Exploring Toronto Neighborhoods to open a Pizzeria

As a part of the IBM Data Science professional program Capstone Project, I worked on the real datasets to get an experience of what a data scientist goes through in real life. Main objectives of this project were to define a business problem, look for data in the web and, use Foursquare location data to compare different neighborhoods of Toronto to figure out which neighborhood is suitable for starting a new Pizzeria business. In this project, I will go through all the process in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

1. Description of the Business Problem & Discussion of the Background (Introduction Section):

Problem Statement: Identifying Ideal locations to open a Pizzeria in Toronto, Canada.

Toronto is the capital city of the Canadian province 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 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, while the Greater Toronto Area (GTA) proper had a 2016 population of 6,417,516. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

The city continues to grow and attract immigrants. A study by Ryerson University showed that Toronto was the fastest-growing city in North America. The city added 77,435 people between July 2017 and July 2018. The Toronto metropolitan area was the second-fastest-growing metropolitan area in North America, adding 125,298 persons, compared with 131,767 in Dallas-Fort Worth-Arlington in Texas. The large growth in the Toronto metropolitan area is attributed to international migration to Toronto.

Source : https://en.wikipedia.org/wiki/Toronto

In this project I will go through step by step process to make a decision whether it is a good idea to open an Pizzeria. I analyze the neighborhoods in Toronto to identify the most profitable area since the success of the restaurant depends on availability of a niche in the market in terms of competition and easy access to material sourcing.

Why Pizzeria?

91% of Americans eat pizza at least once a month, and if we're honest, it's not going anywhere. Pizza has been around since the 1800s, and it has incredible universal appeal. People from around the world love their pizza! Opening a pizza franchise means that you're offering a high-demand product that ensures you'll almost always see a profit. Obviously, demand alone doesn't mean a business will see success, but it does make a business a viable option.

Pizza is no longer just a restaurant food, although many patrons love a sit-down pizza experience. People can order it from anywhere- while coming home from work, sitting at home, or for things like events. The convenience of pizza is easily a favorite. We live in a convenience-oriented society, and pizza has a well-established niche.

https://www.westsidepizza.com/blog/why-pizza-franchises-are-good-investment

Target Audience

Who will be more interested in this project? What type of clients or a group of people would be benefitted?

1. Business personnel who wants to invest or open Pizzeria in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting this space.
2. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.
3. Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

2. Data acquisition and cleaning:

2.1 Data Sources

a) I'm using "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighborhoods.

c) To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

From Foursquare API (https://developer.foursquare.com/docs), I retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

d) Identify neighbourhoods with lesser competition

Using the foursquare data, we will be taking a count of venue locations marked as "Pizza Place", "Italian Restaurant". This will help us identify in which neighbourhoods there is currently a niche in the market that we can take advantage of. The neighbourhoods where the count is zero for these types of venues will be identified as potential candidates for us.

2.2 Data Cleaning

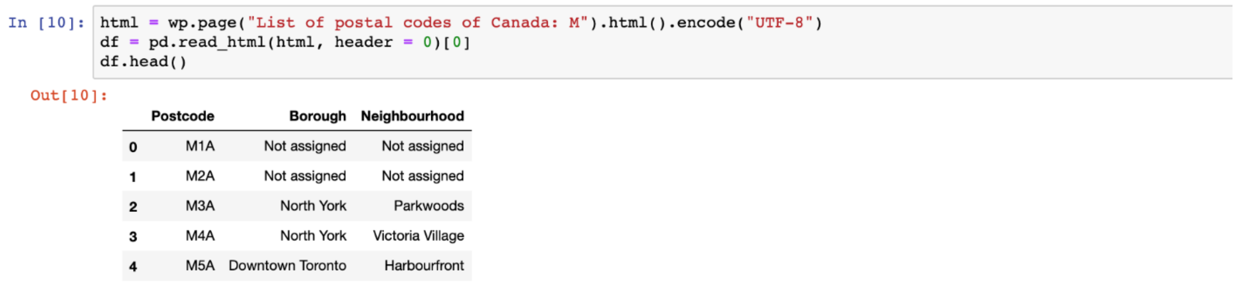
a) Scraping Toronto Neighborhoods Table from Wikipedia

Scraped the following Wikipedia page, "*List of Postal code of Canada: M*" in order to obtain the data about the Toronto & the Neighborhoods in it.

Assumptions made to attain the below DataFrame:

- Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

Wikipedia — package is used to scrape the data from wiki.



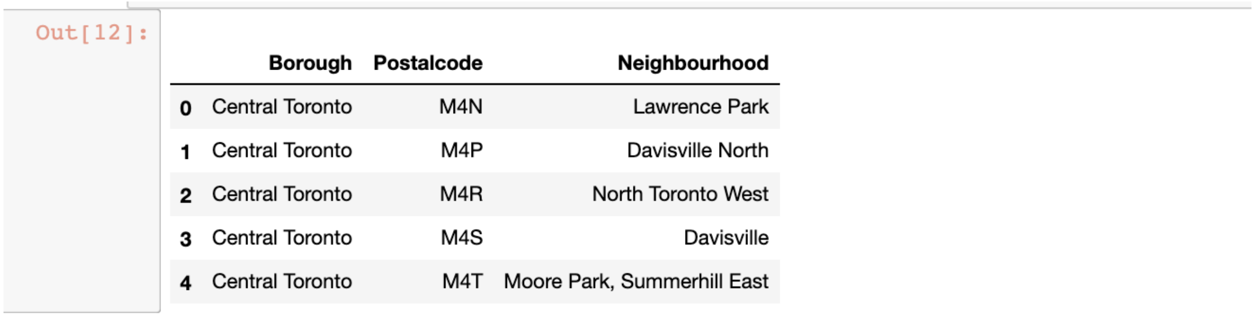
```
In [10]: html = wp.page("List of postal codes of Canada: M").html().encode("UTF-8")
         df = pd.read_html(html, header = 0)[0]
         df.head()
```

Out[10]:

	Postcode	Borough	Neighbourhood
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	Harbourfront

Dataframe formed from the scraped wiki page

After some cleaning I got the proper dataframe with the Postal code, Borough & Neighborhood information.



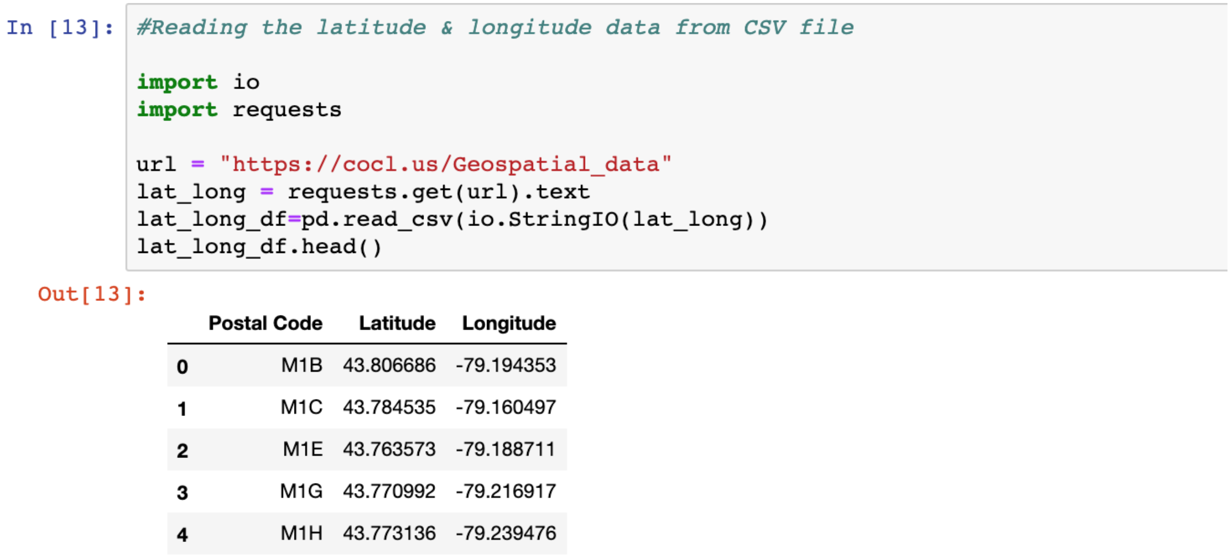
Out[12]:

	Borough	Postalcode	Neighbourhood
0	Central Toronto	M4N	Lawrence Park
1	Central Toronto	M4P	Davisville North
2	Central Toronto	M4R	North Toronto West
3	Central Toronto	M4S	Davisville
4	Central Toronto	M4T	Moore Park, Summerhill East

Dataframe from ‘*List of Postal code of Canada: M’* Wikipedia Table.

b) Adding geographical coordinates to the neighborhoods

Next important step is adding the geographical coordinates to these neighborhoods. To do so I’m extracting the data present in the Geospatial Data csv file and I’m combining it with the existing neighborhood dataframe by merging them both based on the postal code.



In [13]:

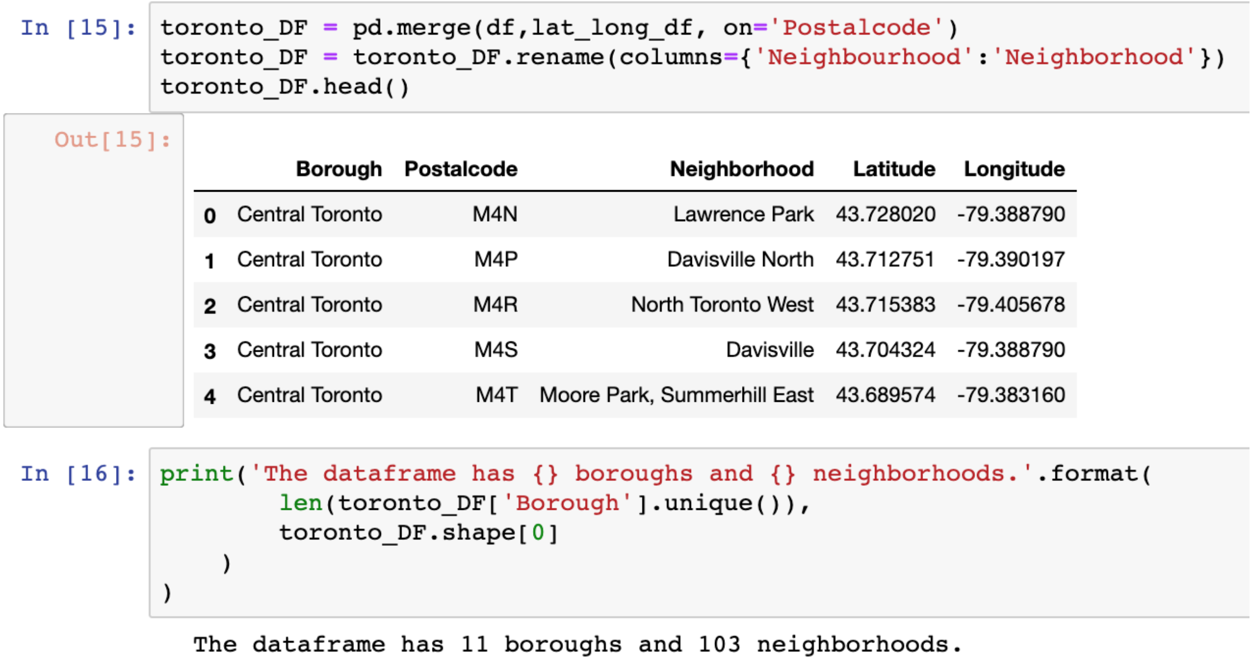
```
#Reading the latitude & longitude data from CSV file

import io
import requests

url = "https://cocl.us/Geospatial_data"
lat_long = requests.get(url).text
lat_long_df=pd.read_csv(io.StringIO(lat_long))
lat_long_df.head()
```

Out[13]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

DataFrame with latitude & longitude of Postal codes in Toronto

I'm renaming the columns to match the existing dataframe formed from '*List of Postal code of Canada: M' wiki page.*After that I'm merging both the dataframe into one by merging on the postal code.



In [15]:

```
toronto_DF = pd.merge(df,lat_long_df, on='Postalcode')
toronto_DF = toronto_DF.rename(columns={'Neighbourhood':'Neighborhood'})
toronto_DF.head()
```

Out[15]:

	Borough	Postalcode	Neighborhood	Latitude	Longitude
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678
3	Central Toronto	M4S	Davisville	43.704324	-79.388790
4	Central Toronto	M4T	Moore Park, Summerhill East	43.689574	-79.383160

In [16]:

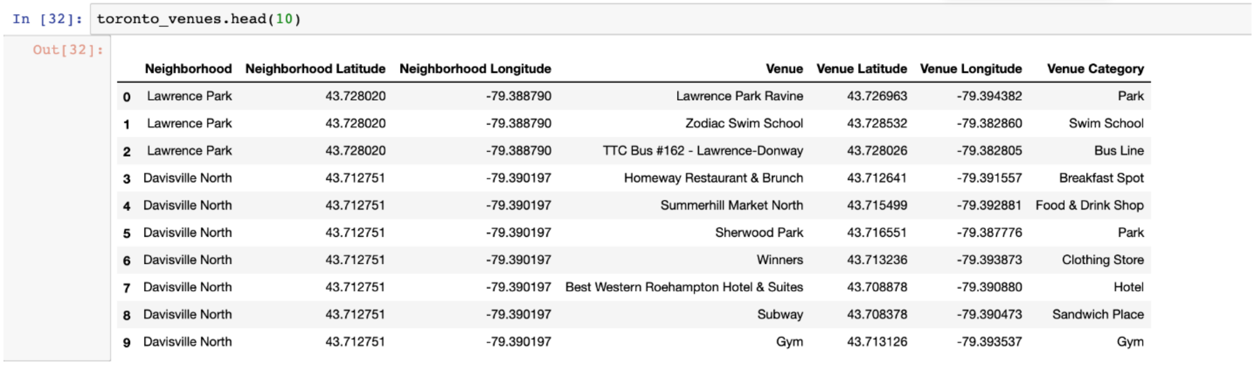
```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(toronto_DF['Borough'].unique()),
        toronto_DF.shape[0]
    )
)
```

```
The dataframe has 11 boroughs and 103 neighborhoods.
```

Merged new dataframe with info about Neighborhoods, borough, postalcode, latitude & longitude in Toronto

c) Get location data using Foursquare

Foursquare API is very useful online application used my many developers & other applications like Uber etc. In this project I have used it to retrieve information about the places present in the neighborhoods of Toronto. The API returns a JSON file and I need to turn that into a data-frame. Here I’ve chosen 100 popular spots for each neighborhood within a radius of 1km.



```
In [32]: toronto_venues.head(10)
```

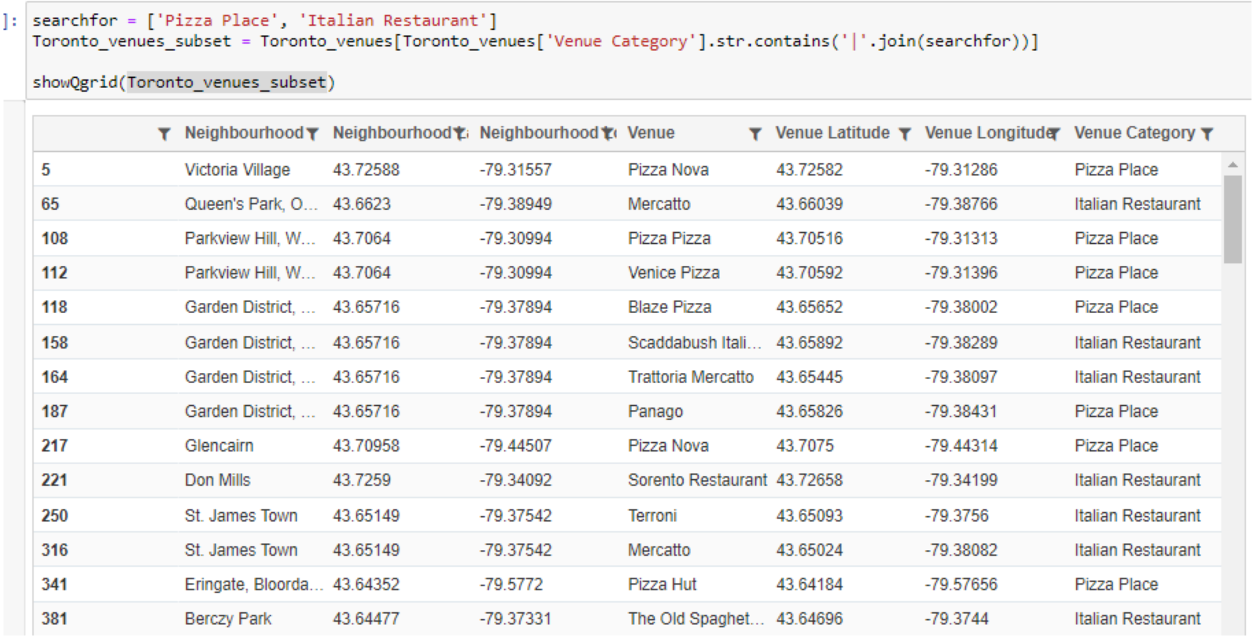
Out[32]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
4	Davisville North	43.712751	-79.390197	Summerhill Market North	43.715499	-79.392881	Food & Drink Shop
5	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park
6	Davisville North	43.712751	-79.390197	Winners	43.713236	-79.393873	Clothing Store
7	Davisville North	43.712751	-79.390197	Best Western Roehampton Hotel & Suites	43.708878	-79.390880	Hotel
8	Davisville North	43.712751	-79.390197	Subway	43.708378	-79.390473	Sandwich Place
9	Davisville North	43.712751	-79.390197	Gym	43.713126	-79.393537	Gym

Dataframe with venues in each neighborhood along with the category info of the venues.

d) Highlight the potential candidates for Neighbourhoods with less competition

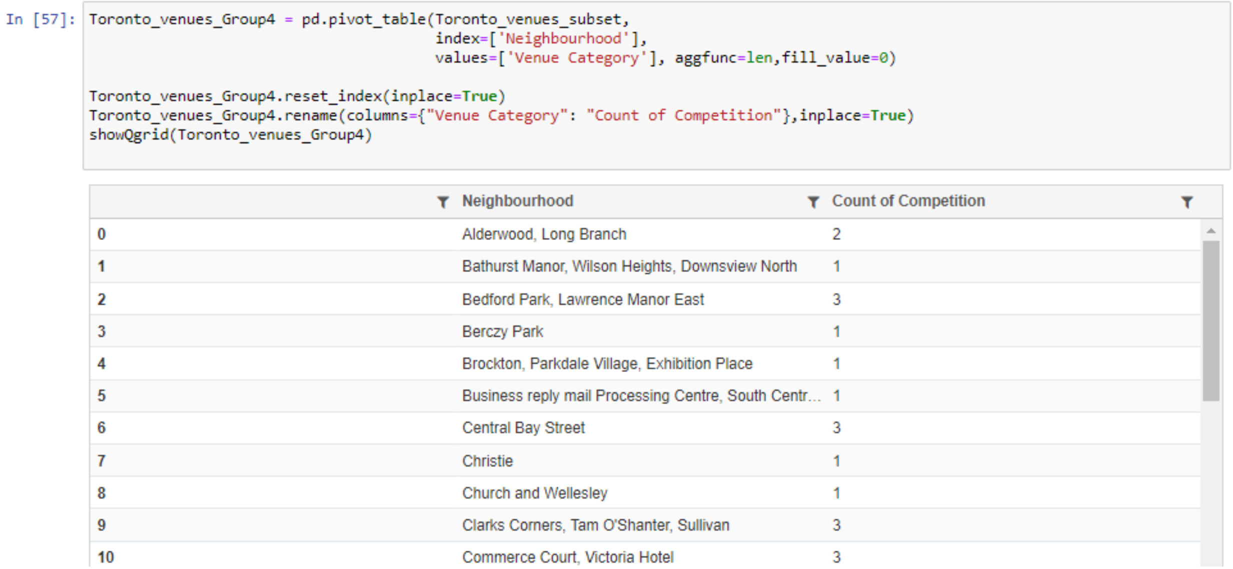
I then create a subset of the existing data frame which only returns the Neighbourhoods which have a “Pizza Place” or an “Italian Restaurant”



```
]: searchfor = ['Pizza Place', 'Italian Restaurant']
   Toronto_venues_subset = Toronto_venues[Toronto_venues['Venue Category'].str.contains('|'.join(searchfor))]

   showQgrid(Toronto_venues_subset)
```

	Neighbourhood	Neighbourhood	Neighbourhood	Venue	Venue Latitude	Venue Longitude	Venue Category
5	Victoria Village	43.72588	-79.31557	Pizza Nova	43.72582	-79.31286	Pizza Place
65	Queen's Park, O...	43.6623	-79.38949	Mercatto	43.66039	-79.38766	Italian Restaurant
108	Parkview Hill, W...	43.7064	-79.30994	Pizza Pizza	43.70516	-79.31313	Pizza Place
112	Parkview Hill, W...	43.7064	-79.30994	Venice Pizza	43.70592	-79.31396	Pizza Place
118	Garden District, ...	43.65716	-79.37894	Blaze Pizza	43.65652	-79.38002	Pizza Place
158	Garden District, ...	43.65716	-79.37894	Scaddabush Itali...	43.65892	-79.38289	Italian Restaurant
164	Garden District, ...	43.65716	-79.37894	Trattoria Mercatto	43.65445	-79.38097	Italian Restaurant
187	Garden District, ...	43.65716	-79.37894	Panago	43.65826	-79.38431	Pizza Place
217	Glencairn	43.70958	-79.44507	Pizza Nova	43.7075	-79.44314	Pizza Place
221	Don Mills	43.7259	-79.34092	Sorento Restaurant	43.72658	-79.34199	Italian Restaurant
250	St. James Town	43.65149	-79.37542	Terroni	43.65093	-79.3756	Italian Restaurant
316	St. James Town	43.65149	-79.37542	Mercatto	43.65024	-79.38082	Italian Restaurant
341	Eringate, Bloorda...	43.64352	-79.5772	Pizza Hut	43.64184	-79.57656	Pizza Place
381	Berczy Park	43.64477	-79.37331	The Old Spaghet...	43.64696	-79.3744	Italian Restaurant

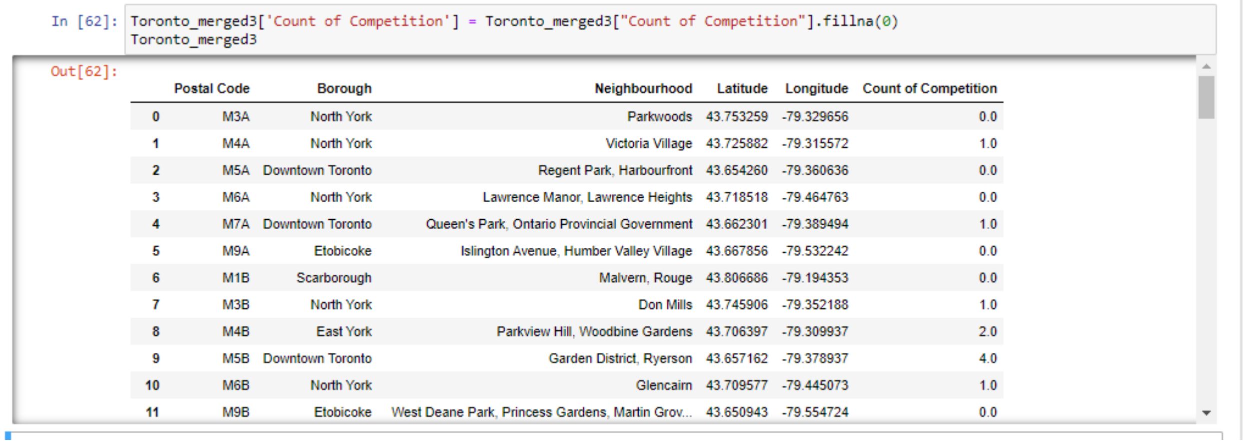
I then create a dataframe that has the count of Neighbourhoods with the count of total number of locations that are either “Pizza Place” or “Italian Restaurant” .



```
In [57]: Toronto_venues_Group4 = pd.pivot_table(Toronto_venues_subset,
                                                index=['Neighbourhood'],
                                                values=['Venue Category'], aggfunc=len,fill_value=0)

         Toronto_venues_Group4.reset_index(inplace=True)
         Toronto_venues_Group4.rename(columns={"Venue Category": "Count of Competition"},inplace=True)
         showQgrid(Toronto_venues_Group4)
```

	Neighbourhood	Count of Competition
0	Alderwood, Long Branch	2
1	Bathurst Manor, Wilson Heights, Downsview North	1
2	Bedford Park, Lawrence Manor East	3
3	Berczy Park	1
4	Brockton, Parkdale Village, Exhibition Place	1
5	Business reply mail Processing Centre, South Centr...	1
6	Central Bay Street	3
7	Christie	1
8	Church and Wellesley	1
9	Clarks Corners, Tam O'Shanter, Sullivan	3
10	Commerce Court, Victoria Hotel	3

I then join the table with the postal codes dataframe and replace the values for the non matching neighbourhoods with “0”. I will be using this dataframe to finally proceed with the clustering analysis.



```
In [62]: Toronto_merged3['Count of Competition'] = Toronto_merged3["Count of Competition"].fillna(0)
         Toronto_merged3
```

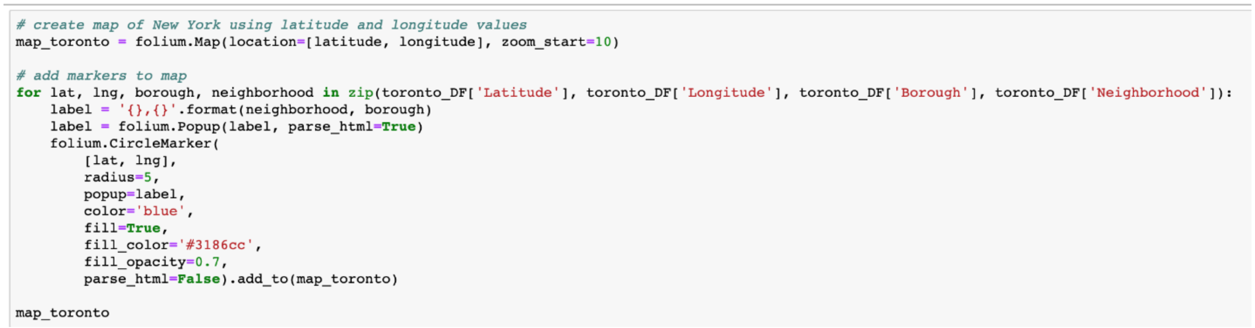
Out[62]:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	0.0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1.0
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	0.0
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	0.0
7	M3B	North York	Don Mills	43.745906	-79.352188	1.0
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937	2.0
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	4.0
10	M6B	North York	Glencairn	43.709577	-79.445073	1.0
11	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650943	-79.554724	0.0

3. Exploratory Data Analysis:

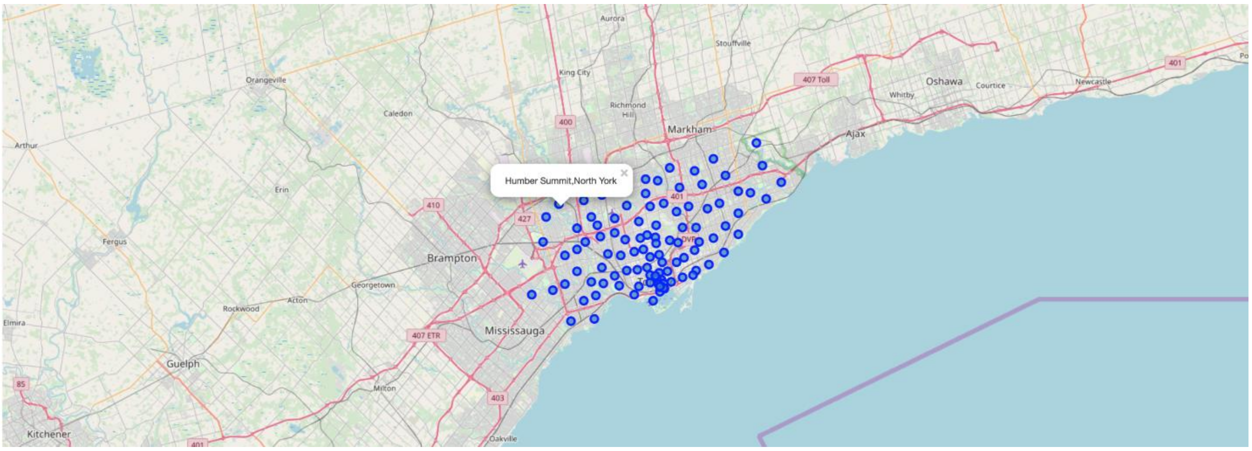
3.1 Folium Library and Leaflet Map

Folium is a python library, I’m using it to draw an interactive leaflet map using coordinate data.



```
# create map of New York using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, borough, neighborhood in zip(toronto_DF['Latitude'], toronto_DF['Longitude'], toronto_DF['Borough'], toronto_DF['Neighborhood']):
    label = '{},{}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

code to draw the folium map



Folium map of Toronto Neighborhood with popup label