kyadav1\_Assignment4

Khushboo Yadav

10/26/2020

## 1.Importing libraries and Loading “University Ranking” data

library(tidyverse) # data manipulation

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

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

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(factoextra) # clustering algorithms & visualization

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

library(caret) # model training process

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(e1071) # supports predict() , plot()  
library(dplyr) # supports data manipulations(select,filter,mutate)  
library(normalr) #normalization of large dataset  
library(fpc) # flexible procedure for clustering  
library(flexclust) # k-centroids ,cluster analysis supporting arbitrary distance measures and centroid computation

## Loading required package: grid

## Loading required package: modeltools

## Loading required package: stats4

##   
## Attaching package: 'flexclust'

## The following object is masked from 'package:e1071':  
##   
## bclust

library(stats) # for statistical calculations and random number generation  
library(ggplot2) # visualization of data  
library(ggfortify) # supports plotting tools for statistical clustering using ggplot2  
library(lattice) # data visualization  
library(ISLR)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(clusterSim)

## Loading required package: cluster

## Loading required package: MASS

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

Universities <- read.csv("~/Downloads/Universities.csv")  
  
View(Universities)  
summary(Universities)

## College.Name State Public..1...Private..2.  
## Length:1302 Length:1302 Min. :1.000   
## Class :character Class :character 1st Qu.:1.000   
## Mode :character Mode :character Median :2.000   
## Mean :1.639   
## 3rd Qu.:2.000   
## Max. :2.000   
##   
## 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

str(Universities)

## 'data.frame': 1302 obs. of 20 variables:  
## $ College.Name : chr "Alaska Pacific University" "University of Alaska at Fairbanks" "University of Alaska Southeast" "University of Alaska at Anchorage" ...  
## $ State : chr "AK" "AK" "AK" "AK" ...  
## $ 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 ...

## 2. Data cleanup:-

## 2.1 Removing the NA values from the dataset

#Identifying the NA values from the University data  
  
colMeans(is.na(Universities))

## College.Name State Public..1...Private..2.   
## 0.000000000 0.000000000 0.000000000   
## X..appli..rec.d X..appl..accepted X..new.stud..enrolled   
## 0.007680492 0.008448541 0.003840246   
## X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad   
## 0.180491551 0.155145929 0.002304147   
## X..PT.undergrad in.state.tuition out.of.state.tuition   
## 0.024577573 0.023041475 0.015360983   
## room board add..fees   
## 0.246543779 0.382488479 0.210445469   
## estim..book.costs estim..personal.. X..fac..w.PHD   
## 0.036866359 0.139016897 0.024577573   
## stud..fac..ratio Graduation.rate   
## 0.001536098 0.075268817

#Removing the NA values  
Universities\_N<- na.omit(Universities) ## 471 observations

## 3. Data Manipulation :-

## 3.1 Adding a new column to calculate acceptance rate

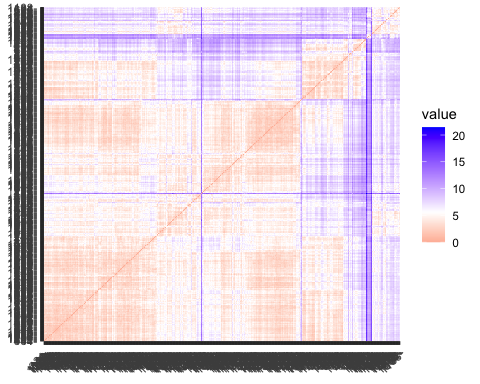
## 3.2 Normalization of all the continuous data

## 3.3 Distance calculation and visualization

#Adding a new column i.e Acceptance rate  
Universities\_N$AcceptanceRate<- Universities\_N$X..appl..accepted/Universities\_N$X..appli..rec.d  
  
#Creating new variable with all continuous measurements from dataset  
n\_Universities<-Universities\_N[,c(4:21)]   
  
#Normalization of the continuous variables in the data frame  
Norm\_Universities<-scale(n\_Universities)  
  
str(Norm\_Universities)

## num [1:471, 1:18] -0.725 -0.737 -0.575 -0.623 0.311 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : chr [1:471] "1" "3" "10" "12" ...  
## ..$ : chr [1:18] "X..appli..rec.d" "X..appl..accepted" "X..new.stud..enrolled" "X..new.stud..from.top.10." ...  
## - attr(\*, "scaled:center")= Named num [1:18] 3147.3 2063 780.7 28 55.7 ...  
## ..- attr(\*, "names")= chr [1:18] "X..appli..rec.d" "X..appl..accepted" "X..new.stud..enrolled" "X..new.stud..from.top.10." ...  
## - attr(\*, "scaled:scale")= Named num [1:18] 4073.1 2503.8 915.6 18.5 20.3 ...  
## ..- attr(\*, "names")= chr [1:18] "X..appli..rec.d" "X..appl..accepted" "X..new.stud..enrolled" "X..new.stud..from.top.10." ...

#Calculating the distance of the normalized Universities data  
distance\_Norm\_Universities <- get\_dist(Norm\_Universities)  
  
#Visualization of the distance  
fviz\_dist(distance\_Norm\_Universities)



## 4.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?

## 4.1 Kmeans

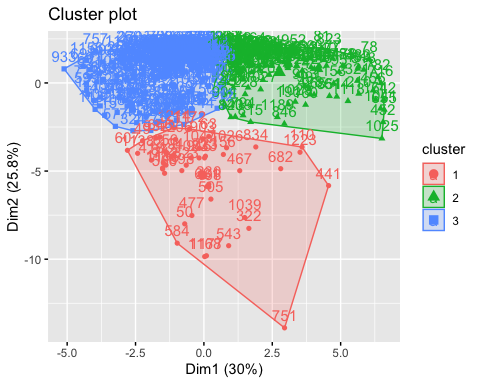
set.seed(3689) # to get the same result in no. of clusters  
  
# Kmeans for the University Ranking dataset  
k3 <- kmeans(Norm\_Universities, centers = 3, nstart = 25) # k = 3, number of restarts = 25  
  
k3$centers

## X..appli..rec.d X..appl..accepted X..new.stud..enrolled  
## 1 1.98179657 2.22992267 2.4447222  
## 2 0.09882751 -0.01778281 -0.1514459  
## 3 -0.36678205 -0.35001620 -0.3196022  
## X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad  
## 1 0.1334215 0.2545856 2.5228452  
## 2 0.9616995 0.9409988 -0.2190855  
## 3 -0.4888588 -0.4982859 -0.2992937  
## X..PT.undergrad in.state.tuition out.of.state.tuition room board  
## 1 1.7486849 -1.0500277 -0.4918168 -0.0388330 -0.1745795  
## 2 -0.3234307 1.0854152 1.1545148 0.7411038 0.7755731  
## 3 -0.1240652 -0.3586414 -0.4820069 -0.3539409 -0.3488602  
## add..fees estim..book.costs estim..personal.. X..fac..w.PHD  
## 1 0.49531762 0.16358567 0.93858632 0.6840794  
## 2 0.01245452 0.08204906 -0.47572563 0.8020973  
## 3 -0.08571954 -0.06618797 0.08024787 -0.4998573  
## stud..fac..ratio Graduation.rate AcceptanceRate  
## 1 0.6139980 -0.2538234 -0.1681130  
## 2 -0.7017481 0.8851790 -0.4848103  
## 3 0.2423045 -0.3893847 0.2626637

k3$size

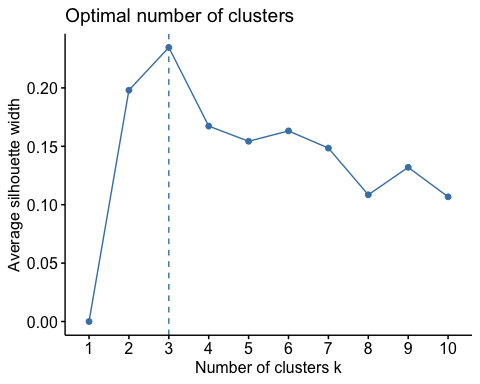
## [1] 46 139 286

fviz\_cluster(k3, data = Norm\_Universities) # Visualize the output



## 4.2.Determining the Optimum value of K

#To determine the optimum K value , using Silhouette Method  
  
  
fviz\_nbclust(Norm\_Universities, kmeans, method = "silhouette")

 The Optimal no. of clusters = 3.

## 5. Analyzing the characteristics of clusters:

Since now we have 3 clusters, we can first get some statistical details about them by using cluster.stats() which returns a list containing many components useful for analyzing the intrinsic characteristics of a clustering:

#To compare the summary statistics of each cluster  
  
cluster.stats(distance\_Norm\_Universities,k3$cluster,alt.clustering = NULL)

## $n  
## [1] 471  
##   
## $cluster.number  
## [1] 3  
##   
## $cluster.size  
## [1] 46 139 286  
##   
## $min.cluster.size  
## [1] 46  
##   
## $noisen  
## [1] 0  
##   
## $diameter  
## [1] 17.385867 9.801045 15.854194  
##   
## $average.distance  
## [1] 6.362113 4.356129 4.295070  
##   
## $median.distance  
## [1] 5.704298 4.185399 4.199208  
##   
## $separation  
## [1] 2.307958 1.108709 1.108709  
##   
## $average.toother  
## [1] 8.035682 6.187104 6.340442  
##   
## $separation.matrix  
## [,1] [,2] [,3]  
## [1,] 0.000000 2.313558 2.307958  
## [2,] 2.313558 0.000000 1.108709  
## [3,] 2.307958 1.108709 0.000000  
##   
## $ave.between.matrix  
## [,1] [,2] [,3]  
## [1,] 0.000000 8.378738 7.868953  
## [2,] 8.378738 0.000000 5.834604  
## [3,] 7.868953 5.834604 0.000000  
##   
## $average.between  
## [1] 6.560206  
##   
## $average.within  
## [1] 4.514966  
##   
## $n.between  
## [1] 59304  
##   
## $n.within  
## [1] 51381  
##   
## $max.diameter  
## [1] 17.38587  
##   
## $min.separation  
## [1] 1.108709  
##   
## $within.cluster.ss  
## [1] 5458.119  
##   
## $clus.avg.silwidths  
## 1 2 3   
## 0.1526675 0.2350492 0.2476107   
##   
## $avg.silwidth  
## [1] 0.234631  
##   
## $g2  
## NULL  
##   
## $g3  
## NULL  
##   
## $pearsongamma  
## [1] 0.4755281  
##   
## $dunn  
## [1] 0.0637707  
##   
## $dunn2  
## [1] 0.9170859  
##   
## $entropy  
## [1] 0.8902657  
##   
## $wb.ratio  
## [1] 0.6882355  
##   
## $ch  
## [1] 128.6964  
##   
## $cwidegap  
## [1] 9.819063 4.328084 8.251368  
##   
## $widestgap  
## [1] 9.819063  
##   
## $sindex  
## [1] 1.614774  
##   
## $corrected.rand  
## NULL  
##   
## $vi  
## NULL

All the above elements can be used to evaluate the internal quality of clustering: 1.cluster.number: 3 cluster.size: 46 ,150 ,275 , size of the three clusters respectively average.distance, median.distance: vector containing the cluster-wise within average/median distances average.between: 6.344849, We want it to be as large as possible average.within: 4.314419 ,average distance within clusters. We want it to be as small as possible Simmilarly , we can check the dunn value , avg silwidth and within.cluster.ss.

## 6. Comparing the summaries of the clusters:

#Now we can compare the summaries of the clusters  
cluster\_description<-cluster.Description(Norm\_Universities,k3$cluster,sdType = "sample")  
  
print( "Arithmetic Mean")

## [1] "Arithmetic Mean"

cluster\_description[1:3,,1] # this will show the Arithmetic mean of the 3 clusters

## [,1] [,2] [,3] [,4] [,5] [,6] [,7]   
## [1,] "1.9818" "2.2299" "2.4447" "0.1334" "0.2546" "2.5228" "1.7487"   
## [2,] "0.0988" "-0.0178" "-0.1514" "0.9617" "0.941" "-0.2191" "-0.3234"  
## [3,] "-0.3668" "-0.35" "-0.3196" "-0.4889" "-0.4983" "-0.2993" "-0.1241"  
## [,8] [,9] [,10] [,11] [,12] [,13] [,14]   
## [1,] "-1.05" "-0.4918" "-0.0388" "-0.1746" "0.4953" "0.1636" "0.9386"   
## [2,] "1.0854" "1.1545" "0.7411" "0.7756" "0.0125" "0.082" "-0.4757"  
## [3,] "-0.3586" "-0.482" "-0.3539" "-0.3489" "-0.0857" "-0.0662" "0.0802"   
## [,15] [,16] [,17] [,18]   
## [1,] "0.6841" "0.614" "-0.2538" "-0.1681"  
## [2,] "0.8021" "-0.7017" "0.8852" "-0.4848"  
## [3,] "-0.4999" "0.2423" "-0.3894" "0.2627"

print("Standard Deviation")

## [1] "Standard Deviation"

cluster\_description[1:3,,2] #this will show the Standard Deviation of the 3 clusters

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]   
## [1,] "1.6917" "1.5945" "1.2616" "0.8257" "0.8449" "1.1956" "2.2411" "0.6663"  
## [2,] "0.7391" "0.5696" "0.4808" "1.0718" "0.7563" "0.3959" "0.3402" "0.62"   
## [3,] "0.3988" "0.4061" "0.4556" "0.5399" "0.7536" "0.4549" "0.4899" "0.715"   
## [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16]   
## [1,] "0.6862" "1.0092" "0.8043" "1.4592" "0.7301" "1.0926" "0.3739" "1.0731"  
## [2,] "0.7115" "1.0009" "0.8744" "0.9304" "0.9987" "0.6348" "0.5139" "0.7533"  
## [3,] "0.6383" "0.7806" "0.8697" "0.9202" "1.0344" "1.0033" "0.9269" "0.9125"  
## [,17] [,18]   
## [1,] "0.8026" "0.8665"  
## [2,] "0.6347" "1.1423"  
## [3,] "0.8987" "0.8428"

print("Median")

## [1] "Median"

cluster\_description[1:3,,3] # This will show the Median of the 3 three clusters

## [,1] [,2] [,3] [,4] [,5] [,6] [,7]   
## [1,] "1.6061" "1.8642" "2.1224" "-0.136" "0.0663" "2.1917" "1.3324"   
## [2,] "-0.1773" "-0.1929" "-0.3186" "0.7569" "0.952" "-0.3585" "-0.4473"  
## [3,] "-0.5265" "-0.5036" "-0.4977" "-0.5418" "-0.5241" "-0.4873" "-0.2963"  
## [,8] [,9] [,10] [,11] [,12] [,13] [,14]   
## [1,] "-1.2025" "-0.7118" "-0.4853" "-0.2804" "0.2121" "0.2526" "0.7202"   
## [2,] "1.1786" "1.2371" "0.6575" "0.5258" "-0.2502" "-0.1764" "-0.6042"  
## [3,] "-0.1527" "-0.4456" "-0.4222" "-0.3562" "-0.3499" "-0.2989" "-0.0776"  
## [,15] [,16] [,17] [,18]   
## [1,] "0.7676" "0.8816" "-0.1412" "-0.0766"  
## [2,] "0.9476" "-0.6829" "0.9058" "-0.1881"  
## [3,] "-0.3725" "0.1507" "-0.4167" "0.4522"

## 7. Assigning clusters to the rows of the University dataframe to analyze data based on clusters:-

#Adding new column 'cluster' to mention the cluster no. in dataset  
Norm\_Universities\_cluster <- data.frame(Norm\_Universities,  
cluster = as.factor(k3$cluster))  
  
#Binding the 2 categorical measurements into the normalized University data  
Norm\_Universities\_cluster<-cbind.data.frame(Norm\_Universities\_cluster,Universities\_N[,c(2:3)])  
head(Norm\_Universities\_cluster)

## X..appli..rec.d X..appl..accepted X..new.stud..enrolled  
## 1 -0.7253139 -0.7656329 -0.7925715  
## 3 -0.7368529 -0.7772155 -0.7554388  
## 10 -0.5750612 -0.5890979 -0.5391950  
## 12 -0.6234268 -0.6162571 -0.7139374  
## 22 0.3109878 -0.2248447 -0.4867723  
## 26 -0.3315143 -0.3207008 0.1717883  
## X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad  
## 1 -0.6500683 -0.5732933 -0.7097404  
## 3 -1.2994472 -1.5573355 -0.6576975  
## 10 2.1097921 1.5915994 -0.4683728  
## 12 -0.1089192 -0.4256870 -0.6478457  
## 22 0.1075405 0.2139404 -0.5686035  
## 26 -0.2171490 -1.0161123 0.7275427  
## X..PT.undergrad in.state.tuition out.of.state.tuition room board  
## 1 0.0462840 -0.3347297 -0.6993021 -0.8428467 0.6669350  
## 3 0.6802614 -1.3893276 -1.2406234 0.4106795 0.2259098  
## 10 -0.3819742 0.4084555 0.2516051 -0.2399203 0.5434479  
## 12 -0.4343744 -0.2404720 -0.5786992 -1.1793639 0.7374990  
## 22 -0.4389028 -0.6780450 -1.1385749 -1.1176691 -1.0266018  
## 26 2.5233243 -1.3026831 -1.4229193 -0.4011680 1.9723696  
## add..fees estim..book.costs estim..personal.. X..fac..w.PHD  
## 1 -0.6997824 1.5394532 0.2758088 0.16752615  
## 3 -0.9695550 -0.2989446 -0.2199034 -2.05260943  
## 10 -0.7278837 -0.9117438 -0.6041537 0.04751883  
## 12 -0.7840863 -0.2989446 -0.3108329 -0.61252148  
## 22 0.1095355 2.7650518 0.1291484 -1.03254714  
## 26 -0.2473512 1.2330536 1.3024317 1.36759945  
## stud..fac..ratio Graduation.rate AcceptanceRate cluster State  
## 1 -0.529035524 -2.7862940 0.12436460 3 AK  
## 3 -1.144600894 -1.4637549 0.40419384 3 AK  
## 10 0.009584174 0.3547362 -0.03796041 2 AL  
## 12 -0.657278310 -1.1882260 0.74012282 3 AL  
## 22 0.394312530 -1.0780144 -2.47270114 3 AL  
## 26 -1.862760492 -1.7943897 -0.22038300 3 AL  
## Public..1...Private..2.  
## 1 2  
## 3 1  
## 10 2  
## 12 2  
## 22 2  
## 26 1

#Creating subsets for each cluster based on the cluster number  
Norm\_Universities\_cluster1<-subset(Norm\_Universities\_cluster,Norm\_Universities\_cluster$cluster==1)  
Norm\_Universities\_cluster2<-subset(Norm\_Universities\_cluster,Norm\_Universities\_cluster$cluster==2)  
Norm\_Universities\_cluster3<-subset(Norm\_Universities\_cluster,Norm\_Universities\_cluster$cluster==3)

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

#Using parallel plot for University data  
parallelplot(Norm\_Universities\_cluster)

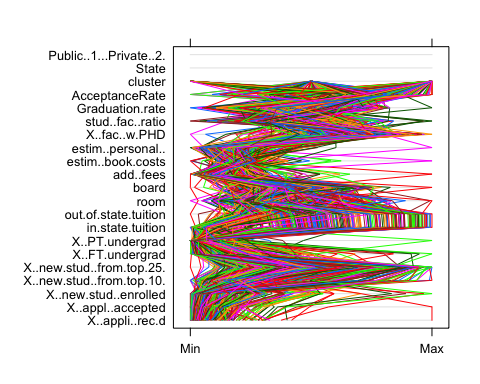
## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in min(as.numeric(x), na.rm = TRUE): no non-missing arguments to min;  
## returning Inf

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in max(as.numeric(x), na.rm = TRUE): no non-missing arguments to max;  
## returning -Inf

## Warning in (function (x, y, z, subscripts, groups = NULL, col =  
## superpose.line$col, : NAs introduced by coercion  
  
## Warning in (function (x, y, z, subscripts, groups = NULL, col =  
## superpose.line$col, : NAs introduced by coercion



## 9.1 Categorizing the data into below parts to understand the relation better:-

## 1. Application received and Application accepted , Acceptance rate per cluster

## 2. Total Students enrollment, students enrolled from top 10 and 25 schools

## 3. FT undergraduate , PT undergraduate

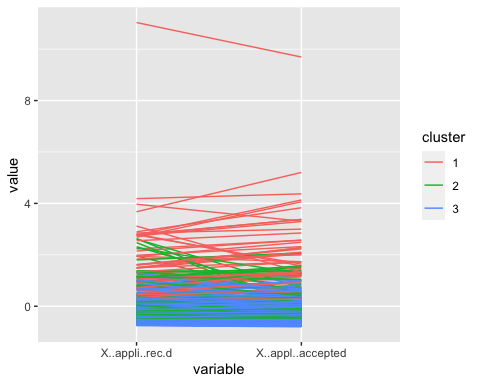
## 4. Tuition Fees

## 5. Additional Expenses

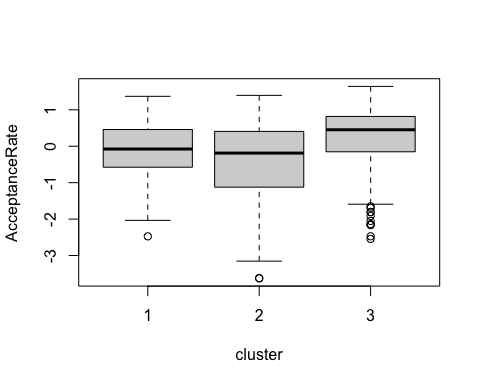
## 6. Student Faculity ratio and PHD Faculity

## 7. Graduation Rate

# 1.Applications received and Application accepted by clusters  
ggparcoord(Norm\_Universities\_cluster,columns=1:2,groupColumn = 19)



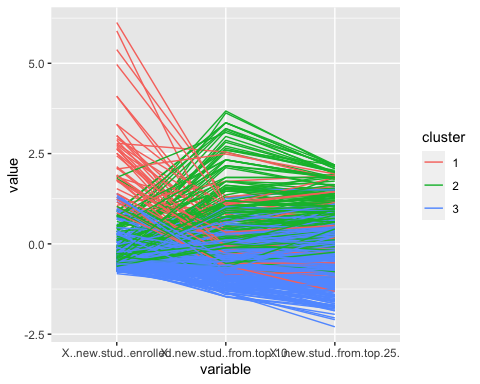
#determining the acceptance rate of universities by clusters  
boxplot(AcceptanceRate~cluster, data=Norm\_Universities\_cluster)

 ##Interpretation of graph: Cluster 1: Highest level of application received and moderate level in acceptance rate

Cluster 2: Moderate level of application received and Lowest in the acceptance rate (less no of students got acceptance from these universities)

Cluster 3: least no. application received and highest level in acceptance rate (more students get admission into these universities)

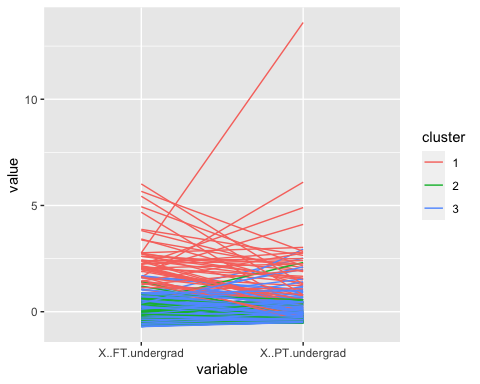
# 2.Total Students enrollment,Students enrolled from top 10 and top 25  
ggparcoord(Norm\_Universities\_cluster,columns=3:5,groupColumn = 19)

 ##Interpretation of graph: Cluster 1: Highest no. of new students enrollment however, least from the top 10 and top 25 schools schools

Cluster 2: Moderate no. of new students enrollment to the Universities however, highest in number of students from top 10 and 25 schools

Cluster 3: Moderate in the new students enrollment and moderate number of students from top 10 and (moderate-lowest)from top 25

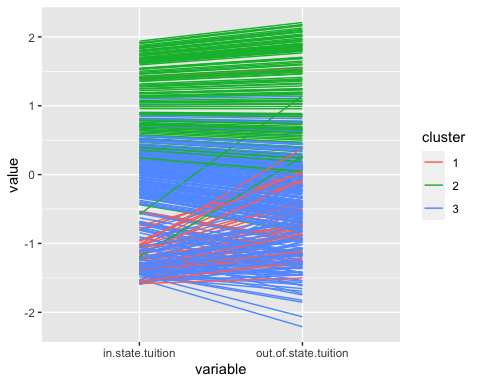
# 3.FT undergraduate , PT undergraduate  
ggparcoord(Norm\_Universities\_cluster,columns=6:7,groupColumn = 19)

 ##Interpretation of graph: Cluster 1: Highest in Full-Time under graduation and as well as part-time under graduation

Cluster 2: Moderate level of FT undergrad and PT undergrad

Cluster 3: Lowest level of Full time under graduation and lowest level in part time under graduation

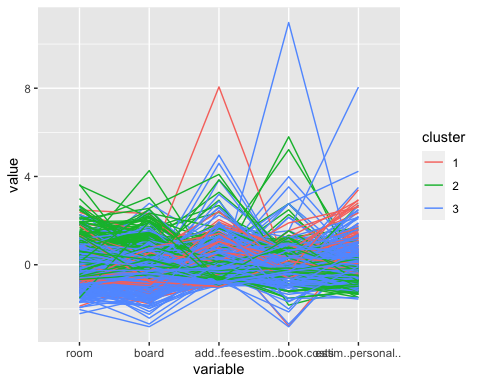
# 4.Tuition Fees  
ggparcoord(Norm\_Universities\_cluster,columns=8:9,groupColumn = 19)

 ##Interpretation of graph: Cluster 1: In state tuition fees seems to be lowest and moderate level of out of state tuition fees

CLuster 2: In state tuition fees seems to be highest and highest in out of state tuition fees

Cluster 3: In state tuition fees seems to be between moderate to lowest and lowest out of state fees

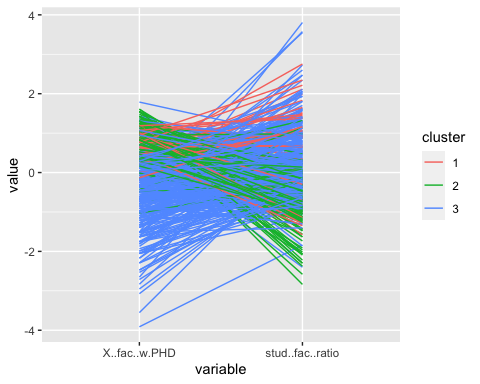
# 5.Additional Expenses  
ggparcoord(Norm\_Universities\_cluster,columns=10:14,groupColumn = 19)

 ##Interpretation of graph: Cluster 1:moderate room and board expenses except for additional fees, books cost and personal expenses

Cluster 2:Highest in room , board and estimated personal expenses however, lower in estimated personal expenses and moderate additional fees (College might be providing some facilities to the students which leads lower personal expenses)

Cluster 3:Lowest in room , board and books cost as well some spike in estimated book costs and personal expenses(some colleges might not be having sufficient books in the libraries to support all students)

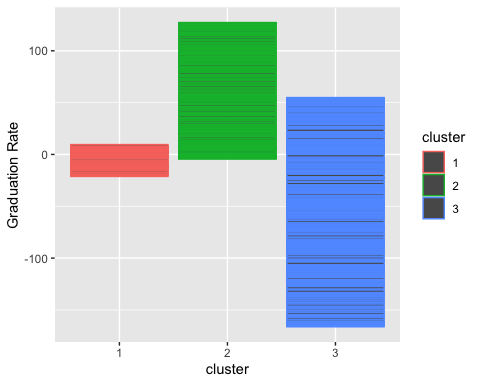
# 6.Student Faculity ratio and Faculity PHD  
ggparcoord(Norm\_Universities\_cluster,columns=15:16,groupColumn = 19)

 ##Interpretation of graph: Cluster 1: Moderate to high in no. of faculty with PHD degree and moderate level of no. of students per faculty(moderate level in quality of education for student)

Cluster 2: Highest in no. of faculty with PHD and less students per faculty(Better education for the students)

Cluster 3: Lowest in no. of faculty with PHD and more no. of students per Faculty (lowest in quality of education for student )

# 7.Graduation Rate  
Norm\_Universities\_cluster%>% group\_by(cluster,Graduation.rate)%>%   
 ggplot(aes(Norm\_Universities\_cluster,x= cluster, y = Graduation.rate,col=cluster)) +  
 geom\_col() +  
 ylab("Graduation Rate")

 ##Interpretation of graph: Cluster 1: Moderate level of graduation rate

Cluster 2: Highest level of graduation rate

Cluster 3: Lowest level of graduation rate

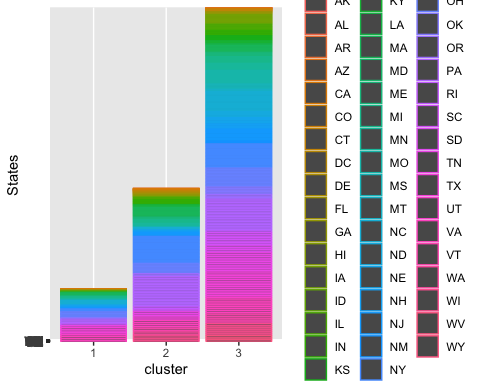
With the help of above graphs , we can see that there is no positive association of public/private universities with clusters1 and cluster 3 however, there is strong relation between cluster 2 and Private university .Therefore with the help of table () , we can see ratio of public/private universities within clusters and understand more about the participation .

## 10. 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?

# Relationship with categorized variable with clusters  
  
#Converting categorical variable to factors  
Norm\_Universities\_cluster$Public..1...Private..2.<- as.factor(Norm\_Universities\_cluster$Public..1...Private..2.)  
Norm\_Universities\_cluster$cluster<- as.factor(Norm\_Universities\_cluster$cluster)  
Norm\_Universities\_cluster$State <-as.factor(Norm\_Universities\_cluster$State)  
  
#Creating table of Public/Private Universities and cluster to determine the relation  
StateCluster\_table= table(Norm\_Universities\_cluster$State,Norm\_Universities\_cluster$cluster)  
print(StateCluster\_table)

##   
## 1 2 3  
## AK 0 0 2  
## AL 0 1 3  
## AR 0 0 4  
## AZ 2 0 0  
## CA 2 10 3  
## CO 0 1 5  
## CT 1 4 5  
## DC 0 4 0  
## DE 1 0 1  
## FL 1 4 3  
## GA 1 3 3  
## HI 0 0 1  
## IA 0 2 16  
## ID 0 0 2  
## IL 2 4 9  
## IN 0 7 8  
## KS 0 0 7  
## KY 0 2 4  
## LA 1 2 2  
## MA 3 11 8  
## MD 1 1 1  
## ME 0 2 4  
## MI 2 4 7  
## MN 1 3 7  
## MO 1 1 13  
## MS 0 0 5  
## MT 0 0 2  
## NC 4 2 17  
## ND 0 0 5  
## NE 1 1 5  
## NH 1 1 4  
## NJ 1 3 9  
## NM 0 0 2  
## NY 2 19 17  
## OH 4 7 13  
## OK 1 0 5  
## OR 0 2 3  
## PA 3 19 20  
## RI 1 2 1  
## SC 0 1 8  
## SD 0 0 4  
## TN 1 3 11  
## TX 4 2 14  
## UT 1 0 1  
## VA 3 4 8  
## VT 0 2 5  
## WA 0 2 0  
## WI 0 3 6  
## WV 0 0 2  
## WY 0 0 1

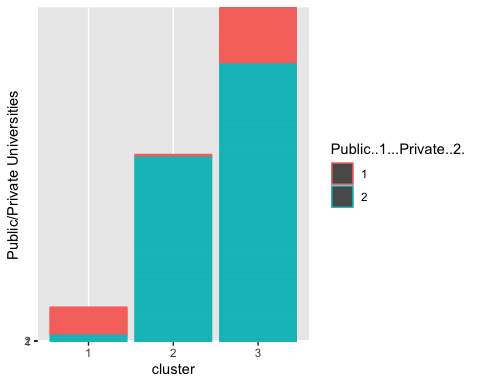
#Plotting Bar graph  
Norm\_Universities\_cluster%>% group\_by(cluster,State)%>%   
 ggplot(aes(Norm\_Universities\_cluster,x= cluster, y = State,col=State)) +  
 geom\_col() +  
 ylab("States")



#Creating table of State and cluster to determine the relation  
PublicPrivatecluster\_table = table(Norm\_Universities\_cluster$Public..1...Private..2.,Norm\_Universities\_cluster$cluster)  
print(PublicPrivatecluster\_table)

##   
## 1 2 3  
## 1 41 4 83  
## 2 5 135 203

#Plotting Bar graph  
Norm\_Universities\_cluster%>% group\_by(cluster,Public..1...Private..2.)%>%   
 ggplot(aes(Norm\_Universities\_cluster,x= cluster, y = Public..1...Private..2.,col=Public..1...Private..2.)) +  
 geom\_col() +  
 ylab("Public/Private Universities")

 As per the above analysis, we can see that:- Cluster 1: It is mainly public university Cluster 2: It is mainly private university Cluster 3: It is mostly private universities(with few public universities as well)

however , there is was no specific relation between the states and the clusters

Overall ,I have interpreted the clusters with below mentioned main features:-

Cluster 1:- 1. It is mainly public university

1. Highest level of application received and moderate level in acceptance rate
2. Moderate level in the new students enrollment and moderate number of students from top 10 and (moderate-lowest)from top 25 schools (seems to be known good public universities)
3. Highest in Full-Time under graduation and as well as part-time under graduation
4. In state tuition fees seems to be lowest and moderate level for out of state tuition fees (students outside states enrolling for the courses chances of International students)
5. Moderate room and board expenses except for additional fees, books cost and personal expenses
6. Moderate to high in no. of faculty with PHD degree and moderate level of no. of students per faculty(moderate level in quality of education for student)
7. Moderate level of graduation rate

Cluster 2:- 1. It is mainly private universities (Ivy League school)

1. Moderate level of application received and Lowest in the acceptance rate (less no of students got acceptance from these universities)
2. Moderate no. of new students enrollment to the Universities however, highest in number of students from top 10 and 25 schools
3. Moderate level of FT undergrad and PT undergrad
4. Highest in room , board and estimated personal expenses however, lower in estimated personal expenses and moderate additional fees (College might be providing some facilities to the students which leads lower personal expenses)
5. In state tuition fees seems to be highest and highest in out of state tuition fees
6. Highest in no. of faculty with PHD and less students per faculty(Better education for the students-These Universities have more qualified faculty and assigning them into small groups of students which provides better learning experience )
7. Highest level of graduation rate

Cluster 3:- 1. It is partially private/public universities(with majority as private universities)

1. least no. application received and highest level in acceptance rate (more students get admission into these universities)
2. Highest no. of new students enrollment however, least from the top 10 and top 25 schools
3. Lowest level of Full time under graduation and lowest level in part time under graduation
4. In state tuition fees seems to be between moderate to lowest and lowest out of state fees
5. Lowest in room , board and books cost as well some spike in estimated book costs and personal expenses(some colleges might not be having sufficient books in the libraries to support all students)
6. Lowest in no. of faculty with PHD and more no. of students per Faculty (lowest in quality of education for student )
7. Lowest level of graduation rate

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

Below are some external information which explains the content of some or all these clusters:-

1. International Students :- International students are likely to get themselves enrolled in the Elite colleges(Public/Private) and they are more focused towards their graduation completion.
2. Ethnicity:- There are certain factors like Ethnicity which tells that students from a certain ethnic background performs better than other students.
3. Male/Female students :- Many studies shows that Female perform better in achieving educational success because they are more focused and better organized.
4. Instructional Expenditure per student :- In cluster 2 , based on the results we can assume that student have provided with better facilities on the campus and therefore their personal and books expenses are less.for example: library , gym , transportation. Also , more no. of highly qualified faculties in the college helps them in getting better quality of education.

## 12.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.

#Filtering the Tufts University data from the data frame  
Tufts\_data<-filter(Universities,Universities$College.Name=="Tufts University")  
  
#selecting the continuous data  
Tufts\_data<- Tufts\_data[,c(4:20)]  
  
University\_new <- n\_Universities[,c(1:17)]  
  
#Normalization of the data  
train\_norm\_df <- University\_new  
Tufts\_norm\_df <- Tufts\_data   
  
# use preProcess() from the caret package to normalize   
norm\_values <- preProcess(University\_new, method=c("center", "scale"))  
  
train\_norm\_df <- predict(norm\_values, University\_new)   
Tufts\_norm\_df <- predict(norm\_values, Tufts\_data)   
  
  
# To determine the distance between centers and Tufts University,Binding the data with k3 centers  
k3centre\_Tufts<-data.frame(rbind(k3$centers[,-18],Tufts\_norm\_df))  
  
#Distance  
distance\_Tufts<- dist(k3centre\_Tufts)  
print(distance\_Tufts)

## 1 2 3  
## 2 6.509304   
## 3 5.960895 4.044232   
## 11 6.905137 2.598750 6.561214

#Minimum distance = closest cluster  
min(distance\_Tufts) #2.75

## [1] 2.59875

#Since I have found the closest cluster to the Tufts University which cluster 2 , we can now add the mean value of the cluster 2 to NA value of Tuft University record.  
  
# Using cluster\_description to determine the mean of the 2nd cluster , 7th column (X..PT.undergrad)  
  
cluster\_description[2,7,1]

## [1] "-0.3234"

#Imputing the missing value in Tufts University data(Normalized data)  
Tufts\_norm\_df[,7]<- cluster\_description[2,7,1]  
  
head(Tufts\_norm\_df)

## X..appli..rec.d X..appl..accepted X..new.stud..enrolled  
## 1 1.096623 0.6158933 0.4633898  
## X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad  
## 1 1.730988 1.690004 0.2216773  
## X..PT.undergrad in.state.tuition out.of.state.tuition room board  
## 1 -0.3234 1.866005 2.116543 1.145408 1.425498  
## add..fees estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio  
## 1 0.3483966 0.3138547 -0.5630888 1.54761 -0.9394124  
## Graduation.rate  
## 1 1.456852