**Recommending Neighborhoods in Toronto**

Yan Jiang

Jan 14, 2020

1. **Introduction**
   1. **Problem Statement**

Analyze Toronto postal code and location data and recommend top neighborhoods that match customer expectation.

* 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. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

Toronto is a prominent center for music, theatre, motion picture production, and television production, and is home to the headquarters of Canada's major national broadcast networks and media outlets. Its varied cultural institutions, which include numerous museums and galleries, festivals and public events, entertainment districts, national historic sites, and sports activities, attract over 43 million tourists each year. Toronto is known for its many skyscrapers and high-rise buildings, in particular the tallest free-standing structure in the Western Hemisphere, the CN Tower.

One of the tough questions when people moving into a new city is which neighborhood is best? In this project, I would like to create a hypothetical business scenario to analyze Toronto postal code and location data and recommend top neighborhoods that best match customer expectation.

* 1. **Audience**

The objective is to analyze neighborhoods of Toronto city. This would interest following audiences:

* City planning officers who want to see what the top business venues are in each neighborhood and plan/approve new city projects accordingly
* Business owners who want to start or move a business into Toronto city
* Anyone who wants to purchase or rent a home to live in Toronto city.
* Tourists who want to explore different lives in Toronto city.

1. **Data Description**

Following dataset will be used for this project:

* Use raw postal code data provided by Wikipedia,
* Geographical coordinates from https://cocl.us/Geospatial\_data
* Location data from FourSquare

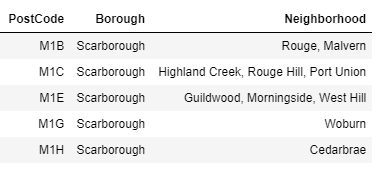
Wikipedia dataset contains a list of Canadian postal codes. The postal code data will be extracted from the Wikipedia article by scraping a table from the website. I will use this dataset to create a geographical segmentation of Toronto based on postal code and link it to coordinates found from https://cocl.us/Geospatial\_data. This dataset will be further enriched from location and venue data retrieved from Foursquare using RESTful API.

1. **Methodology**
   1. Data Collection/Cleaning/Processing

Postal Code data:

* Read the Wikipedia article and save to local copy
* Use BeautifulSoup API html.parser to parse the html, read into tables
* Identify header table containing Postcode, Borough, Neighborhood from Wikipedia article
* Extract Postcode, Borough, Neighborhood and write to semicolon-delimited file
* Load the semicolon-delimited file into pandas dataframe
* Ignore not assigned boroughs
* Assign borough value to neighborhood if the neighborhood is not assigned
* Group dataframe by PostCode and combine neighborhood values

At the end of this step, the postal data was in following shape:

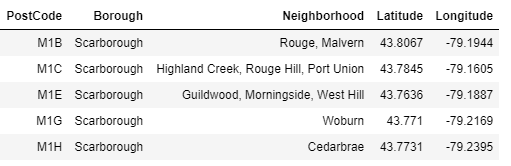


Geographical Coordinates

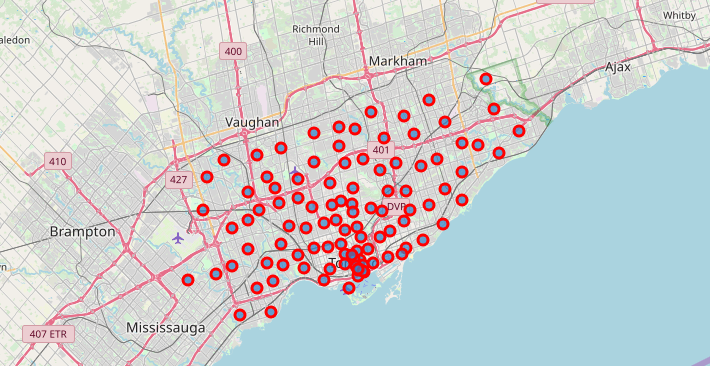
Geospatial\_data contains postal code coordinates. I will link postal code with coordinates.

* Load the csv file from source at <https://cocl.us/Geospatial_data>
* Extract latitude and longitude for postal codes
* Add coordinates information to the pandas dataframe containing postal code data.

At the end of this step, the postal data was in following shape:



Below is a map of the Toronto City with our postal code dataset:



Location Data

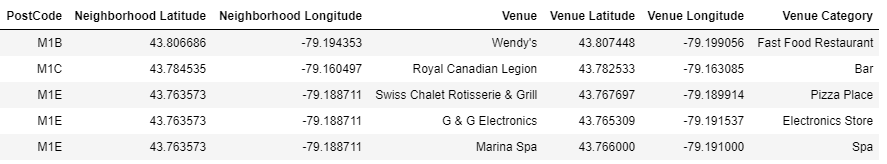
Foursquare is one of the leading providers for location data. I will use Foursquare RESTful API to retrieve venue data for Toronto city

* urllib.request is used to send requests to Foursquare
* Foursquare explore endpoint is used for requesting the venue data. Following is the request url format:

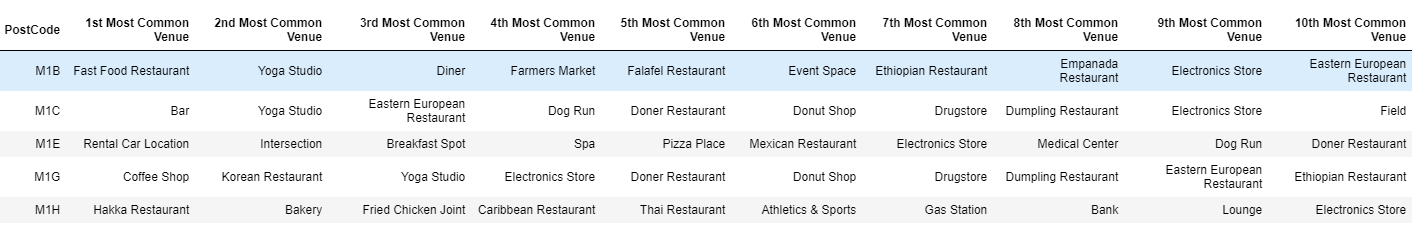
[https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}&ll={},{}&radius={}&limit={}](https://api.foursquare.com/v2/venues/explore?&client_id=%7b%7d&client_secret=%7b%7d&v=%7b%7d&ll=%7b%7d,%7b%7d&radius=%7b%7d&limit=%7b%7d)

* I developed a function for getting nearby venues by latitude and longitude for all postal codes postal code from the postal code data frame, with limit of 100 venues and radius of 300 per postal code.
* JSON data is received from the Foursquare endpoint. Venue name and venue categories have been parsed from the results.

Below is the sample data after combining postal code data with location data from Foursquare:



We can see the top 10 venues for each postal code:

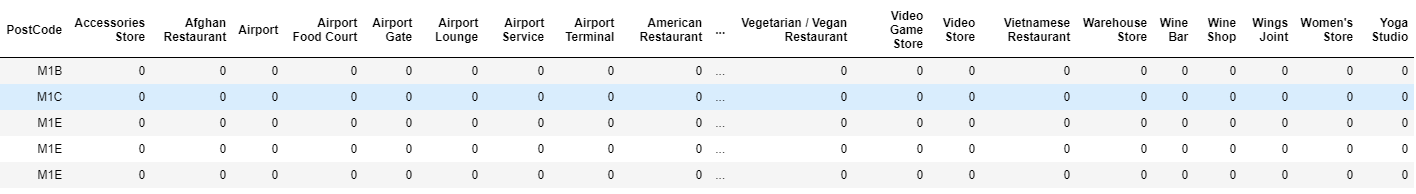


* 1. Modeling

The postal code data is an unsupervised data. For the purpose of this project, I used

**K-Means clustering** **algorithm** from sklearn clustering package to cluster the postal codes/neighborhoods.

In order to facilitate the K-Means clustering, one hot encoding was used to create dummy features for all venue categories:



Group rows by Post Code and take the mean of the frequency of occurrence of each category:



1. **Results**

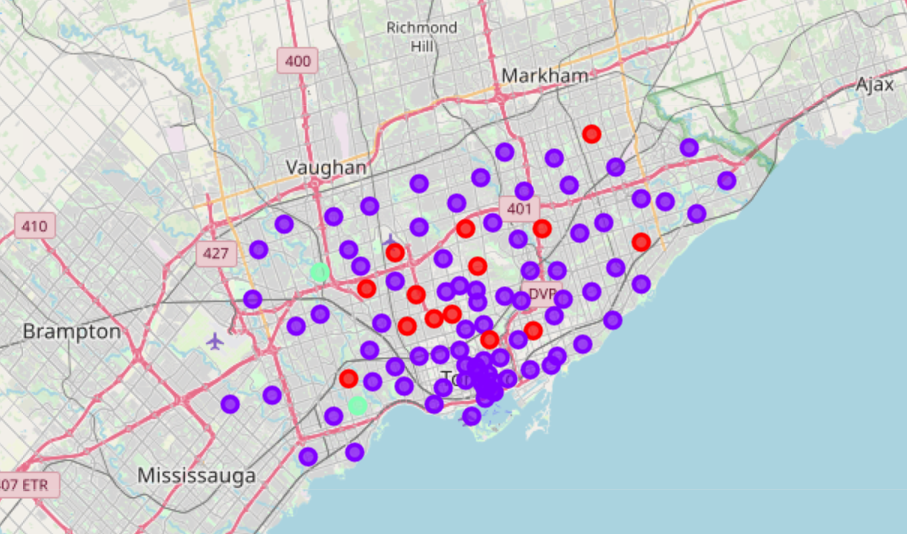
I’ve fit the data into different number of clusters, 3 was the optimal number of clusters found from testing.

NOTE: This small number of clusters was in part caused by insufficient venue data due to limitation of Foursquare personal account. With commercial account, this data should be more complete and more clusters maybe found.

Final data with postal code/neighborhoods assigned with cluster labels with 10 top venues for each postal code:



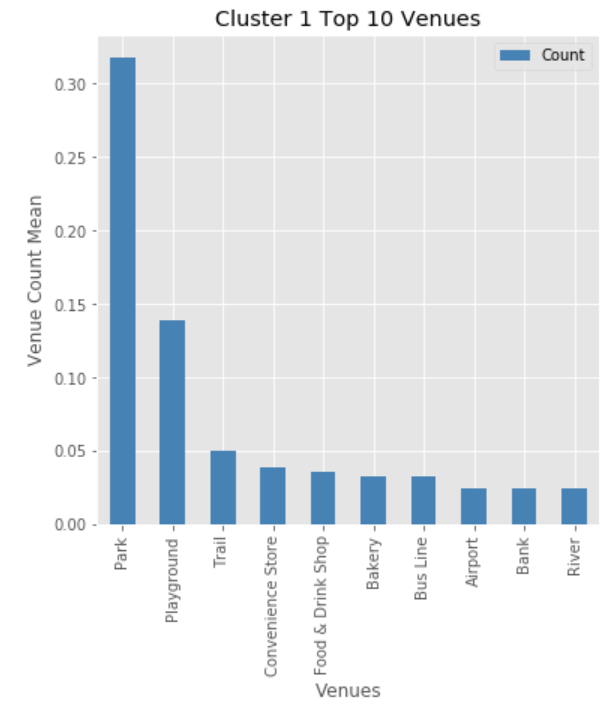
Below is a visualization of the clusters within Toronto city, with each cluster color coded, a click on each postal code would yield the corresponding postal code and cluster label:



Through data analysis and modeling, I’ve found 3 distinct clusters within city of Toronto, each with its unique business venues and life styles:

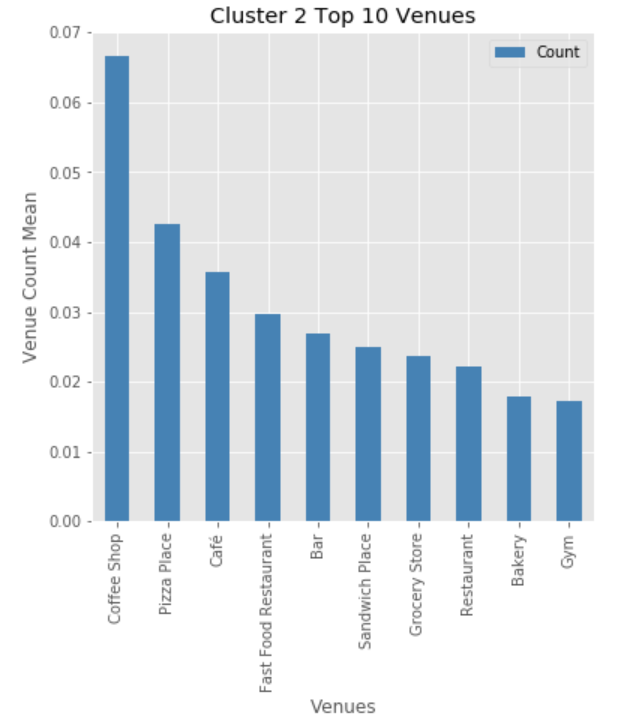
**Cluster 1: Recreation and Residential Area**

This cluster contains recreation venues such as parks, playgrounds, trails, convenient store, food and drink shops, bakery, and gym.



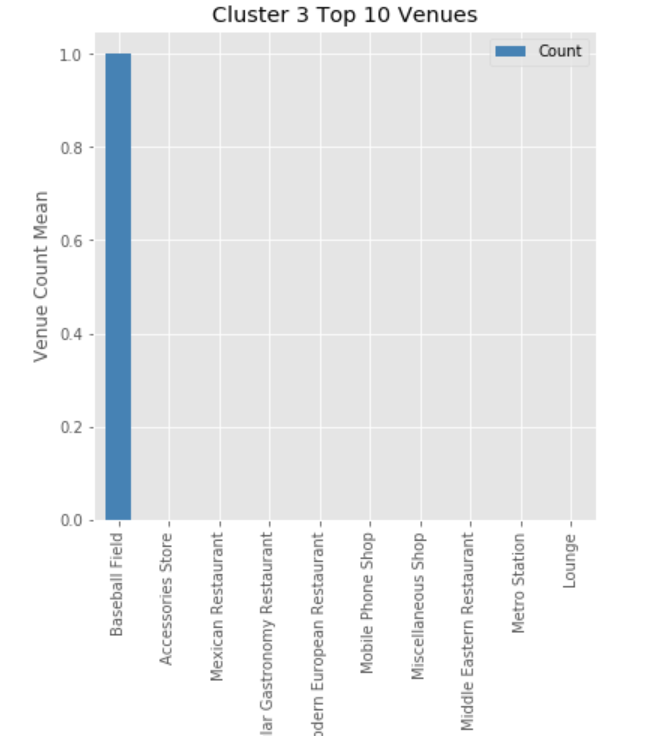
**Cluster 2: Restaurants Area**

This cluster is full of a variety of food venues, from coffee shop, pizza place, café, restaurant, bar, to grocery store, bakery, and gym.



**Cluster 3: Baseball Fields**

This cluster contains solely of one venue: Baseball Field. Even though there is only one venue category in this cluster, we can see the count value is very high. You can confirm this by comparing the same venue count mean values in Cluster 1 and 2.



1. **Discussion**

Data used in this project is limited. We could have more data such as housing data and labeled data. This would add more features and we can then use supervised regression models for further modeling of the Toronto neighborhoods.

Account used for connecting to Foursquare API is a free personal account. By upgrading to commercial account with full premium access, we can retrieve more venue data. As found in cluster 3, there is only one venue category. This was caused by lack of Foursquare location data.

1. **Conclusion**

In this study, I’ve analyzed the Toronto city postal code data and built K-Means classification model to identify 3 distinct types of business venues. This analysis model and results can be very useful in helping city managers for city planning, business owners for where to start business, people who’s purchasing/renting homes, and tourists looking for best places to visit and entertain in city of Toronto.

1. **References**

[Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

[Four Square API](https://developer.foursquare.com/)

[BeautifulSoup](https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

[Geospacial Data](https://cocl.us/Geospatial_data)