Opening a Gym where Best seems Fit

Chris Hamian

1 Introduction

1.1 Background:

New York City is such an illustrious and complex city that provides so much opportunity for wealth and success because of the influx of people, traffic, and businesses that it contains. A big part of success is expansion, but obviously location in a city as big as New York City is primitive and must be nailed to a tee in order to ensure success. (NOTE: I use “location”, “area”, and “zip code” interchangeably throughout my report)

1.2 Problem:

A Client is looking to open up a gym/fitness center in Manhattan but he has never lived in the city and doesn’t know too much about the specifics of the city so he’s not sure about the proper region to open his gym that will be most optimal for the gym’s success.

1.3 Approach:

The key is to look at specific key information surrounding every neighborhood. The initial step is to gather sufficient information on each region such as total income, total population, and the total number of existing gyms already in that region.

2 Data Sources and Extraction:

My first source contains the bulk of my dataset includes specific information pertaining to separate areas in Manhattan. Every row indicates a different zip code in Manhattan and is followed by key information such as city, latitude/longitude coordinates, population, average household income, and national rank. From basic web scraping practices, I was able to extract the proper coordinates, income and population for each zip code in Manhattan. I don’t have the full (purchased) version of the dataset so some zip codes don’t have all of the information. To account for that, I decided to work with only the zip codes that contain all the necessary values and information so that’s about 44 out of the 64 total zip codes from the data source. The link to the data source is here.

The second source is the Foursquare website. I’ll be utilizing the Foursquare API in order to get detailed information on all the venues in each zip code and then I can narrow it down to just the different gyms and fitness centers in the area. This information will help us narrow down the market availability in Manhattan. Since I have all of the latitude and longitude coordinates for each location, I simply created a function that extracts all of the venues surrounding a

particular area and ran that through every location in the dataset and then it returned a full list of every venue and its location zip code, venue category, etc. Then I simplified all of the venues by narrowing them down to just gyms. Once it's simplified, we just add up the number of gyms per zip code and append that to our original data frame

3 Methodology

3.1 Data Compression:

Once we've gathered all of the data and simplified all of the variables, we have a data frame that consists of a zip code, population, average household income, and gyms in that area. To begin the data analysis, we can narrow down our locations even more. Ideally, we wouldn't want to open a gym in an area that:

1. Doesn't have a high population
2. Doesn't produce enough income

The reason why we don't want to open a gym in an area with a low population is simple, there's not a big enough market. Gyms are very simple; they have equipment, weights, locker rooms, music, etc. There's not too many variables that separate gyms (unless specialty gyms), but in our case we have to remember that people place a high value on convenience when it comes to signing up for a gym. Most people in Manhattan work a 9-5 job so we have to be in a location where people can come both before and after work because time is a very valued entity to everyone so if we're in a location that's saving people time, then we are of value to that person as opposed to a different gym.

As well, we definitely don't want to open a gym in a low-income location. At the end of the day, the gym is a luxury; it's not something necessary to live so people will only pay for the gym if they can afford it. In some of those low-income areas, people have to worry more about keeping a roof over their heads and keeping food on the table rather than making sure they reach 10,000 steps for the day. From this analysis, I decided to calculate the lower quartile for both the total populations and the average incomes. The resulting values were 15,887 people and $28,692. I then narrowed down the dataset to locations that have a population higher than 15,887 AND an average household income higher than $28,692. Now, our dataset just compressed from 44 locations to 23 applicable and more viable areas.

3.2 Analysis:

3.2.1 Segmentation:

For further analysis, we'll now focus on the number of gyms in each location and there's definitely a lot of variance in this category. This data has a maximum of 8 and a minimum of 0 (current gyms in an area) so I decided to segment the locations into three different groups. I added an extra column to the data frame called "Gym Frequency" and this basically labels each location as low, med, or high – pertaining to that location's number of gyms. I thought three

groups would be best because there's only 23 zip codes we're looking at so each group won't be visibly hard to analyze and determine the location's with the best characteristics. I chose to use 3 (gyms) as the condition value for labeling each zip code. I thought 3 would make the most sense because 3 is the median and mode of the current number of gym per zip code so I thought it would segment the dataset well.

Once we divide the data, we now have three separate datasets so that now we will continue further analysis on each dataset separately.

Data Frame	Number of Zip Codes in DF
Low Frequency Gyms	7
Med Frequency Gyms	5
High Frequency Gyms	9

Figure 1 – Table of Total Number of Zip Codes per Data Frame

The idea is to draw the top candidates from one dataset and then compare those with the top locations from the other two datasets. We'll find the best candidates from each dataset by utilizing a scatter plot. I chose a scatter plot because I thought that a scatter plot would be easiest to understand how points (locations) compare to each other.

3.2.2 Low Gym Frequency:

We'll start by looking at the scatter plot of the low frequency dataset which is the dataset containing zip codes that have less than 3 gyms in that area.

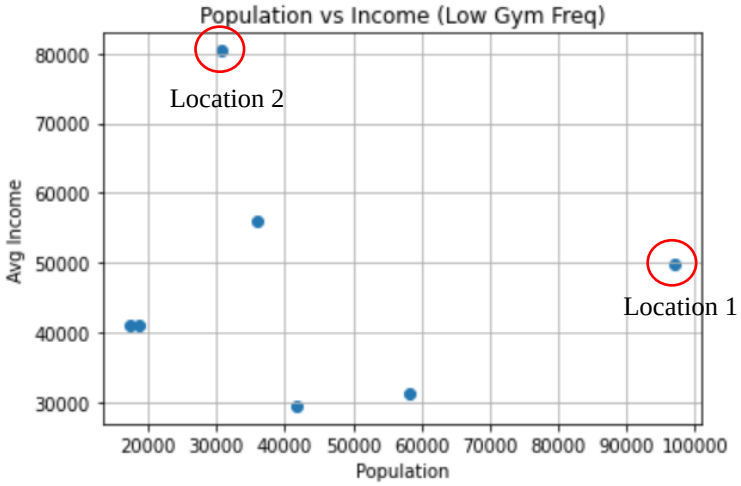

Figure 2 – Population vs Income for Zip Codes with less than 3 gyms

Looking at the plot, we can see how each of those areas compare to each other. So when determining the top candidates from the bunch, we have to think about what we're looking at in this graph; we're looking at the total population and the average income in that area. If this graph

was a direct line, we'd want it to go towards the top-right corner because that would be a maximized population and income. So we want to look at points closest to the top right of the chart. The best data points (circled in red) to further analyze is the point all the way to the right and the point closest to top left of the graph. I chose the point all the way to the right because it's definitely the closest to the top right corner of the graph. I also chose the point to the top because it's (arguably) second closest to the top right.

Once we have our top points, we'll be looking at side by side comparisons of the location's characteristics. I've labelled the points "Location 1" and "Location 2 "for easy viewing. Once labelled, now we calculate the difference in population, average household income, and the current number of gyms.

```
Low Freq- Location 1 comparison to Location 2
Population Difference:  -66444
Average Income Difference:  30673.0
Total Gyms Difference:  0
```

Figure 3 – Statistical Values after Location Comparison

3.2.3 Medium Gym Frequency:

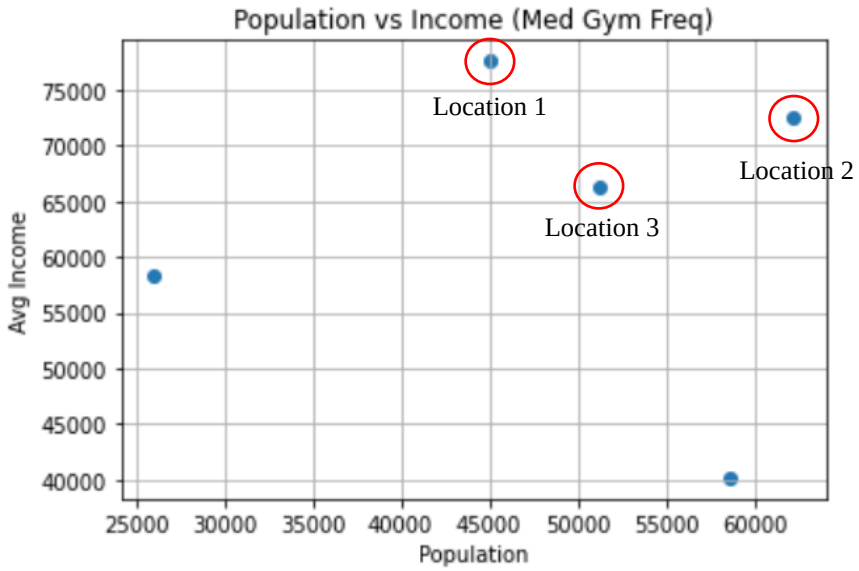

Figure 4 – Population vs Income for Zip Codes with exactly 3 gyms

We'll repeat same procedure that we did for the first plot, so we're going analyze points closest to the top right (maximized population and average income) of the graph. Looking at the plot, we can see that a couple points really hit the mark. We'll use the point closest to the top right of the graph, as well we'll also grab the other two points closest to the right corner. It doesn't matter that we took two samples from the first dataset and three from this one because we don't want to eliminate any good options and those three points definitely stand out.

Similarly to before, we'll look at the comparison of each location:

```
Med Freq- Location 1 comparison to Location 2
Population Difference:  -17219
Average Income Difference:  5141.0
```

Figure 5 – Statistical Values after Location Comparison

```
Med Freq- Location 1 comparison to Location 3
Population Difference:  -6230
Average Income Difference:  11223.0
```

Figure 6 – Statistical Values after Location Comparison

```
Med Freq- Location 2 comparison to Location 3
Population Difference:  10989
Average Income Difference:  6082.0
```

Figure 7 – Statistical Values after Location Comparison

3.2.4 High Gym Frequency:

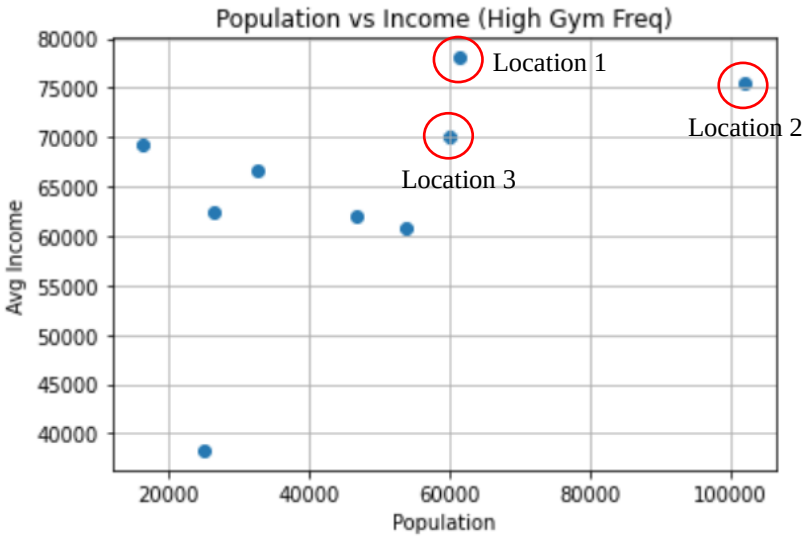

Figure 8 – Population vs Income for Zip Codes with more than 3 gyms

Continuing the same protocol as for the first two datasets, we'll utilize the 2 circled points for further comparison and analysis. I chose 3 points for comparison because the number of gyms for this dataset has a high variance; each zip code has anywhere from 4-8 gyms so we want enough points to have variance to give us more options.

Now repeating the same procedure as before, here's a location to location comparison:

```
High Freq- Location 1 comparison to Location 3
Population Difference:  1558
Average Income Difference:  8035.0
Total Gyms Difference:  2
```

Figure 9 – Statistical Values after Location Comparison

```
High Freq- Location 1 comparison to Location 2
Population Difference:  -40664
Average Income Difference:  2594.0
Total Gyms Difference:  3
```

Figure 10 – Statistical Values after Location Comparison

```
High Freq- Location 2 comparison to Location 3
Population Difference:  42222
Average Income Difference:  5441.0
Total Gyms Difference:  -1
```

Figure 11 – Statistical Values after Location Comparison

4 Results

4.1 High Frequency Gyms:

From our data, we can infer from our final analysis that location 2 will be the most optimal location for this dataset. This can be concluded from a few observations: location 2 has the least number of gyms, location 2 has the highest population by a large margin, and location 2 has the 2nd highest average household income and is only marginally behind location 1.
Location 2 statistics:

Zip Code: 10021
Population: 102,078
Average Income: $75,472
Existing Gyms: 4

4.2 Medium Frequency Gyms:

From this data, we can infer that location 2 will be the most optimal location for this dataset. Location 2 has the highest population amongst the three by a fair decent margin (17,000 more people than location 1 and 11,000 more people than location 3). Then looking at average income, location 1 has the highest, but location 2 is close second and location 3 has the lowest average income. Also remember, location 1 has the lowest population of the three. We only need to look at population and average income for these locations because they all have 3 gym open.
Location 2 statistics:

Zip Code: 10023
Population: 62,206
Average Income: $72,424
Existing Gyms: 3

4.3 Low Frequency Gyms:

From the low frequency dataset, location 1 and location 2 have the same number of gyms, but location 1 has $30,673 more annual income and location 2 has 66,444 more people. So which one is more valuable? Well honestly it doesn't matter because neither location is as optimal as we'd like, and neither location is even close in opportunity and value comparison to the two final locations from the medium and the high gym frequency datasets.

5 Discussion

The results for this project are very interesting. Comparing the two final locations, it's clear that location with zip code 10021 is the best amongst the bunch. The two regions have very similar average household incomes, but zip code 10021 has such a bigger population, nearly 40,000 more people. That location does have one more gym than the other zip code, but I don't think it makes that much of a difference from that much of a variance in population.

Some other observations that I noticed during my research is that this design, process, and analysis could very easily be used for any type of venue. We were able to narrow it down to just gyms, but we could've easily changed some syntax to the observe the same statistics for a café/coffee shop per say.

5 Conclusion

As a whole, this process was successful in achieving the task at hand. In conclusion, my client should open the gym at a location within the zip code 10021. This area deemed to be the most optimal because it has the most opportunity and the most assurance. This result was drawn from multiple series of simplification, analysis, and interpretation.

This process could be improved in the future by getting access to the whole dataset with information for all 64 regions in Manhattan as opposed to just 44 (couldn't purchase full version of dataset). Another way this could be improved would be to closely look at every gym in the remaining zip codes. We could pick apart characteristics such as type of gym/services, the companies of the gyms in the areas, the pricing of these gyms, etc.