COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a Restaurant in Toronto Neighborhood

By

Snehalatha Balakrishnan

(December, 2020)

1. Introduction:

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. The diverse population of Toronto reflects its current and historical 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. Food plays an important role reflecting the culture and the food business is the more profitable sector in the city like Toronto which has people with different ethnic backgrounds. Similar to any business, decision making in choosing the appropriate cuisine and location will play the dominant role for opening a restaurant.

Business problem:

The objective of this capstone project is to analyze and suggest the appropriate location for opening the restaurant and the type of the cuisine in order to run the restaurant successfully. The project will use the data science methodologies and approaches in order to provide the insight and suggestion to the business question: In the Toronto city, if a restauranteur wants to open the restaurant, where can restaurateur open the restaurant? hat category or cuisine can the restauranteur?

Target Audience:

- People looking to open the new restaurant
- Explore the restaurants available across the city
- Travelers who love the Indian / Chinese restaurants.

2. Data

Data Collection

For this project the following data is used.

- List of neighborhoods in the city of Toronto with the postal code
    - Data Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
    - Description: This data set contains the 3 columns Postal code, borough and neighborhoods. The data can be obtained by web scraping technique in python. It has 180 rows.

```
In [4]: df_list[0].head()
```

Out[4]:

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Fig 1: List of neighborhoods in the city of Toronto with the postal code

- Postal code wise Geospatial coordinates for the city of Toronto
    - Data Source: https://cocl.us/Geospatial_data/Geospatial_Coordinates.csv

- Description: Postal code, latitude and longitude columns are selected from the Geospatial_Coordinates csv file.

```
In [4]: neigh_df = pd.DataFrame(df_list[0].Neighbourhood.str.split(',').to_list(), index=df_list[0]['Postal Code']).stack()
neigh_df = neigh_df.reset_index([0, 'Postal Code'])
neigh_df.columns = ['Postal Code', 'Neighbourhood']
neigh_df = pd.merge(neigh_df,df_list[0][['Postal Code','Borough']], on=['Postal Code'])
df_geo = pd.read_csv("https://cocl.us/Geospatial_data/Geospatial_Coordinates.csv")
df_Pos_Geo = pd.merge(neigh_df,
                      df_geo[['Postal Code', 'Latitude', 'Longitude']],
                      on='Postal Code')
df_Pos_Geo['Neighbourhood'] = df_Pos_Geo['Neighbourhood'].str.strip()
df_Pos_Geo.head()
```

Out[4]:

	Postal Code	Neighbourhood	Borough	Latitude	Longitude
0	M3A	Parkwoods	North York	43.753259	-79.329656
1	M4A	Victoria Village	North York	43.725882	-79.315572
2	M5A	Regent Park	Downtown Toronto	43.654260	-79.360636
3	M5A	Harbourfront	Downtown Toronto	43.654260	-79.360636
4	M6A	Lawrence Manor	North York	43.718518	-79.464763

Fig 2: Postal code wise Geospatial coordinates for the city of Toronto

- Neighborhood wise Population distribution:
  - Data Source: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/CompFile.cfm?Lang=Eng&T=1201&OFT=FULLCSV
  - Description: The list contains the population (as on 2016) for each neighborhood for all the province of Canada. We have the geographic code, geographic name, province / territory and population (2016). Geographic code and geographic name contain the postal code.

```
url_pop = 'https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/CompFile.cfm?Lang=Eng&T=1201&OFT=FULLCS

# read into dataframe
pop = pd.read_csv(url_pop)


# drop irrelevant columns and make sure we are looking at Ontario
pop = pop[['Geographic code','Geographic name','Province or territory','Population, 2016']]
pop=pop[(pop['Province or territory']== "Ontario")]


# change population column type to float
pop['Population, 2016'].astype(float)

pop.head()
```

Out[5]:

	Geographic code	Geographic name	Province or territory	Population, 2016
650	K0A	K0A	Ontario	103474.0
651	K0B	K0B	Ontario	20945.0
652	K0C	K0C	Ontario	52154.0
653	K0E	K0E	Ontario	38903.0
654	K0G	K0G	Ontario	37097.0

Fig 3: Neighborhood wise population with geo spatial coordinates

- The top 20 ethnic origins population in Toronto for the years 2011 and 2016.
  - Data Source: https://en.wikipedia.org/wiki/Demographics_of_Toronto
  - Description: The data set is extracted from the above source which lists the top 20 ethnic groups and their percentage of their contribution to the total population of Toronto for the years 2011 and 2016. This gives the trend of change in the ethnic group population over years

```
dt_Ethn.set_index('Ethnicity', inplace=True)
dt_Ethn.head()
```

Out[8]:

Ethnicity	2016	2011
Canadian	12.7	13.2
English	12.5	14.1
Chinese	12.0	10.8
East Indian	11.0	10.4
Irish	9.3	9.8

Fig 4: Top 20 ethnic origin population in Toronto for the years 2011 and 2016

- List of Restaurants in and around Toronto neighborhood:
    - Data Source: Foursquare API
    - Description: The search endpoint of the foursquare API returns the list of the venues around the specified geospatial coordinate. It also provides the details like venue name, venue category, venue coordinates (latitude /longitude).

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

Out[17]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
1	Victoria Village	43.725882	-79.315572	Pizza Nova	43.725824	-79.312860	Pizza Place
2	Victoria Village	43.725882	-79.315572	The Frig	43.727051	-79.317418	French Restaurant
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Wingburger	43.725580	-79.312851	Burger Joint

Fig 5: List of restaurants around Toronto Neighborhood

- List of neighborhoods of Toronto, their average salary, second top language spoken and their population.
    - Data Source: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
    - Description: The dataset is obtained from the above source using the web scraping technique using the BeautifulSoup library. The Dataset has borough wise neighborhood data, their population, average salary, second top spoken language other than English. Since the language has strong relationship with the ethnicity, this information gives the largest ethnic group population in each neighborhood.

Out[39]:

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
0	Crescent Town	EY	0190.01	8,157	0.4	20,393	-10.0	23,021	24.5	20.3	Bengali (18.1%)	18.1% Bengali	
1	Governor's Bridge/Bennington Heights	EY	0186.00	2,112	1.87	1129	4.0	129,904	7.1	13.3	Polish (1.4%)	01.4% Polish	
2	Leaside	EY	0195.00, 0196.00	13,876	2.81	4938	3.0	82,670	9.7	10.5	Bulgarian (0.4%)	00.4% Bulgarian	
3	O'Connor–Parkview	EY	0189.00, 0190.02, 0191.00, 0192.00, 0193.00	17,740	4.94	3591	-6.1	33,517	15.8	19.4	Urdu (3.2%)	03.2% Urdu	
4	Old East York	EY	0180.00, 0181.01, 0181.02, 0182.00, 0183.00, 0...	52,220	7.94	6577	-4.6	33,172	22.0	19.1	Greek (4.3%)	04.3% Greek	

Fig 6: List of neighborhoods of Toronto, their average salary, second top language spoken and their population

Data preprocessing:

The information obtained from the above various data sources provide all the information which is required or not required, which can be directly used or to be processed in order to use them. Hence some processing is done on the data frames to convert them to be the usable data.

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson

Fig 7: List of Neighborhood and their postal code

In the above dataset we observe that lot of the postal code doesn't have Boroughs assigned. So, the Boroughs with "Not assigned" value are removed and only the postal codes with valid Boroughs, neighborhood information is taken for consideration as shown below.

```
In [3]: url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
        df_list = pd.read_html(url)
        df_list[0].drop(df_list[0][df_list[0].Borough == 'Not assigned'].index, inplace =True)
        print(df_list[0].shape)

        (103, 3)
```

```
In [4]: df_list[0].head()
```

Out[4]:

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Fig 8: Neighborhood data after cleansing

Since the postal code is considered as the index field rather than the neighborhood name, the neighborhood information data frame and the population information data frame are merged together for all the future references and analysis.

```
In [6]: df_Pos_Geo = pd.merge(df_Pos_Geo,pop[['Geographic code','Population, 2016']], left_on=['Postal Code'], right_on =['Geographic coc
        df_Pos_Geo=df_Pos_Geo[['Postal Code','Neighbourhood','Borough','Latitude','Longitude','Population, 2016']]
        df_Pos_Geo.head()
```

Out[6]:

	Postal Code	Neighbourhood	Borough	Latitude	Longitude	Population, 2016
0	M3A	Parkwoods	North York	43.753259	-79.329656	34615.0
1	M4A	Victoria Village	North York	43.725882	-79.315572	14443.0
2	M5A	Regent Park	Downtown Toronto	43.654260	-79.360636	41078.0
3	M5A	Harbourfront	Downtown Toronto	43.654260	-79.360636	41078.0
4	M6A	Lawrence Manor	North York	43.718518	-79.464763	21048.0

Fig 9: Population information after merging with neighborhood data

The neighborhood average salary, second top spoken language and the percent of people speaking the language are all provided in columns in the readable format. The same has to be processed and segregated to individual columns for analyzing those parameters to arrive at a decision.

```
In [52]: tor_neighbouhood_data.head()
```

Out[52]:

	Postal Code	Neighbourhood	Borough	Latitude	Longitude	Population, 2016	Average Income	Percent	Language	lan_pop
0	M3A	Parkwoods	North York	43.753259	-79.329656	34615.0	34,811	3.4	Chinese	1176.910
1	M4A	Victoria Village	North York	43.725882	-79.315572	14443.0	29,657	3.2	Urdu	462.176
2	M5A	Regent Park	Downtown Toronto	43.654260	-79.360636	41078.0	19,521	10.5	Bengali	4313.190
3	M5A	Harbourfront	Downtown Toronto	43.654260	-79.360636	41078.0	69,232	2.4	Chinese	985.872
4	M6A	Lawrence Manor	North York	43.718518	-79.464763	21048.0	36,361	7.2	Filipino	1515.456

Fig 10: Neighborhood Data after segregation

3. Methodology and Exploratory Data Analysis:

Now, having collected the various data which will directly or indirectly influence the restaurant business, we have to explore the data and apply the systematic approach/methodologies to arrive at the solution for the problem statement i.e. where to open the restaurant, what will be best cuisine or where we can find our preferred choice of restaurant and what is the most preferred restaurant in the neighborhood.

Population:

As for as any business concerned, the target will be the customers. The service-oriented business sectors like food and recreation also targets the consumers. We get the overall population status of each neighborhood in Toronto which gives the scope for starting restaurant business. We have retrieved the population statics neighborhood wise which can be visualized in the below figure

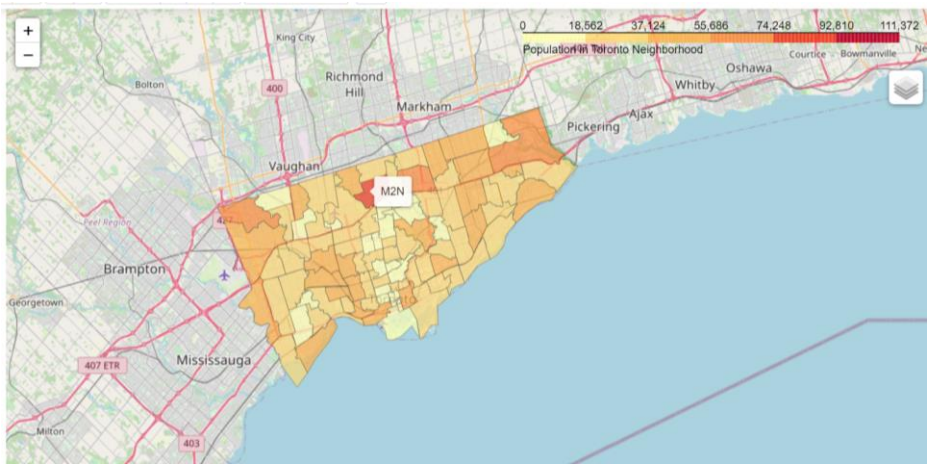

Fig 11: choropleth showing Toronto Neighborhood's Population

The figure above is the choropleth map which clearly gives us the population distribution across the neighborhoods of Toronto. The densely populated area has the population between 92,810 to 111,372 and the least populated has the population between 0 to 18,562. We observe that most of the neighborhoods are with average population ranging between 37,124 to 55,686.

Restaurants in Neighborhood/ Competitors:

Having got the idea of the consumers or customer population in the neighborhood, the next data of interest for analysis is to know the restaurants in the neighborhood. This helps us to identify the potential competitors for the restaurant business. This information is obtained

by from Foursquare API, which gives the list of food venues and their categories in the neighborhood of Toronto. The below visualization gives the distribution of the restaurants.

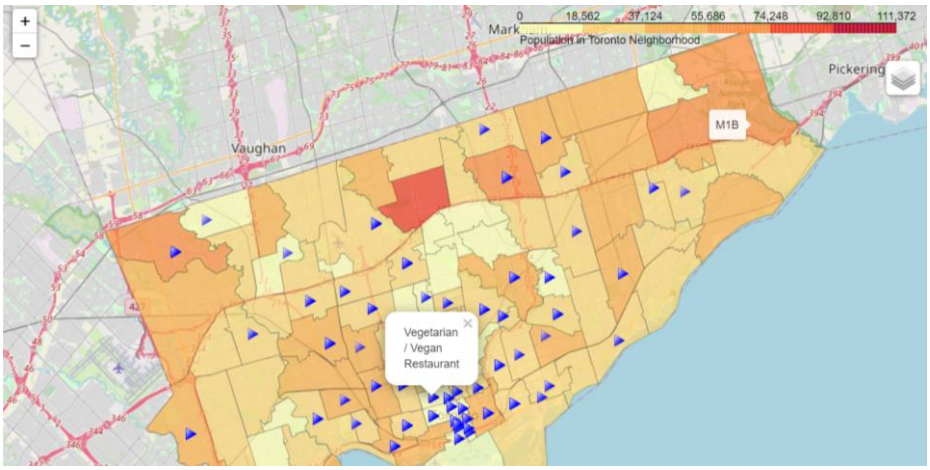

Fig 12: List of Restaurants in Toronto Neighborhood plotted on the Choropleth

The figure above gives the details of the restaurants spread across the neighborhood in and around 200 mts of each neighborhood which is retrieved from Foursquare API. This visual representation gives us further insight of which are all the neighborhoods flooded with restaurants and the neighborhoods which are densely populated but with minimal restaurants. E.g. The densely populated Willowdale neighborhood and other averagely populated neighborhoods like Rouge, West Humber – Clairville etc doesn't seem to have the restaurants, which will be the area or locations of focus for setting the restaurant.

Ethnic Population:

Since Toronto is an international centre for business and has the diverse ethnic population across the neighborhood, it will be a good idea for the restaurateur to target for serving the specific ethnic group. The percentage of the ethnic group population for the years 2011 and 2016 is plotted in the below bar graph.

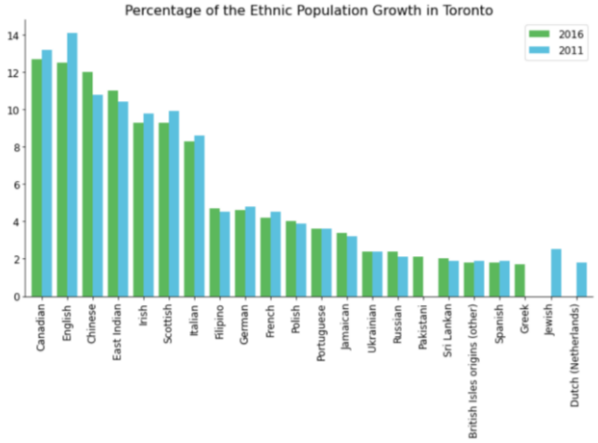

Fig 13: Change in the ethnic population for the years 2016 and 2011

From the above bar graph which compares the change in the ethnic group wise population for the years 2011 and 2016, we can infer the following information

We see the increase in the population of the following ethnic groups like Chinese (1.2%) , East Indian (0.6%) ,Pakistani (0.5%) ,Russian(0.3%), Sri Lankan (0.1%), Filipino (0.2%).

Also, we have retrieved the second largest spoken language in each neighborhood of Toronto and their population and plotted top 10 highest language wise populated neighborhoods.

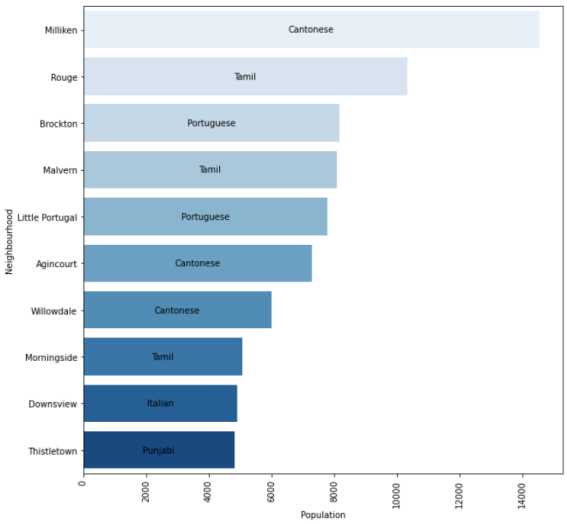

Fig 14: Top 10 Language based population in Toronto's neighborhood

The neighborhoods like Milliken, Agincourt and Willowdale are populated with Cantonese speaking people who are basically Chinese and Rouge, Malvern and Morningside are populated with Tamil people and Thistletown with Punjabi people. This majorly constitutes the East Indian population.

Average Income:

Apart from the above parameters like Population and competitors, there are many other parameters which may have impact on the restaurant business. Demographic data of the Toronto neighborhood also gives the average income of the people in each neighborhood. The spending capability will be a driving parameter for deciding the location for opening the restaurant.

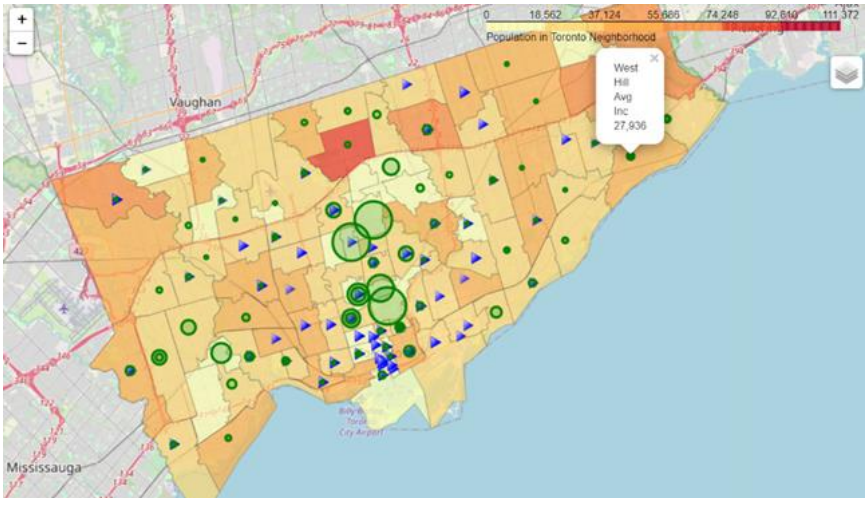

Fig 15: Plot of the average income of neighborhoods in Toronto

We have plotted the average income of each neighborhood in green circles and size of the circle shows the magnitude of the Avg. income based. From the plot we can easily come to know the average income of the neighborhoods without restaurant. E.g. West Hill doesn't have a restaurant but have an average population and also the average income of 27,936.

Clustering:

Based on the various analysis, the Chinese and East Indian ethnic groups are majorly available across the various neighborhoods. So, we cluster the Indian and Chinese restaurant venues based on their location, category and identify the relationship between them. I used the KMeans clustering to cluster the venues.

One of the trickier tasks in clustering is identifying the appropriate number of clusters k. We will use elbow method as a way to estimate the value k. For each k value, we will initialise k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster centre.

Below is a plot of sum of squared distances for k in the range specified above. If the plot looks like an arm, then the elbow on the arm is optimal k.

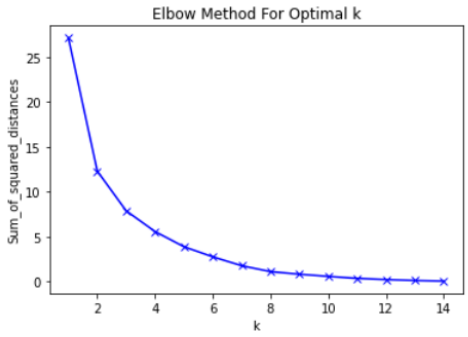

Fig 16: Elbow Plot of sum of the squared distance for optimal k

In the plot above the elbow is at k=6 indicating the optimal k for this dataset is 6. Having identified the k=6, we cluster the Indian and Chinese restaurants based on the category and location into 6 clusters and plotted the cluster.

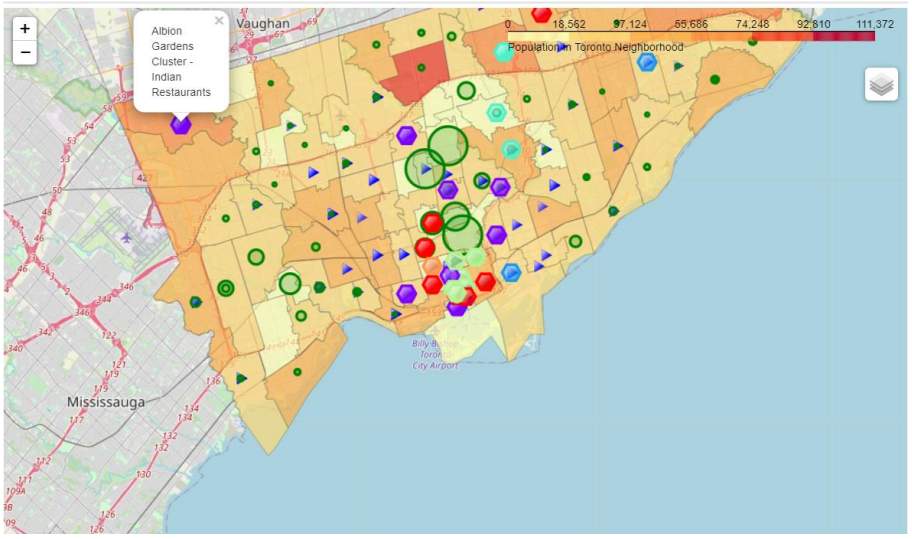

Fig 17: Plot of clusters of the Indian Chinese Restaurants

In the figure, we have 6 clusters plotted as the hexagon marker.

Cluster 1: Red colour hexagon showing the clusters with a greater number of Chinese restaurants.
Cluster 2: Purple colour hexagon showing the clusters with a greater number of Indian restaurants.
Cluster 3: Light Green colour hexagon showing the cluster with 65% of Indian restaurants and 35% of Chinese restaurants.
Cluster 4: Light Blue colour hexagon showing the cluster with a greater number of Hakka restaurants.
Cluster 5: Cyan colour hexagon showing the cluster with a greater number of Dim Sum restaurants.
Cluster 6: Cyan colour hexagon showing the cluster with a greater number of food trucks.

4. Results & Discussion:

The following are the observations after analysing the data.

1. The most popular ethnic groups in the neighbourhoods in Toronto are Chinese, East Indian, Pakistani and Sri Lankan. Chinese and Indian Restaurants will be the ideal choice of cuisine.
2. The most populous neighborhood in Toronto is Willowdale which does not have any restaurants and it has the average income of 39,895. Also, it has more Cantonese speaking population who belong to Chinese ethnic group. So, it will be the ideal location for opening the Chinese restaurant.
3. The other neighborhoods for opening the Chinese restaurants are Agincourt and Milliken with more Cantonese speaking population and average income of 25,750 and 25,243 respectively.
4. South Indian cuisine restaurant can be opened in the neighborhoods like Rouge and Morning side, which have more Tamil speaking population with the average income of 29,230 and 27,936 respectively.

Apart from the above observations, we have the cluster 5 which has a greater number of Dim Sum restaurants located mainly around Don Mills, Oriole/Henry farm neighborhood. Those neighborhoods have more Chinese and Mandarin speaking population and Restaurateur can open a Chinese restaurant. In the Oakridge and Scarborough Village West neighborhoods has Bengali and Tamil speaking population respectively and can be considered for opening the Indian restaurant which can server both North and South Indian dishes. If the Restaurateur plans to open the Chinese cuisine restaurant, he/she should avoid the neighborhoods in the cluster 1 and cluster 3 which already have a greater number of the Chinese restaurants and competitors.

5. Suggestions for future research:

We have considered the population, trend in the change of ethnic group population and average income of the neighborhoods for analyzing and decide for the ideal location for opening the restaurateur. Other parameters also impact the business like the Tourist attractions in the neighborhood which attracts the domestic and foreign visitors, Age wise population which also have major impact on opening the chain of restaurants, Ratings of the existing restaurants to know the popularity of each restaurant which in turn helps to know the potential competitors.

Also, we have to include the Business centers, work premises which also attract many customers towards the restaurants. All the above stated parameters may help in further precise analysis.

6. References:

Foursquare API Endpoint references

https://developer.foursquare.com/docs/

Neighborhood geo spatial coordinates of Toronto

https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm

Neighborhood wise postal codes of Toronto

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Neighborhood wise population of Toronto

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/CompFile.cfm?Lang=Eng&T=1201&OFT=FULLCSV

Demographics of Toronto – change in population of ethnic group

https://en.wikipedia.org/wiki/Demographics_of_Toronto

Toronto's Neighborhood wise Average income and second highest spoken language and population

https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods