Principal Component Analysis on Istanbul Airbnb Dataset

```
library(data.table)#Data. table is an extension of data. frame package in R.
It is widely used for fast aggregation of large datasets,
```

```
## Warning: package 'data.table' was built under R version 3.6.2
```

```
library(Hmisc)#data analysis funs
```

```
## Warning: package 'Hmisc' was built under R version 3.6.2
```

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

```
## Warning: package 'lattice' was built under R version 3.6.2
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.6.2
```

```
## Loading required package: Formula
```

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

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ---------------------------------------------- ti
dyverse 1.3.0 --

## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'stringr' was built under R version 3.6.2

## -- Conflicts ------------------------------------------------- tidyvers
e_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::src()       masks Hmisc::src()
## x dplyr::summarize() masks Hmisc::summarize()
## x purrr::transpose() masks data.table::transpose()

library(ggplot2)
library(plotly)

## Warning: package 'plotly' was built under R version 3.6.2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:Hmisc':
##
##     subplot

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.6.2
```

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

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 3.6.2
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.2
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(relaimpo)
```

```
## Warning: package 'relaimpo' was built under R version 3.6.2
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:plotly':
##
##     select
```

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

```
## Loading required package: boot

## Warning: package 'boot' was built under R version 3.6.2

##
## Attaching package: 'boot'

## The following object is masked from 'package:psych':
##
##     logit

## The following object is masked from 'package:survival':
##
##     aml

## The following object is masked from 'package:lattice':
##
##     melanoma

## Loading required package: survey

## Warning: package 'survey' was built under R version 3.6.2

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

##
## Attaching package: 'survey'

## The following object is masked from 'package:Hmisc':
##
##     deff

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loading required package: mitools

## Warning: package 'mitools' was built under R version 3.6.2

## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric
pmvd is available
```

```
## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.

library(e1071)

## Warning: package 'e1071' was built under R version 3.6.2

##
## Attaching package: 'e1071'

## The following object is masked from 'package:Hmisc':
##
##     impute
```

Loading Dataset and getting overall feel of data

```
AirbnbIstanbul<-read.csv("C:/Alok/OneDrive/Rutgers_MITA/Semester2/MVA/R/AirbnbIstanbul.csv",stringsAsFactors = FALSE)
Istanbul <- copy(AirbnbIstanbul)
View(Istanbul)
str(Istanbul)

## 'data.frame':    16251 obs. of  16 variables:
##  $ id                            : int  4826 20815 25436 27271 28277 28308 28318 29241 30697 33368 ...
##  $ name                          : chr  "The Place" "The Bosphorus from The Comfy Hill" "House for vacation rental furnutare" "LOVELY APT. IN PERFECT LOCATION" ...
##  $ host_id                       : int  6603 78838 105823 117026 121607 121695 121721 125742 132137 135136 ...
##  $ host_name                     : chr  "Kaan" "GÃ¼lder" "Yesim" "Mutlu" ...
##  $ neighbourhood_group           : logi  NA NA NA NA NA NA ...
##  $ neighbourhood                 : chr  "Uskudar" "Besiktas" "Besiktas" "Beyoglu" ...
##  $ latitude                      : num  41.1 41.1 41.1 41 41 ...
##  $ longitude                     : num  29.1 29 29 29 29 ...
##  $ room_type                     : chr  "Entire home/apt" "Entire home/apt" "Entire home/apt" "Entire home/apt" ...
##  $ price                         : int  554 100 211 237 591 237 633 264 596 295 ...
##  $ minimum_nights                : int  1 30 21 5 3 1 3 3 1 2 ...
##  $ number_of_reviews             : int  1 41 0 2 0 0 0 0 1 1 ...
##  $ last_review                   : chr  "6/1/2009" "11/7/2018" "" "5/4/2018" ...
##  $ reviews_per_month             : num  0.01 0.38 NA 0.04 NA NA NA NA 0.01 0.02 ...
##  $ calculated_host_listings_count: int  1 2 1 1 13 1 1 1 1 2 ...
##  $ availability_365              : int  365 49 83 228 356 365 365 365 365 232 ...

#Checking number f rows and columns
dim(Istanbul)
```

```
## [1] 16251    16

class(Istanbul)

## [1] "data.frame"

names(Istanbul)

##  [1] "id"                                "name"
##  [3] "host_id"                           "host_name"
##  [5] "neighbourhood_group"               "neighbourhood"
##  [7] "latitude"                          "longitude"
##  [9] "room_type"                         "price"
## [11] "minimum_nights"                    "number_of_reviews"
## [13] "last_review"                       "reviews_per_month"
## [15] "calculated_host_listings_count" "availability_365"

attach(Istanbul)
#head(Istanbul,25)
#Removing neighbourhood_group and last_review

#Creating new data table with all the quantitative column named Istanbul_num
Istanbul_num2 <- Istanbul[,c("latitude","longitude","price","minimum_nights",
"number_of_reviews","calculated_host_listings_count","availability_365")]
```

Correlation and applying PCA on our data

```
#Correlation
cor(Istanbul_num2)

##                                    latitude    longitude       price
## latitude                        1.000000000 -0.184363006  0.03253566
## longitude                      -0.184363006  1.000000000 -0.02208862
## price                           0.032535661 -0.022088620  1.00000000
## minimum_nights                  0.006076484 -0.006376785  0.01658485
## number_of_reviews              -0.025142526 -0.001883323 -0.01926175
## calculated_host_listings_count  0.001482799 -0.033866953  0.03009963
## availability_365               -0.001116257 -0.034483205  0.04701523
##                                minimum_nights number_of_reviews
## latitude                          0.006076484      -0.025142526
## longitude                        -0.006376785      -0.001883323
## price                             0.016584846      -0.019261753
## minimum_nights                    1.000000000      -0.015149114
## number_of_reviews                -0.015149114       1.000000000
## calculated_host_listings_count   -0.020916303       0.174662879
## availability_365                  0.015297298       0.043230154
##                                calculated_host_listings_count availability
_365
## latitude                                          0.001482799     -0.00111
6257
## longitude                                        -0.033866953     -0.03448
3205
```

```
## price                                            0.030099631      0.04701
5234
## minimum_nights                                  -0.020916303      0.01529
7298
## number_of_reviews                                0.174662879      0.04323
0154
## calculated_host_listings_count                   1.000000000      0.17306
8073
## availability_365                                 0.173068073      1.00000
0000
```

```
#Very little correlation between 'Number of reviews and  calculated host list
ing' & 'calcHostlisting and availability365'
#PCA
#Applying PCA on numeric data as it's not much recommended for categorical da
ta
Istanbul_ip_pca <- prcomp(Istanbul_num2,scale=TRUE)
Istanbul_ip_pca
```

```
## Standard deviations (1, .., p=7):
## [1] 1.1329266 1.0916193 1.0156849 0.9917434 0.9620390 0.9009618 0.8788757
##
## Rotation (n x k) = (7 x 7):
##                                         PC1         PC2         PC3
PC4
## latitude                        -0.13671624  0.67300964 -0.17720115 -0.0385
79150
## longitude                        0.24384922 -0.62742158  0.19983543  0.0991
97960
## price                           -0.15664516  0.18106565  0.61612379  0.5492
57096
## minimum_nights                   0.02663696  0.09422947  0.58495893 -0.7914
22636
## number_of_reviews               -0.45778510 -0.27942467 -0.32052131 -0.2365
68651
## calculated_host_listings_count -0.64683557 -0.17371645 -0.04188819 -0.0071
09716
## availability_365                -0.51830063 -0.05899714  0.31998679  0.0679
92594
##                                         PC5         PC6         PC7
## latitude                         0.068157824  0.69490244 -0.08912466
## longitude                        0.047268165  0.70233520 -0.03995339
## price                            0.504699877 -0.07281020 -0.03691007
## minimum_nights                   0.129066213  0.01560196  0.07059059
## number_of_reviews                0.567198941 -0.03157838 -0.48056653
## calculated_host_listings_count -0.008167183  0.12443764  0.73080076
## availability_365                -0.632422989  0.04253481 -0.46808248
```

```
#PC1--> Dominated by negative effect of calculated_host_listings_count and av
ailability_365 and no of reviews
```

```r
#PC2--> major +ve effect of latitude and negative effect of longitude
#PC3 --> Major +ve effect of minimum_nights and Price
#PC4 -->  Major negative effect of minimum_nights and +ve effect of price
#PC5 -->  Major negative effect of availability_365 and +ve effect of number_
of_reviews

#Summary pf PCAs
summary(Istanbul_ip_pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation     1.1329 1.0916 1.0157 0.9917 0.9620 0.9010 0.8789
## Proportion of Variance 0.1834 0.1702 0.1474 0.1405 0.1322 0.1160 0.1104
## Cumulative Proportion  0.1834 0.3536 0.5010 0.6415 0.7737 0.8897 1.0000
```

```r
#As per Summary output, 'Cumulative Proportion' field, 88.97% of Cummulative
variance is explained by PC1, PC2,----PC6
#So we will have to include PC1 till PC6 to prevent loss of Information.
```

Insights from Above PCA Output

Contents of Principal Components:

PC1–> Dominated by negative effect of calculated_host_listings_count and availability_365 and no of reviews

PC2–> major +ve effect of latitude and negative effect of longitude

PC3 –> Major +ve effect of minimum_nights and Price

PC4 –> Major negative effect of minimum_nights and +ve effect of price

PC5 –> Major negative effect of availability_365 and +ve effect of number_of_reviews

From Summary of Pincipal components,

Proportion of Variance, does not explain much of variance individually.

'Cumulative Proportion' field, 88.97% of Cummulative variance is explained by PC1, PC2,—-PC6

So we will have to include PC1 till PC6 to prevent loss of Information.

Plotting PCs

```
# A table containing eigenvalues and %'s accounted, follows
# Eigenvalues are sdev^2

(eigen_Istanbul <- Istanbul_ip_pca$sdev^2)

## [1] 1.2835228 1.1916328 1.0316158 0.9835550 0.9255190 0.8117322 0.7724225

names(eigen_Istanbul) <- paste("PC",1:7,sep="")
eigen_Istanbul

##       PC1       PC2       PC3       PC4       PC5       PC6       PC7
## 1.2835228 1.1916328 1.0316158 0.9835550 0.9255190 0.8117322 0.7724225

names(eigen_Istanbul)

## [1] "PC1" "PC2" "PC3" "PC4" "PC5" "PC6" "PC7"

#Taking Sum of all Eigen values
sumlambdas1 <- sum(eigen_Istanbul)
sumlambdas1 #sum of Eigenvalues is total var of ur dataset

## [1] 7
```

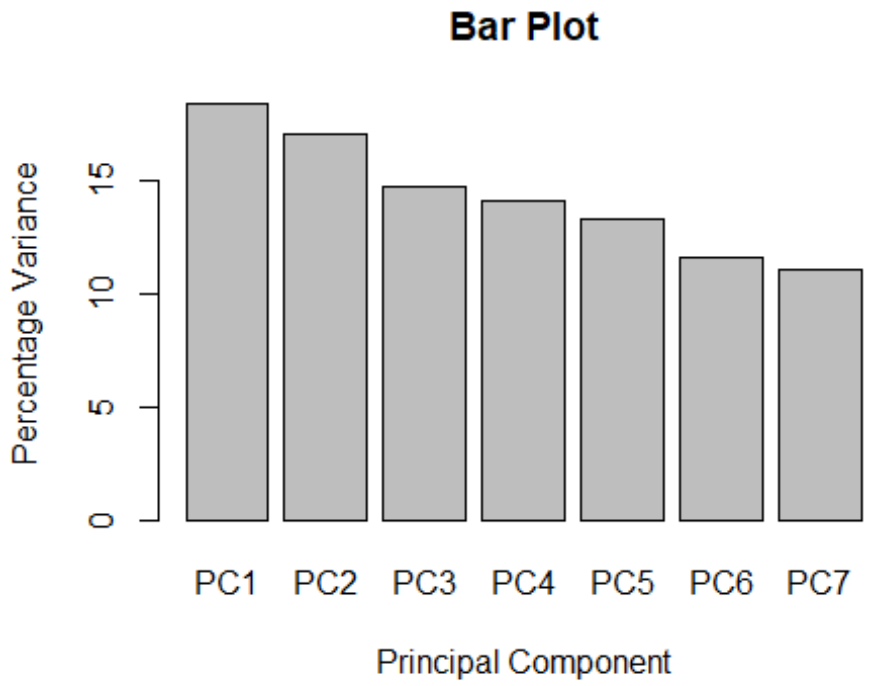
```
propvar1 <- eigen_Istanbul/sumlambdas1
propvar1 #Propvar1 gives the percentage of variance for each PC component
```

```
##       PC1       PC2       PC3       PC4       PC5       PC6       PC7
## 0.1833604 0.1702333 0.1473737 0.1405079 0.1322170 0.1159617 0.1103461
```

```
#Percentage of total variance
percentvar <- (eigen_Istanbul/sumlambdas1) *100
percentvar
```

```
##      PC1      PC2      PC3      PC4      PC5      PC6      PC7
## 18.33604 17.02333 14.73737 14.05079 13.22170 11.59617 11.03461
```

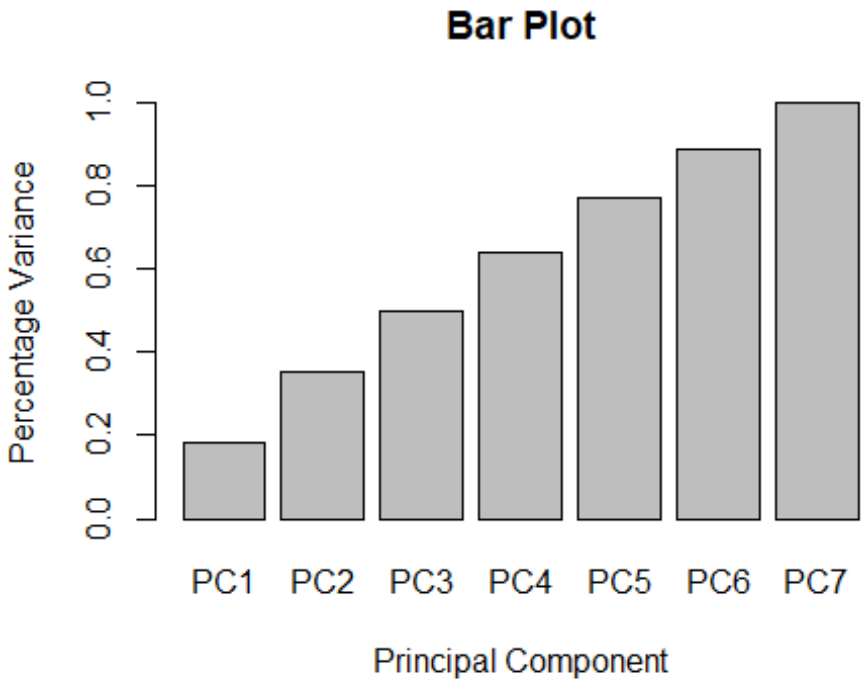
```
#Bar plot of Percentage variance
barplot(percentvar, main = "Bar Plot", xlab = "Principal Component", ylab = "
Percentage Variance")
```




```
#[1] 0.1833604 0.1702333 0.1473737 0.1405079 0.1322170 0.1159617 0.1103461
#OP says none of the component explains much variance so we will have to imcl
ude all
#Can do plotting at this stage
cumvar_Istanbul <- cumsum(propvar1)
cumvar_Istanbul #This variable has cummulative sum of variance
```

```
##       PC1       PC2       PC3       PC4       PC5       PC6       PC7
## 0.1833604 0.3535937 0.5009673 0.6414752 0.7736922 0.8896539 1.0000000
```

```r
#PC1 to PC6 explains 88.96% of variance

#Bar plot of Cummulative Percentage variance
barplot(cumvar_Istanbul, main = "Bar Plot", xlab = "Principal Component", yla
b = "Percentage Variance")
```




```r
matlambdas <- rbind(eigen_Istanbul,propvar1,cumvar_Istanbul)
rownames(matlambdas) <- c("Eigenvalues","Prop. variance","Cum. prop. variance
")
round(matlambdas,4)
```

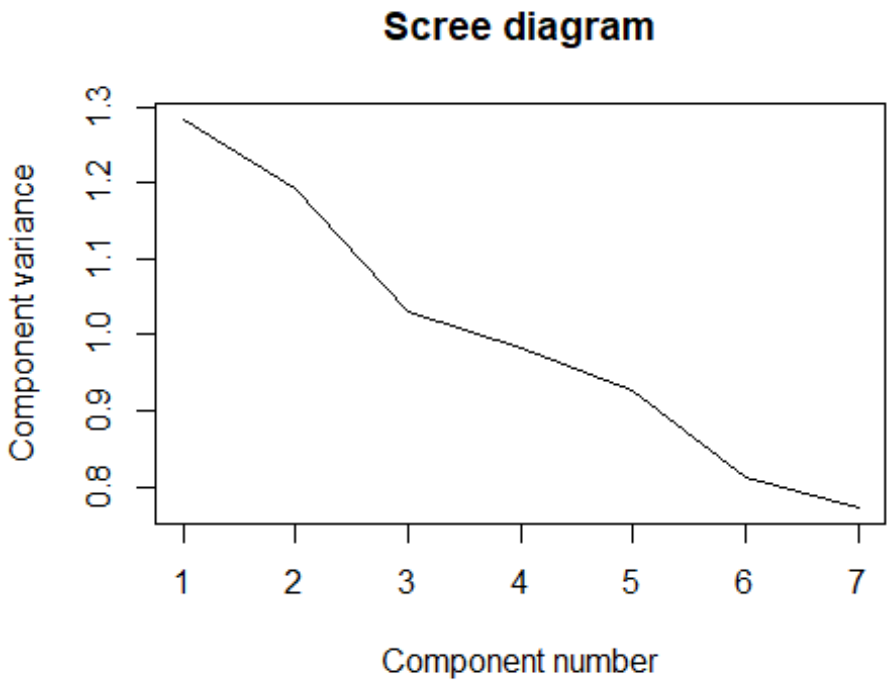
```
##                          PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Eigenvalues         1.2835 1.1916 1.0316 0.9836 0.9255 0.8117 0.7724
## Prop. variance      0.1834 0.1702 0.1474 0.1405 0.1322 0.1160 0.1103
## Cum. prop. variance 0.1834 0.3536 0.5010 0.6415 0.7737 0.8897 1.0000
```

```r
summary(Istanbul_ip_pca)
```

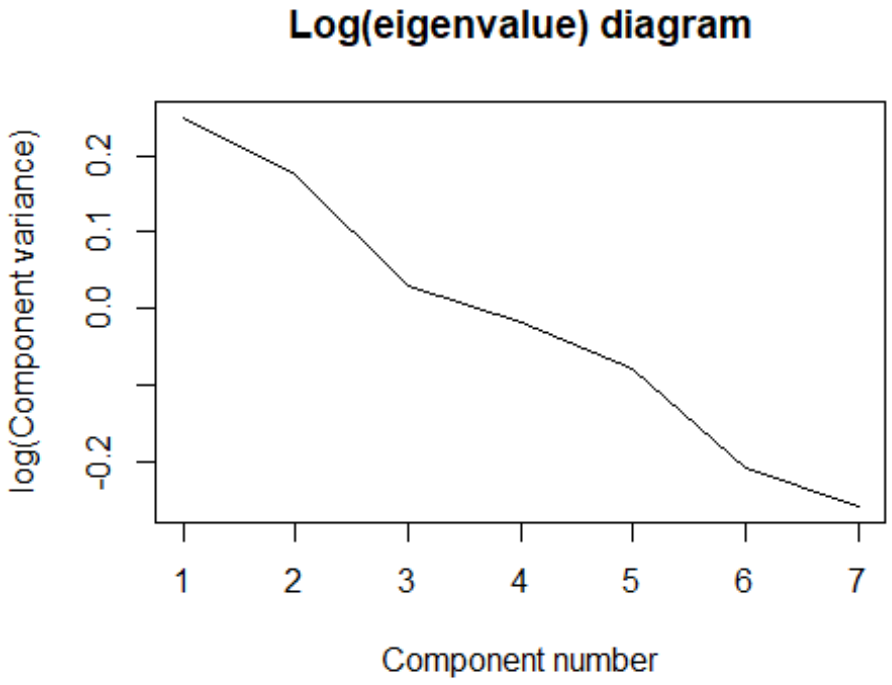
```
## Importance of components:
##                            PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation     1.1329 1.0916 1.0157 0.9917 0.9620 0.9010 0.8789
## Proportion of Variance 0.1834 0.1702 0.1474 0.1405 0.1322 0.1160 0.1104
## Cumulative Proportion  0.1834 0.3536 0.5010 0.6415 0.7737 0.8897 1.0000
```

```r
# PCA Rotation
Istanbul_ip_pca$rotation #= print(Istanbul_ip_pca)
```

```
##                                        PC1         PC2         PC3
PC4
## latitude                          -0.13671624  0.67300964 -0.17720115 -0.0385
79150
## longitude                          0.24384922 -0.62742158  0.19983543  0.0991
97960
## price                             -0.15664516  0.18106565  0.61612379  0.5492
57096
## minimum_nights                     0.02663696  0.09422947  0.58495893 -0.7914
22636
## number_of_reviews                 -0.45778510 -0.27942467 -0.32052131 -0.2365
68651
## calculated_host_listings_count -0.64683557 -0.17371645 -0.04188819 -0.0071
09716
## availability_365                  -0.51830063 -0.05899714  0.31998679  0.0679
92594
##                                           PC5         PC6         PC7
## latitude                           0.068157824  0.69490244 -0.08912466
## longitude                          0.047268165  0.70233520 -0.03995339
## price                              0.504699877 -0.07281020 -0.03691007
## minimum_nights                     0.129066213  0.01560196  0.07059059
## number_of_reviews                  0.567198941 -0.03157838 -0.48056653
## calculated_host_listings_count -0.008167183  0.12443764  0.73080076
## availability_365                  -0.632422989  0.04253481 -0.46808248

#Scree Plot
plot(eigen_Istanbul, xlab = "Component number", ylab = "Component variance",
type = "l", main = "Scree diagram")
```




```r
plot(log(eigen_Istanbul), xlab = "Component number",ylab = "log(Component var
iance)", type="l",main = "Log(eigenvalue) diagram")
```



```
#Conclusion : As per Summary output, 'Cumulative Proportion' field, 88.97% of
Cummulative variance is explained by PC1, PC2,----PC6
#So we will have to include PC1 till PC6 to prevent loss of Information.
```

From Summary of Pincipal components,

Proportion of Variance, does not explain much of variance individually.

'Cumulative Proportion' field, 88.97% of Cummulative variance is explained by PC1, PC2,—-PC6

So we will have to include PC1 till PC6 to prevent loss of Information.

Binding Principal Components and categorical columns together

Printing our new Dataset after PCA

```
Istanbultyp_pca <- cbind(data.frame(neighbourhood,room_type),Istanbul_ip_pca$
x)
names(Istanbultyp_pca)
```

```
## [1] "neighbourhood" "room_type"     "PC1"           "PC2"
## [5] "PC3"           "PC4"           "PC5"           "PC6"
## [9] "PC7"
```

```
#Istanbultyp_pca This is our new dataset
head(Istanbultyp_pca,5)
```

```
##   neighbourhood       room_type        PC1        PC2        PC3         P
C4
## 1       Uskudar Entire home/apt -0.0198208 0.20770197  0.3882550  0.338905
61
## 2      Besiktas Entire home/apt  0.2611837 0.09049684 -0.6360656 -1.248468
31
## 3      Besiktas Entire home/apt  1.0337257 0.76983155 -0.1140098 -0.501509
68
## 4       Beyoglu Entire home/apt  0.4432095 0.24065558 -0.0297182 -0.012200
13
## 5         Sisli Entire home/apt -1.0889860 0.15267916  0.3022794  0.245316
84
##           PC5        PC6        PC7
## 1 -0.56734954 0.88516538 -0.6549925
## 2  1.94434798 0.94668791 -0.3117832
## 3  0.71092336 1.07091534  0.3568210
## 4 -0.06654501 0.02912375 -0.1133420
## 5 -0.58952855 0.47097856  0.5969920
```

```
#Renaming Principal components
names(Istanbultyp_pca) <- c("Neighbourhood", "Room_Type", "calc_Review_365_Ne
gative", "Lattitude_Positive_Longi_Negate",
                            "MinNight_Price_Positive","MinNightNegative_Price
```

```
Pos","availabilityNegate_Reviews_Pos",
                            "Positive_Lat_Long","CalcHostListing_Pos")

#This is Our new dataset
names(Istanbultyp_pca)

## [1] "Neighbourhood"                   "Room_Type"
## [3] "calc_Review_365_Negative"        "Lattitude_Positive_Longi_Negate"
## [5] "MinNight_Price_Positive"         "MinNightNegative_PricePos"
## [7] "availabilityNegate_Reviews_Pos"  "Positive_Lat_Long"
## [9] "CalcHostListing_Pos"

#View(Istanbultyp_pca)
dim(Istanbultyp_pca)

## [1] 16251     9

head(Istanbultyp_pca,5)

##   Neighbourhood       Room_Type calc_Review_365_Negative
## 1       Uskudar Entire home/apt               -0.0198208
## 2      Besiktas Entire home/apt                0.2611837
## 3      Besiktas Entire home/apt                1.0337257
## 4       Beyoglu Entire home/apt                0.4432095
## 5         Sisli Entire home/apt               -1.0889860
##   Lattitude_Positive_Longi_Negate MinNight_Price_Positive
## 1                      0.20770197               0.3882550
## 2                      0.09049684              -0.6360656
## 3                      0.76983155              -0.1140098
## 4                      0.24065558              -0.0297182
## 5                      0.15267916               0.3022794
##   MinNightNegative_PricePos availabilityNegate_Reviews_Pos Positive_Lat_Lo
ng
## 1                0.33890561                     -0.56734954        0.885165
38
## 2               -1.24846831                      1.94434798        0.946687
91
## 3               -0.50150968                      0.71092336        1.070915
34
## 4               -0.01220013                     -0.06654501        0.029123
75
## 5                0.24531684                     -0.58952855        0.470978
56
##   CalcHostListing_Pos
## 1          -0.6549925
## 2          -0.3117832
## 3           0.3568210
## 4          -0.1133420
## 5           0.5969920
```

Finding Means and standard deviations by Room Types

```
# Means of scores for all the PC's classified by Survival status so da u can
perform ttest on that
tabmeansPC1 <- aggregate(Istanbultyp_pca[,c(3,4,5,6,7,8,9)],by=list(room_type
=Istanbul$room_type),mean)
tabmeansPC1 #Means of all the columns per Room Type
```

```
##         room_type calc_Review_365_Negative Lattitude_Positive_Longi_Negate
## 1 Entire home/apt               -0.1756988                      0.02117150
## 2    Private room                0.1297107                     -0.02394794
## 3     Shared room                0.3080373                      0.10680772
##   MinNight_Price_Positive MinNightNegative_PricePos
## 1              -0.01821017              -0.072516130
## 2               0.01739272               0.060570262
## 3              -0.03640266               0.005412508
##   availabilityNegate_Reviews_Pos Positive_Lat_Long CalcHostListing_Pos
## 1                      0.1342928       -0.01136218         -0.02408167
## 2                     -0.1050311        0.01863751          0.01675934
## 3                     -0.1335527       -0.15742400          0.05985366
```

```
#In this op coz of +ve -VE signs u can see the means are different
tabmeansPC1 <- tabmeansPC1[rev(order(tabmeansPC1$room_type)),]
tabmeansPC1
```

```
##         room_type calc_Review_365_Negative Lattitude_Positive_Longi_Negate
## 3     Shared room                0.3080373                      0.10680772
## 2    Private room                0.1297107                     -0.02394794
## 1 Entire home/apt               -0.1756988                      0.02117150
##   MinNight_Price_Positive MinNightNegative_PricePos
## 3              -0.03640266               0.005412508
## 2               0.01739272               0.060570262
## 1              -0.01821017              -0.072516130
##   availabilityNegate_Reviews_Pos Positive_Lat_Long CalcHostListing_Pos
## 3                     -0.1335527       -0.15742400          0.05985366
## 2                     -0.1050311        0.01863751          0.01675934
## 1                      0.1342928       -0.01136218         -0.02408167
```

```
tabfmeans1 <- t(tabmeansPC1[,-1]) #transpose
tabfmeans1
```

```
##                                          3           2           1
## calc_Review_365_Negative         0.308037254  0.12971065 -0.17569882
## Lattitude_Positive_Longi_Negate  0.106807721 -0.02394794  0.02117150
## MinNight_Price_Positive         -0.036402663  0.01739272 -0.01821017
## MinNightNegative_PricePos        0.005412508  0.06057026 -0.07251613
## availabilityNegate_Reviews_Pos  -0.133552683 -0.10503106  0.13429281
## Positive_Lat_Long               -0.157424001  0.01863751 -0.01136218
## CalcHostListing_Pos              0.059853663  0.01675934 -0.02408167
```

```
colnames(tabfmeans1) <- t(as.vector(tabmeansPC1[1]))
tabfmeans1 #This is means for all PCs per room Type
```

```
##                                  Shared room Private room Entire home/apt
## calc_Review_365_Negative          0.308037254   0.12971065     -0.17569882
## Lattitude_Positive_Longi_Negate   0.106807721  -0.02394794      0.02117150
## MinNight_Price_Positive          -0.036402663   0.01739272     -0.01821017
## MinNightNegative_PricePos         0.005412508   0.06057026     -0.07251613
## availabilityNegate_Reviews_Pos   -0.133552683  -0.10503106      0.13429281
## Positive_Lat_Long                -0.157424001   0.01863751     -0.01136218
## CalcHostListing_Pos               0.059853663   0.01675934     -0.02408167
```

```
# Standard deviations of scores for all the PC's classified by Room Type
tabsdsPC1 <- aggregate(Istanbultyp_pca[,c(3,4,5,6,7,8,9)],by=list(room_type=I
stanbul$room_type),sd)
tabfsds1 <- t(tabsdsPC1[,-1])
colnames(tabfsds1) <- t(as.vector(tabsdsPC1[1]))
tabfsds1 #This is Std Deviation for all PCs per room Type
```

```
##                                 Entire home/apt Private room Shared room
## calc_Review_365_Negative              1.1362586    1.1226645   0.8624386
## Lattitude_Positive_Longi_Negate       1.1418609    1.0397350   1.2051611
## MinNight_Price_Positive               0.9987636    1.0233101   1.1184727
## MinNightNegative_PricePos             1.0239683    0.9748299   0.6917807
## availabilityNegate_Reviews_Pos        0.9581406    0.9562878   0.8705255
## Positive_Lat_Long                     0.9391580    0.8631261   0.9572905
## CalcHostListing_Pos                   0.9043482    0.8717852   0.5592181
```

```
class(tabfsds1) #changed to matrix
```

```
## [1] "matrix"
```

Levens Tests

```
# Levene's tests (one-sided)
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## The following object is masked from 'package:purrr':
##
##     some

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

library(carData)
names(Istanbultyp_pca)

## [1] "Neighbourhood"                   "Room_Type"
## [3] "calc_Review_365_Negative"        "Lattitude_Positive_Longi_Negate"
## [5] "MinNight_Price_Positive"         "MinNightNegative_PricePos"
## [7] "availabilityNegate_Reviews_Pos"  "Positive_Lat_Long"
## [9] "CalcHostListing_Pos"

(LTPC1 <- leveneTest(calc_Review_365_Negative~Istanbul$room_type,data=Istanbu
ltyp_pca))

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     2   32.39 9.141e-15 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(p_PC1_1sided1 <- LTPC1[[3]][1]/2)

## [1] 4.570698e-15

(LTPC2 <- leveneTest(Lattitude_Positive_Longi_Negate~Istanbul$room_type,data=
Istanbultyp_pca))

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value  Pr(>F)
## group     2  15.797 1.4e-07 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(p_PC2_1sided=LTPC2[[3]][1]/2)

## [1] 6.999728e-08

(LTPC3 <- leveneTest(MinNight_Price_Positive~Istanbul$room_type,data=Istanbul
typ_pca))
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     2  15.166 2.628e-07 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(p_PC3_1sided <- LTPC3[[3]][1]/2)

## [1] 1.313987e-07

(LTPC4 <- leveneTest(MinNightNegative_PricePos~Istanbul$room_type,data=Istanb
ultyp_pca))

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value   Pr(>F)
## group     2   26.52 3.17e-12 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(p_PC4_1sided <- LTPC4[[3]][1]/2)

## [1] 1.585111e-12

(LTPC5 <- leveneTest(availabilityNegate_Reviews_Pos~Istanbul$room_type,data=I
stanbultyp_pca))

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     2  18.916 6.229e-09 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(p_PC5_1sided <- LTPC5[[3]][1]/2)

## [1] 3.114674e-09

(LTPC6 <- leveneTest(Positive_Lat_Long~Istanbul$room_type,data=Istanbultyp_pc
a))

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value   Pr(>F)
## group     2  9.7007 6.16e-05 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(p_PC6_1sided <- LTPC6[[3]][1]/2)
```

```
## [1] 3.079758e-05
```

```
(LTPC7 <- leveneTest(CalcHostListing_Pos~Istanbul$room_type,data=Istanbultyp_
pca))
```

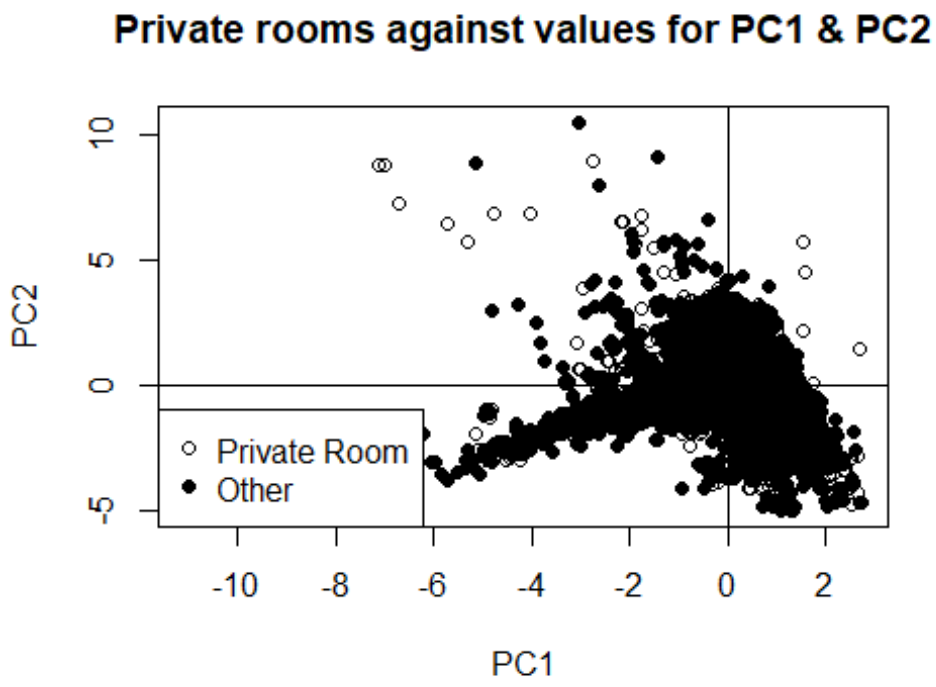
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     2  19.239 4.515e-09 ***
##       16248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(p_PC7_1sided <- LTPC7[[3]][1]/2)
```

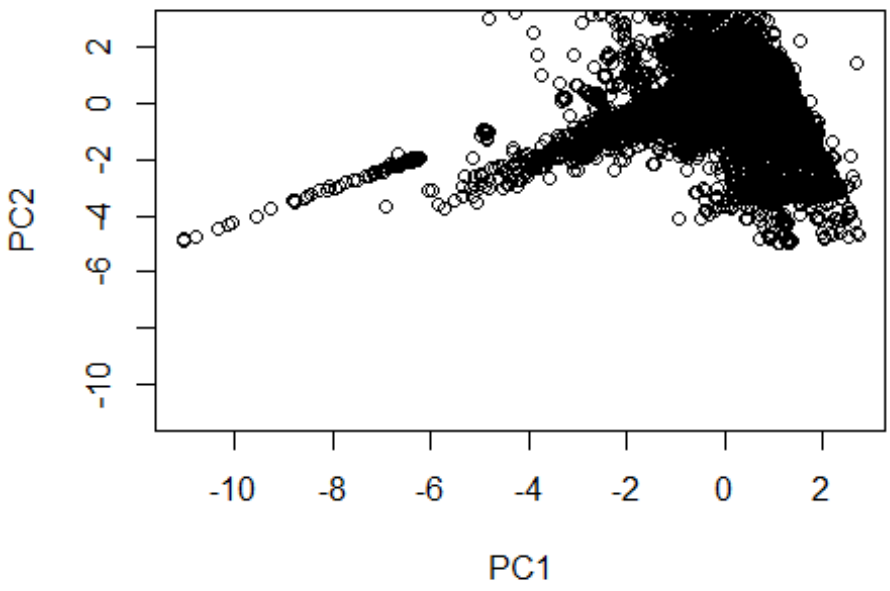
```
## [1] 2.257301e-09
```

Plotting the scores for the first and second components for Private Rooms

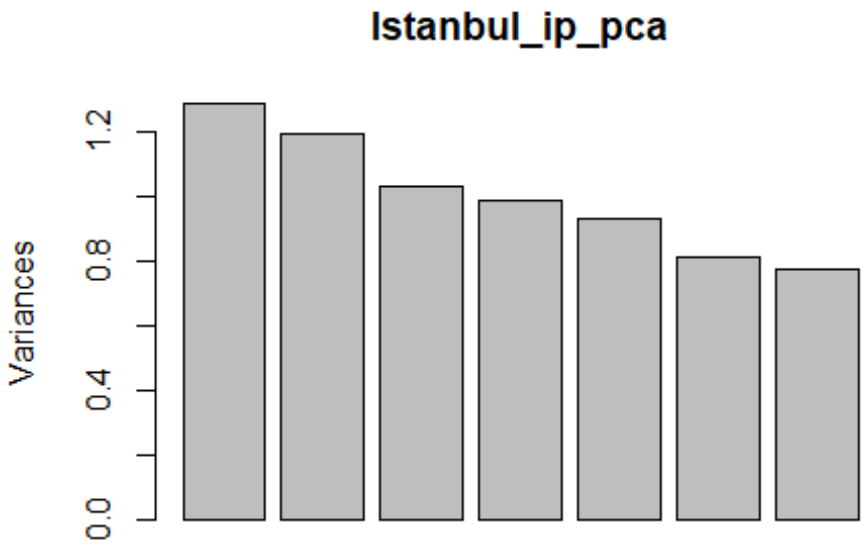
```
# Plotting the scores for the first and second components for Private Rooms

plot(Istanbultyp_pca$calc_Review_365_Negative, Istanbultyp_pca$Lattitude_Posi
tive_Longi_Negate,
     pch=ifelse(Istanbultyp_pca$Room_Type == "Private room",1,16),xlab="PC1",
ylab="PC2", main="Private rooms against values for PC1 & PC2")
abline(h=0)
abline(v=0)

legend("bottomleft", legend=c("Private Room","Other"), pch=c(1,16))
```




```
#View(Istanbul_ip_pca)
diag(cov(Istanbul_ip_pca$x))

##       PC1       PC2       PC3       PC4       PC5       PC6       PC7
## 1.2835228 1.1916328 1.0316158 0.9835550 0.9255190 0.8117322 0.7724225

xlim <- range(Istanbul_ip_pca$x[,1])
#xlim
#Istanbul_ip_pca$x[,1]
#Istanbul_ip_pca$x
plot(Istanbul_ip_pca$x,xlim=xlim,ylim=xlim)
```




```
#Istanbul_ip_pca$rotation[,1]
#Istanbul_ip_pca$rotation
#plot(Istanbul[,-1])
#Plotting Variances
plot(Istanbul_ip_pca)
```



To get the original value of the data based on PCA

```
#get the original value of the data based on PCA
center <- Istanbul_ip_pca$center
scale <- Istanbul_ip_pca$scale
new_Istanbul <- as.matrix(Istanbul[,-1])
head(new_Istanbul,5)
```

```
##      name                                    host_id     host_name
## [1,] "The Place"                             "     6603" "Kaan"
## [2,] "The Bosphorus from The Comfy Hill"     "    78838" "GÃ¼lder"
## [3,] "House for vacation rental furnutare"   "   105823" "Yesim"
## [4,] "LOVELY APT. IN PERFECT LOCATION"       "   117026" "Mutlu"
## [5,] "Duplex Apartment with Terrace "        "   121607" "Alen"
##      neighbourhood_group neighbourhood latitude   longitude  room_type
## [1,] NA                  "Uskudar"     "41.05650" "29.05367" "Entire home/
apt"
## [2,] NA                  "Besiktas"    "41.06984" "29.04545" "Entire home/
apt"
## [3,] NA                  "Besiktas"    "41.07731" "29.03891" "Entire home/
apt"
## [4,] NA                  "Beyoglu"     "41.03220" "28.98216" "Entire home/
apt"
## [5,] NA                  "Sisli"       "41.04471" "28.98567" "Entire home/
apt"
##      price   minimum_nights number_of_reviews last_review reviews_per_mont
h
```

```
## [1,] "  554" "   1"         "  1"             "6/1/2009"  " 0.01"
## [2,] "  100" "  30"         " 41"             "11/7/2018" " 0.38"
## [3,] "  211" "  21"         "  0"             ""          NA
## [4,] "  237" "   5"         "  2"             "5/4/2018"  " 0.04"
## [5,] "  591" "   3"         "  0"             ""          NA
##      calculated_host_listings_count availability_365
## [1,] " 1"                           "365"
## [2,] " 2"                           " 49"
## [3,] " 1"                           " 83"
## [4,] " 1"                           "228"
## [5,] "13"                           "356"
```

Predict

```
predict(Istanbul_ip_pca)[,1]
```

```
##     [1] -1.982080e-02  2.611837e-01  1.033726e+00  4.432095e-01 -1.088986e
+00
##     [6] -3.392863e-02 -1.087338e-01 -1.549697e-01 -1.166976e-01  3.196913e
-01
##    [11] -2.810846e-01 -1.220210e-01 -2.331986e-01 -5.027803e-02 -8.092017e
-01
##    [16] -7.545404e-02 -7.402963e-01  5.570317e-01 -2.858194e-01  1.008379e
+00
##    [21] -3.839433e-01 -2.057558e-01 -3.360221e-02 -3.840837e-02 -3.324720e
-01
##    [26] -4.421108e-01 -5.355561e-01  1.134305e+00 -2.323594e-01 -1.611261e
+00
##    [31] -3.276429e-02  6.685052e-01 -6.898364e-02 -9.518836e-01 -2.161913e
-01
##    [36] -1.789718e+00  1.327175e+00 -1.113483e+00 -1.128657e-01 -1.648513e
-01
##    [41] -9.395780e-02  1.237452e-01 -1.092384e+00 -3.950938e-01  2.464463e
-01
##    [46]  1.155281e+00 -2.574804e-01 -1.205699e+00  2.545080e-01 -7.782369e
-02
##    [51]  8.137489e-01  7.795472e-02 -2.967713e+00 -3.875881e-01 -1.004389e
-01
##    [56] -1.972576e-01  7.501653e-01 -2.355558e-01  4.169608e-01 -1.075607e
-01
##    [61]  1.006752e+00 -9.757401e-01 -3.250128e+00  5.891643e-01 -1.195963e
+00
##    [66] -1.557194e-01 -6.603129e-01 -1.430680e-01 -5.384394e-01  1.348814e
+00
##    [71] -7.498563e-01 -8.312277e-01 -1.135657e+00 -9.759375e-01 -1.067088e
+00
##    [76] -1.715597e+00 -2.232542e+00 -8.404877e-01  9.172502e-02 -3.306277e
-01
##    [81] -6.902405e-01 -1.193715e+00 -1.462379e-01 -6.713765e-01  1.300603e
-01
##    [86] -3.557103e-01 -7.488786e-02  3.369594e-02 -1.126006e+00 -2.377932e
```

```
+00
##    [91] -6.540497e-01 -4.368060e-01  3.464927e-01 -9.154547e-02 -8.320408e
-01
##    [96] -6.584618e-01  8.499784e-01 -1.084718e+00 -2.310425e+00 -7.230624e
-01
##   [101] -1.296381e-01  9.539511e-01  1.153902e-01 -6.239424e-02  1.088967e
+00
##   [106] -8.386576e-02 -1.708867e+00 -7.740343e-01 -2.687296e-02 -1.417552e
+00
##   [111] -7.784377e-02 -6.864227e-02 -4.141733e+00 -6.473901e-01 -2.282851e
+00
##   [116]  9.311354e-01 -7.955900e-02 -2.428303e-01 -2.839345e+00 -9.086922e
-02
##   [121] -4.219141e+00 -1.109395e-01 -2.062707e-01  8.589579e-01 -2.786242e
-01
##   [126]  2.409104e-01 -3.965124e+00 -5.998445e-02 -1.069070e+00 -3.863835e
-01
##   [131] -1.245134e+00 -1.112636e+00 -1.278711e+00 -4.173623e-01 -5.494119e
-01
##   [136] -1.570588e+00 -2.572670e-01 -2.821681e-01 -3.825581e-01 -1.037351e
-01
##   [141] -8.235494e-01 -1.497293e-02 -9.974701e-01 -1.011142e+00 -8.767852e
-02
##   [146] -1.901515e-01 -9.717040e-02 -1.150255e-01 -1.600301e-02 -1.167109e
-01
##   [151] -1.300358e-01 -1.352541e-01 -5.569849e-02  1.013302e+00 -2.116819e
-01
##   [156] -4.404628e-01 -1.015299e-01 -3.565791e-02 -2.388967e-01 -3.039721e
-02
##   [161] -2.442102e+00 -3.449361e+00  2.020249e-01 -7.870325e-01 -3.331865e
+00
##   [166] -1.704644e-01 -1.711917e-01  1.055289e+00  3.258597e-01 -1.471583e
-01
##   [171] -4.302418e+00 -1.067108e+00 -2.599282e+00 -8.883889e-02  1.078873e
-01
##   [176] -4.005199e-01 -4.304152e+00 -1.464351e+00 -4.993317e-01 -1.098409e
+00
##   [181] -1.480237e+00 -1.420427e+00 -5.425331e-01 -1.103854e-01  6.259847e
-01
##   [186]  2.167565e-01 -5.842942e-01 -2.688149e+00 -4.174007e-01 -6.300420e
-02
##   [191] -4.065615e-01 -3.576724e+00 -5.068692e-01 -1.127862e-01 -6.599756e
-01
##   [196]  9.549614e-01  6.011233e-01 -4.556286e-01 -8.016814e-01  1.347489e
+00
##   [201] -8.158979e-02 -1.241353e-01 -2.547744e-02 -3.650825e-01  3.639016e
-01
##   [206] -1.739817e-01 -9.669459e-01 -1.433957e+00 -1.859376e+00  2.704572e
-01
##   [211] -2.088320e+00 -3.377852e+00 -1.705615e+00 -6.423824e-01 -9.442331e
```

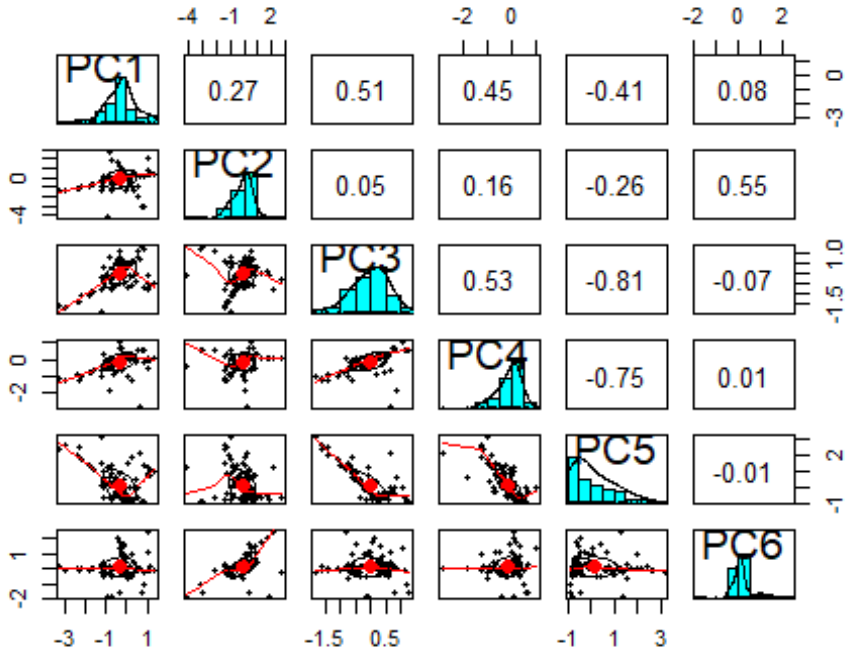
```
-01
##   [216] -1.695883e-01  1.704450e+00  1.722005e-01 -1.209673e+00  1.235584e
+00
##   [221] -3.744665e-01 -1.151915e-01 -1.567328e-01  1.250163e-01  1.321930e
+00
##   [226] -6.529419e-01 -9.019988e-01 -4.988458e+00  2.844205e-01 -6.461650e
-01
##   [231]  1.505858e-01  1.062841e+00 -1.273287e+00 -1.368076e-01 -6.926739e
-01
##   [236] -3.078794e-02 -3.879273e-01  3.679952e-01  6.217408e-01 -3.428627e
-01
##   [241] -2.479444e+00 -6.326378e-01 -5.851018e-01 -5.946826e-01 -6.816518e
-01
##   [246] -3.392980e+00  5.329706e-01  7.083182e-01  1.232485e-01 -4.705729e
-02
##   [251]  4.365217e-01  7.388403e-01  1.737495e-01 -9.778856e-01 -1.057725e
-01
##   [256]  1.296131e-01 -2.230399e-02 -2.645520e+00 -2.555077e-01  1.079501e
-02
##   [261]  8.932809e-02  1.252459e+00 -3.167480e-01 -3.252802e+00 -6.371606e
-01
##   [266] -4.895416e-01 -1.102454e+00 -4.941450e-01 -9.253630e-02 -1.161590e
+00
##   [271] -5.195123e-02 -4.753659e-02 -3.475402e+00 -1.693226e-01 -8.808210e
-02
##   [276] -3.390209e+00 -1.864724e+00  3.035989e-01 -2.653189e+00 -5.548788e
-01
##   [281] -2.570770e-01 -1.054588e-01 -3.676471e-01 -3.265438e-02 -5.033397e
-02
##   [286] -2.068026e-01 -2.734400e+00  3.666016e-02  1.003833e+00  3.696921e
-01
##   [291] -3.328837e-02  1.105063e+00 -1.774027e+00  8.655015e-01  1.237508e
+00
##   [296] -6.484451e-01 -1.212418e+00 -9.981482e-01  1.321613e-01 -9.706043e
-01
##   [301] -9.527078e-01 -4.711859e-01 -3.654901e+00 -1.417470e-01 -3.194388e
+00
##   [306] -3.286248e+00  1.864465e-01 -9.818818e-01 -2.832022e-02 -2.302121e
-01
##   [311] -3.197792e+00 -3.802189e+00 -2.982941e-01 -1.005432e-01 -6.894561e
-01
##   [316]  9.268751e-01 -1.241733e+00  1.388262e+00 -1.180826e+00 -1.237381e
-01
##   [321] -3.921369e-01 -1.077849e+00 -5.909146e-02 -2.619316e-01 -2.625444e
-01
##   [326] -2.921921e-01 -5.081994e-01 -6.784224e-01 -2.223503e-01 -8.669702e
-01
##   [331] -5.860475e-01 -4.250465e+00 -2.536849e+00 -1.965285e+00 -5.142430e
+00
##   [336] -6.029848e-01 -1.838733e-01  1.959318e+00 -7.901505e-03 -7.072318e
```

```
-01
##   [341]  6.204968e-01  3.638214e-01 -1.236216e-01 -3.464861e-01 -8.105086e
-03
##   [346] -7.684888e-01 -1.467215e-01  5.105006e-01 -5.988965e+00 -9.026224e
-02
##   [351] -4.104784e-02 -2.022596e-01 -1.830572e-01 -7.652918e-02 -4.949180e
-01
##   [356]  1.323803e+00 -8.430715e-02 -4.168333e-01 -5.186920e-01 -3.985782e
-03
##   [361] -3.244058e-01  2.008787e-01 -2.015891e-01  1.158165e+00  1.436465e
+00
##   [366] -1.970279e-01 -5.286266e-01 -2.539343e-01 -1.324746e+00 -3.056241e
-01
##   [371] -8.340517e-01 -3.065072e+00 -1.583484e+00 -6.829869e-02 -5.910895e
-01
##   [376] -1.034744e+00 -1.569323e+00  2.408498e-01 -5.123845e-01  3.142843e
-02
##   [381] -5.445187e-01 -5.486889e-01 -5.454307e-01  1.108570e-01 -1.719849e
+00
##   [386] -1.311266e-01 -1.912987e-01  8.749633e-01 -4.131435e-01 -6.605668e
-02
##   [391] -2.897227e+00 -6.278987e-01 -2.338910e+00  1.041685e-01 -1.033187e
-02
##   [396]  1.320492e+00 -2.096394e-01 -2.803434e-01  1.486372e-01 -1.748606e
-01
##   [401] -4.251138e+00 -2.556311e-01 -1.745456e-01 -1.693310e-01 -2.765124e
+00
##   [406] -7.285400e-01 -1.436064e+00  1.686599e-01  1.069908e-02 -1.142294e
-01
##   [411] -9.751302e-01 -4.209342e-01 -2.720712e-01 -5.879651e-02 -7.073787e
-02
##   [416] -2.630785e-01 -1.646618e-01 -4.060008e-02 -5.097411e-02 -2.915763e
-01
##   [421] -1.598913e+00 -1.111920e+00 -2.530662e-01 -3.873790e-02 -7.204242e
-01
##   [426] -2.412394e+00  3.510422e-01 -1.587573e+00  5.638859e-01 -4.564205e
-01
##   [431] -8.057868e-02  2.967512e-01  1.135373e+00 -8.670239e-01 -1.000450e
-01
##   [436] -2.540858e-01  3.240985e-01  9.318218e-02  6.010804e-01  2.182560e
-01
##   [441] -5.176646e-01 -1.002451e+00 -1.387413e-01 -1.228382e-01 -1.592737e
-01
##   [446] -4.616297e-01 -5.226009e-01 -5.555407e-02 -6.964795e-02 -1.412537e
+00
##   [451] -2.114266e+00 -1.580878e+00 -2.067595e+00 -3.106964e-02 -1.825734e
-01
##   [456]  1.828780e+00 -8.708678e-01 -1.124852e-01 -2.573184e+00 -2.977100e
-01
##   [461] -3.306565e+00 -1.183614e-01  4.225446e-02 -6.663781e-01 -5.965215e
```

```
-01
## [16186] -7.569162e-01  1.839086e+00  8.859075e-01  1.331249e+00 -4.882132e
-01
## [16191] -1.106459e-01  8.982957e-01 -5.219667e-01 -4.082493e-01 -2.031369e
-02
## [16196]  4.376895e-02  8.745035e-01  1.054471e+00  1.177472e+00 -1.463083e
-01
## [16201]  9.272812e-01 -1.600413e-01  9.126685e-01  8.721023e-01 -1.385621e
-01
## [16206]  7.008098e-02  1.052904e+00  8.685269e-02 -1.572937e+00 -7.178983e
-01
## [16211] -1.051766e+00 -7.875838e-02 -1.182372e-01 -1.644315e-01 -1.358004e
-01
## [16216] -2.710485e-01 -2.251084e-01 -2.388787e-01  1.031306e+00  7.217415e
-01
## [16221]  2.602975e-01 -4.444977e-01 -8.079784e-02  9.315375e-01 -6.898640e
-01
## [16226]  2.468887e-01  1.027389e+00  1.269236e+00  1.036332e+00  6.079825e
-01
## [16231] -2.324746e-01  1.203770e-01 -2.185624e-01  1.084634e+00  9.868980e
-01
## [16236]  1.195266e+00  1.100947e+00  6.673032e-01  1.079111e+00 -9.962855e
-01
## [16241] -6.105469e-01  6.412052e-01  4.413661e-01 -5.921832e-01 -4.883565e
-01
## [16246]  1.633144e+00 -3.206824e-01 -2.940352e-01  4.240007e-02 -3.057399e
-02
## [16251]  2.519838e-01
```

Pairs

```
pairs.panels(Istanbul_ip_pca$x[1:100,c(1,2,3,4,5,6)])
```



Summarizing our Principal Component Analysis for Istanbul Airbnb Data:

From Summary of Principal components:

Principal components do not explain much of variance individually.

As per, 'Cumulative Proportion' field, 88.97% of Cummulative variance is explained by PC1, PC2, PC3, PC4, PC5 and PC6.

So we will have to include PC1 till PC6 to prevent loss of Information.