Exploring Toronto Neighborhoods – to open a Chinese Restaurant

As a part of the IBM Data Science program capstone project, I worked on the real datasets to get an experience of what a data scientist goes through in the real life, Main object of this project were to define business problem, using the web data through web scraping, and using Foursquare location data to compare different neighborhoods of Toronto to figure out which neighborhood is profitable for starting a new restaurant. In this project, I will go through the whole process step by step manner from problem targeting, data preparation to final analysis result. Last but not the least, I will provide my suggestion which can be leveraged by the business stakeholders to make their decisions.

1. **Description of the Business Problem**
   In this project, I will go through all steps to make a decision whether it is a good idea to open a Chinese restaurant.
   I analyze the neighborhoods in Toronto to identify the most profitable area to place it.
   Target Audience:
   Business personnel who wants to invest or open a Chinese restaurant in Toronto;
   Freelancers who loves to have their own restaurant as a side business;
   Chinese crowd wants to find neighborhoods with lots of option for Chinese restaurants.

2. **Data acquisition & cleaning:**
   2.1 Data Sources
   I used (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.
   Then I used ("https://cocl.us/Geospatial_data")csv file to get all the geographical coordinates of the neighborhoods.
   To get information about the distribution of population by their ethnicity I'm using "Demographics of Toronto" (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) wiki page. Using this page, I'm going to identify the neighborhoods which are densely populated with Indians as it might be helpful in identifying the suitable neighborhood to open a new Chinese restaurant.
   To get location and other information about various venues in Toronto I'm using Foursquare's explore API.(https://developer.foursquare.com/docs)

   2.2 Data Cleaning
   Firstly, I scraped Toronto neighborhoods table from Wikipedia.

```
df.head()
```

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

Then I added geographical coordinates to the neighborhoods by extracting the data present in the Geospatial Data csv file.

```
lat_lng = lat_lng.rename(columns={'Postal Code':'Postcode'})
toronto_df = pd.merge(df, lat_lng, on='Postcode')
toronto_df.head()
```

	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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

Thirdly, I scraped the distribution of population from Wikipedia, examining those neighborhood's population to identify the densely populated neighborhoods with Chinese population.

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

Fourthly, I got location data from Foursquare, by choosing 100 popular spots for each neighborhood within a radius of 1km.

```
print('{} venues were returned by Foursquare.'.format(toronto_venues.shape[0]))
toronto_venues.head()
```

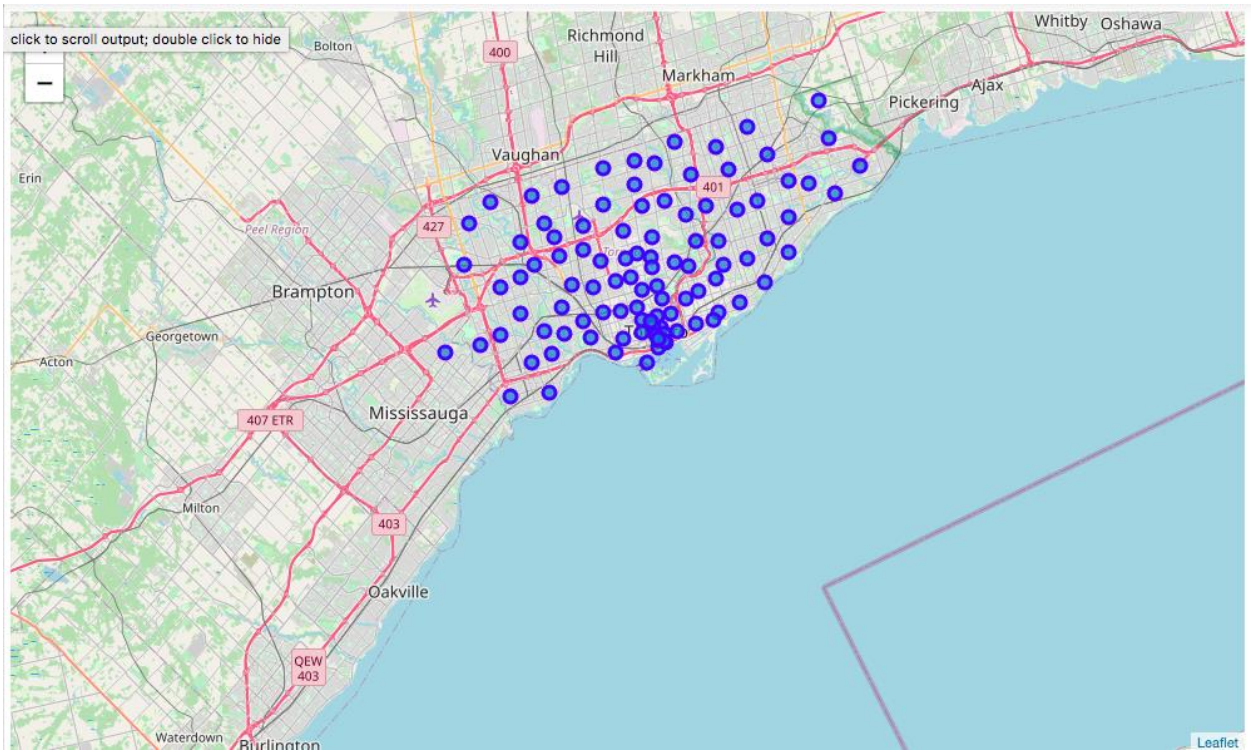
```
2120 venues were returned by Foursquare.
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Malvern, Rouge	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Chris Effects Painting	43.784343	-79.163742	Construction & Landscaping
3	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank

3. Exploratory Data Analysis

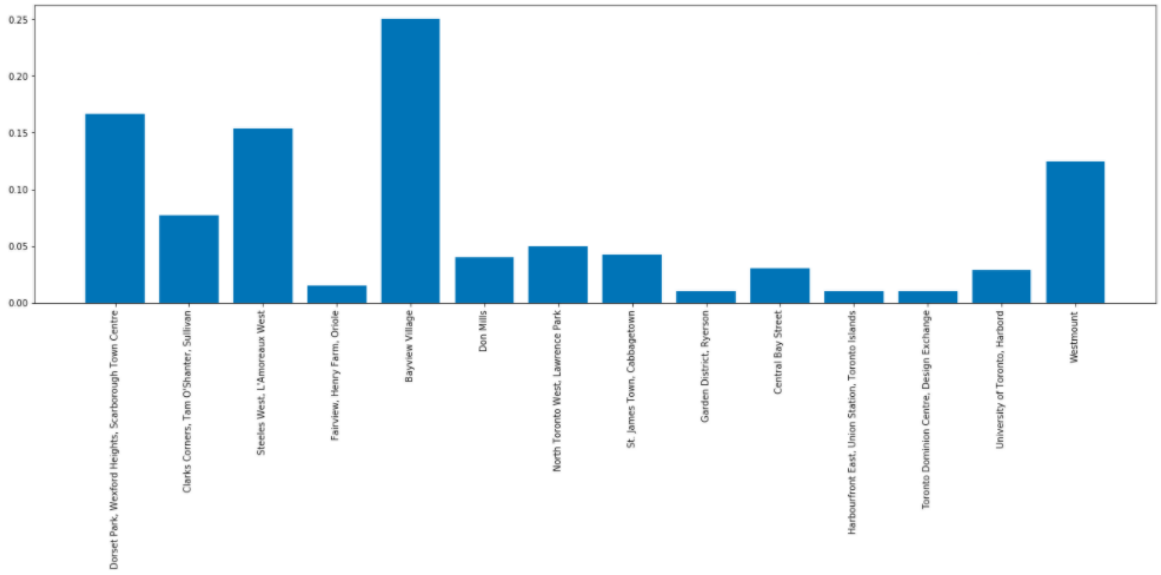
3.1 Visualization

The next step, I used Folium Library and Leaflet Map to draw an interactive map using coordinate data.



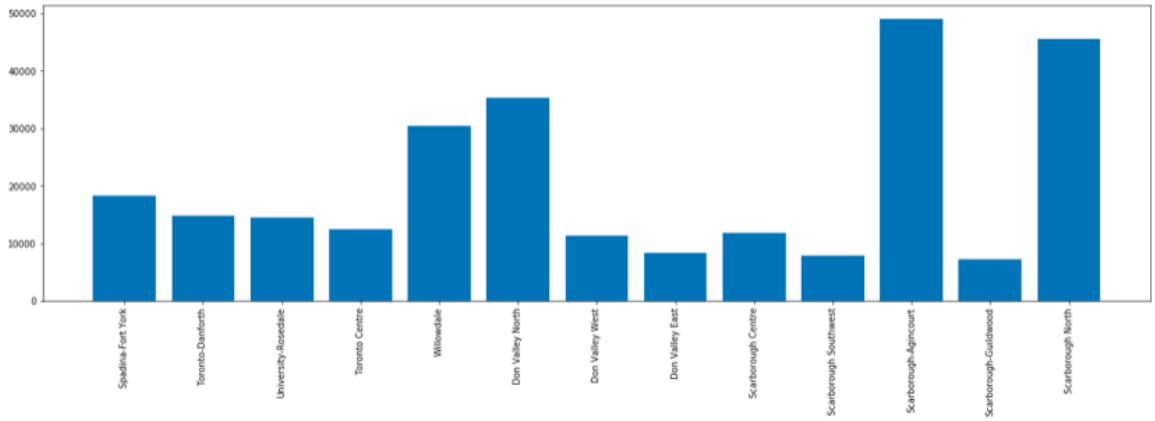
3.2 Relationship between neighborhood and Chinese restaurants
I used bar plots to identify the boroughs with densely populated Chinese restaurants.

```
neighbor_form = toronto_merged[toronto_merged['Chinese Restaurant']>0]
plt.figure(figsize = (22,6))
ax2 = plt.bar(x = 'Neighborhood', height = 'Chinese Restaurant', data = neighbor_form)
plt.xticks(rotation=90)
plt.show()
```



3.3 Relationship between neighborhood and Chinese population
I analyzed the neighborhoods and identified the neighborhoods with the highest number of Chinese populations.

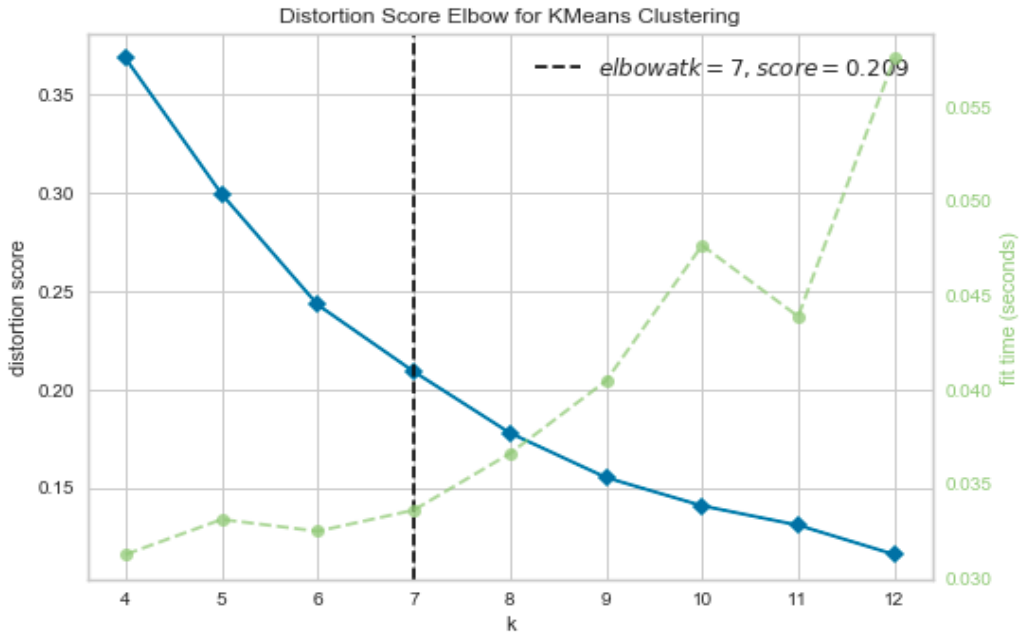
```
plt.figure(figsize = (22,6))
ax2 = plt.bar(x = 'Riding', height = 'Chinese_Population', data = df_Chinese_population)
plt.xticks(rotation=90)
plt.show()
```



4. Predictive Modeling

4.1 Clustering neighborhoods of Toronto

First step in k-means clustering is to identify best K value, which is the number of clusters in a given dataset. To do so I used elbow methods on the Toronto dataset with Chinese restaurants' percentage, aiming to do k-means cluster.



After analyzing using elbow method with distortion score and squared error for each k value, K=7 is the best value.

4.2 Examining the clusters

Cluster 4 contains Boroughs with large numbers of Chinese Restaurants.

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

	Cluster Labels	Postcode	Borough	Neighborhood	Latitude	Longitude	Chinese Restaurant
10	4	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304	0.166667
15	4	M1W	Scarborough	Steeles West, L'Amoreaux West	43.799525	-79.318389	0.153846
18	4	M2K	North York	Bayview Village	43.786947	-79.385975	0.250000

5. Results

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

- In those 10 boroughs we identified that only North York & Scarborough & Etobicoke have high amounts of Chinese restaurants with the help of bar plots between Number of Chinese restaurants in Borough of Toronto.

- In those neighborhoods in Toronto, Bayview Village & Dorset Park Wexford Scarborough Town Centre & Steeles West L'Amoreaux West have high amounts of Chinese restaurants with the help of bar plots between Number of Chinese restaurants in neighborhoods of Toronto.

- In all the ridings, Scarborough & Don Valley North (North York) & Willowdale (North York) & Spadina_Fort York (East York) are densely populated with Chinese crowd ridings.

- Since Scarborough and North York are taken by a large number of Chinese Restaurants, is it is better idea to leave those boroughs out and consider East York for the new restaurant's location.

- After careful consideration it is a good idea to open a new Chinese restaurant in East York borough since it has high number of Chinese populations which gives a higher number of customers possibility and lower competition since very less Indian restaurants in the neighborhoods.