# Matthew Hawkins

## BAN 502: Module 6, Assignment 1

library("tidyverse")  
library("cluster")

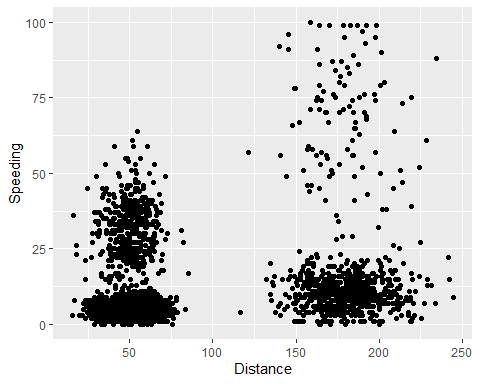
## Warning: package 'cluster' was built under R version 3.6.2

library("factoextra")

## Warning: package 'factoextra' was built under R version 3.6.2

library("dendextend")  
trucks <- read\_csv("trucks.csv")

ggplot(trucks,aes(x=Distance,y=Speeding)) +   
 geom\_point()

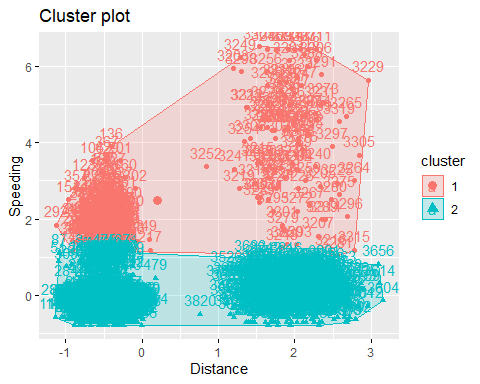


With regard to distance, we see two main clusters. One clusters around 50 miles and the other is around 150 miles. Most of the data points fall under 25% of the time spent speeding. Drivers that spend the most time speeding, all fall into the cluster that drives more per day.

trucks2 = select(trucks,-Driver\_ID)  
trucks2 = as.data.frame(scale(trucks2))  
summary(trucks2)

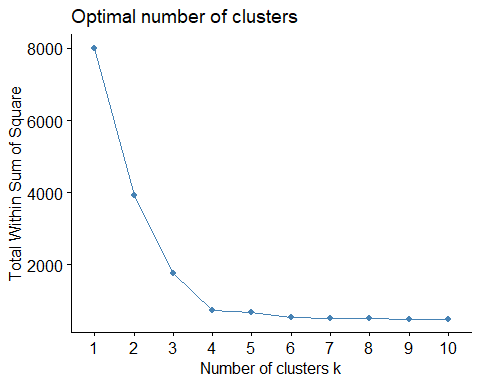
## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

set.seed(1234)  
clusters1 <- kmeans(trucks2,2)  
fviz\_cluster(clusters1,trucks2)

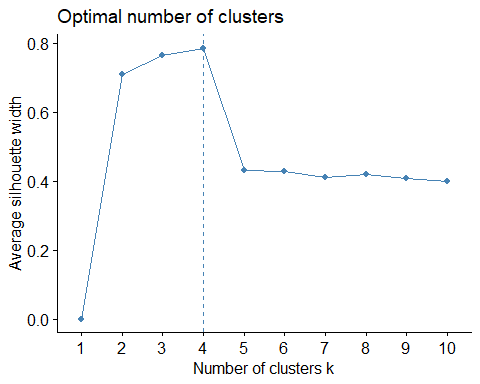


The clusters created divide the data into those that rarely speed and those that speed a lot.

set.seed(123)  
fviz\_nbclust(trucks2, kmeans, method = "wss")

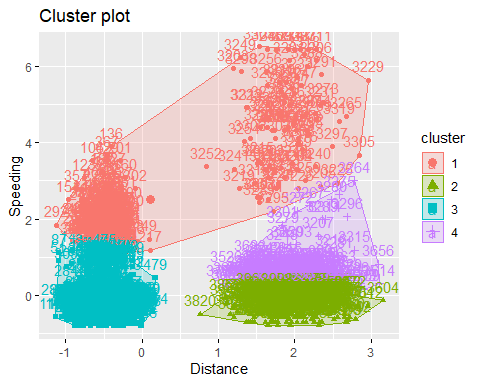


set.seed(123)  
fviz\_nbclust(trucks2, kmeans, method = "silhouette")



Both methods seem to show that 4 is the optimal number of clusters.

set.seed(1234)  
clusters2 <- kmeans(trucks2,4)  
fviz\_cluster(clusters2,trucks2)



The four clusters created in task 5 more or less follow the clusters I identified when first looking at the data. The drivers are clustered on distance and then speeding. There is a cluster for low distance/low speeding, high distance/low speeding, and high distance/moderate speeding. The most interesting cluster to me is cluster one, which encompasses the highest speeders from both distances.

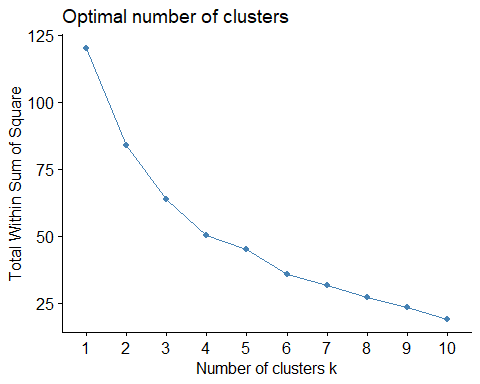
wine <- read\_csv("wineprice.csv")

## Parsed with column specification:  
## cols(  
## Year = col\_double(),  
## Price = col\_double(),  
## WinterRain = col\_double(),  
## AGST = col\_double(),  
## HarvestRain = col\_double(),  
## Age = col\_double(),  
## FrancePop = col\_double()  
## )

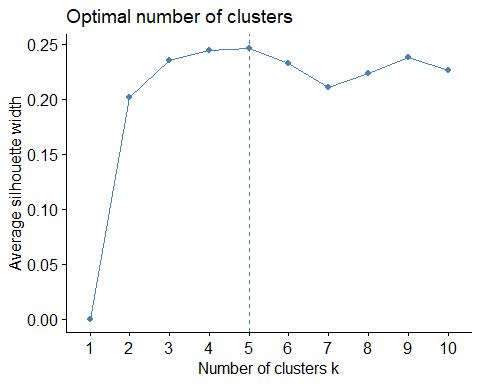
wine2 = select(wine,-Year,-FrancePop)  
wine2 = as.data.frame(scale(wine2))  
summary(wine2)

## Price WinterRain AGST   
## Min. :-1.32596 Min. :-1.73332 Min. :-2.25947   
## 1st Qu.:-0.84329 1st Qu.:-0.52375 1st Qu.:-0.45801   
## Median : 0.08284 Median :-0.03992 Median : 0.03548   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.65777 3rd Qu.: 0.69339 3rd Qu.: 0.82524   
## Max. : 2.19343 Max. : 1.69885 Max. : 1.68888   
## HarvestRain Age   
## Min. :-1.4856 Min. :-1.586   
## 1st Qu.:-0.8003 1st Qu.:-0.806   
## Median :-0.2494 Median :-0.026   
## Mean : 0.0000 Mean : 0.000   
## 3rd Qu.: 0.5165 3rd Qu.: 0.754   
## Max. : 1.9275 Max. : 1.794

set.seed(123)  
fviz\_nbclust(wine2, kmeans, method = "wss")

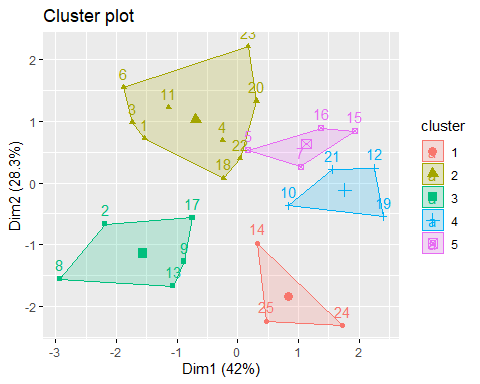


set.seed(123)  
fviz\_nbclust(wine2, kmeans, method = "silhouette")



While I don’t think it was as obvious as with the trucks data, there still is consensus between these two methods for the wine data. For the wss method, I have a hard time deciding between 4 or 5. However, since the silhouette method says 5, one of my choices, I’ll say there is consensus.

set.seed(1234)  
clusters3 <- kmeans(wine2,5)  
fviz\_cluster(clusters3,wine2)

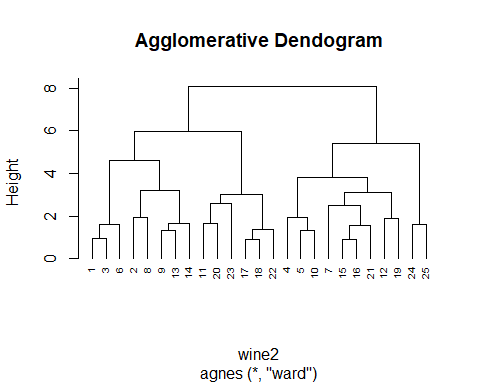


m = c("average","single","complete","ward")  
names(m) = c("average","single","complete","ward")  
ac = function(x) {agnes(wine2,method = x)$ac}  
map\_dbl(m,ac)

## average single complete ward   
## 0.5666719 0.2920143 0.7196616 0.8112139

Ward’s is the highest, at 0.81.

hc = agnes(wine2,method = "ward")  
pltree(hc,cex=0.6,hang=-1,main="Agglomerative Dendogram")



hc2 = diana(wine2)  
pltree(hc2, cex = 0.6, hang = -1, main = "Divisive Dendogram")

