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'

Done in 1a.SQL

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”).

Done in 1bc.py

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

Done in 1bc.py

name_table mark_table

StudentID	Name
V001	Abe
V002	Abhay
V003	Acelin
V004	Adelphos

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:

Done in part_2.ipynb

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?

The easiest way would be to **drop the null values**, but we can also **substitute missing values with a statistical value for each column**. I would **use Dask library** for it, which is a tool that allows you to manipulate big data by using similar interface to numpy arrays and pandas data frames. **I would not use numpy or pandas** to calculate a statistical value because I cannot load all the data to RAM. Dask allows to do lazy computations (compute values only when we need them); also it automatically loads data from the drive and compresses it for faster access, and it uses multithreading.

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

It highly depends on the services I am allowed to use. I can schedule tasks in AWS, Azure, and Airflow services. There is the "shedule" library even in pure python, which allows to run code at specific time every day.

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

My favorite computer scientist is Donald Knuth because I started my programming path with his books. He also contributed a lot in the computer science field by creating Tex, popularazing the asymptotic notation, and lots of other innovations.

Thanks for taking the time to participate in this exercise!