Capstone Project - Where is the best place to open a restaurant in Toronto?

1. Introduction/Business Problem

Before opening a new restaurant in any place, it is essential to investigate the demand for it. It is crucial to investigate the criteria such as population density, availability of other restaurants (competition), income diversity of people around that area to identify whether opening a restaurant would be profitable. These features collectively enable identifying the hotspots for a restaurant that would generate the best income. In this project, we consider these aspects in identifying the best location in Toronto to open a restaurant by a certain client.

The business problem is identifying and recommending the client with a certain location/s in Toronto that would produce the best profit.

In order to achieve this task, we will investigate the population density, availability of other restaurants, and income diversity of the people in the Toronto neighborhoods. The population density will allow us identifying the densest areas that will roughly tell how frequently people will order food from a restaurant. For example, when a restaurant is located in a highly condensed area, the frequency of buying food from a certain restaurant is high. The availability of other restaurants will determine the competition. If a particular location has too many restaurants, it is not easy to attract many food buyers. While population density and competition play two significant roles, it is also essential to identify how affluent the area is. If the household income is lower in an area, buying food from restaurants might not be preferred. Based on these factors, we will use Foursquare API to explore neighborhoods in Toronto.

2. Data

For this project, we will first need neighborhood data of Toronto. Hence, we scrape neighborhood data from the Wikipedia page available at,
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Next, we wrangle the data and clean them to generate the following dataframe (Please note that it is only an abstract view of the entire dataset, and the original dataset has 103 records). The obtained dataset is composed of three columns: PostalCode, Borough, and Neighborhood. Note that we do not consider the records where the borough is "Not assigned". The records where one postal code having more than one neighborhood are merged into one record.

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills North
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson
10	M6B	North York	Glencairn
11	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...

Once the above dataframe is created, we merge the latitude and the longitude data of each location (each record) using the CSV file that has the geographical coordinates of each postal code available at https://github.com/courseraassignmentcapstone/Coursera_Capstone/blob/main/Geospatial_Coordinates.csv.

Once the latitude and the longitude data are available, we can utilize Foursquare location data to further analyze the dynamics of the dataset to accomplish the necessary tasks of the project. An abstract view of the dataset with the latitude and longitude information is given below.

Note: In order to access Foursquare services, we need to have a valid developer account at https://foursquare.com/.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304
11	M1R	Scarborough	Wexford, Maryvale	43.750072	-79.295849

We obtain the data related to population density, availability of other restaurants (competition), income in Toronto from http://map.toronto.ca/wellbeing/#eyJ0b3Itd2lkZ2V0LWNsYXNzYnJlYWsiOsSAcGVyY2VudE9wYWNpdHnElzcwfSwiY3VzxIJtYcSTYcSXxIBuZWlnaGJvdXJob29kc8S2fcSrxIHEg8SFxIfEicSLdGFixYXEmCLEo3RpdmVUxZBJZMSXxYnEhMWPYi1pbmRpY2HEgnLFhcWIYWdzTWFwxLYiesWCbcSXMTPErHjEly04ODM3NzYzLjXGhDcyN8SsxKc6NTQxMjkzMS4yNMaDMjg1xYjFpMWmxajFqsWSxIDFmMWraW9uxJcyxKxzxaRnbGXFhsSsxZZtZXPEm2nGtsayxK3Ev8STxJ9JxaXFp8WpxINNxYPGsToixq1uxq%2FGscWH

3. Methodology

In this section, we will discuss the steps used to identify the best location for a restaurant based on the data collected.

We will first conduct a basic descriptive analysis of the datasets collected to get a brief idea of the dataset (collected from https://github.com/courseraassignmentcapstone/Coursera_Capstone/blob/main/toronto_data.csv) characteristics related to the project.

This project aims to conduct a suitability analysis for a new restaurant to open in Toronto neighborhoods. We will limit our analysis to the top 100 venues that are in Toronto within a radius of ~8km. Next, we will run cluster analysis to identify the best place to open a restaurant. This analysis will plot the points and clusters in maps, generate bar plots, and conduct basic descriptive analysis to suggest the best neighborhoods to open a new restaurant. The stakeholders can then utilize this information to open a restaurant that produces a good income.

For this task, we need to fist extract the data.
Using https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050 we first create the following data frame.

	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
6	M6A	North York	Lawrence Manor

Next, we import the geographical data of the postal codes from https://raw.githubusercontent.com/courseraassignmentcapstone/Coursera_Capstone/main/Geospatial_Coordinates.csv and create the following data frame.

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Then, we import the Toronto project related data (on Income, Competition, and Population) from https://raw.githubusercontent.com/courseraassignmentcapstone/Coursera_Capstone/main/toronto_data.csv , and generate the following data frame.

	Neighbourhood	Competition	Population	Income
0	West Humber-Clairville	2550	32265	63415
1	Mount Olive-Silverstone-Jamestown	273	32130	48145
2	Thistletown-Beaumond Heights	236	9925	55030
3	Rexdale-Kipling	155	10725	52430
4	Elms-Old Rexdale	70	9440	53780

To obtain the main dataset for the analysis, we add the geographical information to the above data fame using the previous data frames and create the following data frame.

	Neighbourhood	Latitude	Longitude	Competition	Population	Income
0	Victoria Village	43.725882	-79.315572	398.0	17050.0	42315.0
1	Rouge	43.806686	-79.194353	832.0	43180.0	76945.0
2	Malvern	43.806686	-79.194353	665.0	44315.0	56610.0
3	Highland Creek	43.784535	-79.160497	91.0	12845.0	86850.0
4	Flemingdon Park	43.725900	-79.340923	220.0	21290.0	41280.0

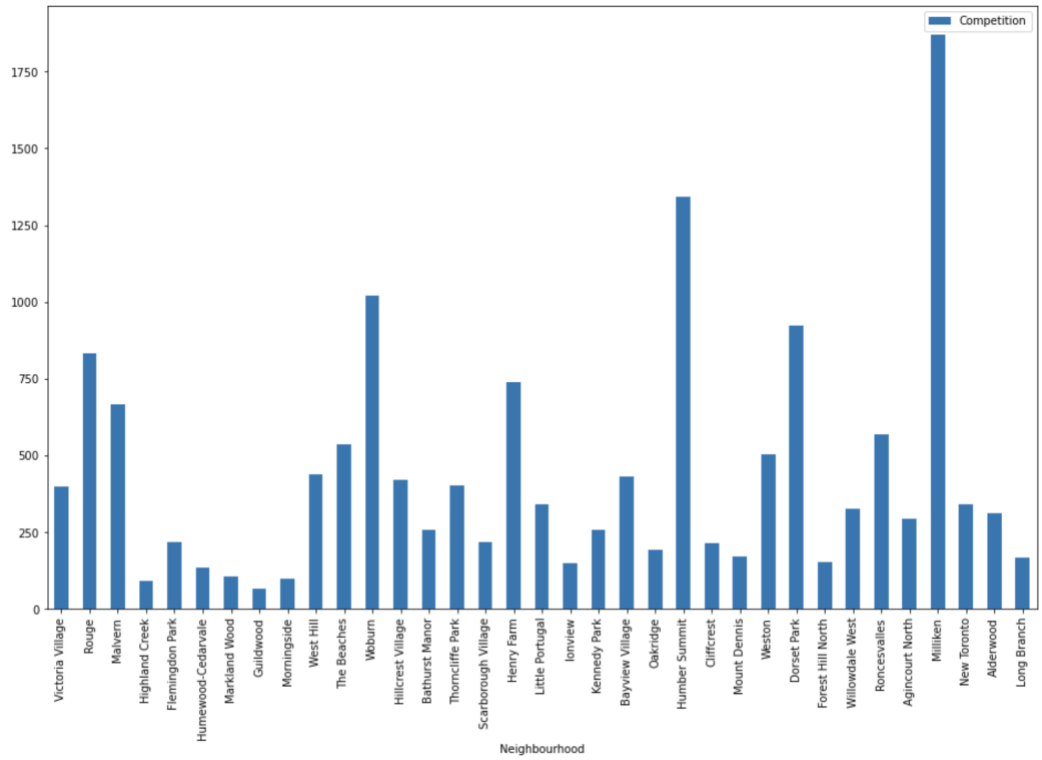
To obtain the restaurant data around the neighbourhoods of interest, we need to use the Foursquare API and extract information. We restrict the analysis to the top 100 venues in Toronto within a radius of 8km. The data frame that is produced with Foursquare data is as follows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Victoria Village	43.725882	-79.315572	Aga Khan Museum	43.725105	-79.332076	History Museum
1	Victoria Village	43.725882	-79.315572	Naan & Kabob Halal	43.742903	-79.305148	Middle Eastern Restaurant
2	Victoria Village	43.725882	-79.315572	Toronto Climbing Academy	43.709362	-79.315006	Rock Climbing Spot
3	Victoria Village	43.725882	-79.315572	Adonis	43.729188	-79.290391	Grocery Store
4	Victoria Village	43.725882	-79.315572	Ghadir Mid-Eastern Grocery	43.743638	-79.304233	Grocery Store

4. Analysis, Results and Discussion

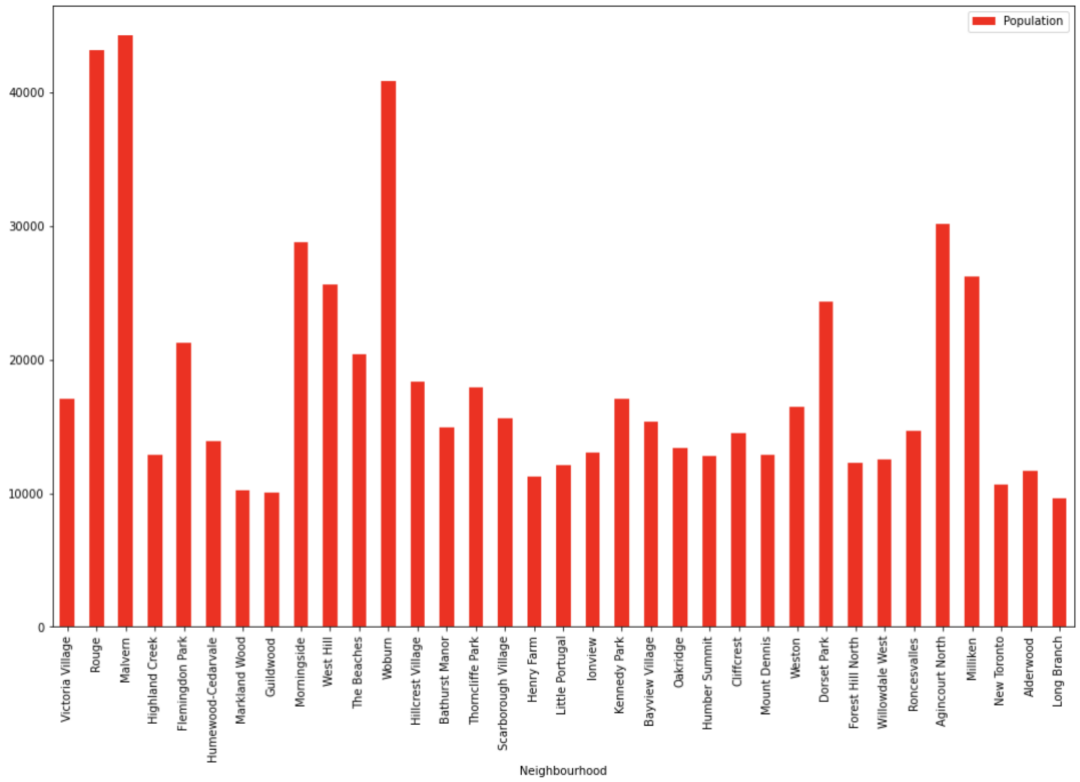
In this section, we will run the analysis and generate results for the different experiments. The results are then discussed accordingly.

Let's first evaluate the competition of the neighbourhoods by plotting a bar plot on the "Competition" data.



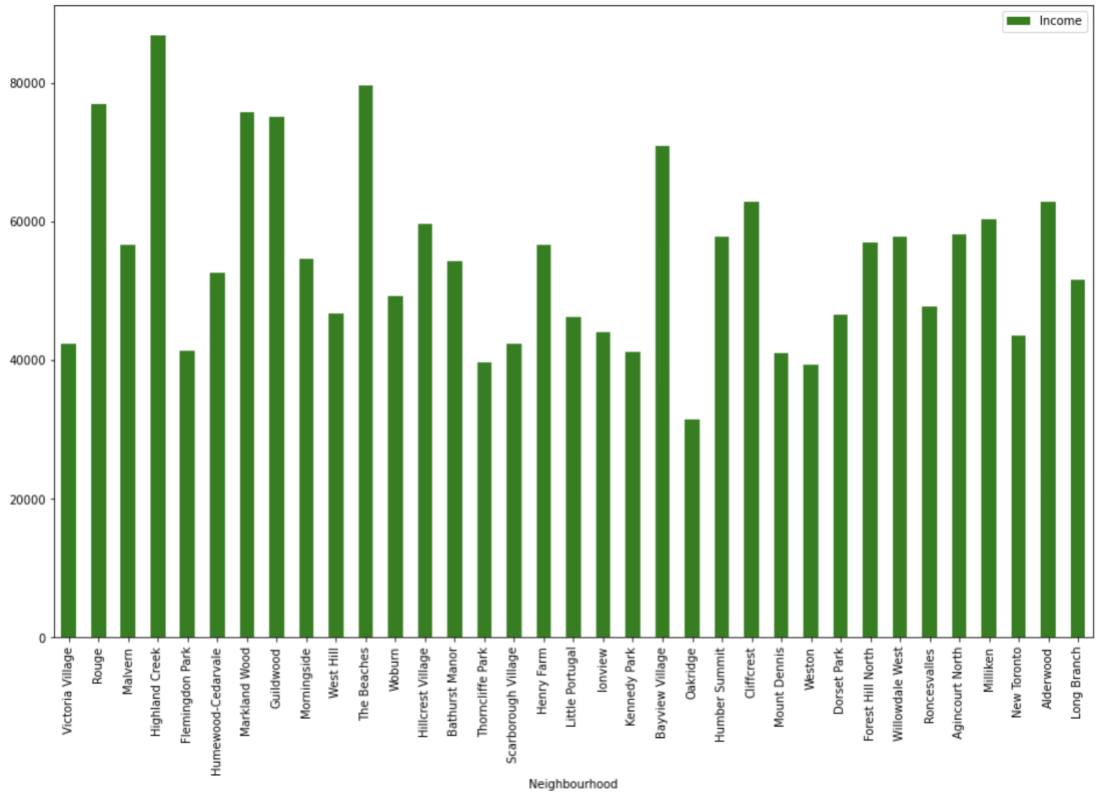
Milliken shows the highest competition for a new business as shown in the bar chart and the above record. This suggests that it is challenging to open a new business in this area.

Let's evaluate the population statistics of the neighbourhoods by plotting a bar plot on the "Population" data.



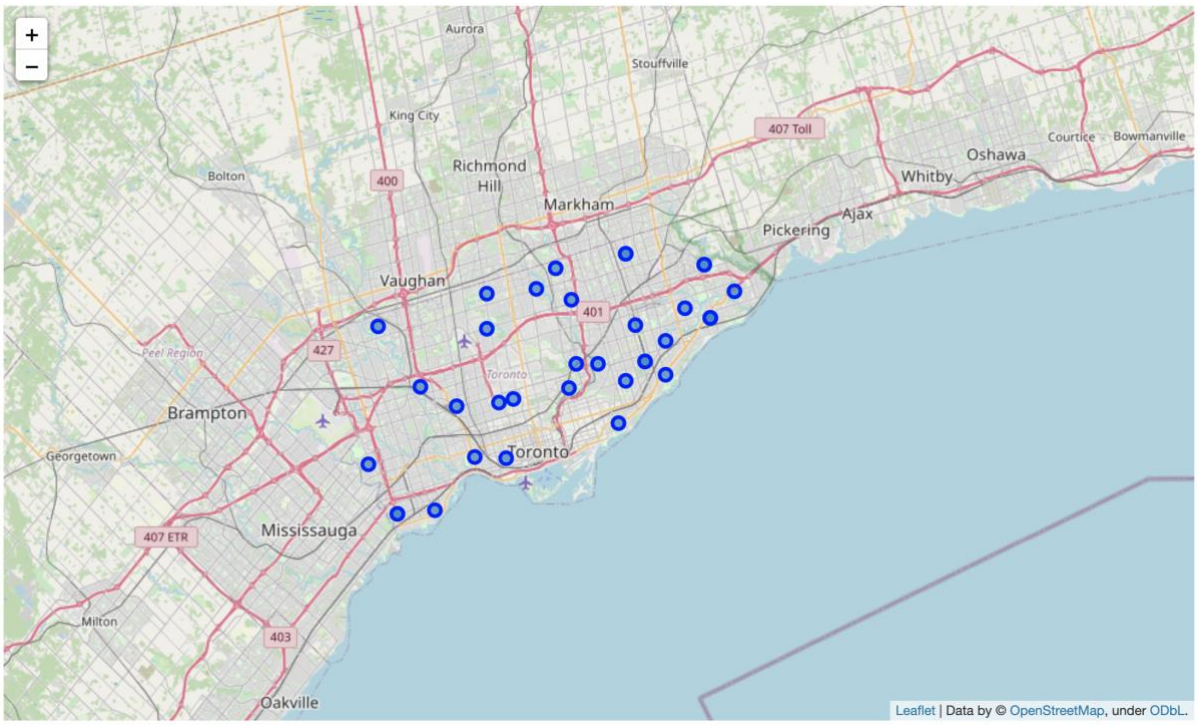
Malvern has the highest population. This means that a restaurant in Malvern may attract many customers.

Let's evaluate the income statistics of the neighbourhoods by plotting a bar plot on the "Income" data.



As shown in the bar graph and the above record, Highland Creek is the area with the highest income. This may have an impact on people to buy from restaurants regularly.

Let's plot the data in a map using the Folium API



Analysing the neighbourhoods

For this task, we need to first create a one hot encoded representation of the dataset, which generates the following data frame.

	Zoo Exhibit	African Restaurant	Airport	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	BBQ Joint	...	Train Station	Turkish Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Warehous Sto
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	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

Next, let's group rows by neighbourhoods and by taking the mean of the frequency of occurrence of each category to obtain the following data frame.

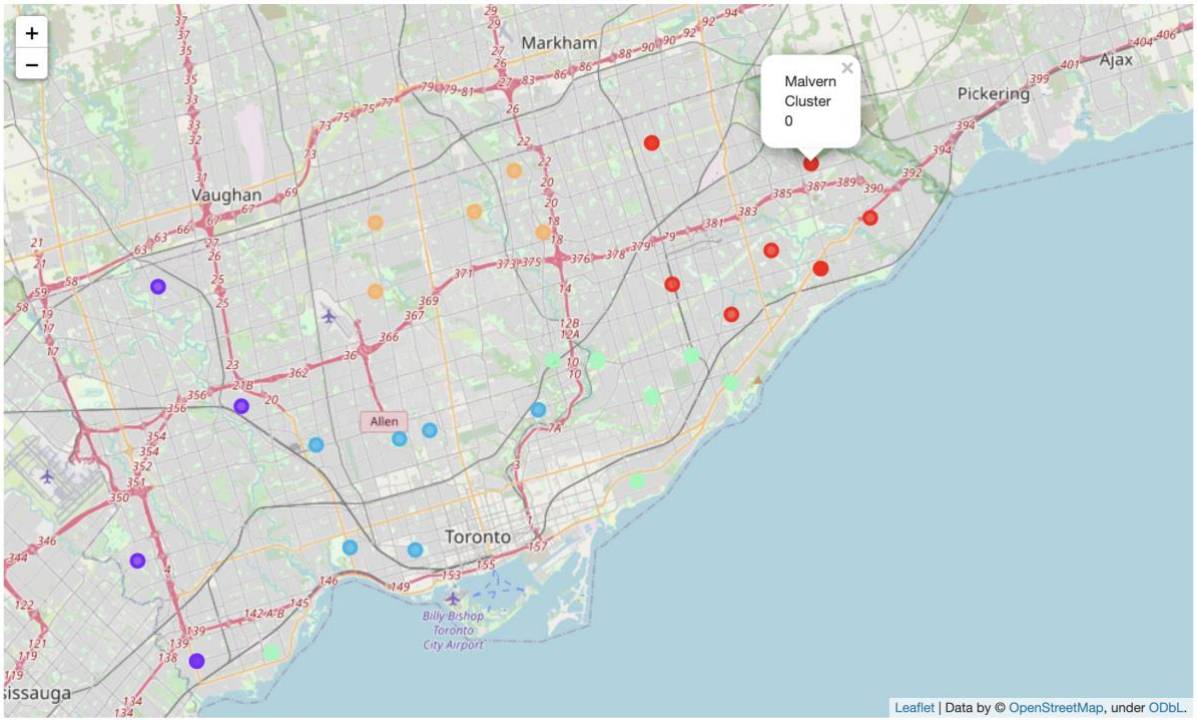
	Neighborhood	Zoo Exhibit	African Restaurant	Airport	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	...	Train Station	Turkish Restaurant	Vegetarian / Vegan Restaurant	Vietna Resta
0	Agincourt North	0.073684	0.00	0.00	0.000000	0.000000	0.010526	0.021053	0.010526	0.00	...	0.00	0.00	0.00	0.0
1	Alderwood	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.00	0.00	0.00	0.0
2	Bathurst Manor	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.010000	0.01	...	0.00	0.00	0.00	0.0
3	Bayview Village	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.01	...	0.00	0.00	0.01	0.0
4	Cliffcrest	0.000000	0.00	0.00	0.000000	0.000000	0.010000	0.010000	0.000000	0.00	...	0.00	0.01	0.02	0.0
5	Dorset Park	0.000000	0.00	0.00	0.000000	0.000000	0.010000	0.020000	0.010000	0.00	...	0.00	0.01	0.00	0.0

Next, we create the following data frame using the above information and considering the top venues

	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 North	Zoo Exhibit	Caribbean Restaurant	Coffee Shop	Supermarket	Bakery	Grocery Store	Restaurant	Burger Joint	Bubble Tea Shop	Chinese Restaurant
1	Alderwood	Park	Bakery	Coffee Shop	Café	Seafood Restaurant	Burger Joint	Grocery Store	Breakfast Spot	Ice Cream Shop	Burrito Place
2	Bathurst Manor	Coffee Shop	Café	Park	Grocery Store	Italian Restaurant	Korean Restaurant	Japanese Restaurant	Supermarket	Middle Eastern Restaurant	Bakery
3	Bayview Village	Coffee Shop	Supermarket	Middle Eastern Restaurant	Bakery	Café	Caribbean Restaurant	Japanese Restaurant	Sushi Restaurant	Korean Restaurant	Park
4	Cliffcrest	Park	Middle Eastern Restaurant	Coffee Shop	Beach	Burger Joint	Bakery	Breakfast Spot	Pub	Indian Restaurant	Grocery Store

Clustering neighbourhoods

Let's run k-means to cluster the neighbourhoods into 5 clusters. The clustering outputs can be plotted in a map as shown below.



We can check the records under each cluster as follows,

Cluster 1 data:

	Latitude	Income	Cluster Labels	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
1	43.806686	76945.0	0	Zoo Exhibit	Park	Indian Restaurant	Caribbean Restaurant	Coffee Shop	Pub	Hakka Restaurant	Supermarket	Burger Joint	Breakfast Spot
2	43.806686	56610.0	0	Zoo Exhibit	Park	Indian Restaurant	Caribbean Restaurant	Coffee Shop	Pub	Hakka Restaurant	Supermarket	Burger Joint	Breakfast Spot
3	43.784535	86850.0	0	Zoo Exhibit	Park	Ice Cream Shop	Breakfast Spot	Coffee Shop	Pub	Pharmacy	Mexican Restaurant	Indian Restaurant	Restaurant
7	43.763573	75025.0	0	Park	Zoo Exhibit	Coffee Shop	Caribbean Restaurant	Pharmacy	Japanese Restaurant	Indian Restaurant	Grocery Store	Restaurant	Pub
8	43.763573	54590.0	0	Park	Zoo Exhibit	Coffee Shop	Caribbean Restaurant	Pharmacy	Japanese Restaurant	Indian Restaurant	Grocery Store	Restaurant	Pub
9	43.763573	46740.0	0	Park	Zoo Exhibit	Coffee Shop	Caribbean Restaurant	Pharmacy	Japanese Restaurant	Indian Restaurant	Grocery Store	Restaurant	Pub
11	43.770992	49305.0	0	Caribbean Restaurant	Park	Indian Restaurant	Zoo Exhibit	Coffee Shop	Burger Joint	Middle Eastern Restaurant	Liquor Store	Italian Restaurant	Sandwich Place
15	43.744734	42415.0	0	Park	Caribbean Restaurant	Middle Eastern Restaurant	Indian Restaurant	Coffee Shop	Burger Joint	Restaurant	Grocery Store	Beach	Bookstore
26	43.757410	46485.0	0	Caribbean Restaurant	Park	Coffee Shop	Middle Eastern Restaurant	Japanese Restaurant	Italian Restaurant	Indian Restaurant	Burger Joint	Grocery Store	Sandwich Place
30	43.815252	58095.0	0	Zoo Exhibit	Caribbean Restaurant	Coffee Shop	Supermarket	Bakery	Grocery Store	Restaurant	Burger Joint	Bubble Tea Shop	Chinese Restaurant
31	43.815252	60350.0	0	Zoo Exhibit	Caribbean Restaurant	Coffee Shop	Supermarket	Bakery	Grocery Store	Restaurant	Burger Joint	Bubble Tea Shop	Chinese Restaurant

Cluster 2 data:

	Latitude	Income	Cluster Labels	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
6	43.643515	75700.0	1	Burrito Place	Grocery Store	Burger Joint	Bakery	Coffee Shop	Restaurant	Brewery	Seafood Restaurant	Hotel	Sandwich Place
22	43.756303	57810.0	1	Steakhouse	Coffee Shop	Italian Restaurant	Hotel	Sporting Goods Shop	Bakery	Ice Cream Shop	Asian Restaurant	Gym	Burger Joint
25	43.706876	39300.0	1	Coffee Shop	Italian Restaurant	Café	Liquor Store	Brewery	Bar	Bakery	Asian Restaurant	Grocery Store	Park
33	43.602414	62840.0	1	Park	Bakery	Coffee Shop	Café	Seafood Restaurant	Burger Joint	Grocery Store	Breakfast Spot	Ice Cream Shop	Burrito Place
34	43.602414	51505.0	1	Park	Bakery	Coffee Shop	Café	Seafood Restaurant	Burger Joint	Grocery Store	Breakfast Spot	Ice Cream Shop	Burrito Place

Cluster 3 data:

	Latitude	Income	Cluster Labels	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
5	43.693781	52590.0	2	Café	Park	Coffee Shop	Bakery	Pizza Place	Grocery Store	Bar	Furniture / Home Store	French Restaurant	Liquor Store
14	43.705369	39580.0	2	Park	Coffee Shop	Café	Brewery	Italian Restaurant	Bakery	Concert Hall	Ice Cream Shop	BBQ Joint	Grocery Store
17	43.647927	46195.0	2	Park	Café	Coffee Shop	Bakery	Sandwich Place	Pizza Place	Bar	Dessert Shop	Farmers Market	Plaza
24	43.691116	41060.0	2	Park	Café	Bakery	Coffee Shop	Pizza Place	Sandwich Place	Bar	Ice Cream Shop	Liquor Store	Dog Run
27	43.696948	56970.0	2	Café	Park	Coffee Shop	Bakery	Brewery	Japanese Restaurant	Grocery Store	Pizza Place	French Restaurant	Sandwich Place
29	43.648960	47685.0	2	Park	Café	Bakery	Coffee Shop	Ice Cream Shop	Pizza Place	Italian Restaurant	Indian Restaurant	Sandwich Place	Scenic Lookout

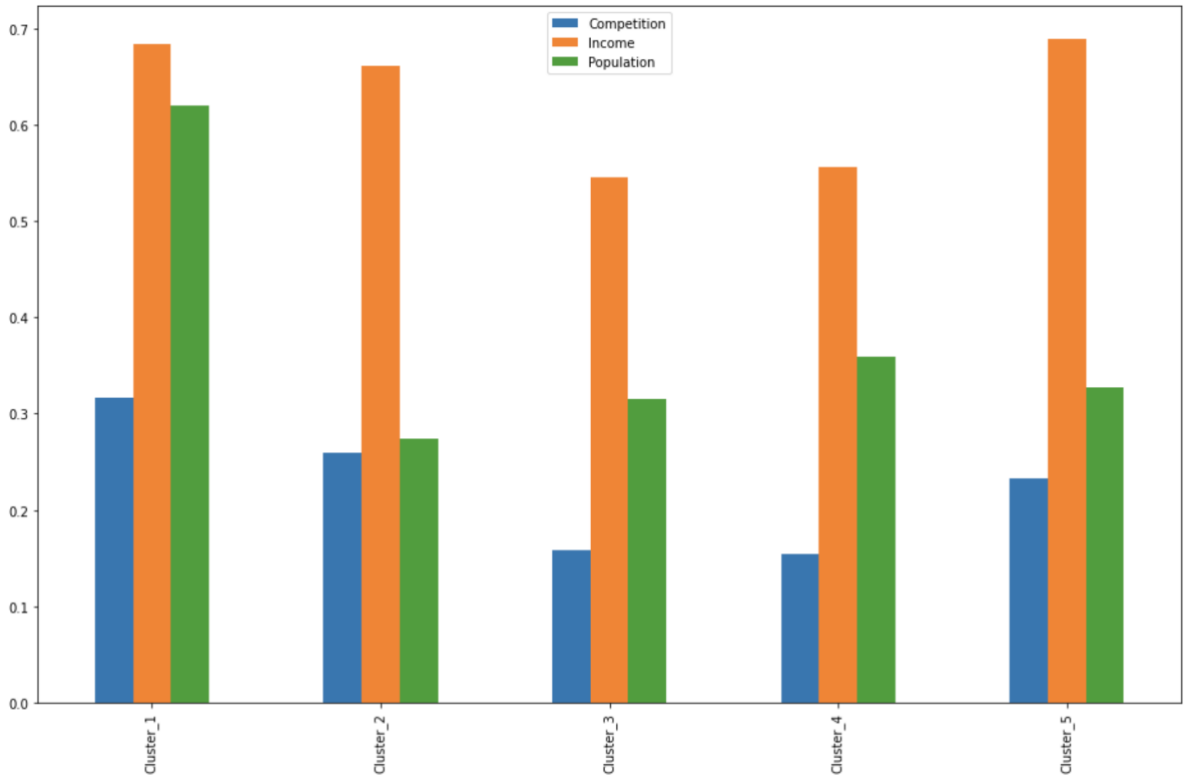
Cluster 4 data:

	Latitude	Income	Cluster Labels	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	43.725882	42315.0	3	Park	Café	Grocery Store	Bakery	Beach	Brewery	Sandwich Place	Coffee Shop	Pizza Place	Greek Restaurant
4	43.725900	41280.0	3	Park	Café	Bakery	Grocery Store	Middle Eastern Restaurant	Supermarket	Italian Restaurant	Coffee Shop	Brewery	Turkish Restaurant
10	43.676357	79665.0	3	Coffee Shop	Park	Grocery Store	Beach	Farmers Market	Brewery	Historic Site	Liquor Store	Bakery	Other Great Outdoors
18	43.727929	44090.0	3	Park	Coffee Shop	Beach	Burger Joint	Caribbean Restaurant	Bakery	Middle Eastern Restaurant	Grocery Store	Café	Burrito Place
19	43.727929	41130.0	3	Park	Coffee Shop	Beach	Burger Joint	Caribbean Restaurant	Bakery	Middle Eastern Restaurant	Grocery Store	Café	Burrito Place
21	43.711112	31365.0	3	Park	Beach	Coffee Shop	Brewery	Bakery	Middle Eastern Restaurant	Café	BBQ Joint	Pizza Place	Sandwich Place
23	43.716316	62770.0	3	Park	Middle Eastern Restaurant	Coffee Shop	Beach	Burger Joint	Bakery	Breakfast Spot	Pub	Indian Restaurant	Grocery Store
32	43.605647	43440.0	3	Park	Coffee Shop	Café	Bakery	Ice Cream Shop	Italian Restaurant	Seafood Restaurant	Scenic Lookout	Restaurant	Burger Joint

Cluster 5 data:

	Latitude	Income	Cluster Labels	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
12	43.803762	59610.0	4	Chinese Restaurant	Bakery	Japanese Restaurant	Caribbean Restaurant	Hotel	Supermarket	Middle Eastern Restaurant	Greek Restaurant	Bubble Tea Shop	Coffee Shop
13	43.754328	54315.0	4	Coffee Shop	Café	Park	Grocery Store	Italian Restaurant	Korean Restaurant	Japanese Restaurant	Supermarket	Middle Eastern Restaurant	Bakery
16	43.778517	56645.0	4	Supermarket	Bakery	Caribbean Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Grocery Store	Shopping Mall	Japanese Restaurant	Coffee Shop	Fish Market
20	43.786947	70855.0	4	Coffee Shop	Supermarket	Middle Eastern Restaurant	Bakery	Café	Caribbean Restaurant	Japanese Restaurant	Sushi Restaurant	Korean Restaurant	Park
28	43.782736	57735.0	4	Coffee Shop	Café	Middle Eastern Restaurant	Grocery Store	Supermarket	Bagel Shop	Park	Liquor Store	Korean Restaurant	Furniture / Home Store

Let's investigate how the income varies in each cluster using line plots for each cluster



In this section, we discuss the results. According to our analysis, the first cluster (numbered as cluster 0 on the map) provides the best place to start a new restaurant. According to the bar chart, it can be seen that there is a high population and income although the competition is a little bit high. However, the competition compared to other clusters is not drastically high in the first cluster. As it is visible, the competition is close to that of clusters 2 and 5. Hence, we suggest that the area around the Malvern (Malvern cluster 0 as depicted in the map) is the best place to open a restaurant.

5. Conclusion

In this project, we tried to identify a suitable place for a new restaurant based on three factors; competition, population, and income. We first extracted data on these statistics, and then we merged the geographical data to the corresponding data to conduct geographical analysis based on clustering. In our analysis, we showed the dynamics of different places in Toronto. Then we also suggested a certain area that shows the best performance in terms of competition, population, and income. However, the final decision to open a restaurant in a certain place will be taken by the stakeholders. The stakeholders can effectively use the insights produced in the project to take this decision effectively.