Clustering Neighborhoods in Toronto Using Crime and Location Data

by Lynx Delta

Introduction:

The aim of this analysis is to explore the venues and the incidents in the neighborhoods of the City of Toronto. Is there a relationship between neighborhoods that share similar venues (e.g. restaurant, park, bank) 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.

Data:

To answer the questions, the following data was fetched, prepared, and analyzed:

- Data from Wikipedia (list of postal codes of Canada; M) [1]
- Geospatial coordinates data [3]
- Location data from Foursquare (venues) [4]
- Crime data from Toronto Police Service [2]

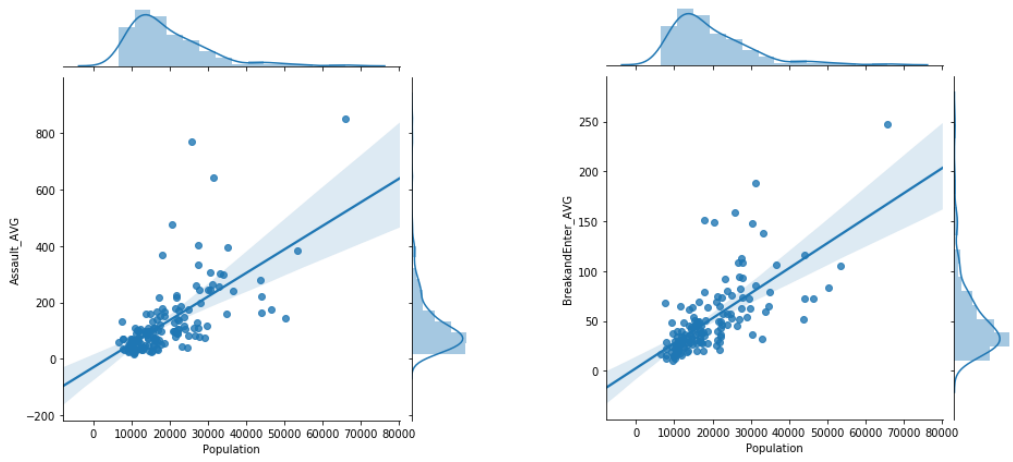
Methodology:

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

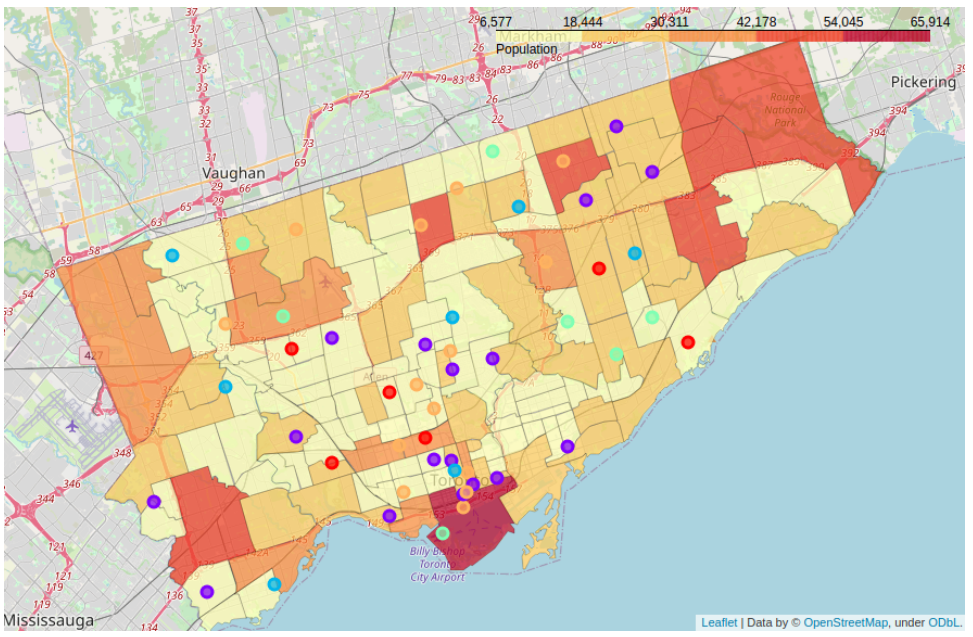
1. The raw data sets were parsed, cleaned, and loaded into data frames
2. The top ten venues per neighborhood were calculated.
3. Joint regression plots of crime types versus population were created.
4. By means of a cluster analysis, the neighborhoods were grouped into five clusters. Two seperate analyses were performed, one using the location (venues) data set and one using the crime data set.
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.

Results:

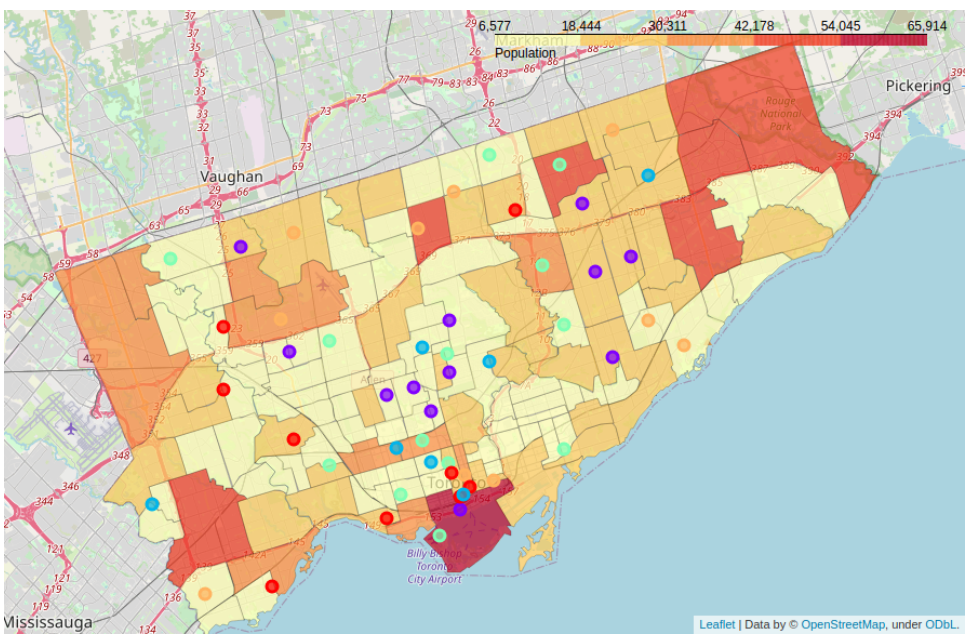
The joint regression plots indicated that counts per crime type seemed to be positively correlated with the population of a particular neigborhood.
However, the neighborhoods were not clustered in a similar way. Thus, the cluster pattern did not reveal a correlation between neighborhoods with similar venues and neighborhoods with similar counts of the same type of crimes.



Joint regression plots of crime versus population



Map with population and cluster markers for location data



Map with population and cluster markers for crime data

Discussion and Conclusion:

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://en.wikipedia.org/wiki/List of postal codes of Canada: M
[2] http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-
[3] http://cocl.us/Geospatial data
[4] https://developer.foursquare.com/