Università degli Studi di Milano-Bicocca
Corso di Laurea Magistrale in Data Science

The Airbnb experience in Milan

ENRICO MANNARINO 850859
SIMONE MASSARDI 812455
GIORGIA PRINA 858740

ACADEMIC YEAR 2022/2023

Contents

Introduction

Online booking platforms are a service used daily by millions of users worldwide. They allow you to easily book flights, hotel rooms, private accommodations, cars, etc.

Certainly, the development of mass tourism is one of the main factors behind the success of these platforms. Among these, one of the most important is Airbnb, which we have taken as a reference in this work to analyze the availability of short-term private accommodations in the city of Milan.

0.1 What is Airbnb?

Airbnb is an online platform for short-term rental of various types of private accommodations founded in California in 2008. Originally named Airbedandbreakfast.com with the aim of offering an alternative to the saturated hotel market, it has rapidly expanded worldwide and today it is present in 100,000 cities and has more than 7 million listings. In 2020, Airbnb's revenue was around 10.2 billion dollars, making the company one of the largest online booking portals.

Through the platform, users can choose from various types of accommodation, from a single room to the entire home and guide their choice based on price, available services and especially based on the reviews left by other users. To become a host, on the other hand, it is sufficient to register the accommodation that you intend to make available on the platform. Hosts who meet certain specific criteria established by Airbnb are assigned the status of *superhost* as a guarantee of quality and reliability.

0.2 Our case study: the city of Milan

As students of a university in Milan, we think that an analysis of the Airbnb phenomenon in the city of Milan is interesting for several reasons. Milan is one of the main tourist destinations in Italy (in 2019, the last year before the COVID pandemic, the city recorded about 8 million arrivals[1]) and also the site of numerous international events that attract thousands of people such as the Milano

Fashion Week, the Salone del Mobile and the upcoming Winter Olympics in 2026. Currently, the number of listings on Airbnb is around 19,000[2]. Tourism and events are not the only factors contributing to this number; the pandemic has stimulated the growth of so-called digital nomads - workers who choose destinations for short periods of time to work remotely. In addition, in an interview[3] with Sole24ore, Airbnb Italy CEO Giacomo Trovato reveals that recent inflation is one of the main reasons behind the decision to become a host; in fact, in the second quarter of 2022, there was a 60% increase in new hosts compared to the same quarter of 2021.

1. Dataset presentation

For the development of the project, we based ourselves on two datasets, both made available by the Inside Airbnb website[2], an independent project created with the aim of analyzing the impact of Airbnb on residential communities. The first dataset consists of a list (updated on September 2022) of all Airbnb listings in the city of Milan; it contains 19,248 listings with 75 variables each. The variables contain all the information related to the listing: general information about the accommodation, information about the host, reviews, price, geographical data. Of these, we selected 14 that we considered to be the most substantial for the purposes of the project and which will now be presented in more detail:

- Host: host registration **date** and whether they are a **superhost**.
- Neighbourhood: the name of the **neighbourhood** where the accommodation is located.
- Geographical data: **latitude** and **longitude** of the accommodation.
- Price: the **price** in Euro for a single night.
- Room type: on Airbnb, you can find different **categories** of rooms (entire place, hotel room, private room, shared room).
- Overall review: a rating from 0 to 5 expressing the **overall satisfaction** of the guest.
- Specific review: a rating from 0 to 5 expressing the guest's satisfaction with six different categories: **accuracy** (does the accommodation correspond to the information provided by the host?), **cleanliness**, **check-in** (how simple is it?), **communication** (does the host respond promptly?), **location** (was the guest informed about safety, transportation, points of interest and special considerations such as noise or other situations that may affect their stay?), **value** (was it worth it?).

The second dataset, on the other hand, contains the price per night for the period from September 14, 2022 to September 14, 2023, and the availability of the accommodation for each listing.

Finally, we have joined a GEOJSON file to the first dataset, containing the geographical data of the boundaries of each neighbourhood, useful for dividing the city territory as we will see later.

Both datasets look clean and show no critical missing values in correspondence of the variables that we considered. However, after a first explorative analysis, some records stood out for having an unusually high price (€9999 or higher). In order to verify the nature of this data we checked the corresponding listing on Airbnb and, as it turns out, some accomodations are never available on any future date and some actually have a much lower price. We then assumed that those prices are purely indicatives and proceded to exclude them from our evaluations by setting as a range for the prices the average of all the prices $\pm$ two times the standard deviation.

2. The creation of the data visualization

The work we intend to realise consists of a story characterised by five dashboards. This chapter will analyse the individual graphics that make them up, describing the communicative purposes of each one.

2.1 Goal and target

The main goal of this work is to analyse the Airbnb market in the city of Milan. Starting with assumptions based on our knowledge, we intend to verify whether or not these are satisfied by the data at our disposal. We want to give a sense of information through immediately comprehensible graphs, allowing users to answer their own questions, which can range from the distribution of listings (*"Which neighbourhoods have the most listings?"*) to their respective prices (*"Which is the most expensive neighbourhood?"*), from host information (*"What was the year with the most new host registrations?"*) to user reviews (*"Which neighbourhood has the best reviews?"*). Each dashboard will then be self-explanatory in order to enable targeted research.

The target audience are ordinary users and possible new hosts who, before registering and making their accommodation available on the website, prefer to inform themselves on market trends, studying prices and availability for each neighbourhood in the city.

2.2 Dashboard 1: Airbnb listings in Milan and distinction by room type

In order to answer the first question that comes to mind, namely "Which is the neighbourhood in Milan with the most Airbnb listings?", it was created a choroplethic map (Figure 2.1) that divided the city into neighbourhoods according to the geospatial data available. The darker the colour associated with the delineated neighbourhood becomes (in our case approaching red as we worked on a divergent colour scale ranging from green to red), the more Airbnb are present. Thanks to this we see that in the Buenos Aires-Porta Venezia area there is the highest concentration of Airbnb,

with a total of 1444, as well as in central neighbourhoods such as Duomo (1109), Navigli (858), Ticinese (827) and so on. These areas are very close to the centre, as we expected. In fact, the further you move towards the suburbs, the more the number of Airbnb decreases. The lowest number is recorded for Ronchetto delle Rane, a neighbourhood in the far south of Milan, with only 2 listings.

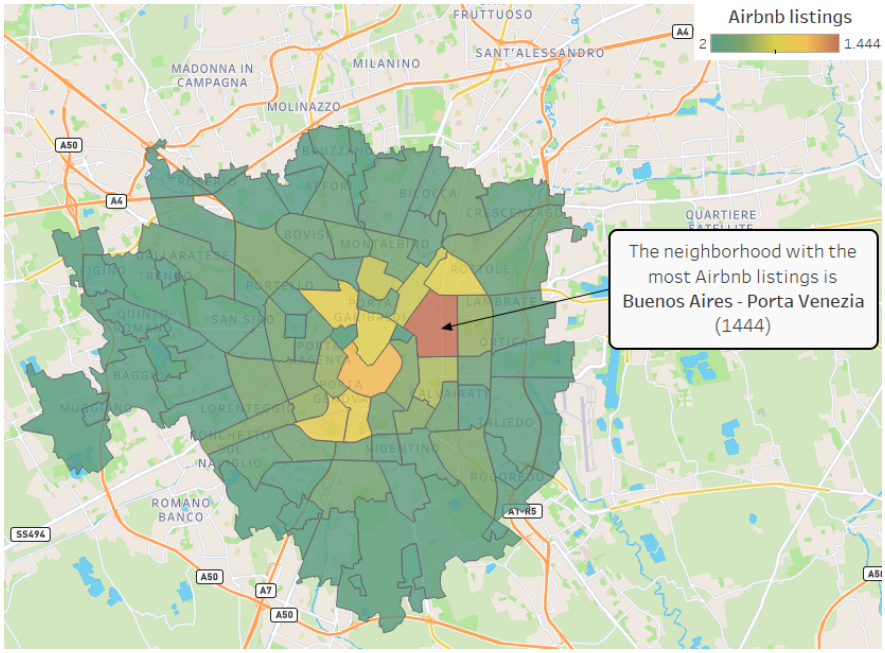

Figure 2.1: Milan neighbourhoods by number of Airbnb listings

Analysing the listings by type of housing solution proposed, we produced a donut chart (Figure 2.2) and a horizontal stacked bar chart. The first is very similar to a pie chart and shows the total of Airbnbs for each category of the room type variable, i.e. entire home/apartment, private room, shared room and hotel room, which take on different colours in the display. If you are interested in one or more neighbourhoods in particular, you can select them using a drop-down filter at the top, for which the total will be recalculated. Considering the entire city with a total of 19,185 listings, we see that most of the proposed solutions (about 79%) are Entire home/apartments, followed by Private rooms (about 20%) and Shared rooms with Hotel rooms, which together make up about 1% of the total.

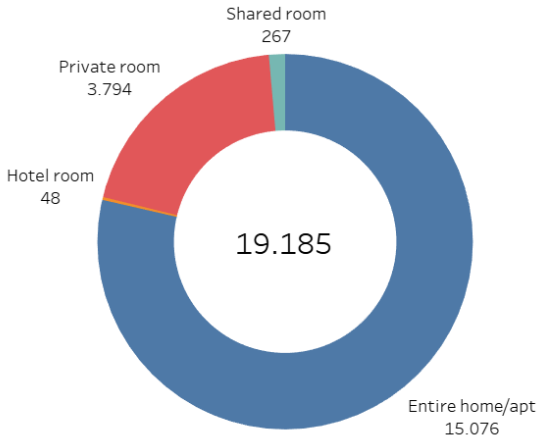

Figure 2.2: Proportions of room types

In the horizontal stacked bar chart (Figure 2.3), on the other hand, each bar, subdivided and coloured according to the categories of rooms present per neighbourhood, is stacked on top of each other, and the total length indicates the number of Airbnb listings present. Since the filter is in common with the donut chart, it is possible to directly display the numbers of each category for as many neighbourhoods as we want.

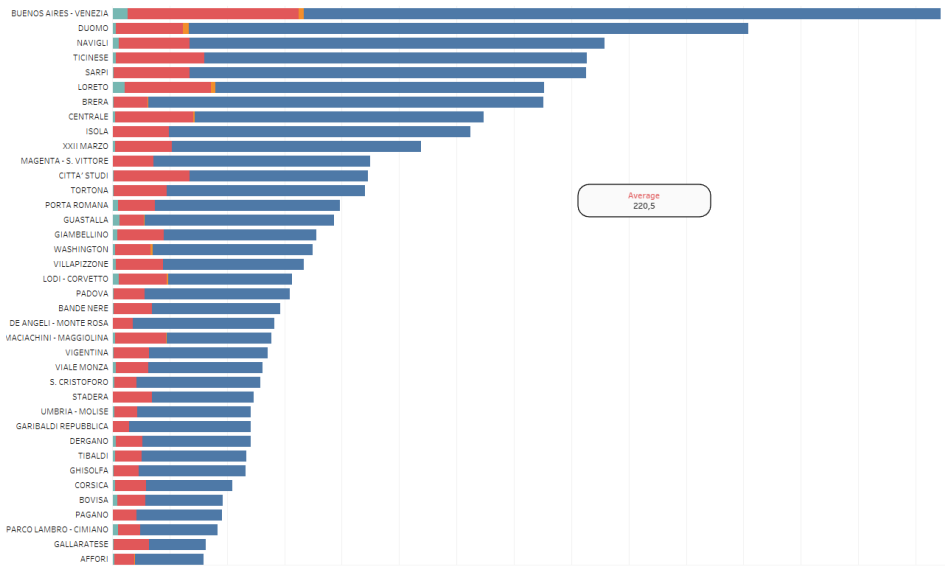

Figure 2.3: Airbnb listings by neighbourhood and room type

2.3 Dashboard 2: analysis and distribution of average Airbnb prices in Milan by neighbourhood

The second dashboard shows the distribution of average Airbnb prices in Milan by neighbourhood. The first time we've represented the choropleth map (Figure 2.4) we noticed that there was a little neighbourhood called "Ronchetto delle Rane" whose average price was higher than we expected. Our curiosity led us to investigate the fact and we discovered that, for that neighbourhood, only two structures are registered and all of them are entire home.

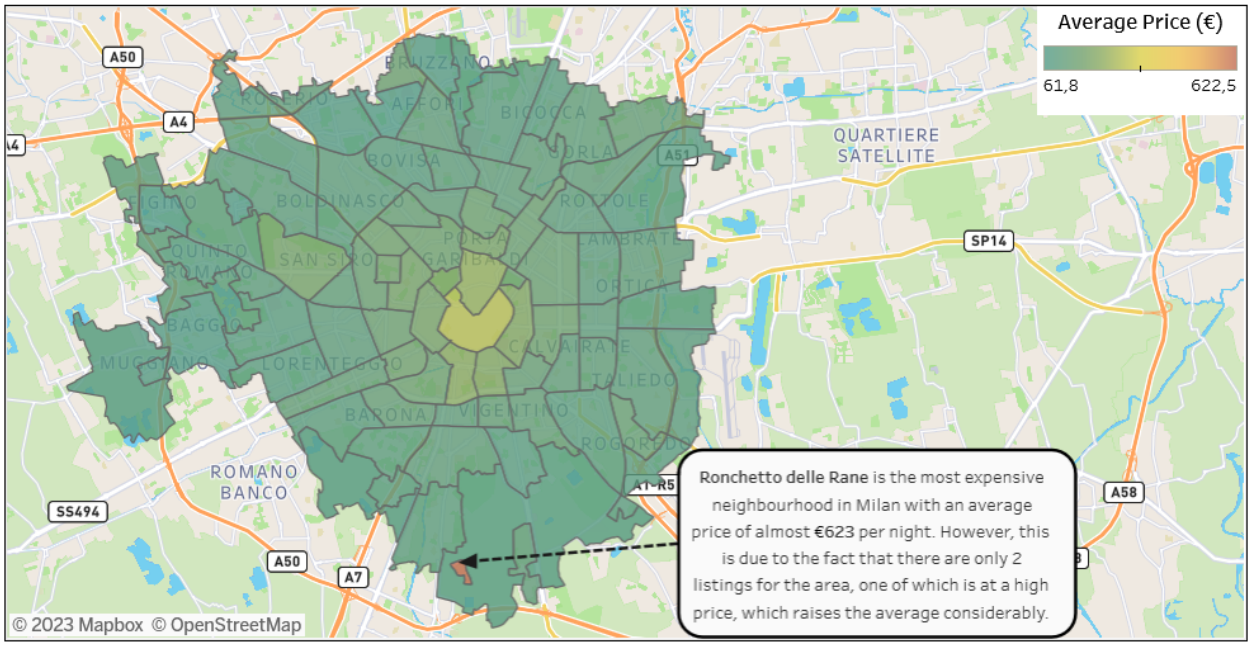

Figure 2.4: The distribution of average Airbnb prices in Milan by neighbourhood considering Ronchetto delle Rane

To solve the visualization problem, we have excluded Ronchetto delle rane from the map; then the colors were redistributed and showed a result more similar to reality. Duomo has therefore become the most expensive neighbourhood and the nuance of color has shown that the central areas are precisely those with higher average prices. We have done this in order to give more prominence to the neighbourhoods with the most listings; if we had only considered these neighbourhoods, we would have obviated this problem, but for the sake of completeness in the map visualization, we

have nevertheless decided to display them all by making this clarification.

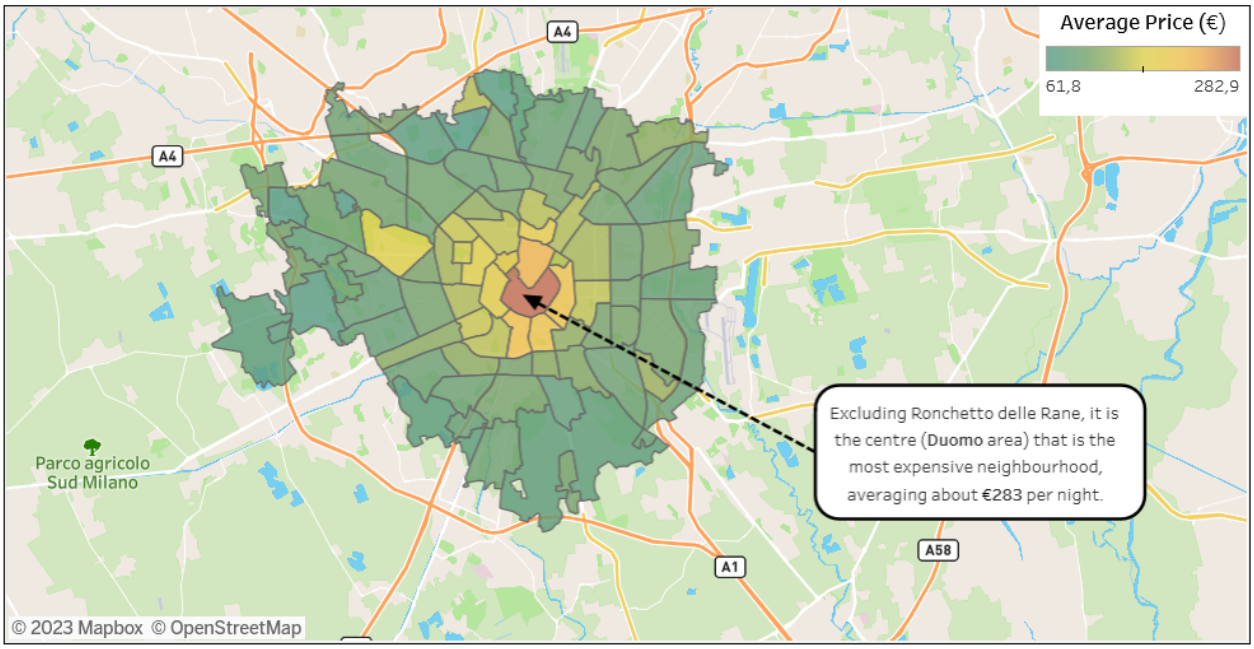

Figure 2.5: The distribution of average Airbnb prices in Milan by neighbourhood excluding Ronchetto delle Rane

For this reason we thought it would be wise to propose a box plot, so that the user could combine it with the maps and better understand the variability of prices. We choose to order the neighbourhoods from the most expensive in average to the less expensive. To add a greater degree of information as well as completeness, we have allowed the user to filter by room type. The filter provides only one option to display only one category at a time and compare prices by neighbourhood.

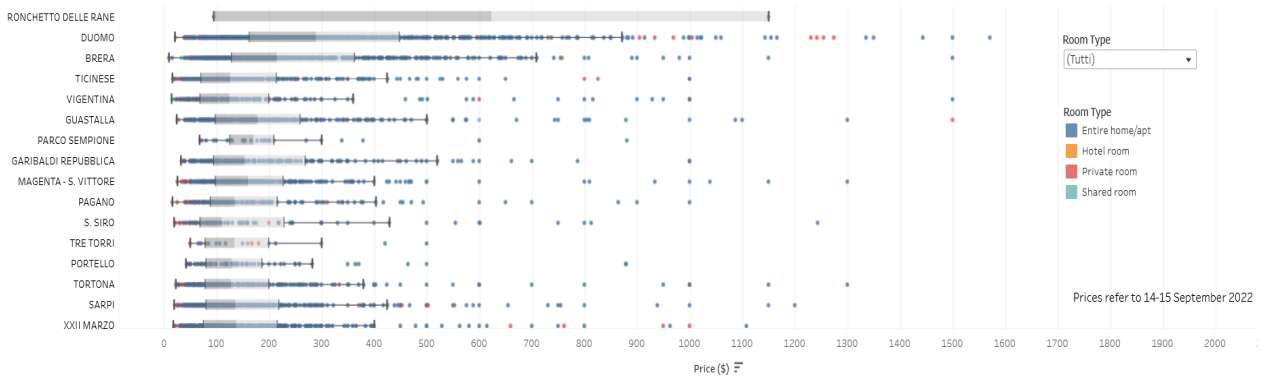

Figure 2.6: Box plot of Airbnb prices in Milan neighbourhoods by room type

2.4 Dashboard 3: price analysis of Airbnb listings in Milan for the coming year

The third dashboard is a representation of the listing prices for the coming year. The time frame foreseen by our data source goes from September 2022 to September 2023. After a few attempts we decided to display the average daily prices as they made us discover not only interesting peaks in conjunction with some specific dates, but also a very clear weekly pattern (Figure 2.7). The time series shows the first pick on September 2022, it's a relative small one in conjunction with Milano Fashion Week. The second peak occurred in conjunction with the New Year's Eve and the last one, which is the highest, during the Salone del Mobile. As it seems Salone del Mobile is the main attraction of the year in Milan. Visitors come from all over the world to participate and this exodus leads to a considerable increase in hotel and flat prices every year. In fact, there have been several analyses [4] concerning mainly the increase of short-term rentals by real estate companies. This phenomenon affects the rental market and all related platforms such as Airbnb, which is why the figures for 2023 tend to reconfirm this trend and predict a peak on those very dates.

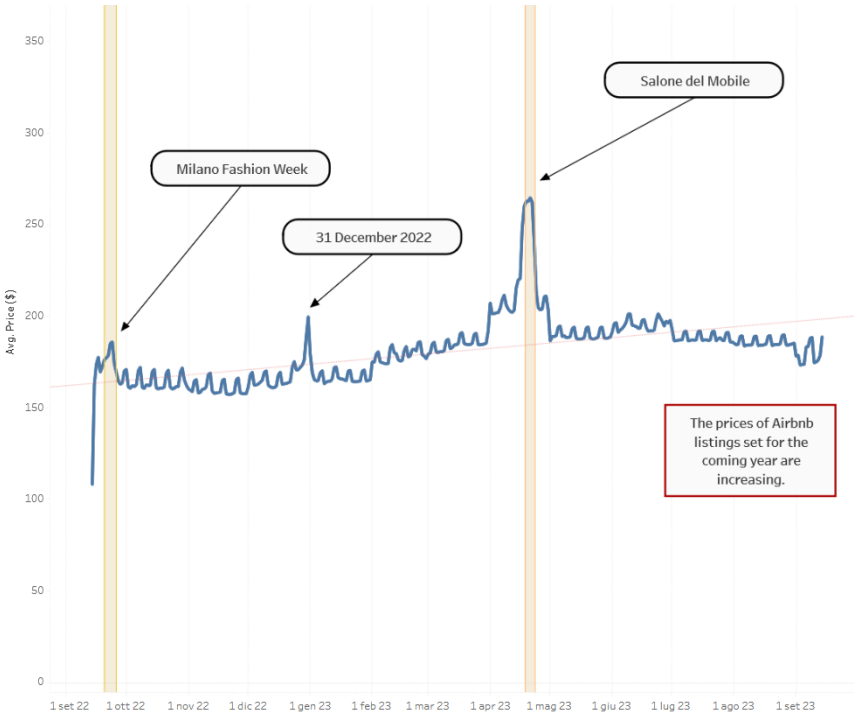

Figure 2.7: Average daily price analysis of Airbnb listings in Milan for the coming year

The other significant element for our analysis is well represented in the bar graph (Figure 2.8). As we could expect prices increase over the weekend on average. We wanted to create a weekly focus to show the repetition of this pattern throughout the time series.

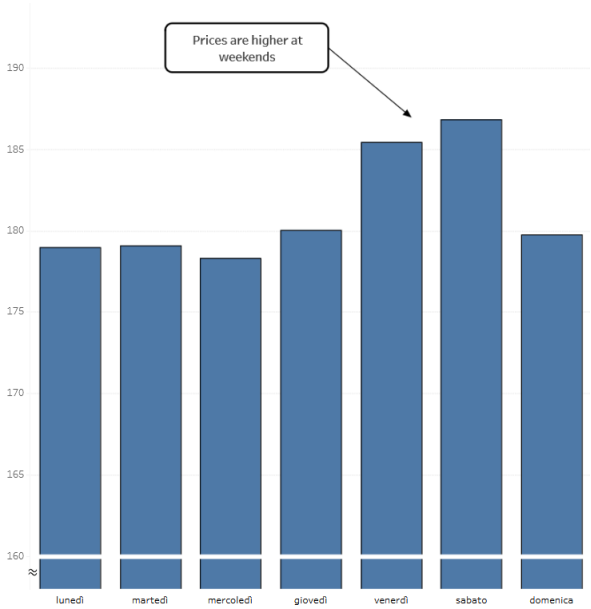

Figure 2.8: Average prices by day of the week of Airbnb listings in Milan

2.5 Dashboard 4: hosts registration and analysis

Let us now move on to the part more dedicated to the hosts, i.e. the owners of the listings; this is usually the owner of the accommodation or the person who lives there, and who must deal with any problems that may arise during the guests' stay. A dashboard has been created in which we find in the foreground the monthly historical series of registrations of new hosts in Milan on the Airbnb website (Figure 2.9), starting from February 2009, the date of the first registration, until September 2022. Thanks to the table on the side showing the total number of registrations per year, highlighted according to a grey colour scheme, it is useful to note that the peak of registrations was reached in 2015 (+3536), probably beacause of the Expo [5].

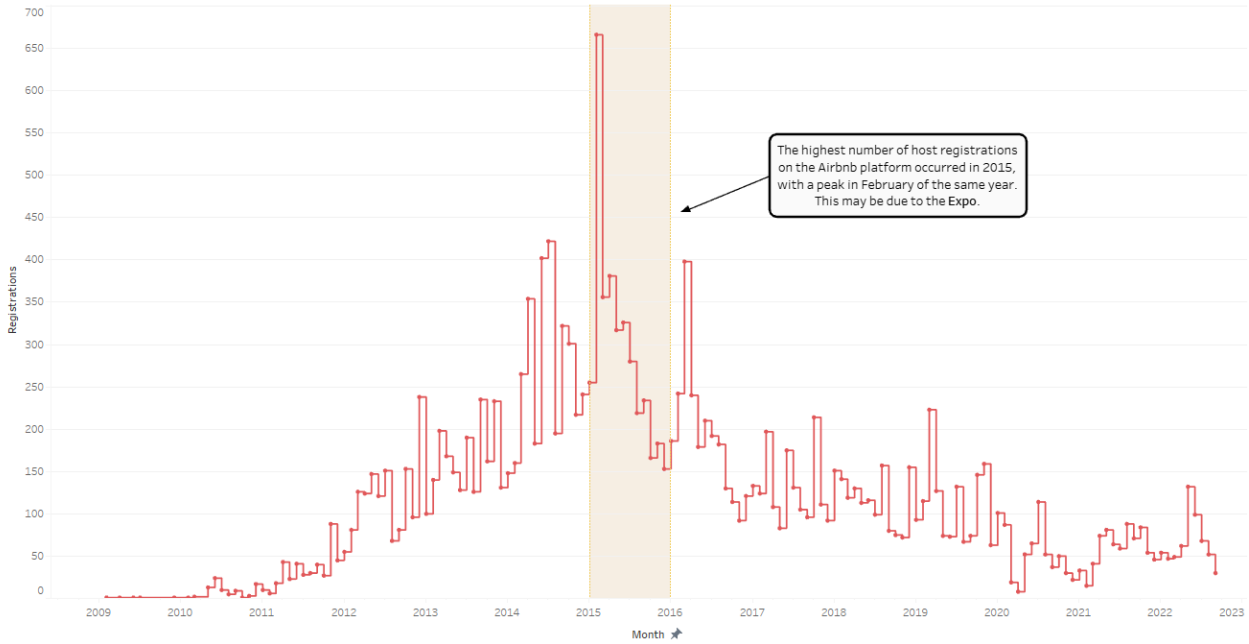

Figure 2.9: Monthly time series of new host registrations (2009-2022).

To accentuate the decline in registrations in recent years, a similar graph was also made, but showing the cumulative absolute frequencies from month to month (Figure 2.10). Here it is much more evident how the slope of the curve was much steeper at the turn of 2012-2015 and started to come to a halt soon after, until 2020 when registrations halved, probably due to Covid-19 pandemic.

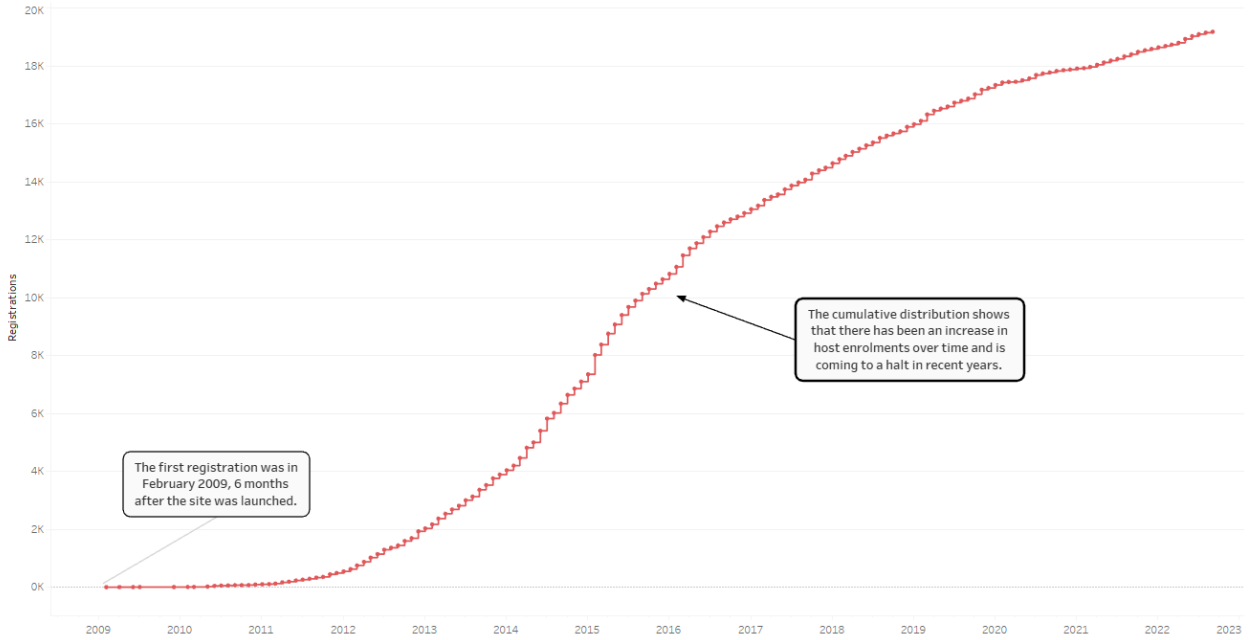

Figure 2.10: Monthly time series of cumulative absolute frequencies of new host registrations (2009-2022).

Secondly, it was made a donut chart to show the proportions between hosts and superhosts (Figure 2.11); specifically, to become a superhost, four requirements must be fulfilled [6]:

- **Maintain a high overall rating**

  You must have an overall average rating of 4.8 or higher, which is calculated based on your Airbnb guests' reviews over the previous year.

- **Have experience**

  In the last year you must have hosted at least 10 stays or, in the case of long-term bookings, at least 3 stays totalling 100 nights.

- **Avoid cancellations**

  You must have cancelled bookings less than 1% of the time, except in extenuating circumstances.

- **Being responsive**

  You must have replied within 24 hours to at least 90% of the new booking-related messages.

We note from the graph below that about 18% of the total listings in Milan are by superhosts .

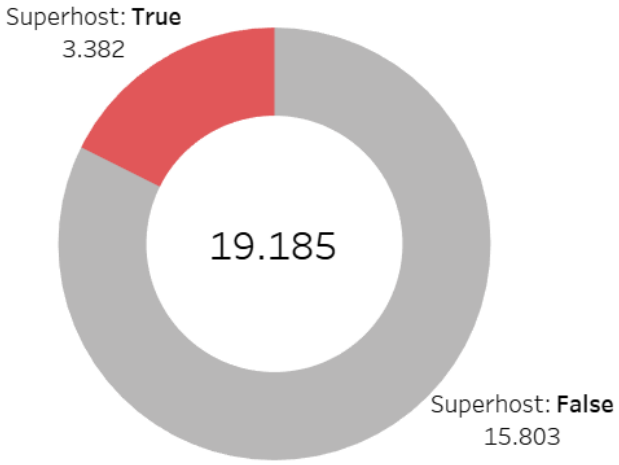

Figure 2.11: Proportion of hosts (grey) and superhosts (red).

Finally, we find the 3 hosts with the most listings listed: *Italianway* (410), *The Best Rent* (139) and *Cleanbnb* (125). These are external websites that act as intermediaries and provide management services for short-term rental properties, enhancing the visibility of listings on online platforms such as Airbnb.

2.6 Dashboard 5: customer reviews by neighbourhood

In the last dashboard we wanted to represent customer satisfaction through three different graphs. The choropleth map offers us, as before, a spatial view of the average overall grade per neighbourhood (Figure 2.12). Even on a visual and color level, a certain homogeneity and a low variability of votes between the different neighborhoods can be noted.

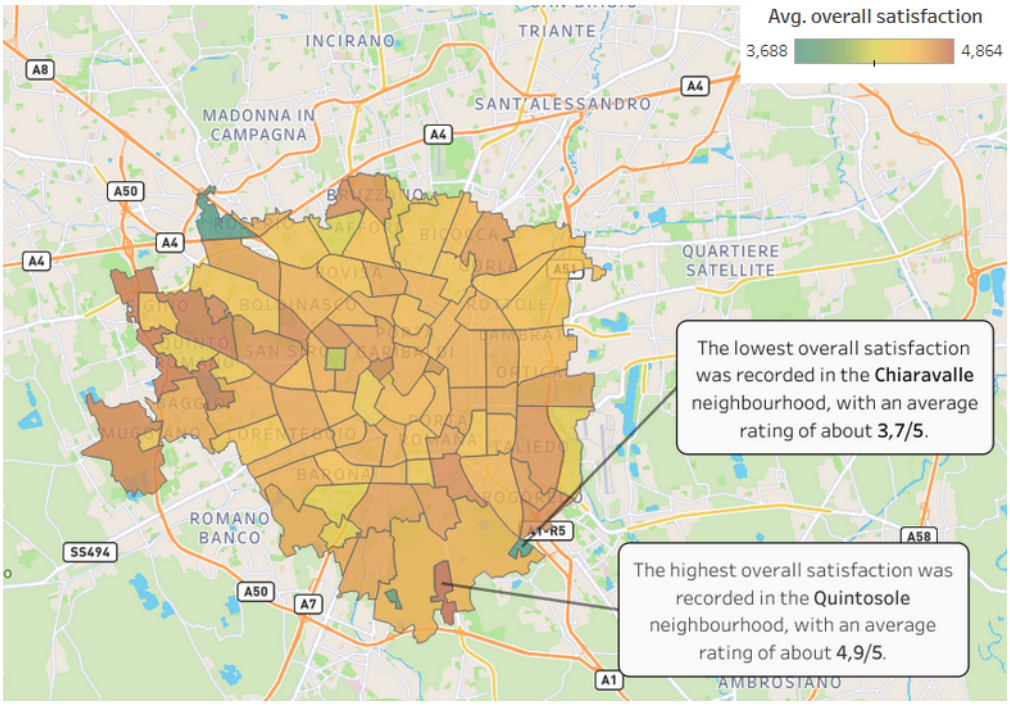

Figure 2.12: Customer reviews by neighbourhood

This statement is confirmed by both the radar chart and the box plot. With the radar chart (Figure 2.13) the user can compare different neighbourhoods through the six typical evaluation methods of the Airbnb platform, already described in Chapter 1: *Cleanliness*, *Accuracy*, *Check-in*, *Communication*, *Location*, *Value for money*.

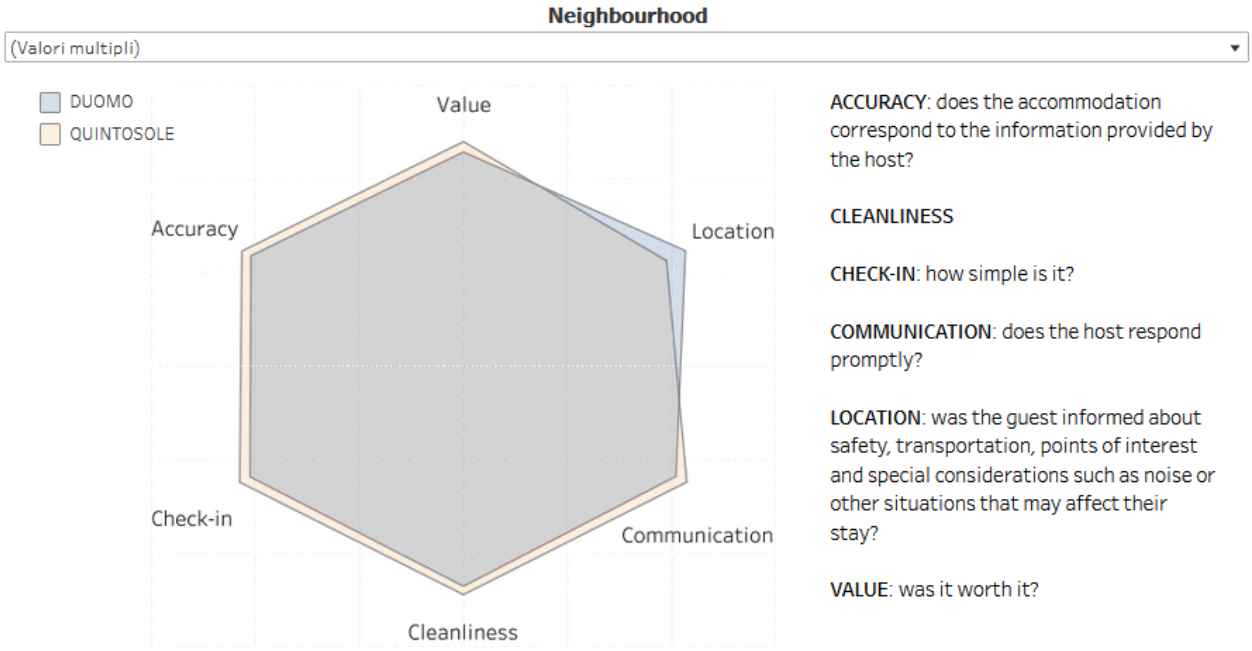

Figure 2.13: Radar chart of the six review parameters for Duomo and Quintosole neighbourhood (ex.)

The second one offers a visual explanation of the real variability of the votes for each evaluation method. With the box plot (Figure 2.14) the user is able to compare each method filtering by neighbourhood providing some additional information such as the median and also the quartiles.

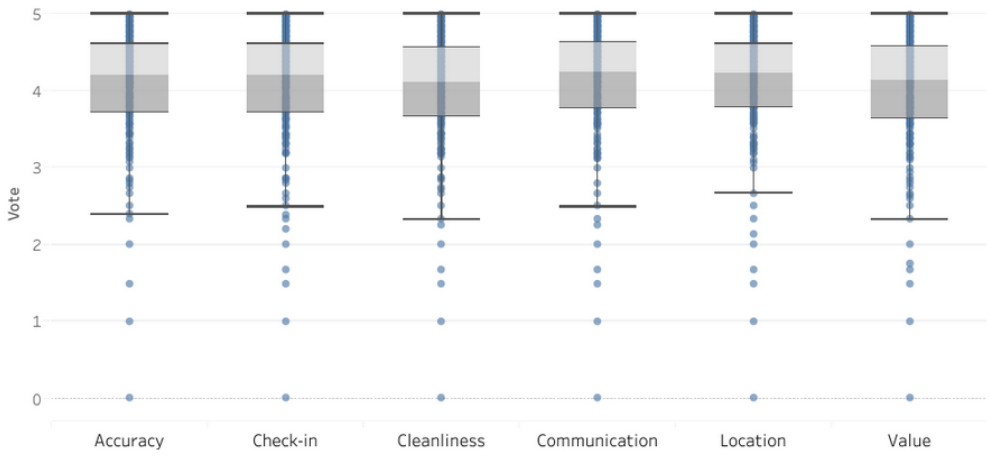

Figure 2.14: Variability of the six Airbnb evaluation methods by neighbourhood

Considering all the neighbourhoods, the result is that there is a real small difference between the distribution of each vote and that the median is around 4 for each method. The result is perceptible at a glance from the colors of the map.

3. Evaluation

In this chapter we show the processes through which we evaluated the quality of our data viz project and the results that we obtained. We made different people interact with the visualizations and run three tests: an heuristic evaluation, a psychometric questionnaire and a user test. In the following sections we go deep into each one and present the results.

3.1 Heuristic evaluation

The heuristic evaluation was conducted on a sample of 4 people, two of whom were familiar with charts and statistical data representation. The evaluation was conducted individually with each person by briefly presenting the project topic and then having them interact with the visualizations, commenting on their experience with personal opinions and making questions. The results of the evaluations are the following.

3.1.1 Results

With regards to the dashboards, the first one regarding the number of listings per neighborhood (Figure 2.1, 2.2, 2.3) did not pose any difficulties in understanding from any of the users. We decided to remove the red line representing the average number of listings that was initially present in the bar chart in Figure 2.3 as users commented it is not very informative, in particular when considering the comparison between few neighbourhoods. As for the second dashboard, less expert users asked for the meaning of the chart in Figure 2.6. However, a brief clarification on the functioning of box plots was sufficient to make the visualization positively explanatory, so we did not consider it necessary to make changes.

Dashboard number 3 was clear and immediate for everyone.

Dashboard number 4 raised some doubts about the choice of color used for the donut chart (Figure 3.1). Users found the distinction between normal hosts and superhosts not very functional. Therefore, we decided to increase the emphasis on the portion of superhosts compared to the total, obtaining the final chart in Figure 2.11.

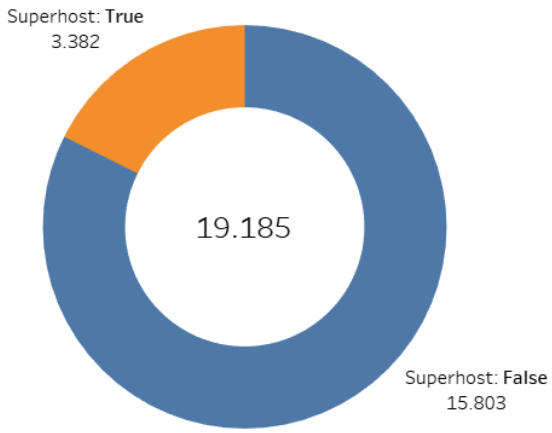

Figure 3.1: Previous version of donut chart

Finally, the final dashboard highlighted the following issue: users instinctively clicked on the map in Figure 2.12 to get information on individual neighborhoods but nothing happened. We then decided to set the map as an active filter for the box plot in Figure 2.14; this way, users can click on a single neighborhood on the map and the box plot updates automatically.

3.2 Psychometric questionnaire

Continuing with the assessment of the quality of data viz, we administered the psychometric questionnaire to 24 subjects, in the age range 20-30 years, using the Cabitza-Locoro scale [7]. The latter consists of rating the work on a scale ranging from 1 (very little) to 6 (very much) the following adjectives:

- Usefulness
- Beauty
- Clarity
- Informativity
- Intuitivity

In addition, using the same scale, it requires a value as a whole to be given to the data viz.

3.2.1 Results

Below are the results obtained from the data collected (Figure 3.2 and Figure 3.3):

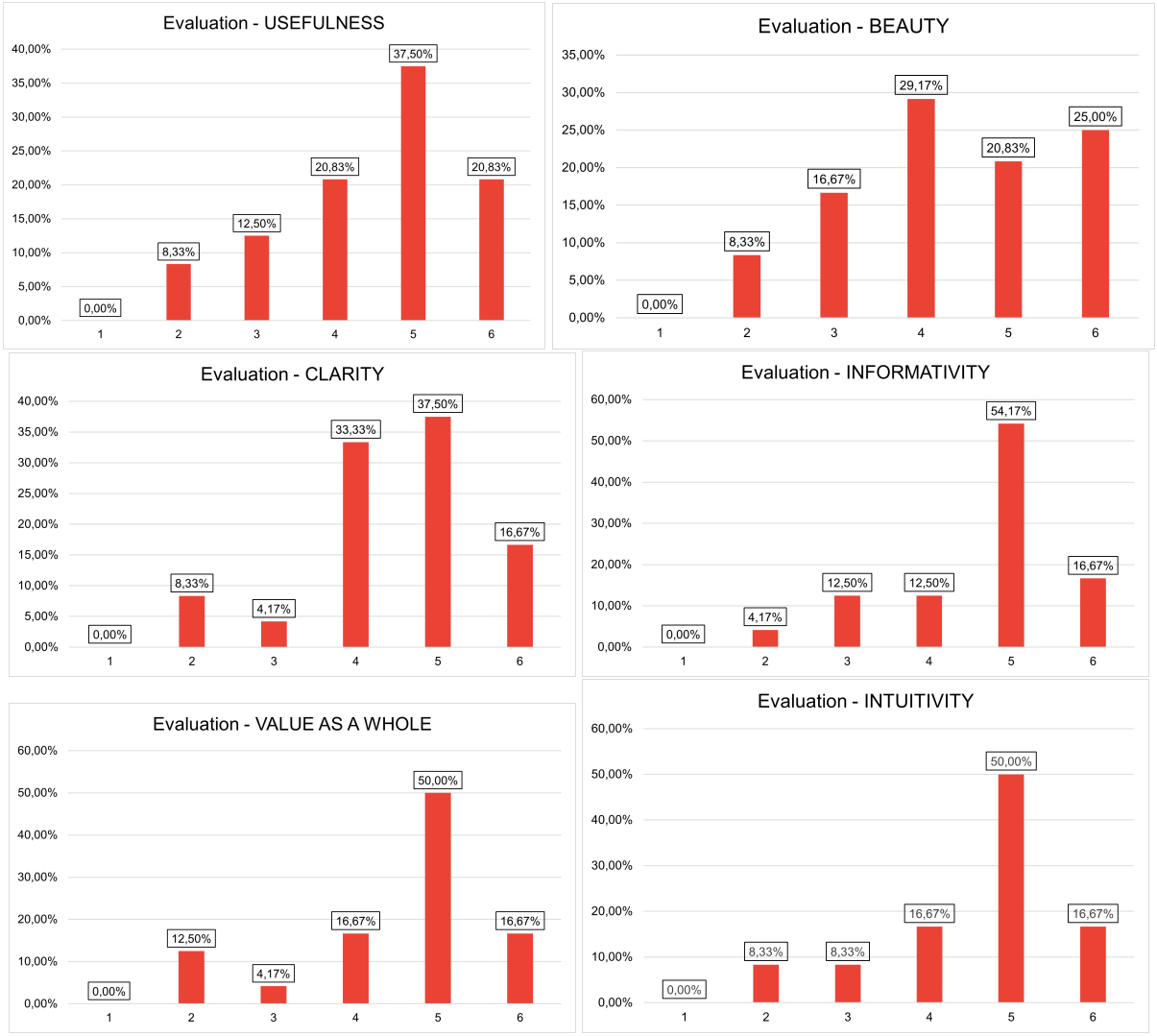

Figure 3.2: Results of the psychometric questionnaire

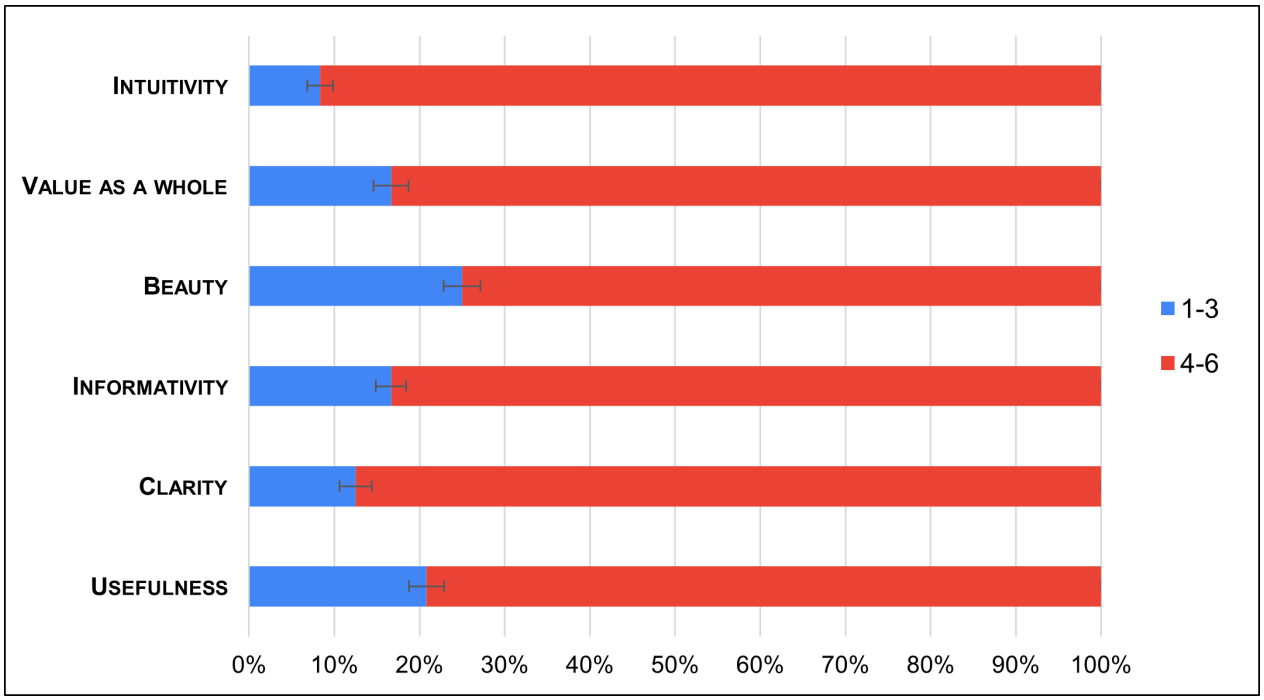

Figure 3.3: Results of the psychometric questionnaire

The results obtained are very positive in all areas, mainly with regard to intuitivity and clarity; less positive is the beauty of the data viz. In particular, in the Figure 3.3 is shown a horizontal 100% stacked bar chart with confidence intervals in which the values are grouped into two classes.
Finally, we conclude with the correlogram of the data collected (Figure 3.4), which is useful for highlighting the (Pearson) correlations between the parameters considered. In this case they are all positive and the most significant are the correlations between intuitivity and informativity ($R = 0.89$) and between clarity and intuitivity ($R = 0.86$), while the least correlated are beauty and usefulness ($R = 0.49$).

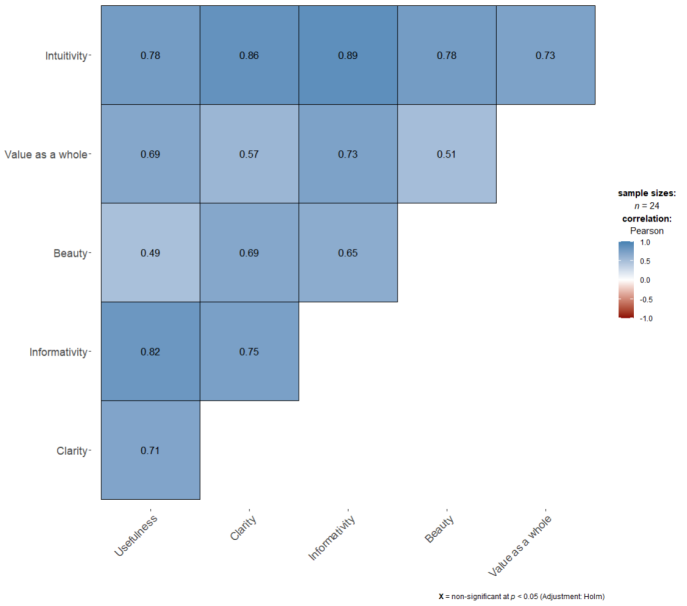

Figure 3.4: Correlogram of the results of the psychometric questionnaire

3.3 User test

For the user test we selected a sample of 6 users of different ages, genders and experience. These values are summarized in the Table 3.1 and graphically represented using pie charts (Figure 3.5).

Table 3.1: Users features

ID	Gender	Experience level	Age
U1	F	Inexperienced	Over 46
U2	M	Expert	18-25
U3	M	Expert	18-25
U4	M	Inexperienced	Over 46
U5	M	Very experienced	18-25
U6	F	Inexperienced	18-25

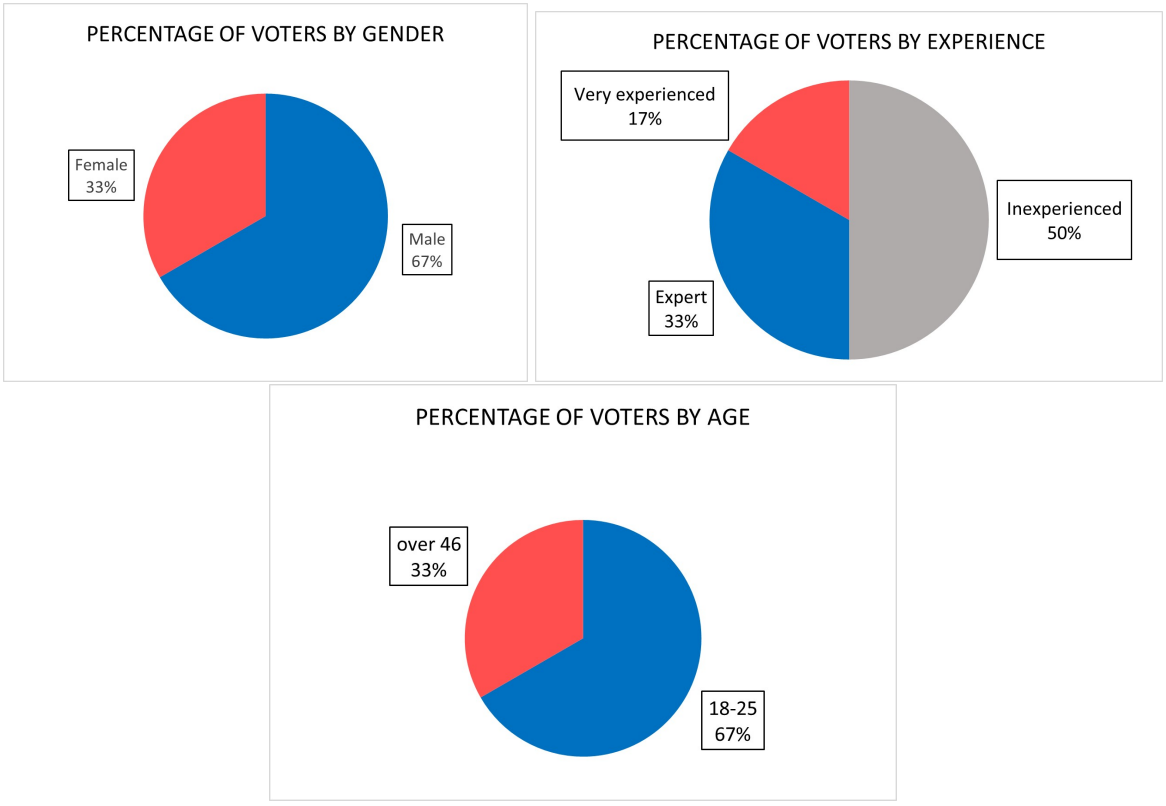

Figure 3.5: Results of the user test

The six users were asked 3 questions of different difficulty, both to understand how they interacted with graphs and legends and to understand which visualizations had the most impact. The first two tasks are simpler and more immediate, the third requires a comparison and greater attention. All three are described in the Table 3.2.

Table 3.2: Description of tasks

	Goal	**Task**
Task 1	Interaction with the map and bar chart	What is the neighbourhood with the lowest number of listings?
Task 2	Comparison between the two maps	What is the neighbourhood with the highest average prices?
Task 3	Interaction with filters and legends	Choose 2 different neighbourhood and find the accuracy with the highest median.

To evaluate the efficiency we measured the execution times of each task performed by the user. The Table 3.3 shows the times expressed in seconds and the Table 3.4 sums up mean and median for each task:

Table 3.3: Time to complete the tasks (in seconds)

ID	**Task 1**	**Task 2**	**Task 3**
U1	38	3	80
U2	33	5	50
U3	10	12	90
U4	10	10	83
U5	23	4	24
U5	29	9	33

Table 3.4: Summary of the execution times (in seconds)

	Task 1	**Task 2**	**Task 3**
Average	23,83	7,17	60
Median	26	7	65

3.3.1 Results

To show the results obtained we built a violin plot (Figure 3.6), which shows us the execution times for the 3 tasks:

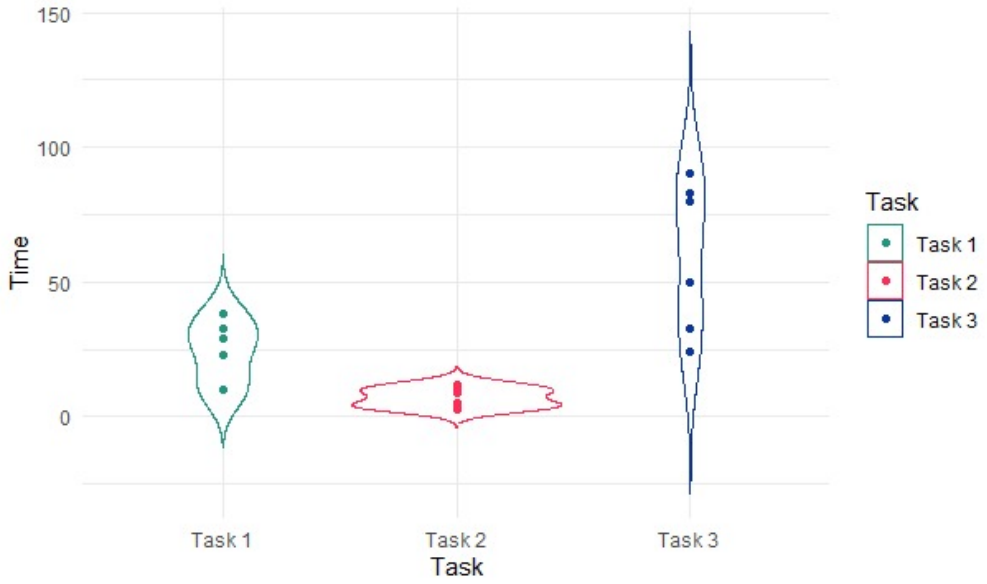

Figure 3.6: Distribution of task execution times

The graph is composed of both a violin plot and a scatterplot. We can see how the first two tasks (the simplest) are represented by more flattened violins, i.e. they have less variability. The difference between non-expert users and very experienced users is of little importance. On the contrary, the blue violin has a high variability and the distribution of points is much more varied. The third task seemed to be the one that required a higher attention span and a more trained eye.

To evaluate the effectiveness, however, we measured the error rate for each of the three tasks submitted to users. The results are shown in the following Table 3.5:

Table 3.5: Division of results (green, blue, orange) by task and type of answer given (correct, with help, wrong)

ID	Task 1	Task 2	Task 3
U1	38	3	80
U2	33	5	50
U3	10	12	90
U4	10	10	83
U5	23	4	24
U6	29	9	33

All the green cells indicate that the task was performed autonomously. The blue ones indicate that the user asked for a clarification or some help. The orange ones indicate that the user made a mistake in order to complete the task. The Figure 3.7 below summarizes the results of the table, showing the percentage of answers given independently (green), with help (blue) or with errors (orange) for each task.

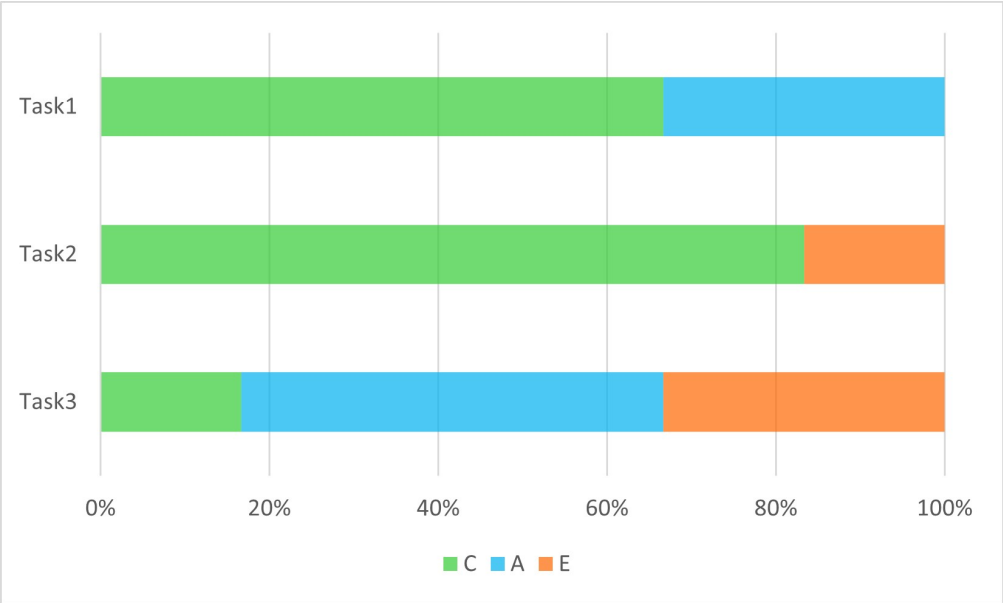

Figure 3.7: Percentage division of results (green, blue, orange) by task and type of answer given (correct, with help, wrong)

In the end the third task was the most complex in terms of execution for the user. The main problem was to choose the right chart from the three on the dashboard and navigate with the right filter. Often the user tried to use the radar chart to make the comparison, but the median was only found on the box plot. The second task seemed to be the most readable. The question we chose was deliberately very simple, because the objective of that slide was exactly to explain why Ronchetto delle Rane was an anomalous detection. Almost all users have added information beyond the simple answer they had to provide. On the contrary, there were no errors in the first one, just a few requests for clarification to find out if it was possible to filter from the map.

Conclusions

In conclusion, following the results obtained from the various evaluation tests we carried out, we can say that we have designed a useful and informative tool that meet our initial intention.

We faced a large number of variables on which we made a thoughtful choice, knowing that we could significally expand the work done so far by allowing us to add more and more levels of detail. This choice allowed us to analyze themes that were absolutely significant and relevant to us and to create small stories within a larger one.

If the initial goal was to propose a visualization that was mostly informative, as well as visually pleasing and useful, during the drafting of the work, additional questions and interesting insights arose that led us to use more data than we initially planned. In some cases, we would have preferred to have even more information. For example, it would have been very interesting to have historical series of prices of years even prior to 2022 and not only a 'picture' of a specific period. With that type of information we could have analyzed the evolution of the Airbnb phenomena from the beginning until now, highlighting the *clou* periods of Milan real estate market transformations. Having available also the content of the reviews written by the users, it would have been possible to carry out a sentiment analysis througout a wordcloud of the most used words to show the themes, both positive and negative, that recur the most.

References

[1] URL: https://polis.lombardia.it/wps/wcm/connect/07290172-0662-4353-b520-53025eb24688/WP+10-2022+-+Flussi+turistici+in+Lombardia_anno2021_cavedo_ed202204.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-07290172-0662-4353-b520-53025eb24688-o2EOLoC.

[2] URL: http://insideairbnb.com/milan.

[3] URL: https://www.ilsole24ore.com/art/airbnb-il-caro-vita-cresciuto-60percento-chi-affitta-casa-grandi-citta-prenotazioni-superano-precovid-AE30E0GC.

[4] URL: https://amp24.ilsole24ore.com/pagina/AES7eLcB.

[5] URL: https://st.ilsole24ore.com/art/casa/2015-04-22/affitti-brevi-vista-expo-160022.shtml.

[6] URL: https://www.airbnb.it/resources/hosting-homes/a/why-strive-for-superhost-status-50.

[7] URL: https://www.mdpi.com/2076-3417/10/18/6189.