The battle of neighborhoods - office in Toronto.

Kamil K.

July 20, 2021

Table of Contents

1. Introduction.

1.1 Background

As reported by Wikipedia, New York is the most populous city in the United States [1]. Its population in 2020 was estimated as 8,253,213 people. What is important New York City is also the most densely populated major city in the United State (more than 27 thousand people per square kilometer). As one of the most important cities of the United States and even the whole world, New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.
The second, undoubtedly interesting city in North America, is Toronto. Toronto is the capital city of the Canadian province of Ontario. As Wikipedia says [2] with a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. Toronto is a prominent centre for music, theatre, motion picture production, and television production. Its varied cultural institutions with numerous museums and galleries, festivals and public events, entertainment districts, national historic sites, and sports activities.
As we can see from this encyclopedic review, both cities are quite large and have many common features. The common features of both cities certainly include [3]:

- Both are the biggest cities in their respective countries.
- Both are the financial capitals in their respective countries.
- Both are the most crowded cities in their respective countries.
- Both cities have similar weather.
- Both cities are safe.
- Both cities are entertainment centers and big on the arts.

However, according to ranking Best Places to Live in North America [4] Toronto seems to be better place to live - in the mentioned ranking Toronto is ranked on the first place, while New York is eighteenth.

1.2 Problem

My client, Data-X Co., is considering relocating its business from today's location in New York to Toronto. He confirmed the fact that most of its employees have expressed such a desire (will move to Toronto with the company). Generally, all believe that with all the similarities between cities, Toronto is a bit less crowded and will be easier to maintain a work-life balance. As everyone appreciated the previous location in terms of available attractions the client wants to move the office to similar, nice place if it is possible.
The problem comes down to answering the question:
Where would I recommend to open the new office taking into account similarity of the new and old location (venues availability)?

1.3 Interest

In the era of high mobility of employees, striving to improve the quality of life by employees, looking for work-life balance but also seeking lower costs of running a business, finding the answer on asked question can be interesting for a large group of entrepreneurs who want to move their headquarters to other countries or open new offices in new location with certain factors unchanging.

2. Data

2.1 Data sources

For the project I will use the following data:

- Toronto City data that contains Borough, Neighborhoods

  Data Source: Wikipedia

  Description: The Wikipedia site contain all the information I need to explore and cluster the neighbourhoods in Toronto. I will scrape the Wikipedia page using pandas package, clean it, and then read it into a pandas data frame.

- Geographical Location data using Geocoder Package

  Data Source:

  for API: https://cocl.us/Geospatial_data

  for latitudes and longitudes of Toronto: Geospatial_Coordinates.csv

  Description: The second source of data was provided with the Geographical coordinates of the neighbourhoods with the Postal Codes of Toronto. I will combine the above data with latitudes and longitudes of Toronto's neighbourhoods.

- New York data that contains Borough, Neighbourhoods, latitude and longitude

  Data Source: newyork_data.json

  Description: The third source of data was provided with the Geographical coordinates of the neighbourhoods of NY. I will transform the json file into pandas data frame. Both data (NY and Toronto) I will use to find venues of each locations.

- Venue Data using Foursquare API

  Data Source: https://foursquare.com/developers/apps

Description: From Foursquare API we can get the name, category, latitude, longitude for each venue. Then I will this feature to group the neighbourhoods into clusters. I will use the k-means clustering algorithm to complete the task. Finally, I will also use the Folium library to visualize the neighbourhoods in Toronto and their emerging clusters.

2.2 Data aquisition and cleaning

2.2.1 Downloading and Exploring Toronto Dataset

To clean, prepare and present the data I used the following libraries: numpy, pandas, json, geopy, matplotlib, requests, sklearn and folium.

To download the data from Wikipedia I used pandas library. First, I imported all needed libraries and connected to url: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

There are several problems with the datasets:

- the initial table downloaded directly from the website had a shape 20x9 – therefore, it required transformation
- In the first version of the table with Postal Code, Borough and Neighborhood I saw that 5 rows have not very nice text in Borough column which should be modify:

	One_column	Postal Code	Borough	Neighborhood
40	M4JEast YorkEast Toronto(The Danforth East)	M4J	East YorkEast Toronto	The Danforth East
69	M5WDowntown TorontoStn A PO Boxes25 The Esplan...	M5W	Downtown TorontoStn A PO Boxes25 The Esplanade	Enclave of M5E
86	M7RMississaugaCanada Post Gateway Processing C...	M7R	MississaugaCanada Post Gateway Processing Centre	Enclave of L4W
87	M7YEast TorontoBusiness reply mail Processing ...	M7Y	East TorontoBusiness reply mail Processing Cen...	Enclave of M4L
102	M9WEtobicokeNorthwest(Clairville / Humberwood ...	M9W	EtobicokeNorthwest	Clairville, Humberwood, Woodbine Downs, West H...

In the code I did some modifications of the lines presented above, I:

- add the space between East York and East Toronto for M4J
- change Borough and Neighborhood to Downtown Toronto for M5W
- change Borough and Neighborhood to Mississauga for M5W
- change Borough and Neighborhood to East Toronto for M7Y
- add the space between Etobicoke and Northwest for M9W
- delete One_column

Then I created a data frame of the latitude and longitudes of the Toronto Neighborhoods from the file provided by Coursera and join them with the table created on a basis of Wikipedia website.

2.2.2 Adding the New York data to basic dataset

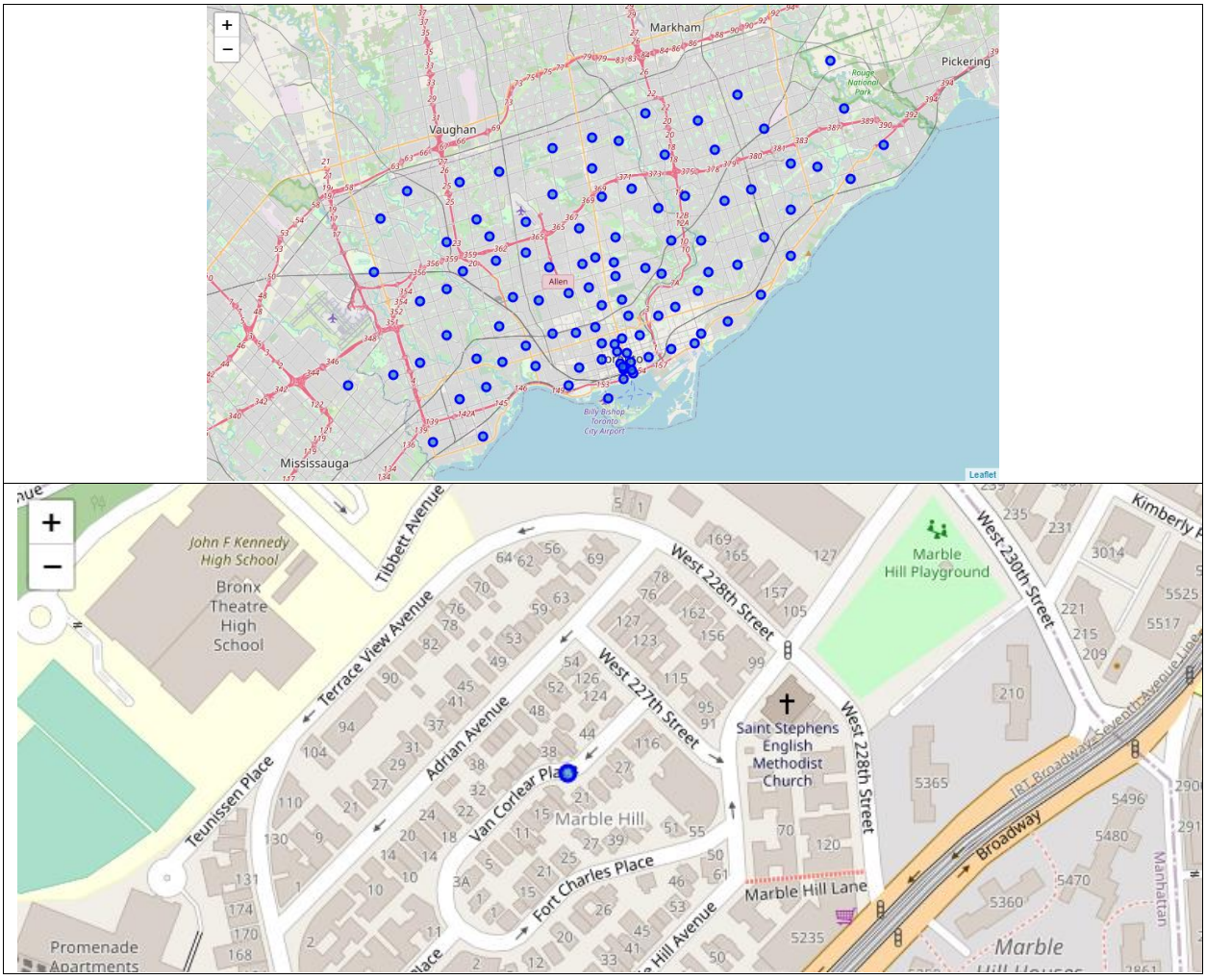
In this stage I've added manually one additional row to the table with Borough and Neighborhood of my client address. It allows me to cluster the address together with Toronto data:

Borough: Manhattan
Neighborhood: Marble Hill
Latitude: 40.876551
Longitude: -73.910660

The resulting data frame had finally 14 boroughs and 104 neighborhoods.

2.2.3 Presentation of the locations on a map.

To check if the data are ready to be cluster I prepared map with all locations from my data frame. The maps are presented below.



2.2.4 Data acquisition from Foursquare

In this stage first I defined Foursquare Credentials and then I created a function **getNearbyVenues** to explore the nearby venues by given latitudes, longitudes and radius 750 m. I used this function on my Toronto data frame.

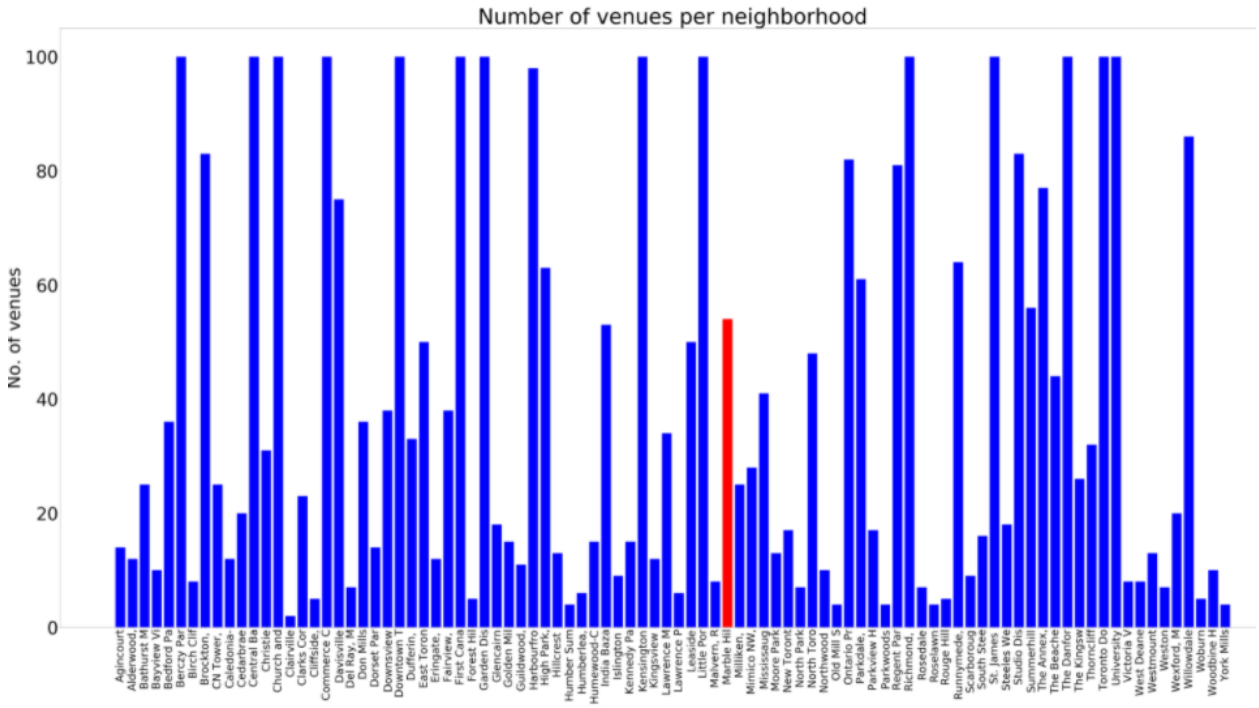
As the result of using **getNearbyVenues** I got data frame with 3717 rows and 7 columns. It's important to remember that Marble Hill is also a part of the data.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
2	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
3	Malvern, Rouge	43.806686	-79.194353	Staples Morningside	43.800285	-79.196607	Paper / Office Supplies Store
4	Malvern, Rouge	43.806686	-79.194353	Lee Valley	43.803161	-79.199681	Hobby Shop

```
Toronto_venues.tail()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
3712	Marble Hill	40.876551	-73.91066	Ray's Pizza Express	40.878901	-73.916558	Pizza Place
3713	Marble Hill	40.876551	-73.91066	Caridad Spanish Restaurant	40.871832	-73.906700	Spanish Restaurant
3714	Marble Hill	40.876551	-73.91066	Dick Savitt Tennis Center	40.873845	-73.917391	Tennis Court
3715	Marble Hill	40.876551	-73.91066	MTA Bx7 Bx20 - 218th & Bway	40.870979	-73.914778	Bus Station
3716	Marble Hill	40.876551	-73.91066	Elsa's Hair Salon	40.871959	-73.904651	Salon / Barbershop

Then I prepared the bar graph which present me a number of venues per neighborhood. As we can see from the graph presented below, for Marble Hill we have 54 Venues and that is roughly halfway in the 0-100 range.



In this stage I also check if I have sufficient number of unique categories. It turned out that there are 327 unique categories which is a good number for further analysis.

3. Methodology

In this paper I will try to detect areas of Toronto which are similar to the area of my **client's office - Marble Hill**. For data clustering I will use **max number of venues equal to 100** and **radius of exploration equal to 750 meters** (as shown in the previous paragraph).

In first step I **have collected** the required data: Toronto locations with one New York locations and the list with types of available venues (with usage of Foursquare categorization).

Next, I will group rows by neighborhood and I will take the mean of the frequency of occurrence of each category in the neighborhood. This data I will use to cluster the locations. **The clustering** is the grouping of a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). In this case the similarity will be done on the venues in given locations.

In a third step I will **check the top 10 most common venues** for all locations and I will compare it with to the Marble Hill most common venues. It will allows me to get the first feeling if I can use the clustering or not.

In fourth step **I will cluster neighborhoods**: I will use **k-means clustering**. But before I will find the best K value using **the Elbow Point method**.

In the last step I will **examine each cluster** and the I will focus on cluster with Marble Hill location. I will also **visualize** the cluster where Marble Hill belongs to. The visualization should be a starting point for final decision of the client to exploration and search for optimal venue location for the new office.

3.1 Analysis of each neighborhood

At the beginning of my analysis I grouped rows by neighborhood and I took the mean of the frequency of occurrence of each category in my data frame. I did a review of each neighborhood along with the top 10 most common venues. For Marble Hill the data are as follow:

```
----Marble Hill----
venue  freq
0         Pizza Place  0.07
1          Donut Shop  0.06
2      Sandwich Place  0.06
3  Spanish Restaurant  0.06
4                Bank  0.04
5  Mexican Restaurant  0.04
6       Grocery Store  0.04
7                 Gym  0.04
8               Café  0.04
9         Coffee Shop  0.04
```

The set of the data abow I compared with the most popular venues in our main dataset.

It's worth noting that out of 10 the most common venues for Marble Hill, 5 overlap with the categories that are popular for all locations. These are (highlighted in yellow):

```
----Marble Hill----
                venue  freq
0         Pizza Place  0.07
1         Donut Shop  0.06
2      Sandwich Place  0.06
3  Spanish Restaurant  0.06
4                Bank  0.04
5  Mexican Restaurant  0.04
6       Grocery Store  0.04
7                 Gym  0.04
8               Café  0.04
9         Coffee Shop  0.04
```

Based on this analysis we can expect that the current office location (Marble Hill) will fall into a large cluster.

Here I also present the most popular categories in our data frame. To prepare it I put them into a *pandas* data frame and write a function to sort the venues in descending order.

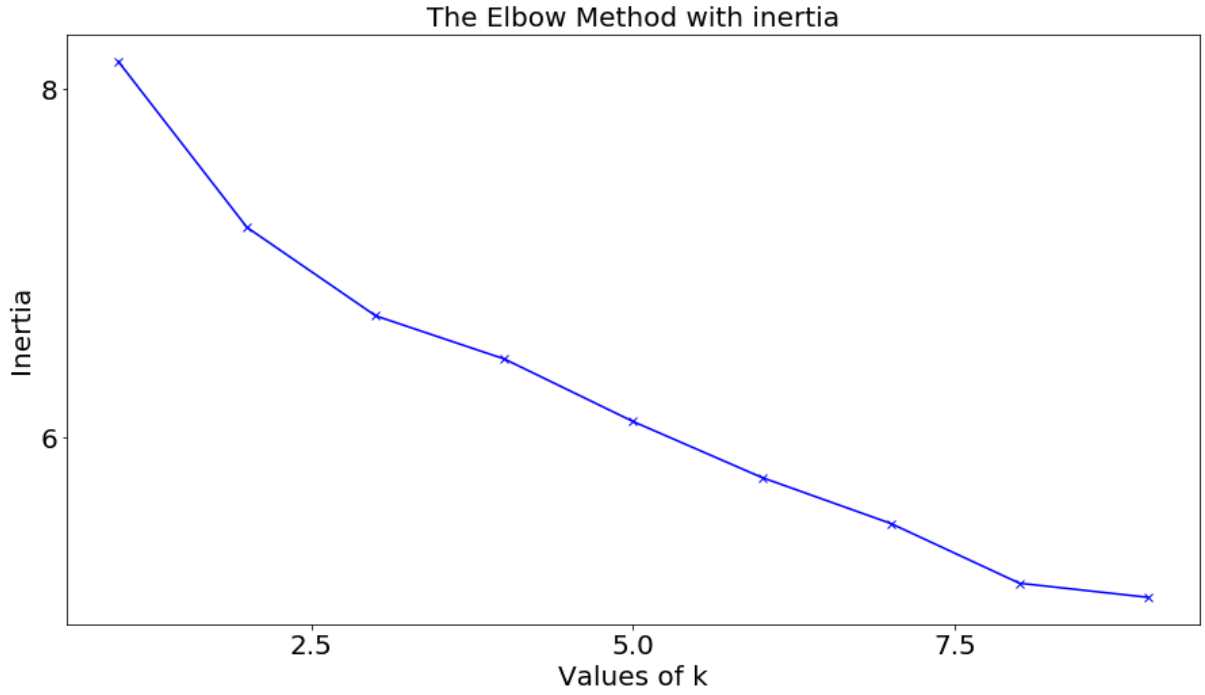
	Categories	Number of venues
0	Coffee Shop	277
1	Café	149
2	Restaurant	113
3	Pizza Place	109
4	Park	105
5	Italian Restaurant	92
6	Sandwich Place	82
7	Bakery	82
8	Japanese Restaurant	75
9	Grocery Store	65
10	Hotel	60

3.2 Cluster Neighborhoods

I use k-means clustering to find similar places to the actual office location, but first I will find the best *K* value using the **Elbow Point** method.

3.2.1 Elbow Method

From the plot below we can see that the optimum K value is equal to 8, so I will have a resulting of 8 clusters.



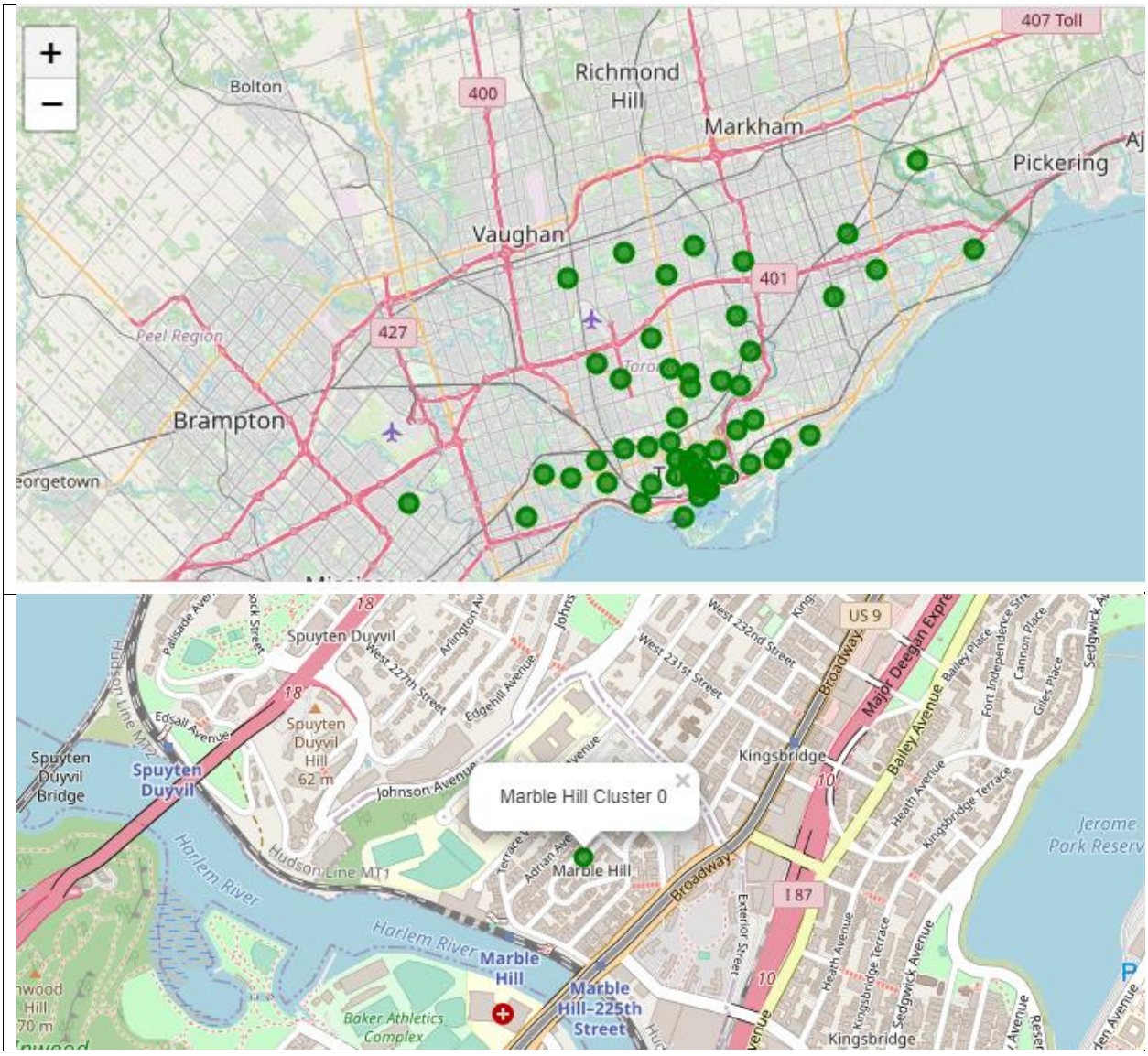
3.2.2 KMeans

Then on the basis of information gather in the point above I created a new data frame that includes the cluster as well as the top 10 venues for each neighborhood.

	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	6th Most Common Venue	7th Most Common Venue
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	3.0	Fast Food Restaurant	Trail	Spa	Martial Arts School	Paper / Office Supplies Store	Hobby Shop	Coffee Shop
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	0.0	Breakfast Spot	Burger Joint	Italian Restaurant	Bar	Yoga Studio	Eastern European Restaurant	Doner Restaurant
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	3.0	Fast Food Restaurant	Restaurant	Pizza Place	Sports Bar	Beer Store	Bank	Fried Chicken Joint
3	M1G	Scarborough	Woburn	43.770992	-79.216917	1.0	Park	Coffee Shop	Business Service	Dumpling Restaurant	Distribution Center	Dive Bar	Dog Run
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	0.0	Coffee Shop	Indian Restaurant	Yoga Studio	Pharmacy	Flower Shop	Burger Joint	Fried Chicken Joint

3.2.3 Examine Clusters

Finally, I examined and visualized the cluster where Marble Hill belongs (it is a Cluster number 0).



4. Results and Discussion

There is a great number of locations in Toronto which are similar to Marble Hill if we take availability of venues only (56 places). It seems that there are too many locations to make an unambiguous decision. In the next step, I would suggest the client to choose important venues for him or add other factors that would lead to a shortlist.

My analysis shows that elbow method was not clear in this case. Certainly, it would be useful to analyze the number of clusters once again with different method. Although I chose 8 clusters as k, from **The Elbow Method with inertia** plot both k = 3 and k = 8 formed the elbow. It is possible that other techniques for finding k would have worked better.

The analysis also shows that the venue itself does not constitute an unequivocal decision as to whether a given venue is similar to the current office location. Usually, in the vicinity of offices, there are the same attractions - so taking only them into account may not be a good location filter. There is no doubt that other factors that I mention in the next point should be taken into account.

5. Conclusion

Purpose of this paper was to identify Toronto's areas which are similar to New York location - Marble Hill in order to aid stakeholders in finding an optimal location for a new office. By joining Toronto and New York location and clustering them I choose those locations which are quite similar to the office actual location in terms of available venues.

Final decision on optimal office location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), traffic, public transport, cost of the rented space, space availability, social and economic dynamics of every neighborhood, neighborhood safety etc.

The analysis performed may be the first step in the further search for the best place, depending on stakeholders decision.

References

[1] https://en.wikipedia.org/wiki/New_York_City

[2] https://en.wikipedia.org/wiki/Toronto

[3] https://www.quora.com/How-different-or-similar-is-Toronto-to-New-York-City

[4] https://nomadlist.com/north-america