Introduction

Toronto is the most populous city in Canada with almost 3million inhabitants and its number of restaurants per capita is bellow many other cities (16th place worldwide with 272 restaurants per 100.000 inhabitants).

For many expats having a local restaurant nearby is a great way to remember their roots and taste some home dishes and this can present great business opportunities.

Business Problem

The goal for this Capstone project is to suggest the best locations for a new Portuguese restaurant in Toronto, Canada. Portuguese Canadians account for more than 170.000 people just in Toronto, that alone can be a driver for new businesses in the restauration domain. Apart from that Portuguese cuisine is one of the best in the world so that could attract investors too.

In summary the goal is to answer the following question: If an investor is looking to open a new Portuguese restaurant where would you recommend, they open it?

Target Audience

This project can be useful both for new investors looking to open new Portuguese themed restaurants and also for existing owners looking to expand or have an insight on the current concentration of businesses around Toronto.

Data

In order to solve the problem we need the following data:

- List of neighbourhoods in Toronto, defining the boundaries in terms of segmentation for attractiveness for new restaurants.
- Latitude and longitude coordinates for those neighbourhoods in order to be able to plot the maps and also to get existing restaurants in the area
- Venue data related to Portuguese Restaurants in order to perform clustering on the neighbourhoods.

Sources of data:

Toronto Neighbourhoods on Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

Toronto Geospatial coordinates (https://cocl.us/Geospatial_data)

Foursquare API for Venue data regarding the Portuguese Restaurants https://developer.foursquare.com/docs/resources/categories Category Id: 4def73e84765ae376e57713a

Methodology

In order to answer the question posed in the Business Problem section we need to go through a series of steps so we can generate some insightful data.

Toronto Neighbourhoods Data Exploratory Analysis

First we have scraped data from Wikipedia as described in the Data section regarding the neighbourhoods in Toronto. After that we have merged that information with geo data by Postal Code using the file mentioned in the Data section, reaching the following consolidated data frame:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

With that information we were able to plot the complete list of neighbourhoods to a map:

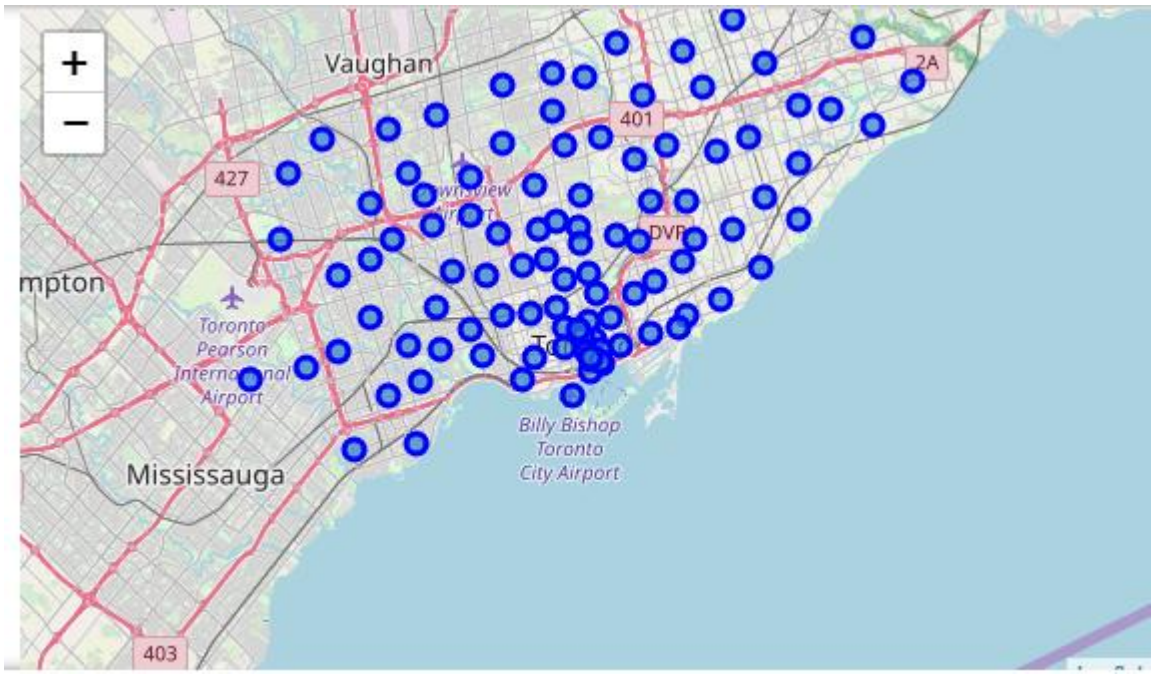

Figure 1 - Toronto Neighbourhoods

Portuguese Restaurants Data Exploratory Analysis

To gather information about existing Portuguese restaurants we have used the Foursquare API to retrieve data. We have queried the first 100 venue within 1km radius from each neighbourhood.

This list included 16 different categories that are siblings of the foursquare category

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Cedarbrae	43.773136	-79.239476	Nando's Flame-Grilled Chicken	43.773113	-79.281166	Portuguese Restaurant

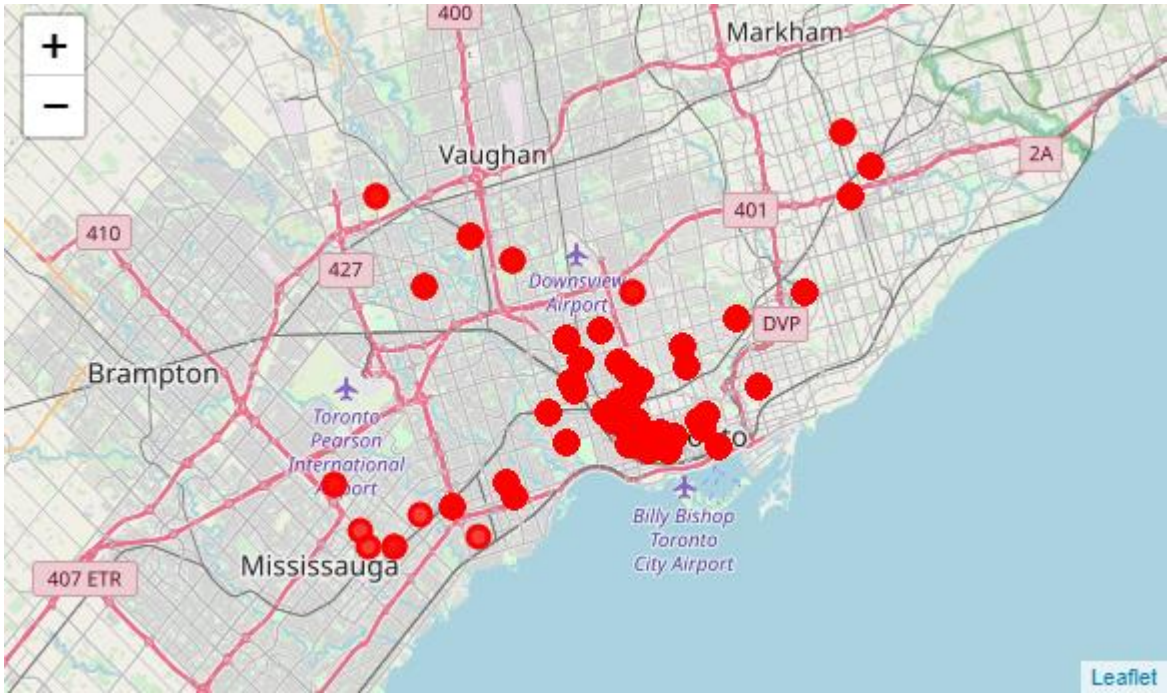

Figure 2 - Portuguese Restaurants in Toronto

After this we have used one hot encoding in order to have a data frame with venue category columns per neighbourhood and then we have aggregated the results so we have a frequency based data frame for each venue category.

	Neighborhood	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Bar	Café	Dessert Shop	Gay Bar	Italian Restaurant	Medit R
0	Adelaide,King,Richmond	0.000000	0.0	0.290323	0.032258	0.032258	0.0	0.0	0.032258	0.032258	
1	Agincourt	0.333333	0.0	0.333333	0.000000	0.000000	0.0	0.0	0.000000	0.000000	
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	0.333333	0.0	0.333333	0.000000	0.000000	0.0	0.0	0.000000	0.000000	
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	
4	Alderwood,Long Branch	0.000000	0.0	0.250000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	

Using this data frame we built a new one with the top 10 venues per neighbourhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	
0	Adelaide,King,Richmond	Portuguese Restaurant	Bakery	Restaurant	Wine Bar	Sandwich Place	Mediterranean Restaurant	Italian Restaurant	
1	Agincourt	Portuguese Restaurant	Bakery	Asian Restaurant	Wine Bar	Seafood Restaurant	Sandwich Place	Restaurant	
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	Portuguese Restaurant	Bakery	Asian Restaurant	Wine Bar	Seafood Restaurant	Sandwich Place	Restaurant	
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	Portuguese Restaurant	Wine Bar	Seafood Restaurant	Sandwich Place	Restaurant	Pizza Place	Mediterranean Restaurant	
4	Alderwood,Long Branch	Portuguese Restaurant	Restaurant	Bakery	Wine Bar	Seafood Restaurant	Sandwich Place	Pizza Place	M

Used Machine Learning Algorithms

In order to reach a conclusion regarding the best places to open a Portuguese restaurant we have used two complementary approaches:

- K-Means Clustering on the top 10 venues per neighbourhood data to extract areas with less Portuguese restaurants
- DBScan on the Venues in order to cluster the existing venues per density and revealing areas there the concentrarion is lower and where there is an opportunity to open a new restaurant.

In order to compare both approaches we have then applied the DBScan results onto the corresponding neighbourhoods so we can take a more informed decision about the best areas to invest.

Results

K-means Clustering

The results from the k-means clustering using 4 clusters based on the frequency of Portuguese restaurants show the following results:

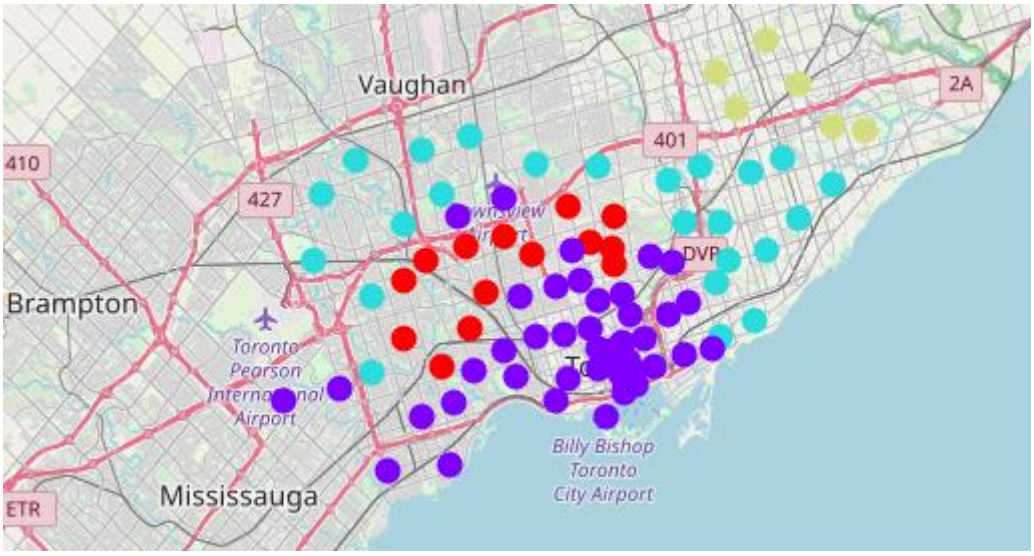

Figure 3 - K-means Clustering

- Cluster 0 (Red)
    - Low concentration of Portuguese restaurants

	MostCommon	All
Portuguese Restaurant	9.0	14.0
BBQ Joint	1.0	NaN
Bakery	1.0	NaN
Sandwich Place	1.0	14.0
Seafood Restaurant	1.0	14.0
Restaurant	NaN	14.0
Wine Bar	NaN	14.0

- Cluster 1 (Purple)
    - High concentration of Portuguese restaurants

	MostCommon	All
Portuguese Restaurant	46.0	46
Bakery	NaN	44
Mediterranean Restaurant	NaN	43
Restaurant	NaN	46
Wine Bar	NaN	42

- Cluster 2 (Cyan)
    - Medium concentration of Portuguese restaurants

	MostCommon	All
Portuguese Restaurant	24.0	NaN
Italian Restaurant	NaN	24.0
Pizza Place	NaN	24.0
Restaurant	NaN	24.0
Seafood Restaurant	NaN	24.0
Wine Bar	NaN	24.0

- Cluster3 (Green)
    - Low Concentration of restaurants

	MostCommon	All
Portuguese Restaurant	5.0	NaN
Asian Restaurant	1.0	6.0
Italian Restaurant	NaN	6.0
Pizza Place	NaN	6.0
Restaurant	NaN	6.0
Seafood Restaurant	NaN	6.0

This suggest that neighbourhoods in clusters 0 and 3 are the suggested to invest in with cluster 2 coming after:

- Cluster 0
    - Lawrence Park
    - Davisville North
    - North Toronto West
    - Davisville
    - Bedford Park,Lawrence Manor East
    - Lawrence Heights,Lawrence Manor
    - Glencairn

        - Downsview,North Park,Upwood Park
        - Del Ray,Keelesdale,Mount Dennis,Silverthorn
        - The Junction North,Runnymede
        - The Kingsway,Montgomery Road,Old Mill North
        - Islington Avenue
        - Weston
        - Westmount
    - Cluster 3
        - Woburn
        - Cedarbrae
        - Agincourt
        - Clarks Corners,Sullivan,Tam O'Shanter
        - Agincourt North,L'Amoreaux East,Milliken
        - L'Amoreaux West
    - Cluster 2
        - Scarborough Village
        - East Birchmount Park,Ionview,Kennedy Park
        - Clairlea,Golden Mile,Oakridge
        - Dorset Park,Scarborough Town Centre,Wexford He...
        - Maryvale,Wexford
        - York Mills West
        - Parkwoods
        - Don Mills North
        - Flemingdon Park,Don Mills South
        - Bathurst Manor,Downsview North,Wilson Heights
        - Northwood Park,York University
        - Downsview West
        - Downsview Northwest
        - Victoria Village
        - Woodbine Gardens,Parkview Hill
        - Woodbine Heights
        - The Beaches
        - The Beaches West,India Bazaar
        - Cloverdale,Islington,Martin Grove,Princess Gar...
        - Humber Summit
        - Emery,Humberlea
        - Kingsview Village,Martin Grove Gardens,Richvie...
        - Albion Gardens,Beaumond Heights,Humbergate,Jam...
        - Northwest

DBSCAN Clustering

We have performed a DBSCAN clustering on the venues data to extract clusters that reflect the concentration of Portuguese restaurants.

Using an epsilon of 0.5km as the minimum distance to represent a high density area and setting the minimum number of venues to represent a cluster as 2 we have reached the following results:

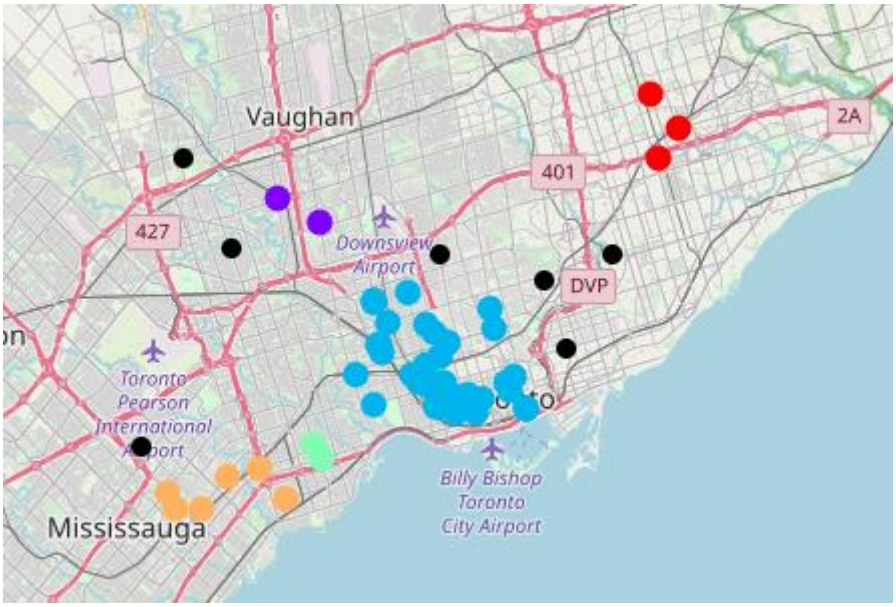

Figure 4 - Venues DBSCAN

The results show 5 clusters and several venues as outliers, these areas are suitable for investment.

Applying these results to the neighbourhoods we get the following distribution:

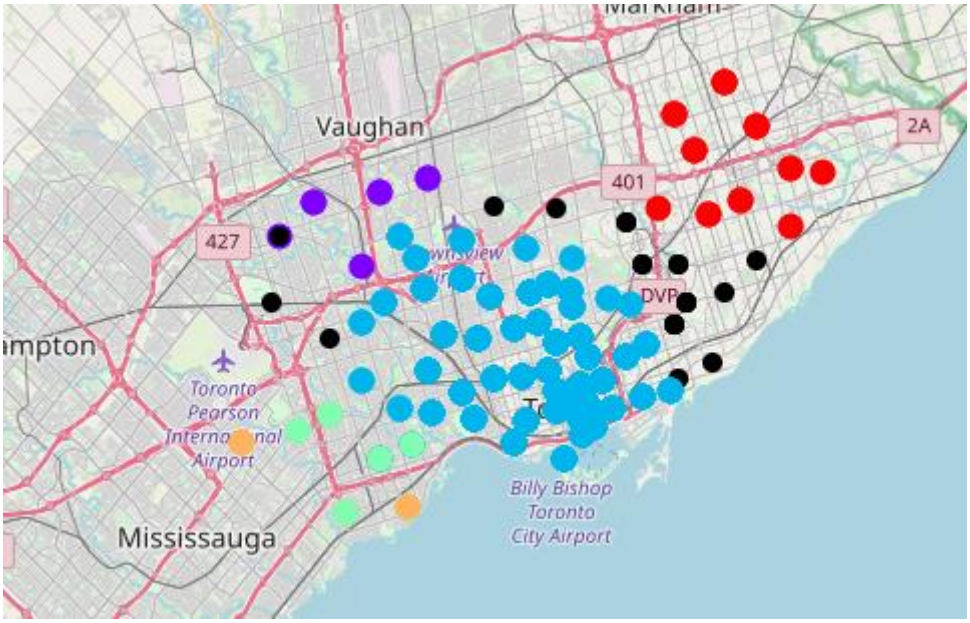

Figure 5 - Neighbourhoods DBSCAN

Discussion

The analysis shows that there is an high concentration of Portuguese restaurants in the center of Toronto. This is visible both in the results from K-means (Purple cluster) and also using the DBSCAN (blue clusters).

Recommendations in terms of invest are discouraged for these areas and the investors should focus on the lower density areas shown by the outliers and then in smaller size clusters on the DBSCAN

analysis. The smaller size clusters 0 and 3 as shown in the K-means analysis overlap significantly with these suggesting that both approaches are complementary.

Conclusion

Although the main goal for this capstone projects were achieved and allow the investors to have a clearer picture about the areas more suitable for opening restaurants without much competition the study uses relatively few data and more variables should be considered such as transportation and population density so we don't recommend areas that have no business to offer.

Another interesting point could be a cross check between the existing venues and the distribution of the Portuguese community, allowing to identify high potential areas for opening new restaurants.