Identifying Locality Types in Toronto

Contents:

Introduction/Business Problem

This project attempts to identify potential business opportunities for restauranteurs in the city of Toronto, Canada. An analytical approach leveraging machine learning techniques such as clustering and data analysis have been used. Wherever applicable, data visualization techniques have been used for better interpretation of results.

Data transformations were performed to derive the best form of data to be fed as inputs to the machine learning model. Following the data preparation and model training, locations were clustered into one of the many possible locality types, such as residential, business district or recreation.

Data

This project consumes data from two sources:

1. Foursquare API

   Data from foursquare API helps identify the most popular restaurant venues. This makes it possible to understand how they are distributed across the city of Toronto and recognize new business opportunities.

2. Geocoder package

   This package provides latitude and longitude values for all neighborhoods in the City of Toronto. Data thus retrieved aids in data visualization (maps).

Methodology

Statistical exploration and data visualization have been used to conduct exploratory analysis of the data. Clustering approach has been used to identify business locations and other localities. In the process, prospective business opportunities for restauranteurs. K-Means algorithm was implemented in Python.

To get started with this experiment, data was scraped from the following wikipage that contains a list of postal codes for the city of Toronto.

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

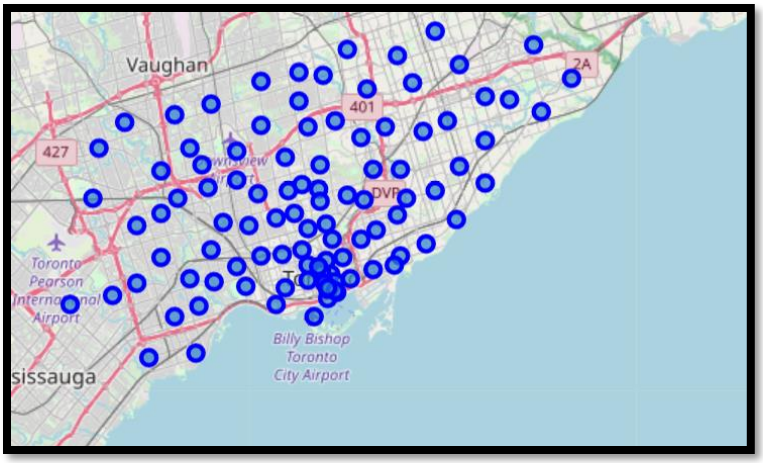
The scraped data was cleaned and inserted into a dataframe with 3 columns (Postcode, Borough, Neighborhood). A sample is shown below.

	Postcode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Geocoder package was used to pull latitude and longitude details for the neighborhoods in the previously created dataframe. After including the latitude and longitude details, we get the following dataframe:

	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

The data to be analyzed can be visualized using choropleth maps, where each blue dot represents a tuple in the above dataframe:



The foursquare API was used to retrieve venue details for each of the above locations. Popular venues along with their category were retrieved and the following "venues" data frame was created.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	Rouge, Malvern	43.806686	-79.194353	Caribbean Wave	43.798558	-79.195777	Caribbean Restaurant
2	Rouge, Malvern	43.806686	-79.194353	Staples Morningside	43.800285	-79.196607	Paper / Office Supplies Store
3	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
4	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant

One-hot encoding was used to transform the above data frame to a format more suitable for the machine learning model K-Means clustering. Moreover, the data was filtered out to include only restaurants. Encoding and filtering generates the following "venue_onehot" dataframe:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Lounge	American Restaurant	Amphitheater	...	Video Store	Vietnamese Restaurant
0	Rouge, Malvern	43.806686	-79.194353	0	0	0	0	0	0	0	...	0	0
1	Rouge, Malvern	43.806686	-79.194353	0	0	0	0	0	0	0	...	0	0
2	Rouge, Malvern	43.806686	-79.194353	0	0	0	0	0	0	0	...	0	0
3	Rouge, Malvern	43.806686	-79.194353	0	0	0	0	0	0	0	...	0	0
4	Rouge, Malvern	43.806686	-79.194353	0	0	0	0	0	0	0	...	0	0

The frequency of each venue category for all locations was calculated to create the below dataframe, which will be fed as input to the clustering algorithm:

	Neighborhood Latitude	Neighborhood Longitude	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Lounge	American Restaurant	Amphitheater	Animal Shelter	...	Video Store	Vietnamese Restaurant	Warehouse Store
0	43.602414	-79.543484	0.0	0.000000	0.0	0.0000	0.0000	0.000000	0.0	0.0	...	0.0	0.0	0.0
1	43.605647	-79.501321	0.0	0.000000	0.0	0.0000	0.0000	0.050000	0.0	0.0	...	0.0	0.0	0.0
2	43.628841	-79.520999	0.0	0.017241	0.0	0.0000	0.0000	0.034483	0.0	0.0	...	0.0	0.0	0.0
3	43.628947	-79.394420	0.0	0.000000	0.0	0.0625	0.0625	0.000000	0.0	0.0	...	0.0	0.0	0.0
4	43.636258	-79.498509	0.0	0.000000	0.0	0.0000	0.0000	0.000000	0.0	0.0	...	0.0	0.0	0.0

Before training the model, the optimal number of clusters was identified to be 2 using the metric "Silhoutte Coefficient"

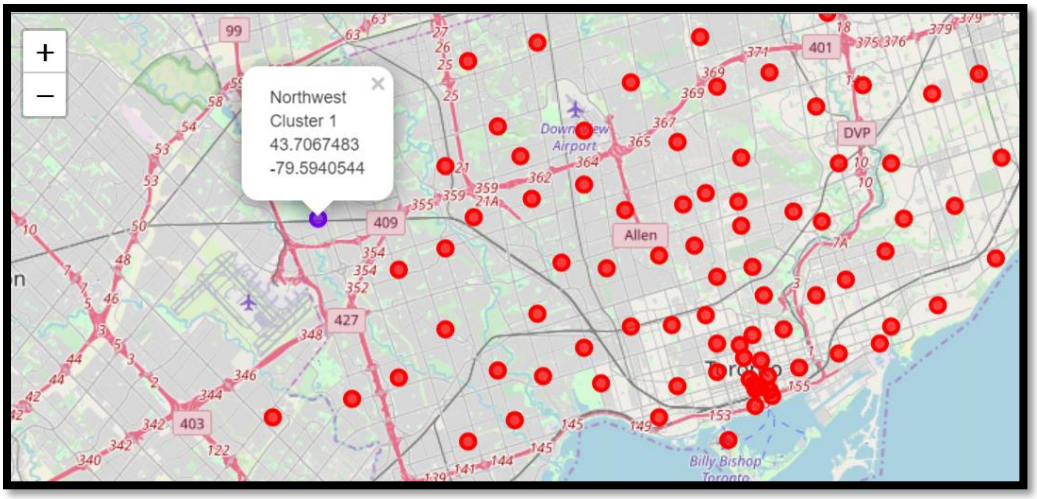
```
For n_clusters=2, The Silhouette Coefficient is 0.755766447754
For n_clusters=3, The Silhouette Coefficient is 0.344986005479
For n_clusters=4, The Silhouette Coefficient is 0.0950850090063
For n_clusters=5, The Silhouette Coefficient is 0.0962752875768
For n_clusters=6, The Silhouette Coefficient is 0.105353802099
For n_clusters=7, The Silhouette Coefficient is 0.0924335909059
For n_clusters=8, The Silhouette Coefficient is 0.111104959923
For n_clusters=9, The Silhouette Coefficient is 0.11232141954
```

Tuples along with their corresponding cluster, were stored to clustered_df dataframe:

Tapas Restaurant	Thai Restaurant	Theme Restaurant	Tibetan Restaurant	Turkish Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Cluster
0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0
0.0	0.025641	0.0	0.0	0.0	0.0	0.0	0
0.0	0.000000	0.0	0.0	0.0	0.0	0.5	0
0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0
0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0

Results

We observe that the localities are clustered into 2 and cluster 1 has only one location "Lawrence Park, Central Toronto". Upon further analysis, we find that there are no popular restaurants in this location (frequency zero for all restaurant categories).



This location could be a potential business opportunity for restauranteurs.

Cluster 1:

- Location with 0 frequency for restaurants

Cluster 2:

- Other locations with at least one

Discussion

It is an interesting finding that there is a place without many restaurants. It will be a potential business opportunity for restaurants.

However, before we draw any conclusions, to make good predictions, we need to take several factors into account. Factors such as other data sources and not just the foursquare API should be considered. Moreover, there could be restaurants that didn't show up in online databases. It is also possible, that this place has very few inhabitants with less demand for restaurants. More market analysis/research needs to be done before making a business decision.

Conclusion

It can be observed that the restaurant chains are densely populated in the downtown area and sparsely populated on the suburbs. Also, we were able to identify at least one neighborhood with potential business opportunity for restauranteurs. However, this result needs to be taken with a grain of salt, since we have relied on a single data source for venue details and as mentioned in the discussion section, there could be restaurants in the locality that weren't captured in data sources for numerous reasons.

Note to the reviewer:

Thank you for your time and patience.