K-MEANS

rajendra

10/31/2019

```
library(readr)
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(ISLR)
library(tidyverse)

## -- Attaching packages ---------------------------------------- tidyver
se 1.2.1 --

## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v purrr   0.3.2     v forcats 0.4.0

## -- Conflicts ------------------------------------------------ tidyverse_con
flicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()

library(factoextra)

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at http
s://goo.gl/13EFCZ

set.seed(15)
```

Qa :- Eliminating null values from the dataset

```
# Importing dataset in r
Universities <- read_csv("C:/Users/rajendra/Downloads/Universities.csv")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `College Name` = col_character(),
##   State = col_character()
## )

## See spec(...) for full column specifications.
```

```
uni<-na.omit(Universities)
summary(uni)
```

```
##  College Name          State           Public (1)/ Private (2)
##  Length:471         Length:471         Min.   :1.000
##  Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :2.000
##                                        Mean   :1.728
##                                        3rd Qu.:2.000
##                                        Max.   :2.000
##  # appli. rec'd  # appl. accepted  # new stud. enrolled
##  Min.   :   77   Min.   :   61.0   Min.   :  27.0
##  1st Qu.:  802   1st Qu.:  635.5   1st Qu.: 264.0
##  Median : 1646   Median : 1227.0   Median : 443.0
##  Mean   : 3147   Mean   : 2063.0   Mean   : 780.7
##  3rd Qu.: 3862   3rd Qu.: 2456.0   3rd Qu.: 896.5
##  Max.   :48094   Max.   :26330.0   Max.   :6392.0
##  % new stud. from top 10% % new stud. from top 25% # FT undergrad
##  Min.   : 1.00            Min.   :  9.00           Min.   :  249
##  1st Qu.:15.00            1st Qu.: 40.00           1st Qu.: 1018
##  Median :23.00            Median : 54.00           Median : 1715
##  Mean   :28.01            Mean   : 55.65           Mean   : 3563
##  3rd Qu.:36.00            3rd Qu.: 69.00           3rd Qu.: 4056
##  Max.   :96.00            Max.   :100.00           Max.   :31643
##  # PT undergrad    in-state tuition out-of-state tuition      room
##  Min.   :    1.0   Min.   :  608    Min.   : 1044        Min.   : 640
##  1st Qu.:   81.5   1st Qu.: 3650    1st Qu.: 7290        1st Qu.:1740
##  Median :  299.0   Median : 9858    Median :10100        Median :2090
##  Mean   :  797.5   Mean   : 9407    Mean   :10575        Mean   :2221
##  3rd Qu.:  869.0   3rd Qu.:13246    3rd Qu.:13286        3rd Qu.:2663
##  Max.   :21836.0   Max.   :20100    Max.   :20100        Max.   :4816
##      board        add. fees      estim. book costs estim. personal $
##  Min.   : 531   Min.   :  10.0   Min.   :  90.0    Min.   : 250
##  1st Qu.:1750   1st Qu.: 137.5   1st Qu.: 500.0    1st Qu.: 850
##  Median :2082   Median : 280.0   Median : 500.0    Median :1200
##  Mean   :2122   Mean   : 379.0   Mean   : 548.8    Mean   :1312
##  3rd Qu.:2420   3rd Qu.: 486.0   3rd Qu.: 600.0    3rd Qu.:1600
##  Max.   :4541   Max.   :3247.0   Max.   :2340.0    Max.   :6800
##   % fac. w/PHD    stud./fac. ratio 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
```

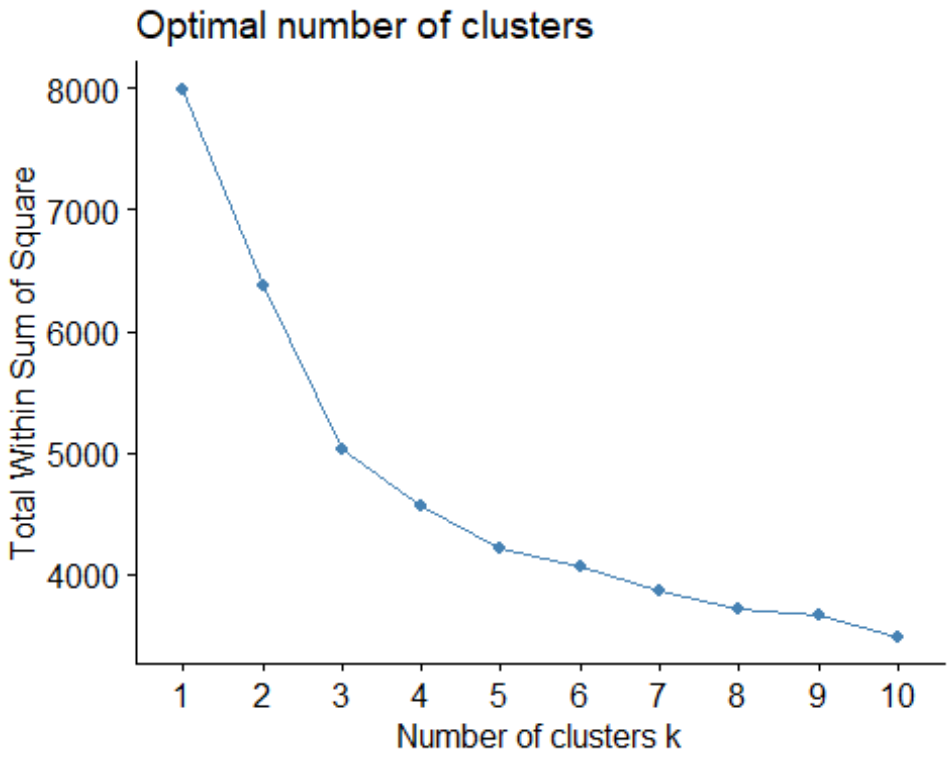
Qb :- Obtaining the opitmal value of k and number of clusters

```
# Removing Categorical values from the dataset
```

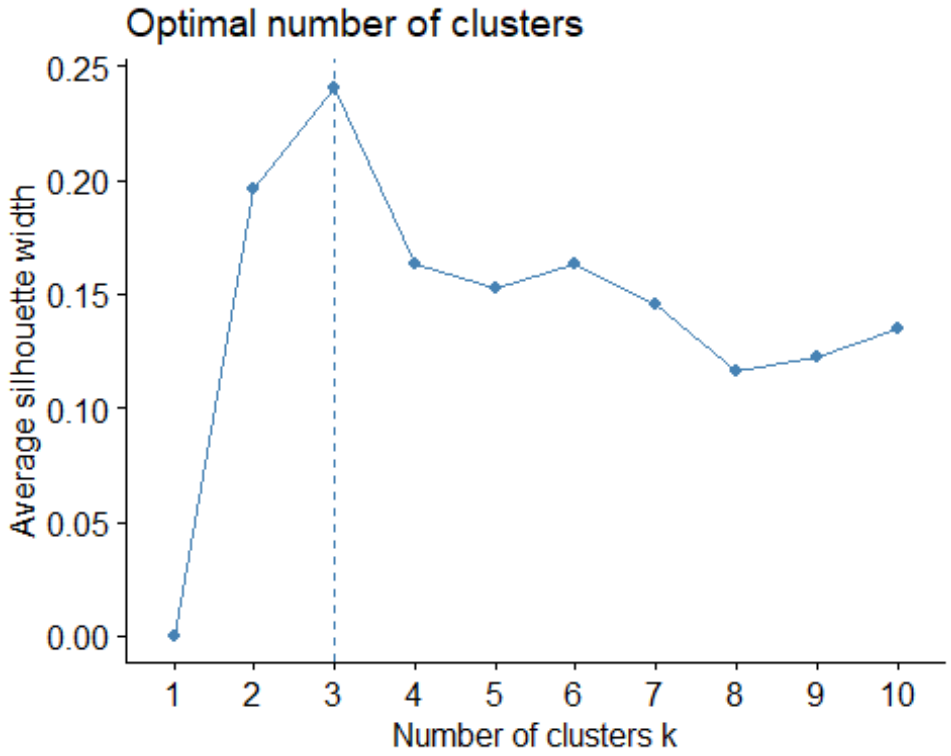
```r
a<-uni[,c(-1,-2,-3)]

#Scaling  the data frame (z-score)

b<-scale(a)

summary(b)
```

```
##  # appli. rec'd    # appl. accepted  # 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
##  % new stud. from top 10% % new stud. from top 25% # 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
##  # PT undergrad     in-state tuition   out-of-state tuition
##  Min.   :-0.51524   Min.   :-1.59488   Min.   :-2.2105
##  1st Qu.:-0.46316   1st Qu.:-1.04338   1st Qu.:-0.7619
##  Median :-0.32246   Median : 0.08182   Median :-0.1102
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.04628   3rd Qu.: 0.69594   3rd Qu.: 0.6287
##  Max.   :13.61017   Max.   : 1.93833   Max.   : 2.2091
##       room              board            add. fees       estim. book costs
##  Min.   :-2.2170   Min.   :-2.80658   Min.   :-1.0370   Min.   :-2.8114
##  1st Qu.:-0.6746   1st Qu.:-0.65614   1st Qu.:-0.6787   1st Qu.:-0.2989
##  Median :-0.1838   Median :-0.07046   Median :-0.2783   Median :-0.2989
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6196   3rd Qu.: 0.52581   3rd Qu.: 0.3006   3rd Qu.: 0.3139
##  Max.   : 3.6384   Max.   : 4.26746   Max.   : 8.0594   Max.   :10.9766
##  estim. personal $  % fac. w/PHD     stud./fac. ratio  Graduation rate
##  Min.   :-1.5574   Min.   :-3.9127   Min.   :-2.8374   Min.   :-2.7863
##  1st Qu.:-0.6775   1st Qu.:-0.6125   1st Qu.:-0.6829   1st Qu.:-0.6923
##  Median :-0.1642   Median : 0.1675   Median :-0.1443   Median : 0.0241
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4225   3rd Qu.: 0.8276   3rd Qu.: 0.6380   3rd Qu.: 0.7405
##  Max.   : 8.0488   Max.   : 1.7876   Max.   : 3.8056   Max.   : 2.8896
```

```r
#Finding the best k value using Elbow Chart

fviz_nbclust(b,kmeans,method = "wss")
```




```r
#Using silhouette method to determine k

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



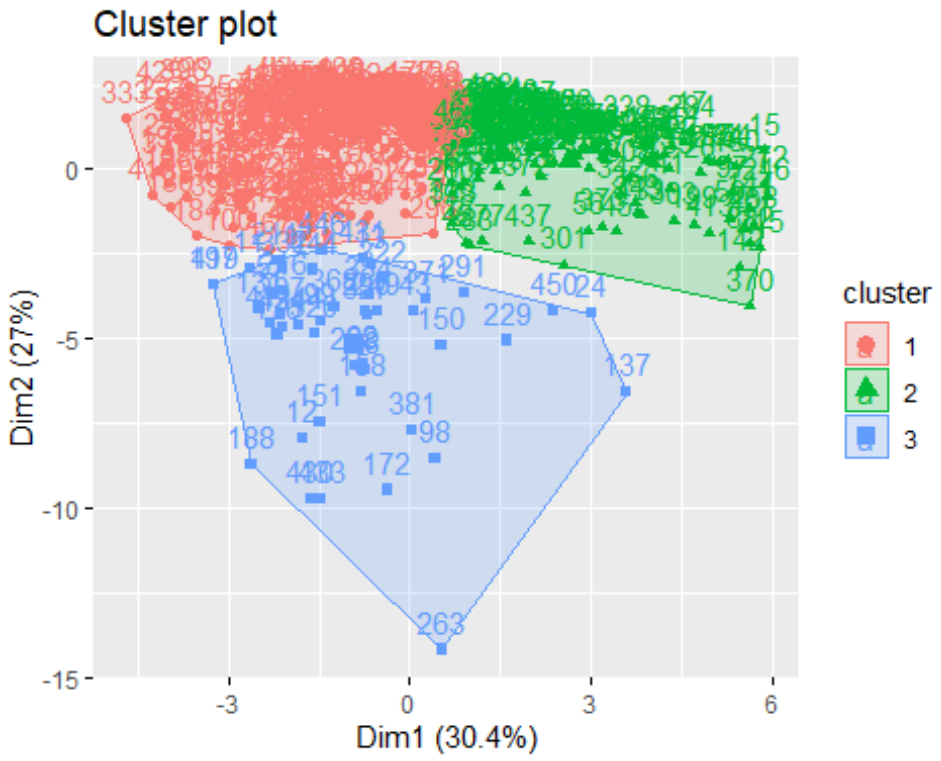
```r
#Kmeans cluster algorithm
k4 <- kmeans(b, centers = 3, nstart = 1)

k4$centers # Centroids
```

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

```r
k4$size    # Size of each cluster
```

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

```r
# Visualizing the clusters

fviz_cluster(k4,data=b)
```

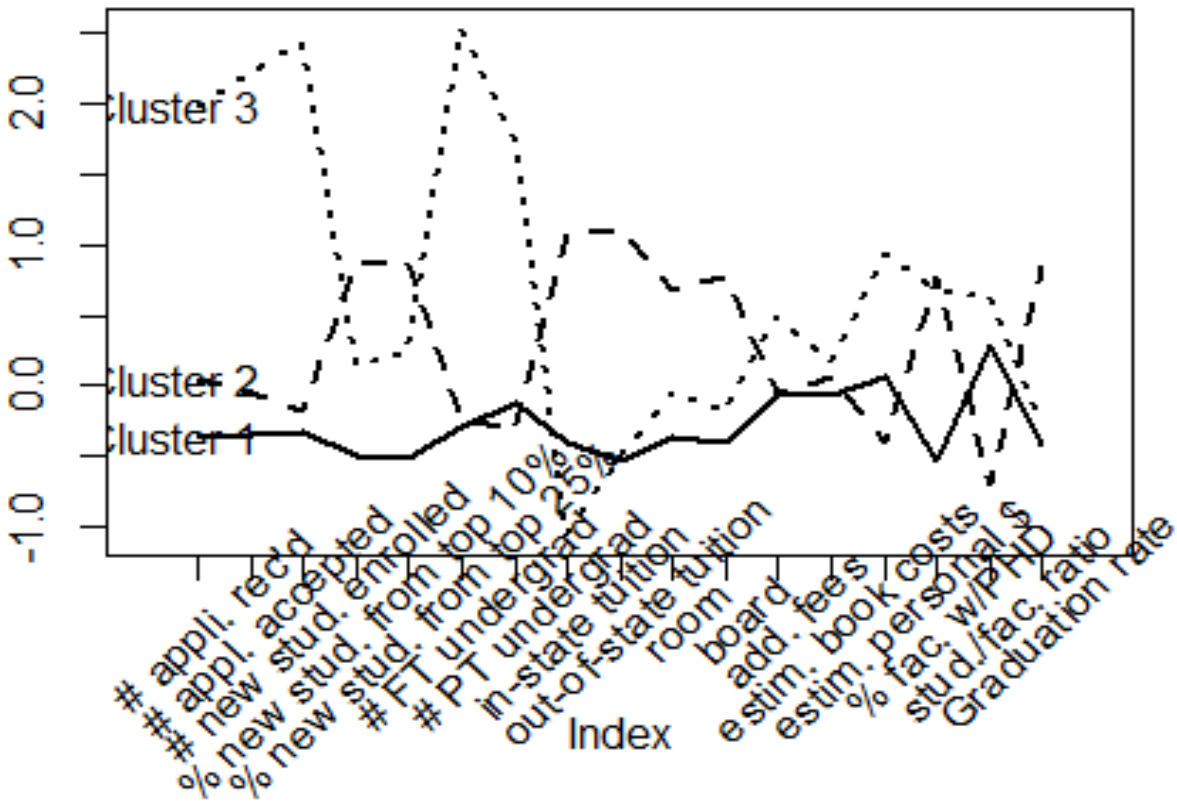


From the graph,we can say that k=3 is the best value and the dataset is divided into three different clusters

Qc :- Combining cluster information with dataset to determine the statistics of each cluster

```
com<-cbind(uni,k4$cluster)

# Assigning labels to three clusters (Tier 1, Tier 2, Tier3 Universities)

com$`k4$cluster`<-factor(com$`k4$cluster`,levels = c(1,2,3),labels = c("T3","
T1","T2"))

c1<-com[com$`k4$cluster`=="T1",]#Cluster 2

c2<-com[com$`k4$cluster`=="T2",]#Cluster 3

c3<-com[com$`k4$cluster`=="T3",]#Cluster 1



plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(k4$centers), max(k4$centers)), xlim = c(0, 18))

axis(1, at = c(1:17), labels = FALSE)
text(seq(1,17,by=1),par("usr")[3]-0.2,labels = colnames(b),srt = 45,pos =1,xp
d = TRUE) # X Measurement names
```

```
# plot centroids
for (i in c(1:3))
  lines(k4$centers[i,], lty = i, lwd = 2)

# name clusters
text(x = 0.5, y = k4$centers[, 1], labels = paste("Cluster", c(1:3))) # Clust
er Names
```



Classifying the clusters based on their mean values of respective columns

Cluster 1 Statistics

```
#From the graph shown above we can infer the characterstics of cluster 1 and
named it as Tier 3 Universities

#Tier 3 universties (Cluster 1)

#Low applications recieved
#Low applications accepted
```

```
#Minimum number of students enrolled
# Minimum number of new students from top 25%
# Minimum number of new students from top 10%
# Less number of full time under graduate
# Out of station tution fee are less expensive
# Very few number of rooms
#Low cost of books
# Number of boards are low
# Graduation rate is low
# less percent of faculty ratio wrt phd
```

Cluster 2 statistics

```
#From the graph shown above we can infer the characterstics of cluster 2 and
named it as Tier 1 Universities

# Tier 1 Universties (Cluster 2)
# Higher Graduation rate
# High percentage faculty ratio w.r.t PHD
# High in-state tutuion fee
# More number of rooms
# Number of boards are high
# More number of new students from top 25%
# More number of new students from top 10%
```

cluster 3 statistics

```
#From the graph shown above we can infer the characterstics of cluster 3 and
named it as Tier 2 Universities

#Tier 2 universties (Cluster 3)
#Maximum number of applications recieved
#Maximum number of applications accepted
#Maximum number of new students enrolled
#More number of part time graduates
#More number of full time graduates
#Huge amount of estimated book cost
#Huge amount of personal expenses
#Higher Student to faculty ratio
```

Qd :- Finding the relationship between the clusters and the categorical information?

```
head(com)
```

```
##                          College Name State Public (1)/ Private (2)
## 1           Alaska Pacific University    AK                       2
## 2      University of Alaska Southeast    AK                       1
## 3         Birmingham-Southern College    AL                       2
## 4                   Huntingdon College    AL                       2
## 5                    Talladega College    AL                       2
## 6 University of Alabama at Birmingham    AL                       1
```

```
##   # appli. rec'd # appl. accepted # new stud. enrolled
## 1            193              146                   55
## 2            146              117                   89
## 3            805              588                  287
## 4            608              520                  127
## 5           4414             1500                  335
## 6           1797             1260                  938
##   % new stud. from top 10% % new stud. from top 25% # FT undergrad
## 1                      16                      44            249
## 2                       4                      24            492
## 3                      67                      88           1376
## 4                      26                      47            538
## 5                      30                      60            908
## 6                      24                      35           6960
##   # PT undergrad in-state tuition out-of-state tuition room board
## 1            869             7560                7560 1620  2500
## 2           1849             1742                5226 2514  2250
## 3            207            11660               11660 2050  2430
## 4            126             8080                8080 1380  2540
## 5            119             5666                5666 1424  1540
## 6           4698             2220                4440 1935  3240
##   add. fees estim. book costs estim. personal $ % fac. w/PHD
## 1       130              800             1500           76
## 2        34              500             1162           39
## 3       120              400              900           74
## 4       100              500             1100           63
## 5       418             1000             1400           56
## 6       291              750             2200           96
##   stud./fac. ratio Graduation rate k4$cluster
## 1             11.9              15         T3
## 2              9.5              39         T3
## 3             14.0              72         T1
## 4             11.4              44         T3
## 5             15.5              46         T3
## 6              6.7              33         T3
```

```
head(com[com$`k4$cluster`=="T1",c(1,2,3,21)])
```

```
##                   College Name State Public (1)/ Private (2) k4$cluster
## 3  Birmingham-Southern College    AL                       2         T1
## 14   Claremont McKenna College    CA                       2         T1
## 15         Harvey Mudd College    CA                       2         T1
## 16              Pitzer College    CA                       2         T1
## 17             Scripps College    CA                       2         T1
## 18          Occidental College    CA                       2         T1
```

```
head(com[com$`k4$cluster`=="T2",c(1,2,3,21)])
```

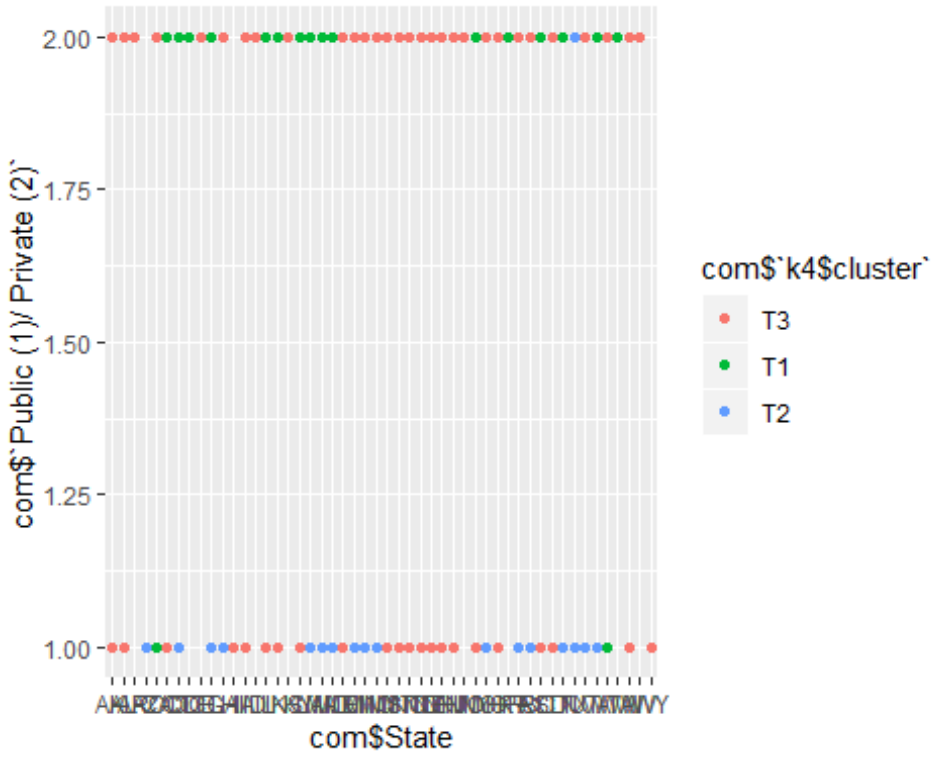
```
##                           College Name State Public (1)/ Private (2)
## 11         Northern Arizona University    AZ                       1
## 12               University of Arizona    AZ                       1
```

```
## 13     California Polytechnic-San Luis    CA                        1
## 24   University of Southern California    CA                        2
## 43 University of Connecticut at Storrs    CT                        1
## 48              University of Delaware    DE                        2
##    k4$cluster
## 11         T2
## 12         T2
## 13         T2
## 24         T2
## 43         T2
## 48         T2
```

```
head(com[com$`k4$cluster`=="T3",c(1,2,3,21)])
```

```
##                          College Name State Public (1)/ Private (2)
## 1           Alaska Pacific University    AK                        2
## 2      University of Alaska Southeast    AK                        1
## 4                  Huntingdon College    AL                        2
## 5                   Talladega College    AL                        2
## 6 University of Alabama at Birmingham    AL                        1
## 7     Arkansas College (Lyon College)    AR                        2
##   k4$cluster
## 1         T3
## 2         T3
## 4         T3
## 5         T3
## 6         T3
## 7         T3
```

```
library(ggplot2)
ggplot(com,aes(x=com$State,y=com$`Public (1)/ Private (2)`,color=com$`k4$clus
ter`))+geom_point()
```



cluster 1:-It has both public and private universities cluster 2:- maximum number of the universities are private cluster 3::-maximum number of the universities are pubic

Qe :-

```
k4

## K-means clustering with 3 clusters of sizes 275, 150, 46
##
## Cluster means:
##   # appli. rec'd # appl. accepted # new stud. enrolled
## 1    -0.35953828      -0.34918455           -0.3171053
## 2     0.05140256      -0.04367128           -0.1683551
## 3     1.98179657       2.22992267            2.4447222
##   % new stud. from top 10% % new stud. from top 25% # FT undergrad
## 1               -0.5020886               -0.5128195     -0.2952142
## 2                0.8795798                0.8620961     -0.2324464
## 3                0.1334215                0.2545856      2.5228452
##   # PT undergrad in-state tuition out-of-state tuition       room
## 1     -0.1217682       -0.4036544           -0.5263964 -0.3588740
## 2     -0.3130216        1.0620416            1.1158839  0.6698444
## 3      1.7486849       -1.0500277           -0.4918168 -0.0388330
##        board   add. fees estim. book costs estim. personal $ % fac. w/PHD
## 1 -0.3938990 -0.05832646       -0.06621454        0.05935933   -0.5322257
## 2  0.7756859 -0.04496556        0.07122705       -0.39665857    0.7659627
## 3 -0.1745795  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
##
## Clustering vector:
##   [1] 1 1 2 1 1 1 1 1 1 1 3 3 3 2 2 2 2 2 1 2 1 2 2 3 2 2 1 1 2 1 1 1 1 1
2
##  [36] 1 2 2 2 2 2 1 3 2 2 2 2 3 1 1 2 1 1 2 2 2 3 1 2 1 2 3 1 1 1 1 1 1 2
1
##  [71] 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 1 2 1 1 1 1 2 3 3 1 1 2 2 2
1
## [106] 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 3 2 2 2 2 2 3 1 2
2
## [141] 1 2 1 2 1 1 1 1 2 3 3 2 2 2 2 2 1 3 1 2 2 1 1 1 1 2 2 1 2 2 1 3 1 1
1
## [176] 3 1 2 1 1 2 2 2 1 1 1 1 3 1 1 1 1 1 1 1 1 3 2 2 1 1 1 1 1 1 1 1 1 1
1
## [211] 3 1 1 2 2 3 1 1 1 1 1 1 1 1 1 1 2 3 3 1 1 1 1 1 1 1 1 1 2 1 1 1 1 3
1
## [246] 2 1 1 3 1 1 2 1 1 1 1 1 2 1 2 1 1 3 1 1 1 2 1 1 2 2 2 2 1 2 2 2 1 2
2
## [281] 1 1 2 2 2 2 1 1 2 3 3 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 3 2 1 1 1 2 1 2
1
## [316] 2 1 1 2 3 1 3 2 1 2 1 3 1 1 1 1 3 1 1 1 2 2 2 2 2 2 2 2 2 2 1 2 1
1
## [351] 2 2 1 2 2 2 1 1 2 2 1 1 1 1 1 1 2 3 1 2 3 2 2 2 1 1 1 1 1 1 3 2 2 3
1
## [386] 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 3 2 2 1 1 1 1 1 3
1
## [421] 1 1 1 1 1 1 2 2 1 3 3 1 3 3 1 1 2 1 2 1 2 1 1 3 1 3 1 1 2 3 1 1 1 1
2
## [456] 1 2 2 2 2 1 2 2 2 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 2562.342 1424.892 1044.680
##  (between_SS / total_SS =  37.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"

# 1st cluster has 150 colleges and has a variability of 1424.8

# 2nd cluster has 46 colleges  and has a variability of 1044.68 (the members
are closer to each other in terms of distance hen compared to other clusters)

# 3rd cluster has 275 colleges  and has a variability of 2562.3
k4$withinss
```

```
## [1] 2562.342 1424.892 1044.680

# It describes the distance between centroids and mean of all the points with
in a cluster

k4$betweenss

## [1] 2958.086

# Itdescribes the distance between centriods of all the customers
```

Qf :-

```
Km<-kmeans(a,centers = 3)
b1<-mean(Km$centers[1,]) # Mean of Cluster 1
b2<-mean(Km$centers[2,]) # Mean of cluster 2
b3<-mean(Km$centers[3,]) # Mean of cluster 3
a1<-Universities[Universities$`College Name`=="Tufts University",]
View(a1)
a2<-apply(a1[,-c(1:3,10)],1,mean) # Mean of record
dist(rbind(a2,b1)) # Euclideam distance betweewn  cluster 1 mean and Tufts un
iversity data

##          a2
## b1 713.8496

dist(rbind(a2,b2))

##          a2
## b2 1314.998

dist(rbind(a2,b3))

##          a2
## b3 2452.064

a1$`# PT undergrad`<-1596.33 # From the above, Mean value which is near to cl
uster 2. Hence replacing the missing value with mean value
uni2<-rbind(uni,a1)
View(uni2)
uni2_z<-scale(uni2[,-c(1:3)])
uni2_cluster<-kmeans(uni2_z,3)
uni2<-cbind(uni2,uni2_cluster$cluster)
uni2[472,] # From the model, this uniersity falls under Cluster 2

##         College Name State Public (1)/ Private (2) # appli. rec'd
## 472 Tufts University    MA                       2           7614
##     # appl. accepted # new stud. enrolled % new stud. from top 10%
## 472             3605                 1205                       60
##     % new stud. from top 25% # FT undergrad # PT undergrad
## 472                       90           4598        1596.33
##     in-state tuition out-of-state tuition room board add. fees
```

```
## 472            19701                19701 3038  2930       503
##     estim. book costs estim. personal $ % fac. w/PHD stud./fac. ratio
## 472               600               928           99             10.3
##     Graduation rate uni2_cluster$cluster
## 472              92                    2
```

Tufts University belongs to cluster 2 from the above information and it belongs to Tier 1 University