Using Applied Data Science to Find Ideal Locations to Open a Coffee Bar

# Introduction

## Background

For the Applied Data Science Capstone project, I have a client who wants to expand their business by opening another coffee bar in the Toronto area. The client would like to be the only coffee bar on the block or street to limit their competition. Location is always an important factor in whether a business succeeds or fails.

## Business Problem

The objective of the project is to find the best location for the client to open a new coffee bar in Toronto. Data science methods and machine learning methods like clustering will be used in this project to answer the question: If an owner of several coffee bars wants to open a new location in Toronto, where would this location be?

## Target Audience

The target audience is for an owner that is looking for a new location for an additional coffee bar.

# Data

The following data will be needed to solve the problem:

* Neighborhoods in Toronto
* Latitude and longitude of the neighborhood.
* Venue data related to coffee bars

## Extracting Data

* Using Wikipedia to scrap Toronto neighborhoods
* Using Geocoder package to get the latitude and longitude
* Using Foursquare API to get venue data

# Methodology

# The first step is obtaining a list of neighborhoods in Toronto by extracting from the Wikipedia page [https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M,](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) a list of neighborhoods. I scraped the data from the web page using beautiful soup function importing into a pandas dataframe. Next, I needed location data, longitude and latitude coordinates for each neighborhood. I used a csv file from [http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data) to pull the geographical coordinates for each of the postal codes. Then I generated a map to visualize the neighborhoods and how they cluster together in Toronto.

Then to pull in data for the venues, I used the Foursquare API for a list of 10,000 places within 50,000 meter radius. Foursquare supplies the names, locations, and venue type of the surrounding businesses. To quickly investigate the amount of each venue type, I grouped all the venues by each neighborhood. The last steps were to use the k-means clustering method. By using k-mean clustering, I created 5 clusters based on the frequency of venues in the neighborhoods. Based on these clusters, I will be able to recommend where to open a new business.

# Results

Clusters

A picture containing text, map

Description automatically generated

The results from the clustering show that there are 5 clusters based on how many venues are in each neighborhood.

* Cluster 0 (red) has none or few coffee bars.
* Cluster 1 (purple) has a high number of coffee bars.
* Cluster 2 (dark blue) has none or few coffee bars.
* Cluster 3 (light blue) has none or few coffee bars.
* Cluster 4 (yellow) has none or few coffee bars.

# Discussion

Cluster 1 has a high number of coffee bars which is mostly centered around the downtown Toronto area and clusters 0, 2, 3, and 4 has a low number of coffee bars which is in the area of east Toronto (The Beaches) and central Toronto (Lawrence Park, Forest Hill, Moore Park, and Roselawn). The clusters show that recommending to the client that opening in these east Toronto and central Toronto neighborhoods will eliminate a battle between the competition.

# Conclusion

The Capstone Project stepped through the process of identifying a problem, recognizing the data that is required, extracting and cleaning the data, using machine learning, and delivering a recommendation to the client.

# References

List of Toronto neighborhoods:

[https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M,](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)