IBM Applied Data Science Capstone Project

Coursera Capstone Project

“The Battle of the Neighborhood: For all the Coffee lovers”

By: Tshering Sherpa

July 4, 2020

Introduction

Coffee is a brewed drink prepared from roasted coffee beans. It contains caffeine that acts as a central nervous system stimulant and when reached to the brain gives high sense of alertness. As an American I drink coffee first thing in the morning. I consider myself as a coffee lover. If I am not brewing in my home, then I usually drink coffee from a café. Everyone has their own preference for the coffee. Not everyone likes same flavor or brand nor they like same café or coffee shop. That is why I like having lots of coffee shops around me so that I can try out and have a favorite one. In the future I would like to open my own coffee shop and therefore I am researching about other shops and its locations leveraging the Foursquare location data to explore the market. For any decision, we need lots of information and facts to support the decision. This project will help us gather information, analyze and conclusion to support the facts to help making decision.

Business problem

To find out coffee shops around the area we need to gather information about the area and competitors. This capstone project purpose is to collect information, analyze and select the best locations where we are interested in to open a coffee shop. For this project, I chose to research in Toronto and New York City and compare these two big cities. I will be using data science methodology, Foursquare API location data and machine learning technique which I learned from coursera to answer the questions like if a particular region is safe to open a coffee shops or where I would recommend to open it. We know that "New York City" is considered as city that never sleeps so I think New York city should have more coffee shops and more competition comparing to Toronto. New York city is also densely populated so there will be lots of opportunities even though it has high number of coffee shops.

Target audience

This project is useful for anyone who loves coffee. It can be used for tourist or residents who wants to try out and know where the good things are. It can be used for property developers i.e. investors, entrepreneurs or business consultancies who wants to open a new coffee shops in the city or their desired location. One of the target audience is myself since I would like to open my own coffee shop in the future. This project will help me to make decision or to support my decision.

Data

Data are the crucial part of any project. We need data that has four dimensions: volume, variety, velocity, and veracity. These four dimensions are the foundation of data science. For this project, we will need a data with the list of neighborhoods of the city, then we will need data with latitude and longitude coordinates of those neighborhood and finally venue data related to the coffee shops in the city which we will get from Foursquare API.

Sources of data and methods to extract

This Wikipedia page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M") has the list of neighborhoods in Canada. We will use web scraping techniques to extract the data from this page with the help of built in function of python pandas library. Then we will use this website (http://cocl.us/Geospatial_data) which was given by coursera to get the geographical coordinates of the neighborhoods of the Toronto. We will use geopy library to convert an address into latitude and longitude values. After that, we will use Foursquare API to get the venue data for those neighborhoods of the Toronto city. Foursquare has one of the largest databases of places of 105 million places and has multiple well-known partners such as Samsung, TouchTunes, Airbnb, uber. The Foursquare API allows application developers to interact with the Foursquare platform.

The API itself is a RESTful set of addresses to which you can send requests, so there is really nothing to download onto your server. In order to extract data from Foursquare API, we will need to open an account in Foursquare API then use credentials given by Foursquare API to request the venue data. In this way we can extract venue data of interested location. Foursquare API will provide many categories of the venue data, but we will focus more on coffee shops or shops related to coffee for this project. For the New York city we will download the data with borough, neighborhood, latitudes and longitudes for the website given by coursera (https://cocl.us/new_york_dataset). We will use the web scraping technique to extract the data. For New York city, we will use wget library to download the file first and convert it to data frame. After that, we will use Foursquare API to get the venue data of the city focusing on coffee shop.

Methodology

The first thing we need to do is to get the list of neighborhoods in the city of Toronto and New York city. For Toronto I used Wikipedia page. The list is available in the Wikipedia page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"). We will do web scraping using Python with built in function of pandas library to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to leverage the use Foursquare API. To do so, we will use the Geocoder library that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas data frame and then visualize the neighborhoods in a map using Folium library. This will help us to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Toronto.

Next, we will use Foursquare API to get the top venues limiting 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Coffee shop” data, we will filter the “Coffee shop” as venue category for the neighborhoods. Also, I will use a list that contains filter words like 'Coffee shop', 'Bagel Shop', 'Donut', 'Breakfast Spot', 'Cafeteria', 'Cafe' which are related to coffee shop that will help to get venue that are related to coffee shop.

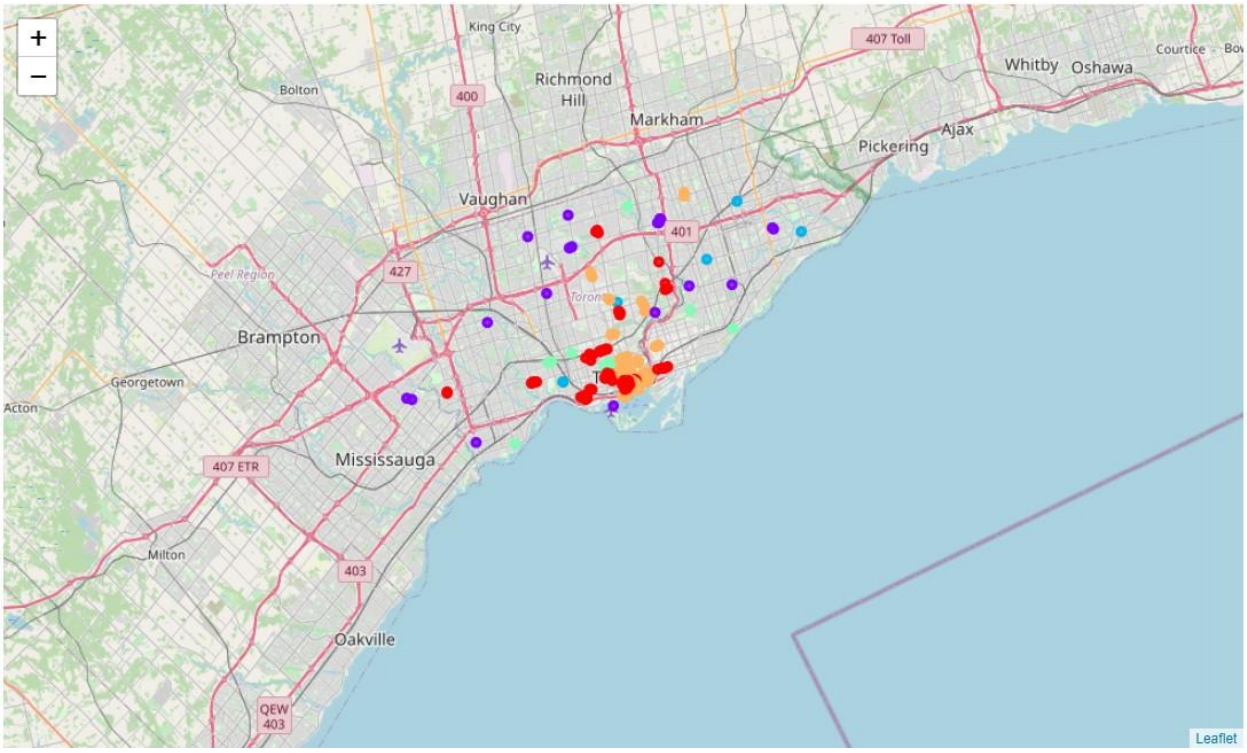
Finally, we will perform clustering on the data by using k-means clustering. K-means clustering is a machine learning algorithm that identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this kind of project. We will cluster the neighborhoods into 5 clusters based on their frequency of occurrence for “Coffee shop”. When I chose 3 k clusters it gave me equal amount of data points and when I chose 6 k clusters it gave me no data point for cluster 6. Therefore, I chose to work with 5 clusters since it gave me the best results. The results will allow us to identify which neighborhoods have higher concentration of coffee shops and which neighborhoods have lower number of coffee shops. Based on the occurrence of coffee shops in different neighborhoods, it

will help us to answer the question as to which neighborhoods are most suitable to open new coffee shops.

Results

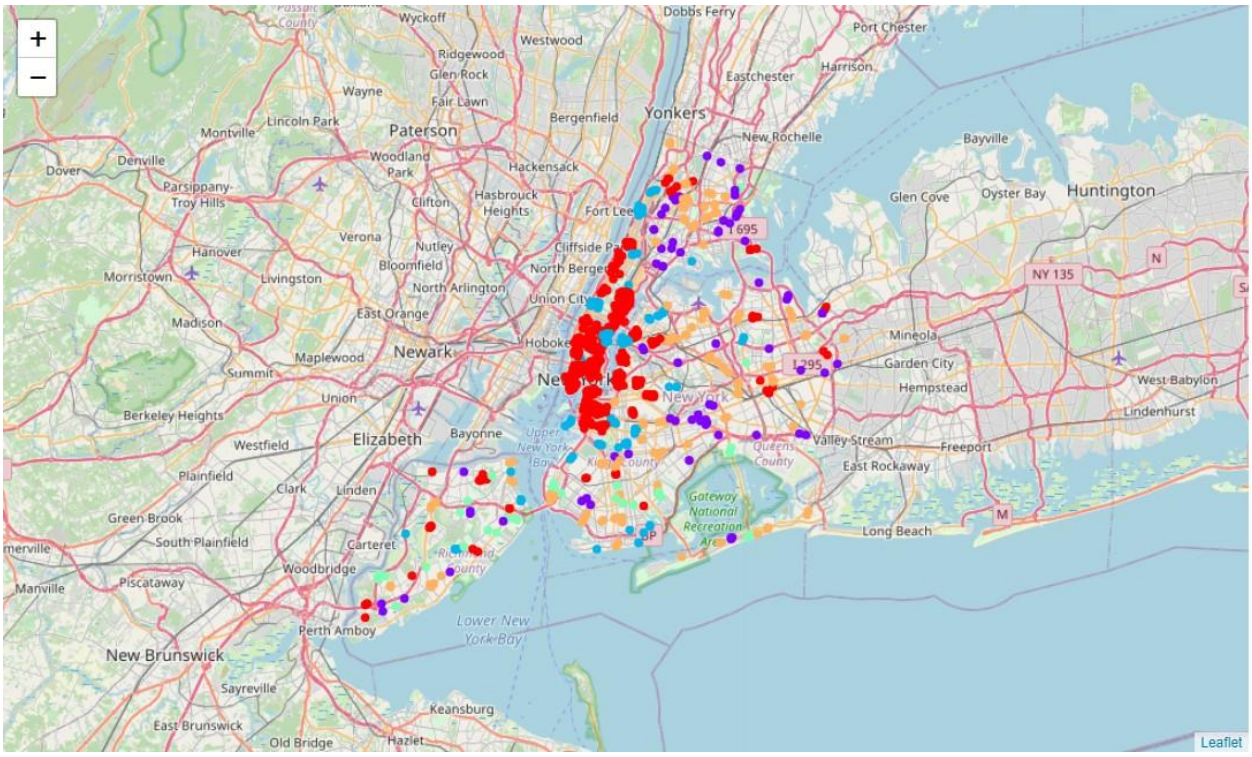
For Toronto: the results from the k-means clustering show that we can categorize the neighborhoods into 5 clusters based on the frequency of occurrence for top category of the venue:

- Red Cluster 0 has high concentration of coffee shops with more diverse venue.
- Purple Cluster 1 has moderate concentration of coffee shops.
- Blue Cluster 2 has low concentration of breakfast Spots with breakfast spots as a top venue.
- Green Cluster 3 has low concentration of cafe.
- Orange Cluster 4 has high concentration of coffee shops.



For New York city: like Toronto the results from the k-means clustering show that we can categorize the neighborhoods into 5 clusters based on the frequency of occurrence for top category of the venue":

- Red Cluster 0 has high concentration of coffee shops.
- Purple Cluster 1 has high concentration of donut shops with top venue as donut shop.
- Blue Cluster 2 has high concentration of cafe.
- Green Cluster 3 has moderate concentration of bagel shops with top venue as bagel shop.
- Orange Cluster 4 has high concentration of donut shops with top venue as donut shop.



Fun fact about New York city result: donut shops listed as top venue for more than one cluster.

Discussion

As observations noted from the map in the results section, for Toronto most of the coffee shops are concentrated in the downtown area of Toronto city, with the highest number in clusters 0 and 4 and moderate number shops are concentrated throughout city in cluster 1. Clusters 2 and 3 has very low number of coffee shops in the neighborhoods. This represents a great opportunity and high potential areas to open new coffee shops as there is very little to no competition from existing shops. Meanwhile, coffee shops in clusters 0, 1 and 4 are likely suffering from intense competition due to oversupply and high concentration of coffee shops. From another perspective, the results also show that the oversupply of coffee shops mostly happened in the middle area of the city, with the suburb area still have very few coffee shops Therefore, this project recommends investor or entrepreneurs to capitalize on these findings to open new coffee shops in neighborhoods in clusters 2 and 3 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new coffee shops in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of coffee shops and suffering from intense competition.

For New York city, the only cluster that has low number of coffee shop is in cluster 3 which has bagel shop as the top venue. Therefore, investors are advised to open coffee shop in the neighborhoods of cluster 3. Any other cluster will be high risk investment and are advised to avoid these neighborhoods. Clusters 2 and 4 has high concentration of coffee shop with donut shop as a top venue, therefore property developers are advised to open coffee shop specifically focused on coffee rather than donut or bagel shop.

New York City is densely populated so even though there are high concentration of coffee shops comparing to Toronto, New York city might have better opportunity to open and have better outcome of opening a coffee shops than in Toronto.

Limitations and Suggestion for future research

In this project, we only consider one factor i.e. frequency of occurrence of coffee shops, there are other factors such as population and income of residents that could influence the location decision of a new coffee shops. The crime rate in the city and ratings of the coffee shop can also influence the decision. Since I was using free account for Foursquare API, I had limited number of requests and response to use each day from Foursquare API which was not good since I must submit my project before the deadline. Also, I did not know about this until I googled about the intermittent error that was giving me by the code. For the future research we can add multiple data to include crime rate, population, ratings, and area of the city to be used in the clustering algorithm to determine the preferred locations to open a new coffee shop. We could also make use of paid account to bypass the limitations and obtain more results from Foursquare API. This project could use longer deadline or get a head start from the beginning of the course or capstone can be very helpful to finish the project which can be use to go for different style of model evaluation from the data science methodology. Finally, use of multiple machine learning algorithm to get different and better results can be very helpful as well.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing k means clustering machine learning algorithm by clustering the data into 5 clusters based on their similarities, and finally providing recommendations to the relevant stakeholders i.e. property developers, entrepreneurs and investors regarding the best locations to open a new coffee shops. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The

neighborhoods in clusters 2 and 3 for Toronto and cluster 3 for New York city are the most preferred locations to open a new coffee shop since it has lower concentration of coffee shops. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding competitive areas in their decisions to open a new coffee shop.

References

Data set for New York city via coursera

https://cocl.us/new_york_dataset

Data of coordinates of Toronto neighbor via coursera

http://cocl.us/Geospatial_data

Foursquare Developers Documentation. Foursquare. Retrieved from

https://developer.foursquare.com/

The Wikipedia

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Weaver's coffee and tea

https://weaverscoffee.com/blogs/blog/the-worlds-top-coffee-consuming-nations-and-how-they-take-their-cup