Study of business viability in Toronto

Luiz. U. R. Sica

June 19, 2020

Abstract

The present study will analyze commercial patterns in each particular borough in Toronto, focusing on to scope out the competition.

1 Introduction

1.1 Background

Before choosing a location, entrepreneurs need to know their future businesses, understanding their needs, and identifying who your customers are and how you can best meet their needs. However, if your new business does not manufacture products, wholesale, or sell strictly over the Internet, then finding the right location to set up your business, from your customers' point of view, can be critical to your success.

1.2 Problem

To identify a location (borough) with a higher probability of success to start a new business in Toronto.

1.3 Interest

The present study will analyze commercial patterns in each particular borough in Toronto, focusing on to scope out the competition.

Case of study: Restaurant.

2 Methodology

The methodology steps adopted in this project are summarized below:

- Web scrape to obtain the postal codes, borough, and neighborhood names.
- Get the geographical coordinates of each postal code using the Pgeocode Python library.

- Join all the previous information obtained so far in a pandas dataframe.
- Use the Foursquare API to explore the 100 most common venues for each particular borough.
- Analyze the most common venues by borough considering the mean of the frequency as our metric and, assing the results to a dataframe.
- Apply k-means clustering to group boroughs with similar venues' data and discover underlying patterns.
- Interactively visualize spatial data that has been manipulated so far, using the Folium library.
- Discuss the insights observed in each cluster.

3 Data acquisition, cleaning and visualization

The data used in the present analysis was a combination of:

- Postal codes, borough, and neighborhood names obtained by web scrapping.
- Geographical coordinates of each postal code obtained using the Pgeocode Python library.
- Venues data of each borough obtained using the Foursquare API.

Each source of data is commented in the next sections.

3.1 Web scraping

To obtain the postal codes, borough, and neighborhood names of Toronto, I have web scraped the data from this webpage URL into a pandas dataframe.

The data had 77 rows with "Not assigned" objects. Therefore, these rows were dropped and the indexes reset.

3.2 Geographical coordinates

We needed the geographical coordinates of each postal code in Toronto. For the present project, we are interested in a reliable free package. Since the Google Maps Geocoding API started charging we have excluded this option. We have tested the Geocoder Python package with the following piece of code:

```
lat_lng_coords = None

while(lat_lng_coords is None):
```

```
    g = geocoder.google('{}, Toronto, Ontario'.format(postal_code))

    lat_lng_coords = g.latlng

latitude = lat_lng_coords[0]

longitude = lat_lng_coords[1]
```

However, this Package has a problem in which we have to be persistent (while loop), in a sense that we can make a call and the result would be None, and then make a call again and get the coordinates. Unfortunately, even with the while loop, the Geocoder Python package was very unreliable. Finally, a very good alternative used here was the Pgeocode Python library designed for high-performance off-line querying of GPS coordinates, region name and municipality name from postal codes. This package has the advantage of returning a pandas dataframe, as a result of a geo-location query.

After this step, we proceeded to combine (merge) the data frames resulted from web scraping and the Pgeocode Python library. In the resulting data frame inconsistencies were observed between columns "Neighborhood" and "$place_name$".$For simplicity, we chose the column "Neighborho$

3.3 Geospatial visualization

To visualize the geospatial data I used the highly intuitive folium library, which allows us to interactively (leaflet map) visualize spatial data that has been manipulated so far. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. The library has several built-in tilesets from OpenStreetMap, Mapbox, and Stamen, and supports custom tilesets with Mapbox or Cloudmade API keys. Furthermore, it supports both Image, Video, GeoJSON, and TopoJSON overlays (check this documentation link for further details).

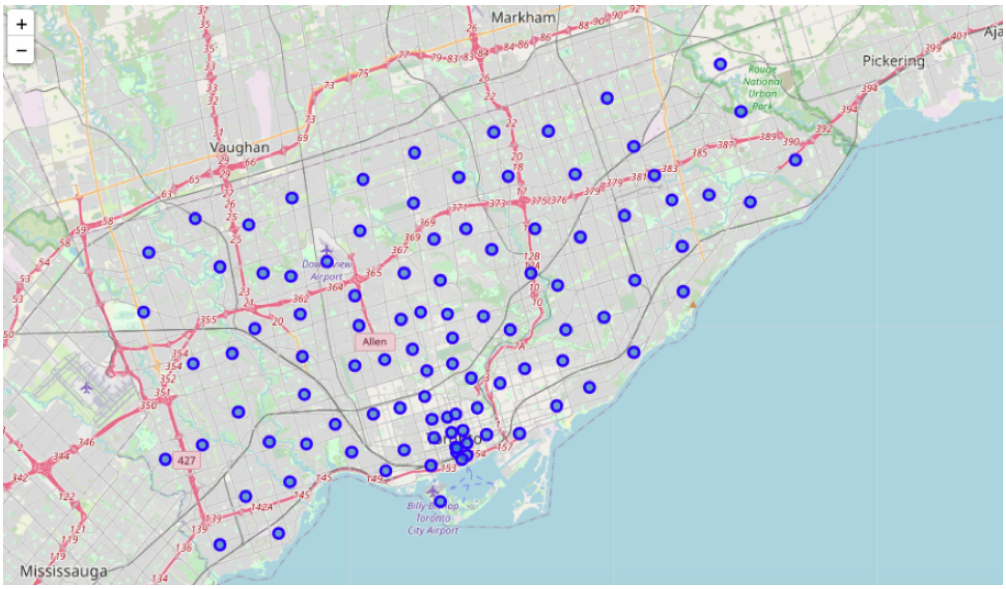


Figure 1: Visualizing boroughs in Toronto.

4 Data analysis

4.1 The most common venues by borough

We have used the Foursquare API to explore the venue's data for each particular borough. After that, I have analyzed the most common venues by borough considering the mean of the frequency as our metric and, assigned the results to a pandas dataframe.

	Postal Code	Borough	Neighborhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	M3A	North York	Parkwoods	43.7545	-79.3300	4	Clothing Store	Coffee Shop	Pizza Place	Park	Sandwich Place
1	M4A	North York	Victoria Village	43.7276	-79.3148	4	Clothing Store	Coffee Shop	Pizza Place	Park	Sandwich Place
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6555	-79.3626	0	Coffee Shop	Café	Restaurant	Hotel	Japanese Restaurant
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504	4	Clothing Store	Coffee Shop	Pizza Place	Park	Sandwich Place
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.6641	-79.3889	0	Coffee Shop	Café	Restaurant	Hotel	Japanese Restaurant
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.6662	-79.5282	1	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Bank
6	M1B	Scarborough	Malvern, Rouge	43.8113	-79.1930	1	Coffee Shop	Pizza Place	Pharmacy	Bank	Convenience Store
95	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.6684	-79.3689	0	Coffee Shop	Café	Restaurant	Hotel	Japanese Restaurant
96	M5X	Downtown Toronto	First Canadian Place, Underground city	43.6492	-79.3823	0	Coffee Shop	Café	Restaurant	Hotel	Japanese Restaurant
97	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.6518	-79.5076	1	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Bank
98	M4Y	Downtown Toronto	Church and Wellesley	43.6656	-79.3830	0	Coffee Shop	Café	Restaurant	Hotel	Japanese Restaurant
99	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.7804	-79.2505	2	Greek Restaurant	Restaurant	Italian Restaurant	Coffee Shop	Pub
100	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.6325	-79.4939	1	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Bank
101	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.6256	-79.5231	1	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Bank

Figure 2: The most common venues per borough.

4.2 Clustering

I am interested in grouping together boroughs with similar venues' data and discover underlying patterns. Therefore, we proceeded to cluster the boroughs in Toronto using an unsupervised machine learning algorithm. The one used was called k-means, where its purpose is to minimize the distance of data points from the centroid of its clusters and maximize the distance from other cluster's centroids. The k-means basic steps are summarized below:

- Set the number of clusters K.
- Randomly initialize the centroids coordinates of each cluster.
- Calculate the distance of each data point from the centroid points. Depending on the nature of the data, different metrics for the distance may be used to place items into clusters.
- Set the distance matrix, which is a matrix where each row represents the distance of a data point from each centroid.
- Use the distance matrix to find the nearest centroid to each data point, then assign each data point to the corresponding cluster.
- Compute the error, which is the sum of the squared distance between data points and all centroids.

- Update each centroid as the mean of all data points contained in its clusters.
- Repeat the process from step 3 until the centroids have stabilized, i.e., no significant change in their values is encountered.

For further details, check the documentation, and these articles analyticsvidhya, towards-datascience1 and, towardsdatascience2. It is worth noting that, as it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial clusters. In other words, this algorithm is guaranteed to converge to a result, but the result may be a local optimum, not necessarily the best possible outcome.

To overcome this issue, it is common to run the whole process multiple times with different starting conditions. The application of randomized starting centroids may give a better outcome. Besides, the algorithm is usually very fast, being not a problem running it multiple times.

Before we proceed with applying the k-means algorithm, we have to choose the number of clusters. For this purpose, we apply the "elbow" method, using the KElbowVisualizer from the Yellowbrick library, to select the optimal number of clusters by fitting the model with a range of values for k.

We will apply three different scoring parameter metric, with the main features summarized below:

- Distortion (default): It computes the sum of squared distances between each instance and its closest centroid. The desirable choice would be the lowest inertia. However, as pointed out by Jyoti Yadav this approach has a limitation, as the number of clusters increases, the closest will be the clusters from the centroids and lower will be the inertia.
- Silhouette: The silhouette score calculates the mean Silhouette Coefficient of all samples, defined as:

$$\mathrm{SC} = \frac{(p-q)}{\max(p,q)} \tag{1}$$

  where p is the mean distance to the points in the nearest cluster and q is the mean intra-cluster distance to all the points. The silhouette score raging from -1 to 1, with a brief description given below:

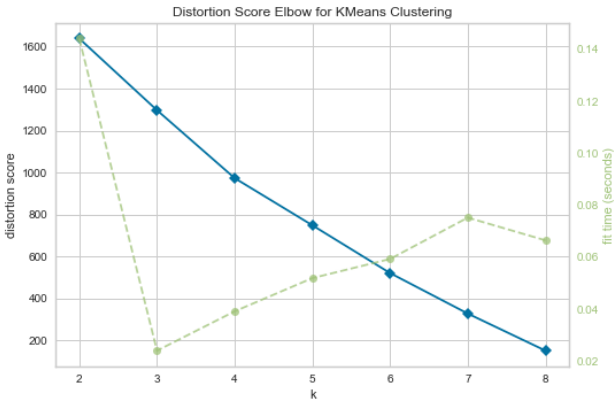
  - 1: It indicates that the sample is far away from its neighboring cluster.
  - 0: It indicates that the sample is on or very close to the decision boundary separating two neighboring clusters.
  - -1: It indicates that the samples have been assigned to the wrong clusters.

  See for further details this article.
- Calinski-harabasz: The Calinski Harabasz score or variance ratio is the ratio between within-cluster dispersion and between-cluster dispersion. Milligan et al. compared 30

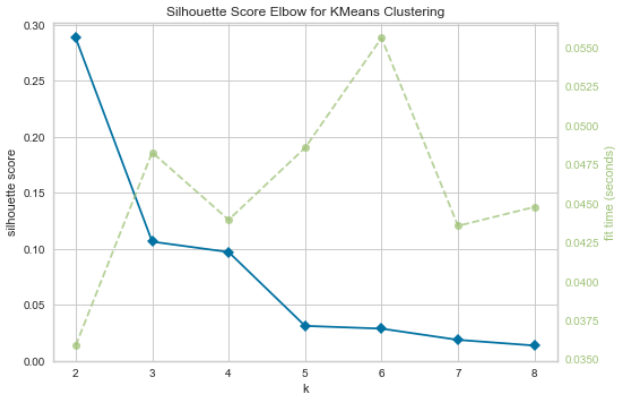
different approaches estimating the number of clusters in a dataset and found that the best performing method is given by Calinski and Harabasz, with its score defined as:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)(n-k)} \tag{2}$$

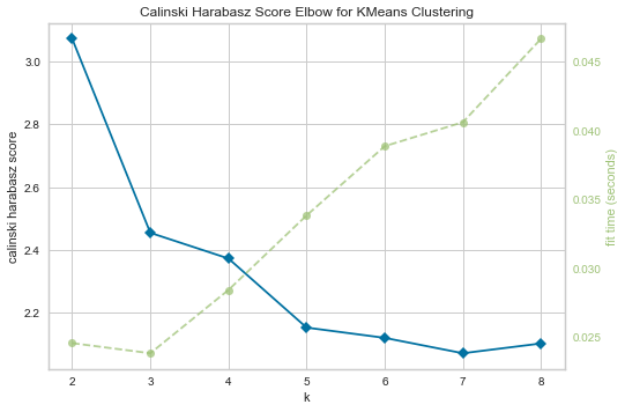
where B(k) is the inter-cluster variance (i.e. the sum of squared distances for the k clusters), and W (k) is the intracluster variance. Maximising CH(k) against different values of k gives the estimated number of clusters (Kingrani and Dell Zhang).



(a) Distortion score elbow for k-means clustering.



(b) Silhouette score elbow for k-means clustering.



(c) Carlinski-Harabasz score elbow for k-means clustering.

Figure 3: The optimal number of clusters analysis.

We could not estimate an optimal k parameter from the inconclusive results obtained by the Elbow method, for the three different metrics used. A further investigation to choose a more suitable method is out of the scope of the present project, which is to illustrate the application of web scrapping, Foursquare API and, k-means clustering to explore neighborhoods in Toronto. However, to continuing our analysis, we will chose k based on the amount of time to train the clustering model. Therefore, four clusters will be chosen, since it corresponds to the lowest value of the amount of time to train the clustering model considering the distortion metric and, it also corresponds to reasonable low values considering the silhouette and Calinski- Harabasz metrics. A more accurate investigation, with an advanced methodology to estimate the optimal number of clusters, is out of the scope of the present work. However, to the ones interested in this topic, I strongly recommend the article published by Boris Mirkin.

4.3 Clusters insights

The insights observed, Fig. 4,considering the five most common venues categories per cluster, are summarized below:

- **Cluster 0:** The boroughs contained in this cluster have shown balanced venues categories, in which you can find restaurants, fast foods, coffee shops, pharmacies, banks, and home services pretty much almost everything that any person could need. Besides, entertainment (Hockey arena) can also be found.
- **Cluster 1:** It seems to be a business center, with a focus on coffee shops, cafés, having hotels and restaurants available for a possible need.
- **Cluster 2:** There is a focus on gastronomical variety here, you can find several types of restaurants. Besides, entertainment for adults (pubs) is also available.
- **Cluster 3:** It seems to be a wealthy residential location with a lot of places to hang out with friends (bar), to chill (coffee shops), and good gastronomical options (restaurants, cafés, and bakeries).
- **Cluster 4:** This cluster appears to be a typical middle-class residential location, with parks for recreation, clothing stores for a possible need, coffee shops to chill, and some gastronomical options (pizza place and restaurant).

:

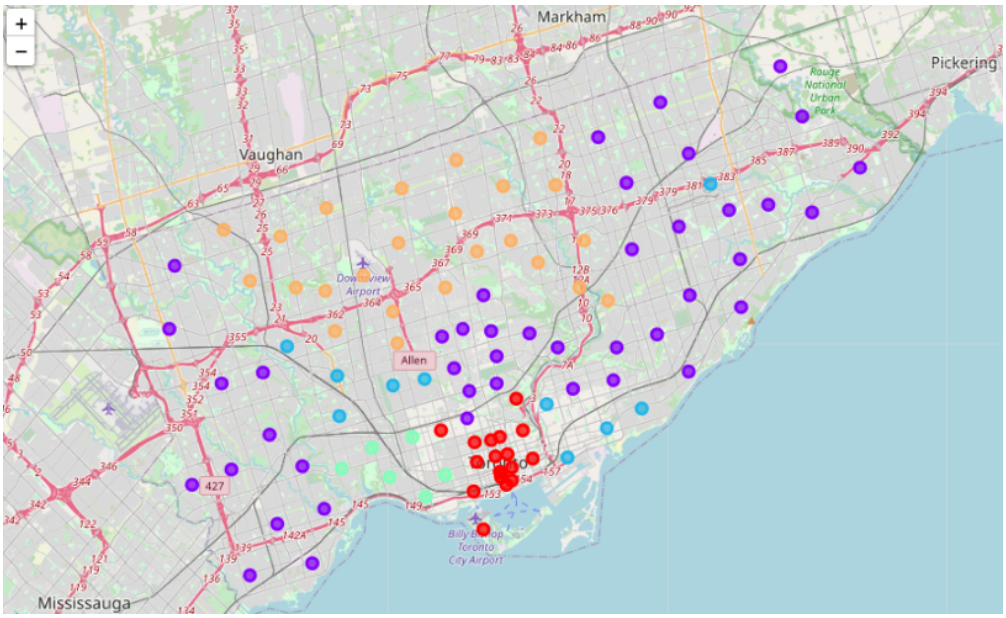


Figure 4: Visualizing clusters in Toronto.

5 Conclusion

The present study aimed at the analysis of commercial patterns in each particular borough in Toronto, focusing on to scope out the competition, to find the optimal location to open a new restaurant. Based on the patterns encountered, we can conclude the following:

- Cluster 2 boroughs would represent the hardest competition due to the gastronomical variety.
- The higher investment would be needed at boroughs in Clusters 1 and 3, due to the high quality provided by competing restaurants.
- Even though Cluster 3 boroughs represent a balanced environment, it can be seen a reasonable variety of restaurants and fast food. High creativity and quality advertising would be necessary to attract customers.
- The most reasonable choice, based in our clustering analysis, would be boroughs at Cluster 4. They have not so many gastronomical options, summed with the middle class-like environment, would not require a very sophisticated restaurant and excessive investment in advertisements.