Predicting Listing Prices on Airbnb during the Summer

Industry: Consulting, Tourism

Team C36: Brandon Zhang, Zoey Hart, Xuebin Zhu, Yashasvi Kathotia, Joanna Hou

Introduction

The datasets we have chosen to analyze contain observations of Airbnb listings from Singapore and New York City. Airbnb is an online marketplace that allows users to rent various types of lodging, such as apartments, homes, and rooms. Airbnb's business is expanding at an exponential rate with more listings added to the website every month. We are particularly interested in the Singapore and New York markets for two key reasons:

1. New York and Singapore are both popular destinations for tourists as well as work related visits. They are both metropolitan cities with a large market for homestays, and are the cultural and financial epicenters of their respective countries with comparable geographic size and population. As a result, they make for a good comparison to explore if there are any differences between the US market and the Asian market for short-term rentals.
2. These are both cities that we either would like to travel to or have lived in, hence we are familiar with the markets and the location, enabling us to curate more nuanced insights about the significance or contextual implications of our findings.

Business Context

In this project, we would like to explore what factors affect the listed price of a house on Airbnb. Specifically, our business goal is to create a model that can predict the price of listings in both Singapore and New York City. This will provide hosts with more information about how they can improve their listing information and set their prices to attract more stayers. For renters, our model can be a useful tool to check for any overpriced deal when searching for an Airbnb in either city and can help ensure the quality of a considered listing.

The Dataset

The datasets[1] were retrieved from Inside Airbnb, which sources publicly available data from the Airbnb site. This ensured the reliability of the data used for our analysis.

The datasets contain 106 columns including information about price, review scores for listings, descriptions, neighborhood, type of accommodation, etc. The types of variables were both qualitative and quantitative and required a great deal of data cleaning and preparation.

Furthermore, the limitation of these datasets is that they are specific to particular dates. The Singapore dataset was retrieved on June 25th, 2019 and the NYC dataset was retrieved on July 9th, 2019. This poses several issues such as inaccurate information and limited applicability; since the data was pulled on specific days it does not account for the fact that prices can fluctuate during the year based on seasons and holidays. Therefore, our model will not accurately reflect these variations over the year. Our model could be improved if we could incorporate variables such as

[1] Inside Airbnb. (2019). *Inside Airbnb. Adding data to the debate.*. [online] Available at: http://insideairbnb.com/get-the-data.html [Accessed 20 Jul. 2019].

seasons, day of week, and booking records to capture price fluctuations over the course of the year, making our model more adaptable and applicable.

Data Cleaning

Our first step in the data cleaning process was converting the variables to the desired type and format so that that could be used in our model. For variables such as “Total Price”, “Cleaning Fees” and “Extra People” we used the gsub() function to drop the “$” sign and any another non-numeric characters to converted the them to the numeric data type. Since “Total Price” is our dependent variable, we investigated it further and removed any outliers to make it more normally distributed. We then dealt with missing values and variables of different types, such as strings, integer, factor and numeric. We first identified variables with null values and missing values to evaluate their impact on the data; “Host Response Rate” and “Review Score Rating” had the highest number of null values. However, these missing values could not be imputed or estimated based on other variables, hence we had to drop the missing values using the drop.na() function.

Exploratory Data Analysis

Total Price Distribution Across Neighborhood Groups

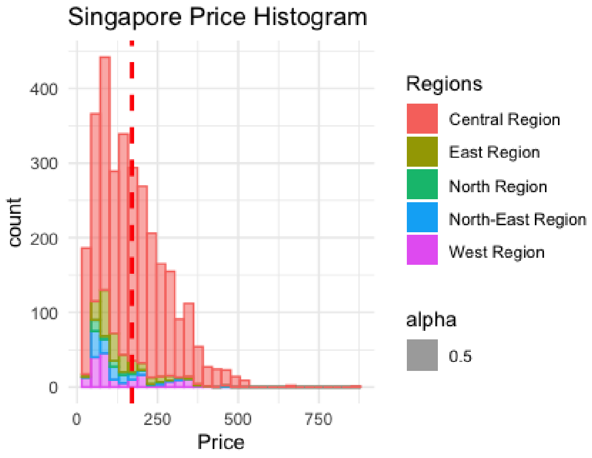

Figure 1.

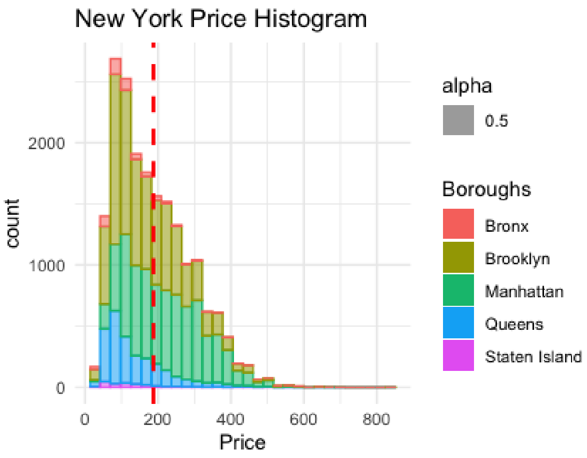

Figure 2.

These histograms show the count of listings in each of the neighborhoods in Singapore and in NYC (colloquially referred to as “regions” and “boroughs”) as well as their respective prices.

Singapore:

Based on this information it can be seen that rentals in the Central and West Regions are higher priced than in other regions and have the largest range in price. There are also a larger number of upper end outliers in the Central Region that indicate higher end rentals are available in this area. This coincides with what we know about Singapore’s city sectors; the Central Region is the downtown metropolitan area of Singapore and therefore a popular area for tourists and an expensive, high-density area to live in. This is also the most populous region in Singapore, which would validate the large range in prices as there are lots of different properties available. The West

Region is Singapore's largest and second most populous region and is primarily residential, and as a result there is also a large range of prices because of the copious living accommodations.

New York:

In New York, Manhattan has the highest and largest price range out of the boroughs, followed by Brooklyn. In addition, all the groups show a significant amount of upper end outliers which indicates that highly priced luxury housing units are available in all areas. Manhattan is the most densely populated of the boroughs and the commercial and administrative center of the city. It is also the most expensive place to live and the most popular tourist destination in the United States. Therefore, it makes sense that Manhattan shows the highest prices of the five boroughs along with the highest outlier prices as well. Brooklyn is the most populous of the boroughs and has been growing in popularity in recent years, which has led to a dramatic increase in house prices reflected in the high prices and large count of listings show the histogram.

Similarities:

Given the two histograms and neighborhood analyses, it can be seen that there are many similar tendencies between the two cities. The commercial and financial district of each city is the most expensive area (Central Region and Manhattan), and the next priciest groups are the heavily populated residential areas located adjacent to the downtown districts (West Region and Brooklyn). These two group types also contain the vast majority of the rentals available as well as the most upper end outliers.

Room Type

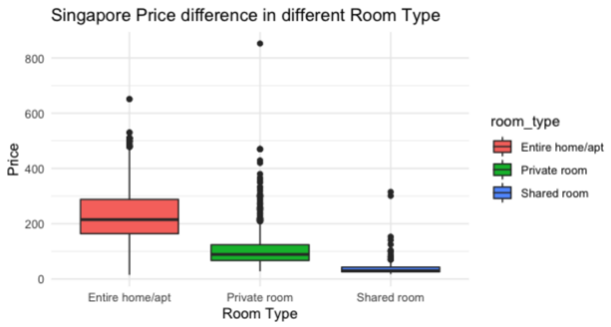

Figure 3.

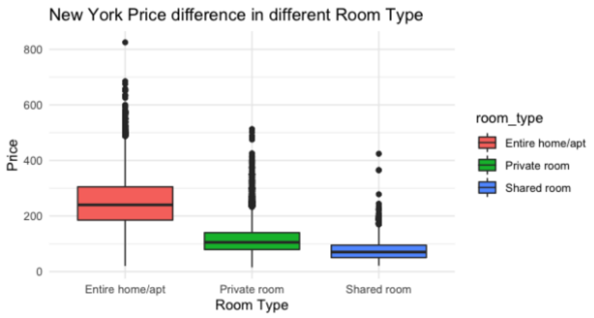

Figure 4.

The two boxplots illustrate the difference in total prices across room type in Airbnb listings.

Similarities:

It can be discerned that the total price listed for an entire home or apartment is higher than the listed price for a private room or shared room for both cities. The total price range for entire homes and apartments is also much larger than the other room types, suggesting that customers have a varied range of price options when choosing an apartment or home. This finding is logical because an entire apartment would likely cost more than just a private room or a shared room given the extra costs of maintaining, cleaning, and renting an entire apartment from the renter's perspective. It also is reasonable that private rooms are more expensive than shared rooms because in a private room you do not have the option of splitting the rent, as is the case for a shared room.

Differences:

However, what is interesting in this data exploration is that when compared the average listing price for all three property types is higher in New York than in Singapore. This finding is contradictory to the knowledge that Singapore is a more expensive city than NYC. Further analysis would be required to reconcile this finding or discover other factors causing this discrepancy.

Property Type

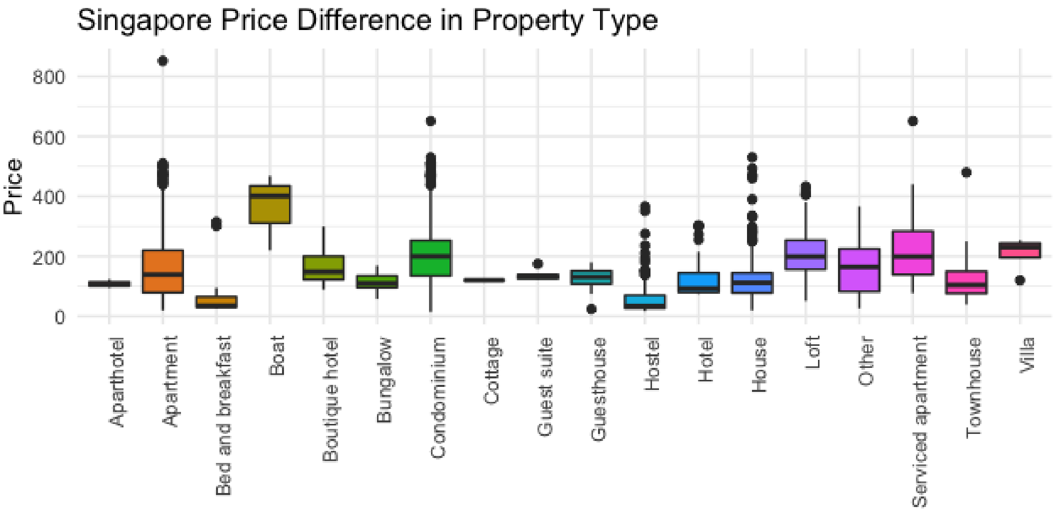

Figure 5.

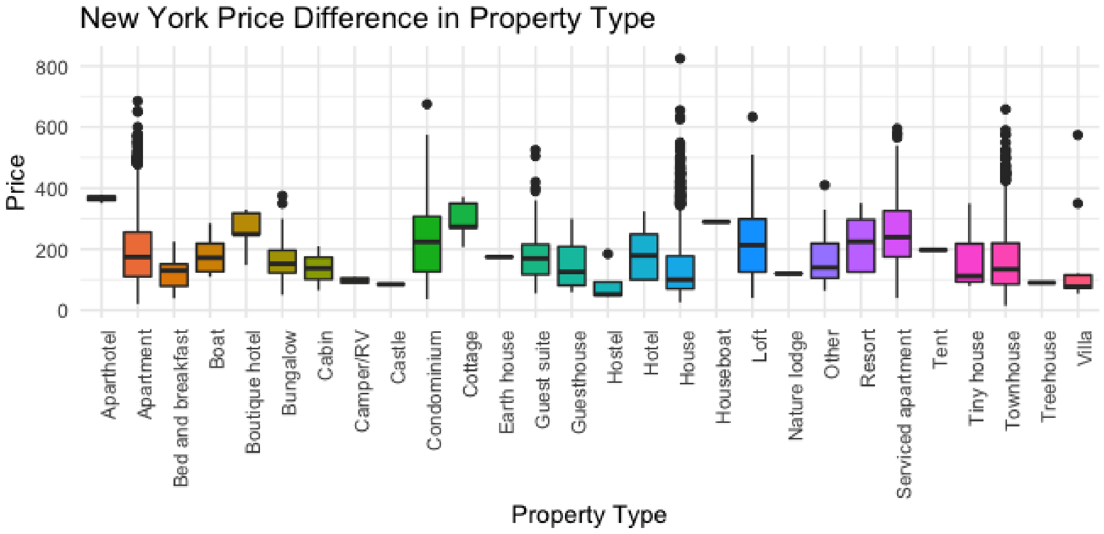

Figure 6.

Singapore:

The most expensive property type in the Singapore Airbnb market is "Boat" which has an average price of around 400 Singapore dollars. This is probably due to the unique experience of living on a boat and the scarcity of boat listings on the website. This type is followed by apartment types (serviced and regular), “Botique Hotel”, “Condominuim”, “Loft” etc. The least expensive property types in the Singapore Airbnb market are “Beds and breakfast” and “Hostel” which have an average price of around 50 Singapore dollars. These types of accommodations involve sharing the space with hosts and/or other strangers but are more affordable for students and travelers who are short in travel budget.

New York:

The above boxplot shows more varieties of property type in the New York Airbnb market in comparison to the Singapore market. However, since there is too little information and data about niche property types including “Camper/RV”, “Castle”, “Earth house”, etc., we are unable to make any valuable interpretations on the price differentiation. Other than that, the relatively more expensive property types in the New York Airbnb market are various apartments, “Boutique Hotel”, “Loft”, “Condominium”, and “Resort”. This is largely because these property types are usually located in the bustling areas with convenient transits. Property type such as “Villa”, usually located in the countryside, do not have higher prices as it will be difficult for travelers to transit to downtown areas. Unlike Singapore, which is a smaller city with higher density and convenient transportation, the location of the properties matters more (which reflects in price) in New York. However, it can still be concluded that there are some similarities between cities in property types, especially regarding the types with higher prices.

Interaction Exploration

Number of Bedrooms VS Room Type

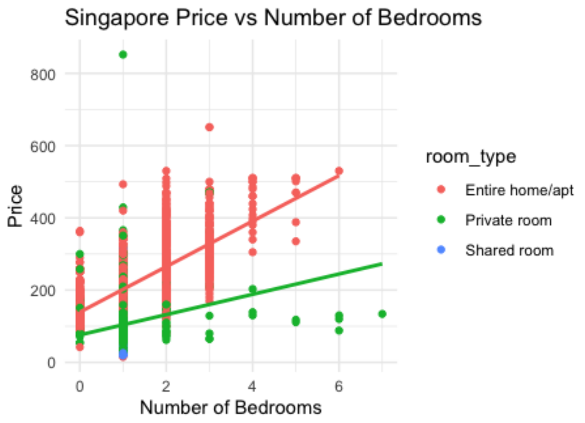

Figure 7.

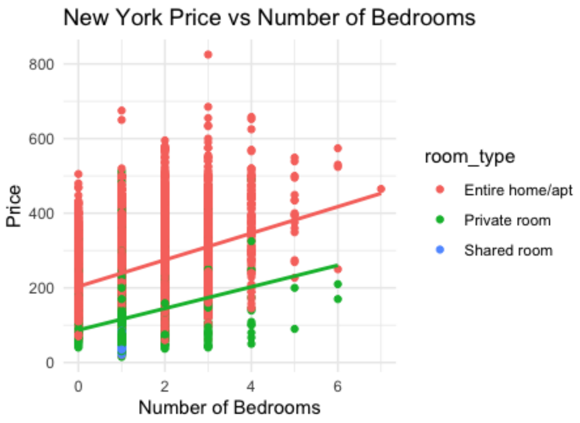

Figure 8.

Singapore:

We grouped the data by room type to explore the effect of room type on the relationship between number of bedrooms and the price of an Airbnb in Singapore. Based on the graph, entire home/apt have a higher rental price than private room. In addition, while both room type groups (shared rooms do not have sufficient data to plot) reveal a positive relationship between the number of bedrooms and the rental price, the slopes of the two lines are significantly different. This indicates that there is a possible interaction between room type and the number of bedrooms when determining price.

New York:

In contrast, the New York room type graph indicates that there is almost no interaction between room type and the relationship between number of bedrooms and the price of an Airbnb. Both room type groups (shared rooms do not have sufficient data to plot) reveal a positive relationship between the number of bedrooms and the rental price with a similar slope.

Trends with "Superhost" Status

When further considering our objectives for our model, we agreed that we not only wanted a comprehensive understanding of what effects listing prices but also an easy way to ensure a high quality and effortless renting experience. We reviewed the significant variables in our model and determined that review scores rating and host response time were important not only for price but also for the customer experience. In searching for commonalities between listings with high rankings in both of these categories, we discovered that they seem to both be correlated with the variable "is host a superhost", another significant variable in our regression.

Airbnb differentiates its members by awarding its exemplary hosts "superhost status". When examining the data, these are the trends we discovered amongst the variables.

Review Scores

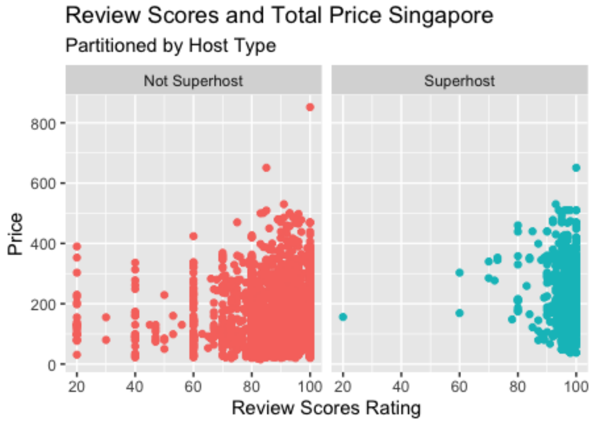

Figure 9.

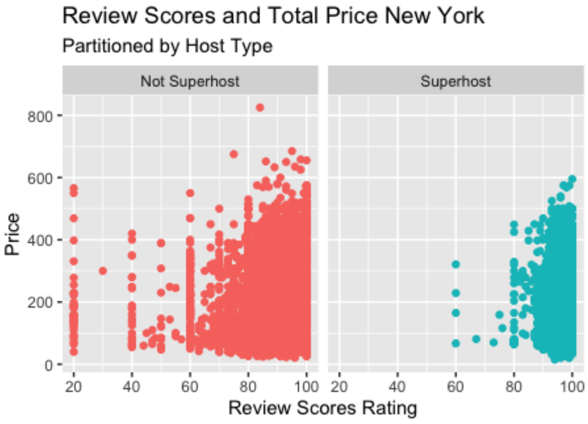

Figure 10.

Similarities

In both Singapore and New York, it can be seen that the majority of review ratings are over 60%. In fact, the interquartile range for Singapore is 85-100 and for New York is 92-99, illustrating the majority of our listings are already up to ideal standards. However, if a customer wanted to only view ratings above 80%, choosing to look at only listings owned by a superhost is a near perfect way to eliminate all unsatisfactory options. In addition, there is still a range of prices available when this is taken into consideration, so renters of all different budgets can still benefit.

Response Times

	A few days or more	*Within the day*	*Within a few hours*	*Within the hour*
Singapore				
Not a Superhost	0.0149	0.0681	0.2142	0.7028
Superhost	0.0000	0.0106	0.1472	0.8422
New York				
Not a Superhost	0.0292	0.1627	0.2251	0.5830
Superhost	0.0015	0.0769	0.1886	0.7329

<u>Similarities</u>

In both Singapore and New York, it can be seen that the proportion of hosts responding within the hour is higher for superhosts than for regular hosts and that they are highly unlikely to take more than a few hours to get back to the customer. In addition, we ran a Pearson's Chi-squared test to check if there was indeed a difference between the response time of superhosts and non-superhosts. The output confirms our statement.

In the Singapore data the superhosts account for 21% of the total listings and in the New York data they account for 30%, and in both cases being a superhost is a significant factor to determining price and the variables have positive coefficients.

In conclusion, if a customer wants to expedite their search to browse listings that have higher ratings and better response times from hosts, they should limit their listings to only those with a superhost owner. The correlation between these variables and the superhost identifier allow us to make this distinction.

<u>Modeling</u>

Full Model

At the beginning of our research, we obtained a data set with 106 variables for both Singapore and New York City. After careful examination, we noticed that lots of variables in our original data set may not be able to explain our dependent variable, which is the total price. Therefore, we need to eliminate these unnecessary variables.

Intuitive Variable Selection Model

Having explored the different variables in the datasets, we used intuition and business context to identify independent variables that could potentially have an effect on price. By discussing and analyzing how factors such as host response rate and room type can impact the price of the listing, we were able to narrow down the independent variables from 106 to 42. We eliminated variables that provided no extra information to our model, such as “id” and “listing url”, and certain qualitative variables that were difficult to include in our model, such as strings like “description” and “summary”. The string variables could not be coded into numeric or categorical data due to there being no pattern in the descriptions of the house.

Having narrowed down are search for independent variables, we first computed our dependent variable named “total price” which is defined as the sum of the price and the cleaning fee.

Logically, we thought it made most sense to add the two values because it was a mandatory payment that the renter would have to pay.

Stepwise Regression

We then used the Stepwise Regression Method and gained two multiple linear regression models to predict the Airbnb price in Singapore and New York City. At the beginning, we had 42 independent variables of interest. Using AIC as the selection standard in two-way direction, we obtained a model with 26 independent variables to predict the total price in Singapore and a model with 31 independent variables to predict the total price in New York City.

The model for Singapore outputs a F-statistic of 170 on 50 and 3017 degrees of freedom and the p-value for this model is less than 2.2e-16, indicating that the overall model is statistically significant. Furthermore, this model has a 0.738 multiple R^2 value, which means this model can explain 73.8% of the variability in price. On the other hand, the model for New York City outputs a F-statistic of 529.5 on 68 and 19052 degrees of freedom and the p-value for this model is less than 2.2e-16, indicating that the overall model is also statistically significant. This model also has a 0.654 multiple R^2 value, which means this model can explain 65.4% of the variability in price.

However, these high R^2 values are likely due to the large number of independent variables included in the model; we may have issues of overfitting and explaining too much on the sample rather than the population. Besides, as the data visualization output illustrates, some independent variables may have interaction effects on the others, but they are not considered in these two models. Last but not least, we still need to generate plots and calculate other statistics to prove that our model indeed satisfies the basic assumptions of linear regression. Therefore, further adjustments are needed for these two models.

Final Model

Finally, we again used business context and intuition to narrow down our variables to 11 for Singapore and 12 for New York City to safeguard our model from the risk of overfitting. In addition, we also discovered a few independent variables with problematic values and removed them accordingly.

- We removed the variable “square feet” because there were only 33 valid observations in the Singapore dataset and 408 in the NYC dataset. There was no other way to make any valuable interpretation or prediction on the missing values.
- We removed the variable “Host has profile pic” in the Singapore dataset because all values under this column were “TRUE” thus providing no additional information.
- We removed the variable “security deposit” because even though the variable was significant it has values that don’t logically make sense when compared to the price of that listing. For example, a third of the data had security deposits over a $1000 but the most expensive listing was only $800 in NYC. We were concerned about it inaccurately influencing the total price.

The variables we chose all have a significant effect on listing price and intuitively we believed were good predictors. We also ran a final regression including the interaction terms we explored previously, but we discovered the coefficients were not significant in determining price so we disregarded them in the final model. To be thorough, we calculated the variance inflation factor

(VIF) to check if we should include any interaction term in our models. None of the VIF values for both models are greater than 5, indicating no multicollinearity problem.

Models:

Singapore

$$\begin{aligned} \mathit{Total\ price} = {} & \text{Room type} + \text{Neighbourhood group} + \text{Bedrooms} + \text{Propety type} \\ & + \text{Guests included} + \text{Cancellation policy} + \text{Total listings count} + \text{Number of reviews} \\ & + \text{Scores rating} + \text{Host is superhost} + \text{Host response time} + \text{Extra people} \end{aligned}$$

New York

$$\begin{aligned} \mathit{Total\ price} = {} & \text{Room type} + \text{Bedrooms} + \text{Extra people} + \text{Host is superhost} + \text{Guests included} \\ & + \text{Total listings count} + \text{Number of reviews} + \text{Neighbourhood group} \\ & + \text{DayDiff reviews} + \text{Scores rating} + \text{Host response time} + \text{Extra people} \end{aligned}$$

Differences in the Final Models

After calculating the final models, we noticed some interesting deviations between the two city datasets and their important explanatory variables. For example, we included property type and cancellation policy in New York because they are highly significant, but this is not the case in Singapore. This could be because there is a larger variety of property types in New York which creates greater differences in total price. The NY data also has stricter tiers of cancellation policies available for hosts and we speculate this greater separation could explain the difference in significance. These are just a few of the insights that can be made studying the differences in model makeups between the two cities.

Evaluation

We constructed our final regression by improving the following models:

- **Intuitive Variable Selection Model** provides further intuitions on the Airbnb market in summer in business context. It is highly interpretable, but the relatively smaller R^2 Value implies that the model is still not accurate enough to capture the whole picture.
- **Stepwise Regression Model** provides more accurate explanation of the data but is not intuitive enough or easily comprehensible. Moreover, it still includes too many variables that are not significant or informative.
- The **Final Model** provides the most accurate and intuitive explanation to the data. It should be noted that the final Singapore model has a higher AIC than New York because there are less observations available.
- On analysis of **Residual Plots** for both models, there are no outliers damaging the validity of our models, assumptions of constant variance, independent error terms, and no multicollinearity and normality of residuals are all met. To make sure the residuals are actually normally distributed, we ran a normality test on our residuals. We got p-value < 2.2e-16 for both models. Therefore, we can say the normality assumption is met.

Model	# of variables	R^2 Value	Adjusted R^2 Value	AIC
Intuitive Variable Selection Model				
Singapore	41	0.7388	0.7333	24513.05
New York	41	0.6629	0.6615	156378.00
Stepwise Regression Model				
Singapore	26	0.7380	0.7336	24496.85
New York	31	0.6540	0.6527	156858.80
Final Model				
Singapore	11	0.7025	0.7009	24820.37
New York	12	0.6096	0.6086	159126.20

Conclusion

These models allow a consumer to evaluate the price of their Airbnb listing depending on the characteristics they desire. The insights gathered will expedite the renter search and allow for well-informed decision-making and a better understanding of the Airbnb market. In addition, this model does not breach any privacy rights of consumers or hosts on Airbnb; in fact, it may help create a more level playing field in this market as customers will be more aware of what they are getting when they choose to pay a certain price for a rental, eliminating imperfect information.

Furthermore, these models allow us to compare two diverse locations in the same market. We found many similarities between the two cities, but there were also some interesting deviations such as the significance of certain variables in one location vs. another, which led to interesting insights about the US and Asian Airbnb markets.

However, it must be noted that these models are a snapshot of each market in a particular point in time; they do not reflect the fluctuations in prices throughout the seasons which are highly probable given the trends seen in the hospitality industry. Further analysis with more variables would be necessary to develop better models.

Bibliography

Inside Airbnb. (2019). *Inside Airbnb. Adding data to the debate.*. [online] Available at: http://insideairbnb.com/get-the-data.html [Accessed 20 Jul. 2019].

Airbnb. (2019). *Holiday Lets, Homes, Experiences & Places - Airbnb*. [online] Available at: https://www.airbnb.com/ [Accessed 20 Jul. 2019].