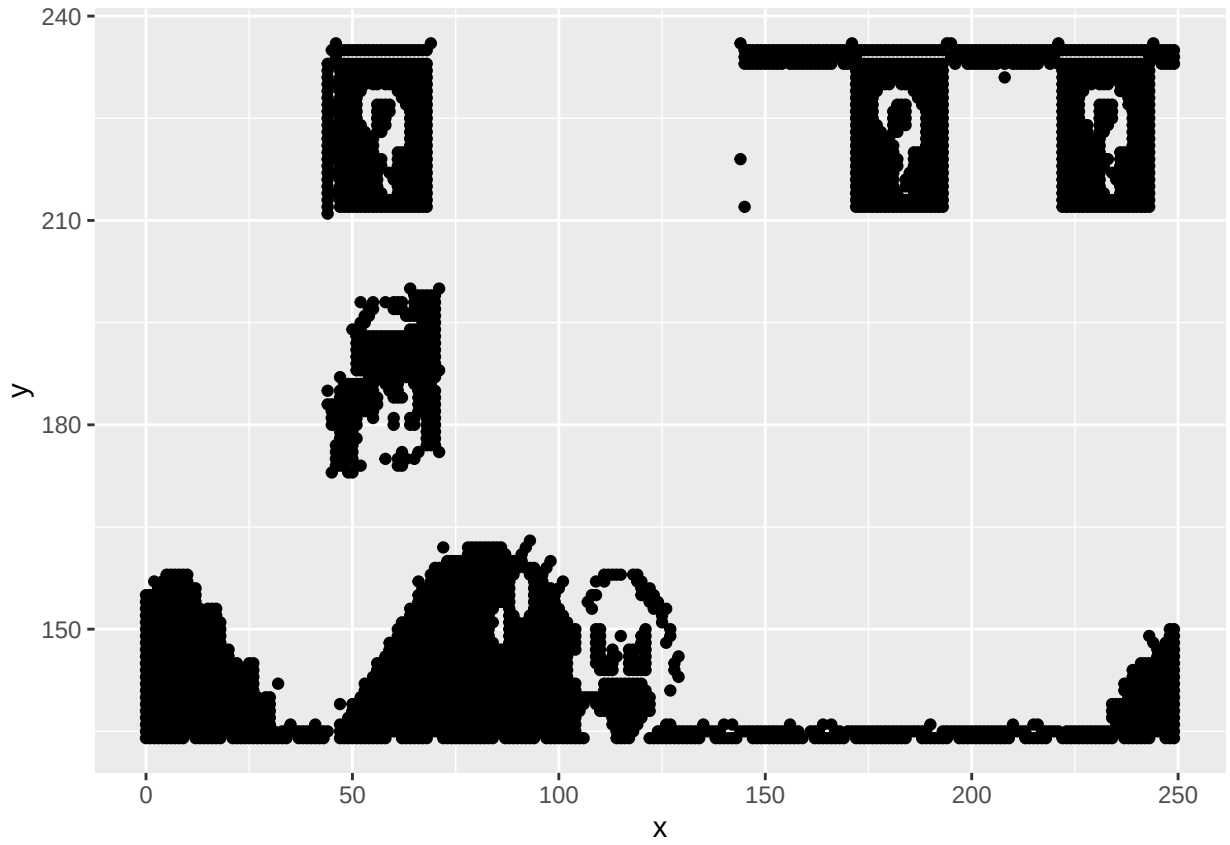
Exercise 10.2 Part 2

Alan Donahue

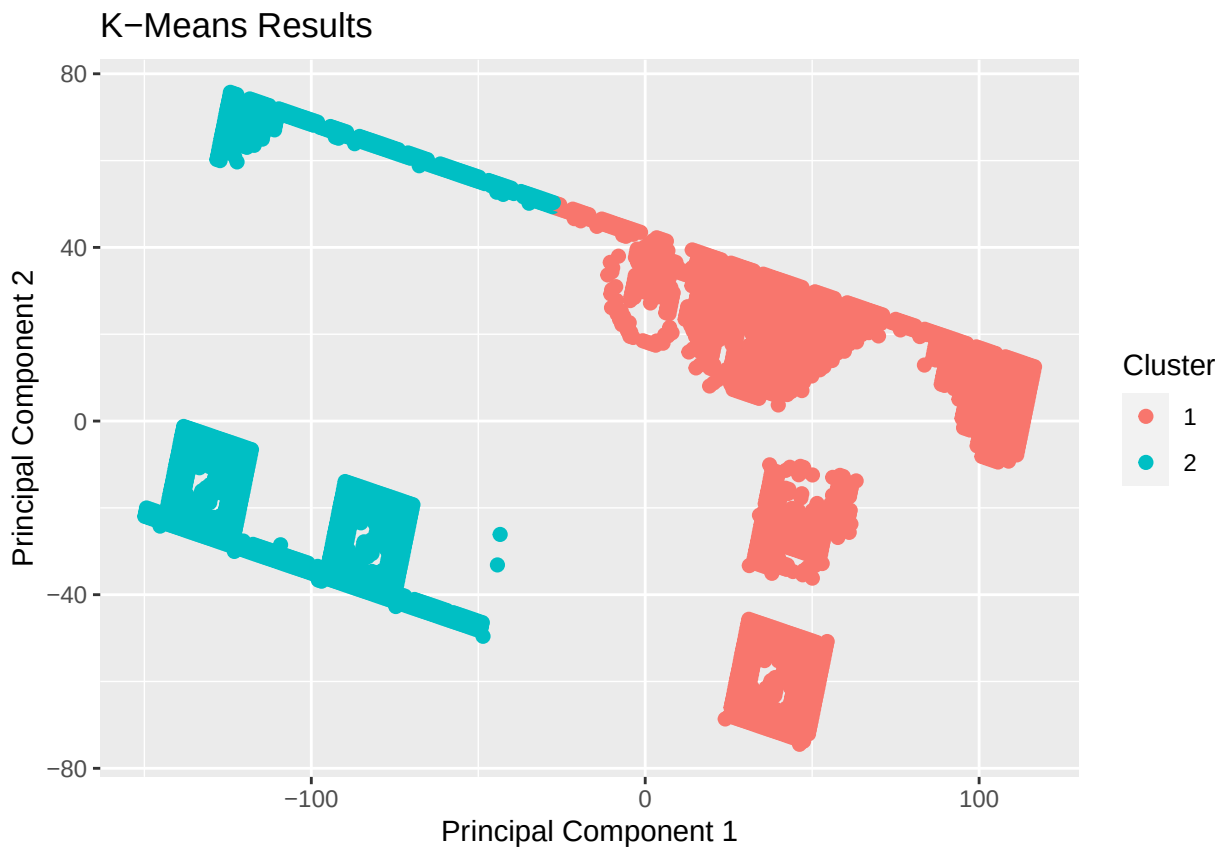
8/14/2021

Section 2

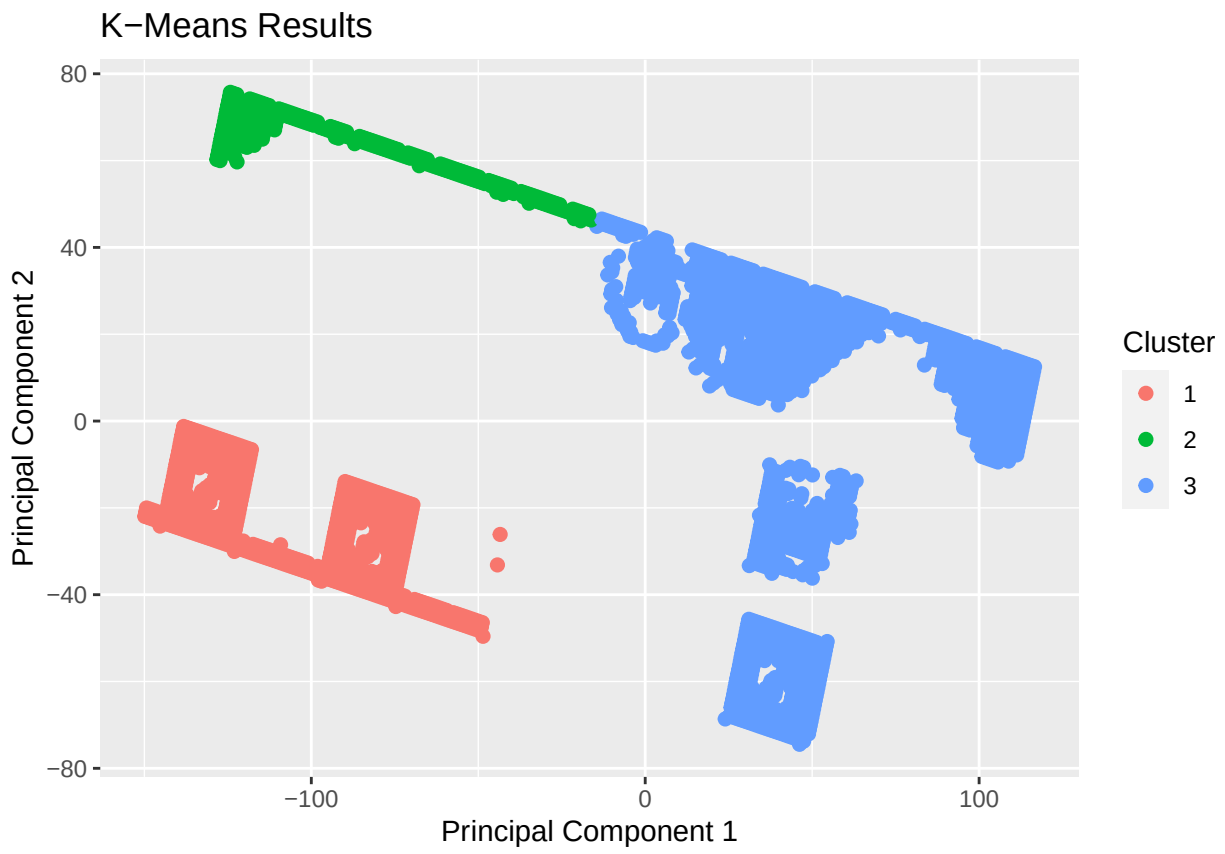
```
library(ggplot2)
library(foreign)
library(caTools)
library(class)
library(plyr)
library(useful)
library(cluster)

#setting the working directory
setwd("C:/Users/Alan Donahue/Documents/data science masters/DSC 520 Stats/GIT/dsc520")

#load the data
clustering_df <- read.csv("data/clustering-data.csv")

#scatter plot
ggplot(clustering_df, aes(x=x, y=y)) + geom_point()
```



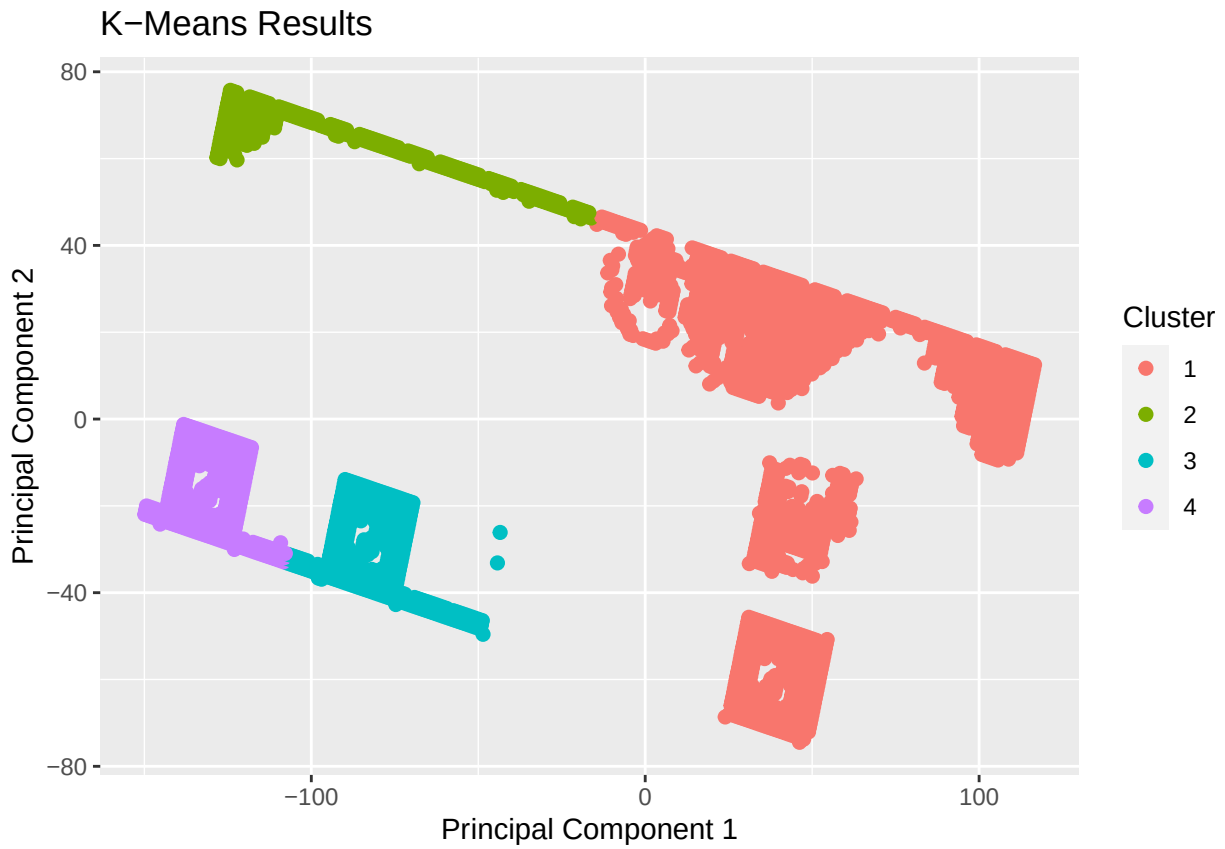

```
#fit data from k = 2 to k = 12
cluster_2 <- kmeans(clustering_df, centers = 2)
cluster_3 <- kmeans(clustering_df, centers = 3)
cluster_4 <- kmeans(clustering_df, centers = 4)
cluster_5 <- kmeans(clustering_df, centers = 5)
cluster_6 <- kmeans(clustering_df, centers = 6)
cluster_7 <- kmeans(clustering_df, centers = 7)
cluster_8 <- kmeans(clustering_df, centers = 8)
cluster_9 <- kmeans(clustering_df, centers = 9)
cluster_10 <- kmeans(clustering_df, centers = 10)
cluster_11 <- kmeans(clustering_df, centers = 11)
cluster_12 <- kmeans(clustering_df, centers = 12)

#plotting the scatter plots
plot(cluster_2, data = clustering_df)
```



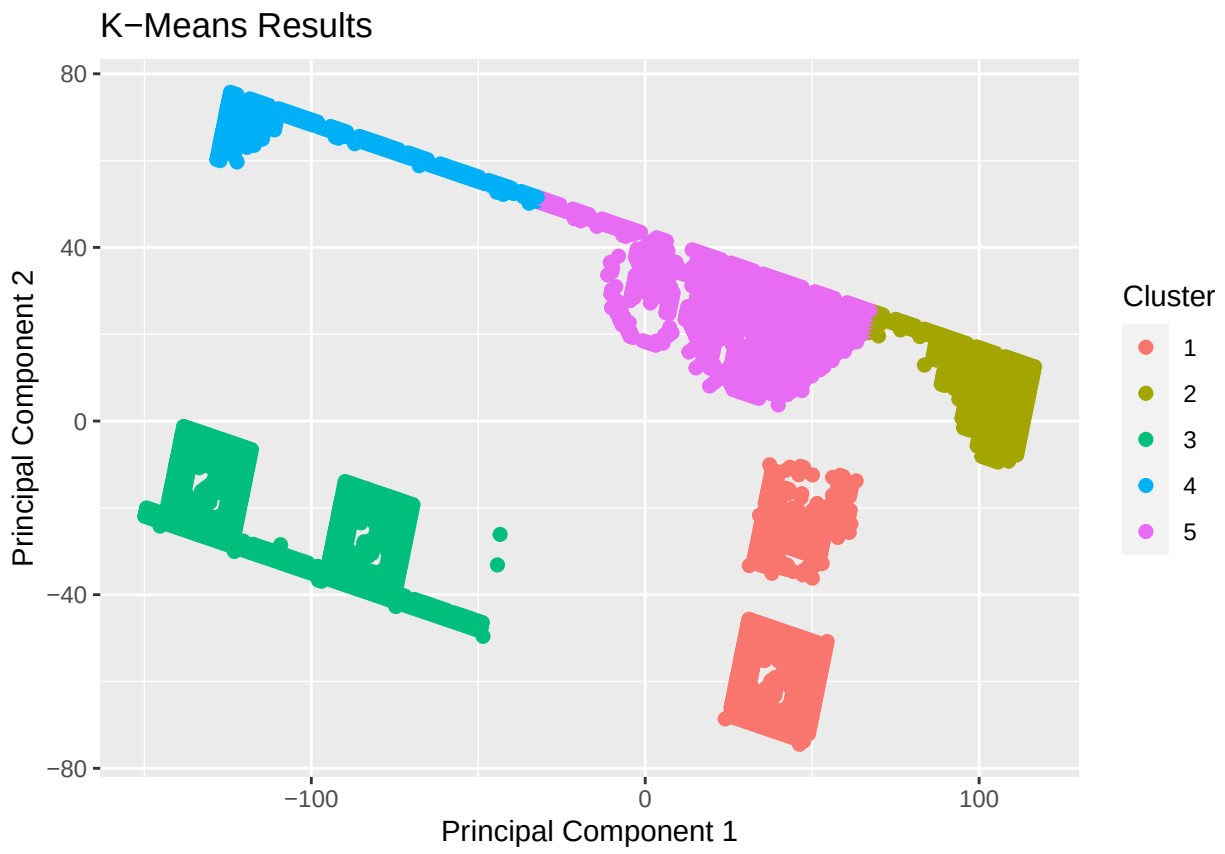

```
plot(cluster_3, data = clustering_df)
```



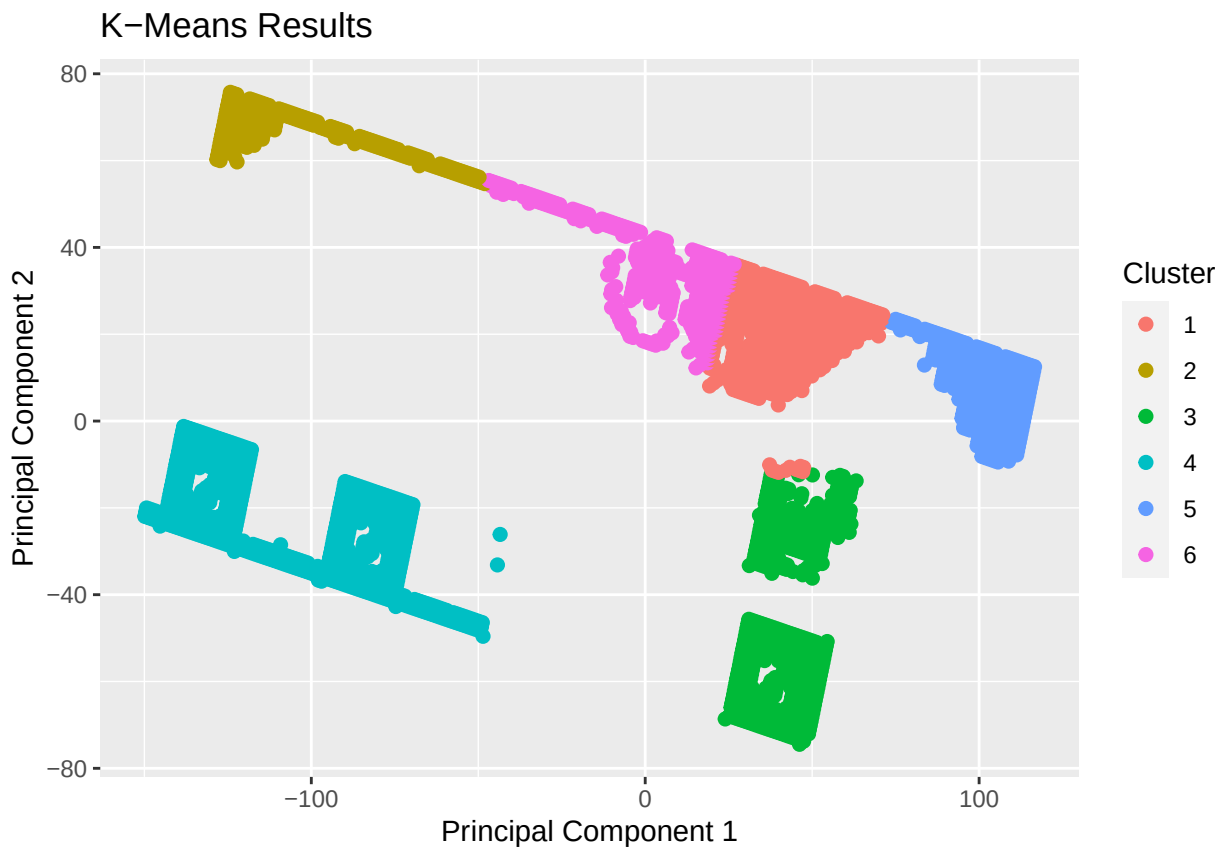

```
plot(cluster_4, data = clustering_df)
```



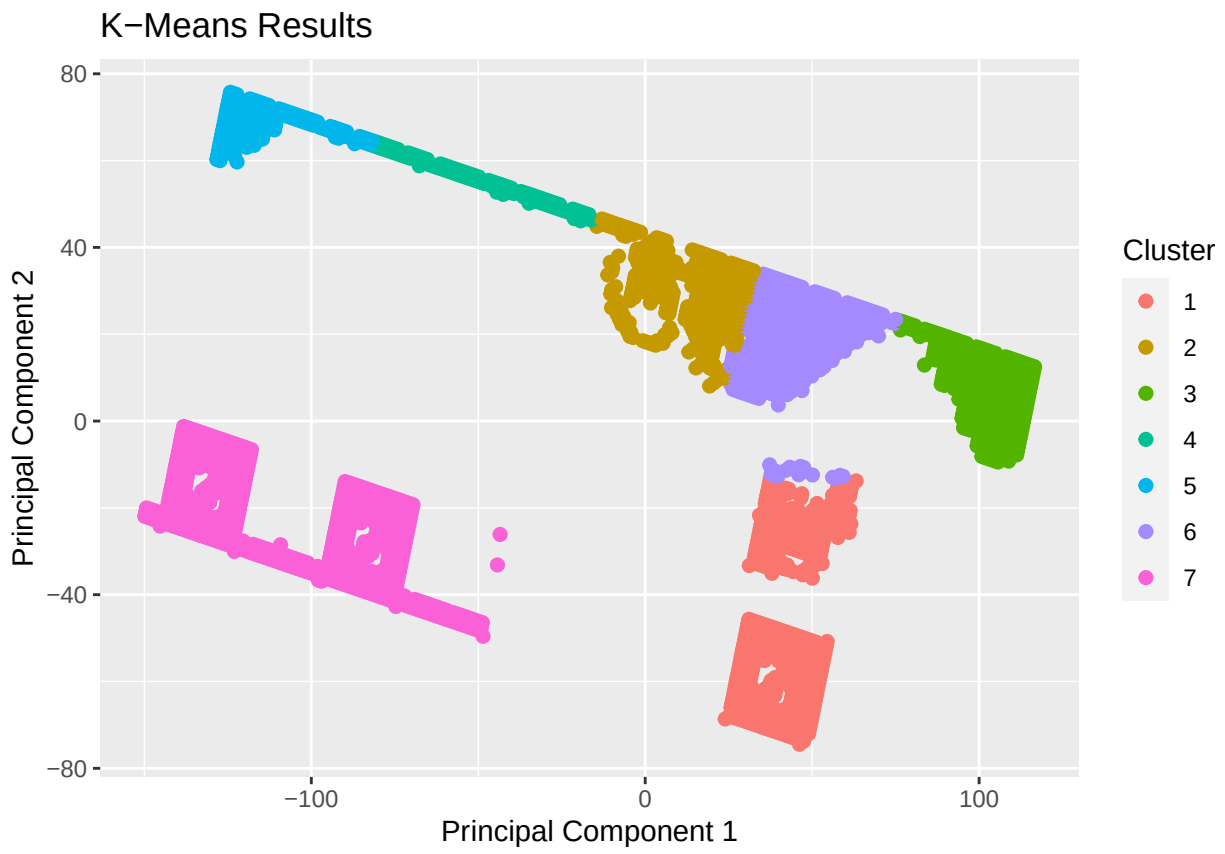

```
plot(cluster_5, data = clustering_df)
```



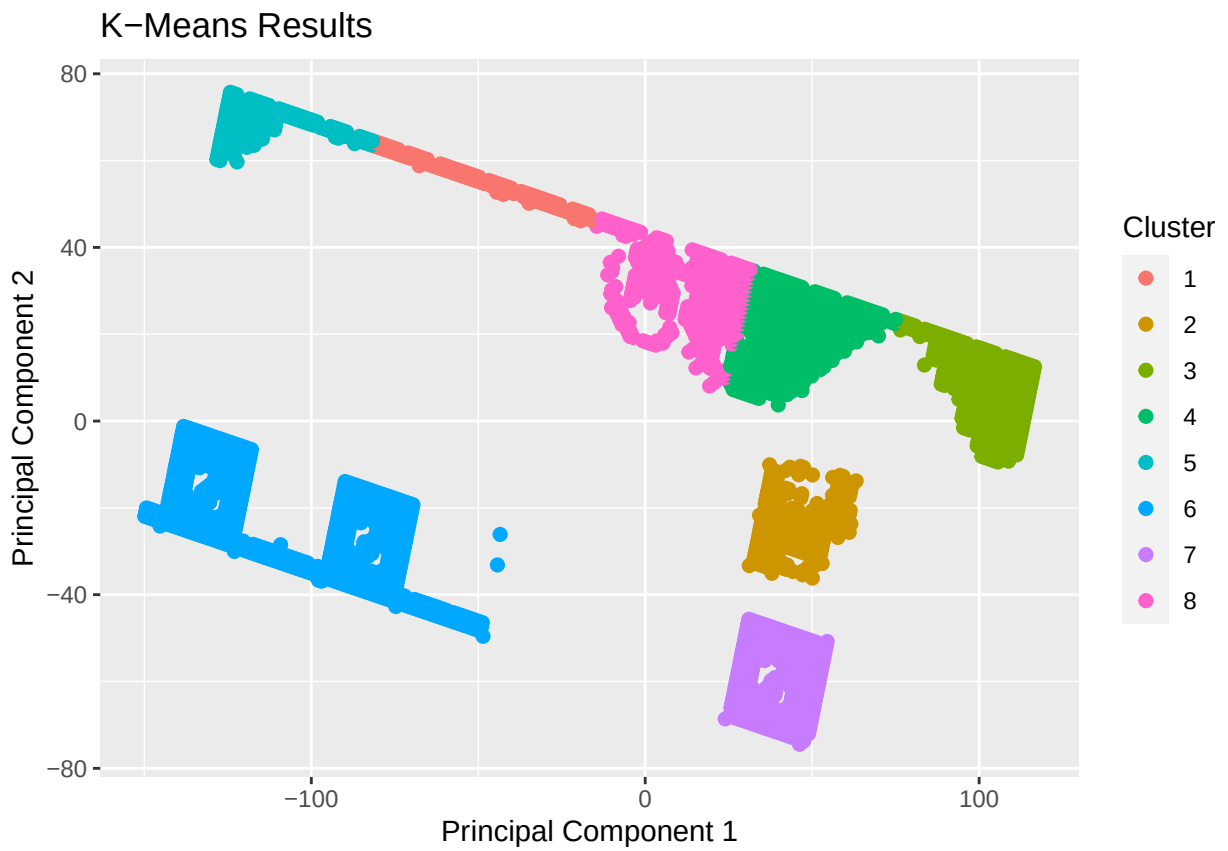

```
plot(cluster_6, data = clustering_df)
```



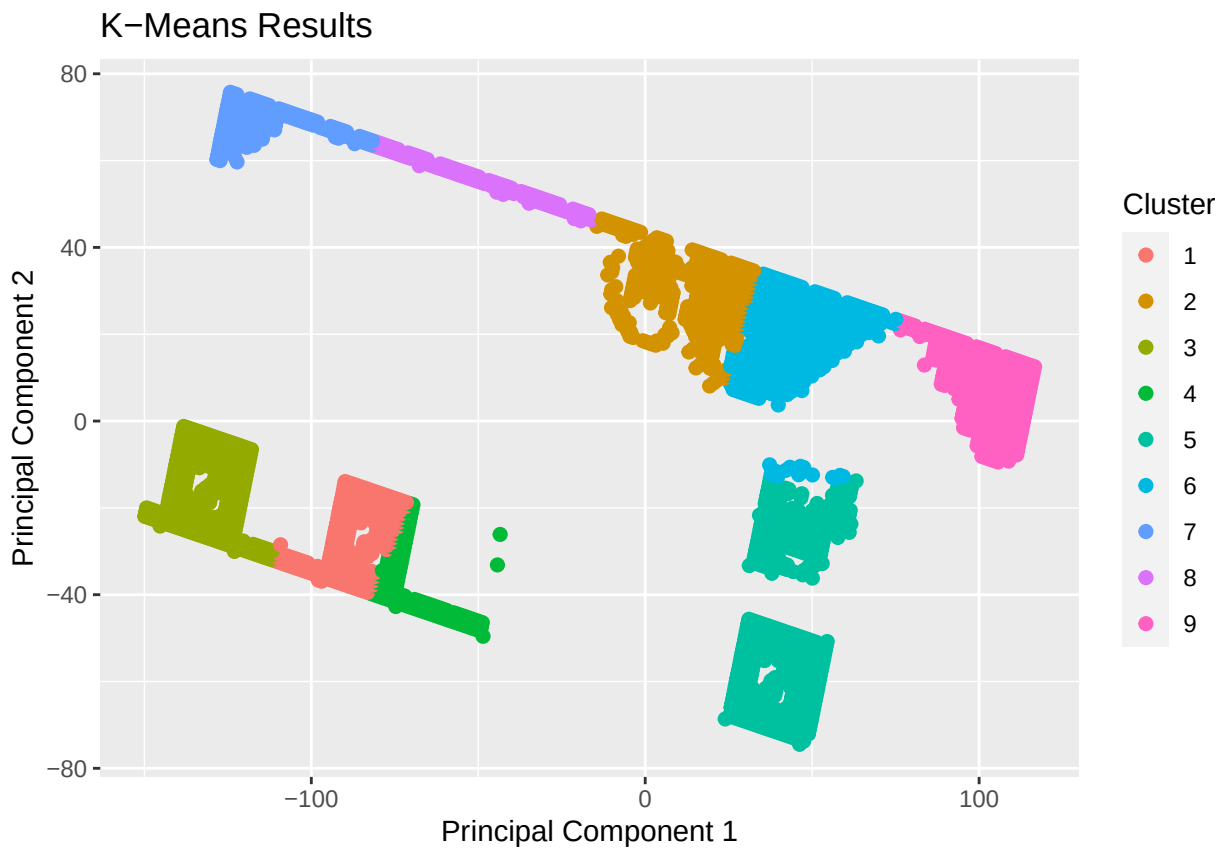

```
plot(cluster_7, data = clustering_df)
```



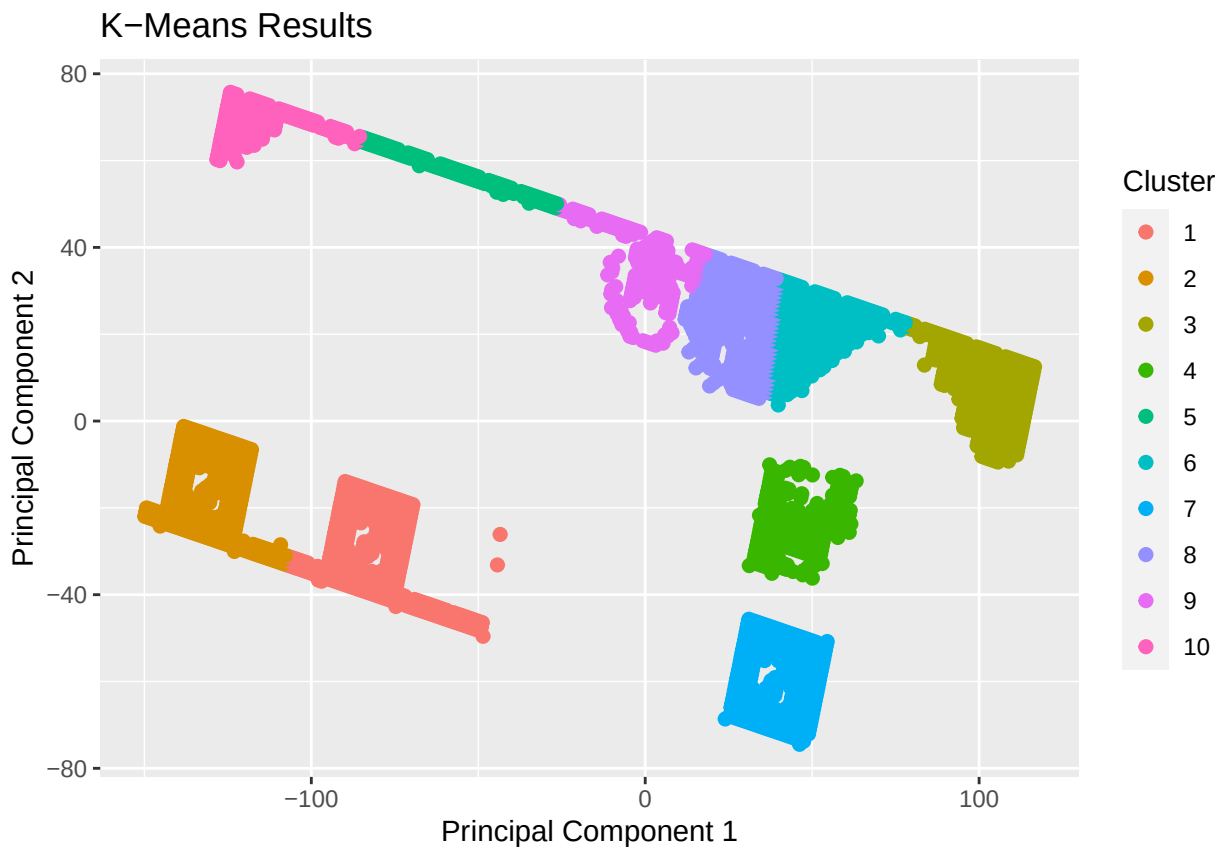

```
plot(cluster_8, data = clustering_df)
```



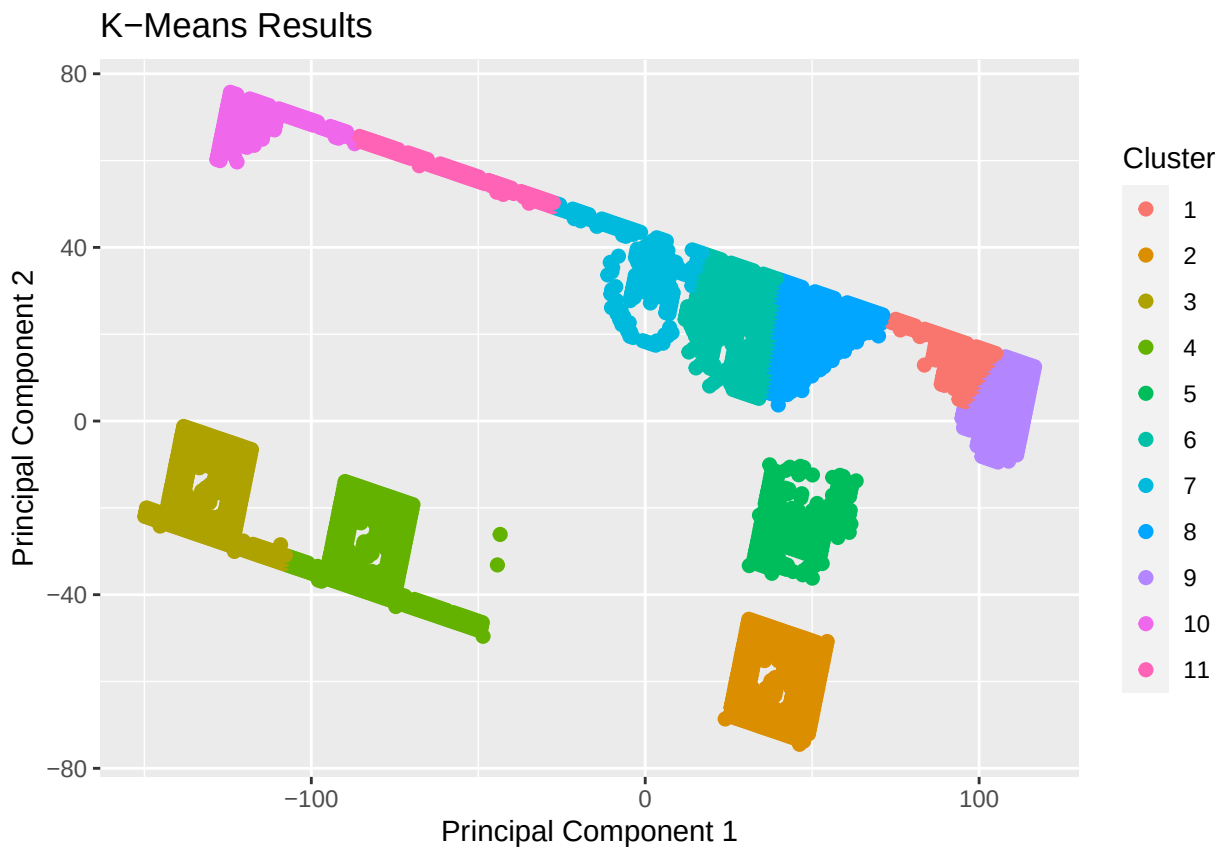

```
plot(cluster_9, data = clustering_df)
```



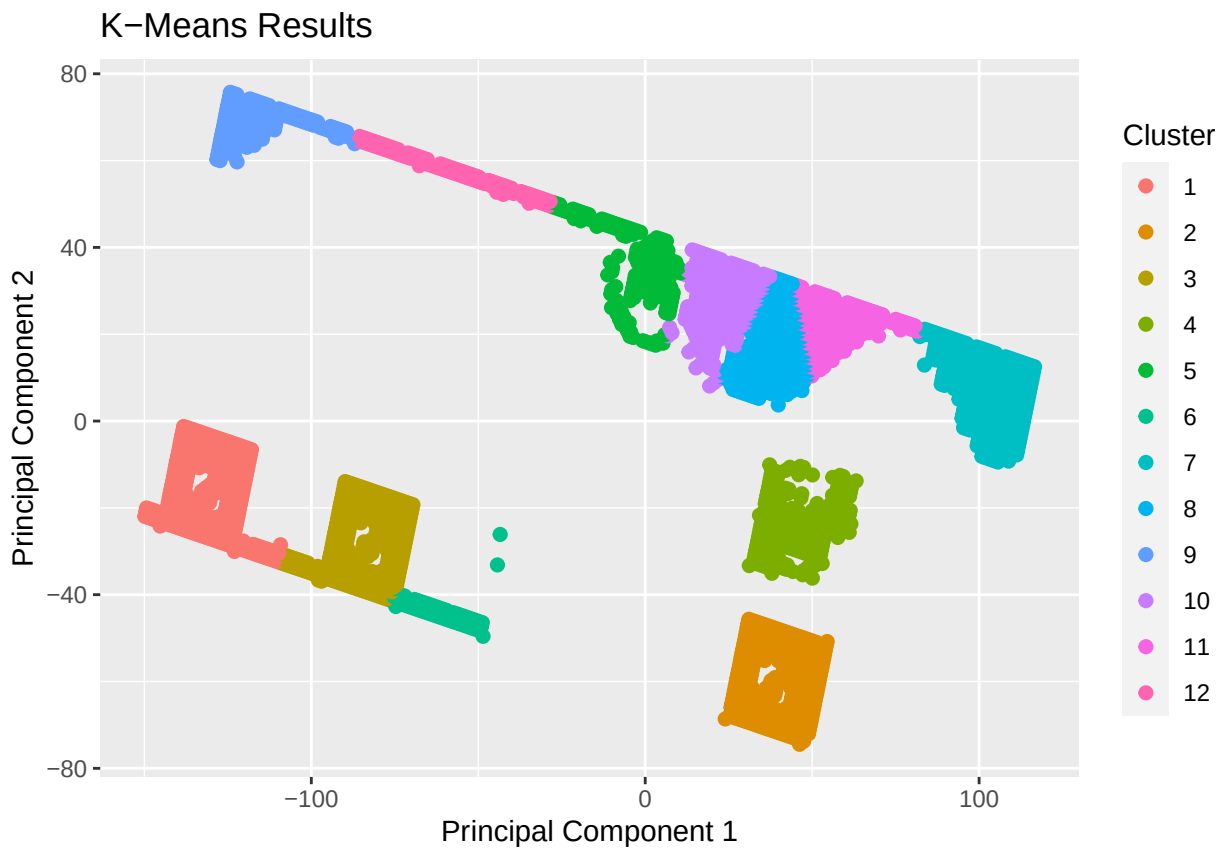

```
plot(cluster_10, data = clustering_df)
```



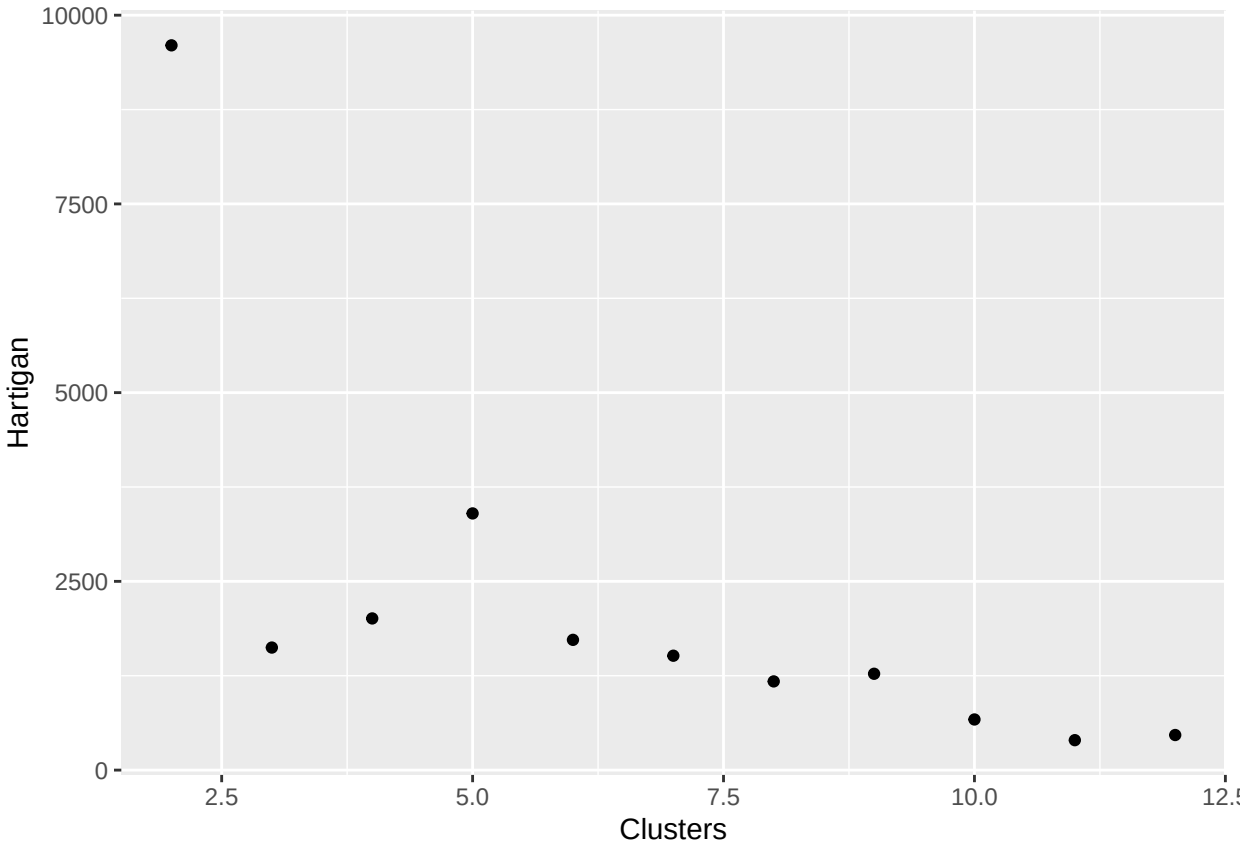

```
plot(cluster_11, data = clustering_df)
```




```
plot(cluster_12, data = clustering_df)
```




```
clusterfit <- FitKMeans(clustering_df, max.clusters = 12, nstart = 20)
ggplot(clusterfit, aes(x=Clusters, y=Hartigan)) + geom_point()
```




```
#finding the elbow point
#Commented out due to personal computer struggling to compute
#elbow <- clusGap(clustering_df, FUNcluster = pam, K.max = 12)
```

I was not sure how to get the average distance from the center of each cluster for each value of k. I think it has to deal with the Euclidean distance, but I wasn't sure how to implement that in R.