Assignment 8: K-Means Clustering

43141 (Sahil Naphade)

18/10/2020

Installing required packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)
#install.packages("FactoMineR")
#install.packages("factoextra")
library(FactoMineR)
library(factoextra)
```

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

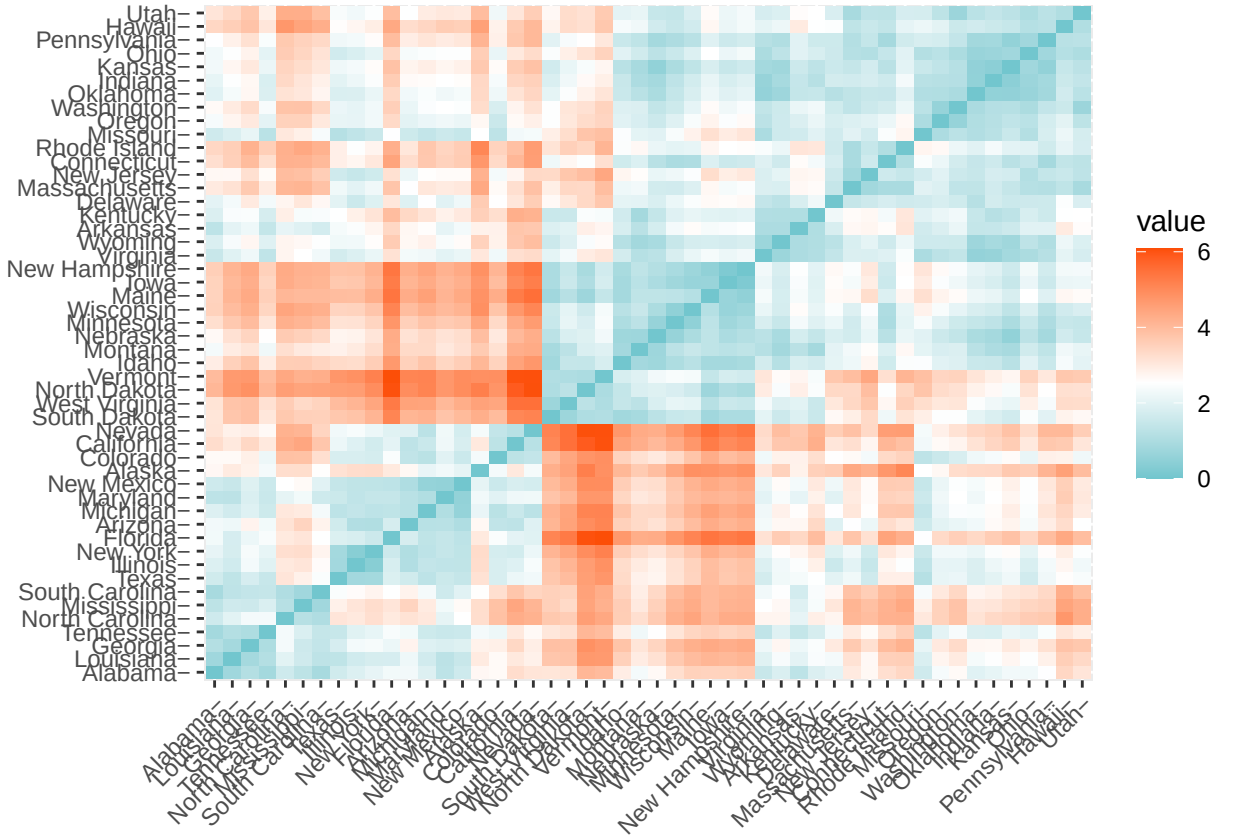
Reading the dataset and removing null values

```
df <- USArrests
df <- na.omit(df)
df <- scale(df)
head(df)
```

```
##               Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

Compute distance matrix between the rows of a data matrix. Visualizr the created distance matrix.

```
distance <- get_dist(df)
fviz_dist(distance, gradient = list(low = "#00AFBB",mid = "white",
                                    high = "#FC4E07"))
```



Perform K-means clustering on the dataset with 3 clusters

```
k3 <- kmeans(df, centers = 3, nstart = 25)
str(k3)
```

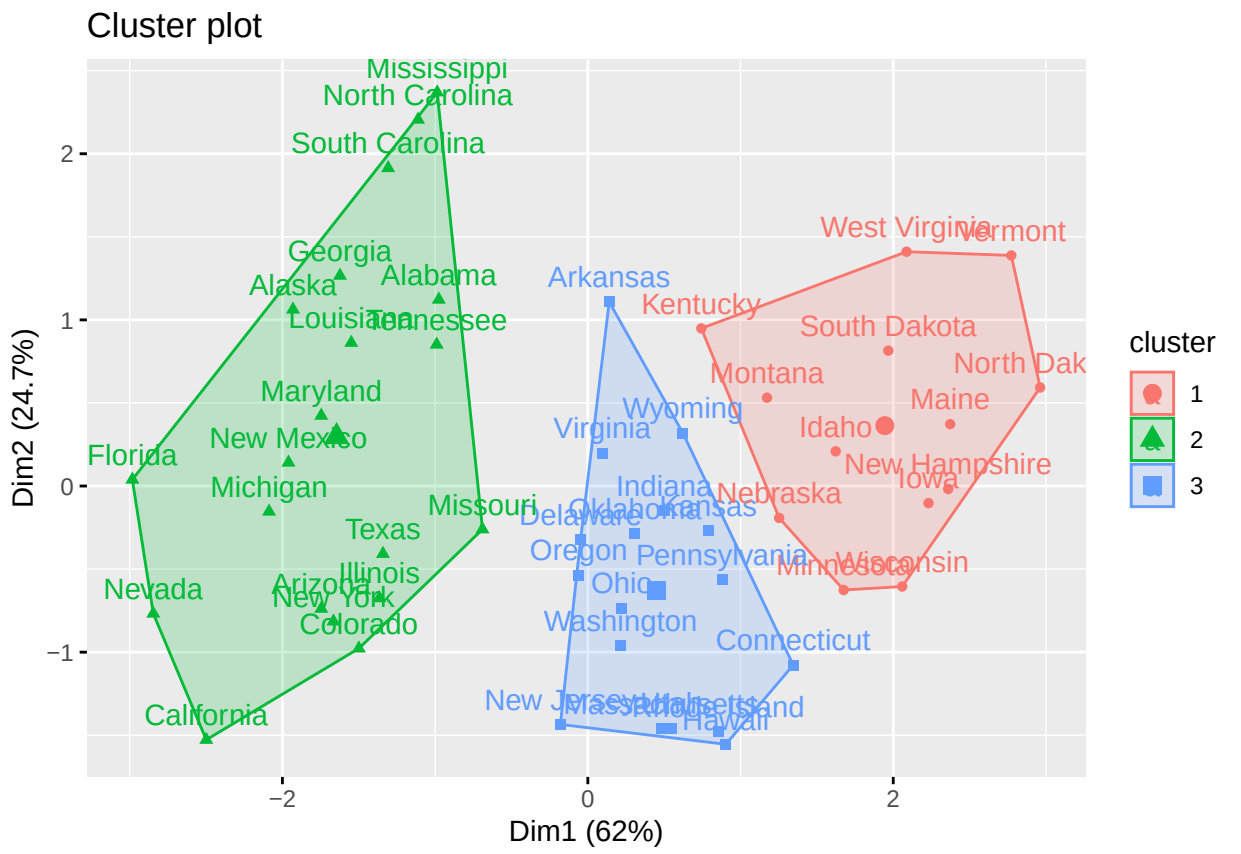
```
## List of 9
##  $ cluster     : Named int [1:50] 2 2 2 3 2 2 3 3 2 2 ...
##   ..- attr(*, "names")= chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ centers     : num [1:3, 1:4] -0.962 1.005 -0.447 -1.107 1.014 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ totss       : num 196
##  $ withinss    : num [1:3] 12 46.7 19.6
##  $ tot.withinss: num 78.3
##  $ betweenss   : num 118
##  $ size        : int [1:3] 13 20 17
##  $ iter        : int 2
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
k3
```

```
## K-means clustering with 3 clusters of sizes 13, 20, 17
##
## Cluster means:
##       Murder    Assault   UrbanPop       Rape
## 1 -0.9615407 -1.1066010 -0.9301069 -0.9667633
## 2  1.0049340  1.0138274  0.1975853  0.8469650
## 3 -0.4469795 -0.3465138  0.4788049 -0.2571398
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              2              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              1              2              3              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              1              2              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              1              2              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              2              1              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              2              2              3              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              1              1              3
##
## Within cluster sum of squares by cluster:
## [1] 11.95246 46.74796 19.62285
##  (between_SS / total_SS =  60.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Visualize the created clusters

```
fviz_cluster(k3, data = df)
```



Intra-cluster variation (Known as total within-cluster variation/total # within-cluster sum of square) # Implementation of Elbow method in R

```
set.seed(123)
```

#Function to compute total within-cluster sum of square

```
wss <- function(k){
  kmeans(df, k, nstart = 10)$tot.withinss
}
```

compute and plot wss for k = 1 to k = 10

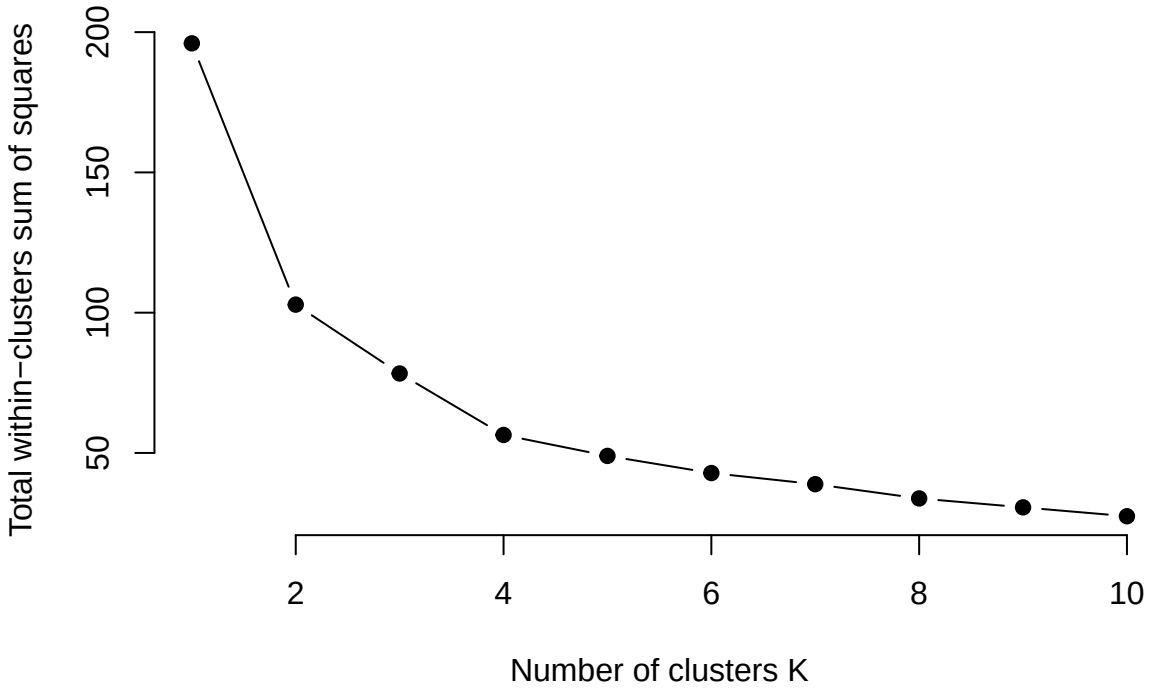
```
k.values <- 1:10
```

Extract wss for 2 - 15 clusters

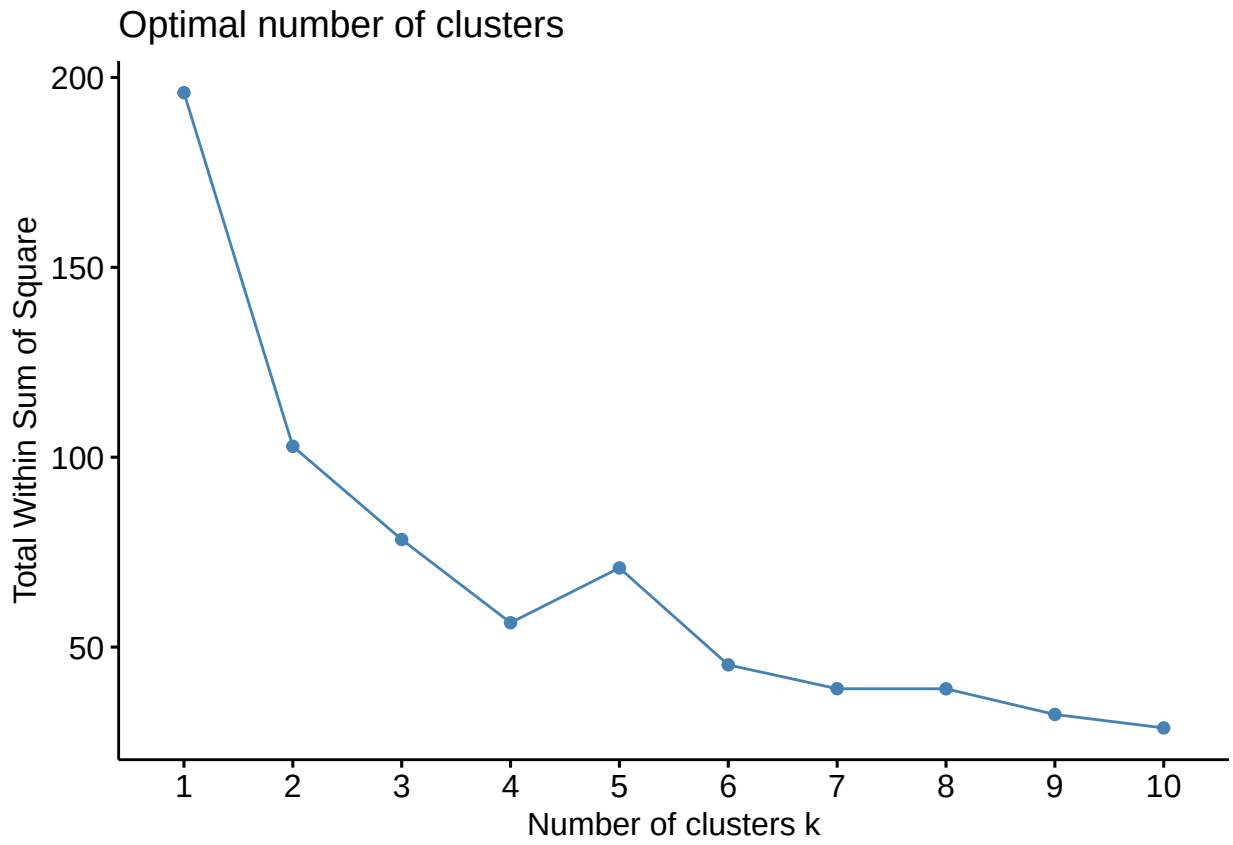
```
wss_values <- map_dbl(k.values, wss)
```

Plot the elbow method (defined in fviz_nbclust)

```
plot(k.values, wss_values, type = "b",
     pch = 19, frame = F, xlab = "Number of clusters K",
     ylab = "Total within-clusters sum of squares")
```




```
set.seed(123)
fviz_nbclust(df, kmeans, method = "wss")
```



Optimal number of clusters = 4, therefore, calc final result using k = 4

```
set.seed(123)
final <- kmeans(df, 4, nstart = 25)
print(final)
```

```
## K-means clustering with 4 clusters of sizes 8, 13, 16, 13
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1  1.4118898  0.8743346 -0.8145211  0.01927104
## 2 -0.9615407 -1.1066010 -0.9301069 -0.96676331
## 3 -0.4894375 -0.3826001  0.5758298 -0.26165379
## 4  0.6950701  1.0394414  0.7226370  1.27693964
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              4              4              1              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              3              3              4              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              2              4              3              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
```

```
##              3              2              1              2              4
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              4              2              1              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              4              2              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              4              4              1              2              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              4              3              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              2              2              3
##
## Within cluster sum of squares by cluster:
## [1]  8.316061 11.952463 16.212213 19.922437
##  (between_SS / total_SS =  71.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(final, data = df)
```