Week 7 Assignment

Christopher Carey Youngbeck

**Please use the sns dataset and apply the k means algorithm as illustrated in your classroom demo. Please try a different k value. Also write a one-two page interpretation of your output results.**

Under the “sns” dataset we have a broad range of individuals' interests gathered using a web crawler to identify a market segment using the K-means algorithm. A full text of social network profiles was put under observation. 36 random words were chosen to identify mentions within each social network profile to predict interests for each Indvidual and identify common interests between individuals across data clusters. The objective is to cluster teenage individuals with similar interests using the K-means algorithm. This data set comprises 30,000 objects with 40 different variables when structuring the data in the R code. Useful variables are taken into consideration such as age, gender, graduation year and friends. When tabulating the data, we have 22054 females and 5222 males. 2724 are of an unknown gender. Many of the ages fell outside of the teen-age range 13 to 20. We try not to take those into consideration, and they should be treated as missing variables. Unfortunately, I ran into an error when attempting to separate non-teen ages so these results may include some ages beyond the teen years. Thankfully I found the mean age for the population in this data set was 17 when executing the code: mean(teenInterests$age, na.rm = TRUE), so we should still be able to draw conclusions for the teen population regardless of the error problem. After classifying the clusters based on interests and omitting other variables I assigned the k-value as 3 or assigned the interest groups into 3 clusters.

From analyzing the data, clusters 1 and 3 had very little interest in sports. Across all sports the results came out pretty evenly for clusters 1 and 3 with the exception of cluster 3 having some interest in tennis. This may be because we have a population to include some older ages that could confound some of the data but with the average age being 17 it’s hard not to draw the conclusion that our young people in these clusters had other interests. On the other hand, cluster 2 had at least some interest in sports with basketball and football being of the highest interest as well as cheerleading to join a high level of interest as well. Cluster 2 was also more active in the areas of socializing and romantic interactions. Columns “cute”, “sex”, “sexy”, “hot” and “kissed” had very high levels of interest with “kissed” being of the highest interest level out of all the variables including the sports variables. The conclusions are honestly depressing when it comes to this population. Most people here don’t seem to have an extremely high level of interest in sports. Even with cluster 2 having some interest in sports, it was nowhere near the level of interest in social and romantic activities. However, this is all my perception. Someone else may view a concentration of interest in romance as way more exciting than I would.