Inference Model

Finding a business opportunity into
Los Angeles City, California at 2019
Analyzing population and active business
by zip code.

Author: Gondres, Jeffry.

May, 2019.

INDEX.

How to find a good option for opening a business in LA

Gondres Jeffry.

May 11, 2019.

1. Introduction.

1.1. Background.

Los Angeles city is located in California State with more than four millions of citizens who some of them have a permanent job, another proportion is looking for some new opportunities of investment and a final part looking for a job. This data exploratory approach, wants to give some insights about how is LA distributing their venues by ZIP code to give a piece of advice about what could be a good investment on it.

1.2. Problem.

The State of Los Angeles has statistics about his population by ZIP code. Some resident of this city wants to know which could be the best zip code for opening a new business. So the problem is about look the best top 10 ZIP codes in order to let him know where could be the best place for opening a new business. He doesn't have any preference about the kind of business, so it could be any type of Venue.

1.3. How Solve the problem.

Once we have Venues per Zip Code, we will check if there any relation between population and quantity of Venues. In order to do that we will make a simple projection of the population using yearly growth rate found in the web site http://worldpopulationreview.com/us-counties/ca/los-angeles-county-population/, and then apply techniques such as PCA, logistic regression and scatter plotting.

After have the estimate population, some models will be applying over final dataset, and the best model performance is going to be selected depending of the test results. Once the model is selected, a new projection will be done in order to know what will be cluster with more growth in the next 2 years, and that cluster shall be the suggest within their zip codes.

2. Data acquisition and cleaning.

2.1. Data sources

Los Angeles has a Open Data web site available from where get the following two datasets and the Venues are going to be collected using Foursquare API.

Dataset 1. Population data regarded from 2010 census grouped by ZIP code.

Dataset 2. ZIP codes with its Latitude and Longitude value (https://catalog.data.gov).

Dataset 3. Foursquare.com will provide us with the quantity of Venues registered on their database that will be query using latitude and longitude for each Zip Code.

2.2. Data cleaning

Within the first dataset I found population data what came with several columns such as population among male and female, households and median age size. For the purpose of this document is just useful the total of the population by zip code, so the other columns were removed.

The Zip code list came with the population data for all California state, so I had to create a list with the list of zip codes just for the Los Angeles city and then make an inner join with the population data above. Once the dataset by LA city zip codes is complete, a multiplier factor was applied in order to get the population for 2019 assuming the same growth rate for all zip code since 2010.

Each zip code contains a lot of data for latitude and longitude, so getting a sample three zip codes, I plotted them in order to see the distribution of each address and get a central point what is going to be use as reference using Foursquare API and build the Venue data set.

Once the dataset with Venues, zip codes a population is joined, there is a lot of different names of venues with some principal categories that can be group. The aggregation of that data was did manually assigning a new parent category depending of the descripcion of the Venue and then make a Pivot table where each column represents the parent category assigned and its quantity of venues per zip code getting NA values what are going to be replacing by zeros understanding that there are zero Venues of that category.

After get some basic statistics about the conformation of each zip code, a limitation of Foursquare popup from the charts, because it has a limitation of 50 venues per address, so I did all the same steps above, but in this time using a dataset with all active business of California update in april 2019, and now the data is more intuitive.

2.2. Feature Selection.

The dataset within zip codes and all address is not useful because when the data request is made using the Foursquare API you may specify a radius, si the key is selecting the middle point of the zip code as shows:

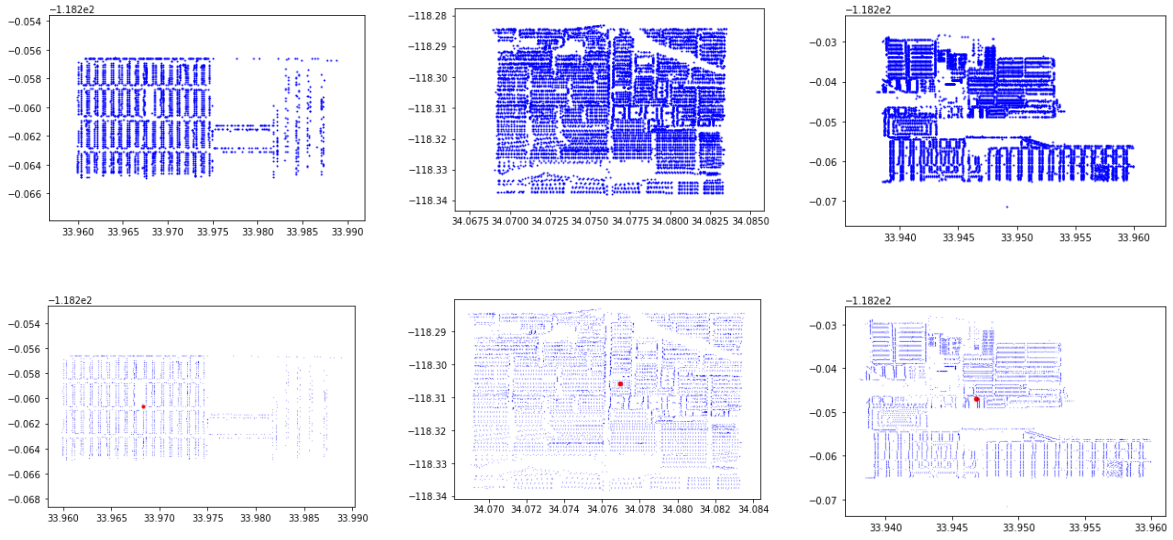

Figure 1. Latitudes and Longitudes by Zip Code (3 samples)

After select the center points by zip code, the following image shows how the distribution looks like into the LA city map.

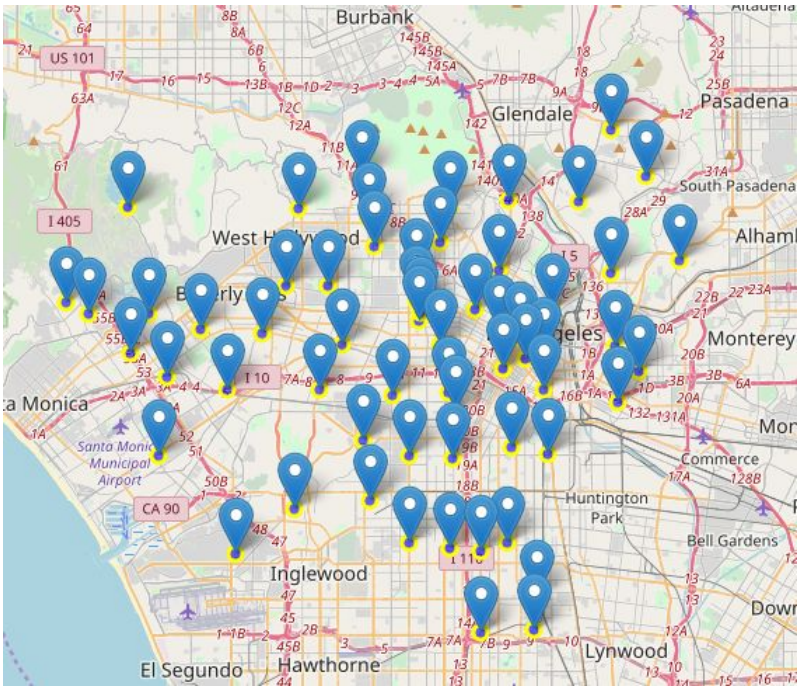

Figure 2. Map plot for all LA city Zip Codes.

The other variable is Parent Category what contains eight subcategories, but getting a distribution of each sub category thru all zip codes with can identify three clusters using singular value decomposition as follows:

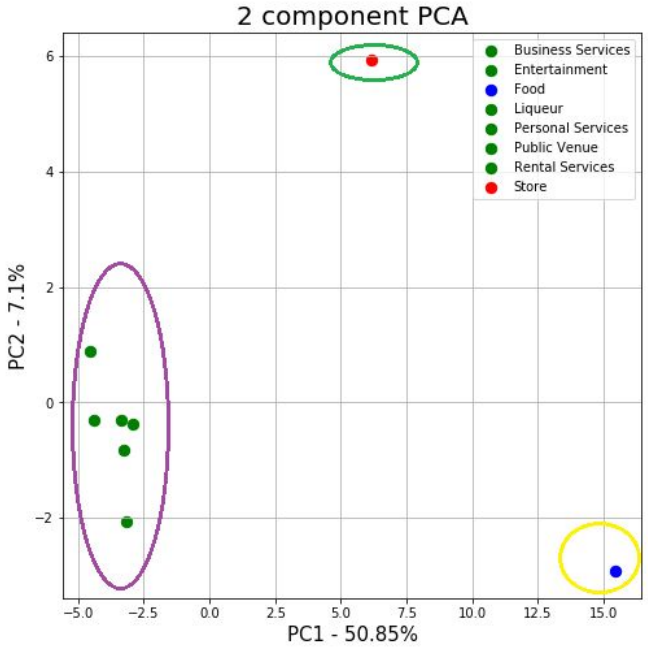

Figure 3. PCA using SVD over data regarding using Foursquare API.

3. Exploratory Data Analysis

3.1 Calculation of target variable

Any investor wants to know which place into LA have a lack of commercial needs in order to supply the market with their own purpose.

The information used in this document is about the relation between population and existing venues, so according to the zip code, its population the model are going to find the top 10 venues that are under the “optimal”, knowing that its optimal value is a inference about the data explored.

3.2 Relationship between population and quantity of venues.

Is expected the the quantity of venues dependes of the size of population, however using Foursquare that is not what we may see into a map, and it makes sense because in the downtown is more common that people who reports venues around food and stores, and is not common get the app for reporting another kind of business how shows the following mat where the size of the circle is according the population on its zip code.

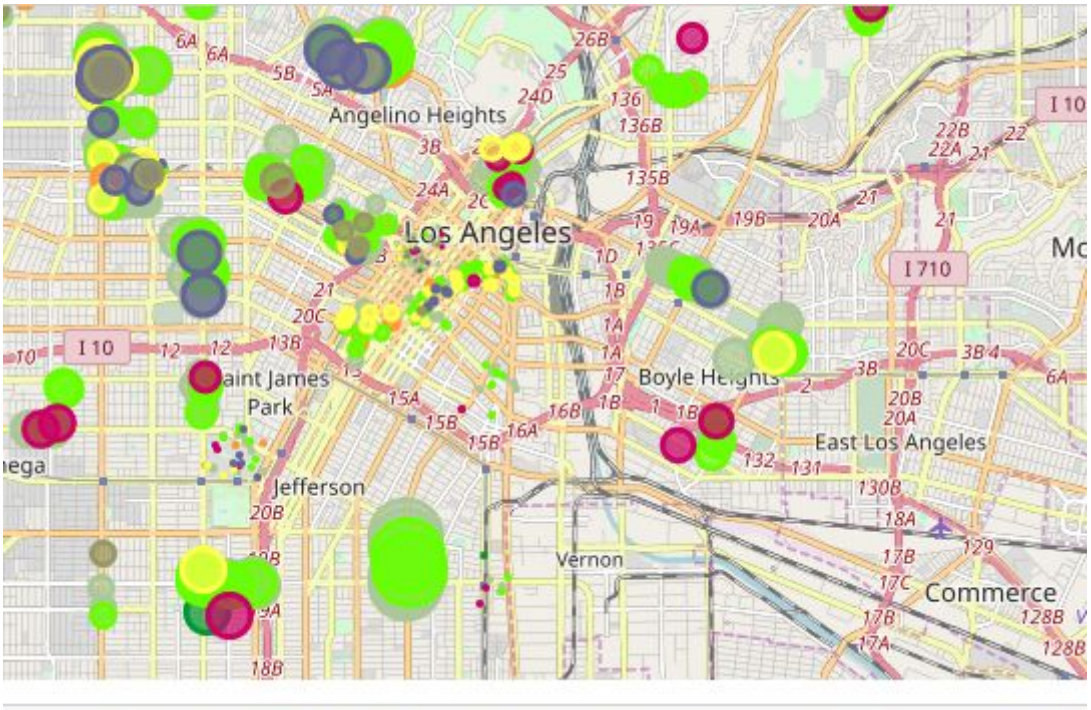

Figure 4. Map plot for LA downtown. The size of the circle represents the population quantity by zip code.

As I said before, the majority number of venues are in the center of Los Angeles, but around of are the biggest circles, what gives us our first insight; downtowns will require

food’s and store venues according the size of working population and that variables even is out of this document, can be searching and find the better fit of the model into.

3.3 Venues distribution across LA city.

These are all categories listed into the data exploratory analysis.

Table 1. Category Distribution within LA city

Parent Category	Quantity of Venues	Color
Business Services	7	
Entertainment	18	
Food	68	
Liqueur	15	
Personal Services	24	
Public Venue	26	
Rental Services	7	
Store	65	

The table 1 shows the quantity of each parent category get from Foursquare API by zip code, accordingly of that aggregation a scatter is as following:

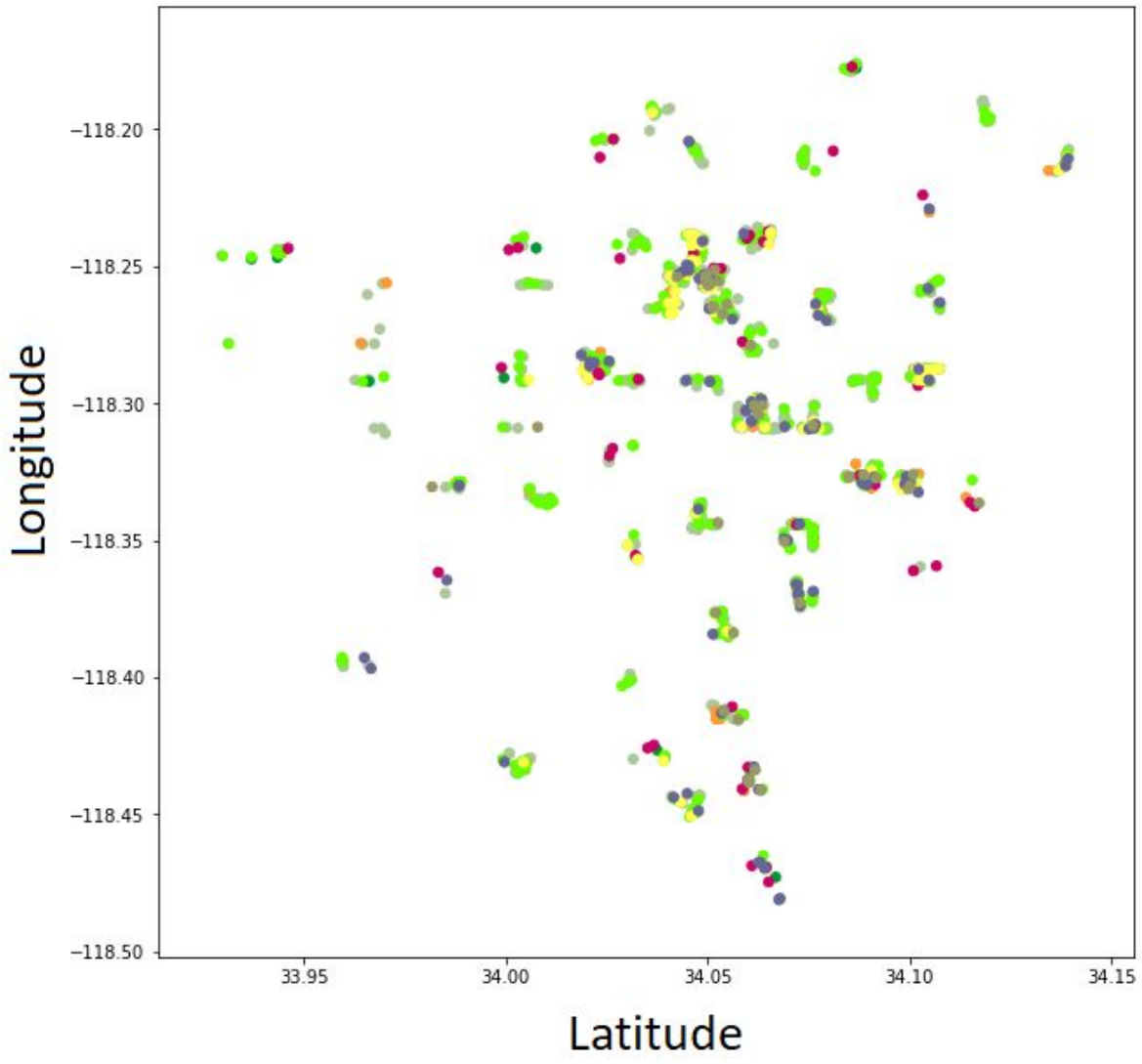

Figure 5. Parent Category distribution among LA city latitudes and longitudes by zip code.

The quantity of venues gave by Foursquare is not a good representation of all business present in LA city, so according to the state dataset for active business I did all the analysis above but now using a now dataset and the distributions is as follows.

Table 2. Total of active business into LA city until april 2019.

Parent Category	**Quantity**
Business Services	**5991**
Entertainment	**135**
Food	**195**
Goods	**1810**
Liqueur	**147**

No category	**62**
Personal Services	**398**
Public Venue	**177**
Rental Services	**1722**
Store	**2762**

Using the new dataset now we have a new PCA chart that explains in a better way the possible clusters we may found into LA city.

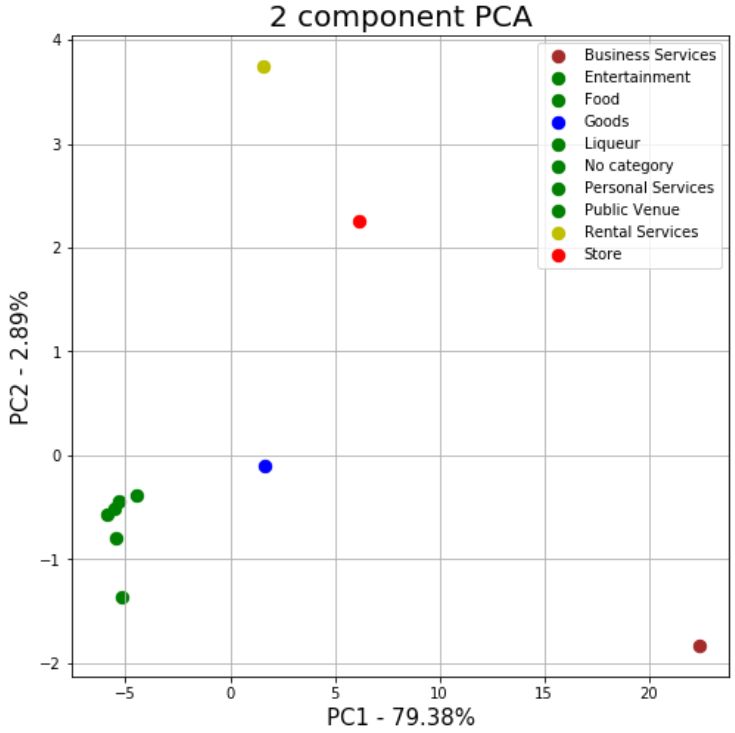

Figure 5. PCA using active business in LA until april 2019.

Comparing Figure 5 against Figure 6, there is two more clusters and food is not a cluster anymore, so this make sense because Foursquare is so useful to find food venues, and its database should contain that kind of business in a major proportion than other ones.

4. Inference Model.

In this kind of scenarios doesn't make sense build a predictive model because the way business are growing depends of how the community is claiming for a certain of goods and services in general and how those good or services find the best delivery performance where final customer can buy it.

4.1. PCA variables performance.

Using Parent Category as tuples and zip codes a columns is how the data was ordered in order to analyst zip codes as samples and categories as characteristics that defines each zip code.

Table 3. Zip codes analysis by Parent Category.

Parent Category	90001	90002	90003	90004	90005	90006	90007	90008	90009	...
Business Services	81.0	81.0	98.0	106.0	104.0	110.0	98.0	89.0	4.0	...
Entertainment	0.0	2.0	3.0	3.0	3.0	2.0	3.0	3.0	0.0	...
Food	2.0	3.0	3.0	4.0	3.0	3.0	3.0	3.0	0.0	...
Goods	36.0	24.0	33.0	33.0	27.0	32.0	33.0	23.0	1.0	...
Liqueur	3.0	1.0	2.0	2.0	2.0	2.0	3.0	2.0	0.0	...
No category	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	...
Personal Services	5.0	7.0	8.0	7.0	7.0	7.0	7.0	7.0	0.0	...
Public Venue	3.0	2.0	2.0	4.0	2.0	2.0	3.0	2.0	1.0	...
Rental Services	23.0	21.0	26.0	33.0	29.0	33.0	28.0	30.0	1.0	...
Store	41.0	34.0	49.0	49.0	47.0	50.0	47.0	46.0	3.0	...

Once the above table is used to feed the PCA algorithm into a loop that changes the the quantity of features and obtaining the measurement of its we can define what how well is used two principal components.

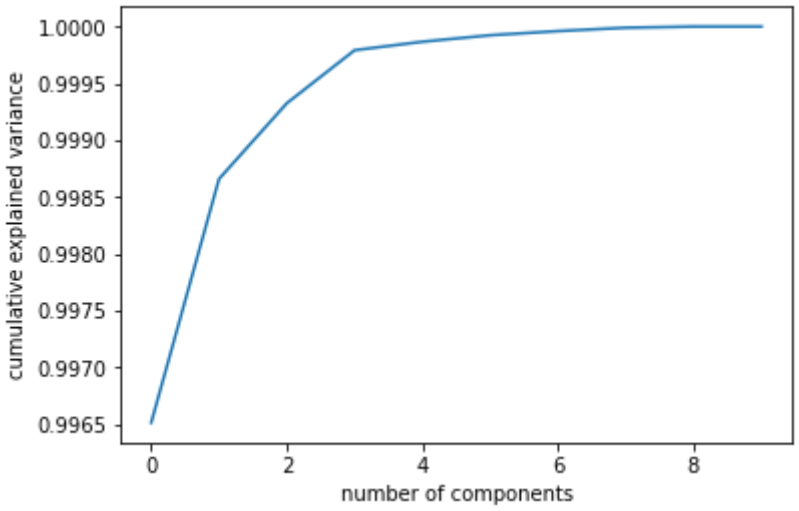

Figure 6. PCA performance by quantity of components.

4.2. Relation between population and quantity of venues.

Analyzing how each category depends or nor of the population size among zip code list is interesting to understand that there is a limit, how is expected, but with this model we can see how real is that assumption.

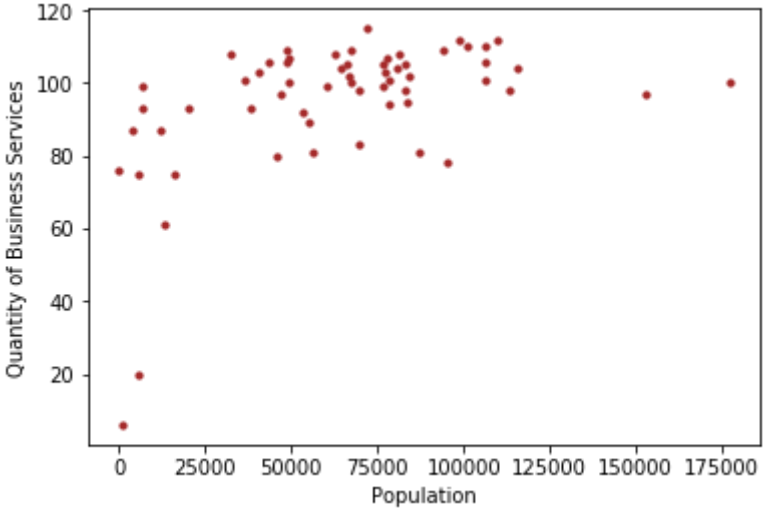

Figure 6. Relation of population against Business Services category.

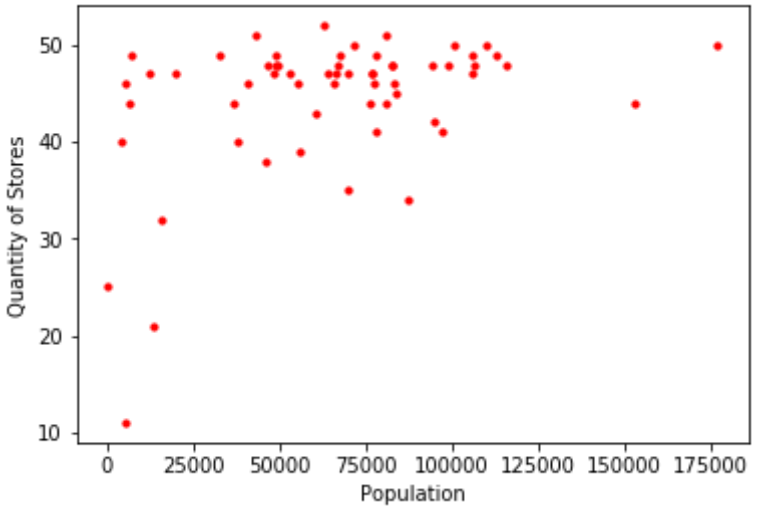

Figure 7. Relation of population against Store category.

After analyst all clusters we found that there is not an confidence relation when the population is less than 25,000 habitants, so in the inference moles I will avoid to use zip codes that contains less than 25K.

4.3. Logistic Regression.

Using Rental Services as selected category for the following inference, lets see first it relationship within population.

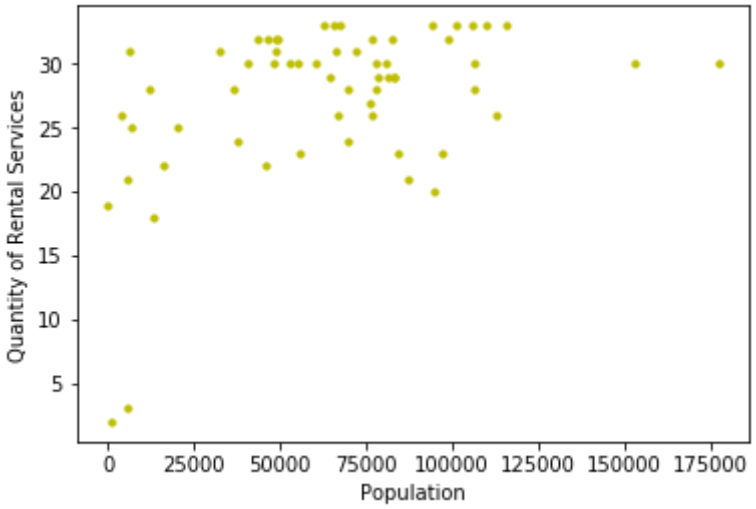

Figure 8. Relation of population against Rental Services category.

After applying Logistic Regression in order to “predict” the quantity of Rental Services across all zip codes we get thirty as a unique result, so it means that as the Figure 8 shows, the best way to fix a values y getting third quintile of each parent category.

5. Conclusions.

Using Foursquare or some of the APIs used for social tracking is not always a good way to understand how a market is build, because each social network have a motivation among its users and is very difficult to create a social network for all kind of business, that'way people use Linkedin as job network and Facebook as good source of news or Tweeter as a good reference about what the people is thinking about a topic. In our case we used Foursquare what could be useful to understand what kind of food venues are more popular and well rated.

In order to know what zip code contains the better opportunity for opening a new business is a well option used zip codes within more than twenty five thousands of habitants and use the third quartile of each category as reference, so all zip codes that contains less that reference quantity could be a well point to start a Market Analysis.