IBM Data Science Capstone

Final Report: The Battle of the Neighborhoods

Project: Neighborhood Similarities Between Cities

Paula Orellana, September 2019

Contents

Introduction 3

Business Problem 3

Target Audience 3

Data 4

Sources and Extraction 4

Methodology 5

Project Objective 5

Analytic Approach 5

Results 9

Discussion 10

Conclusion 10

References 10

Introduction

At some point every Business Owners think about expansion, once their business is locally consolidated the next step is move into international waters. This decision comes with a new set of challenges and opportunities but also with a higher risk since this is literally a New Territory. An analysis of similarities between neighborhoods will allow comparing known successful locations against new candidates, and make and inform decision to where open a new business branch.

Business Problem

Our client is a successful owner of several boutique French Cafes in New York City, which is currently evaluating the possibility to start a new venture in Toronto. For that, we need to answer the following question: which neighborhoods in Toronto have the biggest similarities with New York neighborhoods were the customer enterprise has proved itself successful?.

The customer successful venues are located in the following neighborhoods:

- Brooklyn: Bedford Stuyvesand, Boerum Hill, North Side
- Manhattan: Hamilton Heights, Manhattanville, Murray Hill
- Queens: Hunters Point, Long Island City
- Staten Island: Arden Heights

Target Audience

The current project will be presented to the Business Owner and Directive Board, it would require a high level presentation of the analysis in order for the executives to appreciate the validity of the final recommendations.

Additionally the results can be of interest to any business owner that want to expand their business into Toronto city.

Data

For this project, the following data will be used:

- New York City neighborhood data.
- Toronto neighborhood data
- Venue Data for FourSquare, that will allow us to cluster and classify neighborhoods together and identify the similarities between both cities.

Sources and Extraction

Data 1:

New York City neighborhood data will be extracted from a dataset available for free on the web:

https://geo.nyu.edu/catalog/nyu_2451_34572

This information is structured in 5 boroughs and 306 neighborhoods and contains the latitude and longitude coordinates of each neighborhood

Data 2:

Toronto City information will be extracted from the city's Wikipedia page:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,

We'll scrap the data using BeutifulSoup, and we were leveraging the Google Maps Geocoding API to get the latitude and the longitude coordinates of each neighborhood.

The raw data will be loaded in dataframes , where can be analyses and cleaned in order to apply machine learning technics and map visualization. In the next sections of Methodology and Results the details of the analysis will be presented and discussed.

Methodology

Project Objective

Or main goal is answer the business question and provide a recommendation of the optimum location, based in the similarities with already successful venues, of anew French Café boutique in Toronto.

Analytic Approach

The steps that will be taken to complete the project will be as follow:

1. Obtain access to the list of neighborhoods in the cities on interest. As specified in the Data section, these will be acquired from the web from NYU and Wikipedia.

2. New York data was loaded into Pandas dataframes workable in python. This provided us with the names of the neighborhoods and boroughs. Data was then cleaned and adjusted to create a consistent set to work with.

3. While NYC data already contain geographical information that was not the case for Toronto. Therefore the geographic coordinated had to be obtained in an additional step in order to be able to create map visualizations of both cities using Folium package.

4. GeoPy package was used to convert addresses into geographical coordinates. Again data was reviewed to modify or discard any inconsistent information.

5. Thru exploratory data analysis we were able to study the neighborhoods and their venues, we had to use FourSquare information to complete this step.

    a. Venue information for each of the neighborhoods was obtained by using FourSquare API. Information was restricted to 100 venues from 500 meters of each zone.

    b. The JSON file provided by FourSquare was load into a Panda dataframe to allow us to explore the venue information.

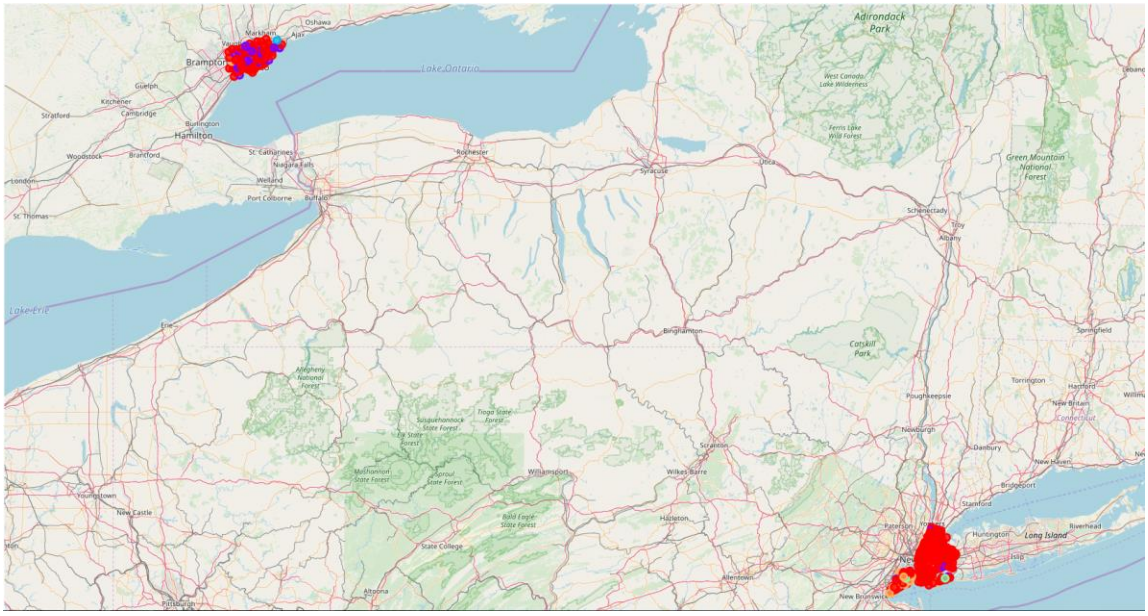
    c. We then examine how many unique categories can be selected, organized, and presented from all the returned venues,

    d. Analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category

6. Then, we created the set of neighborhoods and venues combined from both cities. This allowed us to:

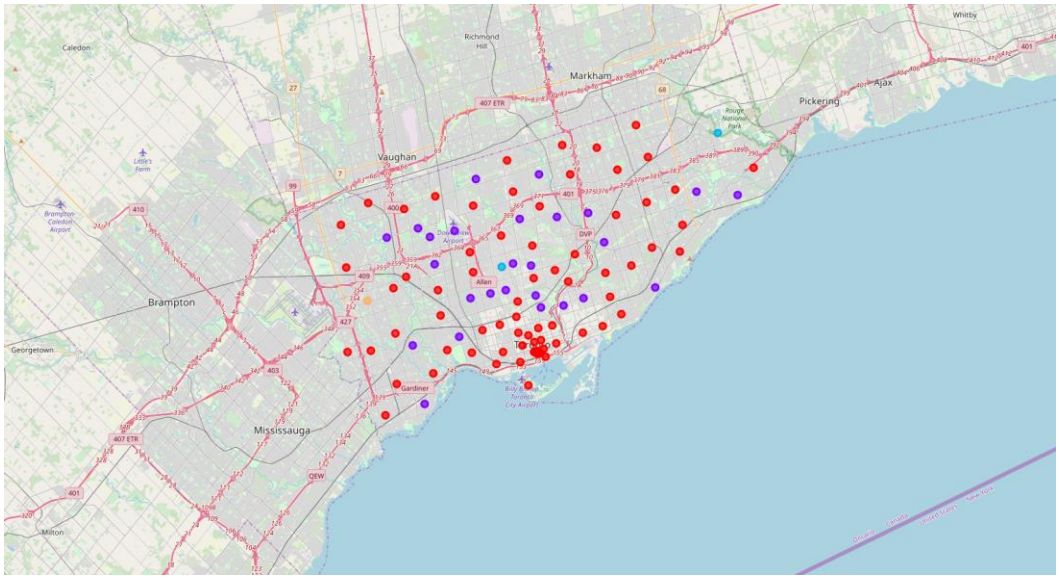
    a. Cluster the neighborhoods based in their similarities. K-mean cluster algorithm was our tool in this step, this algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster

    b. Five clusters were created in our combined dataset. The combined cluster are showed in the following map

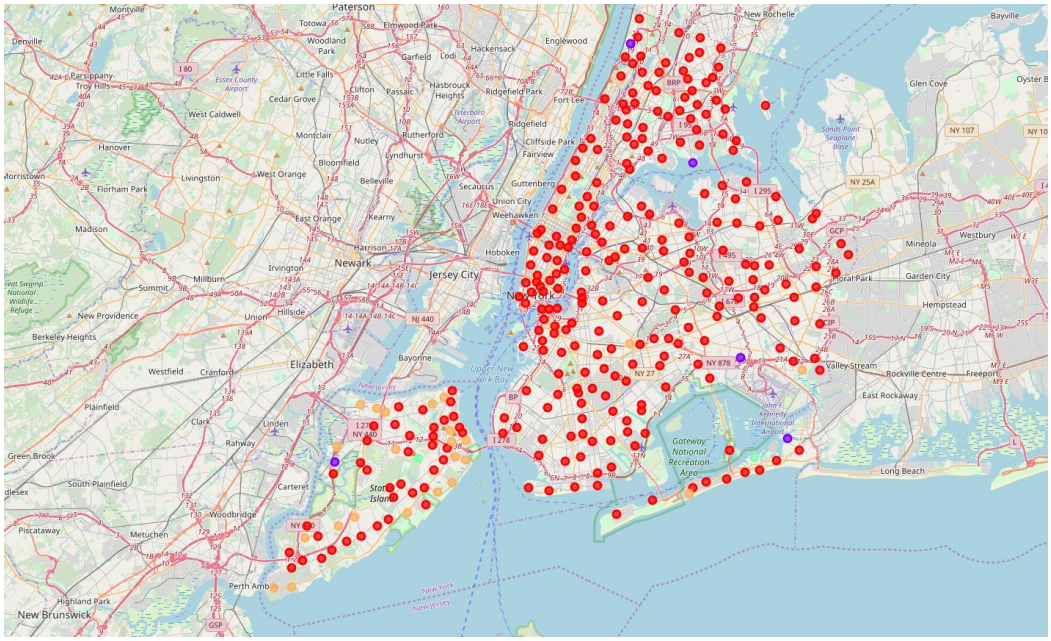


Detail of the cities the show us that the distribution if the clusters is different for Toronto and New York.

Toronto



New York



c. A top 5 ranking was compiled from that statistic data, for each of the neighborhoods to identify the cluster characteristics based in their more popular venues.

- Cluster 1 consist of Neighborhoods with casual eating locations such as Pizza Place and Coffee Shops as the most common venues

- Cluster 2 consist of Neighborhoods with Parks and open Fields most common venues

- Cluster 3 consist of Neighborhoods with Construction and Home improvement as the most common venues

- Cluster 4 consist of Neighborhoods with Gift Shops and Parks as the most common venues - Probably touristic attractions close by

- Cluster 5 consist of Neighborhoods with Bus Stops as the most common venues

d. Finally, we used the know location of already successful Cafes, owned by our customer, to identify the cluster were they are located and their venues distribution.

    - Our customer know successful are located in Cluster 1, and we can describe them as neighborhoods were the most popular venues are Coffee Shop with 2nd and 3rd places also allocated by food related services.

	City	ClusterLabel	Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
54	NYC	0	Brooklyn	Bedford Stuyvesant	Café	Coffee Shop	Pizza Place	Bar	Vietnamese Restaurant
57	NYC	0	Brooklyn	Boerum Hill	Coffee Shop	Dance Studio	French Restaurant	Spa	Sandwich Place
80	NYC	0	Brooklyn	Flatbush	Bakery	Coffee Shop	Caribbean Restaurant	Mexican Restaurant	Chinese Restaurant
101	NYC	0	Brooklyn	North Side	Coffee Shop	Pizza Place	Bar	Jewelry Store	American Restaurant
131	NYC	0	Manhattan	Financial District	Coffee Shop	Hotel	Pizza Place	Wine Shop	Gym
135	NYC	0	Manhattan	Hamilton Heights	Café	Pizza Place	Mexican Restaurant	Deli / Bodega	Coffee Shop
143	NYC	0	Manhattan	Manhattanville	Coffee Shop	Park	Mexican Restaurant	Seafood Restaurant	Fried Chicken Joint
148	NYC	0	Manhattan	Murray Hill	Coffee Shop	Japanese Restaurant	Hotel	Sandwich Place	French Restaurant
198	NYC	0	Queens	Hunters Point	Café	Italian Restaurant	Japanese Restaurant	Wine Shop	Thai Restaurant
209	NYC	0	Queens	Long Island City	Coffee Shop	Hotel	Bar	Pizza Place	Café
213	NYC	0	Queens	Murray Hill	Korean Restaurant	Coffee Shop	Bar	Bank	Supermarket
244	NYC	4	Staten Island	Arden Heights	Coffee Shop	Bus Stop	Pharmacy	Pizza Place	French Restaurant

e. We brought these characteristics into the Toronto cities in the same cluster 1, to obtain the neighborhoods with the biggest similarities, these neighborhoods are our recommended places for open a new venue.

Results

Thru the previously detailed analysis Cluster 1 was the one identified as the one with the potential characteristics to support a successful new Café for our client. Since over 90% of the known venues reside in that cluster.

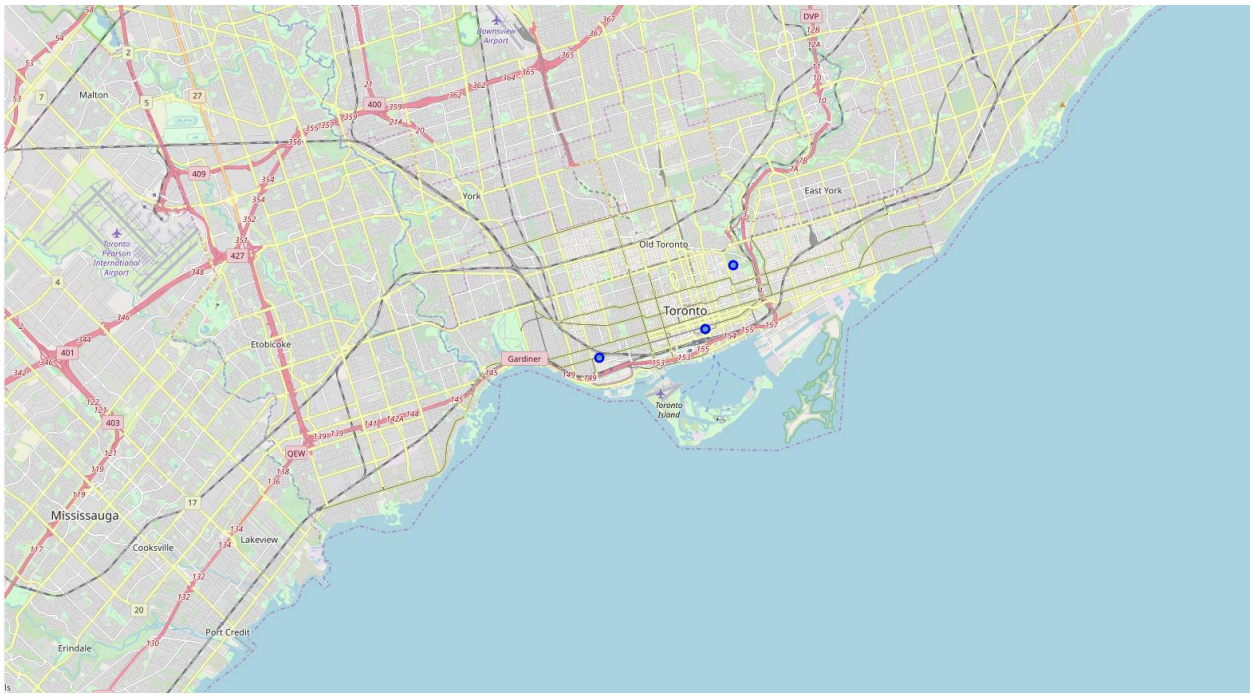
Using that finding as main filter we then selected only the Toronto cities that have Coffee Shops and Café as the First most common venues. To reduce even more the set, to a number easily manageable by the stockholders, we also include in the selection the 2nd and 3rd most popular venues characteristics, which are food related business, such as Pizza Place, Restaurants and Bakery

Our final set of recommendations for possible new locations consists in 3 neighborhoods:

- Cabbagetown,St. James Town in Downtown Toronto
- Commerce Court,Victoria Hotel in Downtown Toronto
- Brockton,Exhibition Place,Parkdale Village in West Toronto

City	Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Toronto	Downtown Toronto	Cabbagetown,St. James Town	Coffee Shop	Restaurant	Café	Italian Restaurant	Pizza Place
Toronto	Downtown Toronto	Commerce Court,Victoria Hotel	Coffee Shop	Restaurant	Café	Hotel	American Restaurant
Toronto	West Toronto	Brockton,Exhibition Place,Parkdale Village	Coffee Shop	Café	Restaurant	Sandwich Place	Bakery

Visualization in the following map, show how close are one of each other.



Discussion

Since two of the neighborhoods are in Downtown Toronto, we will point that borough as the one most promising to our client.

While our three recommendations are the ones supported by previous experiences, would be interesting to explore other Clusters that also seem to have a high presence of business alike to the ones belonging to our customer. In particular Cluster 2, where the market can be less saturated than our current selection, since Coffee Shops are for the most part only in 2nd or 3rd place in the Top 5 venues, but it also have a respectable number of food related business.

Conclusion

The current analysis was performed with limited data, only frequency of occurrence of various venue categories was considered as the similarity factor. There are other interesting data that can enrich the criteria, such as population density, demographics, and cuisine type distribution in the cities that could influence the selection of a future venue location. More in deep analysis can be made bringing these data sets into play and final recommendation can change accordingly.

Nevertheless, our 3 recommendations are supported by a methodological complete analysis that can be explained in detail if need it. These results also respond the business question identified at the beginning of the project in a concise and direct way easy to communicate to the stakeholders.

References

Neighborhoods of Toronto Wikipedia page:

- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Neighborhoods of New York City:

- https://geo.nyu.edu/catalog/nyu_2451_34572