**PRINCIPAL COMPONENT ANALYSIS**

# **DATATYPES: Here the data is discrete type of data**

## **SYNTAX:**

**Mydata<-read\_csv # Load the university dataset**

**View(mydata) # To load the dataset we have to check all the columns**

**Help(princop) # To understand the principle component analysis**

**View(mydata[-1]) # The first column in my data has university names**

**Mydata[-1] # Considering only numerical columns for applying PCA**

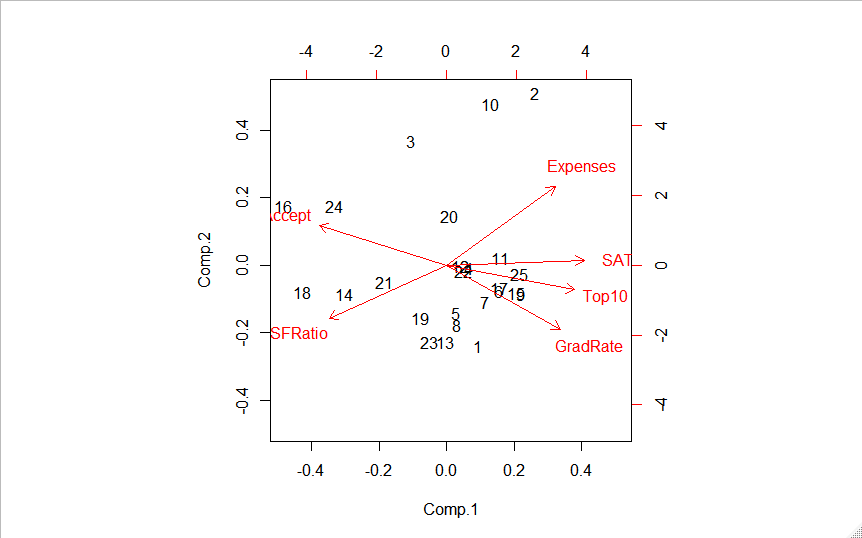
**Data<-mydata[-1] # We have to remove the first column in the data**

**Attach(data) # We have to attach the data**

**Cor(data) # We have to calculate the correlation of the data**

|  |
| --- |
| SAT Top10 Accept SFRatio Expenses GradRate  SAT 1.0000000 0.9225222 -0.8858496 -0.8125517 0.7789760 0.7477120  Top10 0.9225222 1.0000000 -0.8591811 -0.6434351 0.6114666 0.7459420  Accept -0.8858496 -0.8591811 1.0000000 0.6316636 -0.5584395 -0.8195495  SFRatio -0.8125517 -0.6434351 0.6316636 1.0000000 -0.7818394 -0.5609217  Expenses 0.7789760 0.6114666 -0.5584395 -0.7818394 1.0000000 0.3935914  GradRate 0.7477120 0.7459420 -0.8195495 -0.5609217 0.3935914 1.0000000  **Pcaobj<-princomp(mydata[-1],cor=true,scores=true,covmat=null)**  **# For building we have to use the command**  **Summary(pcaobj) # We have to find out the summary of the data**  Importance of components:  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  Standard deviation 2.1475766 0.8870266 0.53531473 0.40469755 0.3525708 0.162636495  Proportion of Variance 0.7686808 0.1311360 0.04776031 0.02729668 0.0207177 0.004408438  Cumulative Proportion 0.7686808 0.8998169 0.94757718 0.97487386 0.9955916 1.000000000  **Loadings(pcaobj) # We have to find out the loadings of the data**  Loadings:  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  SAT 0.458 0.187 0.131 0.858  Top10 0.427 -0.200 0.498 0.375 0.482 -0.396  Accept -0.424 0.321 -0.156 0.801 0.217  SFRatio -0.391 -0.433 0.606 -0.507 0.172  Expenses 0.363 0.634 0.205 -0.623 -0.174  GradRate 0.379 -0.516 -0.532 -0.439 0.338  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  SS loadings 1.000 1.000 1.000 1.000 1.000 1.000  Proportion Var 0.167 0.167 0.167 0.167 0.167 0.167  Cumulative Var 0.167 0.333 0.500 0.667 0.833 1.000  **Plot(pcaobj) # Graph showing importance of principal components** |
|  |
| |  | | --- | |  | |

**Biplot(pcaobj) # The differences between the plots in my dataset**



**Pcaobj$scores[,1:3] # Top 3 PCA scores which represents the whole data**

Comp.1 Comp.2 Comp.3

[1,] 1.00987445 -1.06430962 0.08106631

[2,] 2.82223781 2.25904458 0.83682883

[3,] -1.11246577 1.63120889 -0.26678684

[4,] 0.74174122 -0.04218747 0.06050086

[5,] 0.31191206 -0.63524357 0.01024052

[6,] 1.69669089 -0.34436328 -0.25340751

[7,] 1.24682093 -0.49098366 -0.03209382

[8,] 0.33874978 -0.78516859 -0.49358483

[9,] 2.37415013 -0.38653888 0.11609839

[10,] 1.40327739 2.11951503 -0.44282714

[11,] 1.72610332 0.08823712 0.17040366

[12,] 0.45085748 -0.01113295 -0.17574605

[13,] -0.04023814 -1.00920438 -0.49651717

[14,] -3.23373034 -0.37458049 -0.49537282

[15,] 2.23626502 -0.37179329 -0.39899365

[16,] -5.17299212 0.77991535 -0.38591233

[17,] 1.69964377 -0.30559745 0.31850785

[18,] -4.57814600 -0.34759136 1.49964176

[19,] -0.82260312 -0.69890615 1.42781145

[20,] 0.09776213 0.65044645 0.10050844

[21,] -1.96318260 -0.22476756 -0.25588143

[22,] 0.54228894 -0.07958884 -0.30539348

[23,] -0.53222092 -1.01716720 -0.42371636

[24,] -3.54869664 0.77846167 -0.44936332

[25,] 2.30590032 -0.11770432 0.25398866

**Mydata<-cbind(mydata,pca$scores[,1:3]) # Consdering top 3 principal component scores and combining with my data**

**View(mydata) # To bind the dataset we have to check all the columns**

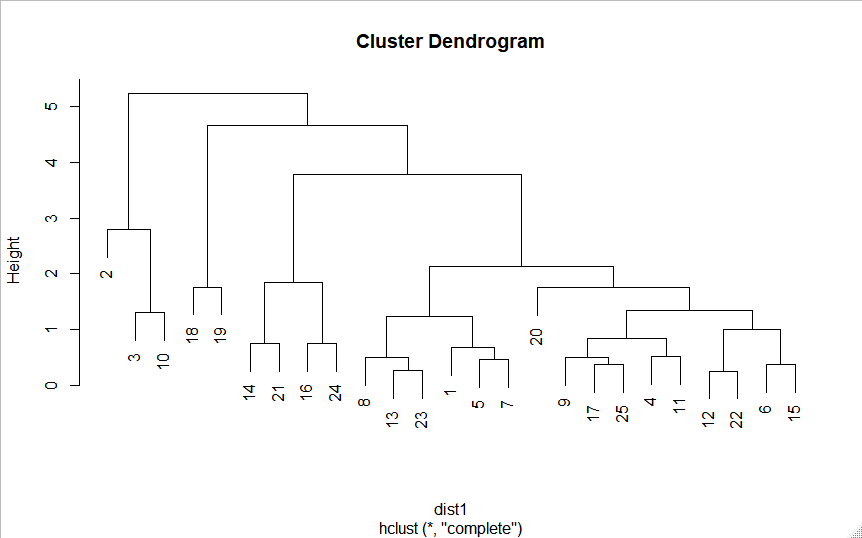
**Clust\_mydata[,8:10] # Preparing data for clustering**

**Norm\_clus<-scale(clus\_data) # Scale function is used to normalize the data**

**Dist1<-dist(norm\_clus,method=”Euclidean”) # The method for finding the distance**

**Fit1<-hclust(dist1,method=”complete”) # Clustering the data H cluster**

**Plot(fit1) # Displaying Dendogram**



**Groups<-cutree(fit1,3) # Cutting the dendogram for 3 clusters**

**Membership1<-as.matrix(groups) # Cluster membering**

**View(membership1) # After cluster membering we have to check the columns**

**Final1<-cbind(membership1,mydata) # Binding column wise with original data**

**View(final1) # To bind we have to check all the columns**

**View(aggregate(final1,-c[2,9:11],by=list(membership\_1),FUN=mean))**

**# Interferences can be drawn from the aggregate of the universities**

**Write.csv(final,file=”universities\_clustered\_csv”,row.names=F,column.names=F) # Wehave to save the life on the desktop**

**Getwd() # To check the file we have to use the command**