W2PT – Welcome to Portugal!

By Filipe Coutinho | 21st of June 2020

Part of the IBM Data Science Professional Certificate Capstone Project

Notebook available here | Github Link

---

1. Introduction

The Business Problem

Tourism is a big deal in Portugal. Elected as the best European destination for three years in a row (2017 to 2019), it is a great place to visit and to stay.

But where should you go?

This project intends to leverage Foursquare ratings on accommodation for the most popular locations of Portugal to clearly rate and classify the best destinations and help to guide you and other tourists towards the most preferred places.

On the other hand, the rating of different locations is especially relevant for new hotel businesses or tourism agencies, allowing them to invest on the tourism hotspots and focus on improving the touristic experience where it is currently less attractive.

2. Data Sources

Our main source of data will be *Foursquare* (Places API). The Foursquare Places API is used by more than 150.000 developers and makes available over 62 million venues throughout the whole world, therefore providing a high degree of confidence in our research results. Data will be extracted regarding the main accommodation venues and associated ratings for the different district capitals in Portugal:

1. Aveiro
2. Beja
3. Braga
4. Bragança
5. Castelo Branco
6. Coimbra
7. Évora
8. Faro
9. Guarda
10. Leiria
11. Lisboa
12. Portalegre
13. Porto
14. Santarém
15. Setúbal
16. Viana do Castelo
17. Vila Real
18. Viseu

This will give us a good overview of the most preferred regions and the most popular. Other variables like number of comments and photos can also serve as a measure of popularity. We will associate these popularity metrics to the coordinates of each district capital.

We will then **rank the different locations and cluster them in 5 groups**, based on their popularity and location. **The output will be a clear recommendation of where you should place your bet for your next vacation!**

3. Methodology

Testing Run

As described in the notebook we started by importing data from Foursquare, more specifically by using the Places API.

In a first stage we've collected data from 30 venues in a single city (Porto) and tested the algorithm to append the number of likes to each of the recommended venues.

Extract from the resulting dataframe can be seen below:

	venue.id	venue.location.address	venue.location.city	venue.location.lat	venue.location.lng	venue.name	likes
0	4cbc54f07a5d9eb0ed5b31e9	R. Tenente Valadim 146	Porto	41.161064	-8.640411	Sheraton Porto Hotel & Spa	265
1	4bcc5b3aaeaaeee151ec3d6d	R. Guedes de Azevedo, 179	Porto	41.152183	-8.607009	Hotel Dom Henrique	73
2	4e15cb9bc65b14b6ca369fa4	Praça Da Liberdade, 25	Porto	41.145869	-8.611540	InterContinental	111
3	4b9572c0f964a52025a334e3	Av. Boavista, 1269	Porto	41.159408	-8.638681	Hotel Porto Palácio	123
4	4da69f026e81162ae782263e	R. Maria Feliciana, 100	Matosinhos	41.187602	-8.597501	Axis Porto Business & SPA Hotel	59

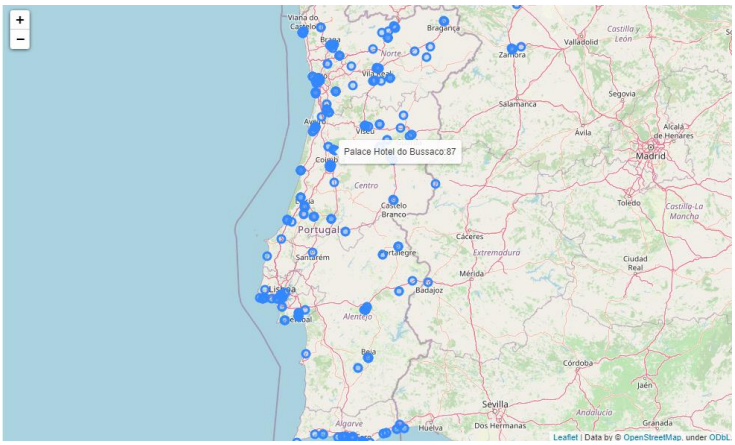
Full-scale Algorithm Application

After testing the algorithm with a single city sample we've expanded the methodology to the whole list of district capitals, achieving a dataset with 499 venues, classified with their respective likes.

After the results were exported to a CSV for safety, we can start visualizing the results.

Visualization & Clustering

In order to retrieve insights from the data, we've used *Folium* as the library for interactively visualizing the results on the map. Initial iteration displayed the venues, without any classification or clustering, including popups for each marker with the venue name and classification (number of likes).

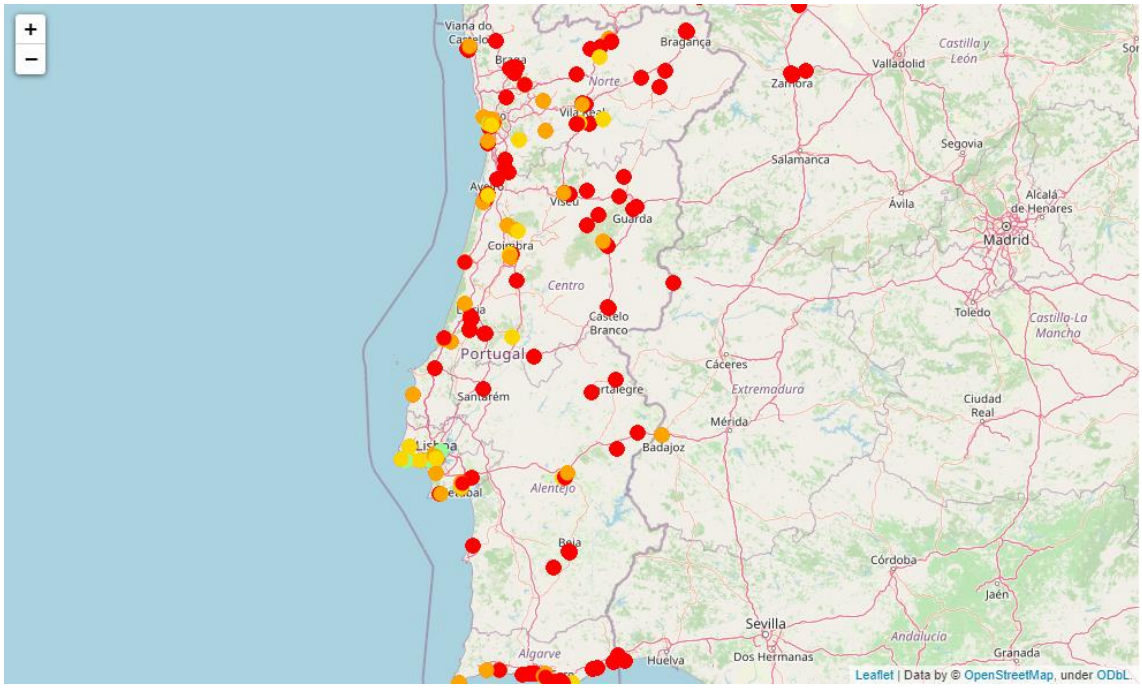


In order to facilitate clustering of the different locations, we will apply the K-Means clustering algorithm. We've opted for 5 groups (clusters), based on location and ranking (number of likes). Averaged results of the clustering can be seen below:

	venue.location.lat	venue.location.lng	likes
cluster			
0.0	39.931064	-8.076390	13.495238
1.0	40.436021	-8.517542	78.460000
2.0	40.118389	-8.859974	263.500000
3.0	39.854827	-8.587103	43.868056
4.0	39.660465	-8.985743	122.838710

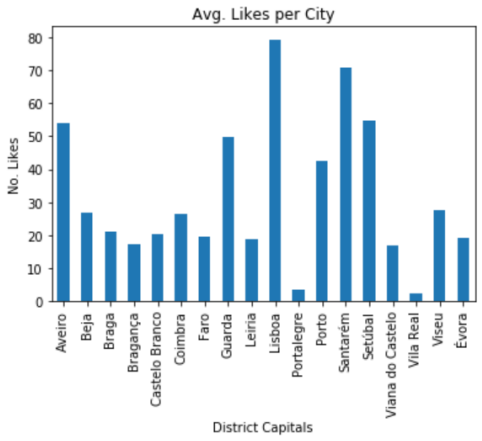
4. Results

After clustering the results, we've attributed a color to each of the clusters, based on the average of likes (from red to green (worst to best), being cluster 0 the worst and cluster 2 the best – as seen in the table above). Final map can be seen below:

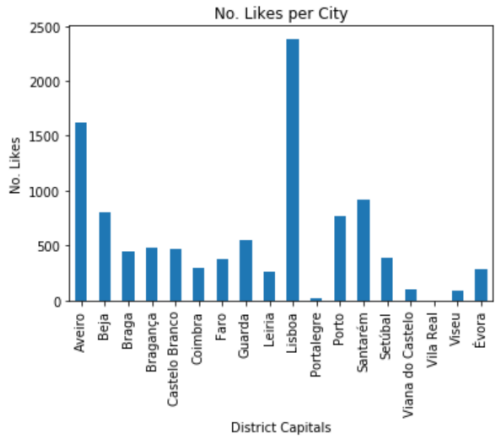


5. Discussion

We can critically analyze the distribution of interest venues per distrinct capital:



venue.location.city	likes
Lisboa	79.533333
Santarém	71.000000
Setúbal	54.714286
Aveiro	54.100000
Guarda	49.818182



venue.location.city	likes
Lisboa	2386
Aveiro	1623
Santarém	923
Beja	800
Porto	763

From the above data we can understand that the most appealing city would be Lisbon, with an average rating of 79.5. In any case, big touristing hotspots like Algarve or Porto are not so well represented here, which may imply that insufficient data was supplied from the Foursquare API. We should consider that, as this information is community-driven, we may have missing ratings for less-visited locations (which may not be an indicator of higher or lower quality.

Also, if we look at the top 5 venues, we can notice some discrepancies:

	likes	color	venue.id	venue.location.address	venue.location.city	venue.name
203	267	limegreen	4b83208af964a52041f930e3	Rua Latino Coelho 1	Lisboa	Sheraton Lisboa Hotel & Spa
3	265	limegreen	4cbc54f07a5d9eb0ed5b31e9	R. Tenente Valadim 146	Aveiro	Sheraton Porto Hotel & Spa
207	256	limegreen	51226c72e4b002b6fa144b52	Av. Engenheiro Duarte Pacheco, 5	Lisboa	EPIC SANA Lisboa Hotel
202	165	palegreen	4bc34d7c74a9a5933753d4f6	Av. da Liberdade, 185	Lisboa	Hotel Tivoli Avenida Liberdade Lisboa
238	164	palegreen	4b76ea5af964a520236a2ee3	Av. dos Aliados, 85-89	Porto	Café Guarany

The second contender (Sheraton Porto) is assumed as located in Aveiro. This is because it came up in, at least, two different queries and registered as being part of Aveiro (the last query), given its proximity. In order to increase the quality of these results we can reduce the radius of the search using the Foursquare API, thus limiting the venues to their own cities (although possibly reducing the sample size).

On the other hand, the fifth contender (Café Guarany) is not a hotel. Although we've specifically narrowed our query to the venue id's that were classified as hotels on Foursquare, some of the results are incorrectly classified. This is because of the community classification of venues on Foursquare which may be prone to error sometimes.

6. Conclusion

We were able to import a selection of venues (hotels) from Foursquare, for each of the capital districts of Portugal and their nearby locations. With the Folium map representation, we are able to visualize these locations and quickly access their scoring (number of likes). With the use of K-Means algorithm we automatically clustered locations and differentiated them based on likes (red being the worst and green the best).

Visually we can conclude that Porto and Lisbon stand out with high-ranking places. We also have a wide variety of choices to the south (Algarve) and the north (around Vila Real and Braga), although with fewer number of likes. However, if we deep-dive on the data and analyze the venue rankings, some discrepancies start to come up, mostly due to the quality of the exported data from Foursquare (misclassification of venues) and the broad radius of research.

Although very subjective, this first approach enables us to discover new touristic places in Portugal. Make sure you pay us a visit and hopefully this will come in handy when checking out your next hotel!

Thank you very much for visiting this notebook and feel free to get in touch with any questions via: fil.coutinho@gmail.com

Filipe Coutinho
IBM Professional Certificate Capstone Project