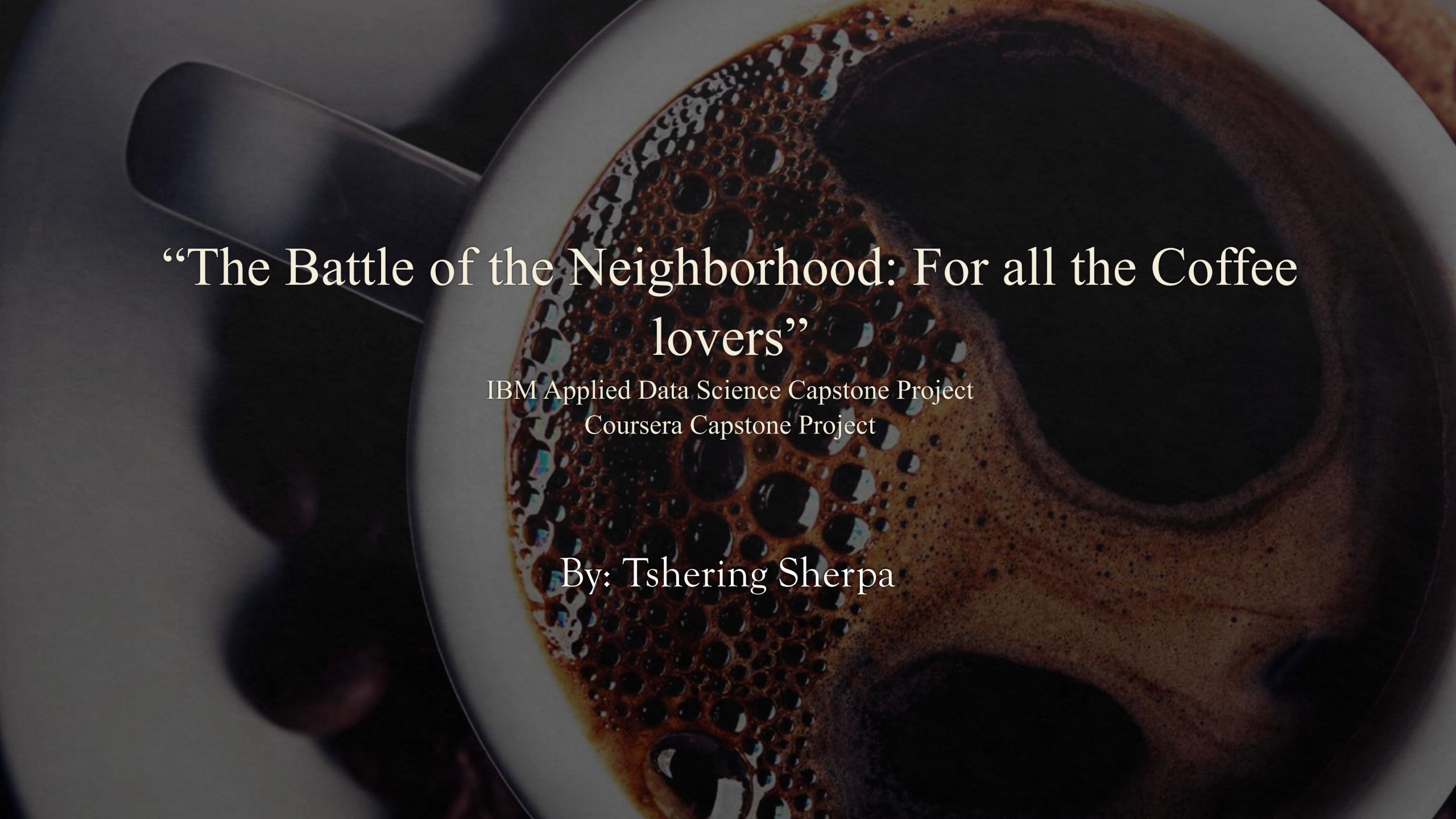


“The Battle of the Neighborhood: For all the Coffee lovers”

IBM Applied Data Science Capstone Project

Coursera Capstone Project

By: Tshering Sherpa

Agenda

- Introduction
- Business Problem
- Target Audience
- Data Description
- Methodology
- Results
- Discussion
- Limitation and Suggestion
- Conclusion
- References

Introduction

- Coffee is a brewed drink prepared from roasted coffee. If not, in-home people usually get coffee from coffee shops.

- Coffee is one of the top consumed beverage by people from all over the world and second most popular drink next to water.

- So if we need to open /own a coffee shop where can we start?

- The first thing is we need to gather data about the location and the competition around the desired location.

- This project will help gather information, analyze and have a conclusion to support the facts to help make a decision.

Business Problem

- The purpose of this capstone is to find out coffee shops around the area, so we need to gather information about the area and competition at the desired location.
- In this project we will collect information, analyze and select the best locations where we are interested in to open a coffee shop.
- I will be use data science methodology, Foursquare API location data services and machine learning technique to complete the project.
- For this project, I chose to research in Toronto and New York City and compare these two big cities.

Target Audience

- The outcome of this project can be used by anyone who loves coffee.
- It can be used for tourist or residents.
- This project will help investors or entrepreneurs who would like to open a new coffee shops in the city or at their desired location with similar model.

Data Description

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

Methodology

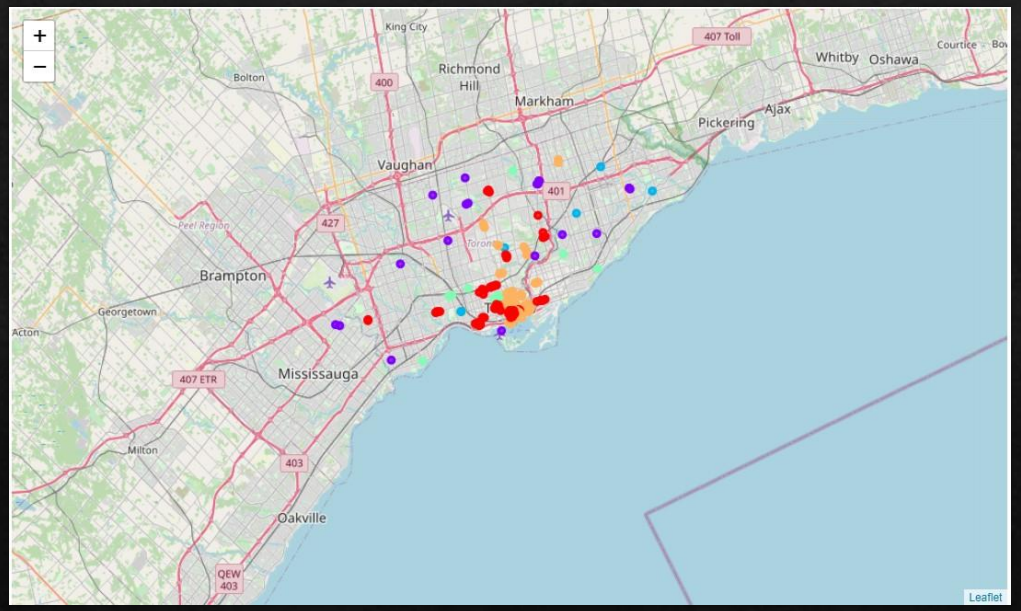
- We will do web scraping using Python with built in function of pandas library to extract the list of neighborhoods data.
- We will use the Geocoder library that will allow us to convert address into geographical coordinates in the form of latitude and longitude.
- We will populate the data into a pandas dataframe and then visualize the neighborhoods in a map using Folium library
- Finally we will use Foursquare API to get the nearby venue of neighborhoods.

Methodology (contd.)

- We will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.
- We will filter the “Coffee shop” as venue category for the neighborhoods. Also create list with filter words like 'Coffee Shop', 'Bagel Shop', 'Donut', 'Breakfast Spot', 'Cafeteria', 'Cafe’ to put it in coffee shop category.
- Finally, we will use k-means clustering (popular unsupervised machine learning algorithm) to cluster neighborhood with similar behavior.

Results

Visualization of clustered data

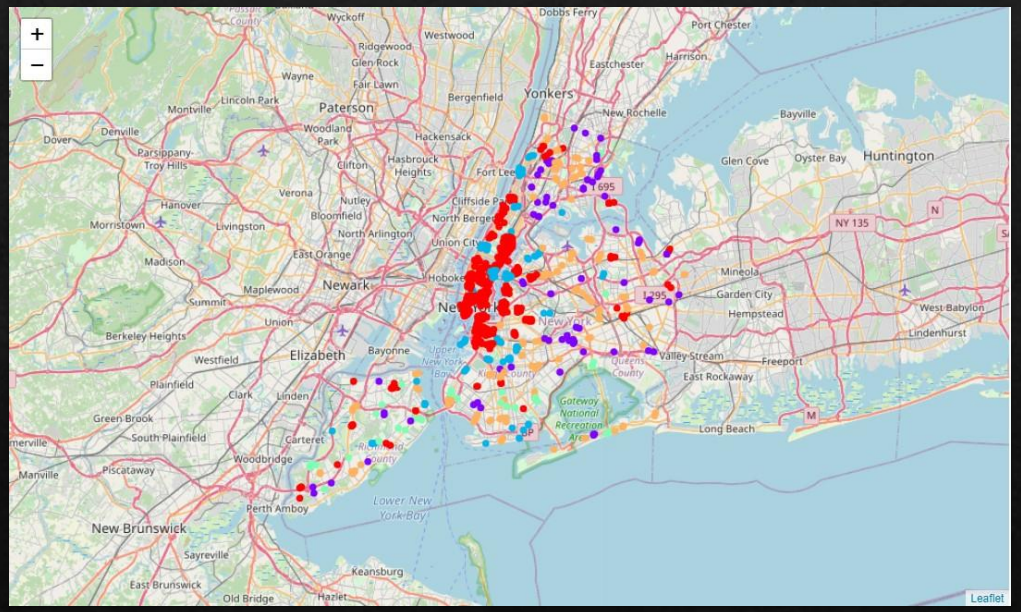


Description of clusters

- 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.
- Green Cluster 3 has low concentration of Cafe.
- Orange Cluster 4 has high concentration of Coffee shops.

Results (contd.)

Visualization of clustered data



Description of clusters

- Red Cluster 0 has high concentration of coffee shops.
- Purple Cluster 1 has high concentration of Donut shops.
- Blue Cluster 2 has high concentration of Cafe.
- Green Cluster 3 has moderate concentration of Bagel shops.
- Orange Cluster 4 has high concentration of Donut shops

Discussion

- Most of the coffee shops are concentrated in the downtown area of Toronto and New York city, with the highest number in cluster 0 for both cities.
- Comparing to Toronto and New York city, my assumptions were true. New York has far more coffee shops than Toronto. New York has more categories of venue that are related to coffee as well.
- 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 so are advised to avoid these neighborhoods.
- For Toronto:

    1. Clusters 2 and 3 has very low number of coffee shops and are widely spread out the city . Therefore there will be good opportunity to open a coffee shop.
    2. Clusters 0,1 and 4 has moderate to high concentration of coffee shop suffering from intense competition and property developers are advised to avoid neighborhoods in cluster.

Limitation and Suggestion

Limitation

- There are other factors such as population and income of residents that could influence the location decision of a new coffee shops which is not include in this project.
- The crime rate in the city and ratings of the coffee shop can also influence the decision which are also not included.
- With the free account of Foursquare API, there are limited number of request and response in a day which made this project difficult to complete in time.
- With short amount of time it is difficult to complete this project before the deadline.

Suggestion

- We need more data that has population, area of the neighborhood, crime rate, other trendy venue that will support the outcomes.
- We can use paid account for Foursquare API for more requests and response.
- Longer deadline or start of the project in the beginning of the course or capstone can be very helpful.
- Use multiple machine learning algorithm to get different and better results can be very helpful as well.

Conclusion

- 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.
- Clustered the data into 5 clusters based on their similarities which was best suited for data.
- Provided recommendations to the relevant stakeholders i.e. property developers, entrepreneurs and investors regarding the best locations to open a new coffee shops.
- Answered the business question that was raised in the introduction section which 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.

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