**Principal Component Analysis**

* Performing to reduce dimensionality of our dataset.

**Code:**

library(factoextra)

x<-data1[11:32]

names(x)

pc<-princomp(x,cor=TRUE,score=TRUE)

summary(pc)

**output:**

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

Standard deviation 2.0491100 1.7543605 1.5017542 1.37854155 1.13977177 1.1311820

Proportion of Variance 0.1908569 0.1398991 0.1025121 0.08638076 0.05904908 0.0581624

Cumulative Proportion 0.1908569 0.3307560 0.4332681 0.51964887 0.57869795 0.6368603

Comp.7 Comp.8 Comp.9 Comp.10 Comp.11

Standard deviation 1.1003670 1.07476536 0.97929807 0.9077245 0.88510508

Proportion of Variance 0.0550367 0.05250548 0.04359203 0.0374529 0.03560959

Cumulative Proportion 0.6918970 0.74440253 0.78799456 0.8254475 0.86105705

Comp.12 Comp.13 Comp.14 Comp.15 Comp.16

Standard deviation 0.80867974 0.77093383 0.66899178 0.60052233 0.56392244

Proportion of Variance 0.02972559 0.02701541 0.02034318 0.01639214 0.01445493

Cumulative Proportion 0.89078264 0.91779805 0.93814123 0.95453337 0.96898830

Comp.17 Comp.18 Comp.19 Comp.20 Comp.21

Standard deviation 0.468134799 0.462438297 0.373266075 0.330421623 2.742191e-02

Proportion of Variance 0.009961372 0.009720417 0.006333071 0.004962657 3.418005e-05

Cumulative Proportion 0.978949675 0.988670092 0.995003163 0.999965820 1.000000e+00

Comp.22

Standard deviation 6.276525e-07

Proportion of Variance 1.790671e-14

Cumulative Proportion 1.000000e+00

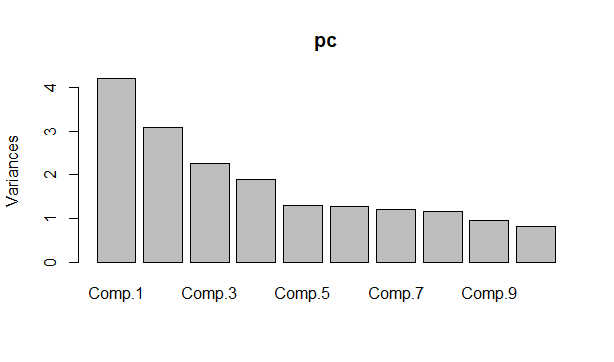
***Conclusion:***

***# 22 components will be there as we have 22 numerical variables.***

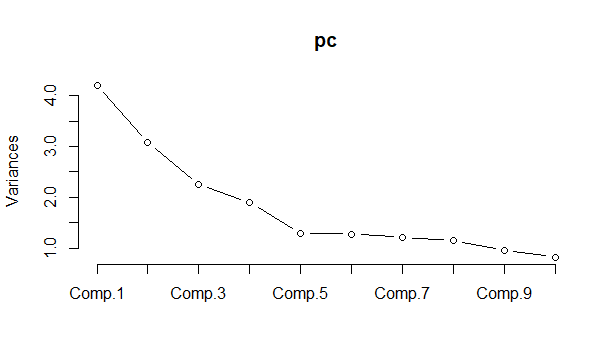
***# Component 1 explains maximum variance in data, followed by component 2 and so on.***

***# We can see that 80 % of the variance of data is explained by 10 components, hence we can take 10 principal components here.***

* plot(pc)



* plot(pc,type="l")



* attributes(pc)

**output:**

$names

[1] "sdev" "loadings" "center" "scale" "n.obs" "scores" "call"

$class

[1] "princomp"

**# Below code will give the contribution by each variable for each component**

* pc$loadings

**# Below code will give the principal component score for each observation.**

* pc$scores

**# Below code will give scores of all components**

* sc<-pc$scores
* dim(sc)

**Output:**

|  |
| --- |
| [1] 69598 22 |
|  |
| |  | | --- | |  | |

* df\_pc<-data.frame(sc)
* names(df\_pc)

**output:**

[1] "Comp.1" "Comp.2" "Comp.3" "Comp.4" "Comp.5" "Comp.6" "Comp.7" "Comp.8"

[9] "Comp.9" "Comp.10" "Comp.11" "Comp.12" "Comp.13" "Comp.14" "Comp.15" "Comp.16"

[17] "Comp.17" "Comp.18" "Comp.19" "Comp.20" "Comp.21" "Comp.22"

**# Creating final data-frame along with categorical variables and principal components**

* final\_df=cbind.data.frame(df\_pc[,1:10],data1$UID,data1$PPQD\_Bucket,data1$PPID,data1$Period,data1$RTM,data1$Country\_ISO\_Code,data1$Partner\_Name,data1$World\_Country\_Region,data1$QCV\_Segment)
* names(final\_df)

**output:**

[1] "Comp.1" "Comp.2"

[3] "Comp.3" "Comp.4"

[5] "Comp.5" "Comp.6"

[7] "Comp.7" "Comp.8"

[9] "Comp.9" "Comp.10"

[11] "data1$UID" "data1$PPQD\_Bucket"

[13] "data1$PPID" "data1$Period"

[15] "data1$RTM" "data1$Country\_ISO\_Code"

[17] "data1$Partner\_Name" "data1$World\_Country\_Region"

[19] "data1$QCV\_Segment"

* dim(final\_df)

**Output:**

[1] 69598 19

**# We can also see the correlation between our actual variables and**

**principal components**

* datad<-data1
* datad<-cbind.data.frame(datad,df\_pc[,1:10])
* names(datad)
* cor(x,datad[,36:45])

***Conclusion:***

***Thus final data-frame consist of categorical variables along with 10***

***principal components.***