Analysis of Airbnb Listings in Beijing

——BIS 634 Final Project Report

Han Yu

M.S. Candidate in Biostatistics

Contents

1 Introduction

1.1 Background

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities[1]. Airbnb officially entered the market of China in 2015, setting off a wave of urban homestay boom in China.

Beijing is the capital of China and one of cities that Airbnb has reached great success. Beijing is referred as the political, cultural, scientific and technological as well as international exchange center of China[2]. It owns 7 World Heritage Sites, which makes Beijing the city with the largest number of cultural heritage projects in the world. The abundant functions of Beijing make it a city with a large population flow, which has contributed to Airbnb's success in China.

Based on Airbnb's public dataset, this project visualized Airbnb's operational data in Beijing from the view of hosts, and created a data panel to explore the distribution, basic information of listings and hosts in Beijing. What's more, most current analysis were from the perspective of tenants. In this project, perspective from hosts is adopted to gave the advice on price of their listings by using K-NN.

1.2 Data Resources and FAIRness Principle

1.2.1 Date Resources

The data used for this project was retrieved from the official website of Airbnb. They published dataset from different regions and countries. The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion[3]. This data was stored in .csv format and does not contain any personal privacy data.

1.2.2 FAIRness Principle

This dataset follows the FAIRness principle.

- Findability: The dataset used in this project is a public dataset. Airbnb official website provides the operation listings data in different countries and regions to public. Everyone can easily get these data from the Airbnb official website.
- Accessibility: This dataset do not need permission to retrieve because it’s a public dataset.
- Interoperability: The dataset is stored in .csv format. Each listing and each host have their own identified number to help this dataset be integrated with other data.
- Resuability: Data are richly described with a plurality of accurate and relevant attributes. People can easily understand the meanings of each variables.

2 Exploratory Data Analysis

2.1 Data overview and data preprocessing

2.1.1 Data Description

This data is stored in .csv format. 74 variables and 25026 listings are contained in the dataset. The variables can be divided into 5 parts and are shown below.

Table 1 Variable descriptions

Parts	Description	Number of variables	Examples
Basic information of the listings	Describe the basic information of each listing	8	listing id, listing url, listing picture url, listing name, listing description ……
Basic information of the host	Describe the basic information of the host	22	host_id, host_url, host_name, host_location, host_response_time ……
Geo information of the listing	Describe the basic Geo information of the listing	5	neighborhood, latitude, longitude……
Booking information of	Describe the booking condition of the listing	26	room_type, accommodates, amenities, price,

the listing			availability ……
Review information of the listing	Describe the review information of the listing	13	number_of_reviews, number_of_reviews, reviews_per_month, review_scores_accuracy ……

2.1.2 Data Preprocessing

Data preprocessing works are completed in the following four steps.

- Remove outliers of the listing data.

a. 24 listings that are under test, delisted and not rent are deleted by viewing the descriptions of the listings and find the key words like “test”, “delisted”, “not rent”.

b. 4 listings with unusual prices (higher than 990k or lower than 10) are removed.

- Standardization and data correction. Change the categorical variables into standard format.

- Define new variables to better describe the information of the listing.

a. host_type: A categorical variable used to define the type of the host based on the number of the listings the host has.

b. availability: A categorical variable used to define the availability of the listing based on the available days per month.

c. income: A numerical variable used to define the income of the host from each listing. The income can be calculated by multiplying the minimum number of days that can be booked, review per month, available days per month and price of the listing.

- Select variables and deal with the missing values. 16 variables and 24998 listings are selected after data preprocessing to do the future analysis.

2.2 Listing Data Analysis

To give a overview of the information of the whole city and each district, I conducted data analysis from several different perspectives.

- Listing Distribution

Based on the position of each listings and the geojson file of Beijing provided by Airbnb official website, I drew a map that describes the distribution of listings in Beijing. Each green point refers to a listing.

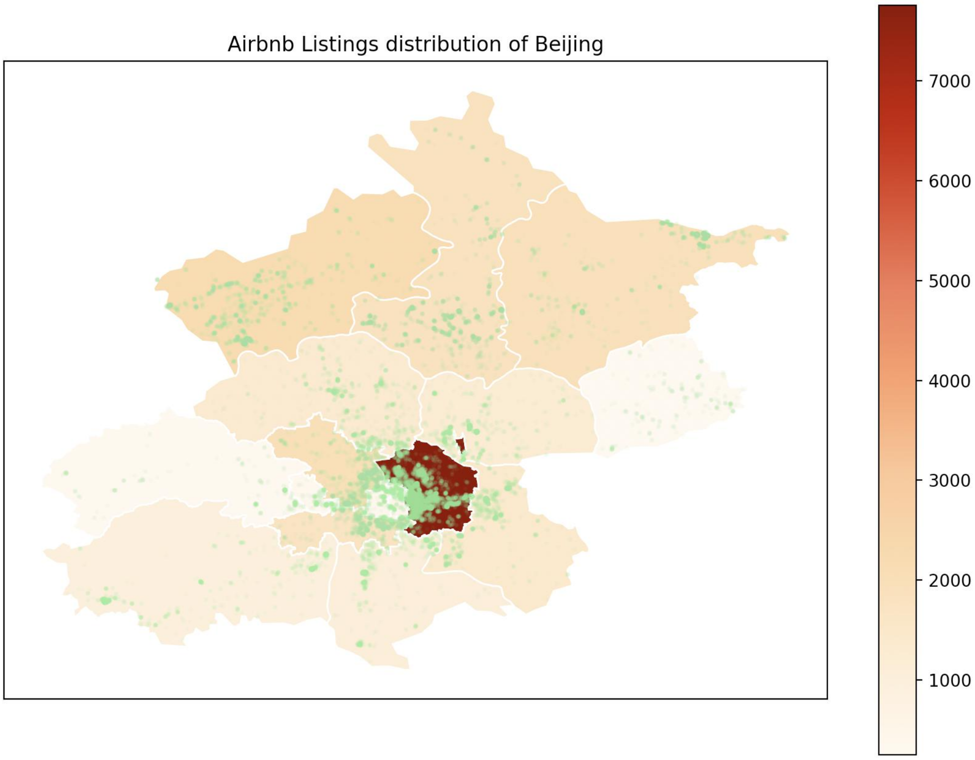


Figure 1: Airbnb listings distribution of Beijing

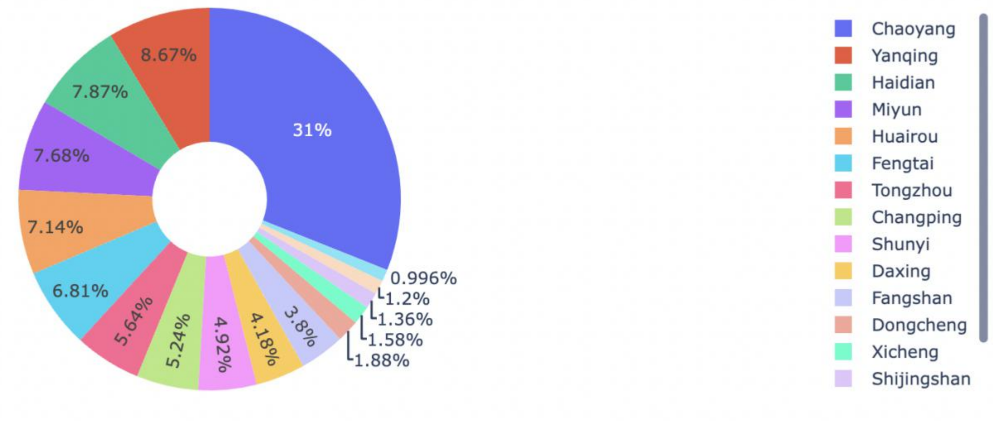


Figure 2: Airbnb listings distribution of each district in Beijing

- Other analysis

1) Room Type

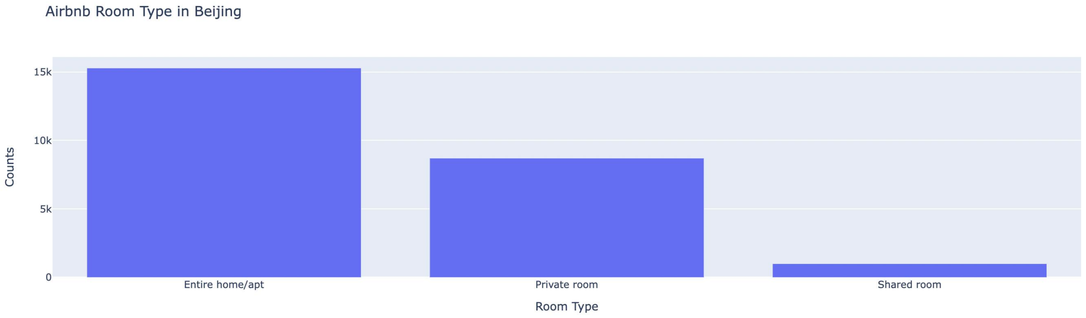


Figure 3: Airbnb room type in Beijing

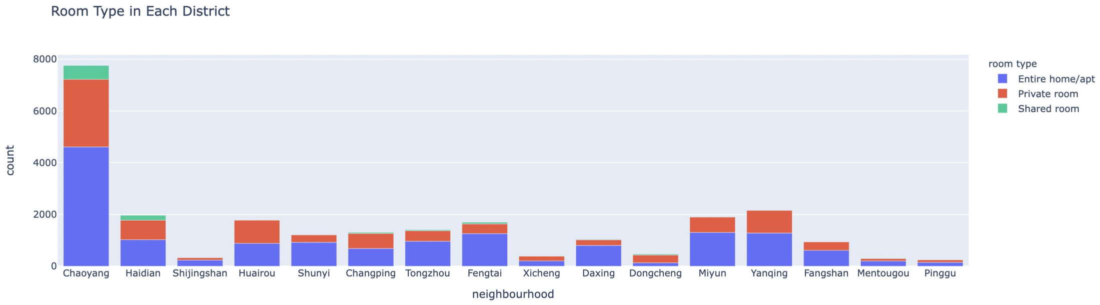


Figure 4: Airbnb room type in each district of Beijing

From the two figures above, I can see that Entire home/apt accounts for the largest proportions of listings in Beijing. What's more, shared room almost only exists in Chaoyang and Haidian district. This can be explained by the fact that Chaoyang and Haidian are two of the more densely populated districts.

2) Host Type

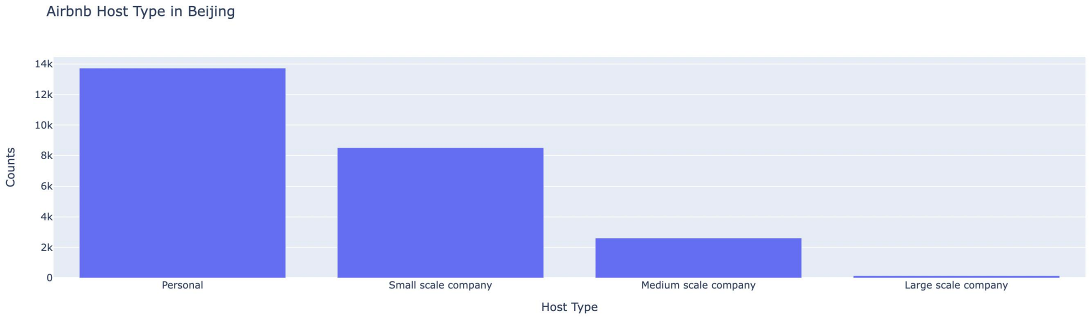


Figure 5: Airbnb host type in Beijing

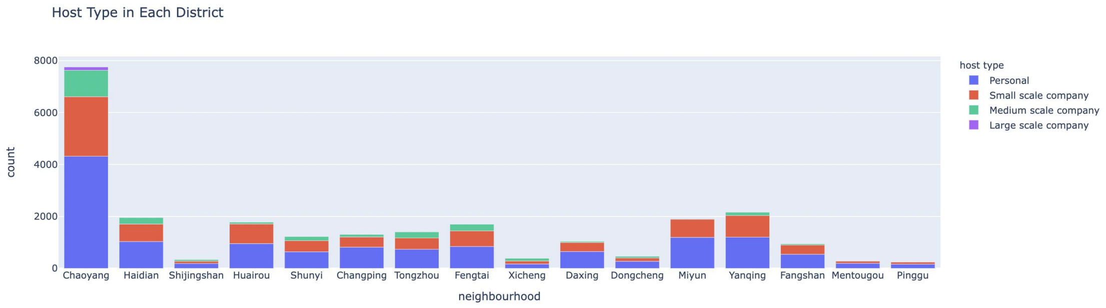


Figure 6: Airbnb host type in each district of Beijing

From Figure 5 and Figure 6, I can draw the conclusion that most host in Beijing is personal host, which means they rent their own listings. What's more, large scale company almost only runs their business in Chaoyang and Haidian district. As I stated in the explanation in the part of room type, the population densities of Chaoyang and Haidian are very high, which means the probability of people renting their listings is high. As a result, Chaoyang and Haidian will be a good choice for them to run their business

3) Availability

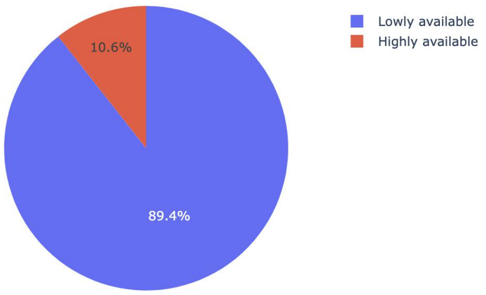


Figure 7. The proportion of availability of listings in Beijing

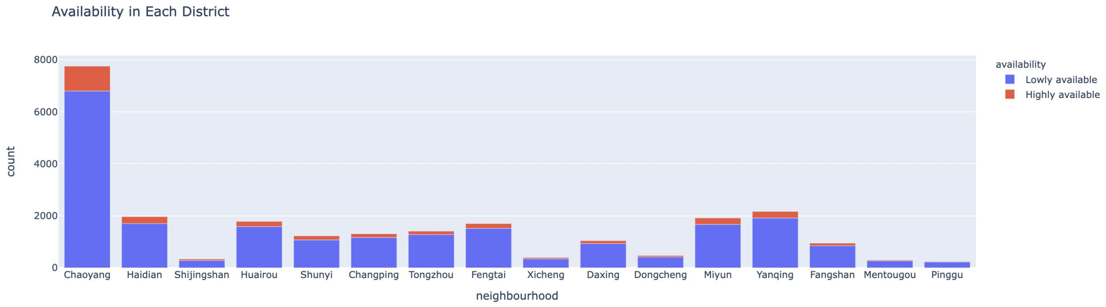


Figure 8. Airbnb availability in each district of Beijing

From the figures above, I can see that nearly 90% of the listings are under low available condition. This situation may due to the effect of Covid19. Last year, Airbnb stopped their business in China for a long time to fight against COVID-19. Thus, the available days of a large percent of listings are under 60 per year.

4) Price

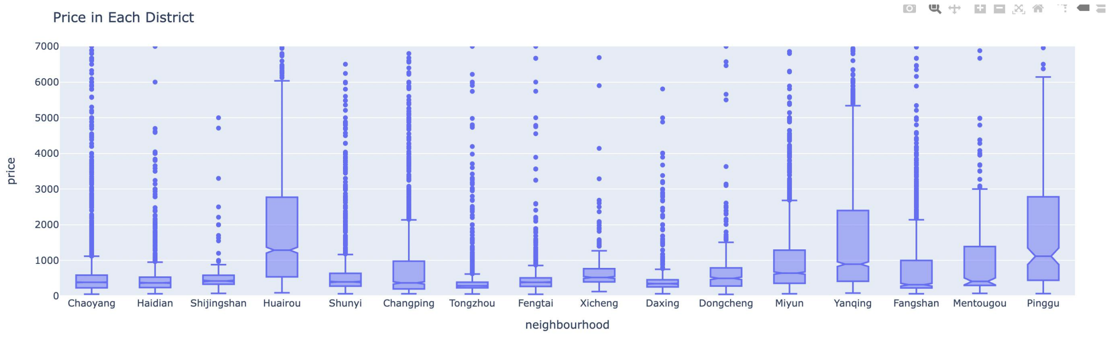


Figure 9. Airbnb price of the listings in each district

From the Figure 9, it's easy for us to get the conclusion that the average price of the listings in suburbs is much higher, like Huairou, Yanqing, and Pinggu. There are two reasons to explain this finding. On the one side, after reviewing the detailed information of the listings in these suburbs, I found that these listings are large enough to accommodate more people and have great environment which is great for group activities. On the other side, due to Covid-19, people prefer to escape from the center of the city and relax in suburbs, which means that the increase in demand has led to an increase in prices.

5) Income

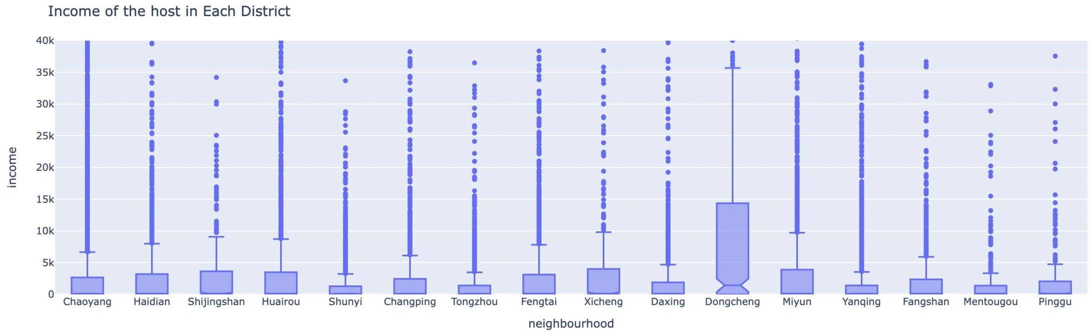


Figure 10. Income of host in each district of Beijing

From the above figure, we can reach a conclusion that the income of host in Dongcheng is the largest. Several reasons led to this result. First, Dongcheng is located in the very center of Beijing, close to the Forbidden City, and has many places of interest and historical sites. As a result, tourists prefer to live in Dongcheng district. Second, transportation and shopping in Dongcheng are very convenient compared with other districts. Last, the price of the housing in Dongcheng is in top place of Beijing.

3 Listing price modeling

3.1 Preprocessing of the data

I decided to use K-NN algorithm to get the k most similar listings (nearest neighbors) of the new listing based on some basic information of the new listings.

Two steps are taken to preprocess the data used for K-NN algorithm.

- Change the categorical variables expressed in word to numbers. For example, in variable “neighbourhood”, change the name of 16 districts to number 1 to 16.
- Normalize the data between 0 and 1 to eliminate the impact of magnitude in different variables.

3.2 Variables selected to retrieve the K nearest neighbors

The variables used to build the K-NN algorithm are shown below. The basic information of the listing and the host are both included in these variables.

Table 2. Variables selected to retrieve the K nearest neighbors

Variable name	Description	Information
neighbourhood	The district this listing belongs to.	Location of the listing.
latitude	The latitude of the listing.	Represent the position, the traffic information, and the surroundings of the listing.
longitude	The longitude of the listing.	
room_type	The room type of the listing. ['Entire home/apt', 'Private room', 'Shared	Basic information of the listing.

	room']	
accommodates	The largest number of people that can accommodate. Indicate the area of the listing.	
minimum_nights	The minimum number of nights that can book.	
availability_30	The largest number of days that your listing is available per month.	
host_type	The host type of the listing. ['Small scale company', 'Personal', 'Medium scale company', 'Large scale company']	Basic information of the host.

4 Back-end API and front-end website

4.1 Back-end API

Several steps are taken to develop the back-end API.

First, a Flask API was generated in back-end development to achieve the aim of interactive.

Second, several routes are generated to render pages from the HTML template. “index” function under route (“/”) renders “index.html”, which is the homepage of the website. “wholecityinonepage” function under route (“/wholecity”) renders “wholecityinonepage.html”, which shows the analysis result of the whole city. “district” function under route (“/district”) renders “district.html”, which shows the analysis result of a certain district. “knn” function under route (“/knn”) renders “advice.html”, which shows the information of the retrieved k nearest neighbors and the suggested price of the listing. Among these four pages, “district” and “knn” receive parameters that are inputted by the user in the homepage to conduct their further tasks.

Third, import the listing data into the main server and preprocessed the data. Create the functions to build the K-NN models and generate the visualization plots of the data.

Last, after all the routes of the flask API were designed. Run the “server_airbnb.py” to start the server.

4.2 Front-end website

The front-end website has 4 pages.

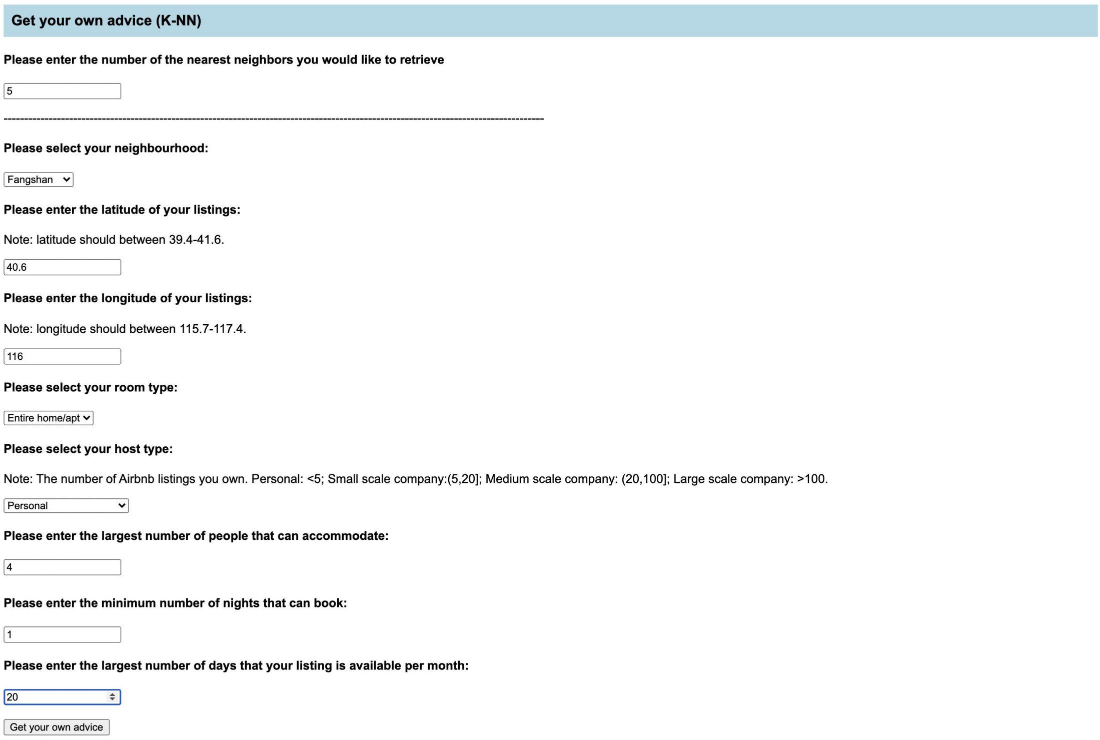
1) Homepage includes 6 parts. The first three parts provide some background information and data description of this project. The forth part is the analysis of the whole city, and there will be a link to jump to the Whole City Analysis page. The last two parts are interactive parts. Users enter or select the parameter and then click the button to jump to the result page.

2) Whole City Analysis page shows the results of data analysis and visualization of the whole city.

3) Each District Analysis page shows the results of data analysis and visualization of a certain district. The name of the district is selected by the host in the homepage and then transferred to the back-end API. Last, the returned result from the back-end API will be shown on the front-end website.

4) Get Your Own Advice page shows the information of the retrieved k nearest neighbors and the suggested price of the listing. The information of the new listing are entered in the homepage and then transferred to the back-end API.

For example, you are a host who has a new listing in Fangshan district. The latitude and longitude of the listing is 40.6 and 116 respectively. This is an entire home which can accommodate 4 people. The minimum number of nights that can be book is one and 20 days are available per month. You have other two listings besides this new listing, so you are a personal host. You want to get some suggestions of the price from the top 5 nearest neighbors. See the detailed information you entered in Figure 11.



Get your own advice (K-NN)

Please enter the number of the nearest neighbors you would like to retrieve

5

Please select your neighbourhood:

Fangshan

Please enter the latitude of your listings:

Note: latitude should between 39.4-41.6.

40.6

Please enter the longitude of your listings:

Note: longitude should between 115.7-117.4.

116

Please select your room type:

Entire home/apt

Please select your host type:

Note: The number of Airbnb listings you own. Personal: <5; Small scale company:(5,20]; Medium scale company: (20,100]; Large scale company: >100.

Personal

Please enter the largest number of people that can accommodate:

4

Please enter the minimum number of nights that can book:

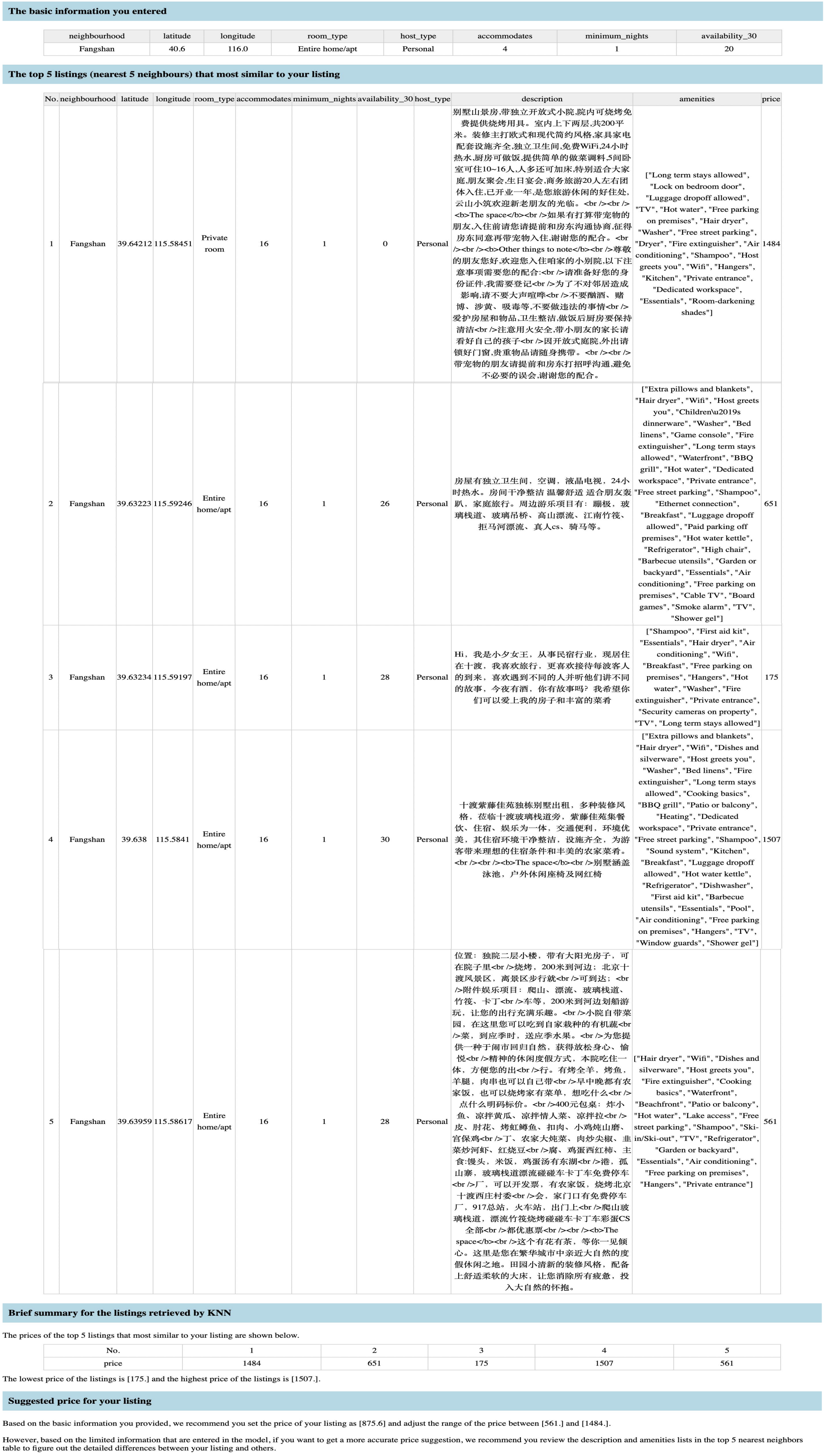
1

Please enter the largest number of days that your listing is available per month:

20

Figure 11. The demo of the entered information of KNN

After submitting the parameters you entered, you will jump to Get Your Own Advice page. There are four parts in this page. The First part shows the basic information you just entered. The second part returns the information of the 5 listings that are most similar to your listing. The third part will summarize the price of the retrieved listings. The last part aims to provide the suggestions about the price of the listing. Based on the basic information you entered, our model suggest you set the price of your listing to 875.6 and adjust the price between 561 and 1484. All these information are shown in Figure 12.



The basic information you entered

neighbourhood	latitude	longitude	room_type	host_type	accommodates	minimum_nights	availability_30
Fangshan	40.6	116.0	Entire home/apt	Personal	4	1	20

The top 5 listings (nearest 5 neighbours) that most similar to your listing

No.	neighbourhood	latitude	longitude	room_type	accommodates	minimum_nights	availability_30	host_type	description	amenities	price
1	Fangshan	39.64212	115.58451	Private room	16	1	0	Personal	别墅山景房,带独立开放式小院,院内可烧烤免费提供烧烤用具。室内上下两层,共200平米。装修主打欧式和现代简约风格,家具家电配套设施齐全,独立卫生间,免费WiFi,24小时热水,厨房可做饭,提供简单的做菜调料,5间卧室可住10~16人,人多还可加床,特别适合大家庭,朋友聚会,生日宴会,商务旅游20人左右团体入住,已开业一年,是您旅游休闲的好住处,云山小筑欢迎新老朋友的光临。<br /><br /><b>The space</b><br />如果有打算带宠物的朋友,入住前请您请提前和房东沟通协商,征得房东同意再带宠物入住,谢谢您的配合。<br /><br /><b>Other things to note</b><br />尊敬的朋友您好,欢迎您入住咱家的小别院,以下注意事项需要您的配合:<br />请准备好您的身份证件,我需要登记<br />为了不对邻居造成影响,请不要大声喧哗<br />不要酗酒、赌博、涉黄、吸毒等,不要做违法的事情<br />爱护房屋和物品,卫生整洁,做饭后厨房要保持清洁<br />注意用火安全,带小朋友的家长请看好自己的孩子<br />因开放式庭院,外出请锁好门窗,贵重物品请随身携带。<br /><br />带宠物的朋友请提前和房东打招呼沟通,避免不必要的误会,谢谢您的配合。	["Long term stays allowed", "Lock on bedroom door", "Luggage dropoff allowed", "TV", "Hot water", "Free parking on premises", "Hair dryer", "Washer", "Free street parking", "Dryer", "Fire extinguisher", "Air conditioning", "Shampoo", "Host greets you", "Wifi", "Hangers", "Kitchen", "Private entrance", "Dedicated workspace", "Essentials", "Room-darkening shades"]	1484
2	Fangshan	39.63223	115.59246	Entire home/apt	16	1	26	Personal	房屋有独立卫生间，空调，液晶电视，24小时热水。房间干净整洁 温馨舒适 适合朋友轰趴，家庭旅行。周边游乐项目有：蹦极，玻璃栈道、玻璃吊桥、高山漂流、江南竹筏、拒马河漂流、真人cs、骑马等。	["Extra pillows and blankets", "Hair dryer", "Wifi", "Host greets you", "Children\u2019s dinnerware", "Washer", "Bed linens", "Game console", "Fire extinguisher", "Long term stays allowed", "Waterfront", "BBQ grill", "Hot water", "Dedicated workspace", "Private entrance", "Free street parking", "Shampoo", "Ethernet connection", "Breakfast", "Luggage dropoff allowed", "Paid parking off premises", "Hot water kettle", "Refrigerator", "High chair", "Barbecue utensils", "Garden or backyard", "Essentials", "Air conditioning", "Free parking on premises", "Cable TV", "Board games", "Smoke alarm", "TV", "Shower gel"]	651
3	Fangshan	39.63234	115.59197	Entire home/apt	16	1	28	Personal	Hi，我是小夕女王，从事民宿行业，现居住在十渡，我喜欢旅行，更喜欢接待每波客人的到来，喜欢遇到不同的人并听他们讲不同的故事，今夜有酒，你有故事吗？我希望你们可以爱上我的房子和丰富的菜肴	["Shampoo", "First aid kit", "Essentials", "Hair dryer", "Air conditioning", "Wifi", "Breakfast", "Free parking on premises", "Hangers", "Hot water", "Washer", "Fire extinguisher", "Private entrance", "Security cameras on property", "TV", "Long term stays allowed"]	175
4	Fangshan	39.638	115.5841	Entire home/apt	16	1	30	Personal	十渡紫藤佳苑独栋别墅出租，多种装修风格，莅临十渡玻璃栈道旁，紫藤佳苑集餐饮、住宿、娱乐为一体，交通便利，环境优美，其住宿环境干净整洁，设施齐全，为游客带来理想的住宿条件和丰美的农家菜肴。<br /><br /><b>The space</b><br />别墅涵盖泳池，户外休闲座椅及网红椅	["Extra pillows and blankets", "Hair dryer", "Wifi", "Dishes and silverware", "Host greets you", "Washer", "Bed linens", "Fire extinguisher", "Long term stays allowed", "Cooking basics", "BBQ grill", "Patio or balcony", "Heating", "Dedicated workspace", "Private entrance", "Free street parking", "Shampoo", "Sound system", "Kitchen", "Breakfast", "Luggage dropoff allowed", "Hot water kettle", "Refrigerator", "Dishwasher", "First aid kit", "Barbecue utensils", "Essentials", "Pool", "Air conditioning", "Free parking on premises", "Hangers", "TV", "Window guards", "Shower gel"]	1507
5	Fangshan	39.63959	115.58617	Entire home/apt	16	1	28	Personal	位置：独院二层小楼，带有大阳光房子，可在院子里<br />烧烤，200米到河边；北京十渡风景区，离景区步行就<br />可到达；<br />附件娱乐项目：爬山、漂流、玻璃栈道、竹筏、卡丁<br />车等，200米到河边划船游玩，让您的出行充满乐趣。<br />小院自带菜园，在这里您可以吃到自家栽种的有机蔬<br />菜，到应季时，送应季水果。<br />为您提供一种于闹市回归自然，获得放松身心、愉悦<br />精神的休闲度假方式，本院吃住一体，方便您的出<br />行。有烤全羊，烤鱼，羊腿，肉串也可以自己带<br />早中晚都有农家饭，也可以烧烤家有菜单，想吃什么<br />点什么明码标价。<br />400元包桌：炸小鱼、凉拌黄瓜、凉拌情人菜、凉拌拉<br />皮、肘花、烤虹鳟鱼、扣肉、小鸡炖山磨、宫保鸡<br />丁、农家大炖菜、肉炒尖椒、韭菜炒河虾、红烧豆<br />腐、鸡蛋西红柿、主食:馒头，米饭，鸡蛋汤有东湖<br />港，孤山寨，玻璃栈道漂流碰碰车卡丁车免费停车<br />厂，可以开发票，有农家饭，烧烤北京十渡西庄村委<br />会，家门口有免费停车厂，917总站，火车站，出门上<br />爬山玻璃栈道，漂流竹筏烧烤碰碰车卡丁车彩蛋CS全部<br />都优惠票<br /><br /><b>The space</b><br />这个有花有茶，等你一见倾心。这里是您在繁华城市中亲近大自然的度假休闲之地。田园小清新的装修风格，配备上舒适柔软的大床，让您消除所有疲惫，投入大自然的怀抱。	["Hair dryer", "Wifi", "Dishes and silverware", "Host greets you", "Fire extinguisher", "Cooking basics", "Waterfront", "Beachfront", "Patio or balcony", "Hot water", "Lake access", "Free street parking", "Shampoo", "Ski-in/Ski-out", "TV", "Refrigerator", "Garden or backyard", "Essentials", "Air conditioning", "Free parking on premises", "Hangers", "Private entrance"]	561

Brief summary for the listings retrieved by KNN

The prices of the top 5 listings that most similar to your listing are shown below.

No.	1	2	3	4	5
price	1484	651	175	1507	561

The lowest price of the listings is [175.] and the highest price of the listings is [1507.].

Suggested price for your listing

Based on the basic information you provided, we recommend you set the price of your listing as [875.6] and adjust the range of the price between [561.] and [1484.].

However, based on the limited information that are entered in the model, if you want to get a more accurate price suggestion, we recommend you review the description and amenities lists in the top 5 nearest neighbors table to figure out the detailed differences between your listing and others.

Figure 12. The demo of returned result of KNN page

5 Discussion

5.1 Interesting findings

Compared the distribution of listings of Beijing this year (2021) with that of 2019 (Figure 13), there are some interesting findings. First, the number of listings in Chaoyang District in 2019 is 3,000 more than this year (2021). This may shows the effect of Covid-19. Second, the center of the city accounts for more than 60% percent of the listings in 2019. However, this year (2021), the number of listings in suburb is pretty high, especially in Yanqing, Miyun and Huairou. I think two reasons may explain this finding. One the one hand, due to Covid-19, people prefer to go to suburbs to take vacations. One the other hand, Yanqing district is the venue for the Winter Olympics in China, so the tourism in Yanqing has developed a lot.

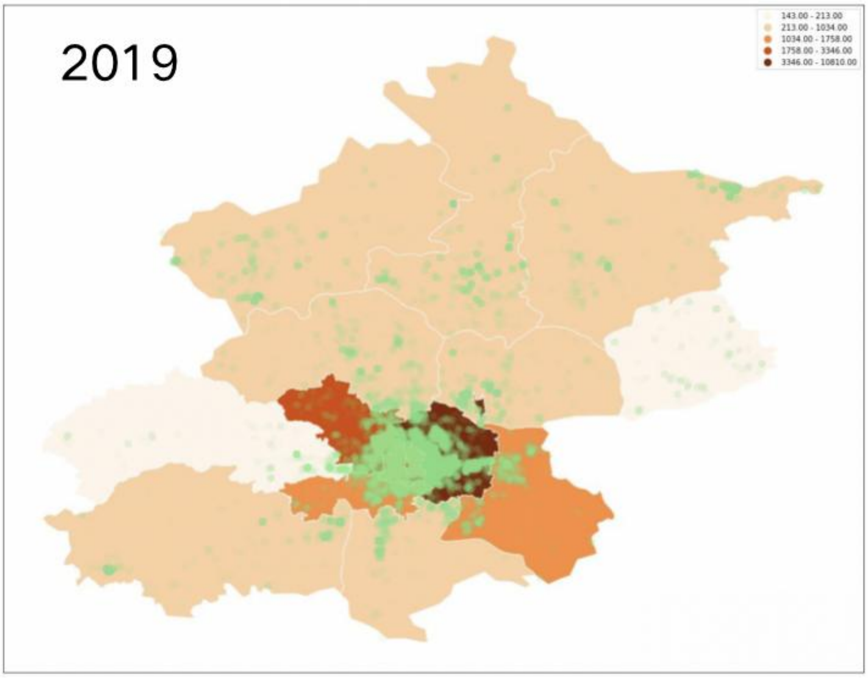


Figure 13. Listing distribution of Beijing in 2019

5.2 Limitations and Future work

There are some limitations of this project.

Firstly, because of the lack of information about the surrounding environment of the listing, the geographic location of the listing can only be expressed by using latitude and longitude. However, even if the latitude and longitude are very similar, the sensitivity of them to traffic and shopping is very high, that is to say, a little change in latitude and longitude will cause great changes in the surrounding environment of the listing. This may lead to inaccurate analysis of KNN algorithm.

Second, due to lack of detailed information of the listings, such as housing size, surrounding environment and so on, the price of the retrieved nearest neighbors may vary largely from each other. If this situation happens, users are suggested to review the detailed information of the listings in the column of “description” and “amenities” to compare the differences between the retrieved listings and their own listings.

The further work should focus on provide information that can better describe the detail of the listings, such as the amenities, the traffic information, the surrounding environment, the area of the listing and so on. What’s more, improve the API and the front-end website to make it more interactive with users.

References

[1] Wikipedia. Airbnb. URL: https://en.wikipedia.org/wiki/Airbnb.

[2] Wikipedia. Beijing. URL: https://en.wikipedia.org/wiki/Beijing.

[3] Inside Airbnb. URL: http://insideairbnb.com/get-the-data.html.

Code

Code and data are available on github. The demos of the front-end website are also provide on github.

See here:

https://github.com/Hannah-Yu-0816/Airbnb_Listing_in_Beijing_Analaysis