Assignment Week 7 Cluster Analysis

Nitin

October 17, 2020

Read Data

```
library(tidyverse)
library(readxl)
Data <- read_excel("data.xlsx")
```

Eliminated variables we are not intrested in dataset.

```
df <- select(Data, -c(3,5,7,9,11,13,15,17,19,21,23,25))
```

Format column names

```
names(df)<-str_replace_all(names(df), c(" " = "." , "," = "" ))
```

Remove rows will missing data

```
df = na.omit(df)
```

Scale your data

Scale continious verible Month.

```
df$Months <- scale(df$Month)
```

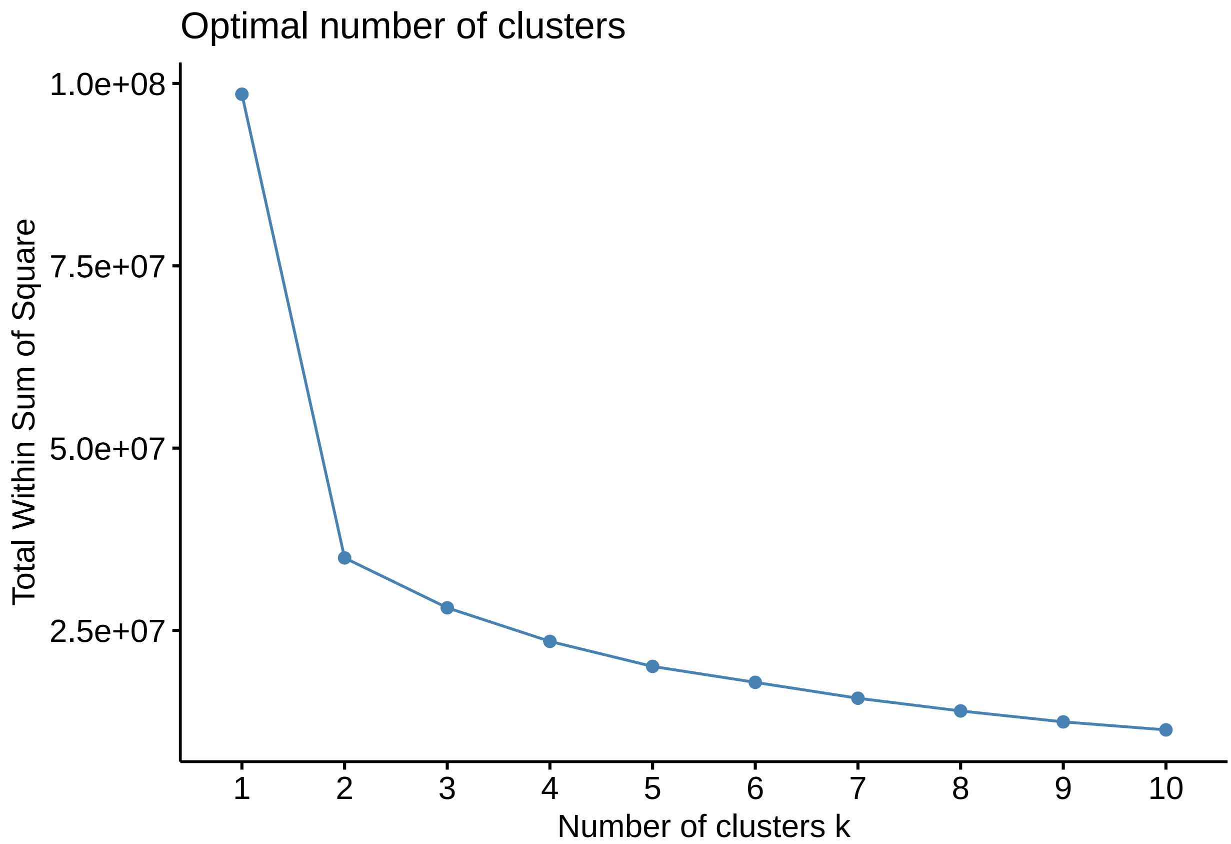
Eliminated unscaled continious verible "Month".

```
data_df <- select(df, -c(1))
```

Find the optimal number of clusters (elbow, gap or silhouette methods).

Elbow Method

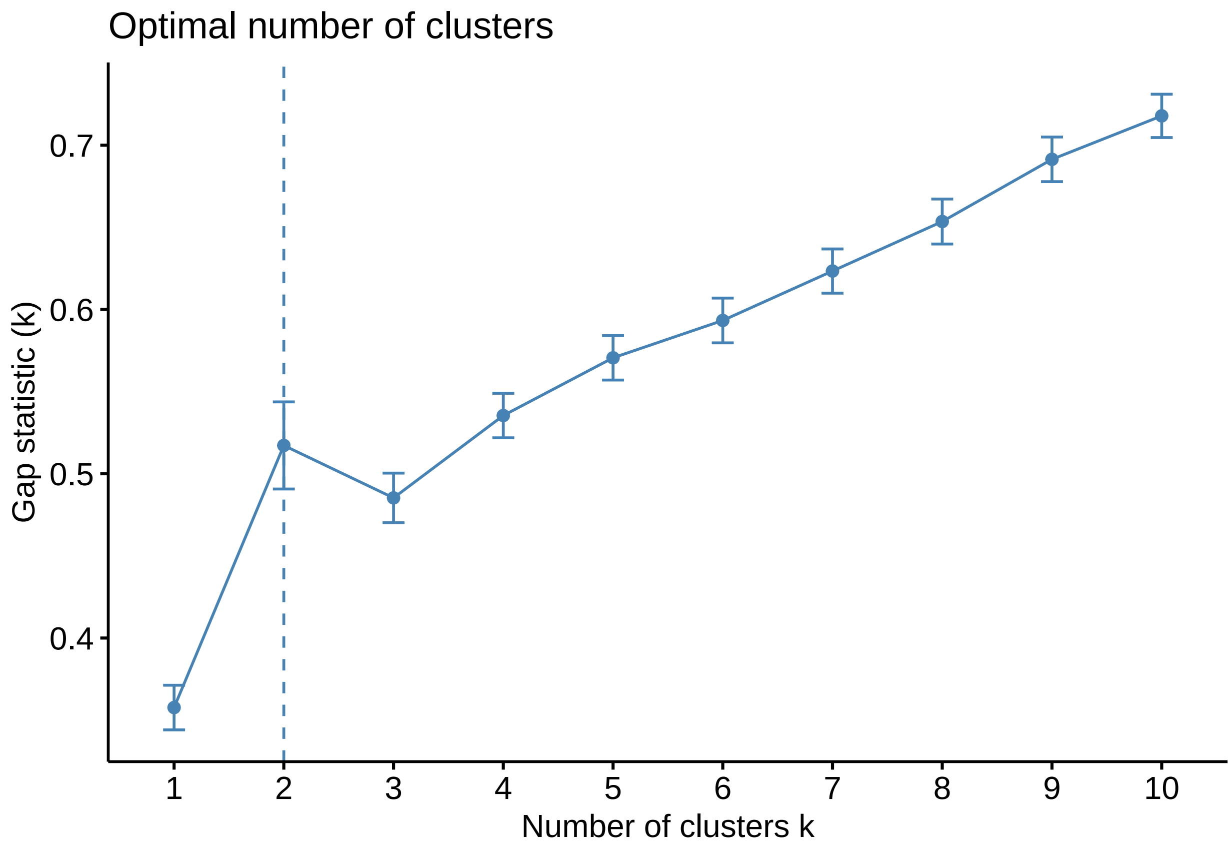
```
library(psych)
library(cluster)    # clustering algorithms
library(factoextra) # clustering visualization
fviz_nbclust(data_df, FUN = hcut, method = "wss")
```



Optimal number of clusters using Elbow methods is 2 clusters.

Gap Method

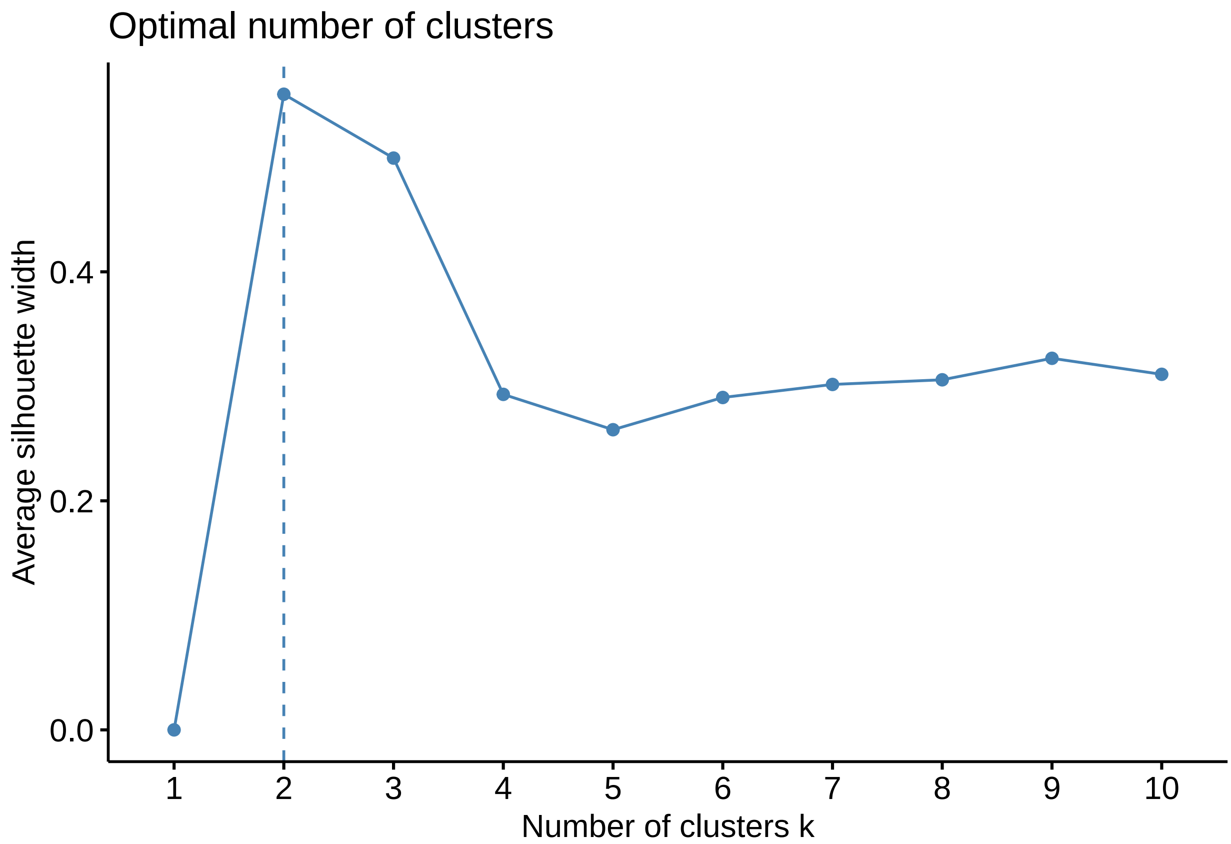
```
gap_stat <- clusGap(data_df, FUN = hcut, nstart = 25, K.max = 10, B = 50)
fviz_gap_stat(gap_stat)
```



Optimal number of clusters using Gap methods is 2 clusters.

silhouette Method

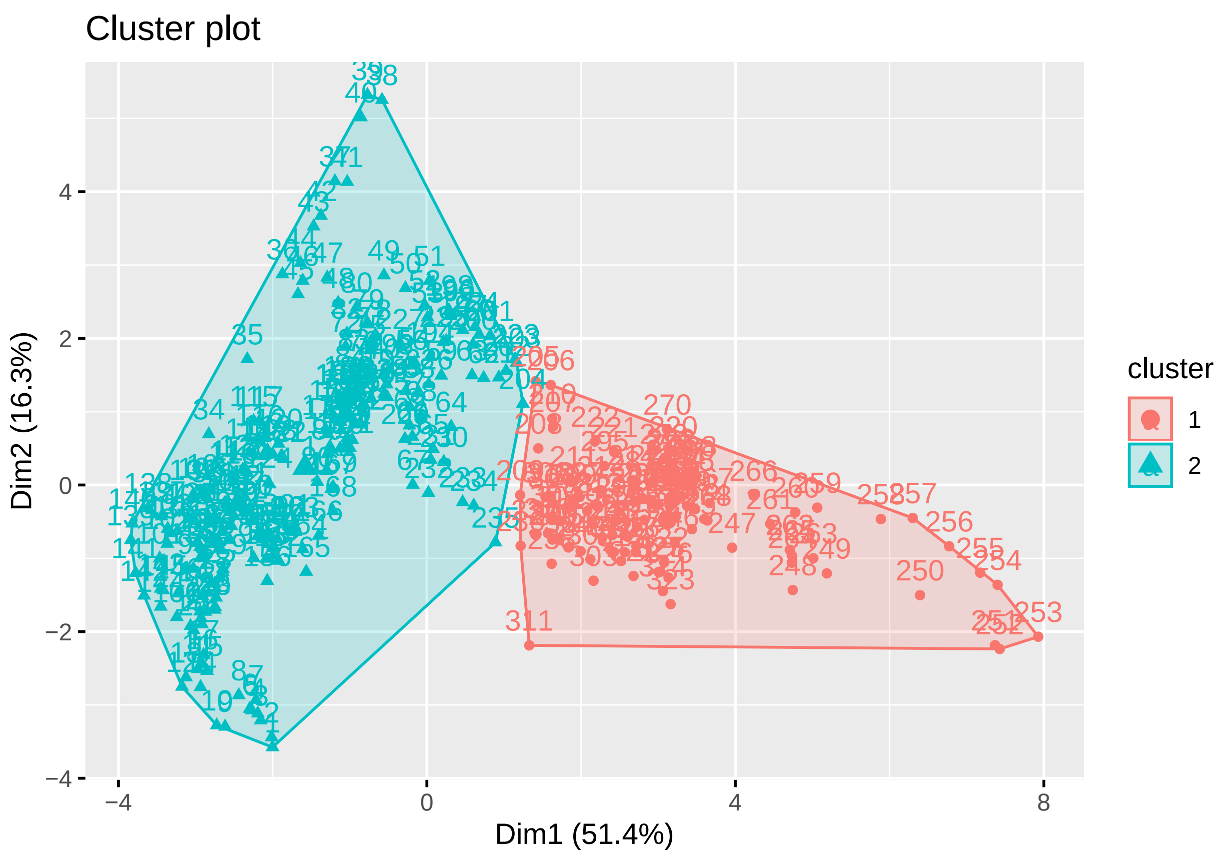
```
fviz_nbclust(data_df, FUN = hcut, method = "silhouette")
```



Optimal number of clusters using Silhouette methods is 2 clusters.

Perform the K-Means cluster analysis and visualize the results

```
k2 <- kmeans(data_df, centers = 2, nstart = 25)
fviz_cluster(k2, data = data_df)
```



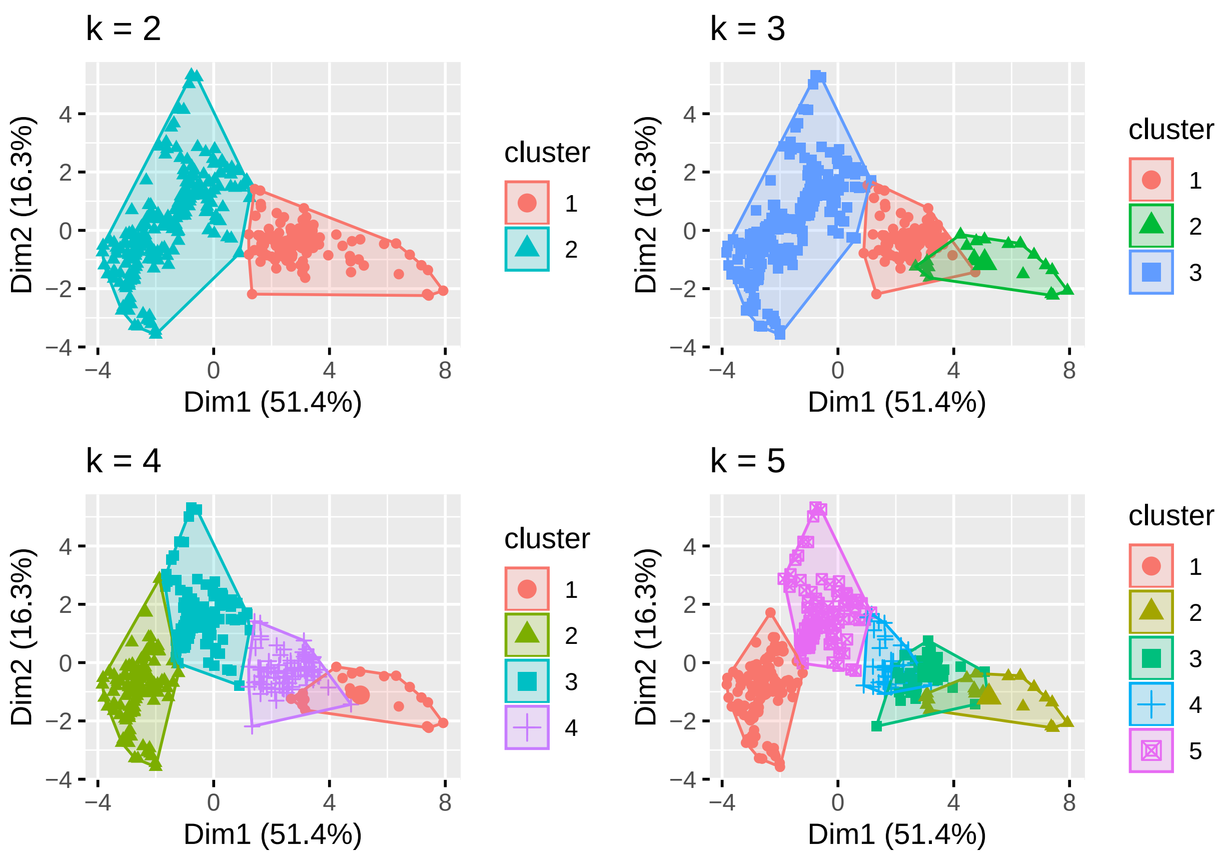
Use several different values of k and examine the differences overlap of cluster and clear separation of cluster.

```r
k3 <- kmeans(data_df, centers = 3, nstart = 25)
k4 <- kmeans(data_df, centers = 4, nstart = 25)
k5 <- kmeans(data_df, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = data_df) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = data_df) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = data_df) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = data_df) + ggtitle("k = 5")

library(gridExtra)

grid.arrange(p1, p2, p3, p4, nrow = 2)
```



It could be observed that with cluster number 2 (K=2)there is no overlap of clusters they are clearly separated.

Perform the hierarchical analysis

Calculate agglomerative coefficient (AC)

Calculate agglomerative coefficient (AC) measures the strength of the clustering structure.

```
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")
ac <- function(x) {
  agnes(data_df, method = x)$ac
}
map_dbl(m, ac)
```

```
##   average    single  complete      ward
## 0.9344819 0.7742130 0.9692225 0.9940206
```

It could be observed that Ward's method identifies the strongest clustering structure of the four methods assessed.

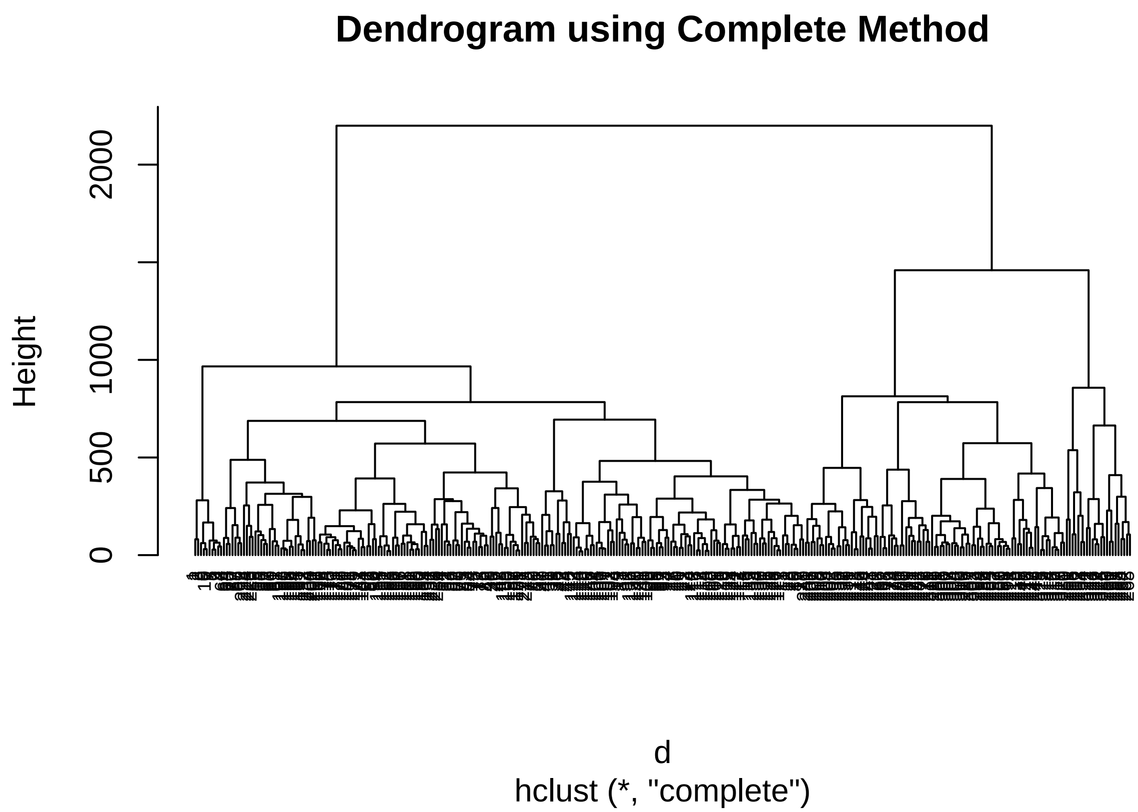
Hierarchical clustering using Complete Linkage

First compute the dissimilarity values with dist() function.

```
d <- dist(data_df, method = "euclidean")
```

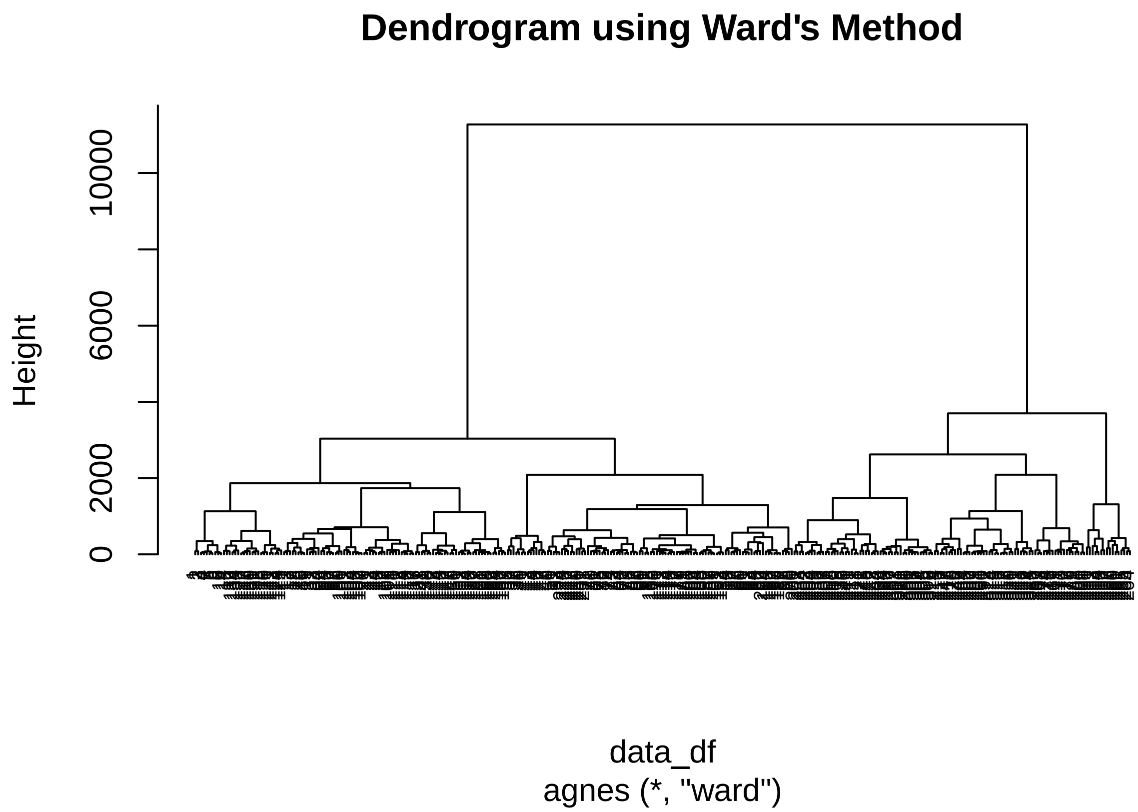
Plot the obtained dendrogram usung Complete method

```
hc1 <- hclust(d, method = "complete" )
plot(hc1, cex = 0.6, hang = -1, main = "Dendrogram using Complete Method")
```



Hierarchical clustering using ward method

```
hc3 <- agnes(data_df, method = "ward")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram using Ward's Method")
```



Describe the results

With the help of various methods like Elbow, Silhouette, Gap it is determined that optimum number of clusters is 2. Performed the K-Means cluster analysis and visualize the results using fviz_cluster function. Performed the hierarchical analysis and calculate agglomerative coefficient (AC). Using agglomerative coefficient (AC) it is determined that Ward's method identifies the strongest clustering structure of the four methods assessed.