Definition of the gym opening potential by zip code areas

Author: Boris Silantev
May 27, 2019

Table of contents:

1. Introduction

1.1. Background

Florida is the southernmost contiguous state in the United States. With a population of more than 18 million, according to the 2010 census, Florida is the 3rd-most populous (21,312,211 inhabitants) state in USA. Florida's $1.0 trillion economy is the fourth largest in the United States. If it were a country, Florida would be the 16th largest economy in the world. The most part of Florida's population is concentrated in metropolitan areas, the most populous of them are Miami-Fort Lauderdale-West Palm Beach, Tampa-St. Petersburg-Clearwater, Orlando-Kissimmee-Sanford, Jacksonville while there are a large number of smaller municipalities.
Florida is very diverse ethnically and racially. For example Hispanic and Latinos of any race made up 22.5% of the population in 2010.
You can say Florida is very sportive state because Florida has three NFL teams, two MLB teams, two NBA teams, two NHL teams, and one MLS team. According to 2014 State Indicator Report on Physical Activity of National Center for Chronic Disease Prevention and Health Promotion 29.2% of adults in Florida met muscle-strengthening guideline that is close to average value for U.S. According to non-confirmed data the number of health and fitness clubs in USA permanently increases from 2012. So it can be interesting for investors which places in Florida are worse or better for fitness club or gym opening.

1.2. Problem

In this project I'm focusing on determination of the potential of fitness/gym opening in area based on the demographics, tax and some other information about the areas. I consider zip codes as areas.

1.3. Interest

The main interested players are investors of any size who can be choosing the right place for opening the new fitness club or gym. From other hand it can be interested for current owners of fitness clubs chains who may make a decision about reduction of chain. Also it can be useful for some contiguous businesses that is related to fitness industry.

2. Methodology

2.1. Data sources

In order to solve the problem I use following data sources:

- basic and demographics information about zip codes areas available here. Unfortunately the demographics based on the Census 2010 and business information is based on the Business Census 2011 but it's most current official information. I also use prediction of current population provided on this website;
- individual income tax statistics in the context of zip code areas for 2016 provided by IRS here;
- information about different venues placement available via Foursquare API

2.2. Data preparation

I requested information about most popular venues in Florida from Foursquare using consequentially latitude and longitude of all zip codes and then grouped them using zip code got from the address field of the Foursquare records. Because Foursquare API returns information about only 100 most popular venues in the given radius I used 5 different radii from 350 m to 27 km in order to get as more venues as possible. Separately I requested the information about 'Gym' and 'Fitness' using this time zip code as an area identificator. Of course I've got not only gyms, among most popular categories there were: Gym / Fitness,

Gym, Martial Arts Dojo, Yoga Studio, Weight Loss Center, College Gym, Gymnastics Gym, Pilates Studio, Gym Pool and even Hotel. Which of these categories to consider as a Gym and which ones not it's a matter of choice. I decided to consider as Gym all categories consisted the word 'Gym' because it includes different kind of gyms and exclude such venues as martial arts dojo, for sure it will also includes pool gym and maybe some other categories that shouldn't be considered as a gym but their quantity is too small, the main splitting will be correct. After that I added all gyms into all venues table and drop duplicates from this table, changing the all categories consisted the word 'Gym' to 'Gym' in advance. So I collected information about 58 thousands venues. In this table I merged some similar categories, for example joined all type of restaurants (Mexican, Italian etc.) into one category "Restaurant", the same with joints and museums, combined 'Wine Bar', 'Pub', 'Cocktail Bar', 'Beer Bar' and 'Beer Garden' in 'Bar' category but left 'Juice Bar' in a separate category. Finally, I made one-hot encoding and saved only categories that at least correlated with the number of gyms and counted at least 80 venues. As a result, I had the table with number of venues by 76 categories in each zip code area.

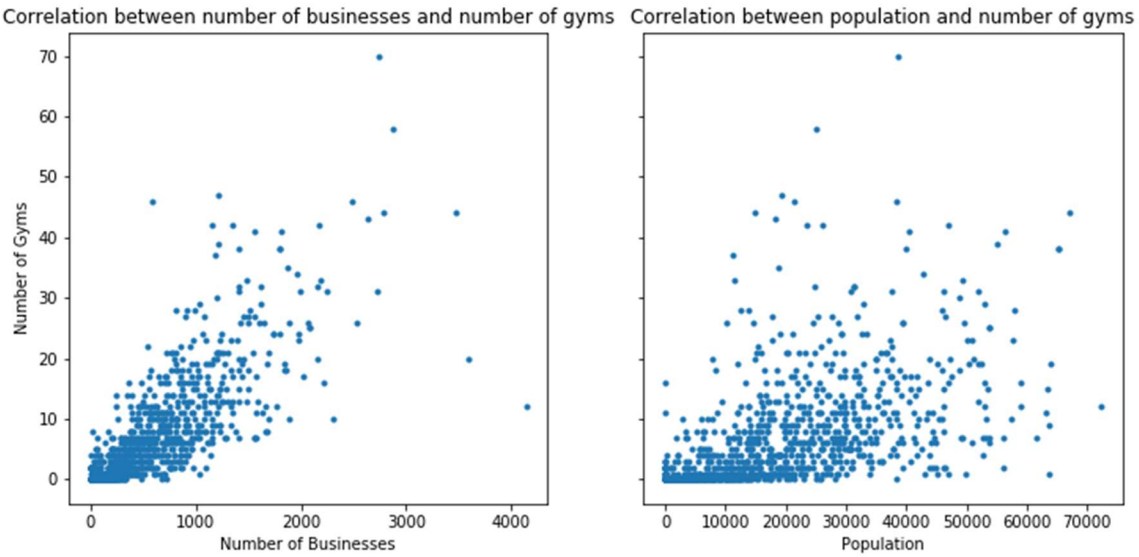
Then venues data were combined with demographics data and taxes data into one table where zip code was an index of rows while different features was the names of columns.

I dropped all rows with missed values: among 1476 rows there was 559 rows with NaN values but most of them was in the rows with Type 'P.O. Box' (491 rows) or 'Unique' (51 rows), these data was unimportant for me because both P.O. Box and Unique type are not the code of any area and exists only for postal purposes. There was also 17 standard zip code areas with missed values but examined it I found out that some of them has no population, some of them has no venues, anyway the data looks strange, maybe incorrect, so I dropped all rows with missed values.

2.3.Exploratory data analysis

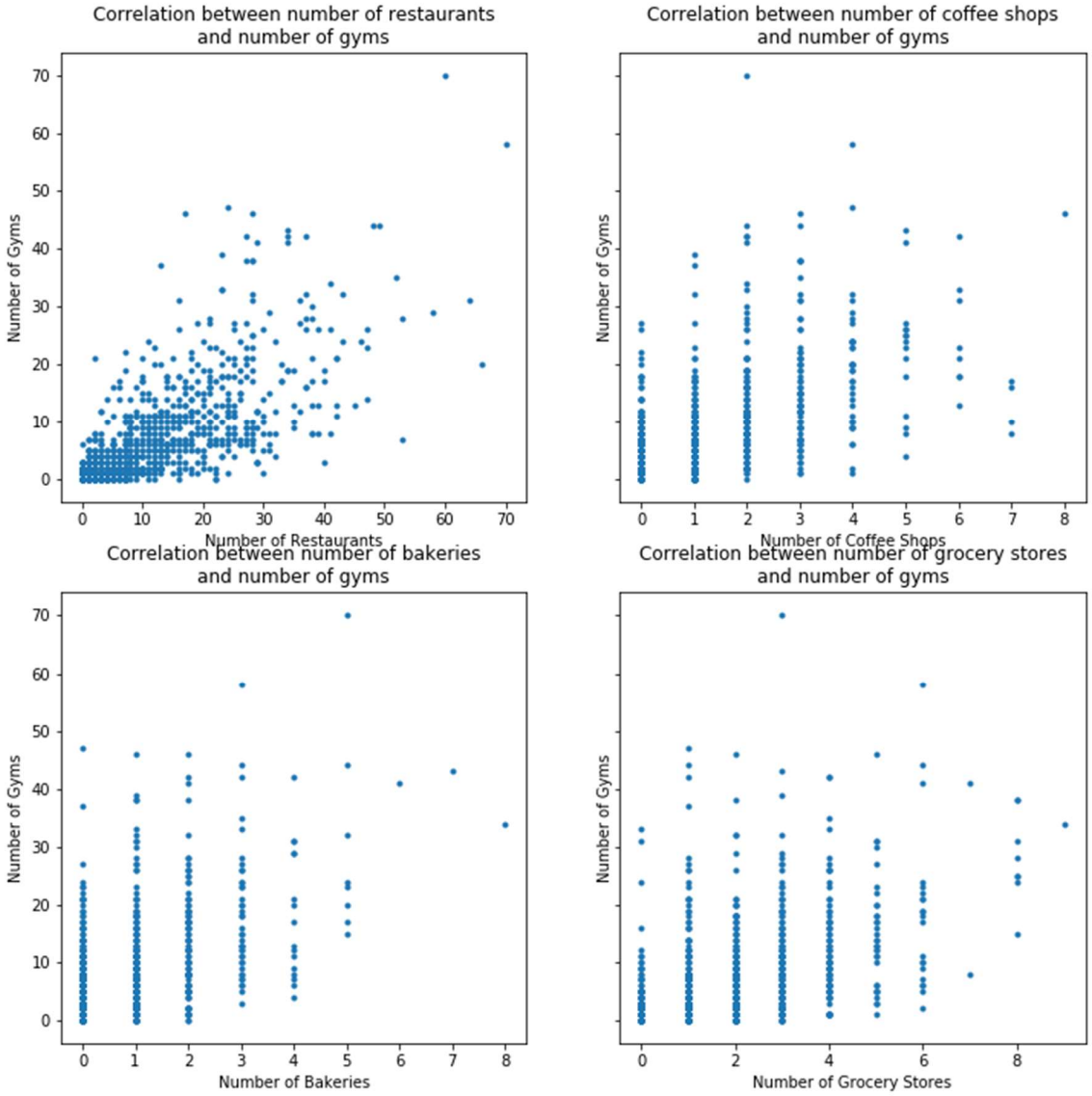
Number of businesses and population

It's not evident but the correlation between number of businesses and number of gyms is much stronger than correlation between population and number of gyms. We can see it on the image above. Among top 5 correlated features there are also Number of Employees and Annual Payroll. It actually means that people prefers to do exercises not near the place of living but near the office or near the shops.



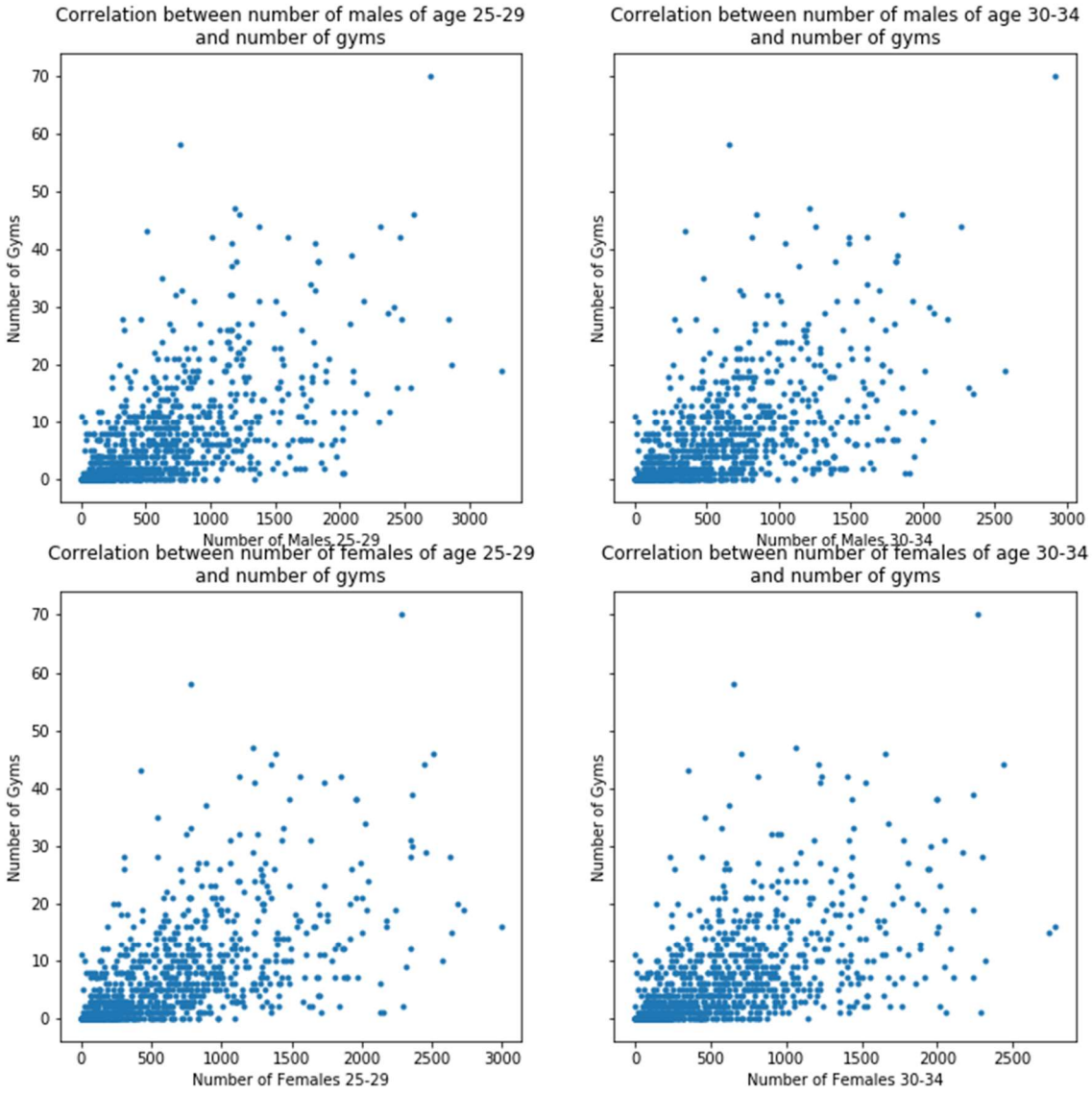
Venues

Among the venue categories most correlated with number of gyms I found Restaurant, Coffee Shop, Bakery, Grocery Store. While the correlation between number of gyms and number of restaurants can be well presented on the image, it's not clear for other 3 categories because the number of venues of these categories differs not a lot.



Demographics

The most correlated with number of gyms demographic groups are Male 25-29 and Female 25-29, next ones are Male 30-34 and Female 30-34 but plots looks are very similar with Total Population plot, so it seemed that just there are strong correlation between demographic groups and total population that is logically sound.

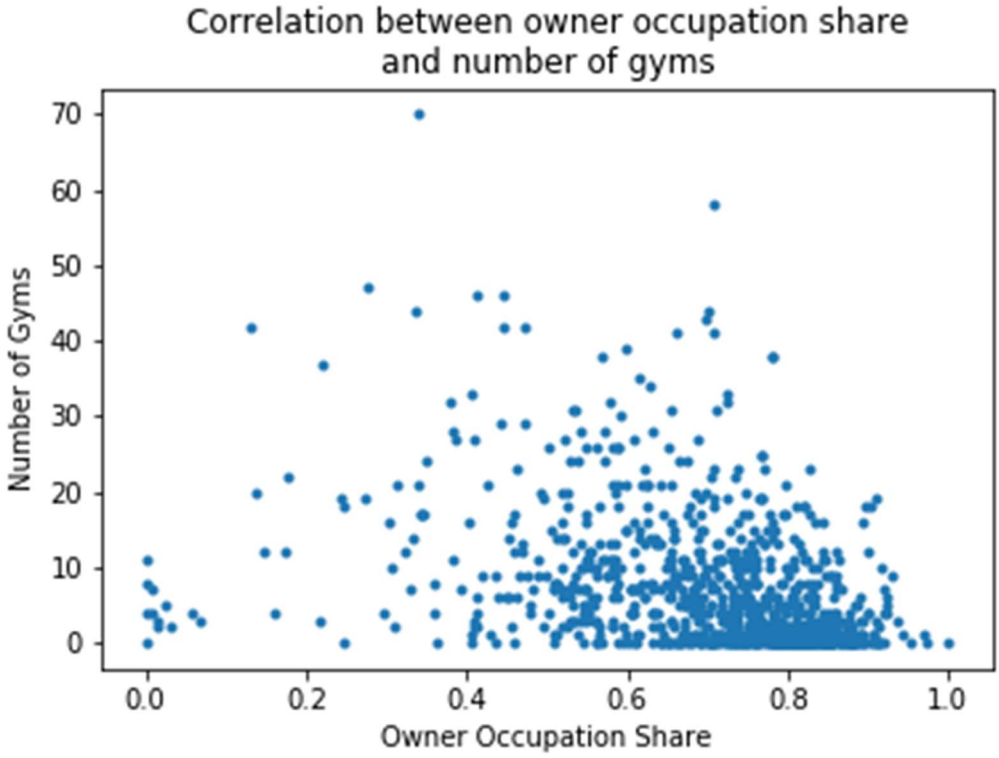


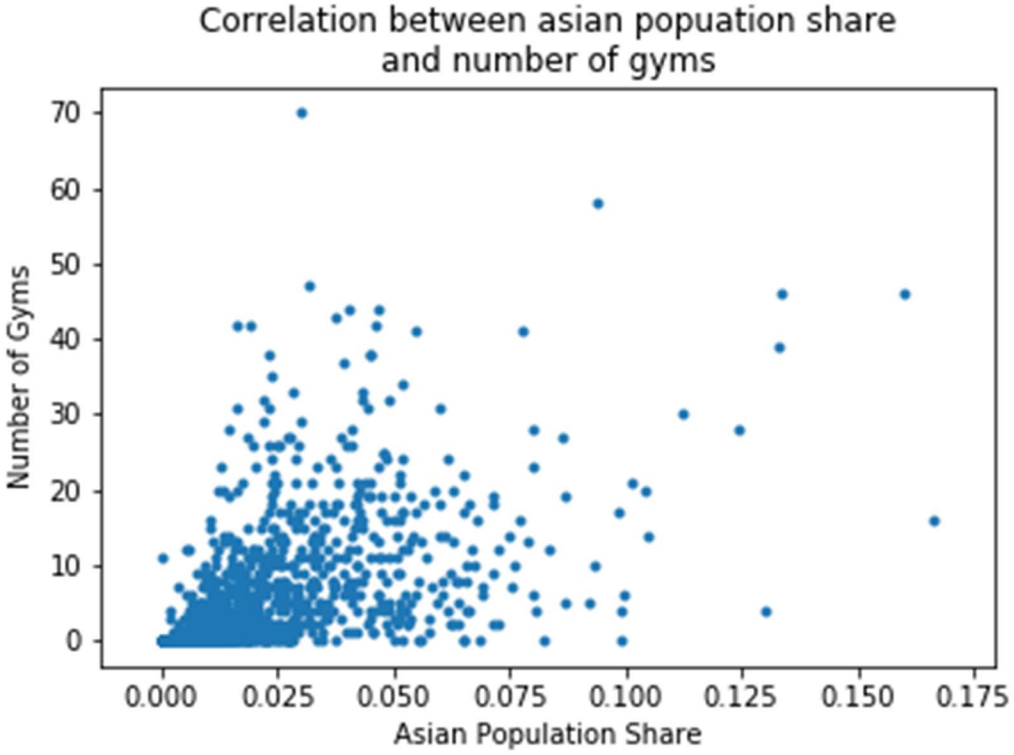
Owner Occupation Share

It's a little unexpectedly but there is a negative correlation between owner occupation share and number of gyms, in other words where owners prefer to live rather than renting housing, the number of gyms is usually less. I can suppose that residents who lives in their own houses and apartments prefer quiet locations where there is no a lot of shops, restaurants and other businesses.

Asian Population Share

Also there is relatively strong correlation between Asian population share and number of gyms, moreover I'm talking about the share but not absolute value, so it cannot be explained by the correlation of the Asian population share and total population. Unfortunately I can't explain it and will use it as is.





2.4.Research methods

In order to determine the potential of the gym opening in an area I decided firstly to build the model that predicts number of gyms in an area basing on the collected features, and after that just to compare the prediction with real data.

Because the data was very diverse, of different scale and of different nature, a lot of features had high correlation I made a decision to use gradient boosting regressor, the model that is insensitive to such a complexities in the data. In order to split the data to train and test sets and at the same time to have a prediction for every zip code I used KFold technique with 4

splits but to smooth the prediction I applied 5 different splits and averaged the results for each zip code.
I defined the gym opening potential index as a ratio of the predicted number of gyms and actual number incremented by one. Increment was needed because the new opening would change the ratio and also it lets to avoid the division by zero in the case when actual number of gyms is equal to zero.