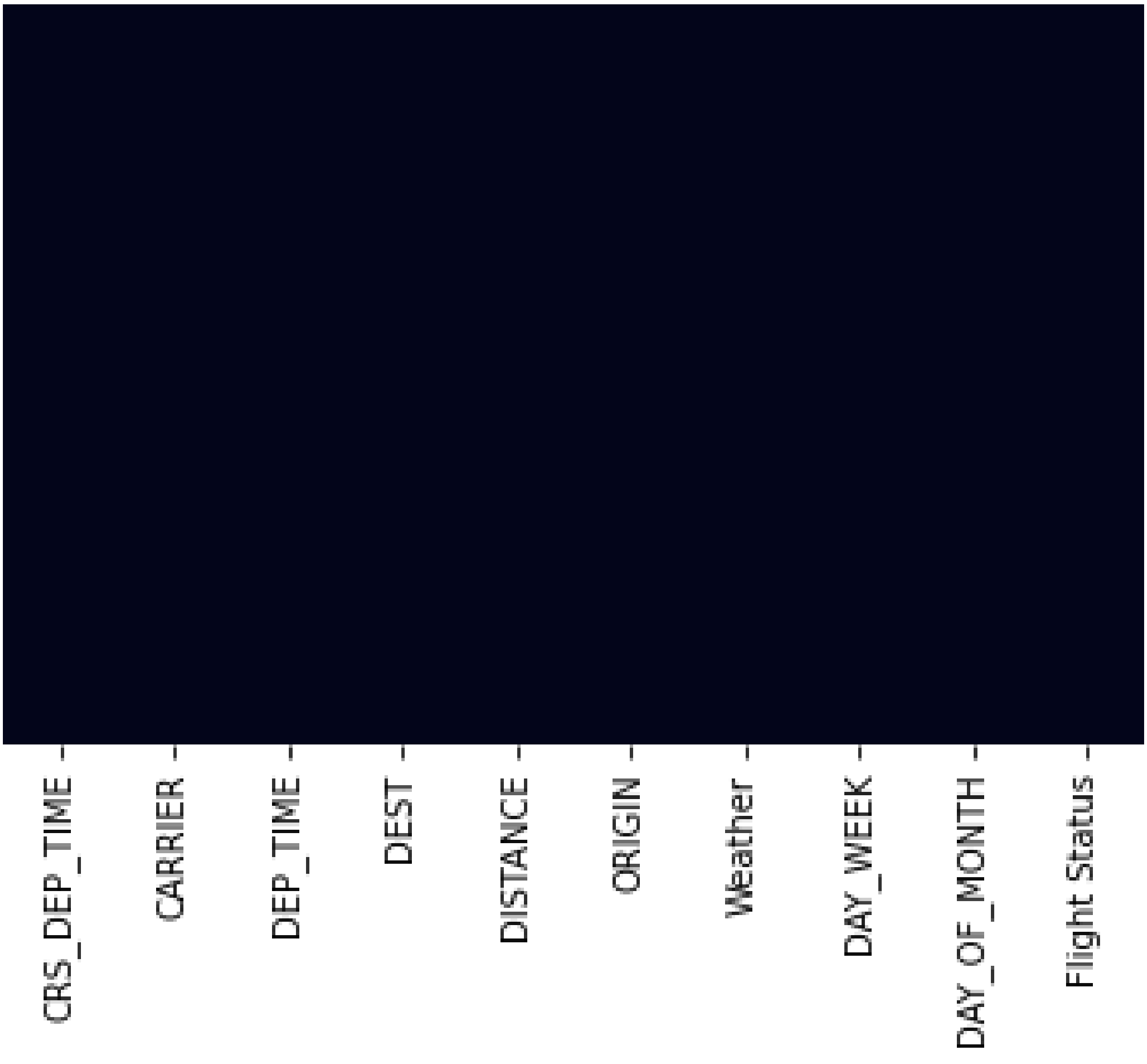
## Honour Code:

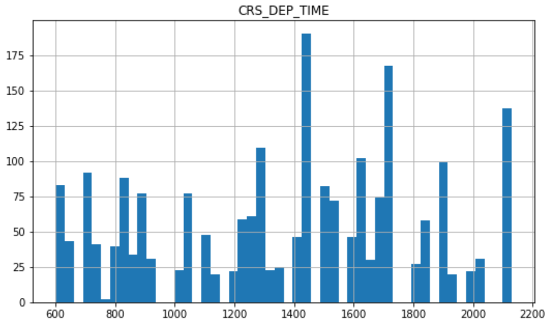
I accept sole responsibility for this entire work, and I have adhered to the following:
- worked independently on this assignment, with utmost sincerity.
- not copy/falsify/fabricate information/ideas in this assignment.
- not disseminate information gathered/submitted in the course of this assignment with a view to facilitate unfair practices.

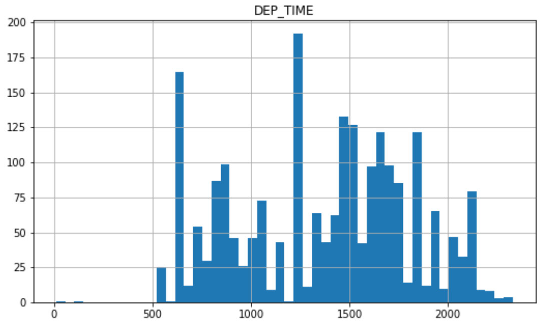
Before beginning, I have dropped 3 columns (*FL_NUM*, *TAIL_NUM* & *FL_DATE*) because they have a large number of categories and do not seem related to flight delay.

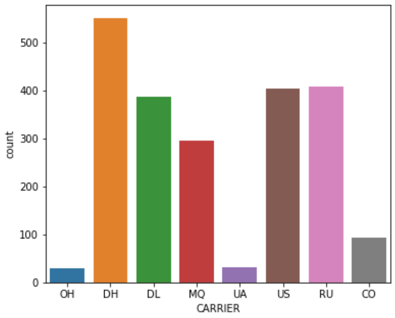
1. I started by plotting a heatmap to analyse for missing attributes. As we can see from out heat map, there are no missing values.

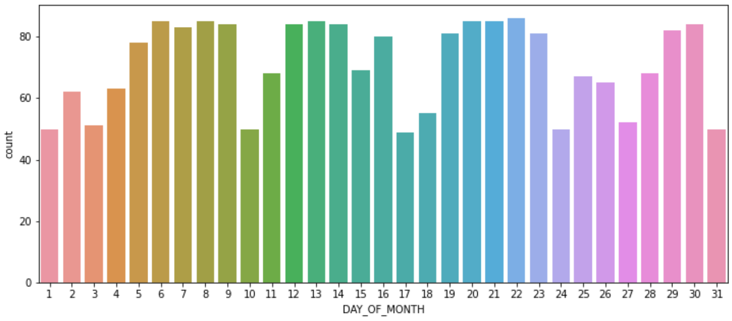


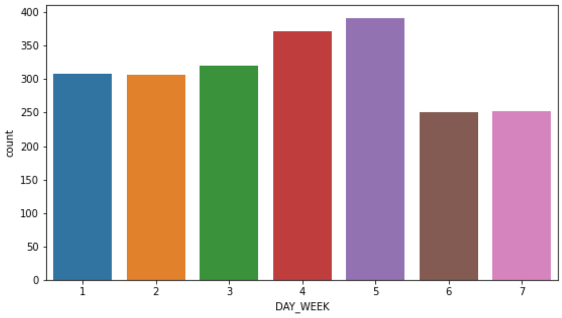
Next I plotted histograms (continuous data) and count plots (categorical data) for visualising the data distribution.

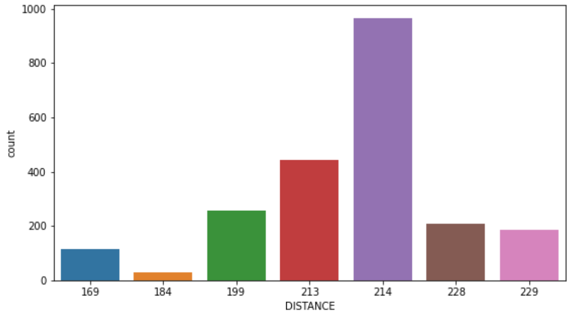


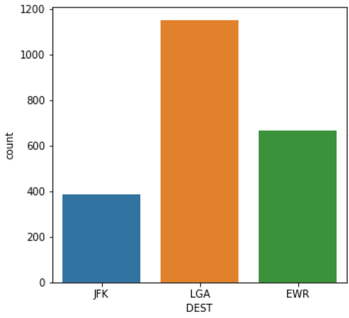


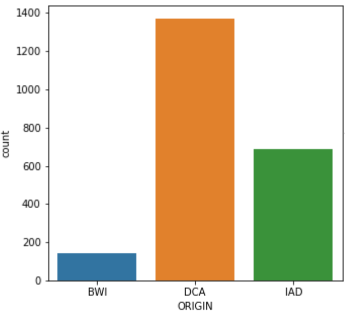


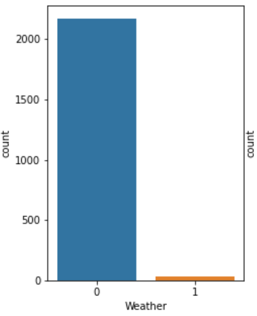


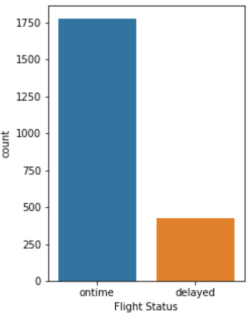






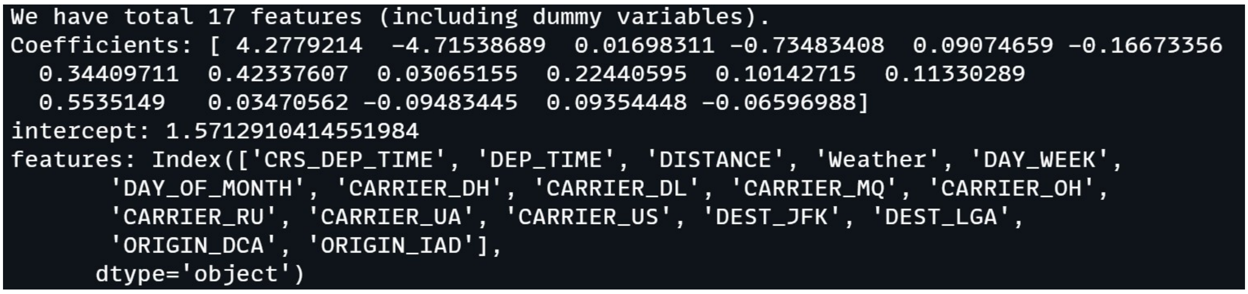






2. We can begin preprocessing our data now that we have seen what our data looks like. Our dataset doesn't have any null values upon closer analysis. Next we create dummy variables for our categorical features. We will use one hot encoding as the number of unique categories in a feature aren't many (max 8). After dummies and deleting original feature col, our columns have increased to 20. To avoid multicolinearity (dummy variables trap), we will drop the first column. The last step we will perform before building our model would be to scale down are variables. We can finally proceed to build the logistic regression model on our training dataset and run it on our test data. We get an accuracy of **87%**.

3. Proceeding to analyse the models coefficients. In general, a positive coefficient means that as our feature increases our output increases, and a negative value means as the feature increases our output decreases. Higher value of a coefficient means that the output will be more influenced by that feature, and a lower (closer to 0) value means that the feature doesn't affect our classification greatly and can thus be dropped.



```
We have total 17 features (including dummy variables).
Coefficients: [ 4.2779214  -4.71538689  0.01698311 -0.73483408  0.09074659 -0.16673356
  0.34409711  0.42337607  0.03065155  0.22440595  0.10142715  0.11330289
  0.5535149   0.03470562 -0.09483445  0.09354448 -0.06596988]
intercept: 1.5712910414551984
features: Index(['CRS_DEP_TIME', 'DEP_TIME', 'DISTANCE', 'Weather', 'DAY_WEEK',
       'DAY_OF_MONTH', 'CARRIER_DH', 'CARRIER_DL', 'CARRIER_MQ', 'CARRIER_OH',
       'CARRIER_RU', 'CARRIER_UA', 'CARRIER_US', 'DEST_JFK', 'DEST_LGA',
       'ORIGIN_DCA', 'ORIGIN_IAD'],
      dtype='object')
```

From our data, we can see that *CRS_DEP_TIME* and *DEP_TIME* have large positive and negative values respectively. Thus we can say that delayed or not has to mainly do with it.
*DISTANCE*, *DAY_OF_MONTH*, *DEST* (dummies) and *ORIGIN* (dummies) have very small coefficients and hence we can drop them during feature selection.

4. For performing variable selection, we will create the following new features:

   (i) *WEEKEND*: Value 1 if the flight was on a weekend, else value 0. We can attain this information using the *DAY_WEEK* feature.

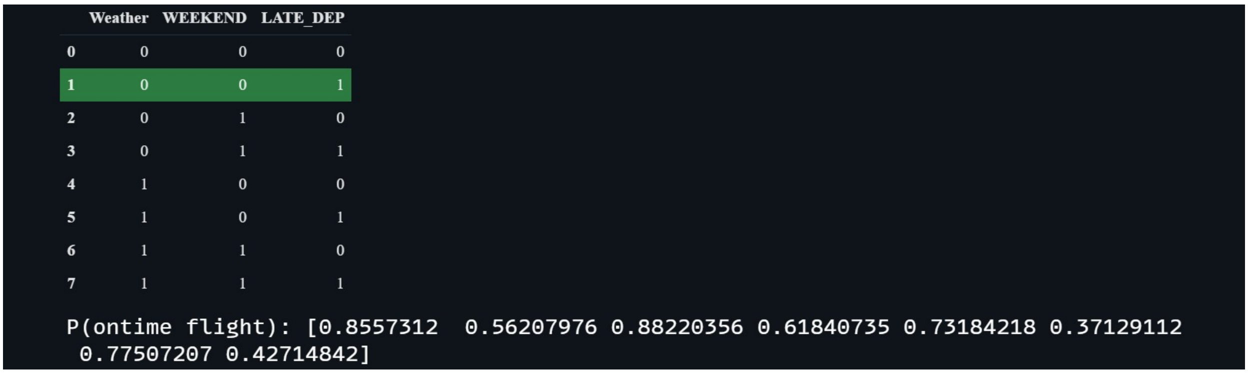
   (ii) *LATE_DEP*: Value 1 if the actual flight departure happened 15 minutes after the scheduled time, else value 0. 15 minutes is the threshold value for deciding on-time or delayed. We can attain this information using the *DEP_TIME* and *CRS_DEP_TIME* features.

   After introducing these new features, we can proceed to drop *DAY_WEEK*, *DEP_TIME* and *CRS_DEP_TIME.*

I tried to use sin-cos encoding for the cyclic *DAY_OF_MONTH* feature but didn't get any satisfactory results. Thus we can drop this feature too.
Through the interpretation of the logistic regression coefficients, we can drop the *DISTANCE*, *DEST* and *ORIGIN* features.
*CARRIER* had comparatively large coefficients and a lot of categories. Thus I used Count Encoding (replace each category by the number of times it has occurred) for it. This did not improve my accuracy and so I decided to drop the feature.

5. We can now finally fit a new model on our dataset which has undergone feature selection. After doing the same, we attain an accuracy of **91%**.

6. For the highest chance of an on-time flight from DC to New York, we will brute-force all states (0 to 7):



| | Weather | WEEKEND | LATE_DEP |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 1 | 1 |
| 4 | 1 | 0 | 0 |
| 5 | 1 | 0 | 1 |
| 6 | 1 | 1 | 0 |
| 7 | 1 | 1 | 1 |

```
P(ontime flight): [0.8557312  0.56207976 0.88220356 0.61840735 0.73184218 0.37129112
 0.77507207 0.42714842]
```

From the probability array we can see that the factors for highest probabilities (0.8822 & 0.8557) are: **no weather delay** and **no departure delay**.
Day could be anything (preferably Saturday/Sunday) and carrier could be anything.

## BONUS:

1. AIs by Tony Stark: **H.O.M.E.R.** (Heuristically Operative Matrix Emulation Rostrum) or **P.L.A.T.O.** (Piezo-Electrical Logical Analytical Tactical Operator).

3. X is the **Rule of Two**.