# **Find Best Place to Open a Cafe in Toronto Canada**

Aswini Chinnenahalli Siddareddy

February 22, 2021

1. **Introduction**
   1. **Background**

Toronto is capital city of the Ontario, Canada with a population of more than 6 Million in 2016 and with area of 243.3 sq mi. Toronto is international center for business, finance, arts, and culture, and is recognized as one of the most multicultural cities in the world. Toronto economy is diversified with technology, design, financial services, life sciences, education, arts, fashion, aerospace, environmental innovation, food services, and tourism. One of my friends who was having Subway restaurant had to close his Subway business during May 2020 due to Covid-19 Pandemic which led to reduced sales in his business. Now he is planning to open a Cafe business in Toronto, Canada region. Starting a Cafe business in such a multicultural and diversified place is not an easy task. We need to consider several factors such as accessibility, visibility, target customers before opening Cafe to be successful in business. Places near business centers, malls, areas with a lot of foot traffic, and tourist attractions will guarantee the steady flow of customers that we need to make a good profit.

* 1. **Problem**

Finding a location is one of the most important things in starting a cafe. The location search could take months if we start searching and analyzing the area manually. We can reduce this time to few hours/ days by using machine learning techniques and Four-square location data to find best suitable location to open a Cafe. In this project, we will find the best and most suitable location to open a Cafe in Toronto, Canada for business personal or entrepreneurs.

**1.3 Interest**

* This project would be interested to business personnel who wants to open a Café in Toronto, Canada.
* The analysis will help entrepreneurs to obtain necessary information in finding the best location for opening a Cafe.

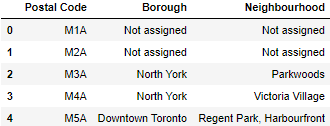
1. **Data sources**

We will use the following datasets for this project:

1. Toronto data that contains Borough, Neighborhoods along with there latitudes and longitudes
   * Data Source: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
   * Description: This data set contains the required information such as postal code, borough and the name of the neighborhoods in city of Toronto
2. Geographical location of the neighborhoods
   * Data Source: <https://cocl.us/Geospatial_data>
   * Description: This dataset provides the Geographical coordinates of the neighborhoods for the respective Postal Codes
3. Venue Data using Foursquare API
   * Data Source: [https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}&ll={},{}&radius={}&limit={}](https://api.foursquare.com/v2/venues/explore?&client_id=%7B%7D&client_secret=%7B%7D&v=%7B%7D&ll=%7B%7D,%7B%7D&radius=%7B%7D&limit=%7B%7D)
   * Description: We will provide ClientID, Client Secret, and version details to get all the venue information for each neighborhood and group data by name of the neighborhood.
4. **Methodology**

**3.1 Data cleaning**

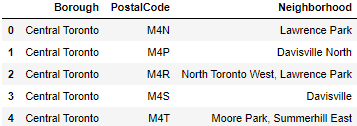
Data downloaded from above data sources are created as panda’s data frame. We can see the top rows of Toronto dataset below.



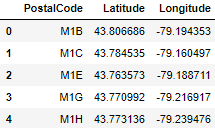
Toronto data scraped from Wikipedia page has Boroughs that were not assigned to any neighborhood. The following assumptions were made to clean the Toronto Wikipedia dataset.

* Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
* More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
* If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

The above assumptions are implemented, and rows are grouped by borough which is shown below.



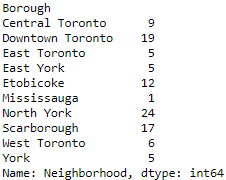
The top rows of geographical location of the neighborhood’s dataset which is extracted from <https://cocl.us/Geospatial_data> website is shown below.



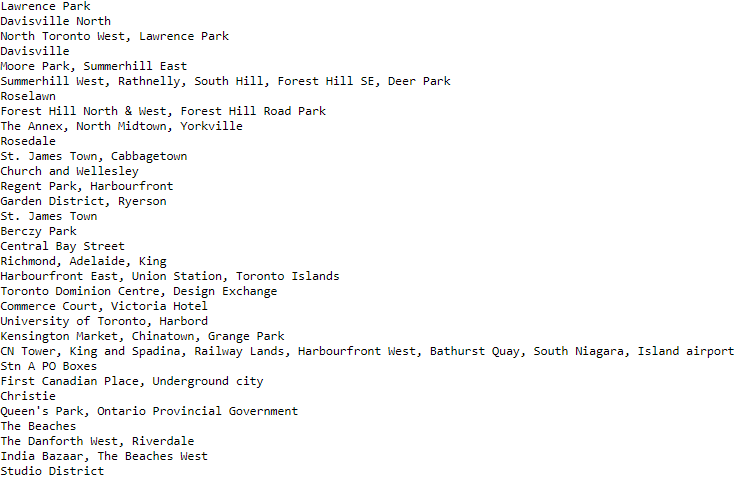
We merged the above two datasets into single dataset and the merged dataset is shown below.



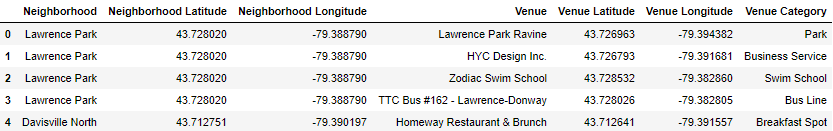
We have grouped the neighborhoods by borough and the count of neighborhoods by borough can be seen below.



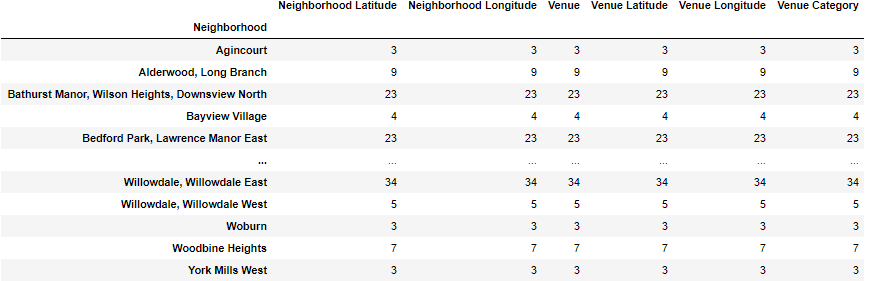
We have retrieved all the venue details such as parks, islands, hotels, restaurants, café, etc. in Toronto using Four Square API to analyze number of cafés in Toronto. We must create an account in Four square API and provide client id, client secret and version details to retrieve the data from Four Square API. The top few rows of venue details of neighborhood is shown below.



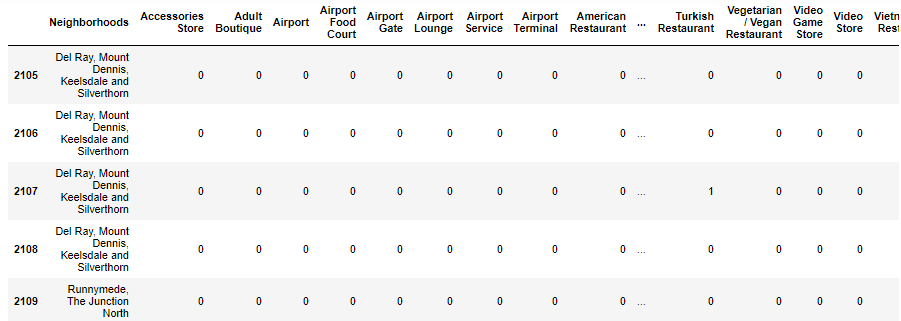
We merged the above venue data with the previous two datasets which shows the nearest venue for each neighborhood. We can see the merged dataset in below table.



We then grouped the venues by neighborhood to see the total number of venues in each neighborhood.



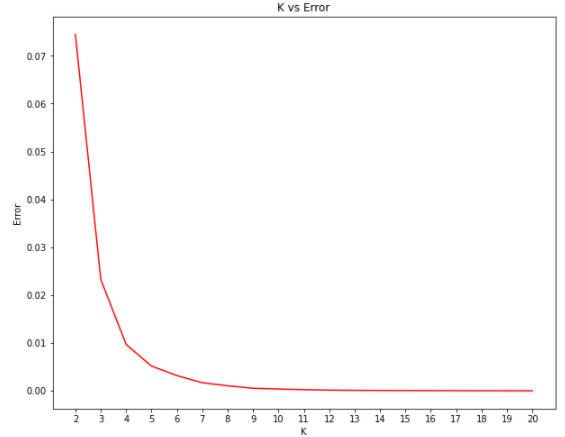
It is difficult to analyze the categorical data. So, we converted the categorical data to numerical data using a technique called one hot encoding. If the venue is present in the neighborhood, it returns the count of how many times that venue is present in neighborhood else zero. We can see the conversion in below screenshot.

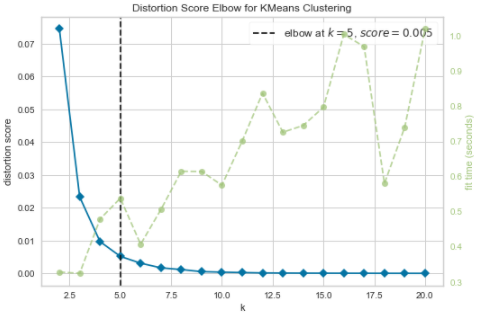


We group the above data by neighborhood to see average number of Café’s present in each neighborhood.

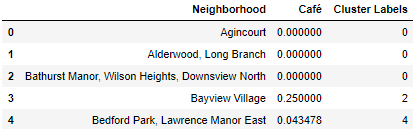


We have clustered the neighborhoods based on the similar number of cafés in that neighborhood. We have used K-means clustering. We have used elbow technique to find the best k value that neither overfit nor under fits our model. We chose the best value of k by providing different k values. We have chosen the best k value based on the point at which the line in the below graph has sharp turn. From the below screenshot we can see that our elbow point is at k=5 which means we have 5 clusters.



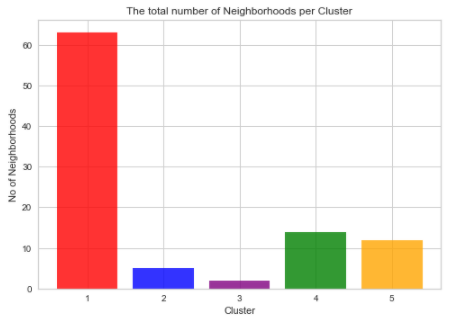


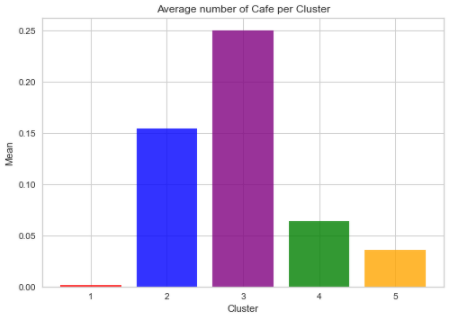
We will put the similar objects of a variable in the same cluster in k-means clustering. So, the neighborhoods that has similar frequency of café’s were added to 5 clusters and are labelled from 0 to 4.



**3.2 Analysis**

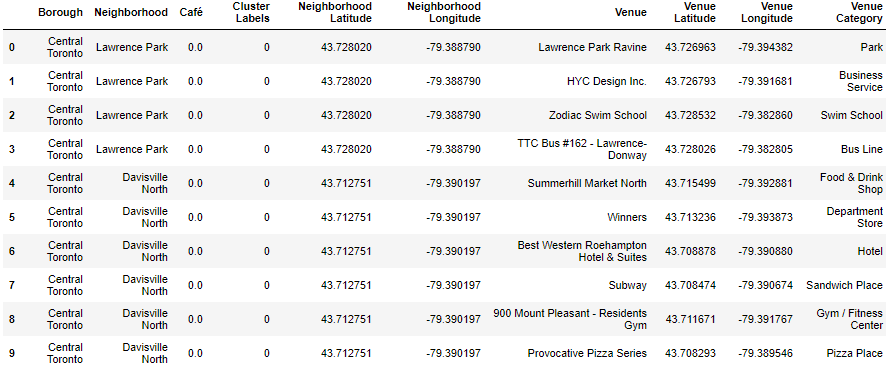
We analyze the five clusters by finding the total number of neighborhoods in each cluster and average number of cafés in that cluster. We create a bar chart to see the total number of neighborhoods per cluster and average number of cafés per cluster. From the below graph we can see that the total number of neighborhoods in cluster 1 is 63, cluster 2 has 5, cluster 3 has 2, cluster 4 has 14 and cluster 5 has 12. We can also see that the average number of café in cluster 3 is higher even though the number of neighborhoods in cluster 3 are 2 and the average number of café in cluster 1 is lower even though the number of neighborhoods in cluster 1 are 63.



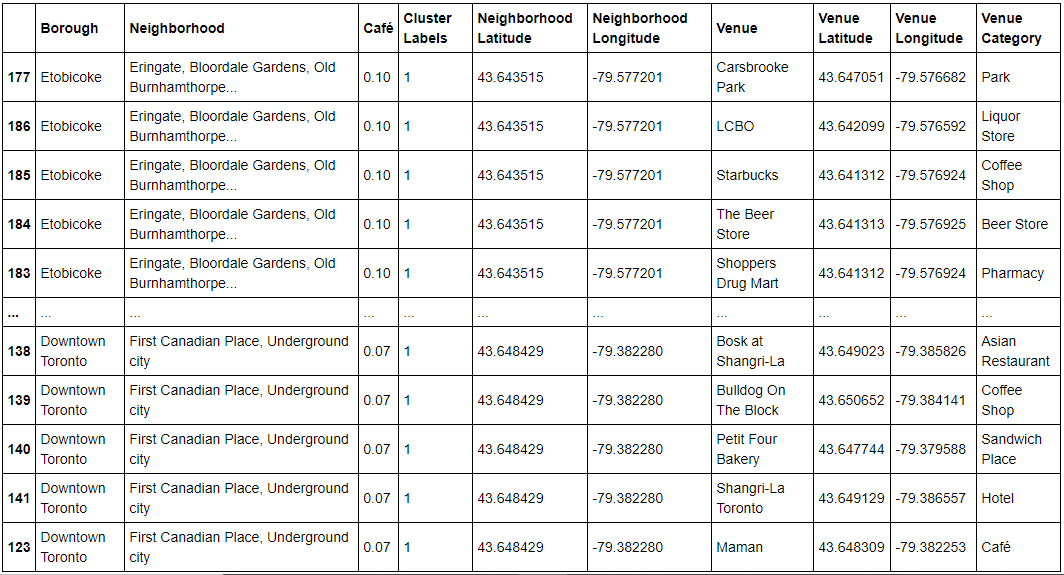


Let us analyze the individual clusters by looking at their data frames.

Cluster 1:



We can see that cluster 1 is in Central Toronto area and it has 65 neighborhoods. We have 22 unique venue locations in this cluster and there are very few cafés. Cluster 1 has highest number of neighborhoods.

Cluster 2: 

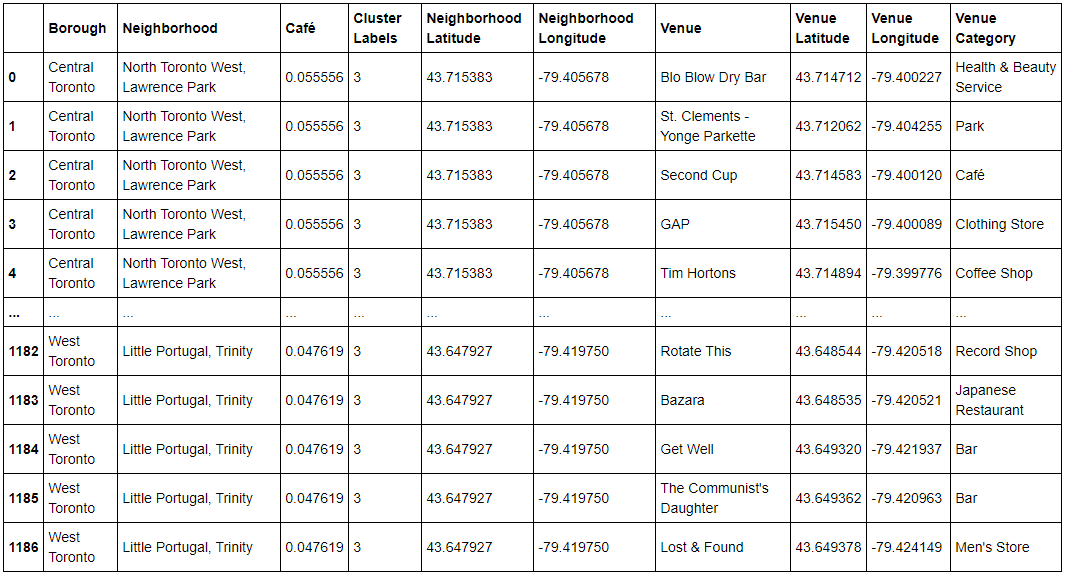
We can see that cluster 2 is in Etobicoke, Downtown Toronto, etc. area and Erin gate, Bloor dale Gardens, Old Bumhamthrope, First Canadian Place, Underground city etc., are few neighborhoods in that cluster. We have 19 unique venue locations in this cluster and there are a greater number of Cafés in this cluster when compared to cluster 1.

Cluster 3:



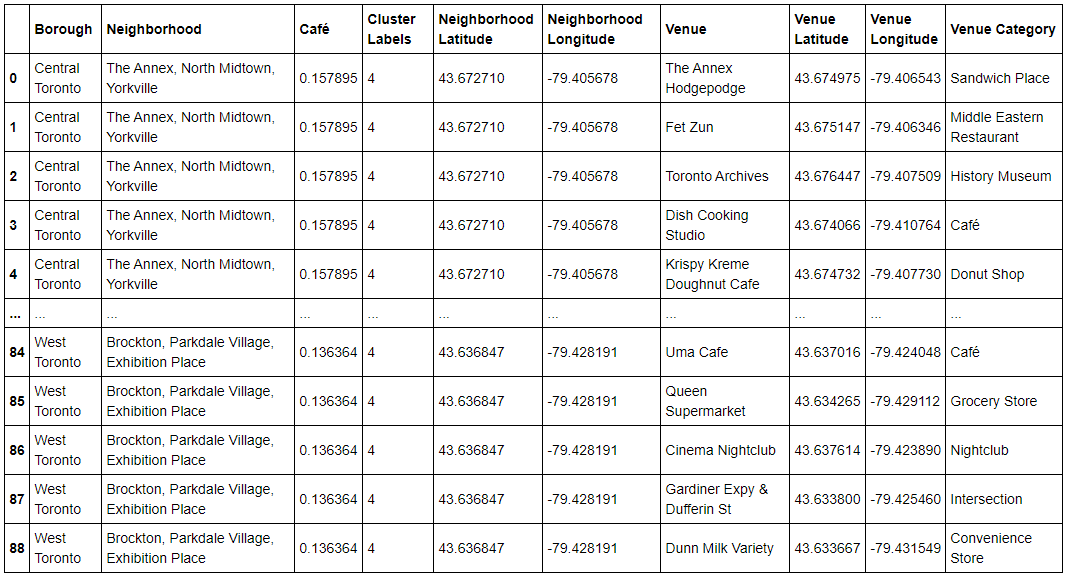
We can see that cluster 3 is in North York, Scarborough area and Bayview Village, Birch Cliff, Cliffside West are few neighborhoods in that cluster. We have 8 unique venue locations in this cluster and there are highest number of cafés in this cluster.

Cluster 4:



We can see that cluster 4 is in Central Toronto, West Toronto area and North Toronto West, Lawrence Park, Little Portugal, Trinity are few neighborhoods in that cluster. We have 56 unique venue locations in this cluster and there are only few cafés. Cluster 4 has third highest average for Café’s.

Cluster 5:



We can see that cluster 5 is in Central Toronto and West Toronto area and The Annex, North Midtown, Yorkville are few neighborhoods in this cluster. We have 26 unique venue locations in this cluster and there are very few cafés.

The average number of cafés in each cluster is shown below:

Cluster 1 - 0.01

Cluster2 - 0.16

Cluster3 - 0.25

Cluster4 - 0.065

Cluster5 - 0.04

1. **Discussion**

The greater number of Café’s are in cluster 3 and is represented by purple color. The neighborhoods located in North York and Scarborough area t have highest number of Cafes. There are very few Café’s in cluster 1 even though there are so many neighborhoods in cluster 1. From the above graphs we can see that cluster 5 has second lowest average number of cafés. By looking at the venues, we can say that the best place to open new café is Central Toronto area as there are many neighborhoods in that area and little to no cafés. The second-best place to open a Café is in West Toronto area which is in Cluster 5. It has 12 neighborhoods in the area with very few number of cafés. We should note that, we have not included population data in this analysis and population plays a huge role in choosing a place to open a new café.

1. **Conclusion**

In conclusion, we have used several python libraries, Four Square API to find the best area to open a Café in Toronto, Canada. We created several visualizations to better understand our dataset and analysis using seaborn, matplotlib libraries. By using all these libraries, machine learning techniques, we were able to find the best location to open a Café in Toronto, Canada.

1. **Future Directions**

In this study, we mainly focused on analyzing neighborhoods, venue details. However, population, other demographic details might also contribute to evaluate Café business success criteria. By adding demographic data into analysis, we can increase the confidence and accuracy of our analysis in finding best place to open Café in Toronto.

<https://github.com/umarkhan2000/IBM-DataScience-Capstone/blob/master/IBM-Capstone/Opening%20up%20a%20new%20Italian%20Restaurant%20in%20Toronto.pdf>