Evaluation of the Feasibility of Opening a Multi-Sport-Discipline Center in Manhattan (New York City)

Enrico Tomassoli, September 5th, 2020

1. Introduction

1.1. Description & Discussion of the Background

1.1.1. Background

New York city is one of the biggest cities in the world. Among its 5 boroughs, Manhattan is the most densely populated one, which counts 1.629 million people (2019) that occupy 22.82 square miles of land, an average of approximately 72,000 inhabitants per square mile. With a GDP of more than $ 600 billion (2018), it holds the 1st place in the per-capita GDP ranking. Probably one of the most famous neighborhoods in US, if not in the world, Manhattan has always been seen as the place where very wealthy and successful Americans and tycoons live. Having lived in this city for more than 10 years, it easy to notice that people who lives in Manhattan tend to be very healthy too, regularly practicing physical activities of different type, ranging from running to yoga, from basketball to hokey. Due to its multicultural character, the city of New York, and in particular its most famous borough, several sports are practiced in the area. In this study our attention will be focused mainly on multisport-center. The main activities that this new sport hub will provide includes, but not limited to, boxing, gym, spin classes, cardio activities, dance and yoga. The idea is having a multi space that, when properly managed and organize, can provide more that one activity. For example, a room can host yoga classes in the morning, cardio activities during the day, and dance class in the evening. Other rooms might have more sport-specific character. For example, a specific space can be designated just for boxing all the time, due to the nature of the equipment that is required for its practice.

1.1.2. Business Problem

The main goal of this report is to evaluate the opportunity of opening a multi-sport center targeting all ages (i.e. not limited to children and teenagers, but also including adults). The main benefit of this kind of business is providing a type of membership that includes more than one activity in one place. Due to the limited space and amenities available in the city, one of the main assumption is that the cost associated with practicing multiple physical activities (which includes but not limited to training, organization, amenities and their maintenance, managing cost, equipments, etc.) can be hardly afford by mid-class segment of the population, fact that forces us to move our attention towards high-end segment instead. This is only an assumption made in this study and it is based on the fact that this type of business needs high revenues to be run in a city where the cost of land, rent, ect., are a significant portion of the expenses. Most important, among the several neighborhoods in

the island, it is crucial selecting the appropriate location to open the mentioned center in order to be sure that proper amenities/spaces are available.

1.1.3. Interest

The main figure interested in this study would be any sport association or organization that intends to start or expand their business in the borough of Manhattan. With this aim, this study will focuses the attention on some econometrics and parameters that will be use classify the different neighborhoods, finding similarities between areas and evaluating the best location/s where the business can be run. We also assume that similar econometrics are available for other successful business in other part of the world (for example in city like Toronto, Paris, Berlin, Los Angeles, etc.), fact that allows the owner of the business to better understand similarities between existing and new locations, selecting the ones that better fit their requirements.

2. Data Description and Preliminary Description of the Methodology

2.1. Raw data

The data necessary to address the business problem illustrated in the previous paragraph is collected from different sources (in addition to Foursquare). In the following, a brief description of the raw data and the process to obtain it is reported.

- **Zip codes and areas**

  In this study, zip codes are used for the areas in lieu of neighborhoods. A mentioned, the only island of Manhattan is considered. With this aim, a list of 42 zip codes is compiled using the New York City health website indicated below.
  https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
  It is important to stress that only 42 zip codes have been considered in this study, despite in Manhattan there are many more. The main reason is that it is common to have very small areas, sometime just building, that have their own zip code. We have removed these records from our study, not being representative of the area. Below an image representing the zip codes considered.

Manhattan	Central Harlem	10026, 10027, 10030, 10037, 10039
	Chelsea and Clinton	10001, 10011, 10018, 10019, 10020, 10036
	East Harlem	10029, 10035
	Gramercy Park and Murray Hill	10010, 10016, 10017, 10022
	Greenwich Village and Soho	10012, 10013, 10014
	Lower Manhattan	10004, 10005, 10006, 10007, 10038, 10280
	Lower East Side	10002, 10003, 10009
	Upper East Side	10021, 10028, 10044, 10065, 10075, 10128
	Upper West Side	10023, 10024, 10025
	Inwood and Washington Heights	10031, 10032, 10033, 10034, 10040

Figure 1 – Neighborhoods and zip codes in Manhattan

- **Foursquare**

  Foursquare is used to evaluate the types of businesses in each zip code. In particular, since we are interested in locating the best area to open a multi-activity center, we define, for each zip code area, the numbers of venues to consider within a certain radius. We assume a maximum of 100 venues within a radius of 600 meters. When the results for each zip code are available, a further filtering will be necessary to just consider some specific businesses that are correlated with the sport center. For example, for each neighborhood (zip code) we count the number of gyms, the filed courts, and any other similar entities that can be correlated to a sport center. In this study the following key-words contained in the business category are considered: *Gym, Athletics, Studio, Martial Arts, Bike, Golf, Tennis, Soccer, Volleyball, Basketball, Sports Club*. The selection of these words come from a preliminary investigation of the data gathered from Foursquare. Below an example of the first 11 of 100 venues found in the 10001 zip code area. For each venue we have the venue category and the related name.

```
-------------------------------------
ZIP CODE 10001 (Similar Area no. 1 )
-------------------------------------
            Venue Category                                      Venue
0              Pizza Place                      New York Pizza Suprema
1            Dance Studio          You Should Be Dancing…..! / Club 412
2             Coffee Shop                             Bluestone Lane
3             Music Venue                               Music Choice
4       Basketball Stadium                      Madison Square Garden
5            Camera Store                            B&H Photo Video
6             Music Venue                               Hulu Theater
7       Chinese Restaurant                              Panda Express
8      Peruvian Restaurant                                      Chirp
9                   Hotel  Fairfield Inn & Suites by Marriott New York Mi...
10                Bakery                                  Magnolia Bakery
```

Figure 2 – First eleven venues in the 10001 zip code area

Among the venues listed above, the ones that are part of our business interest are filtered using the keywords aforementioned.

```
-------------------------------------
ZIP CODE 10001 (Similar Area no. 1 )
-------------------------------------
          Venue Category                                    Venue
0              Boxing Gym                    Renzo Gracie Academy
1   Gym / Fitness Center                         Fly Fitness NYC
2   Gym / Fitness Center                     Foxy Fitness and Pole
3                    Gym                  Crossfit Hell's Kitchen
4              Boxing Gym                          iLoveKickboxing
5                    Gym                    Orange Theory Fitness
6           Dance Studio       You Should Be Dancing…..! / Club 412
7           Dance Studio                      Piel Canela Dancers
8           Dance Studio                Banana Skirt Productions
9           Dance Studio                           Joel Salsa NY
10          Dance Studio                           Pearl Studios
11          Dance Studio                     Ripley-Grier Studios
12           Yoga Studio  AntiGravity® Aerial Yoga NYC Headquarters
13           Yoga Studio     Sivananda Yoga Vedanta Center New York
14   Martial Arts School  Marcelo Garcia Brazilian Jiu-Jitsu Academy
15          Tennis Court                      Midtown Tennis Club
16    Basketball Stadium                   Madison Square Garden
```

Figure 3 - Business-target related venues in the 10001 zip code area

Since all the info is available for each area, the level of saturation of the business in a specific area can be considered i.e. comparing the sport-related business to the total present in the area of interest.

- **Json file for zip code boundaries**
  A .json file was found on Internet with the zip code boundaries that will be used later for graphical porpoises to show the results. The .json file contains all the zip code of NYC. As mentioned already, only the areas in Manhattan will be considered in this study.

- **Income**
  Incomes grouped by zip code are considered as a proxy for the "wealth" of the area. In particular the *Median Household Income*, *Average Household Income*, and *Per-Capita Income* are considered with this aim. The data is collected from the web site https://www.incomebyzipcode.com/newyork/10026 (here referring to zip code 10026). This data will be transformed to generate econometrics as described in the following sections.

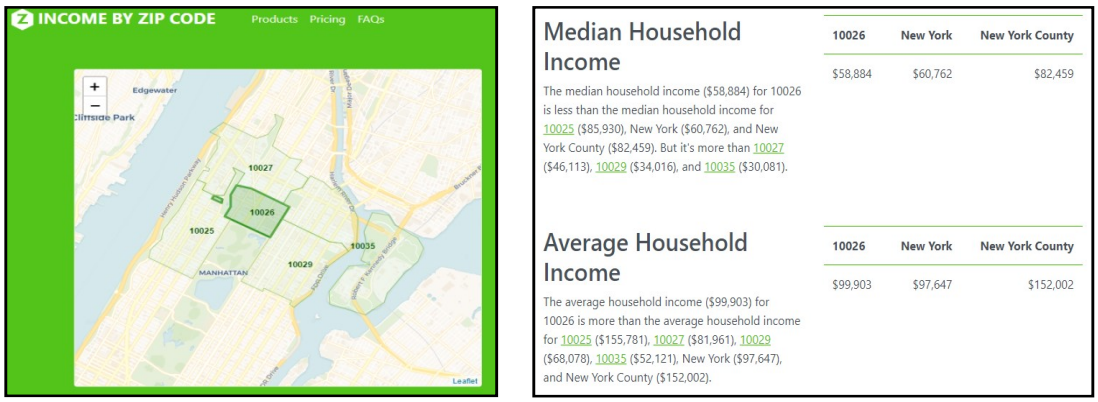

Figure 4 – Income per zip code. Map (left) and data (right)

- **Csv file for area land by zip code**
  In the website below a csv file was downloaded. The file contain the area land, in square mile, organized by zip code. This file will be imported as DataFrame and the data will manipulated as indicated in the following sections.
  https://blog.splitwise.com/2014/01/06/free-us-population-density-and-unemployment-rate-by-zip-code/

- **Demographic by zip codes**
  From the web site https://www.zipdatamaps.com/10026 several statistic can be collected or each zip code. In the figure below, an example of the webpage for the zip code 10026 is proposed.

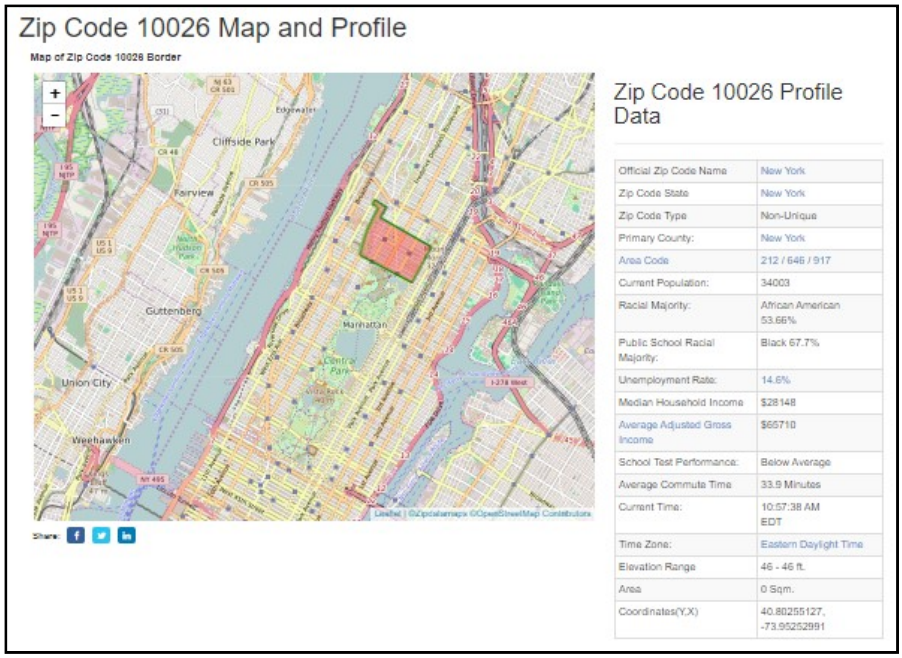

Figure 5 – ZipDataMaps web site

Using pandas, among all the info available on the website, the following are considered in this study: *Current Population*, *School Test Performance*, *Coordinate, Average Real Estate Asking Price, Average Real Estate Sale Price*.

When all the preliminary data mentioned above is available, a DataFrame is generated to organize the info in a structure way. In the figure below the first 5 rows of the DataFrame is shown for clarity.

	ZIP Code	String ZIP Code	Latitude	Longitude	Population	Area	Population Density	Median Household Income	Average Household Income	Per-Capita Income	Average Real Estate Asking Price	Average Real Estate Sale Price	Test Score	Numeric Test Score	Sport Business Ratio
0	10001	10001	40.751	-73.997	21102	0.614	34368.0	88526.0	151628.0	84765.0	1.604e+06	1.823e+06	Above Average	4	17
1	10002	10002	40.716	-73.986	81410	0.878	92722.0	35859.0	68315.0	32694.0	1.195e+06	5.530e+05	Average	3	2
2	10003	10003	40.732	-73.989	56024	0.576	97264.0	112131.0	189885.0	92781.0	1.588e+06	1.647e+06	Above Average	4	5
3	10004	10004	40.704	-74.014	3089	0.560	5516.0	157645.0	218650.0	122165.0	8.783e+05	6.928e+05	Poor	1	4
4	10005	10005	40.706	-74.008	7135	0.071	100493.0	173333.0	208186.0	106702.0	9.958e+05	7.306e+05	Poor	1	7

Figure 6 - DataFrame with collected raw data

2.2. Data used in this study

The sources and the methodology used to collect the raw data are illustrated above. Since the info comes from different sources and in different formats, it is necessary to organized it in a structured way. Furthermore, not all the info gathered will be used directly i.e. some of it will be just employed to generate econometrics. For each zip code, the following info will be considered in this study:

- **Zip code** – primary key to locate the area. This info is an integer. An additional column is considered to also have the zip code in the form of string.
- **Latitude and Longitude** – geographic coordinates are used to locate the area associated to each zip code. Locations have been mainly employed to gather info from Foursquare database and for the post processing of the data to show the results on maps. This info is also associated to the json file used with this purpose.
- **Population** – number of people that live in a particular area.
- **Area** – area land in square miles of the neighborhood identified by zip codes.
- **Population Density** – defined as the ratio between the population and area.
- **Median Household Income** – used as a proxy of the "wealth of the area". The main assumption is that the higher the income of an household, the wealthier the area. This parameter is not considered alone thought. Per-Capita and Average Household income are employed in the analysis as well. All the three parameters together give a better representation of the distribution of the income in each zip code area.
- **Average Household Income** – used as a proxy of the "wealth of the area" in concert with the other two parameters mentioned above.
- **Per-Capita Income** – used as a proxy of the "wealth of the area" in concert with the other two parameters mentioned above.
- **Average Real Estate Asking Price** – another indicator used to represent the "wealth" of an area in term of real estate. This parameter can be considered as a proxy of the cost related to buy or rent a place in a specific area
- **Average Real Estate Sale Price** – another indicator used to represent the "wealth" of an area and considered to be more representative of the current real estate value since comes from recent transactions.
- **School Test Performance** – this parameter is used to characterize the area in case there is a substantial concentration of families. The main assumption is that the target audience of the business we are evaluating would be families with wealthy parent/s that invest significant economic resources in their children activities, in addition to their own too.. The main idea is that if a student perform well (on average) in school, it is also likely that the parents might encourage their children to engage extracurricular activities (like sports) as well. As already mentioned, these activities are associated to an high cost that most probably will be afford by high-end income segment of the population. This parameter is also translated in numeric format to be manipulated later in the analysis.
- **Sport Business Ratio** – Using Foursquare, when the zip code coordinates are known, an investigation of the activities in the area can be performed. In particular, businesses like gyms, sport centers, courts, etc., will be considered for each area, considering a fixed limit for venues and radius. This represents the "level of saturation" of the offer/demand in a specific area and it can be estimated as the ratio of the sport-related businesses divided by the total number of businesses.

2.3. Additional consideration of econometrics used

Regarding the parameters introduced above, additional clarifications on the assumptions made are reported in the following.

2.3.1. Income parameters:

As mentioned above, median and average household income, and per-capita income are considered as a proxy of the wealth of a specific area. Those parameters are obviously related and they represent the income distribution among a certain population. Generally speaking, when median and average household income are approximately the same, this represents a sigh that there may be an uniform distribution of the wealth. In case that the median is greater than average, this may indicate that there is a concentration of high-income people in the area in comparison to their peers. On the contrary, where the median is lower than average, sign that there may be a concentration of low or mid-income people in the area. It is important to mentioned that the distribution of the income should be based on the standard deviation that is an indicator of the income spread. In this study this is neglected and the factor defined above, as mentioned, is used just a proxy for a preliminary representation of the area. It is important to notice that also the relationship between the average household income and the per-capita one might give a preliminary indication of the type of population. In order to clarify the relationships between these two numbers, let's considered the following example. Per-Capita Income is calculated dividing the total income in a zip code area by its population. If the total income is $1 million (made up numbers just to explain the point), and the zip code area count 200 people, the per-capita income will be $1,000,000 / 200 = $5,000 per person. If there are only 50 households, the average income would be $1,000,000 / 50 = $20,000 per household. In this example, their ratio will be $20,000 divided by $5,000 and equal to 4, meaning that, on average, each household correspond to 4 people. This might mean that in the area there is a "average" family where one of the parent is the household, and the other parent and two children are the dependants. On the contrary, if this factor is close to one (one is the minimum value possible by definition), it means that the area might be characterized by singles or married couples without children that file tax separately.

2.3.2. Real estate parameters

Average Real Estate Asking Price and Average Real Estate Sale Price are considered as a proxy for the real estate market condition of the area. When coupled together, these may represent different situations. For example, when the asking price is higher that the sale one, it may show the it may be "difficult" moving to the area since the real estate on sale is very expensive and only few units (affordable ones) are actually sold. On the other side of the spectrum, i.e when the asking price and the selling price are very close, it may represent the fact that real estate market in the area is pretty alive and there may show a trend that several people wants to move to the area, being also a positive incentive to open a business there.

2.4. How data will be used to address the business problem

When a clean and structured data is available, a categorization of the different zip code areas can be performed using the K-Means clustering method. Each record, associated to each zip code, will be characterized by the parameters listed in the previous section. Those parameters will be normalized to have them within the zero to one range.

The main scope of this study is to cluster zip code areas with similar characteristics, trying different number of clusters, in order to have an understanding of where it may be more

convenient expand the business. This can be achieved assuming that similar info is already available in other cities that have similar characteristics and they already proved to be successful. In this way, by comparison, we can find the most-similar neighborhood in Manhattan where there will be reasonably good chances of having a satisfying economic return. Before this final steps, preliminary exploration of the date will be performed to better understand the input and spot possible correlation among the different parameters.

By comparison, when the most similar neighborhood/s are identified, venues from Foursquare are considered to better understand the type of area and select the best-fit place where to start the business.

3. Methodology

3.1. Preliminary exploration of the raw data

Once the raw data mentioned above is collected, we perform a preliminary exploratory analysis to see if there is any correlation of the info gathered. In particular, we consider the parameters below and a correlation matrix is generated. Only numerical values of the correlation factor greater that 0.5 or lower that -0.5 are considered. Note that "reds" represents positive correlation, "blues" indicate negative correlation. In both cases, heavier the color higher the correlation. This preliminary exploration clearly show the heterogeneous nature of the island on Manhattan that, despite its limited land extension, contains a generous variety of characteristics. The results is that, generally speaking each zip code areas is pretty diverse under several aspects.

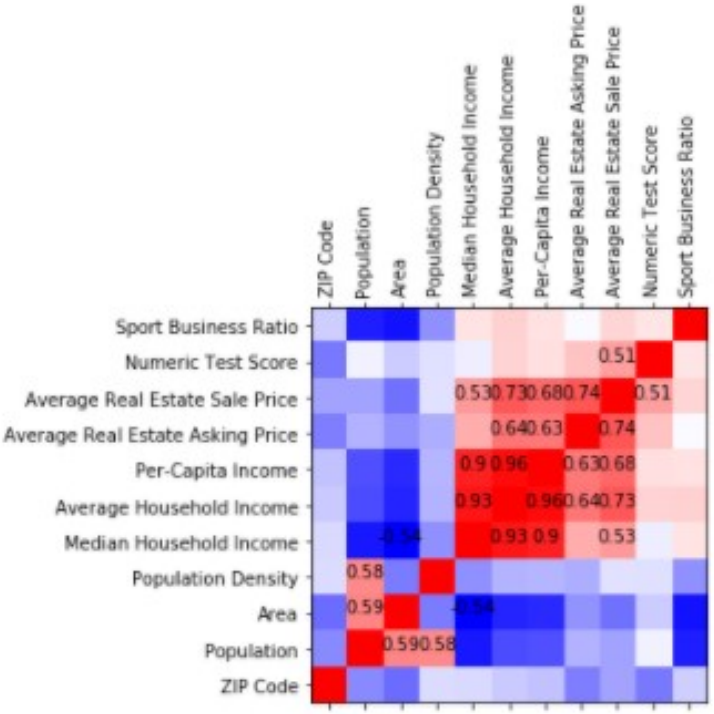

Figure 7 - Correlation matrix

In the figure below the most significant correlations among the data are reported. As the results show, there are strong and obvious correlations between *average household*, *median household* and *per-capita income*. It is the same for *real estate asking* and *sale price*. There is another obvious correlation between the *average real estate sell price* and the *average household income*.

```
Number of unique correlations > 0.5 or < -0.5 = 13
-----------------------------------------------------------------------------------------
Case #   Corr           Parameters
-----------------------------------------------------------------------------------------
1        -0.54          Area  vs.  Median Household Income
2        0.513          Average Real Estate Sale Price  vs.  Numeric Test Score
3        0.532          Average Real Estate Sale Price  vs.  Median Household Income
4        0.584          Population  vs.  Population Density
5        0.595          Area  vs.  Population
6        0.629          Average Real Estate Asking Price  vs.  Per-Capita Income
7        0.641          Average Household Income  vs.  Average Real Estate Asking Price
8        0.685          Average Real Estate Sale Price  vs.  Per-Capita Income
9        0.73           Average Household Income  vs.  Average Real Estate Sale Price
10       0.739          Average Real Estate Asking Price  vs.  Average Real Estate Sale Price
11       0.897          Median Household Income  vs.  Per-Capita Income
12       0.929          Average Household Income  vs.  Median Household Income
13       0.958          Average Household Income  vs.  Per-Capita Income
-----------------------------------------------------------------------------------------
```

Figure 8 – Most significant correlations (with correlation factor less that -0.5 or greater than 0.5)

Even if very week (-0.54), a correlation between the *area* of each zip code and the *median household income* is recognizable. It appears that the people with higher level of income try to live in smaller areas. This can be partially explained by the fact that high-income people try to live in high-rise and skyscrapers, fact that allows a significant number of people in a relative small area.

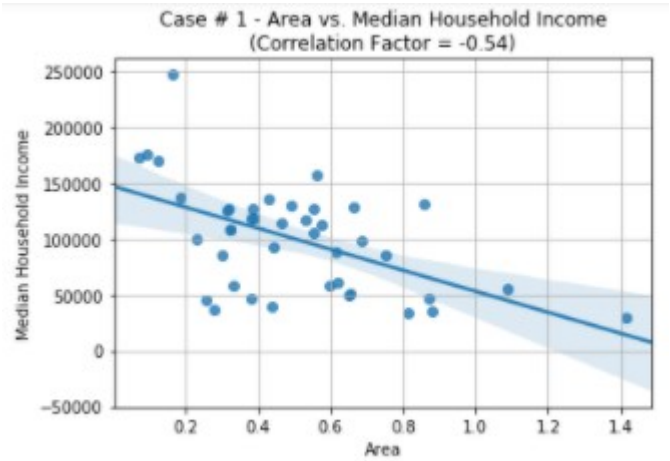

Figure 9 - Median Household Income versus Area correlation

Another very week correlation, but worth of notice, is that the number of sport related businesses tries to concentrate on zip core area with less extension. Then, there is no correlation between this ratio and the population density, which allows us to drop the hypothesis that level of saturation of sport business tend to be higher in more dense populated area.

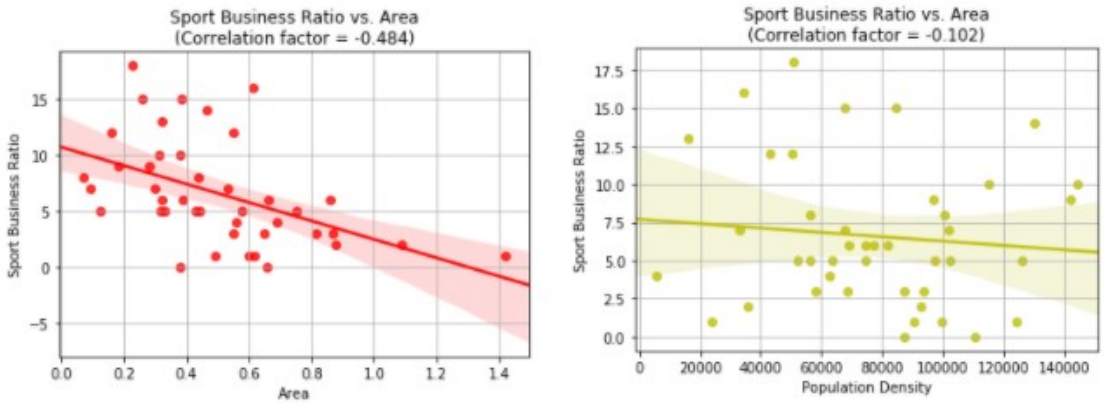

Figure 10 - Sport Business Ratio to Area (left) and to Population Density (right) correlation

3.2. **Representation of the results**

After a preliminary analysis of the data is performed, the collected data is presented in a graphical way to better explain the findings. A general map is provided and marker with the major info are reported. An example for the 10028 zip code is shown.

Figure 11- Map of Manhattan where the main info for each zip code are proposed

3.2.1. Population data

The maps below report the findings in terms of population and population density. It is clear to identify the most populate zip code areas in the island. It can be noticed that not necessarily a very populated area is the one that is also have the higher density.

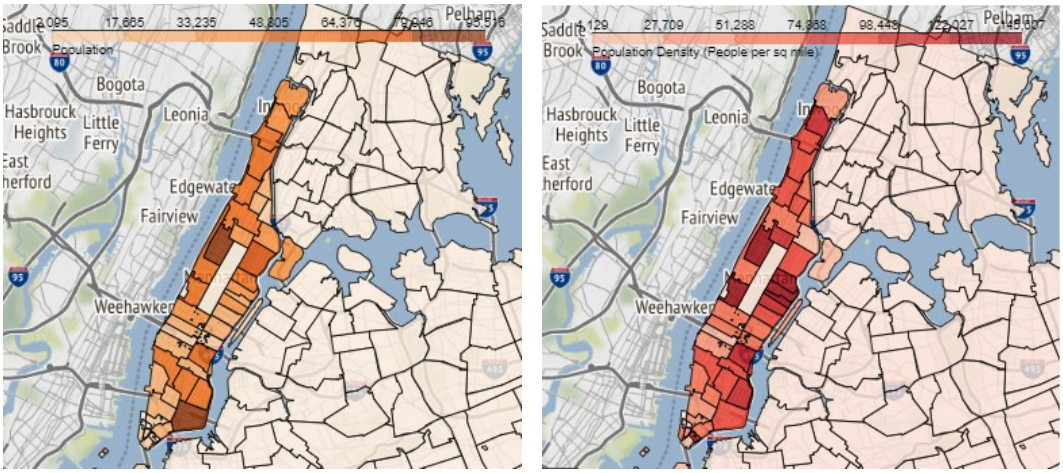

Figure 12 – Population (left) and Population Density per zip code

3.2.2. Income data

The maps below report the findings in terms of income. In particular, only the median household and per-capita income are considered.

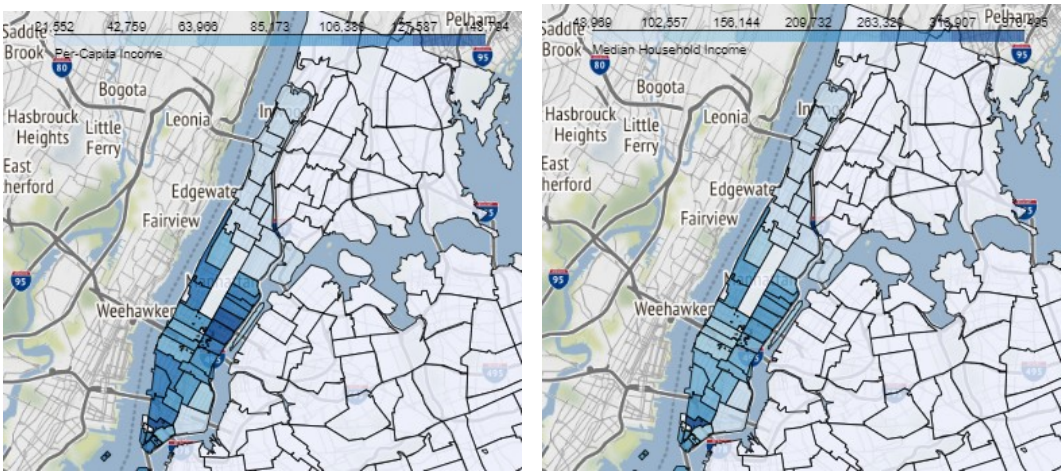

Figure 13 – Per-Capita (left) and Median Household Income (right)

3.2.3. Real estate data

The maps below report the findings in terms of average real estate asking and sell price

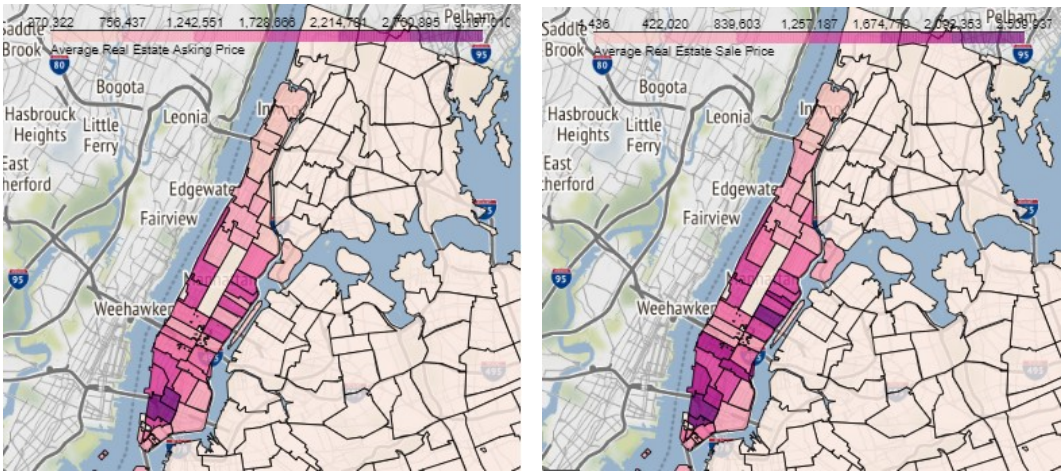

Figure 14 – Average Real Estate Asking (left) and Sell (right) Price

3.2.4. School test score data

The map below reports the findings for the test score results, where 1 is "poor", 2 is "below average", 3 is "average", 4 is "above average" and 5 is "excellent".

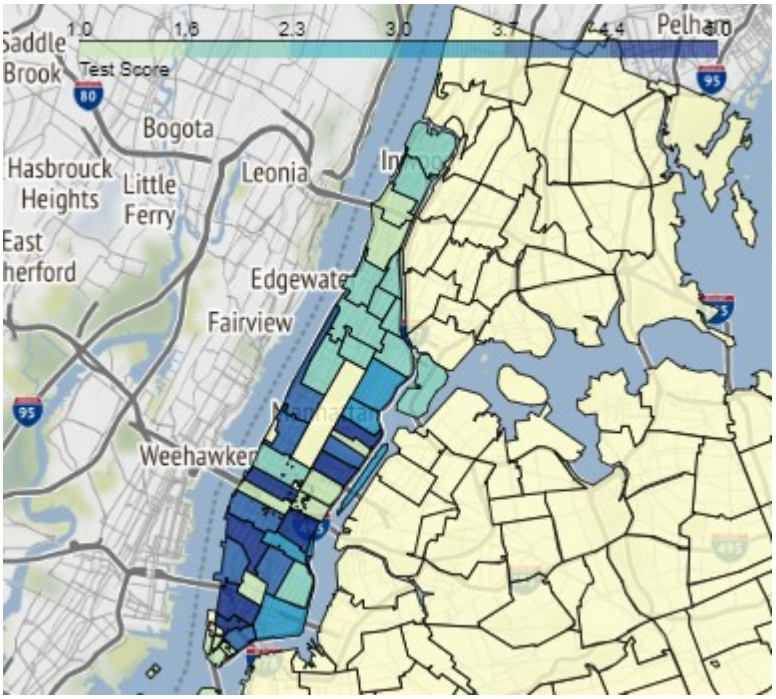

Figure 15 – Test Score Results

3.2.5. Sport business ratio

The map below reports the percentage of sport-related businesses (gym, studio, etc.) per each zip code. This info gives a preliminary information of the presence of the type of business we are interested in.

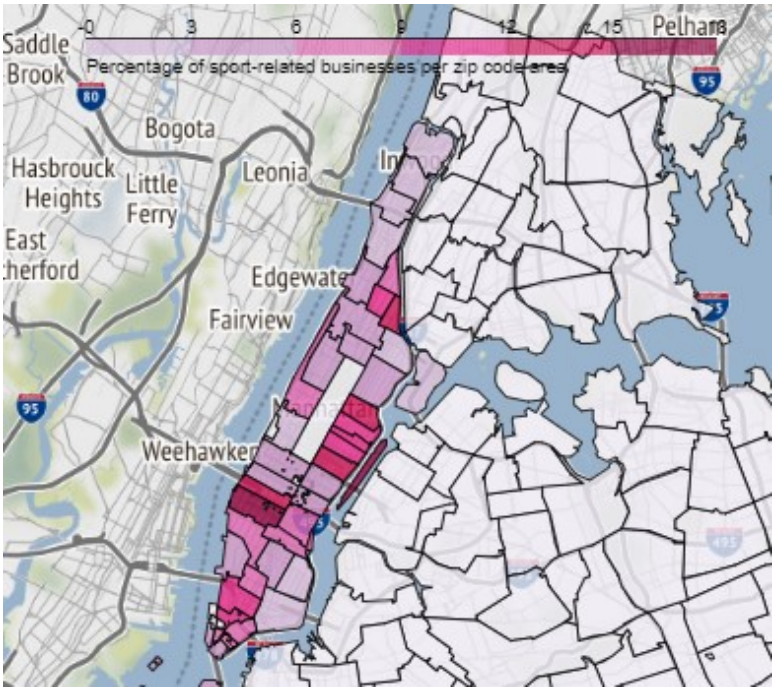

Figure 16 – Sport business ratio

3.3. Parameters used for clustering

Among the data available, the most significant parameters are used to characterized the different zip codes areas. In particular, *Population*, *Population Density*, *Median* and *Average Household Income*, *Per-Capita Income*, *Average Real Estate Sale Price*, *Numeric Test Score* and *Sport Business Ratio* are considered as main parameters for clustering using K-mean method. Below, the first 10 rows of the DataFrame used for clustering is shown. Each row is consistent with the original DataFrame and it correspond to a specific zip code, but not reported here.

	Population	Population Density	Median Household Income	Average Household Income	Per-Capita Income	Average Real Estate Sale Price	Numeric Test Score	Sport Business Ratio
0	21102	34368.0	88526.0	151628.0	84765.0	1.823e+06	4	17
1	81410	92722.0	35859.0	68315.0	32694.0	5.530e+05	3	2
2	56024	97264.0	112131.0	189885.0	92781.0	1.647e+06	4	5
3	3089	5516.0	157645.0	218650.0	122165.0	6.928e+05	1	4
4	7135	100493.0	173333.0	208186.0	106702.0	7.306e+05	1	7
5	3011	32728.0	176250.0	214543.0	114611.0	9.320e+05	1	6
6	6988	43136.0	246813.0	367343.0	147547.0	2.350e+06	5	12
7	61347	99428.0	61548.0	95947.0	51655.0	7.805e+05	2	1
8	31834	81626.0	117923.0	188064.0	93605.0	1.870e+06	3	6
9	50984	77132.0	128613.0	207287.0	121991.0	1.649e+06	5	7

Figure 17 – Parameters used for clustering

The data indicated above is also normalized. Each value is within 0 and 1 since the data is scaled considering minimum and maximum values for each parameter. Below the normalized data.

	Population	Population Density	Median Household Income	Average Household Income	Per-Capita Income	Average Real Estate Sale Price	Numeric Test Score	Sport Business Ratio
0	0.198	0.208	0.270	0.316	0.497	0.730	0.75	0.889
1	0.856	0.629	0.027	0.051	0.079	0.213	0.50	0.111
2	0.579	0.661	0.379	0.437	0.561	0.659	0.75	0.278
3	0.001	0.000	0.589	0.528	0.797	0.270	0.00	0.222
4	0.045	0.685	0.661	0.495	0.673	0.286	0.00	0.444
5	0.000	0.196	0.674	0.515	0.736	0.368	0.00	0.389
6	0.043	0.271	1.000	1.000	1.000	0.945	1.00	0.667
7	0.637	0.677	0.145	0.139	0.231	0.306	0.25	0.056
8	0.315	0.549	0.405	0.431	0.568	0.750	0.50	0.333
9	0.524	0.516	0.455	0.492	0.795	0.659	1.00	0.333

Figure 18 – Normalized parameters used for clustering

In the following, the main statistical info related to the normalized parameters used in the clustering process is shown.

	Population	Population Density	Median Household Income	Average Household Income	Per-Capita Income	Average Real Estate Sale Price	Numeric Test Score	Sport Business Ratio
count	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000
mean	0.374119	0.517690	0.322381	0.337905	0.466929	0.430857	0.422619	0.366476
std	0.243975	0.238999	0.219517	0.234975	0.323854	0.257815	0.364123	0.259079
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.206000	0.364000	0.118500	0.080000	0.079000	0.232500	0.062500	0.167000
50%	0.328500	0.507000	0.356000	0.357500	0.551500	0.374000	0.250000	0.305500
75%	0.574000	0.673000	0.449750	0.523250	0.763750	0.601750	0.750000	0.500000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 19 – Statistical information for the normalized parameters used for clustering

3.4. K-Mean method and optima number of clusters

K-Mean clustering procedure requires to define the number of clusters to be used in the analysis. An iterative preliminary analysis, considering a minimum of 3 to a maximum of 10 clusters, is performed to identify the best value to use. The figure below represents, for each value of clusters assumed, average, minimum and maximum number of items in each cluster. Percentile of 25 and 75 are also included. In addition, the standard deviation is reported too. The main idea is selecting the number of clusters in a way that each cluster contains more or less the same numbers of items. Form the graph, we decide to use for the 42 zip area codes a total of 5 clusters since it seems that is the number that better distribute the items among the clusters.

When the number of clusters is defined, we can analyze the normalized data to classify all 42 zip area codes.

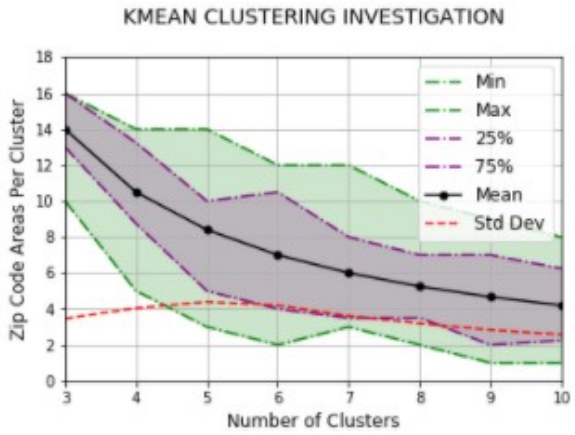

Figure 20 –Evaluation on the optimal number of clusters to consider in K-Mean method

4. Results of clustering using K-Mean method

The map below shows the results of clustering process based on the parameters described above. It appears that the cluster are location independent, i.e. not necessarily the same type of zip code areas are gathered together, characteristic that shows again the heterogeneous nature of the island of Manhattan.

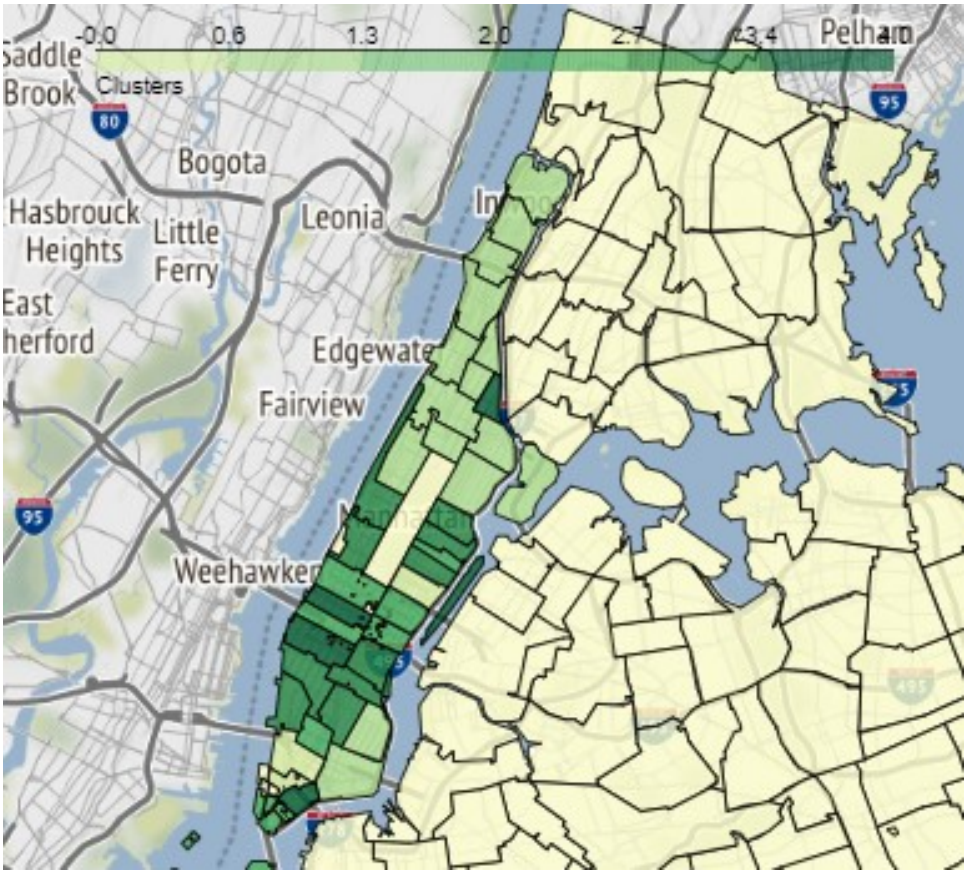

Figure 21 – Results of the classification of the zip code areas using K-Mean method

The figure below reports the average of each parameter to provide at glance a better idea of each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Population	2.231933e+04	52933.785714	16789.7	49449.0	19038.0	3.727638e+04
Population Density	5.932533e+04	83588.857143	62412.9	99865.0	55351.4	7.732752e+04
Median Household Income	1.598767e+05	49290.357143	139387.1	122183.1	82400.4	9.993807e+04
Average Household Income	2.900497e+05	78941.857143	201042.9	206549.1	122253.2	1.586315e+05
Per-Capita Income	1.294887e+05	34106.142857	110643.6	111765.3	62774.2	8.104552e+04
Average Real Estate Sale Price	2.394282e+06	593094.928571	1072727.7	1528863.7	832568.8	1.087260e+06
Numeric Test Score	5.000000e+00	2.000000	1.1	4.1	3.6	2.690476e+00
Sport Business Ratio	1.300000e+01	3.071429	6.6	6.8	12.2	6.595238e+00

Figure 22 – Results of the classification of the zip code areas using K-Mean method
Average values for each cluster

The data above can be normalized to the average of all info considered. In particular each entry is divided by the average value, for each parameter. For example, Cluster #1 has a population that is 1.21 time the average, while the Cluster #2 has 1.54 time the average, making it the one characterized by the large population per zip code. In the same way we can evaluate the other values.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Population	1.21	1.54	0.48	0.46	0.59
Population Density	1.21	1.10	0.87	0.74	0.69
Median Household Income	1.17	0.51	1.43	0.66	1.42
Average Household Income	1.23	0.51	1.30	0.60	1.61
Per-Capita Income	1.30	0.43	1.40	0.56	1.46
Average Real Estate Sale Price	1.28	0.56	1.04	0.46	2.07
Numeric Test Score	1.55	0.74	0.41	0.74	1.77
Sport Business Ratio	0.99	0.37	0.82	2.00	2.06

Figure 23 – Normalized results of the classification of the zip code areas using K-Mean method

5. Recommendations

Since the zip code areas in Manhattan have been classified, and the related characteristics normalized, it is possible to use existing data coming from other locations where an owner or investor already has a successful business. These parameters are consistent with the data indicated in Figure 18. In particular we have as existing normalized data:

Population = 0.05 – Population density = 0.30,
Median Household Income = 0.90 – Average Household Income 0.95 – Per-Capita Income = 0.85
Average Real Estate Sale Price = 0.95
Numeric Test Score = 0.85
Sport Business Ration = 0.70

With this info, still using the K-Mean model we used to train the data, we can predict what cluster this existing data belong too, provide meaningful information to where a business should be open.

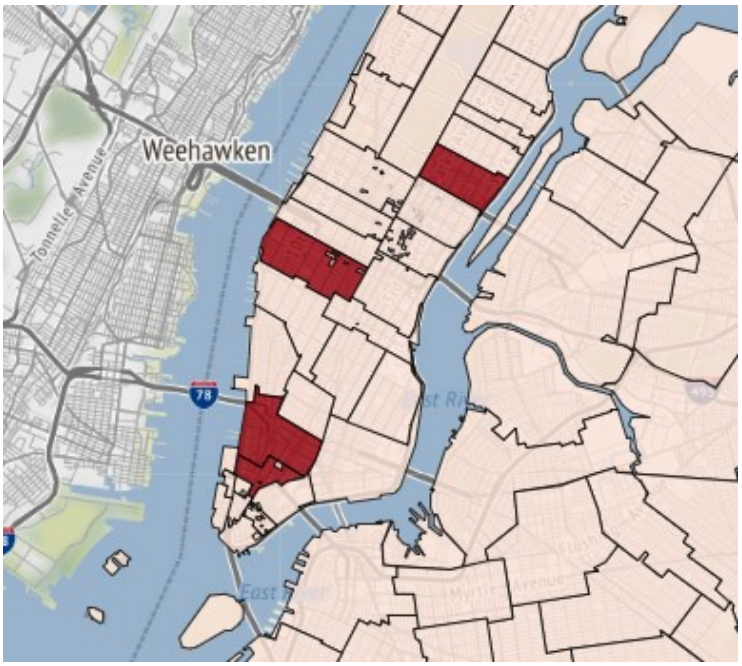

Figure 24 – Zip code area that are similar to the data of an existing successful business

For each of the four zip code areas that are similar to the existing business condition, it is possible to get from Foursquare the name and the categories of the venues. Below the venues for the first cluster are listed (first 15 ones our of 100). The same can be done for the other three.

```
-------------------------------------
ZIP CODE 10001 (Similar Area no. 1 )
-------------------------------------
```

	Venue Category	Venue
0	Pizza Place	New York Pizza Suprema
1	Dance Studio	You Should Be Dancing…..! / Club 412
2	Coffee Shop	Bluestone Lane
3	Music Venue	Music Choice
4	Basketball Stadium	Madison Square Garden
5	Camera Store	B&H Photo Video
6	Music Venue	Hulu Theater
7	Chinese Restaurant	Panda Express
8	Peruvian Restaurant	Chirp
9	Hotel	Fairfield Inn & Suites by Marriott New York Mi...
10	Bakery	Magnolia Bakery
11	Lounge	Delta Sky360° Club
12	Indie Theater	Magnet Theater
13	Music Store	Sam Ash Music
14	Other Great Outdoors	US Post Office Stairs

Figure 25 – Venues for the first cluster

Among all the venues, we can also show the ones that are sport-related, giving a better understanding of what king of sport activities are commonly practiced in the area. Looking at the results, it seems that in the area dance studios and gym/fit centers are very popular.

```
-------------------------------------
ZIP CODE 10001 (Similar Area no. 1 )
-------------------------------------
          Venue Category                                  Venue
0             Boxing Gym                  Renzo Gracie Academy
1   Gym / Fitness Center                       Fly Fitness NYC
2   Gym / Fitness Center                  Foxy Fitness and Pole
3                    Gym                Crossfit Hell's Kitchen
4             Boxing Gym                        iLoveKickboxing
5                    Gym                  Orange Theory Fitness
6           Dance Studio  You Should Be Dancing…..! / Club 412
7           Dance Studio                    Piel Canela Dancers
8           Dance Studio               Banana Skirt Productions
9           Dance Studio                           Joel Salsa NY
10          Dance Studio                           Pearl Studios
11          Dance Studio                  Ripley-Grier Studios
12           Yoga Studio   AntiGravity® Aerial Yoga NYC Headquarters
13           Yoga Studio      Sivananda Yoga Vedanta Center New York
14  Martial Arts School  Marcelo Garcia Brazilian Jiu-Jitsu Academy
15          Tennis Court                   Midtown Tennis Club
16   Basketball Stadium                  Madison Square Garden
```

Figure 26 – Sport-related venues for the first cluster

6. Conclusions

The study proposed above can be useful to provide preliminary indication of where a new sport-related business can be open. The parameters included in this study are not only venue related, but also other econometrics are used to better identify the characteristics of each zip code area. This study can be easily expanded adding other parameters that may provide a more accurate results and help with a more precise classification of the area of interest. Furthermore, the same approach can be easily applied to other cities, being the methodology suitable for this kind of investigations.

7. References

All the references are saved in the https://github.com/EnrTom/Coursera_Capstone repository In particular:

- ***CapstoneProjectNotebook.ipynb*** – Jupiter notebook used to analyze and process data.
- ***CapstoneProjectPresentation.pdf*** – Presentation with data considered and findings in the report.
- ***CapstoneProjectReport.pdf*** – This report is also saved in the repository indicated above