Opening a Mediterranean Restaurant in Toronto

Adarsh Ravikumar

November 22, 2020

1. Introduction

1.1 Background

In a large and highly populated city, there are so many different shops and venues which are all around. One of the biggest and most common cities to start a business in is Toronto, Canada. There are thousands of businesses already there, and so many more which are being opened. Being able to stay afloat among all of the competition could be a huge struggle for some. The first and most defining step of owning and operating a successful business is choosing the right place to open up shop. There are many different factors to consider and methods to use in deciding the most optimal location for a new and thriving business.

1.2 Problem

For new businesses being opened, one of the biggest issues which they have to deal with is keeping up with their competitors. For example, if someone wanted to start a Pizzeria, they would definitely have to consider their competition. Within a neighborhood, if there were multiple different shops which sold Pizza, then it could be difficult to keep a brand new Pizzeria afloat. However, if they could find a spot in the city which doesn't have any other places selling Pizza, has ample foot traffic, and the proper demographic, then the shop owner wouldn't have to worry about competition and the business would have a better chance at staying successful.

1.3 Audience of Interest

The target audience would be the shop owners and any stakeholders they have for their business. Being able to determine a prime location for a shop is important because it would help make the business as successful as possible. Success is the biggest thing which stakeholders and owners look for because profit is the main goal of opening a business and it would make them very happy.

2. Importing and Wrangling Data

2.1 Data Source

For this project, the data being used is geographical data which is called from the Foursquare API. This API uses geographical coordinates to center in on a location. When latitude and longitude are entered, using the API's explore endpoint leads to a json file with many different venues. When using a call from Foursquare, some of the returned results include filters such as relevance, popularity, and distance. It is also possible to get ratings, reviews, nearby streets, and other relevant details about a venue. Other data being used for this project comes from CKan. This data includes information about Pedestrian and Vehicle traffic is neighborhoods and streets around Toronto. Lastly, is another dataset from CKan. This set includes many different features and information about neighborhoods and their profiles/demographic information.

Dataset 1: Pedestrian and Vehicle Traffic Volumes
https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/ae4e10a2-9eaf-4da4-83fb-f3731a30c124/resource/ea4d9b68-f645-4878-bd1d-d7273450255c/download/traffic-signal-vehicle-and-pedestrian-volumes-data.xlsx

Dataset 2: Neighbourhood Profiles
https://open.toronto.ca/dataset/neighbourhoods/

Dataset 3: Postal Codes with Neighborhoods and Boroughs
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Dataset 4: CSV with the Latitude and Longitude of Postal Codes
http://cocl.us/Geospatial_data

2.2 Importing Data

To create the necessary data for this project, there were many different steps which had to be taken. First, was a list of all the neighbourhoods and postal codes from a wikipedia page. Scraping that html provided a dataframe which contained a list of all the postal codes with their neighbourhood and boroughs. After that, the latitude and longitude were necessary to call the foursquare API to get the neighbourhood venue data. By using the csv file which had each postal code and their corresponding latitude and longitude, it could be read into its coordinate dataframe. Then, the two tables were merged based on the criteria 'postal code' to create a combined dataset which had all of the necessary information to properly utilize the API.

Another dataset which was imported was information from ckan about Pedestrian and Vehicle Traffic volumes based on 8 hour averages and roads throughout Toronto. This set was downloaded as an Excel file and was then read into a dataframe.

Lastly, was another dataset imported from ckan which contained information about neighborhoods and their profiles/demographics. This file was downloaded as a csv and was then read into a dataframe.

2.3 Data Wrangling

The first set to work with was the combined information about the Neighborhoods with the postal codes, boroughs, latitude, and longitude. For the most part, the data was almost fully clean enough to work with. The first step in cleaning it was removing any rows which had unassigned boroughs. Another thing which was notable was that some postal codes included more than one neighborhood. To avoid duplicate rows, the neighbourhoods were grouped together in the data based on the postal code. The rows without a neighbourhood, but not an assigned borough, were assigned the same borough as the neighbourhood.

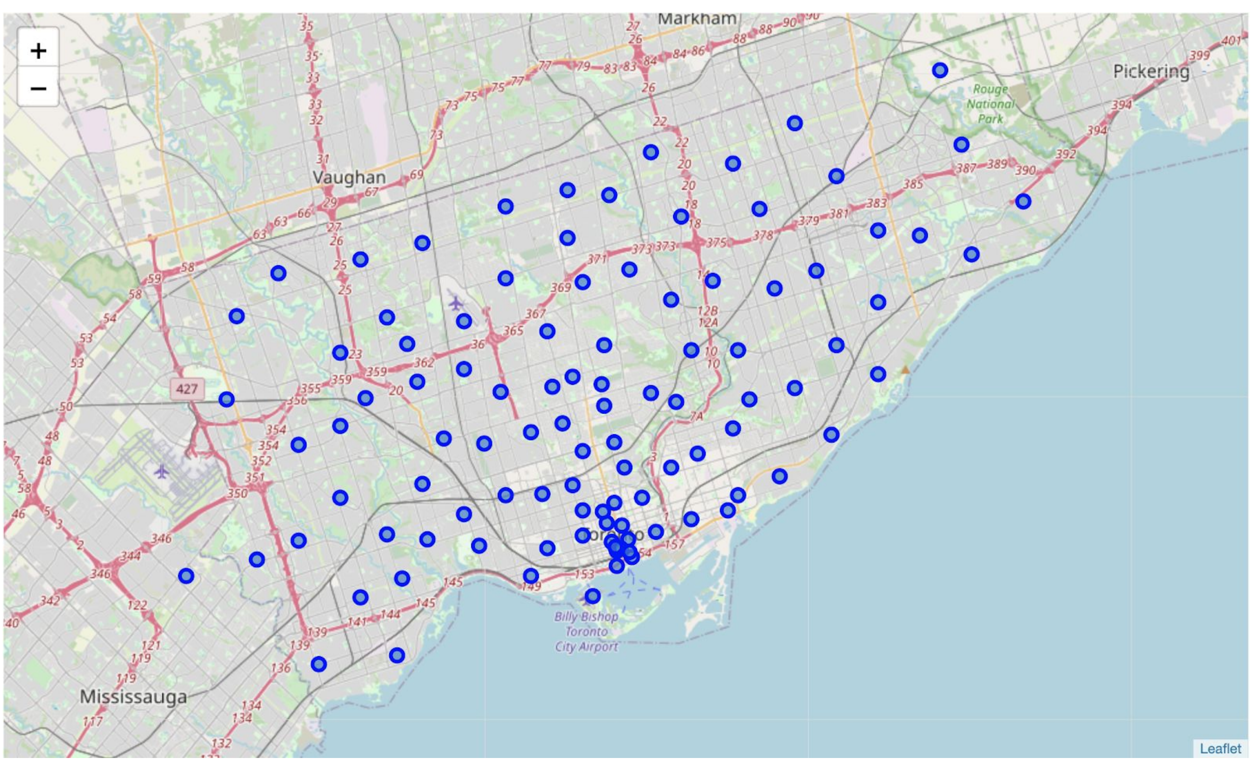
The next set being worked with was the information about the Vehicle and Pedestrian Traffic Volumes. The relevant columns were selected, which included the roads, their latitude and longitudes, and their 8 Peak Hour Averages for both Vehicles and Pedestrians.

Finally, was the dataset about each neighborhood and their profiles/demographics. In this specific situation, since we are looking to open a Mediterranean Restaurant, the population being gathered was middle eastern ethnicities in the area. The neighborhoods with the largest total population of these different ethnicities were of interest for the results of this project.

3. Methodology

3.1 Exploratory Data Analysis

In exploring the data, one of the first steps was to gather information about each neighborhood and the different venues surrounding them. Using the foursquare API, this information could be found for each neighborhood and put into a dataframe. The first thing which was done is that all the neighborhoods were graphed onto a map using Folium and circle markers, and that looked like this:



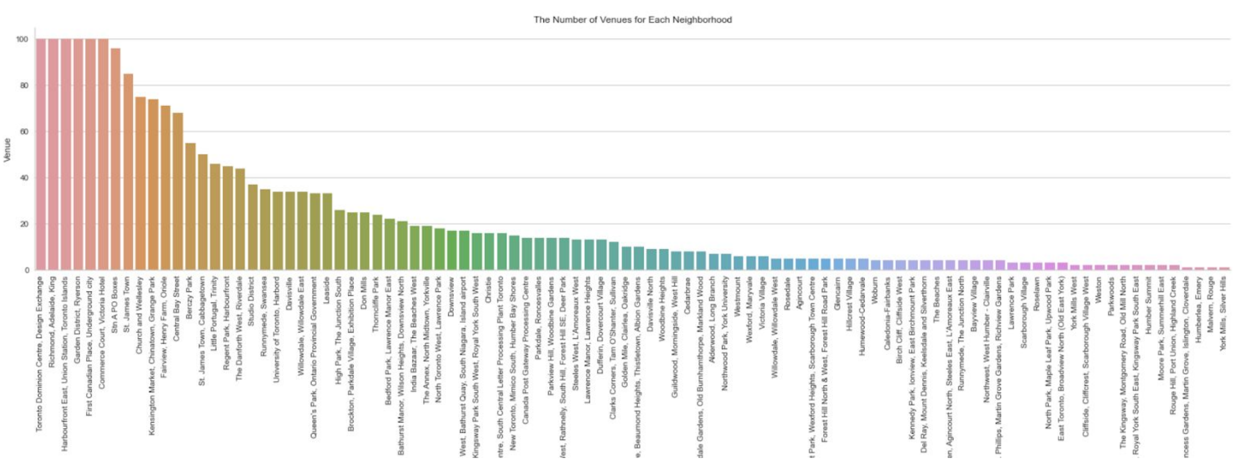
Then, the venue information was gathered for each neighborhood. Here is an example of what the venue information which was returned looked like for a certain neighborhood:

	name	categories	lat	lng
0	Economy Rent A Car	Rental Car Location	43.708471	-79.589943
1	Logistics Distribution	Bar	43.707554	-79.589252
2	Saand Rexdale	Drugstore	43.705072	-79.598725
3	PC Garden	Garden Center	43.706539	-79.599359

By gathering data like this, the information for all of the venues in Toronto could be collected into a dataframe like so:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	SEBS Engineering Inc. (Sustainable Energy and ...	43.782371	-79.156820	Construction & Landscaping
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

Using this collective data, the neighborhoods could be graphed into showing the number of venues in descending order.



After that, the next step was to analyze the venues in each neighborhood by using the different venue categories as labels and then using that information to create a table displaying the number of each different venue category in the neighborhood, like this:

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Res
0	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

The frequency of each venue could also be listed individually for each neighborhood and the top 5 values could be listed for each like this:

```
----Del Ray, Mount Dennis, Keelsdale and Silverthorn----
                     venue  freq
0        Turkish Restaurant  0.25
1                Restaurant  0.25
2            Sandwich Place  0.25
3            Discount Store  0.25
4  Mediterranean Restaurant  0.00


----Don Mills----
                venue  freq
0                 Gym  0.12
1          Beer Store  0.08
2         Coffee Shop  0.08
3  Japanese Restaurant  0.08
4    Asian Restaurant  0.04
```

Finally, a dataframe could be made which contained each neighborhood with their most common venues which was insight into the kind of places each neighborhood had alot of, and it looked like this:

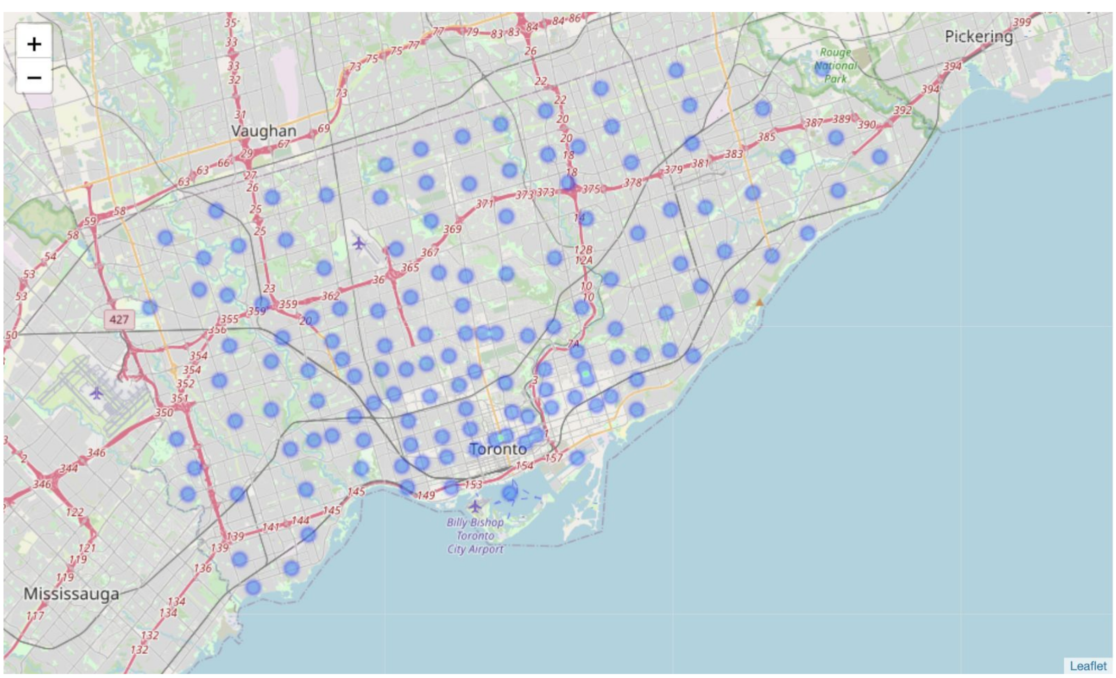
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Lounge	Breakfast Spot	Latin American Restaurant	Skating Rink	Clothing Store	Ethiopian Restaurant	Event Space	Escape Room	Electronics Store	Discount Store
1	Alderwood, Long Branch	Pizza Place	Sandwich Place	Coffee Shop	Pub	Pharmacy	Gym	Greek Restaurant	Discount Store	Department Store	Dessert Shop
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Pharmacy	Deli / Bodega	Shopping Mall	Bridal Shop	Sandwich Place	Diner	Restaurant	Middle Eastern Restaurant
3	Bayview Village	Japanese Restaurant	Café	Bank	Chinese Restaurant	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Yoga Studio
4	Bedford Park, Lawrence Manor East	Sandwich Place	Italian Restaurant	Coffee Shop	Greek Restaurant	Thai Restaurant	Liquor Store	Comfort Food Restaurant	Juice Bar	Butcher	Café
5	Berczy Park	Coffee Shop	Restaurant	Bakery	Cocktail Bar	Beer Bar	Farmers Market	Cheese Shop	Seafood Restaurant	Sandwich Place	Beach

3.2 Target Venue Information - Mediterranean

Following that, the different categories which could be considered Middle Eastern were grouped into a variable called 'Middle Terranean.' By adding these up, the frequencies of 'Middle Terranean' restaurants could be calculated for each neighborhood. The venue categories which were considered 'Middle Terranean' included: ['Doner Restaurant', 'Mediterranean Restaurant', 'Middle Eastern Restaurant', 'Falafel Restaurant', 'Greek Restaurant', 'Turkish Restaurant']. The result dataframe which contained all of these looked like this:

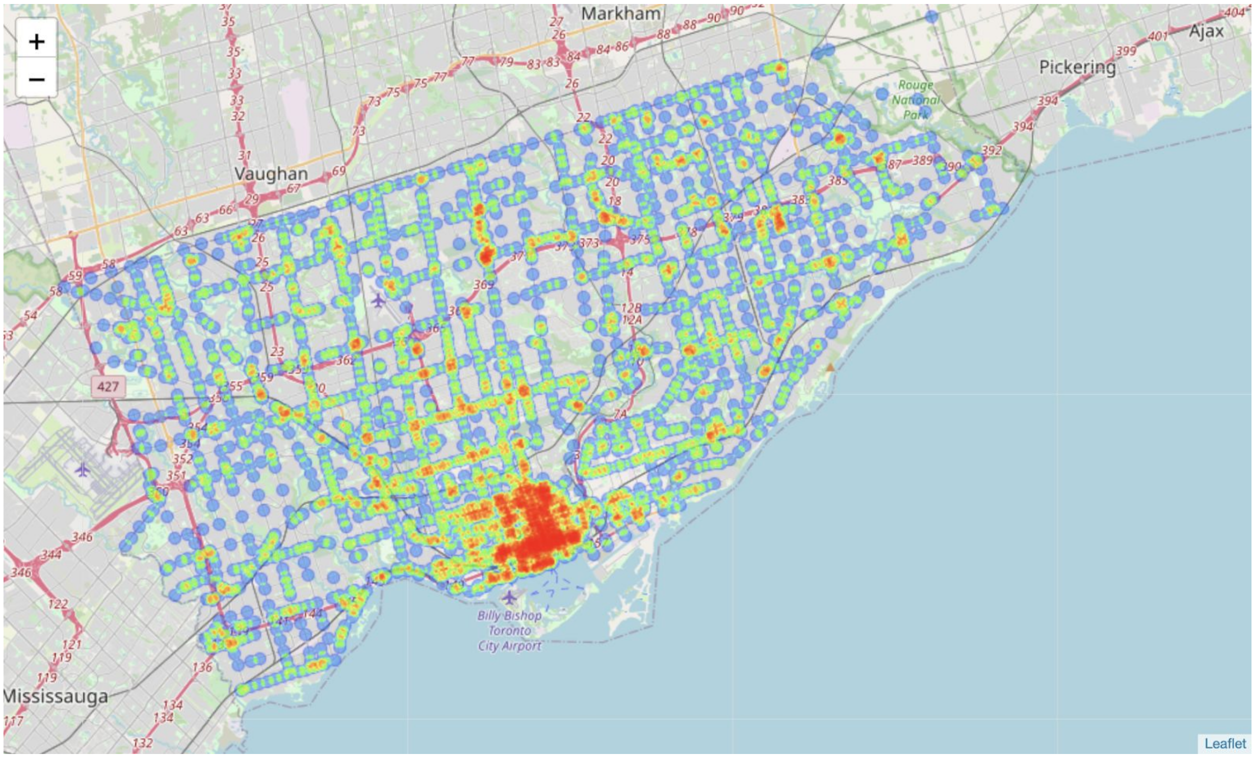
	Middle Terranean	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum
21	0.250000	Del Ray, Mount Dennis, Keelsdale and Silverthorn	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
37	0.200000	Hillcrest Village	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
89	0.166667	Wexford, Maryvale	0.166667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
80	0.159091	The Danforth West, Riverdale	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.022727	0.0	0.0	0.0	0.0
11	0.142857	Canada Post Gateway Processing Centre	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.071429	0.0	0.0	0.0	0.0

There are other factors which need to be taken into account while finding a good place to open up a shop. Some of these include foot traffic, vehicle traffic, and target demographics. The Toronto Neighborhood Profiles helped provide some insight into which neighborhoods Middle Eastern people are more populated in. By adding ethnicities to a group, this number can be added up and divided by the total neighborhood population in each of the neighborhoods. By using folium, a heatmap was created which showed which areas had a higher target population than the others.

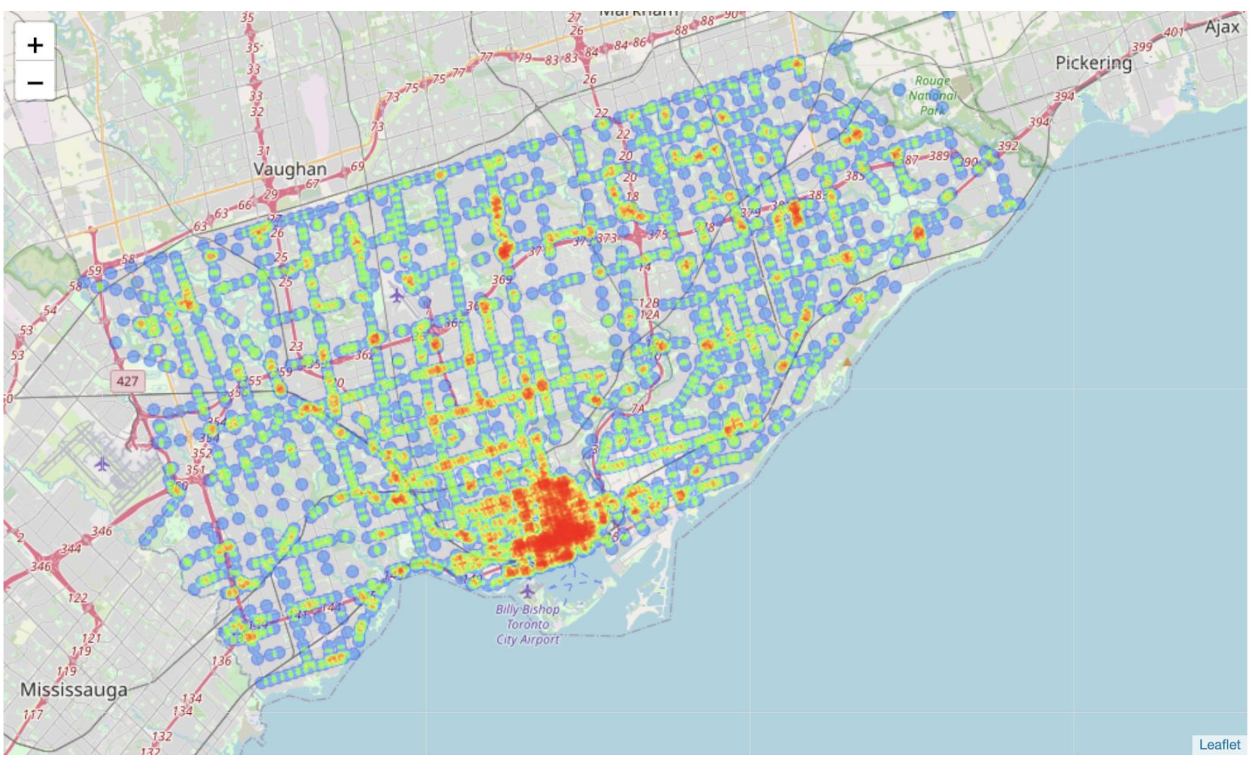


Also, another important factor is the vehicle and pedestrian foot traffic in each of the areas. Finding places with heavier traffic are better in making sure a business can keep busy. The dataset provided 8 Hour Peak Vehicle and Pedestrian Traffic Volume. By dividing the values by the max value for each, a heatmap was created for both to show the traffic density for each.

For Vehicles:

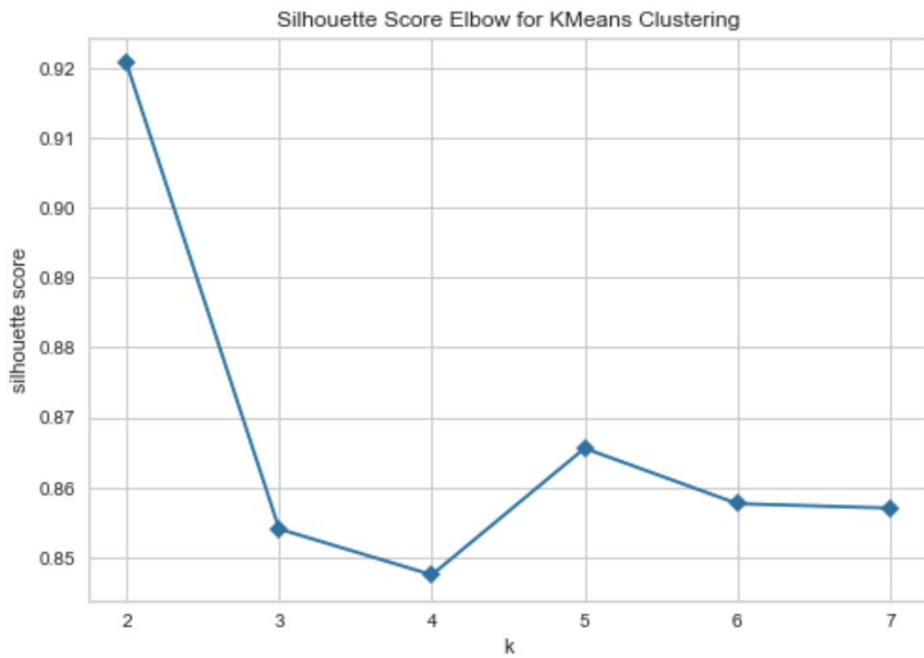


For Pedestrians:



3.3 K-Means Clustering

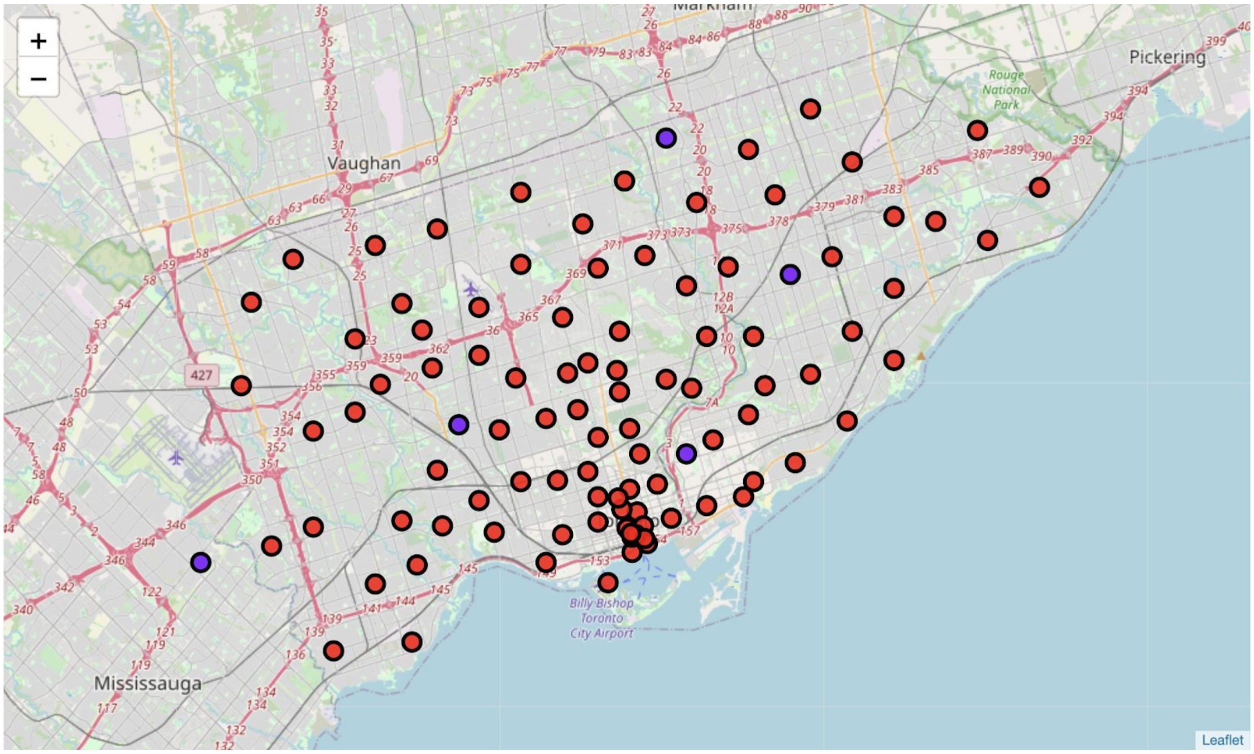
Now, based on the table which included the sum of 'Middle Terranean' venues frequencies, a k-means clustering model was able to be run. The first step of the clustering was to use the KElbow visualizer to figure out which number of clusters would provide the optimal results.



This graph gives a silhouette score for each number of k clusters. A silhouette score signifies the similarity which a value inside a cluster has to the cluster itself. A higher silhouette score means it is a better number of clusters to use. In this case, the optimal number of clusters was 2. Now, the clustering could be run using k number of clusters, which in this case meant k = 2. The resulting cluster labels were given in an array like this.

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0], dtype=int32)
```

Then, when merging the labels with the dataframe being used, the result was a combined dataframe which had all the proper information about each neighborhood with their latitude, longitude, most common venues, and cluster labels. Finally, the clusters were mapped using circle markers in folium and the resulting map looked like this with Cluster_0 being red and Cluster_1 being purple:



4. Results

4.1 Cluster Exploration

Based on the resulting clusters, there are a few observations which can be made. First is that, in Cluster 1, the neighborhoods can be displayed and include these 5:

	Cluster Labels	Postal Code	Latitude	Longitude	Borough	Neighbourhood	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
11	1.0	M1R	43.750072	-79.295849	Scarborough	Wexford, Maryvale	Accessories Store	Sandwich Place	Bakery	Middle Eastern Restaurant	Auto Garage	Smoke Shop	Electronics Store
17	1.0	M2H	43.803762	-79.363452	North York	Hillcrest Village	Mediterranean Restaurant	Athletics & Sports	Pool	Dog Run	Golf Course	Drugstore	Distribution Center
41	1.0	M4K	43.679557	-79.352188	East Toronto	The Danforth West, Riverdale	Greek Restaurant	Coffee Shop	Italian Restaurant	Bookstore	Restaurant	Cosmetics Shop	Ice Cream Shop
80	1.0	M6M	43.691116	-79.476013	York	Del Ray, Mount Dennis, Keelsdale and Silverthorn	Restaurant	Discount Store	Sandwich Place	Turkish Restaurant	Yoga Studio	Doner Restaurant	Diner
86	1.0	M7R	43.636966	-79.615819	Mississauga	Canada Post Gateway Processing Centre	Intersection	Coffee Shop	Hotel	Fried Chicken Joint	Sandwich Place	Gym	Mediterranean Restaurant

It can be observed that of these Neighborhoods, they have a decent amount of most common venues which fall under the previously created 'Middle Terranean' column. For example, Wexford has Middle Eastern Restaurant as its 4th most common venue, Hillcrest Village has Mediterranean Restaurant as its 1st most common venue, The Danforth West, Riverdale has Greek Restaurant as its 1st most common venue, Del Ray has Turksish and Doner Restaurant respectively as its 4th and 6th most common venue, and so on. Thus it can be seen why these different places were put into the same cluster. Another observation which can be made is that when comparing the top 10 Neighborhoods with the highest 'Middle Terranean' frequency with the top 10 Neighborhoods with the highest target population percentage, the results for each of them provide a few different Neighborhood values in common.

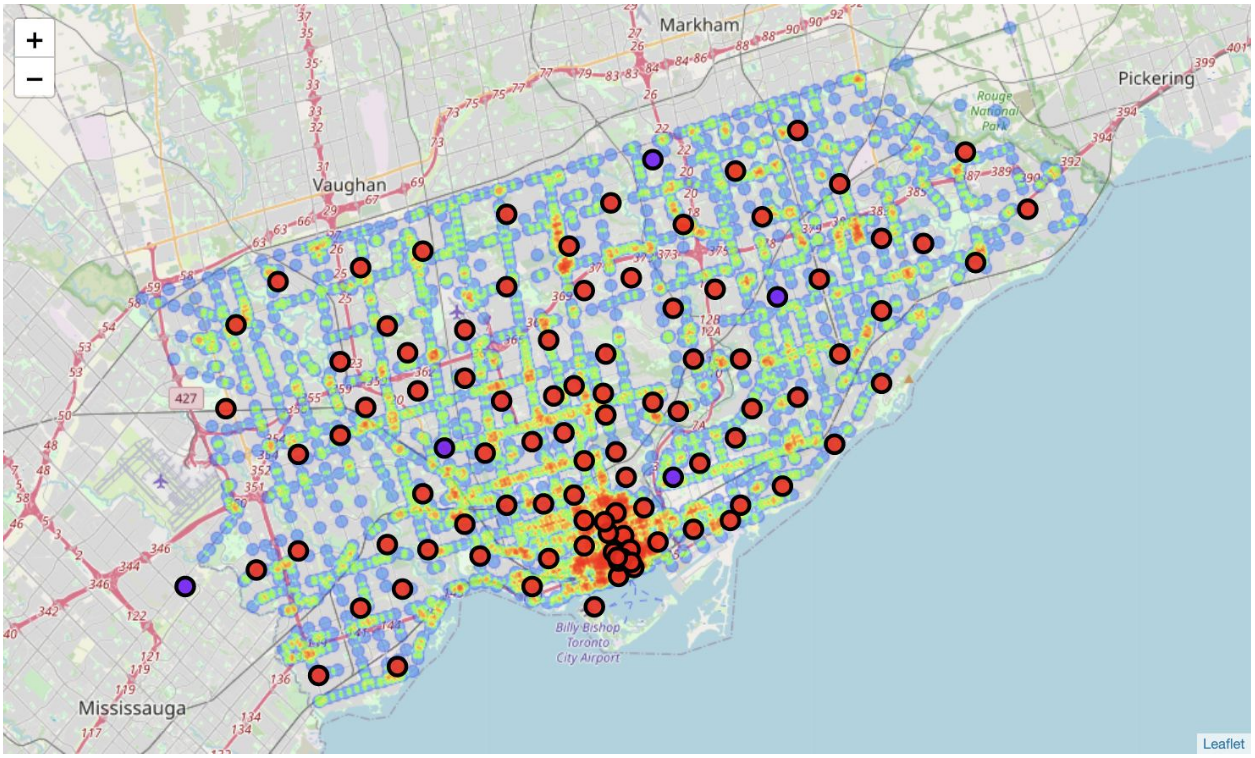
Newtonbrook East
Old East York
Willowdale East
Broadview North
Flemingdon Park
Bayview Village
Victoria Village
Henry Farm
Thorncliffe Park
Wexford/Maryvale

Del Ray, Mount Dennis, Keelsdale and Silverthorn
Hillcrest Village
Wexford, Maryvale
The Danforth West, Riverdale
Canada Post Gateway Processing Centre
Dufferin, Dovercourt Village
The Annex, North Midtown, Yorkville
Bathurst Manor, Wilson Heights, Downsview North
Bedford Park, Lawrence Manor East
Thorncliffe Park

Some of the common neighborhoods include Thorncliffe Park, Wexford/Maryvale, and Yorkville. Based on these results, it could possibly be assumed that there is indeed some slight correlation between the target demographic and the target venue categories.

5. Discussion

When the cluster mapped is overlaid onto the pedestrian traffic map, the following is obtained:



Based on this map, it can clearly be seen that the area with the most dense pedestrian traffic is downtown where a large portion of the neighborhoods from Cluster_0 are. While Cluster_1 contains more venues which are considered 'Middle Terranean,' the pre-existing businesses there could make it difficult for a new business of a similar category to be successful. Downtown has a red heat spot with many overlapping markers.

6. Conclusion

This project was done with the goal of deciding where the best place to open a Mediterranean Restaurant would be in Toronto. It would be reasonable to say an ideal spot would be a neighborhood in Cluster_0, preferably near downtown. Also, the areas with the heaviest Pedestrian and Vehicle traffic are downtown and along the main roads. Finding a neighborhood which satisfies those criteria would likely be the most suitable location to open up a new Mediterranean restaurant. An example of a great neighborhood to pick would be Old York because of its proximity to traffic and the target demographic.

7. Future Improvements

There are a few different things which can be done to improve these results in the future. Some other important factors to consider include accessibility, parking, competitor reviews, customer dwellings/queue, and nearby retail support. All of these could make a difference in picking the ideal spot to open up any shop. Other useful information would be data about consumer demographics, in order to target more specific communities. Specific location information such as whether it is in a shopping complex or a strip mall could also help influence this decision. Last but not least, would be surrounding neighborhood socioeconomic information. This could play a significant role in determining if the business would fit the proper neighborhood economic profiles.