Dies ist ein Artikel über mein Abschlussprojekt zur "Applied Data Science" Spezialisierung, aber auch zum IBM Data Science Professional Zertifikat. Ich habe untersucht, durch welche Parameter sich eine gute Örtlichkeit für ein neues Fitnessstudio in New York City beschreiben lassen könnte. Der Quelltext zur Datenaggregation, -analyse und -visualisierung kann unter Referenz [1] eingesehen werden. Der Artikel ist auf Englisch, da er Teil der Prüfung war.

[1]
https://nbviewer.org/github/markusjaeckle/Coursera_Capstone/blob/master/Capstone_Project_Notebook.ipynb

Opening a Fitness Center in New York City - does it make sense?

An article about the findings of the capstone project by Markus Jäckle for the IBM "Applied Data Science" specialization on coursera.org

Please note:

- The case introduced in this report is fictitious, created by Markus Jäckle, and tries to mimic a real world scenario.
- The respective report can be found here, the respective notebook here.

Abstract

Fitness centers are generating more and more profit in the US.[1] Thus, an example study for finding a suitable borough in New York City to open a new fitness center has been conducted. Based on specific properties, a model has been created and an attempt to propose highly lucrative boroughs has been made.

Introduction

Over the past few years, the revenue generated in the fitness center industry has been on a steady rise.[1] Motivated by this trend, and already benefiting from huge profit gains over the last years, a large fitness center group wants to take hold in New York City.

Expectedly, it want's to make sure that it invests into the right location, because the stakeholders expect it to generate profit from the get-go. However, New York already has a lot of fitness centers and the question is, whether a suitable location can be determined, in this case based on the venues per population based on the density of fitness centers and population density, the amount of likes of the fitness centers within a borough, and the GDP per capita.

Data

Details about the data can be found in the report.

Methodology

Pandas Dataframes were used to aggregate data for further processing.

Folium choropleth maps were used to visualize borough-specific calculated data of the five boroughs of New York City.

The queried venues had to be cleaned, because two duplicate venues have been queried, also 9 venues which were queried by the Manhattan borough location were part of the Queens borough. They had to be removed to not falsify the number of venues per area, since the venues per area were calculated by the venues of the borough within the query radius, and -area respectively. Lastly, 5 venues did not lie in any borough, they also had to be removed from the dataset, which finally consisted of 396 venues.

Venues of original query

Queried venues with removed duplicate entries.

Venues of cleaned query

Queried venues after cleaning the data as described above.

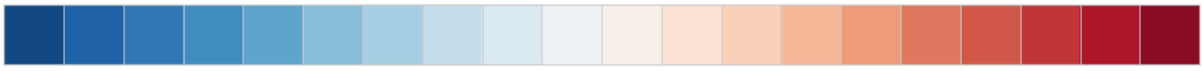
Folium CircleMarkers were used to visualize the fitness center locations, in order to check for locations which are not part of a NYC borough.

Correlation between two kinds of data was investigated by calculating the coefficient of determination R^2 using SciKitLearn LinearRegression and by visualizing the dependencies using Seaborn regression plots.

In order to verify, whether the likes can be used as a substitute for the ratings, such a coefficient of determination was calculated. However, because no sufficient correlation was established, weighed rating signals, were calculated and introduced, in order to compare them with the number of likes, and to investigate a possible correlation, which would make it possible to substitute the weighed rating signals, and thus the rating, with the amount of likes. Please refer to the report for an equation.

The folium plugin BeautifyIcon was used to visualize the number of likes of a venue, using seaborn generated color palettes exceeding the default 256 color scheme of matplotlib, or folium, respectively.

In the following, the color palette used for the 20 clusters mentioned below is shown, please note however, that the color palette used to color-code the likes of a venue consisted 393+1 entries, since this was the maximum amount of likes plus one entry for 0 likes.



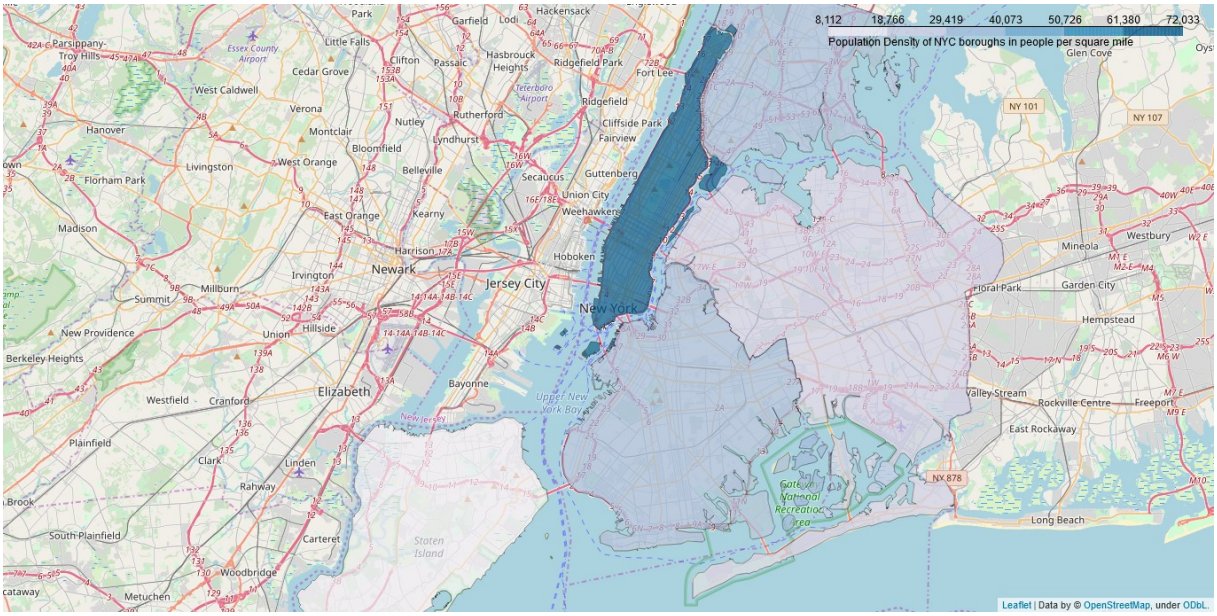
A boxplot has been used to investigate the distribution of venues by their weighed likes, please refer to the report for an equation.

In order to better visualize the distribution of venues with different ranges of weighed likes, 20 Clusters of venues, grouped by their weighed likes were calculated using the SciKitMeans module. Also, they were color-coded by their rating value and visualized using a folium choropleth map and BeautifyIcons.

Lastly, the final comparison included the feature scaled venues per population, weighed likes, and GDP per capita. As mentioned before, unfortunately, it did not include the price tier of the venue, data from crime stats which could help in inferring a measure for safety, and also the data on the traffic was not included.

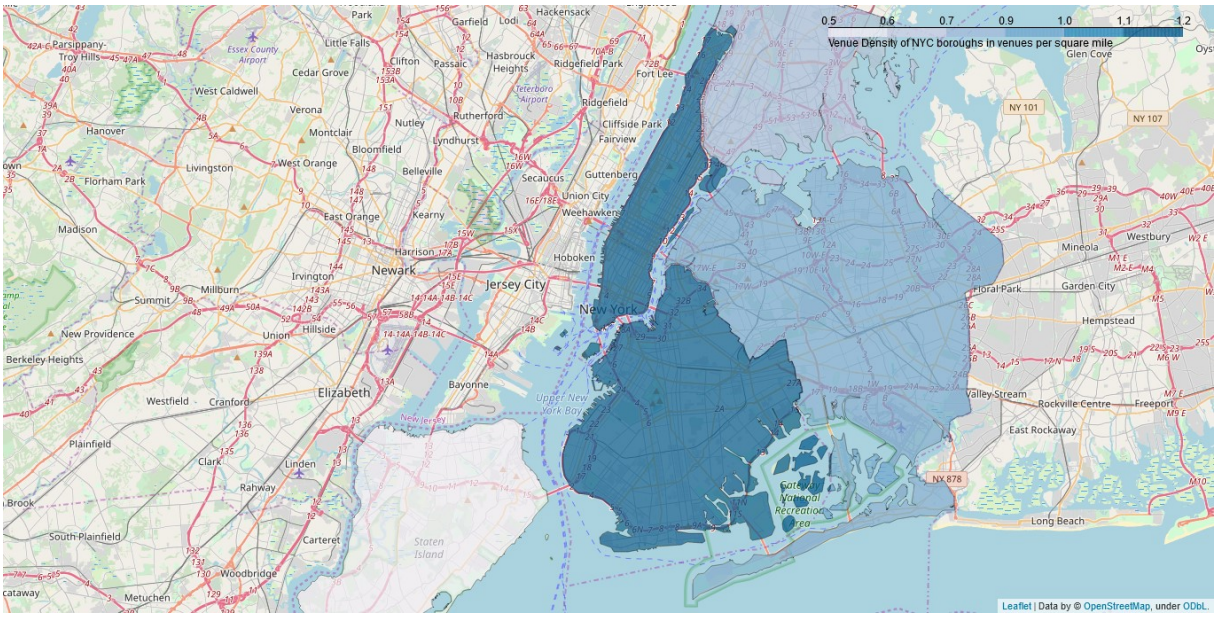
Results and discussion

Visualization and comparison of the population density of the different boroughs:



As can be seen in the figure above, Manhattan has the highest population density of all boroughs. Brooklyn and The Bronx have roughly the same population density, followed by Queens. Staten island has 9 times less people per square mile than Manhattan.

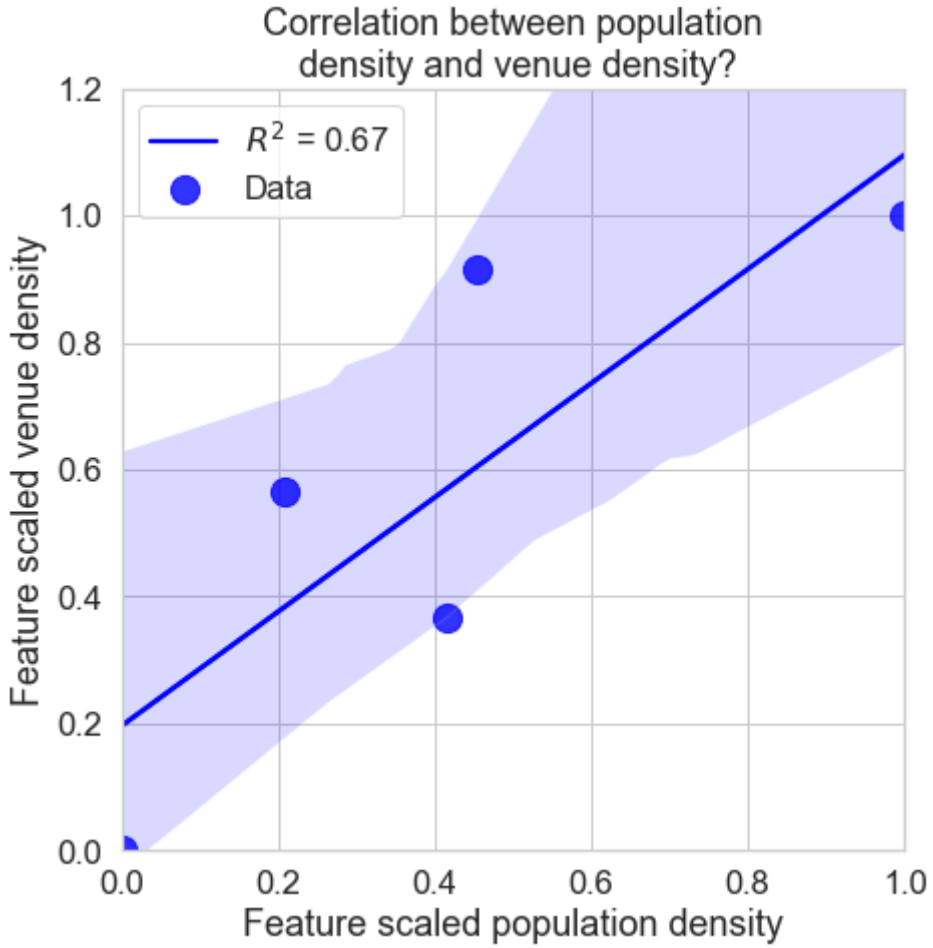
Visualization and comparison of the venue density of the different boroughs:



Suprisingly, according to the figure above, while Manhattan still has the highest venue density of all boroughs, Brooklyn almost has the same density. Queens has double as much venues per area than Staten Island, and The Bronx is in between both.

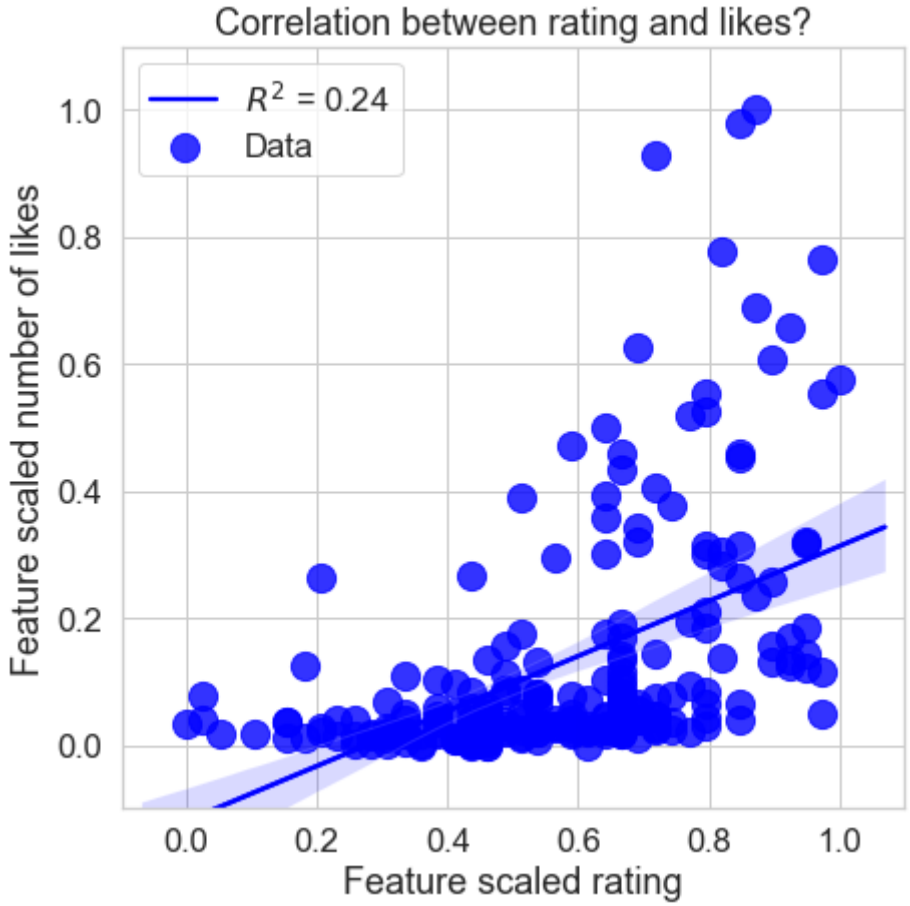
Another interesting find is, that Manhattan is leading in population density and venue density, while Staten Island is the last in both features, which would make sense, if both features correlate.

Correlation between the population density and the venue density:



The correlation analysis shown in the figure above didn't yield a conclusive answer to that question. However, there are good reasons for a correlation, since the more people a borough has per area, if there is a constant percentage of people going to a fitness center, the demand in fitness centers should rise, which then of course would lead to an increase in fitness centers. One reason, why it was not possible to conclusively establish a correlation between both, and why the coefficient of determination R^2 was only 0.67 could be, that there were not enough data points to counter the effect of scattering, which was inherent to the data.

Correlation between ratings and likes?

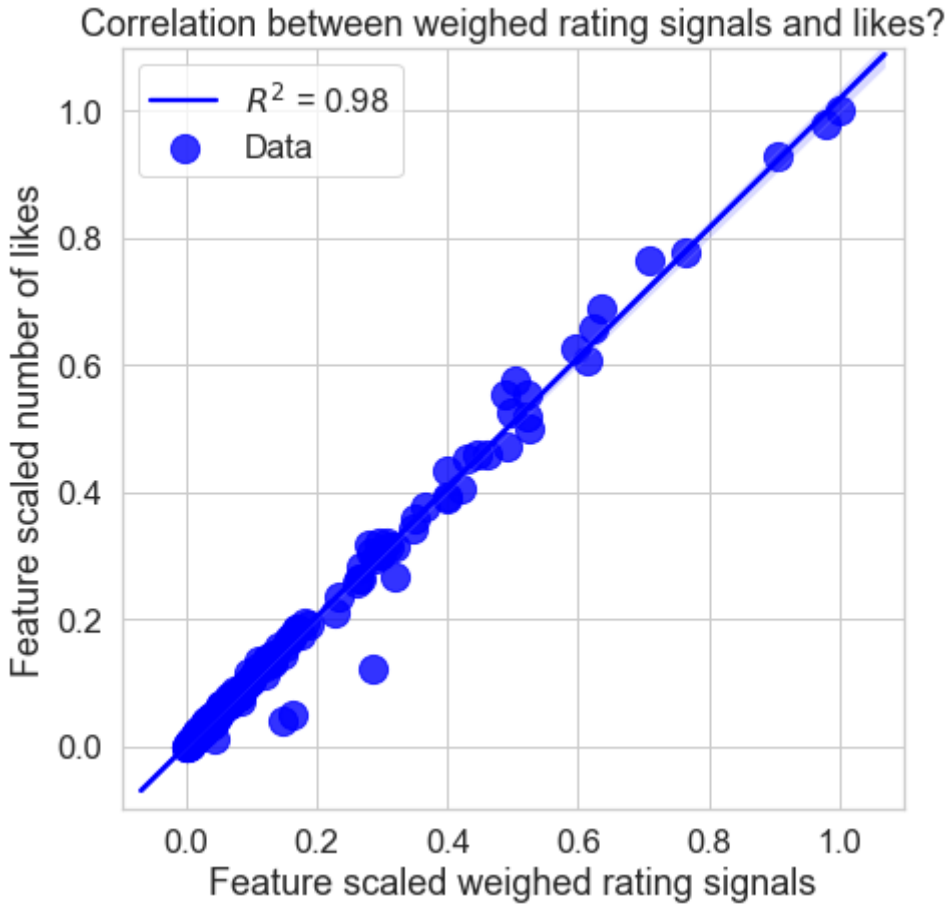


This figure indicates that there is no correlation between the rating and the number of likes.

Correlation between weighed rating signals and likes!

However, a like is an expression of a good rating. Thus, it would make sense that the number of ratings times the normalized rating, which is the rating divided by 10, correlates with the number of likes.

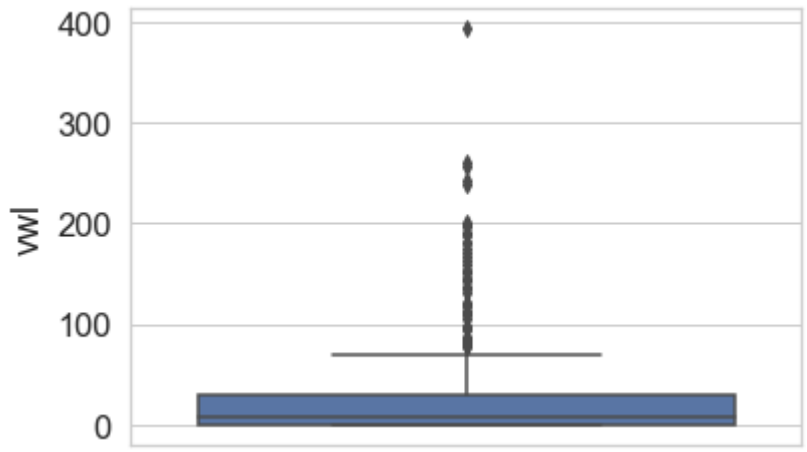
In the next regression plot, let's plot the number of likes vs weighed rating signals, as defined in the report:



No doubt, an almost perfect correlation with only three small outliers is found. Indeed, the established correlation was conclusive in such a way, that the weighed rating signals calculated as mentioned almost directly equaled the number of likes in most cases.

Visualized weighed likes per borough and weighed likes per venue

Since some boroughs have a higher population density, obviously the like counts will be higher in those boroughs. Thus, the weighed likes are the likes per venue divided by the population density of the respective borough. However, these weighed likes per venue have been scaled to the original likes maximum.

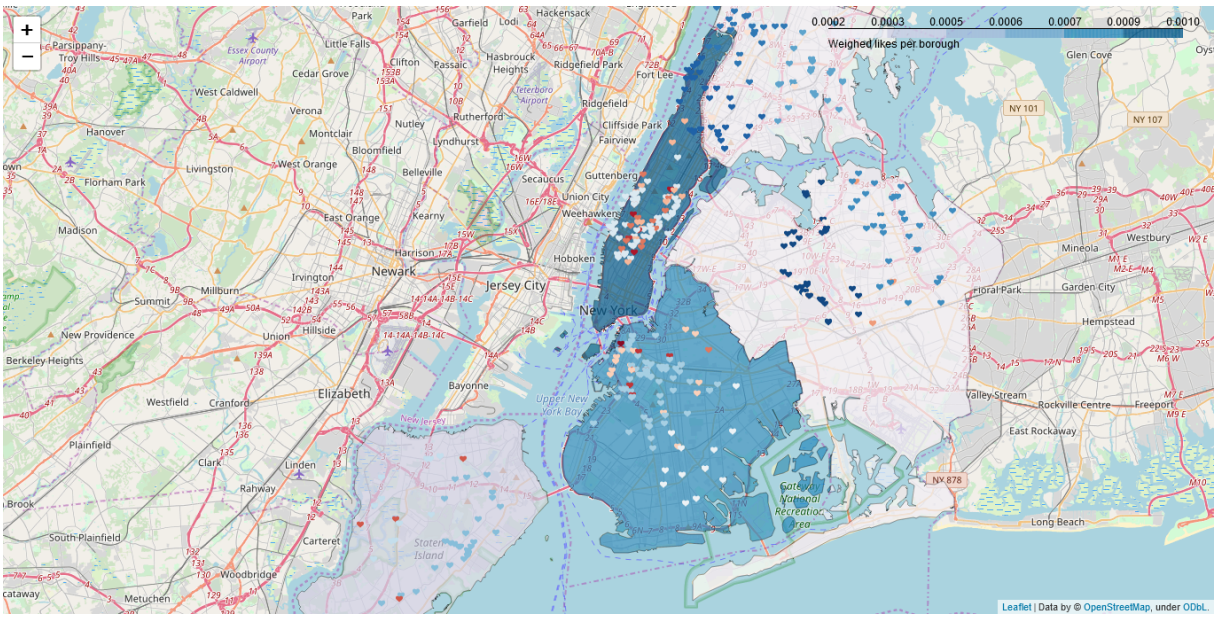


The boxplot clearly shows that 75% of the venues don't get beyond 30 weighed likes per venue, which is the upper quartile threshold. However, a considerable amount of outliers is also present.

A future investigation could show, whether those venues are of such a high quality that they excel their competitors in quality, or whether there are benefits to their location, which could lead to finding an attractive potential location for the fitness center.

Clusters of venues by their rating

Since most venues do not have a high rating count and thus have been hard to distinguish in the likes per venue plots, let's use the skmeans algorithm, which in this case gets us 20 clusters, which are sorted by their feature-scaled weighed likes and their feature-scaled coordinates:



The plotted weighed likes per venue in the figure above on one hand suggest, that an attractive location for founding a fitness center is either in the south of Manhattan or in the north of Brooklyn. However, the high venue density and high amount of weighed likes could also be indicative of a saturated market.

Conclusion

Borough	Inverse Venues per Population	Weighed Likes	GDP per capita	Score
Brooklyn	0.70	0.72	0.02	1.44
Manhattan	1.00	1.00	1.00	3.00
Queens	0.46	0.12	0.03	0.61
Staten Island	0.00	0.28	0.00	0.28
The Bronx	0.88	0.00	0.00	0.88

Comparison of the feature scaled values of the inverse venues per population, weighed likes, and GDP per capita; the 'Score' column equals to the sum over all entries along a row.

The table above clearly shows that Manhattan is leading in every feature.

If the weighed likes are viewed as positive, Brooklyn is also a rather good location to invest in, since they have a rather low fitness centers per population value and since the people in Brooklyn seem to like fitness centers in general, since they have the second highest weighed likes value of all boroughs. However, the low GDP per capita could mean, that, compared with Manhattan, maybe also a lower ratio of people is inclined to go to fitness centers.

According to the scoring, surprisingly, Staten Island has the highest amount of venues per population, the rather low weighed likes value per venue, and the low GDP per capita are counter indicative of it being an prospective borough to invest in.

The Bronx has a rather low venues per population value, however it has the lowest amount of weighed likes per venue and the lowest GDP per capita.

Queens has the second lowest overall score, due its average venues per population, low weighed likes value and seemingly rather low GDP per capita. However, it has to be noted, that Manhattan seems to tower all other GDP per capita values, thus this feature maybe shouldn't be features scaled, but rather standardized.

Generally, Manhattan seems to be the best choice for opening a fitness center, however, Brooklyn should possibly considered as well, based on this model.

Since a high likes count can be indicative of a saturation of fitness centers with a high rating, the analysis will be repeated with the likes count viewed as negative. Remember that this weighed likes count is normalized per venue and population density:

Borough	Inverse Venues per Population	Inverse Weighed Likes	GDP per capita	Score
Brooklyn	0.70	0.28	0.02	1.00
Manhattan	1.00	0.00	1.00	2.00
Queens	0.46	0.88	0.03	1.37
Staten Island	0.00	0.72	0.00	0.72
The Bronx	0.88	1.00	0.00	1.88

Comparison of the feature scaled values of the inverse venues per population, weighed likes, and GDP per capita; the 'Score' column equals to the sum over all entries along a row.

Even though losing the advantage of its high weighed likes count, according to the table above, Manhattan still gets the best rating, followed by The Bronx.

The Bronx takes the second place, however, an investigation into the reasons why customers in The Bronx barely give any likes should be performed, possibly there is a correlation with the GDP per capita, however the data presented above only suggests such a correlation for the extreme cases of Manhattan and The Bronx.

However, an interesting find is, that due to its average venues per population count, and due to its low weighed likes value, which can be indicative of a desire for a non-saturated marked, and due to its second highest GDP per capita, Queens now scores rather well.

Overall, while a low weighed likes count can be indicative of a non-saturated market, it can also be a hint, that customers in those areas are either hard to satisfy, or don't care much about rating a location.

Depending on which reasons applies to the Bronx, should also be considered as a location, but since Manhattan and Queens hold their respective positions rather well, they should be the first boroughs to consider.

The main problem in this analysis was, that there was no available target variable.

The number of venues per population per borough on one hand would have been such a target variable, since it is an indicator of the attractiveness of such a location for other fitness centers, however, a borough with a lack of fitness centers is far more attractive than one saturated with fitness centers, and thus, the mentioned feature could also be counter indicative of an attractive location to invest in.

On the other hand, a combination of a high number of venues and a small weighed-likes value is indicative of a desire for a high-quality fitness center in this area. However, in certain boroughs, there also could be a natural inclination to not give likes as much as in other boroughs. Thus, even in this case, the fitness center might not be profitable, for one reason or another.

Additional data, like price tiers of the venues, traffic stats, and crime stats, might have been able to alleviate this issue, however they were not readily available for all compared boroughs.

References

[1] https://www.statista.com/statistics/236120/us-fitness-center-revenue/

References for the queried Data:

[2] https://en.wikipedia.org/wiki/New_York_City#Boroughs

[3] https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

[4] https://developer.foursquare.com/docs/api/endpoints

[5] https://data.ny.gov/Transportation/Annual-Average-Daily-Traffic-AADT-Beginning-1977/6amx-2pbv

[6] https://data.ny.gov/Public-Safety/Index-Crimes-by-County-and-Agency-Beginning-1990/ca8h-8gjq