Coursera IBM Data Science Capstone Project

Finding best location for Fitness centar in Toronto

Barbara Brzic

Introduction

For the final project I'm trying to find best location in Toronto area to open Fitness centar. With modernization of society we have been less and less active which started affecting health of the population. Increasing number of people are becoming overweight in developed countries. But on the bright side, with social media, fitness and health has become popular so there is demand for the fitness centers especially in big cities like Toronto.

Business problem

Objective is to find the most suitable location for fitness centar in Toronto, Canada. Idea is to use data science methods learned during this Capstone, such as clustering, segmentation and with the help of Foursquare API. Question is where to open Fitness centar and idea is to open it in the area where there is least number of Fitness centers and gyms, just to have least number of competition.

Location, why Toronto?:

Toronto is the most populous city in Canada and the fourth most populous city in North America. It is recognized as one of the most multicultural and cosmopolitan cities in the world. According to Ipsos survey one in three (33%) Canadians say improving their personal fitness and nutrition is their top new year's resolution, compared with only 21 per cent who chose to focus on financial goals. More than half (53%) of Canadians say improving their overall quality of life is the primary motivation for pursuing a health and wellness resolution. Preventing health risks (45%), losing weight (42%) and increasing their energy (41%) ranked as other top reasons to exercise more and eat better. One in five Canadians (18%) say they would join a gym. Which means that demand for the fitness facilities in Toronto.

Foursquare API:

Foursquare API is used in this project as source of data, as it has a database of millions of places, with API which provides the ability to perform location search, location sharing and details about business.

Libraries used:

Pandas: For creating and manipulating dataframes
Folium: Python visualization library – used to visualize cluster distribution
Scikit Learn: importing k-means clustering
JSON: Library to handle JSON files
XML:
Geocoder: To retrieve Location Data
Beautiful Soup: to scrap and library http requests
Matplotip: Python Plotting Module

Data used:

- List of neighborhoods in Toronto, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Latitude and Longitude of neighborhoods
- Venue dana related to Gym/Fitness centers.

Methods for extracting data:

- Web scraping of Toronto neighborhoods from Wikipedia
- Using Geocoder package for latitude and longitude data of neighborhood
- Using Foursquare API to get venue data

Methodology:

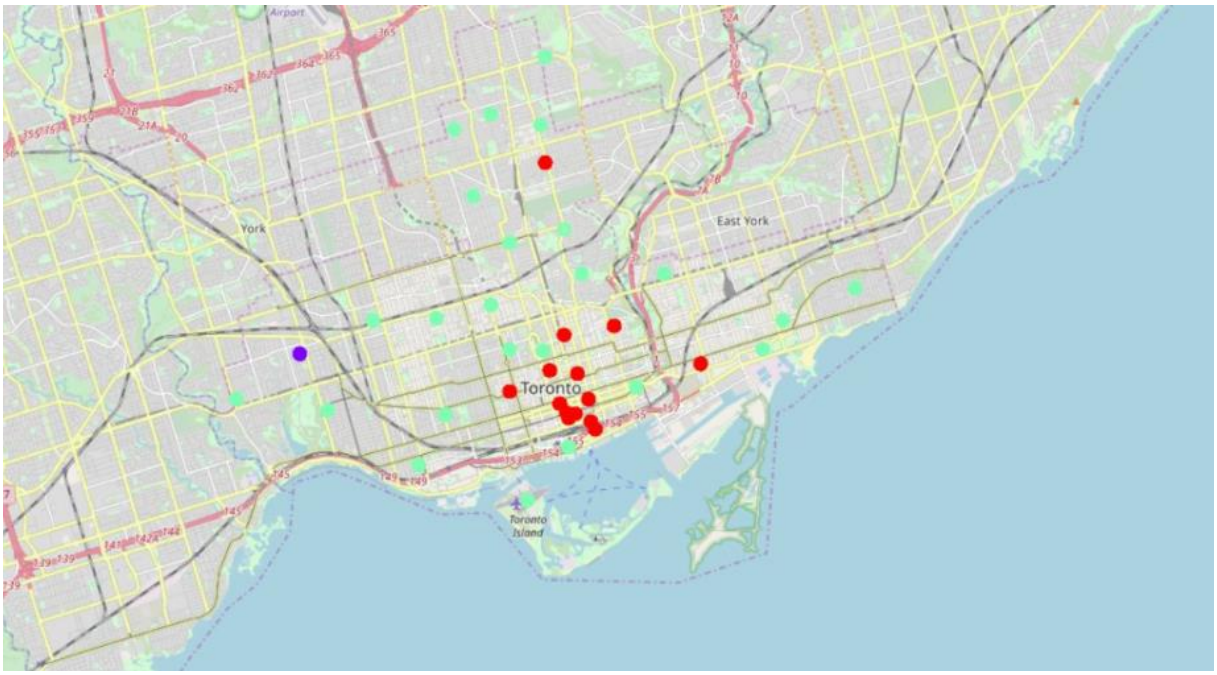
First challenge of this project was finding list of neighborhoods in Toronto, Canada. Since there are no data table on the internet, list was scrapped from Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), by using pandas html table scrapping method and Beautiful Soup. Data is than pulled in data frame with postal codes.

Next step was to find coordinates of the neighborhoods and venues. Coordinates are pulled from Foursquare, which is database of million places with API which provides the ability to perform location search, location sharing and details about business. To get coordinates Geocoder package is used and then map is created and visualized using Folium package to better understand positions and density of venues in different neighborhoods.

After coordinates are extracted and merged with table of neighborhoods, Foursquare API is used to extract list of the top 100 venues within 500 meters radius. Foursquare developer account is created to obtain account ID and API key to pull the data. From Foursquare name, categories, latitude and longitude of the venues is obtained. With this data number of unique categories is extracted so we can search location of categories that we are interested of presenting and to see the frequency of each specific category.

Now data is ready to search for category of venue that interest us. Category “Gym/Fitness centar” is searched so we can see where are they located, where is good place to open new fitness centar. This is done with clustering method by using k-means clustering, which identifies k number of centroids and then allocated every data point to the nearest cluster, while keeping the centroids as small as possible. This is one of the unsupervised machine learning algorithms that we learned during this course. Neighborhoods were clustered in 3 clusters based on their frequency of occurrence of “Gym/Fitness venue”.

Results:



Results show that lowest number of fitness centers are in clusters 0 and 1, while largest in cluster 2.

Cluster 0 is blue, cluster 1 red and cluster 2 green. Based on this it would be smart to open fitness centar in cluster 0 neighborhoods.

Because of all there results my recommendation would be too look at possible locations for opening fitness centar in cluster 1 and cluster 0, which is around neighborhoods such as Agincourt, Moore Park, Harbourfront East, Golden Mile, Clairlea, Oakridge in cluster 0 or Downsview, Parkview Hill, Woodbine Gardens, Davisville North in cluster 1. Cluster 2 with neighborhood such as Garden District, Ryerson, Underground city, etc. Is too busy with fitness centers and gym, so it would be really hard to attract new customers. Especially because most of clients prefer gym that they know, and they have some sort of brand loyalty and they prefer to go to closest gym to their home.

Conclusion:

Since most of the people prefer fitness centers and gyms that are close to their home it would be smart to open fitness centar in the neighborhood where there is least number of similar venues to attract new customers who are maybe driving to the gym that is far away from them because they don't have better option. By building initial customer base of people who live closest to fitness centar, business will have enough profit to return initial investment and to invest more in marketing to attract customers that live little further away.