# Module 6: Assignment 1

## Clustering

### Rachel Urban

library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.3.0  
## v tibble 2.0.1 v dplyr 0.7.8  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.1 v forcats 0.3.0

## -- 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 1.9.0  
## 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")

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

str(trucks)

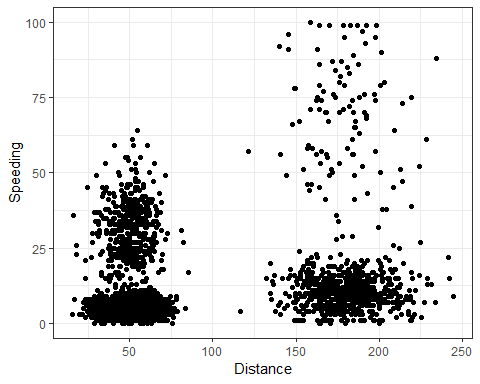
## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 4000 obs. of 3 variables:  
## $ Driver\_ID: num 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num 28 25 27 22 25 10 20 8 34 19 ...  
## - attr(\*, "spec")=  
## .. cols(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

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

### Task 1

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



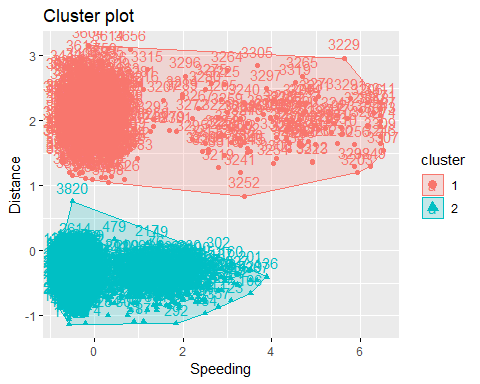
Yes, there appears to be maybe 3-4 natural clusters. One where some drivers speed for short distances, some for long distances, some where they dont speed for long distances, and some where they don’t speed for short distances.

### Task 2

trucks2 <- select(trucks,"Speeding","Distance")  
trucks2 <- as.data.frame(scale(trucks2))

### Task 3

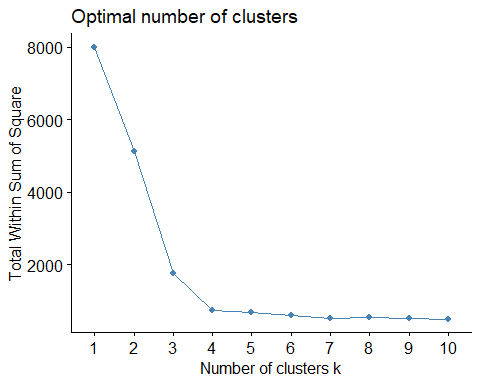
set.seed(1234)  
trucks\_cluster <- kmeans(trucks2,2)  
fviz\_cluster(trucks\_cluster,trucks2)



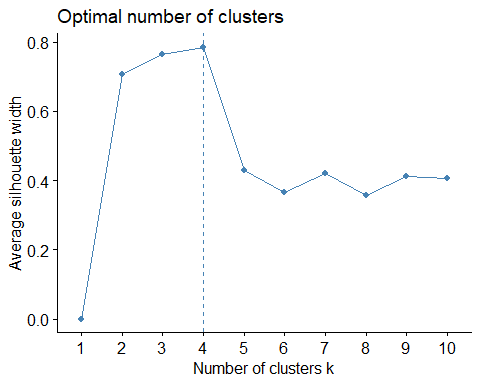
Cluster 1 looks like it might be able to be broken into two seperate clusters and Cluster 2 looks okay but could probably be splitten into two seperate clusters as well.

### Task 4

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



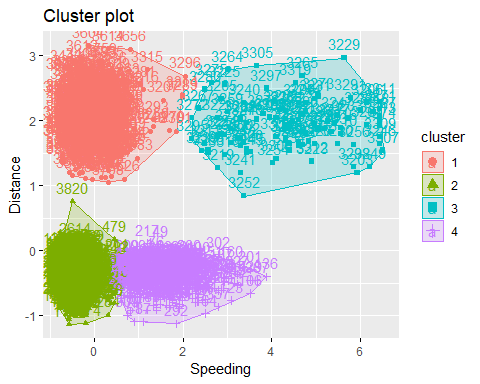
set.seed(123)  
fviz\_nbclust(trucks2,kmeans,method="silhouette")



Yes there is a consensus between the two methods as both show the optimal number of clusters to be 4.

### Task 5

set.seed(1234)  
trucks\_cluster2 <- kmeans(trucks2,4)  
fviz\_cluster(trucks\_cluster2,trucks2)



### Task 6

I would characterize Cluster 1 (pink cluster) as the drivers who drive far but do not speed. I would characterize Cluster 2 (green cluster) as the drivers that don’t speed for short distances. I would characterize Cluster 3 (blue cluster) as the drivers who speed when they drive long distances. Finally, I would characterize Cluster 4 (purple cluster) as the driver who speed while going short distances.

wine <- read\_csv("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()  
## )

str(wine)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 25 obs. of 7 variables:  
## $ Year : num 1952 1953 1955 1957 1958 ...  
## $ Price : num 7.5 8.04 7.69 6.98 6.78 ...  
## $ WinterRain : num 600 690 502 420 582 485 763 830 697 608 ...  
## $ AGST : num 17.1 16.7 17.1 16.1 16.4 ...  
## $ HarvestRain: num 160 80 130 110 187 187 290 38 52 155 ...  
## $ Age : num 31 30 28 26 25 24 23 22 21 20 ...  
## $ FrancePop : num 43184 43495 44218 45152 45654 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Year = col\_double(),  
## .. Price = col\_double(),  
## .. WinterRain = col\_double(),  
## .. AGST = col\_double(),  
## .. HarvestRain = col\_double(),  
## .. Age = col\_double(),  
## .. FrancePop = col\_double()  
## .. )

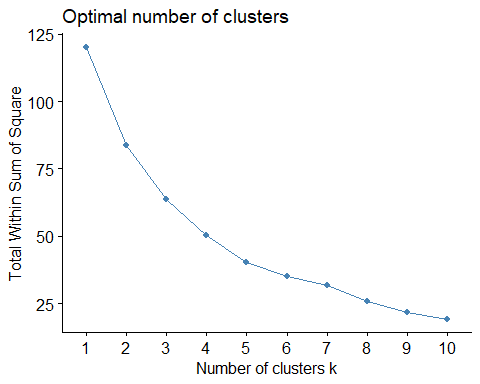
summary(wine)

## 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

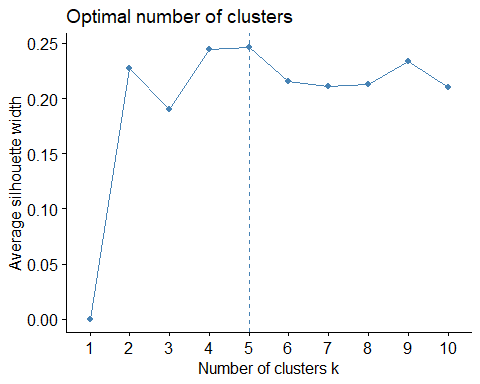
wine2 <- select(wine,"Price","WinterRain","AGST","HarvestRain","Age")  
wine2 <- as.data.frame(scale(wine2))

### Task 7

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



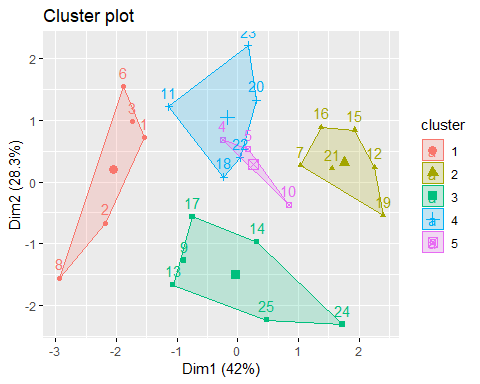
set.seed(123)  
fviz\_nbclust(wine2,kmeans,method="silhouette")



There looks to be a consense between the two methods for the optimal number of clusters. The first method looks like the optimal cluster number could be 5 or 6 and the optimal cluster number with the second method is 5 clusters.

### Task 8

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

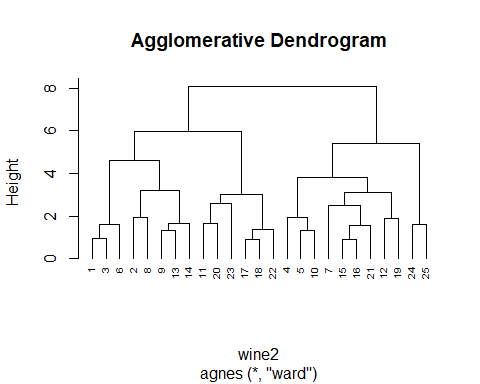


### Task 9

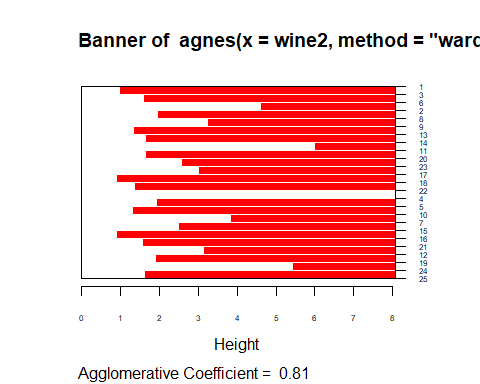
m = c( "average", "single", "complete", "ward")  
names(m) = c( "average", "single", "complete", "ward")  
  
ac = function(x) {  
 agnes(wine2, method = x)$ac  
}  
map\_dbl(m, ac)

## average single complete ward   
## 0.5666719 0.2920143 0.7196616 0.8112139

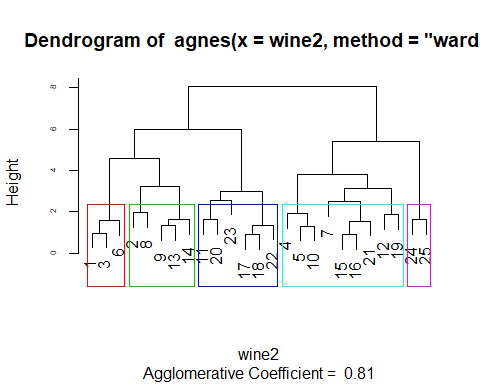
hc = agnes(wine2, method = "ward")  
pltree(hc, cex = 0.6, hang = -1, main = "Agglomerative Dendrogram")



plot(hc, cex.axis= 0.5)

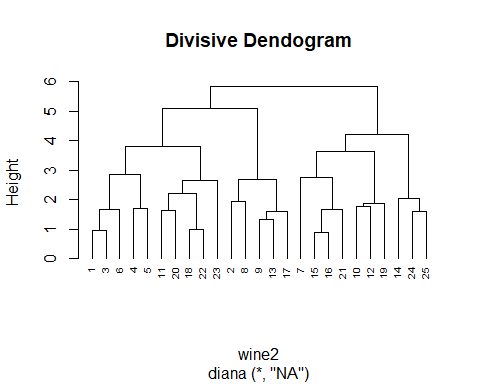


rect.hclust(hc, k = 5, border = 2:6)

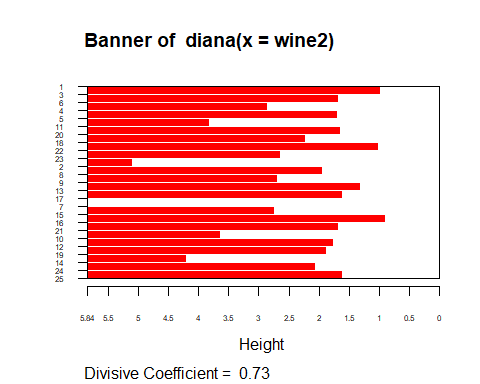


### Task 10

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



plot(hc2, cex.axis= 0.5)



rect.hclust(hc2, k = 5, border = 2:6)

