INTRODUCTION - PROBLEM AND BACKGROUND

Problem: Identifying best Neighborhoods for opening a new gym / fitness centre in New York City

- Fitness chains could use the analysis to decide on which locations in the city would be optimal for expanding their presence
- Relatively newer categories of fitness chains such as cycling studios and yoga studios
- While I've used "gym" as the keyword for searching for data to analyse, the methodology in this analysis can be replicated for other keywords such as "restaurant" to identify new locations
- From a customer perspective, it could be also be used by fitness enthusiasts to figure the best neighborhoods to live in based on density / quality of gyms in the area

DATA DESCRIPTION

Data used include:

- json file with neighborhoods data for New York City
- Foursquare API – to gather information on locations
- "**Likes**" data for each location was used as an indicator popularity of the gyms as well as the areas which have maximum footfalls at gyms
- Key assumption here is that people would generally have only a single gym subscription at a time

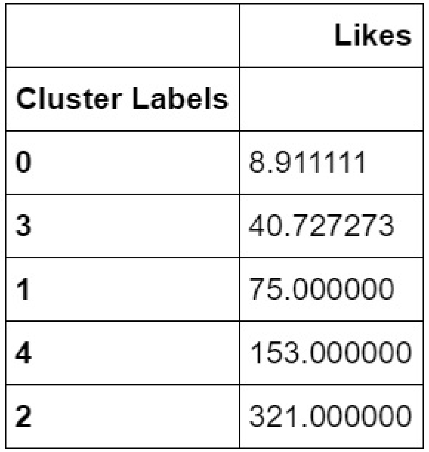
METHODOLOGY – KEY STEPS

1. Retrieving venue data for all neighborhoods in NY, including the venue IDs
2. Create a subset of the venue data to include only rows with "Gym" in the Venue Category

   - While there are other categories of fitness venues such as Yoga Studios, etc, I've chosen Gyms to limit the number of API calls required for further analysis
3. Use the Venue ID column in the new data set to retrieve information on number of likes from the Foursquare API – using the **stats endpoint**
4. Cluster analysis on the average number of likes for each neighborhoods
5. Plot cluster maps and summarize findings

RESULTS (1/2)

Clusters in declining order of average likes



	Likes
Cluster Labels	
0	8.911111
3	40.727273
1	75.000000
4	153.000000
2	321.000000

Most of the Neighborhoods belong to cluster 0, followed by cluster 3

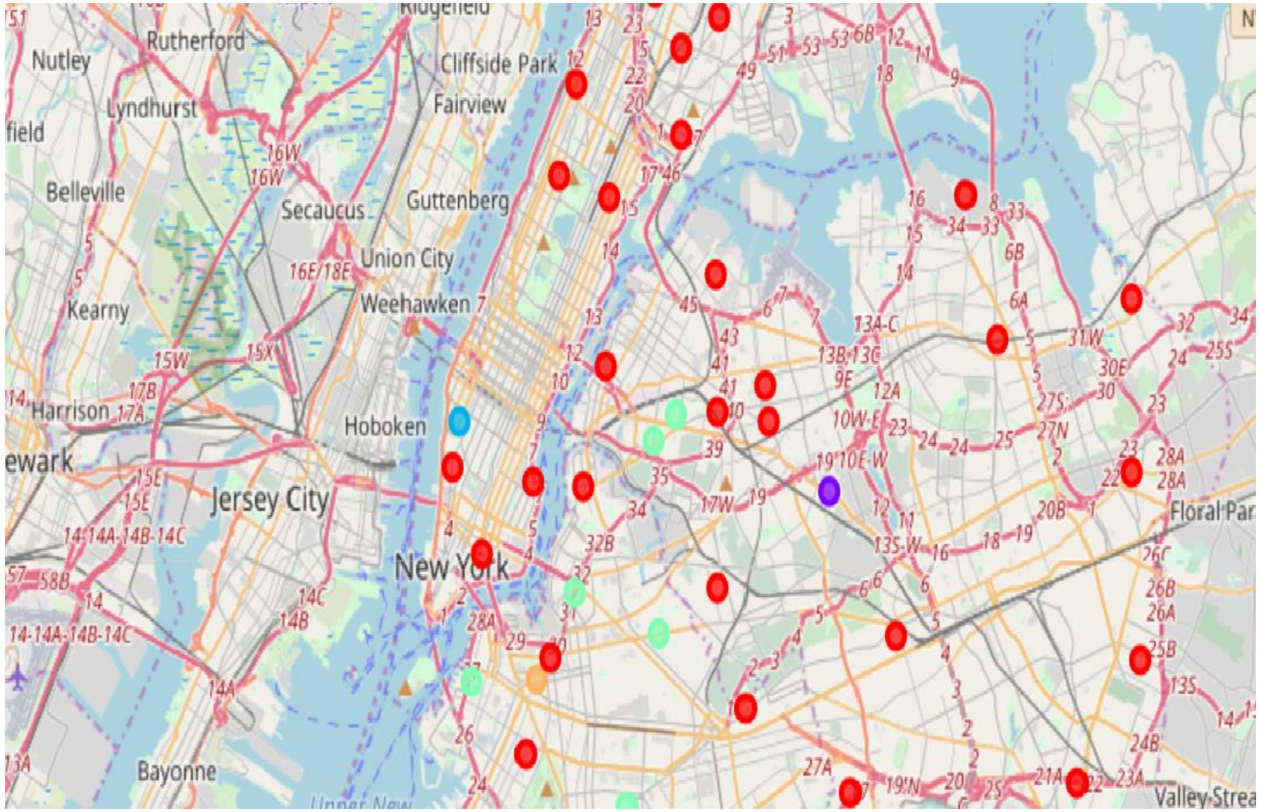
Neighborhoods in Cluster 3 and Cluster 1 have higher ratings than those in cluster 0

Clusters 4 and 2 have a limited number of data points and can be treated as outliers

Please see the notebook provided for further results

RESULTS (2/2)

Snapshot of Cluster Map



DISCUSSION

Based on this analysis, some of the best neighborhoods in terms of likes for existing gyms (see final_df in Jupyter notebook provided) would be (top 5 from Cluster 3):

- Hammels
- Bushwick
- Williamsburg
- Cobble Hill
- Bay Ridge

Also, lower data availability for Neighborhoods in Clusters 2 and 4, along with high likes, could indicate potential pockets of high demand in these areas

While Cluster 0 has a large number of neighboorhoods, these generally have zero or significantly lower number of likes – that could also indicate that these are relatively newer gyms, so may be worth exploring for new launches

CONCLUSION

While clustering may be a good preliminary approach to identifying locations for new gym launches, there may be a number of other factors involved such as

- Tiering (high end vs low priced gyms)
- Availability of other facilities in the area (such as outdoor sports venues, swimming pools, etc.)
- Real estate prices /rentals
- Office locations in the vicinity

Footfall / checkin data may have been a better indicator for new customers – I tried retrieving the total checkin data in the stats endpoint but couldn', so I used the likes data instead