The Battle of Neighborhoods: Starting a Coffee Shop Business

Build models for segmenting the neighborhoods to find the most conducive locations for starting a business in Toronto City

Diardano Raihan
23 November 2020

A. Introduction

1. Background

Toronto is Canada's largest city with a population of more than 2,7 million and a density of 4,334.4 people per square kilometer. The city is renowned as one of the most multicultural cities globally due to its large population of immigrants from all over the globe. This leads the city to become a world leader among other metropolitan and cosmopolitan cities from many sectors, including business.

2. Business Problem

Kopiasli
AUTHENTIC COFFEE TASTE

Now, imagine that you own a coffee shop called **Kopiasli** that has been doing business successfully in New York. This year, you and your team plan to expand the business and decide to look for a city that shares the same trait as New York, and one of the cities is **Toronto**.

To ensure this project's success, the team requires insights into the demographics, neighboring businesses, and crime rates. For each neighborhood, we can ask:

- How many cafes exist?
- What are the most popular venues?
- Can we get information about the vehicle and foot traffic?
- What is the neighborhoods' crime rate? And so on.

Thus, the **project goal** is to figure out the best locations for opening up a new coffee shop in Toronto City.

3. Target Audience

Entrepreneurs who are passionate about opening a coffee shop in a metropolitan city would be very interested in this project. The project is also for **business owners** and **stakeholders** who want to expand their businesses and wonder how data science could be applied to the questions at hand.

B. Data Description

1. Data Requirements and Collection

We need historical data about crime incidents, busiest roads, and popular venues. Luckily, Toronto has an open data portal that makes it public. We can also leverage Foursquare Location data to compare neighborhoods in terms of service. Hence, the followings are data sources that we can use for this project:

- **1st Data**: https://tinyurl.com/vehicle-foot-traffic
  The most updated record of traffic **signal vehicle and pedestrian volumes** in Toronto City.
- **2nd Data**: https://tinyurl.com/toronto-mci
  The most updated **record of crime incidents** reported in Toronto City provided by Toronto Police Services.
- **3rd Data**: https://tinyurl.com/toronto-postal-code
  The list of Toronto neighborhoods represented by postal codes and their boroughs.
- **4th Data**: https://developer.foursquare.com/
  The popular or most common venues of a given neighborhood in Toronto.

2. Data Cleaning and Feature Extraction

- The first data is in a CSV file. It contains 2280 rows and 11 columns. The data is typically collected between 7:30 a.m. and 6:00 p.m at intersections where there are traffic signals. Each intersection holds vehicle and pedestrian volumes data, along with its coordinates. We will focus on 5 columns; those are **Main**, **8 Peak Hr Pedestrian Volume**, **8 Peak Hr Vehicle Volume**, **Lattitude**, and **Longitude**. We will use these features to diagnose each main road's characteristics and locate the busiest main roads in the city.
- The second data is also in a CSV file. It contains 206,435 rows and 9 columns. The rows represent crime incidents that reported from 2014 to 2019. It has 5 Major Crime Indicators (MCIs) scattered to 17 divisions and 140 listed neighborhoods. We will group the data based on division and get statistics about crime rates.
- The third data is a Wikipedia page about Toronto postal code. We will scrape the page and create a data frame consisting of three columns; **PostalCode**, **Borough**, and **Neighborhood**. We remove any rows that do not have borough assigned. Then, we will be using the **Geocoder** python package to retrieve the **postal code's coordinates**. It will return 103 rows and 5 columns.

- The fourth data is stored inside **Foursquare Location Data**, and we will use **Foursquare API** to access it. We utilize the postal coordinates to retrieve popular venues around a specific radius. As a result, the same venue categories will be returned to different neighborhoods. We can use this idea to cluster the neighborhoods based on their venues representing services and amenities.
- We will run the **k-Means** algorithm to perform this clustering with a different number of clusters (k). The **features will be the mean of the frequency of occurrence of each venue category.** Finally, we can visualize the cluster model using the **Folium** module.

To sum up, we will use the 1st and 2nd data to analyze the pedestrian/vehicle volume and crime rates. Then, we load the 3rd data to obtain the exact coordinates for each neighborhood based on the postal code, allowing us to explore and map the city. Finally, we will use the coordinates and Foursquare credentials to access the 4th data source through its API and retrieve the popular venues along with their details, especially for coffee shops. The venue frequency in each neighborhood will be the features of the clustering model.