IBM Data Science Professional Certificate

Capstone Project – Ajay Bhatnagar

Battle of the Neighborhoods

Comparing and Clustering Neighborhoods of New York City and Toronto

Jun 2020

1. Introduction / Business Problem

1.1 Introduction

New York City and Toronto are both very densely populated and diverse cities which are also the financial capitals of their respective countries. New York City is located in the state of New York of the United States of America while Toronto is located in the Ontario province of Canada.

New York City (NYC), often called New York (NY), is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined statistical area, it is one of the world's most populous mega-cities. 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. Home to the headquarters of the United Nations, New York is an important center for international diplomacy.

Situated on one of the world's largest natural harbors, New York City is composed of five boroughs, each of which is a county of the State of New York. The five boroughs—Brooklyn, Queens, Manhattan, the Bronx, and Staten Island—were consolidated into a single city in 1898. The city and its metropolitan area constitute the premier gateway for legal immigration to the United States. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. New York is home to more than 3.2 million residents born outside the United States, the largest foreign-born population of any city in the world as of 2016. As of 2019, the New York metropolitan area is estimated to produce a gross metropolitan product (GMP) of $2.0 trillion. If the New York metropolitan area were a sovereign state, it would have the eighth-largest economy in the world. New York is home to the highest number of billionaires of any city in the world.

Toronto is the provincial capital of Ontario. 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. The Greater Toronto Area (GTA) as a whole had a 2016 population of 6,417,516. The city covers an area of 630.20 square kilometers (243.32 sq mi) and comprises six districts – East York, Etobicoke, North York, Old Toronto, Scarborough

and York – which were amalgamated to form Toronto's present boundaries in 1998. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. While the majority of Torontonians speak English as their primary language, over 160 languages are spoken in the city.

As can be seen above, both of these cities have an impressive set of characteristics that appear to be very similar to each other so let us begin the “Battle of the Neighborhoods” and determine if our analysis and clustering of neighborhoods of these two cities provides reasonable justifications and confirmation for these similarities and that we are also able to outline what, if any, the dissimilarities are.

1.2 Problem Statement

It would be interesting to **compare and contrast the neighborhoods of New York City and Toronto**.

Specifically, we would like to find out answers to questions such as:

- What are the most common venue categories in each city?
- What venue categories are most widespread in each city?
- How many neighborhoods of these two cities are similar to each other?
- How many neighborhoods of these two cities are quite unique to themselves?
- Are these two cities very similar to each other or very different?

1.3 Audience and Interest

People who are considering moving from either New York City to Toronto or vice-versa will be very interested in this comparison to be able to find neighborhoods in the city they want to move to, similar to where they are living in their current city or aspiring to get to in the other city.

This type of comparison would also be of interest to **planners and officials of these cities and their neighborhoods** to be able to look at data trends and decide which types of businesses to promote based on what appears to be working in the similar neighborhoods of the other city.

2. Data Acquisition and Wrangling

In order to achieve the project objectives, we will need two major sets of data that will need to be acquired, cleaned, augmented/supplemented, mapped, transformed and prepared for each of the stages of our data analysis methodology. In broad terms, these activities are referred to as Data Wrangling or Data Munging. The two required datasets are Neighborhood Data and Venues Data. These are explained below.

2.1 Neighborhood Data

This is a dataset of the neighborhoods of New York City and Toronto with their Latitude and Longitude coordinates. In the case of New York City, the dataset already exists in a readily programmatic consumable format like a structured file and will just need to be cleaned, mapped and transformed while in other case for Toronto, we need to scrape portion of the required data from a website and then augment it with data from other sources before cleaning, mapping and transforming. We will also visualize the neighborhoods of each city on a map.

2.1.1 New York City

A dataset has been made available as a JSON file [1] that contains the following details about each neighborhood of New York City:

Name, Borough, Coordinates as Latitude and Longitude and some more data

A sample neighborhood from the file is shown in Figure 1 below

{'type': 'Feature',

'id': 'nyu_2451_34572.1',

'geometry': {'type': 'Point',

'coordinates': [-73.84720052054902, 40.89470517661]},

'geometry_name': 'geom',

'properties': {'name': 'Wakefield',

'stacked': 1,

'annoline1': 'Wakefield',

'annoline2': None,

'annoline3': None,

'annoangle': 0.0,

'borough': 'Bronx',

'bbox': [-73.84720052054902,

40.89470517661,

-73.84720052054902,

40.89470517661]}}

Figure 1: Part of JSON file showing one Neighborhood

The contents of this JSON file will be read and transformed into a dataframe as shown in Figure 2 below. This dataframe contains 306 Neighborhoods of New York City in the 5 Boroughs.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Figure 2: New York City Neighborhoods

We then check to see if there are any duplicate Neighborhood names. There are 4 Neighborhoods (Sunnyside, Bay Terrace, Murray Hill & Chelsea) with the same name that exist in multiple Boroughs. We will update the Neighborhood name to add the Borough name as a suffix. Sample updates for Chelsea are shown in Figure 3 below.

Before:

	Borough	Neighborhood	Latitude	Longitude
116	Manhattan	Chelsea	40.744035	-74.003116
244	Staten Island	Chelsea	40.594726	-74.189560

After:

	Borough	Neighborhood	Latitude	Longitude
116	Manhattan	Chelsea, Manhattan	40.744035	-74.003116
244	Staten Island	Chelsea, Staten Island	40.594726	-74.189560

Figure 3: New York City - Unique Neighborhood Names

We then prefix all New York City Neighborhoods with "NYC_" to ensure that further analysis tasks that require us to combine New York City neighborhoods with Toronto neighborhoods can be merged without any ambiguity and still preserve their relation to New York City. The first 5 rows of the final dataframe **nyc_neighborhoods** are shown in Figure 4 below. **There are 306 neighborhoods in this final dataframe**.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	NYC_Wakefield	40.894705	-73.847201
1	Bronx	NYC_Co-op City	40.874294	-73.829939
2	Bronx	NYC_Eastchester	40.887556	-73.827806
3	Bronx	NYC_Fieldston	40.895437	-73.905643
4	Bronx	NYC_Riverdale	40.890834	-73.912585

Figure 4: New York City Neighborhoods – final (nyc_neighborhoods)

2.1.2 Toronto

There is no readily available location data for all neighborhoods of Toronto. There is a Wikipedia page titled “List of postal codes of Canada: M” [2] that provides a list of Postal Code prefixes, Borough and Neighborhood names of Toronto. We will scrape this data from the wiki page and create a dataframe as shown in Figure 5 below. This dataframe contains 180 Neighborhoods of Toronto in 11 Boroughs.

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 5: Toronto Neighborhoods

On closer inspection, it is seen that each row represents a unique value of the Postal Code prefix with the following additional caveats:

- A Borough may contain more than one Postal Code but these are represented as separate rows.
- A Borough and Postal Code prefix combination may contain more than one Neighborhood and these are represented as a comma-separated list of neighborhoods in the Neighborhood column.
- There are 77 records where the Borough is “Not assigned” - these need to be dropped from the data set and further analysis as they do not have any neighborhoods assigned to them as indicated by the Neighborhood being “Not assigned” value.
- Finally, any records that still have Neighborhood with “Not assigned” value, they should be updated with their corresponding Borough name. In our final dataframe, there are no such records remaining that will require this update.

The final dataframe after all the above wrangling activities are completed contains 103 neighborhoods. The first 5 rows of this dataframe are shown in Figure 6 below.

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 6: Toronto Neighborhoods – after wrangling

We then check to see if there are any duplicate Neighborhood names. There are 2 Neighborhoods (Downsview & Don Mills) with the same name that exist in multiple Postal Code prefixes. We will update the Neighborhood name to add the Postal Code prefix name as a suffix. Sample updates for Downsview are shown in Figure 7 below.

Before:

	PostalCode	Borough	Neighborhood
40	M3K	North York	Downsview
46	M3L	North York	Downsview
53	M3M	North York	Downsview
60	M3N	North York	Downsview

After:

	PostalCode	Borough	Neighborhood
40	M3K	North York	Downsview, M3K
46	M3L	North York	Downsview, M3L
53	M3M	North York	Downsview, M3M
60	M3N	North York	Downsview, M3N

Figure 7: Toronto Neighborhoods – Unique Neighborhood Names

The next wrangling step requires us to find a way to obtain Latitude and Longitude coordinates for each of the Toronto neighborhoods. We will use the geocoder package with the ArcGIS provider (Google doesn't work) to obtain these coordinates and add to our dataframe. The first 5 rows of the updated dataframe is as shown below in Figure 8.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.752935	-79.335641
1	M4A	North York	Victoria Village	43.728102	-79.311890
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.650964	-79.353041
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.723265	-79.451211
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.661790	-79.389390

Figure 8: Toronto Neighborhoods – dataframe after adding coordinates

We will now remove the PostalCode column from the dataframe since it is no longer of relevance. Finally, we then prefix all Toronto Neighborhoods with "YYZ_" to ensure that further analysis tasks that require us to combine New York City neighborhoods with Toronto neighborhoods can be merged without any ambiguity and still preserve their relation to Toronto. YYZ was selected since it is the airport code of the Toronto International Airport. The first 5 rows of the final dataframe **toronto_neighborhoods** are shown in Figure 9 below. **There are 103 neighborhoods in this final dataframe**.

	Borough	Neighborhood	Latitude	Longitude
0	North York	YYZ_Parkwoods	43.752935	-79.335641
1	North York	YYZ_Victoria Village	43.728102	-79.311890
2	Downtown Toronto	YYZ_Regent Park, Harbourfront	43.650964	-79.353041
3	North York	YYZ_Lawrence Manor, Lawrence Heights	43.723265	-79.451211
4	Downtown Toronto	YYZ_Queen's Park, Ontario Provincial Government	43.661790	-79.389390

Figure 9: Toronto Neighborhoods – final (toronto_neighborhoods)

2.2 Venues Data

This is a dataset of the top 100 venues within a 500 meter radius of each neighborhood's Latitude and Longitude coordinates. Ideally, this dataset should list the Venue Name and Venue Category at the very least and could also contain the Venue Latitude and Venue Longitude for more detailed analysis. This data will be obtained from Foursquare, which is the most trusted, independent location data platform for understanding how people move through the real world. We will use the Foursquare API to get the Venues and their Categories in each neighborhood for the data analysis and clustering.

Foursquare defines a three-level hierarchical structure of categories for venues. At this time, there are **ten** first level categories: **Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service and Travel & Transport**. There are a total of **459 second level categories** and **370 third level categories**. The Figure 10 below shows the first level categories and the number of second level categories under each of them.

	Level1	Level2 Count
0	Arts & Entertainment	36
1	College & University	23
2	Event	12
3	Food	92
4	Nightlife Spot	7
5	Outdoors & Recreation	62
6	Professional & Other Places	43
7	Residence	5
8	Shop & Service	145
9	Travel & Transport	34

Figure 10: Foursquare Venue Categories

To retrieve the venues data, we need to create and submit a URL is follows:

https://api.foursquare.com/v2/venues/search?&client_id=9999&client_secret=9999&v=YYYYMMDD&ll=40.89470517661,-73.84720052054902&radius=500&limit=100

where,

search is the Foursquare API endpoint being called

client_id & **client_secret** are the developer credentials used to access the API

v is the API version to be used

ll is the Latitude and Longitude of the specified location around which to get venues

radius is the maximum distance between the specified location and the venues

limit is the maximum number of venues to be retrieved

2.2.1 New York City

We use the above API call for each of New York City Neighborhood's Latitude and Longitude coordinates to retrieve a list of venues, venue category, venue latitude and venue longitude. This is then formatted into a dataframe. In case of New York City, we get 26, 504 Venues for the 306 Neighborhoods that had 579 unique Venue Categories. **We determined that there was one venue category called "Building" that was too generic to play any role in the clustering decision-making and hence we decided to drop all venues of this venue type. The final venues dataset contained 25,659 Venues across the 306 Neighborhoods that had 578 unique Venue Categories**. The first 5 rows of the final dataframe **nyc_venues** is shown below in Figure 11.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NYC_Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
1	NYC_Wakefield	40.894705	-73.847201	Pitman Deli	40.896744	-73.844398	Food
2	NYC_Wakefield	40.894705	-73.847201	Julio C Barber Shop 2	40.892648	-73.855725	Salon / Barbershop
3	NYC_Wakefield	40.894705	-73.847201	Pittman Ave bodega	40.896744	-73.844398	Convenience Store
4	NYC_Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop

Figure 11: Venues for New York City - final (nyc_venues)

2.2.2 Toronto

We use the above API call for each of Toronto Neighborhood's Latitude and Longitude coordinates to retrieve a list of venues, venue category, venue latitude and venue longitude. This is then formatted into a dataframe. In case of Toronto, we get 8, 761 Venues for the 103 Neighborhoods that had 502 unique Venue Categories. **We determined that there was one venue category called "Building" that was too generic to play any role in the clustering decision-making and hence we decided to drop all venues of this venue type. The final venues dataset contained 8,500 Venues across the 103 Neighborhoods that had 501 unique Venue Categories.** The first 5 rows of the final dataframe **toronto_venues** is shown below in Figure 12.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	YYZ_Parkwoods	43.752935	-79.335641	Church Of Our Saviour	43.751496	-79.337078	Church
1	YYZ_Parkwoods	43.752935	-79.335641	Three Valleys Public School	43.750595	-79.337341	School
2	YYZ_Parkwoods	43.752935	-79.335641	GTA Restoration | Emergency Water Damage Plumb...	43.753567	-79.351308	Construction & Landscaping
3	YYZ_Parkwoods	43.752935	-79.335641	Mo's Ride	43.755123	-79.334583	General Travel
4	YYZ_Parkwoods	43.752935	-79.335641	Bruno's Fine Foods	43.745608	-79.336772	Grocery Store

Figure 12: Venues for Toronto - final (toronto_venues)

3. Methodology

The objective of this project is to compare and contrast the neighborhoods of New York City and Toronto.

In previous step we have collected all the required data:

- **Neighborhood Data** for New York & Toronto using JSON file, web scraping a wiki page & latitude/longitude coordinates using geocoder with ArcGIS provider.
- **Venues Data** for New York City and Toronto using Foursquare API and removing venues of any categories not relevant to our analysis.
- Details about **Venue Categories** used by Foursquare to categorize Venues - these categories will be used as the features for our clustering analysis.

In the next step, we will begin our analysis of the neighborhoods of New York City and Toronto by first **visualizing the neighborhoods of each city**, exploring venues of each city by finding out the **total number of distinct venue categories** in each city, **most common and most widespread venue categories** across the neighborhoods of each city. We will then **prepare the data needed for the clustering algorithm** and we will also generate the **top 10 most common categories in each neighborhood**.

In the final step we will carry out the clustering of all neighborhoods of both cities by first combining all relevant datasets for both cities into one. In order to use **K-Means Clustering**, we will generate a dataset of all features (all venue categories) to do the clustering on and we will use the **Elbow Method** as well as the **Silhouette Score Method** to determine the **optimal number of clusters for our dataset**. We will then run the clustering on the optimal number of clusters. Finally, we will then do **analysis of each cluster to name and explain them** so that they can be used appropriately in our Results and Discussion section to meet the project objective.

4. Exploratory Data Analysis

In this step, we will begin our analysis of the neighborhoods of New York City and Toronto by first **visualizing the neighborhoods of each city**, exploring venues of each city by finding out the **total number of distinct venue categories** in each city, **most common and most widespread venue categories** across the neighborhoods of each city. We will then **prepare the data needed for the clustering algorithm** and we will also generate the **top 10 most common categories in each neighborhood**.

4.1 Visualize New York City and Toronto Neighborhoods

4.1.1 New York City

We then plot all the neighborhoods of New York City from **nyc_neighborhood** dataframe on a map using the Folium package as shown in Figure 13 below. Each neighborhood is shown on the map with a blue circle. We will use this blue color to represent all data related to New York City in this project report.

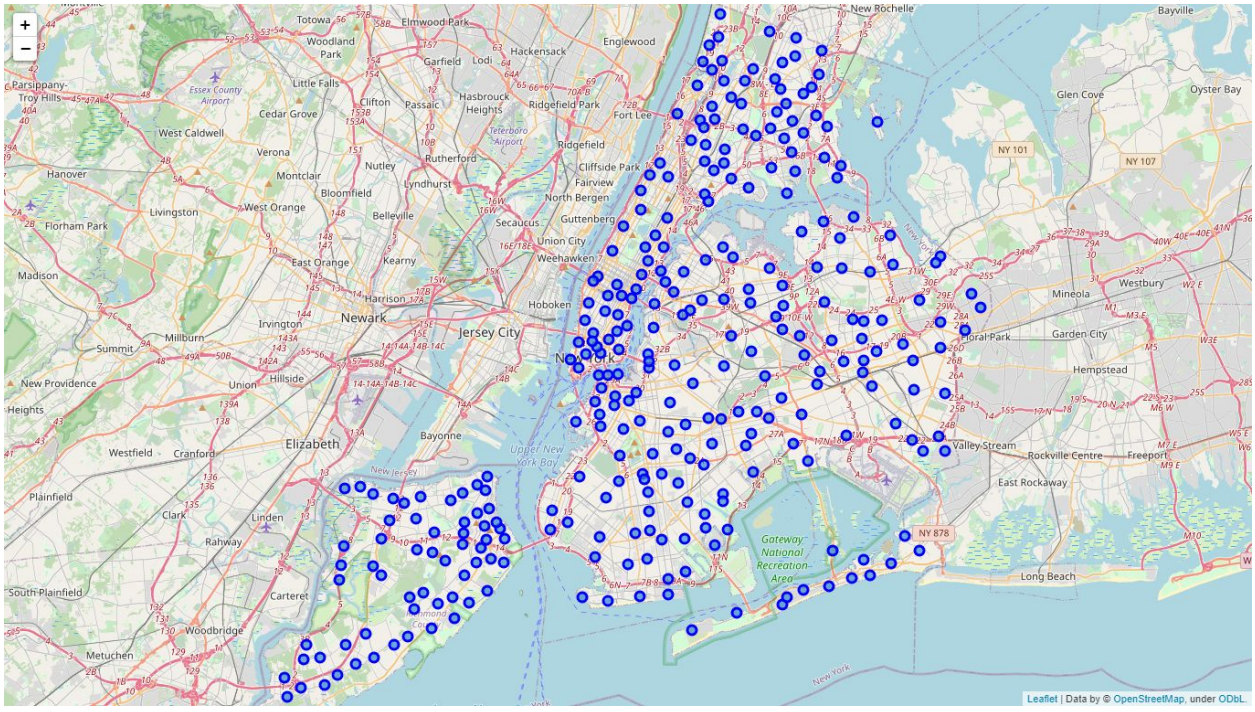

Figure 13: Map of New York City and its Neighborhoods

4.1.2 Toronto

We then plot all the neighborhoods of Toronto from **toronto_neighborhoods** dataframe on a map using the Folium package as shown in Figure 14 below. Each neighborhood is shown on the map with a green circle. We will use this green color to represent all data related to Toronto in this project report.

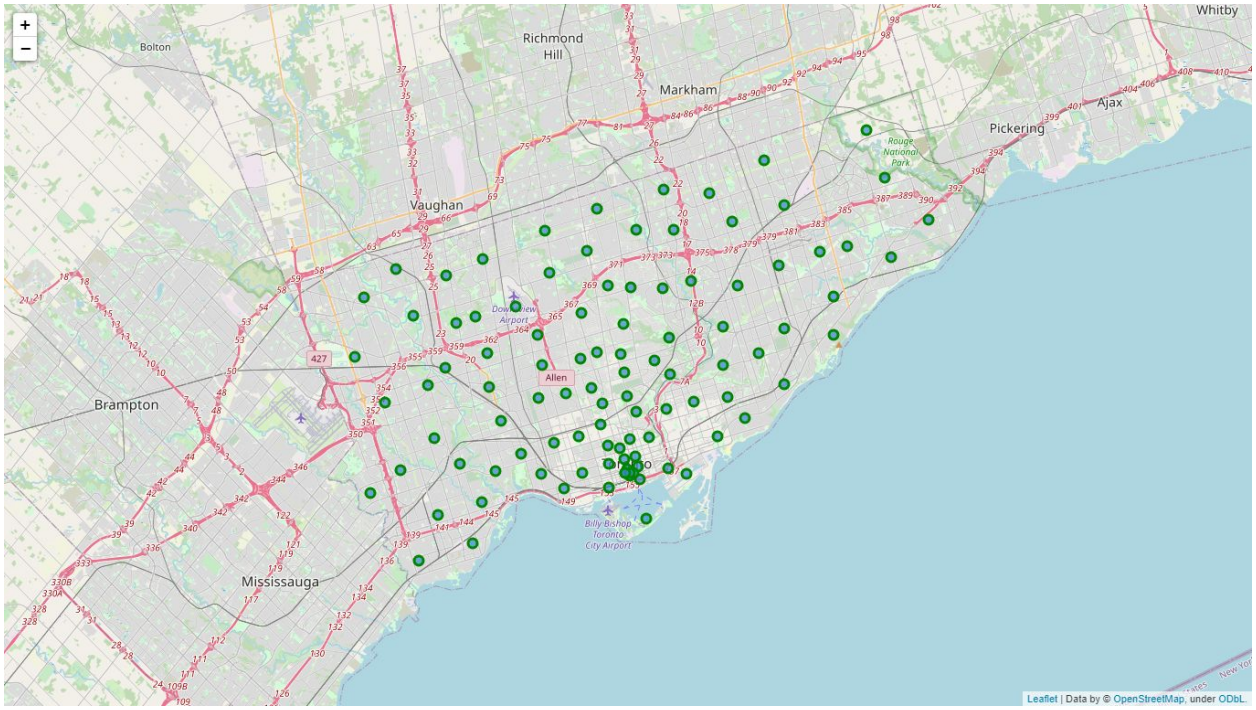

Figure 14: Map of Toronto and its Neighborhoods

4.2 Explore New York City and Toronto Neighborhoods

4.2.1 Most Common Venue Categories

Most common venue categories are defined as the set of venue categories that have the highest number of venues across all the neighborhoods of the city. These can be easily obtained by getting the top n rows of value counts after grouping on the Venue Category field in the venues dataset for each city.

4.2.1.1 New York City

There were 576 unique venue categories for the 306 New York City Neighborhoods. The top 15 most common venue categories from the **nyc_venues** dataframe for New York City Neighborhoods are shown in Figure 15 below.

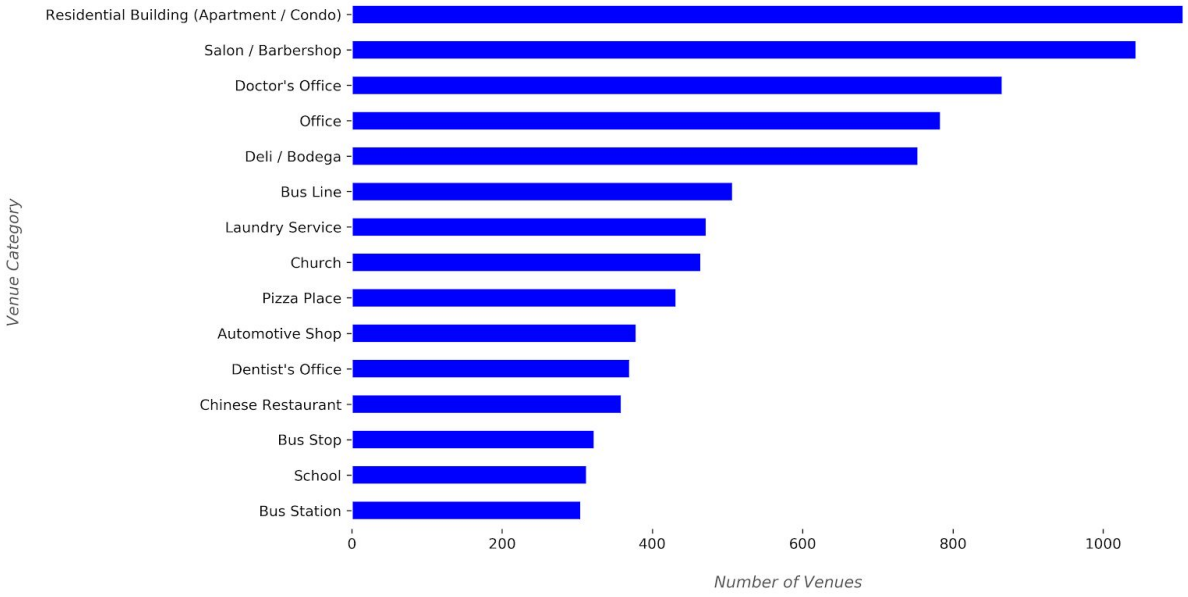

Figure 15: Most Common Venue Categories - New York City

4.2.1.2 Toronto

There were 499 unique venue categories for the 103 Toronto Neighborhoods. The top 15 most common venue categories from the **toronto_venues** for Toronto Neighborhoods are shown in Figure 16 below.

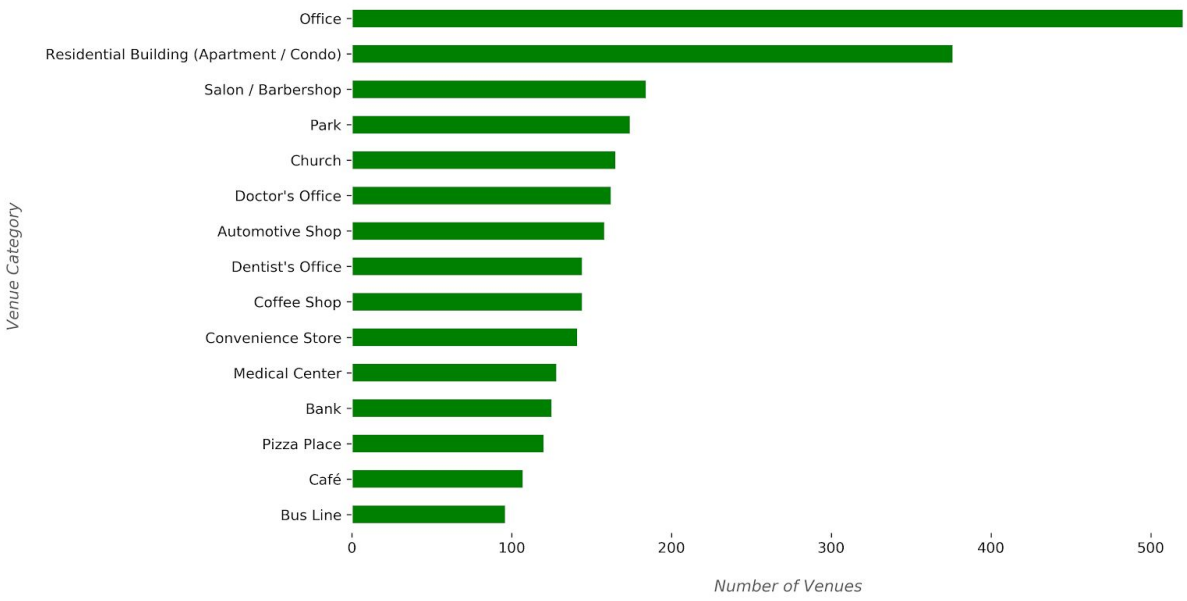

Figure 16: Most Common Venue Categories - Toronto

4.2.2 Most Widespread Venue Categories

Most widespread venue categories are defined as the venue categories that exist in most neighborhoods of the city. We need to first extract the Venue Category from the venues dataset of each city and then “One Hot Encode” to turn each unique venue category value into a separate column. We then get top n rows after getting the sum of venues grouped by neighborhood for each of the venues datasets for each city.

4.2.2.1 New York City

There were 576 unique venue categories for the 306 New York City Neighborhoods. The **nyc_onehot** dataframe was created as described above. The top 15 most widespread venue categories from **nyc_onehot** dataframe for New York City Neighborhoods are shown in Figure 17 below.

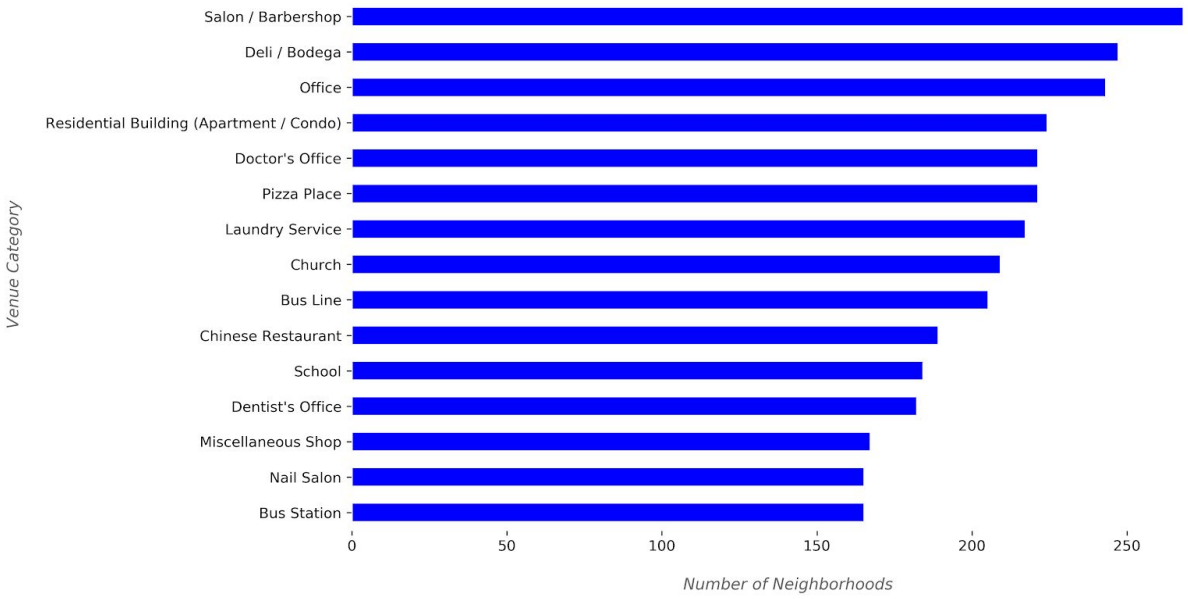

Figure 17: Most Widespread Venue Categories - New York City

4.2.2.2 Toronto

There were 499 unique venue categories for the 103 Toronto Neighborhoods. The top 15 most common venue categories from **toronto_onehot** dataframe for Toronto Neighborhoods are shown in Figure 18 below.

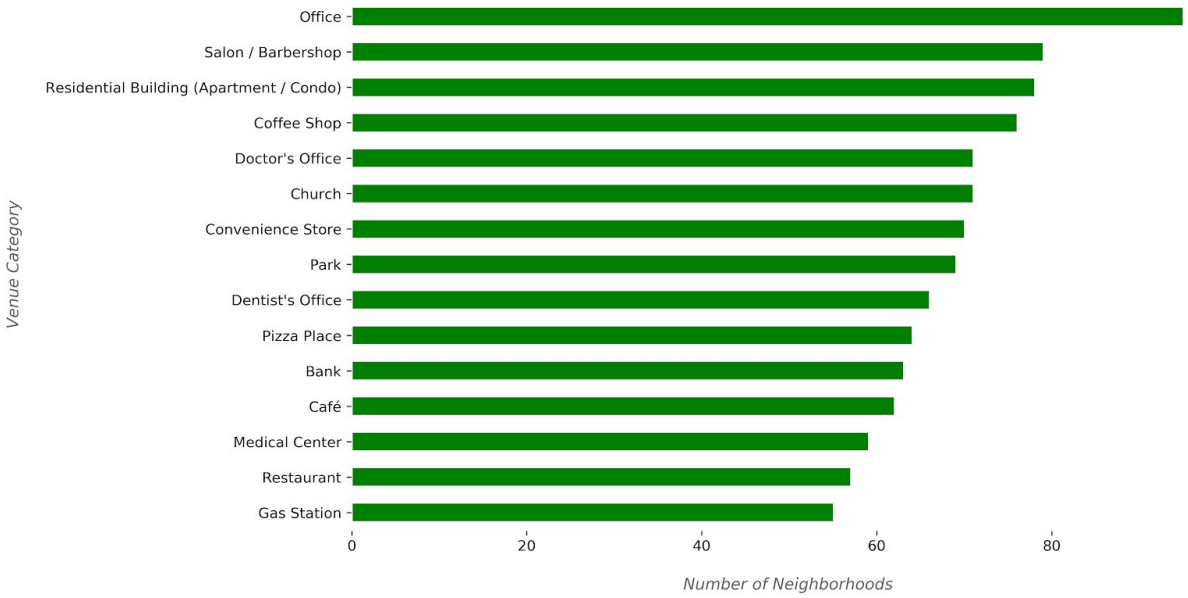

Figure 18: Most Widespread Venue Categories - Toronto

5. Cluster New York City and Toronto Neighborhoods

5.1 Feature Selection

The “One Hot Encoding” done on each of the venue datasets results in each unique value of the venue category becoming a separate column of its own. This was created in the prior step used for determining most widespread venue categories and is also the structure that will be required for the K-Means Clustering algorithm that we will apply to cluster the neighborhoods. We then group the “One Hot Encoded” venue category data by Neighborhood and calculate the mean of the frequency of occurrence of each category. We generated the **nyc_grouped** and **toronto_grouped** dataframes that represented these.

5.2 Combining New York City and Toronto Data

First we combined the New York City and Toronto Neighborhood data (note that columns are the same in both datasets). **There are a total 409 Neighborhoods across New York City and Toronto combined**.

combined_neighborhoods = **nyc_neighborhoods** concatenated with **toronto_neighborhoods**

Figure 19 below shows 10 rows from the combined_neighborhood dataframe.

	Borough	Neighborhood	Latitude	Longitude
301	Manhattan	NYC_Hudson Yards	40.756658	-74.000111
302	Queens	NYC_Hammels	40.587338	-73.805530
303	Queens	NYC_Bayswater	40.611322	-73.765968
304	Queens	NYC_Queensbridge	40.756091	-73.945631
305	Staten Island	NYC_Fox Hills	40.617311	-74.081740
306	North York	YYZ_Parkwoods	43.752935	-79.335641
307	North York	YYZ_Victoria Village	43.728102	-79.311890
308	Downtown Toronto	YYZ_Regent Park, Harbourfront	43.650964	-79.353041
309	North York	YYZ_Lawrence Manor, Lawrence Heights	43.723265	-79.451211
310	Downtown Toronto	YYZ_Queen's Park, Ontario Provincial Government	43.661790	-79.389390

Figure 19: combined_neighborhoods dataframe

Next we combined the New York City and Toronto Grouped data (note that the “One Hot Encoded” venue category columns will be different for each dataset and we replace all Null values generated for non-overlapping category columns with 0)

combined_grouped = **nyc_grouped** concatenated with **toronto_grouped** data

5.3 Most Common Venue Categories of Combined Neighborhoods

We also create datasets for each city that list each neighborhood and it’s top 10 venue category names - **nyc_neighborhoods_categories_sorted** & **toronto_neighborhoods_categories_sorted**

We finally combine the top 10 venue category datasets

combined_neighborhoods_categories_sorted = **nyc_neighborhoods_categories_sorted** concatenated with **toronto_neighborhoods_categories_sorted**

5.4 Clustering the Combined Neighborhoods

Finally we create the dataset that only contains the Features to be used for the K-Means Clustering algorithm. We will drop the Neighborhood from the **combined_grouped** dataset to create the **combined_grouped_clustering** dataset.

Now we are ready to run the K-Means Clustering algorithm but we need to determine the optimal value for the number of clusters to run it for first. We first used the **Elbow Method** to plot the Within-Cluster-Sum-of-Squares (WCSS) generated with clusters ranging from 2 to 10. The Figure 20 below shows the plot.

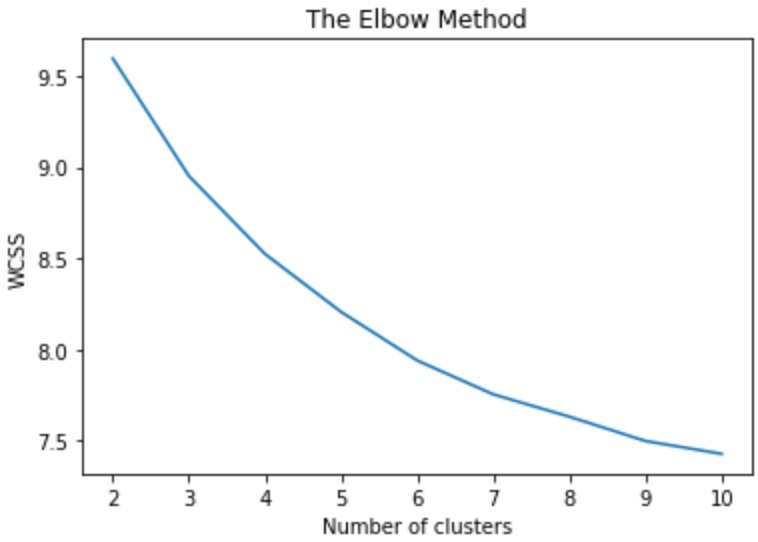

Figure 20: The Elbow Method for K-Means Clustering

As you can see from the above plot, it is not very easy to conclude what the optimal number of clusters should be for this dataset. We could offer perfectly reasonable explanations to pick 4, 5 or even 6. To reduce this ambiguity, we can try to use the **Silhouette Score Method**. Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

The Figure 21 below shows the plot for the Silhouette Score Method for this dataset.

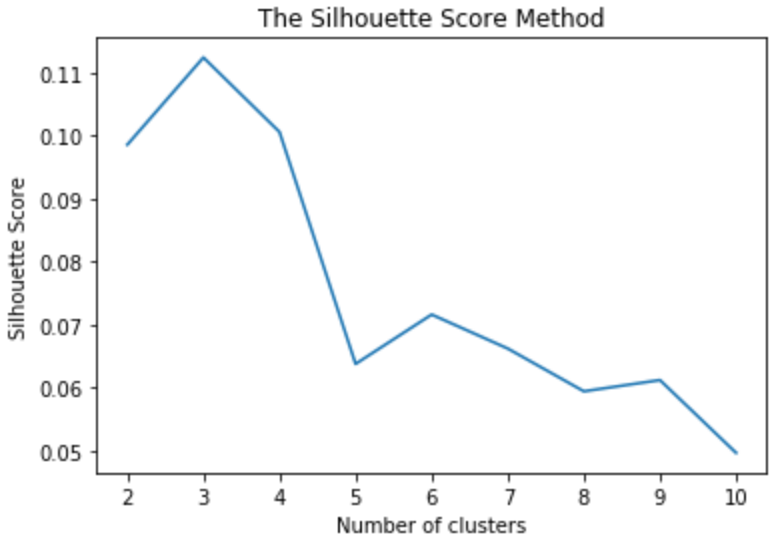

Figure 21: The Silhouette Score Method

As we can see from the above plot, **the optimal number of clusters for this dataset is definitely 5, after which we will likely see overfitting** and that is what we will use to run our final K-Means Clustering algorithm on.

Finally, we run the K-Means Clustering algorithm on the **combined_grouped_clustering** dataset with the optimal 5 clusters and obtain the cluster labels in a dataset named **kmeans**.

To prepare for our cluster analysis we now insert the cluster labels from **kmeans** into the **combined_neighborhoods_categories_sorted** dataset and then create a **combined_merged** dataset by joining it with **combined_neighborhoods** on Neighborhood to get the Neighborhood, Neighborhood Latitude and Longitude, Cluster Label and the Top 10 Venue Categories in our final dataset to be used for our cluster analysis.

5.5 Cluster Analysis

5.5.1 Visualizing the New York City Neighborhood Clusters

We now plot the neighborhoods of New York City to visually depict the clusters on a map as shown in Figure 22 below.

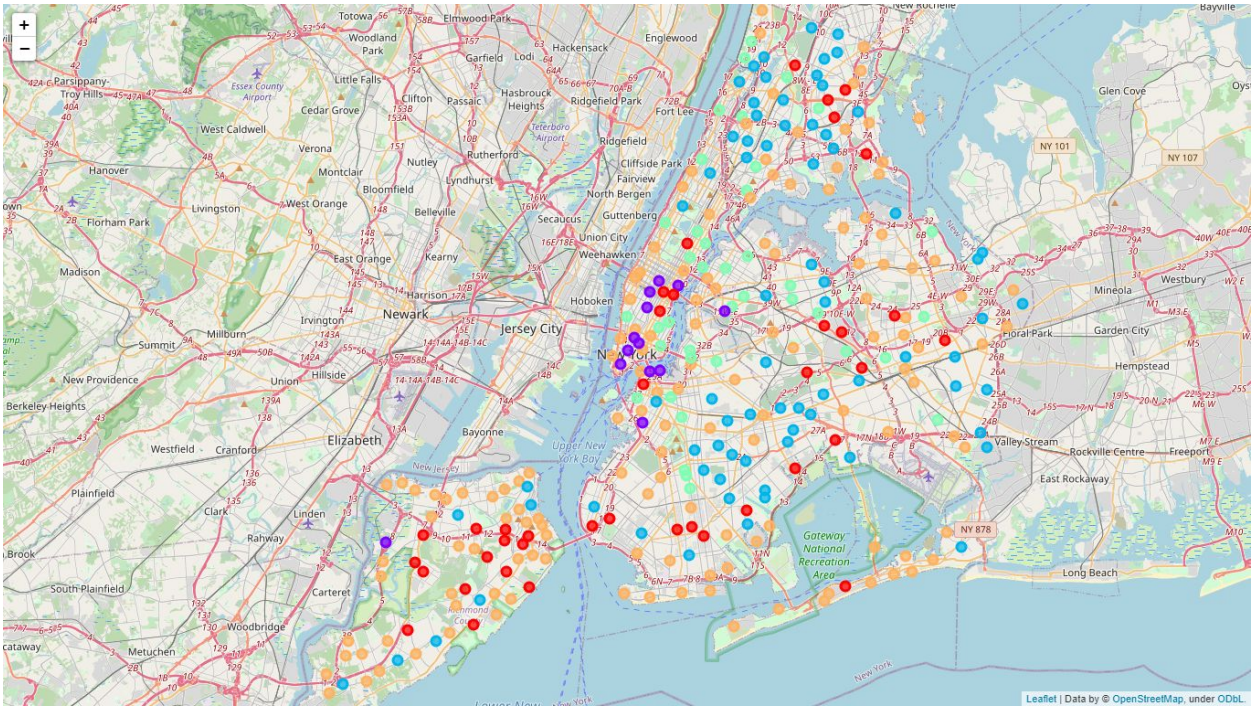

Figure 22: Resulting Clusters of Neighborhoods in New York City

5.5.2 Visualizing the Toronto Neighborhood Clusters

We now plot the neighborhoods of Toronto to visually depict the clusters on a map as shown in Figure 23 below.

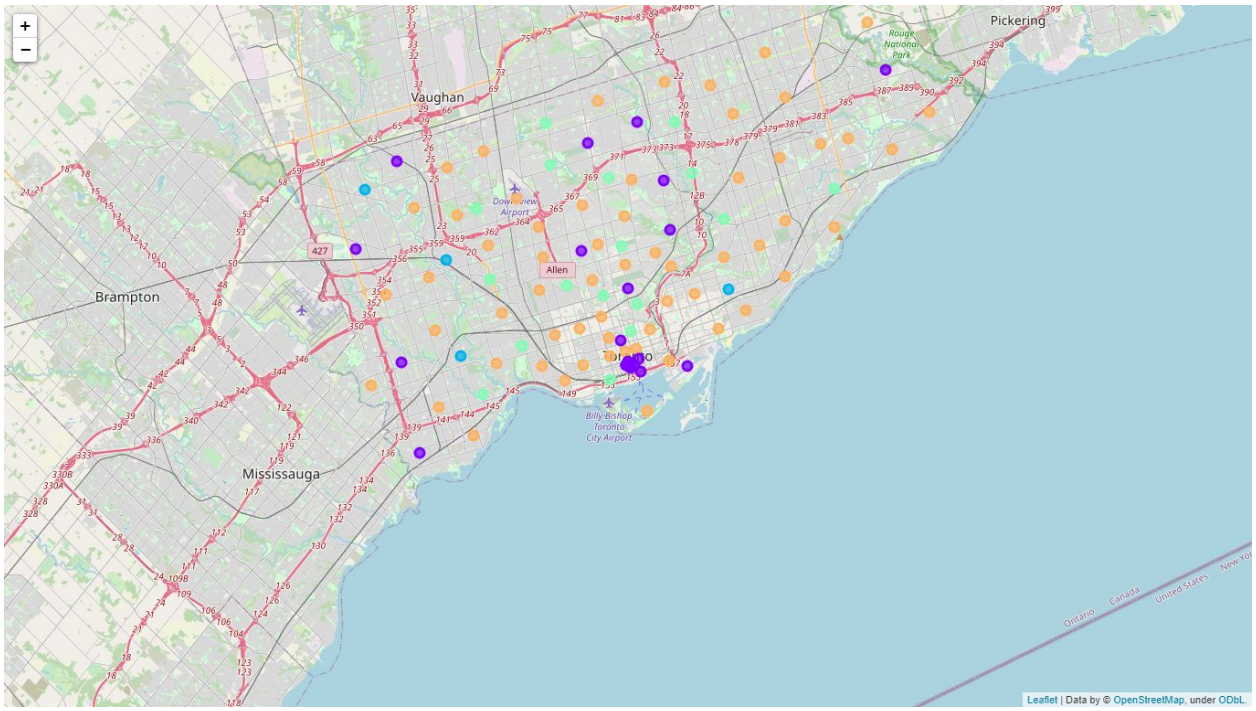

Figure 23: Resulting Clusters of Neighborhoods in Toronto

5.5.3 Analyzing the Neighborhoods of each city across the Clusters

We get a breakout of the number of neighborhoods of each city that were assigned to each cluster as in Figure 24 below.

Cluster Labels	City	
0	New York City	39
1	New York City	13
	Toronto	22
2	New York City	74
	Toronto	4
3	New York City	39
	Toronto	17
4	New York City	141
	Toronto	60

Figure 24: Neighborhoods by City in each Cluster

The Figure 25 below shows this in a visual chart form.

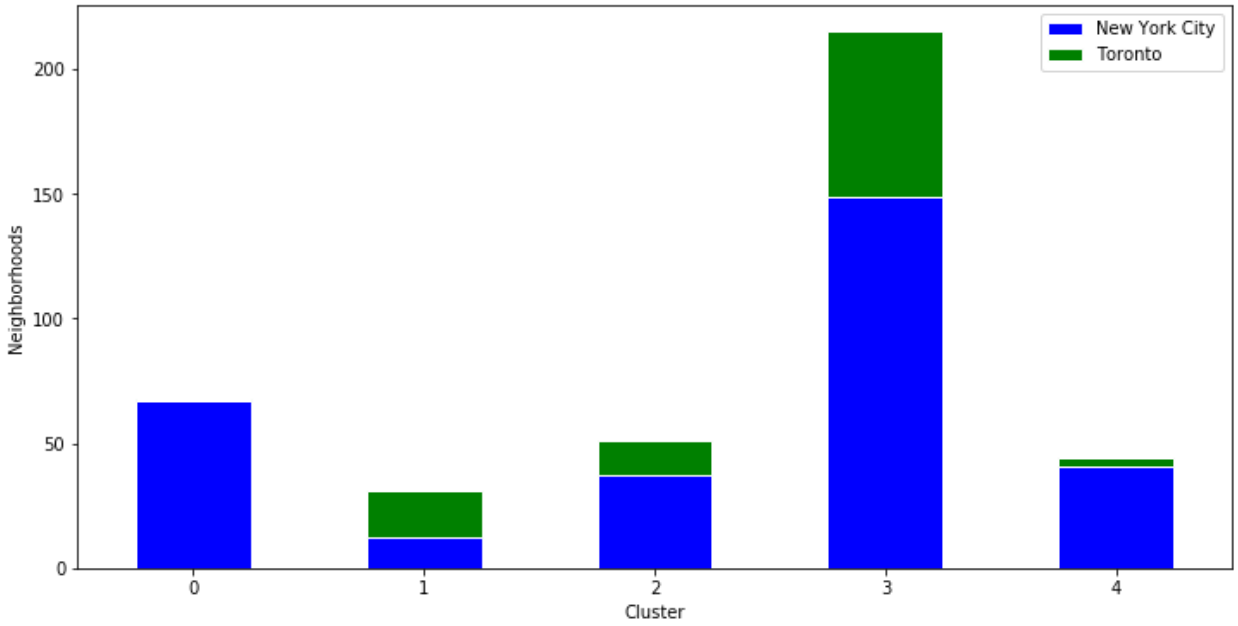

Figure 25: Chart of Neighborhoods by City in each Cluster

The Figure 26 below shows the chart of unique venue categories and the mean frequency of occurrence within each cluster. We can visually see the bigger bars for each cluster and use that to form the basis of naming that cluster. We will elaborate more on this for each cluster in the section following this.

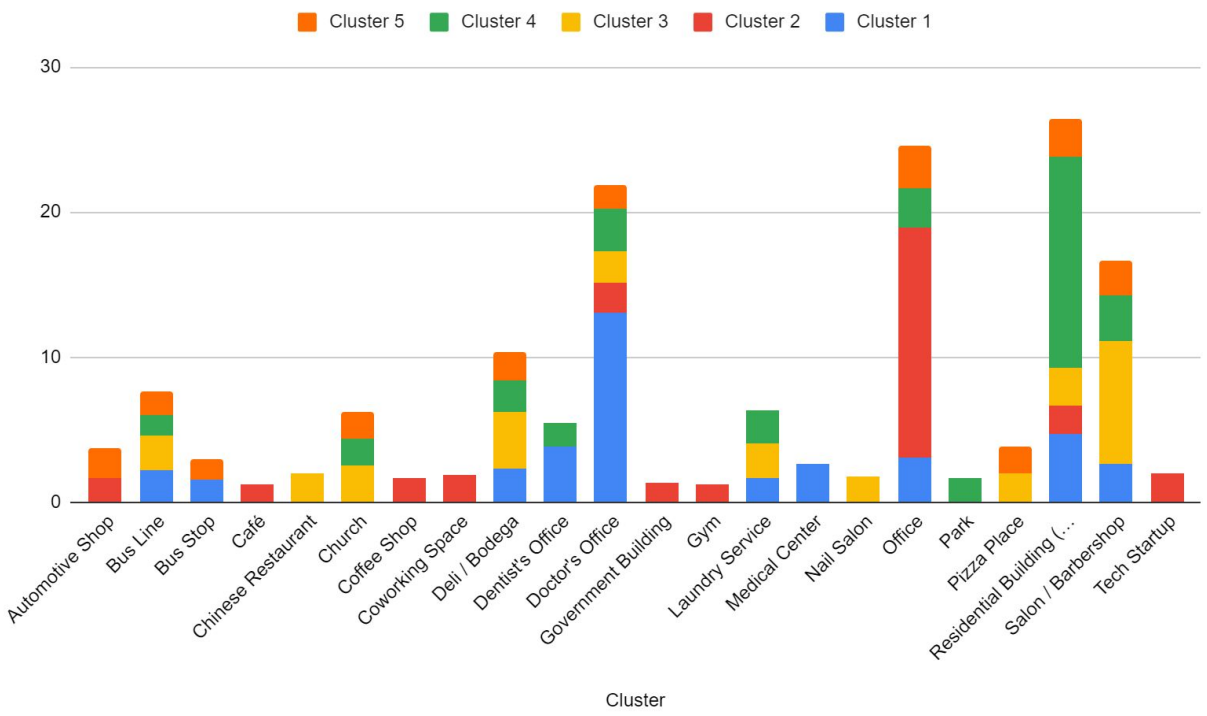

Figure 26: The Clusters and Venue Category Frequencies

5.5.4 Analyzing and Naming Cluster 1 (Cluster Label = 0)

This cluster has a combination of Doctor's Office, Dentist's Office and Medical Center (~ 20% aggregate) as the highest set of venue category frequencies and thus can be named as **"Neighborhoods around Medical Facilities"**. These neighborhoods are unique to New York City since there were no neighborhoods from Toronto in this cluster.

Doctor's Office	13.113695
Residential Building (Apartment / Condo)	4.748062
Dentist's Office	3.779070
Office	3.068475
Medical Center	2.680879
Salon / Barbershop	2.616279
Deli / Bodega	2.357881
Bus Line	2.164083
Laundry Service	1.647287
Bus Stop	1.518088

Figure 27: Top 10 Venue Categories in Cluster 1 (Cluster Label = 0)

5.5.5 Analyzing and Naming Cluster 2 (Cluster Label = 1)

This cluster has a heavy concentration of Office and related categories like Tech Startup, Coworking Space, Event Space & Government Building (~ 20% aggregate) as the highest set of venue category frequencies and thus can be named as **"Neighborhoods that are Business and Commercial Centers"**.

Office	15.815486
Doctor's Office	2.009885
Tech Startup	2.009885
Coworking Space	1.878089
Residential Building (Apartment / Condo)	1.878089
Coffee Shop	1.647446
Automotive Shop	1.614498
Government Building	1.285008
Café	1.252059
Gym	1.252059

Figure 28: Top 10 Venue Categories in Cluster 2 (Cluster Label = 1)

5.5.6 Analyzing and Naming Cluster 3 (Cluster Label = 2)

This cluster has a combination of Salon / Barbershop, Deli / Bodega, Laundry Service, Church & Residential (~ 23% aggregate) as the highest set of venue category frequencies but the services dominate this cluster more than Residential and we further note that Residential is sparse instead of dense and thus this cluster can be named as **"Neighborhoods that are sparsely Residential with above average access to services"**.

Salon / Barbershop	8.503961
Deli / Bodega	3.900762
Residential Building (Apartment / Condo)	2.660290
Church	2.510835
Bus Line	2.465999
Laundry Service	2.436108
Doctor's Office	2.196981
Pizza Place	2.032581
Chinese Restaurant	2.002690
Nail Salon	1.808399

Figure 29: Top 10 Venue Categories in Cluster 3 (Cluster Label = 2)

5.5.7 Analyzing and Naming Cluster 4 (Cluster Label = 3)

This cluster has a combination of dense Residential, Salon / Barbershop, Deli / Bodega, Laundry Service, Church & (~ 25% aggregate) as the highest set of venue category frequencies and thus can be named as **"Neighborhoods that are densely Residential with good access to services"**.

Residential Building (Apartment / Condo)	14.517189
Salon / Barbershop	3.109262
Doctor's Office	2.912196
Office	2.758923
Laundry Service	2.255310
Deli / Bodega	2.123933
Church	1.839282
Park	1.642216
Dentist's Office	1.642216
Bus Line	1.335669

Figure 30: Top 10 Venue Categories in Cluster 4 (Cluster Label = 3)

5.5.8 Analyzing and Naming Cluster 5 (Cluster Label = 4)

This cluster has a balanced presence of Residential and Office as the highest set of venue category frequencies and thus can be named as **"Neighborhoods with balanced Residential and Business presence"**.

Office	2.927847
Residential Building (Apartment / Condo)	2.605844
Salon / Barbershop	2.391175
Automotive Shop	2.134764
Deli / Bodega	1.967800
Church	1.920095
Pizza Place	1.747168
Bus Line	1.717352
Doctor's Office	1.669648
Bus Stop	1.443053

Figure 31: Top 10 Venue Categories in Cluster 5 (Cluster Label = 4)

6. Results and Discussion

Our analysis of **Most Common Venue Categories** of the two cities shows that New York City has Residential as the topmost common venue category overall while Toronto has Office as the topmost common venue category. In New York City, Salon / Barbershop, Doctor's Office, Office and Deli / Bodega are the other categories in top 5 common while Toronto has Residential, Salon / Barbershop, Park and Church in top 5 common.

Our analysis of **Most Widespread Venue Categories** (that exist in most neighborhoods) shows that New York City has Salon / Barbershop as the most widespread venue category while Toronto has Office as the most widespread venue category. In New York City, Deli / Bodega, Office, Residential & Doctor's Office are the other widespread categories in top 5 widespread while Toronto has Salon / Barbershop, Residential, Coffee Shop and Doctor's Office in top 5 widespread.

Some **interesting observations** based on the above:

- In New York City, Residential is the Most Common Venue Category but it is only the 4th Most Widespread Venue Category which means that this venue category is highly concentrated in some of the 306 neighborhoods.
- In Toronto, Office is the Most Common Venue Category and it is also the Most Widespread Venue Category which means that this venue category is highly concentrated in most of the 103 neighborhoods.

The **highlights of the analysis of the 5 clusters** are explained below.

There are **3 clusters which have similar neighborhoods from both New York City and Toronto**.

- **“Neighborhoods with balanced Residential and Business presence”**
  - New York City = 141 & Toronto = 60
- **“Neighborhoods that are Business and Commercial Centers”**
  - New York City = 13 & Toronto = 22
- **“Neighborhoods that are densely Residential with good access to services”**
  - New York City = 39 & Toronto = 17

There are **2 clusters which have neighborhoods that are quite distinct in New York City**.

- **“Neighborhoods around Medical Facilities”**
  - New York City = 39 & Toronto = 0
- **“Neighborhoods that are sparsely Residential with above average access to services”**
  - New York City = 74 & Toronto = 4

7. Conclusion

Specifically, we wanted to find out answers to questions such as:

- What are the most common venue categories in each city?

  New York City has Residential as the topmost common venue category overall while Toronto has Office as the topmost common venue category. In New York City, Salon / Barbershop, Doctor's Office, Office and Deli / Bodega are the other categories in top 5 while Toronto has Residential, Salon / Barbershop, Park and Church in top 5

- What venue categories are most widespread in each city?

  New York City has Salon / Barbershop as the most widespread venue category while Toronto has Office as the most widespread venue category. In New York City, Deli / Bodega, Office, Pizza Place & Doctor's Office are the other widespread categories in top 5 widespread while Toronto has Salon / Barbershop, Residential, Coffee Shop and Doctor's Office in top 5 widespread

- How many neighborhoods of these two cities are similar to each other?

  There are 3 sets of neighborhoods of these two cities that are very similar to each other

  - *"Neighborhoods with balanced Residential and Business presence"*
    - *New York City = 141 & Toronto = 60*
  - *"Neighborhoods that are Business and Commercial Centers"*
    - *New York City = 13 & Toronto = 22*
  - *"Neighborhoods that are densely Residential with good access to services"*
    - *New York City = 39 & Toronto = 17*
- How many neighborhoods of these two cities are quite unique to themselves?

  There are 2 sets of neighborhoods of these two cities that are very unique to themselves and all but one of them are in New York City

  - *"Neighborhoods around Medical Facilities"*
    - *New York City = 39 & Toronto = 0*
  - *"Neighborhoods that are sparsely Residential with above average access to services"*
    - *New York City = 74 & Toronto = 4*
- Are these two cities very similar to each other or very different?

  The fact that 3 out of the 5 clusters have a significant number of each city's neighborhoods (~ 63% of New York City & ~ 96% of Toronto) in the similar cluster buckets indicates that these two cities are very similar to each other.

There are 2 sets of neighborhoods almost all in New York City (they represent ~ 37% of the neighborhoods of that city) that are quite unique to themselves and have insignificant (~ 4%) similar neighborhoods in Toronto.

8. References

[1] New York City Neighborhood Names

[2] List of postal codes of Canada: M