## Week 6 Assigment 1

### DEY, Sankha

#### Clustering Assignment

# message=FALSE done before last knit  
library(tidyverse)  
library(cluster)  
library(factoextra)  
library(dendextend)

Read Dataset

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

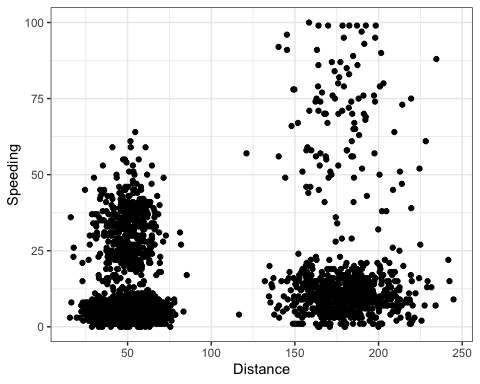
## Parsed with column specification:  
## cols(  
## Driver\_ID = col\_double(),  
## Distance = col\_double(),  
## Speeding = col\_double()  
## )

str(trucks)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 4000 obs. of 3 variables:  
## $ Driver\_ID: num 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num 28 25 27 22 25 10 20 8 34 19 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Driver\_ID = col\_double(),  
## .. Distance = col\_double(),  
## .. Speeding = col\_double()  
## .. )

#### Task 1

ggplot(trucks,aes(Distance,Speeding)) + geom\_point() + theme\_bw()



Distance and Speeding are kind of positively correlated with average Distance traveled each drive each day increasing, there is also a increasing trend of Speedings. From above plot, there are four natural clusterings of drivers. Frist cluster is at Distance traveled around 25 miles to 75 miles per day with number of Speedings between zero and twelve. Second cluster is at Distance traveled around 25 miles to 75 miles per day with number of Speedings clusterd around 25 range. Third cluster is at Distance traveled around 150 miles to 225 miles per day with number of Speedings between zero and twenty five. Fourth cluster at Distance traveled around 150 miles to 225 miles per day with number of Speedings between twenty five and hundred range. This last cluster is distributed over a long range of Speeding variable.

#### Task 2

trucks2 = trucks %>% select(Distance, Speeding)  
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

#### Task 3

set.seed(64)  
clusters1 <- kmeans(trucks2, 2)  
#clusters1 #Commented before final knit to reduce overall dcoument size.

Visualize the clustering

fviz\_cluster(clusters1, trucks2)

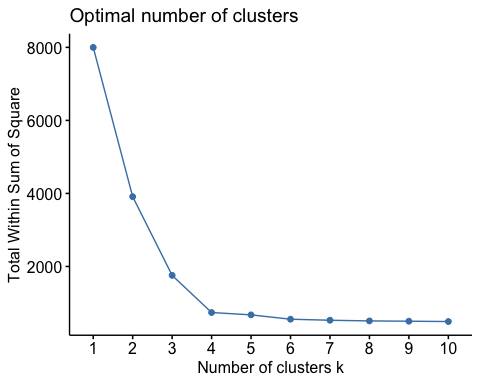


Using k-Means clustering with only two clusters, we see data have been divided (clustered) into two sections based on the Distance traveled per day variable. One cluster is representing lower range of distance traveled whereas the other representing greater numbers.

#### Task 4

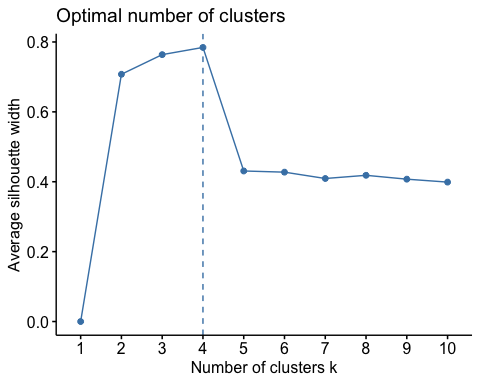
Visually identify optimal number of clusters

set.seed(64)  
fviz\_nbclust(trucks2, kmeans, method = "wss") #minimize within-cluster variation



Another method

set.seed(64)  
fviz\_nbclust(trucks2, kmeans, method = "silhouette") #maximize how well points sit in their clusters



In WSS meethod, The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.That point here is 4. Similarly, the Silhouette method is also showing the optimal number of clusters as 4. So, these two methods are showing same output.

#### Task 5

Let’s try 4 clusters

set.seed(64)  
clusters2 = kmeans(trucks2, 4)  
fviz\_cluster(clusters2, trucks2)



Four clusters are shown in four different colors.

#### Task 6

From above plot, there are four clusterings of drivers. - The frist cluster is for drivers who travel less distance with less number of Speeding instances. They can be considerd as safe drivers.  
- The second cluster is for drivers who travel less distance with more number of Speeding instances. They are in moderate risk zone interms of safety.  
- The third cluster is for drivers who travel more distance with less number of Speeding instances. They can also be considered as safe drivers.  
- The fourth cluster is for drivers who travel more distance with more number of Speeding instances. This last cluster is distributed over a long range of Speeding variable. These drivers in high risk zone interms of safety.

### 2nd Assignment

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)  
str(bball)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 353 obs. of 12 variables:  
## $ TeamName: chr "Abilene Christian" "Air Force" "Akron" "Alabama" ...  
## $ AdjTempo: num 68.4 68.8 69.4 74.8 67.7 ...  
## $ AdjOE : num 99.1 107.8 108.5 111 87.9 ...  
## $ AdjDE : num 100.2 110.3 98.9 99.5 108.2 ...  
## $ eFGPct : num 49.6 52.9 51.7 52.2 42.4 ...  
## $ TOPct : num 20.6 18.1 17.9 19.2 18.8 ...  
## $ ORPct : num 28.8 22.5 27.4 29.9 24.9 ...  
## $ FTRate : num 41.1 35 35.5 36.8 32.7 ...  
## $ eFGPctD : num 50.4 56.2 45.9 47.9 51.3 ...  
## $ TOPctD : num 26.1 16.4 18.4 18.2 18.2 ...  
## $ ORPctD : num 29.5 22.8 29.2 30.8 27 ...  
## $ FTRateD : num 48.9 29.8 29.1 35.9 35.2 ...  
## - attr(\*, "spec")=  
## .. 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()  
## .. )

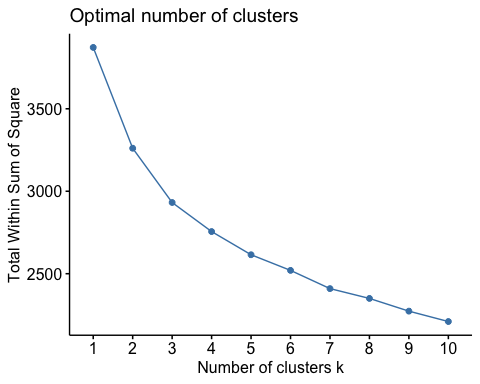
#### Task 7

bball2 = bball %>% select(-TeamName)  
bball2 = as.data.frame(scale(bball2))  
summary(bball2)

## AdjTempo AdjOE AdjDE eFGPct   
## Min. :-3.39100 Min. :-3.45199 Min. :-2.6678124 Min. :-3.68381   
## 1st Qu.:-0.67255 1st Qu.:-0.73679 1st Qu.:-0.6475704 1st Qu.:-0.69461   
## Median :-0.02436 Median : 0.01391 Median :-0.0002663 Median :-0.01141   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.0000000 Mean : 0.00000   
## 3rd Qu.: 0.58621 3rd Qu.: 0.67648 3rd Qu.: 0.6199679 3rd Qu.: 0.68688   
## Max. : 3.31633 Max. : 2.79447 Max. : 3.1546440 Max. : 3.62001   
## TOPct ORPct FTRate eFGPctD   
## Min. :-2.61437 Min. :-3.420873 Min. :-2.25314 Min. :-3.03315   
## 1st Qu.:-0.57807 1st Qu.:-0.676410 1st Qu.:-0.72045 1st Qu.:-0.71103   
## Median :-0.05269 Median :-0.005597 Median :-0.01699 Median :-0.02033   
## Mean : 0.00000 Mean : 0.000000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.66307 3rd Qu.: 0.690177 3rd Qu.: 0.65294 3rd Qu.: 0.68576   
## Max. : 3.78116 Max. : 3.061926 Max. : 3.17766 Max. : 3.17549   
## TOPctD ORPctD FTRateD   
## Min. :-2.1515 Min. :-3.0625871 Min. :-2.1989   
## 1st Qu.:-0.6650 1st Qu.:-0.6857607 1st Qu.:-0.6469   
## Median :-0.1072 Median :-0.0001725 Median :-0.1504   
## Mean : 0.0000 Mean : 0.0000000 Mean : 0.0000   
## 3rd Qu.: 0.6255 3rd Qu.: 0.6534689 3rd Qu.: 0.5913   
## Max. : 3.9482 Max. : 3.0506866 Max. : 3.4186

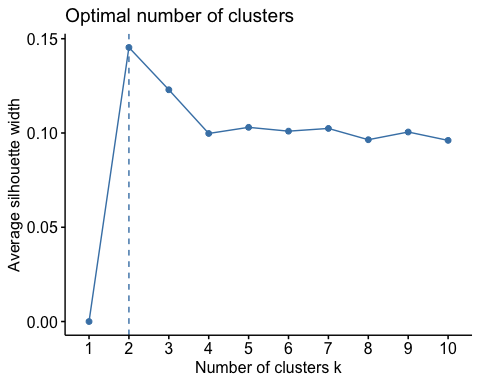
Visually identify optimal number of clusters

set.seed(123)  
fviz\_nbclust(bball2, kmeans, method = "wss") #minimize within-cluster variation



Another method

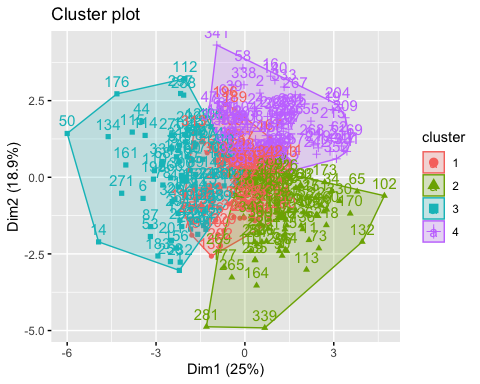
set.seed(123)  
fviz\_nbclust(bball2, kmeans, method = "silhouette") #maximize how well points sit in their clusters



Optimal number of clusters is not very clear from WSS meethod. From the plot, its difficult to determine the location of bend. May be 2 or may be 3. However, the Silhouette method is showing the optimal number of clusters as 2. Not a very clear visible consensus between two methods.

#### Task 8

set.seed(1234)  
clusters3 <- kmeans(bball2, 4)  
fviz\_cluster(clusters3, bball2)



Four clusters are created with one cluster is overlapping on relaining three.

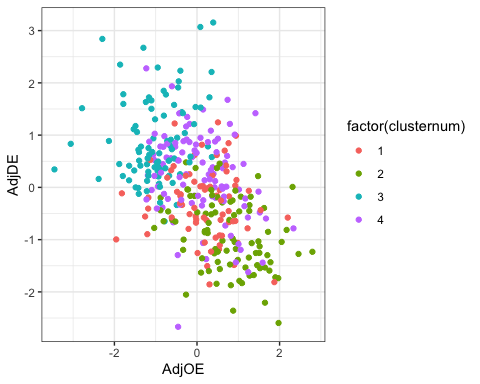
#### Task 9

Attach cluster to dataset

cluster = data.frame(clusters3$cluster)  
#bball2 = bind\_cols(bball2,cluster)  
bball2 = bball2 %>% mutate(clusternum = clusters3$cluster)  
str(bball2)

## 'data.frame': 353 obs. of 12 variables:  
## $ AdjTempo : num 0.0183 0.1577 0.4078 2.4344 -0.235 ...  
## $ AdjOE : num -0.492 0.796 0.898 1.266 -2.132 ...  
## $ AdjDE : num -0.337 1.214 -0.531 -0.449 0.888 ...  
## $ eFGPct : num 0.0254 1.1821 0.7709 0.9326 -2.5489 ...  
## $ TOPct : num 0.8177 -0.4197 -0.48 0.1371 -0.0689 ...  
## $ ORPct : num 0.227 -1.349 -0.135 0.503 -0.746 ...  
## $ FTRate : num 1.73346 0.487 0.59672 0.86277 0.00375 ...  
## $ eFGPctD : num 0.278 2.386 -1.323 -0.619 0.61 ...  
## $ TOPctD : num 3.176 -1.094 -0.202 -0.33 -0.327 ...  
## $ ORPctD : num 0.507 -1.769 0.416 0.973 -0.342 ...  
## $ FTRateD : num 2.719 -0.501 -0.622 0.518 0.405 ...  
## $ clusternum: int 3 4 2 2 3 3 4 3 1 1 ...

#ggplot(bball,aes(AdjOE,AdjDE,color=factor(clusternum))) + geom\_point() + theme\_bw()  
ggplot(bball2,aes(AdjOE,AdjDE,color= factor(clusternum))) + geom\_point() + theme\_bw()



There are four clusters -  
- The first cluster is for teams with more points allowed and less points scored.  
- The second and third cluster (in red and purple) are kind of overlapped. They are the teams allowing similar number of points allowed vs points scored.  
- The fourth cluster is for winners (in light green). They scored more points and allowed less.