Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Coffee Shop in the city of Toronto

By: Jaekisen Agarwal

January 2020

Introduction

For many customers, visiting coffee shop is a great way to relax and enjoy themselves during weekends and holidays. It provides a friendly, comfortable atmosphere where the customer can receive quality food, service and entertainment at a reasonable price. Opening a successful coffee shop can be a rewarding experience. Because of this, hundreds of friends will have great conversations. Because of this, mornings will be brighter and afternoons will seem less stressful. Opening a cafe takes a big investment in both time and money. It's essential to spend time reaching out to coffee business owners and learning from their experience; finding out what works, and what doesn't. And here's the fun part — it also means visiting lots of cafés to get an insight into what you want your business to be like. Consider what you will take from other businesses and what will make you different. Learn about your customer base. Who will they be? What are their needs? What time of the day will be busiest? Knowing your customers well will assist with planning, creating a menu, price points. Particularly, the location of the coffee shop is one of the most important decisions that will determine whether it will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Toronto to open a new coffee shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Recommend a good place to open a new coffee shop in the city of Toronto.

Data

To solve the problem, we will need the following data:

• List of neighborhoods in Toronto. This defines the scope of this project which is confined to the city of Toronto.

• Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to coffee restaurants. We will use this data to perform clustering of the neighborhood.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighborhoods in the city of Toronto, with a total of 103 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup package. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the coffee restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Toronto. Fortunately, 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 requests and beautifulsoup packages 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 be able to use Foursquare API. To do so, we will use the Geocoder package 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 DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check 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 100 venues that are within a radius of 2000 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 analyse 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 analysing the “Coffee Shop” data, we will filter the “Coffee Shop” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. 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 one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Coffee Shop”. The results will allow us to identify which neighborhoods have higher concentration of coffee shop while which neighborhoods have fewer number of coffee shop. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

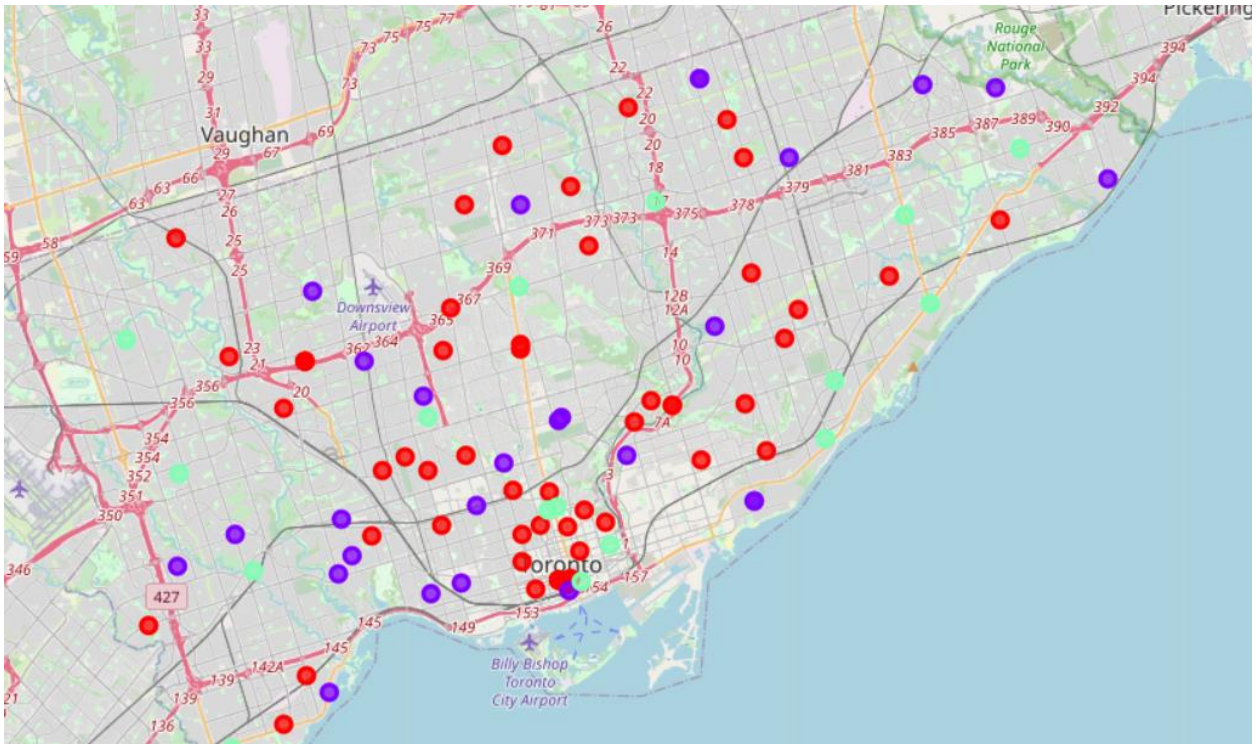
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Coffee Shop”:

• Cluster 0: Neighborhoods with moderate number of shopping malls

• Cluster 1: Neighborhoods with low number to no existence of shopping malls

• Cluster 2: Neighborhoods with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

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

Limitations and Suggestions 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. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new coffee shop. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

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 machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new coffee shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new coffee shop. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new coffee shop.

References

Category:Suburbs in the city of Toronto. Wikipedia. Retrieved from

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

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs