**RESEARCH REPORT:**

**HOUSING PRICE ON AIRBNB IN BOSTON, UNITED STATES**

**ECON 271 Business Analytics**

**By Professor Disha Shende**

**Prepared by**

**Chamnan Suon, and Dung Pham**

**Introduction**

The purpose of this research report is to analyze the data on housing price on Airbnb in Boston, United States. The data is compiled and made available online on February 19, 2021. Our research question in this analysis is ***does neighborhood have an impact on housing prices on Airbnb in Boston?*** We will examine how variable *price* is affected by other variables such as *neighbourhood, number of minimum nights, number of available days for listing per year, and room types.* In an economic and business perspective, understanding how rental price is affected by other factors such neighbourhood may allow the company to better understand issues and be able to determine effective strategies and solutions for business improvement and development.

**Review of Literature**

When a host would like to list their house on the Airbnb marketplace, they would first need to decide a reasonable and suitable price for it. This process of setting price may be based on a lot of different factors, and many researches have proved that neighbourhood is one of the main determinants of housing price. In particular, there are certain neighbourhood features that are more appealing to customers, and those features may significantly increase the value of the homes near them (“Are you paying too much for a house in your neighborhood?,” n.d). Some examples of the more favouring neighbourhood features might include greenways, social connections, activity levels, etc. A study of Lee Chun Yang and Hui-Yu Lin has also shown that when determining the location of residence, consumers will prioritise the environment and public facilities and services offered by the house’s neighbourhood (“The impact of neighborhood characteristics,” n.d). Moreover, location is another important factor that may determine whether a neighbourhood is more valuable or less valuable for the clients. Frequently, those neighbourhoods that are located near famous tourist attractions, or have proximity to shopping, entertainment, and recreational centers tend to be much more attractive than those that are not (“8 critical factors that influence a home’s value,” 2019). Thus, houses that are located in prime areas with favoring neighborhood features will result in a more positive impact on rental prices for the house’s owner. That is, the owner would be able to increase their revenue by setting a higher rental price for their house while not losing any of their customers. According to those analyses, we can conclude that previous researches have found out a positive relationship between house prices and neighbourhood.

**Hypothesis**

Based on literature review that we have discussed above, we believe that location of neighbourhood affects rental price. We hypothesize that location of a neighbourhood affects rental price. Specifically, we hypothesize that neighbourhoods in the north area of Boston affect rental prices more than neighbourhoods in the south area of Boston.

**Methodology**

**Variables**

If we take a look at our dataset, there are 16 variables, and they are divided into two main types of data, numeric or categorical data types. Since the purpose of our analysis is to examine the effect on the variable *price,* we only select variables that we think would give effect to the chosen dependent variable. Other non-selected variables include *ID name,* *listing name, host id, host name, neighbourhood group, latitude, longitude, number of reviews, last review, review per month, and calculated host listing count.*

As a result, five variables will be used. Those variables include *price, neighbourhood, number of minimum nights, number of available listing days per year, and room type.* These variables are also divided into three categories. We have *price* as our outcome variable and *neighbourhood* as our explanatory. Finally, we choose the *number of minimum nights, number of available listing days per year, and room type* as our control variables. We believe that the explanatory variable and control variables are important variables and have an effect on the rental prices.

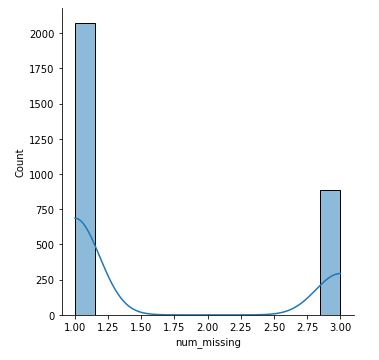
**Data cleaning**

Before we proceed in our analysis and prove or disprove our hypothesis, data cleaning is our first important step. We use Python on Jupyter Notebook to perform our data cleaning. As shown in the Figure 1 below, we see that there are three columns, *neighbourhood group, last view, and reviews per month* that have missing values. We also create a histogram for missing values as shown in Figure 2. Additionally, we use a for loop to find the percentage of missing values in each column. As shown in Table 1 highlighted in yellow, all values are missing in the column *neighborhood group* and thus 100 percent of missing values. Moreover, columns *last\_review* and *reviews\_per\_month* have 30% of missing values in each column.

**Figure 1** showing missing values in listing dataset in Boston



**Graph 2** showing histogram plot for missing values



**Table 1** showing percentage of missing values in each column

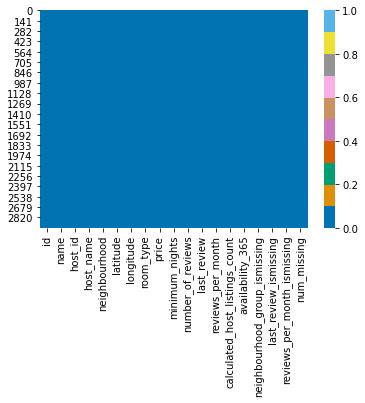
|  |  |
| --- | --- |
| **Variable name** | **Percentage of missing value** |
| id name | 0% |
| name | 0% |
| host\_id | 0% |
| host\_name | 0% |
| neighbourhood\_group | 100% |
| neighbourhood | 0% |
| latitude | 0% |
| longitude | 0% |
| room\_type | 0% |
| price | 0% |
| minimum\_nights | 0% |
| number\_of\_reviews | 0% |
| last\_review | 30% |
| reviews\_per\_month | 30% |
| calculated\_host\_listings\_count | 0% |
| availability\_365 | 0% |

**Data Handling**

We will handle missing values in each column differently. First, we decide to drop the *neighbourhood group* column since there is no value at all. Second, we replace the missing values in the *last\_review* column with '\_MISSING\_' since the variable of this column is a categorical type. Finally, we replace the missing values in *reviews\_per\_month* with its median value as the variable of this column is a numeric type.

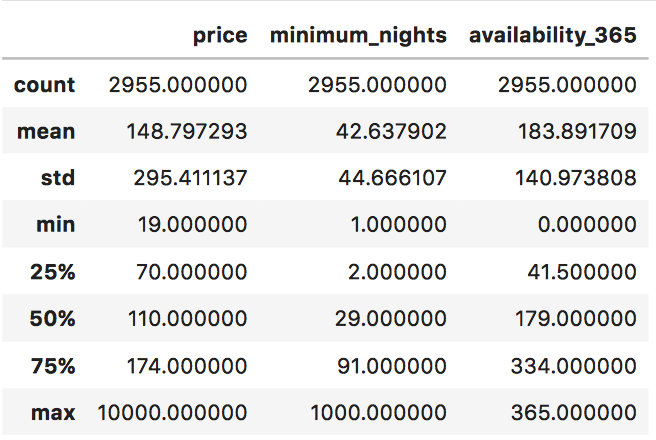
Figure 3 below shows our dataset after we handled the missing values. The dark blue color shows that there are no missing values in any row of the dataset.

**Figure 3**

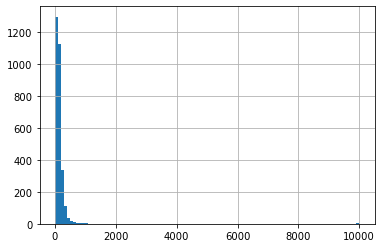
****

We created a summary statistics of the numeric variables in our chosen variables. As shown in Figure 4, the minimum rental price is $19 dollar per night while maximum price is $10,000 dollar per night. The number of minimum nights is 1 and the maximum minimum night is 1000. Figure 5 and 6 showing the boxplots of price and number of minimum nights. Each boxplot shows that there are outliers in each variable. Additionally, the figures also distributions of each variable in each bar graph. Figure 7 shows no outlier in the column and the bar graph shows the spread of the number of available listing days throughout a year. As we mentioned, there are five variables chosen for our analysis, in which three of them are numeric variables, and other two are categorical variables. Figure showing bar graph of neighbourhood name in Boston. The graph shows that the top 5 neighbourhoods with high listing are Dorchester, Downtown, Jamaica Plain, Roxbury, and Brighton. These neighbourhoods are known to be the most popular area for both domestic and international tourism. Similarly, Figure 9 shows the bar graph of room type in Airbnb listing. There are four room types: entire home/apartment, private room, hotel room, and shared room. As shown in the graph, the entire home/apartment is the most popular rental type for Airbnb in Boston with the private room as the second popular room type.

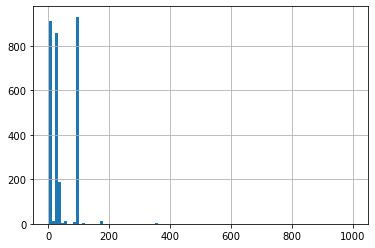
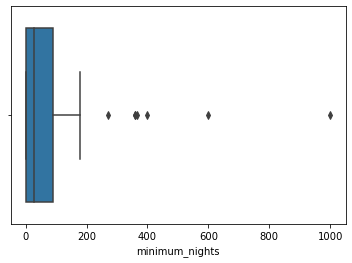
**Figure 4** Showing Summary Statistics of Numeric Variables



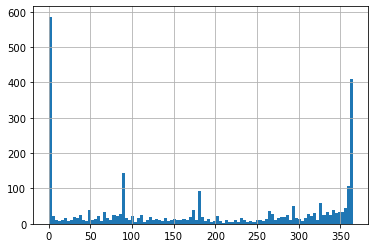
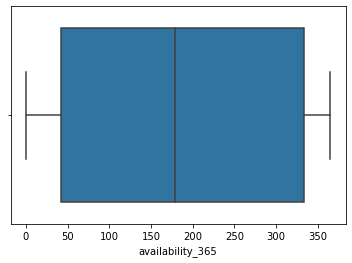
**Figure 5** Showing Boxplot and Histogram Distribution of Price



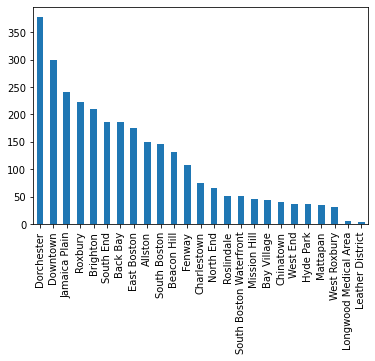
**Figure 6** Showing Boxplot and Histogram Distribution of Number of Minimum Nights



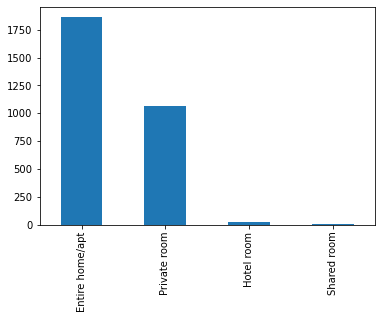
**Figure 7** Showing Boxplot and Histogram of Number of Available Listing Day per Year



**Figure 8** Showing Bar Graph of Neighbourhood Name in Boston



**Figure 9** Showing Bar Graph of Room Type



**Result and Discussion**

**Limitation**

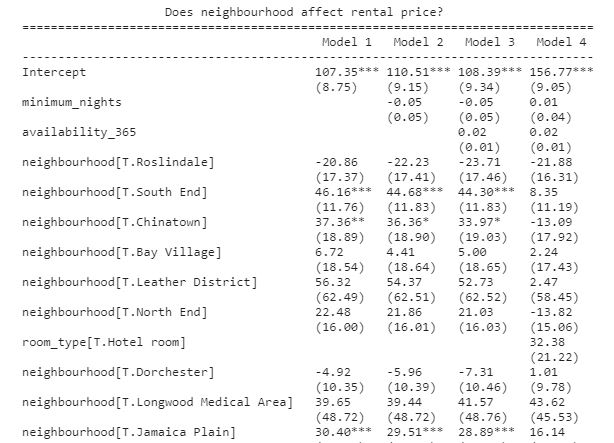
There are some limitations in our analysis. First, our dataset had missing values. We found three columns had missing values. As detaily explained in our methodology, we took care of the missing values by performing and applying a data cleaning method on each column. Second, our dataset is not perfect. There are so many non-significant coefficients between outcome variables and explanatory variables. Third, there are also outliers in our dataset that decrease the value of correlation coefficient and weaken regression relationship. Fourth, our dataset that we obtained from Airbnb website does not provide the year of when this data was being collected. It also does not provide a dictionary that explains each variable in the dataset. This data was compiled and uploaded on the website on February 19, 2021 and we assume that 2018 was when the dataset was being collected. Without the dictionary explaining the meaning of each variable, we also make an assumption of the meaning of our selected variables as well.

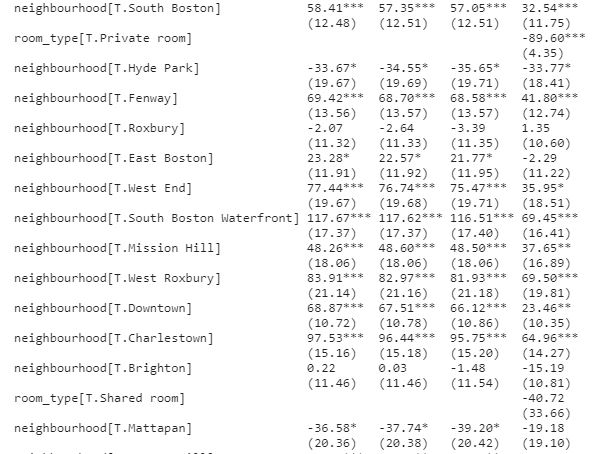
**Correlation Coefficients Table**

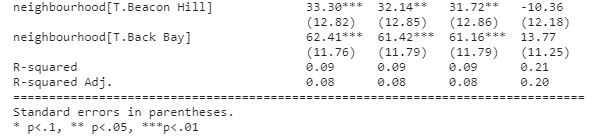
**Table 2** Correlation coefficients in regression analysis

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | |
| **coef** | **p-value** | **coef** | **p-value** | **coef** | **p-value** | **coef** | **p-value** |
| **Charlestown** | 97.5333 | 0.018 | 96.0690 | 0.021 | 98.196 | 0.018 | 71.0725 | 0.087 |
| **Downtown** | 68.8674 | 0.019 | 67.0470 | 0.023 | 71.305 | 0.016 | 33.6899 | 0.262 |
| **Leather District** | 1015.15 | 0.000 | 1015.37 | 0.000 | 1021.9 | 0.000 | 975.5303 | 0.000 |
| **South Boston** | 70.9410 | 0.037 | 69.5452 | 0.042 | 70.534 | 0.039 | 48.7431 | 0.153 |
| **South Boston Waterfront** | 117.673 | 0.013 | 117.605 | 0.013 | 121.03 | 0.011 | 79.5615 | 0.095 |
| **West Roxbury** | 115.435 | 0.043 | 114.282 | 0.045 | 117.29 | 0.040 | 105.4276 | 0.064 |
| **minimum\_nights** | - | - | -0.0722 | 0.557 | -0.0714 | 0.561 | -0.0175 | 0.887 |
| **availability\_365** | - | - | - | - | -0.0489 | 0.215 | -0.0475 | 0.225 |
| **Hotel room** | - | - | - | - | - | - | 31.3029 | 0.612 |
| **Private room** | - | - | - | - | - | - | -78.8764 | 0.000 |
| **Shared room** | - | - | - | - | - | - | -43.3483 | 0.658 |

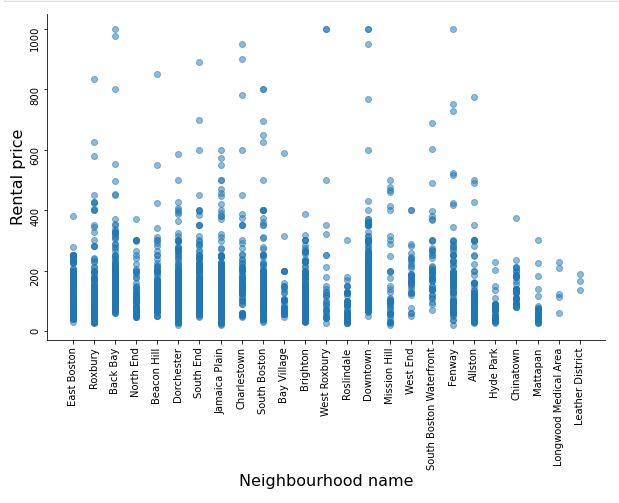
**Figure 10**

****

****

****

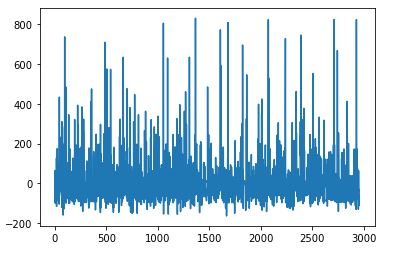
**Figure 11**

****

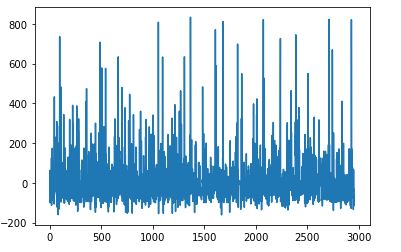
**Residuals**

The residual increase from model 1 to model 4 from 0.02 to 0.04. This shows that there is an improvement in residual.

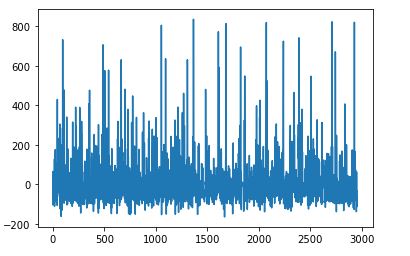
**Residual 1**

****

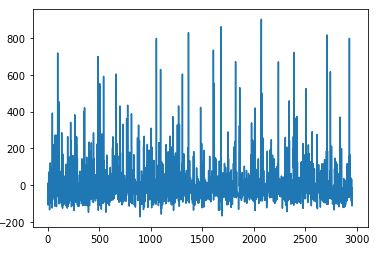
**Residual 2**

****

**Residual 3**

****

**Residual 4**

****

**Regression**

In order to examine the correlations between the outcome variable, the explanatory variable, and the control variables, we used regression models. We created four regression models in total to carefully observe the effects of adding one more control variable to the regression, and from that concluded the mathematical equation for the most effective model. For all four models, our base value is Allston neighborhood. The regression results are shown on Table 2.

In our first model, we only examined the correlation between the outcome variable, price, and the explanatory variable, neighborhood. From the regression result, we found 6 neighborhoods are significant, including Charlestown, Downtown, Leather District, South Boston, South Boston Waterfront, and West Roxbury. Among these neighborhoods, listings belonging to neighbourhood Leather District have the highest increase in rental price, that is by 1015 dollars, holding other variables constant. Similarly, the other variables also have positive coefficient numbers as listed on Table 2. The r-squared value for this first model is 0.027.

For the second model, we added one control variable along with the original two variables, and that is minimum\_nights. The r-squared value remained the same in this model. Again, there are 6 significant neighborhood variables. However, the coefficients of 5 neighborhoods have slightly decreased when a control variable is added, and Leather District is the only exceptional variable that has a slight increase in coefficient. Looking at the control variable, we found out that an increase of 1 unit in minimum nights will decrease the price of rental by 0.07 dollars. Yet, based on p-value, this effect is not significant to the regression model, holding everything else constant.

Moving to the third regression model, we again added one more control variable to the calculation of housing price, in order to examine its effects on the regression results. The control variable that we added this time is availability\_365. When this variable is included, the r-squared value has increased to 0.028. The coefficients of the 6 neighborhood variables have slightly increased again. We also found that an increase of 1 unit in the number of listing days will decrease the price of rental by 0.049 dollars. However, based on p-value, this effect is again not very significant, holding everything else constant.

In our last regression model, we added room\_type as a control variable to the model. The r-squared value for this model has significantly increased to 0.041. For this model, the only significant variable is the Leather District neighborhood variable. The results showed that there is no significant effect to claim that renting a hotel room type or a shared room type will increase or decrease the rental price. However, we noticed that renting private room type significantly decreased rental price by 78.88 dollars. Thus, the final mathematical equation that we can conclude from this regression model is price = \* (neighborhood) +\* (minimum\_nights) +\* (availability\_365) +\* (room\_type) +, where is the random error.

**Conclusion**

After performing those four regression models, we have come to a conclusion of the correlations between the outcome variable, the explanatory variable, and the control variables. Firstly, after controlling for minimum\_nights, availability\_365 and room\_type, the model still shows a significant relation between price and Leather District neighbourhood. Therefore, we can say that Listings belonging to the neighbourhood Leather District had a significant increase in rental price by 976 dollars compared to base variable Allston, holding other variable constant. There was no significant level of correlation between price and number of minimum nights, as well as no significant level of correlation between price and number of listing days. There was also no significant effect to claim that renting a hotel room type or a shared room type will increase or decrease the rental price. However, renting private room type significantly decreased rental price by 78.88 dollars.

Based on the regression conclusion, we have made some suggestions to the clients related to this market of Airbnb in Boston. That is, since the Leather District neighbourhood is the only neighborhood that has a significant effect on rental price, it should maintain its reputation to attract more rental clients. Rental price in Leather District is expensive because of the location of this neighbourhood. Leather District is located near the financial district and Chinatown, which are the popular places for tourists and visitors. On the other hand, those neighbourhoods that have yet to have any significant effect on price should try to improve their neighbourhood standard so as to be able increase their rental price. Lastly, because renting private room type costs less rental price compared to other room types, rental clients should consider renting private room type to save their budget.

**References**

“Inside Airbnb. Adding Data to the Debate.” *Inside Airbnb*, insideairbnb.com/.

“Chinatown-Leather District.” *Boston.gov*, 12 July 2016, [www.boston.gov/neighborhood/chinatown-leather-district](http://www.boston.gov/neighborhood/chinatown-leather-district).

“Are You Paying Too Much For A House In Your Neighborhood?” Trulia's Blog, October 14, 2019. <https://www.trulia.com/blog/home-prices-by-neighborhood/>.

Gomez, Joe. “8 Critical Factors That Influence a Home's Value.” Opendoor, September 19, 2019. <https://www.opendoor.com/w/blog/factors-that-influence-home-value>.

Lee, Chun Yang, and Hui-Yu Lin. “The Impact of Neighborhood Characteristics on Housing Prices .” Accessed April 5, 2021. <http://www.aessweb.com/pdf-files/31-44.pdf>.