Machine Learning Assignment#4

Rakhee Moolchandani

11/01/2020

#load all the required libraries

```
library(readr)
library(gmodels)
library(ISLR)
library(dplyr)
library(tidyverse)
library(factoextra)
library(caret)
library(compareGroups)
library(data.table)
library(fpc)
library(ggplot2)
library(GGally)
```

Import the Universities Data

```
#Read the data set
UniversityData <- read.csv("Universities.csv")
#Show the first few rows of the data set
head(UniversityData)
```

```
##                         College.Name State Public..1...Private..2.
## 1         Alaska Pacific University    AK                       2
## 2 University of Alaska at Fairbanks    AK                       1
## 3    University of Alaska Southeast    AK                       1
## 4 University of Alaska at Anchorage    AK                       1
## 5       Alabama Agri. & Mech. Univ.    AL                       1
## 6               Faulkner University    AL                       2
##   X..appli..rec.d X..appl..accepted X..new.stud..enrolled
## 1             193               146                    55
## 2            1852              1427                   928
## 3             146               117                    89
## 4            2065              1598                  1162
## 5            2817              1920                   984
## 6             345               320                   179
##   X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                        16                        44             249
## 2                        NA                        NA            3885
## 3                         4                        24             492
## 4                        NA                        NA            6209
## 5                        NA                        NA            3958
```

```
## 6                        NA                        27           1367
##   X..PT.undergrad in.state.tuition out.of.state.tuition room board add..fees
## 1             869             7560                 7560 1620  2500       130
## 2            4519             1742                 5226 1800  1790       155
## 3            1849             1742                 5226 2514  2250        34
## 4           10537             1742                 5226 2600  2520       114
## 5             305             1700                 3400 1108  1442       155
## 6             578             5600                 5600 1550  1700       300
##   estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
## 1               800              1500           76             11.9
## 2               650              2304           67             10.0
## 3               500              1162           39              9.5
## 4               580              1260           48             13.7
## 5               500               850           53             14.3
## 6               350                NA           52             32.8
##   Graduation.rate
## 1              15
## 2              NA
## 3              39
## 4              NA
## 5              40
## 6              55
```

a)Remove all records with missing measurements from the dataset. Also, remove all the categorical values

```
#Remove all the missing values from the data set by using na.omit
UniData <- na.omit(UniversityData)
#Remove all the categorical variables from the data set
UniData1 <- UniData[, c(-1, -2, -3)]
```

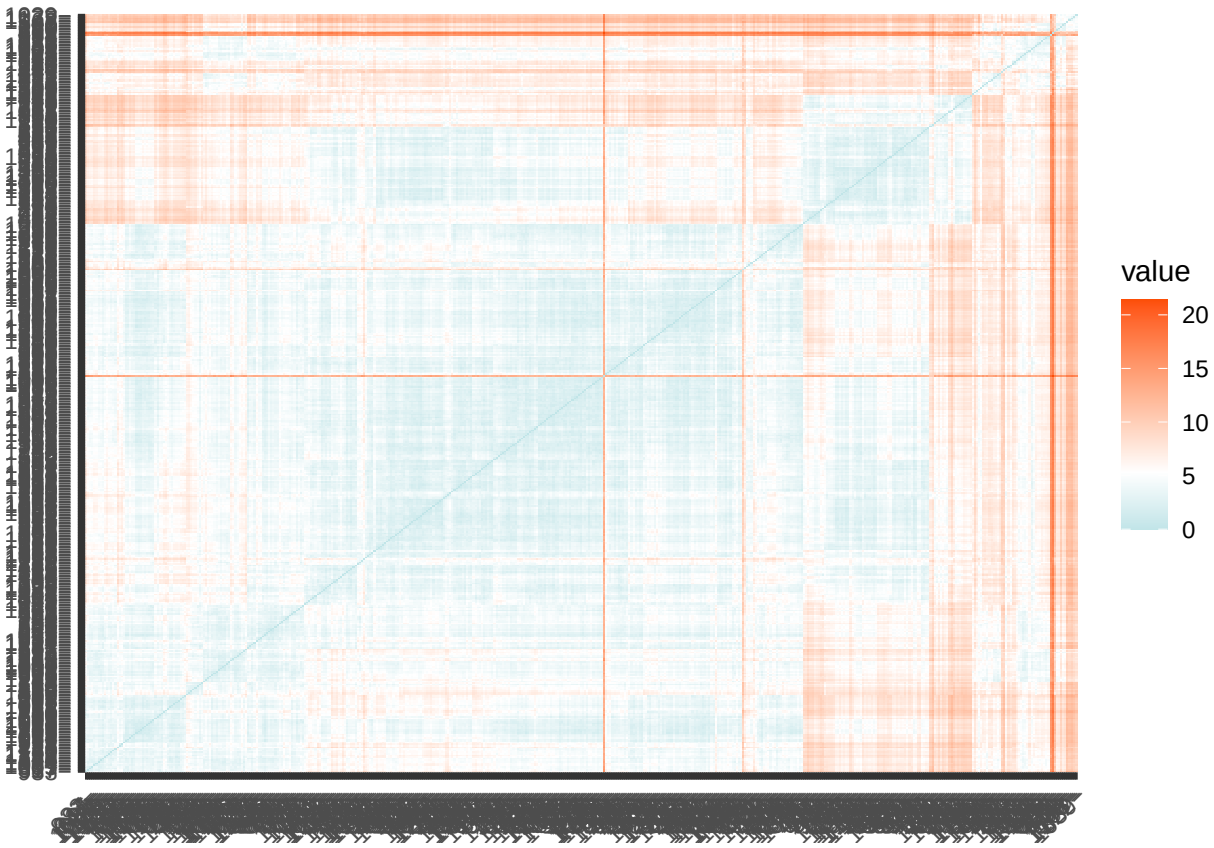
b)For all the continuous measurements, run K-Means clustering. Make sure to normalize the measurements. How many clusters seem reasonable for describing these data? What was your optimal K?

```
#Scale the data set
ScaledData <- scale(UniData1)
#Look at the summary of the scaled data
summary(ScaledData)
```

```
##  X..appli..rec.d   X..appl..accepted X..new.stud..enrolled
##  Min.   :-0.7538   Min.   :-0.7996   Min.   :-0.8232
##  1st Qu.:-0.5758   1st Qu.:-0.5701   1st Qu.:-0.5643
##  Median :-0.3686   Median :-0.3339   Median :-0.3688
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.1755   3rd Qu.: 0.1570   3rd Qu.: 0.1265
##  Max.   :11.0349   Max.   : 9.6923   Max.   : 6.1283
##  X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
##  Min.   :-1.4618           Min.   :-2.29537          Min.   :-0.7097
```

```
##  1st Qu.:-0.7042          1st Qu.:-0.77010          1st Qu.:-0.5450
##  Median :-0.2713          Median :-0.08127          Median :-0.3958
##  Mean   : 0.0000          Mean   : 0.00000          Mean   : 0.0000
##  3rd Qu.: 0.4322          3rd Qu.: 0.65676          3rd Qu.: 0.1055
##  Max.   : 3.6791          Max.   : 2.18202          Max.   : 6.0139
##  X..PT.undergrad    in.state.tuition   out.of.state.tuition      room
##  Min.   :-0.51524   Min.   :-1.59488   Min.   :-2.2105      Min.   :-2.2170
##  1st Qu.:-0.46316   1st Qu.:-1.04338   1st Qu.:-0.7619      1st Qu.:-0.6746
##  Median :-0.32246   Median : 0.08182   Median :-0.1102      Median :-0.1838
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000      Mean   : 0.0000
##  3rd Qu.: 0.04628   3rd Qu.: 0.69594   3rd Qu.: 0.6287      3rd Qu.: 0.6196
##  Max.   :13.61017   Max.   : 1.93833   Max.   : 2.2091      Max.   : 3.6384
##      board            add..fees       estim..book.costs estim..personal..
##  Min.   :-2.80658   Min.   :-1.0370   Min.   :-2.8114   Min.   :-1.5574
##  1st Qu.:-0.65614   1st Qu.:-0.6787   1st Qu.:-0.2989   1st Qu.:-0.6775
##  Median :-0.07046   Median :-0.2783   Median :-0.2989   Median :-0.1642
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.52581   3rd Qu.: 0.3006   3rd Qu.: 0.3139   3rd Qu.: 0.4225
##  Max.   : 4.26746   Max.   : 8.0594   Max.   :10.9766   Max.   : 8.0488
##  X..fac..w.PHD     stud..fac..ratio  Graduation.rate
##  Min.   :-3.9127   Min.   :-2.8374   Min.   :-2.7863
##  1st Qu.:-0.6125   1st Qu.:-0.6829   1st Qu.:-0.6923
##  Median : 0.1675   Median :-0.1443   Median : 0.0241
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.8276   3rd Qu.: 0.6380   3rd Qu.: 0.7405
##  Max.   : 1.7876   Max.   : 3.8056   Max.   : 2.8896
```

```
#Calculate the distance for the scaled data
distance <- get_dist(ScaledData)
#plot the distance using fviz_dist
fviz_dist(distance,gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



The above graph shows the distance between variables. Let us now determine the clusters.

Determining Optimal Clusters:

K Means clustering is a simple algorithm used to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-means clustering requires that you specify in advance the number of clusters to extract. A plot of the total within-groups sums of squares against the number of clusters in a k-means solution can be helpful. A bend in the graph can suggest the appropriate number of clusters.

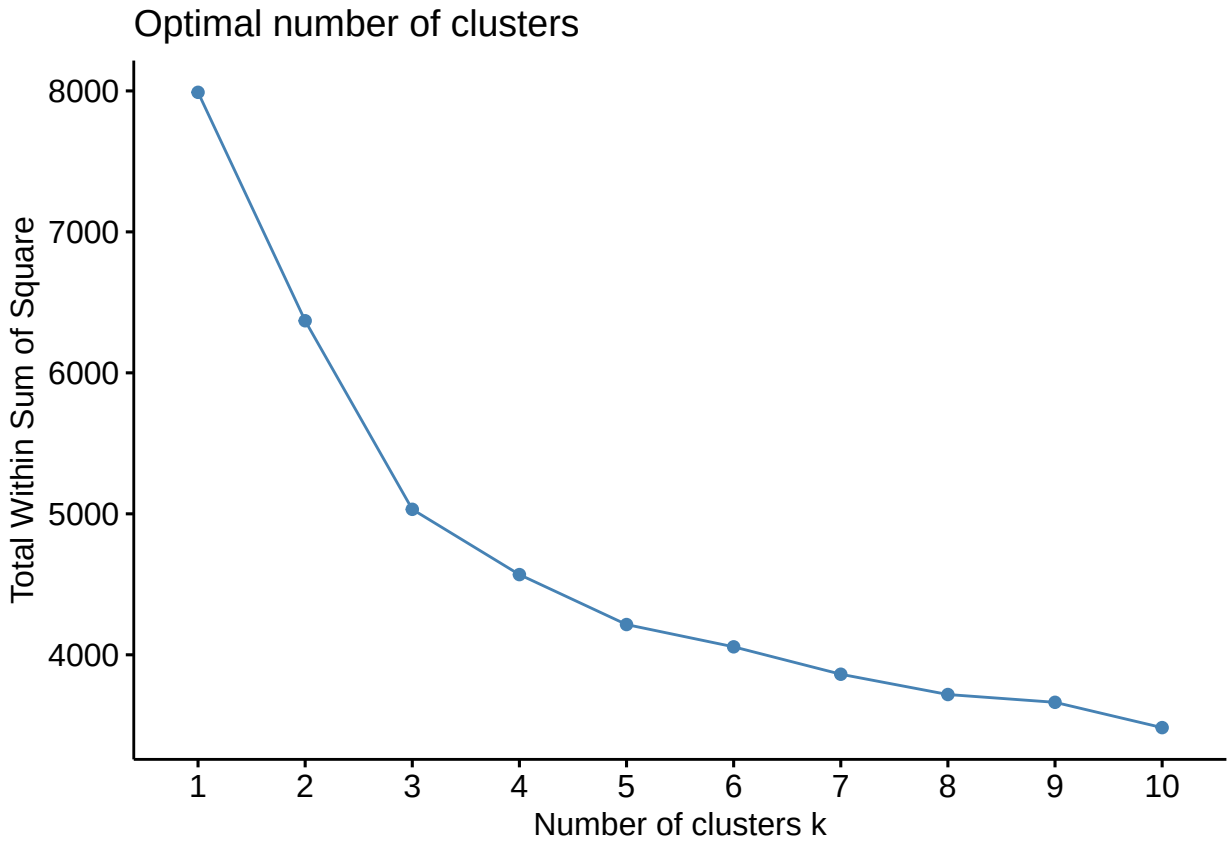
Below are the methods to determine the optimal number of clusters:

Elbow method

Silhouette method

Let us use an "elbow chart" to determine k

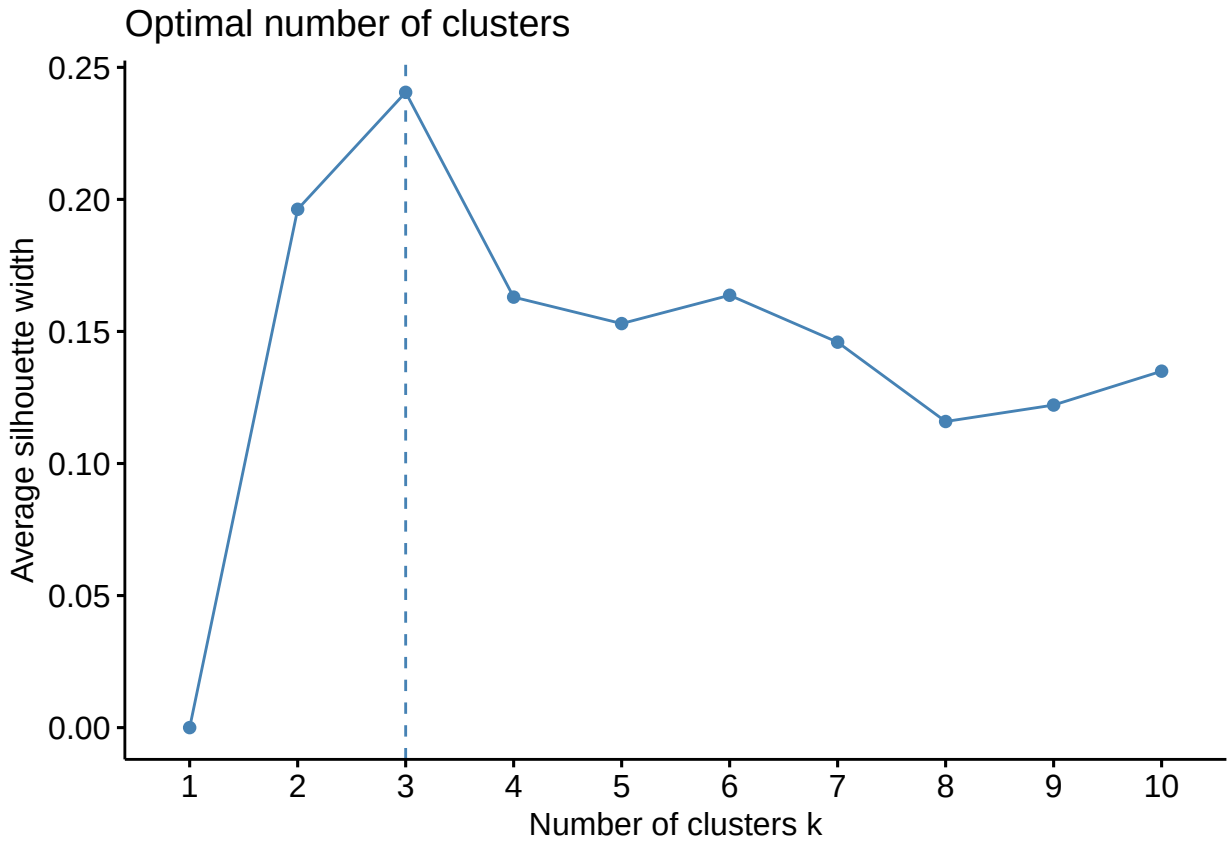
```
set.seed(123)
fviz_nbclust(ScaledData, kmeans, method = "wss")
```



The chart shows that the elbow point 3 provides the best value for k. While WSS will continue to drop for larger values of k, we have to make the trade off between over fitting, i.e., a model fitting both noise and signal, to a model having bias. Here, the elbow point provides that compromise where WSS, while still decreasing beyond k = 3, decreases at a much smaller rate. In other words, adding more clusters beyond 3 brings less improvement to cluster homogeneity.

Now, Let us also apply the Silhouette Method to determine the number of clusters

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



Again, we see that 3 is the ideal number of clusters. Here we look for large values for the Silhouette Width (Y Axis)

Run the Kmeans algorithm for clustering

```
#We will choose the value of k = 3 as observed in the above methods, number of restarts =15
k3 <- kmeans(ScaledData, centers = 3, nstart = 15)
#output the centers
k3$centers
```

```
##   X..appli..rec.d X..appl..accepted X..new.stud..enrolled
## 1     -0.35953828       -0.34918455            -0.3171053
## 2      0.05140256       -0.04367128            -0.1683551
## 3      1.98179657        2.22992267             2.4447222
##   X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                -0.5020886               -0.5128195      -0.2952142
## 2                 0.8795798                0.8620961      -0.2324464
## 3                 0.1334215                0.2545856       2.5228452
##   X..PT.undergrad in.state.tuition out.of.state.tuition       room      board
## 1      -0.1217682       -0.4036544           -0.5263964 -0.3588740 -0.3938990
## 2      -0.3130216        1.0620416            1.1158839  0.6698444  0.7756859
## 3       1.7486849       -1.0500277           -0.4918168 -0.0388330 -0.1745795
##     add..fees estim..book.costs estim..personal.. X..fac..w.PHD
## 1 -0.05832646       -0.06621454        0.05935933    -0.5322257
## 2 -0.04496556        0.07122705       -0.39665857     0.7659627
## 3  0.49531762        0.16358567        0.93858632     0.6840794
```

```
##   stud..fac..ratio Graduation.rate
## 1        0.2810858      -0.4171456
## 2       -0.7036167       0.8426062
## 3        0.6139980      -0.2538234
```

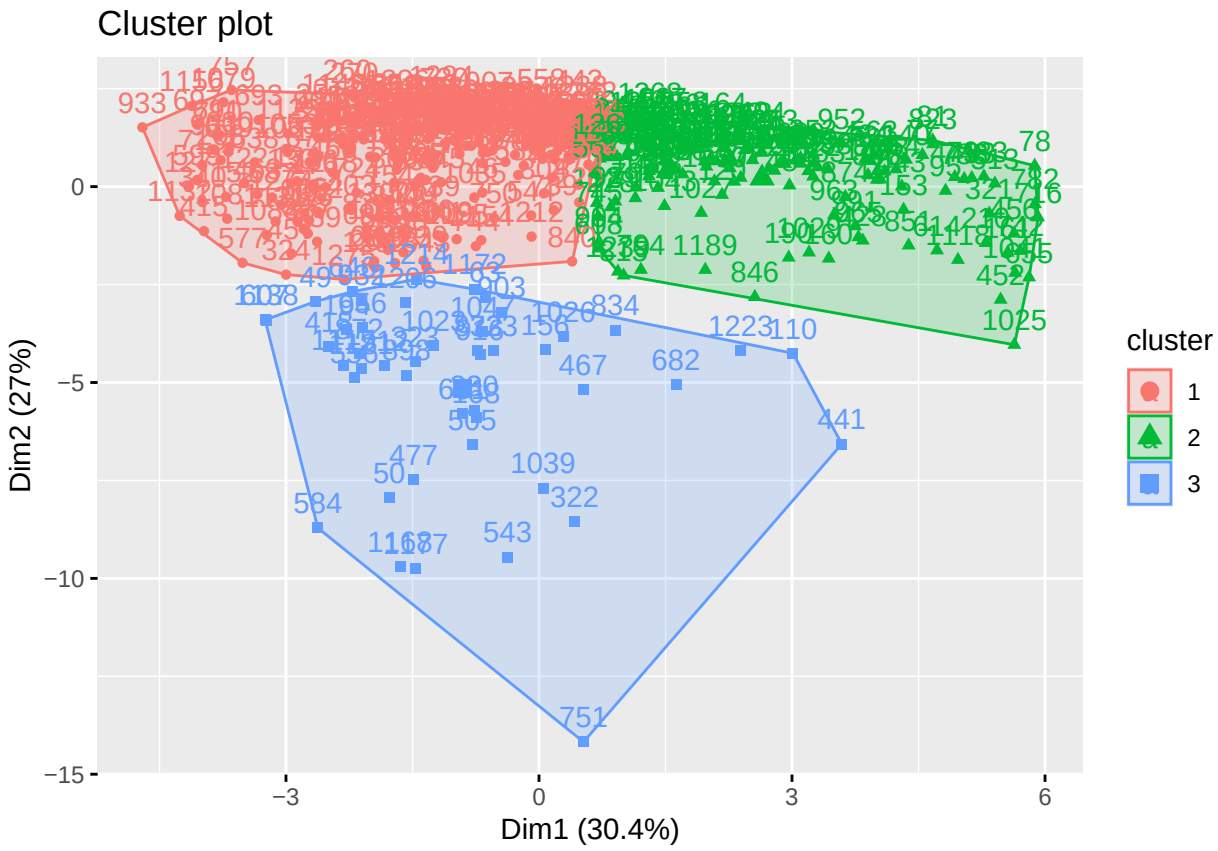
```
#No. of Universities in each cluster i.e. the size of the clusters
k3$size
```

```
## [1] 275 150  46
```

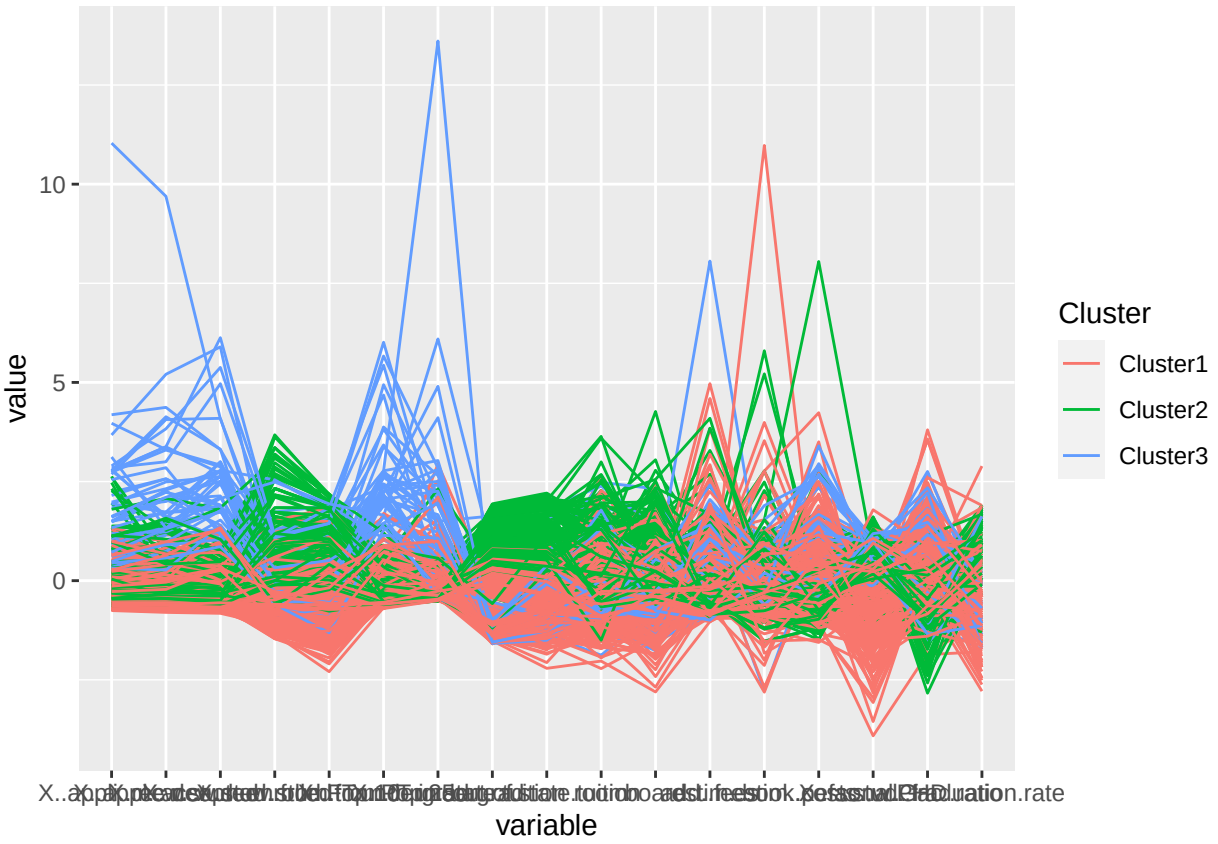
```
#Identify the cluster of the 325th observation as an example
k3$cluster[325]
```

```
## 911
##   2
```

```
#Visualize the output
fviz_cluster(k3, data = ScaledData)
```




```
#Add the cluster column in the University Data set
UniData1$Cluster = k3$cluster
# Label the Clusters
UniData1$Cluster <- factor(UniData1$Cluster,levels = c(1,2,3),labels = c("Cluster1","Cluster2" , "Cluste
#Plot the graph cluster wise
ggparcoord(UniData1, column= 1:17, groupColumn = "Cluster")
```



There are 3 clusters formed with size 275, 150 and 46.

Based on the graph plotted, we can identify some cluster groupings here:

1) Orange samples (cluster 1) with a high proportion of student faculty ratio and lower number of faculties with PHD.

2) Green samples (cluster 2), some of which have a high number of in state ans out of state tution and low portion of student faculty ratio.

3) Blue samples (cluster 3) with high number of students enrolled and Full time graduates and low in state tution

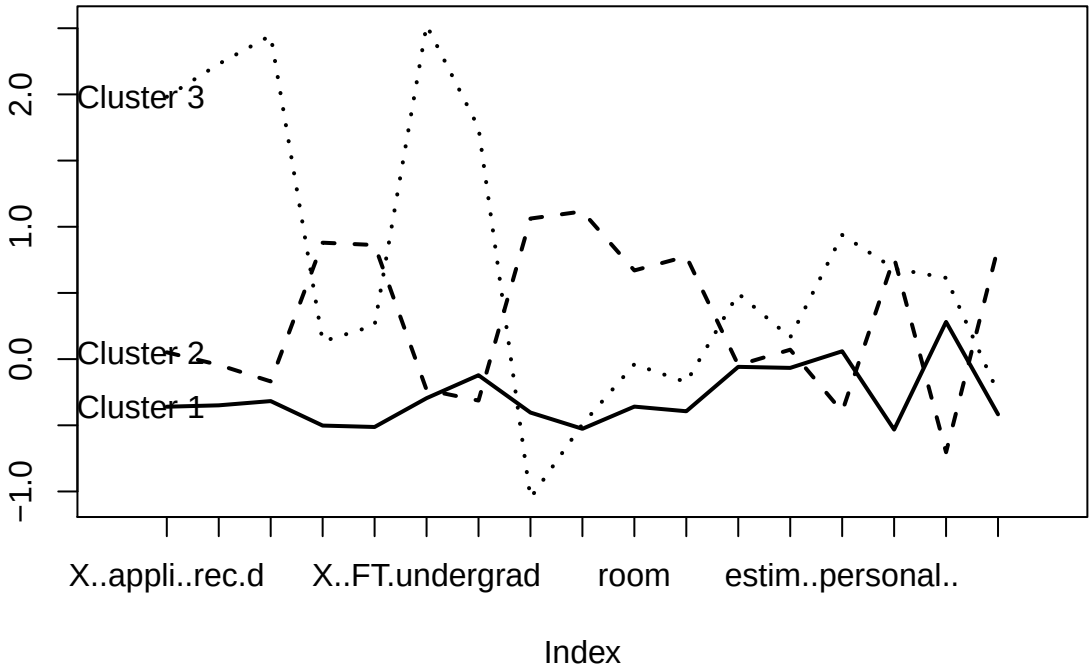
c)Compare the summary statistics for each cluster and describe each cluster in this context

```
#Use comparegroups function to compare the data set cluster wise
comparegroups.main = compareGroups(formula = Cluster~., data = UniData1[,c(1:18)])
comparegroups.main = createTable(x= comparegroups.main,show.all = T)
#View the descriptive summary by Cluster
comparegroups.main
```

```
##
## --------Summary descriptives table by 'Cluster'---------
##
## _____________________________________________________________________________
##                              [ALL]      Cluster1     Cluster2     Cluster3   p.overall
```

```
##                                N=471         N=275         N=150          N=46
## ------------------------------------------------------------------------------------
## X..appli..rec.d              3147 (4073)   1683 (1691)   3357 (2935)   11219 (6890)   <0.001
## X..appl..accepted            2063 (2504)   1189 (1032)   1954 (1398)   7646 (3992)    <0.001
## X..new.stud..enrolled         781 (916)     490 (422)     627 (433)    3019 (1155)    <0.001
## X..new.stud..from.top.10.    28.0 (18.5)   18.7 (10.1)   44.3 (19.8)   30.5 (15.3)    <0.001
## X..new.stud..from.top.25.    55.7 (20.3)   45.2 (15.5)   73.2 (16.0)   60.8 (17.2)    <0.001
## X..FT.undergrad              3563 (4669)   2185 (2155)   2478 (1813)   15343 (5582)   <0.001
## X..PT.undergrad               797 (1546)    609 (765)     314 (529)    3501 (3464)    <0.001
## in.state.tuition             9407 (5517)   7180 (3876)   15266 (3197)  3614 (3676)    <0.001
## out.of.state.tuition         10575 (4312)  8306 (2668)   15386 (2951)  8455 (2959)    <0.001
## room                          2221 (713)    1965 (563)    2699 (718)    2193 (720)     <0.001
## board                         2122 (567)    1899 (473)    2562 (500)    2023 (456)     <0.001
## add..fees                     379 (356)     358 (335)     363 (318)     555 (519)       0.002
## estim..book.costs             549 (163)     538 (171)     560 (159)     575 (119)       0.202
## estim..personal..             1312 (682)    1352 (612)    1041 (639)    1952 (745)     <0.001
## X..fac..w.PHD                73.2 (16.7)   64.3 (15.4)   86.0 (8.89)   84.6 (6.23)    <0.001
## stud..fac..ratio             14.0 (3.90)   15.1 (3.54)   11.2 (2.84)   16.4 (4.18)    <0.001
## Graduation.rate              65.6 (18.1)   58.0 (16.2)   80.9 (12.1)   61.0 (14.6)    <0.001
## ------------------------------------------------------------------------------------
```

```r
#Plot the summary statistics for each cluster
#Plot an empty scatter plot
plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(k3$centers), max(k3$centers)), xlim = c(0, 18
#Label x-axes
axis(1, at = c(1:17), labels = colnames(k3$centers))
#Plot centroids
for (i in c(1:3))
lines(k3$centers[i,], lty = i, lwd = 2, col = ifelse(i %in% c(1, 2, 3), "black", "dark grey"))
#Name clusters
text(x = 0.5, y = k3$centers[, 1], labels = paste("Cluster", c(1:3)))
```



We can now more clearly see the variation across the variables for each of the clusters found by the k-means algorithm.

The graph and the comparison table interprets that most of the variables are high in cluster 3 and 2 and none of the variables is high in Cluster 1.

d)Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information?

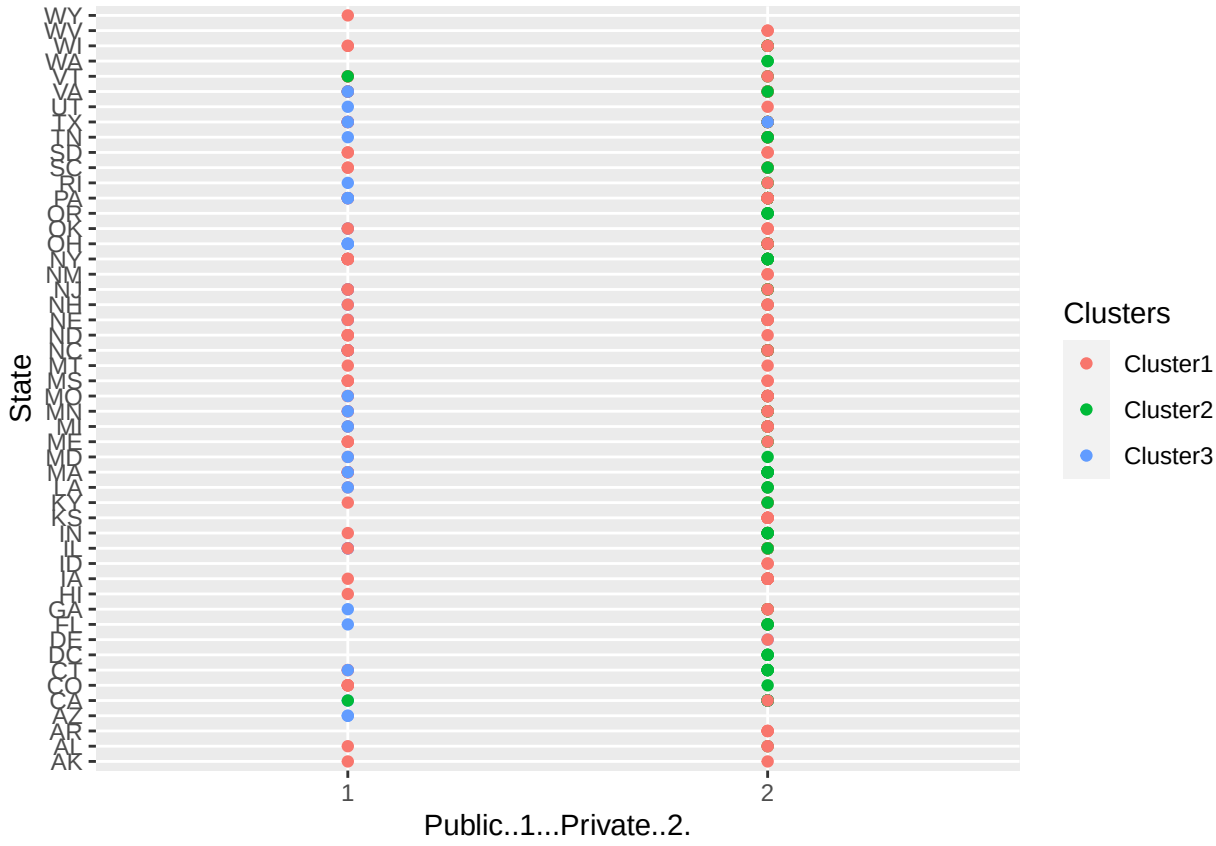
```
#Combine the University name, State and Private/Public columns along with the clusters
CombinedData <- cbind(UniData[,c(1:3)],k3$cluster)
#Label the column names
colnames(CombinedData) <- c("College.Name","State","Public..1...Private..2.","Clusters")
#Label the clusters with names like Cluster1, Cluster2, etc.
CombinedData$Public..1...Private..2. <- factor(CombinedData$Public..1...Private..2.)
CombinedData$Clusters <- factor(CombinedData$Clusters,levels = c(1,2,3),labels = c("Cluster1","Cluster2
#See the first few rows of the Combined data
head(CombinedData)
```

```
##                           College.Name State Public..1...Private..2. Clusters
## 1            Alaska Pacific University    AK                       2 Cluster1
## 3       University of Alaska Southeast    AK                       1 Cluster1
## 10         Birmingham-Southern College    AL                       2 Cluster2
## 12                   Huntingdon College    AL                       2 Cluster1
## 22                    Talladega College    AL                       2 Cluster1
## 26 University of Alabama at Birmingham    AL                       1 Cluster1
```

```
#Find all the Public Universities in each cluster
PubUni <- CombinedData %>% group_by(Clusters) %>% filter(Public..1...Private..2. ==1) %>%summarise(Publi
#Find all the Private Universities in each cluster
PrivUni <- CombinedData %>% group_by(Clusters) %>% filter(Public..1...Private..2. ==2) %>% summarise(Pri
#Combine the output
PubPriv <- cbind(PubUni,PrivUni[,2])
#Display the number of private and public universities in each cluster
PubPriv
```

```
##   Clusters PublicUniverity PrivateUniverity
## 1 Cluster1              84              191
## 2 Cluster2               3              147
## 3 Cluster3              41                5
```

```
#Plot the graph to show the number of public and private universities in each State for every cluster
ggplot(CombinedData, aes(x=Public..1...Private..2., y=State, color=Clusters)) + geom_point()
```


State
WY
WV
WI
WA
VT
VA
UT
TX
TN
SD
SC
RI
PA
OR
OK
OH
NY
NM
NJ
NH
NE
ND
NC
MT
MS
MO
MN
MI
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GA
FL
DE
DC
CT
CO
CA
AZ
AR
AL
AK
1
2
Public..1...Private..2.
Clusters
Cluster1
Cluster2
Cluster3

From the plot and from the table, it can be determined that:

The Cluster 1 has more portion of private Universities and less portion of public Universities.

The Cluster 2 has more Private Universities.

And the Cluster 3 has more Public Universities.

e)What other external information can explain the contents of some or all of these clusters?

The external factors that can impact the contents of the clusters could be following:

1) Climate : Climatic conditions of the University location.

2) Safety : Schools that have high levels of violence and poor student-teacher relations are considered not safe.

3) Recognization: In order to promote success, success needs to be recognized.

4) Enviroment: Interactions between adults and students, environmental factors, academic performance and feelings of trust and respect among educational stakeholders.

f)Consider Tufts University, which is missing some information. Compute the Euclidean distance of this record from each of the clusters that you found above (using only the measurements that you have). Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements.

Apparently, in clustering in which the distance measure is Euclidean distance, the data must be first normalized or standardized to prevent the covariate with the highest variance from driving the clustering. Why is this?

It depends on the data. And actually it has nothing to do with clustering, but with the distance function. The problem is when we have mixed attributes. For example, we have data on persons. Weight in grams and shoe size. Shoe sizes differ very little, while the differences in body mass (in grams) are much much larger. Similarly, we have this University data set. We just cannot compare 1 g and 1 shoe size difference.Usually in these cases, Euclidean distance just does not make sense. But it may still work, in many situations if we normalize our data.

```
#Read the Tufts University Data from the University data
TuftUni <- UniversityData[UniversityData$College.Name=="Tufts University",]
#Save the University data (remove categorical values and missing values)
UniData2 <- UniData[,-c(1,2,3)]
#Scale the University data and The tufts University data
norm.values <- preProcess(UniData2,method=c("center","scale"))
UniData2 <- predict(norm.values,UniData2)
TuftUni <- predict(norm.values,TuftUni)
#View the Tufts University data
```

```
TuftUni
```

```
##         College.Name State Public..1...Private..2. X..appli..rec.d
## 476 Tufts University    MA                       2        1.096623
##     X..appl..accepted X..new.stud..enrolled X..new.stud..from.top.10.
## 476         0.6158933             0.4633898                  1.730988
##     X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad in.state.tuition
## 476                  1.690004       0.2216773              NA         1.866005
##     out.of.state.tuition     room    board add..fees estim..book.costs
## 476             2.116543 1.145408 1.425498 0.3483966         0.3138547
##     estim..personal.. X..fac..w.PHD stud..fac..ratio Graduation.rate
## 476        -0.5630888       1.54761       -0.9394124        1.456852
```

```
# Compute the Euclidean Distance
dist(rbind(TuftUni[, -c(1,2,3)], k3$centers[1,]))
```

```
##        476
## 2 6.640413
```

```
dist(rbind(TuftUni[, -c(1,2,3)], k3$centers[2,]))
```

```
##       476
## 2 2.75131
```

```
dist(rbind(TuftUni[, -c(1,2,3)], k3$centers[3,]))
```

```
##        476
## 2 6.905137
```

```
#Cluster 2 is closest
#Impute the missing value
TuftUni$X..PT.undergrad <- k3$centers[2,7]
TuftUni
```

```
##         College.Name State Public..1...Private..2. X..appli..rec.d
## 476 Tufts University    MA                       2        1.096623
##     X..appl..accepted X..new.stud..enrolled X..new.stud..from.top.10.
## 476         0.6158933             0.4633898                  1.730988
##     X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad in.state.tuition
## 476                  1.690004       0.2216773      -0.3130216         1.866005
##     out.of.state.tuition     room    board add..fees estim..book.costs
## 476             2.116543 1.145408 1.425498 0.3483966         0.3138547
##     estim..personal.. X..fac..w.PHD stud..fac..ratio Graduation.rate
## 476        -0.5630888       1.54761       -0.9394124        1.456852
```

```
####
```

The tufts University is closest to the Cluster 2. The mean value of the column X..PT.undergrad for the cluster 2 is imputed to the missing value of that column for the Tuft University data.