SEGMENTING AND CLUSTERING LISBON'S PARISHES

RICARDO ENES

1. Discussion and Background

When you hear of Europe's hot destinations, Portugal is surely at the top of the charts. At the center of this beautiful country you find its capital, Lisbon. The world is full of wonderful and exciting cities to visit, and Lisbon is one that should be on everyone's bucket list. It has everything within easy reach, from galleries and museums to the seaside and charming nearby villages. Lisbon is also extremely budget-friendly, making it easy to navigate and explore in depth.

Soon I will be leaving my hometown, Braga, in the north of Portugal to start my adult life in Lisbon, working on 9 to 6 job. I am that kind of person that needs to go out and socialize on a daily basis, in order to maintain its sanity. This is where we have a problem... I know nothing about Lisbon's hot spots, due to only have visited the city three times. As such, I am "combining business with pleasure": IBM's data science capstone project with the urge to know more about Lisbon's most cultural and coolest neighborhoods.

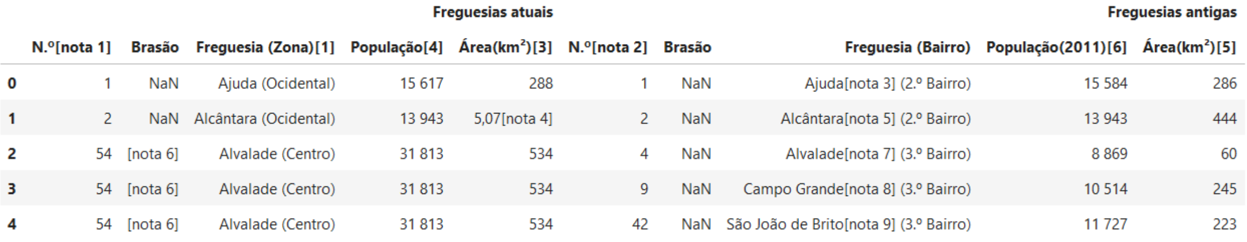
This project aims to nullify these urges by segmenting and clustering Lisbon's major metropolitan area, in terms of social and cultural venues. With this in mind, I will gather data by the means of web scraping and Foursquare's API.

This endeavour will mostly be useful to myself, but it would be very rewarding if by any chance (close to none), common people like myself could make use of this data and get a bigger picture about what Lisbon has to offer and where to find it.

2. Data Gathering and Methodology

The first step is acquiring Lisbon's metropolitan area neighborhoods names and clean the data. The website from where the neighborhoods names were collected using BeaufifulSoup library was,

- `https://pt.wikipedia.org/wiki/Lista_de_freguesias_de_Lisboa`

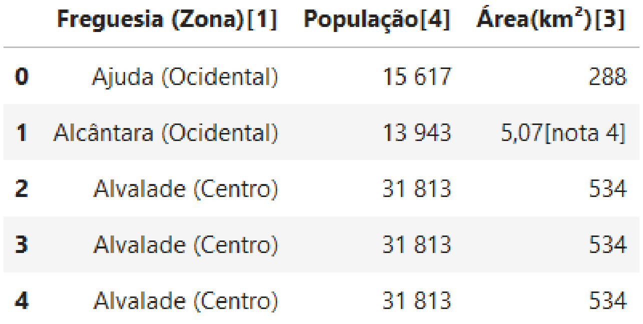


	Freguesias atuais					Freguesias antigas				
	N.º[nota 1]	Brasão	Freguesia (Zona)[1]	População[4]	Área(km^2)[3]	N.º[nota 2]	Brasão	Freguesia (Bairro)	População(2011)[6]	Área(km^2)[5]
0	1	NaN	Ajuda (Ocidental)	15 617	288	1	NaN	Ajuda[nota 3] (2.º Bairro)	15 584	286
1	2	NaN	Alcântara (Ocidental)	13 943	5,07[nota 4]	2	NaN	Alcântara[nota 5] (2.º Bairro)	13 943	444
2	54	[nota 6]	Alvalade (Centro)	31 813	534	4	NaN	Alvalade[nota 7] (3.º Bairro)	8 869	60
3	54	[nota 6]	Alvalade (Centro)	31 813	534	9	NaN	Campo Grande[nota 8] (3.º Bairro)	10 514	245
4	54	[nota 6]	Alvalade (Centro)	31 813	534	42	NaN	São João de Brito[nota 9] (3.º Bairro)	11 727	223

Figura 1. First data acquired.

Some of the features were filtered, more precisely, the columns ('N.º[nota 1]','Brasao','N.º[nota 2]', 'Populaçao(2011)[6]', 'Area(km^2)[5]','Freguesia (Bairro)'), observed in figure 1 were dropped. These columns were ignored because the number of Lisbon's parishes was reduced in 2012, from 53 to 24, hence only the new aggregated parishes were considered.

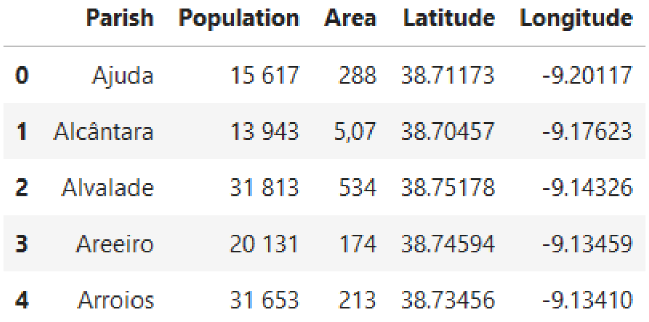
It is noticeable in figure 2, that several parishes have duplicates, also that some of the data is associated with a given note, "[nota x]"and finally some of the parishes have some information between parenthesis. All of these unwanted values were removed, since they would interfere with proper data analysis and ultimately with the collection of coordinates using geocoding.



	Freguesia (Zona)[1]	População[4]	Área(km^2)[3]
0	Ajuda (Ocidental)	15 617	288
1	Alcântara (Ocidental)	13 943	5,07[nota 4]
2	Alvalade (Centro)	31 813	534
3	Alvalade (Centro)	31 813	534
4	Alvalade (Centro)	31 813	534

Figura 2. Removal of unwanted columns.

First, the duplicated parishes were dropped, keeping only the first appearance of a given parish. To efficiently remove the ["nota x"] a function named "clean_area"was created. After the data was properly cleaned, it was possible to acquire the coordinates of all Lisbon's parishes using Geolocator form the GeoPY library. Finnaly, a dataframe including all of the information gathered so far was created, being the latter depicted in figure 3



	Parish	Population	Area	Latitude	Longitude
0	Ajuda	15 617	288	38.71173	-9.20117
1	Alcântara	13 943	5,07	38.70457	-9.17623
2	Alvalade	31 813	534	38.75178	-9.14326
3	Areeiro	20 131	174	38.74594	-9.13459
4	Arroios	31 653	213	38.73456	-9.13410

Figura 3. Data cleaned.

Having finished the basic data gathering and cleaning, the python folium library was used to visualize the if the parish's coordinates were rightly obtained. In figures 4 and 5 we can see that some of the coordinates of the parishes belonging the metropolitan area of Lisbon were wrongly collected.

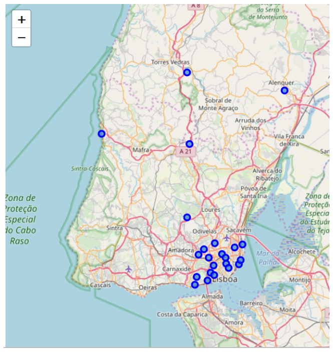


Figura 4. Lisbon Metropolitan area

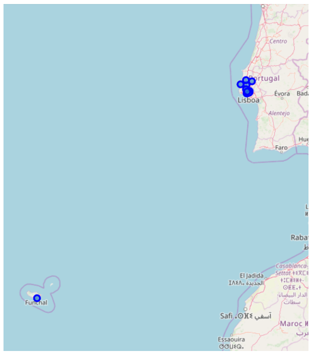


Figura 5. Portugal and Açores

Figura 6. Geographical view of the collected parish coordinates

To circumvent this problem the proper coordinates of these six parishes were collected in their wikipedia pages, and a function called "chang_coord"was created to modify the dataframe.

- `https://pt.wikipedia.org/wiki/Santo_Antonio_(Lisboa)`
- `https://pt.wikipedia.org/wiki/Misericórdia_(Lisboa)`
- `https://pt.wikipedia.org/wiki/Avenidas_Novas_(Lisboa)`
- `https://pt.wikipedia.org/wiki/Santa_Clara_(Lisboa)`
- `https://pt.wikipedia.org/wiki/Sao_Vicente_(Lisboa)`
- `https://pt.wikipedia.org/wiki/Santa_Maria_Maior_(Lisboa)`

In figure 7 we can observ that the problem was solved.

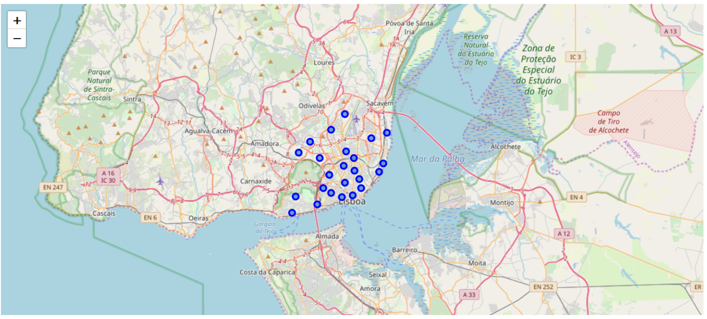
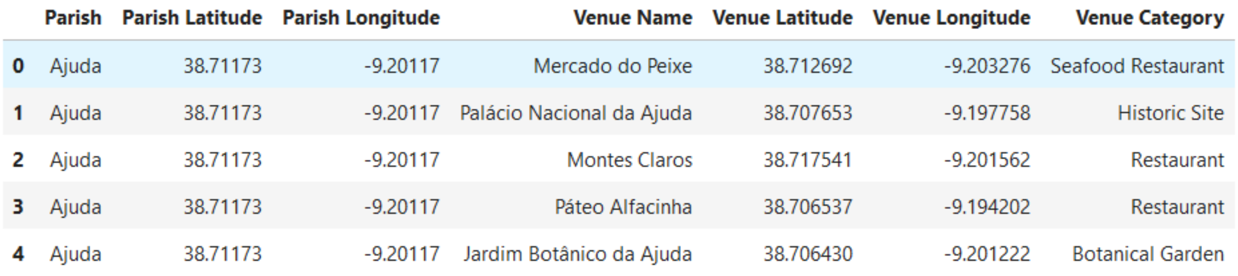


Figura 7. Parishes coordinates corrected.

We then utilized the Foursquare's API to query about the venues of each one of Lisbon's parishes, within a 1000 meters radius and 100 venues per parish. All of the data gathered was aggregated in a single dataframe (figure 8) including the following columns:

(1) 'Parish'
(2) 'Parish Latitude'
(3) 'Parish Longitude'
(4) 'Venue Name'
(5) 'Venue Latitude'
(6) 'Venue Longitude'
(7) 'Venue Category'



	Parish	Parish Latitude	Parish Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Ajuda	38.71173	-9.20117	Mercado do Peixe	38.712692	-9.203276	Seafood Restaurant
1	Ajuda	38.71173	-9.20117	Palácio Nacional da Ajuda	38.707653	-9.197758	Historic Site
2	Ajuda	38.71173	-9.20117	Montes Claros	38.717541	-9.201562	Restaurant
3	Ajuda	38.71173	-9.20117	Páteo Alfacinha	38.706537	-9.194202	Restaurant
4	Ajuda	38.71173	-9.20117	Jardim Botânico da Ajuda	38.706430	-9.201222	Botanical Garden

Figura 8. Data acquired using Forusquare's API.

The total number of unique venues collected is 212, and the number of venues per parish is depicted in figure 1 below,

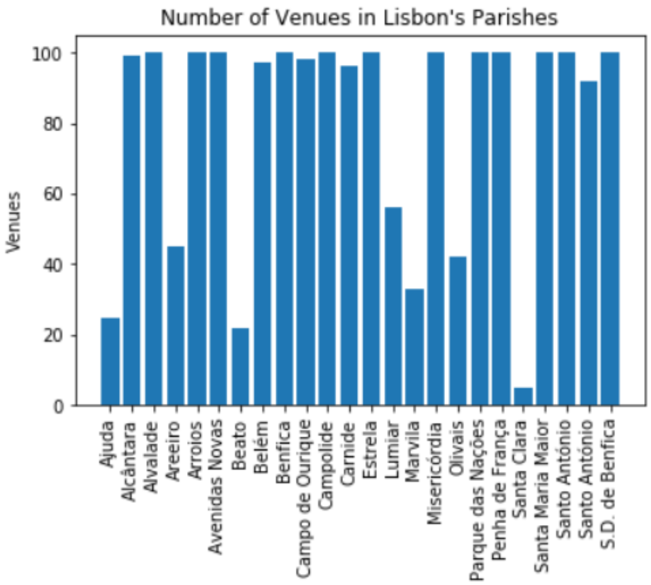


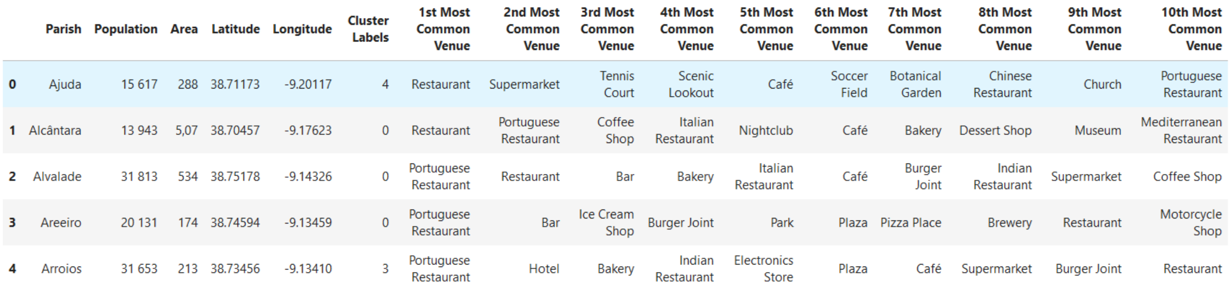
Figura 9. Bar graph depicting the number of venues per parish.

Almost half of the considered parishes reached the 100 venue cap, being those,

(1) Alvalade;
(2) Arroios;
(3) Avenidas Novas;
(4) Benfica;
(5) Campolide;
(6) Estrela
(7) Misericórdia;
(8) Parque das Nações;
(9) Penha de França;
(10) Santa Maria Maior;
(11) Santo António;
(12) São Vicente.

and the perish with the least number of venues is Santa Clara, with 5.

Being the ultimate objective the cluster and segmentation of Lisbon's parishes, a dataframe containing the top-10 venues of each parish was created (figure 10).



	Parish	Population	Area	Latitude	Longitude	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ajuda	15 617	288	38.71173	-9.20117	4	Restaurant	Supermarket	Tennis Court	Scenic Lookout	Café	Soccer Field	Botanical Garden	Chinese Restaurant	Church	Portuguese Restaurant
1	Alcântara	13 943	5,07	38.70457	-9.17623	0	Restaurant	Portuguese Restaurant	Coffee Shop	Italian Restaurant	Nightclub	Café	Bakery	Dessert Shop	Museum	Mediterranean Restaurant
2	Alvalade	31 813	534	38.75178	-9.14326	0	Portuguese Restaurant	Restaurant	Bar	Bakery	Italian Restaurant	Café	Burger Joint	Indian Restaurant	Supermarket	Coffee Shop
3	Areeiro	20 131	174	38.74594	-9.13459	0	Portuguese Restaurant	Bar	Ice Cream Shop	Burger Joint	Park	Plaza	Pizza Place	Brewery	Restaurant	Motorcycle Shop
4	Arroios	31 653	213	38.73456	-9.13410	3	Portuguese Restaurant	Hotel	Bakery	Indian Restaurant	Electronics Store	Plaza	Café	Supermarket	Burger Joint	Restaurant

Figura 10. Most common venues per parish.

For proper clustering of the parishes we utilized the unsupervised learning, K-means algorithm. First the optimal number of clusters was found, by the means of the the elbow method (figure 11). The optimal value found was 5 clusters.

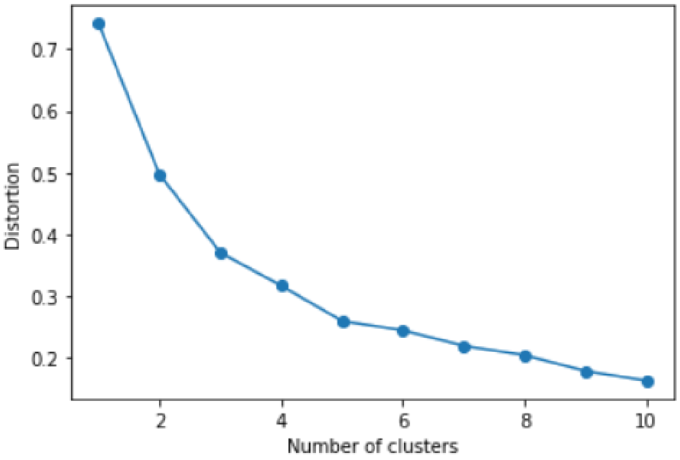


Figura 11. Elbow method - acquiring the optimal value for the number of clusters.

3. Results and Discussion

In figure 12, there are depicted the 5 clusters, obtained given the venues similarities between the different parishes.

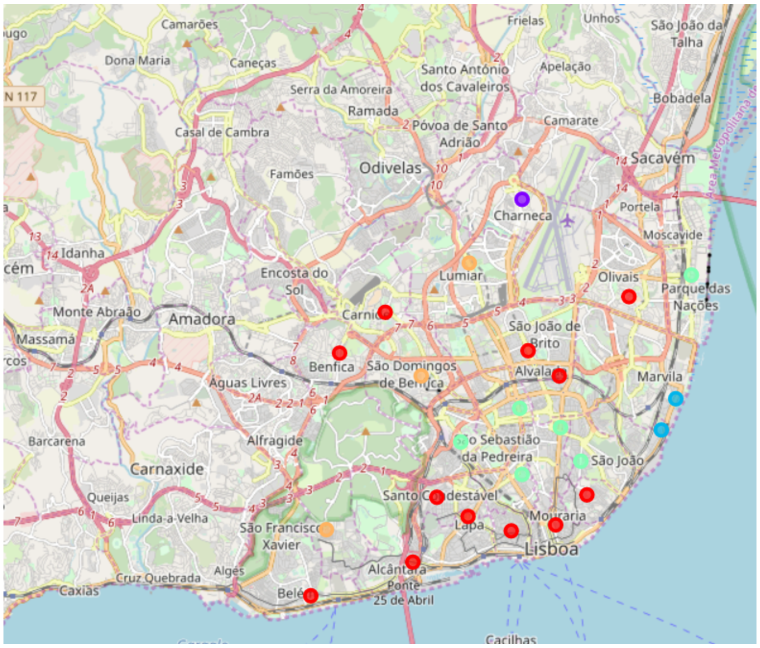
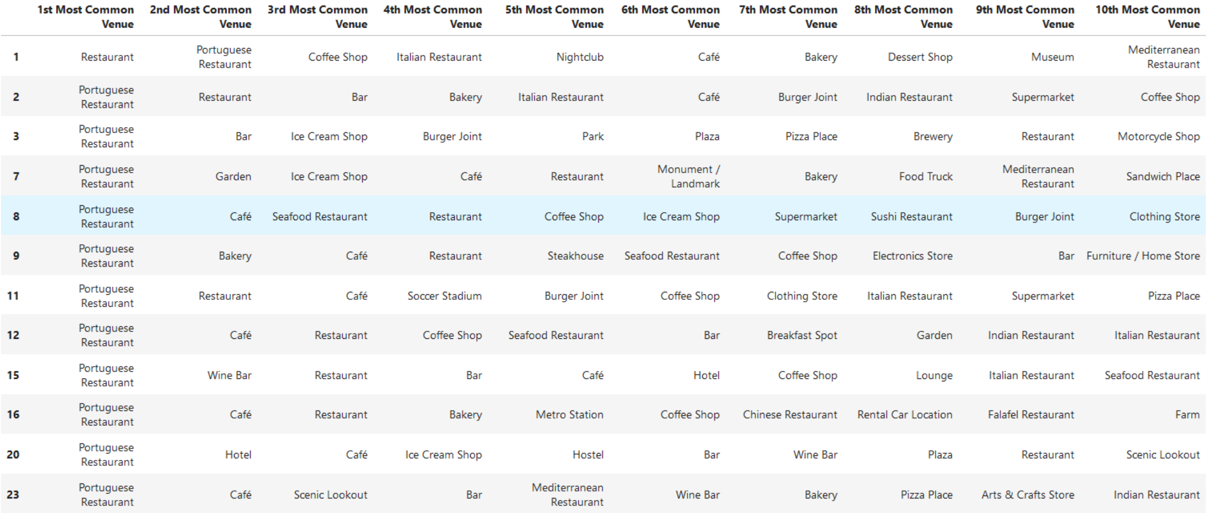


Figura 12. Lisbon's clusters.

As it was mentioned before, the objective of this project is to determine the most social and cultural parishes in the city of Lisbon. The most common venues for on the obtained clusters are:

- Cluster 0 (**red**)

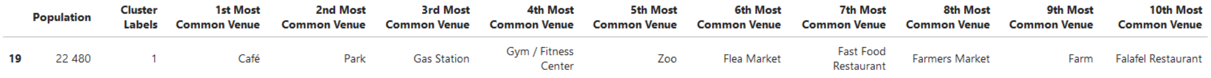


	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Restaurant	Portuguese Restaurant	Coffee Shop	Italian Restaurant	Nightclub	Café	Bakery	Dessert Shop	Museum	Mediterranean Restaurant
2	Portuguese Restaurant	Restaurant	Bar	Bakery	Italian Restaurant	Café	Burger Joint	Indian Restaurant	Supermarket	Coffee Shop
3	Portuguese Restaurant	Bar	Ice Cream Shop	Burger Joint	Park	Plaza	Pizza Place	Brewery	Restaurant	Motorcycle Shop
7	Portuguese Restaurant	Garden	Ice Cream Shop	Café	Restaurant	Monument / Landmark	Bakery	Food Truck	Mediterranean Restaurant	Sandwich Place
8	Portuguese Restaurant	Café	Seafood Restaurant	Restaurant	Coffee Shop	Ice Cream Shop	Supermarket	Sushi Restaurant	Burger Joint	Clothing Store
9	Portuguese Restaurant	Bakery	Café	Restaurant	Steakhouse	Seafood Restaurant	Coffee Shop	Electronics Store	Bar	Furniture / Home Store
11	Portuguese Restaurant	Restaurant	Café	Soccer Stadium	Burger Joint	Coffee Shop	Clothing Store	Italian Restaurant	Supermarket	Pizza Place
12	Portuguese Restaurant	Café	Restaurant	Coffee Shop	Seafood Restaurant	Bar	Breakfast Spot	Garden	Indian Restaurant	Italian Restaurant
15	Portuguese Restaurant	Wine Bar	Restaurant	Bar	Café	Hotel	Coffee Shop	Lounge	Italian Restaurant	Seafood Restaurant
16	Portuguese Restaurant	Café	Restaurant	Bakery	Metro Station	Coffee Shop	Chinese Restaurant	Rental Car Location	Falafel Restaurant	Farm
20	Portuguese Restaurant	Hotel	Café	Ice Cream Shop	Hostel	Bar	Wine Bar	Plaza	Restaurant	Scenic Lookout
23	Portuguese Restaurant	Café	Scenic Lookout	Bar	Mediterranean Restaurant	Wine Bar	Bakery	Pizza Place	Arts & Crafts Store	Indian Restaurant

Figura 13. Red cluster most common venues.

The red cluster is the cluster with the most social venues, since it presents the highest number of food related venues, bars and cafés. The parishes belonging to this cluster are the ones one should pick to enjoy a quick lunch, a nice diner or just some socliazing time in a café or a bar.

- Cluster 1 (**purple**)

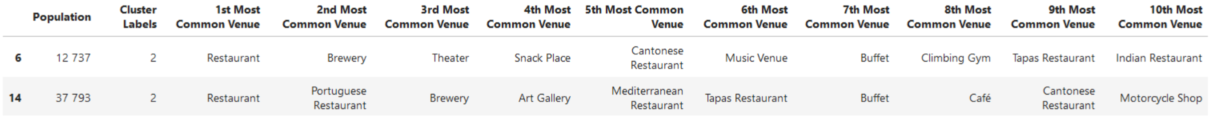


	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
19	22 480	1	Café	Park	Gas Station	Gym / Fitness Center	Zoo	Flea Market	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant

Figura 14. Purple cluster most common venues.

The purple cluster, due to its vicinity to an airport and its distance to Lisbon's center is the cluster that presents a more mixed variety of venues. Since it is far from Lisbon's hot spots, I would only consider this parish if I was staying in town for 1 or 2 days, while staying at an hotel near the airport.

- Cluster 2 (**light blue**)

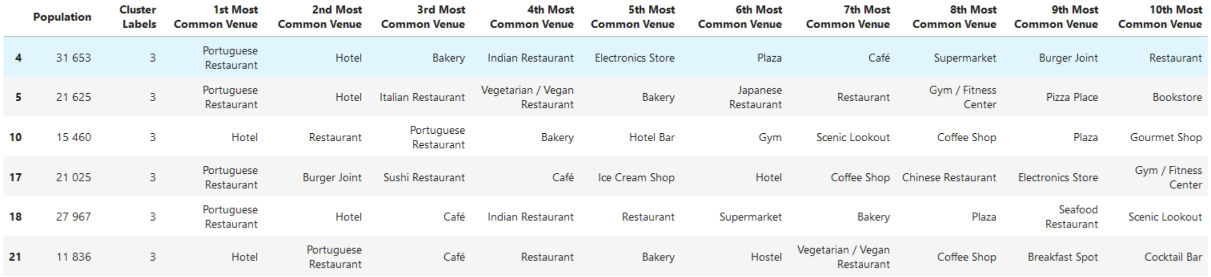


	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
6	12 737	2	Restaurant	Brewery	Theater	Snack Place	Cantonese Restaurant	Music Venue	Buffet	Climbing Gym	Tapas Restaurant	Indian Restaurant
14	37 793	2	Restaurant	Portuguese Restaurant	Brewery	Art Gallery	Mediterranean Restaurant	Tapas Restaurant	Buffet	Café	Cantonese Restaurant	Motorcycle Shop

Figura 15. Light blue cluster most common venues.

The light blue cluster seems like a suitable residential, since if presents a good variety of venues and at the same time is close to Tejo's river bank.

- Cluster 3 (**light green**)

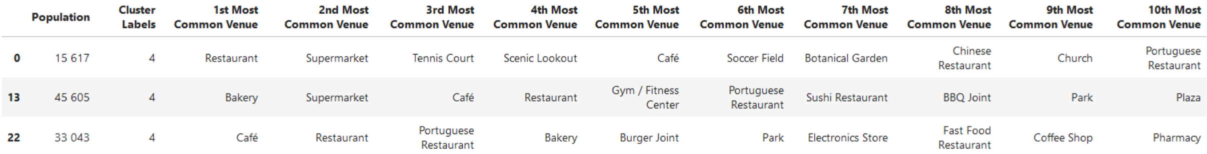


	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	31 653	3	Portuguese Restaurant	Hotel	Bakery	Indian Restaurant	Electronics Store	Plaza	Café	Supermarket	Burger Joint	Restaurant
5	21 625	3	Portuguese Restaurant	Hotel	Italian Restaurant	Vegetarian / Vegan Restaurant	Bakery	Japanese Restaurant	Restaurant	Gym / Fitness Center	Pizza Place	Bookstore
10	15 460	3	Hotel	Restaurant	Portuguese Restaurant	Bakery	Hotel Bar	Gym	Scenic Lookout	Coffee Shop	Plaza	Gourmet Shop
17	21 025	3	Portuguese Restaurant	Burger Joint	Sushi Restaurant	Café	Ice Cream Shop	Hotel	Coffee Shop	Chinese Restaurant	Electronics Store	Gym / Fitness Center
18	27 967	3	Portuguese Restaurant	Hotel	Café	Indian Restaurant	Restaurant	Supermarket	Bakery	Plaza	Seafood Restaurant	Scenic Lookout
21	11 836	3	Hotel	Portuguese Restaurant	Café	Restaurant	Bakery	Hostel	Vegetarian / Vegan Restaurant	Coffee Shop	Breakfast Spot	Cocktail Bar

Figura 16. Light green cluster most common venues.

The light green cluster encompasses 5 of the most central parishes in Lisbon, and it is a good cluster to find a high number of hotels and a rich gastronomy variety, displaying cuisine from around the globe.

- Cluster 4 (**orange**)



	Population	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	15 617	4	Restaurant	Supermarket	Tennis Court	Scenic Lookout	Café	Soccer Field	Botanical Garden	Chinese Restaurant	Church	Portuguese Restaurant
13	45 605	4	Bakery	Supermarket	Café	Restaurant	Gym / Fitness Center	Portuguese Restaurant	Sushi Restaurant	BBQ Joint	Park	Plaza
22	33 043	4	Café	Restaurant	Portuguese Restaurant	Bakery	Burger Joint	Park	Electronics Store	Fast Food Restaurant	Coffee Shop	Pharmacy

Figura 17. Orange cluster most common venues.

The orange cluster seems to have the best from all the other clusters with the addition of a considerable number of supermarkets, making it a good place to live. It is reasanable to assume this, since this cluster presents the highest level of population from the bunch.

4. Conclusion

Finally to conclude this project, we have gotten a small glimpse of how real life data-science projects look like. I have made use of some frequently used python libraries to scrap web-data, visualize and analyse my data, use Foursquare API to explore the major districts of Lisbon and saw the results of segmentation using Folium library. Potential for this kind of analysis in a real life business problem is discussed in great detail. Finally, since my analysis was mostly concentrated on the possibility to know a bit more about what Lisbon has to offer and where to find it. Specially cafes, bars, pubs, restaurants. Hopefully, you will find this project useful and I am delighted to get to know a bit more about the world of data science and expect to take more real-life challenges using data-science.

May your hopes be as pure and strong as your perseverance,

Ricardo