Clustering

Zohaib Sheikh

21/11/2021

Loading relevant libraries

```
library(tidyverse)
library(lubridate)
library(pROC)
library(readxl)
library("writexl")
library(data.table)
library(tidytext)
library(SnowballC)
library(textstem)
library("textdata")
library(factoextra)
library(dbscan)
library(cluster)
```

Reading Data

```
data<-read_excel("D:/Fall'21 - UIC/IDS 572 - Data Mining/Assignments/Clustering/market_data_cluster.xls
df<-data
head(df)
```

```
## # A tibble: 6 x 25
##   `Member id`   SEC   FEH   SEX   AGE    HS CHILD `Affluence Index`
##         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>              <dbl>
## 1     1010010     4     3     1     4     2     4                  2
## 2     1010020     3     2     2     2     4     2                 19
## 3     1014020     2     3     2     4     6     4                 23
## 4     1014030     4     0    NA     4     0     5                  0
## 5     1014190     4     1     2     3     4     3                 10
## 6     1017020     4     3     2     3     5     2                 13
## # ... with 17 more variables: No. of Brands <dbl>, Brand Runs <dbl>,
## #   No. of  Trans <dbl>, Value <dbl>, Trans / Brand Runs <dbl>, Vol/Tran <lgl>,
## #   Pur Vol Promo 6 % <dbl>, Pur Vol Other Promo % <dbl>,
## #   Br. Cd. 57, 144 <dbl>, Br. Cd. 55 <dbl>, Br. Cd. 481 <dbl>,
## #   Br. Cd. 352 <dbl>, Br. Cd. 5 <dbl>, Others 999 <dbl>, Pr Cat 1 <dbl>,
## #   Pr Cat 2 <dbl>, Pr Cat 4 <dbl>
```

Cleaning data

```
x<-colnames(df)
x<-gsub(' ','',x)
x
```

```
##  [1] "Memberid"          "SEC"               "FEH"
##  [4] "SEX"               "AGE"               "HS"
##  [7] "CHILD"             "AffluenceIndex"    "No.ofBrands"
## [10] "BrandRuns"         "No.ofTrans"        "Value"
## [13] "Trans/BrandRuns"   "Vol/Tran"          "PurVolPromo6%"
## [16] "PurVolOtherPromo%" "Br.Cd.57,144"      "Br.Cd.55"
## [19] "Br.Cd.481"         "Br.Cd.352"         "Br.Cd.5"
## [22] "Others999"         "PrCat1"            "PrCat2"
## [25] "PrCat4"
```

```
x<-gsub(',','.',x)
x
```

```
##  [1] "Memberid"          "SEC"               "FEH"
##  [4] "SEX"               "AGE"               "HS"
##  [7] "CHILD"             "AffluenceIndex"    "No.ofBrands"
## [10] "BrandRuns"         "No.ofTrans"        "Value"
## [13] "Trans/BrandRuns"   "Vol/Tran"          "PurVolPromo6%"
## [16] "PurVolOtherPromo%" "Br.Cd.57.144"      "Br.Cd.55"
## [19] "Br.Cd.481"         "Br.Cd.352"         "Br.Cd.5"
## [22] "Others999"         "PrCat1"            "PrCat2"
## [25] "PrCat4"
```

```
x<-gsub('/','.',x)
x
```

```
##  [1] "Memberid"          "SEC"               "FEH"
##  [4] "SEX"               "AGE"               "HS"
##  [7] "CHILD"             "AffluenceIndex"    "No.ofBrands"
## [10] "BrandRuns"         "No.ofTrans"        "Value"
## [13] "Trans.BrandRuns"   "Vol.Tran"          "PurVolPromo6%"
## [16] "PurVolOtherPromo%" "Br.Cd.57.144"      "Br.Cd.55"
## [19] "Br.Cd.481"         "Br.Cd.352"         "Br.Cd.5"
## [22] "Others999"         "PrCat1"            "PrCat2"
## [25] "PrCat4"
```

```
colnames(df)<-x
head(df)
```

```
## # A tibble: 6 x 25
##    Memberid   SEC   FEH   SEX   AGE    HS CHILD AffluenceIndex No.ofBrands
##       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>          <dbl>       <dbl>
## 1  1010010     4     3     1     4     2     4              2           3
## 2  1010020     3     2     2     2     4     2             19           5
## 3  1014020     2     3     2     4     6     4             23           5
## 4  1014030     4     0    NA     4     0     5              0           2
## 5  1014190     4     1     2     3     4     3             10           3
```

```
## 6  1017020     4     3     2     3     5     2            13           3
## # ... with 16 more variables: BrandRuns <dbl>, No.ofTrans <dbl>, Value <dbl>,
## #   Trans.BrandRuns <dbl>, Vol.Tran <lgl>, PurVolPromo6% <dbl>,
## #   PurVolOtherPromo% <dbl>, Br.Cd.57.144 <dbl>, Br.Cd.55 <dbl>,
## #   Br.Cd.481 <dbl>, Br.Cd.352 <dbl>, Br.Cd.5 <dbl>, Others999 <dbl>,
## #   PrCat1 <dbl>, PrCat2 <dbl>, PrCat4 <dbl>
```

```
df<-df%>%subset(select=-c(Vol.Tran))
colSums(is.na(df))
```

```
##          Memberid               SEC               FEH               SEX
##                 0                 0                 0                30
##               AGE                HS             CHILD    AffluenceIndex
##                 0                 0                 0                 0
##       No.ofBrands         BrandRuns        No.ofTrans             Value
##                 0                 0                 0                 0
##   Trans.BrandRuns     PurVolPromo6% PurVolOtherPromo%      Br.Cd.57.144
##                 0                 0                 0                 0
##          Br.Cd.55         Br.Cd.481         Br.Cd.352          Br.Cd.5
##                 0                 0                 0                 0
##         Others999            PrCat1            PrCat2            PrCat4
##                 0                 0                 0                 0
```

```
df[1:8]<-sapply(df[1:8],as.factor)
df1<-df %>% rowwise() %>% mutate(maxBr=max(Br.Cd.57.144, Br.Cd.55, Br.Cd.481,Br.Cd.352, Br.Cd.5))
```

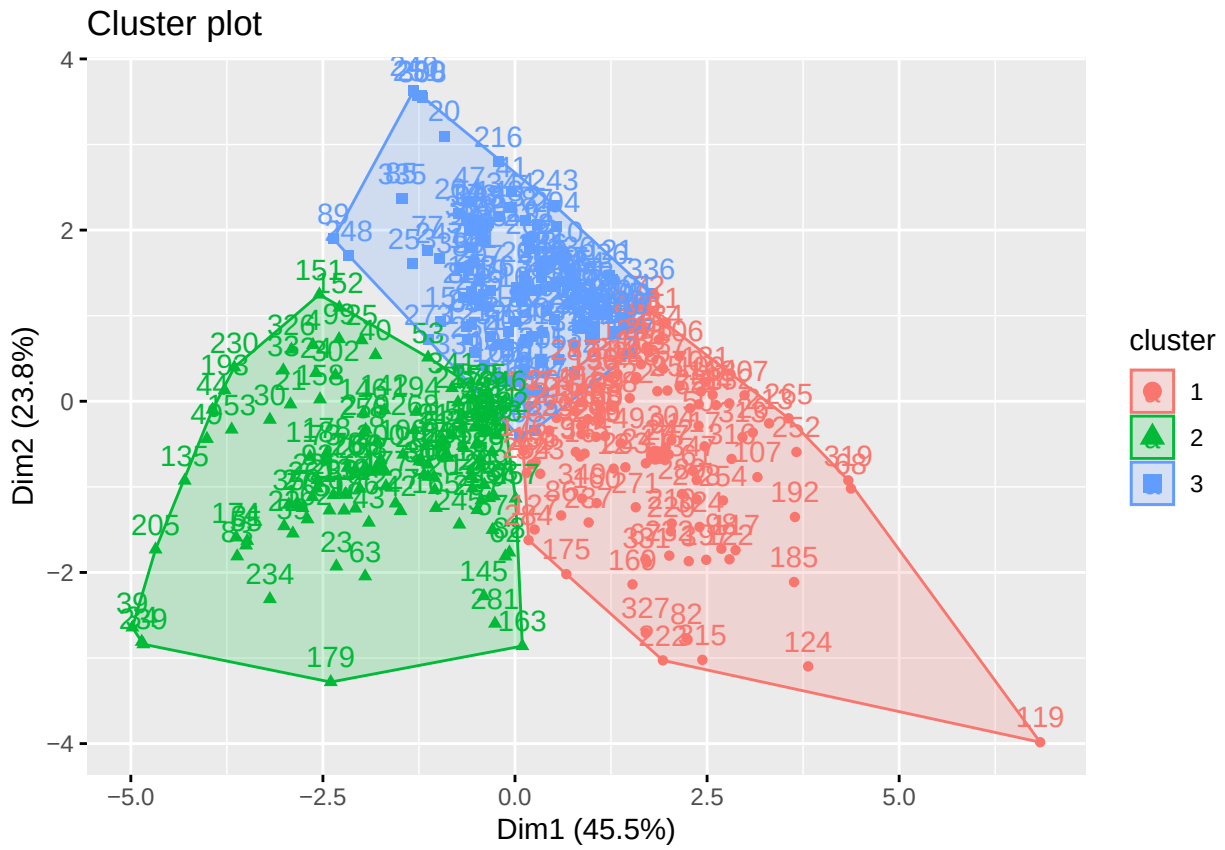
Clustering using K-means

```
PURCHASE_BEHAVIOR <- c('No.ofBrands', 'BrandRuns', 'No.ofTrans', 'Value', 'Trans.BrandRuns','maxBr', 'O
kmClus_pb<- df1 %>% select(PURCHASE_BEHAVIOR) %>% scale() %>% kmeans(centers=3, nstart=25)
kmClus_pb
```

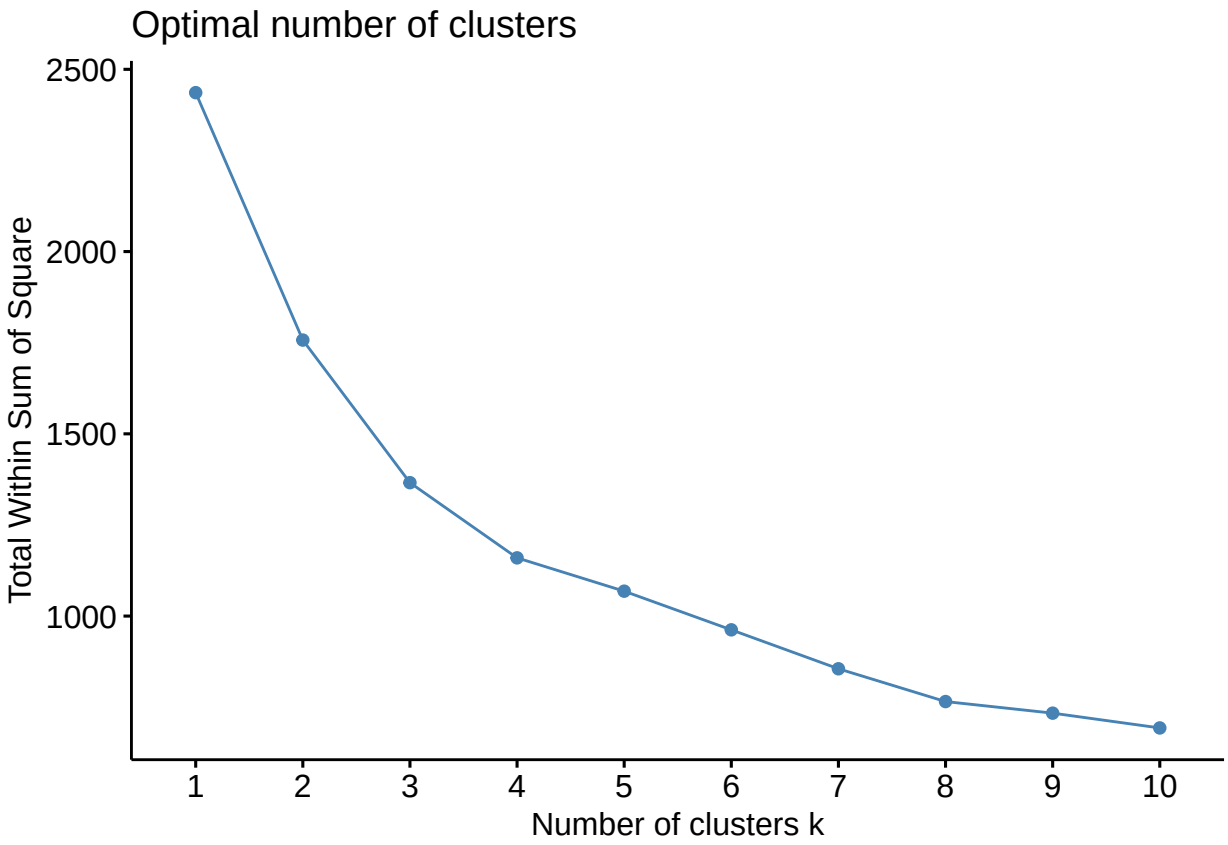
```
## K-means clustering with 3 clusters of sizes 108, 117, 124
##
## Cluster means:
##   No.ofBrands  BrandRuns No.ofTrans      Value Trans.BrandRuns      maxBr
## 1   1.0165418  1.0734809  0.9361059  0.6003165      -0.3345629 -0.4434448
## 2  -0.4453943 -0.7290218 -0.4318510 -0.2547526       0.5565117  1.1716581
## 3  -0.4651240 -0.2470999 -0.4078458 -0.2824849      -0.2337022 -0.7192900
##    Others999
## 1  0.2020041
## 2 -1.0591556
## 3  0.8234256
##
## Clustering vector:
##   [1] 3 1 1 2 3 1 3 2 3 3 1 1 3 2 3 1 1 2 3 3 2 2 2 2 2 1 2 2 3 2 2 3 2 2 2 3 3
##  [38] 3 2 2 3 2 2 2 3 2 3 2 2 1 2 2 2 1 2 2 2 1 2 3 1 2 2 2 1 2 1 1 3 2 2 1 2 3
##  [75] 3 3 3 2 3 3 3 1 2 3 3 1 3 2 3 2 1 1 2 3 1 1 2 1 2 2 3 3 3 1 3 1 1 3 1 3 2
## [112] 3 3 3 3 1 1 2 1 1 3 1 3 1 3 2 1 2 3 3 1 1 2 3 2 3 3 1 1 3 3 2 3 2 2 2 1 3
## [149] 1 2 2 2 2 2 1 1 2 2 3 1 2 2 2 1 2 1 3 3 1 3 1 3 3 2 1 1 2 2 2 3 3 3 1 1 1
## [186] 2 3 1 2 3 3 1 2 2 1 1 1 1 1 2 2 3 3 3 2 3 3 3 2 1 2 3 1 2 1 3 2 1 2 1 3 1
## [223] 2 3 3 2 2 1 3 2 2 1 2 2 2 2 1 2 2 3 1 1 3 2 2 2 3 3 3 3 3 1 3 1 3 2 2 2 1
```

```
## [260] 1 2 3 1 1 1 2 3 2 1 1 1 3 3 3 3 3 1 1 2 3 2 3 1 1 3 3 1 2 1 3 3 3 2 3 3 3
## [297] 2 1 1 3 3 2 3 1 3 2 1 3 3 3 3 1 1 1 1 1 3 3 1 2 1 1 1 1 2 2 1 2 1 3 1 2 3
## [334] 3 3 3 1 1 3 3 2 2 3 3 1 3 1 3 1
##
## Within cluster sum of squares by cluster:
## [1] 479.2368 569.9031 316.6788
##  (between_SS / total_SS =  43.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

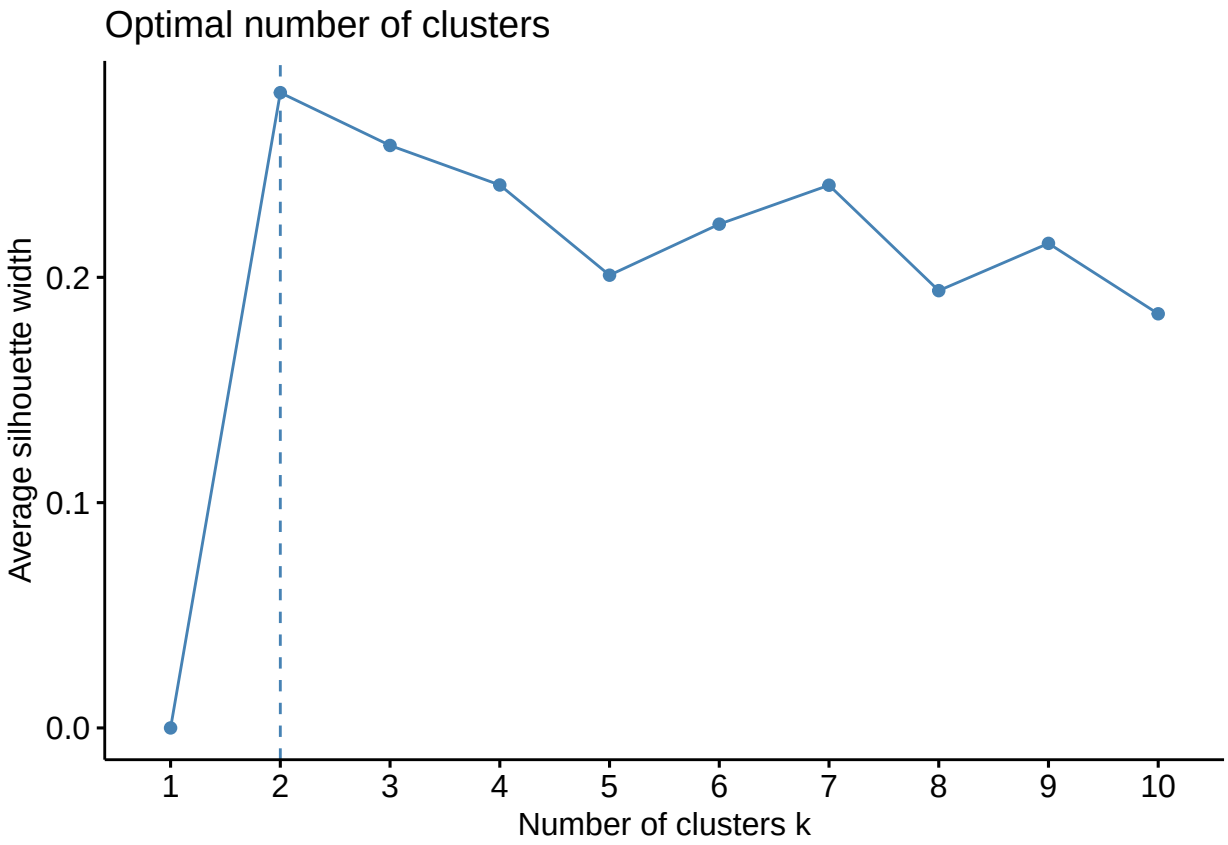
```
fviz_cluster( kmClus_pb, data=df1 %>% select(PURCHASE_BEHAVIOR))
```




```
fviz_nbclust( df1 %>% select(PURCHASE_BEHAVIOR) %>% scale(), kmeans, method = "wss")
```



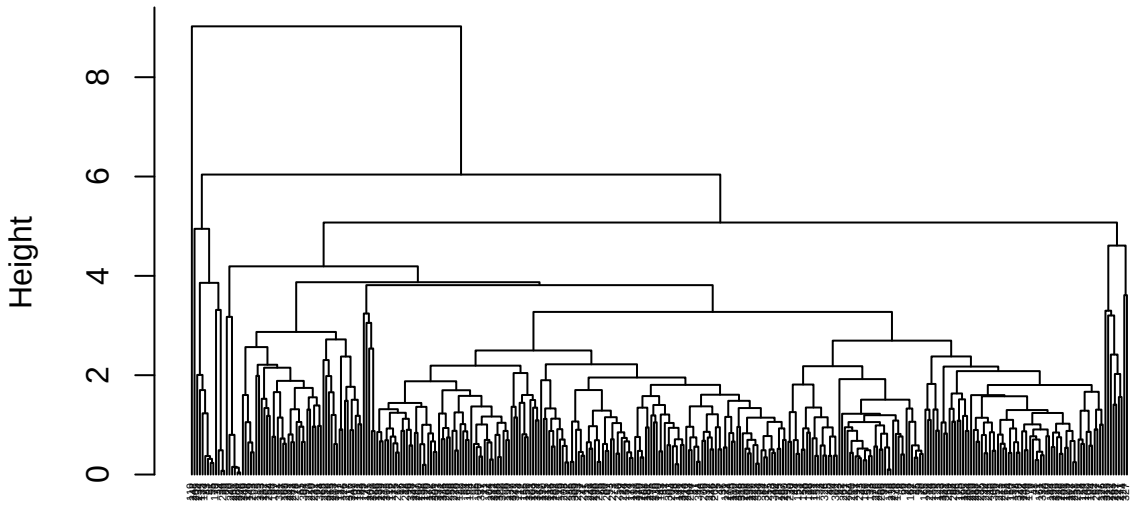

```
fviz_nbclust( df1 %>% select(PURCHASE_BEHAVIOR) %>% scale(), kmeans, method = "silhouette")
```



Clustering - Hierarchial

```
xpb<-df1 %>% select(PURCHASE_BEHAVIOR) %>% scale()
xdist <- dist (xpb, method = "euclidean")
#using hclust
hierC_pb <- hclust (xdist, method = "average" )
hierC_pb_w <- hclust(xdist, method = "ward.D" )
hierC_pb_c <- hclust(xdist, method = "complete" )

plot(hierC_pb, cex=0.3, hang=-3, main='hclust - average')
```




```
#check the agglomerative coeff given by agnes
hierC_pb_ag_c <- agnes(xdist, method = "complete" )
hierC_pb_ag_c$ac
```

```
## [1] 0.934203
```

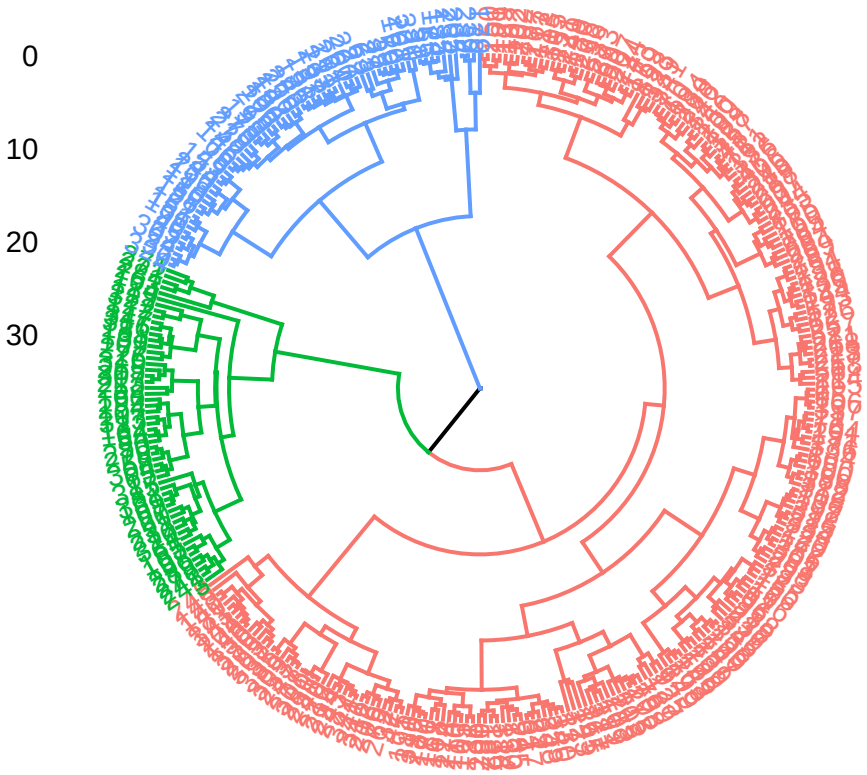
```
hierC_pb_ag_a <- agnes(xdist, method = "average" )
hierC_pb_ag_a$ac
```

```
## [1] 0.9086666
```

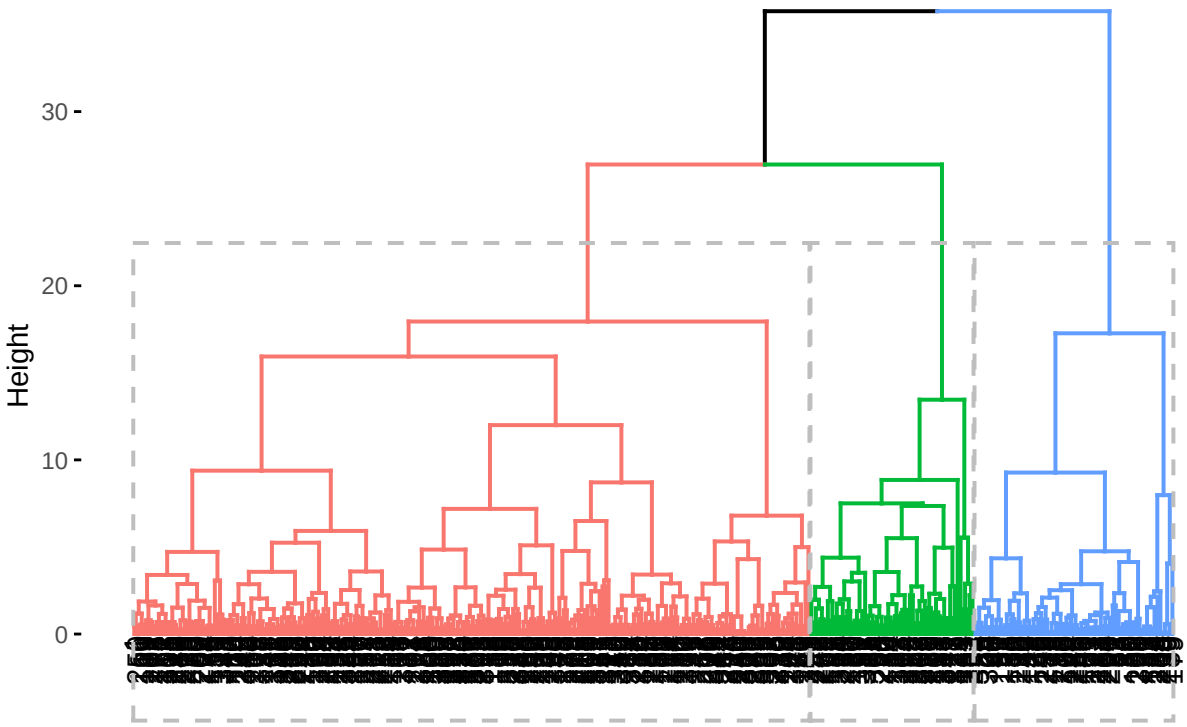
```
hierC_pb_ag_w <- agnes(xdist, method = "ward" )
hierC_pb_ag_w$ac
```

```
## [1] 0.9763677
```

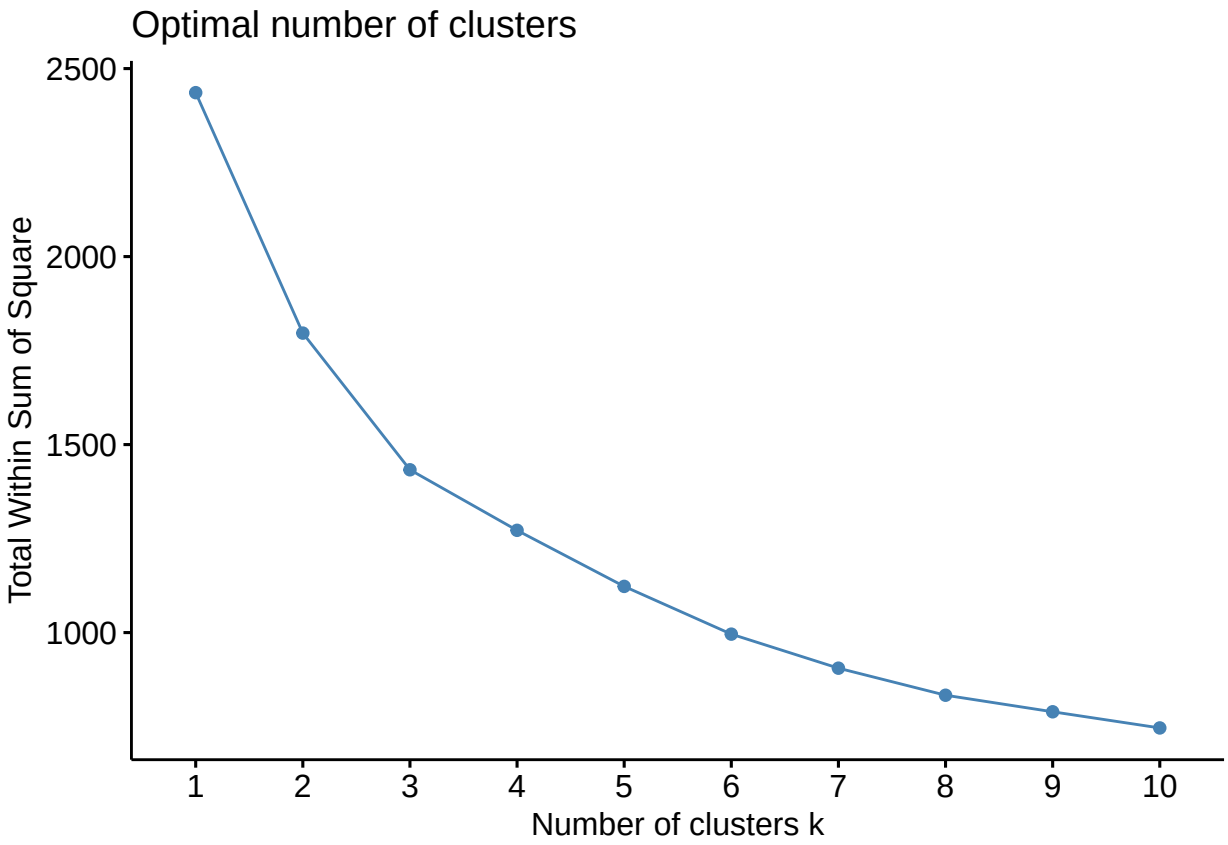
```
fviz_dend(hierC_pb_ag_w, k=3, color_labels_by_k = TRUE, type="circular",
rect=TRUE, main="agnes - Wards")
```



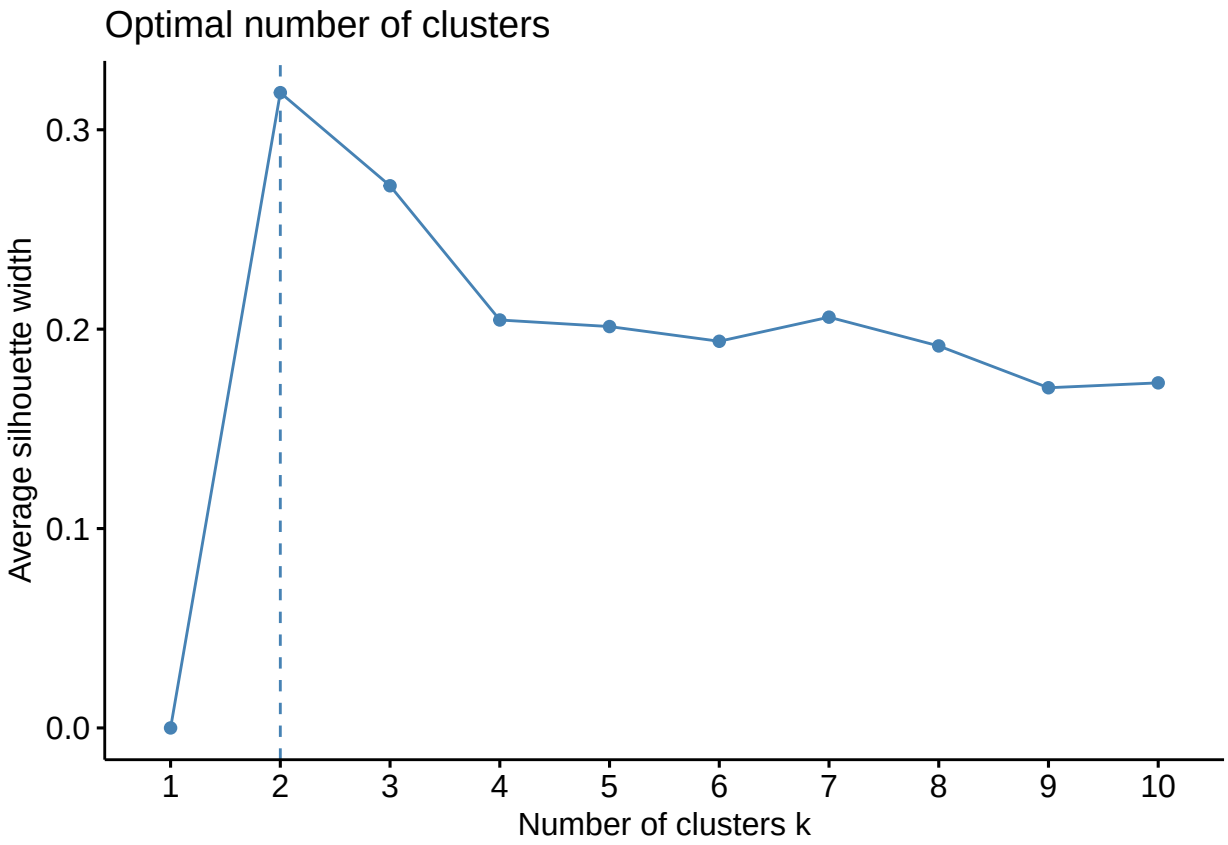

```
fviz_dend( hierC_pb_ag_w, k=3, rect=TRUE, color_labels_by_k = FALSE, main="agnes - Wards")
```




```
fviz_nbclust ( xpb, FUN = hcut, method = "wss")
```




```
fviz_nbclust ( xpb, FUN = hcut, method = "silhouette")
```

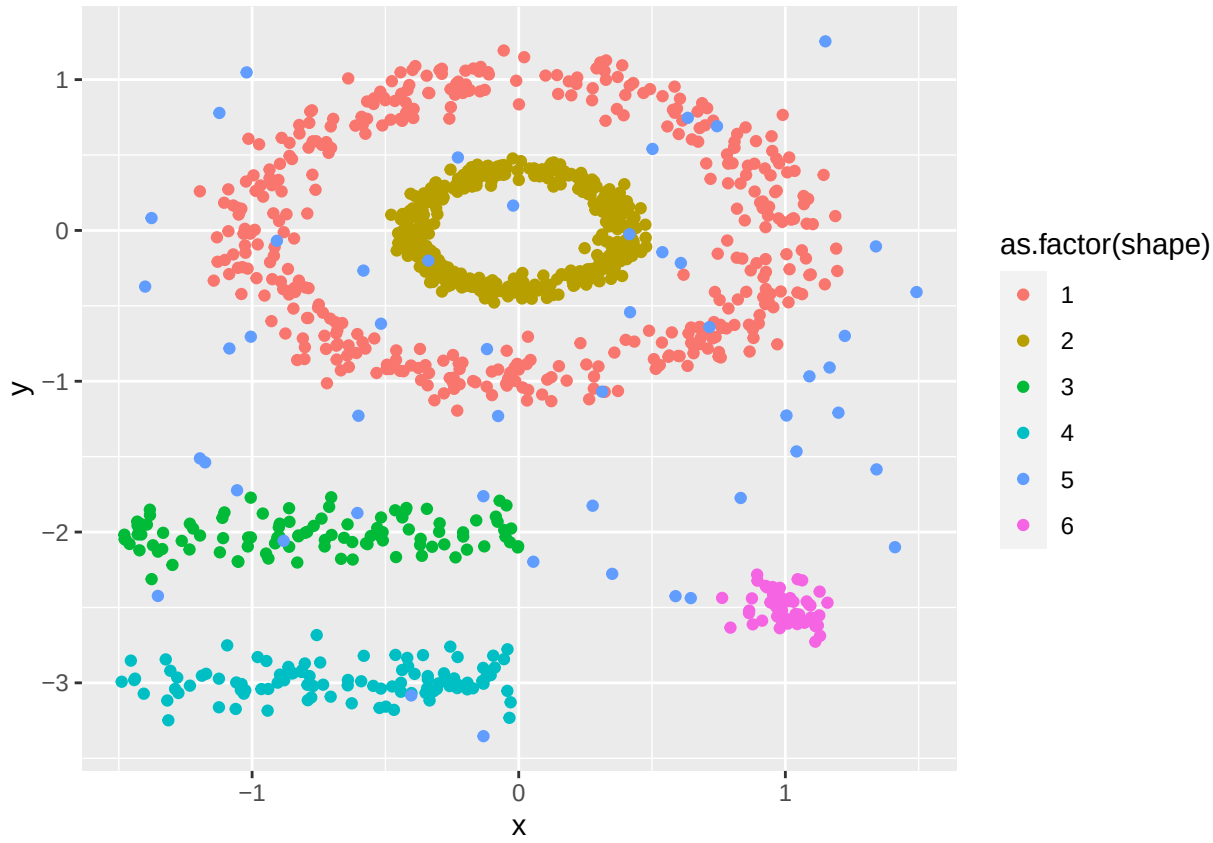


Clustering - Density Based (DBSCAN)

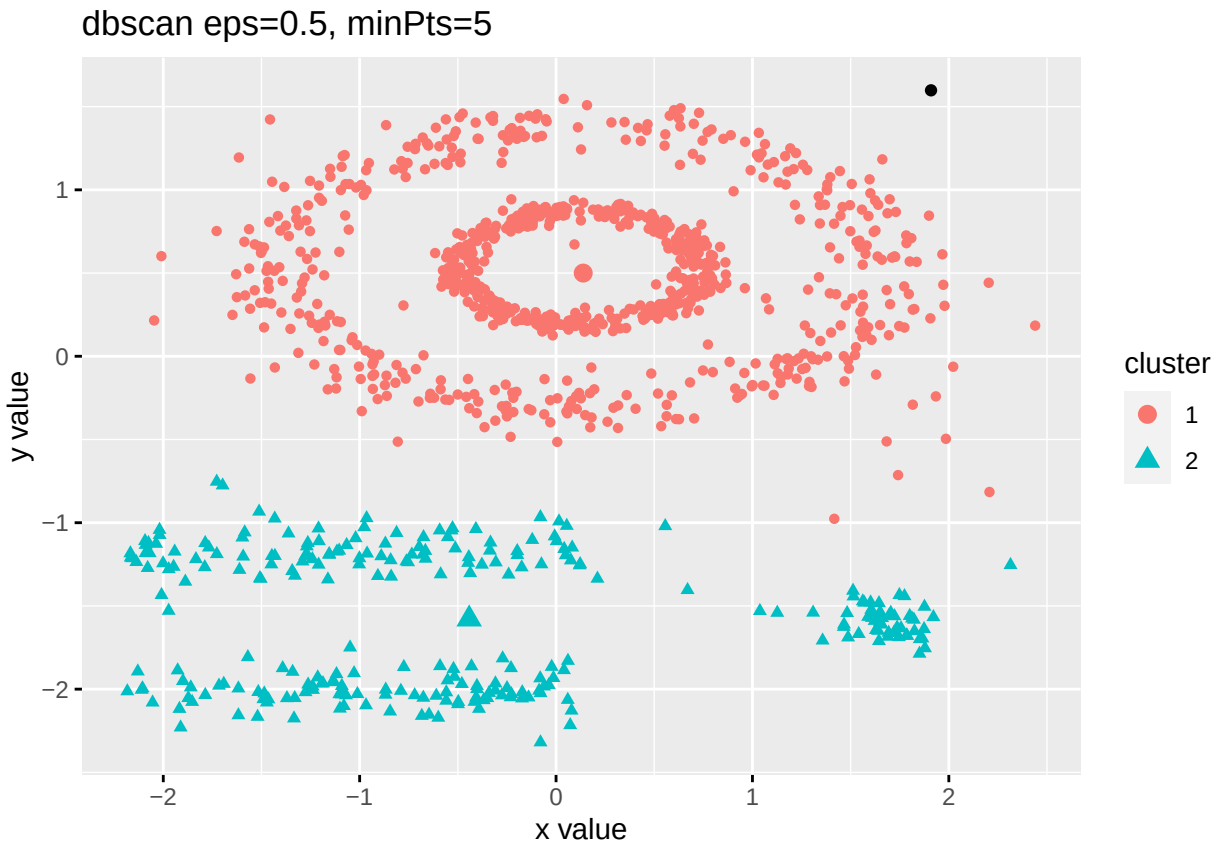
```
data("multishapes")
head(multishapes)
```

```
##            x          y shape
## 1 -0.8037393 -0.8530526     1
## 2  0.8528507  0.3676184     1
## 3  0.9271795 -0.2749024     1
## 4 -0.7526261 -0.5115652     1
## 5  0.7068462  0.8106792     1
## 6  1.0346985  0.3946550     1
```

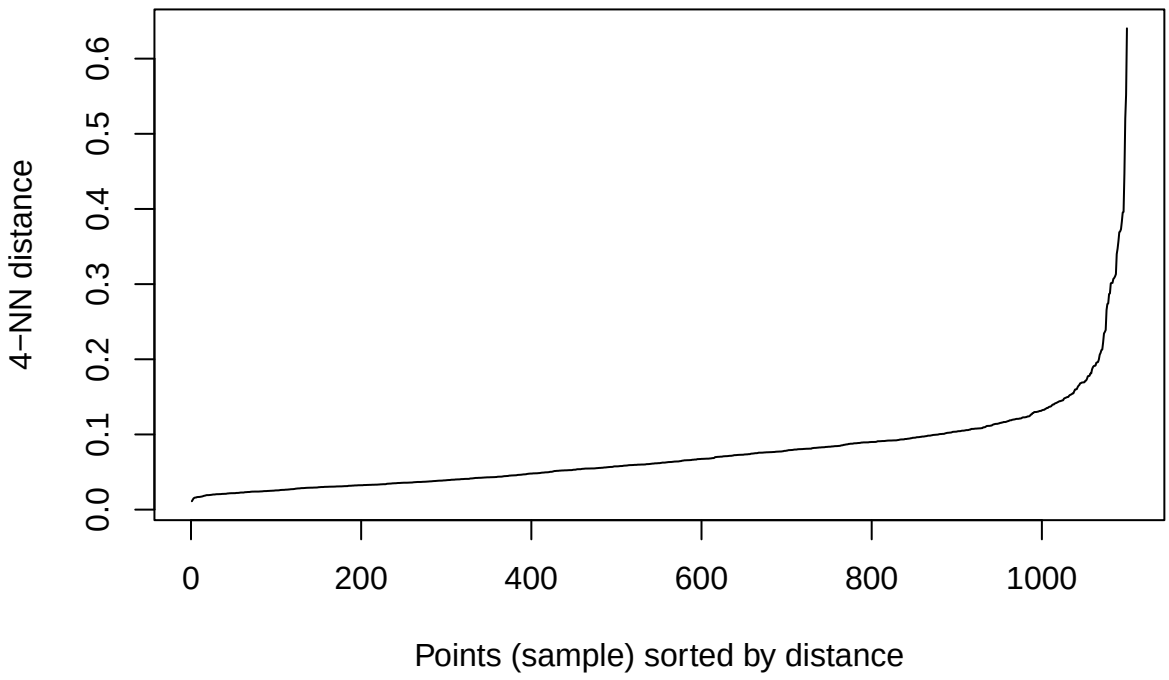
```
multishapes %>% ggplot(aes(x=x,y=y, col=as.factor(shape)))+geom_point()
```



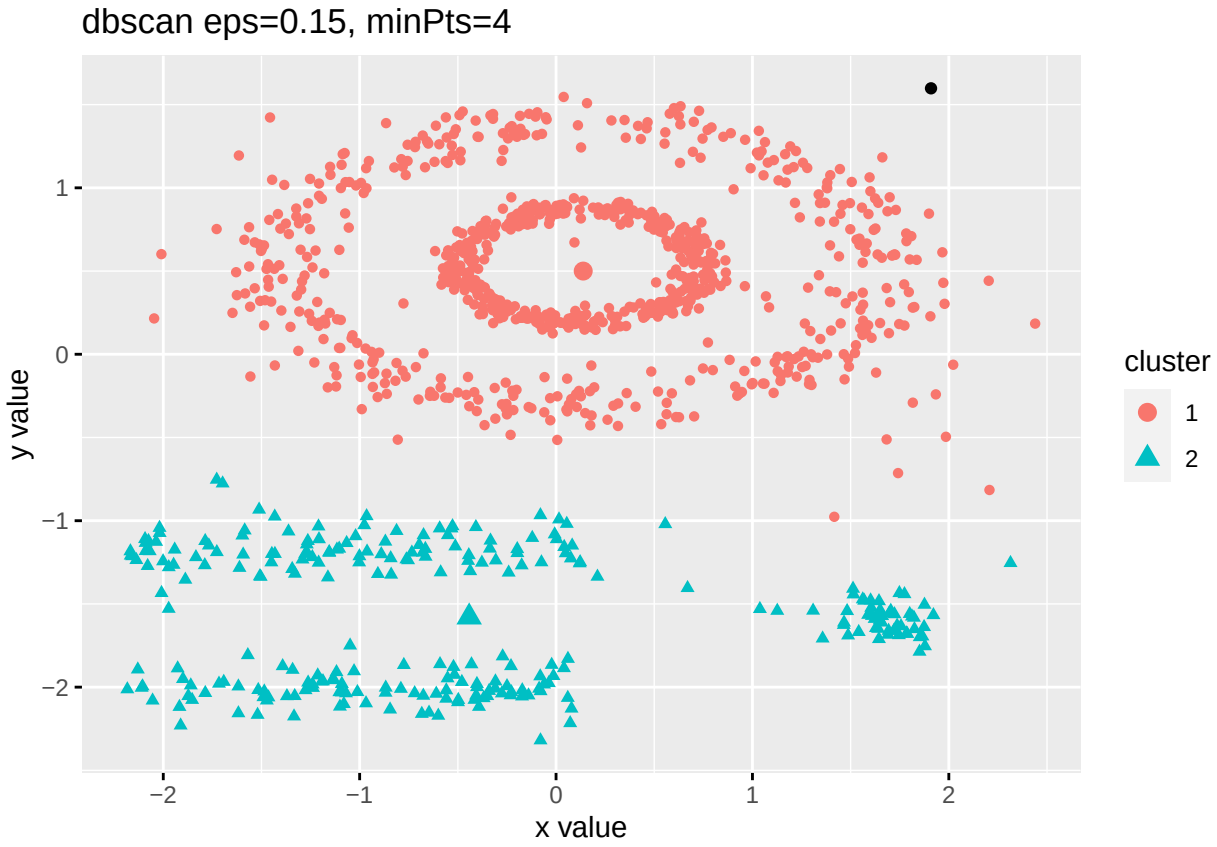

```
msDbscan <- dbscan( multishapes[,1:2], eps = 0.5, minPts = 5)
fviz_cluster(msDbscan, data=multishapes[,1:2], geom="point", ellipse = FALSE, main="dbscan eps=0.5, minl
```




```
kNNdistplot( multishapes[,1:2], k=4)
```




```
fviz_cluster(msDbscan, data=multishapes[,1:2], geom="point", ellipse = FALSE, main="dbscan eps=0.15, mi
```


dbscan eps=0.15, minPts=4
y value
x value
1
0
−1
−2
−2
−1
0
1
2
cluster
1
2