Visualization

Arturo Fredes, David Íñiguez and Jaume Sánchez

1 Exploratory Data Analysis (EDA): Data (bottom-up): Study and describe the data + explore relationships

1.1 The data

We decided to used a dataset that is about Airbnb. Particularly we decided to do the dashboard of the city of Barcelona. The whole information that is going to be used to build the dashboard is divided in four different files. Each file contains specific information about different aspects that have to be taken into account.

- "Listings": Contains data related to every Airbnb, such as its location, webpage of the listing, and information about the host.
- "Calendar": It includes information about the prices each listing has every day, and the minimum and maximum nights that is possible to rent that Airbnb.
- "Reviews": List of all the reviews written by guests.
- "Neighbourhoods": Includes the name of the different districts of Barcelona, and how they are divided in neighbourhoods.

It is reasonable to think that most variables are categorical, such as dates, reviews, and ids. The main numerical variables are the prices, the number of nights, and the number of rooms of each accomodation. Note that most categorical variables have a temporal nature (except the ids), because they depend on when they were published, and even the prices are time-dependent. Finally, it is worth it mentioning that the location of the Airbnb logically have a geographical nature.

As well as this is a dataset related to Airbnb, it should be updated frequently, although it is known that this information is updated quarterly. Note that "Neighbourhoods" is static, and must be updated only if there is a redistribution of the neighbourhoods, or if some Airbnb are listed in areas not incorporated in this file yet. "Listings" is a bit more dynamic, and must be updated when new Airbnb appear that are avaliable. Other aspects like changes in the information of the lodgings must be taken into account, too. Finally, "Reviews" and "Calendar" are the most dynamic files, because are the files that must have the information as recent as possible, because contain the latest information about the different Airbnb, such as prices and reviews, that are very helpful for guests to decide which Airbnb to book.

From the latter it can be drawn as a conclusion that the data has a short life span. From the point of view of a customer, he would like to have the information as up-to-date as possible, regarding all prices and reviews. These are important because they show the

current state of the accommodation, and how much it will cost to them to stay in that Airbnb.

One important issue is always the existence of outliers. In the Airbnb dataset, an outlier can be find in many different contexts, e.g. when comparing prices of accommodations in close locations and with similar characteristics, or reading reviews of an lodging, and the way they are interpreted.

In case of the first example commented, as a customer, having an outlier way below the average (cheaper) is very important if the price is the most important characteristic at the time of booking, so this outlier is taking into account. If the outlier is over the average price, it is quickly dismissed. Instead, as a host that wants to set a price, outliers are not very important, regardless of whether it is cheaper or more expensive.

2 Audience (top-bottom): Who is your audience?

Our dashboard's target audience are Airbnb hosts (new or experienced) who want to get some insights on current prices, demand and how to offer a better service.

Because of this, we assume the users will not be experts in charts, therefore, our dashboard should be as easy to use and understandable as possible. That is why we chose to have minimal interaction, and just by selecting your neighbourhood, you are able to get general information on how people close to you are setting prices, the type of accommodations you are competing against, the seasonality of tourism and what the accommodates value most and other overall information. For further information on competition, we developed a second map, where users can select each individual Airbnb and check specific data.

The dashboard should be hosted in a webpage for any host to check whenever they want to. To make the information relevant, it should contain the tendencies of the previous year as well as recent information, therefore the dashboard should probably be updated monthly. As a proof of concept, we used the data of 2022.

This dashboard was created to help hosts as we commented earlier. These are some potential questions a new host make ask himself:

a. Where and when is Barcelona most visited?

b. Are there strategies for adjusting prices for seasons or special events?

c. How do I manage my calendar and set availability?

d. How should I set my nightly rate?

e. What is important for a guest to have a comfortable stay?

f. How important are guest reviews, and how can I encourage positive reviews?

g. Are there any special amenities that make a listing stand out?

To answer this questions, first we created a map of the neighborhoods gives information about visits at a glance and answers the first question. The map also allows the user to filter information by area, enabling them to get more relevant information for their purpose. Our dashboard also includes time-related information to answer questions about seasons

and fluctuations of visitors and pricing. Our word cloud leverages all the reviews from guests and gives information on what is important to them and what stands out in good accommodations. Finally, the second map, contains all detailed information about every lodging, so new hosts can look at competition more thoroughly. In a nutshell, if we were hosts, we would consult the visualisation if we want to compare our accommodation with other lodgings close to us (e.g. price), if we want to start renting it, or to decide which time of the year should we put it on the market.

Nevertheless, the dashboard is also useful for clients, since it has information about prices and availability of accommodations throughout the year. In the second slide we show each Airbnb and how the lodgings are distributed on the map of Barcelona. If we were guests, we would check the dashboard every time we want to book an Airbnb.

2.1 Create a "napkin" design, low-fi, and explain how this design will answer the proposed questions. Include any contextual information that is relevant to your message.

In order to try to answer this questions, we designed a low-fi "napkin" design, shown in figure 1:

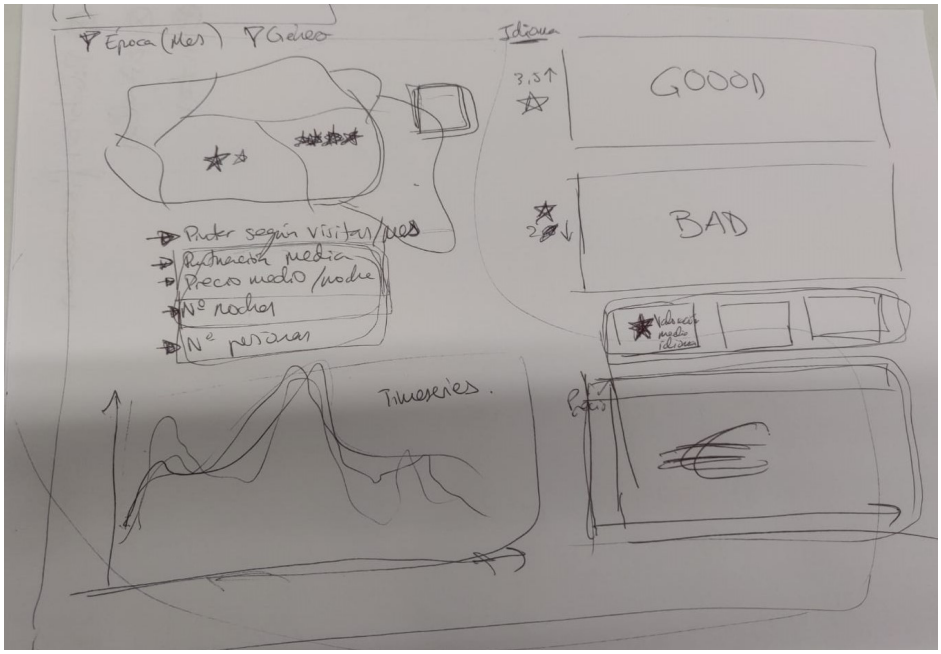


Figure 1: *Napkin design. Low-fi*

We can see that the design is divided in different parts:

- Map of the city distributed by neighbourhoods, adding information about them, like rating, average price, occupation,number of accomodates
- Time series.
- Wordclouds using good and bad reviews.
- Charts related to prices

We think that including this charts and visualizations, we can solve the questions previously commented. Having a look at the map of the city and the information provided from there can be helpful in order to answer questions a and d. By analyzing the time series, we can solve a,b,c. The wordclouds contain information about the reviews, which are the answers to questions e,f,g. And finally, having information about the prices is helpful for question b and d.

3 Selection of chart and encoding

In this dashboard, we use multiple charts and encodings to represent different attributes of the different accomodations. Depending on the kind of information we want to represent, we will select one chart or another. And we also have to pay attention to the encoding of each chart, so that all the charts are in harmony.

3.1 First slide

In the first slide, we decided to divide the dashboard in three parts or sections, containing the following charts:

- Left-side:
  - Interactive map of Barcelona
- Middle-side:
  - Time series
  - Wordclouds
- Right-side:
  - Doughnut charts

3.1.1 Map of Barcelona

In he first section, located in the left side of the dashboard, we put an interactive map of Barcelona, where it is possible to differentiate the neighbourhoods of the city. At the beginning, we added a feature that consisted on that if we moved the cursor to one neighbourhood, a small box that contained the name of the neighbourhood and the numbers of Airbnbs that are in that area appeard. At the beginning, we put much more information in the box, but in the end we simplified that, because it contained information that will be displayed in other parts of the dashboard. Furthermore, we took into account that the bigger the size of the box, the bigger part of the map is covered, and we tried to avoid it. Finally, we eliminated that box because in the end it contained information that is shown in other parts of the dashboard.

We have added a comment that says "press 'ESC' to undo your selection", to make it more dynamic the map and to guide people who may have sensitivity problems, so that they have an easier time choosing other neighborhoods, giving them the opportunity to use the keyboard to undo the selection, which can be easier for them, instead of using the mouse.

We followed the principles of interaction: "Overview first, zoom and filter, details on demand". First, when you enter the dashboard, a general view is fixed where you can see all the neighbourhoods. Once you enter the map you can navigate it as you like zooming in and out and translating the map. If you click a neighbourhood, it will filter the information of the whole dashboard giving information of the user's area, more relevant when comparing your house to others. As we know interactions suppose an effort for the user, we gave a short explanation on how it works.

Furthermore, we included a second interaction to change what is displayed in the map. The colour gradient chosen to display where these values (number of Airbnbs, price per accommodate, rating) are greater or smaller was of red tones, since these are more easily distinguishable by the human eye. For map selector there were different ways to display it. To choose one of them we have focused on the option that has a larger selection area, in line with what was mentioned regarding the "press 'ESC' to undo your selection" box, to facilitate interaction for people with sensitivity problems.

The maps obtained are shown in figure 2

Decisions about how to show the map

First of all, we set the context color to grey because we are going to label with a color gradient. Then, in order to make the map accessible to color blind (assuring yellow-blue distinction), we set the color gradient as a scale of reds, from less saturated (smaller) to most saturated (bigger). To separate neighbourhoods, we draw a black border. Finally, in order to better differentiate colors, we discretized the color spectrum, instead of having a continuous spectrum, which in theory is more complete, but on the other hand it is more difficult to tell differences between neighbourhoods.

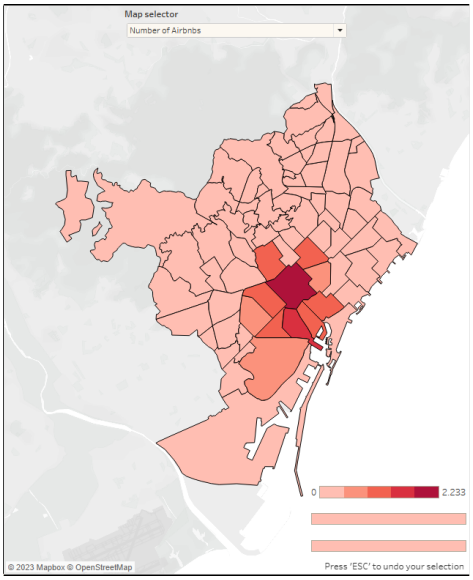
3.1.2 Time series

We also included two time series of the occupation of the accommodations and the overall price per person.

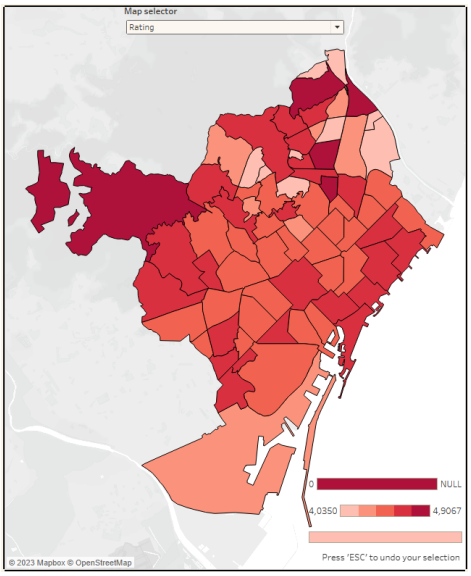
It is known that prices and occupation vary throughout the year, and that is why we thought that it could be a good idea to represent that variations. It can be helpful for both hosts and guests. For the first one, this provides supporting information that will help them to set a price for his accommodation or to know what time of the year his neighbourhood is more demanded. On the other hand, a guest, will be able to compare prices between neighbourhoods, or to know when a neighbourhood is more occupied.

We used a line chart to represent this time series because it is the most common way to represent temporal evolution of variables. We thought about using a bar chart instead, but we finally dismissed it, because an bar chart usually represents how a magnitude is distributed between different answers, which is not the case here. We represented a whole year to see these variation through the months. Moreover, we placed the time series together with a shared axis so user can observe the correlation between demand and pricing. we used vertical lines to see this comparison more easily. Since sometimes it is difficult to see the exact number of a point in a time series, we included an emergent description with the values. This way the user can have more detailed information on demand whilst also being able to see the overall trend at a glance.

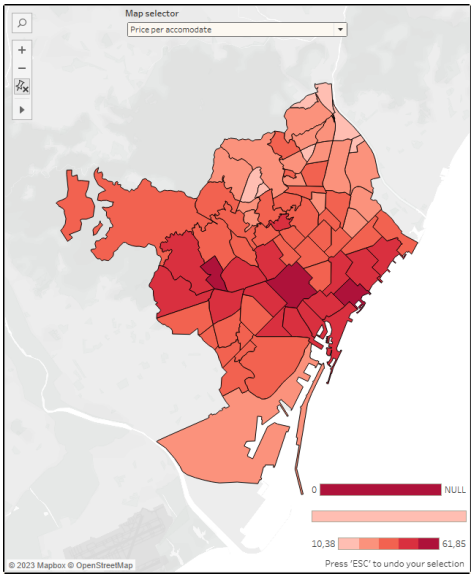
In relation to the colours used here, we chose blue and pink so that they contrast with



(a) *Map based on the number of Airbnb*



(b) *Map based on the Rating*



(c) *Map based on the Price per accommodate*

Figure 2: *Maps shown in the dashboard based on different attributes*

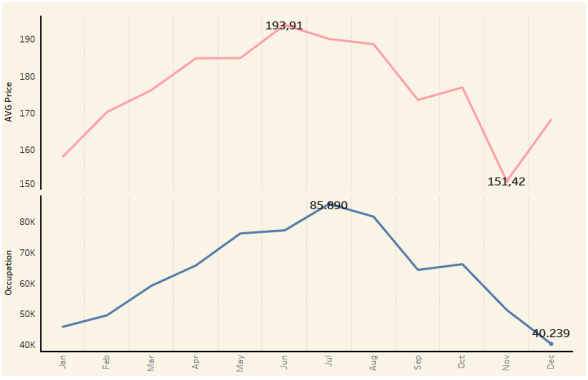


Figure 3: *Timeseries chart*

the background.

3.1.3 Wordclouds

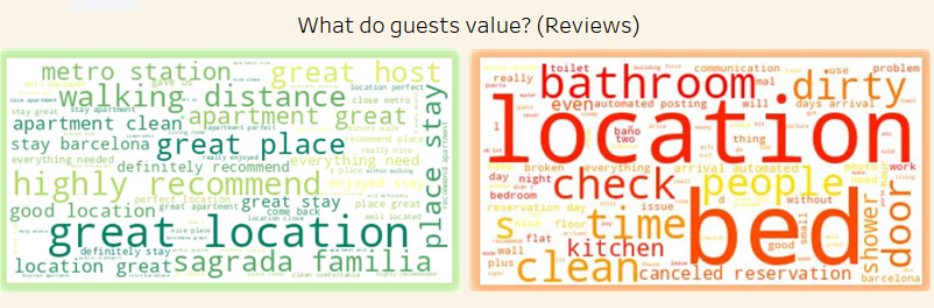


Figure 4: *Word clouds obtained from positive and negative reviews*

We also added two wordclouds, obtained from the positive and negative reviews. We think these are important because they contain the keywords that define if a lodging is good or bad. To decide if a review is positive or negative, we pay attention to the star rating of each Airbnb. This star rating is obtained from the listing of them. If the star rating is below or equal to 3.5 (in a scale from 0 to 5), we consider that the review is bad. In the other case, we have a good one. Note that we preprocess the comments, by just eliminating stop words and punctuations.

One important issue is that the star rating is referred to the Airbnb, and not to its reviews, so its likely to have good reviews in bad Airbnbs and vice versa. This has an small impact in the good lodgings, because the amount of bad reviews is much less than the number of good valorations. However, in the other group the proportion is much bigger, so we found necessary to apply some kind of filter, so we can discard the good reviews of that set. That is why we decided to use the API of ChatGPT, in order to make him decide whether if a review is good or bad. The results we obtained were extremely positive, because the API was totally capable of discriminating the positive and negative reviews. This made us able to have much better wordclouds.

Finally, in the group of bad lodgings, we decided to dismiss some of the words that are the most common ones when we are talking about Airbnbs. These words appear in a lot of reviews and do not provide precise information. Some of that "Airbnb stop words" are: apartment, room, place, us, Airbnb, stay and host.

Finally, we added a contour in order to make

Decisions about how to show the wordclouds

Here, the size of the words included in the wordcloud are related with the frequency of appearance of that word in the reviews, so that the most important ones are going to be bigger in the wordcloud. We think this is key because we tend to pay attention to the big words, instead of the details.

We also thought about the color of the wordclouds. Typically green is associated with good things, and red with the bad ones. So, we decided to plot the wordcloud with positive reviews in a scale of greens, and the plot of negative valorations in a scale of reds.

3.1.4 Doughnut charts

In the right-side part of the dashboard we included doughnut charts we thought gave key information. The information we want to show can be easily compacted in four or five categories, and we think that a doughnut chart is an appropriate way to show that. Note that we show charts of characteristics of each Airbnb, which are not time-dependent, like the room type, the number of the accommodates and the star rating of the accommodation, which reinforces the idea that this type of chart is appropriate to represent these qualities of Airbnbs.

We also discussed the use of a pie chart or a doughnut chart. We finally chose the second one, because we wanted to write in the hole of the doughnut the property that would be represented in the chart, and important statistics of them, in some cases is the mean, and in others the number of elements, for example. Also, in doughnut charts people perceive length versus angle perceived in pie charts, helping with understandabilty

The charts are linked with the map, so that if a neighbourhood is selected in the map, the charts show the distribution of answers in that area, and if there is no one selected, we show the stats of Barcelona.

We think these charts provide very important information because in the blink of an eye you get an overview of the type of Airbnb in the neighbourhood and the general quality of these accommodations. If you want to see an specific Airbnb, you can just go to the second slide and you will have all the information of that particular lodging.

Room type chart Star rating chart

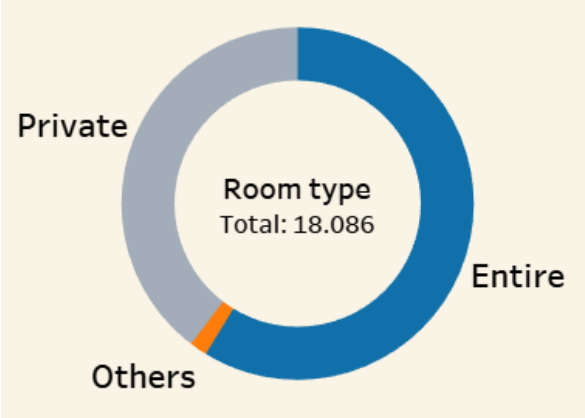
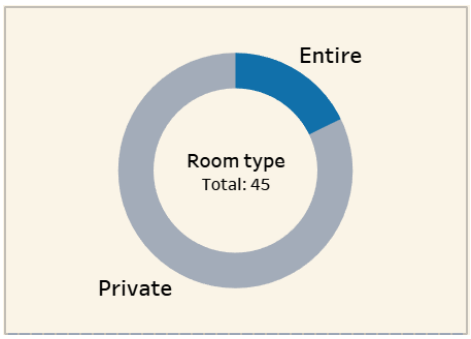


Figure 5: *Room type doughnut chart*

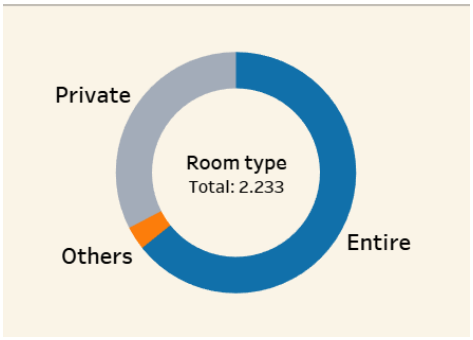
From the first moment we were clear that this chart had to appear. It is an important characteristic and one that fluctuates quite a bit from one neighbourhood to another (see figure 6). We had four categories: private room, entire home/apt, shared room and hotel room, but we decided to put the last two together because the proportion of these two was very small in relation to the other ones.

In the hole of the doughnut, we decided to write the number of elements of the chart, which means, the number of Airbnbs in the area selected.

Note that the variables represented in this chart are categorical, which means that each category should be represented with a colour that is different from the other ones. In this case, having a colour gradient is not recommended, as well as it can be relationed to a ordinal



(a) *Chart for La Verneda i la Pau, at Sant Martí district*



(b) *Chart for La Dreta de l'Eixample, at Eixample district*

Figure 6: *Comparison of two room type doughnut charts*

chart (just like the star rating chart). We took into account color-blindness, so that we chose two different tones of blue for private room and entire home/apt, and orange for the other category.

Number of accommodates chart

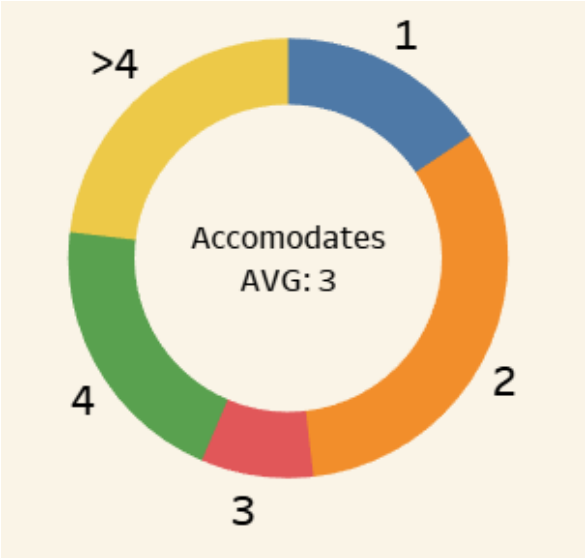


Figure 7: *Number of accommodates doughnut chart*

In this case, we discussed whether to represent the number of accommodations by Airbnb or the number of rooms, but in the end we opted for the former since we consider it to be more representative. We compacted the number of accommodates in $1, 2, 3, 4$ and > 4, because if we apply this partition, the proportion is similar between them.

In this case, in the hole of the donut we write the average accomodates per neighbourhood. This gives an idea of the size of the houses of that area.

As in the previous chart, the variables here are categorical, so that the colour choice here is made in order to differentiate one category from the rest (no colour gradient). We chose colours that are not used in any previous chart, to avoid relating one chart with the other ones, since they represent different attributes.

Star rating chart

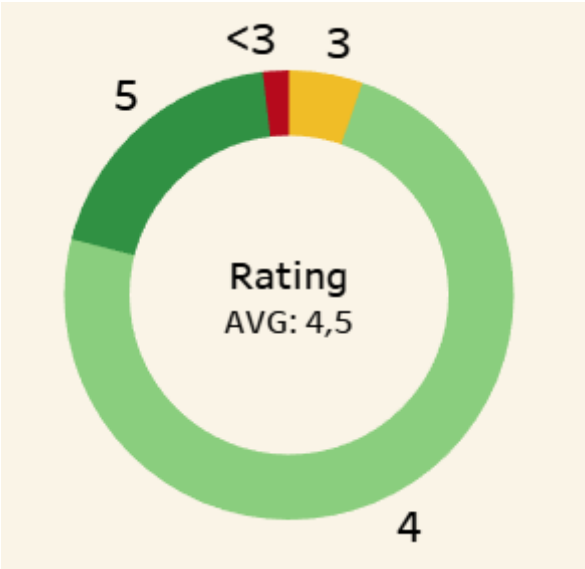


Figure 8: *Ratings doughnut chart*

The last doughnut chart we created represents the star rating. The categories are $< 3, 3, 4$ and 5. We placed it there so that it is close to the wordclouds we made, since it and the star rating are correlated (as will be explained later, the wordcloud is built from good and bad reviews). Having a look at the data, we noticed that there were very few valorations with less than 3 stars, so we decided to put all those together in a category. Note that this is reasonable because each of the components of this group represents the same idea, that is that the quality of that Airbnb is low.

In the hole of the doughnut we write the average star rating of the neighbourhood. We show that because which gives an overview of the quality of the Airbnbs of that area.

In relation to the colors chosen for the chart, we decided to emulate a traffic light (like in the wordclouds), since red is commonly associated to bad reviews, yellow to intermediate valorations, and green to positive ones. We consider < 3 as a bad valoration, so that we decided to label it with green

3.2 Second Dashboard

In this dashboard, we wanted to give more details about each individual Airbnb. To do that, we did a second map with points for each of the Airbnbs. This time, we also coloured the map according to neighbourhood, but the colour is not indicative of any value, only a delimiter. That is why we used the daltonic colour palette of Tableau to make this limits as clear as possible for the maximum amount of users.

Furthermore, we wanted to make the map as easy to navigate as possible. Regarding this issue, we added a layer to the map with streets and important information, making the users able to find their house in the map.

The map is placed on the left because it is what the user should see first, since the information on the right depends on what we select on the map.

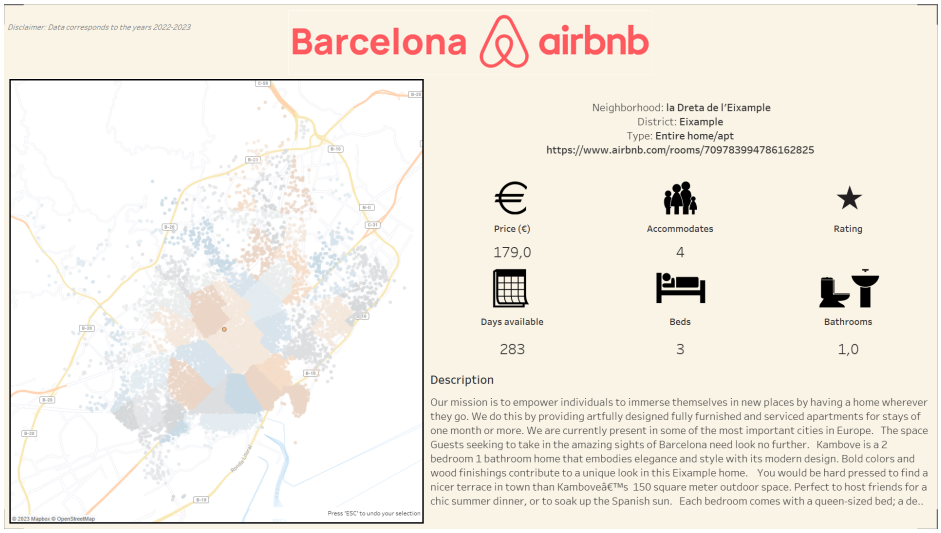


Figure 9: *Second slide of our dashboard*

On the right we wanted to display the description of each individual lodging, and general information when nothing is selected. The information chosen was:

- Neighbourhood information
- Type of accommodation
- URL
- Price
- Accommodates
- Rating
- Beds and Bathrooms
- Availability
- Description

We displayed this as KPIs, and to make the information more clear at first glance, we made the title redundant with symbols we obtained from copyright free image sources (pixabay). In addition, to make everything more clear in our light background, symbols and characters are black for higher contrast. This way they stand out more.

3.3 Create a 2nd "napkin" design, hi-fi.

With all these charts, we elaborated a napkin design of the first slide, obtaining the figure 10:

Note that the colours of the doughnut charts here are the ones that will be used in the dashboard, and the same with the map of Barcelona.

In the case of the second slide, we came up with the idea from the napkin of the first slide, going from filtering by neighborhoods to working with each Airbnb individually, so we started working on it in Tableau. The first napkin design is the one shown in figure 11

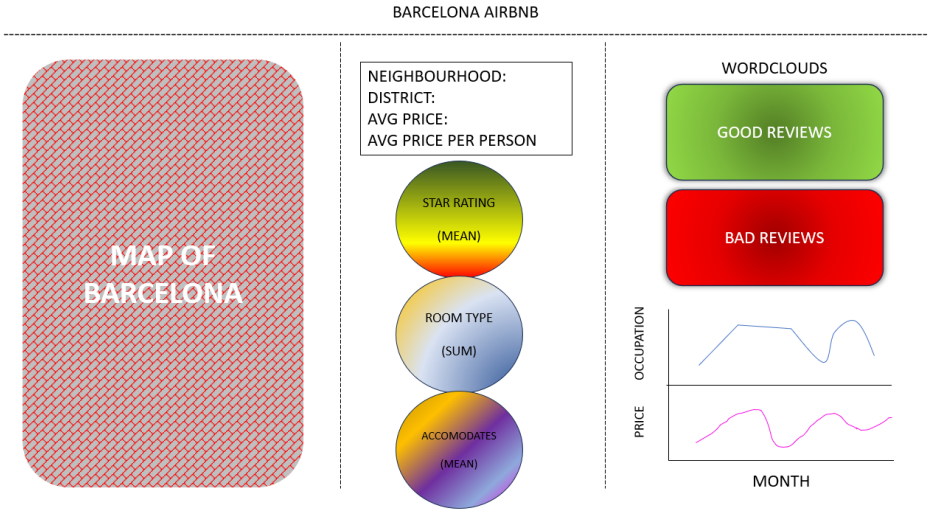


Figure 10: *Napkin design of the first slide. Hi-fi*

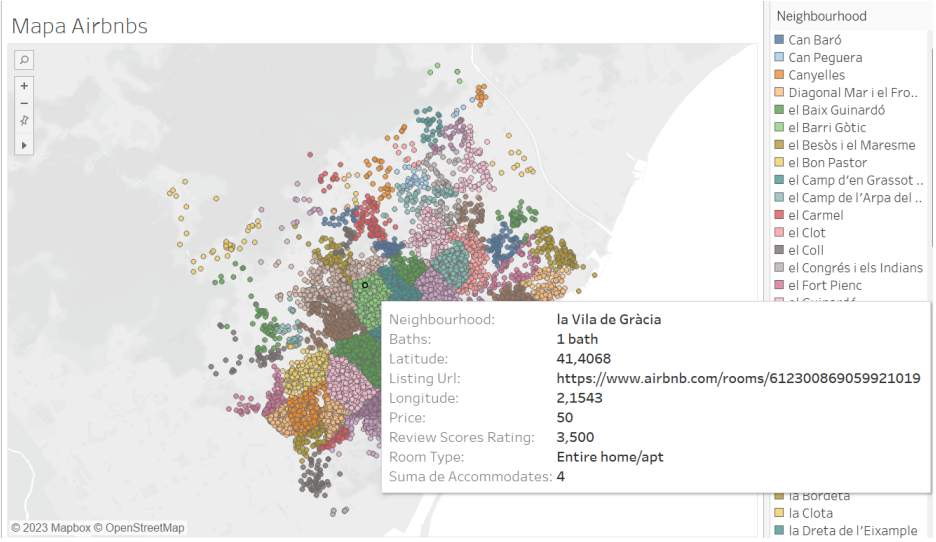


Figure 11: *Napkin design of the second slide. Hi-fi*

4 Implementation

Once we built the first slide of the dashboard, we realized the distribution designed in 10 was not the optimal one, because the Gestalt Law about closure was not satisfied. G6.1 enuntiates the following: "Place symbols and glyphs representing related information close together". That is why we decided to reorder the slide, obtaining figure 12.

With this distribution, the information is divided in 4 modules:

- Information about the neighbourhoods and its distribution (map and name of hthe neighbourhood and district)
- Economic reports (Information about prices and time series)
- Analysis of the reviews (wordclouds and rating chart)

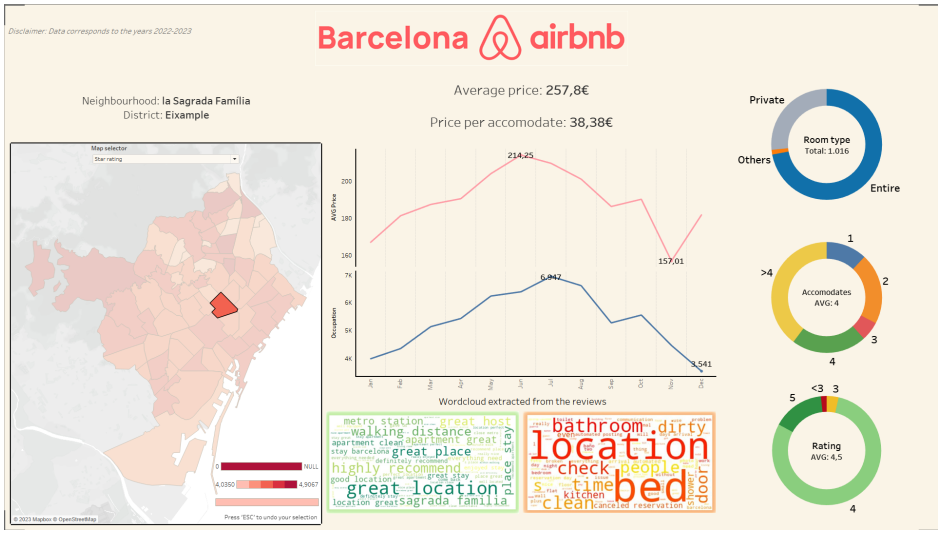


Figure 12: *Final design of the first slide*

- How are the accommodations in the neighbourhoods (room type and accommodation charts)

As explained before, excepting the wordclouds, all charts and information interact with the map of the city, so that the information shown corresponds to the neighbourhood selected on it.

In the second slide, however, everything interacts with the map. There is no static information, as if it occurred in the first slide with the wordclouds. The final dashboard is shown in figure 9