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.

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.

e) Identify neighbourhoods with higher popularity in eating options

Using the foursquare data, we will be taking a count of venue locations marked as *Any Eatery*. This will help us understand how popular the destination is for eating options, and also help us better understand the size of the cachement population.

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

e) Highlight the potential candidates for Neighbourhoods with several eating options, highlighting that it could be a popular destination

```
'Sushi Restaurant',
'Taiwanese Restaurant',
'Theme Restaurant',
'Vegetarian / Vegan Restaurant',
'Vietnamese Restaurant']
```

In [23]:
```
Toronto_venues_subset2 = Toronto_venues[Toronto_venues['Venue Category'].str.contains('|'.join(eateries_list))]
```

In [25]:
```
Toronto_venues_Group3 = pd.pivot_table(Toronto_venues_subset2,
                                       index=['Neighbourhood'],
                                       values=['Venue Category'], aggfunc=len,fill_value=0)

Toronto_venues_Group3.reset_index(inplace=True)
Toronto_venues_Group3.rename(columns={"Venue Category": "Count of Eateries"},inplace=True)
showQgrid(Toronto_venues_Group3)
```

	Neighbourhood	Count of Eateries
0	Agincourt	2
1	Alderwood, Long Branch	5
2	Bathurst Manor, Wilson Heights, Downsview North	8
3	Bayview Village	3
4	Bedford Park, Lawrence Manor East	18
5	Berczy Park	38
6	Birch Cliff, Cliffside West	1
7	Brockton, Parkdale Village, Exhibition Place	12
8	Business reply mail Processing Centre, South Centr...	5
9	CN Tower, King and Spadina, Railway Lands, Harbo...	2
10	Canada Post Gateway Processing Centre	7

I then join the tables 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 [34]:
```
#The Final Dataframe ready to be clustered
Toronto_merged3 = df_loc.join(Toronto_venues_Group4.set_index('Neighbourhood'), on='Neighbourhood')
Toronto_merged3 = Toronto_merged3.join(Toronto_venues_Group3.set_index('Neighbourhood'), on='Neighbourhood')

Toronto_merged3['Count of Competition'] = Toronto_merged3["Count of Competition"].fillna(0)
Toronto_merged3['Count of Eateries'] = Toronto_merged3["Count of Eateries"].fillna(0)
showQgrid(Toronto_merged3)
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition	Count of Eateries
0	M3A	North York	Parkwoods	43.75326	-79.32966	0	1
1	M4A	North York	Victoria Village	43.72588	-79.31557	0	2
2	M5A	Downtown Toronto	Regent Park, Har...	43.65426	-79.36064	0	25
3	M6A	North York	Lawrence Manor,...	43.71852	-79.46476	0	2
4	M7A	Downtown Toronto	Queen's Park, O...	43.6623	-79.38949	1	22
5	M9A	Etobicoke	Islington Avenue,...	43.66786	-79.53224	1	1
6	M1B	Scarborough	Malvern, Rouge	43.80669	-79.19435	0	1
7	M3B	North York	Don Mills	43.74591	-79.35219	1	13
8	M4B	East York	Parkview Hill, W...	43.7064	-79.30994	2	3
9	M5B	Downtown Toronto	Garden District, ...	43.65716	-79.37894	4	48
10	M6B	North York	Glencairn	43.70958	-79.44507	1	5
11	M9B	Etobicoke	West Deane Par...	43.65094	-79.55472	0	0
12	M1C	Scarborough	Rouge Hill, Port ...	43.78454	-79.1605	0	1
13	M3C	North York	Don Mills	43.7259	-79.34092	1	13
14	M4C	East York	Woodbine Heights	43.69534	-79.31839	0	0

3. Exploratory Data Analysis:

3.1 Nature of the Data

Competition outlets range from a minimum of “0” to a maximum of 5 per location. With 50% of the data having no competition outlets at all.
All Eateries range from a minimum of “0” to a maximum of 70 per location. With 50% having 4 Eateries in Total.
Implications : Most of the venues have very low number of eateries in total, which may be perhaps be reflective of the “sparseness” of the population levels living in those neighbourhoods.

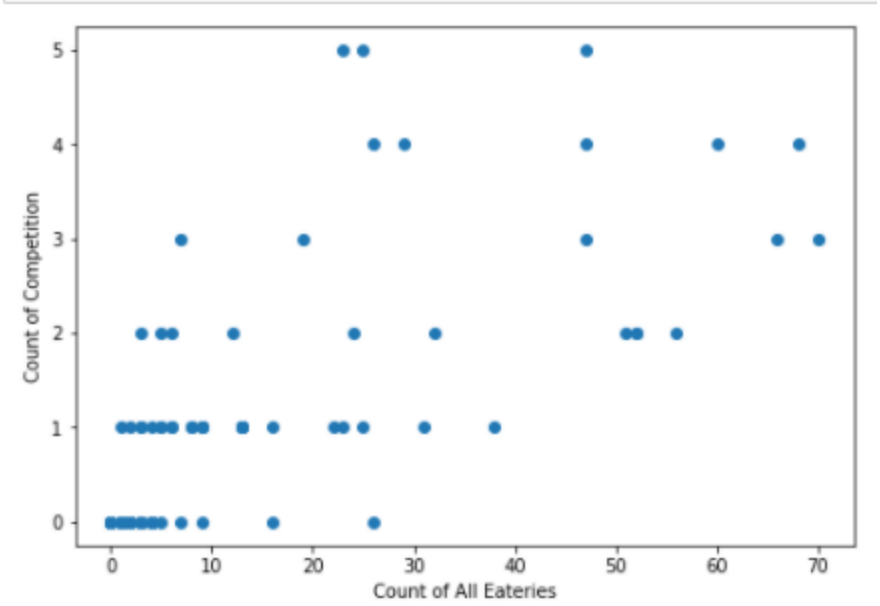
Also, we see that on average, each location contains 1 competition outlet and 12 eateries. Which means that Pizzerias / Italian restaurants represent 8% of all Total eateries.

	Latitude	Longitude	Count of Competition	Count of Eateries
count	103.000000	103.000000	103.000000	103.000000
mean	43.704608	-79.397153	0.922330	12.077670
std	0.052463	0.097146	1.318714	17.723746
min	43.602414	-79.615819	0.000000	0.000000
25%	43.660567	-79.464763	0.000000	1.000000
50%	43.696948	-79.388790	0.000000	4.000000
75%	43.745320	-79.340923	1.000000	14.500000
max	43.836125	-79.160497	5.000000	70.000000

Code:
```
Toronto_merged3.describe()
```

3.2 Relationship between Count of Competition and Count of All eateries

We generally see that locations with a higher number of Competition outlets [Pizzerias / Italian Restaurants] generally have a larger number of eateries. The suitable locations for our Pizzerias will be ones that have a higher presence of eating outlets but those that still do not have as many Competition outlets i.e. a sweetspot



code to draw the scatter plot

The Distribution of Data before the Clustering Analysis

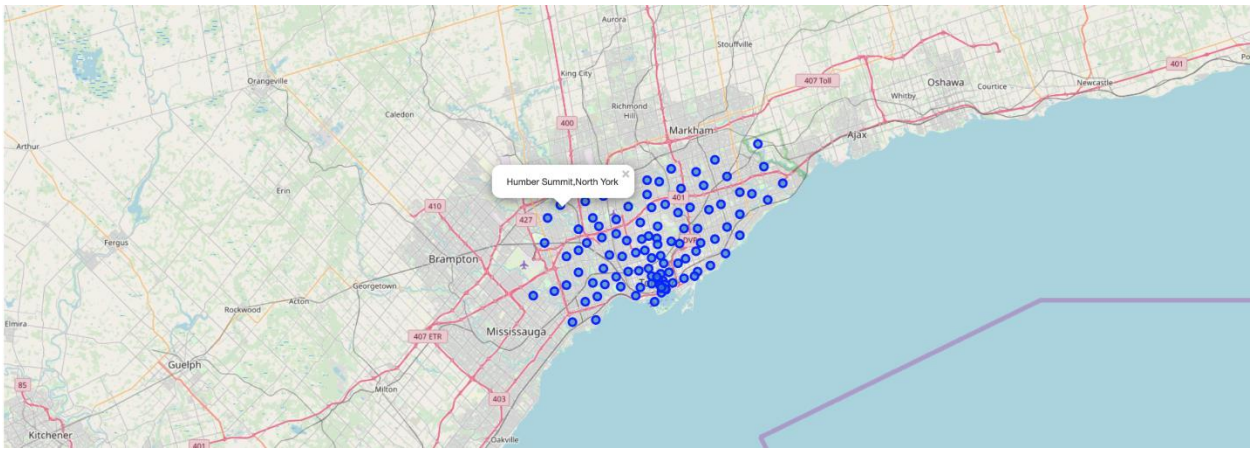
```
In [100]: fig=plt.figure()
          ax=fig.add_axes([0,0,1,1])

          y = Toronto_merged3['Count of Competition']
          x = Toronto_merged3['Count of Eateries']
          ax.set_ylabel('Count of Competition')
          ax.set_xlabel('Count of All Eateries')
          scatter = ax.scatter(x, y)
          plt.show()
```

3.3 Folium Library and Leaflet Map

Folium is a python library, I'm using it to draw an interactive leaflet map using coordinate data. We will be color coding the shortlisted locations later as suitable locations to open our Pizzeria.

```
# 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

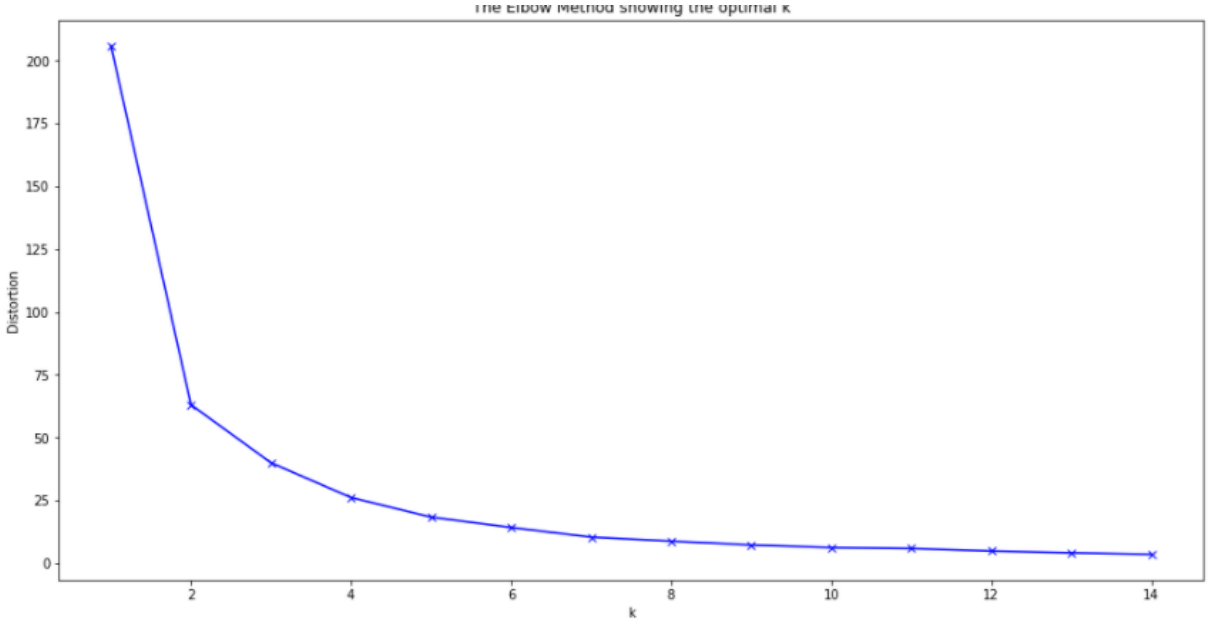


Folium map of Toronto Neighborhood with popup label

4. Predictive Modelling:

4.1 Clustering Neighborhoods of Toronto:

First step in K-means clustering is to identify best K value meaning the number of clusters in a given dataset. To do so we are going to use the elbow method on the Toronto dataset



Code snippet —

```
In [19]: from sklearn.preprocessing import StandardScaler
         from sklearn import preprocessing

         X = Toronto_merged3.iloc[: ,5:7]
         X.head()
```

Out[19]:

	Count of Competition	Count of Eateries
0	0.0	1.0
1	0.0	3.0
2	0.0	26.0
3	0.0	2.0
4	1.0	22.0

```
In [20]: X = preprocessing.StandardScaler().fit(X).transform(X)
```

3.3 Cluster the Neighborhoods

In [101]:
```
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import KMeans

distortions = []
K = range(1,15)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(X)
    distortions.append(kmeanModel.inertia_)

# Plot the elbow
plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

After analysing using elbow method using distortion score & Squared error for each K value, looks like K = 7 is the best value.

Clustering the Toronto Neighborhood Using K-Means with K =7

In [29]:
```
# set number of clusters
kmeanModel = KMeans(n_clusters=7)
kmeanModel.fit(X)

Toronto_merged3.insert(7, 'Cluster Labels', kmeanModel.labels_)
```

4.2 Examine the Clusters:

- Neighbourhoods with low number of eateries : Cluster 1,2,5
  Cluster 1: Contains the most number of locations in Toronto, but the average number of eateries in those locations is very low. Likely reflective of the size of the populations in those locations.
  While cluster 2 and 8 are better off, on average they contain less than 10 eateries, which again represents a low business opportunity.

- Neighbourhoods with higher level of competition : Cluster 3,4.
  Cluster 3 particularly his is the cluster that contains the most number of competition outliers – with an average of 4.5 per location.

While Cluster 4 has a better ratio of competition outlets to eateries (5.8%), each location already contains a minimum number of 3 competition outlets. Which is not ideal for us.

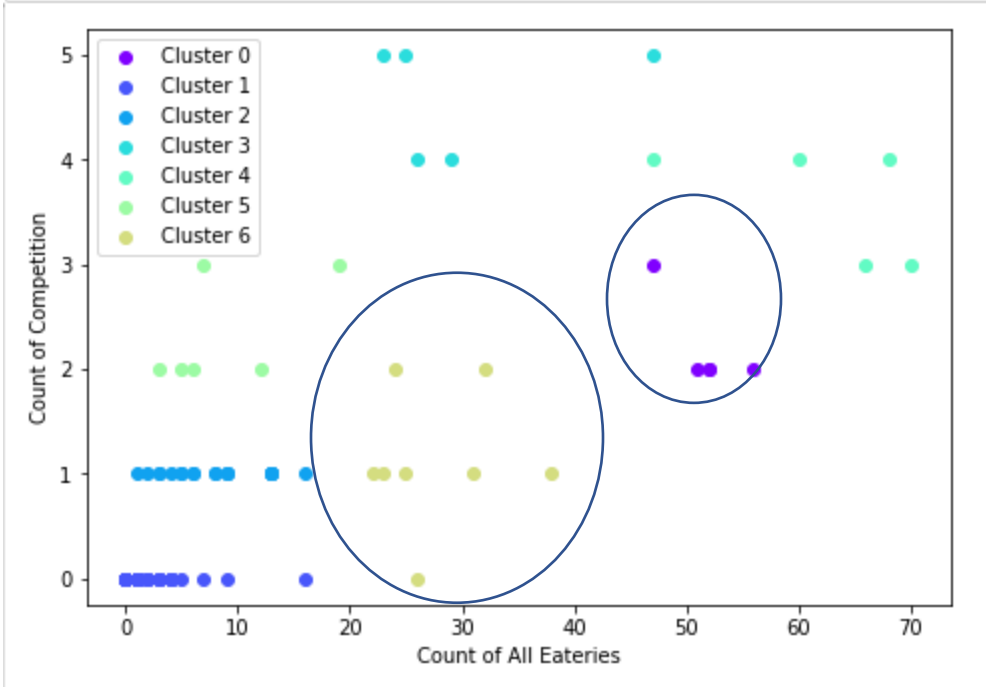
- Optimal Neighbourhoods with high level of eateries and low competition : Cluster 0,6
  In both of these clusters, the Avg # of Competition outlets / Avg # of Total Eateries is 4%. Which is half the overall average of 8%. So quite optimal.

	Count of Competition				Count of Eateries			
	mean	count	min	max	mean	count	min	max
Cluster Labels								
0	2.200000	5	2.0	3.0	51.600000	5	47.0	56.0
1	0.000000	54	0.0	0.0	1.722222	54	0.0	16.0
2	1.000000	20	1.0	1.0	7.950000	20	1.0	16.0
3	4.600000	5	4.0	5.0	30.000000	5	23.0	47.0
4	3.600000	5	3.0	4.0	62.200000	5	47.0	70.0
5	2.333333	6	2.0	3.0	8.666667	6	3.0	19.0
6	1.125000	8	0.0	2.0	27.625000	8	22.0	38.0

Code snippet –

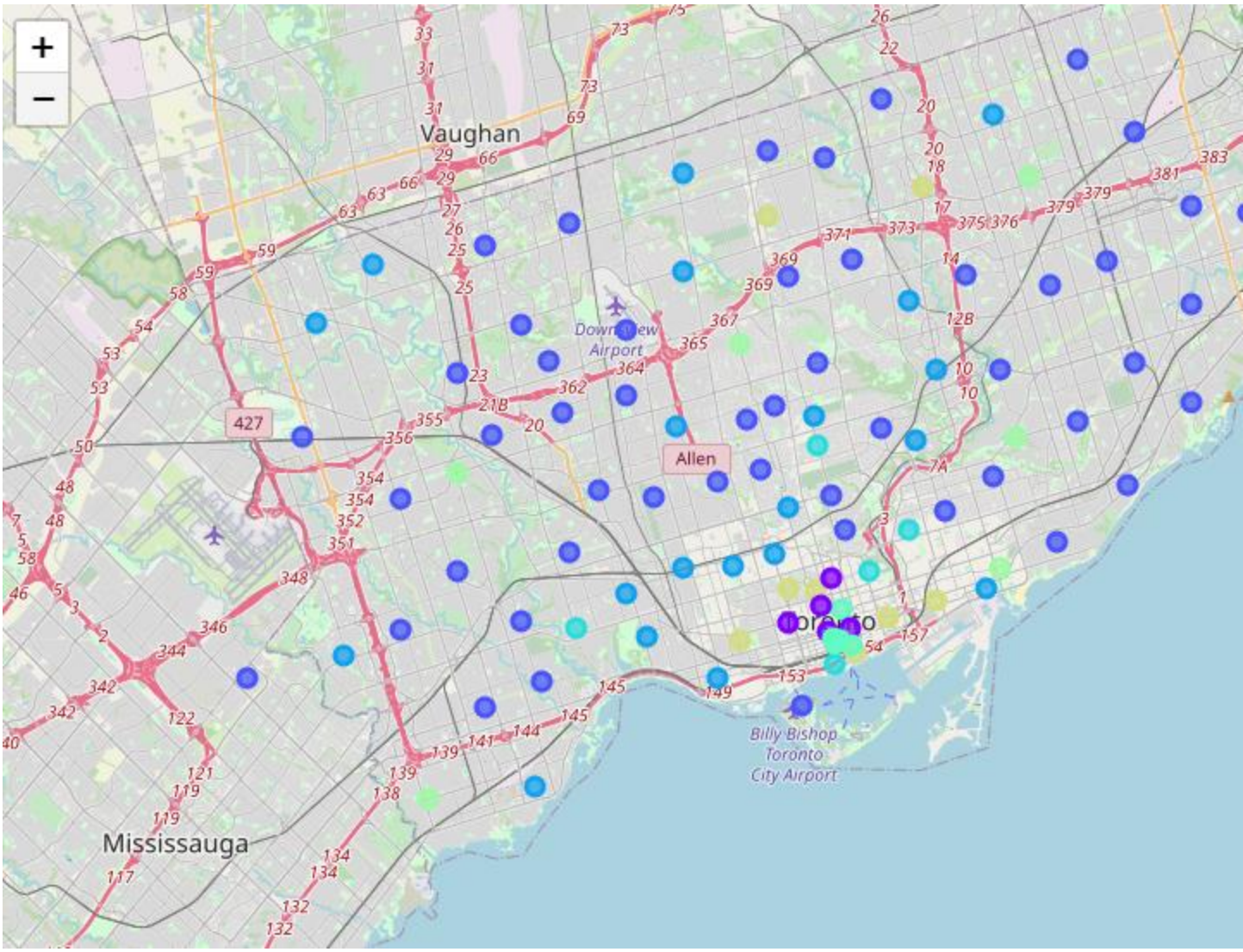
Distribution of Clusters

```
Toronto_merged3[['Cluster Labels',
                 'Count of Competition',
                 'Count of Eateries']].groupby('Cluster Labels').agg(['mean', 'count','min','max'])
```



Code snippet

```
fig=plt.figure()
ax=fig.add_axes([0,0,1,1])

# set color scheme for the clusters
x = np.arange(k)
ys = [i + x + (i*x)**2 for i in range(10)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

for k in range(7):
    Toronto_Temp = Toronto_merged3.loc[Toronto_merged3['Cluster Labels'] == k]
    y = Toronto_Temp['Count of Competition']
    x = Toronto_Temp['Count of Eateries']
    colorlabel = int(k)
    legendlabel = "Cluster " + str(k)
    scatter = ax.scatter(x, y, color=rainbow[colorlabel],label=legendlabel)

ax.set_ylabel('Count of Competition')
ax.set_xlabel('Count of All Eateries')

ax.legend()

plt.show()
```



4.3 A Deeper Look into the Optimal Clusters:

Both these clusters represent good potential. The major difference in both these clusters is the Total number of overall eateries present in those sets of clusters. Cluster 0 avg is 52 eateries and Cluster 6 average is 28 eateries. So the average for both of these is much higher than overall which is quite good.

The Optimum Locations Selected to Open the Pizzeria

```
Toronto_merged3.loc[Toronto_merged3['Cluster Labels'] == 0]
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition	Count of Eateries	Cluster La
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2.0	52.0	
24	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	3.0	47.0	
30	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	2.0	56.0	
84	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	2.0	51.0	
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	2.0	52.0	

```
Toronto_merged3.loc[Toronto_merged3['Cluster Labels'] == 6]
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition	Count of Eateries	Cluster L
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0	26.0	
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1.0	22.0	
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	1.0	38.0	
33	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556	1.0	31.0	
37	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	2.0	32.0	
54	M4M	East Toronto	Studio District	43.659526	-79.340923	1.0	23.0	
59	M2N	North York	Willowdale, Willowdale East	43.770120	-79.408493	2.0	24.0	
80	M5S	Downtown Toronto	University of Toronto, Harbord	43.662696	-79.400049	1.0	25.0	

5. Results and Discussion:

5.1 Results

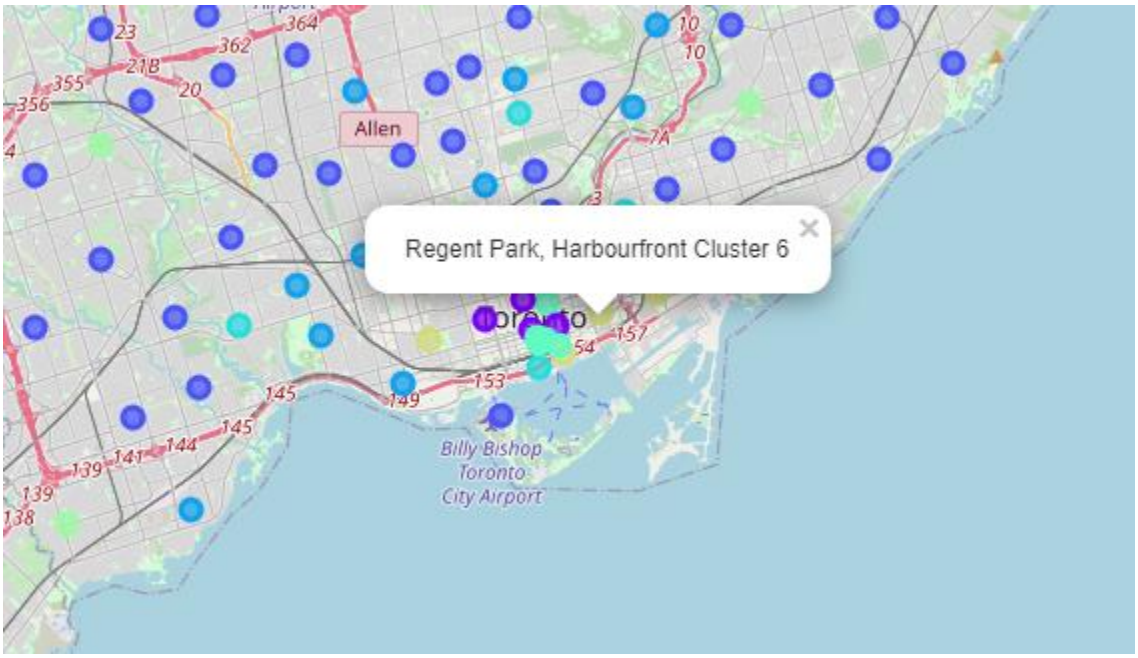
We have reached the end of the analysis, in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighborhood to open a new Pizzeria. To achieve that we looked into all the neighborhoods in Toronto, analysed the neighborhood for competition venues and total number of eateries in each neighborhood to come to conclusion about which neighborhood would be a better spot. We have used variety of data sources to set up a very realistic data-analysis scenario. We have found out that —

- On average, 8% of eateries in Toronto Neighbourhoods are competition outlets [Pizza Places , Italian Restaurant]
- The cluster analysis suggests that there are 3 types of Neighbourhoods
    - Sparse Neighbourhoods – Ones with less than 10 eateries on average
    - High Competition Neighbourhoods – ones where there is already sufficient presence of Competition outlets.
    - Optimal Neighbourhoods – Where the ratio of competition outlets is 4% (half the the overall average) and the average number of eateries is more than 25 (More than double the overall average)
- The Optimal Neighbourhoods represent two clusters which have varying degrees of total number of eateries, 28 and 52 on average. While both represent good choices; since we are a new brand, it might make more sense to choose those Neighbourhoods where there is only 1 competition outlet at most.
- Assuming that the number of eateries in a location is indicative of population, the lower number of eateries might suggest a smaller size of a outlet that we might have to open. Thus also reflective of a lower investment cost. So cluster 6 might be better in that regard.
- After careful consideration of Cluster 6 locations, it maybe a good idea to open a new Pizzeria in Regent Park, Harbourfront since it has high number of Eateries and not a single competition outlet, which likely gives it a greater probability of success.

5.2 Discussion

According to this analysis, Regent park, Harbourfront will provide the best location for the new Pizzeria. As it has no competition, and also a larger potential market for its product. We can then potentially expand into other neighbourhoods mentioned as potential within the optimal cluster selections. The drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API and the data about the restaurants in each neighborhood which may not be not up-to

date. We are also not looking into several other dimensions of determining the selection of location i.e. availability of sourcing, local legislation, land cost etc.



6. Conclusion:

We have got a chance to on a business problem like how a real like data scientists would do. We have used many python libraries to fetch the data, to manipulate the contents & to analyze and visualize those datasets. We have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized using various plots present matplotlib. We also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Hopefully, this project helps acts as initial guidance to take more complex real-life challenges using data-science. There are lots of areas where this analysis can be improved. But this analysis has certainly has given us a head start into the business problem and start discussions with the relevant stakeholders.