Analysis of Hotel Room Pricing In the Indian Market

DATA ANALYSIS REPORT



Aviral Bhatnagar

July 11, 2017

**Contents**

[Aim 3](#_Toc10573)

[Objectives 3](#_Toc10574)

[Abstract 3](#_Toc10575)

[Review of the Data 4](#_Toc10576)

[Dataset Used 4](#_Toc10577)

[Tools Used 5](#_Toc10578)

[Important Observations Made Based on The Initial Review of DataSet 5](#_Toc10579)

[Selection of the Important Variables 8](#_Toc10580)

[1. Boruta Analysis 8](#_Toc10581)

[2. Correlation Identification based on Corrograms 8](#_Toc10582)

[Star Rating: 8](#_Toc10583)

[Hotel Capacity: 9](#_Toc10584)

[Swimming Pool: 9](#_Toc10585)

[Checking Correlation of Selected Factors 9](#_Toc10586)

[Relationship between Room Rent and Star Rating 10](#_Toc10587)

[Relationship between Room Rent and Availability of Swimming Pool 10](#_Toc10588)

[Relationship between Room Rent and Hotel Capacity 10](#_Toc10589)

[Relationship between Star Ratings and Hotel Capacity 10](#_Toc10590)

[Also checking Relationship between Star ratings and Swimming pool 10](#_Toc10591)

[Also checking Relationship between Star ratings and Hotel Capacity 11](#_Toc10592)

[Also checking Relationship between Hotel Capacity and Availability of Swimming Pool 11](#_Toc10593)

[Regression Analysis in the Hotels pricing strategy study [5] [6] 11](#_Toc10594)

[Model 1: 11](#_Toc10595)

[Model 2: 12](#_Toc10596)

[Model 3: 14](#_Toc10597)

[Conclusion 15](#_Toc10598)

[References 16](#_Toc10599)

**Aim**

To identify the factors that contribute to the fluctuations in the prices of room per night of Hotels in India based on Data Analysis of data from 42 cities.

# Objectives

The major objectives of this report are as follows:

1. Identification of factors that drive up the hotel room rents. These factors may be identified based on both machine and human interpretation via:
   1. Correlations shown from Tests and Data Models
   2. Manager Intuition in identifying correlations amongst the variables
2. Run analysis on the selected variables to verify their correlations
3. Form Regression Models on the selected variables

# Abstract

Hotel Room prices in India are based on many factors like the location, the hotel star rating (based on Ministry of Tourism classification), facilities available and hotel popularity (on a scale of 1-5) amongst other factors.

In this project we take data from hotels in 42 different Indian cities collected between December 2016 and January 2016. Based on the data, we run analysis to determine the 3 main factors that contribute to the variance of the room rent across all the hotels. Also, we determine the factors which have a direct correlation with the hotel room price and attribute comprehendible reasons to the correlations between these factors and the Room Rents.

# Review of the Data

## DATASET USED

The data set we are working on here is a sequential data set with records for consecutive days for 42 different cities all over the country. (Mumbai, Delhi, Bangalore, Chennai, Hyderabad,

Ahmedabad, Kolkata, Surat, Pune, Jaipur, Thrissur, Lucknow, Kanpur, Amritsar, Indore,

Kanyakumari, Agra, Madurai, Goa, Rajkot, Varanasi, Srinagar, Jodhpur, Chandigarh,

Thiruvathipuram, Guwahati, Mysore, Bhubaneswar, Kochi, Mangalore, Udaipur, Pondicherry, Haridwar, Puri, Shimla, Panchkula, Darjeeling, Rishikesh, Gangtok, Ooty, Jaisalmer, Bodh Gaya, Nainital, Munnar, Manali) India.

The dataset is compiled from www.hotels.in and aggregates the hotel prices on 8 different dates at different hotels across these different cities. To set a common room type for correct comparative analysis, we study the Room Rents for the cheapest room of each hotel with double occupancy for a night. The fields in this data set are as follows:

* Room Rent
* Date
* Whether the day of booking is a weekend or not
* Whether the day of booking is a new year eve
* City name
* Population of that city
* The city’s rank
* Whether the city is a metro city or not
* Whether the city is a tourist destination or not
* Hotel name
* Star rating
* Distance from airport to that hotel
* Hotel address
* Hotel pin code
* Hotel description
* Hotel capacity
* Facilities offered by hotel:
  + Free wifi
  + Free breakfast
  + Whether it has a swimming pool or not

**Note:** On observation of the data in the dataset, we see that there is an inconsistency in the format of the dates. While most dates were in the UK format, a few were input in the US format. Since the US dates used the MMM DD YYYY format, it was easy to identify this inconsistency and convert them to the UK format using the gsub command.

## TOOLS USED

The entire analysis of this dataset was done on RStudio Version <xyz> which has the underlying R Language version <pqr>. RStudio provides a very useful tool for working with R and is the most widely used enterprise ready tool used in the industry. The dataset itself was obtained as csv files segregated by cities. They were then consolidated in R into a single data frame to run analysis on it collectively.

## IMPORTANT OBSERVATIONS MADE BASED ON THE INITIAL REVIEW OF DATASET

We run a few simple commands to visualise the data and see its main statistics. The initial observations based on these summaries and boxplots (refer appendix with code) are listed below: [1] [2]

**General Facts:**

* Max number of Hotels - Delhi
* Highest Population (12442373) – Mumbai
* Lowest Population (8096) – Manali  Room Rent Statistics:

o Average Room Rent (All) – ₹ 5473.99 o Average Room Rent (Metros) – ₹ 4696.07 o Average Room Rent (Non Metros) – ₹ 5782.79 o Average Room Rent (Weekdays) – ₹ 5430.835 o Average Room Rent (Weekends) – ₹ 5500.129 o Average Room Rent (New Years’ Eve) – ₹ 6222.82 o Most Recurrent Room Rent Value – ₹ 4000 o Max Room Rent value (Rambagh, Jaipur) – ₹ 322500

* Average Star Rating - 3.46
* Most Recurrent Star Rating – 3
* Maximum no of 5 Star Rating Hotels – Ahmedabad
* Average Distance of Hotels from Airport - 21.16km
* Average Hotel Capacity - 62.59  Highest Hotel Capacity - 600

**Other Observations:**

* 92.5% of the hotels have Free Wi-Fi, probably owing to the increasing demand of internet by consumers these days.
* Free breakfast is offered by 64.91%.
* Only 33% Hotels have a swimming pool, probably since it drives the room rent up.
* Most no of hotels are in the tourist destinations which seems to be obvious due to more incoming tourists in those cities.
* From the boxplot of Room Rent vs Population of Cities, we observe that as the population increases, initially the average room rent decreases in the lower half of the plot and the becomes constant as it hits a low cap.
* From the boxplot of Room Rent vs the destination city being a tourist attraction or not, as expected, we observe that the average room rent for the destination city being a tourist attraction is greater than the average room rent for the destination not being a tourist attraction.
* The Hotel Star rating has a major impact on the room rents. From the boxplots representing room rents on the basis of star ratings of the hotel a pattern of regular increase in room rents with increase in star ratings can be seen.
* On more inspection of the data, we see that the average room rent of hotels offering free Wi-Fi is relatively lower compared to that don’t offer Wi-Fi. But the number of hotels offering free Wi-Fi (12251) is far more than those not offering it (981).

**Observations for correlations amongst variables based on Corrograms:**

When we plot the corrogram for all the variables given in the data set we found some of these major correlations between the variables: [3] [2]

* Metro city was highly correlated with population of the city
* Hotel capacity was highly correlated with star rating  Star rating was highly correlated with swimming pool  Hotel no was correlated to city no, city rank and airport.
* City no was correlated to city rank and airport.
* City rank was correlated to airport.

Some inverse correlations (which means one quantity increases with decrease of other quantity) were also seen in the plot which are as follows:

* Population was inversely correlated to city rank.
* City rank was inversely correlated to metro city.

# Selection of the Important Variables

The various key factors or variables affecting the room rent were selected by the following steps:

## 1. BORUTA ANALYSIS

Boruta is a package in R which can select variables and features and is capable of working with any classification method that outputs variable importance measure (VIM). Boruta performs a top-down search for relevant features and variables by comparing their attributes’ importance with importance achievable at random w.r.t. a fixed variable. It eliminates variables one by one and lists out the most important variables that correlate to a certain variable. [4]

As per the Boruta Test run, 15 attributes were confirmed important – Airport, CityName, CityRank, FreeBreakfast, FreeWifi and 10 more while 3 attributes were confirmed unimportant and were eliminated: Date, IsNewYearEve, IsWeekend

## 2. CORRELATION IDENTIFICATION BASED ON CORROGRAMS

As per the corrograms explained in the previous section, the highest correlation of the Room Rent is with the following factors:

### Star Rating:

In India, the Ministry of Tourism has formulated a scheme for classification of operational hotels using a “Star” rating. Hotels are rated as either 5 Star, 4 Star, 3 Star, 2 Star or 1 Star. Accordingly, we classified the hotels in our dataset using their star rating. The reason for doing this is that the star rating of a hotel has a direct, strongly positive correlation with the price of its hotel rooms. Therefore, it is important to control for price variation because of the star rating.

### Hotel Capacity:

The number of rooms in a hotel denotes the available supply in terms of rooms available and it is therefore expected that this will in fact influence the price that a hotel will set for each room. Accordingly, we used HotelCapacity as a control variable to account for the possibility that the room price set by a hotel may depend upon the supply of available rooms.

### Swimming Pool:

The amenities and facilities provided by a hotel can also potentially influence the price of a room. The more the amenities provided, the higher would be the price of the hotel room. Out of these variables, the availability of a Swimming Pool has the most correlation with the Room Rent. The hotels with a pool have an average rent of around ₹ 8549.05 and those without a pool have an average of ₹ 3775.56. These statistics are not surprising since the pools need a lot of investment and maintenance as well accounting for a lot of cost. That cost is compensated by the hotels by increasing their room rents.

According to correlation matrix it has really high correlation value with star rating as well showing that people are really happy with this being in the primary luxury list. Not only this, it also has second highest correlation value (after star rating) with room rents. Hence it will also play an important role in predicting the rents of the rooms.

# Checking Correlation of Selected Factors

Correlogram plots of all the variables in the table helped us conclude that some variables have a strong relation between each other and might be crucial in creating a model for prediction of room rents. These attributes are star ratings, availability of swimming pool and Hotel Capacity.

Chi square test was conducted to test the above hypothesis, with the following results:

## Relationship between Room Rent and Star Rating

Pearson's Chi-squared test

X-squared = 132390, df = 43100, p-value < 2.2e-16

Since p value is less than 0.05 we reject our null hypothesis and this proves a significant relationship between room rent and star rating.

## Relationship between Room Rent and Availability of Swimming Pool

Pearson's Chi-squared test

X-squared = 8394.6, df = 2155, p-value < 2.2e-16

Since p value is less than 0.05 we reject our null hypothesis and this proves a significant relationship between room rent and swimming pool.

## Relationship between Room Rent and Hotel Capacity

Pearson's Chi-squared test

X-squared = 1479800, df = 545220, p-value < 2.2e-16

Since p value is less than 0.05 we reject our null hypothesis and this proves a significant relationship between room rent and Hotel Capacity.

## Relationship between Star Ratings and Hotel Capacity

Pearson's Chi-squared test

X-squared = 77180, df = 5060, p-value < 2.2e-16

Since p value is less than 0.05 we reject our null hypothesis and this proves a significant relationship between star ratings and hotel capacity.

## Also checking Relationship between Star ratings and Swimming pool

Pearson's Chi-squared test

X-squared = 6303.7, df = 253, p-value < 2.2e-16

Since p value is less than 0.05 we reject our null hypothesis and this proves a significant relationship between star ratings and swimming pool

## Also checking Relationship between Star ratings and Hotel Capacity

Pearson's Chi-squared test

X-squared = 77180, df = 5060, p-value < 2.2e-16

## Also checking Relationship between Hotel Capacity and Availability of Swimming Pool

Pearson's Chi-squared test

X-squared = 6303.7, df = 253, p-value < 2.2e-16

## **REGRESSION ANALYSIS IN THE HOTELS PRICING STRATEGY STUDY**

### Model 1:

Call:

lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity - 1, data = hdata.df)

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

P-values of individual independent variables are also < 0.05, So we reject all the null hypotheses and hence, Star Rating, availability of swimming pool and Hotel Capacity all significantly affects Room Rent.

R2 is very small so it may not be good model and some other factors which are not included might affect Room rent.

The coefficients of model 1 implies:

* When Star rating increases by 1 unit Hotel Room rent increases by

Rs 1396.87

* When availability of swimming pool changes then Hotel room rent increases by Rs 3719.69
* When Hotel capacity changes by 1 unit then Hotel room rent decreases by Rs 7.659814

### Model 2:

Call:

lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity + IsWeekend + IsTouristDestination - 1, data = hdata.df)

Residuals:

Min 1Q Median 3Q Max

-8326 -2517 -1212 463 312480

Coefficients:

Estimate Std. Error t value Pr(>|t|)

StarRating 1258.9558 44.4985 28.292 < 2e-16 \*\*\*

HasSwimmingPool 3670.2511 148.8411 24.659 < 2e-16 \*\*\*

HotelCapacity -6.1769 0.9658 -6.396 1.65e-10 \*\*\*

IsWeekend -509.6479 119.1618 -4.277 1.91e-05 \*\*\*

IsTouristDestination 1053.0394 124.7325 8.442 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6792 on 13227 degrees of freedom

Multiple R-squared: 0.4493, Adjusted R-squared: 0.4491

F-statistic: 2159 on 5 and 13227 DF, p-value: < 2.2e-16

The coefficients of the model concludes:

* When Star rating increases by 1 unit Hotel Room rent increases by Rs 3635.81
* When availability of swimming pool changes then Hotel room rent increases by Rs 2285.13
* When Hotel capacity changes by 1 unit then Hotel room rent decreases by Rs 13.965
* When there is no weekend days the charges of room rent increases by Rs 8396.67 and on weekends it increases the room rent by

Rs 8325.09 .

* Is hotel is near a tourist destination then room rent increases by

Rs 1878.94 .

### Model 3:

Call:

lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity + Airport - 1, data = hdata.df)

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

The coefficients of model 3 implies:

* When Star rating increases by 1 unit Hotel Room rent increases by Rs 1248.42
* When availability of swimming pool changes then Hotel room rent increases by Rs 3903.73  When Hotel capacity changes by 1 unit then Hotel room rent decreases by Rs 6.7433.
* When distance of airport from any hotel increases by 1 unit then Hotel room rent increases by Rs 18.8697

# Conclusion

Most influential variables according to the analysis run on the dataset of these 42 cities were found to be:

1. Star rating

1. Swimming pool
2. Hotel Capacity

The correlation of these variables with respect to the room rent have been verified using

lm(formula = RoomRent ~ StarRating + HasSwimmingPool + HotelCapacity - 1, data = hdata.df)

This research helps us identify the major factors which can affect the consumer prices for a service industry. It gives us a clear picture of which factors influence the pricing of rooms in the Hotel Industry in India. As expected, the Star Rating and Premium Facilities offered by a Hotel affect its average room prices. Apart from this, as with any other market the ‘supply’ (available rooms in terms of Hotel Capacity) also determines the premium charged for a room.

# References

1. S. Mathur, “Data Science and Analytics using R programming: Visualising Data,” Udemy, [Online]. Available: https://www.udemy.com/data-science-and-analytics-usingr/learn/v4/t/lecture/6786224?start=0. [Accessed June 2017].
2. S. Mathur, “Data Science and Analytics using R programming: Frequency and Contingency Tables,” Udemy, [Online]. Available: https://www.udemy.com/data-science-and-analyticsusing-r/learn/v4/t/lecture/6786288?start=0. [Accessed June 2017].
3. W. R. R. Miron B. Kursa, “Feature Selection with the Boruta Package.,” *Journal of Statistical Software,* vol. 36, no. 11, pp. 1-13, 2010.
4. S. Mathur, “Data Science and Analytics using R programming: Linear Regression,” Udemy, [Online]. Available: https://www.udemy.com/data-science-and-analytics-usingr/learn/v4/t/lecture/7482226?start=0. [Accessed June 2017].
5. S. Ray, “7 Types of Regression Techniques you should know!,” Analytics Vidhya, 14 August 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/08/comprehensiveguide-regression/. [Accessed June 2017].