# R Notebook

Code

Qs 1. The dataset on American College and University Rankings contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements, including continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school).

Hide

```
library(tidyverse)  # data manipulation
##install.packages("factoextra") # if necessary
library(factoextra) # clustering algorithms & visualization
library(ISLR)
```

Hide

```
udata <- read.csv("Universities.csv")
```

Qs1 Remove all records with missing measurements from the dataset.

Hide

```
udata1 <-na.omit(udata) ##remove all the missing values
```

Hide

```
udata2 <- udata1[,c(-1,-2, -3)] ##remove the categorical variable.
summary(udata2) ##show summary of dataset
```

/

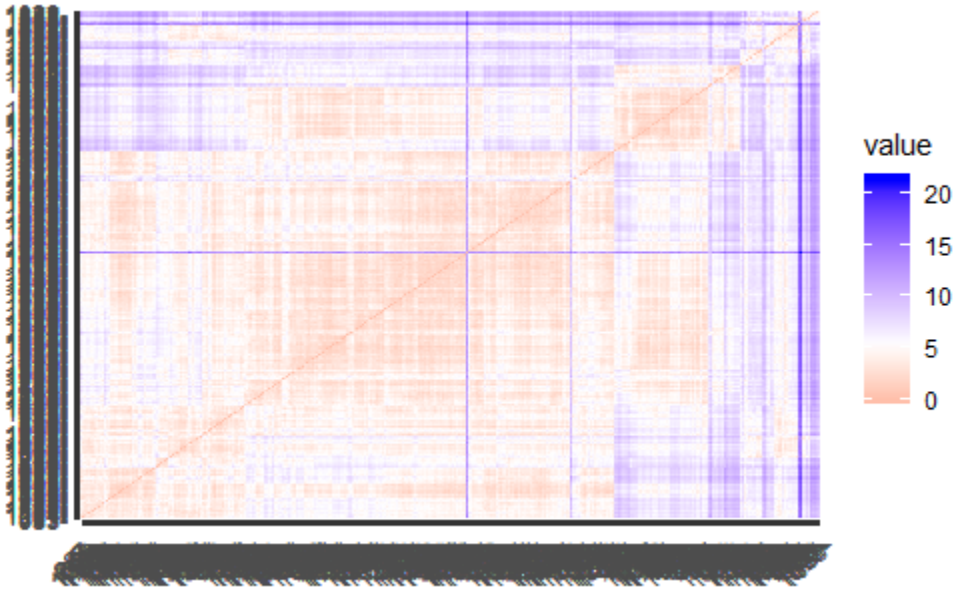
```
X..appli..rec.d X..appl..accepted X..new.stud..enrolled
Min.   :   77   Min.   :   61.0   Min.   :  27.0
1st Qu.:  802   1st Qu.:  635.5   1st Qu.: 264.0
Median : 1646   Median : 1227.0   Median : 443.0
Mean   : 3147   Mean   : 2063.0   Mean   : 780.7
3rd Qu.: 3862   3rd Qu.: 2456.0   3rd Qu.: 896.5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

/

Qs 2. 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?

Hide

```
udata2 <-scale(udata2) ##scale the dataset
distance <- get_dist(udata2)
fviz_dist(distance)
```




Hide

```
##The graph shows the distance between continuous variables
```

Hide

```
##. Lets  run the k-means algorithm to cluster the Universities. lets take initial value of k = 4.
set.seed(123)
k4 <- kmeans(udata2, centers = 4, nstart = 25)
              # k = 4, number of restarts = 25
```

Hide

```
k4$centers# output the centers
```

/

```
  X..appli..rec.d X..appl..accepted X..new.stud..enrolled
1       1.9817966         2.2299227          2.444722e+00
2      -0.3692895        -0.3314846         -3.967692e-01
3      -0.3033156        -0.2989118         -2.276979e-01
4       0.4402622         0.1551461         -2.000371e-05
  X..new.stud..from.top.10. X..new.stud..from.top.25.
1                 0.1334215                 0.2545856
2                 0.0102519                 0.1080080
3                -0.6785172                -0.7279285
4                 1.6526422                 1.4315089
  X..FT.undergrad X..PT.undergrad in.state.tuition
1       2.5228452      1.74868491       -1.0500277
2      -0.4049392     -0.25785122        0.4057712
3      -0.1972688     -0.04353747       -0.7234450
4      -0.1108205     -0.38259215        1.5022093
  out.of.state.tuition        room      board   add..fees
1           -0.4918168 -0.03883300 -0.1745795  0.49531762
2            0.2956208  0.08357902  0.3292398 -0.18996619
3           -0.8237908 -0.53385193 -0.6791344  0.03928218
4            1.6819156  1.19276784  0.9944521  0.07619136
  estim..book.costs estim..personal.. X..fac..w.PHD
1       0.163585669         0.9385863     0.6840794
2      -0.158302104        -0.2978018     0.0835866
3       0.003218005         0.2531393    -0.6684106
4       0.311659604        -0.4921884     1.0478784
  stud..fac..ratio Graduation.rate
1        0.6139980      -0.2538234
2       -0.1828501       0.3971948
3        0.4582141      -0.7769793
4       -1.1189523       1.1188151
```

Hide

```
str(k4)
```

/

```
List of 9
 $ cluster     : Named int [1:471] 3 3 2 3 3 3 2 2 2 3 ...
  ..- attr(*, "names")= chr [1:471] "1" "3" "10" "12" ...
 $ centers     : num [1:4, 1:17] 1.982 -0.369 -0.303 0.44 2.23 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "1" "2" "3" "4"
  .. ..$ : chr [1:17] "X..appli..rec.d" "X..appl..accepted" "X..new.stud..enrolled" "X..new.stud..from.top.10." ...
 $ totss       : num 7990
 $ withinss    : num [1:4] 1045 1271 1679 560
 $ tot.withinss: num 4555
 $ betweenss   : num 3435
 $ size        : int [1:4] 46 183 175 67
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

Hide

```
k4$size
```

```
[1]  46 183 175  67
```

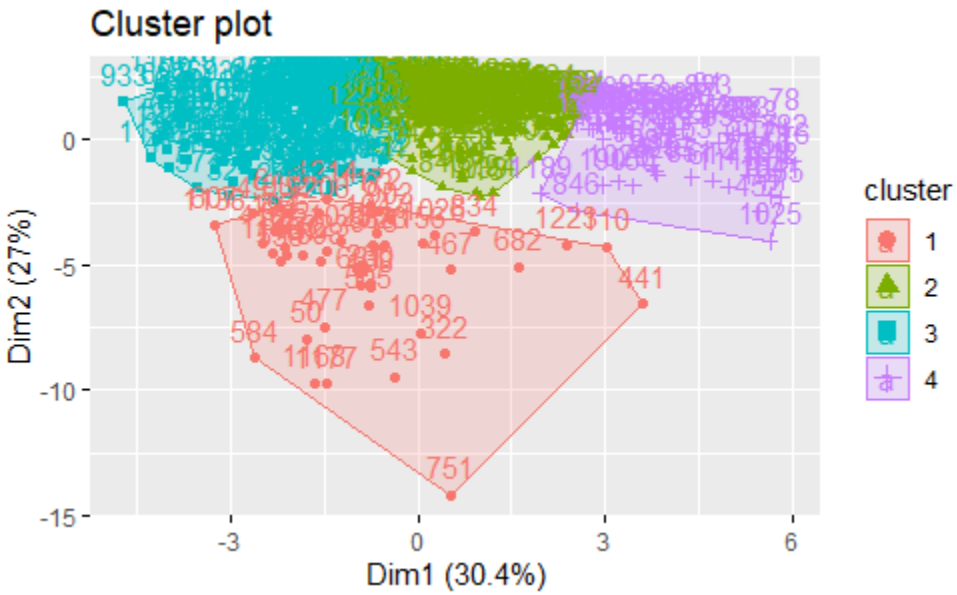
Hide

```
##size of cluster
```

Hide

```
fviz_cluster(k4, data = udata2) ###Visualize cluster plot
```

/

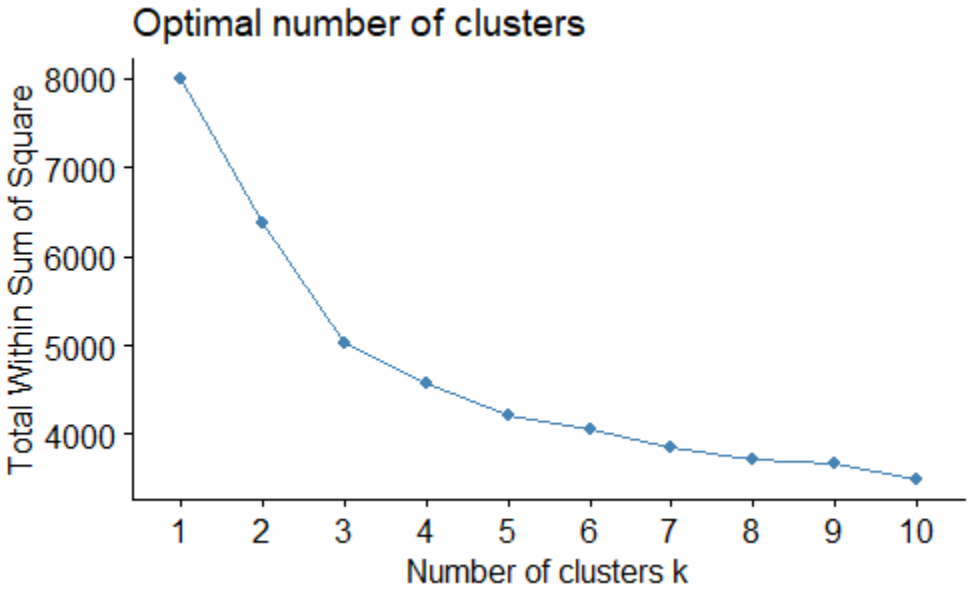



Hide

```
library(tidyverse)  # data manipulation
library(factoextra) # clustering & visualization
library(ISLR)
set.seed(123)

fviz_nbclust(udata2, kmeans, method = "wss")
```

/



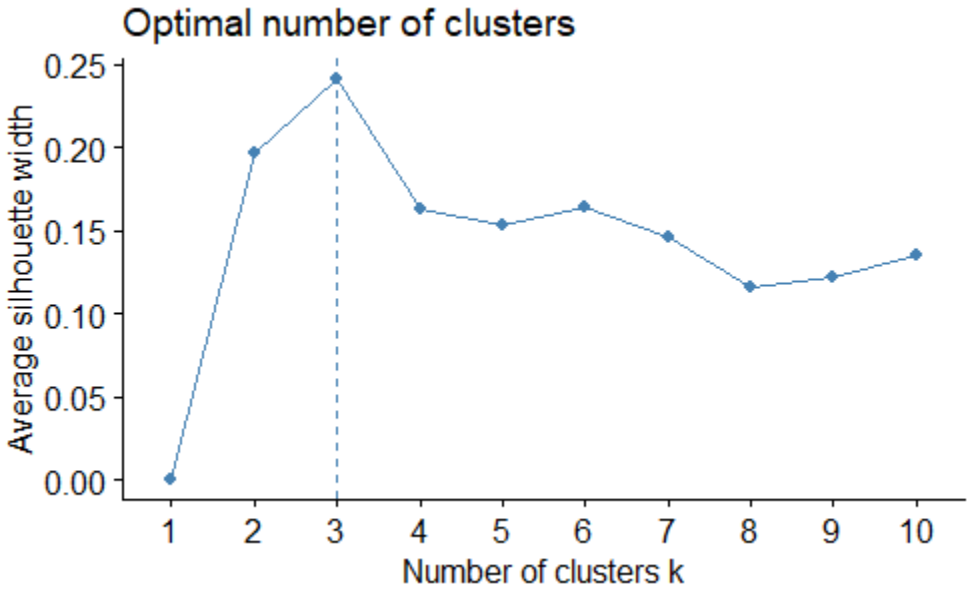

Hide

```
##3 is the ideal number of k.
###Calculating our optimal K using Elbow chart
##The charts shows that the point 3 in Silhouette provides the best value for k.
```

Hide

```
fviz_nbclust(udata2, kmeans, method = "silhouette")
```

/




Hide

```
##calculating optimal k using silhouette method
```

The charts shows that the point 3 in Silhouette provides the best value for k. elbow and Silhouette provides the best value for k. While WSS, Silhouette will continue to drop for larger values of k, we have to make the tradeoff between overfitting, 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.

Hide

```
library(flexclust)
set.seed(123)
#Creating the cluster index for 3 clusters
set.seed(123)
k3 = kcca(udata2, k=3, kccaFamily("kmedians"))
k3
```

/

```
kcca object of family ‘kmedians’

call:
kcca(x = udata2, k = 3, family = kccaFamily("kmedians"))

cluster sizes:

  1   2   3
111 113 247
```

Hide

```
library(dplyr)
library(stats)
set.seed(123)
k3 <- kmeans(udata2, centers = 3, nstart = 25)
              # k = 3, number of restarts = 25
k3$centers# output the centers
```

/

```
  X..appli..rec.d X..appl..accepted X..new.stud..enrolled
1      1.98179657        2.22992267             2.4447222
2      0.05140256       -0.04367128            -0.1683551
3     -0.35953828       -0.34918455            -0.3171053
  X..new.stud..from.top.10. X..new.stud..from.top.25.
1                 0.1334215                 0.2545856
2                 0.8795798                 0.8620961
3                -0.5020886                -0.5128195
  X..FT.undergrad X..PT.undergrad in.state.tuition
1       2.5228452       1.7486849       -1.0500277
2      -0.2324464      -0.3130216        1.0620416
3      -0.2952142      -0.1217682       -0.4036544
  out.of.state.tuition       room      board   add..fees
1           -0.4918168 -0.0388330 -0.1745795  0.49531762
2            1.1158839  0.6698444  0.7756859 -0.04496556
3           -0.5263964 -0.3588740 -0.3938990 -0.05832646
  estim..book.costs estim..personal.. X..fac..w.PHD
1        0.16358567        0.93858632     0.6840794
2        0.07122705       -0.39665857     0.7659627
3       -0.06621454        0.05935933    -0.5322257
  stud..fac..ratio Graduation.rate
1        0.6139980      -0.2538234
2       -0.7036167       0.8426062
3        0.2810858      -0.4171456
```

Hide

```
str(k3)
```

/

```
List of 9
 $ cluster     : Named int [1:471] 3 3 2 3 3 3 3 3 3 3 ...
  ..- attr(*, "names")= chr [1:471] "1" "3" "10" "12" ...
 $ centers     : num [1:3, 1:17] 1.9818 0.0514 -0.3595 2.2299 -0.0437 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:17] "X..appli..rec.d" "X..appl..accepted" "X..new.stud..enrolled" "X..new.stud..from.top.10." ...
 $ totss       : num 7990
 $ withinss    : num [1:3] 1045 1425 2562
 $ tot.withinss: num 5032
 $ betweenss   : num 2958
 $ size        : int [1:3] 46 150 275
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

Hide

```
library(flexclust)
set.seed(123)
#Creating the cluster index for 3 clusters
set.seed(123)
k3 = kcca(udata2, k=3, kccaFamily("kmedians"))
k3
```

```
kcca object of family ‘kmedians’

call:
kcca(x = udata2, k = 3, family = kccaFamily("kmedians"))

cluster sizes:

  1   2   3
111 113 247
```

Hide

```
cluster <- predict(k3)
```

Hide

```
#Merging the clusters to the original data frame
set.seed(123)
cluster <- data.frame(cluster)
udata1 <- cbind(udata1, cluster)
head(udata1)
```

| | College.Name <fctr> | State <fctr> | Public..1...Private..2. <int> | X..appli..rec.d <int> |
|---|---|---|---|---|
| 1 | Alaska Pacific University | AK | 2 | 193 |
| 3 | University of Alaska Southeast | AK | 1 | 146 |
| 10 | Birmingham-Southern College | AL | 2 | 805 |
| 12 | Huntingdon College | AL | 2 | 608 |
| 22 | Talladega College | AL | 2 | 4414 |
| 26 | University of Alabama at Birmingham | AL | 1 | 1797 |

6 rows | 1-5 of 21 columns

Hide

```
NA
NA
```

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

Hide

```
#Summary Statistics for Each Cluster

set.seed(123)
Cluster_Stat <- udata1 %>%
    group_by( cluster ) %>%
    summarise( Univ_InState_Max_Fee=udata1[which.max(in.state.tuition),1],Univ_OutState_Max_Fee=udata1[which.max(out.of.stat
e.tuition),1],low_accept_rate=udata1[which.min(X..appl..accepted),1],Acceptance_rate = sum(X..appl..accepted)/ sum(X..appl
i..rec.d), Avg_out_state_tuition=mean(out.of.state.tuition), Avg_int_state_tuition=mean(in.state.tuition), mean_PHD_fac=mean
(X..fac..w.PHD), mean_stud_fac_ratio=mean(stud..fac..ratio), mean_grad_rate=mean(Graduation.rate), priv_count = sum(Public..
1...Private..2. == 2), pub_count = sum(Public..1...Private..2. == 1))
head(Cluster_Stat)
```

| cluster <int> | Univ_InState_Max_Fee <fctr> | Univ_OutState_Max_Fee <fctr> |
|---|---|---|
| 1 | Adams State College | Hanover College |
| 2 | Catholic University of America | Catholic University of America |
| 3 | Doane College | Doane College |

3 rows | 1-3 of 12 columns

Hide

```
#Summary Statistics For States

Stat_States<-udata1 %>%
             group_by(State) %>%        summarise(Univ_InState_Max_Fee=udata1[which.max(in.state.tuition),1],Univ_OutState_M
ax_Fee=udata1[which.max(out.of.state.tuition),1],low_accept_rate=udata[which.min(X..appl..accepted),1],Acceptance_rate = sum
(X..appl..accepted)/ sum(X..appli..rec.d), Avg_out_state_tuition=mean(out.of.state.tuition), Avg_int_state_tuition=mean(in.s
tate.tuition), mean_PHD_fac=mean(X..fac..w.PHD), mean_stud_fac_ratio=mean(stud..fac..ratio), mean_grad_rate=mean(Graduation.
rate), priv_count = sum(Public..1...Private..2. == 2), pub_count = sum(Public..1...Private..2. == 1))
head(Stat_States)
```

| State <fctr> | Univ_InState_Max_Fee <fctr> | Univ_OutState_Max_Fee <fctr> |
|---|---|---|
| AK | Alaska Pacific University | Alaska Pacific University |
| AL | Alaska Pacific University | Alaska Pacific University |

| State<br><fctr> | Univ_InState_Max_Fee<br><fctr> | Univ_OutState_Max_Fee<br><fctr> |
|---|---|---|
| AR | University of Alaska Southeast | University of Alaska Southeast |
| AZ | Alaska Pacific University | University of Alaska Southeast |
| CA | Hendrix College | Hendrix College |
| CO | University of Alaska Southeast | University of Alaska Southeast |

6 rows | 1-3 of 12 columns

Hide

```
#Summary Statistics for Private Universities

Stat_Private <- udata1 %>%
    filter(Public..1...Private..2. == 2) %>%
    group_by( cluster) %>%
    summarise( Univ_InState_Max_Fee=udata[which.max(in.state.tuition),1],Univ_OutState_Max_Fee=udata1[which.max(out.of.stat
e.tuition),1],low_accept_rate=udata[which.min(X..appl..accepted),1],Acceptance_rate = sum(X..appl..accepted)/ sum(X..appli..
rec.d), Avg_out_state_tuition=mean(out.of.state.tuition), Avg_int_state_tuition=mean(in.state.tuition), mean_PHD_fac=mean
(X..fac..w.PHD), mean_stud_fac_ratio=mean(stud..fac..ratio), mean_grad_rate=mean(Graduation.rate))
head(Stat_Private)
```

| cluster<br><int> | Univ_InState_Max_Fee<br><fctr> | Univ_OutState_Max_Fee<br><fctr> |
|---|---|---|
| 1 | University of Alaska Southeast | Birmingham-Southern College |
| 2 | Williams Baptist College | University of Connecticut at Storrs |
| 3 | Georgia Southwestern College | Duke University |

3 rows | 1-3 of 10 columns

Hide

```
#Summary Statistics for Public Universities

Stat_Public <- udata1 %>%
    filter(Public..1...Private..2. == 1) %>%
    group_by( cluster ) %>%
    summarise(Univ_InState_Max_Fee=udata[which.max(in.state.tuition),1],Univ_OutState_Max_Fee=udata1[which.max(out.of.state.
tuition),1],low_accept_rate=udata[which.min(X..appl..accepted),1], Acceptance_rate = sum(X..appl..accepted)/ sum(X..appli..r
ec.d), Avg_out_state_tuition=mean(out.of.state.tuition), Avg_int_state_tuition=mean(in.state.tuition), mean_PHD_fac=mean(X..
fac..w.PHD), mean_stud_fac_ratio=mean(stud..fac..ratio), mean_grad_rate=mean(Graduation.rate))
head(Stat_Public)
```

| cluster<br><int> | Univ_InState_Max_Fee<br><fctr> | Univ_OutState_Max_Fee<br><fctr> | low_accept_rate<br><fctr> |
|---|---|---|---|
| 1 | Southern California College | Trinity College | John Brown University |
| 2 | Alaska Pacific University | Alaska Pacific University | Alaska Pacific University |
| 3 | Tuskegee University | Hendrix College | Alaska Pacific University |

3 rows | 1-4 of 10 columns

Following observation we have made out of this dataset:

1. From The Dataframe, we can infer that the cluster3 has greater data points compared to other clusters.
2. cluster 1 has highest public universities as compared to other universities in clusters.
3. The cluster2 has greater private universities which also explain the rational behind high instate and out of state tuition fee.
4. The mean PHD faculty ratio is lowest for cluster 1. 5)The mean room, board, and fees is lowest for cluster 1. 6)The average in state tuition is lowest for cluster 3 and same for out of state tuition.
5. The acceptance rate is lowest for cluster 2. 8)Some additional information that could help explain the data would be the state of the school, or the operating budget of the university, or the amount of academic endowments of the university.

QS 4 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?

Hide

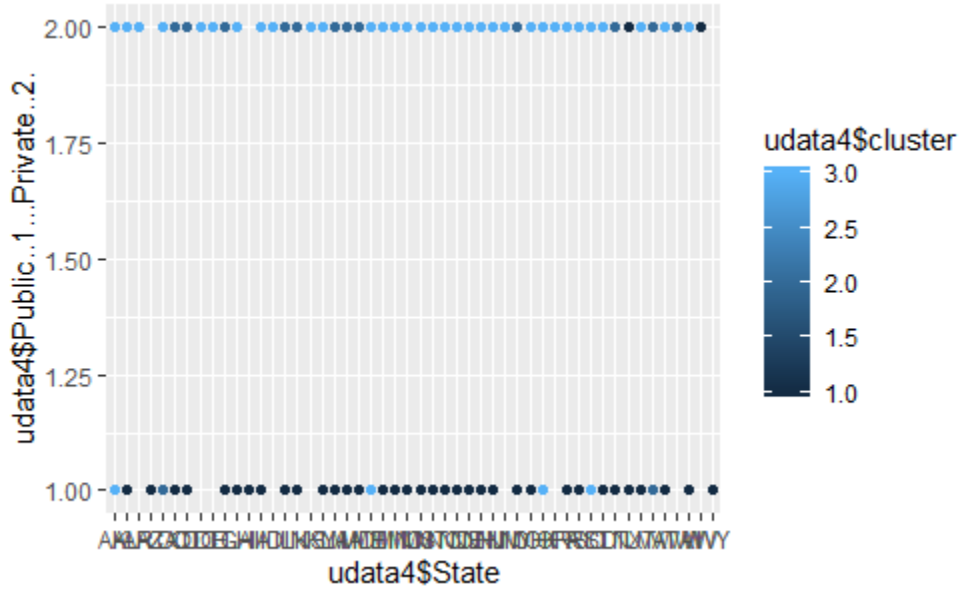
```
clusters<- data.frame(cluster)
udata4 <- cbind(udata1[,c(2,3)], clusters)
udata4
```

| | State<br><fctr> | Public..1...Private..2.<br><int> | cluster<br><int> |
|---|---|---|---|
| 1 | AK | 2 | 3 |
| 3 | AK | 1 | 3 |
| 10 | AL | 2 | 3 |
| 12 | AL | 2 | 3 |
| 22 | AL | 2 | 3 |
| 26 | AL | 1 | 1 |
| 32 | AR | 2 | 3 |
| 38 | AR | 2 | 3 |
| 39 | AR | 2 | 3 |
| 46 | AR | 2 | 3 |



Hide

```
library(ggplot2)
ggplot(udata4, aes(x=udata4$State, y=udata4$Public..1...Private..2., color=udata4$cluster)) + geom_point()
```




This graph shows that cluster 1 has more public university and cluster 3 has more private university, cluster 2 is mixed of public university and private university.

QS 5. What other external information can explain the contents of some or all of these clusters?

1. From cluster_stat dataset, it is inferred that cluster 3 has more datapoint than cluster 1 and cluster2
2. The graduation rate of cluster is highest.
3. The mean PHD faculty ratio is lowest for cluster 1.
4. cluster 1 has highest public universities as compared to other universities in clusters.
5. The cluster2 has greater private universities and high instate and out of state tuition fee.
6. The mean PHD faculty ratio is lowest for cluster 1. 5)The mean room, board, and fees is lowest for cluster 1. 6)The average in state tuition is lowest for cluster 3 and same for out of state tuition.
7. The acceptance rate is lowest for cluster 2.

QS 6. 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.

Hide

```
#centers for clusters
k3 <- kmeans(udata2, centers = 3, nstart = 25)
##Isolating the data to Tufts University
library(dplyr)
library(stats)


Tufts_University <- filter(udata, College.Name == "Tufts University")
#Euclidean distance of this record from Cluster 1
dist(rbind(Tufts_University[, -c(1, 2, 3, 10)], k3$centers[1,]))
```

```
         1
2 29816.76
```

Hide

```
##Euclidean distance of this record from Cluster 2
dist(rbind(Tufts_University[, -c(1, 2, 3, 10)], k3$centers[2,]))
```

```
        1
2 29817.8
```

Hide

```
#Euclidean distance of this record from Cluster 3
dist(rbind(Tufts_University[, -c(1, 2, 3, 10)], k3$centers[3,]))
```

```
         1
2 29819.09
```

The Eucledian Distance from Tufts to Cluster1 is smaller i.e.,29816.76 compared to cluster2 and cluster3. Hence, Cluster1 is Closest to Tufts.

Impute the missing values for Tufts by taking the average of the cluster on those measurements.

Hide

```
NROW(udata)
```

```
[1] 1302
```

Hide

```
library(dplyr)
cluster1 <- filter(udata1, cluster == 1)
cluster1_Avg <- mean(cluster1[,c(10)])
Tufts_University[, c(10)] <- cluster1_Avg
Tufts_University[, c(10)]
```

```
[1] 2260.721
```

The Missing Value in tufts is 2260.721