ISYS 412 Application Development for Data Analytics

Project: Part 1

20pts

Due Date: March 20, 23:59 pm

Overall Project Objective

The objective of this semester project is to perform a data analysis using the concepts covered in class. You will be working with dataset from the website www.kaggle.com.

Your project final report should display the complete analysis with appropriate description. More information about the final deliverable will be in Final Report assignment.

Find your group

On Canvas, under **Modules,** go to the **Project** section and open the page **Group Project Members.** Use this page to indicate your group or eventually find/join a group.
I recommend you set up your group as early as possible.

Work within group

On Canvas, go to **People** menu and select **Groups** on the top tab. You will find a set of **Groups Project**. Once a group is created, I will add the members to a group. You and your team members are **required** to use this group to collaborate on the project (Example: set up meetings, share drafts, …). I will check the groups to make sure everyone contributed on the project.

Important : Project deliverables must be submitted under the respective project assignments. I will not consider any files under your group folder as project submission.

How to get the dataset

1. Create an account on Kaggle.
You are not required to enter your real name or personal information.

2. Kaggle possesses several public datasets available by clicking on **Datasets** in the **top menu. (DO NOT SELECT datasets from Competitions Menu)**

3. From the list of public datasets, select a dataset that you will be using for your project.

How to pick a good dataset:

For each dataset, you have its name, the author name, as well as 4 icons corresponding to when the data set was published, the file size, the dataset usability score and the file format.

- Make sure you can download the dataset file(s) and the format is .csv or .tsv
- The larger the file size, the more data you can analyze
- Pick a dataset with high usability score (at least 8)

Download the dataset, take a look at it, think about what data analysis can be interesting to do. Ideally, I do not want 2 groups working on the same dataset.

There are about 50 000+ available datasets. Find one related to a field you are interested in.

I am aware that many datasets contain example of data analysis. Remember that this project should be your own work. I will check if your work is not your own.

Part 1 Instruction

For the first part of this project you are required to submit a two-page (max) document with the following information:

1. (5 pts) A list of all the group member.
2. (5 pts) A short description of the dataset you will be working on including:
    a. The dataset **name**
    b. The **number of column and rows**
    c. A **description of the data**. This is not a data analysis. Just a short summary of what the data represents.
    d. **Why did you pick this dataset?** Why are you interested by those data ? (Personal interest ? Desire to working in this field ? or any other reason…)
3. (5 pts) The **driving question** of your analysis. It should be a general question that serves as starting point for your analysis and help the reader (me) understand why you performed the various specific analysis you did.

Examples:
- How did the population, life expectancy and GDP per capita evolved over the years across various countries? (for the gapminder dataset)
- Which factors influence the amount of tips given? (for the tips dataset)

4. (5 pts) Data Analysis is about using data to answer **specific analysis questions**.

    You may already have some ideas about which questions your data analysis will answer.

    Let me know about some of those questions (at least 5) so I can give you some feedback.

The objective of this first part is for you to decide which dataset you will use for this course project and share this information with me, so I can give you my approval.

I am likely to **not** validate you part 1 if I consider that your dataset has not enough element to perform a good data analysis.

I may also ask you to pick a different dataset if two groups have the same dataset…

You are welcome to email me if you have any questions.

Your submission should be in one of the following formats: **.doc, .docx or .pdf**
I only want one submission per group.