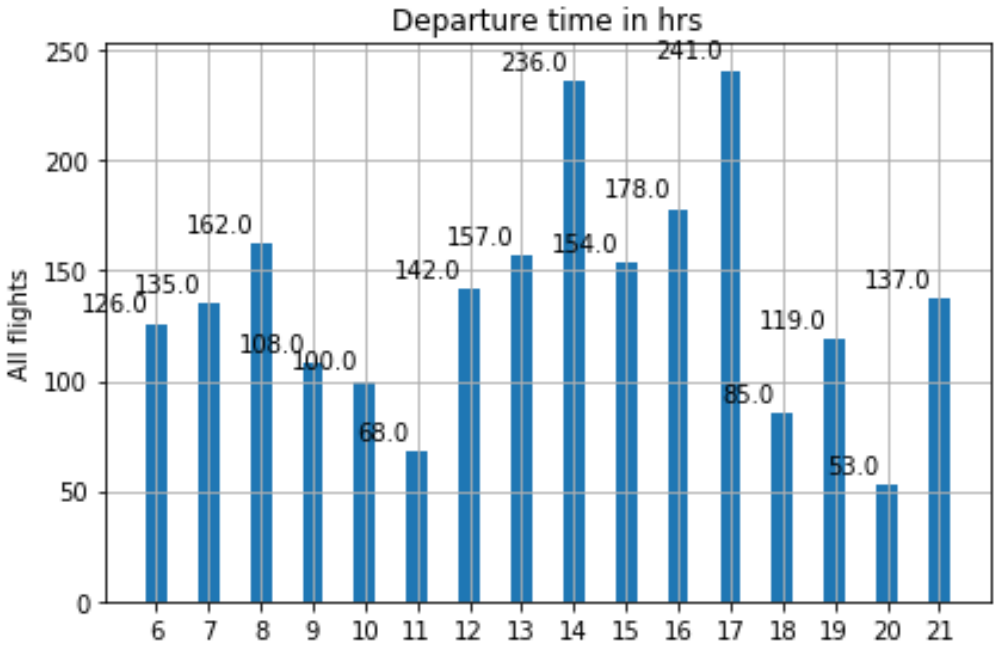
GNR 652 Assignment- Flight Delay Prediction

Samyak Shah, 18D070062

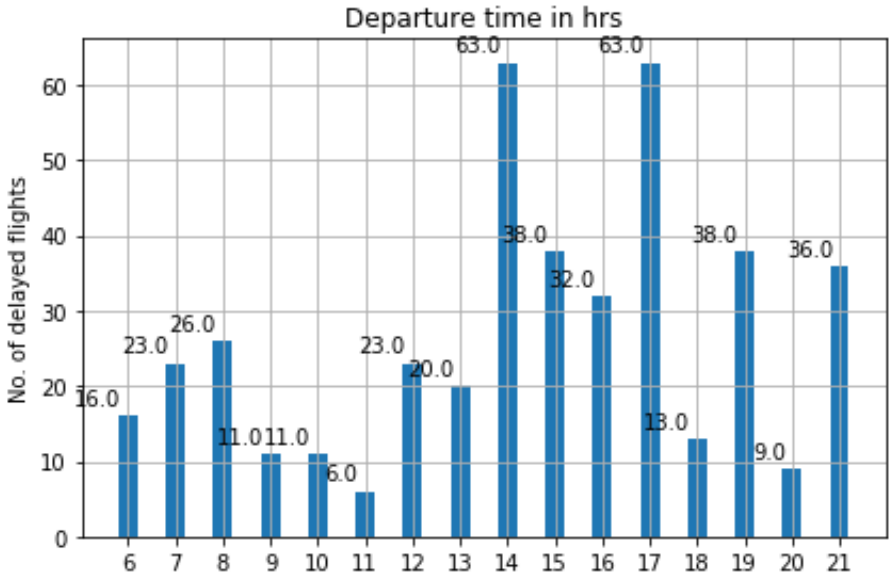
Compulsory Questions

Q1: In this part we plan to explore the data using a combination of line plot, bar graphs, double bar graphs, heat maps, box plots and pie charts. We try to find connection between the given features and the output and between two/three different features,

Plot 1:



Plot 2:



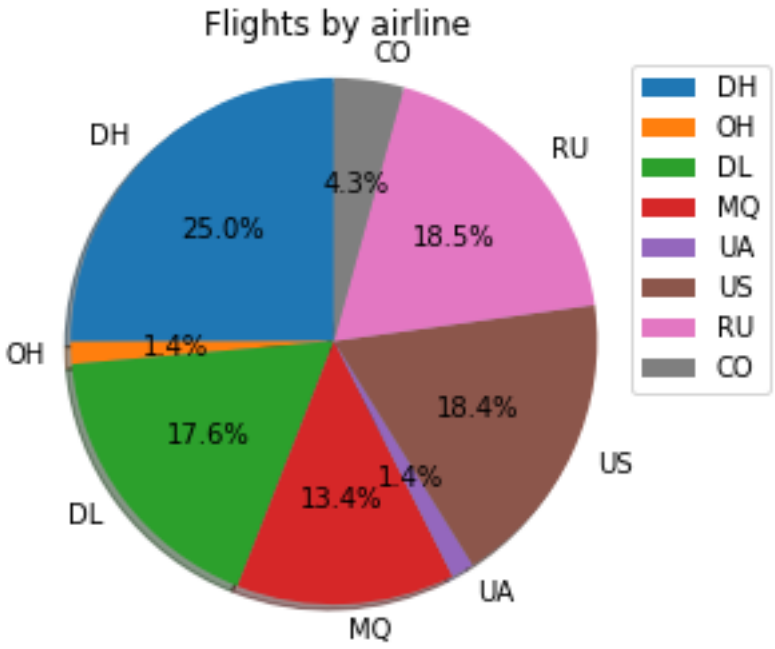
Generated output using Plot 1 and Plot 2:

The fraction is equal to the delay flights in the time interval divided by the total flights in that time interval.

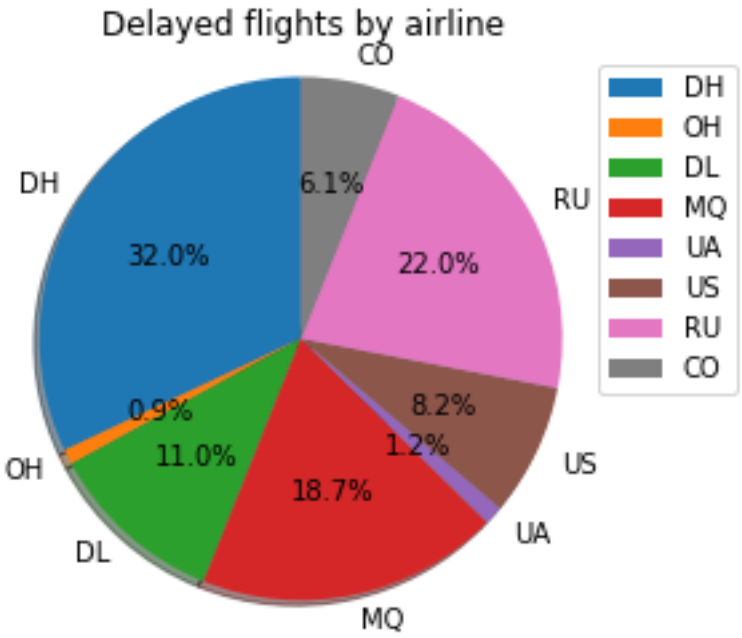
```
Fraction of flights between 600 hrs and 700 hrs delayed is 0.127
Fraction of flights between 700 hrs and 800 hrs delayed is 0.170
Fraction of flights between 800 hrs and 900 hrs delayed is 0.160
Fraction of flights between 900 hrs and 1000 hrs delayed is 0.102
Fraction of flights between 1000 hrs and 1100 hrs delayed is 0.110
Fraction of flights between 1100 hrs and 1200 hrs delayed is 0.088
Fraction of flights between 1200 hrs and 1300 hrs delayed is 0.162
Fraction of flights between 1300 hrs and 1400 hrs delayed is 0.127
Fraction of flights between 1400 hrs and 1500 hrs delayed is 0.267
Fraction of flights between 1500 hrs and 1600 hrs delayed is 0.247
Fraction of flights between 1600 hrs and 1700 hrs delayed is 0.180
Fraction of flights between 1700 hrs and 1800 hrs delayed is 0.261
Fraction of flights between 1800 hrs and 1900 hrs delayed is 0.153
Fraction of flights between 1900 hrs and 2000 hrs delayed is 0.319
Fraction of flights between 2000 hrs and 2100 hrs delayed is 0.170
Fraction of flights between 2100 hrs and 2200 hrs delayed is 0.263
```

The above signifies the lowest chances of delays in 11 am to 12 pm and highest chances of delays in 7 pm to 8 pm. This analysis is done on a single feature and may thus not always hold true in the entire model.

Plot 3:



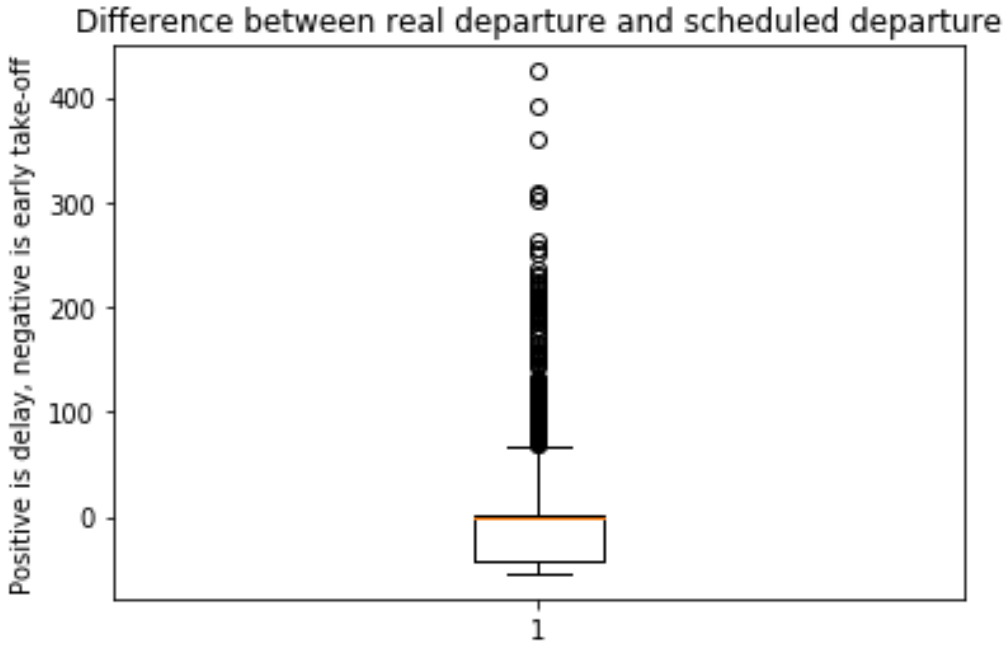
Plot 4:



If the pie size in plot 4 is larger than the corresponding pie size in plot 3 then the airline is more prone to delay. This includes: CO, RU, MQ, DH.

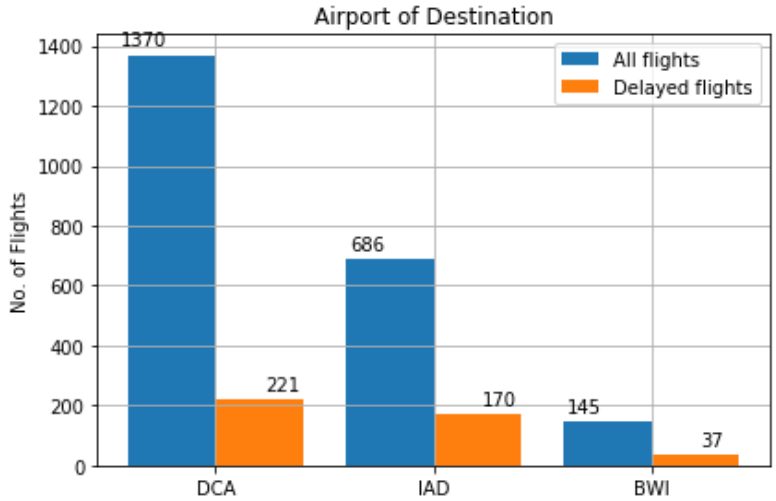
The other case where size of pie in Plot 4 is smaller than the size of pie in Plot 3. Then airline is less prone to delat. This includes: US, UA,OH,DL.

Plot 5:



The above graph shows the following trend: The number of flights with higher delays are decreasing in nature; delays are more common than early departures.

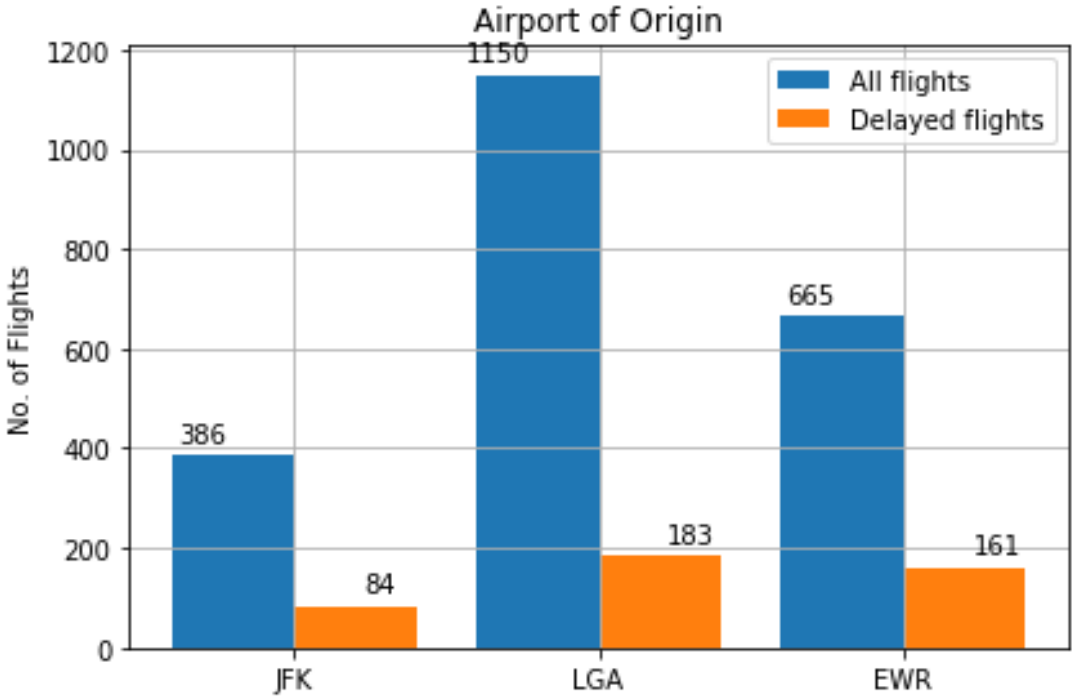
Plot 6:




```
Fraction of flights to DCA delayed is 0.161
Fraction of flights to IAD delayed is 0.248
Fraction of flights to BWI delayed is 0.255
```

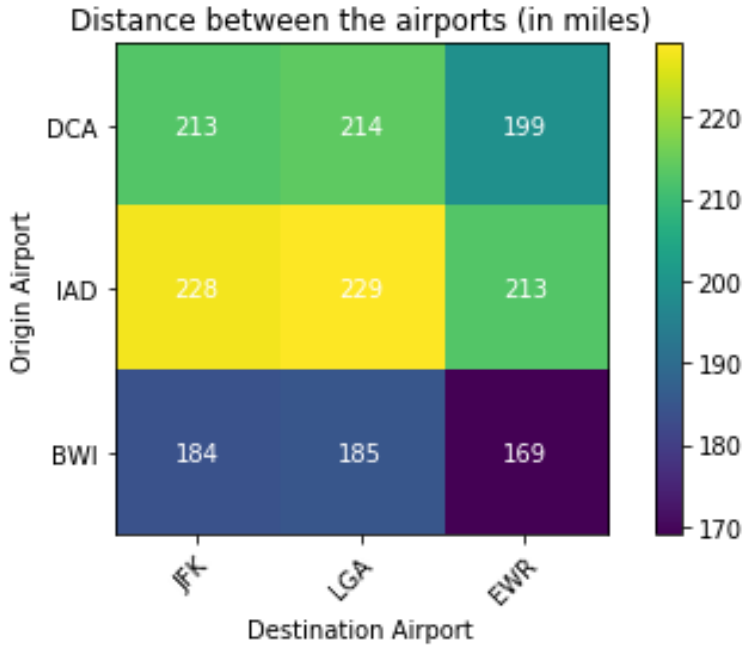
Chances of delays are max when flying to BWI and least when flying to DCA.

Plot 7:



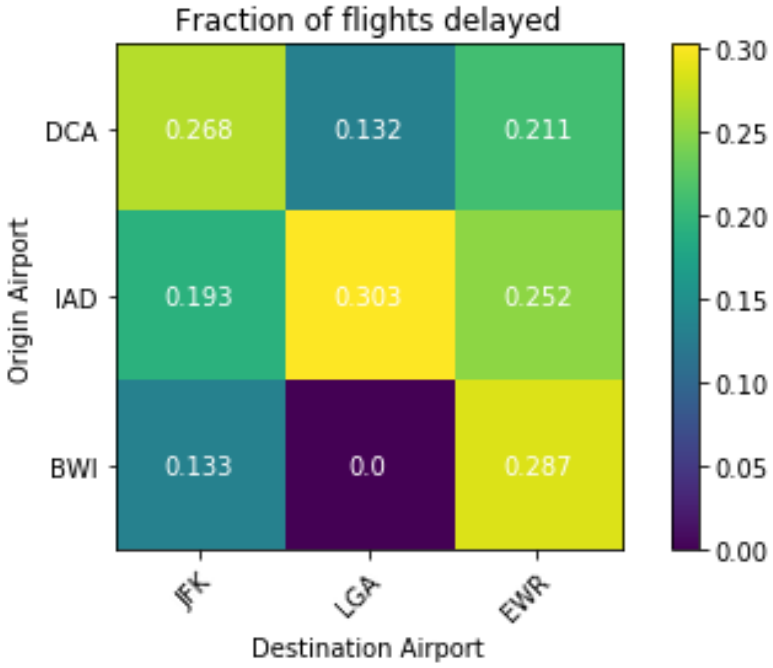
```
Fraction of flights from JFK delayed is 0.218
Fraction of flights from LGA delayed is 0.159
Fraction of flights from EWR delayed is 0.242
```

Plot 8:



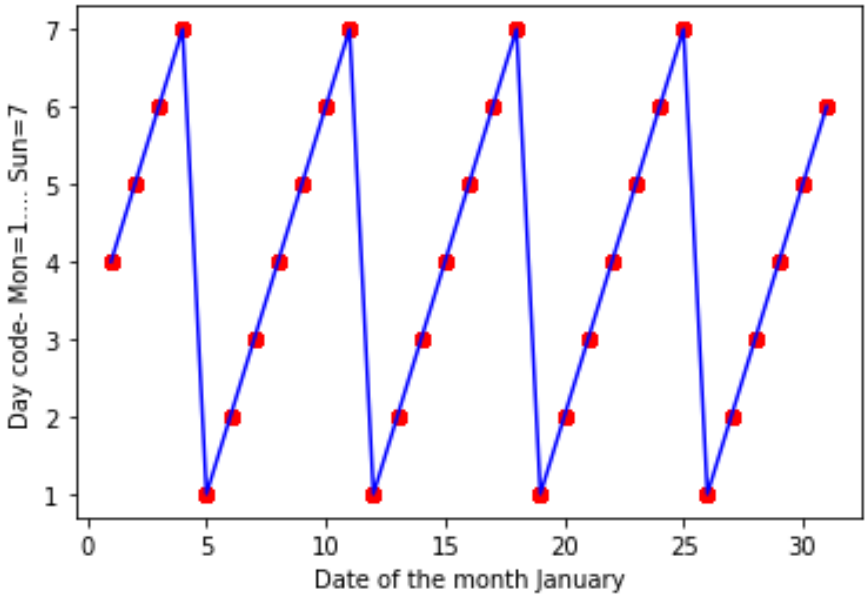
Distances between airports are fixed. The heat map shows the distance when flying between two airports of NY and Washington DC.

Plot 9:



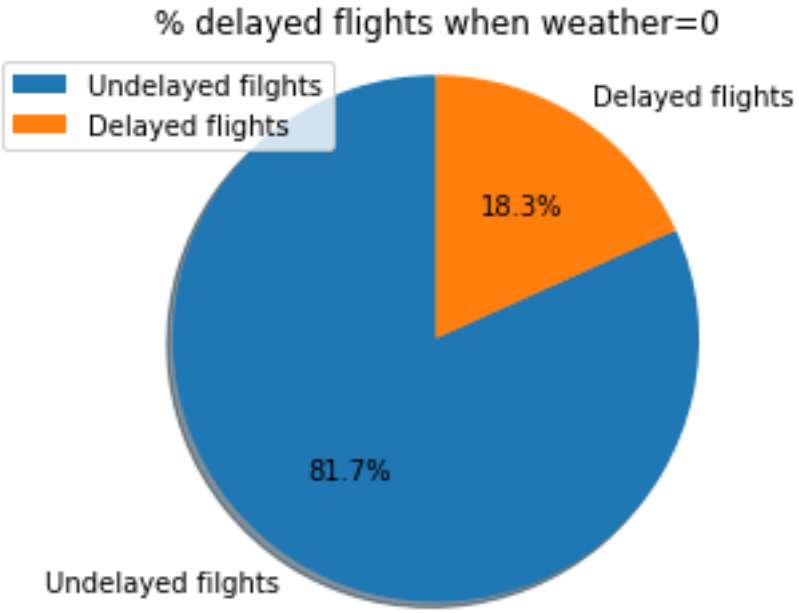
There was no flight in the data from BWI to LGA. Chances of delay are highest between IAD and LGA; lowest between BWI and JFK.

Plot 10:



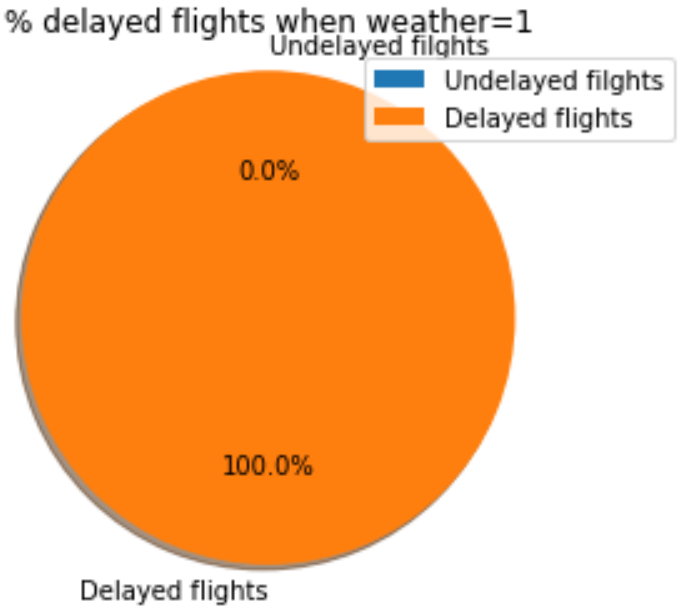
This plot shows the obvious relation between the day of the month and the date.

Plot 11:



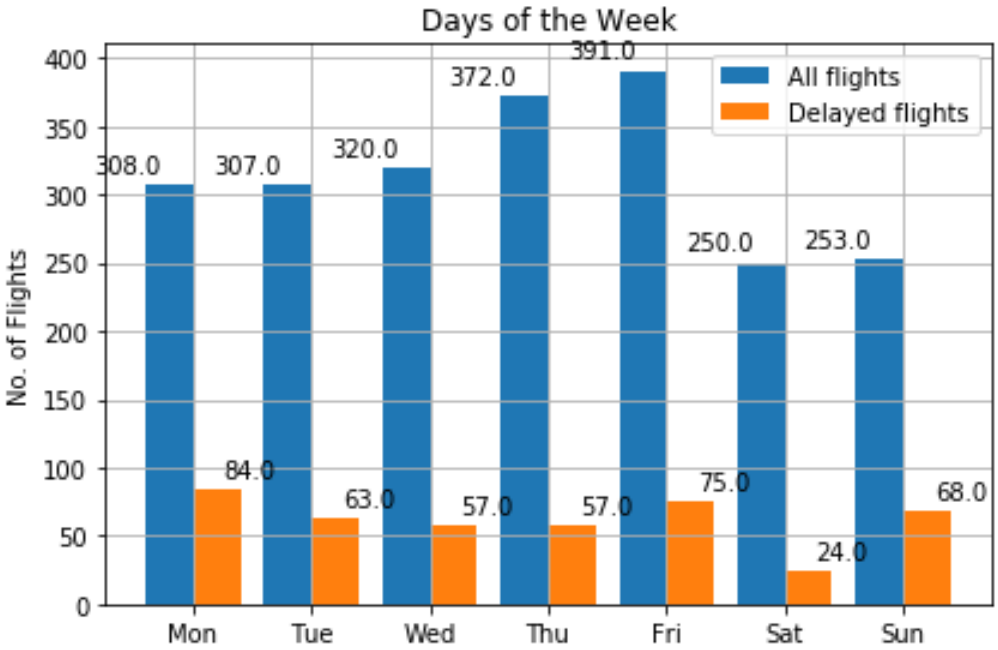
When the weather =0 , 18.3% flights are delayed.

Plot 12:



When weather=1 then all flights are delayed, it manifests that weather is 1 when the delay is weather related hence the delay is sure.

Plot 13:



```
Fraction of flights on Mon delayed is 0.273
Fraction of flights on Tue delayed is 0.205
Fraction of flights on Wed delayed is 0.178
Fraction of flights on Thu delayed is 0.153
Fraction of flights on Fri delayed is 0.192
Fraction of flights on Sat delayed is 0.096
Fraction of flights on Sun delayed is 0.269
```

Least chance of Delay is on Sat, max on Mon and Sun

Q2:

Pandas library was used for reading the given .csv file.

- There was no null value found.
- There were several columns with categorical data. So, we had to use one-hot encoding or assign dummy values to several features. This was realized using **pandas.get_dummies** function. The following features were converted using one-hot encoding.
    - CRS_Departure Time (after converting it into 16 intervals)
    - Departure Time (after converting it into 18 intervals)
    - Carrier airlines (8 airlines)
    - Flight Number (discrete entities)
    - Destination (3)
    - Origin (3)
    - Day of the week (Since the given 1 to 7 encoding is not right during computation as even if Sunday and Monday are adjacent days they are encoded further apart. The given 1 to 7 numbers did not follow the circular relations of days of the week.)
    - Date of the month (January) (Since days do not follow a linear pattern for delays in flights)
    - Tail Number (discrete entities)

Note: Weather was already as 0 and 1 so we didn't need to encode it further.

Since this is a crude approach and we did not perform variable selection yet, we have 744 features.

Using **sklearn.preprocessing – StandardScaler** we standardized the data because it is necessary to bring all feature values around a common center and of similar magnitude. This lead to lower accuracy but it is the right way to go. It is a necessary step before Logistic Regression

Using **sklearn.model_selection -- train_test_split** the given data was converted to training and testing data (60% for training and 40 % testing). I also used a random_seed to shuffle the data before splitting.

Using **sklearn.linear_model – LogisticRegression** we trained the model and tested it on the testing data. Since we have 2201 examples which is less for a Machine

Learning problem on Logistic Regression with 744 features, this leads to significant overfitting with respect to some variables. This leads to a drop in accuracy of our model.

Accuracy obtained with the seed value in the program is **83.42792281498298 %**

Q3: Interpretation of the model and the coefficients:

Since we standardized our features, we can interpret the importance of coefficients using their relative values. Otherwise, we couldn't have done that since the feature data values would have different centers and would have different magnitude orders too.

```
-0.1638688735743953
0.06160805890816023
0.03447007510873061
0.0
0.0
0.1482911784472151
0.0
0.0
0.031261920721256053
0.0
-0.16108108996685336
0.09683320909466961
0.0
0.0
0.05219526002127625
0.0
0.0
```

Above, is the snapshot of the coefficient values for our Logistic Regression model. We observe that our features are extremely high in number causing overfitting of data. Tail Number, Flight Number etc. have a high variety of categorical data and hence create a very sparse feature matrix.

The variance of the data in each of these dummy variable for Flight Number, Tail numbers is very low and a lot of their coefficients are 0.0 . This shows that these are minimally important for our model.

Least Important: Flight Number, Tail Number in general.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta^T X = \beta_1 X_1 + \ldots + \beta_p X_p.$$

$$p(X) = \frac{\exp(\beta^T X)}{1+\exp(\beta^T X)}$$

Using the above equations for the sigmoid function and the log probability odds we see that higher the magnitude of coefficients more will be its effect on the prediction.

Also it is important to note that the negative coefficients contribute to the delay of the flight while the positive coefficients contribute to the flight being on-time.

So the dummy variables of Tail_Num and Flight_Num will be dropped later as their coefficiencts have extremely small magnitude. Larger number of features restrict proper learning of models and hence give lower values of accuracy. In the next parts we expect higher values of accuracy when we perform variable selection

Q4:

Since the variance of some of the features is very low, they make no significant contribution we directly dropped the dummy variables of the following:

- Tail Number
- Flight Number
- Date of the Month (Date of January)

Now we are left with 61 features. We run **RFE** on our model to rank the features. RFE ranks the importance of the features from highest to lowest. Some of the higher ranked coefficients may be negative or positive. So being higher in RFE rank list only signifies the importance in prediction it does not tell us if it would contribute to the flight being delayed or on-time.

We then reduced our model to the top15 features from the 61 features based on RFE ranks. We observed a big jump in the accuracy in case of our model ith the top 15 features.

The top 15 most important features are:

Rank	Feature
1	Weather
2	Dept_Time_22.0
3	Dept_time_23.0
4	Dept_time_19.0
5	Dept_time_14.0
6	CRS_DEPT_TIME_11.0

7	Dept_time_11.0
8	Carrier_CO
9	Carrier_MQ
10	Origin_DCA
11	CRS_DEPT_TIME_12.0
12	Dept_time_20.0
13	Carrier_OH
14	Dep_time_5.0
15	Dep_Time_8.0

It was also observed that most of the Destination dummy variables were very low in the order of importance. We then also printed these coefficients in the rank order to find out if they are positive or negative.

Q5:

A new model was fit for the 15 features:

We get accuracy as: **89.21679909194097 %**

This is a jump of **5.78887627696%** which is large in classification problems in Machine Learning.

Please see the code for more details about the procedure.

Q6:

Some of the higher ranked coefficients may be negative or positive. So being higher in RFE rank list only signifies the importance in prediction it does not tell us if it would contribute to the flight being delayed or on-time.

What we seek for this problem is highest chance of ontime performance. Weather coefficient was negative so it's better to keep weather=0.

OH was the highest ranked airline with a positive coefficient. At Rank 6 CRS Departure time between 11 am and 12 noon is the highest ranked positive coefficient feature. For selecting the day, we used the coefficient values of our old model (Q2.). It reflected that Saturday had the least chance of delay.

So, for highest chance of on time performance we choose the following features

Criterion/ Feature	Feature Value
Weather	0 (since weather= 1 directly entails delay for sure)
Carrier	OH
Day	Saturday
Time	11 am to 12 pm

This selection agrees and is confirmed by our exploratory data analysis done in Q1.

This confirms the intuitiveness of the mathematics behind Logistic Regression.

Limitations: The given data had less examples and the examples with delayed examples are extremely few (20% of the total data). This may lesser accuracy and incorrect conclusions in some cases.

Bonus Questions

Note: Only a few questions were attempted from this section since the marks are capped to a maximum of 2.

Q1: H.E.L.E.N., V.I.R.G.I.L. are some of the other AIs created by Tony Stark in MCU.

Q2: The Data processing inequality is an information theoretic concept which states that the information content of a signal cannot be increased via a local physical operation. This can be expressed concisely as 'post-processing cannot increase information

Q3: X is the Rule of two.

Q4: The name of the robot on the left is R2D2. The other robot (golden) is C3PO.

Q5: In 2019 for Black Friday, the game developers taught a computer how to write Cards Against Humanity cards. It was put to the test. Over the next 16 hours, the writers battled this robust card-writing algorithm to see who could write the most popular new pack of cards. If the writers won, they'd get a $ 5000-holiday bonus. If the A.I. won, they would fire the writers. The writers won, and they have their jobs for a few more years.
