Clustering the Toronto Neighbourhoods based on the food diversity

1. Introduction

1.1 Background

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 as of 2016.

The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. While the majority of Torontonians speak English as their primary language, over 160 languages are spoken in the city.

Food plays a major role in sorting out various people/cultures within a city as people prefer various cuisines based on their culture andethnicity.

1.2 Problem

Given neighborhood details of the city of Toronto such as different venues, category of the venues can we cluster the neighborhoods based on the city's food diversity?

1.3 Interest

So this project can be used by various culture based vendors to start a new business based on the cultural preference in a specific neighborhood. Also someone who is looking to open a new restaurant in the neighborhoods of Toronto can use this analysis or model to understand the food preferences.

2. Data Acquisition and Cleaning

2.1 Data Sources

Neighborhood details of Toronto

The data set consists of the list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

The data set used for the below problem is available in wikipedia.’https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M’

Features:

Postal Code - Postal code of the Neighborhood.

Borough - Borough to which the neighborhood belongs.

Neighborhood - Name of the Neighborhood

Geographical Coordinates

In order to explore the neighborhoods using the Foursquare API, we need the geographical coordinates of neighborhoods. This data is available in https://cocl.us/Geospatial_data.

Foursquare API

After retrieving the data set from Wikipedia and formatted it, we will be using the Foursquare (location data provider) API to explore each neighborhoods and cluster them based on the food practices/diversity.

Some of the features retrieved using Foursquare API:

Categories - Category of the venue

Category ID - unique ID for the category

Name - Name of the venue and so on.

2.2 Data Cleaning

The data is scraped from the Wikipedia page using the web scraping python tool namely

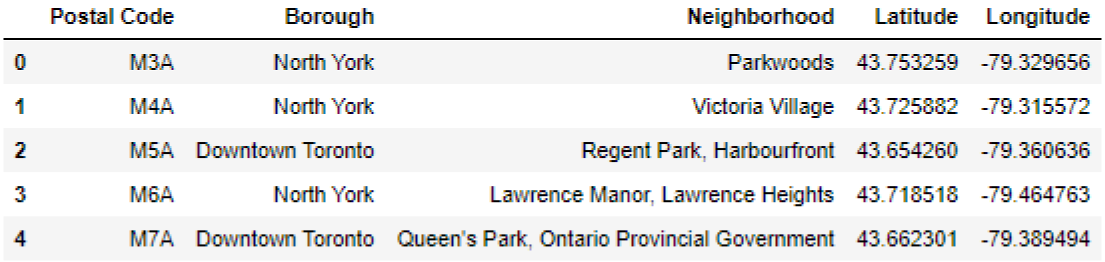
Beautiful soup.

Once the data is loaded into the notebook using Beautiful soup, it is wrangled/cleaned using the Pandas library. The data frame is cleaned in such a way that the resulting data frame has only the columns “Postal Code”, “Borough” and “Neighborhood”.

The rows with the Boroughs “Not assigned” were dropped from the data frame.

Then the location coordinates of the neighborhoods were downloaded from the csv file and merged with the existing data frame using the “Postal Code”.

The final data frame looks like below,



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Image 1 - Final Data Frame

The neighborhoods of Toronto has been visualized using the Folium map.

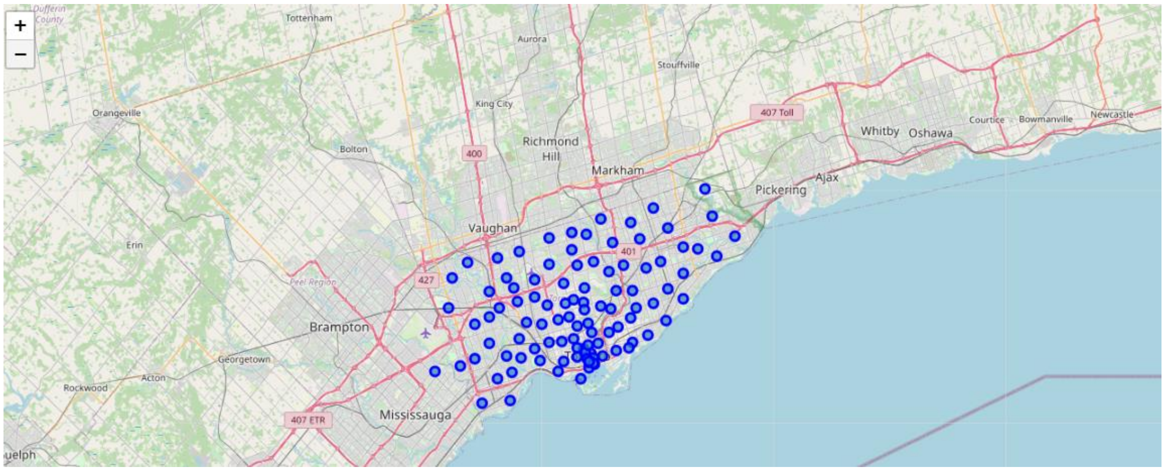
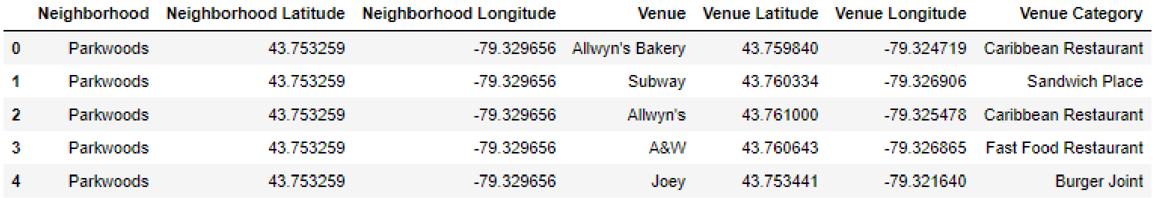


Image 2 - Folium Map

3. Foursquare API

In order to retrieve the unique Category Id for the food Category, the unique categories from the foursquare API is retrieved using the foursquare developer credentials and stored into a list.

Then using the Category Id of the food category, the top 500 venues within the radius of 1000 is retrieved using foursquare API and stored as a data frame.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.753259	-79.329656	Subway	43.760334	-79.326906	Sandwich Place
2	Parkwoods	43.753259	-79.329656	Allwyn's	43.761000	-79.325478	Caribbean Restaurant
3	Parkwoods	43.753259	-79.329656	A&W	43.760643	-79.326865	Fast Food Restaurant
4	Parkwoods	43.753259	-79.329656	Joey	43.753441	-79.321640	Burger Joint

Image 3 - Top Food Category Venues

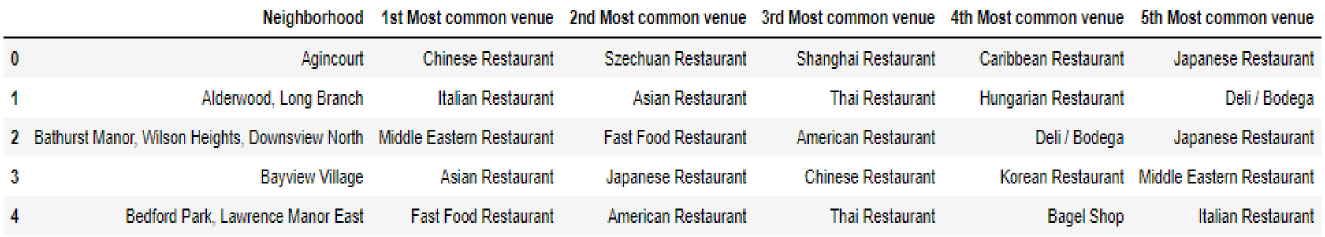
As the model is used to cluster the neighborhoods of Toronto based on the food diversity, the generalized food venues such as cafe, Restaurant were removed from the data frame.

4. Modelling

One hot encoding was done to the data frame using the pandas get_dummies method as we want to analyze each neighborhood based on the food category.

Then a new data frame was created by grouping the neighborhoods by the frequency of occurrence of the venue category.

The data frame is then sorted in descending order and to see the top five most common food venues in the neighborhood.



	Neighborhood	1st Most common venue	2nd Most common venue	3rd Most common venue	4th Most common venue	5th Most common venue
0	Agincourt	Chinese Restaurant	Szechuan Restaurant	Shanghai Restaurant	Caribbean Restaurant	Japanese Restaurant
1	Alderwood, Long Branch	Italian Restaurant	Asian Restaurant	Thai Restaurant	Hungarian Restaurant	Deli / Bodega
2	Bathurst Manor, Wilson Heights, Downsview North	Middle Eastern Restaurant	Fast Food Restaurant	American Restaurant	Deli / Bodega	Japanese Restaurant
3	Bayview Village	Asian Restaurant	Japanese Restaurant	Chinese Restaurant	Korean Restaurant	Middle Eastern Restaurant
4	Bedford Park, Lawrence Manor East	Fast Food Restaurant	American Restaurant	Thai Restaurant	Bagel Shop	Italian Restaurant

Image 4 - Top Five Most Common Food Venues

4.1 K Means Clustering

The Neighborhood column is dropped from the data frame before clustering as the K means clustering is mainly based on the distance between the clusters and the distance calculation for a variable of type object is not valid.

Elbow method was carried out in order to fix the initial number of clusters for the K means clustering algorithm.

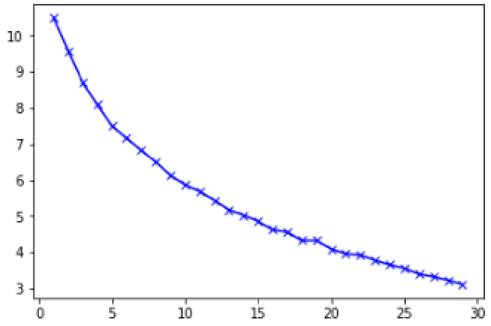
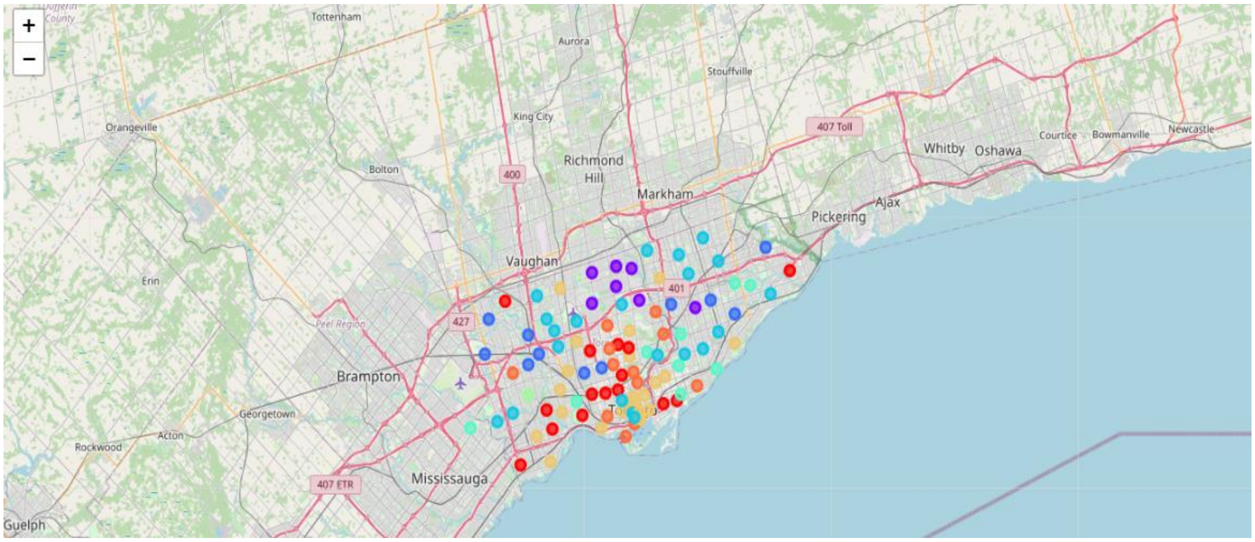


Image 5 - Clusters

The initial number of clusters was selected as 8 and the modeling was done.

Folium map was used to visualize the clusters.



5. Conclusion

In this model, the neighborhoods of Toronto were clustered based on the diversity in the food practice. I used K means clustering method to cluster the neighborhoods.

This clustering helps new food vendors and other culture based vendors to start new restaurant/ business in the neighborhoods of Toronto based on the food practices in the neighborhoods.