Big Data Analytics with Python

End of Course Project

# Introduction

The Big Data Analytics course is a practice focused course being delivered as part of the Master science program at AIMS Rwanda in the 2021-2022. The final assessment for this course involves students working on a mini project to demonstrate the skills they have grabbed. In addition to practicing all the technical skills on handling large datasets and writing good code, this project will also enable students to gain skills in data science project management and problem solving such as:

* The ability to convert high level project requirements from a non Data Scientist client into a solvable data science task(s).
* The ability to break down a complex and ambiguous data science task into smaller problems which can be easily tackled
* The ability to organize and plan a data science project execution in a structured fashion
* The ability to document the analysis process and package data science outputs in a way that can be communicated to clients either using powerpoint slides or an analysis report.

In the rest of this document, we outline the data analysis task as it would be provided by a hypothetical client. The document further provides details on the expected outputs from the project, the dataset which will be used and other pertinent information to help the student accomplish the project.

# Analysis Objectives

In this project, you will pretend that you have a client who is a transport planner in a fictitious city called “the Red city”. At a high level, the client says “I would like to understand mobility patterns within the city. For instance, how many trips do people take on average, how far do they travel everyday. This information is important for me for an upcoming public transport improvement project”. Let's assume that you had several meetings with the client and agreed on two main categories of analysis:

## Individual level mobility patterns

For an average individual in the city, the client would like to understand their travel patterns in terms of quantity (e.g., how many trips do people take) and timing (do they take most trips in morning or afternoon or evenings, what about weekdays vs. weekend). From this description and the description in the previous section, can you come up with at least 3 exact individual metrics to compute from the data[[1]](#footnote-0).

| **Metric Name** | **Exact metric definition** |
| --- | --- |
| Average number of trips per day | For each individual, the mean number of trips they take everyday |
|  |  |
|  |  |
|  |  |

The idea here is to turn generic descriptions above into exact and statistically sound metrics which you can compute from the data. Please see the example provided in the first row.

## Aggregate mobility patterns

The client also wants to understand, at the city level. Which areas receive most trips? What about the area which exports the most trips? The client has a transport background and so he clarifies that what really wants is for you to generate origin and destination matrices where each cell in the matrix indicates the number of trips. The client is not sure at this point how “area” is defined but this will be cleared later. For this category of analysis, your analysis goal is to generate origin and destination matrices often abbreviated to OD.

# Expected Outputs

As the data scientist for this project, please provide suggestions on the exact outputs you can generate for your client[[2]](#footnote-1). The following outputs will be generated at the end of the analysis:

1. Plots and charts
2. Output-2
3. Output-3

# Dataset

In this section, we pretend that the client has given you a dataset from a fictitious mobile company in the country called ‘Dunstan-Cell. There are two data files available as follows:

## Transaction records

This dataset contains historical metadata about each user's cellular activity. The data has the following columns:

* **cdr type**-the type of transaction
* **cdr datetime-**exact time of transaction
* **call duration**-duration of activity
* **last calling cellid-**id representing location of cellular tower
* **User\_id-**a fictitious user\_id

## Geographic location data

There is another file with geographic location details as below:

* Site\_id- an identifier for each unique geographic location
* Cell\_id-This is the cell id which matches with the **l**ast calling cellid in the transaction dataset
* Lat,lon-geographic longitude and latitude

# Methodological and Other Considerations

When tackling often ambiguous data science problems in the wild, it's important to quickly identify what will be challenging aspects of the work whether be it conceptually or due to computational resource limitations. Some of these may also not necessarily be viewed as challenges but rather things which require a lot of thought, consideration and design decisions. Now, list 3 specific such conceptual, methodological and design decisions you need to think about[[3]](#footnote-2):

1. How to define a trip? What should be the minimum distance traveled to constitute a trip? What about minimum time?
2. Consideration-1
3. Consideration-2
4. Consideration-3

# Resources

* [**Source code**](https://drive.google.com/file/d/1U_ubxTNQKLZnL2x0q_ZHeOQCmvTtGe75/view?usp=sharing)with utilities and functions useful for getting started with this analysis.
* Additional packages to install[[4]](#footnote-3): By opening and examining the Python linked above, identify packages you need to install to be able to run all the code in the file.
* **Starter code**- As it's been done in the assignments, starter code asking you to fill in some parts will be provided soon. Furthermore, a detailed run through some of the important code will be provided.

1. Question-1[6 points] [↑](#footnote-ref-0)
2. Question2-these outputs can be categories rather than exact outputs. For instance plots and charts as shown [4 points] [↑](#footnote-ref-1)
3. Question3 [3 points] [↑](#footnote-ref-2)
4. Question4 [1 point per each package installation] [↑](#footnote-ref-3)