Opening a new restaurant in Barcelona

Sala, Albert

Coursera Capstone

6/25/20

Introduction

In Barcelona, food is a lifestyle. Traditional and avant-garde delicacies are put together in an outstanding expression of Mediterranean cuisine and eating out is both a social need and a custom deeply rooted in local culture. From tiny and crowded tapas bars to stunning fine dining venues are spread all over the Condal city territory.

However trendy spots, streets or even neighbourhoods usually change all over the time, and the public for these areas is also different. Barcelona counts with locals, expats and with uncountable tourists all year round, and the preferences for each group are significantly different. This makes the decision on opening a restaurant, a complex project where the type of cuisine, attention, price and targeted public could define the likely location where the business would flourish.

Business Problem

The objective for this project is to analyse with Data Science and Machine Learning the current market of restaurants in Barcelona by small areas within the city, to provide a quantifiable detail rather than a recommendation from a friend telling that one zone is popping. Besides the zonal business analysis, the society will also be considered in the study. Barcelona is an amazing walking city, so people usually wander around their living zone and end up in a restaurant, so that leads to a question: Is the background and income of the society influencing the restaurant industry? How?

Target Audience of this project

Any entrepreneur who is planning to open a restaurant would be the first ones interested in the findings of this project, but they are not the only ones. The insights can be used for opening any business, from ice cream shops to jewellery businesses. The behaviour of the society, together with their income, nationality and age can define the prosperity for the zone studied, and this is universally applicable.

Data

To solve this project, I will count with different sources of data:

Data sets from the Open Data BCN portal, the Ajuntament de Barcelona's open data service

Open Data BCN, a project that was born in 2010, implementing the portal in 2011, has evolved and is now part of the Barcelona Ciutat Digital strategy, fostering a pluralistic digital economy and developing a new model of urban innovation based on the transformation and digital innovation of the public sector and the implication among companies, administrations, the academic world, organizations, communities and people, with a clear public and citizen leadership.

From this source I will use:

- population.csv containing

Year	District.Code	**District.Name**	**Neighborhood.Code**	**Neighborhood.Name**	**Gender**	**Age**	**Number**

- nationalitynsex.csv containing

Year	District.Code	District.Name	Neighborhood.Code	Neighborhood.Name	Gender	**Nationality**	Age	**Number**

- 2017_rendatributariamitjanaunitatconsum.csv

Year	District.Code	District.Name	Neighborhood.Code	Neighborhood.Name	Section	**Income**

Foursquare API

After that, I will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, I am particularly interested in the Restaurant categories.

Geopy

Geopy's geocoding web services will be used to obtain the longitude and latitude of the different neighbourhoods analysed.

Methodology and Analysis

After obtaining the data we will need to review it and clean it to get information about it.

- From the population.csv we want the official neighbourhoods, boroughs and to create a third columns named Address that will be used to find the location of each neighbourhood.
- From nationalitynsex.csv we will obtain modify calculate the rate of foreigners in each neighbourhood.
- From 2017_rendatributariamitjanaunitatconsum.csv we will obtain the average income in the households of each neighborhood.

Location obtention

With the Data we have there is no Latitude and Longitude data to position every neighbourhood in a map, therefore we proceed to use the Geopy libraries to find those values.

	District.Name	Neighborhood.Name	ADDRESS
0	Ciutat Vella	el Raval	el Raval, Ciutat Vella, Barcelona
1	Ciutat Vella	el Barri Gòtic	el Barri Gòtic, Ciutat Vella, Barcelona
2	Ciutat Vella	la Barceloneta	la Barceloneta, Ciutat Vella, Barcelona
3	Ciutat Vella	Sant Pere, Santa Caterina i la Ribera	Sant Pere, Santa Caterina i la Ribera, Ciutat ...
4	Eixample	el Fort Pienc	el Fort Pienc, Eixample, Barcelona

Figure 1 Neighborhood list obtained after cleaning the population.csv

```
[ ] from geopy.extra.rate_limiter import RateLimiter
    locator = Nominatim(user_agent= 'myGeocoder')
    # 1 - conveneint function to delay between geocoding calls
    geocode = RateLimiter(locator.geocode, min_delay_seconds=1)
    # 2- - create location column
    bcn1['location'] = bcn1['ADDRESS'].apply(geocode)
    # 3 - create longitude, laatitude and altitude from location column (returns tuple)
    bcn1['point'] = bcn1['location'].apply(lambda loc: tuple(loc.point) if loc else None)
    # 4 - split point column into latitude, longitude and altitude columns
    bcn1[['latitude', 'longitude', 'altitude']] = pd.DataFrame(bcn1['point'].tolist(), index=bcn1.index)
```

```
bcn1.head(80)
```

	District.Name	Neighborhood.Name	ADDRESS	location	point	latitude	longitude	altitude
0	Ciutat Vella	el Raval	el Raval, Ciutat Vella, Barcelona	(el Raval, Ciutat Vella, Barcelona, Barcelonès...	(41.3795176, 2.1683678, 0.0)	41.379518	2.168368	0.0
1	Ciutat Vella	el Barri Gòtic	el Barri Gòtic, Ciutat Vella, Barcelona	(Barri Gòtic, el Gòtic, Ciutat Vella, Barcelon...	(41.3833947, 2.1769119, 0.0)	41.383395	2.176912	0.0

Figure 2 Geopy code and result of adding the coordinates on the neighbourhood list

To test the veracity of these locations we plot the map with folium with the boroughs and neighbourhoods tagged and find out that all the locations are correctly positioned.

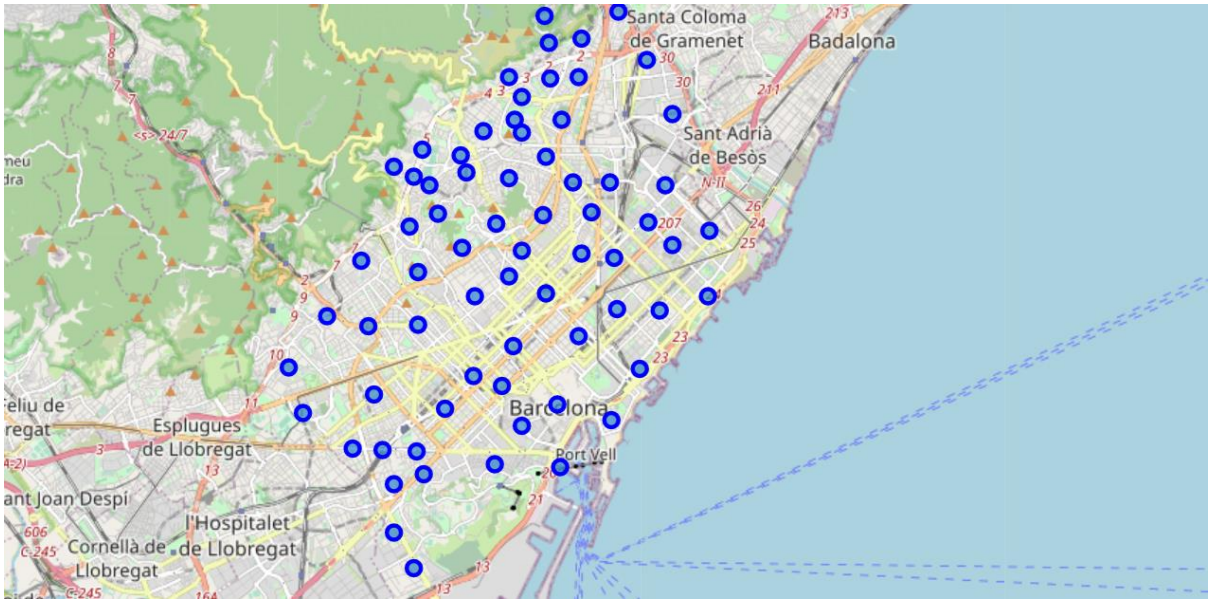

Figure 3 Folium map of the neighborhoods

Foursquare venue data obtention

In this step, we had several discussions on the category we want to obtain for the further clustering. The examples of New York and Toronto took all kind of venues, but we want to obtain an insight mainly in the restaurant sector. Therefore, we did a research on the Foursquare API (https://developer.foursquare.com/docs/build-with-foursquare/categories/) and identified the category that suits us the best; Food category. This category can be found with a unique ID, 4d4b7105d754a06374d81259, that will be placed in the query, and it is also prepared to add other unique ID's in the future. View the code below:

```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)
        categoryId = ["4d4b7105d754a06374d81259"]
        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}&categoryId={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT,
            ",".join(categoryId))
```

Figure 4 Foursquare API URL creation with categoryID included

Cluster Analysis

To identify groups (clusters) with similar characteristics, the unsupervised learning method to our data, namely K-Means algorithm, was applied to our data.

The data used for the clustering method was the 6 most common venues found in each Neighborhood. An important concept for the K/Means algorithm is the number of clusters that are going to be used.

To identify the number of clusters we have used different methods seeing that there was no clear winner:

- Silhouette score
  https://en.wikipedia.org/wiki/Silhouette_(clustering)#:~:text=The%20silhouette%20value%20is%20a,poorly%20matched%20to%20neighboring%20clusters.

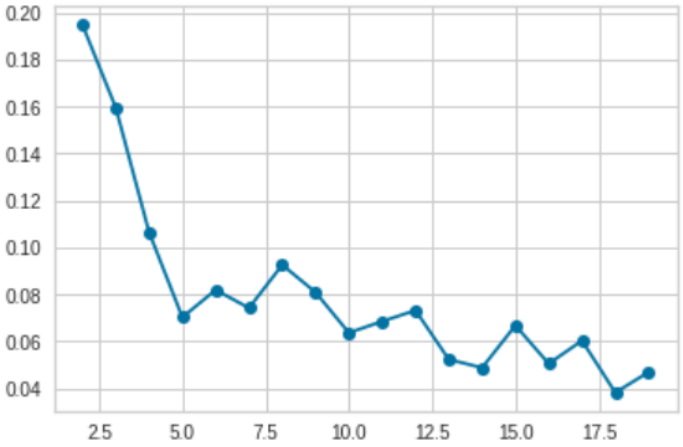

Figure 5 Silhouette Score to get optimal cluster number

- Elbow method
  https://en.wikipedia.org/wiki/Elbow_method_(clustering)

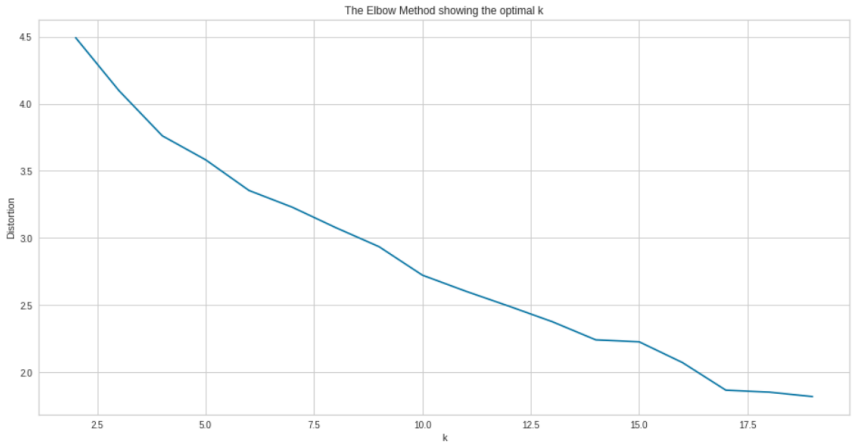

Figure 6 Elbow method for optimal cluster

Among these and other methods tried that can be found in the code like calinski harabasz score or the mixture model, we decided to use 8 clusters, since three were not too many for the diversity of neighborhoods found in Barcelona.

Based on the clustering results, the map below shows you the different clustering found for the neighborhoods of Barcelona.

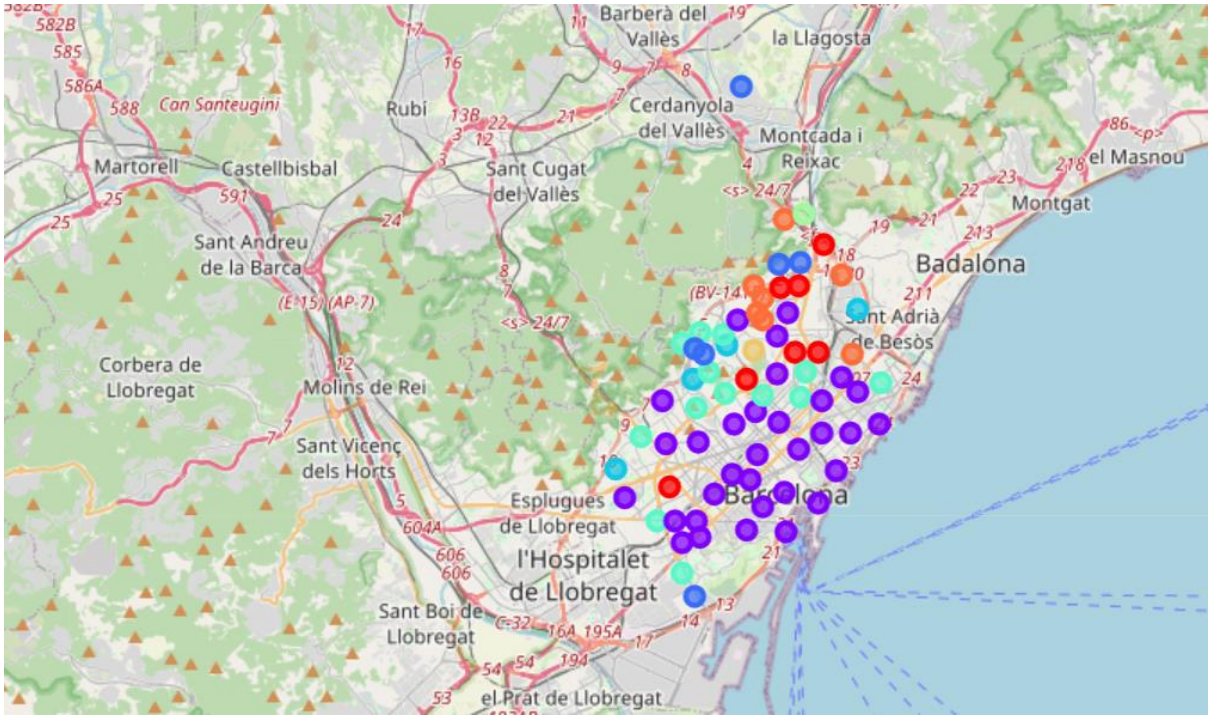

Figure 7 Clustered Map of Neighbourhoods of Barcelona by food

This map shows that the majority of the downtown part of Barcelona and Eixample share the same interests in Food, whereas the Upper-Diagonal zone and residential zones might have another kind of establishments in the food business. There are also differences in the more segregated areas in the north and the ones nearby industrial zones.

The most common types of food venues by cluster are shown in the notebook. Some examples for the most common cluster, the one in purple in the map shown before would be:

```
[ ]  bcn_merged.loc[bcn_merged['Cluster Labels'] == 1, bcn_merged.columns[[1] + list(range(5, bcn_merged.shape[1]))]]
```

	Neighborhood.Name	latitude	longitude	altitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	el Raval	41.379518	2.168368	0.0	1	Spanish Restaurant	Tapas Restaurant	Mediterranean Restaurant	Café	Pizza Place	Seafood Restaurant
1	el Barri Gòtic	41.383395	2.176912	0.0	1	Tapas Restaurant	Spanish Restaurant	Mediterranean Restaurant	Italian Restaurant	Café	Vegetarian / Vegan Restaurant
2	la Barceloneta	41.380653	2.189927	0.0	1	Tapas Restaurant	Mediterranean Restaurant	Paella Restaurant	Spanish Restaurant	Seafood Restaurant	Burger Joint
3	Sant Pere, Santa Caterina i la Ribera	41.372251	2.177532	0.0	1	Mediterranean Restaurant	Restaurant	Chinese Restaurant	Food Truck	Bistro	Diner
4	el Fort Pienc	41.395925	2.182325	0.0	1	Café	Sandwich Place	Chinese Restaurant	Spanish Restaurant	Mediterranean Restaurant	Bakery

Figure 8 Cluster 2, the biggest in Barcelona

As you can see, there is a strong popularity in Spanish, Mediterranean and Tapas restaurants.

We can compare it with a much more residential cluster like the one below to observe the difference.

```
bcn_merged.loc[bcn_merged['Cluster Labels'] == 3, bcn_merged.columns[[1] + list(range(5, bcn_merged.shape[1]))]]
```

	Neighborhood.Name	latitude	longitude	altitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
20	Pedralbes	41.390140	2.112218	0.0	3	Café	Mediterranean Restaurant	Hot Dog Joint	Restaurant	Spanish Restaurant	Pizza Place
27	Vallcarca i els Penitents	41.415712	2.141469	0.0	3	Café	Snack Place	Bakery	Restaurant	Burger Joint	Pizza Place
36	el Carmel	41.425591	2.154959	0.0	3	Restaurant	Café	Bakery	Food Court	Wings Joint	Fish & Chips Shop
58	el Bon Pastor	41.436110	2.204807	0.0	3	Café	Restaurant	Tapas Restaurant	Italian Restaurant	Bakery	Mediterranean Restaurant

Figure 9 Cluster 4, found in residential zones

Income and Nationality relation with the clustering

Another question we wanted to answer was the relation of the income and the nationality with the food offers in each neighbourhood.

After cleaning the rendatributariamitjanaunitatconsum.csv file containing the average income per household in each neighbourhood, we merged it with the clustered data to see if there is any relation.

At first, we plotted the results by using a boxplot from the matplotlib library. But this one presented a lot of outliers in some clusters, so we decided to use also the violin plot from the seaborn library, to observe if there is a second centre in the outliers.

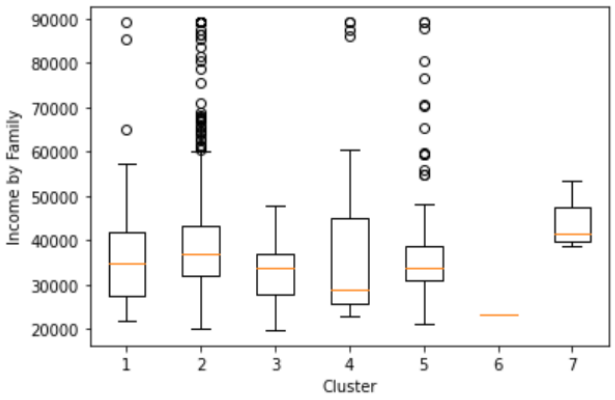

Figure 10 Boxplot of clusters by Income

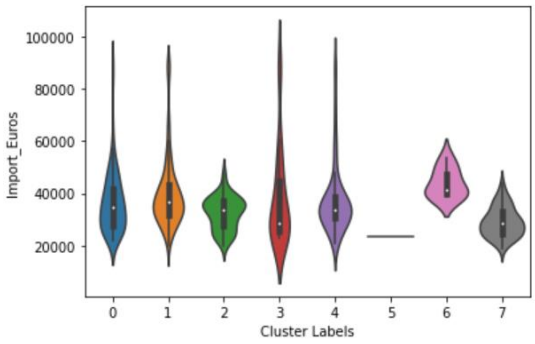

Figure 11 Violin Plot of Income by Cluster

From this graph we cannot obtain any clear relation between the clusters and the economical state of the neighbourhood. This means that the recurrence of certain type of restaurants in a defined zone has nothing to do with the income, just on single cluster is slightly above the others. We also observe some clusters with more economical variability than others.

For the nationality, we obtained a rate by dividing the number of foreigners in a neighbourhood with the total population on that zone in order to obtain a foreigner rate. This rating is compared with the clustering made by food establishment type to see any relation. The comparison is made agau¡in with the violin plot from the seaborn library.

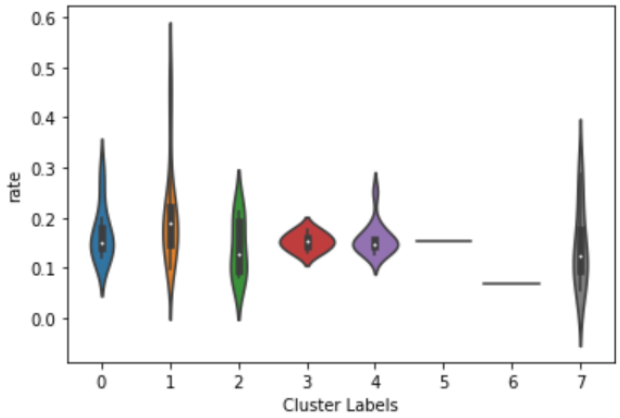

Figure 12 Nationality rate (foreigner) by clusters

We see that the average by cluster is similar, but the variability in the 2nd cluster is much bigger, reaching higher levels of foreigner people.

Results and discussion

The 8 clusters defined represent different types of neighbourhoods that consequently have a different food offering. The second cluster represents all Barcelona's downtown, including oldtown, Raval, Eixample, Poblenou and Gracia. It is also the cluster with more money variability and containing the most foreigners, and despite all these factors the food business in this zone is clearly headed to typical Spanish Restaurants, Tapas and Mediterranean food. Other clusters, located in more residential zones, and with low immigration rate tend to have popular bakeries and cafes.

What could be done better?

Foursquare doesn't represent the full picture, since many venues are not on the list and it is not really used in Spain. For that reason, another map could be utilized such as Google map

Also, Boroughs have too complex geometry, thus defining the closest venues within the certain radius brings additional error to our analysis.

Conclusion

Despite the globalisation and the increase of global food trends in the city, people living in Barcelona, both locals or with other nationalities appreciate the local cuisine. Tapas restaurants, Spanish Food and Mediterranean food is well established and the most visited following the data extracted from Foursquare. The areas that would have an easier introduction of newer trends would be some surrounding residential areas, even though some of them have a high dependency on bakeries and cafes.

Some more analysis could be done by using more reliable sources such as the google maps api, since there is much more info there in Spanish cities.