Sai\_Gopi\_krishna\_Govindarajula\_My\_Project\_Code.R

HansZimmer

Wed Jul 12 23:57:32 2017

# Project Title: Hotel Room Pricing in Indian Market  
# NAME: Sai Gopi Krishna Govindarajula  
# EMAIL: saigopikrishna.g@gmail.com  
# COLLEGE: BITS PILANI  
  
##setting the directory and assigning a variabel to the data frame  
setwd("E:/Studies/Sai Gopi Krishna Govindarajula/Udemy/project/Hotel Analysis/Sai Gopi Krishna Govindarajula Project")  
  
#Reading the dataset and creating a data frame  
hotelData.df<-read.csv(paste("Cities42.csv",sep = ""))  
  
#Viewing the data  
View(hotelData.df)  
  
#Adjusting the dates for consistency using gsub command  
  
hotelData.df$Date<-gsub("Jan 4 2017", "Jan 04 2017", hotelData.df$Date)  
hotelData.df$Date<-gsub("Jan 8 2017", "Jan 08 2017", hotelData.df$Date)  
hotelData.df$Date<-gsub("4-Jan-17", "Jan 04 2017", hotelData.df$Date)  
hotelData.df$Date<-gsub("4-Jan-16", "Jan 04 2017", hotelData.df$Date)  
hotelData.df$Date<-gsub("8-Jan-16", "Jan 08 2017", hotelData.df$Date)  
hotelData.df$Date<-gsub("8-Jan-17", "Jan 08 2017", hotelData.df$Date)  
hotelData.df$Date<-gsub("18-Dec-16", "Dec 18 2016", hotelData.df$Date)  
hotelData.df$Date<-gsub("21-Dec-16", "Dec 21 2016", hotelData.df$Date)  
hotelData.df$Date<-gsub("24-Dec-16", "Dec 24 2016", hotelData.df$Date)  
hotelData.df$Date<-gsub("25-Dec-16", "Dec 25 2016", hotelData.df$Date)  
hotelData.df$Date<-gsub("28-Dec-16", "Dec 28 2016", hotelData.df$Date)  
hotelData.df$Date<-gsub("31-Dec-16", "Dec 31 2016", hotelData.df$Date)  
  
#Changing dates to factors for labelling   
  
hotelData.df$Date<-factor(hotelData.df$Date)  
is.factor(hotelData.df$Date)

## [1] TRUE

#Checking the labelling  
levels(hotelData.df$Date)

## [1] "Dec 18 2016" "Dec 21 2016" "Dec 24 2016" "Dec 25 2016" "Dec 28 2016"  
## [6] "Dec 31 2016" "Jan 04 2017" "Jan 08 2017"

#Analyzing the summary of the data and describing the variables  
  
library(psych)

## Warning: package 'psych' was built under R version 3.3.3

describe(hotelData.df)

## 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 4.50 2.29 4.0 4.50  
## 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 8 7.0 0.00  
## 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\* -1.24 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

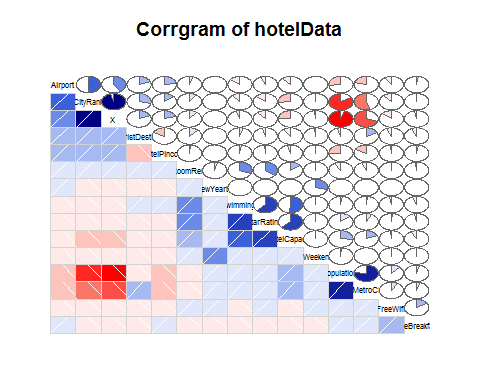
summary(hotelData.df)

## 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 2016:1655 Vivanta by Taj : 32 Min. : 299   
## Dec 24 2016:1655 Goldfinch Hotel : 24 1st Qu.: 2436   
## Dec 25 2016:1655 OYO Rooms : 24 Median : 4000   
## Dec 28 2016:1655 The Gordon House Hotel: 24 Mean : 5474   
## Dec 31 2016:1655 Apnayt Villa : 16 3rd Qu.: 6299   
## Jan 08 2017:1653 Bentleys Hotel Colaba : 16 Max. :322500   
## (Other) :3304 (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   
##

#Identifying the most relevent predictor variables by correlation corrgram  
#Taking Y = RoomRent  
library(corrgram)

## Warning: package 'corrgram' was built under R version 3.3.3

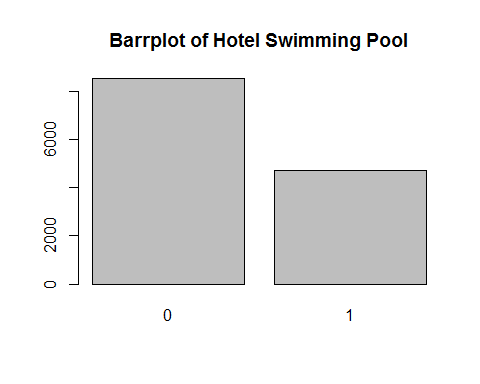
corrgram(hotelData.df, order=TRUE, lower.panel=panel.shade,  
 upper.panel=panel.pie, text.panel=panel.txt,  
 main="Corrgram of hotelData")



##We can see that HasSwimming, StarRating, HotelCapacity are very well   
 #correlated to RoomRent. Hence considering them as predictor variables  
  
  
##Visualizing data for Y as Room rent and X1,X2,X3 as HasSwimmingPool, StarRating and HotelCapacity respectively  
  
#Table for HasSwimmingPool  
table(hotelData.df$HasSwimmingPool)

##   
## 0 1   
## 8524 4708

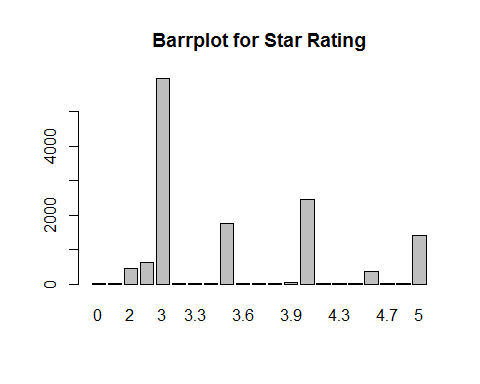
Swim<-table(hotelData.df$HasSwimmingPool)  
barplot(Swim,main="Barrplot of Hotel Swimming Pool")



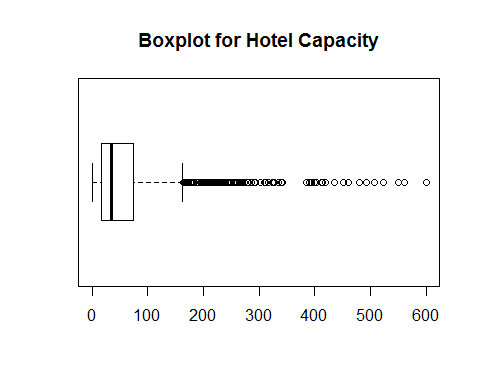
#Table for StarRating  
table(hotelData.df$StarRating)

##   
## 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

starRating<-table(hotelData.df$StarRating)  
barplot(starRating,main = "Barrplot for Star Rating")



#BoxPlot for HotelCapacity  
boxplot(hotelData.df$HotelCapacity, main="Boxplot for Hotel Capacity",horizontal = TRUE)



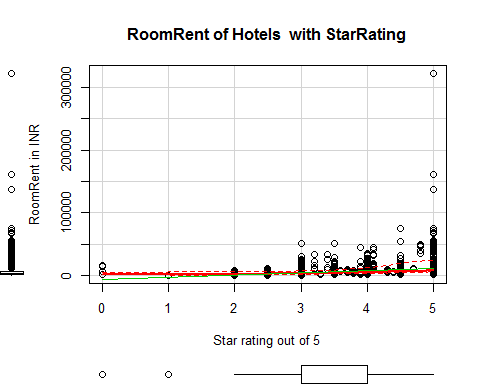
#Scatterplot pair wise for predictor variable  
  
library(car)

## Warning: package 'car' was built under R version 3.3.3

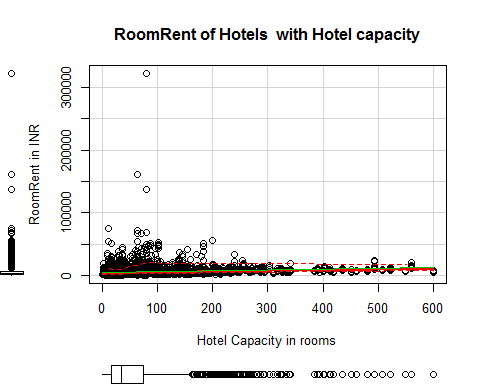
##   
## Attaching package: 'car'

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

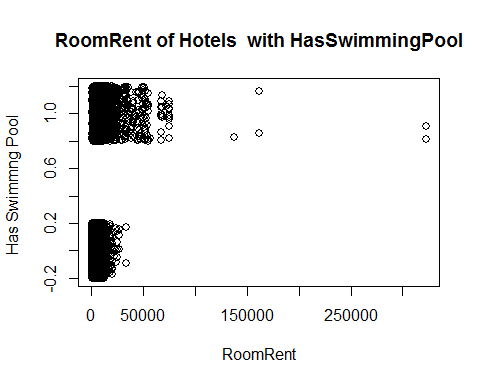
#StarRating Vs RoomRent  
  
scatterplot(hotelData.df$StarRating,hotelData.df$RoomRent,main="RoomRent of Hotels with StarRating",ylab = "RoomRent in INR", xlab="Star rating out of 5",cex=1.1)



#RoomRent Vs HotelCapacity  
  
scatterplot(hotelData.df$HotelCapacity,hotelData.df$RoomRent,main="RoomRent of Hotels with Hotel capacity",ylab ="RoomRent in INR",xlab="Hotel Capacity in rooms",cex=1.1)



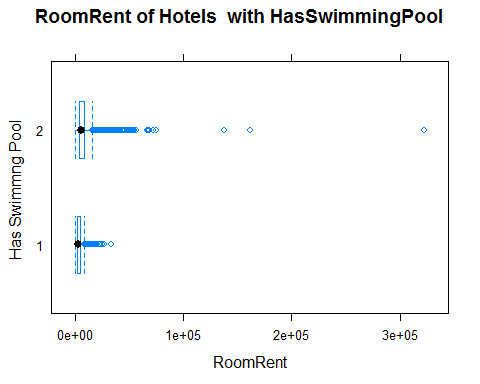
#RoomRent Vs HasSwimmingPool  
  
plot(jitter(hotelData.df$RoomRent),jitter(hotelData.df$HasSwimmingPool),main="RoomRent of Hotels with HasSwimmingPool",ylab = "Has Swimmng Pool ", xlab="RoomRent",cex=1.1)



library(lattice)

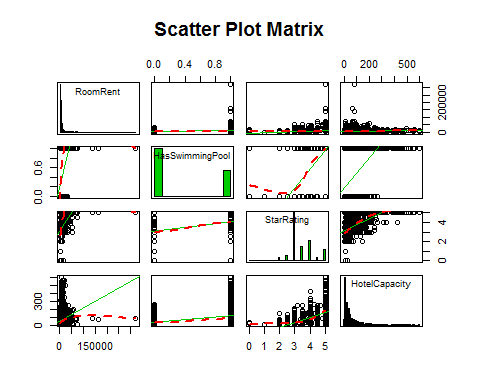
## Warning: package 'lattice' was built under R version 3.3.3

bwplot(HasSwimmingPool~RoomRent, data = hotelData.df,main="RoomRent of Hotels with HasSwimmingPool",ylab = "Has Swimmng Pool ", xlab="RoomRent" )

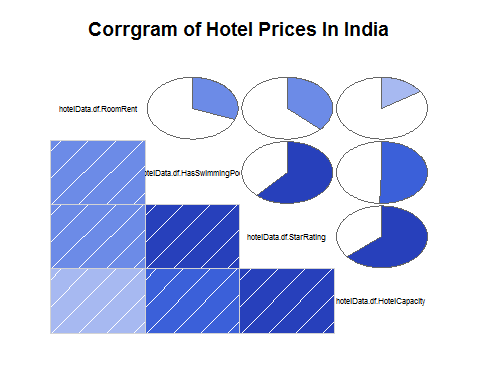


#Scatterplot matrix  
  
scatterplotMatrix(  
 hotelData.df[  
 ,c("RoomRent","HasSwimmingPool","StarRating", "HotelCapacity")],   
 spread=FALSE, smoother.args=list(lty=2),  
 main="Scatter Plot Matrix", diagonal = "histogram")

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



#Corrgram of Y, x1, x2, x3  
  
library(corrgram)  
  
xyz<-data.frame(hotelData.df$RoomRent, hotelData.df$HasSwimmingPool, hotelData.df$HotelCapacity, hotelData.df$StarRating)  
corrgram(xyz, order=TRUE, lower.panel=panel.shade,  
 upper.panel=panel.pie, text.panel=panel.txt,  
 main="Corrgram of Hotel Prices In India")



#Variance-Covariance Matrix for Y, x1, x2, x3  
  
x<-hotelData.df[,c("HasSwimmingPool","StarRating", "HotelCapacity")]  
y<-hotelData.df[,c("RoomRent")]  
cor(x,y)

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

cov(x,y)

## [,1]  
## HasSwimmingPool 1094.202  
## StarRating 2048.375  
## HotelCapacity 88753.413

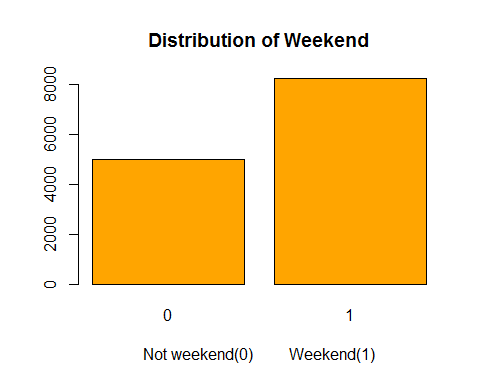
var(x,y)

## [,1]  
## HasSwimmingPool 1094.202  
## StarRating 2048.375  
## HotelCapacity 88753.413

#Forming a variable which is having RoomRent less than 1 lakh because the outliers effect the average  
RoomRent1.df <-hotelData.df[which(hotelData.df$RoomRent<100000),]  
  
#Comparing other factors and their pattern using other trends with roomrent  
  
#Analyzing IsWeekeng effect on RoomRent  
table(hotelData.df$IsWeekend)

##   
## 0 1   
## 4991 8241

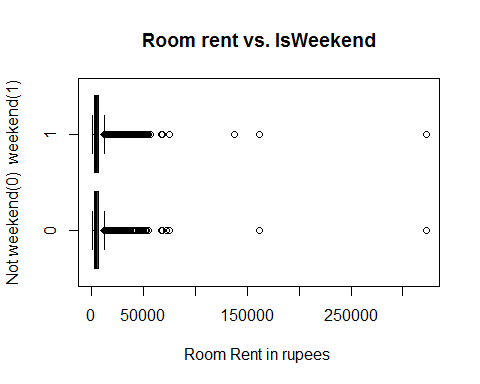
table1<-table(hotelData.df$IsWeekend)  
barplot(table1, main="Distribution of Weekend", xlab="Not weekend(0) Weekend(1)", col="orange")



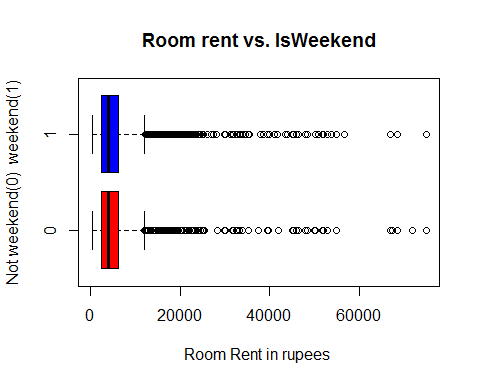
#Effect of Isweekend on RoomRent  
iw= aggregate(RoomRent ~ IsWeekend, data=hotelData.df,mean)  
iw

## IsWeekend RoomRent  
## 1 0 5430.835  
## 2 1 5500.129

boxplot(RoomRent~IsWeekend,data=hotelData.df, main="Room rent vs. IsWeekend", ylab="Not weekend(0) weekend(1)", xlab="Room Rent in rupees ", col=c("red","blue"),horizontal=TRUE)



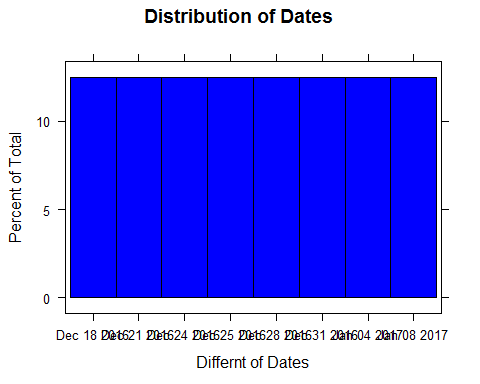
#Without extreme outliers   
boxplot(RoomRent~IsWeekend,data=RoomRent1.df, main="Room rent vs. IsWeekend", ylab="Not weekend(0) weekend(1)", xlab="Room Rent in rupees ", col=c("red","blue"),horizontal=TRUE)



#Comapring RoomRent on different dates  
table(hotelData.df$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

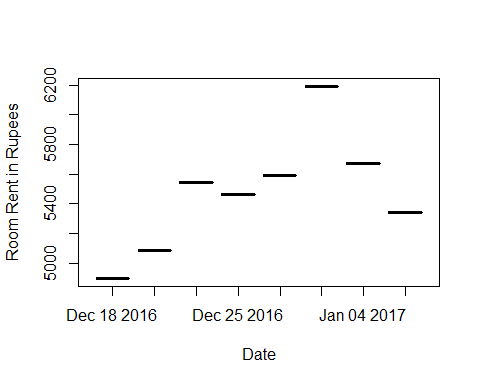
library(lattice)  
histogram(~Date, data = hotelData.df, main="Distribution of Dates", xlab = "Differnt of Dates", col="Blue")



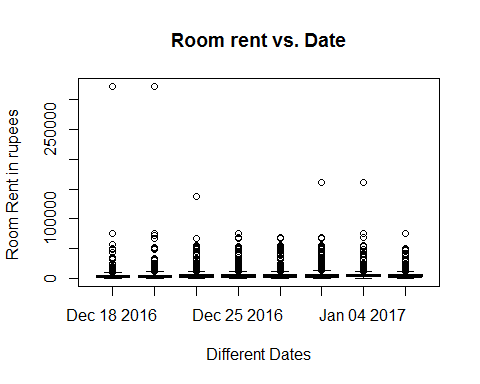
#Effect of different dates on RoomRent  
  
d = aggregate(RoomRent ~ Date, data = hotelData.df,mean)  
d

## Date RoomRent  
## 1 Dec 18 2016 4896.402  
## 2 Dec 21 2016 5085.315  
## 3 Dec 24 2016 5543.236  
## 4 Dec 25 2016 5464.143  
## 5 Dec 28 2016 5593.924  
## 6 Dec 31 2016 6191.776  
## 7 Jan 04 2017 5674.062  
## 8 Jan 08 2017 5342.234

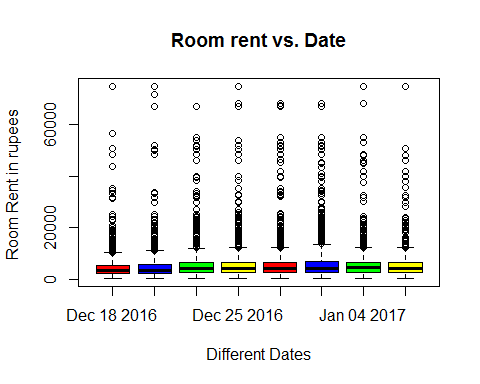
scatterplot(d$Date,d$RoomRent, main="Scatterplot between Date and RoomRent", xlab="Date", ylab = "Room Rent in Rupees")



boxplot(RoomRent~Date,data=hotelData.df, main="Room rent vs. Date", xlab="Different Dates", ylab="Room Rent in rupees ", col=c("red","blue","green","yellow"))



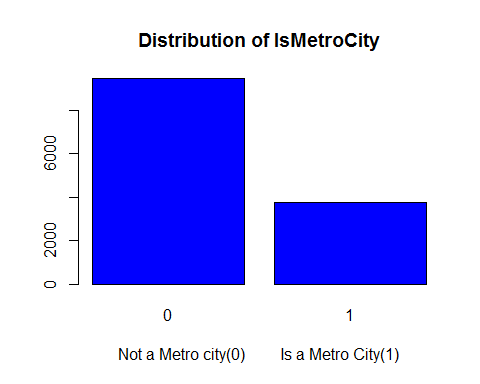
##Without extreme outliers  
boxplot(RoomRent~Date,data=RoomRent1.df, main="Room rent vs. Date", xlab="Different Dates", ylab="Room Rent in rupees ", col=c("red","blue","green","yellow"))



#Analyzing IsMetroCity effect on RoomRent  
table(hotelData.df$IsMetroCity)

##   
## 0 1   
## 9472 3760

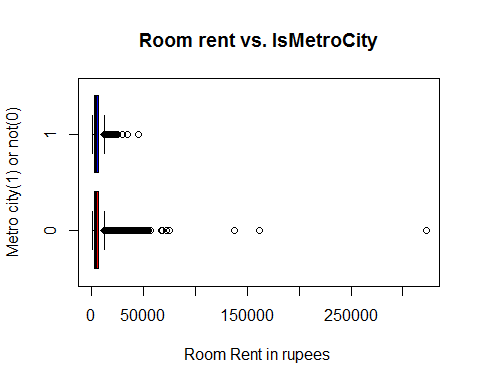
table1<-table(hotelData.df$IsMetroCity)  
barplot(table1, main="Distribution of IsMetroCity", xlab="Not a Metro city(0) Is a Metro City(1)", col="blue")



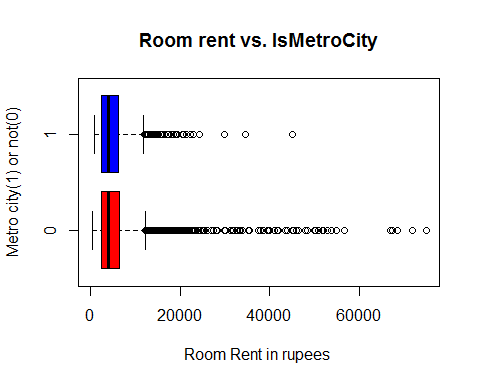
#Effect of IsMetroCity on RoomRent  
imc = aggregate(RoomRent ~ IsMetroCity, data = hotelData.df, mean)  
imc

## IsMetroCity RoomRent  
## 1 0 5782.794  
## 2 1 4696.073

boxplot(RoomRent~IsMetroCity,data=hotelData.df, main="Room rent vs. IsMetroCity", ylab="Metro city(1) or not(0)", xlab="Room Rent in rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



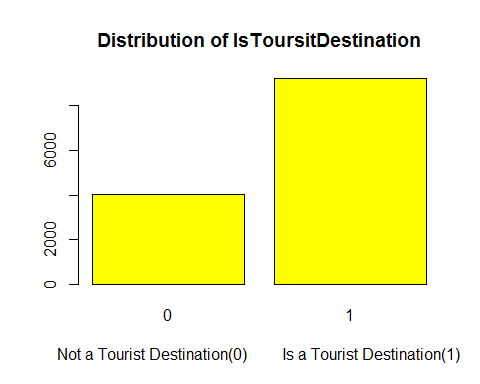
##Without extreme outliers  
boxplot(RoomRent~IsMetroCity,data=RoomRent1.df, main="Room rent vs. IsMetroCity", ylab="Metro city(1) or not(0)", xlab="Room Rent in rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



#Analyzing IsTouristDestination effect on RoomRent  
table(hotelData.df$IsTouristDestination)

##   
## 0 1   
## 4007 9225

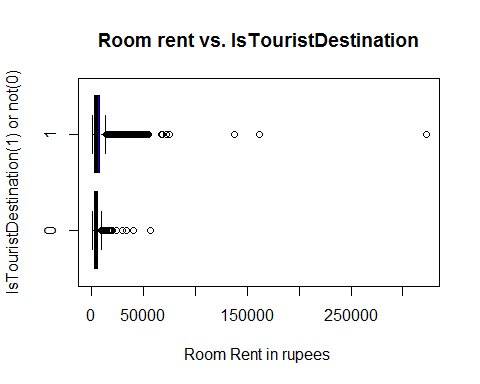
table1<-table(hotelData.df$IsTouristDestination)  
barplot(table1, main="Distribution of IsToursitDestination", xlab="Not a Tourist Destination(0) Is a Tourist Destination(1)", col="yellow")



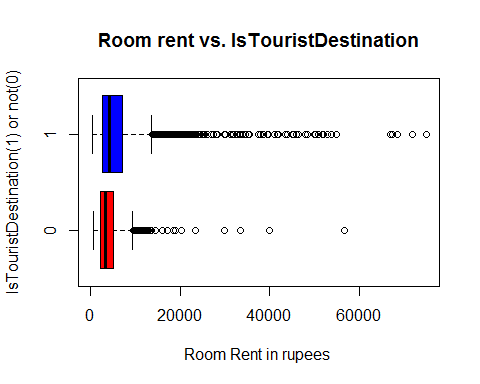
#Effect of IsTouristDestination on RoomRent  
itd = aggregate(RoomRent ~ IsTouristDestination, data = hotelData.df, mean)  
itd

## IsTouristDestination RoomRent  
## 1 0 4111.003  
## 2 1 6066.024

boxplot(RoomRent~IsTouristDestination,data=hotelData.df, main="Room rent vs. IsTouristDestination", ylab="IsTouristDestination(1) or not(0)", xlab="Room Rent in rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



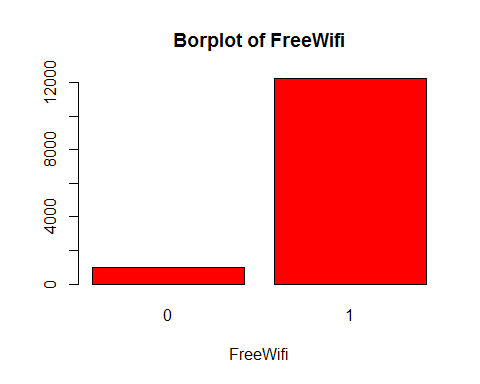
##Without extreme outliers  
boxplot(RoomRent~IsTouristDestination,data=RoomRent1.df, main="Room rent vs. IsTouristDestination", ylab="IsTouristDestination(1) or not(0)", xlab="Room Rent in rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



#Analyzing FreeWifi Vs RoomRent  
table(hotelData.df$FreeWifi)

##   
## 0 1   
## 981 12251

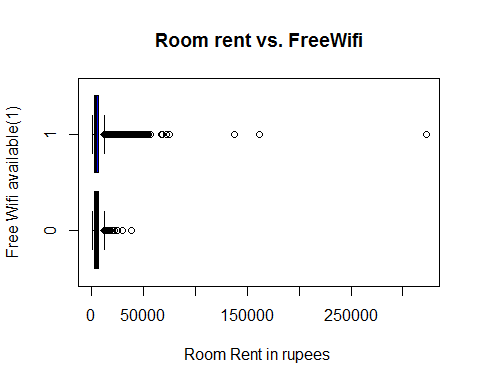
fw<-table(hotelData.df$FreeWifi)  
barplot(fw, main="Borplot of FreeWifi",xlab= "FreeWifi" ,col="red")



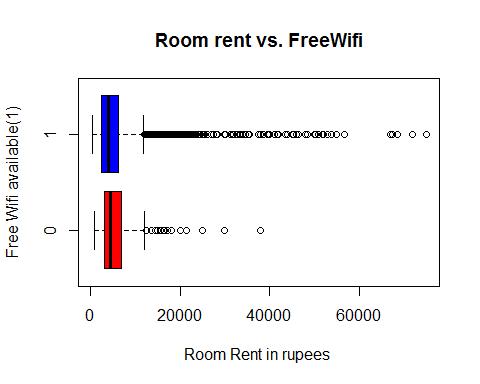
#Effect of FreeWifi on RoomRent  
fw = aggregate(RoomRent ~ FreeWifi, data = hotelData.df, mean)  
fw

## FreeWifi RoomRent  
## 1 0 5380.004  
## 2 1 5481.518

##With extreme outliers of roomrent  
boxplot(RoomRent~FreeWifi,data=hotelData.df, main="Room rent vs. FreeWifi", ylab="Free Wifi available(1)", xlab="Room Rent in rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



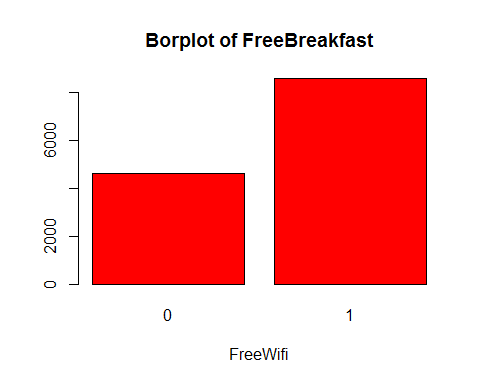
##Without extreme outliers of roomrent  
boxplot(RoomRent~FreeWifi,data=RoomRent1.df, main="Room rent vs. FreeWifi", ylab="Free Wifi available(1)", xlab="Room Rent in rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



#Analyzing FreeBreakfast Vs RoomRent  
table(hotelData.df$FreeWifi)

##   
## 0 1   
## 981 12251

fw<-table(hotelData.df$FreeBreakfast)  
barplot(fw, main="Borplot of FreeBreakfast",xlab= "FreeWifi" ,col="red")



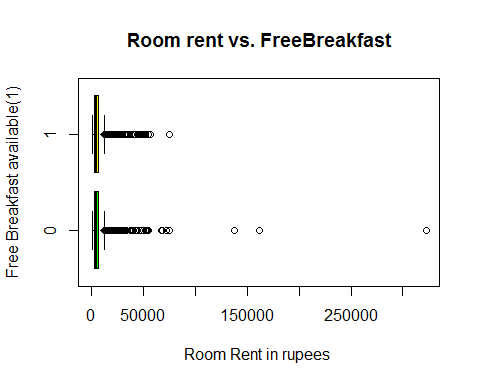
#Effect of FreeBreakfast on RoomRent  
fb = aggregate(RoomRent ~ FreeBreakfast, data =hotelData.df, mean)  
fb1 = aggregate(RoomRent ~ FreeBreakfast, data =RoomRent1.df, mean)  
##Aggregate are affected by outliers a lot in the case of FreeBreakfast on RoomRent  
fb

## FreeBreakfast RoomRent  
## 1 0 5573.790  
## 2 1 5420.044

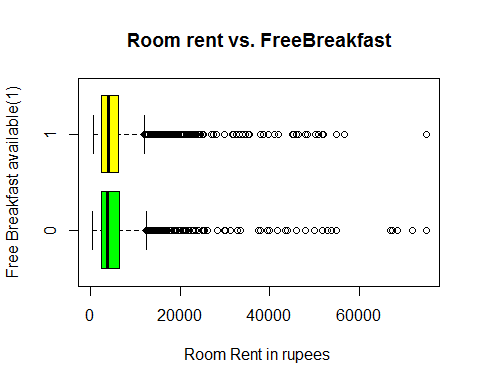
fb1

## FreeBreakfast RoomRent  
## 1 0 5341.260  
## 2 1 5420.044

##With extreme outliers of roomrent  
boxplot(RoomRent~FreeBreakfast,data=hotelData.df, main="Room rent vs. FreeBreakfast", ylab="Free Breakfast available(1)", xlab="Room Rent in rupees ", col=c("green","yellow"),horizontal=TRUE)



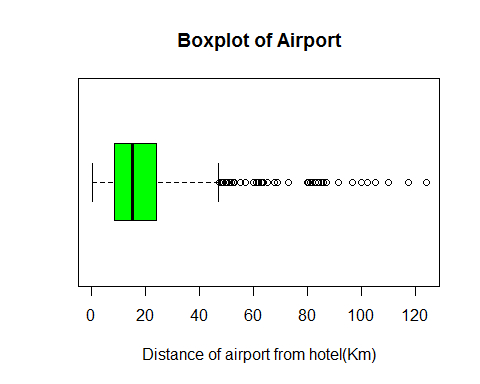
##Without extreme outliers of roomrent  
boxplot(RoomRent~FreeBreakfast,data=RoomRent1.df, main="Room rent vs. FreeBreakfast", ylab="Free Breakfast available(1)", xlab="Room Rent in rupees ", col=c("green","yellow"),horizontal=TRUE)



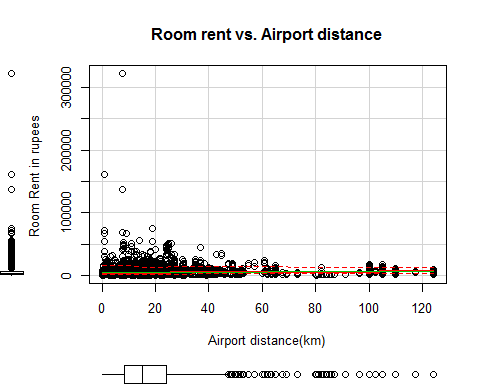
#Analyzing Airport distance from hotel effects in what way on RoomRent  
summary(hotelData.df$Airport)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.20 8.40 15.00 21.16 24.00 124.00

boxplot(hotelData.df$Airport, main="Boxplot of Airport",xlab= "Distance of airport from hotel(Km)" ,col="green",horizontal = TRUE)



#Effect of Airport distance on RoomRent  
  
scatterplot(hotelData.df$Airport,hotelData.df$RoomRent, main="Room rent vs. Airport distance", xlab="Airport distance(km)", ylab="Room Rent in rupees ",cex=1.1)



##Hypothesis  
  
#1.Average RoomRent in hotels having swimming pool is more than that which don't have.  
t.test(RoomRent~HasSwimmingPool,data = hotelData.df, 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.Average RoomRent in hotels with high star rating is high as compared to one which has less star rating.  
t.test(hotelData.df$RoomRent,hotelData.df$StarRating)

##   
## Welch Two Sample t-test  
##   
## data: hotelData.df$RoomRent and hotelData.df$StarRating  
## t = 85.813, df = 13231, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 5345.575 5595.491  
## sample estimates:  
## mean of x mean of y   
## 5473.991838 3.458933

#3.Average RoomRent in hotels providing Free Breakfast is more than that which don't provide.  
t.test(RoomRent~FreeBreakfast, data = hotelData.df, 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.Average RoomRent in metro cities hotels is more than that of non metro cities.  
t.test(RoomRent~IsMetroCity, data = hotelData.df, alternative="less")

##   
## Welch Two Sample t-test  
##   
## data: RoomRent by IsMetroCity  
## t = 10.721, df = 13224, p-value = 1  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 1253.463  
## sample estimates:  
## mean in group 0 mean in group 1   
## 5782.794 4696.073

#5.Average RoomRent in hotels having more hotel capacity is more compared to one with less capacity.  
t.test(hotelData.df$RoomRent,hotelData.df$HotelCapacity)

##   
## Welch Two Sample t-test  
##   
## data: hotelData.df$RoomRent and hotelData.df$HotelCapacity  
## t = 84.882, df = 13234, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 5286.515 5536.445  
## sample estimates:  
## mean of x mean of y   
## 5473.99184 62.51164

#Generating a multiple linear regression model for RoomRent  
#1.  
fit1<-lm(RoomRent~StarRating+HasSwimmingPool+HotelCapacity, data = hotelData.df)  
summary(fit1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity,   
## data = hotelData.df)  
##   
## 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

#Coefficents of the model  
fit1$coefficients

## (Intercept) StarRating HasSwimmingPool HotelCapacity   
## -6896.15387 3597.32177 2528.88529 -15.55769

#Fitted residuals and values are checked and the deviation was around 1000 , because of   
#large data points it's not suitable to show those in the output file.  
  
# Model1: Hotel Rent = b0 + b1\*StarRating + b2\*HasSwimmingPool+ b3\*HotelCapacity  
# b0 = -6896.154, b1 = 3597.322, b2=2528.885, b3= -15.558  
  
  
  
#2.  
fit2<-lm(RoomRent~StarRating+HasSwimmingPool+HotelCapacity+IsWeekend+IsTouristDestination-1, data = hotelData.df)  
summary(fit2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity +   
## IsWeekend + IsTouristDestination - 1, data = hotelData.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8326 -2517 -1212 463 312480   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## StarRating 1258.9558 44.4985 28.292 < 2e-16 \*\*\*  
## HasSwimmingPool 3670.2511 148.8411 24.659 < 2e-16 \*\*\*  
## HotelCapacity -6.1769 0.9658 -6.396 1.65e-10 \*\*\*  
## IsWeekend -509.6479 119.1618 -4.277 1.91e-05 \*\*\*  
## IsTouristDestination 1053.0394 124.7325 8.442 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6792 on 13227 degrees of freedom  
## Multiple R-squared: 0.4493, Adjusted R-squared: 0.4491   
## F-statistic: 2159 on 5 and 13227 DF, p-value: < 2.2e-16

#Coefficents of the model  
fit2$coefficients

## StarRating HasSwimmingPool HotelCapacity   
## 1258.955786 3670.251057 -6.176913   
## IsWeekend IsTouristDestination   
## -509.647863 1053.039364

#Model2: Rent= b0 + b1\*StarRating + b2\*HasSwimmingPool+ b3\*HotelCapacity +b4\*IsWeekend + b5\*IsTouristDestination  
#b0 = -1(assumption), b1 = 1258.9558 , b2=3670.2511, b3= -6.1769 , b4= -509.6479, b5=1053.0394   
  
  
  
#3.  
fit3<-lm(RoomRent~StarRating+HasSwimmingPool+HotelCapacity+Airport, data = hotelData.df)  
summary(fit3)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity +   
## Airport, data = hotelData.df)  
##   
## 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 \*\*\*  
## HasSwimmingPool 2708.400 158.397 17.099 <2e-16 \*\*\*  
## HotelCapacity -14.776 1.006 -14.695 <2e-16 \*\*\*  
## Airport 25.344 2.590 9.786 <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

#Coefficents of the model  
fit3$coefficients

## (Intercept) StarRating HasSwimmingPool HotelCapacity   
## -7288.04830 3522.99002 2708.40013 -14.77562   
## Airport   
## 25.34377

# Model3: Rent = b0 + b1\*StarRating + b2\*HasSwimmingPool+ b3\*HotelCapacity +b4\*Airport  
# b0 = -7288.04830, b1 = 3522.99002 , b2=2708.40013, b3= -14.77562, b4= 25.34377