

FIND MY NEW HOME

by Abhishek Panwar



July 30th, 2020

1. Introduction

1.1 Background

In such a large and heterogenous city like Toronto it might be difficult to figure out which neighborhoods to concentrate on in one's search of a new home. Particularly, considering the current high-house prices in certain areas of the city. When searching for a house with family in mind, one might want to concentrate on neighborhoods with low crime rates, average density of population, access to parks and playgrounds, as well as restaurants and coffee shops. All these criteria are usually considered with a certain budget in mind. This report is a case study that takes Forest Hill South neighborhood as a point of reference for a comfortable, safe, enjoyable living (our case subject, me, is currently renting an apartment here), and is trying to find other neighborhoods in Toronto that are similar to Forest Hill South, but are more affordable (i.e., have an average home price of less than \$800,000) for our subject to be able to go from renting to owning a house (average home price in Forest Hill South in 2017 was \$1.32 million in 2017, see below).

1.2 Problem

There are about 100-140 neighborhoods in Toronto (depending on assignment). In order to determine neighborhoods that are like Forest Hill, these neighborhoods need to be segmented based on their safety scores, population density, access to parks, restaurants, and coffee shops using a clustering technique. Once I determine which neighborhood's properties of interest are like those of Forest Hill South (i.e., the neighborhood my subject knows he likes living in), out of those neighborhoods I'll be able to find neighborhoods that satisfy my budget restriction.

1.3 Interest

Whenever someone meets with a real estate agent or visits a real estate website, it is great to have an idea about what areas to limit your search to. Otherwise, the amount of data can be overwhelming, and one might miss their dream home due to the sheer volume of houses for sales around the city. Anyone buying a house would be interested in learning which neighborhoods are similar to the ones they already love, and for many people, crime rates and population density (i.e., traffic and social environment in the neighborhood) would be more important than venues available in the area.

2. Data

2.1 Crime rate, population density, latitude and longitude of each neighborhood: sources, cleaning, feature selection

In order to determine how safe each neighborhood is, I need to obtain some measure of the crime rates. Toronto Police has an open data access to records about various crime ratings, such as assault, robbery, homicide, auto theft and break-and-enter from 2014 to 2018. I have located the geojson file from their website

(https://data.torontopolice.on.ca/datasets/af500b5abb7240399853b35a2362d0c0_0?geometry=-

80.098%2C43.542%2C-78.659%2C43.890) and extracted the crime data for the year 2018 for each neighborhood in Toronto. The file also included the information about population and size of the neighborhood (area in square-meters), as well as the coordinates of the boundaries of each neighborhood.

I extracted the following properties from the geojson file: *'Neighborhood', 'Assault_Rate_2018', 'AutoTheft_Rate_2018', 'BreakandEnter_Rate_2018', 'Robbery_Rate_2018', 'Homicide_Rate_2018', 'Population', 'Size_of_hood_area'*. Then I used the information about each neighborhood's population and size (area in square-meters) to calculate the population density ("Population"/"Size_of_hood_area") I also used the neighborhoods' population to calculate the crime rate per capita, as opposed to using counts as is. As a result and from those features I created my own data set with the following features:

'Neighborhood', 'Population_Density', 'Assault_per_capita', 'AutoTheft_per_capita', 'BreakandEnter_per_capita', 'Robbery_per_capita', 'Homicide_per_capita'.

I kept different types of crimes separate for my analysis and combined them in the end of the study to make conclusions.

There were 140 neighborhoods in the geojson file from Toronto Police, and the format of their names and assignment differed from the neighborhoods we used earlier in Module 3. Therefore, I had to obtain another measure of the coordinates for the centers of the neighborhoods that would match my data. To do so, I used the coordinates of the boundaries to estimate the center point of each area (i.e., I took averages of all latitudes and all longitudes provided for each neighborhood). Those central coordinates were needed in order to carry out queries later in the report and for plotting neighborhood clusters. I added the "Latitude" and "Longitude" features to my data set as well.

2.2 Venues: sources, cleaning, and feature selection

I queried Foursquare API in order find venues that are most popular in each neighborhood of Toronto, using the central neighborhood coordinates are I extracted from geojson file shared by Toronto Police website. The query returned 1475 venues, in 255 unique venue categories. The venue categories were converted to binary variable, using one-hot encoding, and grouped by neighborhood. The resulting data set had 134 rows and 255 columns (plus the neighborhood index), which means that the query returned no data for 6 out of 140 neighborhoods. When this data set was later merged with the data on crime, I replaced the missing values with zeros, since in this case lack of data most likely means that the venue/category was not found (i.e., equal to zero).

2.3 Average home price: sources, cleaning, and feature selection

After segmentation of the data I wanted to overlay it with the information about home prices for each neighborhood. It was not included in the segmentation process because I didn't want to be one of the criteria for similarity between neighborhoods. The home prices data were scraped from a blog post on Toronto home prices by neighborhood for 2017, which luckily has the same format for the neighborhood names and assignment (<https://www.moneysense.ca/spend/real-estate/where-to-buy-2019-toronto/>). There were, however, 6 neighborhoods missing from the dataset, and the missing values in those cases were replaced with average home price for all neighborhoods. In future plots, those neighborhoods can be easily spotted since their home prices have decimal points.

The average home prices were listed as strings with commas and dollar signs. Dollar signs and commas were removed, and the prices were converted to floats. The resulting data frame was merged with the entire dataset, to be used for plotting maps.