Analyzing Yoga Studio Location Data in the Greater Toronto Area

By Ana White
November 14th 2020

1. Introduction

a. Background

Yoga is a mind and body practice that originated in India. Originally, it's a group of physical, mental, and spiritual discipline. However, it's becoming more popular as a workout routine and thus, yoga studios are popping up in cities like Toronto.

b. Problem

If one were to open a yoga studio, one must do a thorough market analysis. But most importantly, location is key for most businesses. This project will attempt to answer where a yoga studio is most likely to be situated. There are many facets to this question but we would be analyzing the geographical and nearby venues of all yoga studios in Toronto.

2. Data Acquisition and Cleaning

a. Neighborhood Location Data

The Toronto neighborhood name and postal codes were obtained from scraping the wikipedia page https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto. This was then combined with the Toronto geographical coordinates obtained from the previous assignment. In this project, instead of using the Borough aggregate, the dataframe was modified to break out the neighborhoods that are in the same Borough into individual rows. The result is a list of neighborhoods with postal codes. names, latitude and longitude.
Final data below:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
0	M3A	North York	Parkwoods-Donalda	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
2	M5A	Downtown Toronto	Moss Park	43.654260	-79.360636

b. Yoga Venues in Each Neighborhood

For each neighborhood in the neighborhood data previously mentioned, a Foursquare API is sent to retrieve all yoga studios within 5km of the neighborhood coordinates. The resulting json response is then converted into a dataframe. After the duplicates were removed, 62 unique yoga venues remained.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
0	Parkwoods	43.753259	-79.329656	LA Fitness	43.747665	-79.347077
1	Parkwoods	43.753259	-79.329656	Breathe in Yoga	43.777095	-79.348193
2	Parkwoods-Donalda	43.753259	-79.329656	LA Fitness	43.747665	-79.347077
3	Parkwoods-Donalda	43.753259	-79.329656	Breathe in Yoga	43.777095	-79.348193
4	Victoria Village	43.725882	-79.315572	LA Fitness	43.747665	-79.347077

c. Venues near Yoga Studios

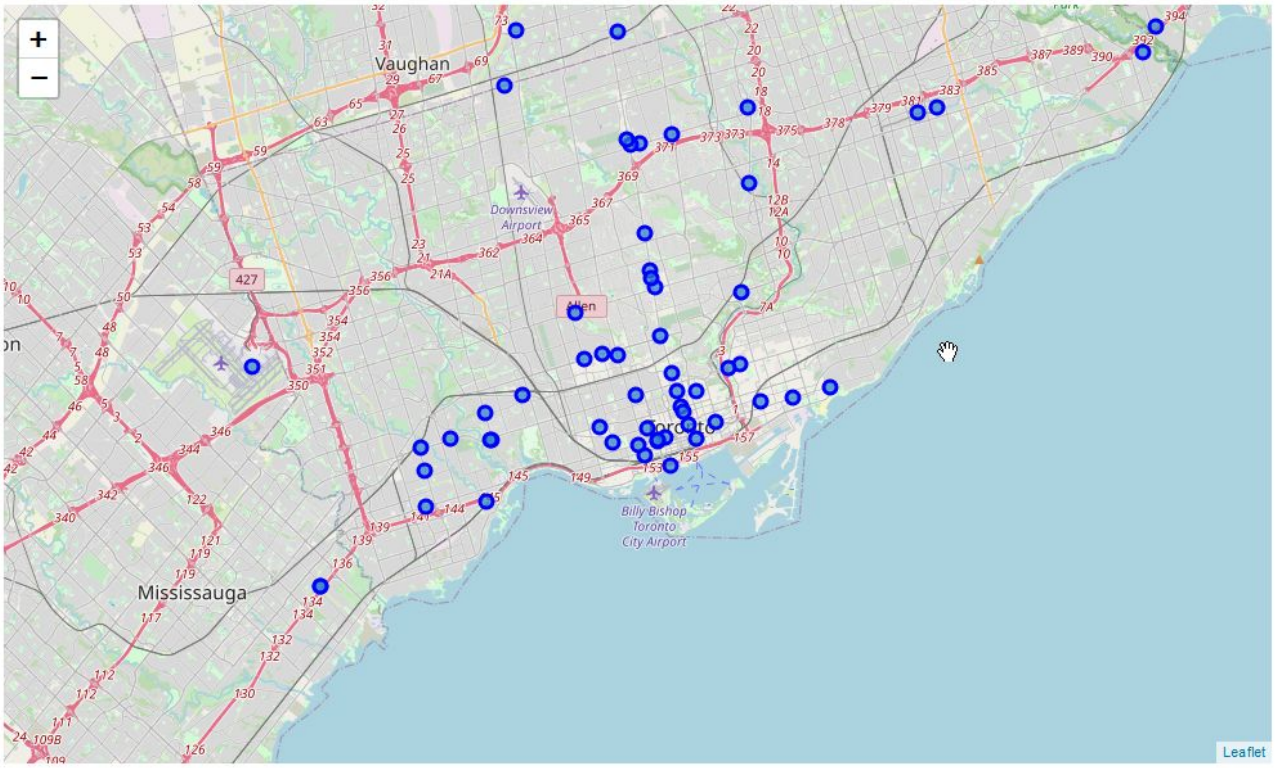
To build this dataframe, I created a function that loops through the Yoga Venues Latitude and Longitude and passed the values in the Foursquare API and stored the results in a dataframe. I set the radius to 200m. The rows tagged with venue type "Yoga Studio" were dropped from the resulting dataframe.

	Neighborhood	Yoga Place	Yoga Place Latitude	Yoga Place Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	LA Fitness	43.747665	-79.347077	LA Fitness	43.747665	-79.347077	Gym
1	Parkwoods	LA Fitness	43.747665	-79.347077	Subway	43.748423	-79.347207	Sandwiches
2	Parkwoods	LA Fitness	43.747665	-79.347077	the REALTOR® Store @ TREB	43.749428	-79.347339	Office Supplies
3	Parkwoods	LA Fitness	43.747665	-79.347077	Wooffles & Cream	43.748159	-79.348235	Desserts
4	Parkwoods	Breathe in Yoga	43.777095	-79.348193	Tommy Hilfiger	43.777427	-79.345893	Apparel

3. Exploratory Data Analysis

a. Aggregation and Visualization

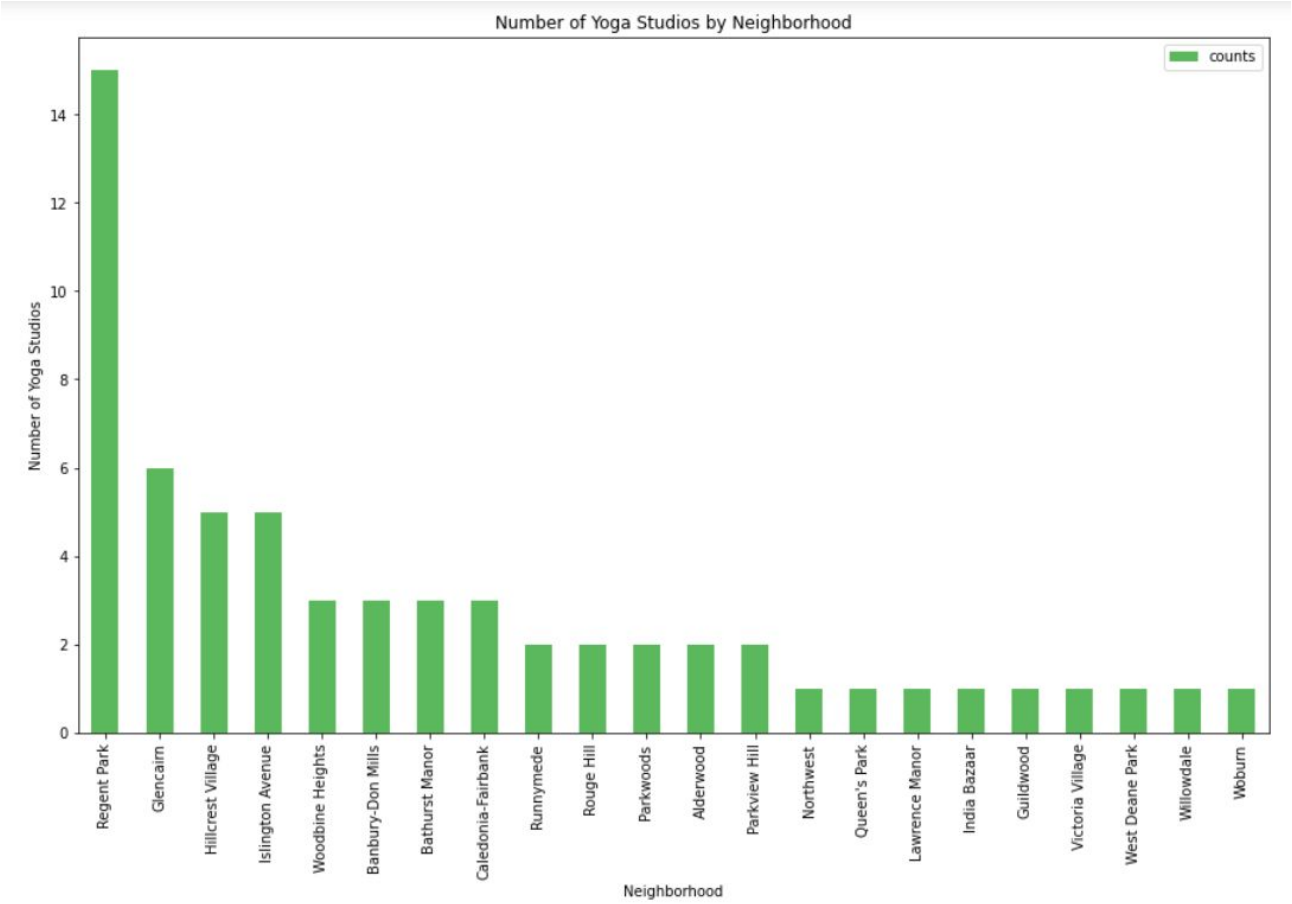
The easiest way to explore the data was to visualize the location of all the yoga studios on a map. As observed in the map below, the majority of the yoga studios are in Central/Downtown Toronto and around the main subway lines. There are visible outliers in the Greater Toronto Area and neighborhoods that do not show any Yoga Studio.



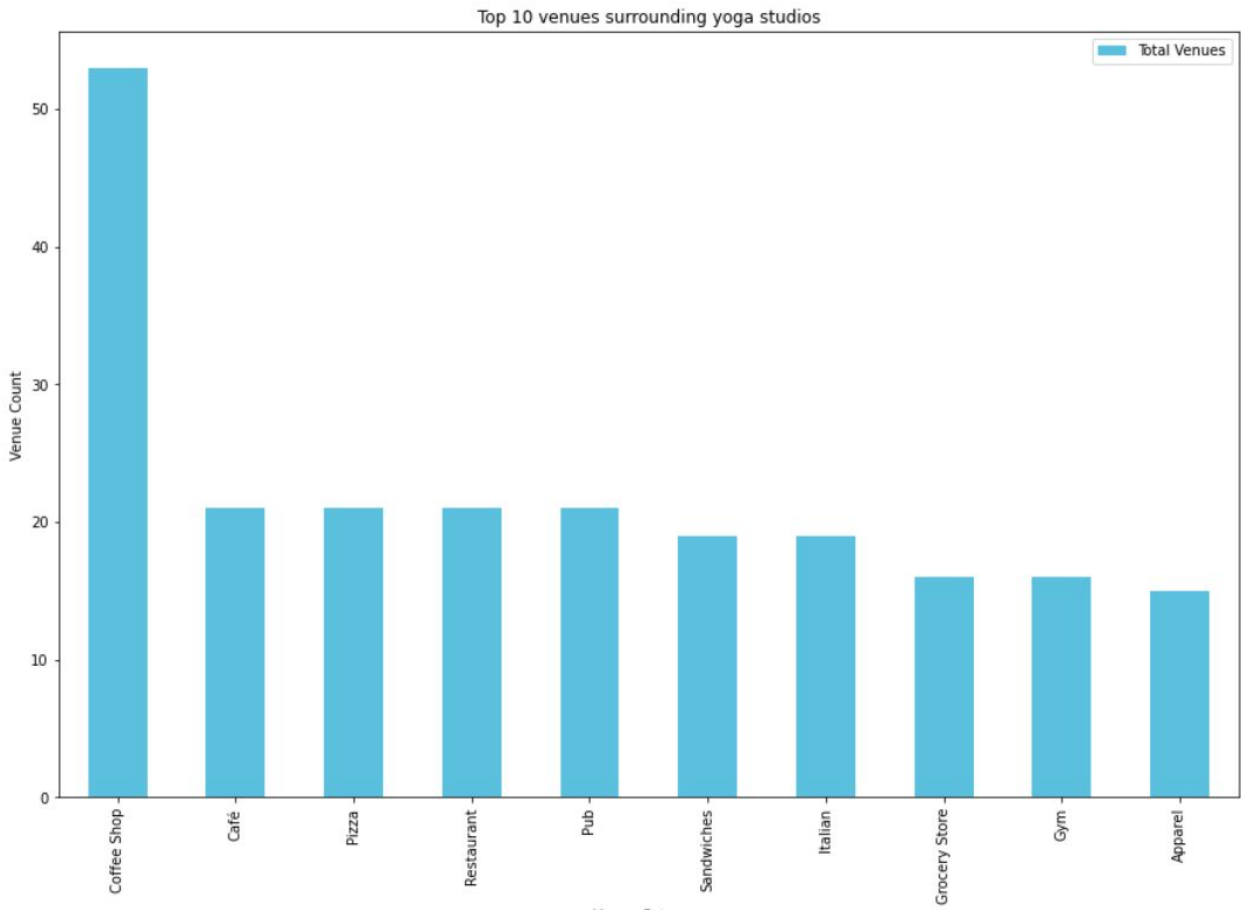
I then grouped the data by neighborhood and thus generated a bar chart by showing the number of yoga studios per each neighborhood. I sorted the data based on the highest count first and below is the 10 neighborhoods with the highest counts.

	Neighborhood	counts
0	Regent Park	15
1	Glencairn	6
2	Hillcrest Village	5
3	Islington Avenue	5
4	Woodbine Heights	3
5	Banbury-Don Mills	3
6	Bathurst Manor	3
7	Caledonia-Fairbank	3
8	Runnymede	2
9	Rouge Hill	2

I created a bar chart off this data to visualize 22 different neighborhoods



I also summarized the number of venue categories surrounding yoga studios. Below is a chart of the 10 most frequent categories.



b. K-means Clustering

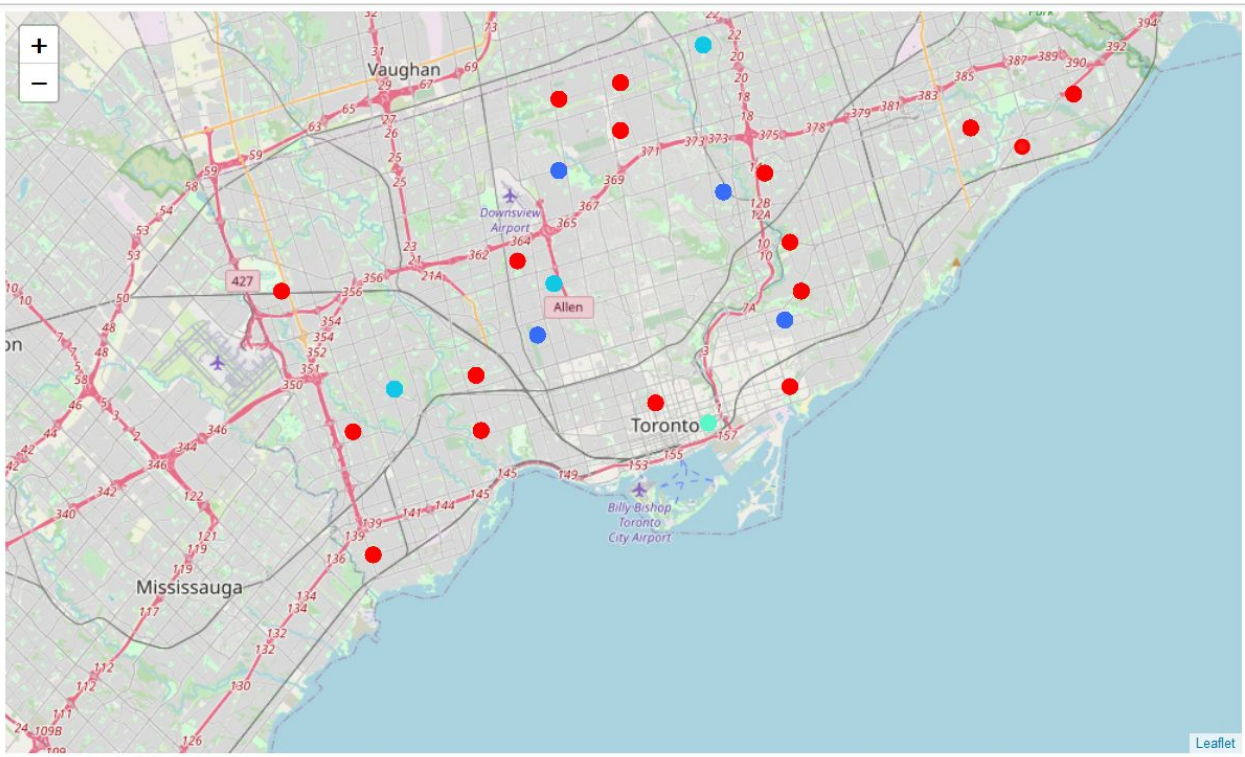
The result above is an aggregate so I wanted to find if I could categorize the neighborhood based on the common venues that are around yoga studios. In order to achieve this, I merged the count of yoga studios with the surrounding venues data and stored the result as a new dataframe.

	Neighborhood	Venue Category	Category Count	#Yoga Venues
454	India Bazaar	Event Space	1	1
455	India Bazaar	Greek	1	1
456	India Bazaar	Sandwiches	1	1
457	India Bazaar	Pharmacy	1	1
458	Guildwood	Martial Arts	1	1

I then used k-means clustering to cluster the neighborhood and uncover the most common surrounding venues. In order to make this analysis work, I had to use One Hot Encoding to change the Venue Category to numerical data by changing it into dummy variables. I tried several values of k, the smaller values did not segment the data as well as expected. I settled on 8 which seemed to be partitioning the data better.
Below is the breakdown by Cluster label after the k-means cluster analysis.

Neighborhood Cluster Label	**Top Venue in that Category**
0	Restaurant
1	Vegetarian/Vegan
2	Spa
3	Fast Food
4	Movie Theater
5	Coffee Shop
6	Coffee Shop
7	Pub

I created a map of the neighborhood clusters but only 4-5 colors appear. Each color represents one cluster.



4. Conclusion

Through the exploratory analysis, it has been determined that the neighborhood of Regency Park has the most yoga studios in Toronto. Coffee shops are the most popular nearby venue within a 200m radius of a yoga studio. I believe this is due to the fact that coffee shops are popular in downtown areas where there is a high concentration of people. Using the data visualization, we can conclude that yoga studios are concentrated in neighborhoods that surround public transits such as subways. From the cluster analysis, most of the nearby venues are food venues (restaurants, bars, coffee shops etc..) with the outliers being movie theaters and spas. Some future consideration of this analysis would be to incorporate real estate rental and yoga customer data to study the likelihood of "success" of a yoga studio in one particular area.