Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Introduction

Background

Toronto is Canada's largest city, the fourth largest in North America, and home to a diverse population of about 2.8 million people. It is a global centre for business, finance, arts and culture and is consistently ranked one of the world's most livable cities.

Problem

When you are looking to open a restaurant in a popular city as Tonronto city, how to build a successful restaurant. Of course, food and service are important to the success of a restaurant, but the location can be just as crucial. Therefore, target audience of this project will be people who are looking to open a new restaurant. This project will segment the neighborhoods of Toronto into major clusters and examine their food. This quantifiable analysis can be used to understand the distribution of different cultures and food over Canada's largest city. Also, it can be utilized by a new **food vendor** who want to open his or her restaurant or by a **government authority** to examine and study their city’s culture diversity better.

Data

Toronto City Dataset

Data will be scraped from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. After Toronto City data is scraped, data will be preprocessed. Data is consist of **Post Code**, **Borough**, and **Neighborhood**.

Example of Toronto City Dataset

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Downtown Toronto	Queen's Park

Geographical Coordinates

Toronto City data will be mapped with the geographical coordinates of each postal code of Toronto City. Geographical Coordinates data is consist of **Post Code**, **Latitude**, and **Longitude**. Link: http://cocl.us/Geospatial_data

Example of Geographical Coordinates

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Foursquare API

Foursquare API, a location data provider, will be used to find the venues on each postal code zone using a radius based on the area cover by each neighborhoods. Data from Foursquare API is consist of **Venue Name**, **Venue Latitude**, **Venue Longitude**, and **Venue Category**.

Example of Foursquare API

Venue	Venue Latitude	Venue Longitude	Venue Category
Brookbanks Park	43.751976	-79.332140	Park
Variety Store	43.751974	-79.333114	Food & Drink Shop
Victoria Village Arena	43.723481	-79.315635	Hockey Arena
Tim Hortons	43.725517	-79.313103	Coffee Shop
Portugril	43.725819	-79.312785	Portuguese Restaurant

Methodology

Data Cleaning

Because of our objective is to understand the distribution of different cultures and food, so we have to remove all the venues which is generalized categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
5	Victoria Village	43.725882	-79.315572	The Frig	43.727051	-79.317418	French Restaurant
7	Harbourfront	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery

Feature Engineering

Using one hot encoding to convert categorical variables which are venue categories into a form that could be provided to ML algorithms to do a better job in prediction.

	Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Bar	Beer Store	Belgian Restaurant	...
1	Parkwoods	0	0	0	0	0	0	0	0	0	...
3	Victoria Village	0	0	0	0	0	0	0	0	0	...
4	Victoria Village	0	0	0	0	0	0	0	0	0	...
5	Victoria Village	0	0	0	0	0	0	0	0	0	...
7	Harbourfront	0	0	0	0	1	0	0	0	0	...

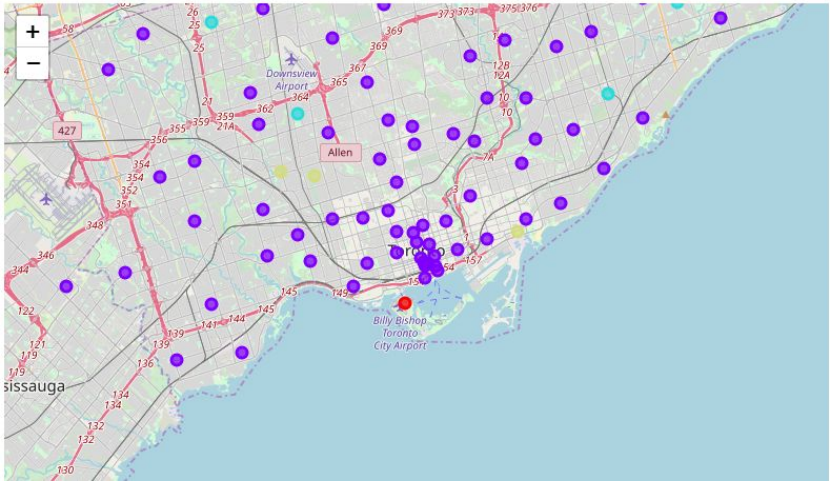
Feature Engineering

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

	Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Bar	Beer Store	Belgian Restaurant	...
0	Adelaide, King, Richmond	0.0	0.032787	0.04918	0.0	0.032787	0.065574	0.0	0.0	0.0	...
1	Agincourt	0.0	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.0	...
2	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.2	0.0	...
3	Alderwood, Long Branch	0.0	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.0	...
4	Bathurst Manor, Downsview North, Wilson Heights	0.0	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.0	...

Cluster Neighborhoods

- Using K-means with k=4



Results and Discussion

Coffee Shop is the most common venue across all the clusters or neighborhoods.

Cluster	1st Most Common Venue	2nd Most Common Venue	Neighborhood
0	Bar	Wine Shop	Highland Creek, Rouge Hill, Port Union
1	Coffee Shop	Coffee Shop	Queen's Park
2	Coffee Shop	Ethiopian Restaurant	Woburn
3	Fast Food Restaurant	Wine Shop	Caledonia-Fairbanks

Conclusion

In conclusion, the neighborhoods of Toronto City can be segmented into 4 clusters and upon analysis, it was possible to rename them basis upon the categories of venues in and around that neighborhood. Along with Coffee Shop, Fast Food Restaurant, Bar and Wine Shop are very dominant in Toronto City. This project can also be adjusted to use with other business.