IBM Applied Data Science Capstone
Project Final Report

Finding Best Place to date with friends in BeiJing, China

Hong

Feb. 2020

Introduction--Problem definition

In a multicultural city and huge city like BeiJing, China. it might be daunting to find out which places are best to date your friends comfortably and relex after working time , considering the traffic issues of the venues，maybe near the subway station is better in a high traffic city，Not only that, one needs to look into different factors like shopping easy，restaurants, cafes good eating taste and so on.

Here can take me to a subway station that I often use as location information and I go to foursquare to explore the types of venues in the street, intuitively sort the data from the places I used to go, extract features, and help me to visualize the characteristics of the data

After cluster the place to help us to learn about them for different purpose and find out the best place for the coming dating.

Audiences

The People who live in Beijing and have subway transportation hobby want to find a suitable venue to have a gathering with his friends based on difference purpose/interesting.

Data Collection and Preparation

To solve the problem, we need to find the following data:

- List of Neighborhoods/subways station and their properties.
- Latitude and Longitude of the neighborhoods/subways.
- Venue data of the neighborhoods

To find the list of neighborhoods, I used the CSV file "beijingstation2.csv" located on the IBM Cloud Object Storage https://s3.eu-geo.objectstorage.service.networklayer.com', They were more than 20 neighborhoods in the CSV file. Since the coordinates were of boundaries, the centre coordinates of each neighborhood have to be calculated according to the average of all latitudes and longitudes. This would be needed in order to plot the neighborhood clusters.

Further more, we also need to draw a map for the place I often visited with the help of the tool (http://www.gpsspg.com/maps.htm)

To find the venue data, Foursquare API was used. It would show the most popular venues in each neighborhoods while using the central coordinates of the respective neighborhoods.

Methodology

Tools

The following tools were used:

- The Foursquare API was used to obtain venue data.
- The Folium package was used to plot Neighborhoods on the map .
- The KMeans module from sklearn package was used to cluster the data
- The json package was used to open and read the geojson file.
- IBM Cloud Object Storage - Python SDK from ibm_boto3 to get the coordinates of Neighborhoods
- The .read_csv from pandas library to scrape the data from a .csv file

Analytic Approach

First of all, the data we get is the street comprehensive data based on the latitude and longtitude from foursquare. This data contains many dimensions. The problem to be solved is exploring the feature in clusters. There are not limited standard for weather it is right or false. Because we do not need to predict at this stage. We just want to find out the result of clusters from the data in entertainment facilities. The

algorithm selected in this project is K-means. Clustering is the division of data into groups such that data points in the same group are more similar than data points in other groups. In short, clustering is the division of data points with similar characteristics into groups, that is, clusters. The goal of the K-means algorithm is to find a group in the data, the number of groups being represented by the variable K. Each data point is assigned to one of the K groups by an iterative operation based on the characteristics provided by the data.

As for the advantages and disadvantages

Advantages:

1. The algorithm is simple and easy to implement;
2. For processing large data sets, the algorithm is relatively scalable and efficient because its complexity is approximately O(nkt), where n is the number of all objects, k is the number of clusters, and t is the number of iterations. Usually k<<n. This algorithm usually converges locally.
3. The algorithm attempts to find the k partitions that minimize the value of the squared error function. When the clusters are dense, spherical or lumpy, and the difference between clusters and clusters is obvious, the clustering effect is better.

Disadvantage:

1. High requirements on data types, suitable for numerical data;
2. May converge to a local minimum and converge slowly on large-scale data.
3. K value is more difficult to select;
4. Sensitive to the cluster value of the initial value, which may result in different clustering results for different initial values;
5. Not suitable for finding clusters with non-convex shapes, or clusters with large differences in size.
6. Sensitive to "noise" and outlier data, a small amount of this type of data can have a significant impact on the average.

Conclusion for selecting K-means

Compared to the DBscan, DBscan is based on density calculation clustering, which will eliminate exceptions (noise points). The noise point will be deleted and clustered by the DBscan algorithm (not in the core point and not in the neighborhood of the core point). As for the station, most of them are the same, we should pay more attention finding out most common features in different station to help to make choice for the purpose according

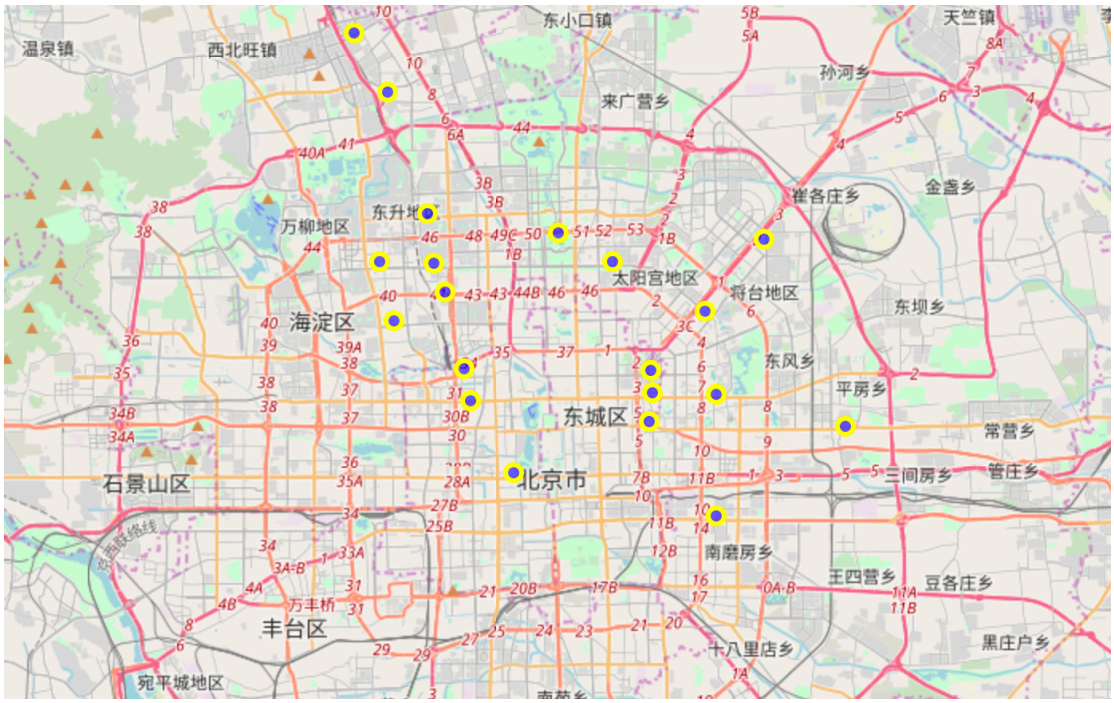
Result

I have selected about 20 subway stations where people visit very often and explore the possibilities around the subways.

[39]: (20, 3)

	Station	Latitude	Longitude
0	XIERQI	40.052243	116.306144
1	SHANGDI	40.032958	116.320519
2	WUDAOKOU	39.992833	116.337780
3	ZHICHUNLU	39.976424	116.340141
4	DAZHONGSI	39.966923	116.345126
5	XIZHIMEN	39.941856	116.353234
6	CHEGONGZHUANG	39.931445	116.356180
7	DONGZHIMEN	39.941255	116.433859
8	HUIXINXIJIENANKOU	39.976998	116.417644
9	SANYUANQIAO	39.960879	116.457055
10	TUANJIEHU	39.933450	116.461705
11	CHAOYANGMEN	39.924540	116.433433
12	HAIDIANHUANGZHUANG	39.976909	116.317043
13	AOTIZHONGXIN	39.986407	116.394155
14	WANGJINGNAN	39.984704	116.482311
15	DONGSISHITIAO	39.933852	116.434333
16	XIDAN	39.907769	116.374751

And visualize them on map below:



Using Foursquare API to obtain the facilities venue data around each subway stations/Neighborhoods I selected before, there are 1769 venues item results. And there are in 141 unique categories from all the returned venues

Venues

(1769, 7)

[6]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	XIERQI	40.052243	116.306144	李记潮汕砂锅粥	40.041352	116.335103	Cantonese Restaurant
1	XIERQI	40.052243	116.306144	Shantou Baheli Hai's Beef Restaurant (汕头八合里海记牛肉店)	40.040046	116.333920	Cantonese Restaurant
2	XIERQI	40.052243	116.306144	Starbucks (星巴克)	40.052671	116.296030	Coffee Shop
3	XIERQI	40.052243	116.306144	Starbucks (星巴克)	40.046375	116.292217	Coffee Shop
4	XIERQI	40.052243	116.306144	Haidilao Hot Pot (海底捞火锅)	40.027357	116.305631	Hotpot Restaurant
...	...	...	...	...	...	...	...
1764	WEIGONGCUN	39.957765	116.323136	Sculpting in Time (雕刻时光)	39.955710	116.305167	Café
1765	WEIGONGCUN	39.957765	116.323136	麻辣诱惑 Spice Spirit	39.977089	116.309981	Chinese Restaurant
1766	WEIGONGCUN	39.957765	116.323136	Carrefour (家乐福)	39.979649	116.307309	Supermarket
1767	WEIGONGCUN	39.957765	116.323136	雕刻时光 Sculpting In Time	39.978723	116.308185	Café
1768	WEIGONGCUN	39.957765	116.323136	Costa Coffee (咖世家)	39.979337	116.306323	Coffee Shop

1769 rows × 7 columns

[8]:
```
print('There are {} uniques categories.'.format(len(Venues['Venue Category'].unique())))
```
```
There are 141 uniques categories.
```

I changed the data of string for categories into number and group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

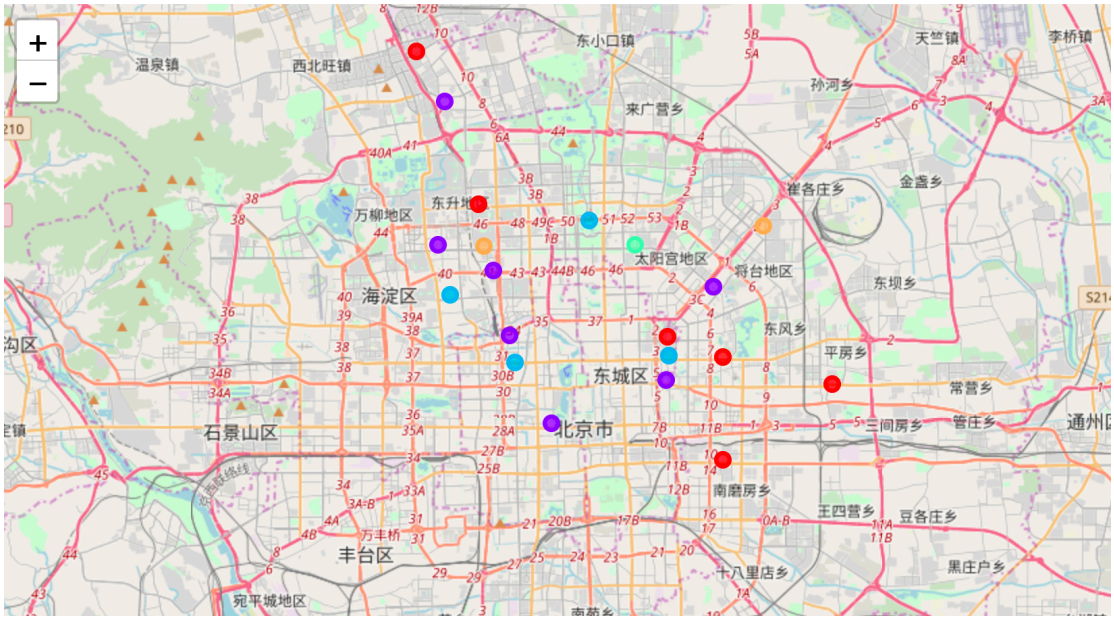
	Neighborhood	American Restaurant	Anhui Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	...	Th
0	AOTIZHONGXIN	0.012658	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.012658	0.000000	...	0.00
1	CHAOYANGMEN	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.0
2	CHEGONGZHUANG	0.000000	0.000000	0.00	0.010000	0.000000	0.00	0.00	0.000000	0.010000	...	0.0
3	DAZHONGSI	0.000000	0.000000	0.00	0.010000	0.000000	0.00	0.00	0.000000	0.010000	...	0.
4	DONGSISHITIAO	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.
5	DONGZHIMEN	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.010000	...	0.
6	HAIDIANHUANGZHUANG	0.000000	0.000000	0.00	0.000000	0.000000	0.01	0.00	0.000000	0.000000	...	0.00
7	HUIXINXIJIENANKOU	0.010753	0.000000	0.00	0.000000	0.010753	0.00	0.00	0.043011	0.010753	...	0.00
8	QINGNIANLU	0.040000	0.000000	0.04	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.00
9	SANYUANQIAO	0.010000	0.000000	0.00	0.000000	0.010000	0.00	0.00	0.020000	0.010000	...	0.00
10	SHANGDI	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.045455	0.000000	...	0.00
11	SHUANGJING	0.000000	0.000000	0.00	0.000000	0.020000	0.00	0.01	0.020000	0.000000	...	0.00
12	TUANJIEHU	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.0
13	WANGJINGNAN	0.050000	0.000000	0.00	0.000000	0.040000	0.00	0.01	0.010000	0.010000	...	0.00
14	WEIGONGCUN	0.010000	0.000000	0.00	0.010000	0.000000	0.00	0.00	0.030000	0.000000	...	0.0

I decided to choose TOP 10 features to Neighborhood cluster in 5 clusters using KMeans module. (Here we didn’ t find the best K using the Elbow Analysis)

13]:

	Neighborhood	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	AOTIZHONGXIN	Chinese Restaurant	Hotel	Coffee Shop	Fast Food Restaurant	Pizza Place	Hotpot Restaurant	Café	Park	Szechuan Restaurant	New American Restaurant
1	CHAOYANGMEN	Hotel	Shopping Mall	Chinese Restaurant	Café	Dumpling Restaurant	Italian Restaurant	Bar	Brewery	Peking Duck Restaurant	French Restaurant
2	CHEGONGZHUANG	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Hotel	Pizza Place	Café	Historic Site	Hotpot Restaurant	Department Store	Szechuan Restaurant
3	DAZHONGSI	Fast Food Restaurant	Pizza Place	Coffee Shop	Café	Chinese Restaurant	Sandwich Place	Bar	Clothing Store	Hotel	Hotpot Restaurant
4	DONGSISHITIAO	Hotel	Chinese Restaurant	Brewery	Japanese Restaurant	Café	Shopping Mall	Bar	Pizza Place	Szechuan Restaurant	Dumpling Restaurant
5	DONGZHIMEN	Hotel	Chinese Restaurant	Japanese Restaurant	Café	Brewery	Coffee Shop	Shopping Mall	Pizza Place	Szechuan Restaurant	Dumpling Restaurant
6	HAIDIANHUANGZHUANG	Chinese Restaurant	Fast Food Restaurant	Café	Sandwich Place	Coffee Shop	Pizza Place	Bakery	Bar	Korean Restaurant	Xinjiang Restaurant
7	HUIXINXIJIENANKOU	Fast Food Restaurant	Chinese Restaurant	Coffee Shop	Hotel	Pizza Place	Hotpot Restaurant	Asian Restaurant	Multiplex	Shopping Mall	Park
8	QINGNIANLU	Supermarket	Coffee Shop	Hotel	Clothing Store	American Restaurant	Noodle House	Park	Farm	Electronics Store	Department Store
9	SANYUANQIAO	Hotel	Japanese Restaurant	Bakery	Italian Restaurant	Park	Chinese Restaurant	Café	Shopping Mall	Cocktail Bar	Brewery

And the final result is like this there 5 clusters in the map below”



Discussion

Checked the details of 5 clusters.

There are 2 clusters (cluster 4&cluster 5)only contain 1 or 2 items, so we just discuss the rest 3 clusters (cluster1,cluster2 & cluster 3).

[17]:
```
##=============================================
##Cluster1
##=============================================
test0=Merged.loc[Merged['Cluster Labels'] == 0]
test0
```

[17]:

	Station	Latitude	Longitude	Cluster Labels	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
0	XIERQI	40.052243	116.306144	0	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Hotel	Cantonese Restaurant	Pizza Place	Shopping Mall	Sporting Goods Shop
2	WUDAOKOU	39.992833	116.337780	0	Fast Food Restaurant	Café	Sandwich Place	Chinese Restaurant	Hotel	Coffee Shop	Bar	Korean Restaurant
7	DONGZHIMEN	39.941255	116.433859	0	Hotel	Chinese Restaurant	Japanese Restaurant	Café	Brewery	Coffee Shop	Shopping Mall	Pizza Place
10	TUANJIEHU	39.933450	116.461705	0	Hotel	Italian Restaurant	Dumpling Restaurant	Shopping Mall	Brewery	Japanese Restaurant	Café	Bar
17	SHUANGJING	39.893557	116.461962	0	Hotel	Coffee Shop	Shopping Mall	Chinese Restaurant	Café	Italian Restaurant	Dumpling Restaurant	Fast Food Restaurant
18	QINGNIANLU	39.923018	116.517454	0	Supermarket	Coffee Shop	Hotel	Clothing Store	American Restaurant	Noodle House	Park	Farm

In [18]:
```
##=============================================
##Cluster2
##=============================================
test1=Merged.loc[Merged['Cluster Labels'] == 1]
test1
```

Out[18]:

	Station	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Com Venu
1	SHANGDI	40.032958	116.320519	1	Fast Food Restaurant	Hotel	Coffee Shop	Chinese Restaurant	Cantonese Restaurant	Asiar Resta
4	DAZHONGSI	39.966923	116.345126	1	Fast Food Restaurant	Pizza Place	Coffee Shop	Café	Chinese Restaurant	Sanc Place
5	XIZHIMEN	39.941856	116.353234	1	Fast Food Restaurant	Chinese Restaurant	Coffee Shop	Hotel	Pizza Place	Café
9	SANYUANQIAO	39.960879	116.457055	1	Hotel	Japanese Restaurant	Bakery	Italian Restaurant	Park	Chin Resta
11	CHAOYANGMEN	39.924540	116.433433	1	Hotel	Shopping Mall	Chinese Restaurant	Café	Dumpling Restaurant	Italia Resta
12	HAIDIANHUANGZHUANG	39.976909	116.317043	1	Chinese Restaurant	Fast Food Restaurant	Café	Sandwich Place	Coffee Shop	Pizza Place
16	XIDAN	39.907769	116.374751	1	Historic Site	Hotel	Coffee Shop	Shopping Mall	Park	Café

In [20]:
```
##=============================================
##Cluster4
##=============================================
test3=Merged.loc[Merged['Cluster Labels'] == 3]
test3
```

Out[20]:

	Station	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7 C V
8	HUIXINXIJIENANKOU	39.976998	116.417644	3	Fast Food Restaurant	Chinese Restaurant	Coffee Shop	Hotel	Pizza Place	Hotpot Restaurant	A R

In [21]:
```
##=============================================
##Cluster5
##=============================================
test4=Merged.loc[Merged['Cluster Labels'] == 4]
test4
```

Out[21]:

	Station	Latitude	Longitude	Cluster Labels	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 Co Ver
3	ZHICHUNLU	39.976424	116.340141	4	Fast Food Restaurant	Sandwich Place	Chinese Restaurant	Coffee Shop	Café	Clothing Store	Piz Pla
14	WANGJINGNAN	39.984704	116.482311	4	Café	Korean Restaurant	Coffee Shop	Chinese Restaurant	Hotel	American Restaurant	Jap Res

Observe the first cluster and I found most of them are Coffee Shop, Café, Chinese Restaurant, Fast Food Restaurant, Dumpling Restaurant, Japanese Restaurant, Asian Restaurant, So I call it as an Asia style cluster as below

```
[29]: ##=============================================
      ##Cluster1 Count Catalogies
      ##=============================================
      test0=test0.drop(['Station','Latitude','Longitude','Cluster Labels'],axis=1)
      df2=pd.concat(test0.iloc[:,i] for i in range(test0.shape[1]))
      #适当修改索引
      df2.head(20)
      df2.index=np.arange(len(df2))
      print(df2.shape)
      df2.value_counts()
```

```
(60,)
```

```
[29]: Coffee Shop               6
      Hotel                     6
      Café                      4
      Chinese Restaurant        4
      Shopping Mall             4
      Pizza Place               3
      Dumpling Restaurant       3
      Fast Food Restaurant      3
      Bar                       2
      Japanese Restaurant       2
      Italian Restaurant        2
      Brewery                   2
      New American Restaurant   1
      Farm                      1
      Asian Restaurant          1
      American Restaurant       1
      Cocktail Bar              1
```

Observe the second cluster below and get the result: most of them are Hotel, Chinese Restaurant, Japanese Restaurant, Italian Restaurant, French Restaurant, Pizza Place, Xinjiang Restaurant, Shopping Mall, Korean Restaurant, Fast Food Restaurant, Cocktail Bar, Fast Food Restaurant, Peking Duck Restaurant, Brewery, Asian Restaurant and etc. there are many diversity tasty, So We call it as diversity cluster.

```
##Cluster2 Count Catalogies
##================================================
test1=test1.drop(['Station','Latitude','Longitude','Cluster Labels'],axis=1)
df2=pd.concat(test1.iloc[:,i] for i in range(test1.shape[1]))
#适当修改索引
df2.index=np.arange(len(df2))
print(df2.shape)
df2.value_counts()
```

```
(70,)
```

[24]:

```
Chinese Restaurant        7
Café                      7
Hotel                     6
Coffee Shop               5
Bar                       4
Fast Food Restaurant      4
Shopping Mall             3
Sandwich Place            3
Pizza Place               3
Park                      2
Brewery                   2
French Restaurant         2
Italian Restaurant        2
Hotpot Restaurant         2
Bakery                    2
Japanese Restaurant       2
Xinjiang Restaurant       2
Historic Site             1
Bus Stop                  1
Peking Duck Restaurant    1
```

Observe the third cluster below most of them are Chinese Restaurant, Hotpot Restaurant, Coffee Shop, Fast Food Restaurant, Szechuan Restaurant, Café, So I call it as a local style cluster. So I call it as a local style cluster as below

[25]:

```
##================================================
##Cluster3 Count Catalogies
##================================================
test2=test2.drop(['Station','Latitude','Longitude','Cluster Labels'],axis=1)
df2=pd.concat(test2.iloc[:,i] for i in range(test2.shape[1]))
#适当修改索引
df2.index=np.arange(len(df2))
print(df2.shape)
df2.value_counts()
```

```
(40,)
```

[25]:

```
Chinese Restaurant        4
Pizza Place               4
Hotel                     4
Café                      4
Hotpot Restaurant         3
Coffee Shop               3
Fast Food Restaurant      3
Szechuan Restaurant       3
Japanese Restaurant       1
Park                      1
Bar                       1
Historic Site             1
Brewery                   1
Asian Restaurant          1
```

Conclusions

Based on for subway transportation hobby people, we request the facilities around the subway to cluster it. And we cluster them into 5 clusters. Besides the other 2 clusters contain only one or two items, so ignore them

As for the first cluster (the Asia style cluster.), people can choose when he wants to have the Asia friends to date and which is more comfortable to get them.

As for the second cluster (Diversity cluster.), people can choose from them when he wants to taste something new and walk around for more. It should be more expensive and funny.

As for the third cluster (local style cluster), people can choose them when he wants to have a family feely