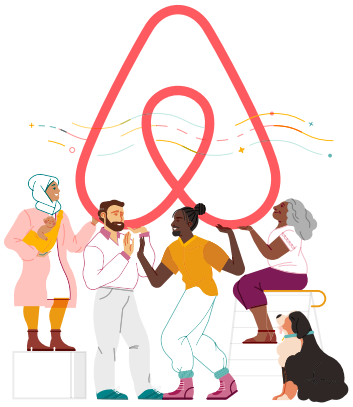
EBA 5002 Project Report

Host Assistance Platform for Airbnb Beijing



Date of Report

1st October 2020

Team Name

MyView (Team 8)

Team Members

Li Keqi(A0215494B), Ye Ruchen(A0215522R)

Xu Haotianan(A0215437H), Wen Jingtian(A0215275H), Huang Jianhao(A0215518H)

Contents:

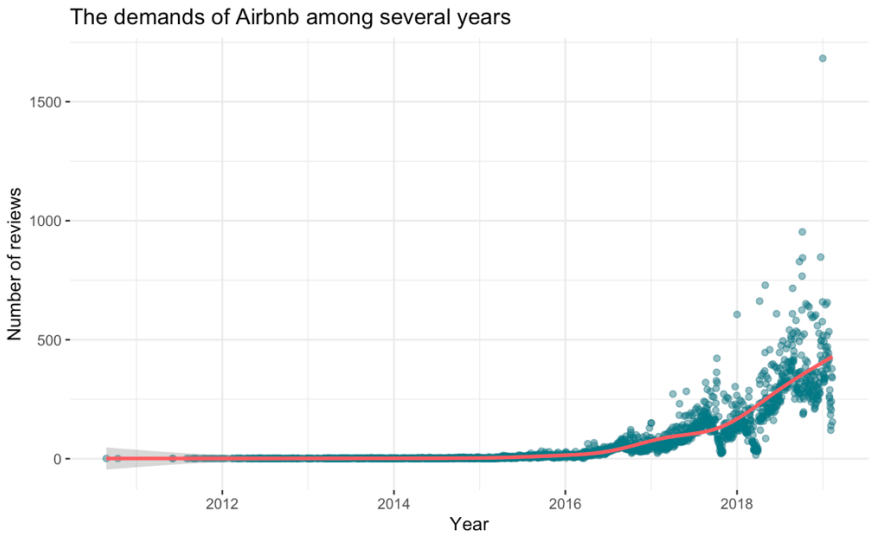
1 Project Background and Introduction

1.1 Introduction of Airbnb

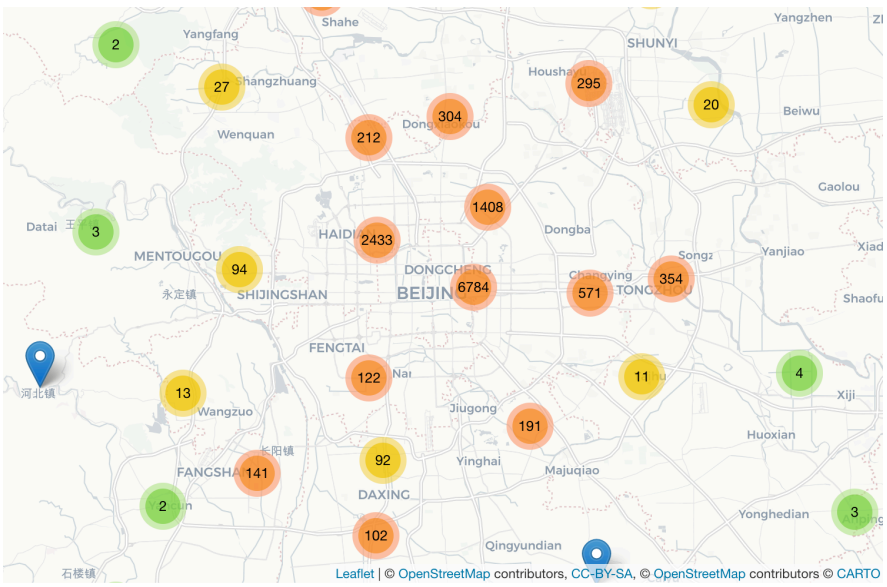
Airbnb is a website for people to rent out their accommodation, which has more than 3 million homes in 65,000 cities in 191 countries. Airbnb's profit's main source is renting commission: A transaction fee of 3% is charged to the landlord and a service fee of 6% to 12% is charged to the residents. Airbnb's business model is mainly platform-related activities. It requires companies to constantly develop and maintain their platforms, the most important of which are marketing, community operations and product development.

1.2 Problems Airbnb Has Now

Since 2010 when Airbnb started the business journey in Beijing, the increase of customers' reviews is going more and more rapidly over and over years. Airbnb becomes lots of customers' first choice when they are enjoying their vacation.



Also, we can see available Airbnb properties are distributed all around Beijing City.



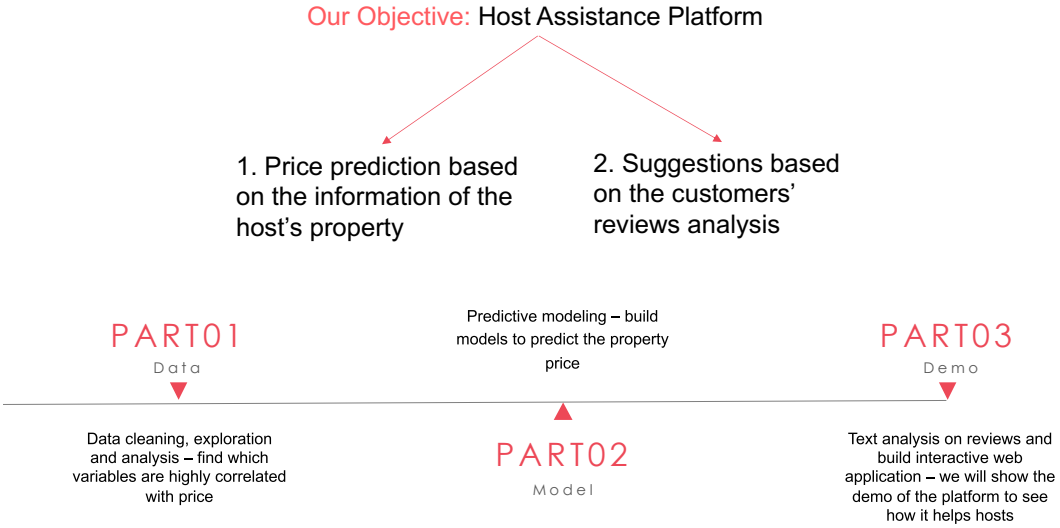
Because of the popularity of Airbnb in Beijing, we are interested in finding lots of people starting to join the group of Airbnb's landlords by all kinds of reasons. But then there comes with a problem.

Since lots of properties owners are new to the Airbnb platform, they don‘t have much experience on how to give the proper price of their houses or what the customers care when they rent a house. And this leads to the result that, some of the landlords cannot rent out their properties because of high price or don’t care on what the customers care about, also in the same time, some customers find sometime the price of the houses are strange with their experience.

And this will absolutely cause the drop of users’ experience with Airbnb and will reduce the trading volume which is import for Airbnb to make profit. So if there is a Host Assistance Platform within the system, it will help landlords to give prices of their properties properly and give suggestions to them on what should take care on the properties and achieve good occupancy rates, which is profit hosts, users and platform, also improving the user experience with Airbnb.

1.3 Project Objective

Our objective of the project is to build a host assistance platform, and it is made up with two parts. First part is the price prediction based on the information of the host’s property. And the second part is some suggestions based on the customers’ review analysis.



2 Data Exploration and Analysis

2.1 The overview of data

2.1.1 Data source

The data is sourced from the "Kaggle" website `https://www.kaggle.com/merryyundi/airbnb-beijing-20190211`, which hosts publicly available data.

2.1.2 Data content

The dataset comprises of three main tables, which are listings, reviews and calendar.

The "listings" file includes detailed listings data showing 106 atttributes for each of the listings. Some of the attributes used in the analysis are `price` (continuous), `longitude` (continuous), `latitude` (continuous), `room_type` (categorical), `is_superhost` (categorical), `neighbourhood_cleanse` (categorical), `bedrooms` (categorical) among others.

The "reviews" file contains detailed reviews given by the guests with 6 attributes. Key attributes include `date` (datetime), `listing_id` (discrete), `reviewer_id` (discrete) and `comment` (textual).

The "calender" file provides details about booking for the next year by listing. Seven attributes in total including `listing_id` (discrete), `date` (datetime), `available` (categorical) and `price` (continuous).

2.1.3 Analysis of data quality

The data quality of the data is not perfect. There are more than half missing data in columns'square_feet','weekly_price', 'monthly_price', 'security_deposit','cleaning_fee' and about 40% of the data in the feature about reviews. We had to drop this column. For other columns that are relatively complete, we need to perform a few imputations and transformations on our dataset for us to create the desired visualizations. Some of the column in file 'Listings' miss a lot of data, some of them contain outliers, and most of the columns/features we were interested in did not contain data in the required format and hence were manipulated in a way that their meanings are retained.

2.2 Data cleaning

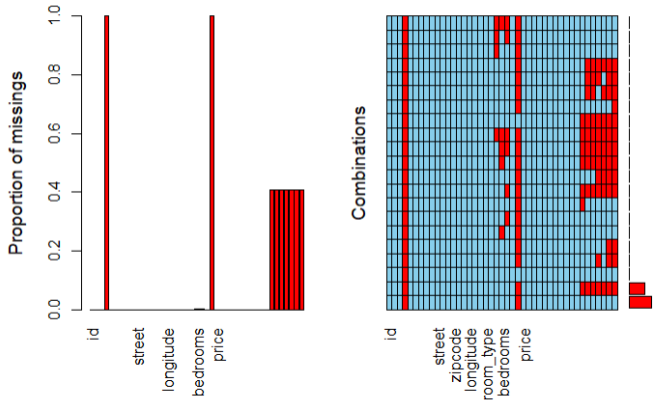
2.2.1 Select useful columns

Data in the listing files is the main data we use to build predictive model. There are 106 columns in this dataset and of course there is no way that we can put them all in the model because of the existence of some meaningless feature. So after the preview, we select 45 columns that we think might be useful to do the data cleaning and transformation.

2.2.2 Dealing with Missing Values

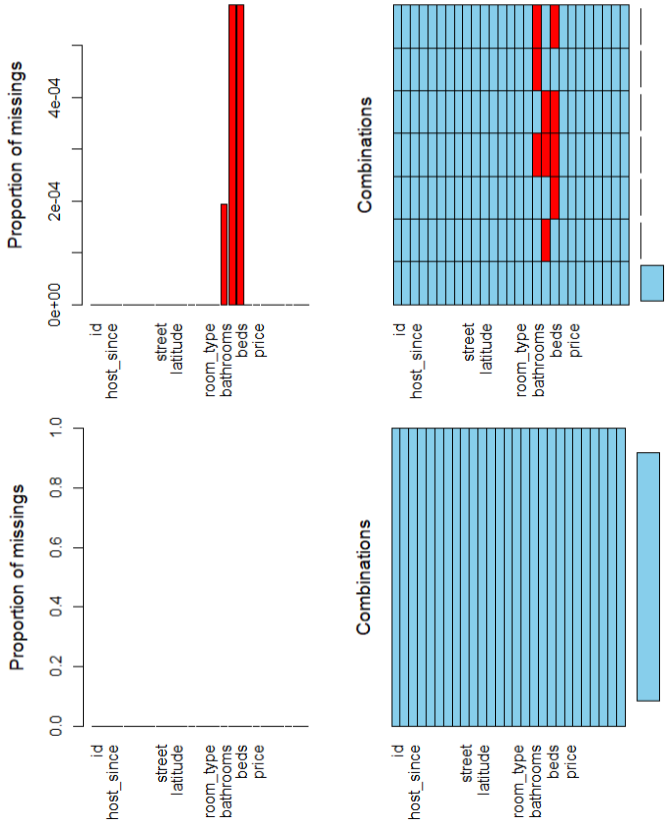
The selected data also had null values. To preserve all the information, we imputed or dropped the rows and columns containing null values while conducting exploratory analysis that made use of these features.

We construct a 'aggr' graph (we are sorry that the graph cannot display all the column names because their length of name is too long) to analysis the missing values for the variables that we would be using in our exploratory analysis.



Several observations from the graph:
1)Most of the rows have no missing values.
2)Columns 'square_feet','weekly_price','monthly_price','security_deposit','cleaning_fee' are missing more than half of the data, and columns 'review_scores_rating','review_scores_accuracy','review_scores_cleanliness','review_scores_checkin','review_scores_communication','review_scores_location','review_scores_value' are missing 40% of the data.

We can conclude from the above graph that the Airbnb dataset contains several missing value which will affect our analysis if not handle well. For datasets that containing more than 40% missing value, we choose to drop them. And for the rest columns, we use function "complete.case" to drop the row because their volume are small and will not affect the predicting accuracy very much.

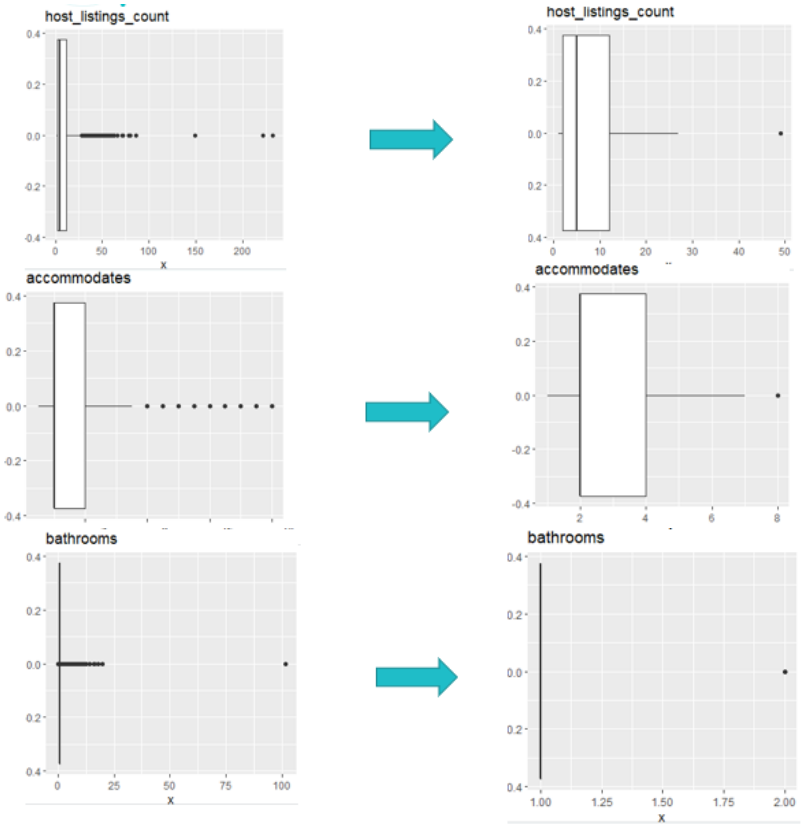


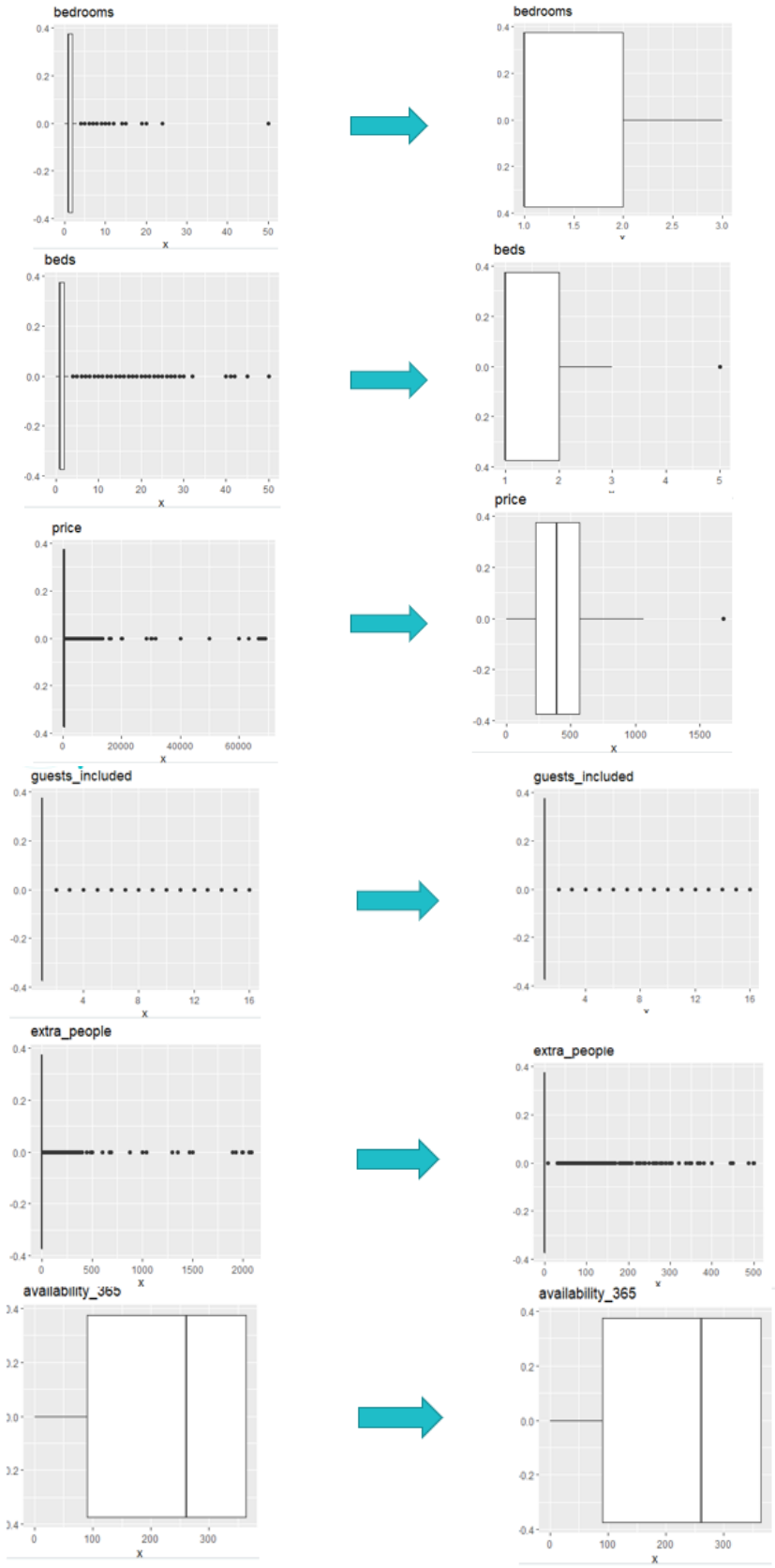
Some columns need to be deal with individually. The "date" and "host_since" column need to transformed into date format so that we can use it as time feature. The columns

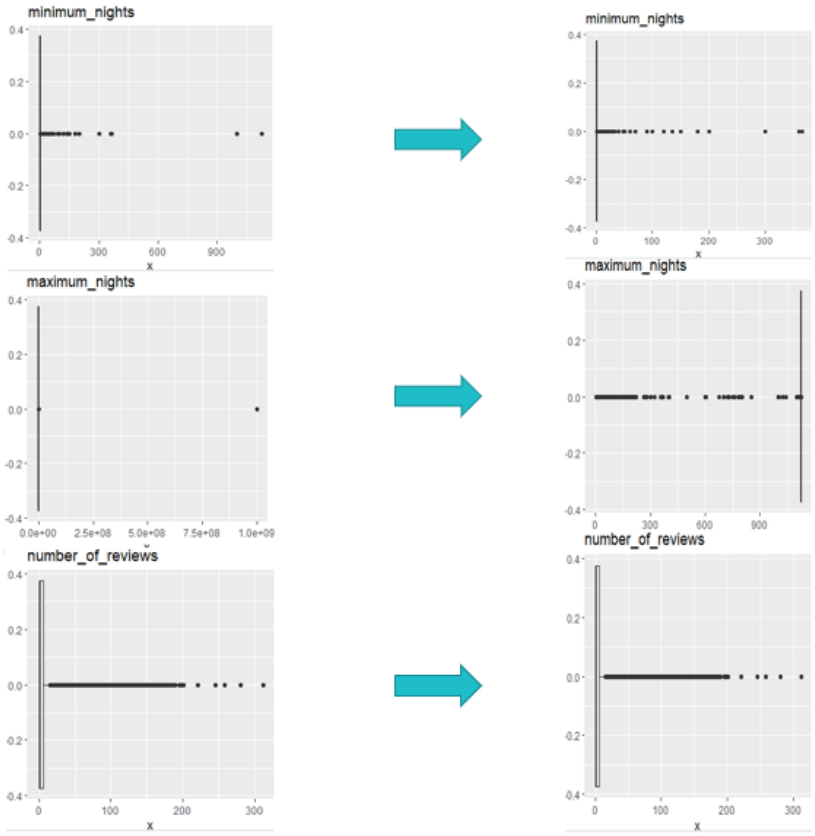
"price" and "extra_people" are in character format containing "$" and ",", so we drop these special symbols and transform them into numeric data. Column "host_listings_time" contains 15 zero values, which is not in accord with reality, so these row need to be drop.

2.2.3 Dealing with outliers

We draw a histogram and boxplot for the cleaned numeric data to view their distribution and outliers. Through these graph, we found that the columns 'host_listings_count','accommodates','bathrooms','bedrooms','beds' and 'price' are obviously positive skewed and containing many outliers. So we decided that the data that are outside the 1.5 IQR plus 75 percentile are outliers and use the 95th percentile to replace them. Other columns need to be deal with individually to simulate real situation better. For columns "extra_people", we thought that the 95th percentile is too small and cannot reflect all the situation because there are lots of room charging high cost for extra people. So we set a top price $500 and replaced the data biggest than this price. For columns "minimun_night" and "maximun_night", we thought that minimnn night will not larger than 365 and maximun will not larger than 1125, so we set them as the cap.




bedrooms
bedrooms
beds
beds
price
price
guests_included
guests_included
extra_people
extra_people
availability_365
availability_365



2.2.4 Combining small district with big district

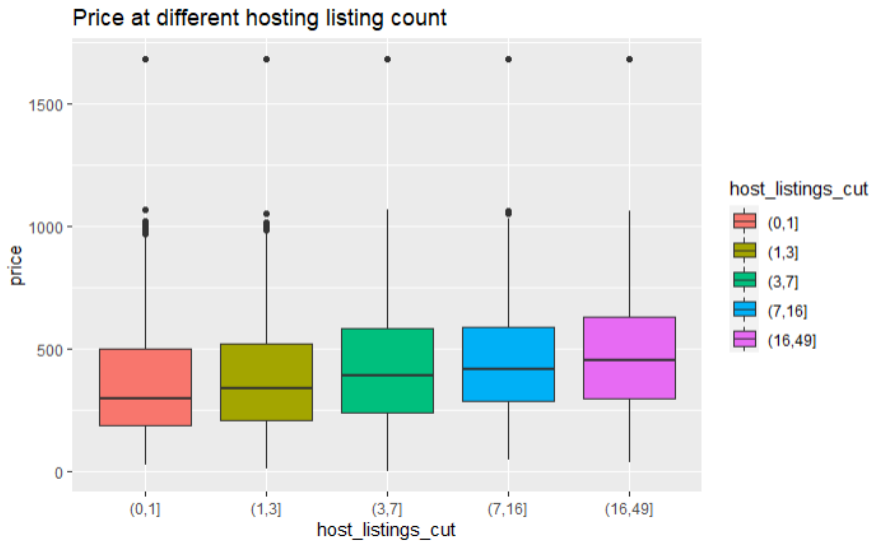
Because some district are relatively less developed by Airbnb Beijing, there are few listing located in these districts. So in order to ensure the significance of the model input variables, we decide to combine the listings of districts containing less than 500 with a closed district. Through the form below and map of Beijing, we decide to combine Shijingshan and Mentougou district with Fangshan district, and combine Pinggu district with Miyun District. These combination are based on their geographical proximity and therefore do not have a significant impact on regional and price relationships.

neighbourhood_cleansed <chr>	Freq <int>
朝阳区 / Chaoyang	10319
东城区 / Dongcheng	3175
海淀区 / Haidian	2983
西城区 / Xicheng	1565
丰台区 / Fengtai	1458
通州区 / Tongzhou	1167
昌平区 / Changping	968
大兴区 / Daxing	725
密云县 / Miyun	721
顺义区 / Shunyi	701

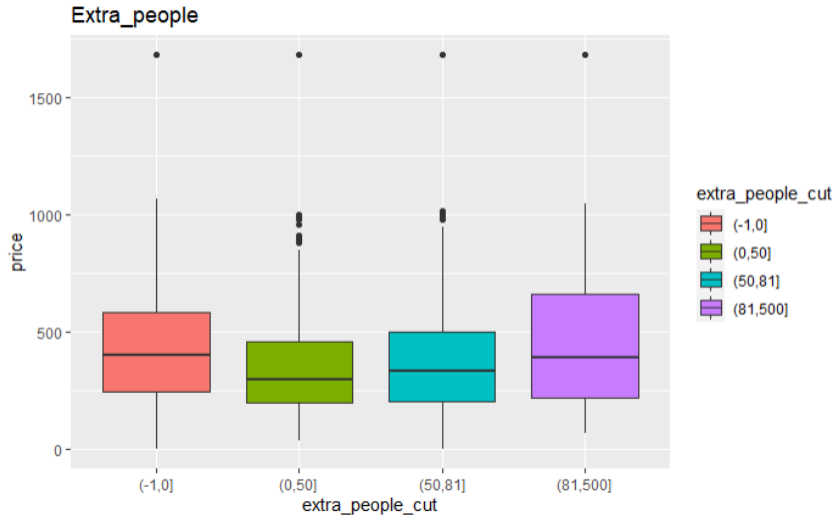
怀柔区 / Huairou	649
延庆县 / Yanqing	562
房山区 / Fangshan	463
石景山区 / Shijingshan	179
门头沟区 / Mentougou	129
平谷区 / Pinggu	116

3.3 Feature exploration

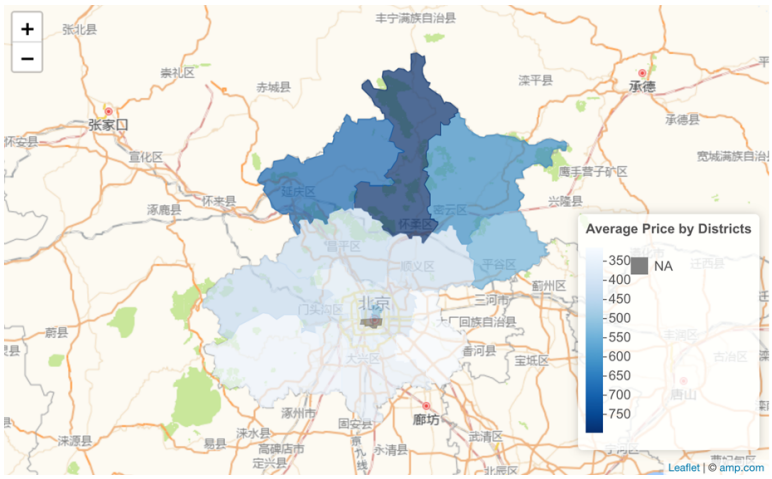
We use visualisation method to discover relation between features and price and put the chosen variables in model.



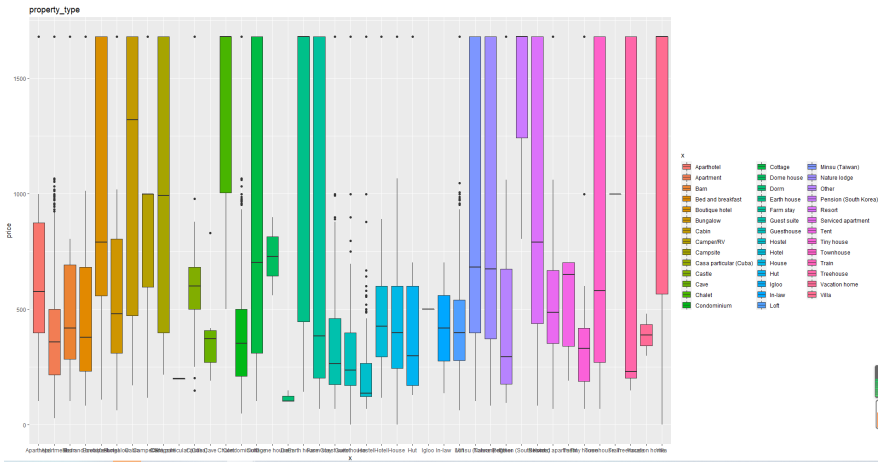
The boxplot indicates that the more property listings owners have, the higher they tend to set the price. According to our analysis last project, owner who have more houses are more likely to be a professional agent or even company. So it's likely that the professional agent tend to set higher listings price to ensure their profits.



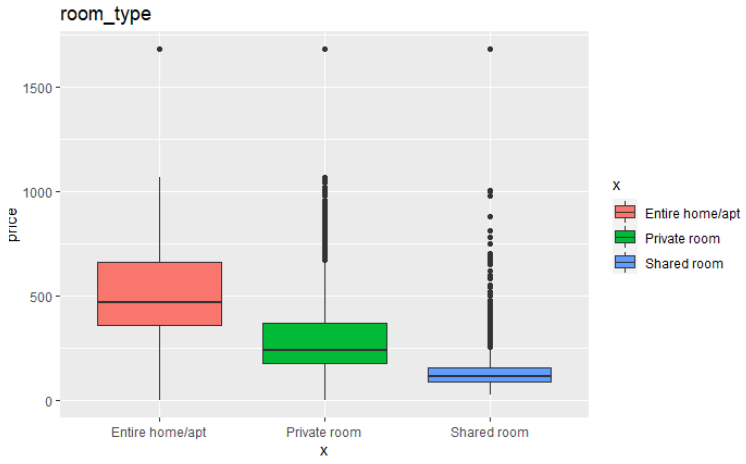
Some property will set an occupancy limit. And sometimes every people that cross that limit need to pay extra fee. For those don't charge extra fee, we think that may be the listing price is relatively high. And for those charging extra fee, higher extra fee probably indicate higher price.

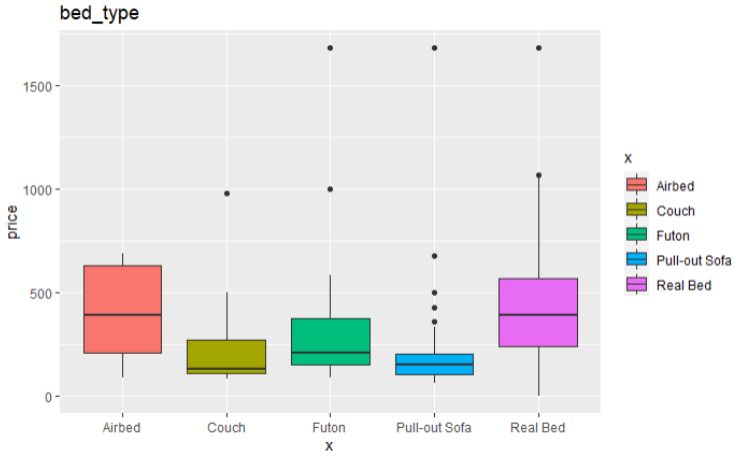


According to this map, we can clearly see that the overall price varies greatly from region to region, which indicates that region is a good indicator of price.

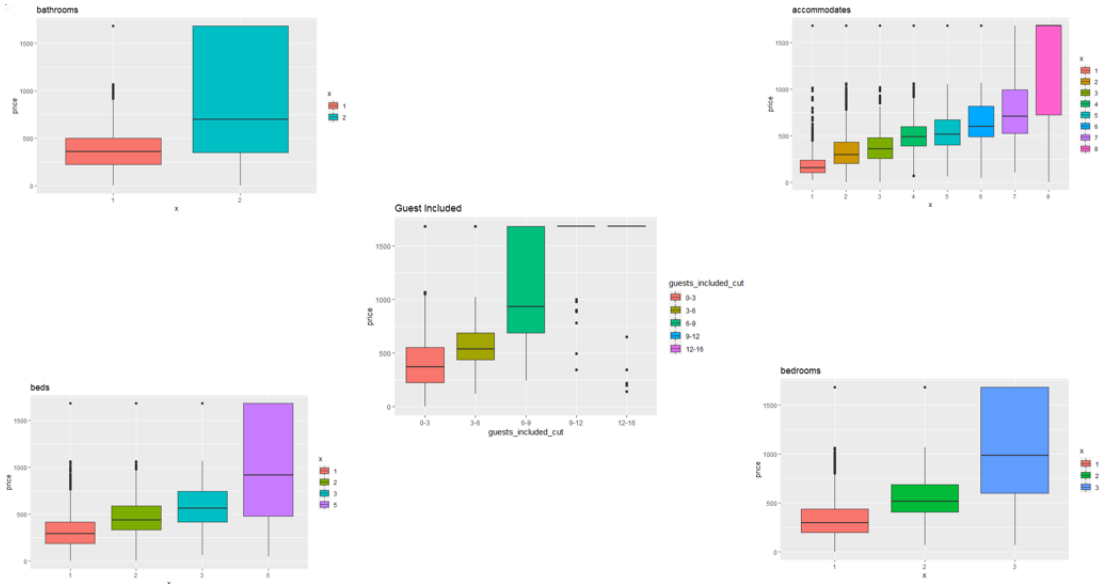


According to this graph, we can clearly see that the overall price varies greatly between different property type, which indicates that this variables is also a good indicator of price.





Through analysis, we realized that room type and bed type may also be correlated with rent price. According to common sense, we can know that entire home and private room are definitely more expensive, because customers can enjoy more space and better privacy. And more advanced rooms are more likely to equipped better beds.



Here are some variables that can reflect the size of the room laterally. We know from that the bigger the house is, the more expensive it is, and this is also well reflected in the boxplot.

Through exploration, we find 11 features relative to price and put them into the model. In the next session, we are going to show the construction and performance of the model.

3 Price Prediction Model

3.1 Linear Regression Model

In this section, we build our prediction models with the data we have cleaned and explored. Firstly, we select the highly correlated variables and put their names into an array. Then we use the function cor and pairs to ensure the relationship between them. Also, in order to test the model, we split the data into train data set and test data set, with the ratio of 0.85.

	price	host_listings_count	guests_included	extra_people	accommodates	bathrooms
price	1.000000000	0.13591553	0.25434923	0.002562351	0.65282219	0.33301050
host_listings_count	0.135915525	1.00000000	0.03260546	-0.036772053	0.09842096	-0.01272256
guests_included	0.254349228	0.03260546	1.00000000	0.198300085	0.35746490	0.13851545
extra_people	0.002562351	-0.03677205	0.19830008	1.000000000	0.02232499	0.08883137
accommodates	0.652822188	0.09842096	0.35746490	0.022324987	1.00000000	0.37814053
bathrooms	0.333010502	-0.01272256	0.13851545	0.088831372	0.37814053	1.00000000
bedrooms	0.612044944	0.08081379	0.31881991	-0.003199113	0.80496167	0.39199258
beds	0.522722846	0.06419433	0.31349620	0.044784098	0.84597424	0.37222742

	bedrooms	beds
price	0.612044944	0.52272285
host_listings_count	0.080813793	0.06419433
guests_included	0.318819906	0.31349620
extra_people	-0.003199113	0.04478410
accommodates	0.804961674	0.84597424
bathrooms	0.391992581	0.37222742
bedrooms	1.000000000	0.72864300
beds	0.728643001	1.00000000

According to the form, variables accommodates, bedrooms and beds are highly relative. This might cause a multicollinearity situation. After finishing all the preparation works, we build the linear regression model first.

We use backward method to let the system automatically select the best model. We start the model with all independent variables. After training, the system will store the best linear regression model in the variable lr_model. We explore the model result by summary, plot and analyze its residuals.

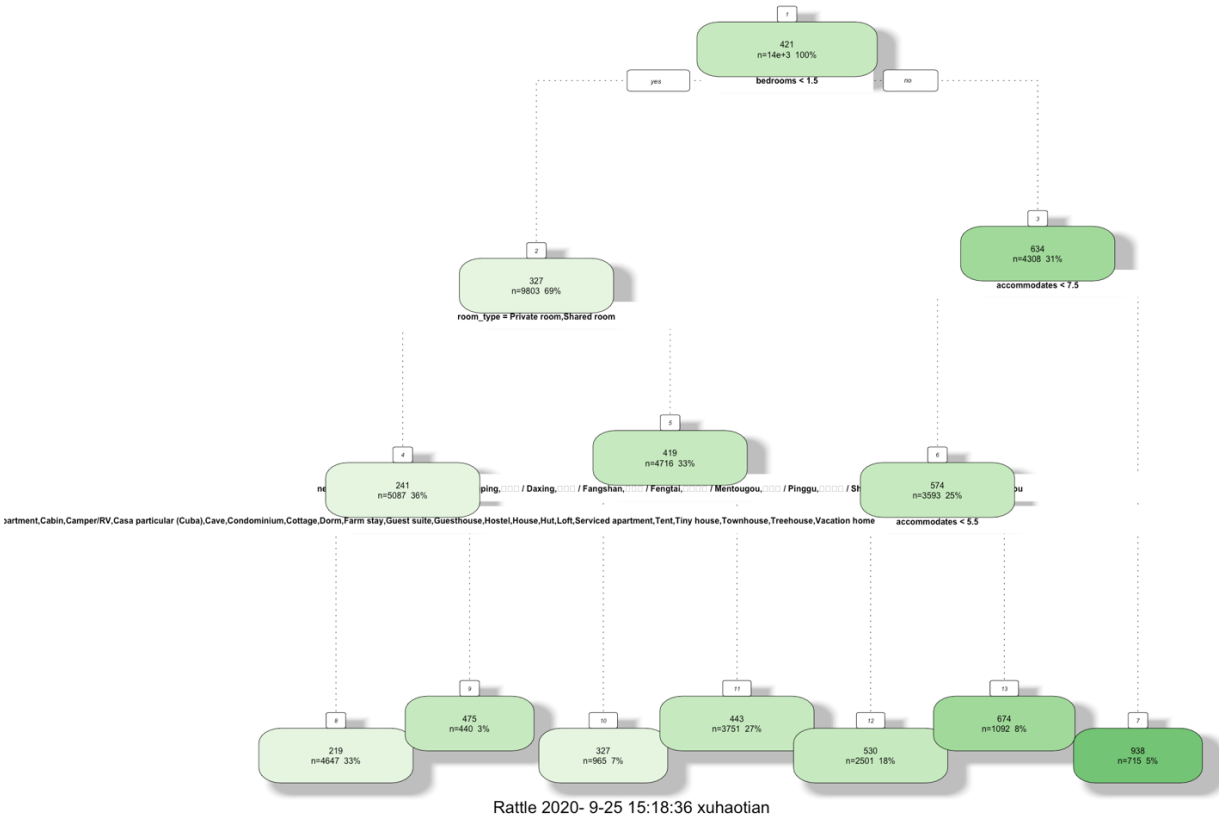
This figure shows the part of the result of the lr model. We can see in this model, the price is dependent on several variables including host_listings_count, extra_people, neighbourhood_cleansed, property_type, accomodates, bathrooms and bedrooms. Also we find that most of the variables are highly significant in the model, which have P-value less than 0.001. Take one for example, when the host have one more property, his/her price of the property is most likely to raise a little.

```
lm(formula = price ~ host_listings_count + guests_included +
    extra_people + neighbourhood_cleansed + property_type + room_type +
    accommodates + bathrooms + bedrooms, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-1344.97  -138.90   -35.40    73.49  1738.35

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             251.74000   41.72862   6.033 1.64e-09 ***
host_listings_count                       1.84717    0.13262  13.928  < 2e-16 ***
guests_included                          -7.21088    1.80278  -4.000 6.36e-05 ***
extra_people                              0.21130    0.04042   5.228 1.73e-07 ***
neighbourhood_cleansed朝阳区 / Chaoyang     8.43812   10.45961   0.807 0.419828
neighbourhood_cleansed大兴区 / Daxing     -86.74696   14.94634  -5.804 6.56e-09 ***
neighbourhood_cleansed东城区 / Dongcheng   92.69333   11.39248   8.136 4.28e-16 ***
neighbourhood_cleansed房山区 / Fangshan   -49.24038   14.65966  -3.359 0.000784 ***
neighbourhood_cleansed丰台区 / Fengtai    -85.00929   12.77608  -6.654 2.92e-11 ***
neighbourhood_cleansed海淀区 / Haidian    -16.39681   11.45695  -1.431 0.152396
neighbourhood_cleansed怀柔区 / Huairou    257.38754   15.74502  16.347  < 2e-16 ***
neighbourhood_cleansed密云县 / Miyun       73.42096   14.45553   5.079 3.82e-07 ***
neighbourhood_cleansed顺义区 / Shunyi     -69.93483   15.08554  -4.636 3.57e-06 ***
neighbourhood_cleansed通州区 / Tongzhou  -156.80776   13.40048 -11.702  < 2e-16 ***
neighbourhood_cleansed西城区 / Xicheng     30.85977   12.54833   2.459 0.013929 *
neighbourhood_cleansed延庆县 / Yanqing    113.56446   16.84618   6.741 1.61e-11 ***
property_typeApartment                  -313.42154   39.65201  -7.904 2.81e-15 ***
property_typeBarn                       -294.31134  114.40803  -2.572 0.010104 *
property_typeBed and breakfast          -198.16331   44.14938  -4.488 7.21e-06 ***
property_typeMinsu (Taiwan)             -125.65631   72.42839  -1.735 0.082771 .
property_typeNature lodge                -36.08406   48.88111  -0.738 0.460400
property_typeOther                      -219.78246   48.96378  -4.489 7.20e-06 ***
property_typePension (South Korea)        -1.76875  287.07254  -0.006 0.995084
property_typeResort                       80.07475   58.39724   1.371 0.170323
property_typeServiced apartment         -190.68114   40.49601  -4.709 2.51e-06 ***
property_typeTent                       -241.88585   98.28023  -2.461 0.013855 *
property_typeTiny house                 -292.78650   55.32878  -5.292 1.22e-07 ***
property_typeTownhouse                  -151.77973   42.69771  -3.555 0.000379 ***
property_typeTrain                        99.29817  286.95444   0.346 0.729315
property_typeTreehouse                   -71.07754  147.51716  -0.482 0.629934
property_typeVacation home              -212.32974  204.77416  -1.037 0.299795
property_typeVilla                        51.56472   42.00854   1.227 0.219654
room_typePrivate room                   -134.58321    4.73107 -28.447  < 2e-16 ***
room_typeShared room                    -264.40773    8.73390 -30.274  < 2e-16 ***
accommodates                              42.86129    1.76492  24.285  < 2e-16 ***
bathrooms                                220.27070    6.36275  34.619  < 2e-16 ***
bedrooms                                 121.06896    4.86770  24.872  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 284 on 23233 degrees of freedom
Multiple R-squared:  0.5103,    Adjusted R-squared:  0.509
F-statistic: 403.4 on 60 and 23233 DF,  p-value: < 2.2e-16
```

The model’s adjusted R square is 0.61, which is not very high because it is less than 0.75. We also plot its qq-plot but find not all the point just right on the line. But when we try to do more with the data like more transforming, the model result is getting worse. And we find some modeling in Kaggle which’s building models with Airbnb in New York, all can’t get a very satisfactory result. So we still think this model has some price prediction suggestion to the new hosts.

3.2 Decision Tree Model

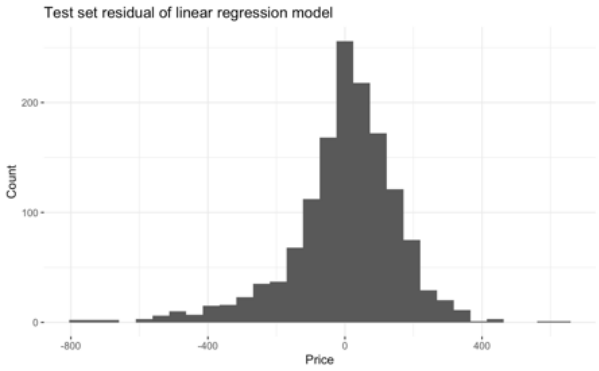
Next, we tried to build a decision tree model with the data. Because the output variable price is continuous. So we set the method of training the tree as ANOVA.After training the tree, we also visualize the model and use the test set to find out how well it works.

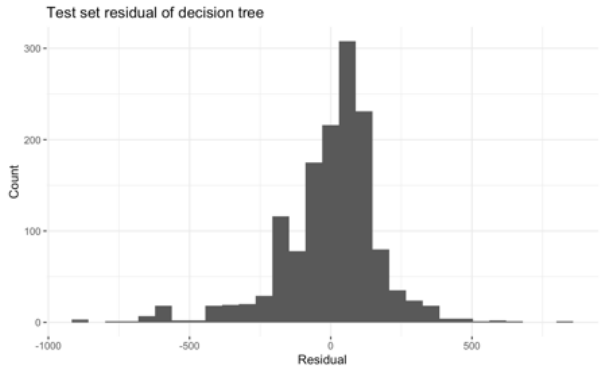


3.3 Compare

After getting the model, we use them to predict on the test set, and try to find how well it can predict the price.

From the two plots shown in the slides, we know that both of linear regression model and decision tree model works well with the data, and most of the predicted prices are within 100 dollars compared to the real prices of the properties, though there are still some have large difference compared to the real ones.





After doing all of this, we will put the price prediction model into the host assistance platform, to help the new hosts to know what price ranges are suitable for their properties.

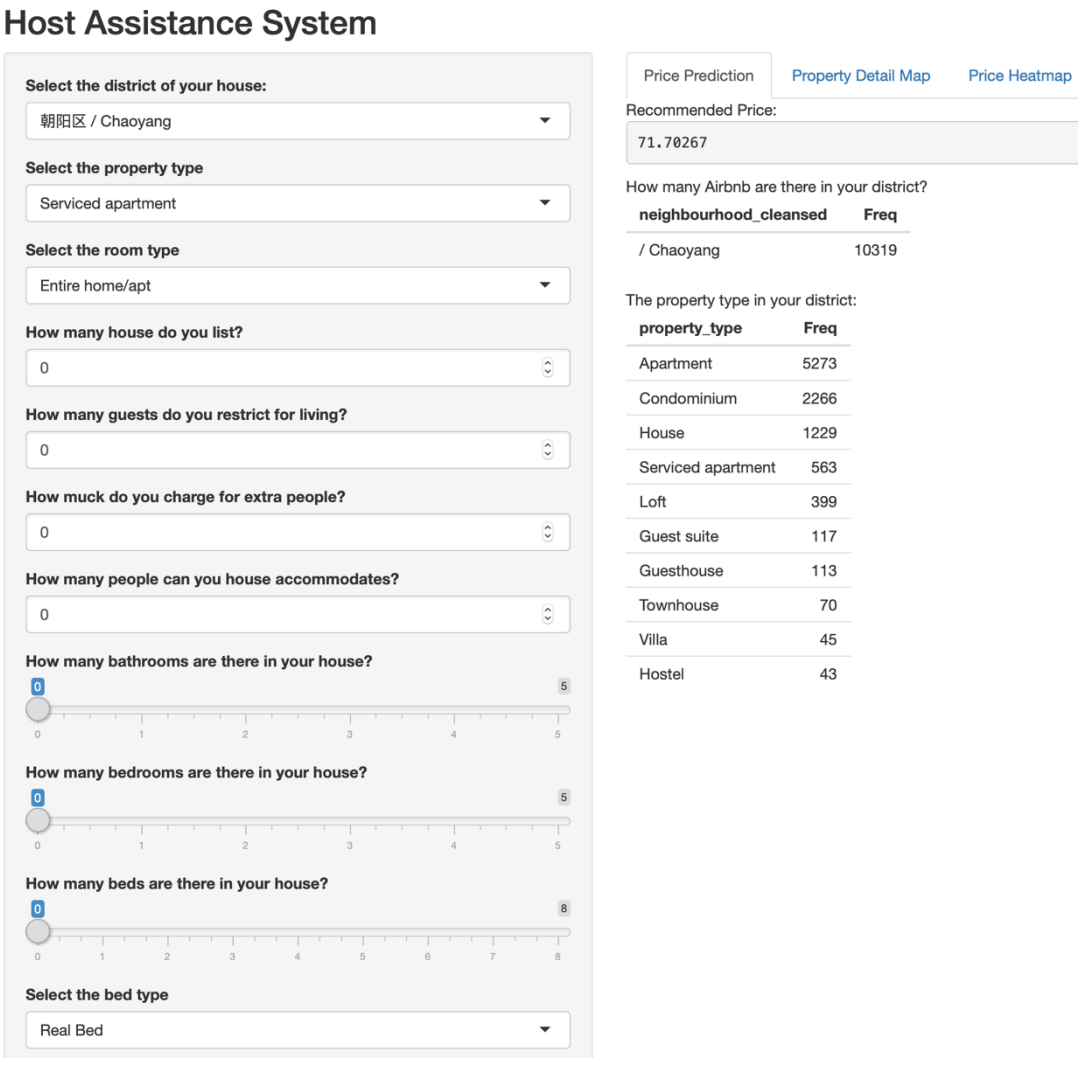
3.4 Host Assistance Platform Demo

We have made an interactive component called Host Assistance System as below, which display the recommended price based on your input information. Here are the parameters that the user can control and filter on:

a) Select the district of your house
b) Select the property type
c) Select the room type

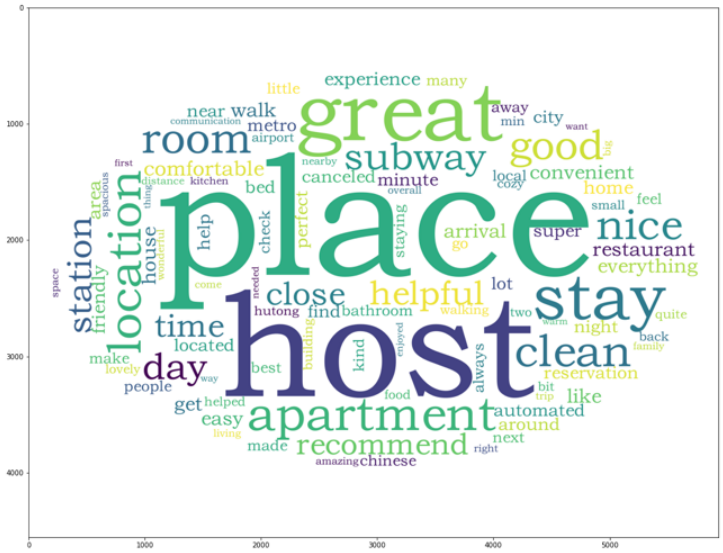
d) How many houses do you list?
e) How many guests do you restrict for living?
f) How much do you charge for extra people?
g) How many people can you house accommodates?
h) The number of the bathrooms
i) The number of the Bedrooms
j) The number of the beds
k) Select the bed type

If all the information has been correct given, the host will get a recommended price on the right top, which is relatively a rational price, to help them pricing. Besides, there are still two functions: Property Detail Map and Price Heatmap. 'Property Detail Map' and 'Price Heatmap' show the other properties and its price in the certain district, which can also be regard as the assisted tools.



4 User Review (Textual Data) Mining

The dataset provides us with a ton of data, but nothing as insightful and close to the customer as their reviews or feedback. If mined properly, they can tell us a lot about the customer mindset, their expectations and how well those were met. For the final result to make sense, the review text data requires a lot of cleaning - e.g. The words need to be stemmed, commas-full stop-percentages etc need to be removed, common English words and stop words need to be removed etc. And the reviews should be separated into different language, in our text analytics we just used comments in English and Chinese.



Wordcloud in English

Wordcloud in Chinese

4.1 Comment Analysis Using Word Cloud

Start first with looking at the dominant themes in the reviews; simply building a word cloud should solve this purpose. Word Cloud take a frequency count of the words in the corpus as input, and return a beautiful display of dominant (frequently occurring) words, with their size being proportional to their relative frequency. In this case we choose to analysis all the reviews. Although we only need the generic words here, but the future analysis on 'positive' and 'negative' reviews needs more data as we'll see in the next section.

Analysis of the word cloud shows interesting trend; For the both versions, they show almost the same results: The word 'Host' finds a lot of mention indicating the important role that hosts play in shaping the Airbnb experience. Transport options like 'subway', 'station' also find frequent mention. Airbnb are short term rentals, yet people seem to care about the 'place', 'room' and 'house' and it would better be clean.

```
INFO  [08:58:20.005] epoch 5, loss 0.0522
INFO  [08:58:23.583] epoch 6, loss 0.0478
INFO  [08:58:27.130] epoch 7, loss 0.0445
INFO  [08:58:30.749] epoch 8, loss 0.0418
INFO  [08:58:34.379] epoch 9, loss 0.0396
INFO  [08:58:38.000] epoch 10, loss 0.0378
INFO  [08:58:41.631] epoch 11, loss 0.0362
INFO  [08:58:45.238] epoch 12, loss 0.0348
INFO  [08:58:48.881] epoch 13, loss 0.0336
INFO  [08:58:52.661] epoch 14, loss 0.0325
INFO  [08:58:56.408] epoch 15, loss 0.0315
INFO  [08:59:00.241] epoch 16, loss 0.0307
```

Building Word Vectors from Reviews

4.2 Building Word Vectors from Reviews

The word cloud generated previously does a good job of finding what customers are looking for, but it is very generic. Wouldn't it be great if could find what people think of the room sizes? How about seeing what makes customers 'comfortable'?

Word vectors simply place any given word in an n-dimensional space; and the proximity of any two words in this vector space is proportional to their 'similarity'. We use the review data to construct such a vector space to build a word cloud of similar words, derive interesting insights. The first word cloud is for the word 'comfortable', we hope to see things that led to a positive experience. Prominently featured are words like 'cosy', 'clean', 'spacious' indicating the importance of the environment, location.

Similarly, querying by the keyword 'uncomfortable', which are usually those that occur in conjunction with it frequently, i.e. reasons for the discomfort. The word cloud shows just that - notice words like 'crowded' indicating that lack of space is one of the most common complaints.

'Inconvenient' and 'hard' are some of the issues about location and transport. And 'Expensive', which means the price is relatively too high, will prompt people to write negative feedback. As we can see, word vectors add so much more meaning to our analysis.

Wordcloud for comfortable

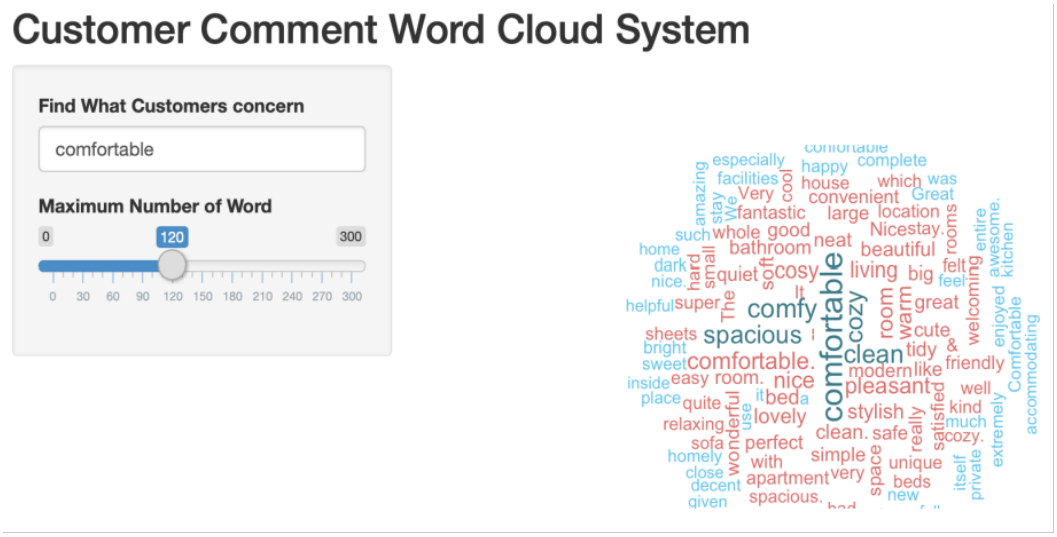
Wordcloud for uncomfortable

4.3 RShiny Word Cloud Generator by Query String

Now we have seen how insightful word vectors can be, it would make sense to develop a tool that allows you to query any word and generate its corresponding word cloud. Words like ‘happy’ - ‘sad’ can again show us the things that matter most people. However, more

specific queries can also give us a wealth of insights. Querying by the term 'bathroom' may show that people seek and talk about: 'towels', 'shower'.

The Customer Comment Word Cloud System serves as an extension to our previous Word-Vector analysis of the customer reviews. Once the word vector has been built, then anyone can query the vector to find similar words and built a wordcloud. User can enter any valid query word and set the 'max words' in the word cloud. A valid query word would be one that is present in the corpus, else it wouldn't be part of the word vector. It is easy to think of such words. But this is an exploratory function actually, that means not all the given words will get a useful information at all.



4.4 Suggestion for the hosts

Based on the results yet, we would like to give 3 advise to the hosts in Airbnb Beijing as below:

1. Please be friendly and enthusiasm to the guestsxw
2. Please keep the house/apartment/room clean and cozy
3. Please provide some information about the transport for the guests