Peer-graded Assignment: Capstone Project- The Battle of Neighborhoods (Week 1)

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

For this project, we will use data mainly from two sources:

1. 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.

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

Data acquisition, processing and cleaning

A. Toronto data neighborhoods and statistics

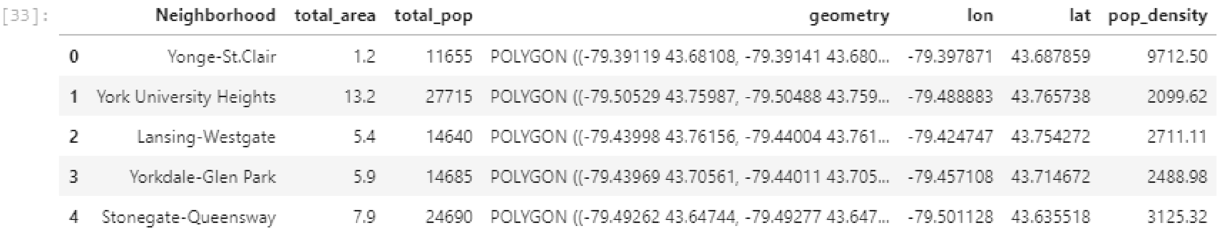
The first dataset will be retrieved in a 3-steps procedure that will include:

1. downloading the data in a zipped file from the site opendata.arcgis.com
2. unarchive the zip file to lead to a set of related ESRI shapefiles (these are geospatial vector data format that can be read by geographic information system (GIS) software).
3. read the shapefile "Toronto_Neighbourhoods.shp"
4. From the geometry field we will create two extra fields (latitude and longitude)

We can see that the data contains 140 neighborhoods and 42 data fields and will retrieve the first 5 rows to explore them.

We will clean the data to keep only fields that are related to our project which will include only Neighborhood, Total Population, Total Area, geometry, latitude, and longitude and rename these fields to give more sense names.

We will also create extra field called "pop_density" that will calculate population density per neighborhood based on neighborhood total area and total population.



[33]:

	Neighborhood	total_area	total_pop	geometry	lon	lat	pop_density
0	Yonge-St.Clair	1.2	11655	POLYGON ((-79.39119 43.68108, -79.39141 43.680...	-79.397871	43.687859	9712.50
1	York University Heights	13.2	27715	POLYGON ((-79.50529 43.75987, -79.50488 43.759...	-79.488883	43.765738	2099.62
2	Lansing-Westgate	5.4	14640	POLYGON ((-79.43998 43.76156, -79.44004 43.761...	-79.424747	43.754272	2711.11
3	Yorkdale-Glen Park	5.9	14685	POLYGON ((-79.43969 43.70561, -79.44011 43.705...	-79.457108	43.714672	2488.98
4	Stonegate-Queensway	7.9	24690	POLYGON ((-79.49262 43.64744, -79.49277 43.647...	-79.501128	43.635518	3125.32

B. Toronto pharmacy and medical centers venues

For obtaining the venue details of each neighborhood, we have used the Foursquare location API, restricting our data to two main categories, **pharmacy** venues (4bf58dd8d48988d10f951735) and **medical centers** venues (4bf58dd8d48988d104941735). The following details have been retrieved:

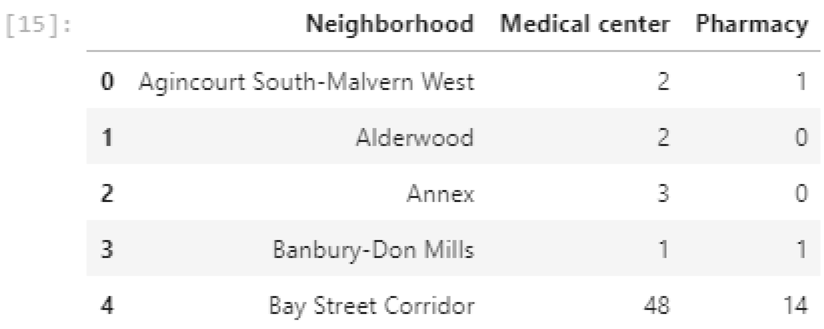
- venue name
- category id
- category name
- venue latitude and longitude.

We will use the foursquare API function that will search for the pharmacies and medical centers venues with a limit of 100 venues within 450m radius of our neighborhoods.

By exploring the data retrieved We can see that some venues are wrongly flagged in foursquare.

Our next step is to remove unrelated venues and rename all medical centers back to their parent category (medical centers) so that we will end up in two categories (pharmacies and medical centers)

For the sake of our analysis, we will depend on the count of pharmacies and medical centers in each neighborhood. Therefore, we convert our foursquare datasets into counts of each category in each neighborhood.



[15]:

	Neighborhood	Medical center	Pharmacy
0	Agincourt South-Malvern West	2	1
1	Alderwood	2	0
2	Annex	3	0
3	Banbury-Don Mills	1	1
4	Bay Street Corridor	48	14

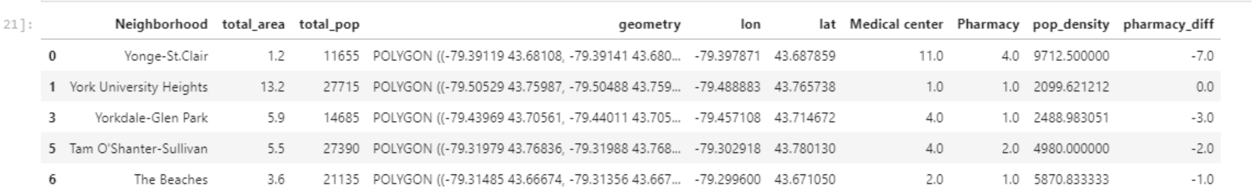
Data merging

The final steps in data preparation include:

1. Merge data from both data sources into one final dataset

2. Remove any neighborhood that does not contain any pharmacy or medical centers.

3. Create extra field called pharmacy difference (the difference between the count of pharmacies and medical centers)



21]:

	Neighborhood	total_area	total_pop	geometry	lon	lat	Medical center	Pharmacy	pop_density	pharmacy_diff
0	Yonge-St.Clair	1.2	11655	POLYGON ((-79.39119 43.68108, -79.39141 43.680...	-79.397871	43.687859	11.0	4.0	9712.500000	-7.0
1	York University Heights	13.2	27715	POLYGON ((-79.50529 43.75987, -79.50488 43.759...	-79.488883	43.765738	1.0	1.0	2099.621212	0.0
3	Yorkdale-Glen Park	5.9	14685	POLYGON ((-79.43969 43.70561, -79.44011 43.705...	-79.457108	43.714672	4.0	1.0	2488.983051	-3.0
5	Tam O'Shanter-Sullivan	5.5	27390	POLYGON ((-79.31979 43.76836, -79.31988 43.768...	-79.302918	43.780130	4.0	2.0	4980.000000	-2.0
6	The Beaches	3.6	21135	POLYGON ((-79.31485 43.66674, -79.31356 43.667...	-79.299600	43.671050	2.0	1.0	5870.833333	-1.0

Methodology

As the dataset has been retrieved from the different sources, cleaned and finalized, data analysis started with the following methods.

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.

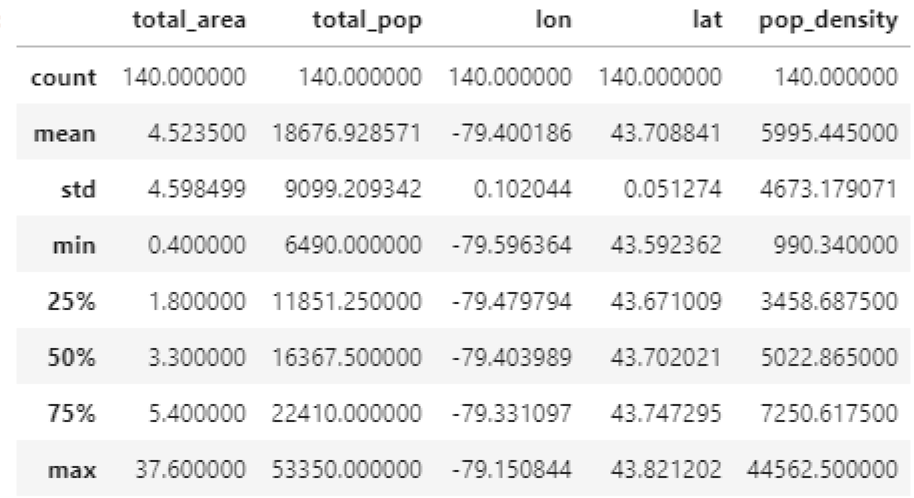
The folium library to show the Toronto neighborhoods and venues. Folium is a python library that can create interactive leaflet map using coordinate data. We used both the choropleth method to show the neighborhoods polygons classified by population density of each neighborhood overlayed by the venues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

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.

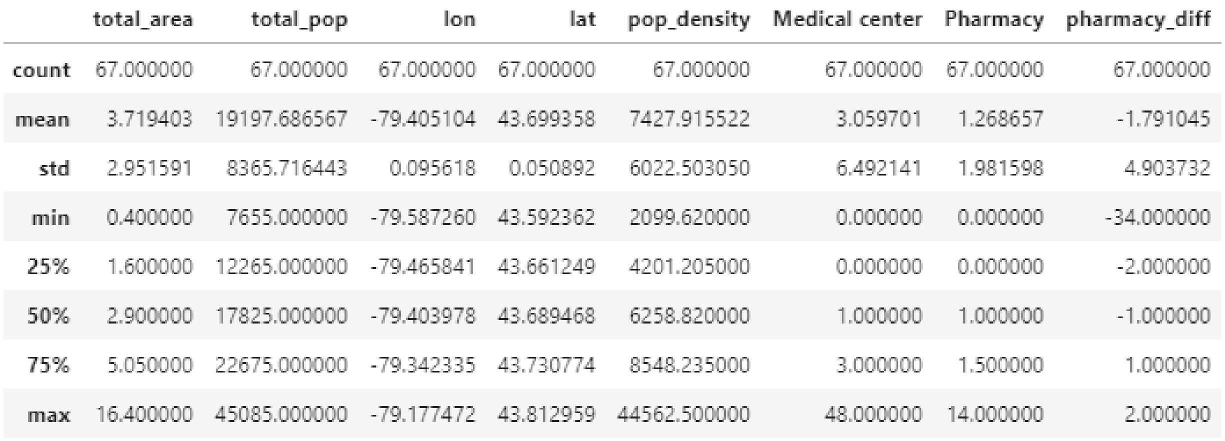
Results

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.



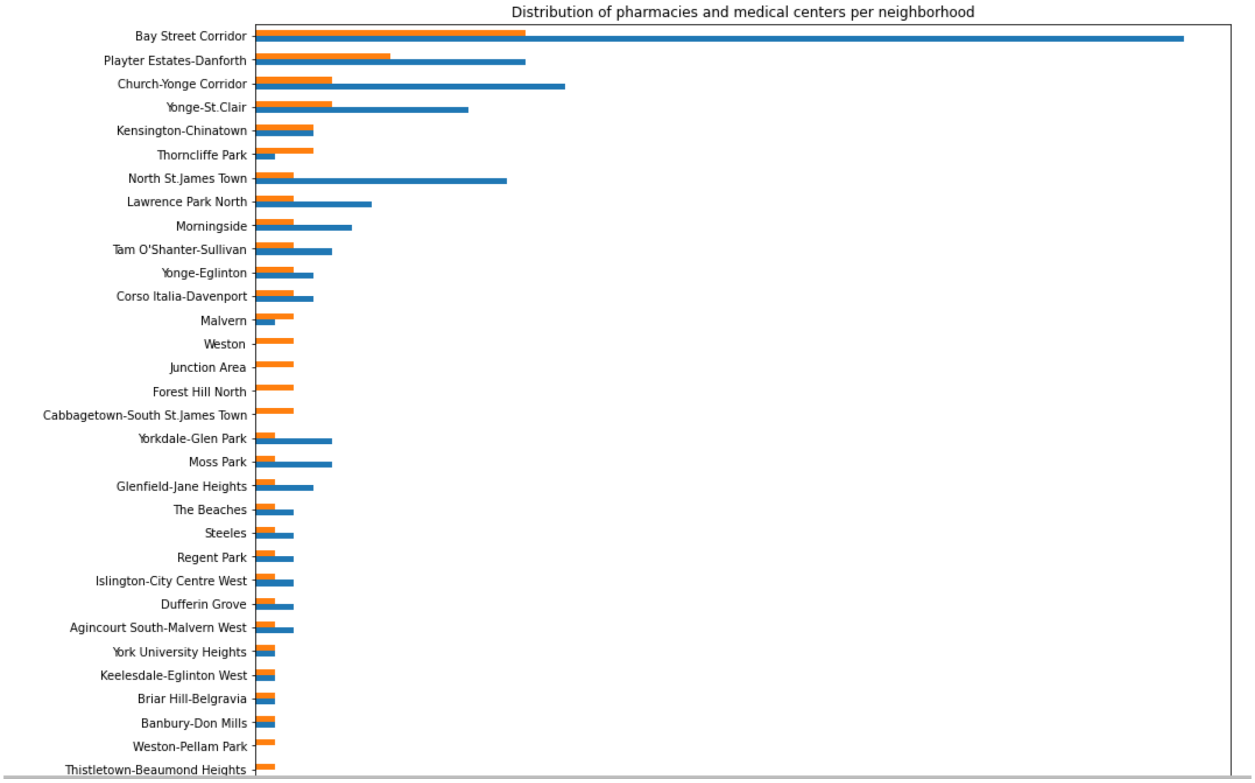
	total_area	total_pop	lon	lat	pop_density
count	140.000000	140.000000	140.000000	140.000000	140.000000
mean	4.523500	18676.928571	-79.400186	43.708841	5995.445000
std	4.598499	9099.209342	0.102044	0.051274	4673.179071
min	0.400000	6490.000000	-79.596364	43.592362	990.340000
25%	1.800000	11851.250000	-79.479794	43.671009	3458.687500
50%	3.300000	16367.500000	-79.403989	43.702021	5022.865000
75%	5.400000	22410.000000	-79.331097	43.747295	7250.617500
max	37.600000	53350.000000	-79.150844	43.821202	44562.500000

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.



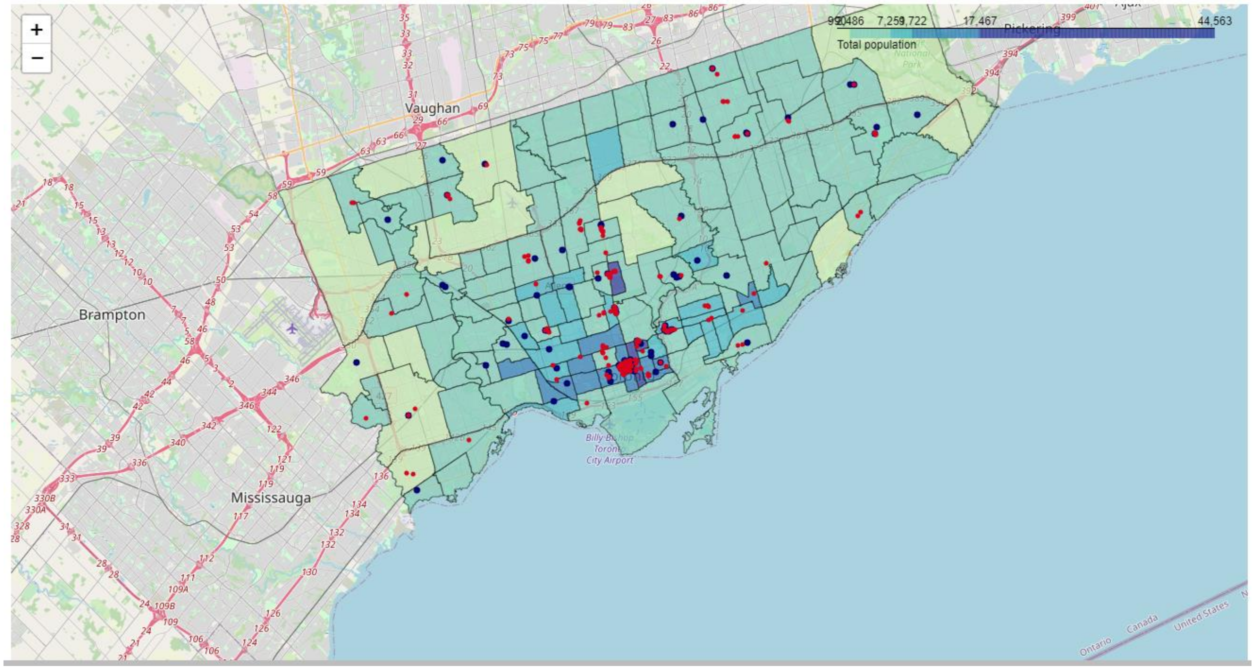
	total_area	total_pop	lon	lat	pop_density	Medical center	Pharmacy	pharmacy_diff
count	67.000000	67.000000	67.000000	67.000000	67.000000	67.000000	67.000000	67.000000
mean	3.719403	19197.686567	-79.405104	43.699358	7427.915522	3.059701	1.268657	-1.791045
std	2.951591	8365.716443	0.095618	0.050892	6022.503050	6.492141	1.981598	4.903732
min	0.400000	7655.000000	-79.587260	43.592362	2099.620000	0.000000	0.000000	-34.000000
25%	1.600000	12265.000000	-79.465841	43.661249	4201.205000	0.000000	0.000000	-2.000000
50%	2.900000	17825.000000	-79.403978	43.689468	6258.820000	1.000000	1.000000	-1.000000
75%	5.050000	22675.000000	-79.342335	43.730774	8548.235000	3.000000	1.500000	1.000000
max	16.400000	45085.000000	-79.177472	43.812959	44562.500000	48.000000	14.000000	2.000000

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.

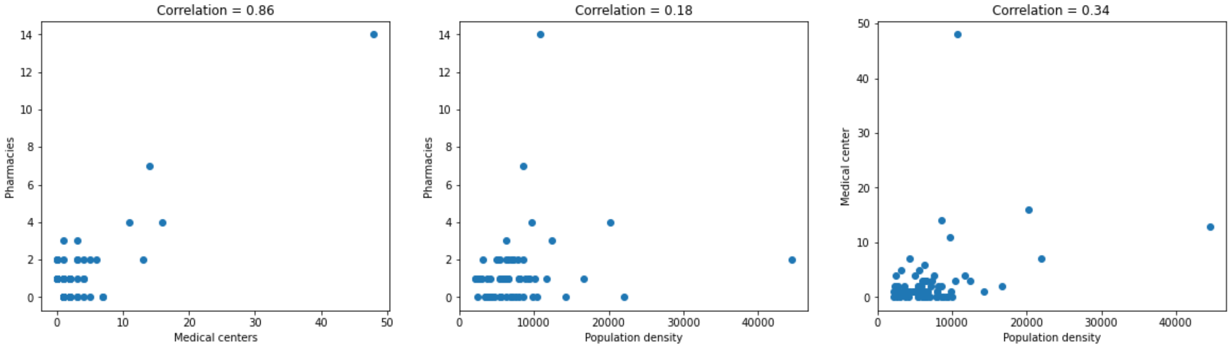


Mapping the datasets

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.



Relationship between variables

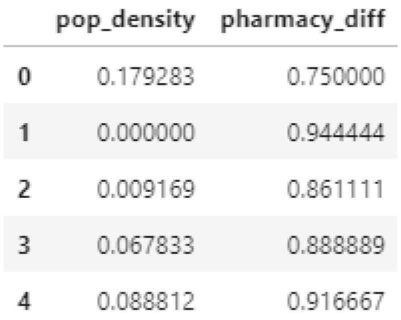


We found that there is a strong relationship (correlation = 0.86) between the count of pharmacies and the count of medical centers within a neighborhood, no relationship between population density and pharmacies (correlation = 0.18) while there is weak relationship between population density and medical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

Clustering of neighborhoods

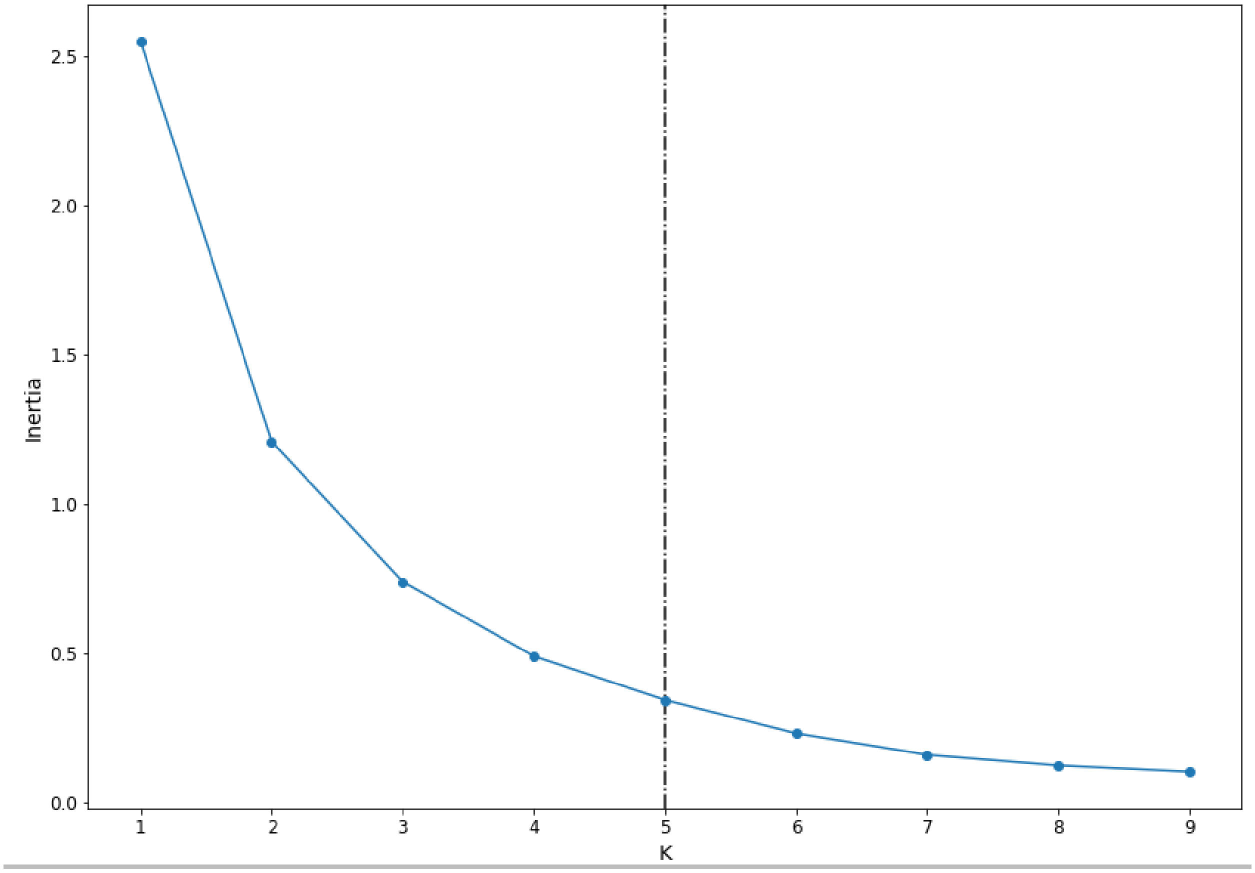
To start clustering neighborhoods, we need first to keep only columns we are interested in which are population density and the difference between medical centers count per neighborhood and pharmacies.

Since the range of values among the categories is significant, our next step is to normalize our values to minimize the distortion differences in the ranges of values and give them all equal importance. Scaling is also important from a clustering perspective as the distance between points affects the way clusters are formed.

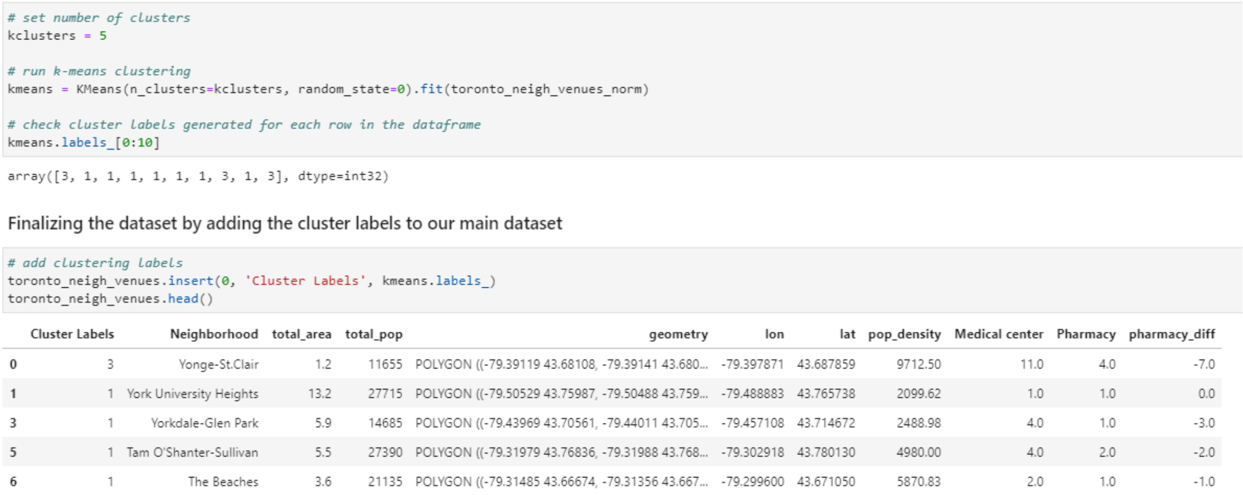


	pop_density	pharmacy_diff
0	0.179283	0.750000
1	0.000000	0.944444
2	0.009169	0.861111
3	0.067833	0.888889
4	0.088812	0.916667

Elbow method to determine k value



Clustering the neighborhoods



```
# set number of clusters
kclusters = 5

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_neigh_venues_norm)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

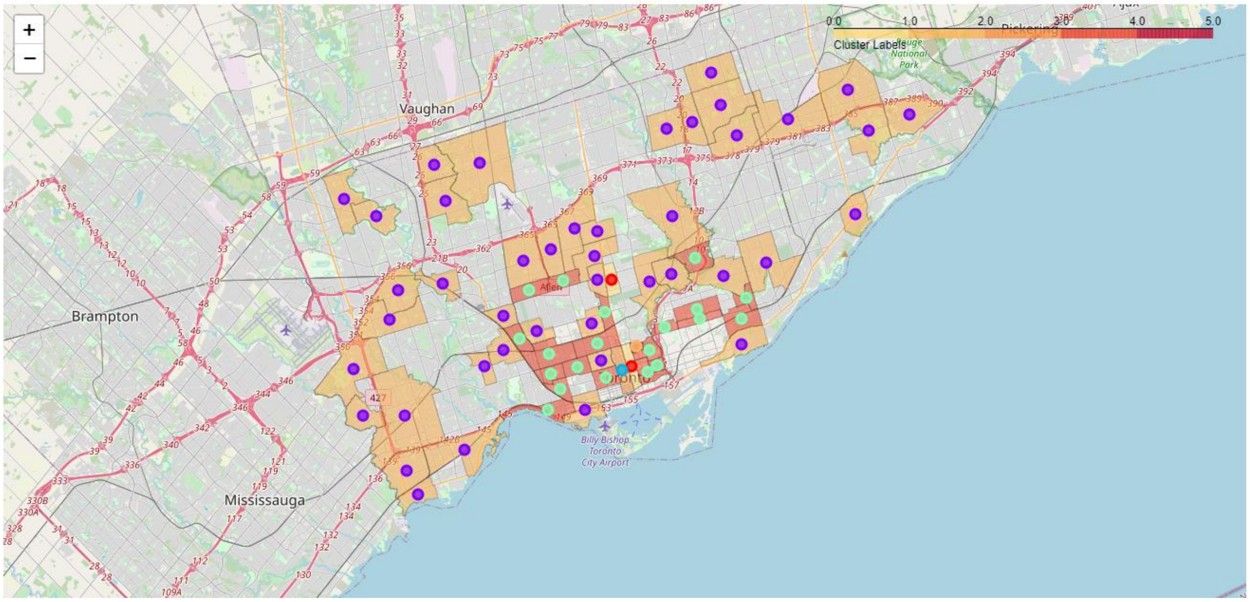
```
array([3, 1, 1, 1, 1, 1, 1, 3, 1, 3], dtype=int32)
```

Finalizing the dataset by adding the cluster labels to our main dataset

```
# add clustering labels
toronto_neigh_venues.insert(0, 'Cluster Labels', kmeans.labels_)
toronto_neigh_venues.head()
```

	Cluster Labels	Neighborhood	total_area	total_pop	geometry	lon	lat	pop_density	Medical center	Pharmacy	pharmacy_diff
0	3	Yonge-St.Clair	1.2	11655	POLYGON ((-79.39119 43.68108, -79.39141 43.680...	-79.397871	43.687859	9712.50	11.0	4.0	-7.0
1	1	York University Heights	13.2	27715	POLYGON ((-79.50529 43.75987, -79.50488 43.759...	-79.488883	43.765738	2099.62	1.0	1.0	0.0
3	1	Yorkdale-Glen Park	5.9	14685	POLYGON ((-79.43969 43.70561, -79.44011 43.705...	-79.457108	43.714672	2488.98	4.0	1.0	-3.0
5	1	Tam O'Shanter-Sullivan	5.5	27390	POLYGON ((-79.31979 43.76836, -79.31988 43.768...	-79.302918	43.780130	4980.00	4.0	2.0	-2.0
6	1	The Beaches	3.6	21135	POLYGON ((-79.31485 43.66674, -79.31356 43.667...	-79.299600	43.671050	5870.83	2.0	1.0	-1.0

Visualizing and examining clusters



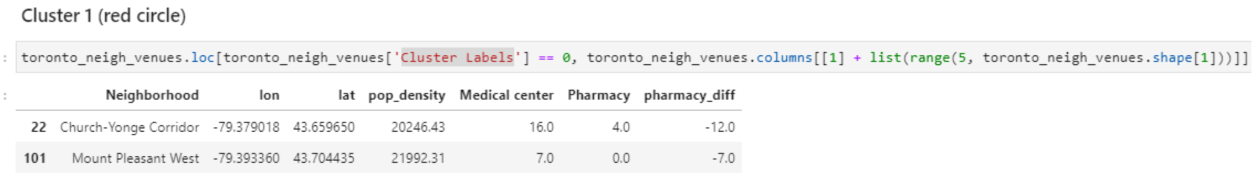
There are 5 clusters

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.

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

1. 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.

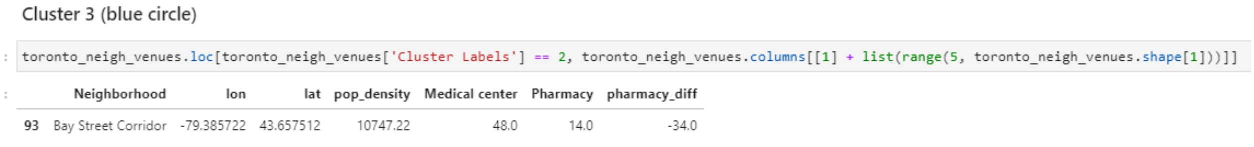
2. cluster 3 that contain Bay Street Corridor has a high population density (10747.22) with the highest number of medical centers (48) compared to 14 only pharmacies.
3. cluster 5 that contain North St.James Town which has the highest population density (44562.5) with 13.0 medical centers and only 2 pharmacies.



Cluster 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]))]]
```

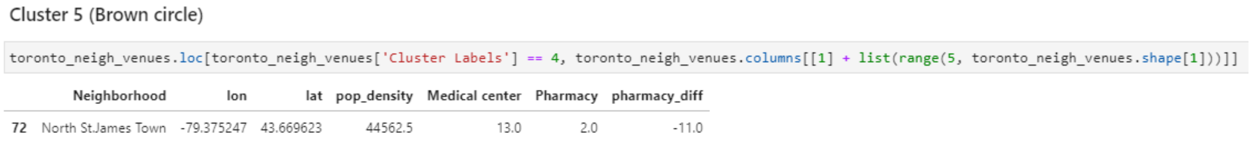
	Neighborhood	lon	lat	pop_density	Medical center	Pharmacy	pharmacy_diff
22	Church-Yonge Corridor	-79.379018	43.659650	20246.43	16.0	4.0	-12.0
101	Mount Pleasant West	-79.393360	43.704435	21992.31	7.0	0.0	-7.0



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



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.

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.