Lab 4H - Finding clusters

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

## Clustering data

* We've seen previously that data scientists have methods to predict values of specific variables.
  + We used *regression* to predict numerical values and *classification* to predict categories.
* *Clustering* is similar to classification in that we want to group people into categories. But there's one important difference:
  + In *clustering*, we don't know how many groups to use because we're not predicting the value of a known variable!
* In this lab, we'll learn how to use the k-means clustering algorithm to group our data into clusters.

## The k-means algorithm

* The k-means algorithm works by splitting our data into *k* different clusters.
  + The number of clusters, the value of *k*, is chosen by the data scientist.
* The algorithm works *only* for numerical variables and *only* when we have no missing data.
* To start, use the data function to load the futbol data set.
  + This data contains 23 players from the US Men's National Soccer team (USMNT) and 22 quarterbacks from the National Football League (NFL).
* Create a scatterplot of the players ht\_inches and wt\_lbs and color each dot based on the league they play for.

## Running k-means

* After plotting the player's heights and weights, we can see that there are two clusters, or different types, of players:
  + Players in the NFL tend to be taller and weigh more than the shorter and lighter USMNT players.
* Fill in the blanks below to use k-means to cluster the same height and weight data into two groups:

kclusters(\_\_\_\_~\_\_\_\_, data = futbol, k = \_\_\_\_)

* Use this code and the mutate function to add the values from kclusters to the futbol data. Call the variable clusters.

## k-means vs. ground-truth

* In comparing our football and soccer players, we *know* for certain which league each player plays in.
  + We call this knowledge *ground-truth*.
* Knowing the *ground-truth* for this example is helpful to illustrate how k-means works, but in reality, data-scientists would run k-means not knowing the *ground-truth*.
* **Compare the clusters chosen by k-means to the ground-truth. How successful was k-means at recovering the league information?**

## On your own

* Load your class' timeuse data (remember to run timeuse\_format so each row represents the mean time each student in spent participating in the various activities).
* Create a scatterplot of homework and videogames variables.
  + Based on this graph, identify and remove any outliers by using the subset function.
* Use kclusters with k=2 for homework and videogames.
  + **Describe how the groups differ from eachother in terms of how long each group spends playing videogames and doing homework**