Assignment 1
GNR 652 Marks 40 [+2]

Guidelines about the assignment:

- Submit the assignment report as a pdf only detailing the analysis and answers. No code in the report.
- Submit the code alongwith (in Julia/R/Python) in separate files.
- Submit all the files as a zip with your roll no as name. Eg: “120010012.zip”
- Any doubts regarding the assignment will be clarified upto 1 week before the deadline via moodle. Do NOT wait till last moment to start the assignment. Doubts will not be addressed thereafter.
- Do NOT submit assignments as late. The deadline will close irrespective of internet delays or any excuse. Ensure to submit your work well before the deadline. No work submitted afterwards will be evaluated, at all.

You are a consultant at KPMG Denver, assigned with the task of analysing flight delays for the airport authorities, to assist them in revamping their operations. Predicting flight delays can be useful to a variety of organizations: airport authorities, airlines, aviation authorities. At times, joint task forces have been formed to address the problem. Such an organization, if it were to provide ongoing real-time assistance with flight delays, would benefit from some advance notice about flights likely to be delayed.

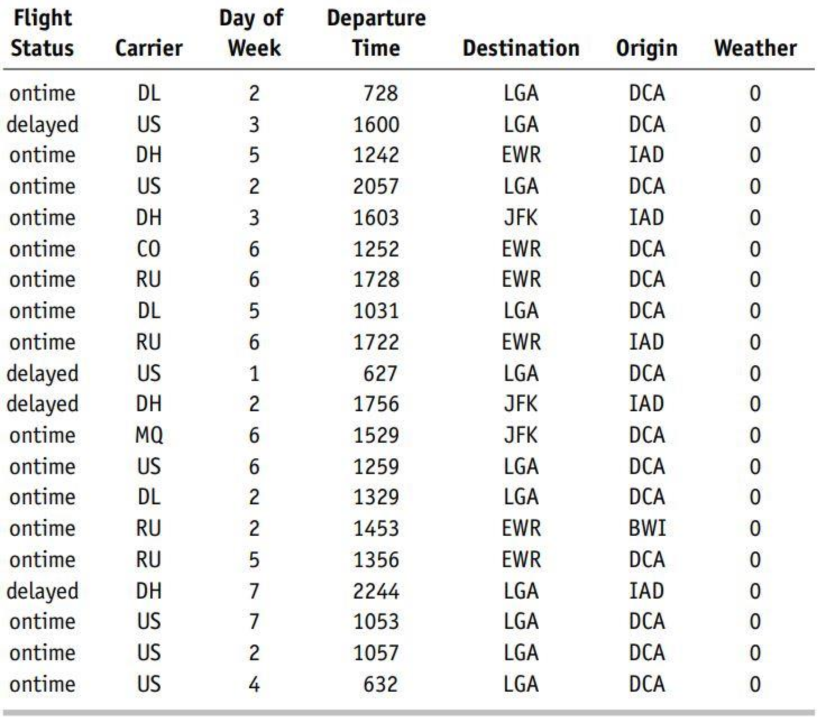
The outcome of interest is whether the flight is delayed or not (*delayed* means more than 15 minutes late). Our data consist of all flights from the Washington, DC area into the New York City area during January 2004. The percent of delayed flights among these 2201 flights is 19.5%. The data were obtained from the Bureau of Transportation Statistics website (www.transtats.bts.gov).

You decide to interpret the problem as a Machine Learning problem and define the goal as to predict accurately whether a new flight, not in this dataset, will be delayed or not. The outcome variable is a variable called Flight Status, coded as *delayed* or *on time*.

Through this analysis you want to provide your recommendations to the Committee, which is not an expert in Machine Learning. The objective of this assignment is to understand that a ML model with high accuracy is often not the only objective, and qualitative analysis is required and reported for ease of understanding.

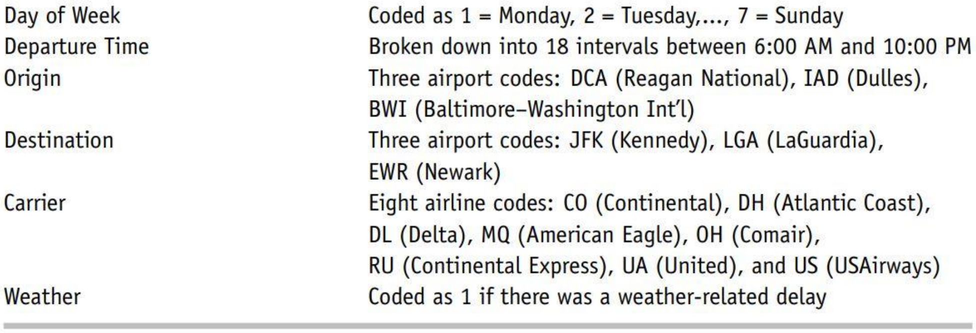
Sample of dataset:

Assignment 1
GNR 652 Marks 40 [+2]



Flight Status	Carrier	Day of Week	Departure Time	Destination	Origin	Weather
ontime	DL	2	728	LGA	DCA	0
delayed	US	3	1600	LGA	DCA	0
ontime	DH	5	1242	EWR	IAD	0
ontime	US	2	2057	LGA	DCA	0
ontime	DH	3	1603	JFK	IAD	0
ontime	CO	6	1252	EWR	DCA	0
ontime	RU	6	1728	EWR	DCA	0
ontime	DL	5	1031	LGA	DCA	0
ontime	RU	6	1722	EWR	IAD	0
delayed	US	1	627	LGA	DCA	0
delayed	DH	2	1756	JFK	IAD	0
ontime	MQ	6	1529	JFK	DCA	0
ontime	US	6	1259	LGA	DCA	0
ontime	DL	2	1329	LGA	DCA	0
ontime	RU	2	1453	EWR	BWI	0
ontime	RU	5	1356	EWR	DCA	0
delayed	DH	7	2244	LGA	IAD	0
ontime	US	7	1053	LGA	DCA	0
ontime	US	2	1057	LGA	DCA	0
ontime	US	4	632	LGA	DCA	0

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

Questions [Maximum 40 Mark]:

[6Marks] 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.)

[10Marks] 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.

[8Marks] Q3) Interpret the model and coefficients and present some insights.

[7 Marks] 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?)

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

[2Marks] Q6) Find the ideal weather conditions for the highest chance of an ontime flight from DC to New York . (weather, time, day, carrier)

Assignment 1

GNR 652 Marks 40 [+2]

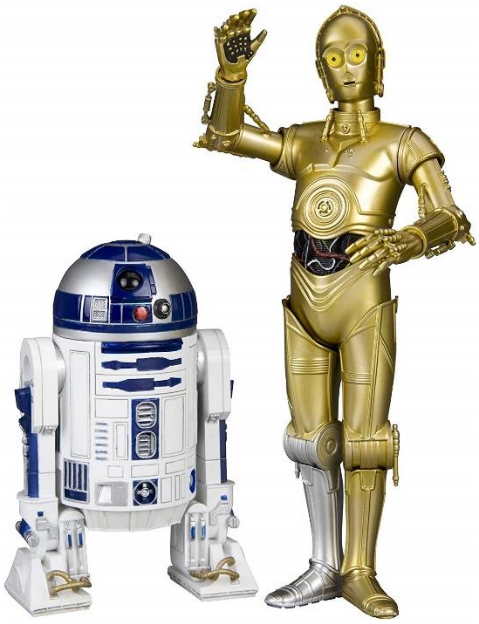
BONUS [Maximum 2 Mark]:

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

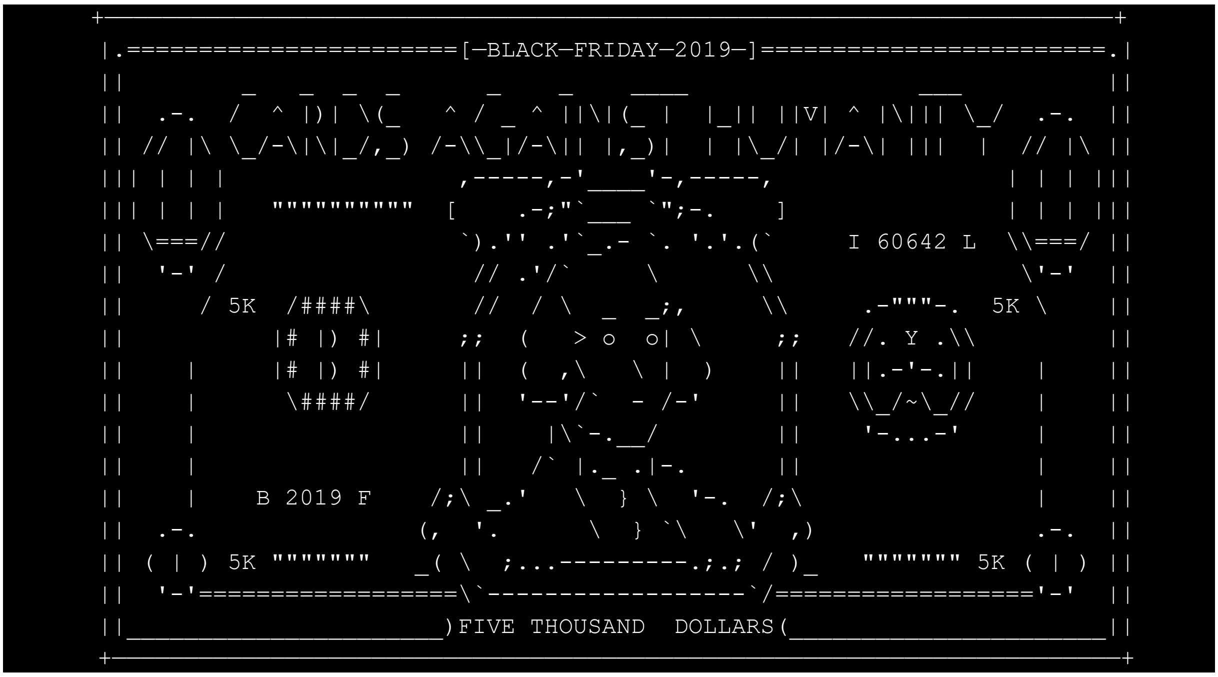
Q2. [2 Mark] Explain the Data processing inequality.

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?**

Q4. [1 Mark] In Star Wars Universe, name this robotic duo:



Q5 [1 Mark] What is special about Cards against Humanity: Black Friday 2019? (Hint: It's related to AI)



```
+-----------------------------------------------------------------+
|.=======================[-BLACK-FRIDAY-2019-]========================.|
||         _   _   _   _       _     _     ____               ___        ||
||  .-.   /  ^ |)| \(_    ^ / _ ^ ||\|(_ |   |_|| ||V| ^ |\||| \_/   .-.  ||
|| // |\ \_/-\|\|_/,_) /-\\_|/-\|| |,_)|   | |\_/| |/-\| |||   |   // |\ ||
||| | | |                ,-----,-'____'-,-----,                  | | | |||
||| | | |   """"""""""   [     .-;"`___  `";-.     ]                | | | |||
|| \===//                `).'' .'`_.- `. '.'.(`      I 60642 L  \\===/ ||
||  '-' /                 // .'/`      \       \\                \'-'  ||
||      / 5K  /####\       //  / \  _  _;,      \\      .-"""-.  5K \    ||
||           |# |) #|     ;;  (   > o  o| \      ;;    //. Y .\\         ||
||     |     |# |) #|     ||  (  ,\   \ |  )     ||    ||.-'-.||    |    ||
||     |      \####/      ||  '--'/`  - /-'      ||    \\_/~\_//    |    ||
||     |                  ||     |\`-.__/        ||     '-...-'     |    ||
||     |                  ||    /` |._ .|-.      ||                 |    ||
||     |    B 2019 F    /;\ _.'    \  } \  '-.  /;\                 |    ||
||  .-.               (,   '.       \  } `\    \'  ,)               .-.  ||
|| ( | ) 5K """""""   _( \  ;...---------.;.; / )_   """"""" 5K ( | ) ||
||  '-'===================\`------------------`/===================='-'  ||
||_________________________)FIVE THOUSAND  DOLLARS(_________________________||
+-----------------------------------------------------------------+
```