

# Using foursquare API and Clustering to Identify a suitable location to open a new Bar in Toronto, Canada

Aishik Mukherjee

26th April, 2020

# Business Problem

A concept manager/owner wants to open a new Bar in Toronto, Canada. With the purpose in mind, finding the location to open a Bar is one of the most important decisions for this entrepreneur.

The idea behind this project is that there may be enough number of Bars in Toronto and it might present a great challenge for this entrepreneur who is based in Canada, to choose a suitable location in Toronto to sustain his business where the competition might be low.

# DATA

Toronto Neighborhood:

We will scrape the Data of neighborhoods in Toronto from a wikipedia page. This data contains a list of neighborhoods and boroughs in Toronto, Canada. After cleaning this data, it looks something like this-

	Postal code	Borough	Neighborhood
0	M1B	Scarborough	Malvern / Rouge
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek
2	M1E	Scarborough	Guildwood / Morningside / West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

## Location Data of Toronto Neighborhood:

After we obtained the list of neighborhoods in toronto, we needed the location data of those neighborhood in order to make calls to foursquare API. For this project, I used the csv file provided by IBM team to match the coordinates of Toronto neighborhoods. It looked something like this-

	Postal Code	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

# Using Foursquare API to get the Venue Data

After getting the location data of the neighborhoods in toronto, I used Foursquare API to pull the list of top 100 venues within 500 meters radius. By using the foursquare API, I pulled the venue data namely- names, categories, latitude and longitude of the venues. This data was particularly useful for this project as we'll be running most of our analysis based on the venue data we received by using Foursqaure API.

# Analysis and Clustering of Data

After getting the venue data of neighborhoods from Foursquare, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This was to prepare the dataset for clustering to be done later. I had looked specifically for bars. Lastly, I performed the clustering method by using k-means clustering. I had created 3 clusters of neighborhood and assigned each respective neighborhood to their corresponding clusters in the process. The resultant data-frame is here as follows:

# Adding cluster labels to Neighborhood

	Neighborhood	Bar	Cluster Labels
0	Berczy Park	0.000000	0
1	Brockton , Parkdale Village , Exhibition Place	0.043478	2
2	Business reply mail Processing Centre	0.000000	0
3	CN Tower , King and Spadina , Railway Lands , ...	0.058824	1
4	Central Bay Street	0.015385	2

# Merging the Dataset

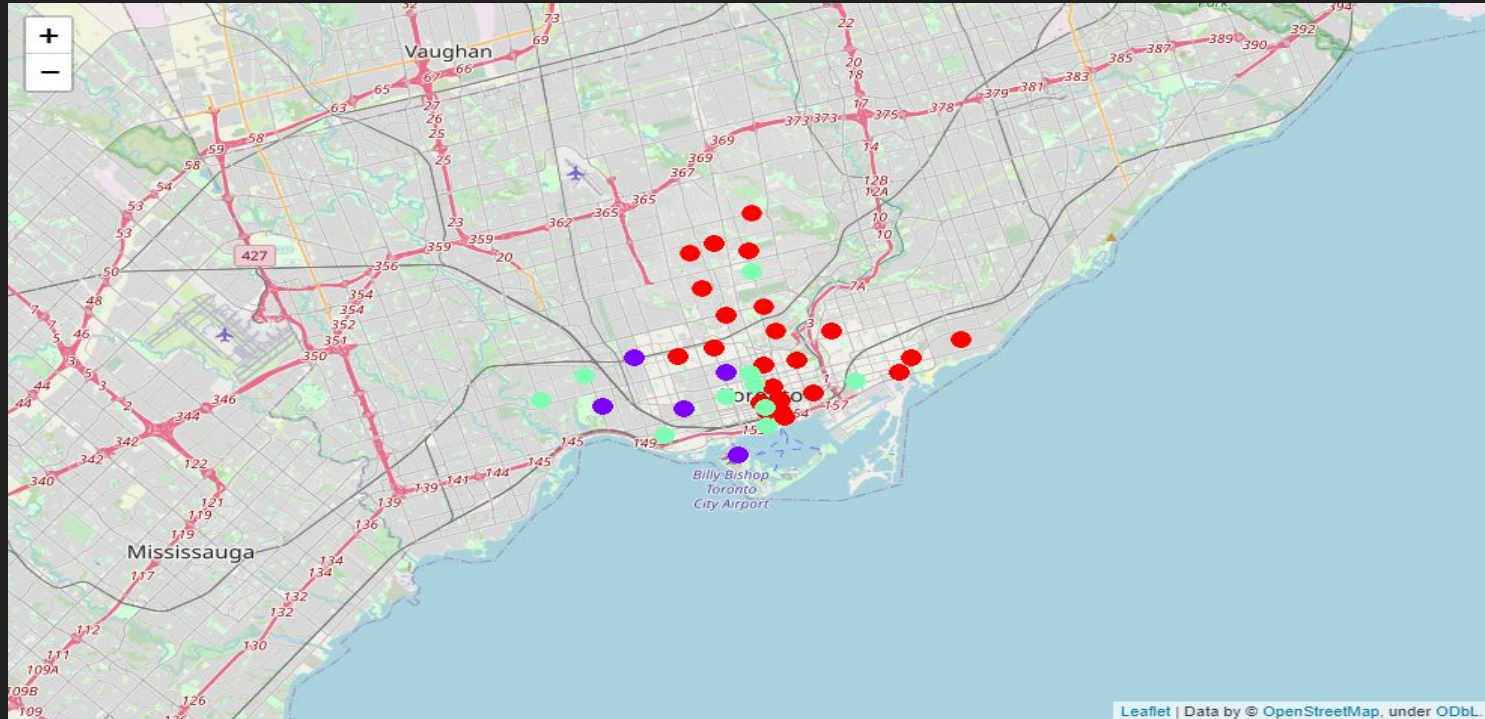
After creating clusters, we merged all the information to a dataset based on venues.

	Neighborhood	Bar	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berczy Park	0.0	0	43.644771	-79.373306	LCBO	43.642944	-79.372440	Liquor Store
0	Berczy Park	0.0	0	43.644771	-79.373306	The Keg Steakhouse + Bar - Esplanade	43.646712	-79.374768	Restaurant
0	Berczy Park	0.0	0	43.644771	-79.373306	Fresh On Front	43.647815	-79.374453	Vegetarian / Vegan Restaurant
0	Berczy Park	0.0	0	43.644771	-79.373306	Meridian Hall	43.646292	-79.376022	Concert Hall
0	Berczy Park	0.0	0	43.644771	-79.373306	Hockey Hall Of Fame (Hockey Hall of Fame)	43.646974	-79.377323	Museum



# Visualization of Clusters

The formed clusters are:



We categorized the Toronto neighborhoods into 3 clusters with the help of K-means based on how many Bars were in the neighborhood(as shown in the previous slide) as follows:

- Cluster 0(Red): Neighborhoods with little or no Bars
- Cluster 1(Blue): Neighborhoods with moderate number of Bars
- Cluster 2(Green): Neighborhoods with significant number of Bars

# Discussion and recommendation

By the analysis of the above 3 clusters, it was noted that most of the Bars were in Cluster 2 which are around First Canadian Place , Underground city, Harbourfront East , Union Station , and Toronto Islands.

The lowest number of bars were in Cluster 0 areas which are Toronto Dominion Centre , Design Exchange, Commerce Court , Victoria Hotel, Garden District, Ryerson, etc.

Hence, there are good opportunities to open the bar in these areas of Cluster 0 as competition seems to be low and it could attract the nearby crowd more often than compared to other clusters.

Based on the results obtained by the analysis, it can be concluded that the neighborhoods of Cluster 0 might be a good location as there are not a lot of Bars in these areas for people to choose from. Therefore, this project recommends the entrepreneur to open a Bar in a suitable neighborhood of Cluster 0.

# Limitations

I only take one factor into consideration: the occurrence / existence of Bars in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a Bar, Age group of people, etc.

Future improvements can certainly be taken into consideration of these factors. In addition, I am relying on the existence of Bars only for this project but future research can also be taken into consideration of other variables such as existence of Pubs, Nightclubs, etc in each neighborhood.

# Conclusion

In this project, we went through the process of identifying the business problem, specifying the data required, extracting and preparing the data, explaining the methodology of performing the machine learning by utilizing k-means clustering and providing best recommendation to the stakeholder.

THANK YOU