Analysis of Venus from Districts of Shanghai and Beijing and A Tentative Clustering Scheme of Those Districts

Yuhui Cao

May 28, 2020

1. Introduction

1.1 Background

During the past 3 decades, China has been the fastest developing country in terms of culture, economy and education. Shanghai, the center of Chinese economy, and Beijing, the capital city of China, are certainly the two main metropolises that catch eyes of people from all over the world. During the past 3 decades, urban expansion, new immigrants and population influx have been the main issues of both cities. With such issues, there are other new topics presenting to not only government, city planners and entrepreneurs from both internal and external companies, but also original residents of the cities and incoming immigrants. In order to have a more appropriate resource allocation, we need to dig deep into the venues data and give some insights to the publics.

1.2 Problem

In order to have a more appropriate resource allocation, and due to the lack of information of each district and the lack of clustering scheme of all the districts from the two cities, this project, aims here to analyzing venues of each districts and try to cluster those districts based on the similarity of common venues. The venues information will be acquired from Foursquare website, by providing location information of a specific district.

1.3 Interest

Obviously, government, city planners, entrepreneurs, original residents and even new immigrants are interested in the developmental information of each district and comparison between two cities. We can give out some insight information about those topics by using one of the most commonly used unsupervised machine learning techniques, K-Means clustering, on venues information of each district from the two main metropolises of China.

2. Data acquisition and cleaning

2.1 Data sources

The data of latitude and longitude of districts of Shanghai and Beijing can be found at https://zh.wikipedia.org/wiki/Template:上海行政区划地图 and https://zh.wikipedia.org/wiki/北

京市行政區劃, respectively. I found 19 districts for Beijing originally, including Beijng downtown, Dongcheng, Xicheng, Chongwen, Xuanwu, Chaoyang, Fengtai, Shijingshan, Haidian, Mentougou, Fanshang, Tongzhou, Shunyi, Changping, Daxing, Huairou, Pinggu, Miyun and Yanqing. I found 18 districts for Shanghai originally, including Huangpu, Xuhui, Changning, Jingan, Putuo, Zhabei, Hongkou, Yangpu, Minhang, Baoshan, Jiading, Pudong, Jinshan, Songjiang, Qingpu, Nanhui, Chongming and Fengxian. Later I will check those district on Foursquare to see if they have enough venues information and decide whether to keep them in the future clustering analysis or not.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were no missing values of the original data. However, there are still one problem with the datasets. I checked the venues around each district on Foursquare website https://foursquare.com, a location technology platform dedicated to improving how people move through the real world. I get venues information of a specific longitude and latitude by getting access to Foursquare database and defining a radius around that location. However, not every district has venues information within its range of location. This will have influence on the later cluster analysis. So at last, I reluctantly decided to delete those districts.

2.3 Area of Interest selection

Below is the information of dropped districts due to the lack of venues information from Foursquare website. Finally, we are going to keep 13 and 8 districts for Shanghai and Beijing, respectively, to do the following venues analysis and tentative clustering, as shown in **Table 1**.

Table 1. Data selection

Cities	Kept Districts	Dropeed Districts	Reason for dropping
Shanghai	13, Shanghai (downtown), Huangpu, Xuhui, Changning, Jingan, Putuo, Zhabei, Hongkou, Yangpu, Minhang, Jiading, Songjiang, Nanahui	5, Baoshan, Jinshan, Qingpu, Chongming, Fengxian	Lack of venues information from Foursquare website
Beijing	8, Beijing (downtown), Dongcheng, Xicheng, Chongwen, Xuanwu, Chaoyang, Haidian, Changping	11, Fengtai, Shijingshan, Mentougou, Fangshan, Tongzhou, Shunyi, Daxing, Huairou, Pinggu, Miyun, Yanqing	

3. Methodology (Shanghai vs. Beijing, separately)

3.1. District Visualization

First of all, we tried to scrape data from the internet and got the following original data, which contains 13 and 8 districts for Shanghai and Beijing, respectively (as shown in **Figure1**).

	Postal Code	City	District	Longitude	Latitude
0	310100	Shanghai	Shanghai	121.487899	31.249162
1	310101	Shanghai	Huangpu	121.496072	31.227203
2	310104	Shanghai	Xuhui	121.446235	31.169152
3	310105	Shanghai	Changning	121.387616	31.213301
4	310106	Shanghai	Jingan	121.454756	31.235381

```
1 Shanghai_df.shape
```

```
(13, 5)
```

	Postal Code	City	District	Longitude	Latitude
0	110100	Beijing	Beijing	116.395645	39.929986
1	110101	Beijing	Dongcheng	116.421885	39.938574
2	110102	Beijing	Xicheng	116.373190	39.934280
3	110103	Beijing	Chongwen	116.424636	39.889292
4	110104	Beijing	Xuanwu	116.369352	39.891531

```
1 Beijing_df.shape
```

```
(8, 5)
```

Figure 1. Original Data shows Postal Code, name of the city, Longitude and Latitude of 13 and 8 districts for Shanghai vs. Beijing, respectively.

Then we plot those district on map by using folium library, to get a first impression of each location for Shanghai and Beijing (as shown in **Figure 2**).

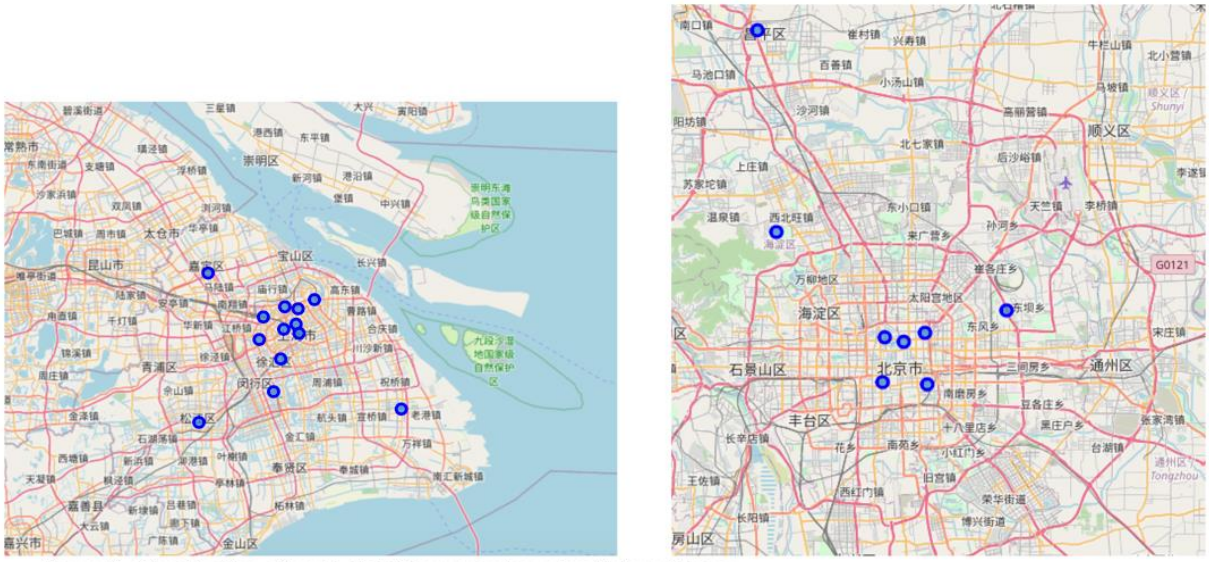


Figure 2. Visualization of each district on map by using Folium library.

3.2. Foursquare Venues

Then we get venues information for each district from both Shanghai and Beijing, separately. As you can see in the following **Figure 3**, we got 115 and 70 venues for the two cities, respectively.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Shanghai	31.249162	121.487899	Bund Club	31.247147	121.488660	Lounge
1	Shanghai	31.249162	121.487899	Hyatt on the Bund Shanghai	31.247206	121.488571	Hotel
2	Shanghai	31.249162	121.487899	Xindalu (新大陆)	31.247296	121.488764	Shanghai Restaurant
3	Shanghai	31.249162	121.487899	Vue Bar	31.246909	121.488319	Hotel Bar
4	Shanghai	31.249162	121.487899	Gym	31.247186	121.488609	Gym

```
1 Shanghai_venues.shape
```

```
(115, 7)
```

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beijing	39.929986	116.395645	TRB Hutong	39.926528	116.397204	French Restaurant
1	Beijing	39.929986	116.395645	The Georg by Georg Jensen	39.933440	116.392740	Scandinavian Restaurant
2	Beijing	39.929986	116.395645	City Walls Courtyard House Beijing	39.928131	116.392442	Hostel
3	Beijing	39.929986	116.395645	Peking Hostel (北平国际青年旅舍)	39.934350	116.396886	Hostel
4	Beijing	39.929986	116.395645	Nanluogu Alley (南锣鼓巷)	39.932498	116.396925	Pedestrian Plaza

```
1 Beijing_venues.shape
```

```
(70, 7)
```

Figure 3. 115 and 70 venues were obtained for Shanghai and Beijing, respectively (from Foursquare).

In **Figure 4**, we counted the unique "venue category" and then group it by districts. There are 61 and 38 unique venue categories for Shanghai and Beijing, respectively.

```
1 Shanghai_venues.groupby('District').count()
```

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Changning	20	20	20	20	20	20
Hongkou	6	6	6	6	6	6
Huangpu	14	14	14	14	14	14
Jiading	4	4	4	4	4	4
Jingan	26	26	26	26	26	26
Minhang	1	1	1	1	1	1
Nanhui	1	1	1	1	1	1
Putuo	4	4	4	4	4	4
Shanghai	23	23	23	23	23	23
Songjiang	2	2	2	2	2	2
Xuhui	6	6	6	6	6	6
Yangpu	4	4	4	4	4	4
Zhabei	4	4	4	4	4	4

Let's find out how many unique categories can be curated from all the returned venues

```
1 print('There are {} uniques categories.'.format(len(Shanghai_venues['Venue Category'].unique())))
```

```
There are 61 uniques categories.
```

```
1 Beijing_venues.groupby('District').count()
```

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Beijing	17	17	17	17	17	17
Changping	5	5	5	5	5	5
Chaoyang	2	2	2	2	2	2
Chongwen	5	5	5	5	5	5
Dongcheng	24	24	24	24	24	24
Haidian	1	1	1	1	1	1
Xicheng	9	9	9	9	9	9
Xuanwu	7	7	7	7	7	7

Let's find out how many unique categories can be curated from all the returned venues

```
1 print('There are {} uniques categories.'.format(len(Beijing_venues['Venue Category'].unique())))
```

```
There are 38 uniques categories.
```

Figure 4. 61 and 38 unique venues categories ("Venue Category") were obtained for Shanghai and Beijing, Respectively.

Then we did onehot encoding of "Venue Category" for Shanghai and Beijing in **Figure 5**.

```
The dimensionality of Shanghai_onehot dataset is (115, 62)
```

	District	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Bakery	Bistro	Breakfast Spot	Burger Joint	Bus Station	...	Spa	Sporting Goods Shop	Steakhouse	Sushi Restaurant	Tea Room	Theater	B
0	Shanghai	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	Shanghai	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	Shanghai	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	Shanghai	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	Shanghai	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

5 rows × 62 columns

```
The dimensionality of Beijing_onehot dataset is (70, 39)
```

	District	Asian Restaurant	Auto Garage	Bakery	Bar	Beer Bar	Beijing Restaurant	Bus Station	Café	Cantonese Restaurant	...	Park	Pedestrian Plaza	Pizza Place	Restaurant	Russian Restaurant	Sandwich Place	Sca R
0	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	Beijing	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	

5 rows × 39 columns

Figure 5. Onehot encoding for "Venue Category".

And then group by "District" and calculate the mean for each district, as shown in **Figure 6**.

```
Shanghai_grouped = Shanghai_onehot.groupby('District').mean().reset_index()
Shanghai_grouped
```

	District	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Bakery	Bistro	Breakfast Spot	Burger Joint	Bus Station	...	Spa	Sporting Goods Shop	Steakhouse	Sushi Restaurant	Tea Room
0	Changning	0.000000	0.05	0.00	0.05	0.000000	0.000000	0.000000	0.05	0.00	...	0.00	0.000000	0.000000	0.000000	0.05
1	Hongkou	0.000000	0.00	0.00	0.00	0.166667	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00
2	Huangpu	0.071429	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.071429	0.000000	0.00
3	Jiading	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.25	0.000000	0.000000	0.000000	0.00
4	Jingan	0.000000	0.00	0.00	0.00	0.000000	0.038462	0.000000	0.00	0.00	...	0.00	0.038462	0.000000	0.000000	0.00
5	Minhang	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00
6	Nanhui	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00
7	Putuo	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00
8	Shanghai	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.043478	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00
9	Songjiang	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00
10	Xuhui	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.166667	0.00
11	Yangpu	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.25	...	0.00	0.000000	0.000000	0.000000	0.00
12	Zhabei	0.000000	0.00	0.25	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.00	0.000000	0.000000	0.000000	0.00

13 rows × 62 columns

Let's confirm the new size

```
Shanghai_grouped.shape
```

```
(13, 62)
```

```
Beijing_grouped = Beijing_onehot.groupby('District').mean().reset_index()
Beijing_grouped
```

	District	Asian Restaurant	Auto Garage	Bakery	Bar	Beer Bar	Beijing Restaurant	Bus Station	Café	Cantonese Restaurant	...	Park	Pedestrian Plaza	Pizza Place	Restaurant	Rus Restau
0	Beijing	0.0	0.0	0.0	0.058824	0.000000	0.058824	0.058824	0.058824	0.000000	...	0.0	0.058824	0.000000	0.058824	0.
1	Changping	0.2	0.0	0.2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000	0.000000	0.000000	0.
2	Chaoyang	0.0	0.5	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.5	0.000000	0.000000	0.000000	0.
3	Chongwen	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000	0.000000	0.000000	0.
4	Dongcheng	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.041667	...	0.0	0.000000	0.041667	0.000000	0.
5	Haidian	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000	0.000000	0.000000	0.
6	Xicheng	0.0	0.0	0.0	0.000000	0.111111	0.111111	0.000000	0.000000	0.000000	...	0.0	0.000000	0.000000	0.000000	0.
7	Xuanwu	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000	0.000000	0.000000	0.

8 rows × 39 columns

Let's confirm the new size

```
Beijing_grouped.shape
```

```
(8, 39)
```

Figure 6. Averaged "Venue Category" for each district.

Then we did the final arrangement to sort out the top 10 most common venues for each district, as shown in **Figure 7** below. This dataset is going to be used to ran the final clustering analysis.

```
22 District_venues_sorted.head()
```

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Changning	Chinese Restaurant	Hotel	Japanese Restaurant	Dongbei Restaurant	Gym Pool	Art Museum	Athletics & Sports	Tea Room	Grocery Store	Convenience Store
1	Hongkou	Ramen Restaurant	Bakery	Chinese Restaurant	Café	Convenience Store	Pool	Furniture / Home Store	French Restaurant	Grocery Store	Farmers Market
2	Huangpu	Clothing Store	Hotel	Hotel Bar	Shanghai Restaurant	Business Center	Café	Coffee Shop	Furniture / Home Store	New American Restaurant	Scenic Lookout
3	Jiading	Spa	Dumpling Restaurant	Chinese Restaurant	Noodle House	Wedding Hall	Convenience Store	Gym / Fitness Center	Gym	Grocery Store	Furniture / Home Store
4	Jingan	Coffee Shop	Chinese Restaurant	Hotel	Convenience Store	French Restaurant	Fast Food Restaurant	History Museum	Vegetarian / Vegan Restaurant	Hotpot Restaurant	Italian Restaurant

Above is each district along with the top 10 most common venues

```
22 District_venues_sorted.head()
```

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Beijing	Hostel	Hotel	Hotpot Restaurant	Metro Station	Pedestrian Plaza	Chinese Restaurant	Bus Station	Beijing Restaurant	Café	Restaurant
1	Changping	Asian Restaurant	Chinese Restaurant	Shopping Mall	Bakery	Korean Restaurant	Fast Food Restaurant	Dessert Shop	Department Store	Convenience Store	Coffee Shop
2	Chaoyang	Park	Auto Garage	Cantonese Restaurant	Fast Food Restaurant	Dessert Shop	Department Store	Convenience Store	Coffee Shop	Cocktail Bar	Chinese Restaurant
3	Chongwen	Chinese Restaurant	Convenience Store	Fast Food Restaurant	Metro Station	Supermarket	Beer Bar	Beijing Restaurant	Bus Station	Café	Cantonese Restaurant
4	Dongcheng	Coffee Shop	Russian Restaurant	Vegetarian / Vegan Restaurant	Hotel	Japanese Restaurant	Chinese Restaurant	Cantonese Restaurant	Fast Food Restaurant	Grocery Store	Cocktail Bar

Above is each district along with the top 10 most common venues

Figure 7. Top 10 most common venues for each district. This data is going to be used for later clustering analysis.

3.3. Clustering Model

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances). For instance, better Euclidean solutions can be found using k-medians and k-medoids. It is easy to apply to learn the inner structure of a dataset without being explicitly programming, so we choose this algorithm in our study.

3.4. Cluster Evaluation

In order to choose the right K number, I run "Elbow" analysis to calculate SSE for each K. As shown below in **Figure 8**, I chose K = 5 and K = 4 for Shanghai and Beijing, respectively, to run the final K means clustering analysis.

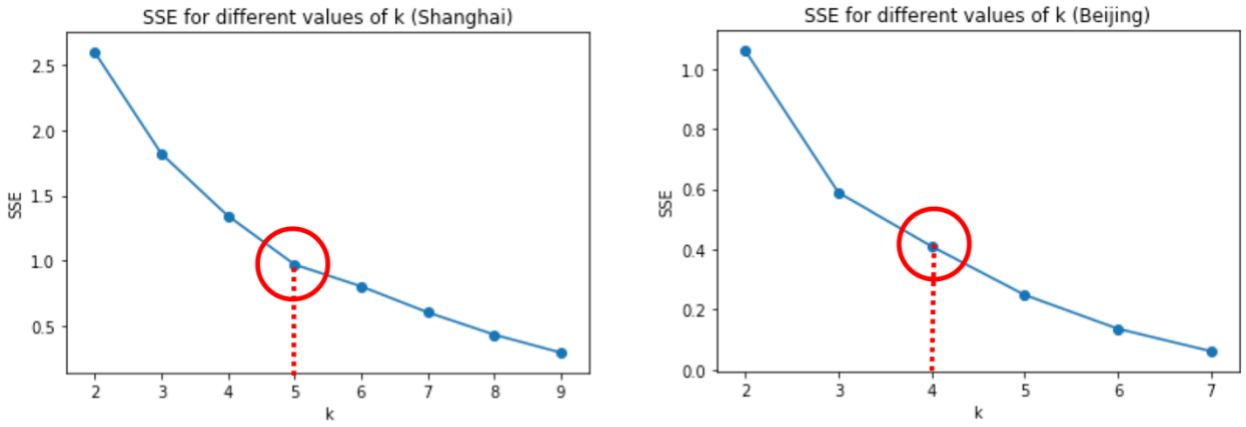


Figure 8. "Elbow" Analysis. Finally I chose K = 5 and K = 4 for Shanghai and Beijing, respectively.

3.5. Clustering result (by Visualization)

Figure 9 is the final results of Shanghai ang Beijing, separately, on map.

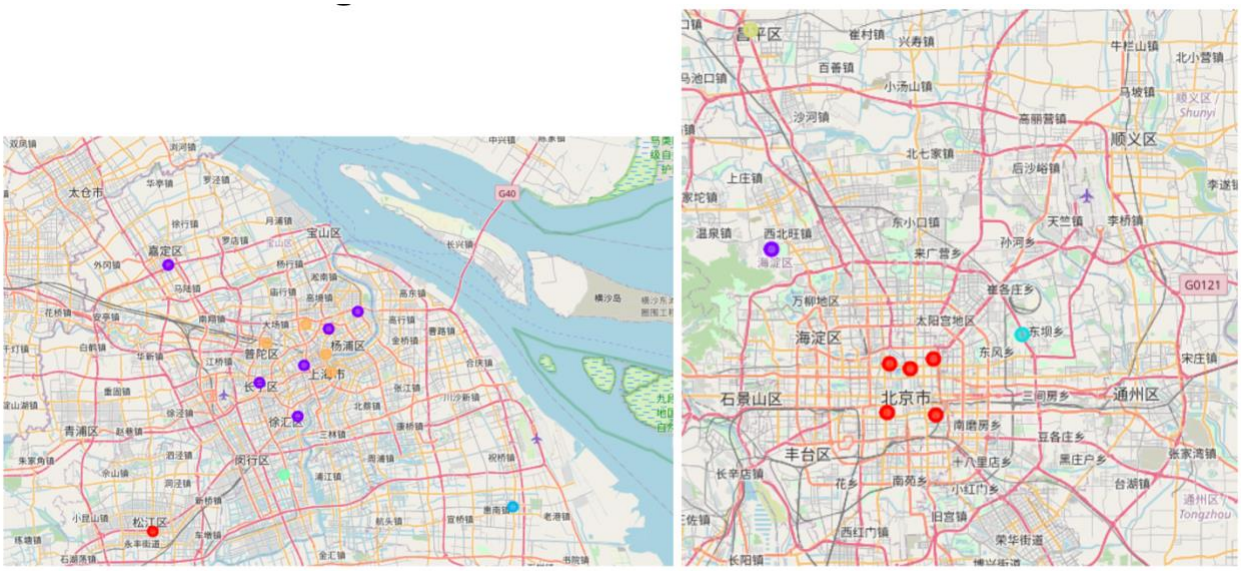


Figure 9. Clustering Visualization of each district on map for Shanghai and Beijing, separately (color-coded). There are 5 clusters (colors) for Shanghai and 4 clusters (colors) for Beijing.

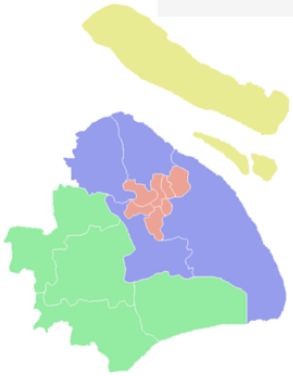
3.6. Examination of Clusters

Following is the summary results of 5 clusters for Shanghai and 4 clusters for Beijing, as shown in **Table 2**.

Table 2. Five clusters of **Shanghai**

	City	District
2	Shanghai	Xuhui
3	Shanghai	Changning
4	Shanghai	Jingan
7	Shanghai	Hongkou
8	Shanghai	Yangpu
10	Shanghai	Jiading

	City	District
11	Shanghai	Songjiang



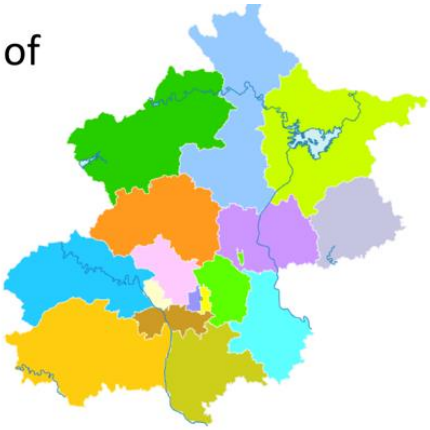
	City	District
0	Shanghai	Shanghai
1	Shanghai	Huangpu
5	Shanghai	Putuo
6	Shanghai	Zhabei

	City	District
12	Shanghai	Nanhui

	City	District
9	Shanghai	Minhang

Four clusters of **Beijing**

	City	District
0	Beijing	Beijing
1	Beijing	Dongcheng
2	Beijing	Xicheng
3	Beijing	Chongwen
4	Beijing	Xuanwu



	City	District
6	Beijing	Haidian

	City	District
7	Beijing	Changping

	City	District
5	Beijing	Chaoyang

4. Results (Shanghai vs. Beijing, together)

4.1. District Data Collection

I combined Shanghai and Beijing original data to a new dataset which contains 21 districts for the 2 cities (as shown in **Figure10**).

	Postal Code	City	District	Longitude	Latitude
0	110100	Beijing	Beijing	116.395645	39.929986
1	110101	Beijing	Dongcheng	116.421885	39.938574
2	110102	Beijing	Xicheng	116.373190	39.934280
3	110103	Beijing	Chongwen	116.424636	39.889292
4	110104	Beijing	Xuanwu	116.369352	39.891531
5	110105	Beijing	Chaoyang	116.521695	39.958953
6	110108	Beijing	Haidian	116.239678	40.033162
7	110114	Beijing	Changping	116.216456	40.221724
0	310100	Shanghai	Shanghai	121.487899	31.249162
1	310101	Shanghai	Huangpu	121.496072	31.227203
2	310104	Shanghai	Xuhui	121.446235	31.169152
3	310105	Shanghai	Changning	121.387616	31.213301
4	310106	Shanghai	Jingan	121.454756	31.235381
5	310107	Shanghai	Putuo	121.398443	31.263743
6	310108	Shanghai	Zhabei	121.457769	31.288044
7	310109	Shanghai	Hongkou	121.491919	31.282497
8	310110	Shanghai	Yangpu	121.535717	31.304510
9	310112	Shanghai	Minhang	121.425024	31.093538
10	310114	Shanghai	Jiading	121.251014	31.364338
11	310117	Shanghai	Songjiang	121.226791	31.021245
12	310119	Shanghai	Nanhui	121.769956	31.052602

Figure 10. Original Data shows Postal Code, name of the city, Longitude and Latitude of 21 districts for the 2 cities.

4.2. Foursquare Venues

Then we get venues information for each district. As you can see in the following **Figure 11**, we got 185 venues for the two cities.

```
(185, 7)
```

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beijing	39.929986	116.395645	TRB Hutong	39.926528	116.397204	French Restaurant
1	Beijing	39.929986	116.395645	The Georg by Georg Jensen	39.933440	116.392740	Scandinavian Restaurant
2	Beijing	39.929986	116.395645	City Walls Courtyard House Beijing	39.928131	116.392442	Hostel
3	Beijing	39.929986	116.395645	Peking Hostel (北平国际青年旅舍)	39.934350	116.396886	Hostel
4	Beijing	39.929986	116.395645	Nanluogu Alley (南锣鼓巷)	39.932498	116.396925	Pedestrian Plaza

Figure 11. 185 venues were obtained for the 2 cities together (from Foursquare).

In **Figure 12**, we counted the unique "venue category" and then group it by districts. There are 80 unique venue categories.

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Beijing	17	17	17	17	17	17
Changning	20	20	20	20	20	20
Changping	5	5	5	5	5	5
Chaoyang	2	2	2	2	2	2
Chongwen	5	5	5	5	5	5
Dongcheng	24	24	24	24	24	24
Haidian	1	1	1	1	1	1
Hongkou	6	6	6	6	6	6
Huangpu	14	14	14	14	14	14
Jiading	4	4	4	4	4	4
Jingan	26	26	26	26	26	26
Minhang	1	1	1	1	1	1
Nanhui	1	1	1	1	1	1
Putuo	4	4	4	4	4	4
Shanghai	23	23	23	23	23	23
Songjiang	2	2	2	2	2	2
Xicheng	9	9	9	9	9	9
Xuanwu	7	7	7	7	7	7
Xuhui	6	6	6	6	6	6
Yangpu	4	4	4	4	4	4
Zhabei	4	4	4	4	4	4

Let's find out how many unique categories can be curated from all the returned venues

```
1 print('There are {} uniques categories.'.format(len(two_city_venues['Venue Category'].u
```

```
There are 80 uniques categories.
```

Figure 12. 80 unique venues categories ("Venue Category") were obtained for the 2 cities, together.

Then we did onehot encoding of "Venue Category" in **Figure 13**.

```
The dimensionality of two_city_onehot dataset is (185, 81)
```

	District	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Auto Garage	Bakery	Bar	Beer Bar	Beijing Restaurant	...	Sporting Goods Shop	Steakhouse	Supermarket	Sushi Restaurant	Tea Room
0	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
3	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	Beijing	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

5 rows × 81 columns

Figure 13. Onehot encoding for "Venue Category".

And then group by "District" and calculate the mean for each district, as shown in **Figure 14**.

	District	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Auto Garage	Bakery	Bar	Beer Bar	Beijing Restaurant	...	Sporting Goods Shop	Steakhouse	Supermarket	S Restau
0	Beijing	0.000000	0.00	0.00	0.00	0.0	0.000000	0.058824	0.000000	0.058824	...	0.000000	0.000000	0.0	0.00(
1	Changning	0.000000	0.05	0.00	0.05	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
2	Changping	0.000000	0.00	0.20	0.00	0.0	0.200000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
3	Chaoyang	0.000000	0.00	0.00	0.00	0.5	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
4	Chongwen	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.2	0.00(
5	Dongcheng	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
6	Haidian	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
7	Hongkou	0.000000	0.00	0.00	0.00	0.0	0.166667	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
8	Huangpu	0.071429	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.071429	0.0	0.00(
9	Jiading	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
10	Jingan	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.038462	0.000000	0.0	0.00(
11	Minhang	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
12	Nanhui	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
13	Putuo	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
14	Shanghai	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
15	Songjiang	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
16	Xicheng	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.111111	0.111111	...	0.000000	0.000000	0.0	0.00(
17	Xuanwu	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
18	Xuhui	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.16(
19	Yangpu	0.000000	0.00	0.00	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(
20	Zhabei	0.000000	0.00	0.25	0.00	0.0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.00(

21 rows × 81 columns

Figure 14. Averaged "Venue Category" for each district.

Then we did the final arrangement to sort out the top 10 most common venues for each district, as shown in **Figure 15** below. This dataset is going to be used to ran the final clustering analysis.

```
(21, 11)
```

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Beijing	Hostel	Hotel	Hotpot Restaurant	Metro Station	French Restaurant	Chinese Restaurant	Restaurant	Café	Scandinavian Restaurant	Bus Station
1	Changning	Chinese Restaurant	Japanese Restaurant	Hotel	Gym Pool	Athletics & Sports	Tea Room	Movie Theater	Art Museum	Ramen Restaurant	Dongbei Restaurant
2	Changping	Shopping Mall	Asian Restaurant	Chinese Restaurant	Bakery	Korean Restaurant	Fast Food Restaurant	Dessert Shop	Dim Sum Restaurant	Dongbei Restaurant	Dumpling Restaurant
3	Chaoyang	Park	Auto Garage	Wedding Hall	Farmers Market	Department Store	Dessert Shop	Dim Sum Restaurant	Dongbei Restaurant	Dumpling Restaurant	Electronics Store
4	Chongwen	Metro Station	Supermarket	Chinese Restaurant	Fast Food Restaurant	Convenience Store	Gym	Grocery Store	Gym / Fitness Center	Furniture / Home Store	French Restaurant

Above is each district along with the top 10 most common venues for, both Shanghai and Beijing.

Figure 15. Top 10 most common venues for each district. This data is going to be used for final clustering analysis ---- districts from 2 cities combined.

4.3 Cluster Evaluation

In order to choose the right K number, I run "Elbow" analysis to calculate SSE for each K. As shown below in **Figure 16**, I chose K = 6 for the two cities together, to run the final K means clustering analysis.

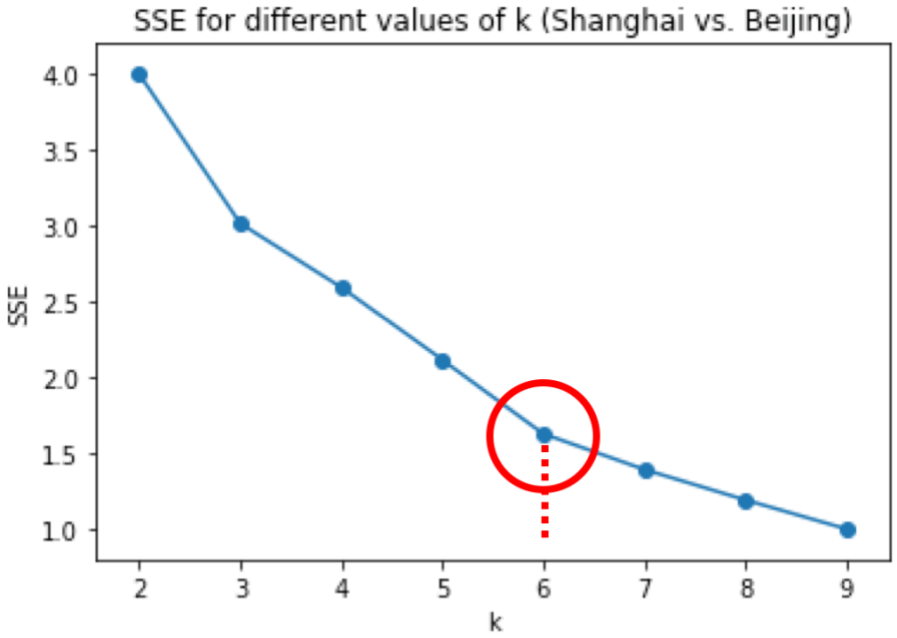


Figure 16. "Elbow" Analysis. Finally I chose K = 6 for the final clustering of the 2 cities together.

4.4. Clustering result (by Visualization)

Figure 17 is the final results of Shanghai ang Beijing, separately, on map.

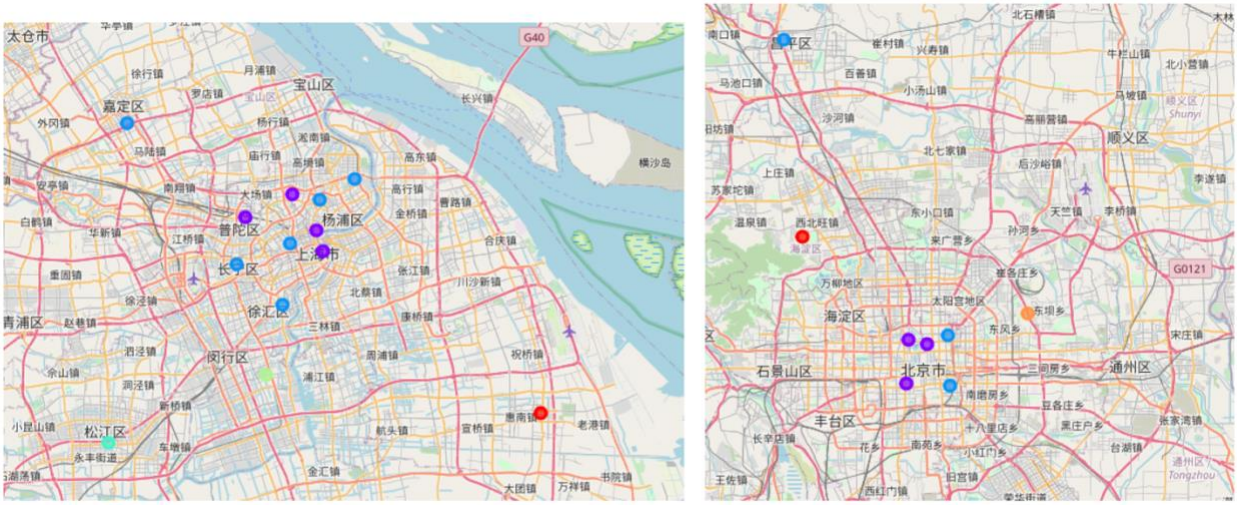


Figure 17. Clustering Visualization of each district for the 2 cities, together (color-coded). There are 6 clusters (colors).

4.5. Examination of Clusters

Following is the summary results of 5 clusters for Shanghai and 4 clusters for Beijing, as shown in **Table 3**. According to the Characteristics of each cluster, I also gave a proper name to each cluster, which is also shown in **Table 3**, and I will explain this further in **Discussion** section.

Table 3. Final clustering of Shanghai and Beijing together. There are 6 clusters and corresponding names.

Cluster 1 "**College Town**"

	City	District
6	Beijing	Haidian
12	Shanghai	Nanhui

Cluster 2 "**Traditional Cosmopolitan Area**"

	City	District
0	Beijing	Beijing
2	Beijing	Xicheng
4	Beijing	Xuanwu
0	Shanghai	Shanghai
1	Shanghai	Huangpu
5	Shanghai	Putuo
6	Shanghai	Zhabei

Cluster 3 "**Newly Developed Cosmopolitan Area**"

	City	District
1	Beijing	Dongcheng
3	Beijing	Chongwen
7	Beijing	Changping
2	Shanghai	Xuhui
3	Shanghai	Changning
4	Shanghai	Jingan
7	Shanghai	Hongkou
8	Shanghai	Yangpu
10	Shanghai	Jiading

Cluster 4 "**Minhang High-tech Development Zone**"

	City	District
9	Shanghai	Minhang

Cluster 5 "**Densely Populated Area**"

	City	District
5	Beijing	Chaoyang

Cluster 6 "**Old Shanghai**"

	City	District
11	Shanghai	Songjiang

5. Discussion

Now we are arriving at the final point that we have partitioned 21 districts from both Shanghai and Beijing into 6 clusters:

CLUSTER 1: 2 districts. Haidian (Beijing) and Nanhui (Shanghai) were partitioned as a cluster. Since the top 2 universities of China, Qinghua University and Peking University, are located in Haidian (Beijing), and there are plenty of universities located in Nanhui (Shanghai), I would like to name CLUSTER 1 as "**College Town**".

CLUSTER 2: 7 districts. Districts Beijing (downtown), Xicheng and Xuanwu from Beijing, and Shanghai (downtown), Huangpu, Putuo, Zhabei from Shanghai, were partitioned as the same cluster. When I check them on map, I realized that those 7 districts are almost all located at the very center of the two cities, I would like to name CLUSTER 2 as "**Traditional Cosmopolitan Area**".

CLUSTER 3: 9 districts. Districts Dongcheng, Chongwen and Changping from Bejing, Xuhui, Changning, Jingan, Hongkou, Yangpu and Jiading from Shanghai, were partitioned as the same cluster. When I check them on map, I realized that those 9 districts are almost all located at the very center of the two cities, except Changping (Beijing) and Jiangding (Shanghai) are located far away from the centers of the two cities. I think those two exception is reasonable since urban expansion was one of the main topics during the past 2 decades in Chinese metropolis development. Therefore, I would like to name CLUSTER 3 as "**Newly Developed Cosmopolitan Area**".

CLUSTER 4: 1 district. Minhang (from Shanghai) was partitioned as a cluster itself. I would like to name CLUSTER 4 as "**Minhang High-tech Development Zone**".

CLUSTER 5: 1 district. Chaoyang (from Beijing) was partitioned as a cluster itself. I would like to name CLUSTER 5 as "**Densely Populated Area**".

CLUSTER 6: 1 district. Songjiang (from Shanghai) was partitioned as a cluster itself. I would like to name CLUSTER 6 as "**Old Shanghai**" since in here live a lot of original residents of Shanghai.

6. Conclusion

As we can from the previous analysis, I was able to found some similarities of the two main big cities of China, in terms of venues around each districts, thus I was able to cluster them by using K-means unsupervised machine learning algorithms. This analysis may provide some information to government, city planners and entrepreneurs from both internal and external companies, but also original residents of the cities and incoming immigrants, when make a decision on moving to these 2 cities.