COMP 4433: Project 1

This project will require you to conduct a visual exploratory analysis on a dataset of your choice using the matplotlib/seaborn ecosystem. The dataset should **not** be one of the built-in Seaborn datasets but should instead be some real-world data of your choosing. You may provide the code necessary to read the data from an online source or you may provide your data file(s) with your submission.

The analysis should be framed towards providing an external audience with an overview of the data in question. You may assume that the audience possesses technical expertise. Think of this project as an exploratory visual analysis for a scientific audience conducted through the lens of visual storytelling. The analysis does not have to be exhaustive but may instead focus on a selection of features from the dataset. The goal here isn't to draw any definitive conclusions. Instead, you should focus on familiarizing yourself and your audience with the data. Consider the following questions when producing your exploratory analysis. You don't have to address these items directly, but they may help guide your work.

- How are the continuous variables distributed?
- Are there strong or weak relationships between any variables?
- Do distributions or central tendency of continuous variables differ across levels of categorical variables?
- Is missing data a concern?
- Is it useful to examine the composition of the data based on any of the categorical features?
- If you were to think about building a predictive model using these data, which feature(s) would be the target(s)?

In order to facilitate an audience-centric analysis, the project should be implemented in a notebook. The notebook should be well-organized and include narrative explanation/clarification of the provided plots. In other words, please discuss what you intend to convey to your audience with each plot and discuss any notable observations. The purpose of your visualizations should be clear, but if you have constructed a particularly complex representation then additional clarification may be necessary.

You are free to use any standard library modules, but limit imported modules to numpy, pandas, matplotlib, seaborn, statsmodels and sklearn. If there are additional modules that you would like to use, please seek approval from your instructor.

Reasonable attention should be given to fine tuning the aesthetics of the plots. For example, effort should be made to reduce unnecessary visual elements, and color should be utilized appropriately. Relevant plot elements such as tick labeling, titles, descriptive string formatting and legends should be included as appropriate.

Be sure to include:

- At least one figure containing multiple plots and using differentially sized axes objects.
- At least three different plots from the matplotlib/seaborn suite.
- At least one on-plot annotation.

- Considerations for analytical approaches that may be applied to these data to further leverage them in an applied context.
- A discussion of anomalies, trends and observations of interest.

Deliverable
Submit a zipped folder containing your notebook and any files necessary to execute the notebook (data, image files, .ttf files, etc.).