Coursera Capstone
IBM Applied Project Report

How to Select Your Next House in Beijing

By: Caining Jin

Date: Aug-2020

1: Introduction

This IBM coursera data science project is about how to choose a house in the beautiful Beijing wisdom. With the development of contemporary society, every country has a very unique landscape, and Beijing (the capital of China) is no exception. Throughout the ages, it has been beautiful. Just like in the cover photo, each house in Beijing has its own unique characteristics. The classic design does not lose sight of the stunning appearance, and the internal structure is even more colorful. I chose this theme for this project mainly because I am preparing a house for marriage. Fortunately, through this project to establish a data model, so as to achieve my ideal analysis of the geographical location of different types of houses will be more helpful for the future selection of house types and the location of houses in Beijing.

1.1: Business Problem:

Beijing's housing prices have been rising very sharply for more than a decade, and it's crazy. Many people have seized this opportunity many years ago and made a lot of correct housing price predictions, which led to several years of profitability. But in 2018, Beijing's Suddenly, housing prices gradually stabilized and even the price per square meter in some districts and counties was slightly lower than before. The emergence of this situation has affected the way many people choose houses and brought more uncertainty. Therefore, I am planning to collect past house prices, house type data and house address information to build a model so that I can get through the model. The incoming information predicts the optional house structure and geographic location within the scope of their capabilities.

The location is especially important to the house. Whether there is convenient transportation around (subway) is very important for commuting to work, and whether you can walk a short distance to the shopping center on weekends, which can greatly affect the price of the house.

The neighborhoods around these houses can be referred to as their supporting facilities. When buying a house, the sales staff will try their best to let you know the buildings and supporting facilities around the house to achieve their goals. Many of them are deceptive. In this project, we can roughly understand the facilities around the house and the difference in housing prices caused by each type of supporting facilities.

1.2: Target Audience:

The ideal target for this report is individuals who are preparing to buy a house, and any customers who are on the sidelines who want to understand the trend of affected housing prices. Of course, it can also be the staff of some house sellers. They can use this method to understand the factors that may affect the house price and get a preliminary understanding of whether the surrounding popular supporting facilities are complete. By establishing different models and analyzing the results, choosing the best housing price prediction model can help us make rational decisions.

2: Data

2.1: For solving this problem, what data we need

In order to establish a housing price prediction model, analyze housing price trend models, and understand the popular supporting facilities around the house, we need to collect past building sales records and data on each type of house, including house details such as decoration style. If you want to know the location of the house, the latitude and longitude of each house is very important. Through the latitude and longitude, we can obtain the surrounding infrastructure and popular location data from the Foursquare website to achieve our purpose.

2.2: Where does the data come from

I collect data through the Internet. Kaggle is a very good website. There are thousands of data lovers who provide high-quality data. In the past, the first thing to collect is whether the kaggle website is feasible or not. related data. https://www.kaggle.com/ruiqurm/lianjia. This website is my data source. This data is very large and contains the information I need. E.g:

About this file

- url: the url which fetches the data
- id: the id of transaction
- Lng: and Lat coordinates, using the BD09 protocol.
- Cid: community id
- tradeTime: the time of transaction
- DOM: active days on market.Know more in https://en.wikipedia.org/wiki/Days_on_market
- followers: the number of people follow the transaction.
- totalPrice: the total price
- price: the average price by square
- square: the square of house
- livingRoom: the number of living room
- drawingRoom: the number of drawing room
- kitchen: the number of kitchen
- bathroom the number of bathroom
- floor: the height of the house. I will turn the Chinese characters to English in the next version.
- buildingType: including tower(1), bungalow(2), combination of plate and tower(3), plate(4).
- constructionTime: the time of construction

- renovationCondition: including other(1), rough(2),Simplicity(3), hardcover(4)
- buildingStructure: including unknow(1), mixed(2), brick and wood(3), brick and concrete(4), steel(5) and steel-concrete composite (6).
- ladderRatio: the proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident have on average.
- elevator have (1) or not have elevator(0)
- fiveYearsProperty: if the owner have the property for less than 5 years,

These are the contents contained in this data file. We need to clean the data afterwards to keep only the information that is needed for us. The data all come from Lianjia.com, which is an intermediary platform for house buying and selling. A large part of the house in China will provide information through this company to establish a connection between the buyer and the seller. This process has led to the company being able to obtain very large and credible house data transfer and entry. This data contains various types of information about house transfers from 2011 to 2017.

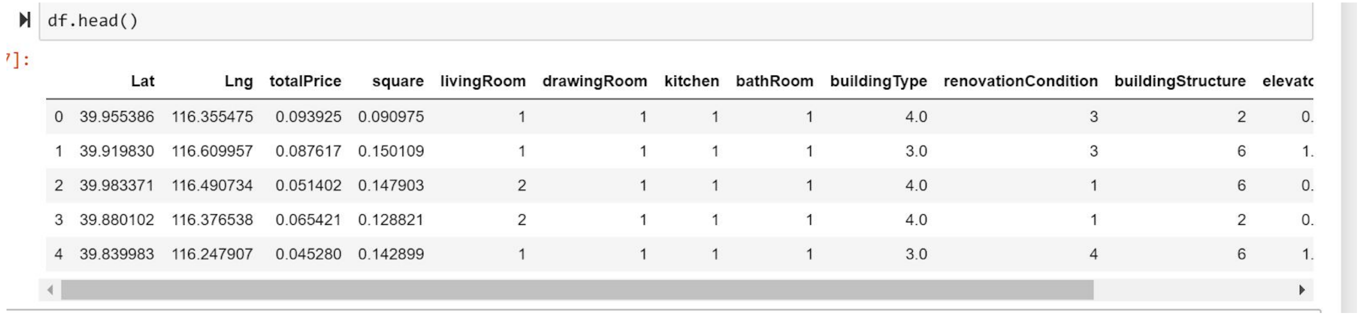
3. Methodology

- After analyzing the business problem level and preliminary self-prediction. Choose to use the popular and simple way pandas package to read the data and store it in the file we created. Put him in the most suitable data frame to start data preprocessing, delete tedious and invalid information such as ladderRatio: the proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident has on average. Although floor information will also affect the price of the house, this column of data is entered in Chinese, which is not friendly to data preprocessing. After consideration, this column of data is selected to be filtered out. After that, the step of removing empty datasets was performed. After reaching the ideal data preprocessing result, we put it into a clean dataframe to wait for the next analysis.

- Because the data set is too large and the calculation time is too long, I randomly select 5000 samples among them. Because housing prices and house size are our most concern, we normalize these two elements. We perform a series of correlation analysis on the dataframe processed by this normalization, including the establishment of boxplot, regression plot, Heatmap, etc., to obtain the impact of each element on housing prices.
- Since we have a large amount of data, we can split the data set. The method used in this step is train, test, and split, and 30% of the total data is used as the test set. The reason why we do not choose to directly build the overall model is because this can greatly reduce out off sample errors.
- Through single linear regression, multiple linear regression, ridge regression analysis, and combined multiple, ridge regression together to establish a model for testing. The method of testing the model basically compares the established model prediction yhat method with the y test group, and sees the mean square error and R-square methods to measure the score status of the model. And we can get the coefficient of the model so that we can calculate the price of the ideal room structure by ourselves. For example, how many bedrooms, how many bathrooms, whether there is an elevator, whether it is close to a subway station, etc.
- After that, by mobilizing the foursquare API, through the latitude and longitude of each house, information about popular places around and their types are obtained. Sort these results and select the 10 most popular surrounding locations and store them in a suitable dataframe to merge with the previous data. Through unsupervised learning, K-mean clustering, the data is divided into three groups. Each group has its own characteristics.
- Add all sample data to the map through latitude and longitude, and folium package for intuitive browsing of results.

4. Results

Figure 1: The basic Dataframe after cleaning.



```
df.head()
```

']:

	Lat	Lng	totalPrice	square	livingRoom	drawingRoom	kitchen	bathRoom	buildingType	renovationCondition	buildingStructure	elevatc
0	39.955386	116.355475	0.093925	0.090975	1	1	1	1	4.0	3	2	0.
1	39.919830	116.609957	0.087617	0.150109	1	1	1	1	3.0	3	6	1.
2	39.983371	116.490734	0.051402	0.147903	2	1	1	1	4.0	1	6	0.
3	39.880102	116.376538	0.065421	0.128821	2	1	1	1	4.0	1	2	0.
4	39.839983	116.247907	0.045280	0.142899	1	1	1	1	3.0	4	6	1.

Figure 2: Some relationships between price and each elements

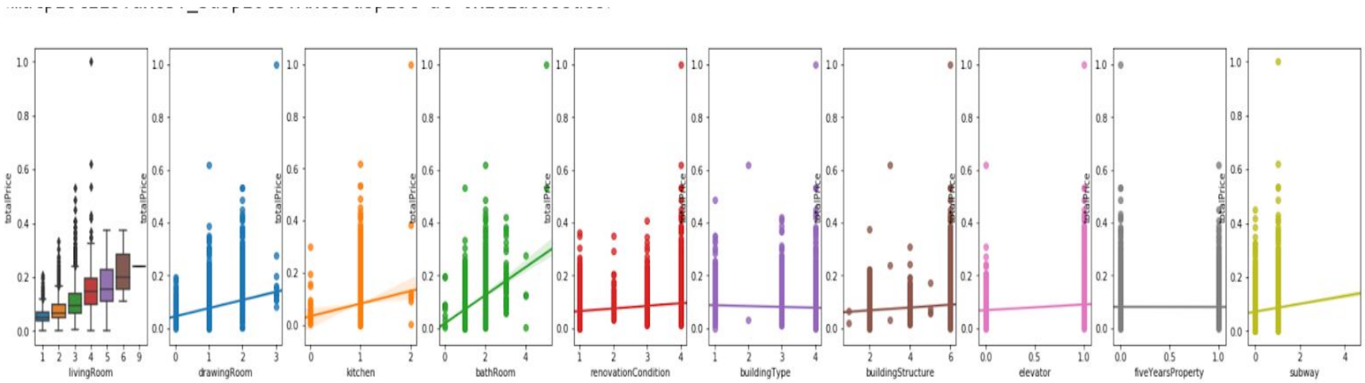
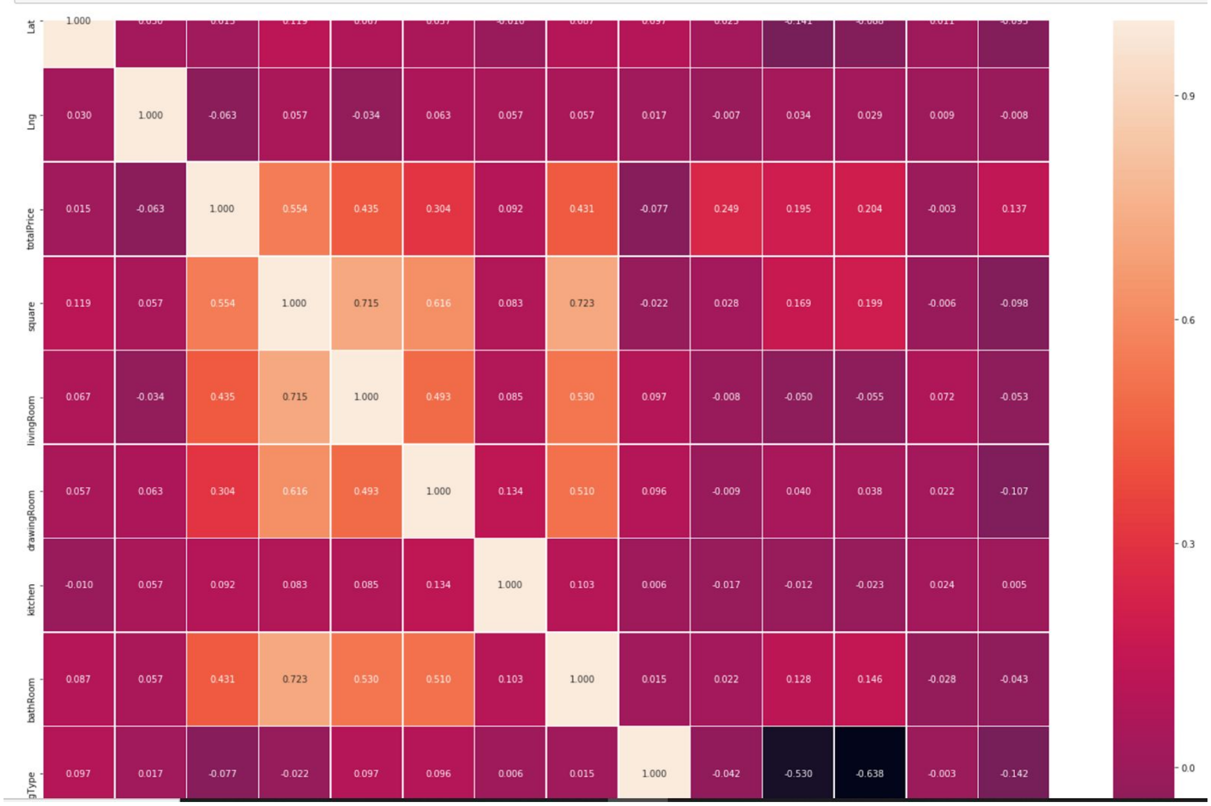
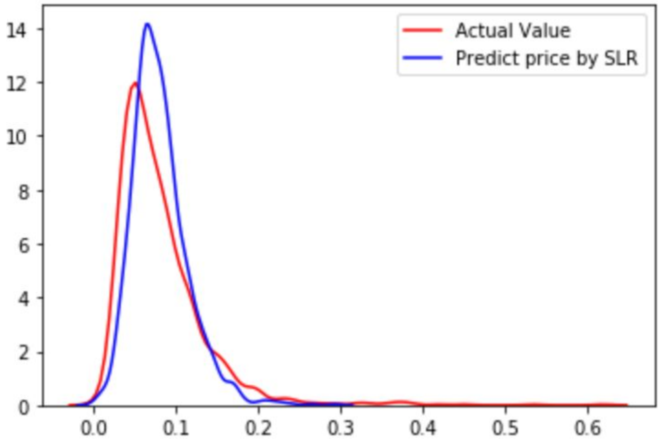


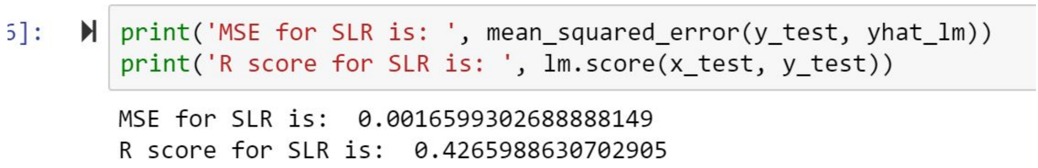
Figure 3: The correlation heatmap



Lat	1.000	0.030	0.015	0.119	0.067	0.057	-0.010	0.087	0.097	0.023	-0.141	-0.088	0.011	-0.093
Lng	0.030	1.000	-0.063	0.057	-0.034	0.063	0.057	0.057	0.017	-0.007	0.034	0.029	0.009	-0.008
totalPrice	0.015	-0.063	1.000	0.554	0.435	0.304	0.092	0.431	-0.077	0.249	0.195	0.204	-0.003	0.137
square	0.119	0.057	0.554	1.000	0.715	0.616	0.083	0.723	-0.022	0.028	0.169	0.199	-0.006	-0.098
livingRoom	0.067	-0.034	0.435	0.715	1.000	0.493	0.085	0.530	0.097	-0.008	-0.050	-0.055	0.072	-0.053
drawingRoom	0.057	0.063	0.304	0.616	0.493	1.000	0.134	0.510	0.096	-0.009	0.040	0.038	0.022	-0.107
kitchen	-0.010	0.057	0.092	0.083	0.085	0.134	1.000	0.103	0.006	-0.017	-0.012	-0.023	0.024	0.005
bathRoom	0.087	0.057	0.431	0.723	0.530	0.510	0.103	1.000	0.015	0.022	0.128	0.146	-0.028	-0.043
gType	0.097	0.017	-0.077	-0.022	0.097	0.096	0.006	0.015	1.000	-0.042	-0.530	-0.638	-0.003	-0.142

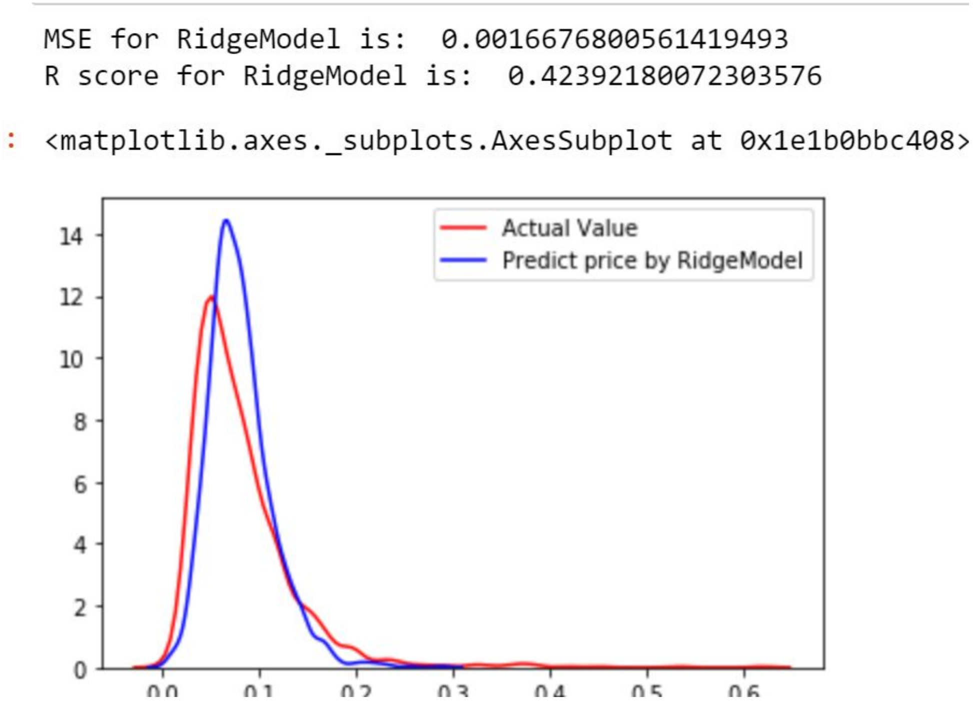
Figure 4: Simple linear regression model results





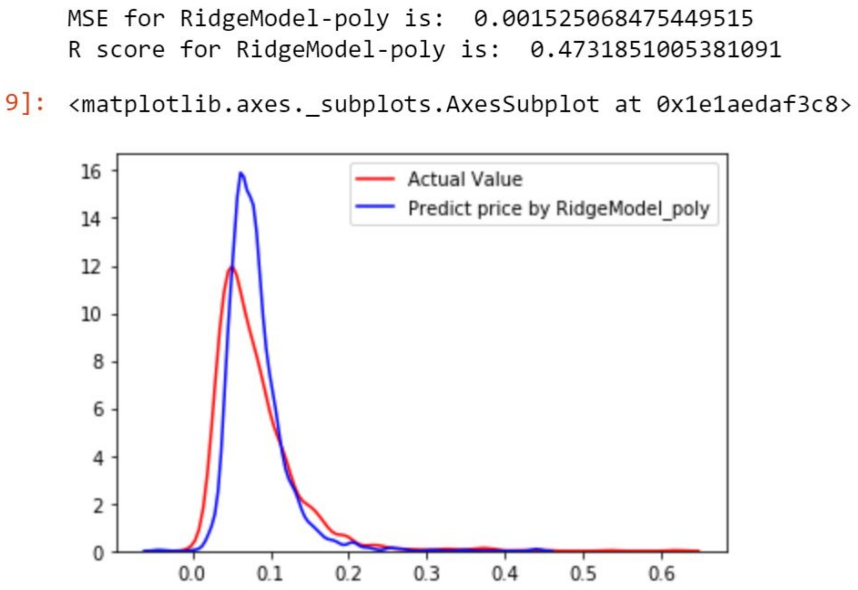
```
5]:  print('MSE for SLR is: ', mean_squared_error(y_test, yhat_lm))
     print('R score for SLR is: ', lm.score(x_test, y_test))

     MSE for SLR is:  0.0016599302688888149
     R score for SLR is:  0.4265988630702905
```

Figure 5: RidgeModel results



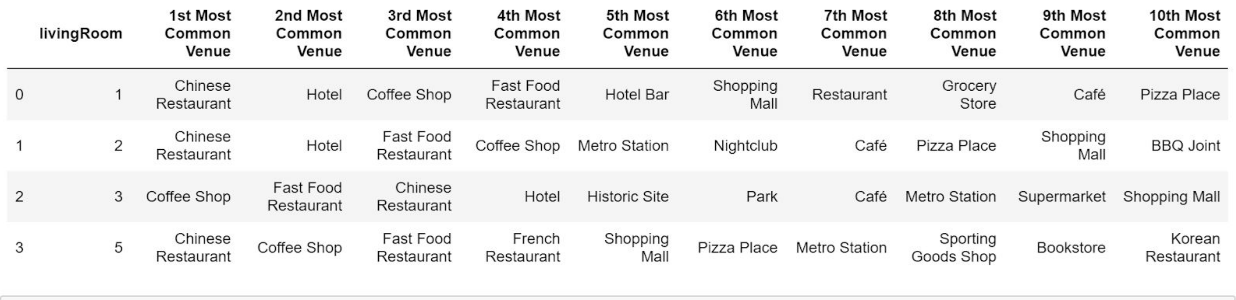
```
  MSE for RidgeModel is:  0.0016676800561419493
  R score for RidgeModel is:  0.42392180072303576

: <matplotlib.axes._subplots.AxesSubplot at 0x1e1b0bbc408>
```

Figure 6: Poly-with Ridge



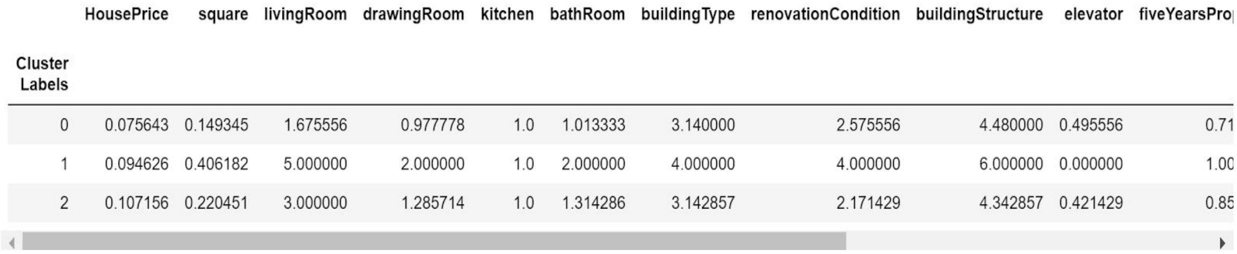
```
    MSE for RidgeModel-poly is:  0.001525068475449515
    R score for RidgeModel-poly is:  0.4731851005381091

9]: <matplotlib.axes._subplots.AxesSubplot at 0x1e1aedaf3c8>
```

Figure 7: Top 10 venue nearby each house



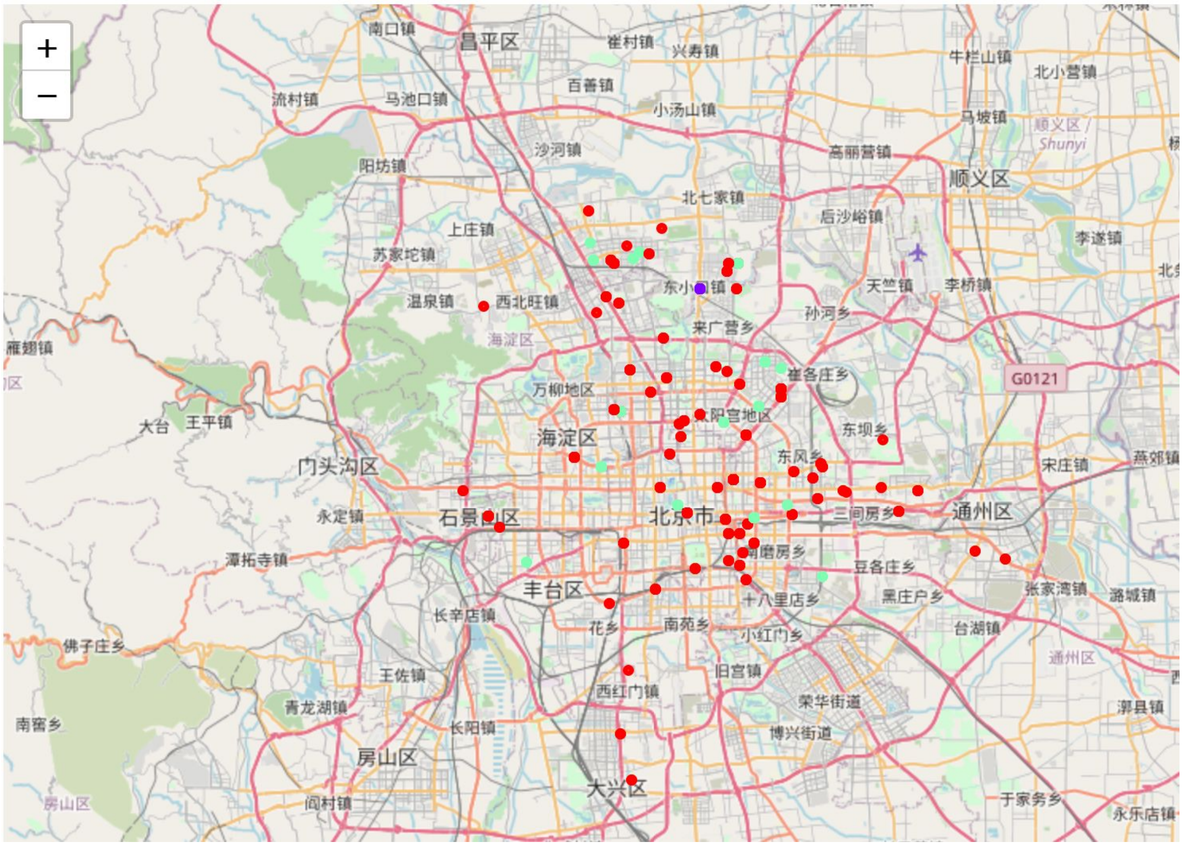
	livingRoom	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1	Chinese Restaurant	Hotel	Coffee Shop	Fast Food Restaurant	Hotel Bar	Shopping Mall	Restaurant	Grocery Store	Café	Pizza Place
1	2	Chinese Restaurant	Hotel	Fast Food Restaurant	Coffee Shop	Metro Station	Nightclub	Café	Pizza Place	Shopping Mall	BBQ Joint
2	3	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Hotel	Historic Site	Park	Café	Metro Station	Supermarket	Shopping Mall
3	5	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	French Restaurant	Shopping Mall	Pizza Place	Metro Station	Sporting Goods Shop	Bookstore	Korean Restaurant

Figure 8: To see the difference between each cluster of houses.



	HousePrice	square	livingRoom	drawingRoom	kitchen	bathRoom	buildingType	renovationCondition	buildingStructure	elevator	fiveYearsPro
Cluster Labels											
0	0.075643	0.149345	1.675556	0.977778	1.0	1.013333	3.140000	2.575556	4.480000	0.495556	0.71
1	0.094626	0.406182	5.000000	2.000000	1.0	2.000000	4.000000	4.000000	6.000000	0.000000	1.00
2	0.107156	0.220451	3.000000	1.285714	1.0	1.314286	3.142857	2.171429	4.342857	0.421429	0.85

Figure 9: The clustered map of each house location, with their price



5. Discussion

Through the observation of the results, we found that the housing prices in Chaoyang District, Haidian District, and areas near Tiananmen Square will have higher prices. By comparing the dense distribution of data, the closer to the center, the greater the transaction volume. The living room of a house has toilets. The more bedrooms there are, the larger the area and the higher the price. However, there is a general perception that houses with FiveYearsProperty are very popular in China, and the price will be slightly higher for the seller. However, by comparing the data, we find that this factor does not affect the trend of much house prices. Areas with subways around will have a beneficial effect on housing prices, because the daily traffic volume in Beijing is too large, and many people are unwilling to drive by themselves, even if they have a car, they are willing to use the subway to reach their destination. This is very different from cities in some countries. Through the overall comparison of each type of house, it can be found that the house price of cluster 0 to cluster 1 to cluster 2 is on the rise, that is, cluster 2 represents the more expensive house type. Cluster 0 is mostly located in the city center. It is suitable for young people working hard. The economic benefits are better, and it is cheaper than the other two. The transaction volume is also the largest. If it is young people in the early stage of business, it should be the most suitable room type. Not only is the price of Cluster 1 not as expensive as the price of Cluster 2, but the size of the units is the largest on average, and the building structure and decoration style of the houses are the best, but the bad thing is that cluster 1 basically has no elevators and no The subway is around and travel is not as convenient as the other two types of houses. It can also be seen from the map that this type of house is basically located in the suburbs of Beijing and there are few residential areas around it. Compared with the other two types of houses, the popular supporting facilities around are more suitable for the elderly to live in a quiet area. The area is large and the price is moderate. The transportation is not so convenient. The most expensive and most popular place around Cluster 2 is nightclub, which is suitable for some wealthy young people or successful people to enjoy Beijing's nightlife, and nightclubs do not frequently appear around the other two types of houses, which also reflects that this type of house is not suitable for elderly people. It may be a bit noisy, but it is suitable for young people with rich families.

6. Conclusion

After the practice of this project, I have a general understanding of Beijing housing price trends, and through the analysis of the results obtained by K-mean clustering, I have obtained some information about the approximate location of the apartment type I want to buy and the surrounding supporting environment. I also hope that this project can provide the most basic assistance to those who want to buy the next house, so that everyone can make smarter decisions.

Of course, there will be some shortcomings in the project. I just started to contact this field and hope to continue to improve and expand my knowledge through many exercises.