AIM 5001 Project 1 (Module 7) (100 Points)

Using PostgreSQL + Pandas for Data Management & Analysis

****You may work in small groups of no more than three (3) people for this project. ****

This project will allow you to demonstrate your ability to: (1) make use of Python's Pandas library; (2) Apply your Python and PostgreSQL skills to the creation of a new database and associated tables within your PostgreSQL server; (3) perform basic exploratory data analysis on a novel data set; (4) create customized Matplotlib and Seaborn graphics as part of your exploratory data analysis work; and (5) present your work in the form of a more "formal" research paper framework.

A key component of this Project will be interacting with your PostgreSQL server via your Jupyter Notebook using whatever Python, Pandas, and SQL skills that you believe may be required for purposes of completing the project. All data storage and retrieval interaction with PostgreSQL **MUST** be facilitated via your Jupyter Notebook (as opposed to using the psql terminal app or some other PostgreSQL tool). However, the psql terminal app may be used for purposes of defining the required database and associated tables prior to your data storage and retrieval work

---

Start by selecting a data set to work with: you are free to work with any data set that has not already been used as part of the course work for this class (e.g., do not use the automotive, diamonds, or Chinook data sets). Your data also ***must NOT be sourced from*** Kaggle.com. Some potential sources of data are listed in the downloadable "**WhereToFindData" PDF** file, though you are ***strongly*** encouraged to seek out others on your own. The web abounds with freely available data sets!

Your selected data set should include at least two (2) numeric variables and one (1) categorical variable. Once you've selected a data set, **define a research question that is answerable with your data**. You will then use that research question to direct your analysis work throughout the remainder of the project.For purposes of making your work reproducible, you must also upload the dataset(s) you are using to your Github account and then load them into your Python environment from directly from your Github repository.

Your deliverable **must** include the following:

Part 1: Introduction(10 Points)– A brief summary of the type of data you've chosen to work with and the research question you hope to answer with it.

Part 2: Data Summary(10 Points)– Explain where you acquired your data from; how many use cases your data set provides; how many attributes are in each use case; what the data types are for each of the attributes; etc.Be sure include any Python code used as part of your Data Summary work.

Part 3: Data Management using PostgreSQL(30 Points) – The required components of Part 3 are as follows:

1. Create a fully normalized SQL database schema for the data you've chosen to work with.

2. Create an ER diagram for your proposed schema.

3. Formulate SQL statements that will create your proposed database and tables within PostgreSQL, including any required primary and foreign keys.

4. Open your PostgreSQLpsql terminal and execute the SQL statements you have written to create your proposed database within your PostgreSQL.**Do not populate the tables you have created with data from your data set as part of this step!!**

5. Connect to your PostgreSQL server from within your Jupyter Notebook environment anduse your knowledge of Python, Pandas, and SQL toload your data set into your new PostgreSQL database. Be sure to include a query that verifies the successful execution of your SQL statements.

6. Using your knowledge of Python, Pandas, and PostgreSQL, read your dataset from your new PostgreSQL database into a new Pandas dataframe within your Jupyter Notebook, using whatever SQL statements and Python logic you believe to be appropriate.

Part 4: **Exploratory Data Analysis (EDA)(20 Points)**– Using your newly created dataframe (see Part 3, #6), provide summary statistics for each attribute within the dataframe. Provide appropriate graphical analysis for each attribute using both the Matplotlib and Seaborn graphics libraries according to the methods specified in Module 8. Include a narrative describing your EDA findings, i.e., what have you learned from your summary statistics and graphics? Be sure include any Python code used as part of your EDA work.

Part 5: Inference (20 Points) – Perform whatever analysis is necessary to answer your research question. Your analysis should include at least one graphic. Include a narrative explaining your research approach and findings and be sure include any Python code used as part of your work.

Part 6: Conclusion (10 Points) – A brief, concise narrative explaining your conclusions.

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Save all of your work for this project within **a single Jupyter Notebook** and upload / submit it within the provided Project 1 Canvas submission portal. Be sure to save your Notebook using the following nomenclature : **first initial_last name_Project1**" (e.g., J_Smith_Project1). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***