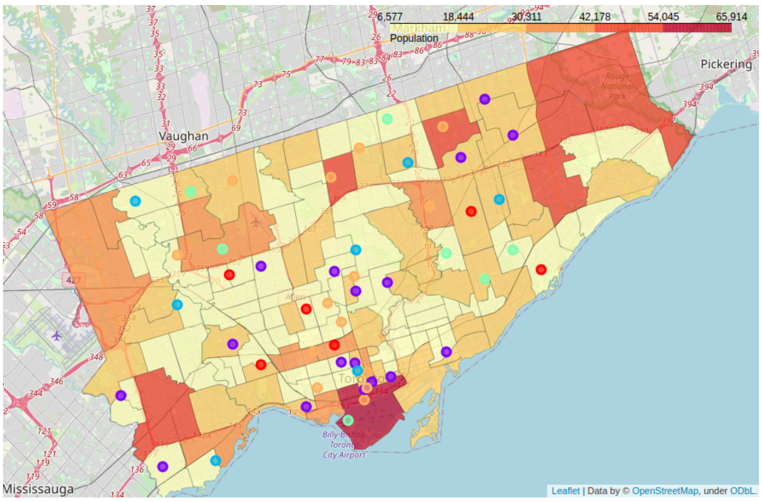
Coursera - IBM Data Science

Capstone Project

Clustering Neighborhoods in Toronto Using Crime and Location Data

Author:
Lynx Delta

Date of Publication:
22 June, 2020



Abstract

The aim of this analysis was 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).
The top ten venues per neighborhood were calculated and analyzed. Furthermore, neighborhoods were clustered into five groups using the location data (venues) and the crime data respectively. A separate map, displaying the population as well as a cluster marker per neighborhood, was created for the location and the crime data set. It could not be observed a correlation between neighborhoods with similar venues and neighborhoods with similar counts of the same type of crimes. There might be too many venues (features) in the location data set for meaningful clustering or an entire lack of correlation between venues and crimes in a particular neighborhood. However, a future respectively extended analysis could include other data sets (e.g. traffic flow, commuting rate etc.) and make use of other clustering algorithms or data transformations.

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.

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*

3 Methodology

This section aims at providing a short description of the steps performed during data cleaning and of the methods and algorithms applied for data analysis.

3.1 Data Cleaning and Preparation

The raw data sets (*HTML*-file, *CSV*-file, *JSON*-file) were 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 was cleaned and only important columns were kept for further analysis.

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

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

Location data: The data of all the GET requests (*JSON*-files) was prepared and combined (one final data frame with data of all neighborhoods). Furthermore, the venue categories (e.g. restaurant, park, etc.) were one-hot encoded, grouped by neighborhood, and averaged (mean venue counts per category and neighborhood).

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

Because of the fact that not every neighborhood entry in the crime data set could be assigned to a neighborhood entry in the postal codes data set, rows without mutual entries were dropped. Finally, all other data sets were adjusted and only the reduced data sets were used for further analysis.

3.2 Data Analysis

The general data analysis workflow was composed of the following steps:

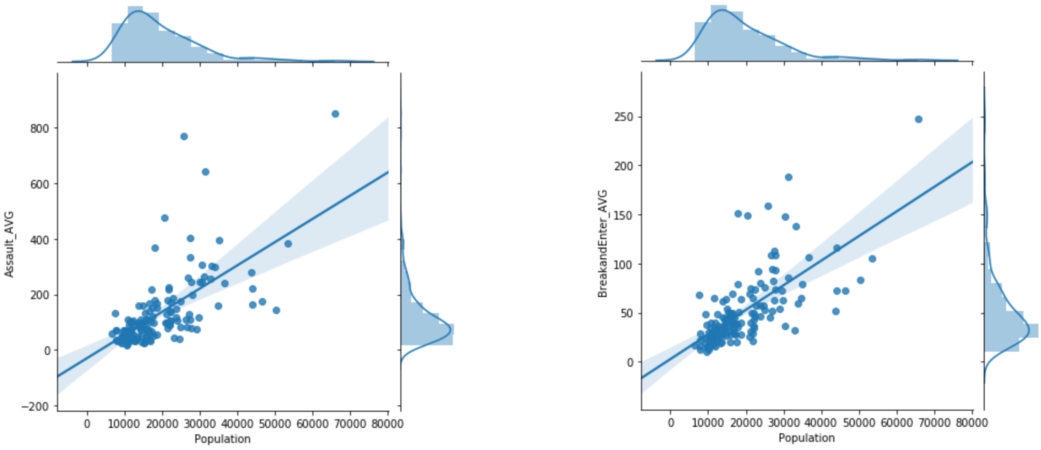
1. The top ten venues per neighborhood were calculated and stored in a data frame.

2. Joint regression plots of crime types (four-year averages) versus population were created. The plots indicated that counts per crime type seemed to be positively correlated with the population of a particular neigborhood. Thus, crime type counts were divided by population of a neighborhood. Two representative plots are depicted in Figure 5.

3. By means of a cluster analysis, the neighborhoods were grouped into five clusters. Two seperate analyses were performed, one using the location (venues) data set and one using the crime data set.

4. The assigned clusters were added to the respective data frames (location and crime data set).

5. A separate map, displaying the population as well as a cluster marker per neighborhood, was created for the location and the crime data set.



(a) assault four-year average **(b)** break and enter four-year average

Figure 5: *Joint regression plots of crime versus population*

3.3 Statistical Inference

In the context of this analysis, no statistical tests (to proof whether the results are significant) were applied. The number of clusters was chosen beforehand and is not guaranteed to be optimal. Furthermore, the data set would need to be randomized and clustered several times, while each time calculating a meaningful test statistic (e.g. overall distance). Finally, it might not be straightforward to determine whether a non-significant result comes from non-significant clusters or from a randomized data set that is not appropriate to perform the intended test.

3.4 Algorithms

In a first approach, the K-Means algorithm was employed to cluster the data sets into five groups. Especially for the high-dimensional (feature space) location data, the algorithm was not able to cluster the data into five meaningful clusters. In a second approach, a Spectral algorithm (graph-based) was used. Graph-based clustering is perhaps most robust for high-dimensional data as it uses the distance on a graph, e.g. the number of shared neighbors, which is more meaningful in high dimensions compared to the Euclidean distance [6]. Cluster analysis and plotting was done using *Python* in combination with the following modules / packages: *sklearn* (*sklearn.cluster.KMeans* and *sklearn.cluster.SpectralClustering*), *seaborn*, and *folium*.

4 Results

Overall, it was not possible to visually detect an obvious pattern between the top ten venues and the cluster of a particular neighborhood. Furthermore, the same holds for the crime data. An excerpt of both resulting data frames is depicted in Figure 6 and Figure 7. The plots in Figure 8 and Figure 9 show the maps with population and cluster markers for location data and crime data respectively. However, the neighborhoods were not clustered in a similar way. Thus, the cluster pattern did not reveal a correlation between neighborhoods with similar venues and neighborhoods with similar counts of the same type of crimes.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	4	Agincourt	_Lounge	_Breakfast Spot	_Latin American Restaurant	_Skating Rink	_Clothing Store	_Drugstore	_Discount Store	_Distribution Center	_Dog Run	_Doner Restaurant
1	3	Alderwood, Long Branch	_Pizza Place	_Coffee Shop	_Gym	_Pharmacy	_Sandwich Place	_Skating Rink	_Dance Studio	_Pub	_Pool	_Diner
2	1	Bathurst Manor, Wilson Heights, Downsview North	_Bank	_Coffee Shop	_Park	_Pizza Place	_Deli / Bodega	_Middle Eastern Restaurant	_Restaurant	_Ice Cream Shop	_Mobile Phone Shop	_Fried Chicken Joint
3	1	Bayview Village	_Café	_Bank	_Japanese Restaurant	_Chinese Restaurant	_Dim Sum Restaurant	_Discount Store	_Distribution Center	_Dog Run	_Doner Restaurant	_Donut Shop
4	1	Bedford Park, Lawrence Manor East	_Sandwich Place	_Italian Restaurant	_Coffee Shop	_Restaurant	_Sushi Restaurant	_Greek Restaurant	_Thai Restaurant	_Comfort Food Restaurant	_Juice Bar	_Butcher
5	4	CN Tower, King and Spadina, Railway Lands, Har...	_Airport Terminal	_Airport Lounge	_Airport Service	_Harbor / Marina	_Bar	_Plane	_Sculpture Garden	_Boutique	_Boat or Ferry	_Airport Gate

Figure 6: *Top ten venues per neighborhood including clusters*

	Cluster Labels	Neighborhood	Population	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG
0	3	Agincourt	52870.0	0.003643	0.001256	0.002521	0.000008	0.001088	0.000340
1	3	Alderwood, Long Branch	12054.0	0.003011	0.001344	0.002049	0.000017	0.000564	0.000564
2	4	Bathurst Manor, Wilson Heights, Downsview North	50925.0	0.008764	0.002641	0.002162	0.000026	0.001453	0.000363
3	3	Bayview Village	21396.0	0.003585	0.000958	0.001879	0.000009	0.000411	0.000388
4	3	Bedford Park, Lawrence Manor East	23236.0	0.001894	0.001971	0.003981	0.000000	0.000559	0.000486
5	4	CN Tower, King and Spadina, Railway Lands, Har...	31180.0	0.008457	0.000792	0.002742	0.000026	0.000657	0.000529
6	0	Church and Wellesley	31340.0	0.020511	0.001206	0.006015	0.000064	0.004330	0.001078

Figure 7: *Crime type (scaled) per neighborhood including clusters*

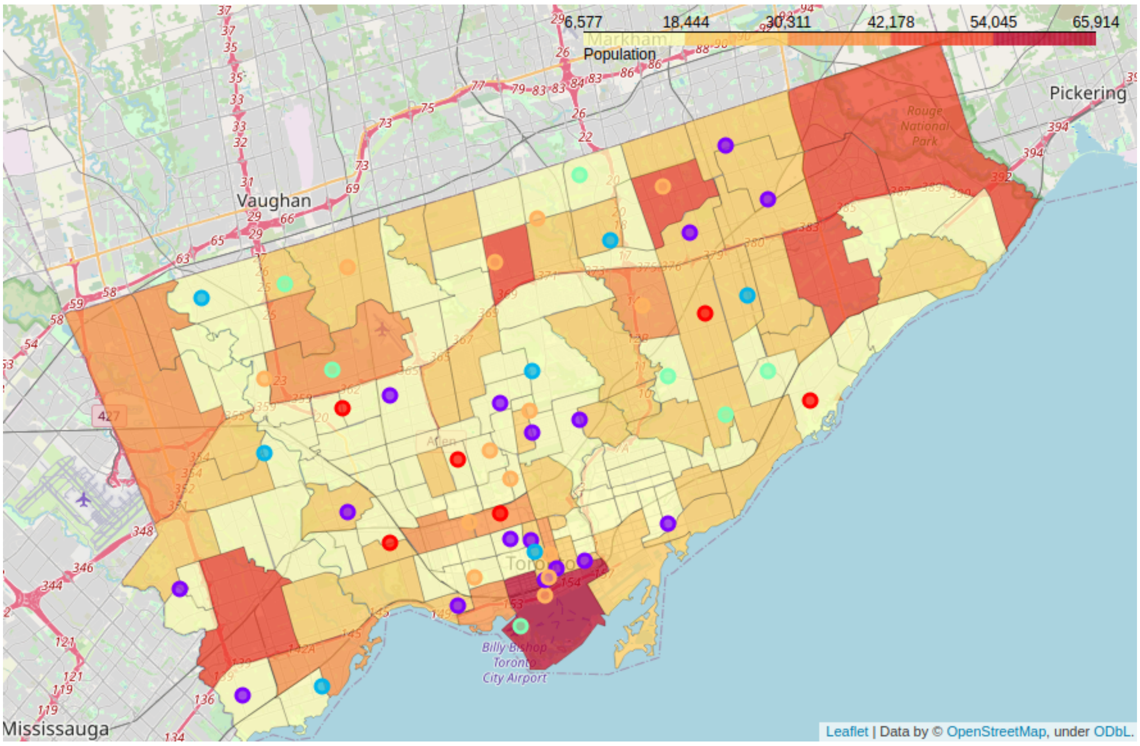


Figure 8: *Map with population and cluster markers for location data*

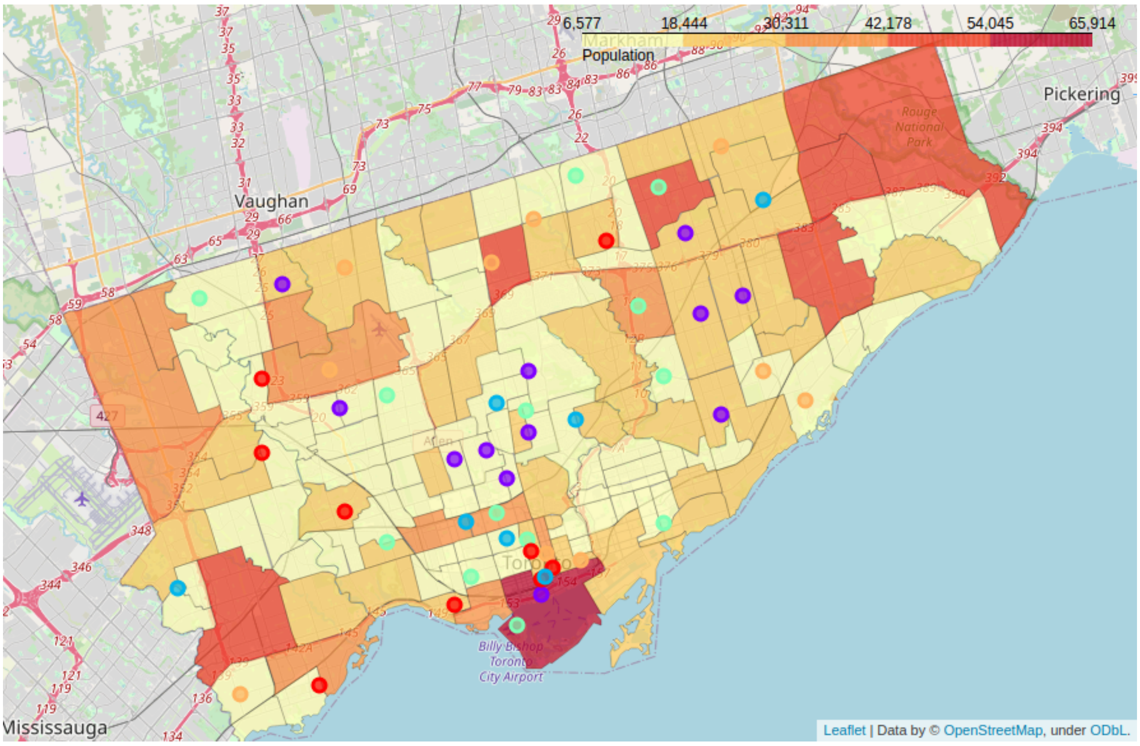


Figure 9: *Map with population and cluster markers for crime data*

5 Discussion

To compare the five clusters, to which the neighborhoods were assigned, the marker colors were chosen the same for both plots (e.g. red means cluster one etc.). This does not imply that neighborhoods assigned to cluster one in the location data set correspond to neighborhoods assigned to cluster one in the crime data set. However, if there exists a correlation, at least a pattern (different colors) should be observable. One reason might be that the location data set contains too many venues (features) and that most of the venues are very similar (e.g. different types of restaurants etc.). It may also be that a correlation between venues and crimes in a particular neighborhood does not exist.

6 Conclusion

The top ten venues per neighborhood were calculated and analyzed. Furthermore, neighborhoods were clustered into five groups using the location data (venues) and the crime data respectively. A separate map, displaying the population as well as a cluster marker per neighborhood, was created for the location and the crime data set. It could not be observed a correlation between neighborhoods with similar venues and neighborhoods with similar counts of the same type of crimes. There might be too many venues (features) in the location data set for meaningful clustering or an entire lack of correlation between venues and crimes in a particular neighborhood. However, a future respectively extended analysis could include other data sets (e.g. traffic flow, commuting rate etc.) and make use of other clustering algorithms or data transformations.

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

[6] https://towardsdatascience.com/how-to-cluster-in-high-dimensions-4ef693bacc6, accessed: 14/06/2020