NEIGHBORHOOD SEARCH FOR BUSINESS

Children Day Care

ABSTRACT
Applied Data Science Capstone Project

Narasimha Koya
Learner

Contents

Introduction

As a part of Applied Data Science Capstone project, assignment is to identify a business problem and use foursquare data along with data available from open data sources or any other sources. Python is to be used for data transformation and analysis and then statistical algorithm like k-means to be used to perform machine learning activities to solve the business problem.

I have spent significant time during Week 3 to analyzing Canada postal codes and associated with data from four square and ran K-means algorithm against it to identify clusters etc. I would like to leverage this knowledge and continue to complete this assignment.

Business Problem

For this assignment I have chosen to solve the following business problem. Identify best neighborhood for establishing a day care center for children in Toronto city, Ontario, Canada to help small business owners to make a informed decision based on data analysis and insights derived from statistical analysis.

Background of the Problem

There are various considerations to be taking into establishing a new business and particularly for a daycare center for children. Some of the features, I would like to use to determine the best neighborhood for a daycare center in Toronto are population, crime rate and existing businesses in neighborhoods to determine their feasibility for establishing a new business.

Apart from this in a real-world scenario, there are various considerations to be taken before finalizing a location, like state laws, zoning laws, parking, and many other stringent safety requirements like, enough space for parking, safe play area, rest area for staff etc.

For the scope of this assignment though, I will be exploring only potential neighborhoods to establish the day care, but will not go in depth about location etc, as that is beyond scope of this exercise.

Data

To do this analysis, I am going to use below data sets

1. Scrape the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the data that is in the table of postal codes of Canada and to transform the data into a *pandas* data frame

	PostalCode	Borough	Neighborhood
0	M5G	Downtown Toronto	Central Bay Street
1	M2H	North York	Hillcrest Village
2	M4B	East York	Parkview Hill, Woodbine Gardens
3	M1J	Scarborough	Scarborough Village
4	M4G	East York	Leaside
5	M4M	East Toronto	Studio District
6	M1R	Scarborough	Wexford, Maryvale
7	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...
8	M9L	North York	Humber Summit
9	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...
10	M1B	Scarborough	Malvern, Rouge
11	M5A	Downtown Toronto	Regent Park, Harbourfront

2. Extract data from the csv file that has the geographical coordinates of each postal code and is located at http://cocl.us/Geospatial_data. Use the Geocoder package or the csv file to create the following data frame

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
1	M2H	North York	Hillcrest Village	43.803762	-79.363452
2	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
3	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
4	M4G	East York	Leaside	43.709060	-79.363452
5	M4M	East Toronto	Studio District	43.659526	-79.340923
6	M1R	Scarborough	Wexford, Maryvale	43.750071	-79.295849
7	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
8	M9L	North York	Humber Summit	43.756303	-79.565963
9	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	43.645711	-79.392732
10	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
11	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636

3. Merge the data from points 1 and 2 and extract only Toronto from it and visualize it in a Map.

4. Using Foursquare API and Toronto latitude and longitude information from above step, get venues data from Foursquare

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

5. Get population data for each neighborhood in Toronto, Canada at https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=25&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV and merge it with data in steps 1,2,3,4 and eliminate less populated areas.

Geographic name	Population, 2016	Total private dwellings, 2016	Private dwellings occupied by usual residents, 2016
A0A	46,587	26,155	19,426
A0B	19,792	13,658	8,792
A0C	12,587	8,010	5,606
A0E	22,294	12,293	9,603
A0G	35,266	21,750	15,200
A0H	17,804	9,928	7,651

6. Get Major Crime Indicator data for each neighborhood in Toronto, Canada from open data at https://opendata.arcgis.com/datasets/98f7dde610b54b9081dfca80be453ac9_0.csv?outSR=%7B%22wkid%22%3A102100%2C%22latestWkid%22%3A3857%7D&session=1751194201.1556194643
and merge with existing data and eliminate neighborhoods with high crime rate

ucr_e	offence	reportedy	reportedm	reportedd	reportedd	reportedd	reportedh	occurrenc	occurrenc	occurrenc	occurrenc	occurrenc	occurrenc	MCI	Division	Hood_ID	Neighbourhood	Lat	Long	ObjectId
200	Robbery - Mugging	2014	April	24	114	Thursday	12	2014	April	24	114	Thursday	11	Robbery	D55	68	North Riverdale (68)	43.66845	-79.3431	1
200	B&E	2014	April	24	114	Thursday	15	2014	April	24	114	Thursday	13	Break and	D31	24	Black Creek (24)	43.75929	-79.5079	2
100	Assault	2014	April	25	115	Friday	13	2014	April	25	115	Friday	13	Assault	D12	30	Brookhaven-Amesbury (30)	43.69755	-79.5017	3
100	Assault	2014	April	25	115	Friday	10	2014	April	24	114	Thursday	17	Assault	D23	4	Rexdale-Kipling (4)	43.7217	-79.5715	4
100	Assault	2014	April	25	115	Friday	16	2014	April	25	115	Friday	16	Assault	D11	114	Lambton Baby Point (114)	43.66389	-79.5035	5
100	Assault	2014	April	25	115	Friday	22	2014	April	25	115	Friday	22	Assault	D51	73	Moss Park (73)	43.65731	-79.3735	6
100	Assault	2014	May	3	123	Saturday	3	2014	May	3	123	Saturday	1	Assault	D55	64	Woodbine Corridor (64)	43.66636	-79.3166	7
100	Assault With Weapon	2014	May	3	123	Saturday	4	2014	May	3	123	Saturday	4	Assault	D14	79	University (79)	43.65811	-79.402	8
100	Assault With Weapon	2014	May	3	123	Saturday	4	2014	May	3	123	Saturday	4	Assault	D14	79	University (79)	43.65811	-79.402	9
100	Assault With Weapon	2014	May	3	123	Saturday	4	2014	May	3	123	Saturday	4	Assault	D14	79	University (79)	43.65811	-79.402	10

Methodology

Methodical approach I have performed for doing this analysis is as follows:

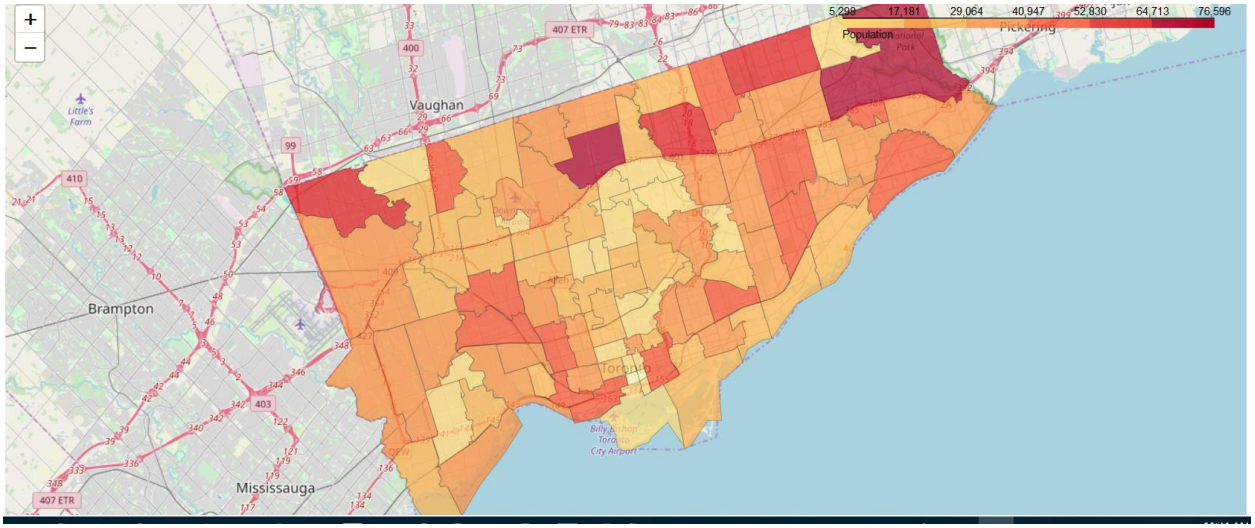
1. Step 1 - I have scraped Wikipedia data and collected Canada Postal Codes that start with M and the respective borough and Neighborhood information and stored it in a data frame and in further steps all the data is stored and manipulated in data frames.

2. Step 2 – I have cleansed the data collected in step 1 above and removed the rows that doesn't have any data and populated borough information in neighborhood column, where ever there is no neighborhood data.

3. Step 3 – I have then grouped the data per Postal code and neighborhoods.

4. Step 4 – I have extracted the needed latitude and longitude information for Toronto, Ontario and then merged the two data frames together.

5. Step 5 – To do my analysis of neighborhoods, I have extracted population data and crime data from Canada Statistics and Toronto police data websites and venue data from four square.



6. Step 6 – I have merged all this data together and analyzed the data to determine potential neighborhoods from Children day care.

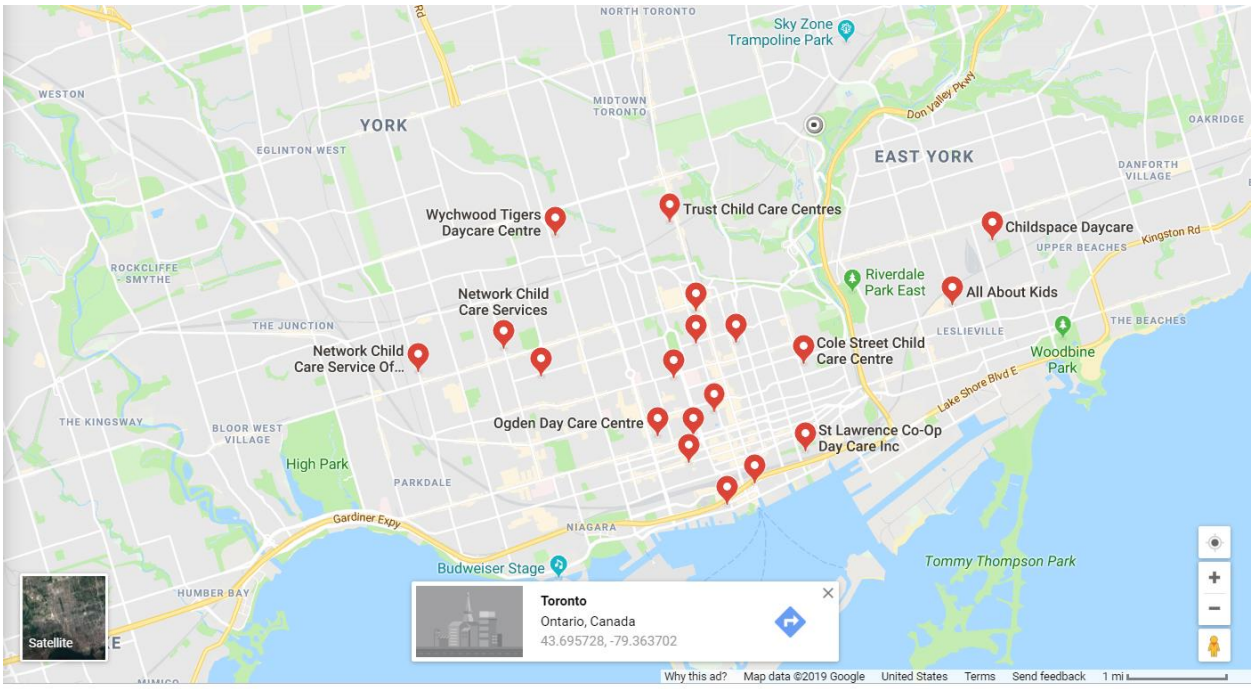
7. Step 7 – I have used K-means clustering algorithm (ML) to determine the cluster of neighborhoods, good for business.

Results

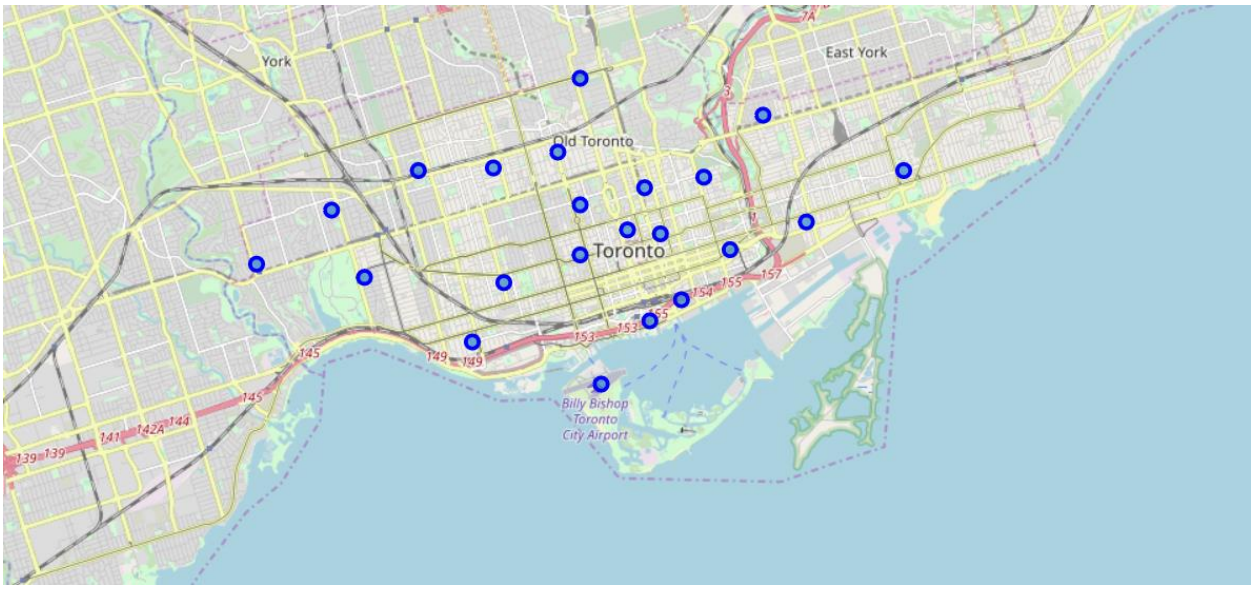
Upon identifying clusters, I have noticed which neighborhood cluster is best suitable for establishing the Child day care business and from examining the data frame results determined that iteration 5 gave me the best suitable cluster.

To validate my results, I have performed a google search to identify, the existing day care centers in Toronto and compared my results to that. Below we can find out both the results.

Results from google search



Results from my analysis



Discussion

- In a real-world scenario, first finding the neighborhoods would help short list potential location options, but after this exercise I felt, it would be interesting to find out the best location for a specific business than finding best neighborhood.

- With just the population and existing venue data, we were able to reduce the potential neighborhoods from 103 to 28 in the end.

- Adding couple of more data sets like number of schools in a neighborhood, office locations and transit options etc. would have enabled us to get more precise conclusions.

Conclusion

This course and assignment have taught me the potential to solve real world business problems, with the help of machine learning and good data sources.

I have understood the amount data cleansing and collection methods that is involved to do proper machine learning. I must admit that at times, I have felt some of the data cleansing operations are quite simple to do in SQL, but I think it depends on the time, I spent on each of these respectively.

From a technical standpoint, I have learned to use many Python functions and I have really liked, how flexible Folium maps are and ease of using Jupyter notebooks where we can at the same place do data analysis and visualizations.

References

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

http://cocl.us/Geospatial_data

https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

https://www.toronto.ca/city-government/data-research-maps/open-data/

https://medium.com/dataexplorations/generating-geojson-file-for-toronto-fsas-9b478a059f04

Google Images