Table of contents

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 calan region, Barcelna 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 Moskau, Paris, London and Madrid are larger european ciities 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 really 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 successfull 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 finally really nail the place they ulitmately open up 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 didnt avail a second michelin-star, they are short on money. Due to this fact, they hired and instructed the soon to be data analyst/scientest Hendrik (cheap labour) to find the perfect place to open the restaurant in and eventually gain another global michelin star.

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 most amount of restaurants related to mediterranean dishes. Addidtionally, after importing the regional data of barcelona into a dataframe, we will use the foursquare API to generate a link, which will support us with a json file which contains informations about all the restaurants and venues of the corresponding region of barcelona. After extracting the data, we have to clean it up and drop some columns and rows, to ultimately only use the necessary data.

After preparing and modelling the data, we will analyze and visualize the data. For the stakeholders it would be really interesting what kind of restaurants in the region they are competing against. If we come to the conclustion, that some regions are full of fastfood, fingerfood 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 surrouding area? Is there a lot of competetion or rather not? With the datapoints we will get from the foursquare API we will find the key to succuess in the corresponding area. After having gained enough informations about the districts of barcelona, we will eventually publish the report to the stakeholders. It will be up to them to decide, wheter 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 informations they need.

Methodology section

The first and most important thing i performed was to import all possible libaries i intented to use, so we can use libraries like the pandas or numpy. Furthermore we imported libaries 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 some kind of 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 have to prepare and model the data. In the first steps we created a pandas dataframe out of the district data and dropped some unnecessary colums and rows. Afterwards we used one of the previously mentioned libraries, to add the latitude and longitude- data to the dataframe. 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 fo 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 really 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 amount 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 on 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 fastfood, 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 no 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 to many categories and only want to get a generel overview.

Results

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 can not 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 resutaurants was plotted using the seaborn library
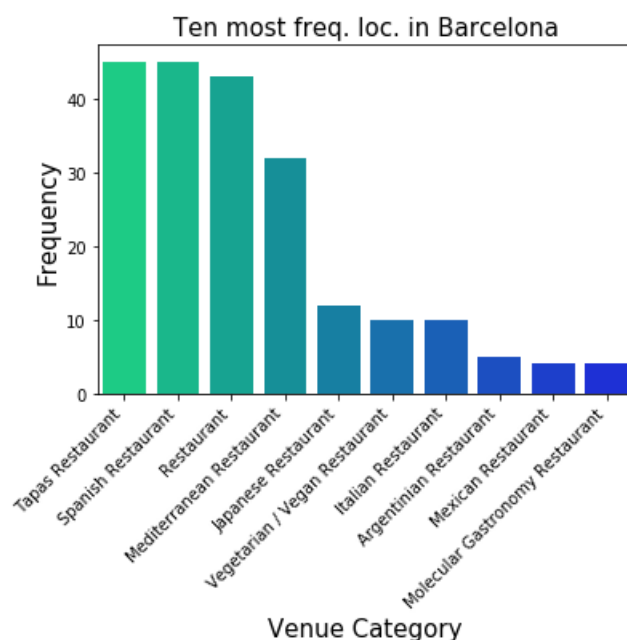
After plotting this, we wrote some code to see how much percent the corresponding categoriy of a restaurant is above the mean of the district. Since we are only interested in the mediterranean restaurants, we see two districts which really stand out in comparison to the other districts: Sants-Montjuic and Les Corts:

```
----Sants-Montjuïc----                          ----Les Corts----
                        venue  freq                              venue  freq
0                   Restaurant  0.22    0                    Restaurant  0.19
1        Mediterranean Restaurant  0.19    1        Mediterranean Restaurant  0.16
2              Spanish Restaurant  0.15    2             Japanese Restaurant  0.16
3              Chinese Restaurant  0.07    3              Spanish Restaurant  0.11
4  Molecular Gastronomy Restaurant  0.07    4                Tapas Restaurant  0.08
5                Tapas Restaurant  0.07    5                 Thai Restaurant  0.05
6          Shabu-Shabu Restaurant  0.04    6              African Restaurant  0.03
7                Asian Restaurant  0.04    7                Asian Restaurant  0.03
8         Comfort Food Restaurant  0.04    8  Vegetarian / Vegan Restaurant  0.03
9           Fast Food Restaurant  0.04    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 fingerfood, aswell as vegetarian restaurants
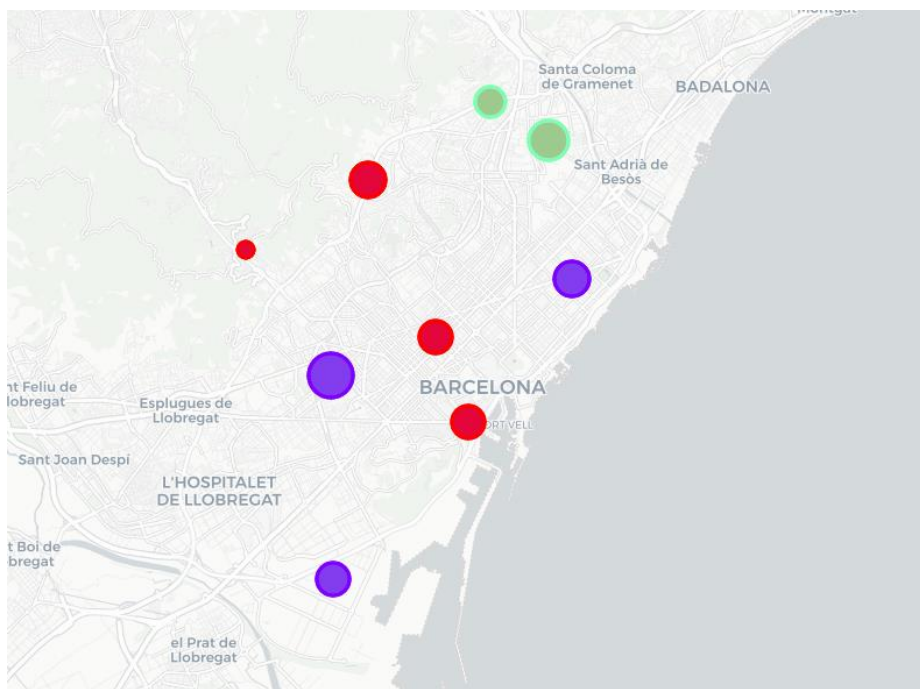
Discussion section and conclusion

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 observate, that the Blue Cluster is on average nearer to the water as the red or the green sector. This is typical for mediteranean restaurants, since one of the main ingredients for the dishes is fish, which they want to get really fresh out of the water. Our last cluster is near to Santa Coloma de Gremenet and already a few hundret meters away from the water. In this cluster the restaurants typically sell fingerfood and tapas.

Since we are interested in opening up a mediteranean restaurant, we will communicate to the stakeholders, that we recommend to open up a restaurant near the the water in the blue cluster. Furthermore we saw in the results section, that two districs really stand out in terms of the number of mediterannean restaurants: Sants Montjuic and Les Corts. Looking at this kind of data, we would communicate to our stakeholders, that these two districts are probaly exceptional for opening up a mediterranean restautrant.

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 atrificial 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 procress. We could improve the model by taking more cities near the cost of spain into account and cluster the data again. For our exmaple, we will recommend the blue cluster to open up a restaurant, so hopefully our stakeholders and our cook can finally gain the second michelin star.