NYC ‘13 Flight Dataset Analysis

Harshil Prajapati, Chirag Jadav
University of Maryland, Baltimore County
Hilltop Circle, Baltimore, MD 21250

(Dated: December 19, 2021)

Abstract

Around the globe, we often see many uncertainties in flights like its cancellation, arrival time and delay, departure time and delay, air_time, performance and its reliability. There are many factors which affect the schedule of various flights. In this project we are doing an analysis of the NYC 2013 Flight dataset. This flight data is about the flights that departed from NYC in 2013. Furthermore, we are only interested in the flights having DMV (Washington DC, Maryland and Virginia) destinations. Along with the flights dataset, there are other datasets for weather, planes, federal holidays and airlines information as well. With the help of these datasets we will carry out the flight analysis and build a predictive model to estimate the arrival delay and canceled flights.

1. INTRODUCTION

The issue of flight delay is very common and is experienced throughout the world. The cost associated with the same is always high and whenever the flight gets canceled or delayed, the airline company faces economic and reputation loss with the same. According to the FAA, the cost to the carrier operator for an hour of delay ranges from $1400 to $4500. If the value of passenger time is included, the cost is increased by $35 to $63 per hour. And weather consisted of 19% of the total delays. If we talk about NYC airports only (JFK, Newark and LaGuardia) consisted of the highest delay in the country having more than 57,000 delays of more than 15 minutes [1].

There are other datasets also to support giving the information about the planes, weather, federal holidays, airports and airlines. These datasets have common column data which link the other datasets with each other. We have carried out analysis of the flights with different attributes to find the trend in delays and cancellations. Data cleaning is done for better analysis purposes. The entire project has been done with the help of Jupyter notebook and Google Colaboratory. Libraries used in the project are matplotlib, seaborn, pandas, numpy, datetime, regex, for data manipulation and analysis whereas libraries like ‘sklearn’ to import and use machine learning features.

From the analysis of the data we try to build the predictive model to estimate the cancellations of flight from the test data set by implementing logistic regression. We also use a linear regression algorithm to train the model to predict the arrival delay of the test flight data.

2. DATASET

NYC13Flights dataset bundle includes a total of 9 datasets. This dataset includes the information about the flights that departed from NYC airports only and have different destinations. In that also, we are only interested in the flights having destinations Washington DC, Maryland and Virginia. They are:

1. flights2DCMDVA
2. airlines
3. airports
4. planes
5. weatherNYourly
6. weatherNYdaily
7. weatherMDdaily
8. Federal-holidays-2013
9. flights_test_data

The last dataset i.e. flights_test_data is used to predict the models for estimating the arrival delays of the flights and determine the canceled flights.

The dataset is very reliable and does not contain much absurd or false information. All the questions listed in the project are answered by analyzing the given dataset only. The datasets are loaded into Google Drive and Google Colaboratory is created for the python programming. This Colab notebook is shared between the team members to work on the project and analyze it.

New data frames are created throughout the project to meet the certain input requirements. These data frames are created by merging, grouping, updating, etc methods.

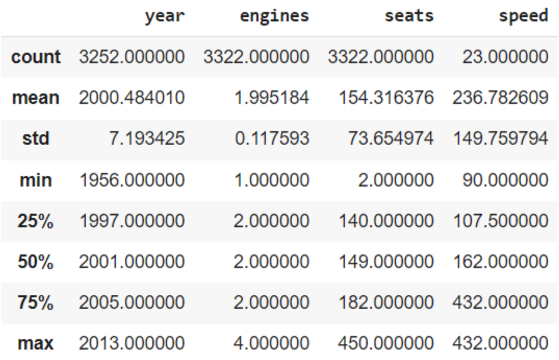
3. METHODOLOGY

EDA (Exploratory Data Analysis) is done on various datasets given to address the given questions. Different questions required different approaches to manipulate the data accordingly to carry out the analysis and build some insight from the same. Most of the analysis is done with the help of pandas library and plots are created with the help of matplotlib, seaborn and plotly.To address the question, we were required to perform operations like grouping, sorting, updating, merging of the dataset. The datetime data given in the dataset is of the object data type which was converted into *datetime[64]ns* datatype to better understand and analyze the trends as well as relation between the attributes.

This project can be divided into two parts. First part includes the analysis of the datasets. Question 1 to 3 is covered in this part. To determine the arrival delay for flights and airlines, impact of weather and federal holidays on delayed and canceled flights, insight on economic loss and reliability are addressed in this part of the project. Finding correlation between two datasets and plots to understand the patterns are performed in this part only. Second part of the project is predictive analysis where the test data is provided and we have to predict the arrival delay and determine the canceled flights from the test data. The estimation of the arrival delay is calculated by deploying a linear regression algorithm and canceled flights are predicted through logistic regression.

4. ANALYSIS

First we have loaded all the datasets into the dataframes in python to have a detailed look of the available data. Pandas library can handle the data frame operations well in python. Pandas are also compatible to use with various other libraries like numpy, matplotlib, seaborn, sklearn, re, etc. To get a brief idea about the dataset, we write ‘df.describe()’ in python cell, where df=data_frame. The example for the same is given below:



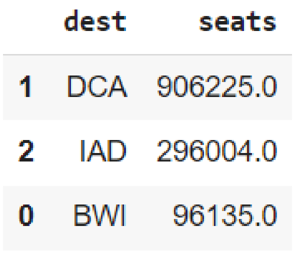
	year	engines	seats	speed
count	3252.000000	3322.000000	3322.000000	23.000000
mean	2000.484010	1.995184	154.316376	236.782609
std	7.193425	0.117593	73.654974	149.759794
min	1956.000000	1.000000	2.000000	90.000000
25%	1997.000000	2.000000	140.000000	107.500000
50%	2001.000000	2.000000	149.000000	162.000000
75%	2005.000000	2.000000	182.000000	432.000000
max	2013.000000	4.000000	450.000000	432.000000

Fig 1.1 (planes.describe() where planes is a dataframe)

This way we can get basic insight about the distribution of the data. After this we start cleaning the dataset to meet the analysis requirement.

A. Basic Analysis

All the questions in question 1 are designed to carry out the basic analysis of the dataset. The main aim of this question is to display the most basic statistics on the data and insights for the flights dataset. However, this question requires merging the planes dataset with the flights dataset in order to determine the number of seats, highest number of seats available on the particular day, etc. The question also includes the total number of all planned flights for each destination separately. This can be achieved by the ‘df.groupby()’ function of the pandas library. See Fig A.1 and A.2



	dest	seats
1	DCA	906225.0
2	IAD	296004.0
0	BWI	96135.0

Fig A.1 (df.groupby with count where df is dataframe)

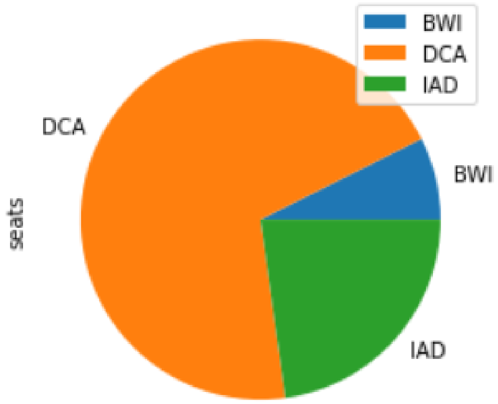


Fig A.2 (plot for the above result of Fig A.1. Distribution of total seats from all flights for each separate destination)

Other questions required combining the ‘year’,’month’ and ‘day’ columns to create a new column of date after converting it into datetime format. This conversion was done in order to determine the day with the highest number of flights. If we consider the day of week, then Wednesday had the highest number of flights but if we consider any particular date, then the day is January 11, 2013 and January 17, 2013.

The day with the highest number of seats available in 2013 for all the planned flights is February 28, 2013 which is on Thursday.

B. Analysis of Canceled Flights

There are many flights in the 'flights' dataset which are canceled. Although there is not a separate column given for the canceled flights, we determined the canceled flights by finding the NaN values in the dep_time column. All the scheduled flights have some departure time, so we assumed that flights having NaN values in the dep_time column would indicate the canceled flights. However, to make sure we returned the dataframe having null values in the dep_time column and noticed that along with dep_time, the values for arr_time, dep_delay, arr_delay, air_time were also NaN. This gave us concrete proof that if we clip the data for NaN values in dep_time we will get all the canceled flights. There were a total of 939 canceled flights in the flights dataset.

On our further analysis we figured that the day with the most cancellations of flights is March 06, 2013 with a total of 46 flights canceled on that day. The question is now to determine why the flights got canceled. We merged the weather dataset with the data frame created for the canceled flights, joining on date. We plotted several plots to find the trend or pattern that can explain the cancellation of flights. See Fig B.1 - B.5:

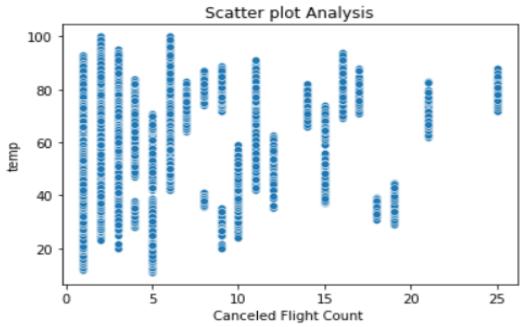


Fig B.1 (temp vs. Canceled flight count scatter plot)

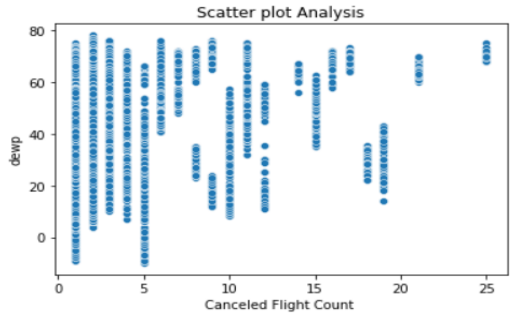


Fig B.2 (dewp vs. Canceled flight count scatter plot)

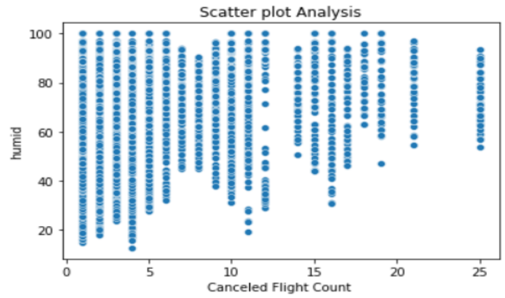


Fig B.3 (humidity vs. Canceled flight count scatter plot)

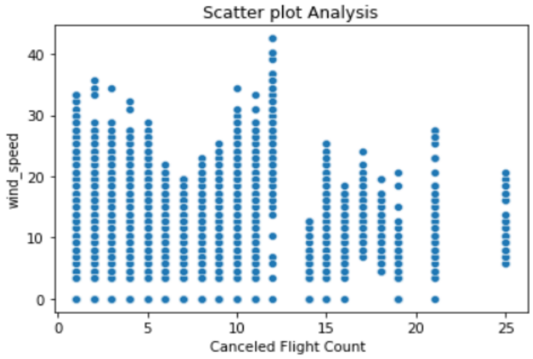


Fig B.4 (wind_speed vs. Canceled flight count scatter plot)

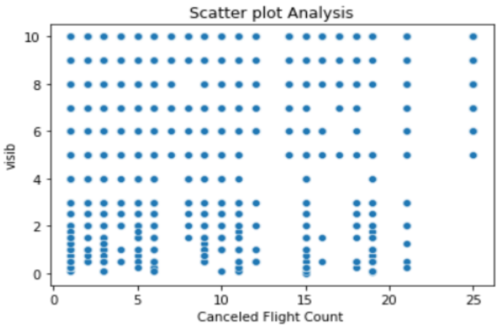


Fig B.5 (visibility vs. Canceled flight count scatter plot)

Studying the above plots from we did not find any linear trend between weather dataset and canceled flights. We were not able to find any concrete number to establish the relationship between the cancellation of flights with the weather dataset. However we can see from figure B.3 (humidity vs. Canceled flights count) that there were many canceled flights when the humidity was higher. Fig B.2 shows that dew point was also present for many flights that were canceled. However fig B.5 showcases that not enough visibility is absent for many canceled flights. This scatter plot analysis was unable to showcase any linear trend or pattern for weather dataset to affect the canceled flights. However, a few attributes like humidity, wind_speed showed some relationship which cannot be established or generalized.

The below correlation heatmap also shows the same result.

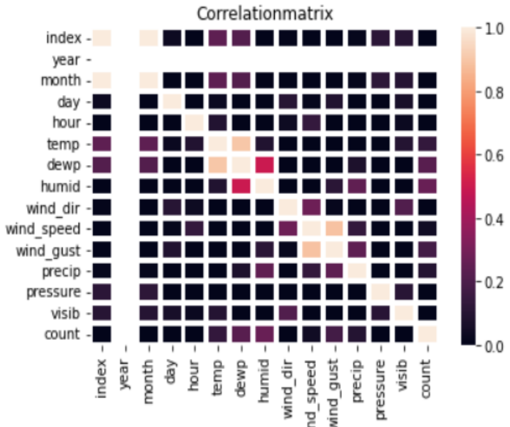


Fig B.6 (correlation heatmap of the merged dataframe)

Similarly we tried to find the trend between the canceled flights and federal holidays. By merging both the data frames and doing analysis on it we found that of all the holidays, the number of canceled flights on Labor Day was the highest. Total flights canceled on holidays were 12 of which 66.67% flights were canceled on Labor Day. If we consider the economic loss for each seat in a canceled flight is $50 then in 2013, total economic loss due to canceled flights was $1,201,600. On further analysis of determining the ratio of canceled flights to the planned flights for each airline company we deduced that there were canceled flights from DL, OO and UA airlines. Considering the reliability of the data accuracy we believe that above stated airlines are most reliable when it comes to canceled to planned flights ratio. The least reliable airline from this analysis is YV airlines.

C. Analysis of Flights Arrival Delay

Question 3 addresses the question of arrival delays of the flights. For this analysis we are not including the canceled flight so we created a dataframe for only planned flights.

We calculated the average arrival delay for each day and added the same values in the dataframe as a new column. We resample the data frame on the ‘date’ column in order to create a regular size interval of the datetime. This interval is used to plot the daily average arrival delay for a year. Average arrival delay determined is also added as a new column to the federal holidays data frame. This would allow us to plot the federal holidays on the same graph as of the daily average arrival delay. The plot is done with the help of the plotly module which has an on hover feature on the dataset. See Fig C.1

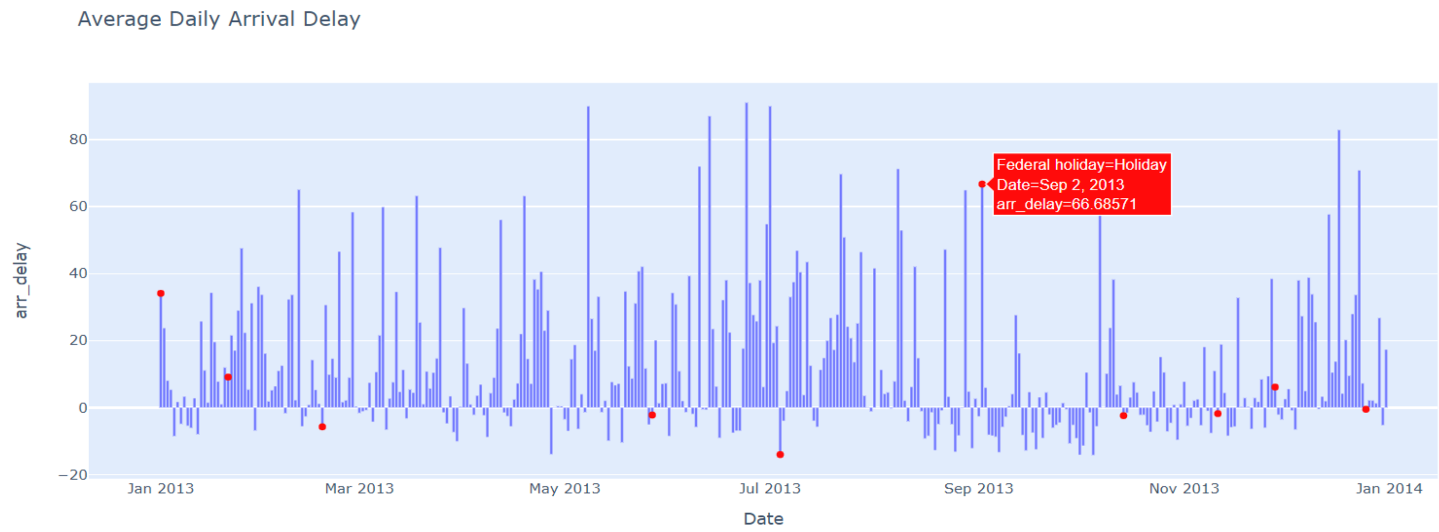


Fig C.1 (Daily Average Arrival Delay plot for 365 days having federal holidays marked on the same plot)

The hover feature for the map reflects the date and arrival delay data when the cursor is hovered over the desired date. From this plot we can see that the daily average arrival delay was higher from June end to August. More delays were experienced during that time of the year. On the other hand, less daily average delay was seen near the end of the year.

Amongst all federal holidays, most delays were seen on Labor Day. Average arrival delay for that was 66.6 minutes. The highest average arrival delay was on June 24, 2013 with a delay of 91 minutes, which is quite high. Negative arrival delay indicates that flight arrived before the expected arrival time. From this plot we can see that flights arrived earlier than the expected arrival time in September and October months. Though the daily average arrival delay is distributed throughout the year.

We also computed the correlation between the weather dataset and average arrival delay of the flights. We merged the weather and planned flights dataframe on date determined the correlation between the same. See Fig C.2 for the preference. We found out that the humidity is strongly correlated with daily average arrival delay whereas visibility has the weakest correlation. Attributes like air time, dew point, wind gust and precipitation shared positive correlation with the average arrival delay of the flights.

Fig C.3 shows the correlation value between the weather and planned flights dataset. There are many attributes which has negative correlation values.

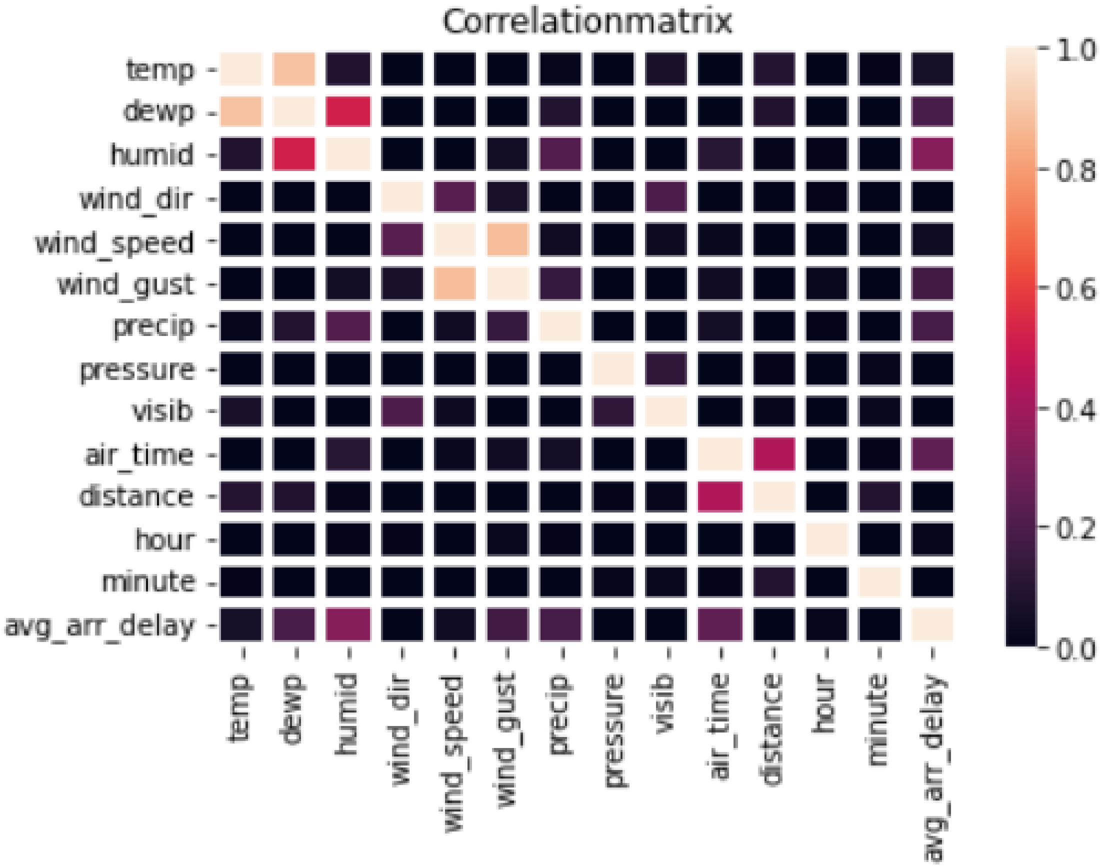
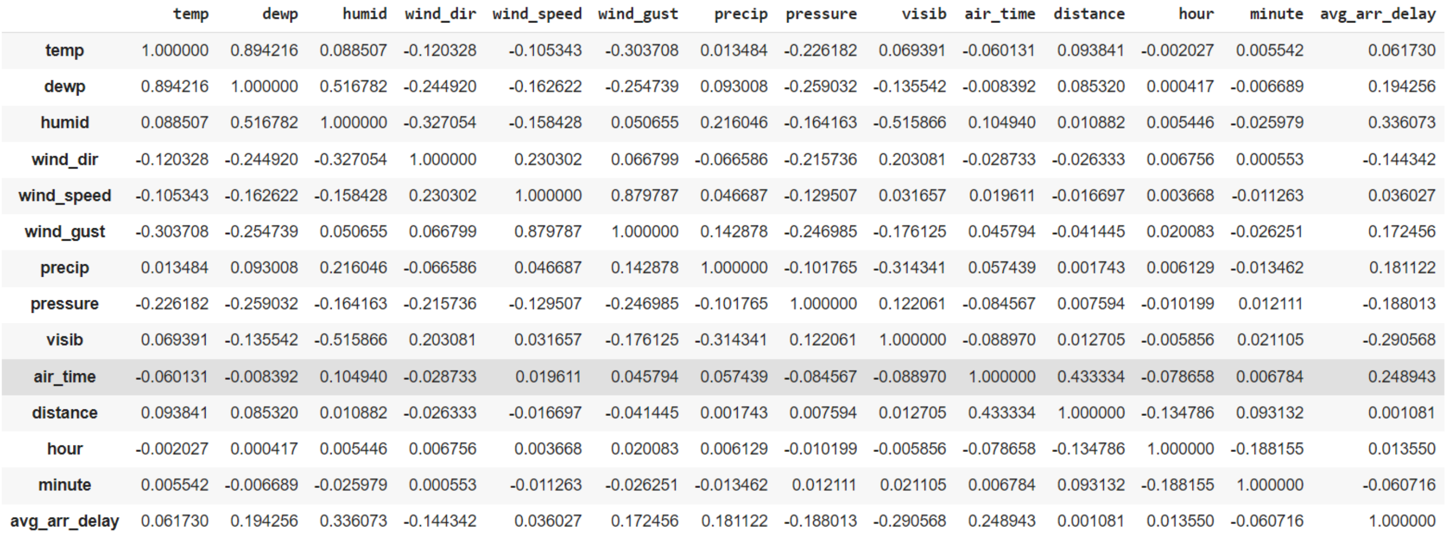


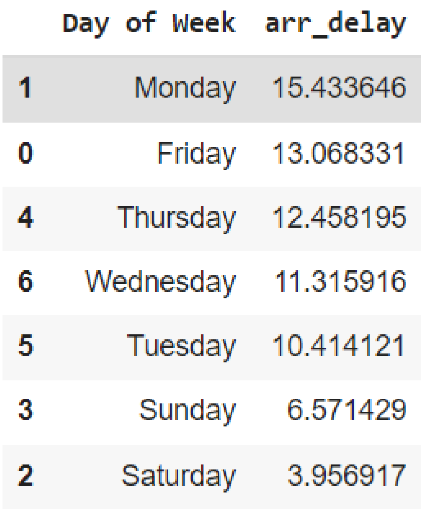
Fig C.2 (Correlation heatmap between weather and average delay)



	temp	dewp	humid	wind_dir	wind_speed	wind_gust	precip	pressure	visib	air_time	distance	hour	minute	avg_arr_delay
temp	1.000000	0.894216	0.088507	-0.120328	-0.105343	-0.303708	0.013484	-0.226182	0.069391	-0.060131	0.093841	-0.002027	0.005542	0.061730
dewp	0.894216	1.000000	0.516782	-0.244920	-0.162622	-0.254739	0.093008	-0.259032	-0.135542	-0.008392	0.085320	0.000417	-0.006689	0.194256
humid	0.088507	0.516782	1.000000	-0.327054	-0.158428	0.050655	0.216046	-0.164163	-0.515866	0.104940	0.010882	0.005446	-0.025979	0.336073
wind_dir	-0.120328	-0.244920	-0.327054	1.000000	0.230302	0.066799	-0.066586	-0.215736	0.203081	-0.028733	-0.026333	0.006756	0.000553	-0.144342
wind_speed	-0.105343	-0.162622	-0.158428	0.230302	1.000000	0.879787	0.046687	-0.129507	0.031657	0.019611	-0.016697	0.003668	-0.011263	0.036027
wind_gust	-0.303708	-0.254739	0.050655	0.066799	0.879787	1.000000	0.142878	-0.246985	-0.176125	0.045794	-0.041445	0.020083	-0.026251	0.172456
precip	0.013484	0.093008	0.216046	-0.066586	0.046687	0.142878	1.000000	-0.101765	-0.314341	0.057439	0.001743	0.006129	-0.013462	0.181122
pressure	-0.226182	-0.259032	-0.164163	-0.215736	-0.129507	-0.246985	-0.101765	1.000000	0.122061	-0.084567	0.007594	-0.010199	0.012111	-0.188013
visib	0.069391	-0.135542	-0.515866	0.203081	0.031657	-0.176125	-0.314341	0.122061	1.000000	-0.088970	0.012705	-0.005856	0.021105	-0.290568
air_time	-0.060131	-0.008392	0.104940	-0.028733	0.019611	0.045794	0.057439	-0.084567	-0.088970	1.000000	0.433334	-0.078658	0.006784	0.248943
distance	0.093841	0.085320	0.010882	-0.026333	-0.016697	-0.041445	0.001743	0.007594	0.012705	0.433334	1.000000	-0.134786	0.093132	0.001081
hour	-0.002027	0.000417	0.005446	0.006756	0.003668	0.020083	0.006129	-0.010199	-0.005856	-0.078658	-0.134786	1.000000	-0.188155	0.013550
minute	0.005542	-0.006689	-0.025979	0.000553	-0.011263	-0.026251	-0.013462	0.012111	0.021105	0.006784	0.093132	-0.188155	1.000000	-0.060716
avg_arr_delay	0.061730	0.194256	0.336073	-0.144342	0.036027	0.172456	0.181122	-0.188013	-0.290568	0.248943	0.001081	0.013550	-0.060716	1.000000

Fig C.3 (Correlation values between weather and average delay)

No concrete correlation was found between federal holidays and daily average arrival delay. Average arrival delay for each airport was also determined. DCA is the most reliable airport when it comes to daily average arrival delay, amongst IAD and BWI. The most reliable airline is DL, Delta Airlines when it comes to the daily average arrival delay. The least reliable is YV (Mesa Airlines Inc.). We also computed the average arrival delay for each day of the week and found out that Monday has the highest daily average arrival delay with the number of 15.43 minutes. See Fig C.4 for reference:

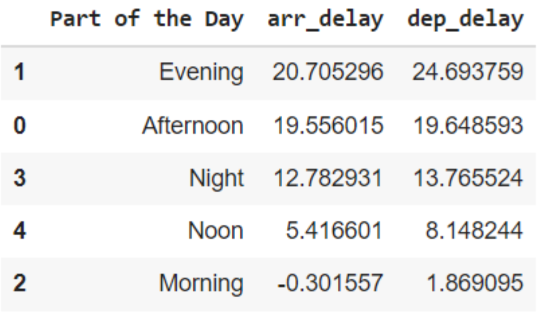


	Day of Week	arr_delay
1	Monday	15.433646
0	Friday	13.068331
4	Thursday	12.458195
6	Wednesday	11.315916
5	Tuesday	10.414121
3	Sunday	6.571429
2	Saturday	3.956917

Fig C.4 (Average Arrival Delay for each day of the week)

On further analysis, we created four groups of morning, noon, afternoon and evening to see which part of the day has the highest average arrival delay. We used a ‘groupby’ with a mean function to carry out the analysis. The result can be seen in Fig C.5.

The result is sorted from high to low. Along with the arrival delay, we have also calculated the departure delay as well for the same and for both the delays the trend is the same. Evening has the highest average arrival and departure delay whereas morning has the lowest.



	Part of the Day	arr_delay	dep_delay
1	Evening	20.705296	24.693759
0	Afternoon	19.556015	19.648593
3	Night	12.782931	13.765524
4	Noon	5.416601	8.148244
2	Morning	-0.301557	1.869095

Fig C.5 (Average Arrival and Departure Delay for each part of the day)

To address the last question we counted the number of airplanes used in these flights which were manufactured by BOEING, AIRBUS and EMBRAER. From all the three, EMBRAER has the highest number of airplanes with a count of 4409.

4. PREDICTIONS

The analysis brought us the basic insight about the datasets. We found correlation between two dataset and understood the trend or pattern for various attributes like daily average delay, cancellation sequence with the help of them. We are given a new dataset for question number 4 and 5. The main of these questions is to estimate the arrival delay for the given test data set and find the canceled flights from the same. Flights_test_data contain 19 rows having different data for each flight. The columns are year, month, day, carrier, origin, distance.

First we create the prediction model with the help of linear regression algorithm by importing the same from sklearn library. We have fed multiple features to X for training the model and get arrival delay in Y of the model. However when we first built the regression model, we saw that there are only four columns (year, month, day, distance) that are being used to predict the arrival delay. As the linear regression model training does not allow string values to be trained in the model, we had to remove all the columns having string values (like origin, destination). We are also getting error processing the date column as well. So we checked the distribution of the data through the time of year and created a plot to get an idea how the data is distributed. See Fig 4.1 - 4.4

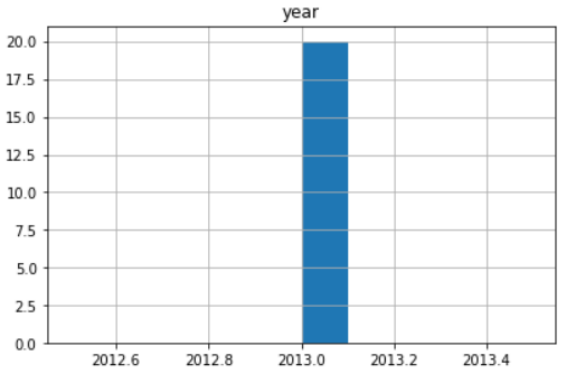


Fig 4.1 (Distribution of the data of the year column in test data)

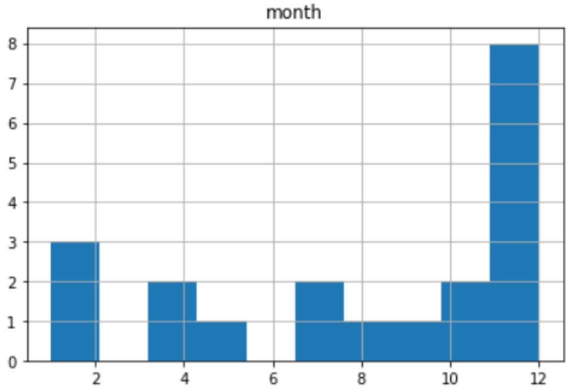


Fig 4.2 (Distribution of the data of the month column in test data)

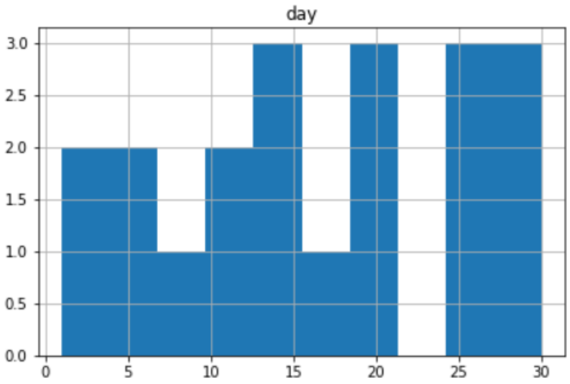


Fig 4.2 (Distribution of the data of the day column in test data)

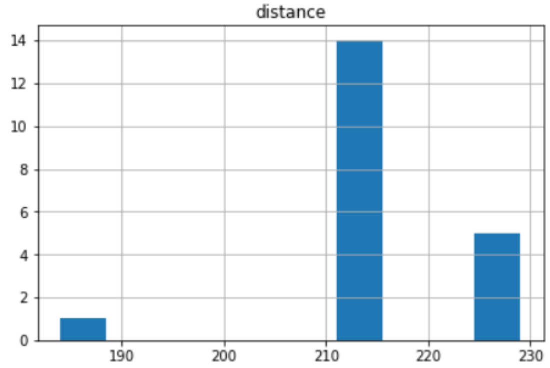


Fig 4.2 (Distribution of the data of the distance column in test data)

This plot shows the data is distributed in a biased manner and we had a doubt whether creating a prediction model based on only these four features will give us a better accuracy model or not. When initially we created a model only giving these four features to the model and then passed the test data to estimate the delay, we were getting the r2 score of 0.0495. Which was quite low. The data was very biased here and the model cannot make predictions with just these four columns. When we did the average arrival delay analysis between the weather dataset and planned flights, we found a strong positive correlation between the two dataframes. It will give us a better prediction model if we use the same features to train the model. So we merged the weather dataset with the planned flights, did some cleaning on it and computed a correlation to check whether the feature selection is good enough for predicting the estimated arrival delay or not. See Fig 4.1 for the heatmap of the correlation. This correlation heatmap shows that there are many features which have strong positive correlation between weather dataset and planned flights arrival delay. So selecting this feature to train the model will yield in to better prediction model. Based on these features we trained the model and we got a r2 score of 0.90, which is quite good compared to the previous scenario.

After this we fed the test data to the model to get the accuracy model but we faced the error that the dimensions of the train data and test data are not similar. This is because we were taking the flight test data having four columns only to deploy in prediction

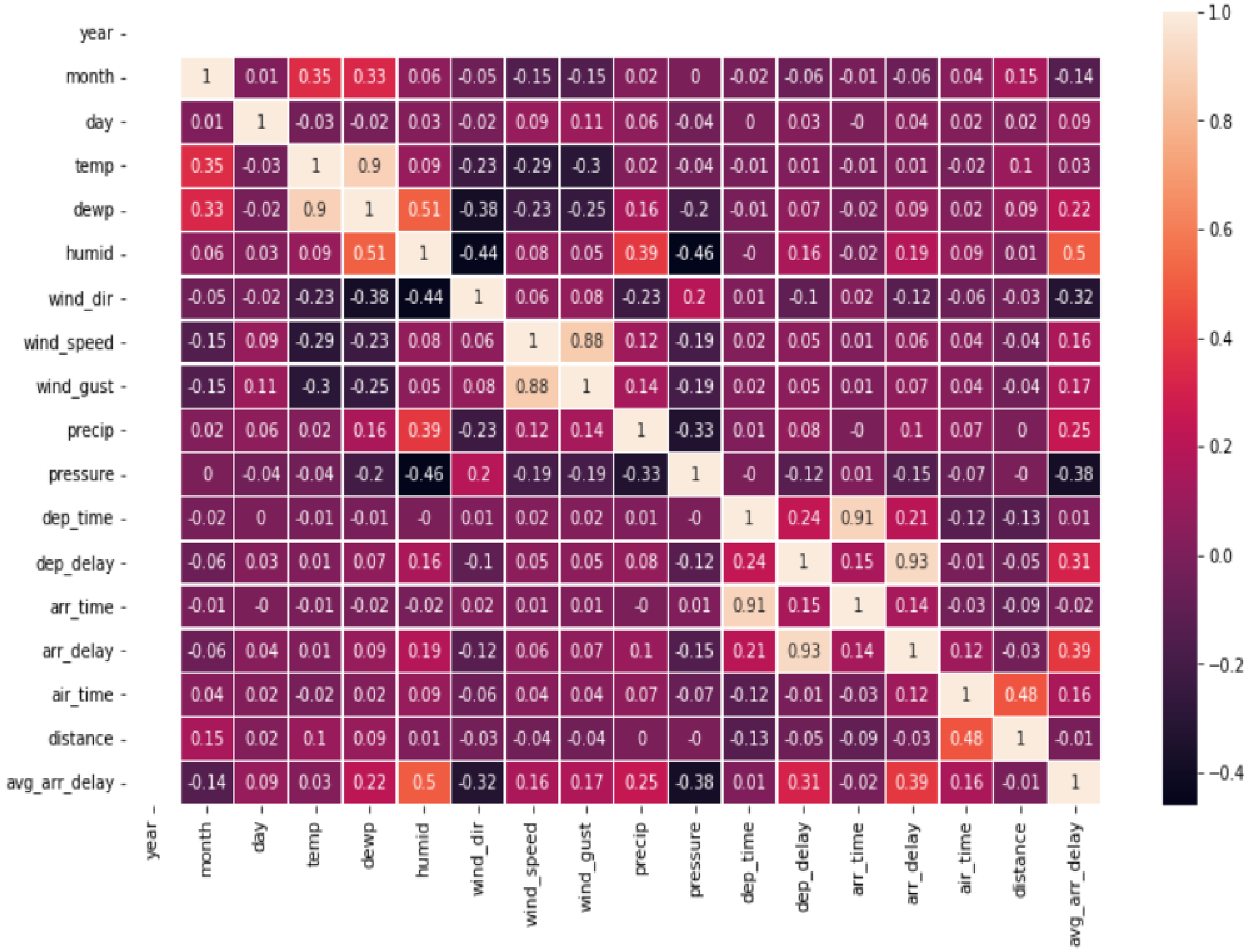


Fig 4.5 (Correlation heatmap of the merged data frame for feature selection)

model. So we decided to merge test data with the data frame created to train the prediction model. With this step we were able to create a data frame for test data having the same shape as of train data. We merged the data with the join condition on the 'date' column to only take the data from the original flight test data. Then we used group by function to get the same number of rows as the original flight test data so that the estimation can be done with each date mentioned in the flight test data. However on groupby, there were 4 columns which were showing null values which we replaced with the mean values so that we do not compromise the model accuracy. Proceeding this, we fed the test data to the model and were able to predict the arrival delay of the given flights. We created a new column of the 'estimated arrival delay' in the flight test data and printed the delay for each record.

Hence by selecting the features that had strong positive correlation between weather and planned flights data, we were able to build a prediction model having 90% accuracy. Based on that accuracy, the estimated arrival delay shows confident values.

Now the aim is to predict the canceled flights from the flight test dataset. Here we deploy a logistic regression algorithm to build the prediction model. Here the output is in objective (canceled flights, planned flights). For this kind of prediction logistic regression is a better algorithm to work on.
Let's understand the basic of logistic regression:

Output = 0 or 1
Hypothesis → $Z = WX + B$
$h\Theta(x) = \text{sigmoid}(Z)$

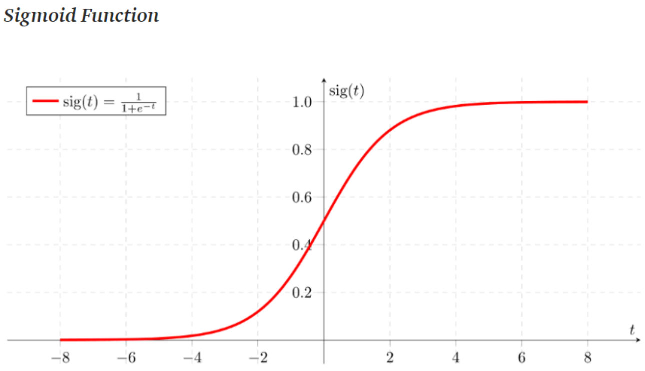
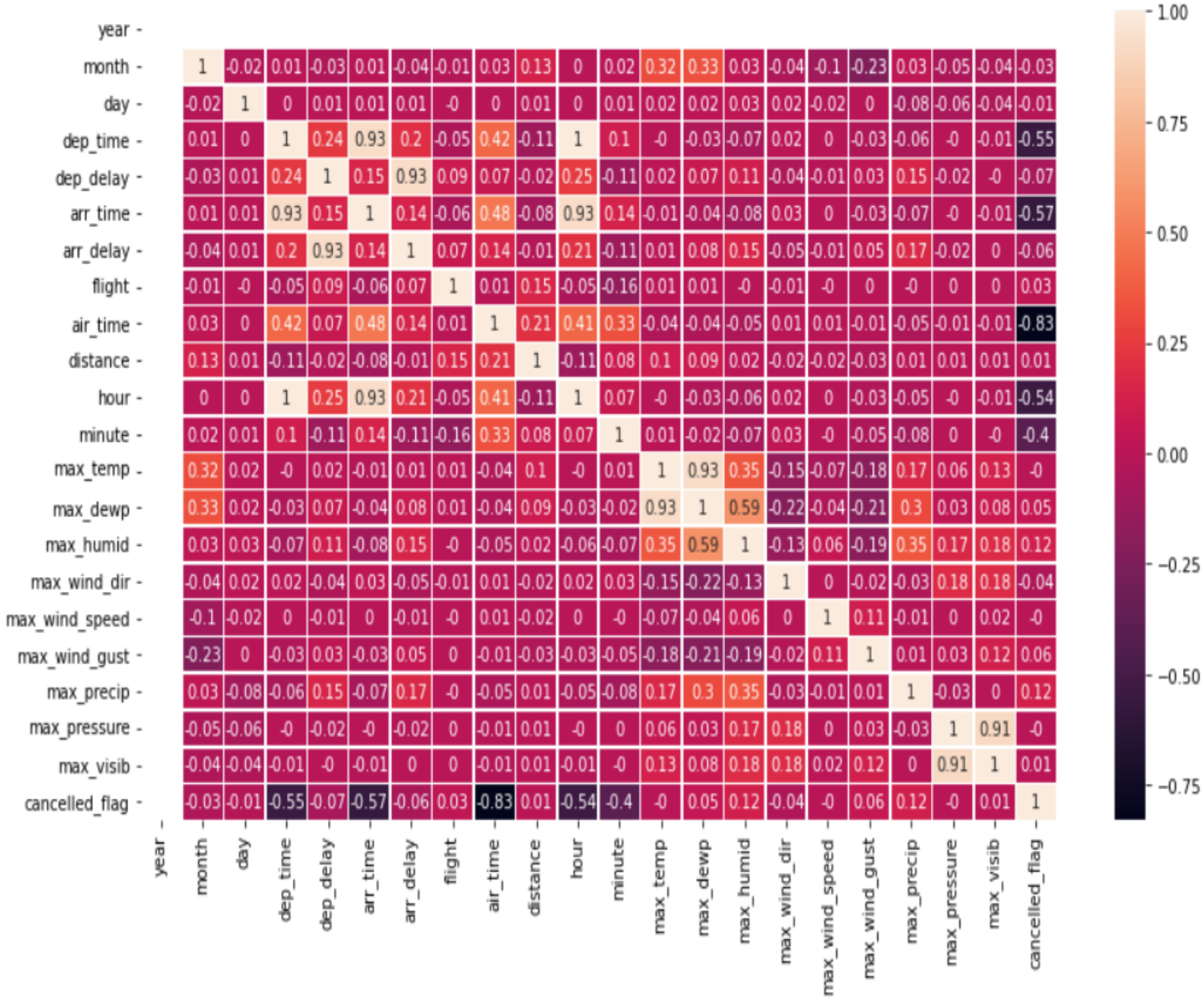


Fig 4.6 (Sigmoid Activation Function)

If 'Z' goes to infinity, Y(predicted) will become 1 and if 'Z' goes to negative infinity, Y(predicted) will become 0. [2]

We created a new dataframe by merging the weather dataset with the flights (including planned and canceled flights) dataset by making a left join in order to get all the data from the weather dataset. Also we have used the aggregation function to count the maximum values for each feature in order to find the correlation between the two dataset for feature selection of the prediction model. From this correlation we figured that there is a strong correlation on wind_gust, temp and wind_speed for the canceled flights. Similarly we prepared the test flight data set so that it has the same number of columns when fed into the prediction model. You can see the correlation heatmap in Fig 4.7 how different attributes are correlated.



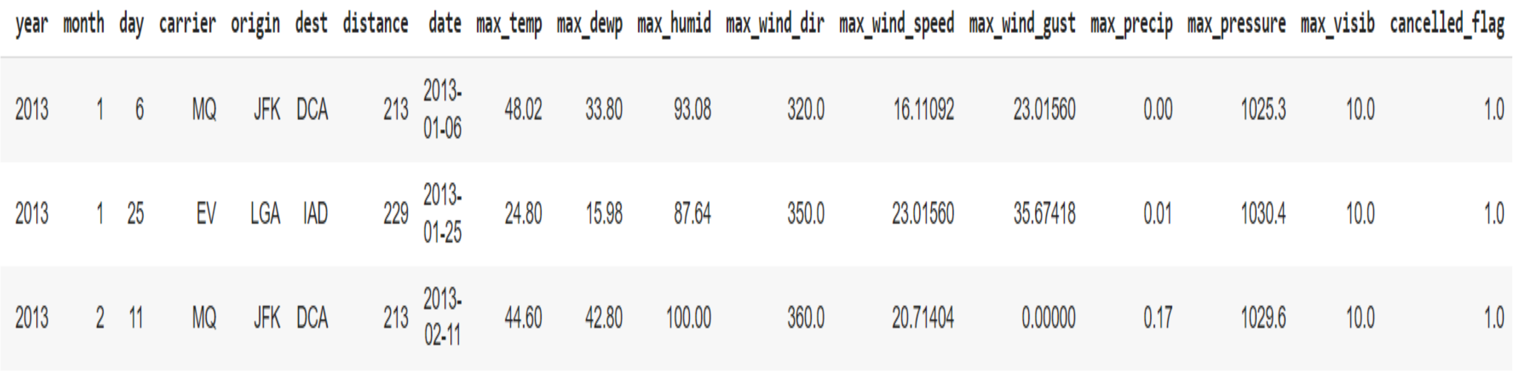
	year	month	day	dep_time	dep_delay	arr_time	arr_delay	flight	air_time	distance	hour	minute	max_temp	max_dewp	max_humid	max_wind_dir	max_wind_speed	max_wind_gust	max_precip	max_pressure	max_visib	cancelled_flag
year																						
month		1	-0.02	0.01	-0.03	0.01	-0.04	-0.01	0.03	0.13	0	0.02	0.32	0.33	0.03	-0.04	-0.1	-0.23	0.03	-0.05	-0.04	-0.03
day		-0.02	1	0	0.01	0.01	0.01	-0	0	0.01	0	0.01	0.02	0.02	0.03	0.02	-0.02	0	-0.08	-0.06	-0.04	-0.01
dep_time		0.01	0	1	0.24	0.93	0.2	-0.05	0.42	-0.11	1	0.1	-0	-0.03	-0.07	0.02	0	-0.03	-0.06	-0	-0.01	-0.55
dep_delay		-0.03	0.01	0.24	1	0.15	0.93	0.09	0.07	-0.02	0.25	-0.11	0.02	0.07	0.11	-0.04	-0.01	0.03	0.15	-0.02	-0	-0.07
arr_time		0.01	0.01	0.93	0.15	1	0.14	-0.06	0.48	-0.08	0.93	0.14	-0.01	-0.04	-0.08	0.03	0	-0.03	-0.07	-0	-0.01	-0.57
arr_delay		-0.04	0.01	0.2	0.93	0.14	1	0.07	0.14	-0.01	0.21	-0.11	0.01	0.08	0.15	-0.05	-0.01	0.05	0.17	-0.02	0	-0.06
flight		-0.01	-0	-0.05	0.09	-0.06	0.07	1	0.01	0.15	-0.05	-0.16	0.01	0.01	-0	-0.01	-0	0	-0	0	0	0.03
air_time		0.03	0	0.42	0.07	0.48	0.14	0.01	1	0.21	0.41	0.33	-0.04	-0.04	-0.05	0.01	0.01	-0.01	-0.05	-0.01	-0.01	-0.83
distance		0.13	0.01	-0.11	-0.02	-0.08	-0.01	0.15	0.21	1	-0.11	0.08	0.1	0.09	0.02	-0.02	-0.02	-0.03	0.01	0.01	0.01	0.01
hour		0	0	1	0.25	0.93	0.21	-0.05	0.41	-0.11	1	0.07	-0	-0.03	-0.06	0.02	0	-0.03	-0.05	-0	-0.01	-0.54
minute		0.02	0.01	0.1	-0.11	0.14	-0.11	-0.16	0.33	0.08	0.07	1	0.01	-0.02	-0.07	0.03	-0	-0.05	-0.08	0	-0	-0.4
max_temp		0.32	0.02	-0	0.02	-0.01	0.01	0.01	-0.04	0.1	-0	0.01	1	0.93	0.35	-0.15	-0.07	-0.18	0.17	0.06	0.13	-0
max_dewp		0.33	0.02	-0.03	0.07	-0.04	0.08	0.01	-0.04	0.09	-0.03	-0.02	0.93	1	0.59	-0.22	-0.04	-0.21	0.3	0.03	0.08	0.05
max_humid		0.03	0.03	-0.07	0.11	-0.08	0.15	-0	-0.05	0.02	-0.06	-0.07	0.35	0.59	1	-0.13	0.06	-0.19	0.35	0.17	0.18	0.12
max_wind_dir		-0.04	0.02	0.02	-0.04	0.03	-0.05	-0.01	0.01	-0.02	0.02	0.03	-0.15	-0.22	-0.13	1	0	-0.02	-0.03	0.18	0.18	-0.04
max_wind_speed		-0.1	-0.02	0	-0.01	0	-0.01	-0	0.01	-0.02	0	-0	-0.07	-0.04	0.06	0	1	0.11	-0.01	0	0.02	-0
max_wind_gust		-0.23	0	-0.03	0.03	-0.03	0.05	0	-0.01	-0.03	-0.03	-0.05	-0.18	-0.21	-0.19	-0.02	0.11	1	0.01	0.03	0.12	0.06
max_precip		0.03	-0.08	-0.06	0.15	-0.07	0.17	-0	-0.05	0.01	-0.05	-0.08	0.17	0.3	0.35	-0.03	-0.01	0.01	1	-0.03	0	0.12
max_pressure		-0.05	-0.06	-0	-0.02	-0	-0.02	0	-0.01	0.01	-0	0	0.06	0.03	0.17	0.18	0	0.03	-0.03	1	0.91	-0
max_visib		-0.04	-0.04	-0.01	-0	-0.01	0	0	-0.01	0.01	-0.01	-0	0.13	0.08	0.18	0.18	0.02	0.12	0	0.91	1	0.01
cancelled_flag		-0.03	-0.01	-0.55	-0.07	-0.57	-0.06	0.03	-0.83	0.01	-0.54	-0.4	-0	0.05	0.12	-0.04	-0	0.06	0.12	-0	0.01	1

Fig 4.7 (Correlation heatmap for the merged data of weather and flights

We filtered the canceled flight from the flight dataset by finding the data having NaN values in dep_time. We created a new data frame for the same. Then we created a 'canceled_flag' which will return the same value as dep_time and added it as a column in the flight dataset. The purpose of this flag is to convert the object data in to binary format for the logistic regression model .If the dep_time has NaN values, then the canceled_flag will return 1 (which indicates canceled flight), otherwise it will return 0, meaning that flight is not canceled. Then data is cleaned for any other NaN values that exist in the dataframe before training it into the predictive model.

We imported logistic regression from sklearn library and built the model from the selected feature where we get the accuracy of 50%. But we were skeptical about this. So we are using the SMOT function importing from the imblearn library which will nullify the biased data for the model and distribute data evenly. So the data is divided into a 50-50 equal ratio. After this we got the estimation for the canceled flights and confirmed flights. See Fig 4.8 for the flights which are estimated as a canceled flight when test data is deployed under the build linear regression model.

The accuracy of the model increased from the previous scenario and the improved accuracy is 66%. The Sensitivity and Specificity score is 62% and 72% respectively. This would be considered an average model and not a very good prediction model.



year	month	day	carrier	origin	dest	distance	date	max_temp	max_dewp	max_humid	max_wind_dir	max_wind_speed	max_wind_gust	max_precip	max_pressure	max_visib	cancelled_flag
2013	1	6	MQ	JFK	DCA	213	2013-01-06	48.02	33.80	93.08	320.0	16.11092	23.01560	0.00	1025.3	10.0	1.0
2013	1	25	EV	LGA	IAD	229	2013-01-25	24.80	15.98	87.64	350.0	23.01560	35.67418	0.01	1030.4	10.0	1.0
2013	2	11	MQ	JFK	DCA	213	2013-02-11	44.60	42.80	100.00	360.0	20.71404	0.00000	0.17	1029.6	10.0	1.0

Fig 4.8 (Estimated canceled flights of the test data set)

ACKNOWLEDGEMENTS

We would like to acknowledge our professor Dr. Ergun Simsek for providing us the dataset and guidance for this project. We have found great help from his lecture notes and concepts he delivered in the class.

References

[1] Federal Aviation Administration, *“FAQ: Weather Delay”*, Retrieved from https://www.faa.gov/nextgen/programs/weather/faq/

[2] Saishruthi Swaminathan. *“Logistic Regression - Detailed Overview"*, Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc