R Notebook

GUOTAI SUN

My Task: I am a data scientist hired by a non-profit organization whose mission is to increase college graduation rates for underpriveleged populations. Through advocacy and targeted outreach programs, my organization strives to identify and alleviate barriers to educational achievement. A

My team is committed to developing a more data-based approach to decision making. As a prelude to future analyses, you are requested to analyze the data to identify clusters of similar colleges and universities

My Steps:

1.Find potential influential factors which could increase/decrease college graduation rates

2.Data clean - make a new table with those essential factors

3.Clustering Algorithm: K-means to identify clusters of similar colleges and universities

4.Report the optimal K-value and make a conclusion

My explanation of the variables:

1. ADM_RATE and SAT_AVG_All, I consider these two are important to college graduation rates. Firstly, if a college with lower ADM_RATE, means it only picks the greatest students, so those nice students will also have more graduation rates in the future. Also, if students with higher SAT grade, proves their hard work, so they will also have more chance to learn better in their college life.

2. LOCALE: Sometimes if a campus is at a big city, or probably surrounding by too many bars, markets, students would be distracted and have lower graduation rates.

3. HIGHDEG: it shows the highest degree of a university, the higher, the better learning environment, proves the graduation rates of a college.

4. CCBASIC, CCSIZSET: I believe these two variables stands for the class management of a campus, which is very related to the quality of education, and proves the graduation rates.

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyve
rse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

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

CollegeScorecard = read_csv("CollegeScorecard.csv")

##
## -- Column specification ------------------------------------------
------------
## cols(
##   .default = col_logical(),
##   UNITID = col_double(),
##   OPEID = col_double(),
##   opeid6 = col_double(),
##   INSTNM = col_character(),
##   CITY = col_character(),
##   STABBR = col_character(),
##   ZIP = col_character(),
##   AccredAgency = col_character(),
##   INSTURL = col_character(),
##   NPCURL = col_character(),
##   HCM2 = col_double(),
##   main = col_double(),
##   NUMBRANCH = col_double(),
##   PREDDEG = col_double(),
##   HIGHDEG = col_double(),
##   CONTROL = col_double(),
##   st_fips = col_double(),
##   region = col_double(),
##   LOCALE = col_double(),
##   LATITUDE = col_double()
```

```
##   # ... with 531 more columns
## )
## i Use `spec()` for the full column specifications.
```

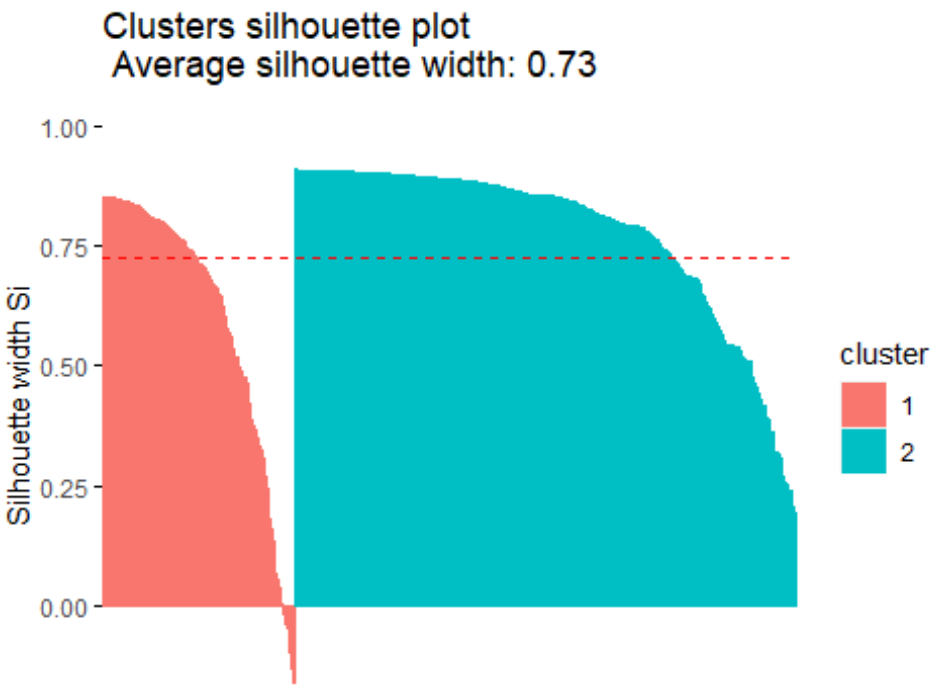
```
new_CollegeScorecard <- select(CollegeScorecard, `UNITID`,`INSTNM`, `AD
M_RATE` ,`SAT_AVG_ALL`,`HIGHDEG`,`LOCALE`,`CCBASIC`,`CCSIZSET`)
new_CollegeScorecard
```

```
## # A tibble: 7,804 x 8
##    UNITID INSTNM            ADM_RATE SAT_AVG_ALL HIGHDEG LOCALE CCBA
SIC CCSIZSET
##     <dbl> <chr>                <dbl>       <dbl>   <dbl>  <dbl>   <d
bl>    <dbl>
##  1 100654 Alabama A & M Un~    0.899         823       4     12
 18       14
##  2 100663 University of Al~    0.867        1146       4     12
 15       15
##  3 100690 Amridge Universi~   NA              NA       4     12
 21        6
##  4 100706 University of Al~    0.806        1180       4     12
 15       12
##  5 100724 Alabama State Un~    0.512         830       4     12
 18       13
##  6 100751 The University o~    0.566        1171       4     13
 16       16
##  7 100760 Central Alabama ~   NA              NA       2     32
  2        2
##  8 100812 Athens State Uni~   NA              NA       3     31
 22        9
##  9 100830 Auburn Universit~    0.837         970       4     12
 18       12
## 10 100858 Auburn University    0.827        1215       4     13
 16       15
## # ... with 7,794 more rows
```

```
CollegeClean <- new_CollegeScorecard %>% na.omit() %>%
  select(-UNITID,-INSTNM)
CollegeClean
```

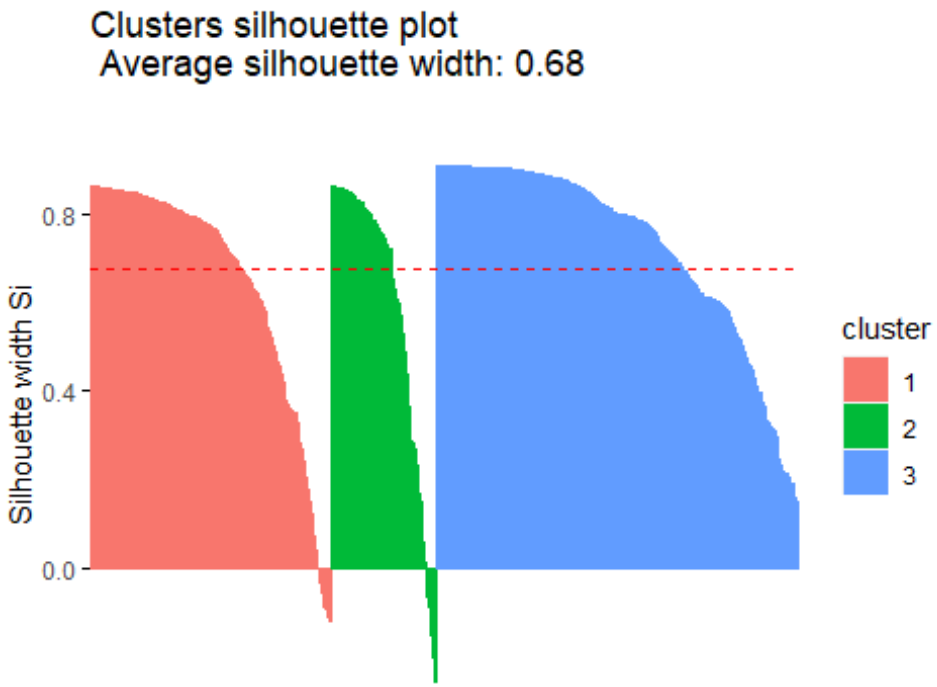
```
## # A tibble: 1,310 x 6
##    ADM_RATE SAT_AVG_ALL HIGHDEG LOCALE CCBASIC CCSIZSET
##       <dbl>       <dbl>   <dbl>  <dbl>   <dbl>    <dbl>
##  1    0.899         823       4     12      18       14
##  2    0.867        1146       4     12      15       15
##  3    0.806        1180       4     12      15       12
##  4    0.512         830       4     12      18       13
##  5    0.566        1171       4     13      16       16
##  6    0.837         970       4     12      18       12
##  7    0.827        1215       4     13      16       15
##  8    0.642        1177       3     12      21       11
##  9    0.628         999       3     12      22        7
```

```
## 10    0.833        1036       4     13      18       12
## # ... with 1,300 more rows

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.0.5

## Welcome! Want to learn more? See two factoextra-related books at htt
ps://goo.gl/ve3WBa

library(cluster)

#two-cluster model,k=2
College2CL <- kmeans(CollegeClean, centers = 2)
dis2CL = dist(CollegeClean)^2
sil2CL = silhouette(College2CL$cluster, dis2CL)
fviz_silhouette(sil2CL) #score was 0.73

##   cluster size ave.sil.width
## 1       1  362          0.60
## 2       2  948          0.78
```

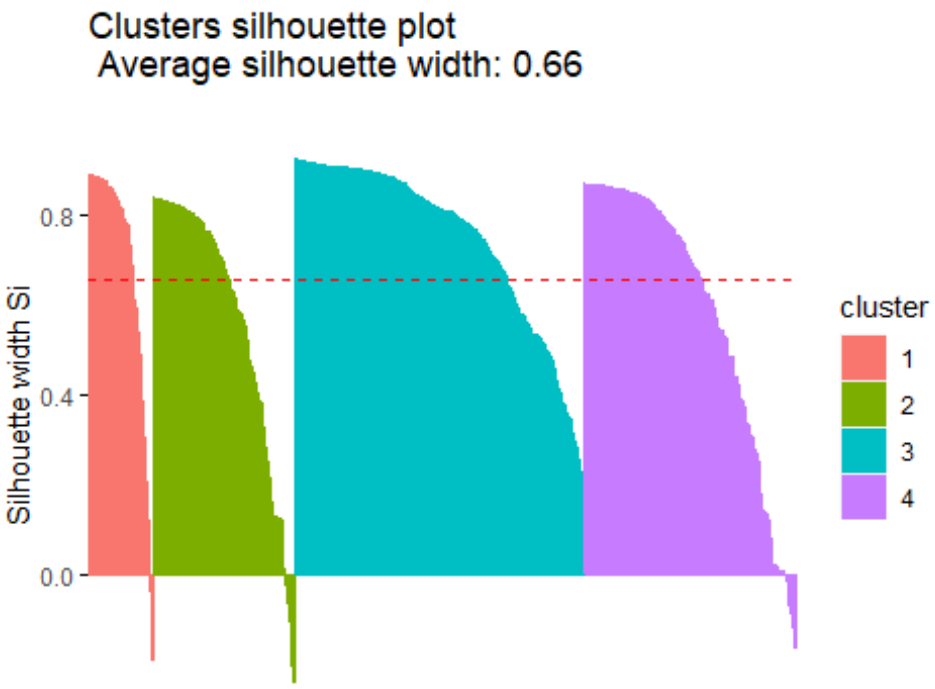



```
#three-cluster model,k=3
College3CL <- kmeans(CollegeClean, centers = 3)
dis3CL = dist(CollegeClean)^2
sil3CL = silhouette(College3CL$cluster, dis3CL)
fviz_silhouette(sil3CL) #score was 0.68
```

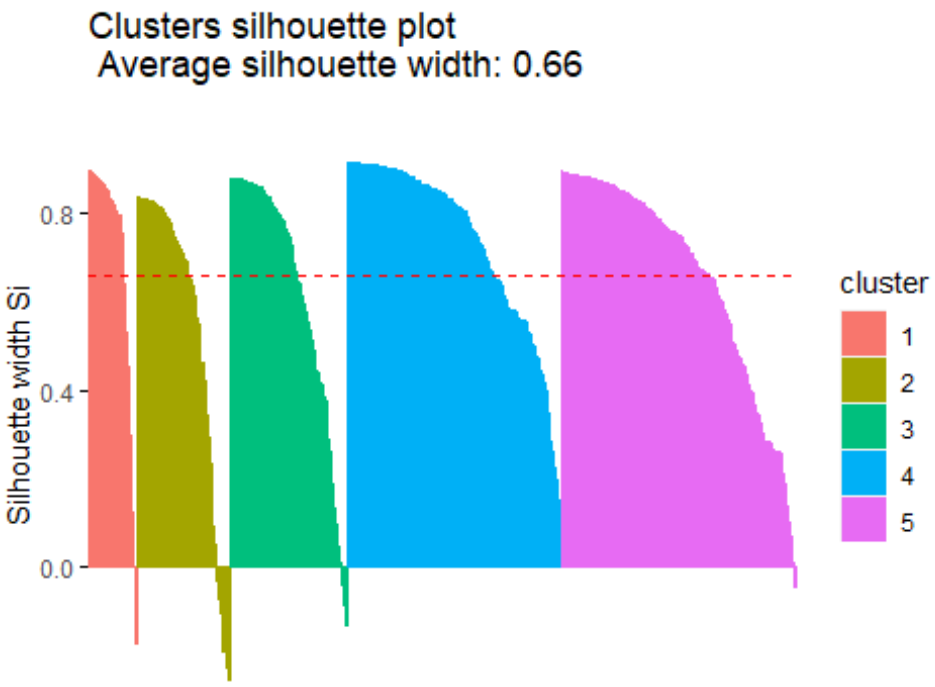
```
##   cluster size ave.sil.width
## 1       1  444          0.64
## 2       2  195          0.58
## 3       3  671          0.73
```



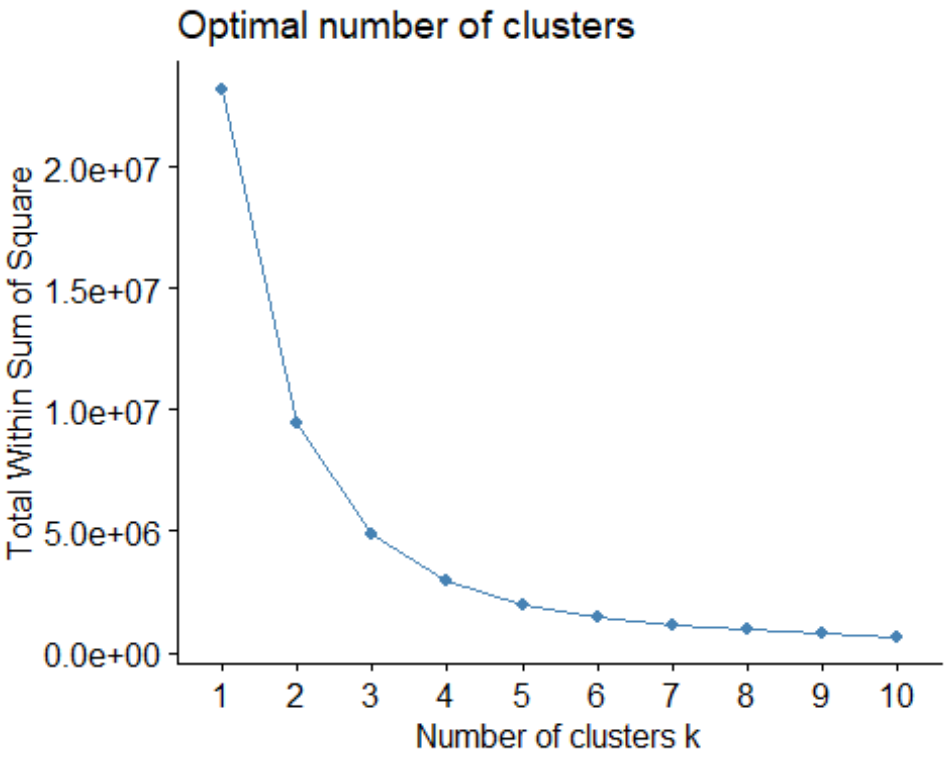

```
#two-cluster model,k=4
College4CL <- kmeans(CollegeClean, centers = 4)
dis4CL = dist(CollegeClean)^2
sil4CL = silhouette(College4CL$cluster, dis4CL)
fviz_silhouette(sil4CL) #score was 0.67

##   cluster size ave.sil.width
## 1       1  118          0.68
## 2       2  263          0.57
## 3       3  536          0.75
## 4       4  393          0.58
```




```
#two-cluster model,k=5
College5CL <- kmeans(CollegeClean, centers = 5)
dis5CL = dist(CollegeClean)^2
sil5CL = silhouette(College5CL$cluster, dis5CL)
fviz_silhouette(sil5CL) #score was 0.65

##   cluster size ave.sil.width
## 1       1   88          0.69
## 2       2  173          0.54
## 3       3  217          0.62
## 4       4  398          0.73
## 5       5  434          0.65
```




```
#I choose the three-cluster model, even though its score is less than t
wo-cluster model, it has more similar intra-cluster similarity and comp
ared to other three models, 0.68 is not too bad.

CollegeClean  %>% mutate(cluster = College3CL$cluster)

## # A tibble: 1,310 x 7
##    ADM_RATE SAT_AVG_ALL HIGHDEG LOCALE CCBASIC CCSIZSET cluster
##       <dbl>       <dbl>   <dbl>  <dbl>   <dbl>    <dbl>   <int>
##  1    0.899         823       4     12      18       14       1
##  2    0.867        1146       4     12      15       15       3
##  3    0.806        1180       4     12      15       12       3
##  4    0.512         830       4     12      18       13       1
##  5    0.566        1171       4     13      16       16       3
##  6    0.837         970       4     12      18       12       1
##  7    0.827        1215       4     13      16       15       2
##  8    0.642        1177       3     12      21       11       3
##  9    0.628         999       3     12      22        7       1
## 10    0.833        1036       4     13      18       12       3
## # ... with 1,300 more rows

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



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.