Applied Data Science Capstone

Gunay Ismayilova

Optimal Location for Opening a Chinese Restaurant in Toronto

1. Introduction:

1.1 Background

While opening a restaurant is one of the profitable businesses, wrong location may lead to failure within a year. Location should be the first consideration before opening a restaurant because it directly determines demand for the restaurant.

1.2 Business Problem:

The objective of this Capstone project is to find an optimal location for opening a Chinese Restaurant in Toronto. With its high income level, being home to diverse nationalities and many successful businesses, Toronto is where many entrepreneurs are willing to open a restaurant. With its unique cuisine and tastes, Chinese restaurant is in high demand all over the world, including in Toronto. So, for this project, we will assume that an entrepreneur plans to open a Chinese restaurant in Toronto. Our focus is to determine neighborhoods where the demand for Chinese restaurant is the highest.

1.3 Target Audience:

The target audience of the project are listed below:

1. Local entrepreneurs planning to open a Chinese restaurant in Toronto
2. Chinese businessmen planning to open a Chinese restaurant in Toronto
3. Chinese restaurant owners willing to open their branch in Toronto

All the stakeholders listed above are interested to find out an optimal neighborhood in Toronto to open their Chinese restaurant. Choosing a wrong neighborhood means failure of a business for them. This analysis will help them to better understand Toronto neighborhoods and find the best location for their restaurant.

2. Data acquisition and cleaning:

2.1 Data sources

a) I’m using “List of Postal code of Canada: M” (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I’m using “https://cocl.us/Geospatial_data” csv file to get all the geographical coordinates of the neighborhoods.

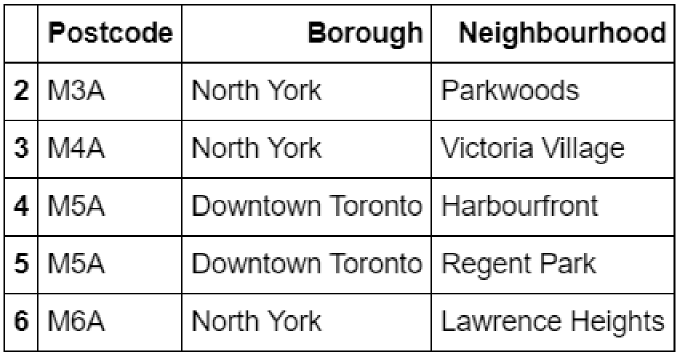
c) To get information about population size, population density and average income in each neighborhood of Toronto, I’m using “Demographics of Toronto neighborhoods” (https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods) wiki page.

d) To extract number of Chinese restaurants in each neighborhood, I’m using Foursquare’s explore API.

2.2. Data cleaning

a) Scraping Toronto Neighborhoods Table from Wikipedia

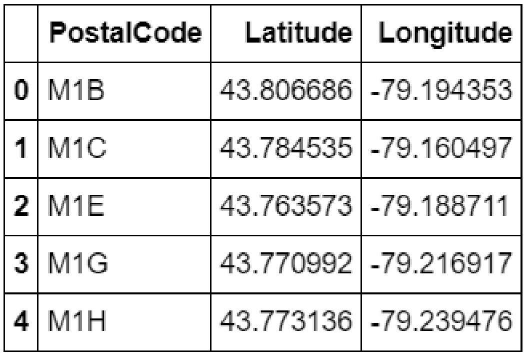
I scraped the following Wikipedia page, “List of Postal code of Canada: M” in order to obtain the data about the Toronto & the Neighborhoods in it. Then I dropped all the rows with missing data. Below is the cleaned dataframe of Toronto Neighborhoods:



	Postcode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights

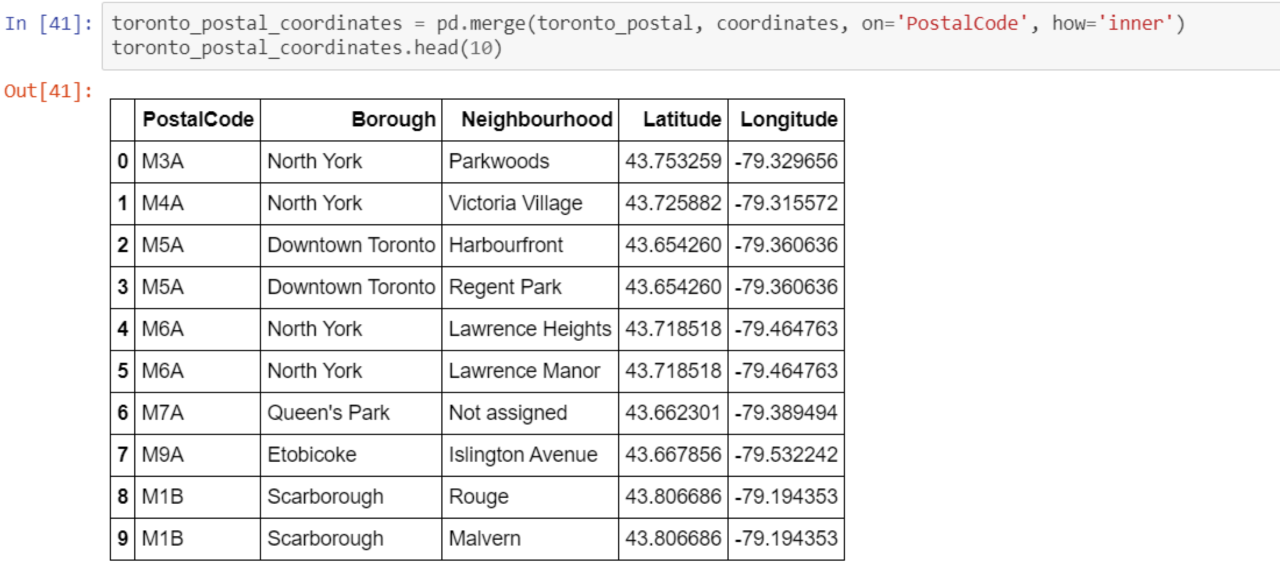
b) Adding geographical coordinates to the neighborhoods

Next step is adding the geographical coordinates to these neighborhoods. To do so I’m extracting geographical data from Geospatial Data csv file.



	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Then I'm combining it with the existing neighborhood dataframe by merging them both based on the postal code.



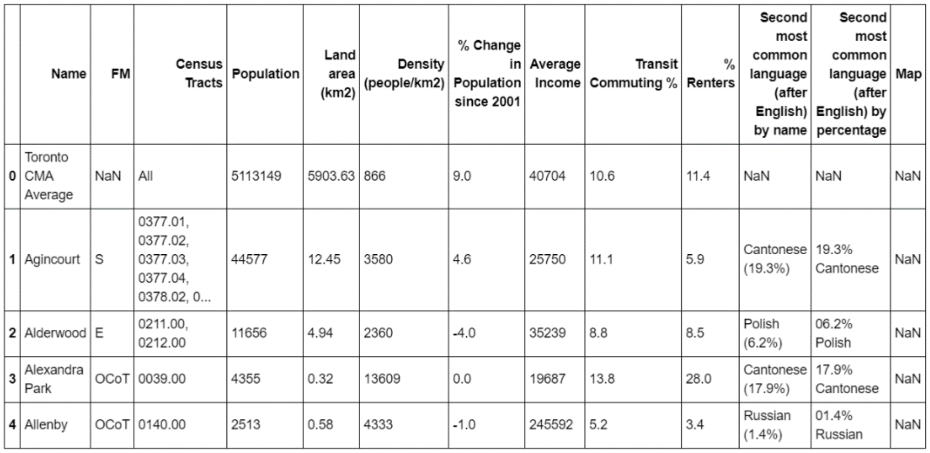
```
In [41]: toronto_postal_coordinates = pd.merge(toronto_postal, coordinates, on='PostalCode', how='inner')
         toronto_postal_coordinates.head(10)
```

Out[41]:

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
4	M6A	North York	Lawrence Heights	43.718518	-79.464763
5	M6A	North York	Lawrence Manor	43.718518	-79.464763
6	M7A	Queen's Park	Not assigned	43.662301	-79.389494
7	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
8	M1B	Scarborough	Rouge	43.806686	-79.194353
9	M1B	Scarborough	Malvern	43.806686	-79.194353

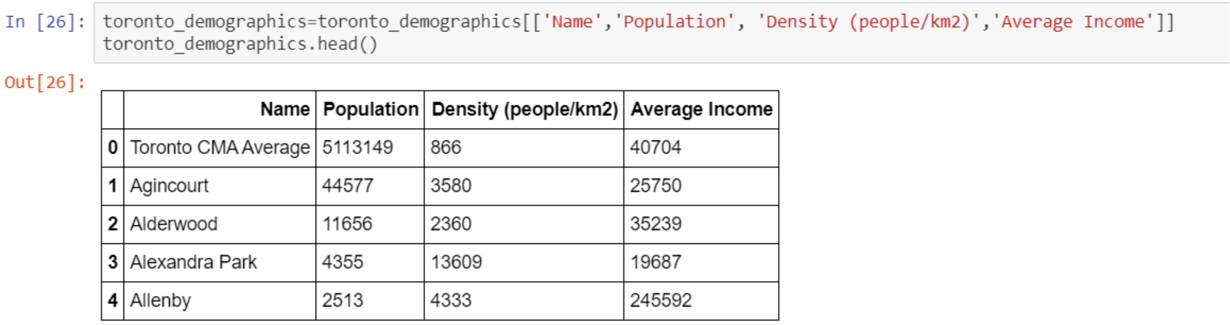
c) Scraping population size, population density and average income for each neighborhood of Toronto

Demand for Chinese restaurants in Toronto is likely to be more where population size is large, population density and average income are high. To acquire the indicators above I import "Demographics of Toronto neighborhoods" table from Wikipedia.



	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
0	Toronto CMA Average	NaN	All	5113149	5903.63	866	9.0	40704	10.6	11.4	NaN	NaN	NaN
1	Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...	44577	12.45	3580	4.6	25750	11.1	5.9	Cantonese (19.3%)	19.3% Cantonese	NaN
2	Alderwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5	Polish (6.2%)	06.2% Polish	NaN
3	Alexandra Park	OCoT	0039.00	4355	0.32	13609	0.0	19687	13.8	28.0	Cantonese (17.9%)	17.9% Cantonese	NaN
4	Allenby	OCoT	0140.00	2513	0.58	4333	-1.0	245592	5.2	3.4	Russian (1.4%)	01.4% Russian	NaN

Then I extract population size, population density and average income from that table.



```
In [26]: toronto_demographics=toronto_demographics[['Name','Population', 'Density (people/km2)','Average Income']]
         toronto_demographics.head()
```

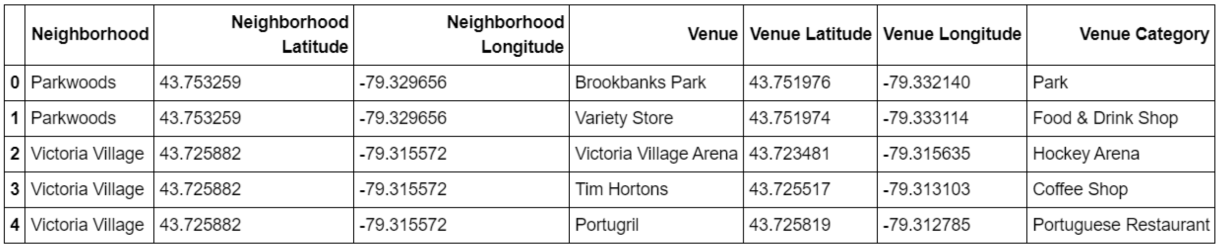
Out[26]:

	Name	Population	Density (people/km2)	Average Income
0	Toronto CMA Average	5113149	866	40704
1	Agincourt	44577	3580	25750
2	Alderwood	11656	2360	35239
3	Alexandra Park	4355	13609	19687
4	Allenby	2513	4333	245592

d) Using Foursquare's *explore API, acquire number of Chinese restaurants for each neighborhood.*

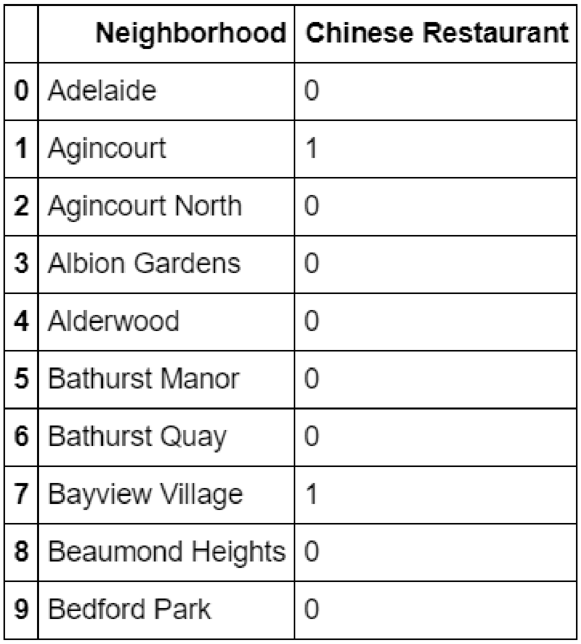
A new Chinese restaurant is likely to more less profitable when there are already many Chinese restaurants in its neighborhood and competition is high. Therefore, a new Chinese restaurant should be opened in a neighborhood with the minimum number of Chinese restaurants. For that, I will acquire number of Chinese restaurants in each neighborhood using Foursquare's API.

Foursquare data is very comprehensive, and it powers location data for Apple, Uber etc. For this business problem I have used it's explore API to get venues in Toronto with their locations and categories. With some data manipulation, I got neighborhood of each venue acquired from Foursquare.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Then by using one hot encoding I found number of venues in each category. After that I grouped venue categories by Neighborhood. At the end I found number of Chinese restaurants in each neighborhood by filtering only Chinese restaurants.



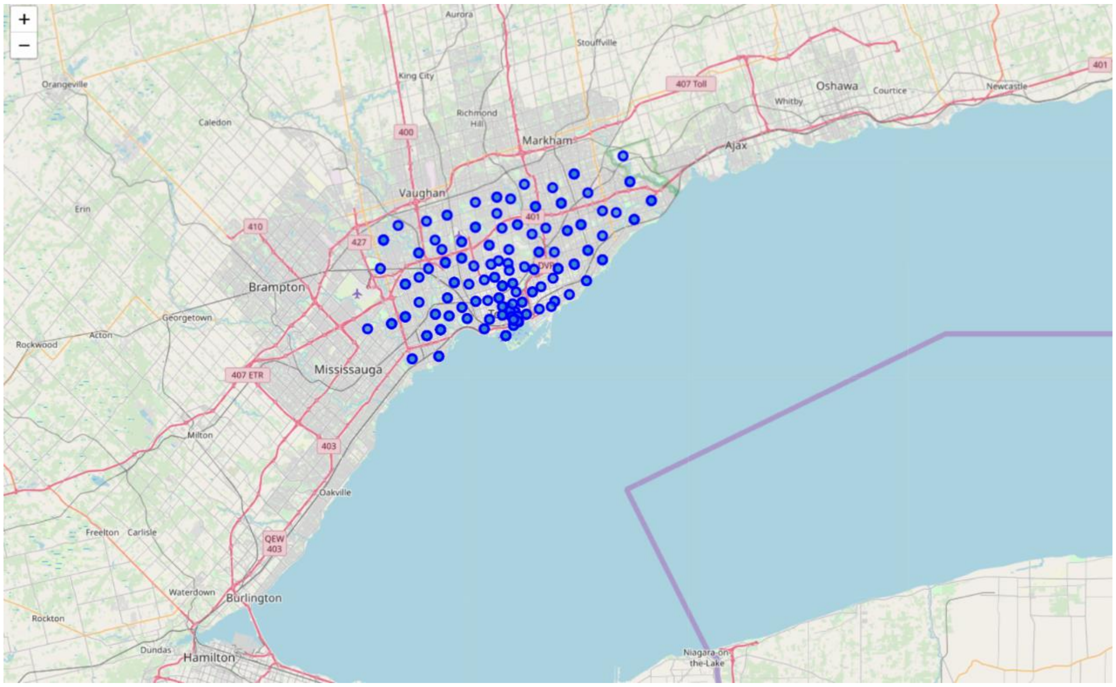
	Neighborhood	Chinese Restaurant
0	Adelaide	0
1	Agincourt	1
2	Agincourt North	0
3	Albion Gardens	0
4	Alderwood	0
5	Bathurst Manor	0
6	Bathurst Quay	0
7	Bayview Village	1
8	Beaumond Heights	0
9	Bedford Park	0

3. Methodology

3.1 Exploratory Data Analysis

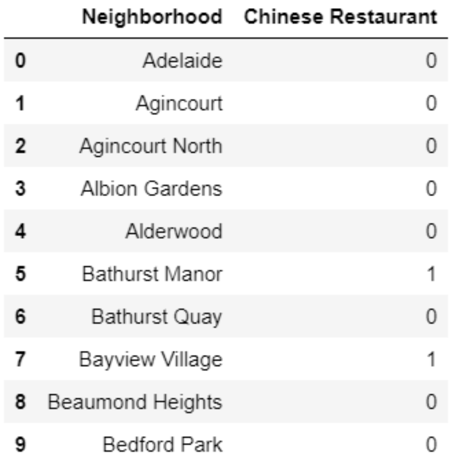
3.1.1 Visualize Toronto neighborhoods

Firstly, it is necessary to visualize the neighborhoods of Toronto to obtain general understanding of their location. Folium map is used to visualize Toronto neighborhoods. The map below shows that neighborhoods are located densely near downtown Toronto and spread out as distance from downtown increases. This is important because while some neighborhoods might not have many Chinese restaurants, if they are located near downtown, adjacent regions may drastically impact profitability of the restaurant.



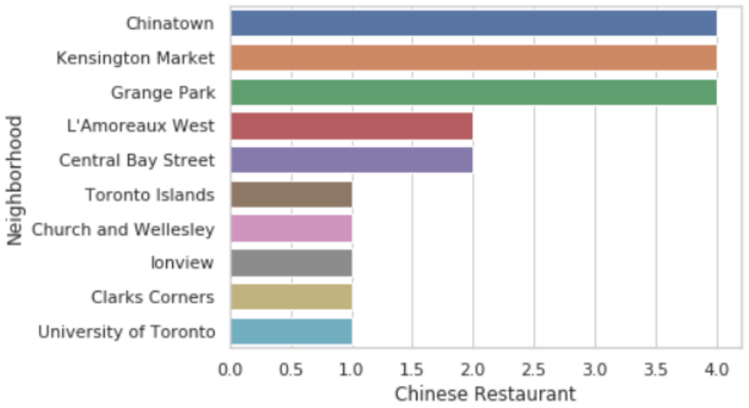
3.1.2 Calculate number of Chinese restaurants in each neighborhood

Now that we have visualized Toronto neighborhoods, let's use Foursquare API to explore the venues in each neighborhood. We use Explore API of Foursquare to return top 200 venues within 2000 meters of the latitude and longitude of each postal code. Then extracted venue categories were encoded using one-hot encoding and total number of Chinese restaurants for each neighborhood is calculated.



	Neighborhood	Chinese Restaurant
0	Adelaide	0
1	Agincourt	0
2	Agincourt North	0
3	Albion Gardens	0
4	Alderwood	0
5	Bathurst Manor	1
6	Bathurst Quay	0
7	Bayview Village	1
8	Beaumond Heights	0
9	Bedford Park	0

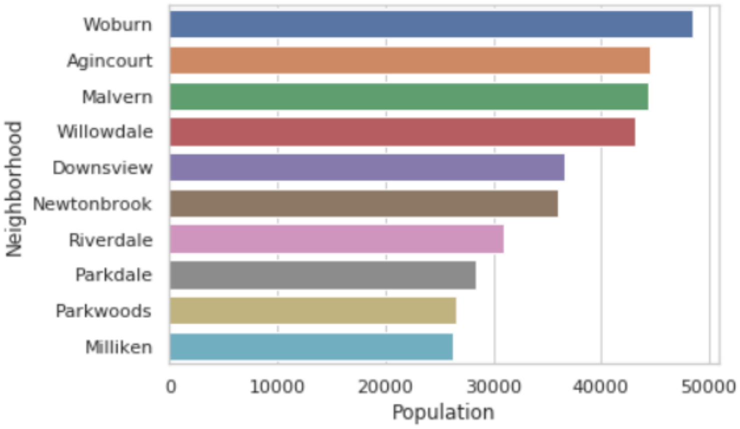
Finding neighborhoods with many Chinese restaurants is important because these are the regions that are not suitable to open a new Chinese restaurant. In these neighborhoods, competition is already high, so if we open a new Chinese restaurant in these regions, profitability of our restaurant will be low. Now let's visualize the top 10 neighborhoods with the maximum number of Chinese restaurants:



The neighborhoods in the bar plot are the ones to avoid opening a new Chinese restaurant in Toronto.

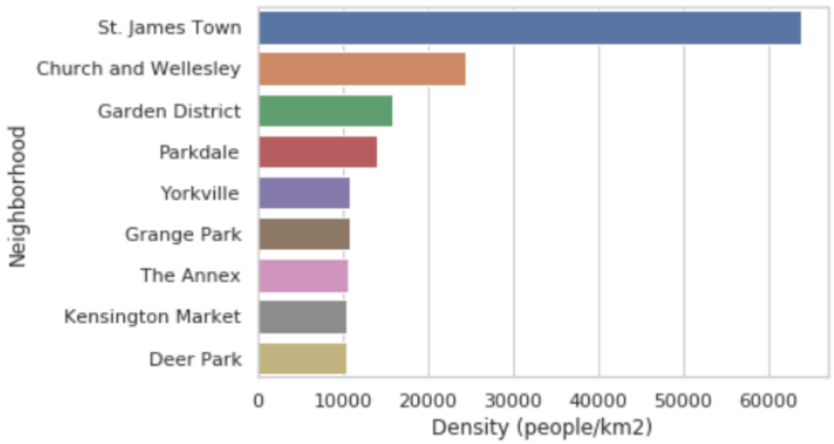
3.1.3 Visualize neighborhoods with high population

Another indicator of an ideal neighborhood to open a new Chinese restaurant is high population. In these regions as there are more people, more customers are expected to our new restaurant. Now let's visualize the top 10 neighborhoods with the maximum number of populations.



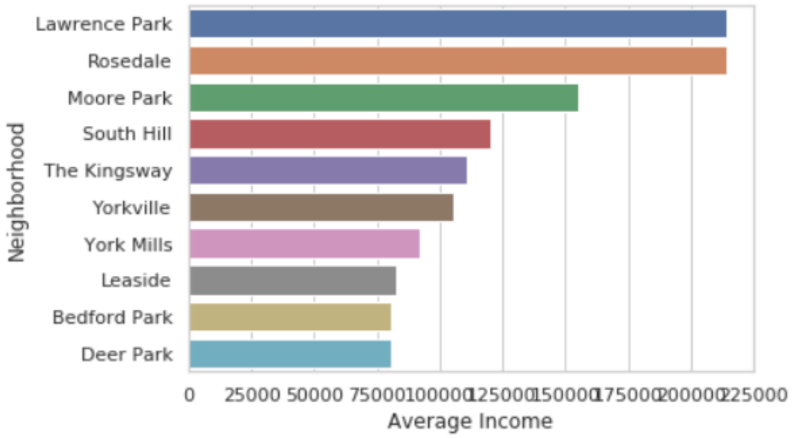
3.1.4 Visualize neighborhoods with high population density

It is also expected that demand for our new restaurant will be high in the neighborhoods with high population density. Now let’s visualize the top 10 neighborhoods with the maximum population density.



3.1.5 Visualize neighborhoods with high average income

It is always a good idea to open a new business in a rich neighborhood because in these neighborhoods spending power of population is high which means our business can make more profit. This general rule is true for opening a new restaurant in a rich neighborhood of Toronto. Here we measure how rich a neighborhood is with the average income of the people living there. So, let’s visualize the top 10 neighborhoods with the maximum average income.

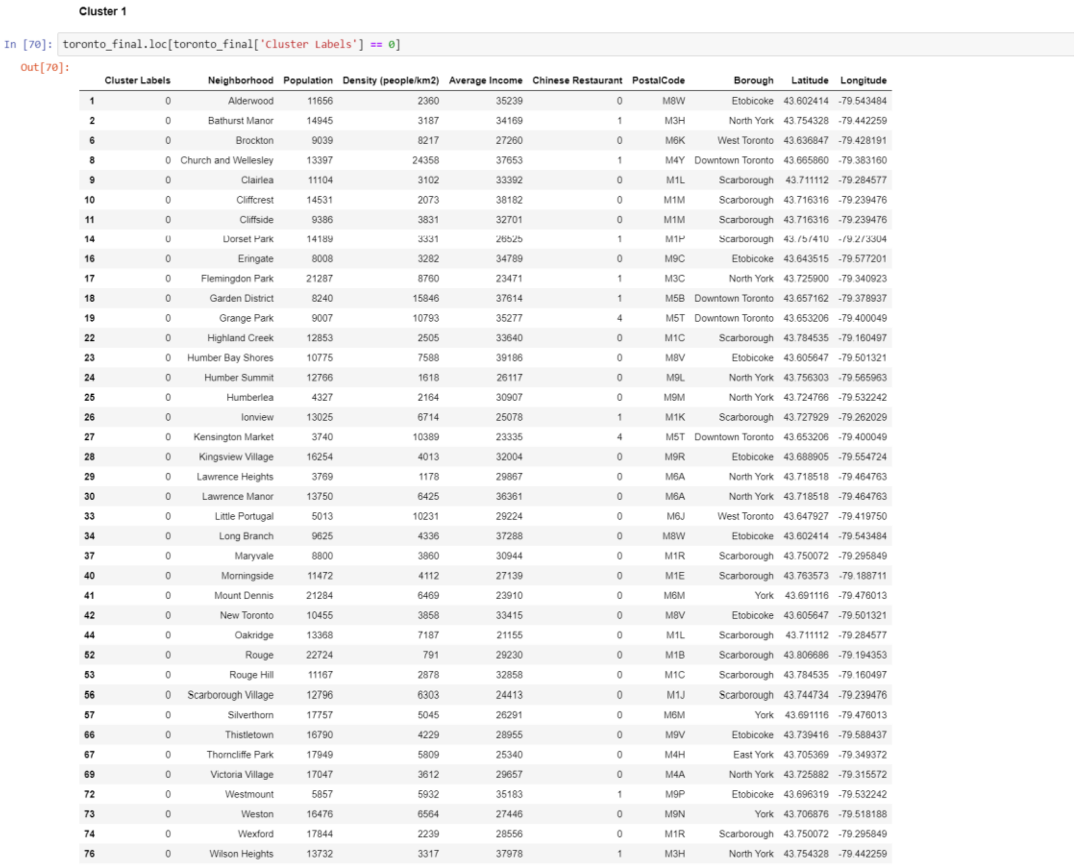


3.2 Modelling

To find ideal neighborhoods to open Chinese restaurant in Toronto, we will be using K-means clustering, one of the most commonly used form of unsupervised machine learning. We will use cluster size of 5 for this project. The reason to conduct K-means clustering is to reveal a bunch of similar neighborhoods that are densely populated, rich and doesn't have any Chinese restaurants. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well.

4. Results

After running K-means clustering we can start analyzing each cluster to find the one to open our Chinese restaurant. Let's see the neighborhoods in the first cluster below:



Cluster 1

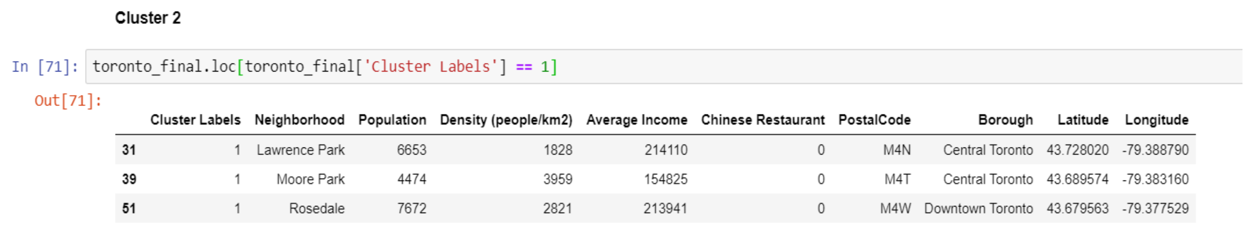
```
In [70]: toronto_final.loc[toronto_final['Cluster Labels'] == 0]
```

Out[70]:

	Cluster Labels	Neighborhood	Population	Density (people/km2)	Average Income	Chinese Restaurant	PostalCode	Borough	Latitude	Longitude
1	0	Alderwood	11656	2360	35239	0	M8W	Etobicoke	43.602414	-79.543484
2	0	Bathurst Manor	14945	3187	34169	1	M3H	North York	43.754328	-79.442259
6	0	Brockton	9039	8217	27260	0	M6K	West Toronto	43.636847	-79.428191
8	0	Church and Wellesley	13397	24358	37653	1	M4Y	Downtown Toronto	43.665860	-79.383160
9	0	Clairlea	11104	3102	33392	0	M1L	Scarborough	43.711112	-79.284577
10	0	Cliffcrest	14531	2073	38182	0	M1M	Scarborough	43.716316	-79.239476
11	0	Cliffside	9386	3831	32701	0	M1M	Scarborough	43.716316	-79.239476
14	0	Dorset Park	14189	3331	26525	1	M1P	Scarborough	43.757410	-79.273304
16	0	Eringate	8008	3282	34789	0	M9C	Etobicoke	43.643515	-79.577201
17	0	Flemingdon Park	21287	8760	23471	1	M3C	North York	43.725900	-79.340923
18	0	Garden District	8240	15846	37614	1	M5B	Downtown Toronto	43.657162	-79.378937
19	0	Grange Park	9007	10793	35277	4	M5T	Downtown Toronto	43.653206	-79.400049
22	0	Highland Creek	12853	2505	33640	0	M1C	Scarborough	43.784535	-79.160497
23	0	Humber Bay Shores	10775	7588	39186	0	M8V	Etobicoke	43.605647	-79.501321
24	0	Humber Summit	12766	1618	26117	0	M9L	North York	43.756303	-79.565963
25	0	Humberlea	4327	2164	30907	0	M9M	North York	43.724766	-79.532242
26	0	Ionview	13025	6714	25078	1	M1K	Scarborough	43.727929	-79.262029
27	0	Kensington Market	3740	10389	23335	4	M5T	Downtown Toronto	43.653206	-79.400049
28	0	Kingsview Village	16254	4013	32004	0	M9R	Etobicoke	43.688905	-79.554724
29	0	Lawrence Heights	3769	1178	29867	0	M6A	North York	43.718518	-79.464763
30	0	Lawrence Manor	13750	6425	36361	0	M6A	North York	43.718518	-79.464763
33	0	Little Portugal	5013	10231	29224	0	M6J	West Toronto	43.647927	-79.419750
34	0	Long Branch	9625	4336	37288	0	M8W	Etobicoke	43.602414	-79.543484
37	0	Maryvale	8800	3860	30944	0	M1R	Scarborough	43.750072	-79.295849
40	0	Morningside	11472	4112	27139	0	M1E	Scarborough	43.763573	-79.188711
41	0	Mount Dennis	21284	6469	23910	0	M6M	York	43.691116	-79.476013
42	0	New Toronto	10455	3858	33415	0	M8V	Etobicoke	43.605647	-79.501321
44	0	Oakridge	13368	7187	21155	0	M1L	Scarborough	43.711112	-79.284577
52	0	Rouge	22724	791	29230	0	M1B	Scarborough	43.806686	-79.194353
53	0	Rouge Hill	11167	2878	32858	0	M1C	Scarborough	43.784535	-79.160497
56	0	Scarborough Village	12796	6303	24413	0	M1J	Scarborough	43.744734	-79.239476
57	0	Silverthorn	17757	5045	26291	0	M6M	York	43.691116	-79.476013
66	0	Thistletown	16790	4229	28955	0	M9V	Etobicoke	43.739416	-79.588437
67	0	Thorncliffe Park	17949	5809	25340	0	M4H	East York	43.705369	-79.349372
69	0	Victoria Village	17047	3612	29657	0	M4A	North York	43.725882	-79.315572
72	0	Westmount	5857	5932	35183	1	M9P	Etobicoke	43.696319	-79.532242
73	0	Weston	16476	6564	27446	0	M9N	York	43.706876	-79.518188
74	0	Wexford	17844	2239	28556	0	M1R	Scarborough	43.750072	-79.295849
76	0	Wilson Heights	13732	3317	37978	1	M3H	North York	43.754328	-79.442259

Cluster one is the biggest cluster among 5 with 39 neighborhoods.

Cluster two on the other hand, is the smallest among 5 with only 3 neighborhoods:



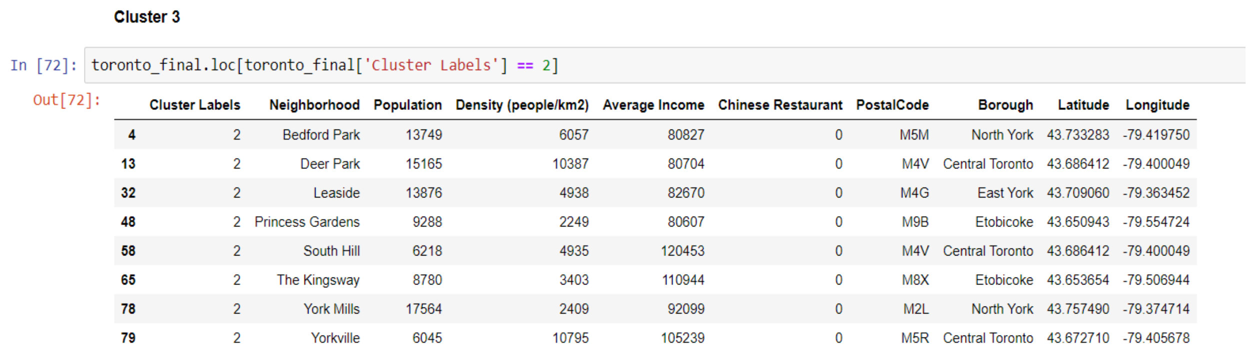
Cluster 2

```
In [71]: toronto_final.loc[toronto_final['Cluster Labels'] == 1]
```

Out[71]:

	Cluster Labels	Neighborhood	Population	Density (people/km2)	Average Income	Chinese Restaurant	PostalCode	Borough	Latitude	Longitude
31	1	Lawrence Park	6653	1828	214110	0	M4N	Central Toronto	43.728020	-79.388790
39	1	Moore Park	4474	3959	154825	0	M4T	Central Toronto	43.689574	-79.383160
51	1	Rosedale	7672	2821	213941	0	M4W	Downtown Toronto	43.679563	-79.377529

Now let’s view cluster three with 8 neighborhoods:



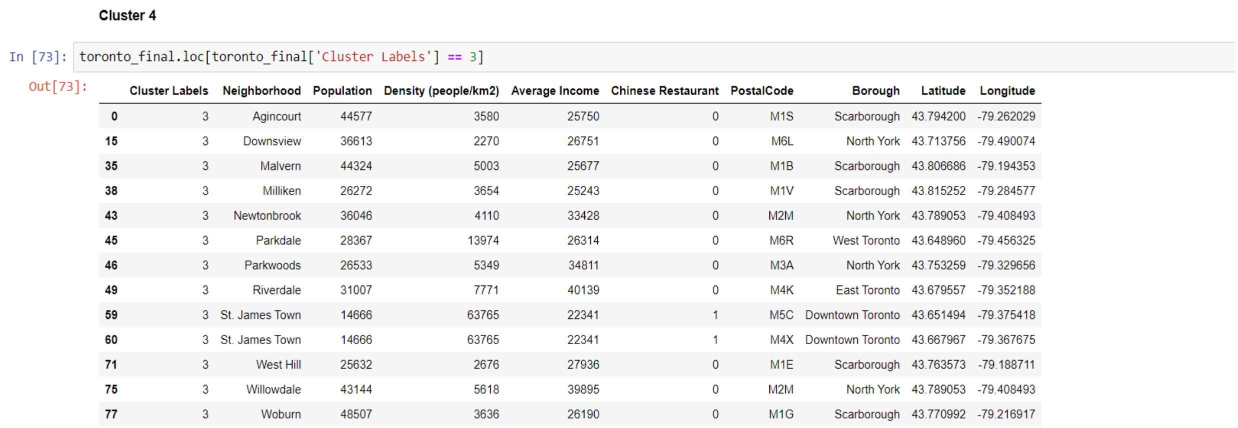
Cluster 3

```
In [72]: toronto_final.loc[toronto_final['Cluster Labels'] == 2]
```

Out[72]:

	Cluster Labels	Neighborhood	Population	Density (people/km2)	Average Income	Chinese Restaurant	PostalCode	Borough	Latitude	Longitude
4	2	Bedford Park	13749	6057	80827	0	M5M	North York	43.733283	-79.419750
13	2	Deer Park	15165	10387	80704	0	M4V	Central Toronto	43.686412	-79.400049
32	2	Leaside	13876	4938	82670	0	M4G	East York	43.709060	-79.363452
48	2	Princess Gardens	9288	2249	80607	0	M9B	Etobicoke	43.650943	-79.554724
58	2	South Hill	6218	4935	120453	0	M4V	Central Toronto	43.686412	-79.400049
65	2	The Kingsway	8780	3403	110944	0	M8X	Etobicoke	43.653654	-79.506944
78	2	York Mills	17564	2409	92099	0	M2L	North York	43.757490	-79.374714
79	2	Yorkville	6045	10795	105239	0	M5R	Central Toronto	43.672710	-79.405678

Cluster four is moderately large with 13 neighborhoods:



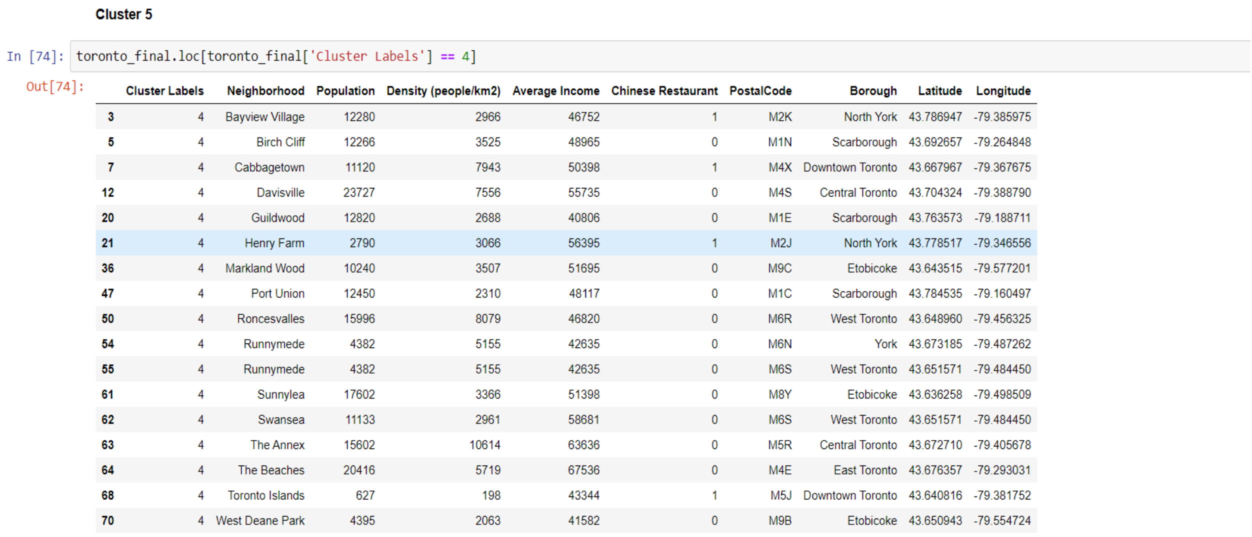
Cluster 4

```
In [73]: toronto_final.loc[toronto_final['Cluster Labels'] == 3]
```

Out[73]:

	Cluster Labels	Neighborhood	Population	Density (people/km2)	Average Income	Chinese Restaurant	PostalCode	Borough	Latitude	Longitude
0	3	Agincourt	44577	3580	25750	0	M1S	Scarborough	43.794200	-79.262029
15	3	Downsview	36613	2270	26751	0	M6L	North York	43.713756	-79.490074
35	3	Malvern	44324	5003	25677	0	M1B	Scarborough	43.806686	-79.194353
38	3	Milliken	26272	3654	25243	0	M1V	Scarborough	43.815252	-79.284577
43	3	Newtonbrook	36046	4110	33428	0	M2M	North York	43.789053	-79.408493
45	3	Parkdale	28367	13974	26314	0	M6R	West Toronto	43.648960	-79.456325
46	3	Parkwoods	26533	5349	34811	0	M3A	North York	43.753259	-79.329656
49	3	Riverdale	31007	7771	40139	0	M4K	East Toronto	43.679557	-79.352188
59	3	St. James Town	14666	63765	22341	1	M5C	Downtown Toronto	43.651494	-79.375418
60	3	St. James Town	14666	63765	22341	1	M4X	Downtown Toronto	43.667967	-79.367675
71	3	West Hill	25632	2676	27936	0	M1E	Scarborough	43.763573	-79.188711
75	3	Willowdale	43144	5618	39895	0	M2M	North York	43.789053	-79.408493
77	3	Woburn	48507	3636	26190	0	M1G	Scarborough	43.770992	-79.216917

Finally, fifth cluster is also large enough with 17 neighborhoods:



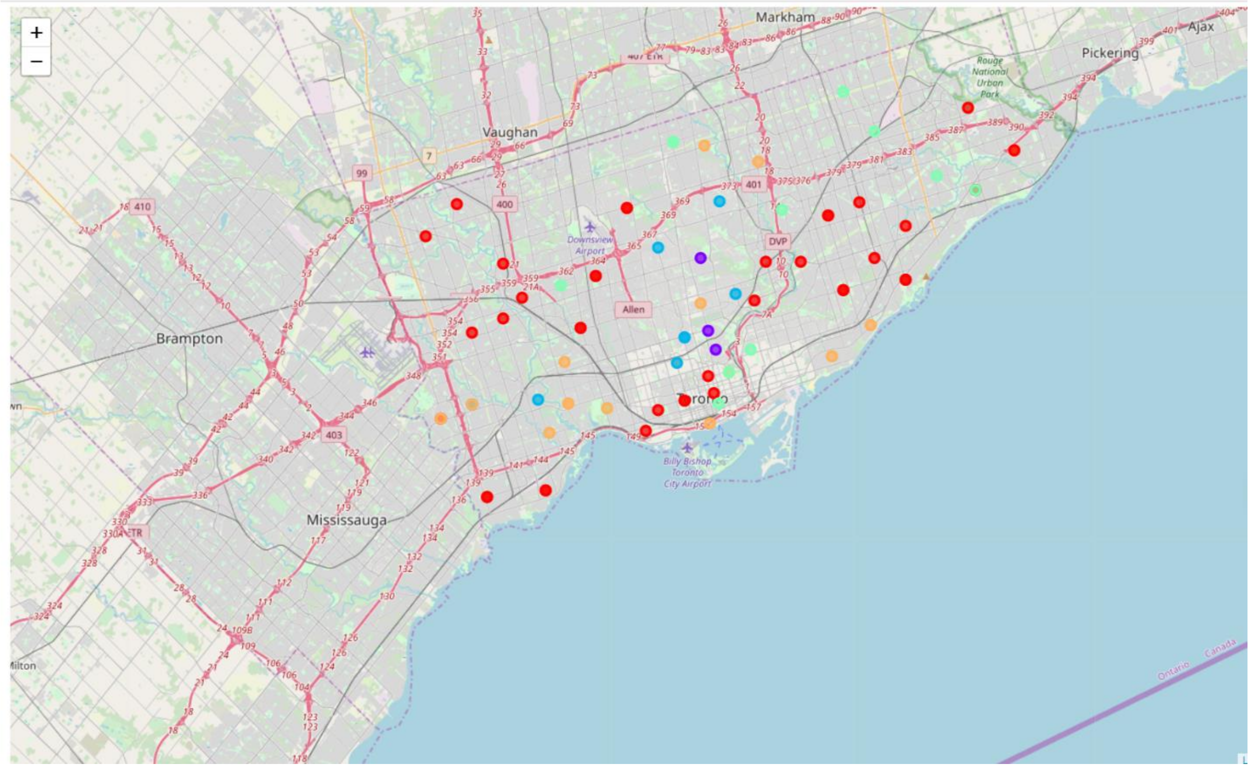
Cluster 5

```
In [74]: toronto_final.loc[toronto_final['Cluster Labels'] == 4]
```

Out[74]:

	Cluster Labels	Neighborhood	Population	Density (people/km2)	Average Income	Chinese Restaurant	PostalCode	Borough	Latitude	Longitude
3	4	Bayview Village	12280	2966	46752	1	M2K	North York	43.786947	-79.385975
5	4	Birch Cliff	12266	3525	48965	0	M1N	Scarborough	43.692657	-79.264848
7	4	Cabbagetown	11120	7943	50398	1	M4X	Downtown Toronto	43.667967	-79.367675
12	4	Davisville	23727	7556	55735	0	M4S	Central Toronto	43.704324	-79.388790
20	4	Guildwood	12820	2688	40806	0	M1E	Scarborough	43.763573	-79.188711
21	4	Henry Farm	2790	3066	56395	1	M2J	North York	43.778517	-79.346556
36	4	Markland Wood	10240	3507	51695	0	M9C	Etobicoke	43.643515	-79.577201
47	4	Port Union	12450	2310	48117	0	M1C	Scarborough	43.784535	-79.160497
50	4	Roncesvalles	15996	8079	46820	0	M6R	West Toronto	43.648960	-79.456325
54	4	Runnymede	4382	5155	42635	0	M6N	York	43.673185	-79.487262
55	4	Runnymede	4382	5155	42635	0	M6S	West Toronto	43.651571	-79.484450
61	4	Sunnylea	17602	3366	51398	0	M8Y	Etobicoke	43.636258	-79.498509
62	4	Swansea	11133	2961	58681	0	M6S	West Toronto	43.651571	-79.484450
63	4	The Annex	15602	10614	63636	0	M5R	Central Toronto	43.672710	-79.405678
64	4	The Beaches	20416	5719	67536	0	M4E	East Toronto	43.676357	-79.293031
68	4	Toronto Islands	627	198	43344	1	M5J	Downtown Toronto	43.640816	-79.381752
70	4	West Deane Park	4395	2063	41582	0	M9B	Etobicoke	43.650943	-79.554724

Now let’s visualize the clusters in the map using Folium maps:



Each cluster is color coded for the ease of presentation. We can observe that majority of neighborhoods fall into red cluster which is Cluster one and only three of the neighborhoods fall into blue cluster which is Cluster two. Some neighborhoods in the same cluster are densely populated while others are quite separate from one another.

5. Discussion

From the results of clustering algorithm, it is determined that neighborhoods corresponding to cluster two are the best choice for opening a new Chinese restaurant. The main factors considered in choosing cluster 2 are number of existing Chinese restaurants, average income and population density. Firstly, none of the three restaurants in cluster two, to name Lawrence Park, Moore Park and Rosedale have any Chinese restaurants. It implies that we will not suffer competition in these neighborhoods. Secondly, cluster two are the only cluster among 5, with neighborhoods that all have 6-digit average income. So, it is the best idea to open a new restaurant in one of these richest neighborhoods. Thirdly, all three neighborhoods are moderately densely populated meaning that we can expect enough number of customers. Overall, I recommend opening a new Chinese restaurant in one of these three neighborhoods: Lawrence Park, Moore Park and Rosedale.

6. Conclusion

In conclusion, opening a restaurant is a complex task that can lead to a large monetary loss if not done properly. Thus, extensive research about the area would greatly increase the likelihood of the restaurant succeeding. From the project above, I demonstrated the workflow for determining the neighborhood to open a new Chinese restaurant. To be more specific I found out that Lawrence Park, Moore Park and Rosedale are the best neighborhoods to open a Chinese restaurant in Toronto. More studies can be conducted to improve my model. For example, ethnic origin of each neighborhood can be determined and neighborhoods with the greatest number of Chinese people can be considered in the model. Adding more relevant variables can increase accuracy of this model.