Capstone Project

The optimal location for great coffee in Lisbon

Applied Data Science Capstone by IBM/Coursera

Report created by Cláudia Morgado

12 September 2020

CONTENTS:

1. INTRODUCTION: BUSINESS PROBLEM

This report describes the analyses performed with the objective to find the optimal location for a coffee shop in Lisbon, Portugal.

Some notes about Lisbon:

Lisbon is not a highly populated city, in the 2019 census the number of inhabitants was approximately 500.000; the city is divided in 24 neighbourhoods occupying an area of about 80 square kilometres. In recent years Lisbon has been visited by an increasing number of tourists - about 5 million guests/year - making it a city where new businesses are opening on a daily basis. Choosing a place to start a new one can be difficult because there is a lot of competition and renting a place in or near the city centre can be quite expensive. Lisbon has a large and rich coffee culture so opening a new *'Café'* will always be a good idea because for locals going out for coffee in the morning, lunch and evening is somewhat of a routine; for tourists it is usually the chosen place to take a break from sightseeing.

The questions:

The objective for this project will be to find a location still close to the city centre where the surroundings are not already crowded with other *'Cafés'*. In order to achieve this there are some questions we need the answer for:

- How many businesses already exist in each neighbourhood?

- What type of businesses are there?

- What are the most popular businesses in each neighbourhood? Is the *'Café'* one of them?

- What neighbourhood have the most population density? (new businesses need clients)

- What neighbourhoods have the least number of *'Cafés'*?

2. DATA

In order to obtain the results we were looking for we needed to collect data from the following data sources:

2.1 Location information for the 24 Lisbon neighbourhoods that can be sourced from local government agencies, the one used for this project will be downloaded from here:
   - http://mapas.dgterritorio.pt/ATOM-download/CAOP-Cont/Cont_AAD_CAOP2019.zip

2.2 Population and neighbourhood density for all neighbourhoods collected here: http://mercadodebemfica.blogspot.com/2011/01/as-novas-freguesias-de-lisboa.html

2.3 Number of coffee shops and their location in every neighborhood will be obtained using Foursquare API

3. METHODOLOGY

3.1 The data collected from Lisbon's 24 neighbourhoods:

- Since the data was in a geographic format (SHP) in polygons I a GIS software – *Quantum GIS* – was used to calculate the coordinates for the centre of each neighbourhood (latitude and longitude);
- The polygon data in geographic format (SHP) was exported in the same GIS software – *Quantum GIS* – to a GeoJSON format to be used in the Clorophet Map;

3.2 The data collected for population density was a table containing only the *Name*, *Population* and *Area(km2)* for each neighbourhood. The population density was calculated in the notebook;

3.3 *Numpy* and *Pandas* were used to process all the data;

3.4 The *Foursquare API* was used to explore the different venues in each neighbourhood;

3.5 The *folium* package was used to map the different neighbourhoods as well as to map the clusters and the population density;

3.6 The clustering was performed using the *K-Means* algorithm;

4. ANALISYS

4.1 We started with two data frames, one for the location data of the neighbourhoods and another for the population data. To this last one a new column was added to insert the calculation for the population density:

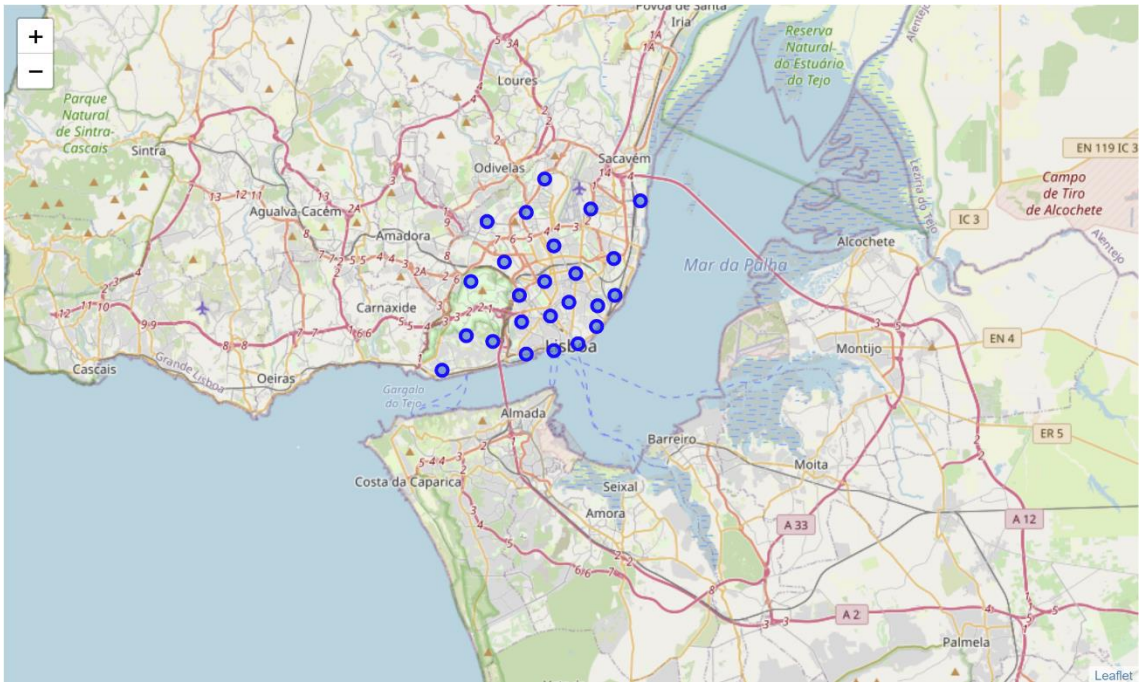
	Neighbourhood	Latitude	Longitude
0	Ajuda	38.712174	-9.198653
1	Alcantara	38.709594	-9.183390
2	Alvalade	38.753884	-9.146498
3	Areeiro	38.741376	-9.133527
4	Arroios	38.727653	-9.137942

Location data frame

	Neighbourhood	Population	Area	Pop_Dens(pop/km2)
0	Ajuda	17958	2.85	6301.052632
1	Alcantara	14443	4.42	3267.647059
2	Alvalade	34217	5.33	6419.699812
3	Areeiro	21035	1.76	11951.704545
4	Arroios	33210	2.15	15446.511628

Population density data frame

4.2 We then map the neighbourhoods to see the location of the different neighbourhoods



Map of Lisbon's 24 neighbourhoods

4.3 Using the *Foursquare API* we get the venues from each neighbourhood and can check the number of venues in each neighbourhood. We can see that there are only a few neighbourhoods with less than 50 venues in the list: *Ajuda*, *Beato*, *Benfica*, *Marvila*, and *Santa Clara*. We can also check the number of venues in each category and we can see *'Café'* is the second venue to appear in the list, so it is a popular category.

Venue Category	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude
Portuguese Restaurant	206	206	206	206	206	206
Café	110	110	110	110	110	110
Restaurant	108	108	108	108	108	108
Hotel	87	87	87	87	87	87
Bakery	65	65	65	65	65	65
Coffee Shop	61	61	61	61	61	61
Plaza	48	48	48	48	48	48
Italian Restaurant	40	40	40	40	40	40
Park	40	40	40	40	40	40
Seafood Restaurant	36	36	36	36	36	36

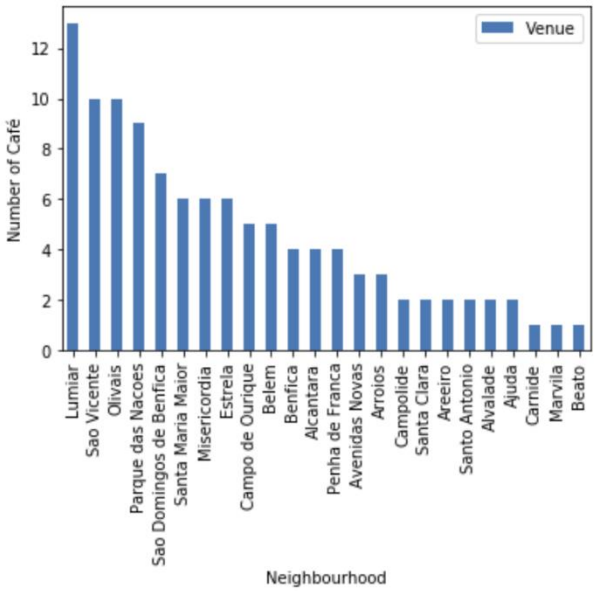
Number of venues in each category

4.4 Then we can check if the *'Café'* is on the Top 10 categories for each of the neighbourhoods. We can see from the list that the category *'Restaurant'* or *'Portuguese Restaurant'* is the number one category for most neighbourhoods.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ajuda	Restaurant	Portuguese Restaurant	Park	Coffee Shop	Soccer Stadium	BBQ Joint	Gym	Café	Supermarket	Soccer Field
1	Alcantara	Restaurant	Portuguese Restaurant	Café	Mediterranean Restaurant	Seafood Restaurant	BBQ Joint	Coffee Shop	Plaza	Museum	Bakery
2	Alvalade	Portuguese Restaurant	Restaurant	Bar	Coffee Shop	Bakery	Burger Joint	Indian Restaurant	Italian Restaurant	Ice Cream Shop	Bookstore
3	Areeiro	Portuguese Restaurant	Hotel	Italian Restaurant	Bakery	Gym / Fitness Center	Vegetarian / Vegan Restaurant	Restaurant	Sushi Restaurant	Ice Cream Shop	Indian Restaurant
4	Arroios	Hotel	Portuguese Restaurant	Plaza	Seafood Restaurant	Cocktail Bar	Breakfast Spot	Scenic Lookout	Café	Restaurant	Park
5	Avenidas Novas	Portuguese Restaurant	Hotel	Italian Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Bakery	Burger Joint	Japanese Restaurant	Steakhouse	Coffee Shop
6	Beato	Restaurant	Brewery	Theater	Coffee Shop	Plaza	Paper / Office Supplies Store	Supermarket	BBQ Joint	Climbing Gym	Music Venue
7	Belem	Portuguese Restaurant	Café	Monument / Landmark	Ice Cream Shop	Garden	Sushi Restaurant	Bakery	Restaurant	Art Museum	Hotel
8	Benfica	Portuguese Restaurant	Restaurant	Café	Bakery	Seafood Restaurant	Park	Pizza Place	BBQ Joint	Burger Joint	Grocery Store
9	Campo de Ourique	Portuguese Restaurant	Hotel	Coffee Shop	Italian Restaurant	Café	Bakery	Restaurant	Garden	Ice Cream Shop	Indian Restaurant
10	Campolide	Hotel	Portuguese Restaurant	Restaurant	Park	Gym	Italian Restaurant	Seafood Restaurant	Garden	Japanese Restaurant	Art Museum
11	Carnide	Portuguese Restaurant	Restaurant	Burger Joint	Sporting Goods Shop	Soccer Stadium	Supermarket	Indian Restaurant	Theater	Soccer Field	Snack Place
12	Estrela	Portuguese Restaurant	Coffee Shop	Café	Restaurant	Seafood Restaurant	Bakery	Italian Restaurant	Ice Cream Shop	Steakhouse	Breakfast Spot
13	Lumiar	Café	Bakery	Portuguese Restaurant	Restaurant	Supermarket	Tapas Restaurant	Chinese Restaurant	History Museum	Park	Pizza Place
14	Marvila	Restaurant	Hotel	Portuguese Restaurant	Gym / Fitness Center	Brewery	Art Gallery	Plaza	Café	Fast Food Restaurant	Farmers Market
15	Misericordia	Portuguese Restaurant	Plaza	Café	Hostel	Tapas Restaurant	Wine Bar	Coffee Shop	Vegetarian / Vegan Restaurant	Italian Restaurant	Theater
				Portuguese	Airport				Fast Food	Rental Car	

Top 10 categories for each neighbourhood

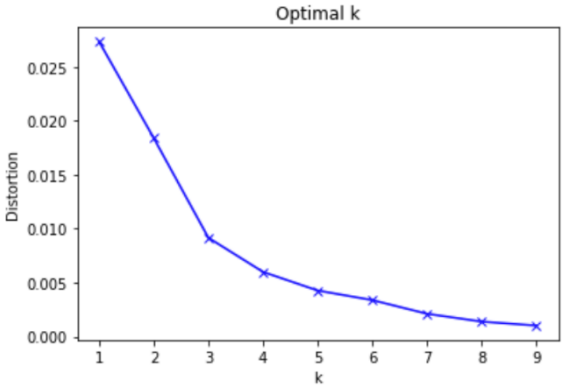
4.5 After analysing the data from all the venues we isolate only the category *'Café'* and see the number of *'Cafés'* that already exist in each neighbourhood. From the data frame and the plotted result we can see that the highest number of *'Cafés'* exists in *'Lumiar'* with ***13*** *'Cafés'* and that there are only two more neighbourhood with a number of *'Cafés'* higher or equal to ***10*** *'Olivais'* and *'São Vicente'*.



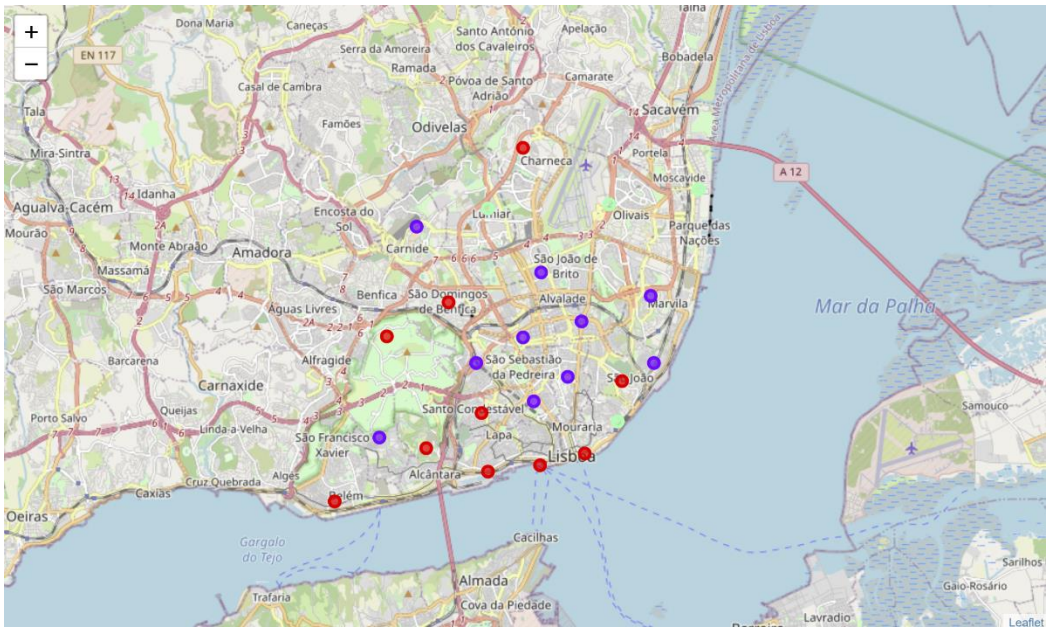
Ajuda	2
Alcantara	4
Alvalade	2
Areeiro	2
Arroios	3
Avenidas Novas	3
Beato	1
Belem	5
Benfica	4
Campo de Ourique	5
Campolide	2
Carnide	1
Estrela	6
Lumiar	13
Marvila	1
Misericordia	6
Olivais	10
Parque das Nacoes	9
Penha de Franca	4
Santa Clara	2
Santa Maria Maior	6
Santo Antonio	2
Sao Domingos de Benfica	7
Sao Vicente	10

Number of *'Cafés'* in each neighbourhood

4.6 Then we use the K-Means algorithm to cluster the *'Cafés'* in groups. Before that the number of clusters to create (k) is calculated using the Elbow Method. In the plot we can see the best choice for the number of clusters is ***3***. Then we can map the clusters.



Plotting for the best choice for ***k***

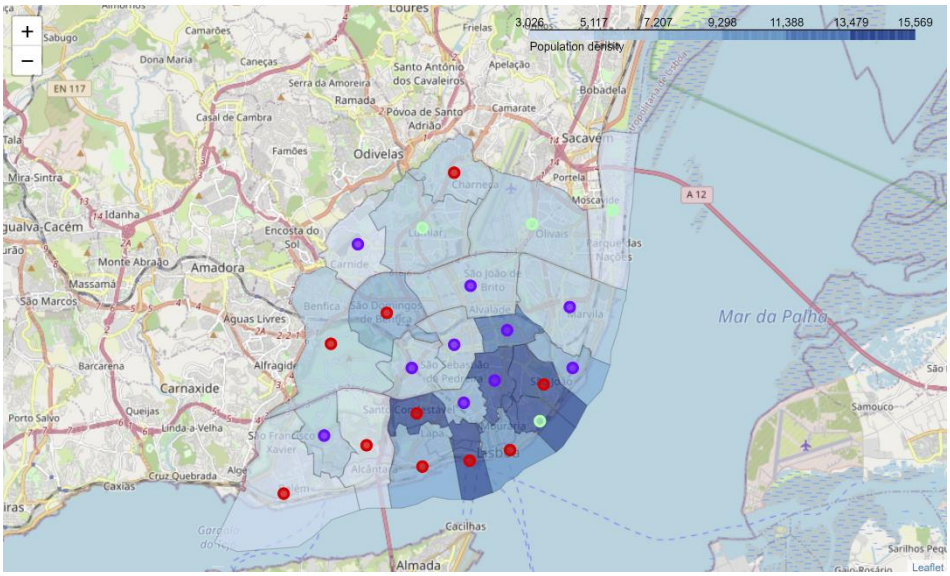


Mapping the clusters

4.7 The map shows the three clusters: Cluster 0 (in red), Cluster 1 (in violet) and Cluster 2 (in green). The three groups can be described as:

- Cluster 0: neighbourhoods with few *'Cafés'*
- Cluster 1: neighbourhoods with an average number of *'Cafés'*
- Cluster 3: neighbourhoods with the highest number of *'Cafés'*

4.8 By merging the population density data frame with the cluster data frame by neighbourhood we can then analyse the different clusters taking into account the population density of each neighbourhood and then map the result.



Mapping the clusters with the population density

5. RESULTS

From the results reached we can ascertain:

- *'Cafés'* are among the top 3 venues found in almost every neighbourhood in Lisbon though rarely the in the first place;

- The neighbourhoods with the highest population density are in Lisbon's city centre (high density = more clients) which could suggest a higher rent price;

- Although more densely populated the city centre does not have the neighbourhoods with the highest number of *'Cafés'*;

- The highest number of *'Cafés'* we can find in Lisbon's (present in Cluster 2) neighbourhoods are mostly in low density areas (*Olivais*, *Parque das Nações*, *Lumiar*) where *São Vicente* is the exception with 10 *'Cafés'*.

- There are a large number of neighbourhoods with few *'Cafés'* (*Alvalade*, *Areeiro*, *Santo Antonio*, *Carnide*, *Marvila*) that could be a good location for a new *'Café'*.

6. CONCLUSION

By looking at the achieved results we can say that **Campo de Ourique** and **Misericórdia** neighbourhoods could be the most likely place to open a *'Café'*; they are both in high densely populated areas and have a low number of *'Cafés'*. However, since the data here analysed comes from different sources and is dated from different years (population density is from 2012) the results may not be the most up to date. The analysis performed could benefit from more recent data regarding the population.

Another relevant data that could be added to this analysis is data from the renting prices for commercial spaces (average rent price/m2) and some mobility data regarding the existence of public transportation and parking spaces could be useful to see if the business could attract clients from outside the neighbourhood.