Please answer the following questions and save your answers in a public GitHub repository. You have 24 hours to submit your answer.

1) Use the table below for problem 1 a - c
   a) Based on the following two tables, write a SQL query that returns the name and student ID of all students that have a higher total marks score than the student that has StudentID of 'V002'

   b) Assume that the two tables are pandas data frame variables. Based on those two data frames--utilizing pandas--write a python function that returns a new data frame version of name_table, where each name containing the letter "e" is uppercased, and lowercased otherwise (e.g. "Edward" → "EDWARD", "Bob" → "bob").

   c) Now write a function that takes in the output of 1) b) and mark_table and returns a data frame that summarizes the average grade of uppercase names and lowercase names

name_table

StudentID	Name
V001	Abe
V002	Abhay
V003	Acelin
V004	Adelphos

mark_table

StudentID	Total_marks
V001	95
V002	80
V003	74
V004	81

2) **Consider the data set below. Write some python code that illustrates some common feature engineering and/or data preparation tasks.**

https://raw.githubusercontent.com/mathcoder3141/blog-data-files/master/Congress_White_House.csv

https://github.com/helloworlddata/white-house-salaries/blob/master/data/converted/2017.csv

Consider the file "data.csv" in the following GitHub repository. What are some descriptive statistics about this set? What can you say about the distribution of this data?

https://github.com/fractalbass/data_engineer

No code is necessary for the following questions:

3) **If you were asked to impute null values in a column of a file that was 365 Gigabytes, what would you do? What tools would you use? What tools would you NOT use?**

4) **What would you do if you were asked to do the above task every Thursday morning at 2:00am?**

5) **Who is your favorite mathematician, statistician or computer scientist and why?**

Thanks for taking the time to participate in this exercise!