Clustering Barcelona neighborhoods

Ian Riera Smolinska

April 18, 2020

Contents

1 Introduction

1.1 Background

Historically cities have been divided administratively into districts and neighborhoods, and so is the case of Barcelona. The city is divided into 10 districts and 73 neighborhoods. However, society is constantly changing and the demography of the cities as well. In the case of Barcelona, there was a huge before and after the Olympic games back in 1992 and nowadays is the destination for 20 million tourists every year. In fact, many decided not just to visit it but to move to Barcelona, as the statistics recall that, in 2019, 20 % of the inhabitants of the city were foreigners. [1]

This constant flux of travelers makes opening a restaurant, shop and other commercial venues, a tempting investment to get profit from the visitors. In 2019, there were 80.500 street-level commercial venues in Barcelona. [2]

As a consequence of the previous two, a huge amount of user rating on local venues is generated every day in platforms like Tripadvisor or Foresquare.

1.2 Problem

The administrative division of the city provides a handful distribution, but not a representative one of the inhabitants' characteristics. The economic power, commercial sectors, age or unemployment among others, would provide a richer approach to segment the neighborhoods of the city. This project aims to cluster the neighborhoods of the city basing on different parameters, detect if there are any correlations between them and test how fit a user-generated social media platform as Foursquare would be for segmenting a city.

1.3 Interest

The target group for this analysis is any company or individual that is considering opening a commercial venue in Barcelona. Basing on the existing data, it can compare how the similar venues are distributed around the city and the characteristics of each neighborhood, choosing to settle in the one that fits more their products.

Note that this data is previous to the Covid-19 pandemic, that might carry an economical crisis and transformation. Several venues might close due to it. This project will be retaken in 2021, when data from 2020 is available and compared with the obtained results for 2019.

2 Data preparation

2.1 Data source

The Barcelona City Council publish statistics openly on their webpage. This includes demographic, economic, political and laboral information grouped by neighborhoods. The following parameters for each neighborhood have been used in this project:

- Nationality. [1]
- Unemployment. [3]
- Local elections voting. [4]
- Age. [5]
- Available Family Income index. [6] [7]
- Street-level commercial venues. [2]
- Covid [8]

In order to be able to work on these datasets, the Beautiful Soup [9] library for web scrapping has been used to obtain the information to fill the created Panda data frames [10].

Additionally, the Foresquare API [11] is used to access the platform information.

2.2 Data processing

First of all, the information published by the Barcelona's city council is either in Catalan or Spanish, so all the tables had to be translated to English. Additionally, the table for commercial had the column and rows transposed so had to be corrected.

For the data analysis, the format of the provided tables allows an easier extraction of information, but in order to segment the population some restructuring is required. For example, the age information is presented by individual years and it will have to be grouped in decades for segmenting. Moreover, the average age for each neighborhood is calculated, as the information is presented as number of individuals for each age.

Available Family Income is an index that represents the clean income a family have after taxes, to either spend or save. In this case the latest information published by neighborhood dates back to 2017. The index for Barcelona is 100, with a reference value of 22.390 € per person. [7] The information about the neighborhoods is presented in relation to the Barcelona index, so the value in euros has to be calculated from the reference value.

For the political scenario, two different approaches can be taken: grouping by left-center-right wings, as of they define themselves, or grouping by Catalan nationalists or Spanish nationalists. For the clustering, the first division is used, counting left as -1 and right as +1 and obtaining the numerical bias.

Finally, we have two datasets that contain information about commercial venues. From the Foursquare API we will need to calculate the more repeated venues per neighborhood and select the top 10 for each one in descending order. From the city council source, two data tables are presented, one with absolute values and another with percentual. The second one will be used.

Once the different segmentations with each single parameter are completed, a new table with the neighborhoods and the corresponding cluster index per segmentation is created and a correlation matrix generated to detect new data relationships and show their scatter plots.

3 Data analysis

The current administrative division of the city of Barcelona is formed by 10 districts and 73 neighborhoods grouped as shown in the following table:

Number	District	Size km²	Neighborhoods
1	Ciutat Vella	4.49	La Barceloneta, El Gòtic, El Raval, Sant Pere, Santa Caterina i la Ribera
2	Eixample	7.46	L'Antiga Esquerra de l'Eixample, La Nova Esquerra de l'Eixample, Dreta de l'Eixample, Fort Pienc, Sagrada Família, Sant Antoni
3	Sants-Montjuïc	21.35	La Bordeta, la Font de la Guatlla, Hostafrancs, la Marina de Port, la Marina del Prat Vermell, El Poble-sec, Sants, Sants-Badal, Montjuïc*, Zona Franca - Port*
4	Les Corts	6.08	les Corts, la Maternitat i Sant Ramon, Pedralbes
5	Sarrià-Sant Gervasi	20.09	El Putxet i Farró, Sarrià, Sant Gervasi - la Bonanova, Sant Gervasi - Galvany, les Tres Torres, Vallvidrera, Tibidabo i les Planes
6	Gràcia	4.19	Vila de Gràcia, el Camp d'en Grassot i Gràcia Nova, la Salut, el Coll, Vallcarca i els Penitents.
7	Horta-Guinardó	11.96	El Baix Guinardó, El Guinardó, Can Baró, El Carmel, la Font d'en Fargues, Horta, la Clota, Montbau, Sant Genís dels Agudells, la Teixonera, La Vall d'Hebron.
8	Nou Barris	8.04	Can Peguera, Canyelles, Ciutat Meridiana, La Guineueta, Porta, La Prosperitat, les Roquetes, Torre Baró, la Trinitat Nova, El Turó de la Peira, Vallbona, Verdum, Vilapicina i la Torre Llobeta
9	Sant Andreu	6.56	Baró de Viver, Bon Pastor, El Congrés i els Indians, Navas, Sant Andreu de Palomar, La Sagrera i Trinitat Vella
10	Sant Martí	10.80	El Besòs i el Maresme, el Clot, El Camp de l'Arpa del Clot, Diagonal Mar i el Front Marítim del Poblenou, el Parc i la Llacuna del Poblenou, Poblenou, Provençals del Poblenou, Sant Martí de Provençals, La Verneda i la Pau, la Vila Olímpica del Poblenou

Table 1: Administrative division of Barcelona. [12]

The statistics presented by the city council, present the districts with the number id and not the name. Therefore, from now on only the number of the districts will be used as id for the processing.

Each division of Barcelona will be represented as a choropleth map using the Folium library, and therefore the first division that needs to be represented is the actual administrative one.

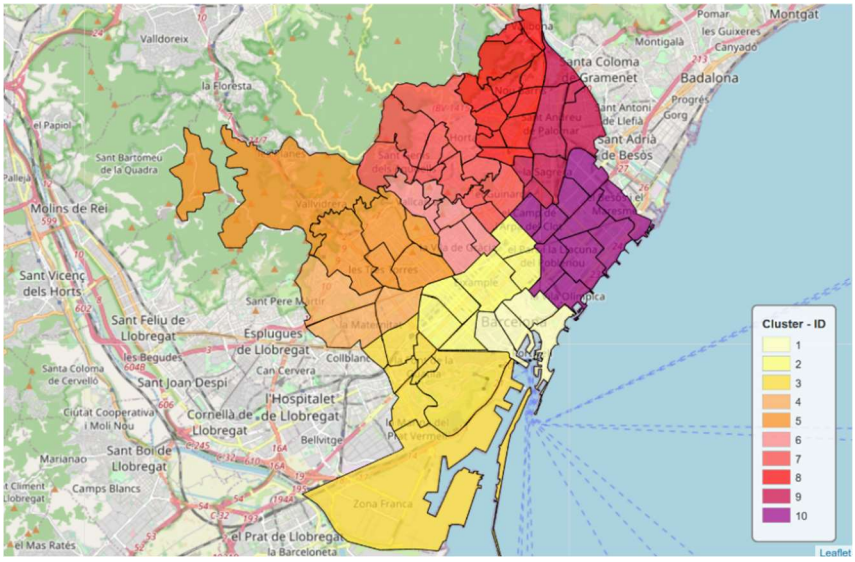


Figure 1: Administrative division of Barcelona

With a first analysis basing on the population, we can obtain the maximum and minimum populated neighborhoods and extract that:

The most populated neighborhood in Barcelona is 'La Nova Esquerra de l'Eixample' with the 3.55 of the population, 58032 inhabitants.

The least populated neighborhood in Barcelona is 'La Clota' with the 0.04 of the population, 683 inhabitants.

We can also do a first alternative clustering of the neighborhoods basing on the population and represent the new map division.

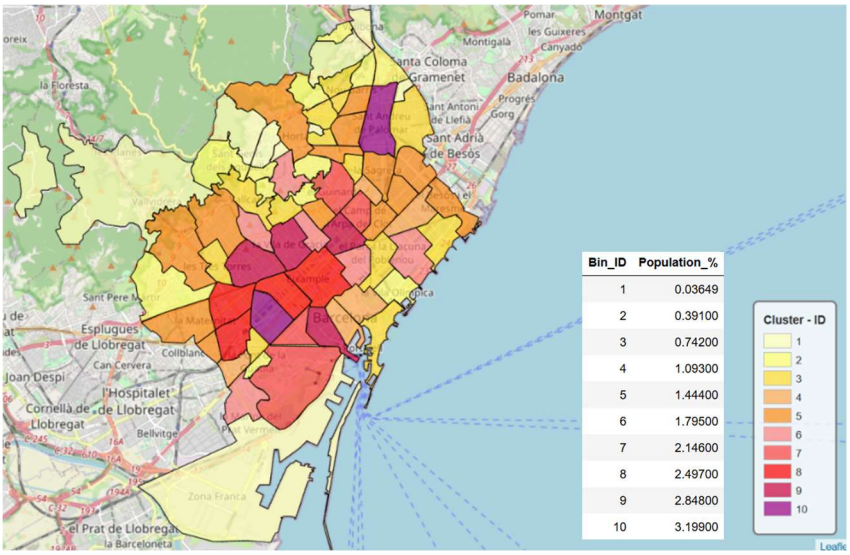


Figure 2: Division by population %.

The following lines present an insight on the demography of the neighborhoods, and the clustering basing on those parameters.

3.1 Age

First, we will analyze the age of the population of Barcelona. The mean age in every neighborhood is calculated and the minimum and maximum calculated:

The neighborhood in Barcelona with the oldest population is 'Canyelles 'with a mean age of 48.

The neighborhood in Barcelona with the youngest population is 'el Raval' with a mean age of 38.

The age pyramids of both neighborhoods are plotted and compared with the total of Barcelona.

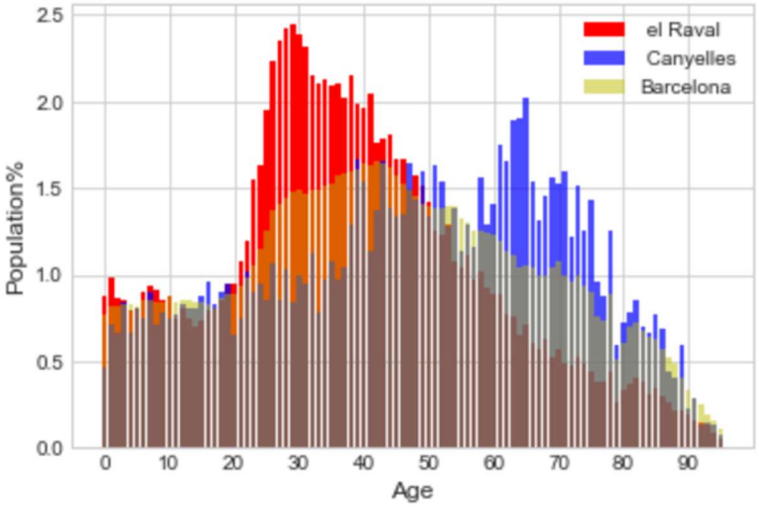


Figure 3: Age pyramids comparison

A segmentation of the city is completed basing on the mean age of every neighborhood, into equidistant 10 bins.

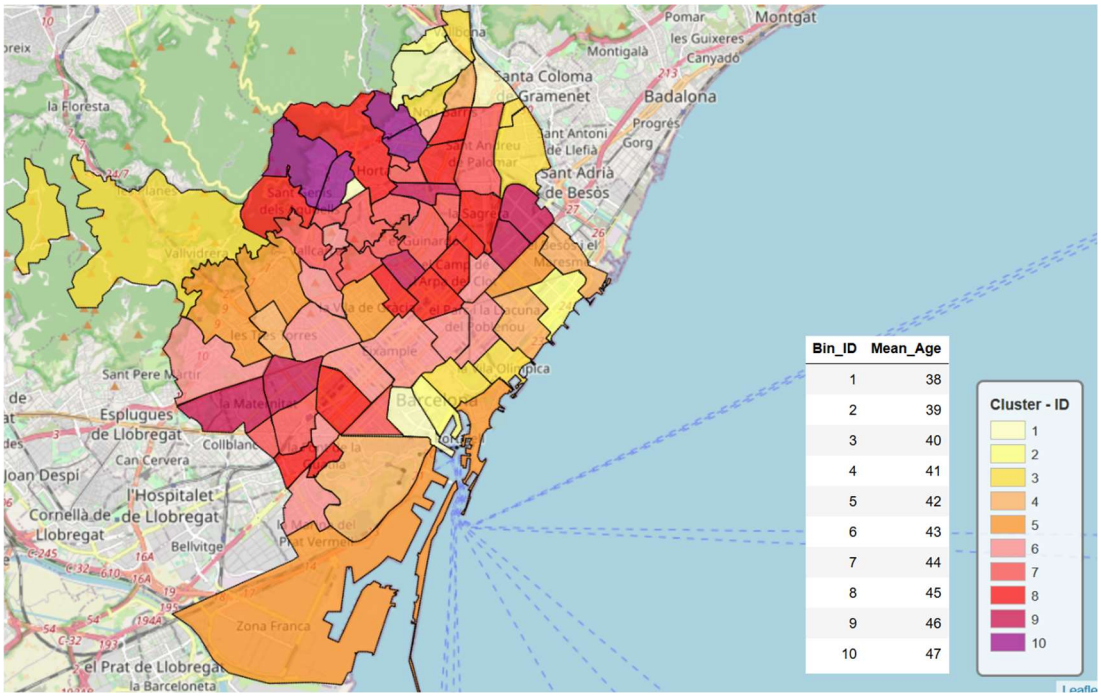


Figure 4: Segmentation by mean age.

3.2 Nationality

Next, we analyze the nationalities of the citizens of Barcelona. The percentage of immigration by neighborhood is calculated and we can extract that:

The neighborhood with more diversity in Barcelona is 'el Barri Gòtic' with the 55.00% of the population being foreigner, 24199 inhabitants.

The neighborhood with more spaniards in Barcelona is 'Canyelles' with just the 6.00% of the population being foreigner, 94 inhabitants.

To make more visual the diversity in Barcelona, a word cloud with the names of the countries of origin at scale by number of immigrants is represented.

Figure 5: Word cloud of nationalities.

We can observe that Italy is the biggest contributor, followed by China, Pakistan and France. Note that none of them is a former Spanish colony, however we can spot several of them in the top 10.

In the map representation we can see that the immigrants reside mainly in the old town.

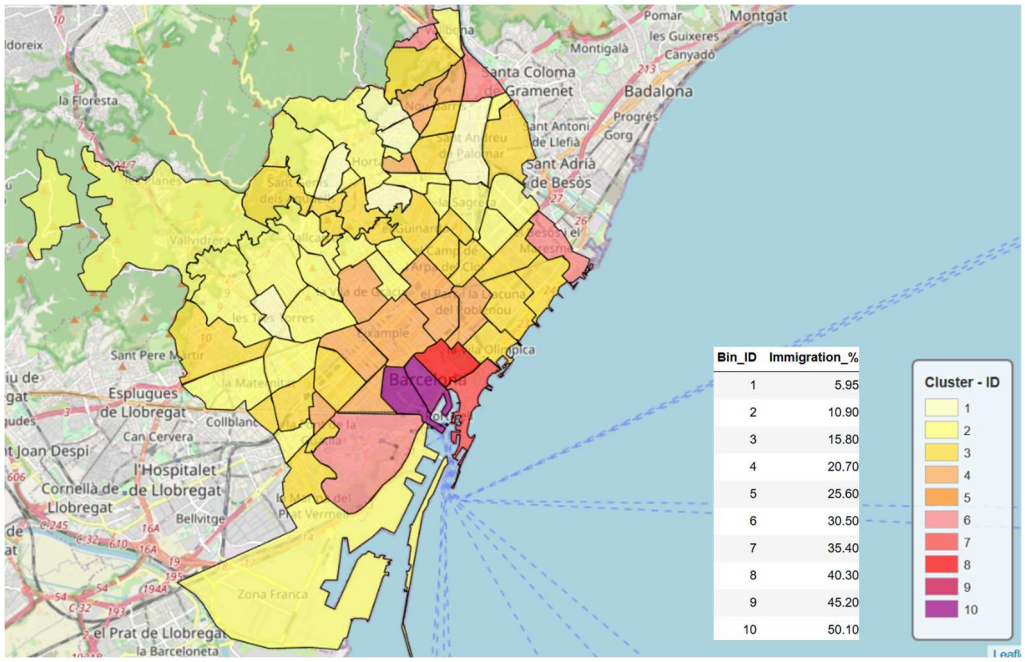


Figure 6: Segmentation by immigration %.

3.3 Political ideology

The 2019 Barcelona local elections voting will be analyzed by participation percentage and political bias. The political bias will be calculated by multiplying the left parties votes by -1 and right parties votes by 1. The grouping of the political parties is the following:

- left = [ERC-AM, BEC-ECG, PSC-CP, CUP-AMUNT, EVE, PACT, UNIDOS SI, PCPC, PUM+J, P-LIB, dCIDE]
- right = [BCN Cs, JUNTS, PP, BCAP, VOX, FC's, CNV , PFiV, FE JONS]

In terms of participation we extract that:

The neighborhood with more participation on the local's elections is la Vila Olímpica del Poblenou where the 77.21% of the population voted.

The neighborhood with less participation on the local's elections is la Marina del Prat Vermell - Zona Franca where the 40.15% of the population voted.

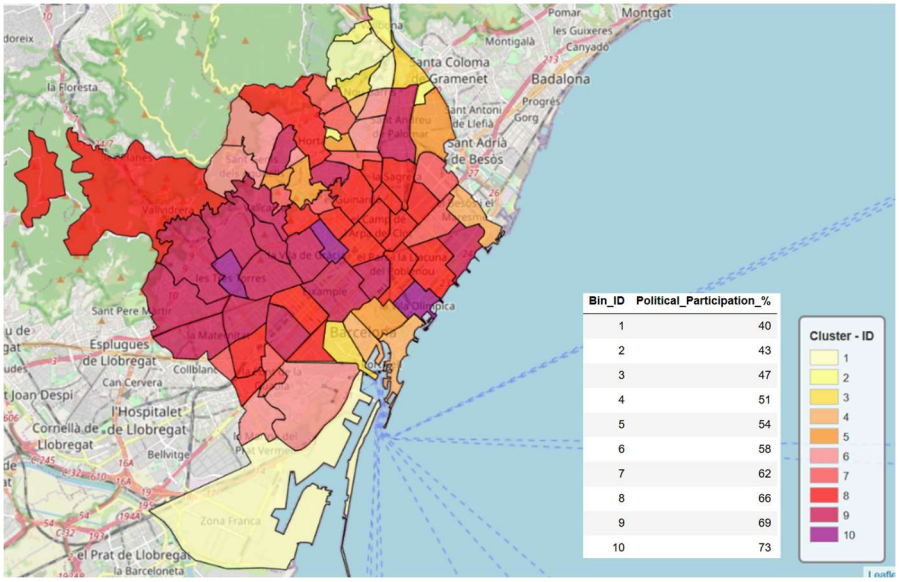


Figure 7: Segmentation by participation %.

Basing on the results of those who voted:

The neighborhood with a bias closer to the right-wing ideology is Pedralbes.

The neighborhood with a bias closer to the left-wing ideology is la Clota.

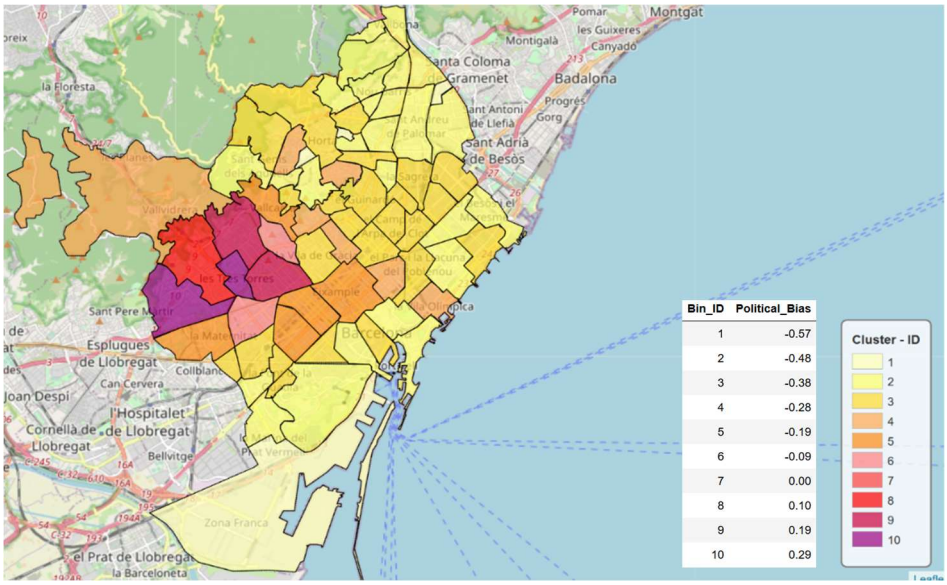


Figure 8: Segmentation by political Bias. -1 left, 1 right

3.4 Household income

Now we will focus on the economical data. The first parameter is the household income, which calculates the average available money after taxes, that every person has to spend on their own or save. The available information dates back to 2017, when the reference index 100 was equivalent to 22390 €.

Calculating the household income in euros for every neighborhood we can extract that:

The neighborhood with the highest available income is Pedralbes with an average of 55706.00 euros per habitant.

The neighborhood with the lowest available income is Ciutat Meridiana with an average of 8642.00 euros per habitant.

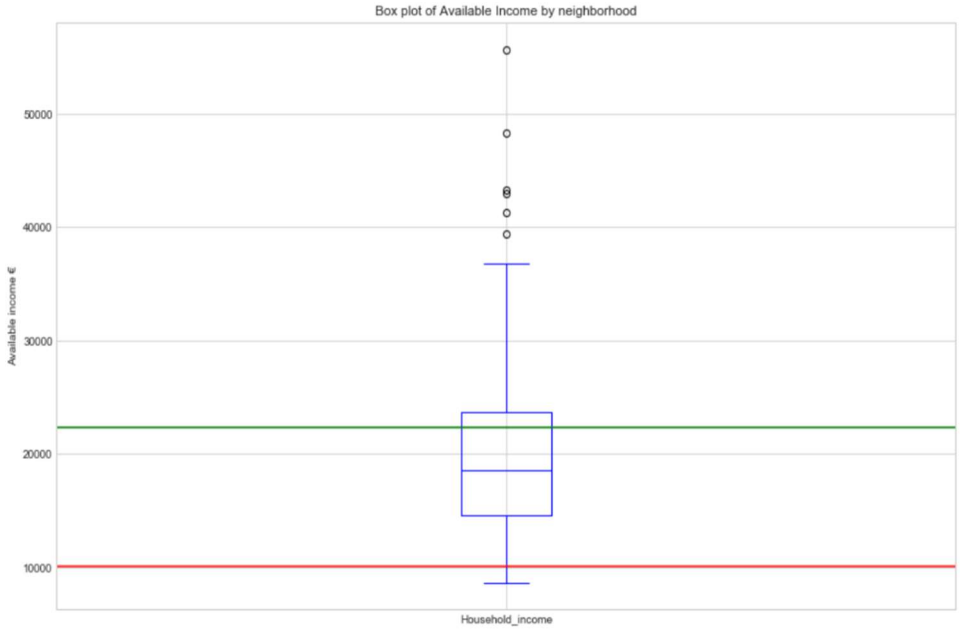


Figure 9: Household income box plot

The green line represents the RFD for Barcelona and the red one the poverty threshold for a home with just one adult living in it.

We can see in the box plot that most of the 75% percentile is barely above the index for the city, and the richest outliers have a household more than two times than three quarters of the neighborhoods. Let's create a bar plot to zoom in. The yellow lines represent the district aggrupation.

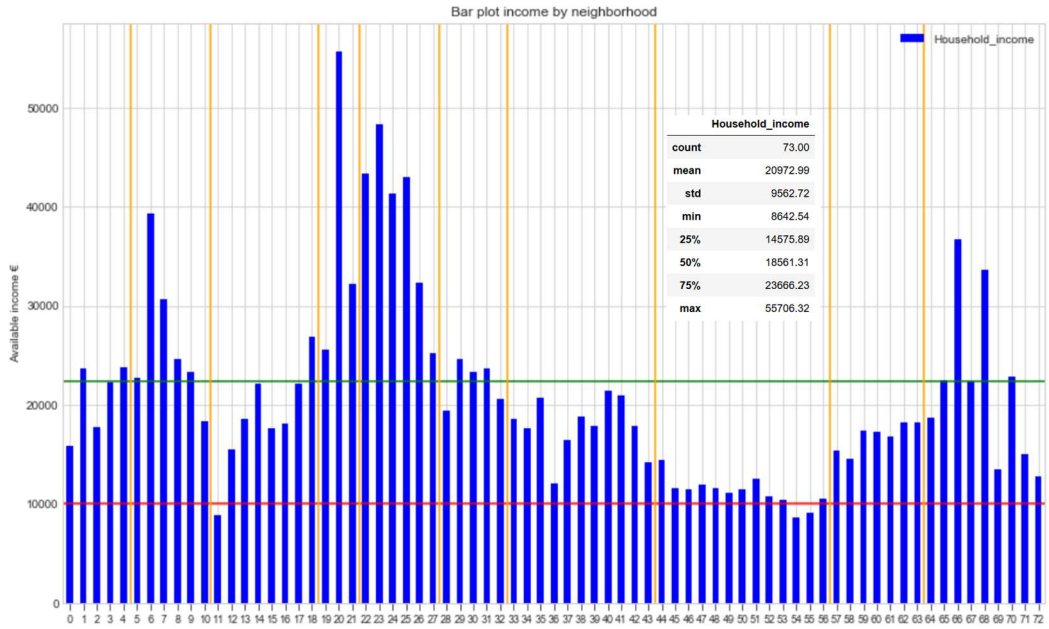


Figure 10: Household income by neighborhood.

We can observe that one out of two neighborhoods is below the city indicator, one out of four is closer to the poverty threshold than to the city average, and only 25% of the city is above the average.

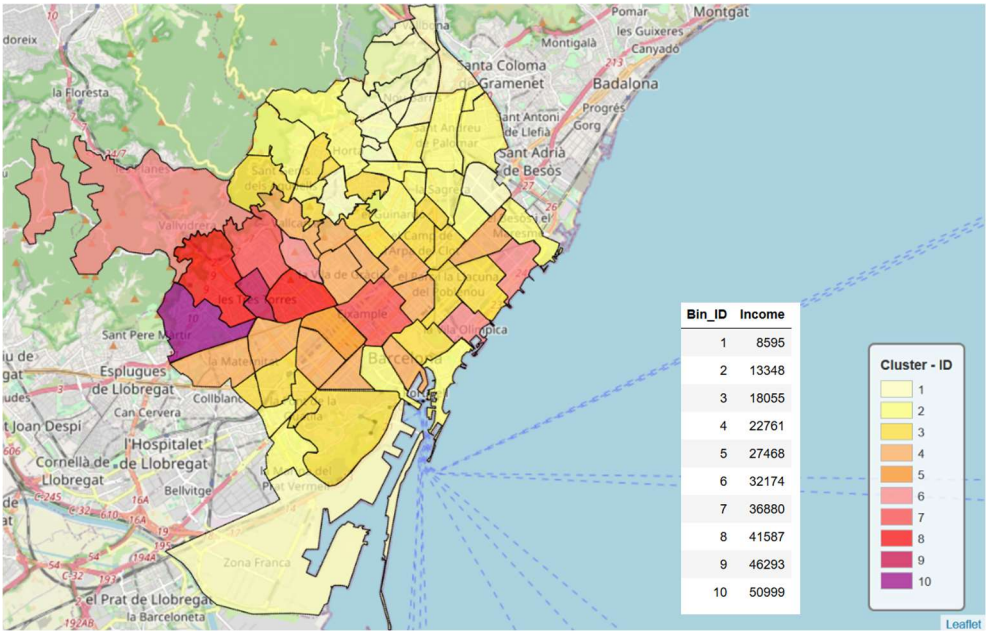


Figure 11: Segmentation by household income.

3.5 Unemployment

In terms of unemployment percentage:

The neighborhood with more unemployment in Barcelona is Ciutat Meridiana with the 12.55% of the population unemployed, 308229 potential workers.

The neighborhood with less unemployment in Barcelona is Vallvidrera. el Tibidabo i les Planes with the 2.87% of the population unemployed, 1604 potential workers.

We can define a box plot by moth to check out how evolves along the year.

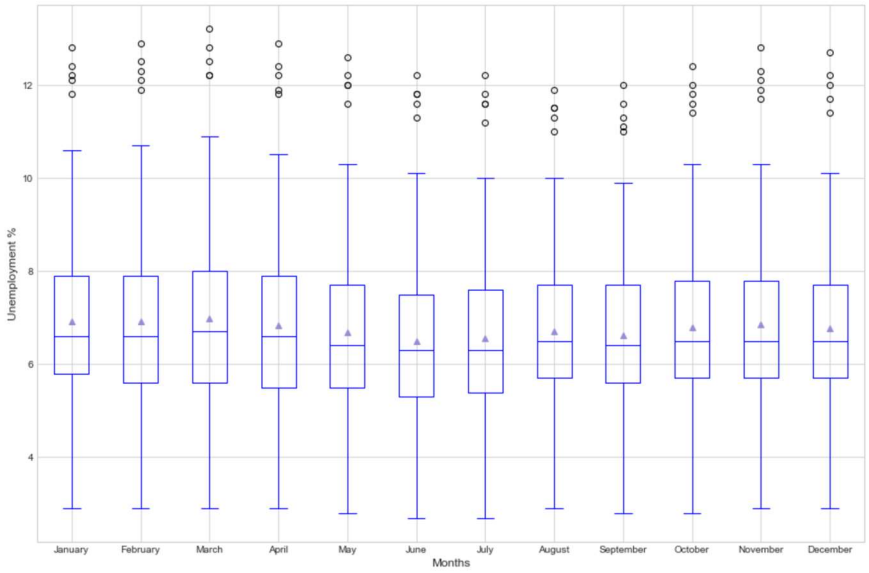


Figure 12: Unemployment evolution along the year.

	January	February	March	April	May	June	July	August	September	October	November	December
count	73.00	73.00	73.00	73.00	73.00	73.00	73.00	73.00	73.00	73.00	73.00	73.00
mean	6.92	6.92	6.99	6.83	6.68	6.50	6.56	6.70	6.62	6.78	6.86	6.78
std	2.17	2.20	2.27	2.18	2.15	2.09	2.05	1.96	1.96	2.07	2.14	2.08
min	2.90	2.90	2.90	2.90	2.80	2.70	2.70	2.90	2.80	2.80	2.90	2.90
25%	5.80	5.60	5.60	5.50	5.50	5.30	5.40	5.70	5.60	5.70	5.70	5.70
50%	6.60	6.60	6.70	6.60	6.40	6.30	6.30	6.50	6.40	6.50	6.50	6.50
75%	7.90	7.90	8.00	7.90	7.70	7.50	7.60	7.70	7.70	7.80	7.80	7.70
max	12.80	12.90	13.20	12.90	12.60	12.20	12.20	11.90	12.00	12.40	12.80	12.70

Table 2: Unemployment statistics by month.

We can observe that in general terms the unemployment behaves in the same way, decreasing during spring and summer which are the touristic high season, and increasing afterwards with a Christmas break. However, in August, while the maximum keeps decreasing, the min and mean values increase as many businesses close for holidays the whole month. We can see that in September recovers as new season begins.

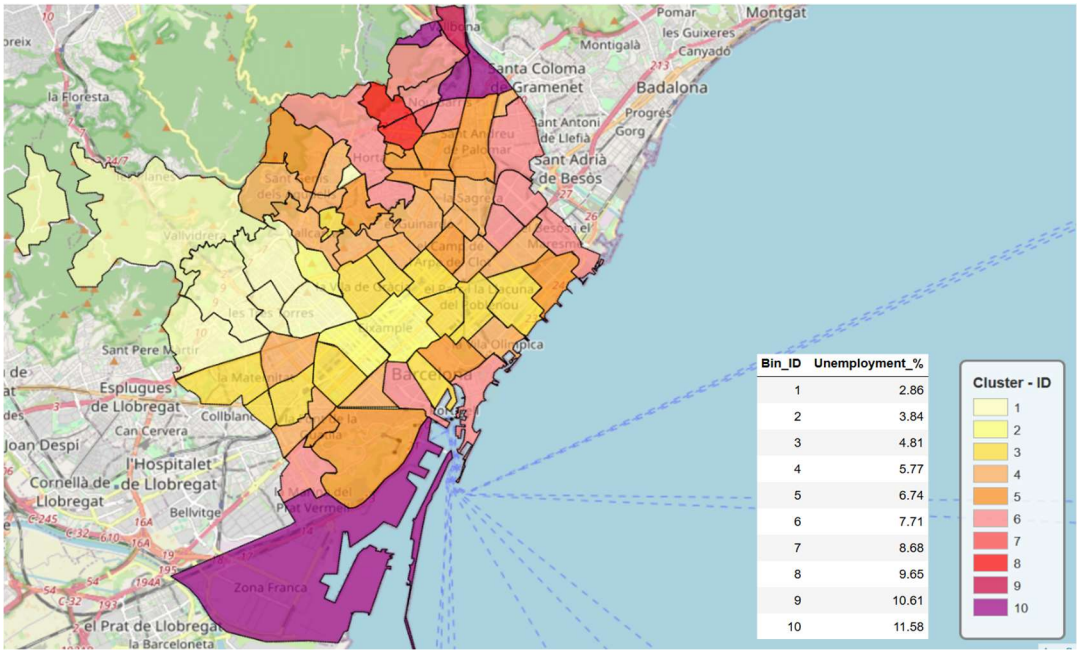


Figure 13: Segmentation by unemployment %.

3.6 Street-level commercial venues

The frequency of each category of commercial venue in each neighborhood is calculated and a top 10 presented. Then we account how many times a category inside the top 5 of each neighborhood:

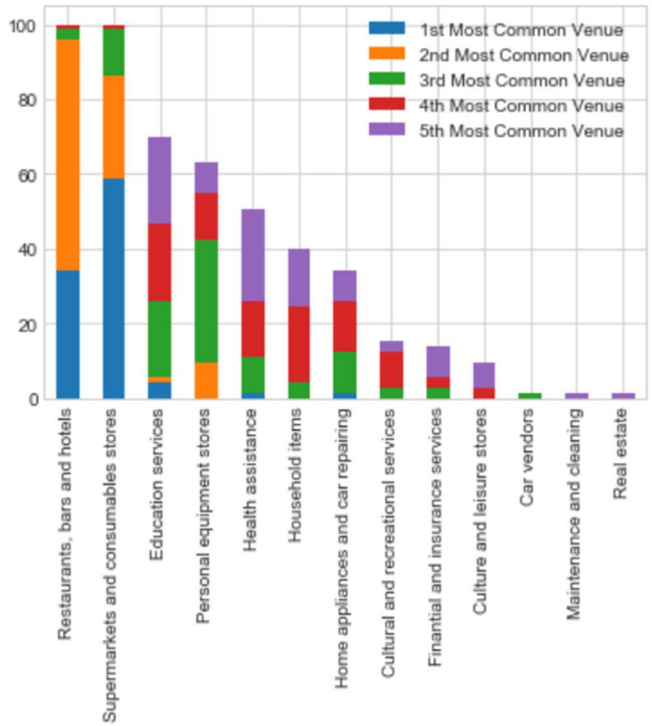


Figure 14: Most common commercial venues categories by neighborhood

As we can see all the neighborhoods have ‘Restaurants, bars and hotels’ and ‘Supermarkets and consumable stores’ in between their top 5 most frequent commercial venues.

K-means algorithm and sklearn library will be used for clustering the neighborhoods by commercial venues. The resultant clusters satisfy that **all the neighborhoods in the cluster** satisfy that in their top 10 of most frequent commercial venue categories:

- A: Top 1 is 'Restaurants, bars and hotels':
    - A.1: Top 2 is 'Supermarkets and consumable stores':
        - Cluster 4: 'Education services' between top 3 and top 5. 'Health assistance' between top 5 and top 7. All contain 'Culture and leisure stores' from top 5, 'Personal equipment stores' from top 7 and 'Finantial and insurance services' from top 8.
        - Cluster 10: 'Home appliances and car repairing' from and mainly as top 3, 'Education services' between top 4 and 5, and all contain 'Finantial and insurance services' and 'Health assistance' in the top 10.
    - A.2: Top 2 and top 3 contain 'Supermarkets and consumable stores' and 'Personal equipment stores':
        - Cluster 5: 'Household items' between top 4 and top 5. 'Culture and leisure stores' between top 7 and top 8. All contain 'Education services' from top 5, 'Finantial and insurance services' from top 6 and 'Home appliances and car repairing' from top 9.
        - Cluster 9: **Contains 'Cultural and recreational services'** in their top 10. Also all contain,'Household items', 'Health assistance' and 'Culture and leisure stores' from top 4.
- B: Top 1 is 'Supermarkets and consumable stores' and top 2 is 'Restaurants, bars and hotels':
    - Cluster 2: 'Personal equipment stores' is top 3 and 'Education services' between top 3 and top 5. 'Health assistance' between top 4 and top 5, and 'Household items' from top 6.
    - Cluster 3: 'Personal equipment stores' between top 3 and 5. 'Household items' between top 4 and 7. 'Health assistance' and 'Home appliances and car repairing' between top 6 and 8.'Education services' from the top 5.
    - Cluster 7: 'Education services' between top 3 and 4. 'Health assistance' between top 3 and top 6.'Home appliances and car repairing' between top 4 and 8.
- Cluster 1: This cluster has just one neighborhood, which has 'Home appliances and car repairing' as top 1, hotels as top 2 and car vendor as top 3. Two of the top3 are automobile related, so this cluster can be defined by that.
- Cluster 6: The top 1 and 2 are shared by 'Restaurants, bars and hotels' and 'Supermarkets and consumable stores'. 'Health assistance' between top 5 and top 6. 'Household items' between top 4 and 7, 'Culture and leisure stores' from top 4, and 'Education services' between top 3 and top 9, mainly in top 7.
- Cluster 8: **Contains 'Real estate'** in their top 10. The top 1, 2 and 3 are shared by 'Restaurants, bars and hotels', 'Supermarkets and consumable stores' and 'Personal equipment stores'. Als contains 'Health assistance' between top 3 and top 6, and 'Finantial and insurance services' between top 5 and top 9.

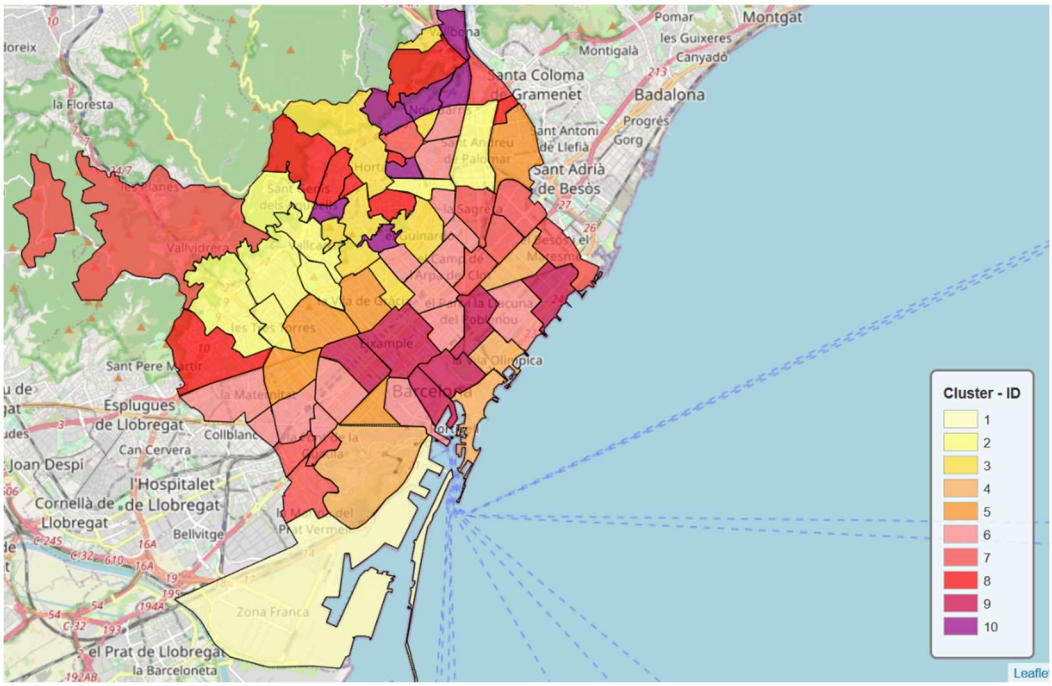


Figure 15: Division by commercial venues clustering

3.7 Foresquare venues

The API is used to explore and obtain commercial venues in the latitudes and longitudes of the neighborhoods. Then we will compare the extracted information with the one in the commercial scope.

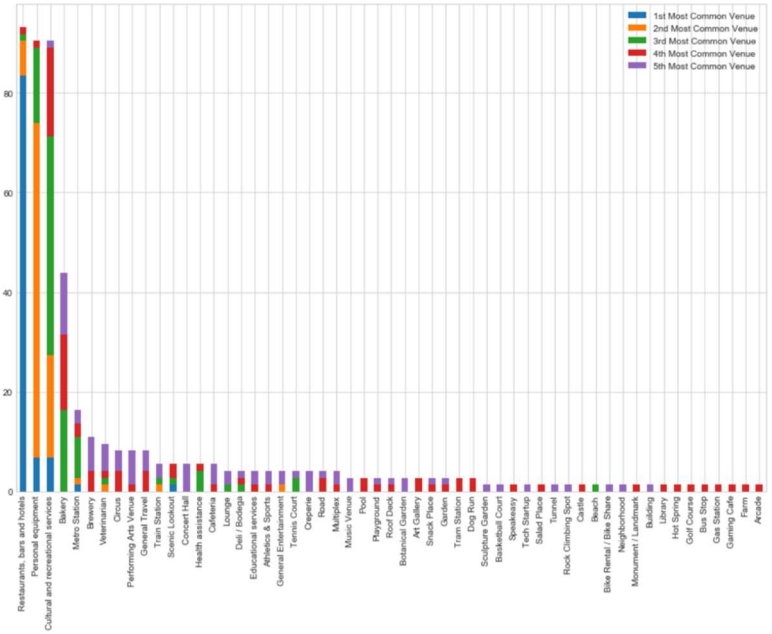


Figure 16: Foresquare most common venues by neighborhood.

First all we had to group all the restaurants and similar venues as Foresquare returned 260 categories for the 14 we had the previous scope. Basing on the graph, we can see that mainly it returned gastronomic, leisure and personal equipment kind of services, the main attractions for visitors. However, it is hard to find household items, educational and health services and more day-to-day commercial venues. In other words, we already miss half of the venues of the city in front of the official commercial venues data presented by the city council. We run the k-means clustering on the Foresquare information.

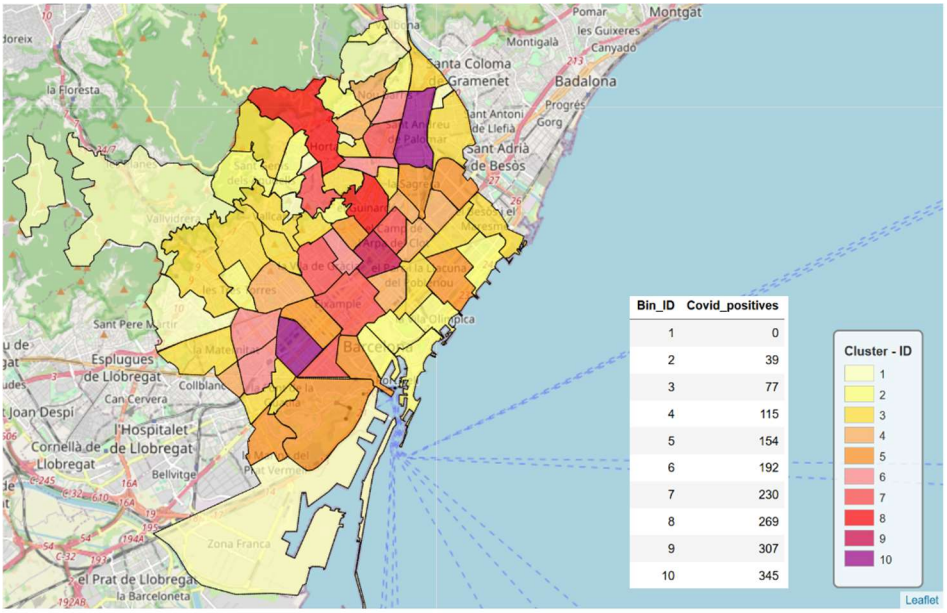
3.8 COVID-19

With the COVID-19 pandemic, loads of information appear daily and this includes the positive cases in Barcelona by neighborhood, up to April 19th 2020. And in the previous cases, we can extract that:

The neighborhood with more positive cases of Covid-19 in Barcelona is Sant Andreu with 384 diagnosed cases.

The neighborhood with more positive cases of Covid-19 in Barcelona is la Clota with 1 diagnosed case.

If we represent it on the city map, we obtain:



Bin_ID	Covid_positives
1	0
2	39
3	77
4	115
5	154
6	192
7	230
8	269
9	307
10	345

Figure 17. Positive cases of COVID-19

4 Results

All the effort on analyzing and segmenting the city of Barcelona into the previous steps, was the preparation for the real goal, which is finding the correlations among them. So, after clustering the city based on different parameters, we will use these values to calculate the correlation heatmap with seaborn and the scatter plot will be calculated for the high absolute value correlations. These plots will set a base for in future development of the project, create regression models and predictions from it.

We calculate the correlation matrix and represent it as heatmap:

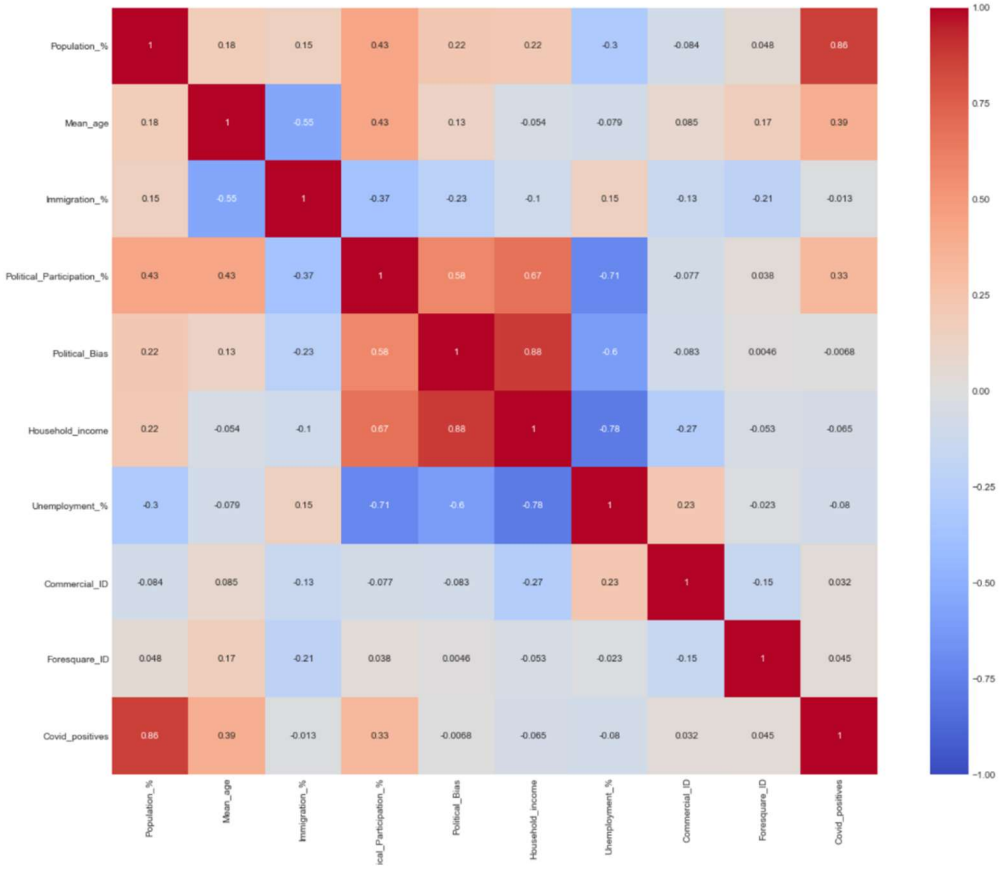


Figure 18: Correlation heatmap of the clustered parameters

By category and considering correlations absolute values over 0.5, we can observe:

- Population %: Logically, we observe a high correlation with the amount of positive cases in covid.
- Mean_age: There is a weak negative correlation between mean age and immigration %.
- Political Participation: We can see a mild correlation of political participation with political bias and household income. There is also a negative correlation with unemployment %.
- Political bias: There is a high correlation between the political bias and the household income. There is also a negative correlation with unemployment %.
- Household income: In addition to the already mentioned, there is a logical strong negative correlation with the unemployment %.
- In terms of Foresquare in relation to official data, there is absolute non correlation, and therefore we can discard Foresquare as a good option for clustering and representing the city.

Now let's see the scatter plots of the mentioned correlations.

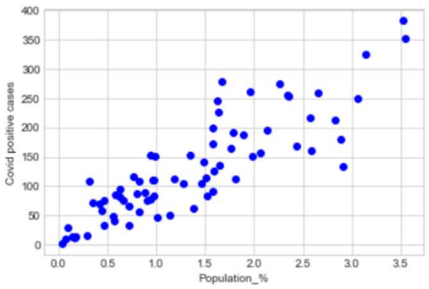


Figure 19: Population % vs Covid

We can observe almost a lineal relationship. A polynomial regression or nonlinear regression could make a good fit.

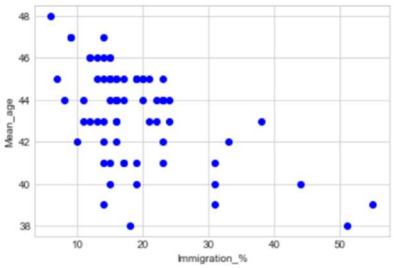
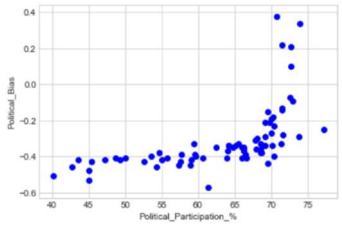
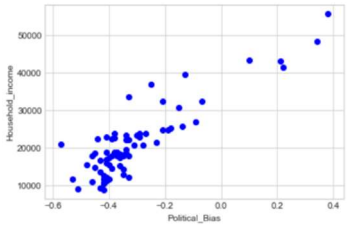
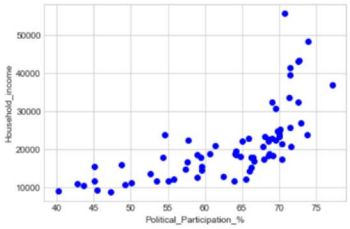


Figure 20: Mean age vs immigration %

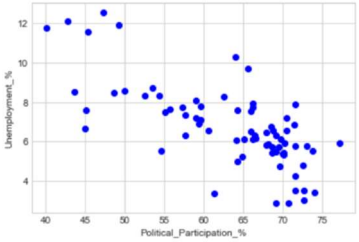
In this case the tendency is not as clear, although the maximums draw a diagonal upper bound.

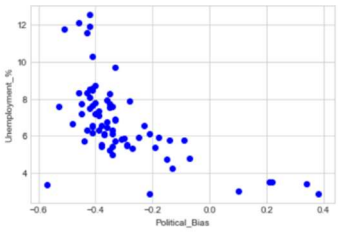


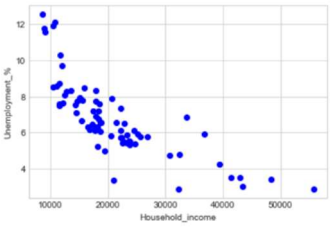




First, we can see when participation increases, the bias falls to the right wing. An exponential or logistic function would provide a good fit. Second, we can see a clear tendency to the right-wing ideology as the household income increases. A non-busted myth. We can observe, as there is a slightly lineal relationship between the political bias and household income, the graphic between the participation and household follows the same shape than the one between bias and participation.







And as one would imagine, the increasement in unemployment generates a decrease in household income. And we found a relationship between political bias and household income, being the rich right winged and the not that rich left winged. And the first correlation we observed was that bias was related to participation. All in all, we find that unemployment and household income, and political participation and bias are correlated and could be modeled for prediction with nonlinear regression functions.

5 Conclusion and further development

This analysis shows that the city of Barcelona can hardly be clustered based in all the demographic parameters, as for each of them you get a different painting. However, a correlation was found between the tuple household income and unemployment, the political participation and the political bias.

With these correlations, a modelling could be applied so, for example, a political party could estimate the potential number of votes that could get in each neighborhood.

In terms of the suitability of Foursquare to cluster a city, it proved to misrepresent the reality of the city, providing a useful tool for tourists and other visitors, or even for local citizens leisure time.

This project aims to be repeated step by step next year to observe the effects of the coronavirus and test if models based on the mentioned correlations would provide accurate predictions.

6 Bibliography

[1] "Barcelona city council. Statistics by neighborhood: Nationality.," 2019. [Online]. Available: https://www.bcn.cat/estadistica/castella/dades/barris/tpob/pad/ine/a2019/ine17.htm.

[2] "Barcelona city council. Statistics by neighborhood: Commercial venues.," 2019. [Online]. Available: https://www.bcn.cat/estadistica/castella/dades/barris/economia/tacteco/gacbarri19.htm.

[3] "Barcelona city council. Statistics by neighborhood: Unemployment.," 2019. [Online]. Available: https://www.bcn.cat/estadistica/castella/dades/barris/ttreball/atur/durada/durbar19.htm.

[4] "Barcelona city council. Statistics by neighborhood: Elections results.," 2019. [Online]. Available: https://www.bcn.cat/estadistica/castella/dades/barris/telec/loc/a2019.htm.

[5] "Barcelona city council. Statistics by neighborhood: Age.," 2019. [Online]. Available: https://www.bcn.cat/estadistica/castella/dades/barris/tpob/pad/ine/a2019/ine02.htm.

[6] "Barcelona city council. Statistics by neighborhood: Family income available.," 2019. [Online]. Available: https://www.bcn.cat/estadistica/castella/dades/barris/economia/renda/rdfamiliar/a2017.htm.

[7] "Diputació de Barcelona. Hermes project.," 2017. [Online]. Available: https://www.diba.cat/hg2/presentacio.asp?prId=888&idioma=cat&codi_any=2017&format=pantalla.

[8] "Covid_Barcelona," 19 04 2020. [Online]. Available: https://aspb.shinyapps.io/COVID19_BCN/.

[9] "Beautiful Soup Documentation.," [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[10] "Panda DataFrame.," [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html.

[11] "Foresquare developer," [Online]. Available: https://developer.foursquare.com/.

[12] "Wikipedia: Districts of Barcelona," [Online]. Available: https://en.wikipedia.org/wiki/Districts_of_Barcelona.