ML_Assignment 1

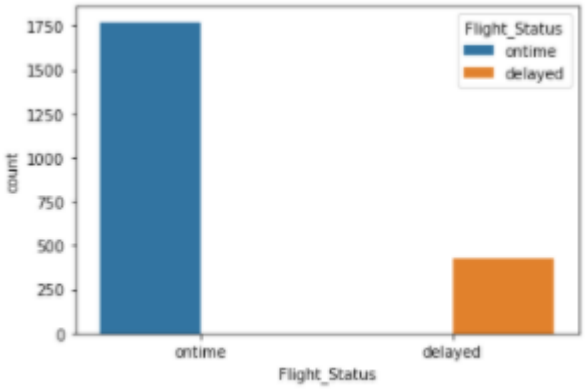
203310012

Q1) 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.)

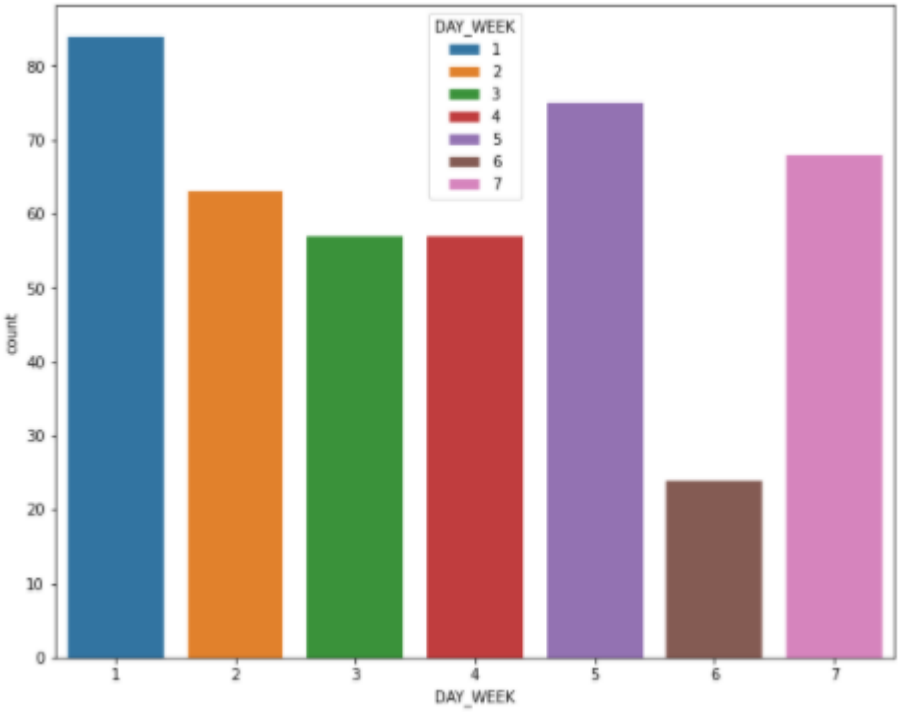
Ans 1)

EXPLORATORY DATA ANALYSIS:

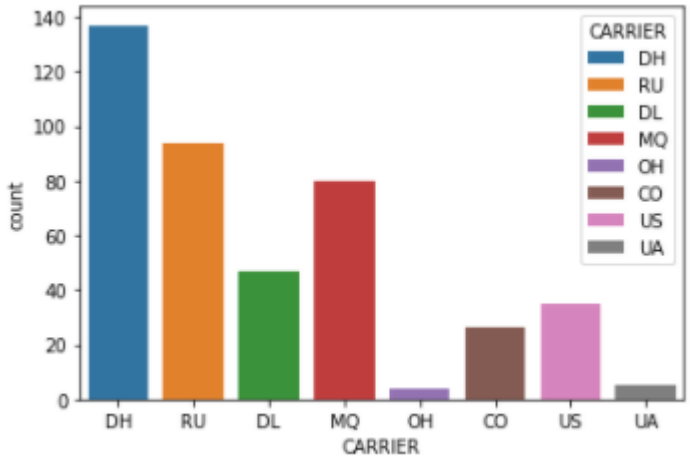
This is a bar chart showing the total number of delayed and on time flights. We can see that on time flights are much more than delayed flights and we can say that the data is imbalanced.



This bar chart shows that the number of flights delayed on each day of the week. We can see that most of the flights are delayed on Mondays and least on Saturday. This may be because less people travel on Saturday since Sunday is a holiday.



The following bar chart shows delayed flights according to the carrier. We can see that DH(Atlantic Coast) has the maximum number of delayed flights whereas OH(Comair) has the least number of delayed flights.



Q2) 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.

Ans.2) Code attached in zip folder. Accuracy achieved upto 89 percent.

Q3) Interpret the model and coefficients and present some insights.

Ans.3) After one hot encoding the feature variables and label encoding of the target variable that is Flight_Status, and finding the coefficients for the columns, I concluded that the coefficients were present in the range between (). Now, the coefficients with high magnitude were important to us. So, I deleted the variables with very low coefficients magnitudes. This significantly increased the magnitude.

Q4) 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?)

Ans.4)

	coef
CRS_DEP_TIME	2.281249
FL_DATE_01/01/2004	0.662677
TAIL_NUM_N225DL	0.525754
FL_NUM	0.407731
CARRIER_US	0.374244
FL_DATE_1/21/2004	0.305341
TAIL_NUM_N744UW	0.277676
TAIL_NUM_N748UW	0.271689
TAIL_NUM_N724MQ	0.262239
TAIL_NUM_N733UW	0.255347
TAIL_NUM_N675BR	0.255218

	coef
TAIL_NUM_N11150	-0.302263
TAIL_NUM_N684BR	-0.333535
CARRIER_MQ	-0.337913
TAIL_NUM_N667BR	-0.344755
CARRIER_DH	-0.350688
FL_DATE_1/18/2004	-0.404967
Weather	-0.765696
DEP_TIME	-2.937174

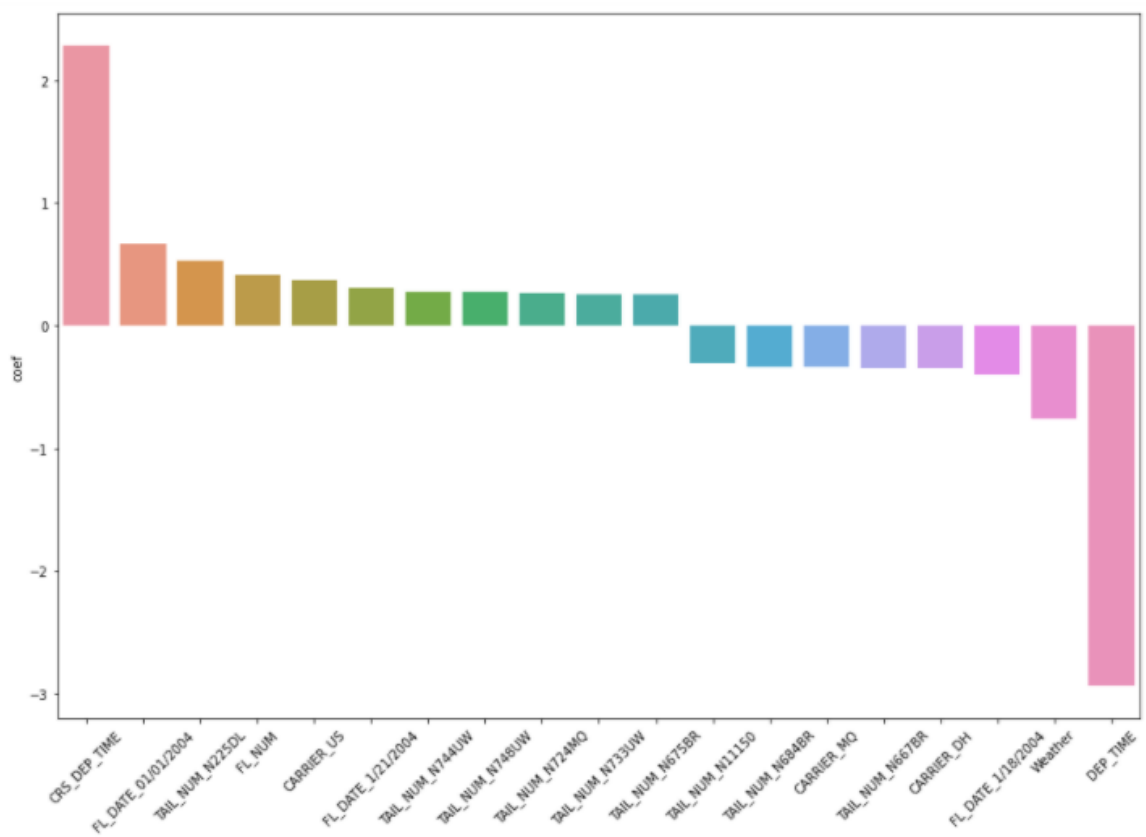


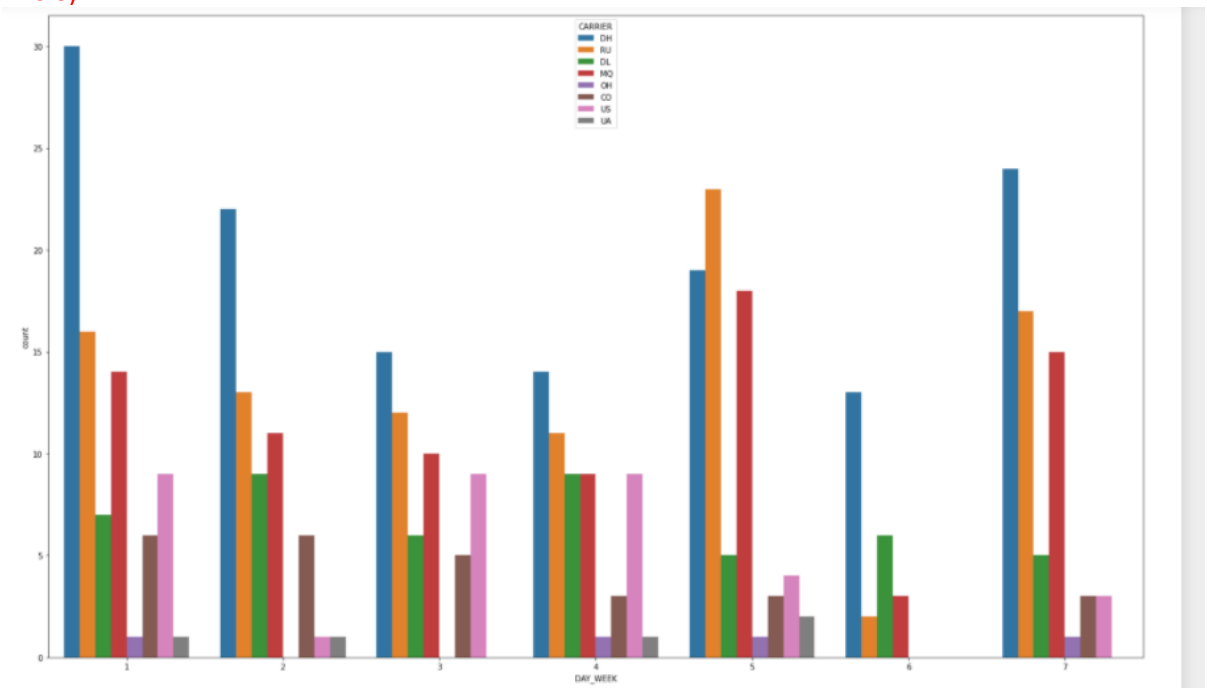
Fig : Graph showing coefficients on y-axis and feature variable on x-axis

Q5) Conclude the analysis by fitting a new model on these selected variables and report the same. Report the accuracy.

Ans 5) Accuracy achieved 88.5%

Q6) Find the ideal weather conditions for the highest chance of an ontime flight from DC to New York . (weather, time, day, carrier)

Ans.6)



From the figure, we can see that the least number of flights delayed are on day 6 (saturday)
And the carriers that are delayed are : DH, RU, DL, MQ.
Therefore, we should prefer any of the remaining flights, i.e., OH, UA, US, CO.

	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	TAIL_NUM	Flight_Status
138	700	US	655	LGA	214	2160	DCA	0	6	3	N760UW	ontime
139	900	US	858	LGA	214	2164	DCA	0	6	3	N710UW	ontime
140	1100	US	1059	LGA	214	2168	DCA	0	6	3	N760UW	ontime
141	1300	US	1256	LGA	214	2172	DCA	0	6	3	N710UW	ontime
142	1500	US	1500	LGA	214	2176	DCA	0	6	3	N760UW	ontime
143	1700	US	1658	LGA	214	2180	DCA	0	6	3	N710UW	ontime
144	1900	US	1857	LGA	214	2184	DCA	0	6	3	N760UW	ontime
667	700	US	655	LGA	214	2160	DCA	0	6	10	N722UW	ontime
668	900	US	857	LGA	214	2164	DCA	0	6	10	N750UW	ontime
669	1100	US	1056	LGA	214	2168	DCA	0	6	10	N722UW	ontime
670	1300	US	1256	LGA	214	2172	DCA	0	6	10	N750UW	ontime
671	1500	US	1458	LGA	214	2176	DCA	0	6	10	N722UW	ontime
672	1700	US	1656	LGA	214	2180	DCA	0	6	10	N750UW	ontime
673	1900	US	1854	LGA	214	2184	DCA	0	6	10	N722UW	ontime
1186	700	US	654	LGA	214	2160	DCA	0	6	17	N736UW	ontime
1187	900	US	859	LGA	214	2164	DCA	0	6	17	N748UW	ontime
1188	1100	US	1100	LGA	214	2168	DCA	0	6	17	N736UW	ontime
1189	1300	US	1259	LGA	214	2172	DCA	0	6	17	N748UW	ontime
1190	1500	US	1456	LGA	214	2176	DCA	0	6	17	N736UW	ontime
1191	1700	US	1658	LGA	214	2180	DCA	0	6	17	N748UW	ontime
1192	1900	US	1854	LGA	214	2184	DCA	0	6	17	N736UW	ontime

The above dataframe shows all the possible cases for day 6 flights filtered for only the 4 carriers from DCA to LGA.
Therefore we conclude that carrier US is the best. Weather – 0(no weather related delay)

BONUS MARKS

Q1. [1 Mark] Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY and EDITH.
Ans.) JOCASTA and TADASHI

Q2. [2 Mark] Explain the Data processing inequality.
Ans.)

Q3. [1 Mark] In Star Wars Universe, X was a Sith philosophy mandating that only two Sith Lords could exist at any given time: a master to represent the power of the dark side of the Force, and an apprentice to train under the master and one day fulfill their role.? What is X?
Ans.) The Rule of Two

Q4. [1 Mark] In Star Wars Universe, name this robotic duo:
Ans.) C-3PO and R2-D2

Q5 [1 Mark] What is special about Cards against Humanity: Black Friday 2019? (Hint: It's related to AI)
Ans.) The employees of Cards against humanity taught a computer to write it. And then competed with the computer for straight 16 hours. The competition was won by humans.