Analysis of neighbourhoods in Barcelona to open an ecological supermarket

Guillermo Fantoni

May 23, 2021

Analysis of neighbourhoods in Barcelona to open an ecological market.

Abstract

This report is part of the final project of the IBM data science course. The only purposes of this report are to demonstrate the skills acquired during the course. The main requirement is to use Foursquare to obtain the locations of interest, this is a problem because using the free version does not obtain all the locations and the results loses precision. It should be considered for the data analysed.

The objective of this report is analysing the Neighbourhoods data to determine **What are the best neighbourhoods to open an Ecological supermarket.**

In this report, machine learning techniques such as clustering with K-Means or multiple linear regression are applied. Techniques such as webscrapting and the use of APIs such as Foursquare and Geopy are used for data extraction. Libraries such as pandas, beautifoulsoup, folium, matplotlib, numpy, sklearn and plotly are used. It works with files, CSV, XLSX, JSON, GEJSON, HTML.

23-05-2021

Table of Contents

Table of figures

Index table

Analysis of neighbourhoods in Barcelona to open an ecological market.

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.

Analysis of neighbourhoods in Barcelona to open an ecological market.

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 is 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)

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.

Analysis of neighbourhoods in Barcelona to open an ecological market.

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).

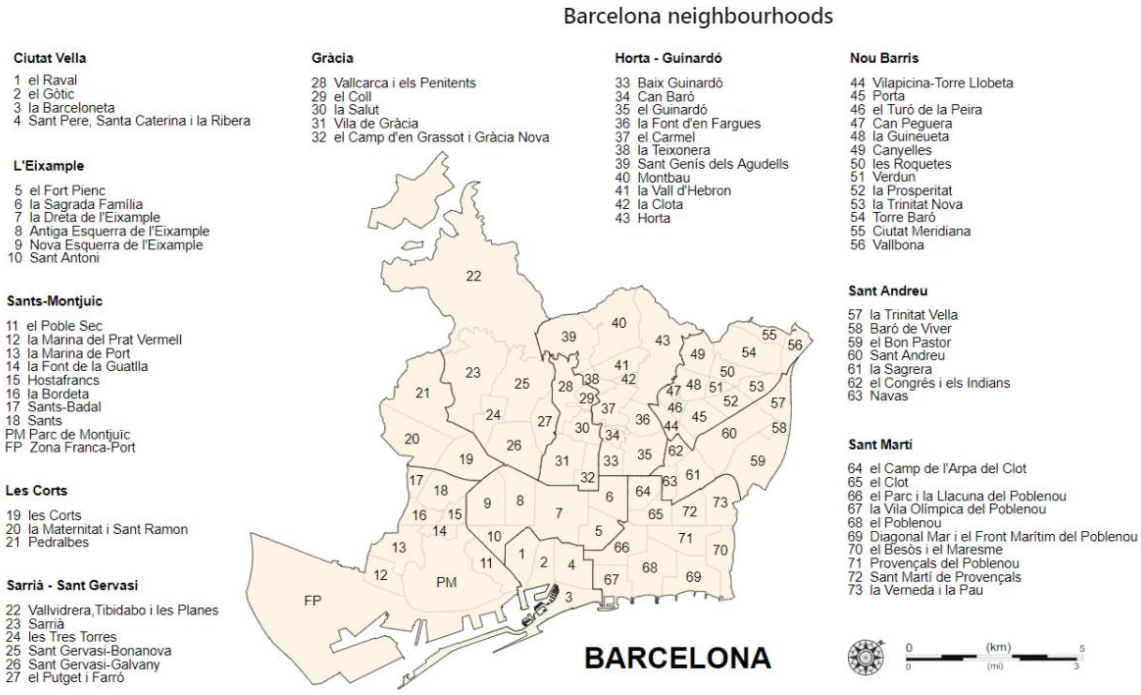

Figure 1 . Barcelona Neighbourhoods

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.

Analysis of neighbourhoods in Barcelona to open an ecological market.

To obtain the coordinates the geopy API is used, but the results aren't 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).

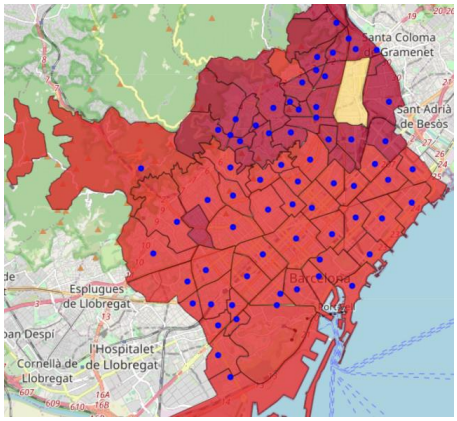

Figure 3. Wrong neighbourhood map

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

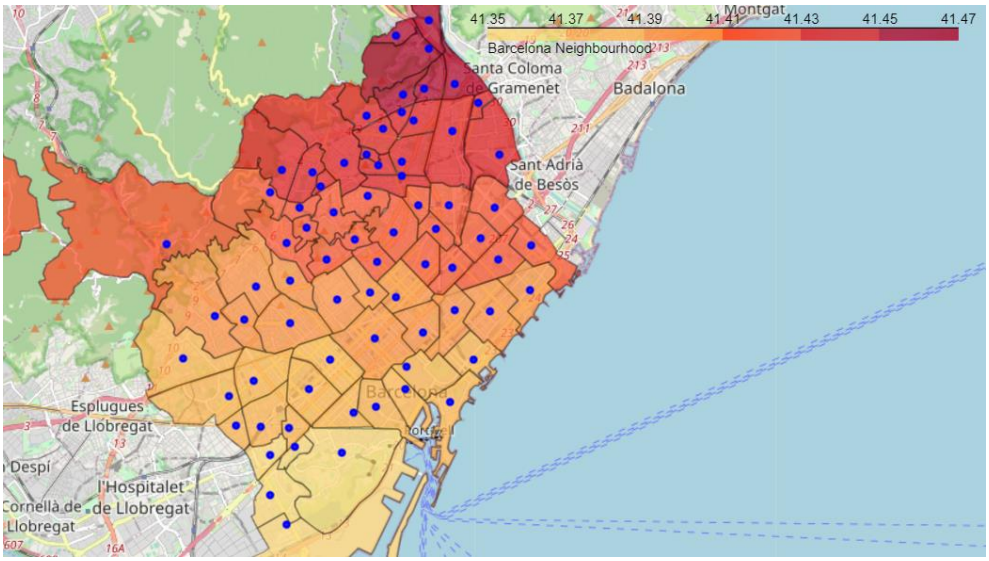

Figure 4. Correct neighbourhood map

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.

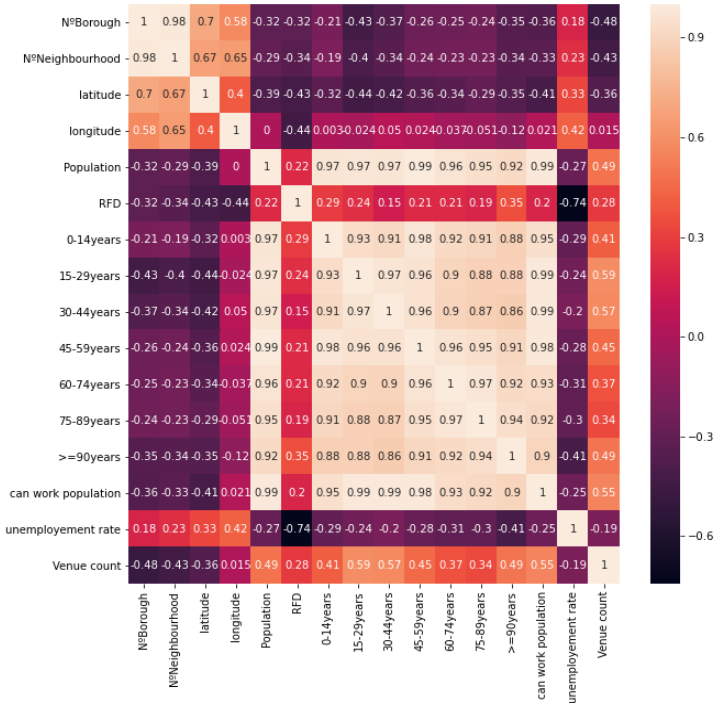

Figure 6. Heatmap for variable correlation

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

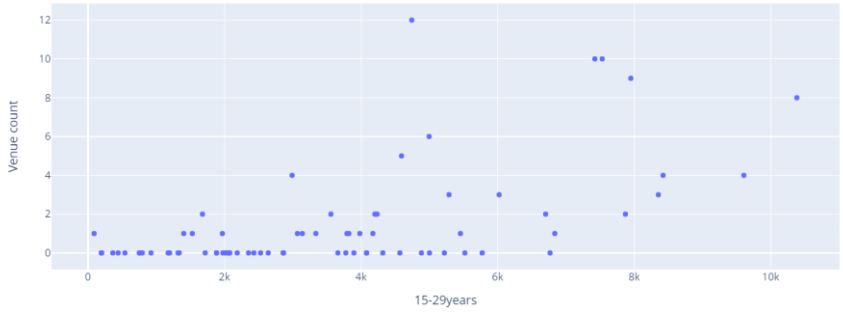

Figure 7. Scatter plot for 15-29 years age group vs venue count

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.

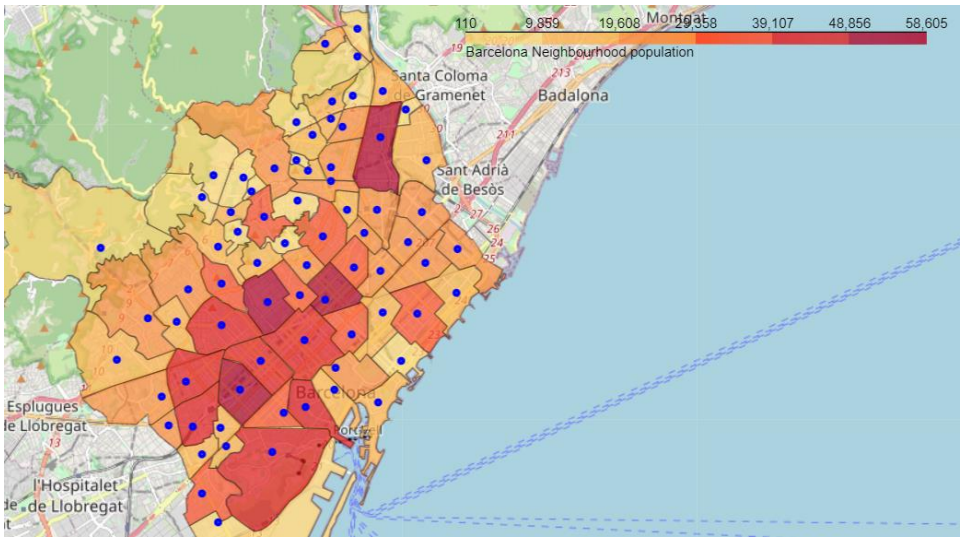

Figure 8. Barcelona population map

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.

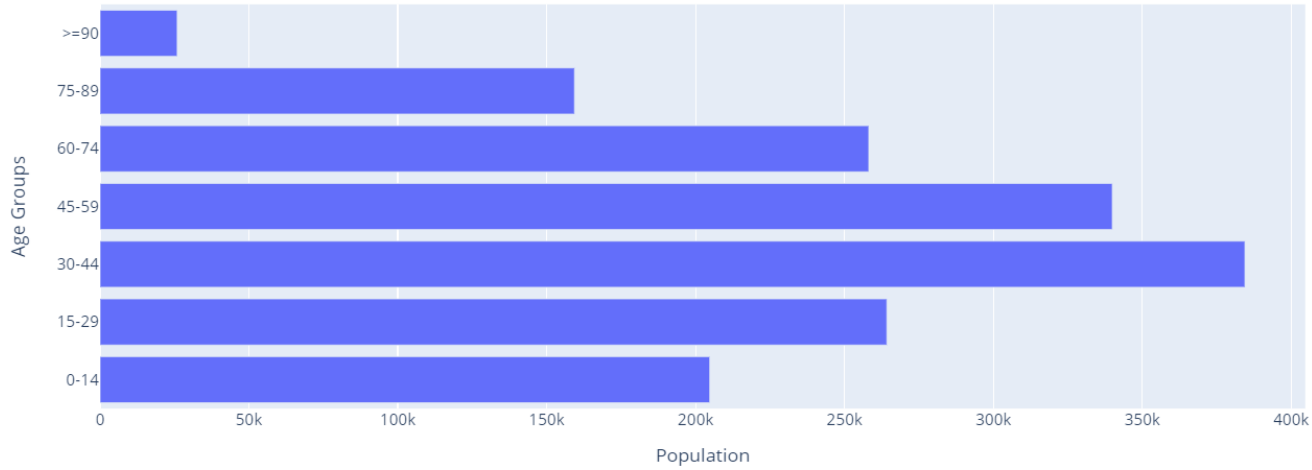

Figure 10. Bar chart population for age groups

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.

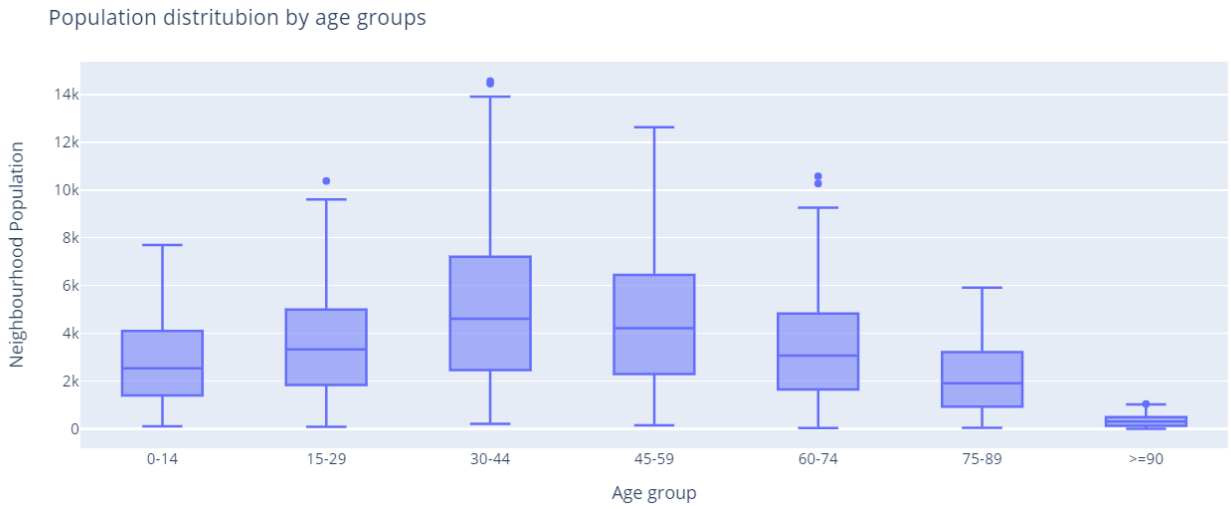

Figure 11. Box plot neighbourhood population for age groups

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%.

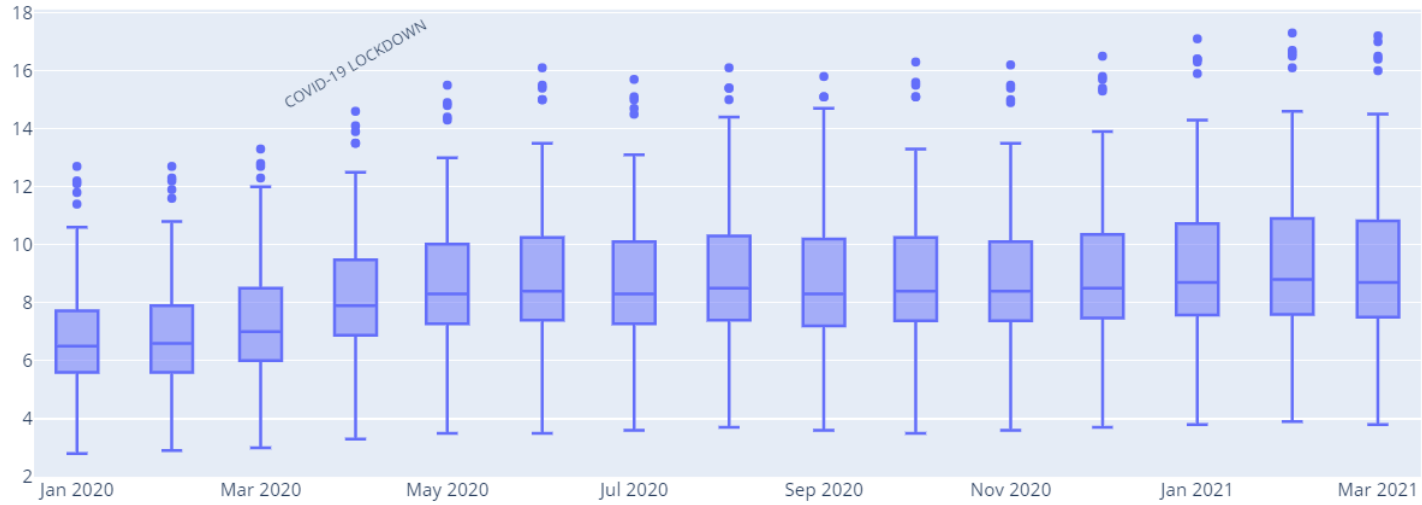

Figure 12. Box plot neighbourhood unemployment rate

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 (Figure 13).

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

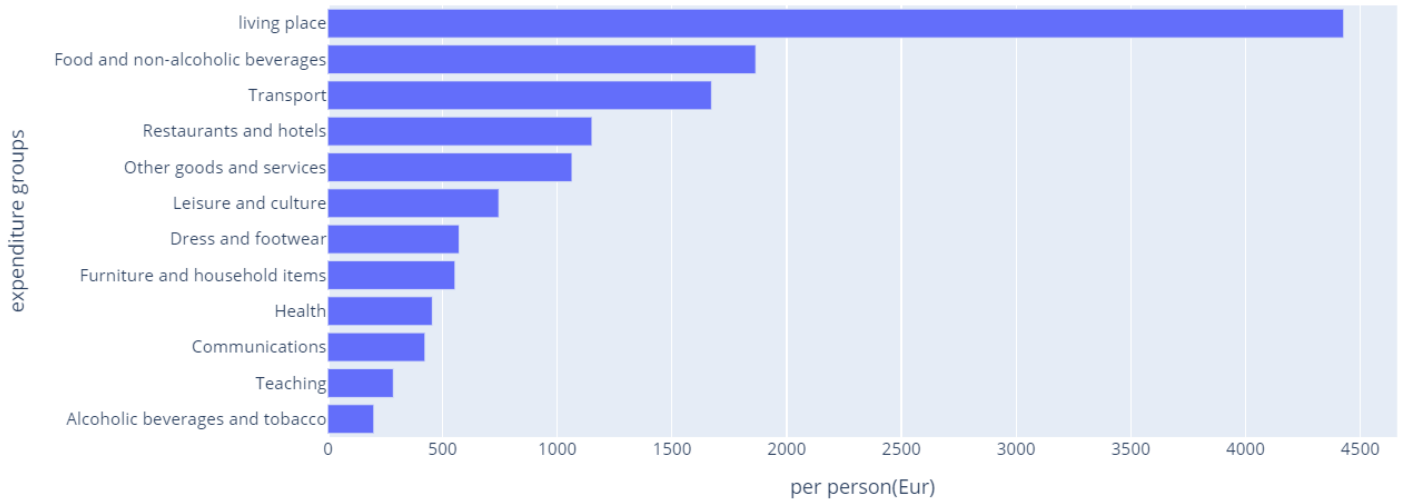

Figure 13. Bar chart of expenditure groups per person

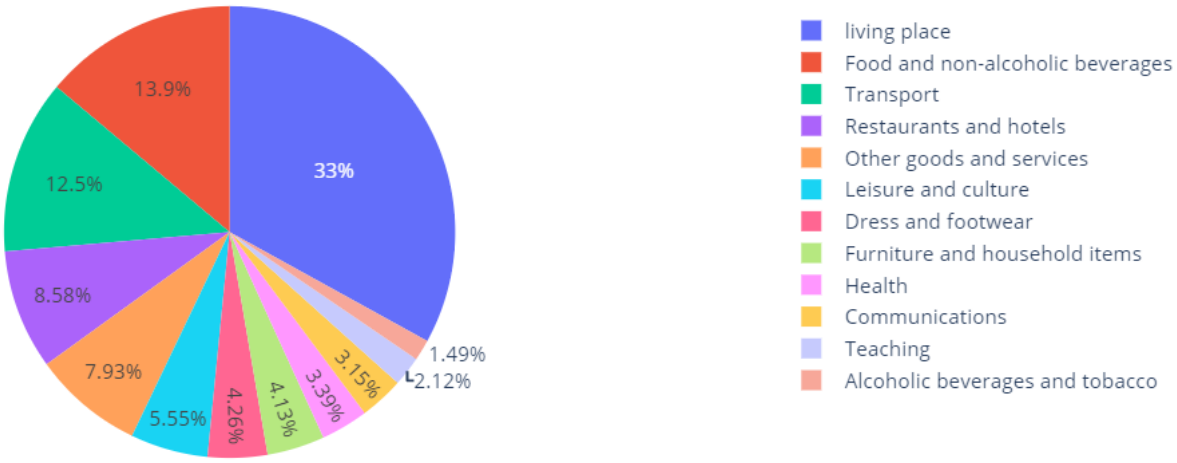

Figure 14. Pie chart of expenditure groups

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.

Analysis of neighbourhoods in Barcelona to open an ecological market.

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.

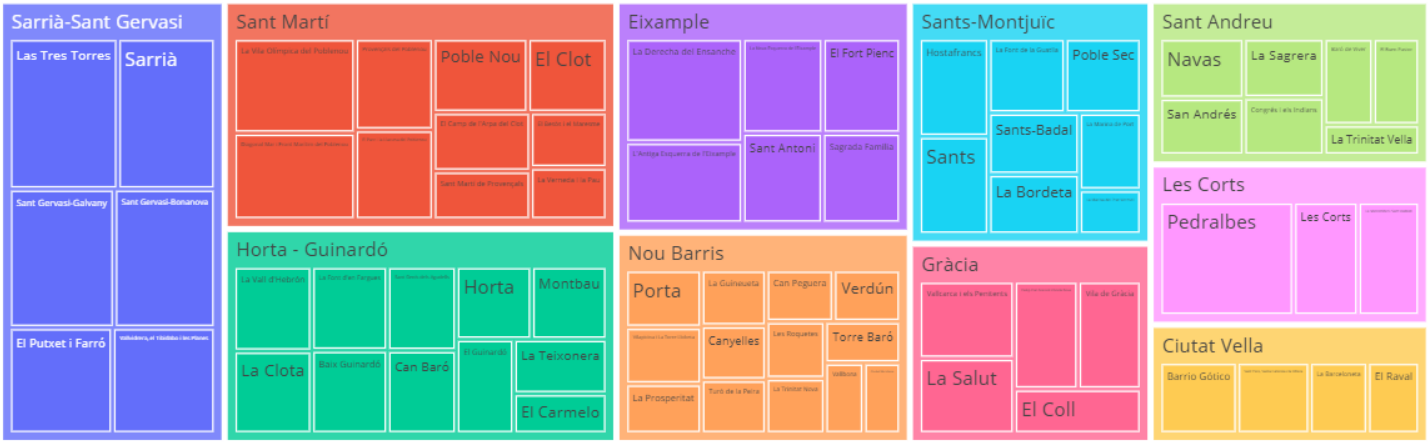

Figure 15. three map chart of Neighbourhood incomes

Visualizing 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.

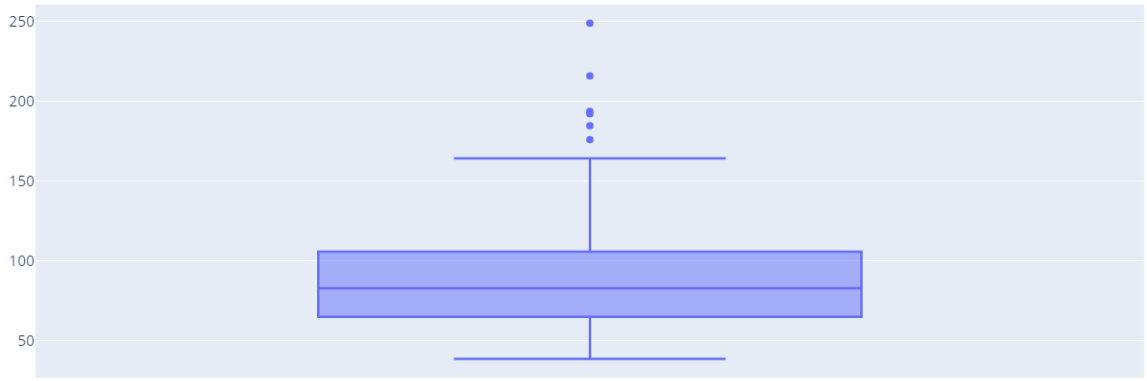

Figure 16. Box plot of neighbourhood income

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.

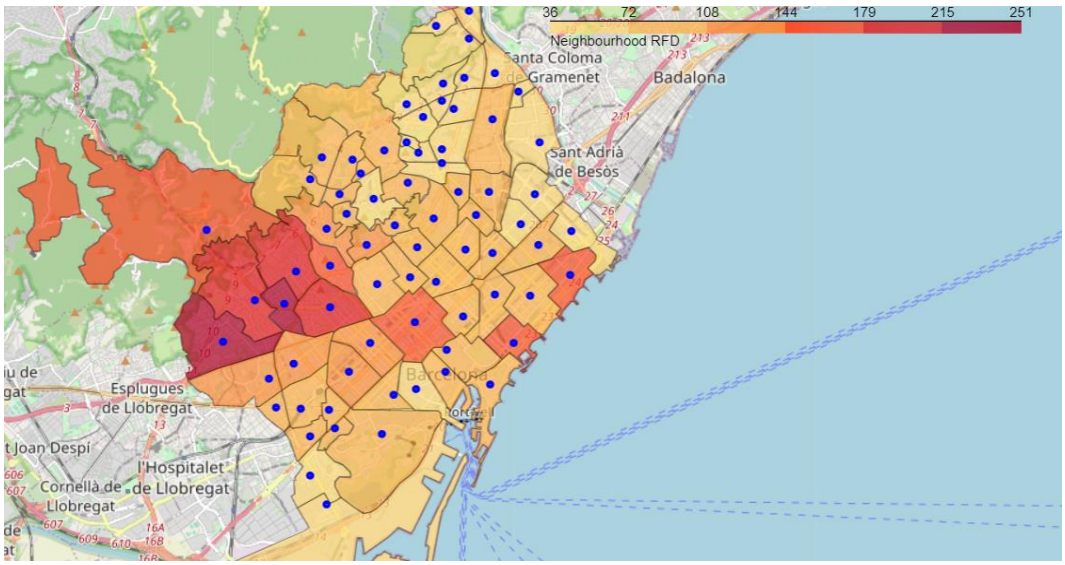

Figure 17. Map of neighbourhood income by neighbourhood

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). Most neighbourhoods have an unemployment rate between 5 and 9 points and have an RFD indicator between 50 and 125.

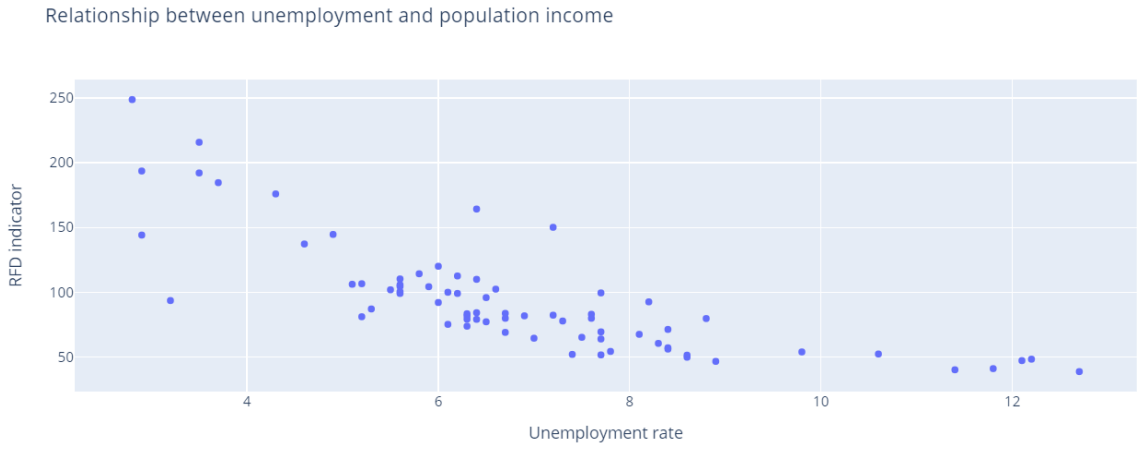

Figure 18. Scatter plot of unemployment rate and incomes by neighbourhoods

3.5 Business data.

3.5.1 Business in Barcelona

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

Figure 19. Barcelona business wordcloud

Table 1. Barcelona most common venues

Business category	N° 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 business 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.

4. Methodology

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 with K-Means method and see which neighbourhoods have similar qualities to those that have green businesses but have not yet opened this type of business.

The weighted factors analysis 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, highlight which neighbourhoods have a higher score depending on the parameters that we interest.

The data are analysed with multiple linear regression techniques to determine how the factors studied affect the number of organic supermarkets. Although in the data obtained from foursquare there are categories in addition to supermarkets, for example vegan restaurants, we are interested in seeing the trend of organic businesses, to have more data to analyse with these trends.

The data used in all machine learning methods are normalized with simple feature scaling method.

$$Xnew = \frac{Xold}{Xmax}$$

Finally, the results obtained from the different analyses are compared to determine the most suitable neighbourhoods.

5. Results

5.1 Cluster neighbourhoods by business

The neighbourhoods are grouped by the number and type of businesses in each one. First of all, analyse the most common business for each neighbourhood.

The neighbourhood's business are sorted by frequency obtaining the most common business for each neighbourhood (Figure 20). The full list can be found at the link https://github.com/gfantonipy/Coursera_Capstone/blob/main/CAPSTONE/dataset/most_common_venues.csv

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Baix Guinardó	Japanese Restaurant	Tapas Restaurant	Bar	Restaurant	Bakery	Café	Grocery Store	Gym	Supermarket	Breakfast Spot
Barrio Gótico	Ice Cream Shop	Tapas Restaurant	Plaza	Italian Restaurant	Bar	Hotel	Mediterranean Restaurant	Spanish Restaurant	Coffee Shop	Pizza Place
Baró de Viver	Asian Restaurant	Metro Station	Supermarket	Deli / Bodega	Furniture / Home Store	Dessert Shop	Pet Store	Restaurant	Falafel Restaurant	Farm
Camp d'en Grassot i Gràcia Nova	Tapas Restaurant	Bakery	Gym	Japanese Restaurant	Bar	Spanish Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Sushi Restaurant
Can Baró	Spanish Restaurant	Scenic Lookout	Tapas Restaurant	Chinese Restaurant	Grocery Store	Italian Restaurant	Soccer Stadium	Soccer Field	Breakfast Spot	Restaurant

Figure 20. Most common venues for neighbourhood

Once the venues data for the neighbourhoods were processed, we grouped the neighbourhoods by clusters with K-Means. With the Elbow method find the optimal number of clusters (Figure 21). Perform the K-means method with 3 clusters.

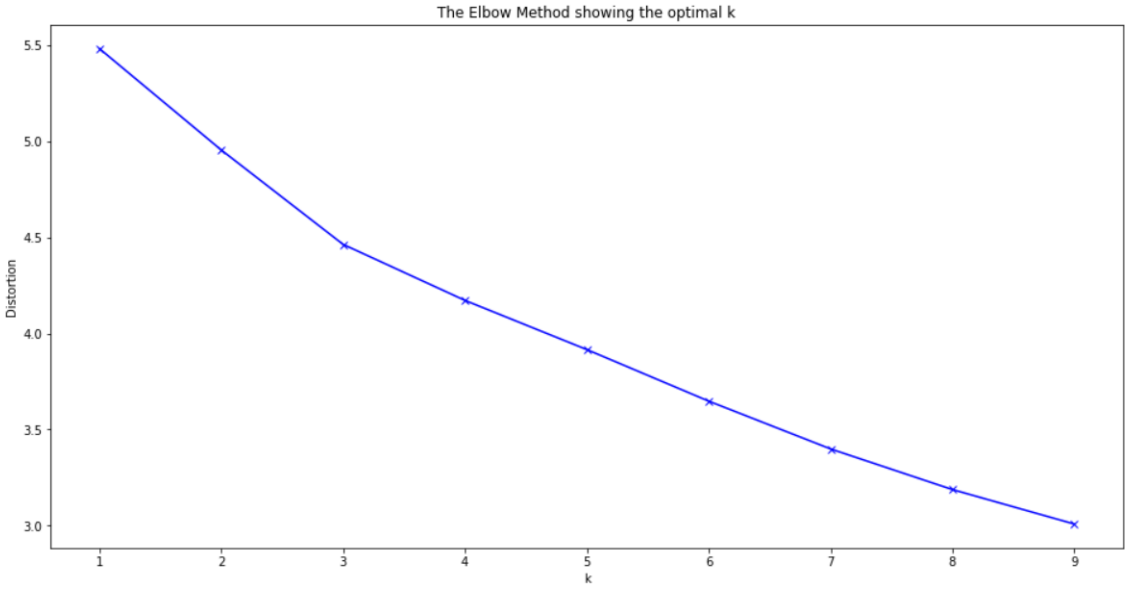

Figure 21. Elbow method

Once the neighbourhood are classified in to 3 clusters, represents it in a colour map (Figure 22). The colour indicates the number of ecological business. Each neighbourhood have a point with colour to indicate the number of cluster (0-Blue ,1-Green ,2-Black)

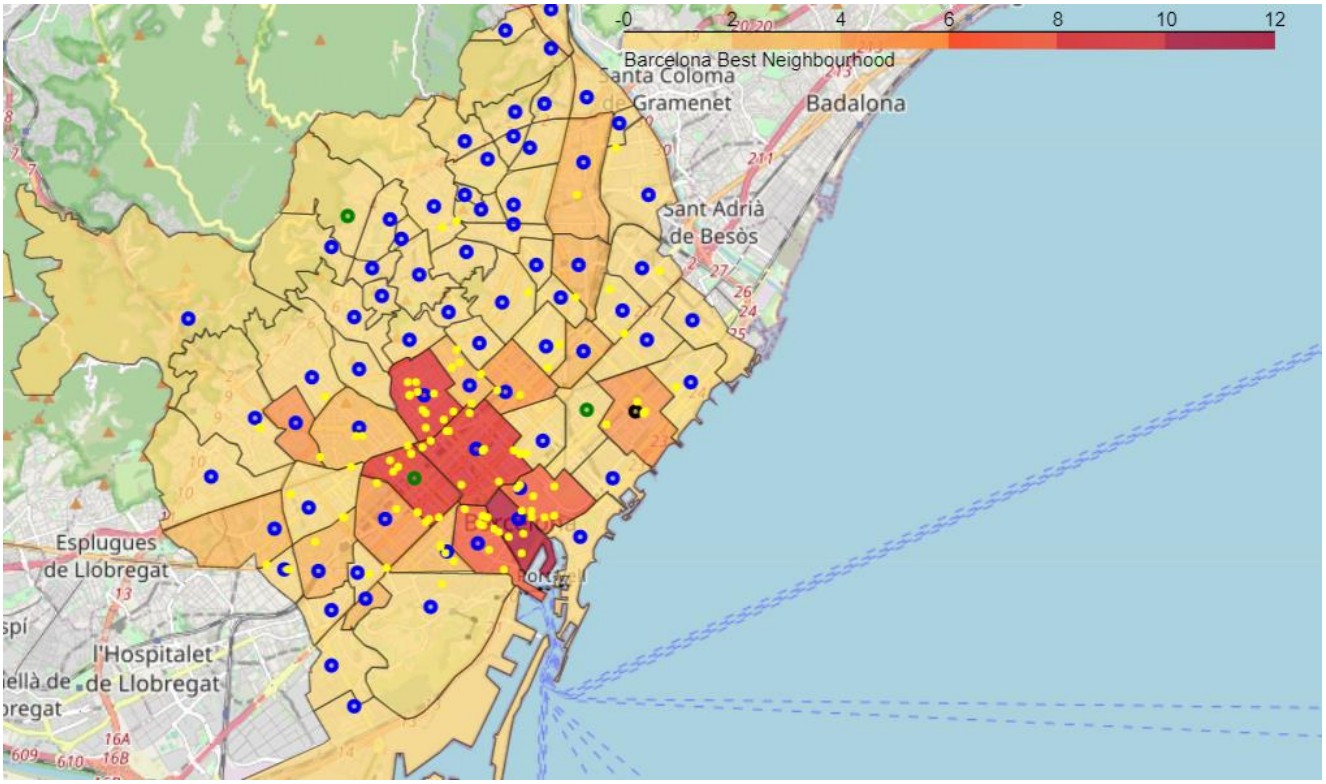

Figure 22. Cluster by business type

Looking at this map, **it is not possible to determine a clear group** of neighbourhoods by the type of businesses they own, in general in Spain, neighbourhoods tend to have the same type of businesses. Furthermore, the limitation of 100 results per neighbourhood reduces any precision.

5.2 Cluster neighbourhoods by socioeconomic indicators

Cluster the different neighbourhoods by the socioeconomic indicators to determine a relationship between the clusters and number of target venues. Perform the K-Means method with 5 clusters. Obtain the total count of ecological business and count of neighbourhoods for each cluster. The cluster 3 have many businesses with 8 neighbourhoods. The neighbourhoods for this cluster are the very interesting for the analysis.

Table 2. Cluster by socioeconomic indicators characteristics

Cluster	N.º Eco business	N.º Neighbourhoods	Business / Neighbourhoods
0	4	19	0.21
1	21	8	2.625
2	26	19	1.368
3	43	8	5.375
4	12	19	0.631

Analysing the cluster 3, we see that the neighbourhood of **"Les Corts"** has similar characteristics to "La derecha del ensache" but 10 times less businesses. So, it is a good candidate to start a business. The neighbourhood of **"San Andres"** too are a good candidate also don't have nearly business to compete.

Table 3. Cluster 3

Borough	Neighbourhood	Population	RFD	unemployement	Venue count
Eixample	La Derecha del Ensanche	43515	175.9	6.3	10
Eixample	L'Antiga Esquerra de l'Eixample	42393	137.2	6.6	10
Gràcia	Vila de Gràcia	50102	104.4	7.5	9
Eixample	Sagrada Familia	51385	101.8	7.5	4
Eixample	La Nova Esquerra de l'Eixample	58032	110.2	7.4	4
Sarrià-Sant Gervasi	Sant Gervasi-Galvany	47588	192.1	4.7	3
Sant Andreu	San Andrés	57843	77.7	9.1	2
Les Corts	Les Corts	46274	120.0	7.8	1

The clusters are represented in a map with folium (Figure 23). The colour represents the number of ecological businesses, and each cluster are represented by a colour (0-blue, 1-green, 2-orange, 3-red, 4-black). The yellow points represent the business.

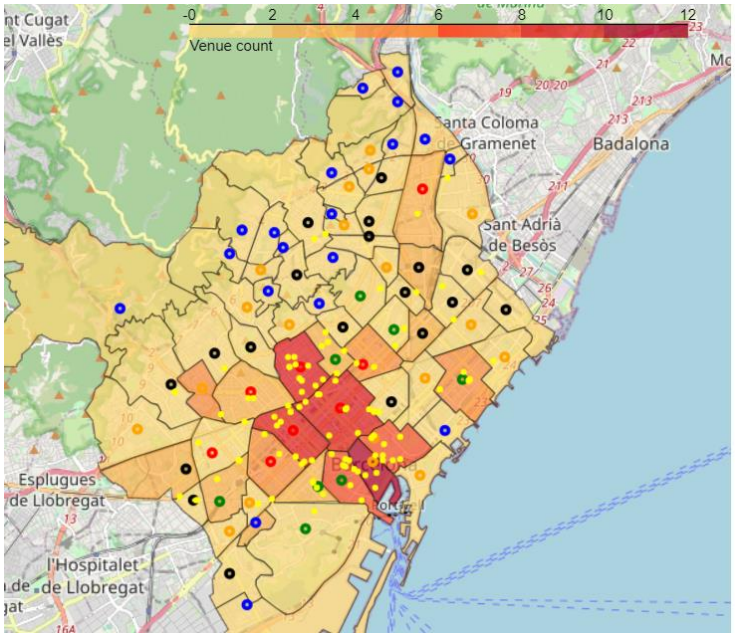

Figure 23. Clusters by socioeconomic indicators

We can see that the clusters with higher number of target venues (cluster 3-RED and 1-GREEN) are principally located in the center zones. Instead the rest of the neighbourhoods with a smaller number of target venues are located around the center of Barcelona. It is determined that the neighbourhoods in the central areas have more possibilities of having a green business. But it does not have to be related to socioeconomic indicators. Perhaps the central area of the city is the reason for the high number of businesses as is logical.

23-05-2021

5.3 Weighed factors analysis.

It is interesting to analyse the neighbourhoods by the indicators that interest us the most. A score is obtained for each neighbourhood by giving each variable a weight. These weights come from market research or intuition itself. For example, ecological trends have a greater impact on young people, so we will give more weight to the young population.

The data of the neighbourhoods are normalized with the simple feature scaling method to perform the analysis with weighed factors.

Table 4. Coefficients for weighed factor analysis

Parameter	Coefficient
Population	0
RFD	0.5
years_0_14	-1
years_15_29	2
years_30_44	2
years_45_59	0.4
years_60_74	0.3
years_75_89	0.1
years_90	0.8
can_work_population	0
unemployement_rate	-0.3
Business	-1

To obtain a score for each neighbourhood, we introduce a multiplication factor for each variable. We give more space to young groups and neighbourhoods with high incomes (Table 4). On the other hand, since we are not interested in unemployment or competition, we introduce a negative multiplier factor.

The first five neighbourhoods with highest score are the showing in the table (Table 5). The full list can be found at https://github.com/gfantonipy/Coursera_Capstone/blob/main/CAPSTONE/dataset/bcn_final_factors.csv

Table 5. Best neighbourhood in weighed factors analysis

Neighbourhood	Score
La Nova Esquerra de l'Eixample	4.27
Sagrada Familia	3.84
San Andrés	3.49
Vila de Gràcia	3.38
Sant Gervasi-Galvany	3.32

Analysis of neighbourhoods in Barcelona to open an ecological market.

We see with the colour scale that the neighbourhoods with the highest scores are those in the center of Barcelona (Figure 24). You can see some **neighbourhoods with high scores and few businesses**. These areas are interesting to open a business according to the priority we have given to the indicators.

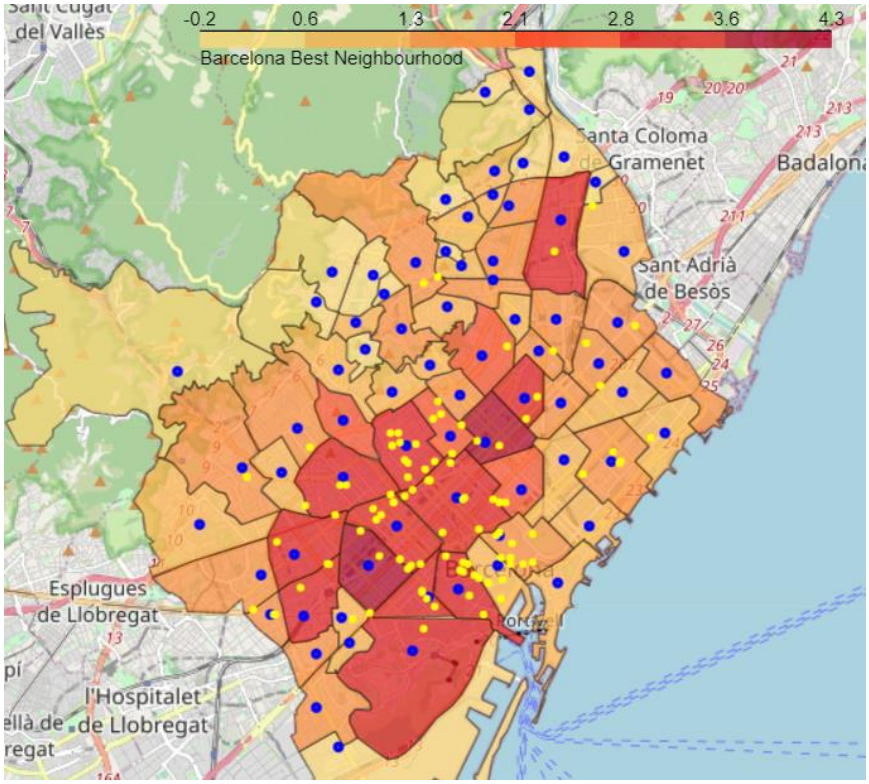

Figure 24. Barcelona map weighed factor analysis.

The neighbourhoods of **San Andrés, les corts and Poble Sec** are the zones with high interest according the selected factors and less numbers of target venues.

5.4 Multiple linear regression analysis.

A multiple linear regression is performed to obtain the correlation between the variables and the target value. In this case we want to obtain the relationship between the socioeconomic indicators and the number of businesses in each neighbourhood, to see if there is any direct relationship.

The multiple linear regression model returns the next coefficients (Table 6). The intercept of the regression model is -0.0592.

Table 6. Regression model coefficients

Parameter	Coefficient
75-89years	-1.142289
0-14years	-0.911339
60-74years	-0.127760
unemployement rate	0.031271
RFD	0.159044
30-44years	0.353543
45-59years	0.624880
>=90years	0.723144
15-29years	0.850739

Showing the parameters in a polar chart can see that the age group of 15 to 29 years and the elderly age group have the highest coefficients (Figure 25). The RFD indicator have a low influence and the unemployment rate have any importance in the model.

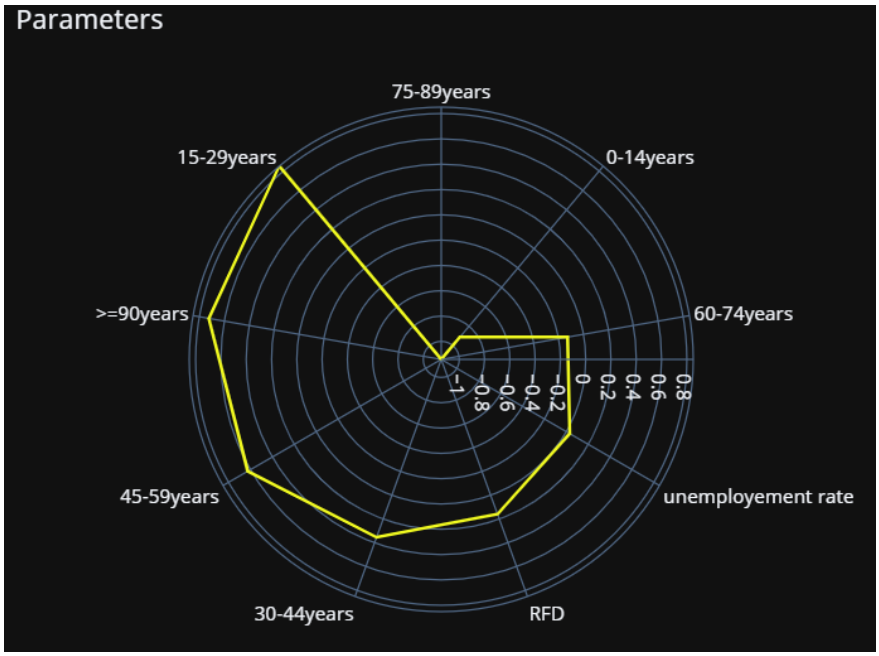

Figure 25. Polar chart of the regression model coefficients

Representing the same values in a bar chart can see the influence of the varablies in the regression model (Figure 26). The age groups of 75-78 years and 0-14 years have a negative influence in the model.

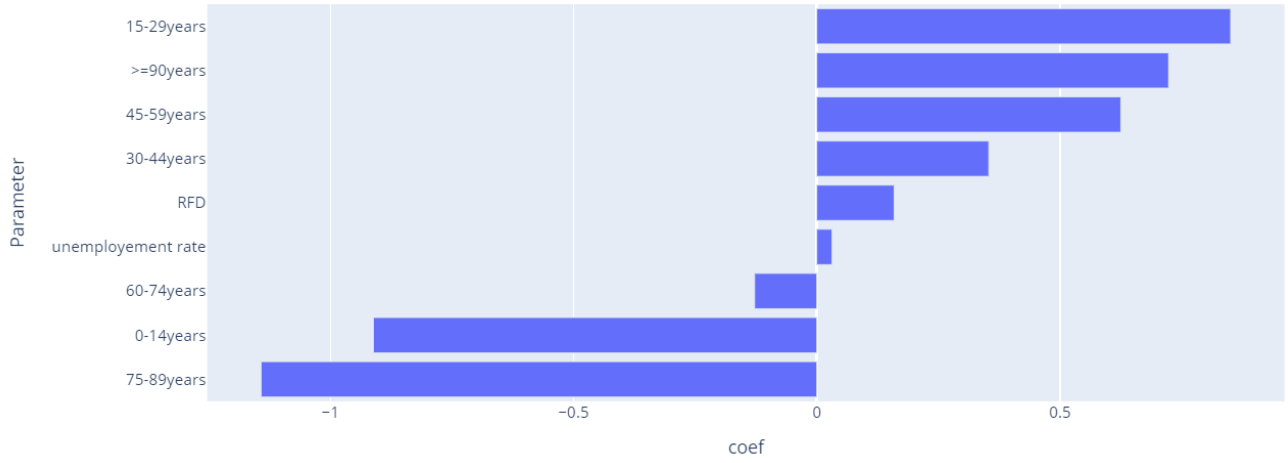

Figure 26. Bar chart of the regression model coefficients

Representing the prediction of number of business vs, the actual number of business for each neighbourhood, we can see a certain relationship (Figure 27). The model isn't very accurate but, can see a neighbourhood with high predictions and any business, this neighbourhoods are interesting for the analysis.

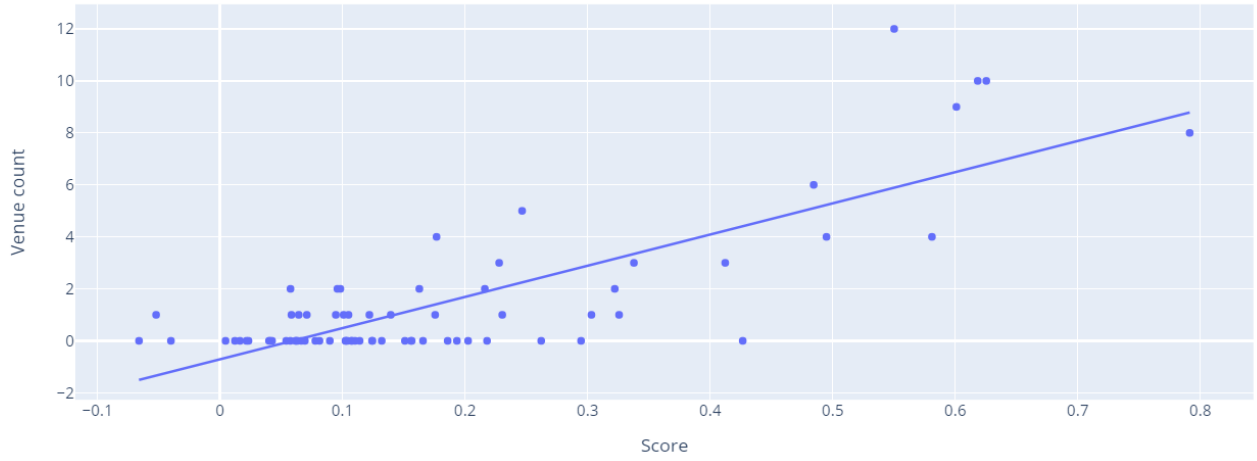

Figure 27. Scatter plot of prediction and actual number of business

Comparing the prediction of business and actual values, an R-Squared of 0.647 is obtained, a not very good value but taking into account that many factors influence the opening of a business, it provides us with a basis for the analysis.

Visualizing the value of the prediction with a colormap and the actual business with a yellow point, we can see that the city center have the most of business (Figure 28).

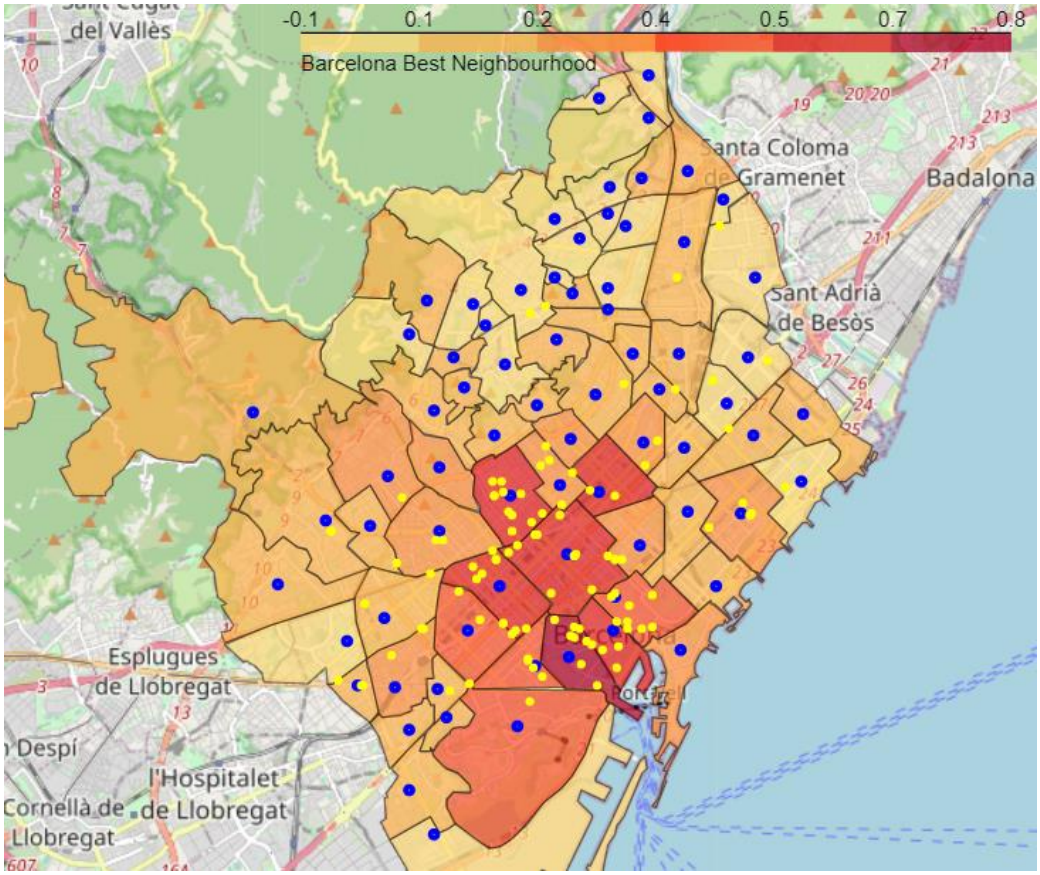

Figure 28. Map of Barcelona for regression model

We collect in a table the predicted values in the regression model and the real values. Comparing we can see that there are neighborhoods that according to the business should have green businesses. Repesenting the neighbourhoods with the highest diference we can see the neighbourhoods with highest probabilities to open a eco business (Table 7). The neighbourhood of **"Poble sec"** is the most interesting under this asumption.

Table 7. Actual values VS prediction values.

Neighbourhood	Venue count	Predict	Predict - real
Poble Sec	0.0	4.407	4.407
El Camp de l'Arpa del Clot	0.0	2.825	2.825
El Putxet i Farró	0.0	2.434	2.434
Sagrada Familia	4.0	6.260	2.260
El Fort Pienc	1.0	3.197	2.197

5.5 Cluster and regression methods by socioeconomic indicators.

The number of businesses and the value obtained from the regression by neighbourhood are obtained for each cluster and they are divided by the total number of neighbourhoods for each cluster. In this way, an average or a trend is obtained by grouping by clusters. Represents these values in a scatter plot (Figure 29), obtain a curious relationship, the trendline have a high correlation of 0.94 points. This can see that in average behaviour for this clusters are very marked. The cluster 1 are under the prediction line, namely have neighbourhoods that admit more eco business according to the prediction model.

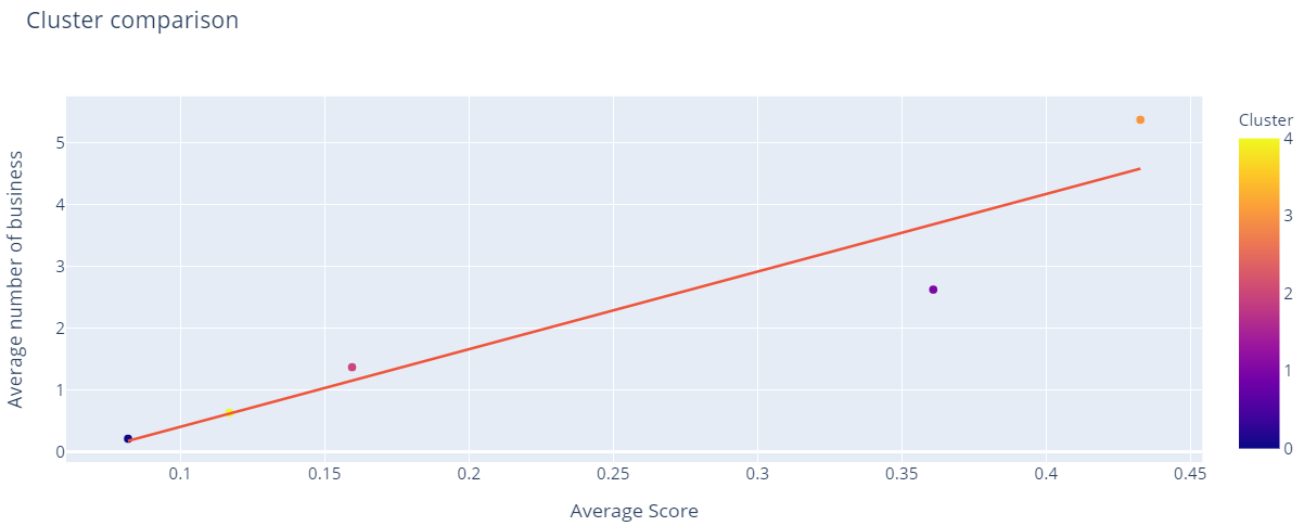

Figure 29. Scatter plot of cluster relationship for predict and actual values

The cluster 1 have the same number of neighbourhoods but the half of target business. It is interesting to analyse this cluster in case there are neighbourhoods with the characteristics sought. Looking at the neighbourhoods, it is determined that "**Poble sec"** is the neighbourhood with the least businesses and the best score (Table 8).

Table 8. Cluster 1 neighbourhoods

Neighbourhood	Population	RFD	unemployement rate	Venue count	Score
El Raval	47353	71.2	14.5	8	0.791
Poble Sec	39995	82.2	10.8	0	0.426
Sant Antoni	38236	104.2	8.2	3	0.412
Sants	42005	99.0	7.4	2	0.322
El Camp de l'Arpa del Clot	38663	81.7	9.0	0	0.294
Poble Nou	33861	99.9	7.8	5	0.246
Camp d'en Grassot i Gràcia Nova	34911	105.7	7.4	3	0.227
El Guinardó	37047	79.1	8.4	0	0.1657

In the map the are of neighbourhoods are coloured by the business prediction and the point of neighbourhood are coloured by number of cluster (0-blue, 1-green, 2-orange, 3-red, 4-black) (Figure30). The yellow points represent the business. We can visualize that the clusters 3 are in the city center and accumulates most of the business. The cluster 1 are the second cluster with most business. This two clusters are the principally objectives for this analysis.

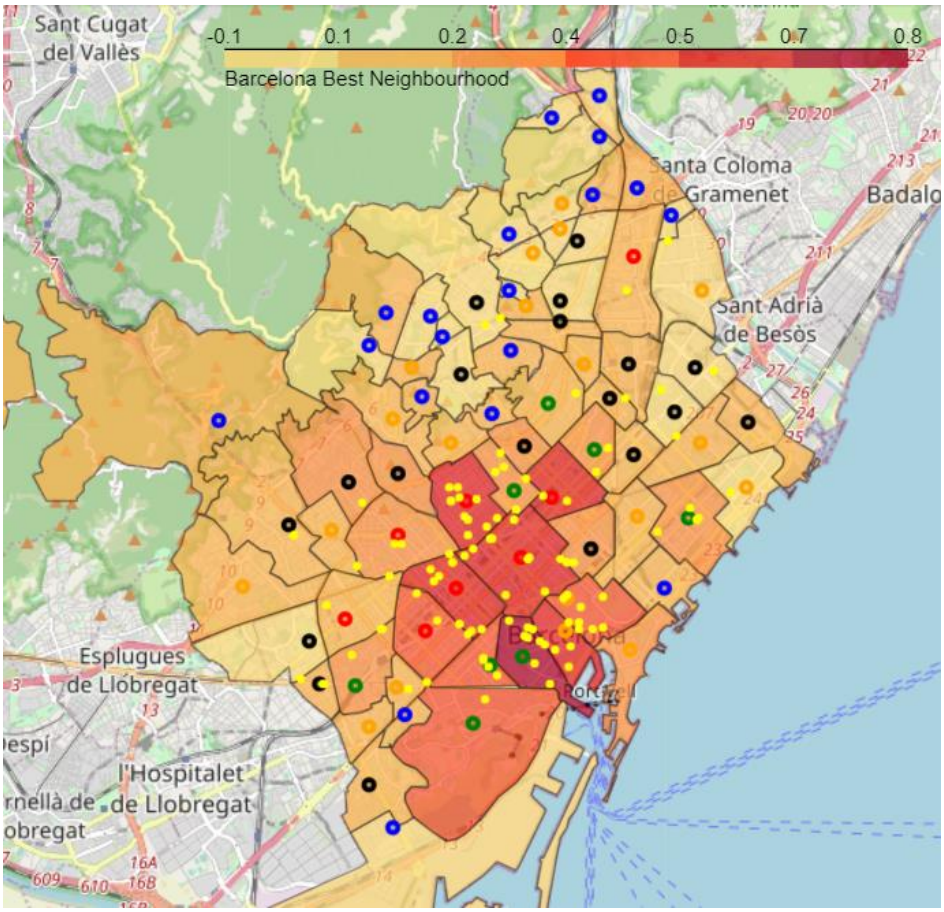

Figure 30. Barcelona map with cluster and regression models

6. Discuss

If we are interested in continuing with the basis that we do not want close competition, the neighbourhoods highlighted in the different analyses applied are the perfect candidates.

Custer by neighbourhood indicators analysis determine that by affinity the neighbourhood of "Les Corts" has characteristics very similar to "La derecha del ensache" but 10 times less businesses. So, it is a good candidate to start a business. The neighbourhood of "San Andres" too are a good candidate also don't have nearly business to compete. The neighbourhood Sant Gervasi-Galvany content in the same cluster have a high RFD indicator, this may be interesting if want a gourmet product.

With the multiple linear regression, we see that the Poble sec neighbourhood has 0 businesses but a prediction of 4. It is seen that this neighbourhood is close to the downtown area and has a high score in the analysis by weighted factors.

with the Weighted factors method, we see that the neighbourhoods of San Andrés, les corts and Poble Sec are the zones with high interest according the selected factors and less numbers of target venues.

7. Conclusion

Once the neighbourhoods with different techniques have been analysed. Several aspects can be concluded.

1.-The location data isn't very good, this is due to the free version of Foursquare. It **would be interesting to obtain very precise location data** and assess the distance to the city center and other factors.

2.- The **city center accumulates most of the businesses**, this is because if there is a good transport system in a few minutes you can reach a business. Or you can accumulate businesses of various types in an area. That is, the location is a determining factor.

3.- Various techniques have been analysed and the neighbourhoods of **Les Corts, San Andrés, Poble Sec, Sant Gervasi-Galvany**, are the most suitable if we **are interested in not having competition nearby**.

4.-If the business we want to set up is focused on **exclusive but not very expensive products** and we are interested in the **downtown** area, then **Poble sec** is the ideal place.

5.-If we are interested in neighbourhoods with the same characteristics as the downtown area but far from it, **San Andrés** is the most suitable place. This neighbourhood can host a type of **local business for daily purchases with high quality but not exclusivity or high prices**.

6.-If the business we have in mind is a **local market for daily purchases and high-end products**, **Sant Gervasi-Galvany** is the best neighbourhood. The **high RFD indicator** and its low number of businesses are a good place to set up a business with the characteristics described.

7.-If the type of business to be opened is **totally exclusive and has no direct competitors due to the differentiation of the product**, **any of the first neighbourhoods obtained in the multiple linear regression** is a good candidate. In this case, it would be interesting to be in the downtown area and have anyone from Barcelona as a client.

8. Future directions

The analysis has been carried out and it has been determined which are the best neighbourhoods to implement an ecological supermarket. The next steps to improve the study are to obtain better quality data and analyse more factors such as distance to the city center. To improve the quality of the study, factors such as the success of the established businesses can be searched to see if an area is prosperous with respect to that type of business. Once these data are obtained that broaden and improve the study, it can be automated to perform the analysis for any type of business in which one is interested.

Analysis of neighbourhoods in Barcelona to open an ecological market.

9. References

[1] 'Getting started with Data Science' Publisher: IBM Press; 1 edition (Dec 13 2015) author: Murtaza Haider.

[2] https://es.wikipedia.org/wiki/Distritos_de_Barcelona - Wikipedia information of neighbourhoods

[3] https://ajuntament.barcelona.cat/estadistica/catala/index.htm - Barcelona ajuntament. Officina municiapal de dades

Documentation of the modules used

BeautifulSoup -> https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Pandas -> https://pandas.pydata.org/docs/

Plotly -> https://plotly.com/python

Foursquare -> https://developer.foursquare.com/docs/

Geopy -> https://geopy.readthedocs.io/en/stable/

Folium -> https://python-visualization.github.io/folium/

10. Datasets and materials

Notebooks -https://github.com/gfantonipy/Coursera_Capstone/tree/main/CAPSTONE/notebooks

Results - https://github.com/gfantonipy/Coursera_Capstone/tree/main/CAPSTONE/Results

Datasets - https://github.com/gfantonipy/Coursera_Capstone/tree/main/CAPSTONE/dataset

23-05-2021