06/2021

Toronto Neighborhoods' Business and their influence on Crime in 2020

Applied Data Science Capstone

IBM Data Science Professional Certificate

Alberto Carrillo Ortega
professional.acarrillo@gmail.com

Index

Introduction

In the next few pages, I intent to gain some insight in the relationship between a given set of business/venues in a neighborhood on the city of Toronto, and the crime rate of that community.

I aim to answer some question such as:

- Is there a correspondence between a predominant type of business and an exceptional high or low criminality?
- Are more diverse neighborhoods less crime prone?
- What characterizes neighborhoods of notable delinquency?

Common sense could provide some answers to some of these questions. For example; one might come to the conclusion that nightlife related neighborhoods rank higher in criminality rates. Or that more "family friendly" ones, those which have a bigger diversity and presence of places like parks and schools and less crime related.

However, is essential to contrast these mental constructs with information extracted from real-world data. It might be the case that nightlife zones are more conflictive, but it might also be the case that parks are hot spots of crime. To answer these questions and give some prescriptive claims we ought to arm ourselves with information.

This information will be extracted from two data sources: Toronto Open Data for crime rates in each neighborhood, and Foursquare API for business and landmarks in each region.

Data Acquisition and Preparation

Data Sources

As I said before, we will be drawing our data from two sources:

TORONTO OPEN DATA

An open data initiative born in 2009 and directed by the government of Toronto City. It offers a wide catalog of data sets ranging from festivals and events to Wellbeing Neighborhoods indexes.

More precisely I will be using a dataset referring to neighborhoods crime rates. This data set contains information about assaults, auto theft, robberies and other crimes. It's recorded annually and there is information from 2014 to 2020.

Here is a link to the dataset: Neighborhoods Crime Rates

FOURSQUARE API

A search-and-discovery service developed by Foursquare Labs Inc. The system provides us with an easy-to-use tool for locating nearby places like restaurants, parks, and all kind of activities.

Once again, to be more precise, we will be using its Places API. This is but a REST interface that allows us to access its services through simple HTTP requests.

I will use this API to obtain a set of venues nearby the coordinates of each neighborhood and relate these locales and their type with the crime rate of the neighborhood. However, is important to acknowledge that this service has some limitations. The most significant one is that there is a limit to the number of venues that be can obtain given a set of coordinates.

Data Cleaning

Fortunately, the data we obtain from Toronto Open Data is a perfectly clean dataset. The data set of Toronto Crimes Rate is this one:

	_id	OBJECTID	Neighbourhood	Hood_ID	F2020_Population_Projection	Assault_2014	Assault_2015	Assault_2016	Assault_2017	...
0	1	1	Yonge-St.Clair	97	14083	16	25	34	25	...

Figure 1 Toronto Crime Rates Data Frame

This is not the case for Foursquare because as it is a REST API, the data we get from it, is in JSON format. We will have to manipulate the HTTP response in order to obtain a useful format.

This is an example of a response we get when we make a call to Foursquare API.

```
{
   "meta": {
      . . .
   },
   "response": {
      . . .
      },
      "groups": [{
            . . .
            },
            "venue": {
               "id": "4be349d763609c7439e11bff",
               "name": "Daeco Sushi",
               "location": {
                  . . .
                  "lat": 43.68783769992881,
                  "lng": -79.39565249242683,
                  . . .
               },
               "categories": [{
                     "id": "4bf58dd8d48988d1d2941735",
                     "name": "Sushi Restaurant",
                     . . .
                  }
               ],
               "photos": {
                  . . .
               }
            },
            "referralId": "e-0-4be349d763609c7439e11bff-0"
         }, {
            "reasons": {
               . . .
            },

                         . . .
```

Figure 2 JSON Response Example

In the hierarchical JSON we can observe some notable elements. We are looking for venues relating to a location. From the JSON it’s easy to extract the venue name, coordinate and category. After calling for each neighborhood we can build a data frame that looks like this:

	Neighborhood	Venue	Venue Category	Venue Latitude	Venue Longitude
0	Yonge-St.Clair	Daeco Sushi	Sushi Restaurant	43.687838	-79.395652

Figure 3 Initial Venue's Data Frame

There are over 350 Venue Categories, and each category only repeats itself in each neighborhood less than 10 times. With this data it would be hard to find any relationship or draw conclusions. Luckily, the categories system follows a hierarchical structure. As so, we can derive a category of a higher order given a specific category. There are still over 100 categories in the second degree of categories, at the highest degree there are 9. After simplifying the venues categorization, the data frame looks like this:

	Neighborhood	Venue	Venue Category	Venue Latitude	Venue Longitude
0	Yonge-St.Clair	Daeco Sushi	Food	43.687838	-79.395652

Figure 4 Simplified Venue's Data Frame

Once we have a set of venues for each neighborhood we can group then by category on each neighborhood and we would get this data frame:

	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Agincourt North	1	0	25	0	1	0	0	14	0

Figure 5 Grouped Venue's Data Frame

Data Preparation and Feature Selection

Once the data frames are ready, is time to merge then and pick the features we wish to use during the rest of the process.

For the Toronto Crime Rates we are only interested in the 2020 data. Therefore, we can get rid of all the data that is not related. Crime on numbers wouldn't be representative of how dangerous a neighborhood is, a larger neighborhood with more population could have more crime due to this bigger citizenry. That doesn't mean that is more conflictive than another with less residents and felonies. The dataset offers us a proportion of crime over the predicted population for that year giving us a tool to compare neighborhoods in an appropriate way.

In addition, we will define a derived variable as the summatory of all weighted crime types in a neighborhood. After that we partition the neighborhoods in three categories: Low, Medium and High Crime. The division will be based on the values of the previous summatory, with Low Crime being the first quartile and below, High Crime being above the third quartile, and Medium Crime the space in between.

	Neighborhood	SumCrimes	SumCrimesBinned
0	Yonge-St.Clair	319.534220	Low

Figure 6 Toronto Crime Rates Curated Data Frame

After curating the Crime Rates data frame, we can merge both data frames on the "Neighborhood" key.

	Neighbourhood	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Nightlife Spot	Travel & Transport	F2020_Population_ Projection	SumCrimes	SumCrimes Cat
0	Yonge-St.Clair	38	10	11	0	3	1	14083	319.534220	Low

Figure 7 Merged Data Frames

Now we are ready to initial approach to analyze the data and gain some insights.