Battle of the Neighbourhoods

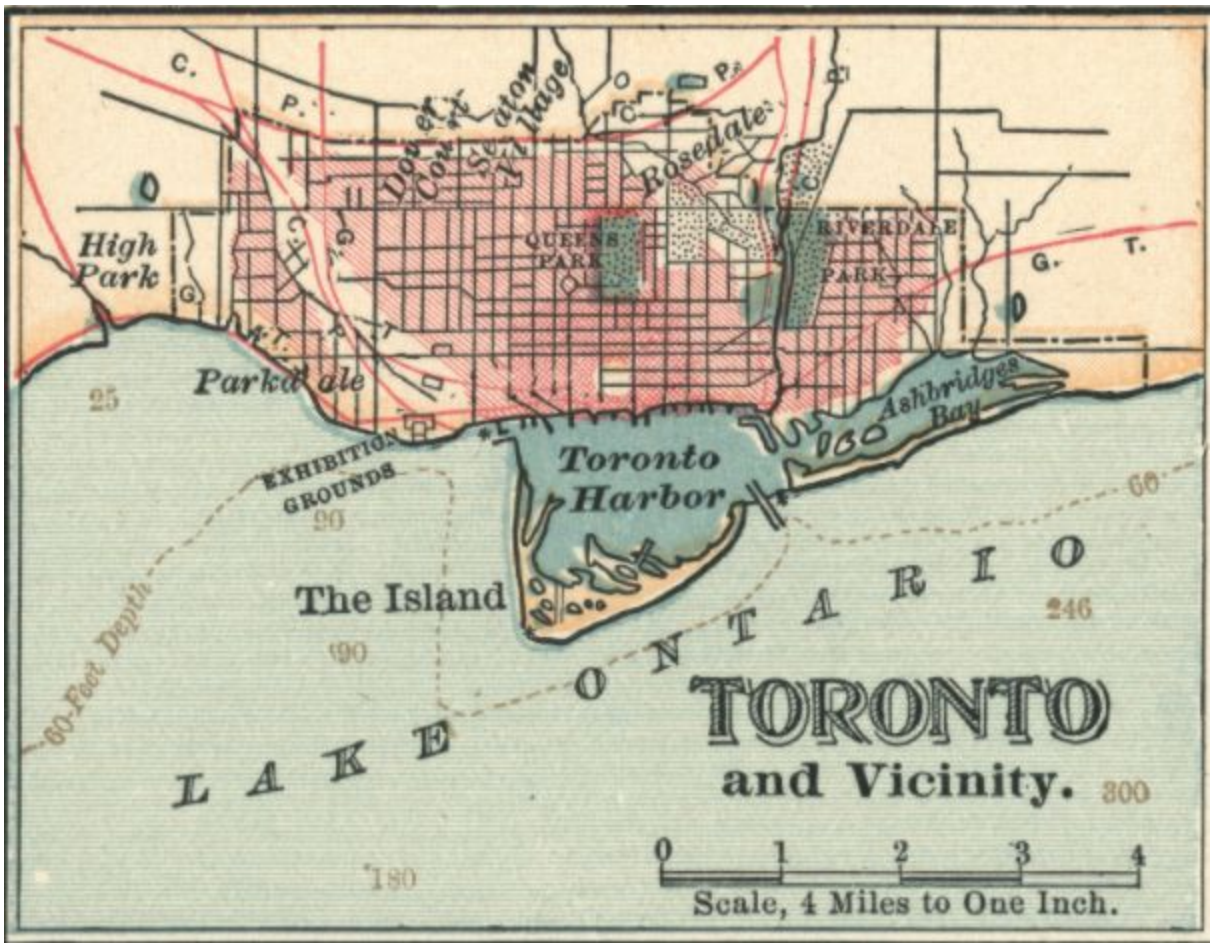


Table of Contents

Introduction - Identify the problem

We have a Canadian based Italian restaurant chain looking to expand into Toronto. The chain is primarily based in the Western Province of Alberta and British Columbia. The chain has been a success. The owner would like to build on that success by expanding into Toronto, which is known for its diversity.

Our objective is to leverage data on Toronto to help determine the best location for the restaurant. The owner is trying to find a location with little competition and as many customers as possible.

Target Audience

Our target audience is our client, the owner of the restaurant chain. Key stakeholders are the owner and his senior management team for the restaurant.

Data Overview

We will use 3 sources of data for this problem:

1. A list of postal codes, boroughs and neighbourhoods in Toronto. This list will be sourced from Wikipedia. These data points provide the information about the neighbourhoods that will be required to analyze the locations.

   Link - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. Geographical coordinates for the neighbourhoods and respective postal codes are the second source of data. This is required to create a map of the neighbourhoods so we can visualise the neighbourhoods. This data is provided in a CSV formatted file.

3. Data from FourSquare is the final component we require. The venue data is necessary to understand what venues and categories of venues are available in each neighbourhood in Toronto. This will help us determine the best location for the restaurant. This data will be combined with data from numbers one and two to help visualise and cluster the data points. API calls will be used to access the data from FourSquare.

Methodology

Data Cleansing and Preprocessing

With the data collected, we worked through a number of normalization processes with it. This included parsing the data into python, cleansing and grouping the data and visualising the data to understand it better.

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

An assumption was made that any row with an empty borough would not be processed from the data we collected in Wikipedia. Since more than one neighbourhood can exist in a postal code we normalised this by ensuring the neighbourhoods were combined into one row.

	Postal Code	Borough	...	Latitude	Longitude
0	M3A	North York	...	43.753259	-79.329656
1	M4A	North York	...	43.725882	-79.315572
2	M5A	Downtown Toronto	...	43.654260	-79.360636
3	M6A	North York	...	43.718518	-79.464763
4	M7A	Downtown Toronto	...	43.662301	-79.389494

[5 rows x 5 columns]
(103, 5)

That this produced 103 unique postal codes for us to analyze. We further explored the data to understand the number of neighbourhoods per borough and the results below were produced.

Borough	Neighbourhood
Central Toronto	9
Downtown Toronto	19
East Toronto	5
East York	5
Etobicoke	12
Mississauga	1
North York	24
Scarborough	17
West Toronto	6
York	5

The data was then merged with the geolocation data to get the latitude and longitude values for each postal code. The geolocation data is then used to collect venue data in FourSquare. The venue data was processed to extract the values we need such as:

- Venue Name
- Latitude Venue Location
- Longitude Venue Location
- Venue Category

```
[39]
print(toronto_venues.shape)
toronto_venues.head()
```

```
(2124, 7)
```

	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	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

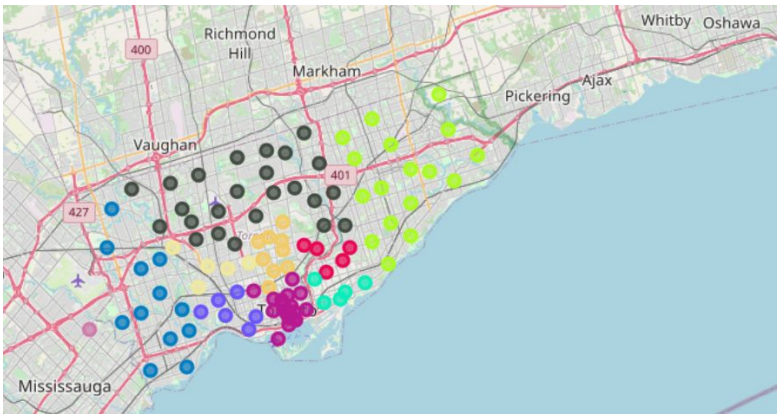
In total we got 2124 entries from FourSquare. We filtered these results to only have the venue category of ‘Italian Restaurants.

```
1485                          Parkdale, Roncesvalles  ...  Italian Restaurant
1520                                      Davisville  ...  Italian Restaurant
1524                                      Davisville  ...  Italian Restaurant
1554                  University of Toronto, Harbord  ...  Italian Restaurant
1591                             Runnymede, Swansea   ...  Italian Restaurant
1595                             Runnymede, Swansea   ...  Italian Restaurant
1620          Clarks Corners, Tam O'Shanter, Sullivan  ...  Italian Restaurant
1801                                 Stn A PO Boxes   ...  Italian Restaurant
1808                                 Stn A PO Boxes   ...  Italian Restaurant
1855                                 Stn A PO Boxes   ...  Italian Restaurant
1877                    St. James Town, Cabbagetown   ...  Italian Restaurant
1899                    St. James Town, Cabbagetown   ...  Italian Restaurant
2015           First Canadian Place, Underground city  ...  Italian Restaurant

[42 rows x 7 columns]
```

Results

We first take a look at the distribution of the data using the postal codes and the geolocation data. The map below shows a distribution of the postal codes.



Exploring the Data

We explored the data by neighbourhood groups to understand how many categories of venues exist in each neighbourhood. This provided a good understanding of how much commerce exists in a particular area.

Machine Learning

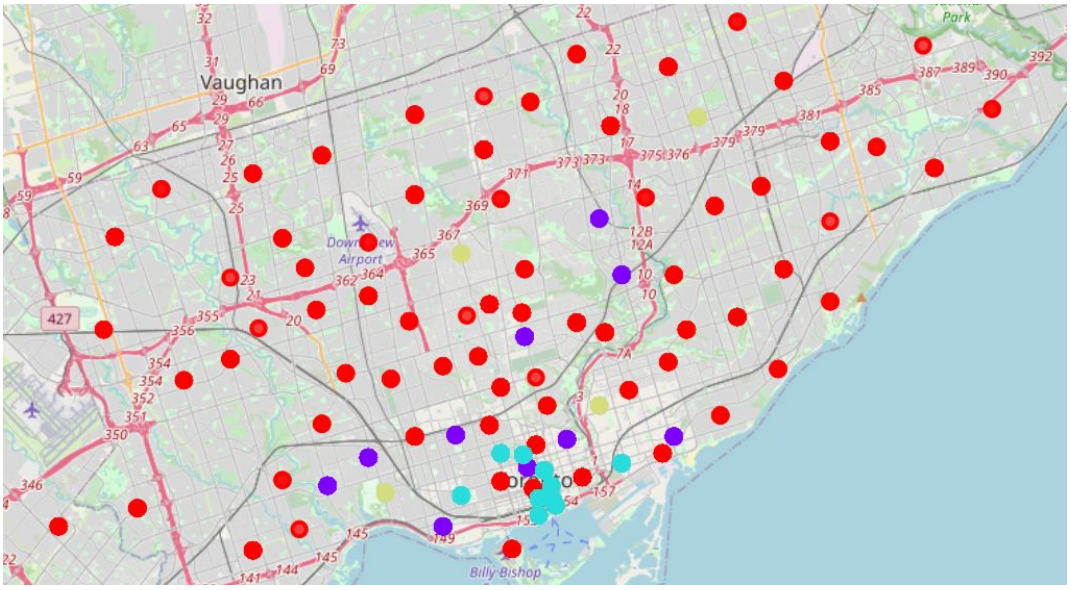
The data was transformed to numerical data for the venue category using a method called one hot encoding.

	Neighborhood	Italian Restaurant
0	Parkwoods	0
1	Parkwoods	0
2	Victoria Village	0
3	Victoria Village	0
4	Victoria Village	0

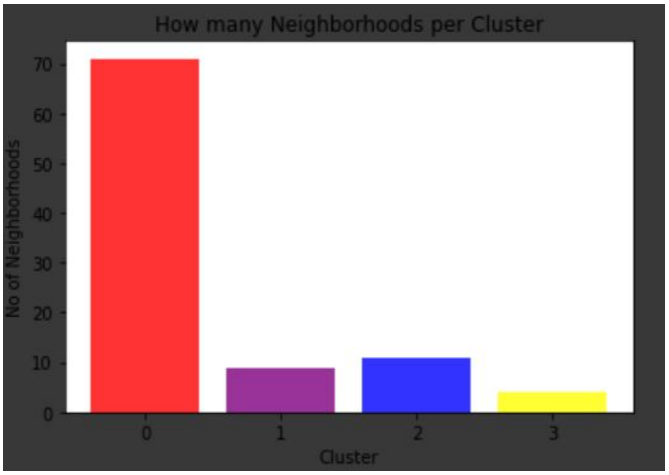
This is required for machine learning algorithms. Each neighborhood's individual venues were turned into the frequency of how many of those venues were located in each neighborhood. These were then grouped by neighborhoods and the mean taken for use in the algorithm.

K-Means clustering was the algorithm used for this analysis. This algorithm looks at the data points and clusters the points together based on how close they are to another point. The optimum K value for the model was chosen as 4.

After the model completed and we had the clusters for each data point, these clusters were then merged back with the original data to understand the distribution of the clusters. The map shows how this looked. This shows Italian restaurants that had a similar mean frequency.



The chart below shows how many neighborhoods per cluster. As you can see, cluster one has the largest number of neighborhoods and Italian restaurants.



Red - Cluster 0
Purple - Cluster 1
Blue - Cluster 2
Yellow = 3

Drilling into this a little deeper, we can see that the average number of Italian restaurants per cluster is as follows:

This shows that cluster 0 does not have a lot of Italian restaurants, while in cluster 3 we have a lot more on average.

Cluster Labels	
0	0.000135
1	0.048285
2	0.024896
3	0.073571

Further breakdown of each cluster can be found below.

Cluster 0

This cluster had over 74 neighbourhoods and a very low average for Italian restaurants.

	Neighborhood	...	Italian Restaurant
33	First Canadian Place, Underground city	...	0.01
0	Agincourt	...	0.00
73	Runnymede, The Junction North	...	0.00
70	Roselawn	...	0.00
69	Rosedale	...	0.00
..	...	...	...
36	Glencairn	...	0.00
34	Forest Hill North & West, Forest Hill Road Park	...	0.00
32	Fairview, Henry Farm, Oriole	...	0.00
31	Eringate, Bloordale Gardens, Old Burnhamthorpe...	...	0.00
98	York Mills West	...	0.00

[74 rows x 3 columns]

Cluster 1

This cluster has less neighbourhoods than cluster 0 but showed that the average number of Italian restaurants were higher.

	Neighborhood	...	Italian Restaurant
79	Stn A PO Boxes	...	0.031250
39	Harbourfront East, Union Station, Toronto Islands	...	0.030000
87	Toronto Dominion Centre, Design Exchange	...	0.030000
66	Queen's Park, Ontario Provincial Government	...	0.027778
88	University of Toronto, Harbord	...	0.027778
80	Studio District	...	0.025000
76	St. James Town	...	0.023529
52	Little Portugal, Trinity	...	0.021277
18	Commerce Court, Victoria Hotel	...	0.020000
35	Garden District, Ryerson	...	0.020000
5	Berczy Park	...	0.017241

[11 rows x 3 columns]

Cluster 2

This cluster has a similar distribution of neighbourhoods and average Italian restaurants to cluster 1

	Neighborhood	...	Italian Restaurant
5	Berczy Park	...	0.017241
18	Commerce Court, Victoria Hotel	...	0.020000
35	Garden District, Ryerson	...	0.020000
39	Harbourfront East, Union Station, Toronto Islands	...	0.030000
52	Little Portugal, Trinity	...	0.021277
66	Queen's Park, Ontario Provincial Government	...	0.027778
76	St. James Town	...	0.023529
79	Stn A PO Boxes	...	0.031250
80	Studio District	...	0.025000
87	Toronto Dominion Centre, Design Exchange	...	0.030000
88	University of Toronto, Harbord	...	0.027778

[11 rows x 3 columns]

Cluster 3
This cluster had the smallest number of neighbourhoods but the largest average number of Italian restaurants.

	Neighborhood	Borough	Italian Restaurant
4	Bedford Park, Lawrence Manor East	North York	0.080000
16	Clarks Corners, Tam O'Shanter, Sullivan	Scarborough	0.071429
63	Parkdale, Roncesvalles	West Toronto	0.071429
84	The Danforth West, Riverdale	East Toronto	0.071429

Discussion

From the analysis conducted we can summarise our findings as follows:

- In cluster 0 we can see that this has the smallest presence for Italian restaurants on average but the largest number of neighbourhoods.
- In cluster 1 we seem to have most of the Italian restaurants.
- Cluster 2 has similar presence to cluster 1
- Cluster 3 has a significant concentration of restaurants

From this data it looks like having a restaurant in cluster 0 may be the best option for opening the new restaurant. With no real presence of Italian restaurants in this cluster it should attract customers. Additional data may be required to further refine the data such as demographics, realestate data etc but given what we have on our data cluster 0, in neighbourhoods such as Fairview, Rosedale etc are good candidates.

Conclusion

This report should provide the high level results required by management. Additional data and analysis may be necessary to look at other factors such as demographics, real estate etc to help understand profitability etc. The tools used in this analysis were python libraries and other open source technologies. Other machine learning strategies may be applicable if additional data is included in the analysis to provide further insights and help drive decisions.