# Clustering

## BAN 502

### Stephen Kiser

library(tidyverse)

## -- Attaching packages ------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

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

library(cluster)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(dendextend)

##   
## ---------------------  
## Welcome to dendextend version 1.13.4  
## 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'

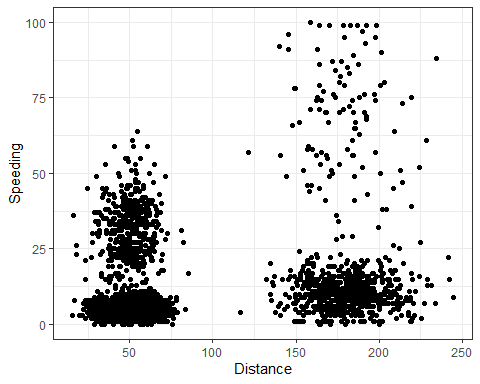
## The following object is masked from 'package:stats':  
##   
## cutree

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

## 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() + theme\_bw()



After plotting Distance versus the time the truck drivers are driving over 5 miles per hour I can see 3 distinct groups, and a fourth cluster that is widely spaced out. We see that most drivers who drove less than 100 miles either clustered in the less than 10 % of the distance or in another cluster from 10% to 65% of the distance. The third and fourth cluster we see is the drivers who drove 100+ miles. The third cluster is those drivers who only drove 5 miles over for 0% to 25% of the distance, and the fourth cluster is from 25% to 100% of the distance.

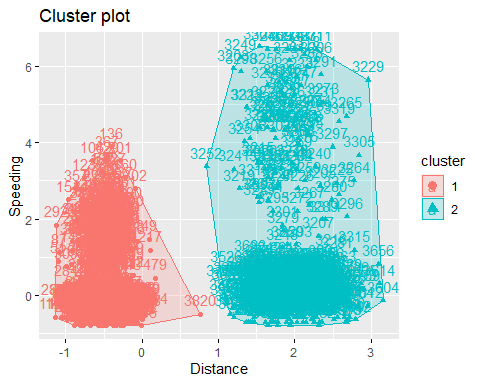
### Task 2

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

## 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

### Task3

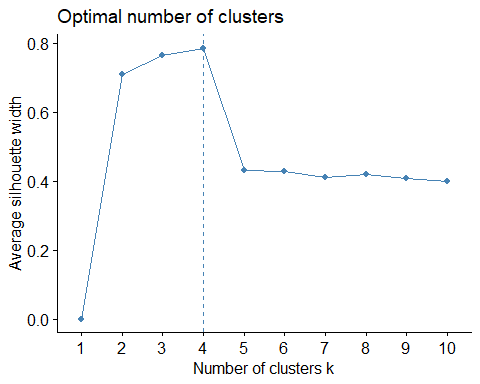
set.seed(64)  
cluster1 <- kmeans(trucks2\_scaled, 2)  
  
fviz\_cluster(cluster1, trucks2\_scaled)



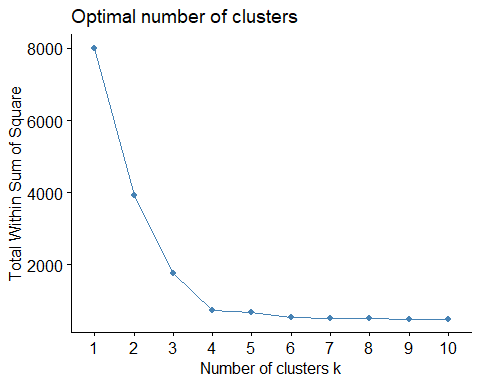
After just putting all the data into two clusters we see that one cluster is the drivers who drove the shorter distance and those who drove the further distance. These are bigger clusters then I would expect. I do not think this is the optimal clustering.

### Task 4

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



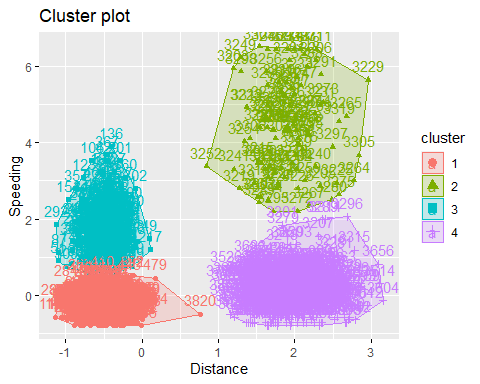
fviz\_nbclust(trucks2\_scaled, kmeans, method = "wss")



The first graph we use to find the optimum amount of clusters we see that after 4 clusters the clustering becomes less efficient. Then in the second graph we do not see this change in clusters as drastic. The second graph decreases after 1 cluster, and we see at 4 clusters is when the graph really starts to level out. Both graphs show an optimal number of 4 clusters.

### Task 5

set.seed(64)  
clusters2 <- kmeans(trucks2\_scaled, 4)  
fviz\_cluster(clusters2, trucks2\_scaled)



### Task 6

I would characterize the red cluster as the group of truckers who traveled a short distance, but rarely went over 5 miles per hour. The blue cluster as the group of truckers who traveled a short distance, and drove over the speed limit longer than the red cluster. The purple cluster is the data collected of truckers who drove a further distance than the first two clusters, and these truckers drove over 5 miles per hour minimal. Then our last cluster, the green one, is truckers who drove a longer distance, and sped for most of the trip to all of the trip.

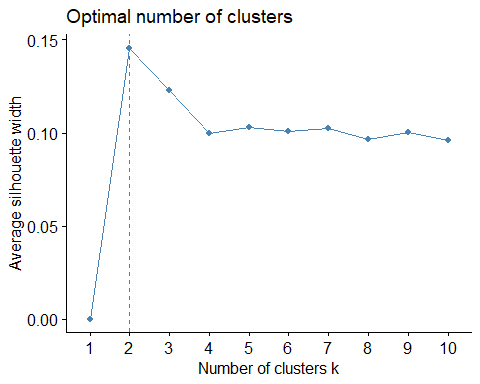
### Task 7

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

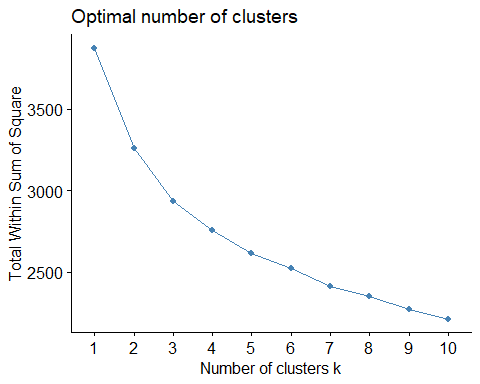
## Parsed with column specification:  
## cols(  
## TeamName = col\_character(),  
## AdjTempo = col\_double(),  
## AdjOE = col\_double(),  
## AdjDE = col\_double(),  
## eFGPct = col\_double(),  
## TOPct = col\_double(),  
## ORPct = col\_double(),  
## FTRate = col\_double(),  
## eFGPctD = col\_double(),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

bball2 <- bball %>% select(-TeamName)  
  
bball2\_scaled <- as.data.frame(scale(bball2))

set.seed(123)  
fviz\_nbclust(bball2\_scaled, kmeans, method = "silhouette")



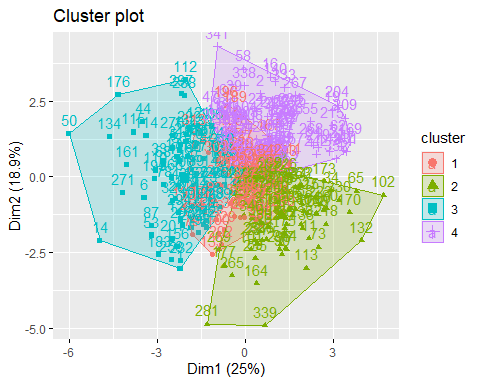
fviz\_nbclust(bball2\_scaled, kmeans, method = "wss")



There is not a consensus between the two graphs. The graph using silhouettes has 2 clusters as the optimal amount. However, once it hits 4 clusters the data starts to remain at the same level. Then in the second graph using wss method we do not see a distinct difference in the graph. There is no jump up or down anywhere that really shows where the optimal amount of clusters are. After cluster 5 we see a slight increase at 6 clusters, and then it goes back to a the pattern decrease. We could say that five clusters could be the optimal amount.

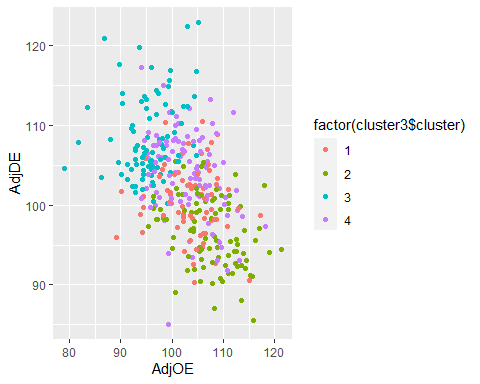
### Task 8

set.seed(1234)  
cluster3 <- kmeans(bball2\_scaled, 4)  
fviz\_cluster(cluster3, bball2\_scaled)



### Task 9

bball2 <- bball2 %>% mutate(clusternum = cluster3$cluster)  
  
ggplot(bball2, aes(AdjOE, AdjDE, color = factor(cluster3$cluster))) +geom\_point()



After creating a ggplot of the data of points scored against points allowed we can see the clusters relationships. The red cluster is the teams who allowed on defense almost the same amount of points they scored. Then the green cluster is the teams that scored more points and did not allow the opposition to score as much. The blue cluster is the teams that allowed the opposition to score more than them, and finally the purple cluster is similar to the red cluster in that the majority of data was that both the opposition and team scored about the same amount of points. HOwever, the purple cluster is more spread out than the red cluster is.