# Analysis of Hotel Room Pricing in the Indian Market

# NAME: Sonal Somani

# EMAIL: [sonalm300@gmail.com](mailto:sonalm300@gmail.com)

# COLLEGE / COMPANY: Antuit India Pvt. Ltd.

#Load packages  
  
library(statsr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(ggplot2)  
library(gplots)

##   
## Attaching package: 'gplots'

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

library(car)

##   
## Attaching package: 'car'

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

library(lattice)  
library(psych)

##   
## Attaching package: 'psych'

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

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

library(corrgram)  
  
#set working directory  
  
setwd("C:/Users/Sonal Somani/Desktop/IIMInternship/R\_code")  
  
#load dataset into R  
  
hotel <- read.csv(paste("Collated\_Hotel\_Data.csv",sep=""))  
  
#View dataset  
View(hotel)  
dim(hotel)

## [1] 13232 19

#Check your dataset's observations and datatypes  
str(hotel)

## 'data.frame': 13232 obs. of 19 variables:  
## $ CityName : Factor w/ 42 levels "Agra","Ahmedabad",..: 7 14 14 14 14 14 14 14 14 14 ...  
## $ Population : int 7088416 6731790 6731790 6731790 6731790 6731790 6731790 6731790 6731790 6731790 ...  
## $ CityRank : int 3 4 4 4 4 4 4 4 4 4 ...  
## $ IsMetroCity : int 1 0 0 0 0 0 0 0 0 0 ...  
## $ IsTouristDestination: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ IsWeekend : int 1 1 0 1 1 0 1 0 1 1 ...  
## $ IsNewYearEve : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ Date : Factor w/ 10 levels "1/4/2016","1/4/2017",..: 4 5 6 7 8 9 10 2 4 5 ...  
## $ HotelName : Factor w/ 1670 levels "14 Square Amanora",..: 1670 22 22 22 22 22 22 22 22 159 ...  
## $ RoomRent : int 3158 3440 3280 3440 3440 3280 3280 4100 4500 4840 ...  
## $ StarRating : num 4 3 3 3 3 3 3 3 3 4 ...  
## $ Airport : num 5.5 21 21 21 21 21 21 21 21 21 ...  
## $ HotelAddress : Factor w/ 2108 levels " H.P. High Court Mall Road, Shimla",..: 657 996 996 996 996 996 996 996 996 21 ...  
## $ HotelPincode : int 600117 500082 500082 500082 500082 500082 500082 500082 500082 500001 ...  
## $ HotelDescription : Factor w/ 1226 levels "#NAME?","10 star hotel near Queensroad, Amritsar",..: 731 64 64 64 64 64 64 64 64 619 ...  
## $ FreeWifi : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ FreeBreakfast : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ HotelCapacity : int 40 88 88 88 88 88 88 88 88 112 ...  
## $ HasSwimmingPool : int 1 0 0 0 0 0 0 0 0 1 ...

#Summarising dataset's mean,std dev , median etc.  
  
describe(hotel)[,c(2,3,4,5,8,9)]

## n mean sd median min max  
## CityName\* 13232 18.07 11.72 16 1.0 42  
## Population 13232 4416836.87 4258386.00 3046163 8096.0 12442373  
## CityRank 13232 14.83 13.51 9 0.0 44  
## IsMetroCity 13232 0.28 0.45 0 0.0 1  
## IsTouristDestination 13232 0.70 0.46 1 0.0 1  
## IsWeekend 13232 0.62 0.48 1 0.0 1  
## IsNewYearEve 13232 0.12 0.33 0 0.0 1  
## Date\* 13232 6.37 2.50 7 1.0 10  
## HotelName\* 13232 841.19 488.16 827 1.0 1670  
## RoomRent 13232 5473.99 7333.12 4000 299.0 322500  
## StarRating 13232 3.46 0.76 3 0.0 5  
## Airport 13232 21.16 22.76 15 0.2 124  
## HotelAddress\* 13232 1202.53 582.17 1261 1.0 2108  
## HotelPincode 13232 397430.26 259837.50 395003 100025.0 7000157  
## HotelDescription\* 13224 581.34 363.26 567 1.0 1226  
## FreeWifi 13232 0.93 0.26 1 0.0 1  
## FreeBreakfast 13232 0.65 0.48 1 0.0 1  
## HotelCapacity 13232 62.51 76.66 34 0.0 600  
## HasSwimmingPool 13232 0.36 0.48 0 0.0 1

#The dependent variable , Y would be RoomRent as it varies based on other factors like star rating, amenties, location etc.  
  
##To determine which 3 independent variables we will be choosing for our analysis, let's Look for the predictor variable that is associated with the greatest increase in R-squared.  
  
#Let us do stepwise regression -  
  
model <- lm(RoomRent ~ StarRating,data=hotel)  
summary(model)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8495 -2311 -1130 811 311505   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6917.79 277.46 -24.93 <2e-16 \*\*\*  
## StarRating 3582.54 78.36 45.72 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6815 on 13230 degrees of freedom  
## Multiple R-squared: 0.1364, Adjusted R-squared: 0.1364   
## F-statistic: 2090 on 1 and 13230 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.1906, Adjusted R-squared: 0.1905  
  
model\_1 <- lm(RoomRent ~ StarRating+IsMetroCity,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + IsMetroCity, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9050 -2424 -976 861 310950   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6724.09 276.47 -24.32 <2e-16 \*\*\*  
## StarRating 3654.89 78.19 46.75 <2e-16 \*\*\*  
## IsMetroCity -1562.28 131.06 -11.92 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6779 on 13229 degrees of freedom  
## Multiple R-squared: 0.1456, Adjusted R-squared: 0.1455   
## F-statistic: 1127 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2023, Adjusted R-squared: 0.2022  
  
model\_1 <- lm(RoomRent ~ StarRating+IsTouristDestination,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + IsTouristDestination, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9244 -2471 -844 893 310756   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8637.42 292.01 -29.58 <2e-16 \*\*\*  
## StarRating 3636.73 77.57 46.88 <2e-16 \*\*\*  
## IsTouristDestination 2197.75 127.62 17.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6740 on 13229 degrees of freedom  
## Multiple R-squared: 0.1554, Adjusted R-squared: 0.1552   
## F-statistic: 1217 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2177, Adjusted R-squared: 0.2175  
  
  
model\_1 <- lm(RoomRent ~ StarRating+IsWeekend,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + IsWeekend, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8507 -2311 -1134 800 311526   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6938.26 287.27 -24.153 <2e-16 \*\*\*  
## StarRating 3582.41 78.37 45.712 <2e-16 \*\*\*  
## IsWeekend 33.64 122.24 0.275 0.783   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6815 on 13229 degrees of freedom  
## Multiple R-squared: 0.1364, Adjusted R-squared: 0.1363   
## F-statistic: 1045 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.1906, Adjusted R-squared: 0.1905  
  
model\_1 <- lm(RoomRent ~ StarRating+CityRank,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9808 -2345 -925 792 311676   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8744.86 292.06 -29.94 <2e-16 \*\*\*  
## StarRating 3771.30 78.11 48.28 <2e-16 \*\*\*  
## CityRank 79.16 4.37 18.11 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6732 on 13229 degrees of freedom  
## Multiple R-squared: 0.1573, Adjusted R-squared: 0.1572   
## F-statistic: 1235 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2228, Adjusted R-squared: 0.2226 \*\*\*\*  
  
model\_1 <- lm(RoomRent ~ StarRating+Airport,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + Airport, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8409 -2243 -1075 732 311753   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7560.13 285.73 -26.459 <2e-16 \*\*\*  
## StarRating 3625.43 78.28 46.316 <2e-16 \*\*\*  
## Airport 23.35 2.60 8.979 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6794 on 13229 degrees of freedom  
## Multiple R-squared: 0.1417, Adjusted R-squared: 0.1415   
## F-statistic: 1092 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2001, Adjusted R-squared: 0.2  
  
model\_1 <- lm(RoomRent ~ StarRating+HotelCapacity,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HotelCapacity, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10257 -2392 -1041 940 310470   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8922.2514 319.0707 -27.96 <2e-16 \*\*\*  
## StarRating 4387.1887 101.0638 43.41 <2e-16 \*\*\*  
## HotelCapacity -12.4577 0.9967 -12.50 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6775 on 13229 degrees of freedom  
## Multiple R-squared: 0.1465, Adjusted R-squared: 0.1464   
## F-statistic: 1135 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2031, Adjusted R-squared: 0.203  
  
model\_1 <- lm(RoomRent ~ StarRating+HasSwimmingPool,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + HasSwimmingPool, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8581 -2238 -999 793 311419   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4855.29 316.64 -15.33 <2e-16 \*\*\*  
## StarRating 2773.44 98.98 28.02 <2e-16 \*\*\*  
## HasSwimmingPool 2068.93 156.30 13.24 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6770 on 13229 degrees of freedom  
## Multiple R-squared: 0.1477, Adjusted R-squared: 0.1476   
## F-statistic: 1146 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2089, Adjusted R-squared: 0.2087  
  
model\_1 <- lm(RoomRent ~ StarRating+FreeWifi,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + FreeWifi, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8489 -2314 -1122 784 311511   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6827.07 344.45 -19.820 <2e-16 \*\*\*  
## StarRating 3583.22 78.38 45.715 <2e-16 \*\*\*  
## FreeWifi -100.53 226.17 -0.444 0.657   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6815 on 13229 degrees of freedom  
## Multiple R-squared: 0.1364, Adjusted R-squared: 0.1363   
## F-statistic: 1045 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.1906, Adjusted R-squared: 0.1905   
  
model\_1 <- lm(RoomRent ~ StarRating+FreeBreakfast,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + FreeBreakfast, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8486 -2310 -1131 802 311523   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6938.10 291.49 -23.802 <2e-16 \*\*\*  
## StarRating 3583.12 78.41 45.699 <2e-16 \*\*\*  
## FreeBreakfast 28.24 124.20 0.227 0.82   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6815 on 13229 degrees of freedom  
## Multiple R-squared: 0.1364, Adjusted R-squared: 0.1363   
## F-statistic: 1045 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.191, Adjusted R-squared: 0.1908  
  
model\_1 <- lm(RoomRent ~ StarRating+IsNewYearEve,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + IsNewYearEve, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9224 -2309 -1116 774 311610   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7018.4 278.1 -25.239 < 2e-16 \*\*\*  
## StarRating 3581.6 78.3 45.740 < 2e-16 \*\*\*  
## IsNewYearEve 834.6 179.4 4.653 3.3e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6810 on 13229 degrees of freedom  
## Multiple R-squared: 0.1378, Adjusted R-squared: 0.1377   
## F-statistic: 1057 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.1927, Adjusted R-squared: 0.1925  
  
model\_1 <- lm(RoomRent ~ StarRating+Population,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + Population, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9704 -2423 -913 894 310890   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.481e+03 2.755e+02 -23.53 <2e-16 \*\*\*  
## StarRating 3.766e+03 7.819e+01 48.16 <2e-16 \*\*\*  
## Population -2.425e-04 1.388e-05 -17.47 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6738 on 13229 degrees of freedom  
## Multiple R-squared: 0.1559, Adjusted R-squared: 0.1558   
## F-statistic: 1222 on 2 and 13229 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2183, Adjusted R-squared: 0.2182  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+IsMetroCity,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + IsMetroCity,   
## data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9821 -2382 -910 794 311537   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8548.978 304.930 -28.036 <2e-16 \*\*\*  
## StarRating 3771.791 78.100 48.294 <2e-16 \*\*\*  
## CityRank 72.554 5.277 13.748 <2e-16 \*\*\*  
## IsMetroCity -350.688 157.169 -2.231 0.0257 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6731 on 13228 degrees of freedom  
## Multiple R-squared: 0.1576, Adjusted R-squared: 0.1574   
## F-statistic: 825.1 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2229, Adjusted R-squared: 0.2227  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+IsTouristDestination,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + IsTouristDestination,   
## data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10112 -2341 -809 888 311067   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9693.261 299.616 -32.35 <2e-16 \*\*\*  
## StarRating 3774.545 77.637 48.62 <2e-16 \*\*\*  
## CityRank 63.085 4.522 13.95 <2e-16 \*\*\*  
## IsTouristDestination 1686.186 131.901 12.78 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6691 on 13228 degrees of freedom  
## Multiple R-squared: 0.1676, Adjusted R-squared: 0.1674   
## F-statistic: 887.8 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2367, Adjusted R-squared: 0.2365  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+IsWeekend,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + IsWeekend, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9826 -2348 -920 789 311706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8774.202 301.338 -29.117 <2e-16 \*\*\*  
## StarRating 3771.127 78.115 48.276 <2e-16 \*\*\*  
## CityRank 79.168 4.371 18.114 <2e-16 \*\*\*  
## IsWeekend 47.780 120.757 0.396 0.692   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6732 on 13228 degrees of freedom  
## Multiple R-squared: 0.1573, Adjusted R-squared: 0.1571   
## F-statistic: 823.2 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2228, Adjusted R-squared: 0.2226  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+IsNewYearEve,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + IsNewYearEve,   
## data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10058 -2369 -901 804 311781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8845.775 292.612 -30.230 < 2e-16 \*\*\*  
## StarRating 3770.394 78.049 48.308 < 2e-16 \*\*\*  
## CityRank 79.163 4.367 18.128 < 2e-16 \*\*\*  
## IsNewYearEve 835.566 177.189 4.716 2.43e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6727 on 13228 degrees of freedom  
## Multiple R-squared: 0.1587, Adjusted R-squared: 0.1585   
## F-statistic: 832 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2249, Adjusted R-squared: 0.2247  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+Population,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + Population, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9870 -2376 -905 856 311339   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.890e+03 3.518e+02 -22.427 < 2e-16 \*\*\*  
## StarRating 3.785e+03 7.813e+01 48.453 < 2e-16 \*\*\*  
## CityRank 5.063e+01 7.879e+00 6.426 1.36e-10 \*\*\*  
## Population -1.088e-04 2.500e-05 -4.350 1.37e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6728 on 13228 degrees of freedom  
## Multiple R-squared: 0.1585, Adjusted R-squared: 0.1583   
## F-statistic: 830.7 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2237, Adjusted R-squared: 0.2235  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+Airport,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + Airport, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9812 -2347 -926 791 311674   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8743.4728 293.0006 -29.84 <2e-16 \*\*\*  
## StarRating 3771.3312 78.1169 48.28 <2e-16 \*\*\*  
## CityRank 79.3089 5.0575 15.68 <2e-16 \*\*\*  
## Airport -0.1779 2.9814 -0.06 0.952   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6732 on 13228 degrees of freedom  
## Multiple R-squared: 0.1573, Adjusted R-squared: 0.1571   
## F-statistic: 823.2 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2228, Adjusted R-squared: 0.2227  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+HotelCapacity,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + HotelCapacity,   
## data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10931 -2388 -949 894 310919   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9971.831 323.095 -30.864 <2e-16 \*\*\*  
## StarRating 4324.275 100.214 43.150 <2e-16 \*\*\*  
## CityRank 70.373 4.472 15.738 <2e-16 \*\*\*  
## HotelCapacity -8.886 1.013 -8.769 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6713 on 13228 degrees of freedom  
## Multiple R-squared: 0.1622, Adjusted R-squared: 0.162   
## F-statistic: 853.6 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2285, Adjusted R-squared: 0.2283  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+HasSwimmingPool,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + HasSwimmingPool,   
## data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9924 -2354 -767 857 311590   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6643.532 327.034 -20.32 <2e-16 \*\*\*  
## StarRating 2936.337 98.108 29.93 <2e-16 \*\*\*  
## CityRank 80.750 4.341 18.60 <2e-16 \*\*\*  
## HasSwimmingPool 2144.759 154.358 13.89 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6684 on 13228 degrees of freedom  
## Multiple R-squared: 0.1694, Adjusted R-squared: 0.1693   
## F-statistic: 899.5 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2422, Adjusted R-squared: 0.242  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+FreeWifi,data=hotel)  
summary(model\_2)

##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + FreeWifi, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9848 -2356 -919 800 311653   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9118.269 362.801 -25.133 <2e-16 \*\*\*  
## StarRating 3770.836 78.106 48.278 <2e-16 \*\*\*  
## CityRank 80.073 4.402 18.191 <2e-16 \*\*\*  
## FreeWifi 390.355 225.030 1.735 0.0828 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6732 on 13228 degrees of freedom  
## Multiple R-squared: 0.1575, Adjusted R-squared: 0.1573   
## F-statistic: 824.4 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.223, Adjusted R-squared: 0.2228  
  
model\_2 <- lm(RoomRent ~ StarRating+CityRank+FreeBreakfast,data=hotel)  
summary(model\_2)

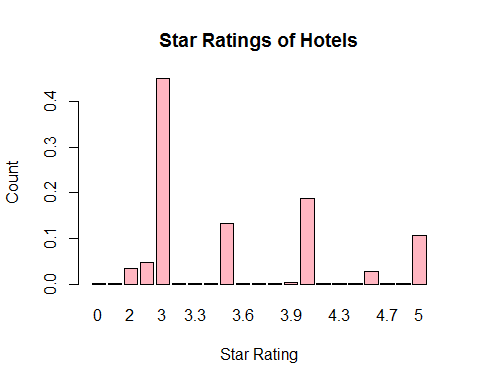
##   
## Call:  
## lm(formula = RoomRent ~ StarRating + CityRank + FreeBreakfast,   
## data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9830 -2348 -922 788 311711   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8786.708 305.495 -28.762 <2e-16 \*\*\*  
## StarRating 3772.520 78.158 48.268 <2e-16 \*\*\*  
## CityRank 79.184 4.371 18.116 <2e-16 \*\*\*  
## FreeBreakfast 57.332 122.707 0.467 0.64   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6732 on 13228 degrees of freedom  
## Multiple R-squared: 0.1573, Adjusted R-squared: 0.1571   
## F-statistic: 823.3 on 3 and 13228 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.2233, Adjusted R-squared: 0.2231   
  
#So, according to the above adjusted R square values, we see that the best results are from model with explanatory variables as StarRating, CityRank ,HasSwimmingPool. Hence, we pick these three variables for further examination.

#Lets visualize the above mentioned variables Y,x1,x2,x3 where Y is Room rent and X1,X2,X3 are StarRating, CityRank, and HasSwimmingPool respectively.  
  
#Since star rating is categorical in nature with ordinal values, we draw a table and barchart for StarRating.  
  
 table(hotel$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 1744 8 24 16 32 2479 24   
## 4.3 4.4 4.5 4.7 4.8 5   
## 16 8 368 8 16 1408

Rating <- prop.table(table(hotel$StarRating))  
 barplot(Rating,main = "Star Ratings of Hotels",xlab="Star Rating",ylab="Count",col="lightpink")

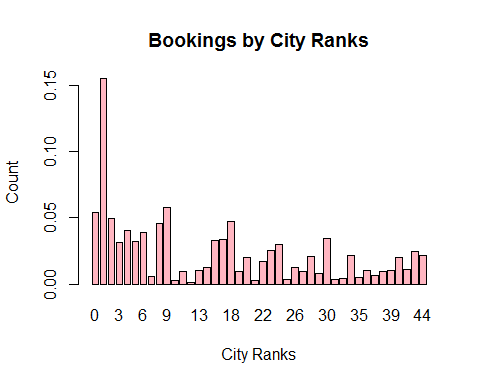


#We see that 3 star rated hotels are the most common with being close to 6000 in number.

#Table for CityRank  
 table(hotel$CityRank)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14   
## 712 2048 656 416 536 424 512 80 600 768 32 128 16 136 160   
## 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30   
## 432 448 624 128 264 40 224 336 392 48 160 120 272 104 456   
## 32 33 34 35 36 37 38 39 40 42 43 44   
## 48 56 280 64 136 88 128 136 264 144 328 288

Ranks<-prop.table(table(hotel$CityRank))  
 barplot(Ranks,main="Bookings by City Ranks",xlab="City Ranks",ylab="Count",col="lightpink")

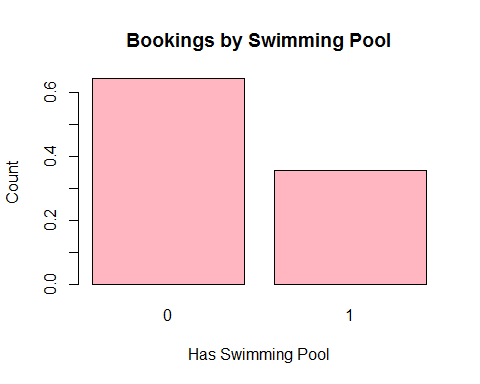


#More than 15% bookings are there in City with Rank 1 which is Delhi (Capital City)

#Table for HasSwimmingPool  
 table(hotel$HasSwimmingPool)

##   
## 0 1   
## 8524 4708

Pool<-prop.table(table(hotel$HasSwimmingPool))  
 barplot(Pool,main="Bookings by Swimming Pool",xlab="Has Swimming Pool",ylab="Count",col="lightpink")



# We have more than 60% of the bookings in hotels without swimming pools.

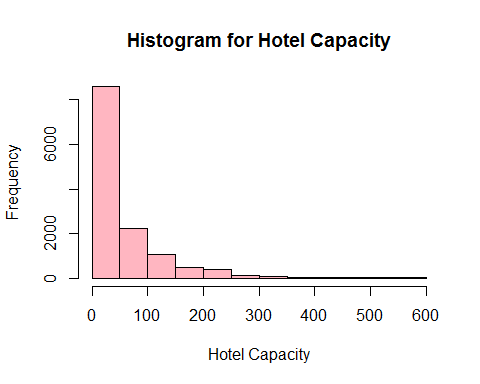
str(hotel$HotelCapacity)

## int [1:13232] 40 88 88 88 88 88 88 88 88 112 ...

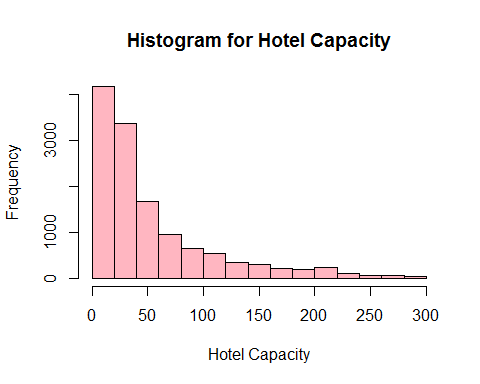
#Since hotel capacity is numeric in nature, we would make a histogram and box plot fot it.   
  
summary(hotel$HotelCapacity)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 16.00 34.00 62.51 75.00 600.00

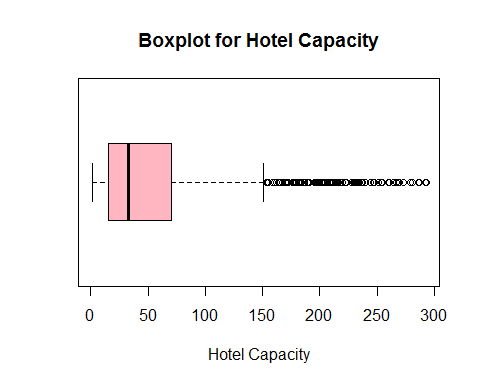
#The median capacity of hotels is 62 while max may even go upto 600.  
  
#Histogram for Hotel Capacity  
  
hist(hotel$HotelCapacity, main="Histogram for Hotel Capacity",xlab="Hotel Capacity",col="lightpink")



#This looks like a right skewed distribution as most of the hotels have capacity less than 50 with several outliers at the right end of the chart.   
  
#Excluding outliers or hotels with capacity > 300 and less than 1 or better results.  
  
hotel <- subset(hotel,hotel$HotelCapacity <= 300 & hotel$HotelCapacity > 0)  
hist(hotel$HotelCapacity, main="Histogram for Hotel Capacity",xlab="Hotel Capacity",col="lightpink")

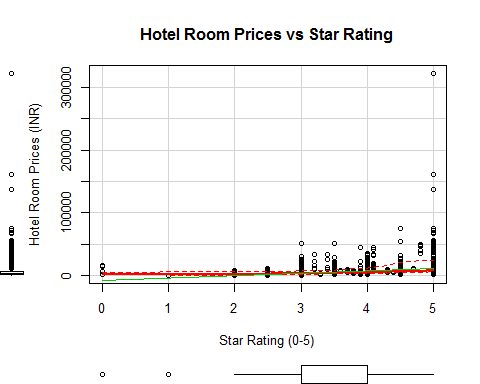


#BoxPlot for Hotel Capacity  
 boxplot(hotel$HotelCapacity, main="Boxplot for Hotel Capacity",horizontal = TRUE,xlab="Hotel Capacity",col="lightpink")



#The box plot clearly shows that there are a lot of outliers in the distribution.

#Scatter Plots to understand how are the variables correlated pair-wise  
  
#StarRating Vs RoomRent  
   
 scatterplot(hotel$StarRating,hotel$RoomRent,main="Hotel Room Prices vs Star Rating",ylab = "Hotel Room Prices (INR)", xlab="Star Rating (0-5)")

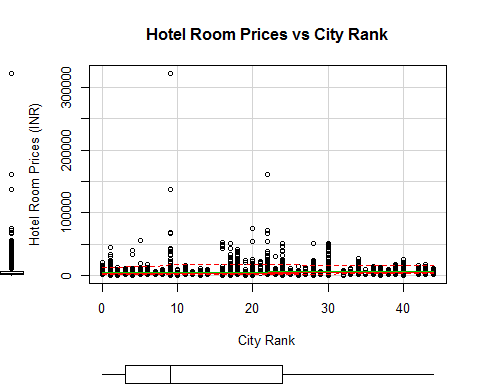


#The scatter plot shows a positive trend in the room prices with increase in star ratings.  
  
#Let's confirm with a correlation test.  
  
cor.test(hotel$StarRating,hotel$RoomRent)

##   
## Pearson's product-moment correlation  
##   
## data: hotel$StarRating and hotel$RoomRent  
## t = 45.02, df = 12944, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3529547 0.3827439  
## sample estimates:  
## cor   
## 0.3679437

#The p value being <2.2e-16 suggests that the result is statistically significant and that 'star rating and room rent affect are independent of each other' can be rejected. Now, the strength of the correlation is moderate with 0.37 and there is a positive correlation between the two.

#CityRank Vs RoomRent  
   
 scatterplot(hotel$CityRank,hotel$RoomRent,main="Hotel Room Prices vs City Rank",ylab = "Hotel Room Prices (INR)", xlab="City Rank")

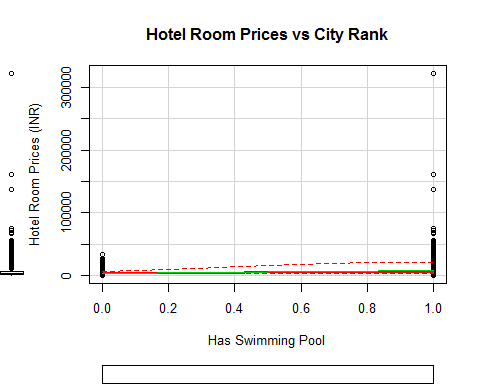


#The scatter plot does not show a clear trend in the room prices with better city ranks.  
  
#Let's do a correlation test.  
  
cor.test(hotel$CityRank,hotel$RoomRent)

##   
## Pearson's product-moment correlation  
##   
## data: hotel$CityRank and hotel$RoomRent  
## t = 11.939, df = 12944, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.08729673 0.12137383  
## sample estimates:  
## cor   
## 0.1043659

#The p value being <2.2e-16 suggests that the result is statistically significant and that the null hypothesis 'city rank and room rent affect are independent of each other' can be rejected. Now, the strength of the correlation is weak with 0.135 and there is a positive correlation between the two.

#HasSwimmingPool Vs RoomRent  
   
 scatterplot(hotel$HasSwimmingPool,hotel$RoomRent,main="Hotel Room Prices vs City Rank",ylab = "Hotel Room Prices (INR)", xlab="Has Swimming Pool")

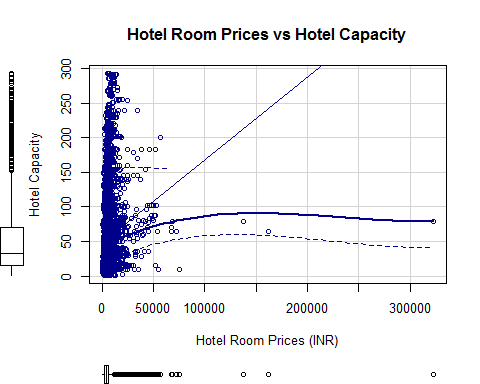


#The scatter plot shows a clear trend in the room prices with swimming pools with the rates being on the higher end.  
  
#Let's do a correlation test to confirm this.  
  
cor.test(hotel$HasSwimmingPool,hotel$RoomRent)

##   
## Pearson's product-moment correlation  
##   
## data: hotel$HasSwimmingPool and hotel$RoomRent  
## t = 36.657, df = 12944, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2909831 0.3221961  
## sample estimates:  
## cor   
## 0.3066721

#The p value being <2.2e-16 suggests that the result is statistically significant and that the null hypothesis 'hotel has swimming pool and room rent affect are independent of each other' can be rejected. Now, the strength of the correlation is moderate with 0.37 and there is a positive correlation between the two.

#RoomRent Vs HotelCapacity  
   
 scatterplot(hotel$RoomRent,hotel$HotelCapacity,main="Hotel Room Prices vs Hotel Capacity",ylab = "Hotel Capacity", xlab="Hotel Room Prices (INR)",col="darkblue")

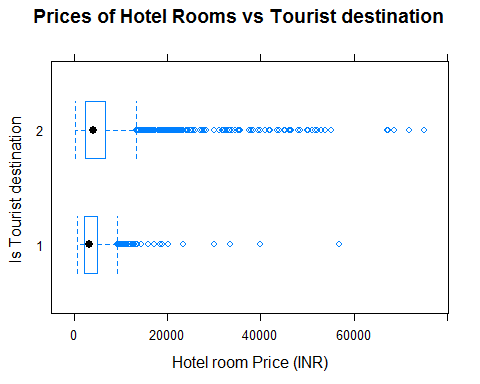


#There is a lot of variabity in the hotel room prices and hotel capacity.  
  
#Let's confirm with a correlation test.  
  
cor.test(hotel$HotelCapacity,hotel$RoomRent)

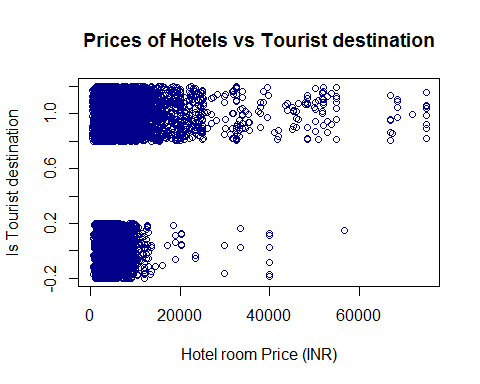
##   
## Pearson's product-moment correlation  
##   
## data: hotel$HotelCapacity and hotel$RoomRent  
## t = 18.205, df = 12944, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1411585 0.1747510  
## sample estimates:  
## cor   
## 0.1580005

#The p value being <2.2e-16 suggests that the result is statistically significant and that 'hotel capacity and room rent affect are independent of each other' can be rejected. Now, the strength of the correlation is weak with 0.16 and a positive correlation is seen between the two.

#RoomRent Vs IsTouristDestination  
  
#Excluding outlier with Room rents more than 100000 -  
  
hotel <- subset(hotel,hotel$RoomRent <= 100000)  
   
#Boxplot of Tourist Destination vs Room Rent  
  
bwplot(IsTouristDestination~RoomRent, data = hotel,main="Prices of Hotel Rooms vs Tourist destination",ylab = "Is Tourist destination ", xlab="Hotel room Price (INR)" )



#Jitter Plot of Tourist Destination vs Room Rent  
  
plot(hotel$RoomRent,jitter(hotel$IsTouristDestination),main="Prices of Hotels vs Tourist destination",ylab = "Is Tourist destination", xlab="Hotel room Price (INR)",col="darkblue")

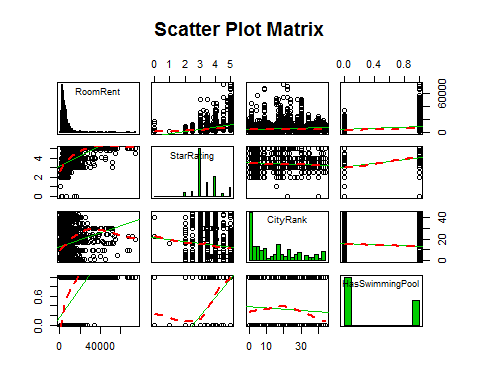


#Both the plots clearly confirms that being a tourist destination shows a positive trend in increase of hotel room prices.  
  
#Let's confirm with a correlation test.  
  
cor.test(hotel$IsTouristDestination,hotel$RoomRent)

##   
## Pearson's product-moment correlation  
##   
## data: hotel$IsTouristDestination and hotel$RoomRent  
## t = 16.613, df = 12939, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1276055 0.1613450  
## sample estimates:  
## cor   
## 0.1445172

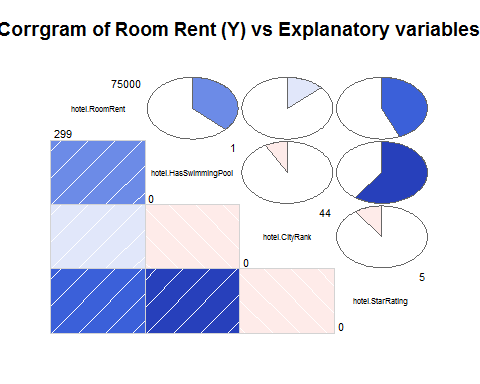
#The p value being <2.2e-16 suggests that the result is statistically significant and that 'being in a tourist destination and room rents are independent of each other' can be rejected. Now, the strength of the correlation is weak with 0.14 and a positive correlation is seen between the two.

#Scatterplot matrix between Y and x1,x2,x3  
   
 scatterplotMatrix(  
 hotel[  
 ,c("RoomRent","StarRating","CityRank","HasSwimmingPool" )],   
 spread=FALSE, smoother.args=list(lty=2),  
 main="Scatter Plot Matrix", diagonal = "histogram")



#With the results of the scatter plot matrix , we can see that with higher Star ratings, we see higher room rates, similarly, if the hotel has a swimming pool, we see higher room rates. But, for city rank, no such clear inferences can be drawn.

#Corrgram of Y, x1, x2, x3  
subset <- data.frame(hotel$RoomRent, hotel$HasSwimmingPool, hotel$CityRank, hotel$StarRating)  
   
corrgram(subset, lower.panel=panel.shade, upper.panel=panel.pie,  
 diag.panel=panel.minmax, text.panel=panel.txt,  
 main="Corrgram of Room Rent (Y) vs Explanatory variables")



#Correlation Matrix  
  
library(Hmisc)

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

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

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

## The following objects are masked from 'package:dplyr':  
##   
## combine, src, summarize

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

colroom <- c("RoomRent", "HasSwimmingPool", "CityRank", "StarRating")  
corMatrix <- rcorr(as.matrix(hotel[,colroom]))  
corMatrix

## RoomRent HasSwimmingPool CityRank StarRating  
## RoomRent 1.00 0.37 0.13 0.44  
## HasSwimmingPool 0.37 1.00 -0.08 0.60  
## CityRank 0.13 -0.08 1.00 -0.10  
## StarRating 0.44 0.60 -0.10 1.00  
##   
## n= 12941   
##   
##   
## P  
## RoomRent HasSwimmingPool CityRank StarRating  
## RoomRent 0 0 0   
## HasSwimmingPool 0 0 0   
## CityRank 0 0 0   
## StarRating 0 0 0

#Another way to do this -  
  
#Variance-Covariance Matrix  
  
x<-hotel[,c("HasSwimmingPool","StarRating", "CityRank")]  
 y<-hotel[,c("RoomRent")]  
 cor(x,y)

## [,1]  
## HasSwimmingPool 0.3694289  
## StarRating 0.4365300  
## CityRank 0.1349080

cov(x,y)

## [,1]  
## HasSwimmingPool 1020.168  
## StarRating 1859.105  
## CityRank 10619.332

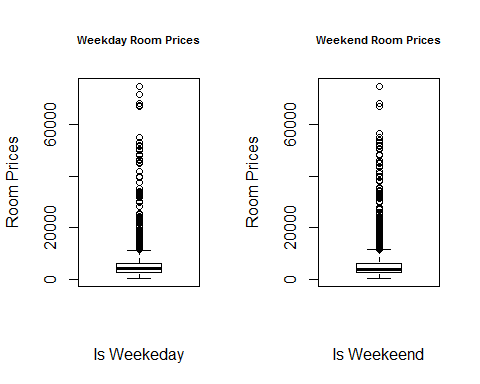
#We prefer using correlations because covariances are hard to compare and thus we normalise these covariances and come up with correlations which lie between -1 to 1, thus making it easier to compare two variables.

A. Hypothesis - H0: The variables ISWeekend,IsTouristDestination,HasSwimmingPool,StarRating,CityRank,HotelCapacity,Airport,IsNewYear,IsMetroCity collectively have no effect on RoomRent. H1: ISWeekend,IsTouristDestination,HasSwimmingPool,StarRating,CityRank,HotelCapacity,Airport,IsNewYear,IsMetroCity together affect RoomRent

B. Hypothesis - H0 : The variables IsTouristDestination,HasSwimmingPool,StarRating,CityRank,Airport ,IsNewYearEve collectively have no effect on RoomRent. H1 : IsTouristDestination,HasSwimmingPool,StarRating,CityRank,Airport ,IsNewYearEve together affect RoomRent

Let's carry out some t-tests :

#Hypothesis 1   
#H1 = On weekends, we have higher room prices  
#H0= Room prices do not vary with being a weekend/weekday  
  
#Let's test this using a t-test:  
#We first subset the booking made on a weekday and those on a weekend.  
  
weekday <- subset(hotel,IsWeekend==0)  
weekend <- subset(hotel,IsWeekend==1)  
par(mfrow=c(1,2))  
boxplot(weekday$RoomRent,main="Weekday Room Prices", xlab="Is Weekeday",ylab="Room Prices",cex.main=0.7)  
boxplot(weekend$RoomRent,main="Weekend Room Prices",xlab="Is Weekeend",ylab="Room Prices",cex.main=0.7)



var.test(weekday$RoomRent,weekend$RoomRent)

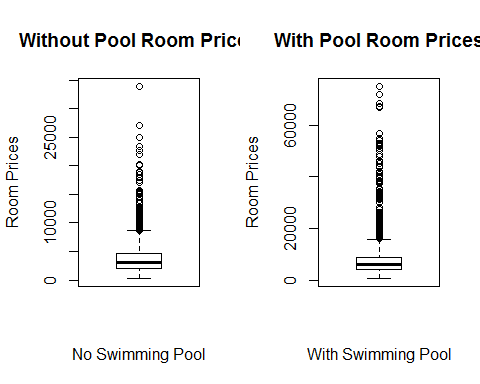
##   
## F test to compare two variances  
##   
## data: weekday$RoomRent and weekend$RoomRent  
## F = 0.98018, num df = 4890, denom df = 8049, p-value = 0.4365  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.9322745 1.0308605  
## sample estimates:  
## ratio of variances   
## 0.9801758

t.test(weekday$RoomRent,weekend$RoomRent,var.equal = TRUE,alternative="less")

##   
## Two Sample t-test  
##   
## data: weekday$RoomRent and weekend$RoomRent  
## t = -0.79432, df = 12939, p-value = 0.2135  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 89.77604  
## sample estimates:  
## mean of x mean of y   
## 5275.122 5358.954

#Since the p-value is 0.2135 which is above the significance level of 0.05, we fail to reject the null hypothesis of independence between weekend and room prices.

#Hypothesis 2  
#H1 = With Swimming Pool, the room prices are higher  
#H0= Room prices do not vary with having a Swimming pool or not   
  
#Let's test this using a t-test:  
#We first subset the bookings with and without swimming pool.  
  
no\_pool <- subset(hotel,HasSwimmingPool==0)  
pool <- subset(hotel,HasSwimmingPool==1)  
  
par(mfrow=c(1,2))  
boxplot(no\_pool$RoomRent,main="Without Pool Room Prices",xlab="No Swimming Pool",ylab="Room Prices")  
boxplot(pool$RoomRent,main="With Pool Room Prices",xlab="With Swimming Pool",ylab="Room Prices")



var.test(no\_pool$RoomRent,pool$RoomRent)

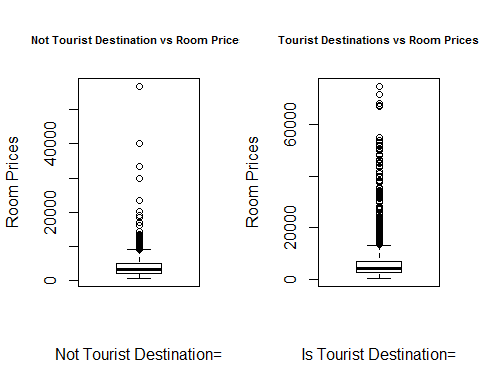
##   
## F test to compare two variances  
##   
## data: no\_pool$RoomRent and pool$RoomRent  
## F = 0.10008, num df = 8515, denom df = 4424, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.09505076 0.10533679  
## sample estimates:  
## ratio of variances   
## 0.1000829

t.test(no\_pool$RoomRent,pool$RoomRent,alternative="less",var.equal=FALSE)

##   
## Welch Two Sample t-test  
##   
## data: no\_pool$RoomRent and pool$RoomRent  
## t = -34.707, df = 4889.2, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -4318.516  
## sample estimates:  
## mean of x mean of y   
## 3777.132 8310.541

#Since this gives a p-value of < 2.2e-16, hence we can reject our null hypothesis in favour of the alternate hypothesis that with swimming pool, the room prices are higher.

#Hypothesis 3  
#H1 = When cities are tourist destinations, the room prices are higher  
#H0= Room prices do not vary with being a tourist destination or not  
  
#Let's test this using a t-test:  
#We first subset the booking made in tourist destinations & those which aren't.  
  
not\_touristy <- subset(hotel,IsTouristDestination==0)  
touristy <- subset(hotel,IsTouristDestination==1)  
  
par(mfrow=c(1,2))  
boxplot(not\_touristy$RoomRent,main="Not Tourist Destination vs Room Prices",cex.main=0.7,xlab="Not Tourist Destination=",ylab="Room Prices")  
boxplot(touristy$RoomRent,main="Tourist Destinations vs Room Prices",cex.main=0.7,xlab="Is Tourist Destination=",ylab="Room Prices")

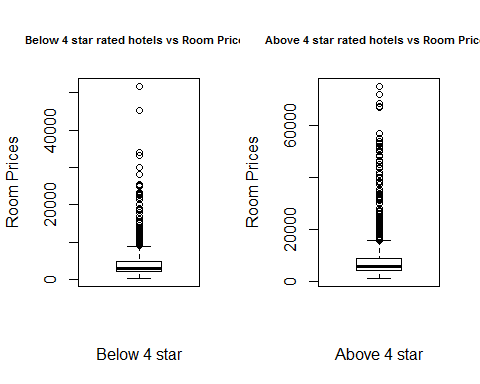


t.test(not\_touristy$RoomRent,touristy$RoomRent,alternative="less")

##   
## Welch Two Sample t-test  
##   
## data: not\_touristy$RoomRent and touristy$RoomRent  
## t = -21.628, df = 12908, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -1691.584  
## sample estimates:  
## mean of x mean of y   
## 4050.878 5881.712

#Since this gives a p-value of < 2.2e-16, hence we can reject our null hypothesis in favour of the alternate hypothesis tourist destinations have higher room prices.

#Hypothesis 4  
#H1 = Are better star rated hotels better priced?  
#H0 = Prices do not vary with Star Rating  
  
#Let's test this using a t-test:  
#We create two buckets, one with bookings made in less than 4 star rated rotels & another bookings with more than 4 star rated hotels.  
  
below\_4\_star <- subset(hotel,StarRating<4)  
above\_4\_star <- subset(hotel,StarRating>=4)  
  
par(mfrow=c(1,2))  
boxplot(below\_4\_star$RoomRent,main="Below 4 star rated hotels vs Room Prices",cex.main=0.7,xlab="Below 4 star",ylab="Room Prices")  
boxplot(above\_4\_star$RoomRent,main="Above 4 star rated hotels vs Room Prices",cex.main=0.7,xlab="Above 4 star",ylab="Room Prices")

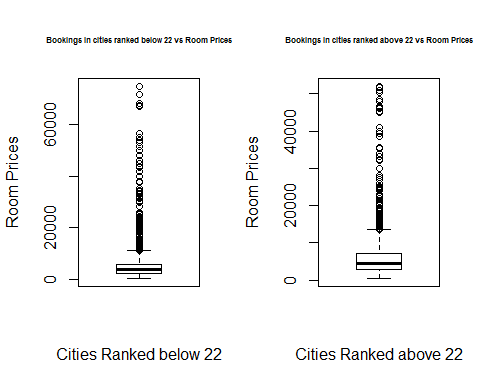


t.test(below\_4\_star$RoomRent,above\_4\_star$RoomRent,alternative="less")

##   
## Welch Two Sample t-test  
##   
## data: below\_4\_star$RoomRent and above\_4\_star$RoomRent  
## t = -33.839, df = 4551, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -4426.69  
## sample estimates:  
## mean of x mean of y   
## 3873.260 8526.163

#Since this gives a p-value of < 2.2e-16, hence we can reject our null hypothesis in favour of the alternate hypothesis that better star rated hotels have higher room rates.

#Hypothesis 5  
#H1 = Room prices with lesser city ranks are higher  
#H0 = Room prices with lesser city ranks are not higher  
  
#Let's test this using a t-test:  
#We create two buckets, one with bookings made in cities ranked below 22 & another with bookings made in cities ranked more than 22.  
  
below\_10 <- subset(hotel,CityRank<=22)  
above\_10 <- subset(hotel,CityRank>22)  
  
par(mfrow=c(1,2))  
boxplot(below\_10$RoomRent,main="Bookings in cities ranked below 22 vs Room Prices",cex.main=0.5,xlab="Cities Ranked below 22",ylab="Room Prices")  
boxplot(above\_10$RoomRent,main="Bookings in cities ranked above 22 vs Room Prices",cex.main=0.5,xlab="Cities Ranked above 22",ylab="Room Prices")



#Our one tailed t-test is as follows -   
  
t.test(below\_10$RoomRent,above\_10$RoomRent,alternative="greater")

##   
## Welch Two Sample t-test  
##   
## data: below\_10$RoomRent and above\_10$RoomRent  
## t = -9.2488, df = 6678.7, p-value = 1  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -1260.681 Inf  
## sample estimates:  
## mean of x mean of y   
## 5009.677 6079.983

#Since this gives a p-value of 1, which although is very unlikely and also lies way above our significance levels, hence we fail to reject our null hypothesis.  
  
#Although, if we do a two tailed test with hypothesis being :  
#Hypothesis 6  
#H1 = Room prices vary with city ranks  
#H0 = Room prices do not vary with city ranks  
  
t.test(below\_10$RoomRent,above\_10$RoomRent)

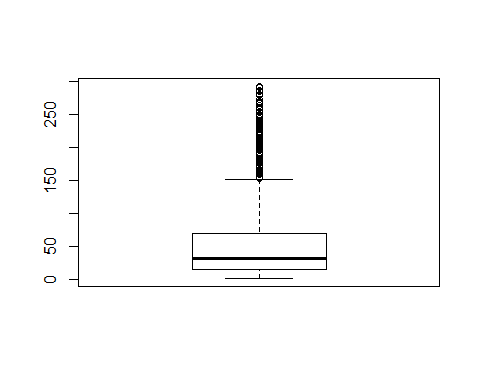
##   
## Welch Two Sample t-test  
##   
## data: below\_10$RoomRent and above\_10$RoomRent  
## t = -9.2488, df = 6678.7, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1297.1610 -843.4508  
## sample estimates:  
## mean of x mean of y   
## 5009.677 6079.983

#Since this gives a p-value of < 2.2e-16, hence we can reject our null hypothesis in favour of the alternate hypothesis that Room prices vary with city ranks.

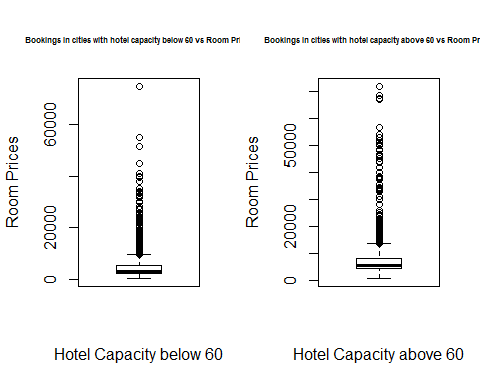
#Hypothesis 6  
#H1 = Hotels with higher capacities have higher room prices  
#H0 = Room prices do not vary with Hotel Capacity  
  
summary(hotel$HotelCapacity)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 15.00 32.00 54.99 70.00 293.00

boxplot(hotel$HotelCapacity)



#Let's test this using a t-test:  
#We create two buckets, one with bookings made in hotels with capacity below 60 & another with capacity more than 60.  
  
below\_60 <- subset(hotel,HotelCapacity<=60)  
above\_60 <- subset(hotel,HotelCapacity>60)  
  
par(mfrow=c(1,2))  
boxplot(below\_60$RoomRent,main="Bookings in cities with hotel capacity below 60 vs Room Prices",cex.main=0.5,xlab="Hotel Capacity below 60",ylab="Room Prices")  
boxplot(above\_60$RoomRent,main="Bookings in cities with hotel capacity above 60 vs Room Prices",cex.main=0.5,xlab="Hotel Capacity above 60",ylab="Room Prices")



t.test(below\_60$RoomRent,above\_60$RoomRent,alternative="less")

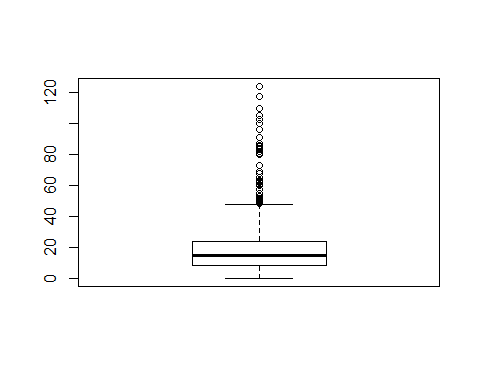
##   
## Welch Two Sample t-test  
##   
## data: below\_60$RoomRent and above\_60$RoomRent  
## t = -22.197, df = 5184.9, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -2644.527  
## sample estimates:  
## mean of x mean of y   
## 4506.888 7363.106

#Since this gives a p-value of < 2.2e-16, hence we can reject our null hypothesis in favour of the alternate hypothesis that higher capacity hotels have higher room rates.

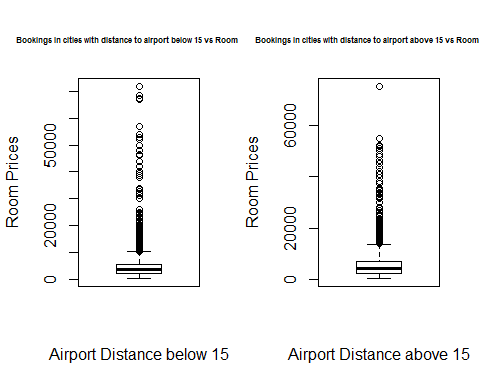
#Hypothesis 7  
#H1 = Hotels with higher distances to airport have higher prices (Probably because airports are outside the city)  
#H0 = Room prices do not vary with distance to Airport  
  
summary(hotel$Airport)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.20 8.50 15.00 21.37 24.30 124.00

boxplot(hotel$Airport)



#Let's test this using a t-test:  
#We create two buckets, one with bookings made in hotels with distance to airport less than 15 kms & another with distance to airport more than 15 kms.  
  
below\_15 <- subset(hotel,Airport<=15)  
above\_15 <- subset(hotel,Airport>15)  
  
par(mfrow=c(1,2))  
boxplot(below\_15$RoomRent,main="Bookings in cities with distance to airport below 15 vs Room Prices",cex.main=0.5,xlab="Airport Distance below 15",ylab="Room Prices")  
boxplot(above\_15$RoomRent,main="Bookings in cities with distance to airport above 15 vs Room Prices",cex.main=0.5,xlab="Airport Distance above 15",ylab="Room Prices")



t.test(below\_15$RoomRent,above\_15$RoomRent,alternative="less")

##   
## Welch Two Sample t-test  
##   
## data: below\_15$RoomRent and above\_15$RoomRent  
## t = -11.086, df = 12349, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -964.1329  
## sample estimates:  
## mean of x mean of y   
## 4768.255 5900.369

#Since this gives a p-value of < 2.2e-16, hence we can reject our null hypothesis in favour of the alternate hypothesis that hotels with greater distance to airport have higher room rates.

We will now make several models and then pick the best of them based on AIC,BIC and R2 values.

#Let's make a regression model for hypothesis A.  
  
model <- lm(log(RoomRent) ~ IsWeekend+IsTouristDestination+HasSwimmingPool+StarRating+CityRank+HotelCapacity+Airport +IsNewYearEve+IsMetroCity,data=hotel)  
  
summary(model)

##   
## Call:  
## lm(formula = log(RoomRent) ~ IsWeekend + IsTouristDestination +   
## HasSwimmingPool + StarRating + CityRank + HotelCapacity +   
## Airport + IsNewYearEve + IsMetroCity, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.00650 -0.35263 -0.04441 0.30027 2.55806   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.3734848 0.0298893 213.236 < 2e-16 \*\*\*  
## IsWeekend -0.0126642 0.0102293 -1.238 0.216   
## IsTouristDestination 0.1341822 0.0121226 11.069 < 2e-16 \*\*\*  
## HasSwimmingPool 0.3496599 0.0131966 26.496 < 2e-16 \*\*\*  
## StarRating 0.4400102 0.0092474 47.582 < 2e-16 \*\*\*  
## CityRank 0.0083505 0.0005351 15.606 < 2e-16 \*\*\*  
## HotelCapacity -0.0001228 0.0001182 -1.039 0.299   
## Airport 0.0032378 0.0002437 13.286 < 2e-16 \*\*\*  
## IsNewYearEve 0.1025027 0.0150330 6.818 9.61e-12 \*\*\*  
## IsMetroCity 0.0138648 0.0142369 0.974 0.330   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5394 on 12931 degrees of freedom  
## Multiple R-squared: 0.4214, Adjusted R-squared: 0.421   
## F-statistic: 1046 on 9 and 12931 DF, p-value: < 2.2e-16

exp(model$coefficients)

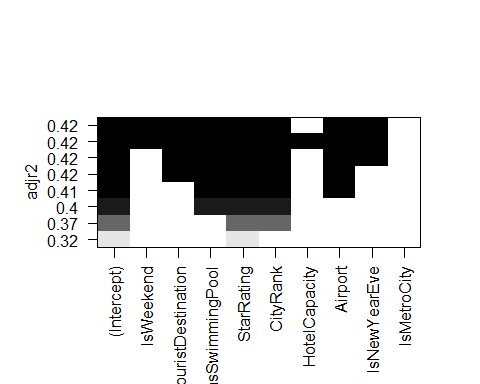
## (Intercept) IsWeekend IsTouristDestination   
## 586.0966921 0.9874157 1.1436012   
## HasSwimmingPool StarRating CityRank   
## 1.4185850 1.5527230 1.0083854   
## HotelCapacity Airport IsNewYearEve   
## 0.9998772 1.0032430 1.1079403   
## IsMetroCity   
## 1.0139614

#Here, the model equation wil look like this -  
#log(y)=B1\*x1 +B2\*x2 ..  
#log(RoomRent) = 6.3734848 + 0.0102293\* IsWeekend + 0.1341822\*IsTouristDestination + 0.3496599\*HasSwimmingPool + 0.4400102 \*StarRating + 0.0083505\*CityRank -0.0001228\*HotelCapacity + 0.0032378\*Airport + 0.1025027\*IsNewYearEve + 0.0138648\*IsMetroCity + e  
  
#Multiple R-squared: 0.4214, Adjusted R-squared: 0.421   
#F-statistic: 1046 on 9 and 12931 DF, p-value: < 2.2e-16  
  
#Since the model's p value is < 2.2e-16, hence we can reject the null hypothesis that the variables collectively do not explain the variation in the RoomRent.  
  
#The results also show that the variable HasSwimmingPool is significant in controlling for the variable RoomRent (p = 2e-16), as are the rest of them except Hotel Capacity,IsMetroCity and IsWeekend.

library(leaps)

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

leap1 <- regsubsets(log(RoomRent) ~ IsWeekend+IsTouristDestination+HasSwimmingPool+StarRating+CityRank+HotelCapacity+Airport +IsNewYearEve+IsMetroCity, data = hotel, nbest=1)  
  
# summary(leap1)  
  
plot(leap1, scale="adjr2")



#The best fit model excludes IsMetrocCity,HotelCapacity,IsWeekend. Therefore, in our next model, we rerun the regression, excluding these variables.

Let's make a regression model for hypothesis B.

model\_1 <- lm(log(RoomRent) ~ IsTouristDestination+HasSwimmingPool+StarRating+CityRank+Airport +IsNewYearEve,data=hotel)  
summary(model\_1)

##   
## Call:  
## lm(formula = log(RoomRent) ~ IsTouristDestination + HasSwimmingPool +   
## StarRating + CityRank + Airport + IsNewYearEve, data = hotel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.99685 -0.35446 -0.04326 0.29898 2.54514   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.3780901 0.0279324 228.340 < 2e-16 \*\*\*  
## IsTouristDestination 0.1409580 0.0108714 12.966 < 2e-16 \*\*\*  
## HasSwimmingPool 0.3448546 0.0126704 27.217 < 2e-16 \*\*\*  
## StarRating 0.4358485 0.0081614 53.404 < 2e-16 \*\*\*  
## CityRank 0.0081063 0.0004185 19.370 < 2e-16 \*\*\*  
## Airport 0.0032428 0.0002427 13.363 < 2e-16 \*\*\*  
## IsNewYearEve 0.0971002 0.0143751 6.755 1.49e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5394 on 12934 degrees of freedom  
## Multiple R-squared: 0.4212, Adjusted R-squared: 0.421   
## F-statistic: 1569 on 6 and 12934 DF, p-value: < 2.2e-16

exp(model\_1$coefficients)

## (Intercept) IsTouristDestination HasSwimmingPool   
## 588.802094 1.151376 1.411785   
## StarRating CityRank Airport   
## 1.546274 1.008139 1.003248   
## IsNewYearEve   
## 1.101971

#Here, the model equation wil look like this -  
#log(y)=B1\*x1 +B2\*x2 ..  
#log(RoomRent) = 6.3780901 + 0.1409580\*IsTouristDestination + 0.3448546\*HasSwimmingPool + 0.4358485 \*StarRating + 0.0081063\*CityRank + 0.0032428\*Airport+ 0.0971002\*IsnewYearEve + e  
  
#Multiple R-squared: 0.4212, Adjusted R-squared: 0.421   
#F-statistic: 1569 on 6 and 12934 DF, p-value: < 2.2e-16  
  
#All variables are significant in this model.  
#Since the model's p value is < 2.2e-16, hence we can reject the null hypothesis that the variables collectively do not explain the variation in the RoomRent.

#Let's choose the better model out of these two -  
  
AIC(model)

## [1] 20761.41

BIC(model)

## [1] 20843.56

#AIC = 20761.41 , BIC = 20843.56, Multiple R-squared: 0.4214, Adjusted R-squared: 0.421  
  
AIC(model\_1)

## [1] 20758.84

BIC(model\_1)

## [1] 20818.59

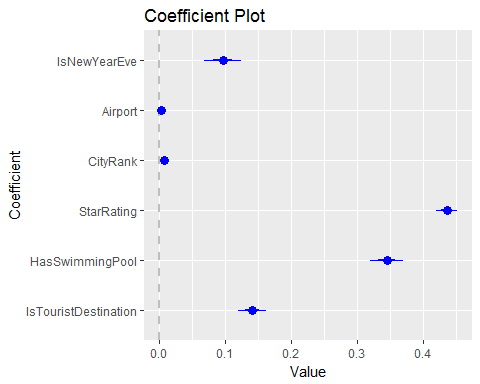
#AIC = 20758.84 , BIC = 20818.59, Multiple R-squared: 0.4212, Adjusted R-squared: 0.421   
  
#Based on these AIC, BIC values & R2 values where R-squared measures the percent of variation in Y explained by variation in X (or combination of Xs), we would choose model\_1 as it has better AIC and BIC values while R2 is the same. Also, by using regsubset, we had removed the insignificant variables which might have resulted in these slightly better results.

#Visualize the beta coefficients and their confidence intervals from model 1 -  
  
library(coefplot)

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

coefplot(model\_1, intercept=FALSE,outerCI=1.96,coefficients=c("IsTouristDestination","HasSwimmingPool","StarRating","CityRank","Airport","IsNewYearEve"))

## Warning: Ignoring unknown aesthetics: xmin, xmax



For inferences, testing a subset of variables using partial F-Test.

We are doing this as we are interested in simultaneously testing whether a certain subset of the coefficients are equal to 0 (e.g. B3 = B4 = 0). We can do this using a partial F-test. This test involves comparing the SSE from a reduced model (excluding the parameters we hypothesis are equal to zero) with the SSE from the full model (including all of the parameters).

We will do this by performing partial F-tests by fitting both the reduced and full mode and then comparing them using the anova function.

#According to model\_1, we are including the variables IsTouristDestination,HasSwimmingPool,StarRating,CityRank,HotelCapacity,Airport in our model and we will be testing whether the airport, IsTouristDestination are significant after taking HotelCapacity,HasSwimmingPool,StarRating,CityRank into consideration.  
  
#The following code performs the partial F-test:  
  
# Reduced model  
reduced = lm(log(RoomRent) ~ IsTouristDestination+HasSwimmingPool+StarRating+CityRank, data=hotel)   
# Full Model  
full = lm(log(RoomRent) ~ IsTouristDestination+HasSwimmingPool+StarRating+CityRank+Airport+IsNewYearEve,data=hotel)   
  
anova(reduced, full)

## Analysis of Variance Table  
##   
## Model 1: log(RoomRent) ~ IsTouristDestination + HasSwimmingPool + StarRating +   
## CityRank  
## Model 2: log(RoomRent) ~ IsTouristDestination + HasSwimmingPool + StarRating +   
## CityRank + Airport + IsNewYearEve  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12936 3829.0   
## 2 12934 3763.7 2 65.298 112.2 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The output shows the results of the partial F-test. Since F=112.2 (p-value< 2.2e-16) we reject the null hypothesis ( B5 = B6 = 0) at the 5% level of significance. It appears that the variables IsNewYearEve and Airport do contribute significant information to the room rent prices once the other variables from the model like IsTouristDestination ,HasSwimmingPool , StarRating ,CityRank have been taken into consideration.

#Another Partial F-Test -  
   
# Reduced model  
reduced = lm(log(RoomRent) ~ IsTouristDestination+Airport+HasSwimmingPool+StarRating+CityRank, data=hotel)   
# Full Model  
full = lm(log(RoomRent) ~ IsTouristDestination+HasSwimmingPool+StarRating+CityRank+Airport+IsNewYearEve,data=hotel)   
  
anova(reduced, full)

## Analysis of Variance Table  
##   
## Model 1: log(RoomRent) ~ IsTouristDestination + Airport + HasSwimmingPool +   
## StarRating + CityRank  
## Model 2: log(RoomRent) ~ IsTouristDestination + HasSwimmingPool + StarRating +   
## CityRank + Airport + IsNewYearEve  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12935 3777.0   
## 2 12934 3763.7 1 13.277 45.627 1.492e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Similarly, the output shows the results of the partial F-test. Since F=45.627 (p-value=1.492e-11)  
#we reject the null hypothesis ( B6 = 0) at the 5% level of significance.  
#It appears that the variables IsNewYearEve do contribute significant information to the room rent prices once the other variables from the model like IsTouristDestination ,HasSwimmingPool , StarRating ,CityRank,IsTouristDestination have been taken into consideration.

Confidence and prediction Intervals

Some brief on how we are using these functions : The function predict() can be used to make both confidence intervals for the mean response and prediction intervals.

To make confidence intervals for the mean response use the option interval="confidence".

To make a prediction interval use the option interval="prediction". By default this makes 95% confidence and prediction intervals. If you instead want to make a 99% confidence or prediction interval use the option level=0.99.

dim(hotel)

## [1] 12941 19

#Let's split the data set with some observations for test dataset  
train <- hotel[1:12900,]  
test <- hotel[12901:12941,]  
  
  
#Taking, log of the price as it should not come to be negative.  
#Log Linear model  
results = lm(log(RoomRent) ~ IsTouristDestination+HasSwimmingPool+StarRating+CityRank+Airport+IsNewYearEve,data=train)  
  
predict(results,data.frame(IsTouristDestination=0,HasSwimmingPool=0,StarRating=2,CityRank=3,Airport=9.6,IsNewYearEve=0),interval="confidence")

## fit lwr upr  
## 1 7.305291 7.278233 7.332348

#Exponentiating back the results  
exp(7.278233) #1448.426

## [1] 1448.426

exp(7.332348) #1528.968

## [1] 1528.968

#A 95% confidence interval is given by (1448.426, 1528.968)  
# With a 95% confidence interval, the results after exponentitaion that we get are really close to the actual room price of 1468. Thus, our model seems to be doing good.

predict(results,data.frame(IsTouristDestination=0,HasSwimmingPool=0,StarRating=2,CityRank=3,Airport=9.6,IsNewYearEve=0),interval="prediction")

## fit lwr upr  
## 1 7.305291 6.246715 8.363867

exp(6.25586) #521.0573

## [1] 521.0573

exp(8.376676) #4344.544

## [1] 4344.544

#A 95% prediction interval is given by (521.0573, 4344.544).  
#The results obtained for prediction have a wider range than confidence interval indicating that the variation about the mean is fairly large.

#Applyting this model in a generic manner to full testing data which is split in a a 70:30 ratio -  
  
# For reproducibility; 123 has no particular meaning  
set.seed(123)   
   
# randomly pick 70% of the number of observations (365)  
index <- sample(1:nrow(hotel),size = 0.7\*nrow(hotel))   
   
# subset weather to include only the elements in the index  
train <- hotel[index,]   
   
# subset weather to include all but the elements in the index  
test <- hotel[-index,]   
   
nrow(train)

## [1] 9058

nrow(test)

## [1] 3883

#Evaluation metrics  
  
#Baseline model  
best.guess <- mean(train$RoomRent)  
RMSE.baseline <- sqrt(mean((best.guess-test$RoomRent)^2))  
RMSE.baseline #5288.912

## [1] 5288.912

MAE.baseline <- mean(abs(best.guess-test$RoomRent))  
MAE.baseline #3063.866

## [1] 3063.866

results = lm(log(RoomRent+1) ~ IsTouristDestination+HasSwimmingPool+StarRating+CityRank+IsNewYearEve+Airport,data=train)  
  
summary(results)

##   
## Call:  
## lm(formula = log(RoomRent + 1) ~ IsTouristDestination + HasSwimmingPool +   
## StarRating + CityRank + IsNewYearEve + Airport, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.98403 -0.36158 -0.04508 0.30082 2.55163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.3514125 0.0336610 188.688 < 2e-16 \*\*\*  
## IsTouristDestination 0.1533133 0.0130371 11.760 < 2e-16 \*\*\*  
## HasSwimmingPool 0.3460052 0.0152291 22.720 < 2e-16 \*\*\*  
## StarRating 0.4415043 0.0098715 44.725 < 2e-16 \*\*\*  
## CityRank 0.0082756 0.0005027 16.461 < 2e-16 \*\*\*  
## IsNewYearEve 0.0945483 0.0174737 5.411 6.43e-08 \*\*\*  
## Airport 0.0031727 0.0002914 10.888 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5436 on 9051 degrees of freedom  
## Multiple R-squared: 0.4215, Adjusted R-squared: 0.4211   
## F-statistic: 1099 on 6 and 9051 DF, p-value: < 2.2e-16

#Multiple R-squared: 0.4213, Adjusted R-squared: 0.421  
  
test.pred <- exp(predict(results,test))-1  
  
#Evaluating the accuracy -  
RMSE <- sqrt(mean((test.pred-test$RoomRent)^2))  
RMSE #1982.59

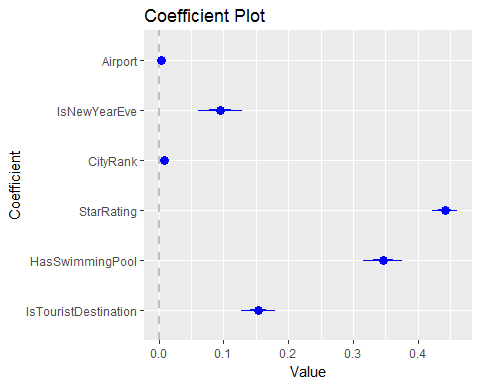
## [1] 4493.422

MAE.lin.reg <- mean(abs(test.pred-test$RoomRent))  
MAE.lin.reg #1472.424

## [1] 2188.373

library(coefplot)  
  
coefplot(results, intercept=FALSE,outerCI=1.96,coefficients=c("IsTouristDestination","HasSwimmingPool","StarRating","CityRank","Airport","IsNewYearEve"))

## Warning: Ignoring unknown aesthetics: xmin, xmax



Here are the main conclusions about the model we have just built:

1. The R-squared is 0.4213, which means that 42% of the variance in our dependent variable can be explained by the set of predictors in the model; at the same time, the adjusted R-squared is not far from that number, meaning that the original R-squared has not been artificially increased by adding variables to the model. Note that the R-squared can only increase or stay the same by adding variables, whereas the adjusted R-squared can even decrease if the variable added doesn't help the model more than what is expected by chance;
2. All the variables are statistically significant (p < 0.05) and the most significant predictor is the StarRating as seen from the coefficient plot. The advantage of doing a log transformation is that, if the regression coefficient is small (i.e. -0.1 to 0.1), a unit increase in the independent variable yields an increase of approximately coeff\*100% in the dependent variable.
3. To be clear, the coefficient of the Star Rating is 0.4369466. It means that a unit increase in the Star Rating (i.e., increasing the rating by 1 point), increases the predicted amount of Room Rent by approximately 44%. One can always exponentiate to get the exact value. By the same token, if a hotel has a swimming pool, the predicted RoomRent increases by 35%)
4. Both the RMSE and MAE have significantly decreased when compared with the baseline model, which means that this linear model, despite all the linearity issues and the fact that it predicts negative values of rain in some days, is still much better, overall, than our best guess.