Capstone Project - A Graduate's Guide to Renting a House in Shanghai

Applied Data Science Capstone by IBM/Coursera

Ji Mao
April 02, 2021

1. Introduction: Background and Problem

After three or four years at university, most students will be thinking about their next steps, and where they will be living. Many graduates choose to rent with friends in their university town, whilst others relocate to the location, they have landed their job or graduate placement.

Choosing a house for rent is not an easy task, as you have to find an optimal one from countless information and advertisements. Thus, this report will be targeted to **stakeholders who need to rent house in new city**. We will try to provide graduates some guides and the target city will be Shanghai, the financial center in China.

We will use our data science powers to generate a few most promising neighborhoods and houses based on **rent fee, house area, neighborhoods, transportation**. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

2. Data acquisition and cleaning

2.1 Data Source

Based on definition of our problem, factors that will influence our decision are:

- *decent rent fees*
- *house area*
- *the convenience of its neighborhoods*
- *distance of house from work location*

Following data sources will be needed to extract/generate the required information:

- houses data like location, neighborhoods, rent fee, house area and url in the house renting website
- coordinates of city, houses and its neighborhoods will be obtained using AMap API
- venues data including their type and numbers in each neighborhood will be obtained using AMap API for neighborhoods segmentation and clustering
- transportation data will be obtained and showed through **Amap JS API**

2.2 Data Cleaning

By using the library BeautifulSoup of Python, house data in house rent website Lianjia will be obtained. After that, data will be cleaned and processed for further analyzing. The houses information including need to be scraped are house price, areas, orientation, neighborhood, address etc., and save all the house data into a csv file for further use.

	name	location	rent	area	orientation	pattern	url
0	长白三村	杨浦-黄兴公园-长白三村	4600	43 ㎡	南	2室0厅1卫	https://sh.lianjia.com//zufang/SH2742772410171...
1	川杨新苑(四期)	浦东-张江-川杨新苑(四期)	5000	60 ㎡	南	2室1厅1卫	https://sh.lianjia.com//zufang/SH2742841630095...
2	浦江馨都	闵行-闵浦-浦江馨都	3700	77 ㎡	南	2室1厅1卫	https://sh.lianjia.com//zufang/SH2743268173280...
3	运光小区	虹口-曲阳-运光小区	4700	40 ㎡	南	1室1厅1卫	https://sh.lianjia.com//zufang/SH2743334892141...
4	宝林九村	宝山-淞宝-宝林九村	3500	58 ㎡	南	1室1厅1卫	https://sh.lianjia.com//zufang/SH2743395188876...

The tasks we'll be done in data cleaning are:

- extract the neighborhood information into a new column;
- check if there is NaN value and replace NaN value with manually searched information
- data type conversion
- select unique neighborhoods data and create new dataframe for them to explore the venues around neighborhoods:

	Neighborhood	District	Postcode	Level	Longitude	Latitude
0	杨浦黄兴公园	杨浦区	310110	兴趣点	121.530003	31.293311
1	浦东张江	浦东新区	310115	乡镇	121.614288	31.205289
2	闵行闵浦	闵行区	310112	兴趣点	121.410704	31.140333
3	虹口曲阳	虹口区	310109	乡镇	121.492828	31.281200
4	宝山淞宝	宝山区	310113	兴趣点	121.498940	31.390274

3. Methodology

we'll cluster our houses based on thress indexs: house price, area and its neighborhood convenience; Since the house price and areas are data we've already scraded from the rental website, the key problem is how to get the neighborhood convenience index. Following are the steps to calculate the neighborhoood convenience:

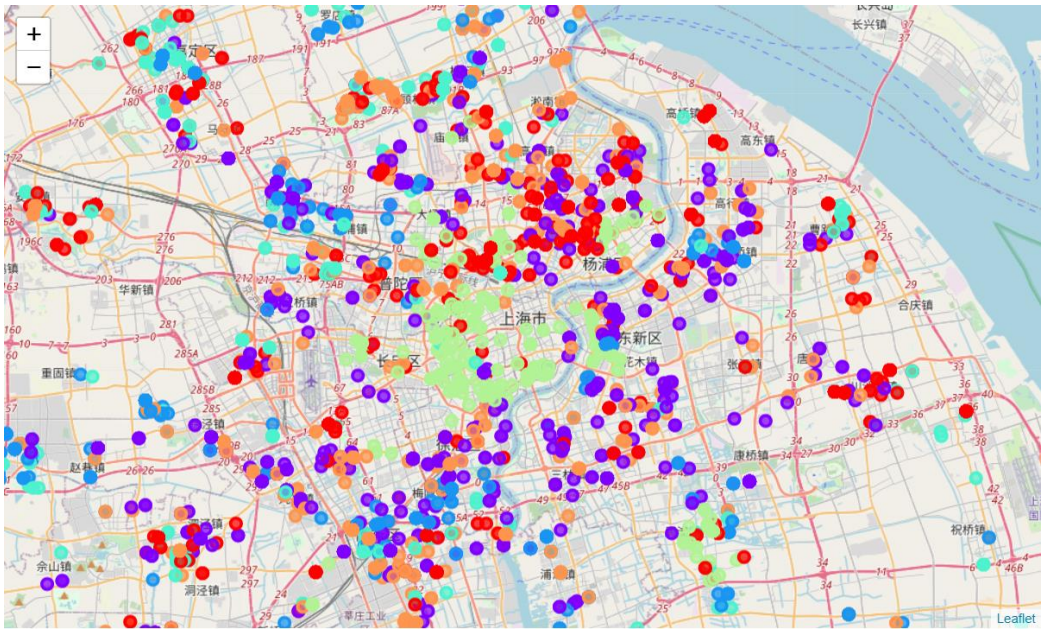
- with help of the AMAP POI API, count the venues within 500m of the neighborhoods.Here we'll **focus on six categories**: restuarants, shoping mall, medical isurance, fitness and entertainment, parks, education and culture
- then we'll allocate different **weighting factors** to different venue categories according their importance for me when choose a house. At the end, we'll calculate a value to represent the convinience of neighborhoods.
- after the calculating of the index, we could apply the **KMeans** to clusering houses into different cluster, and summarize the characteristics of each cluster.

4. Results and Discussion

After clustering of the houses and exploring each clusters, we summarize the characteristics of each cluser in the following table:

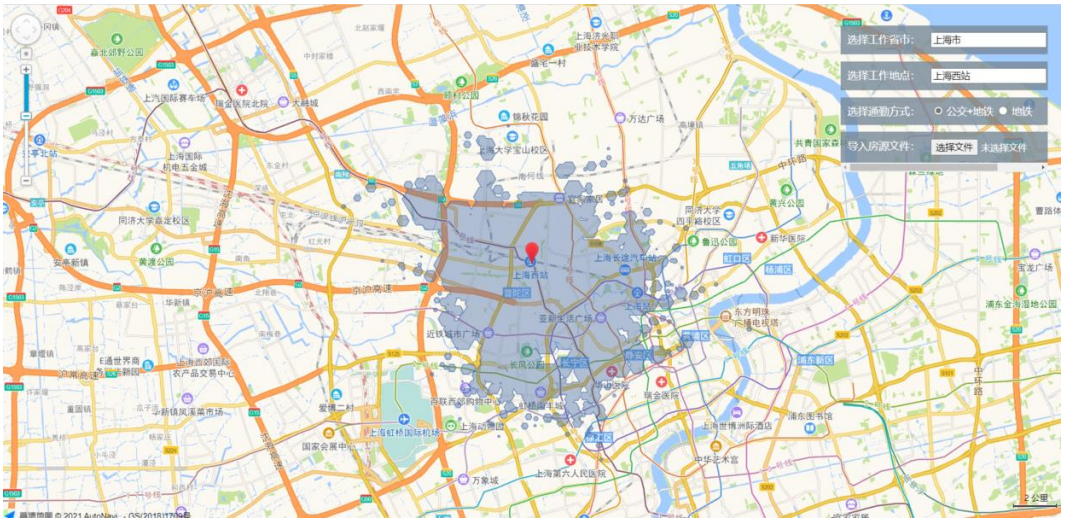
Cluseters	Average Price	Area	Neighborhood Convenience
1	Medium: 4200 ¥/month	Medium: 59㎡	Medium: 0.2
2	High: 5700 ¥/month	Big: 72㎡	Low: 0.15
3	Medium: 4250 ¥/month	Big: 79㎡	Low: 0.01
4	Low: 3500 ¥/month	Big: 72㎡	Low: 0.14
5	High: 5400 ¥/month	Small: 45㎡	High: 0.40
6	Medium: 5000 ¥/month	Medium: 66㎡	Medium: 0.17

And we also visualize all the houses in the Folium map:

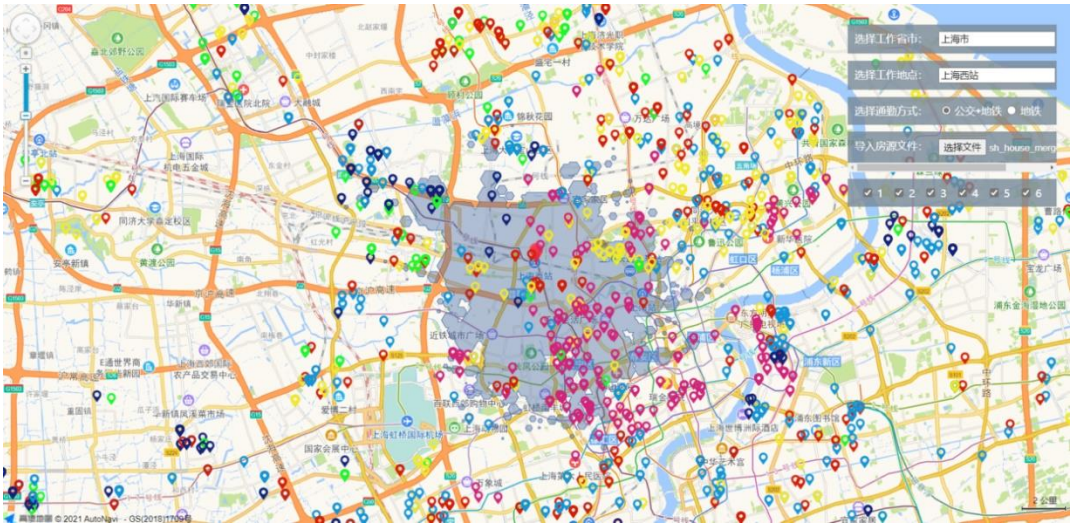


The effect of the above-mentioned visualization by Folium is still relatively static, and the traffic convenience of the hosues is not known, so we will further explore the traffic convenience information to facilitate house choose. With help of AMap JS API, we can visualize our houses in Amap. The advantages of this API allows you to write your own js code to realize everything you want to show in the map. In my project, there are mainly three functions in the Amap:

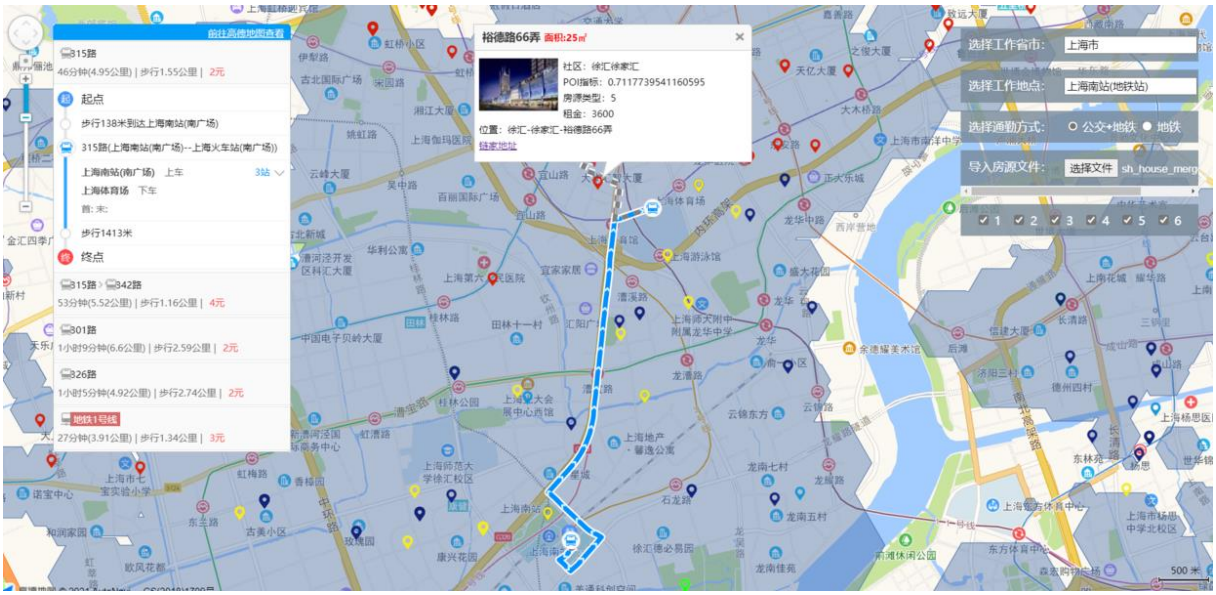
- Workplace input window: input your work place in Shanghai, then it will be marked in the map and show an area within 1 hour drive to the work place;



- Import houses data: import the houses data in csv file, the houses will be marked with different color in the map just like in Folium;



- Traffic navigation: click on one house, the traffic options and time will be automatically shown in the traffic window, which help you determine the traffic convenience around the house. Besides, the pop-up label will show you detailed information of the house, such as address, neighborhood convenience index, price, cluster, house url. By clicking on the url, it willed be directed to the house rental page:



5. Conclusions

Now, based on the cluster types summarized above and the JS Amap functions, we can conveniently and intuitively select the appropriate housing type according to our needs on the map.

For example: Generally, we would like the house to be close enough to the our workplace, so we only consider the houses located in the blue area of the map. Then, we can further consider the community convenience of the house and try to choose a house from cluster 1 and cluster 5. If we still hope that the rent can be a little bit lower when the convenience of the community is in a similar level, then we should choose house of cluster 1. In the few left houses that were filtered, we can click on the house in the map to view the relevant information one by one and make a final decision. After the decision is made, you can click the link of the house to jump to the house rental interface.

Eventually, the house rental process will be much easier than before, this is also the objective of this project!