New shop analysis

Introduction

XYZ Co. is a shop properties owner in Hong Kong. Recently they have a tenant moved out, so they have a shop for lease (at San Ma Tau Street, Ma Tau Kok, To Kwa Wan, Hong Kong, China). According to their relationship with their existing tenants and their internal interests, they have 2 potential targets with the top priority:

1. MCD's Corporation ("MCD"): an American fast food restaurant chain serves over 69 million customers daily in over 100 countries across approximately 40,000 outlets. And there are around 200 outlets in Hong Kong.
2. SBX coffee company ("SBX"): a coffeehouse chain found in Seattle, operating around 28,000 coffee shops worldwide, including around 150 coffee shops in Hong Kong.

And XYZ Co. knows that both of the targets are interested to expand their business. So most likely both of them have a chance to rent the XYZ shop. But XYZ Co. is not sure if his place fits MCD shop or SBX shop, and the locals like which one more. So they came to me for help to solve the problem.

After we have some discussions, and a preliminary plan came out:
As both two chains run their business in Hong Kong successfully. We can analyze the nearby venues of the two chains in Hong Kong, and check if the nearby venues of XYZ Co.'s place are similar to MCD shops or SBX shops.

We can use data analytics to analyze the nearby venues of the two chains in Hong Kong. And we can try to use the machine learning Classification method to classify XYZ shop into MCD shops or SBX shops according to the all nearby venues data of the existing data.

Data

There are around 200 MCD shops and 150 SBX shops in Hong Kong. We need to know where they are, and then, the nearby venues of these shops. In order to get the location of the MCD and SBX shops, I thought they should have lists of all their shops in Hong Kong and we can generate locations from the lists. However, both the websites do not have lists that we can easily import. At the meantime, I cannot found any lists on the internet.

Since there is no list of these shops, next, I used an alternative solution to get the shops locations: Foursquare Search API. After I tested the API, found 2 limitations:

1. If the response has less than 50 results that are exactly matched the name, there may be some results that only similar to the search term. it means that the API may return some results that are not the shops of MCD or SBX.
2. Although maximum "radius" can be set large enough (100,000 meters) to cover Hong Kong, the maximum results to return of my Foursquare account is only 50 results per query. As mentioned in the Introduction, there are around 200 MCD shops and 150 SBX Shops in Hong Kong. We need to use many different locations as the request parameters to get as many shops locations as possible.

The first issue can be solved by adding a name check process. For the second issue, as MTR is one of the most mature public transport network in the world, and MTR stations almost covered Hong Kong urban areas. We can use the MTR station location data to search for MCD & SBX shops near

MTR stations. Although there is no guarantee that to get the full list of MCD and SBX shops in Hong Kong, we still can get most of the results in the urban area.

Once we get the location data of MCD and SBX shops, the following step is to get the venues data near every MCD and SBX shops. The data can be retrieved by Foursquare venue explore API.

Methodology

Data preparation

Location list of MCD and SXB shops

First, we need to get the location data of the MTR stations. we can use GeoPy to convert the addresses into geographical coordinates. And luckily, the stations' names are the addresses, and GeoPy can correctly convert them into geographical coordinates. MTR station names can be found on the website (https://www.exploremetro.com/blog/hong-kong-mtr-station-names-in-cantonese-jyutping/). There is a list on the website and the list has English names, Chinese names and the Cantonese Pinyins of the stations. Once we get the stations names, we can get all the stations names, change them to addresses (just simply add ", Hong Kong" at the end), and use GeoPy library to convert the addresses into geographical coordinates.

Then, we can use each MTR station location, to generate MCD and SBX shops around 1000 meters of each station using Foursquare API. And remove duplicate shops with the same Foursquare ID.

Figure 1 below is the map marks MCD shops (red marks) and SBX shops (blue marks) in Kowloon and Hong Kong Island. I found that there are some MCD shops and SBX shops nearby. The next step is to combine the 2 nearby shops if an MCD shop and an SBX shop are within 50 meters, and the combined location point is in the middle of the 2 shops. And Figure 2 is the map marked 3 types of points: MCD shops (red marks), SBX shops (blue marks), combined points (green marks).

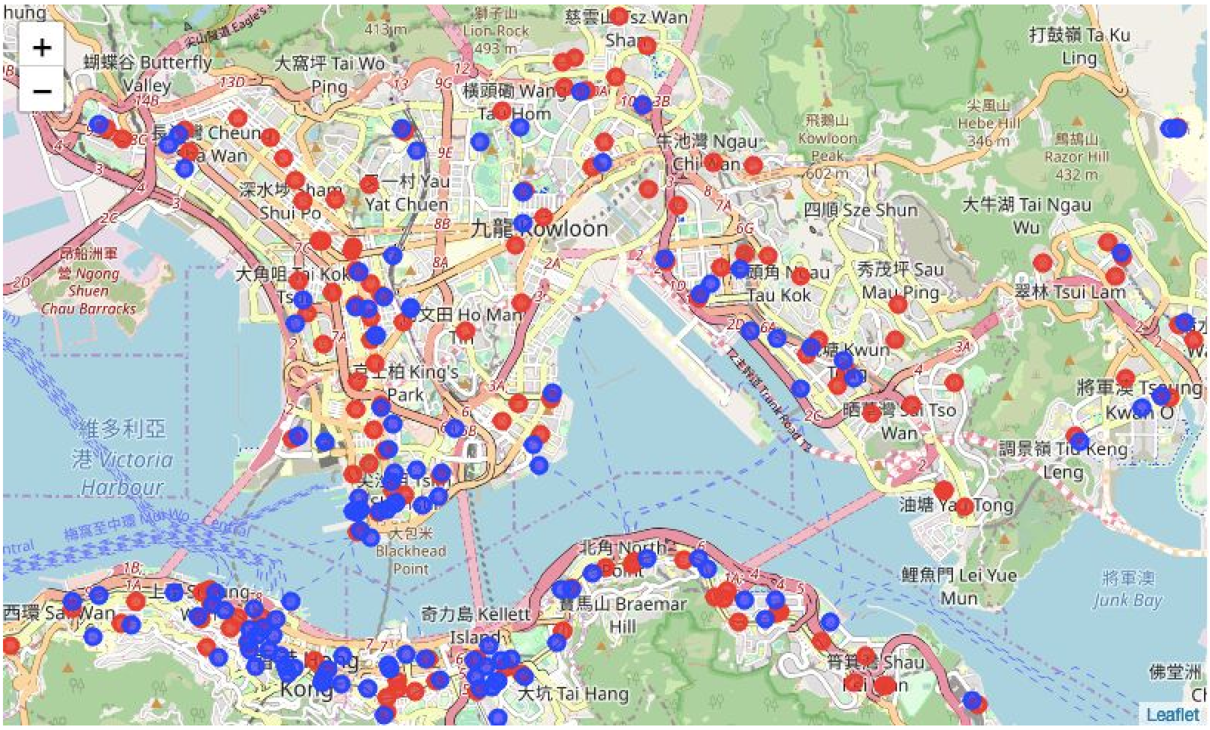


Figure 1

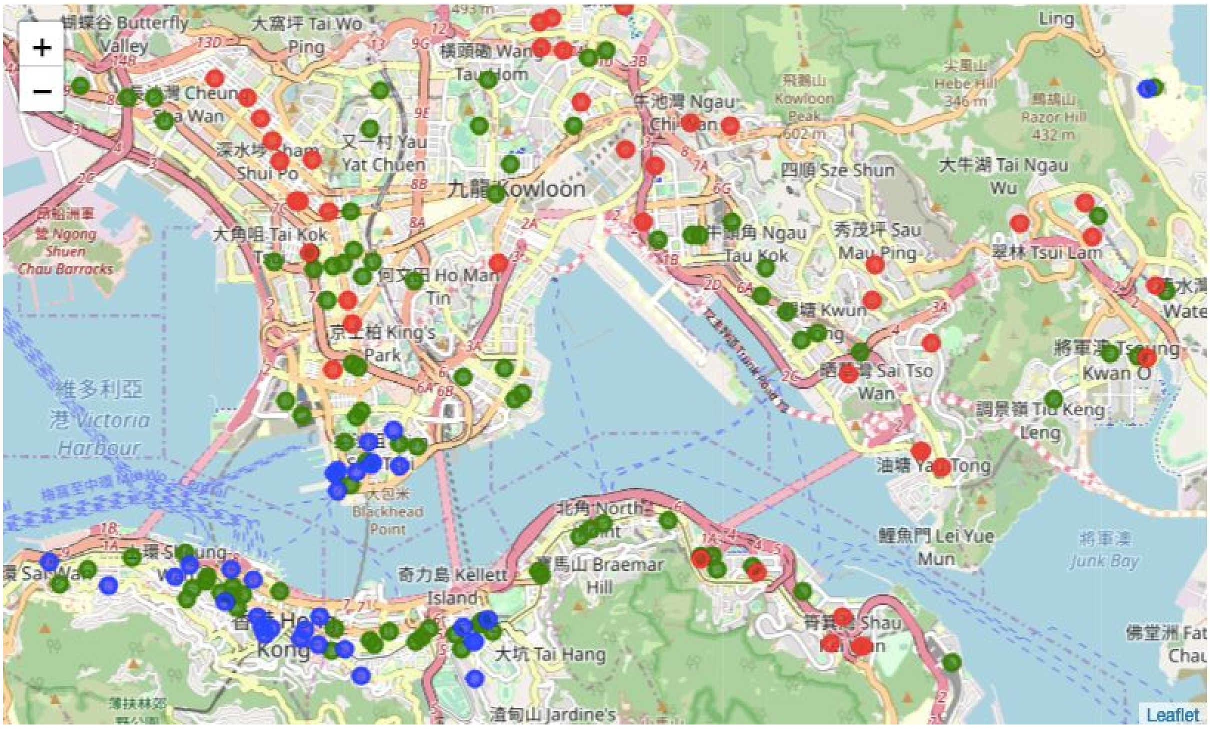


Figure 2

Finally, the data of these 3 types of locations are combined into a list. And the list should at least cover below items: the type of the shop (Type 0 - MCD shops, Type 1 - SBX shops, and Type 2 - both shops), the latitude & the longitude.

Nearby venues list

Once we get the location data of MCD & SBX shops, the following step is to get the venues data near each MCD or SBX shop by sending the request to Foursquare explore venues API, and then convert JSON file of the response into a list. The list should at least cover below columns: the category of the venue near the shop, the type of that shop and the Foursquare ID of the shop. Also, as we need to use the Classification Method to predict XYZ shop, the shop nearby venues data will put it into the Nearby venues list.

Modeling and evaluation

For the modeling step, we will use classification methods to identify to which label (the category) the new observation (XYZ Co.'s shop) belongs, and the training dataset containing observations (existing shops) whose labels are assigned.

We will use nearby venues categories as independent variables "X" and the shop type as dependent variables "y". In order to use the data for machine learning, one-hot encoding is used for representing the venue categories. Next, group rows by each shop and by taking the mean of the frequency of occurrence of each category. As a result, a list that each row represents a shop is generated. Each row has data of frequency of occurrence of each category, and the type of this shop. After the process, the frequency of occurrence of each category is the independent variables "X".

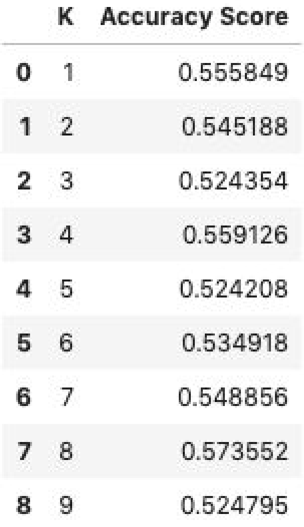
There are some classification algorithms. In the following part, 3 algorithms will be used to build models with different parameters, then compare the accuracies of different parameters under the same algorithms, and finally, compare the algorithms. The 3 algorithms are:

- K-nearest neighbor;
- Logistic regression; and
- Support vector machines.

To evaluate the models built by different algorithms and parameters, we will use the K-fold cross-validation method to estimate accuracy.

K-nearest neighbor

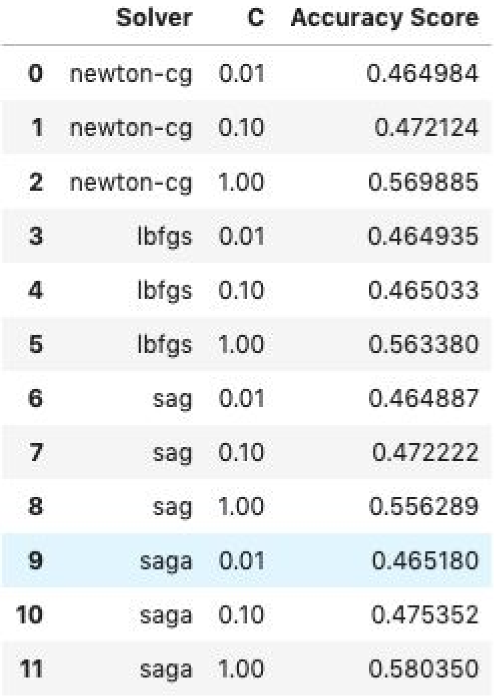
We will assign from 1 to 10 to K, which is the number of nearest neighbors, and calculate the accuracy of predicition to examine the model. The below table is the accuracy when K is 1 to 10. And when K = 8, the model has the best accuracy.



	K	Accuracy Score
0	1	0.555849
1	2	0.545188
2	3	0.524354
3	4	0.559126
4	5	0.524208
5	6	0.534918
6	7	0.548856
7	8	0.573552
8	9	0.524795

Logistic regression

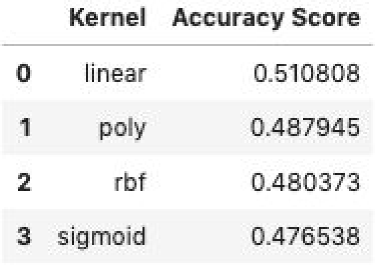
When we examine the Logistic regression, we will assign C to 0.01, 0.1 and 1, and use the solvers: 1) newton-cg, 2) lbfgs, 3) sag and 4) saga. And when solver = saga and C = 1, the model has the best accuracy. But when we future examine the predicted result from the test set, we find that the model only predicts the values either type 0 or type 2.



	Solver	C	Accuracy Score
0	newton-cg	0.01	0.464984
1	newton-cg	0.10	0.472124
2	newton-cg	1.00	0.569885
3	lbfgs	0.01	0.464935
4	lbfgs	0.10	0.465033
5	lbfgs	1.00	0.563380
6	sag	0.01	0.464887
7	sag	0.10	0.472222
8	sag	1.00	0.556289
9	saga	0.01	0.465180
10	saga	0.10	0.475352
11	saga	1.00	0.580350

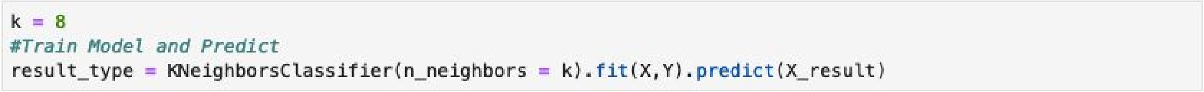
Support vector machines

We will use below kernel to examine the accuracies: Linear, Poly, RBF, and Sigmoid. And the model has the best accuracy when we use the kernel Linear.

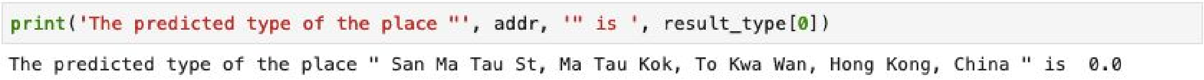


	Kernel	Accuracy Score
0	linear	0.510808
1	poly	0.487945
2	rbf	0.480373
3	sigmoid	0.476538

After we examine the above 3 algorithms with different parameters, we will choose to use the “K-nearest neighbor” algorithm with K = 8 to build the model. After we got the nearby venues data of XYZ shop, we predict the type of the shop is type 0. It means the shop suits MCD more.



```
k = 8
#Train Model and Predict
result_type = KNeighborsClassifier(n_neighbors = k).fit(X,Y).predict(X_result)
```



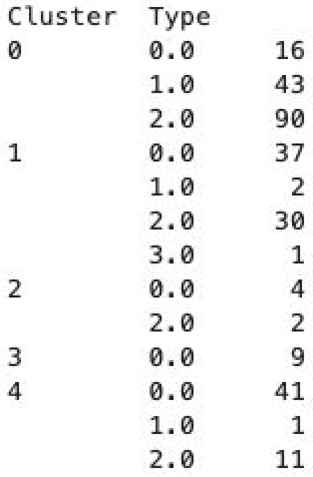
```
print('The predicted type of the place "', addr, '" is ', result_type[0])
```

```
The predicted type of the place " San Ma Tau St, Ma Tau Kok, To Kwa Wan, Hong Kong, China " is  0.0
```

Results

In the above session, we use the K-nearest neighbor algorithm to predict our shop's type, according to the similarity of nearby venues between the shop and the MCD or SBX shops in Hong Kong. So that we can find out the locals like which potential tenant, MCD or SBX, more. At the end of the previous session, we got the result of type 0. And it means the nearby venues is more similar to the nearby venues of MCD shops.

Given that the accuracy of the KNN model we built is only around 0.57, in the following, we will cluster the shops and examine the clusters. We cluster the shops according to the nearby venues using the K-means algorithm with K = 5. After we cluster the shop into 5 categories, we can examine the categories. We combine the shop types and the categories. In this time, there are 4 shop types: 0 = MCD shop; 1 = SBX shop; 2 = the place has both 2 shop; 3 = XYZ Co.'s shop that for lease. Below table shows the number of shops in each cluster.



```
Cluster  Type
0        0.0     16
         1.0     43
         2.0     90
1        0.0     37
         1.0      2
         2.0     30
         3.0      1
2        0.0      4
         2.0      2
3        0.0      9
4        0.0     41
         1.0      1
         2.0     11
```

In Figure 3, it shows the points that cluster into 4 categories: Cluster 0 - red, Cluster 1 - blue, Cluster 2 - green, Cluster 3 - yellow, Cluster 4 - black. And the biggest circle marks the XYZ shop. And in Figure 4, it shows all the shops (including XYZ shop) in Cluster 1. The reds are MCD shops, the blues are SBX shops and the greens are the places of both 2 shops.

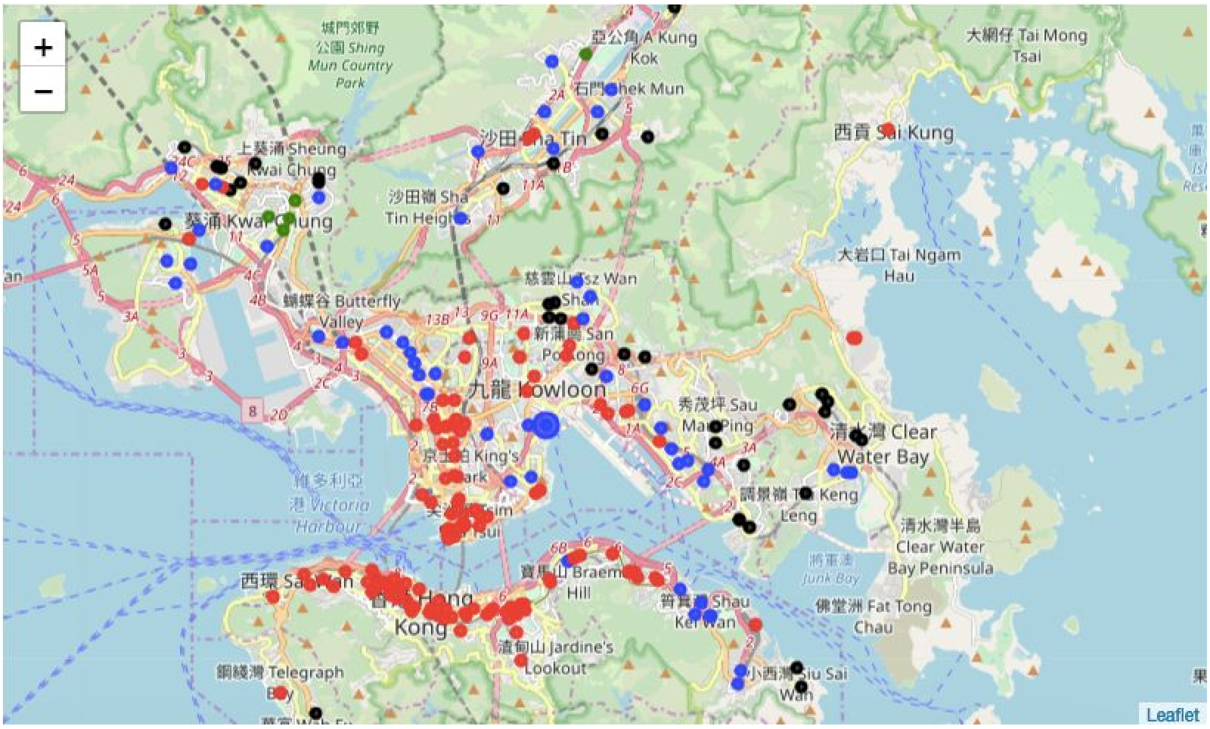


Figure 3

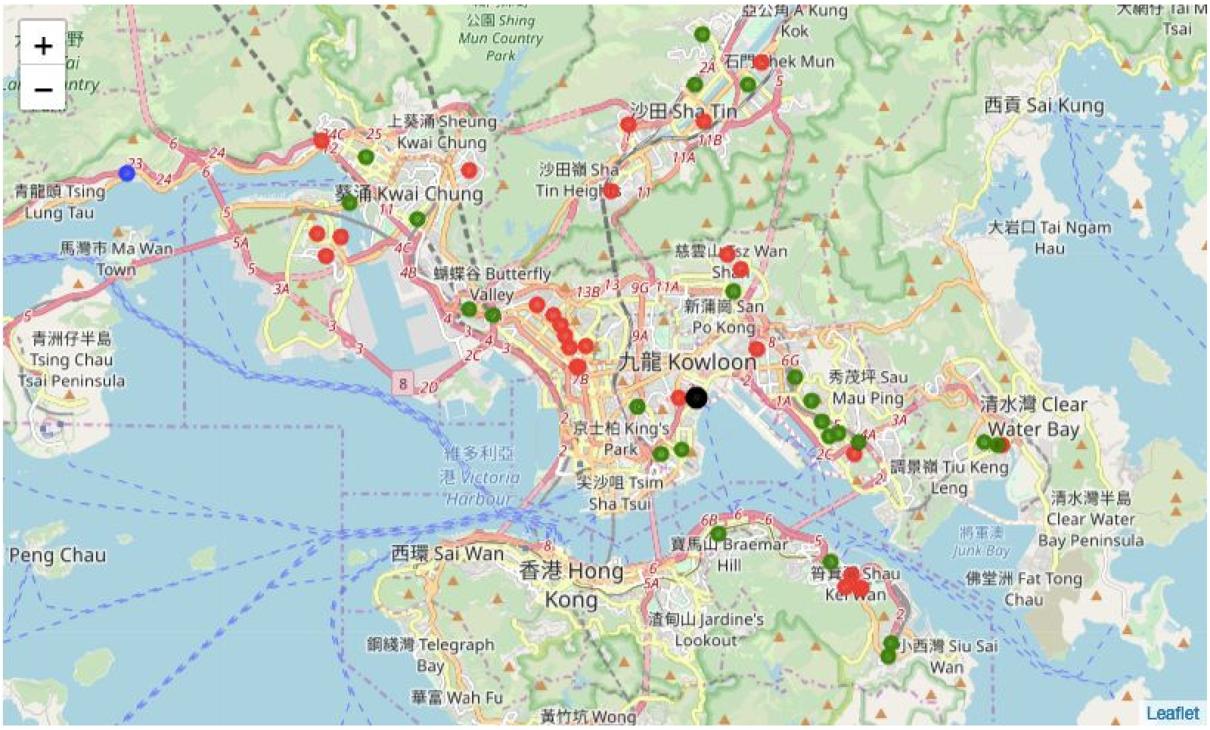


Figure 4

We can see that XYZ shop is in Cluster 1. There are 67 shops (Type 0 = 37 & Type 2 = 30) are MCD shops, and only 32 shops (Type 1 = 2 & Type 2 = 30) are SBX shops. We have a conclusion that XYZ Co.'s place fits MCD shop, as there more MCD shops than SBX shops in Cluster 1's places.

Discussion and Conclusion

After we analyze the data by Classification and Clustering, the data shows that MCD is a better talent. MCD is more likely to open a shop where nearby venues are similar to XYZ shop. The classification result shows the XYZ shop's nearby environment is more similar to MCD shops. And the clustering result shows the XYZ shop is clustered to a category that there are far more MCD shops than SBX shops. However, the accuracy of the Classification is not high. It seems to me that there are other considerations that the chains decide to open the shops.

In the future, we can gather more data, for example, the land value, the population distribution, etc, to find out more key elements that the chains will considerate. In the meanwhile, there are too many venues categories (around 300 categories). The results may be more accurate if we simplify the categories.