ML Assignment 4

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:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Loading the Data

```r
rm(list = ls())

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
```

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

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

```r
#install.packages("factoextra")
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.3
```

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

```r
library(ISLR)
set.seed(123)

DFUniver<-read.csv("Universities.csv")
colnames(DFUniver)
```

```
##  [1] "College.Name"               "State"
##  [3] "Public..1...Private..2."    "X..appli..rec.d"
##  [5] "X..appl..accepted"          "X..new.stud..enrolled"
##  [7] "X..new.stud..from.top.10." "X..new.stud..from.top.25."
##  [9] "X..FT.undergrad"            "X..PT.undergrad"
## [11] "in.state.tuition"          "out.of.state.tuition"
## [13] "room"                      "board"
## [15] "add..fees"                 "estim..book.costs"
## [17] "estim..personal.."         "X..fac..w.PHD"
## [19] "stud..fac..ratio"          "Graduation.rate"
```

```
#summary(DFUniver)
```

Removing missing records from the Dataset (Measurements)

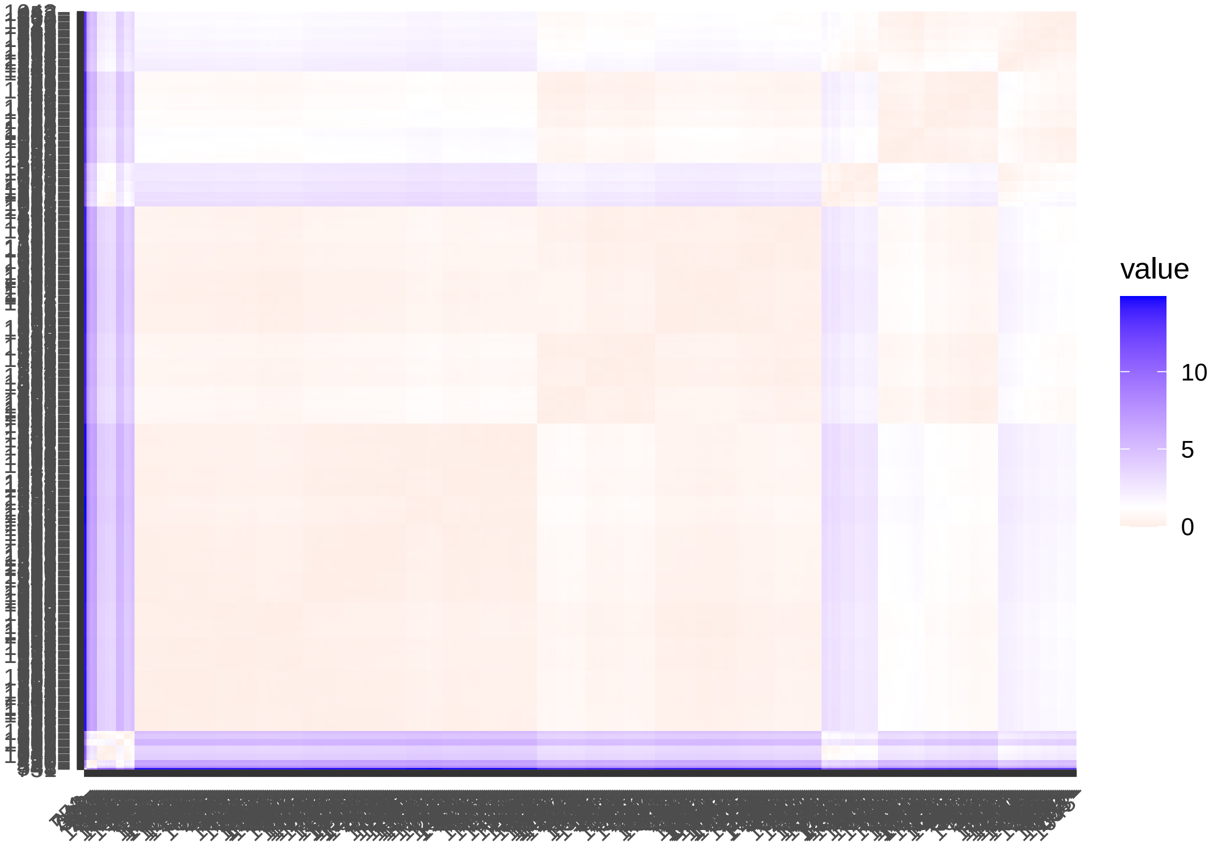
```
DFUniver1<-na.omit(DFUniver, cols=c("in.state.tuition","out.of.state.tuition", "Graduation.rate", "State
```

Scaling the data

```
DFUniver1[,c(-1,-2)]<- scale(DFUniver1[,c(-1,-2)])
distance <- get_dist(DFUniver1[,c(2,5)])
```

```
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion
```

```
fviz_dist(distance)
```



Finding the K mean values

```
DFUniver2<-DFUniver1[,c(9,11)]
k4 <- kmeans(DFUniver2, centers = 4, nstart = 25) # k = 4, number of restarts = 25

# Visualize the output

k4$centers # output the centers
```

```
##   X..FT.undergrad in.state.tuition
## 1      2.55730366       -1.2170108
## 2     -0.46382892        0.1723828
## 3      0.04915656       -1.2152699
## 4     -0.14518440        1.4026492
```

```
#number of Universities in each cluster
k4$size
```

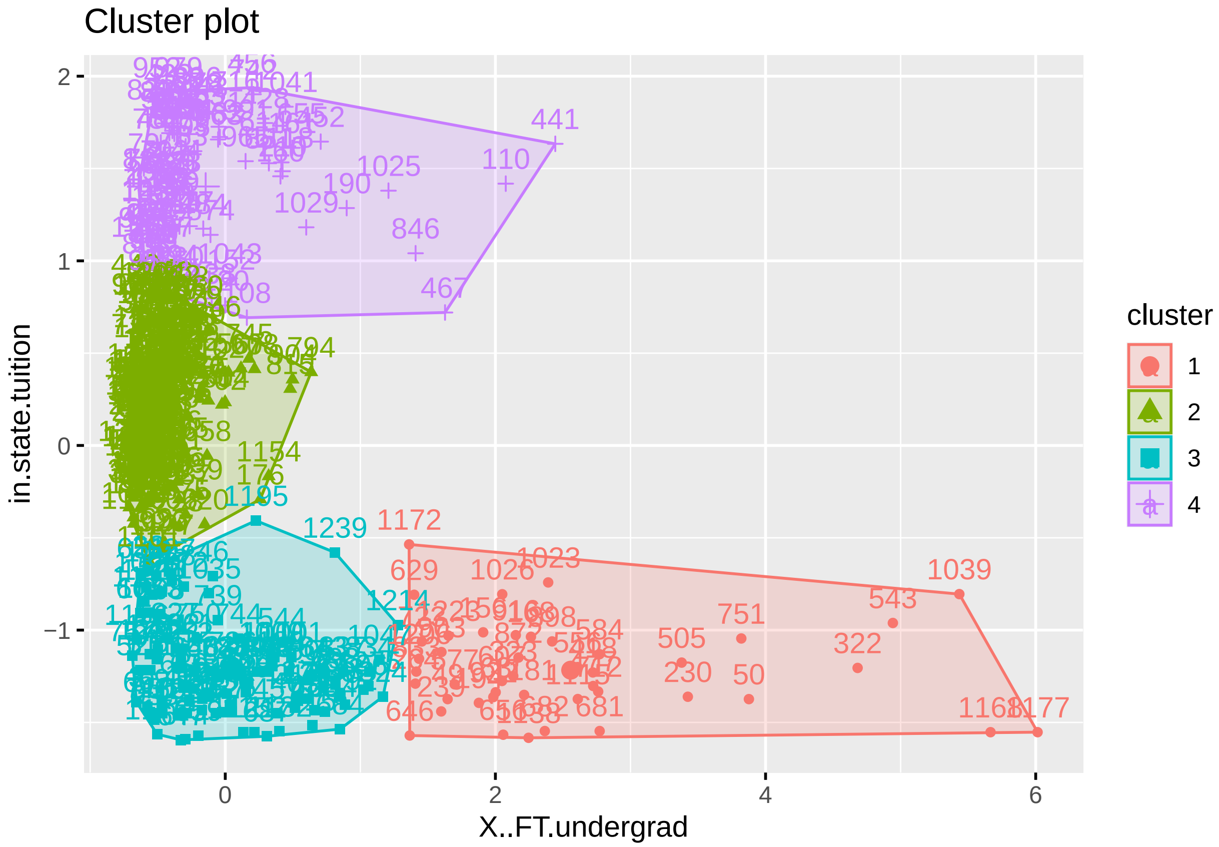
```
## [1]  44 222 104 101
```

```
# Identify the cluster of the 120th observation as an example
k4$cluster[120]
```

```
## 377
##   2
```

```
# Visualize the output

fviz_cluster(k4, data = DFUniver2)
```



It is now easy to see that the bottom right cluster represents Universities with maximum undergrad student with low tution fees.

Usage of manhattan distance

```
#install.packages("flexclust")
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.0.3
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Warning: package 'modeltools' was built under R version 4.0.3
```

```
## Loading required package: stats4
```

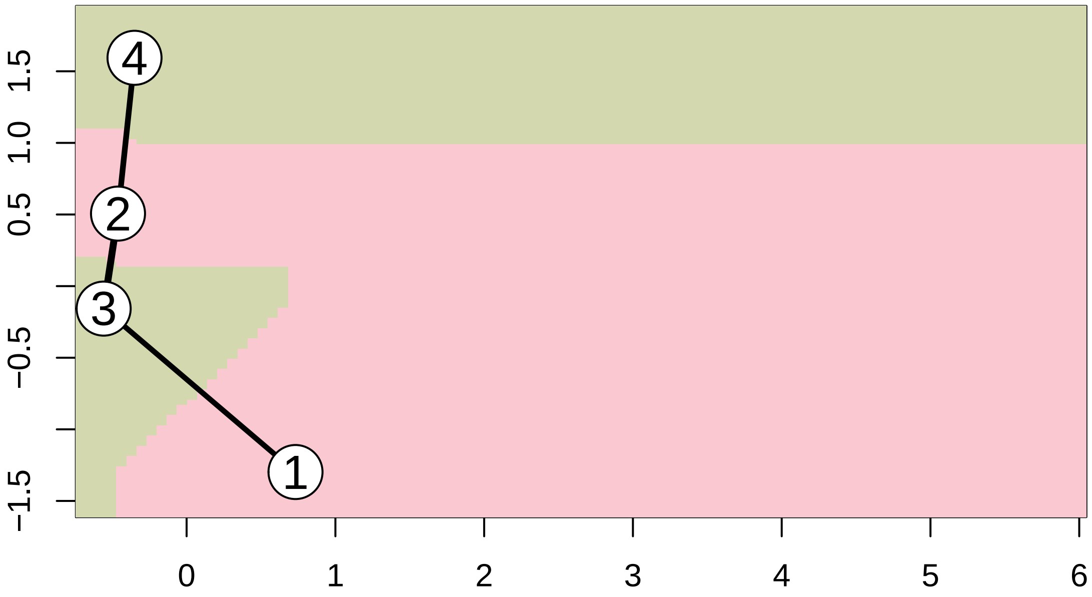
```
set.seed(123)
#kmeans clustering, using manhattan distance
k4 = kcca(DFUniver2, k=4, kccaFamily("kmedians"))
k4
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = DFUniver2, k = 4, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
##   1   2   3   4
## 116 131 145  79
```

```
#Let us now apply the predict function
#Apply the predict() function
clusters_index <- predict(k4)
dist(k4@centers)
```

```
##           1         2         3
## 2 2.1631540
## 3 1.7218736 0.6704232
## 4 3.0878458 1.0932316 1.7632720
```

```
image(k4)
```

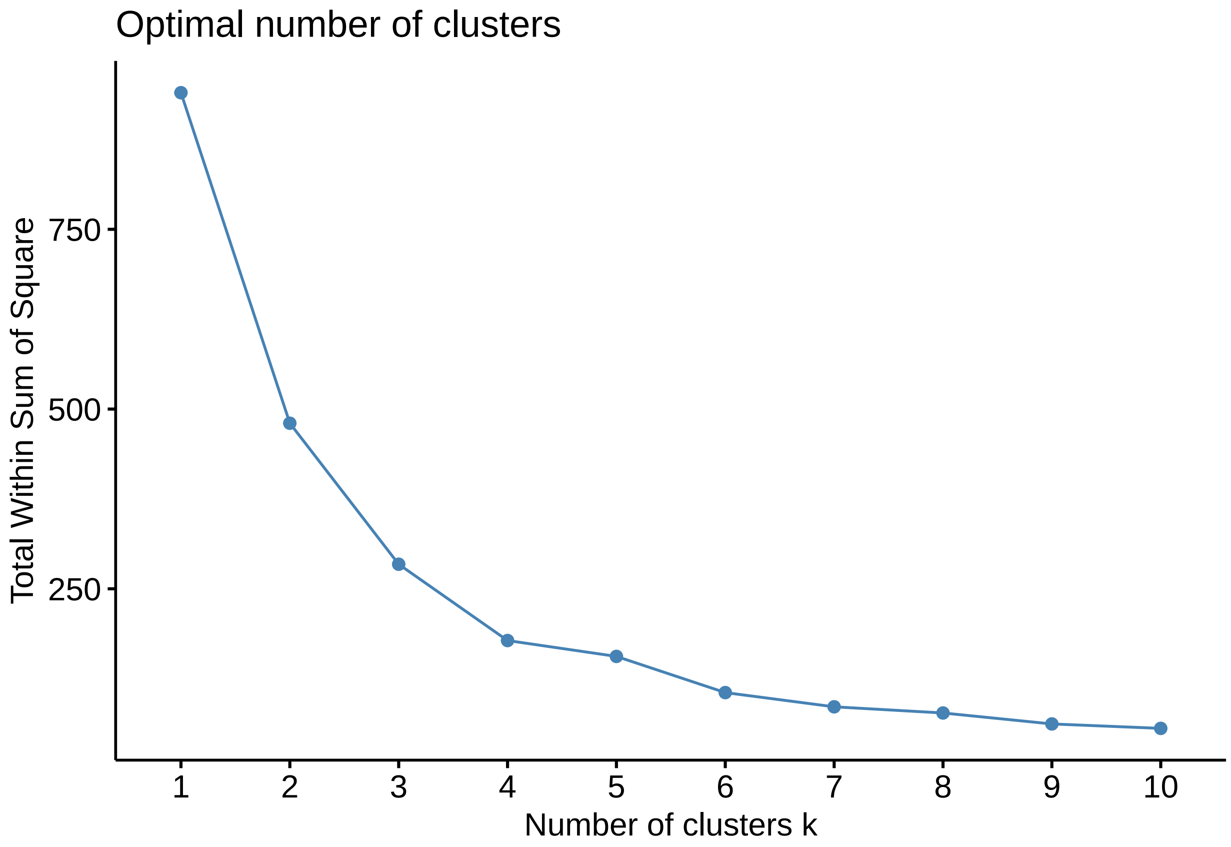


```
#points(df, col=clusters_index, pch=19, cex=0.3)
```

Determining k value using "elbo chart" to determine k

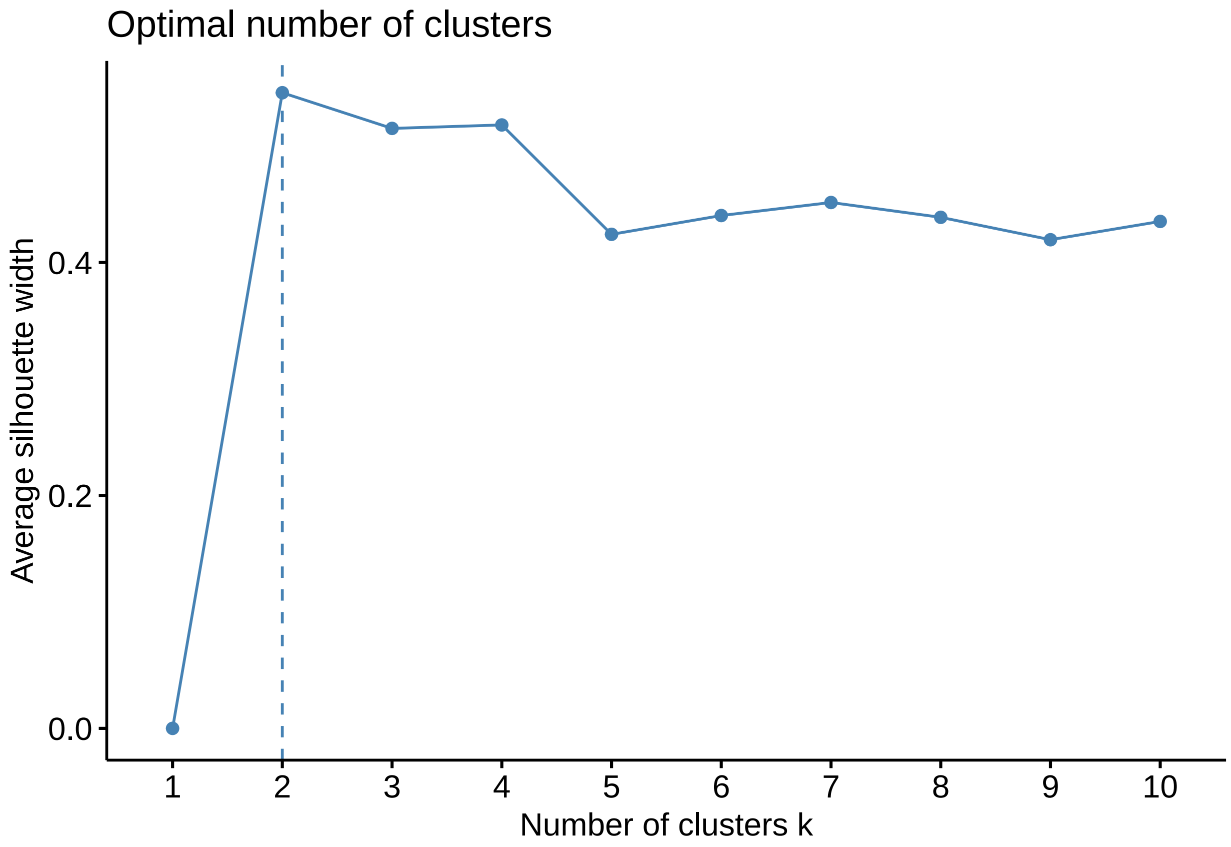
```
library(tidyverse)  # data manipulation
library(factoextra) # clustering & visualization
library(ISLR)
set.seed(123)

df<-DFUniver1[,c(9,11)]
# Scaling the data frame (z-score)
df <- scale(df)
fviz_nbclust(df, kmeans, method = "wss")
```



Let us now apply the Silhouette Method to determine the number of clusters

```
fviz_nbclust(df, kmeans, method = "silhouette")
```


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
1
2
3
4
5
6
7
8
9
10
Number of clusters k