# Capstone project for the IBM Data Science Specialization: The old-man-gym-locator in Toronto

Oliver Emmanuel Argote Brito

March 13, 2020

# Contents



# List of Figures

# 1 Introduction

## 1.1 Background

Many old people is in need of exercising in a different way, at the age of 60 years old many weight lifting (among other) exercises may not be the best ones, in many cases it would be better to hire a physiatry rather than an "all-powerful trainer". To help that people an special place would be needed. A kind of gym that can manage exercise, rehabilitation and some special classes like pilates for persons that are old or have a disability. The personal of the gym needs to be very specialized because of the condition of the target audience.

## 1.2 Problem to solve

A gym like the one that was described would be very difficult to open, the target audience can lead to great income but it is very few people because of many reasons.

The first problem is that hiring health professionals is expensive, not many people (at least at first) will want to invest more money for their physical benefit that the one that they already expend in a regular gym or invest at all.

The second problem is that not every place has many old people nearby, people is more likely to go to a gym that is near their house or work.

The third main problem is that for many reasons, including convenience, comfort or even pride, customers can decide going to a regular gym, even those who really need special treatment, so even if the target audience is different from the one of a regular gym, they still are competence.

## 1.3 Solution

To use data from a given city to know how people is distributed,their age, income and what other options are near of them. This way is possible to make an informed decision of where to place the gym so it has the best chance of succeeding.

# 2 Data and data procedures

## 2.1 Data sources

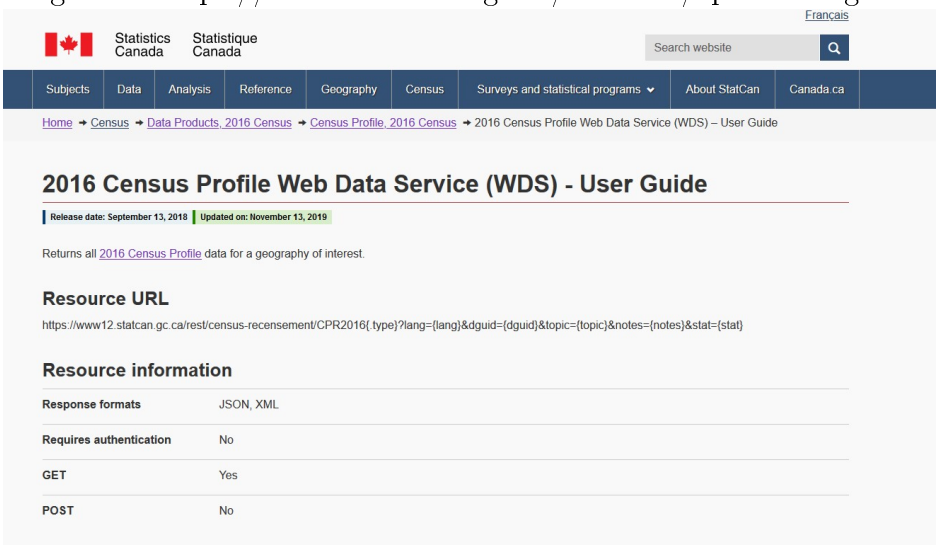
The data analyzed was the distance from other gyms, age, sex and income from people in the city if Toronto, Canada. This information was obtained by consulting the web page of Foursquare and the one of a census from Canada that was made in 2016; Both by using requests.

### 2.1.1 2016 Canadian Census Profile Web Data Service (WDS)

This service is provided trough a page with the link in the image, it contains a lot of information about the Canadian population (Canada, 2019).

To get the information one must request a json file with a link, for the purposes of this project the structure of the link is the default one of the *CPR2016* web service, except that the "DGUID" parameter is 2016A0011xxx, where xxx is the postal code.

Figure 1: https://www12.statcan.gc.ca/wds-sdw/cpr2016-eng.cfm




### 2.1.2 Foursquare

The services of Foursquare are well known so there is not a lot to explain except that those services were used here to know about the competence

around a certain area.

It is very easy to use Foursquare because of the documentation they provide, also the information is vast, so it is very popular among this type of services.

# 3 Methodology

## 3.1 Obtaining, cleaning and preparing data

As described before, the data was obtained by request and then ordered to merge the pandas data frames as one, the neighborhoods are not listed as it is so the postal code was used instead.

For the analysis to be effective, each row must had contained: Location in coordinates, number of people that is 60 years old or more, income and, of course, the postal code; so those rows that did not meet the requirements were discarded.

## 3.2 Exploratory Data Analysis (EDA)

The important parameters found in the EDA were the average per capita income in Toronto (calculated with the data of Toronto), the total quantity of people that is 60 years old or over and how much this number means compared to the raw total of people; the results were: 51,855 dollars, almost 581,000 persons above 59 years and this represents the 21.265% of the total population in Toronto.

## 3.3 Clustering

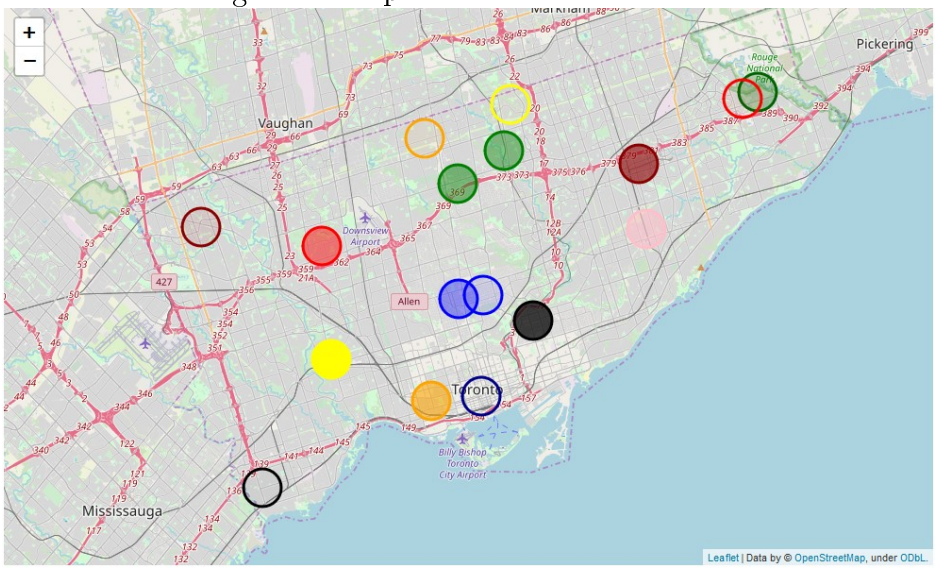
Once the data was ready, a segmentation of the market was needed, not only by physical proximity, but by their characteristics too, also it was a good idea to cluster in a way that could let us see how many clusters were optimal without them being changed each time the clusters were calculated, so K nearest neighbors was not optimal, DBSCAN puts the outliers in a separate group and that is not helpful in this case, so hierarchical clustering was a good option.

Figure 2: Data of the 16 initial clusters

| TEXT_NAME_NOM Cluster | Average total income | Latitude | Longitude | 60 years and over |
|---|---|---|---|---|
| 0 | 36714.636364 | 43.730530 | -79.508468 | 55170.0 |
| 1 | 55766.055556 | 43.647956 | -79.420629 | 58335.0 |
| 2 | 48616.916667 | 43.669826 | -79.493813 | 99540.0 |
| 3 | 132100.000000 | 43.701883 | -79.400751 | 35780.0 |
| 4 | 37506.333333 | 43.779744 | -79.368138 | 40910.0 |
| 5 | 35027.777778 | 43.738153 | -79.264532 | 59780.0 |
| 6 | 58571.076923 | 43.690087 | -79.346684 | 73765.0 |
| 7 | 36635.500000 | 43.772464 | -79.270002 | 60830.0 |
| 8 | 32552.000000 | 43.810330 | -79.183067 | 27685.0 |
| 9 | 113158.000000 | 43.650571 | -79.384568 | 145.0 |
| 10 | 129519.000000 | 43.785895 | -79.425376 | 5030.0 |
| 11 | 209748.000000 | 43.703792 | -79.383160 | 8220.0 |
| 12 | 33962.000000 | 43.803762 | -79.363452 | 2505.0 |
| 13 | 38431.000000 | 43.806686 | -79.194353 | 3935.0 |
| 14 | 45175.000000 | 43.762390 | -79.401912 | 45965.0 |
| 15 | 132686.000000 | 43.602414 | -79.543484 | 3380.0 |

The data was normalized and many maps and tables were made and analyzed, it was concluded that the optimal number of clusters was 16, because the further division of clusters only created uninteresant clusters. For the sake of visualization, *clusters with more possible clients have more opacity in its inside.*

Figure 3: Map of the 16 initial clusters




## 3.4 Analysis of the clusters

With the market segmentation done the analysis took place, the segment that was being looked for was one whose income was not low, had many potential clients and little gyms nearby.

The first filter was the amount of old people nearby, less than 10,000 was not considered optimal. The second filter was the income, less than the 90% than the average per capita would not be good, that left only 4 clusters in the game.

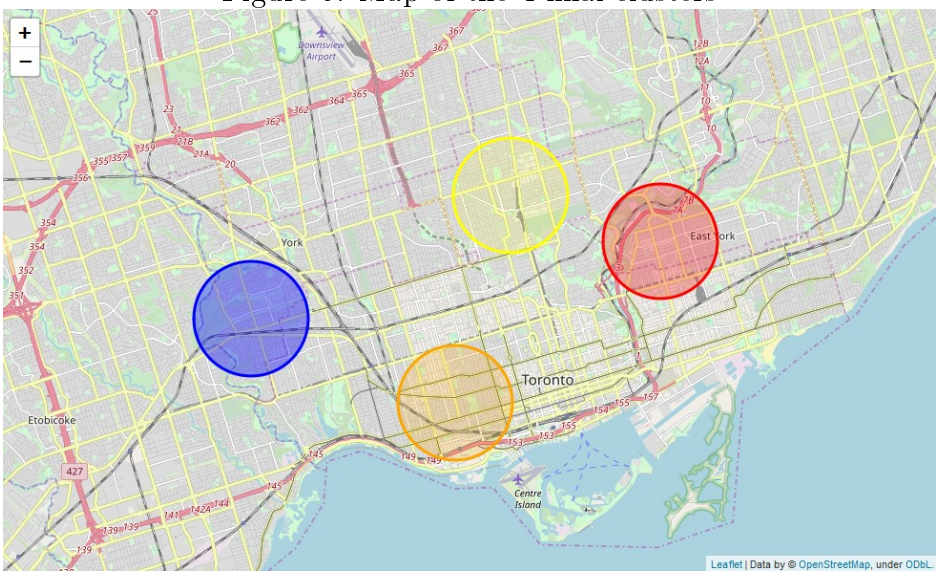
Figure 4: Data of the 4 final clusters

| | TEXT_NAME_NOM | Cluster | Average total income | Latitude | Longitude | 60 years and over |
|---|---|---|---|---|---|---|
| 0 | | 1 | 55766.055556 | 43.647956 | -79.420629 | 58335.0 |
| 1 | | 2 | 48616.916667 | 43.669826 | -79.493813 | 99540.0 |
| 2 | | 3 | 132100.000000 | 43.701883 | -79.400751 | 35780.0 |
| 3 | | 6 | 58571.076923 | 43.690087 | -79.346684 | 73765.0 |

If two clusters were geographically close to each other, then a gym between both could be a possibility; for example, between the yellow and the red

cluster (Figure 5). To match with the table, the orange cluster is the number 0, the blue one the number 1, the yellow is number 3 and the red is number 4.

Figure 5: Map of the 4 final clusters

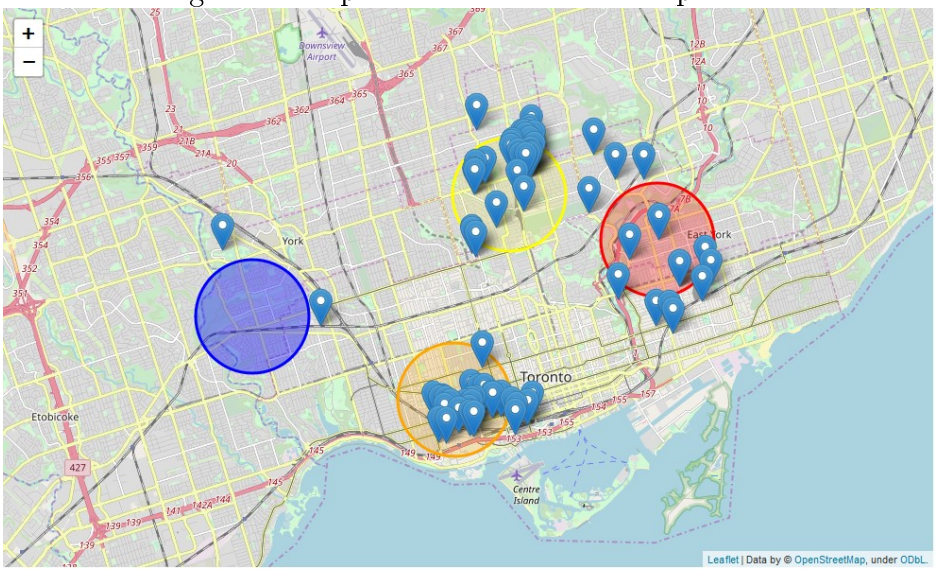



### 3.4.1 Market estimates and analysis

One of the main reason because of which someone goes to a given gym and not any other is the location, this means that having a lot of other gyms nearby could affect the project, even if the kind of clients is not the exactly the same.

For this part the*Foursquare* service was used. The results of nearby gyms was very informative, to the point is almost possible to discard location just for the strong density of gyms in the area. In the yellow cluster there were 32 gyms nearby, in the orange cluster there were 50, in the red cluster there were 14 and in the blue one only 2.

Figure 6: Map of clusters and its competence




# 4 Results

In summary, the yellow cluster has many potential clients and these are wealthy, but there is too much competence nearby, and the approach is not chasing rich people, just avoiding low income zones, so this cluster is not good.

The red cluster has more potential clients than the yellow one, their income is much lower but is not a low income zone and there is less than a half of the competence than in the last cluster, this cluster is cool.

The orange cluster has less potential clients than the red cluster, these persons income is lower compared to the first two options and there is a lot more competence than in the red zone (more than 3.5 times more), so is not a good option either.

At last, the blue cluster has many potential clients, 99540 according to the census, there is almost no competence nearby, but their income is low compared to the others.

# 5 Discussion

The yellow and orange clusters may not be good options for this project. The red clusters seems to be a good place to open the gym because of its people and other gyms nearby so, for me, *the red cluster is the winner*.

The blue cluster is a special case, because is uncertain if these people can afford this idea, however there is a LOT of people over 59 years old and they are not very poor, so surely if this cluster is not fit to have the planed gym into it (it may still be), another kind of project could be developed (like pilates or spinning). But to know these things a deeper study is necessary.

# 6 Conclusion

Even if many statistical assumptions (like using average numbers) were made for the sake of simplicity, the results sure have truth in them, so a decision based on this new information can be considered an informed one. On the other hand many new question arose, like the potential of the blue cluster, for me that means that this is the right line of questioning.

# References

Canada, S. (2019, Nov). *2016 census profile web data service (wds) - user guide.* Retrieved from https://www12.statcan.gc.ca/wds-sdw/cpr2016-eng.cfm