Akhil\_Menon\_Project\_Code.R

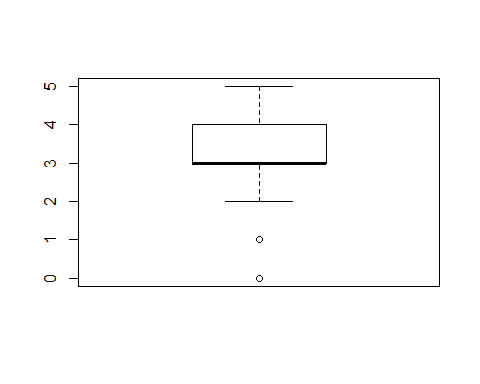
menon

Tue Jul 11 21:52:47 2017

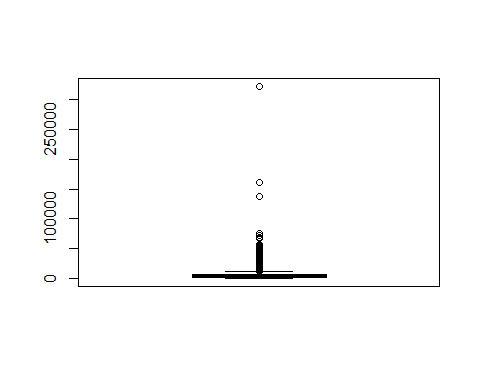
# Project Title: Hotel Room Price Analysis in India  
# NAME: Akhil Menon  
# EMAIL: menonakhil90@gmail.com  
# COLLEGE: Sri Venkateswara College of Engineering  
  
  
# Reading csv file onto a data frame  
hotel<-read.csv(paste("MergedHotelPricingData.csv", sep=""))  
View(hotel)  
  
# Summary Stats of data frame containing Dataset  
summary(hotel)

## 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 OYO Rooms : 32 Min. : 299   
## Dec 24 2016:1611 Vivanta by Taj : 32 1st Qu.: 2436   
## Dec 25 2016:1611 Goldfinch Hotel : 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) :13088   
## 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   
##

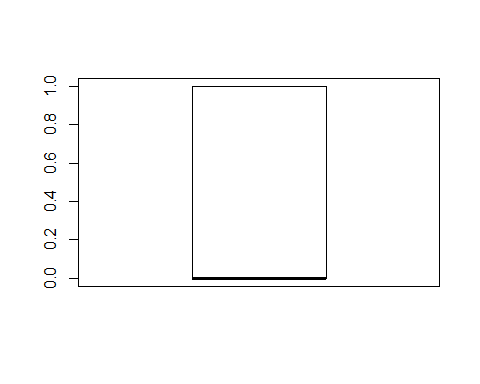
# Boxplot visualizations of Important Variables  
boxplot(hotel$StarRating)



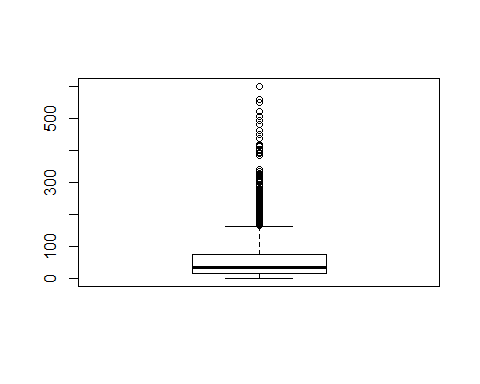
boxplot(hotel$RoomRent)



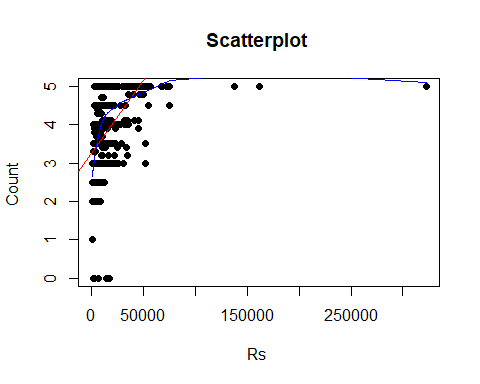
boxplot(hotel$HasSwimmingPool)



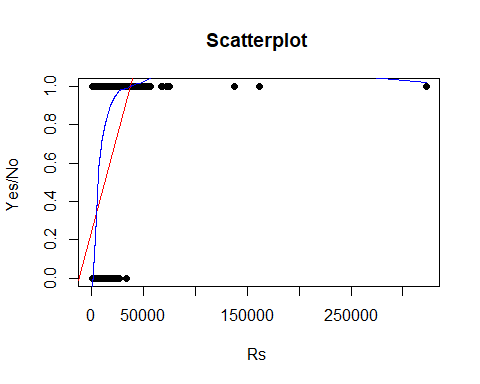
boxplot(hotel$HotelCapacity)



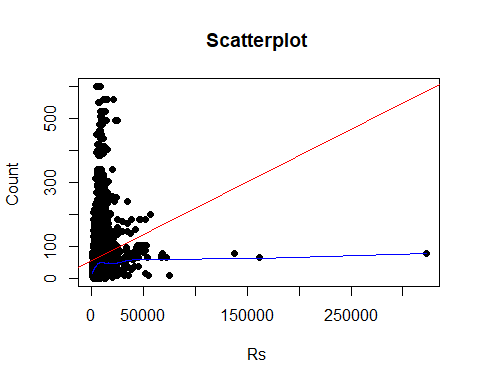
# ScatterPlot Visualizations  
attach(hotel)  
plot(RoomRent, StarRating, main="Scatterplot ",   
 xlab="Rs ", ylab="Count ", pch=19)  
abline(lm(StarRating~RoomRent), col="red") # regression line (y~x)   
lines(lowess(RoomRent,StarRating), col="blue") # lowess line (x,y)



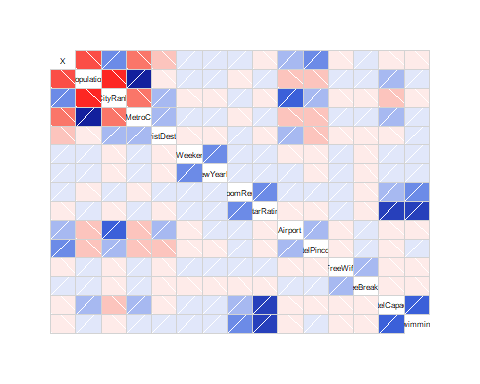
plot(RoomRent, HasSwimmingPool, main="Scatterplot ",   
 xlab="Rs ", ylab="Yes/No ", pch=19)  
abline(lm(HasSwimmingPool~RoomRent), col="red") # regression line (y~x)   
lines(lowess(RoomRent,HasSwimmingPool), col="blue") # lowess line (x,y)



plot(RoomRent, HotelCapacity, main="Scatterplot ",   
 xlab="Rs ", ylab="Count ", pch=19)  
abline(lm(HotelCapacity~RoomRent), col="red") # regression line (y~x)   
lines(lowess(RoomRent,HotelCapacity), col="blue") # lowess line (x,y)  
  
  
  
library(corrgram)



corrgram(hotel)



# Creation of Covariance Matrix  
library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

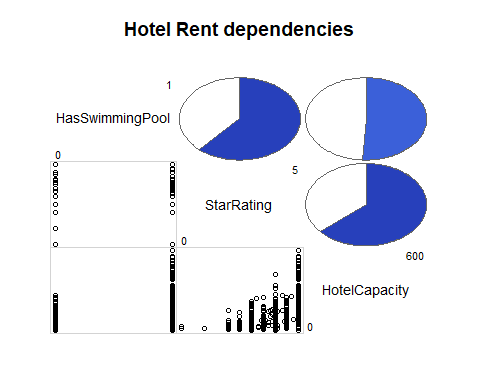
##   
## Attaching package: 'Hmisc'

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

colhotel <- c("StarRating","HasSwimmingPool","HotelCapacity")  
corMatrix <- rcorr(as.matrix(hotel[,colhotel]))  
corMatrix

## StarRating HasSwimmingPool HotelCapacity  
## StarRating 1.00 0.62 0.64  
## HasSwimmingPool 0.62 1.00 0.51  
## HotelCapacity 0.64 0.51 1.00  
##   
## n= 13232   
##   
##   
## P  
## StarRating HasSwimmingPool HotelCapacity  
## StarRating 0 0   
## HasSwimmingPool 0 0   
## HotelCapacity 0 0

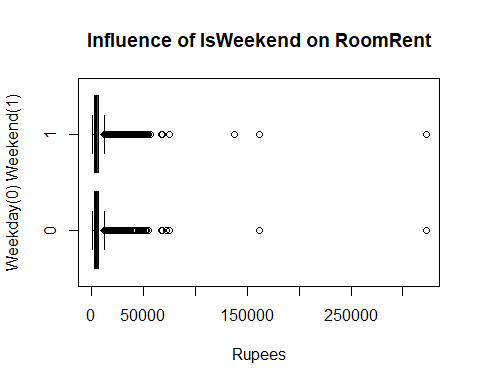
library(Hmisc)  
library(car)  
library(corrgram)  
  
# Creation of Corrgram  
  
corrgram(hotel[,colhotel], order=TRUE,  
 main="Hotel Rent dependencies",  
 lower.panel=panel.pts, upper.panel=panel.pie,  
 diag.panel=panel.minmax, text.panel=panel.txt)



# FACTORS INFLUENCING HOTEL PRICES  
  
# INFLUENCE OF IsWeekend on RoomRent  
  
week <- aggregate(RoomRent ~ IsWeekend, data = hotel, mean)  
week

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

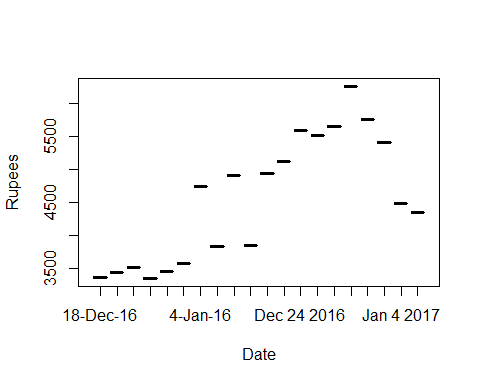
boxplot(RoomRent~IsWeekend, data = hotel, main = "Influence of IsWeekend on RoomRent",   
 ylab = "Weekday(0) Weekend(1)", xlab = "Rupees", col = c("Blue","Red"), horizontal = TRUE)



# INFLUENCE of different dates of the year on RoomRent  
  
dts = aggregate(RoomRent ~ Date, data = hotel,mean)  
dts

## Date RoomRent  
## 1 18-Dec-16 3366.795  
## 2 21-Dec-16 3437.545  
## 3 24-Dec-16 3510.795  
## 4 25-Dec-16 3349.591  
## 5 28-Dec-16 3450.045  
## 6 31-Dec-16 3570.318  
## 7 4-Jan-16 4738.548  
## 8 4-Jan-17 3829.615  
## 9 8-Jan-16 4907.419  
## 10 8-Jan-17 3843.077  
## 11 Dec 18 2016 4938.257  
## 12 Dec 21 2016 5130.320  
## 13 Dec 24 2016 5598.746  
## 14 Dec 25 2016 5521.896  
## 15 Dec 28 2016 5652.478  
## 16 Dec 31 2016 6263.374  
## 17 Jan 04 2017 5754.513  
## 18 Jan 08 2017 5406.821  
## 19 Jan 4 2017 4481.400  
## 20 Jan 8 2017 4347.821

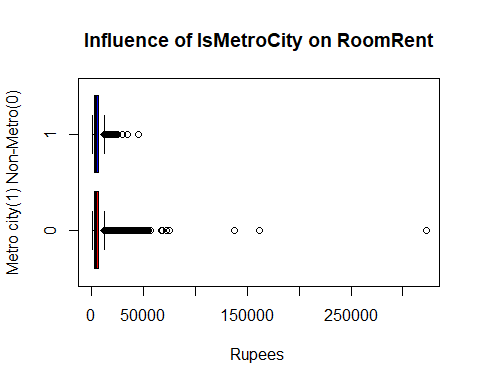
scatterplot(dts$Date,dts$RoomRent, main="Influence of Different Dates on RoomRent", xlab="Date", ylab = "Rupees")



# From above plot we can infer that the Highest RoomRent on New Year's eve (31st Dec 2016); Rent above 6000 Rupees  
  
# INFLUENCE of IsMetroCity on RoomRent  
  
metro = aggregate(RoomRent ~ IsMetroCity, data = hotel, mean)  
metro

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

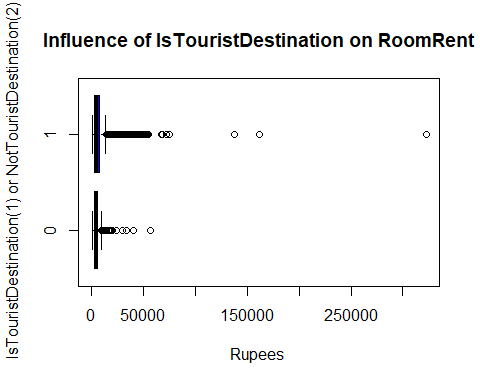
boxplot(RoomRent~IsMetroCity,data=hotel, main="Influence of IsMetroCity on RoomRent", ylab="Metro city(1) Non-Metro(0)", xlab="Rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



# INFLUENCE of IsTouristDestination on RoomRent  
TouDes = aggregate(RoomRent ~ IsTouristDestination, data = hotel, mean)  
TouDes

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

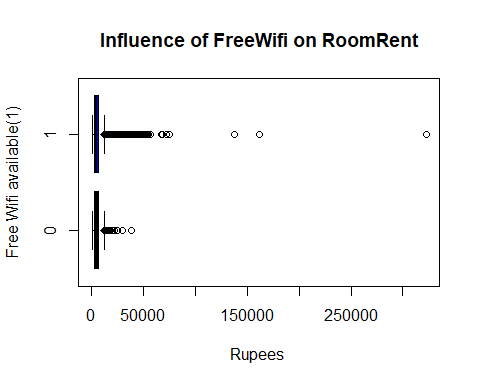
boxplot(RoomRent~IsTouristDestination,data=hotel, main="Influence of IsTouristDestination on RoomRent", ylab=" IsTouristDestination(1) or NotTouristDestination(2)", xlab="Rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



#It is clearly visualized that the RoomRent is more at Tourist destinations than Non Tourist Destinations  
  
#INFLUENCE of FreeWifi on RoomRent  
fw = aggregate(RoomRent ~ FreeWifi, data = hotel, mean)  
fw

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

boxplot(RoomRent~FreeWifi,data=hotel, main="Influence of FreeWifi on RoomRent", ylab="Free Wifi available(1)", xlab="Rupees ", col=c("red","blue","green","yellow"),horizontal=TRUE)



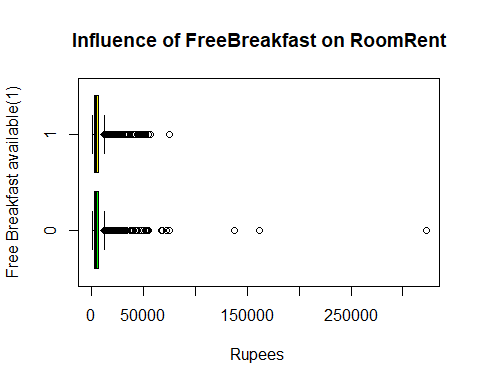
#INFLUENCE of FreeBreakfast on RoomRent  
  
fb = aggregate(RoomRent ~ FreeBreakfast, data =hotel, mean)  
fb1 = aggregate(RoomRent ~ FreeBreakfast, data =hotel, mean)  
fb

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

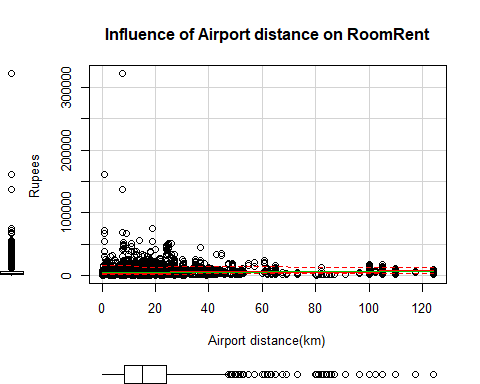
fb1

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

boxplot(RoomRent~FreeBreakfast,data=hotel, main="Influence of FreeBreakfast on RoomRent", ylab="Free Breakfast available(1)", xlab="Rupees ", col=c("green","yellow"),horizontal=TRUE)



# INFLUENCE of Airport Distance on RoomRent  
scatterplot(hotel$Airport,hotel$RoomRent, main="Influence of Airport distance on RoomRent", xlab="Airport distance(km)", ylab="Rupees ",cex=1.1)



# t-tests  
  
# Avg. RoomRents for Hotels with and without Swimming Pools  
  
t.test(RoomRent~HasSwimmingPool,data = hotel, 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

# p-value < 0.05; Rejection of Null Hypothesis of equal mean  
  
# Avg RoomRent based on star-rating  
t.test(hotel$RoomRent,hotel$StarRating)

##   
## Welch Two Sample t-test  
##   
## data: hotel$RoomRent and hotel$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

# p-value < 0.05; Rejection of Null Hypothesis of equal mean  
  
# Avg RoomRent in Metro and other cities  
t.test(RoomRent~IsMetroCity, data = hotel, 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

# p-value > 0.05; Fail to Reject Null Hypothesis of Equal mean  
  
# Avg RoomRent based on Hotels that provide and don't Provide free breakfast  
  
t.test(RoomRent~FreeBreakfast, data = hotel, 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

# p-value > 0.05; Fail to Reject Null Hypothesis of equal mean  
  
#Generating a multiple linear regression model for RoomRent  
  
# Model-1 (Using Variables -> RoomRent,HasSwimmingPool,HotelCapacity)  
fita<-lm(RoomRent~StarRating+HasSwimmingPool+HotelCapacity-1, data = hotel)  
summary(fita)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity -   
## 1, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8039 -2448 -1249 461 312401   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## StarRating 1396.8746 26.1320 53.455 < 2e-16 \*\*\*  
## HasSwimmingPool 3719.6943 148.7835 25.001 < 2e-16 \*\*\*  
## HotelCapacity -7.6598 0.9415 -8.136 4.44e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6813 on 13229 degrees of freedom  
## Multiple R-squared: 0.4457, Adjusted R-squared: 0.4456   
## F-statistic: 3546 on 3 and 13229 DF, p-value: < 2.2e-16

# Model-1 Coefficients  
fita$coefficients

## StarRating HasSwimmingPool HotelCapacity   
## 1396.874562 3719.694300 -7.659814

# Model-2 (Using Variables -> RoomRent,HasSwimmingPool, HotelCapacity, Isweekend, IsTouristDestination)  
fitb<-lm(RoomRent~StarRating+HasSwimmingPool+HotelCapacity+IsWeekend+IsTouristDestination-1, data = hotel)  
summary(fitb)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity +   
## IsWeekend + IsTouristDestination - 1, data = hotel)  
##   
## 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

# Model-2 Coefficients  
fitb$coefficients

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

# Model-3 (Using Variables -> RoomRent,HasSwimmingPool,HotelCapacity,Airport)  
fitc<-lm(RoomRent~StarRating+HasSwimmingPool+HotelCapacity+Airport-1, data = hotel)  
summary(fitc)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity +   
## Airport - 1, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8240 -2380 -1224 384 312742   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## StarRating 1248.4270 33.2220 37.578 < 2e-16 \*\*\*  
## HasSwimmingPool 3903.7369 150.6728 25.909 < 2e-16 \*\*\*  
## HotelCapacity -6.7434 0.9482 -7.112 1.20e-12 \*\*\*  
## Airport 18.8697 2.6157 7.214 5.73e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6800 on 13228 degrees of freedom  
## Multiple R-squared: 0.4479, Adjusted R-squared: 0.4477   
## F-statistic: 2683 on 4 and 13228 DF, p-value: < 2.2e-16

# Model-3 Coefficients  
fitc$coefficients

## StarRating HasSwimmingPool HotelCapacity Airport   
## 1248.426988 3903.736921 -6.743354 18.869726