Week

5

LUCA BENINI

Applied Data Science Capstone

New Gym in New York

APPLIED DATA SCIENCE CAPSTONE

New Gym in New York

Table of Contents

Chapter 1

Introduction

Where we discuss the business problem and who would be interested in this project.

The fitness industry has become hyper-competitive and becoming a successful gym owner is getting harder and harder. In fact, while there is massive potential in owning a fitness business, 8 out of 10 of them will fail in their first year.
To avoid transforming the dream of opening a new gym in a money sink hole it's important to start from the beginning with the right foot and in this report we are going to look on how to select the best location for a new gym in New York

Background

The first step is to identify the requirements for the best location, for this step we have identified an interesting article on the location requirements for identifying the best location for a new gym. The Article is written and published by a renowned source of articles about business development for fitness. The Personal Trainer Development Center or PTDC is a premier source of free information for smart and passionate fitness professionals.

The PTDC archive contains more than 1,000 articles on the practice and business of fitness, written by hundreds of experts from all corners of the globe.

The article identifies 10 main attributes for a new location:

1. High Ceilings & Natural Light
2. Relatively New" Construction
3. Clean Sight Lines
4. Ample Waiting Space for Parents
5. Sufficient Parking
6. Community Restrooms
7. Building Property Manager
8. Availability of Signage
9. ***Minimal local competition***
10. ***Appropriate Adjacent Tenants***

While point 1 to 8 focus on a single lot the last two steps requires an analysis of the neighborhood. Specifically, we want to find a neighborhood that is very good for opening a gym, but where there is no or little gym.

This seems a contradiction; can Data Science help with this step?

Problem

We will proceed by clustering the various areas of the city by similarity in terms of venues, income and population
We will identify which is the cluster with highest gym count: where the gym businesses are more profitable.
In the most gym-profitable cluster, we will locate the suggested candidates: area similar to "high profitable" but with the lowest gym count

Interest

This report can be of interest not only for an entity that what to open a new gym, but also for current gym owner that want to move the business to a more profitable location

Chapter 2

Data

Where we describe the data that will be used to solve the problem and the source of the data.

To save time in the future, print a copy of this document. Click **Print** on the **File** menu, and press ENTER to receive all eight pages of examples and instructions. With the printed document in hand, position yourself in normal view to see the style names next to each paragraph. Scroll through the document, and write the style names next to the paragraphs (press CTRL+HOME to reposition yourself at the beginning of the document).

Overview

We are going to use multiple data sources to cover different aspect of the problem:

1. **Geo-space Data**: The granularity of this study will be a Zip Code area, we are going to use the geometric shape to plot the information in an endearing style. We will use the geographical shape of the administrative area to identify the centroid of the area, that represents the ideal "middle point". These data are available online.

2. **Census Data**: It's important that we take in consideration the number of possible customer, for this reason we are going to use the information about the residents of each Zip Area. The Census data are also available online

3. **Income Data**: The number of possible customer is not enough alone, for this reason we are going to use also the information about the income of each Zip. The income data are provided by IRS divided by zip code and available online the fields and the data are described on the help document

4. **Venues Data**: For the statistics and the information about the various venues in a given Zip we will use the FourSquare API

In the following section we are going to cover for each aspect the data source and the cleaning operation in details.

Geo-space Data

Data Source

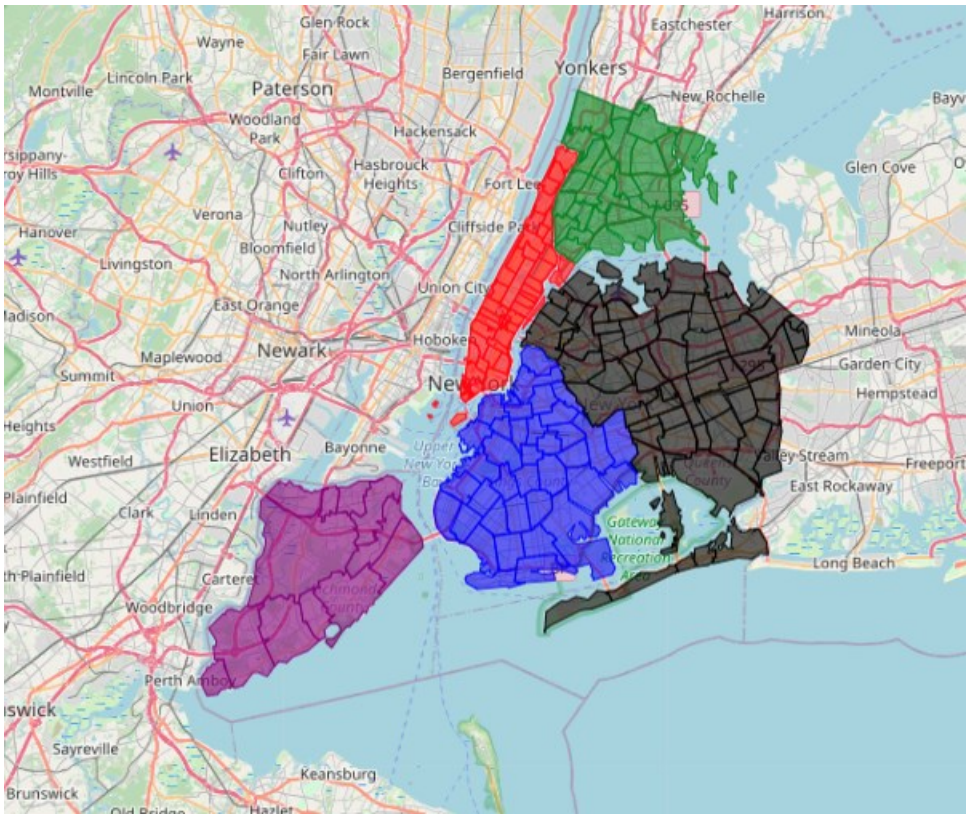
We want to have a plotting that is nicer than simple marker on a map, for this reason, we need the geojson file of the geometries of the various New York City zip codes. Fortunately, also these data are available online

We want to have a greater control on the geographic data compared to what is presented in this course, for this reason we will use geopandas to manipulate the geographic data.

GeoPandas is an open source project to make working with geospatial data in python easier. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types. Geometric operations are performed by shapely. Geopandas further depends on fiona for file access and descartes and matplotlib for plotting.

Data Cleaning

The information contained are good to go and provides us with a good source for Latitude, Longitude, Area and shape for each Zip code that can be plotted with folium



Census Data

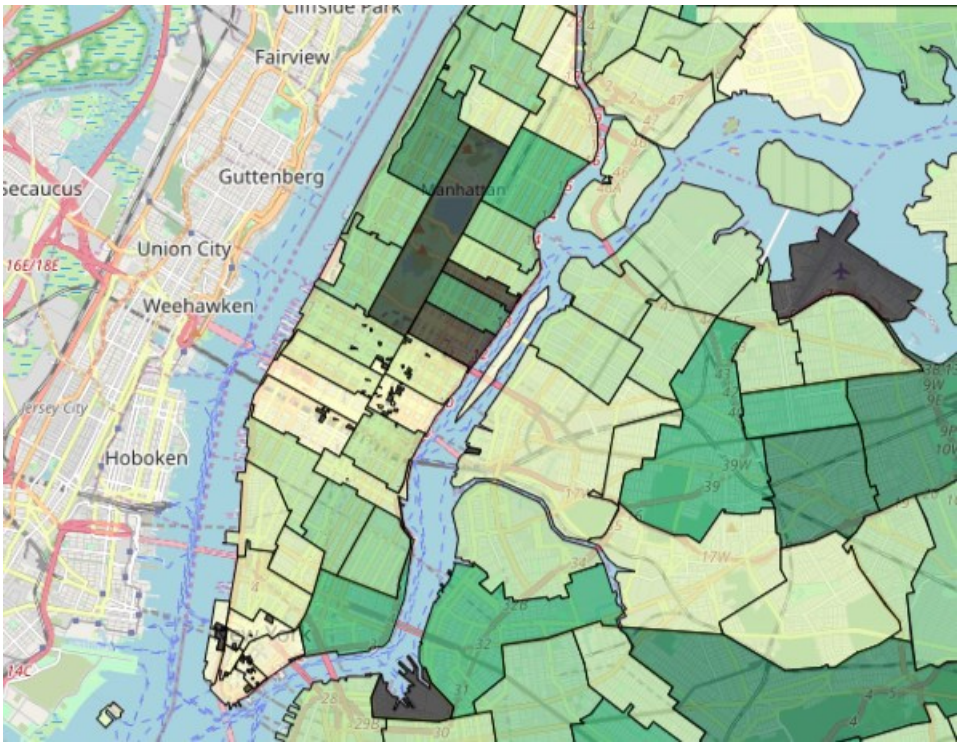
Data Source

The Census data are also available online

The data is distributed on 16 independent pages from http://zipatlas.com/us/ny/zip-code-comparison/population-density.htm to http://zipatlas.com/us/ny/zip-code-comparison/population-density.16.htm and contains also the geographic center of each zip code. For the scraping of the page we have used BeatifulSoup.

Data Cleaning

The majority of the information where good to go, but upon ploting the information has been easy to identify some criticality:



The areas in black have no census information, this missing information can be divided in three category:

1. No census information can exists: These are the codes that referrers to place where no one lives: Airport, Docks and Parks: These area can have an individual Zip Code, but have no residents

2. No census information published: There are the codes that for postal or historical reason refers to a very small location (a building). For privacy reason the census (and we will see also the income) data for these location are published "merged" in the nearest or surrounding zip code

3. Information Missing from our source: the analysis have revealed some Zip Code for which the census information should be available but it's not present in our data sources.

For Zip Code in (1) and (2) no special activity are needed: they are going to disappears from our dataset when we merge the data of the Income.

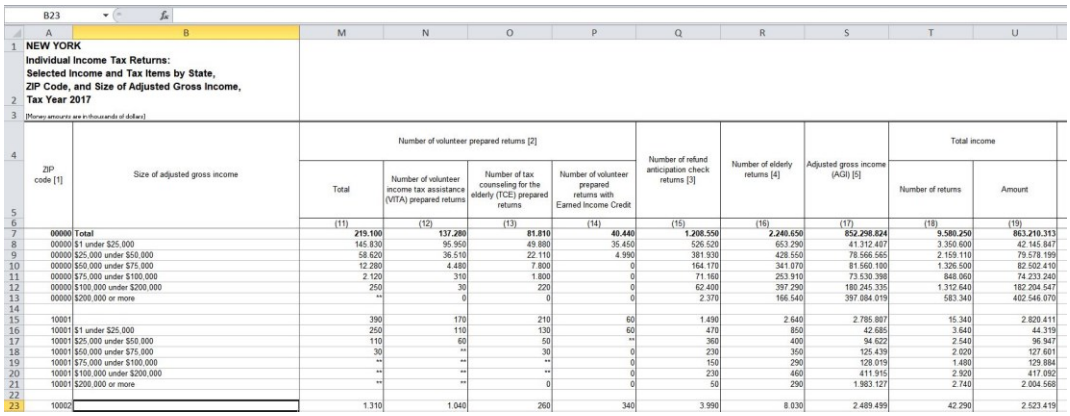
For the items in (3) we have used the data from https://www.zip-codes.com/ with the appropriated scraping code. Should be noted that the original data source was missing the information for only 3 "real" zip code.

Income Data

Data Source

For the Income for Zip Code we will use the data provided by IRS available online the fields and the data are described on the help document

The Excel file structure is human oriented and for this reason quite a lot data wrangling will be required.



NEW YORK
Individual Income Tax Returns:
Selected Income and Tax Items by State,
ZIP Code, and Size of Adjusted Gross Income,
Tax Year 2017

(Money amounts are in thousands of dollars)

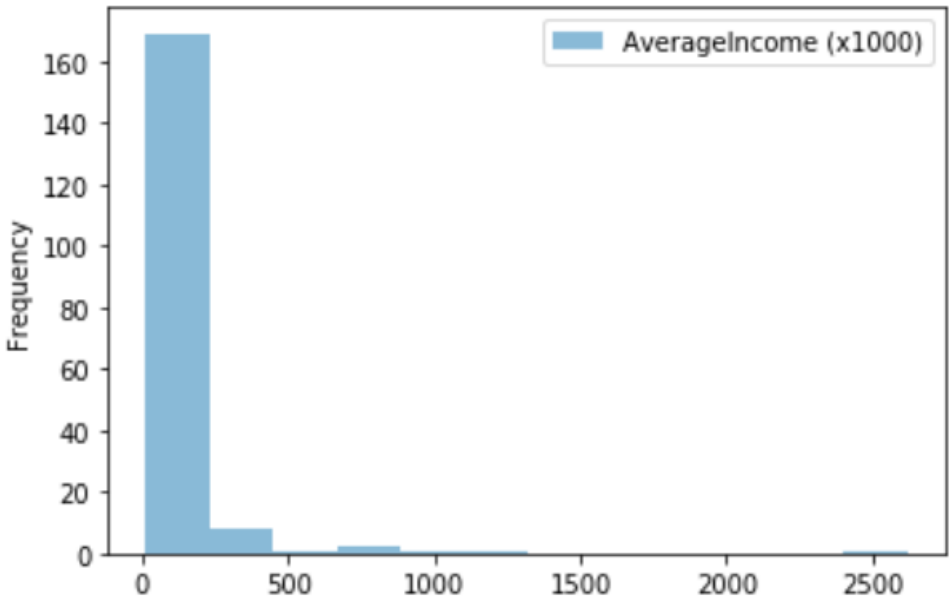
ZIP code [1]	Size of adjusted gross income	Number of volunteer prepared returns [2]				Number of refund anticipation check returns [3]	Number of elderly returns [4]	Adjusted gross income (AGI) [5]	Total income	
		Total	Number of volunteer income tax assistance (VITA) prepared returns	Number of tax counseling for the elderly (TCE) prepared returns	Number of volunteer prepared returns with Earned Income Credit				Number of returns	Amount
		(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
00000	Total	219.100	137.280	81.810	40.440	1.208.550	2.240.650	852.298.824	9.580.250	863.210.313
00000	$1 under $25,000	145.830	95.950	49.880	35.450	526.520	653.290	41.312.407	3.350.600	42.145.847
00000	$25,000 under $50,000	58.620	36.510	22.110	4.990	381.930	428.550	78.566.565	2.159.110	79.578.199
00000	$50,000 under $75,000	12.280	4.480	7.800	0	164.170	341.070	81.560.100	1.326.500	82.502.410
00000	$75,000 under $100,000	2.120	310	1.800	0	71.160	253.910	73.530.398	848.060	74.233.240
00000	$100,000 under $200,000	250	30	220	0	62.400	397.290	180.245.335	1.312.640	182.204.547
00000	$200,000 or more	**	0	0	0	2.370	166.540	397.084.019	583.340	402.546.070
10001		390	170	210	60	1.490	2.640	2.785.887	15.340	2.820.411
10001	$1 under $25,000	250	110	130	60	470	850	42.685	3.640	44.319
10001	$25,000 under $50,000	110	60	50	**	360	400	94.622	2.540	96.947
10001	$50,000 under $75,000	30	**	30	0	230	350	125.439	2.020	127.601
10001	$75,000 under $100,000	**	**	**	0	150	290	128.019	1.480	129.884
10001	$100,000 under $200,000	**	**	**	0	230	460	411.915	2.920	417.092
10001	$200,000 or more	**	**	0	0	50	290	1.983.127	2.740	2.004.568
10002		1.310	1.040	260	340	3.990	8.030	2.489.499	42.290	2.523.419

Data Cleaning

The income data (in thousands$) has been grouped for zip code (collapsing the income brackets) and a new variable AverageIncome has be introduced as the

$$\frac{Total\ Zip\ Income}{Total\ Zip\ Population} = AverageIncome$$

Besides the programming difficulties in parsing a human oriented excel file, some issue have emerged when we have plotted the data:



The income inequality introduce a variable that contains a good number of outlier, a special handling for this variable has been required (see methodology)

Venues Data

Data Source

During the course we have used the foursquare data manually implementing the request, for this project we are going to use the libraries suggested by foursquare.
For Python the suggested library can be installed using pip and in this notebook we assume that it's already installed in the python environment and available to the current running kernel

Two different dataset were created:

1. A dataset containing the list for venues inside a fixed range from the center of each zip code
2. An additional hierarchical dataset containg all the various category and subcategory: for the calculation of how many gym exists in and around a given zip code we want to keep in consideration a "loose" definite of Gym instead of the specific definition given by the "third level subcategory": a "Boxe gym" can easily be a competitor of our new target gym

Data Cleaning

No special cleaning was required for this data, only some small tuning in the request api code to target only the gyms

Chapter 3

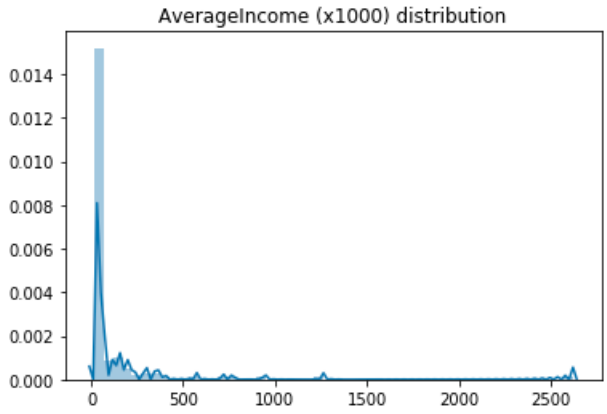
Methodology

Where we discuss and describe the exploratory data analysis that we did and what clustering algorithm were used and why.

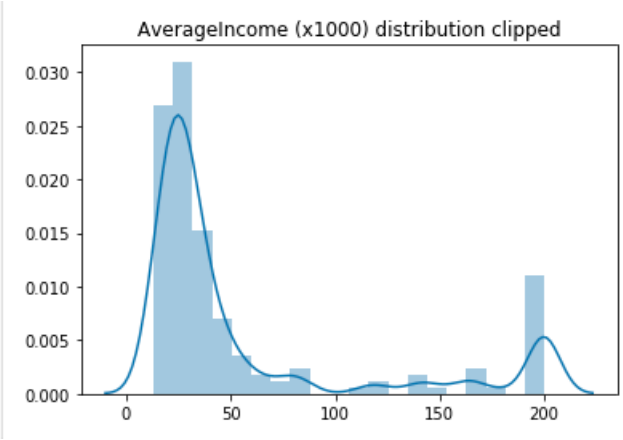
To save time in the future, print a copy of this document. Click **Print** on the **File** menu, and press ENTER to receive all eight pages of examples and instructions. With the printed document in hand, position yourself in normal view to see the style names next to each paragraph. Scroll through the document, and write the style names next to the paragraphs (press CTRL+HOME to reposition yourself at the beginning of the document).

Merging and Exploratory Analysis

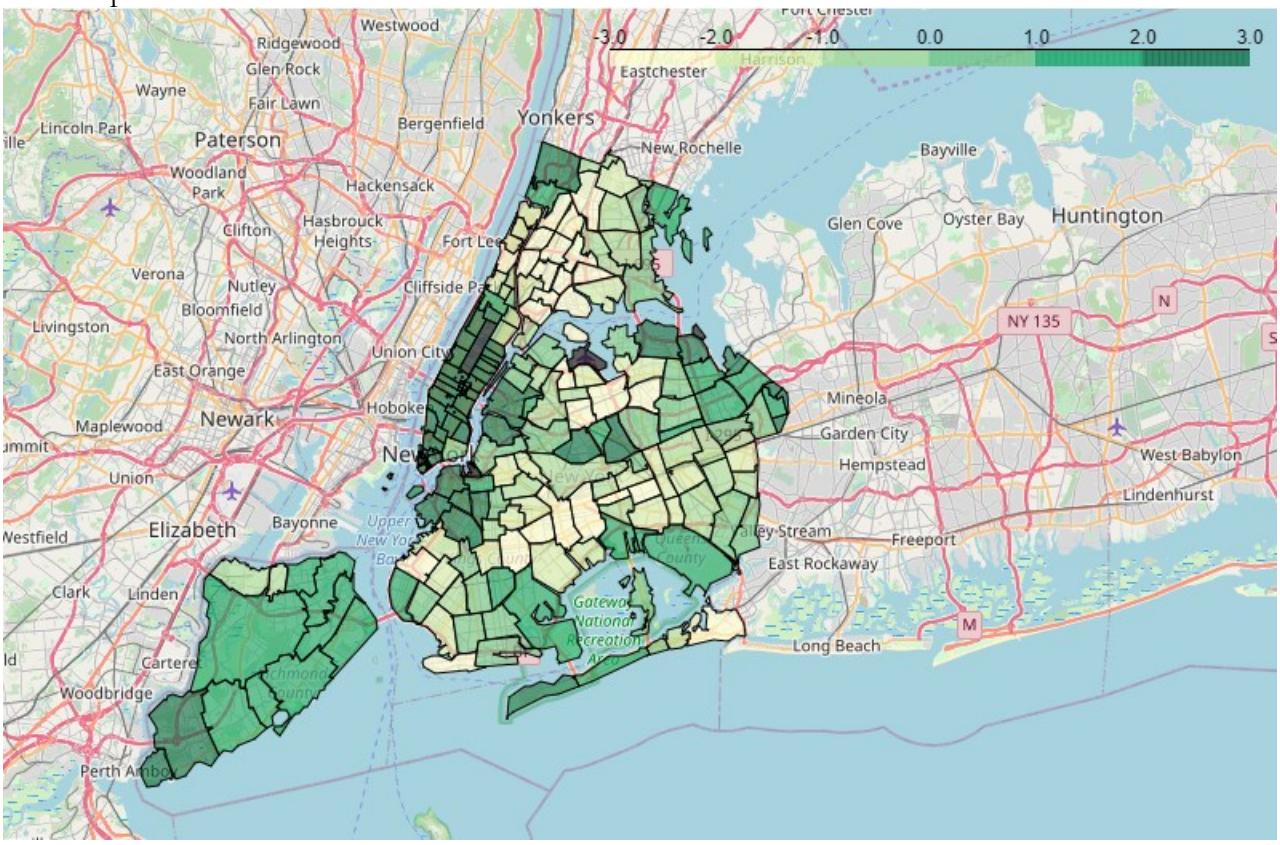
The merge of the income data with the population data allows the calculation of the average income, plotting this data reveals and extremely skewed distribution



Specifically while the median is around 29.7 the standard deviation is around 242.9.
For this reason we handle this outlier clipping the values above >200 to 200, obtaining the following distribution

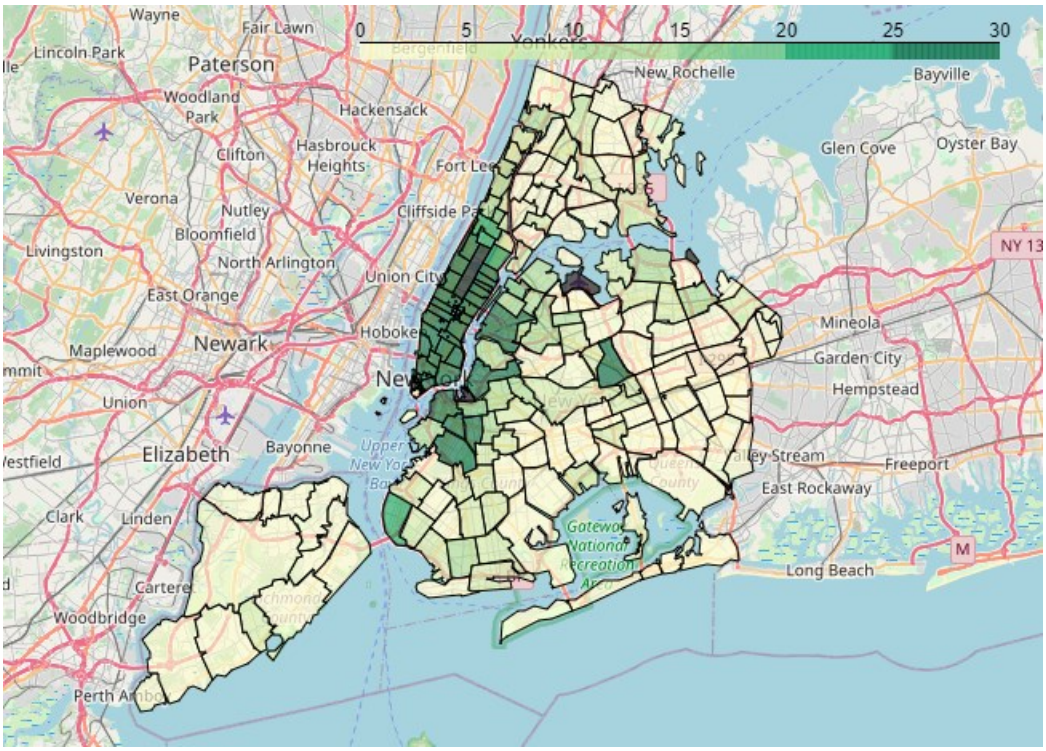


We then split the datas in 7 quartiles and assign to each zip code an Income Index in the range [-3, 3], plotting the Income Index on the map allows us to identify the different economic condition of the various Zip code



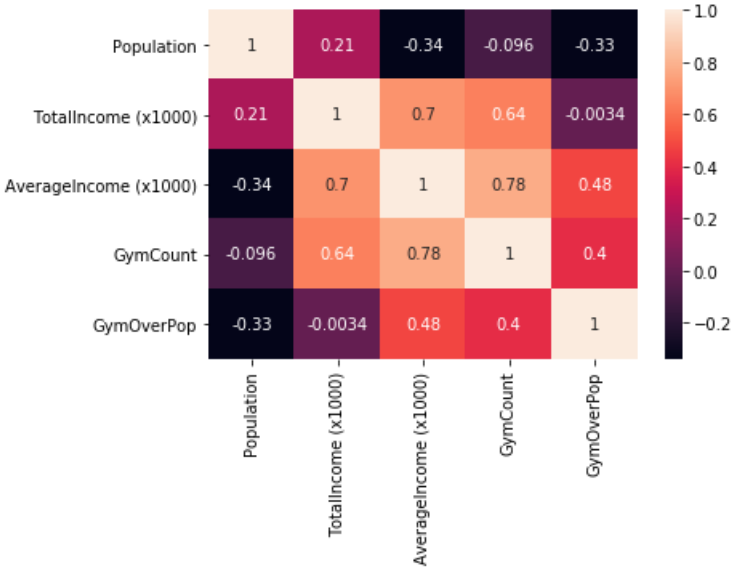
We procede to collect using the foursquare api, on how main gym are near each zip code, some gym will appears in more than a search, but this is fine: the objective is to see "how many gym" are easy to reach (accessible) to a given zip code, not only the ones physically registered there.

These data revels some interesting pattern that are easier to see in map form



Obiously the number of available gym must be taken in consideration together with the population, for this reason we introduce a new variable GymOverPop that rappresent the ration between the number of gym and the population of a zip code.

With the introduction of this variable we can now build a correlation matrix to see if we can refine our search parameters:



Interestingly there is a strong positive correlation between the average income of a zip code and the density of the gym.

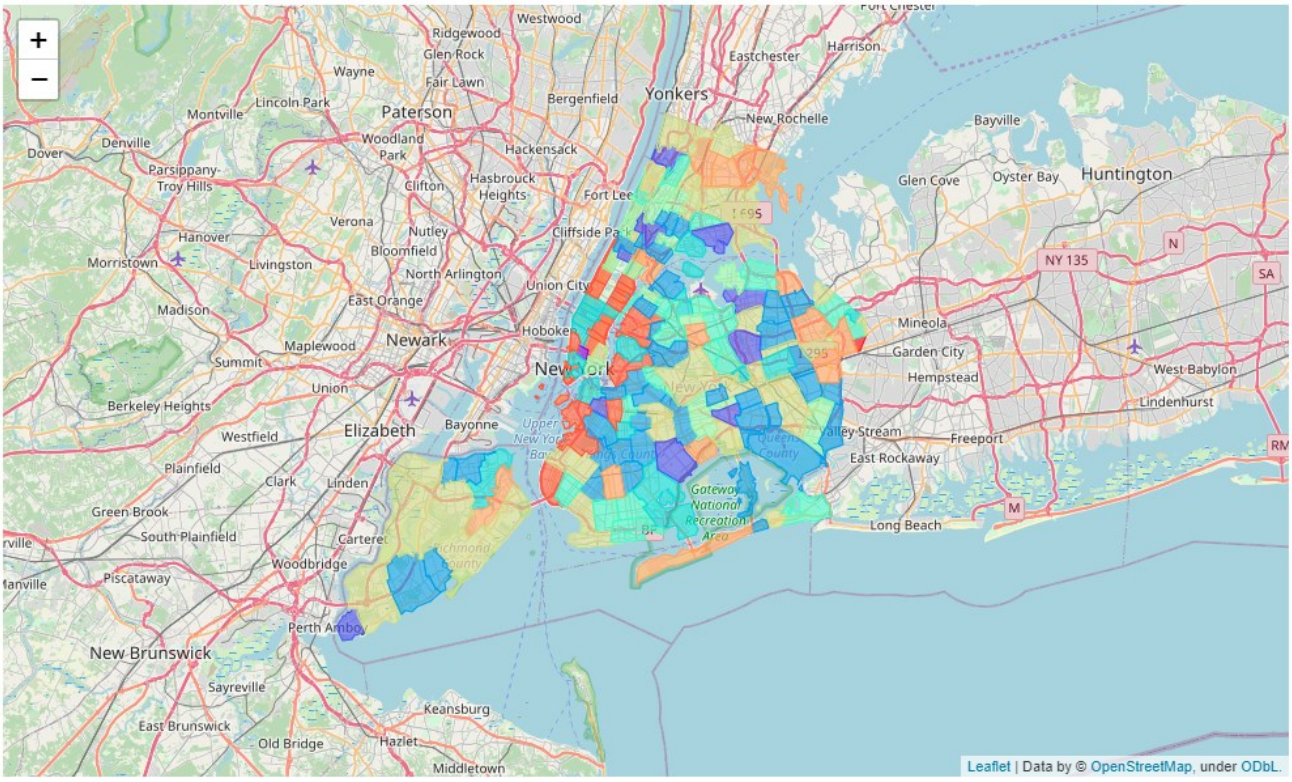
Clustering

We procede now using the venues data (excluding gym) together with the income information to associtate the Zip Codes by similarity.

For this task we are going to use the k-Means clustering algorithm.

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

Plotting the results on the map we can see the zip code similary with a k=10 cluster structure:



Chapter 4

Results

Where we discuss the results obtained from the analysis

Now we study the behavior of the "GymOverPopulation" variable in the various cluster

]:

Cluster	count	mean	std	min	25%	50%	75%	max
0	1.0	30.000000	NaN	30.0	30.00	30.0	30.00	30.0
1	11.0	8.363636	9.330303	0.0	2.00	5.0	9.50	30.0
2	38.0	5.763158	6.478381	0.0	1.25	3.0	8.75	30.0
3	21.0	18.428571	13.643942	0.0	2.00	29.0	30.00	30.0
4	20.0	8.550000	5.942488	1.0	4.00	8.0	11.00	27.0
5	15.0	10.800000	8.945869	0.0	4.00	9.0	14.00	29.0
6	36.0	3.694444	5.338911	0.0	0.75	2.0	5.00	29.0
7	16.0	2.875000	2.305790	0.0	1.00	2.0	5.25	6.0
8	24.0	26.125000	8.131060	1.0	27.00	29.5	30.00	30.0
9	1.0	4.000000	NaN	4.0	4.00	4.0	4.00	4.0

Excluding the two degenerate cluster containing two outiler zip codes the cluster containing the "highest number of gym" in average (and with a very high 1st quartile) is cluster number 8.

Let's extract the 10 zip codes with the lowest number of gym per residents

ZipCode	Population	IncomeIndex	GymCount	GymOverPop
11232	27723	-2	4	0.004252
10021	102078	2	27	0.007794
11209	69840	1	24	0.010126
11211	85089	0	30	0.010389
11231	32974	2	13	0.011617
10024	61414	3	26	0.012475
10023	62206	3	29	0.013737
10128	59856	3	30	0.014769
11238	48965	2	27	0.016248
10003	53673	3	30	0.016470

Chapter 5

Discussion

Where we discuss our observations any recommendations based on the results.

From the analysis we have extracted ten possible Zip Code. These codes are the "most similar" to other "high gym-density" Zip codes, but they have themselves a lower number of gym.

This indicates that this "kind" of zip code generally presents a good business opportunity for gym, but the first two in particular (11232 and 10021) have a lower gym density.

We know from the correlation analysis that the Income has a positive correlation with the number of gym, for this reason the can identify **10021** as the best candidates for opening a new Gym .

Chapter 6

Conclusion

Where we summarize our work and look for further improvement.

Final Consideration

Opening a gym can be a risky business, in this studies we have looked in how Data Science and Area similarity can help mitigate this risks, while in this study we have focused on New York city the same modelling can of course by used on other cities.

Future Developments

This model assumes the ability to open a gym in any possible Zip Codes, further development of this model can include the possibility to use more specific restriction in terms of budget (in some Zip code rent can be extraordinarily high)

Great focus should also be put in analyzing (with other dataset) why the number of Gym seems negatively influenced by the population count.