Beijing Home Cluster

Applied Data Science Capstone Project

Tianxuan Chen

1. Introduction

In this project. I will choose real estate of Beijing to study as my research target to study. As one of the metropolises in the world, has 16 districts and covers around 6336 square miles. However, there are 21.54 million people lived in this city.[1] For now, the population density is 20,462,610.[2] Since 2008, China property price are increasing tremendously. Until today, the average price of house in Beijing has tripled.[3] Considering the average salary of Chinese, buying a house could put you in debt for life. For this reason, this project will analyze Beijing house price and its relevant factors to find out different classes of houses.

2. Data

2.1 Data Explanation

In addition to the homes price and areas, there are lots of other features that affect whether people choose a house or not. As mentioned before, Beijing is so big that the distance between home and the city center will be one of the major factors. Besides the distance, public transportation capacity affects the traffic circumstance directly, so I also need to check if there is a subway station near the home or not. In the other hand, for most Chinese families, they attach great importance to the education of next generation, so I have enough evidence to show that school-nearby homes have higher value in both living and investment. Last, but not the least, Beijing's medical level is the one of the best in China, more medical facilities around homes means better treatment condition for any kind of disease.

2.2 Data Collection

In order to find the most suitable houses for different types of people, I need to collect house data through authoritative website. In this project, I will use web crawler to collect data from

[1] https://simple.wikipedia.org/wiki/Beijing
[2] https://worldpopulationreview.com/world-cities/beijing-population/
[3] https://www.ceicdata.com/en/china/nbs-property-price-monthly/property-price-ytd-avg-beijing

LianJia[4] which is a housing trading platform in China.

In Lianjia, I am able to download the data of community name, price, area and whether it is new or second-hand. As for the other vital data, I need to use Amap APIs[5]. Amap is one of the map services companies in China, which provides a large number of map services and APIs. In this project, I will just the poi API to get these factors including house's price, house's area, the distance to city center, the number of hospitals, schools and subway stations within 1000 meters.

2.3 Data Cleaning

Data downloaded from Lianjia and Amap is combined into a dataframe. In order to proceed our analysis, I need to drop those missing data, non-numeric data, and out- of-range data. I simply use dropna function of pandas library to drop data. As for out-of-range data, I apply histogram to see how each type of data is distributed.

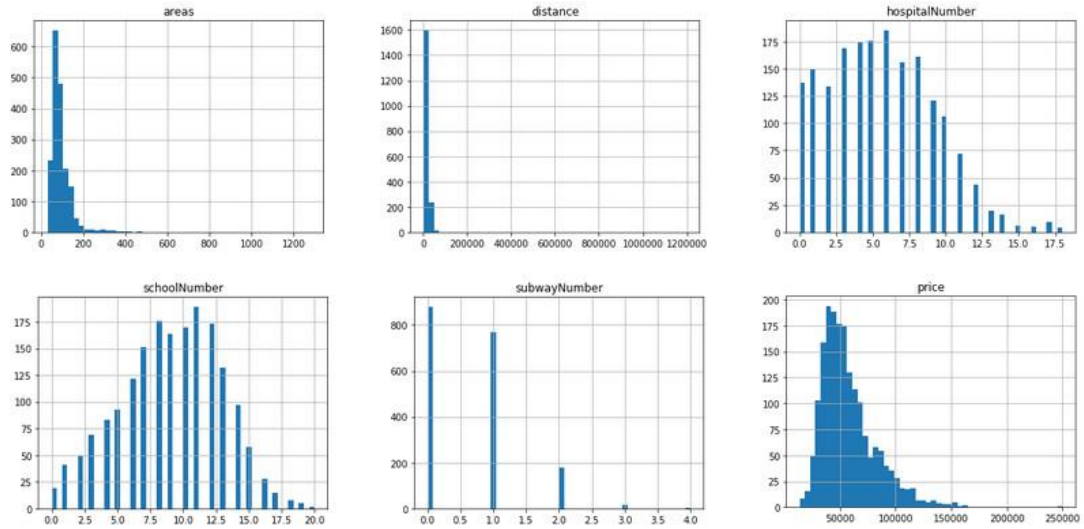

Figure 1

As shown in figure 1, the area and distance data are skewed to the right. Only a few houses have more than 600 square meters and are very far from the city center, which will be create deviation if they remain in our sample. I am only going to keep those houses which are less than 600 square meters and less than 20000 meters away from the city center.

3. Methodology

A dataframe that contains only key features such as *price, area, type, distance, number of schools, subway and hospital numbers and geographic location (latitude and longitude)* will be the main data set I deal with in the next steps.

[4] https://bj.fang.lianjia.com/
[5] https://lbs.amap.com/

	name	price	areas	type	latitude	longitude	schoolNumber	subwayNumber	hospitalNumber	distance
0	10AM新坐标	65492.0	32.28	0	39.868745	116.443373	10.0	1.0	9.0	5928.0
1	8哩岛	44455.0	140.21	0	39.936657	116.641781	3.0	0.0	0.0	21087.0
2	BOBO自由城	53524.5	102.48	0	39.908828	116.693801	8.0	1.0	3.0	25302.0
3	CBD传奇	71839.0	65.95	0	39.887624	116.477012	6.0	1.0	12.0	7185.0
4	CBD总部公寓二期	88126.0	99.29	0	39.899629	116.454618	9.0	0.0	11.0	4982.0

Figure 2: the first five rows of homes data

I firstly use the folium library to see how these houses are distributed in Beijing. Through passing latitude and longitude to folium circle marker function, I could display all houses in Beijing Map according to their latitude and longitude.

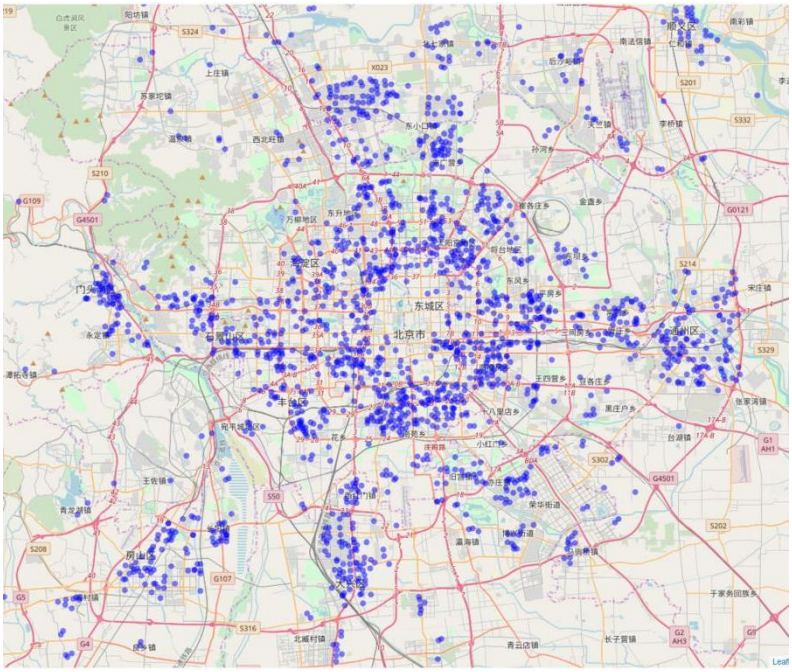

Figure 3: the distribution of real estate in Beijing

Now I have valid data set for 1839 houses. To to cluster these houses through using K-means method, I use normalization method which allow us to evaluate all factors in the same standard. Then I can see the correlation between all factors.

	price	areas	type	latitude	longitude	schoolNumber	subwayNumber	hospitalNumber	distance
price	1.000000	-0.378529	-0.131639	-0.511449	-0.509821	-0.037034	0.180023	0.041613	-0.921856
areas	-0.378529	1.000000	0.439261	0.491037	0.490385	0.062139	-0.028236	-0.000779	0.483352
type	-0.131639	0.439261	1.000000	-0.042376	-0.042669	-0.265164	-0.079769	-0.212601	0.165870
latitude	-0.511449	0.491037	-0.042376	1.000000	0.999932	0.441501	0.026192	0.174661	0.713762
longitude	-0.509821	0.490385	-0.042669	0.999932	1.000000	0.442402	0.027215	0.173633	0.711854
schoolNumber	-0.037034	0.062139	-0.265164	0.441501	0.442402	1.000000	0.110456	0.050047	0.153618
subwayNumber	0.180023	-0.028236	-0.079769	0.026192	0.027215	0.110456	1.000000	-0.040571	-0.153631
hospitalNumber	0.041613	-0.000779	-0.212601	0.174661	0.173633	0.050047	-0.040571	1.000000	-0.003492
distance	-0.921856	0.483352	0.165870	0.713762	0.711854	0.153618	-0.153631	-0.003492	1.000000

Figure 4: correlation

Thirdly, I also want to find optimal number of clusters. In this project, I used elbow method and silhouette score to evaluate. I used a loop function to calculate SSE and silhouette score from cluster 2 to 15. At Last, Matplotlib library could represented how these scores change in figures.

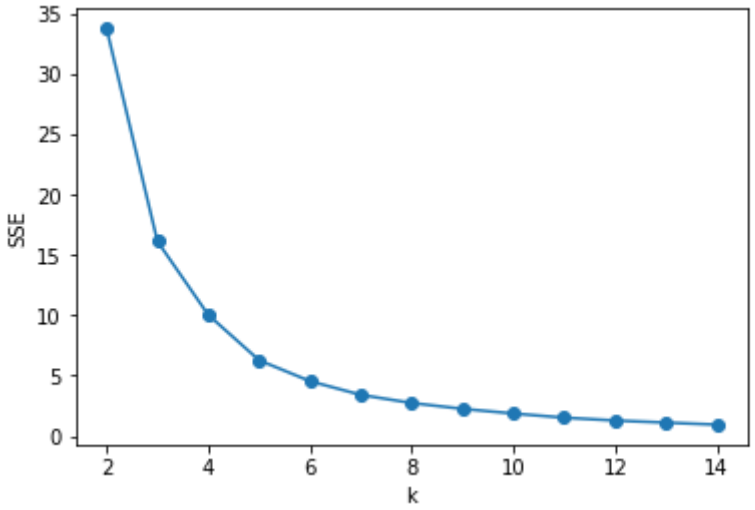

Figure 5: Elbow Method

Let's start with the elbow method. From Figure 5, when the number of clusters is 4 or 5, it is closest to the elbow. In order to find a more accurate number of clusters, let's look at the contour coefficient again. According to the definition of the silhouette coefficient, the higher the score, the better. Combined with the elbow method, it is best when the number of clusters is 5.

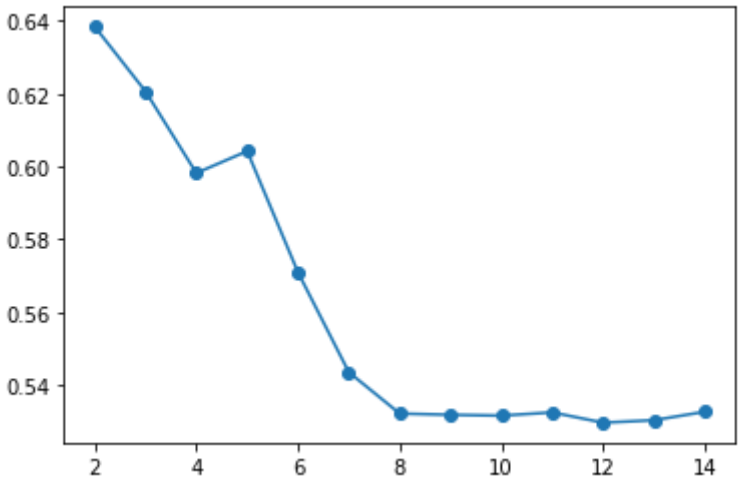

Figure 6: silhouette score

4. Result

Now, Beijing's houses can be divided into 5 clusters and I labeled them as 0 to 4. Also, I use

different colors for each cluster and show them on the map.

	name	price	areas	type	latitude	longitude	schoolNumber	subwayNumber	hospitalNumber	distance	Labels
0	10AM新坐标	65492.0	32.28	0	39.868745	116.443373	10.0	1.0	9.0	5928.0	0
1	8哩岛	44455.0	140.21	0	39.936657	116.641781	3.0	0.0	0.0	21087.0	2
2	BOBO自由城	53524.5	102.48	0	39.908828	116.693801	8.0	1.0	3.0	25302.0	2
3	CBD传奇	71839.0	65.95	0	39.887624	116.477012	6.0	1.0	12.0	7185.0	0
4	CBD总部公寓二期	88126.0	99.29	0	39.899629	116.454618	9.0	0.0	11.0	4982.0	0

Figure 7: the first 5 rows of labeled homes

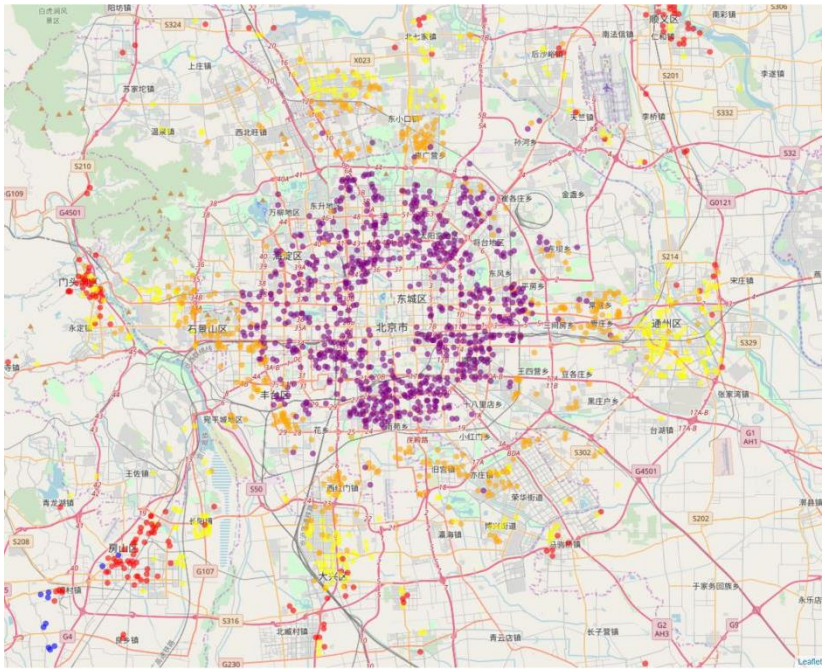

Figure 8: the distribution of real estate in Beijing after cluster analysis

Cluster 1: Label 0 with color purple, as I can see from above table, it owns highest amount of hospitals, subways and schools around and shortest distance from the city center. Of course, it is most expensive cluster. However, most houses are second-hand, which means it is not a good choice for those who want to buy a new house anyway. Also, since most old houses in such central location were built in a decade ago, it means most of them does not have large size.

Cluster 2: Label 1 with color red, it has relatively low price. This cluster is already being counted as far from the city center, in despite of the number of schools and hospitals looks good, far from city center also means the quality of schools and hospitals will not be as good as cluster 1. Moreover, the average amount of subway stations is only 0.21267, the traffic conditions cannot meet young people's commuting requirements.

Cluster 3: Label 2 with color yellow. This cluster stands for the normal commercial houses in Beijing. Most features in this cluster are median. Although the distance is further than Cluster 5, but their price are very similar.

Cluster 4: Label 3 with color blue. From the average distance, I can easily conclude that these houses are in the outskirts of Beijing, the distance to city center of this cluster is more further than other clusters. In addition, the average amount of subway stations is 0.214286. These houses are very inconvenience for people need to work at the city center. Certainly, these house' s price is lowest

Cluster 5: Label 4 with color orange. Due to the development of Beijing, most houses in these locations are new houses. These houses should be most suitable for ordinary people to buy.

5. Discussion

In this project, I used K-Means method to distinguish types of houses in Beijing. From the distribution map and previous analysis, I can conclude that house area and the distance to city center are negative correlated with houses' price and the number of hospitals, schools and subway stations is positive correlated with house price. The new or second-hand houses also affect the price but not as vital as other features.

To be specific, house price is varying. The maximum price of cluster 1 is 250,000 RMB per square meter while the average price is only 77,609 RMB per square meter. Although the houses in cluster 1 possess the best educational, medical resources in Beijing, some of them are too expensive. Cluster 5 (yellow points) and Cluster 3 (orange points) are very similar, most features are very close except for the distance to the city center and the price. In other words, the distance results the difference in price. With sufficient funding, Cluster 5 will be better choice.

As for the Cluster 2 (red points), these houses are in the edge of Beijing. Like I said before, Beijing is a metropolitan city. Considering working and computing costs, these house target customers would be those people who can work at home or working at these communities. For the last cluster (blue points), they are even further than Cluster 2. Only 42 houses are labeled in 4, they are either mansion or cheap houses.

6. Conclusion

This project gives me a basic understanding of clustering, I am pretty satisfied with the results. In fact, there are more factors need to be considered to cluster houses such as land agent company, loan interest rate, number of bedrooms and etc. Due to the limitation of my ability of data collection, I won't be able to implement all my thoughts this time. In the future, I would try to re-do this project by adding more factors and optimizing the model and parameters.

References：

1. https://simple.wikipedia.org/wiki/Beijing

2. https://www.ceicdata.com/en/china/nbs-property-price-monthly/property-price-ytd-avg-beijing
3. https://bj.lianjia.com/
4. https://lbs.amap.com/api/webservice/summary/
5. http://data.stats.gov.cn/easyquery.htm?cn=E0105&zb=A02&reg=110000&sj=2019
6. http://bj.cityhouse.cn/market/
7. https://worldpopulationreview.com/world-cities/beijing-population/