Chirag_My_Project_Output.R

```r
# Hotel Room Pricing In The Indian Market
# NAME: Chirag Gupta
# EMAIL: chirag.gupta364@gmail.com
# COLLEGE : IIT (ISM) Dhanbad

# Loading the Data set
hoteldf <-read.csv(paste("Cities42.csv",sep = ""))

# General view of the entire Dataframe
View(hoteldf)
# Details of datatypes of each variable
str(hoteldf)
```

```
## 'data.frame':    13232 obs. of  20 variables:
##  $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CityName             : Factor w/ 42 levels "Agra","Ahmedabad",..: 26 26
26 26 26 26 26 26 26 26 ...
##  $ Population           : int  12442373 12442373 12442373 12442373 12442373
12442373 12442373 12442373 12442373 12442373 ...
##  $ CityRank             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ IsMetroCity          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ IsTouristDestination: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ IsWeekend            : int  1 0 1 1 0 1 0 1 1 0 ...
##  $ IsNewYearEve         : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Date                 : Factor w/ 20 levels "18-Dec-16","21-Dec-16",..:
11 12 13 14 15 16 17 18 11 12 ...
##  $ HotelName            : Factor w/ 1670 levels "14 Square Amanora",..:
1635 1635 1635 1635 1635 1635 1635 1635 1409 1409 ...
##  $ RoomRent             : int  12375 10250 9900 10350 12000 11475 11220
9225 6800 9350 ...
##  $ StarRating           : num  5 5 5 5 5 5 5 5 4 4 ...
##  $ Airport              : num  21 21 21 21 21 21 21 21 20 20 ...
##  $ HotelAddress         : Factor w/ 2108 levels " H.P. High Court Mall
Road, Shimla",..: 925 928 930 933 935 937 940 941 699 746 ...
##  $ HotelPincode         : int  400005 400006 400007 400008 400009 400010
400011 400012 400039 400040 ...
##  $ HotelDescription     : Factor w/ 1226 levels "#NAME?","10 star hotel
near Queensroad, Amritsar",..: 1030 1030 1030 1030 1030 1030 1030 1030 1006
1006 ...
##  $ FreeWifi             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FreeBreakfast        : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ HotelCapacity        : int  287 287 287 287 287 287 287 287 28 28 ...
##  $ HasSwimmingPool      : int  1 1 1 1 1 1 1 1 0 0 ...
```

```r
# Summarizing the data to understand the statistics of  each variable
summary(hoteldf)
```

```
##        X              CityName      Population          CityRank
##  Min.   :    1   Delhi    :2048   Min.   :    8096   Min.   : 0.00
##  1st Qu.: 3309   Jaipur   : 768   1st Qu.:  744983   1st Qu.: 2.00
##  Median : 6616   Mumbai   : 712   Median : 3046163   Median : 9.00
##  Mean   : 6616   Bangalore: 656   Mean   : 4416837   Mean   :14.83
##  3rd Qu.: 9924   Goa      : 624   3rd Qu.: 8443675   3rd Qu.:24.00
##  Max.   :13232   Kochi    : 608   Max.   :12442373   Max.   :44.00
##                  (Other)  :7816
##   IsMetroCity     IsTouristDestination   IsWeekend       IsNewYearEve
##  Min.   :0.0000   Min.   :0.0000       Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000       1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000       Median :1.0000   Median :0.0000
##  Mean   :0.2842   Mean   :0.6972       Mean   :0.6228   Mean   :0.1244
##  3rd Qu.:1.0000   3rd Qu.:1.0000       3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000       Max.   :1.0000   Max.   :1.0000
##
##           Date                          HotelName        RoomRent
##  Dec 21 2016:1611   Vivanta by Taj        :   32   Min.   :   299
##  Dec 24 2016:1611   Goldfinch Hotel       :   24   1st Qu.:  2436
##  Dec 25 2016:1611   OYO Rooms             :   24   Median :  4000
##  Dec 28 2016:1611   The Gordon House Hotel:   24   Mean   :  5474
##  Dec 31 2016:1611   Apnayt Villa          :   16   3rd Qu.:  6299
##  Dec 18 2016:1608   Bentleys Hotel Colaba :   16   Max.   :322500
##  (Other)    :3569   (Other)               :13096
##    StarRating       Airport
##  Min.   :0.000   Min.   :  0.20
##  1st Qu.:3.000   1st Qu.:  8.40
##  Median :3.000   Median : 15.00
##  Mean   :3.459   Mean   : 21.16
##  3rd Qu.:4.000   3rd Qu.: 24.00
##  Max.   :5.000   Max.   :124.00
##
##
HotelAddress
##  The Mall, Shimla                                                              :
32
##  #2-91/14/8, White Fields, Kondapur, Hitech City, Hyderabad, 500084 India:
16
##  121, City Terrace, Walchand Hirachand Marg, Mumbai, Maharashtra         :
16
##  14-4507/9, Balmatta Road, Near Jyothi Circle, Hampankatta               :
16
##  144/7, Rajiv Gandi Salai (OMR), Kottivakkam, Chennai, Tamil Nadu        :
16
##  17, Oliver Road, Colaba, Mumbai, Maharashtra                            :
16
##  (Other)
:13120
##   HotelPincode         HotelDescription    FreeWifi      FreeBreakfast
##  Min.   : 100025   3           :  120   Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.: 221001   Abc          :  112   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 395003   3-star hotel:  104   Median :1.0000   Median :1.0000
##  Mean   : 397430   3.5          :   88   Mean   :0.9259   Mean   :0.6491
##  3rd Qu.: 570001   4            :   72   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :7000157   (Other)      :12728   Max.   :1.0000   Max.   :1.0000
##                    NA's         :    8
##  HotelCapacity    HasSwimmingPool
##  Min.   :  0.00   Min.   :0.0000
##  1st Qu.: 16.00   1st Qu.:0.0000
##  Median : 34.00   Median :0.0000
##  Mean   : 62.51   Mean   :0.3558
##  3rd Qu.: 75.00   3rd Qu.:1.0000
##  Max.   :600.00   Max.   :1.0000
##

### DATA CLEANING ###

# "hoteldf" contains many variables but there  are some discrepancies in
dataset...
# ...and some variables are unappropriate and absurd which should be removed.


# identifying discrepancies in POPULATION column of dataset
dim(table(hoteldf$CityName))   # output =42

## [1] 42

dim(table(hoteldf$Population)) # output =44

## [1] 44

# as output are not equal it means there is discrepancy in Population
table(hoteldf$Population)

##
##     8096    38471    38472    38473    41377    65471    88430    98658
##      288      325        1        2      144      264      136      128
##   102138   132016   140925   169578   201026   451735   499487   595575
##       88      136       64      280       56      456      104      608
##   744983   755379   885363   957352   960787  1180570  1201815  1286678
##      392      160      120       48      336       40      264      128
##  1447187  1457723  1465625  1637875  1760285  2167447  2490891  2765348
##       48      624      112      224      432      160      136       16
##  2817105  2975440  3046163  3124458  4467797  4496694  5577940  6731790
##      128       32      768      600       80      512      424      536
##  7088416  8443675 11034555 12442373
##      416      656     2048      712

# discrepancy found in city "Munnar" population

# Removing discrepancies in POPULATION column of dataset
```

```
hoteldf$Population[hoteldf$Population==38472 | hoteldf$Population==38473]<-
38471

dim(table(hoteldf$Population)) # output =42

## [1] 42

table(hoteldf$Population)

##
##     8096    38471    41377    65471    88430    98658   102138   132016
##      288      328      144      264      136      128       88      136
##   140925   169578   201026   451735   499487   595575   744983   755379
##       64      280       56      456      104      608      392      160
##   885363   957352   960787  1180570  1201815  1286678  1447187  1457723
##      120       48      336       40      264      128       48      624
##  1465625  1637875  1760285  2167447  2490891  2765348  2817105  2975440
##      112      224      432      160      136       16      128       32
##  3046163  3124458  4467797  4496694  5577940  6731790  7088416  8443675
##      768      600       80      512      424      536      416      656
## 11034555 12442373
##     2048      712

# identifying discrepancies in DATE column of dataset
dim(table(hoteldf$Date))   # output =20 due to different formats (which
should be 8)

## [1] 20

table(hoteldf$Date)

##
##   18-Dec-16   21-Dec-16   24-Dec-16   25-Dec-16   28-Dec-16   31-Dec-16
##          44          44          44          44          44          44
##    4-Jan-16    4-Jan-17    8-Jan-16    8-Jan-17 Dec 18 2016 Dec 21 2016
##          31          13          31          13        1608        1611
## Dec 24 2016 Dec 25 2016 Dec 28 2016 Dec 31 2016 Jan 04 2017 Jan 08 2017
##        1611        1611        1611        1611        1548        1542
##  Jan 4 2017  Jan 8 2017
##          60          67

# removing discrepancies in DATE column of dataset
hoteldf$Date <- as.character(hoteldf$Date)
hoteldf$Date[hoteldf$Date=="18-Dec-16"] <-"Dec 18 2016"
hoteldf$Date[hoteldf$Date=="21-Dec-16"] <-"Dec 21 2016"
hoteldf$Date[hoteldf$Date=="24-Dec-16"] <-"Dec 24 2016"
hoteldf$Date[hoteldf$Date=="25-Dec-16"] <-"Dec 25 2016"
hoteldf$Date[hoteldf$Date=="28-Dec-16"] <-"Dec 28 2016"
hoteldf$Date[hoteldf$Date=="31-Dec-16"] <-"Dec 31 2016"
hoteldf$Date[hoteldf$Date=="4-Jan-16" | hoteldf$Date=="4-Jan-17" |
hoteldf$Date=="Jan 4 2017"] <-"Jan 04 2017"
hoteldf$Date[hoteldf$Date=="8-Jan-16" | hoteldf$Date=="8-Jan-17" |
```

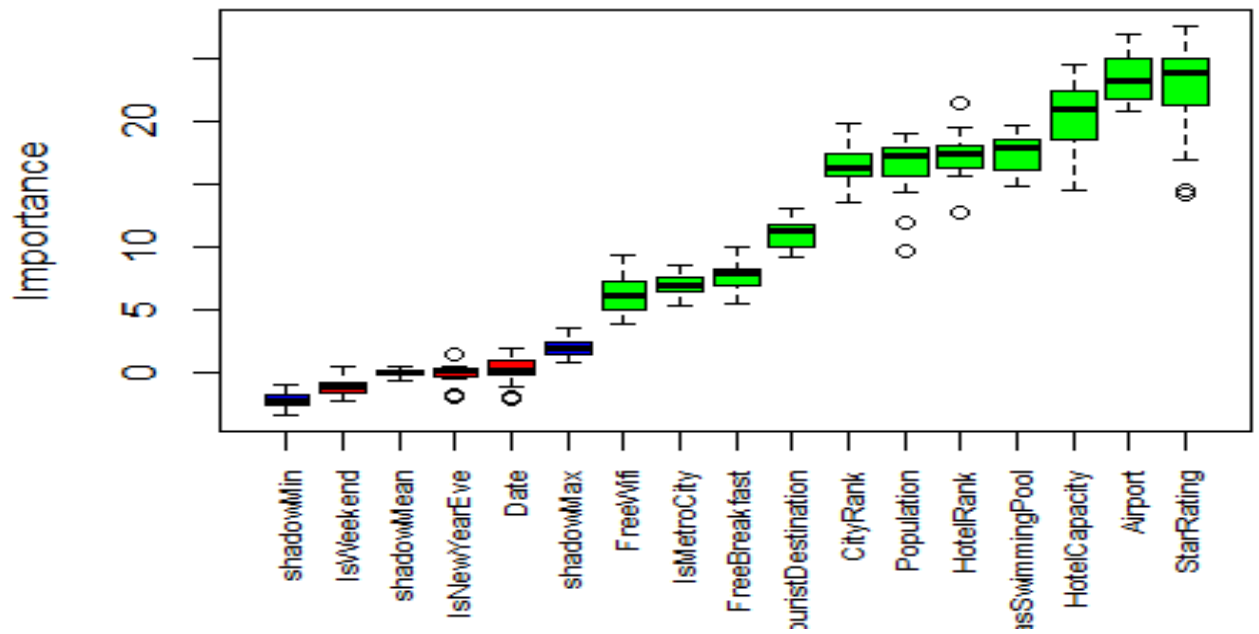
```
hoteldf$Date=="Jan 8 2017"] <-"Jan 08 2017"

dim(table(hoteldf$Date))   # output =8

## [1] 8

table(hoteldf$Date)

##
## Dec 18 2016 Dec 21 2016 Dec 24 2016 Dec 25 2016 Dec 28 2016 Dec 31 2016
##        1652        1655        1655        1655        1655        1655
## Jan 04 2017 Jan 08 2017
##        1652        1653

# Dates are converted into numeric values using dummy variables 1 to 8
hoteldf$Date <- as.numeric(as.factor(hoteldf$Date))

# converting HotelNames to HotelRank using dummy variables 1 to 1670
hoteldf$HotelRank <- as.numeric(as.factor(hoteldf$HotelName))

# removing absurd and irrelevant variables
hoteldata <- hoteldf[,-c(1,2,10,14,15,16)]
View(hoteldata)

# Summarizing the new cleaned dataset to understand the statistics of  each
variable
summary(hoteldata)

##    Population          CityRank      IsMetroCity     IsTouristDestination
##  Min.   :    8096   Min.   : 0.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  744983   1st Qu.: 2.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3046163   Median : 9.00   Median :0.0000   Median :1.0000
##  Mean   : 4416837   Mean   :14.83   Mean   :0.2842   Mean   :0.6972
##  3rd Qu.: 8443675   3rd Qu.:24.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :12442373   Max.   :44.00   Max.   :1.0000   Max.   :1.0000
##    IsWeekend       IsNewYearEve         Date        RoomRent
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.0   Min.   :   299
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.0   1st Qu.:  2436
##  Median :1.0000   Median :0.0000   Median :4.0   Median :  4000
##  Mean   :0.6228   Mean   :0.1244   Mean   :4.5   Mean   :  5474
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:6.0   3rd Qu.:  6299
##  Max.   :1.0000   Max.   :1.0000   Max.   :8.0   Max.   :322500
##    StarRating       Airport          FreeWifi      FreeBreakfast
##  Min.   :0.000   Min.   :  0.20   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:3.000   1st Qu.:  8.40   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :3.000   Median : 15.00   Median :1.0000   Median :1.0000
##  Mean   :3.459   Mean   : 21.16   Mean   :0.9259   Mean   :0.6491
##  3rd Qu.:4.000   3rd Qu.: 24.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :124.00   Max.   :1.0000   Max.   :1.0000
##  HotelCapacity    HasSwimmingPool    HotelRank
##  Min.   :  0.00   Min.   :0.0000   Min.   :   1.0
```

```
##  1st Qu.: 16.00   1st Qu.:0.0000   1st Qu.: 413.8
##  Median : 34.00   Median :0.0000   Median : 827.0
##  Mean   : 62.51   Mean   :0.3558   Mean   : 841.2
##  3rd Qu.: 75.00   3rd Qu.:1.0000   3rd Qu.:1281.0
##  Max.   :600.00   Max.   :1.0000   Max.   :1670.0
```

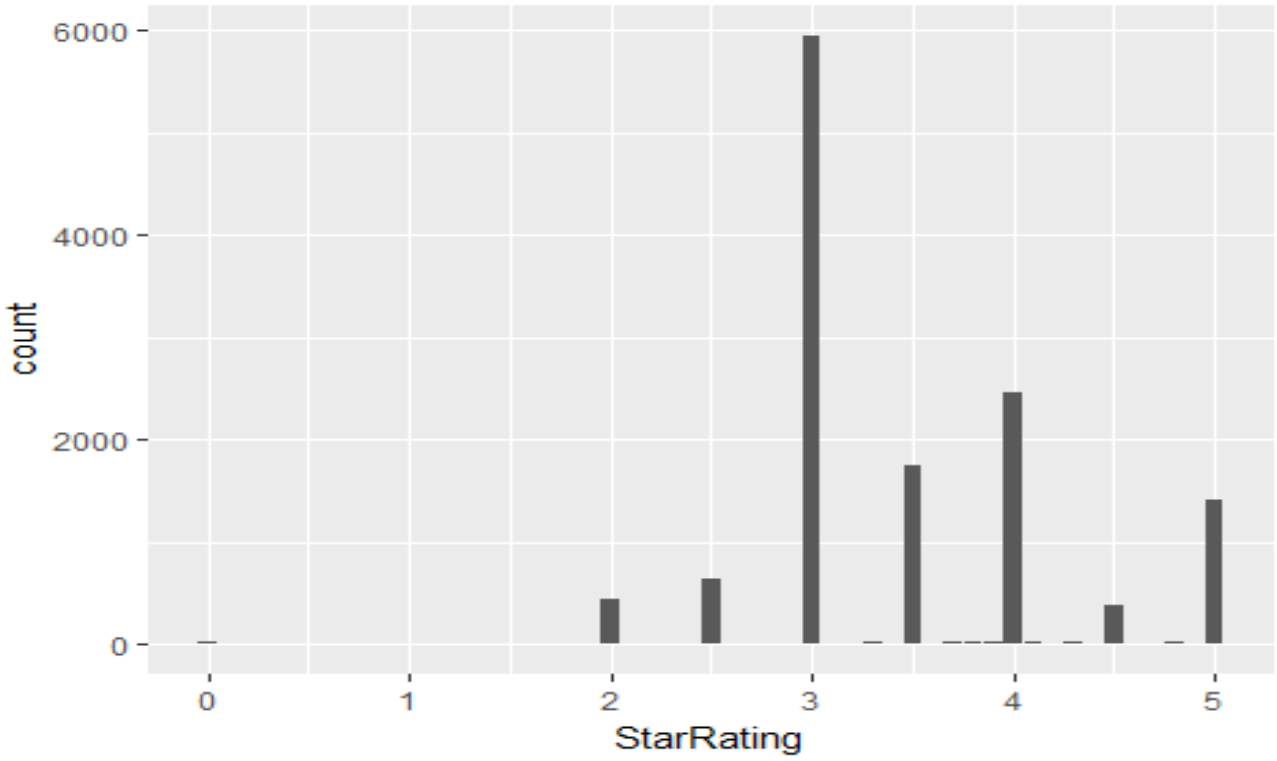
```
library(psych)
describe(hoteldata)
```

```
##                       vars     n       mean         sd  median    trimmed
## Population               1 13232 4416836.87 4258386.00 3046163 4040816.22
## CityRank                 2 13232      14.83      13.51       9      13.30
## IsMetroCity              3 13232       0.28       0.45       0       0.23
## IsTouristDestination     4 13232       0.70       0.46       1       0.75
## IsWeekend                5 13232       0.62       0.48       1       0.65
## IsNewYearEve             6 13232       0.12       0.33       0       0.03
## Date                     7 13232       4.50       2.29       4       4.50
## RoomRent                 8 13232    5473.99    7333.12    4000    4383.33
## StarRating               9 13232       3.46       0.76       3       3.40
## Airport                 10 13232      21.16      22.76      15      16.39
## FreeWifi                11 13232       0.93       0.26       1       1.00
## FreeBreakfast           12 13232       0.65       0.48       1       0.69
## HotelCapacity           13 13232      62.51      76.66      34      46.03
## HasSwimmingPool         14 13232       0.36       0.48       0       0.32
## HotelRank               15 13232     841.19     488.16     827     841.18
##                            mad    min      max      range  skew kurtosis
## Population           3846498.95 8096.0 12442373 12434277.0  0.68    -1.08
## CityRank                  11.86    0.0       44       44.0  0.69    -0.76
## IsMetroCity                0.00    0.0        1        1.0  0.96    -1.08
## IsTouristDestination       0.00    0.0        1        1.0 -0.86    -1.26
## IsWeekend                  0.00    0.0        1        1.0 -0.51    -1.74
## IsNewYearEve               0.00    0.0        1        1.0  2.28     3.18
## Date                       2.97    1.0        8        7.0  0.00    -1.24
## RoomRent                2653.85  299.0   322500   322201.0 16.75   582.06
## StarRating                 0.74    0.0        5        5.0  0.48     0.25
## Airport                   11.12    0.2      124      123.8  2.73     7.89
## FreeWifi                   0.00    0.0        1        1.0 -3.25     8.57
## FreeBreakfast              0.00    0.0        1        1.0 -0.62    -1.61
## HotelCapacity             28.17    0.0      600      600.0  2.95    11.39
## HasSwimmingPool            0.00    0.0        1        1.0  0.60    -1.64
## HotelRank                641.97    1.0     1670     1669.0  0.01    -1.25
##                             se
## Population           37019.65
## CityRank                 0.12
## IsMetroCity              0.00
## IsTouristDestination     0.00
## IsWeekend                0.00
## IsNewYearEve             0.00
## Date                     0.02
## RoomRent                63.75
```

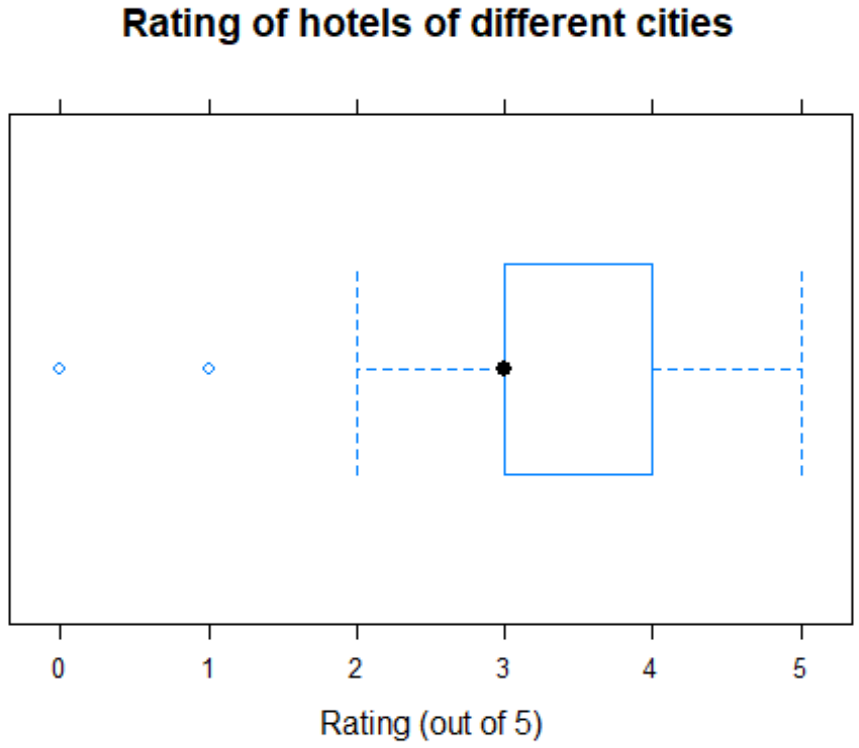
```
## StarRating               0.01
## Airport                  0.20
## FreeWifi                 0.00
## FreeBreakfast            0.00
## HotelCapacity            0.67
## HasSwimmingPool          0.00
## HotelRank                4.24

# selecting most important 3 variables for predicting Hotel Rent using Boruta Package
library(Boruta)

## Warning: package 'Boruta' was built under R version 3.4.1

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.4.1

impout <-Boruta(RoomRent~ Population + CityRank +IsMetroCity +
IsTouristDestination + IsWeekend +
                 StarRating +Airport +  FreeWifi + IsNewYearEve + Date +
HotelCapacity +
                HasSwimmingPool + FreeBreakfast + HotelRank , data= hoteldata)
print (impout)

## Boruta performed 15 iterations in 3.249309 mins.
##  11 attributes confirmed important: Airport, CityRank,
## FreeBreakfast, FreeWifi, HasSwimmingPool and 6 more;
##  3 attributes confirmed unimportant: Date, IsNewYearEve,
## IsWeekend;

plot(impout, xlab = "", xaxt = "n")
lz<-lapply(1:ncol(impout$ImpHistory) ,function(i)
impout$ImpHistory[is.finite(impout$ImpHistory[,i]),i])
names(lz) <- colnames(impout$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),at = 1:ncol(impout$ImpHistory),
cex.axis = 0.7)
```



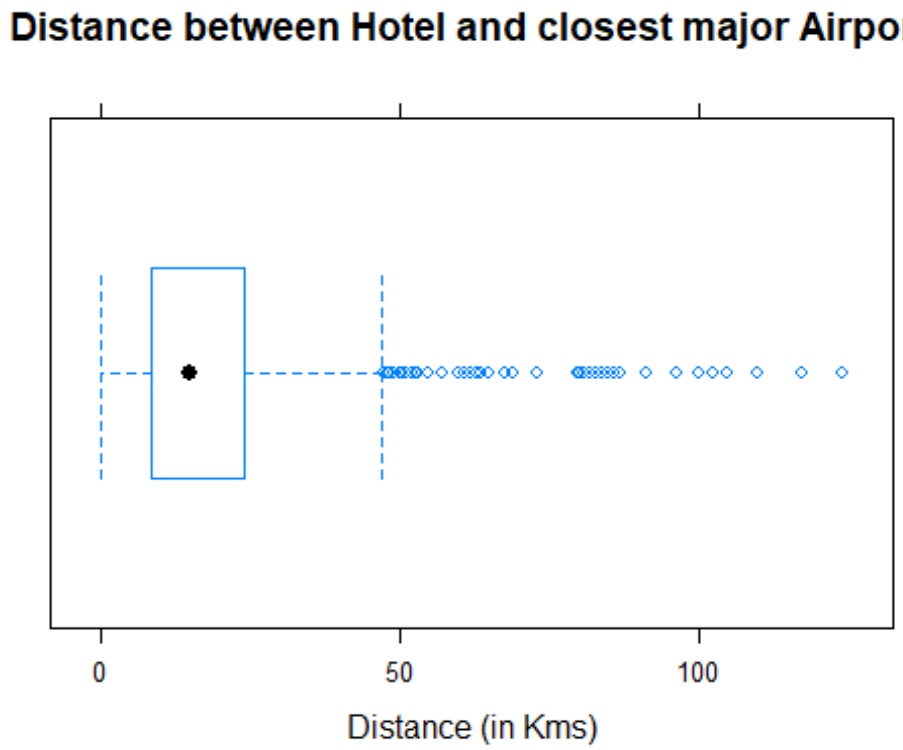

```
## StarRating , Airport , HotelCapacity are 3 most important variables
## RoomRent = F(StarRating , Airport , HotelCapacity)
library(ggplot2)

##
## Attaching package: 'ggplot2'

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

table(hoteldata$StarRating)

##
##    0    1    2  2.5    3  3.2  3.3  3.4  3.5  3.6  3.7  3.8  3.9    4  4.1
##   16    8  440  632 5953    8   16    8 1752    8   24   16   32 2463   24
##  4.3  4.4  4.5  4.7  4.8    5
##   16    8  376    8   16 1408

ggplot(hoteldata, aes(x=StarRating))+ geom_bar()
```



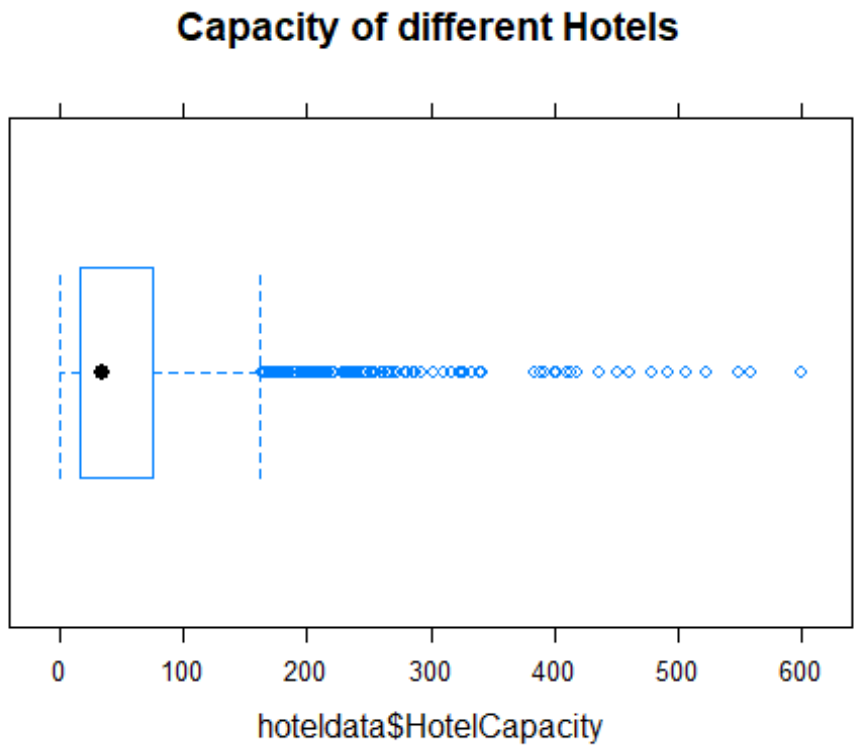

```
library(lattice)
bwplot(hoteldata$StarRating, horizontal = TRUE, xlab = "Rating (out of
5)",main = "Rating of hotels of different cities")
```



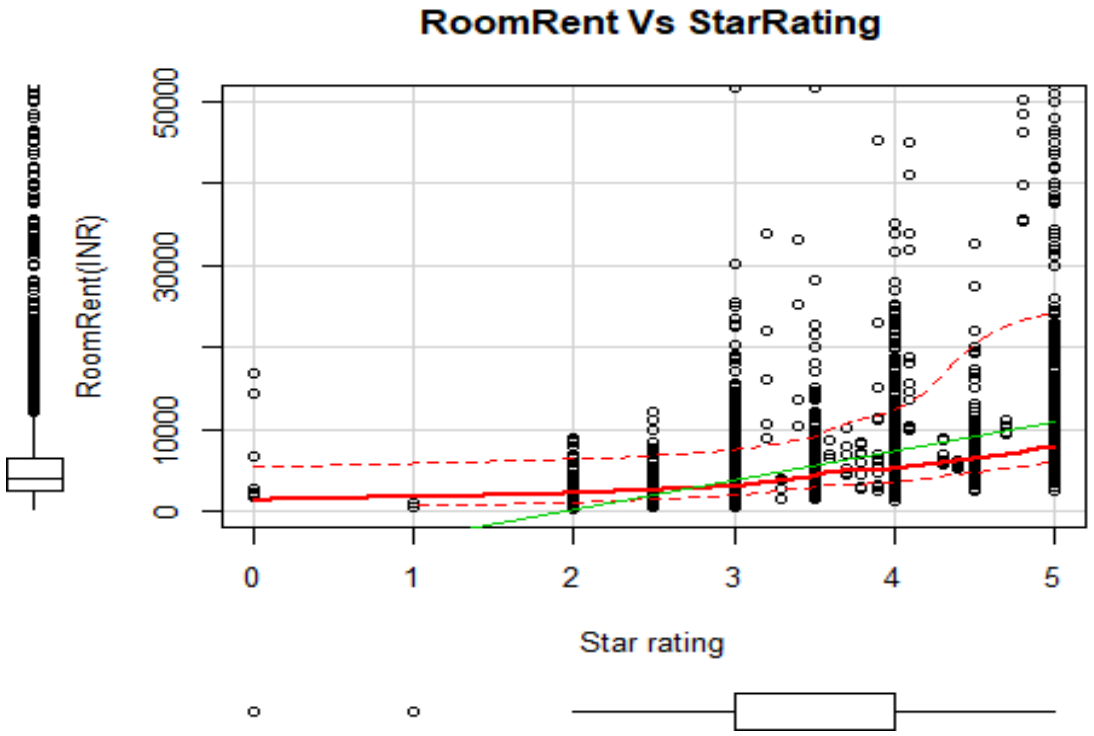
```
bwplot(hoteldata$Airport, horizontal = TRUE, xlab = "Distance (in Kms)", main
= "Distance between Hotel and closest major Airport")
```



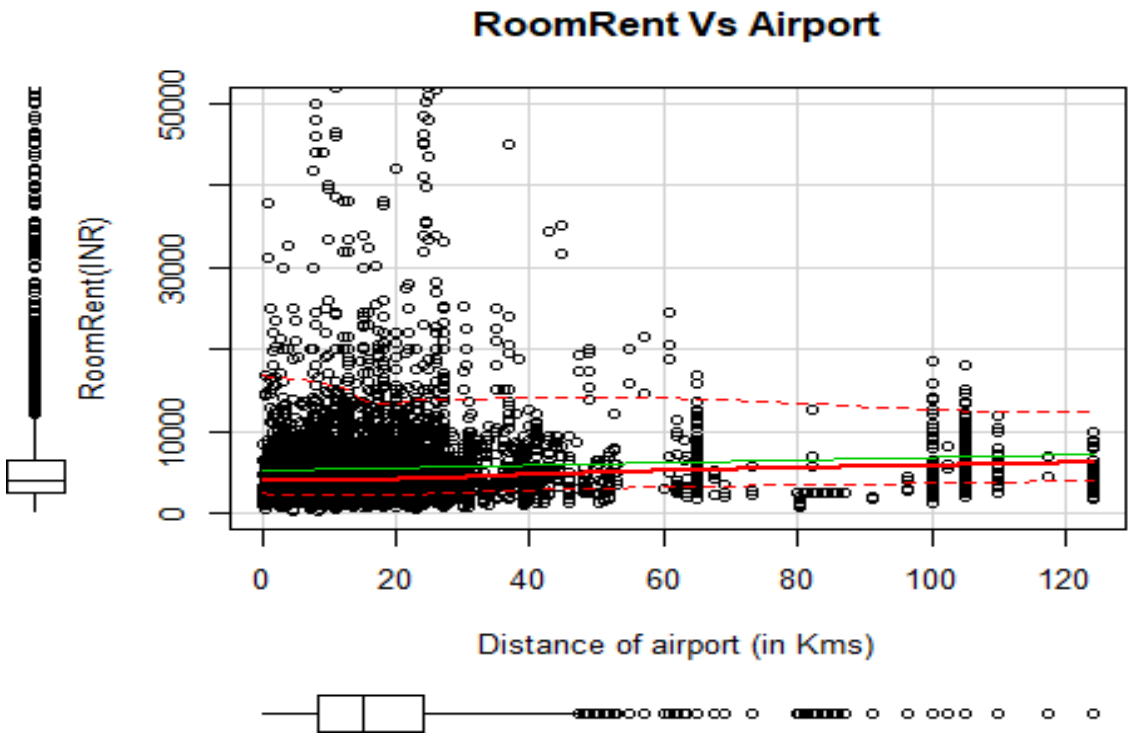

```
bwplot(hoteldata$HotelCapacity, horizontal = TRUE, main = "Capacity of
different Hotels")
```



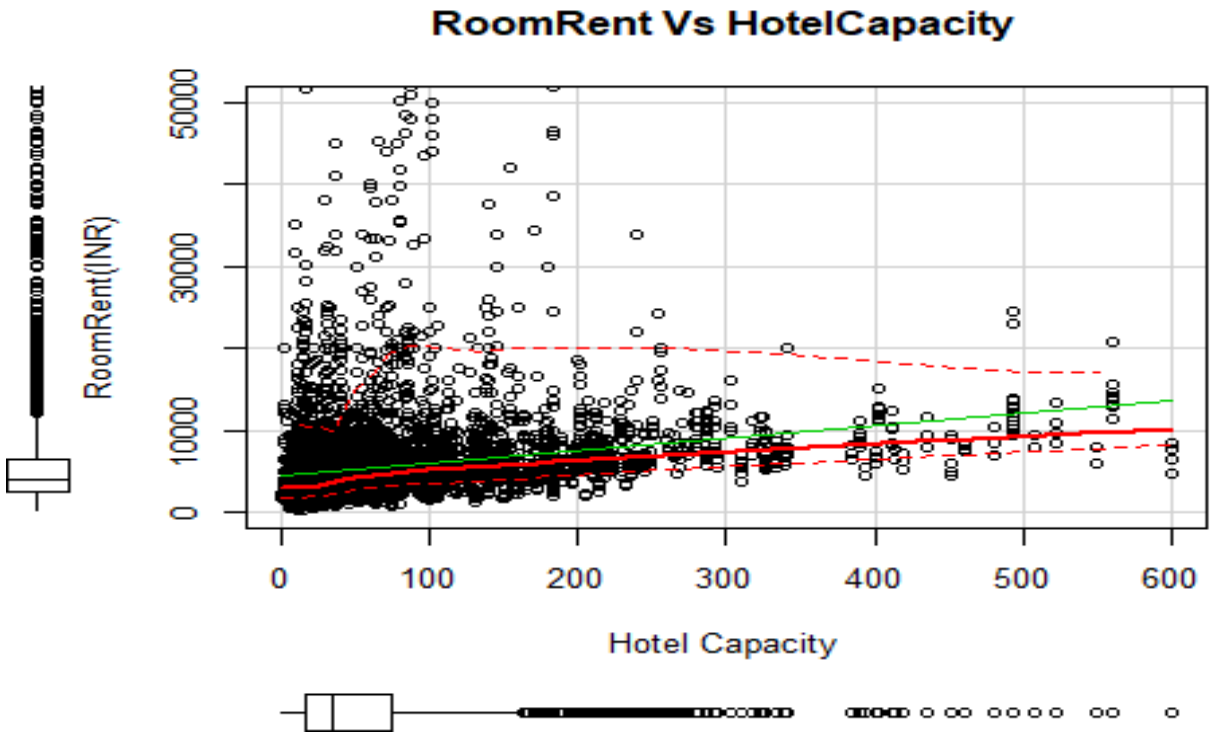

```
library(car)

##
## Attaching package: 'car'

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

#Visualizing relation between StarRating and RoomRent
scatterplot(hoteldata$StarRating,hoteldata$RoomRent,ylim=c(0,50000),main="Roo
mRent Vs StarRating", xlab="Star rating" ,ylab = "RoomRent(INR)")
```



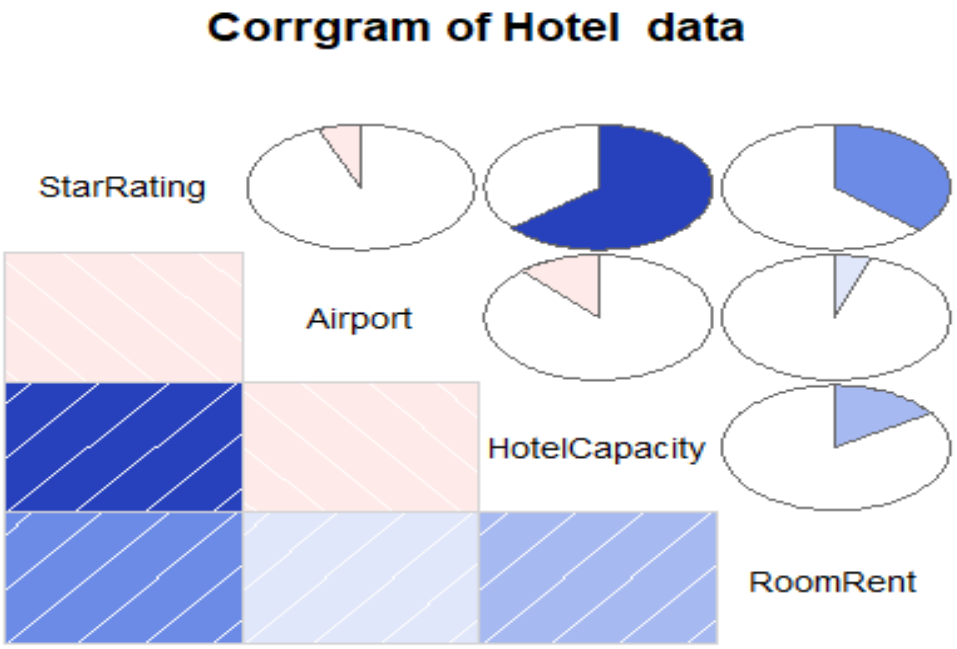

```
#Visualizing relation between RoomRent and Airport
scatterplot(hoteldata$Airport,hoteldata$RoomRent,ylim=c(0,50000),main="RoomRe
nt Vs Airport", ylab="RoomRent(INR)",xlab = "Distance of airport (in Kms)")
```



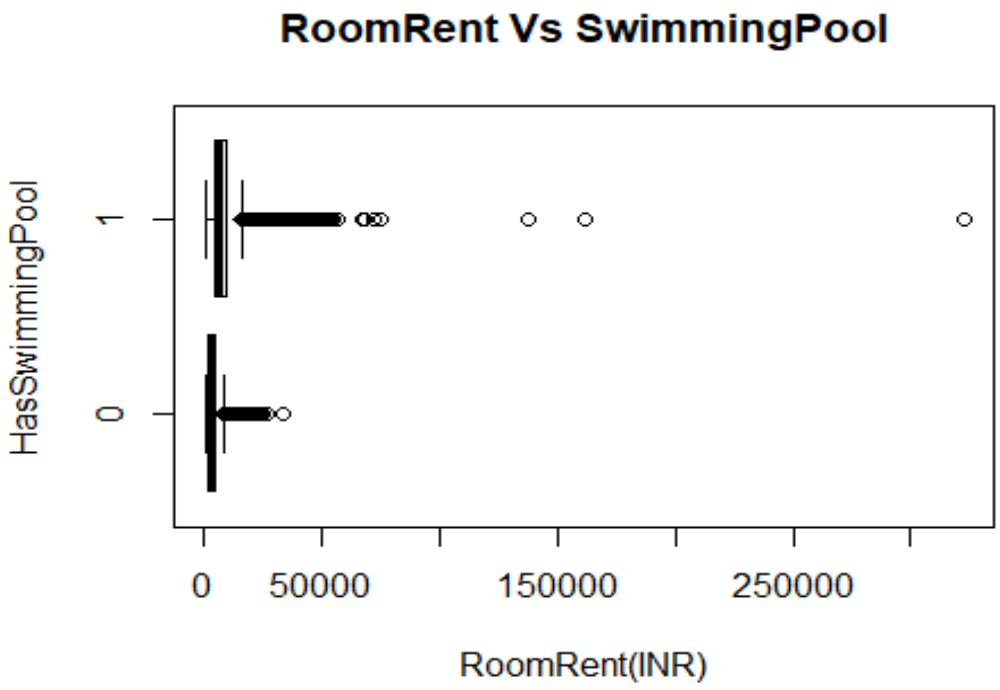
```r
#Visualizing relation between RoomRent and HotelCapacity
scatterplot(hoteldata$HotelCapacity,hoteldata$RoomRent,ylim=c(0,50000),main="
RoomRent Vs HotelCapacity", ylab="RoomRent(INR)",xlab = "Hotel Capacity")
```



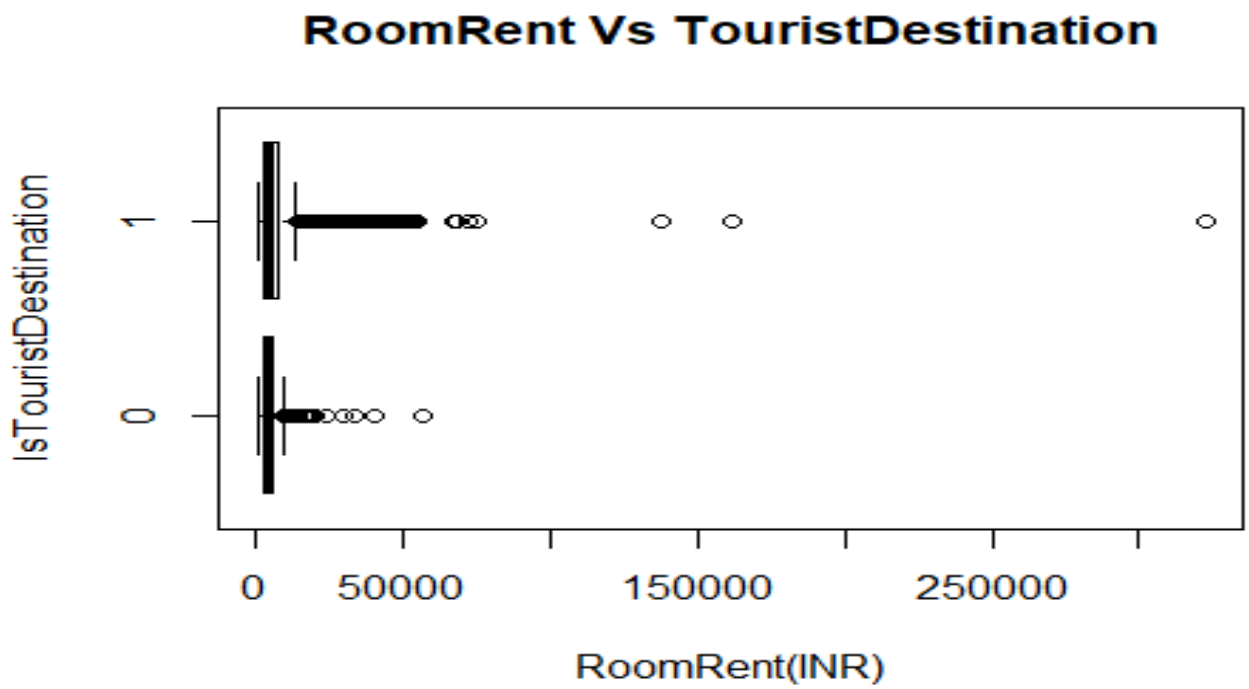

```r
# understanding how are the most important variables correlated pair-wise
library(corrgram)

## Warning: package 'corrgram' was built under R version 3.4.1

corrgram(hoteldata[,c("StarRating","Airport","HotelCapacity","RoomRent")],
lower.panel=panel.shade,
         upper.panel=panel.pie, text.panel=panel.txt,main="Corrgram of Hotel
data")
```



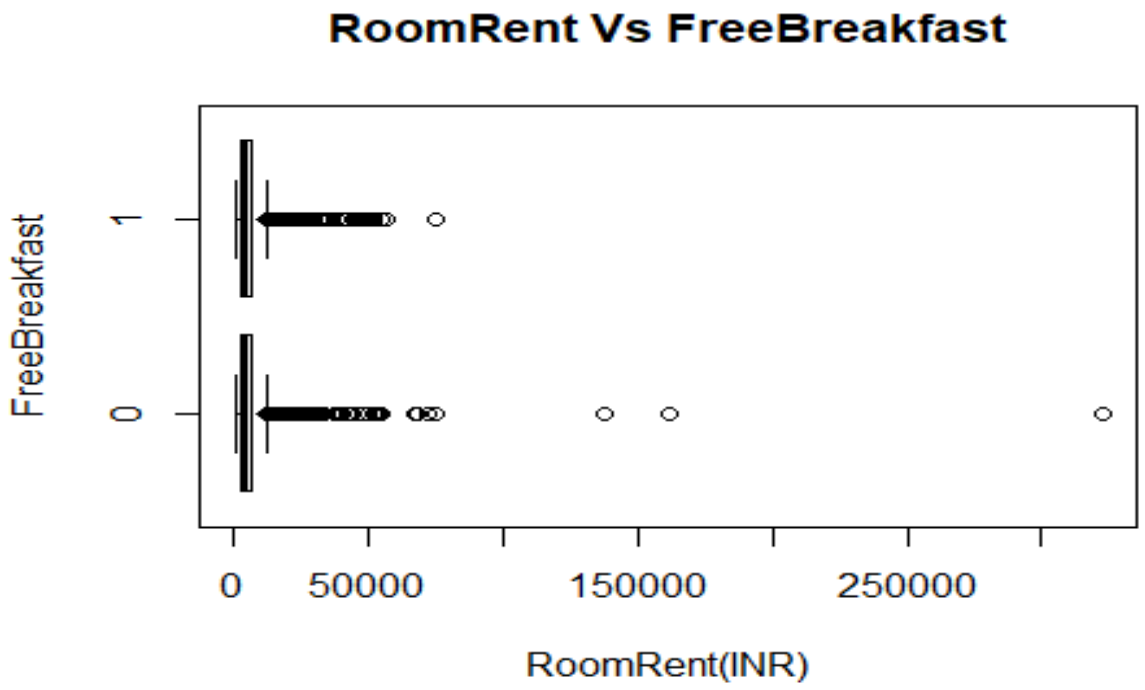

```
# Variance-Covariance Matrix
cov(hoteldata[,c("StarRating","Airport","HotelCapacity","RoomRent")])

##                 StarRating     Airport HotelCapacity     RoomRent
## StarRating       0.5718875   -1.048528      36.95522     2048.375
## Airport         -1.0485276  518.013328    -205.32017     8287.179
## HotelCapacity   36.9552206 -205.320172    5877.26810    88753.413
## RoomRent      2048.3754792 8287.178584   88753.41284 53774601.806

#1. H1 : The Hotels which having swimmingpools have higher RoomRent.
aggregate(hoteldata$RoomRent,
list(swimmingpool=hoteldata$HasSwimmingPool),mean)

##   swimmingpool        x
## 1            0 3775.566
## 2            1 8549.052

boxplot( RoomRent~HasSwimmingPool, hoteldata, horizontal=TRUE,main="RoomRent
Vs SwimmingPool",ylab="HasSwimmingPool" ,xlab = "RoomRent(INR)")
```



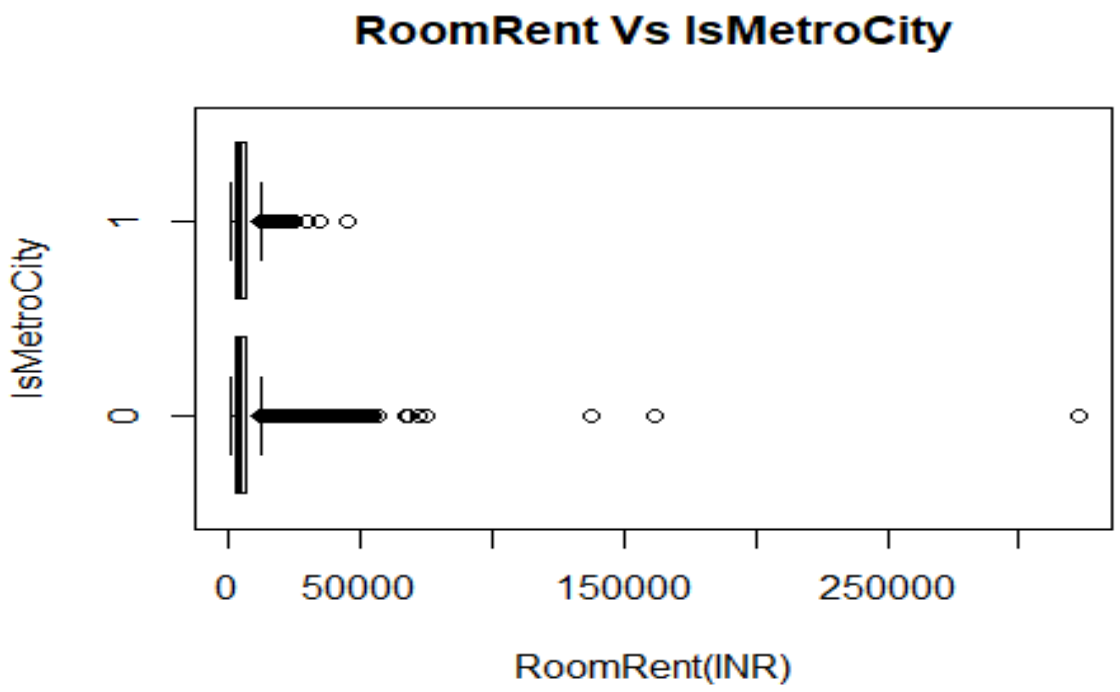

```
t.test(RoomRent ~ HasSwimmingPool , data=hoteldata, alternative ="less")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by HasSwimmingPool
## t = -29.013, df = 5011.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4502.814
## sample estimates:
## mean in group 0 mean in group 1
##        3775.566        8549.052

#2. H1 : The hotels in city having tourist destination have higher RoomRent.
aggregate(hoteldata$RoomRent,
list(TouristDestination=hoteldata$IsTouristDestination),mean)

##   TouristDestination        x
## 1                  0 4111.003
## 2                  1 6066.024

boxplot(RoomRent~IsTouristDestination, hoteldata,
horizontal=TRUE,main="RoomRent Vs
TouristDestination",ylab="IsTouristDestination" ,xlab = "RoomRent(INR)")
```



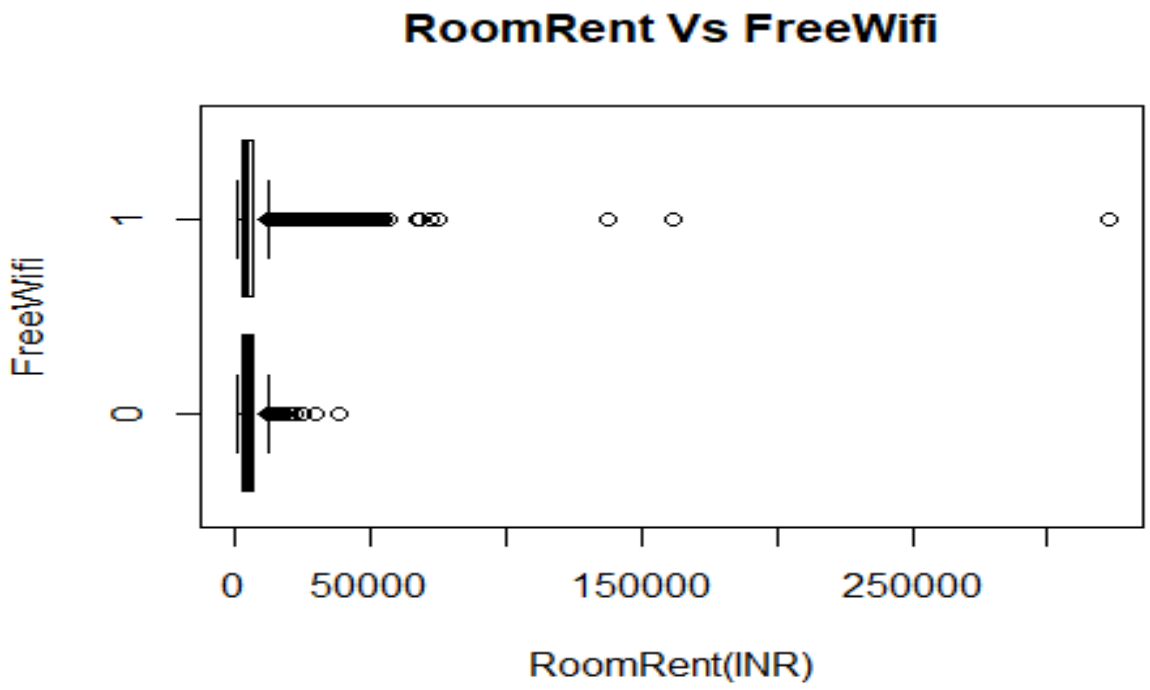

```
t.test(RoomRent ~ IsTouristDestination , data=hoteldata, alternative ="less")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by IsTouristDestination
## t = -19.449, df = 12888, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1789.665
## sample estimates:
## mean in group 0 mean in group 1
##        4111.003        6066.024

#3. H1 : The Hotels having free breakfast have higher RoomRent.
aggregate(hoteldata$RoomRent,
list(FreeBreakfast=hoteldata$FreeBreakfast),mean)

##   FreeBreakfast        x
## 1             0 5573.790
## 2             1 5420.044

boxplot(RoomRent~FreeBreakfast, hoteldata, horizontal=TRUE,main="RoomRent Vs
FreeBreakfast",ylab="FreeBreakfast" ,xlab = "RoomRent(INR)")
```



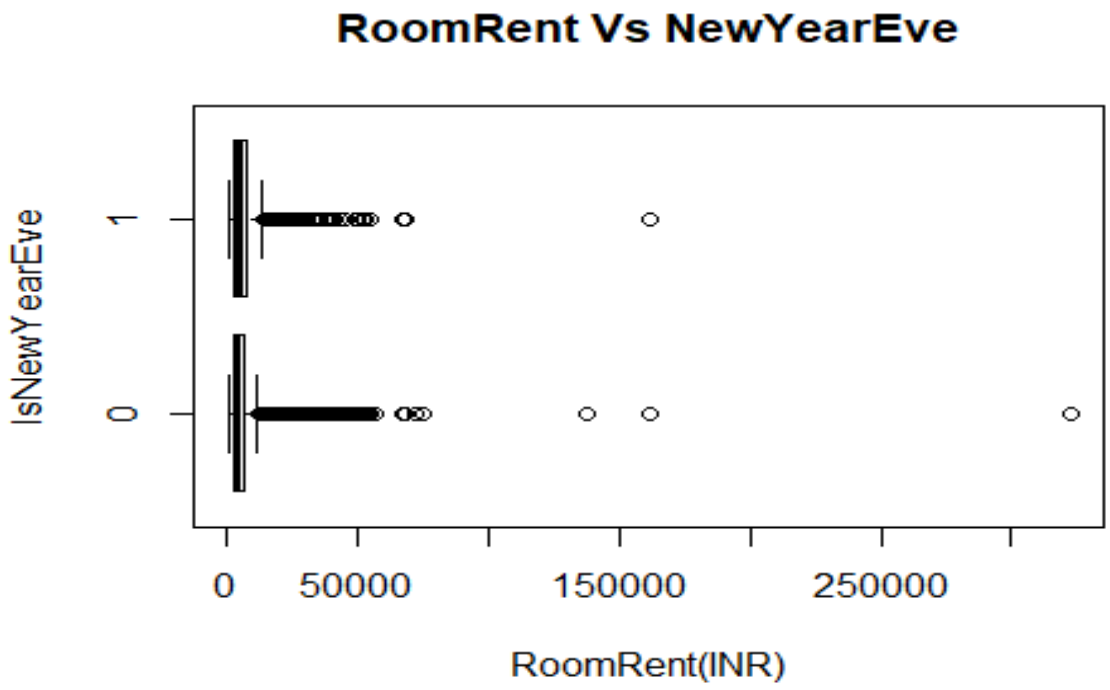

```
t.test(RoomRent ~ FreeBreakfast , data=hoteldata, alternative ="less")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by FreeBreakfast
## t = 0.98095, df = 6212.3, p-value = 0.8367
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 411.5844
## sample estimates:
## mean in group 0 mean in group 1
##        5573.790        5420.044

#4. H1 : The Hotels in non Metro city have higher RoomRent than metro city.
aggregate(hoteldata$RoomRent, list(Metrocity=hoteldata$IsMetroCity),mean)

##   Metrocity        x
## 1         0 5782.794
## 2         1 4696.073

boxplot(RoomRent~IsMetroCity, hoteldata, horizontal=TRUE,main="RoomRent Vs
IsMetroCity",ylab="IsMetroCity" ,xlab = "RoomRent(INR)")
```



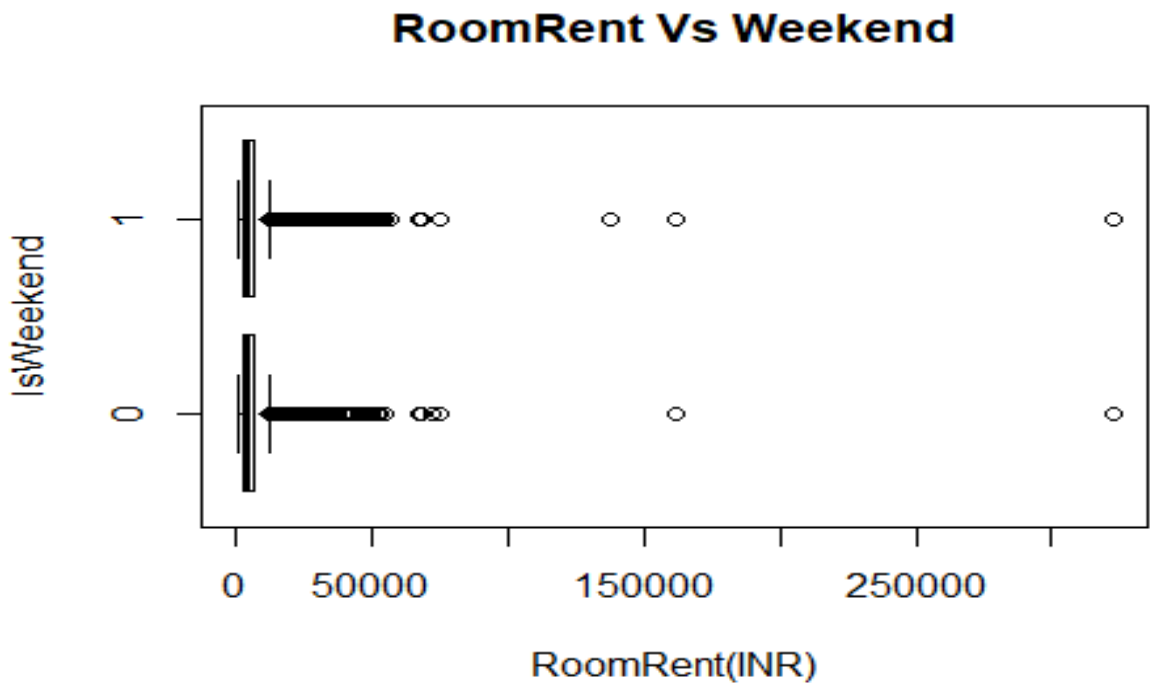

```
t.test(RoomRent ~ IsMetroCity , data=hoteldata, alternative ="greater")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by IsMetroCity
## t = 10.721, df = 13224, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  919.9785      Inf
## sample estimates:
## mean in group 0 mean in group 1
##        5782.794        4696.073

#5. H1 : The Hotels  having free wifi have higher RoomRent.
aggregate(hoteldata$RoomRent, list(Freewifi=hoteldata$FreeWifi),mean)

##   Freewifi        x
## 1        0 5380.004
## 2        1 5481.518

boxplot(RoomRent~FreeWifi, hoteldata, horizontal=TRUE,main="RoomRent Vs
FreeWifi",ylab="FreeWifi" ,xlab = "RoomRent(INR)")
```




```
t.test(RoomRent ~ FreeWifi , data=hoteldata, alternative ="less")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by FreeWifi
## t = -0.76847, df = 1804.7, p-value = 0.2212
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf 115.882
## sample estimates:
## mean in group 0 mean in group 1
##        5380.004        5481.518

#6. H1 : The Hotels on newyeareve have higher RoomRent.
aggregate(hoteldata$RoomRent, list(NewyearEve=hoteldata$IsNewYearEve),mean)

##   NewyearEve        x
## 1          0 5367.606
## 2          1 6222.826

boxplot(RoomRent~IsNewYearEve, hoteldata, horizontal=TRUE,main="RoomRent Vs
NewYearEve",ylab="IsNewYearEve" ,xlab = "RoomRent(INR)")
```

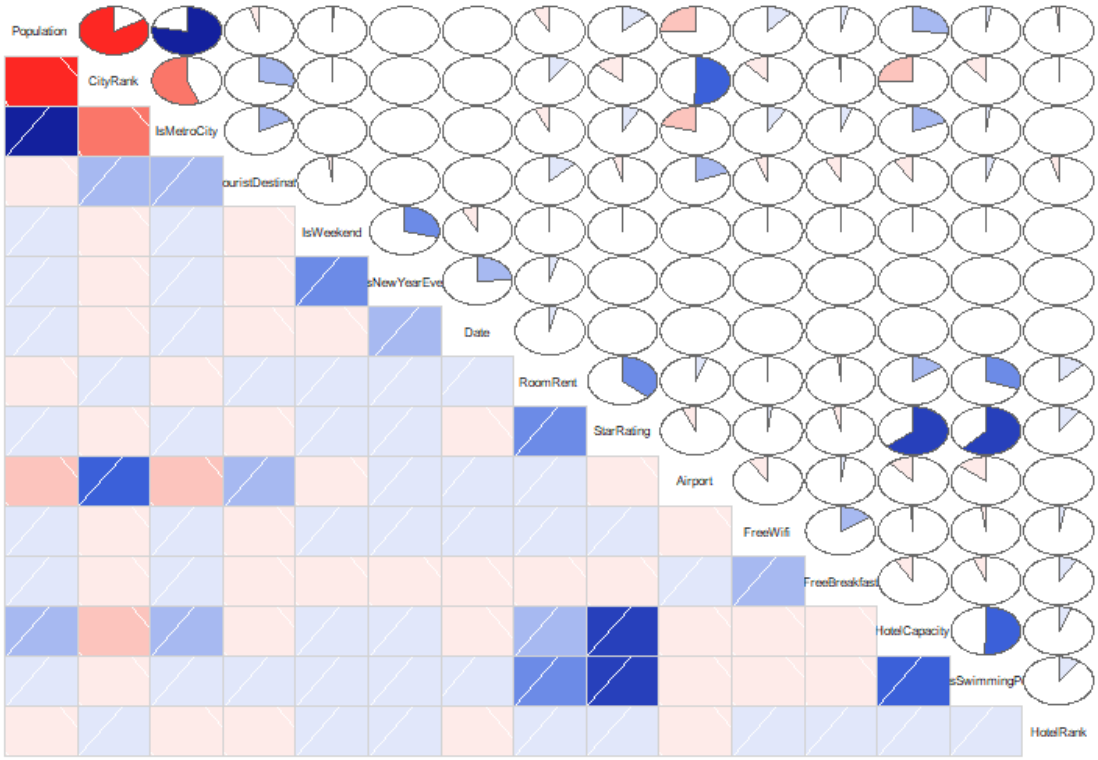



```
t.test(RoomRent ~ IsNewYearEve , data=hoteldata, alternative ="less")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by IsNewYearEve
## t = -4.1793, df = 2065, p-value = 1.523e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -518.4763
## sample estimates:
## mean in group 0 mean in group 1
##        5367.606        6222.826

#7. H1 : The Hotels on weekend have higher RoomRent.
aggregate(hoteldata$RoomRent, list(Weekend=hoteldata$IsWeekend),mean)

##   Weekend        x
## 1       0 5430.835
## 2       1 5500.129

boxplot(RoomRent~IsWeekend, hoteldata, horizontal=TRUE,main="RoomRent Vs
Weekend",ylab="IsWeekend" ,xlab = "RoomRent(INR)")
```

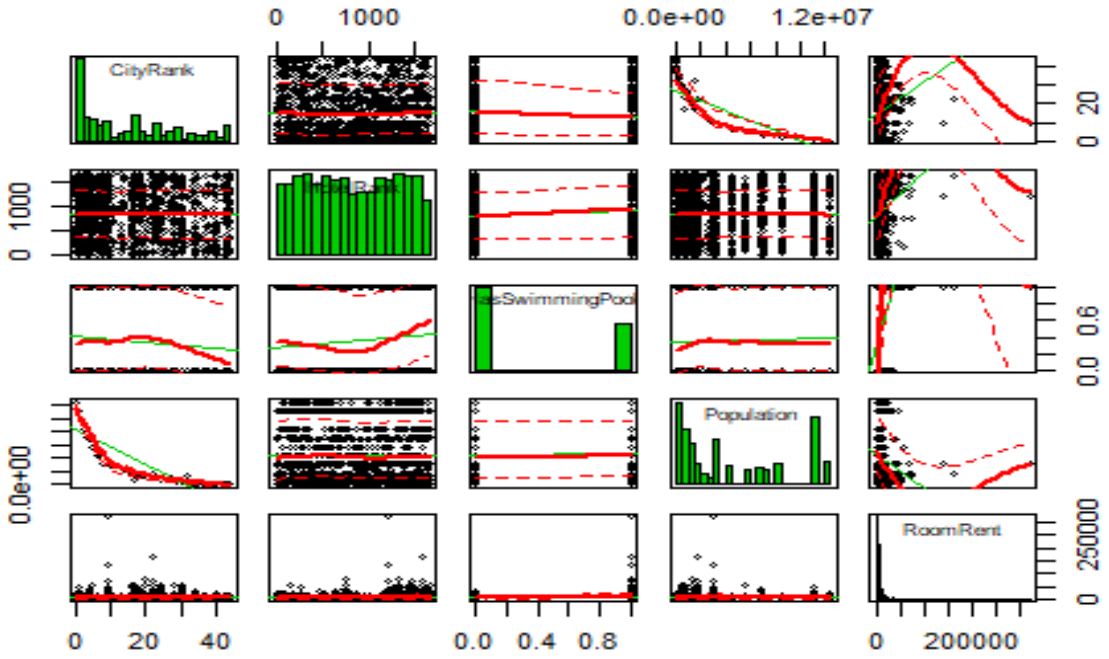



```
t.test(RoomRent ~ IsWeekend, data=hoteldata, alternative ="less")

##
##  Welch Two Sample t-test
##
## data:  RoomRent by IsWeekend
## t = -0.51853, df = 9999.4, p-value = 0.302
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 150.5351
## sample estimates:
## mean in group 0 mean in group 1
##        5430.835        5500.129

corrgram(hoteldata, main = "corrgram of Six airlines variables", lower.panel
= panel.shade,
         upper.panel = panel.pie, text.panel = panel.txt)
```

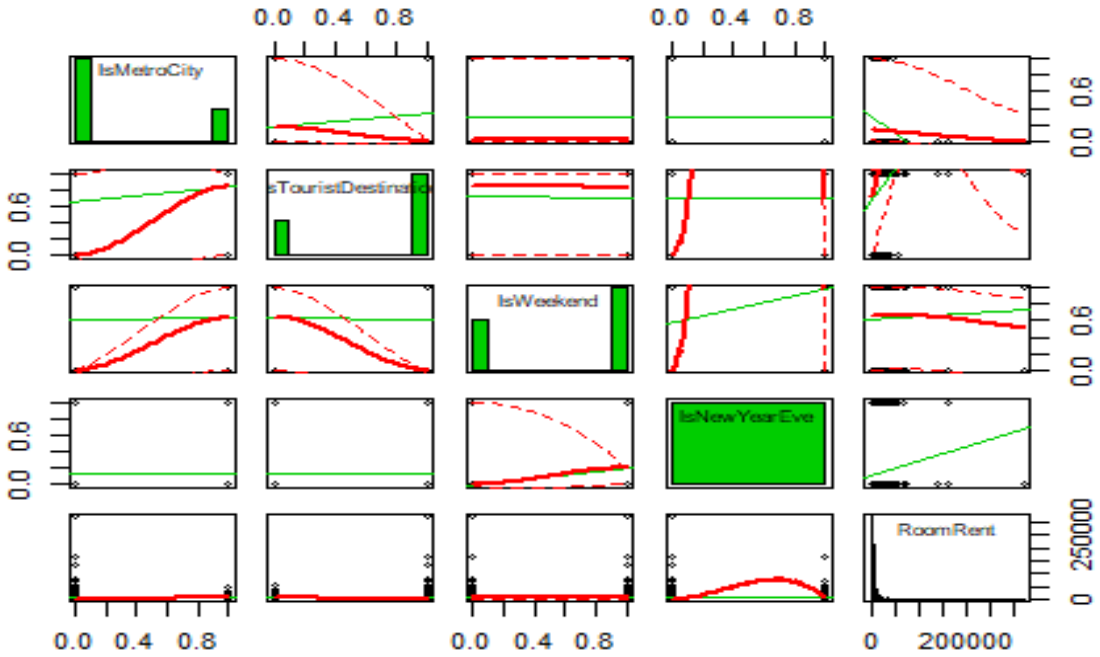
corrgram for factors vs Hotel Room Prices




```
# Analysing correlation of RoomRent with different factors
scatterplotMatrix(formula = ~  CityRank + HotelRank + HasSwimmingPool +
Population + RoomRent,cex=0.6, data=hoteldata, diagonal="histogram")
```




```
scatterplotMatrix(formula = ~  IsMetroCity + IsTouristDestination + IsWeekend
+ IsNewYearEve + RoomRent,
                  cex=0.6, data=hoteldata, diagonal="histogram")

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```




```
# Formulating multivariate linear regression model to fit room rent with
respect to different factors

# model 1 - with only most important features
fit1<-lm(RoomRent ~ StarRating + Airport +HotelCapacity + HasSwimmingPool,
data=hoteldata)
summary(fit1)

##
## Call:
## lm(formula = RoomRent ~ StarRating + Airport + HotelCapacity +
##     HasSwimmingPool, data = hoteldata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10785  -2265   -876    982 310437
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -7288.048    341.691 -21.329   <2e-16 ***
## StarRating       3522.990    111.531  31.588   <2e-16 ***
## Airport            25.344      2.590   9.786   <2e-16 ***
## HotelCapacity     -14.776      1.006 -14.695   <2e-16 ***
## HasSwimmingPool  2708.400    158.397  17.099   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6687 on 13227 degrees of freedom
## Multiple R-squared:  0.1688, Adjusted R-squared:  0.1686
## F-statistic: 671.7 on 4 and 13227 DF,  p-value: < 2.2e-16
```

```
AIC(fit1)
```

```
## [1] 270648.7
```

```
# model 2 - with all features in "hoteldata"
fit2<-lm(RoomRent ~ . , data=hoteldata)
summary(fit2)
```

```
##
## Call:
## lm(formula = RoomRent ~ ., data = hoteldata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12273  -2290   -702   1140 309442
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -9.786e+03  4.691e+02 -20.861  < 2e-16 ***
## Population           -1.130e-04  3.573e-05  -3.162 0.001570 **
## CityRank              3.011e-01  1.029e+01   0.029 0.976656
## IsMetroCity          -7.299e+02  2.153e+02  -3.390 0.000702 ***
## IsTouristDestination  2.007e+03  1.474e+02  13.613  < 2e-16 ***
## IsWeekend            -3.740e+01  1.248e+02  -0.300 0.764337
## IsNewYearEve          7.311e+02  1.884e+02   3.880 0.000105 ***
## Date                  7.305e+01  2.603e+01   2.806 0.005023 **
## StarRating            3.531e+03  1.103e+02  32.022  < 2e-16 ***
## Airport               9.338e+00  3.154e+00   2.961 0.003073 **
## FreeWifi              5.030e+02  2.230e+02   2.255 0.024138 *
## FreeBreakfast         4.462e+01  1.231e+02   0.362 0.717108
## HotelCapacity        -1.018e+01  1.028e+00  -9.907  < 2e-16 ***
## HasSwimmingPool       2.044e+03  1.610e+02  12.699  < 2e-16 ***
## HotelRank             1.389e+00  1.184e-01  11.728  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6565 on 13217 degrees of freedom
## Multiple R-squared:  0.1994, Adjusted R-squared:  0.1985
## F-statistic: 235.1 on 14 and 13217 DF,  p-value: < 2.2e-16
```

```
AIC(fit2)
```

```
## [1] 270173.3
```

```r
# model 3 - best fit model
fit3 <- lm(RoomRent ~ . - CityRank - FreeBreakfast - IsWeekend,
data=hoteldata)
summary(fit3)
```

```
##
## Call:
## lm(formula = RoomRent ~ . - CityRank - FreeBreakfast - IsWeekend,
##     data = hoteldata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12263  -2287   -704   1142 309399
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -9.793e+03  4.254e+02 -23.020  < 2e-16 ***
## Population           -1.140e-04  2.252e-05  -5.061 4.23e-07 ***
## IsMetroCity          -7.235e+02  2.121e+02  -3.411 0.000650 ***
## IsTouristDestination  2.005e+03  1.368e+02  14.650  < 2e-16 ***
## IsNewYearEve          7.127e+02  1.784e+02   3.995 6.50e-05 ***
## Date                  7.427e+01  2.570e+01   2.890 0.003862 **
## StarRating            3.533e+03  1.100e+02  32.122  < 2e-16 ***
## Airport               9.437e+00  2.701e+00   3.494 0.000478 ***
## FreeWifi              5.148e+02  2.206e+02   2.334 0.019625 *
## HotelCapacity        -1.021e+01  1.023e+00  -9.985  < 2e-16 ***
## HasSwimmingPool       2.042e+03  1.592e+02  12.832  < 2e-16 ***
## HotelRank             1.393e+00  1.180e-01  11.807  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6564 on 13220 degrees of freedom
## Multiple R-squared:  0.1994, Adjusted R-squared:  0.1987
## F-statistic: 299.2 on 11 and 13220 DF,  p-value: < 2.2e-16
```

```r
# model 3 is equivalent to (RoomRent ~ StarRating + HotelRank + Airport +
HotelCapacity + HasSwimmingPool + Population+
#             IsMetroCity + IsTouristDestination  + Date+ FreeWifi  +
IsNewYearEve ,data = hoteldata)

# AIC of best model
AIC(fit3)
```

```
## [1] 270167.5
```

```r
#Coefficents of the best model
fit3$coefficients
```

```
##          (Intercept)           Population          IsMetroCity
##        -9.793152e+03        -1.139882e-04        -7.235076e+02
## IsTouristDestination         IsNewYearEve                 Date
##         2.004711e+03         7.127156e+02         7.426967e+01
##           StarRating              Airport             FreeWifi
##         3.532794e+03         9.437204e+00         5.147855e+02
##        HotelCapacity      HasSwimmingPool           HotelRank
##        -1.021488e+01         2.042397e+03         1.392768e+00
```