Where should I open my new Mexican restaurant?

Adán Abuín López

September 15, 2020

1. Introduction

1.1 Background

There are so many factors that helps a new restaurant to succeed. The quality of food, the demand, the brand fame… But one of the most important ones is the emplacement, so the entepreneurs must do whatever it's on their hands to find the perfect place to stablish their enterprise, in this case, their restaurant.

1.2 Problem

The client, proprietary of a known franchise of Mexican restaurants, wants to expand his bussiness to Canada and decided to begin his expansion project in Toronto. He's never lived in the city so he needs help to analyze which is the most appropiate neighbourhood to establish his first restaurant.

I will search for information online about the demographic and economic profile of the population in each neighbourhood to be able to make a cluster analysis which will help me to determine the optimal place or places where the restaurant should be sited.

1.3 Interests

The proprietary it's obviously interested in find the best place to be able to generate the biggest profits.

2. Data acquisition and cleaning

2.1 Data Sources

First I scrapped the wikipedia website to obtain the neighbourhood information (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050).

Then, I obtained the coordinates of each neighbourhood's centroid from the website https://cocl.us/Geospatial_data

I will use the Foursquare API to find the number of Mexican restaurants in each neighbourhood and have a view of the level of competition and offer.

Looking for information online I found in the "Toronto's Open Data Portal" many demographic and economic metrics of the Toronto neighbourhoods. I extracted from it information about the total population, the population from latin american countries and the income in each neighbourhood. Here's the link: https://bit.ly/3airrOJ

2.2 Data Cleaning

After downloading all the data from the previous sources, I combined all into one single table.

There were a lot of missing values for certain neighbourhoods, so I only processed cells with an asigned borough.

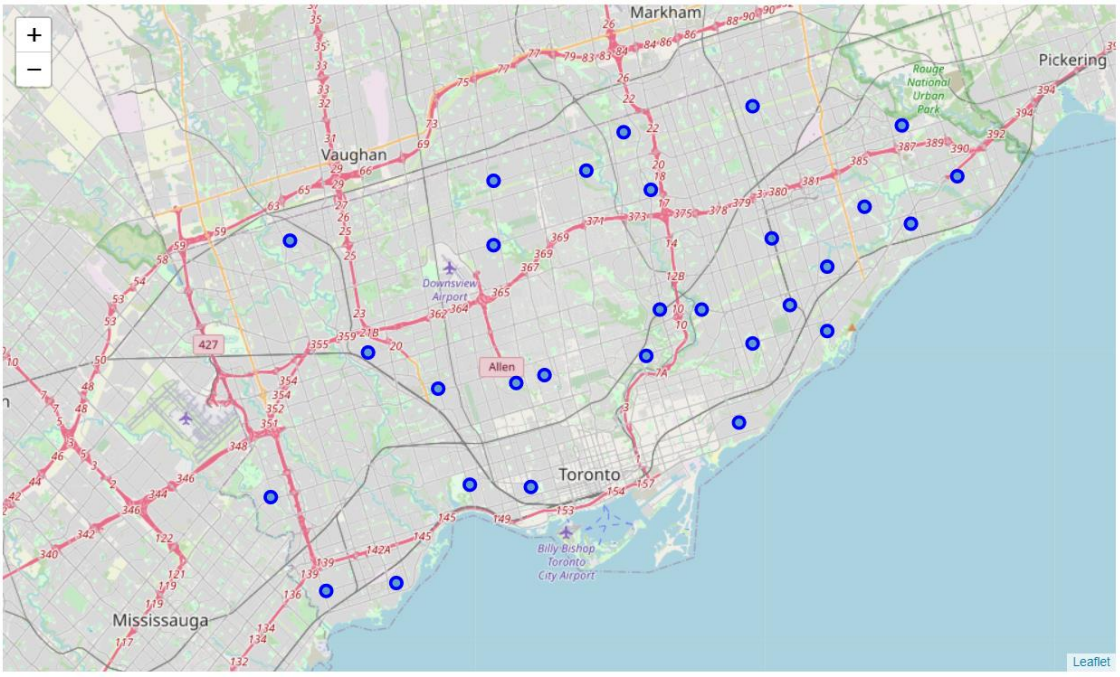
The last column indicates the percentage of Latin american people in each postcode. I obtained that dividing the number of Latin american population by the total population.

	Postcode	Borough	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Total Population	Pct Latin American
0	M1B	[Scarborough, Scarborough]	[Rouge, Malvern]	43.806686	-79.194353	126209.0	90290.0	1.401041
1	M1C	[Scarborough]	[Highland Creek]	43.784535	-79.160497	87321.0	12494.0	1.640788
2	M1E	[Scarborough, Scarborough, Scarborough]	[Guildwood, Morningside, West Hill]	43.763573	-79.188711	164550.0	54764.0	1.862537
3	M1G	[Scarborough]	[Woburn]	43.770992	-79.216917	47908.0	53485.0	1.392914
4	M1J	[Scarborough]	[Scarborough Village]	43.744734	-79.239476	40181.0	16724.0	1.674241

3. Exploratory Data analysis

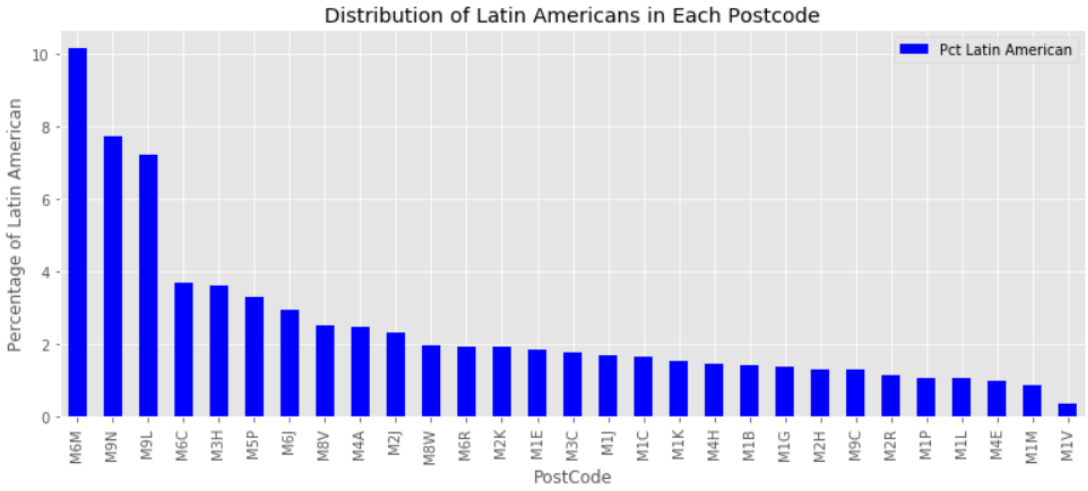
3.1 Mapping

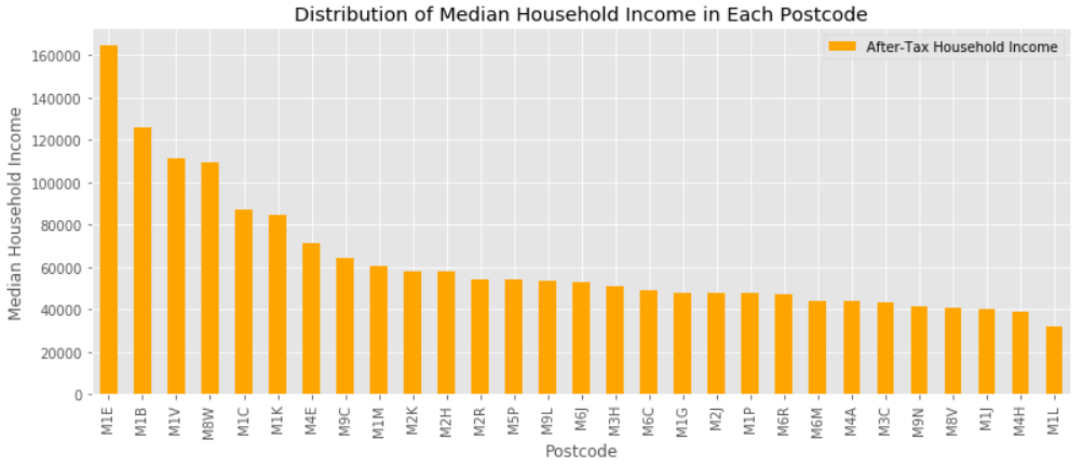
Now using the Folium library we'll represent the points in the centroids of the neighborhoods of Toronto that we'll use for the analysis.



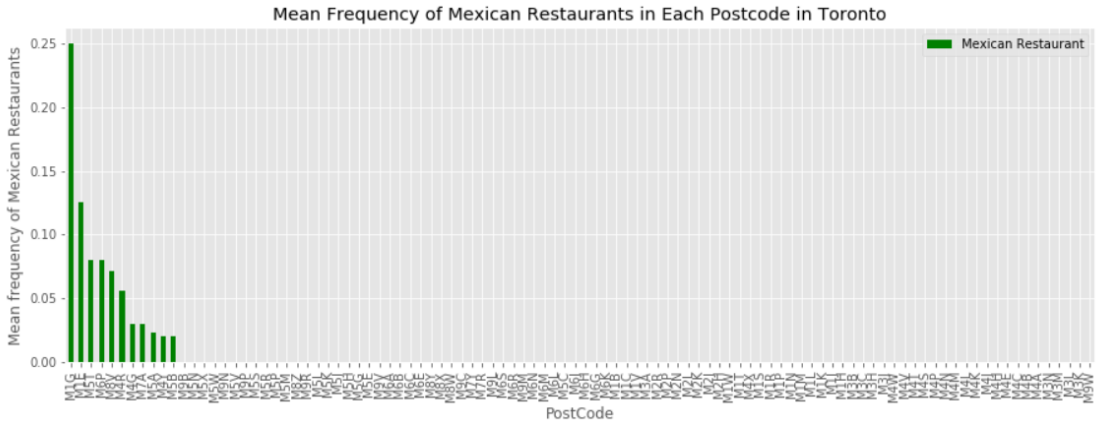
3.2 Distributions

We can build some bar charts to see how the values we collected are distributed among the different postcodes.





Now, to represent the distribution of the competition in each neighbourhood, we'll use the Foursquare API. That will tell us how many mexican restaurants exist in each post code. Once we've made the API call, we can represent the competition distribution:



4. Predictive Modelling

4.1 Normalize data

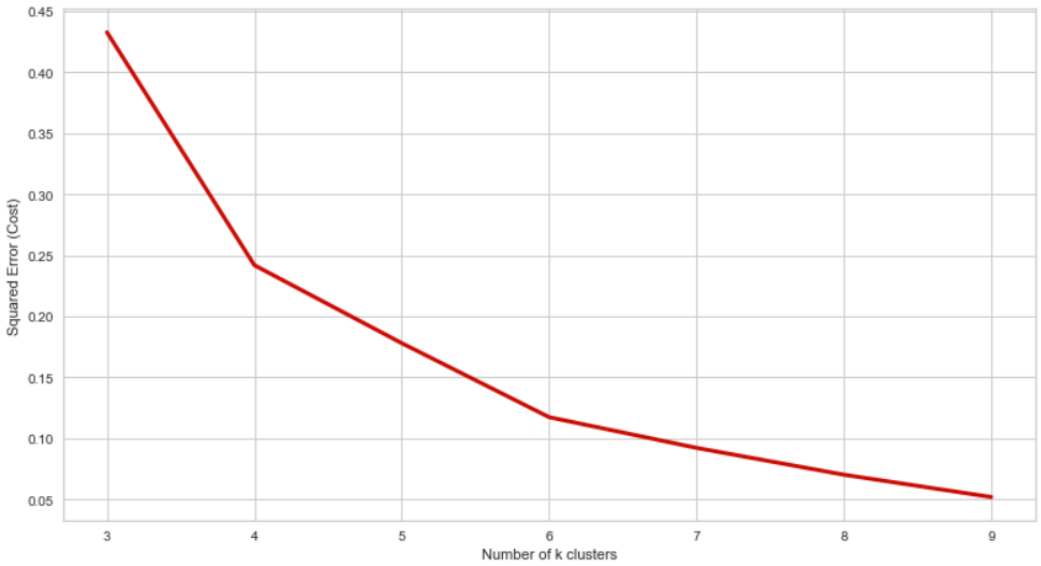
To make our cluster, we must first normalize the data, due to the diferent scale of the metrics (for example , population is much bigger tan percentage of latin american people).

The result of the normalization is shown below:

	Household Income	% Total Population	Pct Latin American	No. of Mexican Restaurants
0	2.125243	3.807544	-0.504272	-0.300010
1	0.820451	-0.632766	-0.394281	-0.300010
2	3.411683	1.779850	-0.292546	2.136068
3	-0.501957	1.706850	-0.508001	4.572146
4	-0.761218	-0.391333	-0.378933	-0.300010

4.2 K-means Clustering

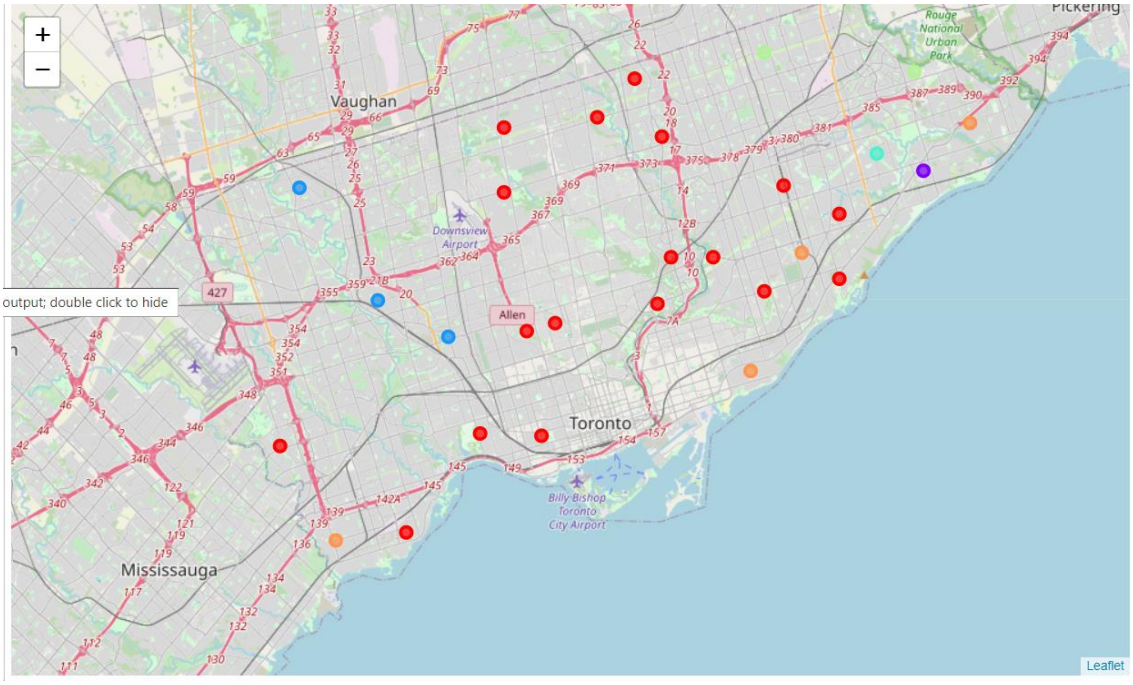
The first thing we must do when we want to make a K-means clustering is to find the optimal number of clusters for our analysis. If we represent the k against the minimal squared error we obtain:



From the graphic we can see that the elbow is sited in k=6, so we'll use 6 different clusters.

4.3 Final Clusters

After running the clustering, we can represent graphically the points. The color represents the cluster which they belong to.



Cluster 0: In this cluster we've got many neighbourhoods. We can see that that they're low populated neighbourhoods, with a medium percentage of Latin Americans and medium Income. The competition is practically null.

Cluster 1: This cluster is made up of one Postcode with 3 neighbourhoods, they have a high income and they're high populated. We can see there are some competitors in the area.

Cluster 2: This Cluster represents 3 neighbourhoods with a high percentage of latin american people, medium income and no competition.

Cluster 3: This cluster is made up exclusively of one neighbourhood. It's much more populated than the average. The income and the percentage of latin american people is medium. Also there's some competitors in the area.

Cluster 4: This cluster it's also high populated, but it has a superior income and there's no competition.

Cluster 5: This last cluster has a high income but the total population is low. The percentage of latin american people is medium and there's no competition.

5. Conclusions

Analyzing all the final clusters we've obtained, we'll recommend our client to open his new Mexican restaurant in one of the neighbourhoods of **Cluster 4**.

That's, in the first place, because of the high income level, that allows people in that areas to eat in restaurants more often than in others.

In second place, we can se that the percentage of latin american people is not very high, but the total inhabitants is. So there's a great community of latinos that are more likely to go to the restaurant.

At last, we can also see that there's no competition, so our client will be the only one who'll offer that kind of food to the neighbours of the area.

6. Future directions

In the future we can improve the clustering by adding more extra information about the population of the zone, for example if they’re young people (that usually go to this type of restaurants).

Other metrics relative to the neighborhoods as its size or the amount of people traffic that it supports (if it’s in an strategic place through which a large number of people pass) etc.