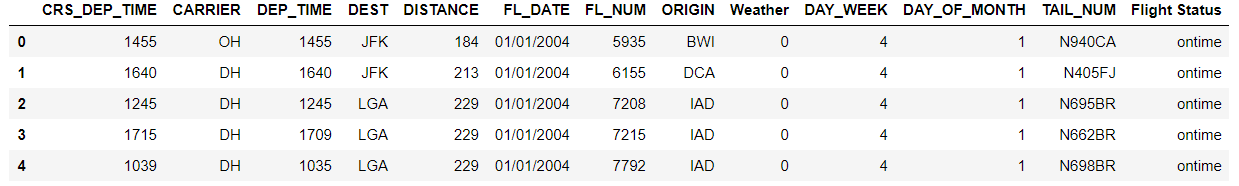
**ASSIGNMENT -1**

**GNR 652 COURSE**

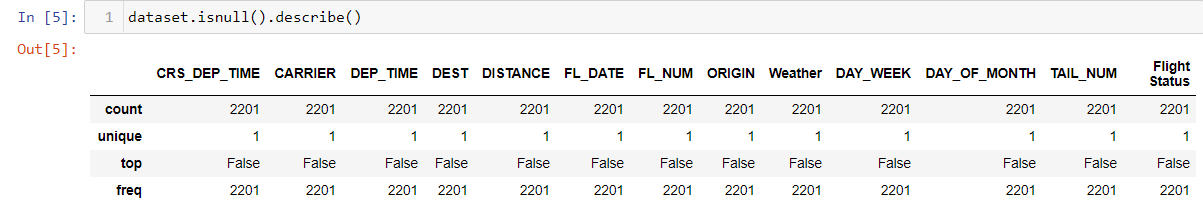
**PREDICTION OF FLIGHT DELAYS**

**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.)**

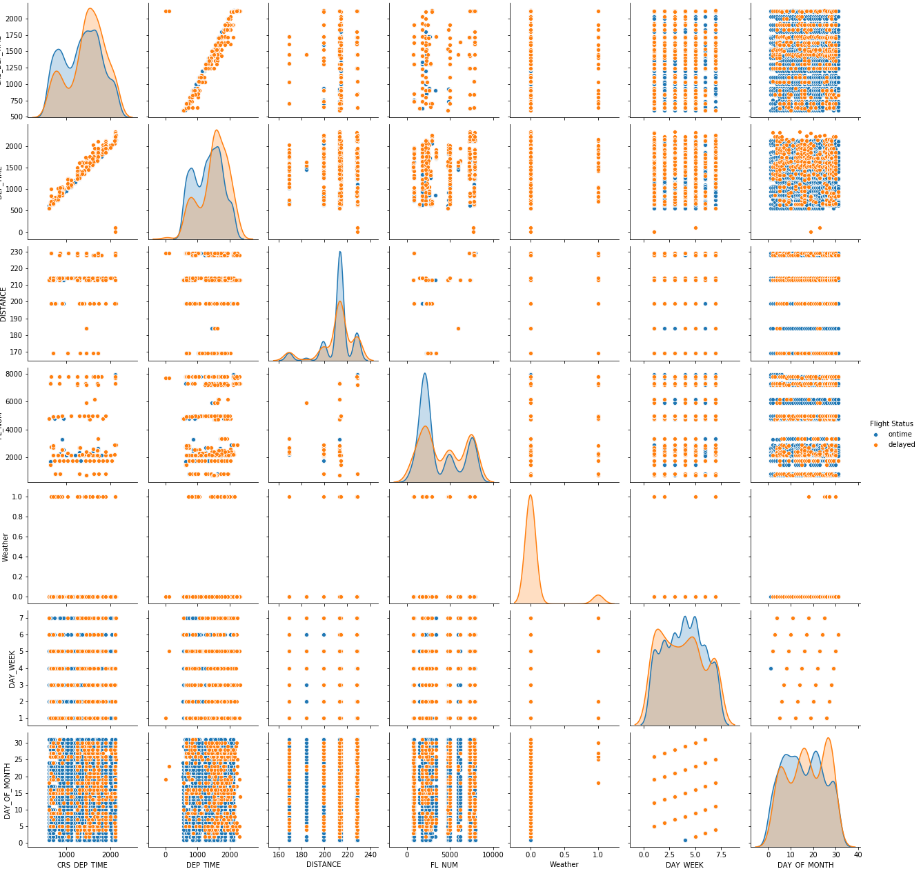
The given dataset has a total of 13 columns and 2201 data rows



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.

****

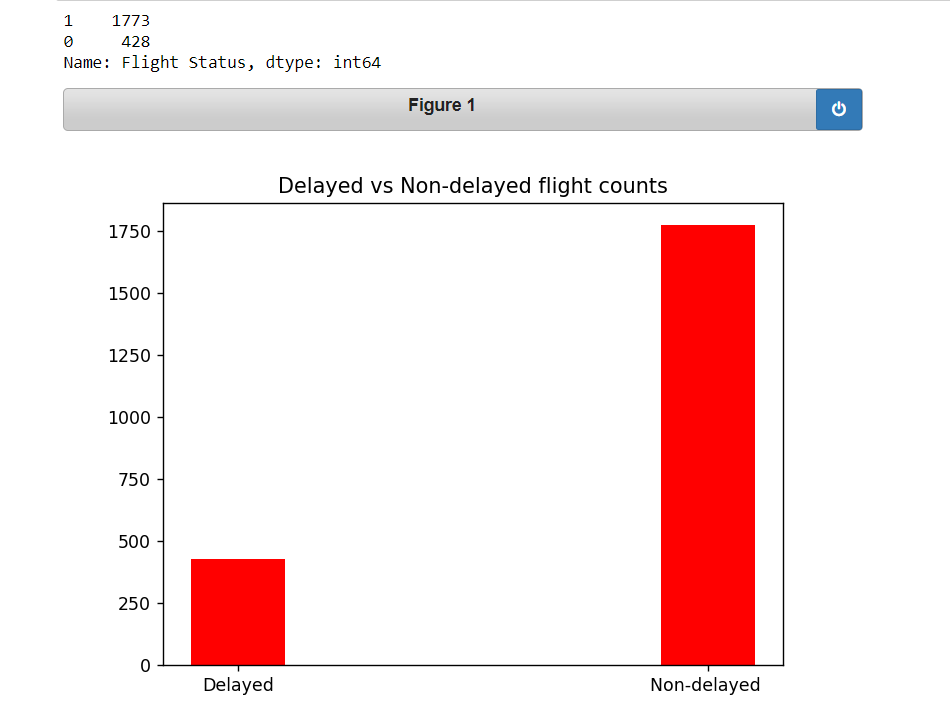
Also, to visualize the trend between categorical feature and flight status we have to convert Carrier, Origin, Destination and Flight status into dummies for modelling, plotting it as histogram we observe the following trends. For an example Tail No. has alphanumeric values and hence cannot be directly 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.

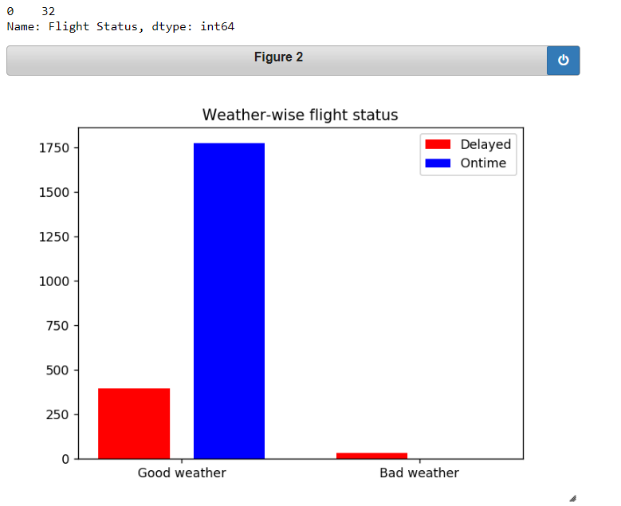
1. **Flight status**

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



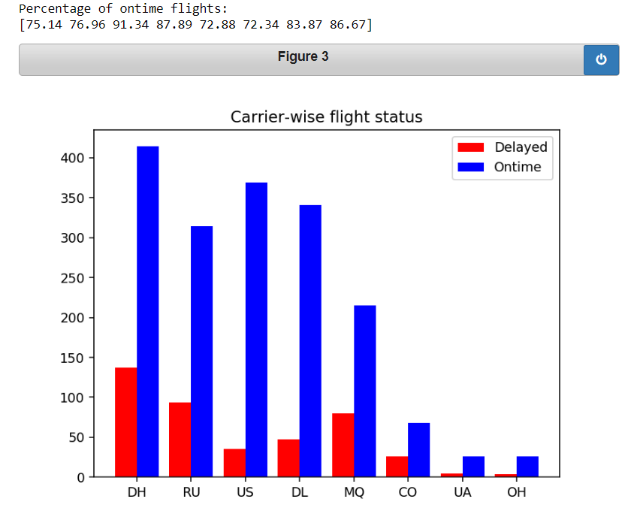
1. **Weather vs Flight Status**

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

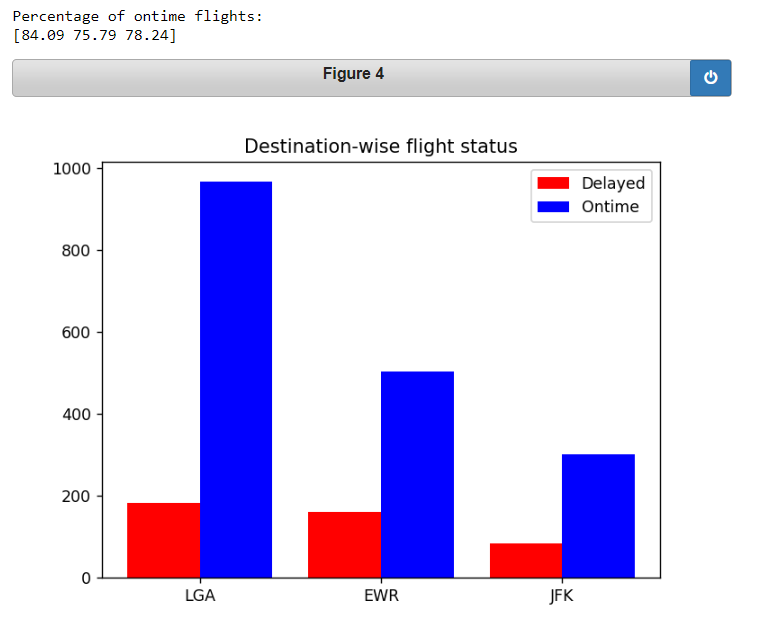
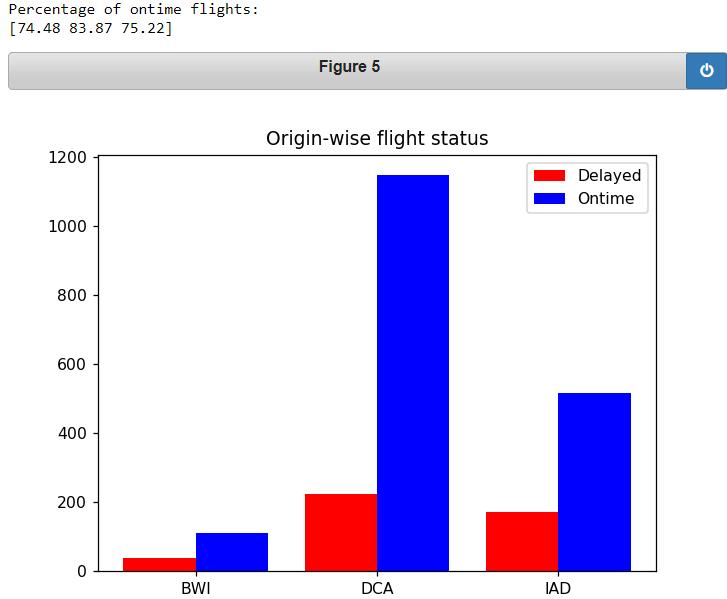


1. **Carrier vs Flight Status**

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.

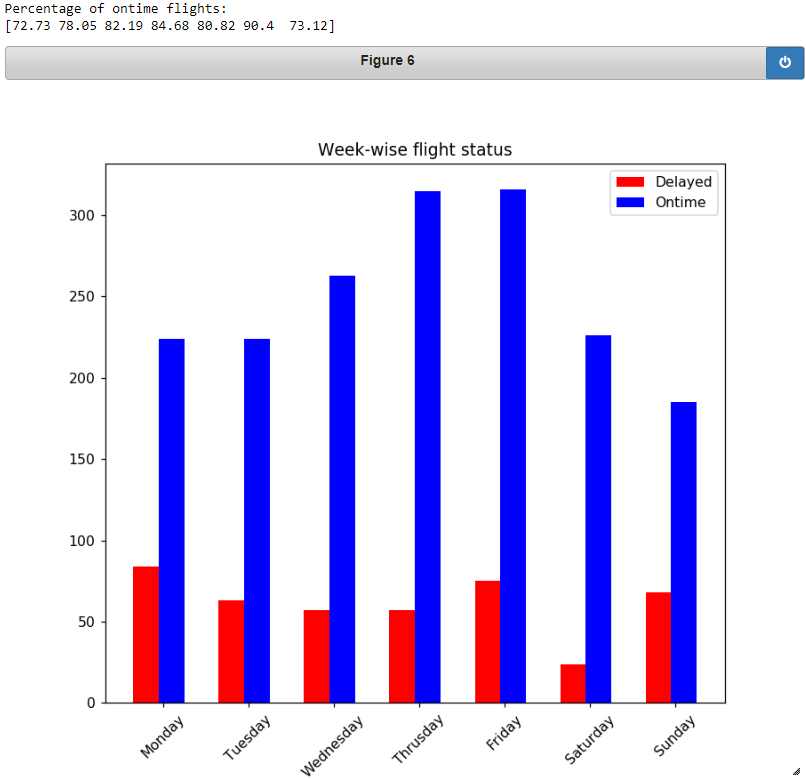
****

1. **Origin and destination airports**



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

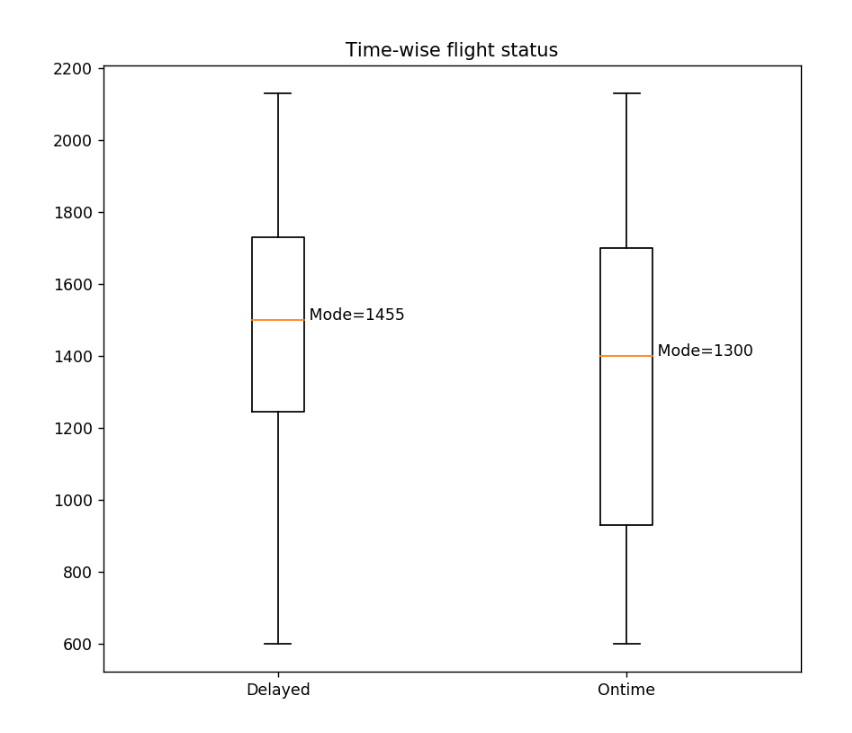
1. **Days of week vs Flight status**

****

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**.

1. **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.

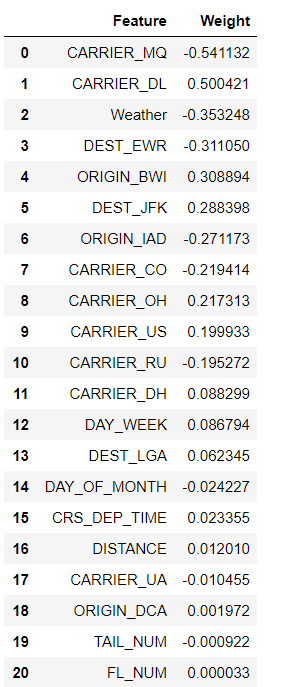
**The accuracy with all columns included** - **0.8978433598183881**

**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: -

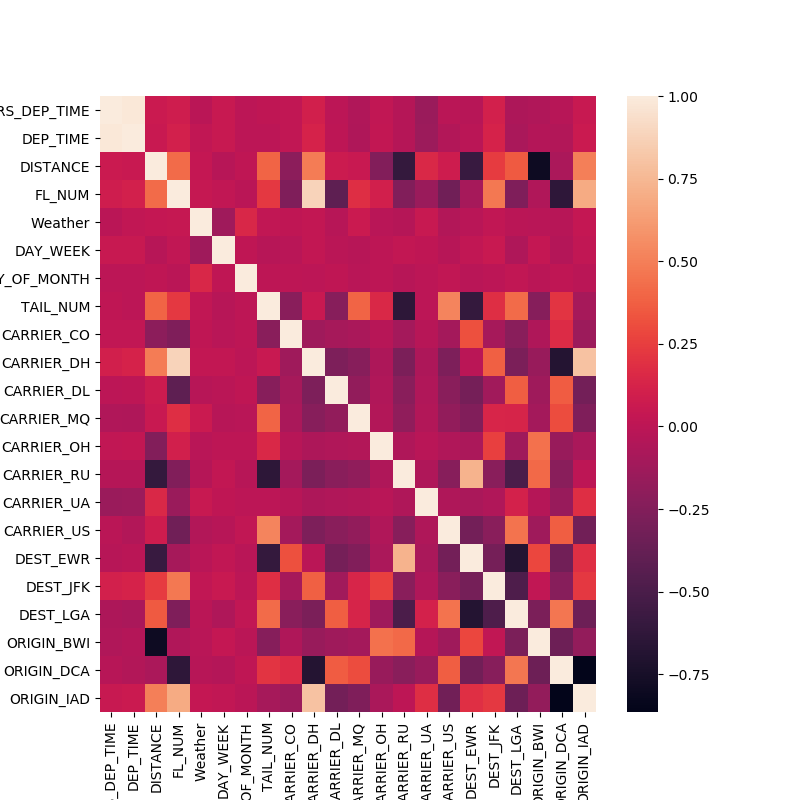


This shows that Carrier\_MQ and Carrier\_DL and weather are the most important features are Carrier column as a whole and weather as we have seen above analysis and Flight and tail no. are not useful in predicting delays and thus can be dropped.

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

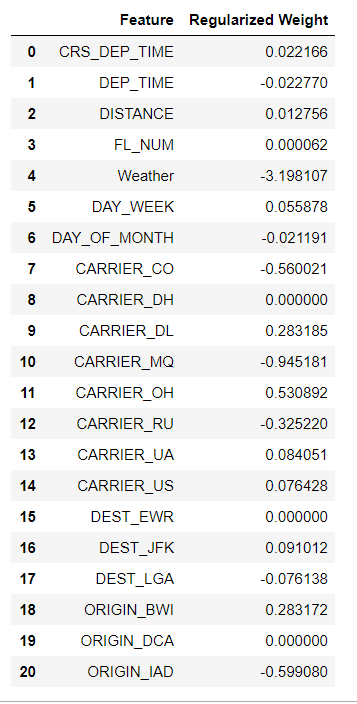
Using a smaller number of features for prediction reduces the model size as well as computation and memory required but also does not overfit the dataset. Only thing we have care about the underfitting while dropping the columns.

1. Based on logistic regression model weights.
   1. Using the weights provided by logistic regression model we can drop the features having low weights like **TAIL\_NUM and FL\_NUM**.
   2. 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.



* 1. 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

1. Based on lasso regularized coefficient of features.



* 1. 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** - **0.934367082860385925**

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**

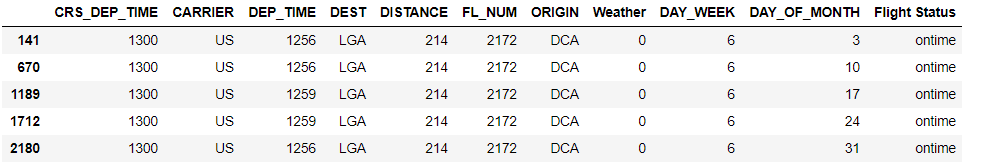
Time should be **1300 Hrs**

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.

****

**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.