The Battle of the Neighborhoods

Applied Data Science Capstone

Glass Recycling Bins in Lisbon

João Romão

IBM Data Science Professional Certificate by Coursera

1. Introduction

1.2 Background

If you are ever in Portugal, just finished a great meal and are now drinking an espresso by the riverside, enjoying the beautiful view, all the while you are listening to a beautiful *Fado* playing from one of the open windows of a house nearby , chances are you are in Lisbon. The capital of Portugal, from where Vasco da Gama set sail to discover the sea route do India on 1497, is a city full of culture, beautiful sites, and nice people. However, not everything in Lisbon is so wonderful. Filled with restaurants, pubs, and nightclubs, the city streets close to the venues are quick to fill with trash, especially on the weekends. This issue is particularly concerning for those who experience the nightlife, where a broken bottle on the ground could spell disaster for anyone who has had a few drinks too many that evening. Unfortunately, waste management in certain areas of Lisbon is still an issue, and even though the Municipality has implemented some measures to reduce the piling up of broken bottles near full bins, not much has been solved. Popular places like *Bairro Alto*, *Cais do Sodré*, *Terreiro do Paço*, are as filled with trash by Saturday and Sunday morning as their bars were full of costumers the night before. If one wants to see where the best pubs are, just follow the trail of broken bottles.

1.2 The Problem

In the present report, I imagined the Lisbon Municipality wanted to implement a better waste management plan, particularly concerning the disposal of glass residues. For that, the stakeholders need to understand which areas have a high density of venues related to the production of glass residues (Bars, Restaurants, etc.), and which are lacking glass recycling bins (GRBs). Although any local resident can pinpoint a lot of candidate places for intervention, let us imagine for the sake of this study that the Municipality only has the budget for a limited number of GRBs. In that case, they need to understand which areas are most severely affected by this problem and invest their efforts and resources accordingly. Therefore, on this report I will present the analysis I used to identify the areas in need of more urgent action concerning the installation of GRBs. This will include the data analysis and leverage of the Foursquare Places API and the knowledge obtained in the various courses of the IBM Data Science Professional Certificate.

2. Data

Before starting to address the problem, it is necessary to collect and prepare the data. The required data is pertaining to both GRBs and relevant venues in Lisbon, particularly geospatial information about individual datapoints. Nonetheless, and since the number of calls with the Foursquare API are limited, I restricted the assessment area to a 2.5 km radius circle centered around the Marquis of Pombal Square, a very famous roundabout in Lisbon which is situated roughly at the geographical center of Lisbon, although it is not the official city center.

GRB data was obtained from the website http://geodados.cm-lisboa.pt/datasets/e4af86f9aabe44a1a036a2c677f2755b_5/data, provided and curated by the Municipality of Lisbon. Although the website displays a table that can be scraped from the html code, I used the .csv file available and uploaded it directly to the Notebook. I selected this approach since it provided two additional columns with Latitude and Longitude data for each GRB, that were absent from the html table. Once I imported the table, I converted it to a dataframe and discarded all information except for the Latitude and Longitude values. Furthermore, I also selected the GRBs located inside our study area and discarded those located outside.

I used the Foursquare Places API to collect venue data. However, since each API call only returns 100 venues, I had to use a different strategy to collect enough information and adequately populate the assessment area with venue data. For that, I first divided the 2.5 km radius circle into a subset of smaller circles with an approximate area of 0.25 km^2, all packed together in a hexagonal fashion (Figure 1). After that, I made individual API calls using the coordinates of the center of each circles and restricted the call to a radius of 284 m, which was the same as the radius of the small circles, and stored the results in a database. From there, I needed to restrict the venue data, excluding irrelevant datapoints from the evaluation. For that, I filtered the venue categories and kept those which had the keywords ‘Bar’, ‘Pub’, ‘Restaurant’, or ‘Café’, storing the relevant data in a new dataframe.

By the end of the data wrangling process, I was left with GRB and venue data located inside the study area. You can have a better sense of the geospatial location of the data by inspecting Figure 2.

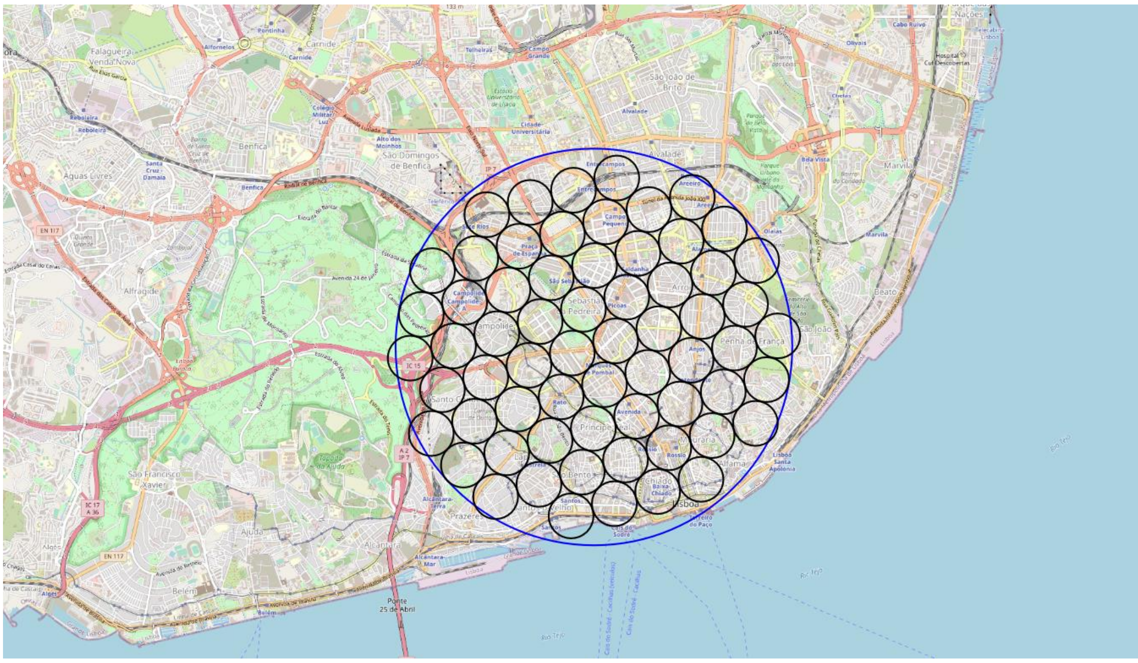
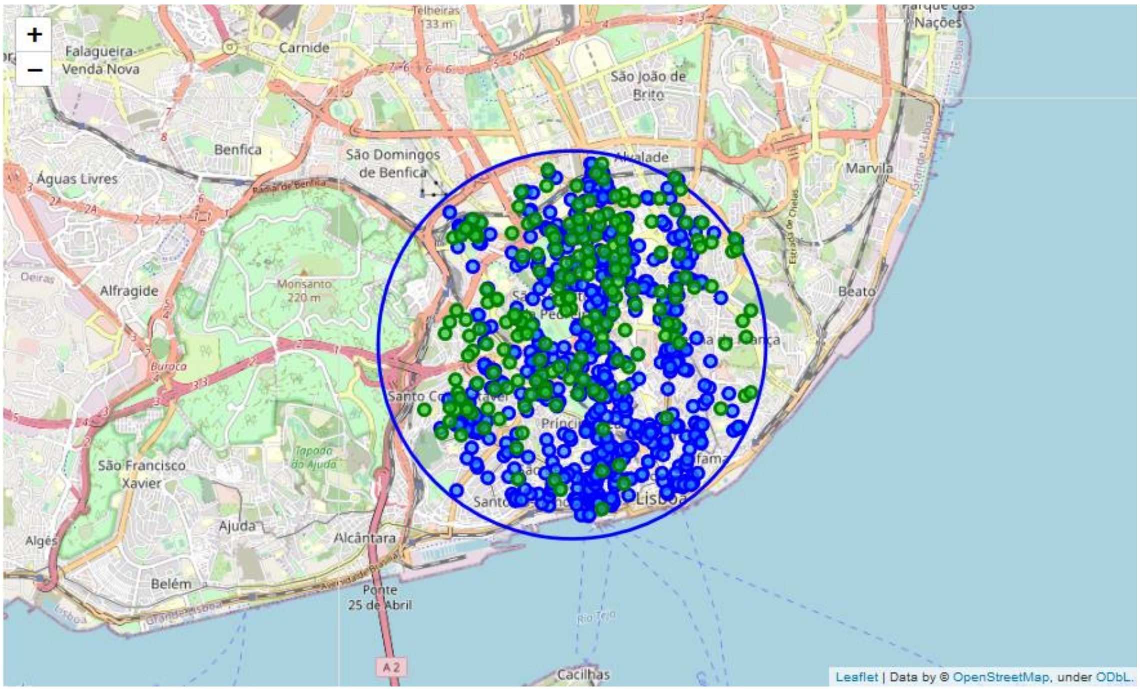


Figure 1. Representation of the subset of small circles, which center coordinates were used to make the API calls.



Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 2. Geospatial data of GRBs (green dots) and venues (blue dots) located inside our study area in Lisbon (blue circumference).

3. Methodology

Before advancing to a more complex evaluation, it is critical to first explore the data, in order to extract meaningful insights for the following analysis. For that purpose, I first wanted to assess how the GRBs were distributed in the study area. Although the simple visualization of the plotted datapoints in a map suffices for an initial appreciation of the GRB distribution, by using a Heatmap it is possible to have a better perception of the density distribution of the bins (Figure 3). By inspecting it, it is evident that there is a shortage of GRBs in the South to Southeast section of the study area. To better assess this distribution across the Lisbon area, I used the same subset of circles displayed in Figure 1 to count the number of GRBs each one contained, and plot this information in a histogram (Figure 4).

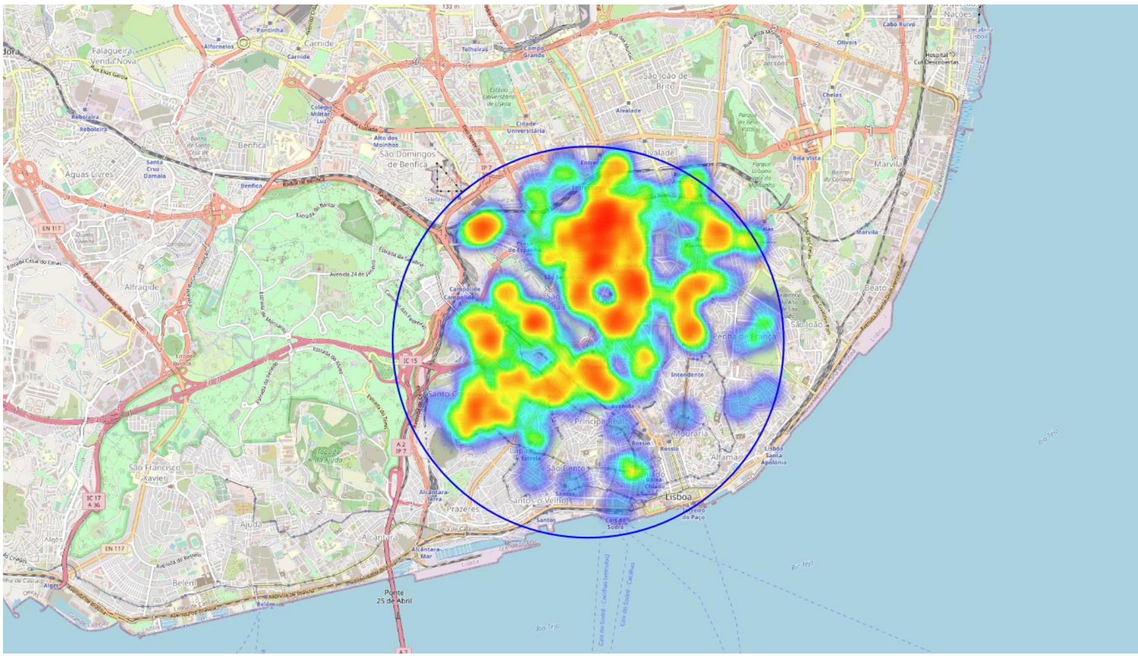


Figure 3. Heatmap of the distribution of GRBs in Lisbon.

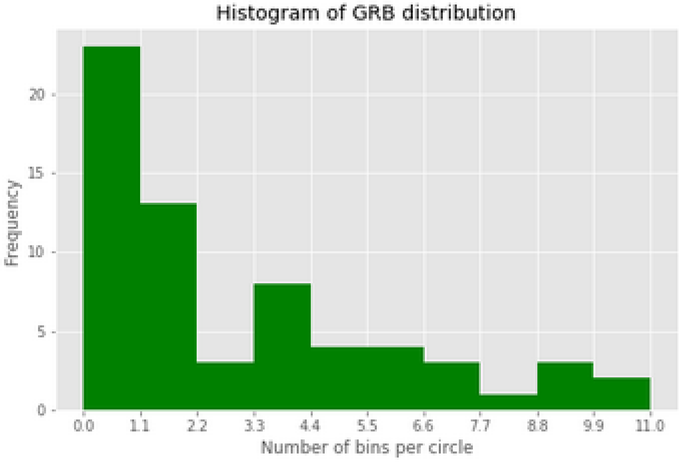


Figure 4. Histogram of the GRB count frequencies across the circle grid in Lisbon.

Likewise, I repeated the same analysis for the venues (Figure 5).

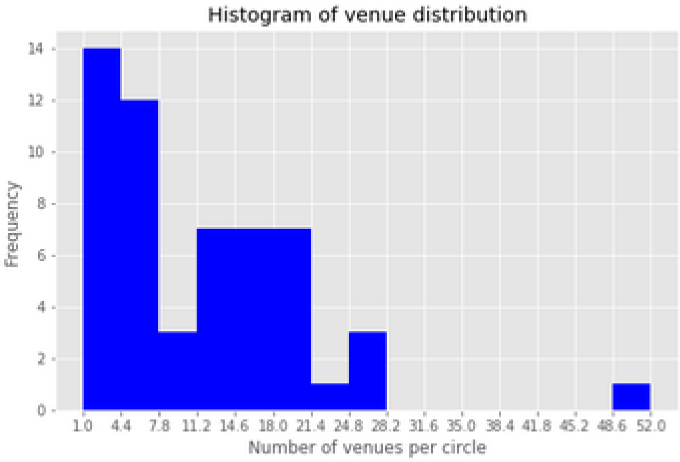
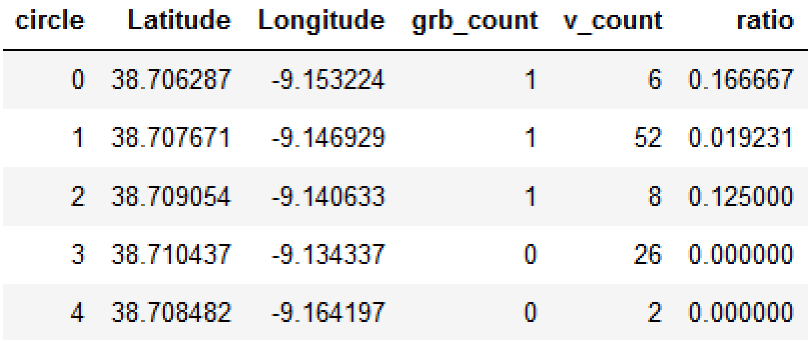


Figure 5. Histogram of the venue count frequencies across the circle grid in Lisbon.

Having the count of both the GRBs and the venues in each circle, it is possible to calculate the GRB/venue ratio for each of them and assess its frequency distribution (Figures 6 and 7). Furthermore, since the new 'ratio' data is specific to each circle, it is easy to plot this information in the map according to a color gradient, displaying circles with lower ratios (few GRBs per venue) in red, and higher ratios (a lot of GRBs por venue) in white (Figure 8). According to this criterion, red circles evidence possible areas of urgent intervention, as the amount of GRBs available is inadequate for the number of venues.



circle	Latitude	Longitude	grb_count	v_count	ratio
0	38.706287	-9.153224	1	6	0.166667
1	38.707671	-9.146929	1	52	0.019231
2	38.709054	-9.140633	1	8	0.125000
3	38.710437	-9.134337	0	26	0.000000
4	38.708482	-9.164197	0	2	0.000000

Figure 6. Dataframe displaying the number of GRBs (grb_count), venues (v_count), as well as their ratio for each of the circles in the subset defined in Figure 1. Only the first five rows are shown.

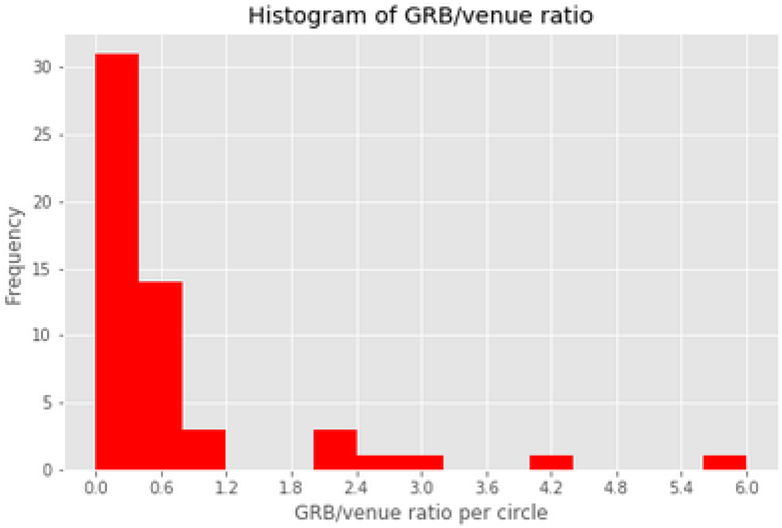


Figure 7. Histogram of the ratio frequencies across the circle grid in Lisbon.

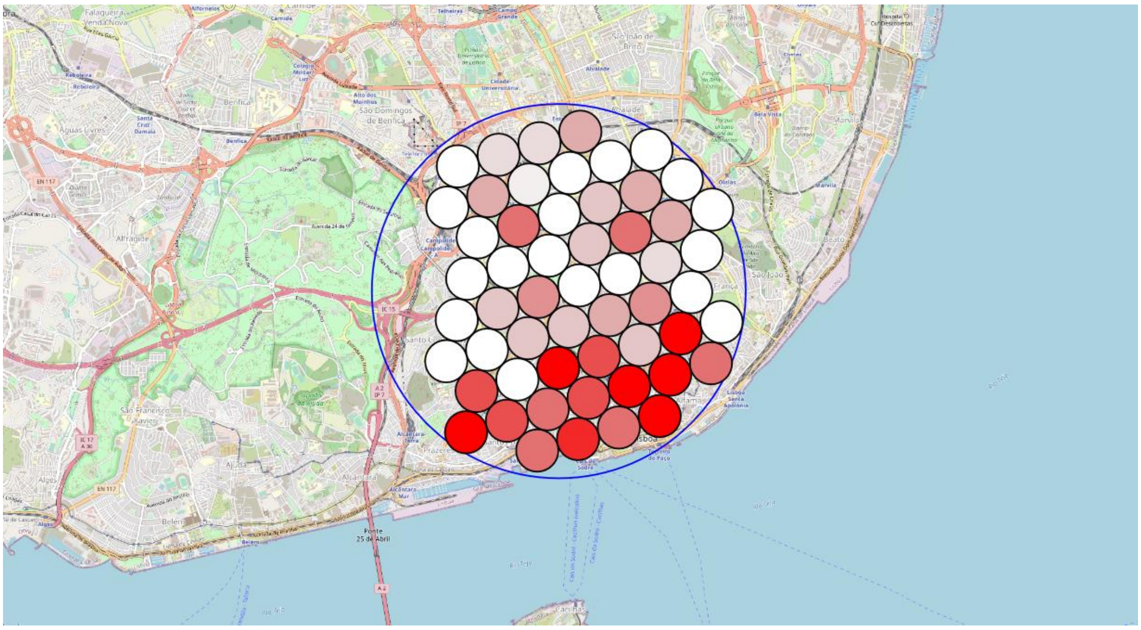


Figure 8. Map displaying the GRB/venue ratio in each circle of the subset contained in the Lisbon area. Red circles correspond to areas where the ratio is zero, while white circles correspond to those where the ratio is grater than 0.5 (i.e., there is at least one GRB for every two venues, which was defined as threshold). Intermediate colors correspond to intermediate ratios, with increasing shades of red for increasingly lower ratios. It is important to note that circles with zero venues were excluded from this representation, as their ratio could not be calculated.

Through visual inspection of Figure 8, we can see that the most imperative cases of GRB shortage are indeed in the South to Southeast region of our study area. Nevertheless, the approach so far has been purely visual, with the information contained in the circle subsets serving as a simplified indication of the real scenario. While simple to visualize, the uniform coloration of each circle is only representative of the mean ratio of the area it contains. To have a more precise notion of the areas in need of intervention, I resorted to the k-mean clustering method to group the different venues in clusters (Figure 9). After that, I calculated the mean GRB/venue ratio of 500 m radius circles centered around the central coordinates of each cluster of venues. This value is more objective, as the assessed circles result from a clustering model that identifies particularly dense areas of venues. Finally, to find the areas in most urgent need of intervention, I sorted the clusters by an ascending order of ratios (Figure 10). As an example, I defined a ratio threshold of 0.1 (i.e., one GRB for 10 venues, at most), and plotted the circles that had a lower value in the map (Figure 11). Thus, we are left with 3 main areas of priority intervention.

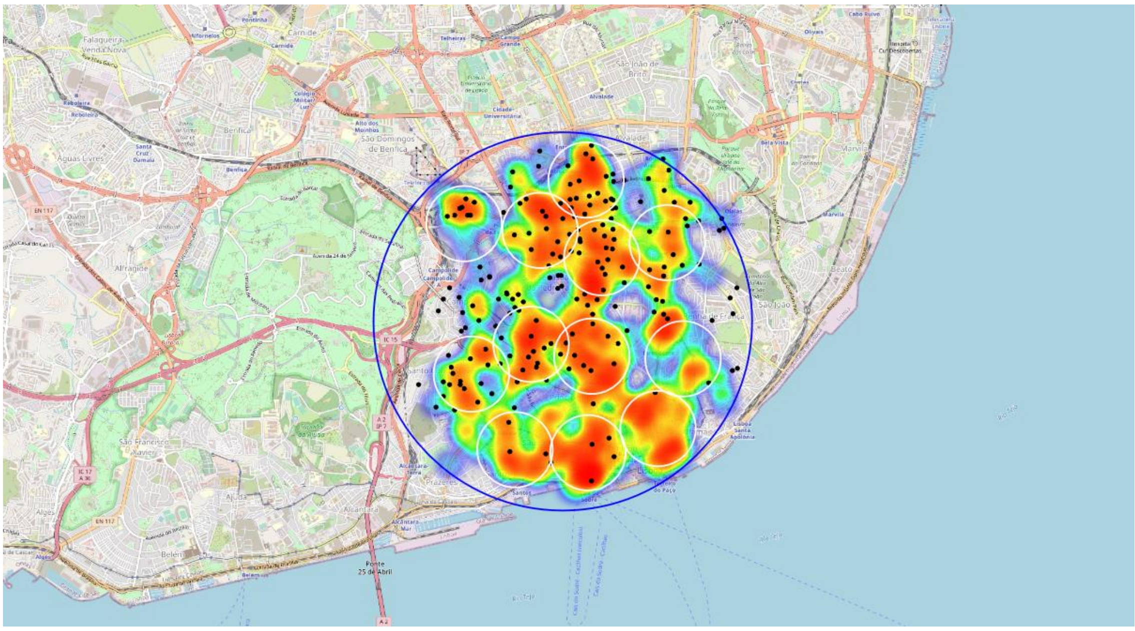
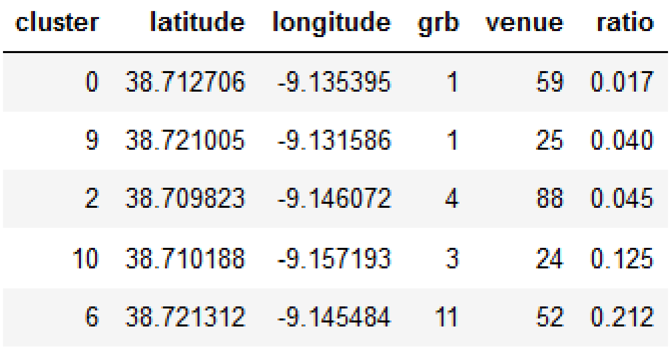


Figure 9. Clusters (white circumferences) formed by k-means clustering (k=12) of venue data, which is represented as a Heatmap. GRBs are plotted as black dots, for ease of visualization and perception of the spatial distribution of the bins.



cluster	latitude	longitude	grb	venue	ratio
0	38.712706	-9.135395	1	59	0.017
9	38.721005	-9.131586	1	25	0.040
2	38.709823	-9.146072	4	88	0.045
10	38.710188	-9.157193	3	24	0.125
6	38.721312	-9.145484	11	52	0.212

Figure 10. Dataframe displaying the number of GRBs, venues, as well as their ratio for each of the circular clusters (radius = 500 m) defined by k-mean clustering, and represented in Figure 9. Only the first five rows are visible and are sorted in ascending order of ratio.

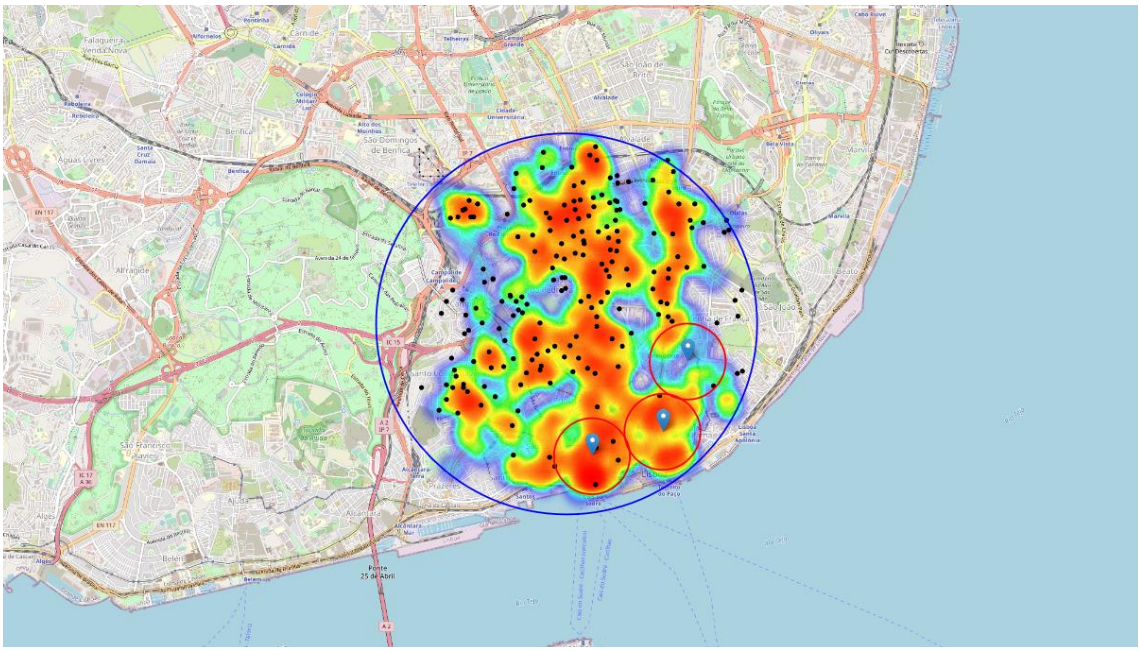


Figure 11. The three main areas in need of urgent intervention are delimited by the red circumferences, and were selected by setting a threshold of 0.1 on the GRB/venue ratio for each cluster.

4. Results and Discussion

It is pertinent to wonder why I did not take that one last step in the analysis and used an API (e.g., the Google Maps API) to retrieve the addresses of the clusters identified in Figure 11. That would, in fact, be appropriate if the resulting addresses were representative of the cluster, which is not true in this case, as it would retrieve the street names where the center point of each cluster is located. However, the general areas delimited by the clusters are referent to well-known areas in Lisbon, which unfortunately are not officially delimited geographically and, as such, an API call would be unable to retrieve the appropriate designation. But, as any local will be able to tell you, these areas are, from South to East, *Cais do Sodré – Bairro Alto*, *Terreiro do Paço – Mouraria*, and *Intendente*. The first area is well-known for its bars and nightlife, resulting in an extremely high number of venues for the four GRBs available in the vicinities (Figure 11). The second is a famous tourist spot, with lots of restaurants and cafés, whilst only having a single GRB available to receive the glass waste of all these venues. Lastly, the eastmost area, while having roughly half of the number of venues as the previous one, also contains only one GRB, leading to the observed low ratio.

The identified areas are coincident with those where residents have previously complained about the piles of broken bottles scattered through the streets, due to the lack of available bins for proper disposal. See, for example, the following news article where residents of the *Cais do Sodré – Bairro Alto* area complain about the amount of garbage that piles up next to the bins:

https://www.dn.pt/cidades/reportagem-moradores-queixam-se-do-lixo-em-lisboa-e-pedem-coimas-para-os-infratores--9812531.html

Although the article is written in Portuguese, you can read it using Google translate if you do not understand the language. Also, please note the GRB in the head image of the article, with garbage piling up next to it, including visible glass bottles. Articles like to this one give strength to the presented results and, more importantly, convey them a proper meaning by serving as an illustrative example.

5. Conclusion and Future Remarks

The present analysis and results allowed me to properly identify areas in need of urgent intervention by the Lisbon Municipality. To summarize, it is imperative to increase the number of GRBs in the Southern region of Lisbon, in order to adequately respond to the overwhelming amount of glass residues generated in these regions. Nevertheless, it is also important to note that this analysis used the presence of specific venues as a proxy of the glass residues produced in an area. This was because data related to the type and amount of residues produced in a given area was not available, and may not be linearly correlated with the number of venues. For example, an area with no bars or restaurants but that is known to have a lot of resident students may also have an associated production of high quantities of glass residues (in Portugal, beer bottles are relatively more available than beer cans). However, to have this type of information one would require a more complex dataset than the one available though Foursquare or other free APIs. For example, data from waste management facilities would be the most appropriate, but that kind of information is private to most waste management companies, and therefore cannot be accessed for free. Nevertheless, it is crucial to remain critical of our work and openly discuss the existing limitations with the analysis with the stakeholders. Only by doing so it is possible to decide what will be the next step to produce better and even more meaningful results.