Capstone\_Project\_Code.R

Tanmay

Wed Jul 12 03:40:36 2017

# CAPSTONE PROJECT ANALYSIS  
# NAME: TANMAY AGARWAL   
# COLLEGE: R.V. COLLEGE OF ENGINEERING  
# EMAIL: tanmay151997@gmail.com   
  
# Setting working Directory  
setwd("C:/Users/Tanmay/Desktop/Data Science and Analytics and Programming in R/Major Project")  
  
# Reading Data set  
caps<-read.csv(paste("Cities42.csv", sep = ""))  
  
# Attching the data set to use the columns easily  
attach(caps)  
  
# Summary  
summary(caps)

## 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   
##

# Loading packages  
library(psych)  
library(car)

##   
## Attaching package: 'car'

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

library(gplots)

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(gmodels)  
library(lattice)  
library(vcd)

## Loading required package: grid

library(Matrix)  
library(corrgram)

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

library(corrplot)  
  
# Desribing Data  
describe(caps)

## vars n mean sd median trimmed  
## X 1 13232 6616.50 3819.89 6616.5 6616.50  
## CityName\* 2 13232 18.07 11.72 16.0 17.29  
## Population 3 13232 4416836.87 4258386.00 3046163.0 4040816.22  
## CityRank 4 13232 14.83 13.51 9.0 13.30  
## IsMetroCity 5 13232 0.28 0.45 0.0 0.23  
## IsTouristDestination 6 13232 0.70 0.46 1.0 0.75  
## IsWeekend 7 13232 0.62 0.48 1.0 0.65  
## IsNewYearEve 8 13232 0.12 0.33 0.0 0.03  
## Date\* 9 13232 14.26 2.82 14.0 14.39  
## HotelName\* 10 13232 841.19 488.16 827.0 841.18  
## RoomRent 11 13232 5473.99 7333.12 4000.0 4383.33  
## StarRating 12 13232 3.46 0.76 3.0 3.40  
## Airport 13 13232 21.16 22.76 15.0 16.39  
## HotelAddress\* 14 13232 1202.53 582.17 1261.0 1233.25  
## HotelPincode 15 13232 397430.26 259837.50 395003.0 388540.47  
## HotelDescription\* 16 13224 581.34 363.26 567.0 575.37  
## FreeWifi 17 13232 0.93 0.26 1.0 1.00  
## FreeBreakfast 18 13232 0.65 0.48 1.0 0.69  
## HotelCapacity 19 13232 62.51 76.66 34.0 46.03  
## HasSwimmingPool 20 13232 0.36 0.48 0.0 0.32  
## mad min max range skew  
## X 4904.44 1.0 13232 13231.0 0.00  
## CityName\* 11.86 1.0 42 41.0 0.48  
## Population 3846498.95 8096.0 12442373 12434277.0 0.68  
## CityRank 11.86 0.0 44 44.0 0.69  
## IsMetroCity 0.00 0.0 1 1.0 0.96  
## IsTouristDestination 0.00 0.0 1 1.0 -0.86  
## IsWeekend 0.00 0.0 1 1.0 -0.51  
## IsNewYearEve 0.00 0.0 1 1.0 2.28  
## Date\* 2.97 1.0 20 19.0 -1.05  
## HotelName\* 641.97 1.0 1670 1669.0 0.01  
## RoomRent 2653.85 299.0 322500 322201.0 16.75  
## StarRating 0.74 0.0 5 5.0 0.48  
## Airport 11.12 0.2 124 123.8 2.73  
## HotelAddress\* 668.65 1.0 2108 2107.0 -0.37  
## HotelPincode 257975.37 100025.0 7000157 6900132.0 9.99  
## HotelDescription\* 472.95 1.0 1226 1225.0 0.11  
## FreeWifi 0.00 0.0 1 1.0 -3.25  
## FreeBreakfast 0.00 0.0 1 1.0 -0.62  
## HotelCapacity 28.17 0.0 600 600.0 2.95  
## HasSwimmingPool 0.00 0.0 1 1.0 0.60  
## kurtosis se  
## X -1.20 33.21  
## CityName\* -0.88 0.10  
## Population -1.08 37019.65  
## CityRank -0.76 0.12  
## IsMetroCity -1.08 0.00  
## IsTouristDestination -1.26 0.00  
## IsWeekend -1.74 0.00  
## IsNewYearEve 3.18 0.00  
## Date\* 2.93 0.02  
## HotelName\* -1.25 4.24  
## RoomRent 582.06 63.75  
## StarRating 0.25 0.01  
## Airport 7.89 0.20  
## HotelAddress\* -0.88 5.06  
## HotelPincode 249.76 2258.86  
## HotelDescription\* -1.25 3.16  
## FreeWifi 8.57 0.00  
## FreeBreakfast -1.61 0.00  
## HotelCapacity 11.39 0.67  
## HasSwimmingPool -1.64 0.00

# Varible types  
str(caps)

## '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 ...

# View data set  
View(caps)  
  
# To organise the data having similar dates  
caps$Date<-gsub("18-Dec-16","Dec 18 2016",caps$Date)  
caps$Date<-gsub("21-Dec-16","Dec 21 2016",caps$Date)  
caps$Date<-gsub("24-Dec-16","Dec 24 2016",caps$Date)  
caps$Date<-gsub("25-Dec-16","Dec 25 2016",caps$Date)  
caps$Date<-gsub("29-Dec-16","Dec 28 2016",caps$Date)  
caps$Date<-gsub("31-Dec-16","Dec 31 2016",caps$Date)  
caps$Date<-gsub("4-Jan-17","Jan 4 2017",caps$Date)  
caps$Date<-gsub("8-Jan-17","Jan 8 2017",caps$Date)  
  
# Checking the changes  
table(caps$Date)

##   
## 28-Dec-16 4-Jan-16 8-Jan-16 Dec 18 2016 Dec 21 2016 Dec 24 2016   
## 44 31 31 1652 1655 1655   
## Dec 25 2016 Dec 28 2016 Dec 31 2016 Jan 04 2017 Jan 08 2017 Jan 4 2017   
## 1655 1611 1655 1548 1542 73   
## Jan 8 2017   
## 80

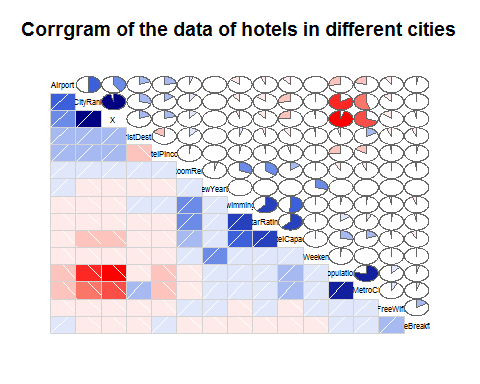
# Changing the dates to factors  
caps$Date<-factor(caps$Date)  
is.factor(caps$Date)

## [1] TRUE

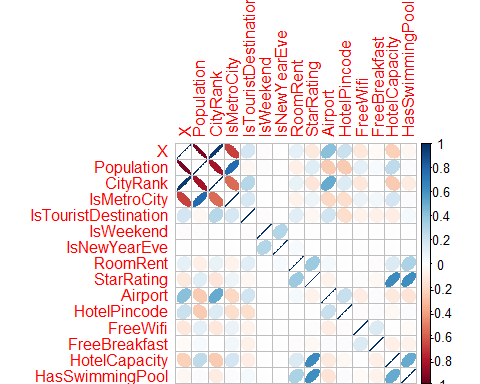
# Checking the labelling   
levels(caps$Date)

## [1] "28-Dec-16" "4-Jan-16" "8-Jan-16" "Dec 18 2016" "Dec 21 2016"  
## [6] "Dec 24 2016" "Dec 25 2016" "Dec 28 2016" "Dec 31 2016" "Jan 04 2017"  
## [11] "Jan 08 2017" "Jan 4 2017" "Jan 8 2017"

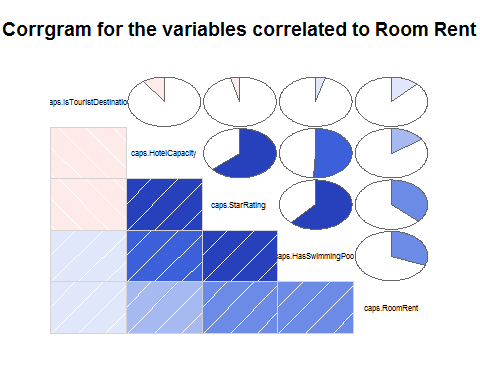
# To check the correlation between different variables of data set  
corrgram(caps,order = TRUE,upper.panel = panel.pie,lower.panel = panel.shade,  
 text.panel = panel.txt, main="Corrgram of the data of hotels in different cities")



# To check the correlation between the variables of the data set through corrplot  
corrplot(corr = cor(caps[,c(1,(3:8),(11:13),15,17,(18:20))],use = "complete.obs"),method = "ellipse")



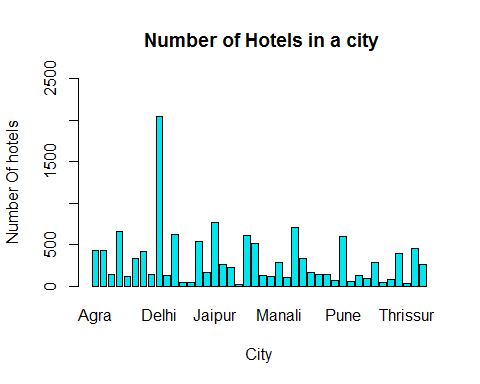
# ROOM RENT ANALYSIS  
# We have found that Room rent of the hotels depend on city rank, tourist destination, Star rating, Hotel Capacity and swimming pool  
  
# Corrgram for the variables correlated to room rent in the hotels  
  
# Data frame of the required variables to analyze  
rr<-data.frame(caps$RoomRent,caps$IsTouristDestination,caps$StarRating,caps$HotelCapacity,caps$HasSwimmingPool)  
corrgram(rr,order = TRUE,upper.panel = panel.pie,lower.panel = panel.shade,text.panel = panel.txt,  
 main="Corrgram for the variables correlated to Room Rent")



# We will analyze the distribution of data in Different variables  
  
# Bar Plots  
# Changing the labels   
tourist1<-factor(caps$IsTouristDestination, levels = c(0,1),labels = c("No","Yes"))  
  
# Hotels in different cities  
table1<-table(caps$CityName)  
table1

##   
## Agra Ahmedabad Amritsar Bangalore   
## 432 424 136 656   
## Bhubaneswar Chandigarh Chennai Darjeeling   
## 120 336 416 136   
## Delhi Gangtok Goa Guwahati   
## 2048 128 624 48   
## Haridwar Hyderabad Indore Jaipur   
## 48 536 160 768   
## Jaisalmer Jodhpur Kanpur Kochi   
## 264 224 16 608   
## Kolkata Lucknow Madurai Manali   
## 512 128 112 288   
## Mangalore Mumbai Munnar Mysore   
## 104 712 328 160   
## Nainital Ooty Panchkula Pune   
## 144 136 64 600   
## Puri Rajkot Rishikesh Shimla   
## 56 128 88 280   
## Srinagar Surat Thiruvanthipuram Thrissur   
## 40 80 392 32   
## Udaipur Varanasi   
## 456 264

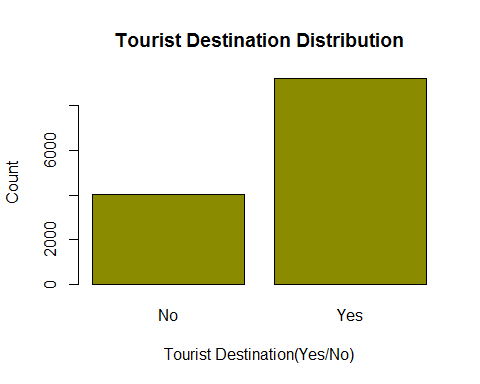
barplot(table1,main = " Number of Hotels in a city",xlab = "City",  
 ylab = "Number Of hotels",  
 col = "Turquoise2",ylim = c(0,2500))



# Visualtion of the number of tourist Destinations  
table2<-table(tourist1)  
table2

## tourist1  
## No Yes   
## 4007 9225

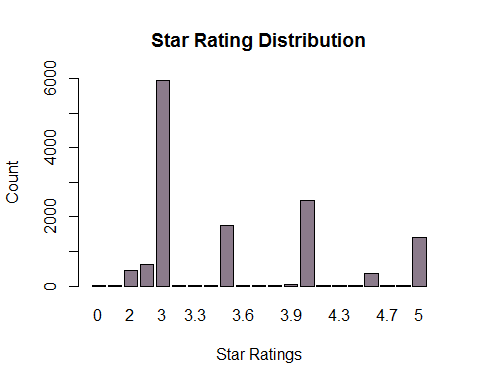
barplot(table2,main = "Tourist Destination Distribution",  
 xlab = "Tourist Destination(Yes/No)",  
 ylab = "Count",col = "Yellow4")



# Star Rating Distribution of the Hotels  
table3<-table(caps$StarRating)  
table3

##   
## 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

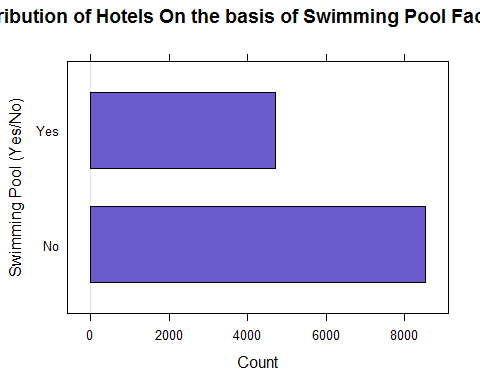
barplot(table3,main = "Star Rating Distribution",xlab = "Star Ratings",ylab = "Count",  
 col = "thistle4",ylim = c(0,6000))



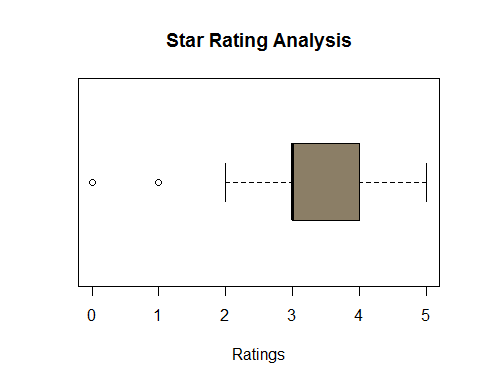
# Visualizing Distribution of hotels having swimmimg pool  
swim1<-factor(caps$HasSwimmingPool,levels = c(0,1),labels = c("No","Yes"))  
table4<-table(swim1)  
table4

## swim1  
## No Yes   
## 8524 4708

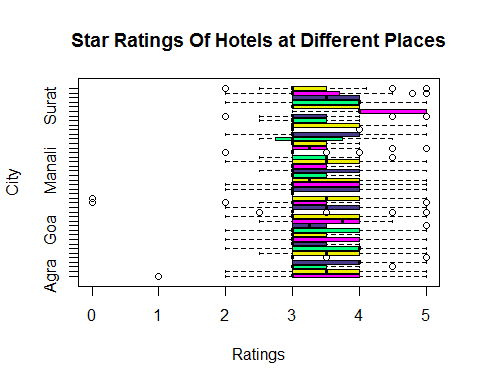
barchart(table4,main="Distribution of Hotels On the basis of Swimming Pool Facility",  
 ylab= "Swimming Pool (Yes/No)",  
 xlab= "Count",  
 col= "Slateblue")



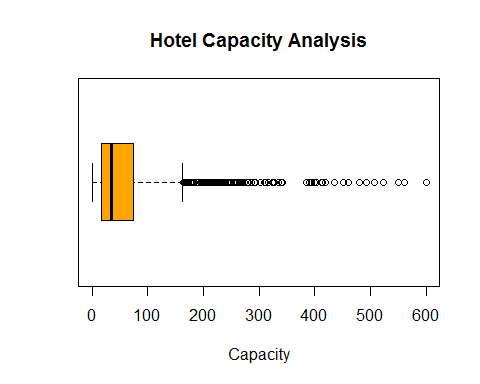
# BOXPLOTS  
  
# Star Rating  
boxplot(caps$StarRating,main="Star Rating Analysis",horizontal = TRUE,xlab="Ratings",  
 col = "wheat4")



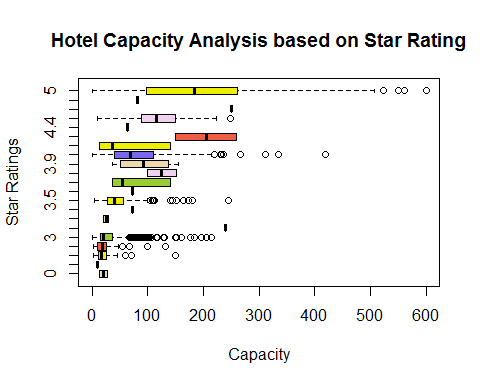
boxplot(caps$StarRating~caps$CityName,main="Star Ratings Of Hotels at Different Places",  
 xlab="Ratings",ylab="City",col= c("magenta","Yellow2","springgreen","slateblue4"),  
 horizontal= TRUE)



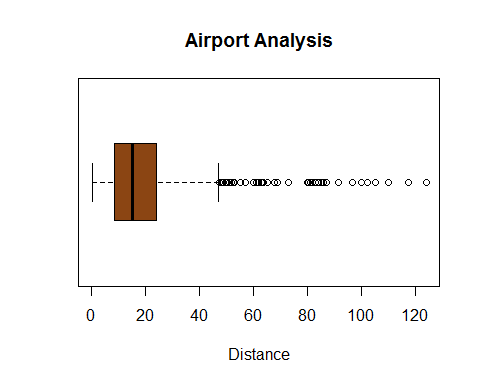
# Hotel Capacity  
boxplot(caps$HotelCapacity,main="Hotel Capacity Analysis",horizontal = TRUE,  
 xlab="Capacity",col = "Orange")



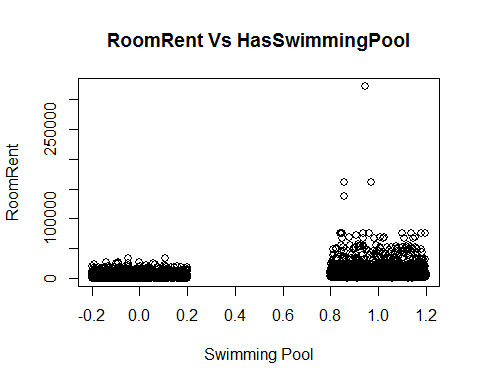
boxplot(caps$HotelCapacity~caps$StarRating,horizontal= TRUE,  
 main="Hotel Capacity Analysis based on Star Rating",  
 xlab="Capacity",ylab="Star Ratings",  
 col= c("wheat2","slateblue2","yellow2","tomato2","yellowgreen","thistle2"))



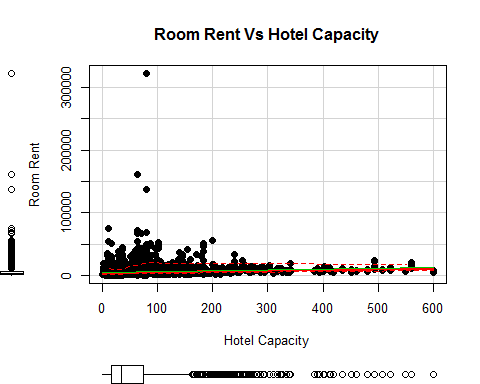
## The above boxplot shows that there are many outliers in the hotels having star rating 3 and 3.5  
  
# Airport  
boxplot(caps$Airport,main="Airport Analysis",horizontal = TRUE,  
 xlab="Distance",col = "Saddlebrown")



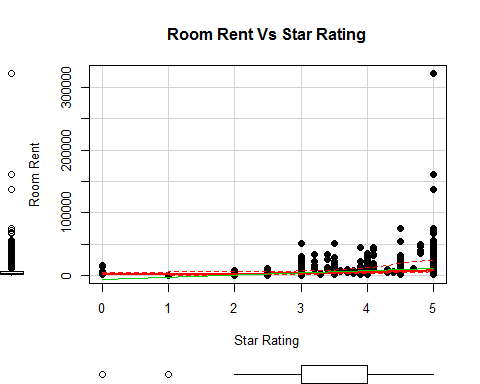
# RoomRent Vs HasSwimmingPool  
plot(jitter(caps$HasSwimmingPool),jitter(caps$RoomRent),main = "RoomRent Vs HasSwimmingPool",xlab = "Swimming Pool",  
 ylab = "RoomRent",cex= 1.1)



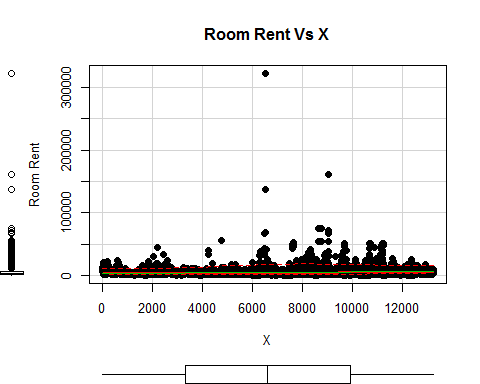
# RoomRent VS HotelCapacity  
scatterplot(caps$HotelCapacity,caps$RoomRent,main="Room Rent Vs Hotel Capacity",  
 xlab="Hotel Capacity",  
 ylab="Room Rent",cex=1.1,pch = 19)



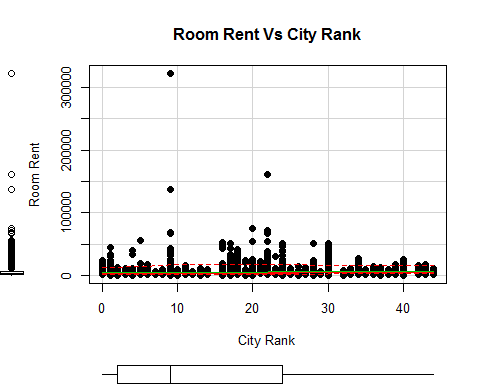
# There are many outliers in the Hotel capacity  
  
# RoomRent Vs StarRating  
scatterplot(caps$StarRating,caps$RoomRent,main="Room Rent Vs Star Rating",  
 xlab = "Star Rating",  
 ylab = "Room Rent",  
 cex = 1.1,pch = 19)



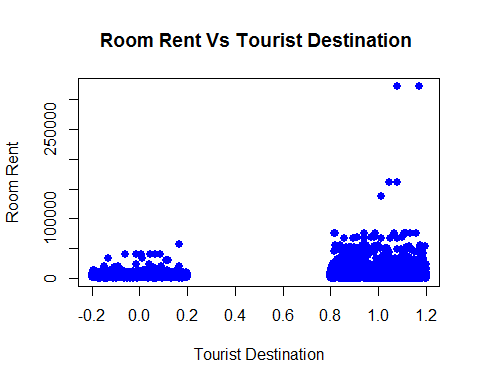
# RoomRent Vs X  
scatterplot(caps$X,caps$RoomRent,main="Room Rent Vs X",  
 xlab = "X",  
 ylab = "Room Rent",  
 cex = 1.1,pch = 19)



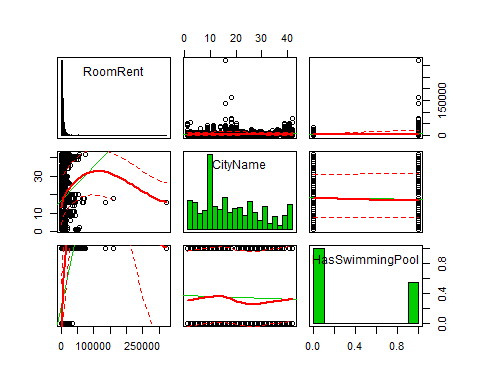
# RoomRent Vs City Rank  
scatterplot(caps$CityRank,caps$RoomRent,main="Room Rent Vs City Rank",  
 xlab = "City Rank",  
 ylab = "Room Rent",  
 cex = 1.1,pch = 19)



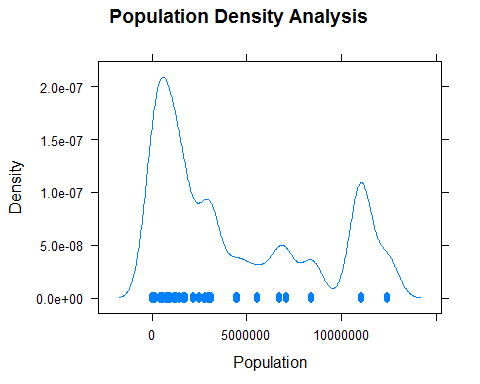
# RoomRent Vs Tourist Destination  
plot(jitter(caps$IsTouristDestination),jitter(caps$RoomRent),  
 main ="Room Rent Vs Tourist Destination ",  
 xlab = "Tourist Destination",  
 ylab = "Room Rent",  
 cex=1.1,pch= 19,col="Blue")



# Scatterplot Matrix  
scatterplotMatrix(~RoomRent+CityName+HasSwimmingPool,data = caps,diagonal= "histogram")



# Density Plots  
densityplot(caps$Population,main="Population Density Analysis",xlab="Population")



# Variance-covariance matrix  
  
# Correlation Matrix  
x<-caps[,c("HasSwimmingPool","HotelCapacity","StarRating","IsTouristDestination")]  
y<-caps[,c("RoomRent")]  
  
cor(x,y)

## [,1]  
## HasSwimmingPool 0.3116577  
## HotelCapacity 0.1578733  
## StarRating 0.3693734  
## IsTouristDestination 0.1225030

cov(x,y)

## [,1]  
## HasSwimmingPool 1094.2017  
## HotelCapacity 88753.4128  
## StarRating 2048.3755  
## IsTouristDestination 412.7803

var(x,y)

## [,1]  
## HasSwimmingPool 1094.2017  
## HotelCapacity 88753.4128  
## StarRating 2048.3755  
## IsTouristDestination 412.7803

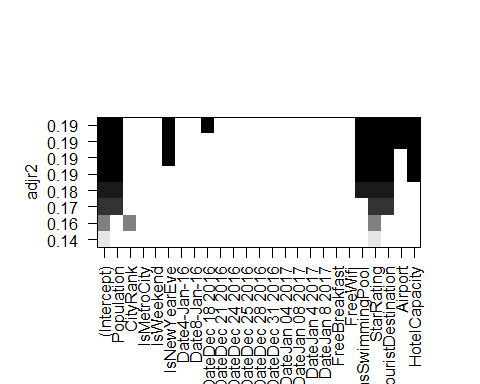
# Multivariable Regression Model   
  
# Regression Model for Room Rent  
  
#Model1  
model1<-RoomRent~Population+CityRank+IsMetroCity+IsWeekend+IsNewYearEve+Date+FreeBreakfast+FreeWifi+HasSwimmingPool+StarRating+IsTouristDestination+Airport+HotelCapacity  
fit1<-lm(model1,data=caps)  
  
#Summary  
summary(fit1)

##   
## Call:  
## lm(formula = model1, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11837 -2345 -692 1043 310128   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.342e+03 1.106e+03 -8.449 < 2e-16 \*\*\*  
## Population -1.220e-04 3.631e-05 -3.361 0.000779 \*\*\*  
## CityRank 9.437e-01 1.036e+01 0.091 0.927440   
## IsMetroCity -6.554e+02 2.196e+02 -2.984 0.002848 \*\*   
## IsWeekend 4.977e+02 4.419e+02 1.126 0.260083   
## IsNewYearEve 1.681e+03 1.418e+03 1.185 0.235875   
## Date4-Jan-16 1.281e+03 1.547e+03 0.828 0.407570   
## Date8-Jan-16 9.526e+02 1.609e+03 0.592 0.553790   
## DateDec 18 2016 -2.705e+02 1.115e+03 -0.243 0.808325   
## DateDec 21 2016 4.022e+02 1.020e+03 0.394 0.693320   
## DateDec 24 2016 3.785e+02 1.115e+03 0.339 0.734329   
## DateDec 25 2016 2.981e+02 1.115e+03 0.267 0.789192   
## DateDec 28 2016 9.304e+02 1.021e+03 0.911 0.362172   
## DateDec 31 2016 -6.388e+02 1.713e+03 -0.373 0.709278   
## DateJan 04 2017 1.022e+03 1.022e+03 1.000 0.317100   
## DateJan 08 2017 2.185e+02 1.109e+03 0.197 0.843842   
## DateJan 4 2017 6.316e+02 1.269e+03 0.498 0.618804   
## DateJan 8 2017 -2.306e+01 1.296e+03 -0.018 0.985806   
## FreeBreakfast 1.788e+02 1.234e+02 1.449 0.147324   
## FreeWifi 5.469e+02 2.242e+02 2.440 0.014708 \*   
## HasSwimmingPool 2.149e+03 1.617e+02 13.290 < 2e-16 \*\*\*  
## StarRating 3.591e+03 1.107e+02 32.433 < 2e-16 \*\*\*  
## IsTouristDestination 1.932e+03 1.507e+02 12.826 < 2e-16 \*\*\*  
## Airport 9.624e+00 3.175e+00 3.031 0.002444 \*\*   
## HotelCapacity -1.025e+01 1.033e+00 -9.917 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6598 on 13207 degrees of freedom  
## Multiple R-squared: 0.1919, Adjusted R-squared: 0.1904   
## F-statistic: 130.7 on 24 and 13207 DF, p-value: < 2.2e-16

# Checking the variables to determine the best fit model for determining the room Rent in hotels  
  
# Loadind Leaps package  
library(leaps)

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

leap1<-regsubsets(model1,data = caps,nbest = 1,really.big = T)  
# Visualizing it through graph  
plot(leap1,scale = "adjr2")



## From the plot it is clear that the Room Rent is depended upon Star Ratings, Availabilty of Swimming Pool,if the Destination is a Tourist Place or not, Hotel Capacity,and some famous places like Delhi,Jodhpur,Udaipur and Varanasi  
  
# REVISED MODEL TAKING THE BEST VARIABLES  
model2<-RoomRent~Population+CityRank+IsNewYearEve+HasSwimmingPool+StarRating+IsTouristDestination+Airport+HotelCapacity  
fit2<-lm(model2,data = caps)  
summary(fit2)

##   
## Call:  
## lm(formula = model2, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11798 -2337 -711 1035 309688   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.900e+03 3.929e+02 -20.107 < 2e-16 \*\*\*  
## Population -1.788e-04 2.869e-05 -6.231 4.77e-10 \*\*\*  
## CityRank -2.676e+00 1.020e+01 -0.262 0.793074   
## IsNewYearEve 8.424e+02 1.740e+02 4.842 1.30e-06 \*\*\*  
## HasSwimmingPool 2.134e+03 1.616e+02 13.207 < 2e-16 \*\*\*  
## StarRating 3.626e+03 1.105e+02 32.805 < 2e-16 \*\*\*  
## IsTouristDestination 1.777e+03 1.421e+02 12.503 < 2e-16 \*\*\*  
## Airport 1.061e+01 3.130e+00 3.390 0.000701 \*\*\*  
## HotelCapacity -1.065e+01 1.027e+00 -10.366 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6604 on 13223 degrees of freedom  
## Multiple R-squared: 0.1894, Adjusted R-squared: 0.189   
## F-statistic: 386.3 on 8 and 13223 DF, p-value: < 2.2e-16

# MODEL WITH THE THREE VARIABLES WHICH AFFECTS THE ROOM RENT THE MOST  
model3<-RoomRent~StarRating+HasSwimmingPool+HotelCapacity  
fit3<-lm(model3,data = caps)  
summary(fit3)

##   
## Call:  
## lm(formula = model3, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10804 -2295 -946 1002 310110   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6896.154 340.549 -20.25 <2e-16 \*\*\*  
## StarRating 3597.322 111.670 32.21 <2e-16 \*\*\*  
## HasSwimmingPool 2528.885 157.894 16.02 <2e-16 \*\*\*  
## HotelCapacity -15.558 1.006 -15.47 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6710 on 13228 degrees of freedom  
## Multiple R-squared: 0.1628, Adjusted R-squared: 0.1626   
## F-statistic: 857.5 on 3 and 13228 DF, p-value: < 2.2e-16

# Coeffecient Plot  
# Loading Package ggplot2 and coefplot  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.1

##   
## Attaching package: 'ggplot2'

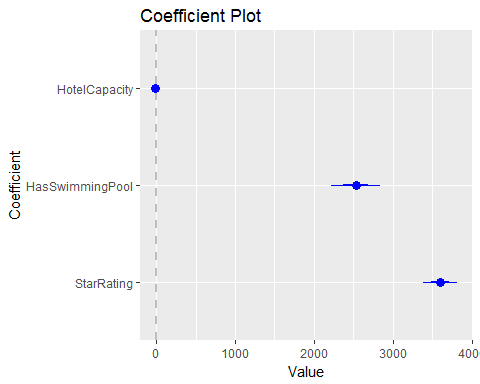
## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

library(coefplot)

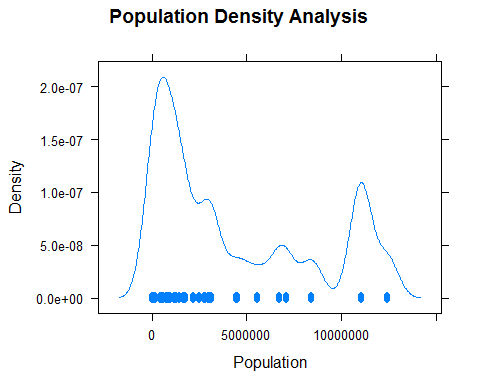
## Warning: package 'coefplot' was built under R version 3.4.1

coefplot(fit3,intercept= FALSE,outerCI=1.96,  
 coefficients= c("StarRating","HasSwimmingPool","HotelCapacity"))

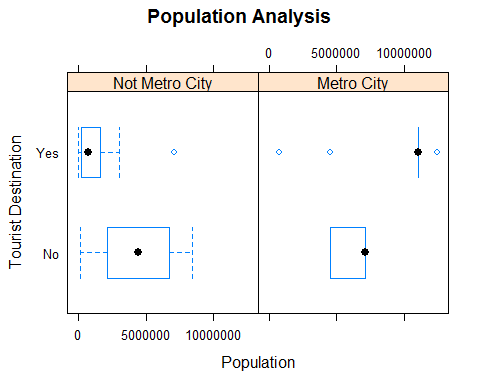
## Warning: Ignoring unknown aesthetics: xmin, xmax



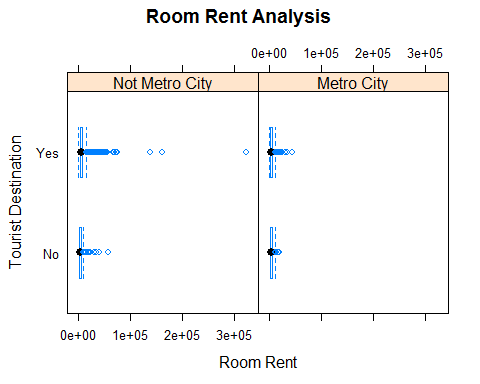
# Model:b0+b1\*StarRating+b2\*HasSwimmingPool+b3\*HotelCapacity  
# b0 = -6896.154, b1 = 3597.322, b2 = 2528.885,b3 = -15.558  
# Final Expression:-6896.154+3597.322\*StarRating+2528.885\*HasSwimmingPool+-15.558\*HotelCapacity  
  
  
# ANALYSIS OF STAR RATING OF HOTELS  
  
# Density Plots  
densityplot(caps$Population,main="Population Density Analysis",xlab="Population")



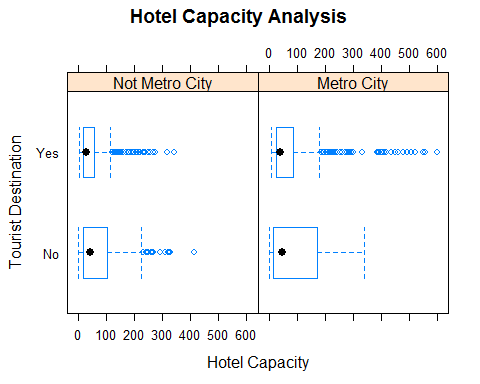
# Changing the labels   
Metro1<-factor(caps$IsMetroCity,levels = c(0,1),labels = c("Not Metro City","Metro City"))  
# BOXPLOTS  
  
# Population of Tourist and Non Tourist Places and Segregation if it is Metro City  
bwplot(tourist1~Population|Metro1,data = caps,xlab = "Population",  
 ylab = "Tourist Destination",horizontal = TRUE,main="Population Analysis")



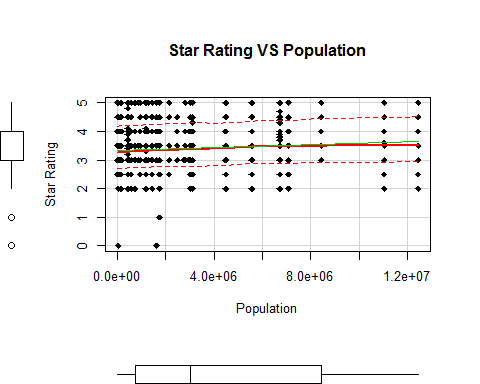
# Room Rent Of Hotels dividing on the basis of Tourist Destination and Metro Cities  
bwplot(tourist1~RoomRent|Metro1,data = caps,xlab = "Room Rent",ylab = "Tourist Destination",  
 main="Room Rent Analysis",horizontal = TRUE)



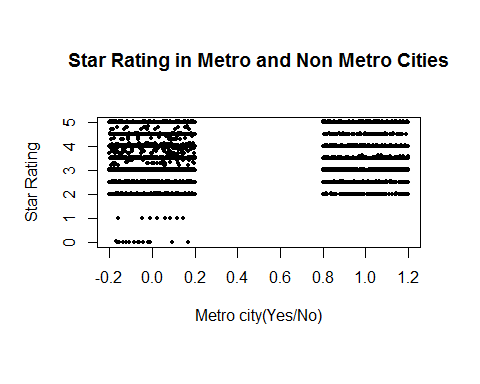
# Hotel Capacity dividing on the basis of Tourist Destination and Metro Cities  
bwplot(tourist1~HotelCapacity|Metro1,xlab = "Hotel Capacity",ylab = "Tourist Destination",  
 main="Hotel Capacity Analysis",horizontal = TRUE)



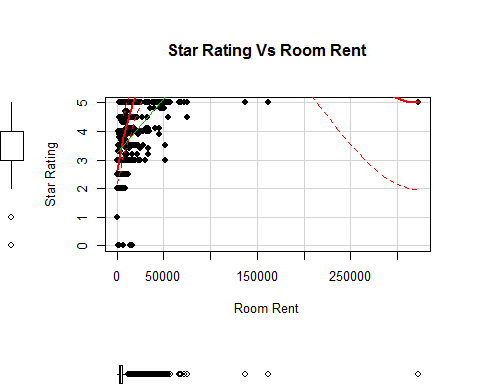
# ScatterPlots  
  
# Star Rating Vs Population  
scatterplot(caps$Population,caps$StarRating,main="Star Rating VS Population",  
 xlab = "Population",ylab = "Star Rating",cex = 1.1,pch = 18)



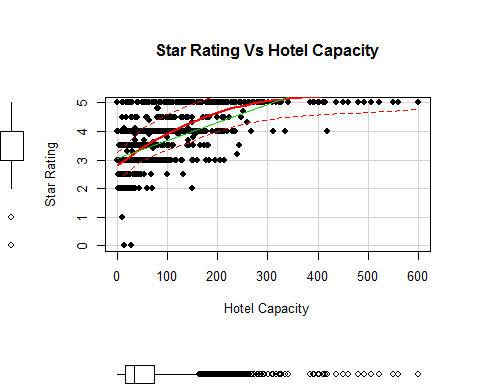
# Star Rating Vs IsMetroCity  
plot(jitter(caps$IsMetroCity),jitter(caps$StarRating),  
 main = "Star Rating in Metro and Non Metro Cities",xlab = "Metro city(Yes/No)",  
 ylab = "Star Rating",cex= 0.5,pch= 19)



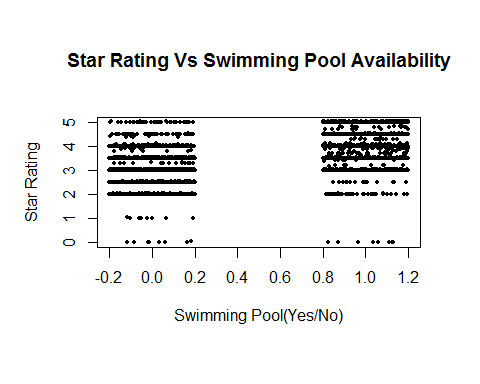
# Star Rating Vs Room Rent  
scatterplot(caps$RoomRent,caps$StarRating,main="Star Rating Vs Room Rent",  
 xlab = "Room Rent",ylab = "Star Rating",cex = 1.0,pch = 19)



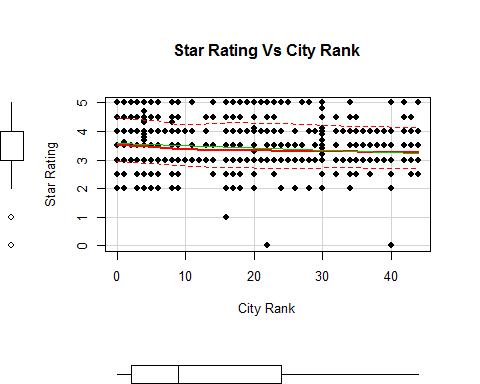
# Star Rating Vs Hotel Capacity  
scatterplot(caps$HotelCapacity,caps$StarRating,main="Star Rating Vs Hotel Capacity",  
 xlab = "Hotel Capacity",ylab = "Star Rating",cex = 1.0,pch = 19)



# Star Rating Vs Availability Of Swimming Pool  
plot(jitter(caps$HasSwimmingPool),jitter(caps$StarRating),  
 main = "Star Rating Vs Swimming Pool Availability",xlab = "Swimming Pool(Yes/No)",  
 ylab = "Star Rating",cex= 0.5,pch= 19)



# Star Rating Vs City Rank  
scatterplot(caps$CityRank,caps$StarRating,main="Star Rating Vs City Rank",xlab= "City Rank",  
 ylab = "Star Rating",cex = 1.0,pch = 19)



# T-Test  
  
# Test 1  
table6<-table(caps$IsMetroCity,caps$StarRating)  
table6

##   
## 0 1 2 2.5 3 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4  
## 0 16 8 344 456 4336 8 16 8 1312 0 24 16 32 1696  
## 1 0 0 96 176 1617 0 0 0 440 8 0 0 0 767  
##   
## 4.1 4.3 4.4 4.5 4.7 4.8 5  
## 0 24 16 8 288 8 16 840  
## 1 0 0 0 88 0 0 568

t.test(caps$StarRating,caps$IsMetroCity,var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: caps$StarRating and caps$IsMetroCity  
## t = 414.75, df = 26462, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 3.159770 3.189777  
## sample estimates:  
## mean of x mean of y   
## 3.4589329 0.2841596

# The above Test Shows that p-value<0.05.Thus it fails the null hypothesis and the two variables are dependent on each other  
  
# Test 2: Population and Metro city   
t.test(Population~IsMetroCity,var.equal= TRUE)

##   
## Two Sample t-test  
##   
## data: Population by IsMetroCity  
## t = -139.35, df = 13230, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -7383923 -7179082  
## sample estimates:  
## mean in group 0 mean in group 1   
## 2347728 9629230

# p-value<0.05  
# Fails Null Hypothesis.Thus,the two variables are related to each other  
  
# Test 3: Star Rating and Hotel Capacity  
t.test(caps$StarRating,caps$HotelCapacity,var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: caps$StarRating and caps$HotelCapacity  
## t = -88.602, df = 26462, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -60.35907 -57.74634  
## sample estimates:  
## mean of x mean of y   
## 3.458933 62.511638

# p-value<0.05  
# Fails Null Hypothesis. Thus, the two variables are dependent on each other  
  
# Test 4: Star Rating and Room Rent  
t.test(caps$StarRating,caps$RoomRent,var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: caps$StarRating and caps$RoomRent  
## t = -85.813, df = 26462, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -5595.485 -5345.581  
## sample estimates:  
## mean of x mean of y   
## 3.458933 5473.991838

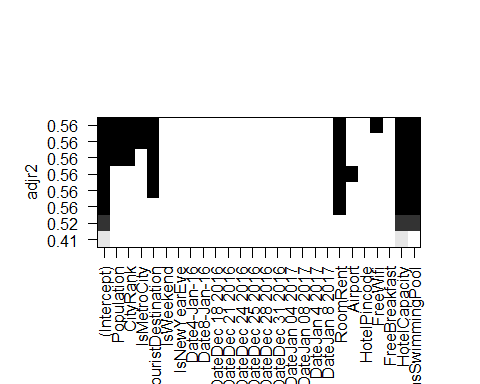
# p-value<0.05  
# Fails Null Hypothesis.Thus,the two variables are dependent on each other  
  
# Test 5: Star Rating and Availability of Swimming Pool  
t.test(caps$StarRating,caps$HasSwimmingPool,var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: caps$StarRating and caps$HasSwimmingPool  
## t = 398.81, df = 26462, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 3.087878 3.118380  
## sample estimates:  
## mean of x mean of y   
## 3.4589329 0.3558041

# p-value<0.05  
# Fails Null Hypothesis.Thus,the two variables are dependent on each other  
  
# Multi variable Regression model for determining Star Rating of a Hotel  
model3<-StarRating~Population+CityRank+IsMetroCity+IsTouristDestination+IsWeekend+IsNewYearEve+Date+  
RoomRent+Airport+HotelPincode+FreeWifi+FreeBreakfast+HotelCapacity+HasSwimmingPool  
fit3<-lm(model3,data = caps)  
summary(fit3)

##   
## Call:  
## lm(formula = model3, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.3158 -0.2169 -0.0241 0.2709 1.7253   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.700e+00 8.204e-02 32.906 < 2e-16 \*\*\*  
## Population 2.719e-08 2.888e-09 9.414 < 2e-16 \*\*\*  
## CityRank 5.829e-03 8.021e-04 7.267 3.89e-13 \*\*\*  
## IsMetroCity -7.095e-02 1.686e-02 -4.208 2.59e-05 \*\*\*  
## IsTouristDestination -9.533e-02 1.158e-02 -8.229 < 2e-16 \*\*\*  
## IsWeekend -2.540e-02 3.342e-02 -0.760 0.44732   
## IsNewYearEve 1.172e-01 1.073e-01 1.092 0.27466   
## Date4-Jan-16 -5.752e-02 1.170e-01 -0.492 0.62304   
## Date8-Jan-16 -3.559e-02 1.217e-01 -0.293 0.76990   
## DateDec 18 2016 1.084e-02 8.440e-02 0.128 0.89783   
## DateDec 21 2016 -1.849e-02 7.721e-02 -0.240 0.81071   
## DateDec 24 2016 -3.215e-03 8.440e-02 -0.038 0.96962   
## DateDec 25 2016 -1.918e-03 8.440e-02 -0.023 0.98187   
## DateDec 28 2016 -3.051e-02 7.729e-02 -0.395 0.69302   
## DateDec 31 2016 -1.329e-01 1.296e-01 -1.026 0.30503   
## DateJan 04 2017 -3.111e-02 7.734e-02 -0.402 0.68746   
## DateJan 08 2017 1.248e-03 8.396e-02 0.015 0.98814   
## DateJan 4 2017 -7.821e-02 9.602e-02 -0.815 0.41536   
## DateJan 8 2017 -7.924e-02 9.804e-02 -0.808 0.41896   
## RoomRent 2.054e-05 6.334e-07 32.430 < 2e-16 \*\*\*  
## Airport 6.832e-04 2.523e-04 2.708 0.00678 \*\*   
## HotelPincode -2.784e-09 1.876e-08 -0.148 0.88202   
## FreeWifi 7.055e-02 1.695e-02 4.162 3.17e-05 \*\*\*  
## FreeBreakfast 2.649e-02 9.332e-03 2.839 0.00454 \*\*   
## HotelCapacity 4.116e-03 6.979e-05 58.982 < 2e-16 \*\*\*  
## HasSwimmingPool 5.665e-01 1.129e-02 50.171 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.499 on 13206 degrees of freedom  
## Multiple R-squared: 0.5654, Adjusted R-squared: 0.5646   
## F-statistic: 687.3 on 25 and 13206 DF, p-value: < 2.2e-16

# Analyzing the best fit variables to determine the star Rating of a Hotel  
leap2<-regsubsets(model3,data=caps,nbest = 1,really.big = T)  
plot(leap2,scale = "adjr2")

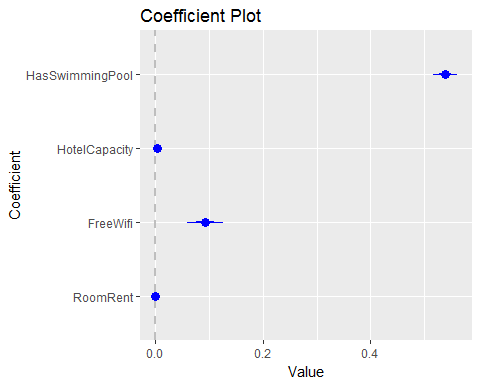


# Best Fit model for determining Star Rating of hotel  
model4<-StarRating~RoomRent+Airport+FreeWifi+HotelCapacity+HasSwimmingPool  
fit4<-lm(model4,data = caps)  
summary(fit4)

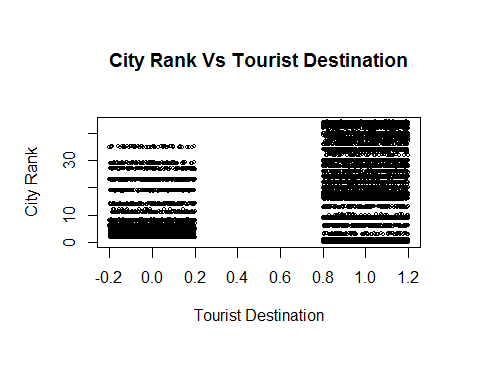
##   
## Call:  
## lm(formula = model4, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1571 -0.1953 -0.0179 0.2780 1.7868   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.780e+00 1.778e-02 156.313 < 2e-16 \*\*\*  
## RoomRent 1.983e-05 6.297e-07 31.494 < 2e-16 \*\*\*  
## Airport 1.076e-03 1.960e-04 5.490 4.10e-08 \*\*\*  
## FreeWifi 9.348e-02 1.675e-02 5.581 2.44e-08 \*\*\*  
## HotelCapacity 4.317e-03 6.625e-05 65.168 < 2e-16 \*\*\*  
## HasSwimmingPool 5.384e-01 1.109e-02 48.528 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5021 on 13226 degrees of freedom  
## Multiple R-squared: 0.5593, Adjusted R-squared: 0.5592   
## F-statistic: 3358 on 5 and 13226 DF, p-value: < 2.2e-16

# Model:b0+b1\*RoomRent+b2\*Airport+b3\*FreeWifi+b4\*HotelCapacity+b5\*HasSwimmingPool  
# b0 = 2.780e+00, b1 = 1.983e-05, b2 = 1.076e-03, b3 = 9.348e-02, b4 = 4.317e-03, b5 = 5.384e-01  
  
# Coeffecient plot of the Regression Model of Star Rating Of the Hotels  
coefplot(fit4,intercept=FALSE,OuterCI=1.96,  
 coefficients= c("CityName","RoomRent","HotelCapacity","HasSwimmingPool","FreeWifi","FreeBreakfast"))

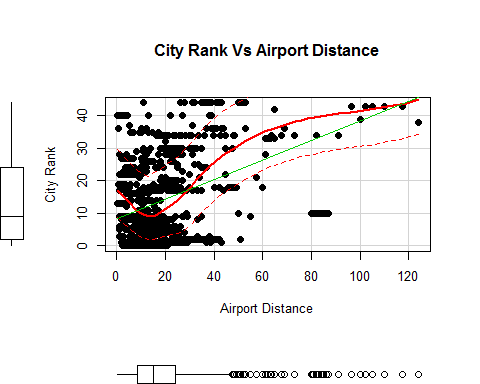
## Warning: Ignoring unknown aesthetics: xmin, xmax



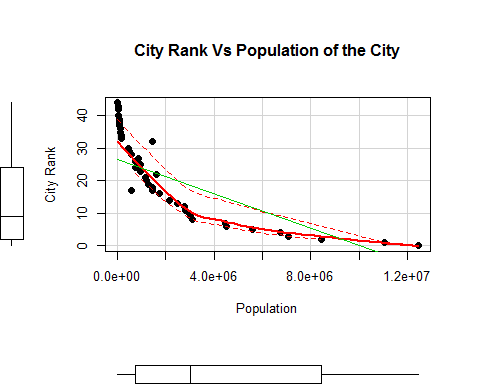
# Analysis Of City Rank  
  
# Scatterplots  
  
# City Rank Vs Tourist Destination  
plot(jitter(caps$IsTouristDestination),jitter(caps$CityRank),main = "City Rank Vs Tourist Destination",  
 xlab = "Tourist Destination",ylab = "City Rank",cex= 0.5)



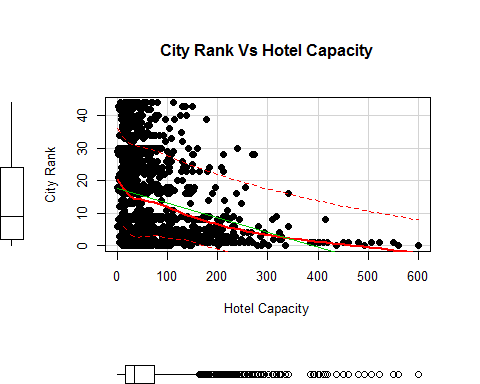
# City Rank Vs Airport  
scatterplot(caps$Airport,caps$CityRank,main="City Rank Vs Airport Distance",  
 xlab = "Airport Distance",ylab = "City Rank",cex = 1.1,pch = 19)



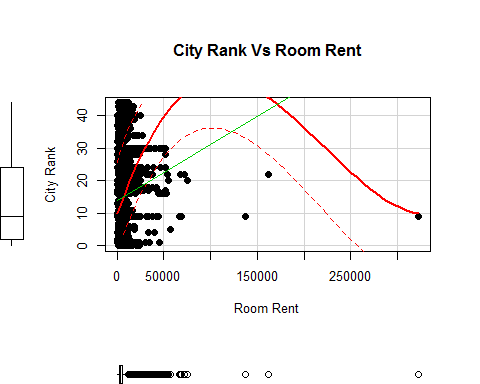
# City Rank Vs Population  
scatterplot(caps$Population,caps$CityRank,main="City Rank Vs Population of the City",  
 xlab = "Population",ylab = "City Rank",cex = 1.1,pch = 19)



# City Rank Vs Hotel Capacity  
scatterplot(caps$HotelCapacity,caps$CityRank,main="City Rank Vs Hotel Capacity",  
 xlab="Hotel Capacity",ylab = "City Rank",cex = 1.1,pch = 19)



# City Rank Vs Room Rent  
scatterplot(caps$RoomRent,caps$CityRank,main="City Rank Vs Room Rent",  
 xlab="Room Rent",ylab="City Rank",cex = 1.1,pch = 19)



# T-Test  
  
# Test1: City Rank Vs Free breakfast  
t.test(CityRank~FreeBreakfast,var.equal= TRUE)

##   
## Two Sample t-test  
##   
## data: CityRank by FreeBreakfast  
## t = 0.99886, df = 13230, p-value = 0.3179  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2366077 0.7283189  
## sample estimates:  
## mean in group 0 mean in group 1   
## 14.99332 14.74747

# p>0.05  
# The test doesn't fails Null Hypothesis. Thus,the two variables are independent of each other.  
  
# Test2: City Rank Vs Free Wifi  
t.test(CityRank~IsTouristDestination,var.equal= TRUE)

##   
## Two Sample t-test  
##   
## data: CityRank by IsTouristDestination  
## t = -33.641, df = 13230, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -8.735927 -7.773948  
## sample estimates:  
## mean in group 0 mean in group 1   
## 9.078612 17.333550

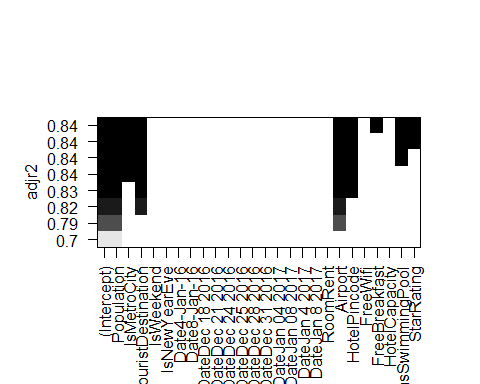
# p-value<0.05  
# Fails Null Hypothesis.Thus,the two variables are dependent on each other  
  
# Test3: City Rank Vs IsWeekend  
t.test(CityRank~IsWeekend,var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: CityRank by IsWeekend  
## t = 0.83467, df = 13230, p-value = 0.4039  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2727733 0.6773636  
## sample estimates:  
## mean in group 0 mean in group 1   
## 14.95973 14.75743

# p-value>0.05  
# Test doesn't fails Null Hypothesis.Thus,the two variables are independent of each other.  
  
# Multi variable regression model for determining City Rank  
model5<-CityRank~Population+IsMetroCity+IsTouristDestination+IsWeekend+IsNewYearEve+Date+  
 RoomRent+Airport+HotelPincode+FreeWifi+FreeBreakfast+HotelCapacity+HasSwimmingPool+StarRating  
fit5<-lm(model5,data = caps)  
summary(fit5)

##   
## Call:  
## lm(formula = model5, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.1404 -3.7519 -0.0256 2.8670 24.2353   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.156e+01 9.047e-01 23.829 < 2e-16 \*\*\*  
## Population -2.829e-06 1.945e-08 -145.455 < 2e-16 \*\*\*  
## IsMetroCity 4.026e+00 1.793e-01 22.457 < 2e-16 \*\*\*  
## IsTouristDestination 4.496e+00 1.195e-01 37.622 < 2e-16 \*\*\*  
## IsWeekend 2.435e+00 3.612e-01 6.741 1.64e-11 \*\*\*  
## IsNewYearEve -1.544e+00 1.162e+00 -1.329 0.18383   
## Date4-Jan-16 2.550e-01 1.267e+00 0.201 0.84052   
## Date8-Jan-16 -2.180e+00 1.317e+00 -1.655 0.09798 .   
## DateDec 18 2016 -4.421e+00 9.129e-01 -4.842 1.30e-06 \*\*\*  
## DateDec 21 2016 -2.063e+00 8.358e-01 -2.468 0.01359 \*   
## DateDec 24 2016 -4.425e+00 9.129e-01 -4.847 1.26e-06 \*\*\*  
## DateDec 25 2016 -4.424e+00 9.129e-01 -4.846 1.27e-06 \*\*\*  
## DateDec 28 2016 -2.108e+00 8.366e-01 -2.519 0.01177 \*   
## DateDec 31 2016 -2.893e+00 1.403e+00 -2.062 0.03919 \*   
## DateJan 04 2017 -2.122e+00 8.371e-01 -2.535 0.01124 \*   
## DateJan 08 2017 -4.393e+00 9.083e-01 -4.837 1.33e-06 \*\*\*  
## DateJan 4 2017 -1.258e+00 1.040e+00 -1.210 0.22626   
## DateJan 8 2017 -3.070e+00 1.061e+00 -2.893 0.00382 \*\*   
## RoomRent -4.657e-07 7.125e-06 -0.065 0.94789   
## Airport 1.703e-01 2.296e-03 74.172 < 2e-16 \*\*\*  
## HotelPincode -5.175e-06 1.980e-07 -26.135 < 2e-16 \*\*\*  
## FreeWifi -1.894e-01 1.836e-01 -1.032 0.30223   
## FreeBreakfast 6.160e-01 1.009e-01 6.104 1.06e-09 \*\*\*  
## HotelCapacity 1.624e-03 8.492e-04 1.912 0.05587 .   
## HasSwimmingPool -2.084e+00 1.322e-01 -15.768 < 2e-16 \*\*\*  
## StarRating 6.833e-01 9.403e-02 7.267 3.89e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.403 on 13206 degrees of freedom  
## Multiple R-squared: 0.8404, Adjusted R-squared: 0.8401   
## F-statistic: 2782 on 25 and 13206 DF, p-value: < 2.2e-16

# Determining Best Fit Regression Model for determing City Rank  
leap3<-regsubsets(model5,data = caps,nbest = 1,really.big = T)  
plot(leap3,scale = "adjr2")

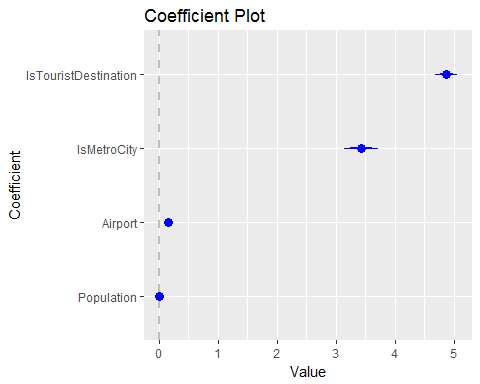


# Fit Regression Model for determing City Rank  
model6<-CityRank~Population+Airport+IsMetroCity+IsTouristDestination  
fit6<-lm(model6,data = caps)  
summary(fit6)

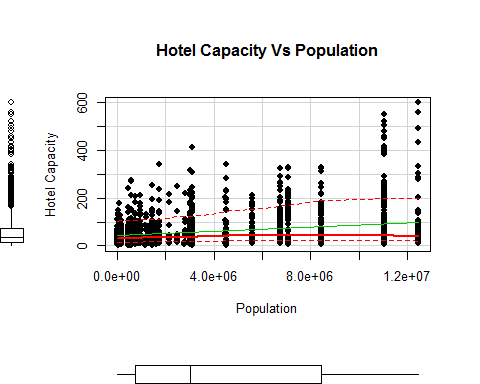
##   
## Call:  
## lm(formula = model6, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0753 -4.0773 0.2561 2.4616 18.5465   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.878e+01 1.202e-01 156.31 <2e-16 \*\*\*  
## Population -2.676e-06 1.894e-08 -141.30 <2e-16 \*\*\*  
## Airport 1.653e-01 2.270e-03 72.81 <2e-16 \*\*\*  
## IsMetroCity 3.431e+00 1.812e-01 18.93 <2e-16 \*\*\*  
## IsTouristDestination 4.874e+00 1.154e-01 42.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.618 on 13227 degrees of freedom  
## Multiple R-squared: 0.8272, Adjusted R-squared: 0.8272   
## F-statistic: 1.583e+04 on 4 and 13227 DF, p-value: < 2.2e-16

# Model:b0+b1\*Population+b2\*Airport+b3\*IsMetroCity+b4\*IsTouristDestination  
# b0 = 1.878e+01, b1 =-2676e-06 , b2 = 1.653e-01, b3 = 3.431e+00,b4 = 4.874e+00  
# Final Expression:1.878e+01+-2676e-06\*Population+1.653e-01\*Airport+3.431e+00\*IsMetroCity+4.874e+00\*IsTouristDestination  
  
# Coefficient plot  
coefplot(fit6,outerCI = 1.6,  
 coefficients = c("Population","Airport","IsMetroCity","IsTouristDestination"))

## Warning: Ignoring unknown aesthetics: xmin, xmax

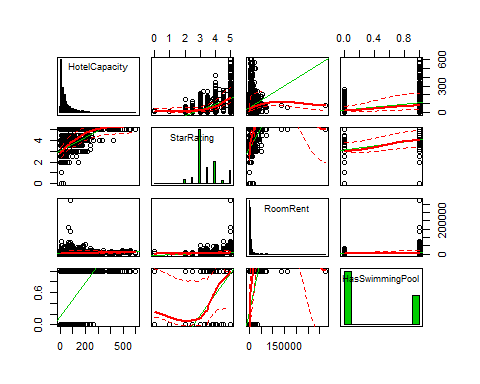


# Analysis Of Hotel Capacity  
  
# Scatterplots  
  
# Hotel Capacity Vs Population  
scatterplot(caps$Population,caps$HotelCapacity,main="Hotel Capacity Vs Population",  
 xlab="Population",ylab = "Hotel Capacity",cex = 1.0,pch = 19)



# Scatterplot Matrix  
scatterplotMatrix(~HotelCapacity+StarRating+RoomRent+HasSwimmingPool,data = caps,diagonal="histogram")

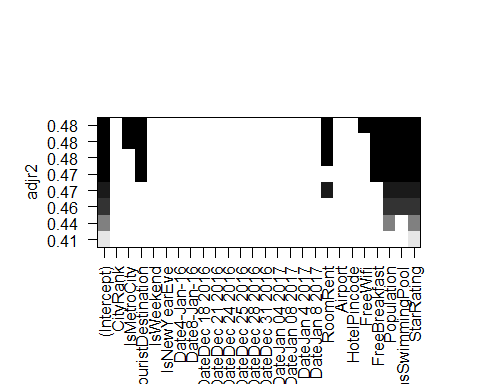
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth



# Multi Variable model for Determining the Hotel Capacity of the Hotels in a city  
  
model7<-HotelCapacity~CityRank+IsMetroCity+IsTouristDestination+IsWeekend+IsNewYearEve+Date+  
 RoomRent+Airport+HotelPincode+FreeWifi+FreeBreakfast+Population+HasSwimmingPool+StarRating  
fit7<-lm(model7,data = caps)  
summary(fit7)

##   
## Call:  
## lm(formula = model7, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -169.70 -27.59 -7.26 15.18 409.44   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.327e+02 9.395e+00 -14.120 < 2e-16 \*\*\*  
## CityRank 1.705e-01 8.914e-02 1.912 0.055873 .   
## IsMetroCity 6.279e+00 1.870e+00 3.357 0.000791 \*\*\*  
## IsTouristDestination -1.503e+01 1.282e+00 -11.726 < 2e-16 \*\*\*  
## IsWeekend -3.156e+00 3.707e+00 -0.851 0.394631   
## IsNewYearEve -2.266e+00 1.190e+01 -0.190 0.849008   
## Date4-Jan-16 -2.531e-01 1.298e+01 -0.019 0.984444   
## Date8-Jan-16 3.024e+00 1.350e+01 0.224 0.822706   
## DateDec 18 2016 1.994e+01 9.360e+00 2.131 0.033140 \*   
## DateDec 21 2016 1.695e+01 8.563e+00 1.980 0.047756 \*   
## DateDec 24 2016 2.035e+01 9.360e+00 2.174 0.029717 \*   
## DateDec 25 2016 2.026e+01 9.360e+00 2.165 0.030405 \*   
## DateDec 28 2016 1.777e+01 8.572e+00 2.073 0.038190 \*   
## DateDec 31 2016 2.301e+01 1.437e+01 1.601 0.109396   
## DateJan 04 2017 1.727e+01 8.577e+00 2.014 0.044060 \*   
## DateJan 08 2017 1.991e+01 9.312e+00 2.138 0.032505 \*   
## DateJan 4 2017 2.688e+01 1.065e+01 2.524 0.011601 \*   
## DateJan 8 2017 2.828e+01 1.087e+01 2.601 0.009317 \*\*   
## RoomRent -7.212e-04 7.273e-05 -9.916 < 2e-16 \*\*\*  
## Airport 1.425e-02 2.800e-02 0.509 0.610629   
## HotelPincode 7.905e-08 2.081e-06 0.038 0.969693   
## FreeWifi -7.611e+00 1.880e+00 -4.048 5.2e-05 \*\*\*  
## FreeBreakfast -1.111e+01 1.031e+00 -10.776 < 2e-16 \*\*\*  
## Population 3.265e-06 3.201e-07 10.199 < 2e-16 \*\*\*  
## HasSwimmingPool 3.497e+01 1.332e+00 26.246 < 2e-16 \*\*\*  
## StarRating 5.065e+01 8.588e-01 58.982 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 55.35 on 13206 degrees of freedom  
## Multiple R-squared: 0.4797, Adjusted R-squared: 0.4787   
## F-statistic: 487 on 25 and 13206 DF, p-value: < 2.2e-16

# Determining the best fit model for estimating Hotel Capacity  
leap4<-regsubsets(model7,data = caps,nbest = 1,really.big = T)  
plot(leap4,scale = "adjr2")



# Best fit model to determine the Hotel Capacity of the Hotel in a city  
model8<-HotelCapacity~Population+IsMetroCity+IsTouristDestination+RoomRent\*StarRating+HasSwimmingPool\*StarRating+StarRating  
fit8<-lm(model8,data = caps)  
summary(fit8)

##   
## Call:  
## lm(formula = model8, data = caps)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -190.60 -25.72 -5.91 13.11 399.31   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.504e+01 4.031e+00 -13.656 < 2e-16 \*\*\*  
## Population 2.599e-06 1.834e-07 14.174 < 2e-16 \*\*\*  
## IsMetroCity 4.083e+00 1.740e+00 2.346 0.01899 \*   
## IsTouristDestination -1.156e+01 1.105e+00 -10.465 < 2e-16 \*\*\*  
## RoomRent 1.545e-03 4.954e-04 3.120 0.00181 \*\*   
## StarRating 2.721e+01 1.217e+00 22.365 < 2e-16 \*\*\*  
## HasSwimmingPool -1.358e+02 6.215e+00 -21.847 < 2e-16 \*\*\*  
## RoomRent:StarRating -5.282e-04 1.052e-04 -5.022 5.17e-07 \*\*\*  
## StarRating:HasSwimmingPool 4.798e+01 1.700e+00 28.231 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 54.02 on 13223 degrees of freedom  
## Multiple R-squared: 0.5037, Adjusted R-squared: 0.5034   
## F-statistic: 1678 on 8 and 13223 DF, p-value: < 2.2e-16

# Model:Hotel Capacity=b0+b1\*Population+b2\*IsMetroCity+b3\*IsTouristDestination+b4\*RoomRent\*StarRating+b5\*HasSwimmingPool\*StarRating+b6\*StarRating  
  
# Coeffecient Plot  
coefplot(fit8,intercept= FALSE,OuterCI=1.6,  
 coeffecients= c("Population","IsMetroCity","IsTouristDestination","RoomRent\*StarRating","HasSwimmingPool\*StarRating","StarRating"))

## Warning: Ignoring unknown aesthetics: xmin, xmax

