DSA 610, Project #3, Spring 2025

Instructions: Use the **cities.xlsx** dataset for this project. You will be creating clustering models for this dataset.

First, perform some general exploratory data analysis. This can include things like pairplots, correlation tables, summary statistics, bar graphs/tables of categorical variables. Convert any categorical data to dummy variables as needed.

Second, you will want to rescale. Assess which rescaling strategy is the best for the models you will be using and given the data you have. Account for any potential outliers and your dummy variables. You may also want to engage in some feature engineering such as creating rate variables (for example, adjusting violent and property crimes to rates relative to the population for better comparison)

Finally, select at least three classification models that are sufficiently different from one another to apply to the dataset. Your modeling should include the following elements:

1. Adjust parameters to obtain the best results. Use elbow plots and silhouette scores as needed to optimize your clustering.

2. Describe your approach to adjusting parameters for each model and how you settled on the optimal settings.

3. Create graphs of the clustering results (include labels).

4. Assess your models for the best clustering results across model types and assess what the clusters could mean in practical terms (are the results regional? Cultural? Related to the size of the city? Another factor?). Creating graphs of the data by the resulting clusters could provide some clues.

5. Optional: since the names of the cities and states are provided, in principle, this data could be plotted on a map. (if you do this, code snippets would be appropriate)

The written portion should be a minimum of five pages, including graphs. Feel free to use Plotly (or another package) to look at three-dimensional scatterplots (or other types of graphs) as part of your analysis.

Submit all supporting exploratory analysis, including Jupyter notebook (or python code file). Cite any additional sources used.

The expectation is that you can complete this analysis project using primarily what we have done in class so far. You may, however, go beyond what we have done in class according to your interests and experience.