DAV 5400 Project 1 (M6)

****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) perform basic exploratory data analysis on a novel data set; (3) create graphics using **Matplotlib** and **Seaborn** as part of your exploratory data analysis work; and (4) present your work in the form of a more "formal" research paper framework.

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 (i.e., do not use the automotive, hflights, diamonds, mushroom, or Pittsburgh Bridges data sets). For example, you might choose to work with the data set you or one of your classmates described as part of your Week 1 Discussion response, or you may choose to look elsewhere for a data source (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 work throughout the remainder of the project.

Your deliverable **must** include the following:

Part 1: Introduction (15 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: **Exploratory Data Analysis (EDA) (25 Points)** – Provide summary statistics for each attribute; provide appropriate graphical analysis for each attribute using both Matplotlib and Seaborn. For example, if you believe it is appropriate to generate a histogram for a particular variable as part of your EDA, create it first using Matplotlib and then once again using Seaborn. Include a narrative describing your EDA findings. Be sure include any Python code used as part of your EDA work.

Part 4: Inference (35 Points) – Perform whatever analysis is necessary to answer your research question. Your analysis should include at least one graphic, and for each graphic you create you must do so using both Matplotlib and Seaborn (as described in **Part 3** above). Include a narrative explaining your research approach and findings and be sure include any Python code used as part of your work.

Part 5: Conclusion (10 Points) – A brief, concise narrative explaining your conclusions.

References (5 Points) - Be sure to include proper citations for any references you may have relied on as part of your work.

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 submit it via the Project 1 page within Canvas. 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.***