Sports PT Clinic in Manhattan

A Data Science Project

Neeraj Baheti, PT

June, 2020

1. Introduction

Originally from a big city (Mumbai, India), I have always had a lure of starting my own sports physical therapy practice in the "Big Apple". I have been working with youth athletes for over 15 years, treating those who play sports such as Football, Soccer, Baseball, and Basketball. There are almost eight million high school athletes that participate in sports in United States.[1] According to National Federation of High School Athletic Associations, more than half a million public high school athletes participated in sports, in 2018-2019, in the state of New York.[2] So, I started looking in to opening a PT clinic in New York City, specifically in Manhattan. I found myself asking the question where exactly, in Manhattan, should I open my sports PT clinic? I wanted a location that would be central to my prospective patients, the high school athletes. So, I decided to use my newly acquired skills of data science analytics, to help me answer this question. Anyone who is interested in starting a sports PT clinic in Manhattan, is the target audience of this project.

2. Data Acquisition

Since my specialty is working with high school athletes, it will be important to get more information about the high school located in Manhattan. Information such as names, location, and sports participation. I was able to acquire some of this information at the following Wikipedia page: https://en.wikipedia.org/wiki/List_of_high_schools_in_New_York_City.[3] In order to get the location data of all these high schools, I would use python libraries geopy to plot the high schools on a map of Manhattan.[4] After that, I can cluster the data to find a location(s) that is/are most centrally located to the high schools in Manhattan. Prior to applying the methods, mentioned above, I was hoping that the data analytics will provide me with a couple of different locations that I could further compare for additional factors such as commercial rental costs and existing competition. The list of existing physical therapy clinics, in Manhattan, can be found on Foursquare,[5] using the search endpoint in the query. Finally, commercial rental prices can be obtained from Metro-Manhattan website.[6]

3. Methodology

The list of all the high schools in Manhattan was obtained from a Wikipedia page at this link https://en.wikipedia.org/wiki/List_of_high_schools_in_New_York_City.[3] This link lists all the highs schools in New York City in a tabular format. The table with Manhattan schools was extracted in to a pandas dataframe. Data slicing was performed to only have the columns *School, Latitude, and Longitude.* Using the geopy python library, I obtained the latitude and longitude data for the high schools and I inserted it in the dataframe. The geopy library was not able to get the latitude and longitude data for some of the schools. I obtained that information using Google search and inserted that manually. In Figure 1, you can see the Manhattan map created using Folium library, with each of the high schools denoted as a marker.

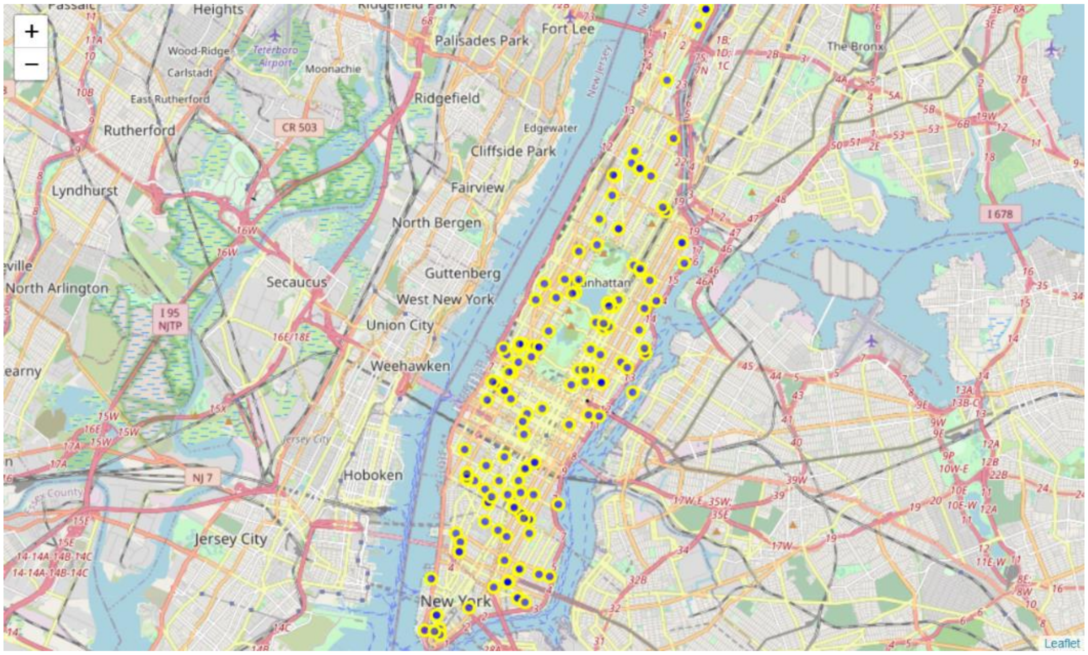

Figure 1: Location of all the high schools Manhattan on a map created using Folium library.

With the goal of finding a central location, for my physical therapy clinic, I decided to calculate K-means. Prior to running the K-means calculations, to find the centroids for the cluster of high schools, I used the Elbow method to determine the appropriate number of clusters for the data at hand.[7,8] Elbow method is one way of determining the ideal number of clusters. Looking at Figure 2, you will see that the graph looks like the elbow joint, with the “tip” of the elbow being at number three. Even though the Elbow method recommended that I create three clusters, I decided to go with four clusters. This is because, with three cluster calculations, one of the centroids was in the middle of Central Park (duh, it’s in the name!). Also, having four centroids seemed like I had more options to choose from.

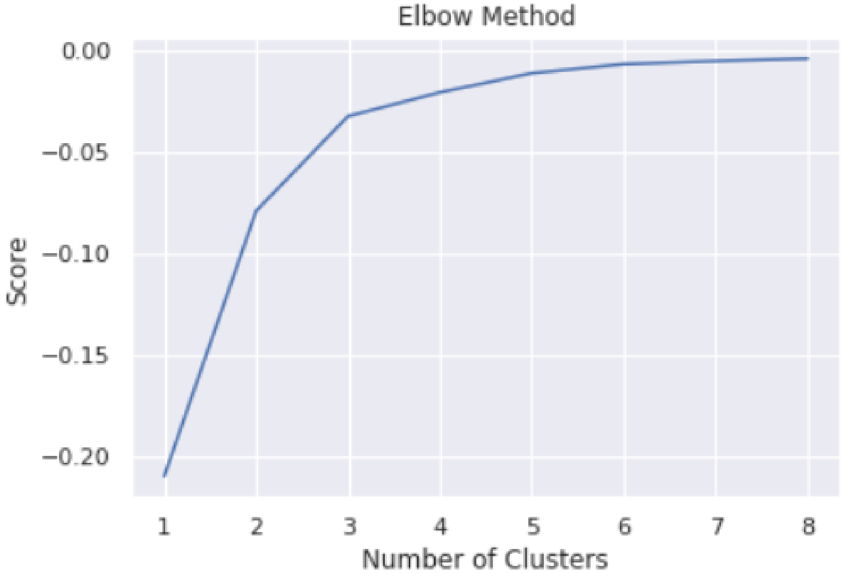

Figure 2: Elbow method to determine the optimal number of clusters.

Figure 3 is a visual representation of the four centroids, using a scatter plot. One can almost see the shape of Manhattan in this scatter plot. For better visualization, I decided to plot the four centroids on an actual map, with their respective neighborhoods.

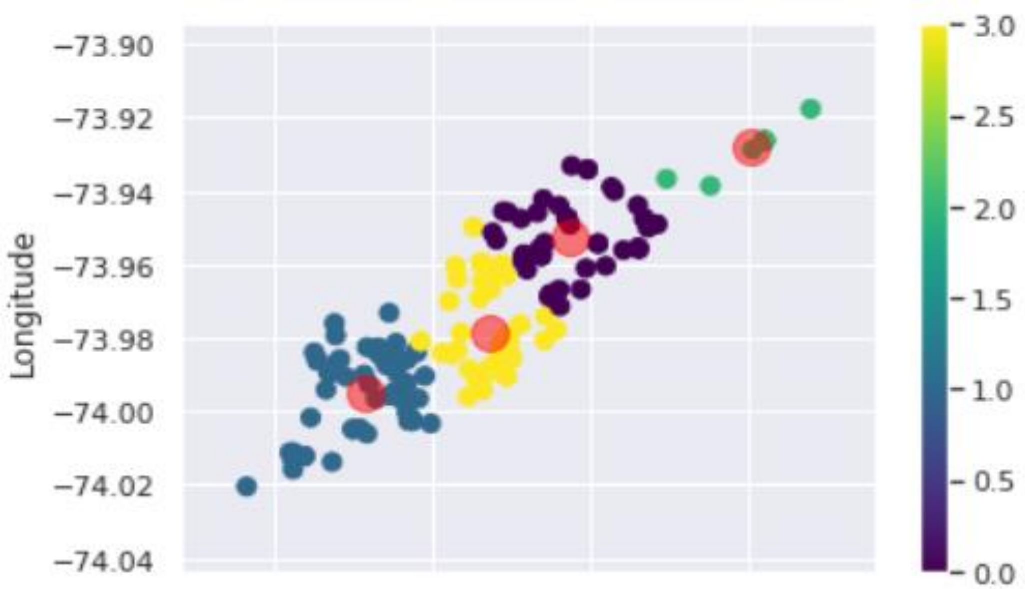

Figure 3: Scatter plot of the 4 centroids (in red), representing the center of clusters of high schools.

Figure 4 is a map of Manhattan with the four centroids placed as markers. The markers at in the following neighborhoods, from north to south, *Hudson Heights*, *East Harlem*, *Columbus Circle*, and *Greenwich Village*.

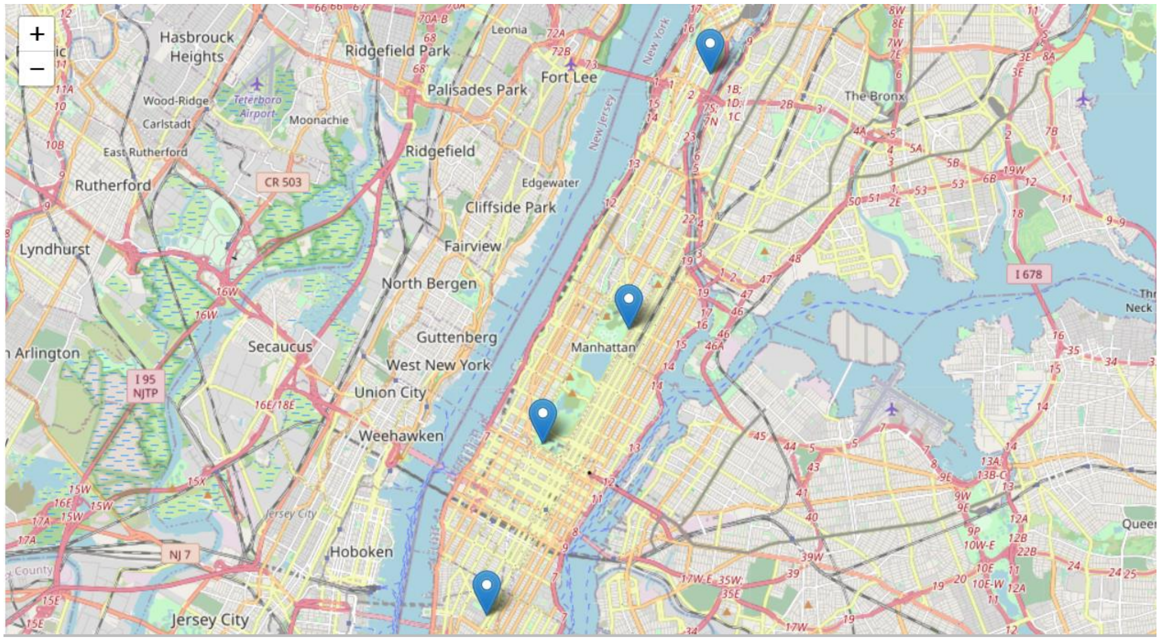

Figure 4: Four centroids, from north to south, *Hudson Heights*, *East Harlem*, *Columbus Circle*, and *Greenwich Village*, respectively.

Once, I had identified the centroids, I wanted to know how many physical therapy clinics already existed around the four locations identified. This will be another data point to determine where I was going to open my sports physical therapy clinic. For this, I used Foursquare's search feature. The keyword "*Physical Therapy"* was used as a search endpoint in the query, to find out how many clinics existed, within 600-meters radius, of the location coordinates. Table 1 below has the results of the number of clinics that exist around each of the four locations.

Additionally, I researched the rates for renting out a commercial space in these respective neighborhoods using the data from Metro Manhattan real estate brokerage firm.[6] Findings of this research are also included in Table 1 below, which includes the average price of renting out a commercial space in the respective neighborhoods.

Neighborhood	East Harlem	Hudson Heights	Greenwich Village	Columbus Circle
Number of PT clinics	3	6	27	37
Avg cost to lease	$50	$118	$80	$69*

Table 1 – Total number of physical therapy clinics and average cost to lease a commercial space in Q1 2019 for the four neighborhoods. * Average of Class A and Class B office space.

The information included in Table 1 above is much easier to explain, by a scatter plot in Figure 5.

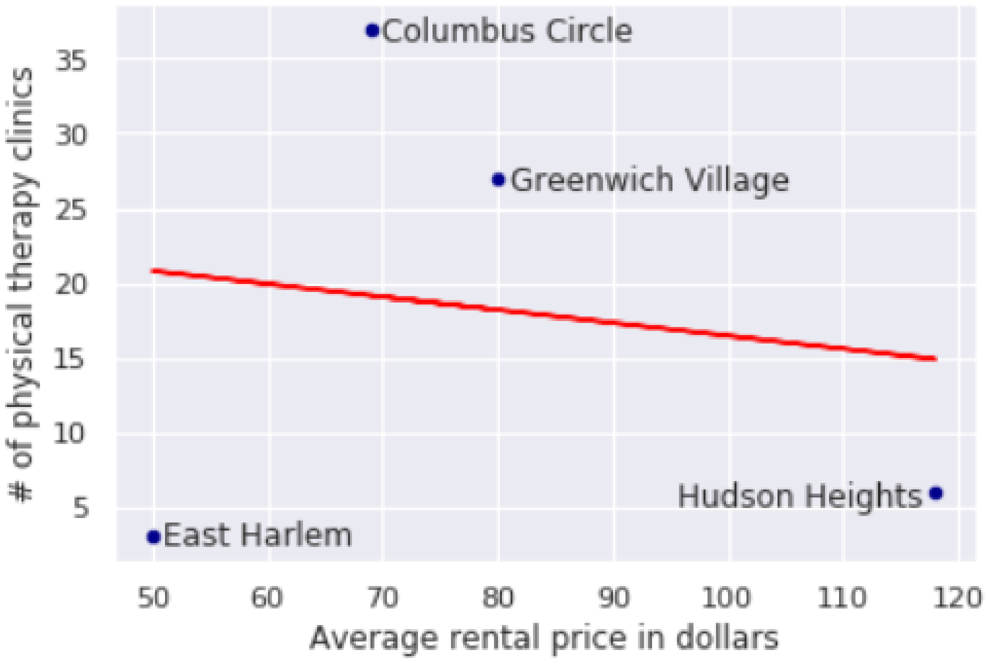

Figure 5: Scatter plot with regression line comparing the number of PT clinics with the average price of renting an office space in the four neighborhoods.

4. Results

Based on the methodology described above, there are four neighborhoods to pick from. I have information regarding the centrally located gps coordinates and also the number of physical therapy clinics around those coordinates. I also have average prices for renting a commercial space, for the respective neighborhoods. Based on all the findings, I believe it makes sense to open a sports physical therapy clinic in the East Harlem neighborhood. Within 600 meters of the East Harlem coordinates, there are only three physical therapy. Additionally, the average price to rent, is the lowest among all the other options. Basically, the East Harlem location has the least competition and the lowest rental prices.

5. Discussion

Prior to discussing the findings, it is important to note that this project has certain limitations. Additional important factors such as reimbursement rates, economic background, social issues, access to healthcare, crime rate, ease of parking, and ease of access, have not been considered. Figure 5 is a scatter plot of total number of physical therapy clinics and the average price to rent a commercial space. As observed, the regression line is sloping downwards slightly from left to right, which means that as the rental prices go up, the number of physical therapy clinics goes down. East Harlem is an outlier, where in the rental prices are very low, however there are very few physical therapy clinics there. Historically, East Harlem has many social issues such as high crime rate, high unemployment rate, homelessness, etc, which could explain the low rental rates and fewer businesses.[9]

If East Harlem was to be ignored, the slope of the regression line would be a lot steeper, in Figure 5. After East Harlem, my second choice would be Greenwich Village, as it seems to have a good balance of rental price and other physical therapy competitors.

6. Conclusion

This information would be helpful to anyone who want to start a sports physical therapy clinic in Manhattan. This methodology can be utilized to start various business across other parts of New York City or other cities. Of course, as I mentioned earlier, additional factors such as reimbursement rates, economic background, social issues, access to healthcare, crime rate, ease of parking, and ease of access should be considered in the decision-making process. Having completed this project, I am understanding how powerful the data science tools are for performing such research and also for presenting the data in a meaningful way, which is easy to understand.

References

1. https://www.nfhs.org/media/1020412/2018-19_participation_survey.pdf
2. https://wibx950.com/new-york-ranks-3rd-in-high-school-athletics-participation/
3. https://en.wikipedia.org/wiki/List_of_high_schools_in_New_York_City
4. https://www.python.org/
5. https://developer.foursquare.com/
6. https://www.metro-manhattan.com/neighborhoods/#
7. Bholowalia, Purnima. “EBK-Means: A Clustering Technique Based on Elbow Method and K-Means in WSN.” International Journal of Computer Applications 105, no. 9 (n.d.): 8.
8. Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1, no. 6 (2013): 90-95.