# Module 6 - Clustering

## Cashwell, Amanda

#### Task 1

Potential natural clusters could be short distance vs long distance.

#### Task 3

There are 2 clusters that we can appear to share similar characteristics.

#### Task 4

The wss model starts the bend at 4 and the silhouette model determined 4 clusters; so there is a consensus between the two models.

#### Task 6

There are blue, purple, pink and lime clusters that are stacked paralled. They do appear to overlap some.

#### Task 7

The wss looks to be between 2 and 3 clusters and the silhouette shows 2 clusters. They are close but I would say there is a consensus between the 2.

library(tidyverse)

## ── Attaching packages ───────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.2.1 ✓ purrr 0.3.3  
## ✓ tibble 2.1.3 ✓ dplyr 0.8.4  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

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

library(cluster)  
#library(factoextend)- says it is not available in version 3.6  
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

library(factoextra)

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

#Reading in data

trucks = read\_csv("trucks.csv")

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

str(trucks)

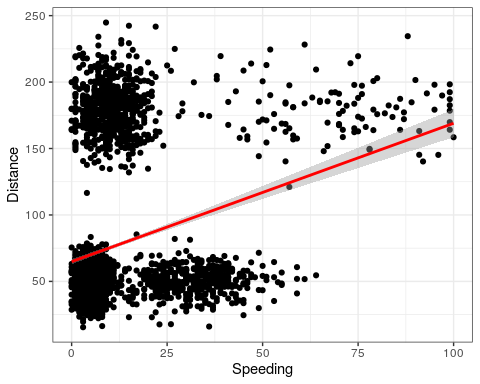
## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 4000 obs. of 3 variables:  
## $ Driver\_ID: num 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num 28 25 27 22 25 10 20 8 34 19 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. 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(x=Speeding,y=Distance)) + geom\_point() + geom\_smooth(method = "lm", color = "red") + theme\_bw()



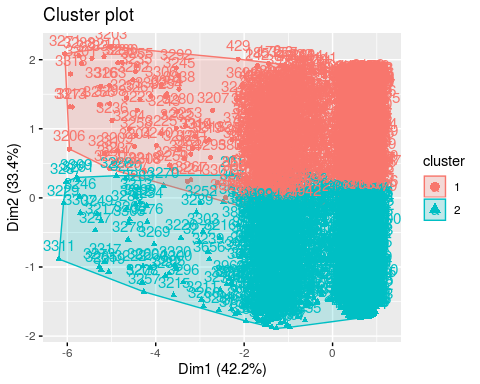
Task 2 create scaled version

trucks2 = as.data.frame(scale(trucks))   
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 3 Use k-means clustering

set.seed(1234)  
clusters1 <- kmeans(trucks, 2)  
  
fviz\_cluster(clusters1, trucks2)



Task 4 determine optimal number of clusters

I had to comment this section out for the R Markdown to work. The chunk runs fine but R Markdown keep crashing when I included this chunk.

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

Task 5 using optimal cluster number

set.seed(1234)  
clusters1 <- kmeans(trucks, 4)  
  
fviz\_cluster(clusters1, trucks2)



read in wineprice data

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 HarvestRain   
## Min. :1952 Min. :6.205 Min. :376.0 Min. :14.98 Min. : 38.0   
## 1st Qu.:1960 1st Qu.:6.519 1st Qu.:536.0 1st Qu.:16.20 1st Qu.: 89.0   
## Median :1966 Median :7.121 Median :600.0 Median :16.53 Median :130.0   
## Mean :1966 Mean :7.067 Mean :605.3 Mean :16.51 Mean :148.6   
## 3rd Qu.:1972 3rd Qu.:7.495 3rd Qu.:697.0 3rd Qu.:17.07 3rd Qu.:187.0   
## Max. :1978 Max. :8.494 Max. :830.0 Max. :17.65 Max. :292.0   
## Age FrancePop   
## Min. : 5.0 Min. :43184   
## 1st Qu.:11.0 1st Qu.:46584   
## Median :17.0 Median :50255   
## Mean :17.2 Mean :49694   
## 3rd Qu.:23.0 3rd Qu.:52894   
## Max. :31.0 Max. :54602

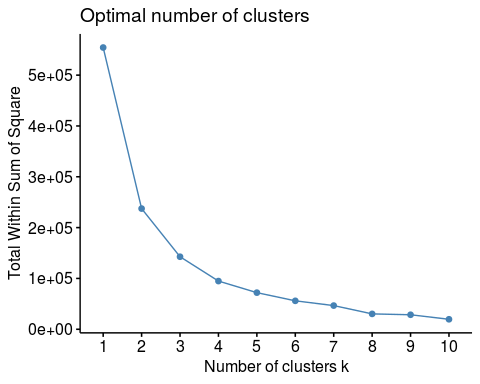
remove columns and scales

wine2 = wine %>% dplyr::select(-c("Year","FrancePop"))%>%  
 as.data.frame(scale(wine))   
summary(wine2)

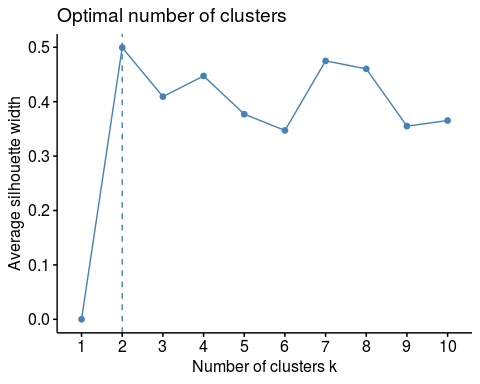
## Price WinterRain AGST HarvestRain Age   
## Min. :6.205 Min. :376.0 Min. :14.98 Min. : 38.0 Min. : 5.0   
## 1st Qu.:6.519 1st Qu.:536.0 1st Qu.:16.20 1st Qu.: 89.0 1st Qu.:11.0   
## Median :7.121 Median :600.0 Median :16.53 Median :130.0 Median :17.0   
## Mean :7.067 Mean :605.3 Mean :16.51 Mean :148.6 Mean :17.2   
## 3rd Qu.:7.495 3rd Qu.:697.0 3rd Qu.:17.07 3rd Qu.:187.0 3rd Qu.:23.0   
## Max. :8.494 Max. :830.0 Max. :17.65 Max. :292.0 Max. :31.0

Task 7 optimal number of clusters

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

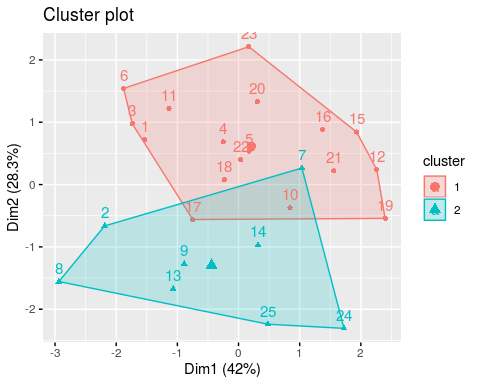


set.seed(123)  
fviz\_nbclust(wine2, kmeans, method = "silhouette")



Task 8 k mean on scaled wine with optimal clusters

set.seed(1234)  
clusters2 <- kmeans(wine2, 2)  
  
fviz\_cluster(clusters2, wine2)



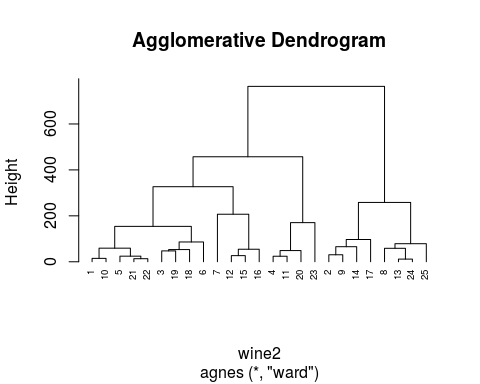
Task 9 agglomerative coefficient

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   
## 0.8125463 0.7384114 0.8973199 0.9330872

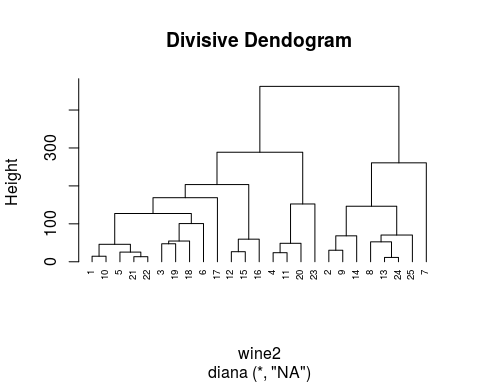
Task 9 agglomerative clustering

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



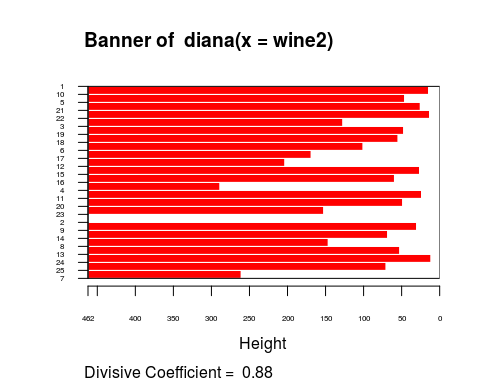
Task 9 agglomerative clustering

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



Task 10 dendogram

plot(hc2, cex.axis= 0.5)



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

