IBM CAPSTONE PROJECT
-
AN ANALYSIS OF MALAGA'S FOOD VENUES

FIDAE EL MORER

Table of contents

1.Introduction/Business Problem

This project is intended to study the most popular food venues of the city of Málaga (Spain) in order to decide which is the most frequent category, in which neighbourhoods are more likely to be found, and what are the preferred venues, according to their ratings.

Before choosing the problem above described, a little research has been conducted in order to decide the following items:

1. What kind of data can we retrieve from Foursquare and how many calls are we allowed to make to the API?
2. Since I live in Spain and I wanted to study a Spanish city, what are the main sources of open data available?
3. Is the available information enough to carry a study related to geolocated venues?

These questions will be answered in the following section.

The business problem that can be proposed is, given an investor interested in the catering sector, what category of venue would be a good choice according to the methodology described below, and where to locate it.

2. Data

2.1. Sources

In order to find out what are the preferred and most frequent food venues, and whether if there are any coincidences in the results, the following data has been used:

1. **Foursquare data retrieved from the "explore" endpoint.** We can get the following information:
   * Name of the venue.
   * Location (latitude and longitude).
   * Category.

There is a limit of 99.500 calls/day, which is a broad limit given the amount of data that is going to take to carry the project

2. **Foursquare data retrieved from the "details" endpoint.** The following information will be used:

- Name of the venue.
- Rating.
- Count of likes.
- Count of tips.
- Count of photos uploaded.

In this case, the number of calls is much more limited, since a Personal account has only 500 calls/day.

3. **Geospatial data from the open data platform of the City Hall of Málaga.** The data is available if .shp format, wich contained the outline of the neighbourhoods of the city. With the help of the software QGIS, the centroids of the polygons have been calculated and their coordinates have been exported in .json format, in order to work with them easily with the json package.

The reason why the city of Málaga was chosen is because it is one of the few Spanish cities that has an open data portal with geospatial data. Moreover, it is a very touristic city, which means that there is a great number of restauration venues that are likely to be rated by customers.

2.2. Methodology description

The process will consist of two parts with the following breakdown:

Part I

1. Download and transformation of data from the City Hall of Málaga.
2. Download of data from Foursquare using the "explore" endpoint to obtain venues from the different neighbourhoods, filtering by category to get only food venues.
3. Obtainment of the clusters in which the recommended venues can be put with the k-cluster method.
4. Discussion of the results obtained (relation between location of the neighbourhood and most frequent venues, most frequent clusters, etc)

Part II

1. Restriction of the data obtained earlier to a maximum sample of 1.000 venues,

2. Download of data from Foursquare using the "details" endpoint to get the ratings and the count of likes associated to each venue.
3. Discussion of the preferred food venues according to the count of likes and the overall rating.

3. Methodology

3.1. Part I

As said in the Data section, the source of geospatial data has been transformed from shapefile polygons into a json file that contains the coordinates of the centroids of said polygons. This has been done with the open source software Quantum GIS or QGIS, a powerful tool for geospatial analysis.

The centroids of the neighbourhoods are a necessary parameter to introduce to the Foursquare query in order to get information on recommended venues around the given points.

Once these points are obtained, the following step is to retrieve all the information about Málaga's venues, particularly restaurants and cafés.

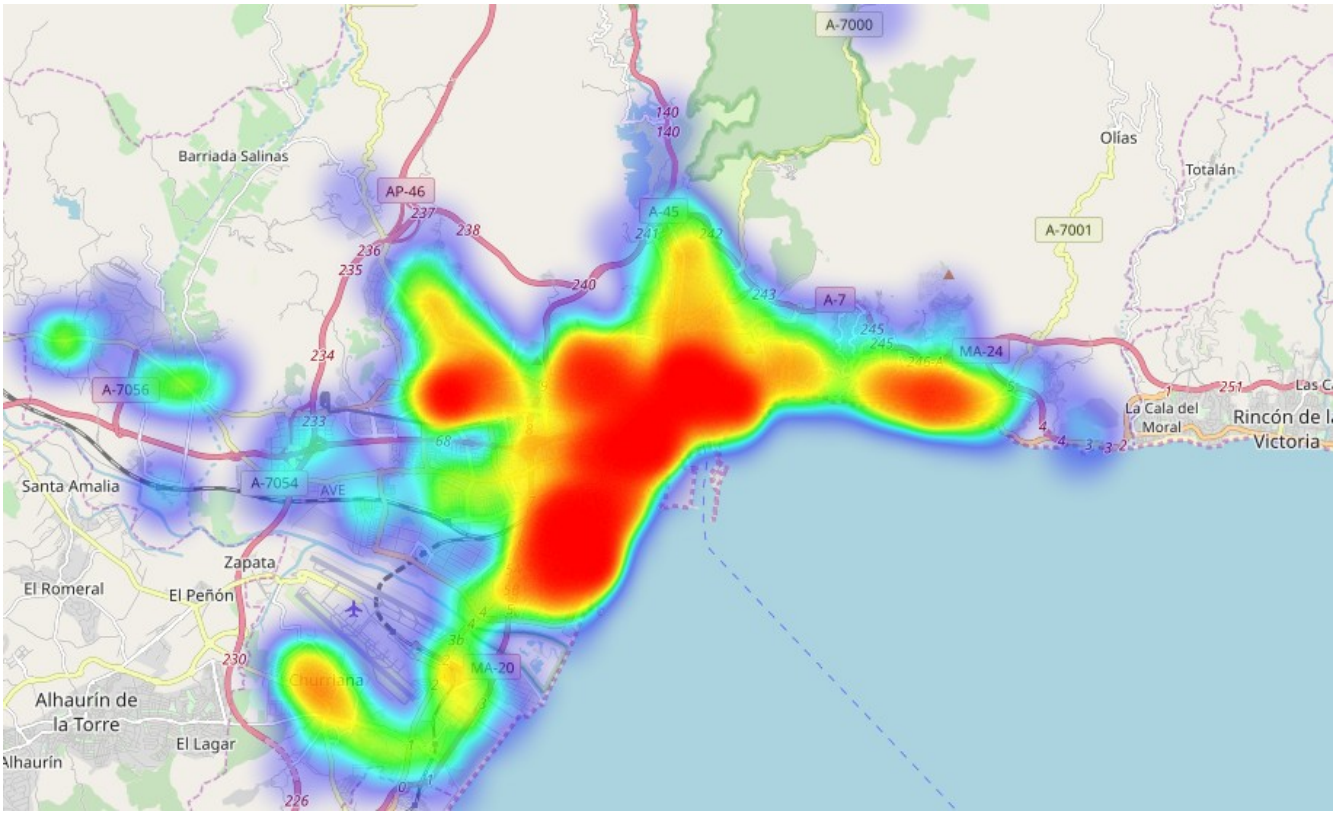

Figure 1: Density of restaurants and cafés in Málaga

Watching Figure 1 we can see 4 major places of concentration of venues:

- The bigger region, that corresponds to the city center.
- The Southwestern region, where the airport and the biggest malls are located.
- The Northwestern region, where there is a university campus.
- The spot on the East side, where there is a great influx of tourists going to the beach

3.1.1. Data Cleaning

The first thing to do once our data is loaded from the Github repository is to clean it in order to avoid duplicates, null values or so. After cleaning the dataframe, we go from 3616 to 1117 venues. The reason of this happening is that when we retrieve data from Foursquare, since it takes values within a radius from a given point, it may take repeated values when there are neighbouring districts.

The second step of data cleaning will be changing or dropping some of the categories, because of the little information they give. For example, there are categories like 'Spanish', 'Tapas' or 'Paella' that basically stand for the same thing, and can all be labeled as Spanish food.

To finish cleaning our data, we can drop categories like 'Food' or 'Restaurant', which give no information about the venues they belong to.

The outcome of this process is represented in the following figure:

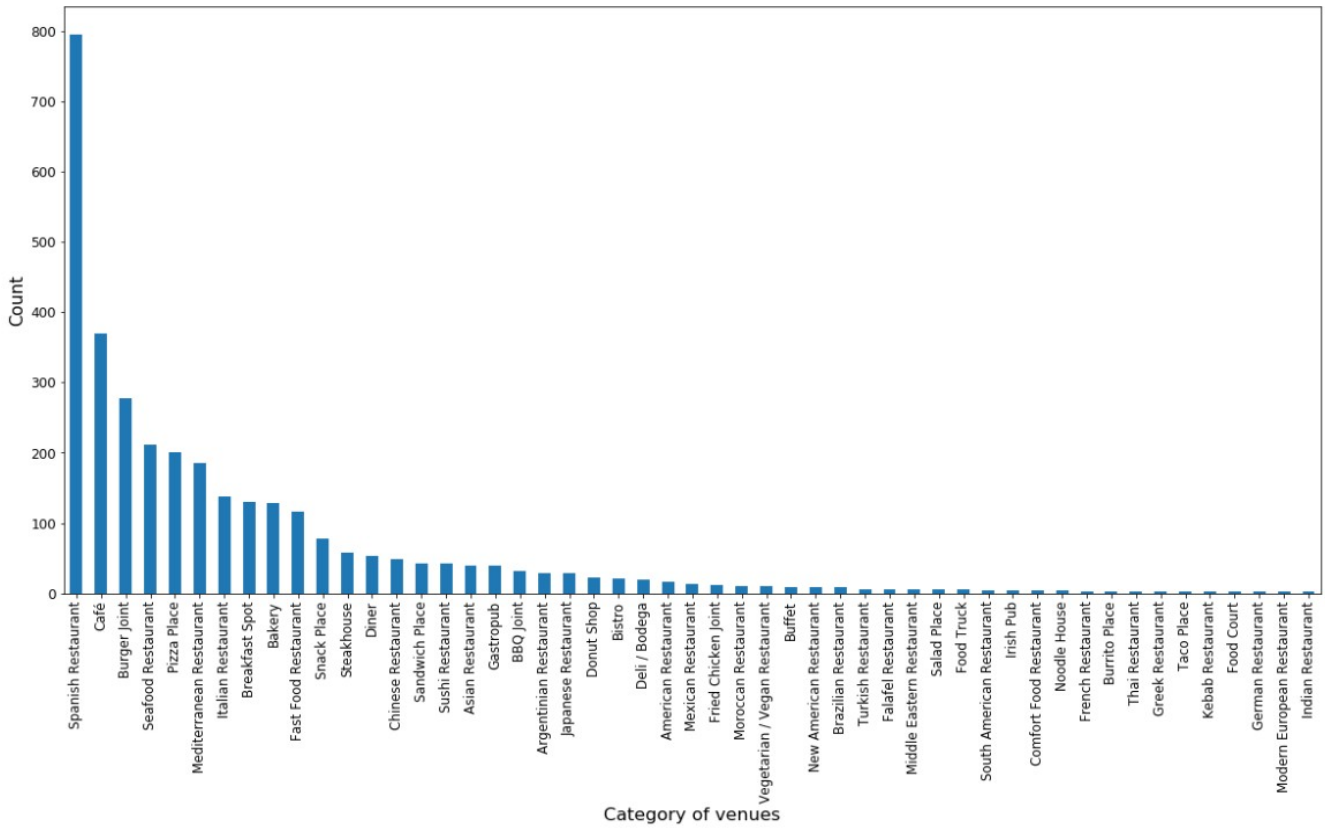

Figure 2: Number of venues per category

The most common venues in Málaga, as it could be expected, are Spanish food restaurants, followed by cafes, and burger joints, which are very common worldwide. At first glance, we might say that Spanish restaurants represent the favorite category of Málaga's citizens. But we have to take in account that it is a very touristic city, where people from all over the world come to, among other things, try Spanish food.

3.1.2. Data clustering

Once our data is clean and ready for analysis, we are going to classify the venues with the K Means method, which is very used to classify and clusterize data.

We are going to classify the data in 10 clusters. Here is a sample of the results obtained:

	Neighb	Latitude	Longitude	Num_neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	26 de febrero	36.741873	-4.428834	106	1	Diner	Vegetarian / Vegan Restaurant	Chinese Restaurant	Fried Chicken Joint	French Restaurant
3	503 viviendas	36.739066	-4.428161	378	8	Fast Food Restaurant	Mediterranean Restaurant	Seafood Restaurant	Burger Joint	Vegetarian / Vegan Restaurant
5	Aeropuerto base aerea	36.678051	-4.495256	331	0	Spanish Restaurant	Fast Food Restaurant	Café	Bakery	Seafood Restaurant
8	Almudena	36.686885	-4.450298	198	0	Spanish Restaurant	Burger Joint	Bakery	Seafood Restaurant	Pizza Place
10	Amoniaco	36.710051	-4.502312	225	9	Bakery	Vegetarian / Vegan Restaurant	Chinese Restaurant	Fried Chicken Joint	French Restaurant
12	Arraijanal	36.661538	-4.466129	327	0	Spanish Restaurant	Seafood Restaurant	Breakfast Spot	Vegetarian / Vegan Restaurant	Chinese Restaurant
13	Arroyo de los angeles	36.730668	-4.432975	112	0	Spanish Restaurant	Diner	Pizza Place	Chinese Restaurant	Vegetarian / Vegan Restaurant
14	Arroyo del cuarto	36.720133	-4.441215	113	0	Café	Spanish Restaurant	Burger Joint	Chinese Restaurant	Fried Chicken Joint
15	Arroyo españa	36.752082	-4.480916	265	7	Mediterranean Restaurant	Vegetarian / Vegan Restaurant	German Restaurant	Fried Chicken Joint	French Restaurant
23	Butano	36.678069	-4.449231	234	7	Mediterranean Restaurant	Vegetarian / Vegan Restaurant	German Restaurant	Fried Chicken Joint	French Restaurant
24	C.t.mercancias	36.715543	-4.511735	519	8	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Chinese Restaurant	Fried Chicken Joint	French Restaurant
28	Camino del colmenar	36.742966	-4.409090	346	9	BBQ Joint	Bakery	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Fried Chicken Joint
29	Campamento benitez	36.652452	-4.483870	298	3	Pizza Place	Vegetarian / Vegan Restaurant	Chinese Restaurant	Fried Chicken Joint	French Restaurant
30	Campanillas	36.726315	-4.543828	287	5	Spanish Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Café	French Restaurant
32	Campos eliseos	36.721488	-4.410584	63	8	Mediterranean Restaurant	Café	Asian Restaurant	Fast Food Restaurant	Irish Pub

Table 1: Clusters and most common venues in each neighbourhood

Once the clustering is finished, we can plot a map showing the distribution of clusters along the city.

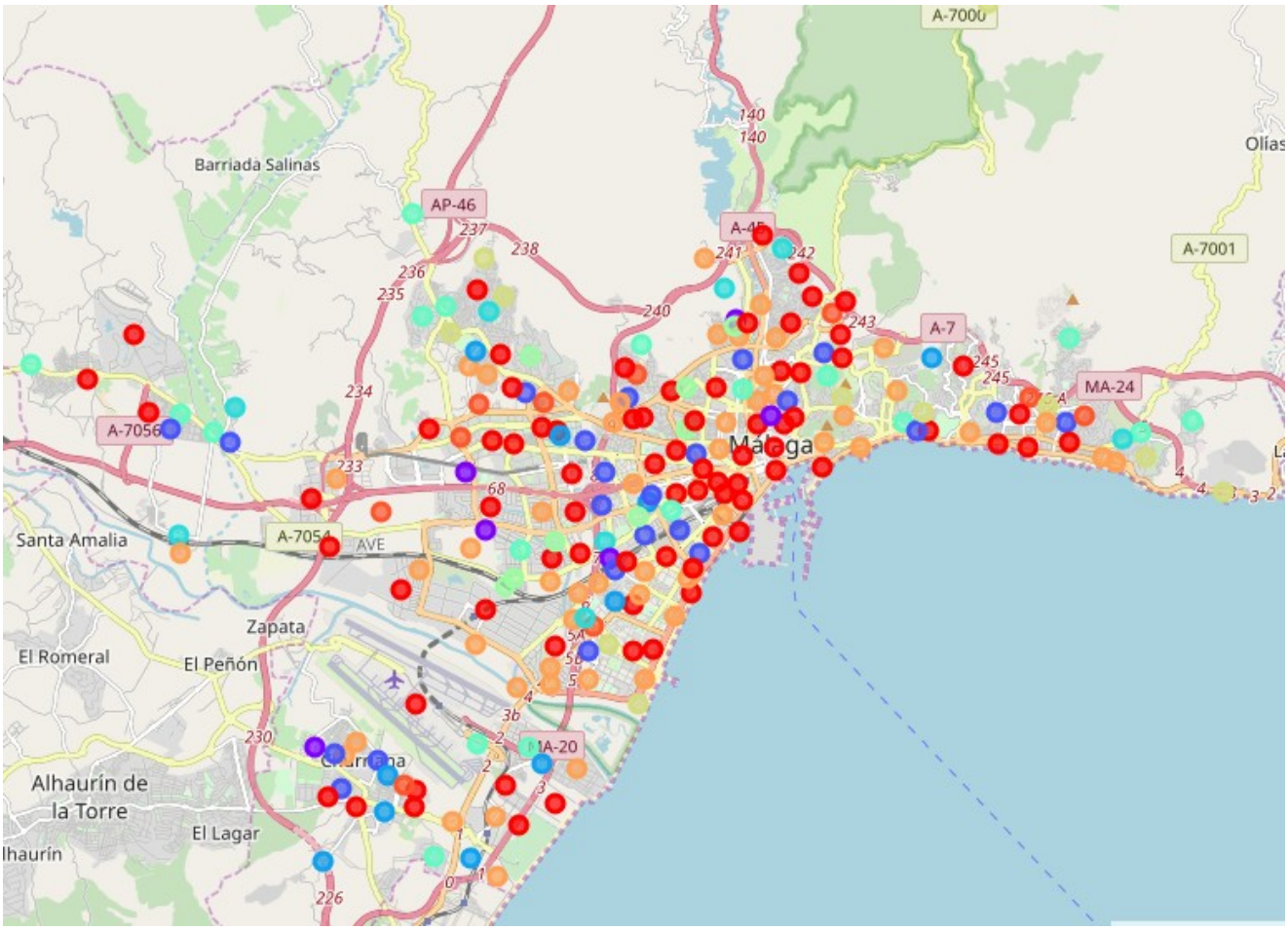

Figure 3: Map showing the distribution of clusters in the city

In the Results sections, we will proceed to discuss the insights of this process.

3.2. Part II

Once we have studied the most common venues and their location, we are going to analyze the prefered venues by the customers, to see if there is a coincidence between both datasets.

In order to do that, let's retrieve the following data from each venue:

- Count of tips
- Rating
- Count of likes
- Count of photos

The goal is to get the preferred venue by the customers, taking account of the data we usually check when we are looking for a place to eat in apps like Foursquare or Tripadvisor.

High ratings do not mean high popularity, but if a venue has a high count of the parameters listed above, it may mean that the customer experience is positive.

Since the Foursquare API has some limitations to the Personal account (500 premium calls/day), we have to split our data in parts of 500 observations. The dataset regarding this matter is stored in the Github repository as well.

To get the best rated venues taking account of all the numerical parameters that we can get from the 'details' endpoint, we are going to divide every column by the maximum value, and get the mean of every row. The outcome is a score from 0 to 1 that will give us a hint about the popularity of a venue.

Category	Mean	Std	Median	Count
Deli / Bodega	0.293389	0.105618	0.293389	2
Vegetarian / Vegan Restaurant	0.234583	0.011946	0.235298	3
Snack Place	0.226270	0.066007	0.226270	2
Sushi Restaurant	0.222839	0.044731	0.215851	8
BBQ Joint	0.220684	0.022997	0.220684	2
Japanese Restaurant	0.215067	0.034536	0.198306	6
Gastropub	0.211376	0.018010	0.212310	13
Spanish Restaurant	0.209763	0.073954	0.203271	161

Table 2: Stats of the resulting dataset

To avoid a bias, we can drop categories with only 1 venue in this dataset. The bias would be caused by a lack of samples, which will give us a null standard deviation. The best results would always be the categories with only 1 venue.

After that, we will create a column called 'Diff', that computes the difference between the mean and the standard deviation. This will return the most solid categories, i.e. the ones with the greatest mean scores and the lowest standard deviations.

4. Results

4.1. Part I

The following figure shows the count of venues per cluster.

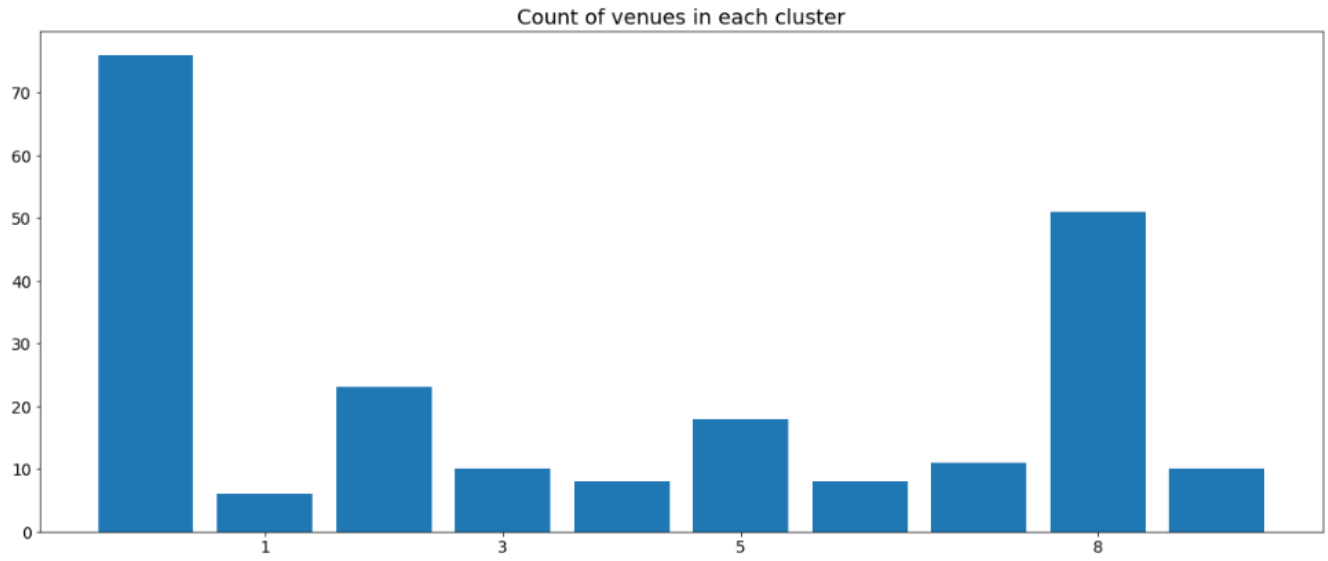

Figure 4: Count of venues per cluster

As we can see, the most populated cluster is `0`, followed by cluster `8`. If we take a glance at the `1`st most common venues of one of the clusters labeled as `0`, we get the following table:

1st Most Common Venue	Frequency
Spanish Restaurant	42
Asian Restaurant	5
Café	5
Bakery	4
Burger Joint	4

Tabla 3: Most common venues in the most populated cluster

It is noticeable that the most common venue is the Spanish food restaurant by far. We can check if there are any similarities between this data sample and the whole population we began the project with.

Category	Frequency
Spanish Restaurant	246
Café	106
Burger Joint	76
Mediterranean Restaurant	60
Seafood Restaurant	60

Tabla 4: Most common venues in the whole dataset

We can see that 3 out of 5 categories are present in both the sample and the population, so it's safe to say that the cluster is a good representation of the whole dataset.

4.2. Part II

The table shown below represents the statistics obtained from the computing process described in Part II of the Methodology section.

The 'Diff' column retrieves the difference between the mean and the standard deviation of each category or neighbourhood. The smallest values mean that the parameters computed (rating, tips count, photos count, likes count) are consistent within the sample.

Category	Mean	Std	Median	Count	Diff
Vegetarian / Vegan Restaurant	0.234583	0.011946	0.235298	3	0.222637
BBQ Joint	0.220684	0.022997	0.220684	2	0.197687
New American Restaurant	0.204292	0.010904	0.204292	2	0.193388
Gastropub	0.211376	0.018010	0.212310	13	0.193366
Deli / Bodega	0.293389	0.105618	0.293389	2	0.187771

Table 5: Results of the top rated categories according to the selected methodology

5. Conclusions

These are a few conclusions that we can extract from the analysis:

1. The clusters are a good representation of the whole studied population.
2. The most common venues, which consist of Spanish (including Tapas and Paella), Mediterranean and Seafood restaurants among others, answer to the fact that Málaga attracts a great amount of tourists that want to try this kind of food.
4. So, following the reasoning of point 2, the neighbourhoods that belong to the most populated cluster are among the most visited places by tourists.
5. However, according to the selected methodology in Part II, the most common venues have no correlation with the preferred ones by the customers. This can be explained by the fact that less common venues are more attractive to people, since there is no variety.

That said, a possible solution to the business problem is to open a venue from the results of Part II in one of the neighbourhoods of the most populated cluster, especially in the neighbourhoods where there is a higher density of venues, or in other words, a higher demand of restaurants and cafés.