The Battle of the Neighborhoods

Kate Pelletier

11/3/2019

Introduction

In this capstone project, I will try to predict the best neighborhood to open a fitness studio (specifically, a pilates studio) in Toronto, Canada. I will use the Toronto neighborhoods dataset from the previous lessons. I will decide which neighborhood to place the fitness studio in based on the proximity of other health-centric venues.

Problem

We need to identify neighborhoods of Toronto that already offer other fitness/health focused venues (other types of gyms, health restaurants such as salad, smoothies, juice, vegetarian) **but** have minimal pilates studios. This will maximize value of the pilates studio while already being in an area that caters to health-focused clientele.

Background

I'll use the skills learned throughout this data science course and during the capstone project to appropriately recommend a suitable location. Neighborhoods will be as previously defined in the last exercises

Data

Determining where to place the pilates studio depends on the following factors:

- the number of other gyms in the neighborhood
- the number of pilates studio in the neighborhood (aim to minimize)
- the number of health-conscious restaurants in the neighborhood (aim to maximize)

First, I will find the number of total gyms in the neighborhood and the total number of pilates gyms in the neighborhood. Then I will calculate what percentage of gyms are NOT pilates gyms, aiming to maximize the %. Next, I will find the number of health-conscious restaurants and venues in the neighborhood, based on specific venue categories within Foursquare (smoothie shop, juice bar, salad place, sandwich place, soup place, vegetarian/vegan restaurant, gluten-free restaurant, health food store, farmers market, organic grocery) Finally, I will assign a weight to total number of gyms, total percent not pilates gyms, and total number of health-related venues to determine the optimal neighborhood for the new pilates studio to be located.

Data Cleaning

Only certain foursquare venues were of value to us while looking at this data:

- Shop & Service
  - Food & Drink Shop

    - Farmers Market
    - Health Food Store
    - Organic Grocery
    - Fruit & Vegetable Store
    - Massage Studio
    - Sauna/Steam Room
    - Smoothie Shop
    - Supplement Shop
- Professional & Other Places
    - Medical Center
    - Acupuncturist
    - Alternative Healer
    - Chiropractor
    - Nutritionist
- Outdoors & Recreation
    - Gym / Fitness Center
    - Boxing Gym
    - Climbing Gym
    - Cycle Studio
    - Gym Pool
    - Gymnastics Gym
    - Gym
    - Martial Arts Dojo
    - Outdoor Gym
    - **Pilates Studio**
    - Track
    - Weight Loss Center
    - Yoga Studio
- Food
    - Vegetarian/Vegan Restaurant
    - Soup Place
    - Salad Place
    - Sandwich Place
    - Juice Bar
    - Gluten-free Restaurant

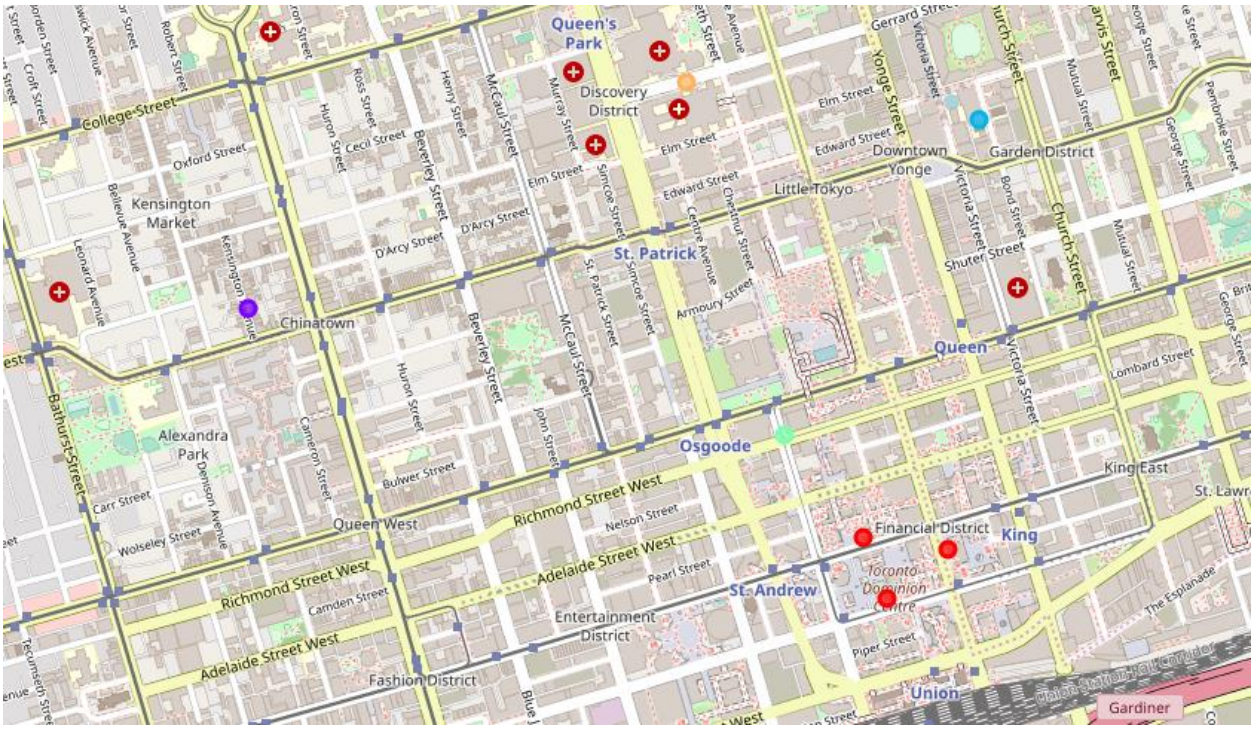
Methodology

I employed two basic methods for solving this problem – one-hot-encoding and k-means clustering. The general workflow was to obtain two lists: 1 list of all health-conscious venues in a 1.2 mile radius, and another list of all currently existing pilates studios in the same radius. We will then exclude any neighborhoods that currently have a pilates studio and choose a location from the remaining neighborhoods. We first found all applicable venues within the radius and displayed that information in

a table. Then, by grouping and sorting the data, we could visualize which neighborhoods had the most health-conscious venues. To take it a step further, we then repeated the process filtering only on pilates studios and filtered neighborhoods that already had pilates out of the list of prospective list of locations.

Results

It's of note that of the final 7 prospective locations for a pilates location, 3 are clustered very close to each other. This could reinforce that any of those 3 neighborhoods are good options for the studio, because clientele may be more likely to travel between neighborhoods.



This image does, however, suggest that centroids and neighborhood radiuses may need to be reassessed or adjusted. It does appear that 3 "neighborhoods" are only about a block apart, which merits a closer look at how the Wikipedia data source classified neighborhoods in the first place. It may be useful in the future to cluster distinct neighborhoods even further than just using their postal code, since postal code is based on population density and not distance.

The other useful bit of information comes from the frequency chart generated from the top prospect list:

```
----Adelaide, King, Richmond----
             venue  freq
0              Gym  0.40
1              Spa  0.16
2   Massage Studio  0.04
3  Supplement Shop  0.04
4  Doctor's Office  0.04


----Central Bay Street----
                  venue  freq
0  Gym / Fitness Center  0.30
1                   Spa  0.15
2                   Gym  0.15
3         Smoothie Shop  0.09
4           Yoga Studio  0.09


----Chinatown, Grange Park, Kensington Market----
            venue  freq
0             Spa  0.44
1  Farmers Market  0.15
2     Yoga Studio  0.07
3   Smoothie Shop  0.07
4  Massage Studio  0.07


----Commerce Court, Victoria Hotel----
                  venue  freq
0                   Gym  0.33
1                   Spa  0.18
2  Gym / Fitness Center  0.11
3     Health Food Store  0.07
4         Smoothie Shop  0.07
```

```
----Design Exchange, Toronto Dominion Centre----
                  venue  freq
0                   Gym  0.29
1                   Spa  0.22
2         Smoothie Shop  0.11
3  Gym / Fitness Center  0.09
4     Health Food Store  0.07


----First Canadian Place, Underground city----
               venue  freq
0                Gym  0.36
1                Spa  0.21
2      Smoothie Shop  0.09
3     Massage Studio  0.06
4  Health Food Store  0.06


----Ryerson, Garden District----
                  venue  freq
0                   Spa  0.21
1                   Gym  0.18
2  Gym / Fitness Center  0.15
3          Yoga Studio  0.09
4            Juice Bar  0.06
```

With the vast frequency of gyms, this data reaffirms that the cluster of 3 red dots on the map may be a great fit for a new pilates studio.

Discussion

This was a good first pass at determining a location to place an up-and-coming pilates studio based on existing venues that may draw the correct crowd. I think good next steps include selecting the top 5 prospects and taking a deeper dive into the ratios of the venues already exist there to see if any particular neighborhood may be more adept to welcoming a pilates studio - for example, the top neighborhood may simply have the highest number of produce stores, but not many gyms, which may indicate a pilates studio may not do as well in that area. It may also be of interest to adjust the radiuses - without being deeply familiar with the Toronto metropolitan area, it's difficult to estimate what a realistic radius is, but I assume a smaller radius would be worthwhile to investigate.

Conclusion

The purpose of this notebook was to help a potential business owner decide the most profitable and successful location for their up and coming pilates studio, based on prior existence of health-conscious type venues and competing pilates studios. I used the foursquare API to filter results by venue types that were health-conscious, and then sorted the data by the number of venues in a 1.2 miles radius of the center of a given neighborhood.