# Exploring Toronto

Capestone Project : Finding the best Neighborhood in Toronto using Data Science

BY PREMA NARAYANAN

# Table of Contents

# Introduction

Identifying the Business Problem :

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the ‘best’ neighbourhood in Toronto to open a restaurant

This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course.

# Target Audience

This project is aimed towards Entrepreneurs or Business owners who want to open a new Italian Restaurant or grow their current business.

The analysis will provide vital information that can be used by the target audience that can be used for decision making purpose.

# Data Overview

The data set required for the following project was acquired from three different data sources as mentioned below :

1. A Wikipedia Page to fetch boroughs and neighborhoods of Toronto city.
2. Geographical Location data using Geocoder Package
3. The foursquare api to fetch different public venues in the vicinity of the neighborhood.

# Methodology

We will perform the following steps

1. Data Cleansing
2. Data Exploration
3. Machine Learning
4. Data Analysis

Please refer to the ‘Exploring Toronto neighbourhood for Italian restaurant’ report for full details on the methodology

# Data Cleansing

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis.

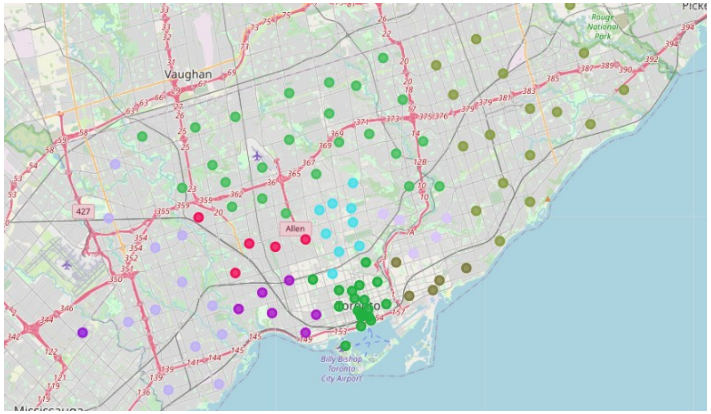
The rows were grouped based on the borough and using the Latitude and Longitude collected from the Geocoder package, I merged the two tables together based on Postal Code.

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

Figure 4: Merging tables together based on Postal Code

# Data Exploration

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located in as below.




Next, we used the Foursquare API to get a list of all the Venues in Toronto. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods as below

| | Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Lawrence Park | 43.728020 | -79.388790 | Lawrence Park Ravine | 43.726963 | -79.394382 | Park |
| 1 | Lawrence Park | 43.728020 | -79.388790 | Zodiac Swim School | 43.728532 | -79.382860 | Swim School |
| 2 | Lawrence Park | 43.728020 | -79.388790 | TTC Bus #162 - Lawrence-Donway | 43.728026 | -79.382805 | Bus Line |
| 3 | Davisville North | 43.712751 | -79.390197 | Homeway Restaurant & Brunch | 43.712641 | -79.391557 | Breakfast Spot |
| 4 | Davisville North | 43.712751 | -79.390197 | Sherwood Park | 43.716551 | -79.387776 | Park |

# Machine Learning

Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding.

For each of the neighbourhoods, individual venues were turned into the frequency based on the neighbourhood and grouped by neighbourhood by taking the average of frequency of occurrence of each Venue Category.

After, we created a new data frame that only stored the Neighborhood names as well as the mean frequency of Italian Restaurants in that Neighborhood.

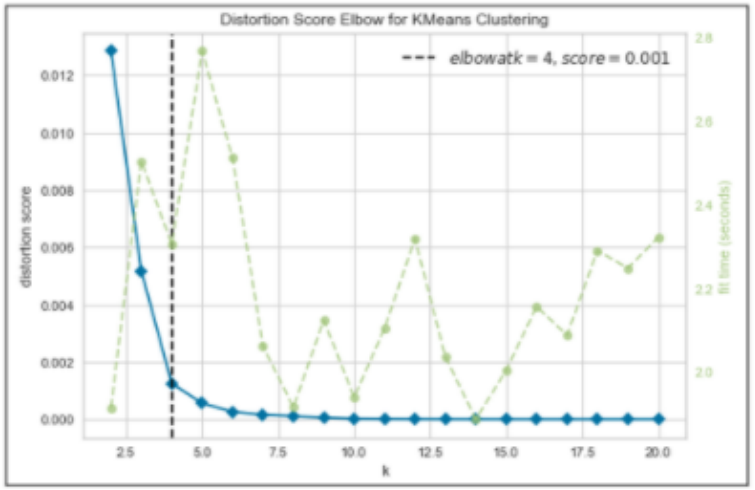
| | Neighbourhoods | Italian Restaurant |
|---|---|---|
| 0 | Agincourt | 0.000000 |
| 1 | Alderwood, Long Branch | 0.000000 |
| 2 | Bathurst Manor, Wilson Heights, Downsview North | 0.000000 |
| 3 | Bayview Village | 0.000000 |
| 4 | Bedford Park, Lawrence Manor East | 0.090909 |

# Machine Learning -K-Means Clustering

We use K-means to cluster the neighbourhoods based on the neighbourhoods that had similar averages of Italian Restaurants in that Neighborhood.

To get our optimum K value that was neither overfitting or underfitting the model, we used the Elbow Point Technique. The best K value is chosen at the point in which the line has the sharpest turn. Then we used a model that accurately pointed out the optimum K value.

We imported 'KElbowVisualizer' from the Yellowbrick package and fit our K-Means model to the Elbow visualizer. We just integrated a model that would fit the error and calculate the distortion score.



# Machine Learning -K-Means Clustering (cont'd)

in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighbourhoods that had a similar mean frequency of Italian Restaurants were divided into 4 clusters. Each of these clusters was labeled from 0 to 3 as the indexing of labels begins with 0 instead of 1.

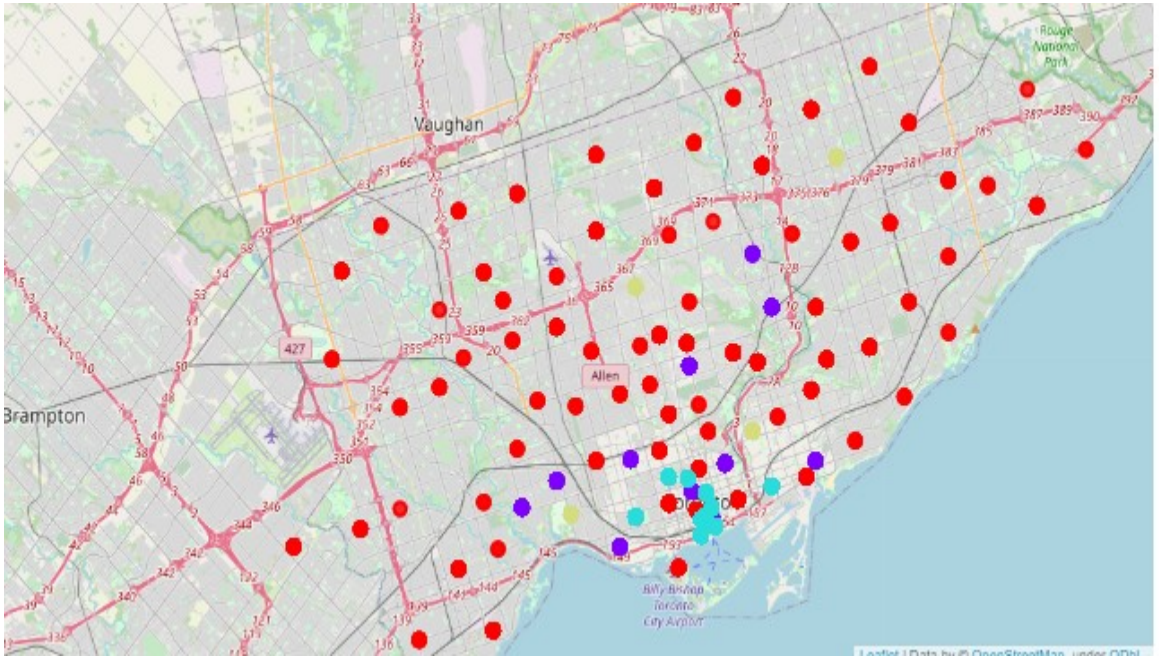
Then we created a map using the Folium package in Python and each neighbourhood was coloured based on the cluster label.

Cluster 1 — Red

Cluster 2 — Purple

Cluster 3 — Turquoise

Cluster 4 — Dark Khaki



# Data Analysis

From the bar graph that was made using Matplotlib (figure 18), we can compare the number of Neighborhoods per Cluster.

We see that Cluster 1 has the maximum neighbourhoods (72) while cluster 2 and cluster 3 have the same no. of neighbourhood i.e. 10. Cluster 4 has only 4 neighbourhoods . Then we compared the average Italian Restaurants per cluster.

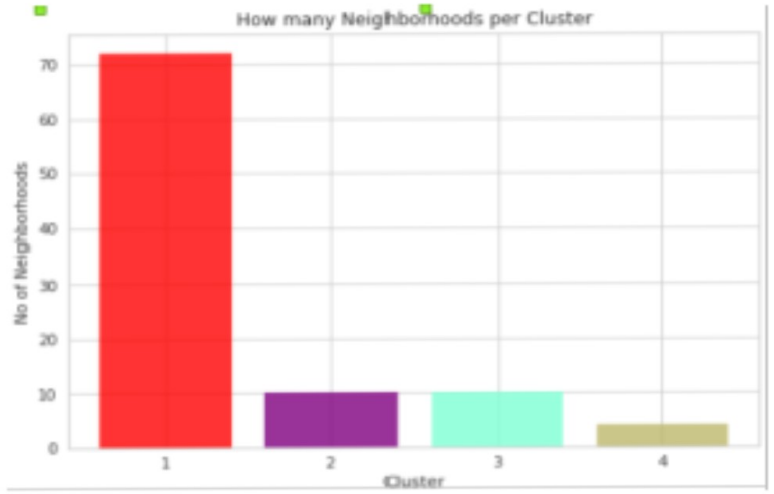



Figure 17: Number of Neighborhoods per cluster

# Results -Cluster Analysis

Cluster 1 is the Central Toronto ,YorK and other areas. Lawrence Park, Davisville North , , The Junction North , Weston etc are the Neighborhoods that were in that cluster. Cluster 1 had the lowest average (0.0) of Italian Restaurants equating with 72 neighbourhood (highest no.) . This seems like a good opportunity given the maximum neighbours and minimum existing italian restaurant.

Cluster 2 has a total of 10 neighborhoods of 19 venue category . It is concentrated around Downtown Toronto , West Toronto etc. This cluster has the 2nd highest no of Avg. number of Italian restaurant (≈ 0.045)

| | Borough | Neighbourhood | Italian Restaurant | Cluster Labels | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Central Toronto | Lawrence Park | 0.0 | 0 | 43.728020 | -79.388790 | TTC Bus #162 - Lawrence-Donway | 43.728026 | -79.382805 | Bus Line |
| 1 | Central Toronto | Lawrence Park | 0.0 | 0 | 43.728020 | -79.388790 | Zodiac Swim School | 43.728532 | -79.382860 | Swim School |
| 2 | Central Toronto | Lawrence Park | 0.0 | 0 | 43.728020 | -79.388790 | Lawrence Park Ravine | 43.726963 | -79.394382 | Park |
| 3 | Central Toronto | Davisville North | 0.0 | 0 | 43.712751 | -79.390197 | 900 Mount Pleasant - Residents Gym | 43.711671 | -79.391767 | Gym / Fitness Center |
| 4 | Central Toronto | Davisville North | 0.0 | 0 | 43.712751 | -79.390197 | Sherwood Off-leash Dog Park | 43.715711 | -79.390118 | Dog Run |

Cluster 1 – sample data

| | Borough | Neighbourhood | Italian Restaurant | Cluster Labels | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|---|
| 248 | Downtown Toronto | Christie | 0.062500 | 1 | 43.669542 | -79.422564 | Marian Engel Park | 43.673754 | -79.423988 | Park |
| 256 | Downtown Toronto | Christie | 0.062500 | 1 | 43.669542 | -79.422564 | Actinolite | 43.667858 | -79.428054 | Restaurant |
| 263 | Downtown Toronto | Christie | 0.062500 | 1 | 43.669542 | -79.422564 | Foto Grocery | 43.667979 | -79.428217 | Grocery Store |
| 262 | Downtown Toronto | Christie | 0.062500 | 1 | 43.669542 | -79.422564 | Garrison Creek Park | 43.671690 | -79.427805 | Park |
| 261 | Downtown Toronto | Christie | 0.062500 | 1 | 43.669542 | -79.422564 | Fiesta Farms | 43.668471 | -79.420485 | Grocery Store |

Cluster 2 – sample data

# Results -Cluster Analysis

Cluster 3 has the same no. of neighbourhood as Cluster 2 (10) . Cluster 3 was mainly located in the Downtown and West Toronto like Garden District, Little Portugal etc. The average of Italian Restaurants in this cluster which was ≈ 0.020. There seems to be a good opportunity to grow in this cluster as well

Cluster 4 venues were located mainly in East and West Toronto area were some of the neighborhoods that made up this cluster. This cluster has the highest average of Italian Restaurants which is ≈ 0.075. There does not seem much opportunity to grow here based on the analysis. However we can further analyze the population type here to understand the reasons for high average of Italian restaurants in this cluster

| | Borough | Neighbourhood | Italian Restaurant | Cluster Labels | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Downtown Toronto | Garden District, Ryerson | 0.020000 | 2 | 43.657162 | -79.378937 | JOEY Eaton Centre | 43.656094 | -79.381878 | New American Restaurant |
| 1 | Downtown Toronto | Garden District, Ryerson | 0.020000 | 2 | 43.657162 | -79.378937 | BMV Books | 43.657047 | -79.381661 | Bookstore |
| 2 | Downtown Toronto | Garden District, Ryerson | 0.020000 | 2 | 43.657162 | -79.378937 | lululemon athletica | 43.653286 | -79.380764 | Clothing Store |
| 3 | Downtown Toronto | Garden District, Ryerson | 0.020000 | 2 | 43.657162 | -79.378937 | Ryerson Square | 43.656988 | -79.376896 | Other Great Outdoors |
| 4 | Downtown Toronto | Garden District, Ryerson | 0.020000 | 2 | 43.657162 | -79.378937 | Disney Store | 43.654248 | -79.381232 | Toy / Game Store |

Cluster 3 – sample data

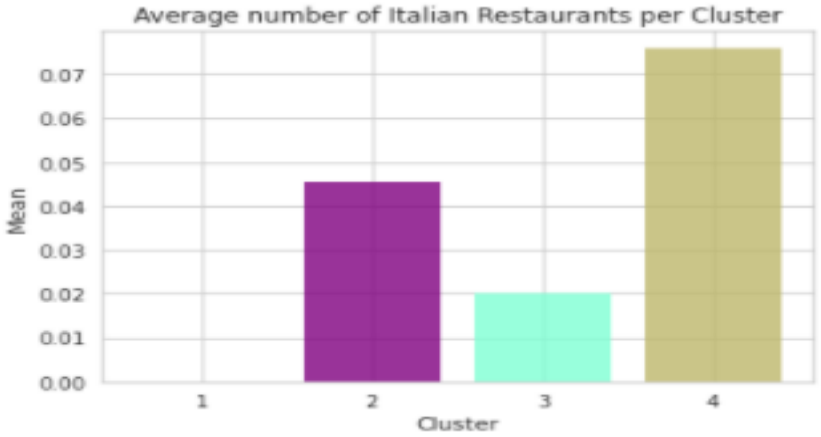
| | Borough | Neighbourhood | Italian Restaurant | Cluster Labels | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | East Toronto | The Danforth West, Riverdale | 0.068182 | 3 | 43.679557 | -79.352188 | Kitchen Stuff Plus | 43.678613 | -79.346422 | Furniture / Home Store |
| 1 | East Toronto | The Danforth West, Riverdale | 0.068182 | 3 | 43.679557 | -79.352188 | Factory Girl | 43.676693 | -79.356299 | American Restaurant |
| 2 | East Toronto | The Danforth West, Riverdale | 0.068182 | 3 | 43.679557 | -79.352188 | Katsu Japanese Restaurant | 43.678619 | -79.347024 | Sushi Restaurant |
| 3 | East Toronto | The Danforth West, Riverdale | 0.068182 | 3 | 43.679557 | -79.352188 | Carrot Commons | 43.677485 | -79.353076 | Restaurant |
| 4 | East Toronto | The Danforth West, Riverdale | 0.068182 | 3 | 43.679557 | -79.352188 | Second Cup | 43.677232 | -79.352898 | Coffee Shop |

Cluster 4 – sample data

# Results - Cluster Analysis

**Therefore, the ordering of the average Italian Restaurant in each cluster as mentioned is as below**

1. **Cluster 1 (Red) ≈ 0.00**
2. **Cluster 3 ( Turquoise) ≈ 0.045**
3. **Cluster 2 (Purple) ≈ 0.020**
4. **Cluster 4 (Dark Kaki) ≈ 0.075**



# Discussion - Final Recommendation

Even though there is a huge no. of Neighborhoods in cluster 1 (70+), there is little to no Italian Restaurant, eliminating any competition. Therefore this cluster will be our first recommendation for opening a new Italian restaurant

The second best Neighborhoods that have a great opportunity would be in areas such as Downtown and West Toronto which is in Cluster 3. Having 10 neighborhoods in the area with only few Italian Restaurants gives a good opportunity as well

Cluster 4 has highest no of Italian restaurant with only 4 neighbhourhood. However we can further analyze the population demographics here to understand the reasons for high average of Italian restaurants in this cluster which could be due to high density of italian immigrants / affluent population who may have preferences for Italian restaurants.

Our current analysis does not take into consideration of the Italian population or other demographics across neighborhoods however this can play a huge factor in selecting the neighbourhood.

# Conclusion

During the course of this capstone project, I was able to apply different data science techniques and tools that I learned in the IBM Data Science course. This helped me unearth meaningful insights from the data analysis that I did on the Toronto data set. Finally to conclude this project, I have got a chance to work on a business problem like how a real like data scientists would do. I have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get data from Wikipedia which was scraped with help of Wikipedia python library and visualized using various plots present in seaborn & matplotlib. I also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Taking account of Italian population data and other life style choice data (population demographics) can also help improve the prediction. Additionally this data can be used for other scenarios like starting other entrepreneur ventures in Toronto with few tweeks in the code. Even with these limitations , this project helps acts as initial guide to take more complex real-life challenges using data-science.

Thank you !