Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods Final

Introduction

Toronto is the capital city of the Canadian province of Ontario. It covers an area of 630 square kilometers, with a recorded 2016 population of 2,731,571 and an annual growth rate of 0.86%., it is the most populous city in Canada and the fourth most populous city in North America.

Toronto is home to twenty public hospitals as well as a host to a wide variety of health-focused non-profit organizations that work to address specific illnesses.

Because of the growth rate and the good number of medical services, I am assuming that there might be great opportunity to open a pharmacy in Toronto area. For this reason, we are exploring Toronto neighborhoods using machine learning techniques to find the best location to open a pharmacy based on the nearby number of medical centers and population density of each neighborhood per square km.

Target audience:

Investors who are trying to open a pharmacy business in the best location based on scientific data science.

Data

Data of Toronto neighborhoods from the site services.arcgis.com which is shared by EsriCanadaEducation that include the arcgis shapefile of Toronto city including a dataset of city geometry, total area, total population, and other statistical information in regard to population per gender and age groups.

Data of pharmacy and medical centers venues that will be retrieved from foursquare site.

Methodology

Exploratory data analysis through:

Simple descriptive analysis using the describe() function to computes a summary of statistics pertaining to the DataFrame columns.

Simple horizontal bar chart to compare between the two venues categories counts against neighborhoods.

The folium library to show the Toronto neighborhoods and venues. Folium is a python library that can create interactive leaflet map using coordinate data. We used both the choropleth method to show the neighborhoods polygons classified by population density of each neighborhood overlayed by the venues markers classified into two categories and colors, the blue markers for the pharmacies and the red marker for the medical centers.

Methodology

Inferential statistical through correlation analysis to see if there are relationships between the variables. Correlation is a measure of a mutual relationship between two variables whether they are related or not. Pearson Correlation is one of the most used correlations during the data analysis process. it measures the linear relationship between variable continuous X and variable continuous Y and has a value between 1 and -1. In other words, the Pearson Correlation Coefficient measures the relationship between 2 variables via a line.

Predictive Modelling through clustering the neighborhoods (by utilizing Sci-kit learn package of python) based on the venues counts and population density using K-Means clustering. In order to determine the optimal number of clusters into which the data may be clustered, we used the elbow method which is one of the most popular methods to determine this optimal value of k. Dataset was first processed to remove unrelated variables which were then normalized used min-max scaling. Again we used folium library to draw the neighborhoods clusters.

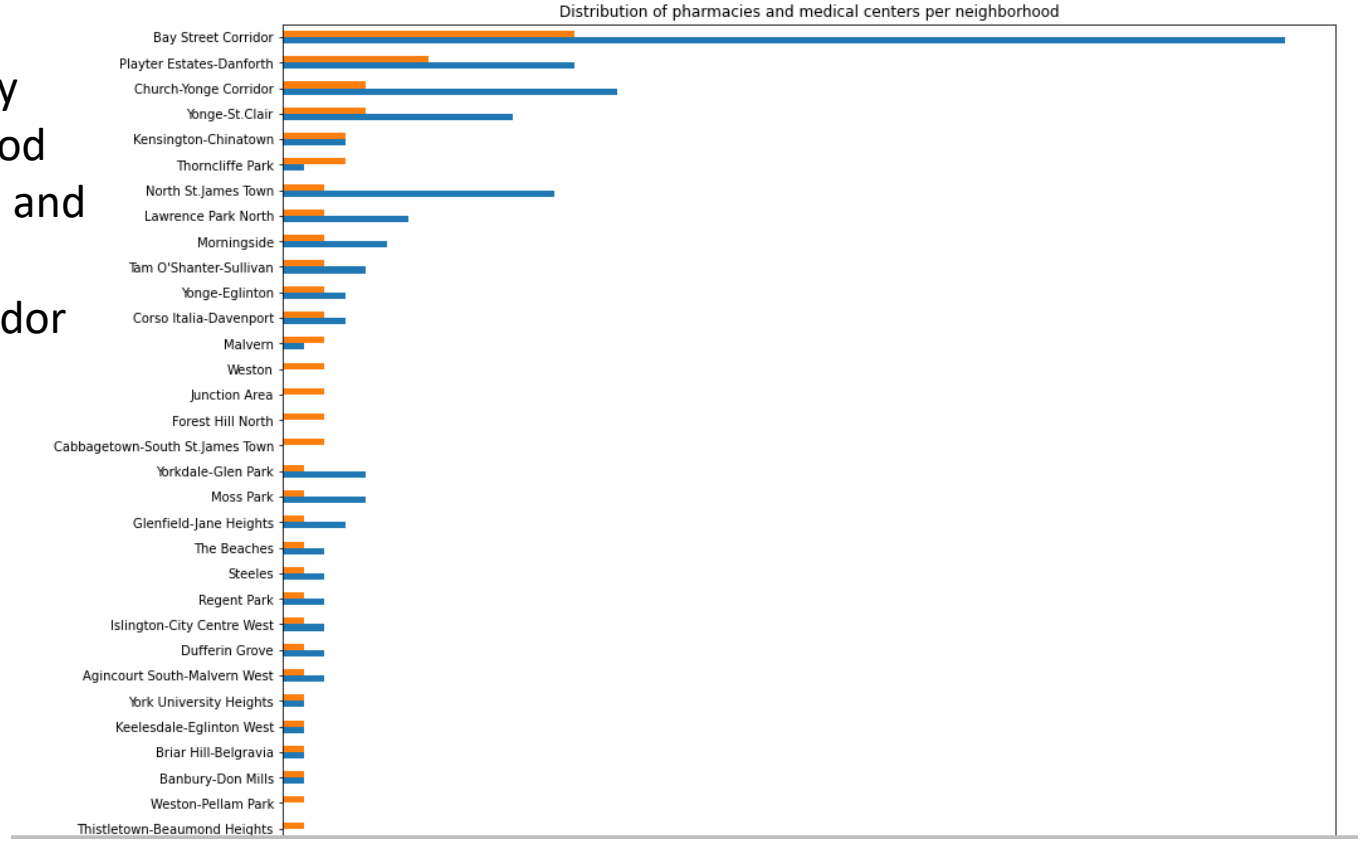
Results

From the original dataset we found that Toronto city has 140 neighborhoods with a total area of 633.29 sq km and a total population of 2,614,770. The average area of the neighborhoods is about 4.5 Square km with standard deviation of 4.6 sq km. The average population per neighborhood is 18676.9 person with standard deviation of 9099 while the average population density is about 5995 people per sq km with a standard deviation of 4673 person per sq km.

After merging and cleaning the data, we find that there are 205 medical centers and 85 pharmacies distributed in 67 neighborhoods. A total of 73 neighborhoods are left uncovered with medical centers and pharmacies. These uncovered neighborhoods were removed from our final data as there are no coordinates to draw on the map.

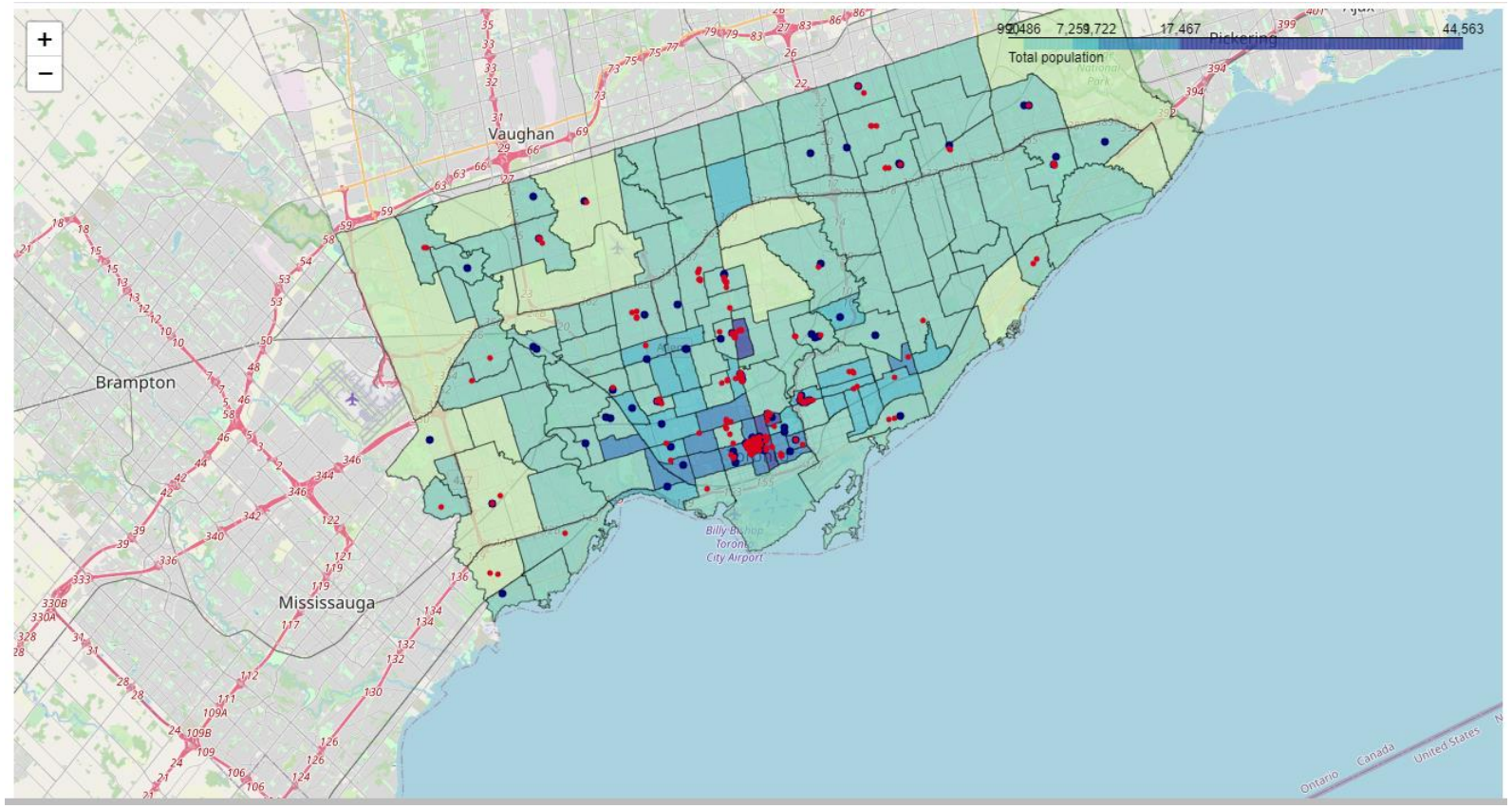
Results

The following bar graph shows that Bay street corridor is the most neighborhood that has the most medical centers (48) and pharmacies (14) followed by Playter Eastates-Danforth, Church-Yonge Corridor and Yonge-St. Clair.

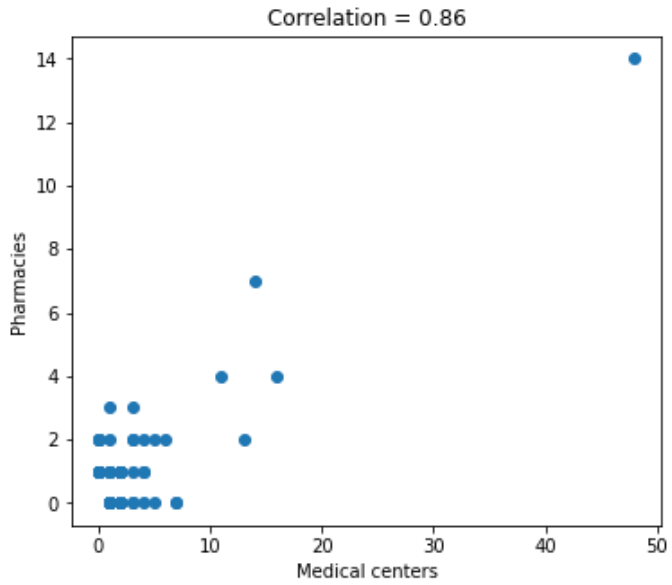


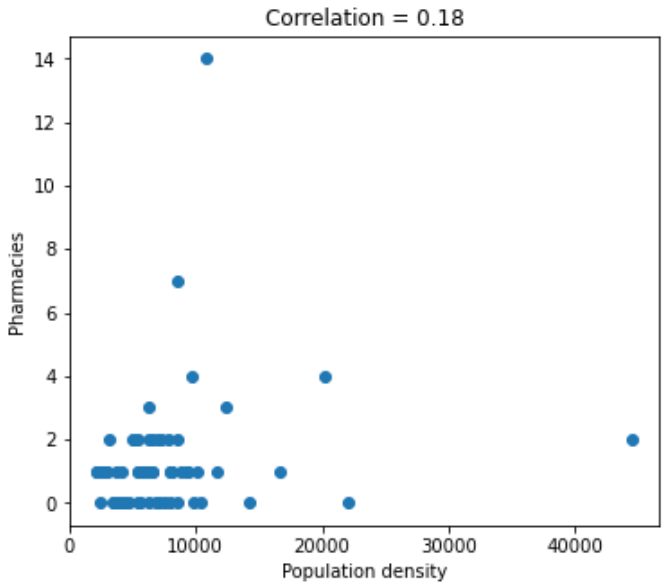
Results

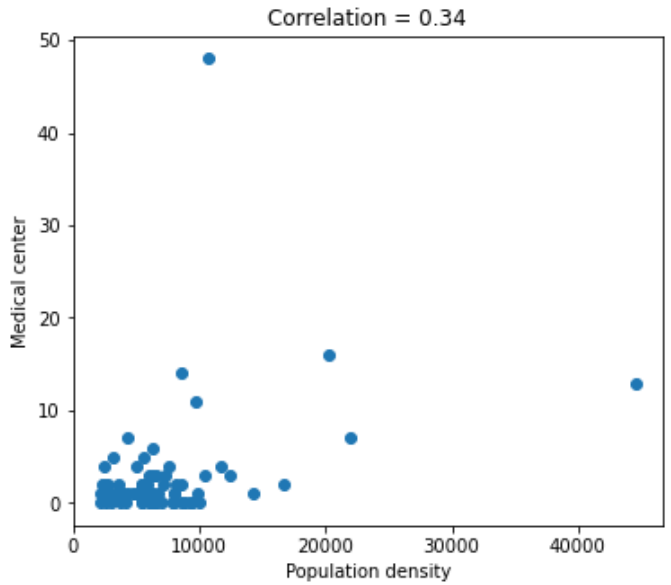
By using Toronto city as the base map, we draw Toronto city neighborhoods polygons based on the population density as well as draw the venues as markers.
Red circles denotes the medical centers venues while the blue circles denotes the pharmacies.



Results





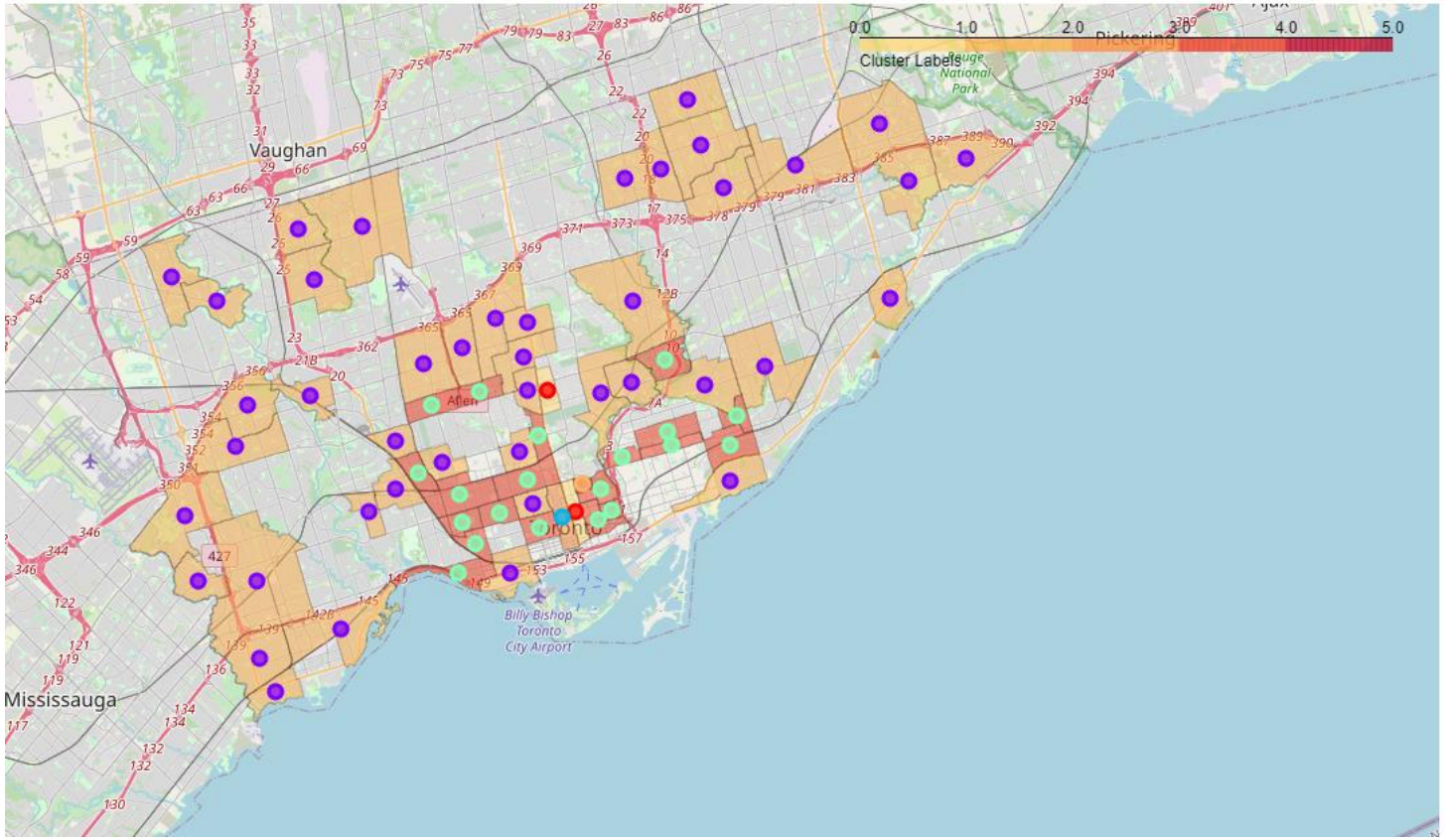


We found that there is a strong relationship (correlation = 0.86) between the count of pharmacies and the count of medical centers within a neighborhood, no relationship between population density and pharmacies (correlation = 0.18) while there is weak relationship between population density and medical centers (correlation = 0.34).

Results - Clustering of neighborhoods

There are 5 clusters

1. cluster 1 (red cycle) with 2 neighborhoods.
2. cluster 2 (magenta circle) contain 43 neighborhoods.
3. cluster 3 (blue circle) with 1 neighborhood.
4. cluster 4 (cyan circle) with 20 neighborhoods.
5. cluster 5 (brown circle) with 1 neighborhood.



Results - Clustering of neighborhoods

With the help of kmeans clustering, we identified three main clusters that share a high population density with more medical centers compared to pharmacies

cluster 1 that contain Church-Yonge Corridor which has a high population density (20246.43) with 16 medical centers and only 4 pharmacies, and Mount Pleasant West which has a high population density (21992.31) with 7 medical centers and no pharmacies.

Cluster 1 (red circle)

```
toronto_neigh_venues.loc[toronto_neigh_venues['Cluster Labels'] == 0, toronto_neigh_venues.columns[[1] + list(range(5, toronto_neigh_venues.shape[1]))]]
```

	Neighborhood	lon	lat	pop_density	Medical center	Pharmacy	pharmacy_diff
22	Church-Yonge Corridor	-79.379018	43.659650	20246.43	16.0	4.0	-12.0
101	Mount Pleasant West	-79.393360	43.704435	21992.31	7.0	0.0	-7.0

Results - Clustering of neighborhoods

2. cluster 3 that contain Bay Street Corridor has a high population density (10747.22) with the highest number of medical centers (48) compared to 14 only pharmacies.

Cluster 3 (blue circle)

```
toronto_neigh_venues.loc[toronto_neigh_venues['Cluster Labels'] == 2, toronto_neigh_venues.columns[[1] + list(range(5, toronto_neigh_venues.shape[1]))]]
```

	Neighborhood	lon	lat	pop_density	Medical center	Pharmacy	pharmacy_diff
93	Bay Street Corridor	-79.385722	43.657512	10747.22	48.0	14.0	-34.0

Results - Clustering of neighborhoods

3. cluster 5 that contain North St.James Town which has the highest population density (44562.5) with 13.0 medical centers and only 2 pharmacies.

Cluster 5 (Brown circle)

```
toronto_neigh_venues.loc[toronto_neigh_venues['Cluster Labels'] == 4, toronto_neigh_venues.columns[[1] + list(range(5, toronto_neigh_venues.shape[1]))]]
```

	Neighborhood	lon	lat	pop_density	Medical center	Pharmacy	pharmacy_diff
72	North St.James Town	-79.375247	43.669623	44562.5	13.0	2.0	-11.0

Discussion

We started our dataset with 140 neighborhoods in Toronto city but unfortunately 73 neighborhoods did not include medical centers nor pharmacies. These neighborhoods were discarded from our final dataset as we are depending on the difference between the count of medical centers and pharmacies, the clustering will be distorded as the difference of zero will be equal in those neighborhoods that have equal number of medical centers and pharmacies and in those neighborhoods that don't have any.

Exploratory analysis showed that Bay street corridor, Playter Eastates-Danforth, Church-Yonge Corridor and Yonge-St. Clair are the neighborhood that have the most medical centers and pharmacies.

Discussion

After studying correlation between the different variable especially the pharmacies count, medical centers count and population density, we found that there is a good correlation between the pharmacies count and the medical centers count and between the medical centers count and population density. This relationship allowed us to include these variables in the kmeans clustering.

According to the kmeans clustering, we were able to identify three clusters, these clusters share a high population density with more medical centers compared to pharmacies. These clusters include (cluster 1, cluster 3 and cluster 5)

These clusters contain four neighborhoods, namely, Church-Yonge Corridor, Mount Pleasant West, Bay Street Corridor and North St.James.

Recommendations

Although there might be good opportunity in opening new pharmacies in clusters 0, 2 and 4 but I recommend that there is a better opportunity to open a new pharmacy in clusters 1, 3 and 5 that contain the following neighborhoods (Church-Yonge Corridor, Mount Pleasant West, Bay Street Corridor and North St.James) since there is high need of pharmacies if compared to the medical centers count and population density in these neighborhoods as well as the good location according to folium maps in the center of Toronto.

Limitations

There might be some opportunities in the neighborhoods that has been discarded because of the lack of pharmacies or medical centers in them. Other data that include the address of visitors to all medical centers might help in determining the needs of pharmacy opening in areas according to the addresses.

The lack of extra information about the number of visitors of medical centers may help limit further analysis and recommendations.

Conclusion

In this project, we proceeded to identify the business problem, specify the required data, extracted and prepared the data, performed machine learning using k-means clustering to provide recommendations to the public.

Finally, we have executed an end-to-end data science project using common python libraries to manipulate data sets, Foursquare API to explore the neighborhoods of Toronto, and Folium leaflet map to cluster and segment neighborhoods to identify the best location in Toronto city to open a new pharmacy business.

Thank you