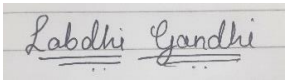
DS303 ASSIGNMENT 1

Honor Code

I declare that I have done the assignment by myself and haven't copied from anywhere.



Q1)

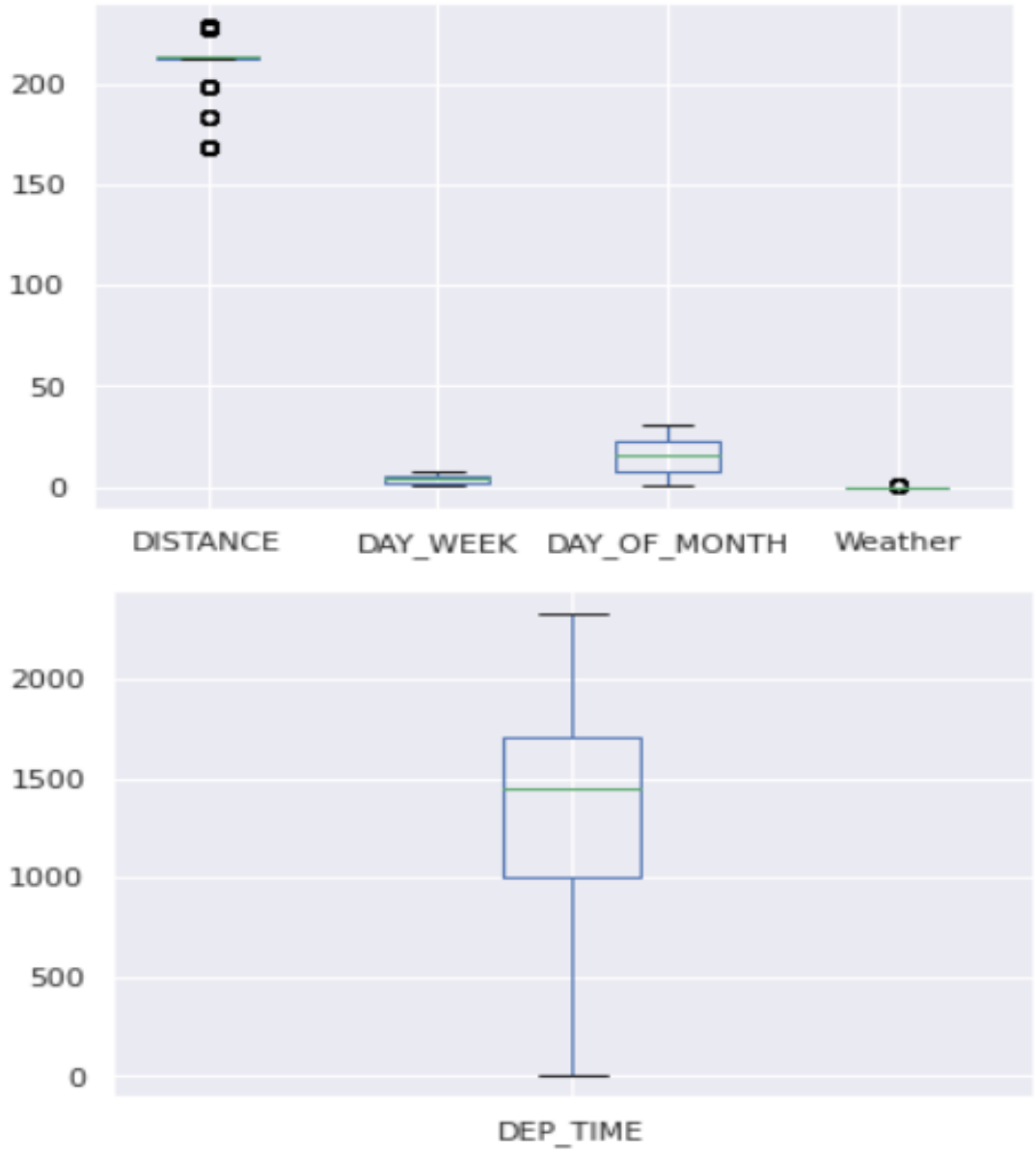
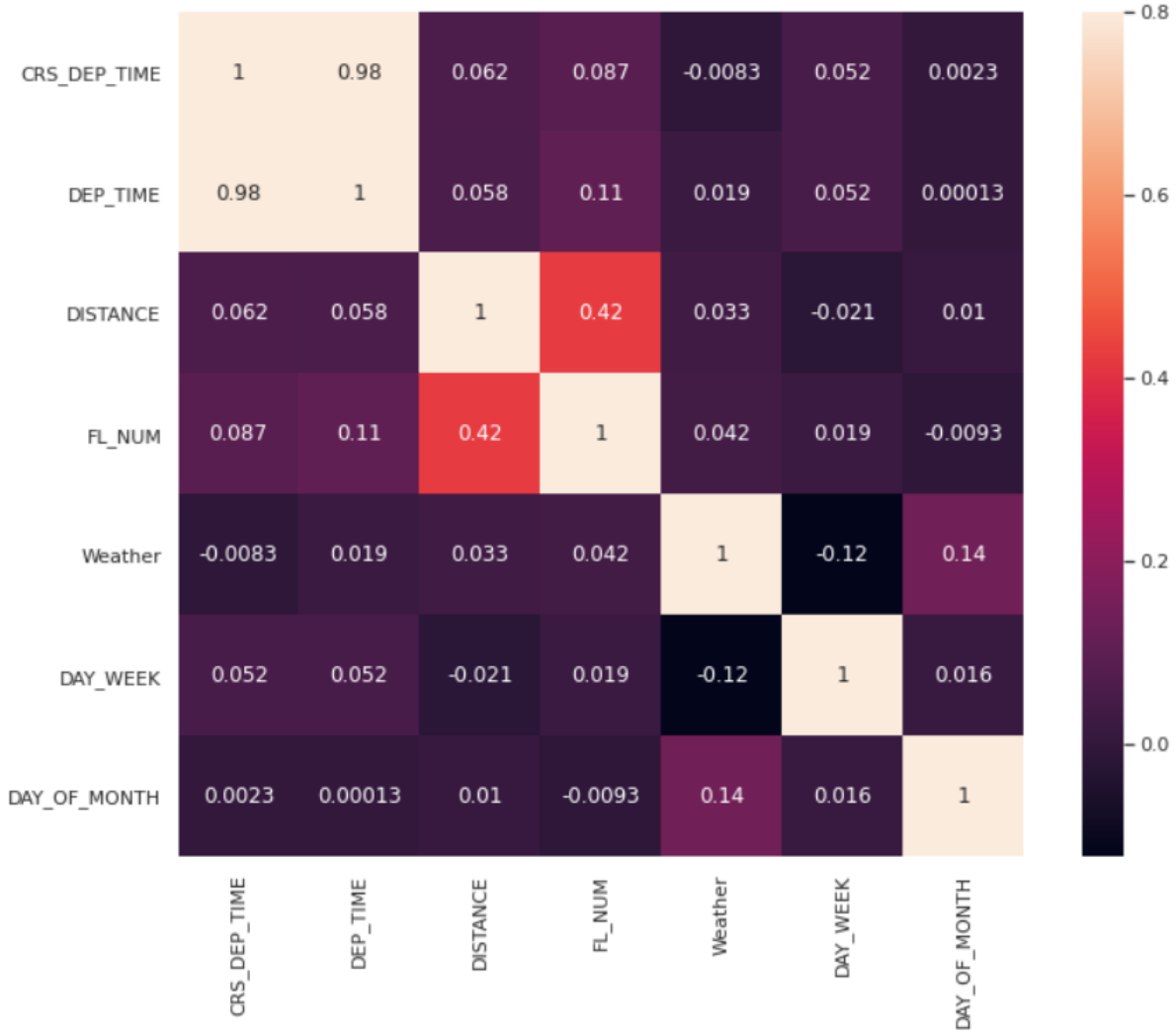
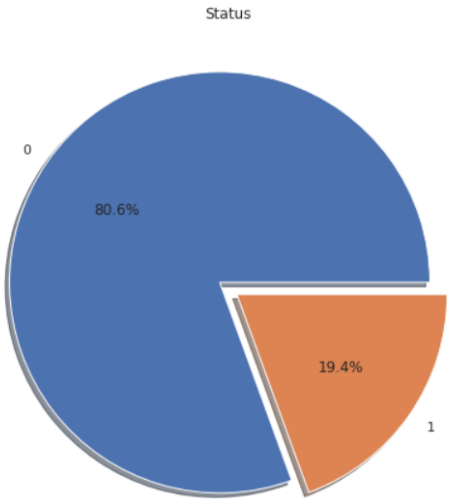
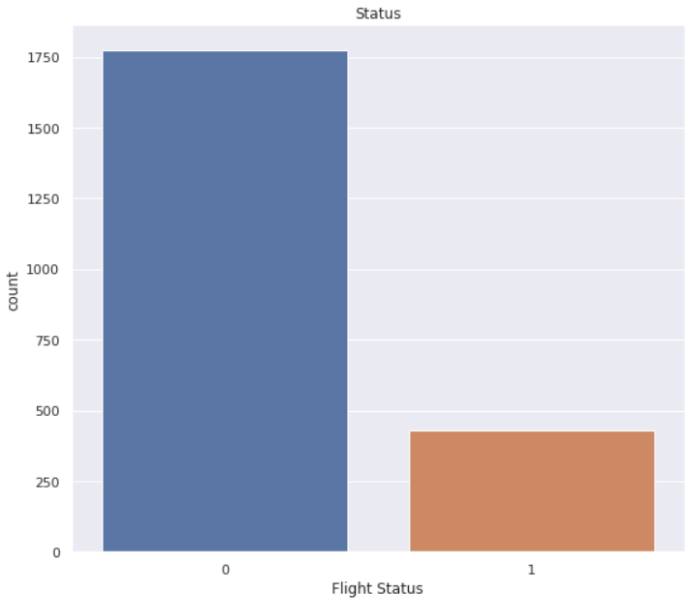


Figure shows the **Boxplot** for different attributes. The distance attribute has many outliers.



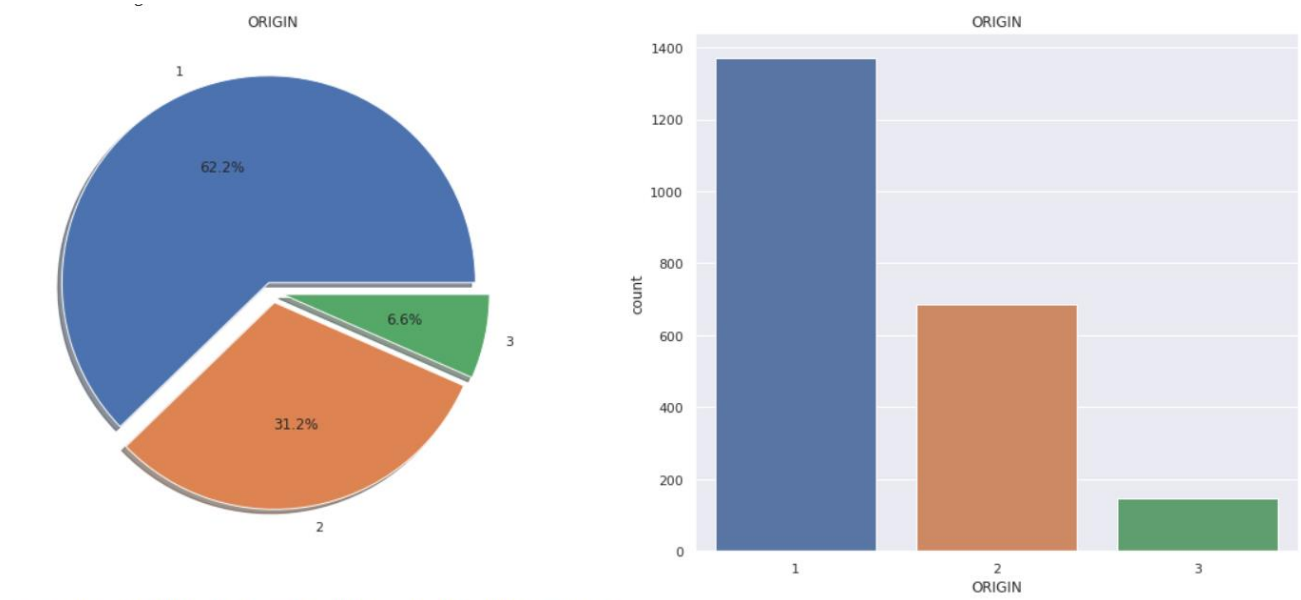
Through the correlation matrix we can see how some of the variables in the dataset present multicollinearity, in other words, can be linearly predicted from the others. The above figure represents a heat map from which we can predict the correlation of one variable to another.





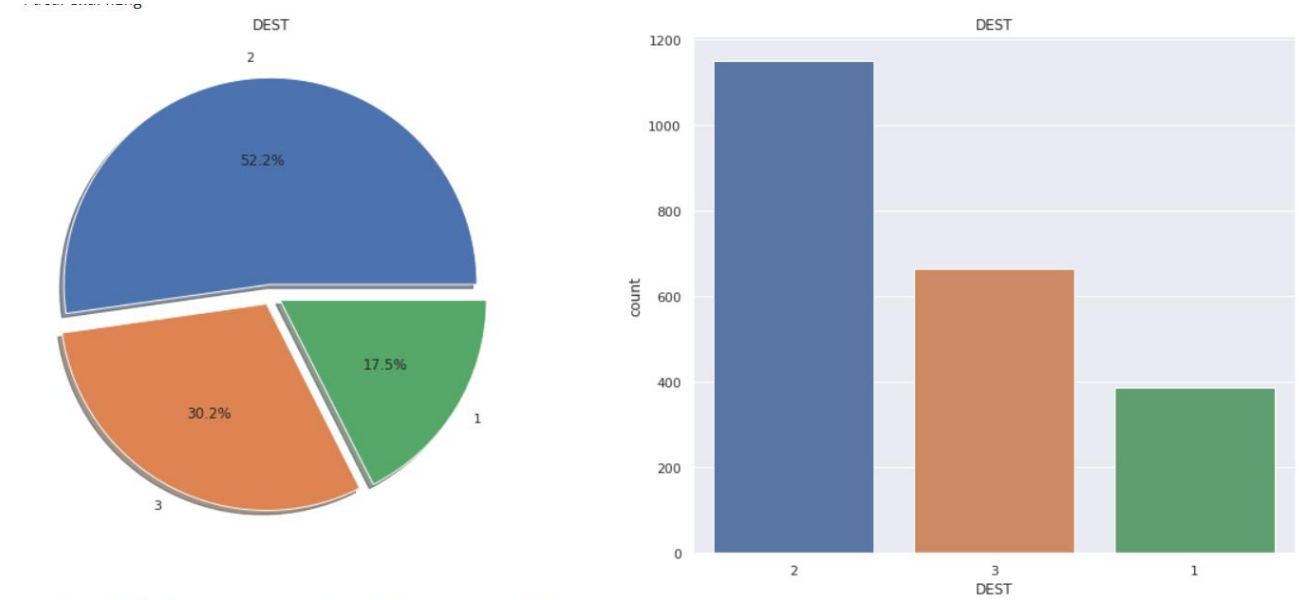
Status represents weather the flight was on time (0) or delayed (1)

The above pie chart shows the percentage of on time and percentage of delayed flights. Also the bar graph beside the above pie chart represents the number (count) of on time and delayed flights.



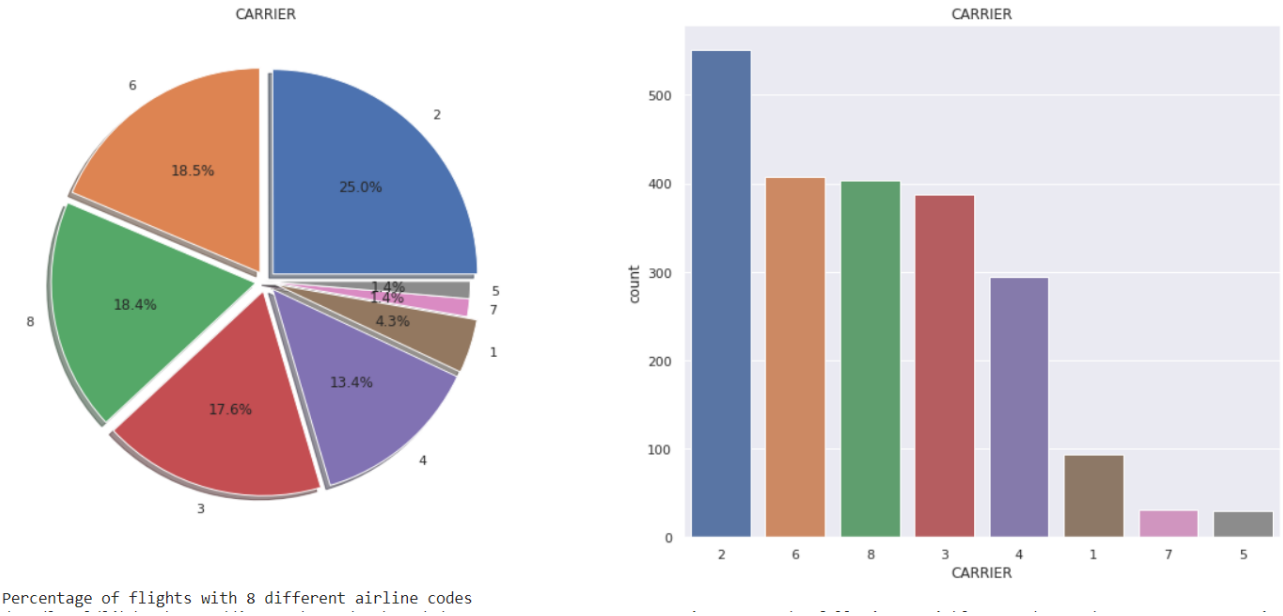

```
Percentage of flights from various origins: DCA(1), IAD(2), BWI(3)
```

The above pie chart shows the percentage of flights from 3 different origins. Also the bar graph beside the above pie chart represents the number (count) of flights from the 3 different origins.



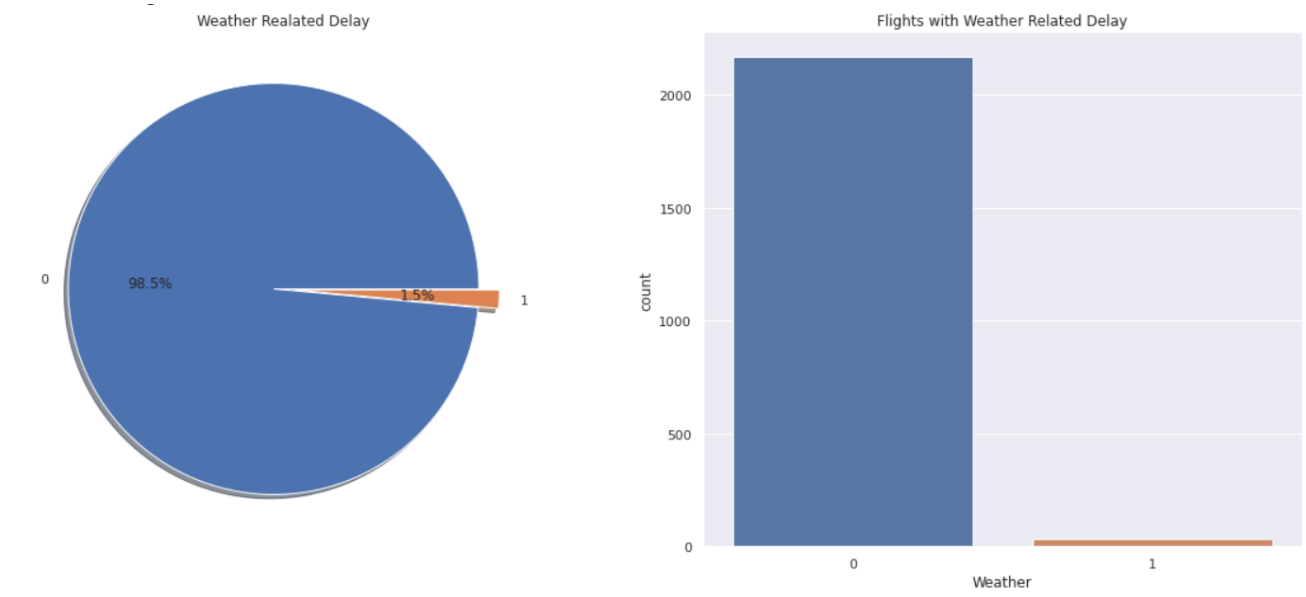

```
Percentage of flights with various destinations: JFK(1), LGA(2), EWR(3)
```

The above pie chart shows the percentage of flights with 3 different destinations. Also the bar graph beside the above pie chart represents the number (count) of flights with the 3 different destinations.



Percentage of flights with 8 different airline codes

The above pie chart shows the percentage of flights with 8 different airline codes. Also the bar graph beside the above pie chart represents the number (count) of flights having these 8 different destinations.



Percentage of flights with Weather related delay(1)

The above pie chart shows the percentage of flights delayed due to bad weather. Also the bar graph beside the above pie chart represents the number (count) of flights having delays due to bad weather conditions.

The above figure shows the scatter plot between any two variables.

Q2)

	CRS_DEP_TIME	CARRIER	DEP_TIME	...	DAY_WEEK	DAY_OF_MONTH	Flight Status
0	1455	5	1455	...	4	1	0
1	1640	2	1640	...	4	1	0
2	1245	2	1245	...	4	1	0
3	1715	2	1709	...	4	1	0
4	1039	2	1035	...	4	1	0
5	840	2	839	...	4	1	0
6	1240	2	1243	...	4	1	0
7	1645	2	1644	...	4	1	0
8	1715	2	1710	...	4	1	0
9	2120	2	2129	...	4	1	0
10	2120	2	2114	...	4	1	0
11	1455	3	1458	...	4	1	0
12	930	3	932	...	4	1	0
13	1230	3	1228	...	4	1	0
14	1430	3	1429	...	4	1	0
15	1730	3	1728	...	4	1	0
16	2030	3	2029	...	4	1	0
17	1530	4	1525	...	4	1	0
18	600	4	556	...	4	1	0
19	1830	4	1822	...	4	1	0
20	900	4	853	...	4	1	0
21	1300	4	1254	...	4	1	0
22	1400	4	1356	...	4	1	0
23	1500	4	1452	...	4	1	0
24	1900	4	1853	...	4	1	0
25	850	7	841	...	4	1	0
26	900	8	858	...	4	1	0
27	1100	8	1056	...	4	1	0
28	1300	8	1253	...	4	1	0
29	1500	8	1458	...	4	1	0
30	1700	8	1655	...	4	1	0

some representative data samples after pre processing the dataset (removeing null values, generating dummy variables etc.)

Q3)

```
intercept: [-1.51203094]
coefficient: [-6.60867061  0.09889824  7.01327985  0.21289918  0.09024365  0.05221481
  0.11725835  0.72642048 -0.03897314  0.19262718]
              precision    recall  f1-score   support

           0       0.88      1.00      0.93       716
           1       1.00      0.39      0.56       165

    accuracy                           0.89       881
   macro avg       0.94      0.69      0.75       881
weighted avg       0.90      0.89      0.86       881

Accuracy: 0.8853575482406356
F1 score: 0.5589519650655022
```

The intercept term from the above logistic regression model is the log odds of our base reference term. Odds of the event implies the times it is likely to occur than not. The 10 coefficients represent the log odds of the 10 attributes in order. The estimated coefficients are the log odds. By exponentiating these values, we can calculate the odds, which are easier to interpret.

We can also interpret that a unit increase in any attribute increases the odds of on time flight arrival exp(coefficient of that attribute) times.

There are four ways to check if the predictions are right or wrong:

1. **TN / True Negative**: the case was negative and predicted negative
2. **TP / True Positive**: the case was positive and predicted positive
3. **FN / False Negative**: the case was positive but predicted negative
4. **FP / False Positive**: the case was negative but predicted positive

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.

Precision:- Accuracy of positive predictions.

Precision = TP/(TP + FP)

In our case from the last figure we can see that the precision for class 0 (on time flights) is 0.88 i.e.88% and that the precision for class 1 (delayed flights) is 1 i.e.100%

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall:- Fraction of positives that were correctly identified.

Recall = TP/(TP+FN)

In our case from the last figure we can see that the recall for class 0 (on time flights) is 1 i.e.100% and that the precision for class 1 (delayed flights) is 0.39 i.e.39%

The accuracy of the model is 88.53% and the f1 score is 55.89%

Q4)Variables whose standardized coefficients have higher values are more significant than variables whose standardized coefficients have lower values.

From the previous figure in Q3 we can therefore say that variables having coefficients `0.09889824,0.21289918,0.09024365,0.05221481,0.11725835,-0.03897314, 0.19262718` are not significant and variables having coefficients `-6.60867061,7.01327985,0.72642048` are significant that is :

Significant variables : CRS_DEP_TIME,DEP_TIME, Weather

Non- significant variables : CARRIER, DEST, DISTANCE, FL_DATE, FL_NUM, ORIGIN,DAY_WEEK,DAY_OF_MONTH,TAIL_NUM

Q5)

On performing logistic regression considering only these significant variables, we get

```
intercept: [-1.56349814]
coefficient: [-7.47575387  7.91025487  0.70296561]
              precision    recall  f1-score   support

           0       0.87      1.00      0.93       354
           1       1.00      0.39      0.56        87

    accuracy                           0.88       441
   macro avg       0.93      0.70      0.75       441
weighted avg       0.90      0.88      0.86       441

Accuracy: 0.8798185941043084
F1 score: 0.5619834710743802
```

Accuracy of model as 87.98% and the f1 score as 56.20%

Therefore there is not much change in accuracy on comparing with previous model as expected.

Q6) [Got from code]

The ideal conditions for the highest chance of an ontime flight from DC to New York are: Date = 2/01/2004 and 9/01/2004 ; Carrier = 6 ; Weather = 0

BONUS QUESTIONS:

1.KAREN, BARF

2. No clever manipulation of the data can improve inference

Theorem. If $X \rightarrow Y \rightarrow Z$, then the

$I(X; Y) \geq I(X;Z)$, $I(Y;Z) \geq I(X;Z)$.

Equality iff $I(X; Y|Z) = 0$.

3. Rule of Two

4.C-3PO and R2-D2