ASSIGNMENT -1

GNR 652 COURSE

PREDICTION OF FLIGHT DELAYS

Question No.1: - **Show visualizations to explore the dataset and understand the underlying trends (Often called Exploratory Data Analysis). Choose visualization methods you think best represent the data (bar graph, pie chart, scatter, boxplot, heatmap etc.)**

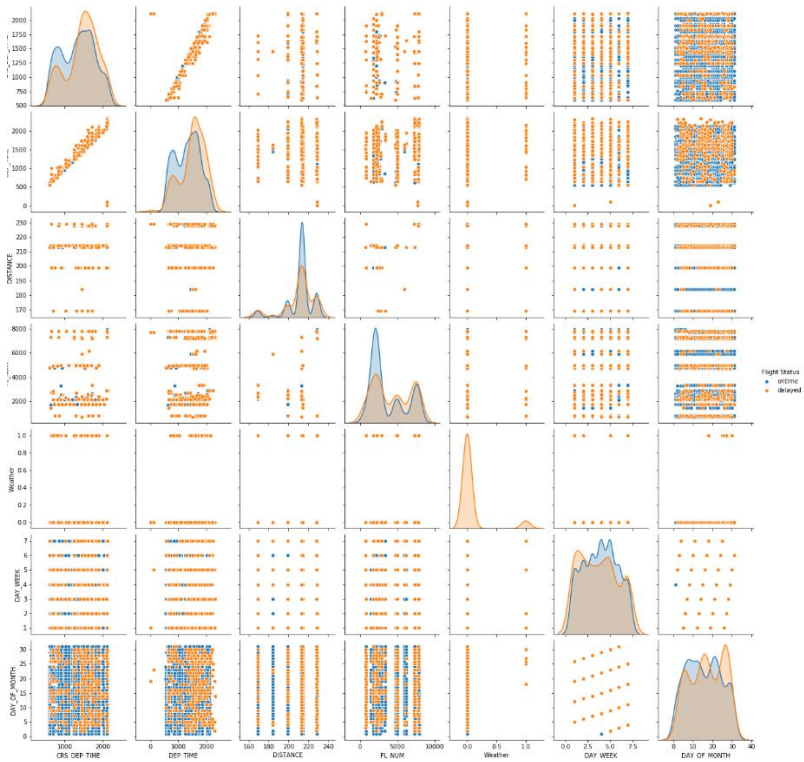
The given dataset has a total of 13 columns and 2201 data rows

	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	TAIL_NUM	Flight Status
0	1455	OH	1455	JFK	184	01/01/2004	5935	BWI	0	4	1	N940CA	ontime
1	1640	DH	1640	JFK	213	01/01/2004	6155	DCA	0	4	1	N405FJ	ontime
2	1245	DH	1245	LGA	229	01/01/2004	7208	IAD	0	4	1	N695BR	ontime
3	1715	DH	1709	LGA	229	01/01/2004	7215	IAD	0	4	1	N662BR	ontime
4	1039	DH	1035	LGA	229	01/01/2004	7792	IAD	0	4	1	N698BR	ontime

First of all, the check of null values is done, **no cell in dataset is null.** All columns have unique values which is false which was returned by function isnull() means there are no null values in our dataset.

	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	TAIL_NUM	Flight Status
count	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201
unique	1	1	1	1	1	1	1	1	1	1	1	1	1
top	False	False	False	False	False	False	False	False	False	False	False	False	False
freq	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201	2201

Also, to visualize the trend between categorical feature and flight status we have to convert Carrier, Origin, Destination and Flight status into dummies for modelling, plotting it as histogram we observe the following trends. For an example Tail No. has alphanumeric values and hence cannot be directly used in Logistic regression, Thus we apply Label encoder to it and convert them to integer features.



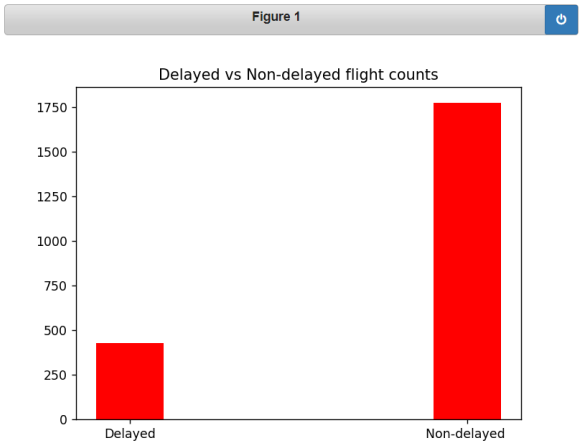
The pair-plot is plotted between in feature to understand the if there is any distribution between two features to classify between the on-time vs delayed status of flight. From above image we can see the

weather has not on-time flight data. Also, departure-time and CRS departure time has collinearity but ot classifying the data point according to on time status. Other than this is no well defined distribution of data point which can classify data by alone itself.

1. Flight status

We have data with 428 entries of delayed and 1773 entries of non-delayed, which is imbalanced

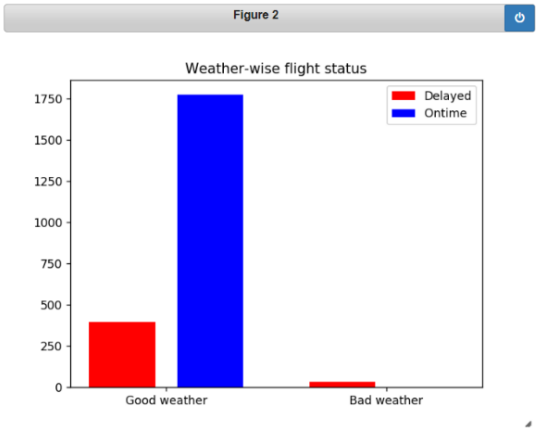
```
1    1773
0     428
Name: Flight Status, dtype: int64
```



2. Weather vs Flight Status

The number of flights of each delayed because of weather can be seen from below join bar graph. So, we can see that when weather is bad then all flights got delayed so weather can be seen as very important feature. But because very imbalanced distribution of data we might not able see the effect of the weather since delayed samples are itself are low

```
0    32
Name: Flight Status, dtype: int64
```

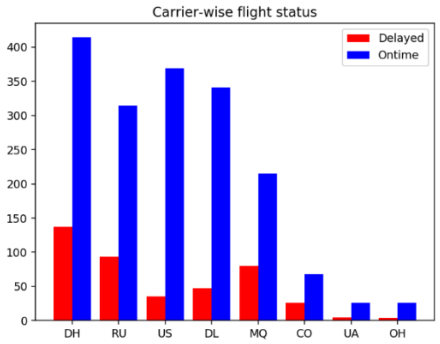


3. Carrier vs Flight Status

The number of flights of each delayed in case of Carrier can be seen from below join bar graph. So, we can see that when Carrier US are having most on time flights.

```
Percentage of ontime flights:
[75.14 76.96 91.34 87.89 72.88 72.34 83.87 86.67]
```

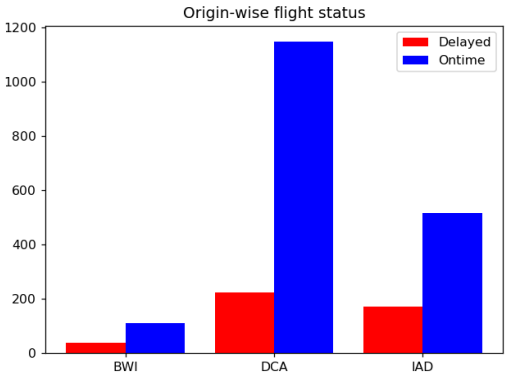
Figure 3



4. Origin and destination airports

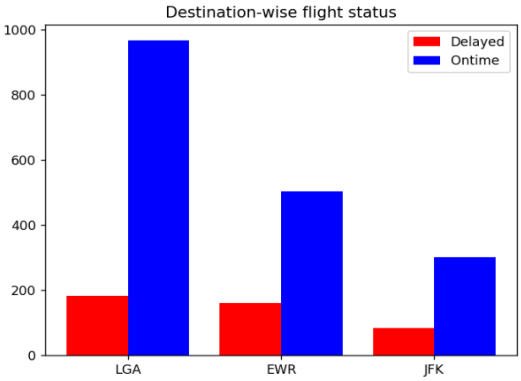
```
Percentage of ontime flights:
[74.48 83.87 75.22]
```

Figure 5




```
Percentage of ontime flights:
[84.09 75.79 78.24]
```

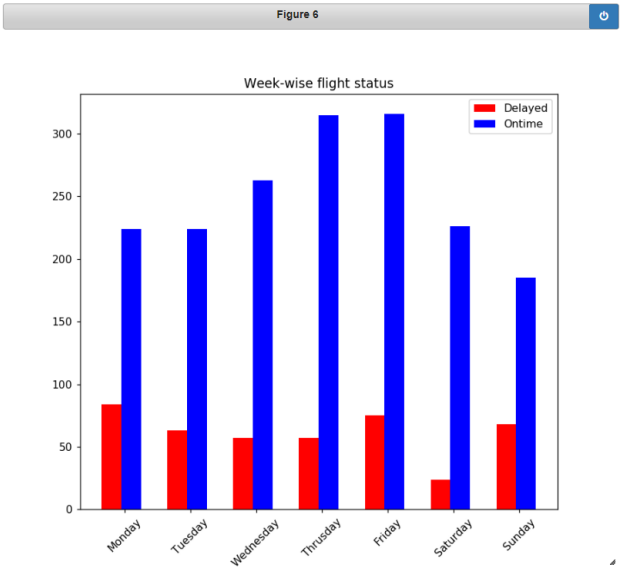
Figure 4



From above plots we can see that **LGA** is a major destination airport and **DCA** is a major origin airport. And the LGA and DCA both having hight percentage of on-time flight

5. Days of week vs Flight status

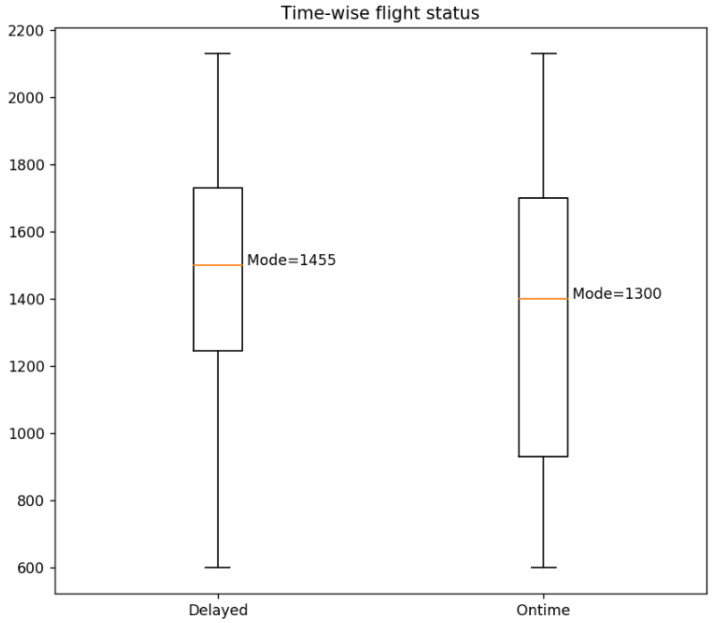
```
Percentage of ontime flights:
[72.73 78.05 82.19 84.68 80.82 90.4  73.12]
```



From above plots Saturday having hight percentage of on-time flight , so that is better day to fly

Next the Feature Flight date does not appear to be significant in predicting delays, This is **because** the day of month and day of week both are already present to account for capturing the variation with respect to dates, Hence it will be good to **drop the Flight date feature**.

6. CRS departure time vs Flight status



The above graph is plotted for the finding out the best time on which the flight is not delayed. Above graph show the max, min and mean of the time on which the flight is delayed and ontime the box plot ranges are overlapping. So mode shows the maximum flight (95) which are on time are CRS_DEPT_TIME 1300.

Question 02: -) Pre-process the dataset (to remove null values, generate dummy variables etc.) and divide the dataset into 60% train and 40% test. Prepare a logistic model that can obtain accurate classifications of new flights based on their predictor information.

we created **dummy variables** for nominal categorical features i.e., Flight status, Carrier, Origin and Destination. (in 23 columns total)
Dataset is randomly split into train and test samples in 60:40 ratio i.e., dataset is divided into x and y features and then applied on model for fitting.
The code is given in a Jupyter notebook named **ML_In_Flight_scheduling(Q_1, Q_2 & Q_3)**.
Algorithm used is Logistic Regression with no hyper-parameter tuning.
The accuracy with all columns included - `0.8978433598183881`

Question Q3) Interpret the model and coefficients and present some insights

The logistic regression is a non-linear binary classifier, it does classification based on the probability values given by model with help of default threshold of 0.50 for classifying the in particular class.
The coefficients given by the model are the weights of features used to predict the best fit hyper-plane in **log(odds)** space [i.e. **log(p/1-p)**] where magnitude of individual weights (slopes) indicate the importance of features(i.e. changing in values of one feature causing change in probability of sample), higher the weight more important is that feature in final classification because it having potential/deciding information to classify unseen sample point., whereas the sign of weight represents its correlation with dependent variable.

The weights for the features for our data obtained from rained logistic model are: -

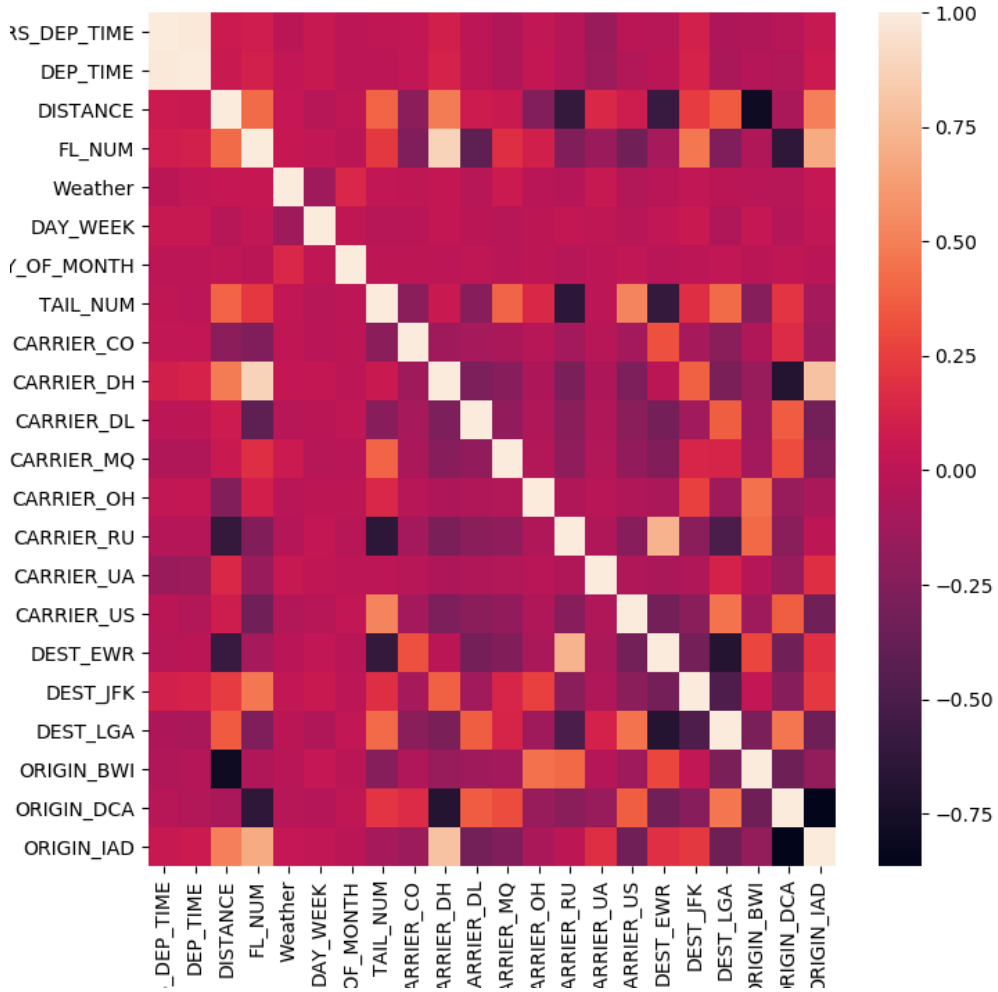
	Feature	Weight
0	CARRIER_MQ	-0.541132
1	CARRIER_DL	0.500421
2	Weather	-0.353248
3	DEST_EWR	-0.311050
4	ORIGIN_BWI	0.308894
5	DEST_JFK	0.288398
6	ORIGIN_IAD	-0.271173
7	CARRIER_CO	-0.219414
8	CARRIER_OH	0.217313
9	CARRIER_US	0.199933
10	CARRIER_RU	-0.195272
11	CARRIER_DH	0.088299
12	DAY_WEEK	0.086794
13	DEST_LGA	0.062345
14	DAY_OF_MONTH	-0.024227
15	CRS_DEP_TIME	0.023355
16	DISTANCE	0.012010
17	CARRIER_UA	-0.010455
18	ORIGIN_DCA	0.001972
19	TAIL_NUM	-0.000922
20	FL_NUM	0.000033

This shows that Carrier_MQ and Carrier_DL and weather are the most important features are Carrier column as a whole and weather as we have seen above analysis and Flight and tail no. are not useful in predicting delays and thus can be dropped.

Q4) Conclude the analysis by fitting a new model on these selected variables and report the same. Report the accuracy

Using a smaller number of features for prediction reduces the model size as well as computation and memory required but also does not overfit the dataset. Only thing we have care about the underfitting while dropping the columns.

1. Based on logistic regression model weights.
    a. Using the weights provided by logistic regression model we can drop the features having low weights like **TAIL_NUM and FL_NUM**.
    b. We can drop **TAIL_NUM, DATE** because No Unique information is encoded in it, the FL_NUM, DAY_OF_MONTH have that information.
2. karl pearson coefficient of variation to find out collinearity between the feature and what should be dropped.



a. From above correlation heatmaps we can see the **CRS_DEP_TIME** and **DEP_TIME** are highly correlated and same with the **CARRIER_DH**, **FL_NUM.** So we can drop either CRS_DEP_TIME or DEP_TIME and FL_NUM

3. Based on lasso regularized coefficient of features.

	Feature	Regularized Weight
0	CRS_DEP_TIME	0.022166
1	DEP_TIME	-0.022770
2	DISTANCE	0.012756
3	FL_NUM	0.000062
4	Weather	-3.198107
5	DAY_WEEK	0.055878
6	DAY_OF_MONTH	-0.021191
7	CARRIER_CO	-0.560021
8	CARRIER_DH	0.000000
9	CARRIER_DL	0.283185
10	CARRIER_MQ	-0.945181
11	CARRIER_OH	0.530892
12	CARRIER_RU	-0.325220
13	CARRIER_UA	0.084051
14	CARRIER_US	0.076428
15	DEST_EWR	0.000000
16	DEST_JFK	0.091012
17	DEST_LGA	-0.076138
18	ORIGIN_BWI	0.283172
19	ORIGIN_DCA	0.000000
20	ORIGIN_IAD	-0.599080

a. After applying lass regularisation penalty with logistic regression, we get the weight for each feature, we can remove the feature we have very low weight like **ORIGIN_DCA, FL_NUM, DEST_EWR, CARRIER_DH**

Based on above point I have removed features.
'ORIGIN_DCA','FL_NUM','DEST_EWR','CARRIER_DH',' TAIL_NUM'.

The accuracy with After removing above columns - `0.8967082860385925`

The code is given in a Jupyter notebook named **New_Model_Fllight_scheduling(Q_4 & Q _5)**.

Q.5 Find the ideal weather conditions for the highest chance of an on-time flight from DC to New York. (weather, time, day, carrier).

Based on the boxplot and bar graphs the ideals condition of flights are as follows:
Weather should be **good**
Time should be **1300 Hrs**
Day should be **Saturday**
Carrier should be **US**
Passenger should start his journey from **DCA** and ends to the **LGA.**
Flight number **2172** is best candidate.

	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	Flight Status
141	1300	US	1256	LGA	214	2172	DCA	0	6	3	ontime
670	1300	US	1256	LGA	214	2172	DCA	0	6	10	ontime
1189	1300	US	1259	LGA	214	2172	DCA	0	6	17	ontime
1712	1300	US	1259	LGA	214	2172	DCA	0	6	24	ontime
2180	1300	US	1256	LGA	214	2172	DCA	0	6	31	ontime

Q1. [1 Mark] Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY and EDITH.
H.E.L.E.N , JOKASTA, H.O.M.E.R, P.L.A.T.O, V.I.R.G.I.L

Q4. [1 Mark] In Star Wars Universe, name this robotic duo:
C-3PO and R2-D2 are robotic duo "robots" Star Wars.