Clustering

Deliverable: All of your work for this assignment should be done in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.

Libraries: You will 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 identifer 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.

Task 1: Plot the relationship between Distance and Speeding. Describe this relationship. Does there appear to be any natural clustering of drivers?

Task 2: Create a new data frame (called trucks2) that excludes the Driver_ID variable and includes scaled versions of the Distance and Speeding variables. **NOTE: Wrap the scale(trucks2) command in an as.data.frame command to ensure that the resulting object is a data frame. By default, scale converts data frames to lists**

Task 3 Use k-Means clustering with two clusters (k=2) to cluster the trucks2 data frame. Use a random number seed of 64. Visualize the clusters using the fviz_cluster function. Comment on the clusters.

Task 4: Use the two methods from the k-Means lecture to identify the optimal number of clusters. Use a random number seed of 64 for these methods. Is there consensus between these two methods as the optimal number of clusters?

Task 5: Use the optimal number of clusters that you identified in Task 4 to create k-Means clusters. Use a random number seed of 64. Use the fviz_cluster function to visualize the clusters.

Task 6: In words, how would you characterize the clusters you created in Task 5?

Before starting Task 7, read in the "kenpom20.csv" file into a data frame called "bball". This is data from kenpom.com, a basketball statistics site run by Ken Pomeroy. The columns in the dataset are as follows:
* TeamName: Name of the university
* AdjTempo: Number of offensive possession per 40 minutes. Adjusted to incorporate free throws and other events
* AdjOE: Points scored (on offense) per 100 possessions, adjusted by opposition
* AdjDE: Points allowed (on defense) per 100 possession, adjusted by opposition
* eFGPct: Effective field goal percentage
* TOPct: Turnover percentage
* ORPct: Offensive rebound percentage
* FTRate: Ratio of free throws attempted to field goals attempted
* eFGPctD, TOPctD, ORPctD, and FTRateD: Same as above, but for the defense

Task 7: Create a new data frame called "bball2" that excludes team name and scales the variables. Then use the two methods from Task 4 to determine the optimal number of k-Means clusters for this data. Use a random number seed of 123. Is there consensus between these two methods as the optimal number of clusters?

Task 8: Create k-Means clusters with a k of 4. Use a random number seed of 1234. Use the fviz_cluster function to visualize the clusters.

Task 9: Extract the cluster number from the k-means algorithm and attach as a new column to your "bball" data frame. Use the code as shown below, but replace XXX with the name of your k-means object. Plot "AdjOE" vs. "AdjDE" (use a scatterplot) and assign point color based on "clusternum". Hint: You should use color = factor(clusternum) to assign colors to the points. What patterns do you see?

```
bball2 = bball2 %>% mutate(clusternum = XXX$cluster)
```