BATTLE OF THE CITIES

A Data Science project

André Barbosa
https://github.com/Barbosabyte/battle-of-the-cities

Title: Battle of the Cities

Author: André Barbosa

Year: 2021

Source: https://github.com/Barbosabyte/battle-of-the-cities

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Content

Introduction

Bustopher Jones is a gentleman who wishes to live in Portugal for a while and asked for my help to choose the best city to live

As Mr. Jones is extremely wealthy, he is not concerned with real estate prices but has some peculiar requirements about the location:

- Mr. Jones is afraid of the consequences of global warming, so the city must not be by the sea;
- He will buy a house by the city centre and will walk everywhere so he can flaunt his designer clothes, so the city must have various restaurants, bars, etc. in a 1km radius of his house;
- It must be a well-known city, as Mr. Jones doesn't like to explain where he lives.

Disclaimer: This work was done during the Covid-19 pandemic; thus, some information may be inaccurate.

Data description

To be able to help Mr. Jones choose his dream city I used data from Foursquare (using their API).

To get the coordinates for the Portuguese cities I used a dataset from simplemaps. I used the free version available at https://simplemaps.com/data/pt-cities.

Methodology

In order to help Mr. Jones choose the best city to live I used a dataset with the coordinates of several Portuguese cities and towns (see the previous section) and read it into a pandas data frame.

	city	lat	lng	country	iso2	admin_name	capital	population	population_proper
0	Lisbon	38.7452	-9.1604	Portugal	PT	Lisboa	primary	506654.0	506654.0
1	Vila Nova de Gaia	41.1333	-8.6167	Portugal	PT	Porto	minor	302295.0	302295.0
2	Porto	41.1495	-8.6108	Portugal	PT	Porto	admin	237591.0	237591.0
3	Braga	41.5333	-8.4167	Portugal	PT	Braga	admin	181494.0	181494.0
4	Matosinhos	41.2077	-8.6674	Portugal	PT	Porto	minor	175478.0	175478.0

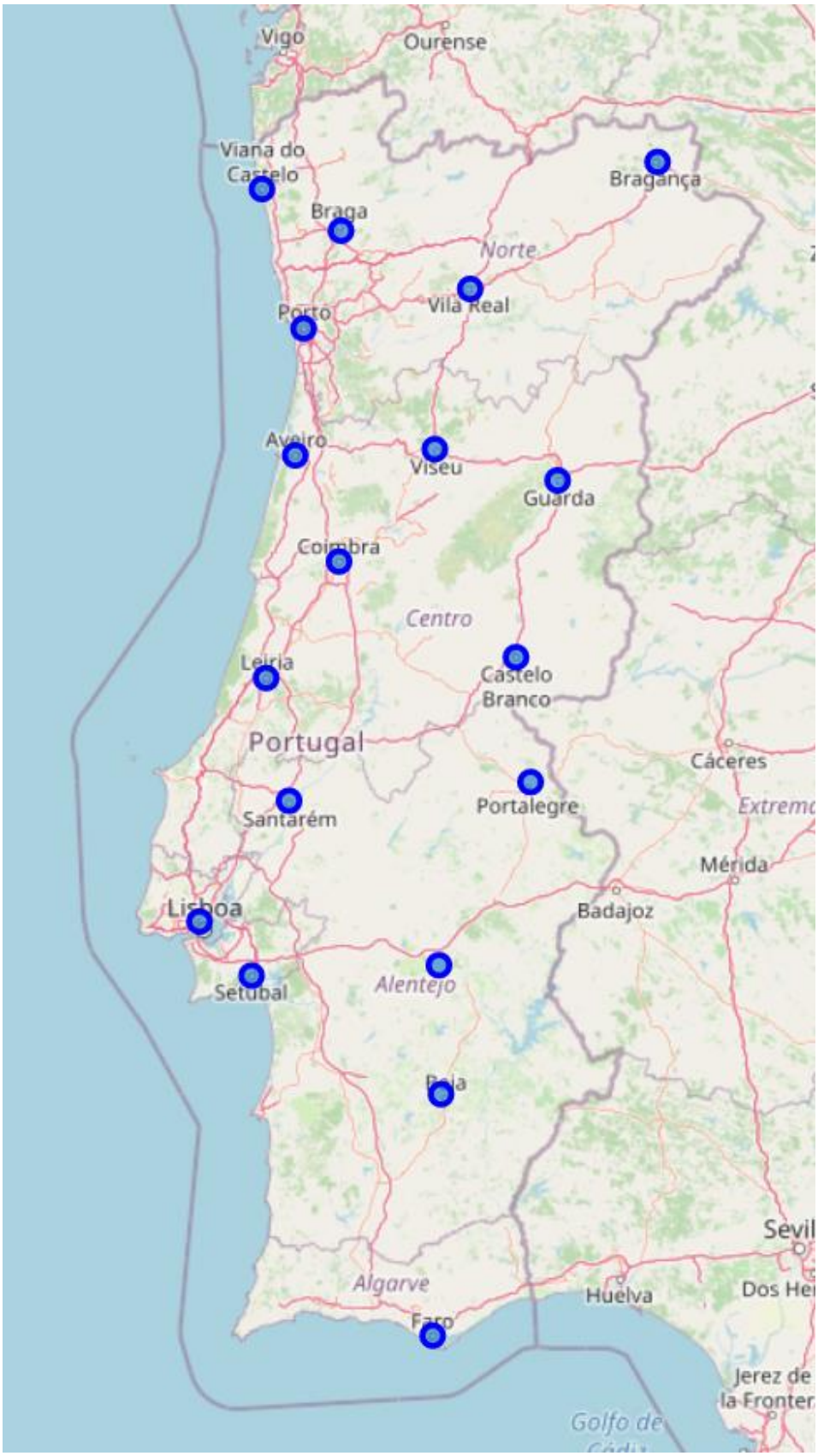
Table 1 - Raw data frame head.

In order to clean the data, I excluded all the lesser known cities, leaving only district capitals (marked as “primary” and “admin” at the capital column), excluded the cities on the two island regions (marked as “Madeira” and “Azores” at the admin_name column) as they are obviously near the sea, and removed the unneeded columns (country, iso2, population, population_proper, admin_name and capital).

	city	lat	lng
0	Lisbon	38.7452	-9.1604
1	Porto	41.1495	-8.6108
2	Braga	41.5333	-8.4167
3	Coimbra	40.2111	-8.4291
4	Leiria	39.7431	-8.8069
5	Setúbal	38.5243	-8.8926
6	Viseu	40.6667	-7.9167
7	Viana do Castelo	41.7000	-8.8333
8	Aveiro	40.6389	-8.6553
9	Faro	37.0167	-7.9333
10	Santarém	39.2369	-8.6850
11	Castelo Branco	39.8230	-7.4931
12	Évora	38.5667	-7.9000
13	Vila Real	41.3002	-7.7398
14	Guarda	40.5364	-7.2683
15	Beja	38.0333	-7.8833
16	Bragança	41.8000	-6.7500
17	Portalegre	39.3167	-7.4167

Table 2 - Clean data frame with the 18 continental district capitals.

To better visualize the cities location on the country and their proximity to the sea, I generated a folium map with markers on the cities' locations.



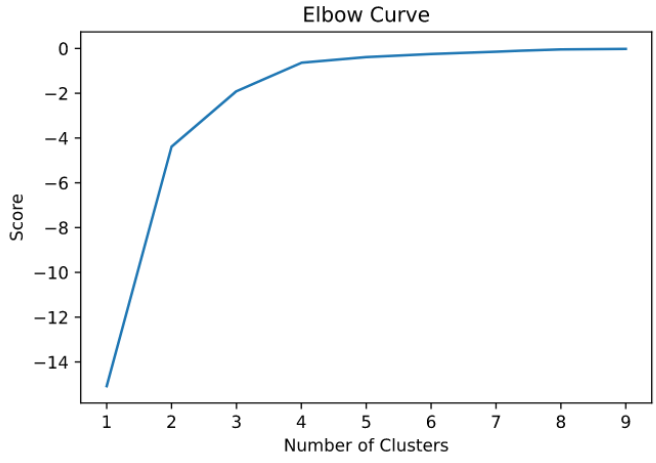
Map 1 - Map of Portugal with markers on the 18 cities of our data frame.

As expected, some of the cities are near the sea, namely Viana do Castelo, Porto, Aveiro, Lisbon, Setúbal and Faro and can be removed from our data frame, leaving 12 cities that satisfy the distance to sea and well-known requisites.

	city	lat	lng
0	Braga	41.5333	-8.4167
1	Coimbra	40.2111	-8.4291
2	Leiria	39.7431	-8.8069
3	Viseu	40.6667	-7.9167
4	Santarém	39.2369	-8.6850
5	Castelo Branco	39.8230	-7.4931
6	Évora	38.5667	-7.9000
7	Vila Real	41.3002	-7.7398
8	Guarda	40.5364	-7.2683
9	Beja	38.0333	-7.8833
10	Bragança	41.8000	-6.7500
11	Portalegre	39.3167	-7.4167

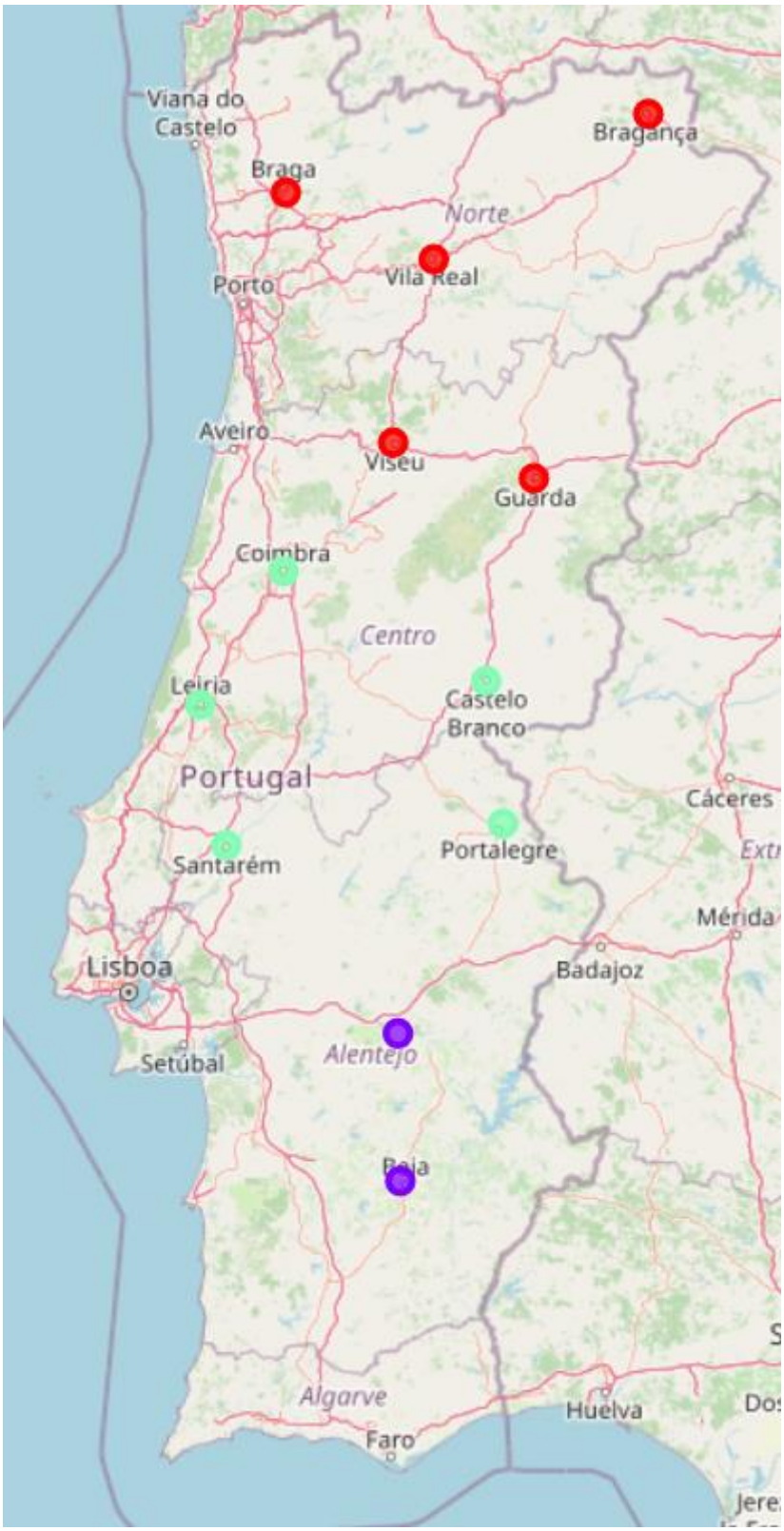
Table 3 - The data frame with the remaining cities

In a country with the size of Portugal, comparing these 12 cities is almost the same as comparing the whole country. For simplicity we will divide the data in different regions and choose the region with the more suitable candidates. For this we will use k-means clustering and plot an elbow curve to find the optimal number of clusters to divide the country.



Graph 1 - Elbow curve.

According to the graph, 3 is the best number of clusters, so I divided the country in 3 clusters and represented them on a folium map.



Map 2 - Map of Portugal with the cluster representation

The clusters mainly followed the traditional regions of North, Centre and South, so it's easier to work.

Next, I fetched data from the Foursquare API, to get all the venues in a 1Km radius of the city centre, performed a One Hot Encoding of the venue categories and grouped them by city.

	Cluster Labels	city	lat	lng	Art Gallery	Art Museum	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	...	Supermarket	Sushi Restaurant	Tapas Restaurant	Tea Room	Theater	Theme Park Ride / Attraction	Train Station	Vegetarian / Vegan Restaurant	Wine Bar	Winery
0	0	Braga	41.5333	-8.4167	0.000000	0.000000	0.040000	0.04	0.000000	0.080000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000
1	2	Coimbra	40.2111	-8.4291	0.010000	0.000000	0.000000	0.00	0.000000	0.060000	...	0.010000	0.000000	0.010000	0.010000	0.000000	0.000000	0.010000	0.01	0.000000	0.000000
2	2	Leiria	39.7431	-8.8069	0.000000	0.000000	0.000000	0.00	0.000000	0.025974	...	0.038961	0.012987	0.064935	0.000000	0.000000	0.012987	0.000000	0.00	0.012987	0.000000
3	0	Viseu	40.6667	-7.9167	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.039216	0.000000	0.000000	0.000000	0.019608	0.000000	0.000000	0.00	0.039216	0.000000
4	2	Santarém	39.2369	-8.6850	0.000000	0.000000	0.033333	0.00	0.000000	0.000000	...	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000
5	2	Castelo Branco	39.8230	-7.4931	0.000000	0.000000	0.000000	0.00	0.000000	0.047619	...	0.000000	0.000000	0.023810	0.000000	0.000000	0.000000	0.047619	0.00	0.000000	0.000000
6	1	Évora	38.5667	-7.9000	0.000000	0.026316	0.000000	0.00	0.000000	0.026316	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.026316	0.00	0.000000	0.000000
7	0	Vila Real	41.3002	-7.7398	0.000000	0.000000	0.000000	0.00	0.018868	0.075472	...	0.000000	0.000000	0.018868	0.018868	0.018868	0.000000	0.000000	0.00	0.000000	0.018868
8	0	Guarda	40.5364	-7.2683	0.000000	0.000000	0.000000	0.00	0.000000	0.074074	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000
9	1	Beja	38.0333	-7.8833	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000
10	0	Bragança	41.8000	-6.7500	0.058824	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000
11	2	Portalegre	39.3167	-7.4167	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000

Table 4 - Venue Categories grouped by city.

To choose the best region I grouped the venues again, this time by cluster, and found the five most common venues of each cluster.

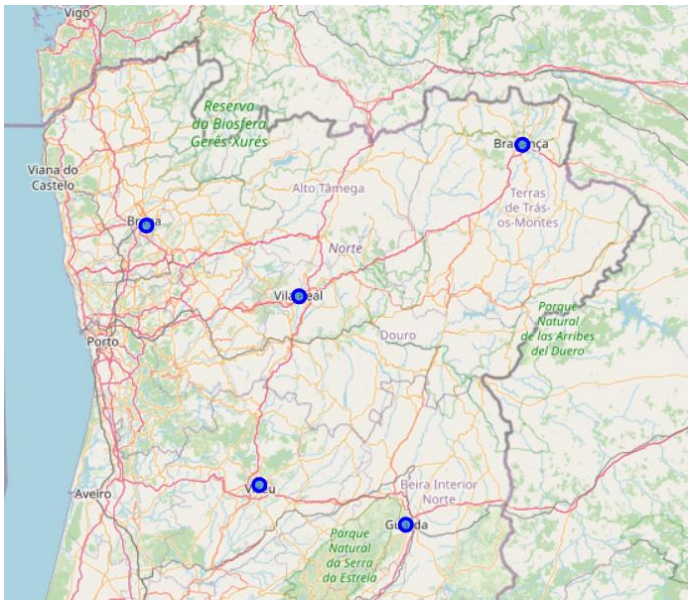
	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	0	Café	Portuguese Restaurant	Restaurant	Bar	Bakery
1	1	Restaurant	Food Truck	Portuguese Restaurant	Historic Site	Café
2	2	Portuguese Restaurant	Pool	Bar	Café	Hotel

Table 5 - The five most common venues on each cluster.

Based on this information we will choose cluster 0 since all the five most common venues are “food places” (in Portugal, most bakeries double as cafés).

	Cluster Labels	city	lat	lng	Art Gallery	Art Museum	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	...	Supermarket	Sushi Restaurant	Tapas Restaurant	Tea Room	Theater	Theme Park Ride / Attraction	Train Station	Vegetarian / Vegan Restaurant	Wine Bar	Winery
0	0	Braga	41.5333	-8.4167	0.000000	0.0	0.04	0.04	0.000000	0.080000	...	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000
1	0	Viseu	40.6667	-7.9167	0.000000	0.0	0.00	0.00	0.000000	0.000000	...	0.039216	0.0	0.000000	0.000000	0.019608	0.0	0.0	0.0	0.039216	0.000000
2	0	Vila Real	41.3002	-7.7398	0.000000	0.0	0.00	0.00	0.018868	0.075472	...	0.000000	0.0	0.018868	0.018868	0.018868	0.0	0.0	0.0	0.000000	0.018868
3	0	Guarda	40.5364	-7.2683	0.000000	0.0	0.00	0.00	0.000000	0.074074	...	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000
4	0	Bragança	41.8000	-6.7500	0.058824	0.0	0.00	0.00	0.000000	0.000000	...	0.000000	0.0	0.058824	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000

Table 6 - The five cities of cluster 0 and respective venues.



Map 3 - The five cities of the cluster

To finish our analysis, we will find the top five venues of each of our selected cities.

	city	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Braga	Café	Restaurant	Portuguese Restaurant	Bakery	Italian Restaurant
1	Viseu	Bar	Café	Fast Food Restaurant	Portuguese Restaurant	Hotel
2	Vila Real	Café	Fast Food Restaurant	Bakery	Bar	Park
3	Guarda	Café	Restaurant	Portuguese Restaurant	Bakery	Hotel
4	Bragança	Portuguese Restaurant	Bar	Plaza	Art Gallery	Café

Table 7 - The top most common venues on our cities.

With this data we can choose the best city for Mr. Jones.

Results

Based on the analysis of the data we can conclude that the best city for Mr. Jones is Braga, as the five most frequent venues are “food places”, namely, cafés, restaurants, Portuguese restaurants, bakeries and Italian restaurants.

We can also recommend as good candidates Viseu and Guarda, in case our client doesn’t like Braga.

Discussion

While the approach used on this analysis focused on simplicity and speed of analysis, a more complete approach could have focused on the clustering based on similar venue frequency, reputation of the respective venues, etc. A more detailed analysis could also have focused on overall distribution (on the map) of the venues on the city or it's absolute number.

Conclusion

Based on the analysis, we can conclude that Mr. Bustopher Jones would do good in choosing the city of Braga, being far enough from the sea, well known, and with eateries being the most common venues.

If Braga is not on the taste of Mr. Jones, he can safely choose Viseu or Guarda, also well known, even further from the sea, and with eateries being also common.