IBM Data Science Professional Program Capstone Project

FEBRUARY 2020

Authored by: Vinay Burhade

Battle of the Neighborhoods: Toronto

Introduction

This report serves as a step by step explanation of the project as a part of the IBM Data Science Capstone course, from problem description to the conclusion, including the analysis involved in the project. We worked on real world datasets and experienced what it takes to be a Data Scientist. The main objectives consisted of defining the business problem, acquiring the data required for analysis through web scraping techniques, cleaning the data, acquiring the neighborhood venue data using the FourSquare API and using the analysis to choose the neighborhoods suitable to open a new Asian/Chinese restaurant.

1. **Description of the Business Problem and the Background:**

   Problem Statement: Target potential neighborhoods holding prospect to open a successful Chinese/Asian Restaurant business.

   The demographics of Toronto, Ontario, Canada make Toronto one of the most multicultural and multiracial cities in the world. In 2016, 51.5% of the residents of the city proper belonged to a visible minority group, compared with 49.1% in 2011, and 13.6% in 1981. Toronto also has established ethnic neighborhoods such as the multiple Chinatowns, Corso Italia, Little Italy, Little India, Greektown, Koreatown, Little Jamaica, Little Portugal and Roncesvalles, which celebrate the city's multiculturalism.

   This project will include steps to decide whether it is a good idea to open a Chinese Restaurant in Toronto and if yes, which neighborhood holds the maximum likelihood to make the business thrive. Toronto is home to a lot of Chinese people; in fact, it the Chinese population holds the maximum contribution to the total ethnic population in Toronto. Using analysis of the neighborhoods using the demographics as well as the venue data, we will try to identify the most profitable areas since the success of such business depends on the population and the ethnic backgrounds of the people.

Target Audience

1) Business personnel who wants to start a new Chinese restaurant and wants to identify the areas where the business will thrive the most targeting the Asian/Chines crowd.
2) A Data Analyst or Data Scientist who wants to practice real world business problems and identify solutions based on statistical, exploratory and visual analytics.
3) A restaurant chain owner who wants to expand his business in various areas of Toronto attracting more crowd and increasing the revenue of his business.

2. Data Gathering and Pre-Processing:

2.1 Data Sources

a. "List of Postal Codes: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) Wikipedia page to extract all the neighborhoods as well as borough in the Toronto city. This page includes tables with postal codes, neighborhoods and the boroughs in Toronto.
b. Geocoder ArcGIS package was used to get the data on all the geographical coordinates of the neighborhoods.
c. To acquire the data on the distribution of population among various neighborhoods in Toronto, the Wikipedia page on Ethnic diversity of Toronto (https://en.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) was scraped. The tables from this page gave the information about the total population of the Ridings and how it was distributed by people from various Ethnic origins one of which was Chinese which will be used in the project.

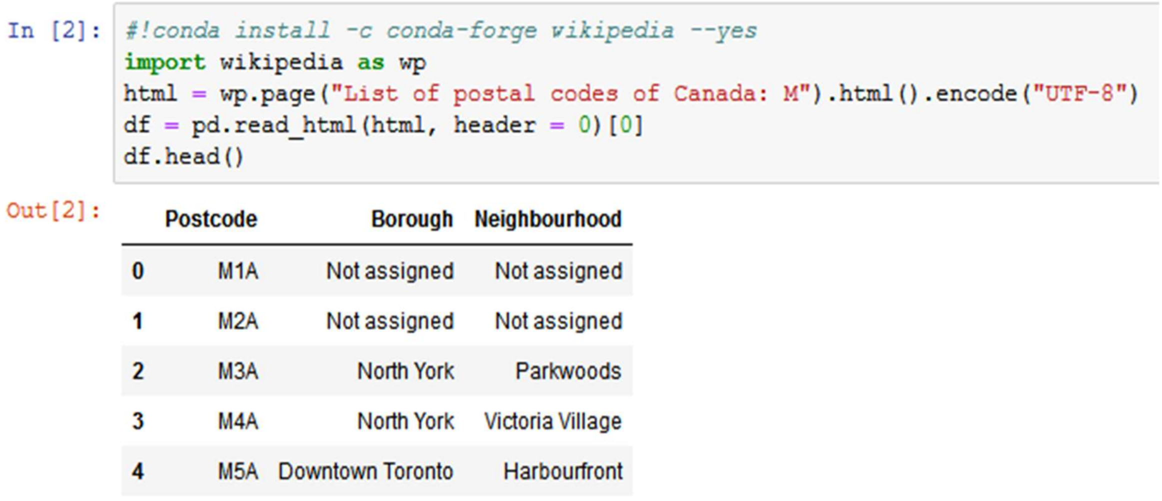
d. To get the information about location and other things of the venues in Toronto, the FourSquare API was used. Using this API, information like name, nearby recommendations, category, latitude, longitude, etc. was fetched.

2.2. Data Cleaning and Pre-Processing:

a. The first and foremost step was to scrape the data from the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), which includes the data about the postal codes and the boroughs with neighborhoods associated with them.

   Once the table was scraped, it had to be processed for redundancy as more than one neighborhood could be present in one postal code area. For example, M5A is listed twice and it has two neighborhoods: Harbourfront and Regent Park. These two will be combined to make one single row separated with comma. If any cell has a borough but neighborhood as 'Not Assigned', then the neighborhood will be same as the borough.

   Scraped data from Wikipedia using Wikipeddia package:

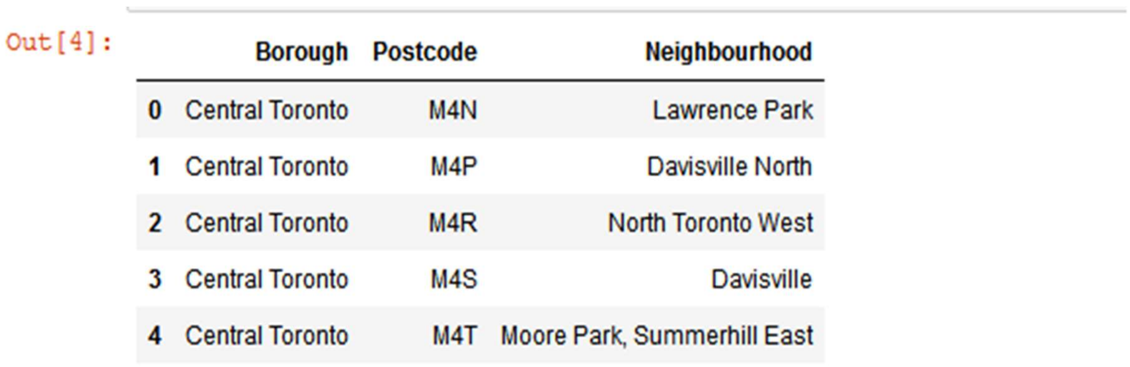


```
In [2]: #!conda install -c conda-forge wikipedia --yes
        import wikipedia as wp
        html = wp.page("List of postal codes of Canada: M").html().encode("UTF-8")
        df = pd.read_html(html, header = 0)[0]
        df.head()
```

Out[2]:

	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 some cleaning tasks, we achieved the proper dataframe with Postal Code, Borough and Neighborhood info.

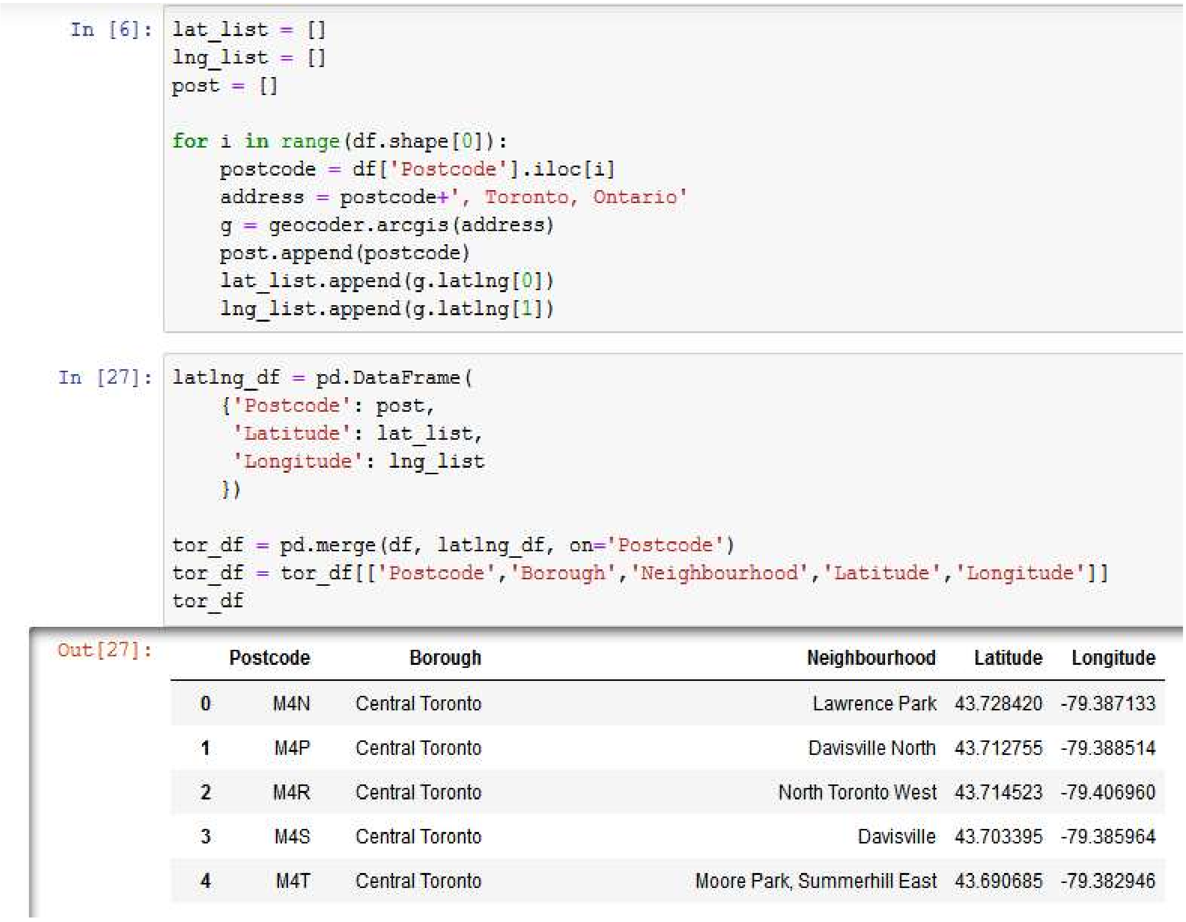


Out[4]:

	Borough	Postcode	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

b. Adding the geographical coordinates to the data:

The next important step is to add the geographical data to the above processed dataframe. The coordinates i.e. latitude and longitude, for each Postal code is acquired using the Geocoder ArcGIS package. This data was merged with the previously created dataframe on the Postal Code column.



```
In [6]: lat_list = []
        lng_list = []
        post = []

        for i in range(df.shape[0]):
            postcode = df['Postcode'].iloc[i]
            address = postcode+', Toronto, Ontario'
            g = geocoder.arcgis(address)
            post.append(postcode)
            lat_list.append(g.latlng[0])
            lng_list.append(g.latlng[1])
```

```
In [27]: latlng_df = pd.DataFrame(
             {'Postcode': post,
              'Latitude': lat_list,
              'Longitude': lng_list
             })

         tor_df = pd.merge(df, latlng_df, on='Postcode')
         tor_df = tor_df[['Postcode','Borough','Neighbourhood','Latitude','Longitude']]
         tor_df
```

Out[27]:

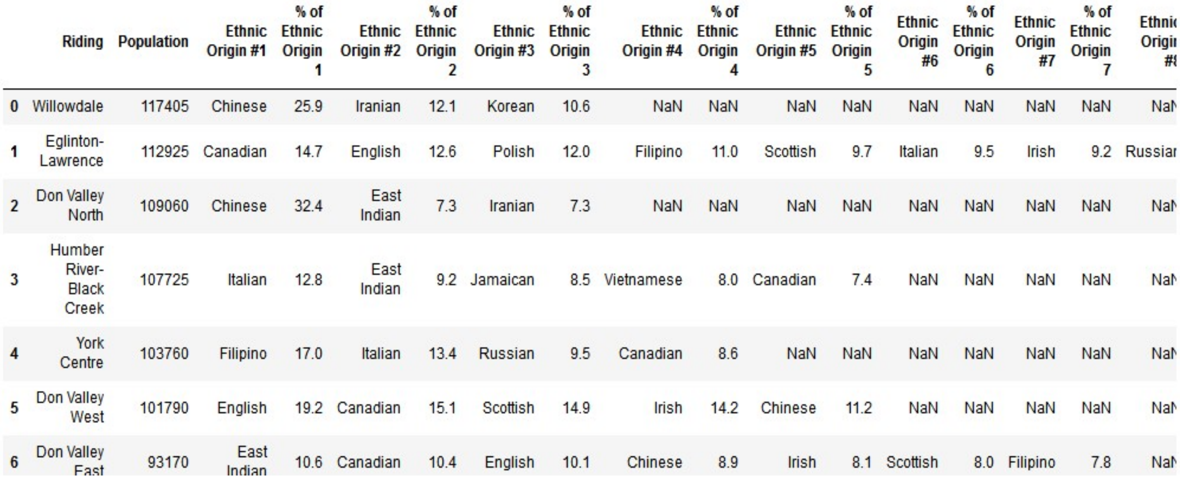
	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728420	-79.387133
1	M4P	Central Toronto	Davisville North	43.712755	-79.388514
2	M4R	Central Toronto	North Toronto West	43.714523	-79.406960
3	M4S	Central Toronto	Davisville	43.703395	-79.385964
4	M4T	Central Toronto	Moore Park, Summerhill East	43.690685	-79.382946

c. Scraping distribution of population from Wikipedia

The data related to how the ethnic population is distributed among various neighborhoods of Toronto will play an important role in deciding where to place our new restaurant. Using this data, we can identify the neighborhoods with highest Chinese ethnic population. This analysis will help us to enlist the areas which have maximum likelihood of the new restaurant's success.

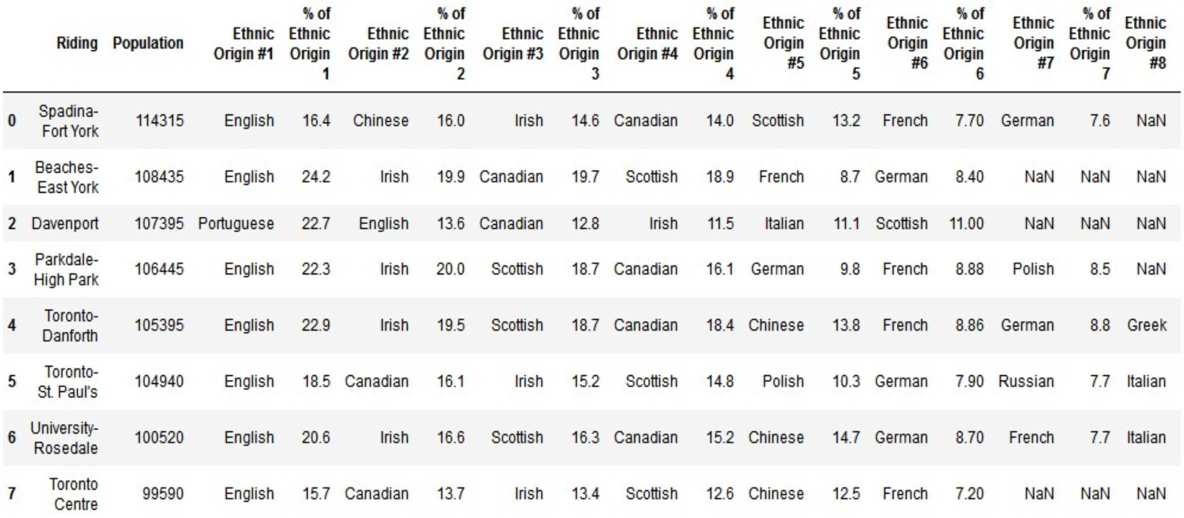
In order to get this data, I had scraped the "Demographics of Toronto" Wikipedia page. This allowed us to collect the distribution of population and the percentage of different ethnicities over the total population. The data for different ridings looks like the following:

North-York:



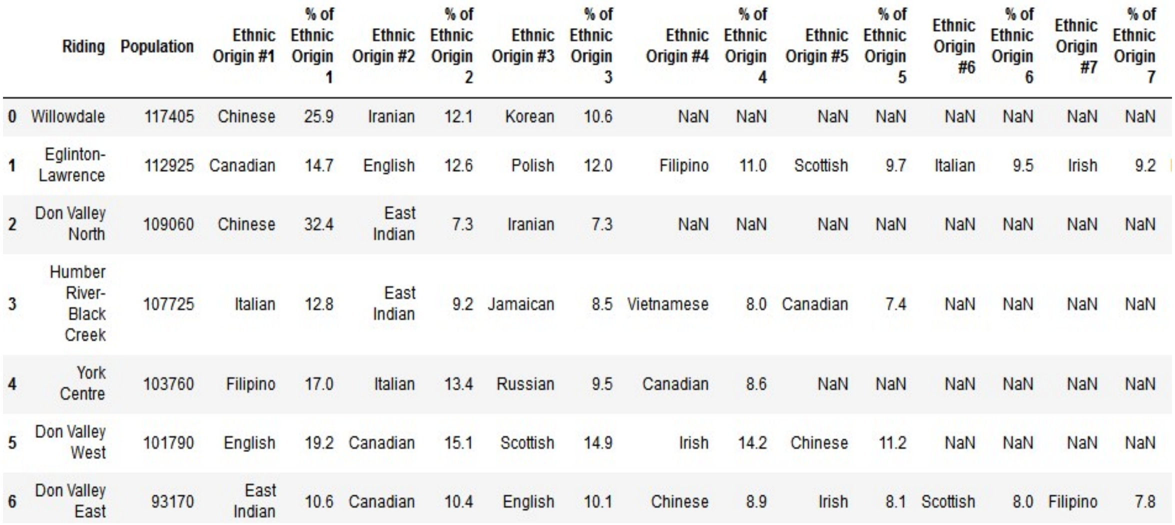
	Riding	Population	Ethnic Origin #1	% of Ethnic Origin 1	Ethnic Origin #2	% of Ethnic Origin 2	Ethnic Origin #3	% of Ethnic Origin 3	Ethnic Origin #4	% of Ethnic Origin 4	Ethnic Origin #5	% of Ethnic Origin 5	Ethnic Origin #6	% of Ethnic Origin 6	Ethnic Origin #7	% of Ethnic Origin 7	Ethnic Origin #8
0	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Eglinton-Lawrence	112925	Canadian	14.7	English	12.6	Polish	12.0	Filipino	11.0	Scottish	9.7	Italian	9.5	Irish	9.2	Russian
2	Don Valley North	109060	Chinese	32.4	East Indian	7.3	Iranian	7.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Humber River-Black Creek	107725	Italian	12.8	East Indian	9.2	Jamaican	8.5	Vietnamese	8.0	Canadian	7.4	NaN	NaN	NaN	NaN	NaN
4	York Centre	103760	Filipino	17.0	Italian	13.4	Russian	9.5	Canadian	8.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	Don Valley West	101790	English	19.2	Canadian	15.1	Scottish	14.9	Irish	14.2	Chinese	11.2	NaN	NaN	NaN	NaN	NaN
6	Don Valley East	93170	East Indian	10.6	Canadian	10.4	English	10.1	Chinese	8.9	Irish	8.1	Scottish	8.0	Filipino	7.8	NaN

Toronto & East York:



	Riding	Population	Ethnic Origin #1	% of Ethnic Origin 1	Ethnic Origin #2	% of Ethnic Origin 2	Ethnic Origin #3	% of Ethnic Origin 3	Ethnic Origin #4	% of Ethnic Origin 4	Ethnic Origin #5	% of Ethnic Origin 5	Ethnic Origin #6	% of Ethnic Origin 6	Ethnic Origin #7	% of Ethnic Origin 7	Ethnic Origin #8
0	Spadina-Fort York	114315	English	16.4	Chinese	16.0	Irish	14.6	Canadian	14.0	Scottish	13.2	French	7.70	German	7.6	NaN
1	Beaches-East York	108435	English	24.2	Irish	19.9	Canadian	19.7	Scottish	18.9	French	8.7	German	8.40	NaN	NaN	NaN
2	Davenport	107395	Portuguese	22.7	English	13.6	Canadian	12.8	Irish	11.5	Italian	11.1	Scottish	11.00	NaN	NaN	NaN
3	Parkdale-High Park	106445	English	22.3	Irish	20.0	Scottish	18.7	Canadian	16.1	German	9.8	French	8.88	Polish	8.5	NaN
4	Toronto-Danforth	105395	English	22.9	Irish	19.5	Scottish	18.7	Canadian	18.4	Chinese	13.8	French	8.86	German	8.8	Greek
5	Toronto-St. Paul's	104940	English	18.5	Canadian	16.1	Irish	15.2	Scottish	14.8	Polish	10.3	German	7.90	Russian	7.7	Italian
6	University-Rosedale	100520	English	20.6	Irish	16.6	Scottish	16.3	Canadian	15.2	Chinese	14.7	German	8.70	French	7.7	Italian
7	Toronto Centre	99590	English	15.7	Canadian	13.7	Irish	13.4	Scottish	12.6	Chinese	12.5	French	7.20	NaN	NaN	NaN

Scarborough:

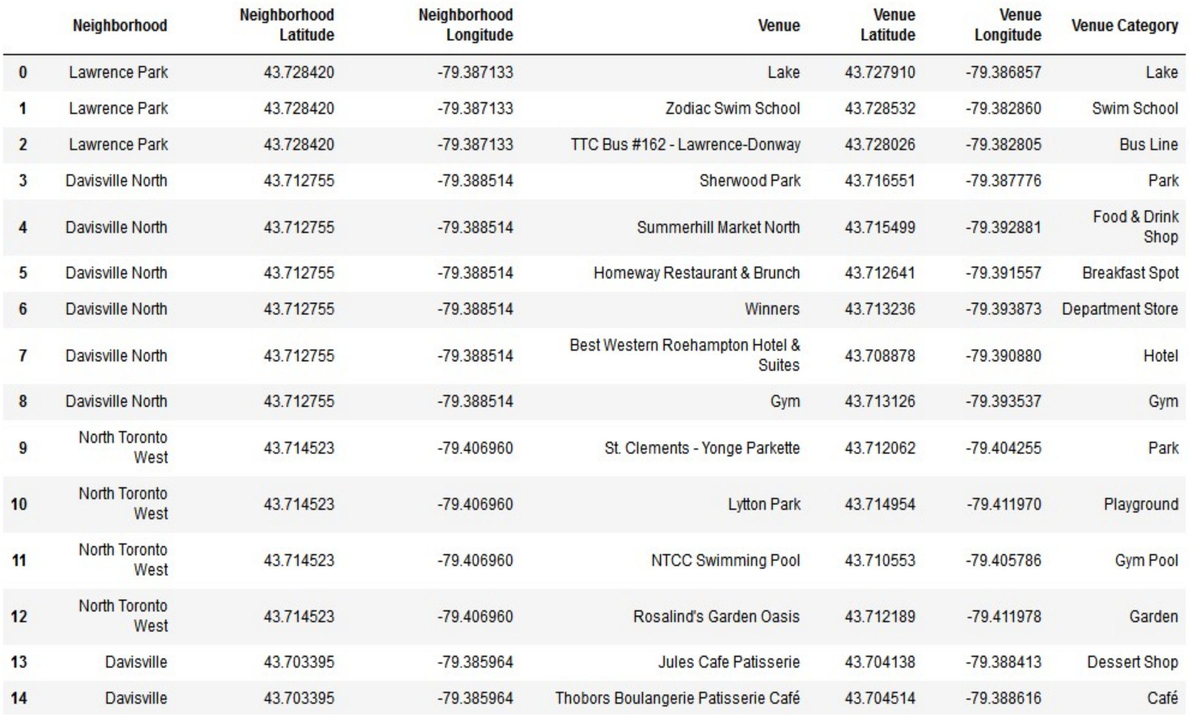


	Riding	Population	Ethnic Origin #1	% of Ethnic Origin 1	Ethnic Origin #2	% of Ethnic Origin 2	Ethnic Origin #3	% of Ethnic Origin 3	Ethnic Origin #4	% of Ethnic Origin 4	Ethnic Origin #5	% of Ethnic Origin 5	Ethnic Origin #6	% of Ethnic Origin 6	Ethnic Origin #7	% of Ethnic Origin 7
0	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Eglinton-Lawrence	112925	Canadian	14.7	English	12.6	Polish	12.0	Filipino	11.0	Scottish	9.7	Italian	9.5	Irish	9.2
2	Don Valley North	109060	Chinese	32.4	East Indian	7.3	Iranian	7.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Humber River-Black Creek	107725	Italian	12.8	East Indian	9.2	Jamaican	8.5	Vietnamese	8.0	Canadian	7.4	NaN	NaN	NaN	NaN
4	York Centre	103760	Filipino	17.0	Italian	13.4	Russian	9.5	Canadian	8.6	NaN	NaN	NaN	NaN	NaN	NaN
5	Don Valley West	101790	English	19.2	Canadian	15.1	Scottish	14.9	Irish	14.2	Chinese	11.2	NaN	NaN	NaN	NaN
6	Don Valley East	93170	East Indian	10.6	Canadian	10.4	English	10.1	Chinese	8.9	Irish	8.1	Scottish	8.0	Filipino	7.8

The above tables show the distribution of population from different ethnic origins and their percentage. In order to focus only on Chinese population, we have scraped the tables for only those Ridings which have comparatively high population densities.

d. Using Foursquare API for location data:

Foursquare API allows developers to acquire the data related to the location. This allows the users to collect the data including the venue name, category, coordinates, menu, recommendations, etc. In this project, we are using the API to collect the venue and category data along with the coordinates. We have limited the results to 100 popular spots in each neighborhood within a radius of 500 meters.

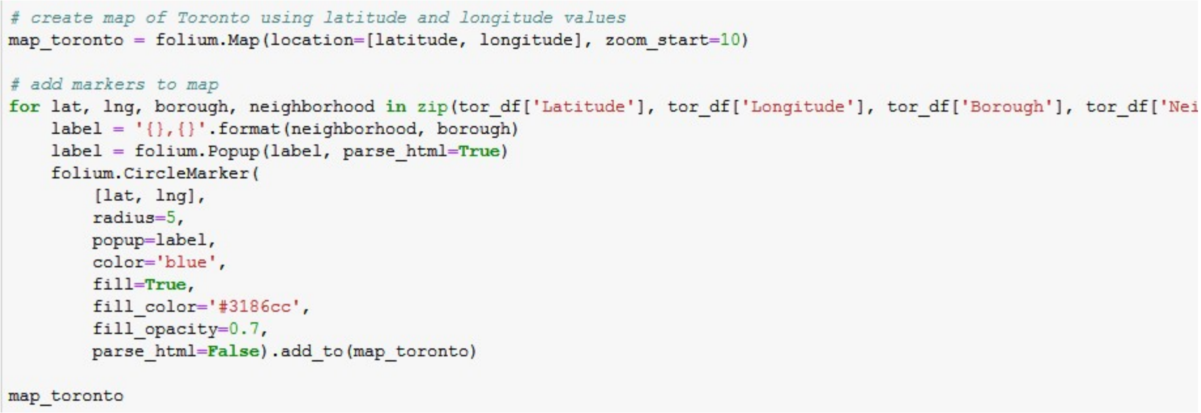


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728420	-79.387133	Lake	43.727910	-79.386857	Lake
1	Lawrence Park	43.728420	-79.387133	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728420	-79.387133	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712755	-79.388514	Sherwood Park	43.716551	-79.387776	Park
4	Davisville North	43.712755	-79.388514	Summerhill Market North	43.715499	-79.392881	Food & Drink Shop
5	Davisville North	43.712755	-79.388514	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
6	Davisville North	43.712755	-79.388514	Winners	43.713236	-79.393873	Department Store
7	Davisville North	43.712755	-79.388514	Best Western Roehampton Hotel & Suites	43.708878	-79.390880	Hotel
8	Davisville North	43.712755	-79.388514	Gym	43.713126	-79.393537	Gym
9	North Toronto West	43.714523	-79.406960	St. Clements - Yonge Parkette	43.712062	-79.404255	Park
10	North Toronto West	43.714523	-79.406960	Lytton Park	43.714954	-79.411970	Playground
11	North Toronto West	43.714523	-79.406960	NTCC Swimming Pool	43.710553	-79.405786	Gym Pool
12	North Toronto West	43.714523	-79.406960	Rosalind's Garden Oasis	43.712189	-79.411978	Garden
13	Davisville	43.703395	-79.385964	Jules Cafe Patisserie	43.704138	-79.388413	Dessert Shop
14	Davisville	43.703395	-79.385964	Thobors Boulangerie Patisserie Café	43.704514	-79.388616	Café

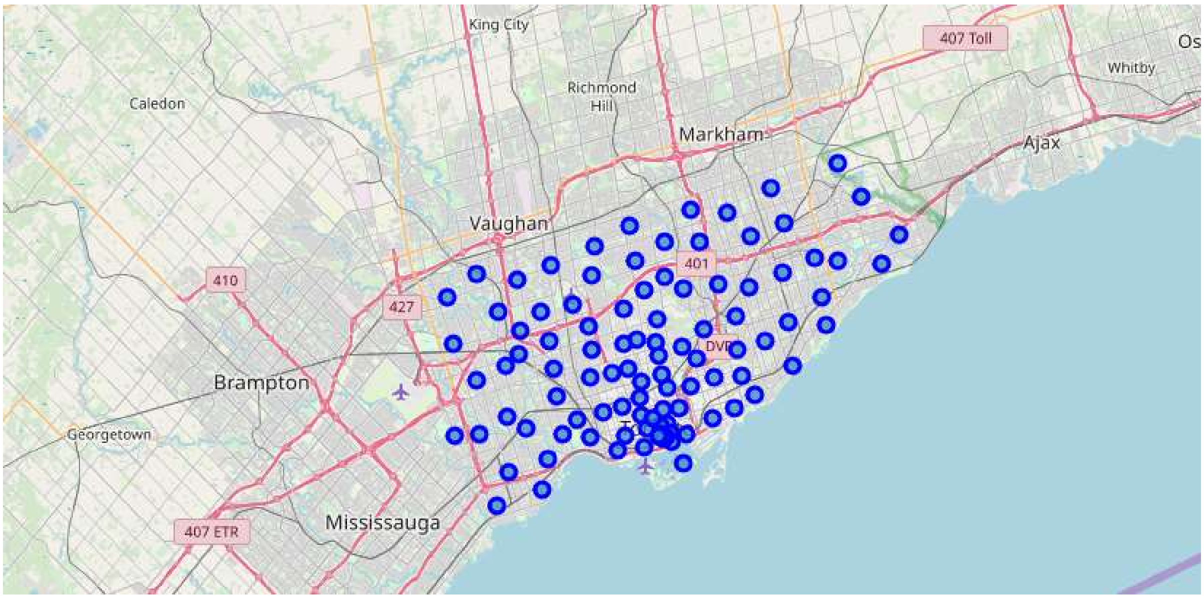
3. Exploratory Data Analysis (EDA)

3.1 Geospatial analysis and visualization with Folium and Leaflet

An interactive map could be drawn using the Folium Python library. This uses the coordinates that we have already fetched and added to our data.

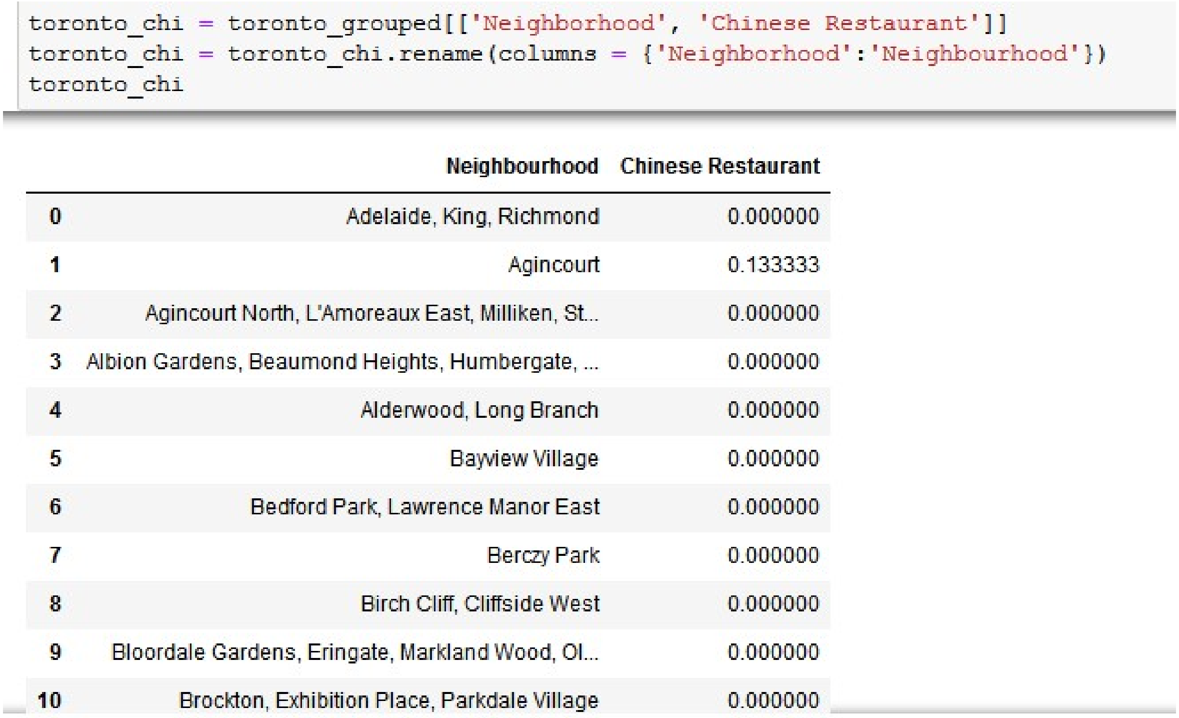


```
# create map of Toronto using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

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



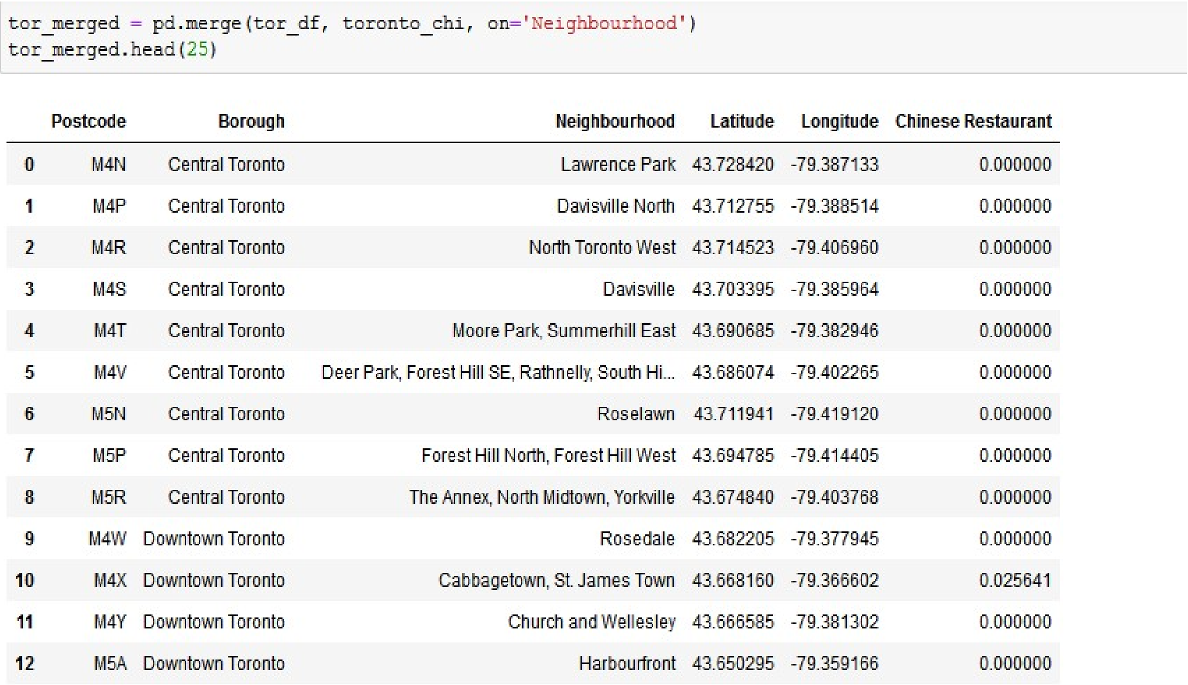
3.2 Discovering relationship between neighborhoods and Chinese Restaurants

To achieve this step, we need to extract the venues data that we already have in our dataframe. We will separate this data and use one hot encoding to create dummy variables. We will merge this data with the neighborhood data having the coordinates.



```
toronto_chi = toronto_grouped[['Neighborhood', 'Chinese Restaurant']]
toronto_chi = toronto_chi.rename(columns = {'Neighborhood':'Neighbourhood'})
toronto_chi
```

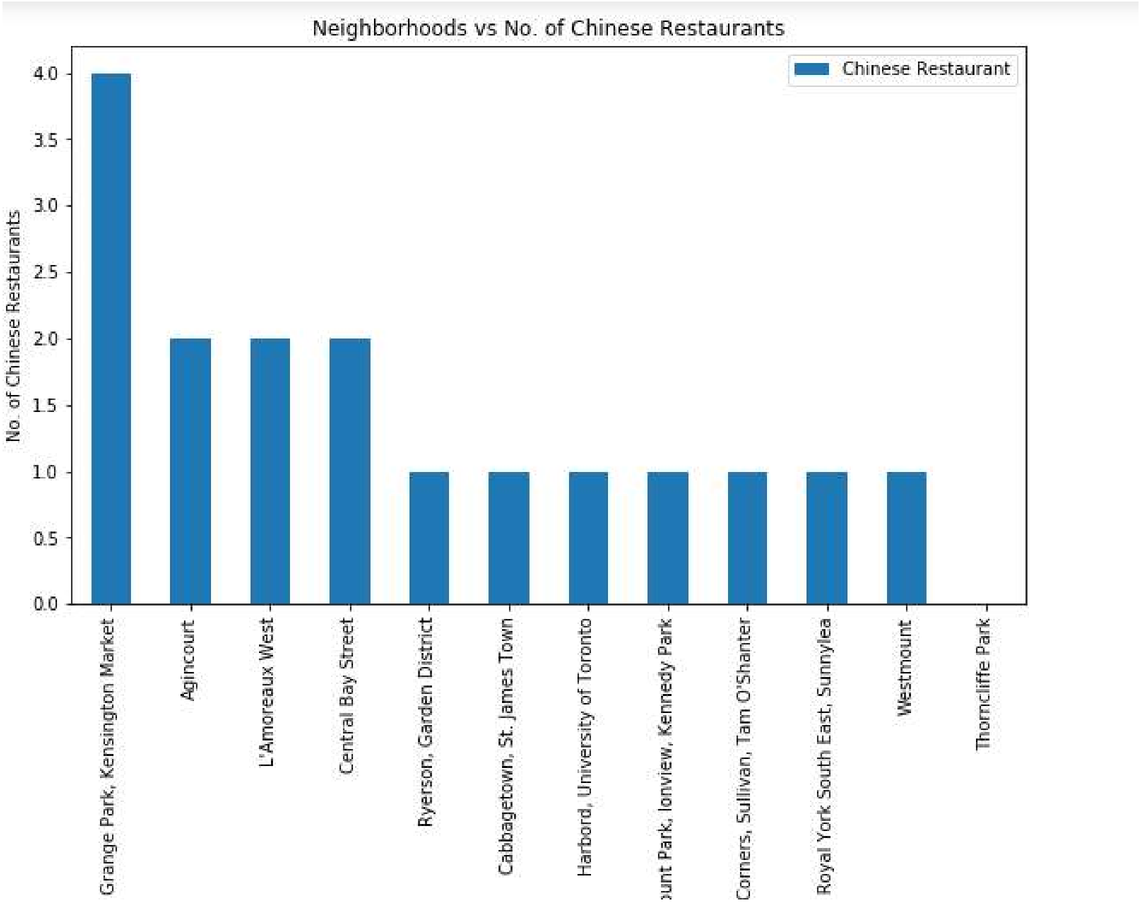
	Neighbourhood	Chinese Restaurant
0	Adelaide, King, Richmond	0.000000
1	Agincourt	0.133333
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.000000
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.000000
4	Alderwood, Long Branch	0.000000
5	Bayview Village	0.000000
6	Bedford Park, Lawrence Manor East	0.000000
7	Berczy Park	0.000000
8	Birch Cliff, Cliffside West	0.000000
9	Bloordale Gardens, Eringate, Markland Wood, Ol...	0.000000
10	Brockton, Exhibition Place, Parkdale Village	0.000000



```
tor_merged = pd.merge(tor_df, toronto_chi, on='Neighbourhood')
tor_merged.head(25)
```

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Chinese Restaurant
0	M4N	Central Toronto	Lawrence Park	43.728420	-79.387133	0.000000
1	M4P	Central Toronto	Davisville North	43.712755	-79.388514	0.000000
2	M4R	Central Toronto	North Toronto West	43.714523	-79.406960	0.000000
3	M4S	Central Toronto	Davisville	43.703395	-79.385964	0.000000
4	M4T	Central Toronto	Moore Park, Summerhill East	43.690685	-79.382946	0.000000
5	M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686074	-79.402265	0.000000
6	M5N	Central Toronto	Roselawn	43.711941	-79.419120	0.000000
7	M5P	Central Toronto	Forest Hill North, Forest Hill West	43.694785	-79.414405	0.000000
8	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.674840	-79.403768	0.000000
9	M4W	Downtown Toronto	Rosedale	43.682205	-79.377945	0.000000
10	M4X	Downtown Toronto	Cabbagetown, St. James Town	43.668160	-79.366602	0.025641
11	M4Y	Downtown Toronto	Church and Wellesley	43.666585	-79.381302	0.000000
12	M5A	Downtown Toronto	Harbourfront	43.650295	-79.359166	0.000000

Next, we will try to visualize this data to compare the neighborhoods and the frequency of Chinese restaurants in each neighborhood.

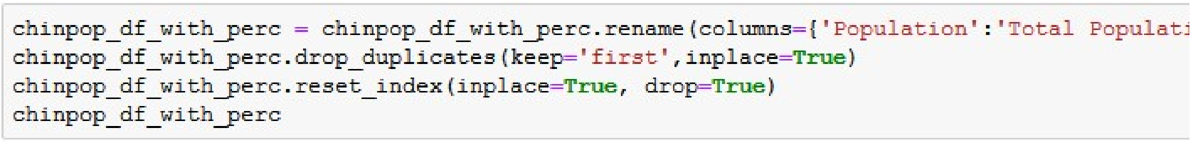


This plot shows us the comparisons clearly. We can clearly see which Neighborhoods have a greater number of Chinese restaurants and which do not. This comparison will further help us identify areas which would be suitable to achieve our goal.

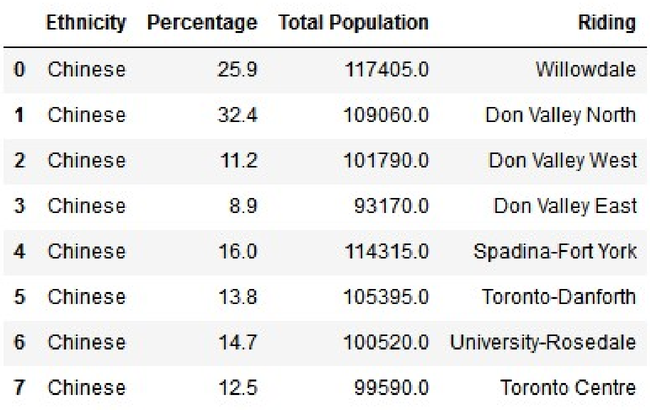
3.3 Discovering relationship between neighborhoods and Chinese Restaurants

Next, we need to understand which areas are more populated with Chinese ethnicity. We will analyze the neighborhoods and identify the ones with high density of Chinese population.

To achieve this, we will gather more data from Wikipedia about the ethnic distribution of population in Toronto and separate out only the Chinese population.



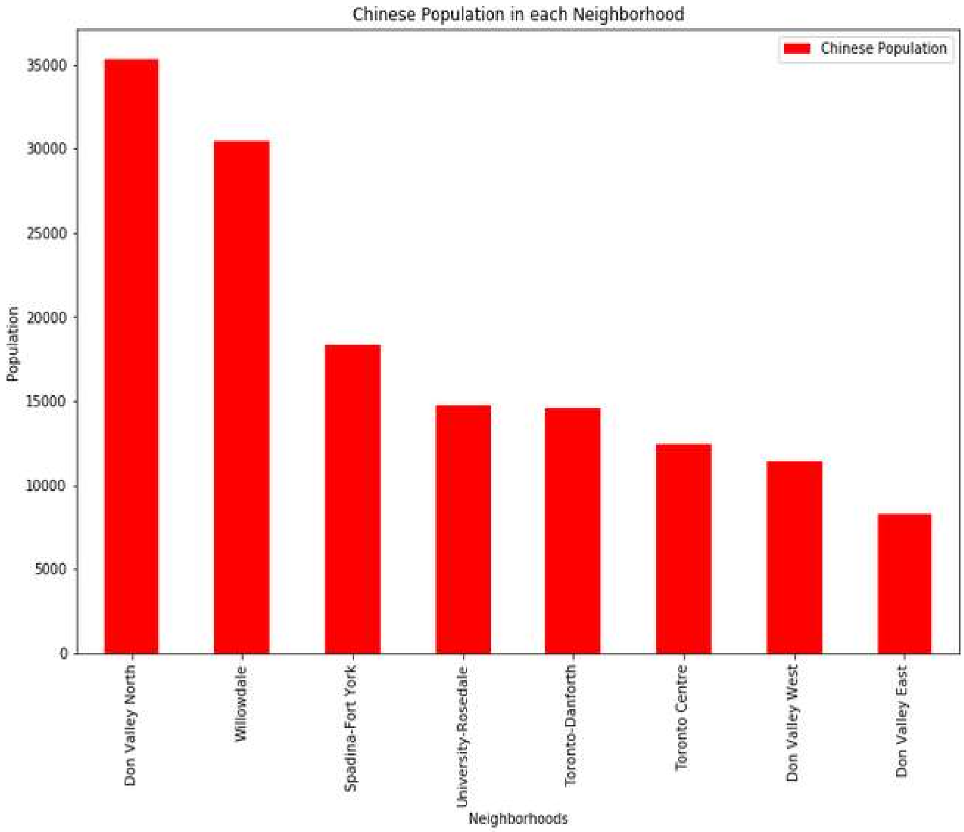
```
chinpop_df_with_perc = chinpop_df_with_perc.rename(columns={'Population':'Total Populati
chinpop_df_with_perc.drop_duplicates(keep='first',inplace=True)
chinpop_df_with_perc.reset_index(inplace=True, drop=True)
chinpop_df_with_perc
```



	Ethnicity	Percentage	Total Population	Riding
0	Chinese	25.9	117405.0	Willowdale
1	Chinese	32.4	109060.0	Don Valley North
2	Chinese	11.2	101790.0	Don Valley West
3	Chinese	8.9	93170.0	Don Valley East
4	Chinese	16.0	114315.0	Spadina-Fort York
5	Chinese	13.8	105395.0	Toronto-Danforth
6	Chinese	14.7	100520.0	University-Rosedale
7	Chinese	12.5	99590.0	Toronto Centre

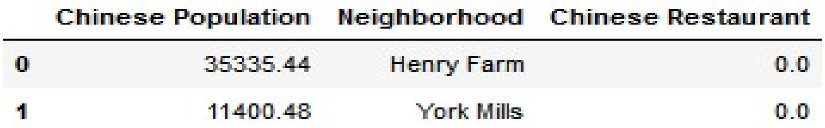
The above code snippet and the table show the Ridings with highest distribution of Chinese population.

We will visualize this data to get a better insight.



3.4 Discovering relationship between Chinese population and Chinese restaurants

We will now compare the Chinese restaurant data and Chinese population data in order to check if there is any relationship between both of them.



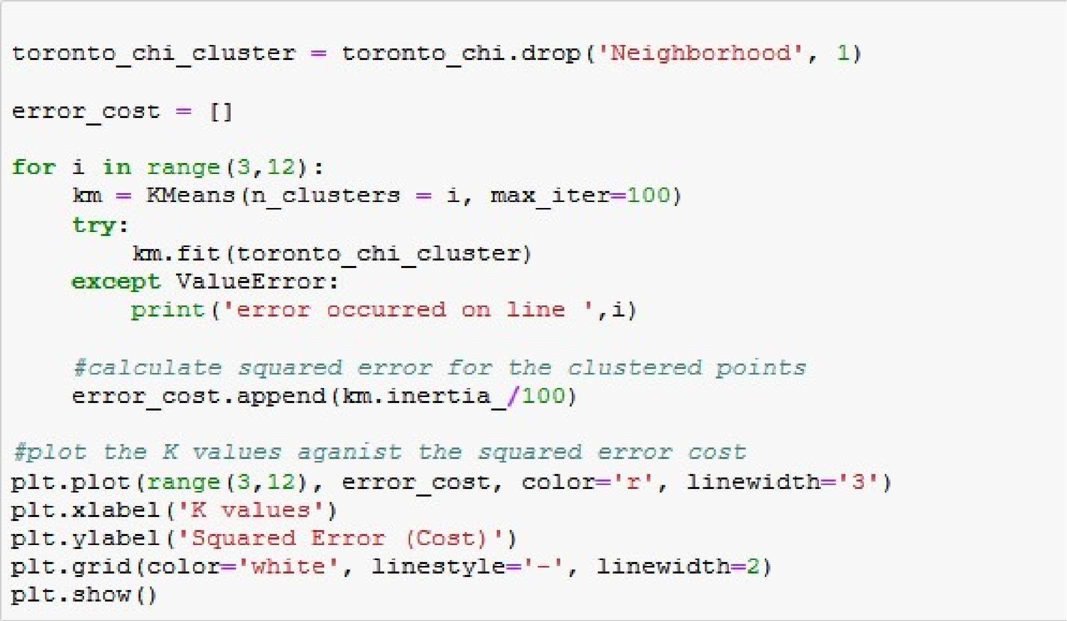
	Chinese Population	Neighborhood	Chinese Restaurant
0	35335.44	Henry Farm	0.0
1	11400.48	York Mills	0.0

After performing merging and comparisons between these two, we found that there was not any significant relationship seen in the analysis. One of the reasons could be different data sources which cause inconsistency in the data and affected the comparisons. This could be improved by further analysis and performing some cleaning tasks.

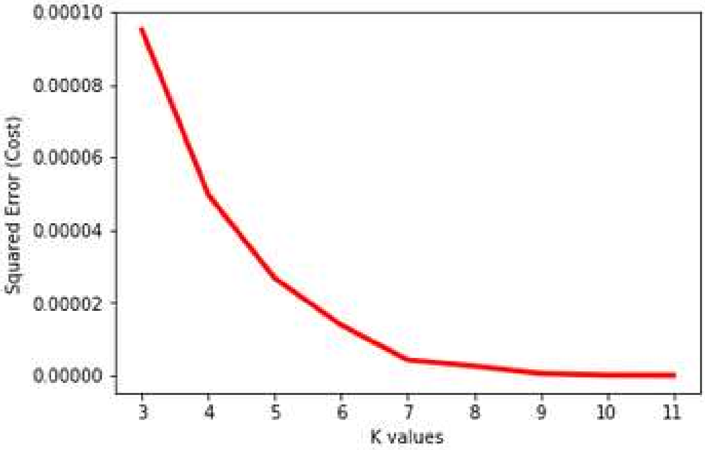
4. Predictive Modelling: K-Means Clustering

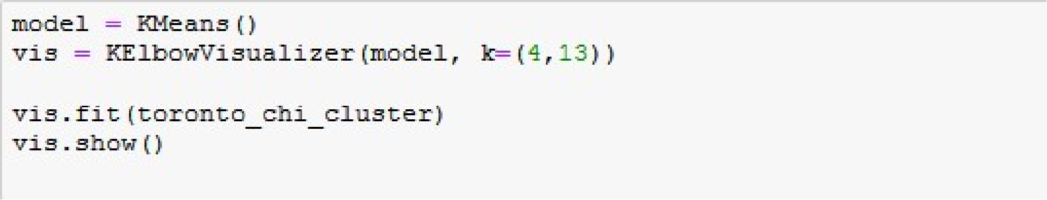
4.1 Clustering the neighborhoods using K-Means Clustering

The initial step in the clustering technique is to identify the best number of clusters as K-Means is an unsupervised technique. We are using the elbow method in order to achieve this goal. We give this method a range of 3-12 clusters and the technique will tell us what number between this range is best for our clustering.

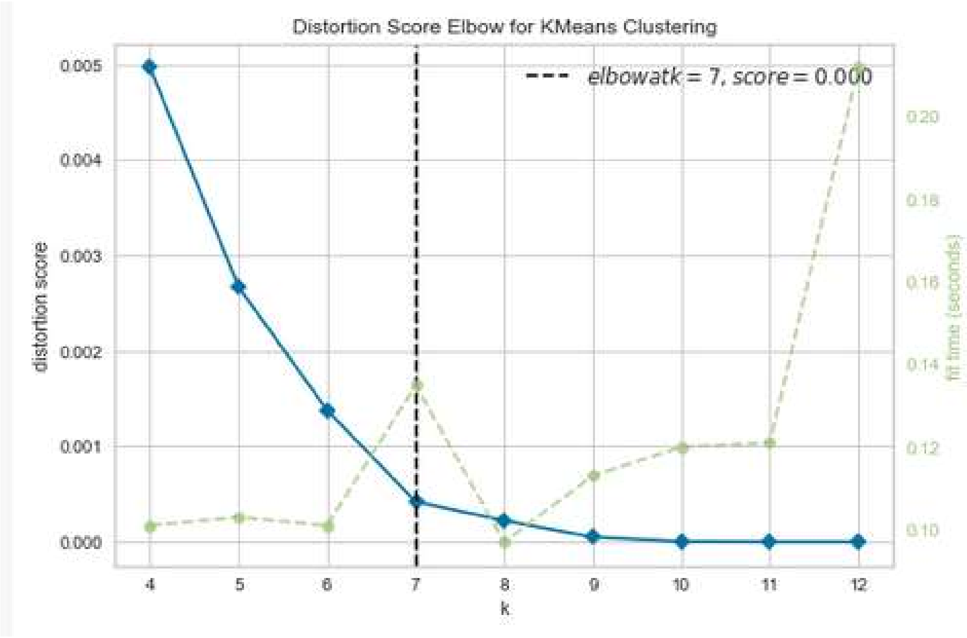


```
toronto_chi_cluster = toronto_chi.drop('Neighborhood', 1)

error_cost = []

for i in range(3,12):
    km = KMeans(n_clusters = i, max_iter=100)
    try:
        km.fit(toronto_chi_cluster)
    except ValueError:
        print('error occurred 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(3,12), error_cost, color='r', linewidth='3')
plt.xlabel('K values')
plt.ylabel('Squared Error (Cost)')
plt.grid(color='white', linestyle='-', linewidth=2)
plt.show()
```



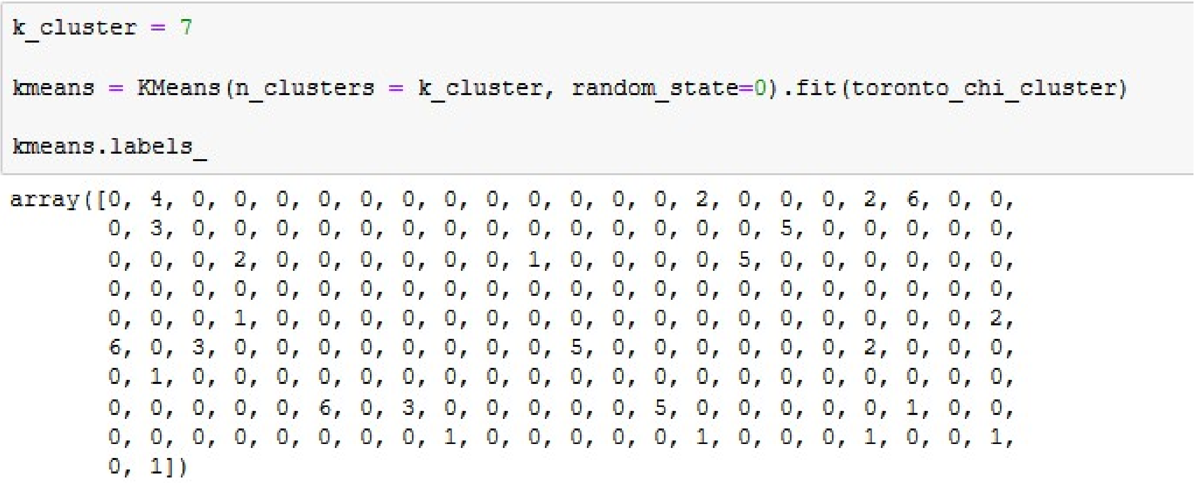


```
model  = KMeans()
vis = KElbowVisualizer(model, k=(4,13))

vis.fit(toronto_chi_cluster)
vis.show()
```



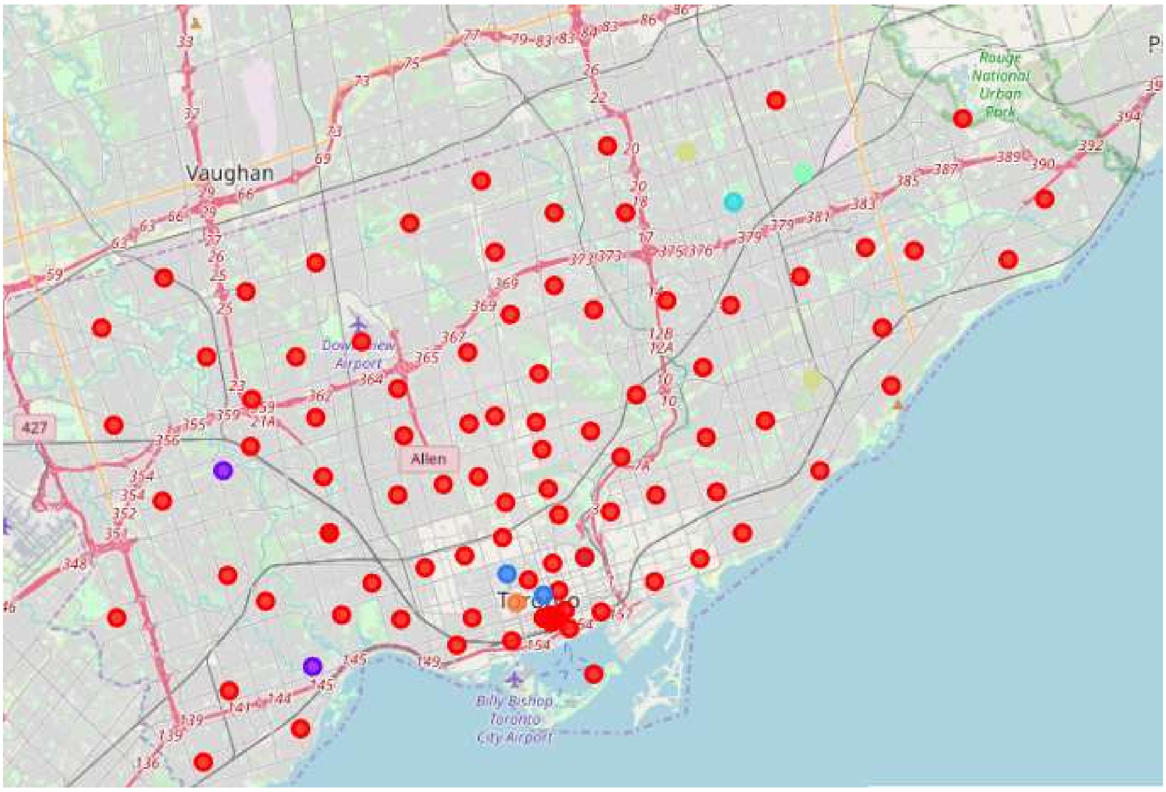
The best number of clusters for our data is shown to be 7. This means we can give this number to our K-Means algorithm and the data will be divided into 7 different clusters.

Applying the K-Means Algorithm with K=7:



```
k_cluster = 7

kmeans = KMeans(n_clusters = k_cluster, random_state=0).fit(toronto_chi_cluster)

kmeans.labels_
```

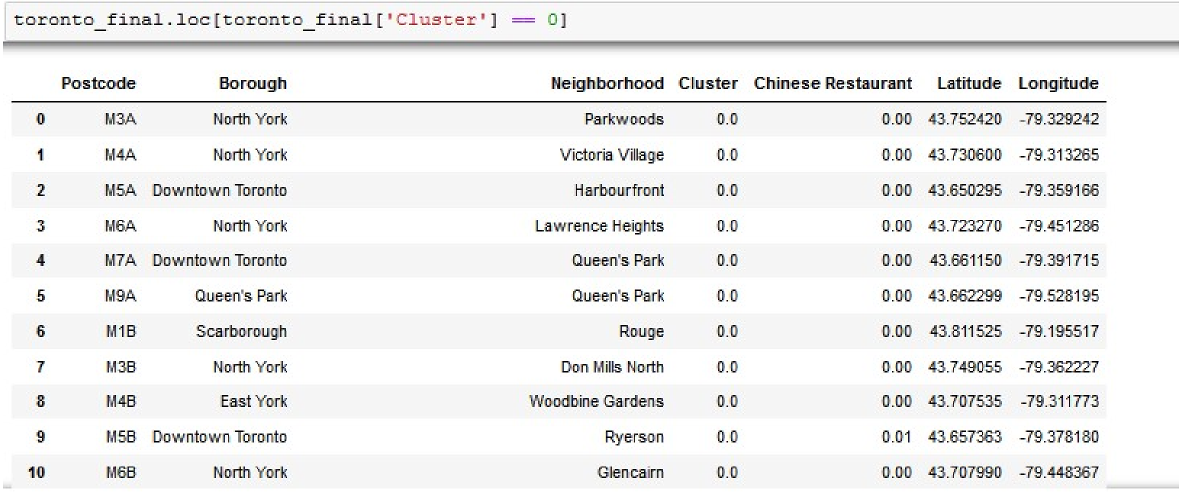
```
array([0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 6, 0, 0,
       0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0,
       0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    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



The above map shows the neighborhoods and the clusters assigned to them. Each cluster has a different color.

Let’s analyze each cluster to identify which neighborhoods are densely populated with Chinese restaurants and how they could be suitable for our mission.

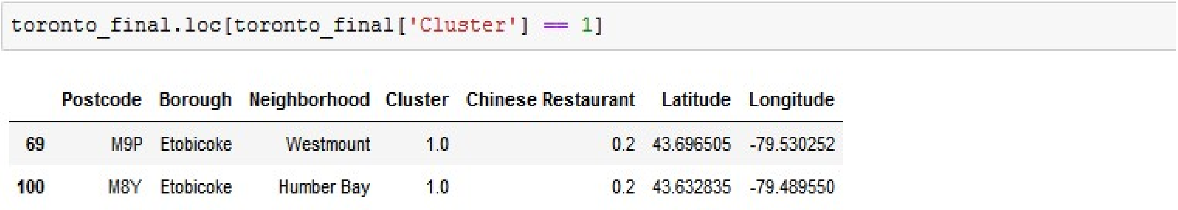
Cluster 0: The cluster 0 consists of those areas which hardly have or do not at all have Chinese restaurants. These areas could be identified by the red markers on the map.



```
toronto_final.loc[toronto_final['Cluster'] == 0]
```

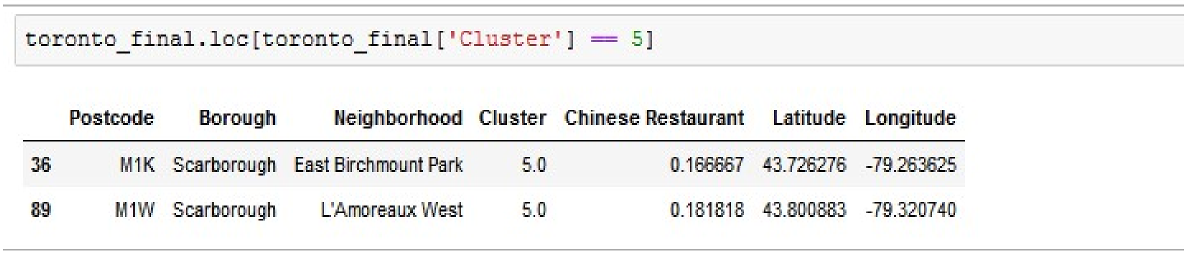
	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
0	M3A	North York	Parkwoods	0.0	0.00	43.752420	-79.329242
1	M4A	North York	Victoria Village	0.0	0.00	43.730600	-79.313265
2	M5A	Downtown Toronto	Harbourfront	0.0	0.00	43.650295	-79.359166
3	M6A	North York	Lawrence Heights	0.0	0.00	43.723270	-79.451286
4	M7A	Downtown Toronto	Queen's Park	0.0	0.00	43.661150	-79.391715
5	M9A	Queen's Park	Queen's Park	0.0	0.00	43.662299	-79.528195
6	M1B	Scarborough	Rouge	0.0	0.00	43.811525	-79.195517
7	M3B	North York	Don Mills North	0.0	0.00	43.749055	-79.362227
8	M4B	East York	Woodbine Gardens	0.0	0.00	43.707535	-79.311773
9	M5B	Downtown Toronto	Ryerson	0.0	0.01	43.657363	-79.378180
10	M6B	North York	Glencairn	0.0	0.00	43.707990	-79.448367

Cluster 1 & Cluster 5: These clusters consists of neighborhoods which are significantly populated with Chinese Restaurants and could be identified by Purple and Yellow colored markers on the map respectively.



```
toronto_final.loc[toronto_final['Cluster'] == 1]
```

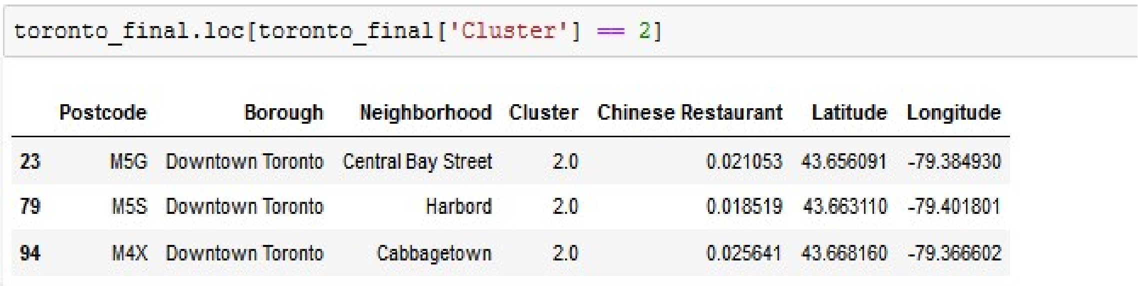
	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
69	M9P	Etobicoke	Westmount	1.0	0.2	43.696505	-79.530252
100	M8Y	Etobicoke	Humber Bay	1.0	0.2	43.632835	-79.489550



```
toronto_final.loc[toronto_final['Cluster'] == 5]
```

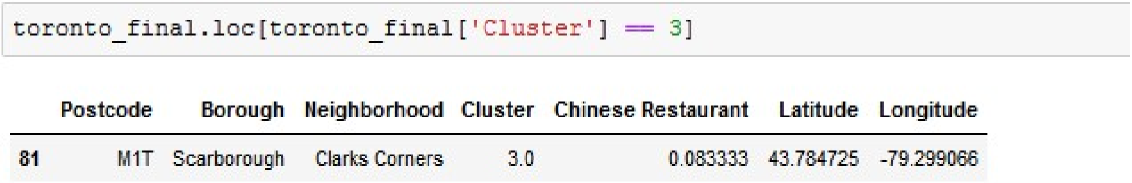
	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
36	M1K	Scarborough	East Birchmount Park	5.0	0.166667	43.726276	-79.263625
89	M1W	Scarborough	L'Amoreaux West	5.0	0.181818	43.800883	-79.320740

Cluster 2, 3 & 6: These clusters consist of neighborhoods which are moderately populated with Chinese Restaurants and could be identified by dark blue and light blue and orange colored markers on the map respectively.



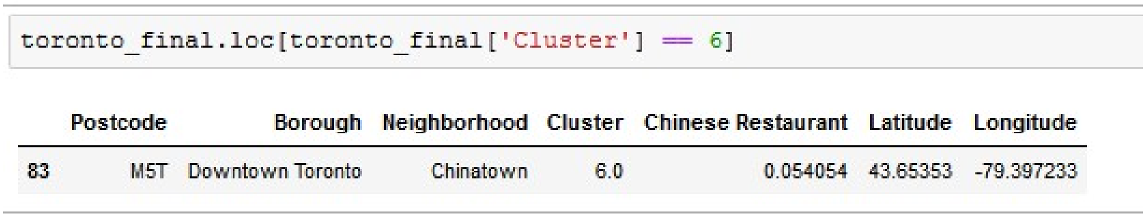
```
toronto_final.loc[toronto_final['Cluster'] == 2]
```

	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
23	M5G	Downtown Toronto	Central Bay Street	2.0	0.021053	43.656091	-79.384930
79	M5S	Downtown Toronto	Harbord	2.0	0.018519	43.663110	-79.401801
94	M4X	Downtown Toronto	Cabbagetown	2.0	0.025641	43.668160	-79.366602



```
toronto_final.loc[toronto_final['Cluster'] == 3]
```

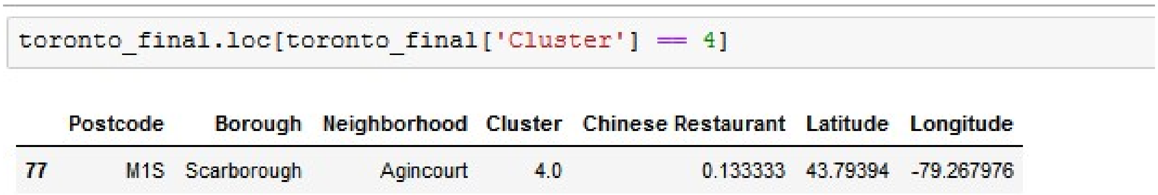
	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
81	M1T	Scarborough	Clarks Corners	3.0	0.083333	43.784725	-79.299066



```
toronto_final.loc[toronto_final['Cluster'] == 6]
```

	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
83	M5T	Downtown Toronto	Chinatown	6.0	0.054054	43.65353	-79.397233

Cluster 4: The Cluster 4 contains the neighborhoods with a little dense population of Chinese Restaurants. It can be identified with sea green colored marks on the map.



```
toronto_final.loc[toronto_final['Cluster'] == 4]
```

	Postcode	Borough	Neighborhood	Cluster	Chinese Restaurant	Latitude	Longitude
77	M1S	Scarborough	Agincourt	4.0	0.133333	43.79394	-79.267976

5. RESULTS AND DISCUSSION

With the analysis of the clusters, we have reached the end of this project. We will be discussing the results and elaborate the findings in the above analysis and visualizations.

In the course of this project, we started off with gathering the data and cleaning it. Then we moved on with the exploratory data analysis and visualizations. This gave us some insights from the data and gave us a clear idea how we should move forward with the analysis and what techniques to use. We looked at the Neighborhood, Population as well as the Venues data and finally combine all of it to find the relationships between them. Finally, we use the K-Means clustering technique in order to distribute the data in different clusters based on the similarities between the features for each record. We have found out that :

- The neighborhoods in Etobicoke, Scarborough, Downtown Toronto and North York have the greatest number of Chinese Restaurants.
- As far as the Chinese population density Is concerned, we have discovered that Willowdale, Don Valley-North, West and East, Spadina-Fort York, Toronto-Dan Forth, University-Rosedale and Toronto Centre hold the highest Chinese population densities.
- After clustering, considering Etobicoke and Downtown Toronto, these areas are mostly populated with Chinese restaurants and there would be a lot of competition around these areas.
- Scarborough and North York have a significantly high number of Chinese populated areas. The neighborhoods based in these areas should be ideal in order to start a new Chinese Restaurant business. Considering Scarborough already has some Chinese Restaurants, the best option would be to choose the areas based under North York as it holds least competition and high population.

The above results are solely based on the data acquired from Wikipedia which might not be updated. However, the data for the venues was acquired from Foursquare API and this data is up to date. The results might differ as there is some inconsistency in the data and if the same procedure is applied to the latest data, we could get better results. Although, there were some inconsistencies in the data, we can say that we have acquired significant results with proper analysis and justifications for each outcome.

6. Conclusion:

We started with a business problem in our minds and step by step, followed the procedure which the real Data Scientists follow. We came through a lot of Python libraries which made our process easier to fetch, visualize, manipulate and analyze the data. We utilized the Foursquare API and Wikipedia to acquire the venues data and scrape tables for demographics and population in Toronto, respectively. Finally, we applied unsupervised machine learning technique, K-Means Clustering using the Scikit-Learn library and Folium to visualize the data on the map.

There were some areas which need improvement and could end up in better results if provided with proper data sources and more time in hand. Additionally, we have only applied one machine learning technique for clustering while there are more techniques which could give better results and we can finalize the final model by comparing these techniques and their results. To conclude, we can say that we have successfully fetched, cleaned, visualized the data and created a model to solve a business problem and hopefully, this analysis could help providing a head-start in solving any similar real-world business problem.