EDA, Visualization, Clustering and Decision Tree of Beijing Airbnb Dataset

Tianyi Li, Jiaqi Yang

2019/12/12

1. Abstract

This paper analyzes the public Beijing Airbnb dataset, builds a decision tree model (including PCA dimensionality reduction processing) for predicting the price of listings and divides the apartments of listings into different groups via clustering. The exploratory data analysis performs a visualized summarization about the statics of listings and type of property, average prices and rating scores sorted by districts, the occupancy rate in the future, word clouds of comments and demand changes overtime. Predicting the price of listing based on provided attributes via DecisionTreeRegressor model, and mean absolute deviation is the measure to evaluate the model. Meanwhile, PCA model is used for reducing dimensions to improve data quality and explained variance ratio is adopted as a measure of the quality of dimensionality reduction. When it comes to clustering, we use K-means model and find the optimal number of clusters based on the Silhouette Coefficient.

2. Introduction

In this project, we – Tianyi Li and Jiaqi Yang, go through the public dataset of Airbnb in Beijing. Our report contains the exploratory data analysis, visualization, clustering work and price prediction by a decision tree. Answer some questions that people concern about while planning to visit Beijing and searching for a place to stay by Airbnb:

1. What are the different types of properties in Beijing? Do they vary by the district?
2. Which area is highly rated by renters?
3. What is the average cost of listings in each district?
4. What is the best time to visit Beijing?
5. How many days/weeks/months should I make the reservation in advance?
6. What do renters care about when searching rooms?
7. What aspects affect the price? Which one is the most important?
8. How to predict apartment prices based on known information?
9. How to group existing apartments based on available information, and how many groups are most suitable?

3. Background

The project referred to some methods used in the NYC Airbnb dataset analysis made by Ankit Peshin, Sarang Gupta, Ankita Agrawal. Their work of map of average price and rating scores map, the map showing Count of Listing by Borough, Number of Reviews across Time, calendar heatmap and word cloud inspired us.

4. Description of Data

The dataset we analyzed is from the website Inside Airbnb http://insideairbnb.com/index.html which provides Airbnb information in cities around the world available to the public.

The dataset consists of several tables and we mainly concentrated on four tables:

1. Listings – Detailed listing data with 96 attributes
2. Reviews – Rating scores with six attributes
3. Calendar – The calendar data for reservation in the next year

The dataset contains 34,745 unique listings in Beijing from Aug.6th 2010 to Sep 9^{th,} 2019. In 9 years, over 250,000 renters left their reviews. The price per night ranges from ￥28 to ￥71,597.

5. Method

Data Preprocessing

Attributes for EDA and Visualization

To perform the EDA and visualization, we selected 14 key attributes from 3 tables that are related to the subsequent work:

1. *host_id* (discrete) – The unique Airbnb ID of hosts
2. *listing_id* (discrete) – The unique Airbnb ID of listings
3. *name* (textual) – Listing's title in Airbnb
4. *neighbourhood_cleansed* (textual) – The name of the district where the listing locates.
5. *latitude* (continuous) – Latitude information in the geographic coordinate system
6. *longitude* (continuous) - Longitude information in the geographic coordinate system
7. *property_type* (discrete) – Type of property of listings
8. *price* (continuous) – Price per night in RMB(¥)
9. *rating* (continuous) – The overall rating scores out of 100
10. *review_date* – The date which renters left reviews
11. *host_since* (discrete) – The date that listings have initially placed on Airbnb
12. *reservation_date* (discrete) – The date of reservation
13. *available* (discrete) – Whether the room is occupied on given reservation_date
14. *comments* (textual) – Comments toward listings

Attributes for Clustering and Decision Tree

Based on the attributes above, we picked extra 21 attributes to perform the clustering and price prediction by a decision tree.

1. *room_type* (textual) - The type of booking room.
2. *bed_type* (textual) - The type of the bed.
3. *cancellation_policy* (textual) - The policy and fee of cancellation.
4. *host_identity_verified* (textual) - Whether the host is verified.
5. *host_is_superhost* (textual) - Whether the host is a super host.
6. *bathrooms* (discrete) - The number of bathrooms.
7. *bedrooms* (discrete) - The number of bedrooms.
8. *beds* (discrete) - The number of beds.
9. *security_deposit* (continuous) - The security deposit.
10. *cleaning_fee* (continuous) - The cleaning fee.
11. *review_scores_rating* (continuous) - The overall rating scores out of 10.
12. *review_scores_accuracy* (continuous) - The rating scores of descriptive accuracy out of 10.
13. *review_scores_cleanliness* (continuous) - The rating scores of cleanliness out of 10.
14. *review_scores_checkin* (continuous) - The rating scores of checking in out of 10.
15. *review_scores_communication* (continuous) - The rating scores of communication out of 10.
16. *review_scores_value* (continuous) - The rating scores of value out of 100.
17. *review_scores_location* (continuous) - The rating scores of location out of 10.
18. *reviews_per_month* (continuous) - The number of reviews per month.
19. *host_response_rate* (continuous) - The response rate of host.
20. *extra_people* (continuous) - The fee of each extra people.
21. *amenities* (textual) - The available amenities in the apartment.

Key Feature Engineering

1. *price* (listings): We noticed that the dataset used US Dollar ($) as the currency in Beijing which mismatched with the local consumption level. We compared the prices listed on the Airbnb website and found the currency is supposed to be CNY(¥). Therefore, we change the currency of price to CNY.
2. *comments* (reviews): Since the official language Beijing is Chinese (Putonghua), so we only selected comments written in Chinese and English.
3. *host_is_superhost* and of *host_identity_verified* (listings): We converted Boolean 'T' and 'F' to 0 and 1.
4. *neighbourhood_cleansed, property_type, room_type, bed_type and cancellation_policy (*listing): We converted the values to Dummy Variables.

Dealing with Missing Data

We detected the null value of the *'host_is_superhost'* and *'host_response_rate*' in the dataset and found that there are few samples with a null value of this attribute, so we removed the values whose attribute is empty. However, we found that some attributes (*'bathrooms', 'bedrooms', 'beds', 'security_deposit', 'cleaning_fee', 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_value', 'review_scores_location', 'reviews_per_month'*) of the sample had more null values, and in order to make the results more accurate, we filled these null values with the average value of each attribute. Next, we uniform all the values by removing the symbol ('$', ',', '%') and duplicate values.

EDA and Visualization

The dataset contains over 30 different types of property, so we built six categories and organized them by the following rules:

1. Apartment: Apartment and Serviced apartment
2. Hotel: Aparthotel, Boutique hotel, and Hotel,
3. Condominium: Farm stay, Condominium, Hostel, Guesthouse, and Guest suite
4. House: Bungalow, Cabin, Castle, Chalet, Cottage, Dome house, Earth house, House, Nature lodge, Resort and Villa
5. Loft: Loft
6. Other: Barn, Bed and breakfast, Campsite, Casa particular (Cuba), hut, Igloo, Kezhan (China), Minsu (Taiwan), Other, Pension (South Korea), Tent, Tiny house and Treehouse

We used pyecharts and matplotlib to implement the visualization.

Spatial Data Analysis

We calculated the total number of listing in each district. Then we calculated how many percents of each type of property that each district has. Then, we studied the price and rating score of each listing and the average values of every district. Then we plotted maps to demonstrate the distribution of prices and rating scores.

Growth of Demand in Airbnb over Time

We took the '*number of reviews*' variable as the demand to reflect the number of users and how popular using Airbnb is in years and months.

Occupancy Rate:

We used the table *calendar* to build a calendar showing the occupancy rate in the future. Every small block represents the percentage occupancy.

Comment Analysis:

We built word clouds in order to figure out what do renters care about when leaving comments. The word cloud takes the frequency of certain keywords and the output graph consists of the most frequent keywords. The size of each word represents how often renter mentioned it in all the comments.

Decision Tree

Establishment of decision trees model：

We use the DecisionTreeRegressor model of sklearn, set the random_state number to 30, and set the criterion to mse, maximum tree depth of 10. Then, we divide the data set, 30% is the test set, 70% is the train set. Next, we call the fit and predict methods to train and predict the dataset, respectively.

We calculated the Deviation of each data and calculated the total Mean Absolute Deviation is 474, and we use this value as a measure of the effectiveness of the decision tree.

We show each sample and the predicted value and the true value in a column chart(Using the sample serial number and house price as the x-axis and y-axis, respectively. And red line is the predicted price, blue line is the real price), and save the output of the comparison result to csv file.

Dimensionality Reduction :

We can see that the Mean Absolute Deviation value is large, which means that the predicted apartment house price result is not ideal. Thus, in order to improve the effect of the decision tree, we use the PCA method to reduce dimension the attributes of the input of the previous decision tree.

Dimension Reduction Model:

We used the PCA model of sklearn, set svd_solver to 'full'. In order to get the best results, we chose to test the model from 1 to 20 when setting the number of dimensions (which is parameter ‘n_components’). Here, we use the total value of ‘explained_variance_ratio_’ as a measure to evaluate how many dimensions the dimension reduction works best. We have drawn graphics to show the results(Using the total value of ‘explained_variance_ratio’ and the number of dimensions as the x-axis and y-axis, respectively). Since this value is larger the better, and we found that when the dimension is reduced to almost 8 dimensions, this value no longer rises, so we choose 8 as the best dimension after dimension reduction.

After the dimensionality reduction is completed, we model and train the decision tree on the dimensionality-reduced data. Here the decision tree training uses the same parameters as before, and also outputs the Mean Absolute Deviation result (which is 541), the Bar chart of the prediction result comparison, and the csv file of the comparison of the prediction result.

Clustering

Building a clustering model:

We applied clustering to divide the apartments in the data into different groups by KMeans model of sklearn. In order to achieve the best clustering effect, we used the Silhouette Coefficient as a measure of the clustering effect and sequentially set k to 2 to 8 as the parameter of the number of clusters for clustering. Then, we draw scatter plots of the effect of different clustering numbers (using regions and prices as the x-axis and y-axis, respectively. And '+' sign of the same color represents the same cluster), and simultaneously plot the Silhouette Coefficient (using clustering numbers and Silhouette Coefficient as the x-axis and y-axis respectively). And, based on the value of Silhouette Coefficient, we choose the best number of clusters

6. Result

1. Number of Listings

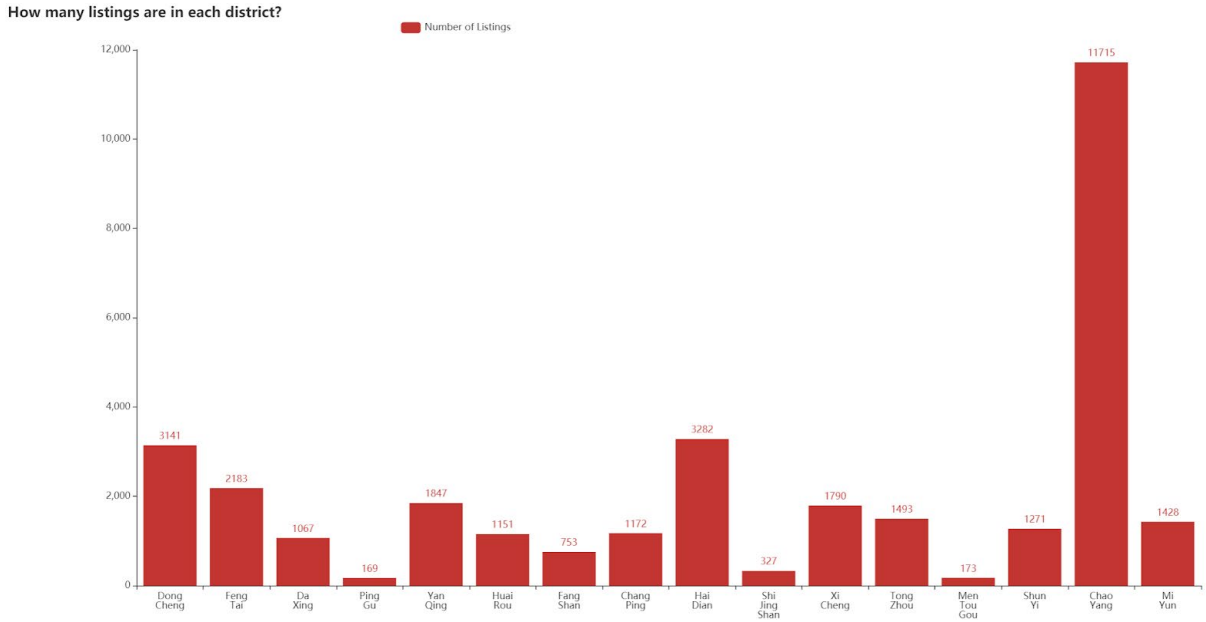

Table 6.1.1 Number of Listing in each District

From table 6.1.1, we found that the Chaoyang District has the largest number of listings of 11,715 which takes about one-third of the whole listing in Beijing.

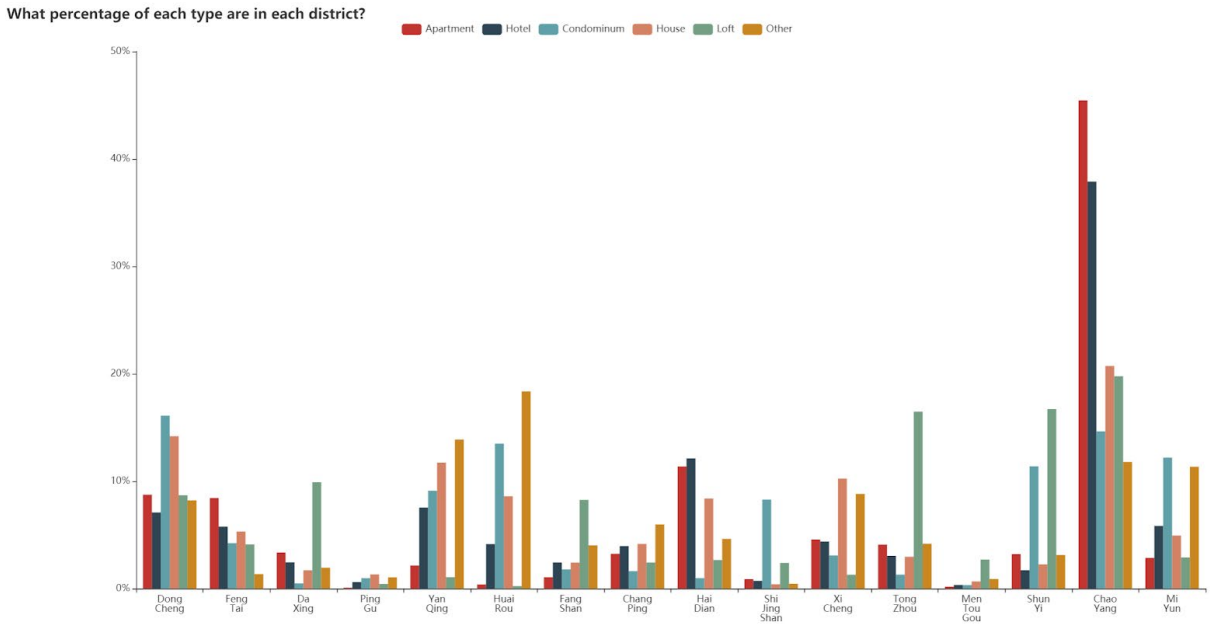

Table 6.1.2 Types of Properties in each District

What we observed:

1. Chaoyang District dominates several types of property. It has the largest number of apartments, hotels, houses and lofts which takes over 45%,40%, 20% and 18%. One possible reason is that Chaoyang has a comparatively larger area than other urban districts with high population density.
2. As two of the most crowded urban areas and the center of Beijing, both of Dongcheng and Xicheng Districts has over 10% of houses because there are many Siheyuan, a traditional type of Chinese courtyard housing, protected by the Beijing government.
3. Some suburb or near suburb districts like Yanqing, Huairou and Miyun District has many listings that their types of property are difficult to categorize including Barn, campsite, Igloo, Treehouse, Kezhan, Minsu and etc.
4. Lofts are popular among Dongcheng, Daxing, Fangshan, Tongzhou, Shunyi and Chaoyang.

2. Review & Rating

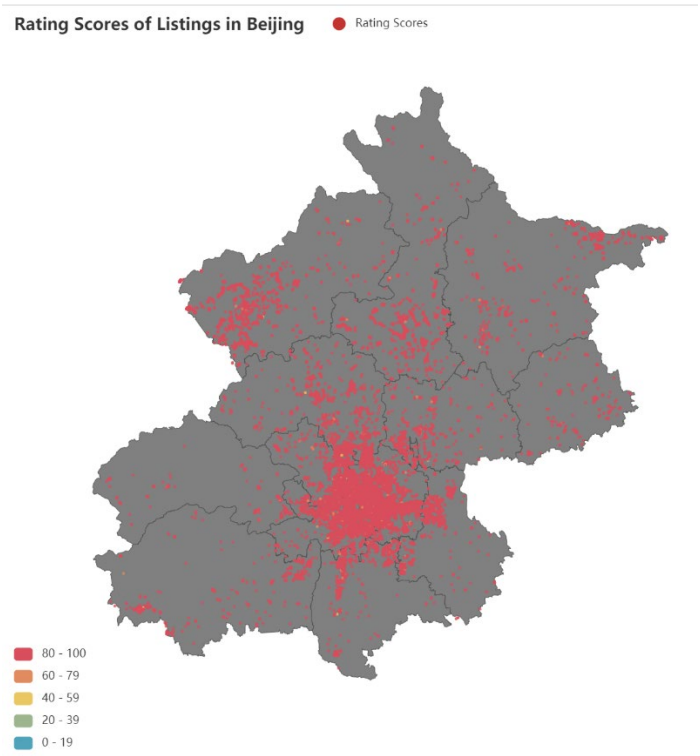

Figure 6.2.1 Rating Scores

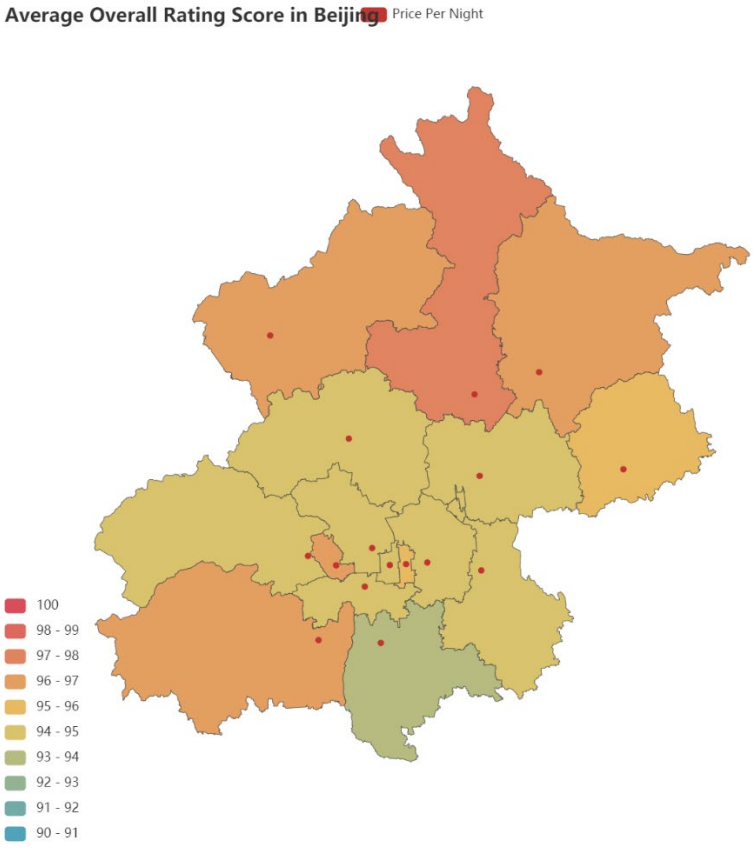

Figure 6.2.2 Average Rating Scores

As the figure 6.2.1 shown, most of the listing are clustered inside the urban area on the center of the map and most of the rating score is above 80 out of 100. For the average rating score, as figure 6.2.2 shown, all of the districts have an average score of over 93 out of 100. What made us surprised is that Huairou, as a suburb area far from the center of Beijing, won the best Districts in Beijing for an average score of 97.28. Daxing District has the lowest score of 93.77.

3. Price

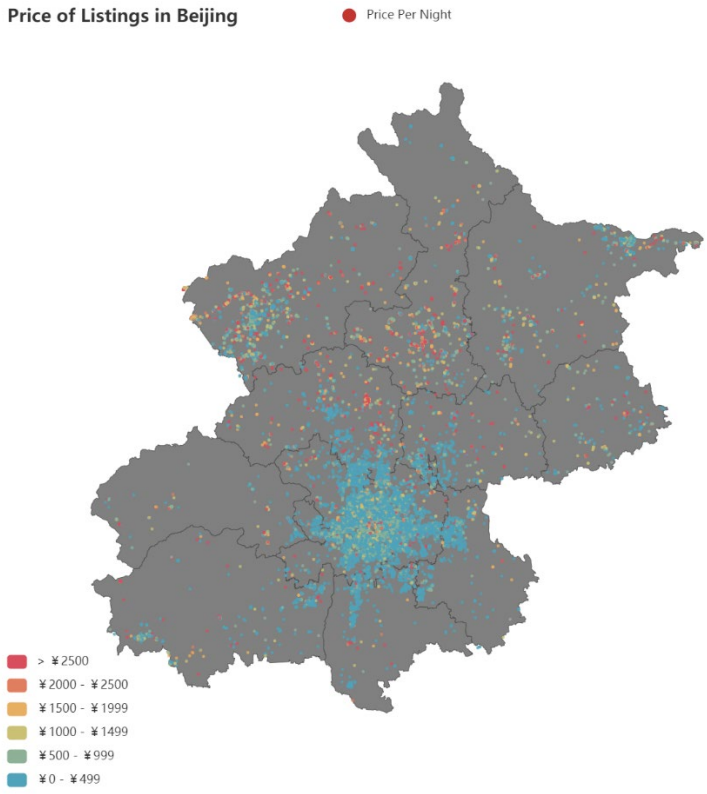

Figure 6.3.1 Price of Listings

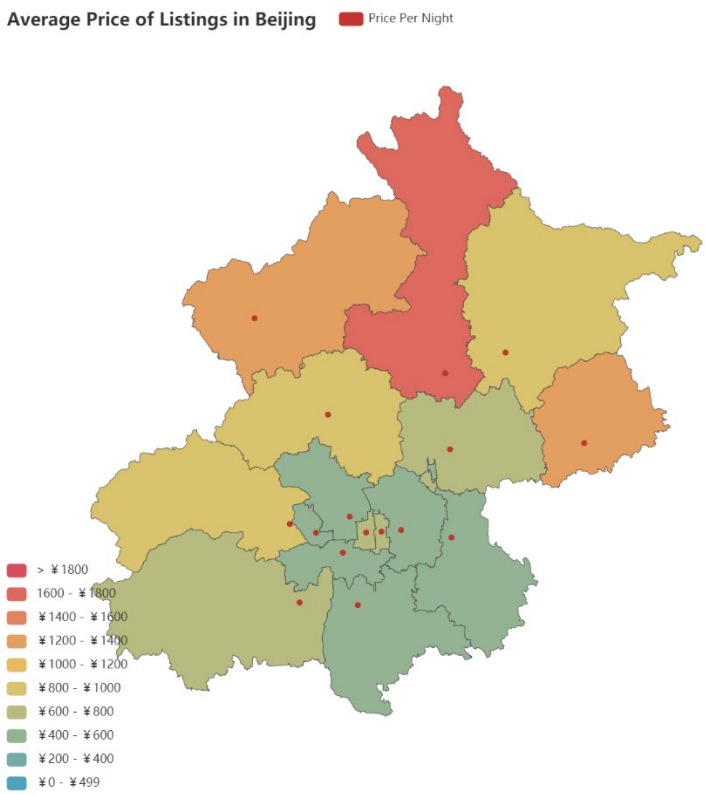

Figure 6.3.2 Average Prices in each District

As figure 6.3.1 shown, most of the cheap listings (below ¥ 500 per night) are in the urban area. Except for the center of the urban area, most of the listings priced over ¥ 1500 in the suburb areas. For the average prices, the result showed some similarities as the average rating scores. Huairou District still takes on first place for an average price of above ¥ 1800 per night. Although Huairou has only 1151 listings, we can see that Huairou seems to have less cheap listings than other districts have. Averages prices in urban areas are diluted by cheap listings.

4. Occupancy Rate: What times of year are the busiest?

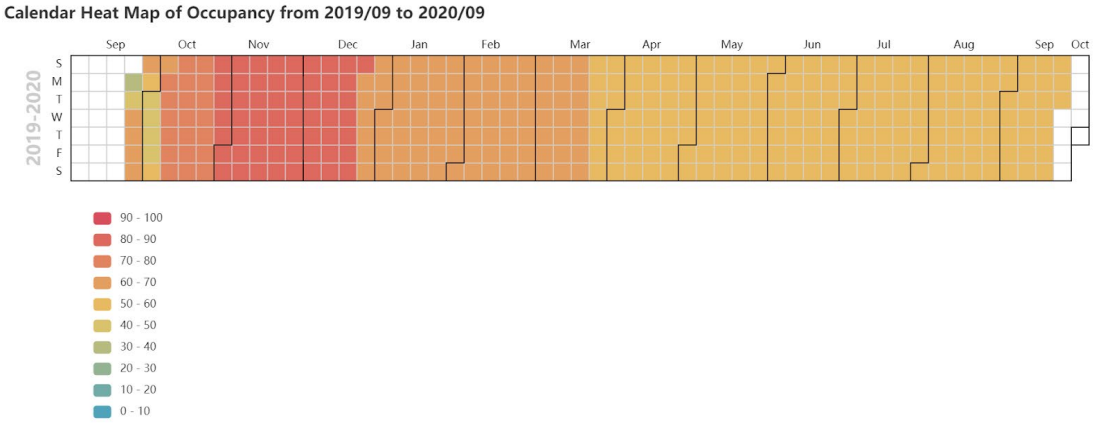

Figure 6.4.1 Heat Map of Occupancy Rate from Sep. 2019 to Sep. 2020

Figure 6.4.1 demonstrates the one-year occupancy rate in the future from September 2019. Starting at the end of September, the occupancy rate continuously increases until the end of 2019. The occupancy rate drops around Christmas. Although people often leave Beijing between January and February due to the Chinses new year, the occupancy rate still holds above 60% until the middle of March. For someone who is planning to travel to Beijing and use Airbnb, users should better make the reservation at least two months in advance.

5. Word Cloud

Figure 6.5.1 Word Cloud Written in Chinese

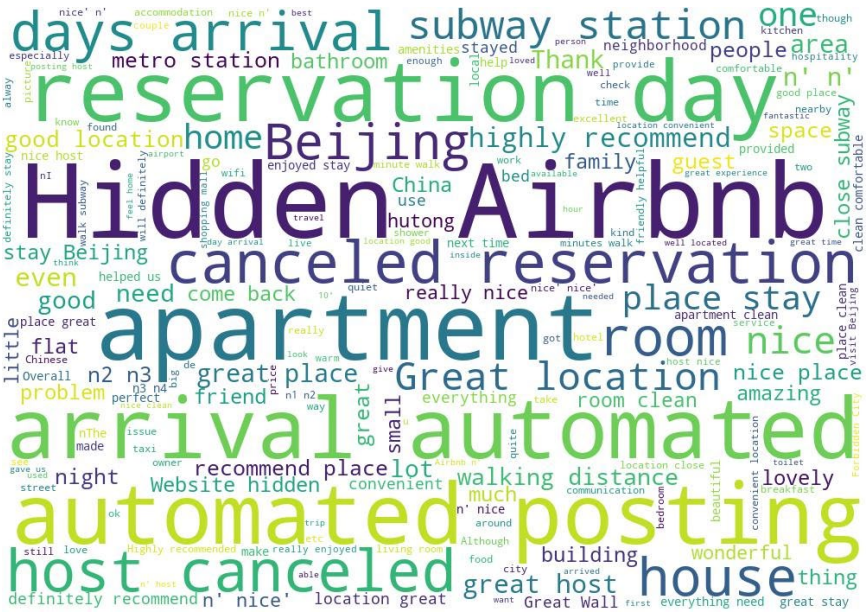

Figure 6.5.2 Word Cloud Written in English

In the Figure 6.5.1, Chinese word cloud, despite some general words saying “good” or “nice” (“好“，”挺好的“，”不错“，“非常好”，“挺好“，“很不错”), renters frequently mentioned “host is nice”（“房东很好”）, “clean”（“干净

“，”整洁“）, “good value for money”（”性价比高“） and “good transportation”（”交通便利“）. Therefore, people who speak Chinese do pay attention to the hosts, cleanness of rooms, the economy of listings and transportation.

In the Figure 6.5.2, English word cloud, “hidden Airbnb” and “hidden website” suggest that renters often face problems with the host hiding their list by various reasons. “arrival automated”, “automated posting”, “cancel reservation” and “reservation day” shows some issues while making reservations because some hosts might cancel the reservation and then renters will receive an automated message about the cancellation. “subway station”, “metro station’, ‘walking distance” and “close (to) subway” indict the renter care about transportation. They also mention a lot about the location.

6. Demand Analysis

In this section, we analyzed the change of demand over time. Since there is no particular data represents the demand, we considered the number of reviews as the demand because each review represents a complete order. So we direct analyzed the demand based on reviews and time.

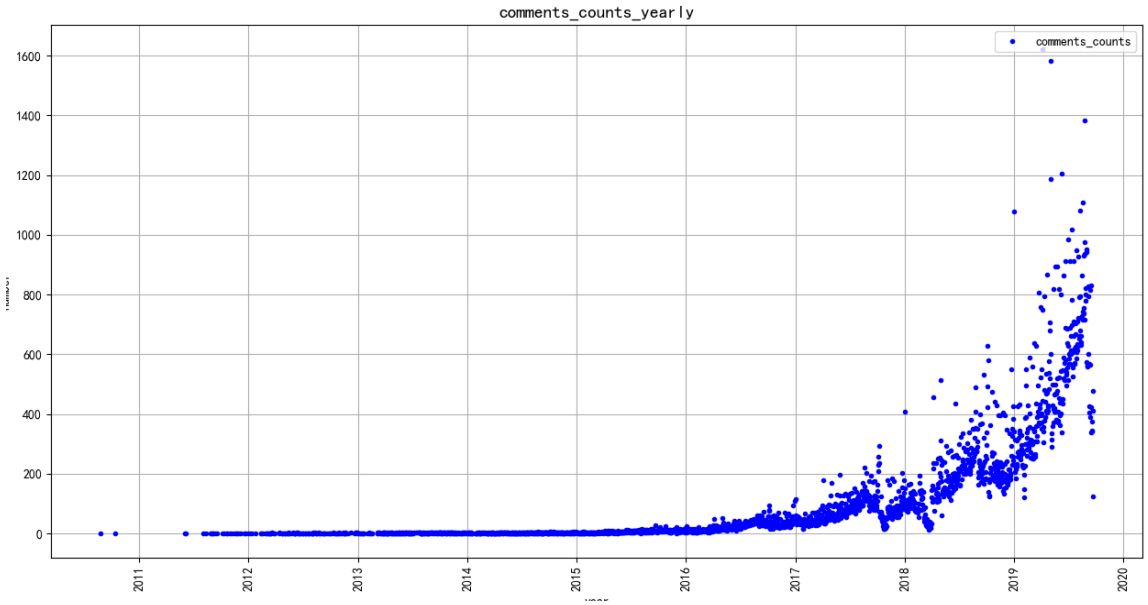

Figure 6.6.1 Number of Reviews vs. Years

Figure 6.6.1 shows an exponential increase in the number of reviews over the years.

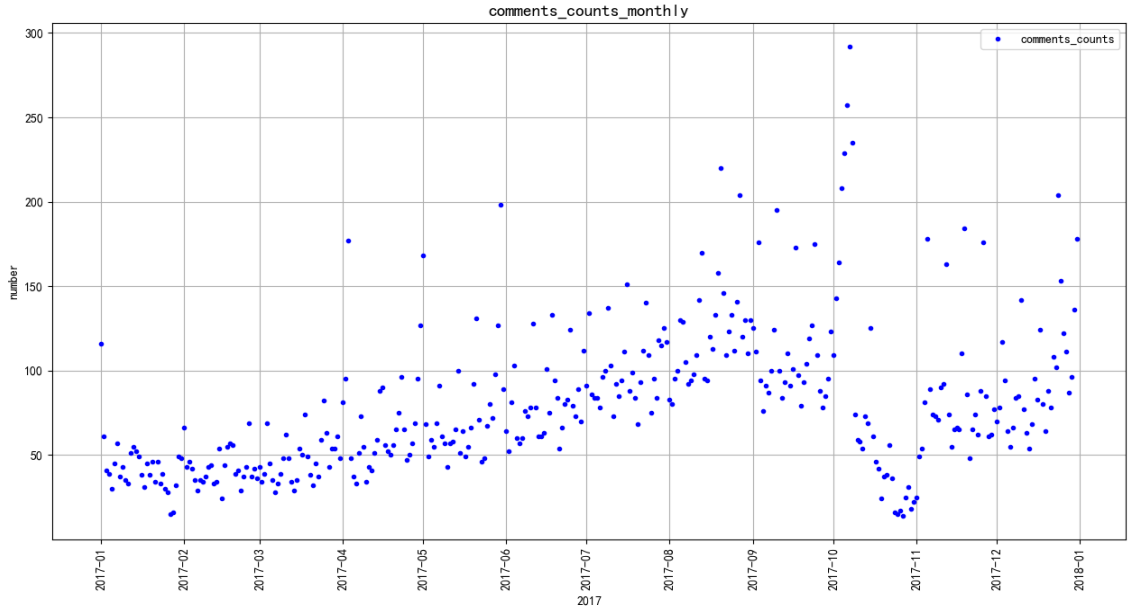

Figure 6.6.2 Number of Reviews vs. Months in 2017

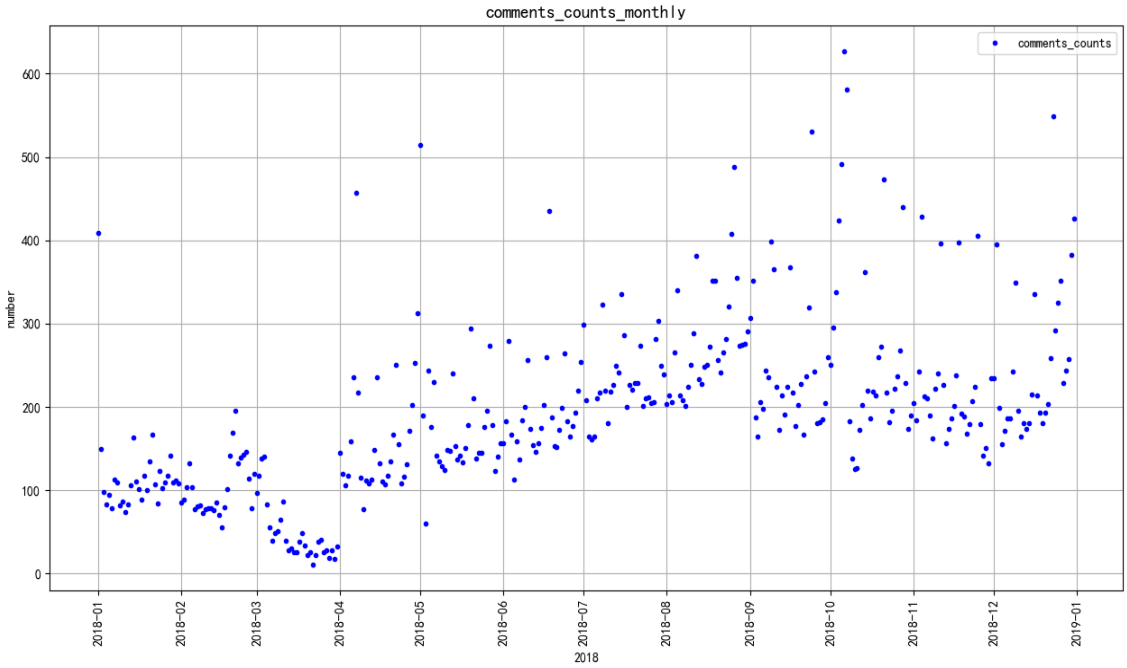

Figure 6.6.3 Number of Reviews vs. Months in 2018

Figure 6.6.2 and 6.6.3 show some similarities. The number of reviews constantly increases until September. After September, the summer break is over and most of the young people must go back to school which causes a decline in demand. After November, the demand starts increasing again to the next year.

7. Correlations between Attributes

```
data.corr()
```

	host_is_superhost	host_identity_verified	accommodates	bathrooms	bedrooms	beds	price	extra_people	minimum_nigh
host_is_superhost	1.000000	0.099205	0.042101	0.049410	0.049774	0.051015	0.033270	-0.044282	0.0181
host_identity_verified	0.099205	1.000000	0.028249	0.021532	0.023490	0.025070	0.003788	-0.053384	-0.0149
accommodates	0.042101	0.028249	1.000000	0.580214	0.778337	0.765258	0.311013	0.003079	-0.0206
bathrooms	0.049410	0.021532	0.580214	1.000000	0.650613	0.591175	0.251823	0.014916	-0.0157
bedrooms	0.049774	0.023490	0.778337	0.650613	1.000000	0.734742	0.284462	-0.002402	-0.0166
beds	0.051015	0.025070	0.765258	0.591175	0.734742	1.000000	0.254719	0.004530	-0.0187
price	0.033270	0.003788	0.311013	0.251823	0.284462	0.254719	1.000000	0.051284	0.0151
extra_people	-0.044282	-0.053384	0.003079	0.014916	-0.002402	0.004530	0.051284	1.000000	-0.0071
minimum_nights	0.018131	-0.014929	-0.020664	-0.015782	-0.016618	-0.018796	0.015128	-0.007114	1.0000
availability_30	0.105321	0.091113	0.126238	0.097904	0.111056	0.110105	0.055950	-0.035110	-0.0026
number_of_reviews	-0.329329	-0.178693	-0.067306	-0.063402	-0.070665	-0.065542	-0.053355	0.082336	-0.0126
review_scores_rating	-0.172519	-0.028412	0.007267	0.033689	0.024120	0.003685	0.003066	0.024613	0.0066
review_scores_accuracy	-0.162267	-0.036105	-0.000090	0.025480	0.014549	-0.009938	0.008932	0.018597	0.0051
review_scores_cleanliness	-0.170150	-0.022922	-0.006399	0.018495	0.008862	-0.017019	-0.005961	0.025023	0.0067
review_scores_communication	-0.128926	-0.039010	-0.001241	0.013613	0.011674	-0.011663	-0.000143	0.025052	0.0115
review_scores_location	-0.130083	-0.041423	-0.016454	-0.001040	-0.013055	-0.017945	0.006889	0.038630	0.0003
review_scores_value	-0.167331	-0.029543	-0.019741	0.014076	-0.000658	-0.014128	-0.004666	0.028410	0.0142
reviews_per_month	-0.249218	-0.009494	-0.077175	-0.050067	-0.088918	-0.083406	-0.053476	0.016168	-0.0256

Table 6.7.1 Listing of correlations between attributes

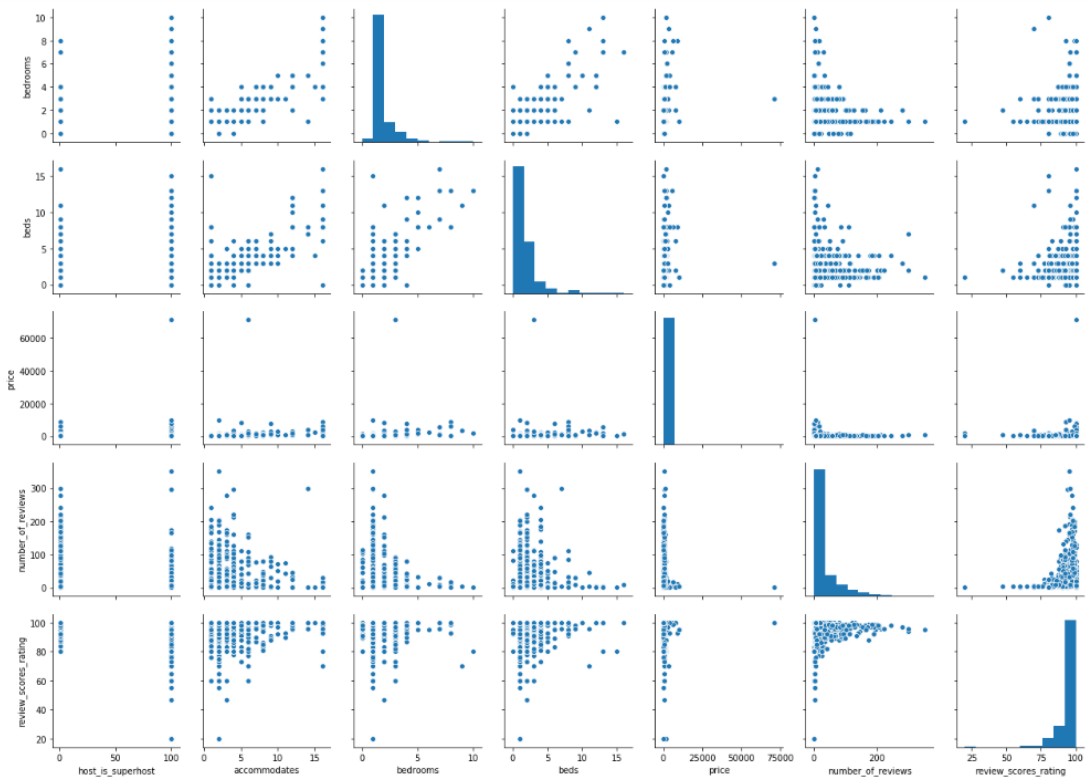

Table 6.7.2 Diagram of correlations between attributes

Table 6.7.1 and Table 6.7.2 are partial diagrams because there are too many attributes and there is limited space to show. According to these diagrams, we can see that there are some clear positive correlations between some attributes. For instance, the more apartments that can accommodate, the more bathrooms, bedrooms, and beds there are, and the higher the price. And these correlations are reasonable to understand.

8. Important Factors to Become a Superhost

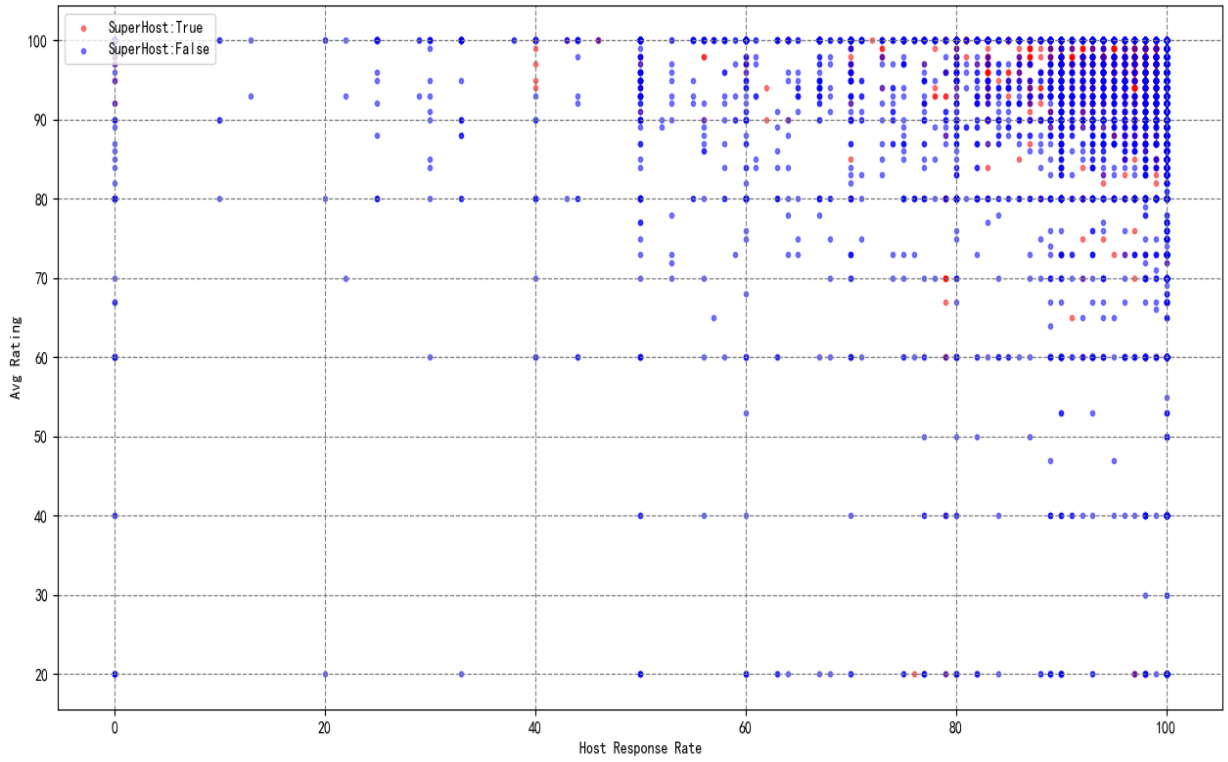

Table 6.8.1 Average Rating by Response Rate

From table 6.8.1, we can see that most super hosts have higher response rates and average ratings than ordinary hosts, and the positive correlation of their average ratings is even higher than their response rate. Thus, we know the importance of these two factors for becoming a super host.

9. Decision Tree

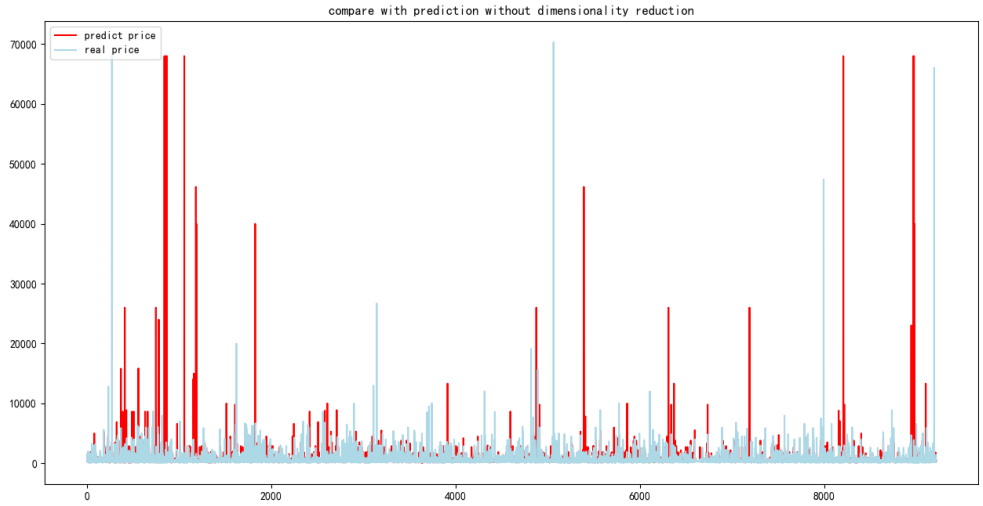

Figure 6.9.1 Diagram of predict price and real price before dimensionality reduction

This graph uses the sample serial number and house price as the x-axis and y-axis, respectively. And the red line is the predicted price, the blue line is the real price. The following types of diagrams are also in the same form.

Mean Absolute Deviation before dimensionality reduction: 474

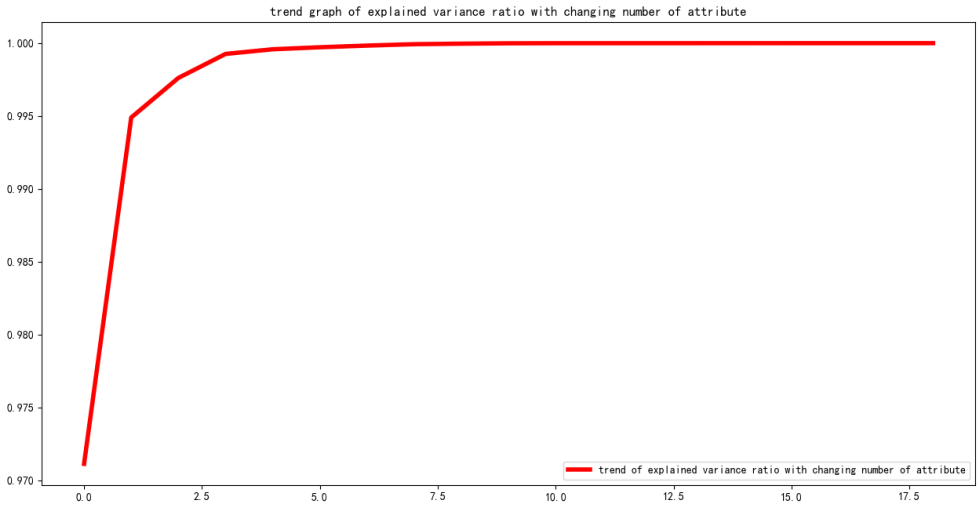

Figure 6.9.2 Trend graph of explained variance ratio with changing number of attributes

This graph uses the number of dimensions and the total value of '*explained_variance_ratio*' as the x-axis and y-axis, respectively.

Since this value is larger the better, and we found that when the dimension is reduced to almost 8 dimensions, this value no longer rises, so we choose 8 as the best dimension after dimension reduction.

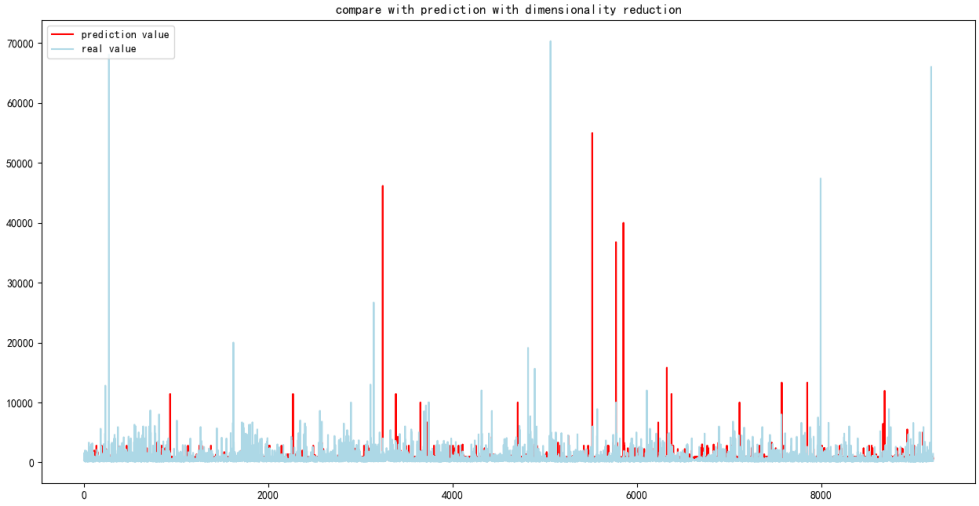

Figure 6.9.3 Diagram of predict price and real price after dimensionality reduction

Mean Absolute Deviation after dimensionality reduction: 541

The specific values of Mean Absolute Deviation of two decision trees:

```
Mean Absolute Deviation without dimensionality reduction:  474.31233846110314
Mean Absolute Deviation with dimensionality reduction:  541.0557187501753
```

10. Clustering

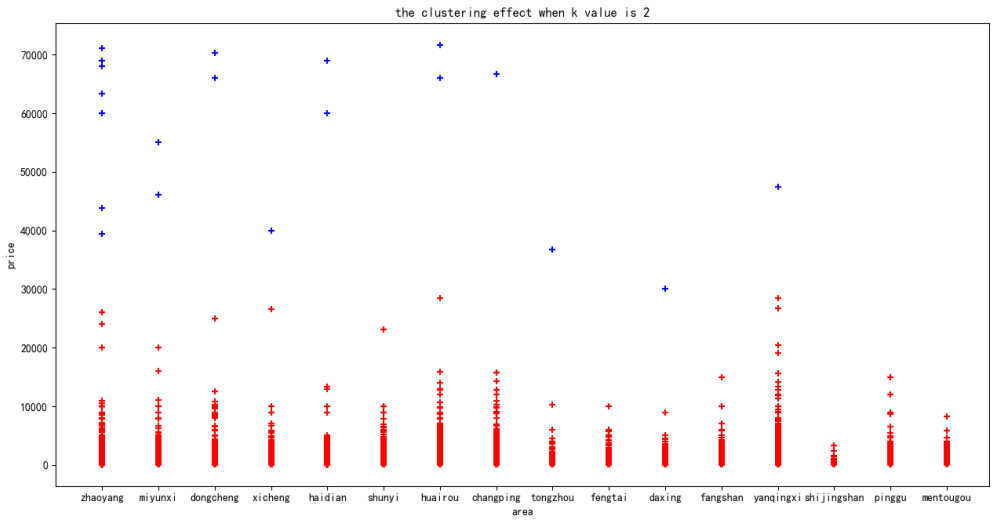

Figure 6.10.1 Graph of the clustering effect when k value is 2

This graph uses the regions and prices as the x-axis and y-axis, respectively. And '+' sign of the same color represents the same cluster — the same for the following pictures of the same type.

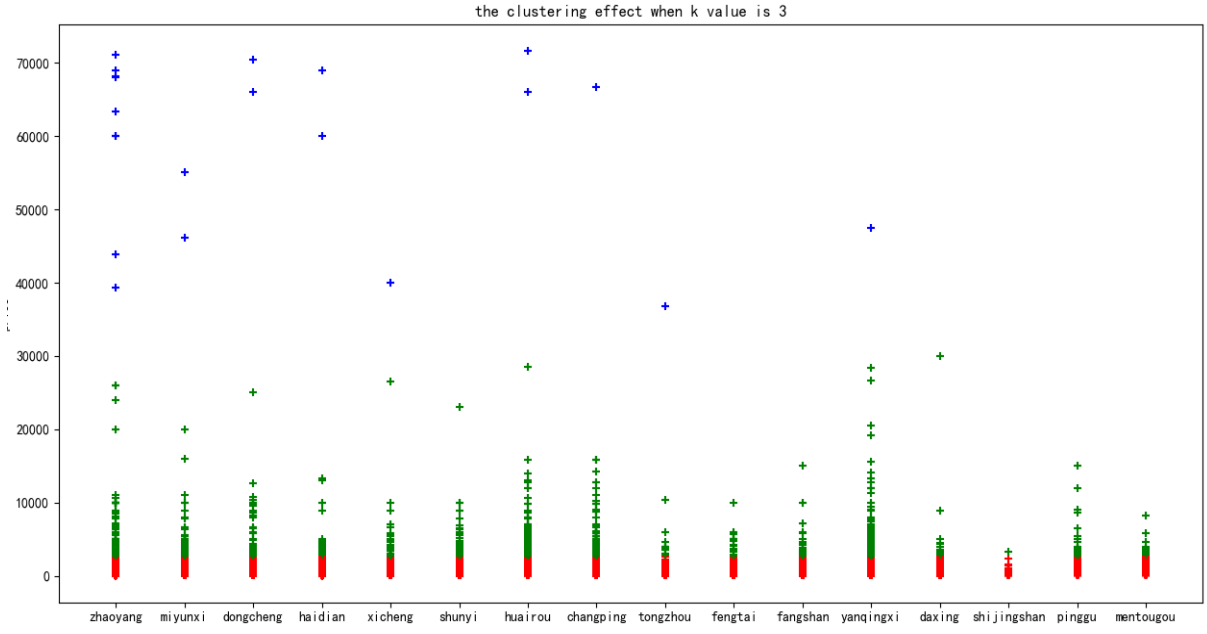

Figure 6.10.2 Graph of the clustering effect when k value is 3

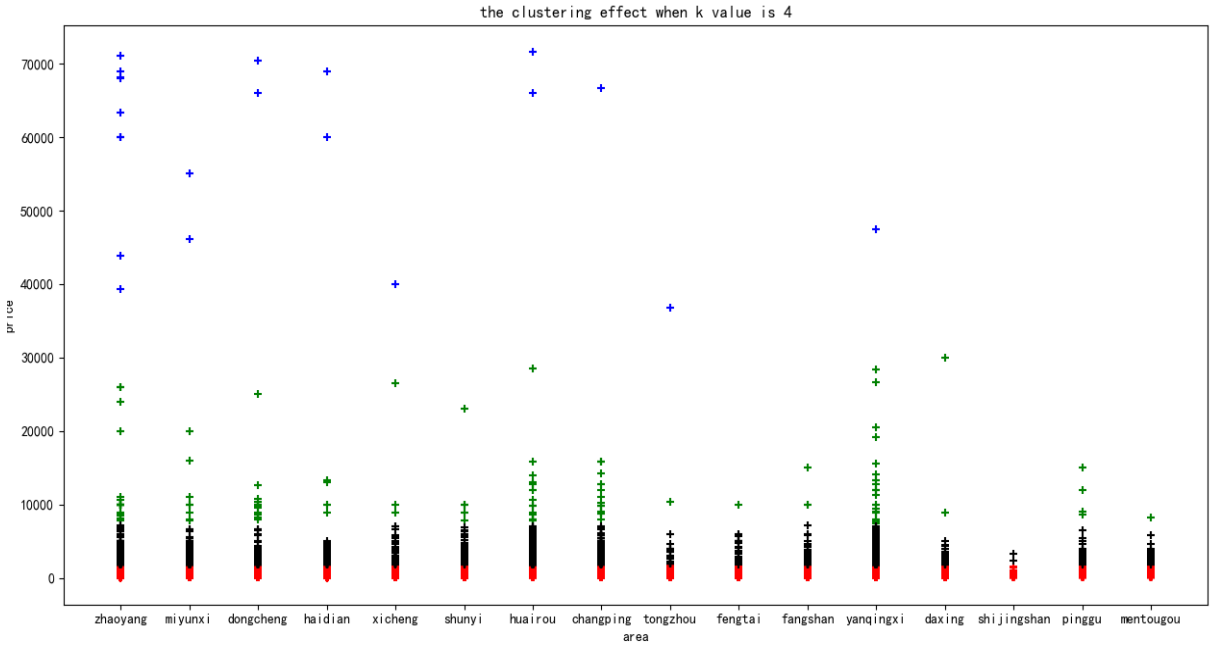

Figure 6.10.3 Graph of the clustering effect when k value is 4

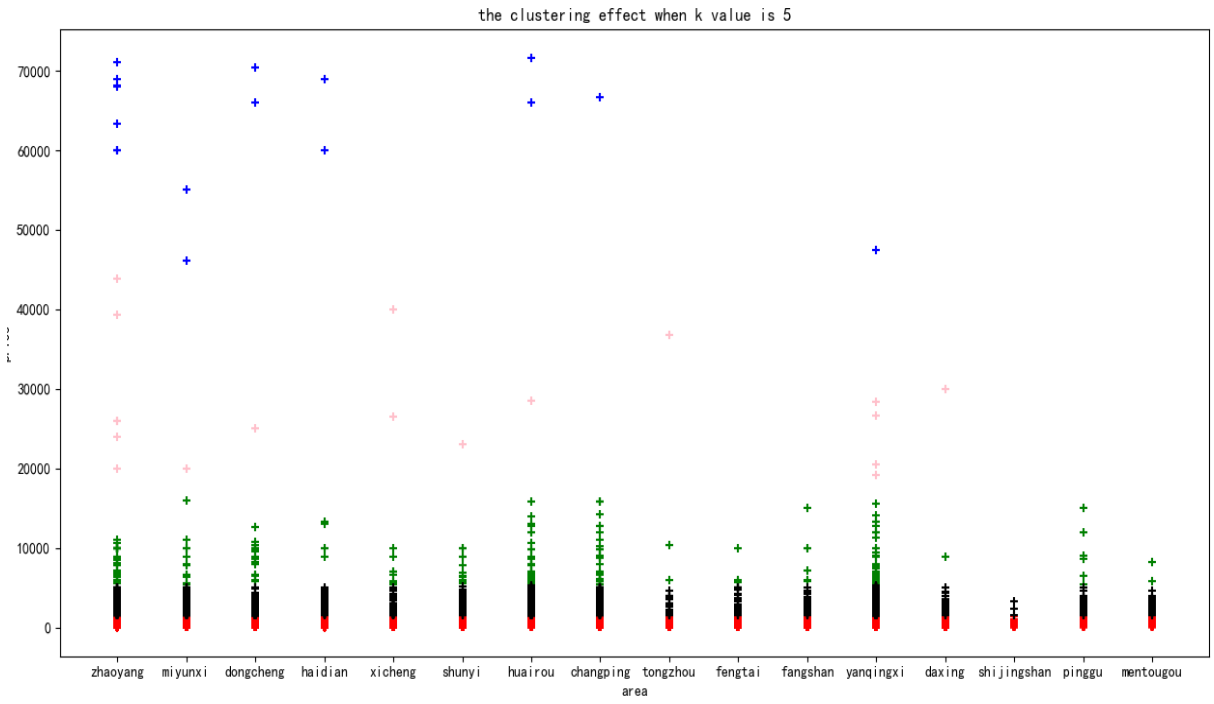

Figure 6.10.4 Graph of the clustering effect when k value is 5

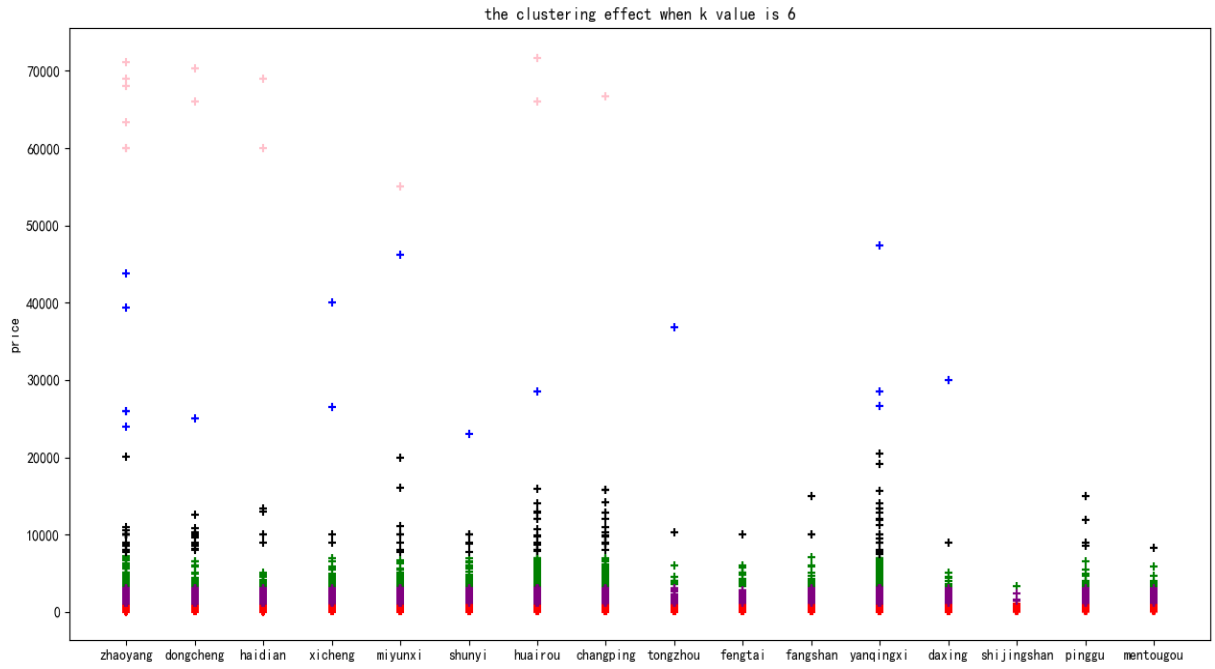

Figure 6.10.5 Graph of the clustering effect when k value is 6

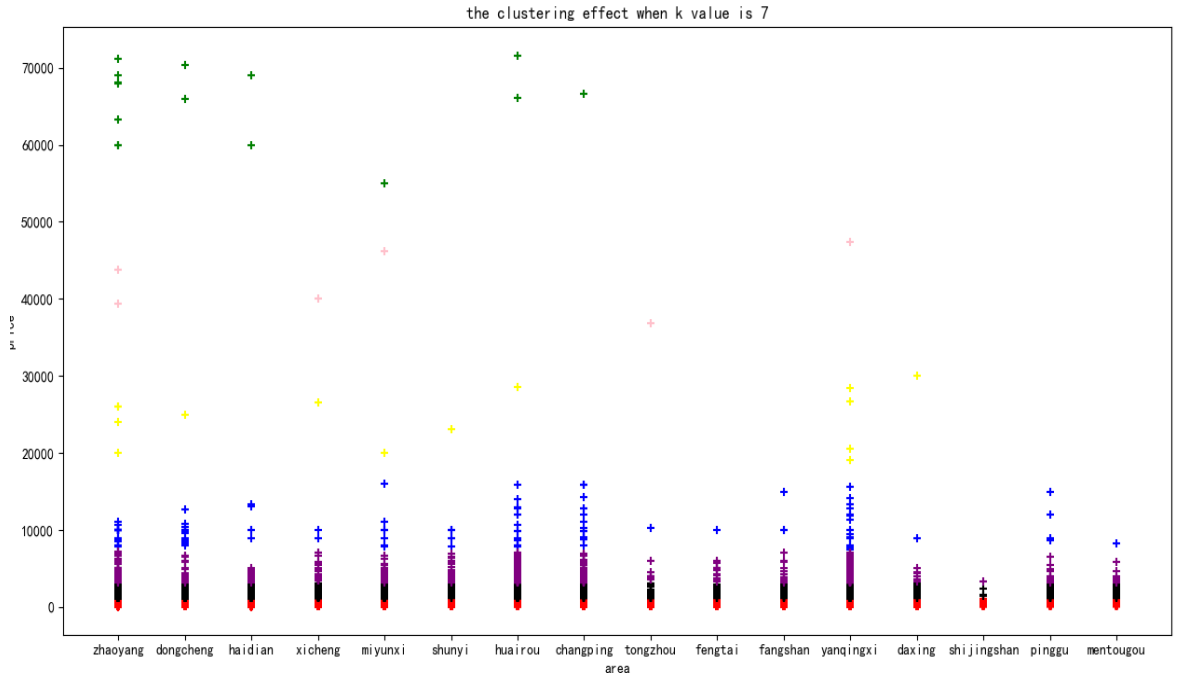

Figure 6.10.6 Graph of the clustering effect when k value is 7

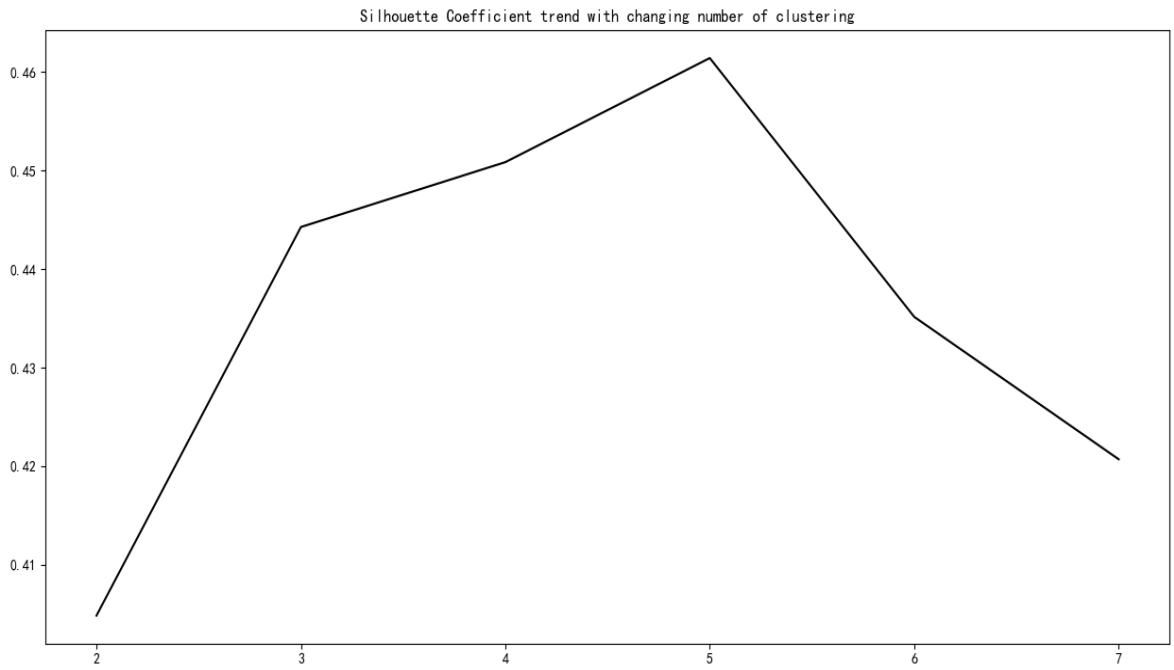

Figure 6.10.7 Graph of Silhouette Coefficient trend with changing the number of clustering

This graph use clustering numbers and Silhouette Coefficient value as the x-axis and y-axis respectively.

When the number of clusters is 5, the Silhouette Coefficient value is max. The clustering effect is the best in this case.

Next, we use the number of clusters to be 5 and redraw the graph of the best clustering effect.

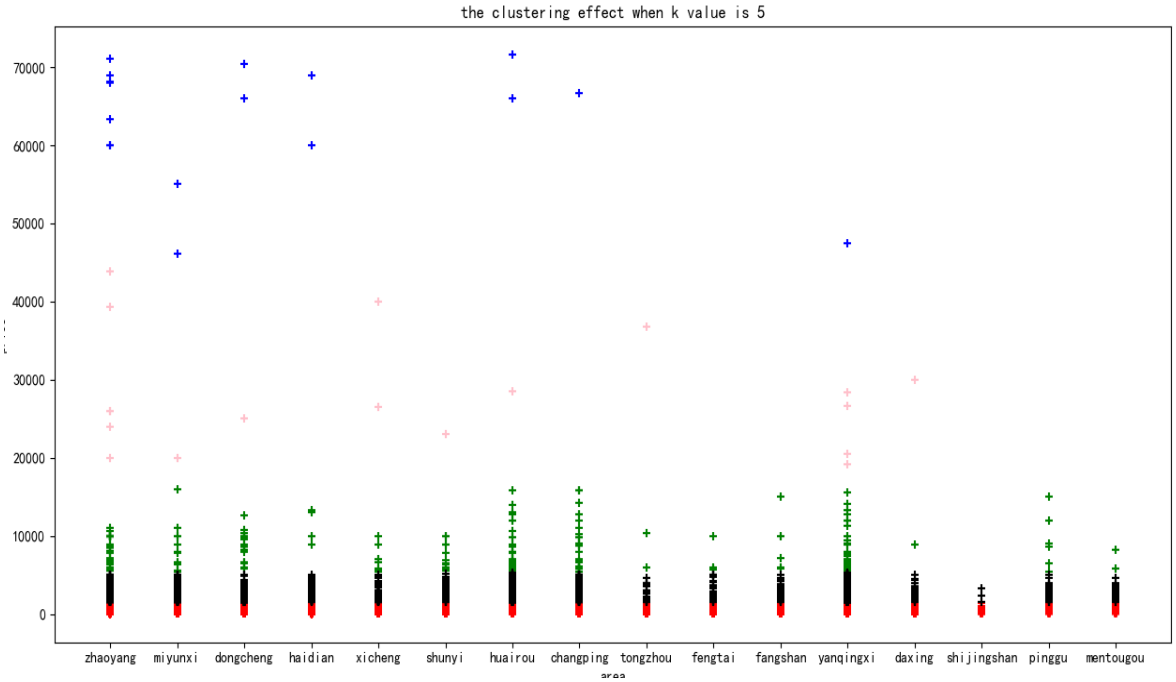

Figure 6.10.8 Graph of the clustering effect when k value is 5

7. Conclusion

1. Answers to Previous Questions

- **What are the different types of properties in Beijing? Do they vary by the district?**

  There are over 30 different types of properties in Beijing. Chaoyang District has the most apartments, hotels, lofts and houses. Dongcheng District has the most condominium. Huairou District has the most other types.

- **Which area is highly rated by renters?**

  Huairou District seems to be the highest but we do not recommend it because it is in the suburb area and has only 1000 listings compared to other urban districts.

- **What is the average cost of listings in each district?**

  The average cost varies from 433 CNY to 1,681 CNY. Most of the urban districts have an average price of around 500 CNY per night which is an affordable price.

- **What is the best time to visit Beijing?**

  Since most of the reviews are left during the summer break, we can conclude that June, July and August are good for visiting.

- **How many days/weeks/months should I make the reservation in advance?**

  Starting from the end of September, since the occupancy rate is over 90% until December, one should better make a reservation for two months in advance.

- **What do renters care about when searching rooms?**

  English speaking renters care about the reservation, cancellation policy, transportation and location.

  Chinese speaking renters care about the prices, transportation, host and cleanness.

- **What aspects affect the price? Which one is the most important?**

From the correlations diagram between attributes, we know that there is a large positive correlation between apartment prices and apartment capacity, the number of bathrooms, the number of bedrooms, and the number of beds. The most important of these is the capacity of the apartment.

- **How to predict apartment prices based on known information?**
  According to the effect of our decision tree, although the predicted prices of some samples are close to the real prices, there are also many samples with inaccurate prediction results. Despite the use of dimensionality reduction to process the data, the prediction effect has not improved. We expect that using boosting may lead to better results.
- **How to group existing apartments based on available information, and how many groups are most suitable?**
  Based on the effect of clustering, we found that, based on some known room properties, we were able to divide apartments into many different numbers of groups. And from the scatter plot results, their groupings have a great positive correlation with their prices. And we found that when divided into five groups, the division effect is the best.

2. Limitations

1. We cannot break Beijing into detailed areas due to over half of the listings missed the detailed information about the neighborhood. The smallest unit we can use for calculating average prices and rating scores is district which is too large compared to the NYU report. This problem caused some inaccurate results.

2. Some hosts' claim that their listings are unavailable in the listing names but still make them available on the list. We cannot detect if a listing is currently inaccessible by reading its name.

8. Future Work

1. Identify the neighborhood name based on the given longitude and latitude.
2. Detect the feature of listings by reading the textual information in the names of the listing.
3. The effect of using decision trees to predict house prices is not very satisfying. We look forward to improving some attributes using more detailed feature extraction and adopting boosting in the future.
4. For the effect of clustering, we look forward to showing it from other key attributes in the future and to build a three-dimensional drawing of some key dimension attributes after dimensionality reduction.

Reference

- Gupta, S. (2019, January 5). Airbnb Rental Listings Dataset Mining. Retrieved from https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec.
- A Python Echarts Plotting Library. (n.d.). Retrieved from https://pyecharts.org/#/.
- Python Data Analysis Library¶. (n.d.). Retrieved from https://pandas.pydata.org/.