FINAL REPORT

CAPSTONE PROJECT- BATTLE OF NEIGBORHOODS

INTRODUCTION

Toronto is on of the most populous cities in Canada. Toronto's demographics show that it is large and ethnically diverse. With its diverse culture comes diverse food item. There are a wide variety of restaurants in Toronto, each belonging categories such as Chinese, Indian, French, Jamaican, Portuguese etc.

In this project we would go through a series of steps to determine whether or not it is a good idea to setup an Italian restaurant in Toronto and if what areas/neighborhoods would be the most profitable for the restaurant. The success of a restaurant depends on the customers and so it is important to cater to the right crowd. Toronto is home to the vast majority of the Italian community in Canada. Toronto is also home to the forth-largest Italian community outside of Italy, behind São Paulo, Brazil, Buenos Aires, Argentina, and New York City, respectively so it already sounds like a good idea to setup a restaurant in Toronto. However, we have to be sure whether it would a profitable idea.

Toronto's diversity is reflected in Toronto's ethnic neighborhoods such as Chinatown, Corso Italia, Greektown, Kensington Market, Koreatown, Little India, Little Italy, Little Jamaica, Little Portugal & Roncesvalles.

PROBLEM:

1. List and visualize all parts of Toronto that have Italian restaurants.
2. What is the best location in Toronto for Italian cuisine?
3. Which areas have potential Italian restaurant market?
4. Which areas lack Italian Restaurants?
5. Which is the best neighborhood to stay in if your preference of food is Italian cuisine?

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 an Indian restaurant in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting the Indian crowd.
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. Indian crowd who wants to find neighborhoods with lots of option for Indian restaurants.
4. 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.

DATA SECTION

For this project, I would be scraping a list of Canada postal codes from the from the following website: **(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)**. This will provide me with the postal code, borough & the name of all the neighborhoods present in Toronto.

Afterwards, I would leverage the following website: (**https://cocl.us/Geospatial_data)** csv file to get all the geographical coordinates of the neighborhoods. This includes a list of **Boroughs, Neighborhoods, postal codes, latitudes and longitudes.**

I would need to get information about various venues in Toronto using Foursquare's explore API (https://developer.foursquare.com/docs).

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.

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.

· 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.

```
In [113]: import pandas as pd
          import requests
          import numpy as np

          url_html='https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050'

          df = pd.read_html(url_html)

          df_postcodes=df[0]

          print("imported dataframe has",df_postcodes['Postcode'].count(), "postcodes entries")

          df_postcodes.head(5)

          imported dataframe has 287 postcodes entries
```

Out[113]:

	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

Importing geospatial data of the neighborhoods in Toronto.

```
In [49]: lat_lon = pd.read_csv('https://cocl.us/Geospatial_data')
         lat_lon.head()
```

Out[49]:

	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

Next I merge both dataframes forming the one below:

```
In [50]: lat_lon.rename(columns={'Postal Code':'Postcode'},inplace=True)
         df3 = pd.merge(df2,lat_lon,on='Postcode')
         df3.head()
```

Out[50]:

	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

Get location data using Foursquare API

Foursquare is an online application that has been utilized by many developers and applications such as Uber amongst others. 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 we need to turn that into a data-frame. Here I've chosen 100 popular spots for each neighborhood within a radius of 1km.

Getting all the nearby venues for all the locations in the data frame

```
In [42]: toronto_venues = getNearbyVenues(names=df3["Neighborhood"],
                                          latitudes=df3['Latitude'],
                                          longitudes=df3['Longitude']
                                         )

         toronto_venues.head(10)
```

```
ylea
Kingsway Park South West, Mimico NW, The Queensway West, Royal York South West, South of Bloor
```

Out[42]:

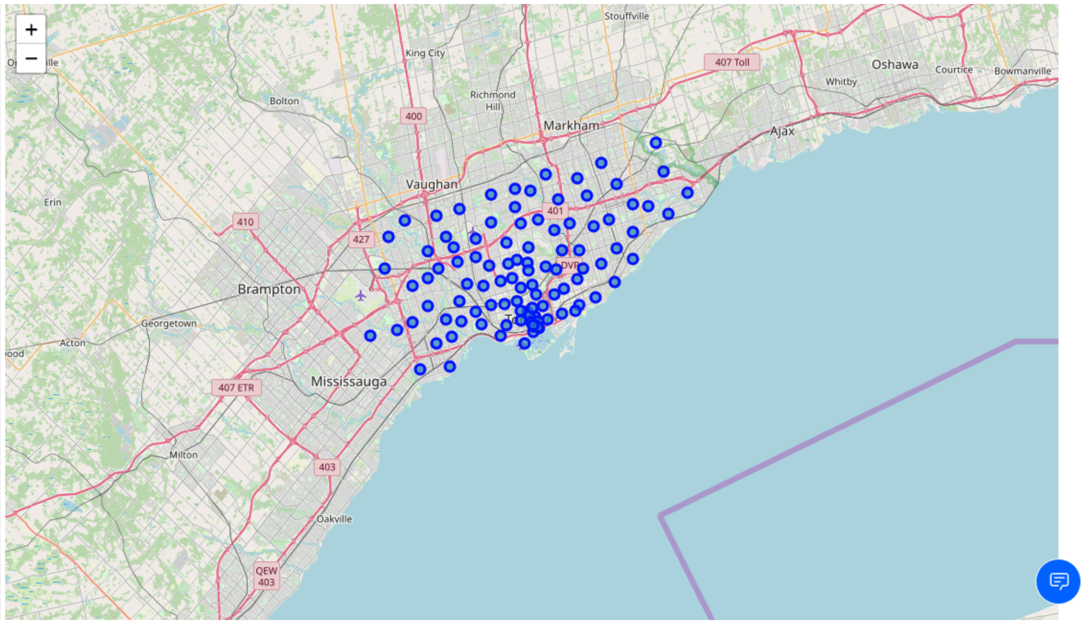
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	649 Variety	43.754513	-79.331942	Convenience Store
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
5	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
6	Victoria Village	43.725882	-79.315572	Eglinton Ave E & Sloane Ave/Bermondsey Rd	43.726086	-79.313620	Intersection
7	Harbourfront	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
8	Harbourfront	43.654260	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop

Exploratory Data Analysis

The python library, Folium, is utilized to create an interactive leaflet map using coordinate data.

```
# create map of New York using latitude and longitude values
map_toronto = folium.Map(location=[43.6532, -79.3832], zoom_start=10)

# add markers to map
for lat, lng, borough, neighborhood in zip(df3['Latitude'], df3['Longitude'], df3['Borough'], df3['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
```

Out[67]:



RELATIONSHIP BETWEEN NEIGHBORHOOD AND ITALIAN RESTAURANTS

The first course of action would be to extract the Neighborhood and Italian Restaurant column from the above Toronto dataframe for further analysis:

```
In [45]: toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

         toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

         fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
         toronto_onehot = toronto_onehot[fixed_columns]
         toronto_grouped = toronto_onehot.groupby('Neighborhood').mean().reset_index()
         toronto_grouped
```

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Turkish Restaurant	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store
0	Adelaide, King, Richmond	0.000000	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	...	0.0	0.020000	0.00	0.000000	0.000000	0.00
1	Agincourt	0.000000	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	...	0.0	0.000000	0.00	0.000000	0.000000	0.00
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.000000	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	...	0.0	0.000000	0.00	0.000000	0.000000	0.00
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.000000	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	...	0.0	0.000000	0.00	0.000000	0.000000	0.00
4	Alderwood, Long Branch	0.000000	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	...	0.0	0.000000	0.00	0.000000	0.000000	0.00

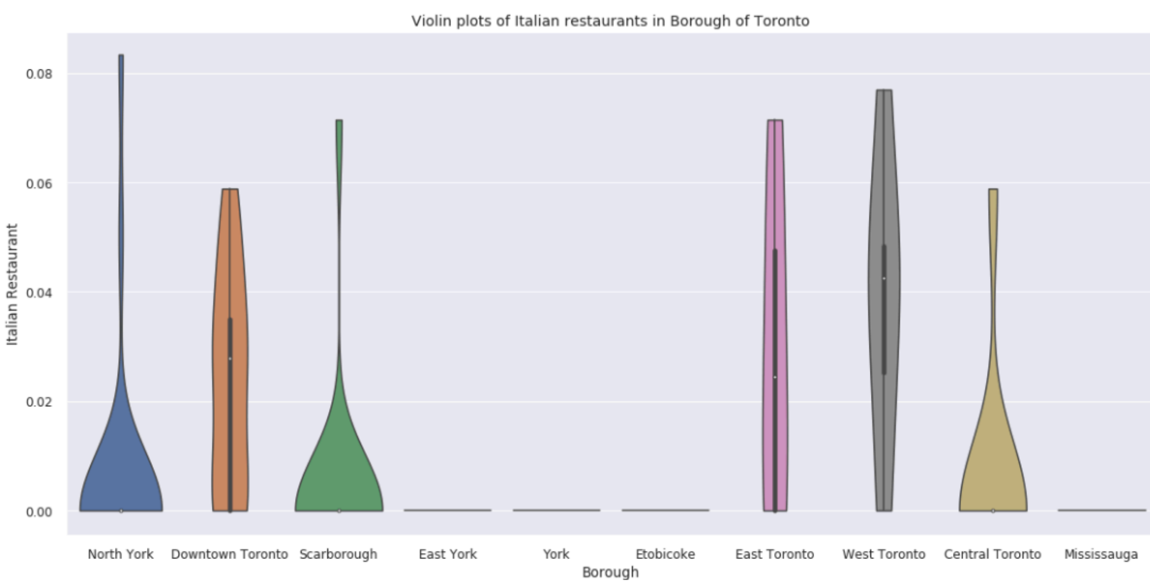
The next thing to do after using pandas one-hot encoding is to merge the dataframe with the Toronto dataframe with neighborhoods latitude and longitude information.

```
In [61]: toronto_merged = pd.merge(df3, toronto_part, on='Neighborhood')
         toronto_merged
```

Out[61]:

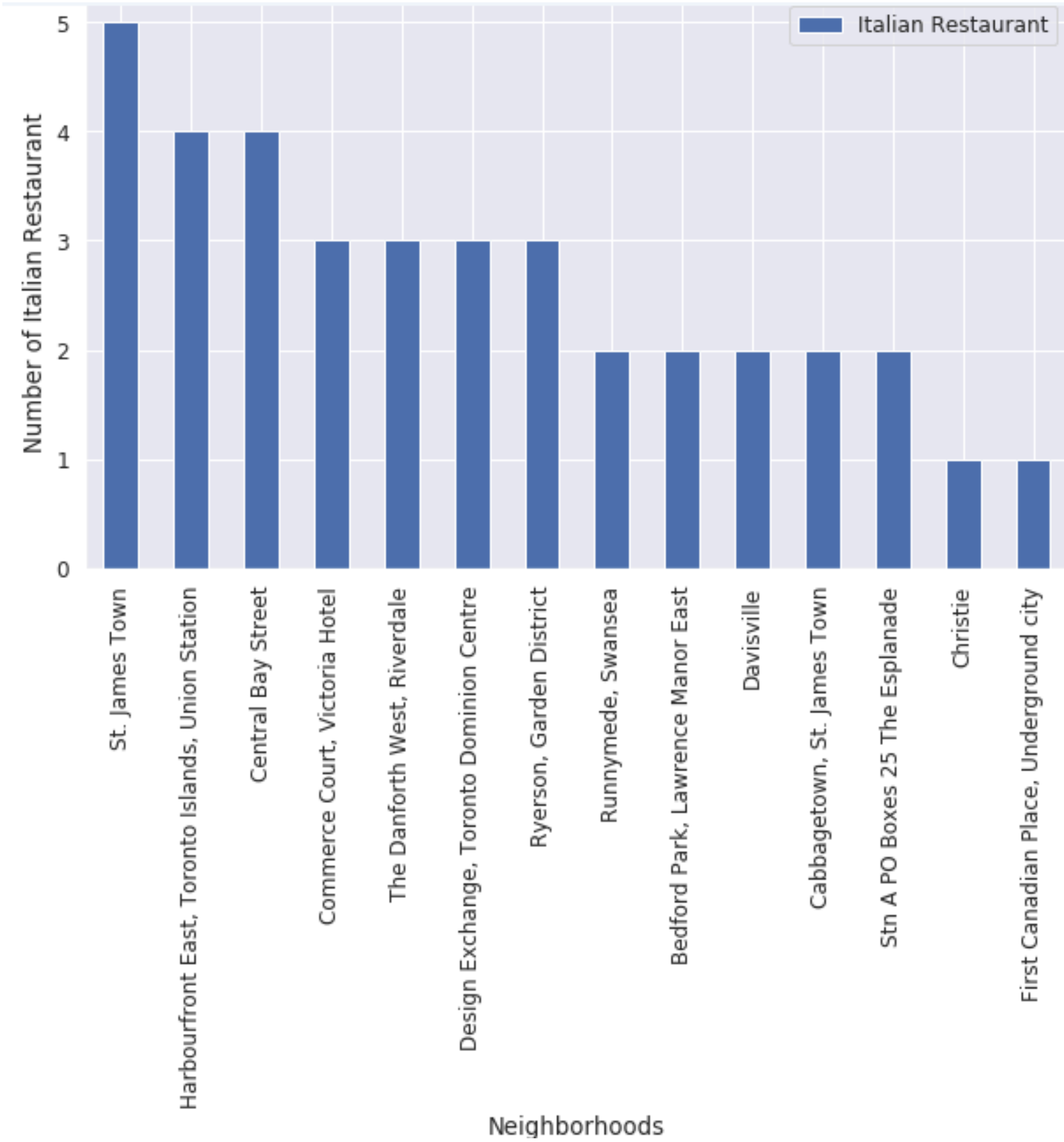
	Postcode	Borough	Neighborhood	Latitude	Longitude	Italian Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.000000
1	M4A	North York	Victoria Village	43.725882	-79.315572	0.000000
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	0.000000
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	0.000000
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	0.027778
5	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	0.000000
6	M3B	North York	Don Mills North	43.745906	-79.352188	0.000000
7	M4B	East York	Woodbine Gardens, Parkview Hill	43.706397	-79.309937	0.000000
8	M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937	0.030000
9	M6B	North York	Glencairn	43.709577	-79.445073	0.000000
10	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0.000000

Now we are going to visualize the above dataframe by drawing some plots. The seaborn library will be used for the following:



We can see the distribution of Italian restaurants in different Boroughs. The plot helps identify the Boroughs with densely populated Italian restaurants.

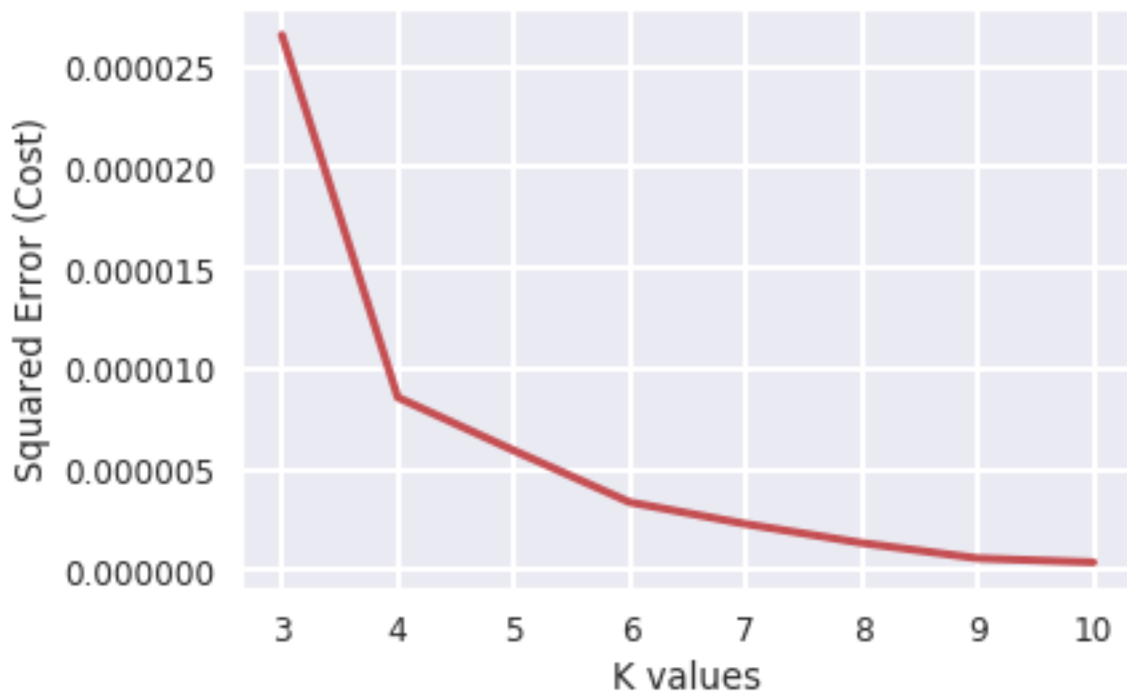
Now that we have visualized the distribution of Italian restaurants In different boroughs, lets visualize the neighborhoods with Italian restaurants:



We can see that the neighborhoods with the highest number of Italian restaurants are St. James Town, Habourfront eat, Toronto islands, Union Station, Central Bay Street, Commerce Count, Victoria Hotel etc. all ranged from 3-5 Italian restaurants in each neighborhood.

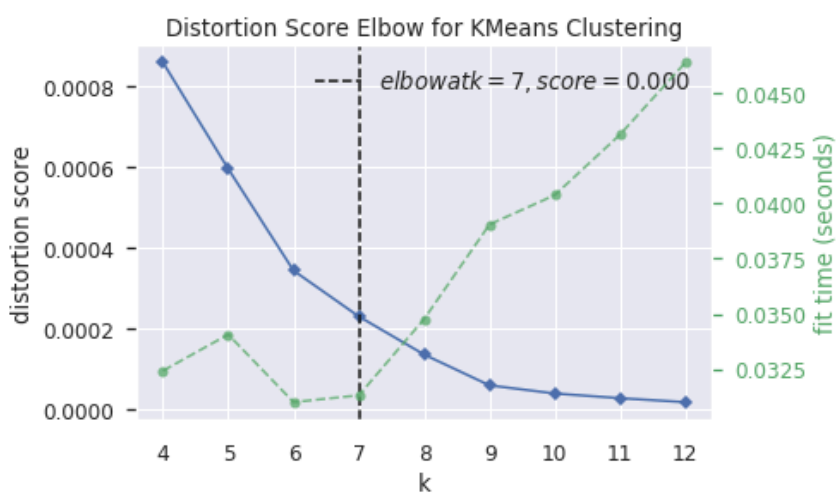
Predictive Modeling

Here we would be clusterig the neigborhoods 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 with Italian restaurant percentage (i.e. toronto_merged dataframe).



```python
from sklearn.cluster import KMeans

toronto_part_clustering = toronto_part.drop('Neighborhood', 1)


error_cost = []

for i in range(3,11):
    KM = KMeans(n_clusters = i, max_iter = 100)
    try:
        KM.fit(toronto_part_clustering)
    except ValueError:
        print("error on line",i)




    #calculate squared error for the clustered points
    error_cost.append(KM.inertia_/100)

#plot the K values aganist the squared error cost
plt.plot(range(3,11), error_cost, color='r', linewidth='3')
plt.xlabel('K values')
plt.ylabel('Squared Error (Cost)')
plt.grid(color='white', linestyle='-', linewidth=2)
plt.show()
```

```python
# Instantiate the clustering model and visualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(4,13))

visualizer.fit(toronto_part_clustering)        # Fit the data to the visualizer
visualizer.show()        # Finalize and render the figure
```



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

CLUSTERING THE TORONTO NEIBORHOOD USING K-MEANS WITH K=7

In [34]:

```
kclusters = 7

toronto_part_clustering = toronto_part.drop('Neighborhood', 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_part_clustering)

kmeans.labels_
```

Out[34]:

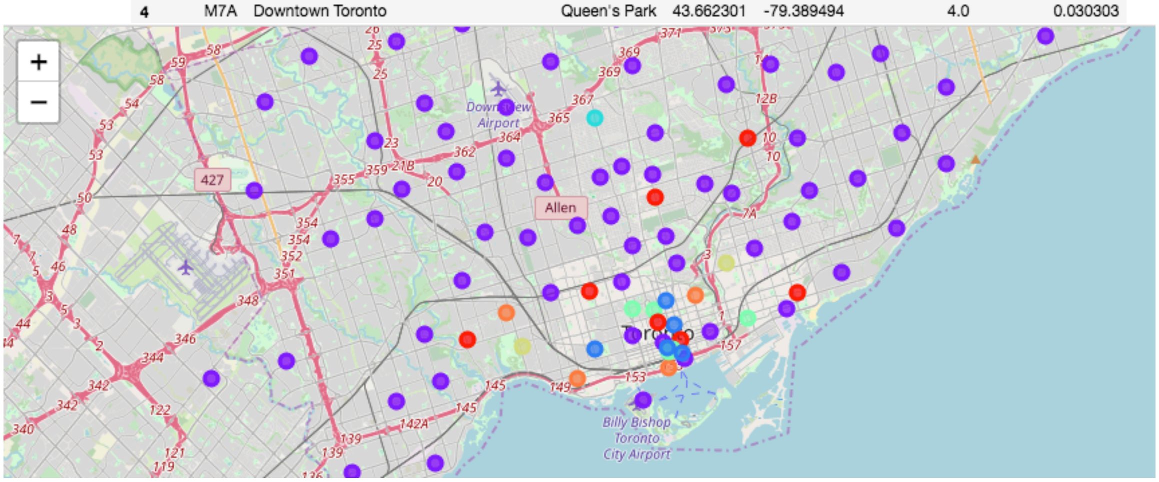
```
array([1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 6, 1, 1, 1, 6, 1, 1, 1, 0, 1, 0,
       2, 1, 5, 1, 4, 0, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,
       0, 1, 1, 1, 4, 1, 6, 6, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,
       1, 1, 1, 1, 5, 1, 4, 1, 1, 1, 0, 2, 1, 0, 2, 4, 1, 1, 0, 5, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

In [35]:

```
#sorted_neighborhoods_venues.drop(['Cluster Labels'],axis=1,inplace=True)
toronto_part.insert(0, 'Cluster Labels', kmeans.labels_)
toronto_merged = df3
# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
toronto_merged = toronto_merged.join(toronto_part.set_index('Neighborhood'), on='Neighborhood')
toronto_merged.dropna(subset=["Cluster Labels"], axis=0, inplace=True)
toronto_merged.reset_index(drop=True, inplace=True)
toronto_merged['Cluster Labels'].astype(int)
toronto_merged.head()
```

Out[35]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	1.0	0.000000
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0	0.000000
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	1.0	0.000000
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	1.0	0.000000
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	4.0	0.030303



Examining the clusters

We have a total of 7 clusters, which are 0,1,2,3,4,5,6.

Cluster 0 contains all the neighborhoods which has least number of Italian restaurants. It is shown in red color in the map

In [37]:
```
#cluster 0
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0]
```

Out[37]:

Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant

Cluster 1 contains all the neighborhoods which has least number of Italian restaurants.These neigbhorhoods have no Italian restaurants. It is shown in purple color on the map

In [81]:
```
#cluster 1
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1]
```

Out[81]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	1.0	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0	0.0
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	1.0	0.0
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	1.0	0.0
5	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	1.0	0.0

Cluster 2 contains all the neighborhoods which are densely populated with Italian restaurants. It is shown in dark blue color on the ma

In [82]:
```
#cluster 2
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2]
```

Out[82]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
39	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	2.0	0.071429
52	M5M	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750	2.0	0.083333
71	M6R	West Toronto	Parkdale, Roncesvalles	43.648960	-79.456325	2.0	0.076923
78	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter	43.781638	-79.304302	2.0	0.071429

Cluster 3 contains all the neighborhoods which are sparsely populated with Italian restaurants. It is shown in light blue color on the ma

In [83]:
```
#cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3]
```

Out[83]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	3.0	0.027778
8	M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937	3.0	0.030000
40	M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576	3.0	0.030000
46	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	3.0	0.030000

Cluster 4 contains all the neighborhoods which are medium populated with Italian restaurants. It is shown in light green color on the map.

In [84]:
```
#cluster 4
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4]
```

Out[84]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
23	M6G	Downtown Toronto	Christie	43.669542	-79.422564	4.0	0.058824
75	M4S	Central Toronto	Davisville	43.704324	-79.388790	4.0	0.058824

Cluster 5 contains all the neighborhoods which are also sparsely populated with Italian restaurants like cluster 3 but even more sparse. It is shown in a darker shade of green color on the map

In [85]:
```
#cluster 5
toronto_merged.loc[toronto_merged['Cluster Labels'] == 5]
```

Out[85]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
35	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	5.0	0.019608
88	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	5.0	0.020833
92	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	5.0	0.010000
94	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	5.0	0.012048

Cluster 6 contains all the neighborhoods which are also medium populated with Italian restaurants like cluster 4. It is shown in orange color on the map

In [86]:
```
#cluster 6
toronto_merged.loc[toronto_merged['Cluster Labels'] == 6]
```

Out[86]:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
11	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923	6.0	0.050000
13	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	6.0	0.050000
22	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	6.0	0.051948
45	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	6.0	0.047619
77	M6S	West Toronto	Runnymede, Swansea	43.651571	-79.484450	6.0	0.050000

RESULTS

At this point we have concluded the analysis, in this section we will put together all the findings from the above clustering and visualization of the dataset. At the start of this project, we began with the business problem of identifying a good neighborhood to open an Italian restaurant. In order to reach this goal we have looked into all the neighborhoods in Toronto, analyzed the relationship between neighborhoods and Italian Restaurants and the number of Italian restaurants in each neighborhood to come to conclusion about which neighborhood would be a good place to open an Italian restaurant. After making use of several data sources and analysis of the data, our observations are as follows:

- Out of the 11 boroughs in Toronto we identified that 6 of the 10 were densely populated with Italian restaurants. The violin plot helped us view the number of Italian restaurants in each borough.
- With the help of the cluster examination and the violin plot we have identified that North York, Downtown Toronto, Scarborough, East Toronto, West Toronto and Central Toronto are densely populated with Italian restaurants.
- Downtown Toronto has potential Italian Restaurant market seeing as it has a high number of Italian restaurants but not as much as boroughs like North York, West Toronto, and Scarborough etc. where the Italian restaurant market is saturated with too much competition for a new restaurant.
- The neighborhood with the highest number of Italian restaurants is St. James Town, having 5 Italian restaurants and the neighborhood with the lowest number of Italian restaurants are Christie, First Canadian Place and Underground City, having just 1.
- For people who have a taste for Italian Cuisine wants a wide variety of Italian Cuisine North York would be the ideal location. North York's neighborhoods are the most densely populated with Italian restaurants.
- The places with the least amount of Italian restaurants are East York, York, Etobicoke and Mississauga.

CONCLUSION

It is exciting to finally reach the end of this project having been exposed to business problems like real life Data Scientists. Through the course of this project I have made use of many python libraries to fetch data, manipulate data and to analyze and visualize real datasets. I have also utilized the Foursquare API to explore all the venues in the neighborhoods of Toronto and also scraped data from Wikipedia and visualized the data using the seasborn and matplotlib python libraries. Clustering, a machine learning technique was used to make predictions give the data and also the folium library was used to create maps.

Through the course of this project I discovered gaps in my problem solution. There is always room for improvement and the improvements would be in the form of more data and more machine learning techniques to provide more accurate results that are fitting for the real world. For future enhancements, population data would help provide more insights that would help develop better analysis for more informed decisions. This project has been a big step into my Data Science future and hopefully we can make the world a better place through informed decisions.