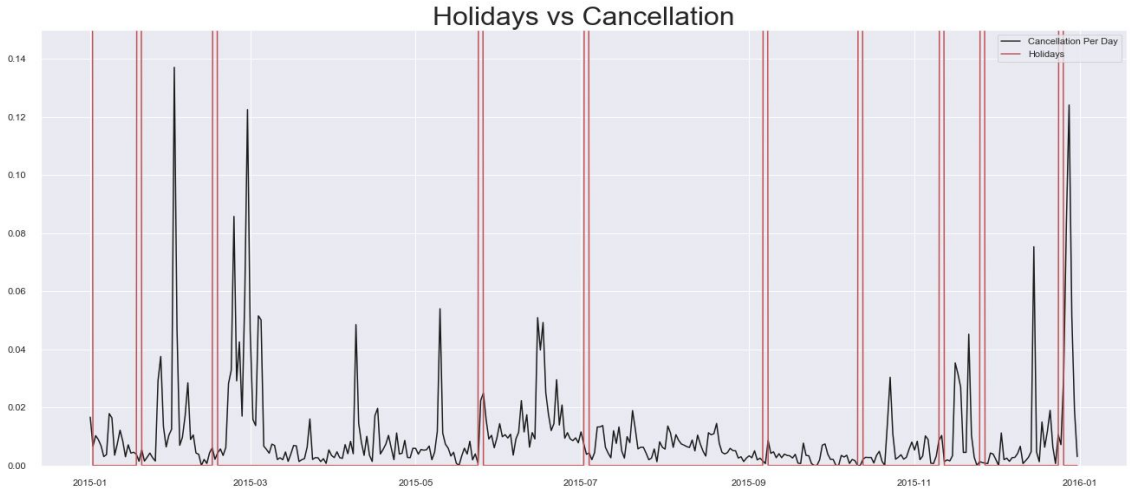

Figure 7. Holiday vs Cancellation

4.3.2 Airports

Due to schedules and the volume of the airport, the cancellation rate varies. According to the analysis in Figure 7, Dallas/Fort Worth International Airport has the highest cancellation throughout the year of 2015 with about 1.6% cancellation. The airport has the least cancellation percentage is Seattle-Tacoma International Airport with 0.5% cancellation rates. See the details at this IPython notebook.

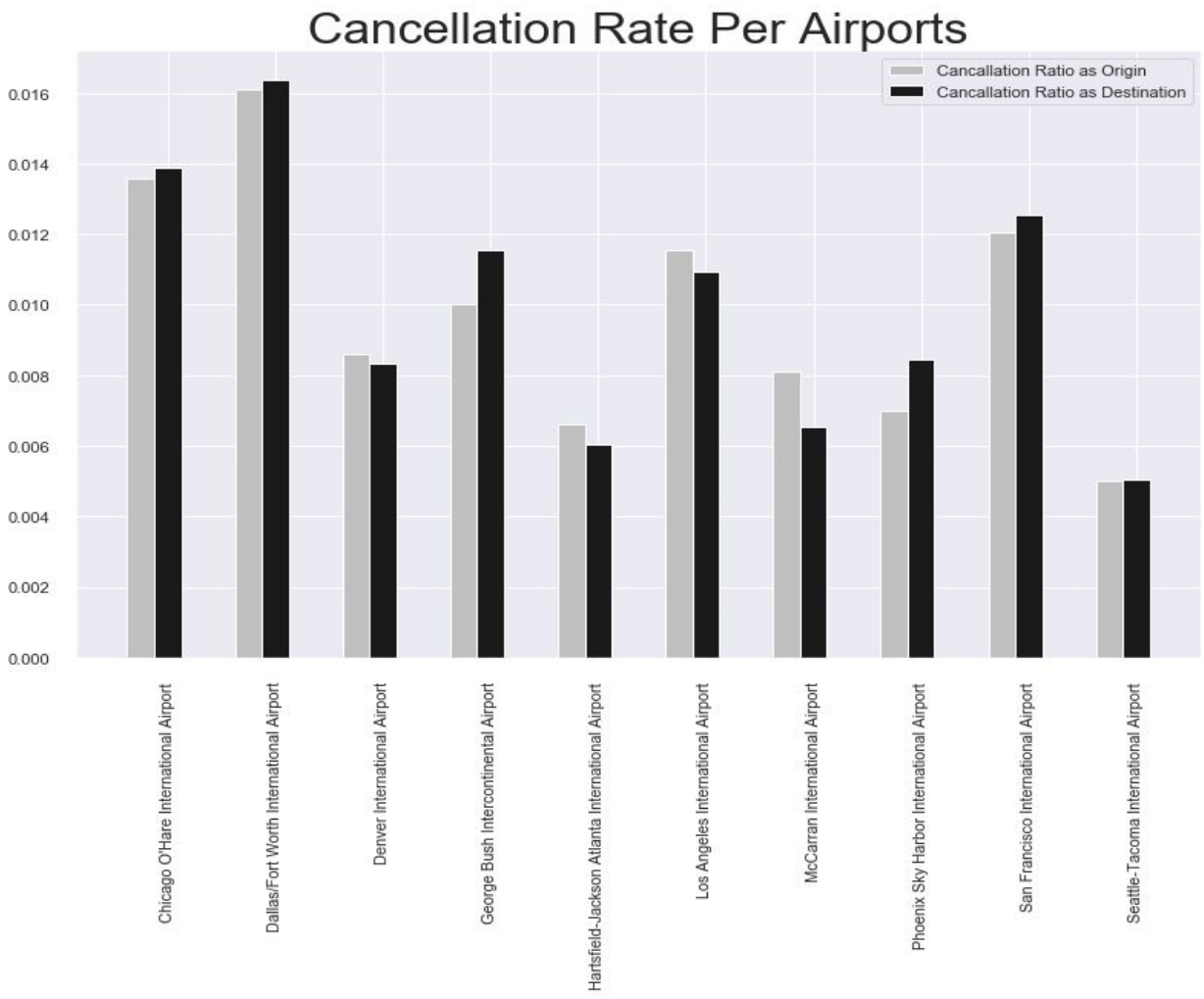

Figure 8. Cancellation Rate Per Airports

4.3.3 Airlines

Similar to analysis on airport, the cancellation rate for each airline have been communicated. Based on the analysis, American Eagle Airlines Inc. has the highest cancellation rate among all other airlines. This should be characteristic for each airline because all other variables kept constant. Meanwhile, the airline has the least cancellation rate is Delta Airlines Inc. Justification of getting the cancellation per airlines can be found at this IPython notebook.

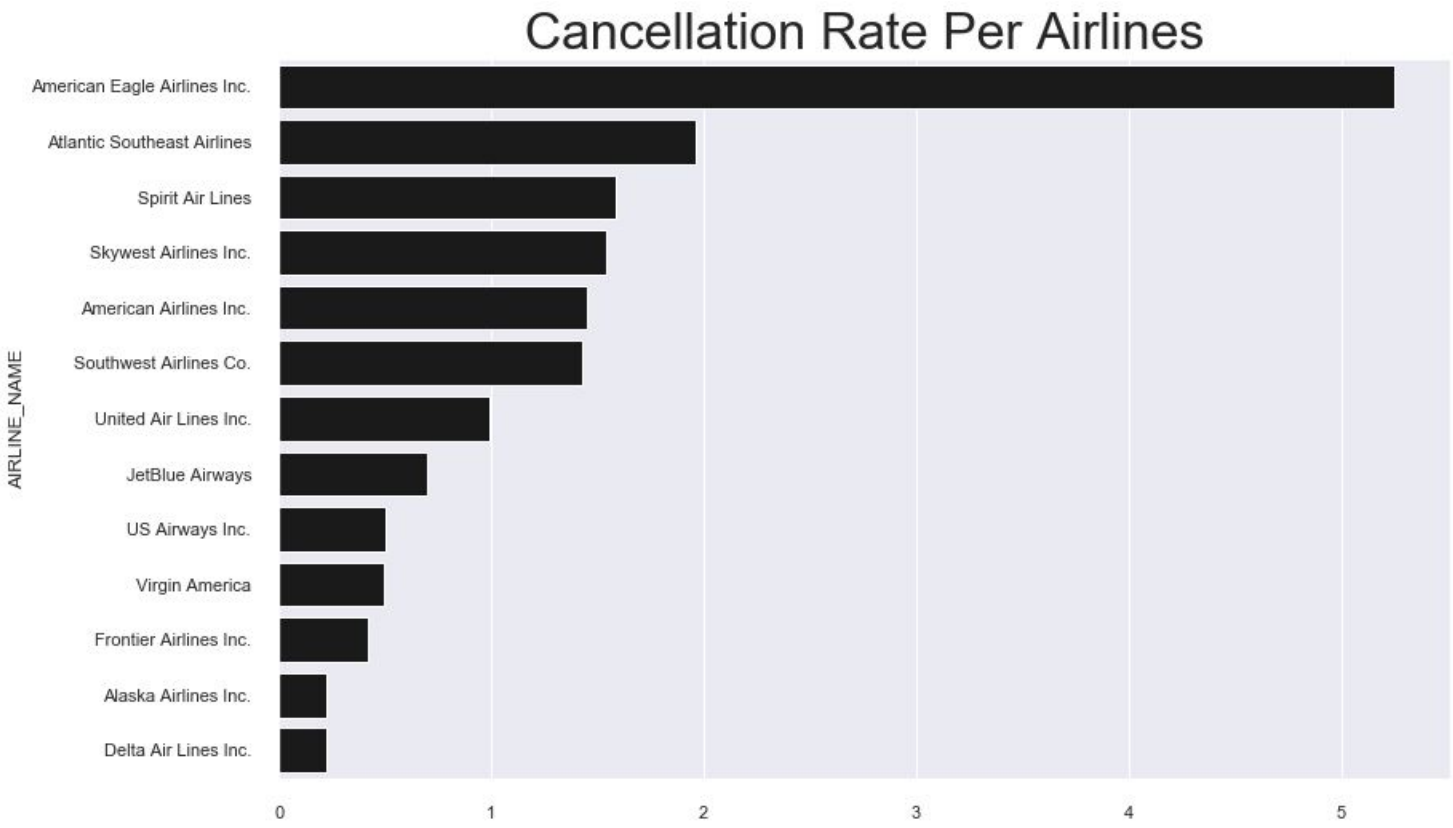

Figure 9. Cancellation Rate Per Airline

4.3.4 Date Analysis

Data set has year, month, day and day of the week.It is evident that this data is important for analysing the cancelation rates. Is there any trend from month the month, each day or each day of the week. Below you will see the sample data from the data set in *Table 8.*

	YEAR	MONTH	DAY	DAY_OF_WEEK
0	2015	1	1	4
1	2015	1	1	4
2	2015	1	1	4
3	2015	1	1	4

Table 8. Date Analysis

As it is seen from the Figure 9, cancellation varies from day to day. It can be concluded from this analysis that there are fewer cancellation in summer days that it is winter days. This is probably caused by bad weather conditions that will be analysed in further sections.

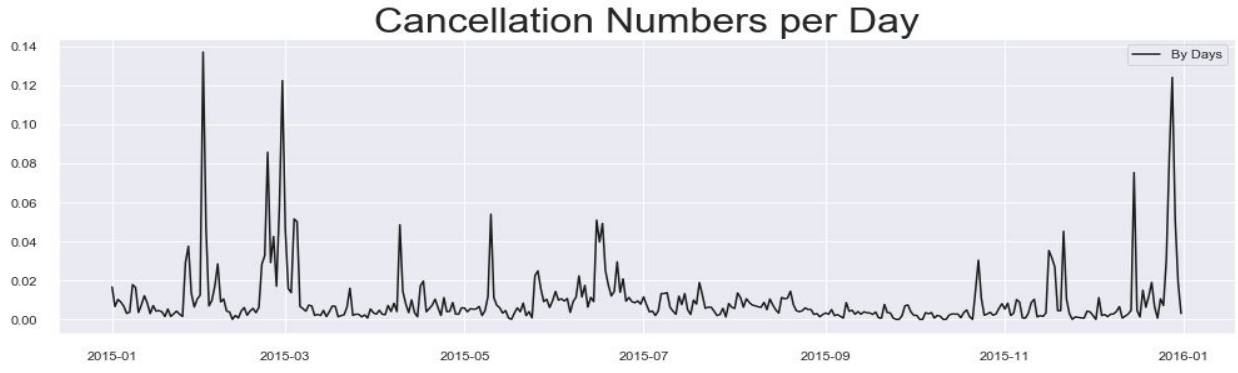

Figure 10 . Cancellation Numbers per Day

Cancelation number per month could be found at Figure 11. That clearly shows that from winter to summer there is a step decline in the trend. The cancellation numbers almost hits the zero around September and October. Cancellation rates higher in winter and very low in fall. There are still a remarkable number of flight cancellation in summer as well.

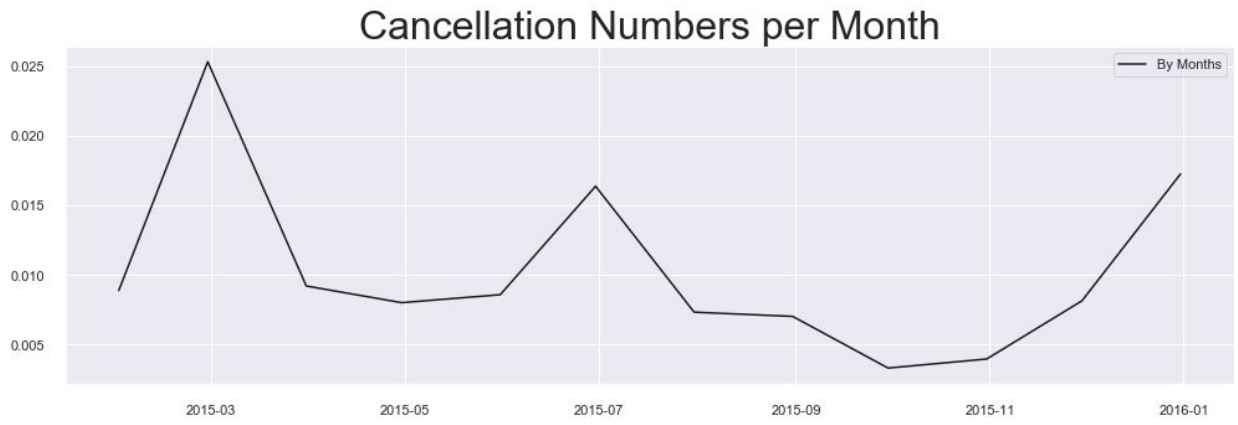

Figure 11.A . Cancellation Numbers per Month

As can be seen in Figure 11.B, there are higher number of cancellation in February and December. Also cancellation numbers are significantly low in September and October.

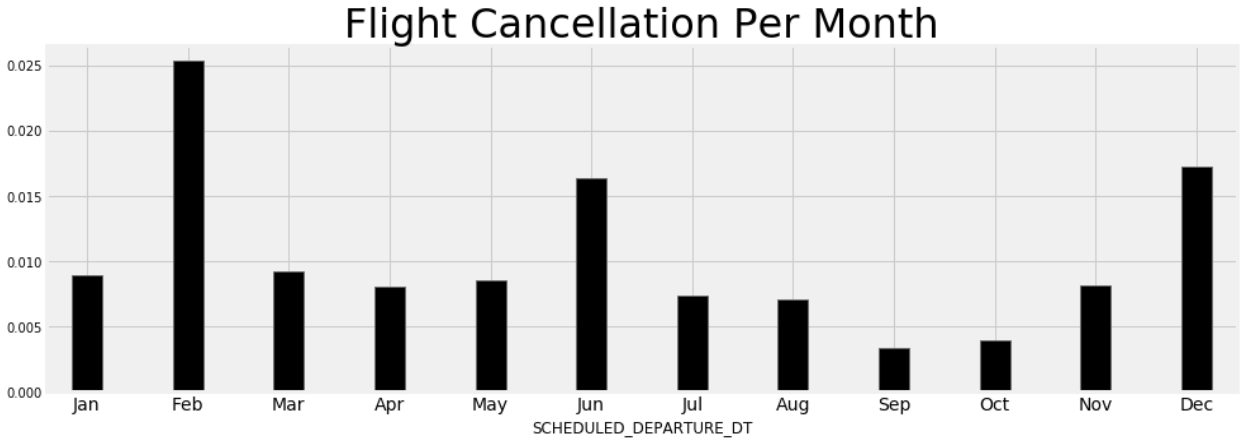

Figure 11.B . Cancellation Numbers per Month

There are more flight cancellation on Monday than any other day. Sunday comes the second for the most cancellations. The least amount of cancellation happened on Fridays. For the details analysis on date, see this IPython notebook.

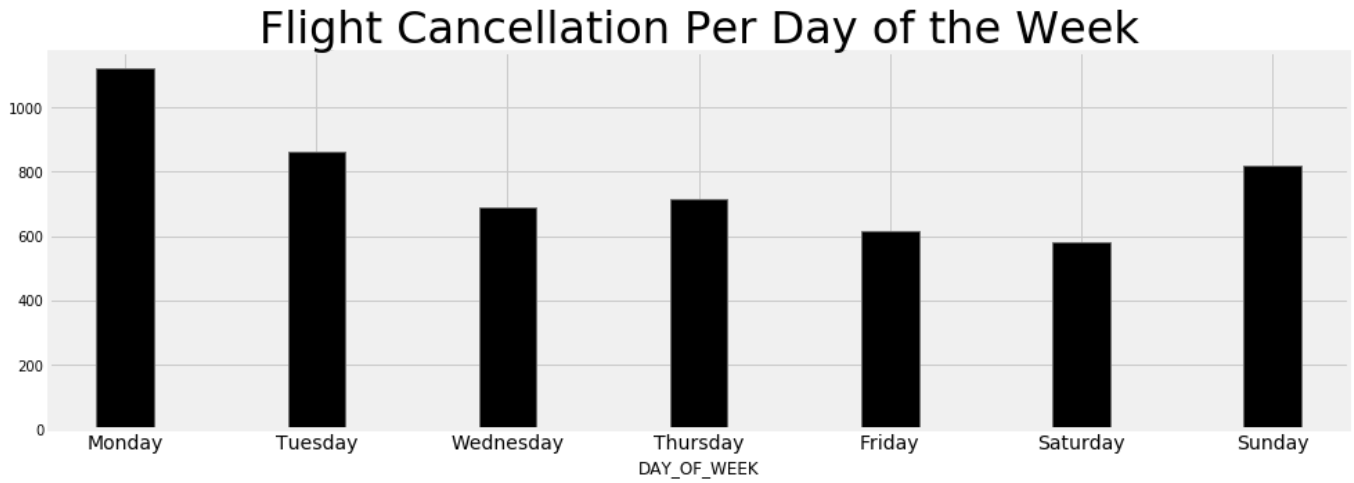

Figure 12 . Cancellation Numbers per Day of the Week

*x*4.3.5 Weather Impact

Based on previous analysis, cancellations are more common in winter times. This is probably caused by the weather. Below in Figure 11, the number of cancellation for each weather condition are shown. Despite general conception that there are more flight cancellation on snowy days, there are more cancellation on overcast and cloudy days (partly cloudy or partly cloudy). This could be about problems the visibility during take off and landings.

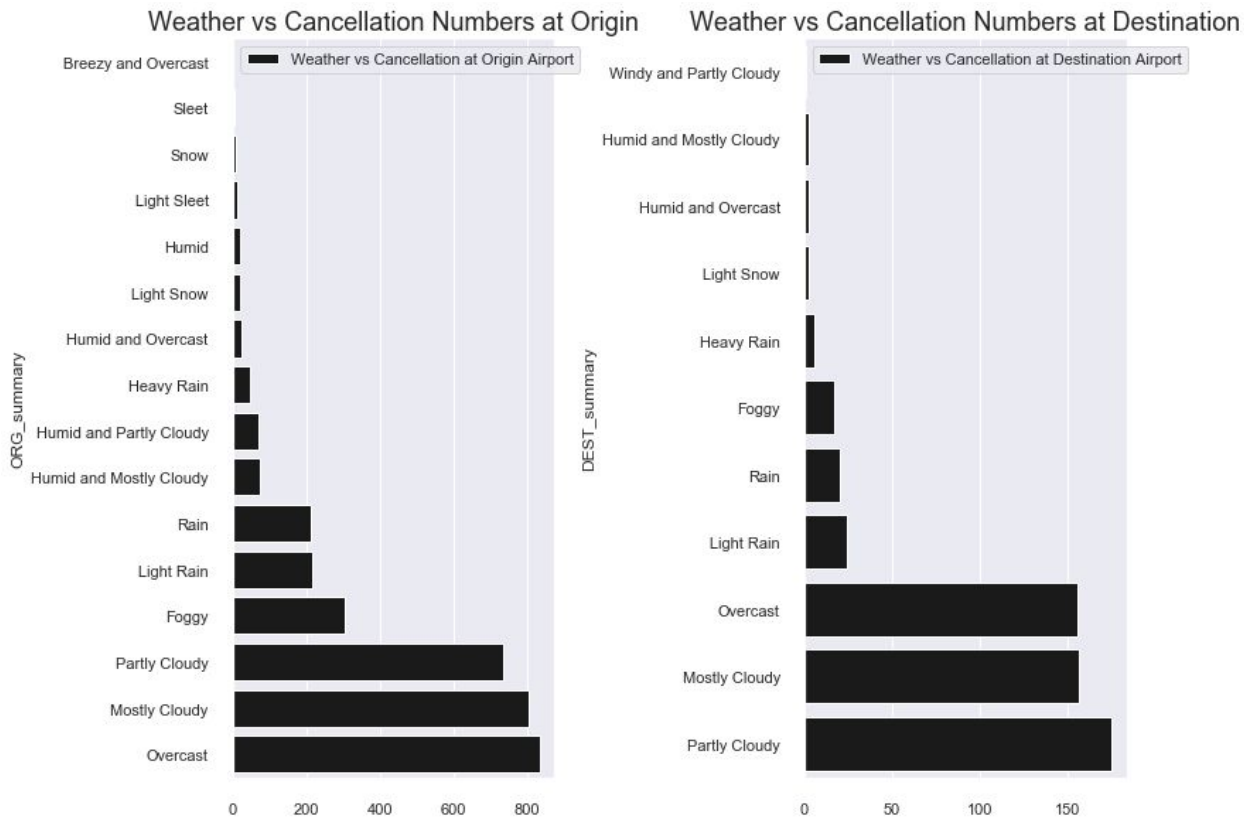

Figure 13 . Cancellation Numbers vs Weather Conditions

There are higher cancellation rates at lower temperatures. Especially, temperature that is lower than 40 degrees F has more cancellation rate. There is also a slight increase in cancellation rate for the temperature higher than 100 degrees F.

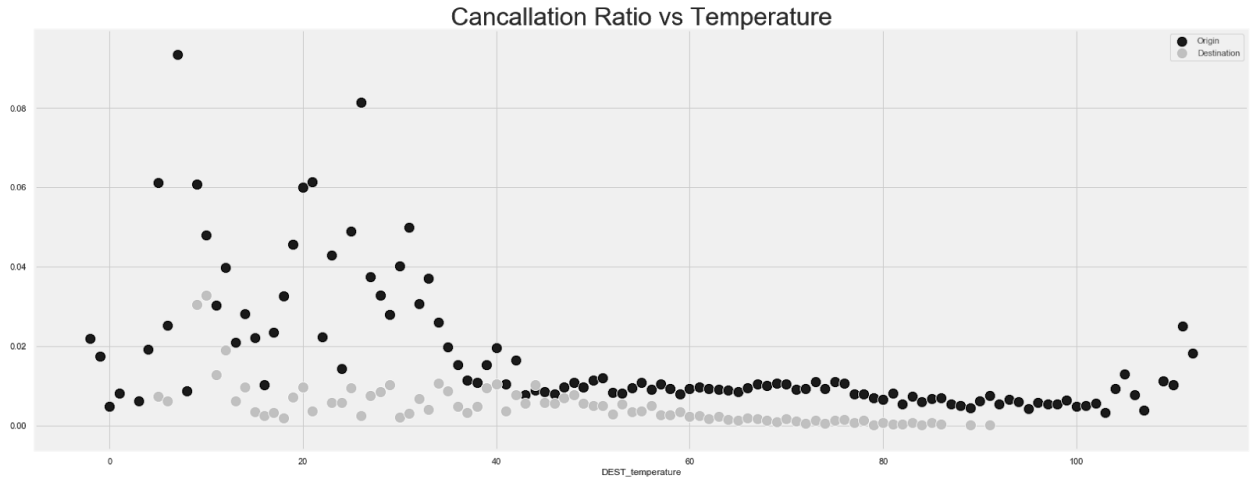

Figure 14 . Cancellation Rates vs Temperature

There is an increasing trend in flight cancellation as the humidity increases at both Origin and Destination airports. The trend of cancellation ratio is an exponential function of humidity.

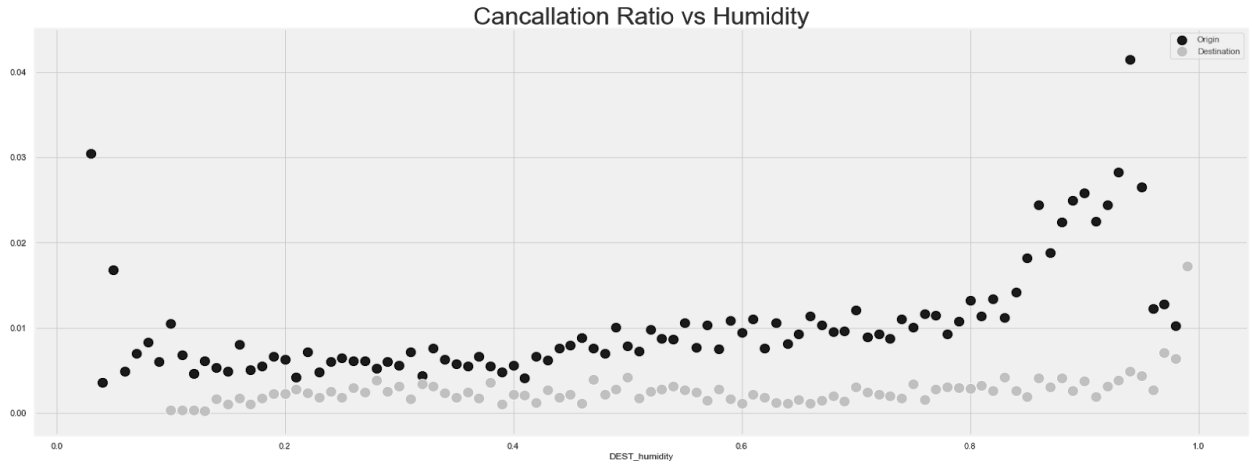

Figure 15 . Cancellation Rates vs Humidity

5. Modelling

5.1 Imbalanced Data

Most machine learning algorithms work best when the number of samples in each class

are about equal, however, imbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio of observations in each class. In this data set, the cancellation rate is 1.6%, which means that even before fitting the model. The accuracy of model will be higher than 98%. Similar problems will be found in medical diagnosis, spam filtering, and fraud detection.

To make sure the model reflects the best performance there are steps will be taken:

1.Performance metrics

1.1 Confusion Matrix**:** a table showing correct predictions and types of incorrect predictions.

1.2 Precision: the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.

1.3 Recall**:** the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.

1.3 F1: Score: the weighted average of precision and recall.

2. Using different algorithms to compare model performance

2.1 KNN Neighbors

2.2 Logistic Reg

2.3 Gradient B

2.4 Random Forest

3. Resampling: resampling module from Scikit-Learn will be used to randomly replicate samples from the minority class.

4. Using SMOTE or Synthetic Minority Oversampling Technique: SMOTE uses a nearest neighbors algorithm to generate new and synthetic data we can use for training our model.

5.2 Reducing the Size

Data set is relatively large. In order to speed up the computation, only ten percent of all data will be used. Once the best algorithm and hyperparameter is chosen, then this can be applied to the whole data set.

5.3 Encoding

Data set has 72 features (columns) like scheduled time, airline, airport etc.. and a target which flight cancellation in binary data type 0-1(0-Not cancelled, 1- Cancelled). Since data set has too many categorical data types. These types should be encoded for the modelling.

Modelling comes with mainly two restrictions:

- Categorical features should be encoded (using One Hot Encoder or pandas get dummies)
- Features and targets have to be numpy arrays or pandas DataFrame

- Futures should not contain any null values

To encode the data set for the non numeric columns from sklearn.preprocessing, One Hot Encoder Standard Scaler(Num) or pandas get dummies functions are used. This turns non numerical, in other word categorical numbers in to numerical values.

5.3 Scaling

By definition “standardize features by removing the mean and scaling to unit variance

The standard score of a sample x is calculated as:

$$z = (x - u) / s$$

where u is the mean of the training samples or zero if with_mean=False, and s is the standard deviation of the training samples or one if with_std=False.”

Data set is standardized by using the StandardScaler form sklearn preprocessing.

5.4 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is an over-sampling method meaning that I takes resamples of minority data to balance it. By definition “SMOTE selects similar records and altering that record one column at a time by a random amount within the difference to the neighbouring records”.

After applying the SMOTE, cancelled ratio 1:1. Data now can be considered as balanced.

5.4 Pipeline

Pipeline is a very useful function sklearn provides.Pipeline uses predetermined steps like StandardScaler(), Model fitting and transforming, and also useful metrics to measure model accuracy like Cross Validation, and Grid Search.

By definition pipeline is "sequentially apply a list of transforms and a final estimator. Intermediate steps of pipeline must implement fit and transform methods and the final estimator only needs to implement fit."

5.5 KNN Neighbors

Cancellation feature of the data-set, which is the 'label' is the quality of the flight data, ranging from 0 (Non-cancelled) to 1 (Cancelled). There is no missing values**.** KNN Classifier takes an unlabeled data as an input and outputs a label. The first step of the classifier is that it learns from already labelled data (training data). KNN predict the label based on the neighbors which are the closed labeled data depending on the neighbors numbers.

After the data set is encoded, it was fitted to the model using KNeighborsClassifier(n_neighbors=6) which means that it uses 6 closest data points to predict the future unlabeled data. To test the performance of the model, the accuracy of the model is tested. All machine learning model uses some sort of data splitting techniques.In this case, from sklearn.model_selection, train_test_split is used as a data splitting tool. It takes some hyperparameters such as test size that is set for 30%, and

random state that is set for 24 to keep the randomness.

Matrics	Before SMOTE	After SMOTE
Cancellation Ratio	405/39057=0.010	39057/39057=1
F1	0.9959	0.9998
Cohen Kappa	0.9959	0.9996
Brier	8.4466	0.00017
LogLoss	0.0030	0.0059

Table 9 .KNN Neighbors and GridSearch

Number of neighbors lower than 4 is under fitting the data and also the number of neighbors higher than 5 is overfitting the data. Number of neighbor 4-5 is the optimal number to use in this case.

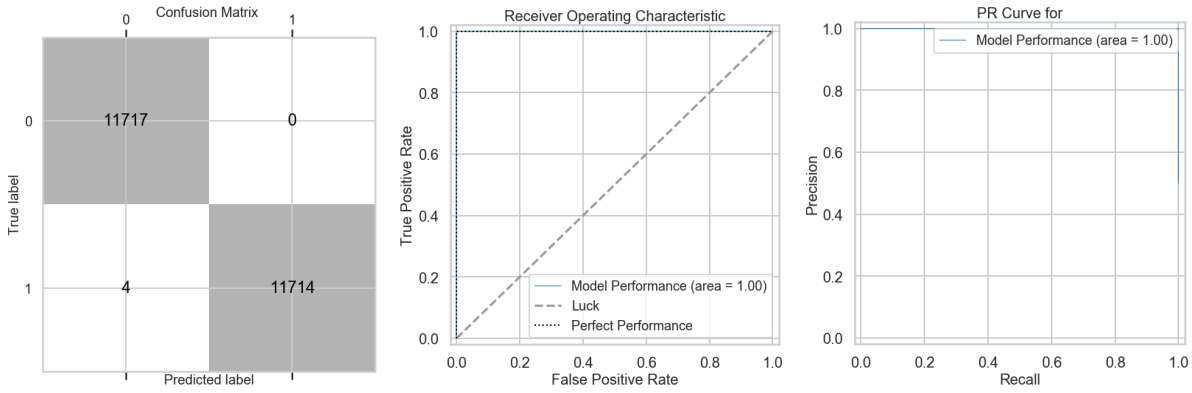

Figure 16 Confusion Matrix ,ROC Curve and PR Curve

5.6 Logistic Regression

Logistic regression is a classification model rather than being a regression model.It is works best with binary classification (0-1). For the prediction on flight cancellation, which is the target data contain 0s and 1s, logistic regression seems like a good model choice.

Best parameters are:

Model C	Model Class Weight	Model Penalty	Best Score
0.1	balanced	l1	0.99995

Table 10 .Logistic Regression Best Parameter and Best Score

Matrics	Before SMOTE	After SMOTE
Cancellation Ratio	405/39057=0.010	39057/39057=1
F1	0.7731	0.9964
Cohen Kappa	0.7703	0.9930
Brier	0.0056	0.0035
LogLoss	0.025	0.012

Table 11..Logistic Regression and GridSearch

Model performed a lot better after using the SMOTE function. SMOTE take resample of the minority to balance the imbalanced dataset.

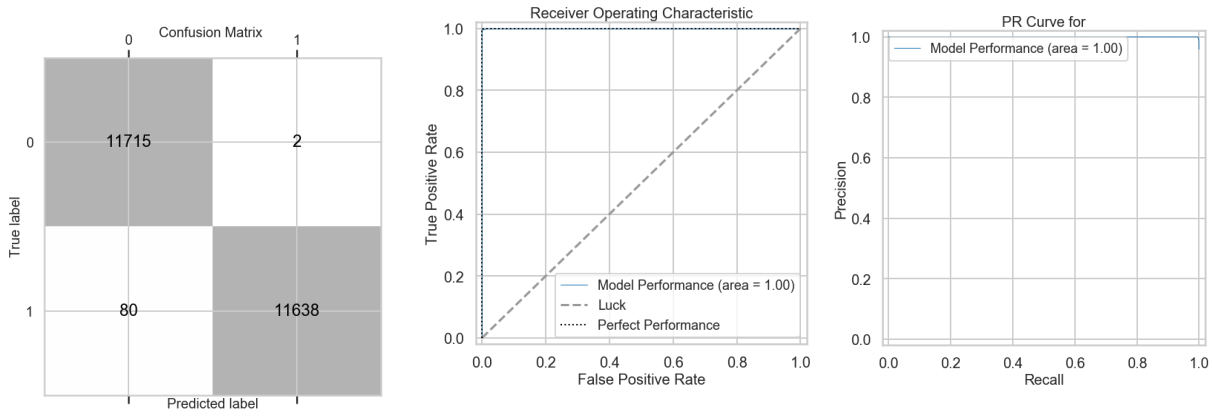

Figure 17 Logistic Regression Confusion Matrix ,ROC Curve and PR Curve

5.7 Random Forest

Random forest is used here because Random Forest works well with a mixture of numerical and categorical features. Also, Random Forest handles unbalanced data pretty well. Results have shown below in Table 10.

Boot strap	Max depth	Max features	Min samples leaf	Min samples split	N estimators	Best score
True	80	3	3	8	100	0.99995

Table 12 .Random Forest Best Parameters and Best score

Matrics	Before SMOTE	After SMOTE
Cancellation Ratio	405/39057=0.010	39057/39057=1
F1	0.9568	0.99995
Cohen Kappa	0.9564	0.99994
Brier	0.00092	4.27
LogLoss	N/A	N/A

Table 13 .Random Forest and GridSearch

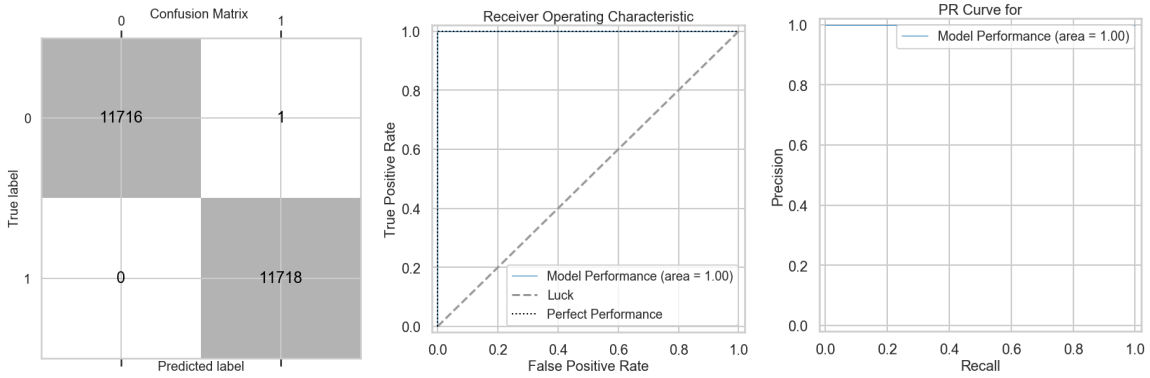

Figure 18 Random Forest and Confusion Matrix ,ROC Curve and PR Curve

5.8 Gradient Boosting Modelling

Boosting is a sequential technique which works on the principle of ensemble. It combines a set of weak learners and delivers improved prediction accuracy.

Max Depth	Min Sample Split	Best Score
5	200	1

Table 14 .Gradient Boosting Best Parameters and Best score

Matrics	Before SMOTE	After SMOTE
Cancellation Ratio	405/39057=0.010	39057/39057=1
F1	0.9918	1
Cohen Kappa	0.9917	1
Brier	0.00016	0.0
LogLoss	N/A	N/A

Table 15 .Gradient Boosting and GridSearch

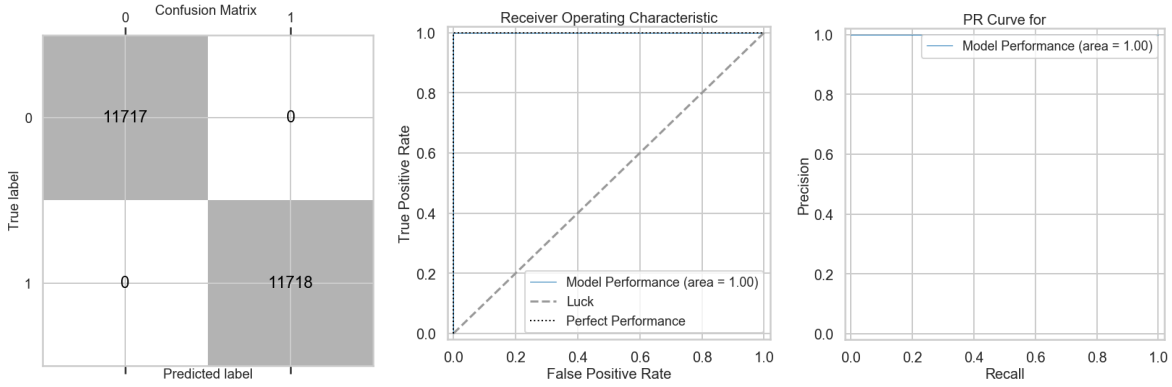

Figure 19 Gradient Boosting and Confusion Matrix ,ROC Curve and PR Curve

5.9 Model Comparison

Gradient Boosting, Random Forest Logistic Regression KNN Neighbors to build a model to predict flight cancellation. Based on testing the models on the holdout dataset (30% of the whole data), The results of various evaluation metrics scores are shown in Tab. 10 for all models.

Matrics	KNN Neighbors	Logistic Regression	Random Forest	Gradient Boosting Modelling
F1	0.99995	0.9964	0.99995	1
Cohen Kappa	0.99991	0.9930	0.99994	1
Brier	4.2671	0.0035	4.27	0.0
LogLoss	0.0015	0.012	N/A	N/A

Table 16 Model Comparison

Between them all, Gradient Boosting is the best model, after one hot encoder (pandas get dummies), sklearn standardscaler preprocessing process.

6. Conclusion

In data exploration part of this project, each column and its relationship with the Cancellation were analysed. The conclusion was that Majority of the flights were cancelled in the winter months and least in the spring and early fall months. Dallas has the most flight cancellation, that is caused by the weather. Other airports with stable weather conditions tend to have less cancellations. In airline perspective, American airline has the most cancellation and Delta has the least. So that, airline played an important role in determining the flight cancellation. For the days analysis, Monday is the day cancellation are more common. Lastly, weather condition vs cancellation was also analysed. Temperature lower than 40 degree F or higher than 100 degree F has more contribution to flight cancellation. Precipitation and clouds are another leading reasons to cancellations.

For modelling

Supervised machine learning and classification models are used for the flight cancellation like Gradient Boosting, Random Forest Logistic Regression KNN Neighbors. Data was split by 70% to train a predictive model and 30% to test the data.

After testing and tuning hyperparameters, Gradient Boosting was the best model to classify a set of data about probability of getting cancelled. ROC AUC to be about 1 for the Gradient Boosting.

7. Future Connection

In order to speed up the computation time, only 10% of the data is used for modelling. Rest of the data will be used since the best model have been selected. In addition, this project was limited to the top ten airports which had the most traffic in 2015. In future connections, more airports will be part of this modelling project.