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

Enrico Tomassoli, September 4th, 2020

1. Introduction

The part that pertains to the introduction/business problem considered in this project is reported in the first part of the preliminary report. Please refer to the link in the Coursera submission

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.

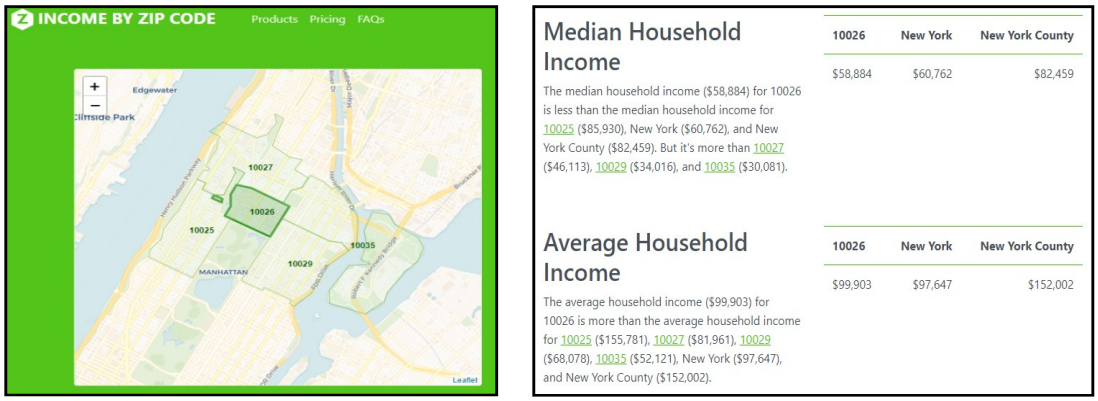

Figure 4 – Income per zip code. Map (left) and data (right)

- **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.

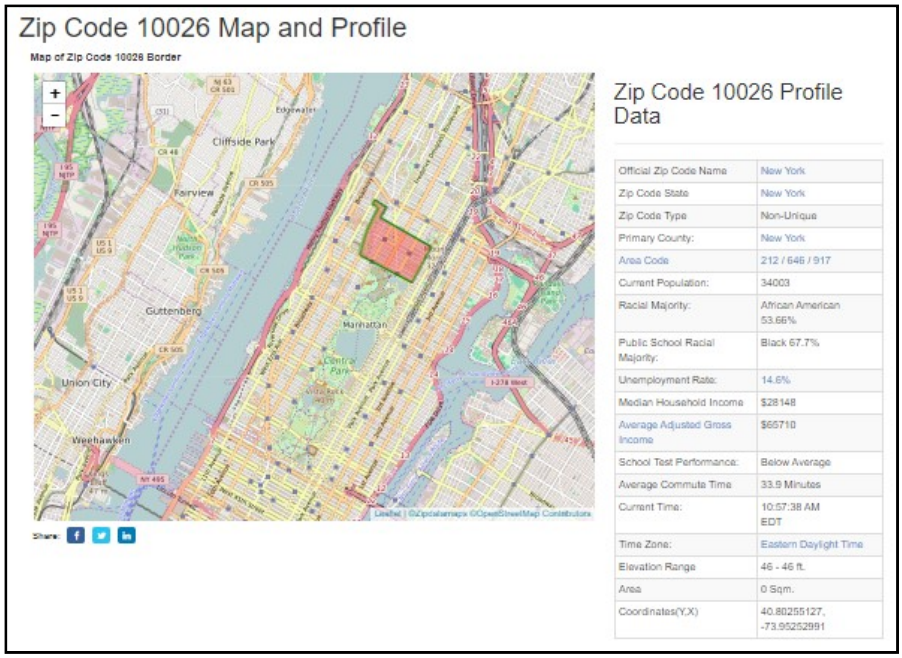

Figure 5 – ZipDataMaps web site

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.