ISYS 412 Application Development for Data Analytics

Project: Final Report

80 pts

Due Date: May 24, 11:59 pm

Overall Project Objective

The objective of this final part is to perform a set of analysis on the dataset you chose. You should develop an overall question that will serve as driving question for your analysis. The first step will consist in cleaning your dataset and derive a dataFrame structure suitable for your analysis. You will then perform a set of data analysis about the variables (columns) or observation (rows) of your choosing. The idea is to derive a set of questions related to each other leading to statistical and comparison analysis. Finally, you will have to observe how two or more variables relate to each other and derive a model from your analysis.

In addition to the final report (one submission pre group) **, each group member is required to submit an additional file** (doc, docx or pdf) with a short paragraph describing how each member contributed to the project. It is expected for each member to equally contribute to the project. I may deduct points to any member who had no or very little contribution to the project.

Part 2 Instructions

I only want one submission per group.

For the final part of this project you are required to submit **two files:**

I. A Python file containing each part of your code and analysis. Use **Markdown** text to indicate and label the different parts of your analysis.

II. A complete report of your project **as a doc, docx or pdf** document with the following information, in the presented order.

Some of the following information may have been submitted in part 1. It is however possible that a few changes have been made since then. This final submission must therefore contain an updated version of those elements.

1. (5 pts) An **introductory paragraph** providing information about the dataset including
   a. The dataset name
   b. A short description of the dataset, including a description of the columns
   c. The number of rows and columns as well as the number of **missing values** in each column.
   d. Optional: Why did you pick this dataset?

2. (10 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)

3. (20 pts) A **description of your dataset cleaning process** and a **screenshot** of the first rows of your final dataset

It is very likely that the original dataset format is not suitable for your data analysis. Prior to analysis, you may want to clean your dataset by perform actions such as:

Removing rows and/or columns; Creating new columns; Splitting, merging and/or melting values; Modifying missing values and other type of required data cleaning.

For each data cleaning step, you must provide a **clear explanation** of why you performed this specific modification. Your python file should contain the code for each of the modification as I should be able to recreate your final dataset by running your code.
You should also include a **screenshot** of the first rows of your final tidy dataset.

4. (20 pts) A description of your **data analysis.**

In alignment with your original driving question, you must formulate a **minimum of 5 questions** that can be analyzed using various **subset**, **comparison** and **statistic methods** covered during the semester. You should make sure that there is a **diversity** in the methods used to answer these questions.
Within your data analysis, you **must** include a **minimum of 2 plots,** each providing additional visual information to answer some of your questions (part of you 5 questions minimum)

5. (20 pts) At **least one data modeling** analyzing the relation between several variables of your data set. Eventually, the model you derive is able to predict the behavior of some variables based on some other variables. This modeling should be used as final answer to your overall analysis question.

6. (5 pts) As conclusion, a paragraph indicating:
   a. The **benefits** you gained from this project (example: better understanding of data analysis, knowledge about dataset manipulation)
   b. The eventual **challenges** you may have encounter during this project and how you faced them.

In addition, each group member is required to submit an additional file (doc, docx or pdf) with a short paragraph describing how each member contributed to the project. It is expected for each member to equally contribute to the project. I may deduct points to any member who had no or very little contribution to the project.