Battle of the Neighbourhoods

The problem

Our client, Canadian based Italian restaurant chain wants to expand into Toronto. The chain is primarily based in the Western Province of Alberta and British Columbia , with 7 restaurants. 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 used 3 sources of data, these are:

- Wikipedia dataset with postal codes, boroughs and neighbourhoods in Toronto.
  Link - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Geolocation coordinates for the neighbourhoods and the respective postal codes. This was provided in CSV format.
- Data on venues from FourSquare. This was important to understand what vensures and categories of venues are in which neighbourhoods.

Methodology - Data Cleaning and Preprocessing

A number of normalisation activities were completed on the data.

- Scrapped the from Wikipedia's site using pandas in python.
- Data points with 'Not assigned' Boroughs were removed.
- Reviewed to ensure each row was a unique postal code and the row included all the neighbourhoods for that code.

	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

Methodology - Data Cleaning and Preprocessing

- The postal code, borough and neighbourhood data was then merged with the geolocation data to get the latitude and longitude points.
- Validation on the number of rows in the data was done to show we had 103 rows in the dataset.

```
  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)
```

Methodology - Data Cleaning and Preprocessing

- The final processing step merged the FourSquare data with the existing dataset we processed.
- We first got the venue name, venue geolocation and venu category data for all venues in Toronto.

```
[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

Methodology - Data Cleaning and Preprocessing

The venue data was cleansed and stripped of data points we don’t need. It was then filtered for only the Italian Restaurants using the venue category column.

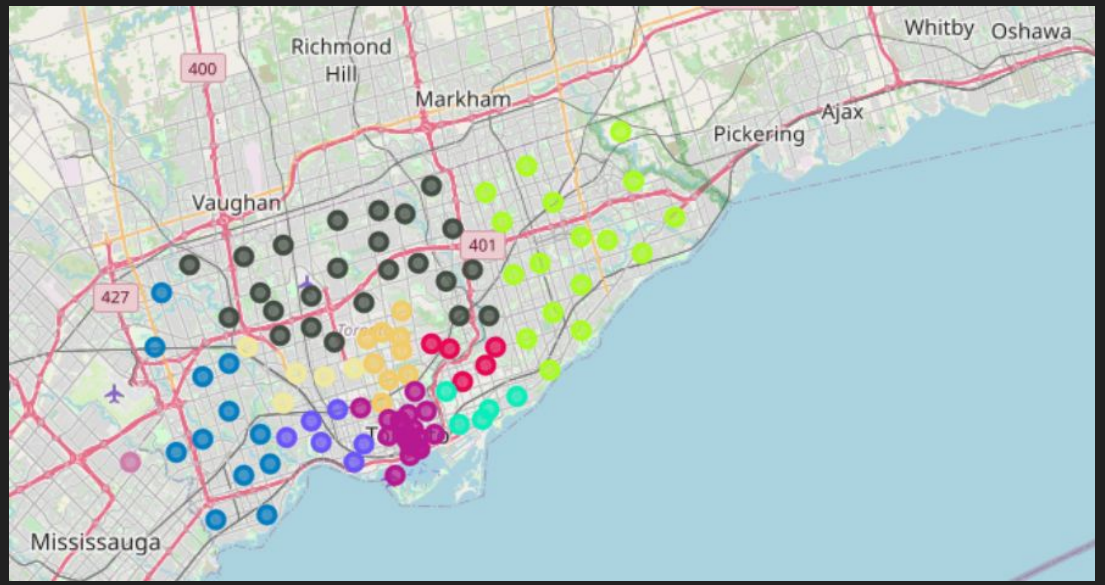
This produced 42 restaurants.

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 - visualise the neighbourhoods

Looking at the neighbourhoods by borough, we can see the concentrations of the neighbourhood to each other.



Machine Learning

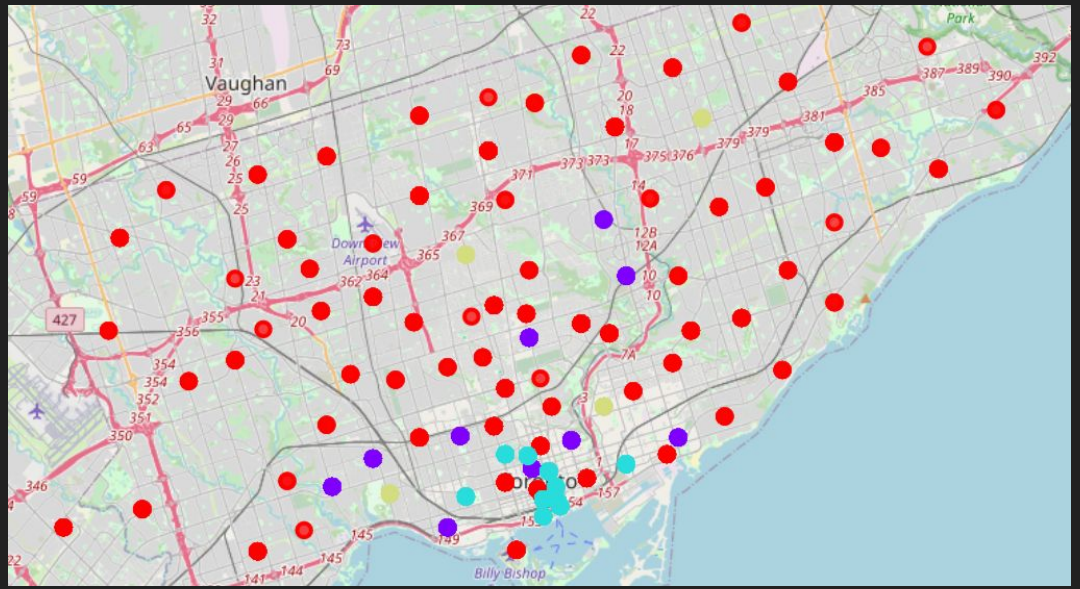
K-Means clustering was used to understand how close together Italian Restaurants are clustered together.

Before running K-Means we have to convert categorical values to numeric values. This was done using one hot encoding. Example of what this looks like is below.

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

Machine Learning - cont’d

We used 4 clusters for K-Means clustering and executed the process. The map below shows the clusters on a map.



- Red - Cluster 0
- Purple - Cluster 1
- Blue - Cluster 3
- Yellow - Cluster 4

Machine Learning - Summary results

The summary results showed that on average cluster 3 has the most amount of Italian restaurants.

Cluster Labels	
0	0.000135
1	0.048285
2	0.024896
3	0.073571

Cluster 3 results:

	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

Machine Learning - Cluster 0

	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]

Machine Learning - Cluster 1

	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]

Machine Learning - Cluster - 2

	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]

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

This suggest that Cluster 0 may be the best option for opening the new restaurant. 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.