# Analysis of Hotel 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 ...

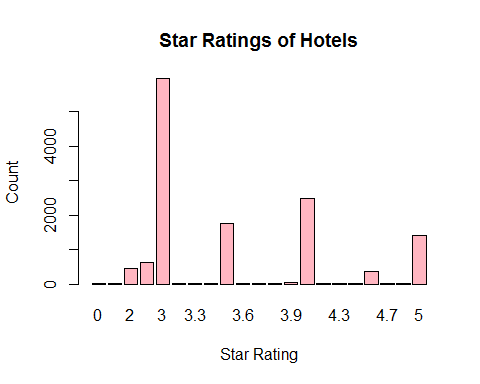
#Summarising dataset's mean,std dev , median etc.  
  
describe(hotel)[,c(2,3,4,5,8,9)]

## 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.  
  
#The three independent variables that are chosen would be StarRating, IsTouristDestination, Hotel Capacity.  
  
#Let's visualize the above mentioned variables Y,x1,x2,x3 where Y is Room rent and X1,X2,X3 are StarRating, IsTouristDestination and HotelCapacity 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 <- table(hotel$StarRating)  
 barplot(Rating,main = "Star Ratings of Hotels",xlab="Star Rating",ylab="Count",col="lightpink")

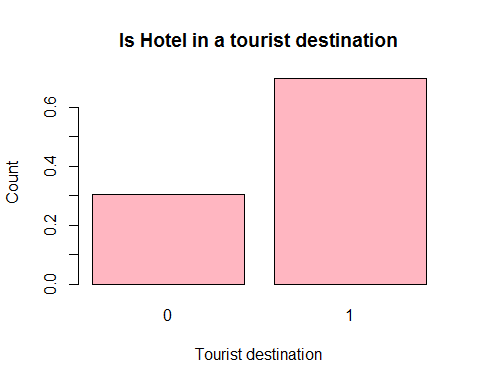


#We see that 3 star rated hotels are the most common with being close to 6000 in number.

#Table for IsTouristDestination  
 table(hotel$IsTouristDestination)

##   
## 0 1   
## 4007 9225

Touristy<-prop.table(table(hotel$IsTouristDestination))  
 barplot(Touristy,main="Is Hotel in a tourist destination",xlab="Tourist destination",ylab="Count",col="lightpink")



# We have more than 60% of the hotels in tourist destinations.  
   
#Table for city name and tourist destination to see how many tourist destinations we do have -  
   
 table(hotel$CityName,hotel$IsTouristDestination)

##   
## 0 1  
## Agra 0 432  
## Ahmedabad 424 0  
## Amritsar 0 136  
## Bangalore 656 0  
## Bhubaneswar 120 0  
## Chandigarh 336 0  
## Chennai 328 88  
## Darjeeling 0 136  
## Delhi 0 2048  
## Gangtok 0 128  
## Goa 0 624  
## Guwahati 0 48  
## Haridwar 0 48  
## Hyderabad 536 0  
## Indore 160 0  
## Jaipur 0 768  
## Jaisalmer 0 264  
## Jodhpur 0 224  
## Kanpur 16 0  
## Kochi 0 608  
## Kolkata 327 185  
## Lucknow 128 0  
## Madurai 0 112  
## Manali 0 288  
## Mangalore 104 0  
## Mumbai 0 712  
## Munnar 0 328  
## Mysore 0 160  
## Nainital 0 144  
## Ooty 0 136  
## Panchkula 64 0  
## Pune 600 0  
## Puri 0 56  
## Rajkot 128 0  
## Rishikesh 0 88  
## Shimla 0 280  
## Srinagar 0 40  
## Surat 80 0  
## Thiruvanthipuram 0 392  
## Thrissur 0 32  
## Udaipur 0 456  
## Varanasi 0 264

#we have 29 cities as Tourist destinations and rest are not. Which make them non-tourist destinations around only 33% of the data.

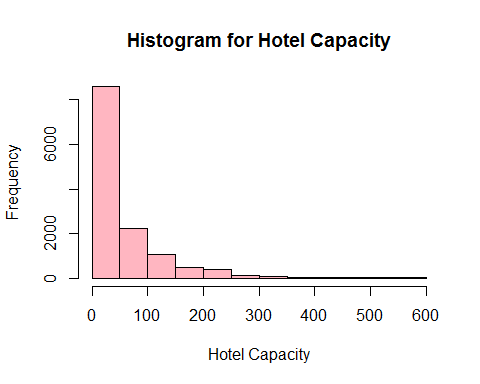
str(hotel$HotelCapacity)

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

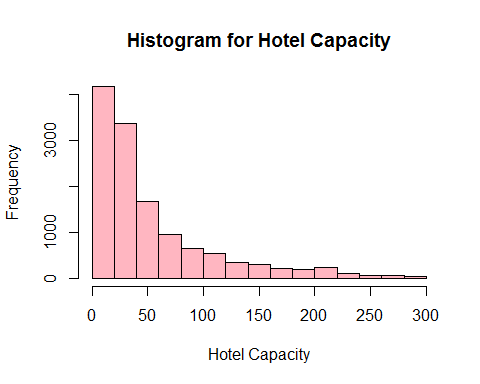
#Since hotel capacity is numeric in nature, we would make a histogram and box plot fot it.   
  
summary(hotel$HotelCapacity)

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

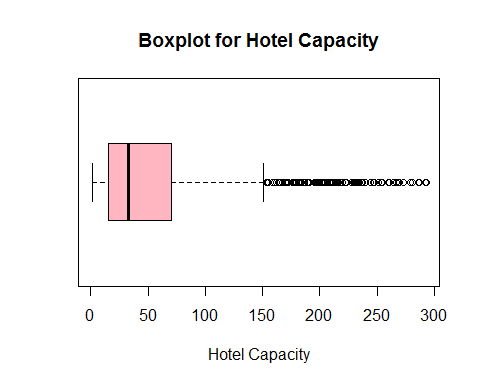
#The median capacity of hotels is 62 while max may even go upto 600.  
  
#Histogram for Hotel Capacity  
  
hist(hotel$HotelCapacity, main="Histogram for Hotel Capacity",xlab="Hotel Capacity",col="lightpink")



#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")

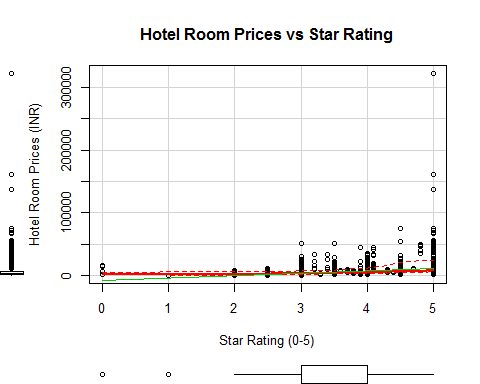


#BoxPlot for Hotel Capacity  
 boxplot(hotel$HotelCapacity, main="Boxplot for Hotel Capacity",horizontal = TRUE,xlab="Hotel Capacity",col="lightpink")



#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)")

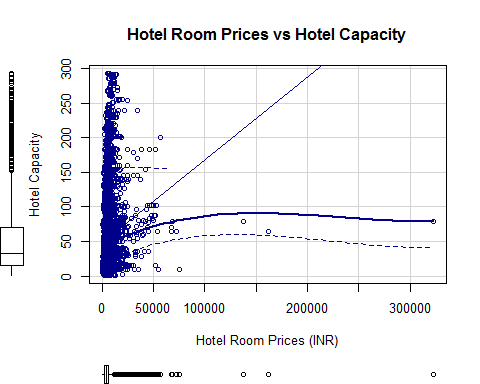


#The scatter plot shows a positive trend in the room prices with increase in star ratings.  
  
#Let's confirm with a correlation test.  
  
cor.test(hotel$StarRating,hotel$RoomRent)

##   
## 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.

#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")

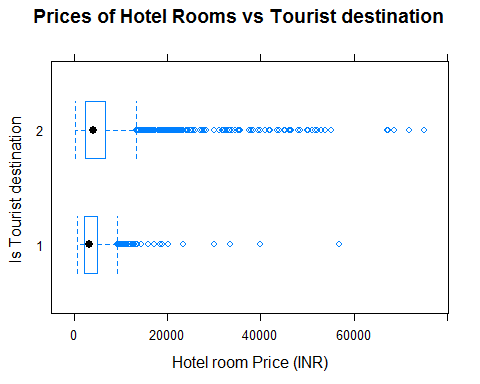


#There is a lot of variabity in the hotel room prices and hotel capacity.  
  
#Let's confirm with a correlation test.  
  
cor.test(hotel$HotelCapacity,hotel$RoomRent)

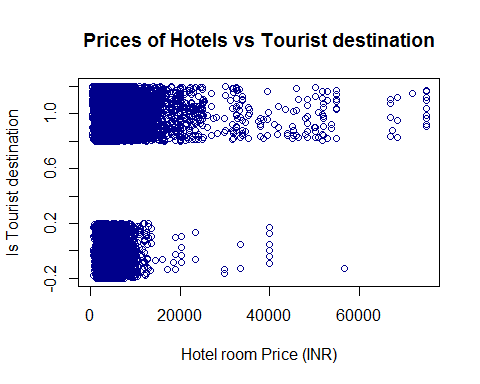
##   
## Pearson's product-moment correlation  
##   
## data: hotel$HotelCapacity and hotel$RoomRent  
## t = 18.205, df = 12944, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1411585 0.1747510  
## sample estimates:  
##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

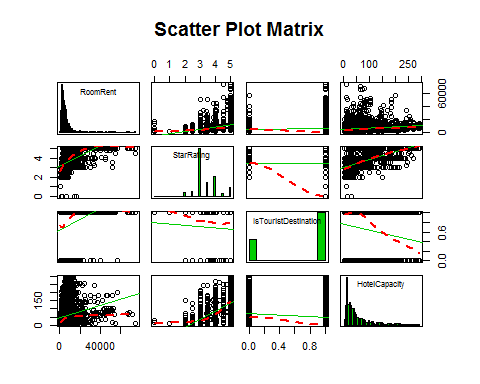


#Both the plots clearly confirms that being a tourist destination shows a positive trend in increase of hotel room prices.  
  
#Let's confirm with a correlation test.  
  
cor.test(hotel$IsTouristDestination,hotel$RoomRent)

##   
## 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

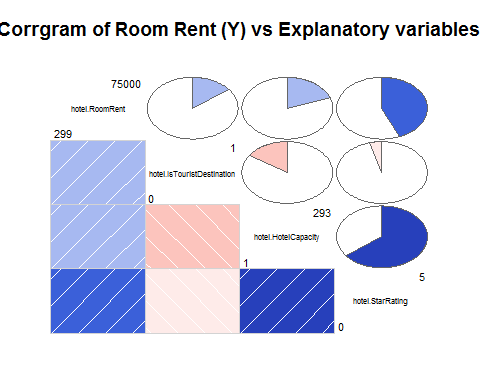
#The p value being <2.2e-16 suggests that the result is statistically significant and that 'being in a tourist destination and room rents are independent of each other' can be rejected. Now, the strength of the correlation is weak with 0.14 and a positive correlation is seen between the two.

#Scatterplot matrix between Y and x1,x2,x3  
   
 scatterplotMatrix(  
 hotel[  
 ,c("RoomRent","StarRating","IsTouristDestination","HotelCapacity" )],   
 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 is in a tourist destination, we see higher room rates. But, for hotel capacity, no such inferences can be drawn.

#Corrgram of Y, x1, x2, x3  
subset <- data.frame(hotel$RoomRent, hotel$IsTouristDestination, hotel$HotelCapacity, hotel$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", "IsTouristDestination", "HotelCapacity", "StarRating")  
corMatrix <- rcorr(as.matrix(hotel[,colroom]))  
corMatrix

## RoomRent IsTouristDestination HotelCapacity  
## RoomRent 1.00 0.14 0.20  
## IsTouristDestination 0.14 1.00 -0.16  
## HotelCapacity 0.20 -0.16 1.00  
## StarRating 0.44 -0.05 0.64  
## StarRating  
## RoomRent 0.44  
## IsTouristDestination -0.05  
## HotelCapacity 0.64  
## StarRating 1.00  
##   
## n= 12941   
##   
##   
## P  
## RoomRent IsTouristDestination HotelCapacity  
## RoomRent 0 0   
## IsTouristDestination 0 0   
## HotelCapacity 0 0   
## StarRating 0 0 0   
## StarRating  
## RoomRent 0   
## IsTouristDestination 0   
## HotelCapacity 0   
## StarRating

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

## [,1]  
## IsTouristDestination 0.1445172  
## StarRating 0.4365300  
## HotelCapacity 0.1954735

cov(x,y)

## [,1]  
## IsTouristDestination 386.5673  
## StarRating 1859.1050  
## HotelCapacity 63712.7318

#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.