

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

```
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()

#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
```

```

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)

#Changing the column names to suitable ones.
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)
count(DFUniver1)

##      n
## 1 471
```

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
```

#Subsetting the data

```
DFNumerical<-subset(DFUNiver1, select = -c(1,2,3))
```

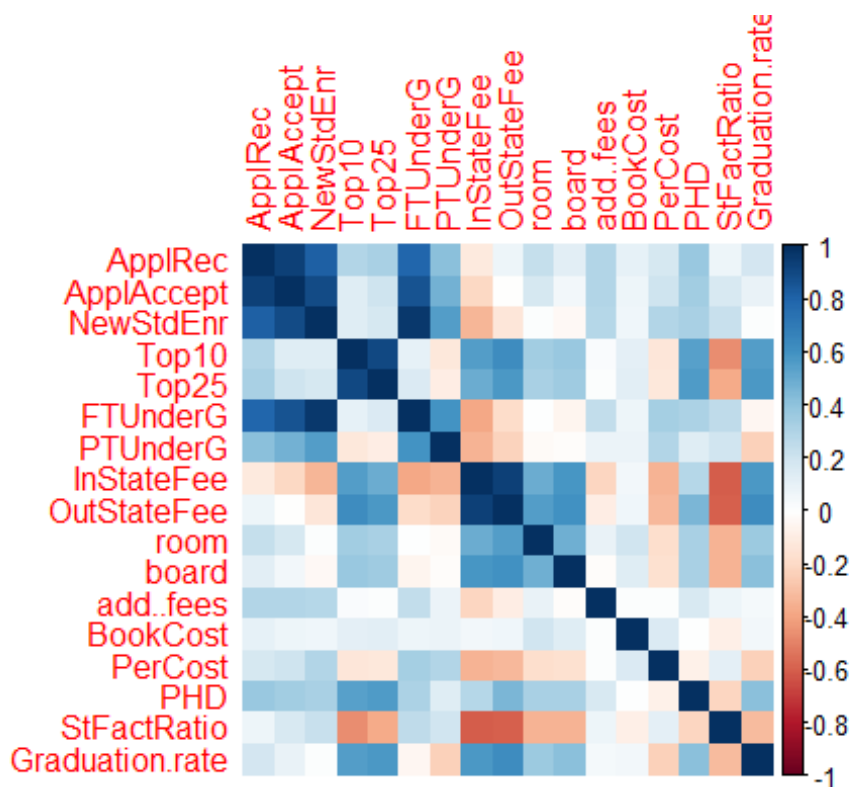
#Finding the correlation between the data set

```
library(corrplot)
```

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

```
## corrplot 0.84 loaded
```

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



In the correlation graph, Darker Blue(+1) and Dark Orange(-1) shows the higher correlated data. Using this data to understand any correlation among the column data.

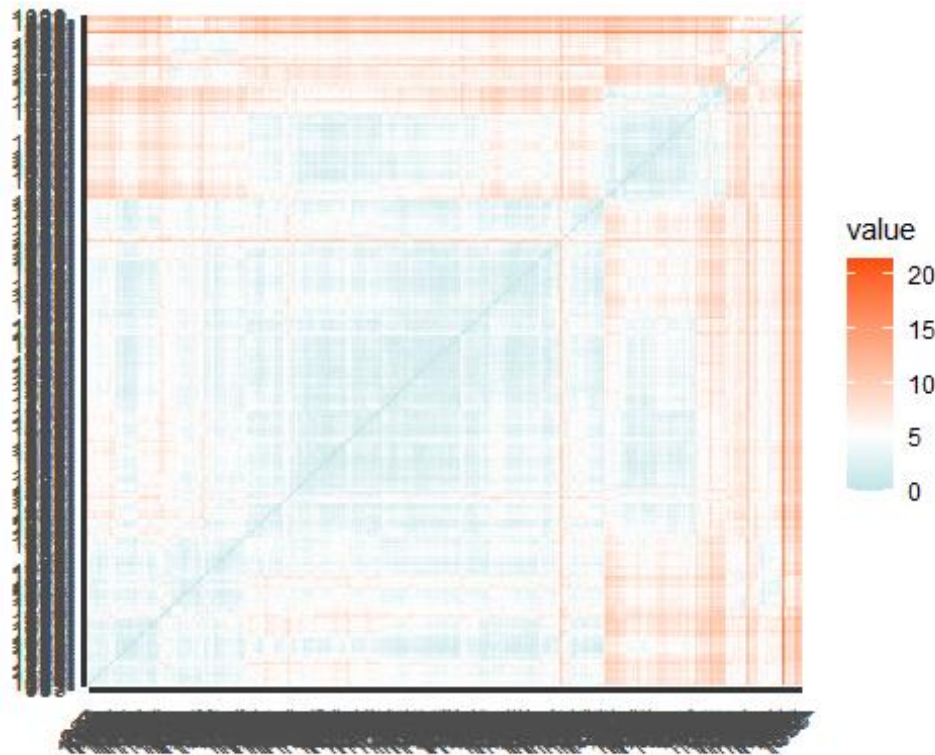
Applying K-means clustering for Numeric Data

#Scaling the Data

```
DFNumerical<-scale(DFNumerical)
```

```
#Distance Between Observations
distance <- get_dist(DFNumerical)
```

```
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high =
"#FC4E07"))
```



Comparison different cluster values

```
k2 <- kmeans(DFNumerical, centers = 2, nstart = 25)
k3 <- kmeans(DFNumerical, centers = 3, nstart = 25)
k4 <- kmeans(DFNumerical, centers = 4, nstart = 25)
k5 <- kmeans(DFNumerical, centers = 5, nstart = 25)
k6 <- kmeans(DFNumerical, centers = 5, nstart = 25)
```

```
# plots to compare
```

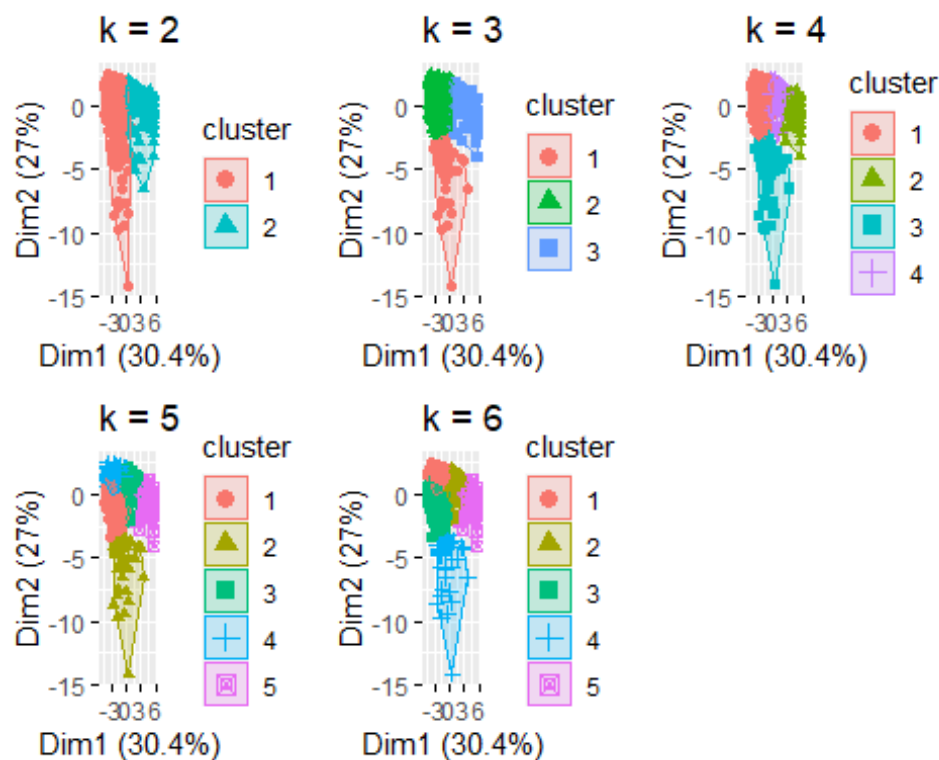
```
p2 <- fviz_cluster(k2, geom = "point", data = DFNumerical) + ggtitle("k = 2")
p3 <- fviz_cluster(k3, geom = "point", data = DFNumerical) + ggtitle("k = 3")
p4 <- fviz_cluster(k4, geom = "point", data = DFNumerical) + ggtitle("k = 4")
p5 <- fviz_cluster(k5, geom = "point", data = DFNumerical) + ggtitle("k = 5")
p6 <- fviz_cluster(k6, geom = "point", data = DFNumerical) + ggtitle("k = 6")
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

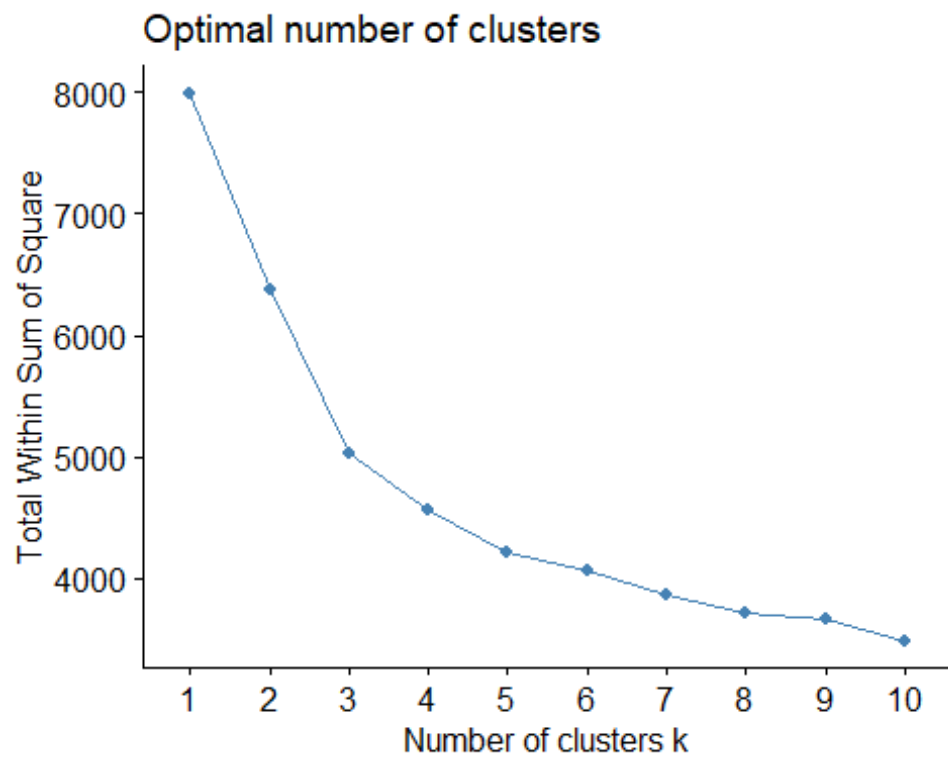
## The following object is masked from 'package:dplyr':
##
##      combine

grid.arrange(p2, p3, p4, p5, p6, nrow = 2)
```

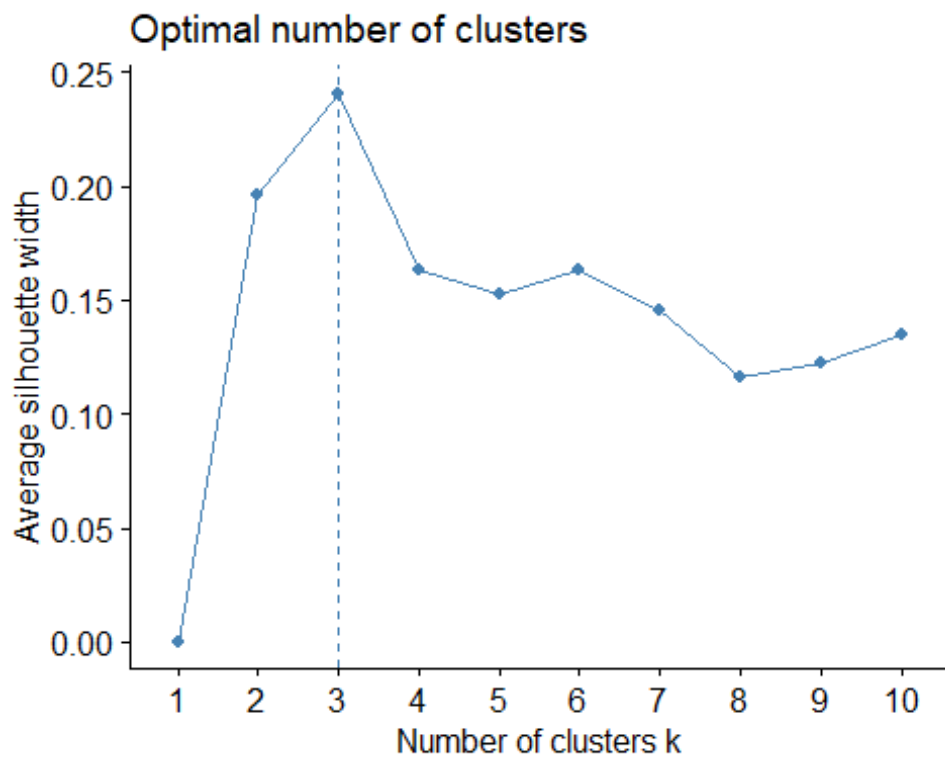


From the above comparison it seems that 3 clusters would be good. Determining Optimal Cluster using Elbow and Silhouette method.

```
set.seed(123)
#Finding optimal number of clusters - Elbow Method
fviz_nbclust(DFNumerical, kmeans, method = "wss")
```



#Determining Optimal Cluster by Average Silhouette Method
`fviz_nbclust(DFNumerical, kmeans, method = "silhouette")`



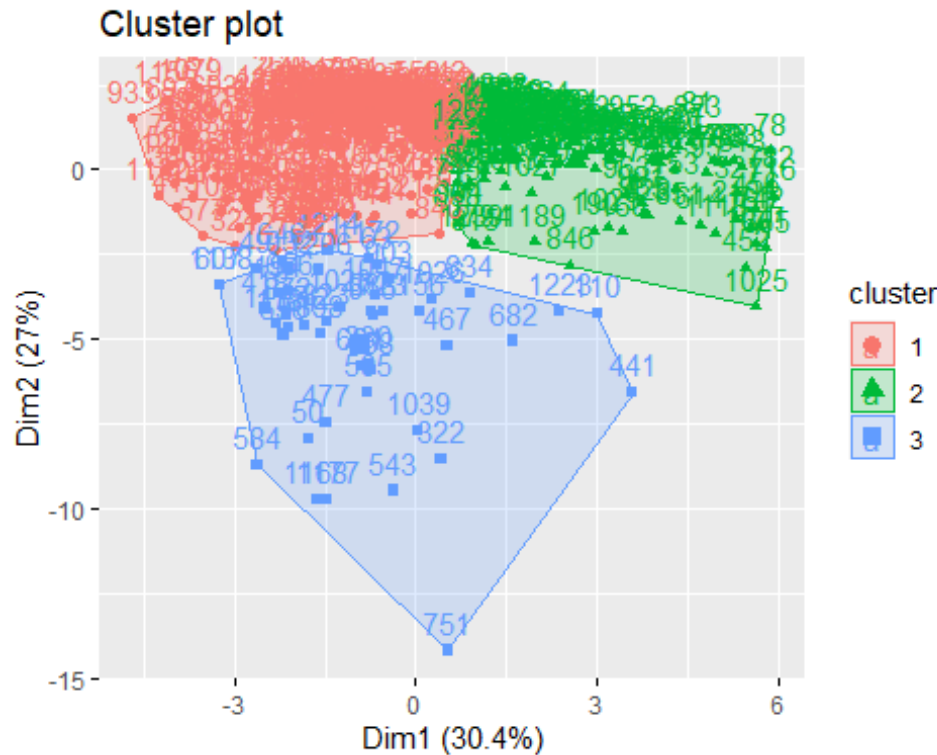
#Silhouette method shows that 3 numbers of clusters would be optimum. From previous cluster plotting we have seen that optimal cluster size would be 3.

#3 clusters are the reasonable for this data and the optimal K is 3.

```
k3 <- kmeans(DFNumerical, centers = 3, nstart = 25)
```

Optimal Visualization

```
fviz_cluster(k3,data = DFNumerical)
```



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

3 is the Optimal Cluster

```
k3 <- kmeans(DFNumerical,centers = 3 ,nstart = 25)
```

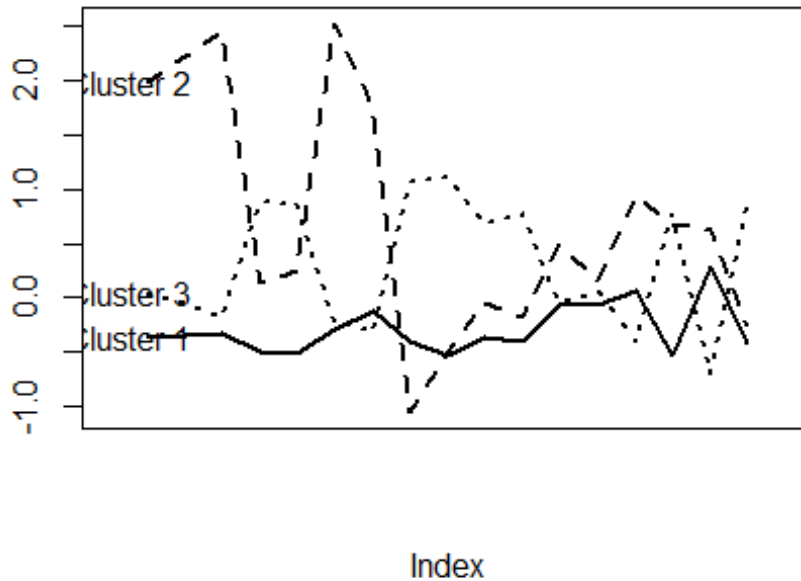
```
#print(k3)
```

```
k3$centers # Description of the centers
```

	ApplRec	ApplAccept	NewStdEnr	Top10	Top25	FTUnderG
## 1	-0.35953828	-0.34918455	-0.3171053	-0.5020886	-0.5128195	-0.2952142
## 2	1.98179657	2.22992267	2.4447222	0.1334215	0.2545856	2.5228452
## 3	0.05140256	-0.04367128	-0.1683551	0.8795798	0.8620961	-0.2324464
	PTUnderG	InStateFee	OutStateFee	room	board	add..fees
## 1	-0.1217682	-0.4036544	-0.5263964	-0.3588740	-0.3938990	-0.05832646
## 2	1.7486849	-1.0500277	-0.4918168	-0.0388330	-0.1745795	0.49531762
## 3	-0.3130216	1.0620416	1.1158839	0.6698444	0.7756859	-0.04496556
	BookCost	PerCost	PHD	StFactRatio	Graduation.rate	
## 1	-0.06621454	0.05935933	-0.5322257	0.2810858	-0.4171456	


```
## 2  0.16358567  0.93858632  0.6840794  0.6139980      -0.2538234
## 3  0.07122705 -0.39665857  0.7659627 -0.7036167      0.8426062

plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(k3$centers), max(k3$centers)), xlim = c(0, 18))
# plot centroids
for (i in c(1:3))
  lines(k3$centers[i,], lty = i, lwd = 2)
# name clusters
text(x = 0.5, y = k3$centers[, 1], labels = paste("Cluster", c(1:3))) #
Cluster Names
```



```
k3$size # Count of Clusters
## [1] 275  46 150

#Merging the clusters to the original Data frame
Clusters<-data.frame(k3$cluster)

Clusters<-Clusters%>%rename(clusters=k3.cluster)

UnivAnalysis<-cbind(DFUNiver1, Clusters)
head(UnivAnalysis)

##              College.Name State Pub.Private ApplRec
ApplAccept
## 1      Alaska Pacific University      AK           2      193
```

```

146
## 3      University of Alaska Southeast      AK      1      146
117
## 10      Birmingham-Southern College      AL      2      805
588
## 12      Huntingdon College      AL      2      608
520
## 22      Talladega College      AL      2      4414
1500
## 26 University of Alabama at Birmingham      AL      1      1797
1260
##      NewStdEnr Top10 Top25 FTUnderG PTUnderG InStateFee OutStateFee room
board
## 1      55      16      44      249      869      7560      7560 1620
2500
## 3      89      4      24      492      1849      1742      5226 2514
2250
## 10     287     67     88     1376     207     11660      11660 2050
2430
## 12     127     26     47     538     126     8080      8080 1380
2540
## 22     335     30     60     908     119     5666      5666 1424
1540
## 26     938     24     35     6960     4698     2220      4440 1935
3240
##      add..fees BookCost PerCost PHD StFactRatio Graduation.rate clusters
## 1      130      800      1500 76      11.9      15      1
## 3      34      500      1162 39      9.5      39      1
## 10     120      400      900 74      14.0     72      3
## 12     100      500      1100 63      11.4     44      1
## 22     418     1000     1400 56      15.5     46      1
## 26     291      750     2200 96      6.7      33      1

```

```

ClusterStat<-
UnivAnalysis%>%group_by(clusters)%>%summarise(Acceptance_rate=sum(ApplAccept)
/sum(ApplRec),
AvgOutStateTution=mean(OutStateFee),AvgInStateTution=mean(InStateFee),
AvgGradRate=mean(Graduation.rate))

```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ClusterStat
```

```

## # A tibble: 3 x 5
##   clusters Acceptance_rate AvgOutStateTution AvgInStateTution AvgGradRate
##   <int>      <dbl>      <dbl>      <dbl>      <dbl>
## 1      1      0.706      8306.      7180.      58.0
## 2      2      0.682      8455.      3614.      61.0
## 3      3      0.582     15386.     15266.      80.9

```

#Cluster 1 - Universities with highest acceptance rate, Lowest Out State fee but average In state fees but the graduation rate is also low.

#Cluster 3- Universities with lowest acceptance rate but having highest Out of state fees. But the graduation rate is significantly higher.

#Cluster 2 - Universities with lowest in state tuition fees and have around 60% graduation rate.

#We can perform all comparision analysis in similar ways.

Using the categorical measurements that were not used i n the analysis (State and Private/Public) to characterize the different clusters.

#State wise values present in the cluster

table(DFUniver1\$State, k3\$cluster)

```
##
##      1  2  3
## AK   2  0  0
## AL   3  0  1
## AR   4  0  0
## AZ   0  2  0
## CA   3  2 10
## CO   5  0  1
## CT   3  1  6
## DC   0  0  4
## DE   1  1  0
## FL   3  1  4
## GA   4  1  2
## HI   1  0  0
## IA  16  0  2
## ID   2  0  0
## IL   7  2  6
## IN   8  0  7
## KS   7  0  0
## KY   4  0  2
## LA   2  1  2
## MA   7  3 12
## MD   1  1  1
## ME   4  0  2
## MI   7  2  4
## MN   6  1  4
## MO  12  1  2
## MS   5  0  0
## MT   2  0  0
## NC  16  4  3
## ND   5  0  0
## NE   5  1  1
## NH   4  1  1
```

```

##    NJ   9   1   3
##    NM   2   0   0
##    NY  18   2  18
##    OH  13   4   7
##    OK   5   1   0
##    OR   1   0   4
##    PA  19   3  20
##    RI   1   1   2
##    SC   7   0   2
##    SD   4   0   0
##    TN  11   1   3
##    TX  14   4   2
##    UT   1   1   0
##    VA   8   3   4
##    VT   5   0   2
##    WA   0   0   2
##    WI   5   0   4
##    WV   2   0   0
##    WY   1   0   0

#View(UnivAnalysis)
Cluster1 <- UnivAnalysis[UnivAnalysis$clusters == 1,]
#View(Cluster1[,c(1,2,3,21)])

Cluster2 <- UnivAnalysis[UnivAnalysis$clusters == 2,]
#View(Cluster2[,c(1,2,3,21)])

Cluster3 <- UnivAnalysis[UnivAnalysis$clusters == 3,]
#View(Cluster3[,c(1,2,3,21)])

library(ggplot2)
legend <- factor(UnivAnalysis$clusters,levels = c(1,2,3),labels = c("Cluster
1","Cluster 2","Cluster 3"))
ggplot(UnivAnalysis,aes(x=State,y=factor(Pub.Private,levels = c(1,2),labels =
c("Public","Private")),
      color=legend))+
  geom_point()+ylab("Public/Private")+xlab("state")+ggtitle("Cluster
Plot")

```



#Plotted the graph which represents all the public and private colleges along with its cluster. Each state has colleges a maximum of 2 out of the 3 clusters. Yes, there is a relationship between clusters and categorical information.

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

k3\$withinss # within cluster sum of squares with high ratio as possible

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

k3\$betweenss # mean of distances between cluster centers with ration Lower as possible

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

k3\$size # number of points in each cluster

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

k3\$centers # CLuster Centers

```
##      ApplRec  ApplAccept  NewStdEnr      Top10      Top25  FTUnderG
## 1 -0.35953828 -0.34918455 -0.3171053 -0.5020886 -0.5128195 -0.2952142
## 2  1.98179657  2.22992267  2.4447222  0.1334215  0.2545856  2.5228452
## 3  0.05140256 -0.04367128 -0.1683551  0.8795798  0.8620961 -0.2324464
##      PTUnderG  InStateFee  OutStateFee      room      board  add..fees
## 1 -0.1217682 -0.4036544 -0.5263964 -0.3588740 -0.3938990 -0.05832646
## 2  1.7486849 -1.0500277 -0.4918168 -0.0388330 -0.1745795  0.49531762
```

```
## 3 -0.3130216 1.0620416 1.1158839 0.6698444 0.7756859 -0.04496556
##      BookCost      PerCost      PHD StFactRatio Graduation.rate
## 1 -0.06621454 0.05935933 -0.5322257 0.2810858 -0.4171456
## 2 0.16358567 0.93858632 0.6840794 0.6139980 -0.2538234
## 3 0.07122705 -0.39665857 0.7659627 -0.7036167 0.8426062
```

- #• Within cluster sum of squares with high ratio as possible*
- #• Mean of distances between cluster centers with ration lower as possible*
- #• Number of points in each cluster*
- #• Cluster Centers*
- #• The k value which the highest \$withinss is the best choice, because we expect the within sum of squares ratio to be as Lower as possible.*

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.

```
library(cluster)

#Finding the Tufts University Data
Tufts<-DFUniver[DFUniver$College.Name == "Tufts University",]
Tufts<-data.frame(Tufts)

#Finding the NULL fields from the Tufts University data frame
Tufts[!complete.cases(Tufts)]

##      College.Name State Pub.Private ApplRec ApplAccept NewStdEnr Top10
Top25
## 476 Tufts University    MA          2    7614        3605    1205    60
90
##      FTUnderG PTUnderG InStateFee OutStateFee room board add..fees BookCost
## 476    4598      NA    19701    19701 3038  2930      503    600
##      PerCost PHD StFactRatio Graduation.rate
## 476    928  99      10.3          92

#PTUnderG field has NULL value

#Euclidean Distance
dist(rbind(Tufts, k3$centers[1,]))

## Warning in dist(rbind(Tufts, k3$centers[1, ])): NAs introduced by coercion
##      476
## 2 32343.31

dist(rbind(Tufts, k3$centers[2,]))

## Warning in dist(rbind(Tufts, k3$centers[2, ])): NAs introduced by coercion
```

```
##          476
## 2 32342.17

dist(rbind(Tufts, k3$centers[3,]))

## Warning in dist(rbind(Tufts, k3$centers[3, ])): NAs introduced by coercion

##          476
## 2 32341.6

#Cluster 1 has shortest distance

Avg_PTUnderG <- mean(Cluster1$PTUnderG)

#Update the value 609.22 for the Tuft University
Tufts[,c(10)]<-Avg_PTUnderG
Tufts[,c(10)]

## [1] 609.2255
```