COURSERA CAPSTONE

IBM Data Science Capstone

FINDING THE LOCATIONS OF PHARMACIES, GROCERY STORES AND DEPARTMENT STORES IN TORONTO DURING COVID-19 LOCK DOWN.

- SAI KRISHNA ADUSUMILLI

I. INTRODUCTION

The lock down due to COVID-19 has really became a primary problem for people. People are finding difficulties for grocery stores and pharmacy stores which are in-active.

This project fetches the list of grocery stores and pharmacy stores in Toronto Area. Therefore it might be helpful for an individual who is in need of groceries and medicines. With the purpose in mind, this project is designed to find all the locations of grocery stores and pharmacy stores in and neighbourhood of Toronto.

BUSINESS PROBLEM

During the lock down, the need of groceries and medicines has become primary concern for people.

The main objective is to find ideal locations in the city for the people looking for grocery stores and pharmacy stores .

TARGET AUDIENCE OF THIS PROJECT

This project is particularly useful for the people living in Toronto, who are seeking to find pharmacies, grocery stores and department stores during this COVID-19 lock down.

II. DATA

The data needed for this project :

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods. (It is provided in COURSERA Data Science Capstone Project week 3 as .csv file)
- Venue data related to Groceries and pharmacy stores.

Sources of data :

- Scrapping of Toronto neighborhoods via Wikipedia. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Getting Latitude and Longitude data of these neighborhoods. (http://cocl.us/Geospatial_data)
- Using Foursquare API to get venue data related to these neighborhoods.

Description and extraction of data :

- Fetching the data from the web page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) which contains a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario.
- Using beautifulsoup package, the data from above web page is scrapped and converted into a pandas data frame.
- Data cleaning for the obtained data frame is done by removing the rows those contain Borough value as 'Not Assigned' and then concatenating the data which contains same value of Postal Code &

Borough.

- Data from (http://cocl.us/Geospatial_data) contains latitude and longitude of places based on the Postal Codes. This link returns a .csv file.
- The data is downloaded from the above web page and the .csv file data is converted into data frame using pandas.
- Now, the two data frames are merged based on Postal Codes.
- Venue data of a place is collected using foursquare.

III. METHODOLOGY

Python Data Science tools are used to analyze data. Complete code can be found here- https://github.com/Krish716/CourseraIBM_Capstone/blob/master/CapstoneFinal.ipynb

First insight using visualization:

- Map of Toronto and its neighbourhood is visualized using the data frame which contains list of places in and around Toronto, Canada.

  The figure.1 shows the map illustrates the above point.

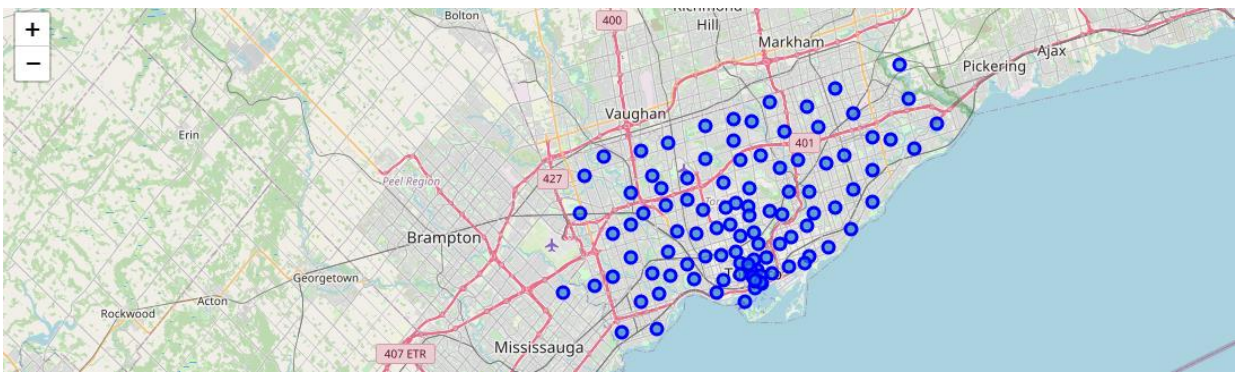

Figure.1

- Now fetching the data which contains Toronto in Borough column and visualizing those places. This is shown in figure.2 .

Figure.2

- Using foursquare, the venue data for the places is collected by the url- "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}" .
- Now **one hot encoding** is performed to analyze neighbourhood using postal codes and venue data.The following table.1 shows the table after one hot encoding.

	PostalCode	Borough	Neighborhoods	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Win Ba
0	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	
1	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
3	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
4	M4K	East Toronto	The Danforth West / Riverdale	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	

5 rows × 233 columns

Table.1

- Then, we fetch top 10 venues for each area noted in the data frame.The data frame looks like below table.2.

	PostalCode	Borough	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
27	M5V	Downtown Toronto	CN Tower / King and Spadina / Railway Lands / Harbourfront West / Bathurst Quay / South Niagara / Island airport	Airport Lounge	Airport Service	Airport Terminal	Airport	Harbor / Marina	Coffee Shop	Plane	Rental Car Location	Sculpture Garden	Boat or Ferry
32	M6J	West Toronto	Little Portugal / Trinity	Bar	Restaurant	Café	Vietnamese Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Men's Store	Yoga Studio	Japanese Restaurant	Juice Bar
25	M5S	Downtown Toronto	University of Toronto / Harbord	Café	Bookstore	Bar	Italian Restaurant	Japanese Restaurant	Bakery	Restaurant	Sandwich Place	Beer Bar	Beer Store
3	M4M	East Toronto	Studio District	Café	Coffee Shop	Gastropub	Brewery	Bakery	American Restaurant	Neighborhood	Sandwich Place	Cheese Shop	Pet Store
15	M5C	Downtown Toronto	St. James Town	Café	Coffee Shop	Gastropub	Cocktail Bar	American Restaurant	Hotel	Italian Restaurant	Clothing Store	Art Gallery	Seafood Restaurant

Table.2

- Then clustering is performed on the data frame to cluster the areas located in the Toronto.
- On the basis of Silhouette scores, the optimum value of k in **K-means** Clustering is obtained. The plot between number of clusters and Silhouette scores is shown in figure.3 .

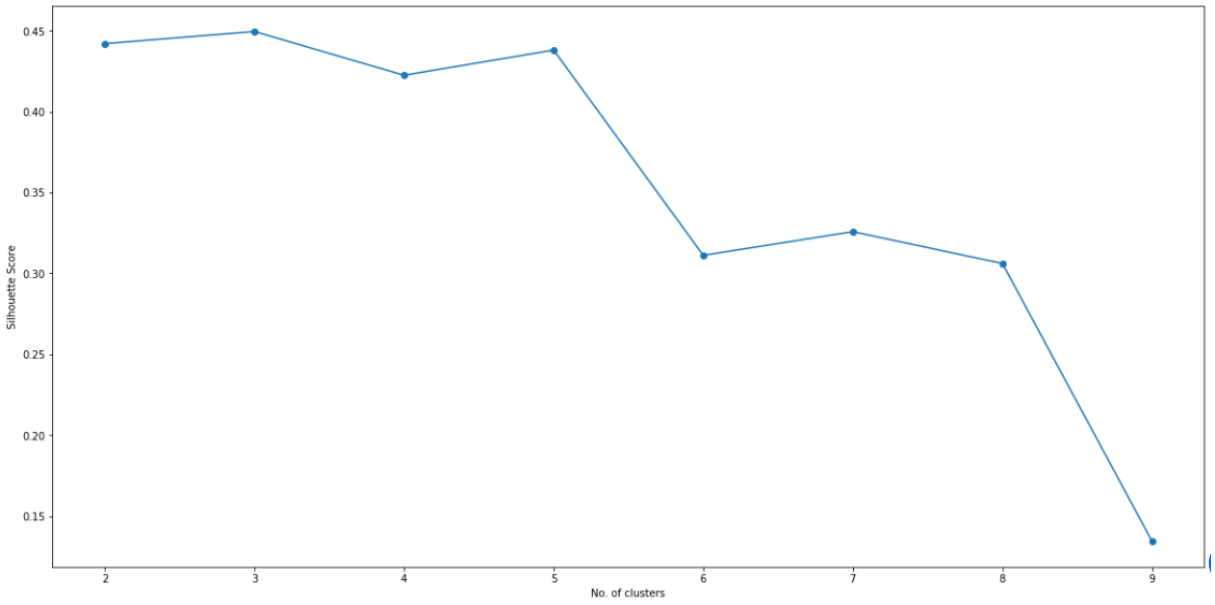

Figure.3

- After getting the k value, respective clusters will be assigned to the areas in the data frame. The clusters obtained is shown in figure.4 .

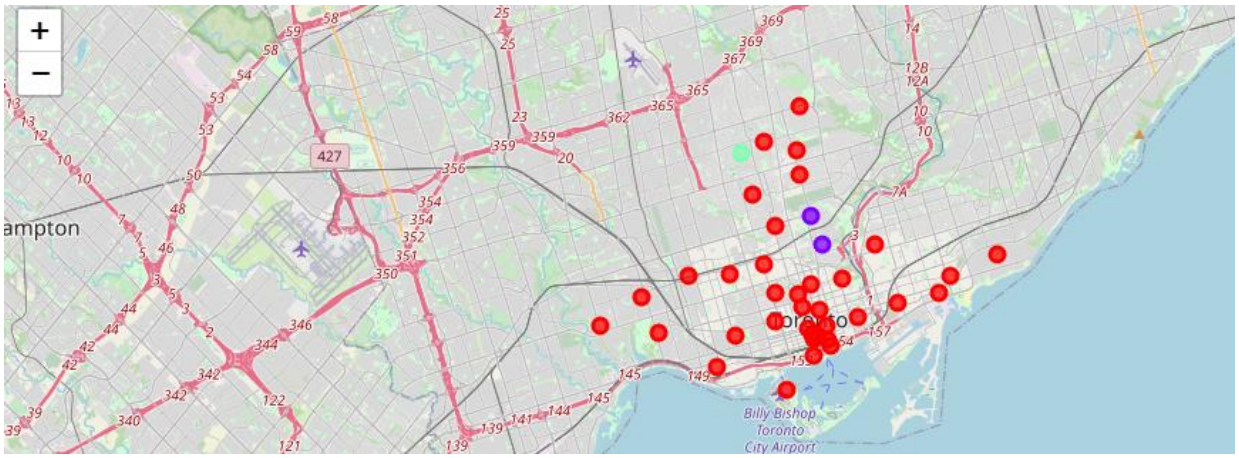

Figure.4

- Now the above steps are performed to identify Pharmacies, grocery stores and department stores in respective clusters.

IV. RESULTS

After obtaining the data frames for each venue category - Pharmacy, Grocery Store and Department Store. The resulting maps by clusters for Pharmacy, Grocery Store and Department Store is shown in figure.5, figure.6 and figure.7 respectively.

Figure.5

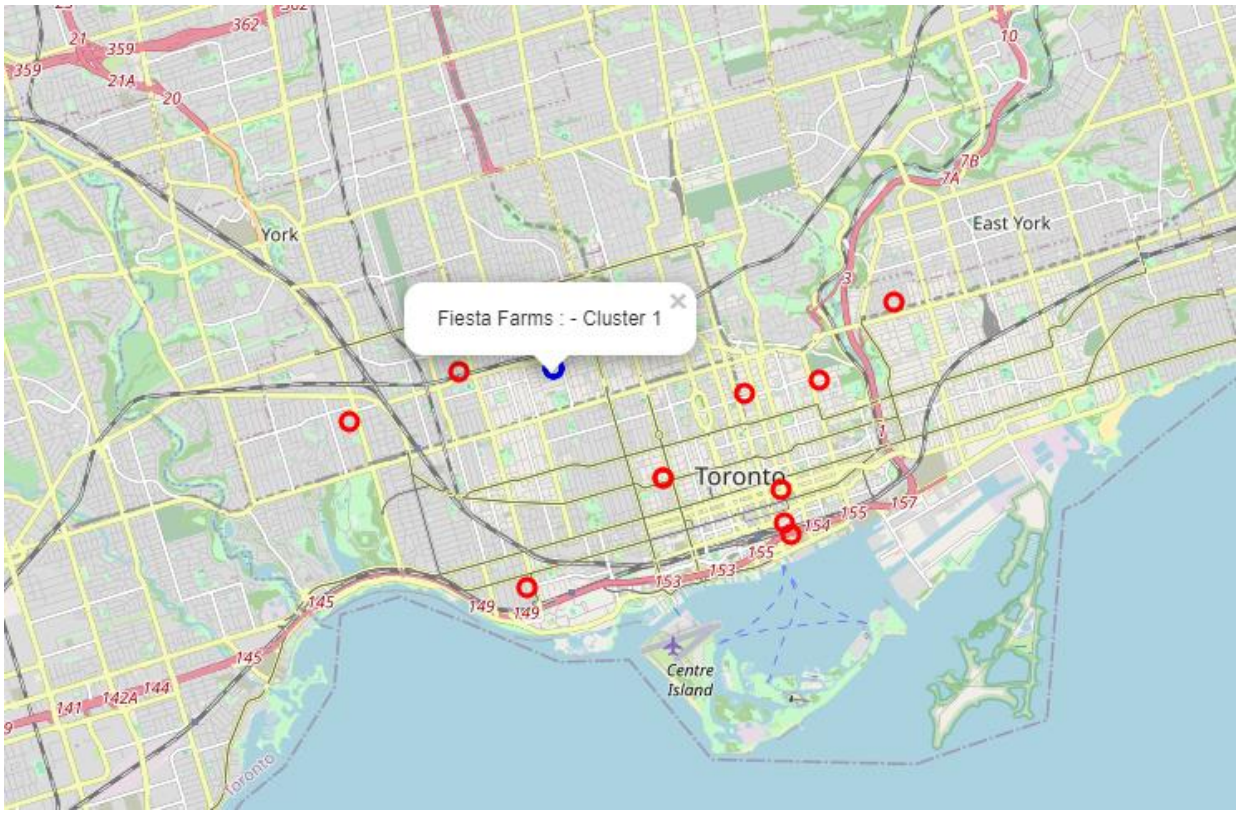

Figure.6

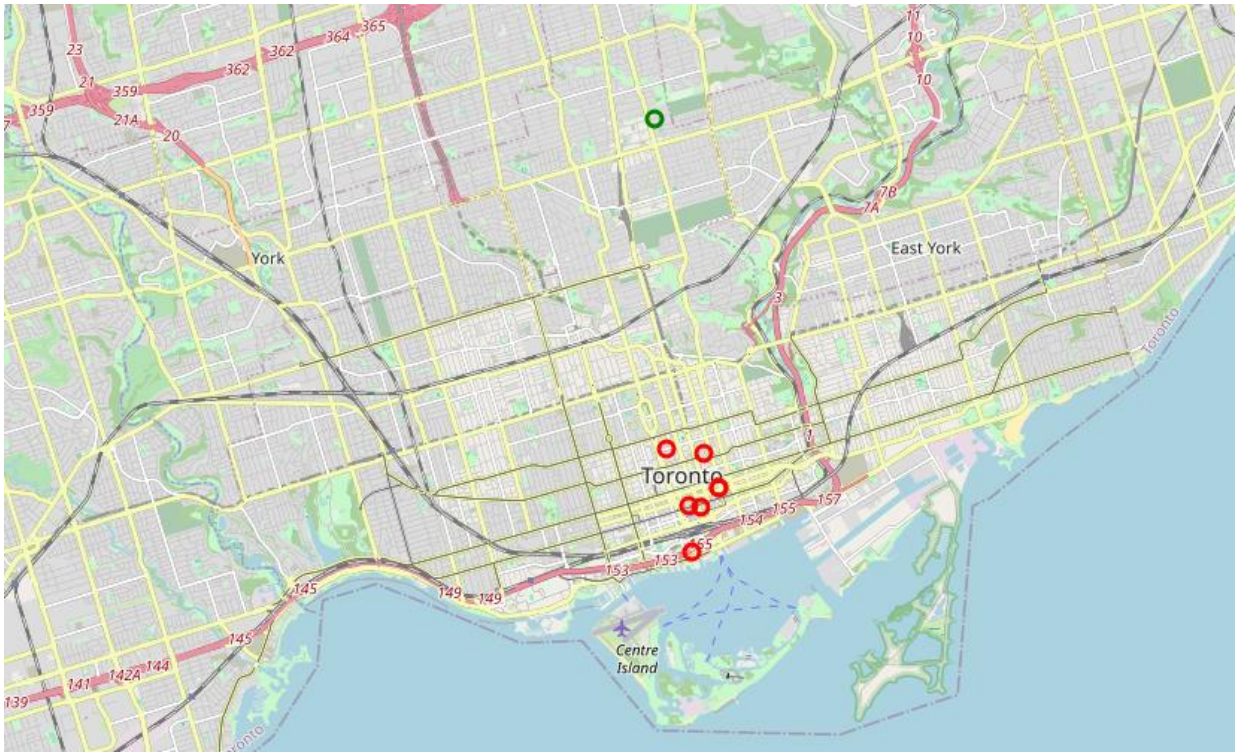

Figure. 7

V. DISCUSSION

The real challenge is constructing the data set-

- The required data is not 100% accurate as it is not available publicly.
- The geocoding API is limited which is not free anymore. Need to find a similar API with free services.
- After the data set has been constructed, different results are returned with different set of parameters at different timestamps because of foursquare API usage.

It can be considered as most important process in a Data Science Pipeline project.

On the other hand, choosing the optimum method to construct a model is totally worth.

VI. CONCLUSION

1.Pharmacy-

There are 7 pharmacies located in cluster 0 where in cluster 1 & cluster 2, there are no pharmacies.

2.Grocery Store-

There are 11 grocery stores located in cluster 0 and 4 in cluster 1. In cluster 2,

there are grocery stores.

3. Department Store-

There are 7 department stores located in cluster 0 and 1 in cluster 0. In cluster 1, there are grocery stores.

For more cluster analysis and result set, visit-

https://github.com/Krish716/CourseraIBM_Capstone/blob/master/CapstoneFinal.ipynb

This provide provides all the locations of Pharmacies, Grocery Stores and Department Stores in Toronto based on the data set obtained.