

# Applied Data Science Capstone of IBM Course

## Introduction

The goal of this project is to help people explore facilities around neighborhoods to make smart and efficient decisions on selecting a good place to live. It can be applied to many different situations such as: exploring the nightlife of a city, buying a house depending on the nearby facilities that surround the property, looking for hotels near good restaurants when you travel, etc.

This project aims to analyze the features for people interested in living in this area to search for the best neighborhood as a comparative analysis between neighborhoods. The features can include housing price, school ratings, crime rates, road connectivity, weather conditions, good management for emergency, water resources both fresh and waste water and excrement conveyed in sewers and recreational facilities.

For this project we are going to work with the following scenario: As an entrepreneur, one of my dreams is to open a craft brewery and since we are going to work with Toronto's data, I'm going to simulate where to open this business. There are several more factors when it comes to opening a business, but this project will help to gather data and get a starting point to locate a good spot.

The location will be Toronto. One of the most diverse and multicultural areas in Canada. Although immigration has become a hot topic over the past few years with more governments seeking more restrictions on immigrants and refugees, the general trend of immigration into Canada has been one of on the rise.

Foursquare API: We will use Four-square API as data source. This platform is a location data provider where we can gather information of all manner of venues and events within our area of interest. The API returns a JSON file that we can convert into a dataframe.

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 1000.

Using a clustering approach we can compare the similarities of neighborhoods, to explore, segment them, and group them into clusters to find similar neighborhoods in Toronto. To be able to do that, we need to cluster data using k-means clustering algorithm.

## Data Description

We are going to use the dataset we scrapped from wikipedia that consist of latitude, longitude and zip codes. This data set can be found [here](#):

[https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

	Borough	Postcode	Neighborhood	Latitude	Longitude
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678
3	Central Toronto	M4S	Davisville	43.704324	-79.388790
4	Central Toronto	M4T	Moore Park, Summerhill East	43.689574	-79.383160

In order to get data about this particular location (Toronto), we are using Foursquare API. This platform it's a location data provider where we can gather information of all manner of venues and events within our area of interest. The API returns a JSON file that we can convert into a dataframe.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

## Exploratory Data Analysis

Once we have gather, prepare and clean our data we can proceed to explore it.

There are many ways you can proceed with, but for this project I did the following:

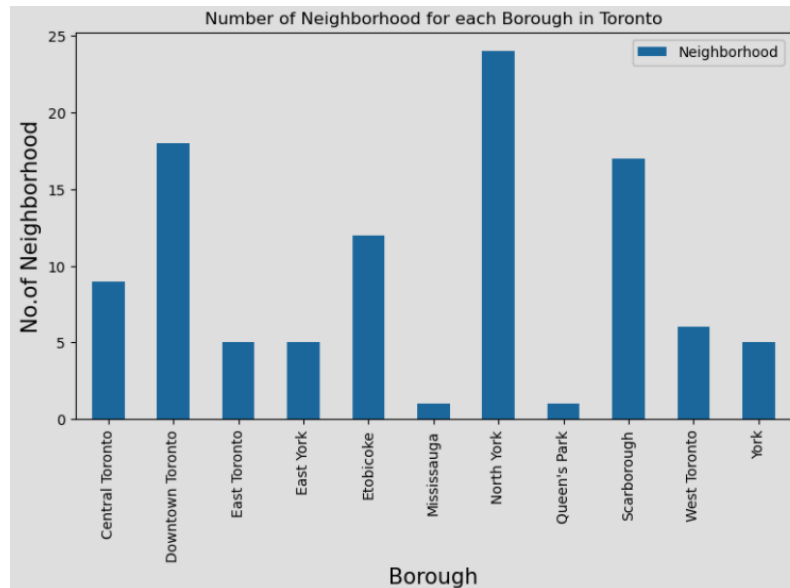
1. Checking how many venues were returned for each neighborhood,
2. Get the unique categories for the returned venues.

3. Count each category in each neighborhood
4. Getting the top 20 venues among all neighborhoods.
5. Visualize the neighborhoods and number of boroughs per neighborhood.
6. Get the number of breweries per neighborhood.

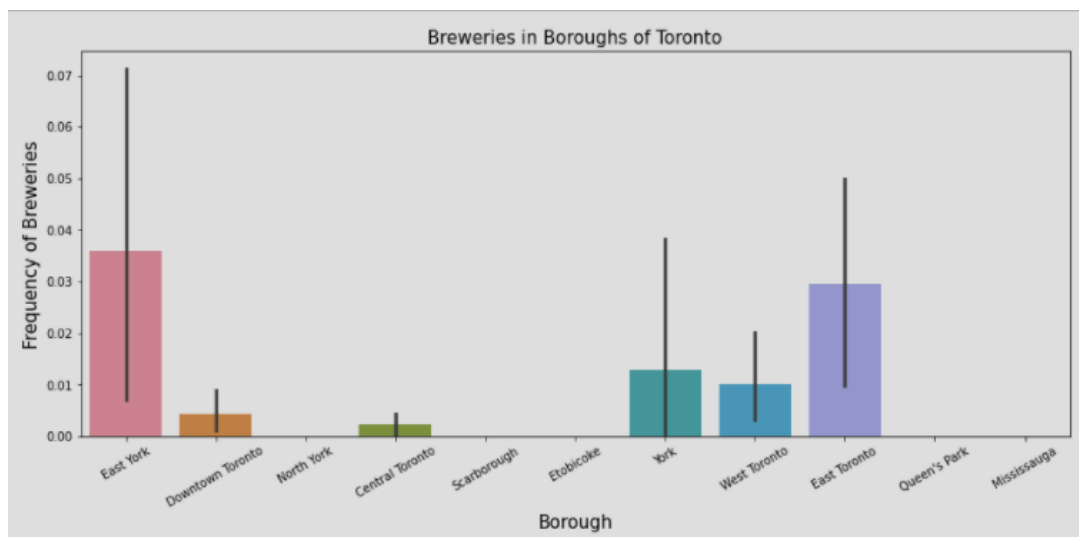
One of the very first interesting insights is that we can gather the top 20 venues at our neighborhoods and some statistics. We can say that Canadians really love coffee and foreign food such as Korean, Greek, Italian, Indian and Chinese. Our point of interest is the Brewery which gets the 15th place in the top. Another interesting point is that there are 325 unique categories.

	count	mean	std	min	25%	50%	75%	max
<b>Coffee Shop</b>	102.0	3.715686	3.404649	0.0	1.0	2.5	5.0	15.0
<b>Korean Restaurant</b>	102.0	0.274510	1.470212	0.0	0.0	0.0	0.0	13.0
<b>Greek Restaurant</b>	102.0	0.343137	1.396566	0.0	0.0	0.0	0.0	12.0
<b>Café</b>	102.0	2.078431	2.809765	0.0	0.0	1.0	3.0	11.0
<b>Italian Restaurant</b>	102.0	1.098039	1.512390	0.0	0.0	1.0	2.0	9.0
<b>Indian Restaurant</b>	102.0	0.460784	0.971592	0.0	0.0	0.0	1.0	7.0
<b>Chinese Restaurant</b>	102.0	0.480392	1.096623	0.0	0.0	0.0	1.0	7.0
<b>Bubble Tea Shop</b>	102.0	0.215686	0.712372	0.0	0.0	0.0	0.0	6.0
<b>Ramen Restaurant</b>	102.0	0.245098	0.788987	0.0	0.0	0.0	0.0	6.0
<b>Bar</b>	102.0	0.500000	1.272325	0.0	0.0	0.0	0.0	6.0
<b>Restaurant</b>	102.0	1.421569	1.524949	0.0	0.0	1.0	2.0	6.0
<b>Sushi Restaurant</b>	102.0	0.784314	1.294524	0.0	0.0	0.0	1.0	6.0
<b>Grocery Store</b>	102.0	0.960784	1.125066	0.0	0.0	1.0	1.0	6.0
<b>Gastropub</b>	102.0	0.480392	0.951605	0.0	0.0	0.0	1.0	6.0
<b>Brewery</b>	102.0	0.343137	0.850103	0.0	0.0	0.0	0.0	5.0
<b>Clothing Store</b>	102.0	0.274510	0.746701	0.0	0.0	0.0	0.0	5.0
<b>Pizza Place</b>	102.0	1.480392	1.200086	0.0	1.0	1.0	2.0	5.0
<b>Diner</b>	102.0	0.392157	0.772765	0.0	0.0	0.0	1.0	5.0
<b>Park</b>	102.0	1.568627	1.270455	0.0	1.0	1.0	2.0	5.0
<b>Vegetarian / Vegan Restaurant</b>	102.0	0.441176	0.970593	0.0	0.0	0.0	0.0	5.0

Once you have finish exploring your data set, you will have a very good idea of what it is going on. For this particular project we are interested in opening a brewery, so based on the conclusions we get we will make a decision.



We can see that North York has the highest number of neighborhoods. Now we want to know how many breweries are in each neighborhood.



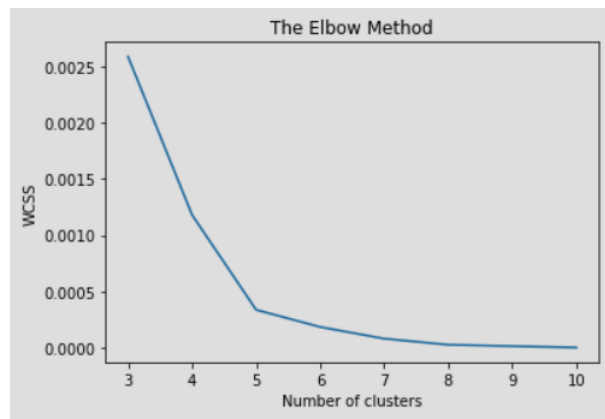
On the contrary than the last plot, we can recall that North York has the greatest number of neighborhoods, however as we may see in this plot, there are no breweries at this location,

such as Etobicoke, Mississauga, Queen's Park and Scarborough, this may indicate that these neighborhoods are most likely residential. On the other hand, East Toronto and East York have the greatest number of breweries by far than the rest.

## Predictive modeling

Since we are going to use the K-means clustering algorithm, the first step is to identify the K value using the elbow method to get the optimal number of K.

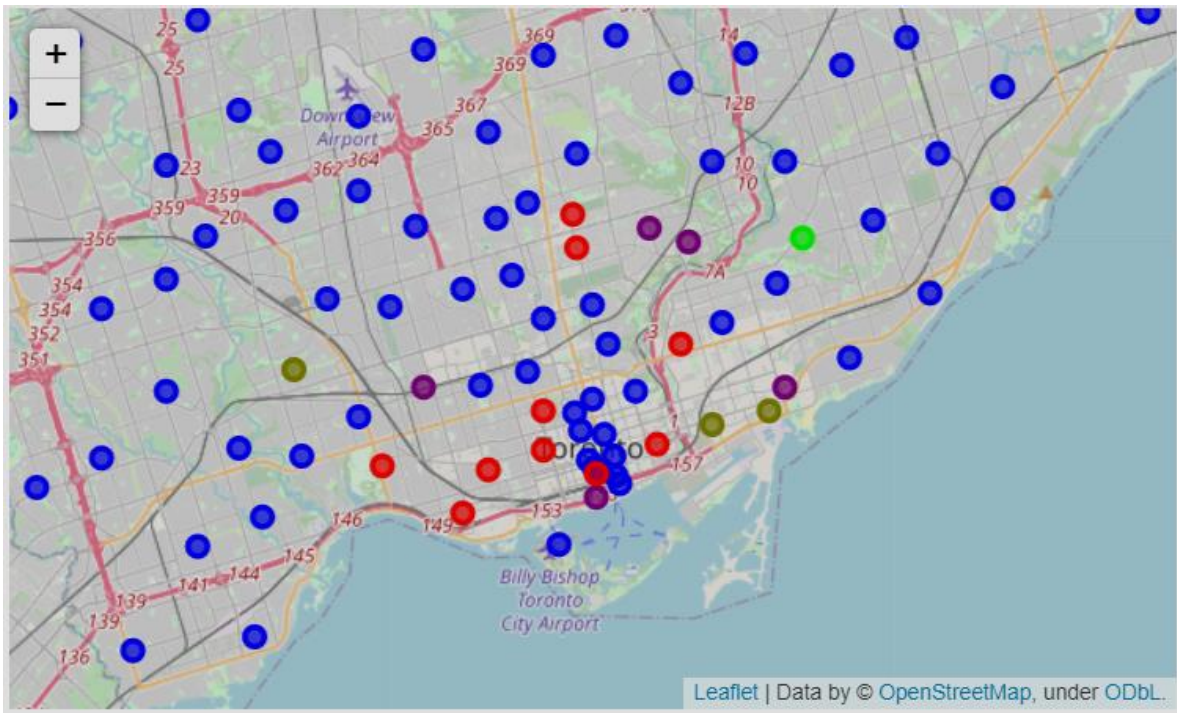
This method calculates the sum of squared distances of samples to their closest cluster center for different values of K. The value of K after which there is no significant decrease in sum of squared distances is chosen.



Resulting from the plot of the elbow method it seems that K=5 is the best value for our model, so we are going to proceed with this value.

## Mapping clusters

Once we applied the algorithm, we get a data set containing the results of each cluster. Visually we get the map of the right with the following insights:



Resulting from clustering we can see cluster 0 in color green, cluster 1 in color blue, cluster 2 in color purple, cluster 3 in color red and cluster 4 in color yellow.

- Cluster 0 has 1 result.
- Cluster 1 has 83 results.
- Cluster 2 has 5 results.
- Cluster 3 has 10 results.
- Cluster 4 has 3 results.

Based on these results and the exploratory data analysis we can make conclusions.

## Conclusions

Based on the visualizations and data sets generated we can conclude that:

- North York Etobicoke, Mississauga, Queen's Park and Scarborough have no breweries at these neighborhoods, which may indicate that these neighborhoods are most likely residential and may be not the optimal place to open a business.
- On the other hand, neighborhoods like East Toronto and East York have the greatest number of breweries which may be strong competitors for our new business. It would

be better to place our business at York or West Toronto that have a humbler number of breweries, or even at Central Toronto and Downtown Toronto that have the least number of breweries.

- Clustering gives us similar results, suggesting that opening brewery will be a good idea at 2 or 4.