Capstone Project - The Battle of Neighborhoods

JUNE 9

Coursera Project
Authored by: Prashanth H Siddaraju

Contents

Introduction:

Problem Description

Lot of stores are selling coffee in New York and Toronto, but there are very fewer coffee shops which only sell coffee (example- Starbucks).

Questions which required to be answered:

1) **If any coffee company or a coffeehouse chain is planning to open a new store in New York or Toronto, where can they open the same? Which are the probable Neighborhoods in which coffee is already being sold by a non-Coffee store, but have a possibility to have a coffee store?**
2) **Which coffee drinking neighborhoods in New York and Toronto are similar, so the same business model can be implemented in the other city as well?**

Business Stakeholders:

1. **Coffee company who would like to open a new coffee shop in New York.**
2. **Existing coffee shops which would like to improve there business model.**

Background of the issue

According to thecoffeebump statistics

- **58% of Americans over the age of 18 drink coffee every single day**
- **The US alone spends $4 billion importing coffee each year**
- **The average coffee drinker spends $164.71 on coffee each year**
- **100 million people in the US drink coffee each day**
- **25% of Americans skip breakfast, yet 50% will still consume a cup of coffee in the morning**

Specialty coffee sales are increasing by 20% per year and account for nearly 8% of the 18 billion-dollar U.S. coffee market. Major coffee companies like Starbucks, Espresso house are planning to open more and more stores. But coffee is already being sold in Neighborhood by general shops, like Donut store, Mc Donald's, KFC etc.… which proves a fact that there is an increasing demand for coffee in these areas. If major

coffee companies can plan to open their outlets in these places, it could lead to better sales and profit.

Second issue being which business model to be selected for a specific neighborhood while opening the store. Business model / Sales strategy place a vital role in starting and improving the business in any place. There is a need to know which business model would help to improve the business here depending on coffee drinking habits, by comparing the same which is already implemented in a different city.

Methodology

Description of Data and Solution

We would be needing the Neighborhood information of 2 Cities here

1) **New York**
2) **Toronto**

New York neighborhood data with latitude and longitudes are already available in the link - https://cocl.us/new_york_dataset
This data is json format, this must be converted into Tabular format.
Data Would look something like this:

	Borough	Neighborhood	Latitude	Longitude	Place
0	Bronx	Wakefield	40.894705	-73.847201	Newyork
1	Bronx	Co-op City	40.874294	-73.829939	Newyork
2	Bronx	Eastchester	40.887556	-73.827806	Newyork
3	Bronx	Fieldston	40.895437	-73.905643	Newyork
4	Bronx	Riverdale	40.890834	-73.912585	Newyork

Toronto dataset is not readily available, but we can find the neighborhood information with zip codes Wikipedia page
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
This must be web scraped and cleansed before the same can be used.

- **Drop the postcodes with Borough's 'Not Assigned'**

- **Group the Neighborhood with more than one post code**
- **Replace Neighborhood names which are 'Not Assigned' with respective Borough names.**

Data would look something like this:

:

	Borough	Neighborhood	Latitude	Longitude	Place
0	Scarborough	Malvern, Rouge	43.806686	-79.194353	Toronto
1	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Toronto
2	Scarborough	Morningside, Guildwood, West Hill	43.763573	-79.188711	Toronto
3	Scarborough	Woburn	43.770992	-79.216917	Toronto
4	Scarborough	Cedarbrae	43.773136	-79.239476	Toronto

Latitude and longitude data for Toronto neighborhoods can be found in this location http://cocl.us/Geospatial_data
Once the above 2 data sets are gathered, we would require Venue details which are selling coffee around 500 meters of the above-mentioned neighborhoods. This data can be acquired from foursquare using explore api with relevant filters.

The Get url would look something like below :-
https://api.foursquare.com/v2/venues/explore?&client_id=<Your Client id> &client_secret=<yourclientsecret>&v=<version>&ll=<latitude>,<longitude?&radius=500&limit=50&section=coffee

Once the data is cleansed and Venue details are captured, the same can be split into 2 data frames

1) **coffee companies selling coffee**
2) **Other Outlets selling coffee**

The same needs to be visualized using Folium on a New York map to see the distribution of Non-Coffee shops, and the same can be used to estimate the possibility of opening a coffee shop in the vicinity.

We can also do the same analysis for Toronto, which later can be used to cluster with New York to see which Coffee Drinking Neighborhood in Toronto is similar to a neighborhood in New York.

Exploratory Data Analysis:

Toronto data on Wikipedia should be first converted into useable data frame, using some web scraping tool. We can here use a python library called 'Beautiful Soup' for accomplishing this task.

Data on Wikipedia:

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor

Web Scraped and processed data frame

	Borough	Neighborhood	Latitude	Longitude	Place
0	Scarborough	Malvern, Rouge	43.806686	-79.194353	Toronto
1	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Toronto
2	Scarborough	Morningside, Guildwood, West Hill	43.763573	-79.188711	Toronto
3	Scarborough	Woburn	43.770992	-79.216917	Toronto
4	Scarborough	Cedarbrae	43.773136	-79.239476	Toronto

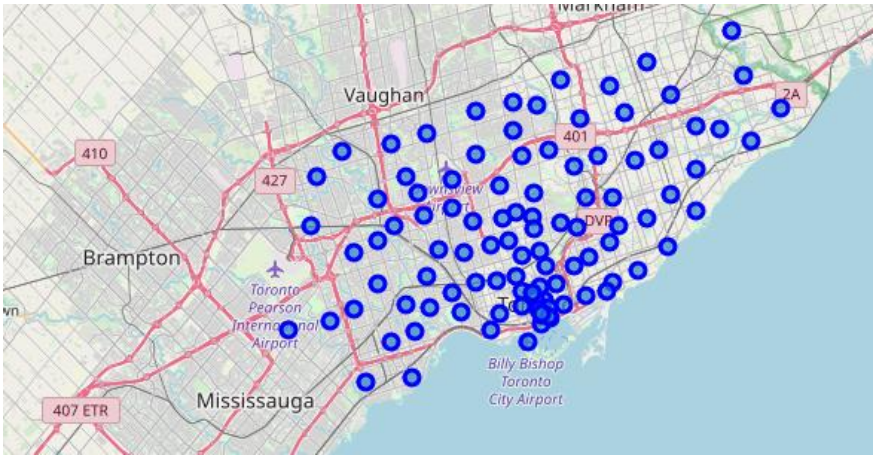
New York data is already preprocessed and is available in the above mentioned link. But the file is in JSON format, which needs to be converted into a data frame. Only consider the columns *['Borough', 'Neighborhood', 'Latitude', 'Longitude'].*

New York data after processing

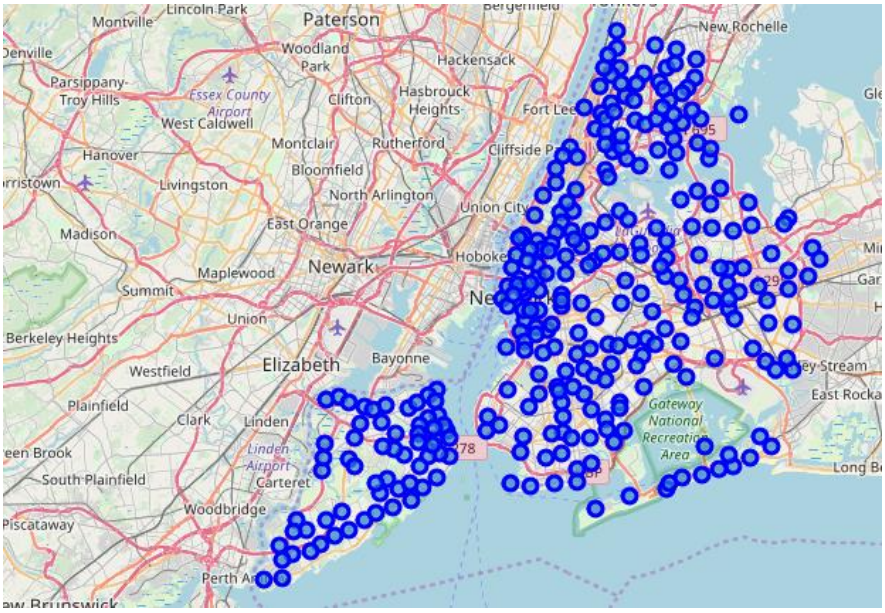
	Borough	Neighborhood	Latitude	Longitude	Place
0	Bronx	Wakefield	40.894705	-73.847201	Newyork
1	Bronx	Co-op City	40.874294	-73.829939	Newyork
2	Bronx	Eastchester	40.887556	-73.827806	Newyork
3	Bronx	Fieldston	40.895437	-73.905643	Newyork
4	Bronx	Riverdale	40.890834	-73.912585	Newyork

Data can be visualized using Folium library as shown below. We can plot each neighborhood in Toronto and New York using the same. Below are the two plots of Toronto and New York Neighborhoods

Toronto



New York



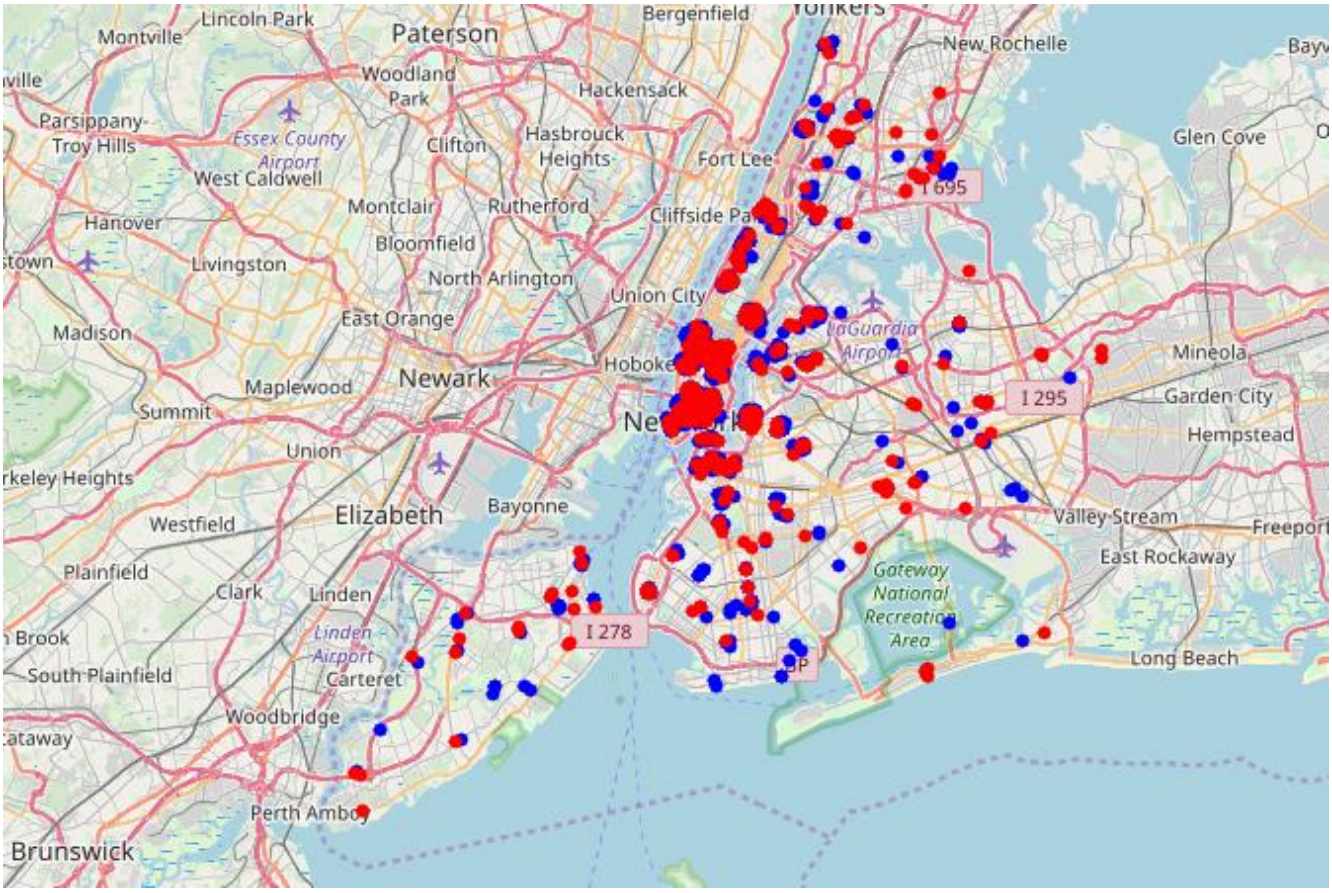
Foursquare will be used here to get the nearby Venues , which serve coffee. We would have to later distinguish the Actual Coffee Shops from Generic stores selling coffee. We can here use Foursquare API explore function to explore the venues around each neighborhood, with section as coffee.

https://api.foursquare.com/v2/venues/explore?&client_id=<Your Client id> &client_secret=<yourclientsecret>&v=<version>&ll=<latitude>,<longitude?&radius=500&limit=50&section=coffee

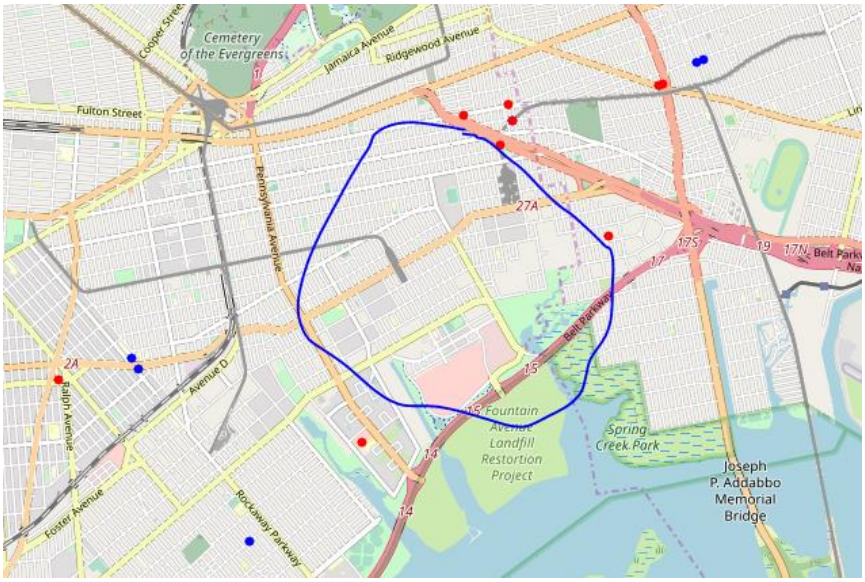
The output of this would be json which needs to be later processed into data frame similarly as we did for New York data.

	Neighborhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
1	Eastchester	40.887556	-73.827806	Dunkin'	40.885384	-73.828099	Donut Shop
2	Kingsbridge	40.881687	-73.902818	Mon Amour Coffee & Wine	40.885009	-73.900332	Coffee Shop
3	Kingsbridge	40.881687	-73.902818	Sugarboy Bakery Cafe	40.877832	-73.902669	Bakery
4	Kingsbridge	40.881687	-73.902818	Dunkin'	40.879308	-73.905066	Donut Shop

Here we can notice that Mon Amour Coffee & Wine is a coffee shop, but rest others are also selling coffee, but are not main stream in coffee business. This is just a sample data, we have approx. 4000 records only its New York based on which the model has to be built.
This requires a descriptive model, and it would take un supervised learning approach, which would lead us to clustering. First part, we have to manually cluster the stores which are coffee shops and differentiate the same from which are not.
The final result would look as in the below image



Here if we zoom in to check the distribution, we can notice that in few areas, we have more non-coffee shops selling coffee than coffee shops, which are hot spots where coffee shops can be opened. Below is one such area where there are no coffee shops



For the next problem, we would first have to cluster the neighborhoods of New York and Toronto together to check which neighborhoods are more similar. Based on the cluster forms we can decide weather the same business model works in similar cities.

Note: Here we are not taking economic factors and population into consideration. It has to be considered for the model to be accurate.
We can consider 4 Clusters to start with, and use KMeans Clustering approach to solve this problem.

We need to one-hot encode the venue details for the same before starting.
Once we cluster the neighborhoods, we go ahead and check the clusters.

Cluster 1

```
36]: new_tor_merged[new_tor_merged['Cluster Labels'] == 0]
```

36]:

	Neighborhood	Place	Cluster Labels
3	Alderwood, Long Branch	Toronto	0
5	Astoria	Newyork	0
8	Bay Ridge	Newyork	0
12	Bedford Park, Lawrence Manor East	Toronto	0
21	Bulls Head	Newyork	0
22	Bushwick	Newyork	0
23	Business Reply Mail Processing Centre 969 Eastern	Toronto	0
24	Butler Manor	Newyork	0
28	Central Bay Street	Toronto	0

We can notice that, above mentioned clusters in New York and Toronto are similar , considering coffee shops and non coffee shops only. Similarly we would be ending up with 3 more clusters. We noticed that 4th Cluster was having only New York neighborhoods, which are not similar to any of the neighborhood in Toronto.

Results

1) **By Clustering Coffee Shops and Non Coffee Shops, which are selling coffee, we were able to plot the distribution of the same on a map. This can be used to with other factors to decide where to open the next Coffee Shop**
2) **By clustering the neighborhoods for Toronto and New York, we were able to find similar neighborhoods in both the Cities. This helps in deciding the business model for coffee shops , provided other factors like economic index, population are considered too.**

Conclusion

We hereby can conclude that there are similar neighborhoods in New York and Toronto based on demand for Coffee shops in both the cities, which would help business to build better profit models. We also proved that by looking at the distribution of coffee and non-coffee shops , and clustering the same, we were able to make decisions on probable places for opening main stream coffee shops.