**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

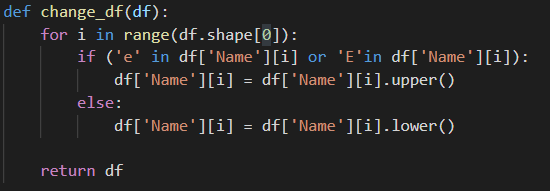
a) Based on the following two tables, write a SQL query that returns the name and student ID of all students that have a higher total marks score than the student that has StudentID of 'V002’

Answer:

SELECT name\_table.Name, name\_table.StudentID FROM name\_table INNER JOIN mark\_table on name\_table.StudentID = mark\_table.StudentID WHERE mark\_table.Total\_marks > (Select mark\_table.Total\_marks FROM mark\_table WHERE StudentID = '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”).

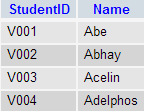
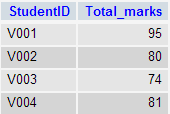
Answer:



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

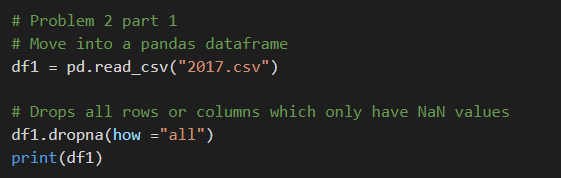


name\_table mark\_table

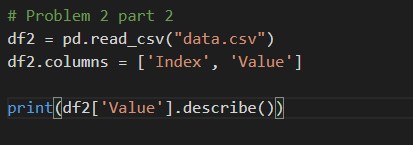
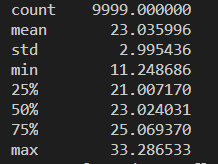
1. **Consider the data set below. Write some python code that illustrates some common feature engineering and/or data preparation tasks.**

<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:

1. **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?**

I would store the data from the file in a database and then make changes to the data. Using other big data analytics tools such make such changes.

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

I would setup a script which runs every Thursday morning which would impute null values in a column of a file.

1. **Who is your favorite mathematician, statistician or computer scientist and why?**

My favorite computer scientist is Guido van Rossum, the creator of python. Python is one of the most powerful programming languages which is incredibly versatile. Python can be used in many fields of computer science such ranging from web development to machine learning.

**Thanks for taking the time to participate in this exercise!**