CAPSTONE PROJECT

The Battle of the Neighborhoods

Business Problem

Problem

In this project, we are going to recommend a location for a pharmacy to be started in a neighborhood.

Background

Toronto is the capital city of the Canadian province of Ontario. It is the fourth most populous city in North America. It is an international and multicultural city center with many businesses. This project will focus on areas in and near Toronto to identify the best choice locations for starting a new pharmacy.

Toronto's 2020 population is very diverse consisting of an international mix of people. The average life expectancy is about 83 years. The urban population has been on an increase(Reference). An aging population needs medical care and appropriate close-by pharmacies to get medicine soon. The pharmacy also typically helps with consultation and sometimes is even equipped with mini-clinics to offer advice to incoming customers with or without insurance at a nominal charge. Starting a pharmacy however can be tricky and should not be done in a place that is already overcrowded with these services. This project allows us to use four square api to get venues near a neighborhood location and apply machine learning techniques to analyse the data in order to recommend a location for starting a pharmacy. Age of people and income also matters as these are considered when pharmacy usage is involved.

Interest

The population that needs medical attention is growing and will constitute a need for more medical care and pharmacies. Also, the world is now experiencing a pandemic called Covid-19. Hence this problem is interesting and will provide a report to help find a good location to start a pharmacy.

Appropriate data gathering, cleaning, analysis using machine learning techniques learnt in the Coursera lessons and labs will be used to identify a solution to the problem.

Data

The project analysis can be applied to any city whose borough and neighborhood information is available, but we will focus on Toronto data that was available in the Coursera labs to analyze and report on a suitable location for a pharmacy setup.

Based on the definition of our problem, factors that will impact the decision are:

- Number of existing pharmacies in the neighborhood and nearby
- Number of existing medical centers in the neighborhood and nearby

- Toronto income and age levels reference study to understand the need for pharmacies. This is useful to understand if it is viable financially and demand wise to thrive well (this will be obtained from the net using different sources of information on the internet)

In this project, we will use the postal code and borough/neighborhood information pertaining to Toronto city. We will then clean the data and apply Foursquare API to get a list of venues related to medical data near the city. We will then use python libraries to create appropriate data frames that relate the neighborhoods to the frequency of nearby available medical centers including pharmacy. Using this data frame, we will then use the machine learning algorithm like k-means clustering to identify or predict which neighborhood may be better suited to start a pharmacy.

Data Acquisition

The data is obtained from the following sources:

- Toronto Neighborhoods and Boroughs
- Toronto Geospatial Data **We could use the Google Geocoding API but chose to use a set that is already available for the geospatial data from Coursera lab**
- **Toronto demographics information is obtained from these sources**
- **Toronto Venues of Interest: The Four-Square API will be used to fetch nearby medical center data including that for pharmacies. Limits are set to 100 with radius 500 due to limitations of free API availability.**

Data Description

Toronto Neighborhoods and Boroughs

The data fetched from this site will provide a list of postal codes along with the borough and neighborhood information for Toronto. This will be accessed using a http request and then the response will be parsed to obtain this list.

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills

Toronto Geospatial Data

The geo spatial data is useful for finding locations and their latitude and longitude that will provide input to Four-Square API to get a list of nearby venues.

Postal Code	Latitude	Longitude
M1B	43.8066863	-79.194353
M1C	43.7845351	-79.160497
M1E	43.7635726	-79.188712
M1G	43.7709921	-79.216917
M1H	43.773136	-79.239476
M1J	43.7447342	-79.239476

Toronto Demographics

This is just used to understand the demographics, age and health status in Toronto. It also provides an insight to the population density and average mortality rate. These help us understand how health and pharmacies are essential.

Toronto Venues of Interest

Four-Square API will return a list of venues that are of interest to the geo spatial locations obtained using the geospatial data for certain borough (Etobicoke in this project) and neighborhoods in Toronto. This JSON formatted result set will be processed to identify close-by medical centers and pharmacies to the location being considered, in this case a borough in Toronto. The JSON data is parsed and converted to a pandas dataframe as follows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Islington Avenue, Humber Valley Village	43.667856	-79.532242	Shoppers Drug Mart	43.663067	-79.531753	Pharmacy
1	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	Shoppers Drug Mart	43.641312	-79.576924	Pharmacy
2	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	Burnhamthorpe Health Centre	43.642328	-79.576959	Medical Center
3	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	Dr Henry Nirenberg Dental Office	43.641895	-79.578301	Dentist's Office
4	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	Medical Clinic - Family Practice	43.641797	-79.576441	Doctor's Office
5	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	Dr Tse	43.641522	-79.576825	Doctor's Office
6	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	John C Kuhlmann	43.641707	-79.574864	Doctor's Office

Data Cleaning

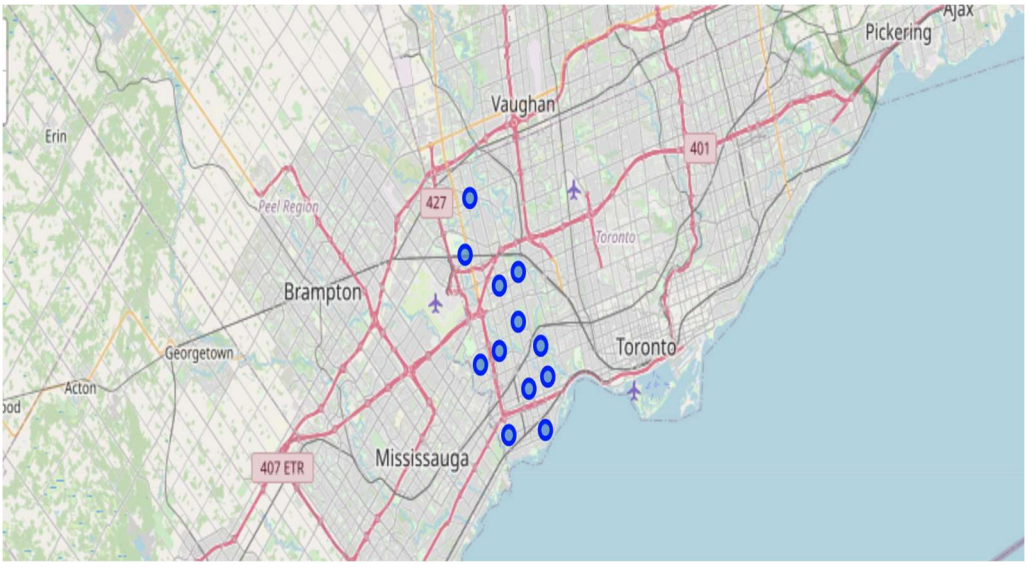
The data is cleaned and modified using the following approaches:

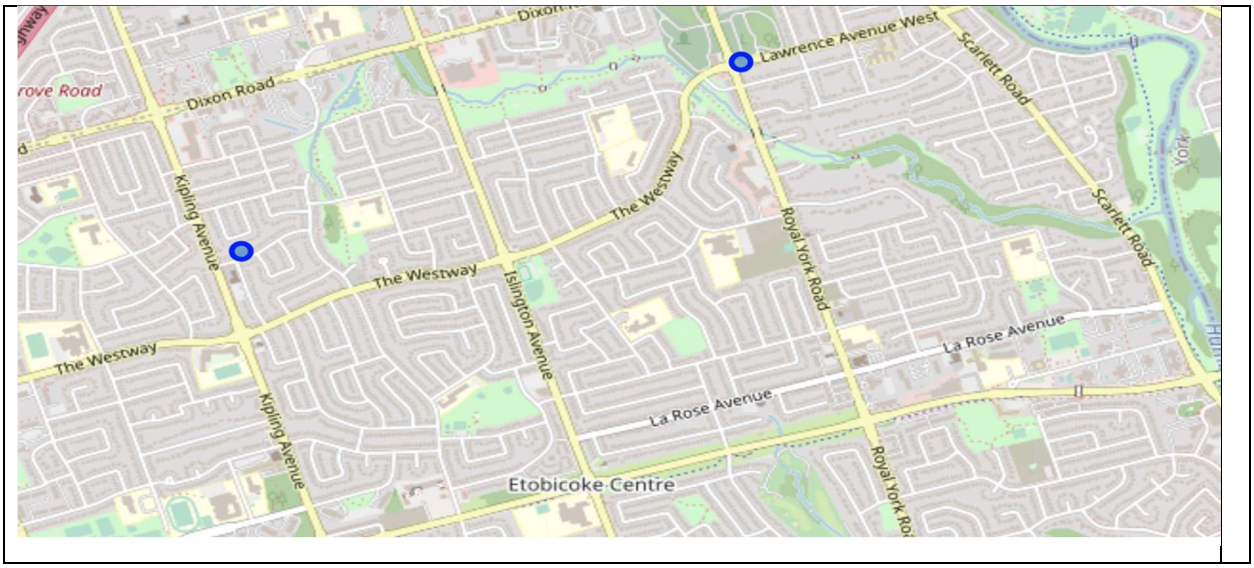
- If the borough is not assigned to any postal code
- If a neighborhood is not assigned, then use the borough name
- Group by postal code with neighborhoods concatenated by comma

Data Visualization

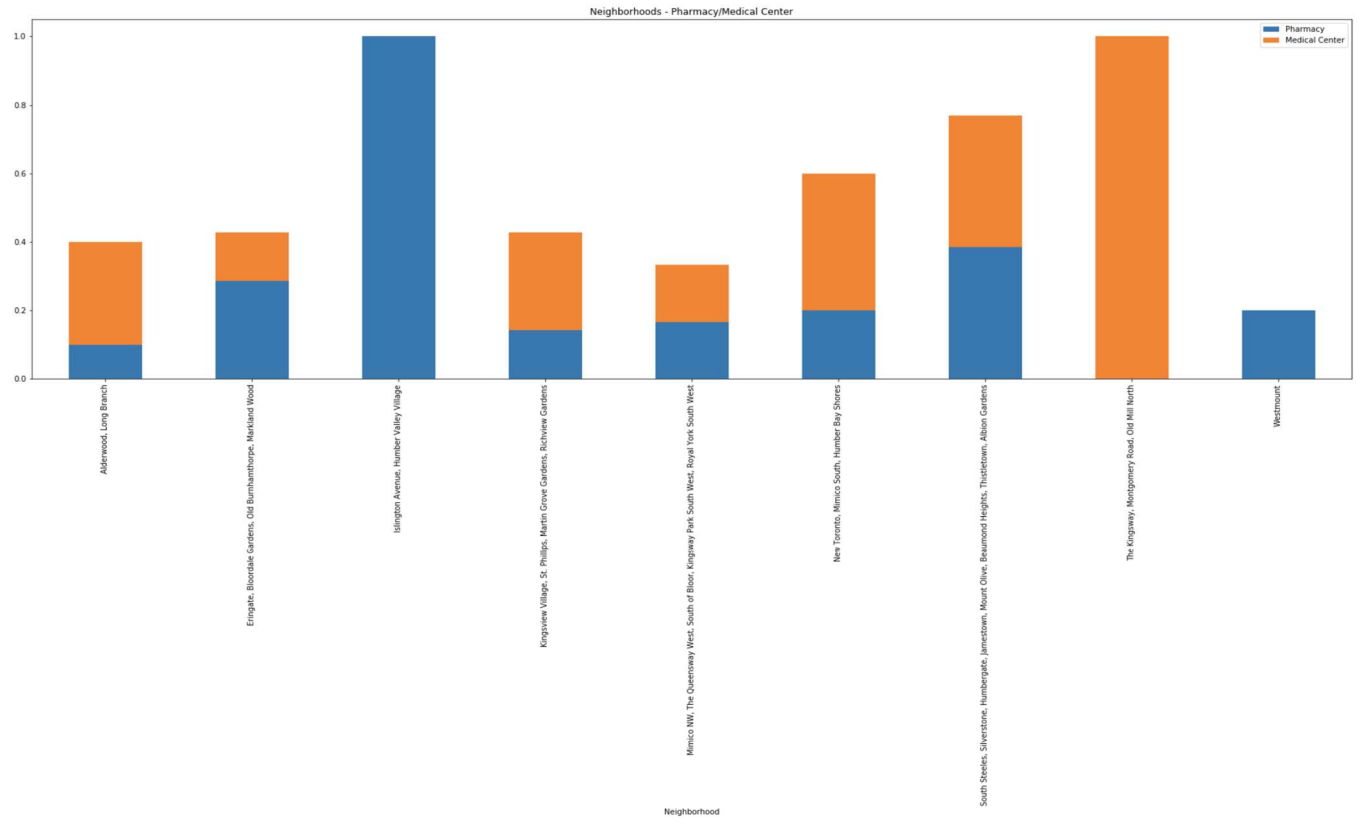
Folium map is used to visualize the neighborhoods and boroughs in the city using the Geo spatial data. Likewise, the venues found in the city using Four-Square API is also visualized using Folium library. Additionally, python bar charts using matplotlib library are used to look at which clusters have a greater concentration of the pharmacies or medical centers. We focus on medical centers and pharmacies, but this can be easily extended to other medical venues.

Folium Maps of Etobicoke - Neighborhoods





Medical Centers and Pharmacies - Neighborhood



Data Feature Selection

The pharmacy venue is chosen for the feature set to be analyzed and cluster the neighborhoods based on the frequency of occurrence of these kinds of venues.

Exploratory Data Analysis

Tools

Library or Tool	Capability Used
Pandas	Dataframe capability
NumPy	Data for arrays and vectors
BeautifulSoup	Parse HTML and get data
Four-Square API	To get venues nearby a location along with latitude and longitude
Requests	Use this for HTTP requests
Folium	Map library
MatPlotLib	Python plotting for bar graphs
Geopy	Suggested but not used to retrieve location data as this was obtained from Coursera suggested link
sklearn	Machine learning k-means

Modeling

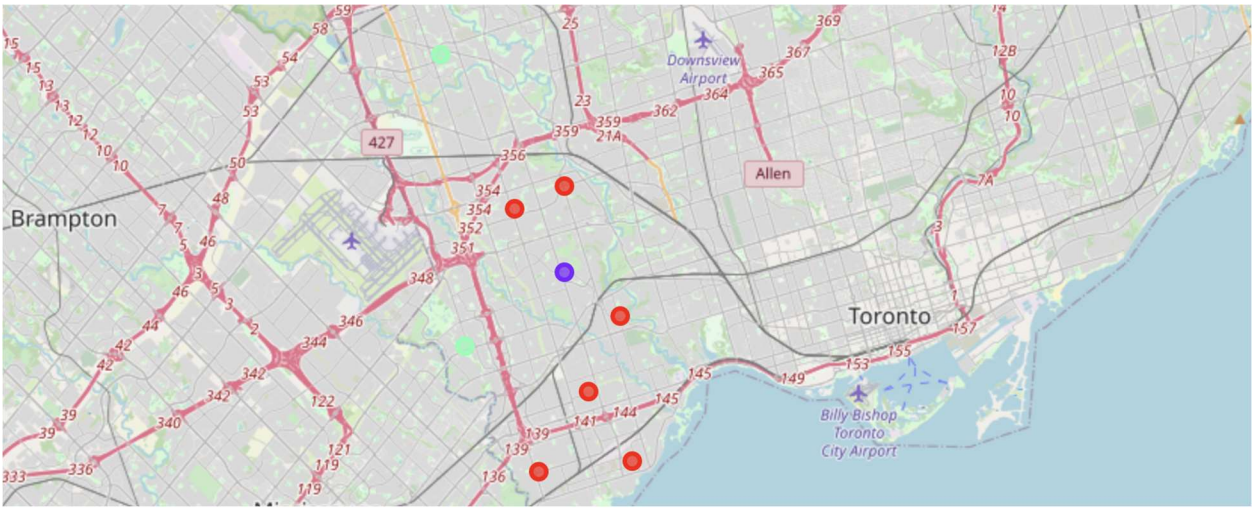
Classification and clustering that are used for pattern identification, have certain similarities. Classification uses already defined classes. Clustering groups items based on c ommon characteristics and also differentiates between objects in different groups. These groups are known as "clusters". To identify a good location for starting a pharmacy, we have to cluster neighborhoods. In this project, the solution requires some predictive analysis based on grouping of neighborhoods and identifying the best group for the problem. Clustering identifies similarities between objects, in this case neighborhoods and groups them based on. Characteristics in this case medical centers and pharmacies. This helps predict what would be the best choice to go forward for a location for a pharmacy startup.

The problem in this project falls more into the clustering category and hence K-means clustering was used, and different k factors were analyzed manually with Folium maps and bar charts. k=3 was found to be optimal on analysis.

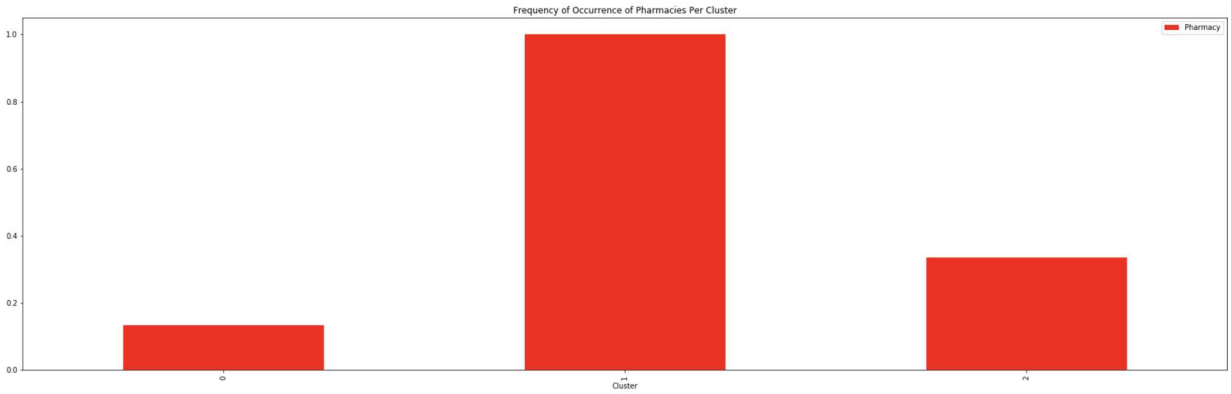
The venues data obtained using Four Square API was used to fetch medical center related venues nearby Etobicoke borough. The API results was then processed (JSON processing) to get the results of venues and their latitude and longitude. One hot encoding technique is used to help find the frequency of the occurrence of each type of venue. The frequency of occurrence of pharmacy, and medical centers (this can be extended to other types of medical centers, but these were considered for this project) was compared to identify which neighborhoods already had a good number of pharmacies and which lacked in this facility. The clustering (k-means) was used to group the data into clusters of neighborhoods for a particular borough. This is used to identify which cluster had a good set of neighborhoods that could potentially be more profitable to start a pharmacy. If the cluster had enough neighborhoods with a low or zero

number of pharmacies already available, then this was considered as a good choice for starting a pharmacy.

Clusters - Folium Map



Clusters - Graph



Clusters - K-Means Data

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Pharmacy
0	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	1.0	1.000000
2	M9C	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201	2.0	0.285714
3	M9P	Etobicoke	Westmount	43.696319	-79.532242	0.0	0.200000
4	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724	0.0	0.142857
5	M8V	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	0.0	0.200000
6	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437	2.0	0.384615
7	M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484	0.0	0.100000
9	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	0.0	0.000000
11	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	0.0	0.166667

Cluster 0 Neighborhoods - Medical center, Pharmacy frequency of occurence

	Neighborhood	Pharmacy_x	Medical Center	PostalCode	Borough	Latitude	Longitude	Cluster Labels	Pharmacy_y
0	Alderwood, Long Branch	0.100000	0.300000	M8W	Etobicoke	43.602414	-79.543484	0.0	0.100000
1	Kingsview Village, St. Phillips, Martin Grove ...	0.142857	0.285714	M9R	Etobicoke	43.688905	-79.554724	0.0	0.142857
2	Mimico NW, The Queensway West, South of Bloor,...	0.166667	0.166667	M8Z	Etobicoke	43.628841	-79.520999	0.0	0.166667
3	New Toronto, Mimico South, Humber Bay Shores	0.200000	0.400000	M8V	Etobicoke	43.605647	-79.501321	0.0	0.200000
4	The Kingsway, Montgomery Road, Old Mill North	0.000000	1.000000	M8X	Etobicoke	43.653654	-79.506944	0.0	0.000000
5	Westmount	0.200000	0.000000	M9P	Etobicoke	43.696319	-79.532242	0.0	0.200000

Discussion

The venues data obtained using Four Square API was not always focusing on getting data for medical centers or pharmacies. Hence the Four-Square API is modified to fetch medical center related venues nearby for a better analysis. The API results was then parsed to get the results of venues and their latitude and longitude. Also, Toronto neighborhoods and boroughs were also analyzed. For this scenario, the data (due to the fact that different boroughs were selected in one analysis set) did not cluster well with k-means. So it was decided to use just one borough for analysis at one time.

Results

The Toronto has 11 boroughs and 103 neighborhoods. A particular borough ‘Etobicoke' was selected for analysis. But this python notebook can just as easily be applied to any borough by changing the parameter to the functions. Boroughs with names ‘Toronto’ and ‘Scarborough’ were also analyzed and graphed. But for illustration purposes, and for the purpose of this project Etobicoke was picked. The focus was on medical center and pharmacy venues. Any medical center would house doctors or medical professionals that would prescribe to patients. So close locations for pharmacies would be ideal. This was the idea behind the project.

The cluster analysis was done using k-means algorithm. Cluster 1 had many pharmacies so was not considered. Cluster 2 had equal distributions for medical center and pharmacy so was not ideal for selection.

Cluster 0 had a few neighborhoods with medical centers but not many pharmacies. This would be an ideal cluster of neighborhoods to look at starting a. pharmacy. Specifically, with the analysis done, the neighborhoods such. as Kingsway, Montgomery, Old Mill North were found to be ideal for opening a pharmacy s there is a medical center close-by.

Conclusion

The location of Etobicoke was chosen to recommend a pharmacy starter location. Medical center data was collected using neighborhoods, venues near the location. K-means was used to cluster similar neighborhoods and a solution was proposed as described in Results. This was a great course with great hands-on that helped me master machine learning techniques.