Assignment 8.3

Scott Breitbach

7/20/2020

## Assignment 8.3: Clustering

clusterDF <- read\_csv("completed/Week8/data/clustering-data.csv")  
head(clusterDF)

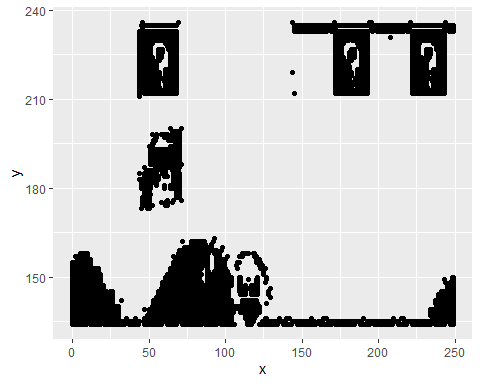
## # A tibble: 6 x 2  
## x y  
## <dbl> <dbl>  
## 1 46 236  
## 2 69 236  
## 3 144 236  
## 4 171 236  
## 5 194 236  
## 6 195 236

summary(clusterDF)

## x y   
## Min. : 0.0 Min. :134.0   
## 1st Qu.: 56.0 1st Qu.:141.0   
## Median : 82.0 Median :154.0   
## Mean :109.6 Mean :175.7   
## 3rd Qu.:180.0 3rd Qu.:218.0   
## Max. :249.0 Max. :236.0

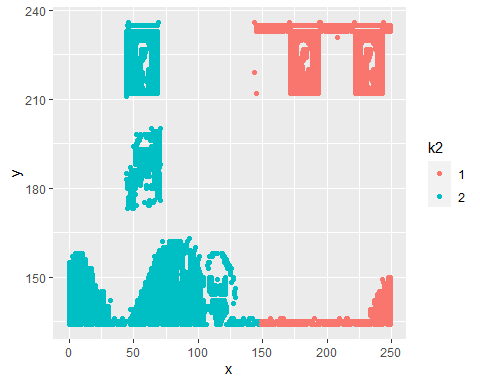
### a. Plot the dataset using a scatter plot.

ggplot(data = clusterDF, mapping = aes(x, y)) + geom\_point()

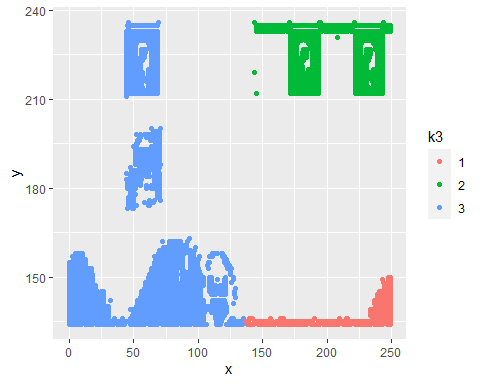


### b. Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

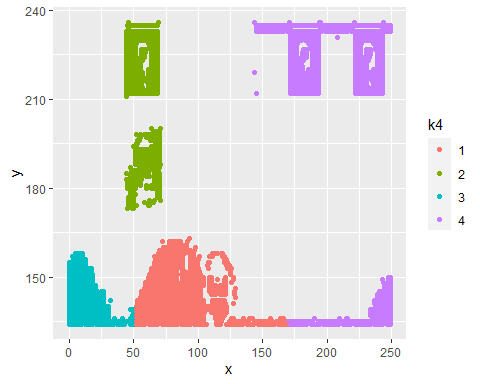
clusters.2 <- kmeans(clusterDF, 2)  
clusterDF$k2 <- as.factor(clusters.2$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k2)) + geom\_point()



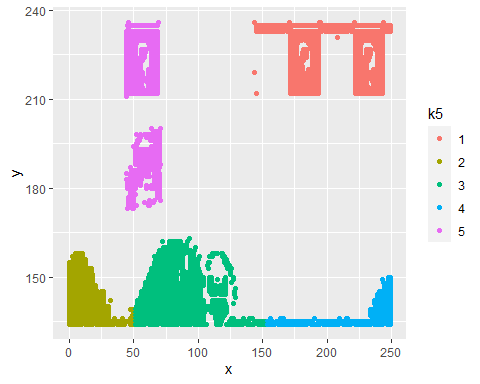
clusters.3 <- kmeans(clusterDF, 3)  
clusterDF$k3 <- as.factor(clusters.3$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k3)) + geom\_point()



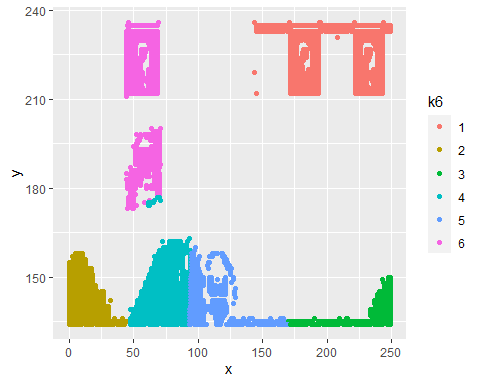
clusters.4 <- kmeans(clusterDF, 4)  
clusterDF$k4 <- as.factor(clusters.4$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k4)) + geom\_point()



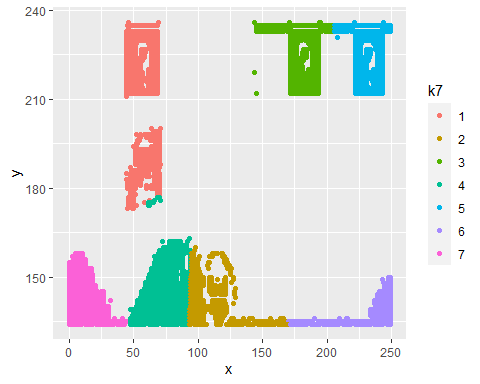
clusters.5 <- kmeans(clusterDF, 5)  
clusterDF$k5 <- as.factor(clusters.5$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k5)) + geom\_point()



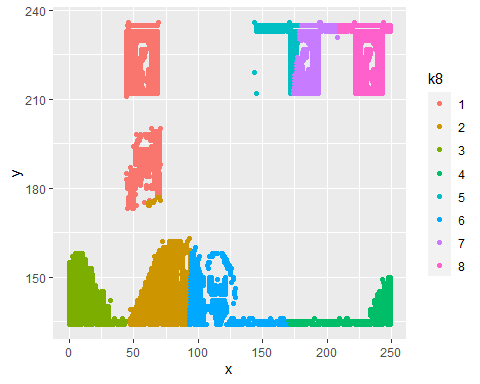
clusters.6 <- kmeans(clusterDF, 6)  
clusterDF$k6 <- as.factor(clusters.6$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k6)) + geom\_point()



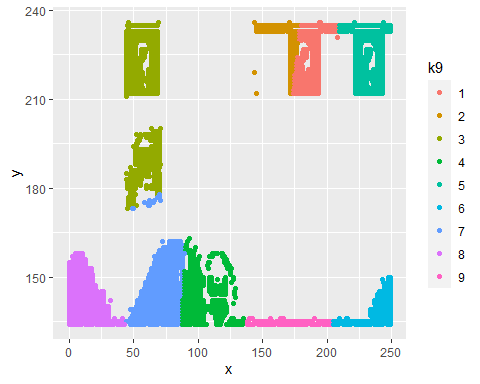
clusters.7 <- kmeans(clusterDF, 7)  
clusterDF$k7 <- as.factor(clusters.7$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k7)) + geom\_point()



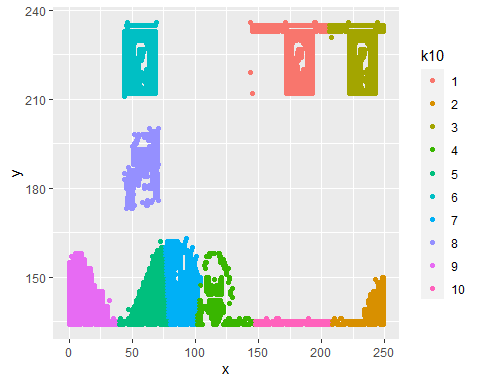
clusters.8 <- kmeans(clusterDF, 8)  
clusterDF$k8 <- as.factor(clusters.8$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k8)) + geom\_point()



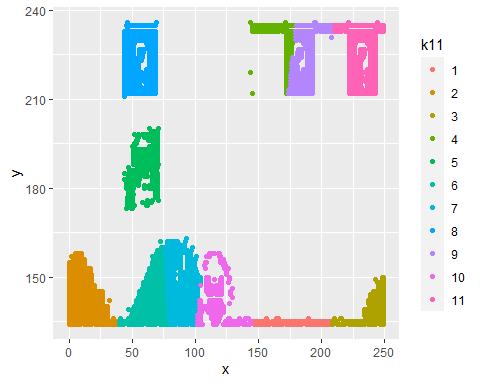
clusters.9 <- kmeans(clusterDF, 9)  
clusterDF$k9 <- as.factor(clusters.9$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k9)) + geom\_point()



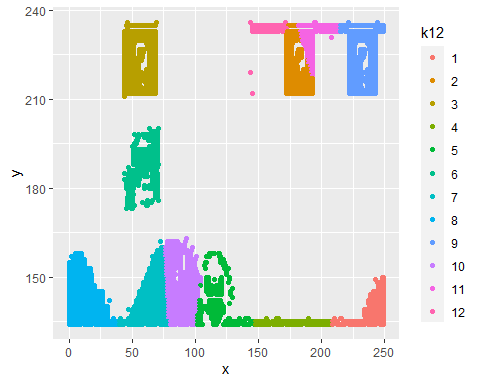
clusters.10 <- kmeans(clusterDF, 10)  
clusterDF$k10 <- as.factor(clusters.10$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k10)) + geom\_point()



clusters.11 <- kmeans(clusterDF, 11)  
clusterDF$k11 <- as.factor(clusters.11$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k11)) + geom\_point()



clusters.12 <- kmeans(clusterDF, 12)  
clusterDF$k12 <- as.factor(clusters.12$cluster)  
  
ggplot(clusterDF, mapping = aes(x, y, color = k12)) + geom\_point()



### c. To calculate the average distance from the center of each cluster, compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.

myDist<-function(p1,p2) sqrt((p1[,1]-p2[,1])^2+(p1[,2]-p2[,2])^2)

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.2$cluster==1,],clusters.2$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.2$cluster==2,],clusters.2$centers[2,,drop=FALSE])  
# Get average distance  
k2.avdist <- sum(dist.1, dist.2)/nrow(clusterDF)  
k2.avdist

## [1] 42.07717

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.3$cluster==1,],clusters.3$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.3$cluster==2,],clusters.3$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.3$cluster==3,],clusters.3$centers[3,,drop=FALSE])  
# Get average distance  
k3.avdist <- sum(dist.1, dist.2, dist.3)/nrow(clusterDF)  
k3.avdist

## [1] 36.53193

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.4$cluster==1,],clusters.4$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.4$cluster==2,],clusters.4$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.4$cluster==3,],clusters.4$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.4$cluster==4,],clusters.4$centers[4,,drop=FALSE])  
# Get average distance  
k4.avdist <- sum(dist.1, dist.2, dist.3, dist.4)/nrow(clusterDF)  
k4.avdist

## [1] 25.78229

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.5$cluster==1,],clusters.5$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.5$cluster==2,],clusters.5$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.5$cluster==3,],clusters.5$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.5$cluster==4,],clusters.5$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.5$cluster==5,],clusters.5$centers[5,,drop=FALSE])  
# Get average distance  
k5.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5)/nrow(clusterDF)  
k5.avdist

## [1] 20.33178

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.6$cluster==1,],clusters.6$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.6$cluster==2,],clusters.6$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.6$cluster==3,],clusters.6$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.6$cluster==4,],clusters.6$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.6$cluster==5,],clusters.6$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.6$cluster==6,],clusters.6$centers[6,,drop=FALSE])  
# Get average distance  
k6.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6)/nrow(clusterDF)  
k6.avdist

## [1] 18.10784

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.7$cluster==1,],clusters.7$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.7$cluster==2,],clusters.7$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.7$cluster==3,],clusters.7$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.7$cluster==4,],clusters.7$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.7$cluster==5,],clusters.7$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.7$cluster==6,],clusters.7$centers[6,,drop=FALSE])  
dist.7 <- myDist(clusterDF[clusters.7$cluster==7,],clusters.7$centers[7,,drop=FALSE])  
# Get average distance  
k7.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6, dist.7)/nrow(clusterDF)  
k7.avdist

## [1] 14.38976

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.8$cluster==1,],clusters.8$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.8$cluster==2,],clusters.8$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.8$cluster==3,],clusters.8$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.8$cluster==4,],clusters.8$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.8$cluster==5,],clusters.8$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.8$cluster==6,],clusters.8$centers[6,,drop=FALSE])  
dist.7 <- myDist(clusterDF[clusters.8$cluster==7,],clusters.8$centers[7,,drop=FALSE])  
dist.8 <- myDist(clusterDF[clusters.8$cluster==8,],clusters.8$centers[8,,drop=FALSE])  
# Get average distance  
k8.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6, dist.7, dist.8)/nrow(clusterDF)  
k8.avdist

## [1] 14.01589

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.9$cluster==1,],clusters.9$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.9$cluster==2,],clusters.9$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.9$cluster==3,],clusters.9$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.9$cluster==4,],clusters.9$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.9$cluster==5,],clusters.9$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.9$cluster==6,],clusters.9$centers[6,,drop=FALSE])  
dist.7 <- myDist(clusterDF[clusters.9$cluster==7,],clusters.9$centers[7,,drop=FALSE])  
dist.8 <- myDist(clusterDF[clusters.9$cluster==8,],clusters.9$centers[8,,drop=FALSE])  
dist.9 <- myDist(clusterDF[clusters.9$cluster==9,],clusters.9$centers[9,,drop=FALSE])  
# Get average distance  
k9.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6, dist.7, dist.8, dist.9)/nrow(clusterDF)  
k9.avdist

## [1] 12.82285

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.10$cluster==1,],clusters.10$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.10$cluster==2,],clusters.10$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.10$cluster==3,],clusters.10$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.10$cluster==4,],clusters.10$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.10$cluster==5,],clusters.10$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.10$cluster==6,],clusters.10$centers[6,,drop=FALSE])  
dist.7 <- myDist(clusterDF[clusters.10$cluster==7,],clusters.10$centers[7,,drop=FALSE])  
dist.8 <- myDist(clusterDF[clusters.10$cluster==8,],clusters.10$centers[8,,drop=FALSE])  
dist.9 <- myDist(clusterDF[clusters.10$cluster==9,],clusters.10$centers[9,,drop=FALSE])  
dist.10 <- myDist(clusterDF[clusters.10$cluster==10,],clusters.10$centers[10,,drop=FALSE])  
# Get average distance  
k10.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6, dist.7, dist.8, dist.9, dist.10)/nrow(clusterDF)  
k10.avdist

## [1] 10.47212

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.11$cluster==1,],clusters.11$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.11$cluster==2,],clusters.11$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.11$cluster==3,],clusters.11$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.11$cluster==4,],clusters.11$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.11$cluster==5,],clusters.11$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.11$cluster==6,],clusters.11$centers[6,,drop=FALSE])  
dist.7 <- myDist(clusterDF[clusters.11$cluster==7,],clusters.11$centers[7,,drop=FALSE])  
dist.8 <- myDist(clusterDF[clusters.11$cluster==8,],clusters.11$centers[8,,drop=FALSE])  
dist.9 <- myDist(clusterDF[clusters.11$cluster==9,],clusters.11$centers[9,,drop=FALSE])  
dist.10 <- myDist(clusterDF[clusters.11$cluster==10,],clusters.11$centers[10,,drop=FALSE])  
dist.11 <- myDist(clusterDF[clusters.11$cluster==11,],clusters.11$centers[11,,drop=FALSE])  
# Get average distance  
k11.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6, dist.7, dist.8, dist.9, dist.10, dist.11)/nrow(clusterDF)  
k11.avdist

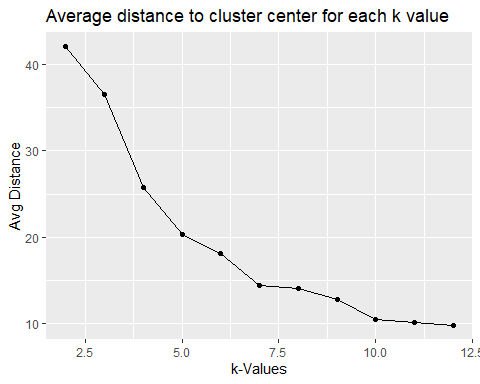
## [1] 10.09825

# Get distances from each cluster  
dist.1 <- myDist(clusterDF[clusters.12$cluster==1,],clusters.12$centers[1,,drop=FALSE])  
dist.2 <- myDist(clusterDF[clusters.12$cluster==2,],clusters.12$centers[2,,drop=FALSE])  
dist.3 <- myDist(clusterDF[clusters.12$cluster==3,],clusters.12$centers[3,,drop=FALSE])  
dist.4 <- myDist(clusterDF[clusters.12$cluster==4,],clusters.12$centers[4,,drop=FALSE])  
dist.5 <- myDist(clusterDF[clusters.12$cluster==5,],clusters.12$centers[5,,drop=FALSE])  
dist.6 <- myDist(clusterDF[clusters.12$cluster==6,],clusters.12$centers[6,,drop=FALSE])  
dist.7 <- myDist(clusterDF[clusters.12$cluster==7,],clusters.12$centers[7,,drop=FALSE])  
dist.8 <- myDist(clusterDF[clusters.12$cluster==8,],clusters.12$centers[8,,drop=FALSE])  
dist.9 <- myDist(clusterDF[clusters.12$cluster==9,],clusters.12$centers[9,,drop=FALSE])  
dist.10 <- myDist(clusterDF[clusters.12$cluster==10,],clusters.12$centers[10,,drop=FALSE])  
dist.11 <- myDist(clusterDF[clusters.12$cluster==11,],clusters.12$centers[11,,drop=FALSE])  
dist.12 <- myDist(clusterDF[clusters.12$cluster==12,],clusters.12$centers[12,,drop=FALSE])  
# Get average distance  
k12.avdist <- sum(dist.1, dist.2, dist.3, dist.4, dist.5, dist.6, dist.7, dist.8, dist.9, dist.10, dist.11, dist.12)/nrow(clusterDF)  
k12.avdist

## [1] 9.779885

#### Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.

k.values <- 2:12  
avDist <- c(k2.avdist, k3.avdist, k4.avdist, k5.avdist, k6.avdist, k7.avdist, k8.avdist, k9.avdist, k10.avdist, k11.avdist, k12.avdist)  
k.dist <- data.frame(k.values, avDist)  
  
ggplot(k.dist, mapping = aes(k.values, avDist)) + geom\_line() + geom\_point() + labs(x = "k-Values", y = "Avg Distance", title = "Average distance to cluster center for each k value")



The results seem to vary slightly depending on the random subset of data used to determine clusters, but with the seed set to 42, I would say that the “elbow” point would appear to be around k = 7 because that is where the chart seems to stop dropping and start to really level out. There are many ways to determine an elbow point so it seems somewhat arbitrary (again, especially considering the built-in variation due to randomness).

As a side note, I started trying to figure out a cleaner way to calculate average distance from cluster center for each value of k by using a function with a built-in loop. I wasn’t able to get it working initially and between life, work, and school I ran out of time to figure it out and implement it, but I’m sure it’s possible.