**Final Project – STA 380 (MATH STAT FOR APPLICS)**

**Project Title : IMPACT OF VARIOUS PARAMETERS ON HOTEL PRICING**

Submitted By : Tushar Bansal (EID : tb34826)

Submitted To : Prof. Jared Murray

**Introduction**

Tourism industry is one of the biggest industries in India and this paper focuses on the pricing of the hotel industry. Pricing is a principal concern for workers in the field of hospitality. It's really the one component of the marketing mix of lodging that explicitly influences returns. Also, pricing strategies are dynamic and complicated and are affected by internal aspects (e.g. near-term marketing targets, long-term marketing plan, along with the expenses and factors required in posting prices and price change) and external aspects (e.g. market requirement and availability conditions, competitive environment and other environmental conditions). This is where the need for *hedonic pricing* : by considering the prices of different attributes come in place.

This paper is based on determining the characteristics relevant for the hotel pricing in India. Here the dependent variable Y: hotel pricing is studied against a set of independent qualitative and quantitative characteristics like star rating, distance from airport, metro/tourist city, availability of other amenities etc. The major goal of this paper is to determine which factors significantly contribute to the pricing of the hotel. Regression analysis has been used to figure out the more relevant characteristics. Also, a city-wise regression analysis is done to understand the effects on other variables when the city variables are removed from consideration. In the later part of the study, shortcomings of this model are also discussed. In general, in this paper, statistical significance argument has been made based on estimates and confidence intervals.

**Methods**

Most Indian hotels can be seen on the lists of more than one internet travel agencies. The data was collected from www.hotels.in by web-scraping as on 12/18/2016 and 12/31/2016 from hotels from 45 Indian cities. The dates were selected to include the effect of holiday (i.e. New Year), and since the New Year was on the weekend, an additional weekend date was considered to adjust for the weekend effect. The dates considered are also close by to eliminate any seasonality effect. Also, the data was collected from a single website to eliminate disparities related to price differences/ bookings. Additionally, cities from all over India were considered to eliminate any effects of regionalities : Mumbai, Delhi, Chennai, Goa, Thrissur, Kolkata, Ahmedabad etc.

The hotels in the study include luxury, moderate and cheap hotels with rent ranging from INR 299 - 39500. The capacity in the hotels range from 5 rooms to 600 rooms with the distance from the airport ranging from 0.2 kms to 124 kms. This information along with more information like locations, facilities is available online. The rates on these platforms typically represent the charge customers are expected to spend willingly instead of the 'rack cost' and so in that context and for the objective of this study, they may be assumed to indicate the hotelier's view of the implicit price for the hotel’s goods and services.

The set of characteristics for this study consists of both quantitative and qualitative features. The quantitative characteristics considered for deciding room price is the star rating, distance from airport, hotel capacity. Qualitative characteristics include factors about the location of the hotel, for when the hotel is booked, amenities provided by the hotel. The amenities considered in this case are free Wi-Fi, free breakfast, swimming pool. These characteristics are calculated on a binary scale with dummy variables, 1 depicting presence of the said characteristic and 0 indicating absence.

The characteristics can also be divided as :

City characteristics: Metro City, Tourist Destination

Time characteristics: New Years’ Eve

Hotel characteristics: Star Rating, Distance from Airport, Free Wi-Fi, Free Breakfast, Hotel Capacity, Swimming Pool

First, in order to study the behaviour of the weekend on the room rent, a regression analysis of all the hotel characteristics is done to understand the statistical significance of the variables. Then, to study the impact of New Year’s Eve on the room rent, a similar regression is done with all the hotel characteristics. Then, to understand the city characteristics, four cities : Delhi, Kolkata, Chandigarh, Goa were chosen to consider all the possible cases of cities being metropolitan and/or tourist destinations or none.

|  |  |  |
| --- | --- | --- |
| Variables | IsMetroCity | IsTouristDestination |
| Delhi | 1 | 1 |
| Kolkata | 1 | 0 |
| Chandigarh | 0 | 0 |
| Goa | 0 | 1 |

These four cities were then regressed individually to understand the impact of hotel characteristics with non-variable city characteristics.

*Table 1: Four Cities and their variables for a City-Wise Analysis*

**Results**

**Part 1 : Regression of variables on Room Rent for all datapoints**

It is a large dataset with 3149 datapoints and with the extreme range of room rent (Rs. 299 – Rs. 39500), it can be observed that most of the datapoints lie in the under Rs. 5000 (Appendix A, Chart 1). This is important to note as the mean of the Room Rent also comes to be around 5068 with a median at 3820. After analysis of each variable using box plots and scatter plots, it can be seen that most covariates do not depend on each other. However, there was one particular pair of covariates (namely, StarRating : The rating of each hotel from 1-5 and HasSwimmingPool : whether a hotel has a swimming pool or not) which exhibited dependence. On average, hotels with a swimming pool got one point more in their rating when compared to the hotels with no swimming pool (Appendix A, Chart 2). This makes sense as well, because hotels with added amenities would be preferred (and hence rated higher). However, such an effect was not observed on comparing Star Rating with free Wi-fi or free breakfast).

Apart from that, on studying correlations (Appendix A, Chart 3) among different covariates and the input variable (Room Rent), it can be seen a few variables are strongly correlated (Hotel Capacity, Star Rating and Star Rating, Swimming Pool) with factor greater than 0.6. The second factor was also discovered in the box plot analysis. Hence, for the initial model, Star Rating is not considered. Also, it makes more sense intuitively to include amenities such as Free Breakfast, Free Wi-Fi, Swimming Pool and other factors including Airport Distance, New Year’s Eve, Tourist Destination are considered.

From Appendix A, Figure 1a and 1b(confidence intervals from bootstrap), it can be seen that while most of the X-variables are statistically significant, there is one which is free Wi-fi which is not. It is surprising but multiple factors can go into this consideration : one of the obvious ones being that with the cheap mobile internet available in India, it isn’t a problem if the hotels don’t have Wi-fi and they can still charge the room rents without providing this service.

After considering all the variables together, it can be seen (Appendix A, Figure 2a and 2b), while most of the variable estimates and their ranges are similar to as in the above case, only swimming pool term change considerably. This can be attributed to multicollinearity in the case of swimming pool as discussed above and in the collinearity chart. (Although Free Wi-fi estimate changes as well, it lies in the confidence interval from the case above and its new confidence interval also doesn’t change as much, again being the only statistically insignificant one). Hotel Capacity also comes out to be statistically insignificant variable (from estimate and confidence intervals)

Another interesting inference from these two fits is the change in R-squared value that goes approximately from 0.26 to 0.36, signifying a better fit. Considering Variance Inflation Factor (Appendix A, Figure 2c), factors are between 1 and 2.12, signifying moderate correlation between variables. It can be observed (Appendix A, Chart 4) that while errors are distributed, they are heteroskedastic. In the chart, they are only plotted with Airport variable, but when, taken other plots as well, the errors show similar trend. Hence, usage of bootstraps helps give valid inferences in this case.

An obvious interaction that could be present is between Tourist Destination and New Year’s Eve. That is, typically at a tourist destination, the prices are increased by a major factor when there on New Year’s Eve because people go celebrate to these destination. As can be seen (Appendix A, Figure 3a and 3b), the interaction term is statistically significant (with a small increase in R-square term), which makes the inclusion more conclusive. It doesn’t have influence on other covariates, as such, but the individual terms of Tourist Destination and New Year’s Eve are considerably changed, which makes sense because there is another factor including them.

**Part 2 : Regression of variables on Room Rent for city-wise Analysis**

Regression Analysis for four cities (Delhi, Kolkata, Chandigarh and Goa) has been done to consider the impact of Tourist Destination as well as being a metropolitan city. An important thing to note here is that the number of observations taken from each of these cities varies with them being 510 (Delhi), 119 (Kolkata), 84 (Chandigarh) and 152 (Goa). This might (and possibly will) affect any comparative analysis conducted between the cities so they are considered independently.

**Case 1 : Delhi**

The hotels in Delhi (Appendix B, Figure 1a and 1b), were regressed with all the covariates to understand their impact on hotel room pricing. In this case, there are four insignificant variables : New Year Eve, Airport, Wi-fi, Hotel Capacity. The R-squared value for the regression increases considerably to 0.49.

Case 2 : Kolkata

For Kolkata (Appendix B, Figure 2a and 2b), there are five insignificant variables : New Year Eve, Airport, Wi-fi, Breakfast, Swimming Pool. The R-squared value increases again to 0.62.

Case 3 : Chandigarh

In case of Chandigarh (Appendix B, Figure 3a and 3b), there are six insignificant variables : New Year Eve, Airport, Wi-fi, Breakfast, Swimming Pool, Star Rating (basically all except Hotel Capacity). The R-squared value reaches 0.71.

Case 4 : Goa

As for Goa (Appendix B, Figure 4a and 4b), there are five insignificant variables : Airport, Wi-fi, Breakfast, Swimming Pool, Hotel Capacity. The R-squared value for the regression again, increases to 0.46 (from the complete regression).

**Discussions**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Expected Effect** | **Effect in this study** |
| **IsTouristDestination** | **+** | **+** |
| **IsNewYearEve** | **+** | **+** |
| **StarRating** | **+** | **+** |
| **Airport** | **-** | **+** |
| **FreeWiFi** | **+** | **IS** |
| **FreeBreakfast** | **+** | **+** |
| **HotelCapacity** | **-** | **IS** |
| **HasSwimmingPool** | **+** | **+** |
| **IsMetroCity** | **+** | **-** |

*Table 2: Expected and Obtained Effects of different Parameters on Room Rent*

The cities that are discussed above might not be able to capture the general characteristics of the defined city and thus might not incorporate all causes and effects of different factors.  
While analysing the model. Further there have been three uncertain parameters which might have contributed to the deviations in the above model:-

* Multicollinearity- Dealing with multiple correlated variables
* Heteroskedasticity- Variability in errors
* Misspecification- Relevant/irrelevant variables

While some part of multicollinearity and heteroskedasticity was discussed in the paper, misspecification is also very important that is the model might not account for everything it could. The model may have biased coefficients and error terms. For example : although the distance from airport was considered, the distance from other important landmarks should also be considered because when we are taking about tourists, an airport might be far away and hence, the model does not consider that. There could have also been other parameters, incorporating for the different price ranges that the hotels provide for different kind of rooms.

Hotel Pricing is one of the most important components in the marketing mix of the industry and yet it remains less investigated than other factors. Therefore, Hotel pricing depends on more than just what we pay (for what we consume). From what is observed from these models, hotel pricing depends on various other factors, and these factors keep on changing with time and city. It is a dynamic, but not a flexible concept, meaning there are factors that affect pricing and have to be accounted for. In order to define a structure for pricing, a primary thing is the city that one is looking at. These results obtained might not be universal and vary from city to city.

**Appendix A :**

Chart, histogram

Description automatically generated

*Chart 1 : Frequency of Room Rent (in Rupees)*

Chart, box and whisker chart

Description automatically generated

*Chart 2 : To show that having Swimming Pool considerably changes Star Rating*

Chart

Description automatically generated

*Chart 3 : To show collinearity in different variables (including input variable)*

Table

Description automatically generated

*Figure 1a : Regression of Room Rent on New Year, Tourist Destination, Airport, Free Breakfast, Free Wifi, Swimming Pool*

Text, table

Description automatically generated

*Figure 1b : Confidence Intervals from bootstrapping above fit*

Text

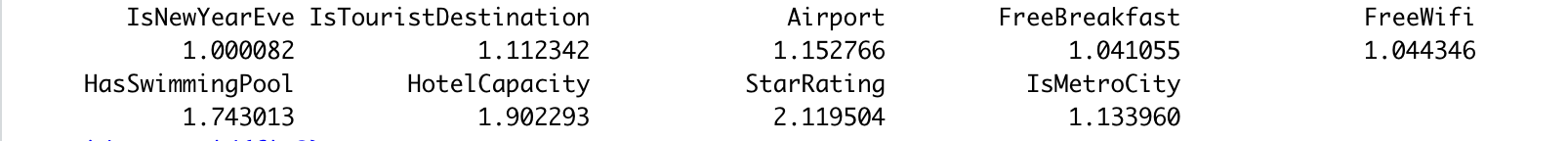
Description automatically generated

*Figure 2a : Regression of Room Rent on all the variables*

Text, table

Description automatically generated

*Figure 2b : Confidence Intervals from bootstrapping above fit*



*Figure 2c : VIF factors for different parameters*

Chart, scatter chart

Description automatically generated

*Chart 3 : To show error distribution for above fit with X-variable being Airport (distance in kms)*

Text

Description automatically generated

*Figure 3a : Regression of Room Rent on all the variables and an additional interaction term*

Text, table

Description automatically generated

*Figure 3b : Confidence Intervals from bootstrapping above fit*

**Appendix B :**

Table

Description automatically generated

*Figure 1a : Regression of Room Rent on all the variables in Delhi*

Text, table

Description automatically generated

*Figure 1b : Confidence Intervals from bootstrapping above fit*

Table

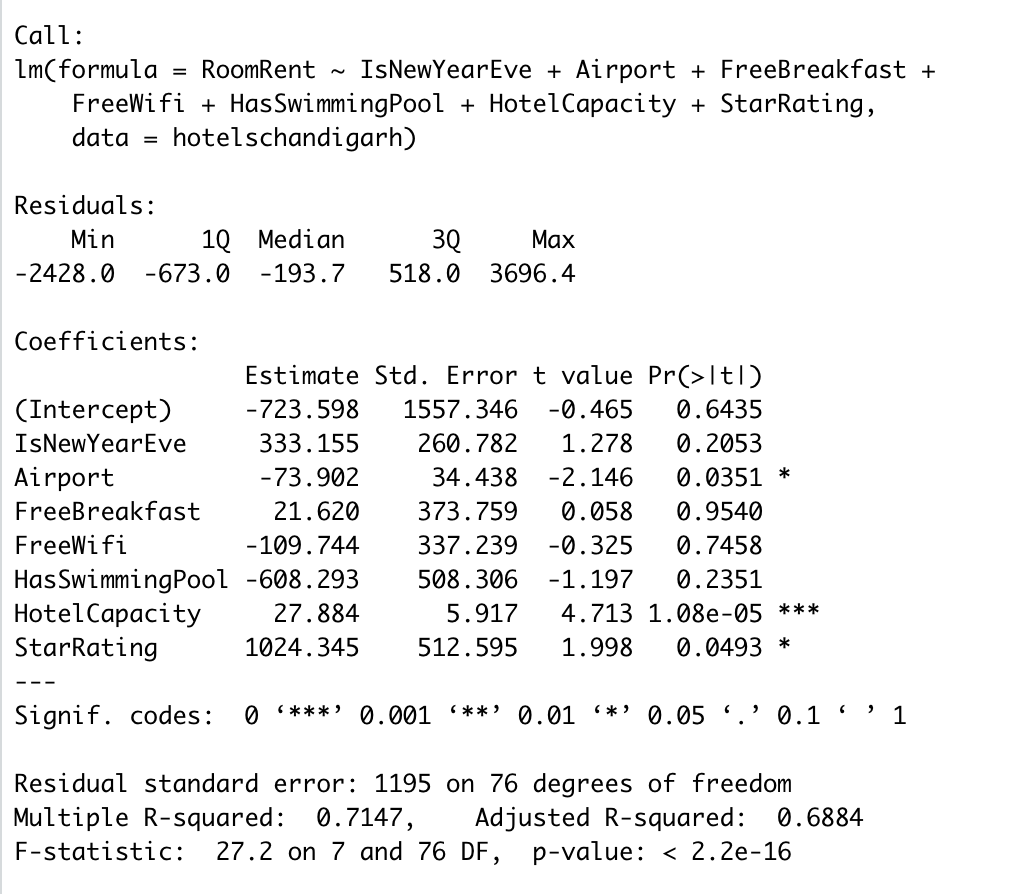
Description automatically generated

*Figure 1a : Regression of Room Rent on all the variables in Kolkata*

Table

Description automatically generated

*Figure 2b : Confidence Intervals from bootstrapping above fit*



*Figure 3a : Regression of Room Rent on all the variables in Chandigarh*

Text, table

Description automatically generated

*Figure 3b : Confidence Intervals from bootstrapping above fit*

Table

Description automatically generated

*Figure 4a : Regression of Room Rent on all the variables in Goa*

Table

Description automatically generated

*Figure 4b : Confidence Intervals from bootstrapping above fit*