ANLY 502-91: Analytical Methods I

Team: Anonymous

Team Member: Hsiao-Ting Tseng

**New York City Airbnb Listing Pricing Analysis**

**Introduction**

Nowadays, Airbnb is one of the best examples of sharing economy. Many people rent or share their home to make money through Airbnb. In New York City, there are more than 40,000 listing on Airbnb. The listing prices vary because of different locations, boroughs and room type. What is the key factor which influences Airbnb listing price most? How do the Airbnb hosts to set prices for their room?

The dataset was released on July 12th, 2017 from <https://s3.amazonaws.com/tomslee-airbnb-data-2/new_york.zip>. This dataset includes 41,245 observations and 21 variables. The primary goals of this analysis are to find out the main factors which affect Airbnb listing price most and build a model to predict listing prices in New York City.

**Exploratory Data Analysis**

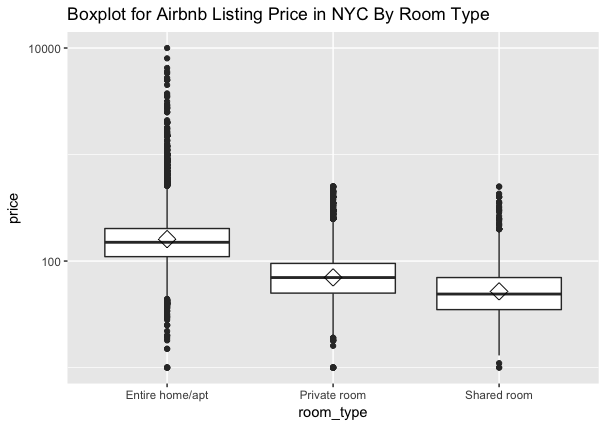
First of all, before analyzing how specific variables influence listing prices, we can have a clear overview about Airbnb listing price in New York City from the following histogram. According the histogram below, we can see price range is huge which is from $0~$9,999 and the mean price, $136, is larger than the median, $100.



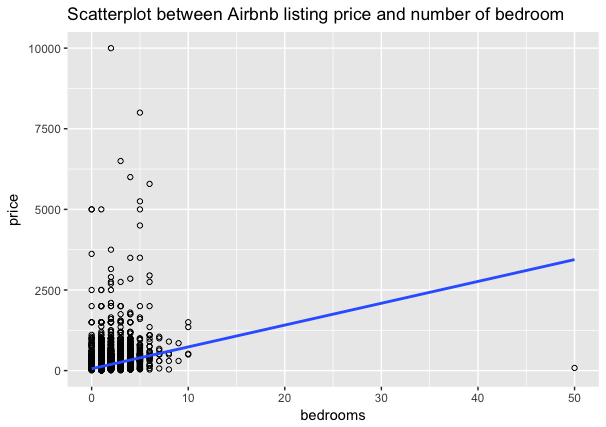
In the second part, we analyze the relationship between listing price and each variable. From the

boxplot for Airbnb listing price in NYC by room type, compared to private room and shared room, the entire home/apt has a higher median price, $150, and the higher average price, $194.32. Because of many outliers for entire home/apt, the standard deviation is 216.02, which is way larger than the standard deviation of private room or shared room. The price range of private room and the price range of shared room are similar. Also, their standard deviations are closed.

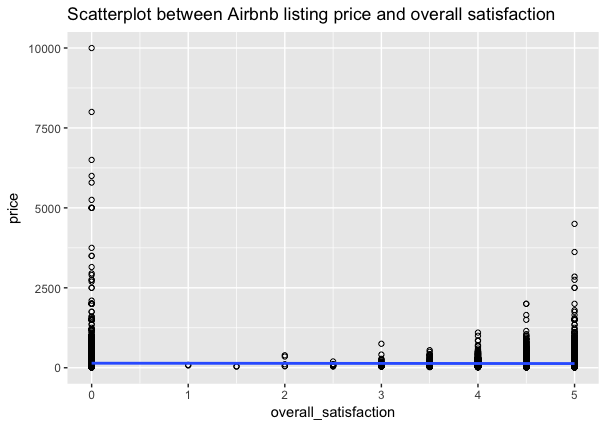
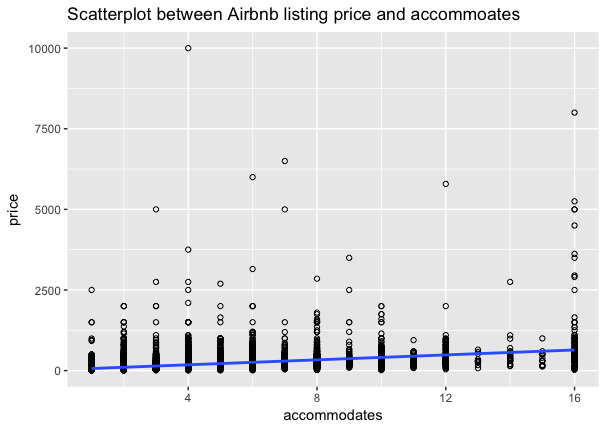
According to the boxplot for Airbnb listing price by borough, Manhattan has the highest median price, $125, and the highest average price, $167.85. Both of Manhattan and Brooklyn have the huge price range and many outliers. Interestingly, Staten Island has the largest standard deviation, 314.17.



In this analysis, we also want to know if the distance between listing and the Empire State Building, the highest building in New York City, affect listing prices so we add distance between listing to the Empire State Building as a variable. From the scatterplot 1, the relationship between price and distance is negative and weak because correlation is -0.21. However, the p value is less than 0.05 so the relationship between price and distance is statistically significant. From the scatterplot 2, the relationship between price and number of bedroom is positive and weak because the correlation is 0.3. The p value is less than 0.05 so the relationship is statistically significant. About the relationship between price and accommodates, correlation is 0.42 so the relationship is modest. Last but not least, correlation between price and overall satisfaction is only -0.02 so the relationship is negative and very weak.



Scatterplot1 Scatterplot 2



Scatterplot 3 Scatterplot 4

**Inference**

In this analysis, the following are our hypotheses.

H0 : Bi = 0 when other explanatory variables are included in the model.

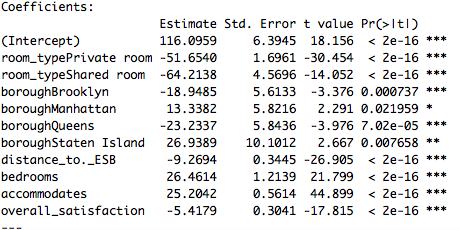
HA : Bi =/= 0 when other explanatory variables are included in the model.

**The “Best” Model**

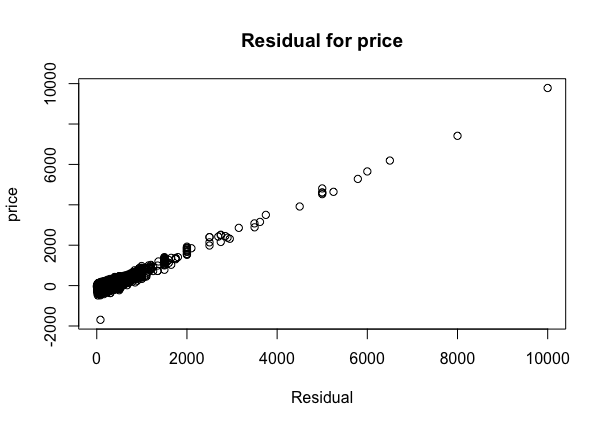
The initial model to predict Airbnb listing price in New York City is shown as below:

Price= y0  (room type) +y1 (borough) + y2(distance between listing and the Empire State Building) + y3  (number of bedroom) +y4 (accommodates) + y5 ( overall satisfaction)+e

According to the following chart below, all of p value of each variable are less than 0.05. We decided to use R2 Approach to select the best model.



After four stages of selection, each variable pass the ANOVA test and p values are less than 0.05, which means rejecting the null hypothesis. Besides, all the residual plot, shown as below, looks near normality. Therefore, we conclude the best model in the following:



Private room: Price= -52.9181(room\_type) + -11.9816  (distance between listing and the Empire State Building) +26.5802 (number of bedroom) + 24.2521 ( accommodates)+111.8427

Shared room: Price= -63.65(room\_type) + -11.9816  (distance between listing and the Empire State Building) +26.5802 (number of bedroom) + 24.2521 ( accommodates)+ 111.8427

**Conclusion**

This paper is about the key factor which influences Airbnb listing price in New York City and building a model to predict the listing price. We use the multiple linear regression to build model and select the best one by R2 approach. However, all of models we build from this dataset may not predict the precise listing prices because R2 for all models are less than 0.7. On the other hand, there are many factors will affect price such as amenities (TV, WIFI, dryer) and other features about the listings so information like amenities and listing description would be meaningful in this analysis. However, the current dataset is lack of listing description, amenity information. The bigger dataset with more variables will be helpful to develop a better model to predict precise listing prices.