# Module 6 - Cluster Assignment

## Cooper, Sarah

library(tidyverse)

library(cluster)  
library(factoextra)

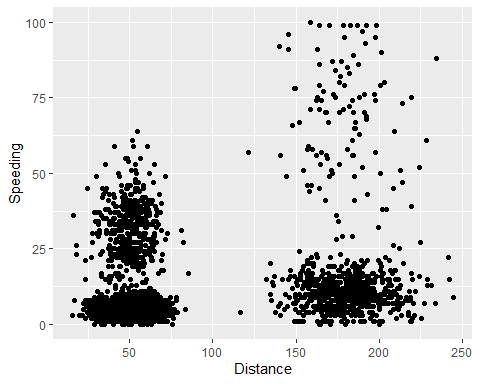
library(dendextend)

trucks <- read\_csv("C:/Users/Sarah/Downloads/trucks.csv")

## Parsed with column specification:  
## cols(  
## Driver\_ID = col\_double(),  
## Distance = col\_double(),  
## Speeding = col\_double()  
## )

# Task 1

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



*There does appear to be two natural clusterings of drivers. One, where the driver travels about 50 miles per day, and the other where the driver travels around 175 miles per day.*

# Task 2

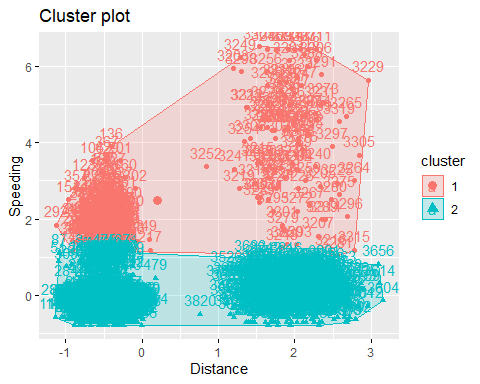
trucks2 = scale(as.data.frame(trucks[c(2:3)]))  
summary(trucks2)

## 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

# Task 3

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

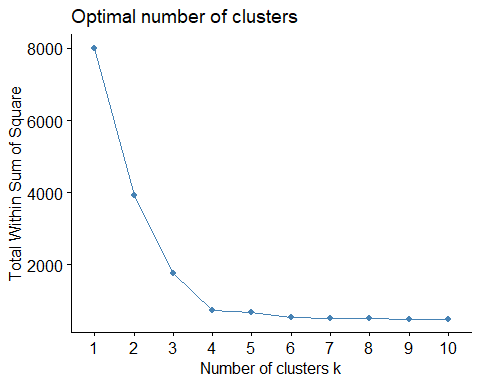
fviz\_cluster(clusters1, trucks2)



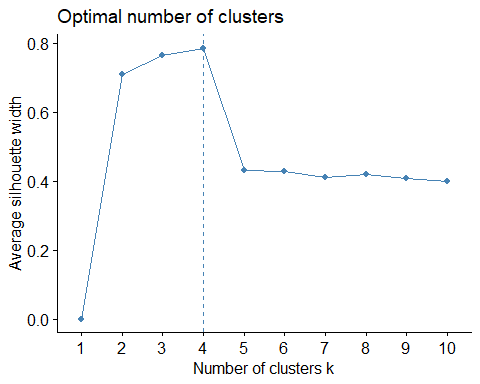
*There are two significant clusters. One belonging to the 1.5 - 2.5 Distance grouping with a Speeding rank of 0. The second also in the Distance grouping of -1 to 0 then separated into Speeding groups of 0 and 2.*

# Task 4

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



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



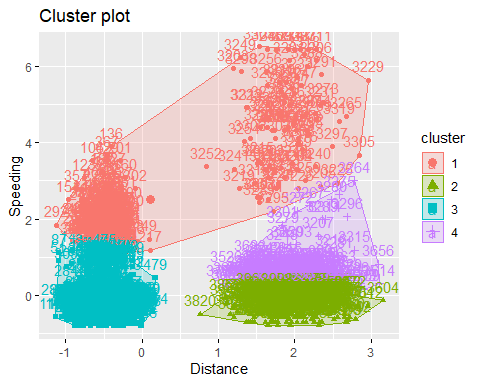
*According to each model, the optimal number of clusters would be 4.*

# Task 5

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

## K-means clustering with 4 clusters of sizes 399, 463, 2900, 238  
##   
## Cluster means:  
## Distance Speeding  
## 1 0.1099584 2.5111614  
## 2 1.9532586 -0.2399259  
## 3 -0.4874490 -0.3463710  
## 4 1.9553359 0.4773451  
##   
  
## Within cluster sum of squares by cluster:  
## [1] 1135.73431 79.87293 385.95989 81.27803  
## (between\_SS / total\_SS = 79.0 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

fviz\_cluster(clusters2, trucks2)



# Task 6

*I would characterize the clusters displayed in Task 5 as clearly defined. We don’t have much, if any, overlap of clusters - They appear to be independent of one another.*

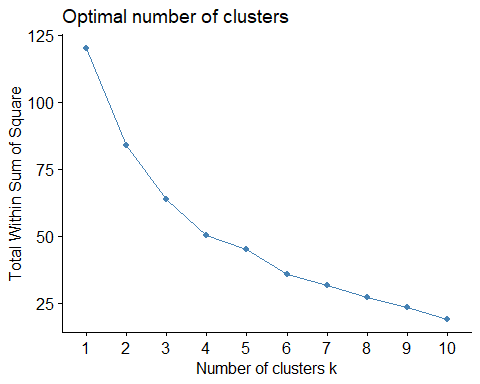
wineprice <- read\_csv("C:/Users/Sarah/Downloads/wineprice.csv")

## 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()  
## )

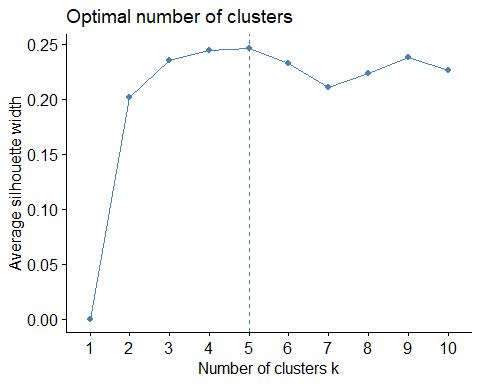
wine <- wineprice  
wine2 = scale(as.data.frame(wine[c(2:6)]))

# Task 7

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



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



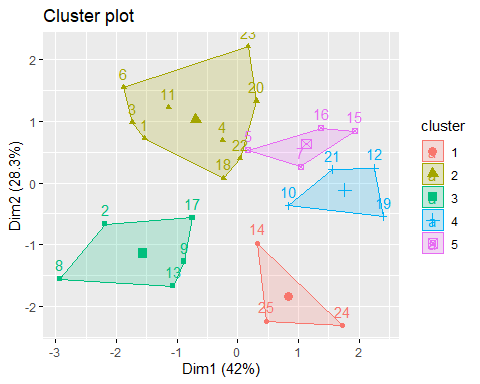
*I would say these two models land on an optimal cluster of 5 (silhouette) or 6 (wss). For purposes of moving forward, let’s choose 5.*

# Task 8

set.seed(1234)  
clusters2 <- kmeans(wine2, 5)  
clusters2

## K-means clustering with 5 clusters of sizes 3, 9, 5, 4, 4  
##   
## Cluster means:  
## Price WinterRain AGST HarvestRain Age  
## 1 -0.4700981 1.2150172 -0.9351085 -0.84959852 -1.0660444  
## 2 0.3739208 -0.9420566 0.8224590 -0.00155276 0.1473395  
## 3 1.2047467 0.9549587 0.3020210 -1.06907515 0.4420184  
## 4 -1.0784329 -0.1911127 -1.3586990 0.57028091 -0.3510146  
## 5 -0.9162487 0.2057788 -0.1680286 1.40675562 0.2665111  
##   
## Clustering vector:  
## [1] 2 3 2 2 5 2 5 3 3 4 2 4 3 1 5 5 3 2 4 2 4 2 2 1 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 3.498321 24.609221 7.991457 5.549316 4.752018  
## (between\_SS / total\_SS = 61.3 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

fviz\_cluster(clusters2, wine2)

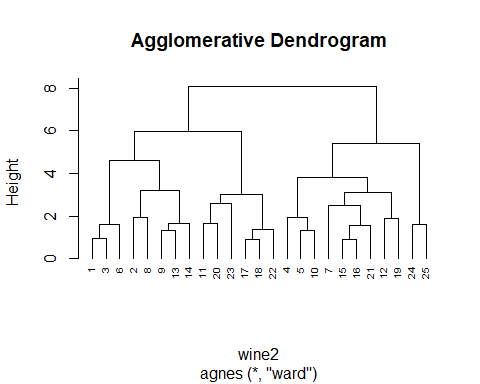


# Task 9

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") #change ward to other method if desired  
pltree(hc, cex = 0.6, hang = -1, main = "Agglomerative Dendrogram")



# Task 10

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

