Instructions:

- You are required to read the datasets uploaded in grader to answer the questions.
- This is an open book exam
- After completing, you are required to upload a zipfile containing both R and Python code with proper comments on Grader.
- The file naming convention is Name_20180421_Batch42_CSE7212c_CUTe.zip

Max Marks: 50
Duration: Total 2 Hrs for R and Python

Section I: R

1. Create a vector that contains a sequence of numbers from 1000 to 10000 and filter out only those elements which are divisible by 7 and 19 [2 Marks]

2. Create an R function that takes an integer "n" as the input and returns the nth number in the fibonacci sequence [3 Marks]

- A Fibonacci sequence is characterized by the fact that every number after the first two is the sum of the two preceding ones. Ex: 1, 1, 2, 3, 5, 8, 13, 21, 34, 55….

About the data:

The data set is a students' dataset, containing data of top 50 students of 5 different colleges.
The data contains info about the students' academic performance, personality type, extra-curricular activities, their overall score measured upon various other parameters including the afore-mentioned scores. The data also contains their placement information in either private/public sector companies.

Data sets Description:

Files COL1.csv through COL5.csv:

These contain data related to top 50 students of colleges 1 to 5. Columns are:

- CollegeID – ID of the college
- Student – ID of the student
- Acad_Score : Academic score of the student (some colleges provided the score as percentage, while some others provided the data as CGPA)
- BehaviourType: Indicates the Myers-Briggs personality type of the students
- Extra_Curr: Indicates the one important extra-curricular activity in which the student is best at
- Overall_Score: An overall score is measured upon various parameters (unknown to us), including their above scores.

Questions:

3. Read the files relating to top 50 students from all the 5 colleges [2 Marks]

4. Observe the data distribution of each column in all the 5 dataframes. For each college,

a. Report if the summary of the variables [2 Marks]

b. Also report if any one category in categorical attribute has dominance over other categories. Hint: Can this be answered by observing the counts of each level [2 Marks]

c. Find attributes that have missing values [2 Marks]

d. Report how many missing values are present in each file [1 Mark]

5. In each of the files fill missing values and explain the imputation strategy [3 Marks]

6. Combine all those data frames into a single, consolidated data frame, and name it as "consolidated_data" [2 Marks]

7. Do range normalisation of the numeric variables [1 Marks]

Section II: Python

1. Write code which accomplishes the following tasks:

a. Create a list of integers from 1 to 100. Using list comprehension, generate a list of squares of the above integers if they are odd. [2 Marks]

b. Write a function that returns the number of vowels in a given string. [3 Marks]

2. Write a function in python that takes a string as an input and returns the sum of the numbers corresponding to each alphabets position from a to z. (Hint: can use a dictionary, handle cases of the input string) [5 Marks]

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

Ex: Regards ---> 18 + 5 + 7 + 1 + 18 + 4 + 19 = 72

3. Read the 'bank.csv' file as DataFrame with Pandas and answer/solve the following questions.

a. Reading in the data, importing modules [2 Marks]

b. Structure, Summary, Head and Tail of the data [2 Marks]

a. What is the most common occupation in the 'job' column? [3 Marks]

b. What is the mean 'balance' for people with marital status as married and also education as tertiary? [2 Marks]

c. What is the mean 'age' for different 'education' categories? [2 Marks]

d. Drop the following columns from the dataframe: job, day, month, pdays, previous, poutcome. [2 Marks]

e. Recode categorical variables with 2 levels as 0 & 1. [3 Marks]

f. Replace categorical variables with more than 2 levels with their respective dummy variables. [2 Marks]

g. Standardize the numeric columns. [2 Marks]