Comparing the cities of Toronto and New York regarding the quantity of Health Centers

Alejandro Somarriba

December 6, 2019

1. Introduction/Business Problem

1.1 Background

New York City, located in the United States, and the city of Toronto, located in Canada, are their respective countries' financial capitals. Both countries are very developed and are among the world leaders. However, despite how similar they may seem, they have some key differences. For one, healthcare in Canada is socialized, while healthcare in the US relies on private health insurance. Given that Toronto and New York are their countries' financial capitals, it would be interesting to compare them using the proximity of their hospitals to each of their neighborhoods.

1.2 Problem

The objective of this project was to visually compare the way neighborhoods are spread throughout both cities, as well as determining which city had more hospitals closer to its neighborhoods. I chose to find out how many hospitals were located within 1 kilometer of the cities' neighborhoods. Then, based on the data I found, I would be able to determine which city is more prepared to provide healthcare services. Furthermore, the project also has the intention of categorizing the neighborhoods based on the most common hospital-related venue by category, such as Hospital, Medical Center, Hospital Ward, Emergency Room, and Urgent Care Center.

1.3 Audience

This project may be of interest to someone who lives on either city or intends to move to one of them, as they may want to see which healthcare options are closest to where they are. It may also be of interest to healthcare providers who may want to set up their own services close to neighborhoods where other venues are too far away.

2. Data

2.1 Data Sources

Parts of the data used in this project are lists of neighborhoods for Toronto and New York. The list used for the Toronto neighborhoods was scraped from a table from

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, which lists all of the neighborhoods in Toronto sorted by their postal codes. This list was cleaned to eliminate rows where values were not assigned, and then complemented by getting the coordinates of the rest of the neighborhoods. The list used for the New York neighborhoods was obtained from https://cocl.us/new_york_dataset, which is a JSON file that contains a list of New York neighborhoods, as well as their coordinates. The rest of the data used is mostly location data provided by the FourSquare API in order to find the closest hospital-related venues to the neighborhoods.

All this data will help me make markers on maps using the Folium Python library, which will allow me to see the distribution of venues for both cities. It will also help me when I use a clustering algorithm to separate the neighborhoods into distinct classes.

2.2 Data Cleaning

In order to clean the data, I first got rid of the useless rows from the Toronto neighborhoods list. Then, I used the GeoPy Python library to obtain the coordinates for all the neighborhoods in Toronto; this required having to eliminate some rows which yielded no results, as well as renaming other rows to get them. I also got rid of the 'Postal Code' and 'Borough' columns, as I only required the 'Neighborhood' column. Then, I cleaned the New York neighborhoods list by getting rid of the 'Borough' column, as I would only need the 'Neighborhood', 'Latitude', and 'Longitude' columns.

More data cleaning occurred further along in the process, once I had found the nearby venues I was looking for.

2.3 Data Examples

For example, a row from the Toronto neighborhoods table would have the neighborhood's name, latitude, and longitude, e.g.:

Neighborhood	Latitude	Longitude
Parkwoods	43.7588	-79.3202

A row from the New York neighborhoods table would be similar to the one from the Toronto neighborhoods table, with the neighborhood's name, latitude, and longitude, e.g.:

Neighborhood	Latitude	Longitude
Wakefield	40.894705	-73.847201

An example of the rows obtained from the FourSquare API would have the neighborhood's name, neighborhood's latitude, neighborhood's longitude, venue's name, venue's latitude, venue's longitude, and venue's category, e.g.:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
South of Bloor	43.667662	-79.394698	CAMH Russell Site	43.659131	-79.399194	Medical Center

3. Methodology

3.1 Exploring the datasets and plotting maps

First, I had the neighborhood lists that had the Neighborhood, Latitude, and Longitude for the all the neighborhoods in Toronto and New York. I had already cleaned the data frames so they didn't have rows that didn't have coordinates. The data frame that had the information about the Toronto neighborhoods had 208 rows, which meant that there were at least 208 unique neighborhoods in Toronto whose coordinates could be obtained using the GeoPy Python library. On the other hand, the data frame that had the information for the neighborhoods in New York had 306 rows, which meant that New York had 306 neighborhoods.

Once, I knew how many neighborhoods there were in each city, I decided to plot them on a map to see how they were distributed. This was the result:

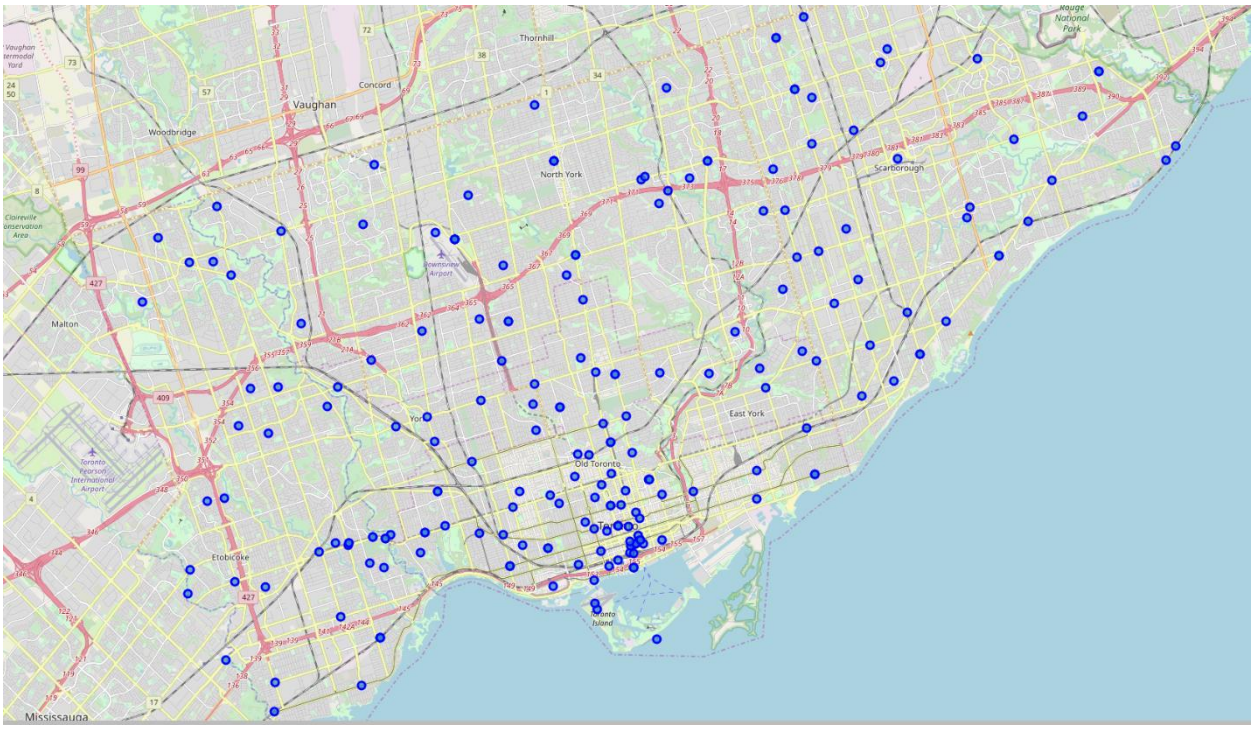

Image 1.1: A map of Toronto with markers indicating the location of its neighborhoods

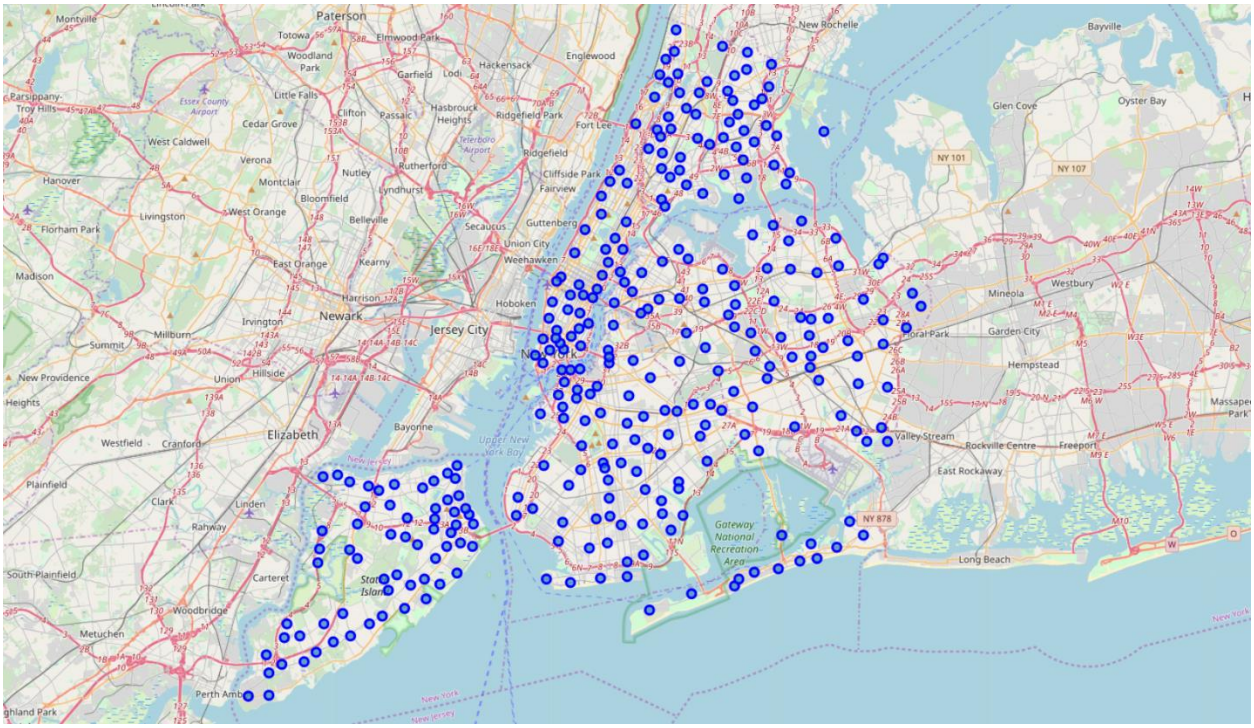

Image 1.2: A map of New York City with markers indicating the location of its neighborhoods

With these images, it is possible to observe that the neighborhoods in Toronto are more spread out than the ones in New York City. This is interesting, since Toronto has an area of 630 km^2, while New York City has a land area of 784 km^2. This makes sense, as even though New York City is almost 1.25 times larger than Toronto, it has nearly 1.5 times the number of neighborhoods.

3.2 Getting the hospital-related venues for each neighborhood

After finishing with the data frames, I proceeded by getting the hospital venues using the FourSquare API. I used the "search" endpoint of the API in order to get a list of all the venues it could find. I set the result to limit to 300 because I thought it was a big enough number that it would get all the venues. I also specified the radius of the search to be 1,000 meters (1 km), so I would only get results for venues that were located within 1 km of the specified location. Among the parameters I specified in the request URL, I added one called "categoryId", which narrows results matching the specific category of venue that is specified. In my case, I used the ID 4bf58dd8d48988d196941735, which corresponded to hospitals according to the FourSquare documentation.

Using the API, I iterated through the coordinates of every neighborhood in Toronto and New York, and then sorted the data into data frames that contained the neighborhood's name, its coordinates, the name of the venue, the venue's coordinates, and the venue's category (see example in Section 2.3). I ended up with a data frame for all the hospital-related venues in Toronto and another one for New York City. The Toronto data frame had 1289 rows, and the New York City data frame had 2120 rows.

Before proceeding with anything else, I wanted to see what kinds of venues had been retrieved, as I noticed that not all of them were hospitals. I ran the .value_counts() method to see all the different venues for both cities. They returned this:

Toronto

Venue	Count
Hospital	1139
Hospital Ward	63
Conference Room	36
Medical Center	18
Emergency Room	18
Veterinarian	9
Building	4
Medical Lab	1
Doctor’s Office	1

New York

Venue	Count
Hospital	1709
Hospital Ward	163
Doctor’s Office	114
Medical Center	46
Office	20
Emergency Room	10
Pharmacy	7
Optical Shop	7
Medical School	7
Government Building	6
Eye Doctor	5
Medical Lab	3
College Science Building	3
Veterinarian	3
Auditorium	3
Spiritual Center	3
Building	3
Urgent Care Center	2
Bus Station	2
Scenic Lookout	2
Dentist’s Office	1
High School	1

I realized that the API had obtained venues whose category wasn’t Hospital, so I had to clean it. I iterated through the data frames and got rid of venues that I thought weren’t sufficiently related to providing general healthcare. I ended up with this:

Toronto

Venue	Count
Hospital	1139
Hospital Ward	63
Medical Center	18
Emergency Room	18

New York

Venue	Count
Hospital	1709
Hospital Ward	163
Medical Center	46
Emergency Room	10
Urgent Care Center	2

Once the data frames were clean, I realized that it was possible that a single neighborhood could have more than 1 hospital-related venue within 1 km, so I ran another method to figure out how many unique neighborhoods matched the criteria for the API. It turned out that Toronto had 94

neighborhoods that had at least 1 hospital-related venue within 1 km, while New York City had 232. As it stood, New York City had almost 2.5 times more neighborhoods than Toronto that had at least 1 hospital-related venue within 1 km.

Similarly, I realized that it was possible for there to be an overlap between venues in different neighborhoods, meaning that a single venue could be within 1 km of more than 1 neighborhood, which would mean that I could have duplicated venues. Because of this, I wanted to figure out how many unique hospital-related venues I had retrieved, so I did something like what I did to get the unique neighborhoods. After running a couple functions, I learned that Toronto had 167 unique venues within 1 km of at least 1 of its neighborhoods, and that New York City had 862 unique venues. Therefore, New York City had just a bit over 5 times more hospital-related venues within 1 km of a neighborhood than Toronto.

3.3 More maps

With the new data I had found, I was able to make two new maps with more markers. I made maps for both cities that had green markers for the positions of all the neighborhoods that had at least 1 hospital-related venue within 1 km, blue markers for the positions of all the other neighborhoods, and smaller red markers for the positions of all the venues. This was the result:

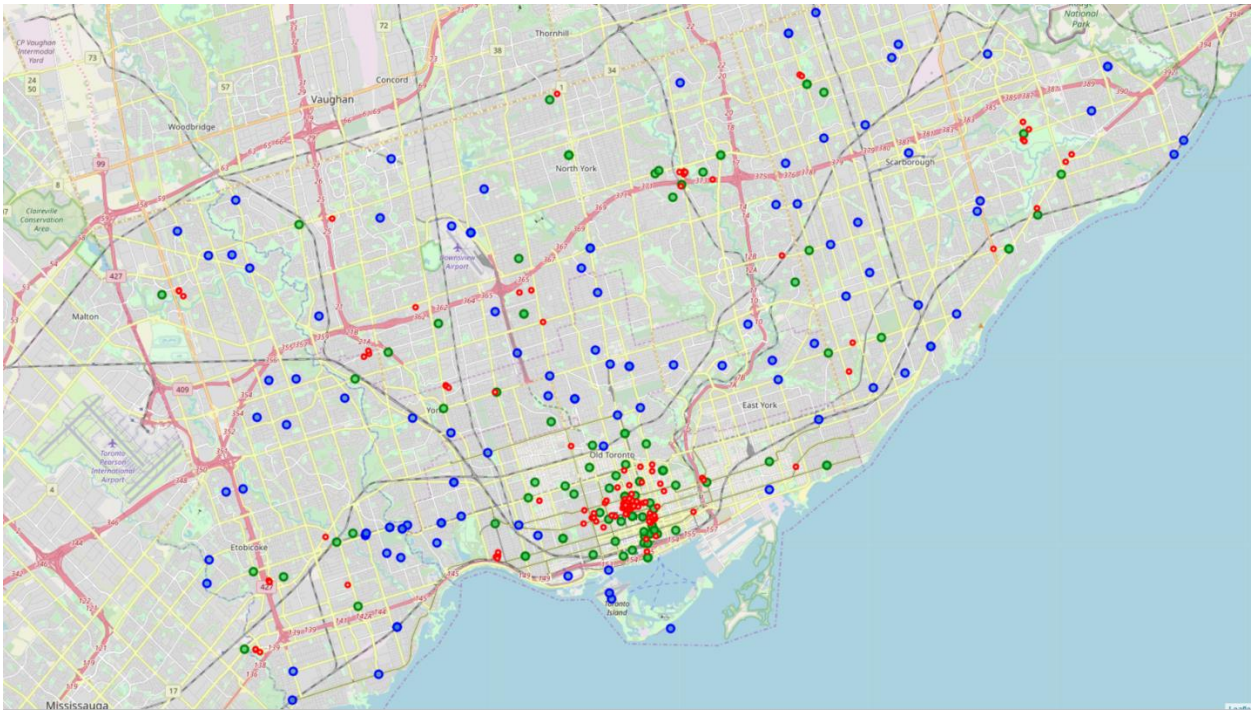

Image 2.1: A map of Toronto with markers indicating the location of its neighborhoods, as well as markers indicating the location of the venues retrieved by the API

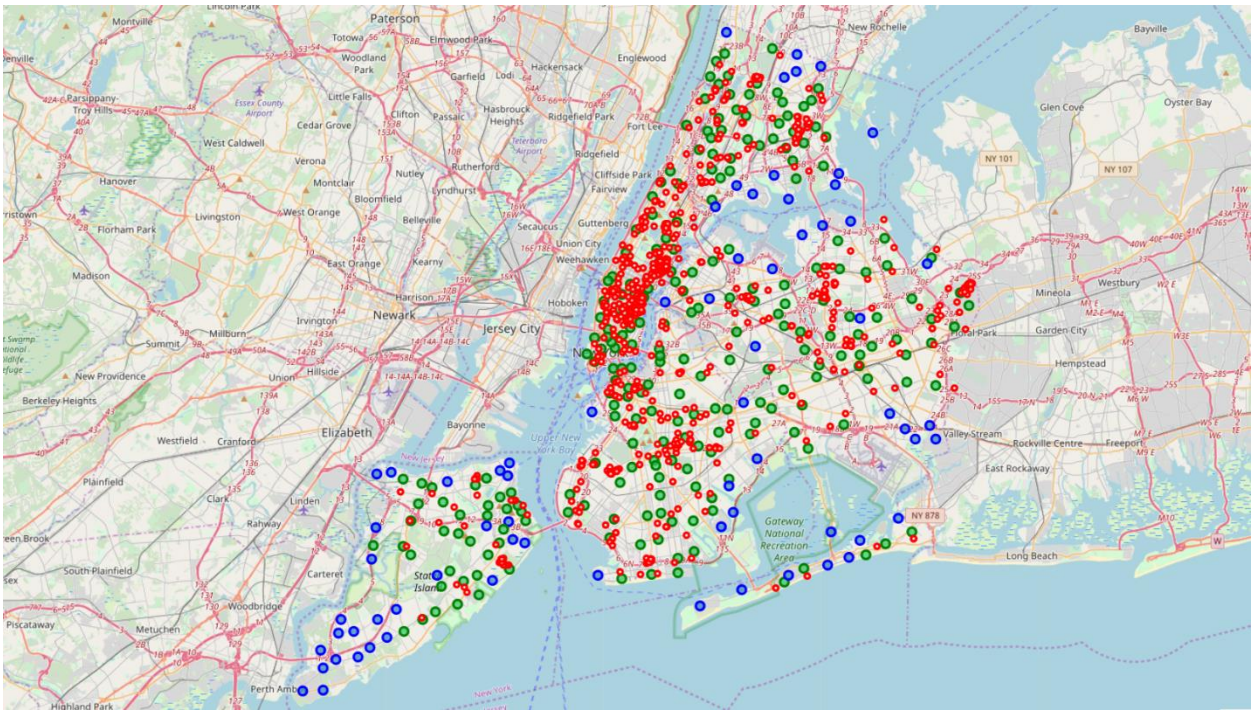

Image 2.2: A map of New York City with markers indicating the location of its neighborhoods, as well as markers indicating the location of the venues retrieved by the API

From these new maps, it's possible to observe that most of the venues in Toronto are gathered near the center while the rest are spread far apart. On the other hand, while there are some venues in New York City that also seem to gather around the middle, the rest appear to be a bit more evenly scattered along the map. The ratio between neighborhoods that have a venue within 1 kilometer and neighborhoods that don't is also more noticeable between the two cities; New York City has a higher ratio than Toronto.

3.4 One-hot encoding and most common venues

After figuring out which were the unique neighborhoods and venues for both cities, and mapping them, I decided it was time to begin with the one-hot encoding. I did this for each venue category so I could then determine the mean occurrence of the venue categories in each neighborhood (for each city). This would help me determine which were the most common venues for each neighborhood later. Eventually, I had two more data frames that contained the neighborhoods and the 4 most common types of venue (5 in the case of New York City).

3.5 KMeans Clustering

I decided to use the KMeans clustering algorithm to see if some pattern emerged from the way the venues were distributed in the maps. I did 5 clusters as a test, and just kept those results because they seemed appropriate. Afterwards, I merged the data frames that contained the cluster labels with the data frames that had the neighborhoods and their coordinates so I could generate a final set of maps:

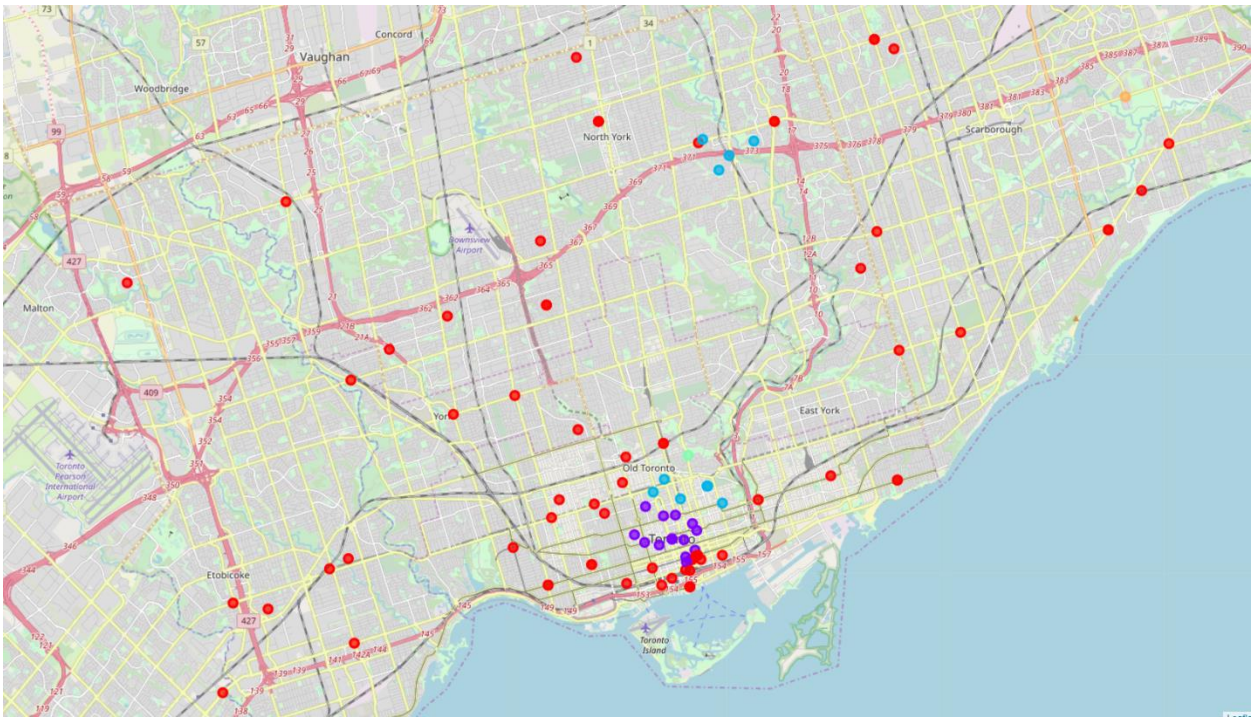

Image 3.1: A map of Toronto with markers indicating the location of the neighborhoods that had hospital-related venues within 1 km, color-coded to distinguish clusters

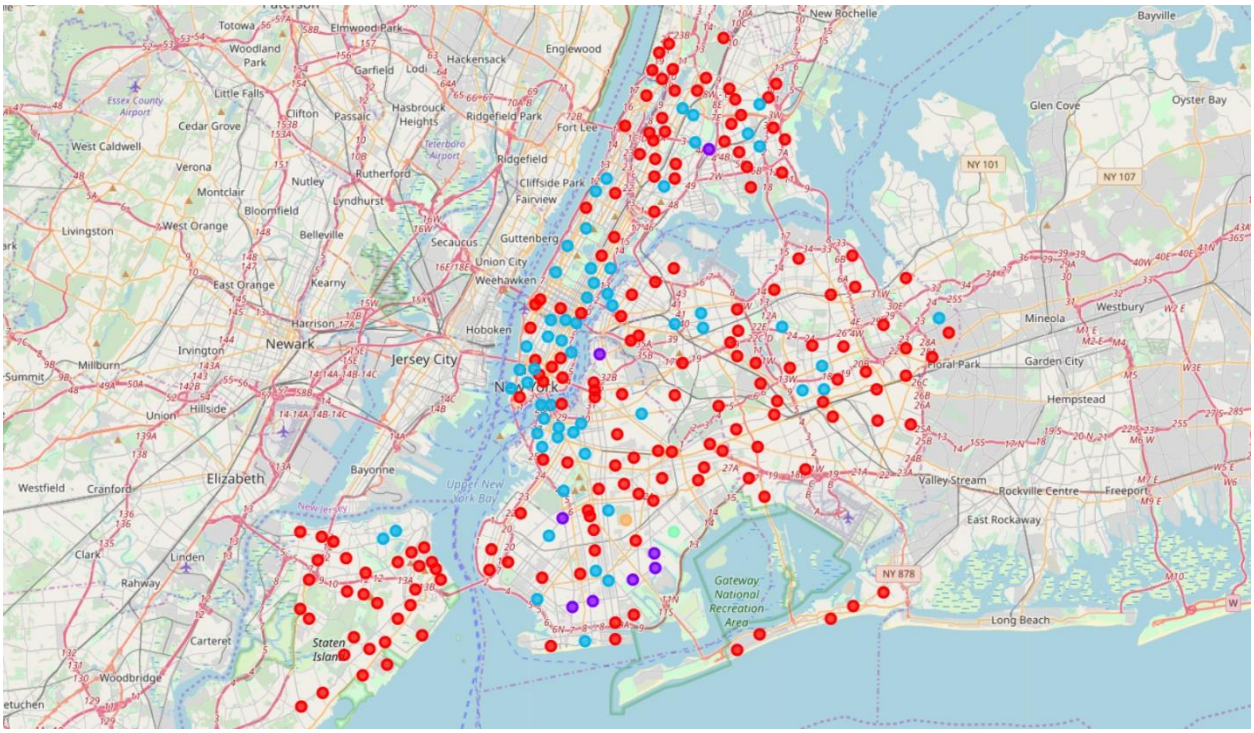

Image 3.2: A map of New York City with markers indicating the location of the neighborhoods that had hospital-related venues within 1 km, color-coded to distinguish clusters

With these visual representations of how the clusters classified the neighborhoods, it's possible to observe that the red cluster appears to be the majority and the most spread out of the neighborhoods in both cities. The blue and purple clusters in the Toronto map appear to be more close to the center than the rest; however, in the New York City map, the blue cluster seems to have more neighborhoods along the middle and top of the map, with the purple cluster being generally more spread out. Finally, it seems that the orange and green clusters for both cities were some sort of outlier in both cases.

After those maps, I selected the data frames that had the neighborhoods displayed for each different cluster. This will be discussed in the next section.

4. Results

4.1 Raw Analysis of the Clusters

From the last data frames, it was possible to select the portions that corresponded to each of the clusters, so I could try to find the pattern that made each cluster unique.

For the Toronto cluster, I managed to surmise the following:

- Cluster 1 (red): had neighborhoods whose most common venues, in order, were:
    - Hospital, Medical Center, Hospital Ward, Emergency Room
- Cluster 2 (purple): had neighborhoods whose most common venues in order, were:
    - Hospital, Hospital Ward, Medical Center, Emergency Room
- Cluster 3 (blue): had neighborhoods whose most common venues in order, were:
    - Hospital, Hospital Ward, Medical Center, Emergency Room
- Cluster 4 (green): had neighborhoods whose most common venues in order, were:
    - Hospital, Hospital Ward, Medical Center, Emergency Room
- Cluster 5 (orange): had neighborhoods whose most common venues in order, were:
    - Hospital, Emergency Room, Medical Center, Hospital Ward

For the New York City cluster, I managed to surmise the following:

- Cluster 1 (red): had neighborhoods whose most common venues, in order, were:
    - Hospital, Urgent Care Center, Medical Center, Hospital Ward, Emergency Room
- Cluster 2 (purple): had neighborhoods whose most common venues, in order, were:
    - Hospital, Hospital Ward, Urgent Care Center, Medical Center, Emergency Room
- Cluster 3 (blue): had neighborhoods whose most common venues, in order, were:
    - Hospital, Hospital Ward, Urgent Care Center, Medical Center, Emergency Room
- Cluster 4 (green): had neighborhoods whose most common venues, in order, were:
    - Medical Center, Urgent Care Center, Hospital Ward, Hospital, Emergency Room
- Cluster 5 (orange): had neighborhoods whose most common venues, in order, were:
    - Hospital Ward, Urgent Care Center, Medical Center, Hospital, Emergency Room

5. Discussion

5.1 Result discussion

After scanning more carefully through the result from the KMeans clustering algorithm, I realized that it was probably possible to have used less clusters to separate the locations, as there were many overlaps in the most common venues between clusters, such as clusters 2, 3, and 4 for Toronto and clusters 2 and 3 for New York City. Other than that, I believe the results that were yielded were quite interesting.

If this project were to be repeated, I would suggest trying to find more sources for the neighborhoods, just make sure there are none missing and so all the coordinates can be retrieved. Also, it might be interesting to take the project further and consider more data points to make more interesting inferences and draw more useful conclusions.

5.2 General observations

Overall, Toronto has less neighborhoods and hospital-related venues than New York.

New York has Urgent Care Centers within 1 kilometer of some of its neighborhoods, while Toronto does not.

New York has its neighborhoods closer together than Toronto.

Looking at the clusters, both the majority of Toronto and New York's neighborhoods most common venues in order are: Hospitals, Urgent Care Centers (New York only), Medical Centers, Hospital Wards, and Emergency Rooms.

5.3 Map observations

Neighborhoods in Toronto are more spread out than the ones in New York City (Images 1.1 and 1.2).

Most of the venues in Toronto are gathered near the center while the rest are spread far apart. On the other hand, while there are some venues in New York City that also seem to gather around the middle, the rest appear to be a bit more evenly scattered along the map. The ratio between neighborhoods that have a venue within 1 kilometer and neighborhoods that don’t is also more noticeable between the two cities; New York City has a higher ratio than Toronto (Images 2.1 and 2.2).

The red cluster appears to be the majority and the most spread out of the neighborhoods in both cities. The blue and purple clusters in the Toronto map appear to be more close to the center than the rest; however, in the New York City map, the blue cluster seems to have more neighborhoods along the middle and top of the map, with the purple cluster being generally more spread out. Finally, it seems that the orange and green clusters for both cities were some sort of outlier in both cases (Images 3.1 and 3.2).

6. Conclusions

Overall, it would seem that New York has more hospital-related venues within 1 kilometer of its neighborhoods, which would suggest that it is better equipped to handle health care.

Please note that this implication is formed solely based on the data compiled in this notebook, and does not take into account other measures such as population, medical insurance policies, quality of health care, or otherwise; for this reason, the implication that New York is more prepared than Toronto regarding health care may not be completely accurate.

The purpose of this project was simply to visually compare the two cities based on how many hospitals their neighborhoods have nearby, and to see which venues were more common. In that regard, the project was successful in revealing how differently distributed are the hospital-related venues in Toronto and New York City.

Finally, the project was able to generate clear enough clusters to distinguish different neighborhoods based on the most common types of venues that were located within 1 kilometer of them.