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)
DFUniver<-DFUniver%>%rename(
  Pub.Private=Public..1...Private..2.,
  ApplRec=X..appli..rec.d,
  ApplAccept=X..appl..accepted,
  NewStdEnr=X..new.stud..enrolled,
  Top10=X..new.stud..from.top.10.,
  Top25=X..new.stud..from.top.25.,
  FTUnderG=X..FT.undergrad,
  PTUnderG=X..PT.undergrad,
  InStateFee=in.state.tuition,
  OutStateFee=out.of.state.tuition,
  BookCost=estim..book.costs,
  PerCost=estim..personal..,
  PHD=X..fac..w.PHD,
  StFactRatio=stud..fac..ratio
)

colnames(DFUniver)
```

```
##  [1] "College.Name"    "State"           "Pub.Private"     "ApplRec"
##  [5] "ApplAccept"      "NewStdEnr"       "Top10"           "Top25"
##  [9] "FTUnderG"        "PTUnderG"        "InStateFee"      "OutStateFee"
## [13] "room"            "board"           "add..fees"       "BookCost"
## [17] "PerCost"         "PHD"             "StFactRatio"     "Graduation.rate"
```

Removing missing records from the Dataset (Measurements)

```
#Total NULL fields in the data frame
count(DFUniver[!complete.cases(DFUniver),])
```

```
##     n
## 1 831
```

```
#Ipute the NULL values
DFUniver1<-na.omit(DFUniver)
```

Finding the Data Summary and Measure of Dependence

```
#Summary Data
summary(DFUniver1)
```

```
##  College.Name          State             Pub.Private       ApplRec
##  Length:471         Length:471         Min.   :1.000   Min.   :   77
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.:  802
##  Mode  :character   Mode  :character   Median :2.000   Median : 1646
##                                        Mean   :1.728   Mean   : 3147
##                                        3rd Qu.:2.000   3rd Qu.: 3862
##                                        Max.   :2.000   Max.   :48094
##    ApplAccept        NewStdEnr          Top10           Top25
##  Min.   :   61.0   Min.   :  27.0   Min.   : 1.00   Min.   :  9.00
##  1st Qu.:  635.5   1st Qu.: 264.0   1st Qu.:15.00   1st Qu.: 40.00
##  Median : 1227.0   Median : 443.0   Median :23.00   Median : 54.00
##  Mean   : 2063.0   Mean   : 780.7   Mean   :28.01   Mean   : 55.65
##  3rd Qu.: 2456.0   3rd Qu.: 896.5   3rd Qu.:36.00   3rd Qu.: 69.00
##  Max.   :26330.0   Max.   :6392.0   Max.   :96.00   Max.   :100.00
##     FTUnderG        PTUnderG         InStateFee     OutStateFee
##  Min.   :  249   Min.   :    1.0   Min.   :  608   Min.   : 1044
##  1st Qu.: 1018   1st Qu.:   81.5   1st Qu.: 3650   1st Qu.: 7290
##  Median : 1715   Median :  299.0   Median : 9858   Median :10100
##  Mean   : 3563   Mean   :  797.5   Mean   : 9407   Mean   :10575
##  3rd Qu.: 4056   3rd Qu.:  869.0   3rd Qu.:13246   3rd Qu.:13286
##  Max.   :31643   Max.   :21836.0   Max.   :20100   Max.   :20100
##       room          board        add..fees         BookCost         PerCost
##  Min.   : 640   Min.   : 531   Min.   :  10.0   Min.   :  90.0   Min.   : 250
##  1st Qu.:1740   1st Qu.:1750   1st Qu.: 137.5   1st Qu.: 500.0   1st Qu.: 850
##  Median :2090   Median :2082   Median : 280.0   Median : 500.0   Median :1200
##  Mean   :2221   Mean   :2122   Mean   : 379.0   Mean   : 548.8   Mean   :1312
##  3rd Qu.:2663   3rd Qu.:2420   3rd Qu.: 486.0   3rd Qu.: 600.0   3rd Qu.:1600
##  Max.   :4816   Max.   :4541   Max.   :3247.0   Max.   :2340.0   Max.   :6800
##       PHD          StFactRatio    Graduation.rate
##  Min.   :  8.00   Min.   : 2.90   Min.   : 15.00
##  1st Qu.: 63.00   1st Qu.:11.30   1st Qu.: 53.00
##  Median : 76.00   Median :13.40   Median : 66.00
##  Mean   : 73.21   Mean   :13.96   Mean   : 65.56
##  3rd Qu.: 87.00   3rd Qu.:16.45   3rd Qu.: 79.00
##  Max.   :103.00   Max.   :28.80   Max.   :118.00
```

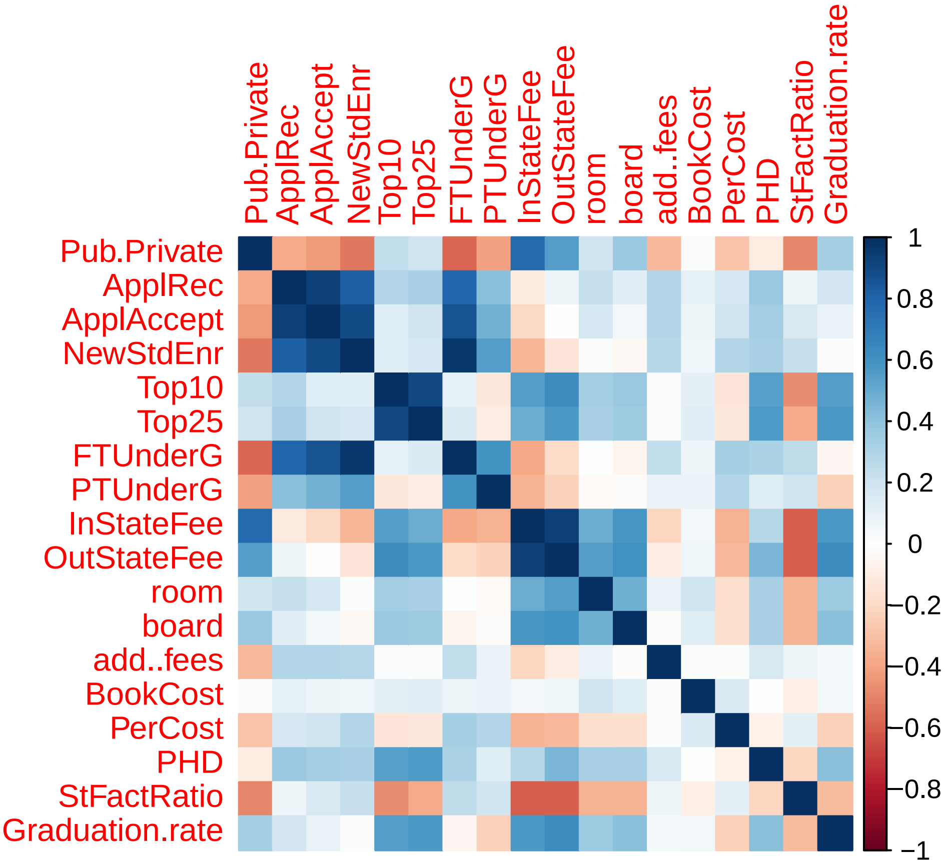
```r
#Finding the correlation between the data set
#Selecting numerical columns only

DFNumerical<-DFUniver1[,c(-1,-2)]
library(corrplot)
```

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

```
## corrplot 0.84 loaded
```

```r
corrplot(cor(DFNumerical), method = "color")
```



In the correlation graph, Darker Blue and Dark Orange shows the higher correlated data. Using this data to understand any correlation among the column data.

Finding the K-means clustering values - Universities of Public & Private type and In State Fee Amount

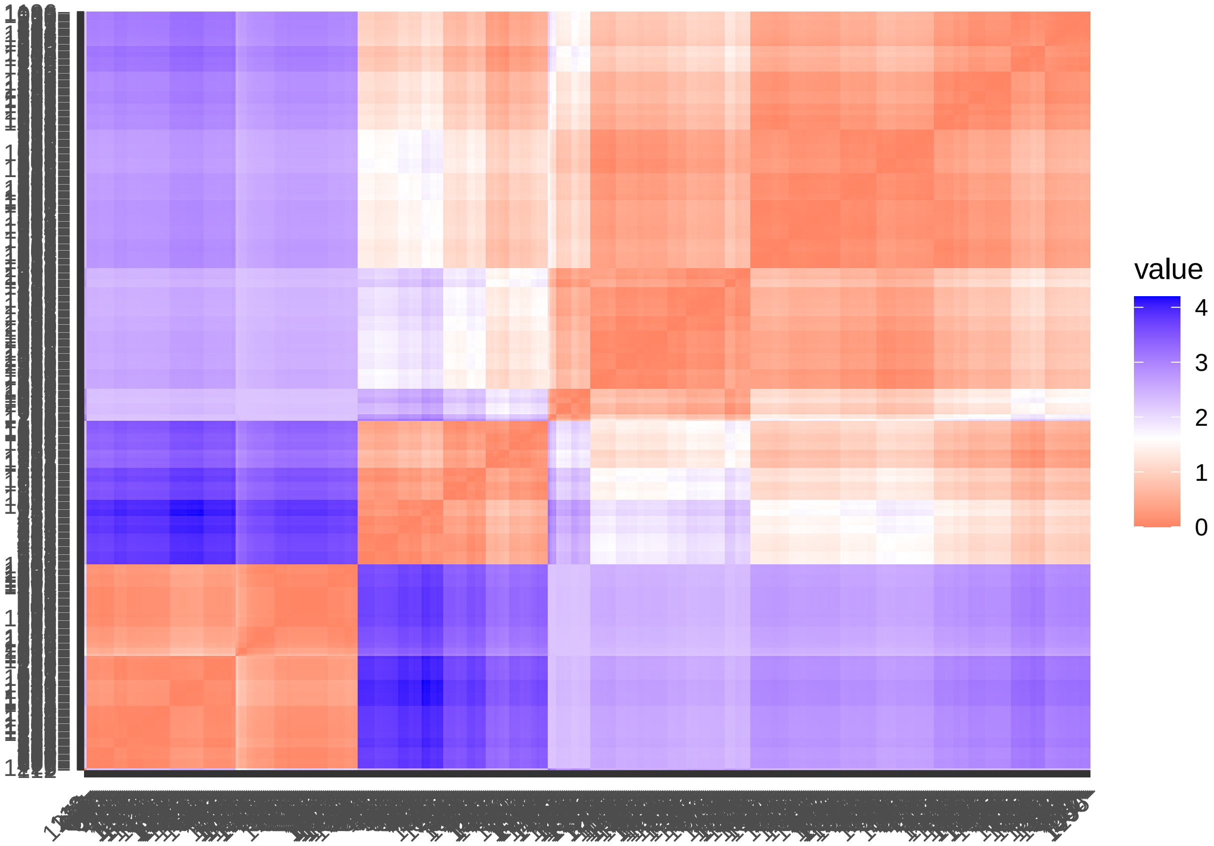
```
colnames(DFNumerical)
```

```
##  [1] "Pub.Private"     "ApplRec"         "ApplAccept"      "NewStdEnr"
##  [5] "Top10"           "Top25"           "FTUnderG"        "PTUnderG"
##  [9] "InStateFee"      "OutStateFee"     "room"            "board"
## [13] "add..fees"       "BookCost"        "PerCost"         "PHD"
## [17] "StFactRatio"     "Graduation.rate"
```

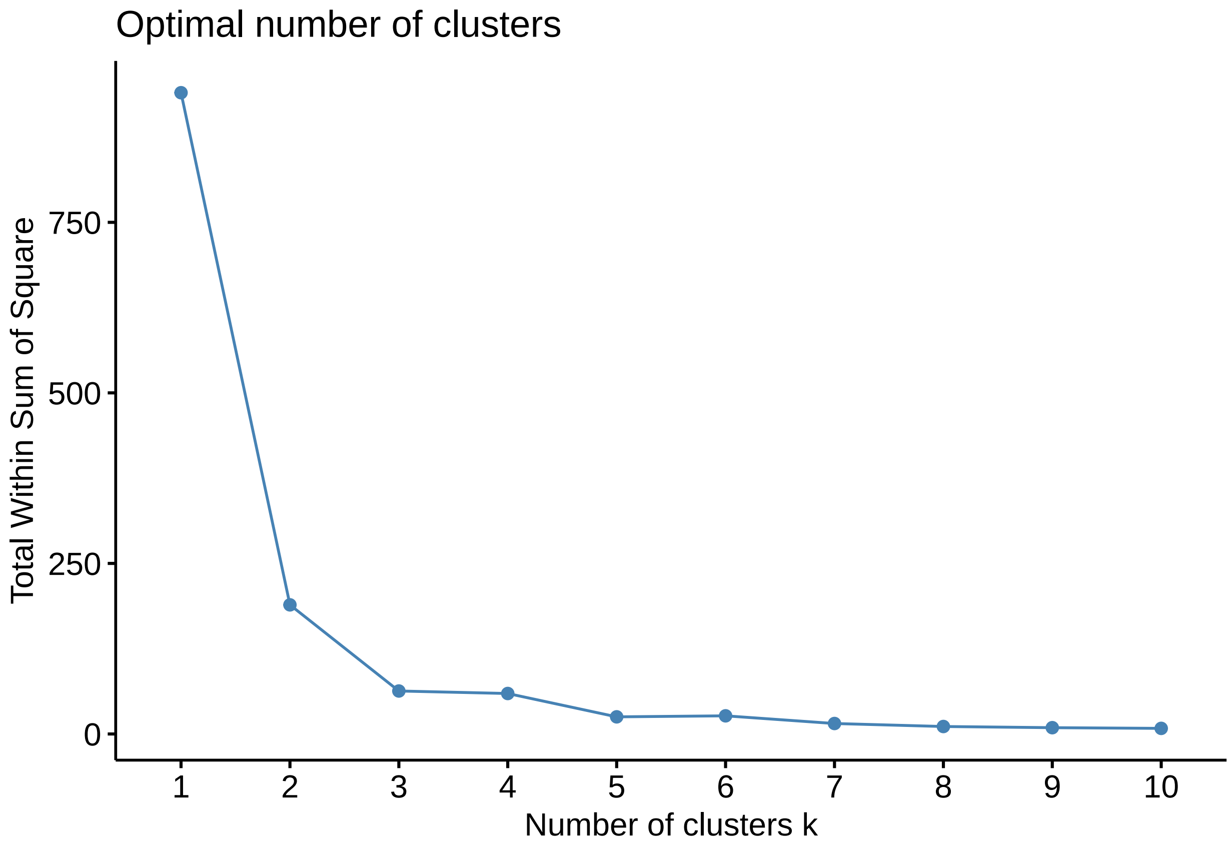
```
DFPubPriInState<-DFNumerical[,c(1,9)]

#Scaling the Data
DFPubPriInState<-scale(DFPubPriInState)

#Distance Between Observations
distance <- get_dist(DFPubPriInState)
fviz_dist(distance)
```



```
#Finding optimal number of clusters - Elbow Method
fviz_nbclust(DFPubPriInState, kmeans, method = "wss")
```



```
#From the Elbow method we can see that optimum no. of cluster size is 7, for this data set

#Clustering of In State Fee data with the relation between Public and Private Universities
k4 <- kmeans(DFPubPriInState, centers = 3, nstart = 25) # k = 3, number of restarts = 25

# Visualize the output

k4$centers # output the centers
```

```
##   Pub.Private  InStateFee
## 1    0.610234  0.02079352
## 2    0.610234  1.28826525
## 3   -1.635236 -1.26377914
```

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

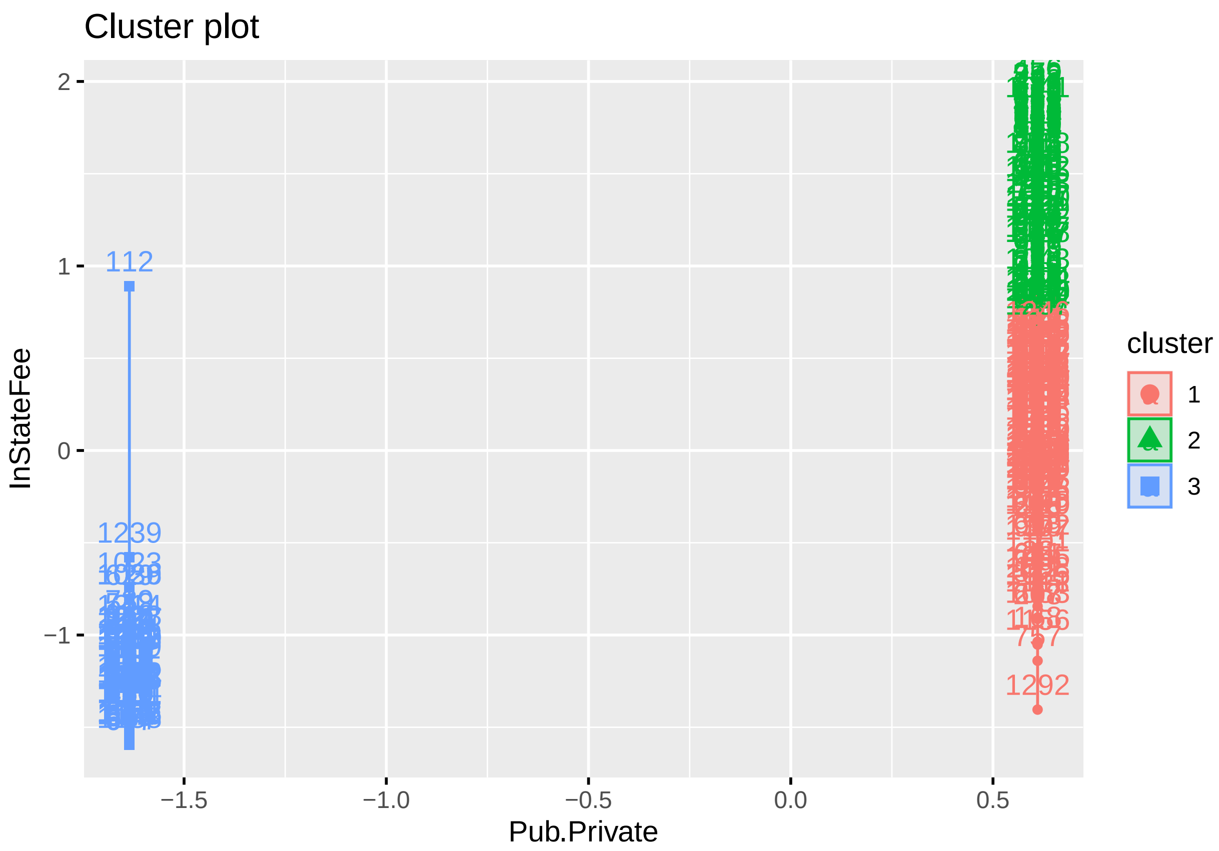
```
## [1] 221 122 128
```

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

```
## 377
##   1
```

```
# Visualize the output

fviz_cluster(k4, data = DFPubPriInState)
```



Making cluster on another data set - New Student Enroll and Out of State Fee

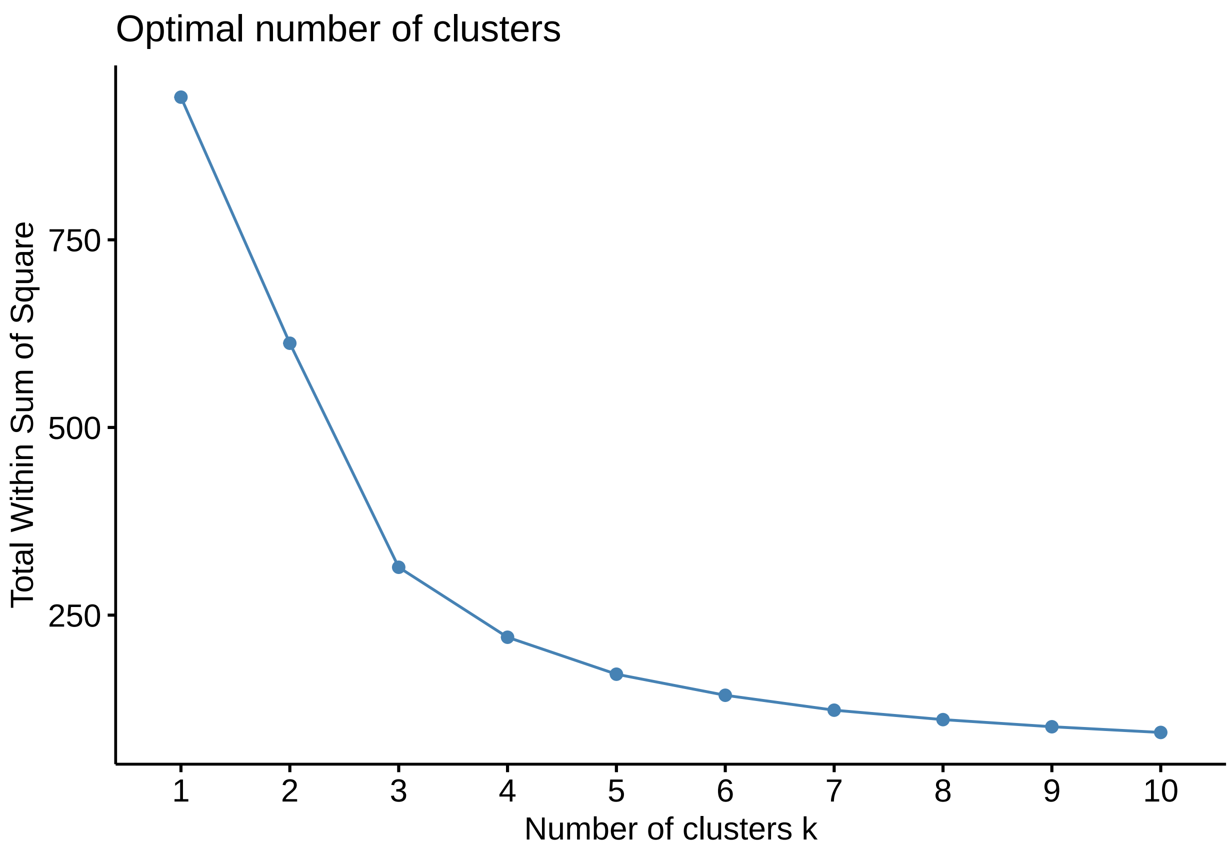
```
colnames(DFNumerical)
```

```
##  [1] "Pub.Private"     "ApplRec"         "ApplAccept"      "NewStdEnr"
##  [5] "Top10"           "Top25"           "FTUnderG"        "PTUnderG"
##  [9] "InStateFee"      "OutStateFee"     "room"            "board"
## [13] "add..fees"       "BookCost"        "PerCost"         "PHD"
## [17] "StFactRatio"     "Graduation.rate"
```

```
DFEnrolOutStFee<-DFNumerical[,c(4,10)]

#Scaling the Data
DFEnrolOutStFee<-scale(DFEnrolOutStFee)

#Finding optimal number of clusters - Elbow Method
fviz_nbclust(DFEnrolOutStFee, kmeans, method = "wss")
```



```
#From the Elbow method we can see that optimum no. of cluster size is 4, for this data set

#Clustering of Out State Fee data with relation between Student Enroll by the Universities
k4 <- kmeans(DFEnrolOutStFee, centers = 5, nstart = 25) # k = 5, number of restarts = 25

# Visualize the output

k4$centers # output the centers
```

```
##     NewStdEnr OutStateFee
## 1  3.4222956  -0.5835170
## 2 -0.2756069  -1.0159117
## 3  1.2669381  -0.6885870
## 4 -0.4671991   0.1021368
## 5 -0.1035616   1.5882927
```

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

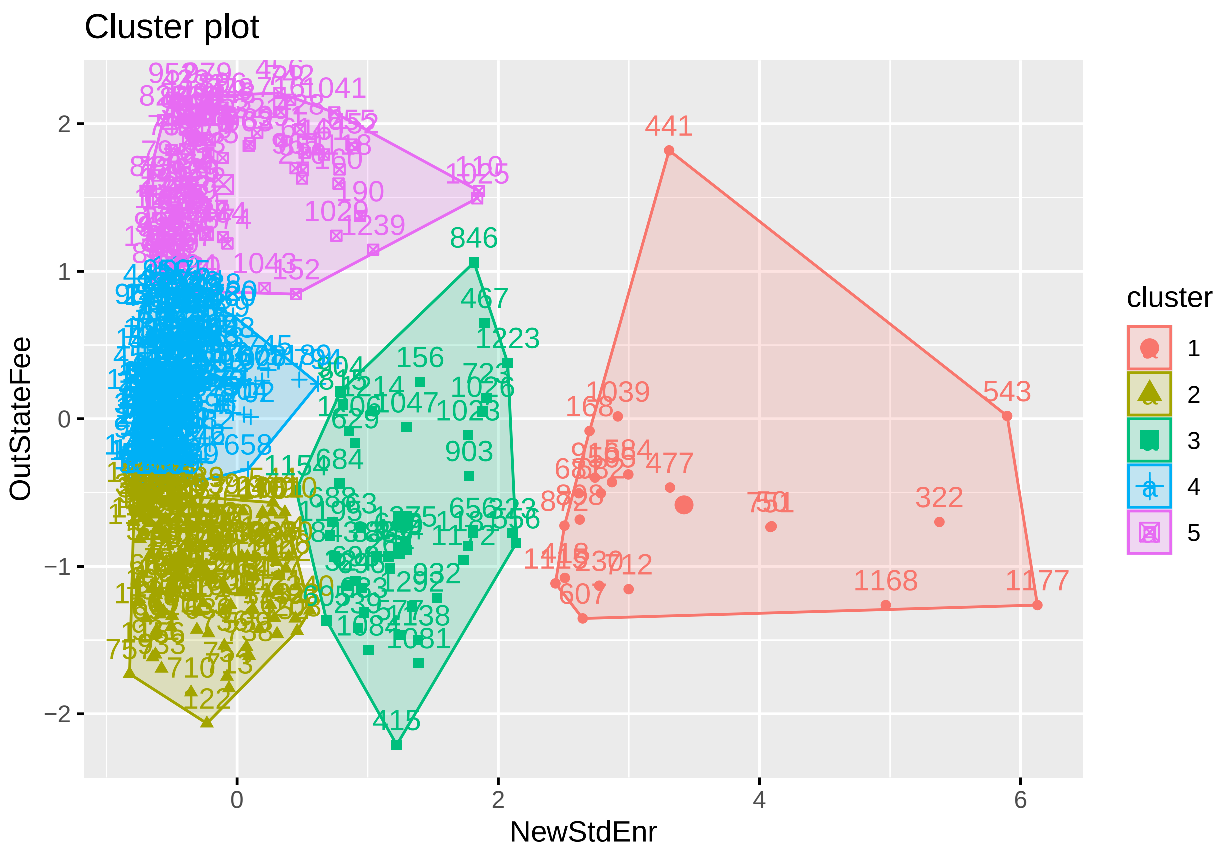
```
## [1]  22 120  45 192  92
```

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

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

```
# Visualize the output
```

```
fviz_cluster(k4, data = DFEnrolOutStFee)
```



Making cluster on another data set - Student Application and Out of State Fee

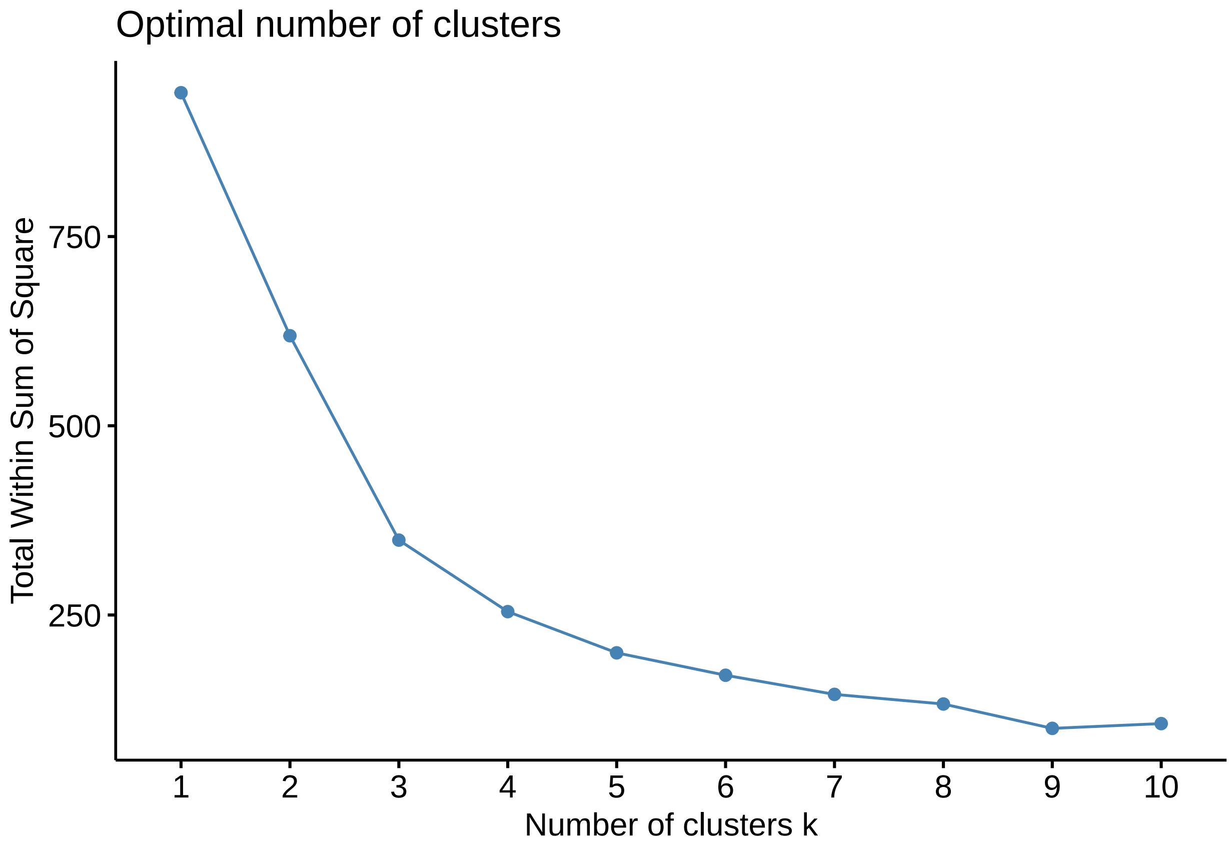
```
colnames(DFNumerical)
```

```
##  [1] "Pub.Private"     "ApplRec"         "ApplAccept"      "NewStdEnr"
##  [5] "Top10"           "Top25"           "FTUnderG"        "PTUnderG"
##  [9] "InStateFee"      "OutStateFee"     "room"            "board"
## [13] "add..fees"       "BookCost"        "PerCost"         "PHD"
## [17] "StFactRatio"     "Graduation.rate"
```

```
DFStAppOutStFee<-DFNumerical[,c(3,10)]

#Scaling the Data
DFStAppOutStFee<-scale(DFStAppOutStFee)

#Finding optimal number of clusters - Elbow Method
fviz_nbclust(DFStAppOutStFee, kmeans, method = "wss")
```



```
#From the Elbow method we can see that optimum no. of cluster size is 5, for this data set
k4 <- kmeans(DFStAppOutStFee, centers = 5, nstart = 25) # k = 5, number of restarts = 25

# Visualize the output

k4$centers # output the centers
```

```
##    ApplAccept OutStateFee
## 1 -0.45441373  0.07885858
## 2 -0.23953710 -1.06140991
## 3  4.33206043 -0.34643331
## 4  1.49079739 -0.38798336
## 5  0.01600671  1.56687709
```

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

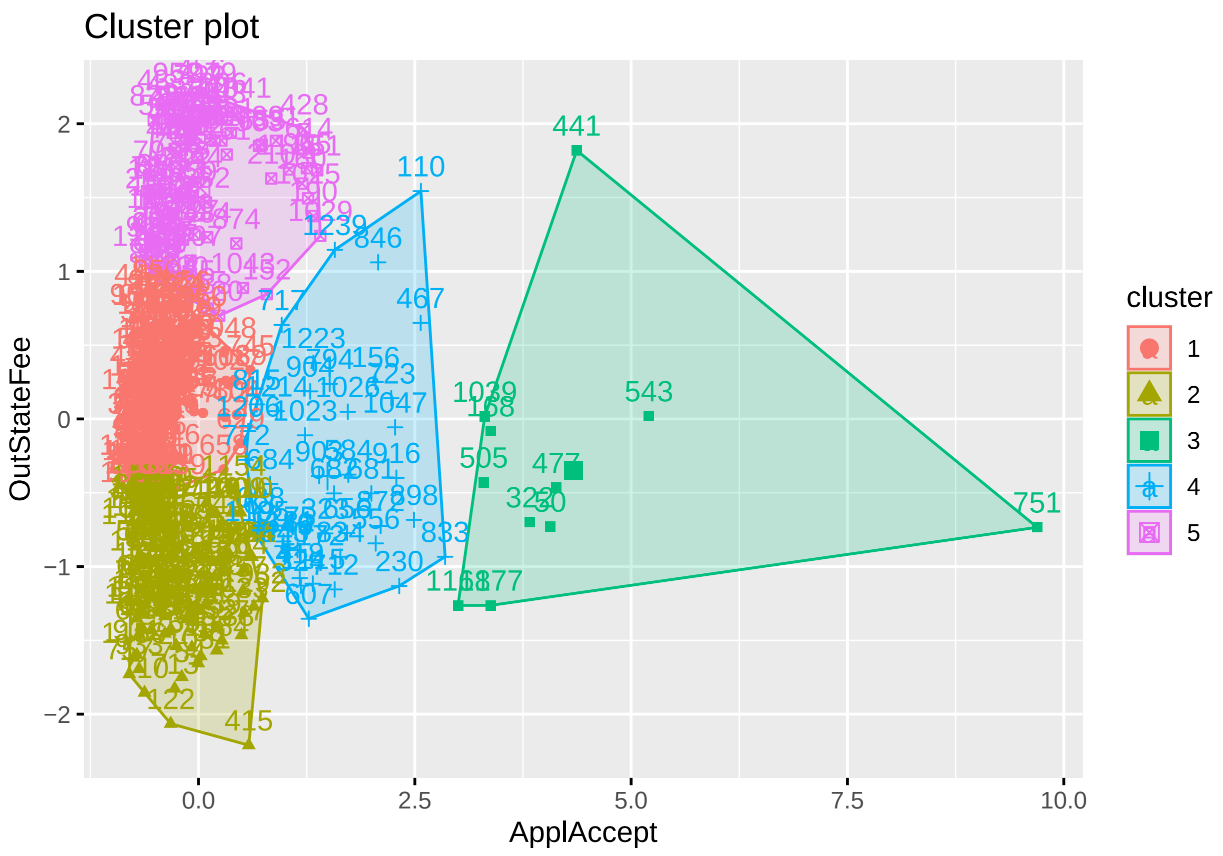
```
## [1] 190 131  11  46  93
```

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

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

```
# Visualize the output

fviz_cluster(k4, data = DFStAppOutStFee)
```



Cluster plot
OutStateFee
ApplAccept
2
1
0
−1
−2
0.0
2.5
5.0
7.5
10.0
cluster
1
2
3
4
5
441
543
751
1039
168
505
477
322
30
1168
877
110
1239
846
467
717
1223
904
156
723
1026
1047
1023
916
898
833
230
607
428
474
1043
132
415
122
1013