Problem4-2

Mehrab Atighi

11/18/2021

problem

University Rankings. The dataset on American college and university rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or a public school).

a. Remove all categorical variables. Then remove all records with missing numerical measurements from the dataset.
b. Conduct a principal components analysis on the cleaned data and comment on the results. Should the data be normalized? Discuss what characterizes the components you consider key

solution

At the first we should add data in R and see head of dataset.

```
Data<-read.csv("F:/lessons/Data mining/Data/Universities.csv")
#View(Data)
head(Data,4)
```

```
##                         College.Name State Public..1...Private..2.
## 1         Alaska Pacific University    AK                       2
## 2 University of Alaska at Fairbanks    AK                       1
## 3    University of Alaska Southeast    AK                       1
## 4 University of Alaska at Anchorage    AK                       1
##   X..appli..rec.d X..appl..accepted X..new.stud..enrolled
## 1             193               146                    55
## 2            1852              1427                   928
## 3             146               117                    89
## 4            2065              1598                  1162
##   X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                        16                        44             249
## 2                        NA                        NA            3885
## 3                         4                        24             492
## 4                        NA                        NA            6209
##   X..PT.undergrad in.state.tuition out.of.state.tuition room board add..fees
## 1             869             7560                 7560 1620  2500       130
## 2            4519             1742                 5226 1800  1790       155
## 3            1849             1742                 5226 2514  2250        34
## 4           10537             1742                 5226 2600  2520       114
##   estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
## 1               800              1500            76             11.9
## 2               650              2304            67             10.0
## 3               500              1162            39              9.5
## 4               580              1260            48             13.7
##   Graduation.rate
## 1              15
## 2              NA
## 3              39
```

solution

Now we want to remove the categorical variables after that remove na (missing data) data.

a

```
data <- Data[,-c(2,3)]
chek_na<- is.na(data)
n=1
index<-c()
for( i in 1:nrow(chek_na)){
  if(sum(chek_na[i,])>=1){
    index[n]=i
    n=n+1
  }
}

data<-data[-index,]
dim(data)
```

```
## [1] 471  18
```

solution

Now we are going to solve part b.
thus we know we had 2 categorical variable and 831 missing data.
i think that we need to normalaizing data cuse we have alot of variable with diffrent scales.
now we want to do a dimention reduction with principal components method:

```
pca<-prcomp(data[,-1] ,scale. = TRUE ,center = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.2749 2.1426 1.09838 1.03247 0.97599 0.87284 0.80327
## Proportion of Variance 0.3044 0.2700 0.07097 0.06271 0.05603 0.04481 0.03796
## Cumulative Proportion  0.3044 0.5745 0.64542 0.70813 0.76416 0.80898 0.84693
##                           PC8     PC9   PC10    PC11    PC12   PC13    PC14
## Standard deviation     0.77279 0.70316 0.6622 0.62788 0.54973 0.4383 0.30389
## Proportion of Variance 0.03513 0.02908 0.0258 0.02319 0.01778 0.0113 0.00543
## Cumulative Proportion  0.88206 0.91115 0.9369 0.96013 0.97791 0.9892 0.99464
##                           PC15    PC16    PC17
## Standard deviation     0.20002 0.17428 0.14388
## Proportion of Variance 0.00235 0.00179 0.00122
## Cumulative Proportion  0.99700 0.99878 1.00000
```

solution

```
head(pca$rotation,2)
```

```
##                            PC1        PC2        PC3         PC4        PC5
## X..appli..rec.d   0.07836149 -0.4201638 0.03198244 -0.07262064 0.01669353
## X..appl..accepted 0.02365875 -0.4344710 0.03142262 -0.11812757 0.08907266
##                            PC6       PC7         PC8        PC9        PC10
## X..appli..rec.d   -0.1123199 0.2681455 -0.09356958 0.03962825 -0.08736098
## X..appl..accepted -0.1143806 0.2662853 -0.08099058 0.02279461  0.03519709
##                           PC11        PC12       PC13       PC14       PC15
## X..appli..rec.d   -0.07302129 0.009995194 -0.6029957 -0.1987904 -0.3467745
## X..appl..accepted -0.16604598 0.062100043 -0.2512570  0.2402318  0.4523467
##                          PC16       PC17
## X..appli..rec.d   -0.3446373  0.2463541
## X..appl..accepted  0.4298300 -0.3922380
```

```
head(pca$x, 2)
```

```
##         PC1      PC2       PC3        PC4          PC5       PC6        PC7
## 1 -1.551795 1.449883 -2.010113  0.3875416 -0.09962324 0.3773497 -1.3796057
## 3 -2.585562 1.863903 -1.445699 -0.8579998  1.03470364 0.6262750  0.3485298
##          PC8       PC9       PC10       PC11        PC12       PC13        PC14
## 1 -0.7440404 0.2899511 -1.0925960 -1.7252170  0.01193685 -0.3753815  0.16509624
## 3 -1.1917674 0.3019966 -0.6615726  0.5796455 -1.46876810 -0.1257514 -0.09656765
##         PC15       PC16       PC17
## 1  0.1042940 -0.1275235 0.03213121
## 3 -0.1626597  0.2989030 0.08019281
```

solution

so now we can see that we can reduce our dimantion to 6 for least 80% of all variance. for example the first component have 30% the second 27% the third 7% and etc.
the pca$x values are our values with new rotation.
the rotation values come back to cofficient of each variable and there we have just for first and second variable.
Now we want to plot the clean, normalize dataset.

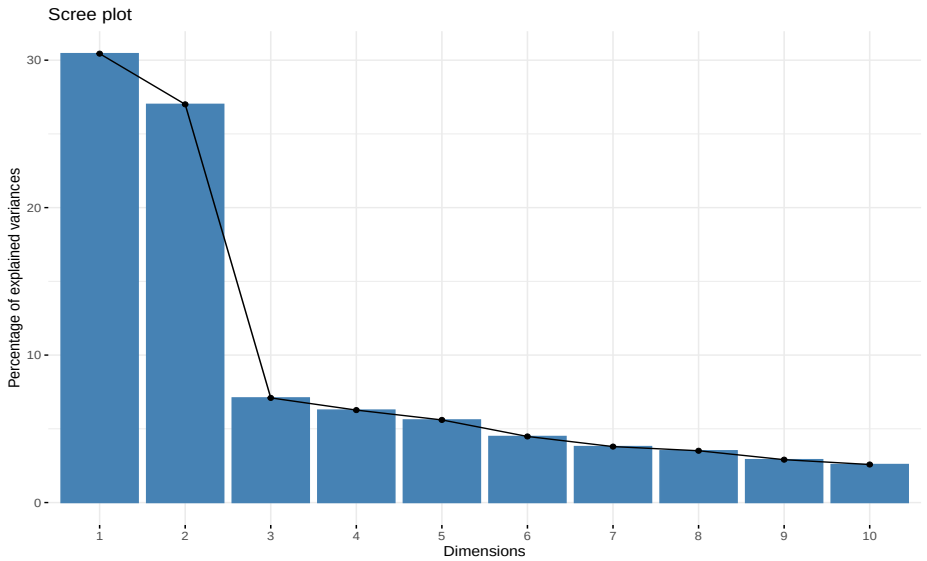
solution

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

```
## Loading required package: ggplot2
```

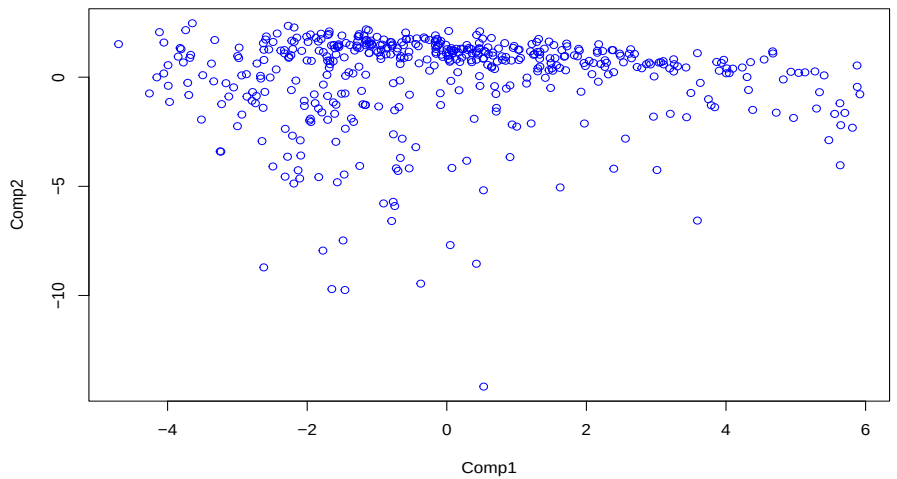
```
## Welcome! Want to learn more? See two factoextra-related
```

```
fviz_eig(pca)#plot(pca)
```



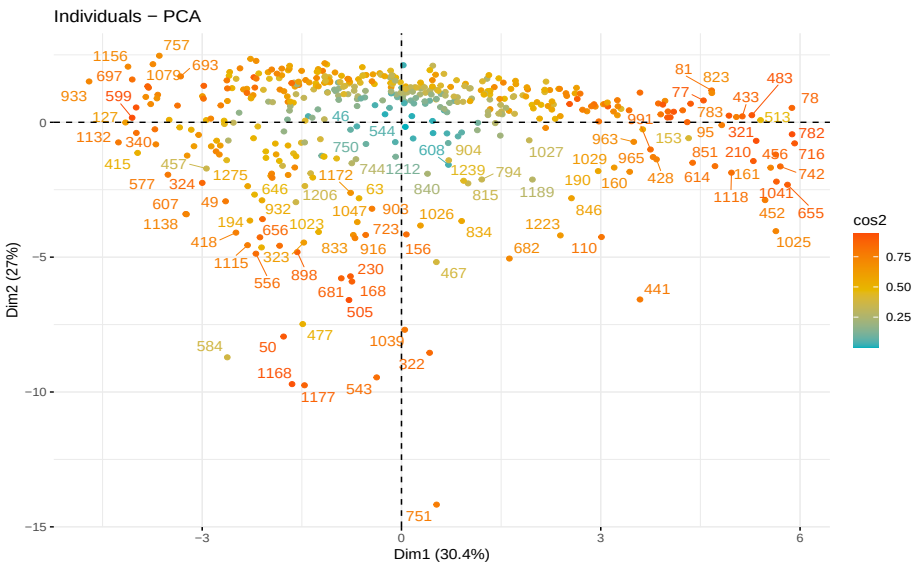
solution

```
plot(pca$x[,1],pca$x[,2]
     ,xlab = "Comp1" , ylab="Comp2" ,col="Blue")
```



solution

```
#biplot(pca)
fviz_pca_ind(pca,
             col.ind = "cos2", # Color by the quality of r
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4
             repel = TRUE)      #Avoid text overlapping)
```



solution

```
fviz_pca_biplot(pca, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
                )
```