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

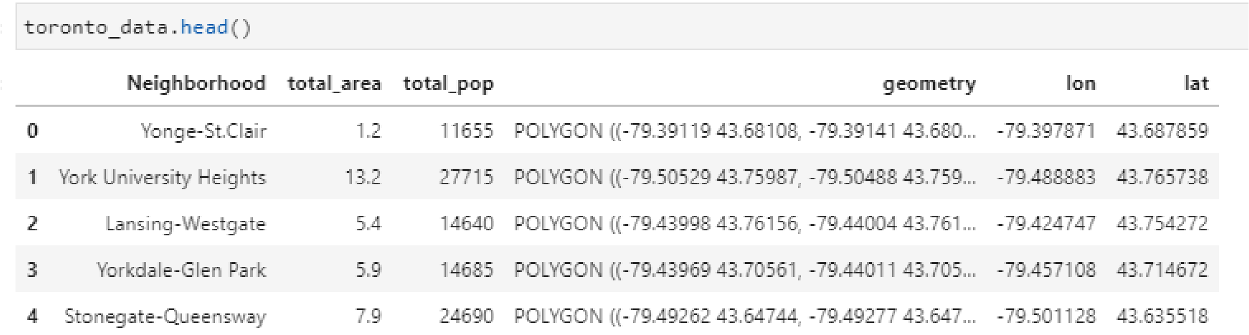
A. Toronto data neighborhoods and statistics

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.



```
toronto_data.head()
```

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

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.

Retrieving foursquare data

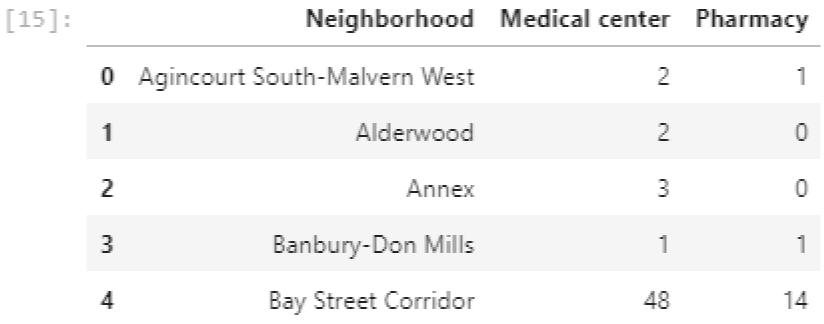
We will use the foursquare API function that will search for the pharmacies and medical centers venues with a limit of 100 venues within 450m radius of our neighborhoods.

We can see that some venues are wrongly flagged in foursquare which we will.

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

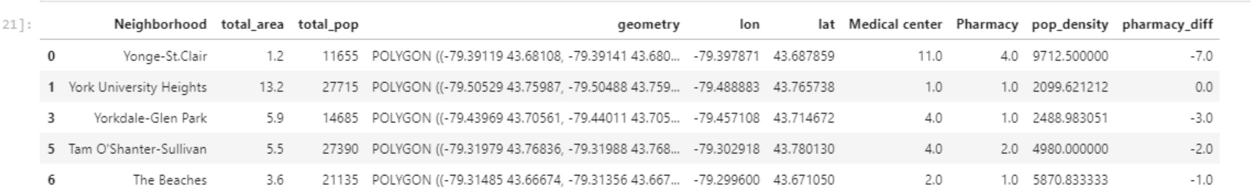
Data merging

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 two extra fields: population density (from total population and total area of each neighborhood) and 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