## Module 6 Clustering Assignment

## Shealyn Wilson

options(tidyverse.quiet=TRUE)  
library(tidyverse)  
library(cluster)  
library(factoextra)  
library(dendextend)

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

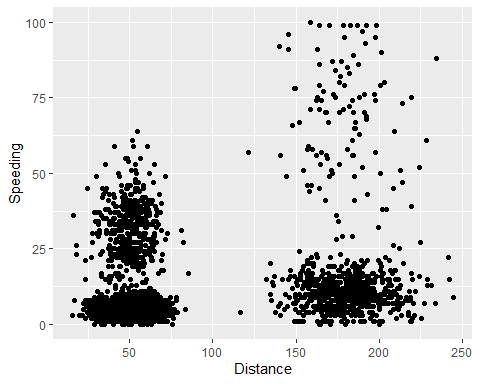
## Parsed with column specification:  
## cols(  
## Driver\_ID = col\_double(),  
## Distance = col\_double(),  
## Speeding = col\_double()  
## )

# summary(trucks)  
# str(trucks)

### TASK 1

There are 3 concentrated groupings of the distance and speeding data with a 4th group of points that have more spread and variance from each other.

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



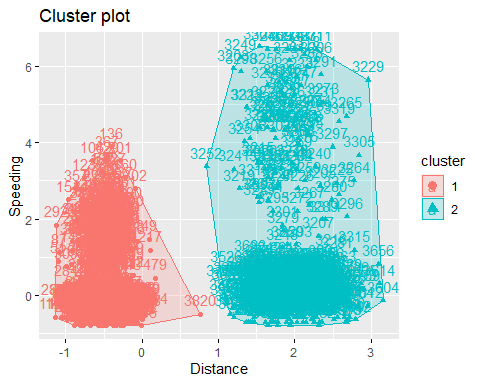
### TASK 2

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

### TASK 3

There are 2 groups of clusters with cluster 2 having a greater spread in speed and distance.

set.seed(64)  
cluster1 = kmeans(trucks2, 2)  
fviz\_cluster(cluster1, trucks2)

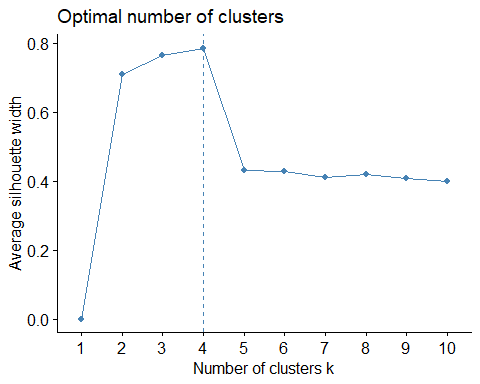
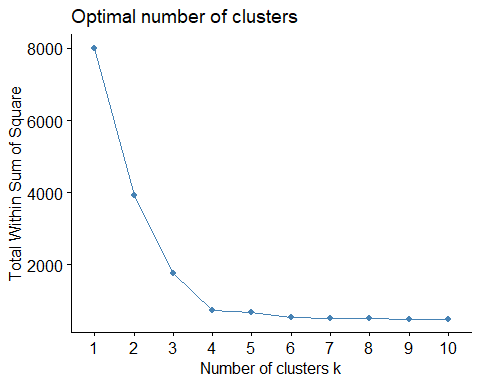


### TASK 4

Looking at other methods gives an suggested optimal cluster of 4.

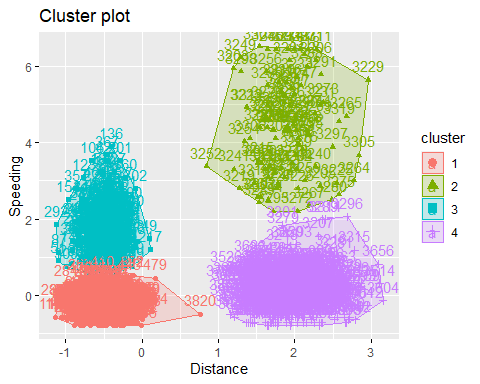
set.seed(64)  
fviz\_nbclust(trucks2, kmeans, method="wss")

fviz\_nbclust(trucks2, kmeans, method="silhouette")



### TASK 5

set.seed(64)  
cluster2 = kmeans(trucks2, 4)  
fviz\_cluster(cluster2, trucks2)



### TASK 6

A cluster of 4 is what was predicted when first plotting the data and is a more optimal grouping. There are the 3 distinct groups of data with points closely grouped and the fourth group with data points more spread out.

bball <- read\_csv("kenpom20.csv")

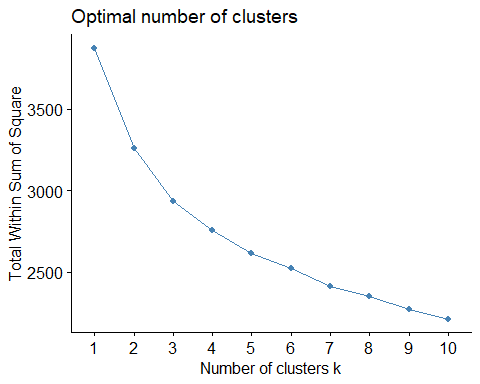
## Parsed with column specification:  
## cols(  
## TeamName = col\_character(),  
## AdjTempo = col\_double(),  
## AdjOE = col\_double(),  
## AdjDE = col\_double(),  
## eFGPct = col\_double(),  
## TOPct = col\_double(),  
## ORPct = col\_double(),  
## FTRate = col\_double(),  
## eFGPctD = col\_double(),  
## TOPctD = col\_double(),  
## ORPctD = col\_double(),  
## FTRateD = col\_double()  
## )

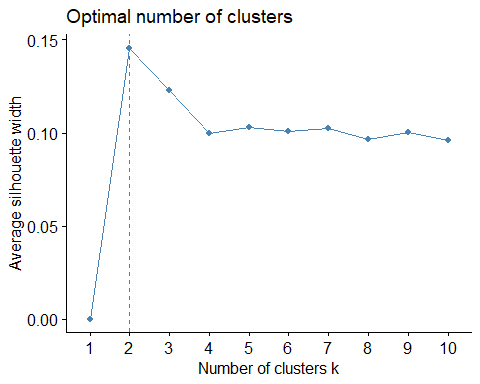
# summary(bball)

### TASK 7

There is not a strong consensus between the two methods of a optimal cluster. Elbow method cluster group could be 3 or 4. The silhouette method suggests 2 and another a leveling at 4. I would suggest trying a cluster of 4 or 3.

bball2 = bball%>%  
 select(-TeamName)  
  
bball2\_scaled =scale(bball2)  
# summary(ball2\_scaled)  
  
set.seed(123)  
fviz\_nbclust(bball2\_scaled, kmeans, method="wss")

fviz\_nbclust(bball2\_scaled, kmeans, method="silhouette")

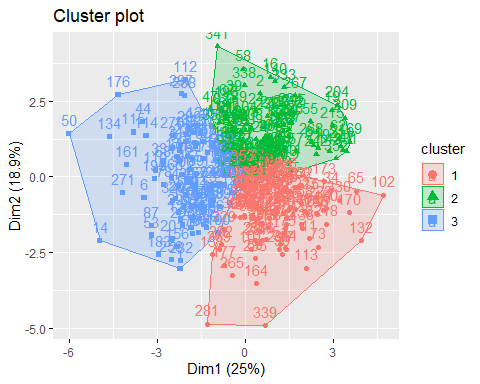
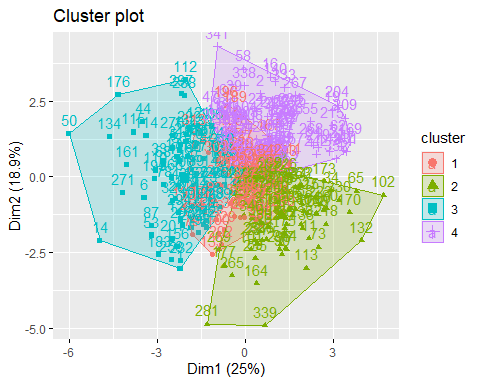


### TASK 8

A cluster of 4 has a lot of overlapping with the first cluster. A cluster of 3 is better and there is minimal if any overlapping.

set.seed(1234)  
bball2\_cluster1 = kmeans(bball2\_scaled, 4)  
fviz\_cluster(bball2\_cluster1, bball2\_scaled)

bball2\_cluster2 = kmeans(bball2\_scaled, 3)  
fviz\_cluster(bball2\_cluster2, bball2\_scaled)



### TASK 9

There is relationship between the AjdOE and AdjDE where cluster 3 is a group of teams that allow a greater number of points on defense but fewest points scored on offense. Cluster 2 is a group of teams that scores the greastest on offense and allows the fewest points on defense. Cluster 1 falls in the middle overlapping the 3rd and 2nd cluster of teams. Cluster 4 of teams is scattered and overlaps the other 3 clusters greatly.

bball2 = bball2 %>% mutate(clusternum = bball2\_cluster1$cluster)  
  
ggplot(bball2, aes(x=AdjOE, y=AdjDE, color=factor(clusternum)))+  
 geom\_point()

