# Midterm Exam

Steve Spence

10/24/2019

## Overview of Packages and Dataset

The dataset on American College and University Rankings contains information on 1,302 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).

First, we will load all of the packages that will be required for this problem. Specifically, "ISLR", "caret", "tidyverse", "factoextra", "ggplot2", "proxy", and "dplyr" will be loaded for this problem.

Next, we will import the "university" data set into the RStudio environment.

```
# Import data set from BlackBoard into the RStudio environment

university <- read.csv("university.csv")
```

A review of the structure of the data set will be displayed to review the data set.

```
# Investigate the structure of the data set

str(university)

## 'data.frame':    1302 obs. of  20 variables:
##  $ College.Name              : Factor w/ 1274 levels "Abilene Christian
University",..: 8 1004 1005 1003 6 327 1084 7 44 88 ...
##  $ State                     : Factor w/ 51 levels "AK","AL","AR",..: 1 1 1
1 2 2 2 2 2 2 ...
##  $ Public..1...Private..2.  : int  2 1 1 1 1 2 1 1 1 2 ...
##  $ X..appli..rec.d           : int  193 1852 146 2065 2817 345 1351 4639
7548 805 ...
##  $ X..appl..accepted         : int  146 1427 117 1598 1920 320 892 3272
6791 588 ...
##  $ X..new.stud..enrolled     : int  55 928 89 1162 984 179 570 1278 3070
287 ...
##  $ X..new.stud..from.top.10.: int  16 NA 4 NA NA NA 18 NA 25 67 ...
##  $ X..new.stud..from.top.25.: int  44 NA 24 NA NA 27 78 NA 57 88 ...
##  $ X..FT.undergrad           : int  249 3885 492 6209 3958 1367 2385 4051
16262 1376 ...
##  $ X..PT.undergrad           : int  869 4519 1849 10537 305 578 331 405
1716 207 ...
```

```
##  $ in.state.tuition         : int  7560 1742 1742 1742 1700 5600 2220 1500
2100 11660 ...
##  $ out.of.state.tuition     : int  7560 5226 5226 5226 3400 5600 4440 3000
6300 11660 ...
##  $ room                     : int  1620 1800 2514 2600 1108 1550 NA 1960
NA 2050 ...
##  $ board                    : int  2500 1790 2250 2520 1442 1700 NA NA NA
2430 ...
##  $ add..fees                : int  130 155 34 114 155 300 124 84 NA 120
...
##  $ estim..book.costs        : int  800 650 500 580 500 350 300 500 600 400
...
##  $ estim..personal..        : int  1500 2304 1162 1260 850 NA 600 NA 1908
900 ...
##  $ X..fac..w.PHD            : int  76 67 39 48 53 52 72 48 85 74 ...
##  $ stud..fac..ratio         : num  11.9 10 9.5 13.7 14.3 32.8 18.9 18.7
16.7 14 ...
##  $ Graduation.rate          : int  15 NA 39 NA 40 55 51 15 69 72 ...
```

From the structure of the dataset, we can see that the data was imported with 2 variables as factors, 1 as numeric, and the remaining 17 as integer. The variable “Public..1..Private..2.” will need to be converted to a factor with two levels.

```
# Convert the variable "Public..1..Private..2." to a factor with two levels

university$Public..1...Private..2. <-
as.factor(university$Public..1...Private..2.)

# Return the structure of the converted variable to confirm

str(university$Public..1...Private..2.)

##  Factor w/ 2 levels "1","2": 2 1 1 1 1 2 1 1 1 2 ...
```

The next chunk of code will return a summary of all the variables in the data set.

```
# Investigate the summary of the data set

summary(university)

##             College.Name      State     Public..1...Private..2.
##  Bethel College   :   4    NY     :101   1:470
##  Concordia College:   4    PA     : 83   2:832
##  Trinity College  :   4    CA     : 70
##  Columbia College :   3    TX     : 60
##  Union College    :   3    MA     : 56
##  Augustana College:   2    OH     : 52
##  (Other)          :1282    (Other):880
##  X..appli..rec.d   X..appl..accepted X..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
##  X..new.stud..from.top.10. X..new.stud..from.top.25. X..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
##  X..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
##  X..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
```

From this summary, it can be concluded that 17 variables contained missing values, designated as “NA’s”. The first part of the assignment will address this issue.

## Midterm Task 1

1. Remove All Records With Missing Measurements from the Dataset.

The first task of the assignment is to remove these missing values. We will do this by utlizing the “complete.cases” function from base R.

```
# Remove all variables with missing values in the dataset

university_na_removed <- university[complete.cases(university), ]
```

```
# Return summary of the dataset to confirm all NA's have been removed

summary(university_na_removed)
```

```
##                     College.Name     State     Public..1...Private..2.
##  Trinity College         :  4   PA     : 42   1:128                  
##  Augustana College       :  2   NY     : 38   2:343                  
##  Monmouth College        :  2   OH     : 24                          
##  University of St. Thomas:  2   NC     : 23                          
##  Westminster College     :  2   MA     : 22                          
##  Adams State College     :  1   TX     : 20                          
##  (Other)                 :458   (Other):302                          
##  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. X..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  
##                                                                     
##  X..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     
##                                                                     
##  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  
##
```

After executing this command, there are a total of 471 observations remaining. Therefore, 830 records were removed for having at least 1 missing value.

## Midterm Task 2

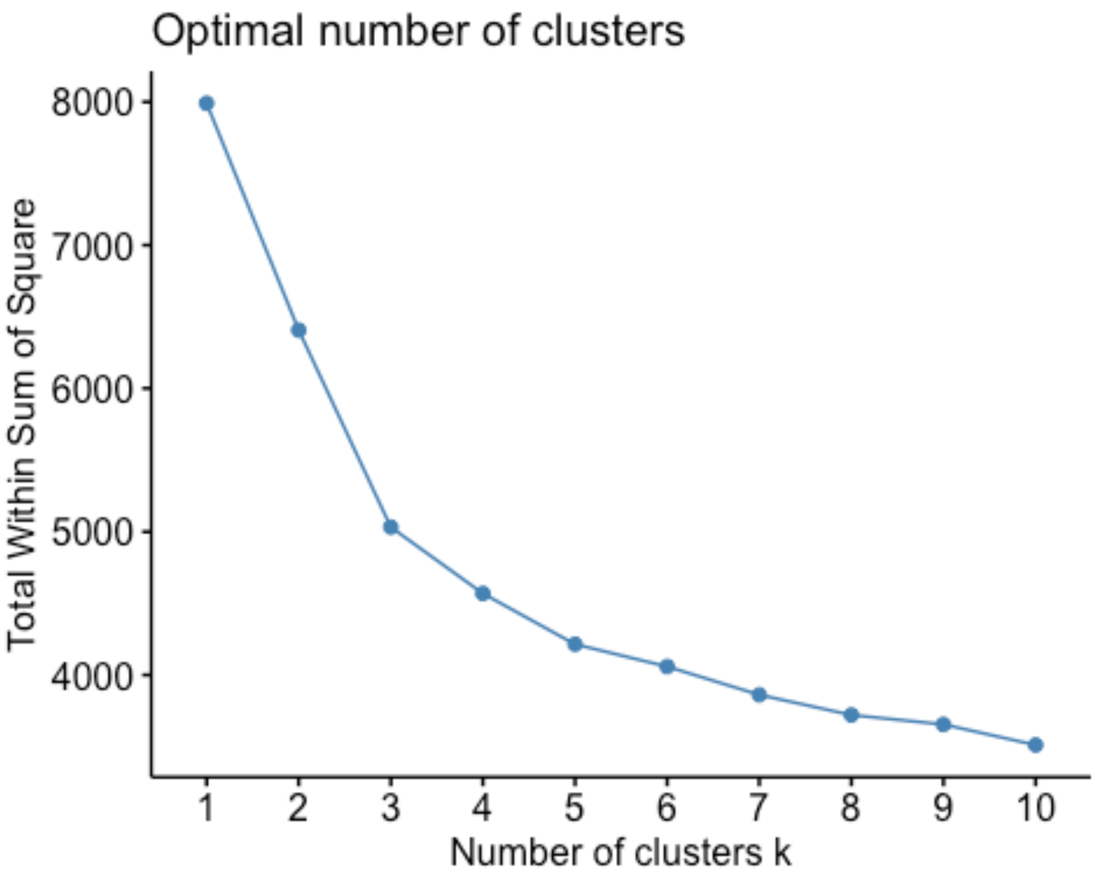
2. For All the Continuous Measurements, Run K-Means clustering. Make sure to normalize the measurements. How many clusters seem reasonable for describing these data? What was your optimal K?

Per the outlined task, the continuous variables will need to be normalized before we place them into the K-means clustering algorithm.

```
# Create a new variable "university_scaled" to perform the normalization
function on

university_scaled <- university_na_removed

# Scale the numeric variables in the dataset

university_scaled[ , c(4:20)] <- scale(university_scaled[ , c(4:20)])
```

Now that the numeric variables are normalized, we will run the k-means clustering algorithm; however, the optimal number of clusters will need to be determined first. This will be completed via the “elbow method”, which looks at the within-cluster sum square (WSS).

```
# Determine the optimal number of clusters for the dataset

fviz_nbclust(university_scaled[ , c(4:20)], kmeans, method = "wss")
```



From this method, it appears that a k value of 4 will be the optimal value to use. Therefore, the kmeans algorithm will now be run with this value.

```
# Set the seed for randomized functions

set.seed(102419)

# k-means algorithm with the numerical variables

km1 <- kmeans(university_scaled[ , c(4:20)], centers = 4, nstart = 25)
```

## Midterm Task 3

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

To get the insights from the k-means algorithm, we will need to look at the centers, sizes, and compare the statistics for insights. This will utilize the "fviz_cluster" function along with others summary views.

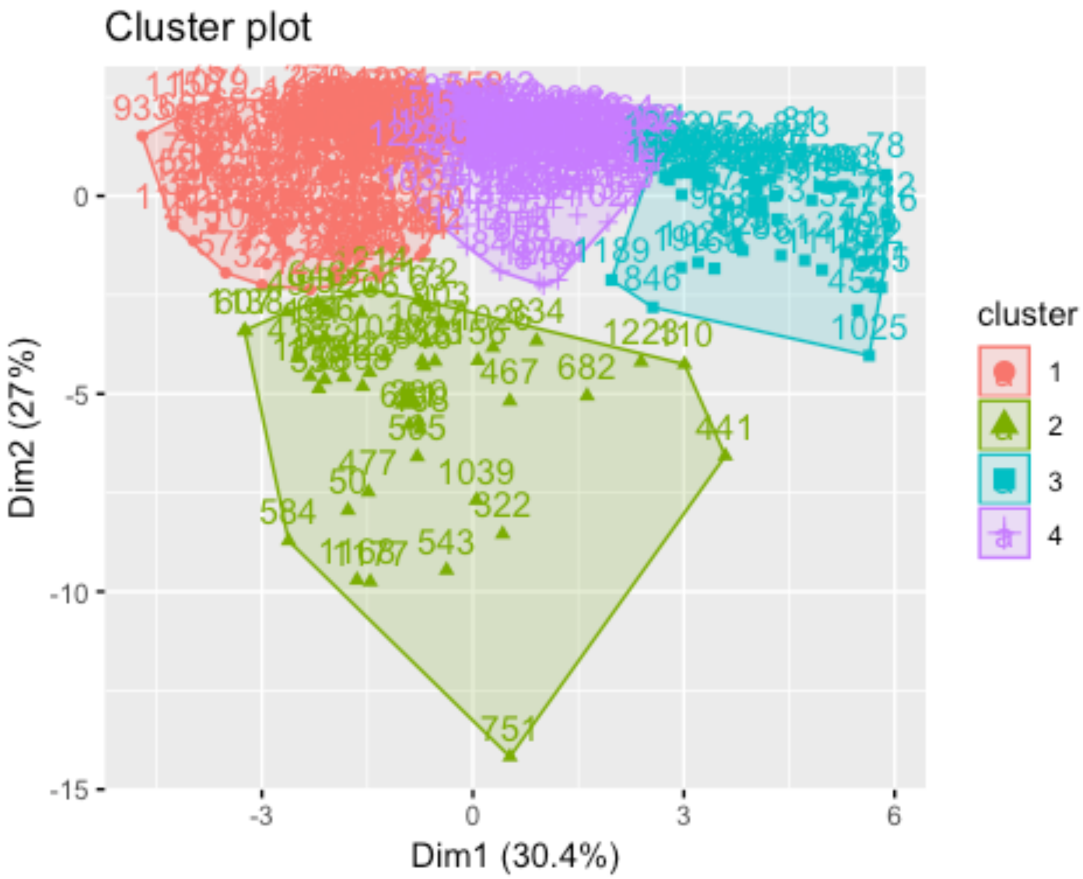
```
# Return the centroids for the variables

km1$centers
```

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

```
# Return the size of each cluster

km1$size
```

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

```
# Visualize the k-means output

fviz_cluster(km1, data = university_scaled[ , c(4:20)])
```



Insights for Cluster 1:

These Universities have fewer applications, acceptabed students, and enrolled students than the other clusters. There are not as many students from the top of their classes, and the tuition rates are lower than the other university clusters. Additionally, their faculty does not have as many with phDs, and the graduation rate is lower than all the other clusters.

Insights for Cluster 2:

Universities with higher number of applications, accepted students, and enrolled students. Additionally, these universities have the lower tuition and room/board, as well as a lower graduation rate than average.

Insights for Cluster 3:

Universities with a lot of applications, but lower acceptance and enrollment rate. These Universities have more students that were top of their class, as well as higher tuition, room, and board costs. The faculity has a lot more professors with phDs, as well as a lower student-to-faculty ratio. Lastly, the graduation rate at these universities has a higher graduation rate.

Insights for Cluster 4:

These Universities have fewer applications, accepted students, and enrolled students. Additionally, all their other metrics are found around the mean for tuition, room, board, graduation rate, etc.

## Midterm Task 4

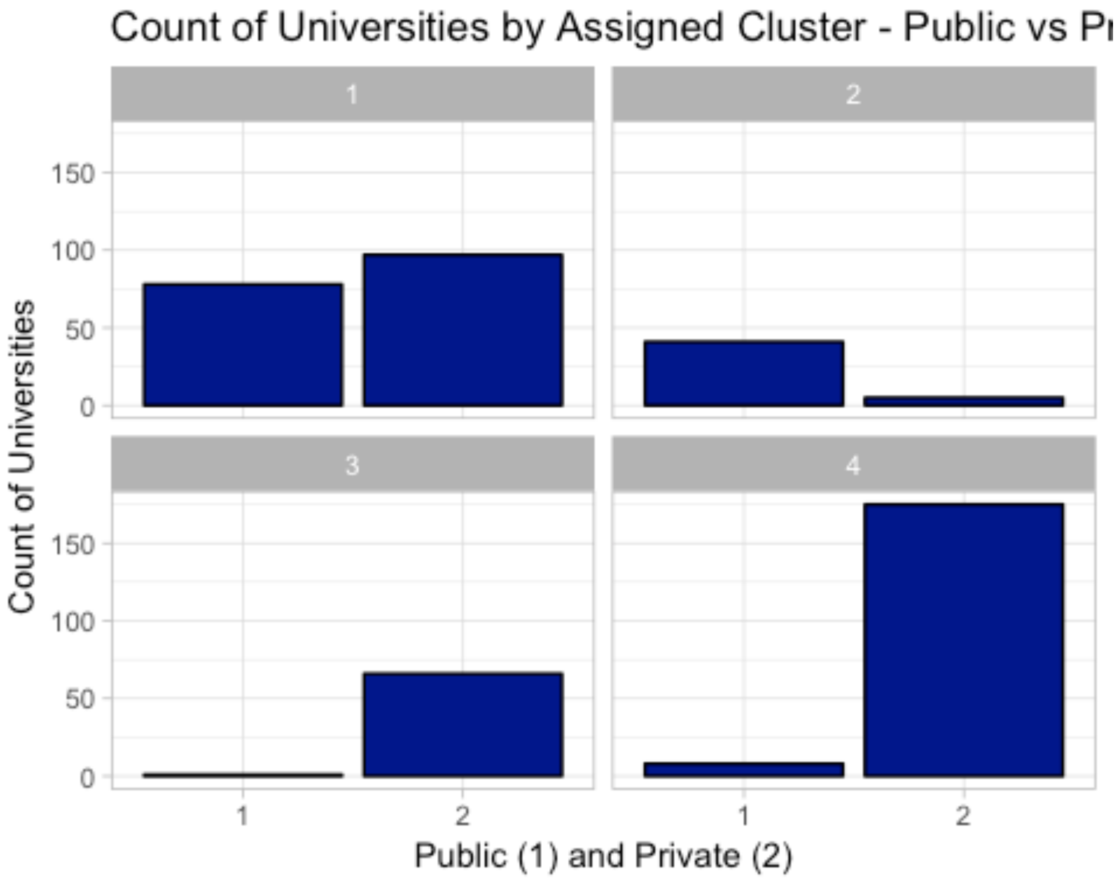
4. Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information?

The gain these insights into the data, the assigned clusters will have to be bound to the datafame. Then ggplot can be used to visual all the clusters with these categorical variables to gain insights.

```
# The assigned cluster group will be bound to the dataframe

university_clustered <- cbind(university_scaled, km1$cluster)
```

Next, the chart of clusters will be compared to the categorical variables that were left out (State and Private/Public) with ggplot2.

```
# Create plot private and public universities, faceted by their assgined
cluster.

ggplot(data = university_clustered) +
  geom_bar(mapping = aes(university_clustered$Public..1...Private..2.),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(university_clustered$`km1$cluster`)) +
  labs(title = "Count of Universities by Assigned Cluster - Public vs
Private") +
  labs(x = "Public (1) and Private (2)", y = "Count of Universities") +
  theme_light()
```



From these charts, it can be seen that clusters 3 and 4 mainly contain private universities, cluster 2 is mainly public universities, and cluster 1 is a fairly equal split between the two.

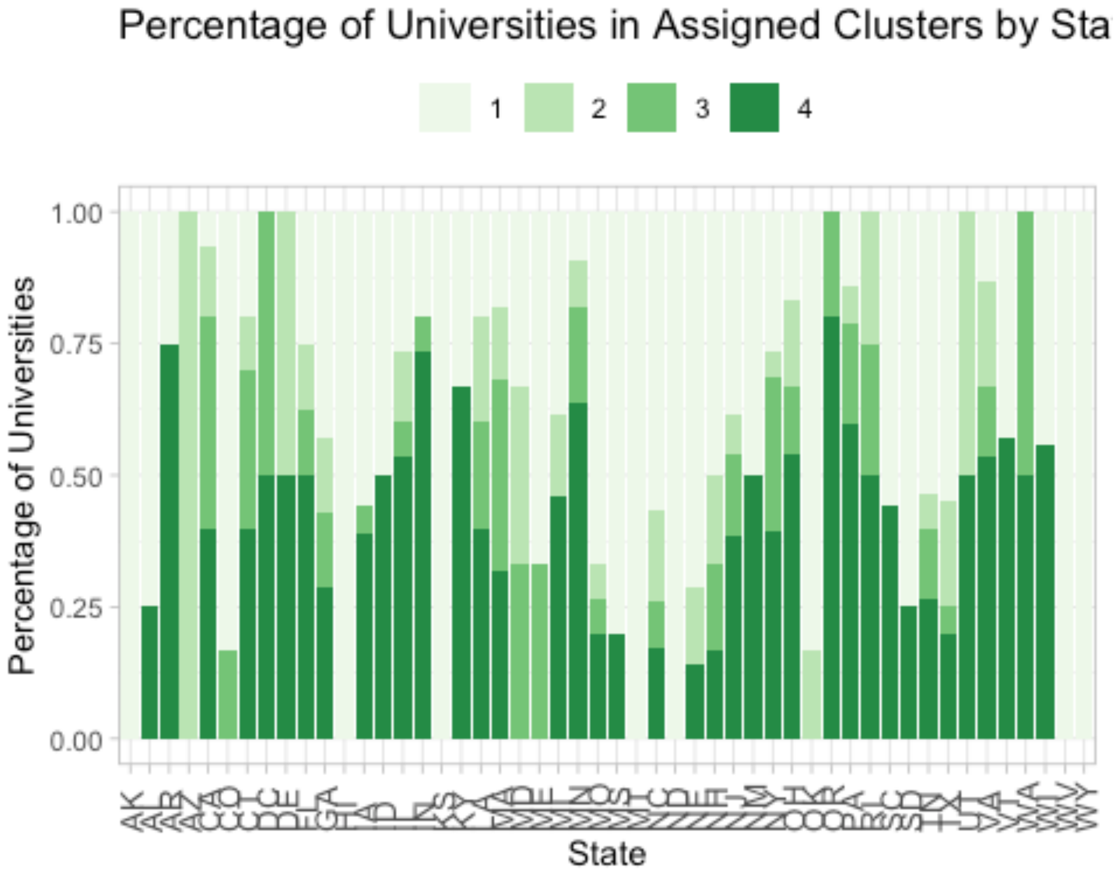
The next charts will look at ways to review the state variable for additional insights.

```
# Convert the assigned cluster values to a factor with four levels for
plotting

university_clustered$`km1$cluster` <-
as.factor(university_clustered$`km1$cluster`)

str(university_clustered$`km1$cluster`)

##  Factor w/ 4 levels "1","2","3","4": 1 1 4 1 1 1 4 4 4 1 ...
```

The first chart for states will look at percentage of universities in a cluster by state.

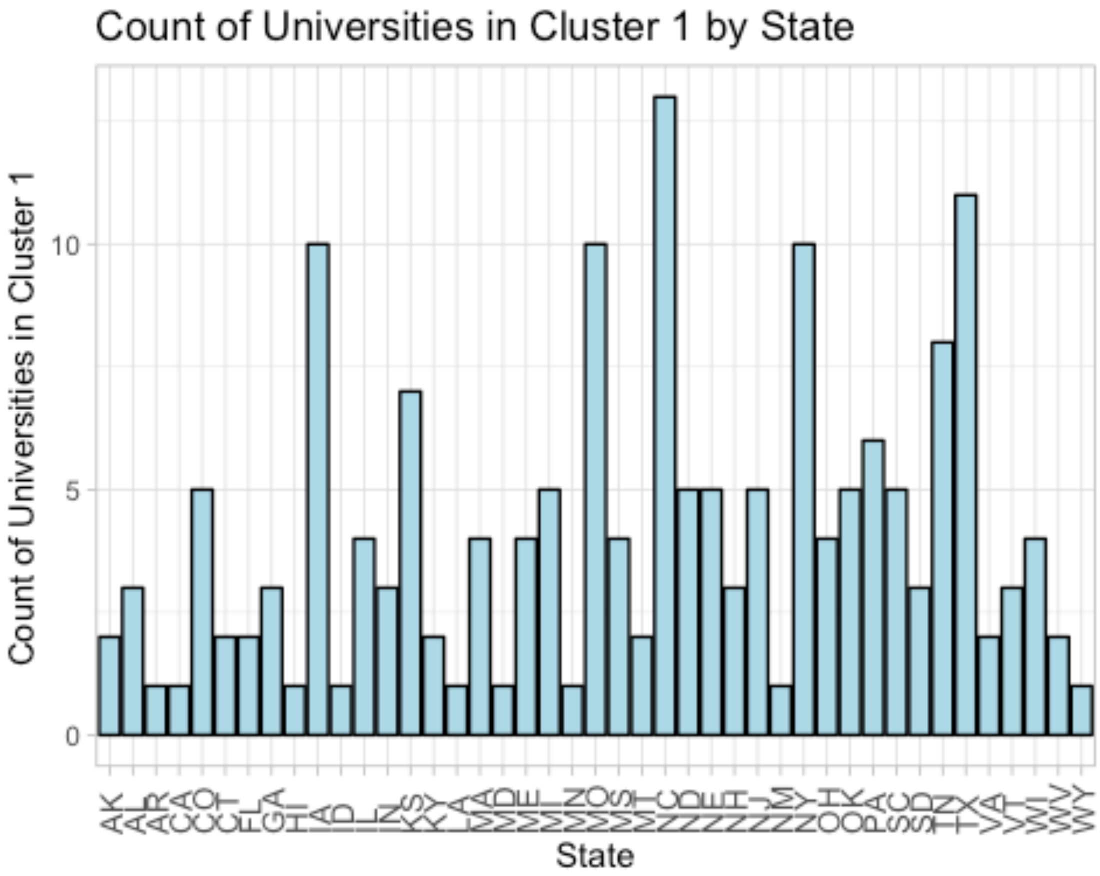
```
# Create plot private and public universities, faceted by their assgined
cluster.

ggplot(data = university_clustered) +
  geom_bar(mapping = aes(State, fill = `km1$cluster`), position = "fill") +
  labs(title = "Percentage of Universities in Assigned Clusters by State") +
  labs(x = "State", y = "Percentage of Universities") +
  theme_light() +
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = 0, vjust =
.60), legend.position="top") +
  scale_fill_brewer(palette = "Set4") +
  guides(fill = guide_legend(title = NULL))
```

```
## Warning in pal_name(palette, type): Unknown palette Set4
```



The additional charts below will also break it down into individual charts for each cluster to view the count of universities by state to get another view of the breakdown.

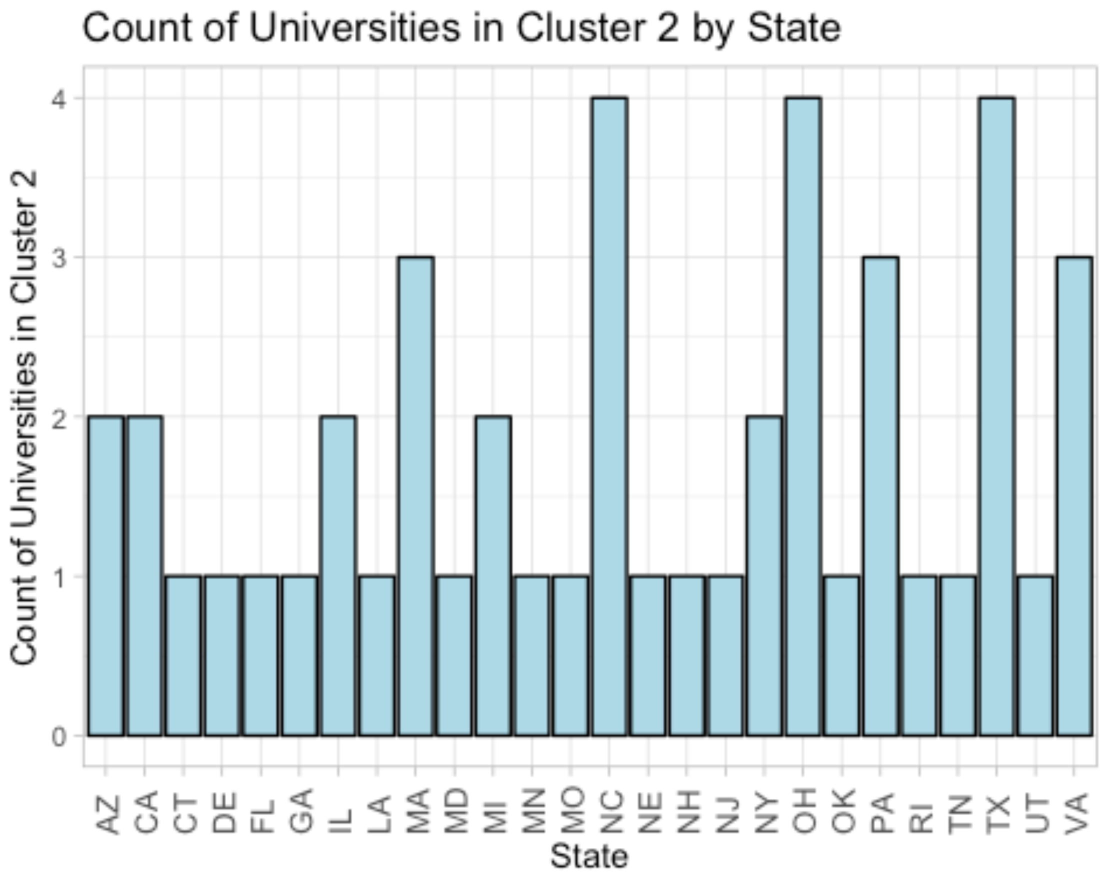
```
# Create plot private and public universities, faceted by their assgined
cluster.

university_clustered_1 <- subset(university_clustered, km1$cluster == 1)

# Create plot private and public universities, faceted by their assgined
cluster.

ggplot(data = university_clustered_1, aes(x = State)) +
  geom_bar(col = "black",
           fill = "lightblue") +
  labs(title = "Count of Universities in Cluster 1 by State") +
  labs(x = "State", y = "Count of Universities in Cluster 1") +
  theme_light() +
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = 0, vjust =
.60))
```



Top 5 States by count of universities in cluster 1:

1. North Carolina
2. Texas
3. Iowa
4. Missouri
5. New York

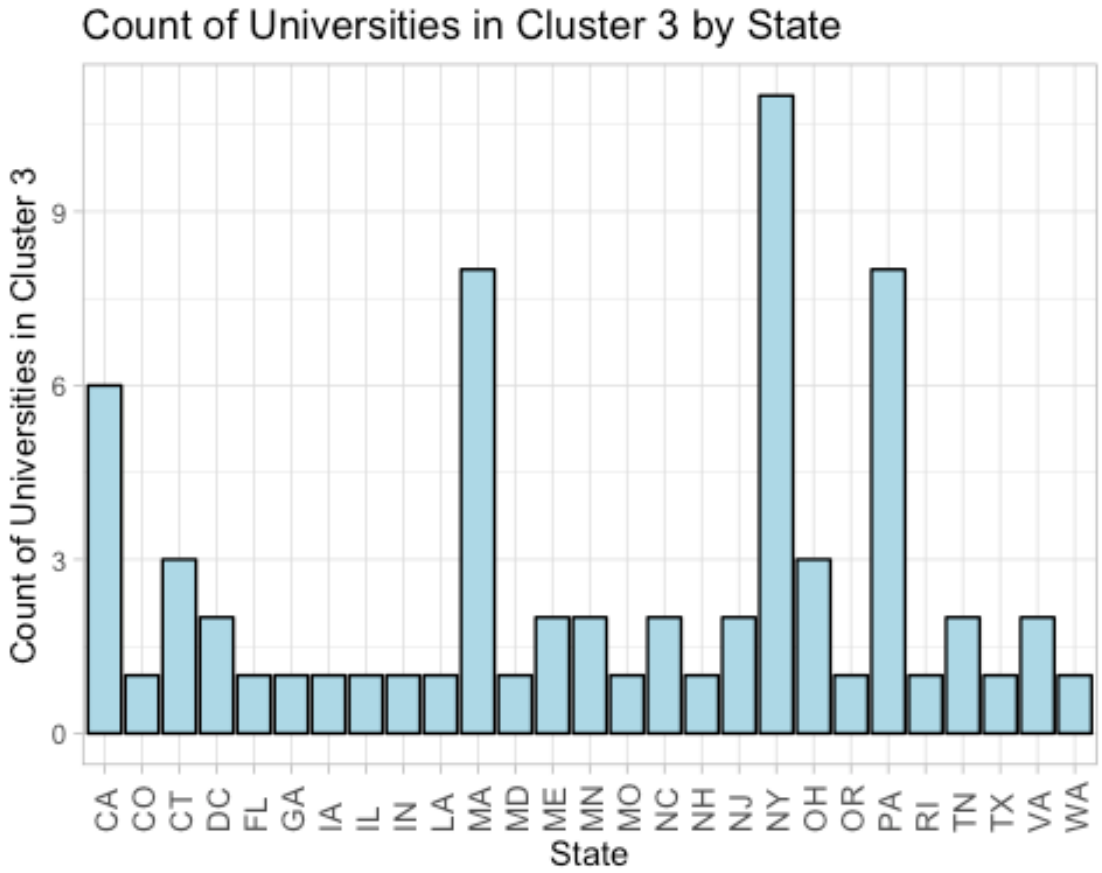
```
# Create plot private and public universities, faceted by their assgined
cluster.

university_clustered_2 <- subset(university_clustered, km1$cluster == 2)

# Create plot private and public universities, faceted by their assgined
cluster.

ggplot(data = university_clustered_2, aes(x = State)) +
  geom_bar(col = "black",
          fill = "lightblue") +
  labs(title = "Count of Universities in Cluster 2 by State") +
  labs(x = "State", y = "Count of Universities in Cluster 2") +
  theme_light() +
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = 0, vjust =
.60))
```



Top 5 States by count of universities in cluster 2:

1. North Carolina
2. Ohio
3. Texas 4(T). Massachusetts 4(T). Pennsylvania 4(T). Virginia

```
# Create plot private and public universities, faceted by their assgined
cluster.

university_clustered_3 <- subset(university_clustered, km1$cluster == 3)

# Create plot private and public universities, faceted by their assgined
cluster.

ggplot(data = university_clustered_3, aes(x = State)) +
  geom_bar(col = "black",
           fill = "lightblue") +
  labs(title = "Count of Universities in Cluster 3 by State") +
  labs(x = "State", y = "Count of Universities in Cluster 3") +
  theme_light() +
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = 0, vjust =
.60))
```

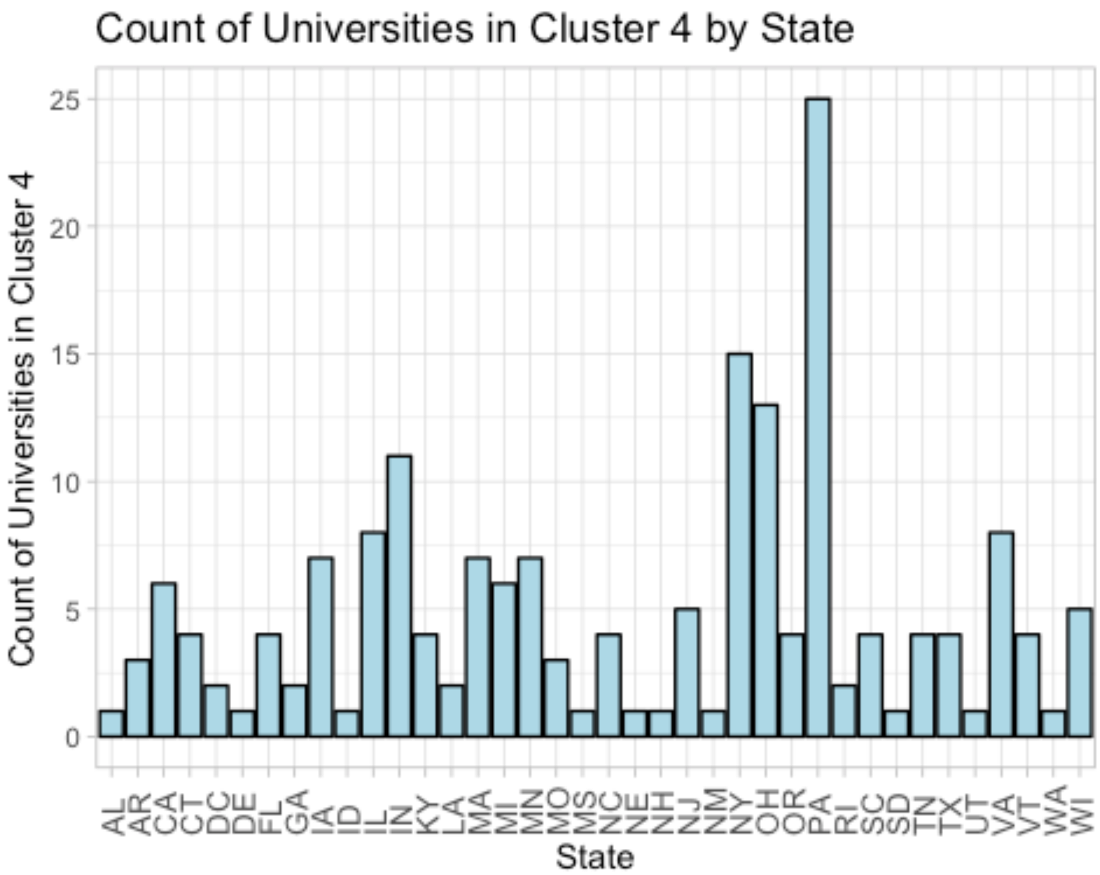


Top 5 States by count of universities in cluster 3:

1. New York
2. Massachusetts
3. Pennsylvannia
4. California
5. Connecticut

```
# Create plot private and public universities, faceted by their assgined
cluster.

university_clustered_4 <- subset(university_clustered, km1$cluster == 4)

# Create plot private and public universities, faceted by their assgined
cluster.
```

```
ggplot(data = university_clustered_4, aes(x = State)) +
  geom_bar(col = "black",
           fill = "lightblue") +
  labs(title = "Count of Universities in Cluster 4 by State") +
  labs(x = "State", y = "Count of Universities in Cluster 4") +
  theme_light() +
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = 0, vjust =
.60))
```



Top 5 States by count of universities in cluster 4:

1. Pennsylvania
2. New York
3. Ohio
4. Indiana
5. Illinois

## Midterm Task 5

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

Other external domain knowledge of the university environment can explain some of the cluster groupings:

Cluster 3 contains mostly private universities with high achieving students, low acceptance rates, and high tuition rates in New York, Massachusetts, Pennsylvannia, California, Connecticut. This aligns with this cluster containing ivy league schools (MIT, Harvard, Stanford, UPenn, NYU, etc.) and other well-known top universities. They typically get a lot of applications and are very selective with their enrollment of students. These schools also have very prestigeous staff members and have higher tuition rates as a result.

Cluster 2 contains mainly public universities, which typically get more students applying and enrolling to them. Additionally, public schools are typically less expensive. The opposite is true for Cluster 4. Cluster 4 is mainly composed of private universities, which are more expensive.

In general, private schools are more expensive than public schools. Schools that have a higher percent of faculty with phDs, and lower ratio of student to faculty, will be more expensive and have a higher graduation rate. More prestigeous schools will have higher number of students being top in their high school classes, which relates to higher graduation rates. Public schools typically have higher number of applications and higher acceptance rates.

## Midterm Task 6

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.

First, the record for Tufts University will be selected from the data frame.

```
# Select the Tufts University record from the original data frame

university[ , c(4:20)] <- scale(university[ , c(4:20)])

university_tufts <- subset(university, College.Name == "Tufts University")

# Return the record for Tufts University

university_tufts

##         College.Name State Public..1...Private..2. X..appli..rec.d
## 476 Tufts University    MA                       2        1.372653
##     X..appl..accepted X..new.stud..enrolled X..new.stud..from.top.10.
## 476         0.7705108             0.4817205                  1.874556
##     X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad
## 476                  1.803047       0.1992003              NA
##     in.state.tuition out.of.state.tuition      room    board add..fees
## 476         2.207062             2.499321 0.4547283 1.313225 0.2364556
##     estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
## 476         0.2989274        -0.6458426      1.702848       -0.8789855
##     Graduation.rate
## 476        1.672645
```

From reviewing the record, it appears the only missing value on this record is "X..PT.undergrad"

Next, the Euclidean distance can be computed against the four cluster values. Per the prompt instructions, the "X..PT.undergrad" variable will be left out for the calculation purposes.

```
# Calculate the Euclidean distance of Tufts University's record versus
Cluster 1

dist(rbind(university_tufts[ , c(4:9,11:20)], km1$centers[1, -7]))

##        476
## 2 7.514044

# Calculate the Euclidean distance of Tufts University's record versus
Cluster 2

dist(rbind(university_tufts[ , c(4:9,11:20)], km1$centers[2, -7]))

##        476
## 2 6.983519

# Calculate the Euclidean distance of Tufts University's record versus
Cluster 3

dist(rbind(university_tufts[ , c(4:9,11:20)], km1$centers[3, -7]))

##        476
## 2 2.101794

# Calculate the Euclidean distance of Tufts University's record versus
Cluster 4

dist(rbind(university_tufts[ , c(4:9,11:20)], km1$centers[4, -7]))

##        476
## 2 5.117285
```

Based on the Euclidean distance results, Cluster 3 is the closest cluster for Tufts University. Per the prompt instructions, the value for "X..PT.undergrad" will be imputed for this record.

```
# Impute the record for "X..PT.undergrad" using the value from Cluster 3

university_tufts[ , 10] <- km1$centers[3, 7]

# Return the record to verify there are no more missing values
```

```
university_tufts

##          College.Name State Public..1...Private..2. X..appli..rec.d
## 476 Tufts University    MA                       2        1.372653
##     X..appl..accepted X..new.stud..enrolled X..new.stud..from.top.10.
## 476         0.7705108             0.4817205                  1.874556
##     X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad
## 476                  1.803047       0.1992003      -0.3825922
##     in.state.tuition out.of.state.tuition      room    board add..fees
## 476         2.207062             2.499321 0.4547283 1.313225 0.2364556
##     estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
## 476         0.2989274        -0.6458426      1.702848       -0.8789855
##     Graduation.rate
## 476        1.672645
```

After reviewing the record for Tuft University, it has been confirmed that the missing value for “X..PT.undergrad” has been imputed with the value of the centroid for Cluster 3, which is -0.3825922.