Capstone Final Project Report

IBM Data Science Professional Certificate

Alejandro Plaza Larrea

Opening a Latino Restaurant in Toronto

In this final project of the capstone of the IBM Data Science Professional Program, I will define a business problem and try to solve it with the tools presented in the previously mentioned program.

I will explain the process step-by-step, and also provide all the necessary background.

1. Introduction/Business Problem

Toronto is the most populous city in Canada and is home to a variety of nationalities. People from other parts love to have their country food, and also locals can enjoy different tastes.

This project will be about figuring out if opening a Latino Restaurant is a good idea, and if it is, where to open it. We will analyze different neighborhoods, the demographics and the current offer to better match-up our proposal. The stakeholders of this project will be:

- *Businessman/Chefs interested in opening/investing restaurants.* The project will be useful to them to identify key areas for the business.
- *Latino community*, who will be grateful to identify areas where they can find food that matches their taste.
- *City guides*, they will gain insight into the different attributes areas have and how they can improve their service to match the demands in different areas.

2. Data

In this section, we will see which data we will need to develop our model. I will describe the use of it and where to find it.

To achieve our objective, we will use different datasets:

- ***List of Postal Code of Canada*** (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M): This is a dataset that lists the different neighborhoods in Toronto, having the Postcode, Borough and the Neighborhood as attributes. For example, MP4 with Borough York and Neighborhood Weston.

- ***Toronto Geospatial Data*** (https://cocl.us/Geospatial_data): This dataset list all the postal codes with latitude and longitude as attributes. For example, MP4 with Longitude -79 and latitude 32.

- ***Demographics of Toronto*** (https://en.wikipedia.org/wiki/Demographics_of_Toronto): We will use this Wikipedia page to have insight into where Latinos live. The neighborhood Latino racial density.

- ***Bicycle Parking Racks Data (***Toronto Open Data): We will use this database to know how many bicycle parking racks are in the Postal code because of the food delivery trends.

- ***Population (***Toronto Open Data): Since we don't have the total number of Latinos per postal code, we will have to calculate it using the population of each neighborhood. Also, it can give us insight of the demographics of each area.

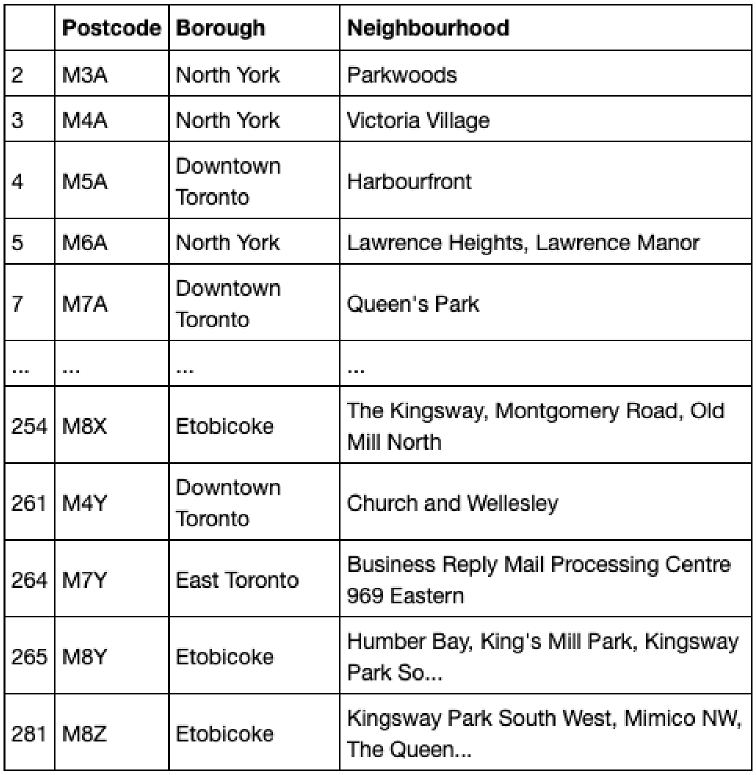
- ***Foursquare API***: We will use this API to extract more information about neighborhoods like restaurants (current offer). For example, the names of Latino restaurants with their latitude and longitude.

3. Methodology

In the following section we will explore how to extract the data from the source, clean it and use it for our purpose.

3.1. Web-scraping Postal Codes

As said before, we will use the Wikipedia page to scrap the data of the postal codes, as recommended we use the Beautiful Soup library to do it.

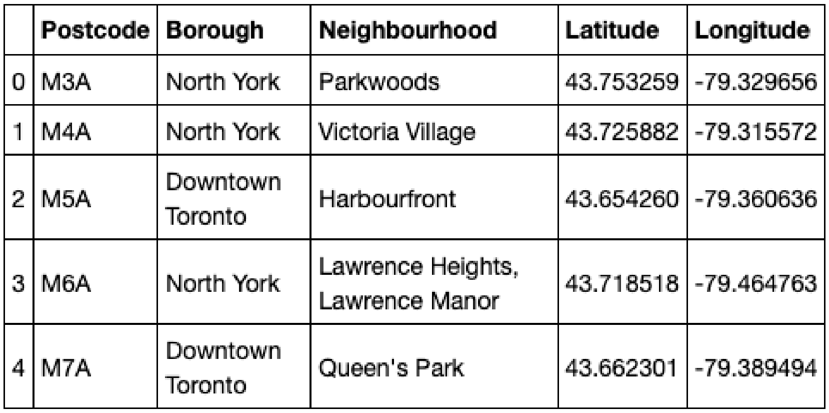


	Postcode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights, Lawrence Manor
7	M7A	Downtown Toronto	Queen's Park
...	...	...	...
254	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
261	M4Y	Downtown Toronto	Church and Wellesley
264	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern
265	M8Y	Etobicoke	Humber Bay, King's Mill Park, Kingsway Park So...
281	M8Z	Etobicoke	Kingsway Park South West, Mimico NW, The Queen...

In the figure above, we have the result.

3.2. Geo-Spatial Data

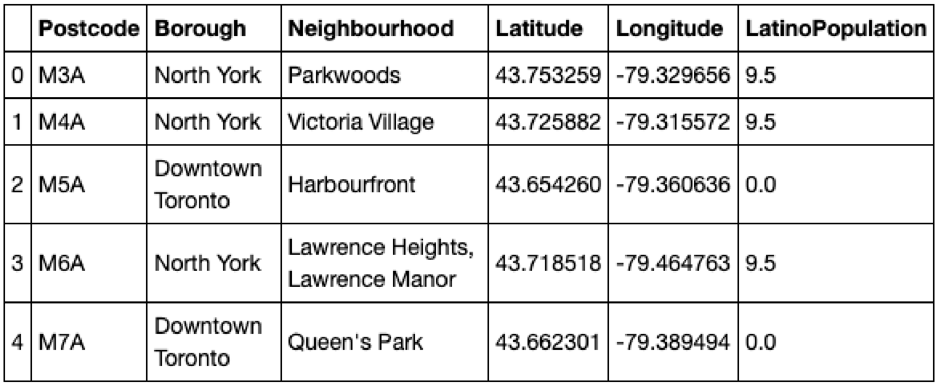
Now we had to read a csv file to have the postal code with the latitude and longitude for each of them. Merging this dataset with the later one we got:



	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

3.3. Web-scrapping Toronto Demographics Data

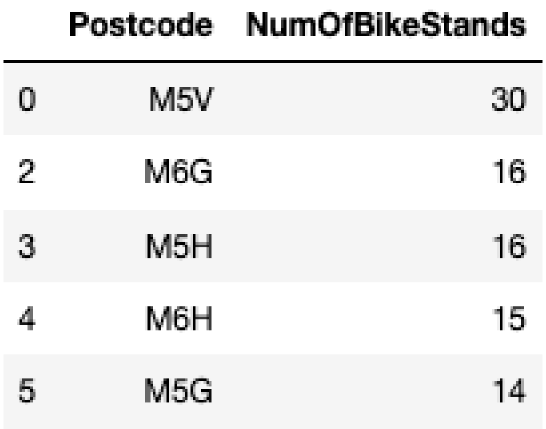
As before, we are scrapped the Wikipedia page to have the total Latino population by postal code. It is important to say that for the ones that weren't in the dataset we gave the value of 0, also that the population was given on percentage.



	Postcode	Borough	Neighbourhood	Latitude	Longitude	LatinoPopulation
0	M3A	North York	Parkwoods	43.753259	-79.329656	9.5
1	M4A	North York	Victoria Village	43.725882	-79.315572	9.5
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	0.0
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	9.5
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	0.0

3.4. Bicycle Parking Racks Data

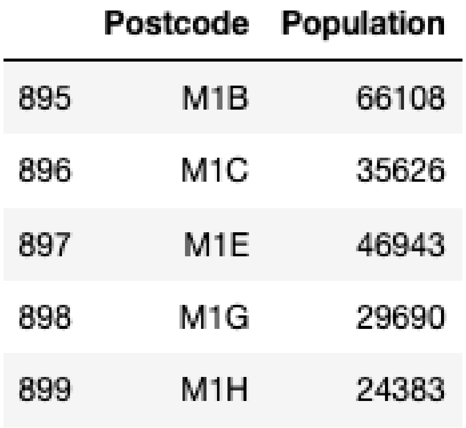
We took a dataset from the Toronto Open data to know how many bike stands per postal code we got, there were only the postal code, coordinates and capacity of each of them so we only had to count the amount of them by postal code. We got the following table, that we then merge it with the one above.



	Postcode	NumOfBikeStands
0	M5V	30
2	M6G	16
3	M5H	16
4	M6H	15
5	M5G	14

3.5. Population Data

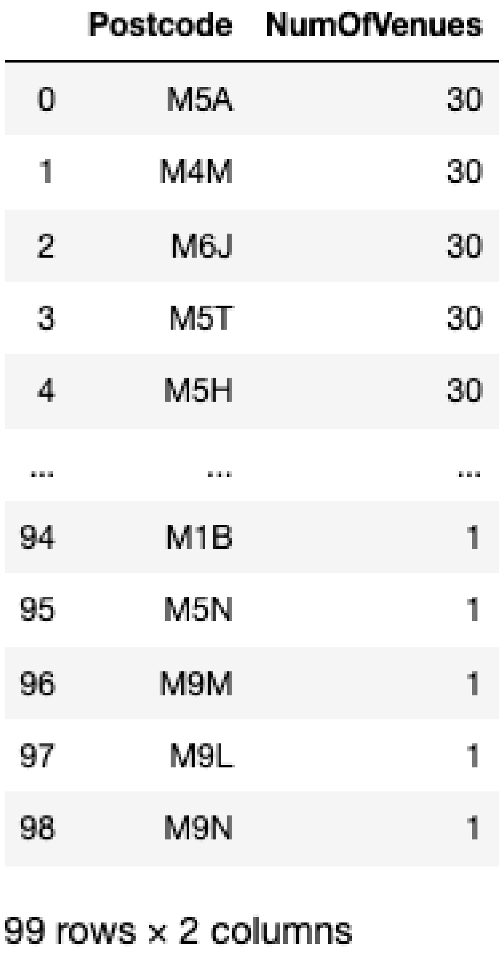
As before, we use the population data to then merge it with the big data frame. We got from a data set containing the postcode with the population of it.



	Postcode	Population
895	M1B	66108
896	M1C	35626
897	M1E	46943
898	M1G	29690
899	M1H	24383

3.6. Foursquare Data

Now we will use the Foursquare Data to know the current offer places around the neighborhoods. After this step, we will begin clustering the different zones.



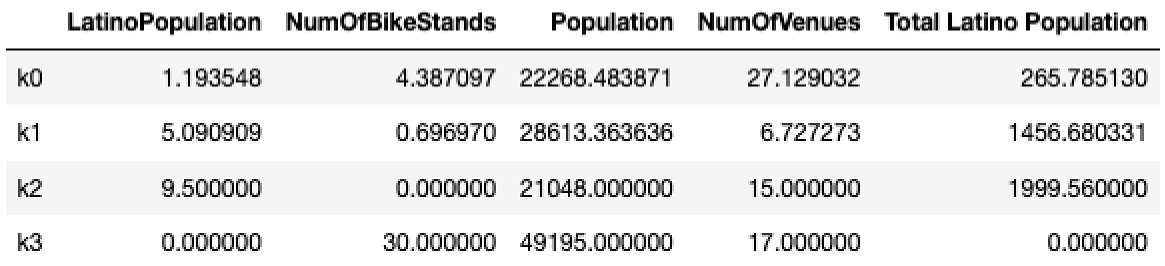
	Postcode	NumOfVenues
0	M5A	30
1	M4M	30
2	M6J	30
3	M5T	30
4	M5H	30
...	...	...
94	M1B	1
95	M5N	1
96	M9M	1
97	M9L	1
98	M9N	1

99 rows × 2 columns

Also, we made another data frame containing the 1st (to 5th) most common venue category.

4. Clustering

We used the k-means clustering method to categorize the different postal codes, and then analyze them by group. Finally we ended up with 4 different clusters, where the differences between them where quite impressing.



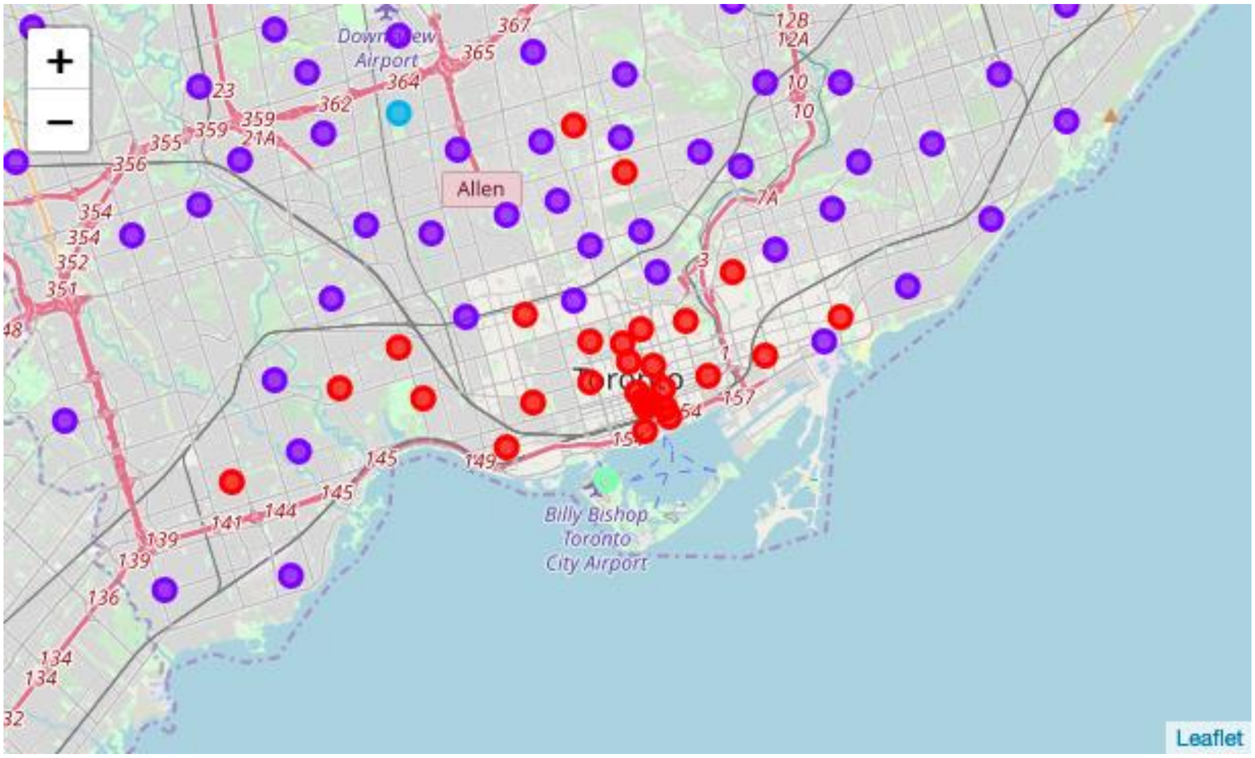
	LatinoPopulation	NumOfBikeStands	Population	NumOfVenues	Total Latino Population
k0	1.193548	4.387097	22268.483871	27.129032	265.785130
k1	5.090909	0.696970	28613.363636	6.727273	1456.680331
k2	9.500000	0.000000	21048.000000	15.000000	1999.560000
k3	0.000000	30.000000	49195.000000	17.000000	0.000000

5. Discussion

We can choose between two different postal code, one is the k0 which is the M5A and the k2 which is the M6A.

The first one has a strong offer of venues and more bike stands, but the Latino population is low. In the other hand, we have half the offer of venues and no bike stands but almost 2000 Latinos living there.

Comparably, the most common venue category of the first postal code is coffee places meanwhile in the only neighborhood of the second option the most common venue is Clothing Store (Coffee Shop is the third most common). It seems that the first option has more offer of restaurants than the second one.


+
−
Downsview Airport
Allen
Toronto
Billy Bishop Toronto City Airport
Leaflet