Applied Data Science Capstone

The Battle of Neighborhoods - Restaurants in Barcelona

Author: Adolfo López-Cerdán Contact: adlpecer@gmail.com

Table of Contents

Introduction

Have you ever imagined setting up your own restaurant?
Opening a new and successful restaurant is probably a tough but exciting challenge. There are several points to consider to achieve it. One of the most important is where the new restaurant will be placed. Indeed, the physical location of the food-service business is a key point to get known, popular and to set the best possible price range.
In this project, we will work with restaurant data from Barcelona (Spain) to determine which neighborhoods are more convenient to start this kind of business.

Why Barcelona?

Barcelona is a coastal city in the northeast of Spain. It is the second most populated spanish city, with a population of 1.6 million, and it heads one of the most populous metropolitan areas in the European Union. Besides, Barcelona is considered one of the world's main tourist, economic and cultural centers. Therefore, it is a trendy and cosmopolitan city, plenty of people from all over the world.
Such a variety of people and cultures makes Barcelona an inviting pool of potential customers but also a competitive market to start a business in the food-service sector. In that sense, this project could be helpful in the first steps of setting up a restaurant.

Business Problem

The objective of this project is to identify the most suitable neighborhoods in Barcelona to open a restaurant. To achieve this goal, data science methodology will be applied to Barcelona neighborhoods and venues data. Exploratory analysis of data and neighborhood clustering will be used to answer the business question: Which neighborhoods of Barcelona would be more advisable to consider if you want to maximize the popularity of your restaurant?

Target Audience

This project could be particularly interesting to:

- New entrepreneurs who wants to start their own restaurant.
- Restaurant owners who are planning to move the location of their business.
- Investors looking for opportunities in the real estate sector.
- Food franchises looking for new locations.

Data

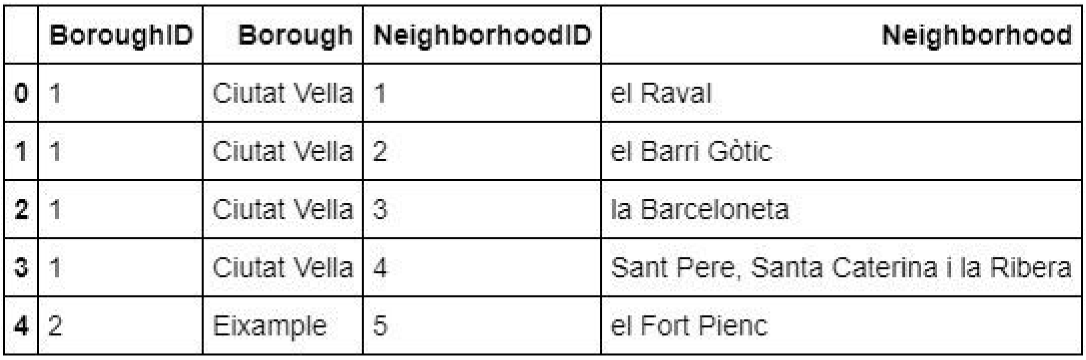
The following data has been used in this project:

Neighborhoods list

The list of neighborhoods in Barcelona has been extracted from the *Open Data BCN* service (https://opendata-ajuntament.barcelona.cat/), a platform managed by the Municipal Data Office. This data source contains a wide and open catalogue of datasets from several public administrations in Barcelona.

In this case, the dataset *Administrative units of the city of Barcelona* has been used. This dataset include a table with a total of 73 records that correspond to each neighborhood of the city with the name of the neighborhood, its ID and the borough to which it belongs (table 1).

Table 1: 5 first rows of *Administrative units of the city of Barcelona dataframe.*



	BoroughID	Borough	NeighborhoodID	Neighborhood
0	1	Ciutat Vella	1	el Raval
1	1	Ciutat Vella	2	el Barri Gòtic
2	1	Ciutat Vella	3	la Barceloneta
3	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera
4	2	Eixample	5	el Fort Pienc

Neighborhoods boundaries

One of the visualization tools most used in this project is the choropleth map. This tool allows shading areas of a map in proportion to a numerical variable. However, the geospatial data of the neighborhood's limits is required to plot them on a map.

To achieve this goal, a *GeoJSON* file containing those limits has been extracted from the GitHub repository *bcn-geodata* (https://github.com/martgnz/bcn-geodata), created by Martín González.

Foursquare data

All the restaurant data have been extracted from *Foursquare Database* by making the right calls to *Foursquare Places API*. Two kinds of data have been obtained in this way:

- Restaurant names, IDs and categories by making the regular "*GET explore*" call for each neighborhood.
- The number of likes, Rating and Price tier by making the premium "*GET* details" call for each restaurant.

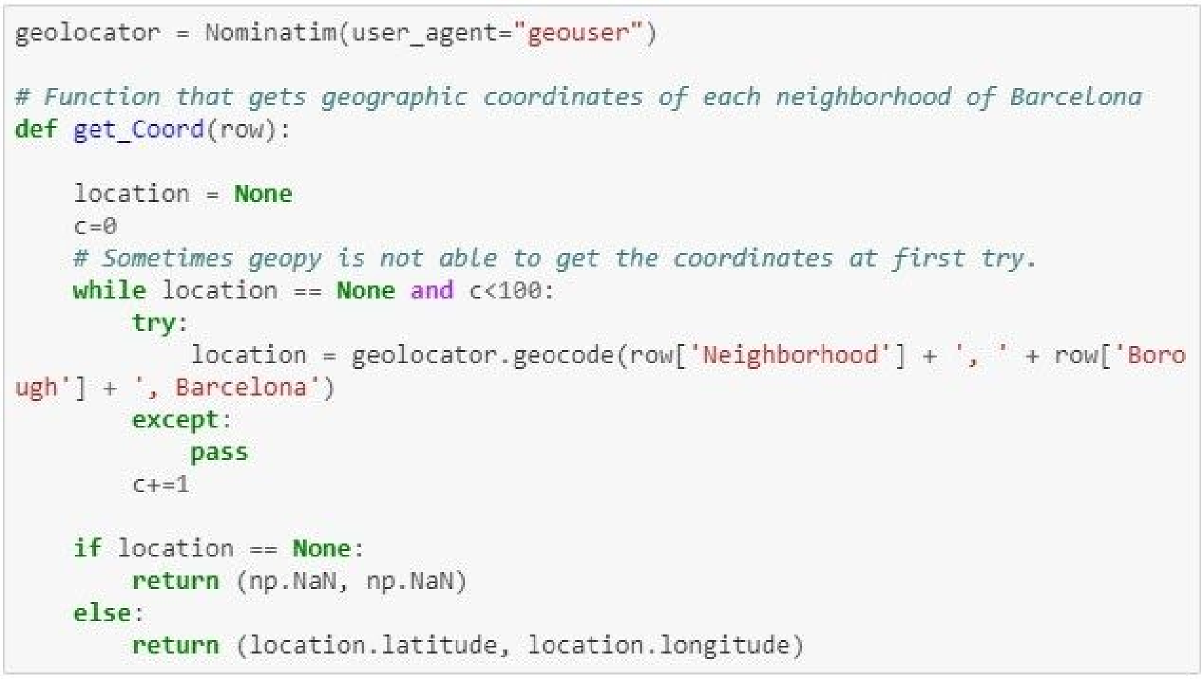
All the data has been concatenated to get a data frame containing all restaurant data.

Methodology

The methodology followed in this project is composed by the next three parts:

Data collection and preprocessing

In the first place, data from the dataset *Administrative units of the city of Barcelona* has been extracted as a data frame applying the *Pandas* function *Pandas.read_csv()* with the CSV table URL as argument. This dataframe has been completed with the geospatial coordinates of each neighborhood using a custom function (figure 1) to apply the geolocator *Nominatim* from the Python library *Geopy*. Finally, the neighborhoods coordinates were reviewed in order to detect and manually correct the NaN values.



```
geolocator = Nominatim(user_agent="geouser")

# Function that gets geographic coordinates of each neighborhood of Barcelona
def get_Coord(row):

    location = None
    c=0
    # Sometimes geopy is not able to get the coordinates at first try.
    while location == None and c<100:
        try:
            location = geolocator.geocode(row['Neighborhood'] + ', ' + row['Borough'] + ', Barcelona')
        except:
            pass
        c+=1

    if location == None:
        return (np.NaN, np.NaN)
    else:
        return (location.latitude, location.longitude)
```

Figure 1: Custom function that applies the geolocator Nominatim.

In the second place, the neighborhoods boundaries were extracted as a *GeoJSON* file applying the *request.get()* function to the file URL. Boundaries data and coordinates were plotted in an interactive map using the python library *folium*. All inconsistencies between both data sources were manually corrected in the neighborhoods data frame.

Finally, venue data was collected from Foursquare Database following two steps:

1. The ID, name and category of all venues were extracted applying a custom function that performs the API call “explore” for each neighborhood. These venues were filtered to keep only the restaurants (venue that contains the word“restaurant”in its category).

2. The price tier, rating and number of likes were extracted applying another custom function to perform the API call “details” for each restaurant. This step was done in two parts because of the Foursquare limit for premium calls (500 per day). All venue data was collected in a dataframe with the following columns: Venue ID, Name, Neighborhood, Category, Price tier, Rating and Likes. Finally, rows with missing values were dropped.

Exploratory analysis

In this part, all variables were summarized with visual methods to understand the variability of the data set.

In the first place, univariate analysis of all features was performed: On the one hand, the count values of the categorical variables (Neighborhoods, Boroughs, Categories and Price Tiers) were plotted as horizontal bar plots. On the other hand, both boxplots and histograms were displayed to represent the numerical variables (Ratings and Likes).

In the second place, bivariate analysis was performed to pairs of variables: In the case of categorical-numerical pairs, boxplots segregated by categories were selected. In the case of categorical-categorical pairs, grouped bar plots were chosen to visualize the possible relationship between groups.

Finally, the restaurants were grouped and aggregated by neighborhoods in a new data frame to study the relative frequency of the restaurant categories for each neighborhood and to prepare the data for cluster analysis.

Cluster analysis

At this point in the project, the machine learning method K-means clustering was applied to grouped data to partition the neighborhoods in separate groups based on their features. This method, considered as one of the most popular methods, uses a K number of centroids to define clusters: Each point belongs to a particular cluster if it is nearer to that cluster centroid than to any other.

The first step to apply this method was to select the best K to this data. To achieve it, K-means models were constructed using the *sklearn.cluster.Kmeans* function with K taking values from 2 to 10. Then, the Silhouette score was computed to each model with the *sklearn.metrics.silhouette_score* function. The scores for each K value were displayed as a line plot and the K with the highest score was selected.

Lastly, the final model was fitted to the data with the best value for K and the predicted labels were extracted and added to the grouped data set.

Results

The methodology applied in this research led to the following results:

Exploratory analysis

Univariate analysis of categorical features revealed a clear imbalance between classes.
As we can see in figure 2, the distribution of restaurants between neighborhoods is asymmetric, with a progressive increase in the number across neighborhoods. It is remarkable that, despite the great differences, clearly differentiated neighborhood groups are not observed.
In the case of categories, the imbalance between classes is even more pronounced. Indeed, the three most represented categories exceed 100 counts while the rest ones do not reach 50 counts.
As for the price tiers, we can find that the middle tiers (2 and 3) include most of the restaurants in Barcelona, while tiers 1 (the cheapest) and 4 are the less represented. As it seems logical, the most expensive restaurants (tier 4) are the least frequent.

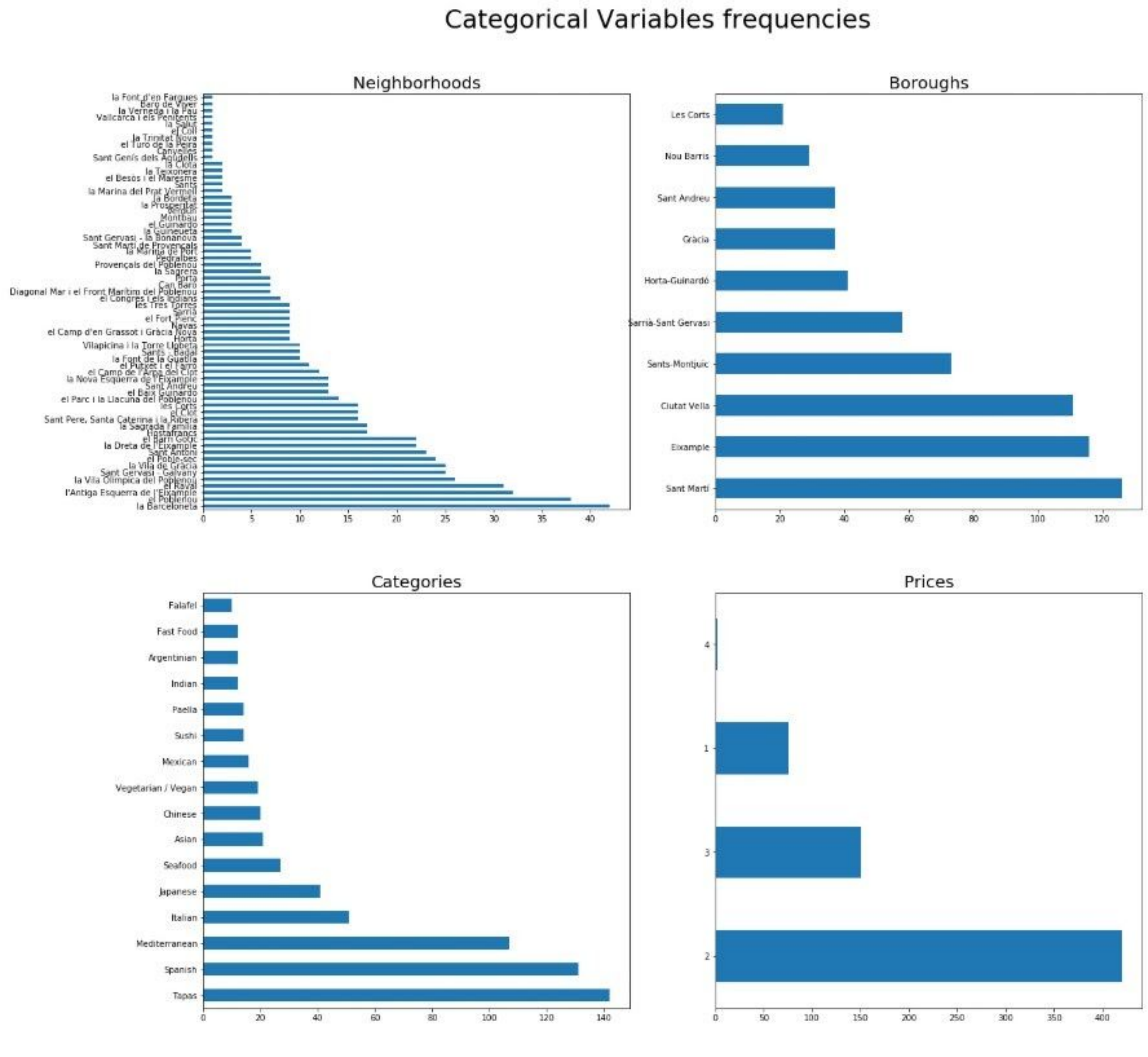


Figure 2: Horizontal bar graphs displaying the value counts for all of the categorical variables.

Meanwhile, univariate analysis of the numerical features shows great differences in values and distributions between both variables (fig. 3).

On the one hand, rating values are in a score range greater than 4 and less than 10. Values tend to cluster around score 8 in a distribution near to normal. Also, there are no outliers in this variable.

On the other hand, the likes count shows a massive number of outliers. Indeed, both plots show a high proportion of restaurants with a modest count of likes (range from 0 to about 250) and a fortunate group of several restaurants with explosive numbers of likes (outliers).

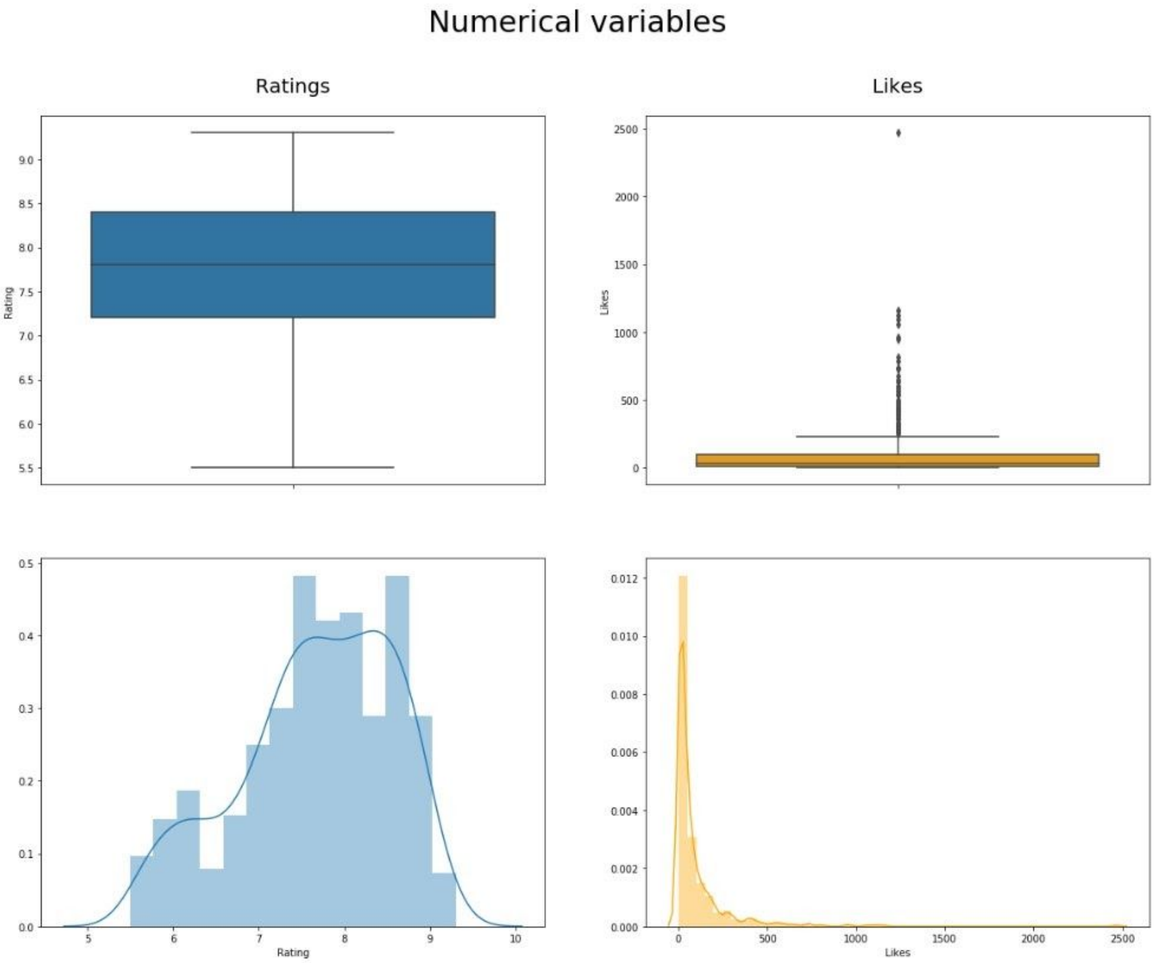


Figure 3: Box plots and Histograms graphs displaying the values of the numerical variable.

These results from univariate analysis were enriched with the contribution of multivariate analysis results.

As we can see in figures 4 and 5, the rating of Barcelona restaurants presents some variation between the classes contained in the category and district variables. In the case of the restaurant categories, the figure shows that some kinds of restaurants like vegan/vegetarian tend to obtain better ratings. On the contrary, other classes like fast food get weaker evaluations. However, the high intraclass variability impedes us to extract more conclusions.

In the case of the city boroughs, results reveal a significant geolocation component in rating variability. Indeed, three boroughs (Ciutat Vella, Eixample and Gràcia) tend to get better rating values than the other.

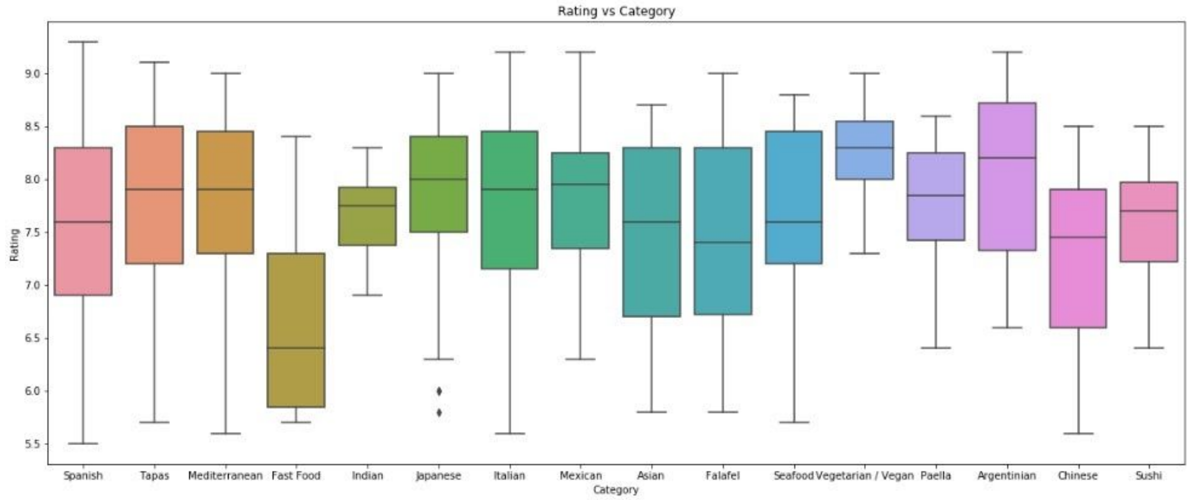


Figure 4: Box plots displaying rating values across restaurant categories.

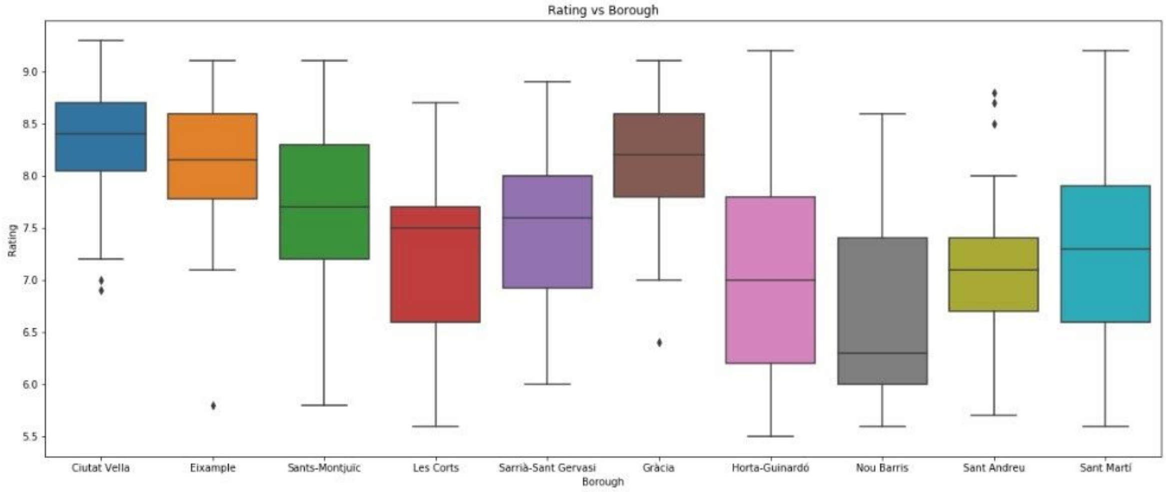


Figure 5: Box plots displaying rating values across Barcelona boroughs.

The bivariate analysis of the number of likes was slightly different. In this case, the number of likes was discretized in 4 groups because of the high quantity of outliers. Those outliers made impossible to get valid results by displaying box plots, so the likes ranges were plotted versus the categorical variables as bar graphs. The results obtained by this way can be observed in figure 6.

Regarding these results, it is remarkable that:

In the first place, the 4 likes tiers are represented in most of the restaurant categories but with very different frequencies. Between the top 3 categories, Tapas restaurants seem to be more popular than Mediterranean and Spanish restaurants. Seafood and Paella categories have important frequencies of tiers 2 and 3. In contrast, Fast Food and Chinese categories tend to accumulate restaurants in the lower ranges of likes.

In the second place, there is a group of three boroughs clearly differentiated with the other in terms of likes ranges frequencies. This group presents the most important frequencies of tiers 2 and 3 restaurants and a minimal frequency of tier 0 (the less popular). Also, this result completely match with the result of the ratings analysis meaning that these three boroughs are undoubtedly the most popular of Barcelona.

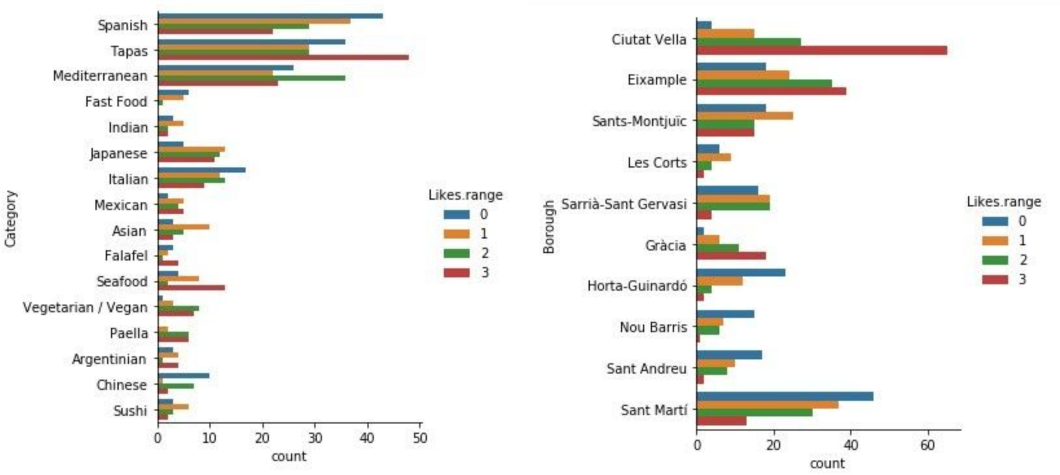


Figure 6. Horizontal bar graphs displaying the value counts for each likes tier separated by: left) Restaurant categories, right) City boroughs.

Cluster analysis

In the first place, it was necessary to calculate the optimal number of centroids for cluster modeling. As a result of the K selection method, the model fitted with a value of 2 centroids achieved the best Silhouette score (figure 7). Therefore, the K value of 2 was selected for the main analysis as the optimal number of clusters.

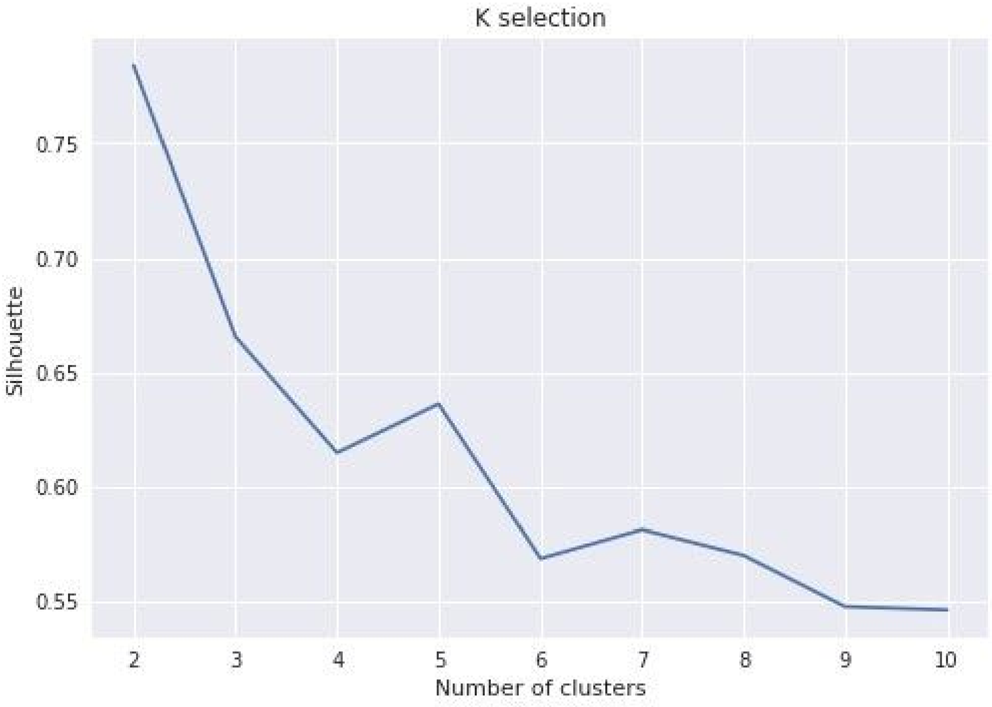


Figure 7: Line plot displaying the Silhouette scores for each value of K (number of centroids).

The model fitted using 2 centroids led to the classification of the neighborhoods in Barcelona in two different clusters according to their similarities. A total of 53 were categorized into cluster 0, while 9 were assigned to cluster 1. The remaining 11 neighborhoods were excluded from the analysis due to the absence of restaurant data.

Once the neighborhoods were classified into the clusters, the possible differences between the two were evaluated. These analysis showed the following results:

- In the first place, neighborhoods from cluster 1 tend to be more popular than the ones from cluster 0. This result is clearly observable in figure 8 were both ratings and likes counts are significantly higher for cluster 1.

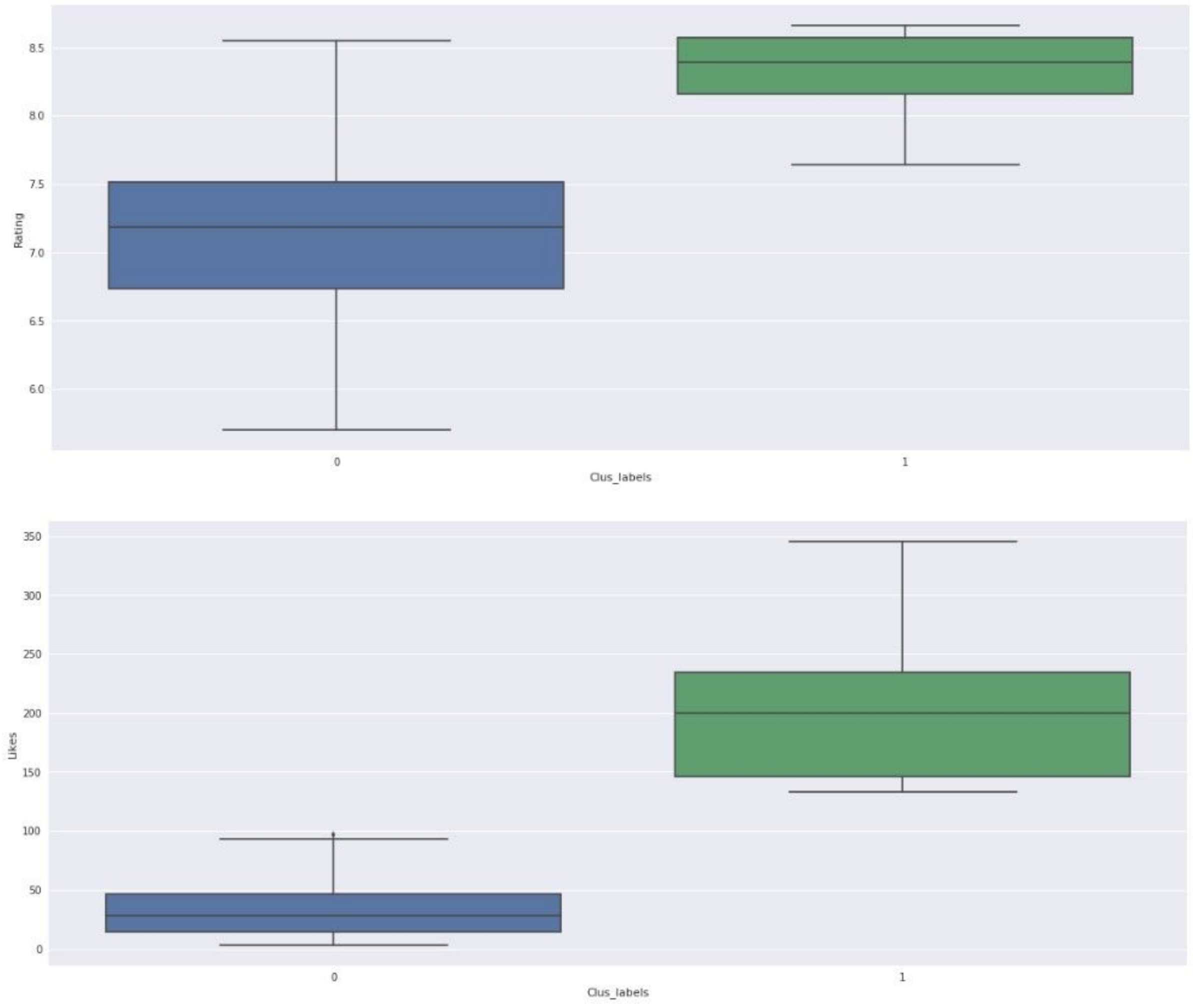


Figure 8: Comparative box plots for each cluster displaying: Top) Rating values, Down) Likes counts.

- In the second place, neighborhoods from cluster 1 are physically adjacent to each other in a concrete and central area of the city as can be seen in figure 9. Cluster 0 neighborhoods, by contrast, fill the rest of the city. These results could point to the existence of a trendy urban area in Barcelona, filled with fashionable restaurants and, possibly, other types of leisure offers.

- Finally, all restaurant categories are represented in both clusters but the relative frequencies are slightly different (figure 10). The categories Tapas, Vegetarian/Vegan, Seafood, Mexican and Japanese are overrepresented in cluster 1. On the contrary, the categories Spanish, Mediterranean and Fast Food tend to be more common in cluster 0.

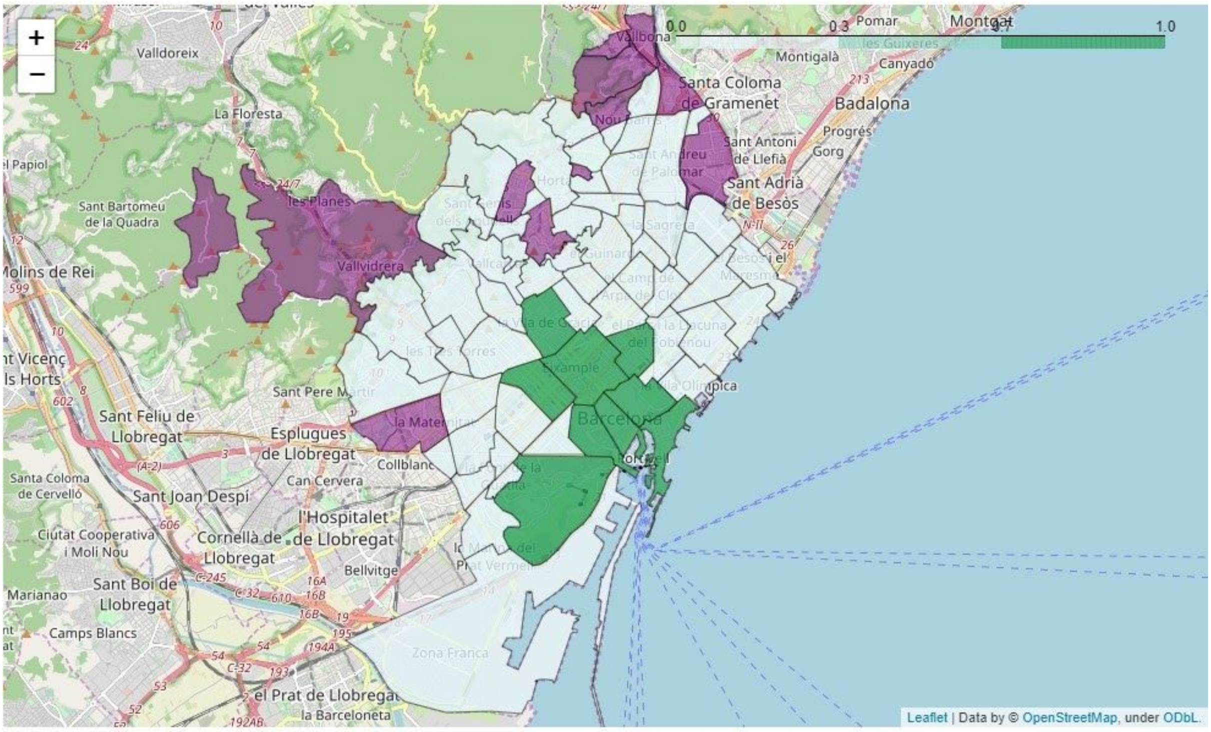


Figure 9: Cluster representation across the neighborhoods of Barcelona. Green neighborhoods belong to the cluster 1 and light blue neighborhoods belong to the cluster 0. Purple neighborhoods are not included in cluster analysis by lack of restaurant data.

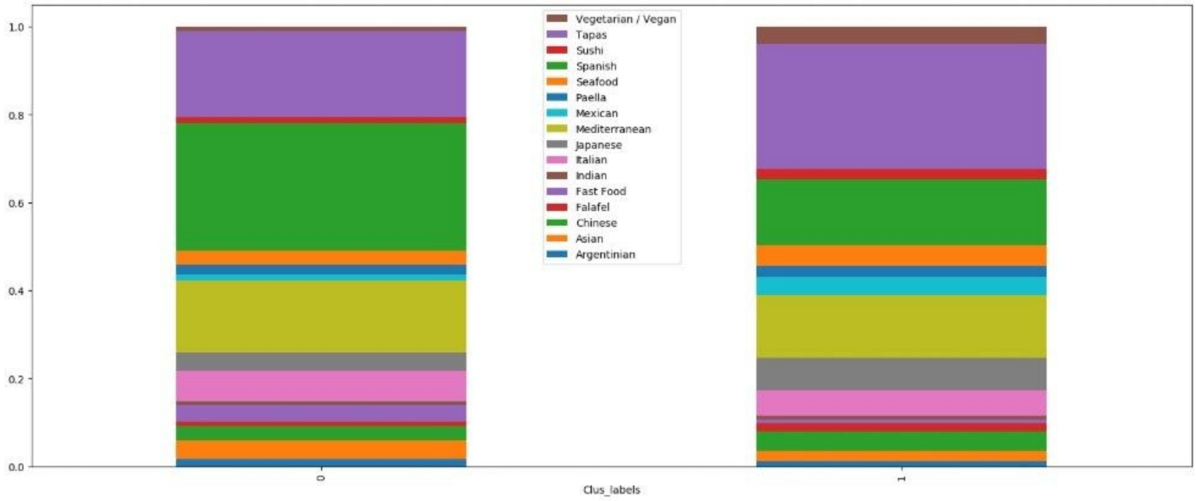


Figure 10: Comparative stacked bar plots displaying the relative frequency of the categories for each cluster.

Discussion

As can be noted in cluster analysis results, neighborhoods of Barcelona can be split into two groups according to its restaurant features. Both clusters differ in terms of the kind of their restaurants, the urban area involved by each one and, especially, the popularity that their restaurants have raised. These results are enormously relevant to answer the proposed business question: Which neighborhoods are more advisable for setting up a highly successful restaurant?

The existence of only two different clusters makes quite easy the solution: The neighborhoods from cluster 1 are the most convenient to consider if you want to start a food-service business in Barcelona. The explanation could be this simple: All the neighborhoods grouped in cluster 1 are part of a greater trendy urban area full of attractions apart from restaurants. This area holds a central and well-communicated position in the city, between three very iconic and touristic boroughs: Ciutat Vella, Eixample, and Gràcia.

However, despite being this popular, this area could have high prices for premises that can compromise the viability of a restaurant. Indeed, one limitation of this project is the lack of real estate data that could lead to a better solution in terms of a balance between popularity and economic cost. Therefore, further analysis that include this kind of data could fill this gap to reach a more precise and profitable solution.

Conclusion

In this project, restaurant data from Foursquare's Database has been retrieved and analyzed to get data-driven recommendations for opening restaurants in Barcelona. The results obtained point to the existence of a highly popular area that enhances its restaurants' popularity. Therefore, setting up a restaurant within this area could improve the probability of business success. Further analysis with additional data could improve the precision of these recommendations, although it is quite clear that the purpose of this project has been satisfied.

In conclusion, the achievement of this objective highlights the effectiveness of the use of data science methodology to improve decision-making in business.