Battle of the Neighbourhoods - Determining the Best Location

Tsungai Mashava

January 16, 2020

1. Introduction

Background and Business Problem

Toronto is the provincial capital of Ontario and also the most populous city in Canada, with a population of 2,731,571 as per 2016 census. Toronto is a city that is rich in history, full of interesting events and cultural ethnicities that have made the city great. The diverse population of Toronto reflects its 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. Toronto is also popularly known as one of the biggest entertainment hubs of the country with a variety of bars, theatres, and restaurants representing a plethora of ethnic cultures. Among these bustling ethnic cultures, are the Italian Canadians, which Canada's Official Statistical office revealed that the Italians were the 6th largest ethnic group in Canada constituting 1,587,970 Canadians with full or partial Italian descent or 4.6% of the country's total population.

A significant part of the Italian heritage enjoyed world over is their Italian cuisine. Italian cuisine is known for its regional diversity, especially between the north and the south of the Italian peninsula. It offers an abundance of taste, and is one of the most popular and copied in the world. It influenced several cuisines around the world, chiefly that of North America. Italian cuisine is generally characterized by its simplicity, with many dishes having only two to four main ingredients. Italian cooks rely chiefly on the quality of the ingredients rather than on elaborate preparation. Given the significant presence of the Italian community and Toronto being the entertainment hub of Canada, Toronto presents a great setting to open an Italian Restaurants. This project will go through step by step process to make a decision whether it is a good idea to open an Italian restaurant. The neighborhoods in Toronto will be analysed to identify the most profitable areas since the success of the restaurant depends on factors such as the presence of a vibrant Italian community and the presence similar establishments in the area.

Target Audience

Audiences that would be interested in this project and the types of clients or groups who stand to benefit from this detailed analysis may include:

1. Existing restaurant owners or restaurant chain owners looking to expand into the Canadian market can benefit from gaining insights through targeted placement of new branches that target the Italian community.
2. Prospective business owners looking to break into the hospitality industry and start their own business can benefit from insights derived from assessing the risks and benefits of certain locations.
3. Italian restaurant enthusiasts who wish to find neighborhoods with options for Italian cuisine.

2. Data Acquisition and Cleaning

To carry out the analysis, secondary data was collected from different sources. The datasets used are as follows:

- For data on the neighborhoods of Toronto, I used "List of Postal code of Canada: M" which was web scraped using wikipedia link https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The page contains the name of each neighborhood including the postal code and borough which I then wrangle, clean and read into a *pandas* dataframe so that it is structured. This dataframe provides the basis of the analysis and give an understanding of the Toronto landscape.

- For the dataset above to be complete, it required geographical coordinates to the mapped to each postal code. The geographical coordinates are pertinent to establishing clusters and the venue available in each cluster. For information on the geographical coordinates of the Toronto neighborhoods I used the csv file - https://cocl.us/Geospatial_data.

- Given that the project requires that we identify the most suitable areas to for an Italian restaurant, the dataset required information on the ethnic distribution of the Toronto population to identify areas that are more densely populated by Italian communities. To obtain this information I downloaded Toronto's neighborhood profile from the city of Toronto's Open Data portal - https://open.toronto.ca/dataset/neighbourhood-profiles/. The CSV file I downloaded has information on a wide range of topics including household income, age demographics, marital status etc.

- To obtain information on venues present in the different locations of Toronto, I used the Foursquare API which provides general information on the various venues such as their name, geographical location and category. The data provided through the Foursquare API included the following:
  - 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.

3. Methodology

3.1 Data Collection and Preparation

Scraping Neighborhood Data from Wikipedia

The initial step of this process involved scraping the Wikipedia page 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M' in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe.

Assumptions

1. The dataframe consisted of only three columns: PostalCode, Borough, and Neighborhood.
2. The only cells that have been processed are cells that have an assigned borough. Cells with a borough that is ‘Not assigned’ have been ignored.
3. More than one neighborhood can exist in one postal code area.

4. If a cell has a borough but a ‘Not assigned’ neighborhood, then the neighborhood will be the same as the borough.

Below is a sample of the initial output of the data after web-scraping:

```
In [3]: url = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text

In [4]: toronto_data = pd.read_html(url, header = 0)[0]
        toronto_data.head()
```

Out[4]:

	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

After applying data cleansing techniques based on the assumptions above, the final output of the data is seen below, where all boroughs are assigned to neighbourhood:

```
In [5]: #As stated in the assumptions the only cells that will be processed are cells that have an assigned
        borough. Cells with a borough that is Not assigned will be ignored.
        toronto_data = toronto_data[toronto_data.Borough != 'Not assigned']

        #To allow us to merge the location and demographics data to the dataframe, the reference column wil
        l need to have a shared name 'PostalCode'
        toronto_data = toronto_data.rename(columns={'Postcode': 'PostalCode'})
        toronto_data = toronto_data.rename(columns={'Neighbourhood': 'Neighborhood'})

        #A cell that has a borough but is Not assigned to a neighborhood, then the neighborhood will be giv
        en the same name as the borough
        for index, row in toronto_data.iterrows():
            if row['Neighborhood'] == 'Not assigned':
                row['Neighborhood'] = row['Borough']

        toronto_data.head()
```

Out[5]:

	PostalCode	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights
6	M6A	North York	Lawrence Manor

Linking Coordinates to the Data

The next step of the process involved adding geographical coordinates to each location. I extracted data from the csv file (https://cocl.us/Geospatial_data) and merged it with the existing neighborhood dataframe using the PostalCode as the reference for the two sets of data. The resultant table is seen below:

In [7]:
```
coordinates = "https://cocl.us/Geospatial_data"

neighborhood_latlon = requests.get(coordinates).text
neighborhood_latlon_data = pd.read_csv(io.StringIO(neighborhood_latlon))

neighborhood_latlon_data.head()
```

Out[7]:

	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

In [8]:
```
#The reference column 'Postal Code' should be renamed to match the same format as column 'PostalCod
e' in the first dataframe to allow for merging of the two dataframes
neighborhood_latlon_data = neighborhood_latlon_data.rename(columns={'Postal Code': 'PostalCode'})

#Merge the two dataframes
toronto_dataframe = pd.merge(toronto_data, neighborhood_latlon_data, on='PostalCode')
toronto_dataframe.head()
```

Out[8]:

	PostalCode	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	43.718518	-79.464763
4	M6A	North York	Lawrence Manor	43.718518	-79.464763

Getting Demographics Data

The next step of the process involved obtaining data on the ethnic demographics of Toronto given that an important assumption in our analysis was that the success of an Italian restaurant would be influenced by the presence of a large customer base of Italian ethnicity. The City of Toronto's Open Data portal - https://open.toronto.ca/dataset/neighbourhood-profiles/ was able to provide downloadable csv and excel files containing data from the 2016 census. Each file contained a vast amount of census data ranging from household income, employment status, age, gender, ethnicity, immigration etc organized into neighborhoods. A good Data Scientist is able to make use of multiple tools to sort, cleanse and analyse data. For the csv file that was downloaded for this exercise, the data was sorted and cleansed to only show the ethnicity population per neighbourhood, after which the csv file was read into a dataframe and wrangled to only show the top 10 ethnic groups per neighbourhood. The final output is shown below.

Out[40]:

	Neighborhood	Population	Ethnic Group #1	Group 1 Count	Ethnic Group #2	Group 2 Count	Ethnic Group #3	Group 3 Count	Ethnic Group #4	Group 4 Count	Ethnic Group #5	Group 5 Count	Ethnic Group #6
0	Agincourt North	32350	Chinese	16950	Sri Lankan	2230	East Indian	2090	Filipino	1465	Canadian	1295	Englist
1	Agincourt	27185	Chinese	11455	East Indian	2180	Filipino	1405	Sri Lankan	1145	Canadian	1125	Englist
2	Alderwood	18985	English	2320	Canadian	2245	Irish	1900	Italian	1275	Polish	1225	Germa
3	The Annex	50305	English	6745	Irish	5235	Canadian	4655	German	3030	French	2665	Polish
4	Don Mills North	38050	Chinese	4850	English	3615	Irish	3075	Canadian	3035	East Indian	2365	Germa

In [41]:

```
neighborhood_profile.columns
```

Out[41]:

```
Index(['Neighborhood', 'Population', 'Ethnic Group #1', 'Group 1 Count',
       'Ethnic Group #2', 'Group 2 Count', 'Ethnic Group #3', 'Group 3 Count',
       'Ethnic Group #4', 'Group 4 Count', 'Ethnic Group #5', 'Group 5 Count',
       'Ethnic Group #6', 'Group 6 Count', 'Ethnic Group #7', 'Group 7 Count',
       'Ethnic Group #8', 'Group 8 Count', 'Ethnic Group #9', 'Group 9 Count',
       'Ethnic Group #10', 'Group 10 Count'],
      dtype='object')
```

In [42]:

```
neighborhood_profile.shape
```

Out[42]: (140, 22)

Foursquare Location Data

The Foursquare API is a location technology platform that allows developer to access location data. Foursquare was very useful throughout this project as it allowed me to retrieve information on the various venues located within Toronto. This was especially important given that I required information on the number and location Italian restaurants within the Toronto area. I chose to look at 100 popular venues in each neighborhood within a radius of 1km.

In [25]:

```
toronto_venues.head()
```

Out[25]:

	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	GTA Restoration	43.753396	-79.333477	Fireworks 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

To analyse the distribution of each venue category across each neighborhood, one hot encoding was used which allowed me to calculate the mean of all venues grouped by their neighborhoods.

```
In [28]: #one hot encoding
trn_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

# adding neighborhood column back to dataframe
trn_onehot['Neighborhood'] = toronto_venues['Neighborhood']

# moving neighborhood column to the first column
fixed_columns = [trn_onehot.columns[-1]] + list(trn_onehot.columns[:-1])
trn_onehot = trn_onehot[fixed_columns]
trn_grouped = trn_onehot.groupby('Neighborhood').mean().reset_index()

trn_grouped.head()
```

Out[28]:

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop
0	Adelaide	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0
2	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0
3	Albion Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0
4	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0

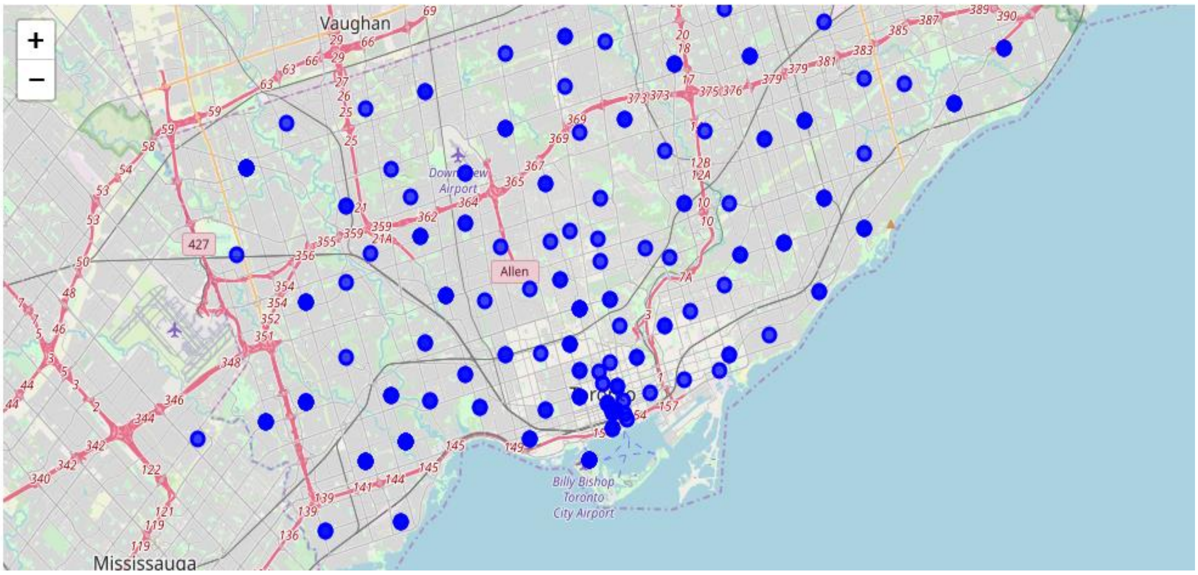
3.2 Exploratory Analysis

Interactive Map

The map below shows the area we will be analysing. The interactive map was generated using the Folium package for Python by making use of the coordinates data collected earlier.

```
In [69]: # Creating a map of Toronto
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=11)

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

map_toronto
```

Out[69]:



Exploring the relationship between Neighborhood and Italian Restaurants

After performing one hot encoding, the next step was to isolate only the Italian restaurants per neighborhood as the resultant table would form the basis of our analysis when determining similar businesses operating within the same cluster.

In [30]:
```
#Extracting Italian Restaurant by Neighborhood
restaurant_by_neighborhood = trn_grouped[['Neighborhood', 'Italian Restaurant']]
restaurant_by_neighborhood
```

Out[30]:

	Neighborhood	Italian Restaurant
0	Adelaide	0.000000
1	Agincourt	0.000000
2	Agincourt North	0.000000
3	Albion Gardens	0.000000
4	Alderwood	0.000000
5	Bathurst Manor	0.000000
6	Bathurst Quay	0.000000
7	Bayview Village	0.000000
8	Beaumond Heights	0.000000
9	Bedford Park	0.076923
10	Berczy Park	0.017857

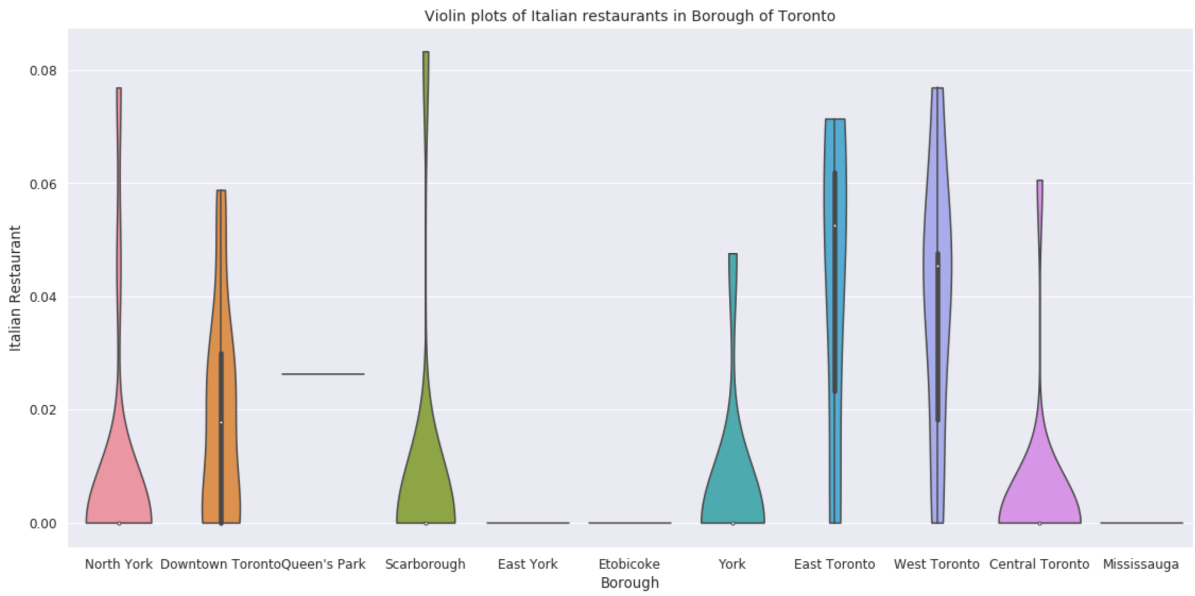
Finally I merged the table showing the Italian Restaurants in each neighborhood to the main dataframe showing the Borough and coordinates of each neighborhood. The resultant table can be seen below.

```
In [31]: #Merging italian restaurants to original dataframe
toronto_merged = pd.merge(toronto_dataframe, restaurant_by_neighborhood, on='Neighborhood')
toronto_merged
```

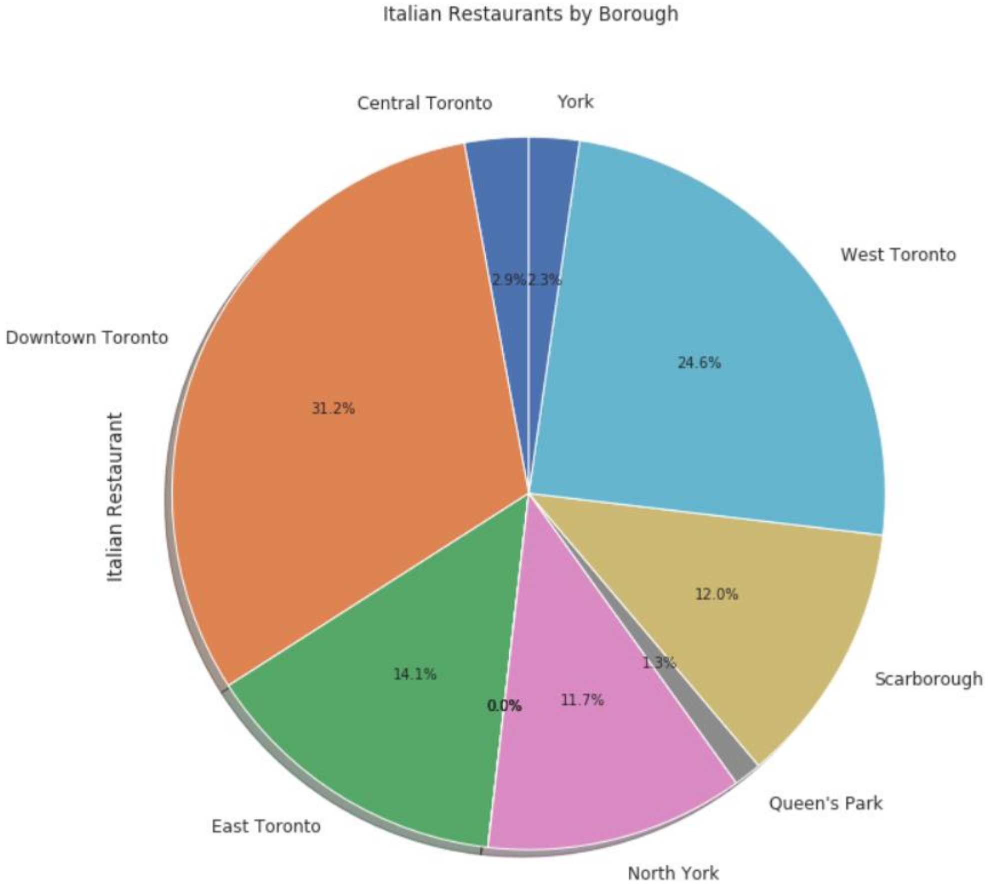
Out[31]:

	PostalCode	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	43.718518	-79.464763	0.000000
4	M6A	North York	Lawrence Manor	43.718518	-79.464763	0.000000
5	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	0.026316
6	M9A	Queen's Park	Queen's Park	43.667856	-79.532242	0.026316
7	M1B	Scarborough	Rouge	43.806686	-79.194353	0.000000
8	M1B	Scarborough	Malvern	43.806686	-79.194353	0.000000
9	M3B	North York	Don Mills North	43.745906	-79.352188	0.000000
10	M4B	East York	Woodbine Gardens	43.706397	-79.309937	0.000000
11	M4B	East York	Parkview Hill	43.706397	-79.309937	0.000000
12	M5B	Downtown Toronto	Ryerson	43.657162	-79.378937	0.020000
13	M5B	Downtown Toronto	Garden District	43.657162	-79.378937	0.020000
14	M6B	North York	Glencairn	43.709577	-79.445073	0.000000
15	M9B	Etobicoke	Cloverdale	43.650943	-79.554724	0.000000
16	M9B	Etobicoke	Islington	43.650943	-79.554724	0.000000

Given that the Italian restaurants had now been mapped to each Postal Code and Borough, it was now possible to visualize the data. Visualization tools allow for a better understanding of data. For the visualisation exercise, I used a categorical Violin plot to identifying the boroughs with densely populated Italian restaurants.



To a better understanding of the distribution percentage of Italian neighborhoods in each Borough, I made use of a pie chart, which allowed me to have an illustration of the general number of Italian restaurants in each Borough in relation to all other Boroughs.



Exploring the relationship between Neighborhood and Italian Population

After determining the relationship between neighbourhoods and Italian restaurants, I proceeded to do the same analysis, except looking at the population of the Italian community in each neighborhood. This analysis is particularly important as it provides the basis for comparison as to whether there exists a relationship between the number of Italian restaurants and size of the Italian population in that neighborhood. This process involved filtering only the Italian population from the overall neighborhood profile dataframe that showed all ethnicities in each neighborhood. The resultant table was then merged with Boroughs to show a final table that shows the Italian population in each neighbourhood and borough.

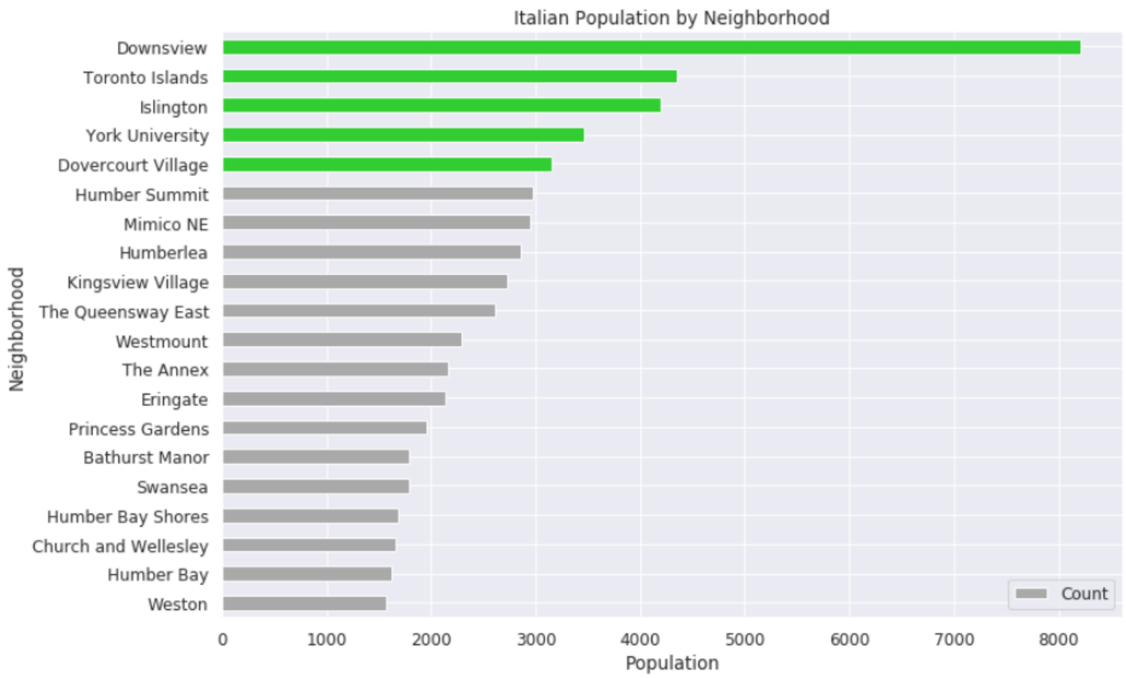
```
In [50]: italian_population = italian_df_count[['Neighborhood','Count']]
         italian_population = pd.merge(italian_population,toronto_merged,on='Neighborhood')
         italian_population = italian_population[['Borough','Neighborhood','Count']]
         italian_population = italian_population.sort_values(by='Count', ascending=False)
         italian_population.head(20)
```

Out[50]:

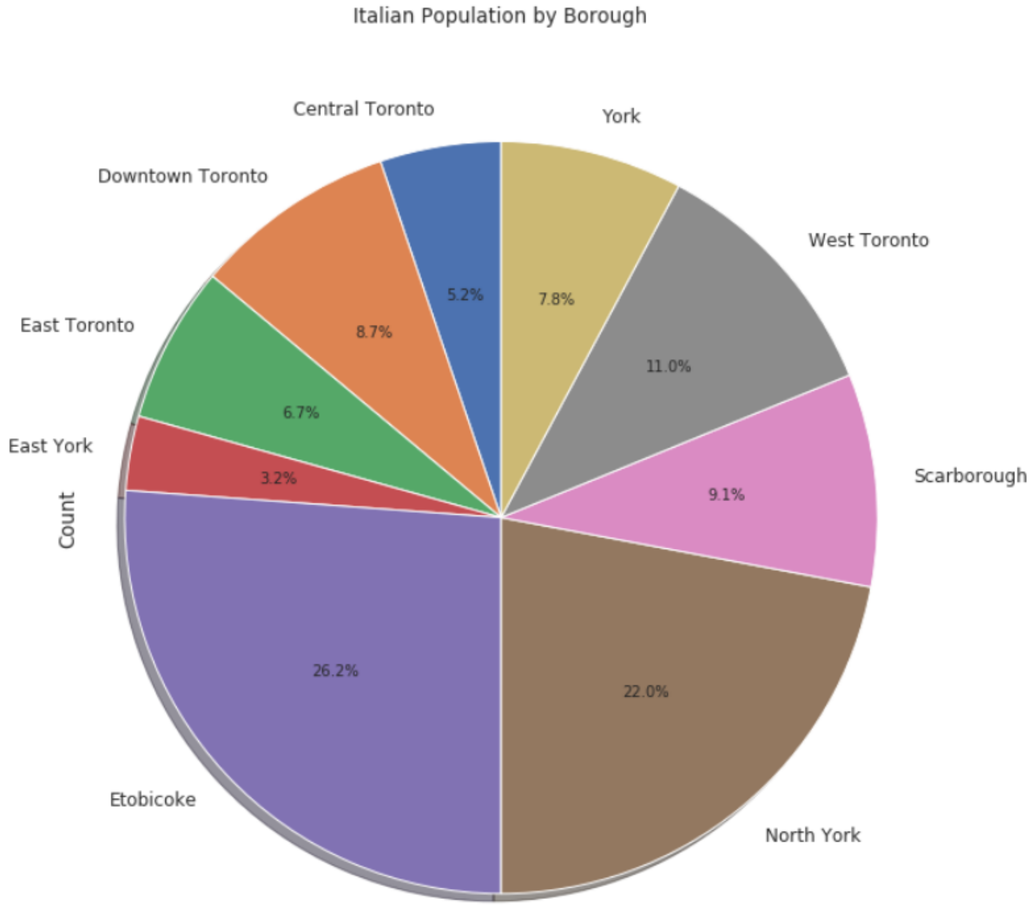
	Borough	Neighborhood	Count
21	North York	Downsview	8205.0
69	Downtown Toronto	Toronto Islands	4350.0
36	Etobicoke	Islington	4195.0
78	North York	York University	3465.0
20	West Toronto	Dovercourt Village	3155.0
33	North York	Humber Summit	2970.0
47	Etobicoke	Mimico NE	2945.0

A horizontal bar graph visualization of the table above was able to show us the neighborhoods with the highest Italian populations.

```
In [52]: # Differentiating the colours of the bars
         colors = ['darkgrey','darkgrey','darkgrey','darkgrey','darkgrey','darkgrey','darkgrey','darkgrey
         ','darkgrey',
                   'darkgrey','darkgrey','darkgrey','darkgrey','darkgrey','darkgrey','limegreen','limegreen
         ','limegreen','limegreen','limegreen']

         # Plotting the chart
         bar_graph = italian_population.head(20).sort_values(by='Count', ascending=True)
         bar_graph.plot(kind='barh',x='Neighborhood', y='Count',figsize=(12,8), color=colors)

         plt.title("Italian Population by Neighborhood")
         plt.xlabel("Population")
         plt.ylabel("Neighborhood")

         plt.show()
```



To a better understanding of the distribution percentage of Italian population in each Borough, I made use of a pie chart, which allowed me to have an illustration of the general number of Italians in each Borough in relation to all other Boroughs. The logic behind the analysis is that placing an Italian restaurant in a densely populated Italian neighborhood would more likely to get more Italian customers than a restaurant placed in a neighborhood with less or no Italian population.



Exploring the relationship between Italian Restaurants and Italian Population

After exploring the different relationships between Italian population, Italian restaurants and the various neighbourhoods, the next logical step was to determine whether there was a relationship between the Italian restaurants and the Italian population. The initial step of this exercise was merging the Italian population dataframe with the Italian restaurant dataframe.

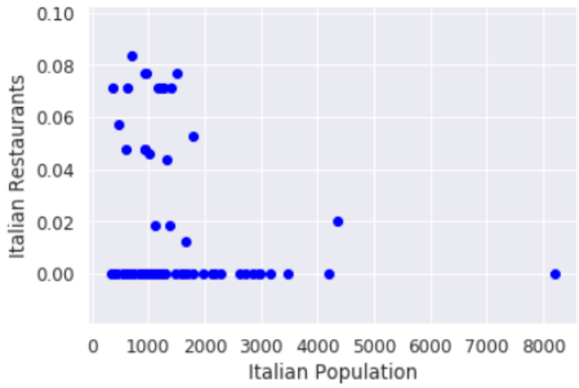
In [54]:
```
merged_italian_restaurant = pd.merge(italian_population, restaurant_by_neighborhood, on='Neighborh
ood')
merged_italian_restaurant = merged_italian_restaurant[['Neighborhood','Count','Italian Restaurant
']]
merged_italian_restaurant.columns = ['Neighborhood','Italian Population','Italian Restaurants']
merged_italian_restaurant
```

Out[54]:

	Neighborhood	Italian Population	Italian Restaurants
0	Downsview	8205.0	0.000000
1	Toronto Islands	4350.0	0.020000
2	Islington	4195.0	0.000000
3	York University	3465.0	0.000000
4	Dovercourt Village	3155.0	0.000000
5	Humber Summit	2970.0	0.000000

A scatter plot was then used to visualise whether there was a strong linear relationship between the Italian population and the number of Italian neighborhoods.

In [55]:
```
# Plotting a scatter plot
plt.scatter(merged_italian_restaurant['Italian Population'], merged_italian_restaurant['Italian Re
staurants'],  color='blue')
plt.xlabel("Italian Population")
plt.ylabel("Italian Restaurants")
plt.show()
```



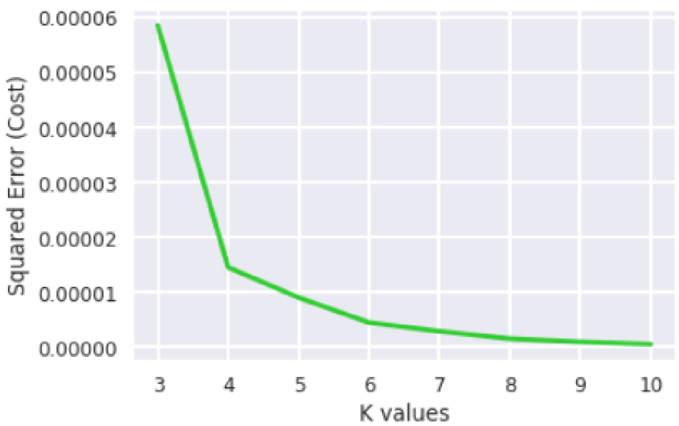
After performing the data cleansing and data analysis we can identify from the scatter plot that there is no strong linear relationship between the Italian population and the number of Italian restaurants, therefore population is not a good indicator of the presence of Italian restaurants. However, this might be because of missing data. This is an area which can be further improved in future analysis to get more meaningful insight.

3.3 Modelling

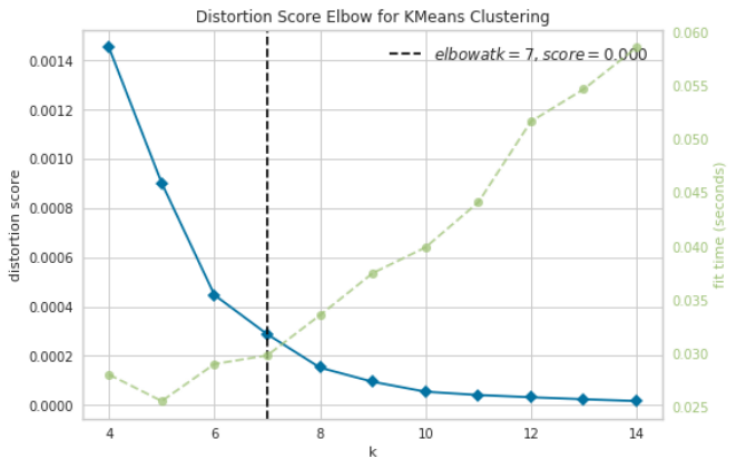
Clustering the Neighborhoods of Toronto

After drawing insights from our exploratory analysis, the next step was to come up with a predictive model to determine the various clusters in which to set up the new Italian restaurant. For this analysis I chose K-means clustering. K-means is vastly used for clustering in many data science applications. It is especially useful for quickly discovering insights from unlabelled data. The initial step in K-means clustering involves identifying the best K value i.e. the number of clusters in a given dataset. To do so I used the elbow method on the Toronto dataset with the Restaurant by neighborhood (i.e. toronto_merged dataframe).

In [56]:
```
from sklearn.cluster import KMeans

toronto_part_clustering = restaurant_by_neighborhood.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='limegreen', linewidth='3')
plt.xlabel('K values')
plt.ylabel('Squared Error (Cost)')
plt.grid(color='white', linestyle='-', linewidth=2)
plt.show()
```



After analysing using elbow method using distortion score & Squared error for each K value, K = 7 was determined to be the K value to use for the clustering exercise as shown below.




```
Out[58]: <matplotlib.axes._subplots.AxesSubplot at 0x7fd3ba3fa128>
```

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

```
In [59]: kclusters = 7

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

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

         kmeans.labels_
```

```
Out[59]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 0, 0, 5, 0, 0, 0, 5, 0, 0, 0, 5,
                0, 1, 6, 0, 3, 0, 0, 0, 0, 2, 1, 0, 0, 0, 2, 0, 5, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 6, 5, 0, 0, 0, 4, 0, 0, 0, 0, 1, 0,
                4, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 5, 0, 0, 0, 0,
                2, 0, 0, 0, 0, 3, 3, 0, 0, 0, 0, 0, 0, 5, 4, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 2, 0, 0, 2, 5, 3, 0, 0, 0, 1, 3, 0, 0, 1, 3, 0, 5, 0, 0,
                0, 0, 0, 2, 4, 4, 6, 4, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0], dtype=int32)
```

```
In [60]: #sorted_neighborhoods_venues.drop(['Cluster Labels'],axis=1,inplace=True)
         restaurant_by_neighborhood.insert(0, 'Cluster Labels', kmeans.labels_)
         toronto_merged = toronto_dataframe
         # merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
         toronto_merged = toronto_merged.join(restaurant_by_neighborhood.set_index('Neighborhood'), on='Nei
         ghborhood')
         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[60]:

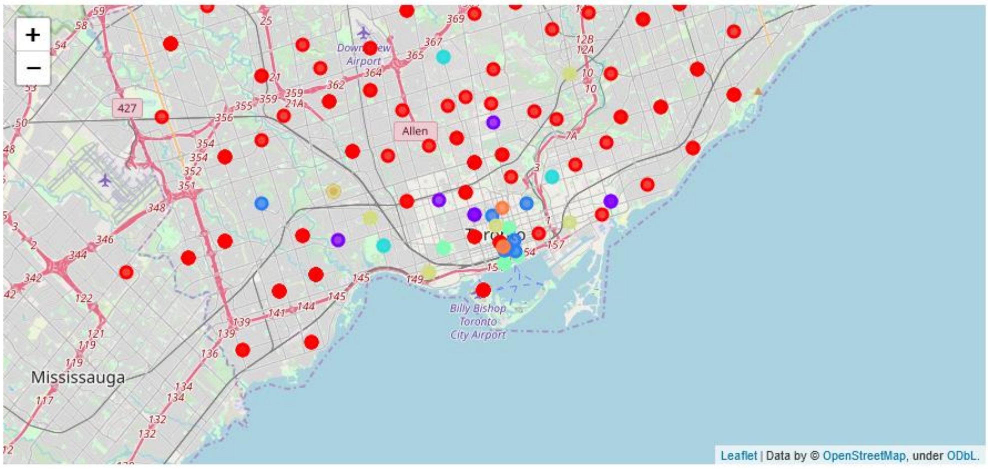
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.0	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	0.0	0.0
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	0.0	0.0
3	M6A	North York	Lawrence Heights	43.718518	-79.464763	0.0	0.0
4	M6A	North York	Lawrence Manor	43.718518	-79.464763	0.0	0.0

Examining the Clusters

After carrying out the K-means clustering, the output was a total of 7 clusters ranging cluster 0 to cluster 6. The map below shows the different clusters spread out over our interactive map that was generated using folium.

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11, width='90%', height='70%')

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

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(toronto_merged['Latitude'], toronto_merged['Longitude'], toronto_merged['Neighborhood'], to
ronto_merged['Cluster Labels'].astype(int)):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)
map_clusters
```

Out[61]:



Cluster 0 contained all the neighborhoods which had the least number of Italian restaurants. It is shown in red color on the map.

In [62]:
```
#Cluster 0
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0]
```

Out[62]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.0	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	0.0	0.0
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	0.0	0.0
3	M6A	North York	Lawrence Heights	43.718518	-79.464763	0.0	0.0
4	M6A	North York	Lawrence Manor	43.718518	-79.464763	0.0	0.0
7	M1B	Scarborough	Rouge	43.806686	-79.194353	0.0	0.0
8	M1B	Scarborough	Malvern	43.806686	-79.194353	0.0	0.0
9	M3B	North York	Don Mills North	43.745906	-79.352188	0.0	0.0

Cluster 3 contained all the neighborhoods which were the most densely populated with Italian restaurants. It is shown in the color teal on the map.

In [65]:
```
#Cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3]
```

Out[65]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
71	M4K	East Toronto	The Danforth West	43.679557	-79.352188	3.0	0.071429
72	M4K	East Toronto	Riverdale	43.679557	-79.352188	3.0	0.071429
97	M5M	North York	Bedford Park	43.733283	-79.419750	3.0	0.076923
98	M5M	North York	Lawrence Manor East	43.733283	-79.419750	3.0	0.076923
131	M6R	West Toronto	Parkdale	43.648960	-79.456325	3.0	0.076923
132	M6R	West Toronto	Roncesvalles	43.648960	-79.456325	3.0	0.076923
144	M1T	Scarborough	Clarks Corners	43.781638	-79.304302	3.0	0.083333
145	M1T	Scarborough	Sullivan	43.781638	-79.304302	3.0	0.083333
146	M1T	Scarborough	Tam O'Shanter	43.781638	-79.304302	3.0	0.083333

All other clusters are shown below.

In [63]:
```
#Cluster 1
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1]
```

Out[63]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
41	M6G	Downtown Toronto	Christie	43.669542	-79.422564	1.0	0.058824
84	M4L	East Toronto	The Beaches West	43.668999	-79.315572	1.0	0.052632
85	M4L	East Toronto	India Bazaar	43.668999	-79.315572	1.0	0.052632
139	M4S	Central Toronto	Davisville	43.704324	-79.388790	1.0	0.060606
140	M5S	Downtown Toronto	Harbord	43.662696	-79.400049	1.0	0.057143
141	M5S	Downtown Toronto	University of Toronto	43.662696	-79.400049	1.0	0.057143
143	M6S	West Toronto	Swansea	43.651571	-79.484450	1.0	0.052632

In [64]:
```
#Cluster 2
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2]
```

Out[64]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
5	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	2.0	0.026316
6	M9A	Queen's Park	Queen's Park	43.667856	-79.532242	2.0	0.026316
26	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2.0	0.028169
73	M5K	Downtown Toronto	Design Exchange	43.647177	-79.381576	2.0	0.030000
74	M5K	Downtown Toronto	Toronto Dominion Centre	43.647177	-79.381576	2.0	0.030000
86	M5L	Downtown Toronto	Commerce Court	43.648198	-79.379817	2.0	0.030000
87	M5L	Downtown Toronto	Victoria Hotel	43.648198	-79.379817	2.0	0.030000
181	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	2.0	0.031579
186	M4X	Downtown Toronto	St. James Town	43.667967	-79.367675	2.0	0.028169

```
In [66]: #Cluster 4
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4]
```

Out[66]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
12	M5B	Downtown Toronto	Ryerson	43.657162	-79.378937	4.0	0.020000
13	M5B	Downtown Toronto	Garden District	43.657162	-79.378937	4.0	0.020000
36	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	4.0	0.017857
60	M5J	Downtown Toronto	Harbourfront East	43.640816	-79.381752	4.0	0.020000
61	M5J	Downtown Toronto	Toronto Islands	43.640816	-79.381752	4.0	0.020000
62	M5J	Downtown Toronto	Union Station	43.640816	-79.381752	4.0	0.020000
63	M6J	West Toronto	Little Portugal	43.647927	-79.419750	4.0	0.018182
64	M6J	West Toronto	Trinity	43.647927	-79.419750	4.0	0.018182

```
In [67]: #Cluster 5
toronto_merged.loc[toronto_merged['Cluster Labels'] == 5]
```

Out[67]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
23	M3C	North York	Flemingdon Park	43.725900	-79.340923	5.0	0.045455
24	M3C	North York	Don Mills South	43.725900	-79.340923	5.0	0.045455
40	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	5.0	0.045977
75	M6K	West Toronto	Brockton	43.636847	-79.428191	5.0	0.045455
76	M6K	West Toronto	Exhibition Place	43.636847	-79.428191	5.0	0.045455
77	M6K	West Toronto	Parkdale Village	43.636847	-79.428191	5.0	0.045455
96	M4M	East Toronto	Studio District	43.659526	-79.340923	5.0	0.046512
112	M6N	York	Runnymede	43.673185	-79.487262	5.0	0.047619
121	M6P	West Toronto	High Park	43.661608	-79.464763	5.0	0.043478
122	M6P	West Toronto	The Junction South	43.661608	-79.464763	5.0	0.043478
142	M6S	West Toronto	Runnymede	43.651571	-79.484450	5.0	0.047619
185	M4X	Downtown Toronto	Cabbagetown	43.667967	-79.367675	5.0	0.047619

```
In [68]: #Cluster 6
toronto_merged.loc[toronto_merged['Cluster Labels'] == 6]
```

Out[68]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Italian Restaurant
187	M5X	Downtown Toronto	First Canadian Place	43.648429	-79.38228	6.0	0.010000
188	M5X	Downtown Toronto	Underground city	43.648429	-79.38228	6.0	0.010000
192	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.38316	6.0	0.012195

4. Results and Discussion:

4.1 Results

The results section documents all the findings from above clustering and visualization exercise on the data provided. The business problem set out to identify a good neighborhood in which to open a new Italian restaurant with the initial assumptions that Italian restaurants would be concentrated in areas with high Italian populations. We looked into all the neighborhoods in Toronto and analysed the Italian population in each neighborhood including the spread of Italian restaurants in those neighborhoods. The aim of the analysis was to come to a conclusion on which neighborhood would be a better location for opening a new Italian restaurant.

For the analysis, data was scraped from web resources like Wikipedia, geospatial coordinates of Toronto neighborhoods, and Foursquare API, to set up a very realistic data-analysis scenario. Observations from the analysis suggested that of the 210 neighborhoods in Toronto, only 114 neighborhoods were identified as having Italian communities presents in them with the highest concentration of Italians being located in Downsview. We further analysed the concentration of Italian restaurants and visualized the data using a Violin plot. The analysis suggested that of the 11 boroughs only North York, Central Toronto, Downtown Toronto, East Toronto, West Toronto, York & Scarborough boroughs had the highest concentration of Italian restaurants in Toronto. Using a scatter plot, we were able to visualize whether there was an distinct relationship between the population density of the Italian community and the number of Italian restaurants present in a neighborhood. The results of the analysis showed that there was no distinct relationship between the population density of Italians and Italian restaurants. The analysis showed that clustering the dispersion of restaurants into 7 clusters would provide a good spread of centroids. From the K-mean clustering analysis, the cluster with the most restaurants is Cluster 3, followed by Cluster 1, then Cluster 5 followed by Cluster 2, then Cluster 4, Cluster 6 and finally Cluster 0

4.2 Discussion

The most ideal location for an Italian restaurant would be one where there is an established market for Italian restaurants without too much competition. Cluster 3, 1 and 5 respectively have the highest concentration of Italian restaurants, which would be problematic for the new business as it would be going up against more established businesses. Cluster 0 and Cluster 6 has the least number of Italian restaurants, which could be a indicator that the market for Italian cuisine/restaurants is very weak in those areas. The options in between the clusters that are most densely populated with Italian restaurants and the areas with the least number of restaurants are clusters 2 and cluster 4. Therefore from a competition perspective, cluster 2 and cluster 4 would represent a good balance.

Another factor to consider when choosing an area to set up a restaurant would be market size. We have taken the Italian population of each neighborhood to be an indicator of our potential market size per neighborhood. Although our scatter plot did not suggest that there is a relationship between Italian ethnic population and the number of restaurants, as evidenced by the results of Downsview having the highest concentration of Italians but having no Italian restaurants in the area, on the surface it would still be a fair assumption that the Italian population would be more familiar with Italian cuisine than all other ethnicities. However, the results of the scatter plot warrant further research and analysis into the factors that influence the number of Italian restaurants in the area. Based on the results presented, the highest number of Italian restaurants is in East Toronto followed by West Toronto then Downtown Toronto. However, East Toronto has a smaller population than Downtown Toronto and West Toronto; furthermore West Toronto has a very high number of Italian restaurants, leaving clusters within Downtown Toronto as the more ideal locations for a new Italian restaurant as it have a fairly sizable population of Italians and does not have a very high number of competitors.

Some of the drawbacks of this analysis are that the clustering is completely based only on data obtained from Foursquare API. Furthermore, the Italian population distribution in each neighborhood is also based on the 2016 census which is not up-to date. Thus population distribution may have changed by the time of this analysis.

5. Conclusion

Throughout this project we made use of many python libraries to fetch data, manipulate its contents and analyse the datasets. We have made use of Foursquare API to explore the venues in neighborhoods of Toronto, and got a good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized the data using various plots present in seaborn & matplotlib. We also applied machine learning technique to to predict the output given the data and used Folium to visualize it on a map. Some of the drawbacks were that this analysis can be improved further with more data and different machine learning techniques to examine relationship. Similarly we can use this project to analyse any scenario for opening any type of business, for example a spa. Hopefully, this project can help act as an initial guide to make complex decisions using data-science.