IBM- COURSERA

APPLIED DATA SCIENCE CAPSTONE PROJECT
FINAL REPORT

By: Shadi Farahani
Date: 5/9/2020

Content:

Introduction

When a family travel to a city they usually look for play areas, playgrounds and entertainment places for their children. Also they would prefer those areas be close to hotels and restaurants. So as long as parents are enjoying hotels and restaurants they are able to go to play areas and have some fun with their children.

The objective of this project is to analyze and select the best locations in Toronto, Canada to choose neighborhood that has play areas and entertainment places. Also we will review which neighborhood would be better to choose if someone need to open play area for kids. At the end of this document will review which parameters is used in this project to compare best neighborhood.

The target audience for this report is anyone who is thinking of opening a play area for kids, or anyone who wants to choose the best location to stay during their journey.

Data Description

To solve this problem we need three different Data set:

1. To list the borough and neighborhoods we need to have a table which contains all name and neighborhood of Toronto City. This table contains Postal Code and Borough and Neighborhood of Toronto City[1] .

2. To get coordination of any location we need it on map in Toronto city. it is used a CSV file which is included all latitude and longitude of each postal code [2].

3. At last to have some information for any location like restaurants, salons,… I used Foursquare API that gives good data like latitude and longitude of any venue category[3].

All those data added to my project on their proper technique. I used scraping technique to extract data from web pages. After extracting data I have focused on Toronto city which it's coordination(latitude and longitude) got from geographical CSV file. This file was downloaded and added to my project. To get any information about locations and venue in Toronto City I used foursquare API. Foursquare API is used by thousands of developers to work with categories of venues. To use of Foursquare API any developer should open an account and then they will have Client_ID and Client_Secret and Version. With those parameters they are able to use information and have some request from that data base. I have used Playground category to solve defined problem by using some Data Science Techniques like data cleaning, clustering, data visualization and map visualization.

Methodology

To extract data from a web page I have used web scraping. To do this I used BeautifulSoup package in python. After extracting the neighborhood details of Toronto in Wikipedia, they are populated in data frame for using easier in future.
Then we have to get the geographical coordination of neighborhoods that we can plot folium graph. After having these two data sets I merge them together to get a table like figure 1. I have to mention that boroughs with Toronto word are selected.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District / Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

(figure1)

Those Boroughs are Downtown Toronto, East Toronto, West Toronto and Central Toronto. I have imported a map rendering library named folium to create Toronto map by using longitude and latitude of Toronto Boroughs. We could see the result in figure 2.

(figure2)

We will use Foursquare API to obtain the 100 top venues which are in radius 1000.
to do this any developer need to open an account in Foursquare to get a unique Client_ID and Client_Secret key. By those two parameters They could send some requests to get information like longitude, latitude, category and name of each venue. We make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a python loop. Foursquare will return a the venue data in JSON format. By these data we can check how many categories are for each venues then we will analyze each neighborhood. Also we could have the number of each category in each neighborhood so we could analyze and solve the problem by clustering.

Our problem is analyzing neighborhood with number of Playground in there so I filter the (Playground) category for neighborhoods. I have analyzed each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.

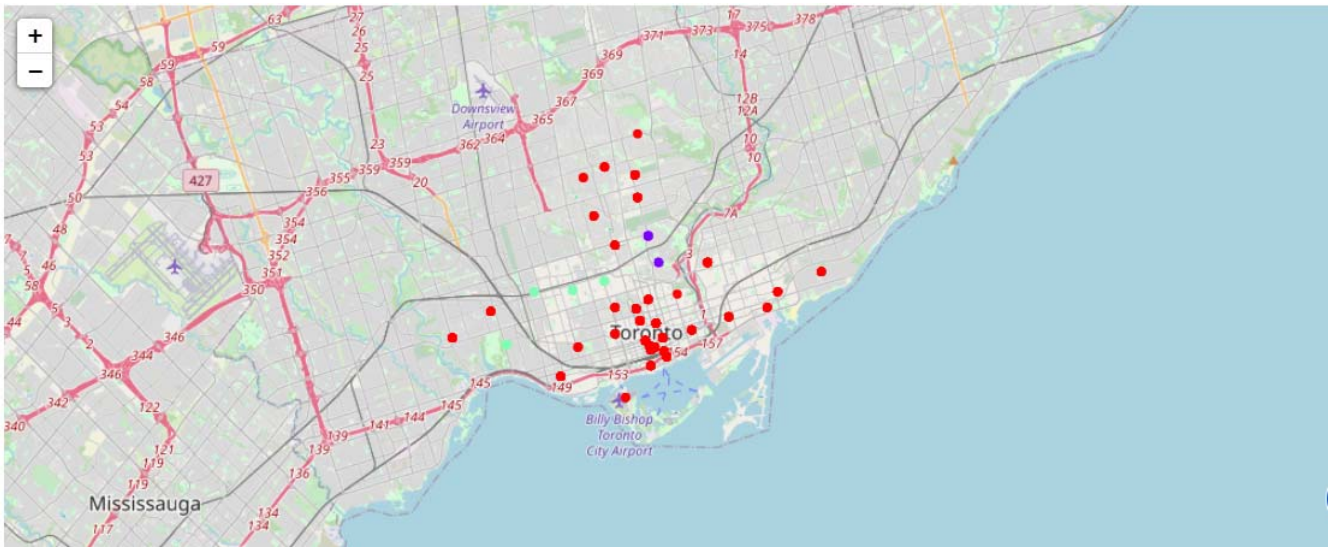
So far I have extracted data then I merge some other data to achieve a table of useful data. I have used a Machine Learning Algorithm to solve defined problem. K-means Clustering Algorithm that I used in this project is an Unsupervised Machine Learning Algorithm. The goal of K-means is to partition the N samples from your data set in to K clusters where each data point belongs to the single cluster for which it is nearest to[4].

For applying K-means clustering Algorithm I used K-means library in python which made it easy to use for developers. For this algorithm I choose number 3 for K and I have tested cluster numbers up to 3 and counted Playground category in each neighborhood. We could visualize clustering the Playground category by folium. I'll show you more results in numbers in continue. I need to explain that I am looking for Playground which is close to restaurants or hotels.

Result:

If you take a look at map in Figure 3 you will see most that of the playgrounds are located in down town. But most of the playground which is close to hotels and restaurants are located in Central Toronto.
we could categorize the result of this project into 3 clusters based on the frequency of occurrence for “Playground” :

1. cluster 0: in this cluster we have a lot of Playground which shows most number of Playground in this cluster.
2. Cluster 1: this cluster contains very low of Playground .
3. Cluster 2: This cluster comprised of neighborhood of neighborhoods with moderate concentration of Playground.



(figure3)

Discussion:

As I mentioned above, there are 362 Playground located in cluster 0 while there are just 7 Playground in cluster 1 that are the most and least playground in Toronto. Cluster number 2 has moderate number of Playground with number 24. Cluster 0 includes Down Town area so we could say most of the playground are located in Down town. While next level is for West Toronto that is in cluster 2. and at last cluster number 1 has a few number of Playground that includes Central and west Toronto.

So we could say If some one wants to open a playground or any play area for kids Down Town and West Toronto is the worst place cause has less benefits. While as we could see in results East Toronto has the most opportunity to open a business for children entertainment like any playground.

Because I compared the playground location and hotels and restaurants, we could say Down town and West Toronto would be the best choose for tourist specially for some one who has kids.

Conclusion

In this project we solved a business and tourism problem. I extracted data and cleaned them to be prepared for analyzing. K-means Clustering was the Machine Learning Algorithm that is used in this project to analyze the neighborhood better. Each cluster includes some neighborhoods. To solve the problem we are looking for neighborhoods which has more opportunities to run a business. As shown in result section cluster 0 has less benefit to open a new Play ground or any children entertainments. But we could say East Toronto is the best choose to open run business. While we could say Down Town is very nice for tourism and some one who travel to Toronto. Specially for who has kids.

References

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

2. https://cocl.us/Geospatial_data

3. www.Foursquare.com

4. https://en.wikipedia.org/wiki/K-means_clustering