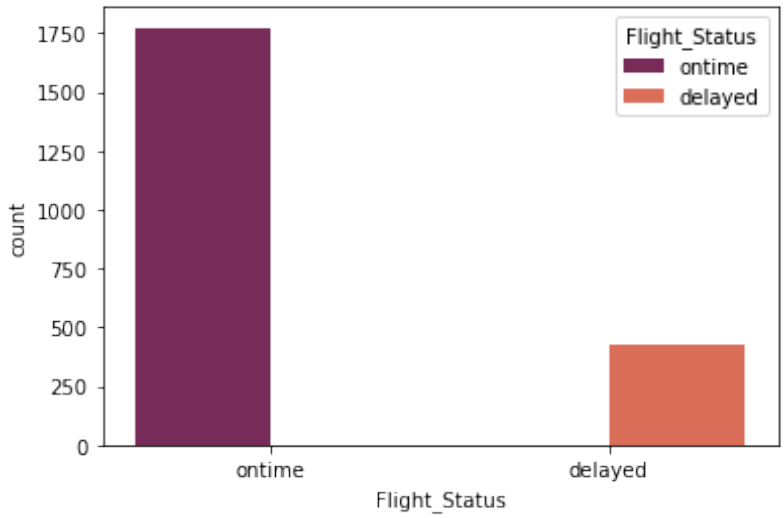
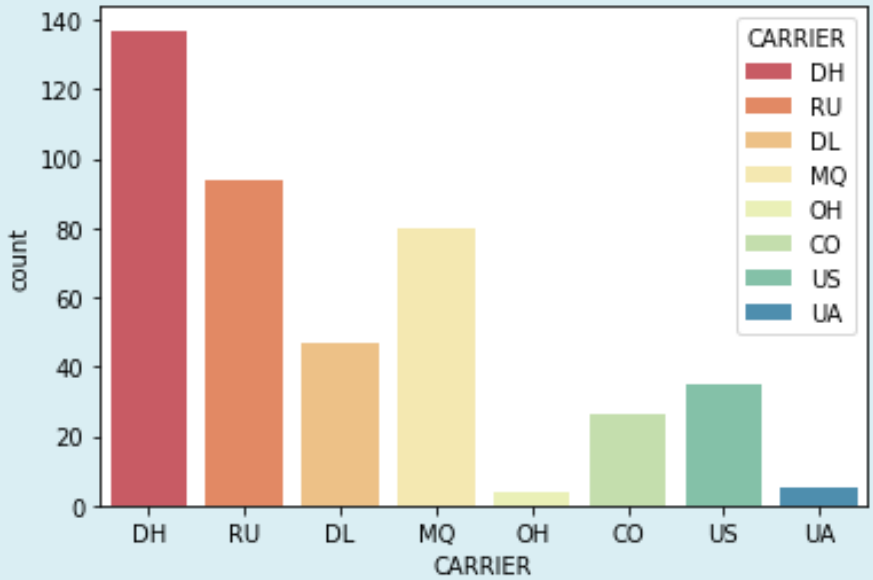
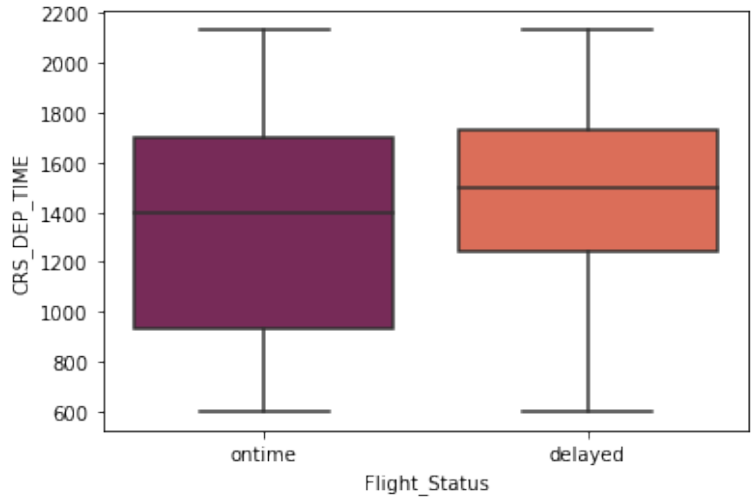
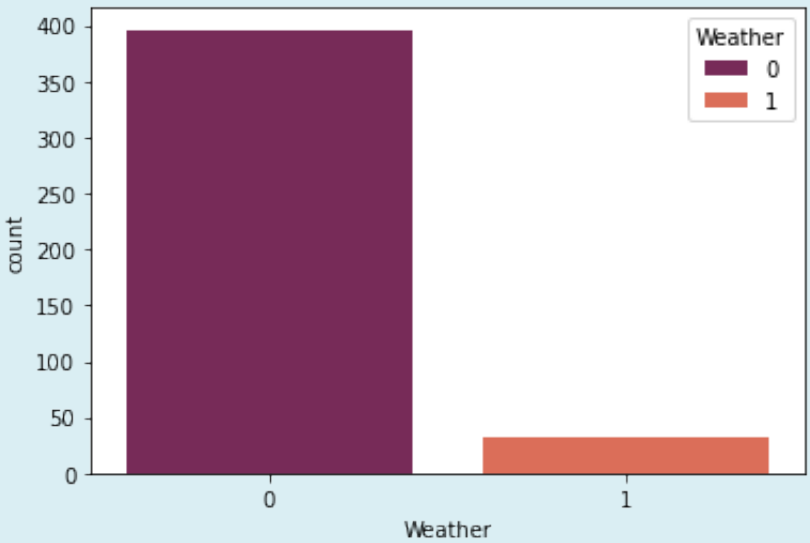
Q.1 Show visualisations to explore the dataset and understand the underlying trends (Often called Exploratory Data Analysis). Choose visualisation methods you think best represent the data (bar graph, pie chart, scatter, boxplot, heatmap etc.

For making visualizations Seaborn, Pandas and matplotlib libraries were used. Various features compare with each other. The outputs and inferences are given below.

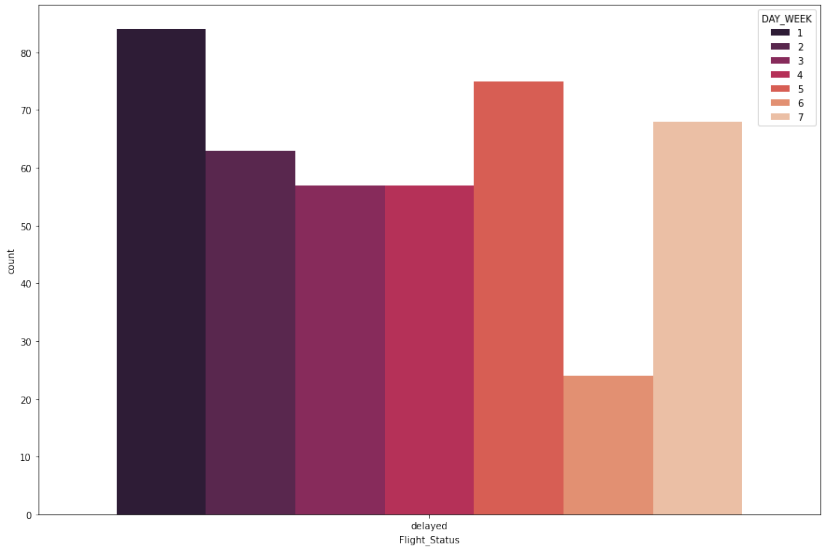
Graph	Analysis
	● This graph shows the number of samples having on time flights vs delayed flights. ● This shows that the data is imbalanced. Around 80% of flights are of "ontime" and around 20% are of "delayed".
	● The graph shows the delayed flights as per the carriers. ● We can deduce that DH (Atlantic Coast) has the highest number of delayed flights. ● So, we could check DH (Atlantic Coast), RU (Continental Express) and MQ (American Eagle) like what is causing the problems.



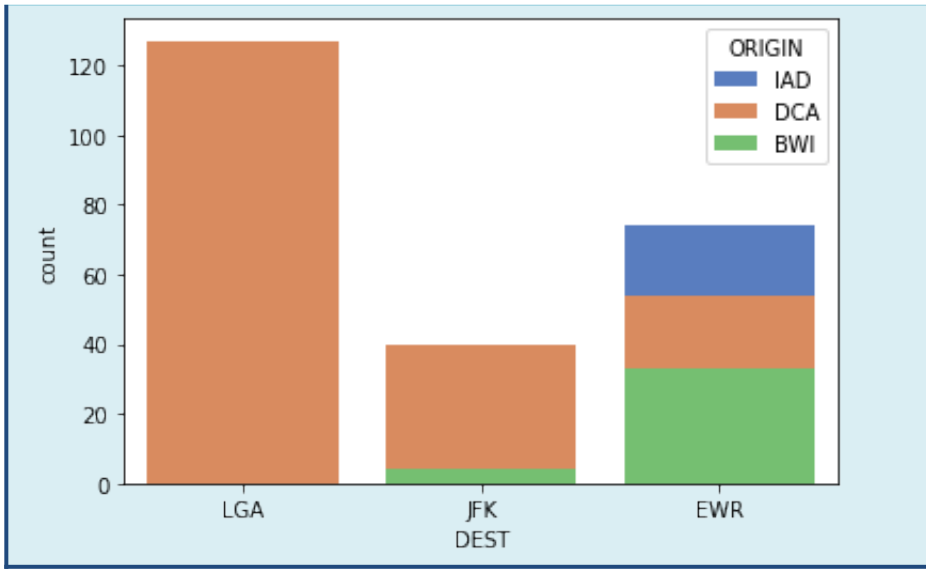
- As we can see in the box plot, IQ range of the delayed flights lies in the late afternoon (after 13:00).



- As per weather feature, we can deduce that all delayed happened due to weather related ituations.



- Most number of delayed flights are on either Monday, Friday or Sunday because most people like to travel on weekends.
- Also as people stays over weekend, it is also shown on Saturday delayed flights are less.



- As we can observe, all flights starting from LGA and ending at DCA are being delayed and around 90% flights starting from JFK going to DCA are delayed.

Q.2 Preprocess 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.

As per the given data, the classes have been divided into 80% 'on-time 'class and 20% 'delayed' flights. The dataset does not contain any null values as shown below. Categorical features are represented by the 'object' data type and the rest of them are numerical features represented by 'int64'.

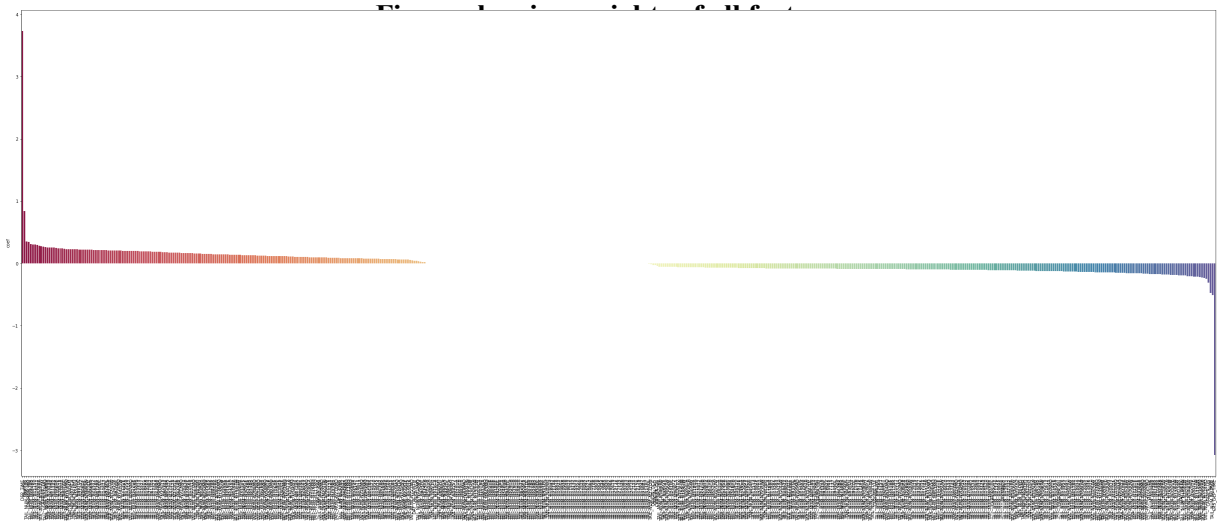
```
flight_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2201 entries, 0 to 2200
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   CRS_DEP_TIME   2201 non-null   int64
 1   CARRIER        2201 non-null   object
 2   DEP_TIME       2201 non-null   int64
 3   DEST           2201 non-null   object
 4   DISTANCE       2201 non-null   int64
 5   FL_DATE        2201 non-null   object
 6   FL_NUM         2201 non-null   int64
 7   ORIGIN         2201 non-null   object
 8   Weather        2201 non-null   int64
 9   DAY_WEEK       2201 non-null   int64
 10  DAY_OF_MONTH   2201 non-null   int64
 11  TAIL_NUM       2201 non-null   object
 12  Flight_Status  2201 non-null   object
dtypes: int64(7), object(6)
memory usage: 223.7+ KB
```

Machine Learning algorithms work only with numerical data. Therefore,we should convert all the categorical data to numerical data. To do so, we can use sklearn's 'OneHotEncoder' and Pandas' 'get_dummies'. I have used 'pd.get_dummies' to convert all categorical data to numerical data(generating dummy variables). And used pd.factorize() to convert target class labels into numerical data.

After that, I have splitted the dataset into 60% train, 40% test randomly using test_train_split from sklearn. Then, I have used all the features with basic hyperparameters to train the Logistic Model. After training, I have tested the model on trained model. The code as well as the results are shown in the ipynb file.

Q.3 Interpret the model and coefficients and present some insights.

Also, the mean and max model score are shown below. I have used full data to train the model after one hot encoding the feature variables and label encoding of the target variable that is Flight_Status. After finding the coefficients for the columns, I observed that the coefficients were present in the range between -0.25 and 0.25 . Now, the coefficients with high magnitude were important to us. So, I dropped the features with very low coefficients magnitudes. This significantly increased the magnitude.



Q.4 Perform variable selection, and reduce the size of the model, only keeping the relevant variables based on the analysis done earlier. (What variables are significant? What variables are not significant?)

Logistic regression coefficients cant be easily interpreted. This is because logistic regression uses the logit link function to "bend" our line of best fit and convert our classification problem into a regression problem. Because of the logit function, logistic regression coefficients represent the log odds that an observation is in the target class ("1") given the values of its X variables. Thus, these log odd coefficients need to be converted to regular odds in order to make sense of them. Happily, this is done by simply exponentiating the log odds coefficients, which we can do with np.exp()

Now we will set threshold for the coefficients to only select the coefficients with higher magnitudes

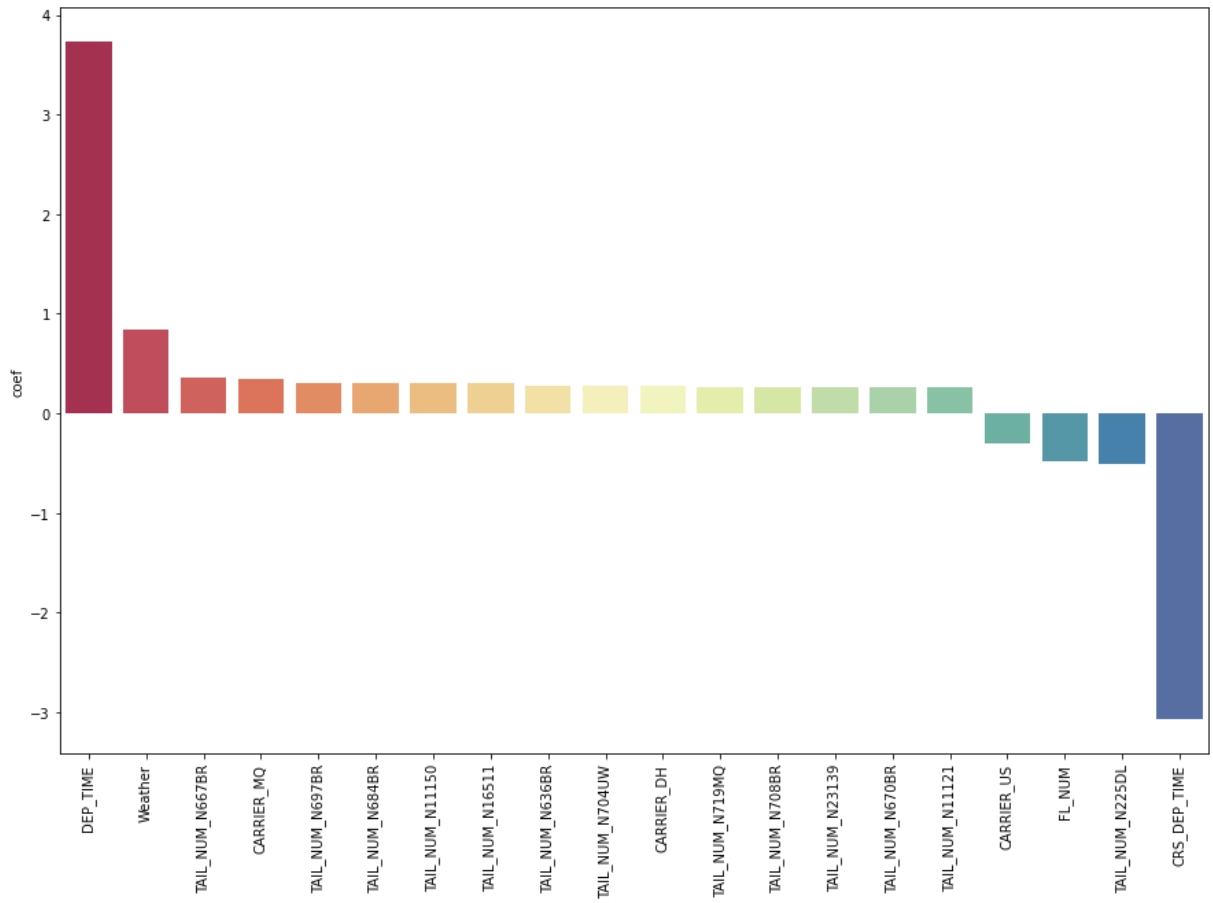


Figure showing features after feature slectionwith coefficients

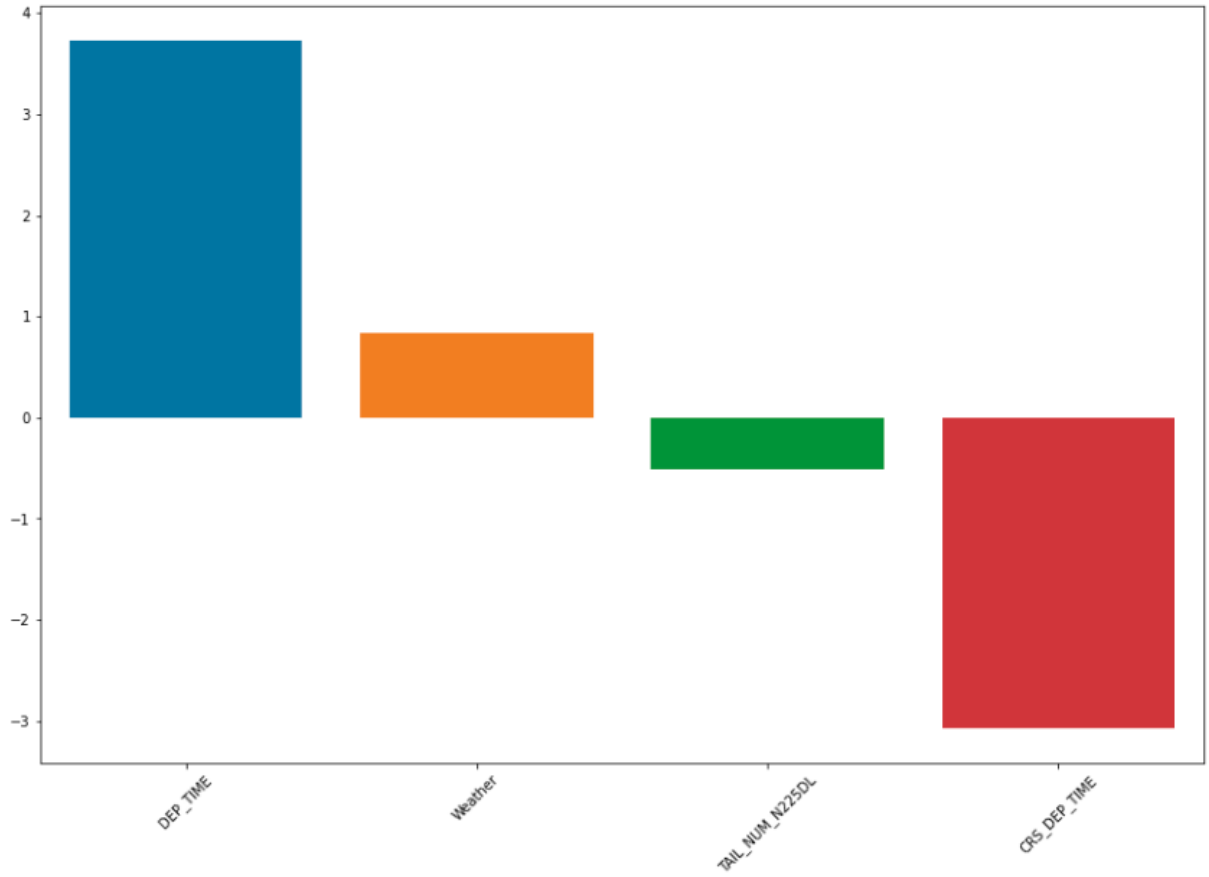


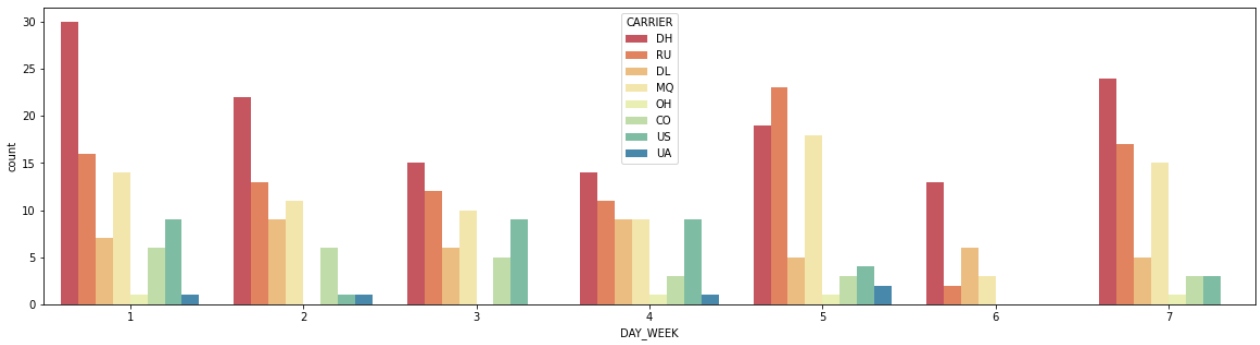
Figure showing feature having coefficients higher than 0.5

Q.5 Conclude the analysis by fitting a new model on these selected variables and report the same. Report the accuracy.

New model was fitted and the accuracy obtained is given below:

- Basic Logistic Regression Model – Accuracy before On Hot Ecoding ☾ 77.52%
- Basic Logistic Regression Model – Accuracy on the top 20 features ☾ 87.74%
- Cross Validation Model having `cv=10` - The mean accuracy on all the validation sets ☾ 81.33% and the max accuracy is ☾ 89.86%

Q.6 Find the ideal weather conditions for the highest chance of an ontime flight from DC to New York . (weather, time, day, carrier)



From the the plots we can observe that the maximum on time flights are between LGA and DCA. Therefore we now find for the Carrier. The dataframe shows all the possible cases for day 6 flights filtered for only the 4 carriers from DCA to LGA. Fromt he above discussion,

Best carrier option: **US**

Departure time: **12:55**

Flight number : **808**

Weather: **0** (no weather related delay)

Bonus Questions

Q.1 Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY and EDITH.

Ans1 – JOCASTA and TADASHI. Also Tony's robots like DUM-E, Butterfingers etc.

Q.2

Q.3 In Star Wars Universe, name this robotic duo:

Ans4 – C-3PO and R2-D2