MATH 656 - Assignment3

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'readr' was built under R version 4.0.5

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

==========

Question 1

==========

```
# Load dataset
Data <- USArrests
```

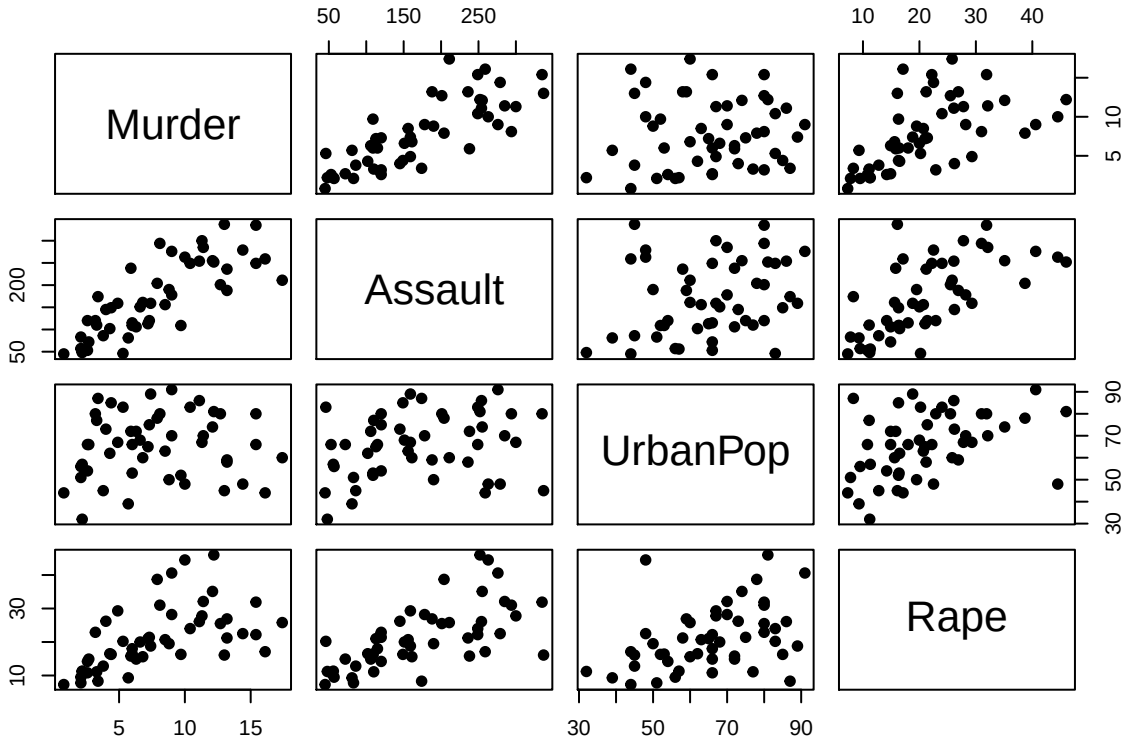
```
## [1] 50  4

## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
##    Murder   Assault  UrbanPop      Rape
##  4.355510 83.337661 14.474763  9.366385
```

```
pairs(Data[,1:4], pch = 19)
```




```
# Murder and Assault have a positive correlation
# Murder and UrbanPop does not have any correlation
# Murder and Rape might have a positive correlation, but we need more tests to confirm it
# Assault and UrbanPop does not have any correlation
# Assault and Rape have a positive correlation
# UrbanPop and Rape have a positive correlation
```

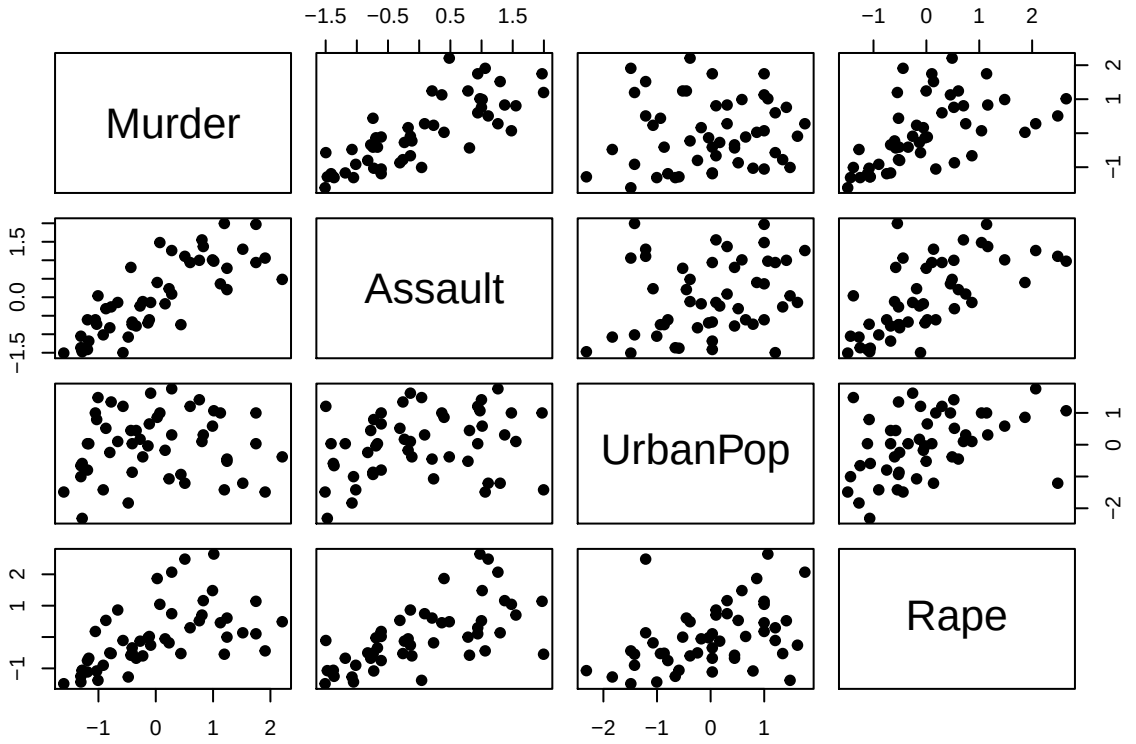
==========

Question 2

==========

```
scale_numeric <- function(x) x %>% mutate_if(is.numeric, function(y) as.vector(scale(y)))

Data_scaled <- Data %>% scale_numeric()
pairs(Data_scaled[,1:4], pch = 19)
```



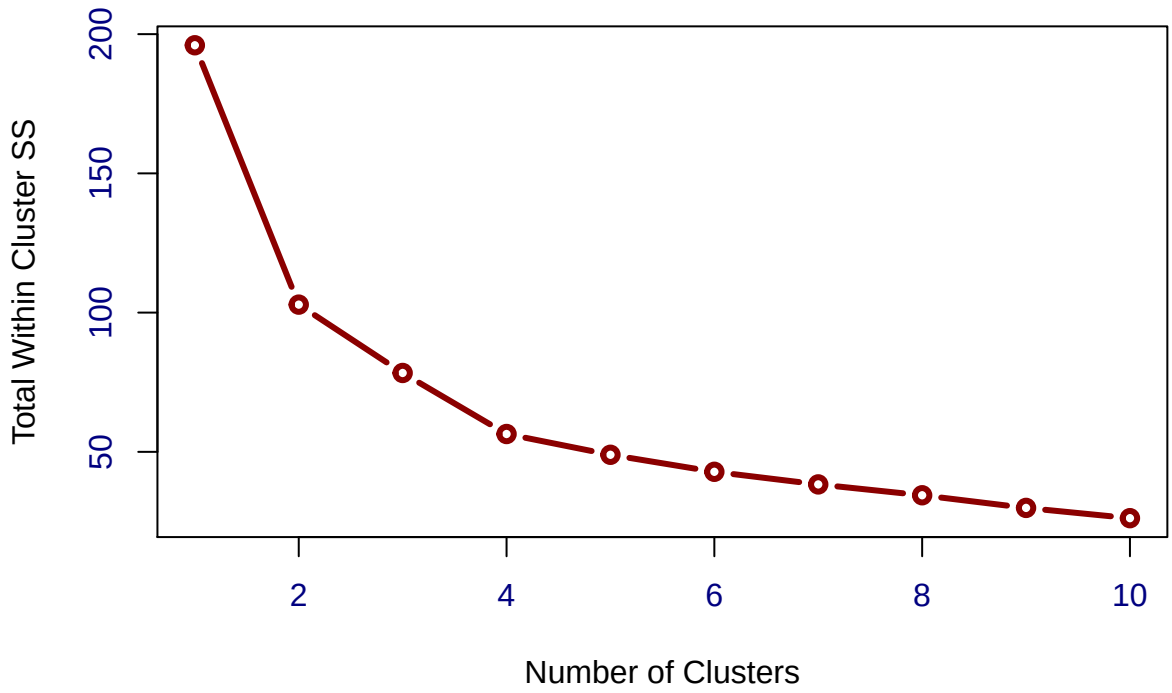

```
# There is no pattern changed in comparison with the results of Q1
```

==========

Question 3

==========

```
# Use the WCSS plot to determine the optimal number of clusters
n = 10
clust = numeric(n)
for (i in 1:n){
  clust[i] = kmeans(Data_scaled, i, nstart = 20)$tot.withinss
}
plot(1:10, clust, col = "darkred", lwd = 3, xlab = "Number of Clusters", type = "b",
     ylab = "Total Within Cluster SS", col.axis = "navyblue")
```




```
# Compute SSE at different numbers of clusters to help determine the best cluster
# 2 clusters
km_2 <- kmeans(Data_scaled, centers = 2, nstart = 10)
km_2
```

```
## K-means clustering with 2 clusters of sizes 30, 20
##
```

```
## Cluster means:
##      Murder    Assault   UrbanPop       Rape
## 1 -0.669956 -0.6758849 -0.1317235 -0.5646433
## 2  1.004934  1.0138274  0.1975853  0.8469650
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              2              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              2              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              2              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              2              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              2              1              2              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              2              1              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              2              2              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              1              1              1
##
## Within cluster sum of squares by cluster:
## [1] 56.11445 46.74796
##  (between_SS / total_SS =  47.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
# 3 clusters
km_3 <- kmeans(Data_scaled, centers = 3, nstart = 10)
km_3
```

```
## 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"
```

```r
# 4 clusters
km_4 <- kmeans(Data_scaled, centers = 4, nstart = 10)
km_4
```

```
## K-means clustering with 4 clusters of sizes 13, 13, 8, 16
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1 -0.9615407 -1.1066010 -0.9301069 -0.96676331
## 2  0.6950701  1.0394414  0.7226370  1.27693964
## 3  1.4118898  0.8743346 -0.8145211  0.01927104
## 4 -0.4894375 -0.3826001  0.5758298 -0.26165379
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              4              4              2              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              4              1              2              4              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              4              1              3              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              4              2              1              3              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              4
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              3              1              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
```

```
##              4              4              4              4              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              3              2              4              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              4              1              1              4
##
## Within cluster sum of squares by cluster:
## [1] 11.952463 19.922437  8.316061 16.212213
##  (between_SS / total_SS =  71.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# 5 clusters
km_5 <- kmeans(Data_scaled, centers = 5, nstart = 10)
km_5
```

```
## K-means clustering with 5 clusters of sizes 11, 7, 12, 10, 10
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1 -0.1642225 -0.3658283 -0.2822467 -0.11697538
## 2  1.5803956  0.9662584 -0.7775109  0.04844071
## 3  0.7298036  1.1188219  0.7571799  1.32135653
## 4 -0.6286291 -0.4086988  0.9506200 -0.38883734
## 5 -1.1727674 -1.2078573 -1.0045069 -1.10202608
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              3              3              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              4              4              3              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              4              5              3              1              5
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              2              5              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              4              3              5              2              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              3              5              4
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              2              5              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              4              4              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              5              2              3              4              5
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              4              5              5              1
##
## Within cluster sum of squares by cluster:
## [1]  7.788275  6.128432 18.257332  9.326266  7.443899
##  (between_SS / total_SS =  75.0 %)
```

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
km_6 <- kmeans(Data_scaled, centers = 6, nstart = 10)
km_6
```

```
## K-means clustering with 6 clusters of sizes 8, 4, 7, 7, 11, 13
##
## Cluster means:
##       Murder    Assault    UrbanPop        Rape
## 1  0.8666035  1.2103171  0.82626566  0.84936722
## 2  0.4562038  0.9358314  0.61900839  2.26533514
## 3  1.5803956  0.9662584 -0.77751086  0.04844071
## 4 -0.6958674 -0.5679476  1.12728218 -0.55096728
## 5 -1.1034717 -1.1654231 -0.99194587 -1.04874074
## 6 -0.2162425 -0.2611064 -0.04793489 -0.06172647
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              1              6              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              4              6              1              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              4              5              1              6              5
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              6              6              3              5              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              4              1              5              3              6
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              5              6              2              5              4
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              3              5              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              6              6              4              4              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              5              3              1              4              5
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              6              6              5              5              6
##
## Within cluster sum of squares by cluster:
## [1]  5.888384  6.257771  6.128432  5.244931  8.499862 10.860162
##  (between_SS / total_SS =  78.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
km_3$tot.withinss - km_2$tot.withinss
```

```
## [1] -24.53913
```

```r
km_4$tot.withinss - km_3$tot.withinss
```

```
## [1] -21.9201
```

```r
km_5$tot.withinss - km_4$tot.withinss
```

```
## [1] -7.45897
```

```r
km_6$tot.withinss - km_5$tot.withinss
```

```
## [1] -6.06466
```

```r
# We compute the change in the SSE for different clusters. We want to see
# how big the change from k=2 to k=3, from k=4 to k=5 are, etc. We generally
# choose the k with the smaller gap. I can see from the WCSS graph above, k=4
# looks to have a smaller change from k=4 to k=5, which confirming by the
# calculation. This means after k=4, the SSE will not be changed too much,
# however, if we set k=2, the SSE would continue to change a lot.
# Therefore, we will choose k=4 as the optimal one in this situation.
```

==========

Question 4

==========

```r
# 4 clusters
set.seed(123)
km_4 <- kmeans(Data_scaled, centers = 4, nstart = 10)
km_4
```

```
## 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"
```

```
# Clusters sizes are 13, 16, 8, 13
# Cluster means are presented right below
```

```
km_4$withinss    # Vector of within-cluster sum of squares
```

```
## [1]  8.316061 11.952463 16.212213 19.922437
```

```
km_4$tot.withinss    # Total within-cluster sum of squares  i.e.sum(withinss).
```

```
## [1] 56.40317
```

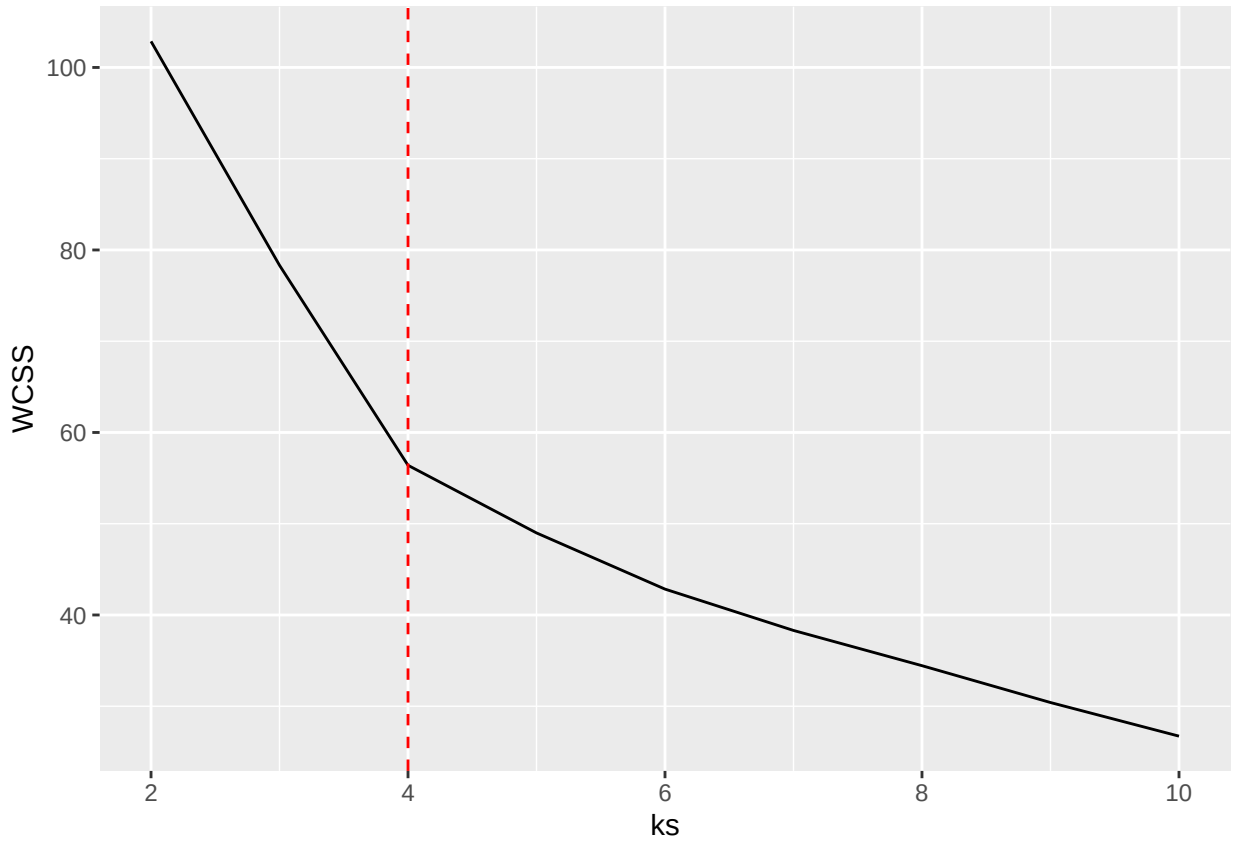
```
km_4$betweenss    # The between-cluster sum of squares, i.e.totss-tot.withinss.
```

```
## [1] 139.5968
```

```
km_4$size    # The number of points in each cluster.
```

```
## [1]  8 13 16 13
```

```
# Calculate the silhouette index, and list the index and also plot the
# silhouette index from k=2 to k=10 and confirm that the optimal cluster is 4
set.seed(1234)
ks <- 2:10
WCSS <- sapply(ks, FUN = function(k) {
kmeans(Data_scaled, centers = k, nstart = 5)$tot.withinss
})
ggplot(as_tibble(ks, WCSS), aes(ks, WCSS)) + geom_line() +
  geom_vline(xintercept = 4, color = "red", linetype = 2)
```



==========

Question 5

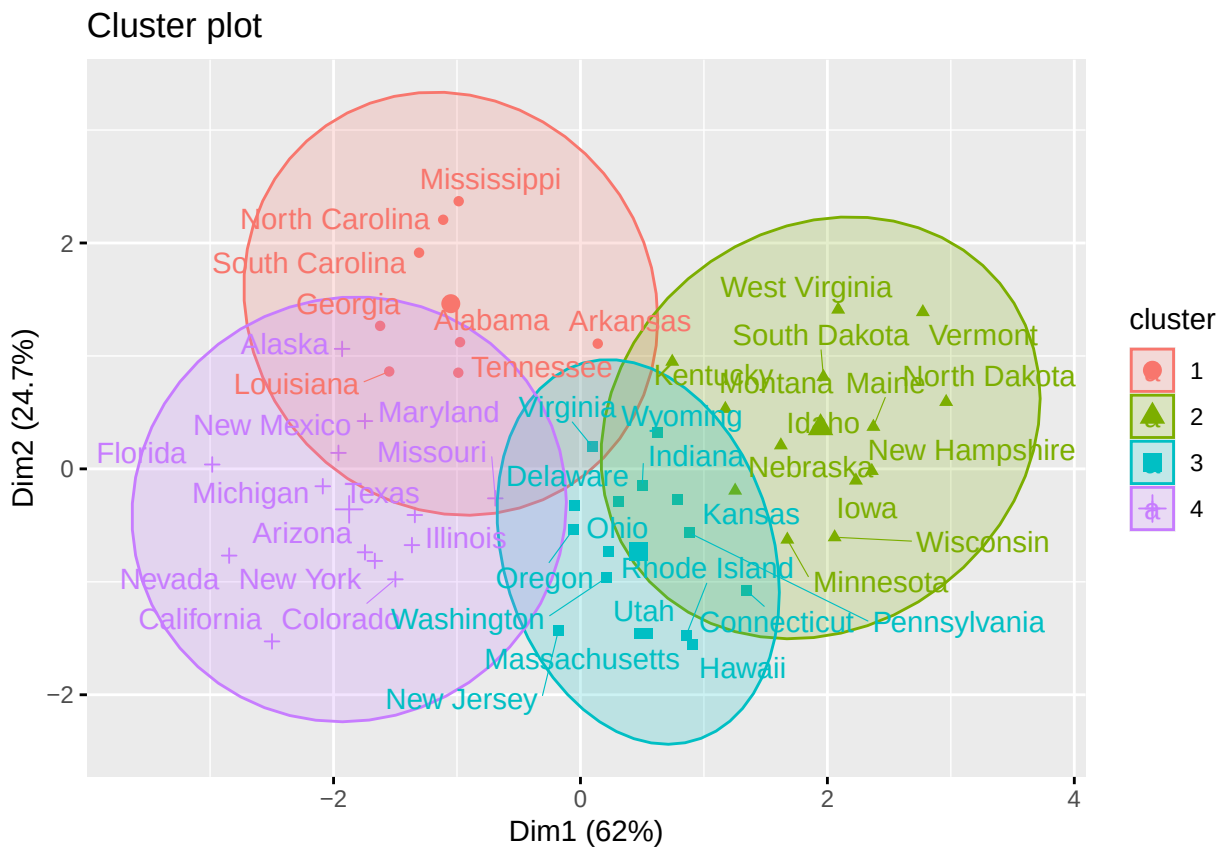
==========

```
library(factoextra)
```

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

```
fviz_cluster(km_4, data = Data_scaled, centroids = TRUE, repel = TRUE, ellipse.type = "norm")
```

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

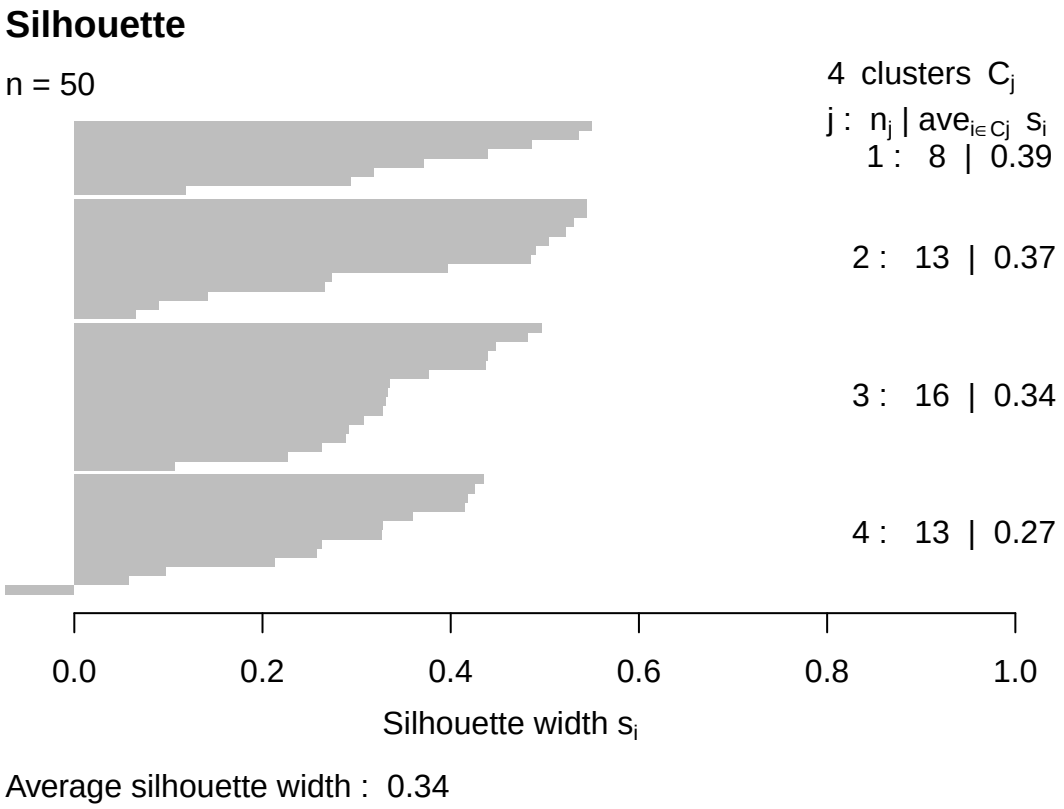


==========

Question 6

==========

```
library(cluster)
plot(silhouette(km_4$cluster, dist(Data_scaled[,1:4])), main="Silhouette")
```



The index with 4 clusters are lowest, 3 clusters are highest. It means that we should use 3 clusters. Nevertheless, all of the index are smaller than 0.4 . An index that smaller than 0.4 is considered not good.