André Barbosa

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

Battle of the Cities

A Data Science project

**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 3](#_Toc64147114)

[Data description 4](#_Toc64147115)

[Methodology 5](#_Toc64147116)

[Results 11](#_Toc64147117)

[Discussion 12](#_Toc64147118)

[Conclusion 13](#_Toc64147119)

# 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.

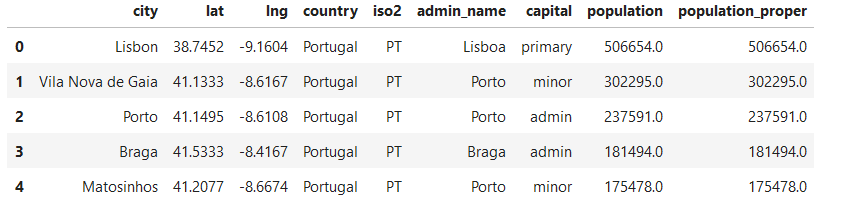


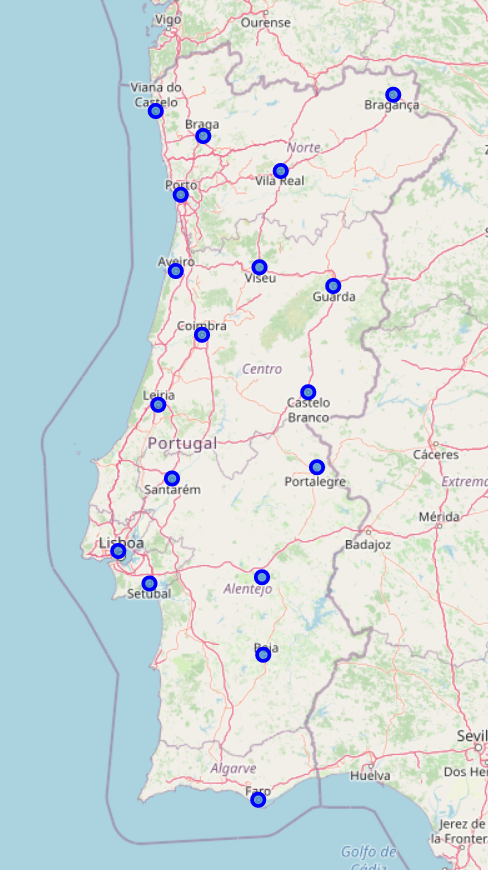
Table 1 - Raw data frame head.

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).



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.



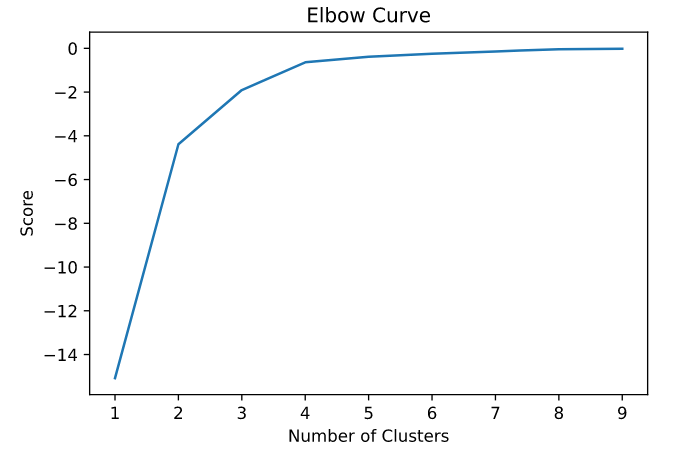
Map 1 - Map of Portugal with markers on the 18 cities of our data frame.

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.



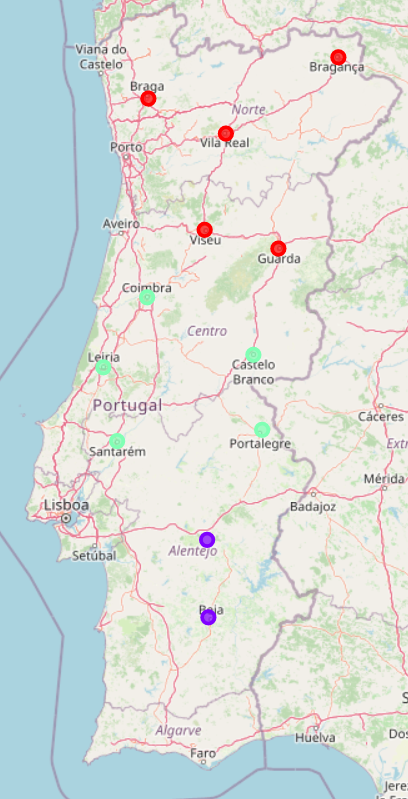
Table 3 - The data frame with the remaining cities

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.



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.

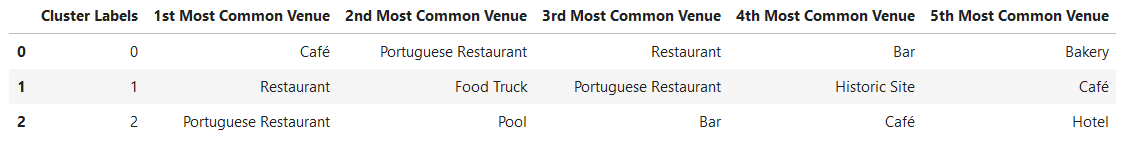


Table 5 - The five most common venues on each cluster.

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).

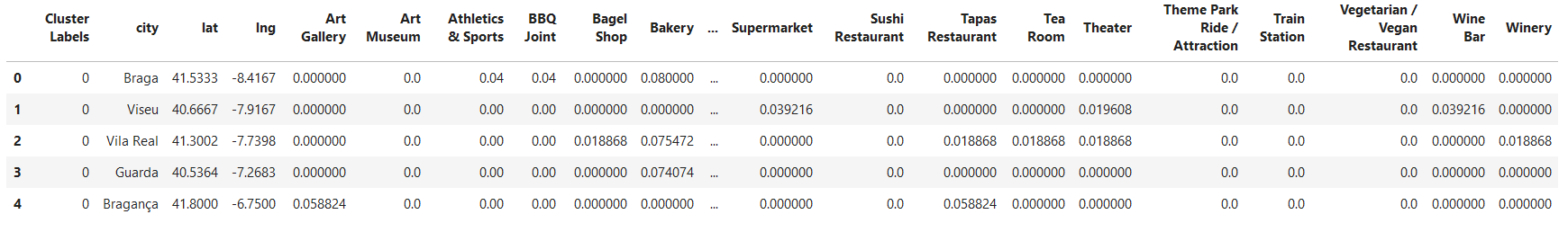
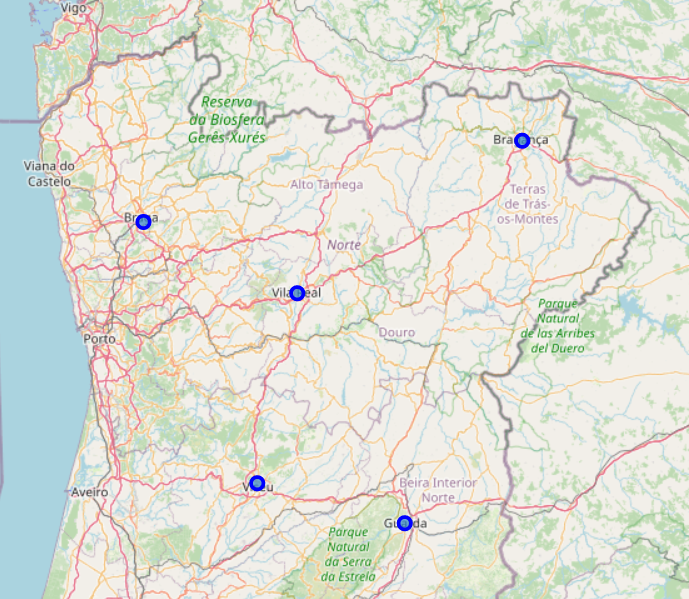


Table 6 - The five cities of cluster 0 and respective venues.



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.



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.