The Battle of Neighborhoods

Meng Yu

March 22, 2019

1. Introduction

1.1 Background

Jinan and Qingdao are the two largest cities in China's Shandong province. There are many similarities and differences between the two cities. To be specific, Jinan is the provincial capital city with a long history and culture and a convenient high-speed rail transportation network. Qingdao is a coastal open city with a better environment and climate, a huge port, developed foreign trade, tourism and financial industries. Jinan and Qingdao have long been a hot topic of discussion. This paper hopes to make an intuitive demonstration of the differences between the two cities through the data science method taught in the course. The following is the basic situation of jinan and Qingdao.

Table 1. Basic situation of Jinan and Qingdao

	Jinan	Qingdao
Prefecture-level area	10244km^2	11067km^2
Population	7321200	9290500
GDP	￥886.2 billion	￥1256.1 billion
Human Development Index	0.811	0.822
Position	36.38N 117.01E	36.5N 120.2E

1.2 Problem

This paper will collect data from Foursquare, list the number of venues owned by each borough of Jinan and Qingdao, and compare them with charts. In addition, they are classified by k-means clustering method, and these clusters are displayed on the map.

1.3 Interest

For those interested in urban governance, this article may be useful to them. The geographical differences between the two cities can be found through the study of the urban boroughs and the venues around the neighborhoods. For those who wish to settle

down in a certain city or a certain area of a certain city, it can also provide reference value for them.

2. Data acquisition and cleaning

2.1 Data sources

There are 12 boroughs and 161 neighborhoods in Jinan, and 10 boroughs and 133 neighborhoods in Qingdao. These can be found on their respective government websites. To get their location, I used the open API provided by Autonavi, a Chinese mapping and geographic information service. The venues near each neighborhood is available through Foursquare's free API. In order to create a choropleth map, I use a json download service provided by Aliyun.

2.2 Data cleaning

Fortunately, I was able to get the data I wanted from Foursquare, Autonavi, and Aliyun. The data is very neat, with no missing values and no outliers. I simply took this data in Python and wrote it to a local CSV file so that I could call it at any time. For details you can examine the complete code in my jupyter notebook. Therefore, there is no special data cleaning required.

2.3 Feature selection

As I said earlier, after getting the data I wanted in Python, I stored it in a local CSV file. This prevents me from rerunning the program every time I get this data. This saves time.

	Borough	Neighborhood	Latitude	Longitude
0	青岛市市南区	八大峡街道办事处	36.056636	120.298179
1	青岛市市南区	云南路街道办事处	36.067401	120.305969
2	青岛市市南区	中山路街道办事处	36.065414	120.322654
3	青岛市市南区	八大关街道办事处	36.058097	120.350388
4	青岛市市南区	湛山路街道办事处	36.054171	120.362353
5	青岛市市南区	香港中路街道办事处	36.071735	120.391082
6	青岛市市南区	八大湖街道办事处	36.079966	120.389510
7	青岛市市南区	金门路街道办事处	36.071063	120.413246
8	青岛市市南区	珠海路街道办事处	36.061707	120.407954

The df_qingdao_geo.csv and df_jinan_geo.csv stores the latitude and longitude of each neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	八大峡街道办事处	36.056636	120.298179	船歌鱼水饺	36.054346	120.298572	Dumpling Restaurant
1	八大峡街道办事处	36.056636	120.298179	前海购物广场 Qianhai Shopping Center	36.057827	120.303856	Shopping Mall
2	八大峡街道办事处	36.056636	120.298179	友誼遊艇 Youyi Yacht	36.052653	120.301900	Harbor / Marina
3	八大峡街道办事处	36.056636	120.298179	飛洋遊艇碼頭 Feiyang Yacht	36.052678	120.302035	Harbor / Marina
4	八大峡街道办事处	36.056636	120.298179	中苑旅游码头	36.050800	120.299210	Pier
5	八大峡街道办事处	36.056636	120.298179	八大峡广场	36.052673	120.304842	Plaza
6	八大峡街道办事处	36.056636	120.298179	团岛农贸市场	36.063656	120.296617	Farmers Market
7	八大峡街道办事处	36.056636	120.298179	金茂湾购物中心	36.065249	120.296662	Shopping Plaza
8	云南路街道办事处	36.067401	120.305969	St. Michael's Cathedral 圣弥爱尔大教堂	36.067386	120.315366	Church

The qingdao_venues.csv and jinan_venues.csv stores the venue, vanue latitude, venue longitude and venue category.

3. Exploratory Data Analysis

3.1 Descriptive statistics

According to Foursquare, there are 458 venues in Jinan and 297 venues in Qingdao. For Jinan, shizhong borough and lixia borough have more venues, while for Qingdao, shinan borough and shibei borough have most venues. This is much the same as our experience. Because Jinan shizhong, lixia boroughs and Qingdao shinan, shibei boroughs are the most populous and prosperous areas. At the same time, we can find that, in terms of the number of venues, regional differences are relatively small in Jinan, while large in Qingdao.

Because matplotlib can not display Chinese very well, I used plotly for Python to generate the chart below.

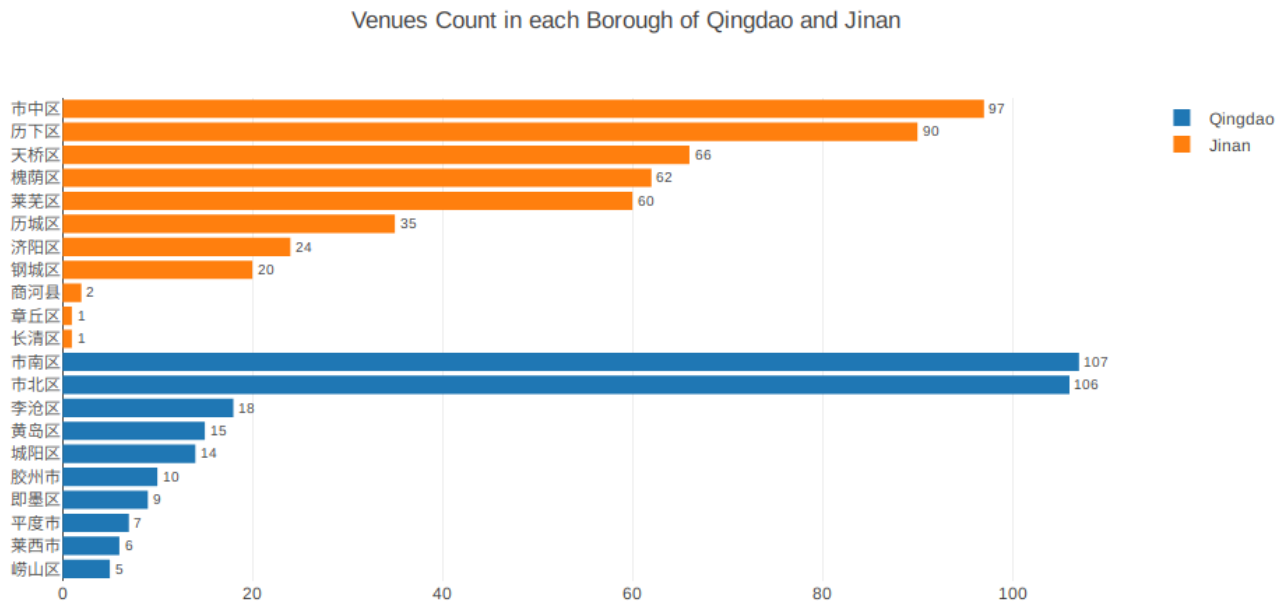

Figure 1. Venues count in each borough of Jinan and Qingdao

I continue to check the types of venues that have the most venues respectively in Jinan and Qingdao. As you can see, the most venues in both cities are hotels, Chinese restaurants and shopping mall. At the same time, Jinan has more movie theater , soccer fields and basketball stadium.

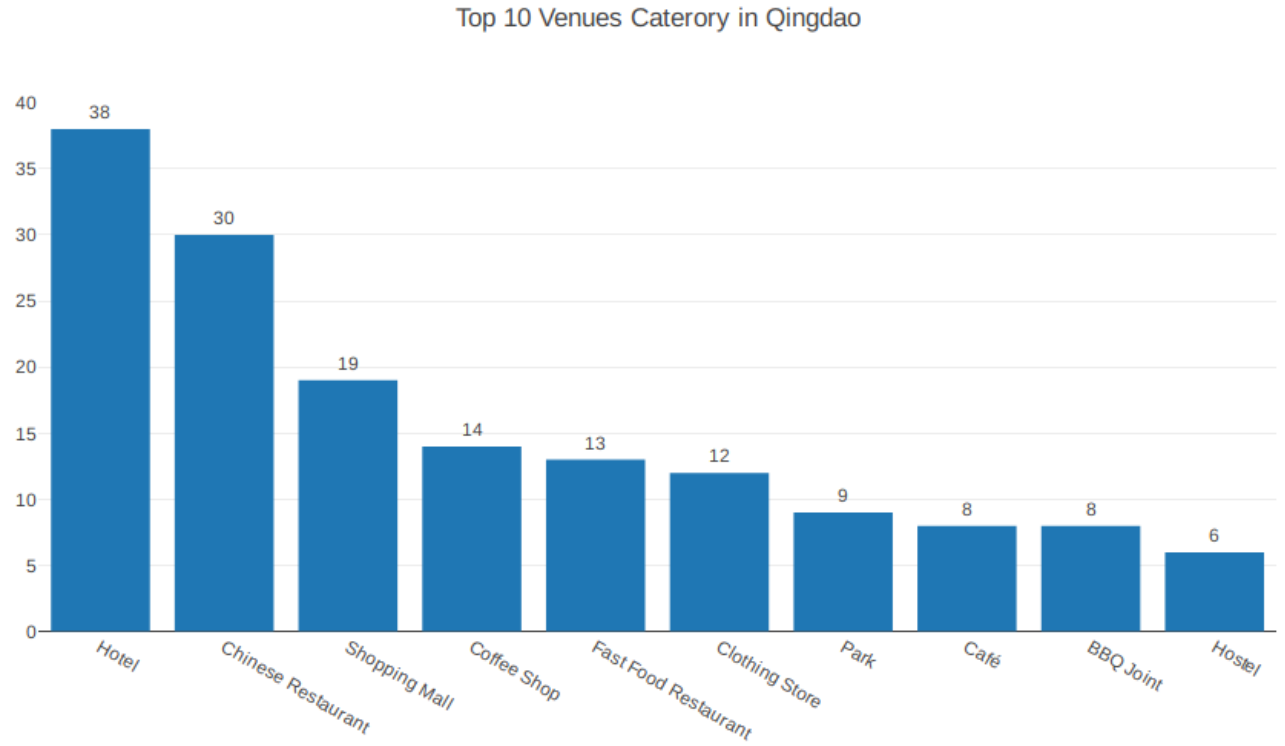

Figure 2. Top 10 venues category in Qingdao

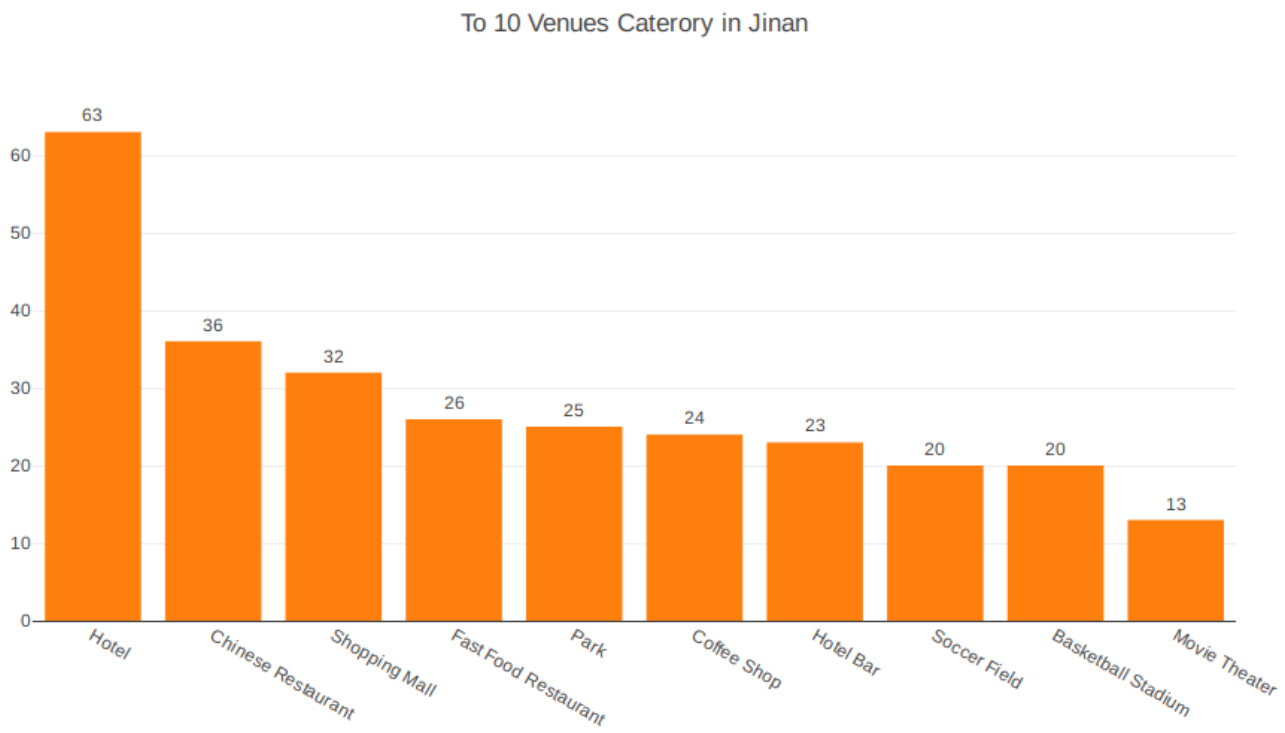

Figure 3. Top 10 venues category in Jinan

3.2 K-means Cluster

In order to find the best k value in K-means cluster, I use elbow method. In general, it is helpful to determine the appropriate value of k by drawing the broken line graph of k and SSE. If you're at a point where SSE goes down very fast from the beginning, and then it goes down very slowly after that point, then this is the appropriate k value. However, sometimes this trend is not so obvious, so it is difficult to judge the value of k. At this point, it is necessary to combine the practical meaning of the problem to choose an appropriate value of k. Here, for the clustering of venues in Qingdao, I selected k = 6.

Similarly, I also used elbow method in the process of clustering venues in Jinan. Unfortunately, the downward trend of SSE is also less pronounced. Well, it looks like k=4 is a good value, and for convenience I chose k=4.

For data preprocessing, I directly adopted the method of course teaching. Finally, the 1st to 10th venues were obtained for each community in each city. The code is available on the jupyter notebook.

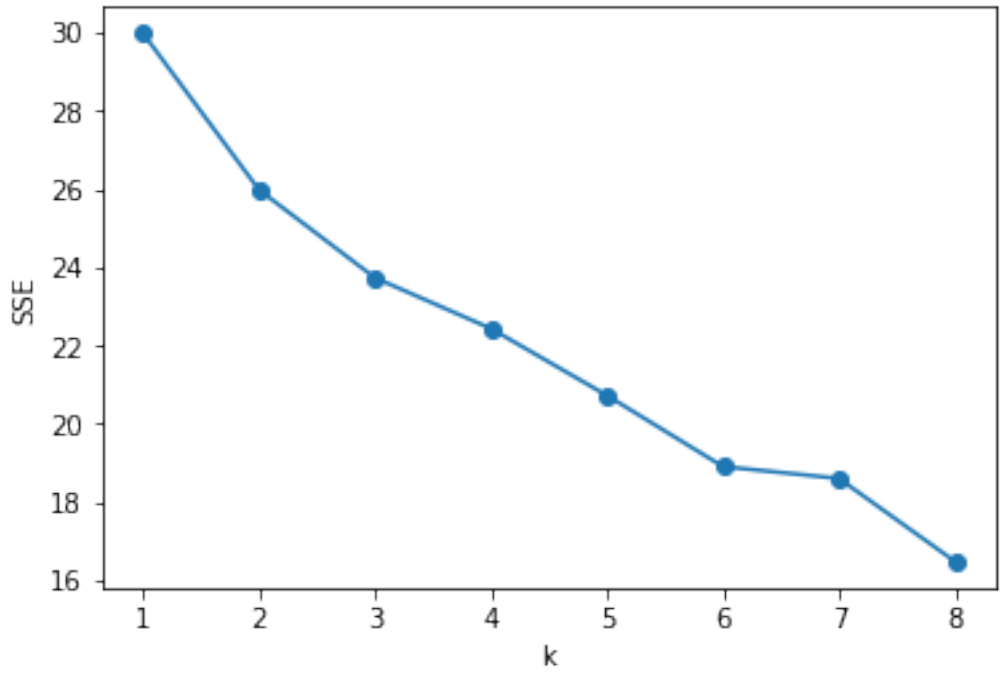

Figure 4. Elbow method for K-means cluster of Qingdao venues

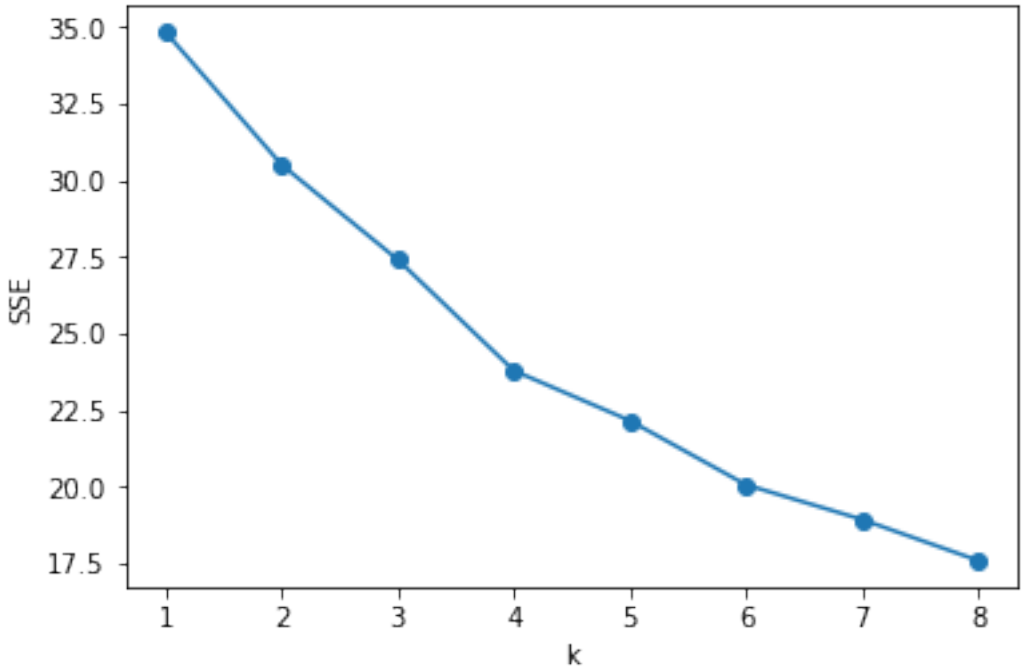

Figure 5. Elbow method for K-means cluster of Jinan venues

Finally, I have obtained the results of clustering venues in Jinan and Qingdao. The sample number of each cluster is as follows.

Table 2. Counts in each cluster in Jinan and Qingdao

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Jinan	26	12	56	5	NA	NA
Qingdao	5	6	41	2	9	2

3.3 Visualizing Geospatial Data

I chose to display the results of k-means clustering on the map. Different colors of each point in the map represent different clustering. For each Borough, the more venues in the region, the closer the color is to red; while the fewer venues in the region, the closer to yellow.

As we have seen, venues in each city cluster in their own city centres. Specifically, venues are clustered in shinan borough and shibei borough, while in Jinan, venues are clustered in shizhong borough and lixia borough.

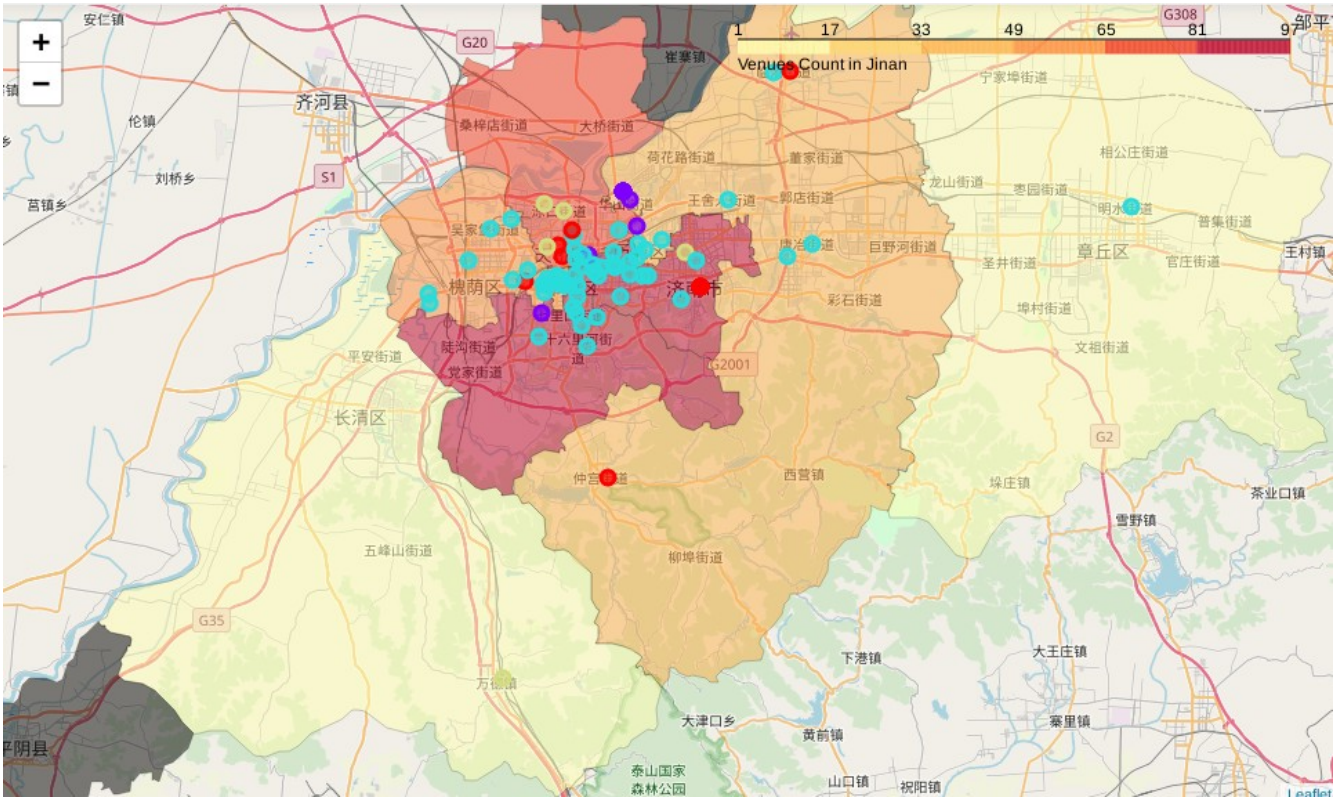

Figure 6. Choropleth and cluster map of Jinan

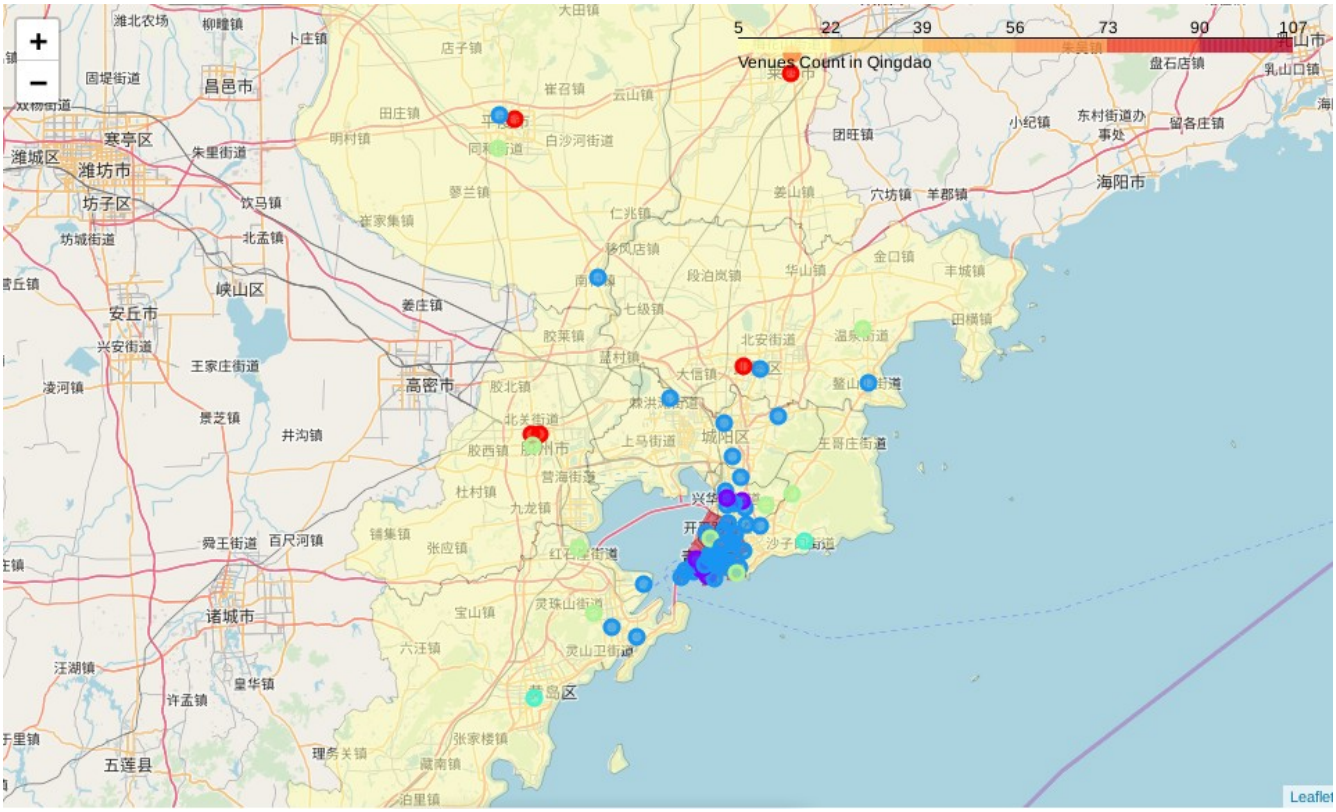

Figure 7. Choropleth and cluster map of Qingdao

4. Conclusions

In this study, I collected Foursquare data of Jinan and Qingdao. Finally, I found that Jinan have more venues, and the differences between boroughs are small. However, Qingdao has fewer venues, with great differences among boroughs. Venues in both cities are concentrated in their respective city centres.

5. Future Directions

There is a gap between my research and people's experience in the real world. Generally speaking, people tend to think that Qingdao is more prosperous, with more population and more venues. Also, laoshan borough, located in a busy area, should have more venues. I think this is because Foursquare, as a foreign service, cannot fully reflect the actual situation in China. In addition, the results of clustering based on k-means are not ideal, which need to be improved in the future.