mchris26_4

Mark Christian

10/25/2020

```
#install.packages("tidyverse")
#install.packages("factoextra")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

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

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

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

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.3
```

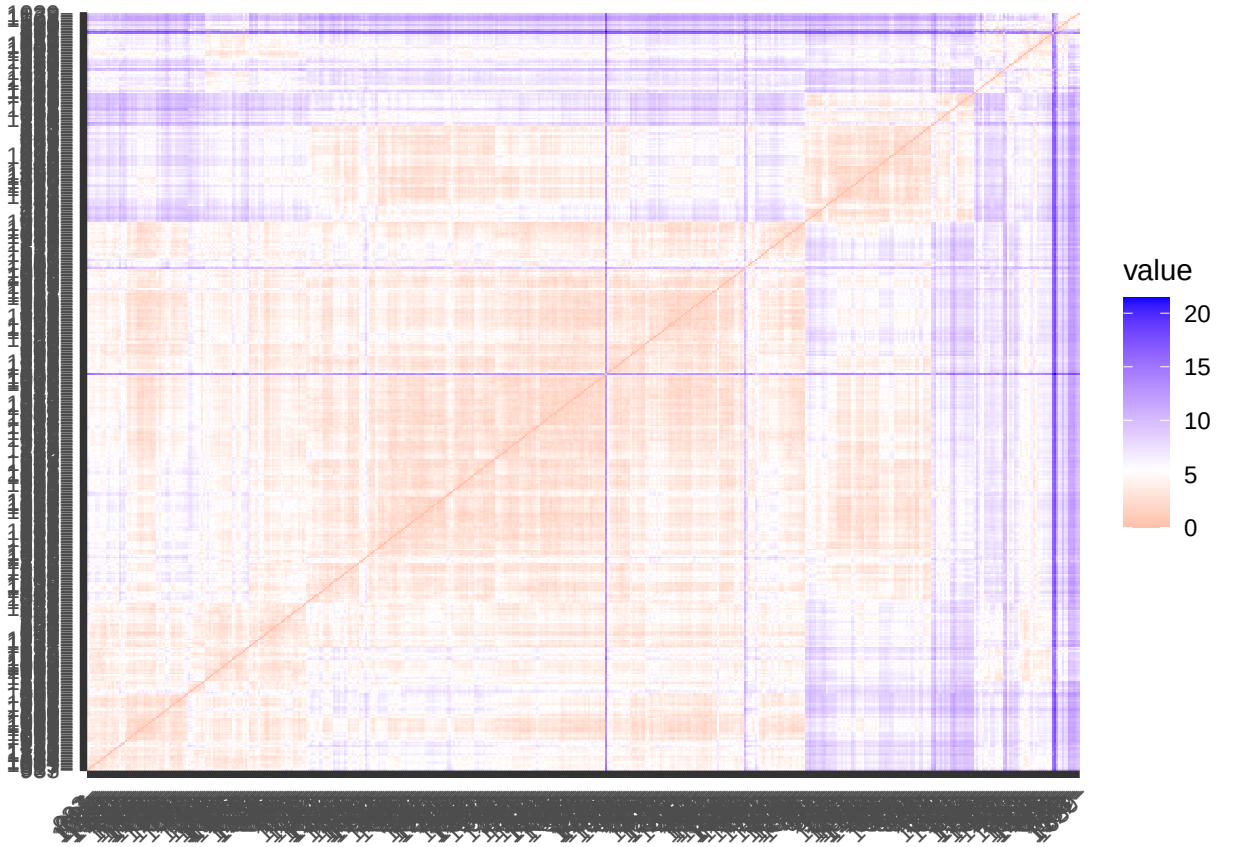
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

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

```
Universities_R <- na.omit(Universities)
Univ_Conti<-Universities_R[,4:20]
```

I removed the records with missing values

```
Universities_Conti<-scale(Univ_Conti)
distance <- get_dist(Universities_Conti)
fviz_dist(distance)
```




```
set.seed(123)
k4 <- kmeans(Universities_Conti, centers =4 , nstart = 25) # k = 4, number of restarts = 25
```

```
k4$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. X..FT.undergrad
## 1                 0.1334215                 0.2545856       2.5228452
## 2                 0.0102519                 0.1080080      -0.4049392
## 3                -0.6785172                -0.7279285      -0.1972688
## 4                 1.6526422                 1.4315089      -0.1108205
##   X..PT.undergrad in.state.tuition out.of.state.tuition        room      board
## 1      1.74868491       -1.0500277           -0.4918168 -0.03883300 -0.1745795
## 2     -0.25785122        0.4057712            0.2956208  0.08357902  0.3292398
## 3     -0.04353747       -0.7234450           -0.8237908 -0.53385193 -0.6791344
## 4     -0.38259215        1.5022093            1.6819156  1.19276784  0.9944521
##     add..fees estim..book.costs estim..personal.. X..fac..w.PHD
## 1  0.49531762       0.163585669         0.9385863     0.6840794
## 2 -0.18996619      -0.158302104        -0.2978018     0.0835866
## 3  0.03928218       0.003218005         0.2531393    -0.6684106
```

```
## 4  0.07619136       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
```

```
k4$size
```

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

```
fviz_cluster(k4, data = Universities_Conti)
```

Size and Center for the Clusters

```
#install.packages("flexclust")
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.0.3
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Warning: package 'modeltools' was built under R version 4.0.3
```

```
## Loading required package: stats4
```

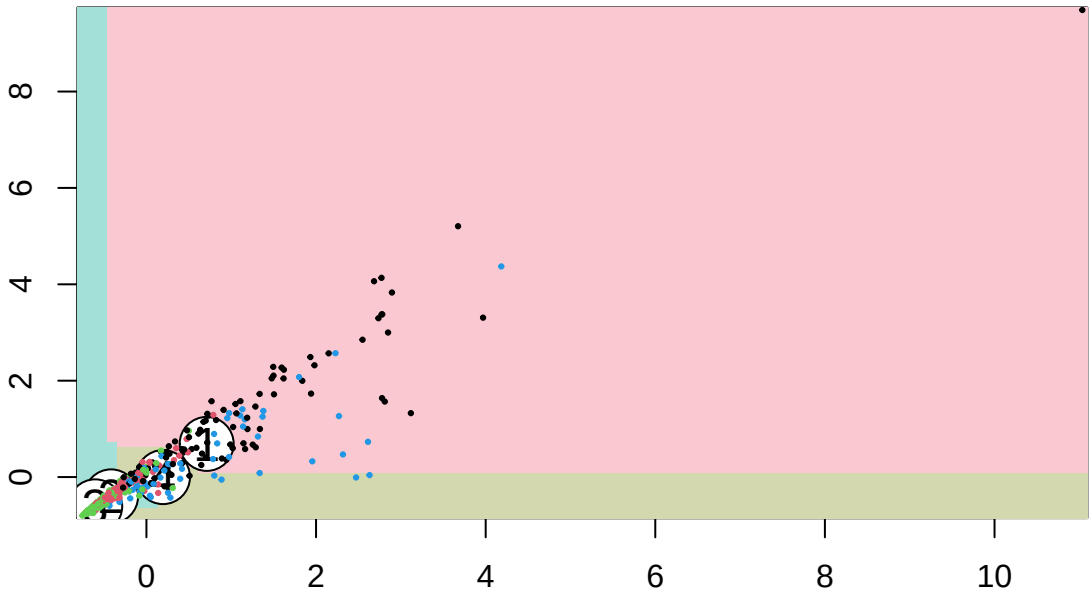
```
set.seed(123)
k4 = kcca(Universities_Conti, k=4, kccaFamily("kmedians"))
k4
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = Universities_Conti, k = 4, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
##   1   2   3   4
##  98 142 165  66
```

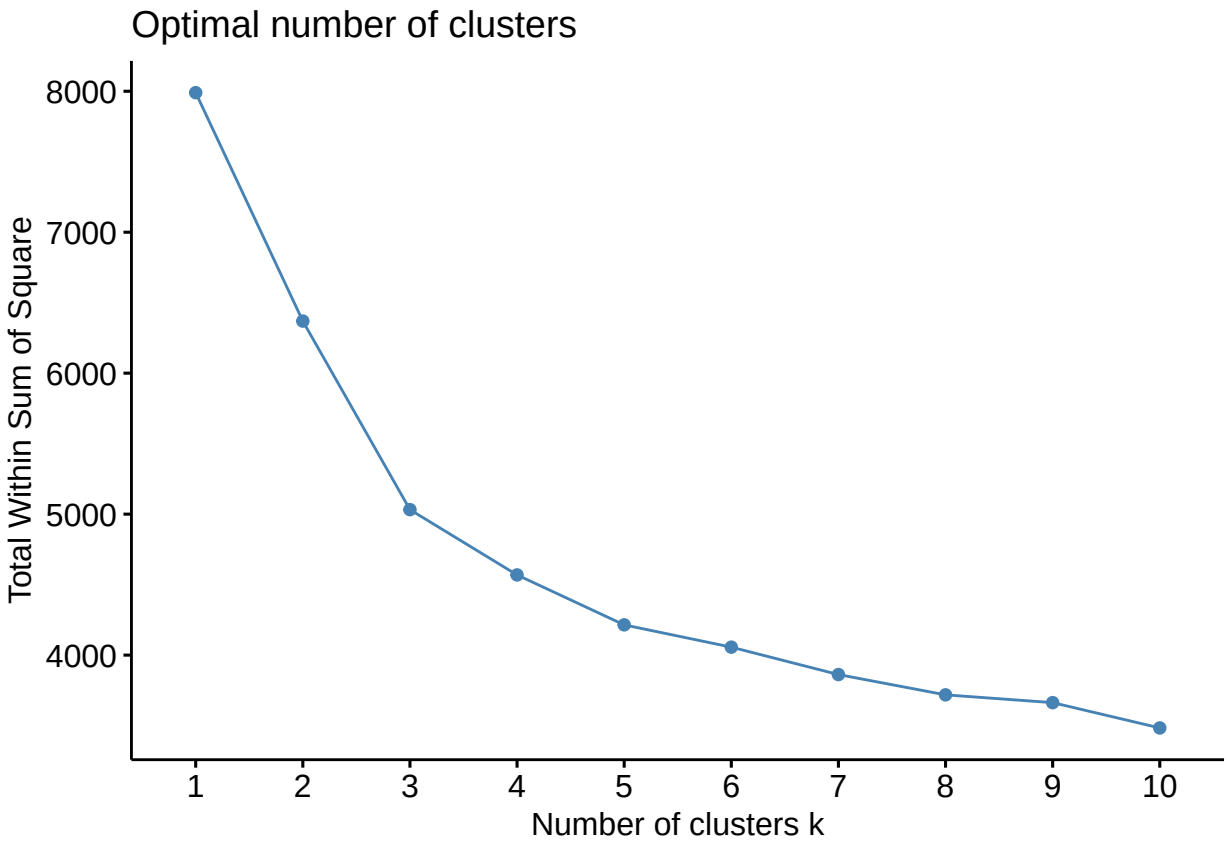
```
clusters_index <- predict(k4)
dist(k4@centers)
```

```
##          1        2        3
## 2 4.194248
## 3 3.854080 2.579616
## 4 6.337718 3.460245 5.874384
```

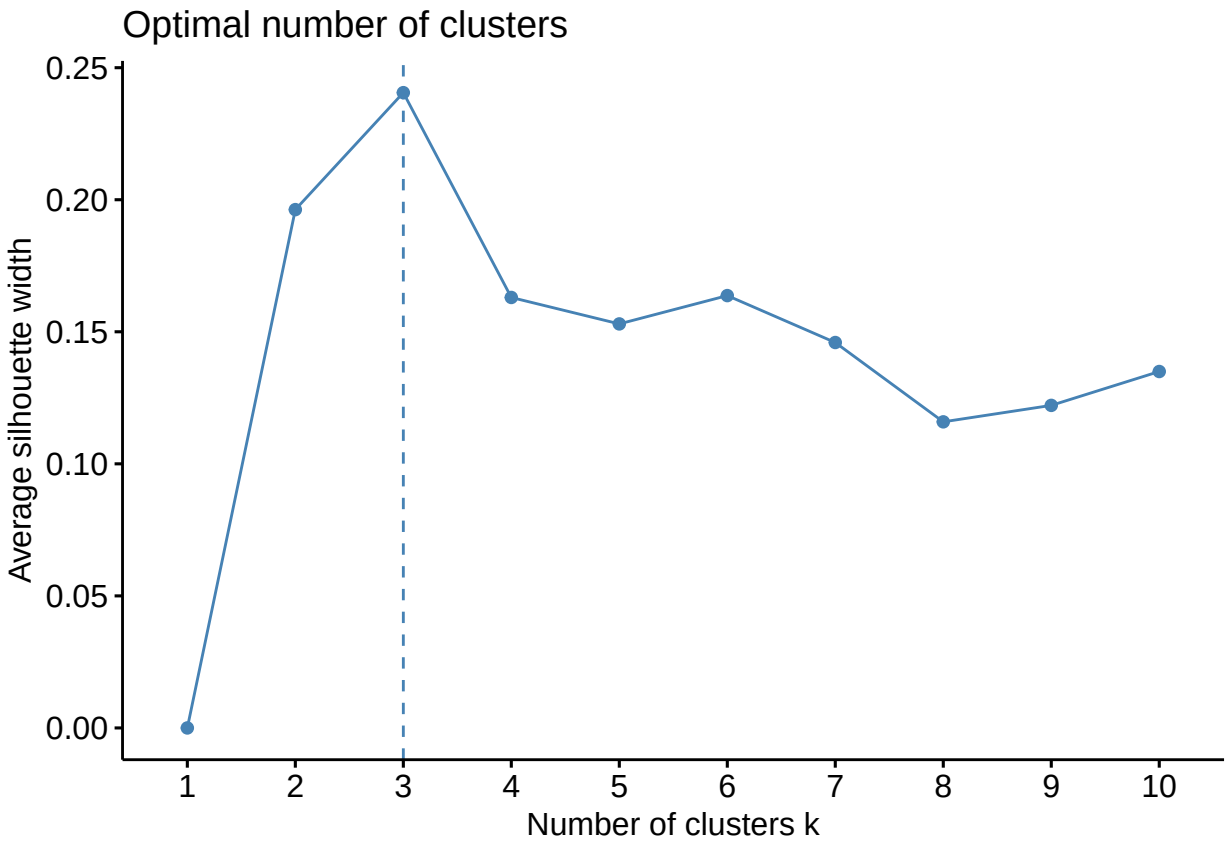
```
image(k4)
points(Universities_Conti, col=clusters_index, pch=19, cex=0.3)
```




```
set.seed(123)
fviz_nbclust(Universities_Conti, kmeans, method = "wss")
```



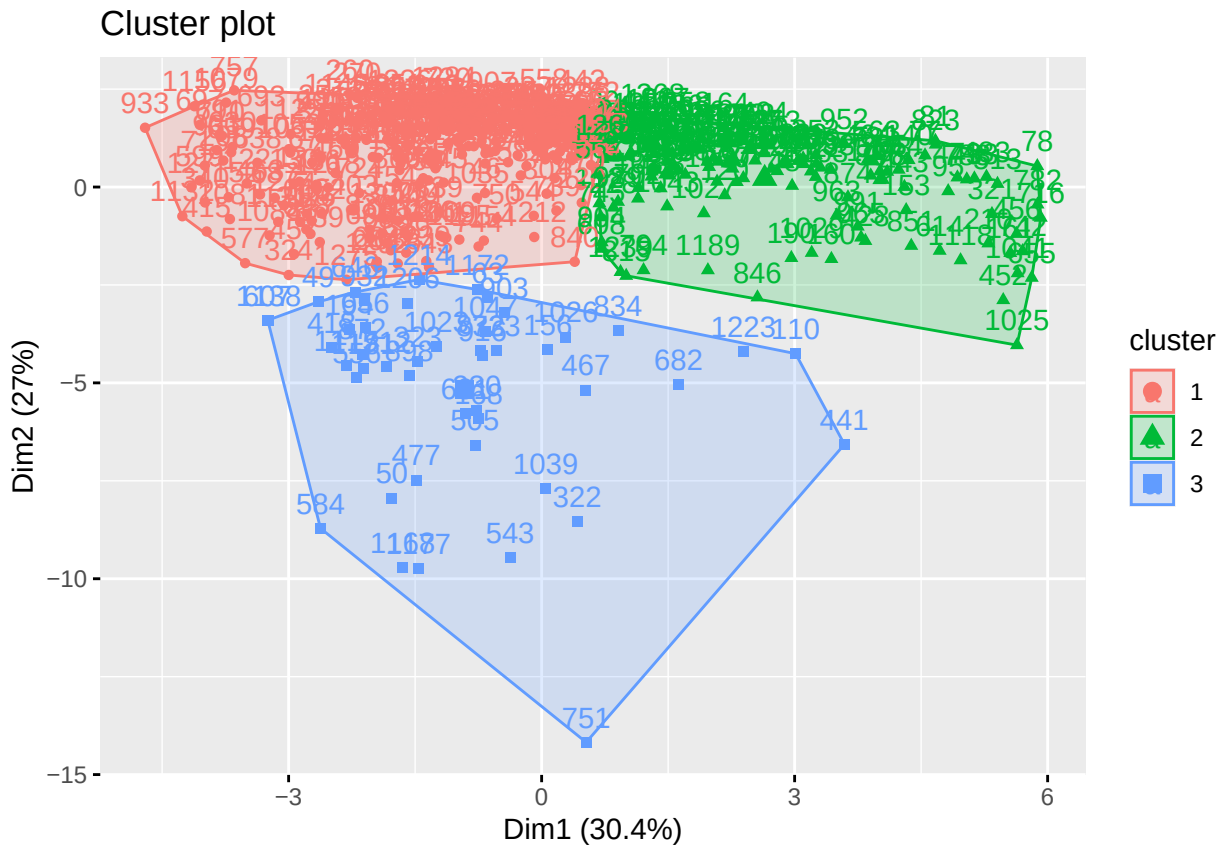

```
fviz_nbclust(Universities_Conti, kmeans, method = "silhouette")
```



Summary Stats for clusture

Sillhouetter chart had the best value ie. 3

```
k3 <- kmeans(Universities_Conti, centers = 3, nstart = 25)
fviz_cluster(k3, data = Universities_Conti)
```




```
#Creating the cluster index for 3 clusters

set.seed(123)
k3 = kcca(Universities_Conti, k=3, kccaFamily("kmedians"))
k3
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = Universities_Conti, k = 3, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
##   1   2   3
## 111 113 247
```

```
cluster_index_3 <- predict(k3)
```

```
set.seed(123)
clusters123<- data.frame(cluster_index_3)
Universities_R <- cbind(Universities_R, clusters123)
head(Universities_R)
```

```
##                          College.Name State Public..1...Private..2.
```

```
## 1            Alaska Pacific University    AK                      2
## 3       University of Alaska Southeast    AK                      1
## 10         Birmingham-Southern College    AL                      2
## 12                 Huntingdon College    AL                      2
## 22                  Talladega College    AL                      2
## 26 University of Alabama at Birmingham    AL                      1
##    X..appli..rec.d X..appl..accepted X..new.stud..enrolled
## 1              193               146                    55
## 3              146               117                    89
## 10             805               588                   287
## 12             608               520                   127
## 22            4414              1500                   335
## 26            1797              1260                   938
##    X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                         16                         44             249
## 3                          4                         24             492
## 10                        67                         88            1376
## 12                        26                         47             538
## 22                        30                         60             908
## 26                        24                         35            6960
##    X..PT.undergrad in.state.tuition out.of.state.tuition room board add..fees
## 1              869             7560                 7560 1620  2500       130
## 3             1849             1742                 5226 2514  2250        34
## 10             207            11660                11660 2050  2430       120
## 12             126             8080                 8080 1380  2540       100
## 22             119             5666                 5666 1424  1540       418
## 26            4698             2220                 4440 1935  3240       291
##    estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
## 1                800              1500            76             11.9
## 3                500              1162            39              9.5
## 10               400               900            74             14.0
## 12               500              1100            63             11.4
## 22              1000              1400            56             15.5
## 26               750              2200            96              6.7
##    Graduation.rate cluster_index_3
## 1               15               3
## 3               39               3
## 10              72               3
## 12              44               3
## 22              46               3
## 26              33               1
```

Comparing the summary stats

```
set.seed(123)


Cluster_Stat <- Universities_R %>%
    group_by( cluster_index_3 ) %>%
    summarise( Univ_InState_Max_Fee=Universities_R[which.max(in.state.tuition),1],Univ_OutState_Max_Fee=
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(Cluster_Stat)
```

```
## # A tibble: 3 x 12
##   cluster_index_3 Univ_InState_Ma~ Univ_OutState_M~ low_accept_rate
##             <int> <chr>            <chr>            <chr>
## 1               1 Adams State Col~ Hanover College  Eastern Connec~
## 2               2 Catholic Univer~ Catholic Univer~ University of ~
## 3               3 Doane College    Doane College    Clark Universi~
## # ... with 8 more variables: Acceptance_rate <dbl>,
## #   Avg_out_state_tuition <dbl>, Avg_int_state_tuition <dbl>,
## #   mean_PHD_fac <dbl>, mean_stud_fac_ratio <dbl>, mean_grad_rate <dbl>,
## #   priv_count <int>, pub_count <int>
```

```
Stat_States<-Universities_R %>%
             group_by(State) %>%       summarise(Univ_InState_Max_Fee=Universities_R[which.max(in.stat
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(Stat_States)
```

```
## # A tibble: 6 x 12
##   State Univ_InState_Ma~ Univ_OutState_M~ low_accept_rate Acceptance_rate
##   <chr> <chr>            <chr>            <chr>                     <dbl>
## 1 AK    Alaska Pacific ~ Alaska Pacific ~ University of ~           0.776
## 2 AL    Alaska Pacific ~ Alaska Pacific ~ University of ~           0.507
## 3 AR    University of A~ University of A~ Alaska Pacific~           0.655
## 4 AZ    Alaska Pacific ~ University of A~ Alaska Pacific~           0.860
## 5 CA    Hendrix College  Hendrix College  Arkansas Colle~           0.614
## 6 CO    University of A~ University of A~ Talladega Coll~           0.721
## # ... with 7 more variables: Avg_out_state_tuition <dbl>,
## #   Avg_int_state_tuition <dbl>, mean_PHD_fac <dbl>, mean_stud_fac_ratio <dbl>,
## #   mean_grad_rate <dbl>, priv_count <int>, pub_count <int>
```

Above is summary for states

Summary for private and public universities respectively

```
Stat_Private <- Universities_R %>%
    filter(Public..1...Private..2. == 2) %>%
    group_by( cluster_index_3 ) %>%
    summarise( Univ_InState_Max_Fee=Universities_R[which.max(in.state.tuition),1],Univ_OutState_Max_Fee=
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(Stat_Private)
```

```
## # A tibble: 3 x 10
##   cluster_index_3 Univ_InState_Ma~ Univ_OutState_M~ low_accept_rate
##             <int> <chr>            <chr>            <chr>
## 1               1 Birmingham-Sout~ Birmingham-Sout~ Hendrix College
```

```
## 2               2 University of C~ University of C~ University of ~
## 3               3 Duke University  Duke University  Georgetown Col~
## # ... with 6 more variables: Acceptance_rate <dbl>,
## #   Avg_out_state_tuition <dbl>, Avg_int_state_tuition <dbl>,
## #   mean_PHD_fac <dbl>, mean_stud_fac_ratio <dbl>, mean_grad_rate <dbl>
```

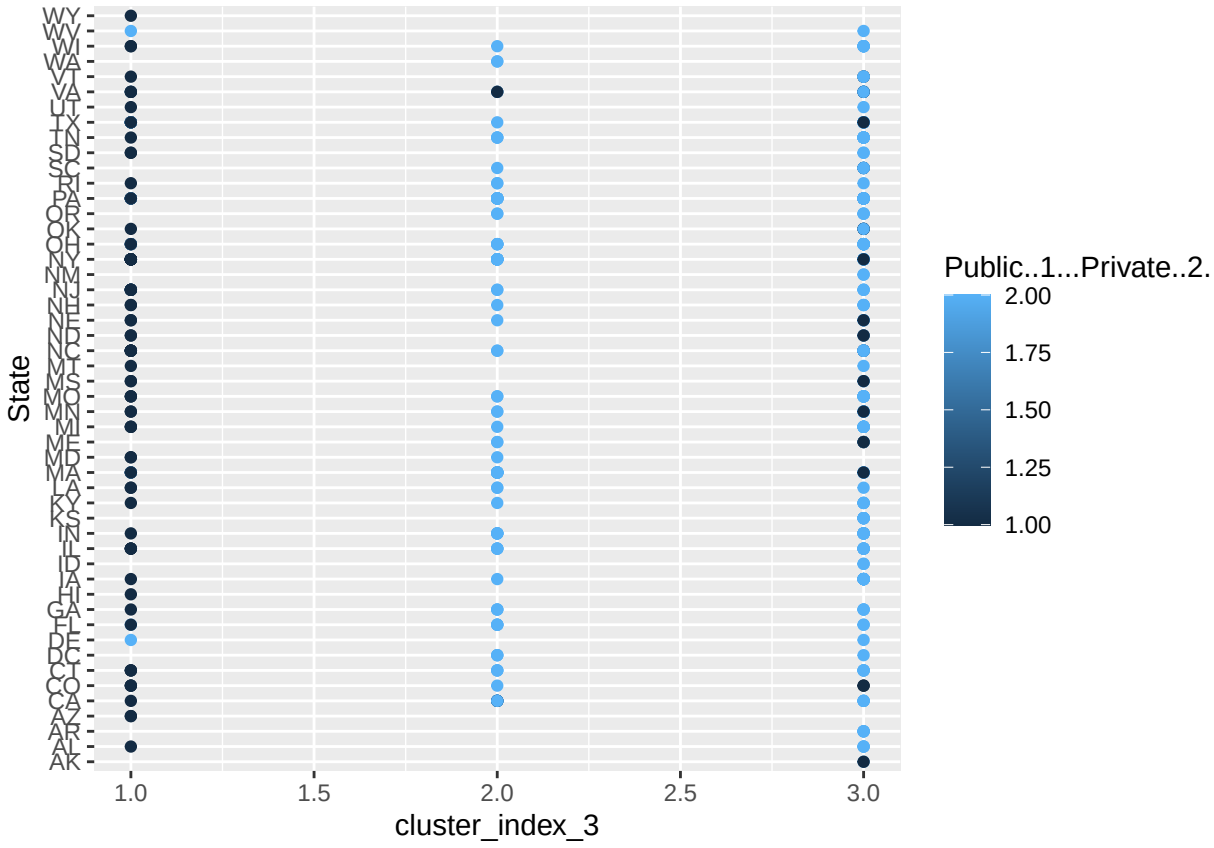
```
Stat_Public <- Universities_R %>%
    filter(Public..1...Private..2. == 1) %>%
    group_by( cluster_index_3 ) %>%
    summarise(Univ_InState_Max_Fee=Universities_R[which.max(in.state.tuition),1],Univ_OutState_Max_Fee=
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(Stat_Public)
```

```
## # A tibble: 3 x 10
##   cluster_index_3 Univ_InState_Ma~ Univ_OutState_M~ low_accept_rate
##             <int> <chr>            <chr>            <chr>
## 1               1 Trinity College  Trinity College  University of ~
## 2               2 Alaska Pacific ~ Alaska Pacific ~ Alaska Pacific~
## 3               3 University of S~ Hendrix College  Alaska Pacific~
## # ... with 6 more variables: Acceptance_rate <dbl>,
## #   Avg_out_state_tuition <dbl>, Avg_int_state_tuition <dbl>,
## #   mean_PHD_fac <dbl>, mean_stud_fac_ratio <dbl>, mean_grad_rate <dbl>
```

```
#plotting cluster
library(ggplot2)
ggplot(Universities_R,aes(x = cluster_index_3, y = State, color =Public..1...Private..2.)) +
  geom_point()
```



Tufts university question with some missing information

```
#centers for clusters
k3 <- kmeans(Univ_Conti, centers = 3, nstart = 25)


#f. Isolating the data to Tufts University
library(dplyr)
library(stats)


Tufts_University <- filter(Universities, 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 26313.12
```

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

```
##          1
## 2 24073.55
```

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

```
##         1
## 2 24664.5
```

The Eucledian Distance from Tufts to Cluster1 is smaller i.e.,24073.55 compared to cluster2 and cluster3. Hence, Cluster1 is Closest to Tufts.

```r
#Impute the missing values for Tufts by taking the average of the cluster on those measurements.
cluster1 <- filter(Universities_R, cluster_index_3 == 1)
cluster1_Avg <- mean(cluster1[,c(10)])
Tufts_University[, c(10)] <- cluster1_Avg
Tufts_University[, c(10)]
```

```
## [1] 2260.721
```

Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements. #2260.721