# Module 6 Assignment 1

## Gregory Desloge

#install.packages("cluster")  
#install.packages("factoextra")  
#install.packages("dendextend")  
library("tidyverse")

## -- Attaching packages -------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.3.2   
## v tibble 2.1.1 v dplyr 0.8.0.1  
## v tidyr 0.8.3 v stringr 1.4.0   
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("cluster")  
library("factoextra")

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library("dendextend")

##   
## ---------------------  
## Welcome to dendextend version 1.12.0  
## Type citation('dendextend') for how to cite the package.  
##   
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## Or contact: <tal.galili@gmail.com>  
##   
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## ---------------------

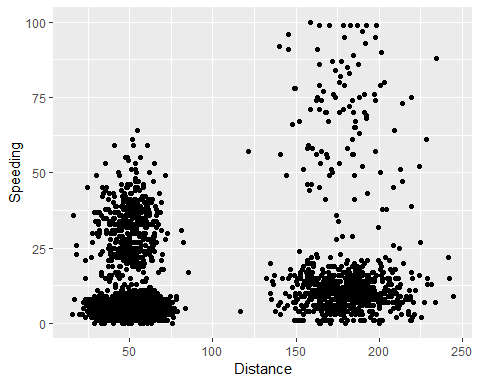
##   
## Attaching package: 'dendextend'

## The following object is masked from 'package:stats':  
##   
## cutree

trucks = read.csv("trucks.csv")  
summary(trucks)

## Driver\_ID Distance Speeding   
## Min. :3.423e+09 Min. : 15.52 Min. : 0.00   
## 1st Qu.:3.423e+09 1st Qu.: 45.25 1st Qu.: 4.00   
## Median :3.423e+09 Median : 53.33 Median : 6.00   
## Mean :3.423e+09 Mean : 76.04 Mean : 10.72   
## 3rd Qu.:3.423e+09 3rd Qu.: 65.63 3rd Qu.: 9.00   
## Max. :3.423e+09 Max. :244.79 Max. :100.00

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

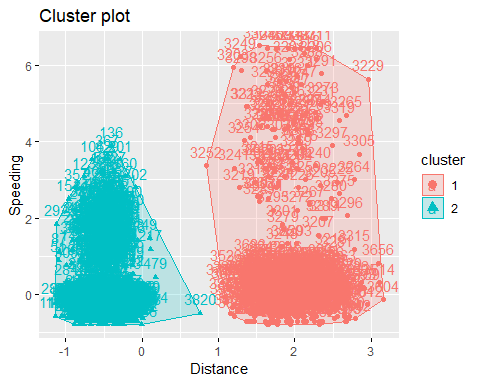


There appears to be serveral relationships when it comes to the data. There a cluster that shows low speeding and low distance, but also low speeding and high distance. This could be based on the amount of time spent driving which is not given. There is a cluster where higher speeding has meant further distance, but also a cluster where there is speeding (25-50) but not much distance gained. The data appears to have 4 separate clusters.

trucks2 = trucks  
trucks2 = trucks2 %>% select(-Driver\_ID)  
trucks2 = data.frame(scale(trucks2))  
summary(trucks2)

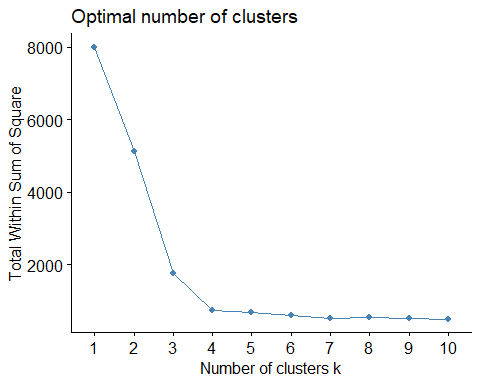
## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

set.seed(1234)  
cluster1 <- kmeans(trucks2, 2)  
fviz\_cluster(cluster1, trucks2)

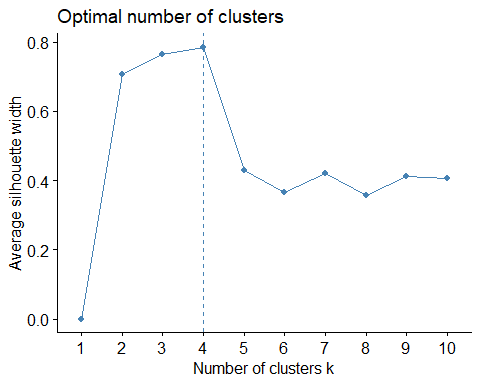


The clustering with only 2 doesn’t seem to fit. There looks like there should be 4 possible clusters as predicted from just plotting the data. The 2 clusters appear to encompass data that don’t belong together, especially in regards to distance. Both clusters have low distance and high distance grouped together.

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

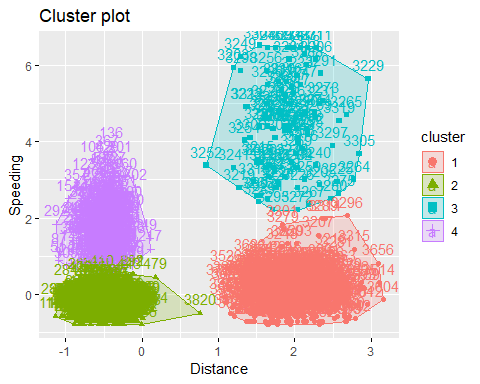


set.seed(123)  
fviz\_nbclust(trucks2, kmeans, method = "silhouette")



The optimal number of clusters is 4 when using wss as a method, it can be subjective but 4 appears to be where the data makes a turn. That would then lead me to believe that the consensus for optimal number of clusters is the same in both methods.

set.seed(1234)  
cluster2 <- kmeans(trucks2, 4)  
fviz\_cluster(cluster2, trucks2)



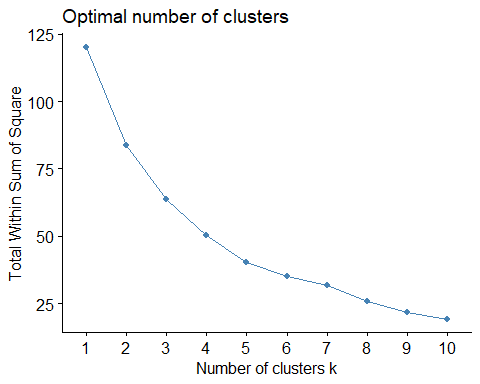
The clusters segmented themselves as I would have anticipated. Cluster 1 is the large distance and low speeding, cluster 2 is low distance and low speeding, cluster 3 is high distance and high speeding, and cluster 4 is low distance high speeding. All 4 combinations of variables were represented by separate clusters.

wine = read.csv("wineprice.csv")  
summary(wine)

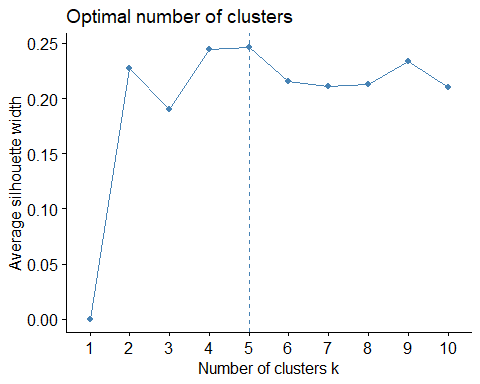
## Year Price WinterRain AGST   
## Min. :1952 Min. :6.205 Min. :376.0 Min. :14.98   
## 1st Qu.:1960 1st Qu.:6.519 1st Qu.:536.0 1st Qu.:16.20   
## Median :1966 Median :7.121 Median :600.0 Median :16.53   
## Mean :1966 Mean :7.067 Mean :605.3 Mean :16.51   
## 3rd Qu.:1972 3rd Qu.:7.495 3rd Qu.:697.0 3rd Qu.:17.07   
## Max. :1978 Max. :8.494 Max. :830.0 Max. :17.65   
## HarvestRain Age FrancePop   
## Min. : 38.0 Min. : 5.0 Min. :43184   
## 1st Qu.: 89.0 1st Qu.:11.0 1st Qu.:46584   
## Median :130.0 Median :17.0 Median :50255   
## Mean :148.6 Mean :17.2 Mean :49694   
## 3rd Qu.:187.0 3rd Qu.:23.0 3rd Qu.:52894   
## Max. :292.0 Max. :31.0 Max. :54602

wine2 = wine  
wine2 = wine2 %>% select(-Year, -FrancePop)  
wine2 = data.frame(scale(wine2))

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

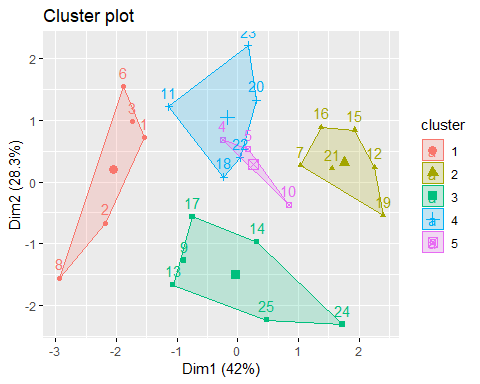


set.seed(123)  
fviz\_nbclust(wine2, kmeans, method = "silhouette")



Silhouette come up with an optimal number of 5 clusters, while looking at wss it appears that the optimal could be 5 or 6. I would not say there is a clear concensus on the exact optimal number of clusters.

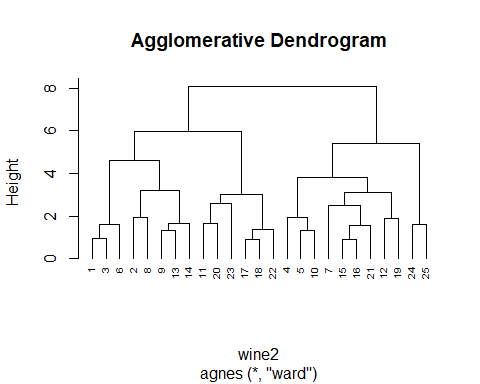
set.seed(1234)  
cluster3 <- kmeans(wine2, 5)  
fviz\_cluster(cluster3, wine2)



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

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



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

