Madrid Airbnb interactive data analysis tool

Big Data - EIT Master in Data Science

Ángel Igareta, Guillermo Antoñanzas and Cristian Abrante

Introduction

The rise of vacation renting and platforms like Airbnb has created a lot of issues in big cities, like Madrid. We have decided to analyze a public dataset which offers information about Airbnb renting, dates and zones. The dataset is public in Kaggle: Madrid Airbnb Data.

This dataset is well structured, giving the opportunity of extracting a lot of insight from it, also it has a considerable volume of data. We chose it because of its great potential and the various visualizations that could be extracted from it.

It is structured in different *csv* files:

- `calendar.csv`: It is the dataset which contains information about the rentals, each row corresponds to a rental in a certain flat (or *listing* in the terminology of the dataset).
- `Listings.csv, listings_detailed.csv`: This file has all the information in detail about all the flats (called listings) in the Madrid area. Contains information such as the location of each listing (latitude, longitude and neighbourhood), name, description and scores (for 6 different categories and a “total” one).
- `Reviews.csv, reviews_detailed.csv`: Detailed reviews for the listings with comment.
- `neighbourhoods.csv`: List of the neighbourhoods of the Madrid area where there are located Airbnb flats.
- `neighbourhoods.geojson`: GeoJSON file of neighbourhoods of the city. Useful for mapping tools.

Data questions

The objective of this work is to answer several questions according to the presented problem domain using a visualization tool. The questions we are trying to answer are:

- **How does the renting price change depending on what neighbourhood of Madrid the place is? Does it vary too depending on the date?**
- **What are the neighbourhoods with the best ratings? (taking into account all the categories of the scores)**
- **What are the neighbourhoods with the most number of Airbnb flats? Has this changed over time?**
- **What insights can we gather from the review comments? What do clients value more from the Airbnb?**

Next, we will approach all these questions using the selected data and explain the chosen design principles for answering them.

Problem characterization in the application domain

In this section, we will explain which specific problem are we trying to tackle each question.

How does the renting price change depending on what neighbourhood of Madrid the place is? Does it vary depending on the date too?

One of the main concerns about vacation renting is the price. With this question, we are trying to analyze which are the variations over the price of renting over different neighbourhoods of Madrid and depending on a specific date range.

The expected outcome of this question should be the mean price for each neighbourhood when the user specifies the dates.

What are the neighbourhoods with the best ratings? (taking into account all the categories of the scores)

In the Airbnb platform we can find different types of scores which are given by the tenants, these are:

- **Location:** This score is referred to the closeness of the flat to important monuments or relevant places, or to public transport connections.

- **Accuracy:** If the flat was accurate to which was expected by the tenants, according to what was published on the advert.
- **Cleanliness:** If the flat was clean or not.
- **Checking:** Regarding the checking process.
- **Communication:** Regarding the communication with the host.
- **Rating:** The overall score that Airbnb gives to the flat.

With this question, we seek to calculate the average punctuation for each of the scores given a certain neighbourhood. This will offer the user relevant insights such as: "*which are the cleanest flats neighbourhoods?*".

What are the neighbourhoods with the most number of Airbnb flats? Has this changed over time?

The increase in Airbnb renting has been significant over the years. This has caused many issues with neighbours who were living in those zones from for a long time.

Here, we are trying to gather the number of Airbnb flats per neighbourhood given a certain date. In this way, users could know which were the most trendy neighbourhoods in this platform.

What insights can we gather from the review comments? What do clients value more from the Airbnb?

As previously explained, the dataset includes information about the reviews over the years for the different listings. However, these reviews only contain comments, not being available the rating the user left.

In this next question, it would be interesting to follow a data analysis approach over the comments and try to get insight from them.

An example of these insights would be an estimated score of the reviews based on the sentiment of the written comments. Besides, the comments could be analyzed to discover what made the customers more or less happy about their experience (e.g. location, noise...)

Data and tasks abstractions

In this section, we are going to explain the necessary data, and the tasks involved to answer the different proposed questions.

How does the renting price change depending on what neighbourhood of Madrid the place is? Does it vary depending on the date too?

First of all, we are going to define which is the data involved to find the answer to this question.

- **Price of the renting:** We can find this information in the `calendar.csv` file, each of the registries of this dataset contains information about the price at which it was rented.
- **Date of the renting:** Also this information could be found on the `calendar.csv` file. Each of the registries has an associated date specifying when the reservation was done.
- **Information about the neighbourhood:** Each of the flats has an associated neighbourhood, and this relationship could be found on the `listings.csv` file.

Then, we are going to point out which are the different tasks that the user have to do introduce in their workflow:

- **Introduce start and end date:** User has to specify the date range from which the average prices are going to be computed.
- **Compare the neighbourhoods:** According to the visualization output, users have to compare the prices per neighbourhood.

What are the neighbourhoods with the best ratings? (taking into account all the categories of the scores)

Like in the previous question, we are going to define which is the data needed to create the visualization tool:

- **Scores per flat:** The different scores depending on the category (location, cleanliness...), are found in the file `listings_detailed.csv` file when each registry represents a flat.
- **Neighbourhood:** As in the other question, the information about the neighbourhood is found on the `listings.csv` file.

Then, we are going to define user tasks:

- **Specify the type of score:** User has to specify the type of score that he wants to be displayed.
- **Compare the results:** Compare the results per neighbourhood.

What are the neighbourhoods with the most number of Airbnb flats? Has this changed over time?

As in the other questions, we are going to define the necessary data:

- **The number of flats:** Each row of the `listings.csv` file represents a flat.
- **Neighbourhood:** The neighbourhood identifier could be found in the `listings.csv` file.
- **Date**: We do not have a direct way to know the specific date when the flat started as housing in the platform, but we could make an estimation. We have the date of the first review that was published on the website for a specific flat (in the `listings_detailed.csv` file), and we can consider this date as the starting date of the flat.

Then, we need to define user tasks:

- **Specify the date:** The date when we want to calculate the number of flats per neighbourhood.
- **Explore and compare results:** Compare per neighbourhood.

What insights can we gather from the review comments? What do clients value more from the Airbnb?

In order to get insight from the comments, we will need the following data:

- **Reviews**: The comments the users made to the Airbnb after their stay. It can be found on `reviews_detailed.csv` file, specifically on the comments column.
- **Neighbourhood:** Can be found on the `listings.csv` file.

The tasks that the user will be able to do would be:

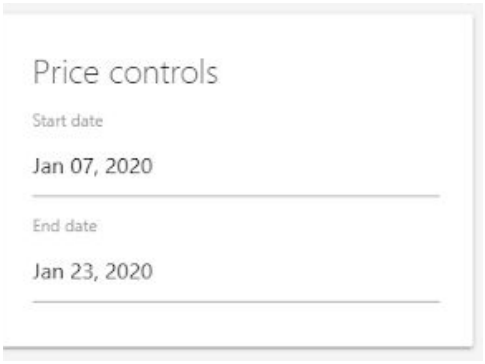
- **Specify the date:** The range of dates the user wants to analyze comments from.
- **Compare the results:** The user would be able to visualize different type of insights from the comments, divided by type of comment.

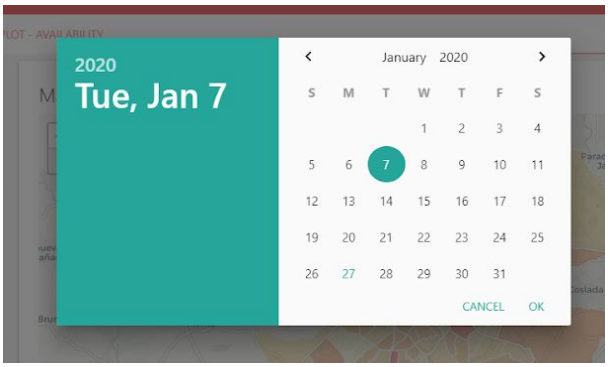
Interaction and visual encoding

In this section, we are going to present the interaction design for each question are the appropriate encoding to show the output information. For the interface design, we used the library *shinymaterial*, in order to improve the visual appearance of the input design.

How does the renting price change depending on what neighbourhood of Madrid the place is? Does it vary depending on the date too?

For the interaction design, we have decided to use a date picker, both for start and end date. In this way, the user can select the necessary dates in an intuitive way.





Figures 1 and 2: Date control and popup, respectively.

The other thing that we have to show is how we presented the information. As we were using geographical data, we decided to use a *choropleth*. In this map were represented the different neighbourhoods of Madrid, and the average price for each of them, codified with a colour scale. The colour scale varies from 0€ (represented in yellow) to more than 500€ (represented in red).

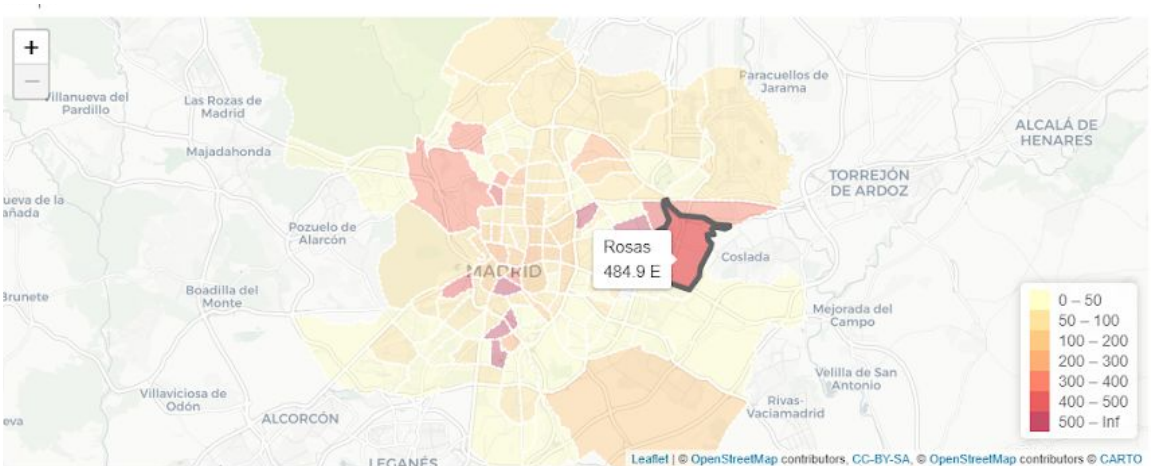


Figure 3: Choropleth of the price of flats per neighborhood

When the user hovers over the map, it is shown the price information for the given neighbourhood. Also, we presented the top 5 neighbourhoods with the given criteria, for doing this we used a bar chart.

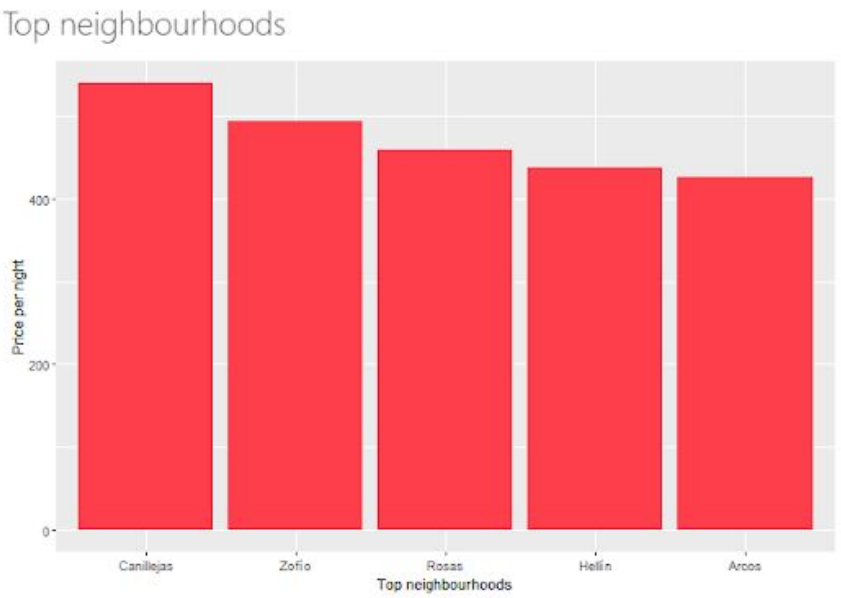


Figure 4: Barchart of top neighbours per price

What are the neighbourhoods with the best ratings? (taking into account all the categories of the scores)

For the interaction we are going to use a select component from the material design library. In this component we predefined a set of options for the user to select, this options are the different types of scores that the user can select.

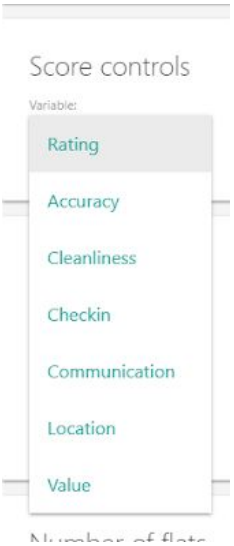


Figure 5: Dropdown of the different Scoring categories.

Showing the information is similar than in the previous question, using a *choropleth*. With this representation, we can present the scores per neighbourhood's, encoded with a color scale which vary from the minimum value to the maximum in steps of five.

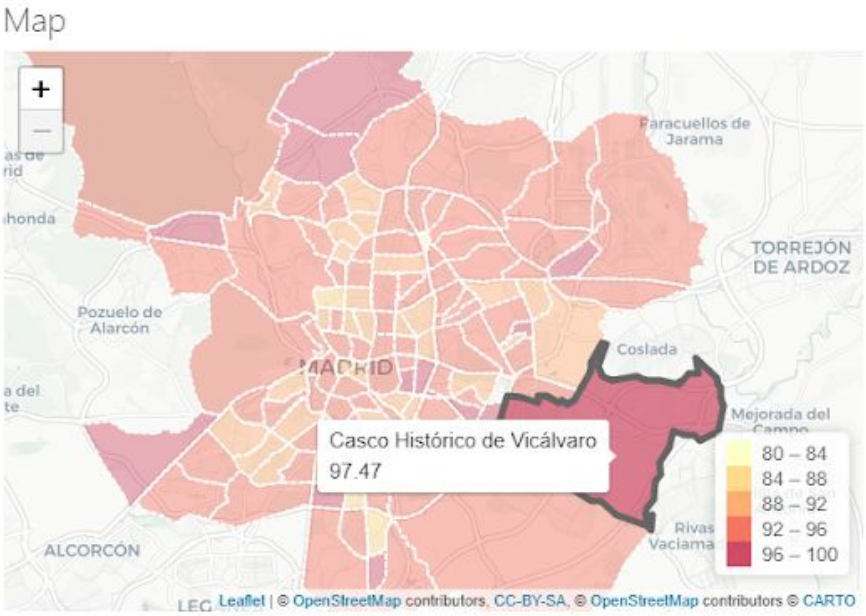
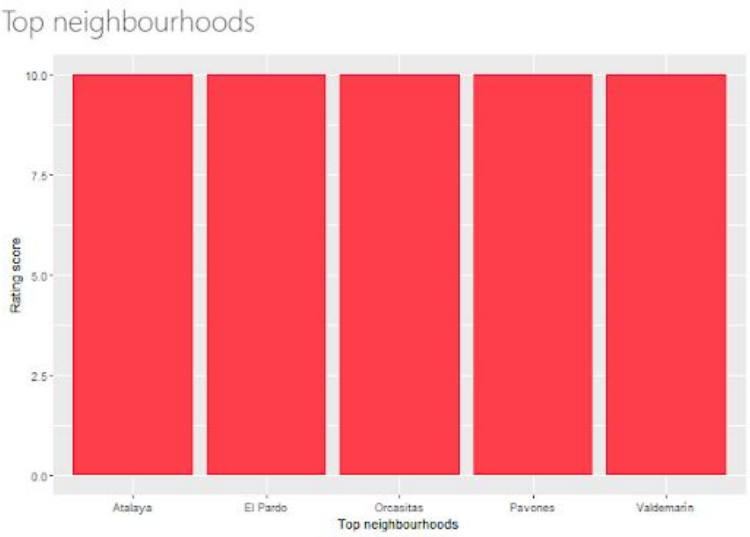


Figure 6: Choropleth of the ratings of the different neighbourhoods.

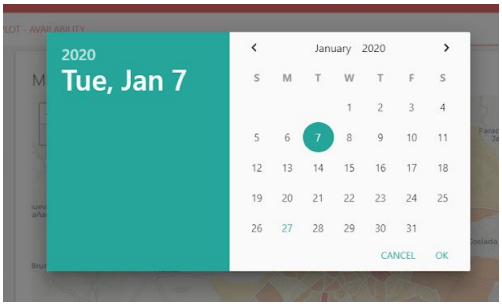
Also, we presented a barchart with the top 5 neighbourhoods given a score type.



Figures 7: Bar plot of the highest rated neighborhoods.

What are the neighbourhoods with the most number of Airbnb flats? Has this changed over time?

The interaction for this question is very similar to the first one, because the user has to introduce a date. We used a date picker from the material package, but this time only for introducing one date instead of a range.



Figures 8 and 9: Date control and popup, respectively for the Number of flats.

For showing the information, again we chose the *choropleth* map. In this case, the color scale was encoded depending on the number of flats, from 0 (encoded in yellow) to more than 1500 flats (encoded in red).

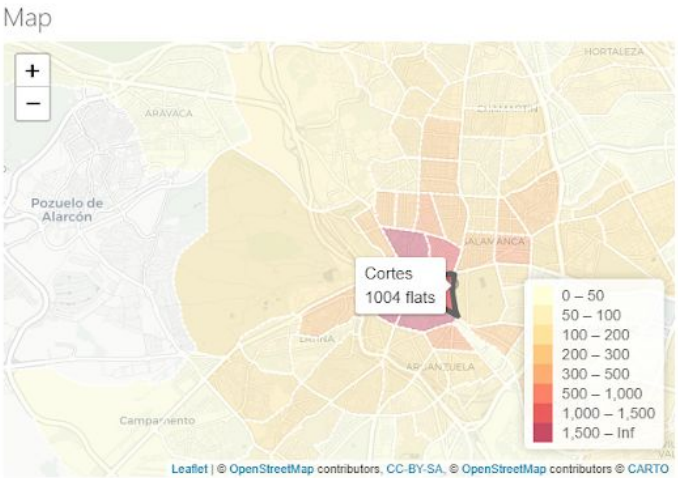


Figure 10: Choropleth of the number of flats per neighbourhood.

As in the other cases, we added a bar plot with the top neighbourhoods in number of flats.

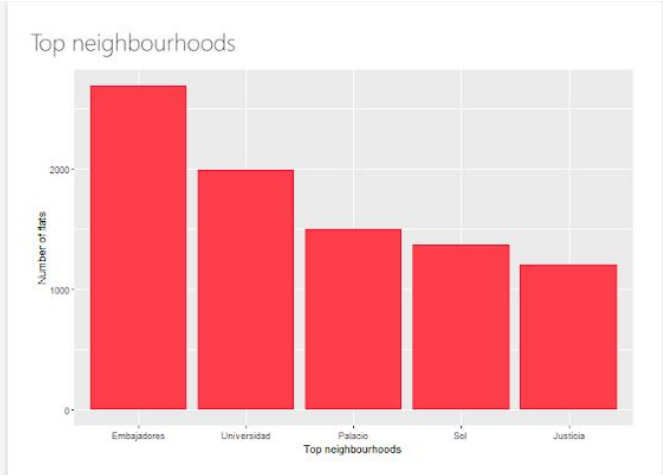


Figure 11: Barchart of top neighbours per number of flats

What insights can we gather from the review comments? What do clients value more from the Airbnb?

As in the previous questions, in this one, we offer the user the possibility to pick a range of dates through DatePickers to filtering review comments according to the time they were written.

In order to visualize an insight of the comments, after performing the algorithms that will be explained in the next section and extracting the sentiment from each of the comments (positive or negative), we generated a word cloud per each comment category.

Figure 12: Cloud of Words of the reviews with a positive or negative sentiment.

Through this visualization, the user can see which words are more frequently used in the positive and in the negative comments. As we removed the stopwords, most of the words would refer to which aspect of the experience they were more or less happy about.

Algorithmic implementation

For this part, we will explain how we decided to implement the solutions to the problems and the visualizations we chose.

How does the renting price change depending on what neighbourhood of Madrid the place is? Does it vary depending on the date too?

For all the Choropleths we decided to use the included files of the Kaggle page that were in .geojson format to get the neighbourhood polygonal data. This, mixed with leaflet and the listing information of all the places meant that we just had to assign to each area of each neighbourhood the aggregated data and set up the necessary filters that would depend on the different types of data we wanted to represent on the map.

What are the neighbourhoods with the best ratings? (taking into account all the categories of the scores)

Because for this part we also had Choropleths in mind, we could use a very similar implementation than the one previously described but with the aggregation done with the rating fields that we extracted from the `listing.csv` file.

What are the neighbourhoods with the most number of Airbnb flats? Has this changed over time?

Again, the implementation for this part is similar to the previous ones. However, this time we will take the information from the field “first_review”, which holds the date of the first review the apartment got, to assume that it would also be the time the apartment first got inaugurated.

What insights can we gather from the review comments? What do clients value more from the Airbnb?

As previously explained, the dataset includes information about the reviews over the years for the different listings. However, these reviews only contain comments, not being available the rating the user left. In order to generate this insight, we are going to follow a data analysis approach called *Sentimental analysis*.

Sentimental analysis is a technique to automatically analyze data and detect the polarity or sentiment of a text, in our case a positive or negative review.

In order to perform it, we made use of a well-known library for this kind of studies called 'sentimentr', which we preferred from others because it takes into account valence shifters (negators, amplifiers, de-amplifiers...) while maintaining high speed.

After running some manual tests and comparing the outcomes, we found out that the algorithm used by this library worked poorly in the sentences that were not written in English, so we decided to filter and analyze only the English comments.

In order to filter by language, we chose the google library cld3 because of its high performance, once it was compared with textcat. After discarding the not English comments and running the sentiment analysis over them, we achieved very impressive results, checking the classification manually.

Once the comments had their sentiment score associated, we cleaned the text data for each comment, removing common stop words in English, numbers or punctuation. After that, a document-term-matrix was created containing each word with their frequency. Finally, we made use of wordcloud2 in other to show the cloud of frequent words per sentiment.

Conclusions

After developing this tool and make and exhaustive study of the dataset we arrived at some interesting conclusions.

First of all, speaking about the price, we can see that the most expensive neighbourhoods are those in the residential and fancy part of Madrid (like Salamanca and Goya), with some exceptions due to high rent places (outliers) located in other parts of the city.

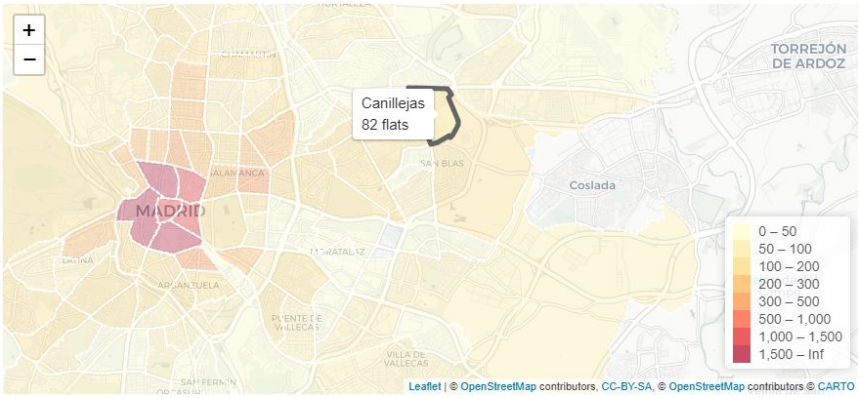


Figure 13: Choropleth view of the # of flats per neighbourhood.

To see if this is due to a problem with the offer and demand in that zone, we can use the other visualization to see the amount of places available in that zone. The number is really small compared to other parts of the city like the centre, but similar to other “extra radius” listings, which would validate our assumption that the zone is overall more expensive.

Considering the score question, we can see that mostly in all the categories, all the zones have a good punctuation. The only exception is when we are considering the location punctuation, because the most central places are the ones which have a highest punctuation (Sol, Recoletos, Justicia...).

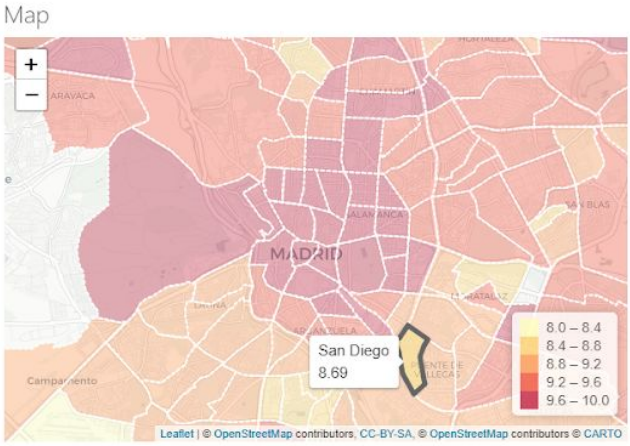
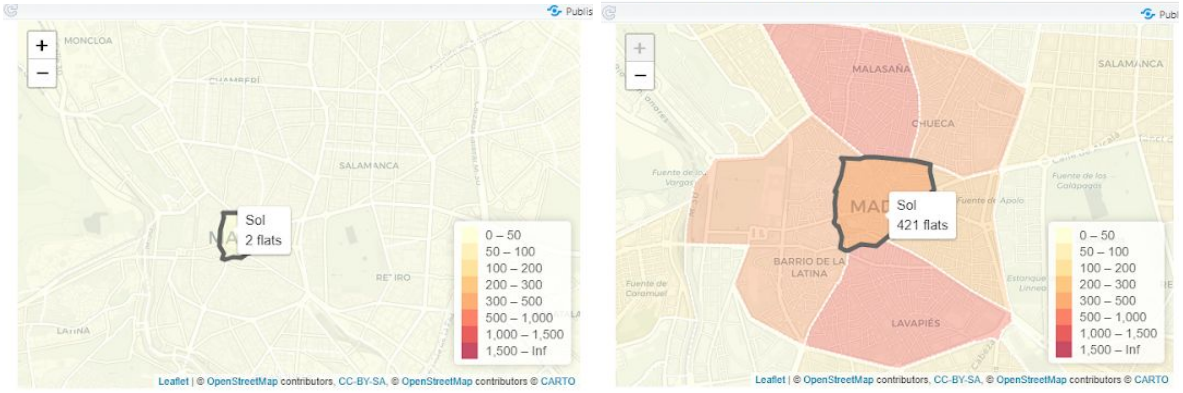


Figure 14: Choropleth view of the average score per neighbourhood.

Regarding the number of flats, we are going to compare for example the number of flats that were in Sol in the year 2012 and in the year 2019.



Figures 15 and 16: Choropleth views of number of flats for different dates.

As we can see, in 2012 there were only two flats in Sol, but in 2019 there are 421. So we can conclude that each year more and more flats are added to the platform.

How to run the tool

The best way to run the Shiny app is to use the deployed app is at angeligareta.shinyapps.io/shinyairbnb/.

In case you want to run it locally, you have to use the `Code` folder uploaded with the report. This folder contains the RStudio project that could be opened.

After opening the project you can execute the command `runApp()` in the R terminal, and the application opens the tool in other window. Once you have run the application, you can find this user interface, divided by tabs.

In the first tab, you can see the sentiment analysis part of the application and in the second tab the display of maps related to price, score and number of flats.

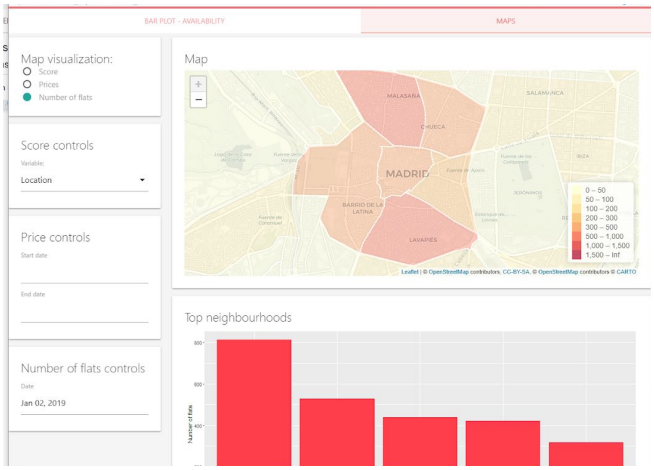


Figure 18: Example of the Tool.

In this tab, you can select (**with a radio button**) if you want to display the prices, the score or the number of flats analysis. Behind this radio button, you can see three cards with the different controls for the different options, they are only useful when the option is selected.

Published tool

The tool will be available in angeligareta.shinyapps.io/shinyairbnb/

Bibliography

Leaflet documentation: https://rstudio.github.io/leaflet/

Shinymaterial: https://ericrayanderson.github.io/shinymaterial/