Airbnb Listing Analytics in Boston and Beijing

Github Repository: https://github.com/elvirazyyan/BA888Team4B.git

Xiaohan Mei, Yuhong Lu, Ziyan Pei, Peng Yuan, Mengqing Zhang, Jiayuan Zou

1. Problem Statement, Setting, and Why it is Important

Airbnb is an online marketplace that arranges or offers primarily homestays or tourism experiences. Instead of owning any of the real estate listings, Airbnb acts as a broker, receiving commissions from each booking, and connects the hosts and the customers. Airbnb is not just a booking platform, it wishes all their hosts can connect with potential guests on an emotional level and hone in on what makes their rental one of a kind.

This study analyzes the comparison of demand and supply between Airbnb homestays in Beijing and Boston by establishing a supervised machine learning model of housing price, obtaining the influencing factors of the prices in the two cities, and implementing unsupervised machine learning models to build Airbnb homeowners(hosts) profiles. With the techniques above, we can dig into the external differences in the Airbnb marketplaces between China and the United States, and provide reference opinions for homestay hosts in the pricing area. At the same time, through sentiment analysis of homestay reviews, this study hopes to obtain relevant factors that affect the preferences of Airbnb consumers in China and the United States, which helps hosts to improve their quality of homestays and be in comparative advantage in market competition.

The reason why Airbnb is chosen as the topic is that it has become extremely popular in recent years and is changing people's travelling habits step by step. More people choose to look for homestays instead of traditional hotels, such as hotels on booking.com, and short-term rentals. Since 2008, Airbnb’ s market share is continuing to grow. Meanwhile, while searching for this topic online, there are very few studies that compare two countries' Airbnb markets together. And since all the group members come from China, we decided to do an in-depth study and find out similarities and dissimilarities between China and the United States Airbnb marketplace.

Since both China and the United States have a really large national territorial area, comparing the whole two countries won't be very effective as there are many dissimilarities between each region, and some remote areas may affect the results, thus Beijing and Boston are chosen as the primary focus. As Beijing is the capital of China, and Boston is the city where all the group members are living and studying right now, both of these two cities are familiar for the group members to study and discover. Further, many similarities have been found between Beijing and Boston during the study. Firstly, their geographic positions are very similar, which makes these two cities have alike weather. For instance, there is lots of snow during winter which may affect the price of the homestays market. Secondly, Beijing is known for its dense academic atmosphere by having lots of great universities, and there are also many universities or liberal arts colleges in the greater Boston area. Both of these two cities would be busy around each year's graduation season and the demand for the homestays will also increase highly. Other than that, Bird's Nest Olympic stadium and TD garden also attracts people from all over the world to visit these two cities around sports season. All these common points build up to two similar marketplaces, and we are looking forward to having more findings.

2. Dataset

a. Description

Beijing and Boston Airbnb datasets both have three sub-datasets: listings dataset, reviews dataset and calendar dataset.

Listing datasets include all homestays in each city since 2009, which record the basic information of houses, such as the number of bedrooms/bathrooms, house types, longitudes and latitudes. Specifically, Beijing listings data contains 108 variables with 25,921 unique listings, and Boston listings dataset contains 106 variables with 3,501 unique listings.

Both Beijing and Boston reviews datasets have 9 variables, including reviews that have been given to each unique listing. Beijing reviews data has 180,561 observations and Boston's data has 148,151 observations, with a time range from 2009 to 2019 without missing values.

Beijing and Boston calendar datasets also contain 9 variables. Each record reports everyday prices on each day of the following year for each listing from Dec. 2018 to Dec. 2019. There are 9,461,165 observations in Beijing data and 2,269,205 in Boston data without missing values.

b. Missing Values Analysis

Both Beijing and Boston Listing datasets have missing values. First, we evaluate the latent business meanings and effects of variables, then we delete variables that have over 60% missing values. For the rest of missing values, to best preserve the information that the dataset provides, we used to adopt the MICE function to calculate estimated values of missing data by regression and classification. However, the MICE method is time-consuming and has high computing-cost. After comparing the results from substitutions made by column means based upon listings' neighborhoods and house type, we then decide to adopt the mean method to be efficient.

c. Data Cleaning

For variables that do not contain useful information, such as URLs of web pictures and variables solely composed by one unique value, we remove them. Then, for the "amenities" variable which lists all types of homestays' facilities in one text paragraph, we manually extract keywords from each text paragraph and transform each keyword into a categorical variable to indicate the availability of such facility (such as Wi-Fi, TV, or Pets allowance) in homestays. We also shrink the geographic scope of Beijing datasets into places within Fifth-Ring Road to exclude villages. Last, we transform all categorical variables into dummy variables.

d. Feature Selection with Boruta

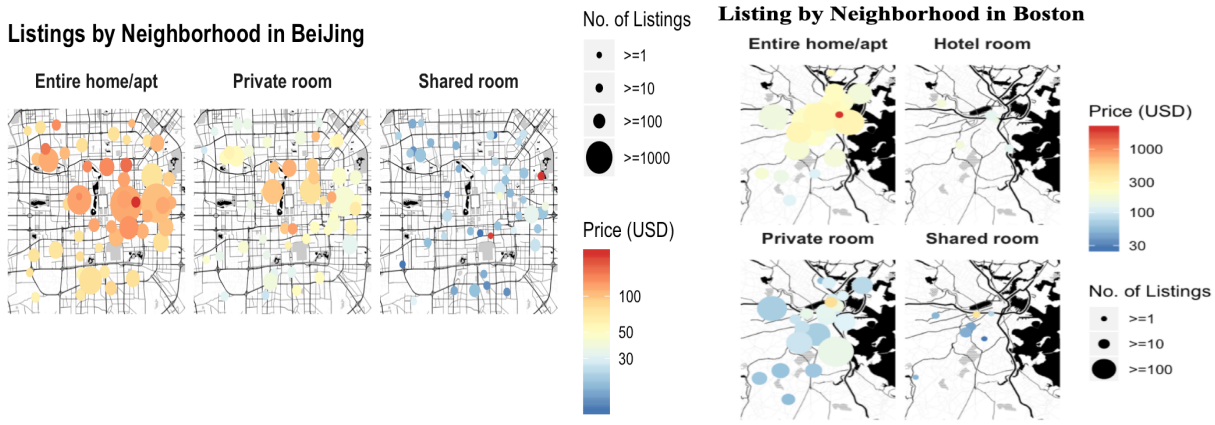
Since we have 106 variables in total, feature selection is crucial for future modelling. As we have stated above, we manually remove variables with too many missing values and variables that do not reveal any useful information. Then we perform the Boruta Algorithm to automatically select the most useful and important variables according to their importance scores. In the end, for both Beijing and Boston Listings dataset, we have 18 significant variables and 1 dependent variable "price" to be considered, along with 15,056 and 3,482 observations without missing values.

3. Exploratory Data Analysis

a. Geographic Features

Though having similar latitude and season patterns, Beijing and Boston are very different in their cityscapes. While Boston is a port city whose downtown has narrow and crowded streets, Beijing is an inland city, which not only follows the checkerboard pattern but also has its unique "ring roads" (See Appendix graph I). In general, people classify the places within "Fifth-Ring Road" as the major part of Beijing city, while other places are rural areas. Thus, we keep the home stays within "Fifth-Ring Road" according to latitudes and longitudes, so that we can directly compare the metropolitan areas of Beijing vs. Boston without involving outliers from rural areas.

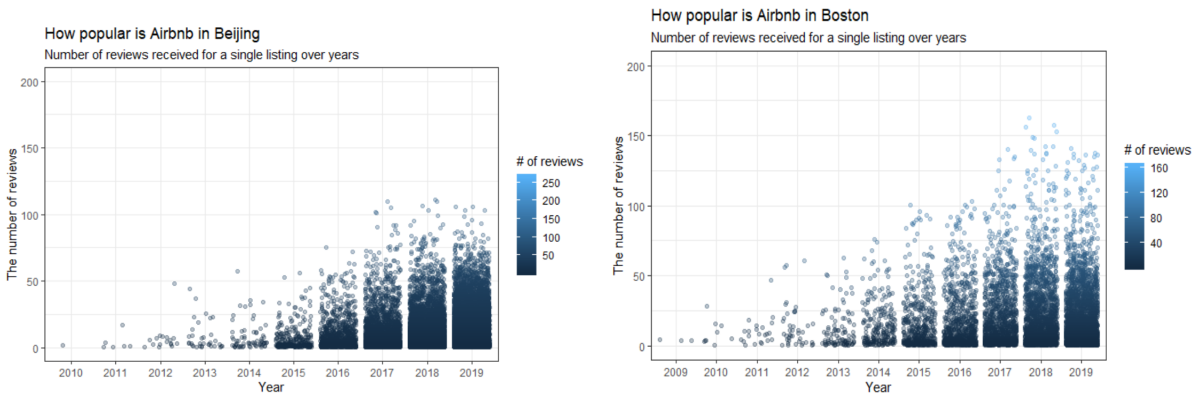
b. Price & Number of Listings by Neighborhoods



Based on the map of Airbnb prices by different room types in Boston, the most expensive type is the entire home/apartment. The price is higher as the location gets closer to downtown. The majority of listings are those type of rooms. There are just a few shared rooms and hotel rooms. Compared with Boston, the price in Beijing is distributed by its intrinsic "Ring-shape" cityscape.

The downtown of Beijing is in the center of these rings, and the price is more expensive as the ring is closer to the center (The Forbidden City & Tiananmen Square). The most expensive type is the entire home/apartment which also has the most number of listings. There are the least number of shared rooms with low prices but there are two red points on the shared room map, which means even if they are shared rooms, the good location makes them unusually expensive.

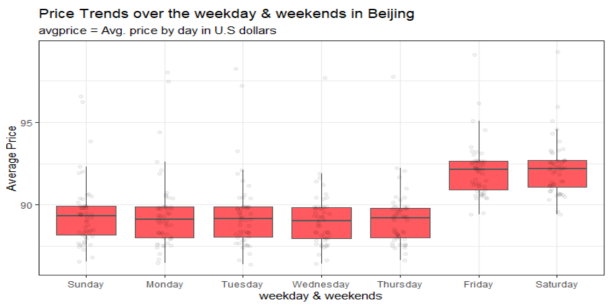
c. Popularity of Airbnb Across Years

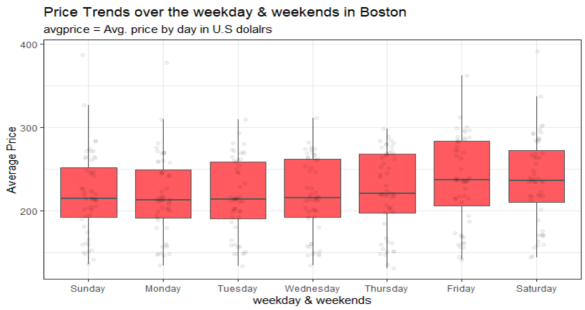


Then, we investigated the pricing trends of the Airbnb marketplace in Boston and Beijing across the last 10 years. Overall, the number of customer reviews for each unique homestay increases on an exponential scale, and customers travelled to Boston are more likely to leave their reviews than customers who went to Beijing's Airbnb houses.

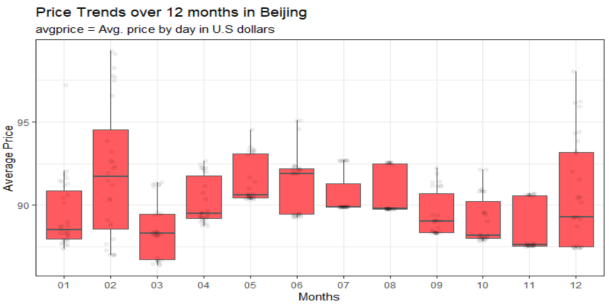
d. Price Trends by Weekday & Weekend and by Seasonality

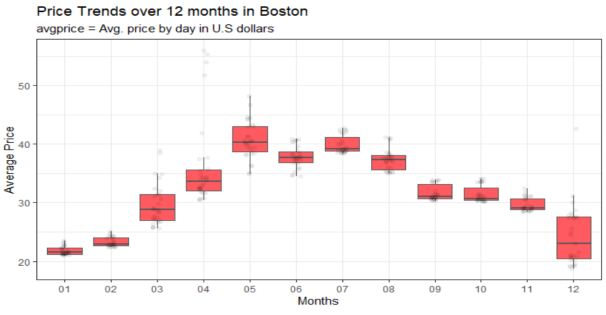
We also make a comparison of pricing patterns from weekday to the weekend for two cities. In general, prices are higher during the period from Friday to Saturday than during Sunday to Thursday. Apparently, there is a demand trend that more customers come to Airbnb on Friday and leave on Sunday. Meanwhile, the pricing fluctuations in Boston is greater than that in Beijing. In other words, from the perspective of week patterns, Beijing Airbnb's pricing is more constant than Boston's.





Thirdly, we compared the seasonality of pricing over 12 months in 2019 for Boston and Beijing. Similarly, the pricing trend in Beijing is more constant than that in Boston. For Boston, summertime (from May to August) has the highest pricing, and in winter the prices are getting down because of the coldness and frequent snow in Boston. Specifically, the deviation of prices in Boston's December is unusually high. The truth is that since Christmas holidays are in December, though the weather is not favorable for tourists, there are still many people coming to Boston to visit their family and friends, or taking 1-2 days rest for skating. The increasing demand during Christmas and winter holidays elevates prices for some homestays, which results in high deviations for the overall listings.





Beijing has a similar seasonal pattern of pricing, followed by a peak in summer and the lowest price in winter. Not surprisingly, like Boston, Beijing's deviations of pricing in January and February, which correspond to New Year and Chinese Lunar New Year, are higher than other months in winter. Needless to say, New Year and Chinese Lunar New Year are the top 2 important festivals and holidays in China, with an unusual increasing demand from customers.

4. Methodology

a. Supervised Machine Learning Model

This study aims to build supervised machine learning models that estimate the influence of geography, date, locations, and other relevant factors on Airbnb house prices. By establishing different models, this study hopes to choose the model with the smallest losses for the estimation of house prices. Specifically, for every modelling method, we generate two models for Beijing and Boston respectively. Since we have numeric outputs and inputs, we adopt Linear regression, Linear regression with Lasso method, and regression trees methods including Decision Tree, Random Forest, and Boosting. In the end, we expect to test multiple models and find the best.
At the operational level, this study will use Beijing and Boston Airbnb house price data in the year 2019. The study will separate the training set and validation set by 75% and 25% to make sure the models are trained perfectly. We will also evaluate the accuracy of the model by using MSE and R-square values to prevent the occurrence of overfitting. At the same time, in order to ensure the accuracy of the training results, we will regard the highest and lowest 5% price as outliers, and replace them by the 5% and 95% quantiles specifically for models that are sensitive to outliers, including Linear regression, Lasso regression, and two boosting models.

b. Sentimental Analysis

Besides the supervised machine learning that finds what factors influence prices most, we also explore the customer expectations and how well their expectations were met. Our research thus conducts a sentiment analysis on 2019 Airbnb consumer reviews in Beijing and Boston, so as to get ten positive and negative factors that affect consumer satisfaction. Specifically, we mainly utilize the Lemmatization method to handle the text analysis. Also, since the members of this group have the ability to read and write in Chinese and English, we can guarantee that this research can be carried out smoothly, and reviews with critical findings of the sentiment analysis will be presented in both Chinese and English.
Though without missing values, the review text data requires complicated data handling, since it contains both Chinese and English comments. We utilize the “cld3” package to identify the language categories of comments and then analyze them separately. Besides the commonly used “tidy” package for English comments, we found the “jiebaR” package to tokenize Chinese comments. Then we use “Baidu”, “cn”, “hit” and “SCU” Chinese stop words dictionaries to remove the stop words. Last, by AFINN polarity dictionary for English, we classify the polarity of words and combine the results back to each comment which can help us to identify whether a single comment is negative or positive.

5. Results

a. Linear & Lasso Regression

Linear Regression

We tried to build a multivariate linear model to predict the prices of Airbnb houses in Beijing and Boston, and observe which variables are significantly related to the price changing. At first, we did not deal with the outliers of the data, which caused the linear regression model to fit poorly. However, after replacing the outliers with 95% and 5% quantiles, the performance of the model has dramatically increased. The R-square value of the Beijing price model has reached 38.86%, and the R-square value of the Boston model has reached 44.18%.

Given the model, we found that the count of host listings, the number of accommodations, bedrooms, the cancellation policy, and the location of Beacon Hill, South Boston and West End are positively related to Boston hosting price in 95% confidence interval. However, the number of beds and the location of Mattapan have a negative relationship with the price. In Beijing, the count of host listings, accommodates, the number of bedrooms, bathrooms, the availability in 30 and 365 days, and the availability of TV have a positive relationship with the hosting price in 95% confidence interval. The number of beds, the number of reviews, and the type of rooms have a negative correlation with the hosting price in the 95% confidence interval.

Lasso Regression

We built lasso regression models to make the linear model much fitter the data and expect to get a lower test MSE. After handling the outlier with the same method, we got the test loss (MSE) of the Boston price lasso model 5628 and Beijing 1240, which are all larger than the test MSE of linear model. As we set the same seed and use the same datasets, this may be caused by the model itself, which indicates that the lasso model does not quite fit this problem.

b. Regression Trees

Before building up the trees, we substitute the outliers. But we didn't clean the outliers for the Decision tree and Random Forest since these two models are not that sensitive to outliers.

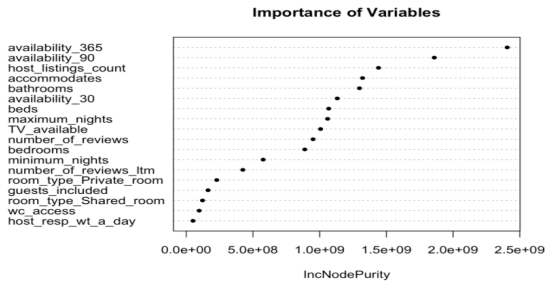
Decision Tree

We try the decision tree first since the decision tree is the simplest one and easy to interpret. After setting up the decision tree model, we prune the tree to avoid overfitting by adjusting the parameter "cp" (complexity parameter), which controls the size of the decision tree. We optimize tree size by choosing the cp value with the least "xerror" (cross-validated error).

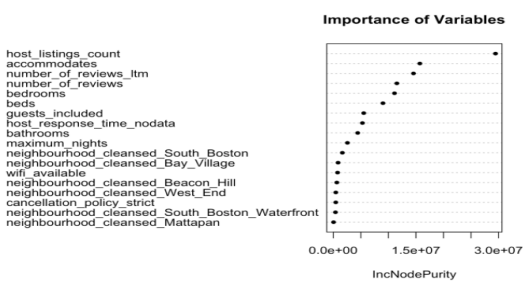
Random Forest Tree

To further optimize our decision tree, we apply the random forest which randomly creates and merges multiple trees into one "forest". In addition, we adjust the appropriate number of trees by looking at the mean square error over 0-500 trees. Then we ranked the variables by their importance as shown in the below graphs.

For Beijing



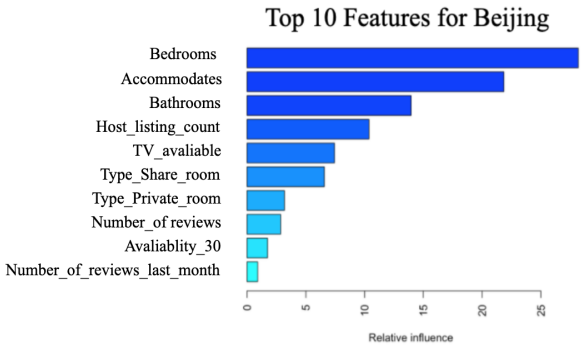
For Boston

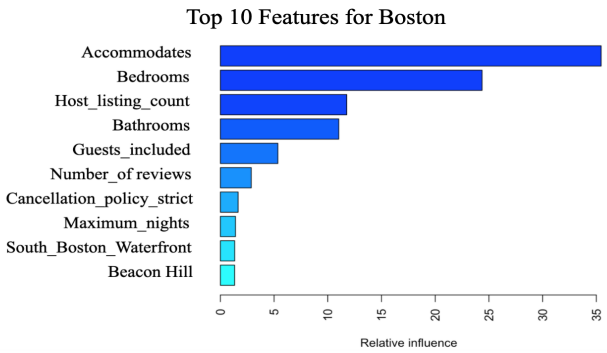


The plots show that the most important variables for Beijing are "availability in 3 months to 1 year", the number of listings a host has, accommodates and the number of bedrooms; while for Boston are the number of listings that a host has, accommodations, the number of reviews and bedrooms. Our modelling indicates that customers in Beijing check long-term availability first, which means they schedule earlier, and prefer long-term staying. For example, it is prevalent that non-local students in Beijing book Airbnb for 1-3 months since they have interns during school breaks when the universities do not provide housing. Meanwhile, visitors in Boston care more about the host response time, whereas Beijing customers do not evaluate host response heavily.

Gradient Boosting Tree

Gradient Boosting Tree also contains many decision trees, but it uses boosting to combine the training processes. We handle the outliers in data, and checked the top 10 features for two cities:





Graphs show that accommodations, the number of bedrooms, and the number of host listings are important for two cities. Still, overall the patterns are similar to results from Random Forest. Meanwhile, modelling in Boston shows the location ("South Boston Waterfront" and "Beacon Hill") does matter. While in Beijing we did not find any factors related to location. Combined with our acknowledgements to Beijing, instead of having one downtown as Boston, due to its special cityscape in Fifth-Ring shape, every district in Beijing has its own downtown, so that the influence of location to prices is not as significant as that for Boston's.

XGBoost (eXtreme Gradient Boosting)

Last, we handle outliers and utilize the XGBoost method. We used to consider the cross-validation methods with 5 or 10 fold. However, both results increase the computing cost but do not lower MSE as expected. Therefore we did not utilize cross validation here. Then we utilize both "XGBoost" and "caret" packages to construct grid spaces to search for the best hyper parameters for XGBoost modeling. Controlling 1000 rounds, the best hyper parameters for Boston Listing are 0.05 eta learning rate with a maximum tree depth of 6, using 50% features and 90% samples per tree. Beijing's are 0.05 eta learning rate with a maximum tree depth of 6, using 70% features and 50% samples per tree. As a result, Boston's XGboost MSE for the training set is 572.13, and 3367.44 for the validation set. Beijing's performance is not as good as Boston, with MSE 647.22 for the training set and 6518.66 for the validation set. Since the importance scores are the same as Gradient Boosting, see Appendix graph Ⅱ Ⅲ.

c . Text and Sentiment Analysis

We construct top words of positive and negative reviews in two cities (See Appendix graph Ⅳ Ⅴ Ⅵ). Negative words in Beijing and Boston can be concluded as "uncomfortable": "crowded" indicates the lack of spaces; "cold" refers to temperature issues; "dusty", "dirty", "trash", "dark" and "smell" unveil the poor conditions. Many customers feel "nervous" and "unsafe", which raises a red flag for future tenants regarding safety issues. Specifically, both cities have "lost", "expensive" and "tricky" that indicate the theft, the poor pricing and the unreliability of hosts. Similarly, top words for positive reviews of both cities can be described as "comfortable". The "helpful", "friendly" and "responsive" hosts give the customers the most decisive impressions, along with the clean, spacious and cozy environment. Words "recommend" and "convenient" attract future customers.

6. Discussion

a. Pricing Models Evaluation

In order to compare the performance of five main pricing models, we construct two tables.

Beijing	Linear	Decision Tree	Random Forest	Gradient Boosting	XGBoost
MSE	Train 1,244 Test 1,221	Train 645,739 Test 1023,998	Train 225,742 Test 1002,295	Train 19,865 Test 19,018	Train 647 Test 6,518
R-Square	39.46%	1.04%	3.14%	49%	55.7%

For Beijing Airbnb listing, the table indicates the XGBoost is the best model with the smallest MSE and highest R-Square, which explains 55.7% of the price. Also, two boosting (Gradient Boosting & XGBoost) methods have the greatest predictability than other methods.

Boston	Linear	Decision Tree	Random Forest	Gradient Boosting	XGBoost
MSE	Train 5,012 Test 5,622	Train 44,362 Test 153,067	Train 21,307 Test 154,589	Train 4,051 Test 4,188	Train 572 Test 3,367
R-Square	44.18%	7.11%	6.19%	53.91%	64.8%

Similarly, XGBoost model performs best for Boston listings, and explains 64.8% of the price. In conclusion, we will adopt XGBoost with optimal hyper parameters into further utilization. Meanwhile, we will use Linear regression for interpretation since its MSE and R-square are also less than Decision Tree and Random Forest, and it is easy to translate into business language and perspectives.

b. Key Differences Between Two Cities

Throughout our analysis, it is apparent that the distinct cityscapes, local cultures and other characteristics of cities lead to the differences of marketplaces for Beijing and Boston. The pricing modelling reflects the fact that Beijing's Airbnb marketplace does not care about response rate and time from hosts as much as Boston's does. The factors related to host responses are not significant enough to be considered into Beijing's models. Moreover, we analyzed the properties of "super host" in Beijing and Boston and found that the average reply time & rate for Boston's super hosts are higher than that of Beijing's.

However, customers in Beijing do care about the attitudes and help received from hosts. For top words of Chinese comments in Beijing (Appendix graph II), some of biggest Chinese characters are "Sister " ("姐姐", female hosts), "Host"(房东), and "Kind"(好心). Customers mention hosts more frequently than mentioning housing conditions, which is different from Boston's situation. Also, it is apparent that the effects of neighborhoods on prices of Beijing listings are smaller than on prices of Boston listings. There are no significant location variables for Beijing. Meanwhile, the downtown areas in Boston do have significant and positive effects on the prices. The special "Ring-Shape" cityscape that Beijing has eliminates the importance of the center of Beijing (The Forbidden City). Instead, every district in Beijing that is distributed around First to Fifth-Ring Road has its own CBD and downtown. Also, the top words of Chinese comments have "subway"(地铁), meaning that access to public transportations in Beijing is more crucial. Last, while the availability of WIFI is positively related to Boston's housing prices, a room with a TV always has a higher price controlling other factors. It implies customers in Beijing prefer TV to WiFi. Maybe customers in Beijing do not need WIFI as much as customers in Boston do since cell phone plans in China usually contain more data at cheaper prices.

7. Criticisms and Future Directions

For Supervised Machine Learning, the best model (XGBoost) explains around 60% of variability of prices. The rest of variability might be correlated with polarities of reviews, and the images of houses, which we do not cover in this project. Also, we considered using the Convolutional Neural Network. However, according to the results from available resources, Neural Network does not perform as well as expected with its high computing cost (Lewis, 2019). Last, models in our project do not account for the outbreak of COVID-19. The main reason is that there is no service, demand and supply during the outbreak of COVID-19 in Airbnb marketplace. Airbnb in Beijing has stopped all services from February 2020 to May 2020. Even though we have the 2020's pricing data from the 2019 calendar datasets, which record hosts' quotes for the next year, the data does not make sense anymore.

For sentiment and text analysis, we will utilize "Chinese hotel reviews semantic dictionary" from Mr. Songbo Tan, which is trained specifically towards hotel reviews in China. We will then label the sentimental polarity of Chinese reviews for Beijing Airbnb's analysis and adopt the overall polarity scores of comments for each listing into the Supervised Machine Learning modeling. In the future, we hope it can improve our predictions towards prices, and help hosts to better identify their potential strength and weakness. Also, we hope that the comparisons of two marketplaces can facilitate the company's understanding towards different customers' tastes in two countries, so that the company can prompt more suitable marketing strategies.

8. Conclusion

Different from our perception, though the two cities have similar latitudes, weather patterns and city structures, Airbnb marketplaces in Beijing and Boston differ on multiple perspectives. By looking at exploratory data analysis, we obtain the differences in price trends and the geographic distributions of the homestays for both cities. By conducting supervised machine learning, we predict the prices of listings, which would assist Airbnb hosts in posting suitable prices accordingly. Meanwhile, by combining with sentiment analysis, we gain insights into the different customers' tastes in Beijing and Boston. For the host in both cities, we have the following suggestions:

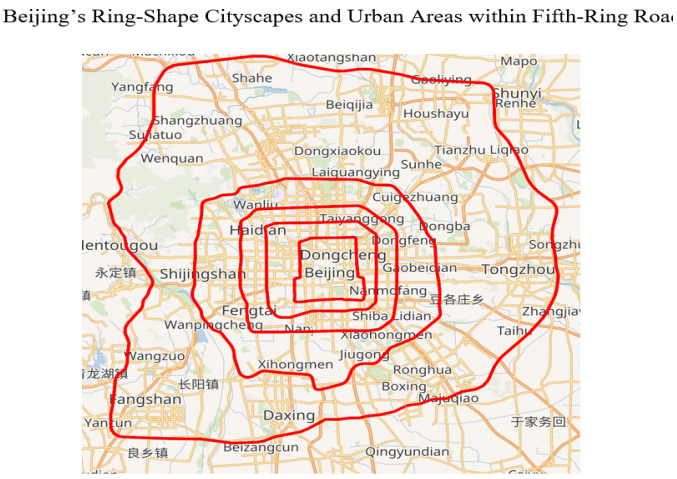
Hosts in Beijing

1. Friendly and helpful attitudes: Homestays in Beijing are more easily recommended by customers if hosts are warmhearted enough, which can even outweigh some shortcomings of the houses.
2. Setting a proper renting period: Hosts in Beijing should consider whether to attract more long-term tenants or short-term tourists because that really influences the price.

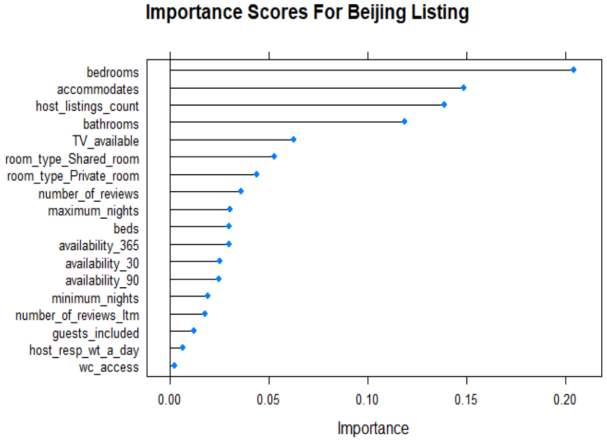
Hosts in Boston

1. Focusing on the quality of the house: In Boston, people tend to write positive feedback for homestays that are cozy and have quiet neighbors which make them feel safe.
2. Short response time: In Beijing, hosts don't need to reply to messages that fast, which is diverse from Boston. If you are a host in Boston, we suggest that you should check the email and reply to the inquiries as frequently as you can.
3. Geographic location: Locations in Boston have a strong relationship with the hosting price. For example, if your houses are in the Mattapan area, you may encounter natural competitive disadvantages compared with other locations.

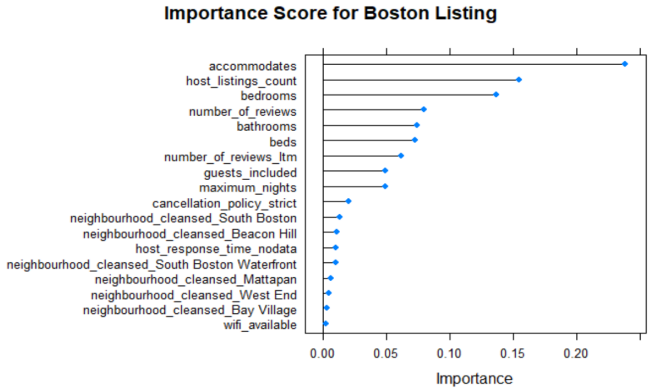
Appendix



Graph I



Graph II



Graph III

Graph IV

Graph V

Graph VI

Acknowledgements

Berhane, Fisseha. *"*Extreme Gradient Boosting with R and Python.*"Datascience-Enthusiast*, datascience-enthusiast.com/R/ML_python_R_part2.html.

Lewis, Laura. "Predicting Airbnb Prices with Machine Learning and Deep Learning." *Medium*, Towards Data Science, 22 May 2019, towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6.