**Exercise 2-1: Unsupervised Machine Learning Approaches**

Instructions for Exercise:

1. Researchers are interested in understanding how information gained during routine clinical exams can reflect overall wellness in women. Using the dataset Clinic\_GA4.csv, use PCA to describe the presence of underlying constructs that explain the shared variance of the tests. Feature names are informative. All features are quantitative, continuous variables.

* Make sure you scale the input variables
* Generate a scree plot
* How many components would you retain in subsequent analyses? What proportion of the total variance do those components explain?
* Create a table showing the factor loadings on each component. Interpret these results in 2-3 sentences.

1. The built-in R dataset USArrests includes the crime statistics for each of the 50 US states in 1973. Incidence of arrest, per 100,000 residents for assault, murder and rape are included along with the proportion of the population that lives in urban communities. Using this dataset, identify clusters of states based on their crime stats using k-means cluster analysis.
   * Determine if you need to scale variables.
   * Evaluate different values of k to determine the optimal number of clusters using a clear, data-driven strategy.
   * Describe the composition of each cluster in terms of the original input features