Don’t get stressed –
Study on Finding a perfect spot to open a Yoga Studio

1. Introduction

1.1. Current Situation

In our stressful time, self-awareness has become more and more important to get resilient and to recover from the daily challenges. In the last couple of years several strategies like sport, meditation and yoga have seen a risen request amongst people. Manhattan is one of the busiest and most stressful cities in the world. Recreation should always be of high demand. Since time is of an essence here as well. Long distances are not acceptable so the venues must be easily to reach from home.

1.2. Business Case

Our customer wants to open a new yoga studio in Manhattan. Special interest should be paid to the already available venues and population and income of the area. Basic research on market size and price should also be conducted.

2. Data Acquisition and Cleaning

2.1. Data Akquisition

The geographical data were gathered using Foursquare.com. The API was programmed in a Jupyter Notebook with a Python Kernel.
The resulting data frame was analyzed in the same notebook.
The statistical data on population and median income were collected manually from www.points2home.com for each neighborhood. This was necessary since census.gov don´t use neighborhoods. Even zip code is not used otherwise the geospatial data could be easily transferred in the zip code.
From the Foursquare dataset the locations for yoga studios, gyms, pilates studios, spas and sport clubs were gathered and combined in one data frame. This was necessary since yoga classes can also be offered in this venue as well.

2.2. Data Cleaning

The data of both frames where combined. Manhattanville and Morningside heights has got no sports venues at all. The missing data (NaN) was replaced with “0”.

3. Exploratory Data Analysis

3.1. Data Grouping

The data set for the venue was first visualized in a folium map (Figure 1).

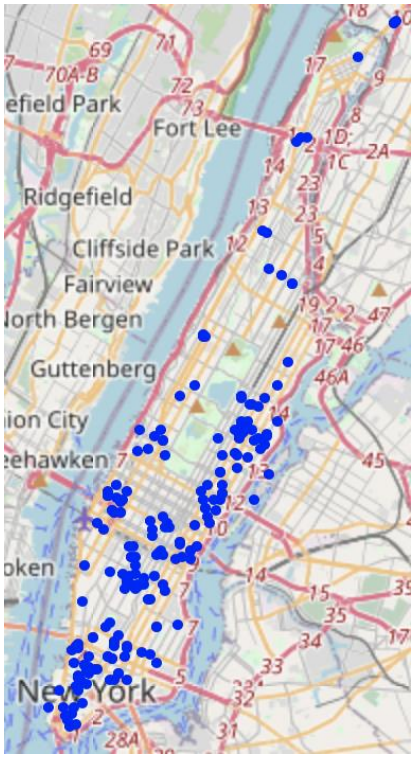


Figure 1: Folium Map of Manhattan

It can be seen that the highest density for sport venues is in the south and east of Manhattan.

The dataset was then grouped in a pivot table and a pie chart was derived.

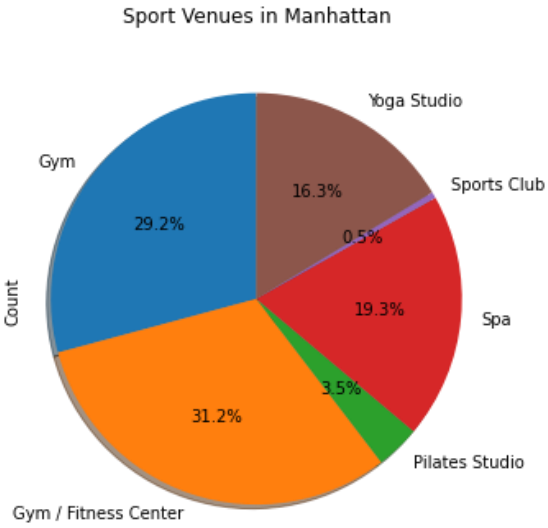


Figure 2: Pie Chat for Sport Venues

As seen in Figure 2, only 16.3% of the sport venues are yoga studios. The Main Part are gyms and gyms/ fitness centers with 29.2% and 31.2%.

3.2. Analysis of the Neighborhoods

NEIGHBORHOOD	MEDIAN INCOME [US$]	POPULATION	GYM	GYM / FITNESS CENTER	PILATES STUDIO	SPA	SPORTS CLUB	YOGA STUDIO	SUM	POPULATION PER SPORT VENUE	POPULATION PER YOGA STUDIO
BATTERY PARK CITY	21.638	10.970	4	-	-	-	-	-	4	2.743	10.970
CARNEGIE HILL	122.969	45.225	3	1	-	-	-	3	7	6.461	15.075
CENTRAL HARLEM	29.059	116.345	1	2	-	1	-	-	4	29.086	116.345
CHELSEA	115.556	48.108	1	-	-	-	-	-	1	48.108	48.108
CHINATOWN	75.086	12.874	-	-	-	3	1	1	5	2.575	12.874
CIVIC CENTER	159.882	5.974	2	6	-	5	-	3	16	373	1.991
CLINTON	103.792	46.648	3	5	-	3	-	-	11	4.241	46.648
EAST HARLEM	33.720	95.589	1	-	-	2	-	-	3	31.863	95.589
EAST VILLAGE	90.939	65.101	-	1	-	1	-	-	2	32.551	65.101
FINANCIAL DISTRICT	186.231	27.834	3	3	-	1	-	-	7	3.976	27.834
FLATIRON	132.988	14.560	2	4	-	4	1	2	13	1.120	7.280
GRAMERCY	125.574	19.342	-	-	2	2	-	1	5	3.868	19.342
GREENWICH VILLAGE	125.831	30.283	3	-	1	1	-	1	6	5.047	30.283
HAMILTON HEIGHTS	40.161	17.745	-	-	-	-	-	2	2	8.873	8.873
HUDSON YARDS	92.840	24.117	2	5	-	-	-	-	7	3.445	24.117
INWOOD	54.406	42.399	-	-	-	-	-	1	1	42.399	42.399
LENOX HILL	143.540	88.306	3	3	1	-	-	-	7	12.615	88.306
LINCOLN SQUARE	68.770	79.444	2	4	-	-	-	1	7	11.349	79.444
LITTLE ITALY	113.191	28.799	1	1	-	1	-	1	4	7.200	28.799
LOWER EAST SIDE	36.982	74.479	-	-	-	-	-	1	1	74.479	74.479
MANHATTANVILLE	41.453	40.568	-	-	-	-	-	-	-	40.568	40.568
MANHATTAN VALLEY	91.624	92.251	-	1	-	-	-	2	3	30.750	46.126
MARBLE HILL	58.408	71.132	2	-	-	-	-	1	3	23.711	71.132
MIDTOWN	122.484	306.638	3	-	1	2	-	-	6	51.106	306.638
MIDTOWN SOUTH	103.792	45.498	1	3	-	-	-	1	5	9.100	45.498
MORNINGSIDE HEIGHTS	91.624	46.942	-	-	-	-	-	-	-	46.942	46.942
MURRAY HILL	128.836	2.599	2	3	-	1	-	1	7	371	2.599
NOHO	112.314	5.532	1	-	1	-	-	2	4	1.383	2.766
ROOSEVELT ISLAND	104.808	12.440	1	1	-	-	-	-	2	6.220	12.440
SOHO	109.829	13.224	2	-	-	-	-	1	3	4.408	13.224
STUYVESANT TOWN	63.717	58.293	-	1	-	-	-	-	1	58.293	58.293
SUTTON PLACE	150.718	31.130	3	4	1	1	-	2	11	2.830	15.565
TRIBECA	168.627	16.236	1	2	-	3	-	1	7	2.319	16.236
TUDOR CITY	131.045	15.846	3	2	-	1	-	1	7	2.264	15.846
TURTLE BAY	140.882	14.827	-	1	-	2	-	-	3	4.942	14.827
UPPER EAST SIDE	130.804	197.935	-	5	-	3	-	3	11	17.994	65.978
UPPER WEST SIDE	132.605	99.773	1	1	-	-	-	-	2	49.887	99.773
WASHINGTON HEIGHTS	53.525	180.158	2	1	-	-	-	-	3	60.053	180.158
WEST VILLAGE	133.501	30.344	1	-	-	-	-	-	1	30.344	30.344
YORKVILLE	122.969	64.404	5	2	-	2	-	-	9	7.156	64.404

Table 1: Data Set for analysis

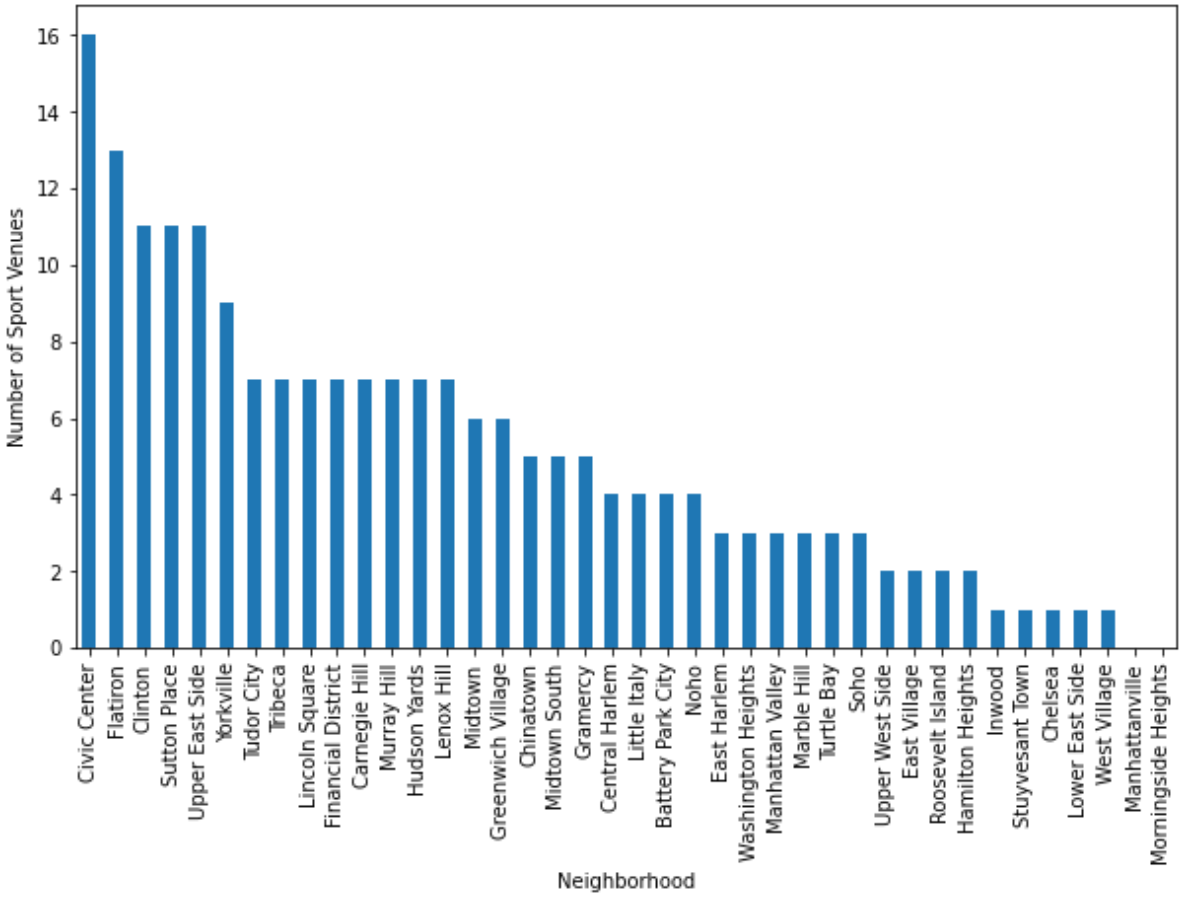


Figure 3: Number of Sport Venues by Neighborhood.

Civic center and Flatiron have the most sport venues with 15 and 14 venues in total. As mentioned before, Manhattanville and Morningside Heights has no venues at all.

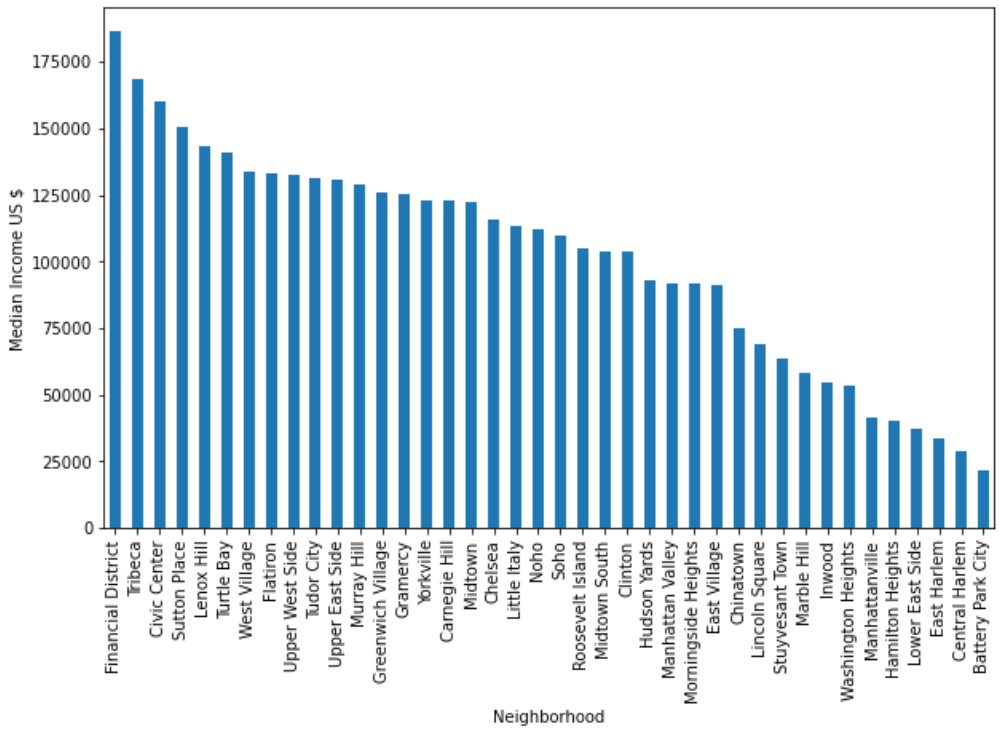


Figure 4: Median Income by Neighborhood.

As seen in Figure 4, Civic Center and Tribeca has the population with the highest median income. The median income was used since it is not affected by extreme values.

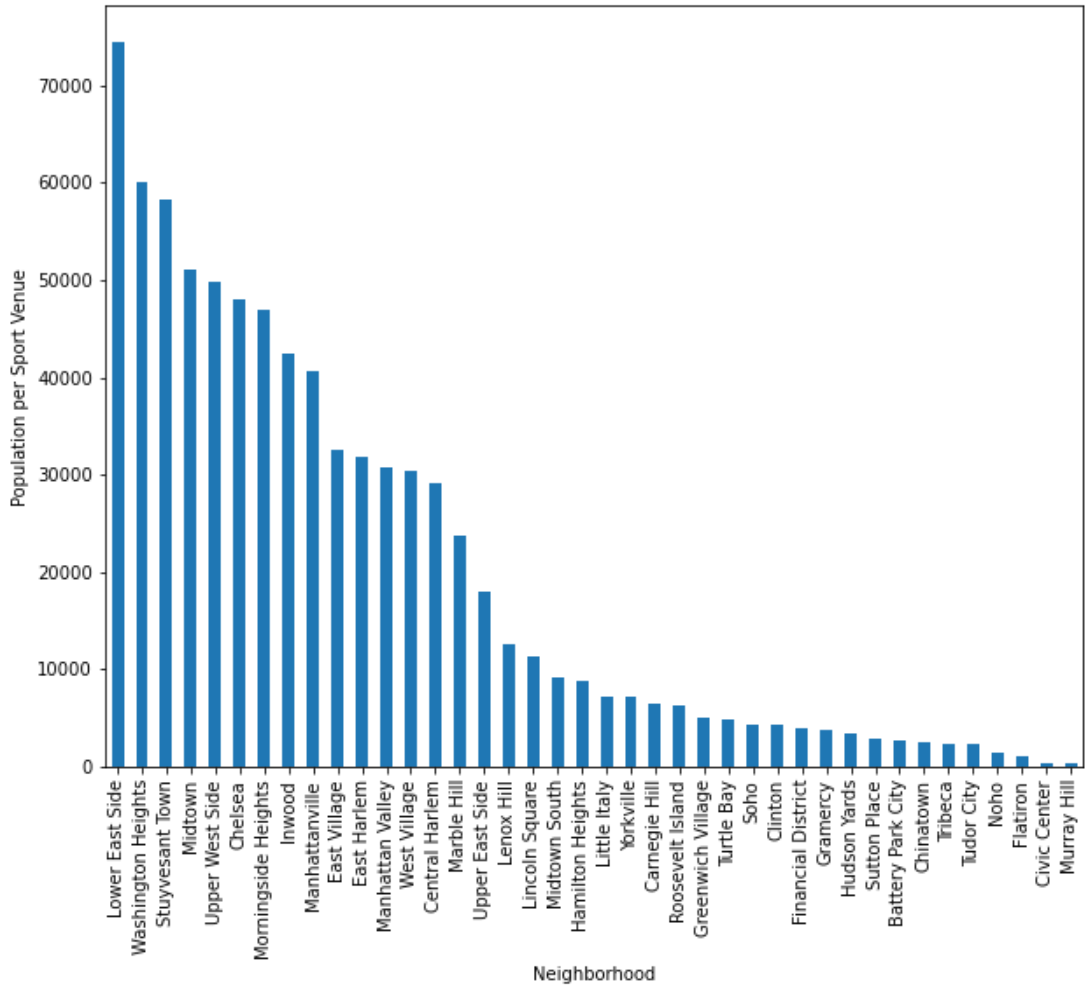


Figure 5: Population per Sport Venue.

Instead of showing the total population, the KPI population per sport unit was used. It was calculated by division of the total population of the neighborhood by the total number of sport venues.

The first observation is that the Civic Center has the highest number of sport venues and the highest median income. Furthermore, it has the lowest population per sport venue. This area shows a very strong competition.

4. Finding the Right Predictor

4.1. Regression

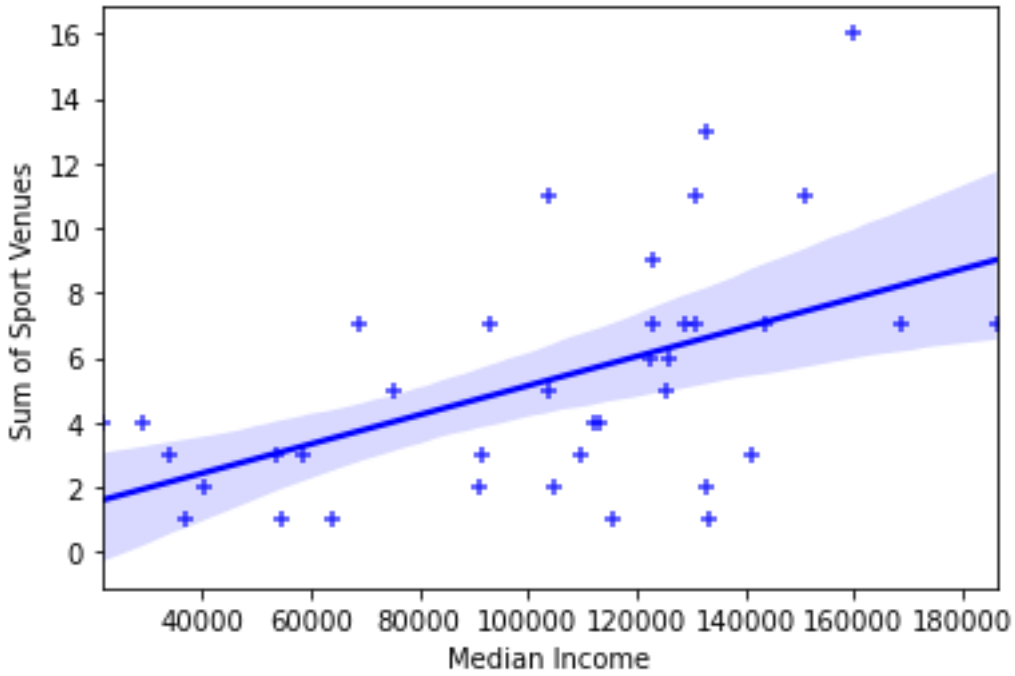


Figure 6: Correlation Median Income with Sum of Sport Venue for each Neighborhood.

It can be seen in Figure 6 that areas with low income less sport venues can be found. The scatter plot and regression line indicate a weak positive correlation between median income and number of sport venues

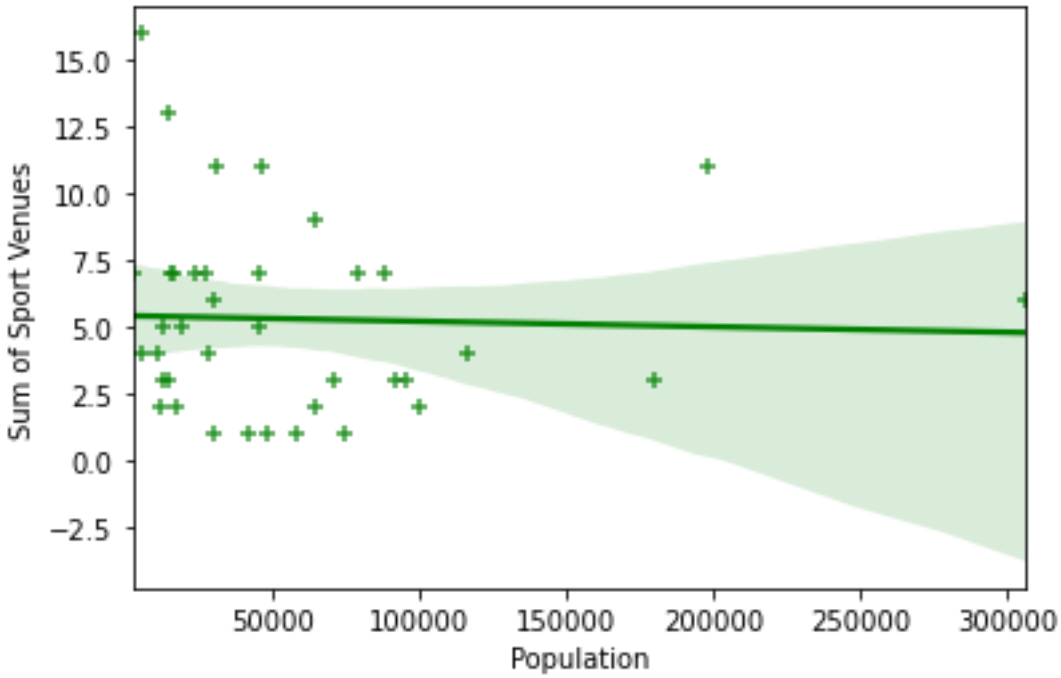


Figure 7: Correlation Population and Number of Sport Venues

The population in a neighborhood shows no correlation with the number of sport venues inside it. It is not a suitable predictor.

4.2. Bubble Plot

The features “Median Income”, “Population” and “Number of Sport Venues” per neighborhood were visualized in a Bubble Plot (Figure 8).

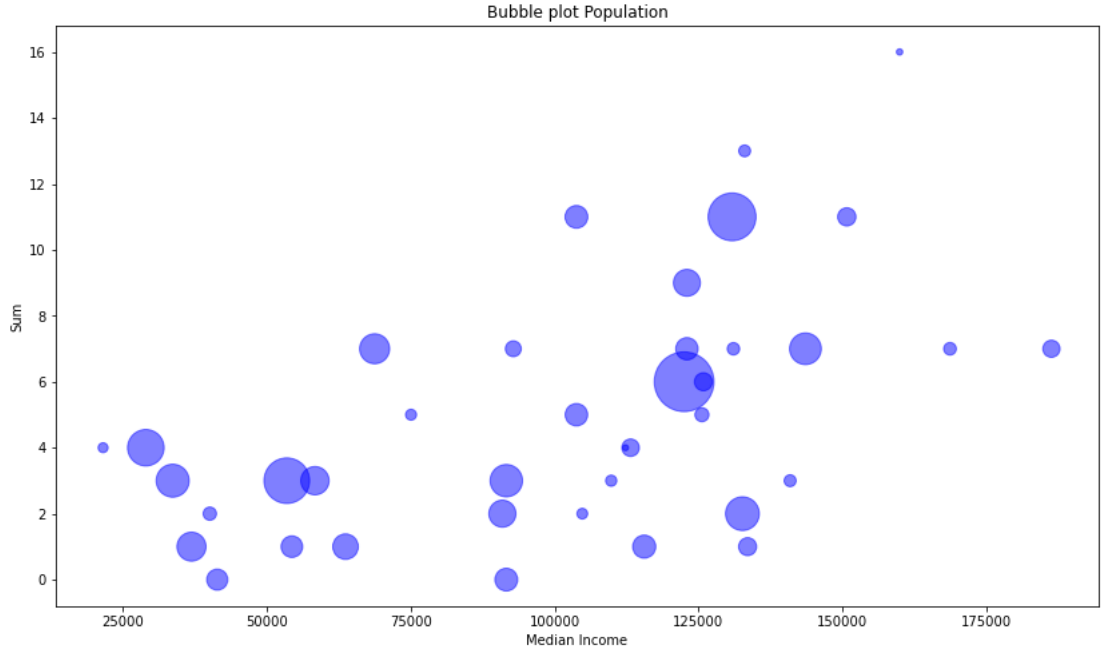


Figure 8: Bubble Plot for Number of Sport Venues.

It can be seen that the neighborhoods of interest are in the top-right area where the median income and population is maximized. In this area, there are three neighborhoods. The diameter of the bubble varies. The neighborhood with the smallest bubble of this set of three neighborhoods is an interesting neighborhood for opening the yoga venue.

4.3. Classification Models

For finding suitable areas for a yoga studio, the classification approach was used as well.

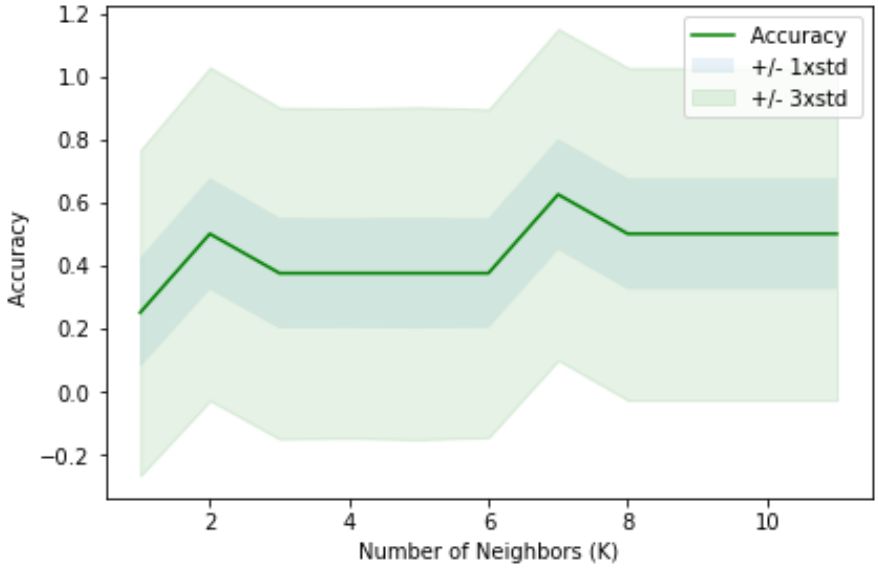


Figure 9: Accuracy Plot

The method of choice was the K-nearest clustering method. So, the neighborhoods were clustered according to these four features.

It was used to build clusters by predicting number of yoga studios by using the median income, population and total number of sport venues in this area.

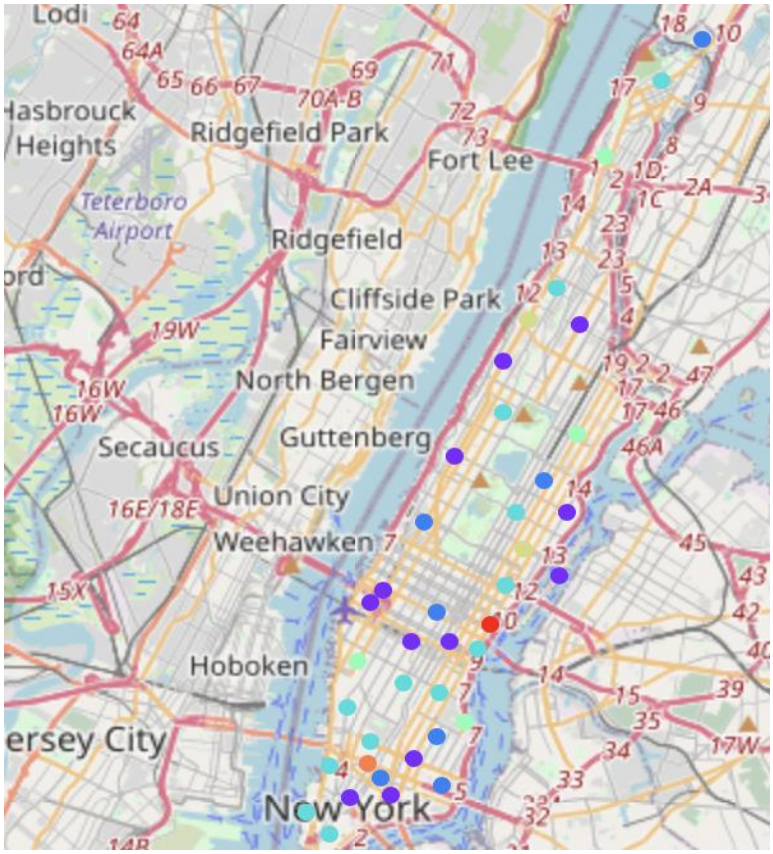
The values were normalized using the standard scalar packages.

CLUSTER LABELS	NEIGHBORHOOD	LATITUDE	LONGITUDE	POPULATION	MEDIAN INCOME	YOGA STUDIO	SPORT VENUES
0	Turtle Bay	40,75204237	-73,96770825	14827	140882	0	3
1	Chinatown	40,71561842	-73,99427936	12874	75086	1	5
1	Central Harlem	40,81597607	-73,94321113	116345	29059	0	4
1	Yorkville	40,77592985	-73,94711784	64404	122969	0	9
1	Roosevelt Island	40,76215961	-73,94916769	12440	104808	0	2
1	Upper West Side	40,787658	-73,97705924	99773	132605	0	2
1	Clinton	40,75910089	-73,99611936	46648	103792	0	11
1	Murray Hill	40,74830308	-73,97833208	2599	128836	1	7
1	Morningside Heights	40,80799974	-73,96389628	46942	91624	0	0
1	Noho	40,72325902	-73,98843368	5532	112314	2	4
1	Civic Center	40,71522892	-74,0054153	5974	159882	3	16
1	Midtown South	40,74850966	-73,98871313	45498	103792	1	5
1	Hudson Yards	40,75665808	-74,00011136	24117	92840	0	7
2	Marble Hill	40,87655078	-73,91065966	71132	58408	1	3
2	Lincoln Square	40,77352889	-73,98533777	79444	68770	1	7
2	Midtown	40,7546911	-73,98166883	306638	122484	0	6
2	East Village	40,72784678	-73,98222617	65101	90939	0	2
2	Lower East Side	40,71780675	-73,98089032	74479	36982	1	1
2	Little Italy	40,71932379	-73,99730467	28799	113191	1	4
2	Carnegie Hill	40,78268257	-73,95325647	45225	122969	3	7
3	Inwood	40,86768396	-73,92121042	42399	54406	1	1
3	Hamilton Heights	40,82360428	-73,94968792	17745	40161	2	2
3	Upper East Side	40,77563857	-73,96050763	197935	130804	3	11
3	Greenwich Village	40,72693289	-73,99991403	30283	125831	1	6
3	Tribeca	40,72152197	-74,01068329	16236	168627	1	7
3	West Village	40,73443394	-74,00617998	30344	133501	0	1
3	Manhattan Valley	40,79730704	-73,96428618	92251	91624	2	3
3	Gramercy	40,73720983	-73,98137595	19342	125574	1	5
3	Battery Park City	40,71193198	-74,01686931	10970	21638	0	4
3	Financial District	40,70710711	-74,01066545	27834	186231	0	7
3	Sutton Place	40,76028033	-73,96355614	31130	150718	2	11
3	Tudor City	40,74691741	-73,97121929	15846	131045	1	7
3	Flatiron	40,73967305	-73,99094711	14560	132988	2	13
4	Washington Heights	40,85190253	-73,93690028	180158	53525	0	3
4	East Harlem	40,79224947	-73,94418223	95589	33720	0	3
4	Chelsea	40,74403471	-74,00311633	48108	115556	0	1
4	Stuyvesant Town	40,73099955	-73,9740517	58293	63717	0	1
5	Manhattanville	40,81693443	-73,95738539	40568	41453	0	0
5	Lenox Hill	40,76811266	-73,95885969	88306	143540	0	7
6	Soho	40,72218384	-74,00065667	13224	109829	1	3

Table 2: Neighborhoods clustered by median income, population and number of sport venues.

The optimum value for K was calculated by finding the “elbow” for the accuracy score (Figure 9).

The plot showed a value for 0,625 for 7. The k-value was used to calculate the clusters. The result is shown in table 2. The clusters were visualized on a folium map of Manhattan (Figure 10).



Cluster	Color
0	*Red*
1	*Blue*
2	*Light blue*
3	*Turquoise*
4	*Light green*
5	*Yellow*
6	*Orange*

Figure 10: Map of Manhattan with 7 Clusters

5. Discussion

The data set obtained from Foursquare indicated that the neighborhoods Manhattanville and Morningside Heights has no sports venue at all. This area should be generally considered to be a good place for the opening of a yoga studio. Both have a comparable population between 40,000 – 47,000 inhabitants. However, the median income of both neighborhoods is significantly different with 41,453 US$ for Manhattenville and 91,624 US$ for Morningside Heights.

By the regression analysis only the median income shows a correlation with the number of sport venues. This means in neighborhoods with higher income more studios can survive because the people have the money to afford yoga lessons. If the neighborhoods are compared by the indicator sport venue per population Morningside heights and Manhattanville drops to position 7 and 9. There are at least 6 neighborhoods with a more favorable sport venue to population ratio. However, the regression analysis showed no correlation between the number of sport venues and the population in a neighborhood. The population is not suitable feature for a prediction.

In the combined visualization method “bubble chart”. The scatterplot of the median income and sum of sport venues was updated by the population. This was achieved by correlation of the bubble size to the population.

Finally, machine learning was used to analyze the data by clustering. The trained algorithm showed for k=7 in the first cluster only one neighborhood. Turtle Bay has the 6th highest median income and three sport venues. The population is around 14,000 inhabitants. The second cluster contains twelve neighborhoods, including Morningside Heights.

Special attention should be paid to the 6th cluster where Manhattanville is present. The median income is relatively low here.

6. Conclusion

It could be shown that the median income of a neighborhood correlate with the number of sport venues. Based on the discussion above we recommend opening a yoga studio at Morningside Heights. The neighborhood has no sport venue at all a high median income and enough inhabitants to support the yoga venue. The data analysis performed in this work can be easily upscaled by another feature. The next steps would include feature like sex ratio in a given neighborhood as well as number of offices. It could be examined if people tend to visit spot venues before or after work and bring their sport equipment along. An indication for this thesis is the high number of sport venues (16) in the Civic Center, where only 6,000 inhabitants live but a high number of offices are located.