R Notebook

```
library(readr)
Universities <- read_csv("C:/Trading detail/STUDY/01_MSBA/02 MSBA ML/03/Universities.csv")
```

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

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

```
View(Universities)


library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------
------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

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

```
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFC
Z
```

```
library(ISLR)
library(caret)
```

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

```
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
## 
##     lift
```

```
summary(Universities)
```

```
##  College Name          State           Public (1)/ Private (2)
##  Length:1302        Length:1302        Min.   :1.000          
##  Class :character   Class :character   1st Qu.:1.000          
##  Mode  :character   Mode  :character   Median :2.000          
##                                        Mean   :1.639          
##                                        3rd Qu.:2.000          
##                                        Max.   :2.000          
##                                                               
##  # appli. rec'd    # appl. accepted  # new stud. enrolled
##  Min.   :   35.0   Min.   :   35.0   Min.   :  18.0      
##  1st Qu.:  695.8   1st Qu.:  554.5   1st Qu.: 236.0      
##  Median : 1470.0   Median : 1095.0   Median : 447.0      
##  Mean   : 2752.1   Mean   : 1870.7   Mean   : 778.9      
##  3rd Qu.: 3314.2   3rd Qu.: 2303.0   3rd Qu.: 984.0      
##  Max.   :48094.0   Max.   :26330.0   Max.   :7425.0      
##  NA's   :10        NA's   :11        NA's   :5           
##  % new stud. from top 10% % new stud. from top 25% # FT undergrad
##  Min.   : 1.00            Min.   :  6.00           Min.   :   59 
##  1st Qu.:13.00            1st Qu.: 36.75           1st Qu.:  966 
##  Median :21.00            Median : 50.00           Median : 1812 
##  Mean   :25.67            Mean   : 52.35           Mean   : 3693 
##  3rd Qu.:32.00            3rd Qu.: 66.00           3rd Qu.: 4540 
##  Max.   :98.00            Max.   :100.00           Max.   :31643 
##  NA's   :235              NA's   :202              NA's   :3     
##  # PT undergrad    in-state tuition out-of-state tuition      room     
##  Min.   :    1.0   Min.   :  480    Min.   : 1044        Min.   : 500  
##  1st Qu.:  131.2   1st Qu.: 2580    1st Qu.: 6111        1st Qu.:1710  
##  Median :  472.0   Median : 8050    Median : 8670        Median :2200  
##  Mean   : 1081.5   Mean   : 7897    Mean   : 9277        Mean   :2515  
##  3rd Qu.: 1313.0   3rd Qu.:11600    3rd Qu.:11659        3rd Qu.:3040  
##  Max.   :21836.0   Max.   :25750    Max.   :25750        Max.   :7400  
##  NA's   :32        NA's   :30       NA's   :20           NA's   :321   
##      board        add. fees      estim. book costs estim. personal $
##  Min.   : 531   Min.   :   9.0   Min.   :  90      Min.   :  75     
##  1st Qu.:1619   1st Qu.: 130.0   1st Qu.: 480      1st Qu.: 900     
##  Median :1980   Median : 264.5   Median : 502      Median :1250     
##  Mean   :2061   Mean   : 392.0   Mean   : 550      Mean   :1389     
##  3rd Qu.:2402   3rd Qu.: 480.0   3rd Qu.: 600      3rd Qu.:1794     
##  Max.   :6250   Max.   :4374.0   Max.   :2340      Max.   :6900     
##  NA's   :498    NA's   :274      NA's   :48        NA's   :181      
##   % fac. w/PHD    stud./fac. ratio Graduation rate 
##  Min.   :  8.00   Min.   : 2.30    Min.   :  8.00  
##  1st Qu.: 57.00   1st Qu.:11.80    1st Qu.: 47.00  
##  Median : 71.00   Median :14.30    Median : 60.00  
##  Mean   : 68.65   Mean   :14.86    Mean   : 60.41  
##  3rd Qu.: 82.00   3rd Qu.:17.60    3rd Qu.: 74.00  
##  Max.   :105.00   Max.   :91.80    Max.   :118.00  
##  NA's   :32       NA's   :2        NA's   :98      
```

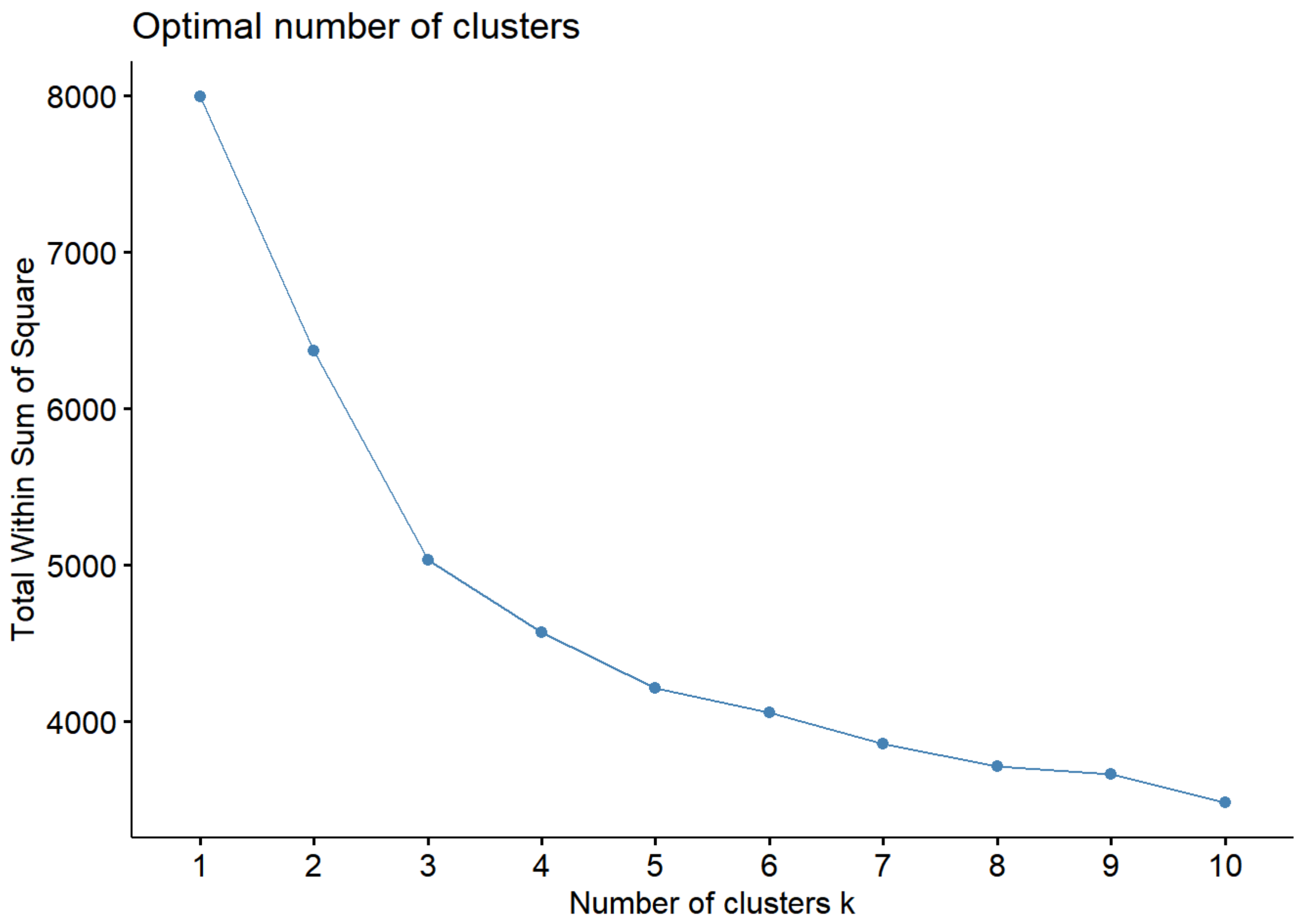
a. Remove all records with missing measurements from the dataset.

```
# remove na in r - remove rows - na.omit function / option

set.seed(123)
univ <- na.omit(Universities)


univ<-univ[,c(-1,-2,-3)]
summary(univ)
```

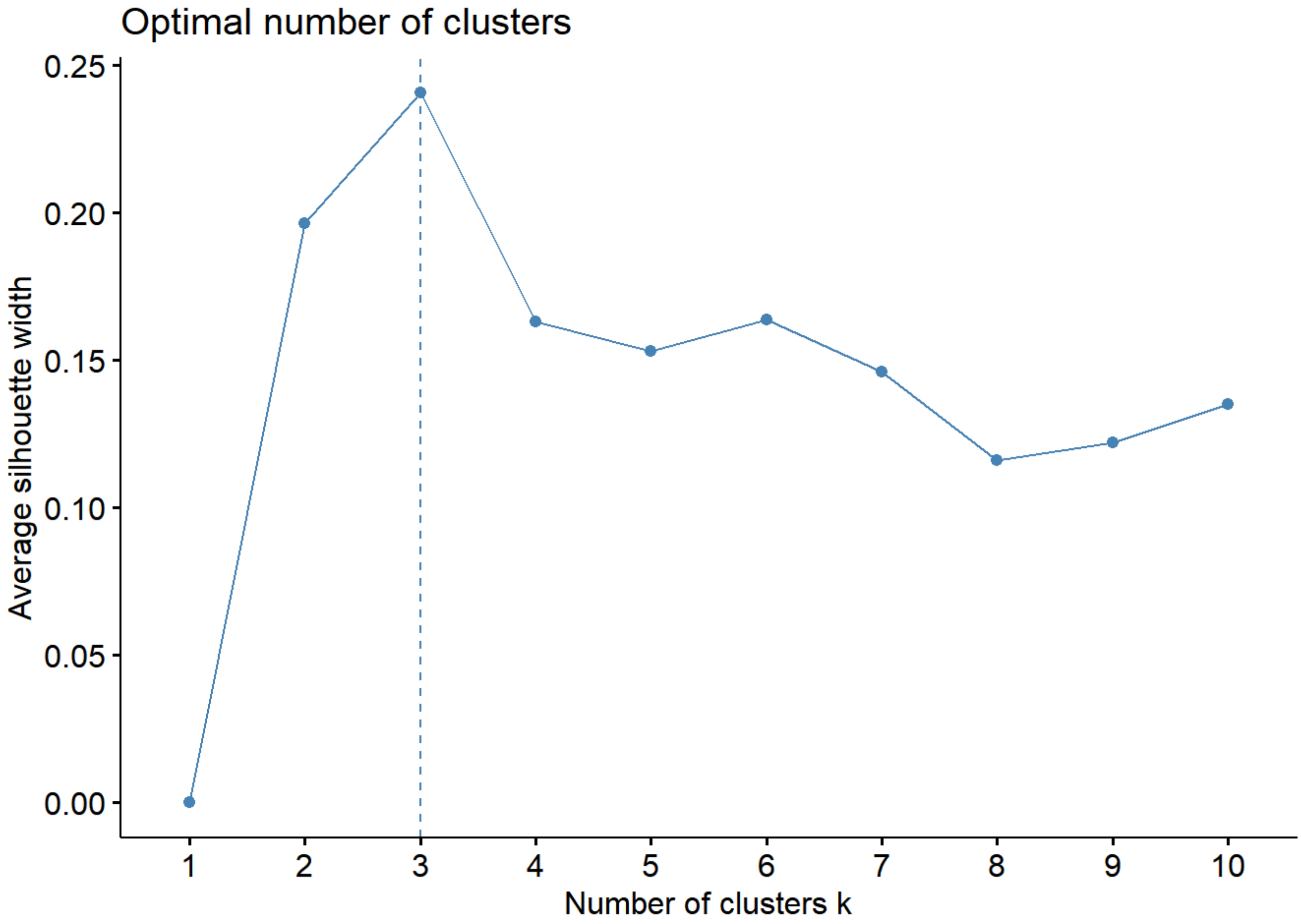
```
##  # appli. rec'd  # appl. accepted  # new stud. enrolled
##  Min.   :   77   Min.   :   61.0   Min.   :  27.0
##  1st Qu.:  802   1st Qu.:  635.5   1st Qu.: 264.0
##  Median : 1646   Median : 1227.0   Median : 443.0
##  Mean   : 3147   Mean   : 2063.0   Mean   : 780.7
##  3rd Qu.: 3862   3rd Qu.: 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
```

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?

```
##Scaling the data frame (z-score)
univ <- scale(univ)
univ<- as.data.frame(univ)

### toFind best value of k by for total within sum of square
fviz_nbclust(univ, kmeans, method = "wss")   ###by applying 2 method "Wss" The chart shows th
at the elbow point 3 provides the best value for k.
```



```
### toFind best value of k by average silhouette width

fviz_nbclust(univ, kmeans, method = "silhouette")  ### applying "silhouette" method we see th
at 3 is the ideal number of clusters.
```



```
fitk <- kmeans(univ,centers = 3)

fitk$size ## shows no. of observation in each cluster
```

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

```
fitk  ## summary of each cluster
```

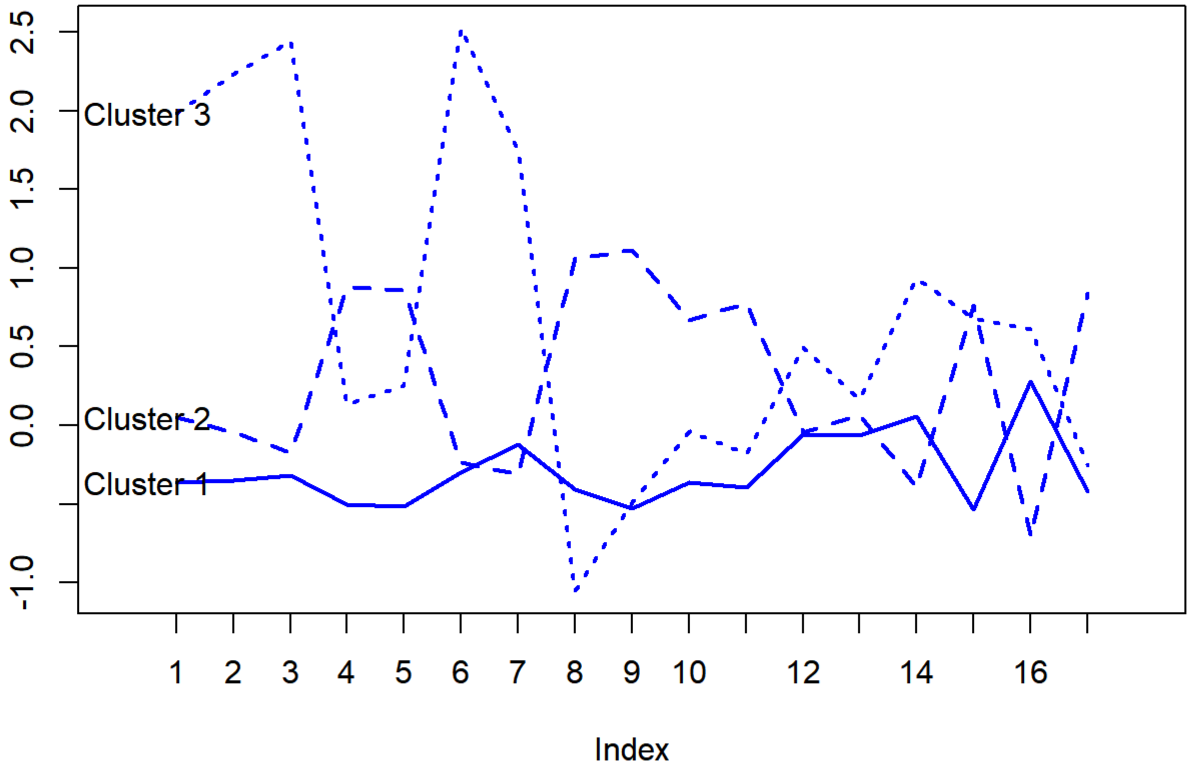
```
## K-means clustering with 3 clusters of sizes 275, 150, 46
## 
## Cluster means:
##   # appli. rec'd # appl. accepted # new stud. enrolled
## 1    -0.35953828      -0.34918455           -0.3171053
## 2     0.05140256      -0.04367128           -0.1683551
## 3     1.98179657       2.22992267            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 2 1
##  [71] 1 1 2 1 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 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 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"
```

c. Compare the summary statistics for each cluster and describe each cluster in this context (e.g., “Universities with high tuition, low acceptance rate…”).

```
fitk$centers  ## means of each cluster
```

```
##   # 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
```

```
##from means of the cluster we can conclude following information about different cluster

# plot an empty scatter plot

plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(fitk$centers), max(fitk$centers)), xlim = c(0, 18))
# label x-axes
axis(1, at = c(1:17), labels = names(fitk$centers))
# plot centroids
for (i in c(1:3))
  lines(fitk$centers[i,], lty = i, lwd = 2, col = ifelse(i %in% c(1, 2, 3),
                                                        "blue"))
# name clusters
text(x = 0.5, y = fitk$centers[, 1], labels = paste("Cluster", c(1:3)))
```



CLUSTER 3 : has received high no. of application received, has received high no. of application accepted, high % of student enrolling for admission, high no. of student with fulltime undergrad, high no. of student with parttime undergrad, low tution fee for in state student , high charges for additionalfees , high cost of books compared to others, high personal expense, good ratio for student to faculty,

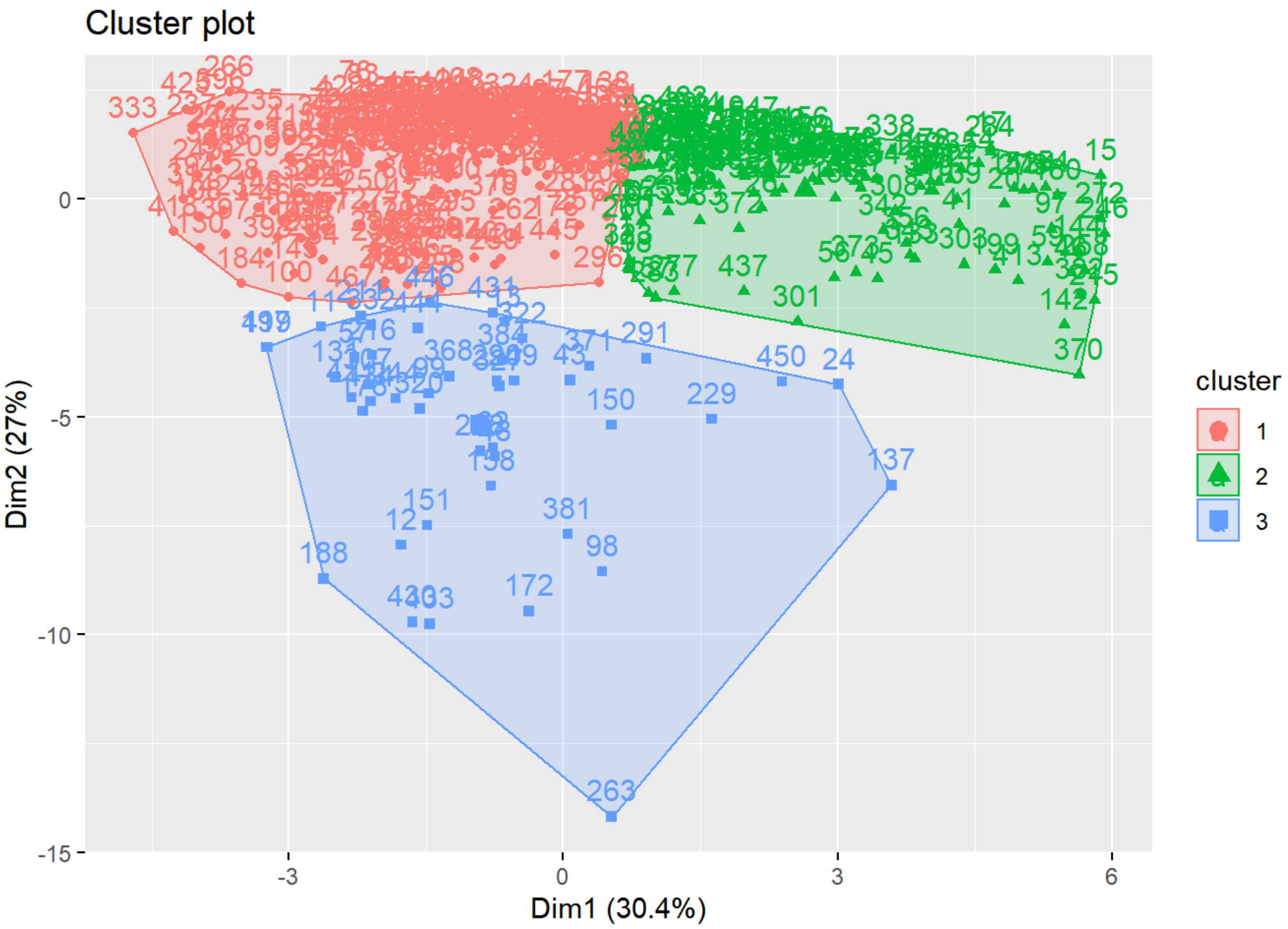
Cluster 2

high no of new student from top 10%, high no of new student from top 25%, low no. of student enrolled in undergrad, high tution fee for in state student , high tution fee for out state student, high occupncy for rooms, high useage of board, low personal expense, high % of faculty with PHD, bad ratio for student to faculty, high graduation rate,

Cluster 1

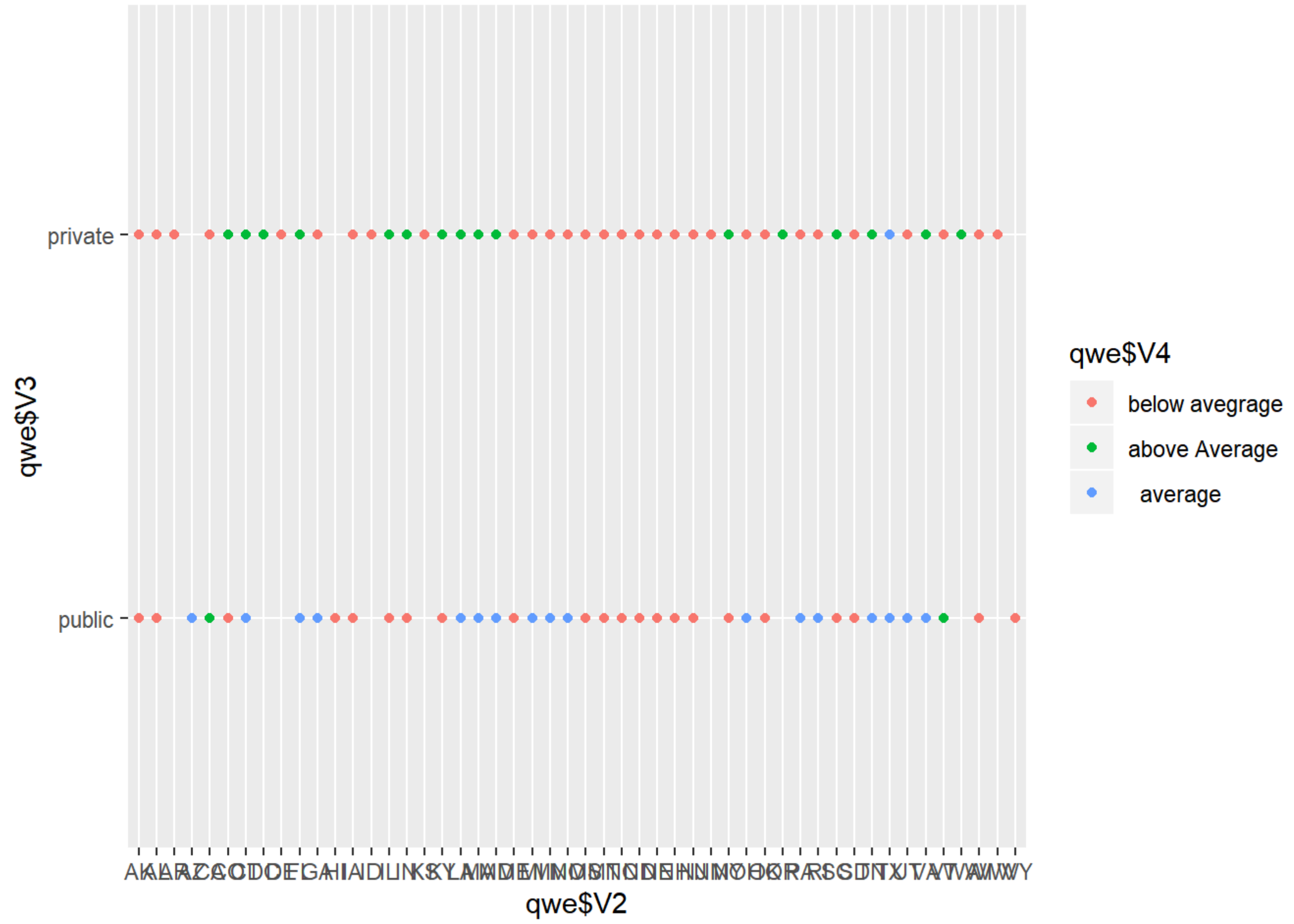
received low application, accepted low application, low % of student enrolling for admission, low no of new student from top 10%, low no of new student from top 25%, low no. of student with fulltime undergrad, low tution fee for out state student, low occupncy for rooms, low useage of board, low charges for additionalfees, low cost of books compared to others, low % of faculty with PHD , low graduation rate,

```
## ploting information about cluster on graph

fviz_cluster(fitk,data = univ)
```



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?

```
## adding catagorical variable into a table format


  univ1 <- na.omit(Universities) ## using orignal dataset to omit NA values

qwe <- cbind(univ1$`College Name`,univ1$State,univ1$`Public (1)/ Private (2)`,fitk$cluster) #
## combining coloumn with cluters information and finding which university fallin which clust
er

qwe <- as.data.frame(qwe) ## converting into dataframe

qwe$V3 <- factor(qwe$V3,levels = c(1,2),labels = c("public","private"))  ### defining levels
 for private ans public

qwe$V4 <- factor(qwe$V4,levels = c(1,2,3),labels = c("below avegrage ","above Average", "  av
erage")) ## name cluster as different catagory

library(ggplot2)

ggplot(qwe, aes(x=qwe$V2,y=qwe$V3, color= qwe$V4)) +
  geom_point()
```



```
head(qwe)
```

	V1 <fctr>	V2 <fctr>	V3 <fctr>	V4 <fctr>
1	Alaska Pacific University	AK	private	below avegrage
2	University of Alaska Southeast	AK	public	below avegrage
3	Birmingham-Southern College	AL	private	above Average
4	Huntingdon College	AL	private	below avegrage
5	Talladega College	AL	private	below avegrage
6	University of Alabama at Birmingham	AL	public	below avegrage

6 rows

```
head(qwe[qwe$V4=="below avegrage ",]) ### this show that good mix of private and  public univ
ersity
```

	V1 <fctr>	V2 <fctr>	V3 <fctr>	V4 <fctr>
1	Alaska Pacific University	AK	private	below avegrage
2	University of Alaska Southeast	AK	public	below avegrage
4	Huntingdon College	AL	private	below avegrage

	V1 <fctr>	V2 <fctr>	V3 <fctr>	V4 <fctr>
5	Talladega College	AL	private	below avegrage
6	University of Alabama at Birmingham	AL	public	below avegrage
7	Arkansas College (Lyon College)	AR	private	below avegrage

6 rows

```
head(qwe[qwe$V4=="  average",])  ### this show that more no. of public university
```

	V1 <fctr>	V2 <fctr>	V3 <fctr>	V4 <fctr>
11	Northern Arizona University	AZ	public	average
12	University of Arizona	AZ	public	average
13	California Polytechnic-San Luis	CA	public	average
24	University of Southern California	CA	private	average
43	University of Connecticut at Storrs	CT	public	average
48	University of Delaware	DE	private	average

6 rows

```
head(qwe[qwe$V4=="above Average",])  ###  this show that more no. of private university
```

	V1 <fctr>	V2 <fctr>	V3 <fctr>	V4 <fctr>
3	Birmingham-Southern College	AL	private	above Average
14	Claremont McKenna College	CA	private	above Average
15	Harvey Mudd College	CA	private	above Average
16	Pitzer College	CA	private	above Average
17	Scripps College	CA	private	above Average
18	Occidental College	CA	private	above Average

6 rows

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

```
fitk$withinss
```

```
## [1] 2562.342 1424.892 1044.680
```

```
fitk$tot.withinss
```

```
## [1] 5031.914
```

```
fitk$betweenss
```

```
## [1] 2958.086
```

`

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.

```
univ1<- univ1[,-c(1:3)]

Km<-kmeans(univ1,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 university data
```

```
##          a2
## b1 713.8496
```

```
dist(rbind(a2,b2))
```

```
##          a2
## b2 1314.998
```

```
dist(rbind(a2,b3))
```

```
##          a2
## b3 2452.064
```

```
a1$`# PT undergrad`<-3255.4528 # From the above, Mean value which is near to cluster 1. Hence
replacing the missing value with mean value
univ3 <- na.omit(Universities)
uniV2<-rbind(univ3,a1)
View(uniV2)
uni2_z<-scale(uniV2[,-c(1:3)])
uni2_cluster<-kmeans(uni2_z,3)
uni2<-cbind(uniV2,uni2_cluster$cluster)
uni2[472,] # From the model, this uniersity falls under Cluster 2("Above Average")
```

	College Name <chr>	State <chr>	Public (1)/ Private (2) <dbl>	# appli. rec'd <dbl>
472	Tufts University	MA	2	7614

1 row | 1-5 of 22 columns

```
### 2 part

univ2 <- Universities[ Universities$`College Name`== "Tufts University",] ## selecting Tufts
 university from original dataset

view(a1)
a1$`# PT undergrad` <- mean(Universities$`# PT undergrad`,na.rm=TRUE) ## selecting means valu
e from coloumn and applying it to NA value for PT undergrad

a3 <- rbind(univ1,a1[,-c(1:3)])  ### combining dataset
view(a3)

a3_scale <- scale(a3) ### applying normalization

fitk2 <- kmeans(a3_scale,3) ### apply kmeans to dataset


a3<-cbind( a3,fitk2$cluster) ## combining cluster information col to dataset to find which cl
uster does tufts fallunder

a3[472,]  ### this university falls in cluster 1
```

	# appli. rec'd <dbl>	# appl. accepted <dbl>	# new stud. enrolled <dbl>
472	7614	3605	1205

1 row | 1-4 of 19 columns