Machine Learning - Midterm

msasnur@kent.edu (mailto:msasnur@kent.edu)

31/10/2019

```
#Machine Learning - Midterm Assignment#
#Email: msasnur@kent.edu#
#Date:31/10/2019#

library(readr)
library(ISLR)
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   0.8.3     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)
```

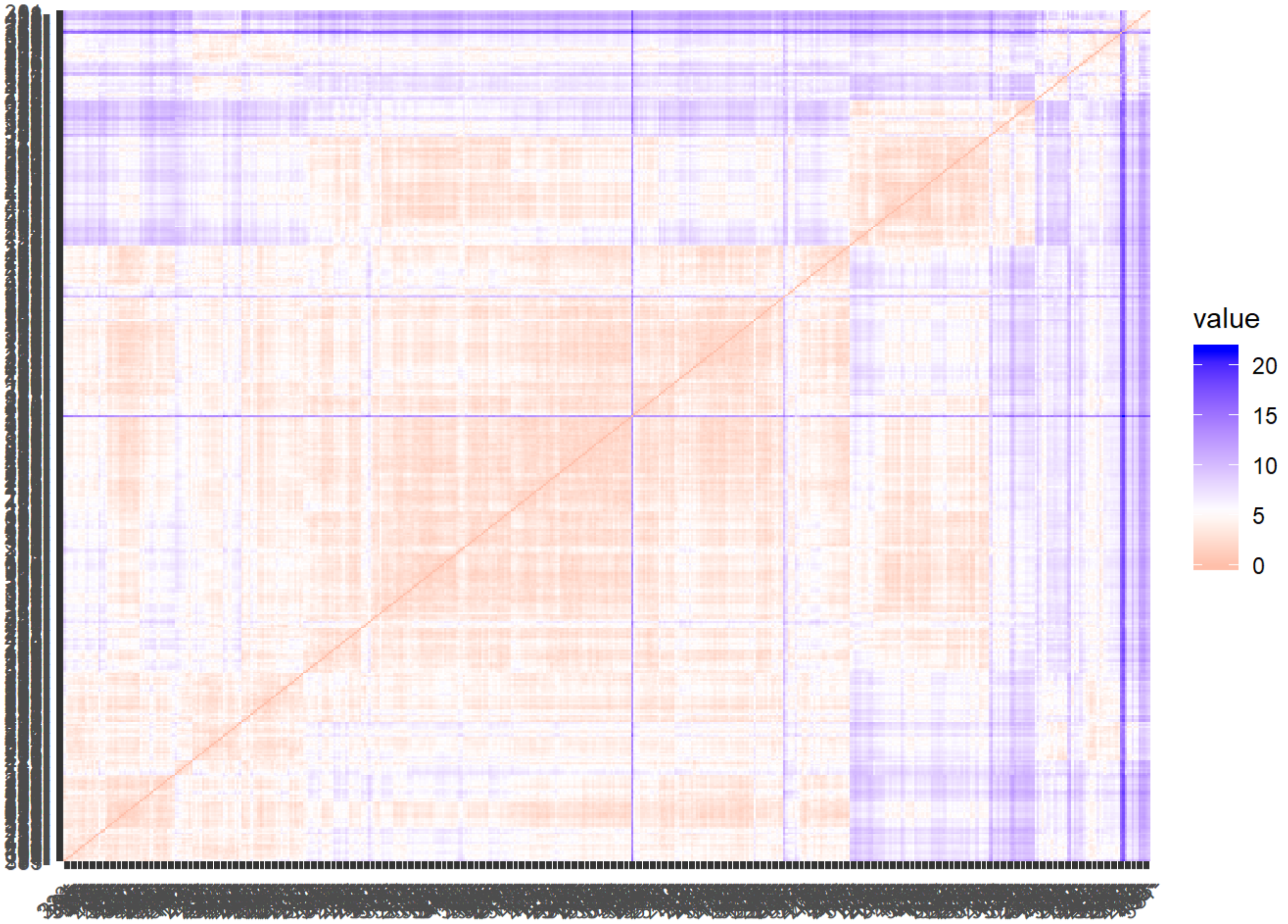
```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFC
Z
```

```
univ<-read_csv("Universities.csv")
```

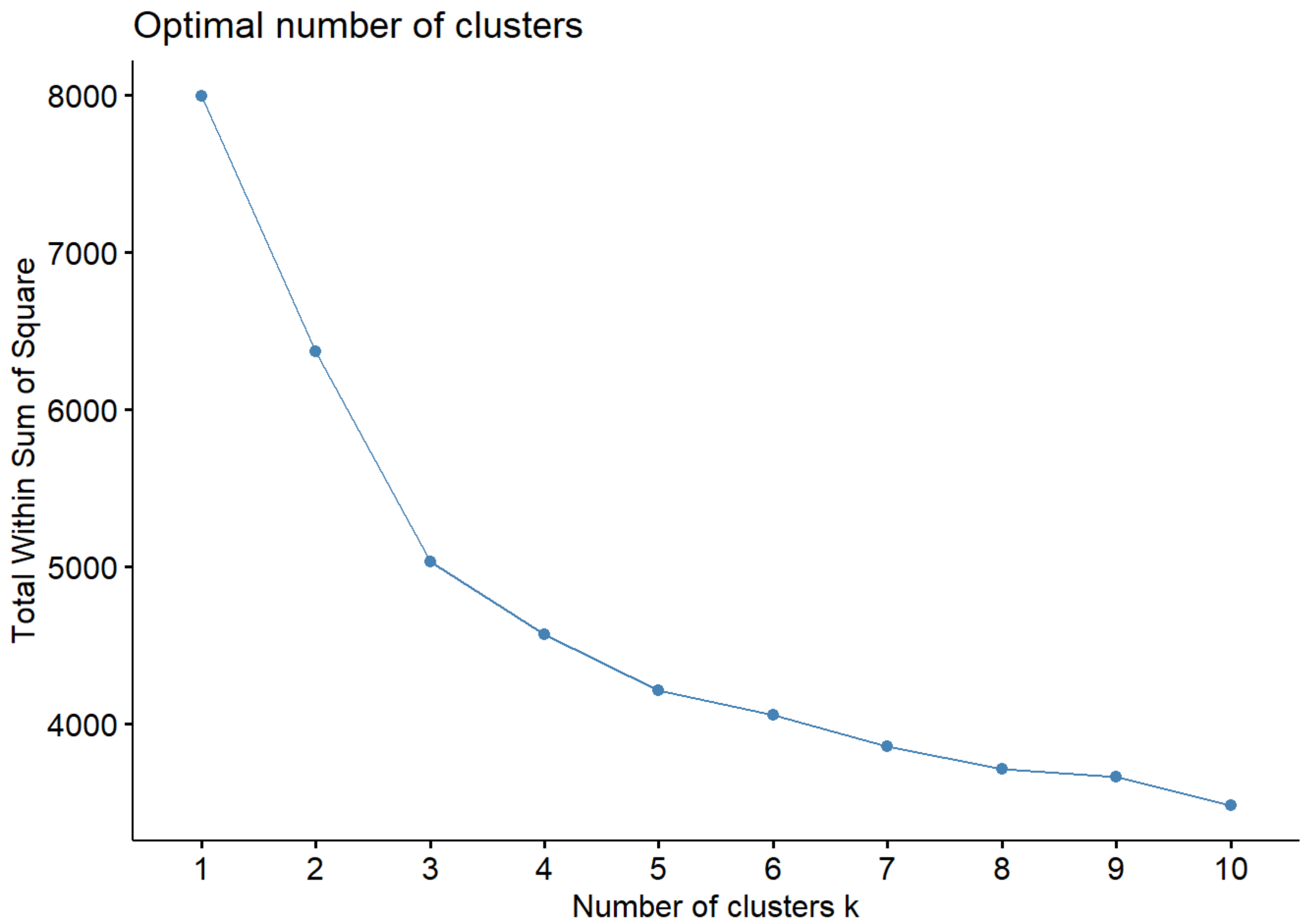
```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `College Name` = col_character(),
##   State = col_character()
## )
```

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

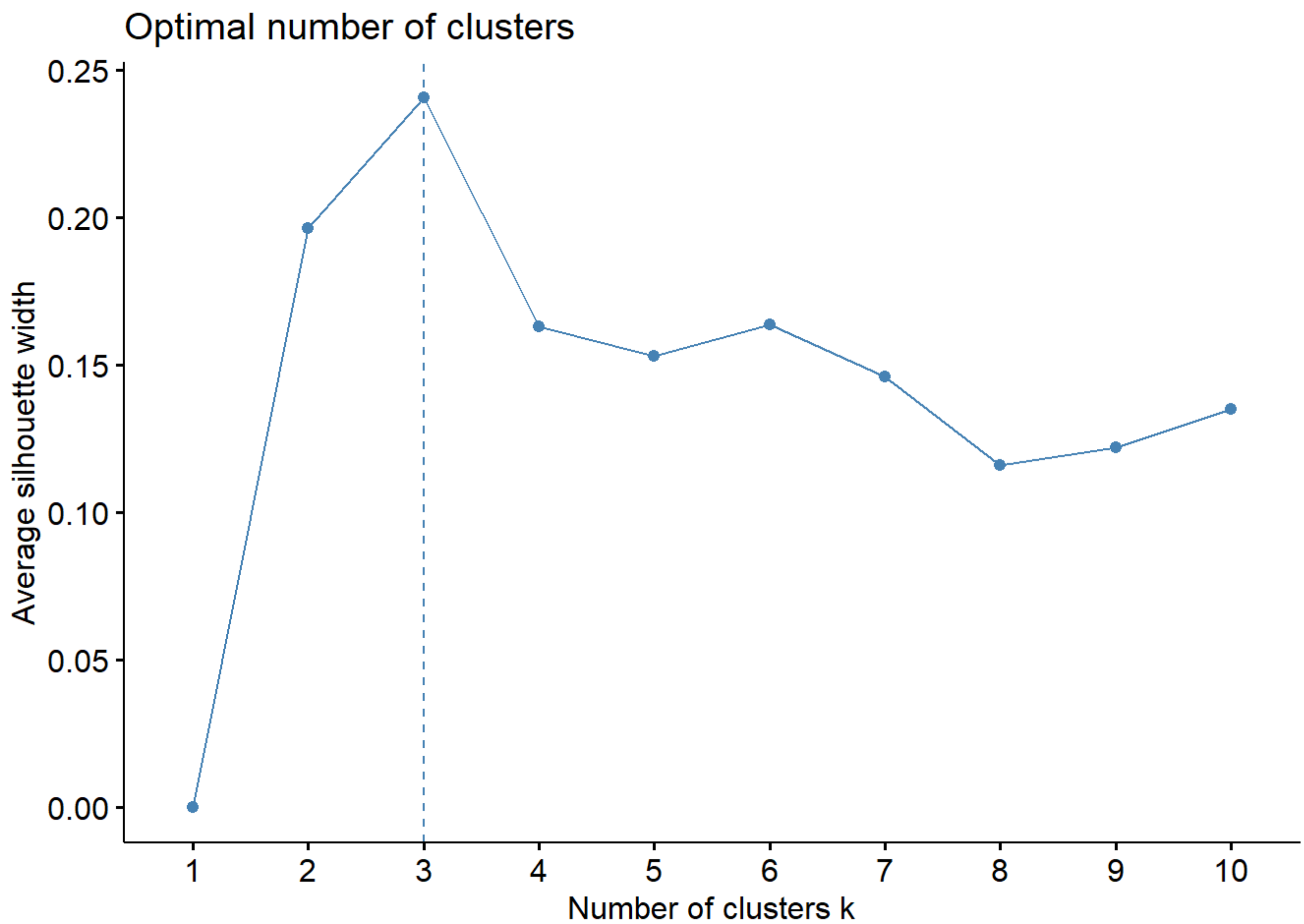
```
#Question 1
#Removing all records with missing measurements from the dataset
univ1<-na.omit(univ)
View(univ1)

#Question 2
# Scaling the data frame (z-score)
uni<-univ1[,c(-1,-2,-3)]
uni<-scale(uni)
distance <- get_dist(uni)
fviz_dist(distance)
```




```
# To find the best K value using Elbow Method and Silhouette Method
fviz_nbclust(uni,kmeans,method = "wss")
```




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



```
# From above two methods we have found out that the Best K value for cluster analysis is 3.

# To run kmeans clustering analysis
k3<- kmeans(uni, centers = 3, nstart = 25)
k3$centers #summary of cluster analysis
```

```
##   # 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
```

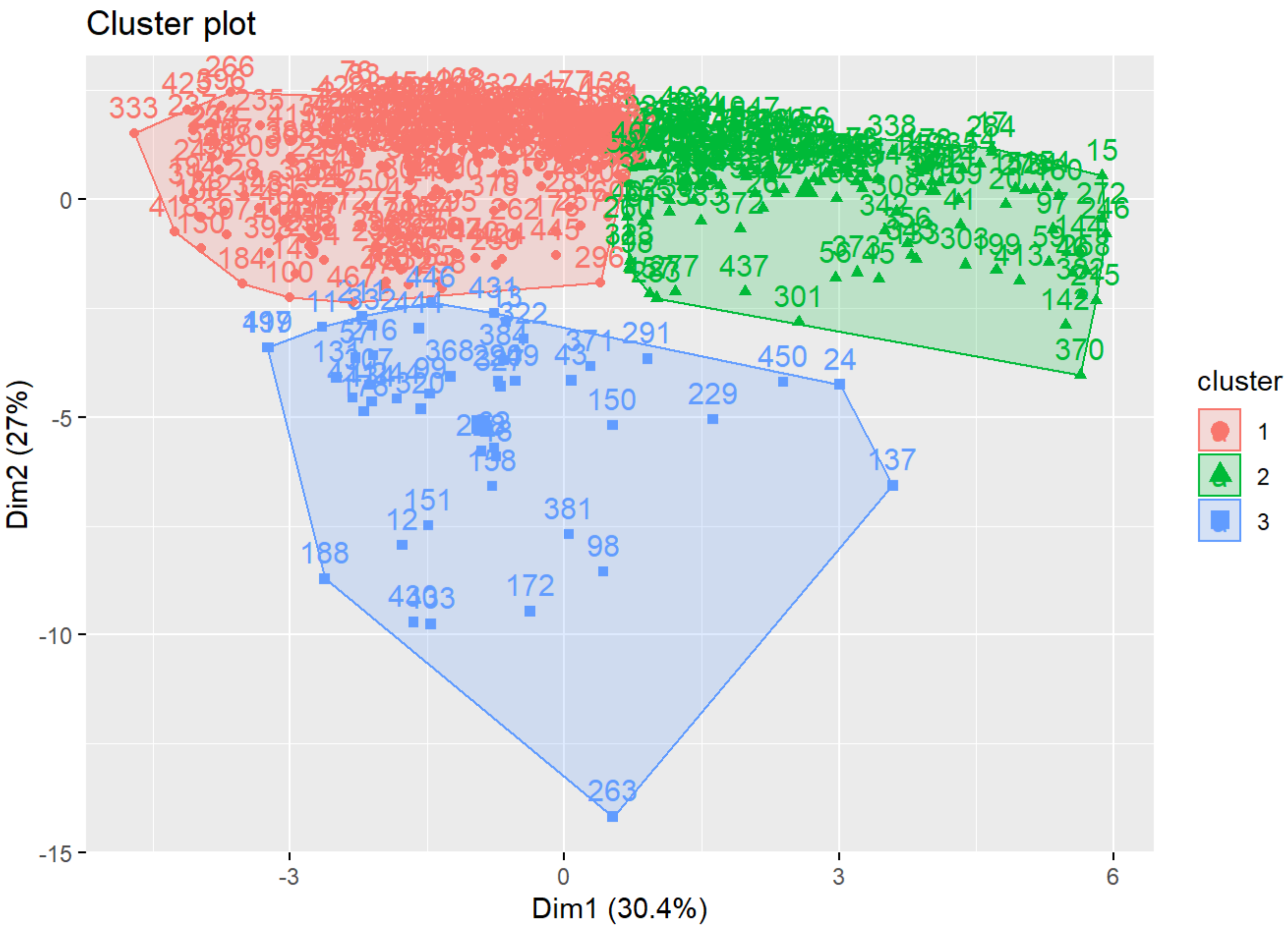
```
k3$size #Size of each cluster
```

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

```
k3$cluster[99] # To see which Cluster does 99th record belong to
```

```
## [1] 3
```

```
fviz_cluster(k3, data = uni) #Drawing cluster graph
```


Cluster plot
Dim2 (27%)
Dim1 (30.4%)
0
-5
-10
-15
-3
0
3
6
cluster
1
2
3

```
#Question 3
# As seen above from summary of k3$centers, we can observe the values for three different clu
sters.

# In cluster 3,
# Columns (Application Rejected,
#             Application Accepted,
#             New Student Enrolled,
#             Full Time underGrad,
#             Part Time underGrad,
#             Additional fees,
#             book costs,
#             estimated personal expenses,
#             student to faculty ratio)
#             have higher values and we can discern this pattern in cluster 3.

#In cluster 2,
# Columns (New student from 10%,
#           New student from top 25%,
#           in-state tution,
#           out-of-station tution,
#           Room,
#           Board,
#           Percentage of faculty with PhD,
#           Graduation Rate)
#           have higher values and we can discern that in Cluster 2.

#In cluster 1,
# Columns (Application Rejected,
#             Application Accepted,
#             New Student Enrolled,
#             New student from 10%,
#             New student from top 25%,
#             Full Time underGrad,
#             Part Time underGrad,
#             in-state tution,
#             out-of-station tution,
#             Room,
#             Board,)
#             have lower values and we can discern that in Cluster 1.

#Question 4
cat<-cbind(univ1[,c(1,2,3)],k3$cluster)
head(cat)
```

	College Name <chr>	State <chr>	Public (1)/ Private (2) <dbl>	k3$cluster <int>
1	Alaska Pacific University	AK	2	1
2	University of Alaska Southeast	AK	1	1
3	Birmingham-Southern College	AL	2	2
4	Huntingdon College	AL	2	1
5	Talladega College	AL	2	1

	College Name <chr>	State <chr>	Public (1)/ Private (2) <dbl>	k3$cluster <int>
6	University of Alabama at Birmingham	AL	1	1

6 rows

```
cat<-as.data.frame(cat)
cat$`Public (1)/ Private (2)`<-factor(univ1$`Public (1)/ Private (2)`, levels=c("1","2"), lab
els = c("Public","Private"))
Cluster1 <- cat[cat$`k3$cluster` == 1,]
View(Cluster1)
Cluster2 <- cat[cat$`k3$cluster` == 2,]
View(Cluster2)
Cluster3 <- cat[cat$`k3$cluster` == 3,]
View(Cluster3)

#After binding the categorical columns with clusters, we observe that
# Cluster 1 has data of both Public and Private Universities
# Cluster 2 has data belonging to Private universities
# Cluster 3 has data belonging to Public Universities mostly

# Using Pivot table we can get detailed information on number of universities belonging to ea
ch cluster,
# represented according to states. Separated by Public and Private Universities.
# We can also see the total number of Public and Private universities in each state.

library(pivottabler)
pt<-PivotTable$new()
pt$addData(cat)
pt$addColumnDataGroups('Public (1)/ Private (2)')
pt$addColumnDataGroups('k3$cluster')
pt$addRowDataGroups('State')
pt$defineCalculation(calculationName= 'Total', summariseExpression = 'n()')
pt$renderPivot()
```

	Public				Private				Total
	1	2	3	Total	1	2	3	Total	
AK	1			1	1			1	2
AL	1			1	2	1		3	4
AR					4			4	4
AZ			2	2					2
CA		1	1	2	3	9	1	13	15
CO	5			5		1		1	6
CT	2		1	3	1	6		7	10
DC						4		4	4
DE					1		1	2	2
FL			1	1	3	4		7	8
GA			1	1	4	2		6	7
HI	1			1					1
IA	1			1	15	2		17	18
ID					2			2	2
IL	2		2	4	5	6		11	15
IN	1			1	7	7		14	15
KS					7			7	7

KY	1			1	3	2		5	6
LA	1		1	2	1	2		3	5
MA	4		1	5	3	12	2	17	22
MD	1		1	2		1		1	3
ME	3			3	1	2		3	6
MI	1		2	3	6	4		10	13
MN	2		1	3	4	4		8	11
MO	2		1	3	10	2		12	15
MS	3			3	2			2	5
MT	1			1	1			1	2
NC	6		4	10	10	3		13	23
ND	4			4	1			1	5
NE	2		1	3	3	1		4	7
NH	1		1	2	3	1		4	6
NJ	6		1	7	3	3		6	13
NM					2			2	2
NY	10		2	12	8	18		26	38
OH			4	4	13	7		20	24
OK	2		1	3	3			3	6
OR					1	4		5	5
PA	4		3	7	15	20		35	42
RI			1	1	1	2		3	4
SC	2			2	5	2		7	9
SD	2			2	2			2	4
TN			1	1	11	3		14	15
TX	4		3	7	10	2	1	13	20
UT			1	1	1			1	2
VA	2	1	3	6	6	3		9	15
VT	3	1		4	2	1		3	7
WA						2		2	2
WI	2			2	3	4		7	9
WV					2			2	2
WY	1			1					1
Total	84	3	41	128	191	147	5	343	471

```r
#Question 5
# Using cluster.stats() function, we can get statastics of the all the clusters.
# This Statistics include Number of Cluster, Cluster Size, Diameter of each cluster, distanc
e, Separation.
library(fpc)
cluster.stats(distance,k3$cluster)
```

```
## $n
## [1] 471
##
## $cluster.number
## [1] 3
##
## $cluster.size
## [1] 275 150  46
##
## $min.cluster.size
## [1] 46
##
## $noisen
## [1] 0
##
## $diameter
## [1] 15.72735 10.83931 17.38478
##
## $average.distance
## [1] 4.102453 4.113867 6.235578
##
## $median.distance
## [1] 4.019750 3.907029 5.489743
##
## $separation
## [1] 1.054636 1.054636 2.106758
##
## $average.toother
## [1] 6.102039 5.977532 7.918952
##
## $separation.matrix
##          [,1]     [,2]     [,3]
## [1,] 0.000000 1.054636 2.106758
## [2,] 1.054636 0.000000 2.769109
## [3,] 2.106758 2.769109 0.000000
##
## $ave.between.matrix
##          [,1]     [,2]     [,3]
## [1,] 0.000000 5.598819 7.742976
## [2,] 5.598819 0.000000 8.241575
## [3,] 7.742976 8.241575 0.000000
##
## $average.between
## [1] 6.344849
##
## $average.within
## [1] 4.314419
##
## $n.between
## [1] 60800
##
## $n.within
## [1] 49885
##
## $max.diameter
## [1] 17.38478
##
```

```
## $min.separation
## [1] 1.054636
##
## $within.cluster.ss
## [1] 5031.914
##
## $clus.avg.silwidths
##         1         2         3
## 0.2503594 0.2484554 0.1560818
##
## $avg.silwidth
## [1] 0.2405454
##
## $g2
## NULL
##
## $g3
## NULL
##
## $pearsongamma
## [1] 0.4736057
##
## $dunn
## [1] 0.06066437
##
## $dunn2
## [1] 0.897883
##
## $entropy
## [1] 0.9057607
##
## $wb.ratio
## [1] 0.6799877
##
## $ch
## [1] 137.5604
##
## $cwidegap
## [1] 8.247873 6.747930 9.655971
##
## $widestgap
## [1] 9.655971
##
## $sindex
## [1] 1.523979
##
## $corrected.rand
## NULL
##
## $vi
## NULL
```

```
#Question 6
# Replacing the NA value
univ$`# PT undergrad`[is.na(univ$'# PT undergrad')] <- mean(univ$'# PT undergrad',na.rm = TRU
E)
tuftuni<-univ[476,]
summary(univ$`# PT undergrad`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   136.5   487.5  1081.5  1286.0 21836.0
```

```
x<- rbind(univ1,tuftuni) #Binding the Tuft University record to our Dataset without NA values
y<- scale(x[,c(-1,-2,-3)]) #Normalizing the dataset
k.tuft<-kmeans(y,centers = 3,nstart = 25) #Performing Cluster Analysis on Dataset
k.tuft$cluster
```

```
##   [1] 1 1 3 1 1 1 1 1 1 1 2 2 2 3 3 3 3 3 1 3 1 3 3 2 3 3 1 1 3 1 1 1 1 1 3
##  [36] 1 3 3 3 3 3 1 2 3 3 3 3 2 1 1 3 1 1 3 3 3 2 1 3 1 3 2 1 1 1 1 1 1 3 1
##  [71] 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3 1 3 1 1 1 1 3 2 2 1 1 3 3 3 1
## [106] 1 1 1 1 1 1 3 3 3 3 1 1 1 1 1 1 1 1 3 1 1 1 3 1 1 2 3 3 3 3 3 2 1 3 3
## [141] 1 3 1 3 1 1 1 1 3 2 2 3 3 3 3 3 1 2 1 3 3 1 1 1 1 3 3 1 3 3 1 2 1 1 1
## [176] 2 1 3 1 1 3 3 3 1 1 1 1 2 1 1 1 1 1 1 1 1 2 3 3 1 1 1 1 1 1 1 1 1 1 1
## [211] 2 1 1 3 3 2 1 1 1 1 1 1 1 1 1 1 3 2 2 1 1 1 1 1 1 1 1 1 3 1 1 1 1 2 1
## [246] 3 1 1 2 1 1 3 1 1 1 1 1 3 1 3 1 1 2 1 1 1 3 1 1 3 3 3 3 1 3 3 3 1 3 3
## [281] 1 1 3 3 3 3 1 1 3 2 2 1 1 1 1 1 1 1 1 1 3 3 3 1 1 1 2 3 1 1 1 3 1 3 1
## [316] 3 1 1 3 2 1 2 3 1 3 1 2 1 1 1 1 2 1 1 1 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1
## [351] 3 3 1 3 3 3 1 1 3 3 1 1 1 1 1 1 3 2 1 3 2 3 3 3 1 1 1 1 1 1 2 3 3 2 1
## [386] 1 1 1 1 1 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 2 3 3 1 1 1 1 1 2 1
## [421] 1 1 1 1 1 1 3 3 1 2 2 1 2 2 1 1 3 1 3 1 3 1 1 2 1 2 1 1 3 2 1 1 1 1 3
## [456] 1 3 3 3 3 1 3 3 3 1 1 1 1 1 1 1 3
```

```
k.tuft$centers
```

```
##   # appli. rec'd # appl. accepted # new stud. enrolled
## 1    -0.36178481      -0.35072087           -0.3183527
## 2     1.97905301       2.23008958            2.4457813
## 3     0.05598931      -0.04063498           -0.1652911
##   % new stud. from top 10% % new stud. from top 25% # FT undergrad
## 1               -0.5046909               -0.5153888     -0.2959828
## 2                0.1294809                0.2505136      2.5249258
## 3                0.8796945                0.8623066     -0.2301411
##   # PT undergrad in-state tuition out-of-state tuition        room
## 1     -0.1222831       -0.4065412           -0.5289356 -0.36118215
## 2      1.7500917       -1.0512232           -0.4944827 -0.04124618
## 3     -0.3104396        1.0606299            1.1139304  0.67034712
##        board   add. fees estim. book costs estim. personal $ % fac. w/PHD
## 1 -0.3964868 -0.05911977       -0.06694360         0.0605963   -0.5347161
## 2 -0.1774061  0.49504158        0.16307690         0.9404620    0.6797982
## 3  0.7761229 -0.04313891        0.07223809        -0.3968559    0.7667298
##   stud./fac. ratio Graduation rate
## 1        0.2831120      -0.4197344
## 2        0.6160664      -0.2566056
## 3       -0.7032771       0.8425881
```

```
which(grepl("Tufts University",x$`College Name`)) #To find the index of Tuft University recor
d
```

```
## [1] 472
```

```
k.tuft$cluster[472] #To find cluster value in which Tuft University belongs to, using the ind
ex value
```

```
## [1] 3
```

```
# From above results, we can see that Tufts University belongs to Cluster 3 and its indexed a
t 472nd record.
```