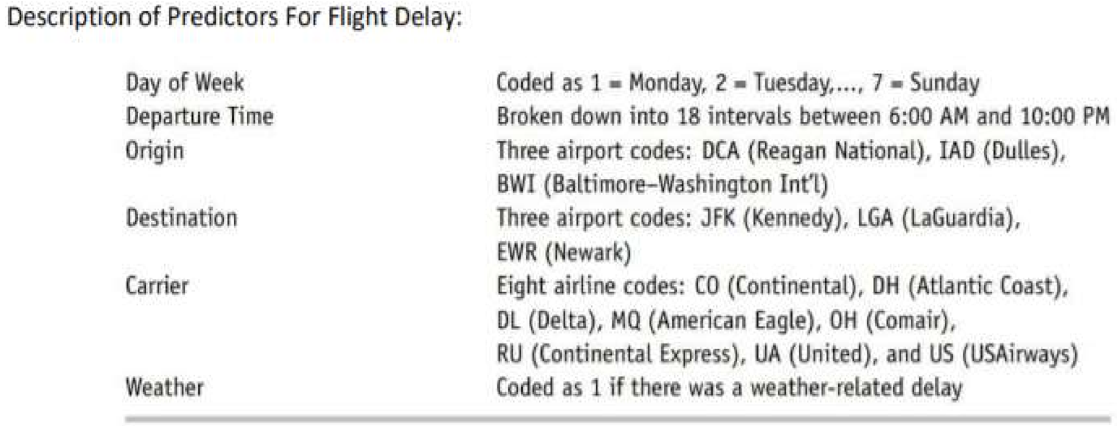
I declare that I have done the assignment on my own, and applied concepts to the best of my understanding and have not copied the report or code from anyone/anywhere.

Satwik Murarka
190020101

Analysis of Flight Delays

Assignment 1 (DS 303)

The objective of this report is to analyse flight delays and predict future delays using data of all flights from **Washington DC to New York** during **January 2004**.
Firstly we review the dataset to find relations between predictors and the final output, then we pre-process the data, apply logistic regression to predict the outcome for future flights.Below is the description for the flight predictors-

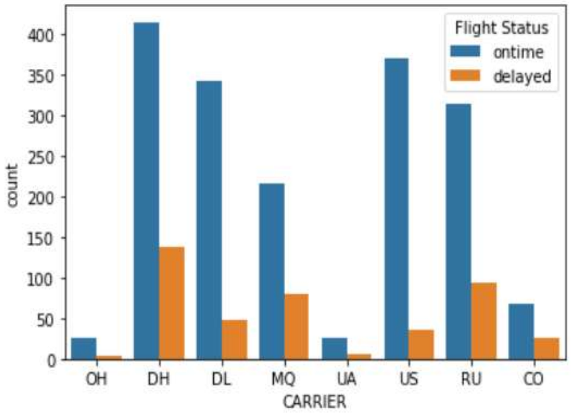


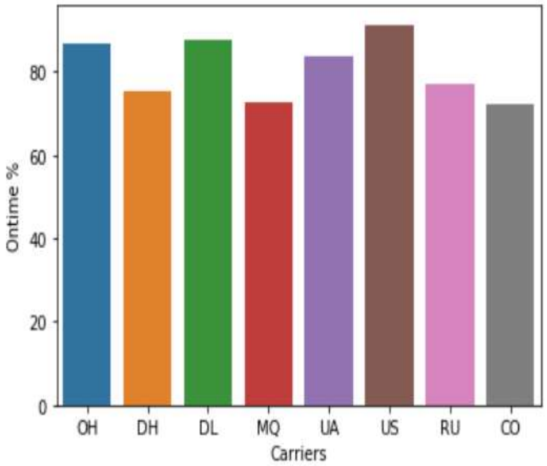
Description of Predictors For Flight Delay:

Day of Week	Coded as 1 = Monday, 2 = Tuesday,..., 7 = Sunday
Departure Time	Broken down into 18 intervals between 6:00 AM and 10:00 PM
Origin	Three airport codes: DCA (Reagan National), IAD (Dulles), BWI (Baltimore–Washington Int'l)
Destination	Three airport codes: JFK (Kennedy), LGA (LaGuardia), EWR (Newark)
Carrier	Eight airline codes: CO (Continental), DH (Atlantic Coast), DL (Delta), MQ (American Eagle), OH (Comair), RU (Continental Express), UA (United), and US (USAirways)
Weather	Coded as 1 if there was a weather-related delay

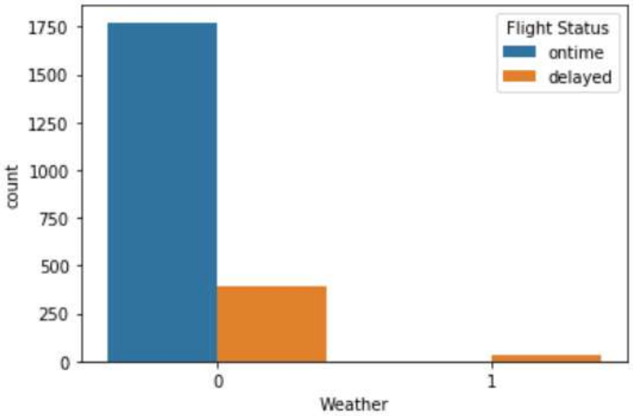
Q1

Firstly we check for null values in the dataset and we find that it does not have any null values. Then from the **2201** flights we find that **1773 (80.45%)** were on time and **428 (19.45%)** were delayed.
The highest number of flights and highest number of delays are of the carrier DL.

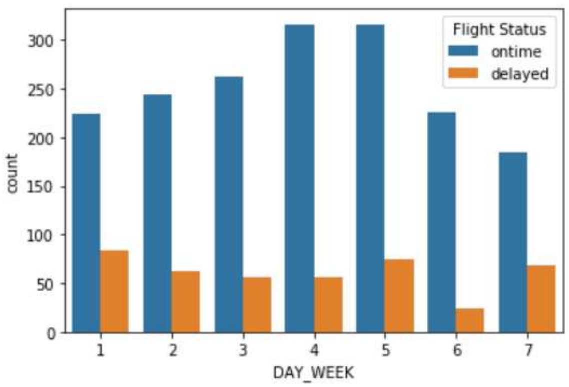


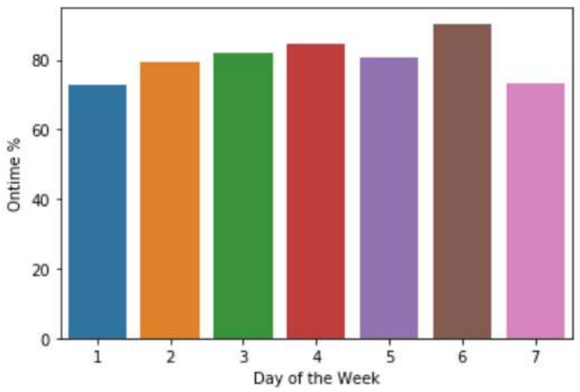


The highest percent of ontime flights are by the carrier **US (91.33%)** and the carrier **DL (87.88%)** and **OH (86.66%)** are just behind.

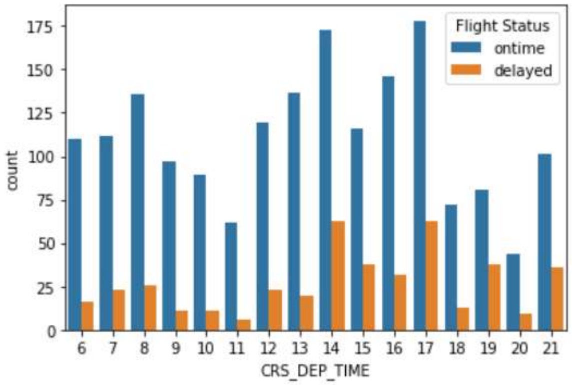


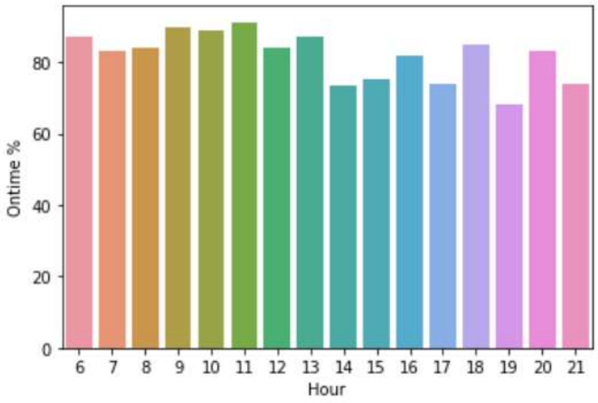
From this plot we infer that the number of weather delays are very less but it always results in a delayed flight and so weather being bad has a significant influence on flight delays.





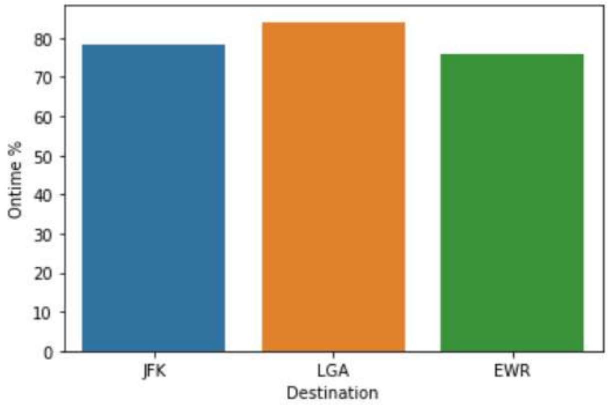
The frequency of flights is highest for **Thursday and Friday** and the highest percent of ontime flights is on **Saturday (90.40%)**.

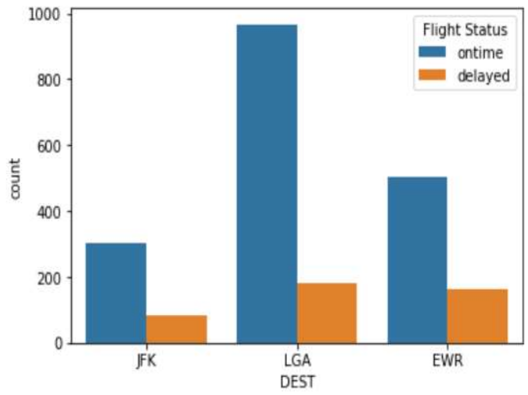


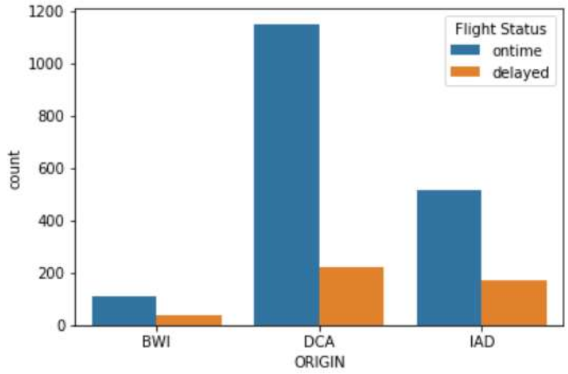


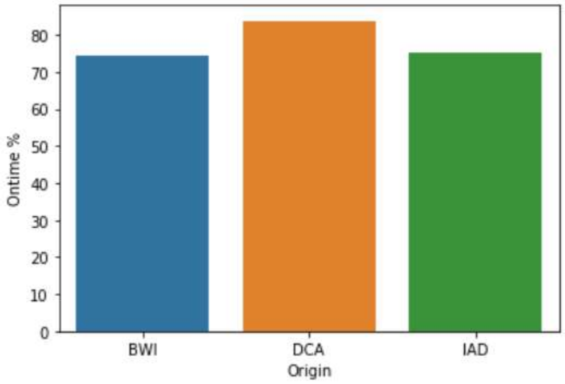
Most flights are during the afternoon hours from **2 to 5 pm** and the least percent delay is observed for flights during **9 to 12 am (90%)**

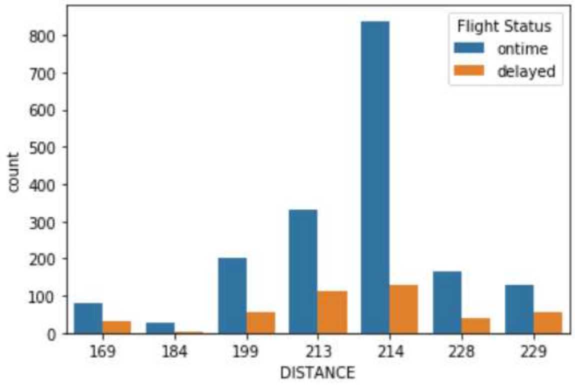
Similarly we get the following plots for the rest of the predictors-

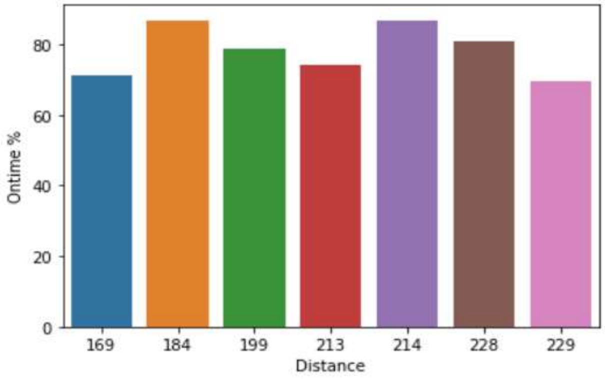












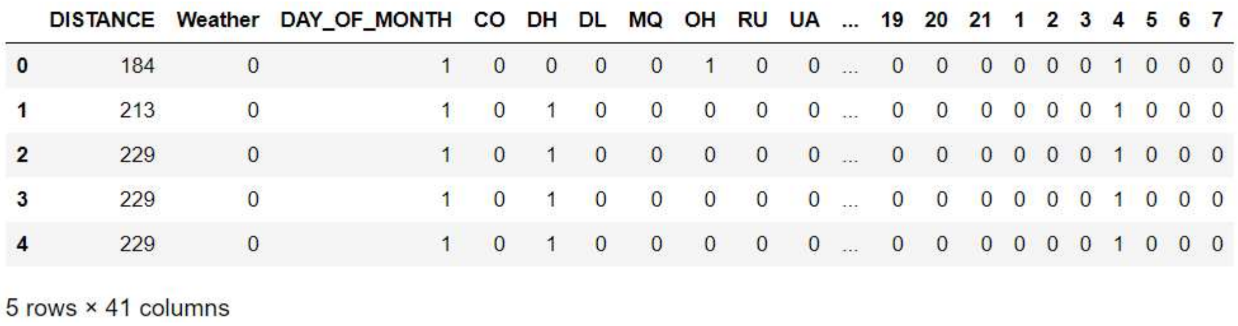
From these we infer that most number,highest percent ontime flights originated from **DCA** and landed at **LGA**. Similarly the distance was **214.**

Some statistics to describe the data-

	MIN	MAX	AVG
CRS_DEP_TIME	600	2130	1371.93
DEP_TIME	10	2330	1369.29

Q2

No dummy variables present. We generate dummy variables for the categorical variables namely **[DEST, CARRIER, DAY_WEEK, ORIGIN, Flight Status, CRS_DEPT_TIME].**
We also drop some variables which have a large number of unique values like **[TAIL_NUM,FL_NUM]** making it difficult for analysis.
Head of final data frame for logistic regression-

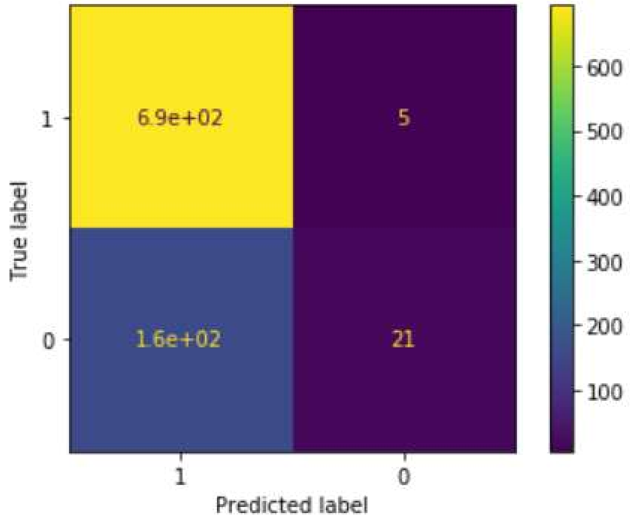


	DISTANCE	Weather	DAY_OF_MONTH	CO	DH	DL	MQ	OH	RU	UA	...	19	20	21	1	2	3	4	5	6	7
0	184	0	1	0	0	0	0	1	0	0	...	0	0	0	0	0	0	1	0	0	0
1	213	0	1	0	1	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
2	229	0	1	0	1	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
3	229	0	1	0	1	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
4	229	0	1	0	1	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0

5 rows × 41 columns

Q3

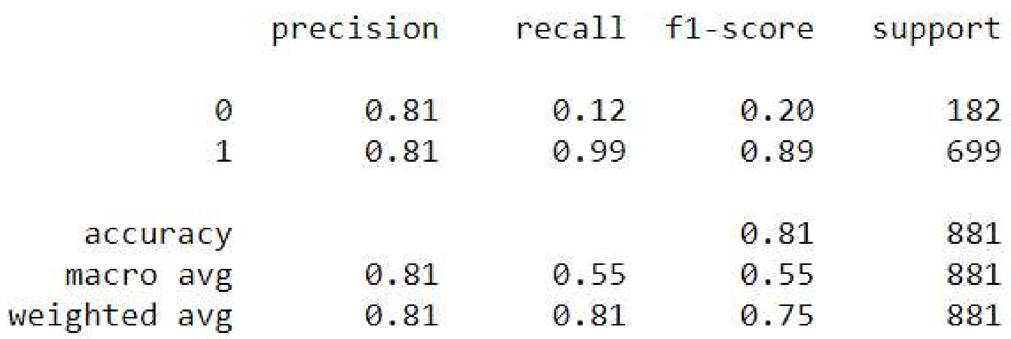
Applied logistic regression by dividing the dataset into (60:40) and obtained the following confusion matrix-



Out of the **880** test samples, **693** are true positives **21** are true negative, **161** are false negative and **5** are false positive.

The accuracy score is around **81%.**

The classification report is as follows-

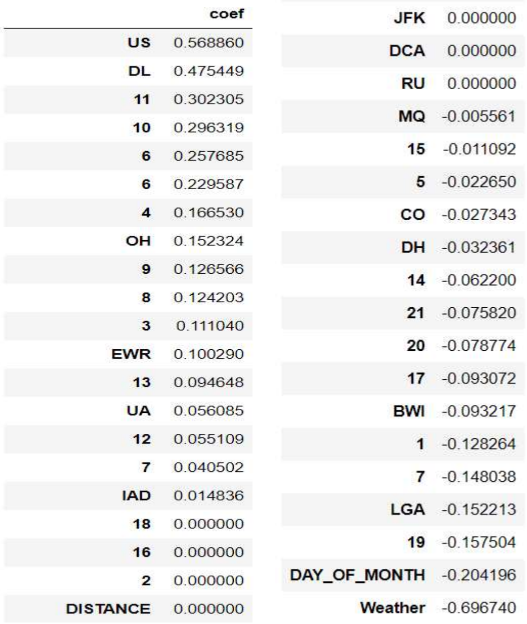


```
              precision    recall  f1-score   support

           0       0.81      0.12      0.20       182
           1       0.81      0.99      0.89       699

    accuracy                           0.81       881
   macro avg       0.81      0.55      0.55       881
weighted avg       0.81      0.81      0.75       881
```

Precision for delayed and non delayed flights are **81%**.

Q4

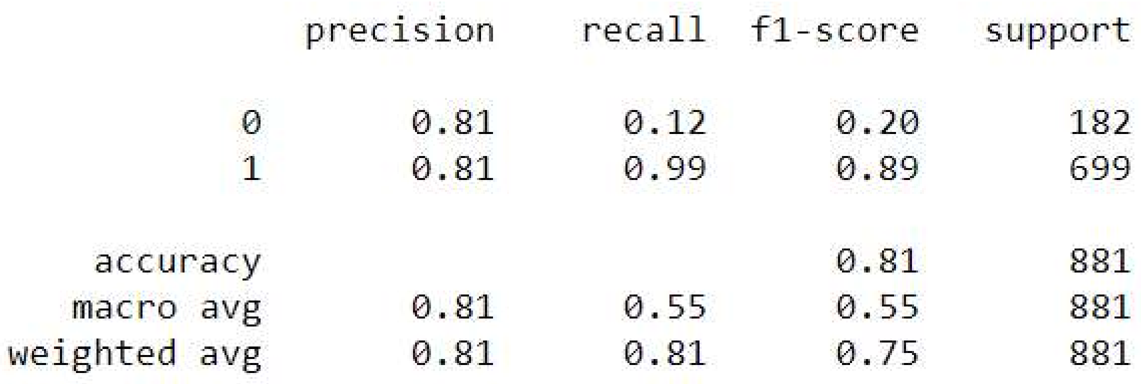
We find the coefficients of the model and remove those whose absolute values are very less as they contribute very less to the final to the final classification but increase the computations of the system.(Helpful when dealing with very large datasets.)



	coef
US	0.568860
DL	0.475449
11	0.302305
10	0.296319
6	0.257685
6	0.229587
4	0.166530
OH	0.152324
9	0.126566
8	0.124203
3	0.111040
EWR	0.100290
13	0.094648
UA	0.056085
12	0.055109
7	0.040502
IAD	0.014836
18	0.000000
16	0.000000
2	0.000000
DISTANCE	0.000000
JFK	0.000000
DCA	0.000000
RU	0.000000
MQ	-0.005561
15	-0.011092
5	-0.022650
CO	-0.027343
DH	-0.032361
14	-0.062200
21	-0.075820
20	-0.078774
17	-0.093072
BWI	-0.093217
1	-0.128264
7	-0.148038
LGA	-0.152213
19	-0.157504
DAY_OF_MONTH	-0.204196
Weather	-0.696740

Q5

After variable selection we repeat the process of logistic regression and obtain the following classification report-



```
              precision    recall  f1-score   support

           0       0.81      0.12      0.20       182
           1       0.81      0.99      0.89       699

    accuracy                           0.81       881
   macro avg       0.81      0.55      0.55       881
weighted avg       0.81      0.81      0.75       881
```

The new model is also **81%** accurate and has the same precision accuracy as before.This helps in **reducing the computations** which helps us in larger datasets.

Also we infer that removing variables(data) which do not have high correlation with the output **does not make the model lose accuracy**.

Q6

Ideal weather conditions for an ontime flight from New York to DC are-

Carrier- US(USAirways)

Weather-0(No weather delay)

Time- 9 am to 12 pm

Day- Saturday(6)

This can be inferred from the percentage graphs and the combination of these attributes will have the highest chance to be ontime.

BONUS QUES-

Q1 AI made by Tony Stark - TADASHI

Q4 Robotic Duo - R2-D2 and C-3PO