WHICH DISTRICT OF BEIJING TO CHOOSE TO SETTLE DOWN?

Cheng Yang

York University, Toronto, Canada

1. INTRODUCTION

Beijing is the capital city of China with a population of 21.54 million [1]. Since Beijing is China's centre of politics, culture, international communication, and technological innovation, it constantly attracts both Chinese and foreign citizens who are willing to settle in and live a life here, especially *the Millennials*.

However, in the aspect of settlement in such a large city, one has to consider the fact that the 16 districts of Beijing have drastically different population, housing prices, average annual income, venue structure, and region area. Therefore, it is of great interest for people who are planning to move to Beijing to analyze the data above, so that one can make wise decision which district to settle in. Motivated by finding directions in settlement in Beijing based on one's financial and living goals, this report provides preliminary results, specifically, visualization and clustering results on the above Beijing district profiles for settlement in one's ideal district.

The remainder of the report is organized as follows. We explain in detail the data in Section 2, methodology on how we analyze the data in Section 3, and the experimental results of 16 Beijing districts in Section 4. We discuss the experimental results in Section 5. Finally, we conclude the notebook in Section 6.

2. DATA

To analyze the population, housing prices, average annual income, venue structure, and region area of Beijing districts, in this section, we will analyze both non-location and location data as follows: A) Non-location data: Housing Price, Income, and Population data for Beijing city, and B) Location data for Beijing city.

2.1. Non-location data

We adopt the income and population data of Beijing districts published by National Bureau of Statistics [1]. We also adopt the housing price data published by AnJuKe [2]. All above non-location data is then manually re-orginized in a csv file in [3].

As shown in Fig. 1, the columns of the csv file represent:
1. 'id': the Beijing district name in Chinese;
2. 'price': the housing price in RMB per square meter;
3. 'price_k': the housing price in K RMB per square meter;
4. 'population': the population in tens of thousands;
5. 'population_k': the population in millions;
6. 'income': the income in RMB per year;
7. 'income_k': the income in K RMB per year.

2.2. Location data

We adopt the Beijing json file published by Geojson-Map-China [4] as our location data. We also adopt the location data in published by Foursquare as our main location data source. As shown in Fig. 3, the json file contains the contour of Beijing districts. We use it for data visualization purpose, specifically, visualization of the population. housing price, and income of each district. We use the Foursquare data to analyze the venue types and structures in each district (neighbourhood). See Section 4 for our experimental results.

3. METHODOLOGY

In this section, we first heuristically define a living-stress-level metric based on the non-location data, so that

	id	price	price_k	population	population_m	income	income_k
2	朝阳区	72139	72.139	360.5	3.605	70746	70.746
6	海淀区	88242	88.242	335.8	3.358	78178	78.178
1	昌平区	41395	41.395	210.8	2.108	45399	45.399
5	丰台区	59185	59.185	210.5	2.105	60144	60.144
3	大兴区	41720	41.720	179.6	1.796	43464	43.464
11	通州区	46757	46.757	157.8	1.578	40553	40.553
4	房山区	28608	28.608	118.8	1.188	39391	39.391
15	西城区	120524	120.524	117.9	1.179	75547	75.547
0	顺义区	40444	40.444	116.9	1.169	36575	36.575
14	东城区	98122	98.122	82.2	0.822	75547	75.547
10	石景山区	49889	49.889	59.0	0.590	71244	71.244
12	密云县	24625	24.625	49.5	0.495	34951	34.951
9	平谷区	24428	24.428	45.6	0.456	36012	36.012
7	怀柔区	31580	31.580	41.4	0.414	36797	36.797
13	延庆县	22788	22.788	34.8	0.348	33887	33.887
8	门头沟区	37878	37.878	33.1	0.331	49298	49.298

Fig. 1. Housing price, population and income of 16 Beijing districts in the descending order of the population.

one can roughly decide if a district is appropirate for settlement. We then categorize the 16 districts (neighbourhoods) into groups according to their venue types and structures., so that one can make a safer settlement decision based on both financial stress levels and preferred living/working styles.

3.1. Population, Housing Price, Income and PHI index

It is difficult to make decision on which district to settle by direct comparison of the population, housing price, and income because of the drastic difference among 16 districts for each of the three attributes. To help decision-making, we heuristically define a metric that measures a 'living-stress-level', named as *PHI index*, which is simply:

$$\mathrm{PHI} = \mathrm{Population} * \frac{\mathrm{HousingPrice}}{\mathrm{Income}}. \qquad (1)$$

The intuiation is that, regardless of the human-living-area of each district (for simplicity), the larger the population, and the larger the division of housing price and income, the higher the living and financial stress. We compute PHI's for each district based on data in Fig. 1. See Fig. 4 for the PHI indices.

3.2. Grouping of the districts

Since there is no groudtruth 'group' for each district, we adopt an unsupervised learning method *clustering* for the grouping task. We do this by preparing the location data from Foursquare, and then choosing an appropirate number of groups for clustering.

3.2.1. Data preparation for grouping

Searching closest venues in each district. When using the Foursquare location data, as shown in Fig. 5, we perform a searching scheme to choose an appropirate searching radius (with respect to the centroid of a district (neighbourhood)) to return the closest 50 venues. See Fig. 6 for the closest 5 venues of Chaoyang district (朝阳区).

Onehot encoding of the closest venues. We now adopt onehot encoding to the closest venues in each district. See Fig. 7 for the closest 5 venues of Chaoyang district (朝阳区). We then apply python 'groupby' process onto the onehot result in terms of the district, and then take the mean of the frequency of occurrence of each venue, for subsequent clustering process. See Fig. 8 for the dataframe that is ready for clustering. Also, see Fig. 9 for the top 3 most common venues

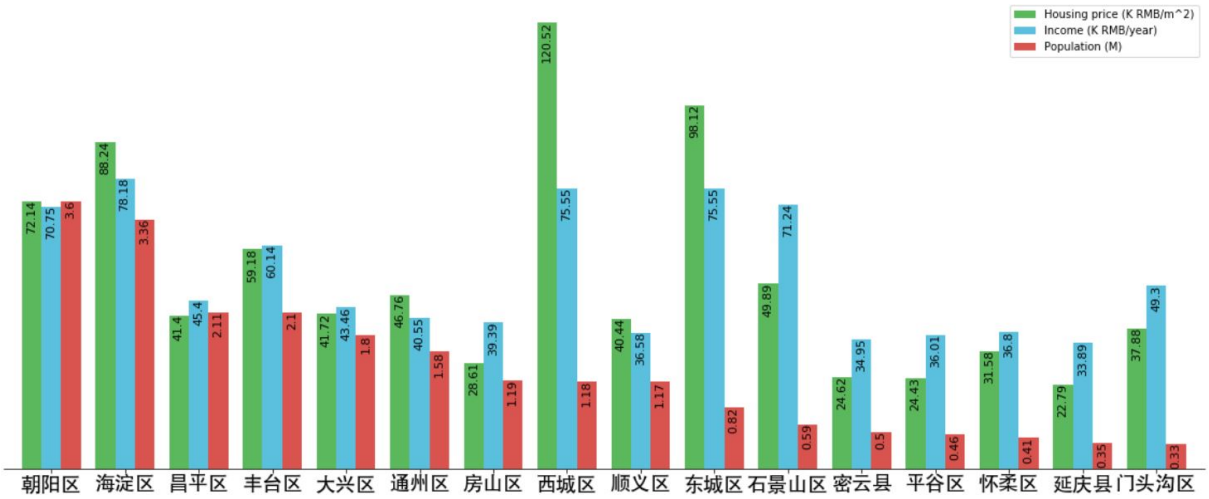

Fig. 2. Visualization of the data in Fig. 1, where 1). Chaoyang district (朝阳区) has the largest population among all 16 districts, 2). Xicheng district (西城区) has the most expensive housing price among all 16 districts, and 3). Haidian district (海淀区) has the highest annual income amonth all 16 districts.

in each district.

3.2.2. Choosing the number of groups for grouping

We deploy k-means to cluster the district (neighbourhood) into k clusters. In order to choose an appropirate group (cluster) number k for k-means clustering, we adopt the 'Elbow' method [5] that runs k-means clustering for a range of k's and for each value, we are calculating the sum of squared distances from each point to its assigned centroid (distortions). We then plot the distortions with respect to k and pick the point of inflection on the curve, *i.e.*, the 'elbow' point, that is the optimal k.

4. RESULTS

4.1. Population, Housing Price, Income and PHI index

We present the visualization results for the population (see Fig. 10), housing price (see. Fig. 11), income (see. Fig. 12), and the PHI index (see Fig. 13). It is clear that Haidian district (海淀区) and Chaoyang district (朝阳区) have the highest PHI indices among all 16 Beijing districts.

4.2. District grouping

As shown in Fig. 14, the optimal k is 2. Therefore, we set the number of clusters to be 2 for k-means clustering. See Fig. 15 for the visualized clustering results, where the colored circles represent the centroids of each district. Red ones represent Cluster 0 and purple ones represent Cluster 1. See also Figs. 16 and 17 for the details of Clusters 0 and 1, respectively.

For Cluster 0:

1. The most common venues are historic site, restaurants and coffee shops.
2. The district housing prices range between 22K and 98K RMB per square meters.
3. The clustered districts mainly lie in the center, west, north, and north-east regions of Beijing.

For Cluster 1:

1. The most common venues are coffee shops, restaurants and shopping malls.
2. The district housing prices range between 40K and 120K RMB per square meters.
3. The clustered districts mainly lie in the center and south-east regions of Beijing.

5. DISCUSSION

PHI index wise: the average housing price of Cluster 1 is much lower than Cluster 2. However, there is not much difference between Clusters 0 and 1 in terms of the annual income. People should consider the financial stress level before

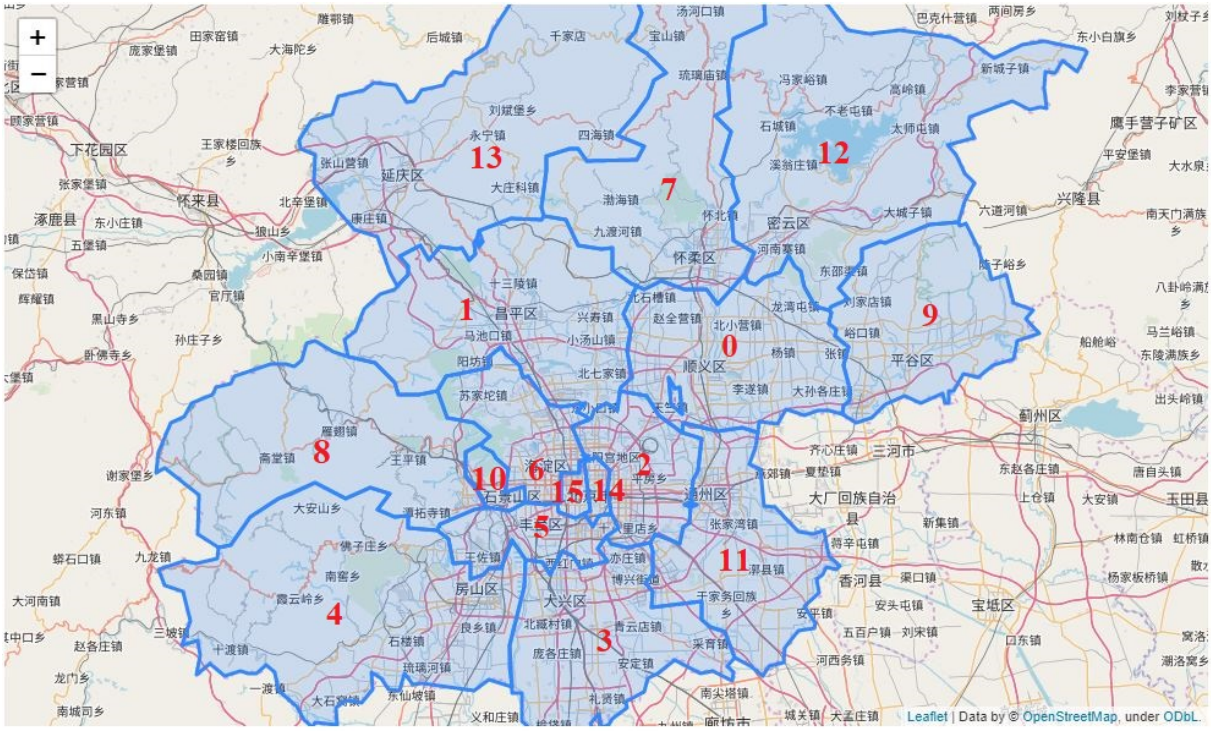

Fig. 3. Contour of Beijing districts. The numbers refer to the indices of the 16 Beijing districts in the dataframe in Fig. 1.

moving in one of these districts. Especially Haidian (海淀区) and Chaoyang (朝阳区) Districts that have very high PHI indices.

Living style wise: Cluster 1 generally has more historic sites than Cluster 0, in addition to the fact that Cluster 1 is closer to the center of Beijing than Cluster 0. Therefore, people who prefer 1) short-distance commuting during weekdays and 2) short-distance travel during weekends may consider move to the districts in Cluster 1. Since the gasoline/electric cost for the commuting/travel would be much lower than living in most of the districts in Cluster 0.

6. CONCLUSION

We analyzed the living/working charactistics of the 16 district of Beijing cities and provided settlement advice for millennials. We did this by the following processes: 1. We used the heuristically defined PHI index, that is based on the housing price, income, and population in 16 districts of Beijing cities, to indicate the potential, financial stress living in a Beijing district. 2. We used clustering results of the top 3 most common venues in each of the 16 Beijing districts to provide commuting/travel cost comparison.

We suggest that Haidian (海淀区) district may be the ideal district for people who can handle the relatively high financial stress and prefer short-distance travel for historic sites. In addition, we suggest that Chaoyang (朝阳区) district may be the ieal district for people who can handle the high financial stress and prefer day-to-day recreational activities rather than distance travel.

7. REFERENCES

[1] "National Bureau of Stats," http://www.stats.gov.cn/tjsj/, Accessed Feb. 13, 2020.

[2] "AnJuKe," https://www.anjuke.com/fangjia/beijing2020/, Accessed Feb. 13, 2020.

[3] "Beijing," https://github.com/bobchengyang/Coursera_Capstone/, Accessed Feb. 13, 2020.

[4] "Bejing json file," https://github.com/longwosion/geojson-map-china, Accessed Feb. 13, 2020.

[5] "K-means elbow method code for python," https://predictivehacks.com/k-means-elbow-method-code-for-python/, Accessed Feb. 13, 2020.

	id	PHI
2	朝阳区	3.675983
6	海淀区	3.790282
1	昌平区	1.922083
5	丰台区	2.071436
3	大兴区	1.723935
11	通州区	1.819410
4	房山区	0.862794
15	西城区	1.880919
0	顺义区	1.292660
14	东城区	1.067631
10	石景山区	0.413151
12	密云县	0.348756
9	平谷区	0.309318
7	怀柔区	0.355304
13	延庆县	0.234020
8	门头沟区	0.254323

Fig. 4. PHI indices of Beijing districts.

```
for beijing_df_current_idx in list(range(0,beijing_df.shape[0])):

    limit_candidate = 0
    radius_temp = 2000

    while limit_candidate < 50:

        try:

            beijing_venues = getNearbyVenues(names=beijing_df['id'][beijing_df_current_idx:beijing_df_current_idx+1],
                                latitudes=beijing_df['latitude'][beijing_df_current_idx:beijing_df_current_idx+1],
                                longitudes=beijing_df['longitude'][beijing_df_current_idx:beijing_df_current_idx+1],
                                radius=radius_temp
                                )
            limit_candidate = beijing_venues.shape[0]

        except ValueError:
            print('value error...' + 'radius_temp = ' + str(radius_temp))

        radius_temp += 2000

    print('limit_candidate = ' + str(limit_candidate) + ' | radius_temp = ' + str(radius_temp))

    if beijing_df_current_idx == 0:
        beijing_venues_final = beijing_venues
    else:
        beijing_venues_final = pd.concat([beijing_venues_final,beijing_venues])
```

Fig. 5. Python code for choosing an appropirate searching radius (with respect to the centroid of a district (neighbourhood)) to return the closest 50 venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	朝阳区	39.955555	116.520269	Maan Coffee	39.971858	116.504744	Café
1	朝阳区	39.955555	116.520269	East Beijing (东隅)	39.967369	116.485615	Hotel
2	朝阳区	39.955555	116.520269	Indigo (颐堤港)	39.968919	116.484762	Shopping Mall
3	朝阳区	39.955555	116.520269	Page One 叶壹堂	39.968382	116.484582	Bookstore
4	朝阳区	39.955555	116.520269	NUO Hotel Beijing (北京诺金酒店)	39.974946	116.474571	Hotel

Fig. 6. 5 closest venues in Chaoyang district (朝阳区).

	Neighbourhood	Airport	Airport Lounge	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bar	Beer Bar	Beer Garden	Beijing Restaurant	Bi Bo Stor
0	朝阳区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	朝阳区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	朝阳区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	朝阳区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	朝阳区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Fig. 7. Onehot encoded 5 closest venues in Chaoyang district(朝阳区).

	Neighbourhood	Airport	Airport Lounge	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bar	Beer Bar	Beer Garden	Beijing Restaurant	I E St
0	东城区	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0
1	丰台区	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
2	大兴区	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0
3	密云县	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.02	0
4	平谷区	0.00	0.00	0.04	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0
5	延庆县	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
6	怀柔区	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
7	房山区	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
8	昌平区	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
9	朝阳区	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0
10	海淀区	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0
11	石景山区	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
12	西城区	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0
13	通州区	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0
14	门头沟区	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.04	0.00	0.00	0.00	0
15	顺义区	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0

Fig. 8. Onehot encoded closest venues in each district.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	东城区	Historic Site	Hotel	Chinese Restaurant
1	丰台区	Coffee Shop	Hotel	Shopping Mall
2	大兴区	Coffee Shop	Park	Shopping Mall
3	密云县	Historic Site	Hotel	Resort
4	平谷区	Hotel	Historic Site	Chinese Restaurant
5	延庆县	Historic Site	Chinese Restaurant	Fast Food Restaurant
6	怀柔区	Historic Site	Resort	Ski Area
7	房山区	Historic Site	Park	Hotel
8	昌平区	Historic Site	Fast Food Restaurant	Coffee Shop
9	朝阳区	Hotel	Café	Japanese Restaurant
10	海淀区	Historic Site	Chinese Restaurant	Coffee Shop
11	石景山区	Fast Food Restaurant	Hotel	Historic Site
12	西城区	Coffee Shop	Chinese Restaurant	Fast Food Restaurant
13	通州区	Hotel	Shopping Mall	Park
14	门头沟区	Historic Site	Park	Hotel
15	顺义区	Coffee Shop	Fast Food Restaurant	Hotel

Fig. 9. Top 3 most common venue in each Beijing district.

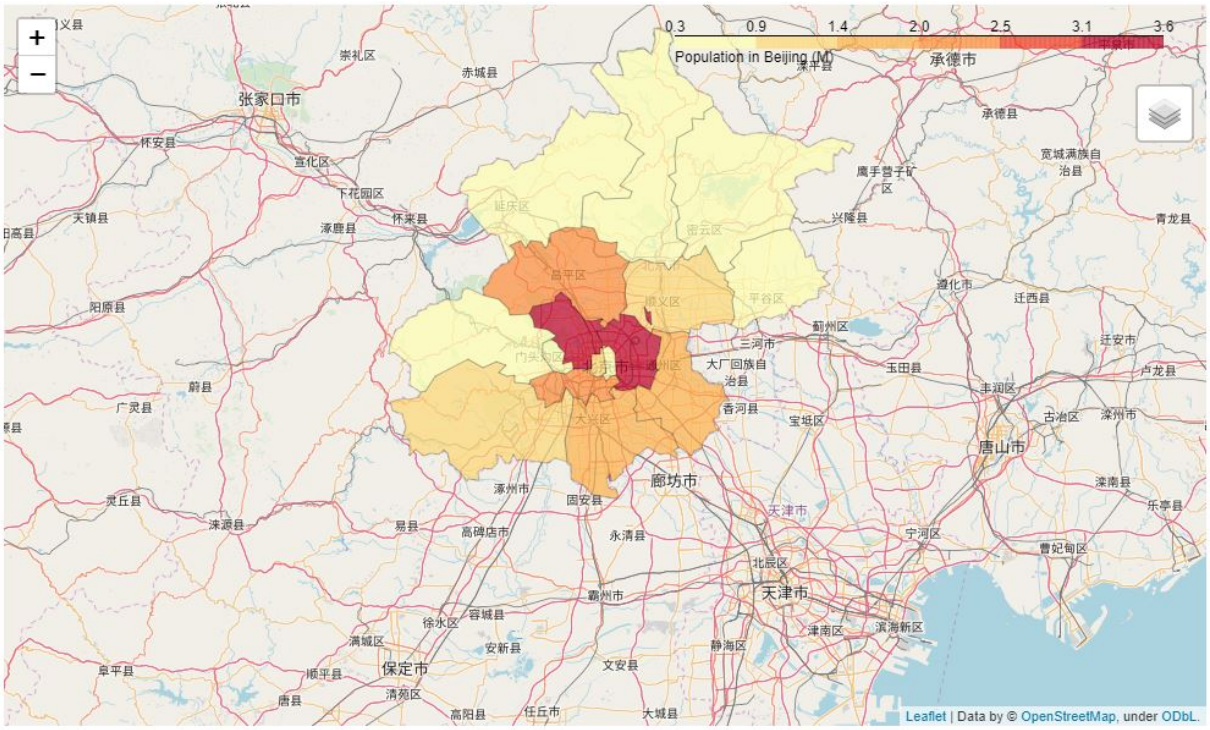

Fig. 10. Visualization of the population in Beijing districts.

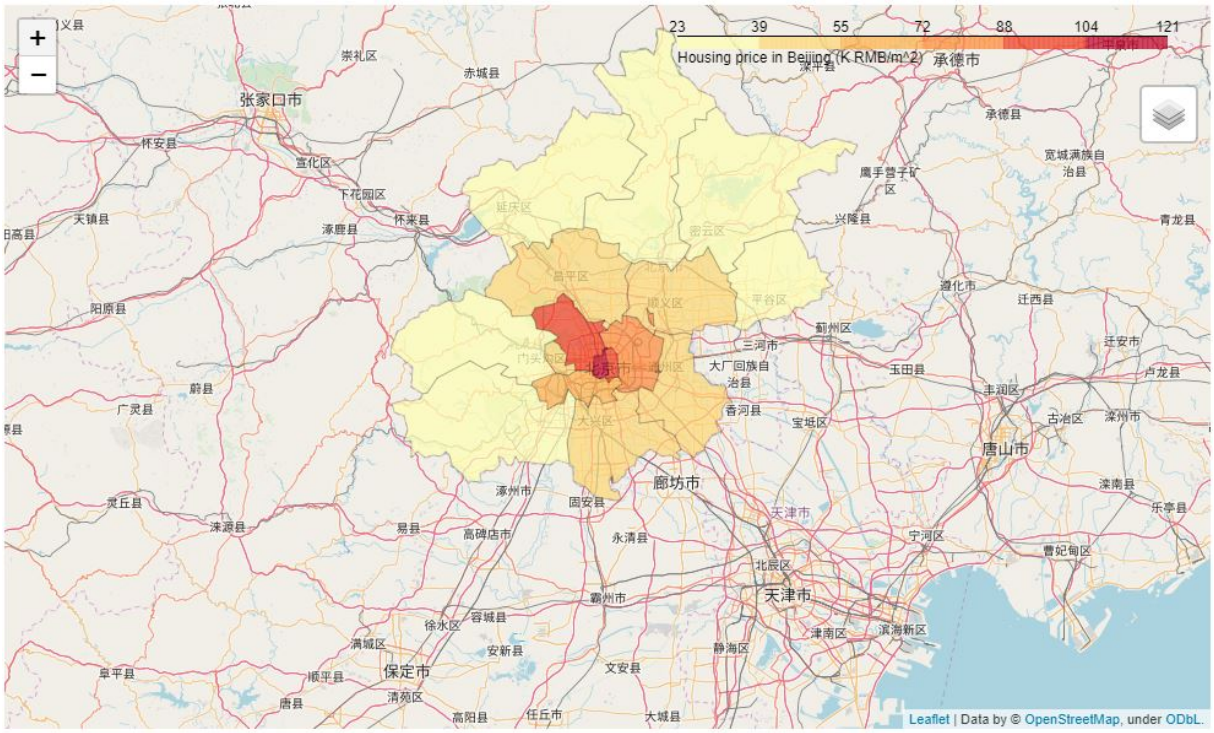

Fig. 11. Visualization of the housing price in Beijing districts.

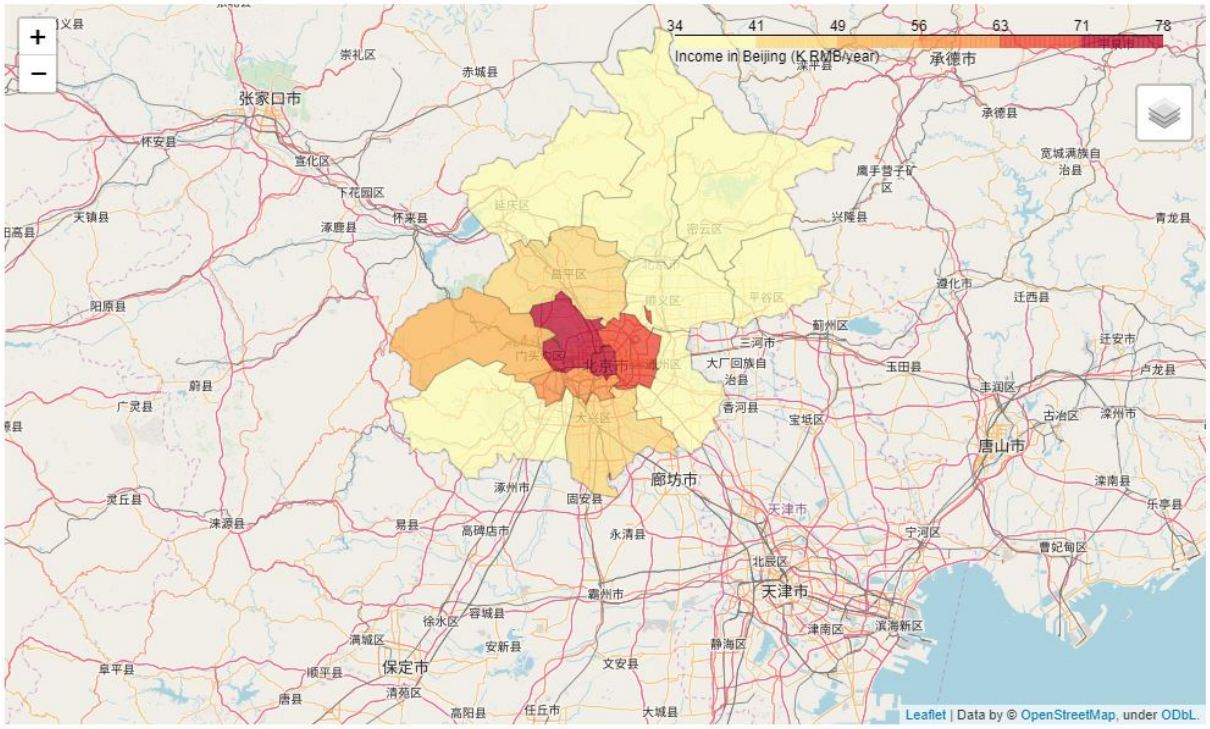

Fig. 12. Visualization of the income in Beijing districts.

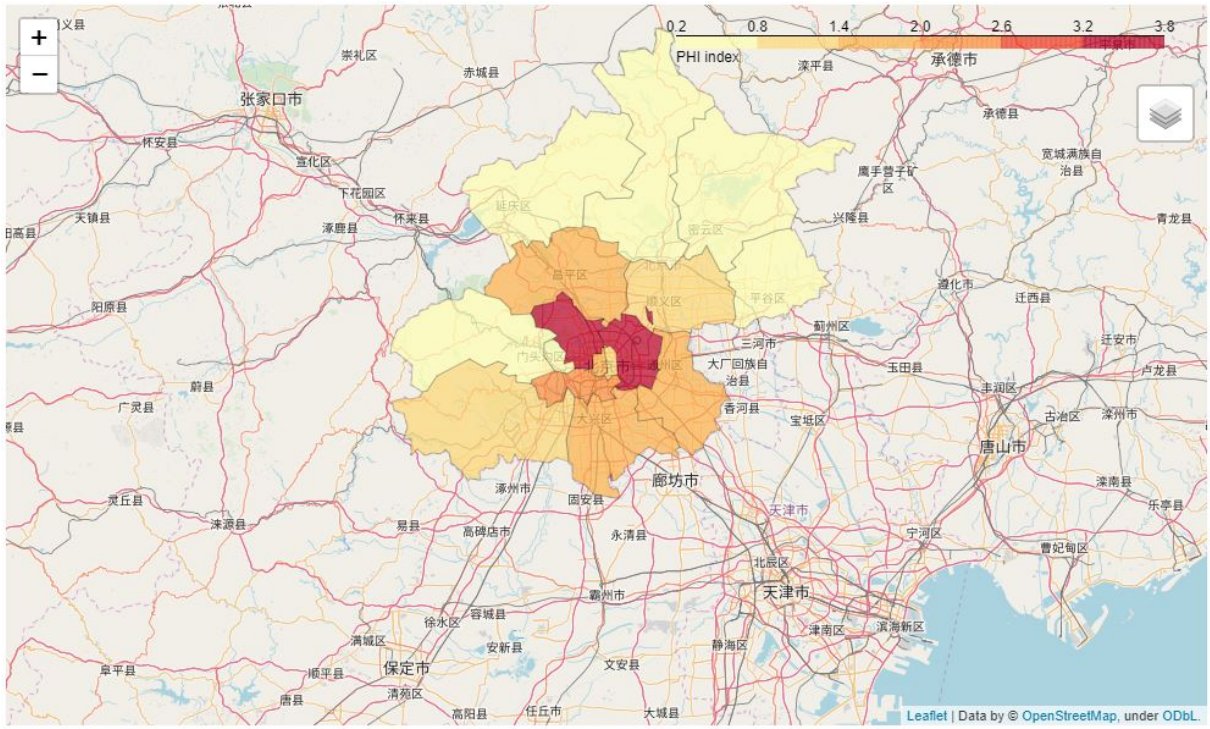

Fig. 13. Visualization of the PHI index in Beijing districts.

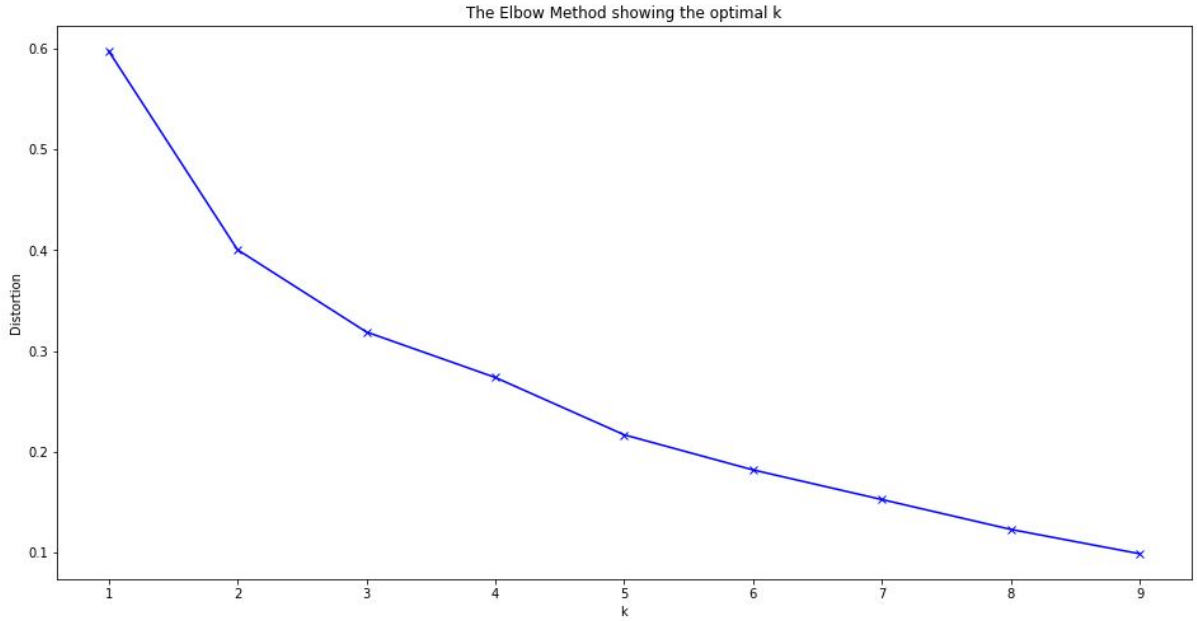

Fig. 14. 'Elbow' plot with 10 different number of clusters using k-means clustering. The 'elbow' appears to be at $k = 2$.

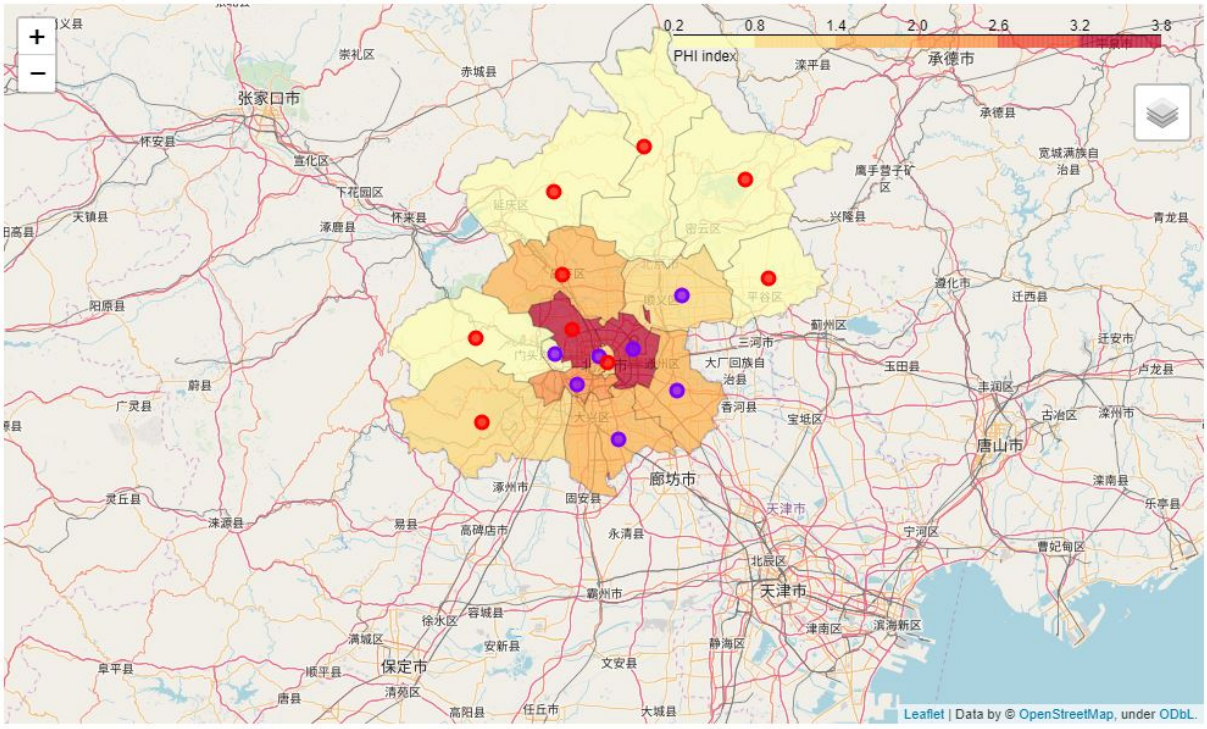

Fig. 15. Visualization of the clustering results, where the colored circles represent the centroids of each district. Red ones represent Cluster 0 and purple ones represent Cluster 1. Visualzation of the PHI index in Beijing districts is also presented.

	id	price	price_k	population	population_m	income	income_k	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
6	海淀区	88242	88.242	335.8	3.358	78178	78.178	40.024252	116.245290	0	Historic Site	Chinese Restaurant	Coffee Shop
1	昌平区	41395	41.395	210.8	2.108	45399	45.399	40.210785	116.201086	0	Historic Site	Fast Food Restaurant	Coffee Shop
4	房山区	28608	28.608	118.8	1.188	39391	39.391	39.701904	115.838850	0	Historic Site	Park	Hotel
14	东城区	98122	98.122	82.2	0.822	75547	75.547	39.909897	116.404088	0	Historic Site	Hotel	Chinese Restaurant
12	密云县	24625	24.625	49.5	0.495	34951	34.951	40.540558	117.026101	0	Historic Site	Hotel	Resort
9	平谷区	24428	24.428	45.6	0.456	36012	36.012	40.198224	117.131640	0	Hotel	Historic Site	Chinese Restaurant
7	怀柔区	31580	31.580	41.4	0.414	36797	36.797	40.649906	116.567993	0	Historic Site	Resort	Ski Area
13	延庆县	22788	22.788	34.8	0.348	33887	33.887	40.495297	116.164841	0	Historic Site	Chinese Restaurant	Fast Food Restaurant
8	门头沟区	37878	37.878	33.1	0.331	49298	49.298	39.994308	115.811084	0	Historic Site	Park	Hotel

Fig. 16. Cluster 0.

	id	price	price_k	population	population_m	income	income_k	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	朝阳区	72139	72.139	360.5	3.605	70746	70.746	39.955555	116.520269	1	Hotel	Café	Japanese Restaurant
5	丰台区	59185	59.185	210.5	2.105	60144	60.144	39.832243	116.266229	1	Coffee Shop	Hotel	Shopping Mall
3	大兴区	41720	41.720	179.6	1.796	43464	43.464	39.642946	116.453457	1	Coffee Shop	Park	Shopping Mall
11	通州区	46757	46.757	157.8	1.578	40553	40.553	39.812867	116.720561	1	Hotel	Shopping Mall	Park
15	西城区	120524	120.524	117.9	1.179	75547	75.547	39.932061	116.364889	1	Coffee Shop	Chinese Restaurant	Fast Food Restaurant
0	顺义区	40444	40.444	116.9	1.169	36575	36.575	40.142372	116.738176	1	Coffee Shop	Fast Food Restaurant	Hotel
10	石景山区	49889	49.889	59.0	0.590	71244	71.244	39.939716	116.169618	1	Fast Food Restaurant	Hotel	Historic Site

Fig. 17. Cluster 1.