Finding A Proper Location To Live In Shanghai, China

Rui Wu

August 1st, 2019

1. Introduction

1.1 Background

In China, Shanghai is a popular place for people to work and live. This city has convenient traffic, various venues and many companies' headquarters are located in Shanghai. Many people working here are coming from other cities, so they need to rent apartments or house here. Also, people may have different preferences about living environment. They care various factors, and these factors have different priorities for individual.

1.2 Problem

Actually this problem is a real problem of myself. I got a job in Shanghai, China, and I came to this city last year. Because I am a consultant, so I often go on business trip to other cities. Living near metro station is important to me, which make easier for me to access to airport or train station. And besides living near metro station, of course, for living easy and convenient, I want to live in a place where there are restaurants, cafe, fitness centers, pharmacies, supermarkets or malls.

Last year, I lived near Hailun road('海伦路' in Chinese) metro station and this place have all these venues. I like this place. However, I have to move to other places because of the end of house contract and some other reasons, so I want to find a similar place in Shanghai to live in.

And my problem is, I want to find a similar place(as Hailun road) to live in, which can meet my following requirements:

1) near metro station, within 500 meters;
2) there are restaurants, cafe, fitness centers, pharmacies, supermarkets or malls around the station;
3) this station has multi-lines across, which can be more convenient than just 1 line;
4) stations near airport and train station are out of consideration because of the daily noise.

1.3 Interest

1) People who needs to rent apartments in Shanghai, China, like me. You can get some insights of these places to help you make decision of where to live;
2) Stakeholders of a renting apartment app, which can provide a recommendation of places to app users based on their filtering conditions and preferences.

2. Data acquisition and cleaning

2.1 Data sources

I found geographical coordinates of almost all metro stations in Shanghai from https://download.csdn.net/download/sinat_29675423/10844042. The data is stored in SQL file. I downloaded the file and copy these data directly into this notebook. And I put these data into a dataframe. Each row represents a station and its coordinates. In column 'id', the 1st number is the line number and the 2nd, 3rd number represent the sequence number of stations of each line.
It also needs to be noticed that because some lines are under continual construction in these years, some new stations appear. And the data I use in this project was collected in 2015/07/28 and some was updated in 2018.
There are line 1 to 13, 16, 17 of Shanghai metro, so first I create a dataframe named sh_station, which contains geographic coordinates of these stations.

2.2 Data cleaning

In sh_station, data is stored by each line. Many lines cross, so some stations are duplicated. I want to separate the information of sh_station into two parts.
In the first part, it includes the information of all these unique stations' geographic coordinates, its name is still sh_station. I drop the duplicated station info and as a result, it has 322 unique stations(322 rows) and 4 columns(station id, station name, longitude, latitude). Its first five rows are shown in Table1.

	id	st_name	lng	lat
0	101	莘庄	121.385379	31.111193
1	102	外环路	121.393020	31.120899
2	103	莲花路	121.402943	31.130986
3	104	锦江乐园	121.414107	31.142217
4	105	上海南站	121.430041	31.154579

Table1 First five rows of sh_station

In the second part, it includes the information of all stations which has over 1 lines. Its name is freq_over1. The reason why I choose the stations with over 1 line is that,

Hailun Road('海伦路') station has 2 line across, so if I want to find a better place to

live in, I had better find a place where the station has over 2 lines across. But 2 lines is also good for me. Figure 1 shows the number of stations of 1, 2, 3, 4 lines across. We can see there are 57 stations with more than 1 line. Then I drop the airport station, train station(4 stations in all). The reason why I drop the airport station, train station is that, I want to find some quiet places to live and airport or train station may be noisy. As a result, freq_over1 has 53 stations, which has multi-lines varying from 2 lines to 4 lines.

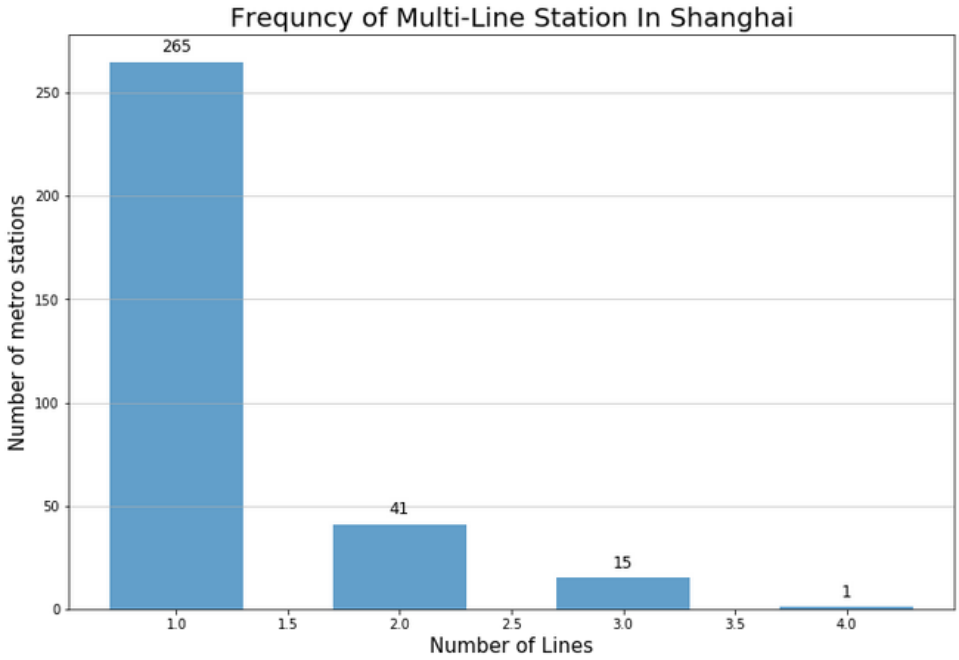

Figure1 Number of metro stations with different lines

2.3 Feature selection and preprocessing

After data cleaning, I will give some details about what kind of datasets I want to use to build clustering model and before building model, I will do some exploratory data analysis.

In order to find places similar with Hailun road, I need to cluster these 322 metro stations(322 samples).

The features I will use is the appearance possibility of each venue category with 500 meters of each station. For example, there are 16 venues around Hailun road(<500 meters). Among these 16 venues, there are 2 hotels, the appearance possibility of hotel is 2/16, that is 0.125. So in each feature column, the value is between 0 and 1. And in this case, there are 262 features(venue categories).

It is noticed that I use Four Square APIs to request venues data based on metro stations' geographical coordinates. Some other factors like renting money is not in these features, which is important for me to consider when I choose apartment.

3. Exploratory Data Analysis

3.1 Explore a metro station in Shanghai

First I get the Shanghai geographical coordinates: 31.2252985(latitude), 121.4890497(longitude). And then I draw a geographic map of these 322 metro stations of Shanghai, as shown in Figure 2.

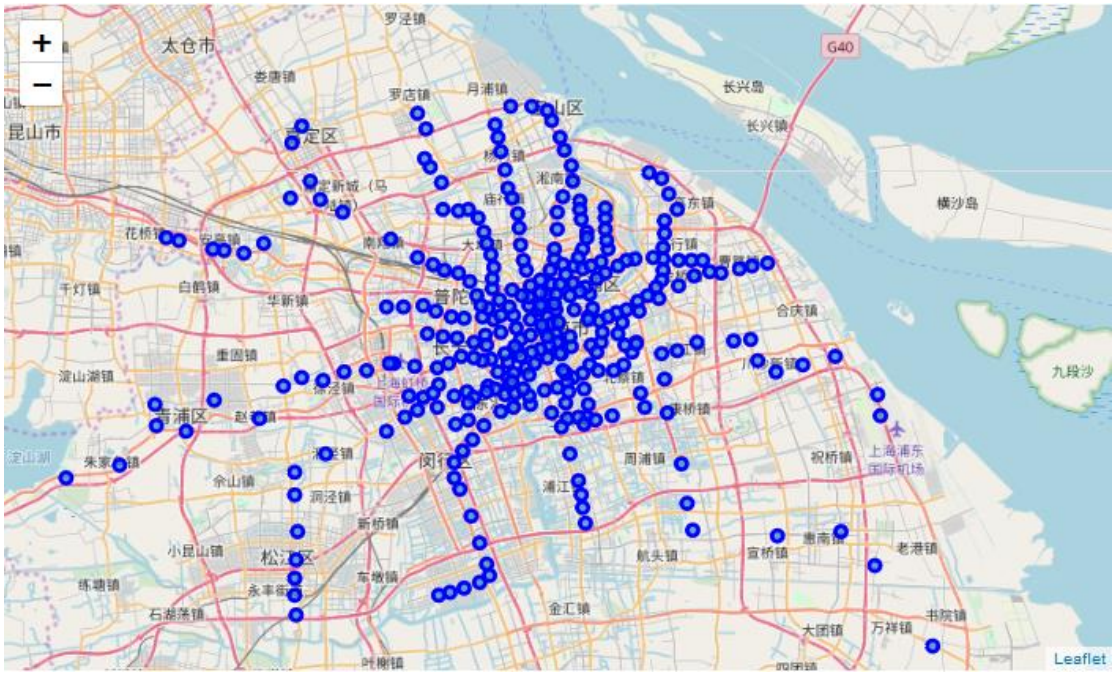

Figure2 Shanghai’s 322 metro stations location

Second, let's get the top 50 venues(maybe less than 50) that are around Hailun road(‘海伦路’)station with a radius of 500 meters. Here we can see the details of these 16 venues in Table2.

And I observe that maybe some small stores like MyFamily(which is a chain supermarket in China) are not shown here. However, these small stores contributed so much to my daily life convenience. So these results can just be a recommendation and reference, which may be a little different from the actual surroundings.

	name	categories	lat	lng
0	1933 Shanghai (19叁叁老场坊)	Art Gallery	31.256540	121.487510
1	Sheraton Shanghai Hongkou Hotel (Sheraton Shan...	Hotel	31.259633	121.483617
2	SNH48 Theater (SNH48星梦剧院)	Theater	31.258674	121.485969
3	Starbucks (星巴克)	Coffee Shop	31.260685	121.484656
4	Central Perk	Café	31.257594	121.485415
5	Noodle Bull	Noodle House	31.256787	121.487941
6	物美 Wu Mart	Department Store	31.260210	121.491589
7	Jade Garden (苏浙汇)	Chinese Restaurant	31.256591	121.487445
8	Hailun Road Metro Station (海伦路地铁站)	Metro Station	31.261200	121.486099
9	Canil Café (狗窝)	Café	31.256430	121.487433
10	Rosso Italiano 红意	Italian Restaurant	31.256425	121.487292
11	Grace Coffee	Coffee Shop	31.257537	121.485459
12	Jiulong Hotel 九龍賓館	Hotel	31.255759	121.487701
13	红麻辣料理	Szechuan Restaurant	31.261436	121.492250
14	半層書店	Bookstore	31.257079	121.484812
15	Sheraton Hongkou Club Lounge	Lounge	31.259664	121.483658

Table2 Venues around Hailun road metro station within 500 meters

3.2 Explore metro stations in Shanghai

Now, let's get the top 50 venues that are around each station(322 stations) within a radius of 500 meters. The process is similar with 3.1. Four Square response 2931 venues in all. Table3 shows the first 5 rows of these venues. We can see that each row represents a venue, which belongs to a station and each row shows the station's name, coordinates, venue's name, coordinates and category.

	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	莘庄	31.111193	121.385379	Starbucks (星巴克)	31.108173	121.382887	Coffee Shop
1	莘庄	31.111193	121.385379	外婆家	31.108776	121.381458	Chinese Restaurant
2	莘庄	31.111193	121.385379	Skymall (仲盛世界商城)	31.109338	121.382424	Shopping Mall
3	莘庄	31.111193	121.385379	Blue Frog (蓝蛙)	31.108128	121.383250	Burger Joint
4	莘庄	31.111193	121.385379	上影CGV莘庄影城	31.108190	121.382946	Multiplex

Table3 Venues' coordinates and categories

Then I group these venues by station and get a list of number of venues of each metro station, as shown in Table4. We can see that, the first 10 stations show 50 venues and actually they may have venues more than 50. And the least number is 1 venue. The difference among these stations are obvious.

Station name	Number of venues within 500m
徐家汇	50
静安寺	50
人民广场	50
陆家嘴	50
淮海中路	50
新闸路	50
黄陂南路	50
上海图书馆	50
天潼路	50
南京东路	50
南京西路	48
中山公园	48
商城路	47
常熟路	43
••••••	……
光明路	1
花桥	1
江月路	1
临港大道	1
临平路	1
淞发路	1
颛桥	1
徐盈路	1
昌吉东路	1

Table4 Number of venues of each station

Finally, I need to generate the dataset as the input to the cluster model. As you can see in the Table5, each row represents a station, and except the first column 'Station', the rest columns are venues category name. And there are 314 stations and 262 features(venue categories). Table5 shows the first 5 rows of the dataset. We can see that 314 stations has data of venues frequency. In previous part, there are 322 unique stations. So using Four Square APIs, there are no venues around 8 stations.

Station	……	Chinese Restaurant	Chocolate Shop	Climbing Gym	Clothing Store	Club House	Cocktail Bar	Coffee Shop	……
七宝	……	0.25	0	0	0	0	0	0.125	……
七莘路	……	0.285714	0	0	0.142857	0	0	0.142857	……
三林	……	0	0	0	0	0	0	0.2	……
三林东	……	0	0	0	0	0	0	0	……
三门路	……	0	0	0	0	0	0	0	……

Table5 Appearance possibility of each venue category of 314 stations

Also, according to Table5, I can get the top 10 venues of each station and put that

into a dataframe, and the first five rows is shown in the Table6.

Station	1st most common venue	2nd most common venue	3rd most common venue	…	9th most common venue	10th most common venue
七宝	Chinese Restaurant	Shopping Mall	Fast Food Restaurant		Food & Drink Shop	Zhejiang Restaurant
七莘路	Chinese Restaurant	Clothing Store	Japanese Restaurant		Frozen Yogurt Shop	Flea Market
三林	Shopping Mall	Coffee Shop	Metro Station		Gaming Cafe	Furniture / Home Store
三林东	Hotel	Metro Station	Fast Food Restaurant		Fruit & Vegetable Store	Frozen Yogurt Shop
三门路	German Restaurant	Business Service	French Restaurant		Food Truck	Frozen Yogurt Shop

Table6 Most common venues of each station

4. Clustering Modeling

In this project, I use K-means algorithm to cluster these 314 stations. This algorithm has some advantages: 1. simple principle, convenient implementation and fast convergence speed; 2. better clustering effect;3. ony the number of clusters k is needed to adjust parameters. However, the choice of k is not easy to guess. So I will use 2 kind of k values to build models, one is 5 and the other is 8.

4.1 Cluster number = 5

4.1.1 Applying k-means algorithms to build model

I set the cluster number as 5 and random_state as 0, use the input dataset as Table5, which has 314 samples and 262 features.

As you can see in Figure3, each spot has a color. One color represents a cluster.

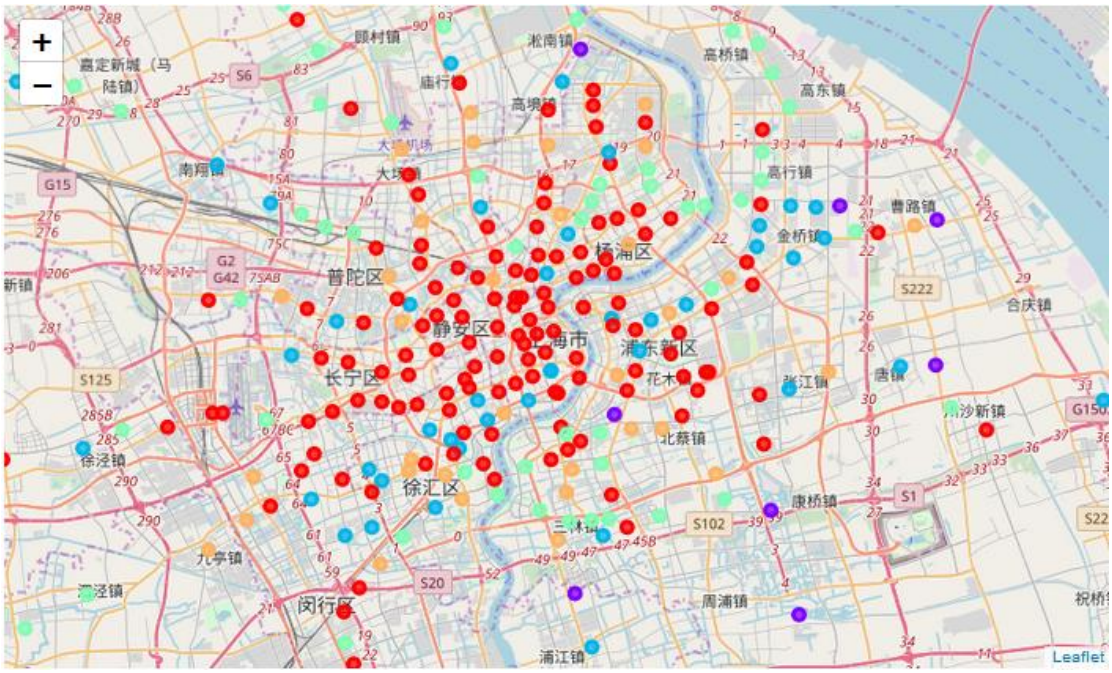

Figure3 Cluster results of 314 stations(cluster number=5)

In this case, red represents cluster1, purple represents cluster2, blue represents cluster3, green represents cluster4, orange represents cluster5.

4.1.2 Solution to the problems

Hailun road station is in cluster1(red), and cluster1 has so many stations. Then we can user freq_over1 to filter the stations of cluster1 with over 1 line. As a result, I got 38 stations, which is in cluster1, and has more than 1 line. Figure4 shows the location of these 38 stations.

Figure4 38 proper places to live in

4.2 Clustering number = 8

4.2.1 Applying k-means algorithms to build model

I set the cluster number as 8 and random_state as 0, use the input dataset as Table5, which has 314 samples and 262 features.

As you can see in Figure5, each spot has a color. One color represents a cluster.

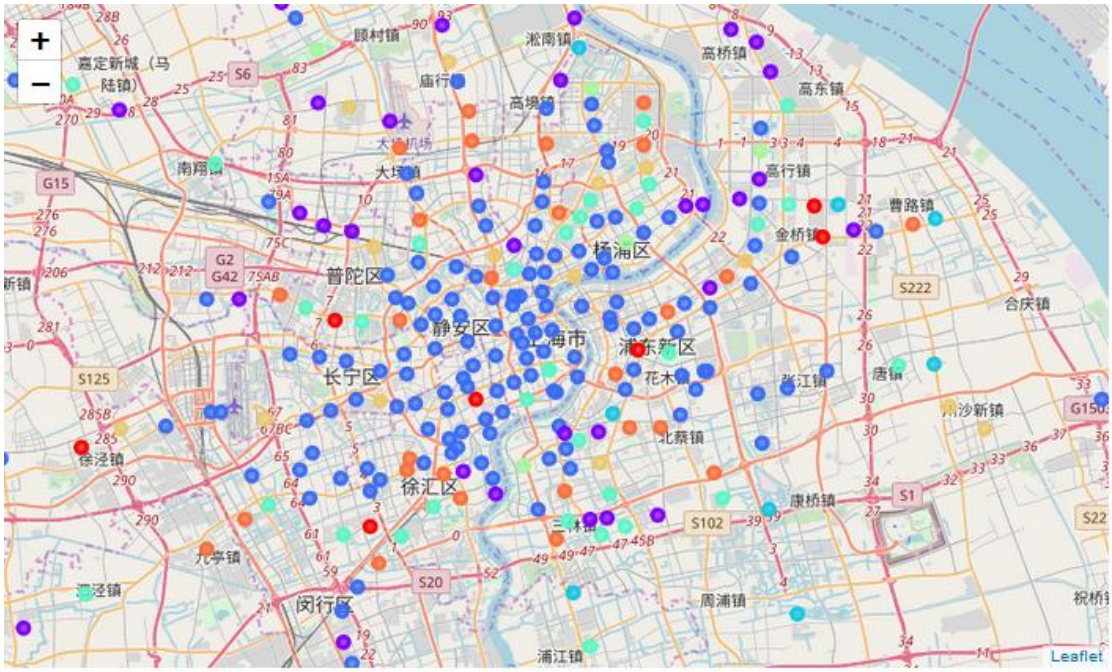

Figure5 Cluster results of 314 stations(cluster number=8)

In this case, red represents cluster1, purple represents cluster2, blue represents cluster3, light blue represents cluster4, light green represents cluster5, grass green represents cluster6, yellow represents cluster7 and orange represents cluster8.

4.2.2 Solution to the problems

Hailun road station is in cluster3(blue), and cluster3 has so many stations. Then we can user freq_over1 to filter the stations of cluster3 with over 1 line. As a result, I got 44 stations, which is in cluster3, and has more than 1 line. Figure6 shows the location of these 44 stations.

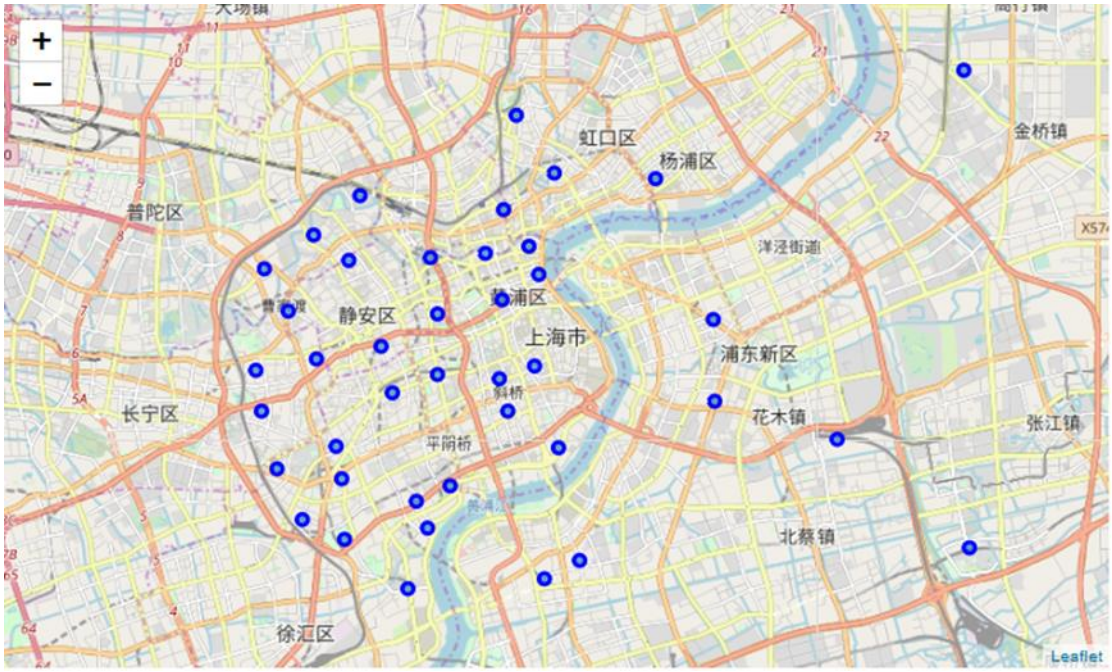

Figure6 44 proper places to live in

4.3 Performances of different parameters

When cluster number is 5, we get 38 places and when cluster number is 8, we get 44 places. Now I will compare these two results.

It shows that proper_places_8 contains the whole stations of proper_places_5. The other 6 stations(‘上海体育馆’,’东安路’,’东方体育中心’,’大木桥路’,’宜山路’,’曹杨路’) may really has some difference to these 38 stations. Because the difference is based on the venues recorded in Four Square, so maybe some other venues are not recorded in it. If you want to make sure about these difference, you had better go on a field visit to explore the actual environment.

5. Conclusions and discussions

From the above analysis, we got over 40 places(stations, including Hailun road) which may be proper to live in.

These places meet my following requirements:

1) they are near metro station, within 500 meters;

2) some small fitness centers and pharmacies are not in our venue lists from Four Square APIs, so we need to do some field visit to explore the actual venues; restaurants location info is relatively more and complete;

3) these stations have multi-lines across(>=2 lines);

4) stations near airport and train station are excluded.

Other important factors like renting money(eg, apartment in Jing An District is more expensive than Bao Shan District), need to be considered, which is not included as a feature in this clustering analysis. So in the next step, I will find some renting money data from the website and add it into the input dataset as a new feature, which will contribute to more accurate clustering results according to my living requirements.