Capstone Project - The Battle of Neighbourhoods
Gym Opening in New York City

Ali Manzouri
April 29, 2020

1. Introduction

1. Background

New York City was home to nearly 8.4 million people in 2018, accounting for over 40% of the population of New York State. Due to New York growing in population each year with a finite amount of space, the state of New York is offering incentives to small fitness owners to open gyms/fitness areas in the New York area to get the population to remain fit. The challenge is to ensure the need for gyms and thus the reason all gym franchises are looking to do an in-depth study of the New York area and determine the best possible solution/area to open a gym.

2. Problem

A new gym franchise is looking to open a gym in one of New York's neighbourhoods. The franchise does not understand the area and the availability of gyms in each neighbourhood and requires an investigation to take place in order to determine the best place for the franchise to open a new gym based on the decision metrics below.

3. Decision Metrics

The following decision metrics are requested in order to make an informed decision by the gym franchise to where the gym should be located.

- Density of people for each Borough
- Number of Neighbourhoods in each Borough
- Number of gyms in each Borough
- 4 Gyms in the Neighbourhood with the best density metric per gym
- Cluster Gyms in Neighbourhood with the best density metric per gym

2. Data Requirements and sources

2.1. Data Sources

For the investigation, the following data sources will be used:

1. Wikipedia to obtain density of each Borough in New York city.

a. Source: https://en.wikipedia.org/wiki/New_York_City

b. Description: New York Boroughs and the density of each Borough in the New York area.

2. New York City data that contains list Boroughs, Neighbourhoods along with their latitude and longitude.

a. Source: https://cocl.us/new_york_dataset

b. This contains the data as mentioned above and will be used for investigating the Borough and Neighbourhoods using Foursquare API.

3. Gyms in each neighbourhood of New York city.

a. Source: Foursquare API

b. The API will return all known gyms in each Borough and Neighbourhood.

2.2. Data cleaning

During the data extraction process from the various sources, it is important to validate and verify clean data to work with throughout the analysis. Each data source required validations and visual confirmation to whether there are discrepancies in the data we received/sourced.
It was vital that these data feeds had validation to ensure the accuracy of the results obtained. Each feed/data source went through its own unique data extractions and validation based on the source of the data. The following was done to each data source to ensure a robust approach in terms of the analysis and the comparing of data sources.

2.2.1 Wikipedia
The data source to obtain density, populations and other key metrics for New York city came from a Wikipedia page. The page required extracting as well as validation on the extraction of the data as seen in the code. It was of utmost importance to ensure the correct extraction criteria occurred to ensure the validity of the data. In this process it was noted that the naming convention of the Boroughs were not the same as the other feeds and this required adaptation to ensure linking to other data sources was seamless and accurate.

2.2.2. New York City Data
This data source was important to obtaining the data for borough and neighbourhoods in the New York city area. We required this data to validate correct geo locations could be sources for the foursquare API. This data source was robust and required minimal changes to align with the other datasets.

2.2.3 Foursquare data for gyms
This data set required a lot of cleaning in order to ensure only gym information was extracted for the analysis in New York City. The cleaning of the data required all other “noise” to be dropped and that only valid data of gyms was required. This process of gym validation required the checks to ensure the gyms we in the correct mapping location based on the New York city Boroughs and ensuring the data received was recent and not old.

3. Methodology and explanation of Data analysis.

In order to answer the various decision metric as stated in 1.3 above, it was vital to ensure each decision metric was fully analysed and a decision could be made. Thus, it is reporting below is the analysis to obtaining each result as well as the data analysis done for each.

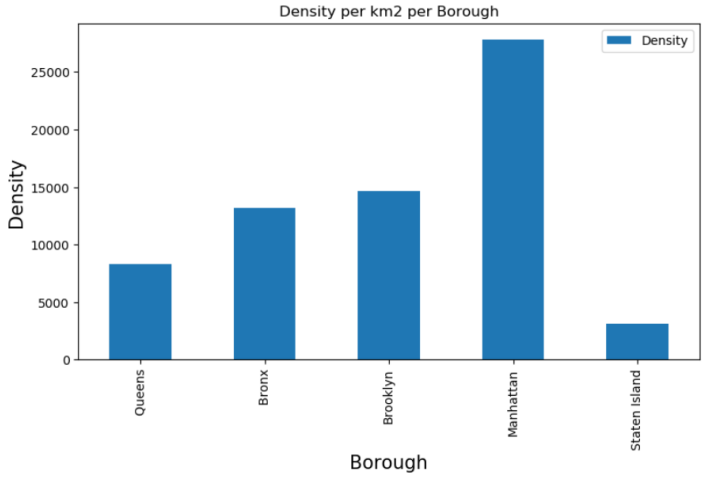
3.1. Density of people for each Borough

The density of each borough in New York city was achieved through the extraction of data from the Wikipedia page as stated in the data source section 2.1 above. With the various scrapping of data from the source the following data could be extracted (Table 1).

	Borough	Density
0	Bronx	13231
1	Brooklyn	14649
2	Manhattan	27826
3	Queens	8354
4	Staten Island	3132

Table 1: Density per Borough in New York City

The density of each borough helps to visualise the density of people per borough and later will be used to ensure the density of people per gym that is currently in each area. With targeting the area with the best density will allow for more potential customers. The graph below helps to visually see the major difference in the boroughs based on density.



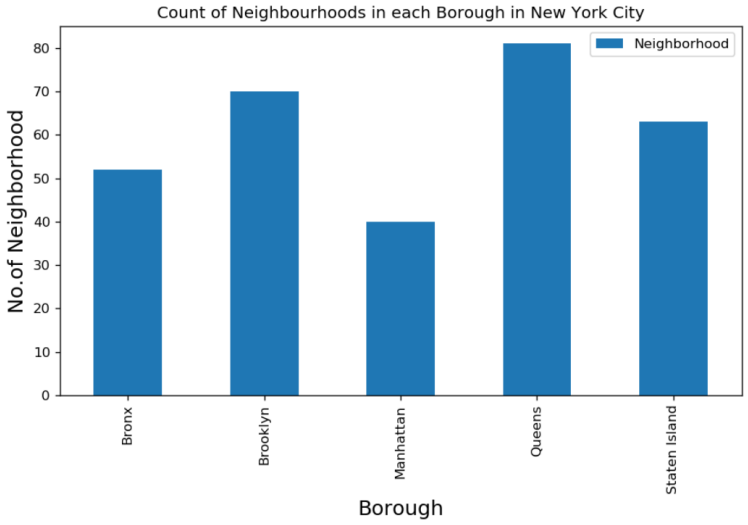
Graph 1: Bar graph of density of people in each borough of New York City.

With graph 1 above, it is easily noticeable that Manhattan has the most people per km2 of any borough. This shows population density that can be used in further analysis of population density per gym in each borough.

3.2. Number of Neighbourhoods in each Borough

The next important metric was to ensure that number of neighbourhoods in each borough. This decision metric aids in the density to determine the spread of the population density. This decision metric aids the ability to open a gym with few neighbourhoods to ensure a greater buy in when opening a new gym in an area.

The following was obtained through the New York city data as seen in section 2.2.2 above. This was used to graph the neighbourhoods in each borough as seen in table 2 below.

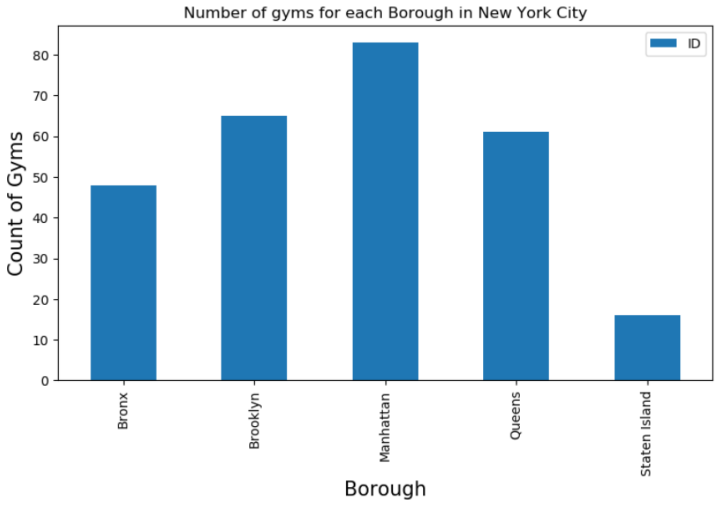


Graph 2: Count of Neighbourhoods in each borough.

As seen above, the Manhattan borough has the fewest number of neighbourhoods whilst potentially having the highest population density as seen in section 3.1. Both allow for valuable insight in decision making further on.

3.3. Number of gyms in each Borough

The next metric was to validate and source the number of gyms per borough. This will give valuable insight of competitors and the location of the competitors. When extracting data from Foursquare the follow number of gyms were noted in each area.

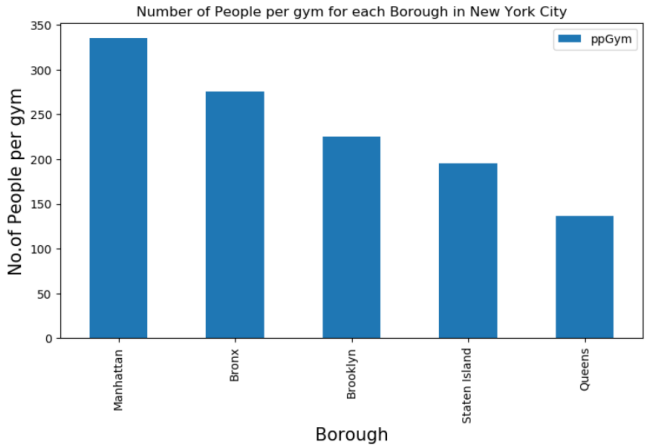


Graph 3: Count of gyms per borough

As seen above, there is major difference in the count of gyms in each borough. This helps us to see a potential gap in the market of the other boroughs whilst we use this data to determine the borough with the maximum density of people per gym. This can also be valuable insight in determining a rating factor for decision making for further analysis on best location.

3.4. Gyms in the Neighbourhood with the best density metric per gym

With all the above complete, a combination of each result was needed to identify a local best location. With merging data and evaluation, the boroughs per density and gym count. It was discovered that Manhattan still produced the highest density of people per gym. This statistic goes against the data represented above if seen in isolation. Thus, was of utmost importance that the combination as seen below in graph 4 was designed to evaluate each borough in the same manner. This allows for opportunity not to be weighed heavier toward one borough creating a bias in the data analysis.

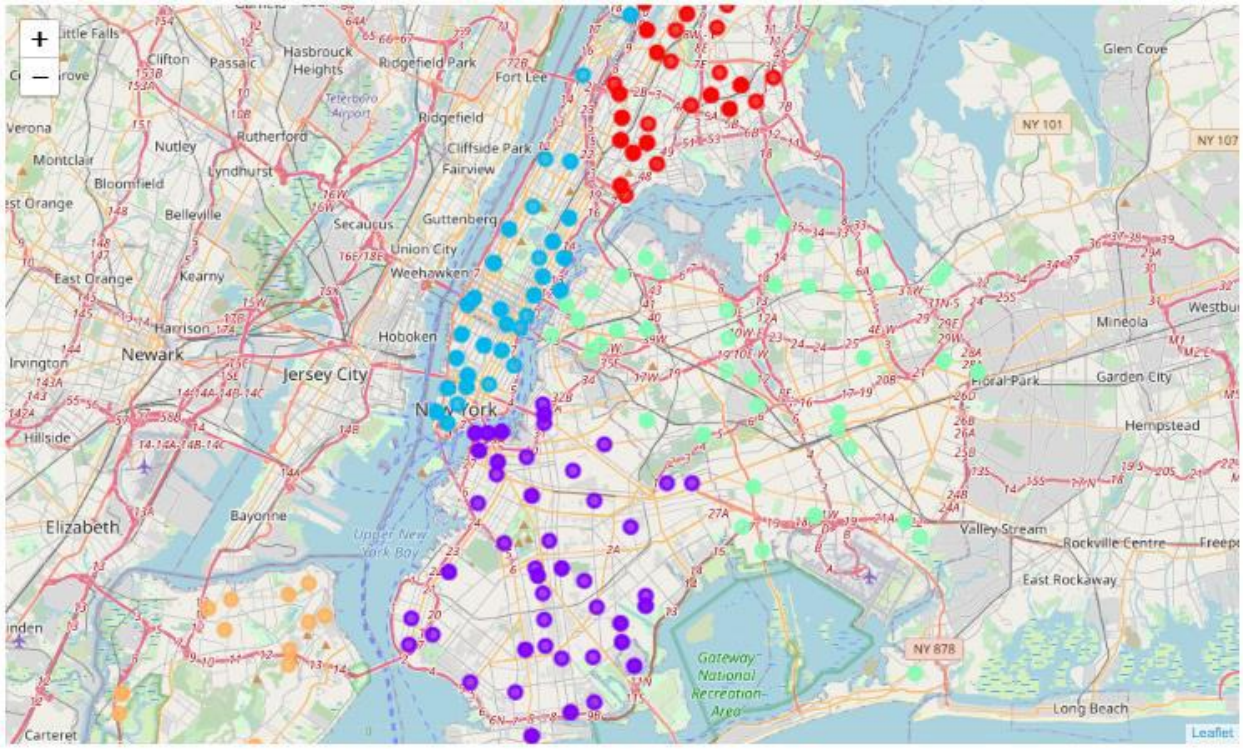


Graph 4: Number of people per gym in each borough

It was also important to see the spread of gyms across the entire network of gyms to check fair distribution and not over population of the gyms in a localised area. The following map displays the density of gyms in each area mapped.

Map 1: Density of gyms across New York to evaluate the overcrowding of a localised area.

With the above seen, it was more important to visualise the borough density individually by clustering them based on their borough. This allowed for further visual investigation of the borough's gym density and possible overcrowding of gyms.

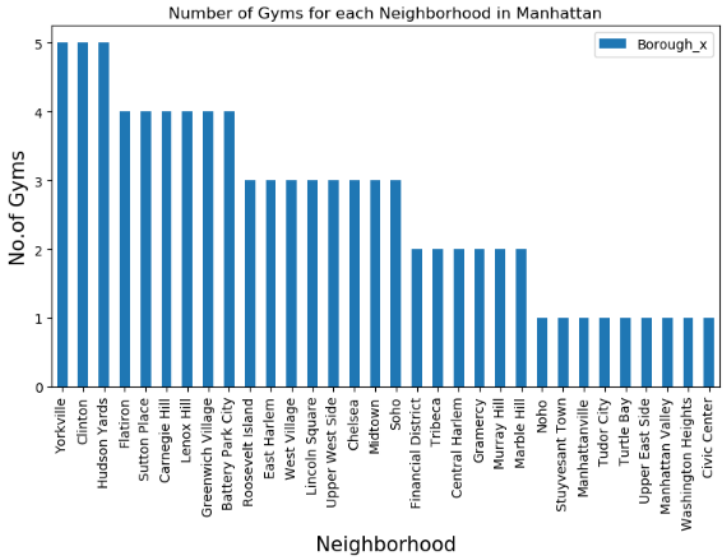


Map 2: Density of gyms across New York to evaluate the overcrowding of a localised area clustered into boroughs

As seen above, the special density of the gyms illustrated in map 2 above, show that all the gyms are evenly spread. This allows for a normalisation of the data to determine the best borough and thus looking into neighbourhood potential further.

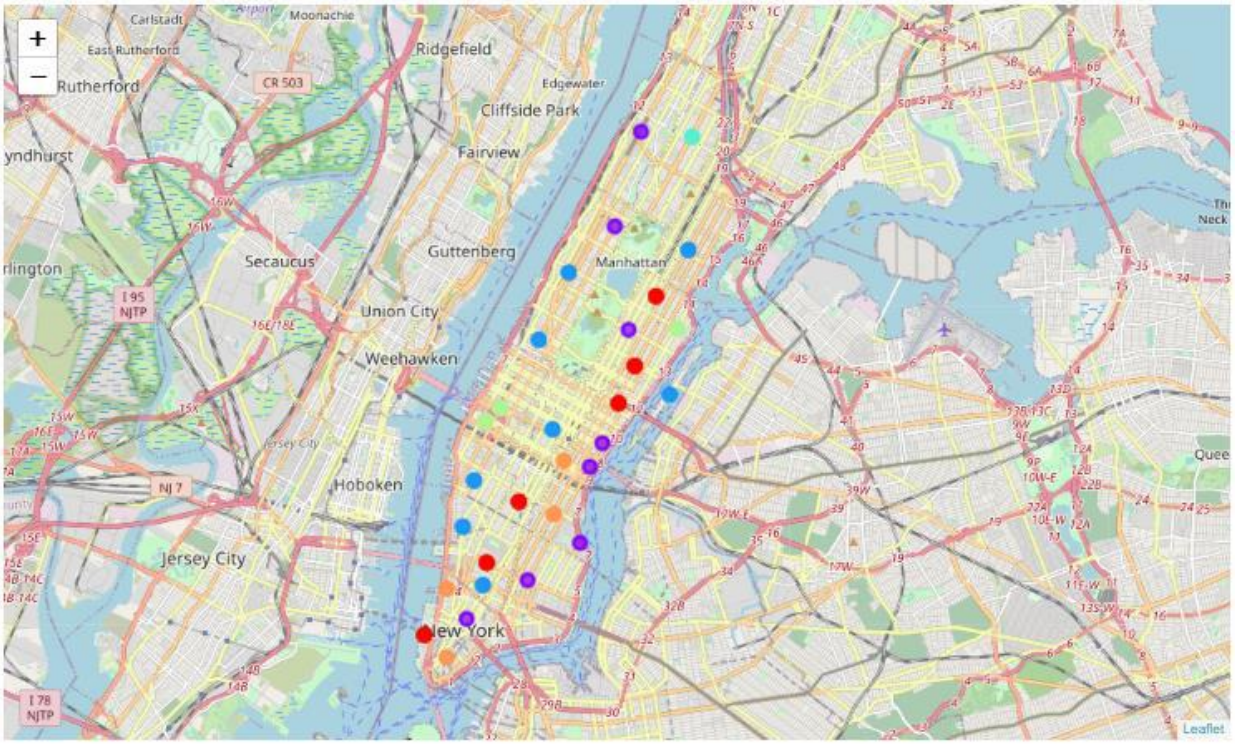
3.5. Cluster Gyms in Neighbourhood with the best density metric per gym

With the above taken into consideration, it is noted that the best potential for the gym franchise was to evaluate the neighbourhoods in Manhattan based on population density per gym as well as gym spread in the Manhattan area.
The count of gyms in each neighbourhood is required to ensure a full picture of the neighbourhood competitors is evaluated. With the location data of each gym, it is possible to normalise the data to view the count of gyms per neighbourhood.



Graph 5: Gyms per neighbourhood in Manhattan

With the above information (Graph 5), the data was used to determine gym spread and clustering of the neighbourhoods to evaluate each neighbourhood on the merit of gym capacity and population density. The top gym ratings were used to evaluate a visual representation of the gyms in the Manhattan area. The following was observer in map 3 below.



Map 3: Best gym ratings in Manhattan clustered for density spread of the gyms.

With confirmation of spread across the neighbourhoods being equal, it can be determined that evaluating the number of gyms in each neighbourhood would be a fair analysis as this is a normalised data set.

4. Results Discussion

With the above decision metrics being consider the following was observed in the data. The population density of each borough is majorly skewed with some boroughs having a much higher population density that other. The gyms per a borough was vital for to determine the possibility of the customers joining the gym. The ratings of each gym ensured a robust approach to determining the impact of opening a gym when competing with some of the best gyms in the area.
Finally, the spread of the gyms aided in deciding if there is a bias in the exact location of the gyms in each borough to ensure the robustness of the analysis driving the decision is not skewed with overpopulated areas of gym with the best rating.
The neighbourhoods in Manhattan also allowed for the full understanding of gyms in each neighbourhood whilst ensure again the spread of the best-known gyms across the neighbourhood was not skewed to overcrowding.

5. Conclusion

With all the above taken into consideration, it can be sent that Manhattan has the best chance of making money when opening a gym. This is confirmed with the density of people per gym in the Manhattan area. The spread of gyms also ensured that there is not a skew/ bias in the data and that the gyms with the best rating are well spread and determination of the