Applied Data Science Capstone – Week 4 – Capstone Project – The Battle of Neighborhoods (Week 1)

Project Title:

The Battle of Neighborhoods (Week 1)

Descriptions (Part 1):

Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data to solve or execute. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.

This submission will eventually become your **Introduction/Business Problem** section in your final report. So I recommend that you push the report (having your Introduction/Business Problem section only for now) to your Github repository and submit a link to it.

Report (Part 1):

Section 1: Introduction and Business Problem

The purpose of this data project is to explore the popular venues or facilities in different neighborhoods in Toronto. Toronto was the most populous metropolitan area in Canada in 2019, with a population or around 6.47 million people (Statista, 2020). Toronto attracts many tourists and new immigrants every year because it is highly economically developed and is one of the most vibrant cities in North America.

Stakeholders. This data project targets these tourists and immigrants who consider moving to Toronto for either recreational visits (travelling and sightseeing) or business purposes (opening and expanding to a new business). Primarily, tourists would like to find popular venues, scenic spots, or superstar cafes and restaurants to visit, while immigrants may find neighborhood(s) with less intense competition in their own interested fields of businesses. For example, an immigrant who runs grocery stores may find a neighborhood with fewer grocery stores of similar kinds more attractive to start their very first business in Canada. All these procedures are costly in terms of time and resources.

Business Problems and Data Project Objectives. This data project is meant to help these groups of stakeholders (tourists and new immigrants) to learn more about different neighborhoods in Toronto. By comparing different neighborhoods in Toronto, these stakeholders would obtain information on which types of venues (e.g. cafes, restaurants, rental car locations, banks, gas stations, gyms, etc.) are the most (or least) popular. With this analysis, stakeholders could decide which neighborhoods have the features of venues that they would like to see. We assume that tourists are more interested in neighborhoods with more popular restaurants and/or tourist spots, while immigrants are more interested in neighborhoods with less intense competition of venues in their preferred fields of businesses.

Toronto is also well known for its higher housing prices. It may affect both tourists and immigrants. Tourists may find hotel or Airbnb accommodation more expensive in certain neighborhoods, while immigrants may find their homes less affordable in certain neighborhoods. This would affect their decisions whether to visit or stay in a neighborhood or not. We will use housing price data in Toronto to give suggestions to our stakeholders, especially those who are sensitive to accommodation/rental expenditures.

Recently, Toronto has been hit hard by the global Covid-19 pandemic. We see soaring Covid-19 active cases and death tolls, unfortunately. We expect that tourists and immigrants may also want to factor in the overall health and safety conditions into their consideration whether to move into a neighborhood or not. We will also use the Covid-19 cases data to provide insights from public health perspectives to potential tourists and immigrants who may want to avoid visiting the hardest hit neighborhoods in Toronto.

Descriptions (Part 2):

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

This submission will eventually become your **Data** section in your final report. So I recommend that you push the report (having your **Data** section) to your Github repository and submit a link to it.

Report (Part 2):

Section 2: Data Sources and Descriptions

Data Sources. In this data project, we will use 5 data sources:

- List of Postal Codes of Canada, obtained from Wikipedia. This data is the same as what we used in previous weeks of this course. The URL is: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. It contains the information of Postal Codes, Boroughs and Neighborhoods in Toronto. There are 103 neighborhoods.
- House Price Data for all neighborhoods in Toronto sorted by Postal Codes, obtained from House Price Hub. The URL is: https://housepricehub.com/cities/city/Toronto. It contains the information of Postal Codes, Average House Prices for all Postal Codes in Toronto.
- Covid-19 Cases for all neighborhoods in Toronto sorted by Postal Codes, obtained from Open Data Portal Toronto. The URL is: https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.
- Geospatial Data (Geographical Coordinates of Each Postal Code in Toronto). This data file is taken from the previous weeks of this course. The URL is: http://cocl.us/Geospatial_data. It contains the information of Postal Codes, Latitude and Longitudes for each of these Postal Codes.
- Foursquare Social Location Service Data, obtained from Foursquare Developer account and using API requests. The requests can be made by specifying Client ID, Client Secret, Version, Latitude and Longitude (of Neighborhoods that you want to search for), Radius, and Limit (number of venues returned by Foursquare API).

Data Descriptions. The first 4 data sources have the following variables:

- Source 1: Toronto, columns = {PostalCode, Borough, Neighborhood}
- Source 2: houseprice, columns = {PostalCode, AvgPrice}
- Source 3: covid, columns = {PostalCode, CovidCases}
- Source 4: geocode, columns = {PostalCode, Latitude, Longitude}

We will then clean each of these dataframes and join/merge all these sources into one single dataframe, which we call ‘df’ containing: columns = {PostalCode, Borough, Neighborhood, Latitude, Longitude, AvgPrice, Count}. The column Count refers to the Covid-19 cases in each neighborhood. Note that we focus on confirmed cases only, probable cases are excluded.

Borough Selection. To simplify our analysis, we focus on the neighborhoods in Scarborough, Toronto. The geographical coordinate of Scarborough, Toronto are 43.773077, -79.257774. There are two reasons to focus on Scarborough. First, Scarborough is a popular destination for new immigrants in Canada, making it one of the most diverse and multicultural areas in the Greater Toronto Area. It particularly suits our context of providing location-based information to new immigrants defined in our Introduction and Business Problem section. Second, there are no missing values of Average House Price and Covid Cases data for neighborhoods in Scarborough, so that we can preserve the most comprehensive information of neighborhoods in Scarborough. A previous version of this project focuses on Downtown Toronto, but then Downtown Toronto seems to have more missing values of Average House Price and Covid Cases data for its neighborhoods.

More Data Descriptions on Foursquare API. By making the Foursquare API requests, we can obtain the detailed information of popular venues within a specified radius of a specified neighborhood (with latitude and longitude values). Take our API requests for Malvern, Scarborough as an example. It has 43.806686…, -79.194354… as neighborhood latitude and longitude values. For the top venues, it gives 'SEPHORA' (the store name) as the name of venue, '300 Borough Drive' as the location, 43.775016…, -79.258109… as the latitude ang longitude values, '217' as the distance, 'M1P 4P5' as the 6-digit postal code, 'Cosmetics Shops' as the category of venue, and etc.

There are, of course, much more information from Foursquare, including menus (for places like restaurants, cafes, etc.), photos, and comments, for all these venues. We restrict the radius to be 1000 meters.

Python Libraries and Packages. This data project requires the following dependencies: NumPy (to handle data in a vectorized manner), Pandas (for data analysis), JSON (to handle JSON files), XML (to process XML), Geocoder (to convert an address into latitude and longitude), Requests (library to handle requests), Matplotlib (plotting tools), Scikit-Learn (use k-means clustering), Beautiful Soup (for parsing HTML and XML documents), and Folium (map rendering library).