IBM Data Science Professional Certificate

Capstone Project – Ajay Bhatnagar

Comparing and Clustering Neighborhoods of New York City and Toronto

Jun 2020

1. Introduction / Business Problem

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. 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?
Are these two cities very similar to each other or very different?
Which neighborhoods of these two cities are similar to each other?
Which neighborhoods of these two cities are quite unique to themselves?

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 provides reasonable justifications for these similarities and that we are also able to outline what, if any, the dissimilarities are.

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 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 some cases, 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 cases 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 that contains the following details about each neighborhood of New York City:

Name, Borough, Coordinates as Latitude and Longitude and some additional 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 Pandas 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 – Pandas dataframe

We then check to see if there are any duplicate Neighborhood names. There are 4 Neighborhoods (Sunnyside, Bay Terrace, Murray Hill & Chelsea) with same name that exists 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

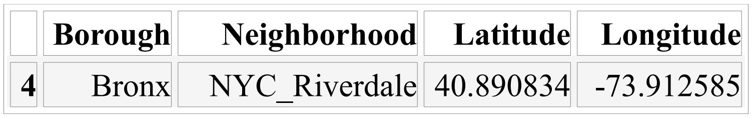
After:

	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 Pandas dataframe are shown in Figure 4 below.

	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



	Borough	Neighborhood	Latitude	Longitude
4	Bronx	NYC_Riverdale	40.890834	-73.912585

Figure 4: New York City Neighborhoods – final Pandas dataframe

We then plot all the neighborhoods of New York City on a map using the Folium package as shown in Figure 4 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.

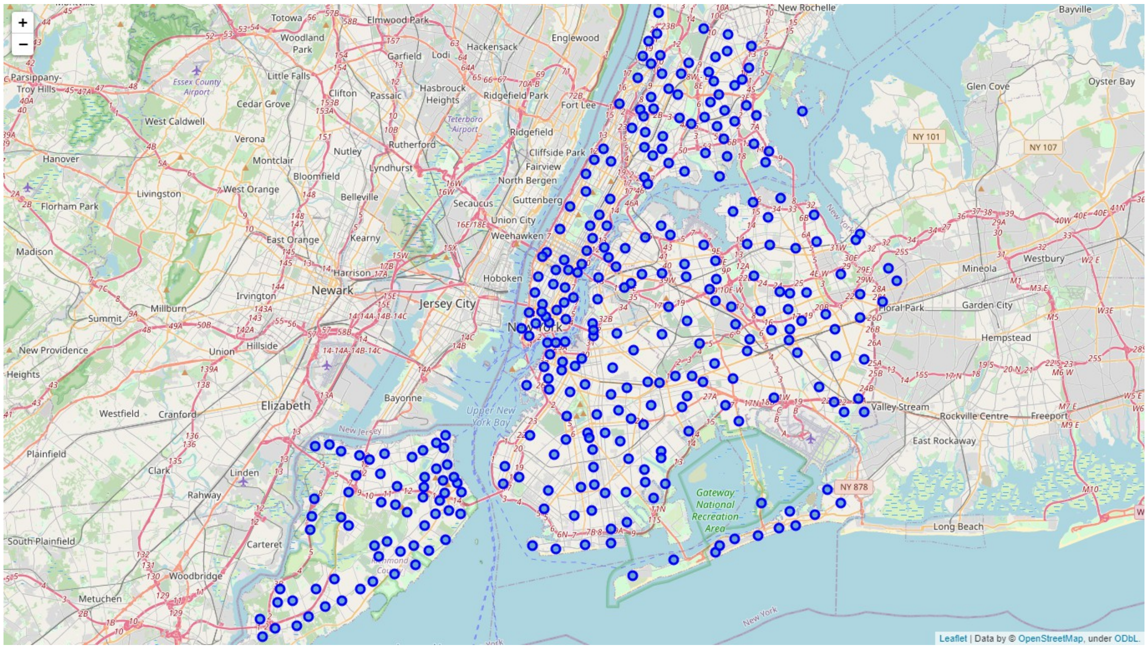
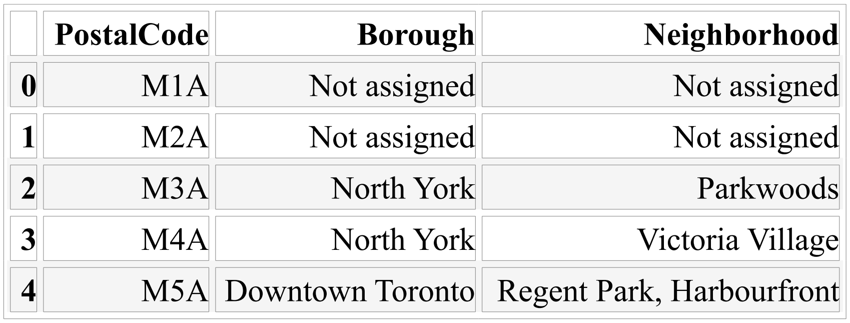


Figure 4: Map of New York City and its neighborhoods

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" at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M 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 Pandas dataframe as shown in Figure 5 below. This dataframe contains 180 Neighborhoods of Toronto in 11 Boroughs.

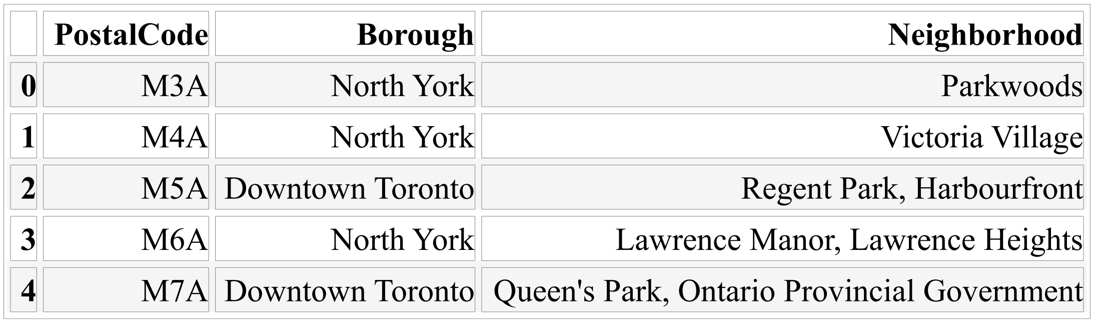


	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 5: Toronto neighborhoods – Pandas dataframe

On closer inspection, it is seen that each row represents a unique value of the Postal Code prefix with the following additional observations:
A Borough may contain more than one Postal Code prefix 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 in this dataframe where the Borough is "Not assigned" - these need to be dropped from the dataset 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.

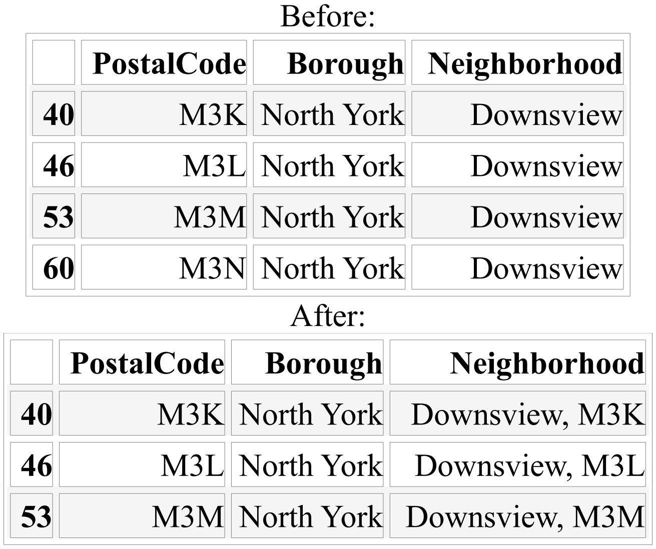
The final Pandas dataframe after all the above wrangling activities are completed contains 103 neighborhoods. The first 5 rows of this dataframe are shown in Figure 6 below.



	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 – Pandas dataframe after wrangling

We then check to see if there are any duplicate Neighborhood names. There are 2 Neighborhoods (Downsview & Don Mills) with same name that exists in multiple Postal Code prefix. 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.

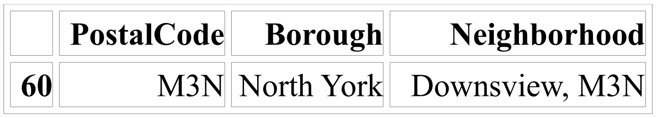


Before:

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

After:

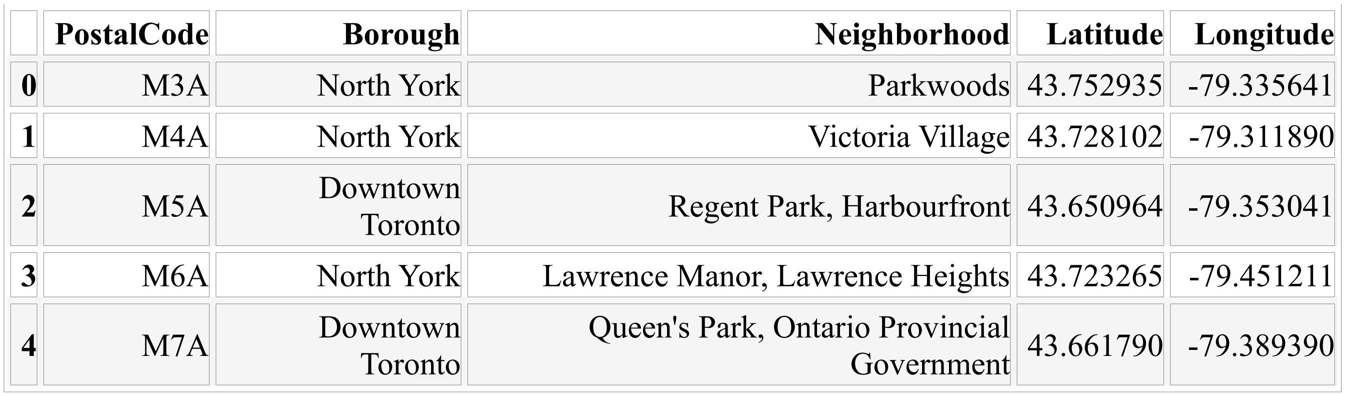
	PostalCode	Borough	Neighborhood
40	M3K	North York	Downsview, M3K
46	M3L	North York	Downsview, M3L
53	M3M	North York	Downsview, M3M



	PostalCode	Borough	Neighborhood
60	M3N	North York	Downsview, M3N

Figure 7: Toronto Neighborhoods – Unique Neighborhood Names

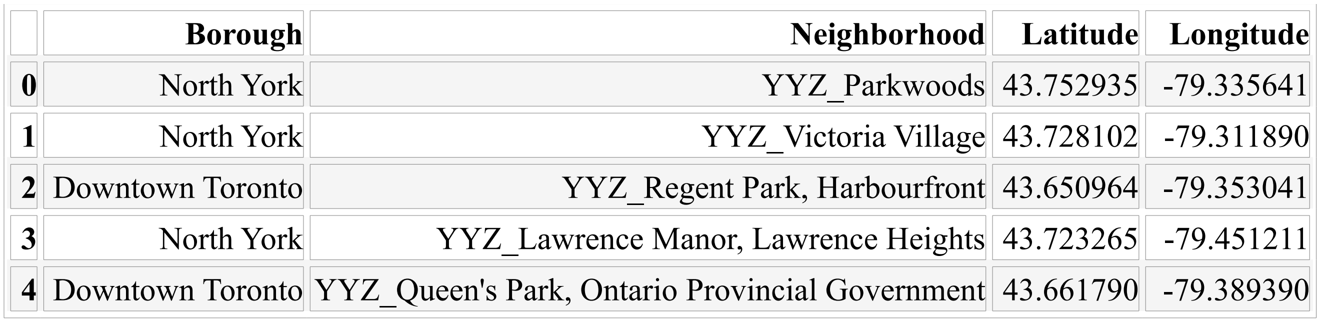
The next wrangling step requires us to find a way to obtain Latitude and Longitude coordinates for each of the Toronto neighborhoods. We will use the geocoder package with the ArcGIS provider (Google doesn't work) to obtain these coordinates and add to our Pandas dataframe. The first 5 rows of the updated Pandas 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 – Pandas 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 Pandas dataframe are shown in Figure 9 below.



	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 Pandas dataframe

We then plot all the neighborhoods of Toronto on a map using the Folium package as shown in Figure 10 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.

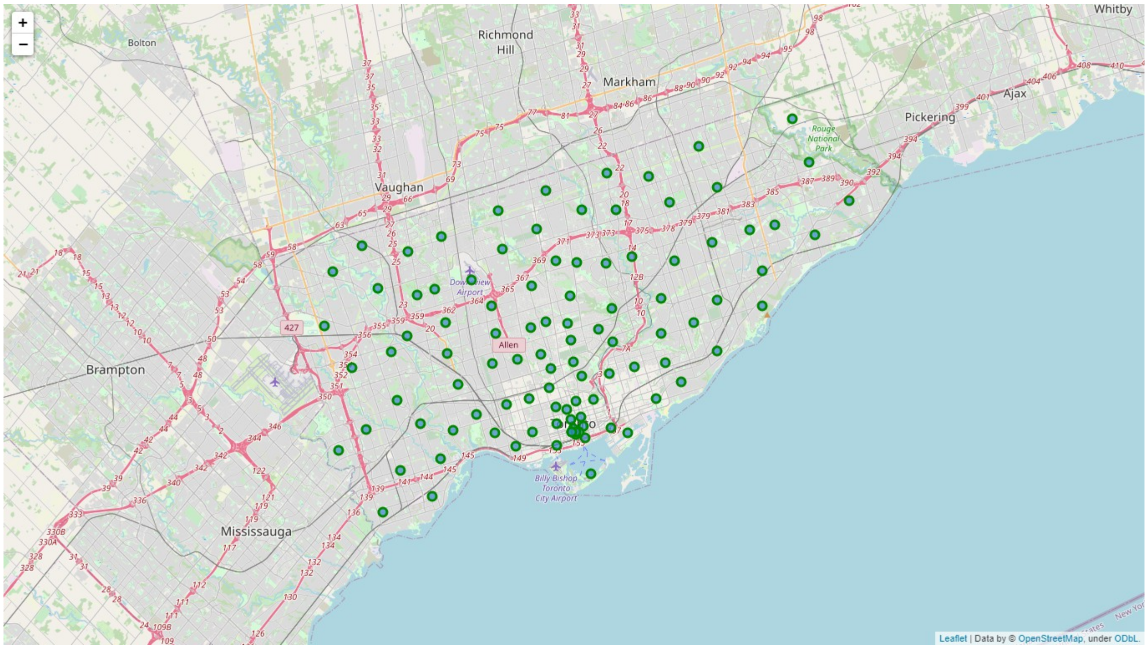


Figure 10: Map of Toronto and its neighborhoods

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.

To retrieve the 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 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 Pandas dataframe. In case of New York City, we get 26, 503 Venues for the 306 Neighorhoods that had 577 unique Venue Categories. The first 5 rows of the Pandas dataframe is shown below in Figure 11.

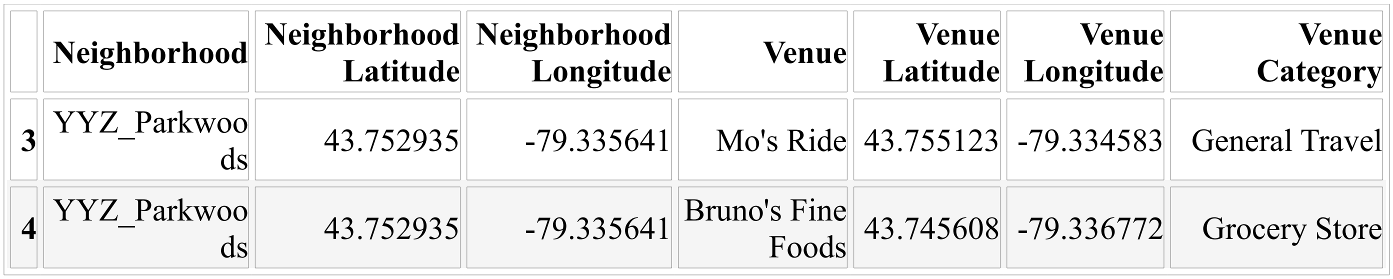
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NYC_Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
1	NYC_Wakefield	40.894705	-73.847201	Pitman Deli	40.896744	-73.844398	Food
2	NYC_Wakefield	40.894705	-73.847201	Julio 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 dataframe for New York City

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 Pandas dataframe. In case of Toronto, we get 8, 757 Venues for the 103 Neighorhoods that had 500 unique Venue Categories. The first 5 rows of the Pandas dataframe 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



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
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 dataframe for Toronto

3 Exploring the Neighborhoods

3.1 Most Common Venue Categories

3.1.1 New York City

3.1.2 Toronto

3.2 Most Widespread Venue Categories

3.2.1 New York City

3.2.2 Toronto

4 Clustering the Neighborhoods

4.1 Feature Selection

4.2 Combining New York City and Toronto Data

4.3 Most Common Venue Categories of Combined Neighborhoods

4.4 Clustering the Combined Neighborhoods

4.5 Cluster Analysis

5 Conclusion

6 Appendices