# Case Study #5: Crime Rate

Consider the data on crimes in the U.S. states (*Crime.csv*) which consists of 50 rows describing each state's crime rate per 100,000 population. These includes (a) violent crimes like murder, rape, robbery, and assaults; and (b) non-violent crimes, i.e., burglary, larceny (theft), and auto. The goal is to form groups of clusters based on these crime rates data and characterize those clusters in a way that will be useful for their analysis.

## Questions

1. Upload, explore, clean, and preprocess data for clustering.
    a. Create a *rates_df* data frame by uploading the original data set into Python. Determine and present in this report the data frame dimensions, i.e., number of rows and columns. Display and present in your report the first 10 records of the *rates_df* data frame.
    b. Use Pandas to normalize the crime data (*rates_df_norm*), display the first 10 records of the normalized data and present the table in your report. Briefly explain how the normalized data was calculated and what it means. Why is the normalized data used in clustering instead of the original data? Briefly explain.

2. Apply hierarchical clustering to classify the states into clusters based on the normalized crime data.
    a. Develop the hierarchical clustering (*hi_complete*) based on the complete (maximum) linkage method (*method='complete'*). Create and display the hierarchical dendrogram with the cluster threshold of *5.0* (*color_threshold=5.0*). Provide the dendrogram in your report and explain how many clusters are shown on the dendrogram. Develop and present in your report the cluster membership based on the number of clusters you received in the dendrogram.
    b. Identify a data frame with the normalized mean values for each cluster and input variable. Display these data frame and provide it in your report. In addition, present in your report the profile plots of the normalized means of the clusters for the input variables. Briefly explain how the clusters can be characterized by their respective means.
    c. Based on the clusters' profile plots and normalized mean values, provide cluster labeling using some common feature(s) or variable(s) means of clusters.

3. Apply *k*-means clustering to classify the states into clusters based on the crime data.
    a. Create *k*-means clustering with *k = 5*. Identify cluster membership for k-means clusters and provide them in your report. What are the two main differences between the algorithms used in hierarchical and k-means clustering?
    b. Develop the *Elbow* chart for k-means clustering (*k* varies from 1 to 12) of the normalized crime data, present the chart in your report, and explain if *k = 5* is an appropriate number of clusters in k-means clustering of the crime data.
    c. Identify a data frame with the normalized cluster centroids for each cluster and input variable. Display these data frame and provide it in your report. In addition, present in

your report the profile plots of the normalized clusters' centroids. Briefly explain how the clusters can be characterized by their respective centroids.

d. Based on the k-means clusters' profile plots and normalized centroids, provide cluster labeling using some common feature(s) or variable(s) means of clusters.

4. Compare the clusters from parts 2 and 3. From your standpoint, which clustering, hierarchical or k-means, provides more useful insights of the states' crime rates.