Strategic Location for Establishing a Mexican Restaurant in Toronto, Canada

Amin Ali

January 27, 2021

1. Introduction

The success of opening a new restaurant depends on several factors: demand, class, quality of food, competition, and more. In most cases, a restaurant's location plays an essential determinant for its success. Therefore, it is very important to determine the most strategic location for establishment in order to maximize business profits.

1.2 Business Problem

A client seeks to establish a new Mexican restaurant in a Toronto neighborhood. There are not that many Mexican restaurants in the city, which means that the market can support one more establishment in this niche. Which neighbourhood would appear to be the optimal and most strategic location for the business operations?

The objective of this project is to find the most suited neighborhood for opening the restaurant. Our decision would be based on spending power, distribution of ethnic group, and competition across each neighbourhood. For this project we will be using the Foursquare API and the geographical and census data from Toronto's Open Data Portal.

1.3 Interests

The findings of this data science project could be interesting to entrepreneurs who seek to either establish a new restaurant of a certain niche or plan to expand their franchised restaurants and therefore would benefit from newly discovered advantages.

2. Data Acquisition and Cleaning

2.1 Data Sources

The neighbourhoods alongside their respective postal codes and boroughs were scraped from Wikipedia. Geographical coordinates for each neighbourhood were extracted from here. As for Toronto's census data - average household income, total population, and population of Mexicans across each neighbourhood - Toronto's Open Data Portal provides all that data here.
For finding the number of Mexican restaurants in the vicinity of each neighbourhood, we will be utilizing Foursquare API, more specifically, its explore function. One has to register for a Foursquare developer account here to access their API credentials.

2.2 Data Cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values for certain neighbourhoods, due to lack of record keeping. Few assumptions were made to achieve the dataframe shown in Fig 1:

• Only the cells that have an assigned borough will be processed; boroughs that were not assigned were ignored.

• Neighbourhoods missing more than two census data values were dropped.

A column that features the percentage of distribution of Mexicans across each neighbourhood was calculated by dividing the population of Mexicans by the total population of each neighbourhood. So, the two latter columns were made redundant and dropped.

	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Mexicans
0	Victoria Village	43.725882	-79.315572	171271	0.171331
1	Rouge	43.806686	-79.194353	729154	0.107536
2	Malvern	43.806686	-79.194353	533202	0.125588
3	Highland Creek	43.784535	-79.160497	196620	0.160077
4	Flemingdon Park	43.725900	-79.340923	261233	0.273560

Figure 1

3. Methodology and Exploratory Data Analysis

3.1 Folium Mapping

The folium library was called to help visualize, geographically, the location of each neighbourhood centered around Toronto.

Fig 2: Neighbourhoods marked on a map of Toronto

3.2 Frequency Distribution of Mexican Restaurants

Using the Foursquare API's explore function, we could return the number of Mexican restaurants located in each neighbourhood. By calculating the mean respectively, it can give us a better understanding of the frequency of occurrence in each neighbourhood. The argument for the use of frequency of Mexican restaurants is that I hypothesize that there would be a correlation between the number of Mexican restaurants and competition. The higher the number of Mexican restaurants in a neighbourhood, the stronger the competition. The assumption of our analysis is that the barrier of entry to establish a new restaurant in a competitive market is high as existing Mexican restaurants may have the competitive advantage of brand loyalty. Though, counterintuitively, the presence of Mexican restaurants may even be an indicator of demand for Mexican cuisine; the presence of competition may even incentivize innovation to reduce cost and increase productivity. Hence, it would be sound to establish business operations in a neighbourhood that consists of a number of restaurants around the median value.

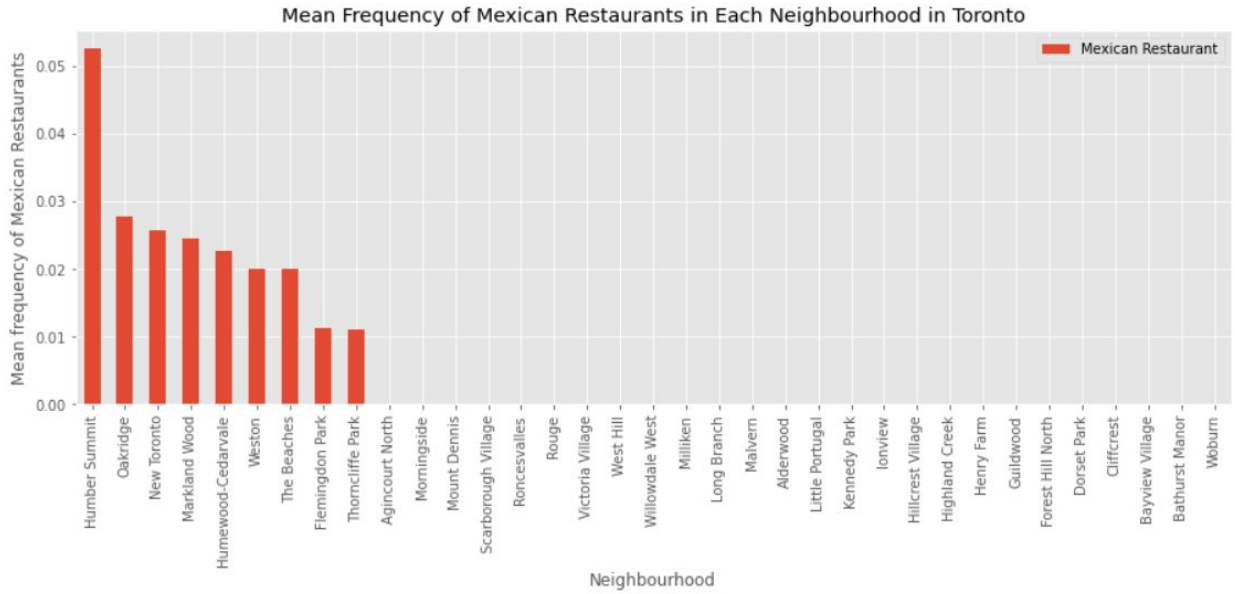

Fig 3: Mean frequency distribution of Mexican restaurants in each neighbourhood

3.3 Distribution of Mexicans

I hypothesize that there would too exist a linear relationship between the population of a specific ethnic group and the demand for its respective cultural cuisine. Hence, it would only be sound for our clients to carry out business operations in neighbourhoods that are relatively more densely populated with Mexicans.

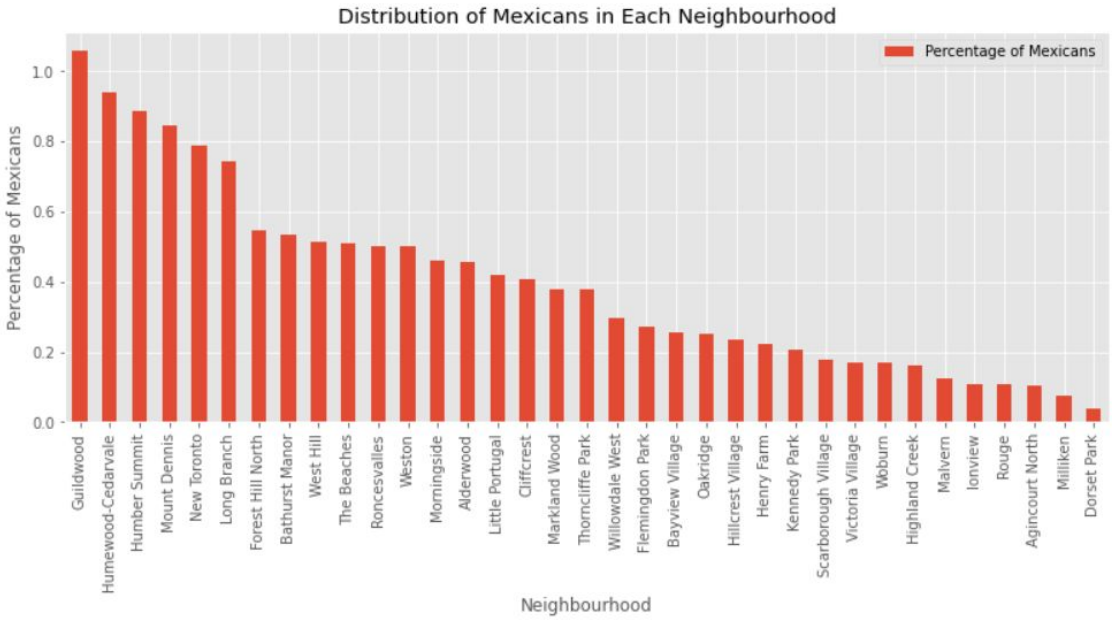

Fig 4: Distribution of Mexicans in each Neighbourhood by percentage

3.4 Distribution of Average Household Income

As the franchised Mexican restaurant could be categorized as casual dining, the target audience is more geared towards the middle class. As can be inferred from the bar chart below, neighbourhoods distributed around the average household income can readily afford and indulge themselves in the aforementioned Mexican cuisine.

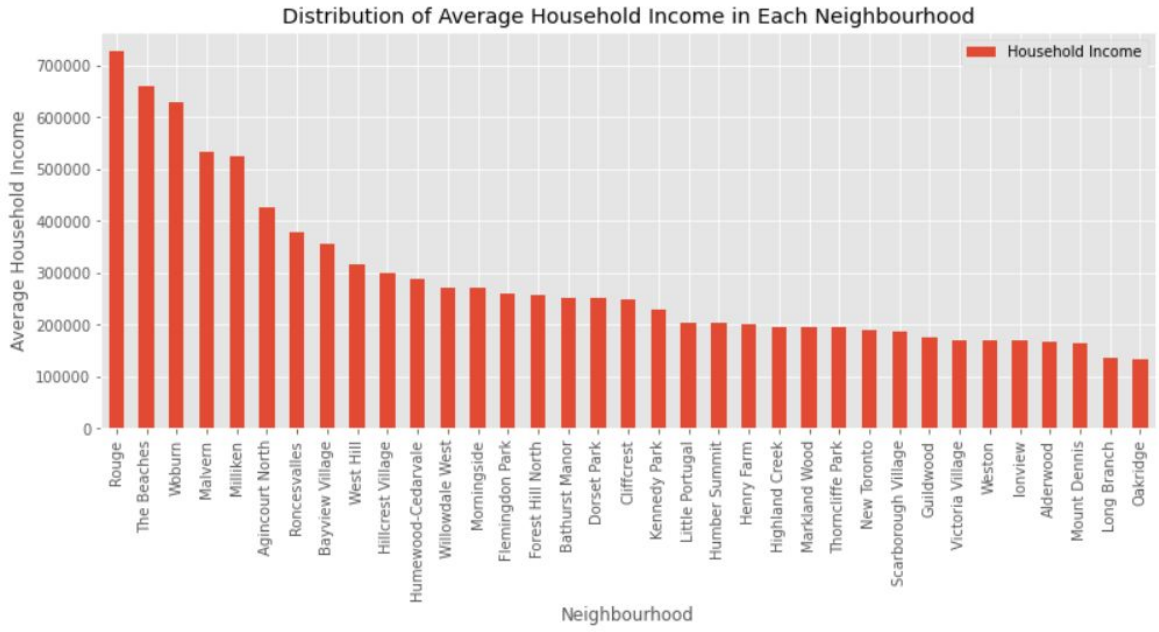

Fig 5: Distribution of Median Household Income in each Neighbourhood

4. Predictive Modeling

4.1 Data Pre-processing

To help mathematical-based algorithms--like our k-Means algorithm in this case---to interpret features with different magnitudes and distributions equally, we will have to normalize our data; as these feature columns are different in scale, we will standardize the values to a common scale. One approach of data normalization is StandardScaler.

	Household Income	% Mexican Population	No. of Mexican Restaurants
0	-0.764233	-0.844961	-0.513484
1	2.914594	-1.085364	-0.513484
2	1.622435	-1.017338	-0.513484
3	-0.597075	-0.887370	-0.513484
4	-0.171000	-0.459719	0.434113

Fig 6: Dataframe of standardised values across all features

4.2 K-Means

Clustering Before we fit the feature values into our model, we have to pre-assign the number of clusters the algorithm should label. To identify the optimal number clusters to use, a range of 3 to 10 clusters were used, then the squared error calculated respectively were used as metrics of their performances.

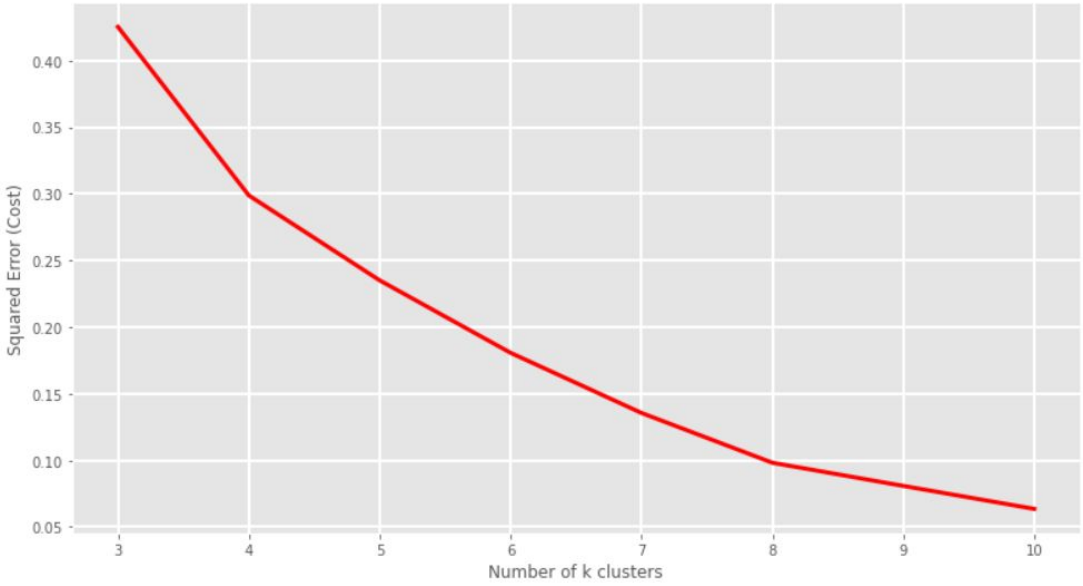

Fig 7: Relationship between the Number of k Clusters and their corresponding Squared Errors

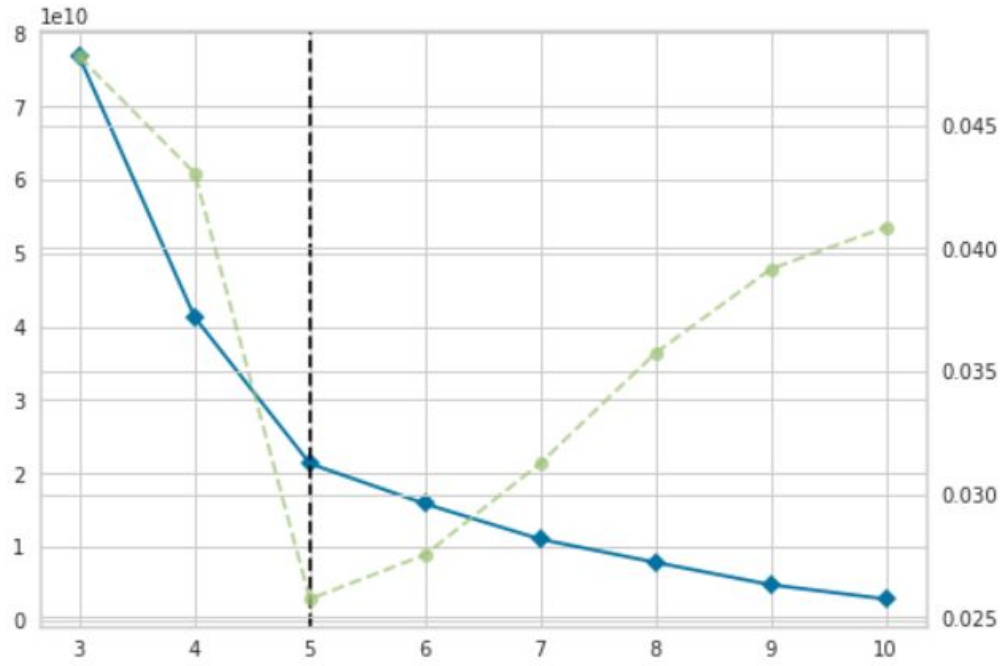

Fig 8: Using K Elbow Visualizer to identify the elbow point for the optimal k value

An analysis using K Elbow Visualizer and Squared error for each k value evident shows that k = 5 would be the best value. After identifying the number of clusters, we will fit the standardized feature values into our k-Means algorithm. The results will be clusters of neighbourhoods of similar characteristics.

4.2.1 Cluster Labels

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Mexicans	No. of Mexican Restaurants
22	0	Humber Summit	43.756303	-79.565963	202677	0.885954	2.786267
32	0	New Toronto	43.605647	-79.501321	188819	0.785135	3.039962

Fig 9.1: Cluster Label 0

Cluster Label 0:
- *MID Spending Power*
- *MID Percentage of Target Customers*
- *HIGH Number of Competitors*

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Mexicans	No. of Mexican Restaurants
0	1	Victoria Village	43.725882	-79.315572	171271	0.171331	-0.511760
3	1	Highland Creek	43.784535	-79.160497	196620	0.160077	-0.511760
4	1	Flemingdon Park	43.725900	-79.340923	261233	0.273560	0.232956
8	1	Morningside	43.763573	-79.188711	272837	0.458321	-0.511760
9	1	West Hill	43.763573	-79.188711	316750	0.511098	-0.511760
12	1	Hillcrest Village	43.803762	-79.363452	298716	0.236211	-0.511760
13	1	Bathurst Manor	43.754328	-79.442259	251583	0.535501	-0.511760
14	1	Thorncliffe Park	43.705369	-79.349372	195114	0.379003	0.217277
15	1	Scarborough Village	43.744734	-79.239476	185967	0.179383	-0.511760
16	1	Henry Farm	43.778517	-79.346556	202357	0.222604	-0.511760
17	1	Little Portugal	43.647927	-79.419750	202912	0.417765	-0.511760
18	1	Ionview	43.727929	-79.262029	169046	0.109963	-0.511760
19	1	Kennedy Park	43.727929	-79.262029	229390	0.204403	-0.511760
20	1	Bayview Village	43.786947	-79.385975	354894	0.257057	-0.511760
23	1	Cliffcrest	43.716316	-79.239476	248846	0.407907	-0.511760
26	1	Dorset Park	43.757410	-79.273304	251476	0.039995	-0.511760
27	1	Forest Hill North	43.696948	-79.411307	258469	0.546619	-0.511760
28	1	Willowdale West	43.782736	-79.442259	272986	0.295229	-0.511760
29	1	Roncesvalles	43.648960	-79.456325	377518	0.500868	-0.511760
33	1	Alderwood	43.602414	-79.543484	168602	0.456280	-0.511760

Fig 9.2: Cluster Label 1

Cluster Label 1:

- *MID Spending Power*
- *LOW Percentage of Target Customers*
- *LOW Number of Competitors*

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Mexicans	No. of Mexican Restaurants
1	2	Rouge	43.806686	-79.194353	729154	0.107536	-0.511760
2	2	Malvern	43.806686	-79.194353	533202	0.125588	-0.511760
10	2	The Beaches	43.676357	-79.293031	659192	0.510038	0.873411
11	2	Woburn	43.770992	-79.216917	629030	0.168271	-0.511760
30	2	Agincourt North	43.815252	-79.284577	427037	0.103047	-0.511760
31	2	Milliken	43.815252	-79.284577	525507	0.075267	-0.511760

Fig 9.2: Cluster Label 2

Cluster Label 2:

- *HIGH Spending Power*
- *LOW Percentage of Target Customers*
- *LOW Number of Competitors*

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Mexicans	No. of Mexican Restaurants
5	3	Humewood-Cedarvale	43.693781	-79.428191	289910	0.939784	1.117853
7	3	Guildwood	43.763573	-79.188711	177062	1.058788	-0.511760
24	3	Mount Dennis	43.691116	-79.476013	164344	0.846024	-0.511760
34	3	Long Branch	43.602414	-79.543484	137480	0.743752	-0.511760

Fig 9.2: Cluster Label 3

Cluster Label 3:
• MID Spending Power
• HIGH Percentage of Target Customers
• LOW Number of Competitors

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Mexicans	No. of Mexican Restaurants
6	4	Markland Wood	43.643515	-79.577201	195412	0.379003	1.310834
21	4	Oakridge	43.711112	-79.284577	134303	0.252799	1.412089
25	4	Weston	43.706876	-79.518188	171044	0.500222	2.315120

Fig 9.2: Cluster Label 4

Cluster Label 4:
• MID Spending Power
• LOW Percentage of Target Customers
• HIGH Number of Competitors

5. Results and Discussion

In this study, I have labeled the neighbourhoods corresponding to their characteristics-- spending power, percentage of target customers, and the number of competitors. The most promising group of neighbourhoods for opening a Mexican restaurant appears to be ‘Cluster Label 3’. The medium spending power of the neighbourhoods in this cluster allows them to readily afford the low to medium prices of the client’s Mexican restaurant menu. The slightly high in comparison distribution of target customers--Mexicans--indicates a relatively reasonable demand for Mexican cuisine. The number of competitors is low and is a good stepping stone for a successful marketing campaign to attract new visitors. Our client could more specifically consider Humewood-Cedarvale as a location for establishment for optimal results. However, wherever there is a shift in the dynamic of business demands, we could always target different clusters of neighborhoods. Case in point, if the client has plans to expand a well-established franchised restaurant, ‘Cluster Label 2’ would be the optimal location due to high spending power in those neighbourhoods.

6. Conclusion

In conclusion, the extensive analysis above would greatly increase the likelihood of the restaurant's success. Similarly, we can use this project to analyze interchangeable scenarios, such as opening a restaurant of different cuisines. I think the model could use more improvements in capturing restaurants' individual traits. For example, two restaurants might have a similar number of competitors, but one might have a smaller geographical radius while the other might have a bigger radius. More data, especially data of different types, would help to improve the model performance significantly.