Best Neighborhood in Toronto to Open Italian Restaurants

1 Introduction

1.1 Background

Toronto is recognized as one of the most multicultural and cosmopolitan cities in the world. Immigrants from all over the word live here together, which makes this city versatile and pluralistic in all respects, including foods. Here, we can find almost all types of restaurants, Italian, Japanese, Mexican, Chinese, Korean, etc., in each neighborhood. For food lovers, Toronto gives them the opportunity to explore the world in one city.

1.2 Problem

The pluralism of Toronto is great, but sometimes this may also lead to confusion when someone is deciding where to go for a certain type of restaurants, especially for tourists and new immigrants.

This project aims to analyze the data from Foursquare and determine which neighborhoods are the best to explore for different types of foods.

1.3 Audience

As mentioned above, the result of this project can first be used to help people decide which neighborhoods to explore when they want to try certain types of foods.

Meanwhile, according to Game Theory, or more specifically, “Hotelling’s Model of Spatial Competition”, these neighborhoods are also the best locations to open specific types of restaurants, so this will also help someone who is starting a new restaurant business.

Beyond this, apps like TripAdvisor, or social media influencers, can also use this result to create specialized articles and posts.

2 Data

2.1 Data Sources

In order to address the goal described above, obviously, we will first need the neighborhoods information of Toronto, more specifically, their names. This can be found in a Wiki page here: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The next step is to get the location of each neighborhood (their latitude and longitude). This information is needed so that we can later visualize our results in a map.

The last step is to use the location information gained above to get the restaurants info in each neighborhood by calling APIs to Foursquare. This includes the restaurant name, the type of restaurant (e.g. Italian, Mexican, etc.), how many people visited this place and how many people liked this place.

2.2 Data Cleaning

2.2.1 Neighborhoods Data

There are several tables in the Wiki page mentioned above but we only need the 1st one, which is the postal code table. We can use built-in method "read_html" from pandas to get all the tables and then only keep the 1st one.

After that, we need to remove those rows that don't contain neighborhood information. Those rows always contain "Not assigned" in the string, so we can simply remove all the rows contain that.

The last step is to split the string in each row into postal code, borough name and neighborhoods names. Note that some boroughs don't have more detailed neighborhood information, in this case, we will use the borough name as the neighborhoods' name.

2.2.2 Restaurants Data

The restaurants data from Foursquare sometimes contains missing information (NaN), as this is a rare case that doesn't happen frequently, we will just remove those restaurants from our data, which should not have too much impact on our final result.

2.3 Data Usage

The neighborhoods data will help us visualize the final results and get the position info of each neighborhood (which will also be used to construct the map). To get restaurants info from Foursquare, we need the location info gained above and pass them in as the parameters of APIs. For the data we get from Foursquare, we will analyze how many restaurants of the same type are there in each neighborhood, how popular are they, and how many people liked those restaurants.

With the result from the above analysis, we can then group all neighborhoods into several groups for different types of restaurants by using K-Means. By looking closely into each label, we should be able to rank the neighborhoods, in terms of where should be the 1st choice to open a restaurant of a specific type.

3 Methodology

3.1 Toronto Borough-Neighborhoods Data

In the Wiki page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, we can find all the postal code – borough – neighborhoods mapping. After removing all the rows with "Not assigned" for borough and using borough name as the neighborhood name for those neighborhoods who don't have names, we can get such a table (only showing first 5 rows):

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park / Ontario Provincial Government	Queen's Park / Ontario Provincial Government

Table 1. Postal Code – Borough - Neighborhoods Table

3.2 *Neighborhood-Location Data*

Now we have all the neighborhoods of Toronto, but only names. In order to later visualize the results and get restaurants information from Foursquare, we also need the location of each neighborhood, more specifically, latitude and longitude.

To achieve this, we can extract the data from the csv file here: http://cocl.us/Geospatial_data. This file maps postal code to latitude and longitude. After using Python to convert this into a dictionary, we can now append the location information to the neighborhoods table above, which will give us such a table (only showing first 5 rows):

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.7532586	-79.3296565
1	M4A	North York	Victoria Village	43.7258823	-79.3155716
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6542599	-79.3606359
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.4647633
4	M7A	Queen's Park / Ontario Provincial Government	Queen's Park / Ontario Provincial Government	43.6623015	-79.3894938

Table 2. Neighborhoods – Location (Latitude and Longitude) Table

Using the above table, we can now generate a map for neighborhoods in Toronto, like the one below. We use Lawrence Park as the centre of the map and use 12 as the initial zoom value, so that most neighborhoods can show up in the map clearly.

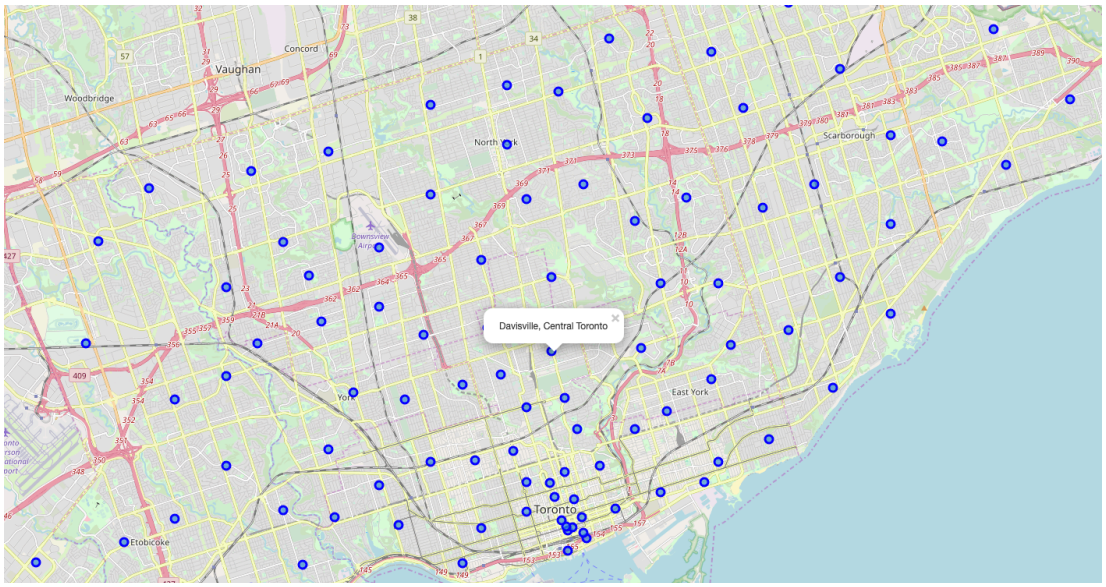

Figure 1. The Map of Toronto Neighborhoods

3.3 Restaurants Data

The last step for collecting the data before we do the real analysis is to get the restaurants information. This information is available in Foursquare, by calling APIs to it, we can get info like the category of this restaurant, price tier, how many people have visited it and how many people liked it.

To get enough data for the analysis, I decide to include around 50 restaurants for each neighborhood, which is the maximum amount of venues the explore API can return. I tried several times and noticed that some neighborhoods don't have too many restaurants, so if I want to get enough data for each neighborhood, I need to set radius as 2000 meters so that each neighborhood can return around 50 restaurants.

We can get the restaurants' information by calling Foursquare's explore API, which is 'GET https://api.foursquare.com/v2/venues/explore'. To utilize this API, we need to pass in the location info of each neighborhood, so that the API knows which area to search for. We already have the location data in Neighborhood – Location table above, so we only need to loop through the table and call the corresponding API one by one.

However, the above API will return all kinds of venues if we don't put a restraint, including parks, gas stations, banks, etc. To avoid this, we need to set the parameter "section" to "food" when calling the API. Now we only have restaurants in our API returns. As we only need the category information from the result, so I only kept restaurants' names and their categories. After combining it with the Neighborhood – Location table above, we now get the below table. The API returned 4489 restaurants in total (around 50 restaurants for each neighborhood), I'm only showing the first 5 rows here:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Restaurant	Main Category
0	Parkwoods	43.753258	-79.329659	Allwyn's Bakery	Caribbean Restaurant
1	Parkwoods	43.753258	-79.329659	Darband Restaurant	Middle Eastern Restaurant
2	Parkwoods	43.753258	-79.329659	Me Va Me Kitchen Express	Mediterranean Restaurant
3	Parkwoods	43.753258	-79.329659	Tim Hortons	Café
4	Parkwoods	43.753258	-79.329659	The Captain's Boil	Seafood Restaurant

Table 3. Restaurant – Name - Category Table

Because stats (how many people have visited this place) and price tier are only available via Foursquare's venues details API, which is a Premium call, and I'm using a Personal Tier account which has a 500 Premium calls limit per day, so I couldn't get such information for all 4489 restaurants here and had to give them up.

But 'Likes' information is available via Foursquare's regular API 'likes' (GET https://api.foursquare.com/v2/venues/VENUE_ID/likes), by passing in the venue id we got from the step above (it's available in the returned JSON), I did get how many liked each restaurant received and appended it to the table above.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Restaurant	Main Category	Likes
0	Parkwoods	43.753258	-79.329659	Allwyn's Bakery	Caribbean Restaurant	18
1	Parkwoods	43.753258	-79.329659	Darband Restaurant	Middle Eastern Restaurant	45
2	Parkwoods	43.753258	-79.329659	Me Va Me Kitchen Express	Mediterranean Restaurant	18
3	Parkwoods	43.753258	-79.329659	Tim Hortons	Café	7
4	Parkwoods	43.753258	-79.329659	The Captain's Boil	Seafood Restaurant	11

Table 3. Restaurant – Name – Category - Likes Table

3.4 Category Analysis and Target Choice

In the above restaurants data, we have 107 unique values for column 'Main Category', which means we have 107 different types of restaurants. We cannot analyze all of them so we will only pick several popular types here. In order to do this, we need to rank the different

categories by the number of restaurants in each category. The following table shows the top 10 categories along with their percentage.

Café	0.075963
Pizza Place	0.072622
Restaurant	0.061929
Italian Restaurant	0.059701
Sandwich Place	0.052127
Bakery	0.050791
Japanese Restaurant	0.040766
Fast Food Restaurant	0.032969
Sushi Restaurant	0.032078
Chinese Restaurant	0.030742

Table 4. Top 10 Restaurant Categories in Toronto

If we visualize the above table into a bar chart, we got this:

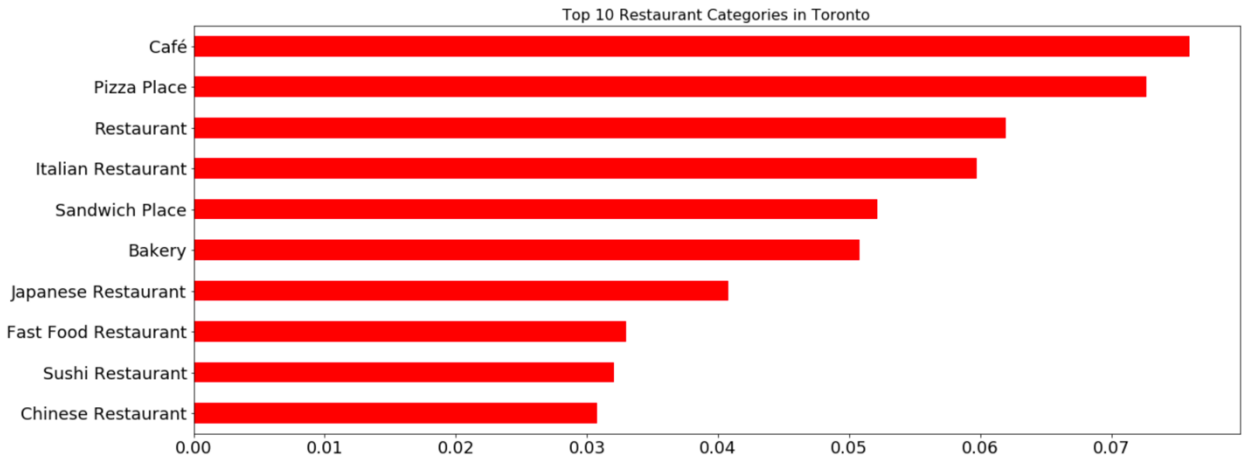

Figure 2. Top 10 Restaurant Categories in Toronto

Among the top 10 categories, we want to choose one type of restaurants to continue analysis. Café, Pizza Place, Sandwich Place are too general, 'Restaurant' is obviously not a correct category, as we want to choose a category that has enough data to analyze, Italian Restaurant seems to be the best candidate.

3.5 Analyze Italian Restaurants

3.5.1 Clean the Restaurants Data

Now we have chosen the type of restaurant we want to analyze, let's first remove all other categories of restaurants from the dataframe and only keep the Italian ones. After removal, only 268 rows are left, and this is the total number of Italian restaurants we have in our data.

Then let's group them by neighborhood, count how many Italian restaurants in each neighborhood and how many likes in total for all Italian restaurants in the neighborhood. The table below shows the first 5 rows of the calculated result.

	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
Parkwoods	1	6	43.753258	-79.329659
Victoria Village	1	3	43.725883	-79.315575
Regent Park, Harbourfront	3	412	43.654259	-79.360634
Lawrence Manor, Lawrence Heights	4	90	43.718517	-79.464760
Queen's Park, Ontario Provincial Government	4	440	43.662300	-79.389496

Table 5. Number of Italian Restaurants and Total Likes Received in Each Neighborhood

3.5.2 Normalize the Data

The data above is clean, but for 'Count' and 'Likes' column, their values will have unbalanced impact if we later directly use this in K-Means, as 'Likes' has much larger value than 'Count'. Hence, we need to normalize the data before further analysis. We will divide each row of 'Count' and 'Likes' by their total, and this will lead us to the below table.

	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
Parkwoods	0.003731	0.000463	43.753258	-79.329659
Victoria Village	0.003731	0.000232	43.725883	-79.315575
Regent Park, Harbourfront	0.011194	0.031805	43.654259	-79.360634
Lawrence Manor, Lawrence Heights	0.014925	0.006948	43.718517	-79.464760
Queen's Park, Ontario Provincial Government	0.014925	0.033966	43.662300	-79.389496

Table 6. Normalized Version of Table 5

3.5.3 Visualize the Data

If we put the above data into a bar chart, we will get this:

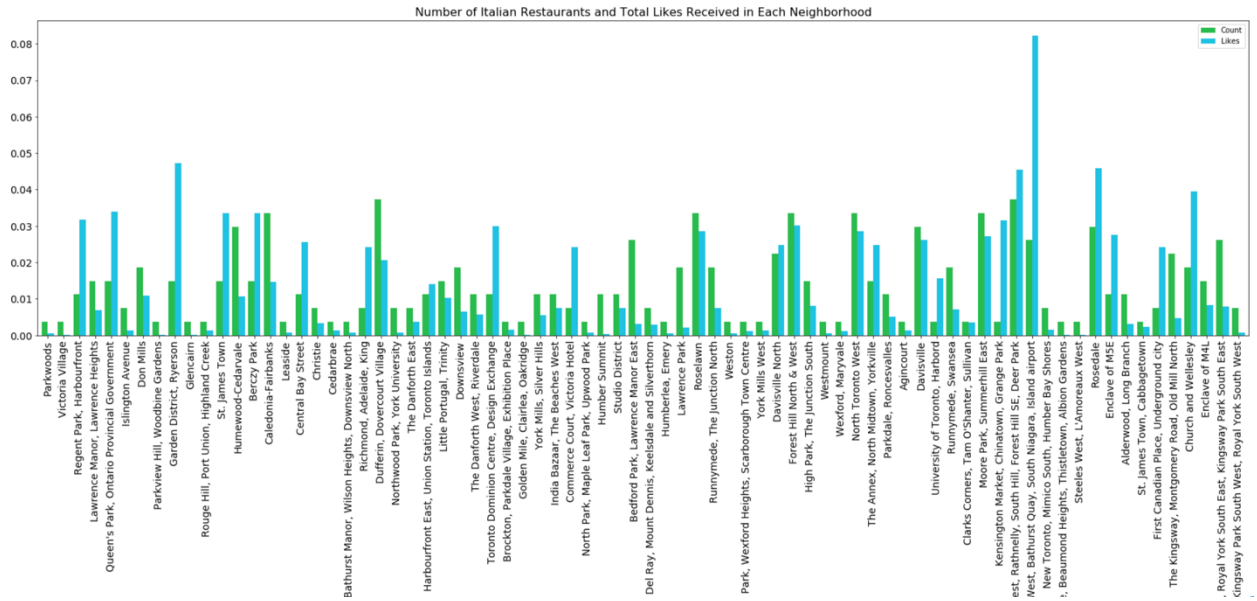

Figure 3. Number of Italian Restaurants and Total Likes Received in Each Neighborhood

3.5.4 Use K-Means to Group the Neighborhoods

Now, we can use 'Count' and 'Likes' to group neighborhoods into different groups, so that we can decide which neighborhoods are the best neighborhoods to open Italian restaurants. This can be done by utilizing the unsupervised Machine Learning method K-Means, which will gather similar candidates into one group. After applying K-Means and adding the result cluster labels into the original table, we got the following table (first 10 rows):

	Cluster Label	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
Parkwoods	0	0.003731	0.000463	43.753258	-79.329659
Victoria Village	0	0.003731	0.000232	43.725883	-79.315575
Regent Park, Harbourfront	1	0.011194	0.031805	43.654259	-79.360634
Lawrence Manor, Lawrence Heights	4	0.014925	0.006948	43.718517	-79.464760
Queen's Park, Ontario Provincial Government	1	0.014925	0.033966	43.662300	-79.389496
Islington Avenue	0	0.007463	0.001312	43.667854	-79.532242
Don Mills	4	0.018657	0.010807	43.745907	-79.352188
Parkview Hill, Woodbine Gardens	0	0.003731	0.000232	43.706398	-79.309937
Garden District, Ryerson	1	0.014925	0.047167	43.657162	-79.378937
Glencairn	0	0.003731	0.000232	43.709576	-79.445076

Table 7. Cluster Label of Each Neighborhood

3.5.5 Visualize the Resulting Clusters

Let's visualize the above cluster labels into a map (note that only 76 neighborhoods have Italian restaurants, so the neighborhoods showed below are less than the initial map we showed):

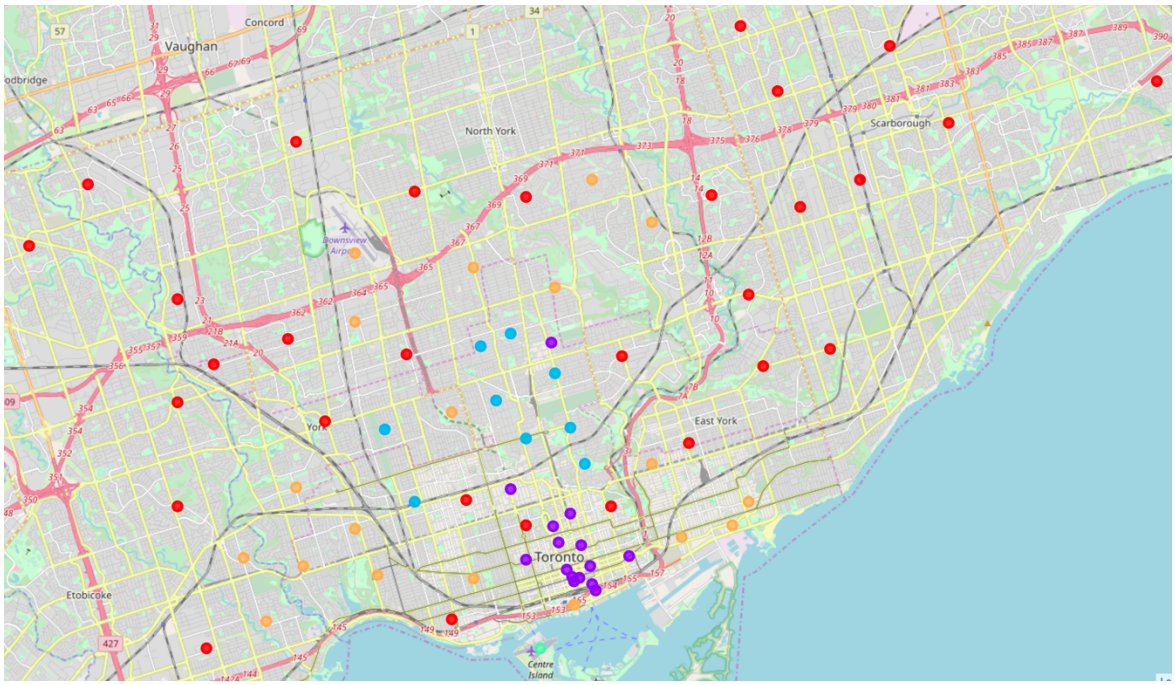

Figure 4. The Map of Resulting Clusters

3.5.6 Determine Which Cluster Represents the Best Neighborhoods

Although we grouped all neighborhoods that have Italian restaurants into 5 groups and also visualized them, we still don't know which cluster represents the best neighborhoods to open a new Italian restaurant. We need to look into the details of each group to decide this.

3.5.6.1 Cluster 0

	Neighborhood	Cluster Label	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
0	Parkwoods	0	0.003731	0.000463	43.753258	-79.329659
1	Victoria Village	0	0.003731	0.000232	43.725883	-79.315575
5	Islington Avenue	0	0.007463	0.001312	43.667854	-79.532242
7	Parkview Hill, Woodbine Gardens	0	0.003731	0.000232	43.706398	-79.309937
9	Glencairn	0	0.003731	0.000232	43.709576	-79.445076

Table 8. Cluster 0 Details

3.5.6.2 Cluster 1

	Neighborhood	Cluster Label	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
2	Regent Park, Harbourfront	1	0.011194	0.031805	43.654259	-79.360634
4	Queen's Park, Ontario Provincial Government	1	0.014925	0.033966	43.662300	-79.389496
8	Garden District, Ryerson	1	0.014925	0.047167	43.657162	-79.378937
11	St. James Town	1	0.014925	0.033503	43.651493	-79.375420
13	Berczy Park	1	0.014925	0.033503	43.644772	-79.373306

Table 9. Cluster 1 Details

3.5.6.3 Cluster 2

	Neighborhood	Cluster Label	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
14	Caledonia-Fairbanks	2	0.033582	0.014590	43.689026	-79.453514
21	Dufferin, Dovercourt Village	2	0.037313	0.020611	43.669006	-79.442261
41	Roselawn	2	0.033582	0.028485	43.711697	-79.416939
47	Forest Hill North & West	2	0.033582	0.030184	43.696949	-79.411308
51	North Toronto West	2	0.033582	0.028485	43.715382	-79.405678

Table 10. Cluster 2 Details

3.5.6.4 Cluster 3

	Neighborhood	Cluster Label	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
62	CN Tower, King and Spadina, Railway Lands, Har...	3	0.026119	0.082291	43.628948	-79.394417

Table 11. Cluster 3 Details

3.5.6.5 Cluster 4

	Neighborhood	Cluster Label	Count	Likes	Neighborhood Latitude	Neighborhood Longitude
3	Lawrence Manor, Lawrence Heights	4	0.014925	0.006948	43.718517	-79.464760
6	Don Mills	4	0.018657	0.010807	43.745907	-79.352188
12	Humewood-Cedarvale	4	0.029851	0.010730	43.693783	-79.428192
24	Harbourfront East, Union Station, Toronto Islands	4	0.011194	0.014127	43.640816	-79.381752
25	Little Portugal, Trinity	4	0.014925	0.010344	43.647926	-79.419746

Table 12. Cluster 4 Details

4 Results

By looking into the details of each cluster, we can now decide the order from the best to the worst:

- 1st Place: Cluster 3
- 2nd Place (Tie): Cluster 1 and Cluster 2
- 3rd Place: Cluster 4
- 4th Place: Cluster 0

Although Cluster 3 only has one neighborhood, it has the 2nd high volume of Italian restaurants and ranked 1st on 'Likes' (and it's much higher than the 2nd). Hence, I think this is the winner of our contest.

Cluster 2 has the highest volume of Italian restaurants, but it doesn't receive as many 'Likes' as Cluster 1, so I would say they tied in this contest and both received the 2nd place.

Obviously, Cluster 4 and Cluster 0 didn't perform very well, no matter in terms of volume or in terms of 'Likes', so they ranked 3rd and 4th respectively.

5 Discussion

Let's see the map of resulting clusters again:

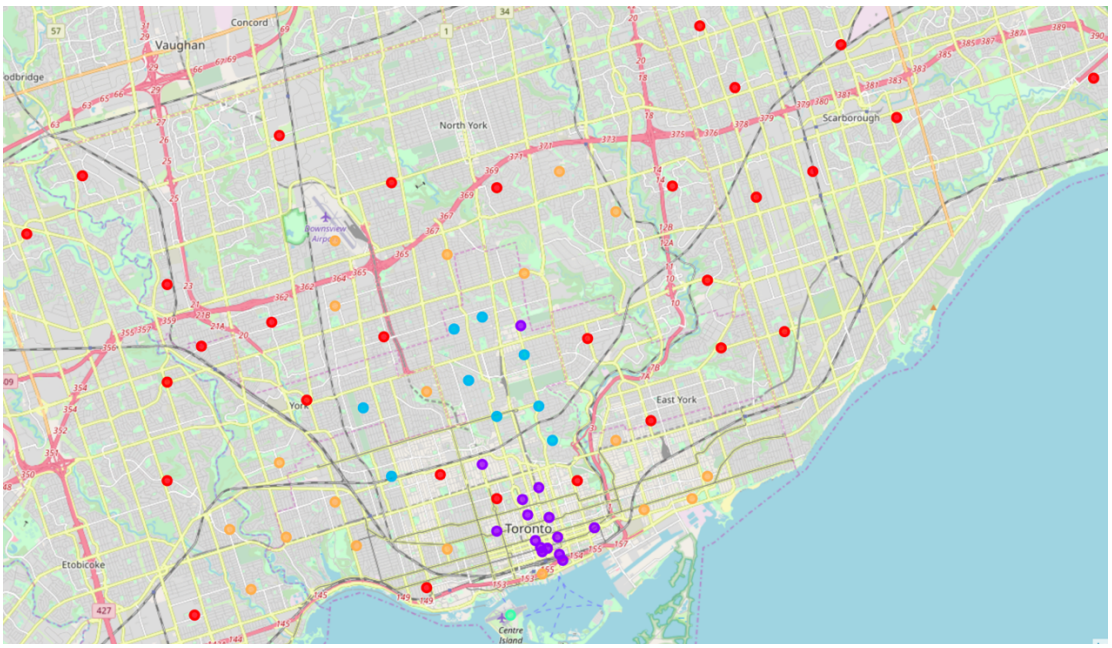

Figure 5. The Map of Resulting Clusters

In the above map, Green represents Cluster 3, Purple represents Cluster 1, Blue represents Cluster 2, Yellow represents Cluster 4 and Red represents Cluster 0.

5.1 Downtown Toronto

Basically, we can see the trend is that the closer to Downtown Toronto (the core of Toronto city), the higher the score will be. Although for each neighborhood in Downtown Toronto, its size is relatively small, there are still too many Italian restaurants opening there and they usually have very good reviews.

Hence, Cluster 3 won this contest, and the only neighborhood it represents is the CN Tower neighborhood. CN Tower is located at the center of Downtown Toronto, and it's a must-see for all the tourists visiting Toronto. It's not surprised at all that this neighborhood got the 1st place.

5.2 Toronto Subway Line 1

TTC Line 1 is the longest subway line in Toronto, it has 38 stations, and is a "U-shaped" route running generally in a south and then north direction. If we look carefully into the above map, we can notice that the blue dots (Cluster 2) represent the neighborhoods around the Line 1 subway stations. The blue dots on the left are along Dufferin Street and the blue dots on the right are along Yonge Street, they are the exact 2 streets the "U-shaped" route is passing through.

Downtown Toronto is located at the bottom of the "U-shaped" route, so it's basically the center of the Line 1, so this opinion also matches the opinion in section 5.1 above.

5.3 Possible Future Improvements

As mentioned above, due to the limitation on the number of API calls I can make each day to Foursquare, data like total visits and price tier are not included in this analysis, but they could.

According to the discussions above in section 5.1 and 5.2, total visits may not have too much impact on the final result here, as restaurants in Downtown Toronto or along the subway lines usually get a lot of visits.

However, I think price tier data may change the winner, as restaurants in Downtown Toronto or along the subway lines are usually high-end and very expensive. If we include the capital expense of opening a restaurant, we may not get such an extreme result above.

6 Conclusion

The conclusion of the analysis result above fits very well to the real life. Downtown Toronto, the center of the Toronto city, won the contest. The next option will be along the Subway Line 1. The further we go away from these 2 places, the less competitive an Italian restaurant will be.