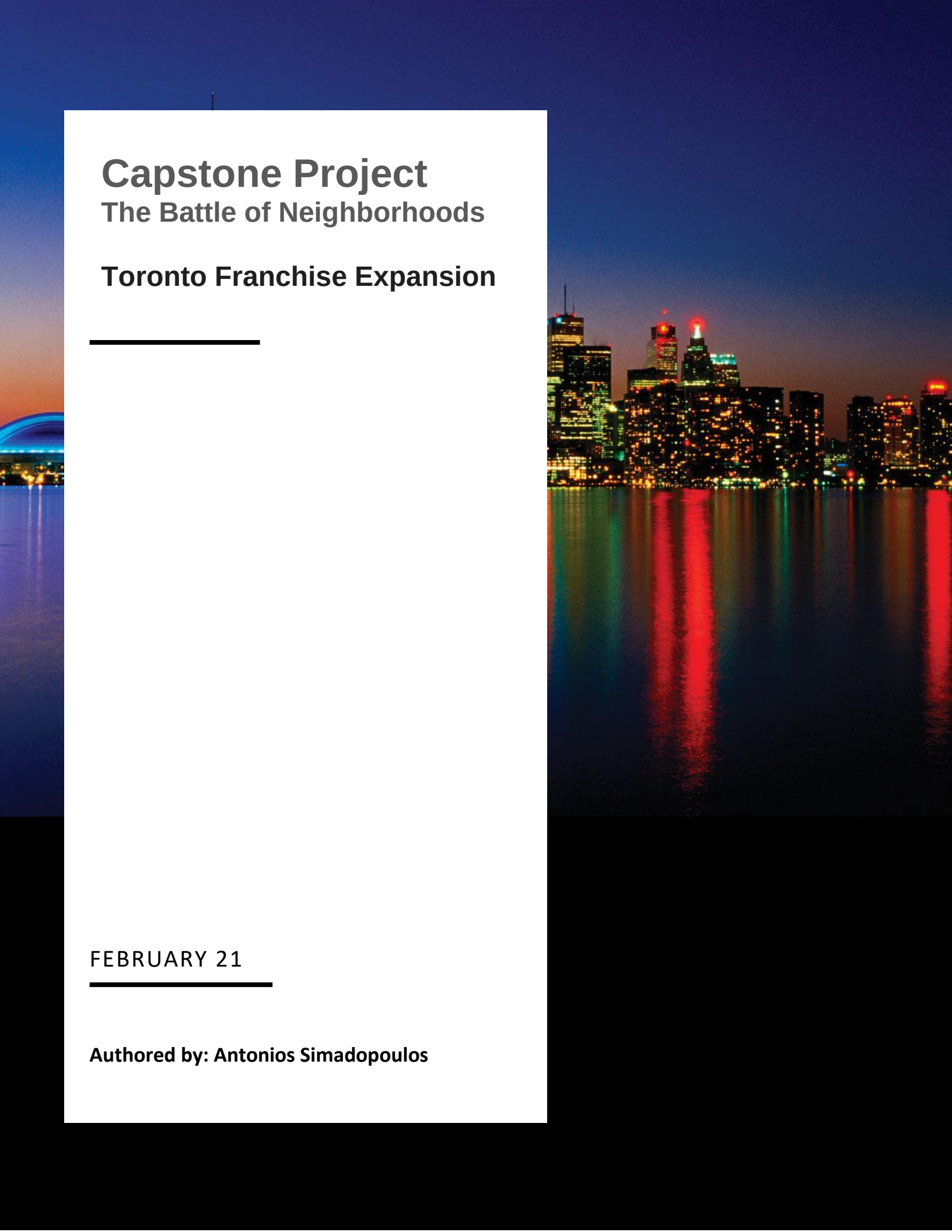


Capstone Project

The Battle of Neighborhoods

Toronto Franchise Expansion

FEBRUARY 21

Authored by: Antonios Simadopoulos

1. Introduction

1.1 Background

Toronto is the financial capital and of the most famous cities in Canada. Hence, it attracts and hosts a variety of different businesses, store chains and franchises. Some of them they have plans for a further international expansion in other key cities around the world with top priority being New York in the USA. The owners and stakeholders of these businesses want to minimize the risk of their investments and maximize their potential returns. For this reason, it would be of a great interest to them to be able to identify whether the two cities are similar and if yes, which specific areas of the cities should prioritize for their selection.

The objective of this analysis is twofold. First, we aim to answer the question of whether Toronto and New York have similar neighborhoods with the same characteristics. Secondly, to develop a consulting tool able to propose the most promising areas in New York for a particular type of business that is performing well in Toronto.

1.2 Audience

The audience of this analysis is chain store and franchise owners in Toronto, who are willing to expand their businesses in New York. This toll aims to work best for businesses with physical presence (brick & mortar stores) and not online services or stores. However, there is no limitation on the type of business the tool can support.

2. Data

In this analysis we use the following data types:

- Population and Age Distribution of the two cities [1] [2]
- The geographical coordinates of New York neighborhoods [3]
- The geographical coordinates of Toronto neighborhoods [4]
- Foursquare API to retrieve the available venues for each neighborhood

The demographic information will help us to establish a good understanding of the population size of the two cities, but also the distribution of the various age groups. This will serve us as the first insight to how similar or not the two cities are. The geographical coordinates of the cities' neighborhoods normally can be retrieved from the goopy package. However, since this package is unreliable most of the time, we will procure this information from existing files. Finally, based on the geographical coordinates collected on the previous step, we will use the API provided by Foursquare to retrieve the venues for each city's neighborhoods. Regarding the venues, we will be able to identify how many venues each neighborhood has, what type of venues, the exact location of each venue, ratings from customers and even comments. After targeting the most promising areas for expansion of a business, this data can also be used to identify and monitor the competitive landscape. Namely, how many businesses of the same type exist and how well they perform.

3. Methodology

3.1. Demographic composition of Toronto and New York

The first part of the analysis focuses on the demographic composition of Toronto and New York. The purpose of this first step is to explore how similar the two cities are in terms of the distribution of their population per age group and enable us to determine whether a company that targets a particular age group will have high likelihood of success by expanding from Toronto to New York. To achieve that we will leverage the Wikipedia entries for the two cities in question, where they include the required demographic data. For each city, the process of cleaning and preparing the data is slightly different due to the format of the raw data. You may find the exact pre-processing steps in the respective notebook. The first table illustrates the age distribution for Toronto and the second one for New York.

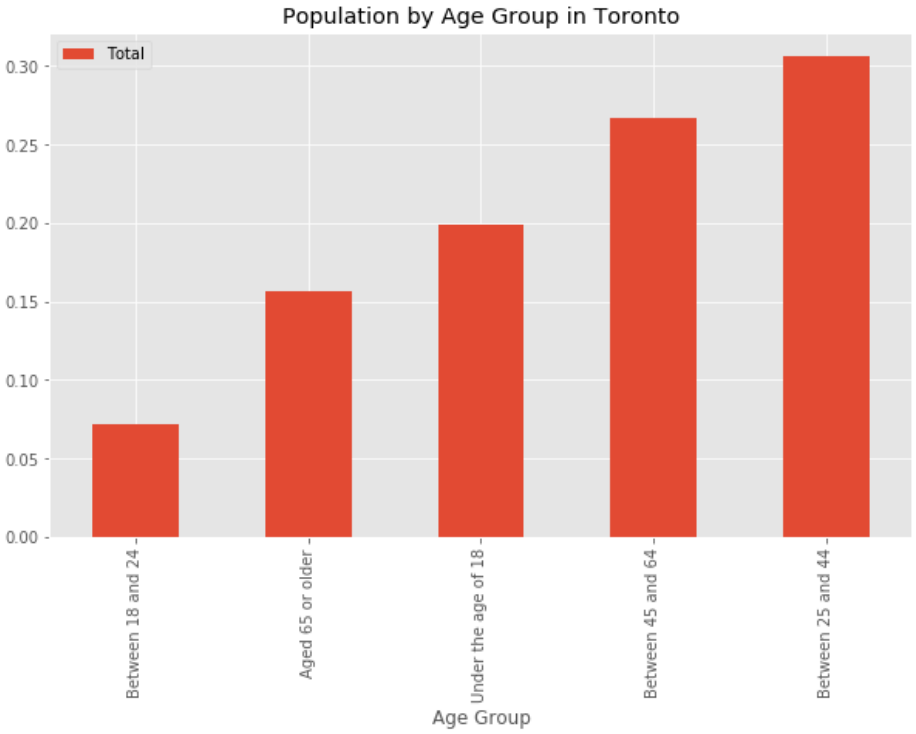
Table 1: Toronto Age Distribution

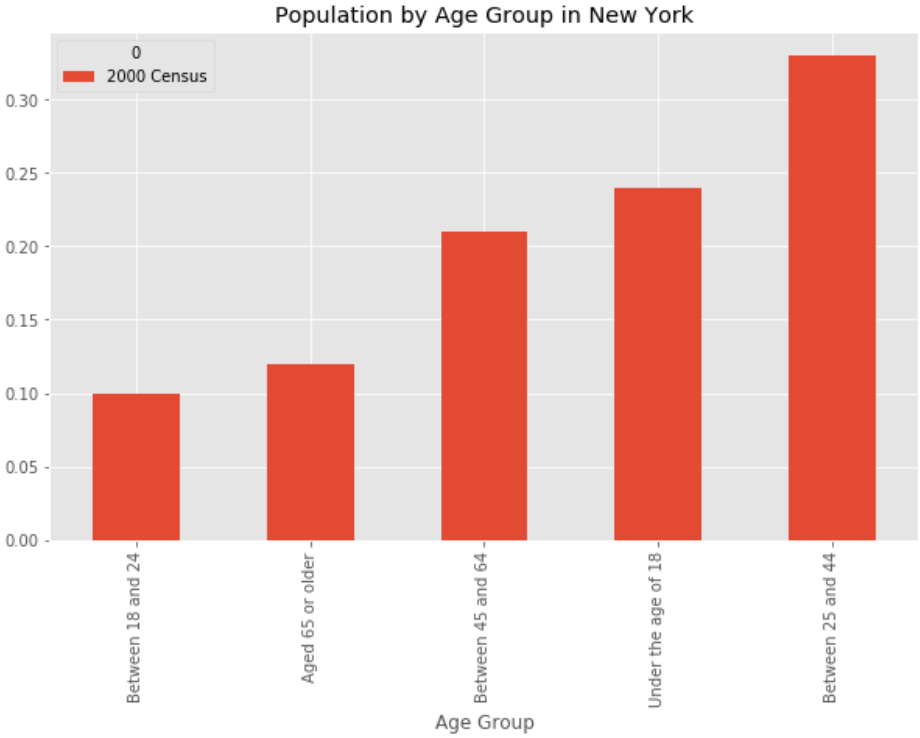
Age Group	Split
Under the age of 18	20%
Between 18 and 24	7%
Between 25 and 44	31%
Between 45 and 64	27%
Aged 65 or older	16%

Table 2: New York Age Distribution

Age Group	Split
Under the age of 18	24%
Between 18 and 24	10%
Between 25 and 44	33%
Between 45 and 64	21%
Aged 65 or older	12%

We can clearly see that Toronto has more people with an age of 45 years old and older than New York. On the contrary, New York's population appears to be younger, especially for the people less than 24 years old. To get a better understanding of the above data, we can visualize the results and draw our conclusions.


Population by Age Group in Toronto
Total
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Between 18 and 24
Aged 65 or older
Under the age of 18
Between 45 and 64
Between 25 and 44
Age Group


Population by Age Group in New York
0
2000 Census
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Between 18 and 24
Aged 65 or older
Between 45 and 64
Under the age of 18
Between 25 and 44
Age Group

3.2. Neighborhoods' Insights

Having concluded the first insights of our analysis, we proceed by having a look at how the neighborhoods of the cities look like. Our first step is to get the geographical coordinates of the neighborhoods of each city by the data sources mentioned in the previous section and visualize the results.

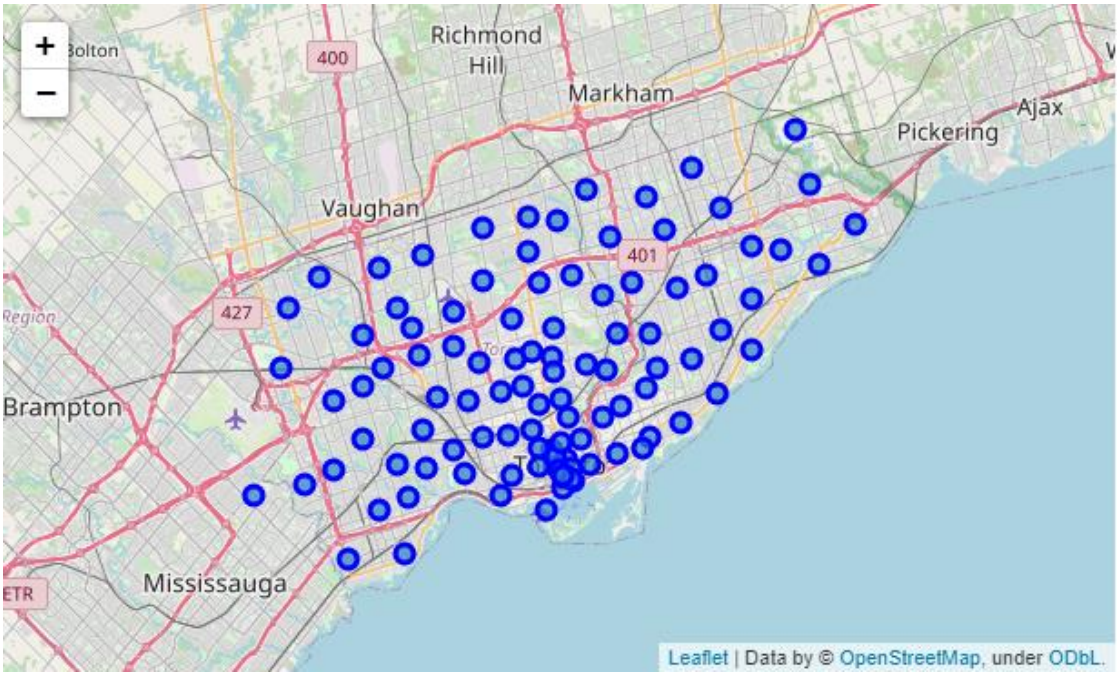


Figure 1: Toronto Neighborhoods

Figure 2: New York Neighborhoods

By looking at the above maps, we can see that Toronto has clearly less neighborhoods that New York. Moreover, the neighborhoods of Toronto appear to be more uniformly distributed across the city, while in New York the pattern does not follow any visible rule. We could infer that New York might have neighborhoods of different sizes and types, namely a bigger variety and culture in the city. Toronto, on the other side, is expected to be more consistent in these terms. To verify these assumptions, we can retrieve the venue types for each city by using Foursquare's API. We can the calculate how many different venue types we have in each city and validate our previous assumption. Indeed, as it appears also in the respective section of the notebook, Toronto has 265 unique venue categories while New York has 440. We can then validate our hypothesis that New York offers a more versatile environment.

3.3. Clustering

Equipped with a better understanding on the environment of the two cities, we proceed with the core part of our analysis. Namely, to find similar neighborhoods across the two cities in terms of the currently available venue types and be able to recommend which neighborhoods in New York will be more suitable for an existing business in Toronto willing to expand their presence in the former city. To achieve this result, we first merge the neighborhoods of the two cities in a common dataframe, along with their geographical coordinates and venues. This dataset will be used for unsupervised clustering using the k-Means algorithm. We will use 10 clusters to get an adequate granularity on the different neighborhood types. Finally, we will plot the results to get a better understanding of how the results look like. As you may see in the below figure, the neighborhoods of the two cities have been clustered based on the same rules.

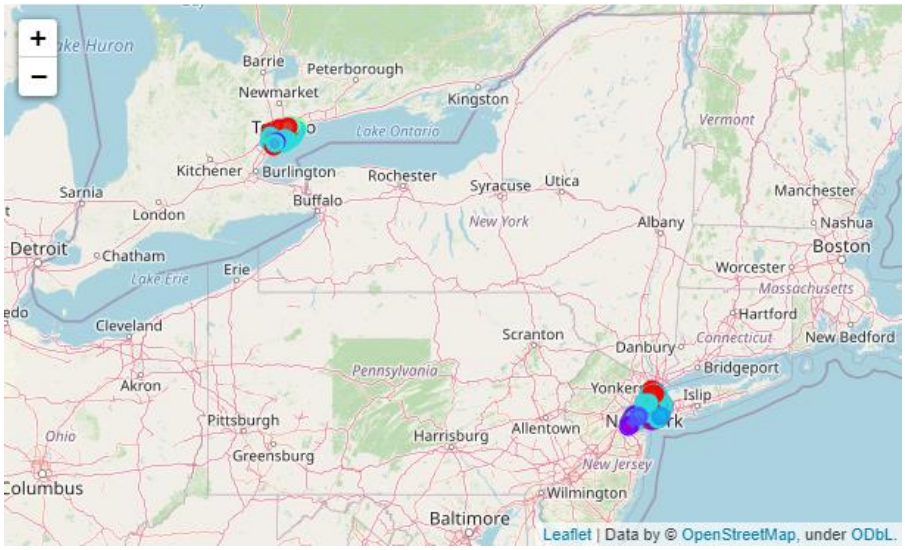


Figure 3: Clustering results for Toronto and New York

Let us now focus on Toronto to see how the clustered neighborhoods look like.

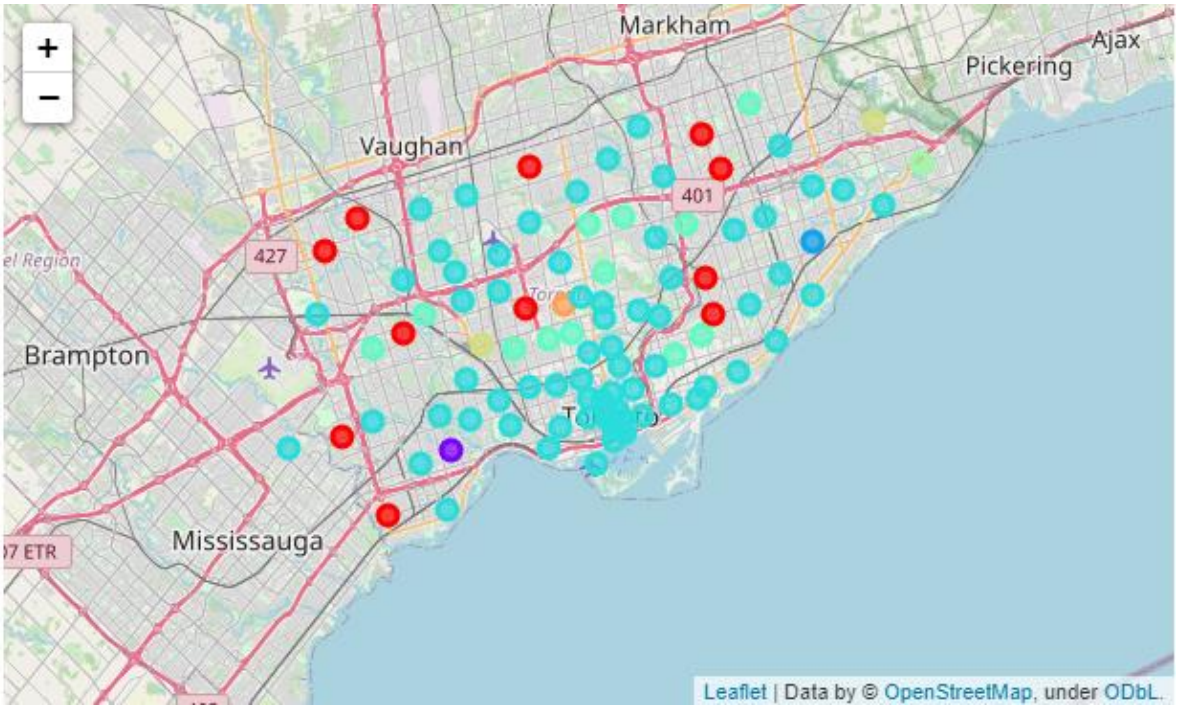


Figure 4: Toronto clusters

And here, the same representation for New York city.

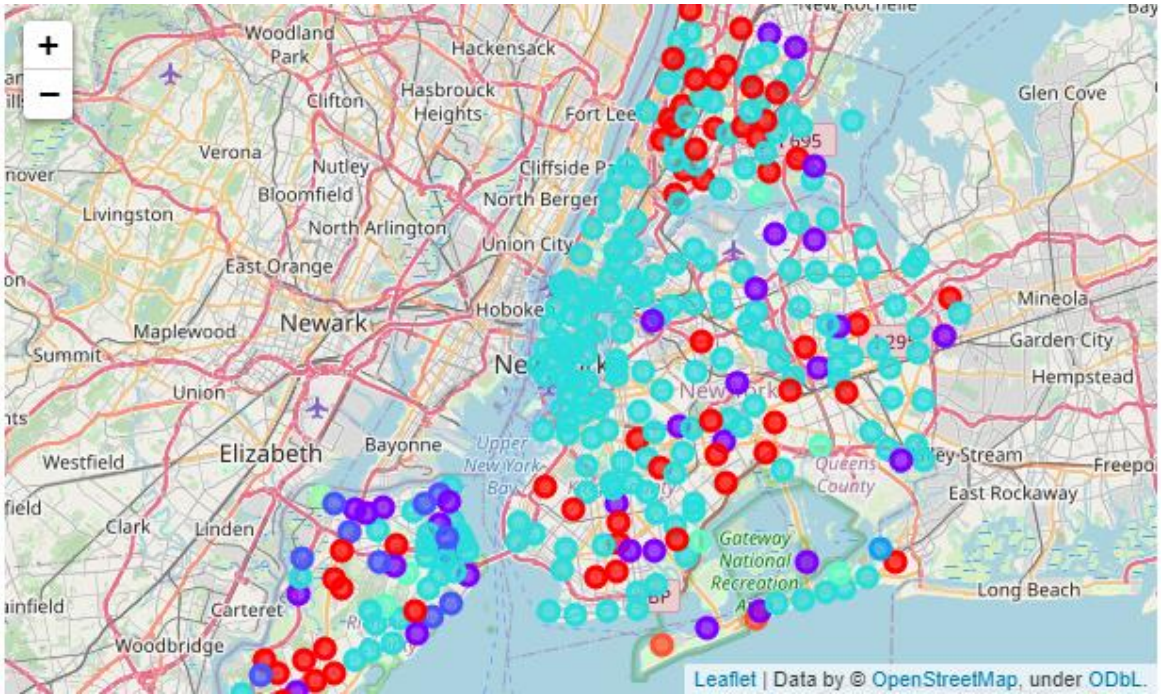


Figure 5: New York clusters

4. Results

We can see that in both cities, we have three main clusters that describe most of the neighborhoods, based on the currently available venue types. Furthermore, we observe that most of the neighborhoods in Toronto's city center belong to Cluster 4, highlighted with a light blue-green color. The same cluster is also present in New York, mainly in Mid/Lower Manhattan. Exploring in depth this cluster, we can identify that the most frequent venue types in these neighborhoods are Coffee Shops, Clothing Stores, Restaurants and Gyms. In addition to that, we can also observe that this type of neighborhoods can be found in both cities in more de-centralized locations but less frequently. If we assume that our client has a Restaurant business in Toronto's city center, we are now able to recommend which locations in New York are more suitable for expanding their business in.

Additionally, we can enhance our recommendation by exploring how the competition looks like in the targeted neighborhoods. Namely, how many restaurants exist, what is ratio of the restaurants per capita for each location, and how many restaurants of a particular type exist in each neighborhood. For example, if our client has a Mexican restaurant, we can identify the suitable areas and then sort them based on how many Mexican restaurants they have. This can provide us with adequate information of choosing a promising area with the lowest possible level of competition.

5. Discussion

This analysis aims to provide a consulting tool on businesses that they want to expand their presence in more geographical locations. More specifically, we leverage our current knowledge of where the business is currently performing well and based on that we can decide which areas in a new location provide a similar environment and hence higher potential returns on the investment. This tool can be applied in more than two cities and in any location if we have adequate data to derive insights for the neighborhoods. Moreover, the tool can be further improved by getting more granular demographic information at a neighborhood level. For example, age group distribution, population, and nationalities. This information can work in parallel with the clustering based on venue types and provide a more solid way of providing a recommendation. This data-driven approach can leverage the power of data analysis to provide evidence-based business recommendations and generate results for our customers.

Bibliography

1. https://en.wikipedia.org/wiki/Demographics_of_New_York_City
2. https://en.wikipedia.org/wiki/Demographics_of_Toronto
3. Coursera Lab Server: “newyork_data.json”
4. https://cocl.us/Geospatial_data