Opening a Chinese Restaurant in Toronto Neighbourhoods

Introduction

Background

Toronto is the most populous city in Canada with an estimate of around half a million Chinese descents residing in it. Chinese are known to be the second largest immigrant group in Canada after south Asians. There are hundreds of Chinese restaurants in GTA which are liked by not just Chinese but by all.

Business Problem

In this project we are looking to explore Toronto neighbourhoods and find the best suitable neighbourhood to open a new Chinese Restaurant.

Target Audience

This project may interest Investors or Entrepreneurs who are looking to open a Chinese restaurant in GTA. It is also useful for Chinese food lovers to find a neighbourhood with many restaurant options.

Data Acquisition

Data Sources

The neighbourhood and borough lists are obtained from Wikipedia page
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The coordinates of each neighbourhoods are obtained from the location
https://cocl.us/Geospatial_data

Population of each neighbourhoods are again obtained from Wikipedia page
https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

All the Chinese restaurants in the Toronto neighbourhoods are extracted from FourSquare API

Data Cleaning

a) Extracted Toronto neighbourhoods from Wikipedia using web scraping techniques:

Scraped the wiki page to obtain a data set containing Postal code, Boroughs and Neighbourhoods.

In [54]:
```
webpage = wp.page("List of postal codes of Canada: M").html().encode("UTF-8")
df = pd.read_html(webpage, header=0)[0]
df.head()
```

Out[54]:

	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

Borough with no Postal codes assigned are ignored. When a post code is assigned to multiple neighbourhoods, they are combined to a single row.

In [55]:
```
df['Neighbourhood'] = np.where(df.Neighbourhood == 'Not assigned', df.Borough, df.Neighbourhood)
df['Neighbourhood'] = df['Neighbourhood'].replace('Not assigned',np.nan)
df['Borough'] = df['Borough'].replace('Not assigned',np.nan)
df = df.dropna(axis=0, how='any', thresh=None, subset=None)
df = df.reset_index()
del df['index']
df_orig = df.copy()
df  = df.groupby(['Postcode', 'Borough']).agg({'Neighbourhood': lambda x: ','.join(tuple(x.tolist()))})
df = df.reset_index()
df.head()
```

Out[55]:

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

b) Geographical coordinates of neighbourhoods

Geo coordinates of the neighbourhoods(Postal Codes) are extracted as CSV from the location https://cocl.us/Geospatial_data

After combining it with the previous neighbourhood data set, we get the below data set.

In [6]:
```
df_coord = pd.read_csv('load_coordinates.csv')
df_coord.columns = ['Postcode','Latitude','Longitude']
df_final = pd.merge(df, df_coord, on='Postcode')
df_final = df_final.drop_duplicates()
df_final.head()
```

Out[6]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

In [7]:
```
print('Number of Broughs in the dataset: {} '.format(len(df_final['Borough'].unique())))
print('Number of Neighbourhoods in the dataset: {} '.format(len(df_final['Neighbourhood'].unique())))
```

```
Number of Broughs in the dataset: 10
Number of Neighbourhoods in the dataset: 103
```

c) Population distribution of neighbourhoods

Since it was difficult to get specifically chinese population in each neighbourhoods, overall population of neighbourhoods are considered. These are extracted from wiki page, “Demographics of Toronto Neighbourhoods”

Get Neighbourhood populations

In [78]:
```
wikipage = wp.page("Demographics of Toronto neighbourhoods").html().encode("UTF-8")
df_pop = pd.read_html(wikipage, header = 0)[1]
df_pop = df_pop[['Name', 'Population']]
df_pop.columns = ["Neighbourhood", "Population"]
df_neigh_pop = pd.merge(df_orig, df_pop, on="Neighbourhood")
df_neigh_pop  = df_neigh_pop.groupby(['Postcode', 'Borough']).agg({'Neighbourhood': lambda x: ','.join(tuple(x.tolist())), 'Population': 'sum' })
df_neigh_pop
```

Out[78]:

		Neighbourhood	Population
Postcode	Borough		
M1B	Scarborough	Rouge,Malvern	67048
M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	36470
M1E	Scarborough	Guildwood,Morningside,West Hill	49924
M1G	Scarborough	Woburn	48507
M1J	Scarborough	Scarborough Village	12796
M1K	Scarborough	Ionview	13025
M1L	Scarborough	Clairlea,Oakridge	24472
M1M	Scarborough	Cliffcrest,Cliffside	23917
M1N	Scarborough	Birch Cliff	12266
M1P	Scarborough	Dorset Park	14189
M1R	Scarborough	Maryvale,Wexford	26644
M1S	Scarborough	Agincourt	44577

d) Get Restaurant list from FourSquare API

Using the location coordinates, a list of venues in the neighbourhoods are obtained from FourSquare API.

In [14]:
```
column_list = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
items = results['response']['groups'][0]['items']

venues = json_normalize(items)
venues =venues.loc[:, column_list]
venues['venue.categories'] = venues.apply(get_categories, axis=1)
venues.columns = [col.split(".")[-1] for col in venues.columns]
venues.head()
```

Out[14]:

	name	categories	lat	lng
0	Downtown Toronto	Neighborhood	43.653232	-79.385296
1	Japango	Sushi Restaurant	43.655268	-79.385165
2	Karine's	Breakfast Spot	43.653699	-79.390743
3	Manpuku まんぷく	Japanese Restaurant	43.653612	-79.390613
4	Nathan Phillips Square	Plaza	43.652270	-79.383516

In [17]:
```
toronto_venues.head()
```

Out[17]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge,Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood,Morningside,West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
3	Guildwood,Morningside,West Hill	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa
4	Guildwood,Morningside,West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant

Filtered the data set to obtain the list of Chinese restaurants in the neighbourhoods

In [18]:
```
toronto_venues[toronto_venues['Venue Category']=='Chinese Restaurant']
```

Out[18]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
47	Dorset Park,Scarborough Town Centre,Wexford He...	43.757410	-79.273304	Kim Kim restaurant	43.753833	-79.276611	Chinese Restaurant
68	Clarks Corners,Sullivan,Tam O'Shanter	43.781638	-79.304302	The Royal Chinese Restaurant 避風塘小炒	43.780505	-79.298844	Chinese Restaurant
82	L'Amoreaux West	43.799525	-79.318389	Mr Congee Chinese Cuisine 龍	43.798879	-79.318335	Chinese

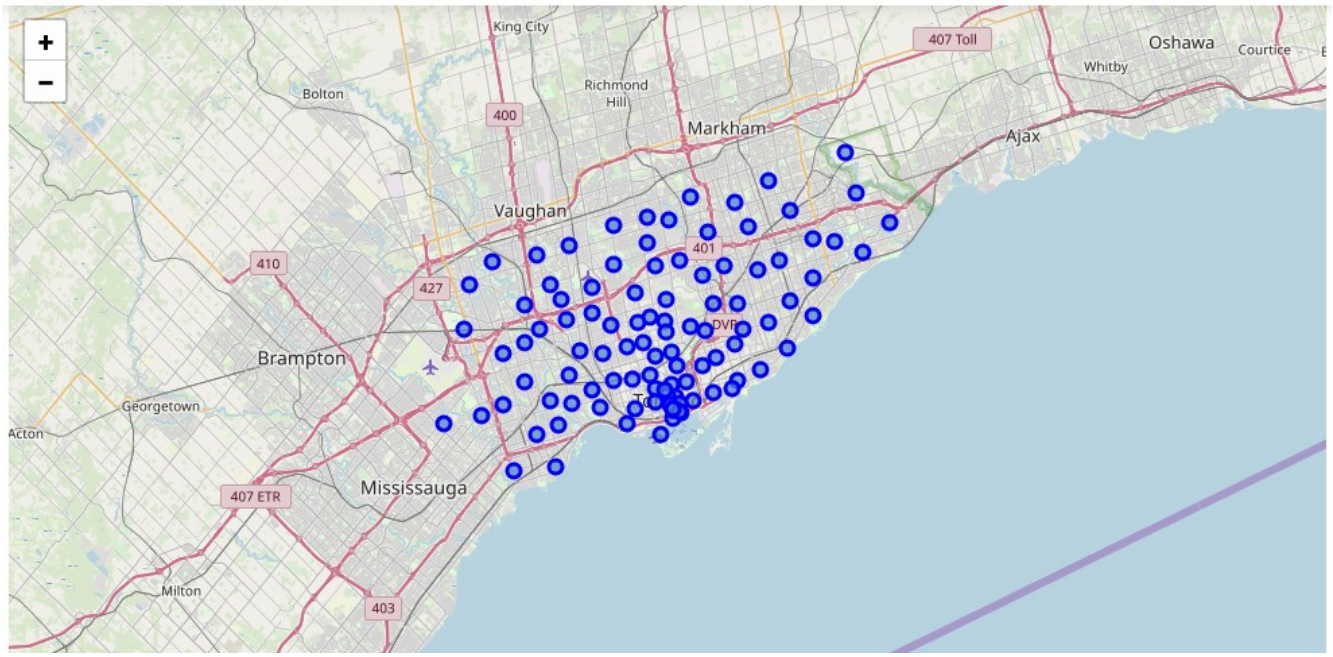
Exploratory Data Analysis

Plotting the neighbourhoods on a map using Folium library

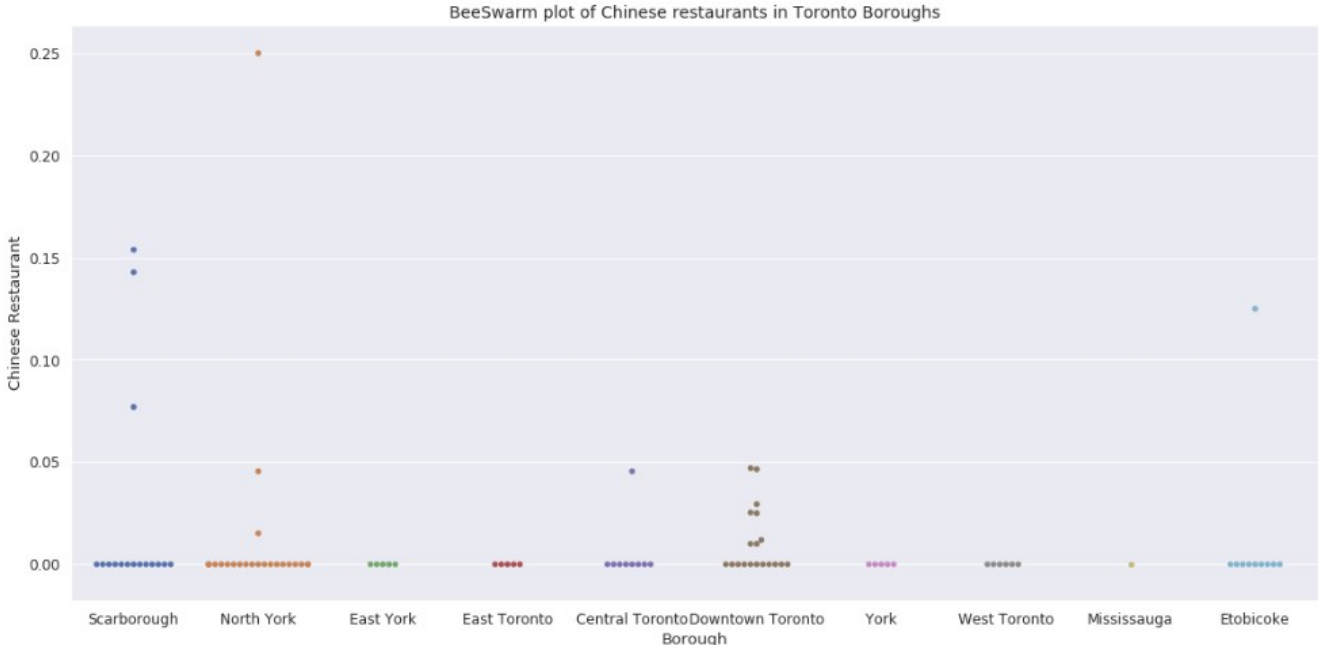
Plot Neighbourhoods on Map

In [20]:
```
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

for lat, lng, borough, neighbourhood in zip(df_final['Latitude'], df_final['Longitude'], df_final['Borough'], df_final['Neighbourhood']):
    label = '{},{}'.format(neighbourhood, 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[20]:

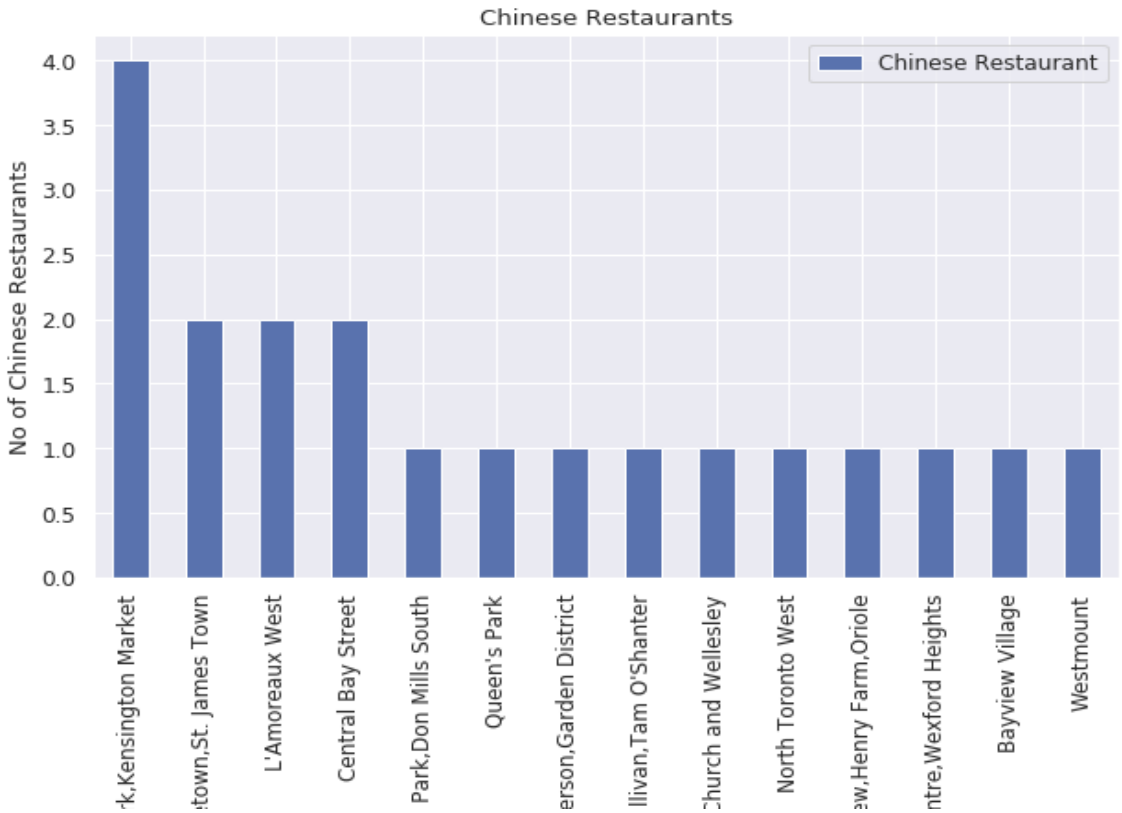


Applying panda one hot encoding and finding relationship between boroughs and restaurants using categorical plotting. Used bee swarm plots.



The plot shows downtown Toronto is densely packed with Chinese restaurants. But boroughs like Mississauga and York not so.

Plotting a graph between neighbourhoods and restaurants

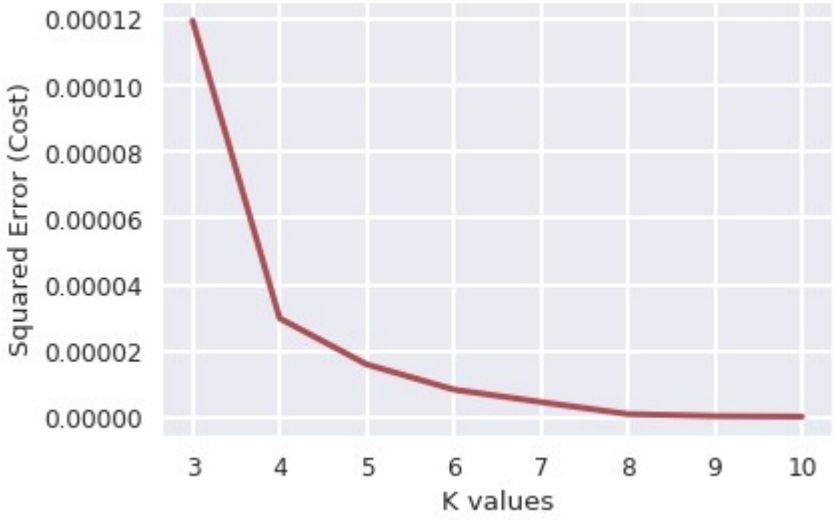


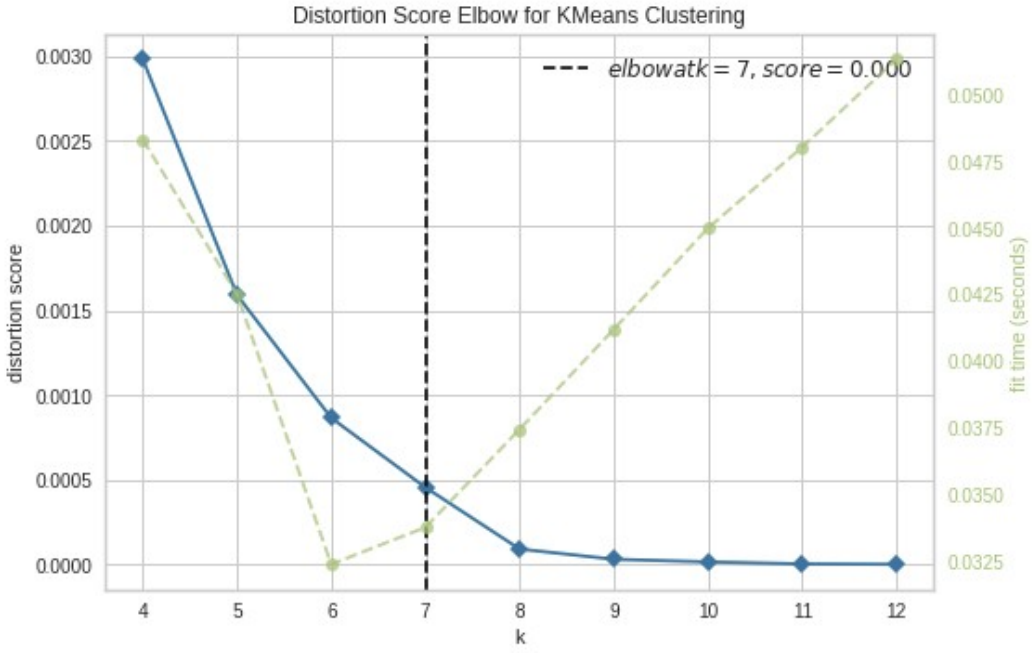
We can notice most of the neighbourhoods with more than 2 restaurants are in Down town Toronto.

Predictive Modelling

Clustering Neighbourhoods of Toronto using K Means Clustering technique/

First in order to find optimum K, we'll plot Squared Mean Error vs K value graph and Distortion Score Elbow Graph



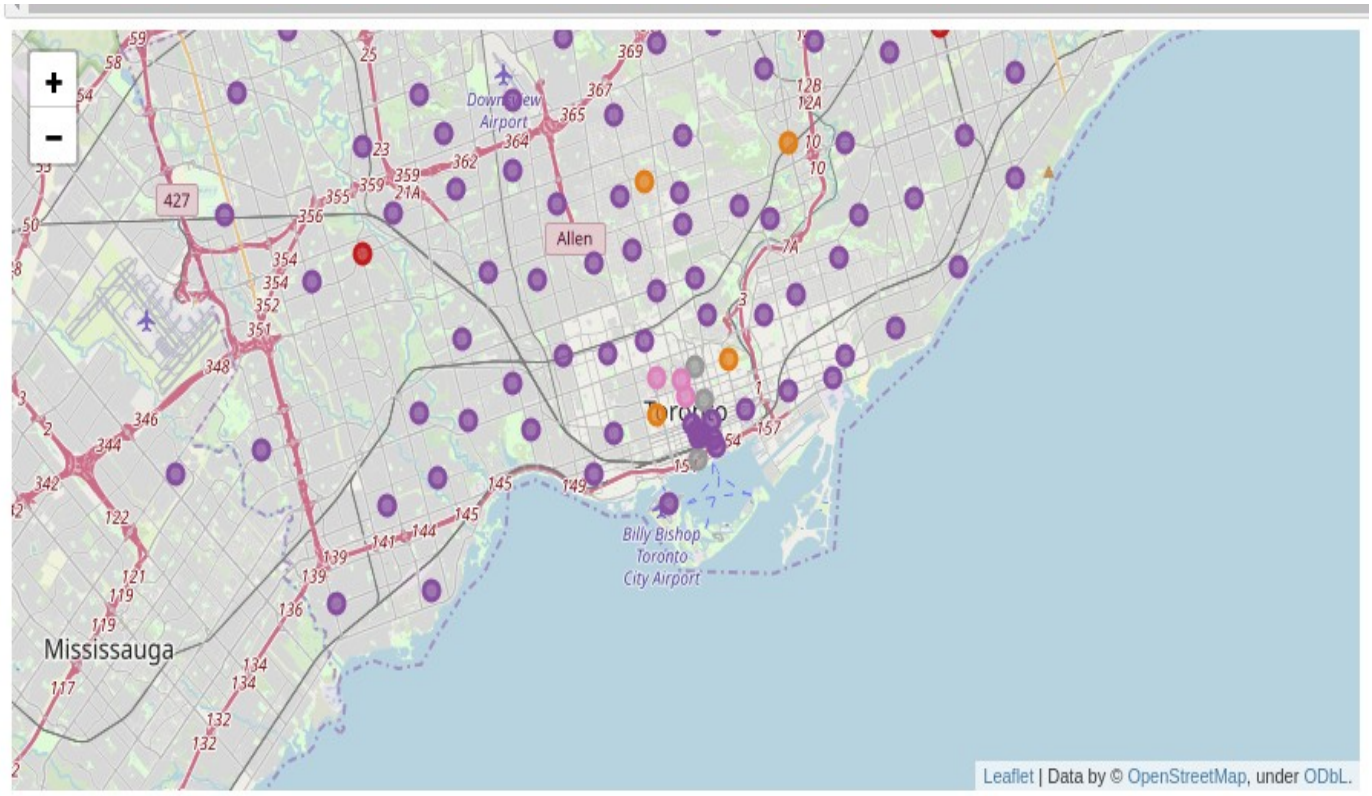


After analyzing the 2 graphs we can confidently say K=7 is the best value for the data set.

Clustering neighbourhoods with K=7,

```
In [37]: # Clustering the Toronto Neighborhood Using K-Means with K = 7
kclusters = 7
toronto_clustering = toronto_subset.drop('Neighbourhood', 1)
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_clustering)
kmeans

Out[37]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=7, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=0, tol=0.0001, verbose=0)
```

Plot the neighbourhoods by Cluster on the Map



Cluster Analysis

We have 7 clusters and cluster 3 has the least number of Chinese restaurants and as expected has maximum number of neighbourhoods in it.

In [44]:
```
# Cluster 3:
toronto_merged.loc[toronto_merged['Cluster'] == 3]
```

Out[44]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster	Chinese Restaurant
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	3	0.0
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	3	0.0
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	3	0.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	3	0.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	3	0.0
...	...	...	...	...	...	...	...
93	M9M	North York	Emery,Humberlea	43.724766	-79.532242	3	0.0
94	M9N	York	Weston	43.706876	-79.518188	3	0.0
96	M9R	Etobicoke	Kingsview Village,Martin Grove Gardens,Richvie...	43.688905	-79.554724	3	0.0
97	M9V	Etobicoke	Albion Gardens,Beaumond Heights,Humbergate,Jam...	43.739416	-79.588437	3	0.0
98	M9W	Etobicoke	Northwest	43.706748	-79.594054	3	0.0

83 rows × 7 columns

Combining population data set with cluster 3 to obtain the neighbourhoods in Cluster 3 with highest population.

Cluster 3 population

In [80]:
```
df_favourable = toronto_merged.loc[toronto_merged['Cluster'] == 3]
df_fav_pop = pd.merge(df_favourable, df_neigh_pop, on='Postcode')
df_fav_pop.sort_values(by=['Population'], ascending=False).head()
```

Out[80]:

	Postcode	Borough	Neighbourhood_x	Latitude	Longitude	Cluster	Chinese Restaurant	Neighbourhood_y
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	3	0.0	Rouge,Malvern
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	3	0.0	Guildwood,Morningside Hill
3	M1G	Scarborough	Woburn	43.770992	-79.216917	3	0.0	Woburn
10	M1S	Scarborough	Agincourt	43.794200	-79.262029	3	0.0	Agincourt
34	M6R	West Toronto	Parkdale,Roncesvalles	43.648960	-79.456325	3	0.0	Parkdale,Roncesvalles

Results

At the end of this analysis, we have found the following results:

1. From the K Means Clustering we could infer that the neighbourhoods in cluster 3 has no or less number of Chinese restaurants.
2. From the Bee Swarm plot and pie chart of boroughs we could deduce that there are higher density of Chinese restaurants located in the Downtown Toronto and very few in Mississauga.
3. Neighbourhoods in cluster 3 from Boroughs East Toronto, West Toronto, East York, York and Mississauga Boroughs have no Chinese restaurants.
4. In Cluster 3, neighbourhoods Rouge & Malvern in Scarborough have the largest population and since there are no Chinese Restaurants at these neighbourhoods, they are suitable for opening a new restaurant.

Conclusion

This is simple and yet a powerful tool to find a suitable/favourable neighbourhood to open any business venture.

The drawback of this method in this scenario is that the neighbourhood population by ethnicity could not be obtained, therefore overall population is considered. Also it makes an assumption that all the population not just Chinese would be interested to try Chinese food in restaurants.