



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

