Applied Data Science Capstone Project

(IBM Data Science Professional Certificate Course)

Topic:

Best neighborhood to operate an independent coffee shop in Toronto

BY: Aditi M Jayakumar

|  |  |
| --- | --- |
|  | Table of Contents |
| 1 | Introduction |
| 2 | Description of Problem/Background |
| 3 | Description of Data |
| 4 | Data Analysis |
| 5 | Results |
| 6 | Conclusion |

**Introduction**

This is a "Capstone project" for the final course in the IBM Data Science Professional Certificate. This project aims to determine the best location to open an independent coffee shop using Python scripts, visualization libraries, pandas & numpy packages, FourSquare API, etc.

This project aims to enable aspiring entrepreneurs in making an informed choice in terms of location identification by taking into account various factors like population demographics, existing competitors and so on.

**Description of Problem/ Background**

For this project, Toronto has been chosen as it has a growing independent coffee shop scene. According to Wikipedia - *"Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario, while the Greater Toronto Area (GTA) proper had a 2016 population of 6,417,516. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world."*

As it is known, choosing a neighborhood and location for a coffee shop is an important and difficult step for the entrepreneurs. Most of time it can be as crucial factor as it determines the coffee shop's success and also affects the type of service, quality, menu, interior design.

To gather necessary insights, the following questions must be answered:

* How many cafes exist?
* What are the different boroughs in Toronto?
* What are the trending coffee shops?

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

### Description of Data

To identify the best location, the following data is required:

1. The list of Toronto neighborhoods and their boroughs. ('<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'>)
2. The postal codes of the neighborhoods. ('<https://cocl.us/Geospatial_data/Geospatial_Coordinates.csv'>)
3. The venues of a given neighborhood in Toronto.(<https://developer.foursquare.com/>)
4. The trending coffee shops in Toronto. (<https://developer.foursquare.com/>)

The List of neighborhood data will be used to obtain the exact coordinates for each neighborhood based on the postal code; this allows exploration of the city map. Then the coordinates and FourSquare credentials will be used to access the popular venues along with their details, especially for coffee shops.

The following sections of the Data have been extracted and used for analysis:

1. Coffee shops in Toronto.
2. Location of the coffee shops in the form of Latitude and Longitude coordinates.
3. Trending coffee shops in Toronto.

**Data Analysis**

The primary data required for analysis are the details regarding coffee shops in Toronto, which are gathered by making API calls to FourSquare services, specifically the 'search' call and ‘trending’ call. In the search call, information like latitude and longitude of the target area, i.e. Toronto, and the query, which was 'coffee', was passed. The get request containing the API call returned a json file with all the relevant data. The data is pre-processed and cleaned to generate a Pandas dataframe. The dataframe contains the following feature sets:

1. Name of the coffee shop
2. Latitude co-ordinate
3. Longitude co-ordinate

The latitude and longitude subset of the dataframe were fitted using KNN clustering algorithm. The number of clusters was set to 5. This creates 5 labeled clusters and the labels are then added to the dataframe.

Using matplotlib rainbow libraries, different colors were assigned to each cluster to help in visualization.

Using Folium, a map of the Toronto area is generated and a marker for each coffee shop is placed on the map. The markers in the each cluster are assigned unique colors in order to better visualize the clustering patterns.

Data Science tools used in this project includes:

* Panda Dataframe
* Numpy Package
* Seaborn Package
* BeautifuSoup
* Sklearn for KNN clustering
* Geopy for latitude and longitude extraction
* FourSquare APIs
* Folium &
* Matplotlib

## Results

## 

The KNN Clustering, Neighborhood segmentation and the Foursquare API calls enable us to segment the existing coffee shops in different boroughs.

Visualizing the KNN clusters in the map shown below shows us the distribution of coffee shops across Toronto.

From the map it can be said that:

There are 2 major clusters with most coffee shops denoted in purple and orange which indicates:

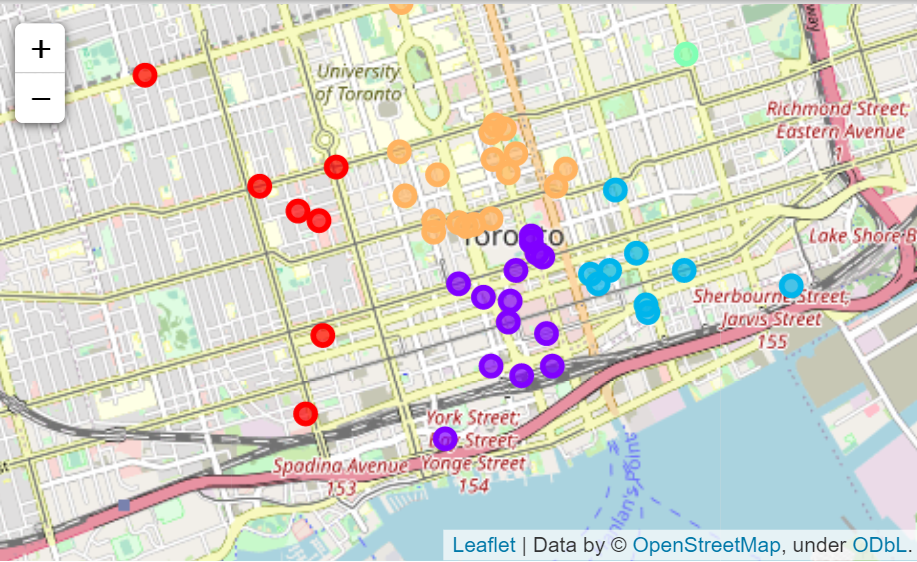
1. That the coffee shop scene is very active and may potentially be highly competitive and result in higher barrier to entry.
2. Also tell us that the coffee shop scene located in the blue, green or red clusters will have lesser competition, therefore reducing the barrier to entry.

It is also important to consider the target consumers' demographics, spending patterns and geographical distribution to derive a more holistic approach for the identification of the most suitable location.

It must be noted that identification and study of the trending coffee shops, their:

* location,
* product offerings,
* price range,
* customer profiles and
* proximity to parking or subway/tram/bus stops.

This will give further insights into developing a potentially profitable and successful business model by giving the entrepreneur a competitive edge.



Map of clustered coffee shops in Toronto

**Conclusion**

For this project python scripts, libraries, virtualization of data, FourSquare APIs are used to understand and extract relevant data.

The location data of coffee shops across Toronto are plotted on a map and clustered using KNN algorithm to show the number of coffee shops in different neighborhoods of Toronto. The data analysis should help stakeholders make decisions about locating and developing the appropriate business model.