**Final Project Report:**

**Hotel Room Pricing In the Indian Market**

1. **Introduction**

A hotel is an establishment that provides paid lodging on a short-term basis. Facilities provided may range from a modest-quality mattress in a small room to large suites with bigger, higher-quality beds, a dresser, a fridge and other kitchen facilities, upholstered chairs, a flat screen television and en-suite bathrooms. Small, lower-priced hotels may offer only the most basic guest services and facilities. Larger, higher-priced hotels may provide additional guest facilities such as a swimming pool, business centre (with computers, printers and other office equipment), childcare, conference and event facilities, tennis or basketball courts, gymnasium, restaurants, day spa and social function services. Hotel rooms are usually numbered (or named in some smaller hotels) to allow guests to identify their room. Some boutique, high-end hotels have custom decorated rooms. Some hotels offer meals as part of a room and board arrangement.

**The purpose of this project is to analyze the pricing strategy of hotels in the Indian hotel industry.**

Many factors drive hotel room prices. The objective of this project is to identify the factors that matter the most. So, we define Room Rent as follows:

Room Rent = FUNCTION (Date(s); Hotel Features; External Factors)

**2. Overview of the Study**

Hotels vary from modest hotels which provide cheaper rooms with basic facilities to high-end hotels which provide luxurious rooms with value several addition facilities. Hotels decide facilities and room price based on expectation of target customer-base and their willingness to pay for it.

So, Internal factors like Swimming Pool, Spa, decorations, free meals improves experiences of target customer-base and their willingness to pay higher price. While, customers looking for room in economy range would not prefer such rooms, but modest rooms with basic facilities, like bed, availability for water, mattresses and blankets, and vicinity to restaurants.

Similarly external factors like vicinity of hotels to airport, heritage sites or tourist locations, hotel’s rating, date etc. also influence experience of customer base and their willingness to pay. A family on vacation may not consider hotel vicinity, but a business person may value this and pay more. For example, in city Kota of Rajasthan, rooms just near IIT coaching “Bansal” would cost almost double than similar room located 1-1.5 km away.

So, we compare hotel room prices on these internal and external factors, and analyze what factors matter most in hotel-room pricing strategy.

**3. An Empirical study of Indian hotel pricing strategy for rooms**

(Empirical data is produced by experiment and observation.)

**3.1 Overview**

In this Empirical study, we compare prices of hotel rooms from 42 cities across several states, based upon internal and external factors, and identify the factors that matter most in hotel-room pricing strategy. We also consider date on which data was recorded.

Accordingly we construct two hypotheses:

H1: Hotels in metro-cities charge higher rent for room than those not in metro cities.

H2: Hotels with swimming pool charge higher rent for room than those hotels without one

**3.2 Data**

**3.2.1 Data description**

For this study, we collected data from [www.hotels.in](http://www.hotels.in/) in October 2016. The dataset tracks hotel prices on 8 different dates at different hotels across 42 cities, located in several states.  
  
**Dependent Variable**

|  |  |  |
| --- | --- | --- |
| **DECISION VARIABLE** | **UNITS** | **MEANING** |
| RoomRent | Rupees | Rent for the cheapest room, double occupancy, in  Indian Rupees.  Some hotels have more than one type of double occupancy room. For simplicity, we picked the cheapest room with double occupancy. |

RoomRent = Function (Date(s), Hotel Features, aka Internal factors, External factors)

**External Factors**

Many external factors can potentially influence the RoomRent. The dataset captures some of these external factors, as explained below.

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **UNITS** | **MEANING** |
| Date | Text | We have hotel room rent data for the following 8 dates  for each hotel:  {Dec 31, Dec 25, Dec 24, Dec 18, Dec 21, Dec 28, Jan  4, Jan 8}  *If a hotel is sold out on a given date, assume that the price of the hotel room on the date it is sold out is the maximum price from the sample of dates for which prices are available.* |
| IsWeekend | Dummy | We use ‘0’ to indicate week days, ‘1’ to indicate  weekend dates (Sat / Sun) |
| IsNewYearEve | Dummy | ‘1’ for Dec 31, ‘0’ otherwise |
| CityName | Text | Name of the City where the Hotel is located e.g. Mumbai |
| Population | Dummy | Population of the City in 2011 (See Table A1 below |
| CityName | Dummy | Rank order of City by Population (e.g. Mumbai = 0, Delhi =  1, so on); (See Table A1) |
| IsMetrocity | Dummy | ‘1’ if CityName is {Mumbai, Delhi, Kolkata, Chennai}, ‘0’  otherwise |
| ISTouristDestination | Dummy | We use ‘1’ if the city is primarily a tourist destination, ‘0’ otherwise. For example, Goa and Agra are primarily tourist destinations. We assume that most people who visit Goa and Agra and stay in their hotels are in these cities primarily for tourism. |

**Hotel features aka Internal Factors**

Many Hotel Features can influence the RoomRent. The dataset captures some of these internal factors, as explained below.

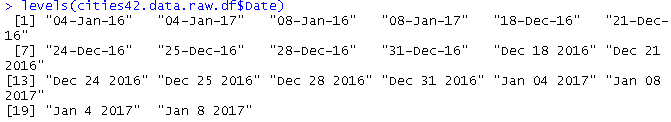
|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **UNITS** | **MEANING** |
| HotelName | Text | e.g. Park Hyatt Goa Resort and Spa |
| StarRating | Number | e.g. 5 |
| Airport | Km | Distance between Hotel and closest major Airport |
| HotelAddress | Text | e.g. Arrossim Beach, Cansaulim, Goa |
| HotelPincode | Number | 403712 |
| HotelDescription | Text | e.g. 5-star beachfront resort with spa, near Arossim  Beach |
| FreeWifi | Dummy | ‘1’ if the hotel offers Free Wifi, ‘0’ otherwise |
| FreeBreakfast | Dummy | ‘1’ if the hotel offers Free Breakfast, ‘0’ otherwise |
| HotelCapacity | Dummy | e.g. 242. (enter ‘0’ if not available) |
| HasSwimmingPool | Dummy | ‘1’ if they have a swimming pool, ‘0’ otherwise |

**3.2.2 Data Cleaning**

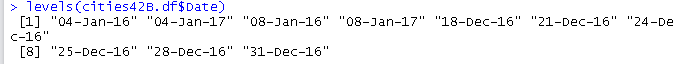
There were 8 entries with NA values, I omitted them.

There were duplicate levels of date. I merged duplicate entries

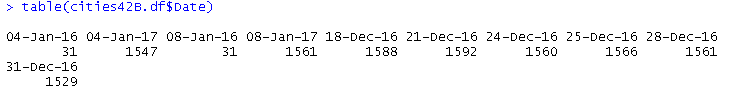
Before,



Later,



Note: It looks like there is data entry error for date 04-Jan-17 and 08-Jan-17. All dates seem to be around Dec 31 2016, but 04-Jan-16 and 08-Jan-16 seems odd. I think dates 04-Jan-16 and 08-Jan-16 were supposed to be 04-Jan-17 and 08-Jan-17. One can see in table that count for 04-Jan-16 is 31 only.



Or maybe recent year data was not available so previous year data was taken. But, I am not sure about that. I am not merging them and leaving it like this.

Also, there are some 0 star hotels which are charging rent Rs 5000-6000 for room. While there are some 5 star hotels which are charging rent Rs 3000. I am not sure about how to interpret and correct this, so I have left it like that.

**3.3 Models**

We analyzed following models to test the hypothesis:

**First Model**

As mentioned in TED talk and HBR articles **(**[Simplify Your Analytics Strategy](https://docs.google.com/document/d/1EjByKggHJq-8kIwld8mCPtTedFfM_Sz1JoZcVuZTMEw/edit?usp=sharing) by Narendra Mulani**,** Making Advanced Analytics Work for You *by Dominic Barton and David Court*, and What People Analytics Can’t Capture by Daniel Goleman)that we should not let data drive us to conclusion, but we should have our questions and then have data support it. So, I made my first model by using 3 most important factors which I found through quick analysis.

This model has multiple R-squared value of 0.23.

Result: Apparently, either my qualitative analysis was incorrect or data-set has limitations. This guided me towards second model, in which I checked the scope of good fit measure, through Linear Regression by OLS method.

**Second Model**

In this Model, I included all variables except HotelDescription, HotelAddress and HotelPincode to see maximum fit measure possible through Linear Regression by Ordinary Least Square method. I am not considering dependence among variables to keep it simple. My aim is to see the scope of prediction and model fitting.

Model has high good fit measure, because multiple R-squared value is 0.76.

Then I removed RoomRent outliers which were beyond 0.95 quantile, since they were causing slope to be more than actual. Also, those outliers were not important like outliers in bank data, which may show fraud. As a result, multiple R-squared value increased to 0.88.

*But at that time I was unaware that HotelName can’t be interpreted in final model. But, this gave me idea*. I tried to look into pattern of regression coefficients of HotelName. I was hoping to create new variable in data-set accordingly, and then use that variable instead of HotelName. I thought about how to give levels to this variable so it can keep good fit measure high. Based on complex nature of task and my technical limitation I left it for another day. This brings me to third model.

**Third Model**

After removing HotelName from second model, I got this model. That is, this model includes all variables except HotelName, HotelDescripton, HotelAddress and HotelPincode.

It has multiple R-squared value of 0.49, after removing RoomRent outlier beyond 0.95 quantile.

**Fourth and Final Model**

After removing HotelName from second model, and eliminating not important regression coefficients, I got this final model. The regression coefficient of variable HotelCapacity is 4, which is very small so I removed it as significant predictor.

After entering coefficients:

ci is regression coefficient for particular City. Term CityNamei is 1 for corresponding city, for other cities it will be zero.

di is regression coefficient for particular date. Term Datei is 1 for corresponding date, for other dates it will be zero.

Multiple R-squared value for this model is 0.49. So, this model is an ok fit. This model can’t predict rent of room precisely, but it clearly explain relation between predictors and response variable, whether predictors (StarRating, HasSwimmingPool) have positive or negative effect.

**3.4 Results**

H1: Hotels in metro-cities charge higher rent for room than those not in metro cities.

H2: Hotels with swimming pool charge higher rent for room than those hotels without one

**Third Model**

I didn’t found empirical support for H1. For beta coefficient of variable “IsMetroCity” P-value was 0.12 (p > 0.05, this means its coefficient is not of statistical significance. Although, I expected that Metro City influence hotel room pricing, but based on results from analysis I reject H1.

I found empirical support for H2. The beta coefficient for predictor “HasSwimmingPool” is +1493 and significant. Therefore, I can conclude that having swimming pool causes increase in hotel room price.

**Fourth and Final Model**

Since Final Model is filter of Third Model, same conclusion applies. We reject H1.

Since Final Model is filter of Third Model, same conclusion applies. Except, beta coefficient is +1037.

**4. Conclusion**

This paper was motivated by the need for research to investigate the most important factors for Hotel room pricing strategy. I observed that one group of customers are willing to pay more for greater experience. The other group is fine with basic amenities. Hotels provide various value-addition services like swimming pool, free-wifi etc. The factors which influence hotel room pricing strategy most is City, date of stay, Star Rating, and availability of swimming pool.

I found empirical support that hotel room price is positively and significantly influenced by its rating. I didn’t found empirical support to conclude that location of hotel with respect to metro-city influence price of room.

**5. References**

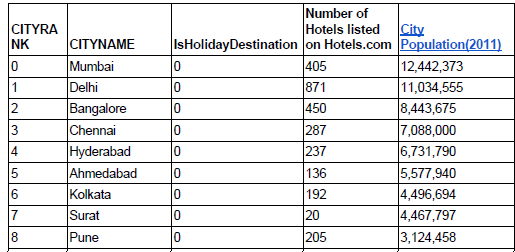
Hotel Industry description: <https://en.wikipedia.org/wiki/Hotel>   
Data-set: [www.hotels.in](http://www.hotels.in/)

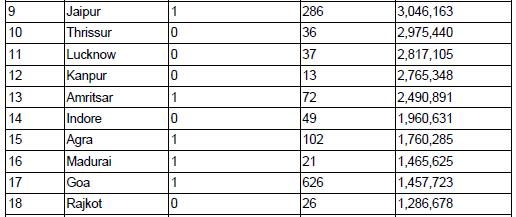
Data Description: <https://docs.google.com/document/d/1_Bq_NhLK8wEHYx43cKXOBqdvxzRuuc1DqdxaXRsaR84/edit>

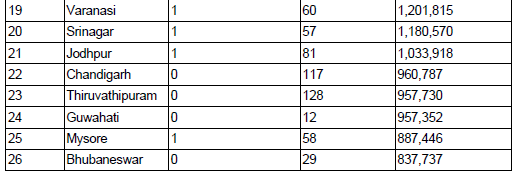
**Table-1 Regression analysis in Hotel room pricing strategy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | **Third Model** | | **Fourth and Final Model** | |
| **Beta coeff** | **Std. error** | **Beta coeff** | **Std. error** |
| Intercept | -1642 | 476.4 | -1940 | 368.5 |
| CityName | 1 coeff per city |  | 1 coeff per city | 1 value per coeff |
| Population | NA | NA |  |  |
| CityRank | -14.5 | 14.8 |  |
| IsMetrocity | 422.18 | 281.9 |  |
| IsTouristDestination | 311.6 | 173.2 |  |
| IsWeekend | 217.4 | 134.1 |  |
| IsNewYearEve | 50.14 | 412.8 |  |
| Date | 1 coeff per date |  | 1 coeff per date | 1 value per coeff |
| StarRating | 1493 | 34.6 | 1679 | 30.7 |
| Airport | 12.27 | 2.3 |  |  |
| FreeWifi | -90 | 66.0 |  |
| FreeBreakfast | 339 | 37.7 |  |
| HotelCapacity | 3.96 | 0.3 |  |
| HasSwimmingPool | 901.06 | 49.5 | 1037 | 48.7 |

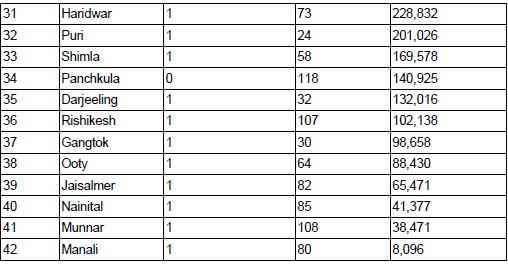
**Table-2: City rank (based on 2011 City Population)**











**Data from following cities and states are taken:**

State City

Maharashtra- Mumbai, pune

Delhi

Goa

Kerala - Thrissur , Kochi, Munnar

Uttar Pradesh- Lucknow, Kanpur, Agra, Varanasi

Tamilnadu- Madhurai, thiruvananthpuram, Chennai, ooty

Madhya Pradesh- Indore

Punjab- Amritsar, chandigarh

Karnataka- banglore, mysore, manglore,

Gujarat- Rajkot, ahemdabad, surat

Rajasthan- jodhpur, Udaipur, Jaisalmer,surat

Jammu and Kashmir- shrinagar

Assam- Guwahati

Odissa- bhubaneshwar, puri

Uttarakhand- haridwar, rishikesh, nainital

Himachal Pradesh- shimla, manali

Haryana- Panchkula

West Bengal- Darjeeling, kolkata

Sikkim- gangtok

Telangana- hyderabad

**Very expensive Hotels:**

**Hotel Umaid Bhavan, Jodhpur, Rajasthan**



**Hotel Taj Rambagh Palace, Jaipur**





**Inexpensive Hotels:**

**Hotel chaman, shimla**

