#### Part 1: Python Programming

**Overview**

Your understanding of how python dictionaries work will later form the basis of its applications during data analysis with python.

In this part of the Independent project, you will be required to perform the following procedures using the given contact\_book dictionary. Initially, this dictionary will have the following contacts as keys and values.

contact\_book = {'Parul' : 'parul@moringa.org',

'Thomas' : 'thomas@moringa.org',

'Ashley' : 'ashley@moringa.org',

'Kellen' : 'kellen@moringa.org',

'June' : 'june@moringa.org',

'Joseph' : 'joe@moringa.org',

'Lillian' : 'lillian@moringa.org',

'Arnold' : 'arnold@moringa.org'

}

You will write python code that performs the following procedures

* + 1. Delete a contact from the dictionary when the user specifies the its key
    2. Print out the first 2 contacts.
    3. Display the total no. of contacts left in the dictionary.
    4. Add 2 new contacts in the dictionary.
    5. Print out all the contacts.

**Deliverable**

* + 1. You will create an Colaboratory Notebook that will contain your responses to the above questions.
    2. Use the following naming convention for the above Google Colaboratory Notebook;
       - "Moringa\_Data\_Science\_Prep\_W2\_Independent\_Project\_2019\_06\_FirstName\_LastName\_Python\_Notebook"
    3. Have the above document in one folder that you will submit on the LMS.

#### Part 2: SQL Programming

**Overview**

In this part of the assessment, you will act as a Data analyst to answer a research question on the US elections.

First, will be required to write a Data Report on the process that will undertake while working on the given research problem. Using the CRISP-DM Methodology, you will document the various stages of the data science lifecycle for the given research problem while also providing your recommendation.

You have been provided with a detailed description of what you will be expected to do below.

Do remember that complex tasks that may seem hard at first can often be broken down into a sequence of simple tasks, and there are workarounds to do what first seems “impossible” with a succession of known operations.

**Problem Description**

You have been asked to help a candidate to become US president. The winner of the election will be the candidate winning the most grand electors.

Grand electors are attributed at the state level: in each of the 51 states, there is a given number of grand electors to win (roughly, but not exactly, proportional to the size of the state) and the presidential candidate receiving the most local votes wins ALL the Grand Electors in that state.

Because the number of grand electors is not exactly proportional to the population, some states can be prioritized to increase the return on investment of the campaign. We assume here there are only 2 candidates, and no history (no trend of certain states to vote for a particular candidate or party). Hence, each vote is equally "expensive" to get, but some states grant more grand elector per capita.

You are provided with 2 tables: one giving the number of Grand Electors per state, the other the population per state.

You are asked to identify the states that should be prioritized to win the election, with a smart but simple algorithm (brute force computation of all possibilities to find the absolute optimum is not accepted, as it would be to computationally expensive). It is ok not to find the overall optimum, we just want a strategy that makes sense.

(This is of course a very simplistic view of reality, but it is a nice starting point to play with data and analyze possibilities).

First take a few minutes to think about what you need to do. Steps will be suggested hereafter, but a big part of the data scientist's job is to identify the flow of operations when being asked a practical question, so it is important you start exercising on that in addition to programming.

Here is what we are suggesting to do: we will rank states by decreasing number of grand electors per capita. The first states in the list will be the most valuable (you get a large number of grand electors by convincing a small number of people to vote for you). We will target all the states at the top of the list until the cumulative sum (also called running total) of grand electors won is larger than half the total number of Grand Electors in the country.

To do that, we need (you are allowed to create as many intermediary tables as you want, to keep queries short):

* + 1. To join the 2 tables:
       - You notice States are not capitalized the same way in both tables (one is in uppercase letters, the other not), so you will first need to convert all to uppercase, for instance.
       - Now you can join the tables on the state key.
    2. Your boss wants you to change the name of the "District of Columbia" state to its short version "DC". Please do that.
    3. To compute the ratio between the number of grand electors and the population. Please create a new column with that ratio.
    4. To order the states by decreasing ratio of Grand Electors per capita. That will make our priority list.
    5. To compute the running total of Grand Electors in that sorted list.
       - Hint: you can get inspiration from here to compute a running total from here: <https://stackoverflow.com/questions/21382766/cumulative-summing-values-in-sqlite>
       - [(Links to an external site.)](https://stackoverflow.com/questions/21382766/cumulative-summing-values-in-sqlite)
    6. Independently, to compute the half of the total of Grand Electors overall (in the whole country):
       - This is the threshold we need to reach for winning the presidential election.
    7. To filter our sorted list of states in order to keep only the (top) ones enabling us to reach the computed threshold. (the other states can be ignored). That is our target list.
       - Hint: You can do that in 2 steps:
         * Select all the states for which the running total is below or equal to the threshold.
         * Add the first state for which the running total is larger than the threshold.

Can you draw some conclusions from the result? Is it in line with your expectations? How many states do you end up with in the target list? Is it a small or a large number? Do you think it would be a good recommendation to target those states?

**Datasets**

Below are two datasets that can be considered for research:

* + 1. Grand electors by State. [[Link](https://drive.google.com/a/moringaschool.com/file/d/1AuPlNXKjmEdRTLGqKbP-OP1XZ5UQPIPw/view?usp=sharing)
    2. [(Links to an external site.)](https://drive.google.com/a/moringaschool.com/file/d/1AuPlNXKjmEdRTLGqKbP-OP1XZ5UQPIPw/view?usp=sharing)
    3. ]
    4. Population by State. [[Link](https://drive.google.com/a/moringaschool.com/file/d/1VKt_hF2pRqPxcNb1DKotkVXWNd2HX_KL/view?usp=sharing)
    5. [(Links to an external site.)](https://drive.google.com/a/moringaschool.com/file/d/1VKt_hF2pRqPxcNb1DKotkVXWNd2HX_KL/view?usp=sharing)
    6. ]

**Deliverables**

* + 1. You will be required to provide a short data report (Google Docs Document) on the process that will undertake while working on the given research problem.
       - You are allowed to make use of any relevant external resources including the learnings from this week’s sessions i.e. CRISP - DM Guide [[Link](https://drive.google.com/a/moringaschool.com/file/d/1rdNSTG9BW_M8ke155GlE6j1kygW3iApU/view?usp=sharing)
       - [(Links to an external site.)](https://drive.google.com/a/moringaschool.com/file/d/1rdNSTG9BW_M8ke155GlE6j1kygW3iApU/view?usp=sharing)
       - ].
       - Your responses towards the given research problem should also be documented in this data report.
       - Use the following naming convention for the above Google Docs Document.
         * "Moringa\_Data\_Science\_Prep\_W2\_Independent\_Project\_2019\_06\_FirstName\_LastName\_DataScience\_Lifecycle"
    2. You will also provide a Notebook that will contain your backend analysis and to the given research problem.
       - All questions, text responses, SQL queries should be included in the Notebook.
       - Use the following naming convention the above Notebook; "Moringa\_Data\_Science\_Prep\_W2\_Independent\_Project\_2019\_06\_FirstName\_LastName\_SQL\_Notebook"
    3. Put both files in one folder then submit the folder’s shared link on the LMS.