## Clustering Assignment

**Libraries**: For this assignment you may need the following libraries: tidyverse, cluster, factoextra, and dendextend.

Before beginning the assignment tasks, you should read-in the "trucks.csv" dataset into a data frame called "trucks". In this dataset, Driver_ID is a unique identifier for each delivery driver, Distance is the average mileage driven by each driver in a day, and Speeding is the percentage of the driver's time in which he is driving at least 5 miles per hour over the speed limit. Note that the Driver_ID variable should NOT be used in your clustering analysis.

**Question 1**: Plot the relationship between Distance and Speeding.

Which characteristics (select all that apply) of the relationship between Distance and Speeding seems most apparent?
A. There appears to be more speeding among the drivers with smaller Distances
B. The data points are arranged in what appear to be four clusters
C. Longer distance drivers appear more likely to speed
D. There are no well-defined clusters of data points

**Question 2**: Create a new data frame called "trucks_cleaned" that contains the scaled and centered variables. Two notes: 1) The "predictor" variables in the recipe are "Distance" and "Speeding" and 2) There is no need to create dummy variables as there are no categorical variables in the data. Be sure that you do NOT include the Driver_ID variable.

What is the maximum value (to four decimal places) of the Distance variable in the scaled dataset?

**Question 3** Use k-Means clustering with two clusters (k=2) to cluster the "trucks_cleaned" data frame. Use a random number seed of 64. Use augment to add the resulting clusters object to the the "trucks" data frame. Design an appropriate visualization to visualize the clusters.

Which statement best describes the resulting clusters?
A. Drivers with shorter distances are in one cluster and those with longer distances are in another
B. Drivers with a higher proportion of speeding are in one cluster and those with a lower proportion of speeding are in another
C. Neither of these statements apply to the resulting clusters

**Question 4**: Create a visualization to show how the clusters appear from values of k from 1 to 8. Use a random number seed of 412. Which value of k appears to be most appropriate for this data?

**Question 5**: Create a plot of k versus within cluster sum of squares. Hint: We did this in the first clustering lecture. What number of clusters appears to be ideal based on this plot?

**Question 6**: Repeat Question 3 for the number of clusters that you correctly identifed in Question 5. Use the same random number seed as in Task 3. Create an appropriate visualization.

Which statements (select all that apply) appear to be most apparent about the clusters created in this question?
A. One cluster is composed of short distance drivers with a low proportion of speeding.
B. One cluster is composed of long distance drivers with a high proportion of speeding.
C. One cluster is composed of long distance drivers with a low proportion of speeding.
D. One cluster is composed of short distance drivers with a high proportion of speeding.