Data Science Capstone Project

A Tale of Two Neighborhoods

Part 1: Research Plan

Sara Wedeman July 11, 2019

Business Problem

There are several interrelated questions here. They follow.

Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to determine whether there is a location in Toronto that:

- Offers a better business environment than does Philadelphia to a company like my own?
- Are there any neighbourhoods in Toronto that are similar to my Philadelphia neighbourhood along 'liveability' such as: green space, low crime rate, and number of venues for socializing (coffee shops, restaurants, etc.)?
- Are there any neighbourhoods in Toronto that meet both criteria?
- If so, what are they?

For this particular analysis, we will focus on postal/zip code – level data, for those locations that fall within Toronto and Philadelphia city limits. We recognize this not necessarily the best approach to take, since while the respective metro areas are almost identical in size, the proportions that fall within city limits are different by more than a million people. The reason for this decision is a practical one: to assess one metro area in comparison with another is beyond the scope of the current assignment, in terms of focus and time required.

Target Audience

The primary audiences for this project are:

- The business owner (me);
- Potential investors or other stakeholders; and
- In the event of a decision to move my business to Toronto, Canadian Immigration authorities.

All businesses have startup costs. Establishing a business in a new location, in a different country, will doubtless carry costs. While bringing such a move to fruition requires a cost estimate, with a list of decision points and criteria, this is beyond the scope of the current undertaking.

Methodology (synopsis). *Please see Part 2 for a more detailed description*)

1. Conduct a review of the relevant literature, using resources available online. Topics include:
    - Toronto history and current state (geographic, demographic, economic, etc.)
    - Business trends
2. Review Data specifications and availability. Steps include:
    - List data Requirements
    - Collect data – locate Web sites offering Zip and or Postal Code information that can be readily scraped.
    - Load Foursquare data for all Zip Codes in Philadelphia and all Postal Codes in Toronto.
    - Data understanding: Are there any particularities in the data set that must be taken into account during setup (e.g., Zip or Postal codes that fall outside the city limits or are assigned only to P.O. Boxes, and therefore should not be counted)?
    - Clean data – for example, remove from analysis any Zip codes or Postal codes that are actually P.O. boxes.
    - In the interest of future research, identify and describe limitations of this research, implications, and suggestions for future researchers.

Data Used

Sources Include but are not limited to[1]:

- U.S. Census
- Canadian Census
- Foursquare Data
- Philadelphia vs. Toronto
- Technical.ly Philly
- History of Toronto
- History of Philadelphia
- The Encyclopedia of Philadelphia
- Toronto Neighbourhoods and Communities
- The Paris Review: America's First Female Map Maker
- Don Valley Historical Mapping Project

For Future Research

One of my major concerns with this analysis is that Postal and/or Zip codes are not appropriate units of measurement for population, demographic, or economic data[2]. They were never intended for this purpose: *they were designed solely to facilitate the delivery of mail*. There are some office buildings in New York City that have their own zip codes simply because they are so big. In these and many other cases, zip code-level statistics are meaningless for a whole variety of assessments. For example, a Zip code-based analysis would make the Seagram's building one of the wealthiest "neighbourhoods" in New York (except that it is not a neighbourhood).

A vastly superior unit of measurement would be one of the US (or Canadian) census categories, like – in the U.S. case – the block group. This is the smallest unit of measurement in the Census (between 600 and 3,000 people). It is more stable than the Zip code (which changes when the volume of mail does). Also, for any population group, the median is a better measure of central tendency than the arithmetic mean, since simple averages (means) can be badly skewed by outliers. In densely populated urban areas like downtown Toronto or Philadelphia, one also finds wealthy and impoverished people living in close proximity, so the danger of such skew is a real one.

Had I the requisite technical skills, I would add a lot more data to the mix: NAICS (North American Industrial Classification System) codes, traffic patterns, etc. I would perform a number of other types of analyses I have not yet mastered (e.g. chi square analyses on standardized data comparing Philadelphia and Toronto on any number of characteristics).

In short, I predict I will be going much farther down this exciting Data Science learning path.

---

[1] This is the *short* list.

[2] See 10 Reasons to use Census Tract Versus ZIP Code Geography & Demographics, among many others, for more on this topic.

Methodology Detail

Our first step will be to get lists of neighbourhoods for the cities of Philadelphia PA and Toronto, ON. Fortunately, such lists are not hard to find on the Internet (see: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and http://ciclt.net/sn/clt/capitolimpact/gw_ziplist.aspx?ClientCode=capitolimpact&State=pa&StName=Pennsylvania&StFIPS=42&FIPS=42101). We will use python's *beautifulsoup* library to extract the needed postal code lists. Then, we will get the geographical coordinates (latitude and longitude) so we can use them to query the Foursquare API database.[1] A geocoder will allow us to do so. We will then be able to load this information into a pandas dataframe, then using *folium*, we will visualize each city's neighbourhoods on the map.[2]

Using the Foursquare API, we will subsequently get the top 100 venues that are within a radius of 500 meters from the center point of each Zip or Postal Code. We do this by making API calls to Foursquare, passing the geographical coordinates until we are done via a Python loop. Foursquare then returns venue data to us in a JSON format, and we extract the venue name, category, latitude, and longitude. With these data, we will be able to check to see how many venues were returned for each neighbourhood and to tally up the number of (somewhat)[3] unique categories can be curated from all the returned venues. The next step will be to conduct k-means clustering – using the mean frequency of occurrence of each venue category to create a centroid for each postal code. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is a simple and popular unsupervised machine learning algorithms. I hope that as I progress in my development as a data science, I will be able to add to my repertoire of analytic abilities, so I can do more with data. In the interim, this represents a colossal leap forward on my part.

The results will allow me to identify which neighbourhoods are most likely to meet my twin aims of building a new practice that will be offer new services based in Data Science, while permitting me to live someplace like, where I can walk to work.

---

[1] At this point, we will have set up Foursquare API accounts and gotten Foursquare credentials.

[2] We will also conduct a 'sanity check' to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the cities of Philadelphia PA and Toronto, ON.

[3] These data are crowd sourced, and the categories are – it seems – far from orthogonal. For example, one category is "food," which could mean any establishment that sells food. How one distinguishes "food" from "grocery store" is a mystery. See: Using Foursquare place data for estimating building block use,