Introduction:
In this capstone project for Applied Data Science, I would use techniques such as clustering to find an area where to open your own business is something that every new business owner struggles with. Toronto is a center of business, finance, arts and culture. Toronto is a city that is not only large but also with diverse communities. My mother is a pharmacist from India and we have been thinking for long time what could be the best location to open a pharmacy for her in the city where we have spent our time since we immigrated. This project would not only help my family make an important decision but also to all other people who want to open their own pharmacy in a beautiful city like Toronto.

Business problem:
I want to find the best area for my family and many other to open a Pharmacy in Toronto which is the provincial capital of Ontario. This project is designed to provide solution to the problem by using data science methods and tools I have learned during the course along with the machine learning algorithm, K-means Clustering.

Target Audience:
Entrepreneurs
Business owners looking for expansion in Toronto
New Business owners
Professional looking for relocation

Data:
Let me provide you with a list of data that would be required to solve this problem.

1. List of Neighborhoods in Toronto
   This dataset provides name of neighborhoods, name of towns and postal code. Using this dataset, we can get list of all neighborhoods grouped by different name of towns.
2. Latitude and Longitude
   This can be found by using the following file: http://cocl.us/Geospatial_data .
   It would result in dataframe with three columns: Postal Code, Latitude, Longitude.
3. Foursquare API
   This is used to find venue data related to Pharmacy in Toronto. And it would also tells us which neighborhood needs to have a pharmacy opened. It would contain data of venues and venue categories for each neighborhood.

Data Extraction:

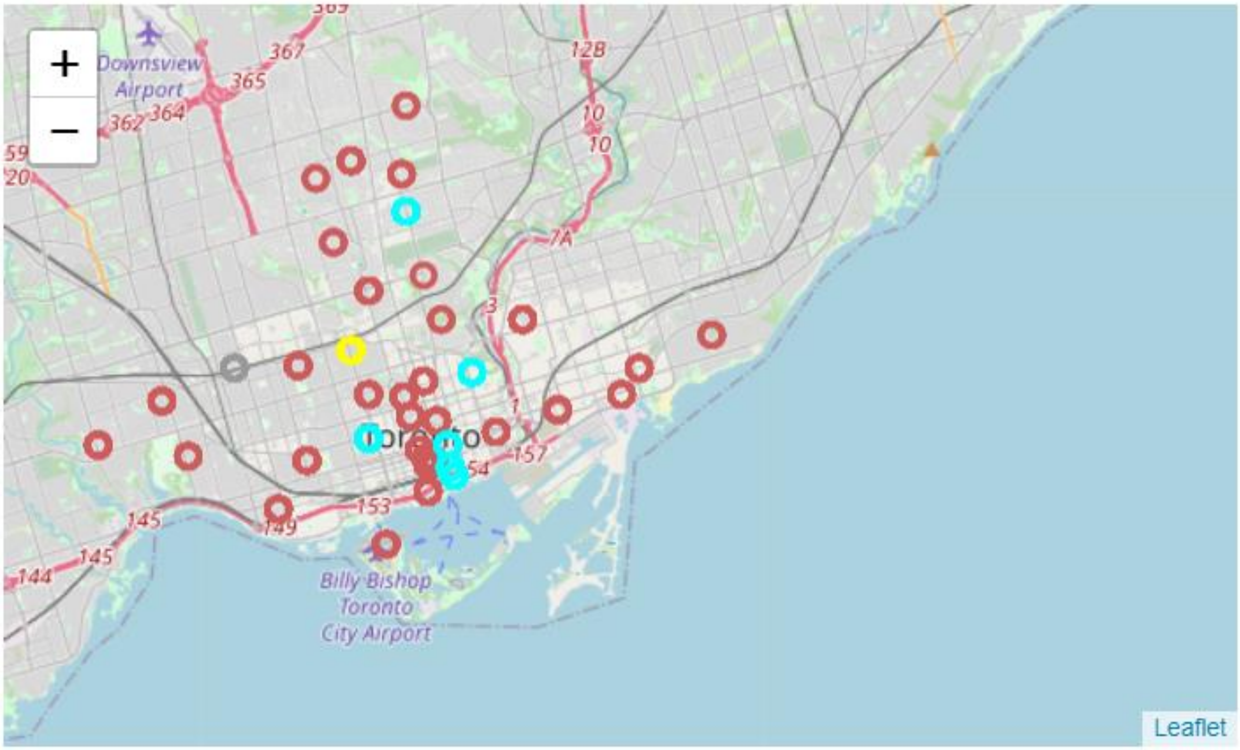
1. Using the wikipedia link which contain postal codes of entire Canada, we can scrape data of neighborhoods located in Toronto.
2. To obtain latitude and longitude for each neighborhood, geocoder package from geopy library in Python was used.
3. To extract data for pharmacy in each neighborhood, Foursquare API was great help.

Methodology:

- To read the data from URL, lxml library is used. The column Borough is of no use so decided to drop it.
- Checked if there is any row in neighborhood column is empty or not
- We need to join neighborhood that has same postal codes to avoid ambiguity
- Merge two data files one which contain postal code, latitude, longitude and another one is postal code, borough, neighborhood

- Save the merged data set into another file
- Read the new file and group dataframe using Borough column by count of neighborhood
- Search for borough containing Toronto which would give data of all different areas of Toronto
- Using folium, map Toronto neighborhoods
- Foursquare API can help us find pharmacy venue
- Calculate mean of all venues in Toronto and separate mean of venue pharmacy for each neighborhood
- Label each neighborhood with a cluster using the calculated mean
- We can map Toronto again using the clusters
- Using the clusters we can see which cluster has lowest number of pharmacy

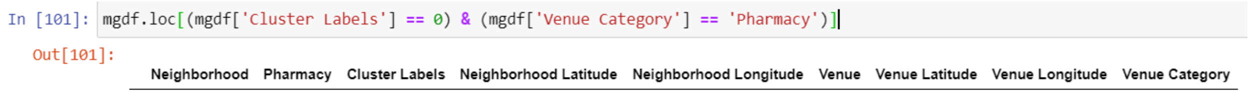
Results:



Discussion:

Seeing the tables below we can see that, it would be better to open a pharmacy in an area with cluster label 0, 1, and even 3 also. Cluster label 1 has only two pharmacies so it is safe to open a pharmacy in that particular neighborhood.

Cluster label 0:

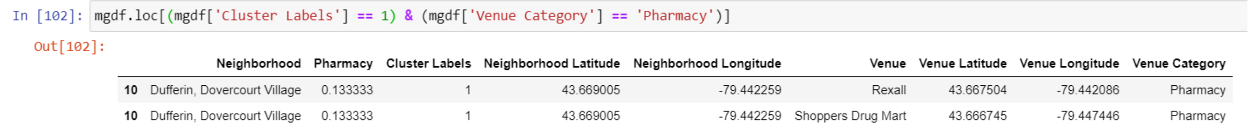


```
In [101]: mgdf.loc[(mgdf['Cluster Labels'] == 0) & (mgdf['Venue Category'] == 'Pharmacy')]
```

Out[101]:

Neighborhood	Pharmacy	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category

Cluster label 1:

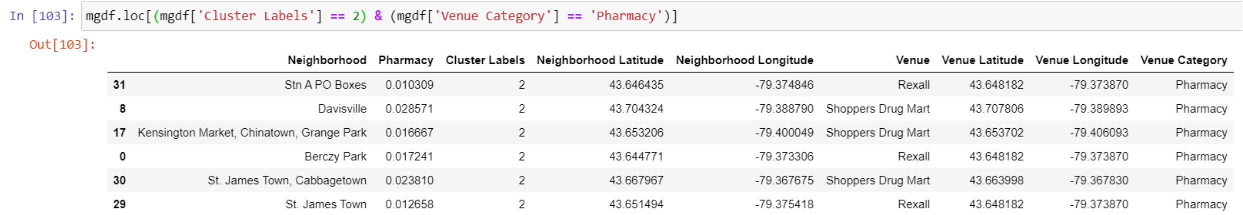


```
In [102]: mgdf.loc[(mgdf['Cluster Labels'] == 1) & (mgdf['Venue Category'] == 'Pharmacy')]
```

Out[102]:

	Neighborhood	Pharmacy	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
10	Dufferin, Dovercourt Village	0.133333	1	43.669005	-79.442259	Rexall	43.667504	-79.442086	Pharmacy
10	Dufferin, Dovercourt Village	0.133333	1	43.669005	-79.442259	Shoppers Drug Mart	43.666745	-79.447446	Pharmacy

Cluster label 2:

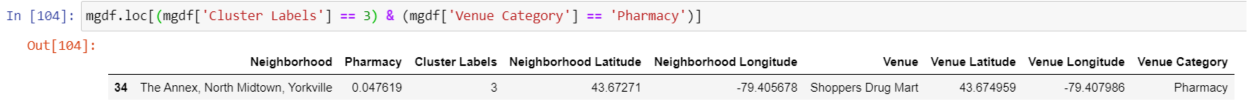


```
In [103]: mgdf.loc[(mgdf['Cluster Labels'] == 2) & (mgdf['Venue Category'] == 'Pharmacy')]
```

Out[103]:

	Neighborhood	Pharmacy	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
31	Stn A PO Boxes	0.010309	2	43.646435	-79.374846	Rexall	43.648182	-79.373870	Pharmacy
8	Davisville	0.028571	2	43.704324	-79.388790	Shoppers Drug Mart	43.707806	-79.389893	Pharmacy
17	Kensington Market, Chinatown, Grange Park	0.016667	2	43.653206	-79.400049	Shoppers Drug Mart	43.653702	-79.406093	Pharmacy
0	Berczy Park	0.017241	2	43.644771	-79.373306	Rexall	43.648182	-79.373870	Pharmacy
30	St. James Town, Cabbagetown	0.023810	2	43.667967	-79.367675	Shoppers Drug Mart	43.663998	-79.367830	Pharmacy
29	St. James Town	0.012658	2	43.651494	-79.375418	Rexall	43.648182	-79.373870	Pharmacy

Cluster label 3:



```
In [104]: mgdf.loc[(mgdf['Cluster Labels'] == 3) & (mgdf['Venue Category'] == 'Pharmacy')]
```

Out[104]:

	Neighborhood	Pharmacy	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
34	The Annex, North Midtown, Yorkville	0.047619	3	43.67271	-79.405678	Shoppers Drug Mart	43.674959	-79.407986	Pharmacy

Conclusion:

I have used K-means clustering algorithm to separate the neighborhoods in Toronto city into discrete clusters k=4 and also used 103 latitudes and longitudes. The coordinates had very much similar neighborhoods around. We have managed to figure out which neighborhood is suitable to open a pharmacy. By completing this capstone project, I have been able to see a realistic approach to address real life problem that has impact on personal as well as financial with the help of data science resources. The most interesting thing I learnt while working on the project is Folium. It is such an amazing technique integrating knowledge, enhancing interpretation of the given data and last but not the least, make decision related to the problem with confidence.