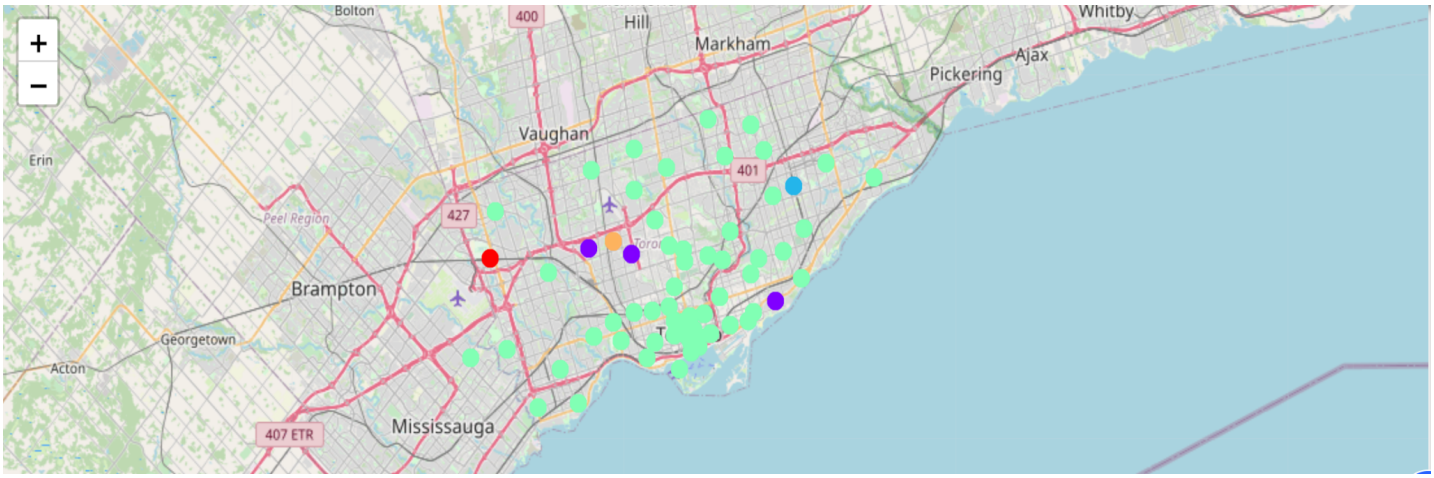
The Battle of Neighborhoods - Toronto



Applied Data Science Capstone by IBM on Coursera - Neha Gupta

1. INTRODUCTION: BUSINESS PROBLEM

This project discusses the Neighborhoods in **Toronto** and what will be a suitable place to start a new business such as **Restaurants, Hotels**, etc.

Toronto is one of the most densely populated areas in **Canada**. Being the land of opportunity, it brings in a variety of people from different ethnic backgrounds to the core city of Canada, Toronto. Being the largest city in Canada with an estimated population of over 6 million, there is no doubt about the diversity of the population. Multiculturalism is seen through the various neighbourhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown and many more. Downtown Toronto being the hub of interactions between ethnicities brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.

The **Foursquare API** is used to access the venues in the neighborhoods. Since it returns less venues in the neighborhoods, we would be analyzing areas for which a countable number of venues are obtained. Then they are clustered based on their venues using Data Science Techniques. Here the **k-means** clustering algorithm is used to achieve the task. The optimal number of clusters can be obtained using silhouette score metrics. **Folium** visualization library can be used to visualize the clusters superimposed on the map of **Toronto** city.

This project is aimed towards **Entrepreneurs or Business** owners who want to start a new **Restaurant or Hotel** or grow their current business. The analysis will provide vital information that can be used by the target audience.

2. DATA REQUIREMENTS

The data required for the neighbourhoods in Toronto will be collected and prepared from Wikipedia using different techniques. The Geographical location of the neighbourhoods will be used from the Geocoder package. The Venue data will be collected from Foursquare which will be used to find a suitable place to start the business.

1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The Wikipedia site shown above gives almost all the information about the neighbourhoods. It includes the postal code, borough and the name of the neighbourhoods present in Toronto. Since the data is not in a format that is suitable for analysis, scraping of the data is done from this site.

	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

2. https://cocl.us/Geospatial_data

The second source of data provided us with the Geographical coordinates of the neighbourhoods with the respective Postal Codes. The file was in CSV format, so it is attached it to a Pandas 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

3. https://foursquare.com/

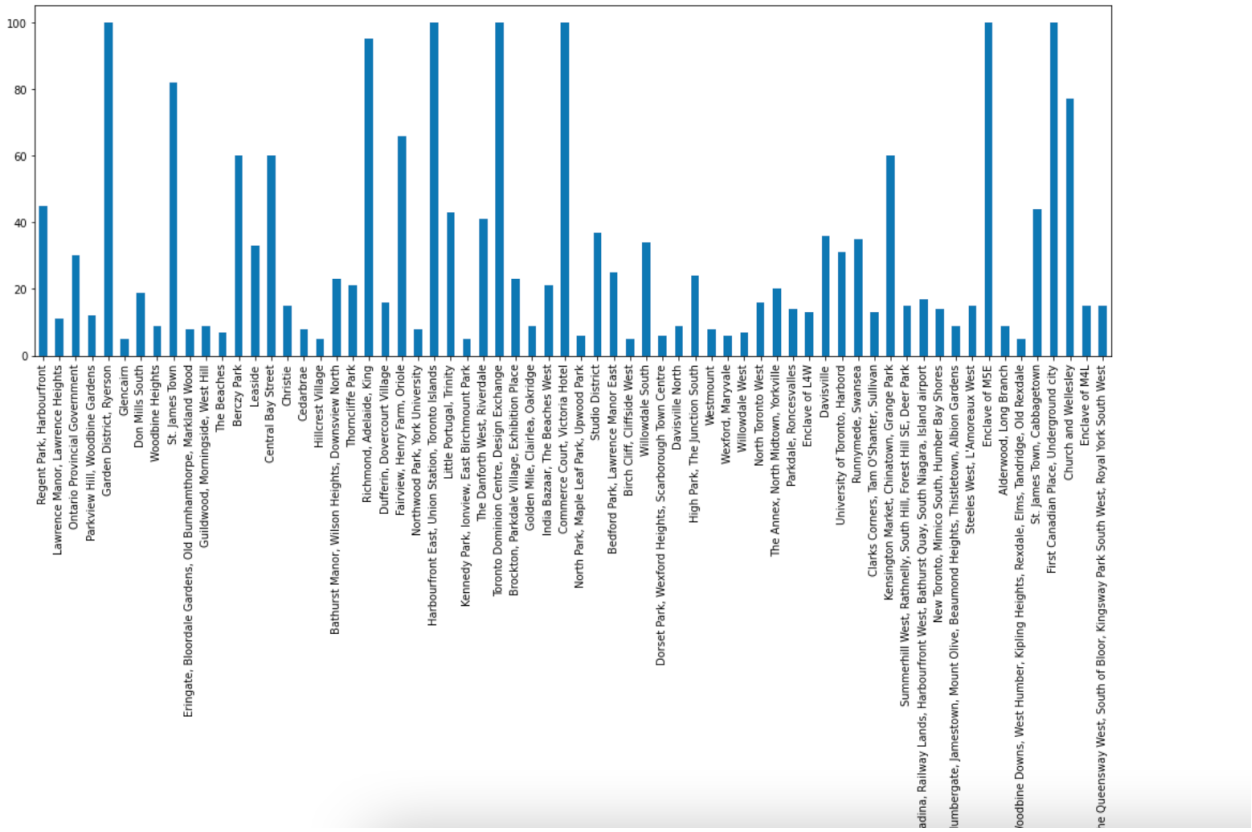
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

3. METHODOLOGY

Now, we have the neighborhoods data of Toronto (**103 neighborhoods**). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of **2130** venues have been obtained in the whole city and **272** unique categories. But as seen we have multiple neighborhoods with less than 5 venues returned. In order to create a good analysis let's consider only the neighborhoods with more than 5 venues. We can perform one hot encoding on the obtained data set and use it to find the 5 most common venue categories in each neighborhood. Then clustering can be performed on the dataset. Here K - Nearest Neighbor clustering technique has been used. To find the optimal number of clusters silhouette score metric technique is used. The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

4. Analysis

Looking into the dataset we found that there were many neighborhoods with less than 5 venues which can be removed before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 5 or more than 5 venues were obtained.



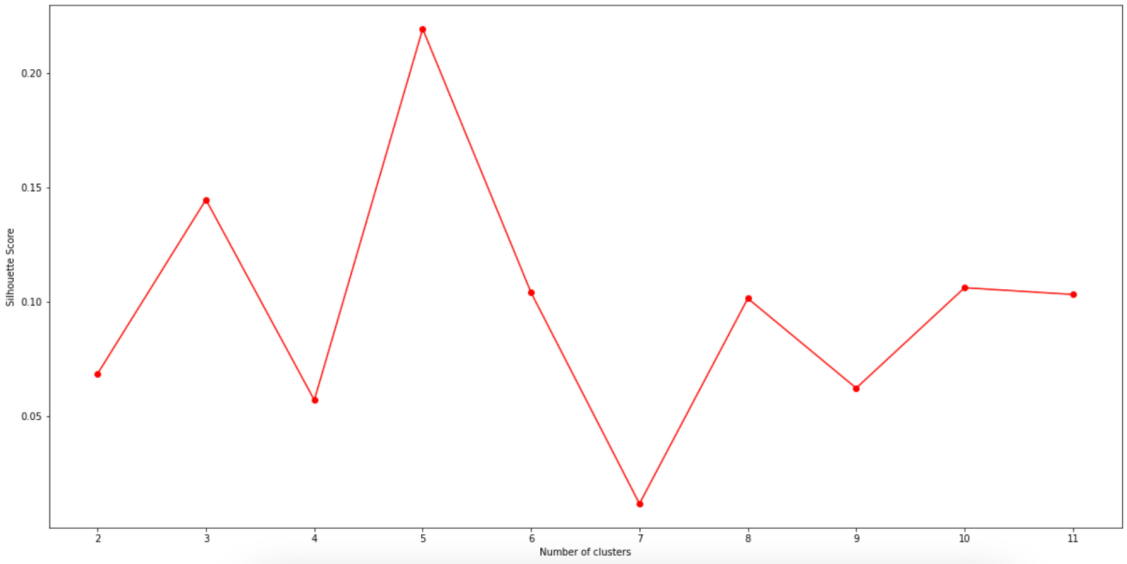
Next, we will perform one hot encoding on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. A sample output is shown in below fig.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Muse
0	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.000000	0.0
1	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.000000	0.0
2	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.000000	0.0
3	Berczy Park	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.016667	0.0
4	Birch Cliff, Cliffside West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.000000	0.0

The above dataset is used to obtain the top 10 most common venues in each neighborhood i.e. the 10 venues with the highest mean of frequency of occurrence. A sample for the first 5 neighborhoods is shown below.

	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	Alderwood, Long Branch	Pizza Place	Pharmacy	Pub	Sandwich Place	Athletics & Sports	Coffee Shop	Playground	Gym	Medical Center	Mediterranean Restaurant
1	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Bridal Shop	Sandwich Place	Frozen Yogurt Shop	Sushi Restaurant	Chinese Restaurant	Restaurant	Gas Station	Diner
2	Bedford Park, Lawrence Manor East	Coffee Shop	Italian Restaurant	Sandwich Place	Restaurant	Pizza Place	Café	Sushi Restaurant	Juice Bar	Pharmacy	Liquor Store
3	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Farmers Market	Seafood Restaurant	Pharmacy	Cheese Shop	Beer Bar	Restaurant	Comfort Food Restaurant
4	Birch Cliff, Cliffside West	College Stadium	Skating Rink	Café	General Entertainment	Farm	Mediterranean Restaurant	Medical Center	Men's Store	Metro Station	Monument / Landmark

This dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters. For optimal result we need to select the best value for K. Here, the silhouette score is used to find the best value for K. A range of values from 2 to 10 was considered, KNN clustering was performed on the dataset and the silhouette score was calculated and plotted on a line plot as shown below. From the plot we can see that a K value of 5 provides the best score. This K value is used for the K-Means Clustering Technique.



The K-Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5 C
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant	3	Coffee Shop	Bakery	Park	Breakfast Spot	C

5. Results

Let’s examine the 5 clusters and find the discriminating venue categories that distinguish each cluster. For this purpose, let's also look into the five most common venue categories in each cluster.

Cluster 1

The top venue categories in Cluster 1 are Garden Center, Truck Stop, Bar, Rental Car Location, Drugstore, Medical Center, Restaurants, etc.

	Neighborhood	Venue Latitude	Venue Longitude	Venue Category	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 Cor
1773	Clairville, Humberwood, Woodbine Downs, West H...	43.708471	-79.589943	Rental Car Location	0	Garden Center	Truck Stop	Bar	Rental Car Location	Drugstore	Medical Center	Mediterranean Restaurant	Mer Stor
1774	Clairville, Humberwood, Woodbine Downs, West H...	43.707554	-79.589252	Bar	0	Garden Center	Truck Stop	Bar	Rental Car Location	Drugstore	Medical Center	Mediterranean Restaurant	Mer Stor
1775	Clairville, Humberwood, Woodbine Downs, West H...	43.705072	-79.598725	Drugstore	0	Garden Center	Truck Stop	Bar	Rental Car Location	Drugstore	Medical Center	Mediterranean Restaurant	Mer Stor
1776	Clairville, Humberwood, Woodbine Downs, West H...	43.706539	-79.599359	Garden Center	0	Garden Center	Truck Stop	Bar	Rental Car Location	Drugstore	Medical Center	Mediterranean Restaurant	Mer Stor
1777	Clairville, Humberwood, Woodbine Downs, West H...	43.704891	-79.599410	Truck Stop	0	Garden Center	Truck Stop	Bar	Rental Car Location	Drugstore	Medical Center	Mediterranean Restaurant	Mer Stor

Cluster 2

The top venue categories in Cluster 2 are Bakery, Pizza place, Parks and Restaurants.

	Neighborhood	Venue Latitude	Venue Longitude	Venue Category	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 M Comn Ve
198	Glencairn	43.707420	-79.443126	Bakery	1	Bakery	Pizza Place	Park	Japanese Restaurant	Asian Restaurant	Mexican Restaurant	Modern Europea Restaura
199	Glencairn	43.709111	-79.443930	Japanese Restaurant	1	Bakery	Pizza Place	Park	Japanese Restaurant	Asian Restaurant	Mexican Restaurant	Modern Europea Restaura
200	Glencairn	43.708828	-79.443366	Asian Restaurant	1	Bakery	Pizza Place	Park	Japanese Restaurant	Asian Restaurant	Mexican Restaurant	Modern Europea Restaura
201	Glencairn	43.707170	-79.442658	Pizza Place	1	Bakery	Pizza Place	Park	Japanese Restaurant	Asian Restaurant	Mexican Restaurant	Modern Europea Restaura
202	Glencairn	43.713550	-79.442482	Park	1	Bakery	Pizza Place	Park	Japanese Restaurant	Asian Restaurant	Mexican Restaurant	Modern Europea Restaura
330	The Beaches	43.676821	-79.293942	Trail	1	Pub	Trail	Health Food Store	Park	Coffee Shop	Asian Restaurant	Modern Europea Restaura
331	The Beaches	43.678879	-79.297734	Health Food Store	1	Pub	Trail	Health Food	Park	Coffee Shop	Asian Restaurant	Modern Europea

Cluster 3

The top venue categories in Cluster 3 are Chinese Restaurant, Indian Restaurant, Pet Store, Vietnamese Restaurant, Gaming cafe, etc.

	Neighborhood	Venue Latitude	Venue Longitude	Venue Category	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
1296	Dorset Park, Wexford Heights, Scarborough Town...	43.753833	-79.276611	Chinese Restaurant	2	Indian Restaurant	Pet Store	Vietnamese Restaurant	Chinese Restaurant	Gaming Cafe	Mediterranean Restaurant	Men's Store
1297	Dorset Park, Wexford Heights, Scarborough Town...	43.754915	-79.276945	Indian Restaurant	2	Indian Restaurant	Pet Store	Vietnamese Restaurant	Chinese Restaurant	Gaming Cafe	Mediterranean Restaurant	Men's Store
1298	Dorset Park, Wexford Heights, Scarborough Town...	43.756042	-79.276276	Indian Restaurant	2	Indian Restaurant	Pet Store	Vietnamese Restaurant	Chinese Restaurant	Gaming Cafe	Mediterranean Restaurant	Men's Store
1299	Dorset Park, Wexford Heights, Scarborough Town...	43.757770	-79.278572	Vietnamese Restaurant	2	Indian Restaurant	Pet Store	Vietnamese Restaurant	Chinese Restaurant	Gaming Cafe	Mediterranean Restaurant	Men's Store
1300	Dorset Park, Wexford Heights,	43.759279	-79.278325	Pet Store	2	Indian Restaurant	Pet Store	Vietnamese Restaurant	Chinese Restaurant	Gaming Cafe	Mediterranean Restaurant	Men's Store

Cluster 4

The top venue categories in Cluster 4 are Coffee Shop, Bakery, Park, Breakfast Spot and Cafe.

	Neighborhood	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	
0	Regent Park, Harbourfront	43.653447	-79.362017	Bakery	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
1	Regent Park, Harbourfront	43.653559	-79.361809	Coffee Shop	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
2	Regent Park, Harbourfront	43.653249	-79.358008	Distribution Center	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
3	Regent Park, Harbourfront	43.654735	-79.359874	Spa	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
4	Regent Park, Harbourfront	43.656369	-79.356980	Restaurant	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
5	Regent Park, Harbourfront	43.655618	-79.356211	Park	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
6	Regent Park, Harbourfront	43.653947	-79.361149	Breakfast Spot	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
7	Regent Park, Harbourfront	43.653313	-79.359725	Gym / Fitness Center	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
8	Regent Park, Harbourfront	43.650244	-79.359323	Historic Site	3	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Thea
	Regent Park,			Chocolate					Breakfast		

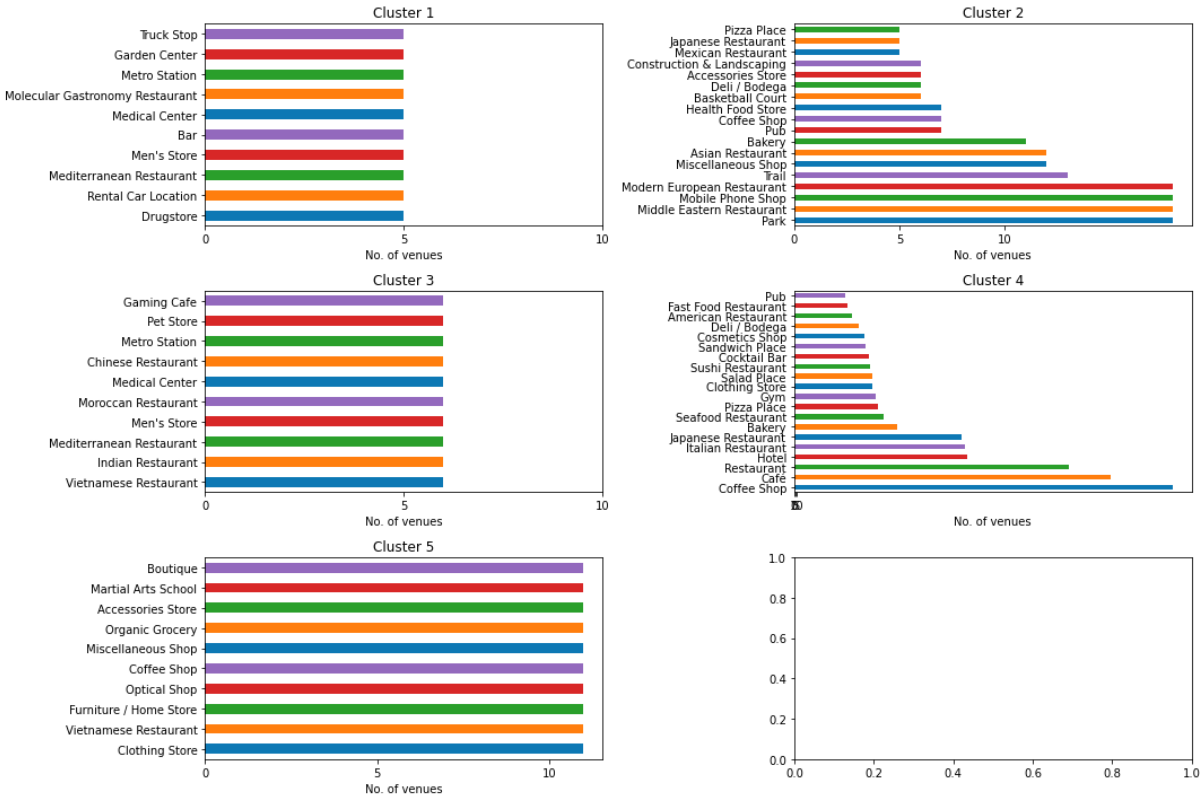
Cluster 5

The top venue categories in Cluster 5 are Clothing Store, Furniture Store, Accessories Store, Restaurant, Miscellaneous Shop, Coffee Shop and boutique.

	Neighborhood	Venue Latitude	Venue Longitude	Venue Category	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
45	Lawrence Manor, Lawrence Heights	43.718214	-79.463893	Boutique	4	Clothing Store	Furniture / Home Store	Accessories Store	Vietnamese Restaurant	Miscellaneous Shop	Coffee Shop	Boutique
46	Lawrence Manor, Lawrence Heights	43.719096	-79.462675	Furniture / Home Store	4	Clothing Store	Furniture / Home Store	Accessories Store	Vietnamese Restaurant	Miscellaneous Shop	Coffee Shop	Boutique
47	Lawrence Manor, Lawrence Heights	43.721259	-79.468472	Vietnamese Restaurant	4	Clothing Store	Furniture / Home Store	Accessories Store	Vietnamese Restaurant	Miscellaneous Shop	Coffee Shop	Boutique
48	Lawrence Manor, Lawrence Heights	43.719045	-79.460849	Clothing Store	4	Clothing Store	Furniture / Home Store	Accessories Store	Vietnamese Restaurant	Miscellaneous Shop	Coffee Shop	Boutique
49	Lawrence Manor, Lawrence Heights	43.719427	-79.467995	Coffee Shop	4	Clothing Store	Furniture / Home Store	Accessories Store	Vietnamese Restaurant	Miscellaneous Shop	Coffee Shop	Boutique
50	Lawrence Manor, Lawrence	43.718892	-79.461344	Accessories Store	4	Clothing Store	Furniture / Home	Accessories Store	Vietnamese Restaurant	Miscellaneous Shop	Coffee Shop	Boutique

6. Discussion

Now that we have the clusters and the top venue categories, let's visualize the top 20 venue categories in each Cluster for comparison.



This plot can be used to suggest valuable information to Business persons. Let's discuss a few examples considering they would like to start the following category of business.

1. **Restaurant**

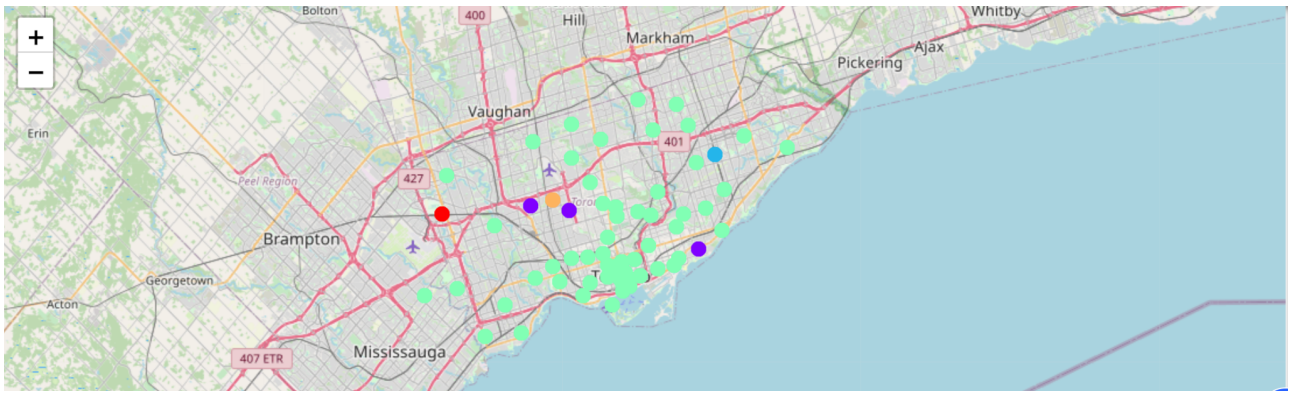
The neighborhood in cluster 2 has the greatest number of restaurants and also it has different kinds of restaurants such as American, Sushi, Seafood,etc. Hence opening one here is not the best choice. One can think of opening a specific type of restaurant in Cluster 1 or 3 since there are only a few specific types of places there. Cluster 5 is also a good choice for a Hotel or restaurants since there are less number of places there. Other factors such as places to be explored in the vicinity by the customers can also be considered by looking at the venues in the plot.

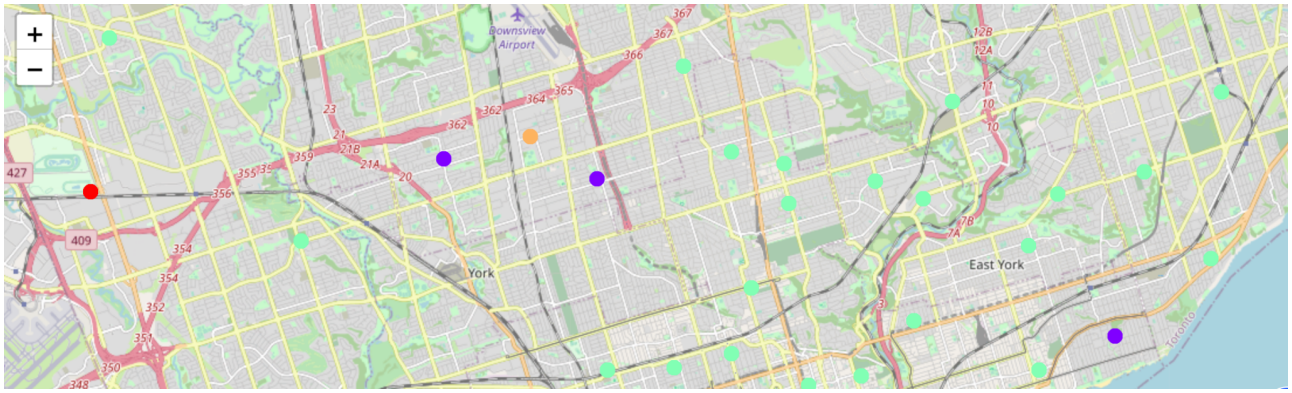
2. **Medical Centre**

The neighborhoods 1 and 3 have a notable number of medical stores whereas other clusters hardly have any. Hence the suitable cluster would be the Cluster 2 and Cluster 4 and 5. Cluster 5 has a Martial Arts School and many other shops which gives an advantage.

Similarly, based on the requirement suggestions can be provided about the neighborhood that would be best suitable for the business.

Below figure shows a map of Toronto with the neighborhood clusters superimposed on top of it. This map can be used to suggest a vast location to start a new business based on the category.





7. CONCLUSION

Purpose of this project was to analyze the neighborhoods of Toronto and create a clustering model to suggest personal places to start a new business based on the

category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 5 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e. 5 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster. A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided. Both these can be used by stakeholders to decide the location for the particular type of business. A major drawback of this project was that the Foursquare API returned only a few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.