Best Neighborhoods to open an Italian restaurant in Toronto

David Mateo Cangrejo Angel

August 10, 2020

1. Introduction/Business Problem

1.1. **Description of the problem**

The problem consist in identifying the optimum place to open an Italian restaurant in the city of Toronto, Canada.

The target audience of this project are entrepreneurs of Italian food or owners of Italian restaurants.

More specifically, stakeholders who are interested in identifying the best Neighborhood candidates in Toronto to open a new Italian restaurant, so with the best group of candidates decide finally where to open it.

1.2. **Discussion of the background**

One important aspect for a restaurant success is the place where it is located. Of course, the food and service are very important, however the location can be as crucial especially in the early years. Therefore it is necessary to make a good analysis of the location, in a first step of the neighborhood, of the restaurant before starting the operation. That's why the stakeholders would care about this project.

Some key aspects to take into account for this analysis are:

1. Parking.
2. Visibility.
3. Number of people who can pass near the restaurant.
4. Income of the Neighborhood.
5. Presence of near similar restaurants.
6. Presence of other business that can attract people to the restaurant: Stadiums, parks, theaters, medical majors.

2. Data

2.1. Data Description

For this analysis the following data is going to be used:

1. List of Neighborhoods in Toronto: from The City of Toronto's Open Data Portal.
2. Latitude and longitude coordinates of neighborhoods: from The City of Toronto's Open Data Portal.
3. Venues near every neighborhood: from Foursquare Api.
4. Population and income of each neighborhood: from The City of Toronto's Open Data Portal

The City of Toronto's Open Data Portal is an open source delivery tool to bring people and data together.

2.2. Data Usage

The data is going to be used in the following way to solve the problem:

- The list of Neighborhoods and its coordinate's data are going to be merged to identify the location of each neighborhood.
- The population and income data of each neighborhood is going to be also merged with the previous data.
- The venues data of every neighborhood is going to be classified in the following way: Parking and presence of other business that can attract people to the restaurant (like stadiums, theaters, medical majors) are going to be count as a "Collaborator Index" and similar restaurants (Italian food) are going to be count as a "Competitor index".
- Finally, K-mean machine learning will be used to cluster the neighborhoods with this data: Population, Income, Collaborator index and Competitor index.

We propose that is a benefit to the new restaurant to have higher values in population, income, Collaborator index and lower in Competitor index.

2.3. Data Cleaning

Based on the previous information we clean the data from de datasets to get the information we required.

The coordinates of each neighborhood:

	Neighbourhood Number	LONGITUDE	LATITUDE
0	94	-79.425515	43.676919
1	100	-79.403590	43.704689
2	97	-79.397871	43.687859
3	27	-79.488883	43.765736
4	31	-79.457108	43.714672

Figure 1 – Coordinates data.

And the demographic data:

Characteristic	index	Neighbourhood Number	Population, 2016	Total - Income statistics in 2015 for the population aged 15 years and over in private households
0	Agincourt North	129	29,113	25,005
1	Agincourt South-Malvern West	128	23,757	20,400
2	Alderwood	20	12,054	10,265
3	Annex	95	30,526	26,295
4	Banbury-Don Mills	42	27,695	23,410

Figure 2 – Demographic data.

After combining both datasets we finally get:

	Neighbourhood Number	Longitude	Latitude	Neighborhood	Population, 2016	Total - Income statistics in 2015 for the population aged 15 years and over in private households
0	94	-79.425515	43.676919	Wychwood	14,349	11,345
1	100	-79.403590	43.704689	Yonge-Eglinton	11,817	9,995
2	97	-79.397871	43.687859	Yonge-St.Clair	12,528	11,170
3	27	-79.488883	43.765736	York University Heights	27,593	23,530
4	31	-79.457108	43.714672	Yorkdale-Glen Park	14,804	12,065

Figure 3 – Coordinates and Demographic data combined.

Plotting the neighborhoods in the Toronto's map we get:

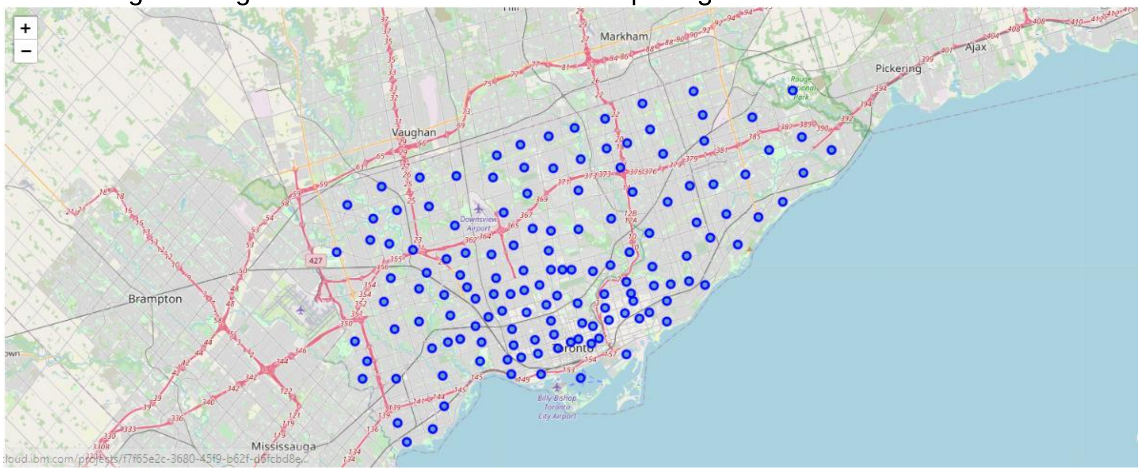

Figure 4 – Neighborhood's location in Toronto.

Also, the venues data from Foursquare API looks like:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wychwood	43.676919	-79.425515	Wychwood Barns Farmers' Market	43.680010	-79.423849	Farmers Market
1	Wychwood	43.676919	-79.425515	Wychwood Barns	43.680028	-79.423810	Event Space
2	Wychwood	43.676919	-79.425515	Hillcrest Park	43.676012	-79.424787	Park
3	Wychwood	43.676919	-79.425515	The Stop	43.679793	-79.423825	Convenience Store
4	Yonge-Eglinton	43.704689	-79.403590	North Toronto Memorial Community Centre	43.706098	-79.404337	Gym

Figure 5 – Toronto's venues data.

3. Methodology

3.1. Exploratory Data Analysis

We try to understand our venues data. First, we plot how many venues we collect from each neighborhood:

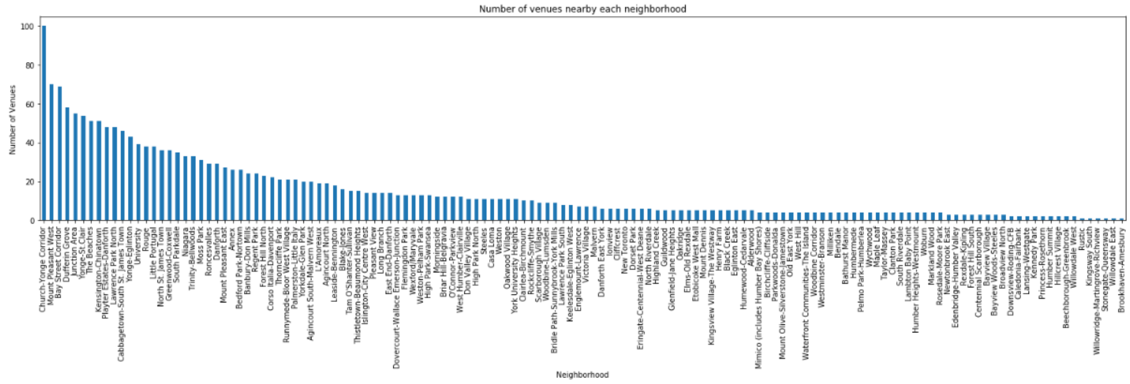

Figure 6 – Number of venues nearby each neighborhood.

The total number of unique neighborhoods are 138, and the total count of venues are 2028. Also, the total number of unique venues categories are 285.

We plot the top 50 venues categories with more number of venues in Toronto:

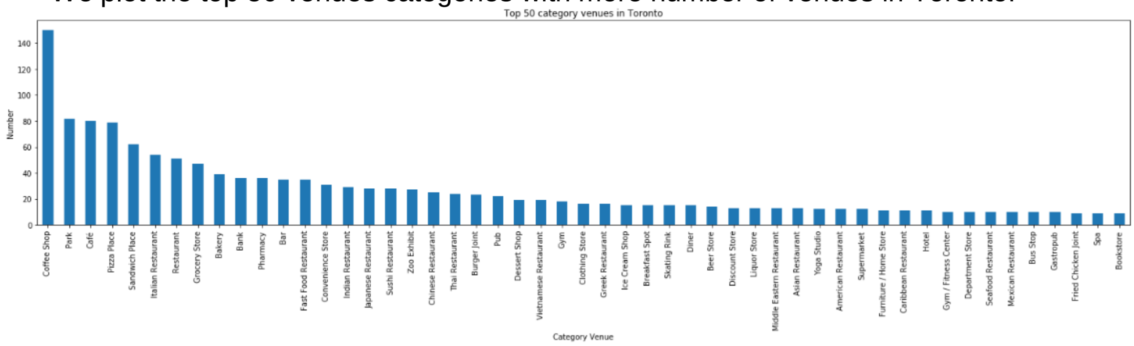

Figure 7 – Top 50 category venues in Toronto.

3.2. **Competitor and collaborator indices**

After understanding better the data, we proceed to identify which venues are restaurants and Italian Restaurants (our competitors).
So we obtain, 472 Restaurants and 54 Italian Restaurants. Then, we define our functions that qualify the competitors and the collaborators indices:

- Competitor's index: if it is a restaurant +1 point; and if is an Italian Restaurant +10 points. We try to penalize a lot higher if it is an Italian Restaurant. As far as this index increase, the sector is very competitive for an Italian Restaurant.
- Collaborator’s index: Every venues which is not a Restaurant we give +1 point. As far as this index increase, the sector have more venues to attract more people, which is good for the restaurant.

We apply the functions to the data and the group them by neighborhood. The final dataset we obtain is:

	Neighborhood	Competitors	Collaborators
0	Agincourt North	5	14
1	Agincourt South-Malvern West	15	5
2	Alderwood	0	5
3	Annex	4	22
4	Banbury-Don Mills	12	22

Figure 8 – Competitors and Collaborators indices data.

3.3. **Clustering Data**

We join the competitors and collaborators indices data with the demographic data. So finally we have the dataset for the clustering modeling:

Neighborhood	Competitors	Collaborators	Population	Income
Agincourt North	5	14	29113	25005
Agincourt South-Malvern West	15	5	23757	20400
Alderwood	0	5	12054	10265
Annex	4	22	30526	26295
Banbury-Don Mills	12	22	27695	23410

Figure 9 – Clustering Dataset.

3.4. **Clustering Algorithm**

We apply the clustering machine learning algorithm to cluster our neighborhoods of Toronto based in the features we have.

First we normalize the data, and then we obtain the optimal number of clusters based on the elbow method:

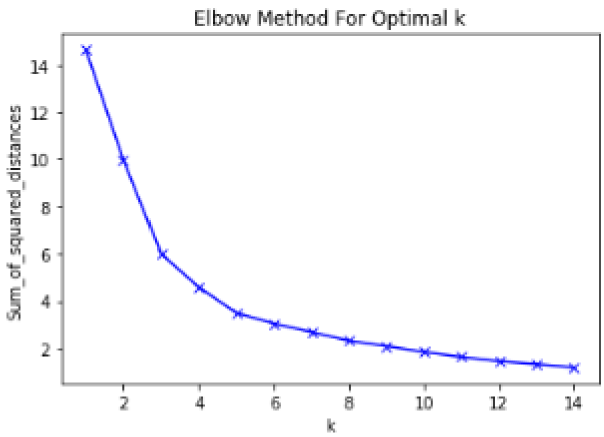

Figure 10 – Elbow method For Optimal k.

The Elbow method shows that k=5 is the optimal k for this dataset. So we proceed with this number of clusters with the k-means algorithm.

We run the k-mean clustering and obtain the cluster labels of each neighborhood:

Neighborhood	Cluster Labels	Competitors	Collaborators	Population	Income
Agincourt North	1	5	14	29113	25005
Agincourt South-Malvern West	1	15	5	23757	20400
Alderwood	0	0	5	12054	10265
Annex	1	4	22	30526	26295
Banbury-Don Mills	1	12	22	27695	23410

Figure 11 – Cluster Labels dataset.

Restore all the data with the original features and with the cluster labels:

	Neighborhood	Cluster Labels	Competitors	Collaborators	Population	Income	Latitude	Longitude
0	Agincourt North	1	5	14	29113	25005	43.805441	-79.266712
1	Agincourt South-Malvern West	1	15	5	23757	20400	43.788658	-79.265612
2	Alderwood	0	0	5	12054	10265	43.604937	-79.541611
3	Annex	1	4	22	30526	26295	43.671585	-79.404001
4	Banbury-Don Mills	1	12	22	27695	23410	43.737657	-79.349718

Figure 12 – Final dataset.

3.5. Cluster Map

We plot the cluster neighborhoods in the Toronto Map:

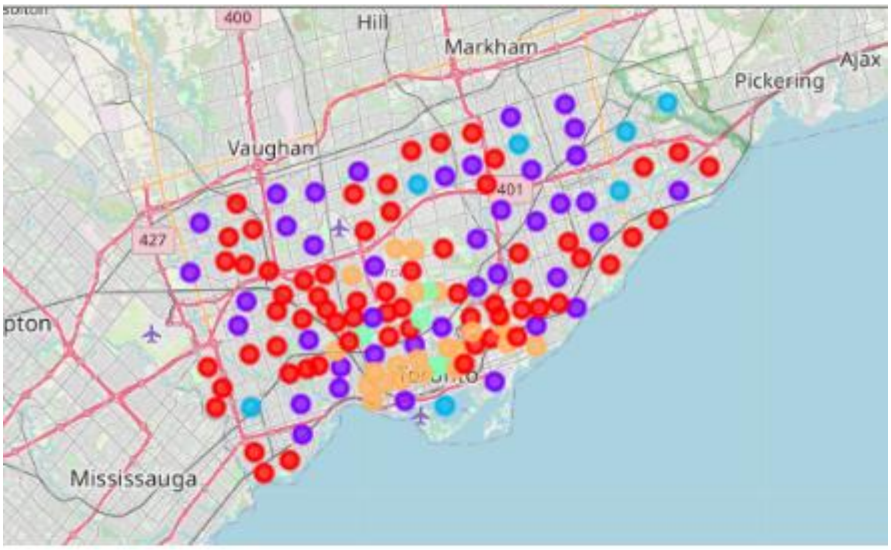

Figure 13 – Cluster map of Toronto's Neighborhoods. Cluster identification: 0-Red / 1-Purple / 2-blue / 3-Green / 4-Orange

4. Results

4.1. Cluster examination

Now, we can examine each cluster and determine the discriminating characteristics that distinguish each cluster. Let's examine the 5 clusters:

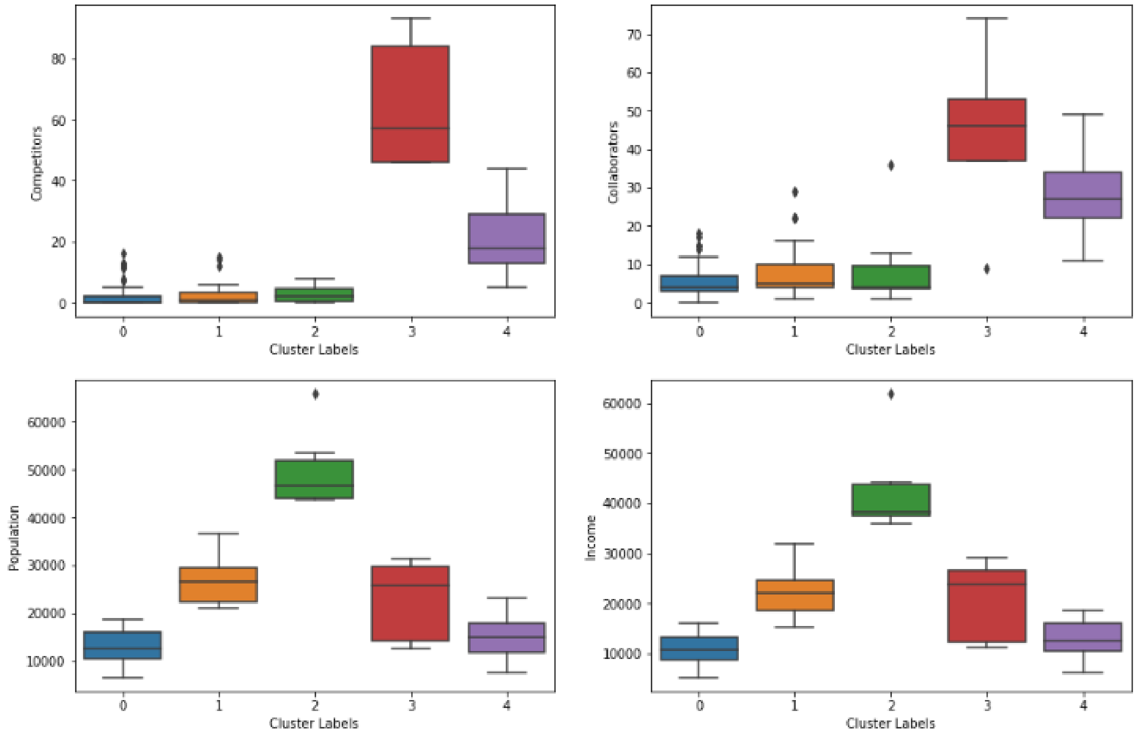

Figure 14 – Features description boxplot for each cluster.

5. Discussion

With the previous graph we can analyze the main characteristics of each cluster and classify them as follows:

1. Cluster 0: Low Competitors, low collaborators, low population and low income.
2. Cluster 1: Low Competitors, low collaborators, medium population and medium income.
3. Cluster 2: Low Competitors, low collaborators, high population and high income.
4. Cluster 3: High Competitors, high collaborators, medium population and medium income.
5. Cluster 4: Medium Competitors, medium collaborators, low population and low income.

We can drive some conclusions from these results. Definitely cluster 0 is not a good option for the restaurant. Cluster 1 has medium population and income, but as it has low collaborators it seems that those neighborhoods does not attract people to the business. So this is not a good option too.

So we have finally 3 options.

Cluster 2, has high population and income, this may be a good option, however it has low collaborators. It seems to be neighborhoods were a lot of people live but with few venues which attracts other people, so this is maybe a risky sector. If the stakeholder is beginning with its business is not a good option but if he is with confidence and experience it could be an excellent place.

Cluster 3, has high collaborators so it is good to attract people, but it also has high competitors. This may be good for a renowned Restaurant which many costumers know about it but may be not a good options for new restaurants.
Cluster 4, may be a good option because it has medium collaborators and also medium competitors (is not too high). So maybe this are the best neighborhoods to start with an Italian Restaurant.

6. Conclusion

We did the clustering of Toronto neighborhoods based on the business problem which consists of identifying the optimum place to open an Italian restaurant in the city.

We cluster successfully the neighborhoods in 5 groups with the data features: population, income, competitor's index and collaborator's index (from venues data).

Not only we obtain the best group of neighborhoods to start the Restaurant, but also, we understand the characteristics of the other clusters.

As a summary we obtain the following groups:

- Good neighborhoods for renowned restaurants.
- Good neighborhoods for new restaurants.
- Good neighborhoods for experienced stakeholders.
- Bad places for a restaurant (for 2 clusters).

The model may be improved in the following aspects:

- Optimizing the functions of Collaborator's and Competitor's index: like giving a higher score to a stadium and a lower to a coffee shop, because the first one attracts more people.
- Adding more features: like customers scores of the existing restaurants and visibility of the places.