Best Neighborhood for Italian restaurant in Toronto

Introduction and Business Problem

The main objective of the project is to find the best location for a new, Italian restaurant in Toronto. This information can be useful for entrepreneurs based in Canada and I am designing this study for them.

The location resolution will be by neighborhood and the parameters I am going to observe in order to recommend the best option(s) are:

- number of restaurants nearby
- number of Italian restaurants nearby

I will be using an unsupervised Machine Learning technique called K-means in order to create groups of Neighborhoods based on the feature parameters (number of restaurants).

I will also visualize the data just to get a whole picture about this topic. This will give us a brief overview, where are the places which are not overcrowded with similar places like ours and where a profitable business can be established.

Of course, there are other factors such as population, net income, number of tourists - just to name a few -, these parameters can be the subject of a potential model refining in the future.

Data

Description

We need the following information:

- List of neighborhoods in City of Toronto, Ontario, Canada
- Geographical coordinates of the neighborhoods (latitude, longitude)
- Venue data related to:
  - Restaurants and
  - Italian restaurants

Preparation

- Scrapping Toronto neighborhood data from Wikipedia

	Postal code	Borough	Neighborhood
0	None	None	None
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Regent Park / Harbourfront
6	M6A	North York	Lawrence Manor / Lawrence Heights
7	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government
9	M9A	Etobicoke	Islington Avenue
10	M1B	Scarborough	Malvern / Rouge
12	M3B	North York	Don Mills
13	M4B	East York	Parkview Hill / Woodbine Gardens
14	M5B	Downtown Toronto	Garden District, Ryerson
15	M6B	North York	Glencairn
18	M9B	Etobicoke	West Deane Park / Princess Gardens / Martin Gr…
19	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek
21	M3C	North York	Don Mills
22	M4C	East York	Woodbine Heights
23	M5C	Downtown Toronto	St. James Town
24	M6C	York	Humewood-Cedarvale

- Getting the coordinates of the neighborhoods by using Geocoder package

	Postal Code	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
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848
10	M1P	43.757410	-79.273304
11	M1R	43.750072	-79.295849

- Connect to Foursquare API to get venue data related to the identified neighborhoods and create a new dataframe with venues and neighborhoods:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot

Based on the above steps we were able to create our new, master dataframe which will be the starting point of our data analysis.

|	Neighborhood	Postal code	Borough	Latitude	Longitude	Number of restaurants	Number of Italian restaurants
0	Regent Park, Harbourfront	M5A	Downtown Toronto	43.654260	-79.360636	5.0	0.0
1	Queen's Park, Ontario Provincial Government	M7A	Downtown Toronto	43.662301	-79.389494	5.0	1.0
2	Garden District, Ryerson	M5B	Downtown Toronto	43.657162	-79.378937	24.0	2.0
3	St. James Town	M5C	Downtown Toronto	43.651494	-79.375418	23.0	3.0
4	The Beaches	M4E	East Toronto	43.676357	-79.293031	0.0	0.0

Methodology and Analysis

EDA - Exploratory Data Analysis

It is useful to visualize Toronto Neighborhoods on a map - this can give us a brief picture about the structure of the city and districts.

We also visualize the numbers of restaurants and Italian restaurants by Neighborhoods:
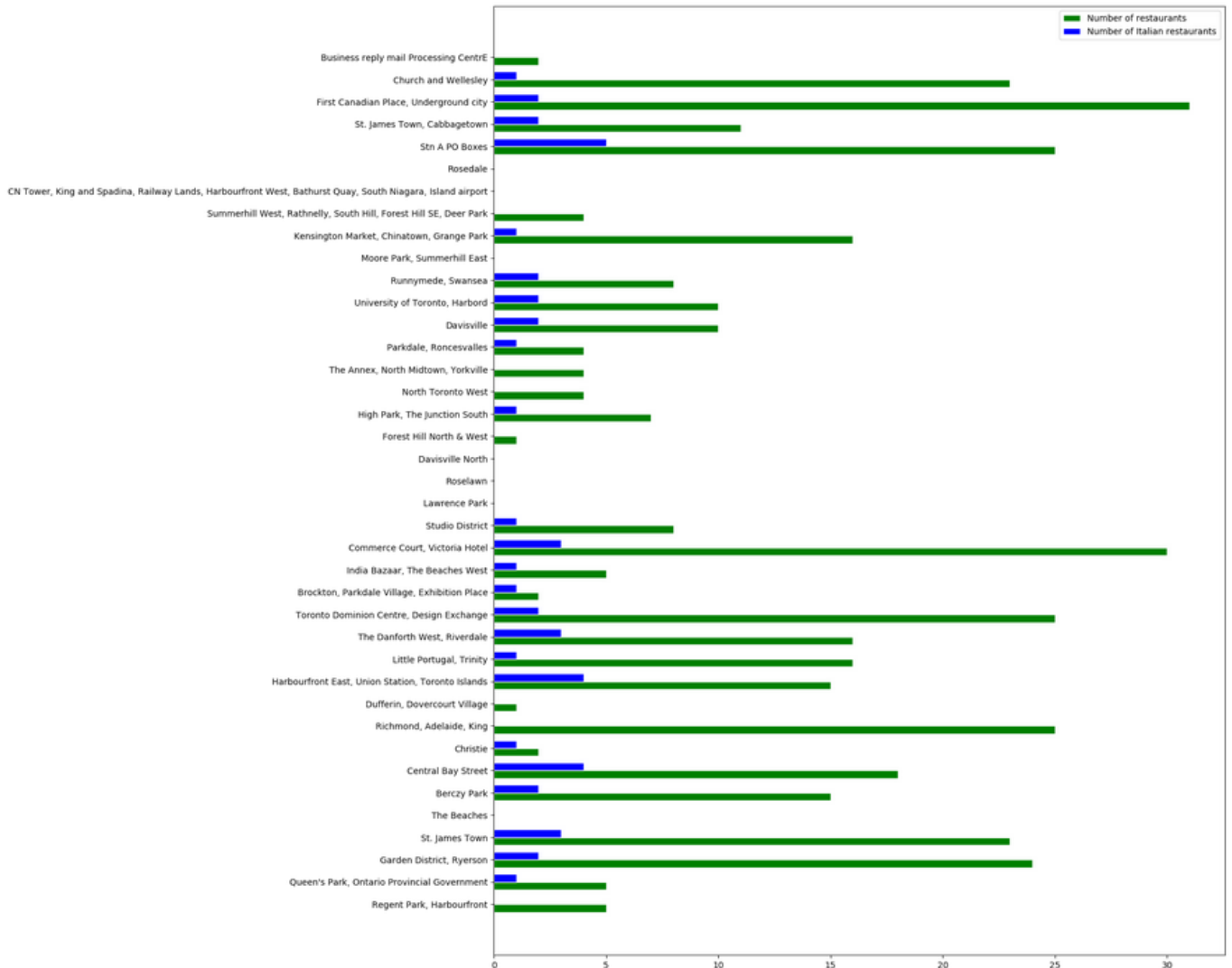

Clustering and Analysis

In order to define groups with similar characteristics, we used K-Means algorithm, which is an unsupervised machine learning method for clustering.

Therefore we reduced our master dataset to just numbers of the parameters (restaurants and Italian restaurants) and created a new dataframe for the algorithm.
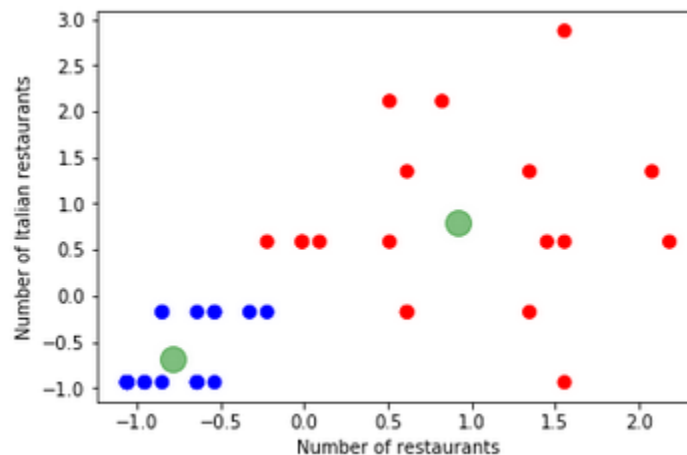
	Number of restaurants	Number of Italian restaurants
0	5.0	0.0
1	5.0	1.0
2	24.0	2.0
3	23.0	3.0
4	0.0	0.0

Afterwards we ran a scaler (Standard Scaler) on our dataset.

Then Elbow method was applied to define optimal k value for clustering, where k means the number of clusters:

The Elbow Method showing the optimal k
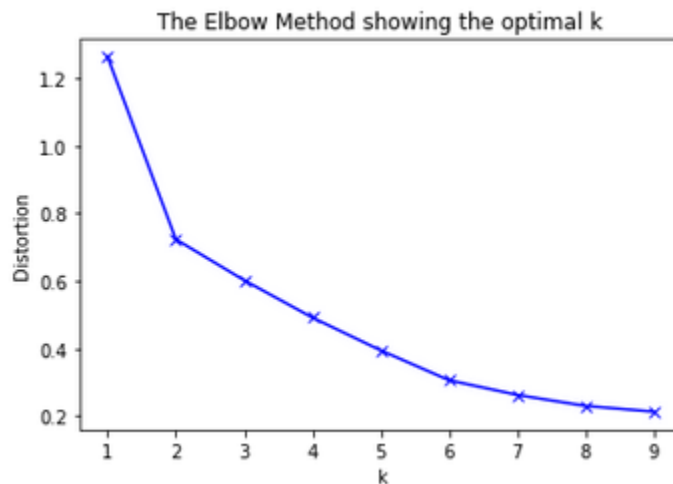
Based on the above plot we identified 2 as the best optional k value (where the plot line shift is the highest). Then we plotted the clusters and put them into dataframe:

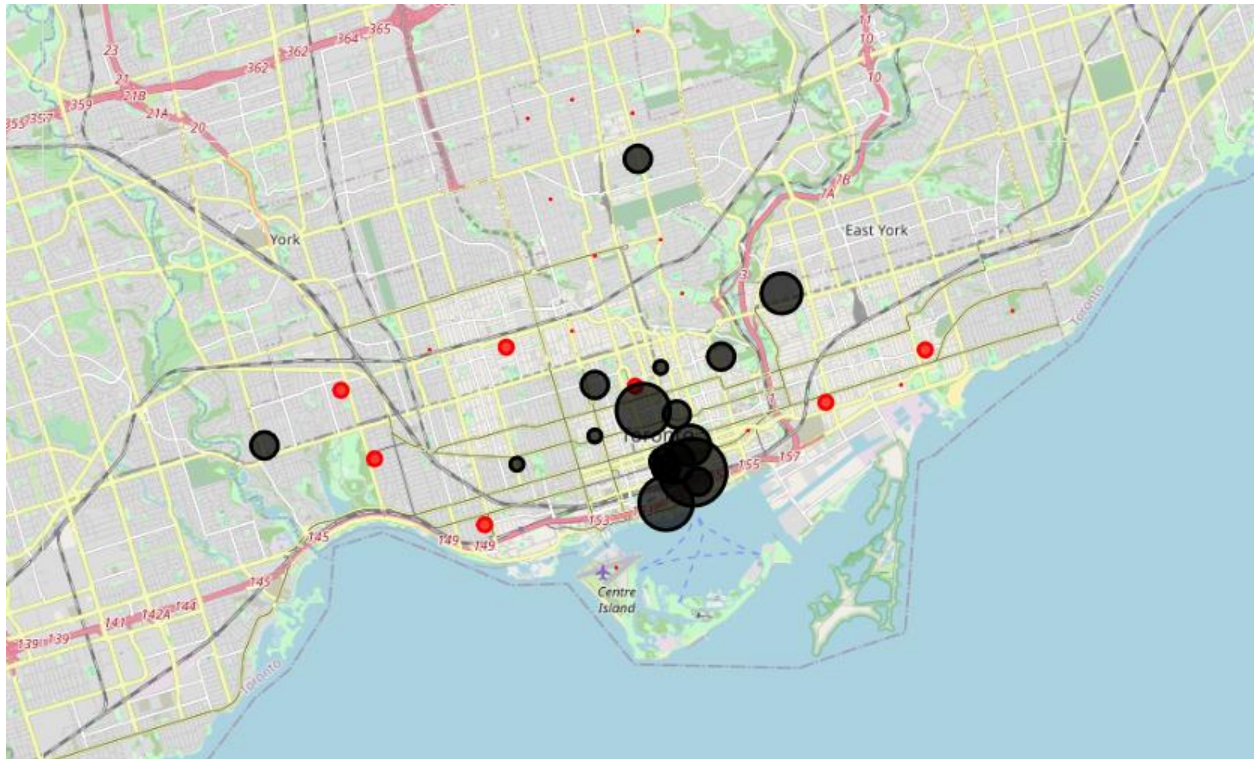|	Cluster Labels	Neighborhood	Postal code	Borough	Latitude	Longitude	Number of restaurants	Number of Italian restaurants
0	0	Regent Park, Harbourfront	M5A	Downtown Toronto	43.654260	-79.360636	5.0	0.0
1	0	Queen's Park, Ontario Provincial Government	M7A	Downtown Toronto	43.662301	-79.389494	5.0	1.0
2	1	Garden District, Ryerson	M5B	Downtown Toronto	43.657162	-79.378937	24.0	2.0
3	1	St. James Town	M5C	Downtown Toronto	43.651494	-79.375418	23.0	3.0
4	0	The Beaches	M4E	East Toronto	43.676357	-79.293031	0.0	0.0

Results

As a result, we got two clusters based on the number of restaurants and Italian restaurants. In Cluster 0 there are less restaurants by neighborhood and in Cluster 1 there are more restaurants.

However, in Cluster 1 there are neighborhoods where the number of Italian restaurants is very low, could be even zero. This will be important later when discussing the results.

Restaurants per neighborhoods on a map (the more values a neighborhood has the bigger the circle is):

Italian restaurants per neighborhoods on a map (the more values a neighborhood has the bigger the circle is):

Neighborhoods with zero Italian restaurants (almost all of them are Cluster 0, with generally less restaurants):

	Cluster Labels	Neighborhood	Postal code	Borough	Latitude	Longitude	Number of restaurants	Number of Italian restaurants
0	0	Regent Park, Harbourfront	M5A	Downtown Toronto	43.654260	-79.360636	5.0	0.0
4	0	The Beaches	M4E	East Toronto	43.676357	-79.293031	0.0	0.0
8	1	Richmond, Adelaide, King	M5H	Downtown Toronto	43.650571	-79.384568	25.0	0.0
9	0	Dufferin, Dovercourt Village	M6H	West Toronto	43.669005	-79.442259	1.0	0.0
18	0	Lawrence Park	M4N	Central Toronto	43.728020	-79.388790	0.0	0.0
19	0	Roselawn	M5N	Central Toronto	43.711695	-79.416936	0.0	0.0
20	0	Davisville North	M4P	Central Toronto	43.712751	-79.390197	0.0	0.0
21	0	Forest Hill North & West	M5P	Central Toronto	43.696948	-79.411307	1.0	0.0
23	0	North Toronto West	M4R	Central Toronto	43.715383	-79.405678	4.0	0.0
24	0	The Annex, North Midtown, Yorkville	M5R	Central Toronto	43.672710	-79.405678	4.0	0.0
29	0	Moore Park, Summerhill East	M4T	Central Toronto	43.689574	-79.383160	0.0	0.0
31	0	Summerhill West, Rathnelly, South Hill, Forest…	M4V	Central Toronto	43.686412	-79.400049	4.0	0.0
32	0	CN Tower, King and Spadina, Railway Lands, Har…	M5V	Downtown Toronto	43.628947	-79.394420	0.0	0.0
33	0	Rosedale	M4W	Downtown Toronto	43.679563	-79.377529	0.0	0.0
38	0	Business reply mail Processing CentrE	M7Y	East Toronto	43.662744	-79.321558	2.0	0.0

Neighborhood with many restaurants but zero Italian restaurant:

	Cluster Labels	Neighborhood	Postal code	Borough	Latitude	Longitude	Number of restaurants	Number of Italian restaurants
8	1	Richmond, Adelaide, King	M5H	Downtown Toronto	43.650571	-79.384568	25.0	0.0

Discussion and Conclusion

During the analysis two clusters could be found.

The first cluster is where there are less restaurants and Italian restaurants and the second cluster is the group with more venues.
First of all, we can identify potential places from **Cluster 0 where the number of Italian restaurants is zero**. These neighborhoods could be a good option, but more information needed for further studies.
Secondly, **from Cluster 1 we can filter the districts where there aren't any Italian restaurants even if there are a high number of restaurants. This could mean that these neighborhoods have a huge potential** (probably with large tourism and higher income per capita), there are many businesses running, but they are different than our plan - so we can also recommend these options for further studies.

To conclude, with the data analysis we identified different clusters in order to help the decision-making process for the placement of a new Italian restaurant in Toronto.

We identified and visualized two strongly different clusters with K-means, based on other restaurants by neighborhood. This can give us a basic picture and we should take-away the following information:
We need to look at **Cluster 0 with less restaurants** for potential placement but **we shouldn't forget about more crowded neighborhoods** - from there we can filter for the places where we found zero Italian restaurants - these two options can identify promising neighborhoods for our business.

Please consider that this study gives us a good starting point, but further examination is needed, adding population, net income, tourism data to our model.