Battle of Neighborhoods –

How Health Conscious Is Your Neighborhood?

Devdatta Kanthe

Mar 30, 2020

Contents

1. Introduction and Problem Statement

As we see the world change around us, one of the noticeable changes is how much health-conscious people have become. They are conscious about working out, eating healthy and staying fit. One of the facilities within a neighborhood that can indicate whether the inhabitants are taking their health seriously is the number of Gyms available AND whether Gyms are among the top 10 recommended venues in that neighborhood.

We are going to find a statistical method to evaluate every neighborhood within borough of Manhattan in New York City and use folium maps to visualize the density. Based on the results, we would try to speculate if we can find top 5 health-conscious neighborhoods within the borough of Manhattan in New York city.

So, let's find out How Health Conscious is your neighborhood?

2. Data

Analysis for problem statement would be based on following factors:

- Number of Gyms, Health Centers in the Neighborhood
- Number of Gyms/Fitness Centers appearing in the Top 10 recommended venues
- Type of fitness center frequented in the recommended list

Data Sources

- **Google Geocoding API** for listing coordinates for Manhattan Neighborhood. - Google API is used for finding out coordinates based on addresses.
- **NYU Spatial Data Repository** for listing neighborhoods in Manhattan - NYU Data contains list of Neighborhoods with coordinates that can be used to find recommended venues around Manhattan.
- The venue details are scraped from **Foursquare** API - We would be using the explore API to figure out recommended venues based on inputs

2.1 Data Extraction and Cleanup

Importing NY Spatial Data Repository in the form of a JSON file provides us with geo-location of various neighborhoods within NY Boroughs. However, JSON would need to be read and information such as Borough, Neighborhood Coordinates and Names would need to be extracted from it.

As we are looking at data for Borough of Manhattan, we would filter the data for other boroughs from the data set. In the end, we would have a dataset of 40 neighborhoods for borough of Manhattan.

For these neighborhoods, we would fetch the recommended venues from FourSquare API. In all, we received 2,872 venues. A venue can be anything from a coffee shop to restaurant. These venues are limited to 100 per neighborhood and within 500 meters from the geo-location.

As we are interested in Gym/Fitness Centers, we would be looking at specific categories from FourSquare API data. Based on the documentation, for a master category of Gym/Fitness Center, there are 13 specific categories viz. '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', and 'Yoga Studio'.

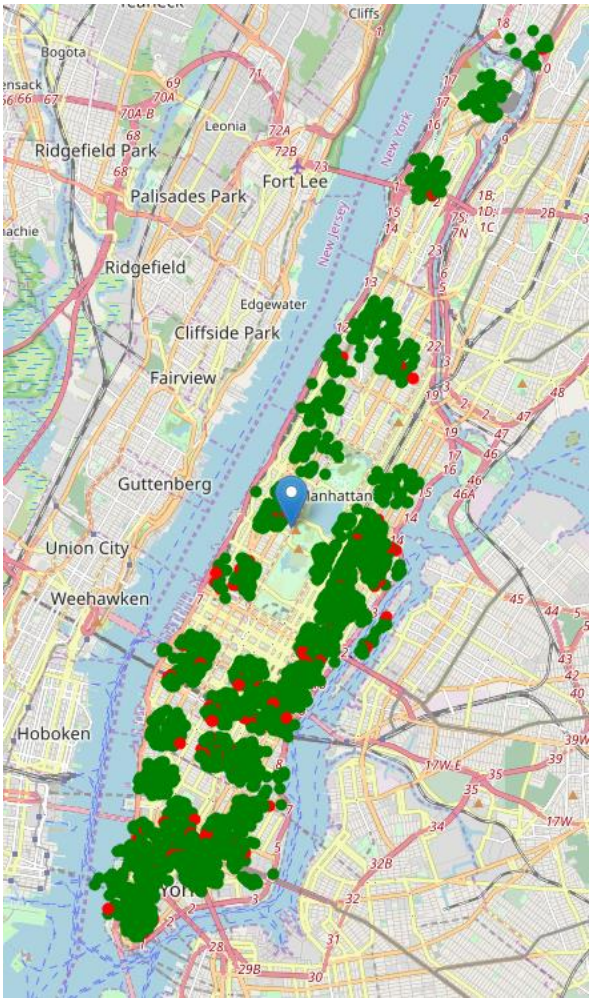
All venues with these categories would be marked specifically for further analysis. We received 198 Gym/Fitness Centers across Manhattan.

2.2 Data Exploration

There are 40 neighborhoods across Manhattan. The total recommended venues of all types are 2,872; out of which 198 are Gym/Fitness Centers of all kinds. At an average, there are close to 5 Gym/Fitness Centers per neighborhood.

We would be analyzing the data further to determine the exact distribution across neighborhoods. At this point, it would be safe to assume, neighborhoods with more than 4 Gym/Fitness Centers can be considered health conscious.

For understanding, we plotted all venues on Manhattan map using Folium. This provided an insight on how distributed the venues are across the borough. We noticed that bulk of the venues are aggregating towards downtown Manhattan; with neighborhoods around midtown having higher frequency of Gyms. In the map, the green markers are all other venues, red markers are Gyms/Fitness Centers.



Our search for health-conscious neighborhood would mostly be around the Midtown area where the density of red markers seems to be more.

3. Methodology

To determine the health-conscious neighborhoods, we would need to determine the density of Gym/Fitness Centers within that neighborhood. We would then differentiate neighborhoods with lower density of Gym/Fitness Centers.

As a first step, we found and stored the **neighborhoods and their coordinates** based on data available in the NYU dataset. We then used the Foursquare API to explore **all venues** within each neighborhood. While doing that, we added an **identifier for Gyms and Health Food centers** within a neighborhood based on venue categories. To visualize the results, we plotted all venues on a map of Manhattan.

As a second step, we would find out top 10 venues in each neighborhood and determine if Gym/Fitness centers are part of the top 10 list. Along with that, we would use **heatmap** to determine the density of Gyms across Manhattan. We would also need to establish the **ratio of Gyms to total venues** across neighborhood. It may happen that a neighborhood has a greater number of Gyms but the total venue count in that neighborhood could be higher resulting in lower ratio.

We would also try to use **K-Means clustering as the unsupervised machine learning algorithm on neighborhoods, based on Gyms/Fitness Centers** and analyze each cluster to understand classification. K-Means would provide us with insights on similarities between neighborhoods in terms of the popular Gym/Fitness Center types.

Lastly, we would determine based on the above metrics, the neighborhoods that can be classified as health conscious.

4. Analysis

We would first find out the top 10 venues per neighborhood based on data returned by FourSquare API. This API would get neighborhood coordinates to provide 100 venues within a 500 meters radius around it. After ranking the neighborhoods, we get a data frame with top 10 venues.

Initially, we are working with 40 neighborhoods with 344 total venue categories for analysis. The below data frame is returned as a sample

Out[11]:

	Neighborhood	Most Common Venue 1	Most Common Venue 2	Most Common Venue 3	Most Common Venue 4	Most Common Venue 5	Most Common Venue 6	Most Common Venue 7	Most Common Venue 8	Most Common Venue 9	Most Common Venue 10
0	Battery Park City	Coffee Shop	Park	Hotel	Wine Shop	Boat or Ferry	Memorial Site	Shopping Mall	Gym	Italian Restaurant	Food Court
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Yoga Studio	Japanese Restaurant	Gym / Fitness Center	Gym	French Restaurant	Cosmetics Shop	Bookstore
2	Central Harlem	African Restaurant	Bar	French Restaurant	American Restaurant	Seafood Restaurant	Chinese Restaurant	Caribbean Restaurant	Spa	Dessert Shop	Beer Bar
3	Chelsea	Coffee Shop	Bakery	Italian Restaurant	American Restaurant	Ice Cream Shop	Hotel	Wine Shop	Breakfast Spot	Tapas Restaurant	Cycle Studio
4	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Spa	Bakery	Hotpot Restaurant	Optical Shop	Vietnamese Restaurant	Salon / Barbershop	Dessert Shop

Now, within this data set, we would need to find neighborhoods that have Gym/Fitness Center categories in top 10 venues. We perform this analysis by searching for Gym Categories (see section 2.1) within any columns of the data frame. We find that there are multiple venue categories appearing, but 3 Gym Categories are popular – Gym/Fitness Center, Gym and Yoga Studio.

```
# of Neighborhoods that have Gym / Fitness Center as top 10 place of interest = 12
# of Neighborhoods that have Boxing Gym as top 10 place of interest = 0
# of Neighborhoods that have Climbing Gym as top 10 place of interest = 0
# of Neighborhoods that have Cycle Studio as top 10 place of interest = 2
# of Neighborhoods that have Gym Pool as top 10 place of interest = 0
# of Neighborhoods that have Gymnastics Gym as top 10 place of interest = 0
# of Neighborhoods that have Gym as top 10 place of interest = 20
# of Neighborhoods that have Martial Arts Dojo as top 10 place of interest = 0
# of Neighborhoods that have Outdoor Gym as top 10 place of interest = 0
# of Neighborhoods that have Pilates Studio as top 10 place of interest = 0
# of Neighborhoods that have Track as top 10 place of interest = 0
# of Neighborhoods that have Weight Loss Center as top 10 place of interest = 0
# of Neighborhoods that have Yoga Studio as top 10 place of interest = 7
```

Based on the above data, we can derive that the popular Gym types are Gym/Fitness Centers (a combined **32** places of interest), Yoga Studio (**7** Venues) and Cycle Studio (albeit with just **2** Venues!).

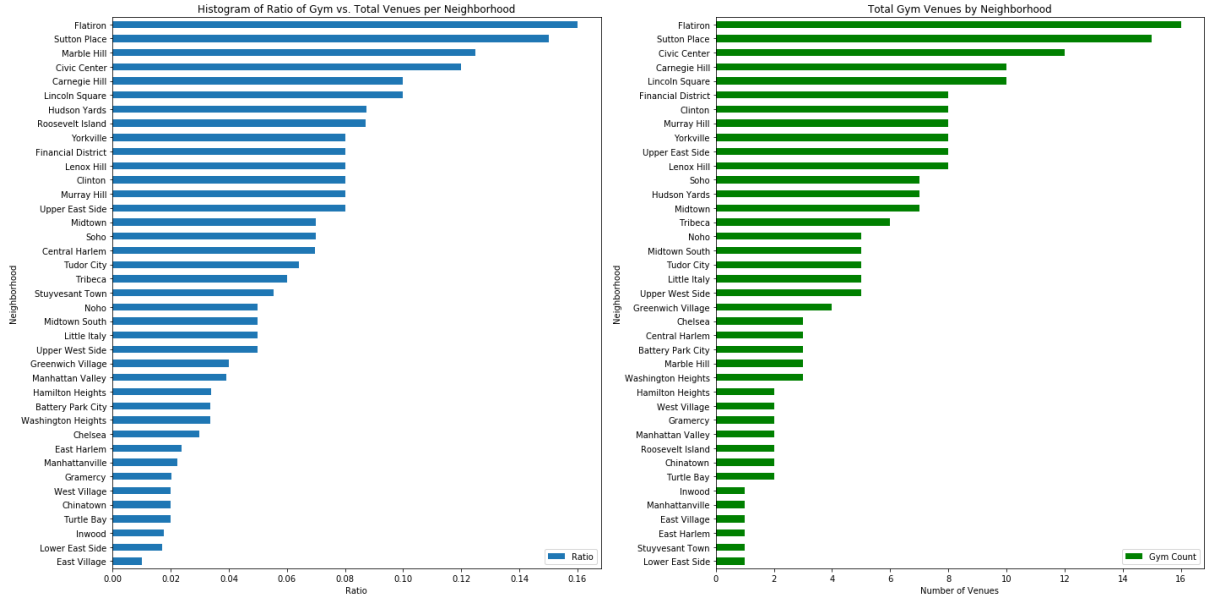
However, filtering the data even further, we find that there are only a few neighborhoods with Gyms as top 3 venues.

- Civic Center
- Clinton
- **Flatiron**
- Marble Hill
- Roosevelt Island
- **Sutton Place**
- Yorkville

Flatiron and **Sutton Place** are highlighted because Gyms are their **topmost common places (Rank = 1)**. These neighborhoods have a higher priority for Gyms than other category types. **Can we claim them to be Manhattan's most Health Conscious Neighborhoods? Probably.**

However, we still have some scope to understand the proportion of venues in these areas and the total fitness centers available in the neighborhoods. So, we would keep digging deeper into the data!

Now, we would find the ratio of Gyms vs. total venues. This would help us determine frequency of gyms and plot on a heatmap to visualize. The ratio is important because a neighborhood may have a greater number of gyms but the ratio to total may be lesser. The comparison is evident from the charts below where we have plotted ratio vs. number of gyms against neighborhood.



Again, we have similar results as obtained earlier. Top 2 Neighborhoods with highest ration of Gyms to Total Venues are Flatiron and Sutton Place.

Interestingly, in the above charts, we would notice that although certain neighborhoods have a **higher number of Gyms, the overall frequency per total venues is lower.** Example, Civic Center and Carnegie Hill. Conversely, Marble Hill and Hudson Yards, scoring lower on overall Gym venues, scores higher on

the Ratio scale. Especially, Marble Hill that is very low on the number of Gyms but in top 3 based on ratio.

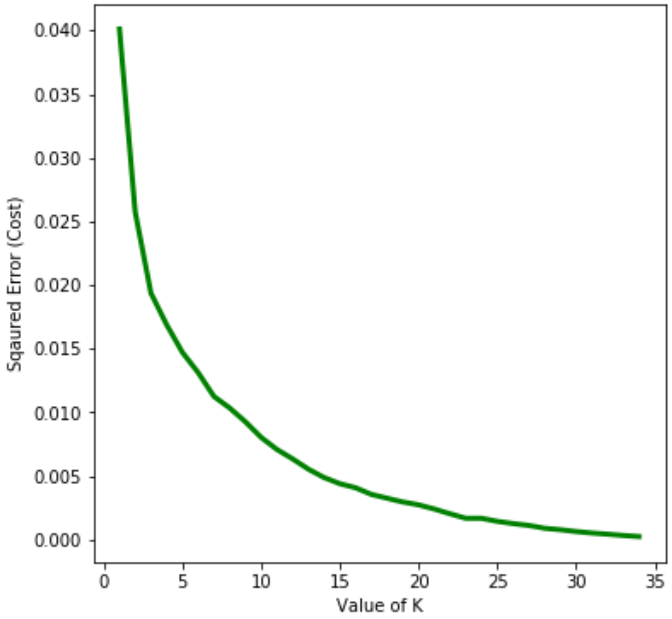
Plotting the Gym venues against Manhattan using heatmap, provides similar results:

Based on the above heatmap, we can observe a few Neighborhoods have a higher density of Gyms. Prominent, as observed before, are Sutton Place, Flatiron, Civic Center, Carnegie Hill, Financial District.

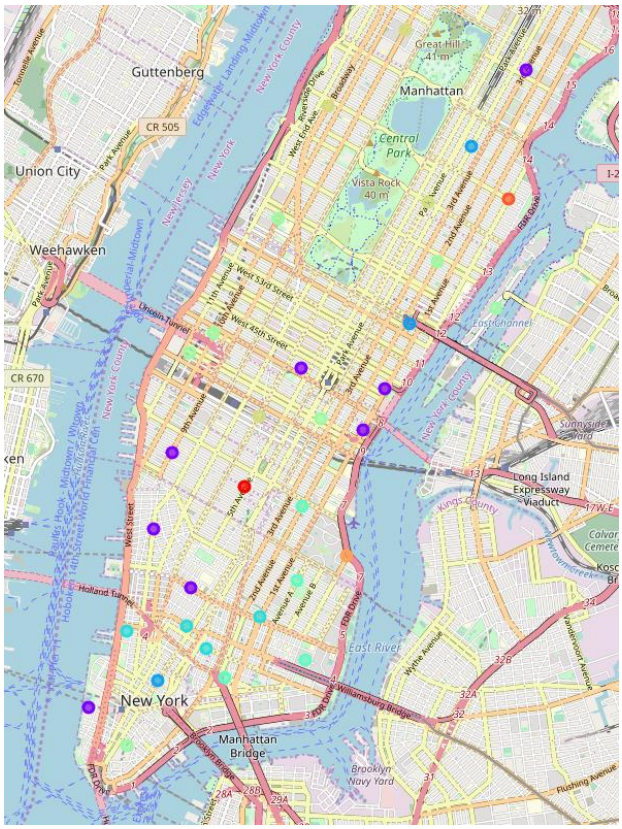
Evidently, places such as Marble Hill, show lower concentration on the heat map due to lesser venues. However, ratio of Gyms vs. All venues is higher for Marble Hill.

4.1 Clustering Neighborhoods

To understand how every neighborhood is associated to each other based on Gym Venues, we created clusters. For this, we used the K-Means clustering algorithm. Initially, we figured the optimum number of clusters to 10 based on the elbow method.



With K=10, we categorized every neighborhood and plotted the data against Manhattan map to get the following results:



After analyzing all 10 clusters, it is observed that neighborhoods are clustered based on top 3 venue categories which in our case were Gym Categories. Clusters 1, 5 and 6 were largest clusters by number

and had Gym/Fitness Center, Gym, Yoga Studio and Cycle Studio in top 3 categories. This is in line with the top 10 analysis performed earlier, where these very categories were among the most popular.

Snapshot of cluster 6 is provided below for reference

Out[36]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Cluster Labels	Most Common Venue 1	Most Common Venue 2	Most Common Venue 3	Most Common Venue 4	Most Common Venue 5	Most Common Venue 6	Most Common Venue 7	Most Common Venue 8	Most Common Venue 9	Most Common Venue 10
374	Central Harlem	40.815976	-73.943211	6	Gym / Fitness Center	Gym	Cycle Studio	Yoga Studio	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool	Climbing Gym
659	Lenox Hill	40.768113	-73.958860	6	Gym / Fitness Center	Gym	Cycle Studio	Yoga Studio	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool	Climbing Gym
759	Roosevelt Island	40.762160	-73.949168	6	Gym / Fitness Center	Gym	Yoga Studio	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool	Cycle Studio	Climbing Gym
882	Lincoln Square	40.773529	-73.985338	6	Gym / Fitness Center	Gym	Cycle Studio	Yoga Studio	Climbing Gym	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool
982	Clinton	40.759101	-73.996119	6	Gym / Fitness Center	Gym	Yoga Studio	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool	Cycle Studio	Climbing Gym
1182	Murray Hill	40.748303	-73.978332	6	Gym / Fitness Center	Gym	Martial Arts Dojo	Boxing Gym	Yoga Studio	Weight Loss Center	Pilates Studio	Gymnastics Gym	Gym Pool	Cycle Studio
2321	Financial District	40.707107	-74.010665	6	Gym	Gym / Fitness Center	Cycle Studio	Yoga Studio	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool	Climbing Gym
3217	Hudson Yards	40.756658	-74.000111	6	Gym / Fitness Center	Gym	Cycle Studio	Yoga Studio	Weight Loss Center	Pilates Studio	Martial Arts Dojo	Gymnastics Gym	Gym Pool	Climbing Gym

5. Discussion

We started with a simple question at hand - How health conscious is your neighborhood? After much statistical analysis, we have come to multiple conclusions as under:

- Based on **popularity of a Gym/Fitness Center venue in a neighborhood**, the following 7 neighborhoods had Gyms in their top 3 venues:
    - Civic Center
    - Clinton
    - **Flatiron**
    - Marble Hill
    - Roosevelt Island
    - **Sutton Place**
    - Yorkville
- **Flatiron** and **Sutton Place** are highlighted because Gyms are their **topmost common places (Rank = 1)**. These neighborhoods have a higher priority for Gyms than other category types.
- Based on **Ratio of Gyms to total number of venues**, the following neighborhoods had the highest ratios:
    - Flatiron
    - Sutton Place
    - Marble Hill
    - Civic Center

    - Carnegie Hill
    - Lincoln Square
    - Hudson Yards
- In terms of popular Gym/Fitness Center Types, the following came up top:
    - Gym/Fitness Centers (a combined **32** places of interest)
    - Yoga Studio (**7** Venues)
- Having said that, there are other Gym/Fitness Centers as well but statistically, they were not among the recommended venues.
- While analyzing the clusters of neighborhoods, we can see the clusters have most popular venue as one of the above-mentioned types.

6. Conclusion

If you are part of the **Flatiron, Sutton Place** then most definitely you are part of a health-conscious neighborhood in Manhattan, NY. However, in addition to these two, there are other neighborhoods that qualify as health-conscious basis other criteria listed in the Discussion section above.

All in all, out of the 40 neighborhoods in Manhattan, about 10 have qualified as health-conscious in our study. Overall, **25% of all neighborhoods in Manhattan can be considered health conscious based on Gym/Fitness Center related data available as of today.**

This analysis is essential so as to help plan neighborhoods, keeping in mind, the health requirements of inhabitants. It is essential for people to get basic needs within the area but fitness related facilities are becoming essential day-by-day. Due to the health hazards, we face, everyday due to our lifestyles, staying fit has become a requirement, a need and an expectation. After all, a **health conscious neighborhood** has the potential to become a **healthy neighborhood**.

7. Assumptions

- Factors such as day-time population of neighborhood along with other related demographics are excluded from the analysis. These are critical for a holistic analysis.
- It is assumed that the facilities available are utilized by people who are living in the neighborhood. It is highly likely that Manhattan being a business district, people coming in to work, may be using the facilities as well.
- By definition, healthy and health conscious are not always synonymous.

- FourSquare API data is assumed to be complete. There may be other sources providing extensive and expansive data.

8. Acknowledgement

This project would not have been completed without the express support of Open Source community, Coursera and IBM professionals. Millions of help articles available on multiple forums help get answers to complex questions. Here's thanking every contributor, tutor, coder and tester who contributes to the community to help rookies grow. Thank you.

The entire world is affected by COVID-19 outbreak. New York city is also among the worst affected in the world. This analysis is done on Manhattan, NY data. I hope the crisis subsides soon and the community places that are shut due to lock-down are opened up again. I would also hope that we see more and more health conscious neighborhoods across the world for a healthy planet and humanity. Thank you to everyone.