Applied Data Science Capstone Week 5

Introduction

Toronto is a fast-growing city in Canada, more and more people plan to start their own business there. Hypothetically, there might not so many gyms in Toronto, because working out is a very important way for everyone to keep fit in current society, therefore, an entrepreneur probably wants to take the good opportunity to start his own gym in Toronto. For the entrepreneur, finding a suitable location to open a gym is a big decision. The purpose of this project is to help people who would like to open a Gym in Toronto Area. It will give people some useful suggestions to make an efficient decision on selecting a good location to open a gym. In this project, some data science methods, tools, and machine learning algorithms will be used. The project aims to create analysis and provide solutions for the business problem: where is the most suitable location to open a gym for an entrepreneur in the Toronto area?

Data

To analyze and solve the business problem, the data we will use is as follows,

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

http://cocl.us/Geospatial_data

- Postal Code
- Borough
- Neighborhood
- Latitude
- Longitude

2.Foursquare API Data

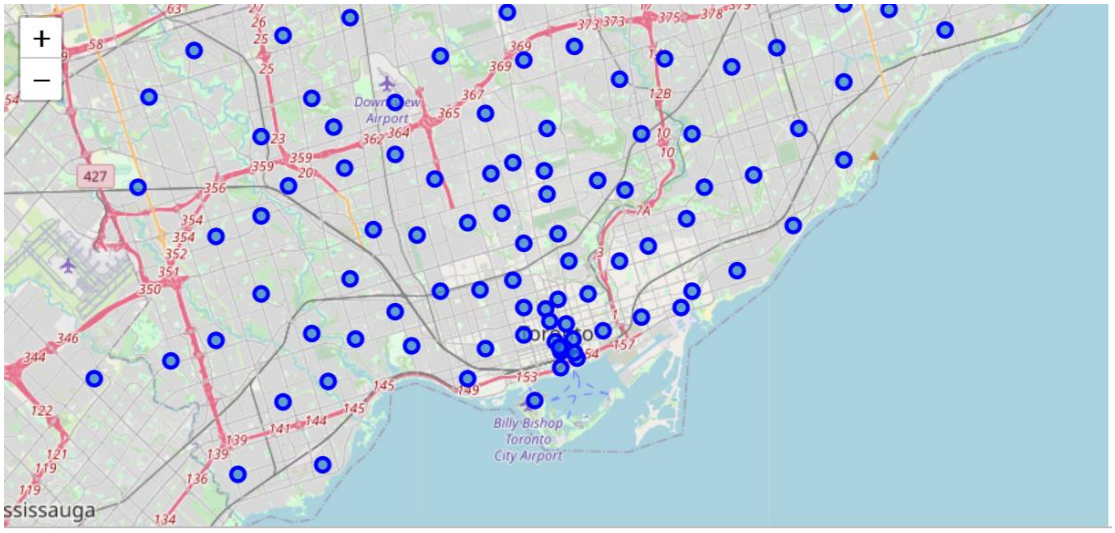
- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category

We will use the data from the Wikipedia website in week three and also the data from the Foursquare API. Specifically, we need to have the list of neighborhoods in the Toronto area and the coordinates of these neighborhoods, we should also find the venues which are related to Gym. To acquire the data, web scraping, Geocoder package, Foursquare API, and some other methods will help us to extract these useful data in this project.

Methodology

First of all, according to this Wikipedia website: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M , we could know the data about many boroughs in Toronto and the neighborhoods in different boroughs. To get the list of the neighborhoods in Toronto, I used web scraping with Python and Beautiful Soup, it is a useful and effective way to pull the data you want from a website and export a CSV.

Now, we just have a list of postal codes, borough, and neighborhood, we still need to get the related coordinates that could be used in Foursquare to explore more details. Then I used the CSV file which is provided by IBM. I merged the two datasets by their same postal codes. In this way, I acquired the coordinates of different neighborhoods. After that, I created a map of Toronto with Folium to test the coordinates.



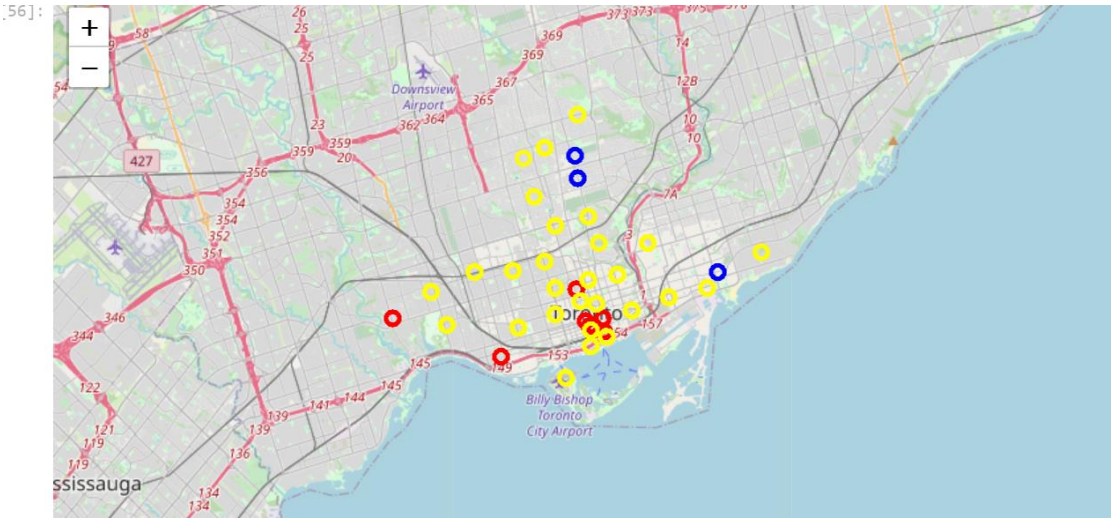
Because I have already created a Foursquare developer account, I could use my personal CLIENT_ ID and CLIENT_SECRET. I only selected boroughs that contain the word “Toronto” in Toronto data to explore. In the Toronto area, I get the list of top 100 venues with a 500 meters radius by Foursquare API, also, I could get these venues’

names, latitude, longitude and categories. In addition, I was able to know the number of unique categories for these different venues. Next, I explored and analyzed each neighborhood by grouping them and getting the mean about the frequency of occurrence of each venue category.

In the end, specially, I searched “Gym” in the venue category to prepare for clustering later. The K-means clustering algorithm is what I used in this project. K-means could divide the data into K non-overlapping clusters and it tries to minimize the intro-cluster distances and maximize the inter-cluster distances. In short, it is a simple unsupervised learning algorithm by defining k centers, one for each cluster. I divided the neighborhoods data in Toronto data into 3 clusters according to the frequency of “Gym”. Therefore, I could give some suggestions about the good locations to open a gym.

Results

See the clusters on the map.



By the K-means clustering, the neighborhoods in the Toronto area were categorized into 3 clusters. They are cluster 0 (red), cluster 1 (yellow), cluster 2 (blue).

Discussion

In cluster 0 (red), the neighborhoods have a normal number of gyms. In cluster 1 (yellow), the neighborhoods have the lowest number of gyms. In cluster 2 (blue), the neighborhoods have the highest number of gyms. Most of the gyms are near Davisville and Davisville North. I would suggest opening a gym in cluster 0 (red) or cluster 1

(yellow), also, not too far from downtown, some places like Church and Wellesley, Harbourfront, St. James Town, and Stn A PO Boxes could be good choices.

Conclusion

This project is to give suggestions for people who would like to open a gym in the Toronto area. Some methods or tools like web scraping, Foursquare API, Folium are utilized in this project. Also, by using the K-means clustering algorithm, the neighborhoods of Toronto area are separated into 3 different clusters for exploring and analyzing the data.