Open a Chinese Restaurant in Toronto

Xizhen Zhang

January 20, 2020

1. Introduction

1.1 Background

Toronto is the most populous city in Canada, and it's also a multicultural city. There are many ethnic neighborhoods such as Chinatown, Koreatown, little India, little Italy and so on, and it will be a significant factor to determine which neighborhood would be the best option to open a new Chinese restaurant. In addition, Ontario is one of the most immigrant-friendly province in North America, hence more and more Chinese are moving to Canada for their rest of life. Furthermore, Chinese food is a very popular food, and people all over the world like to eat it. Therefore, open a Chinese restaurant in Toronto probably will be a good choice to start one's business.

1.2 Problem

In this project, we will analyze the neighborhoods in Toronto to identify the most profitable and least competitive area. Then we can decide the location of the restaurant, where it can yield more profit to the owner.

1.3 Interest

This project focuses on three kinds of people:

- The person who interested in cooking and hope to open a Chinese restaurant as a side business. This analysis will tell them what are the pros and cons about this business.
- The business person who wants to get involved into Chinese restaurant field in Toronto. This analysis will help them to get enough information about the catering industry in Toronto.
- The person who does not have ability of cooking and needs to eat outside. This analysis will help them to know the distributions of Chinese restaurant.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

1) Using “List of Postal code of Canada: M” table from Wikipedia, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, where can get the postal code, borough and neighborhoods in Toronto.
2) Using http://cocl.us/Geospatial_data csv file to get geographical coordinates for each postal code in Toronto.
3) Using statistic data from Wikipedia, https://en.m.wikipedia.org/wiki/Demographics_of_Toronto, where can get the distribution of population by their ethnicity in Toronto.
4) Using Foursquare’s explore API to get the present information about venues in Toronto.

2.2 Data Cleaning

1) In “List of Postal code of Canada: M” table, there are many not assigned values in column “Borough” and “Neighborhood”. First, I decided to scrape the rows that with “Neighborhood” not assigned, and if a cell has value in “Borough” but not assigned in “Neighborhood”, the “Neighborhood” will be the same as the “Borough”. As we can see, there are some neighborhoods exist in on postal code area, then I combined these two rows into one row and separated in “Neighborhood” by “,”.
2) Adding the geographical coordinates to existing table. First extracting the data from csv file, and then combining it with existing table by merging with same postal code.

Table 1: Combining two tables in step 1 and 2

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848

3) Extracting the data from Wikipedia which give us the distribution of population in different neighborhoods in Toronto, and rename some columns to make us to understand easier.

Table 2: Example for distribution of population in Toronto and East York

	Riding	Population	Ethnic Origin #1	Origin #1 %	Ethnic Origin #2	Origin #2 %	Ethnic Origin #3	Origin #3 %	Ethnic Origin #4	Origin #4 %	Ethnic Origin #5	Origin #5 %	Ethnic Origin #6	Origin #6 %	Ethnic Origin #7	Origi #7 %
0	Spadina-Fort York	114315	English	16.4	Chinese	16.0	Irish	14.6	Canadian	14.0	Scottish	13.2	French	7.70	German	7.6
1	Beaches-East York	108435	English	24.2	Irish	19.9	Canadian	19.7	Scottish	18.9	French	8.7	German	8.40	NaN	NaN
2	Davenport	107395	Portuguese	22.7	English	13.6	Canadian	12.8	Irish	11.5	Italian	11.1	Scottish	11.00	NaN	NaN
3	Parkdale-High Park	106445	English	22.3	Irish	20.0	Scottish	18.7	Canadian	16.1	German	9.8	French	8.88	Polish	8.5
4	Toronto-Danforth	105395	English	22.9	Irish	19.5	Scottish	18.7	Canadian	18.4	Chinese	13.8	French	8.86	German	8.8
5	Toronto-St. Paul's	104940	English	18.5	Canadian	16.1	Irish	15.2	Scottish	14.8	Polish	10.3	German	7.90	Russian	7.7
6	University-Rosedale	100520	English	20.6	Irish	16.6	Scottish	16.3	Canadian	15.2	Chinese	14.7	German	8.70	French	7.7
7	Toronto Centre	99590	English	15.7	Canadian	13.7	Irish	13.4	Scottish	12.6	Chinese	12.5	French	7.20	NaN	NaN

4) Using Foursquare's explore API to retrieve the information in Toronto, and it will return a JSON file. Then turn it into dataframe. After that, set the radius of area that we need to find, which the radius using in this project is 1500 meters.

Table 3: The venues in each neighborhood in Toronto

	name	categories	lat	lng
0	Downtown Toronto	Neighborhood	43.653232	-79.385296
1	Nathan Phillips Square	Plaza	43.652270	-79.383516
2	Japango	Sushi Restaurant	43.655268	-79.385165
3	Four Seasons Centre for the Performing Arts	Concert Hall	43.650592	-79.385806
4	Art Gallery of Ontario	Art Gallery	43.654003	-79.392922

Table 4: Neighborhood and venue category for each venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
2	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
3	Rouge, Malvern	43.806686	-79.194353	Caribbean Wave	43.798558	-79.195777	Caribbean Restaurant
4	Rouge, Malvern	43.806686	-79.194353	Tim Hortons	43.802000	-79.198169	Coffee Shop

3. Exploratory Data Analysis

3.1 Show the distribution of neighborhoods in Toronto on map

Using the Folium library in Python to create maps for different neighborhoods. Folium is developed for the sole purpose of visualizing geospatial data. Although other libraries are available to visualize geospatial data, they might have a cap on how many API calls you can make within a defined time frame. Folium is completely free.

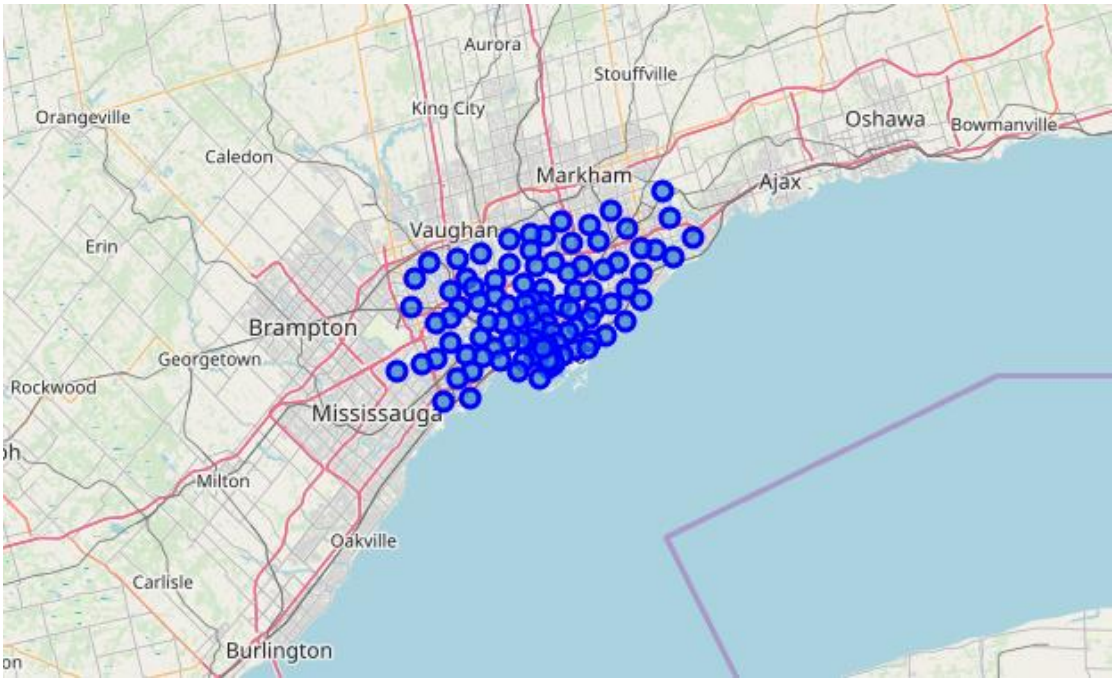

Figure 1: The distribution of neighborhoods in Toronto

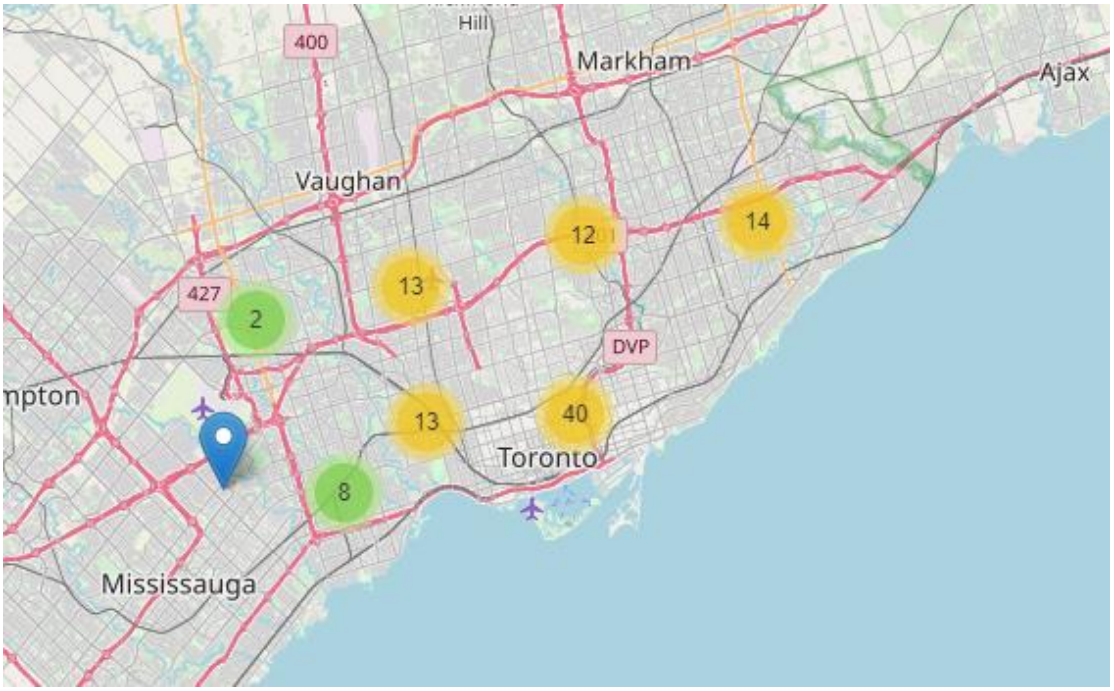

Figure 2: The better view of above visualization

3.2 The distribution of Chinese restaurant in Toronto

Using groupby function to group by Neighborhood, and sum it to count how many Chinese restaurants in each neighborhood. Then extracting the data from the table that we just grouped, and leave 2 columns which are Neighborhood and Chinese Restaurant respectively. Afterward, merge this dataframe with Table 1 we did above with latitude and longitude information by same Neighborhood. Finally, use groupby function one more time to group the table by Borough and extract the column name Borough and Chinese Restaurant, then we get the table below:

Table 5: Brough and Count of Chinese Restaurant

	Borough	Chinese Restaurant
0	Central Toronto	2
1	Downtown Toronto	9
2	East Toronto	2
3	East York	1
4	Etobicoke	2
5	Mississauga	1
6	North York	10
7	Queen's Park	1
8	Scarborough	27
9	West Toronto	1
10	York	0

The data was all set, then let’s visualize this data by bar chart. The functions I used are import the matplotlib and seaborn in Python. These are good libraries to make date visualize. Below is the bar chart showing:

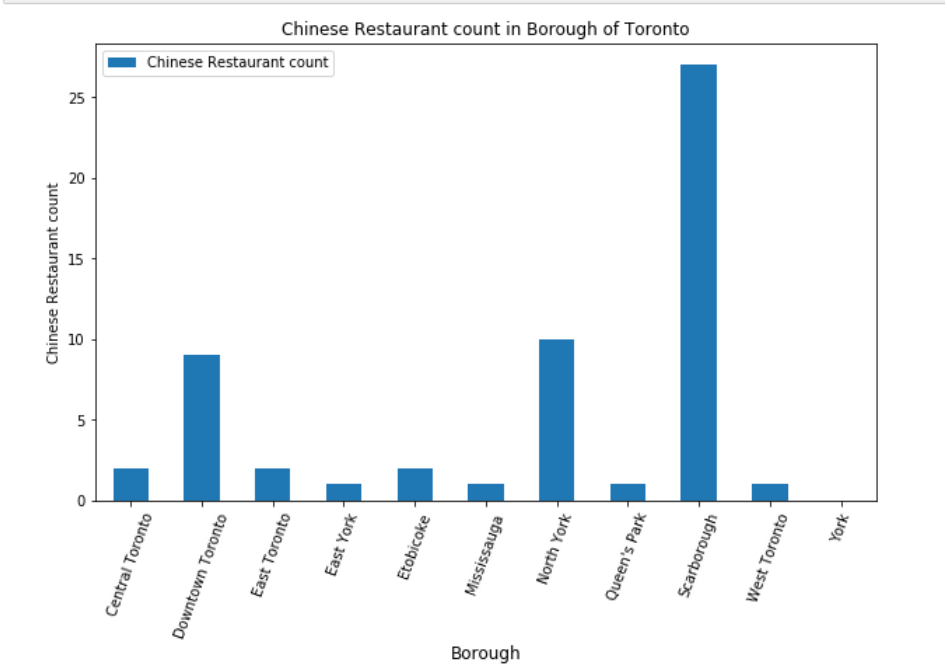

Figure 3: The bar chart for Chinese Restaurant count in Borough of Toronto

3.3 Relationship between Chinese population and Borough

Population is another significant factor to determine whether a best place to open a new Chinese restaurant. Therefore, I analyze the neighborhoods and discover the high Chinese population distributed in Toronto. First, I merge the tables above in data cleaning step 3. There are 4 tables amount and have same column name, then use the concat function in Python to combine all tables in one table, which give us 24 rows. Meanwhile, some values are missing, I just fill the NaN into missing cells. Here is the example of Ethnic table looks like:

	Riding	Population	Ethnic Origin #1	Origin #1 %	Ethnic Origin #2	Origin #2 %	Ethnic Origin #3	Origin #3 %	Ethnic Origin #4	Origin #4 %	Ethnic Origin #5	Origin #5 %	Ethnic Origin #6	Origin #6 %	Eth Ori
0	Spadina-Fort York	114315	English	16.4	Chinese	16.0	Irish	14.6	Canadian	14.0	Scottish	13.2	French	7.70	Germa
1	Beaches-East York	108435	English	24.2	Irish	19.9	Canadian	19.7	Scottish	18.9	French	8.7	German	8.40	NaN
2	Davenport	107395	Portuguese	22.7	English	13.6	Canadian	12.8	Irish	11.5	Italian	11.1	Scottish	11.00	NaN
3	Parkdale-High Park	106445	English	22.3	Irish	20.0	Scottish	18.7	Canadian	16.1	German	9.8	French	8.88	Polish
4	Toronto-Danforth	105395	English	22.9	Irish	19.5	Scottish	18.7	Canadian	18.4	Chinese	13.8	French	8.86	Germa
5	Toronto-St. Paul's	104940	English	18.5	Canadian	16.1	Irish	15.2	Scottish	14.8	Polish	10.3	German	7.90	Russia
6	University-Rosedale	100520	English	20.6	Irish	16.6	Scottish	16.3	Canadian	15.2	Chinese	14.7	German	8.70	Frenc
7	Toronto Centre	99590	English	15.7	Canadian	13.7	Irish	13.4	Scottish	12.6	Chinese	12.5	French	7.20	NaN
8	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	Eglinton-Lawrence	112925	Canadian	14.7	English	12.6	Polish	12.0	Filipino	11.0	Scottish	9.7	Italian	9.50	Irish

Figure 4: Example of Ethnic table

As we can see, there are some rows without Chinese, which means too less to show them on table. Therefore, I need to clean the data and extract the data that only have the information of Chinese population shown in table:

Table 6: Ethnicity only with Chinese population

	Ethnicity	Percentage	Population	Riding
0	Chinese	16.0	114315.0	Spadina-Fort York
1	Chinese	13.8	105395.0	Toronto-Danforth
2	Chinese	14.7	100520.0	University-Rosedale
3	Chinese	12.5	99590.0	Toronto Centre
4	Chinese	11.2	101790.0	Don Valley West
5	Chinese	8.9	93170.0	Don Valley East
6	Chinese	10.7	110450.0	Scarborough Centre
7	Chinese	7.2	108295.0	Scarborough Southwest
8	Chinese	7.1	101115.0	Scarborough-Guildwood

After get the information of percentage and population, I can get the exactly number of Chinese populations in new table:

Table 7: Borough and Chinese population:

	Riding	Chinese Population
0	Spadina-Fort York	18290.400
1	Toronto-Danforth	14544.510
2	University-Rosedale	14776.440
3	Toronto Centre	12448.750
4	Don Valley West	11400.480
5	Don Valley East	8292.130
6	Scarborough Centre	11818.150
7	Scarborough Southwest	7797.240
8	Scarborough-Guildwood	7179.165

Then visualize this date into bar chart:

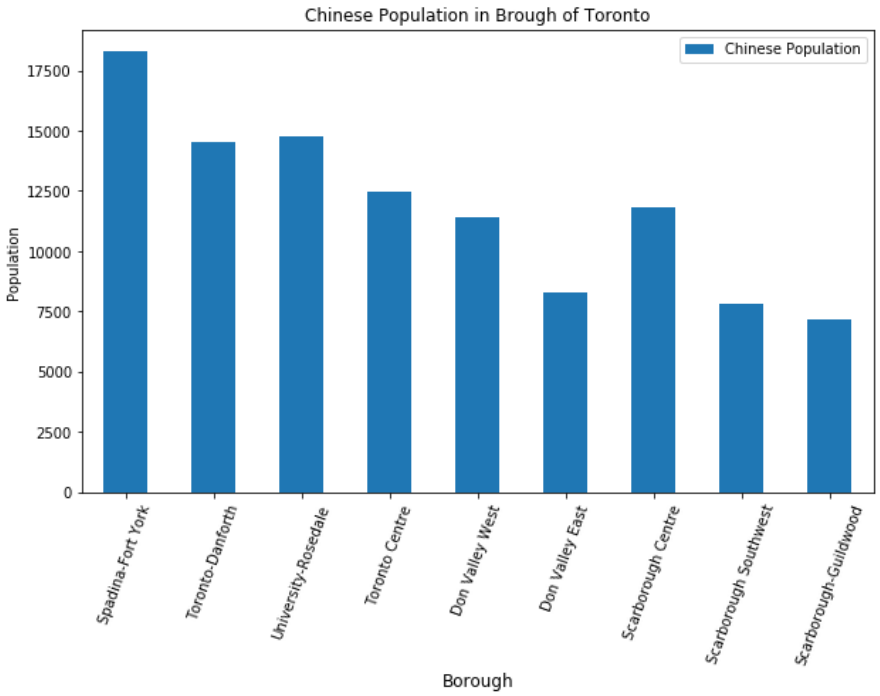

Figure 5: Bar chart for Chinese population in each Brough

4. Predictive Modelling

4.1 Clustering

To start the step of predictive modelling, we need to identify the best K value in modelling. K value stand for the number of clusters in the given dataset. I ploted the graph for K values by using range from 3 to 8, and the figure is shown:

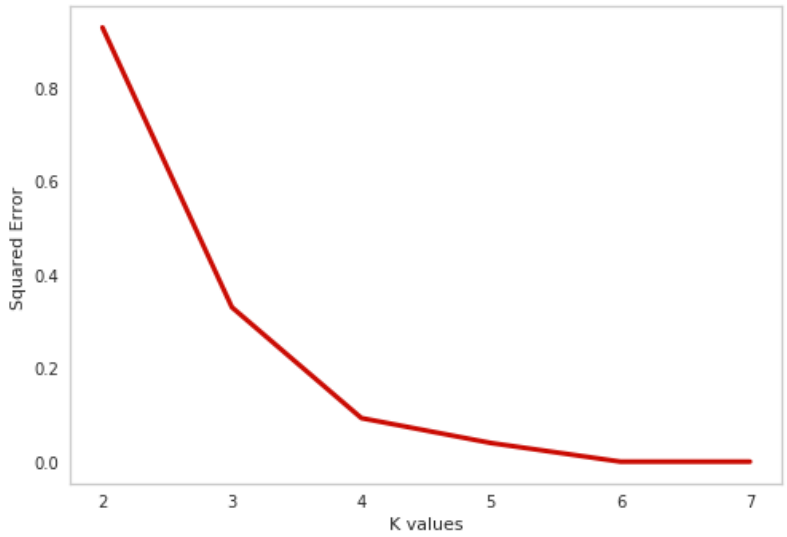

Figure 6: Trend of K values

According to this picture, I found that the K = 6 probably the best value since the squared error reached lowest point. Then cluster the dataset and get the total of 6

clusters labeled from 0 to 5. Here is the example for clustering table:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	2.0	1.0
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0.0	0.0
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	0.0	0.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	2.0	1.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	2.0	1.0

Figure 7: Example for clustering table

Meanwhile, I make a visualization for clustering table:

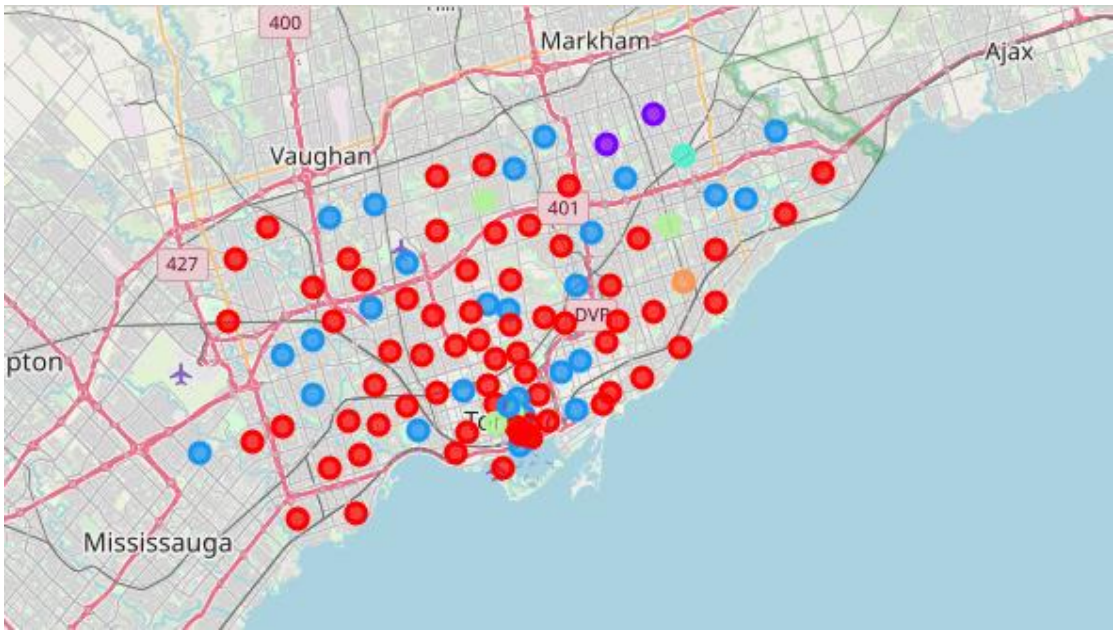

Figure 8: Visualization for clustering table

4.2Examine the Clusters

Let's examine the clusters labeled from 0 to 5 respectively:

Cluster 0:

```
print(df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 0].shape)
df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 0]
```

(67, 7)

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0.0	0.0
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	0.0	0.0
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476	0.0	0.0
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577	0.0	0.0
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476	0.0	0.0
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848	0.0	0.0
11	M1R	Scarborough	Maryvale, Wexford	43.750072	-79.295849	0.0	0.0
17	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556	0.0	0.0
19	M2L	North York	Silver Hills, York Mills	43.757490	-79.374714	0.0	0.0
20	M2M	North York	Newtonbrook, Willowdale	43.789053	-79.408493	0.0	0.0

Cluster 1:

```
print(df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 1].shape)
df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 1]
```

(2, 7)

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
14	M1V	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...	43.815252	-79.284577	1.0	5.0
15	M1W	Scarborough	L'Amoreaux West	43.799525	-79.318389	1.0	5.0

Cluster 2:

```
print(df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 3].shape)
df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 3]
```

(1, 7)

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
12	M1S	Scarborough	Agincourt	43.7942	-79.262029	3.0	7.0

Cluster 3:

```
print(df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 2].shape)
df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 2]
```

(27, 7)

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	2.0	1.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	2.0	1.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	2.0	1.0
13	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter	43.781638	-79.304302	2.0	1.0
16	M2H	North York	Hillcrest Village	43.803762	-79.363452	2.0	1.0
18	M2K	North York	Bayview Village	43.786947	-79.385975	2.0	1.0
24	M3A	North York	Parkwoods	43.753259	-79.329656	2.0	1.0
26	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923	2.0	1.0
28	M3J	North York	Northwood Park, York University	43.767980	-79.487262	2.0	1.0

Cluster 4:

```
print(df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 4].shape)
df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 4]
```

```
(4, 7)
```

0]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...	43.757410	-79.273304	4.0	2.0
21	M2N	North York	Willowdale South	43.770120	-79.408493	4.0	2.0
56	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	4.0	2.0
66	M5T	Downtown Toronto	Chinatown, Grange Park, Kensington Market	43.653206	-79.400049	4.0	2.0

Cluster 5:

```
print(df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 0].shape)
df_new_cn_rest.loc[df_new_cn_rest['Cluster Labels'] == 5]
```

```
(67, 7)
```

.]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Chinese Restaurant
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029	5.0	4.0

5. Results and Discussion

5.1 Results

As we can see in Figure 3, the existing Chinese restaurant were focused in 3 boroughs, which are Downtown Toronto, North York and Scarborough. Adding up the rest of boroughs and still cannot outnumber those three boroughs. I found there are universities around those boroughs, for example, UTSG in Downtown Toronto, York University in North York and UTSC in Scarborough, and there are also some colleges around. If you want to open a new Chinese restaurant in those 3 boroughs, then it will be a challenging work because of the highest competitive area. Due to those boroughs are hot area in Greater Toronto Area, the rental fee should be expensive, which means the higher cost. On the other hand, if your restaurant has fantastic dishes that can attract more people to eat, it will be the best choice since many students will pass through those areas.

Figure 5 reveals that the number of Chinese populations in Spadina-Fort York is the most, and order the rest by descending is University-Rosedale, Toronto-Danforth, Toronto Centre, Don Valley West and Scarborough Centre. From Figure 3 we already known that Downtown Toronto has densely populated with Chinese restaurant, hence leave Downtown area will be the better choice.

5.2 Discussion

According to the analysis, I find that open a new Chinese restaurant in Harbourfront Neighborhood is the best choice, which is located on south of Downtown Toronto in few kilometers. As I seen in cluster, it belongs to cluster labeled 0, which is no Chinese restaurant within a kilometer. This meaning of the least competition. Meanwhile, the Chinese population in Spadina-Fort York is the highest, and the distance between Harbourfront and Spadina-Fort York only few kilometers. The high number of Chinese populations will help a new restaurant to provide high possibility of customer visits. In my perspective, definitely Harbourfront is the optimal place to start a new Chinese restaurant. On the other hand, the population distribution in each neighborhood is based on 2016 census, it may have some distinctions. Also, the venues information I got only from Foursquare's API, this may occur some existing restaurants not shown or updated. These two factors may influence the result in this project. Although there are still some drawbacks can be improved, at least it gave us some useful insights.

6. Conclusion

In this project, I analyzed the different factors to discover the best option to open a new Chinese restaurant in Toronto. First, we used the Python libraries to extract data from Wikipedia or CSV files, and then used the Python library to clean up the data into the form that we wanted. In addition, we used Foursquare's API to explore the venues in neighborhoods of Toronto and get the information that we needed. Afterward, the functions in Python library, which are Seaborn and Matplotlib were used to visualize the data that we transformed before. Furthermore, machine learning is used to predict the output of given data and make the data can be visualized on the map through the Folium library in Python. In my perspective, machine learning has a wide range of applications, and it works on many challenging fields. Data always used to tell a story, but we need to use that story to solve the problem in daily life.