The Battle of the neighborhoods

1. Introduction Section :

1.1 Scenario and backgroud

I am a data scientist living in Shanghai. I have rent an apartment unit in Qingpu Area. However, after several months, I find it is very inconvenient as the apartment is far from subway station and there is no large supermarket within 3 kilometers. Furthermore, recently I was offered a new job in another area(Yangpu Area). So I would like to buy a new apartment which brings more convenience to my work and life.

1.2 Problem to be resolved:

The challenge to resolve is being able to find a second-hand apartment in Yangpu Area that is subject to the following conditions:

1. Apartment with min 2 bedrooms and min 80 square meters with price not to exceed 6,000,000RMB
2. Area with ammenties like subway metro station, supermarket and cinemas within walking distance(<=1.5 km)

1.3 Interested Audience

I believe this is a relevant project for a person who wants to find a better place to live and work, since the approach and methodologies used here are applicable in all cases. The use of Foursquare data, scraping other relevant data from website and mapping techniques combined with data analysis will help resolve the key questions arisen. Lastly, this project is a good practical case toward the development of Data Science skills.

2.Data Section:

2.1 Data Required to resolve the problem

In order to make a good choice of a good apartment in Shanghai, the following data is required:

1. Listed second-hand apartments in Yangpu Area with descriptions (price, location, how many bedrooms, area)
2. Geodata (latitude, longitude) of the second-hand apartments
3. Venues and ammenities in the neighborhoods

2.2 Data source¶

Following steps are used to get data.

1. Scrap second-hand apartment info from a website named Lianjia and filter data with first condition in Section 1.2.

2. Get geodata of the addresses of selected apartments with Baidu ap

3. Use Foursquare data and geodata to map venues

4. Merge selected apartments info with their geodata

5. Select apartment satisfying the second condition in section 1.
Here are some samples of data:

	Address	Price	Bedrooms	Area	Longitude	Latitude	metro station	Supermarket	Cinema	Price/m^2
0	阳明新城	558.0	3	89.79	121.527805	31.275399	7	7	5	6.214501
1	阳明新城	575.0	3	89.79	121.527805	31.275399	7	7	5	6.403831
2	阳明新城	568.0	3	89.79	121.527805	31.275399	7	7	5	6.325871
3	文化名园	570.0	3	92.70	121.498061	31.312913	6	9	5	6.148867
4	鞍山四村第二小区	550.0	2	87.90	121.514984	31.283729	10	6	7	6.257110

3. Methodology Section:

In this project we will direct our efforts on choosing an appropriate apartment in Shanghai Yangpu Area. The apartment meet requirements in the Introduction Section

In first step we have collected the required **data: price, number of bedrooms, area, location and venues** .

Second step we will explore the information of apartment, such as price, number of bedrooms and area to see if my constraint is a common requirement. In this part, we will also figure out the relationship between price and area. We will compare the situation of all apartments in Yangpu Area and the selected convenient apartments.

In third and final step we will locate selected convenient apartments in the map and we will use k-means clustering method to divide them into 3 categories by price, area, location. After that we will find features of each cluster. Finally, we will try to find if there's a **cost efficient** apartment.

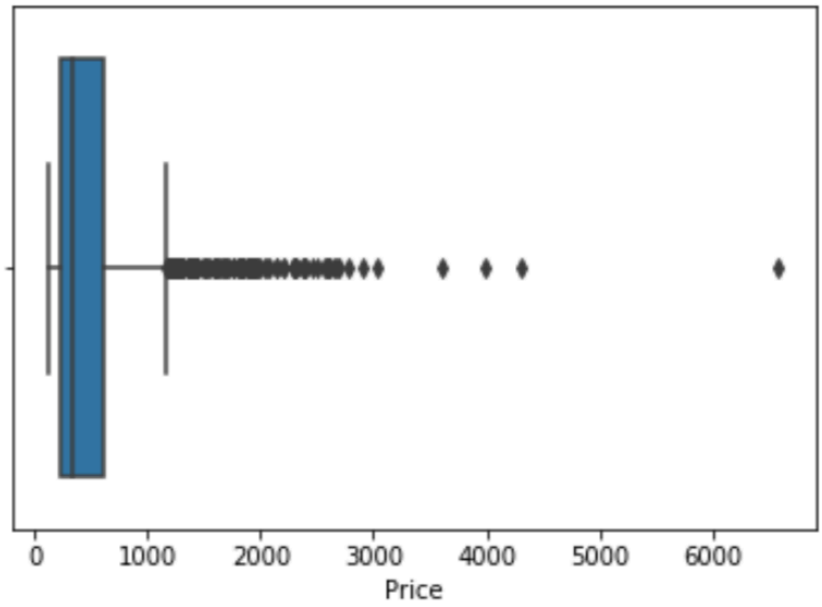
4. Results and Discussion:

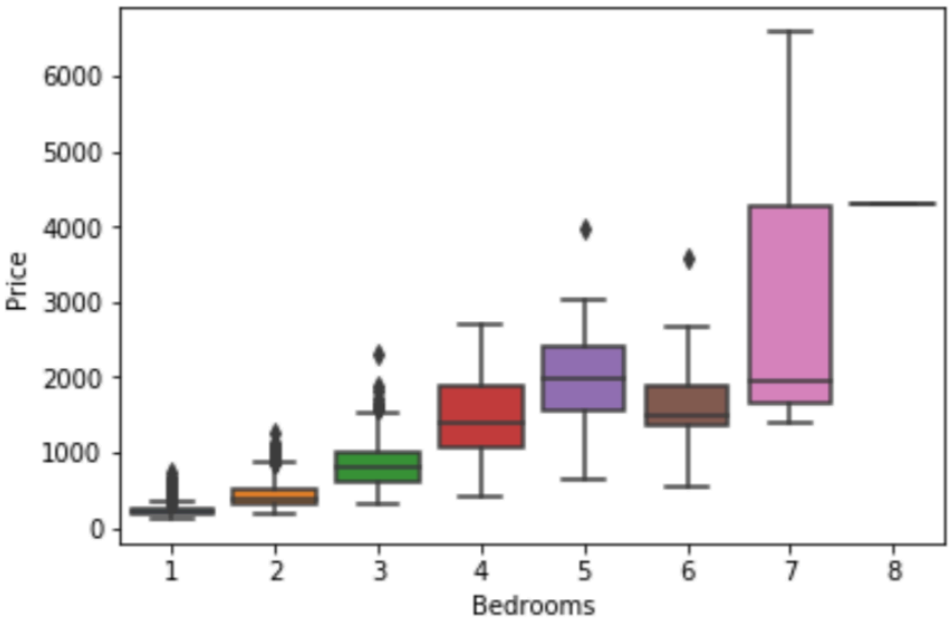
For all apartments in the area:

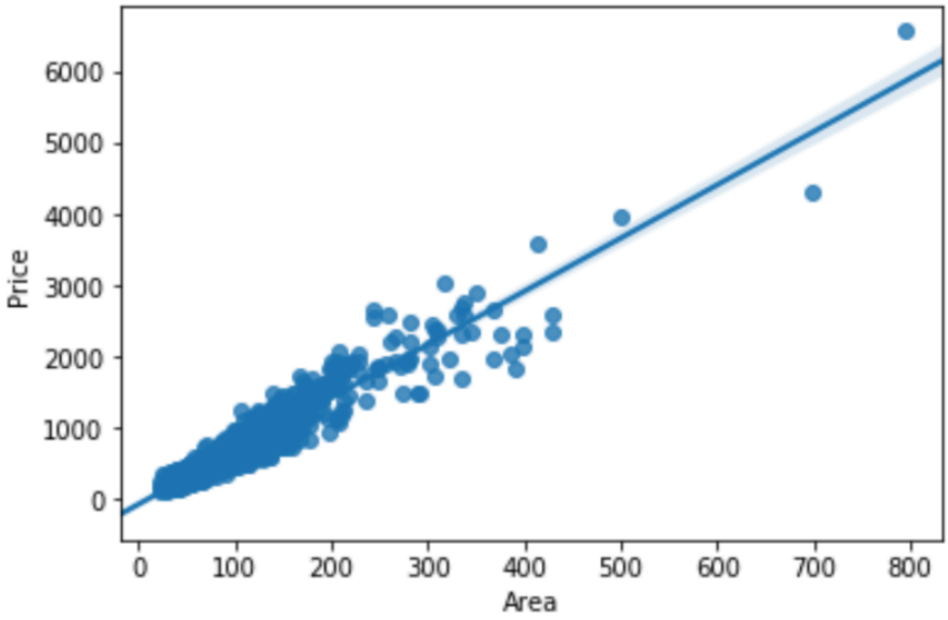
Here are the findings of all apartments:

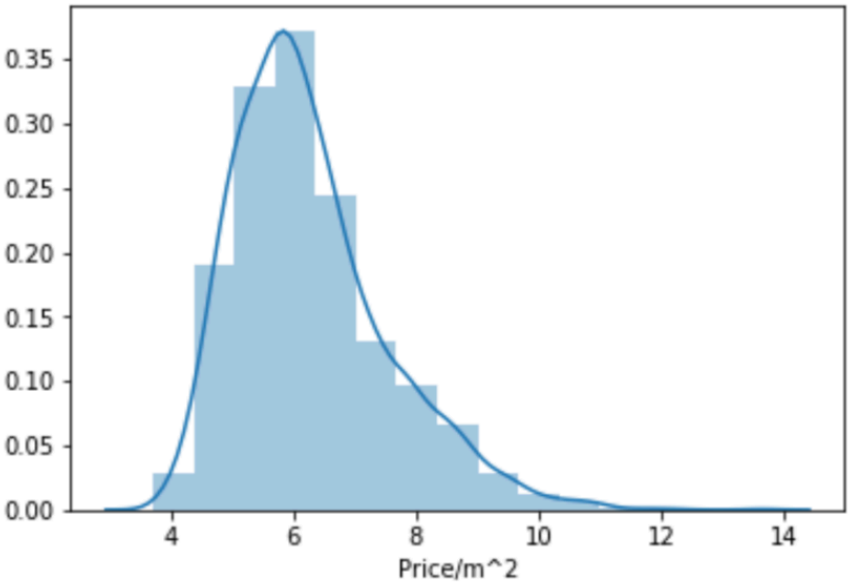
1. the median is around 3,000,000RMB, 3 quarters are less than 7,500,000RMB;

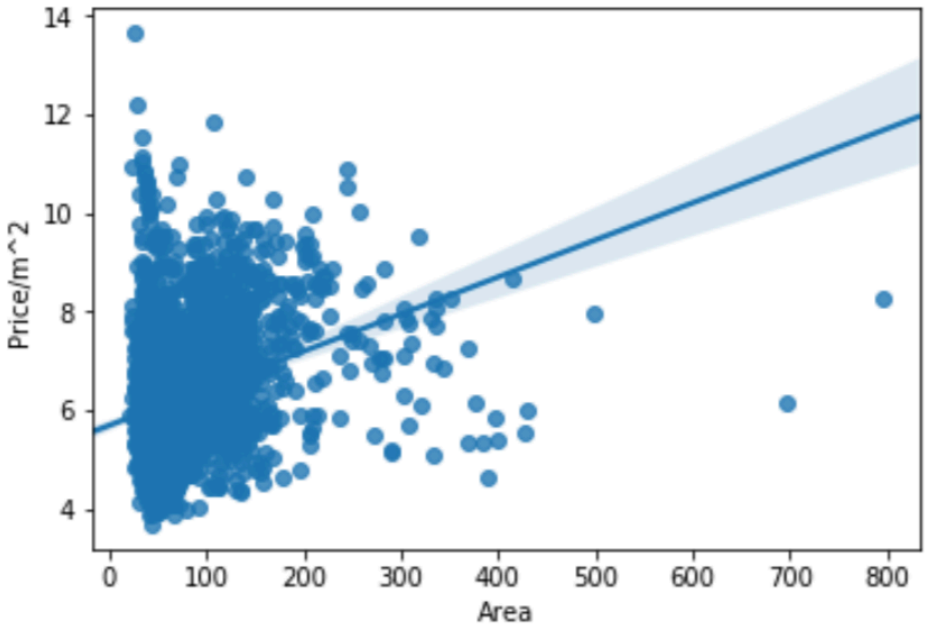
2. Price of apartment with 1,2,3 rooms differs obviously while price of apartment with 4,5,6,7 rooms differs slightly. One interesting finding is that apartments with 6 rooms are a little cheaper than those with 5 rooms;
3. price, bedrooms and area are highly correlated;
4. the most frequent price/m^2 is 60,000 RMB/m^2;
5. no correlation between unit price and area.






Price
Area
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
600
700
800

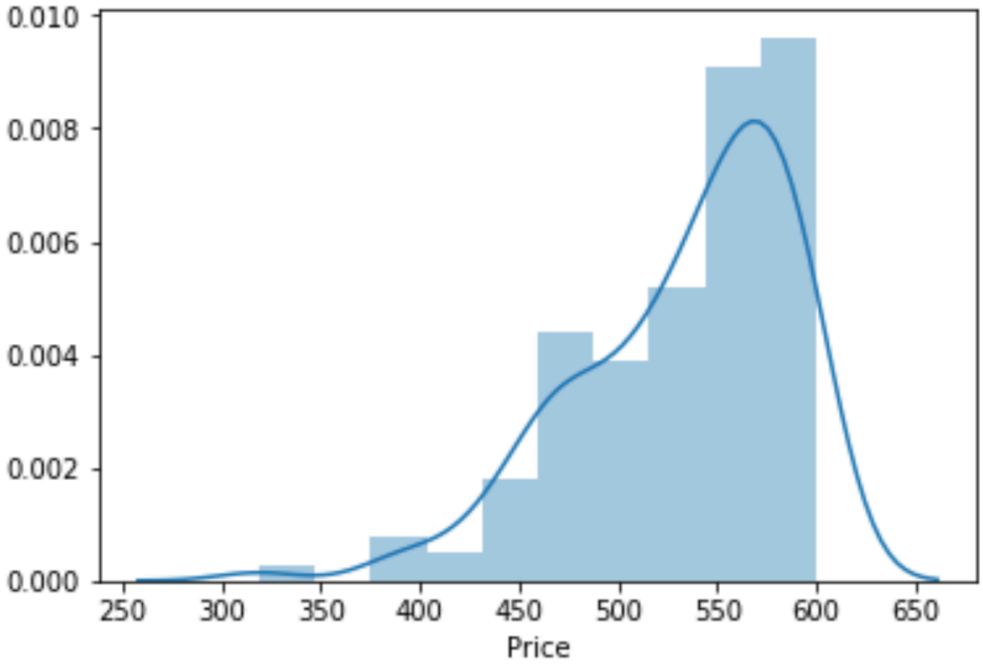

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
4
6
8
10
12
14
Price/m^2

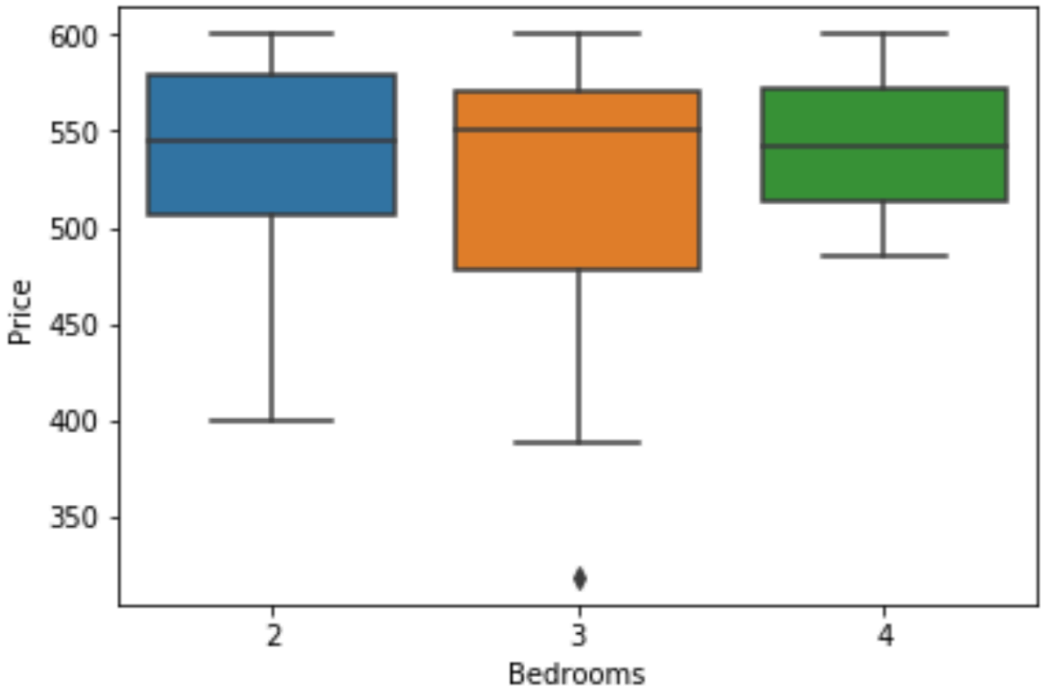


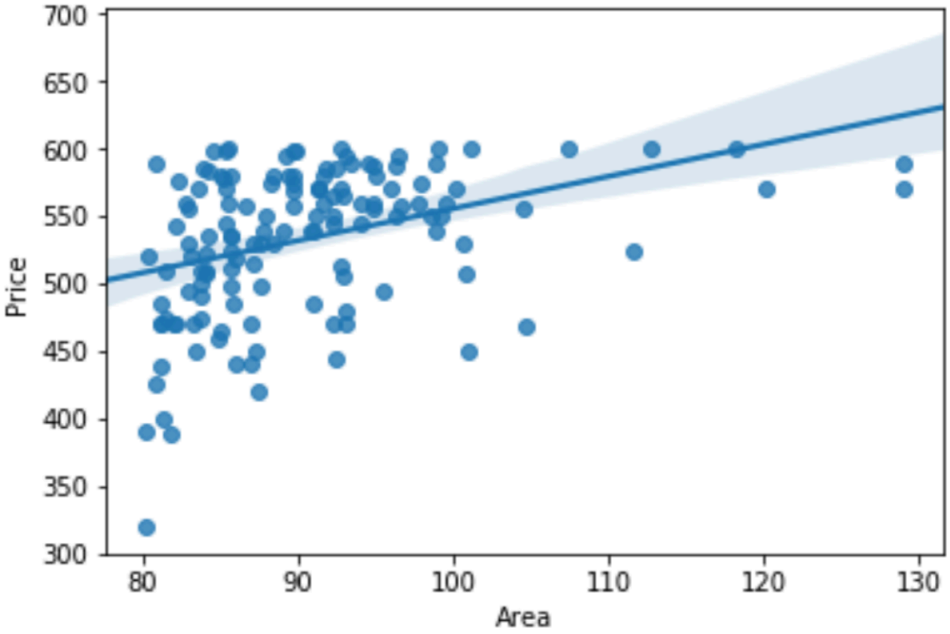
For selected convenient apartments:

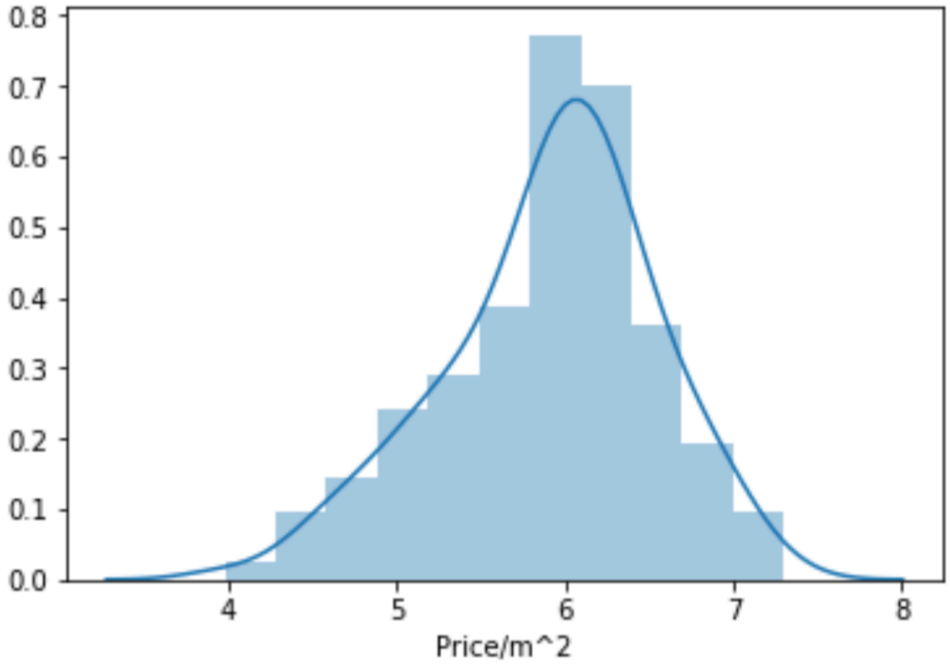
Here are the findings of selected convenient apartments:

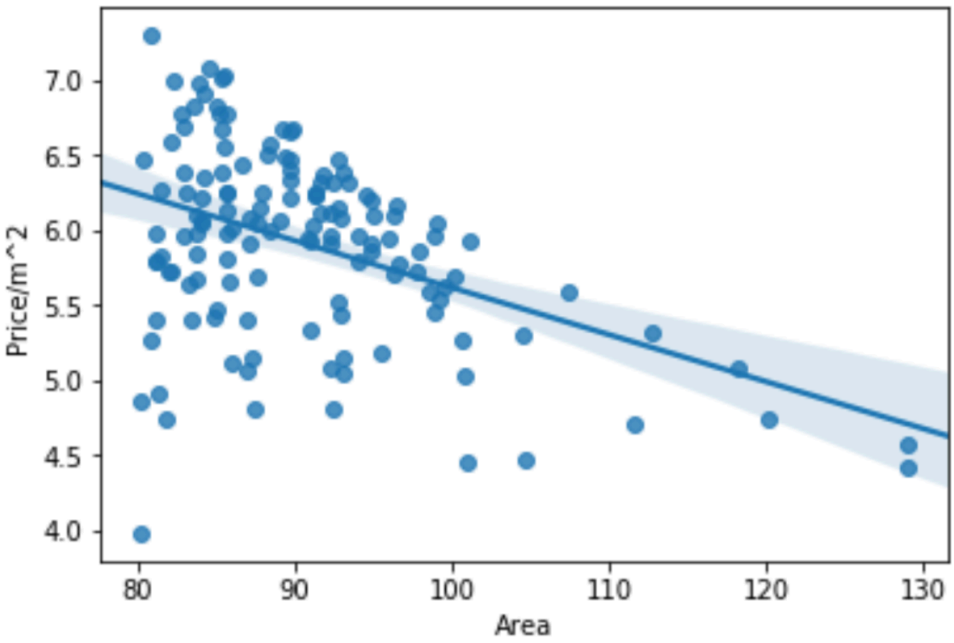
1. most apartments are around 5,800,000RMB;
2. almost no difference between price of apartments with different rooms;
3. price and area are slightly correlated;
4. the most frequent price/m^2 is 60,000 RMB/m^2;
5. no correlation between unit price and area.
6. as for tranportation, most apartments have 4 metro stations within working distance;
7. as for shopping, most apartments have 4 supermarkets within working distance;
8. as for entertainment, most apartments have 6 cinema within working distance.

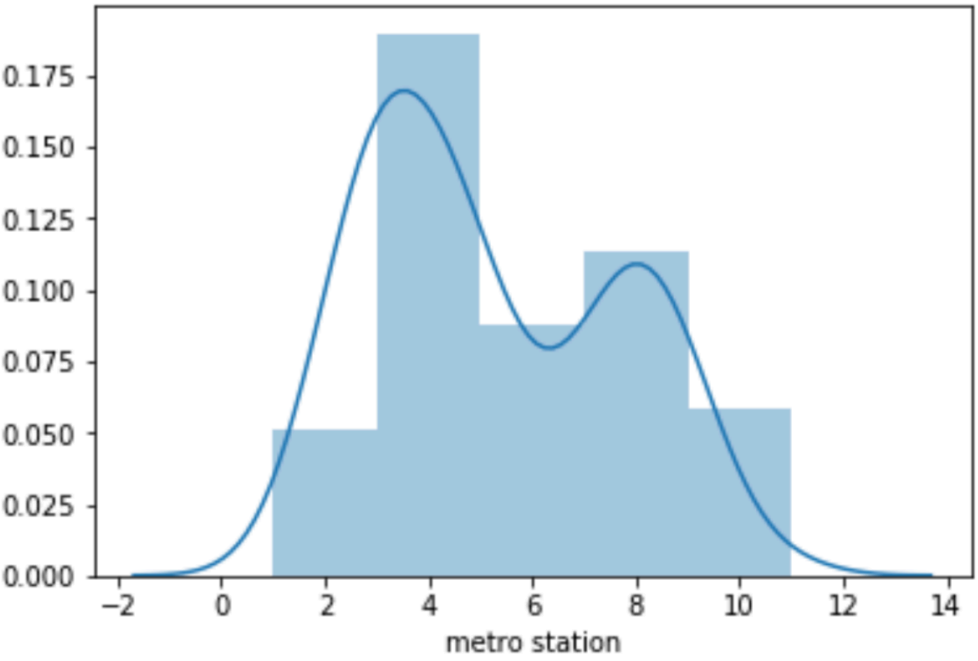

0.010
0.008
0.006
0.004
0.002
0.000
250
300
350
400
450
500
550
600
650
Price

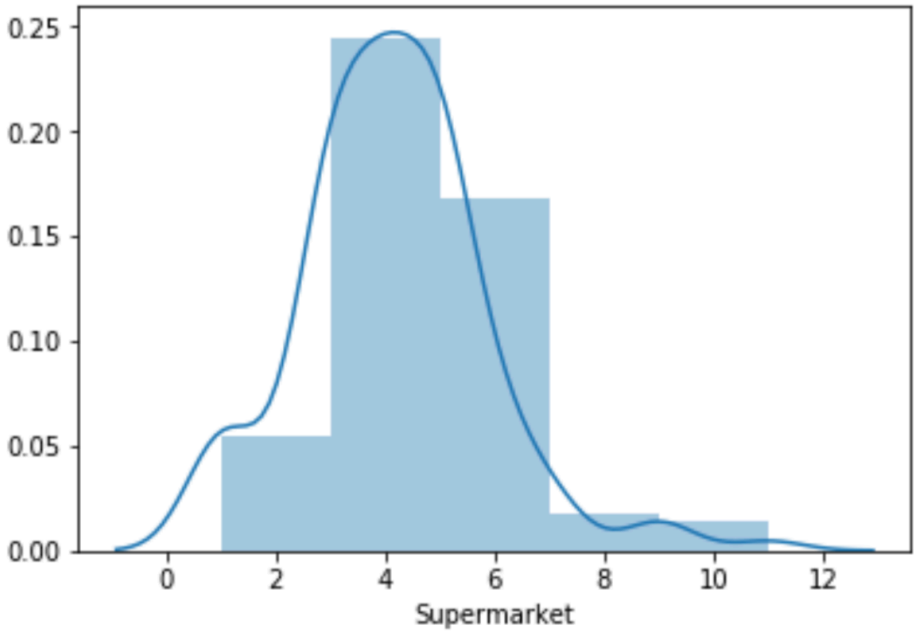

600
550
500
450
400
350
Price
2
3
4
Bedrooms

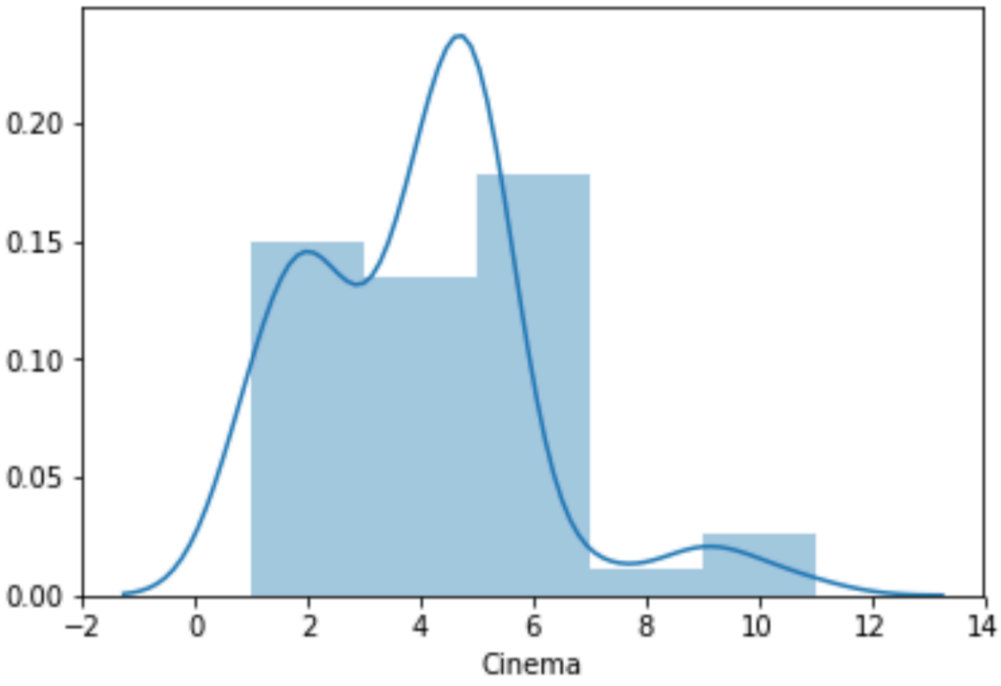

700
650
600
550
500
450
400
350
300
Price
80
90
100
110
120
130
Area


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
4
5
6
7
8
Price/m^2

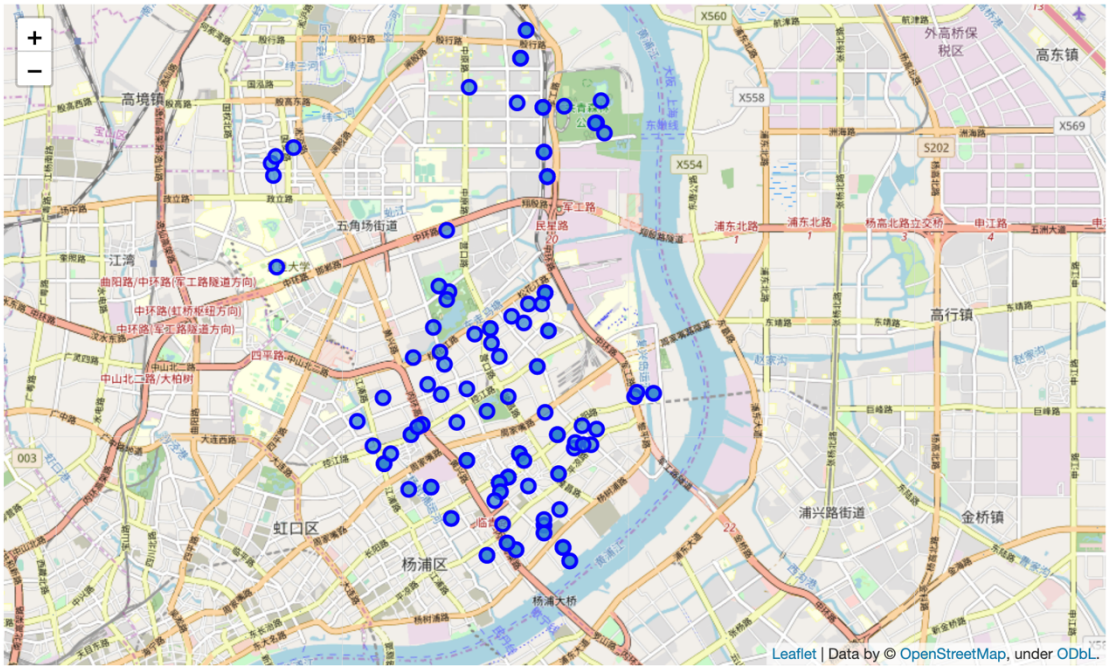

Price/m^2
7.0
6.5
6.0
5.5
5.0
4.5
4.0
80
90
100
110
120
130
Area


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−2
0
2
4
6
8
10
12
14
metro station


0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
Supermarket


0.20
0.15
0.10
0.05
0.00
−2
0
2
4
6
8
10
12
14
Cinema

Location of selected convenient apartments:



Clustering of selected convenient apartments:

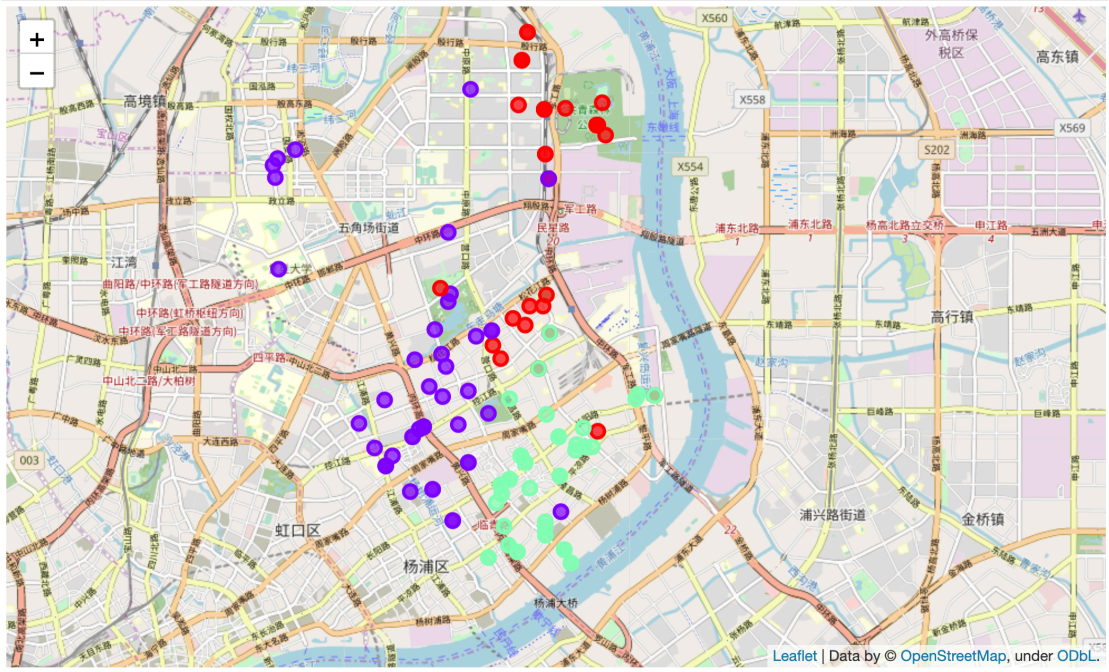
We select price, area, longitude, latitude as clustering features.

The apartments falls into 3 clusters. Cluster0 is red, Cluster1 is purple, Cluster2 is green.

Cluster 0: low price, small area, with fewer metro stations, supermarkets and cinemas, and low unit price.

Cluster 1: high price, large area, with many metro stations, supermarkets and cinemas, and medium unit price.

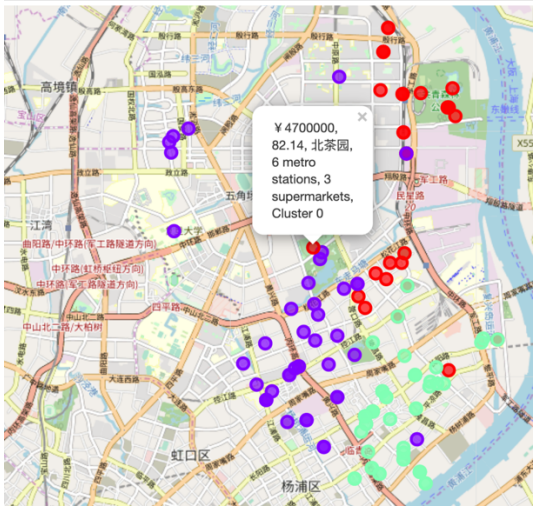
Cluster 2: medium price, medium area, with enough metro stations, supermarkets and cinemas, and high unit price.



Cluster	Price	Bedrooms	Area	Longitude	Latitude	metro station	Supermarket	Cinema	Price/m^2
0	469.735294	2.500000	86.856471	121.539964	31.311409	3.088235	2.676471	2.352941	5.418352
1	567.022727	2.477273	94.818864	121.521307	31.287489	7.181818	5.159091	4.840909	6.050149
2	544.796610	2.169492	89.697627	121.538507	31.273863	5.033898	4.237288	4.237288	6.090598

A cost-efficient apartment:

A red marker means it is cheap (price and unit price). It is close to cluster1, which means it is as convenient as those apartments of Cluster1. As we compare the area, 82.14 m^2 vs 80.8 m^2. Bravo!



Conclusion

Purpose of this project was to find an appropriate apartment in Yangpu Area. The apartment has to meet the following 2 conditions:

1. Apartment with min 2 bedrooms and min 80 square meters with price not to exceed 6,000,000RMB
2. Area with ammenities like subway metro station, supermarket and cinemas within walking distance(<=1.5 km)

First, we collect apartment description data from Lianjia website, get geodata with Baidu api and use Foursquare data to find venues like metro station, supermarket and cinemas.

Secondly, we explored the information of apartment, such as price, number of bedrooms and area and found that constraint is a common requirement. We also figured out the relationship between price and area. We compared the situation of all apartments in Yangpu Area and the selected convenient apartments.

Thirdly, we used k-means clustering method to divide them into 3 categories by price, area, location. After that we found features of each cluster as follows: Cluster 0: low price, small area, with fewer metro stations, supermarkets and cinemas, and low unit price. Cluster 1: high price, large area, with many metro stations, supermarkets and cinemas, and medium unit price. Cluster 2: medium price, medium area, with enough metro stations, supermarkets and cinemas, and high unit price.

Finally, we found a **cost efficient** apartment!