Table of contents

1 Introduction and business problem

Within this paper, we will look at the data of restaurants near the region and districts of Barcelona. While Spain can be subdivided into two categories, the castellan and the Catalan region, Barcelona as a city with currently 5,5 million residents is one of the largest cities in the world. In fact, Barcelona is the fifth largest city in Europe. Only Moskva, Paris, London, and Madrid are larger European cities than Barcelona. With this information in our mind, we can see that Barcelona combined with Madrid is one of the most important cities of Spain. Barcelona seems to be a place people and tourists are interested in and the city is globally known.

In this paper we suppose that there is a group of stakeholders, which want to open a restaurant in the region of Barcelona. The stakeholders have a nice Mediterranean cook called "Rordon Gamsey" who already has one Michelin star within his kitchen. He has further restaurants in Italy and near the coast France. With this successful cook, the stakeholders seek help from a data analyst (me :>) to get a better overview about what kind of restaurants there are in Barcelona and which district of Barcelona they should open their restaurant in. Since the group of stakeholders do not want to earn money with the investment, but gain another Michelin star, they really need to get deep information of the locational data of the region of Barcelona, to really nail the place finally they ultimately open the restaurant in. The group of stakeholders is highly interested in getting the best Mediterranean area of Barcelona. Since the stakeholders already opened a lot of restaurants but did not avail a second Michelin-star, they are short on money. Due to this fact, they hired and instructed the soon to be data analyst/scientist Hendrik (cheap labor) to find the perfect place to open the restaurant in and eventually gain another global Michelin star.

2 Data section

We will be using regional data of Barcelona, since we do want to get a better overview over a full region instead of clustering the data too much into minor districts or streets. This is due to the fact, that we want to limit the number of results we will be eventually getting out of the dataset. We first want to explore which region is demanding and supplying the greatest number of restaurants related to Mediterranean dishes. Additionally, after importing the regional data of Barcelona into a data frame, we will use the foursquare API to generate a link, which will support us with a json file which contains information about all the restaurants and venues of the corresponding region of Barcelona. After extracting the data, we must clean it up and drop some columns and rows, ultimately to only use the necessary data.

After preparing and modelling the data, we will analyze and visualize the data. For the stakeholders it would be interesting what kind of restaurants in the region they are competing against. If we conclude that some regions are full of fast food, finger food or small-menu food, we will drop the regions and carry on with the other districts. What kind of competition do they have and what is nearby the surrounding area? Is there a lot of competition or rather not? With the datapoints we will get from the foursquare API we will find the key to success in the corresponding area. After having gained enough information about the districts of Barcelona, we will eventually publish the report to the stakeholders. It will be up to them to decide, whether they want to take the risk of investing and building up a restaurant or not. Our aim is to support them in their decision-making process and give them insights into the regions, so they have all the information they need.

3 Methodology section

The first and most important thing I performed was to import all possible libraries I intended to use, so we can use libraries like the pandas or NumPy. Furthermore, we imported libraries to handle json files, since the data we gather from the foursquare API will be stored as a json file. Additionally, we imported some libraries to handle minor plotting and machine learning. We are not supposed to forget, that we need a library to visualize locational and mapped data. We did this by importing the folium library.

Prior to performing any kind of data analysis, we must prepare and model the data. In the first steps we created a pandas data frame out of the district data and dropped some unnecessary columns and rows. Afterwards we used one of the previously mentioned libraries, to add the latitude and longitude- data to the data frame. We then created the first map within this assignment, by using the folium library to output the map. The map shows all the exported districts from Barcelona, which is not that much at this point.

After this part, we came to the main part of the analysis, the implementation of the foursquare API in connection with the dataset. Since we intend to open a restaurant, we are only interested in the kind of data which connects to the word „restaurant“. In the next step, we are exporting the venue data from the foursquare API, which was within the exported json file. In this step we already could get some interesting results regarding the distribution and category of the restaurants. Since we do not want to export all data that is available, we limit the radius to 2500 meters and the maximum count of exported venues to a number of 50. With this number, we can guarantee, that we do not cross the limit of queries, since the foursquare API only supports a limited number of exported queries at a day, which is currently 950 queries. After connecting the location data of the corresponding venues with the district data, we already have a lot of interesting insights, which will be shown and discussed later in this paper.

In the last step we calculated the mean of all categories for each region and indexed it as a number. After this, we looked at how the restaurants were distributed. Are there Mediterranean restaurants in our district or is it rather fast food, which we are not interested in?

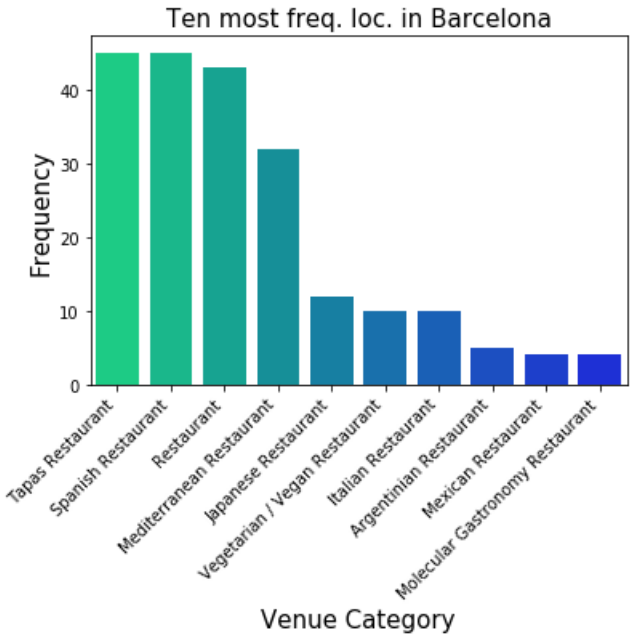
In the last section of the analysis, we used the k-means algorithm to find clusters in the data. The k-means algorithm is a machine learning algorithm, which is implemented in the sklearn library of python. Since we did not want to have a high number of clusters, we stuck with a total number of three clusters, since we do not intend to divide our data into too many categories and only want to get a general overview.

4 Results and clustering

One of the first interesting results we gathered from analyzing the data, was the distribution of different restaurants in the regions of the district. We can clearly see that people in Barcelona tend to prefer tapas and Spanish restaurants over any kind of other restaurants. It is followed by core-Spanish restaurants, and restaurants. Here we can already see a problem with the retrieved data. We cannot distinguish „Spanish restaurants“ and „restaurants“ and do not know where they differ.

	Venue_Category	Frequency
0	Tapas Restaurant	45
1	Spanish Restaurant	45
2	Restaurant	43
3	Mediterranean Restaurant	32
4	Japanese Restaurant	12
5	Vegetarian / Vegan Restaurant	10
6	Italian Restaurant	10
7	Argentinian Restaurant	5
8	Mexican Restaurant	4
9	Molecular Gastronomy Restaurant	4

The distribution of the restaurants was plotted using the seaborn library.



After plotting this, we wrote some code to see how much percent the corresponding category of a restaurant is above the mean of the district. Since we are only interested in the Mediterranean restaurants, we see two districts which really stand out in comparison to the other districts: Sants-Montjuic and Les Corts:

----Sants-Montjuïc----

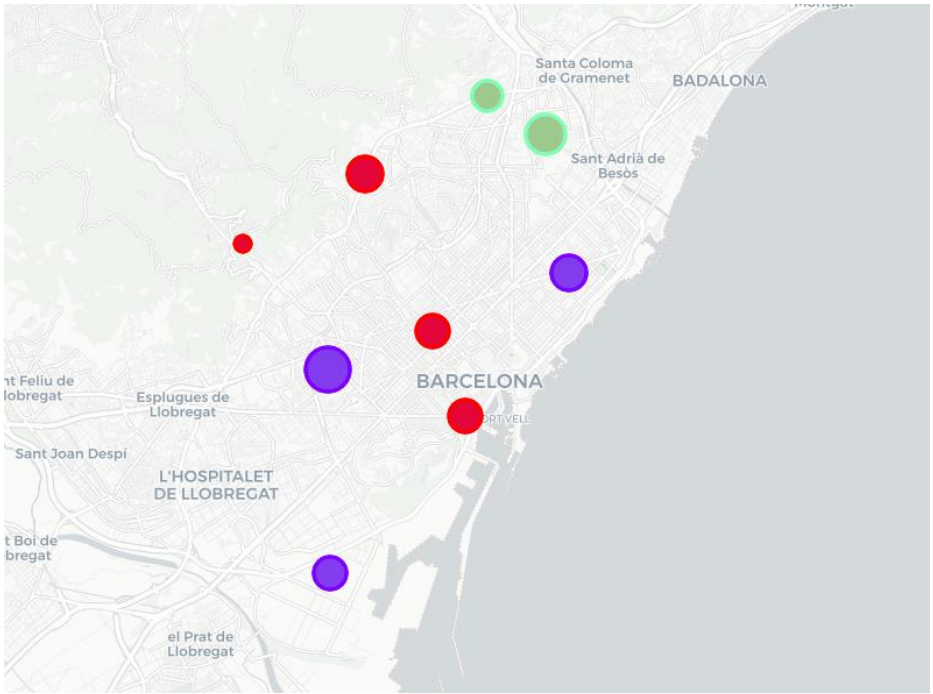
	venue	freq
0	Restaurant	0.22
1	Mediterranean Restaurant	0.19
2	Spanish Restaurant	0.15
3	Chinese Restaurant	0.07
4	Molecular Gastronomy Restaurant	0.07
5	Tapas Restaurant	0.07
6	Shabu-Shabu Restaurant	0.04
7	Asian Restaurant	0.04
8	Comfort Food Restaurant	0.04
9	Fast Food Restaurant	0.04

----Les Corts----

	venue	freq
0	Restaurant	0.19
1	Mediterranean Restaurant	0.16
2	Japanese Restaurant	0.16
3	Spanish Restaurant	0.11
4	Tapas Restaurant	0.08
5	Thai Restaurant	0.05
6	African Restaurant	0.03
7	Asian Restaurant	0.03
8	Vegetarian / Vegan Restaurant	0.03
9	Ethiopian Restaurant	0.03

In the last step, we used machine learning, to cluster the districts into similar ones. The map shows three clusters, which are categorized by the following attributes:

- Cluster 0 - **RED**
  - Typical (average) Spanish restaurants without focus on special food.
- Cluster 1 - **BLUE**
  - Mediterranean restaurants near the water
- Cluster 2 - **GREEN**
  - Tapas restaurants and finger food, as well as vegetarian restaurants



5 Discussion section and conclusion

The most interesting part of this analysis was the clustering of the districts into the clusters [Red, Blue, Green]. We can see, that on average the Red cluster is in the main city of Barcelona and restaurants of this sector typically just sell food and are not specialized in a certain area. On the other hand, we can observe, that the Blue Cluster is on average nearer to the water as the red or the green sector. This is typical for Mediterranean restaurants, since one of the main ingredients for the dishes is fish, which they want to get fresh out of the water. Our last cluster is near to Santa Coloma de Gremenet and already a few hundred meters away from the water. In this cluster the restaurants typically sell finger food and tapas.

Since we are interested in opening a Mediterranean restaurant, we will communicate to the stakeholders, that we recommend opening a restaurant near the water in the blue cluster. Furthermore, we saw in the results section, that two districts really stand out in terms of the number of Mediterranean restaurants: Sant's Montjuic and Les Corts. Looking at this kind of data, we would communicate to our stakeholders, that these two districts are probably exceptional for opening a Mediterranean restaurant.

In conclusion we can see that machine learning and clustering has helped us get deep information from locational data. In the future we will probably see a lot of decisions or at least recommendations being made by artificial intelligence. Python and the sklearn library helped us with this problem. In reality, a lot of data is not taken into consideration yet and there is a lot of room to improve the decision-making process. We could improve the model by taking more cities near the cost of Spain into account and cluster the data again. For our example, we will recommend the blue cluster to open a restaurant, so hopefully our stakeholders and our cook can finally gain the second Michelin star.