Coursera - IBM Applied Data Science Capstone Project

Neighbourhoods that may require new Pharmacies in Toronto

By: Veena C

Introduction

My interest in Data Analysis and Data Science resulted in me pursuing the IBM Data Science Professional certificate course on coursera. The last module of the course is the capstone project which intends to demonstrate the skills ,tools, techniques and knowledge acquired on learning of the course to solve a real world business problem

Business Problem

While analysing the Toronto neighbourhood data in the previous courses, it was observed that most neighbourhoods around Toronto had more recreational venues like restaurants, bars, theatres etc..and did not have many pharmacies and hospitals. This led me to think of the business problem that iam addressing as part of this capstone project, which is to identify potential neighbourhoods to open new pharmacies in the interest of the public.

Methodology

Data

Data used for this project analysis were taken from the following sources.

1. The Toronto neighbourhood data: From Wikipedia page
2. Geospatial Data for neighbourhood coordinates from csv file : Coursera

3. Venues for all Toronto Neighbourhoods via Foursquare API : https://api.foursquare.com/v2/venues

Data Gathering and Preparation

Data was gathered from the sources as mentioned in the Data Section. The Canada postal codes data was filtered for only Toronto neighbourhoods and further analysis and processing was carried on the Toronto Neighbourhood subset data.

Exploratory Data Analysis

The analysis was done using using Jupiter Notebooks, with Python and supporting libraries as follows: pandas, numpy, beautifulsoup, matplotlib, folium

Data was gathered from wikipwdia site, csv files and venue location data from foursquare API. Care was taken to prepare the data by dropping columns that were not required and bridging all data to a format that deemed fit for analysis. This analysis was done only for Toronto neighbourhoods. The bvenues for each of the neighbourhoods were taken suing the Foursquare API and then merged along with the neighbourhood location details to form the complete dataset

The below table shows the structure of the overall dataset.

Out[13]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

Out[12]:

	Neighborhoods	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	...	Tibetan Restaurant	Toy / Game Store	Trail	Tra Static
0	Berczy Park	0.0000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.017241	...	0.00000	0.000000	0.00000	0.00
1	Brockton, Parkdale Village, Exhibition Place	0.0000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000	...	0.00000	0.000000	0.00000	0.00
2	CN Tower, King and Spadina, Railway Lands, Har...	0.0625	0.0625	0.125	0.1875	0.125	0.000000	0.000000	0.00	0.000000	...	0.00000	0.000000	0.00000	0.00

The data was further processed to group the neighbourhoods and the frequency of each of the venues was determined.

Out[11]:

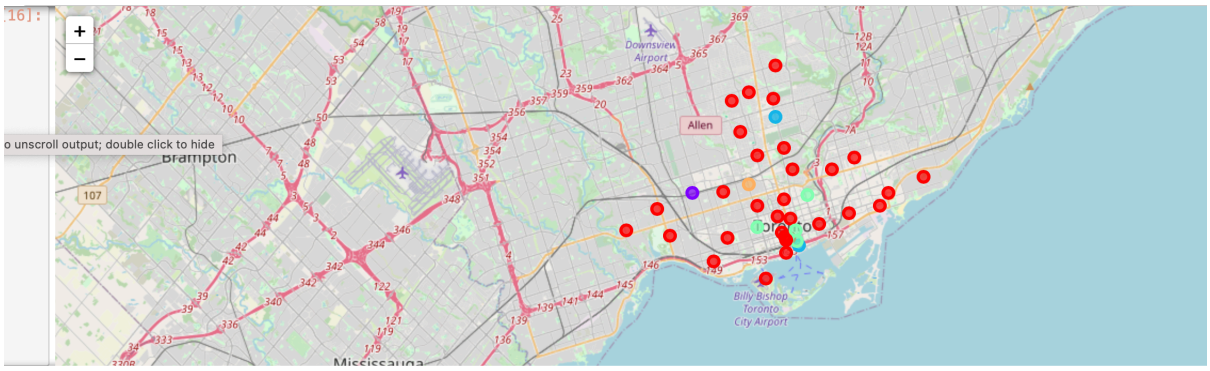
	Neighborhoods	Pharmacy
0	Berczy Park	0.035088
1	Brockton, Parkdale Village, Exhibition Place	0.000000
2	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
3	Central Bay Street	0.000000
4	Christie	0.000000

Model Building

Clustering method, which is an unsupervised machine learning method was used for this analysis. Clustering interprets the input data to find groups or clusters based on the data available.

K-means algorithm was used of this analysis.

For this analysis, 5 clusters were defined and the neighbourhood data was assigned according to these 5 clusters.



Observations

Based on the results of K-means clustering algorithm, all the neighbourhoods were assigned to one of the 5 clusters based on the similarity. Each cluster was further analysed to understand how the current pharmacies were distributed in the neighbourhoods.

Based on the below observations, cluster 1 is where there is potential to put up new pharmacies for the benefit of the people.

The following observations were made,

Cluster 0 : Has a total of 31 neighbourhoods and zero pharmacies.

	Neighborhood	Pharmacy	Cluster Labels	PostalCode	Borough	Latitude	Longitude
1	Brockton, Parkdale Village, Exhibition Place	0.0	0	M6K	West Toronto	43.636847	-79.428191
2	CN Tower, King and Spadina, Railway Lands, Har...	0.0	0	M5V	Downtown Toronto	43.628947	-79.394420
3	Central Bay Street	0.0	0	M5G	Downtown Toronto	43.657952	-79.387383
4	Christie	0.0	0	M6G	Downtown Toronto	43.669542	-79.422564
5	Church and Wellesley	0.0	0	M4Y	Downtown Toronto	43.665860	-79.383160
6	Commerce Court, Victoria Hotel	0.0	0	M5L	Downtown Toronto	43.648198	-79.379817
8	Davisville North	0.0	0	M4P	Central Toronto	43.712751	-79.390197
10	Enclave of M4L	0.0	0	M7Y	East Toronto Business	43.662744	-79.321558
12	First Canadian Place, Underground city	0.0	0	M5X	Downtown Toronto	43.648429	-79.382280
13	Forest Hill North & West	0.0	0	M5P	Central Toronto	43.696948	-79.411307
14	Garden District, Ryerson	0.0	0	M5B	Downtown Toronto	43.657162	-79.378937
15	Harbourfront East, Union Station, Toronto Islands	0.0	0	M5J	Downtown Toronto	43.640816	-79.381752

Cluster 1: Has a total of 1 neighbourhoods and 1 pharmacy.

```
]: Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 1]
18]:
```

	Neighborhood	Pharmacy	Cluster Labels	PostalCode	Borough	Latitude	Longitude
9	Dufferin, Dovercourt Village	0.142857	1	M6H	West Toronto	43.669005	-79.442259

Cluster 2: Has a total of 2 neighbourhoods and 1 pharmacy each.

```
]: Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 2]
19]:
```

	Neighborhood	Pharmacy	Cluster Labels	PostalCode	Borough	Latitude	Longitude
0	Berczy Park	0.035088	2	M5E	Downtown Toronto	43.644771	-79.373306
7	Davisville	0.030303	2	M4S	Central Toronto	43.704324	-79.388790

Cluster 3: Has a total of 4 neighbourhoods and 1 pharmacy each.

```
: Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 3]
```

20]:

	Neighborhood	Pharmacy	Cluster Labels	PostalCode	Borough	Latitude	Longitude
11	Enclave of M5E	0.020408	3	M5W	Downtown Toronto Stn A	43.646435	-79.374846
18	Kensington Market, Chinatown, Grange Park	0.014493	3	M5T	Downtown Toronto	43.653206	-79.400049
29	St. James Town	0.012346	3	M5C	Downtown Toronto	43.651494	-79.375418
30	St. James Town, Cabbagetown	0.023256	3	M4X	Downtown Toronto	43.667967	-79.367675

Cluster 4: Has a total of 1 neighbourhoods and 1 pharmacy

30	St. James Town, Cabbagetown	0.023256	3	M4X	Downtown Toronto	43.667967	-79.367675

```
Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 4]
```

1]:

	Neighborhood	Pharmacy	Cluster Labels	PostalCode	Borough	Latitude	Longitude
33	The Annex, North Midtown, Yorkville	0.047619	4	M5R	Central Toronto	43.67271	-79.405678

Conclusion

This project was to come up with the solution for a hypothetical business problem. The analysis was performed based on the undertsnadoing gained through the learning pf this course. The output of the analysis was the basis for the observations. This project has scope for further improvement and refinement to get into more details . I will continue to explore this project further to strengthen my knowledge on Data Science.