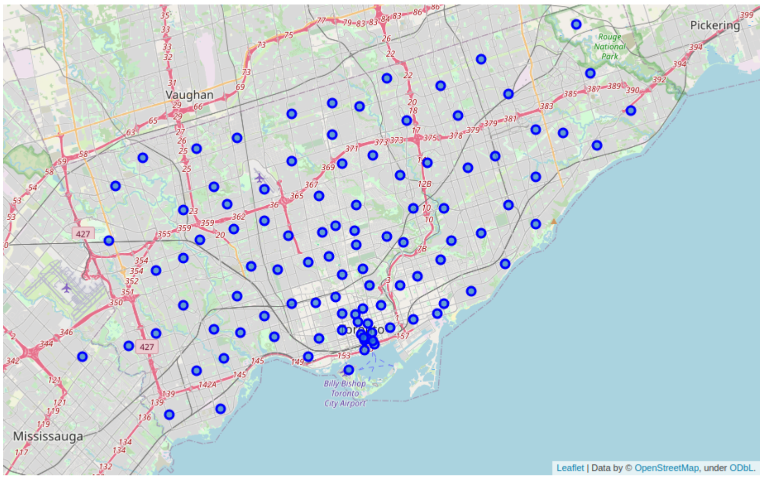
Coursera - IBM Data Science

Capstone Project

Clustering Neighborhoods in Toronto Using Crime and Location Data

Author:
Lynx Delta

Date of Publication:
16 June, 2020


Pickering
Vaughan
Rouge National Park
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

Abstract

This is the abstract...

Contents

1 Introduction

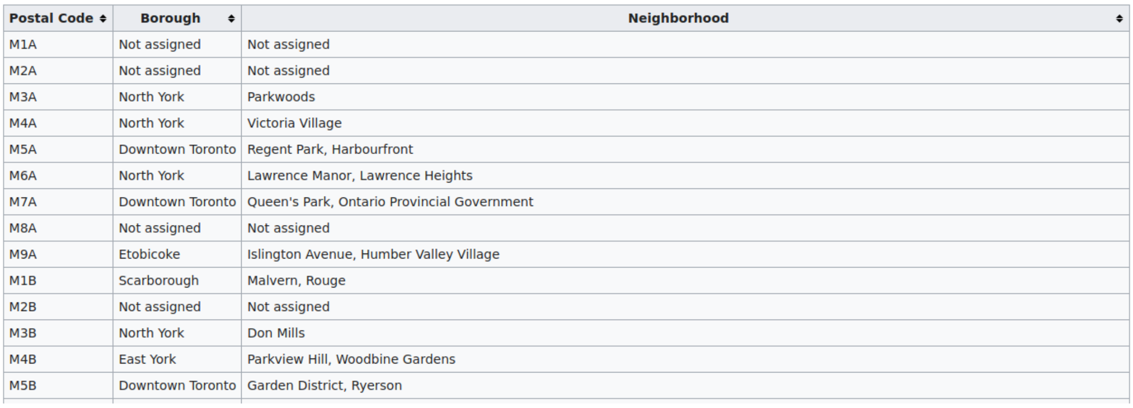
The population in the world is currently growing at a rate of around 1.05% per year and is expected to reach 7.8 billion people by the end of July 2020. More than 50% live in cities or urban areas [1]. It is assumed, that the higher the density of human beings at a certain place or in a certain area, the higher the probability of incidents (there may be other factors apart from people density). The aim of this analysis is to explore the venues (data from Foursquare) and the incidents (data from Toronto Police Service) in the neighborhoods of the City of Toronto, Canada. Is there a relationship between neighborhoods that share similar venues (e.g. airport, park, restaurant) and neighborhoods where similar incidents (e.g. traffic accident, robbery, murder) are reported? How does that compare to the population density? The result of the analysis could help the police improve respectively better target their monitoring in neighborhoods. If in a certain type of neighborhood (similar venues) a certain type of incident is likely to happen, the activity of the police may be adapted to the type of incident and the type of venue nearby (e.g. more traffic surveillance, lower the patrolling frequency).

2 Data

To answer the questions mentioned in the introductory section, data from Wikipedia (list of postal codes of Canada: M), geospatial coordinates data (provided by IBM), location data from Foursquare (venues), and crime data from Toronto Police Service (crime data by neighborhood) is used.

2.1 Description and Source

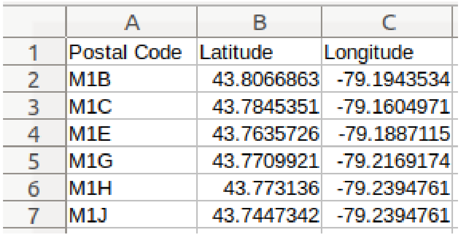
Postal codes, boroughs, and neighborhoods: The postal codes beginning with the letter M are located within the city of Toronto in the province of Ontario. The data is taken from the Wikipedia website [2] through web scraping. It is provided as a table that consists of 180 rows and three columns (”Postal Code”, ”Borough”, and ”Neighborhood”). Multiple postal codes have no particular entry for ”Borough” or ”Neighborhood” (indicated as ”Not assigned”). Furthermore, a borough may span multiple postal codes and may include multiple neighborhoods. If multiple neighborhoods share the same postal code, they are listed in the same row (column ”neighborhood”), separated by commas. The first 14 row entries are presented in Figure 1.



Postal Code	Borough	Neighborhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson

Figure 1: *Postal codes, boroughs, and neighborhoods*

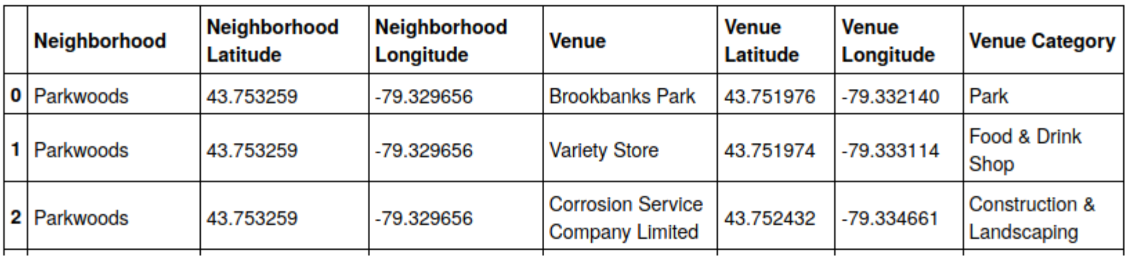
Geospatial coordinates: The geospatial coordinates data set is composed of 180 rows and three columns (”Postal Code”, ”Latitude”, and ”Longitude”). Each row has its distinct entry for the postal code, the latitudinal coordinate, and the longitudinal coordinate of the center of the corresponding borough (represented by the postal code). The data (*CSV*-file) is downloaded from the IBM cognitive class data server [5]. An excerpt of the data is depicted in Figure 2.



	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.8066863	-79.1943534
3	M1C	43.7845351	-79.1604971
4	M1E	43.7635726	-79.1887115
5	M1G	43.7709921	-79.2169174
6	M1H	43.773136	-79.2394761
7	M1J	43.7447342	-79.2394761

Figure 2: *Postal codes and geospatial coordinates*

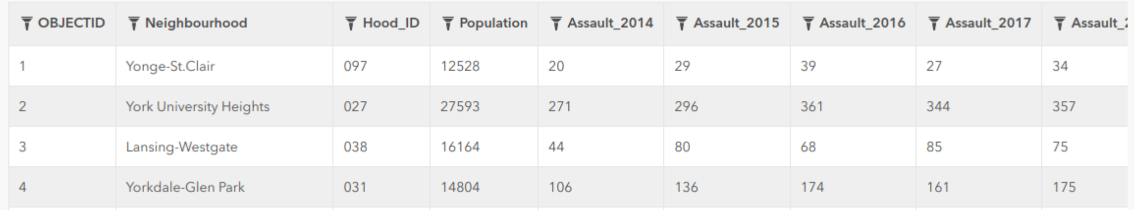
Location data: To get the location data (venues) in a certain radius of a neighborhood (represented through geospatial coordinates of the center of corresponding borough), the Foursquare API is utilized [3]. The data is obtained through a GET request (returned as *JSON*-file). The first three rows of the prepared data of a Foursquare request is shown in Figure 3.



	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	Parkwoods	43.753259	-79.329656	Corrosion Service Company Limited	43.752432	-79.334661	Construction & Landscaping

Figure 3: *Venues: Prepared result of Foursquare GET request*

Crime data: The Toronto neighborhoods boundary file includes 2014-2018 crime data by neighborhood. Counts are available for ”Assault”, ”Auto Theft”, ”Break and Enter”, ”Robbery”, ”Theft Over”, and ”Homicide”. The data set also includes four-year averages and crime rates per 100’000 people by neighborhood based on 2016 census population and is provided as a *CSV*-file from Toronto Police Service [4]. An excerpt of the data set in presented is Figure 4.



OBJECTID	Neighbourhood	Hood_ID	Population	Assault_2014	Assault_2015	Assault_2016	Assault_2017	Assault_:
1	Yonge-St.Clair	097	12528	20	29	39	27	34
2	York University Heights	027	27593	271	296	361	344	357
3	Lansing-Westgate	038	16164	44	80	68	85	75
4	Yorkdale-Glen Park	031	14804	106	136	174	161	175

Figure 4: *Crime data from Toronto Police Service*

2.2 Data Cleaning and Preparation

The raw data (*HTML*-file, *CSV*-file, *JSON*-file) is parsed and loaded into data frames using *Python* in combination with the following modules / packages: *bs4* (*bs4.Beautifulsoup* for *HTML* parsing), *json*, *numpy*, *pandas*, and *requests*. Furthermore, the data is cleaned and only important columns are kept for further analysis.

Postal codes, boroughs, and neighborhoods: If a neighborhood is not present in a row ("Not assigned"), the neighborhood gets the name of the borough. Rows with "Not assigned" values in the "Borough" and "Neighborhood" column are dropped. The final data frame consists only of the "Postal Code" and "Neighborhood" columns.

Geospatial coordinates: The geospatial coordinates data set is combined with the final neighborhoods data frame (inner join on "Postal Codes").

Location data: The data of all the GET requests (*JSON*-files) is prepared and combined (one final data frame with data of all neighborhoods).

Crime data: Only the "Neighborhood", the "Population", and the four-year averages ("AVG") column of each crime type is retained. The data frame is adjusted with the "Neighborhood" column of the final neighborhood data frame (including geospatial data).

3 Methodology

To follow...

3.1 Data Analysis

To follow...

3.2 Statistical Inference

To follow...

3.3 Algorithms

To follow...

4 Results

To follow...

5 Discussion

To follow...

6 Conclusion

To follow...

References

[1] https://www.worldometers.info/world-population/, accessed: 11/06/2020

[2] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, accessed: 12/06/2020

[3] https://developer.foursquare.com/, accessed: 12/06/2020

[4] http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-, accessed: 12/06/2020

[5] http://cocl.us/Geospatial_data, accessed: 12/06/2020