Assignment 4

Jared Stadden

3/15/2020

R Markdown

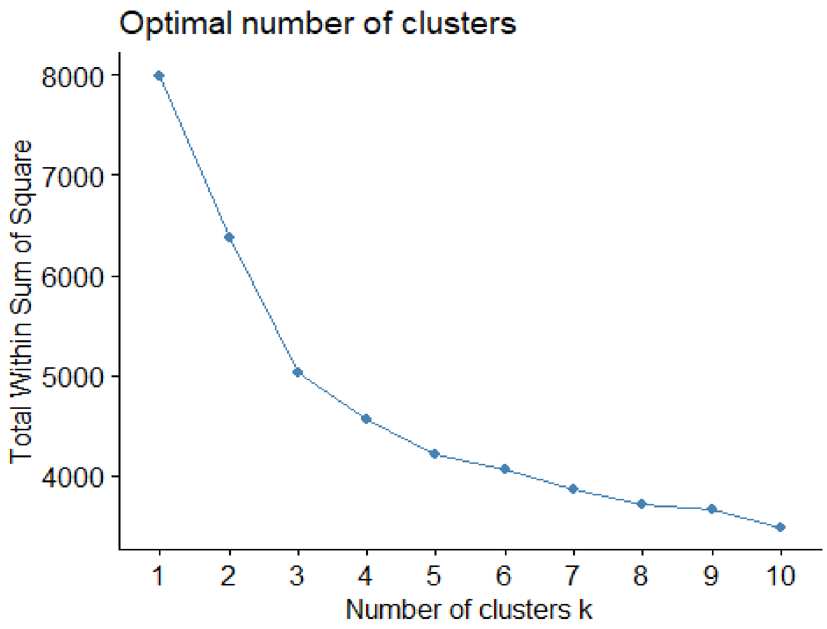
Initial data setup (removing missing records) and loading libraries

```
#Read data into R
original = read.csv("C:\\Users\\jared\\Desktop\\Universities.csv")

#loading libraries
#install.packages("factoextra")
#install.packages("tidyverse")
#install.packages("flexclust")
library(factoextra)

## Warning: package 'factoextra' was built under R version 3.6.3

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages ---------------------------------------------------
------ tidyverse 1.3.0 --

## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3

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

library(ISLR)
library(flexclust)

## Warning: package 'flexclust' was built under R version 3.6.3
## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Warning: package 'modeltools' was built under R version 3.6.3

## Loading required package: stats4
```

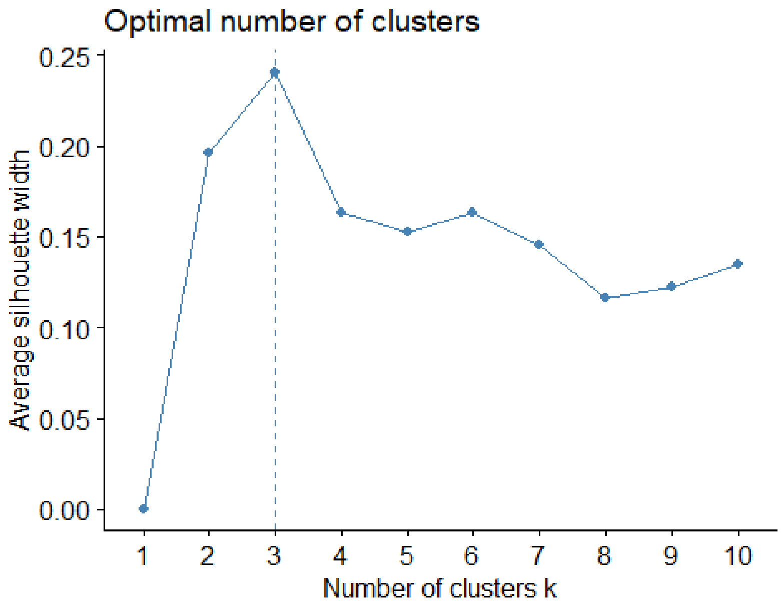
```r
set.seed(123)

#removing NAs
uni_complete<-na.omit(original)

#normalizing data
uni_complete[,4:20]<-as.data.frame(scale(uni_complete[,4:20]))
uni<-uni_complete[,4:20]
```

Determining optimal number of clusters:

```
#Elbow Method
fviz_nbclust(uni,kmeans,method = "wss")
```



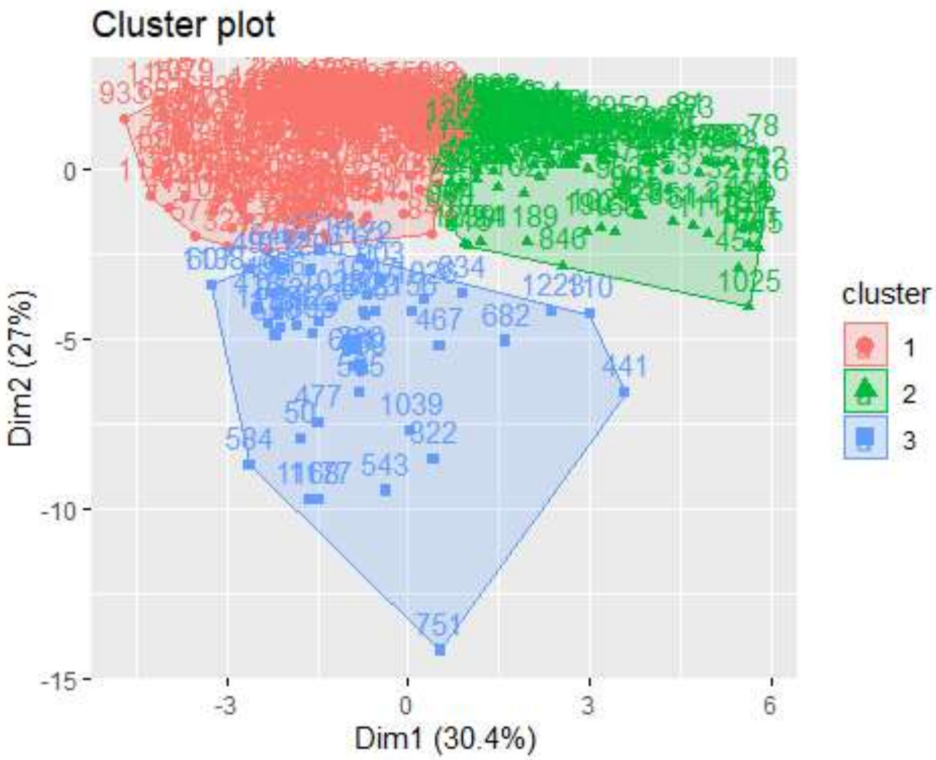
```
#Silhouette Method
fviz_nbclust(uni,kmeans,method = "silhouette")
```



From the methods above 3 clusters seem to be the optimal choice.

K Means Clustering of continuous measurements

```
#Computing K Means
k3 <- kmeans(uni,centers = 3,nstart = 25)

#Creating dataframe of centroids of clusters
centroids <- as.data.frame(k3$centers)

#Seeing the size of each cluster
k3$siz

## [1] 275 150  46

#Visualizing the clusters
fviz_cluster(k3,data = uni)
```



Comparing summary statistics by cluster

```
uni$cluster <-as.factor(k3$cluster)
aggregate(uni,by=list(uni$cluster), FUN=mean)
```

```
## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

##   Group.1 X..appli..rec.d X..appl..accepted X..new.stud..enrolled
## 1       1     -0.35953828       -0.34918455            -0.3171053
## 2       2      0.05140256       -0.04367128            -0.1683551
## 3       3      1.98179657        2.22992267             2.4447222
##   X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                -0.5020886                -0.5128195      -0.2952142
## 2                 0.8795798                 0.8620961      -0.2324464
## 3                 0.1334215                 0.2545856       2.5228452
##   X..PT.undergrad in.state.tuition out.of.state.tuition       room      bo
ard
## 1      -0.1217682       -0.4036544           -0.5263964 -0.3588740 -0.3938
990
## 2      -0.3130216        1.0620416            1.1158839  0.6698444  0.7756
859
## 3       1.7486849       -1.0500277           -0.4918168 -0.0388330 -0.1745
795
##     add..fees estim..book.costs estim..personal.. X..fac..w.PHD
## 1 -0.05832646       -0.06621454        0.05935933    -0.5322257
## 2 -0.04496556        0.07122705       -0.39665857     0.7659627
## 3  0.49531762        0.16358567        0.93858632     0.6840794
##   stud..fac..ratio Graduation.rate cluster
## 1        0.2810858      -0.4171456      NA
## 2       -0.7036167       0.8426062      NA
## 3        0.6139980      -0.2538234      NA
```

Cluster 1: Low enrollment/acceptance rate, Low tuition, Fewer top tier students, Fewer Phd faculty, Moderate student to faculty ratio, Low graduation rate

Cluster 2: Average enrollment/acceptance rate, High tuition, More top tier students, Low student to faculty ratio, High graduation rate

Cluster 3: High enrollment/acceptance rate, Low tuition, Average qualtity students, Larger student to faculty ratio, Lower graduation rate

Characterizing the clusters by catergorical measurements

```
uni_complete$cluster<-as.factor(k3$cluster)

table(uni_complete$cluster, uni_complete$State)

##
##AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO
##1  2  3  4  0  3  5  3  0  1  3  4  1 16  2  7  8  7  4  2  7  1  4  7  6 1
2
##2  0  1  0  0 10  1  6  4  0  4  2  0  2  0  6  7  0  2  2 12  1  2  4  4
2
##3  0  0  0  2  2  0  1  0  1  1  1  0  0  0  2  0  0  0  1  3  1  0  2  1
1
##
##MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV
##1  5  2 16  5  5  4  9  2  0 18 13  5  1 19  1  7  4 11 14  1  8  5  0  5
2
##2  0  0  3  0  1  1  3  0  0 18  7  0  4 20  2  2  0  3  2  0  4  2  2  4
0
##3  0  0  4  0  1  1  1  0  0  2  4  1  0  3  1  0  0  1  4  1  3  0  0  0
0
##
##      WY
##   1  1
##   2  0
##   3  0

round(prop.table(table(uni_complete$cluster, uni_complete$State),margin = 1),
2)

##
##        AK   AL   AR   AZ   CA   CO   CT   DC   DE   FL   GA   HI   IA   ID
IL
##   1 0.01 0.01 0.01 0.00 0.01 0.02 0.01 0.00 0.00 0.01 0.01 0.00 0.06 0.01
0.03
##   2 0.00 0.01 0.00 0.00 0.07 0.01 0.04 0.03 0.00 0.03 0.01 0.00 0.01 0.00
0.04
##   3 0.00 0.00 0.00 0.04 0.04 0.00 0.02 0.00 0.02 0.02 0.02 0.00 0.00 0.00
0.04
##
##        IN   KS   KY   LA   MA   MD   ME   MI   MN   MO   MS   MT   NC   ND
NE
##   1 0.03 0.03 0.01 0.01 0.03 0.00 0.01 0.03 0.02 0.04 0.02 0.01 0.06 0.02
0.02
##   2 0.05 0.00 0.01 0.01 0.08 0.01 0.01 0.03 0.03 0.01 0.00 0.00 0.02 0.00
0.01
##   3 0.00 0.00 0.00 0.02 0.07 0.02 0.00 0.04 0.02 0.02 0.00 0.00 0.09 0.00
0.02
##
##        NH   NJ   NM   NV   NY   OH   OK   OR   PA   RI   SC   SD   TN   TX
```

```
UT
##   1 0.01 0.03 0.01 0.00 0.07 0.05 0.02 0.00 0.07 0.00 0.03 0.01 0.04 0.05
0.00
##   2 0.01 0.02 0.00 0.00 0.12 0.05 0.00 0.03 0.13 0.01 0.01 0.00 0.02 0.01
0.00
##   3 0.02 0.02 0.00 0.00 0.04 0.09 0.02 0.00 0.07 0.02 0.00 0.00 0.02 0.09
0.02
##
##       VA   VT   WA   WI   WV   WY
##   1 0.03 0.02 0.00 0.02 0.01 0.00
##   2 0.03 0.01 0.01 0.03 0.00 0.00
##   3 0.07 0.00 0.00 0.00 0.00 0.00
```

```
table(uni_complete$cluster, uni_complete$Public..1...Private..2.)
```

```
##
##       1   2
##   1  84 191
##   2   3 147
##   3  41   5
```

```
round(prop.table(table(uni_complete$cluster, uni_complete$Public..1...Private
..2.),margin = 1),2)
```

```
##
##        1    2
##   1 0.31 0.69
##   2 0.02 0.98
##   3 0.89 0.11
```

```
#prop.table(table(uni_complete$State,uni_complete$cluster),margin = 1)
#prop.table(table(uni_complete$Public..1...Private..2.,uni_complete$cluster),
margin = 1)
```

It is difficult to determine a relationship between clusters and State. State with the most universities tended to land in clusters 1 and 2. States with a moderately large number of schools seemed to skew to either cluster 1 or cluster 2, but not a fairly even split like those previously mentioned. States with few schools seemed to have a more random spattering of clusters.

However, there seems to be a stronger relationship between Public/Private and the clusters. Cluster 2 is almost exclusively composed of private colleges, while Cluster 3 leans quite heavily into public universities. Cluster 1 has a less extreme split, but contains a majority of private universities.

What external info can explain some or all of the clusters?

The rank of the schools may play a role in the clusters. Highly ranked schools receive many qualified applicants and can make them more selective. This demand can also lead to higher tuition costs. Highly ranked schools also generally have prestigious faculty, which could be captured by the number of faculty with Phds component.

Another piece of information that is missing but could help explain the clusters is the type of degrees/programs offered by the schools. For example, one of the clusters might have a high concentration of liberal arts colleges while another cluster might be full of science focused schools. Based on the characteristics I would expect cluster 1 to be liberal arts schools, cluster 2 to be more prestigious private schools, and cluster 3 to be standard state universities.

Predicting Tufts' cluster:

```
#Retreiving and normalizing the Tufts data
tufts <- subset(original,College.Name == "Tufts University")
uni_tufts<-na.omit(original)
uni_tufts<-rbind(uni_tufts,tufts)
uni_tufts[,4:20]<-as.data.frame(scale(uni_tufts[,4:20]))
tufts_norm<-uni_tufts[472,4:20]

#Imputing and computing distance from cluster 1
clust1_cent<-centroids[1,]
tufts_norm$X..PT.undergrad<-clust1_cent$X..PT.undergrad
tufts_dist1<-dist(rbind(clust1_cent,tufts_norm))
tufts_dist1

##            1
## 476 6.422723

#Imputing and computing distance from cluster 2
clust2_cent<-centroids[2,]
tufts_norm$X..PT.undergrad<-clust2_cent$X..PT.undergrad
tufts_dist2<-dist(rbind(clust2_cent,tufts_norm))
tufts_dist2

##           2
## 476 2.65142

#Imputing and computing distance from cluster 3
clust3_cent<-centroids[3,]
tufts_norm$X..PT.undergrad<-clust3_cent$X..PT.undergrad
tufts_dist3<-dist(rbind(clust3_cent,tufts_norm))
tufts_dist3

##            3
## 476 6.683839
```

Tufts has the smallest distance from and would be best classified into cluster 2.