Coursera IBM Data Science Certification

The Battle of Neighborhoods

New Barbershop in Toronto

Final Report

Prepared by Elias Abou Charanek

Oct 01, 2019

Contents

Problem Statement:

A new comer to Toronto wants to open a new barbershop, preferably in a populated postal code, with no or little competition. For this reason, the new investor is only interested in checking a score that reflects the ratio of population to competitor per postal code, to identify the top 5 postal codes to consider.

Executive Summary:

By looking at the population size and the number of competitors per postal code, the following 5 postal areas ranked the highest. Note that all of them are the periphery of Toronto.

Rank	Postal Code	Population	Competitors	Normalized Score
1	M1B	66,108	0	100.0
2	M2J	58,293	0	88.2
3	M9V	55,959	0	84.7
4	M1V	54,680	0	82.7
5	M1W	48,471	0	73.3
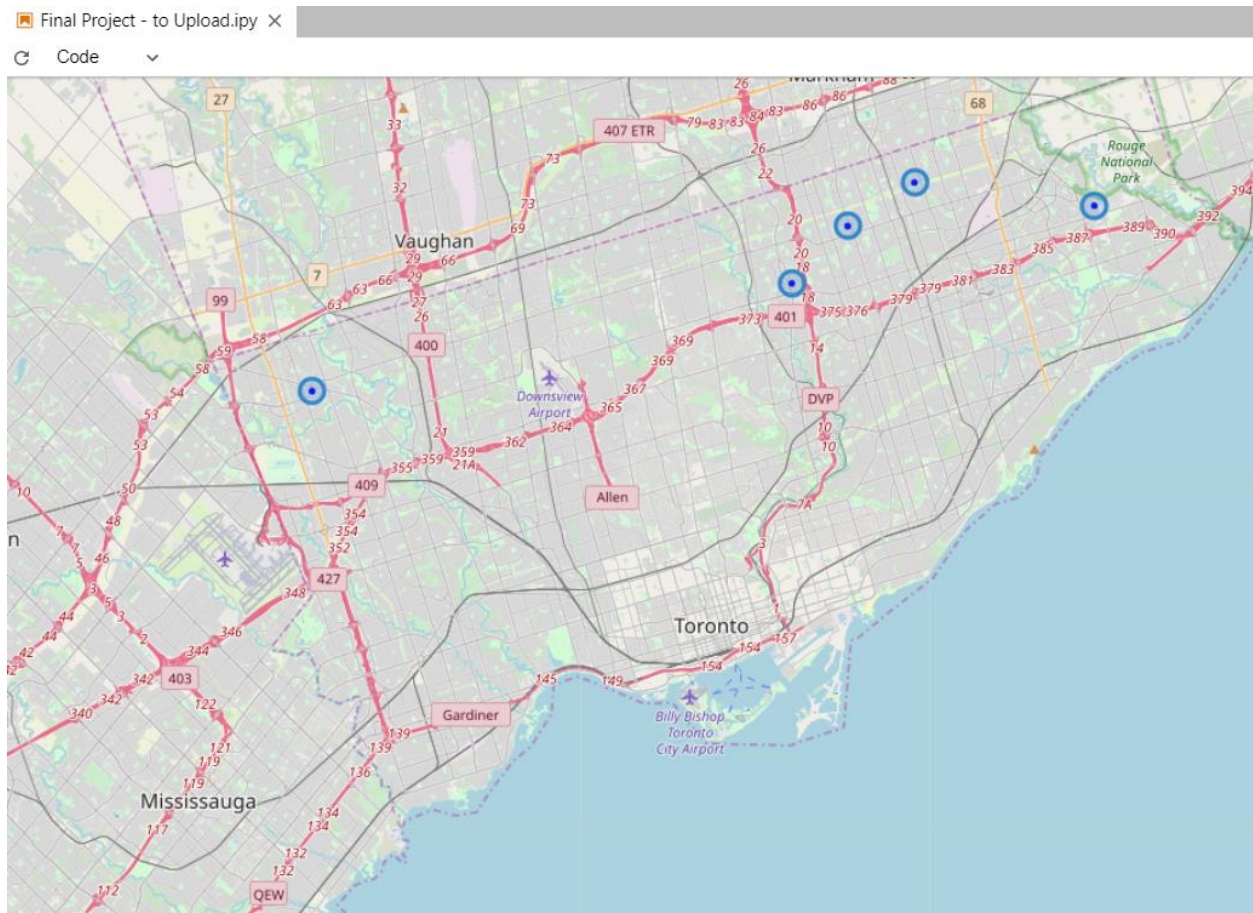
Limitation of this work:

The limitations of this work stem mainly from:

1. The assumption that population and number of competitors per postal code, suffice to make a business decision or start a business plan
2. Accuracy of the data provided by the sources:
    a. Population 2016 figures
    b. Coordinates accuracy depends on geopy
    c. Competitors data depends on the accuracy and limitations of Foursquare API)

Technical Work:

Load the required libraries

The required libraries were loaded and imported.

Data Acquisition

a. Population per postal code
Data per postal code was acquired from Statistics Canada, and I used BeautifulSoup to acquire data from the webpage and load it into a dataframe.

Data had to be manipulated and cleaned from the comma character, to transform the population field from string to float.

b. Coordinates of Postal Codes
Coordinates of postal codes in Toronto area were acquired from https://cocl.us/Geospatial_data and loaded into a panda dataframe.

c. Foursquare businesses
Venues with categories related to salon/barbershops were downloaded for each of the postal codes in Toronto (category = 4bf58dd8d48988d110951735 as shown on this page https://developer.foursquare.com/docs/resources/categories), and results were aggregated in a panda data frame. Duplicates were removed, and I ended up with a dataframe of 64 barbershops. Many had missing postal codes, or were situated in postal codes that start with L. In order to minimize queries to Foursquare, results were saved in a CSV file, which was loaded again when needed in subsequent steps.

Data Cleaning and Preparation

Data from Foursquare needed some manipulation to make it usable. I used GeoPy Nominatin to collect the addresses for businesses that didn't show a postal code. The returned address was parsed, and the postal code was determined. All businesses that had a postal code that does not start with M were dismissed.

Data manipulation

Foursquare data was grouped by postal code to show the number (count) of barbershops per postal code.

One panda dataframe that joins the three tables (population, coordinates, and #barbershops) was created. A new metric, population / (#barbershops +1) was used, and it was normalized to determine the

attractiveness of the area. The dataframe was sorted in a descending way according to the score, to list the most attractive postal codes on top.

Results Display

Data was displayed in a tabular format, and the top 5 locations were also displayed on a folium map as shown in the data executive summary. The top 5 locations are very populated, and lack barbershops according to the used data. All 5 are in the suburbs of Toronto.

Data Sources:

1. Population Data: Link
2. Postal Codes Coordinates: https://cocl.us/Geospatial_data
3. Foursquare API Category 4bf58dd8d48988d110951735 (Salon/Barbershop)