Analysis of neighbourhoods in Barcelona to open an ecological supermarket

Guillermo Fantoni

May 20, 2021

1.-Introduction

1.1 Background

Barcelona is the second city in Spain with 1.6 million people. It is one of the cities with the most trade and tourism in Europe. Its situation with a maritime port in the Mediterranean and communication with Europe makes it interesting on several levels. Given its population and level of commerce, it is an ideal place to open new businesses.

In this study we put ourselves in the shoes of a consultant who is asked for a study in which to select a list of neighbourhoods to set up an ecological supermarket.

Various parameters are analysed such as population, age groups, economic indicators, population by age groups and unemployment rate. These variables are analysed to predict how adequate is a neighbourhood to open a business. The objective its make machines learning model to predict the best zones and find the best neighbourhoods.

1.2 Problem

Barcelona is a huge city with a lot of business. We can use the data to determine what is the best neighbourhood to open a business?

Obviously, each business has its characteristics and you must take them into account to choose the most suitable neighbourhood. The study focuses on generic businesses and therefore neighbourhoods with little competition are sought.

1.3 Interest

The principal interest is knowing the parameters that involves ecologic business, and any person that want invest in this type of business can find high value information.

The study can be carried out for any type of business and the socioeconomic parameters of interest. It is not an accurate study, but it does provide an idea of which sites are ideal to start a more in-depth study. Anyone interested in starting a business can benefit from this study, and even use the notebook in a simple way to obtain personalized results. The foursquare data can be obtained for each type of business and any person can load the interest business data in the notebook to analyse other types of business.

2.- Data acquisition and cleaning

The data collects various points of interest such as population, age ranges, economic indicators, and businesses. In the first place, the neighbourhoods can be grouped by clusters and see which neighbourhoods have similar qualities to those that have green businesses but have not yet opened this type of business. The factors can be used to see which neighbourhood interests us the most, a high population with high incomes, a young population and the absence of similar businesses, giving more weight to a value or others, we can have which neighbourhoods have a higher score depending on the parameters that we interest. Also, multiple linear regression are applied to predict the number of target venues and analyse the importance of each parameter studied.

The data used for this study are:

1.-Barcelona Neighbourhoods data, such as the names and localization. (Webscrapting, geopy API)

2.-Barcelona Demographic data - Population by Neighbourhoods (Statistical Institute of Catalonia)

3.-Barcelona Unemployment data (Statistical Institute of Catalonia)

4.-Barcelona Economic data (Statistical Institute of Catalonia)

5.-Barcelona Business data (Foursquare API)

The data is extracted through various techniques, such as Webscrapting, REST API calling and official statistical reports.

For webscrapting and Foursquare the data is cleaned in the extraction process. For the data obtained from official sites, they are processed with excel to match the key fields and present the appropriate format. With Geocoder, a visual inspection of the locations is carried out using Folium and representing the coordinates obtained on the map, with the support of a colormap and a GEOJSON file, the task is carried out more easily.

2.1 Data sources

1.-Barcelona Neighbourhoods data are obtained applying webscrapting to Wikipedia page https://es.wikipedia.org/wiki/Distritos_de_Barcelona. The locations are obtained from geopy geocoder nominatim API.

2.-Barcelona Demographic data is obtained from the official website of Barcelona website.

2.1.-For population:

https://www.bcn.cat/estadistica/castella/dades/tpob/pad/ine/a2019/sexe/barri.htm (2019)
Source: institut d'Estadística de Catalunya.

2.2.-For quinquennial age neighbourhood population:

https://www.bcn.cat/estadistica/castella/dades/tpob/pad/padro/a2019/edat/edatq05.htm
Source: Ajuntament de Barcelona. Departament d'Estadística i Difusió de Dades. Lectura del Padrón Municipal de Habitantes a 1 enero 2019.

3.-Barcelona Unemployment data is obtained from official Barcelona website

https://www.bcn.cat/estadistica/castella/dades/barris/ttreball/atur/Evolucio/bcnbar.htm (2020-2021)
Source: Departament de Treball, Afers Socials i Famílies. Generalitat de Catalunya.

4.-Barcelona Economic data (Statistical Institute of Catalonia) is obtained from Barcelona official website

https://www.bcn.cat/estadistica/catala/dades/economia/renda/rdfamiliar/evo/rfbarris.htm
Source: Ajuntament de Barcelona. Oficina Municipal de Dades.

5.-Barcelona Business data is obtained using the Foursquare API.

2.2 Data cleaning

For webscrapting and Foursaquare the data is cleaned in the extraction process. For the data obtained from official sites, they are processed with excel to match the key fields and present the appropriate format. With Geocoder, a visual inspection of the locations is carried out using Folium and representing the coordinates obtained on the map, with the support of a colormap and a GEOJSON file, the task is carried out more easily.

2.2.1.-Neighbourhood data is extracted perform webscrapting method to Wikipedia page. After obtaining the data this is compare with the official data (Figure 1).

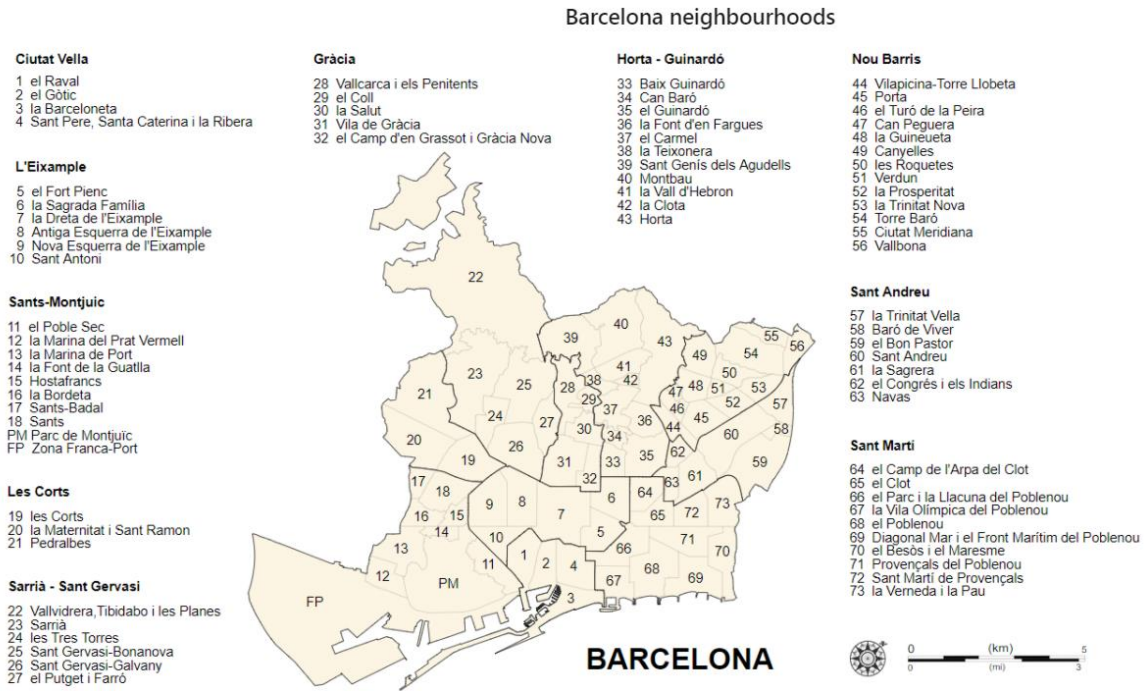

Figure 1 . Barcelona Neighbourhoods

In the process 75 neighbourhoods are obtained, two are drop from the dataframe (Figure 2).

NºBorough	Borough	NºNeighbourhood	Neighbourhood
3	Sants-Montjuïc	*	Zona Franca
3	Sants-Montjuïc	*	Montjuïc

Figure 2. Wrong neighbourhoods

After this step, the data of neighbourhoods are corroborated.

To obtain the coordinates the goepy API is used, but the results arent accurate. First detect the null values and the places wrong located. To easily visualization uses folium with GEOJSON file to give colour to neighbourhoods by the latitude (Figure 3).

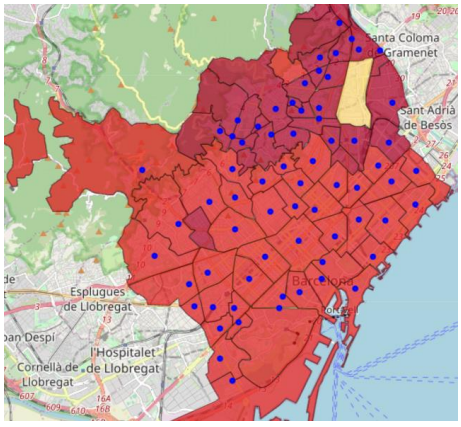

Figure 3. Wrong neighbourhood map

Once the Wrong locations are corrected the map presents a beautiful colorscale. Now the data of the neighbourhoods are perfect(Figure 4), and have a powerfull tool to visualize the results.

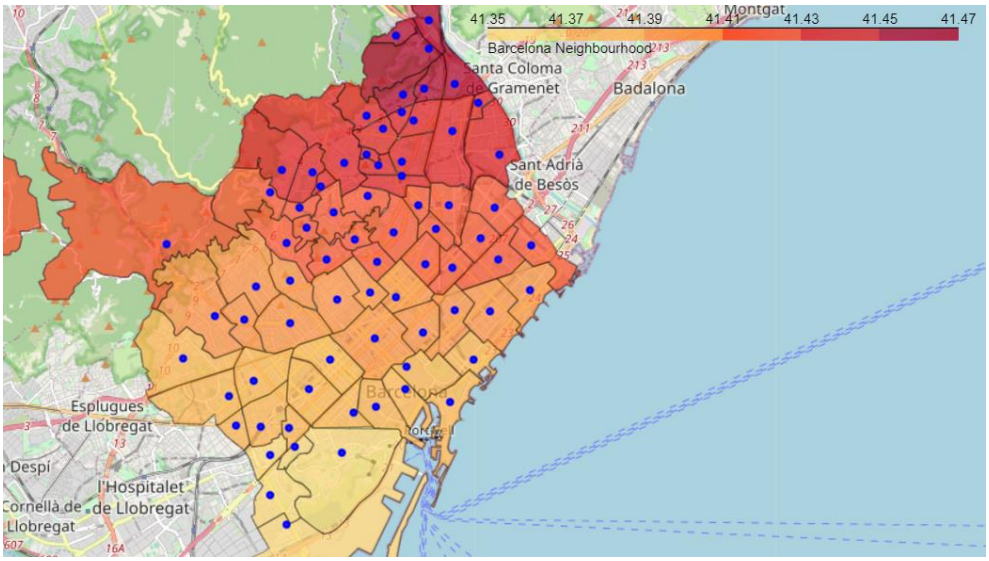

Figure 4. Correct neighbourhood map

2.2.2.- Barcelona Demographic, unemployment and economic data are manipulated with excel to fast formatting of the neighbourhood names.

2.2.3.- Barcelona Business data is obtained with foursquare; the problem is the duplicate values. To delete the duplicates a function is built to associate the venue to the near neighbourhood.

In the next code segment (Figure 5), the foursquare returns 3106 venues, 2753 of these are duplicated, after process the data with the function keep the 2753 correct places.

```
[87]: bcn_bussines[["Venue Latitude","Venue Longitude"]].duplicated().value_counts()

[87]: False    2753
      True      353
      dtype: int64

      365 venues are duplicated.

[88]: bcn_bussines_clean = delete_duplicates(bcn_bussines, bcn_coor) # Remove the duplicate venues
      print("Duplicated data removed")

      Duplicated data removed

[89]: bcn_bussines_clean.shape # show the data after remove the extra rows.

[89]: (2753, 5)
```

Figure 5. Delete duplicates code

2.3 Variable selection

All the variables are analysed to understand their importance. For the machine learning the total population, and population in age of work are drop, because the population by age group gives the same value if they add up.

For the Machine learning methods, the selected variables are:

1.-Economic indicator: RFD

2.- Population by age groups. (in groups of 15 years)

3.- Unemployment rate

4.- Number of target venues.

3.- Exploratory Data Analysis

3.1 Calculation of target variable

Analyse the superficial linear correlation between variables with a heatmap (Figure 6). We can see a strong relationship between the population and the age groups, this identifies that there is a similar percentage of the population for each age group. A strong relationship is also seen for the unemployment rate and the RFD economic indicator.

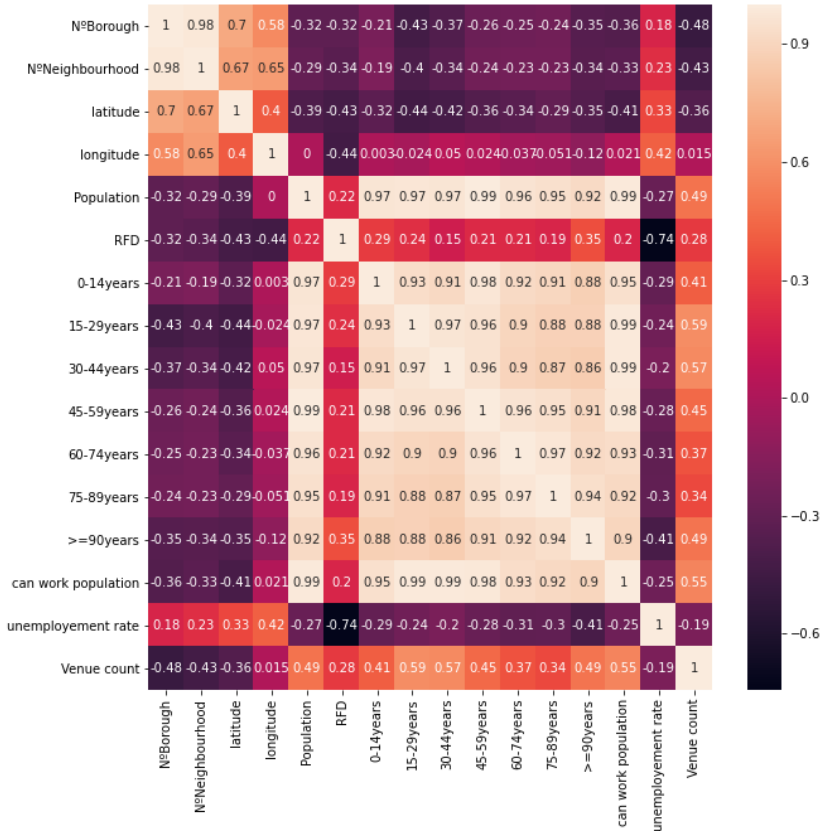

Figure 6. Heatmap for variable correlation

With this correlation can be created a first idea of the weight for each variable. The RFD indicator have a relationship, but the unemployment rate has a negative correlation this means the higher the unemployment, the fewer the businesses, this is totally logical. For other way the age groups that have higher relationship are the age group of 15 to 29 years old. The younger group that can work

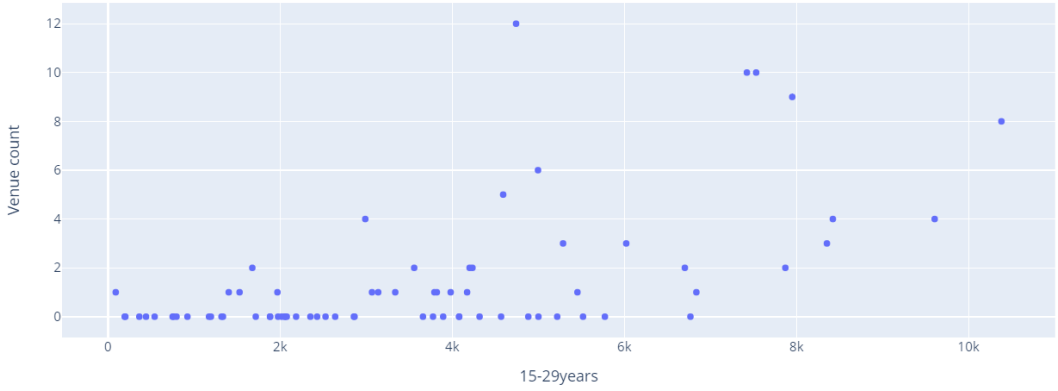

Figure 7. Scatter plot for 15-29 years age group vs venue count

The distribution is very dispersed but there is a certain relationship between the number of green businesses and the number of people between 15 and 29 years old (Figure 7).

3.2 Population Data

With a population of 1,6 million it's very interesting see how this people it's distributed, both by area and by age.

The population by neighbourhood is a magnificent indicator to determine the level of economic activity of the businesses in an area (Figure 8). The greater the population, the greater the number of nearby consumers. We can see in the colormap how the population it's concentrated in the downtown and the San Andrés neighbourhood in the north of the city.

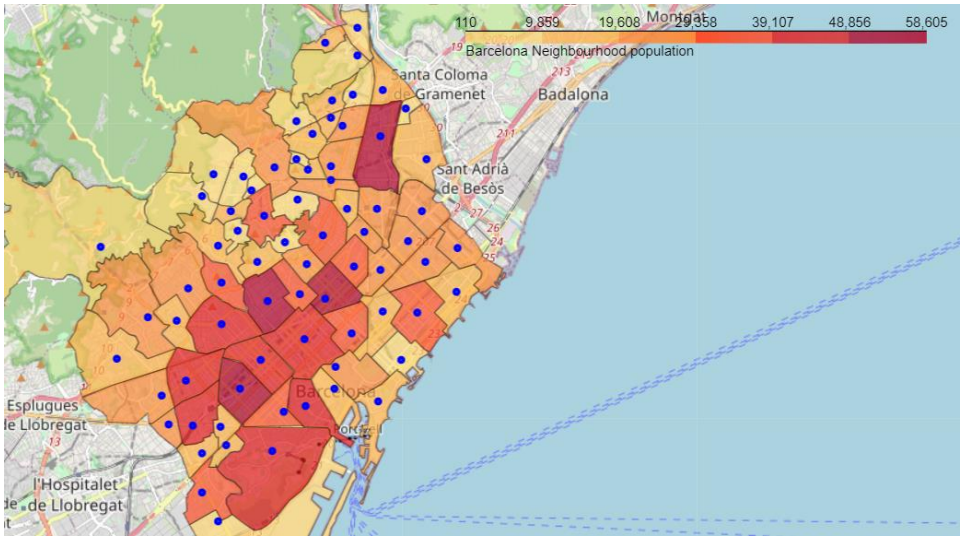

Figure 8. Barcelona population map

Representing the population with a sunsburst chart (Figure 9), the most populated neighbourhoods are easily visualized.

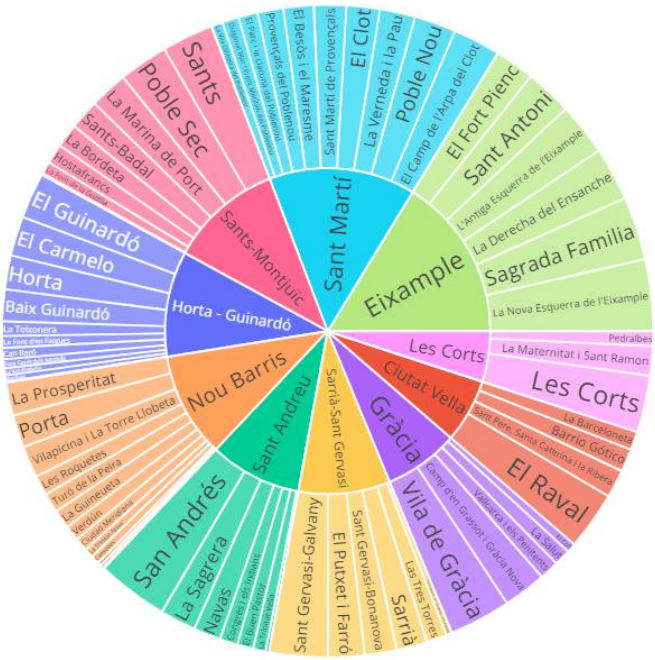

Figure 9. Sunsburst chart for Borough and neighbourhood population.

The predominant age group is 30 to 44 years old (Figure 10). Demographically an uncertain future can be seen. The active population between the ages of 15 and 60 is the majority and maintains the social services and benefits of the rest of the population. When this population ages and retires, given that the Young population is not enough, the retirement payments of the elderly will not be able to be maintained. This phenomenon explains the insistence of some politicians on private retirement plans.

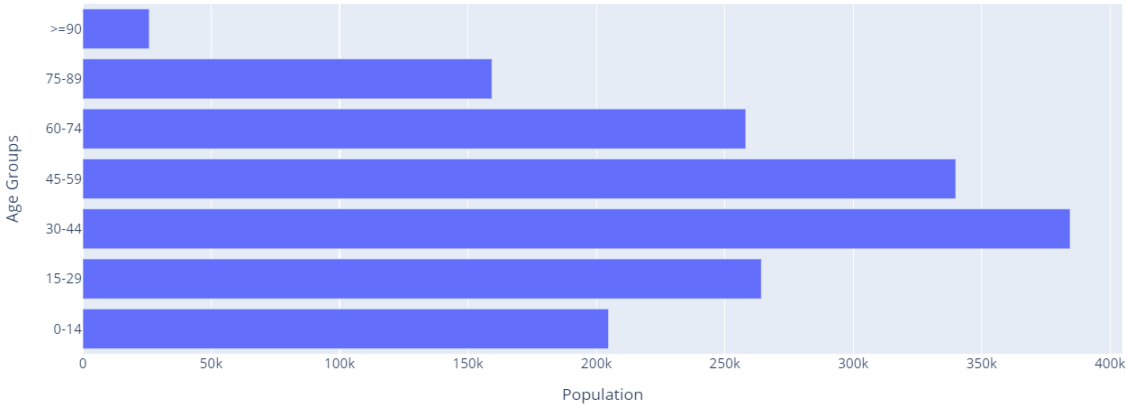

Figure 10. Bar chart population for age groups

It is appreciated that there is almost no elderly population in the neighbourhoods (Figure 11). The largest groups are those between 30 and 60 years old.

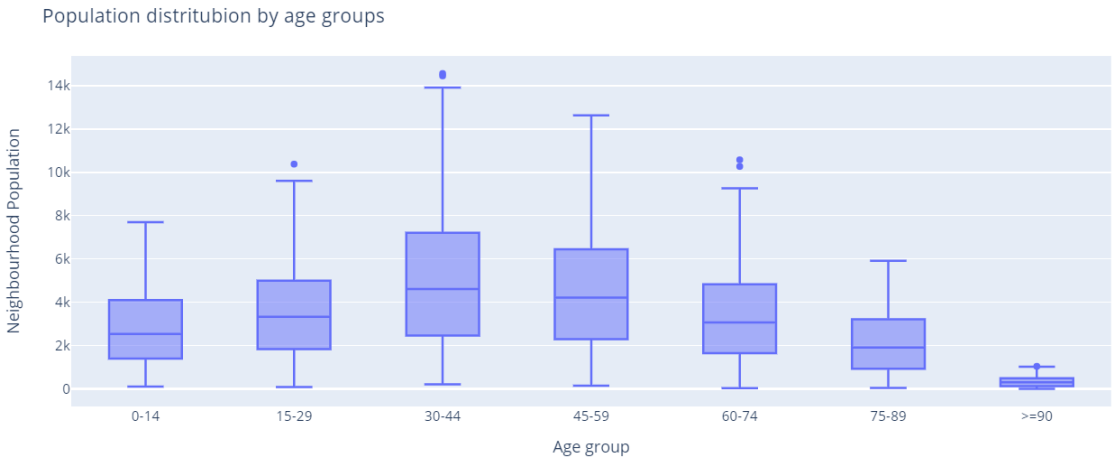

Figure 11. Box plot neighbourhood population for age groups

3.3 Unemployment data

There is an evolution of the unemployment rate (Figure 12), mainly this trend begins with the global pandemic situation. You can see some neighbourhoods that have extreme unemployment rates above 15%. In general, the average unemployment rate is 9%.

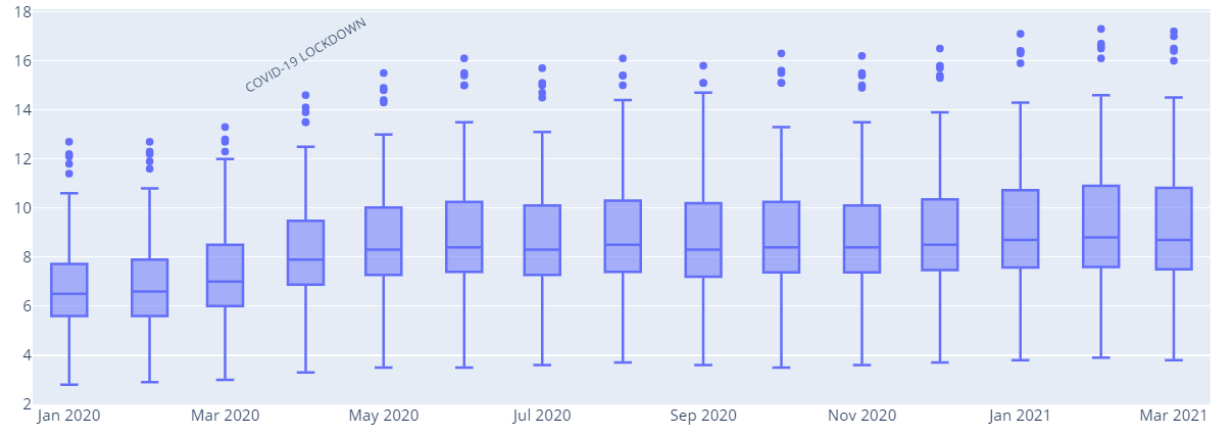

Figure 12. Box plot population neighbourhood unemployment rate

The Neighbourhoods into the boroughs of Les Corts and Sarrià-Sant Gervasi have the lowest unemployment**,** with rates between 4 and 4.5 points.

3.4 Economic data

Visualize the information on expenditure and income level of the population of Barcelona**,** this information tells us which spending groups are the majority and in which it is a good idea to invest. And which neighbourhoods have the greatest economic capacity

3.4.1 Expenditure groups

Bar chart shows the expenditure made per person in different areas of consumption.

This information indicates the spending groups of the population of Barcelona and which could be a business opportunity.

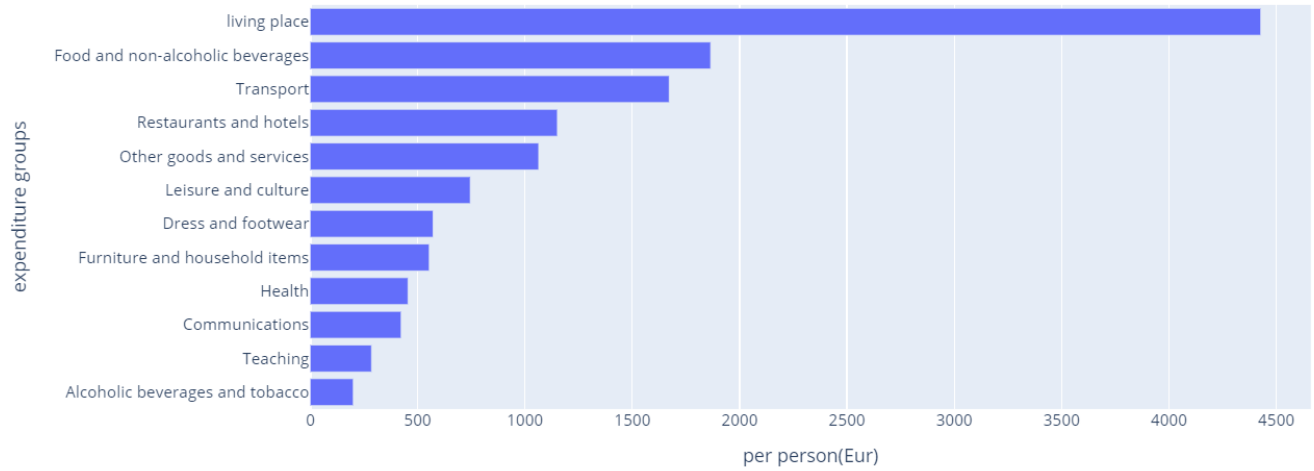

Figure 13. Bar chart of expenditure groups per person

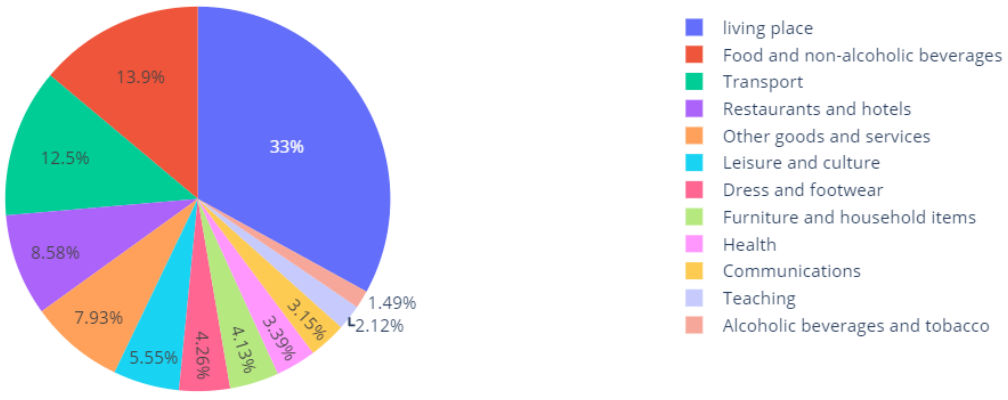

Figure 14. Pie chart of expenditure groups

Given that 33% of spending is on housing (Figure 14). The second group of expenditure is determined to be on food with the 13.9% of the expenditure, this group it's an interesting sector to analyse to open a business.

3.3.2 RFD indicator.

The RFD indicator (Renta Familiar Disponible per cápita/ Per capita Available Family Income) is the amount of income available to resident families for consumption and savings, once the amortizations or consumption of fixed capital in family economic farms and direct taxes have been deducted. This indicator is calculated about the mean of the total population of Barcelona, which is assumed to be 100.

With three map charts perform a easily visualization of the neighbourhood withs highest incomes. (Figure 15) The boroughs of Sarrià-Sant Gervasi have all neighbourhood with high incomes with a RFD indicator between 144 and 215 points. The neighbourhood of Pedralbes have the highest income with RFD indicator of 248.

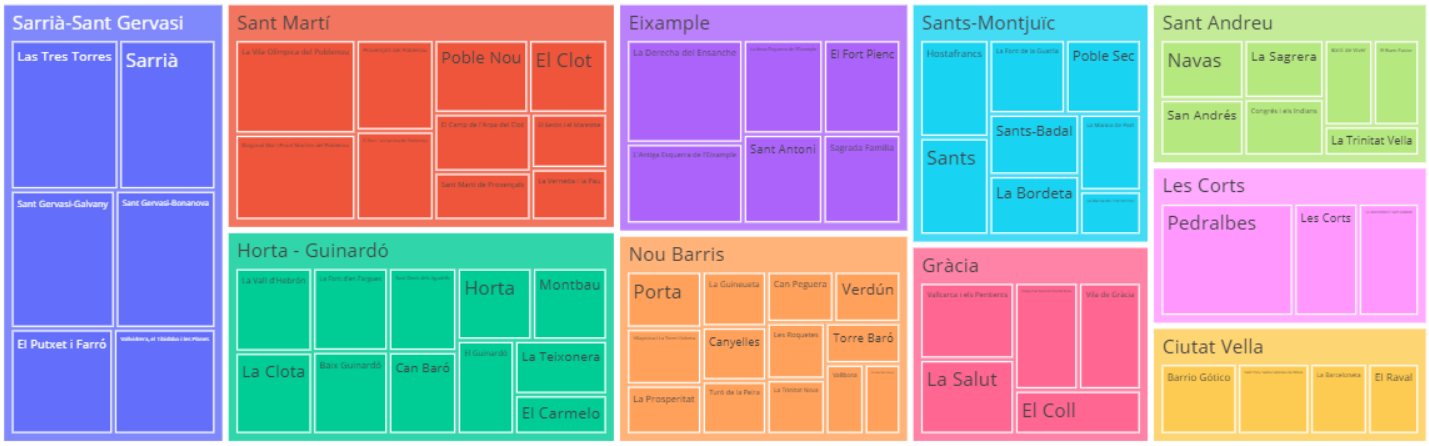

Figure 15. three map chart of Neighbourhood incomes

Visulizing the distribution with a box plot the RFD indicator (Figure 16) presents a sadly results, the third quartile falls into 100 points, namely only the 1 of 4 neighbourhoods are up to the average incomes.

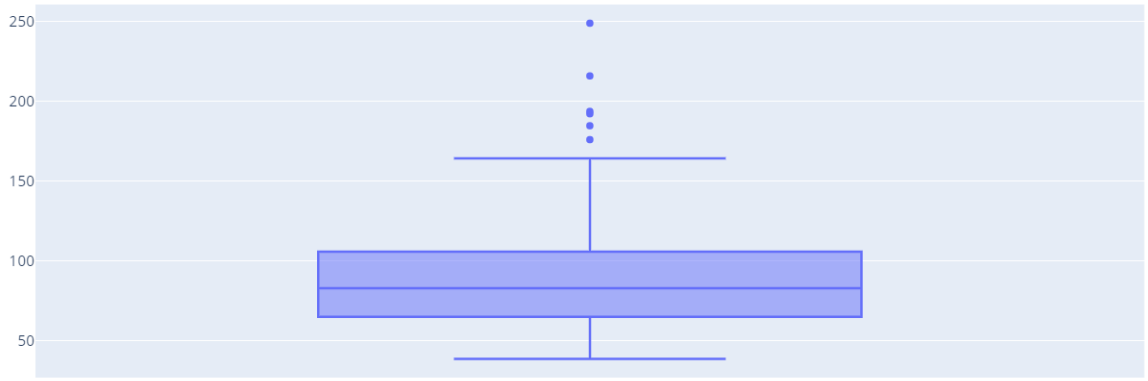

Figure 16. Box plot of neighbourhood income

Visualizing in the map the neighbourhoods with highest incomes are in the outskirts (Figure 17) , this is because the people living in chalets or good apartments.

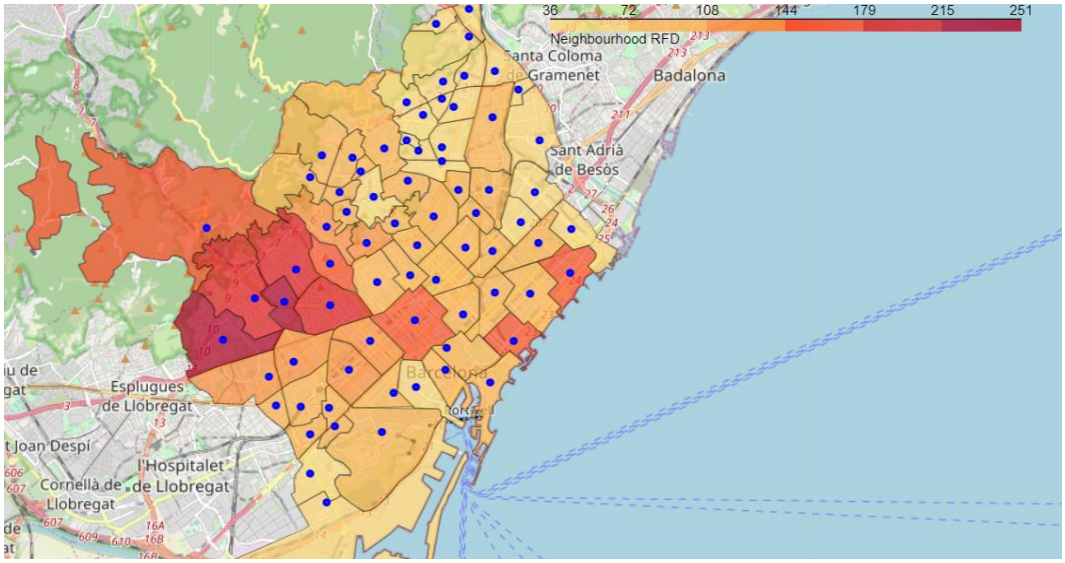

Figure 17. Map of neighbourhood income by neighbourhood

3.4 Relationship between unemployment and RFD indicator.

With these data and the previous point, it can be seen how the neighbourhoods with higher incomes have lower unemployment rates. It may be interesting to analyse the relationship between unemployment and income

We can see how the neighbourhoods with high unemployment rates those with the lowest RFD indicator are (Figure 18). The majority of neighbourhoods have an unemployment rate between 5 and 9 points, and have an RFD indicator between 50 and 125.

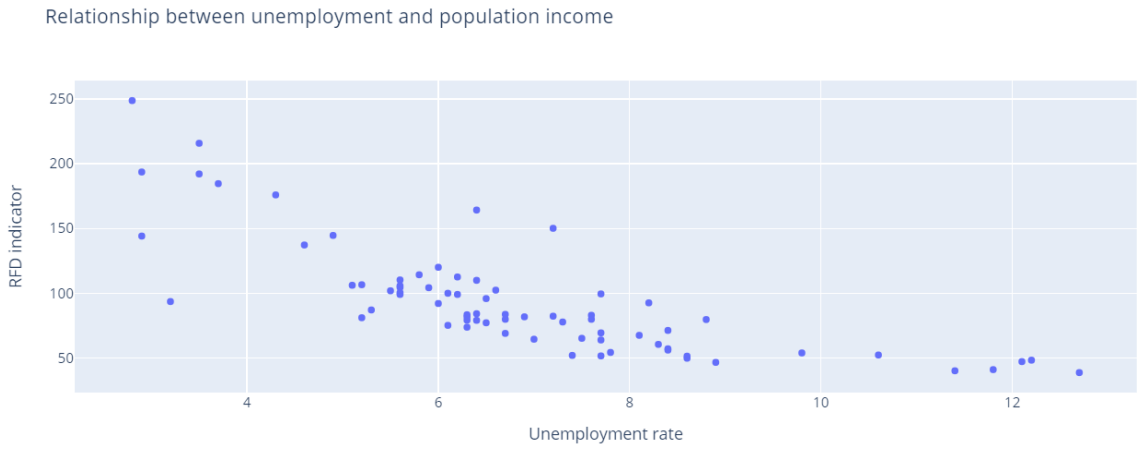

Figure 18. Scatter plot of unemployment rate and incomes by neighbourhoods

3.5 Business data.

3.5.1 Business in Barcelona

2735 venues are extracted with foursquare, representing the frecuency of differents types in a wordcloud(Figure 19) can visualize the most common business in Barcelona. Restaurants, hotels, cafes and bar are the most common business(Table 1).

Figure 19. Wordcloud for Barcelona business

Table 1. Most common venues in Barcelona

Most common venues in Barcelona	
Business category	**Number of business**
Spanish Restaurant	138
Mediterranean Restaurant	128
Tapas Restaurant	121
Hotel	108
Bakery	102
Café	102
Restaurant	102
Plaza	80
Bar	79
Pizza Place	74
Grocery Store	67
Supermarket	65

3.5.2 Ecologic business in Barcelona

106 Business are obtained searching by Organic, Ecologic, and Ecologic Market. Extract data from green busines because the data of ecologic market isn't enough to analyse with statistical methods. the neighbourhood trend towards the ecological is what we are interested in analysing.