Unsupervised Learning II: Clustering

Problem statement: A granting agency wants to identify colleges that have high numbers of low-income, and first generation college attendees to give those colleges additional funding.

1. In the clustering tutorial, we used k-means clustering to identify 3 clusters of colleges using these criteria.

   A. Replicate this analysis using the code in the tutorial to generate those 3 clusters and append the cluster levels to the `college_features` dataset. If you are getting an error using mutate, use the following code instead:

   ```
   college_features$cluster = kmeans_cluster$cluster
   ```

   B. What is the median family income for each cluster (**hint:** see `kmeans_cluster$centers` from the tutorial)?

   C. Subset the colleges_features dataset on the cluster with the lowest `family_income_median`, call this new data grant_candidates. **Note:** in the tutorial, grant_candidates were from Cluster 1, you could find that a different cluster from your analysis has the lowest `family_income_median` when you look at `kmeans_cluster$centers`.

   D. How many universities are in the cluster of grant receivers?

2. Upon review you're informed that there are too many universities receiving grants. The granting agency really likes the cluster approach but suggests you make 5 clusters instead of 3.

   A. Redo the k-means analysis above but create 5 clusters instead of 3. **Note:** If you appended cluster onto your `college_features` dataset, make sure to remove it before redoing the k-means analysis.

   B. Again subset the data on the cluster with the lowest `family_income_median`. How many universities will receive a grant now? What is the median and range of `family_income_median` of these universities and how does it compare to your answers in Question 1?

   C. You will likely find that there were two clusters out of the five with low but similar `family_income_median`. Among these two clusters, what else determined which cluster these universities were assigned to (**hint:** look at the centers again)? Based on those other variables, do you think we made the correct decision to distribute grants considering only `family_income_median`?

3. Hierarchical clustering: Part of the grant is to reformulate curriculums to better match top ten universities.

   A. Subset your colleges dataset using the following code. The !is.na(sat_verbal_quartile_1) removes universities that do not have SAT admission criteria, so we are looking at similar degree-granting universities. What other criteria are we using to subset?

   ```
   grant_colleges = colleges %>% filter(
   (!is.na(sat_verbal_quartile_1) & family_income_median < 40000 & median_earnings < 30000)
     )
   top_ten_schools = colleges %>% filter(top_ten == TRUE)

   heir_analysis_data = rbind(grant_colleges, top_ten_schools)
   ```

   B. Replicate the heirarchical clustering from the tutorial comparing major percentages using `heir_analysis_data` dataset. Which universities are the most different from the top ten schools in terms of majors?

   C. How else can we compare the grantee schools to the top ten schools? Explore using any of the methods we learned in this class.