Department of Informatics

Master Degree in
Data Science

Analysis of guest superhosts on the AirBnB Milan website

Andrea Malinverno 847340, Longo Gloria 864579,
Giorgio Zacchetti 906074, Mattia Proserpio 846858

Febbraio 2023

Contents

1. Introduction

Every year thousands of people go to Milan to spend their holidays or to carry out work activities, staying in different types of hotels. According to ISTAT statistics carried out on data for the year 2019, Milan is the third Italian city with the highest number of tourists, counting 12,474,278 in that year alone.
Milan is not only a city rich in history and culture but it is also an important commercial centre that hosts hundreds of important events every year that allow it to attract many people from all over the world. Given this huge flow of tourism, the city offers numerous hotel services able to satisfy the demand, ranging from simple hostels to luxurious hotels. These structures are available on dozens of websites that are consulted daily by future visitors, among the most famous there are certainly Booking, AirBnB and Expedia.
In this project an analysis will be carried out on the apartments present in Milan on the AirBnB website in the year 2018.

1.1. Purposes of the project

The aim of the project is to build a representative infographic referring to the data containing information related to the structures present on the AirBnB Milan 2018 site.
First of all, the focus will be on the division into districts of the city of Milan to try to obtain a correlation between reviews and the location of the apartment.
Subsequently, an analysis will be carried out on the variable containing information regarding the average price per night in relation to the boolean variable that describes whether a host is a super host in order to reach conclusions of an economic nature. To do this we will use Tableau, a data visualisation program, with the support of Python for the pre-processing of the data.

1.2. AirBnB

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales. The idea behind Airbnb is simple: Find a way for local people to make some extra money renting out their spare home or room to people visiting the area. Hosts using this platform get to advertise their rentals to millions of people worldwide, with the reassurance that a big company will handle payments and offer support when needed. And for guests, Airbnb can offer a homey place to stay that has more character, perhaps even with a kitchen to avoid dining out, often at a lower price than what hotels charge. Travellers can browse the site for the accommodation that's right for them by applying a number of filters to narrow their searches. Filters include basic factors like location, available dates and guest capacity as well as more particular preferences regarding amenities, facilities, property type, house rules and more.
Once guests have found a property that suits their needs, they can send the host an inquiry (unless Instant Book is enabled, in which case they can reserve the rental on the spot) with any questions they may have about the space or simply expressing that they'd like to rent it.
Hosts have 24 hours to either accept or reject booking requests. If a request is accepted or ignored, the calendar dates of the reservation in question will be automatically blocked for that property.
Founded in 2007, Airbnb has grown a community of over 4 million hosts across 100,000 cities worldwide. Today, more than 193.2 million "nights and experiences" are booked on Airbnb annually.

1.2.1 Reviews on AirBnB

Reviews are key to building trust within Airbnb: they allow hosts and guests to exchange feedback and help the community make informed decisions and understand what to expect when making travel plans. After checkout, hosts and guests have 14 days to write a review. They can also do this for some reservations that are canceled on or after the day of check-in (from midnight, in the time zone of the accommodation).

Reviews will be published once:

- both parties will have submitted their review OR
- 14 days will have passed since it was sent, whichever comes first.If a host cancels a reservation before check-in day, neither party will be able to leave a review for the stay.

In addition to the written review, guests can leave a star review for the following categories:

- **Rating**: guests rate the overall experience
- **Accuracy**: guests rate whether the property matches both the description and the photos on AirBnB
- **Cleanliness**: guests rate the cleanliness of the facility where they stayed
- **Check-in**: guests rate their welcome, check-in should be as fast as possible
- **Communication**: guests value the host's availability, he should respond to messages and requests as soon as possible
- **Location**: guests evaluate the position of the structure with respect to the city, if it is well connected by public transport, if it is a safe area or if it is a noisy area
- **Value**: guest value the cost-quality ratio

1.2.2 Super Host

A Super-host is someone who goes above and beyond their hosting duties and is a prime example of what a host should be like. He is easily recognisable by the badge inserted in his profile and in his ads. Every quarter, AirBnB evaluates the performance of hosts over the past 12 months in relation to all listings included in their account. Each quarterly assessment will be carried out over the course of 5 days starting on the following dates:

- January 1st
- April 1st
- July 1st
- October 1st

If the host meets the requirements of the program by the evaluation date, they will automatically become a Super-host, without having to apply.

To meet the requirements, the host must be the owner of the listing and have an account in good standing and meet the following criteria:

- Must have completed at least 10 stays or 3 reservations, totalling at least 100 nights.
- Must have maintained a minimum response rate of 90%.
- Must have maintained a cancellation rate of less than 1%, except when covered by our Extenuating Circumstances Policy.
- Must have maintained an overall rating of 4.8 (Super-host status considers reviews when both the guest and host have submitted theirs or when the period for writing one has expired, i.e. within 14 days after the end of the reservation, whichever comes first).

2. Districts of Milan

Milan is the second largest city in Italy with an area of 181.8 km^2, it is divided into about 9 districts which will be described and illustrated below.

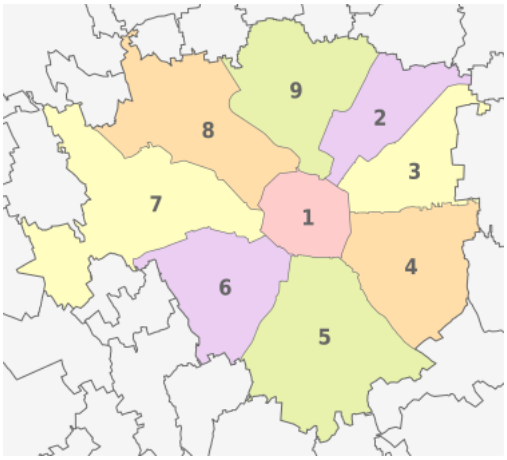

Figure 2.1: Districts of Milan

- Zona 1: Centro storico
- Zona 2: Stazione Centrale, Gorla, Turro, Greco, Crescenzago
- Zona 3: Città Studi, Lambrate, Venezia
- Zona 4: Vittoria, Forlanini
- Zona 5: Vigentino, Chiaravalle, Gratosoglio
- Zona 6: Barona, Lorenteggio
- Zona 7: Baggio, De Angeli, San Siro
- Zona 8: Fiera, Gallaratese, Quarto Oggiaro

- Zona 9: Stazione Garibaldi, Niguarda

In each municipality there is a president and a municipal council, elected at the same time as the mayor and the municipal council.

District 1 of Milan is the most central of all and contains the main attractions of the city such as the Duomo, the Castle and the Galleria Vittorio Emanuele, which is why it is one of the favourite areas for tourists.

3. Description of the datasets

The dataset used for this project was downloaded from the kaggle platform, it contains all the information regarding the apartments available on the AirBnB platform in Milan in 2018 (click here to see it). It consists of 9323 rows and 61 columns.
The attributes available to us are:

- Id: primary key of the dataset (unique code of the apartment)
- Host_id: unique code of the host
- Host_location
- Host_response_time: boolean variable (1 if the host is quick to respond and 0 otherwise)
- Host_response_rate: the percentage of new guest inquiries you responded to within 24 hours in the past 30 days (Hosts have 24 hours to accept or decline your request)
- Host_is_superhost: boolean variable that identifies super hosts (1 for super hosts and 0 for non super hosts)
- Host_total_listings_count:
- Host_has_profile_pic: boolean variable (1 if the host has a profile picture, 0 otherwise)
- Host_identity_verified: boolean variable (1 if host has an verified identity and 0 otherwise)
- Neighbourhood_cleansed: variable that indicates in which district of Milan the apartment is located
- Zipcode: zipcode of the apartment
- Latitude

- Longitude
- Room_type: has only one attribute *Entire home/apt*
- Accommodates: how many rooms has the apartment
- Bathrooms: number of bathrooms
- Bedrooms: number of bedrooms
- Beds: number of beds in the bedroom
- Bed_type: variable containing the type of bed (from 1 to 5): Sigle/twin bed, Double/full bed, Queen bed, King bed, Sofa bed.
- Daily_price: price for a night
- Security_deposit:
- Cleaning_fee: cost of cleaning
- Guests_included: how many guests can the apartment accommodate
- Extra_people:
- Minimum_nights: minimum number of nights
- Availability_30:
- Availability_60:
- Availability_90:
- Availability_365: how many days of the year is the apartment available
- Number_of_reviews
- Review_scores_rating: overall rating of the stay
- Review_scores_accuracy: how accurate was the description of house and location on the online listing
- Review_scores_cleanliness
- Review_scores_checkin: how welcoming was
- Review_scores_communication: how avalilable and clear is the host

- Review_scores_location
- Review_scores_value: price-quality rate
- Instant_bookable: boolean variable that indicates if the apartment is bookable immediately
- Cancellation_policy
- Require_guest_profile_picture
- Require_guest_phone_verification
- TV
- WiFi
- Air_Condition
- Wheelchair_accessible
- Kitchen
- Breakfast
- Elevator
- Heating
- Washer
- Iron
- Host_greets_you
- Paid_parking_on_premises
- Luggage_dropoff_allowed
- Long_term_stays_allowed
- Doorman
- Pets_allowed
- Smoking_allowed
- Suitable_for_events

- 24_hour_check_in

  Since the *Neighbourhood* variable is numeric and contains only values from 1 to 9 corresponding to the areas of Milan, we decided to insert a new column called *Quartiere* of string type which contained the same values as the *Neighbourhood* variable but in the following format: *Zona 1*, *Zona 2*, *Zona 3* etc...

  Through the use of this dataset, an infographic was created capable of displaying analyzes of the available variables with the aim of analyzing the apartments on the AirBnB website in Milan 2018. To do this, pre-processing operations were first carried out using Python and then the infographic was developed in Tableau.
  In the next chapters, the graphs contained within our dashboards will be illustrated and explained one by one

4. Zones, Average Price, Minimum stay

In the first dashboard of the story there are three charts

- **Choropleth map**:
  This map is intended to represent the districts of Milan in a geographic map.The shades of the colors of each area inside it indicate, based on the intensity, whether that particular area has a large number of apartments inside it or not

- **Treemap**:
  Is used to display hierarchical data using nested rectangles.
  In our infographic it was used to compare the average prices per night and the minimum days of stay of the apartments between the different areas

- **Segmented bar chart**:
  Is a type of chart that uses segmented bars that add up to 100% to help us visualize the distribution of categorical data.
  In this project its horizontal version was used, placing the quantity of apartments on the abscissa axis and the price ranges on the ordinate axis. We have thus obtained a graph capable of representing the quantity of houses for each price range for each area

4.1. Distribution of the apartments in Milan

The first analysis performed for our infographic is that of the distribution of apartments within the municipalities of Milan. This analysis was done through a choropleth map which highlights the number of apartments in the 9 districts of Milan based on the colour tone.

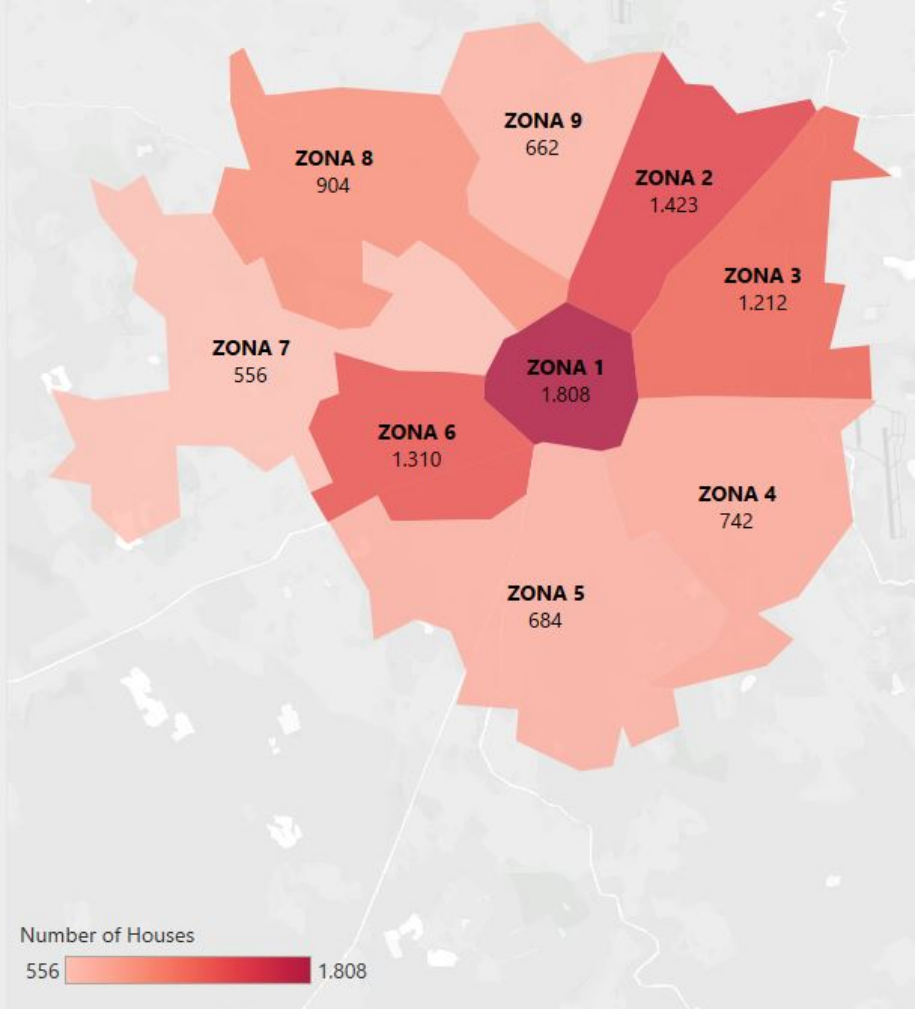

Figure 4.1: Apartments per Zone

As shown in Figure 4.1, zone 1 of Milan is the one containing the greatest number of apartments counting around 1808 even if it is the smallest area. Following are the areas with more apartments are Zone 2 (1423 apartments), Zone 6 (1310 apartments), Zone 3 (1212 apartments), Zone 8 (904 apartments), Zone 4 (742 apartments), Zone 5 (684 apartments), Zone 9 (662 apartments) and Zone 7 (556 apartments). The decreasing order just listed is easily understood by simply looking at the density of the colour of the highlighted area. In fact, the range goes from 556 apartments with the lightest shade up to a maximum of 1808 with the darkest shade.
The result obtained is the expected one, since zone 1 of Milan is the one which, as already mentioned, contains the main attractions of the city and therefore has a high tourist demand.

In any case, from the graph it can be understood that the distribution is heterogeneous given the different nuances that can be observed, in fact, zones 7, 9 and 5 contain about a third of the apartments contained in zone 1

4.2. Average Daily Price and Average Minimum Nights

The subsequent analysis carried out has the purpose of analyzing the average price per night and the number of days of minimum stay of the apartments grouped by area.
To do this, a treemap was created that represented the values obtained for each area, as regards the colours, a palette already available in *Tableau* was chosen with the addition of the missing ninth colour always in harmony with the palette

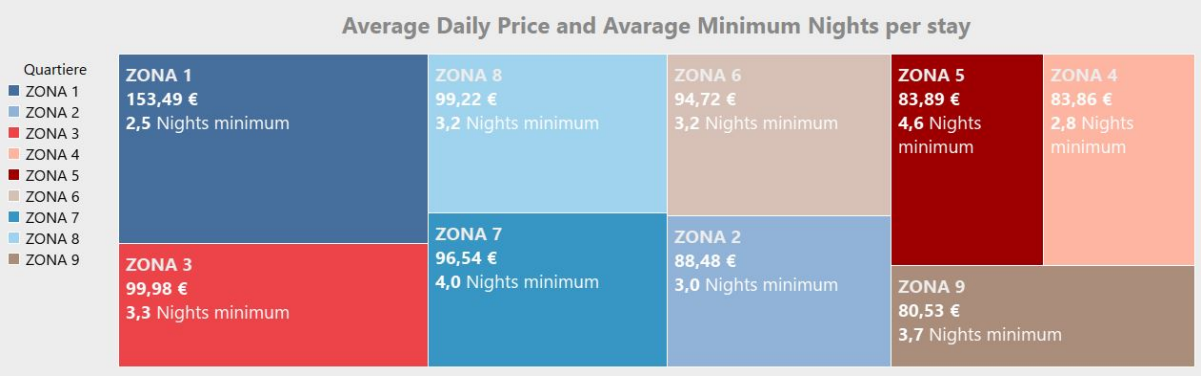

Figure 4.2: Average Daily Price and Average Minimum Nights per stay

As can be seen in Figure 4.2, zone one of Milan is the one that has the highest average price per night (153.49€) compared to all the others, this result does not leave us surprised, since it is the neighborhood with the highest tourist demand. The cheapest neighbourhoods are 9, 5 and 4.
The results obtained are easily deducible by simply looking at the graph, in fact, as can be seen, the rectangle of zone 1 is the largest of all, which highlights the high price per night of its apartments.
As regards the minimum number of nights that guests can stay, zone 1 is, contrary to usual, the one with the lowest number.This could be due to the fact that the historic center of Milan attracts many tourists who stay in the city only for the weekend or even during the day, the host structures therefore have to meet these needs and offer a very low minimum number of nights.

Zone 7 and zone 5, on the other hand, are those with the highest number of minimum nights of stay, the cause could be that outside the historic center there are much larger apartments than those located in the historic center for which it is not advisable to rent the apartment for a few nights given the great cleaning and organization work behind it.

4.3. Number of Houses per Averange Price

The analysis of the average price of the apartments aggregated for the areas of Milan continued with the development of a segment bar chart which allowed us to highlight the price ranges and the quantity of apartments in each district.
Naturally, the chosen palette is the one used previously in order not to create ambiguity and confusion between the graphs and to have the dashboard in harmony
As a first step, the *Daily_ Price* variable was divided into different intervals:

- [0,50]
- [51,70]
- [71,100]
- [101-150]
- [151-250]
- [251-500]
- [500-+inf]

The choice of the width of the intervals is not random but made according to the price ranges of interest. In particular, we wanted to give more weight to the price variations for the lower ranges compared to the higher ones (a difference of €30 can be significant for an average tourist but it is not for those with higher budgets)

For the creation of this graph, the same color palette used previously was used to keep the dashboard in harmony and not to create ambiguity.
The result is a horizontal segmented bar chart from which it is easy to deduce which price ranges of the most common AirBnB apartments in Milan are by simply looking at the length of the bars.
The color of the segments, on the other hand, indicates the area of Milan

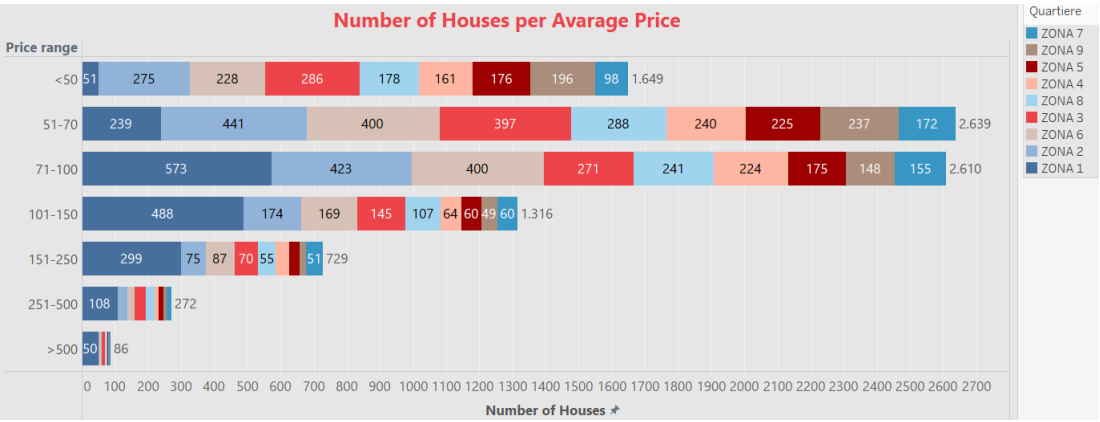

Figure 4.3: Average Daily Price and Average Minimum Nights per stay

in question while the length indicates the number of apartments inside it, in that specific price range.

The graph in Figure 4.3 immediately shows how zone 1 has the highest number of apartments with the average price per night higher than all the others, in fact it contains the highest number of apartments for the first 5 price ranges in descending order.
The price ranges [51,70] and [71,100] are those that contain the greatest number of apartments (respectively 2639 and 2610) which indicates which are the most common average prices per night in Milan.

5. Host Is Super-Host

The second dashboard of the infographic contains a set of graphs relating to the representation of the *Host_Is_SuperHost* variable. The goal of this dashboard is to highlight what are the factors that characterize super hosts compared to normal hosts. The graphs contained are the following:

- **Choropleth map**:
  As before, this graph illustrates the 9 areas that make up Milan on a geographical map. The difference this time is that the color density indicates the number of apartments owned by super-hosts within each geographical area

- **Butterfly chart**:
  A Butterfly Chart is a type of bar chart where two sets of data series are displayed side by side. It gives a quick glance of the difference between two groups with same parameters.
  In this dashboard it will be used to compare the amount of apartments for each price range between super-hosts and non-super-hosts

- **Bar chart**:
  Two Bar charts were used to compare reviews of super-hosts and non-super-hosts i every field, the graphs used are a little different from usual as they do not display the entire bars but only the points corresponding to the value

5.1. Superhosts' houses per Zona

For the Choropleth map it was chosen to use the blue color because it is in harmony with the rest of the dashboard.
In this graph that represents the subdivision of Milan into zones, the areas with a darker shade of blue represent the districts with a higher presence of apartments owned by super-hosts, on the contrary a lighter shade indicates

a low quantity.
The range goes from a minimum of 100 to a maximum of 502.

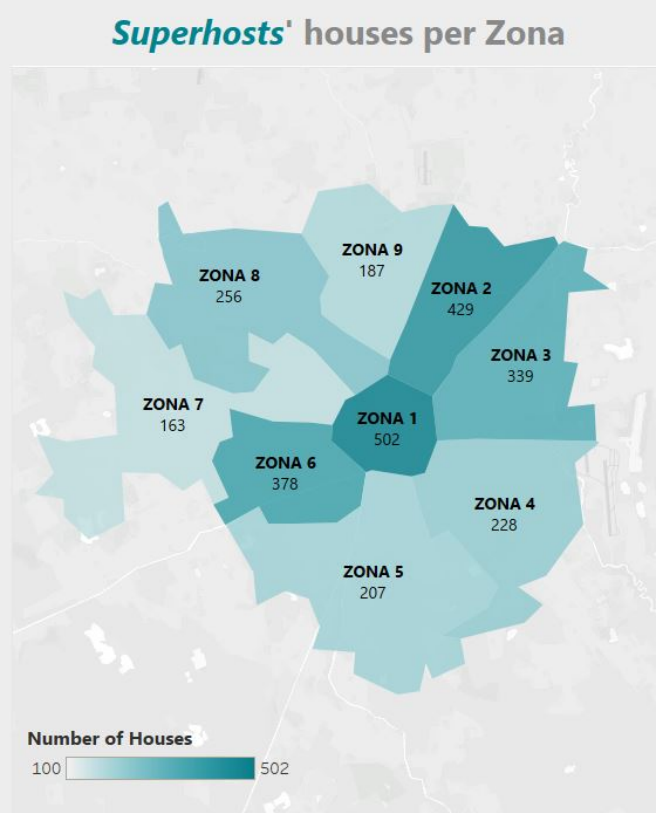

Figure 5.1: Superhosts' houses per Zona

As can be seen in Figure 5.2, the zone that contains the highest number of apartments owned by super-hosts is zone 1 with 502 apartments. Immediately following are the zones 2 with 429 apartments, 6 with 378 and 3 with 339.
Therefore, seen from the analyzes made previously, zone 1 is the one with the highest number of apartments, with the highest price per night and with the highest number of apartments owned by super-hosts.

5.2. Percentage of Superhosts and Non-superhosts houses having Daily Price in a given range

In this graph, the number of apartments in each price range owned by super-hosts and non-super-hosts will be analyzed using a butterfly chart. The color

palette chosen is that of AirBnB which contains shades of blue, red and gray suitable for our visualization.
The intervals for the price ranges used are those already seen previously, in particular they are:

- [0,50]
- [51,70]
- [71,100]
- [101-150]
- [151-250]
- [251-500]
- [500-+ inf]

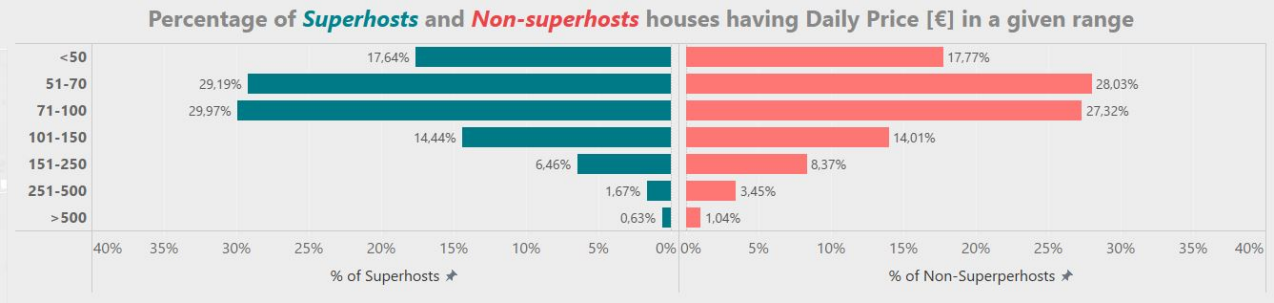

Figure 5.2: Percentage of Superhosts and Non-superhosts houses having Daily Price in a given range

The percentage values reported in this visualisation, that are referred to Superhosts, were calculated on the population of Superhosts houses, not over all Milan's houses. Similarly for numbers referred to Non-superhosts houses. This means that, for example, while 0,63% superhosts accommodation, cost daily 500€ or more, for Non-superhosts the percentage of houses having a daily price of >500€ is 1,04.

If we had confronted the amount of houses in absolute value offered by Supehosts and Non-Superhosts, in any price range the last ones would have appeared as far more frequent, since, in general, Superhosts are an excellent minority among Airbnb Hosts.

In general, we can see that the distribution of accommodations on the two sides of the butterfly chart is very similar. This visualisation, then, shows how becoming a superhost isn't necessarily influenced by the average price of the house.

5.3. Average rating: Non-Superhosts vs Superhosts

In the last part of the second dashboard of our infographic, two bar charts were compared. The bar charts in question contain the display of all the variables relating to the reviews grouped by zone and differentiated by superhost and non-super-host.
In comparison with normal bar charts, ours do not display the entire bar but only the point corresponding to the value. In particular, our graphs display the average of each review category for each area differentiated by super-host and non-super-host The displayed variables are:

- Review_scores_rating: overall rating of the stay
- Review_scores_accuracy: how accurate was the description of house and location on the online listing
- Review_scores_cleanliness
- Review_scores_checkin: how welcoming and on time was the host
- Review_scores_communication: how much available and clear was the host
- Review_scores_location
- Review_scores_value: price-quality rate

As for the color palette used, red from the AirBnB palette was used to visualize the bar chart of non-super hosts and blue for super-hosts

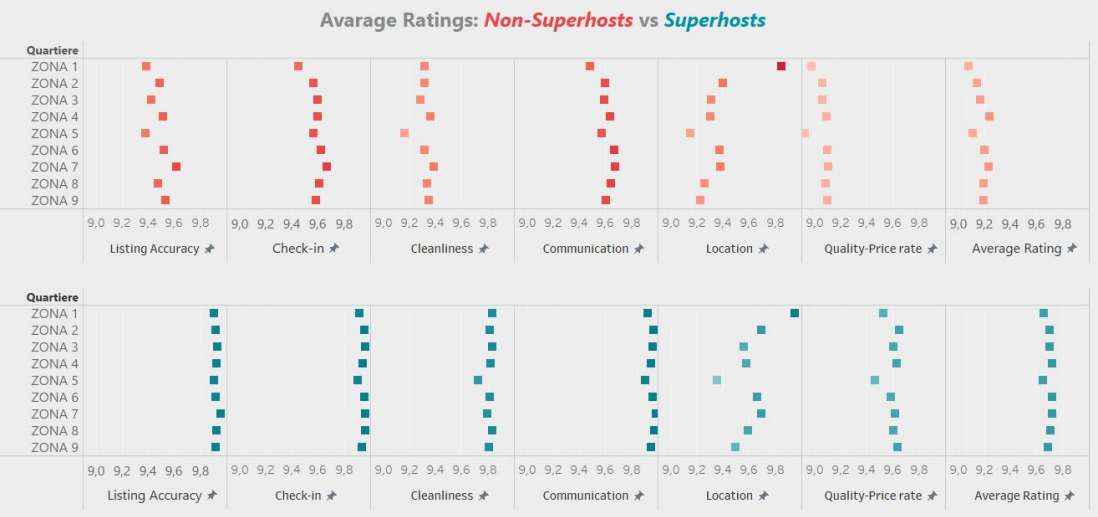

Figure 5.3: Average rating: Non-Superhosts vs Super-hosts

Looking at Figure 5.3, it's immediately apparent how super hosts garner higher reviews across all categories than regular hosts.This data does not leave us surprised since to become a super host it is necessary to have extremely high scores.
We also note that, in general, there is not much difference in the distribution of the average values of the reviews between the two categories belonging to Location. In fact, having an average of reviews higher than 4.8 is a necessary requirement to become a super-host, which becomes almost the only sufficient one since the other three are easily obtainable.
This may also be the reason why most super-hosts are concentrated in zone 1 thus getting very high reviews regarding the location, as opposed to apartments located outside the city.As mentioned in the previous chapters, value for money is also a category of reviews to which the tenant is asked to give an evaluation, and as shown in Figure 5.3, super hosts have very high reviews in this category too. Referring to the segment bar charts displayed above, apartments owned by super-hosts have on average a lower price per night than normal hosts and this could influence the reviews just displayed

6. Serviced Offered by Super-hosts

The last dashboard of the infographic will show the visualizations able to explain what are the attributes that characterize a superhost compared to a normal host.
To do this we will use:

- **Dumbbell chart**:
  A dumbbell chart is a composite chart with circles and lines. It is ideal for illustrating change and comparing the distance between two groups of data points. We used it to visualize the percentage difference between the services offered by normal hosts and super-hosts

- **Bar chart**:
  Two bar charts have been created to show whether hosts who own at least 10 apartments allow pets or allow smoking

6.1. Difference in services offered between Superhost and Non-Superhosts

The following dumbbell chart displays the difference between the percentage of Super-hosts and Non-Super-hosts concerning the services offered to the guests. The palette used is always that of AirBnb where red is used for Non-super-hosts and blue for super-hosts.
The goal of this visualization is to investigate whether superhosts offer more services than non-superhosts.
As can be seen in Figure 6.3, almost all types of services are offered more by super-hosts than by normal hosts, a result that does not leave us surprised. On the other hand, however, the *Smoking allowed* and *Pets allowed* categories are mostly present in the apartments of non-super-hosts.
This result could be due to the fact that, in general, welcoming pets into one's apartment requires a great deal more effort in terms of cleaning, also considering the risk that they could damage the furniture and other objects

in the home.
The same is also true regarding the permission to smoke inside your own apartment since this would generate a big problem for the smell it creates.
This type of service could therefore create a problem for super-hosts who must have a very high review score and must therefore always guarantee a clean and fragrant apartment.

Figure 6.1: Difference in services offered between Superhost and Non-Superhosts

6.2. Hosts allowing pets and smokers

Following the results of the previous visualization, we decided to deepen the analysis concerning the variables *Smoking allowed* and *Pets allowed*.
A simple bar chart displayed the hosts with the highest number of apartments and then checked whether they were superhosts or normal hosts and whether they allowed pets to stay or smoking. The result is that the majority of superhosts who own at least 10 apartments do not allow pets or smoking inside.

Since, generally, all zones have an average minimum of a few nights per stay, hosts (especially those who manage multiple accommodations) need to be able to clean their apartments frequently and quickly to ensure high levels of cleanliness, which it could be much more difficult with consent for pets to stay and smoking allowed.

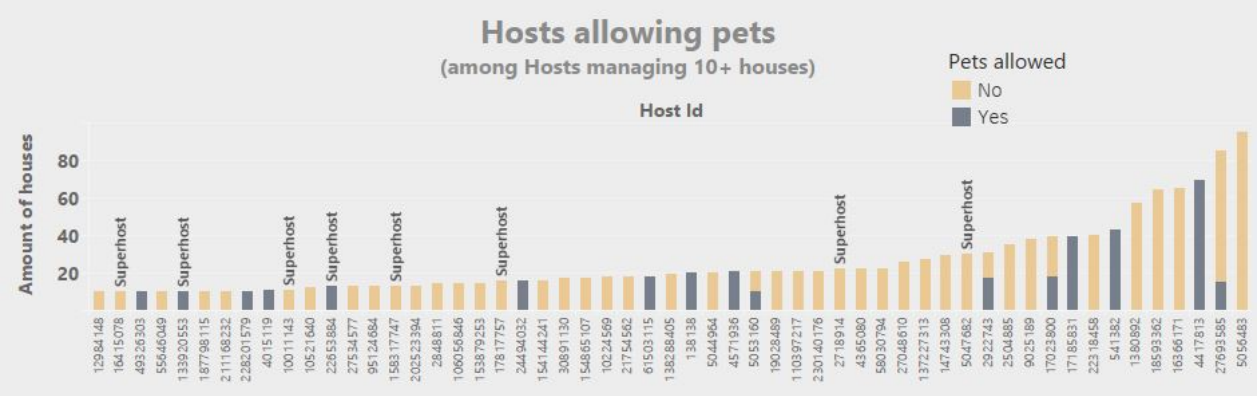

Figure 6.2: Hosts allowing pet with at least 10 houses

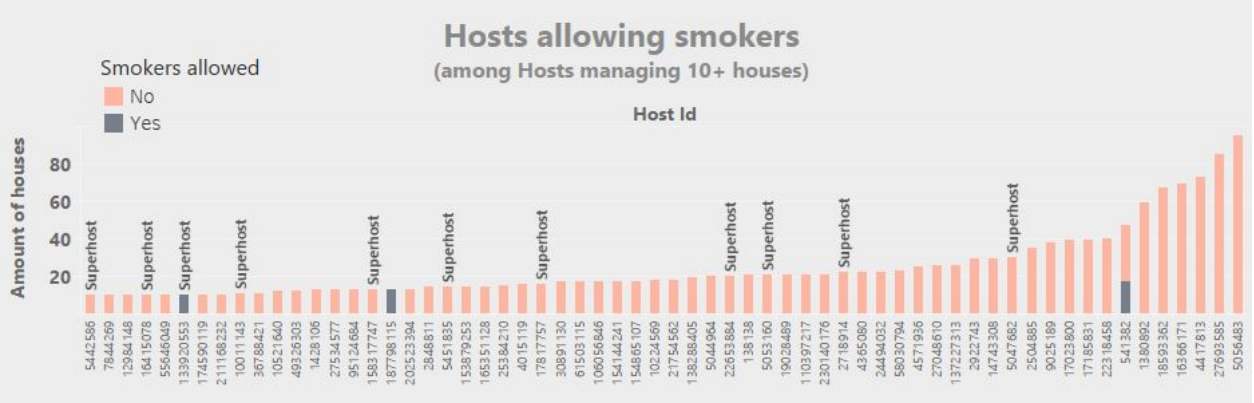

Figure 6.3: Hosts allowing smokers with at least 10 houses

7. Assessment

There are different ways to evaluate an infographic:

- Qualitative - Quantitative
  This types of evaluation are respectively characterized by two different kinds of methods: the heuristic evaluation, which aims to highlight the problems of usability and readability of the infographic with a psychometric questionnaire, which is used to evaluate some quality dimensions of the interaction of the users; and the user test, which has the goal of observing user interactions with the infographic through specially designed tasks.

- Absolute - Comparative
  An absolute rating assesses whether an infographic is 'good', while a comparative evaluation compares two or more infographics to understand if there is a significant difference between the two and which of them is the best one.

- Formative - Summative
  Formative evaluation is carried out during the development phase to correct the infographic and re-evaluate it before the final version.
  Summative evaluation is instead carried out at the end of the development phase to verify that the product meets requirements and expectations.

7.1. Users

The infographic must be consistent with the user's expectations, it has to put him at ease with simple but at the same time complete elements.
It must not contain elements of confusion or that lead to some kinds of mistakes. It also aims to lead the user to ask himself questions and to solve them through it.

To balance the cost of the operation and on the other side its effective accuracy, we decided to involve for our evaluation of the infographic a number of 24 users.

7.2. Heuristic evaluation

The heuristic evaluation involved 6 expert users who were asked to interact with the infographic and to comment out loud what they were doing.
The main comments made for each story dashboard were collected in order to evaluate them for making changes.

- **Comment 1**
  The main comment referred to the second dashboard in the story was regarding the choice of the colour palette used. In fact, some of them have pointed out to us that the colourblind would have difficulty distinguishing red from green.

- **Comment 2**
  The second comment was addressed once again to the colour of the title of the Dashboard graphs since they contained red and green for the distinction of super host and non super host. We were also advised to change the title of the dumbbell chart as it was not very representative.

- **Comment 3**
  Finally, it was suggested that we reverse the order of some charts within the dashboards. In particular, we were told to put the dumbbell chart together with the two bar charts concerning the variables *Smoking allowed* and *Pets allowed*.

7.2.1 Changes to the data visualization

Following feedback from experienced users, changes have been made to the infographic to improve it and facilitate user understanding.
The first change made was on the choice of palette since red and green were initially used to distinguish super hosts from non-super hosts.
It was therefore decided to use the AirBnB colour palette which does not contain particular problems for the colourblind and is more appropriate for the context of the project.

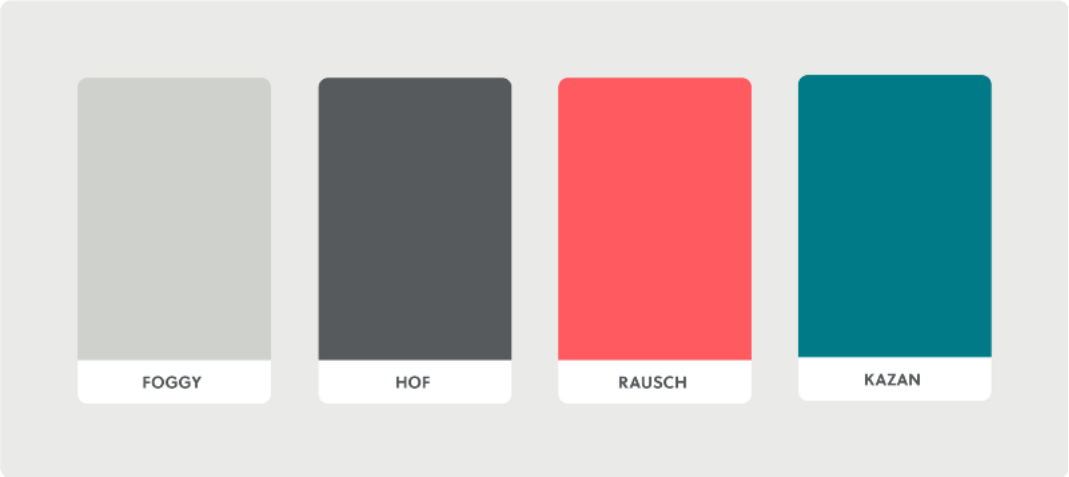

Figure 7.1: AirBnB colour palette

Secondly, a change was made to the titles of the second dashboard since they also contained the colours red and green. Once again we opted to use the AirBnB colour palette.
Furthermore, the title of the dumbbell chart has been changed since it was not very representative of the objective of the chart. In fact, previously the title was "Percentage of host offering a service" but this did not reflect the difference in services offered by super hosts and non-super hosts, that is the key aspect. The new title chosen is therefore "Differences in services offered between super-hosts and no super-hosts"

The last change made in relation to user comments was made on the size of the graphs related to reviews. In fact, previously they were a bit squashed, not allowing a complete view of all 9 districts of Milan, while now it is possible to distinguish them perfectly

Finally, the last change that has been made on the basis of user comments is that relating to the order of the graphs within the infographic. In fact, previously the bar charts referring to reviews were in the third dashboard together with the other two bar charts referring to variables *Smoking allowed* and *Pets allowed*. The two bar charts of the reviews in dashboard 3 were then reversed with the dumbbell chart in dashboard 2 so as to have a more coherent story that better follows a thread of speech.

7.3. User Test

After the heuristic evaluation, the user test was done.
We showed the infographic to 12 sampled people asking them to solve some specific tasks with the aim of determining the clarity of the dashboards. The users were also timed and we offered them our assistance to answer the questions, in this way we could evaluate the efficiency of our infographic.
The tasks created have a rather high level of difficulty, requiring a high degree of concentration and a good study and understanding of the graphs.

1. **Task 1**
   The first task reserved for users is the simplest of all, as it aims to understand if the treemap chart is easy to understand even for less experienced users. Users will have to try to answer the following question:
   Which of the 9 neighborhoods has the highest average price per night?

2. **Task 2**
   The second task is related to the dumbbell chart and requires great skill in reading and understanding it. In fact, the question that users will be asked to answer is the following:
   What are the services offered more by hosts than super hosts as a percentage of their population?
   Therefore users will have to be able to understand that the length of the dumbells represents the difference in percentage of the services offered between super-hosts and non-super-hosts and that the two are differentiated by two different colours

3. **Task 3**
   The third task also requires great ability to read the infographic since users will have to understand how to compare the two horizontal bar charts related to reviews and answer the following question:
   In which of the review categories is there the least difference between super-host and host in terms of average rating?
   With this task we want to understand if the comparison between the two horizontal bar charts is easy to understand or if, on the contrary, it creates confusion.

4. **Task 4**
   The second task wanted to evaluate the clarity of the segment bar chart, in fact users will have to clearly distinguish the price ranges on the ordinate axis and the number of apartments on the abscissa axis

by answering the following question:
For each price range, which area has the most apartments?

5. **Task 5**
   The last task wants to verify if the butterfly chart is easy to read and understand by users of any level of experience. The question that users will have to try to answer is the following:
   What are the most popular price ranges for super-hosts and hosts?

This phase of heuristic evaluation took place entirely through the use of a PC.
The infographic was shown through a story created with Tableau (click here to see it), while the psychometric questionnaire was submitted to users via a google form.
It's also notable that users had full control of the device, they could scroll forward and backward across the history with the only "constraint" of thinking aloud so that we could record all the problems encountered by the users themselves.

7.3.1 Results of the user test

The response times to the tasks calculated in seconds for each individual user will be reported below

User	Task 1	Task 2	Task 3	Task 4	Task 5
User 1	20	73	65	102	70
User 2	23	62	69	110	89
User 3	25	88	78	154	77
User 4	18	77	73	126	85
User 5	15	79	80	140	95
User 6	23	54	55	91	55
User 7	15	52	59	89	63
User 8	30	91	94	156	78
User 9	19	73	61	127	83
User 10	14	67	57	113	89
User 11	25	64	62	119	97
User 12	22	63	71	105	75

We represented the data obtained through a violin plot to immediately get an idea of the distribution of times between the different tasks.

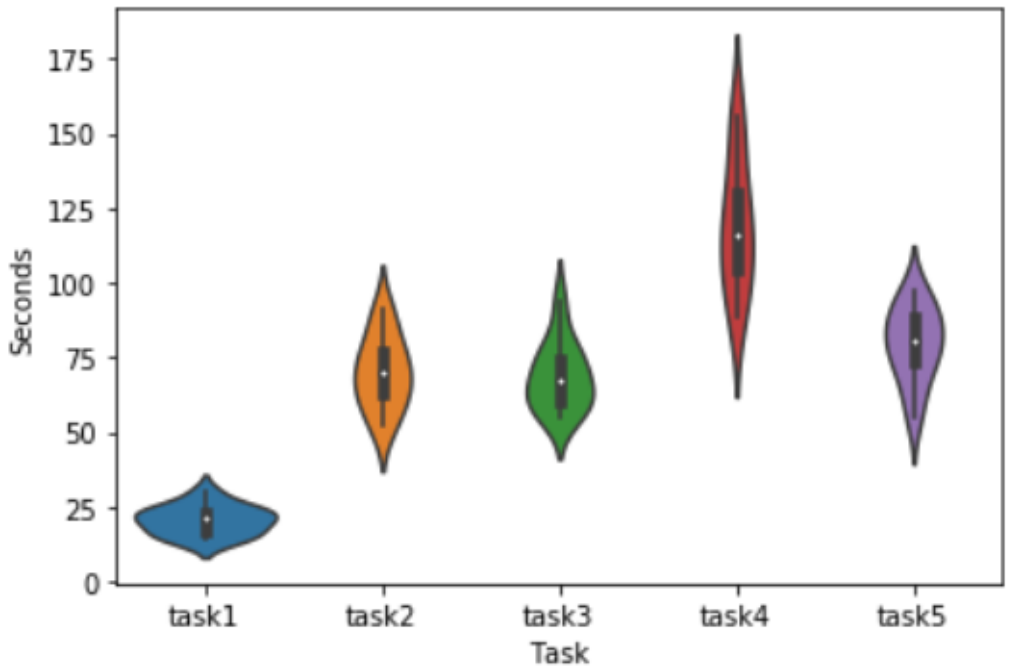

Figure 7.2: Violin plots showing the distributions of the execution times for each task

As can be seen in the violin plot, the response time for task 1 is clearly lower than for all the others, probably due to the ease of the question. Task 4, on the other hand, has a very high variability since users had to search for the number of apartments contained in each price range, an operation that can take a few seconds for more expert people and a few minutes for less expert people.
Box-plots of the results for each task have also been created to check the data distribution in more detail, such as the presence of outlines, comparisons between the median and the range.

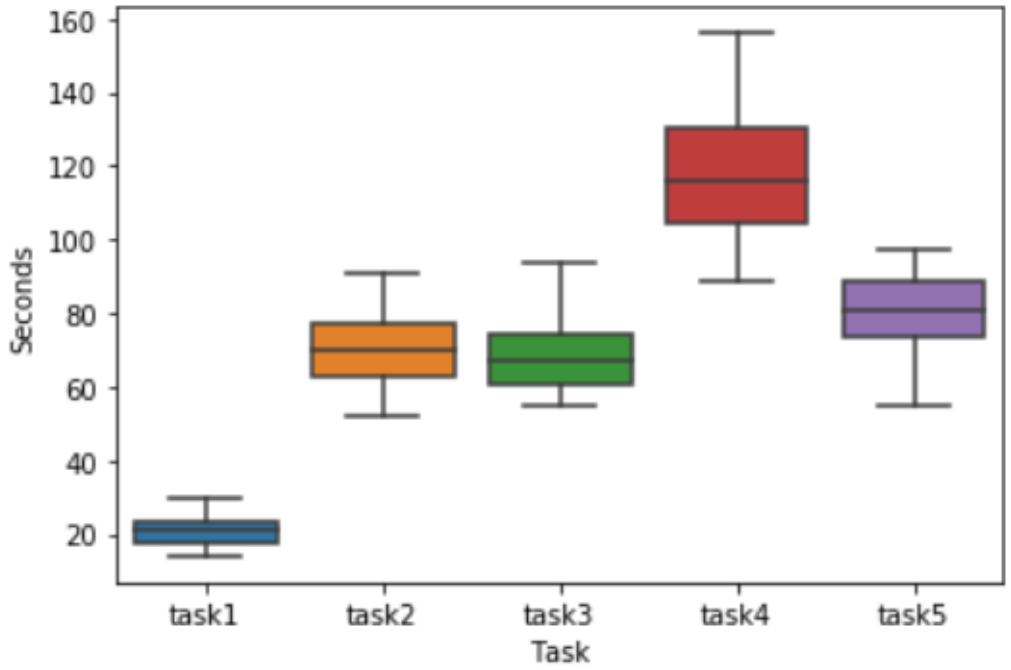

Figure 7.3: Box plots showing the distributions of the execution times for each task

Similarly to what was said previously, looking at the box-plots, task 1 is the one with the shortest response times while task 4 is the one with the longest response times. Tasks 2, 3 and 5 have quite similar median response times, none of the box-plots obtained show outliers.
The average number of seconds taken to respond for each task will now be represented through a bar plot.

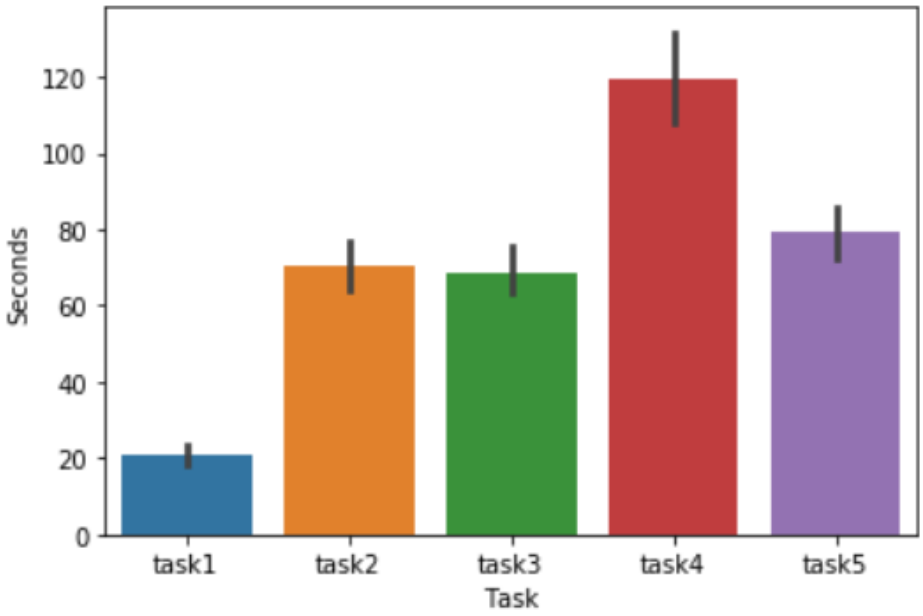

Figure 7.4: Average seconds taken to respond to each task

Now the types of responses obtained on each task by each user will be reported, the execution modality are:

- **C**: Correct
- **E**: Wrong
- **S**: Alone
- **A**: Assisted

The task resolutions registered as "assisted" are obviously the ones characterised by any kind of help or explanation both on the graph and on the question's meaning.

User	Task 1	Task 2	Task 3	Task 4	Task 5
User 1	C-S	C-S	C-S	C-S	C-A
User 2	C-S	C-A	C-A	E-S	E-S
User 3	C-A	E-A	C-A	E-S	C-A
User 4	C-S	C-S	C-S	C-A	C-A
User 5	C-A	C-A	C-S	C-A	C-S
User 6	C-S	C-A	C-A	E-A	E-S
User 7	C-S	C-S	C-A	C-S	C-A
User 8	C-A	E-A	C-A	C-S	E-A
User 9	C-S	E-S	C-S	C-S	C-A
User 10	C-S	C-S	C-S	C-A	C-S
User 11	C-S	C-S	C-S	C-A	C-S
User 12	C-S	C-A	C-A	E-S	E-S

The frequency of the 4 response modalities will now be illustrated through a segment bar chart

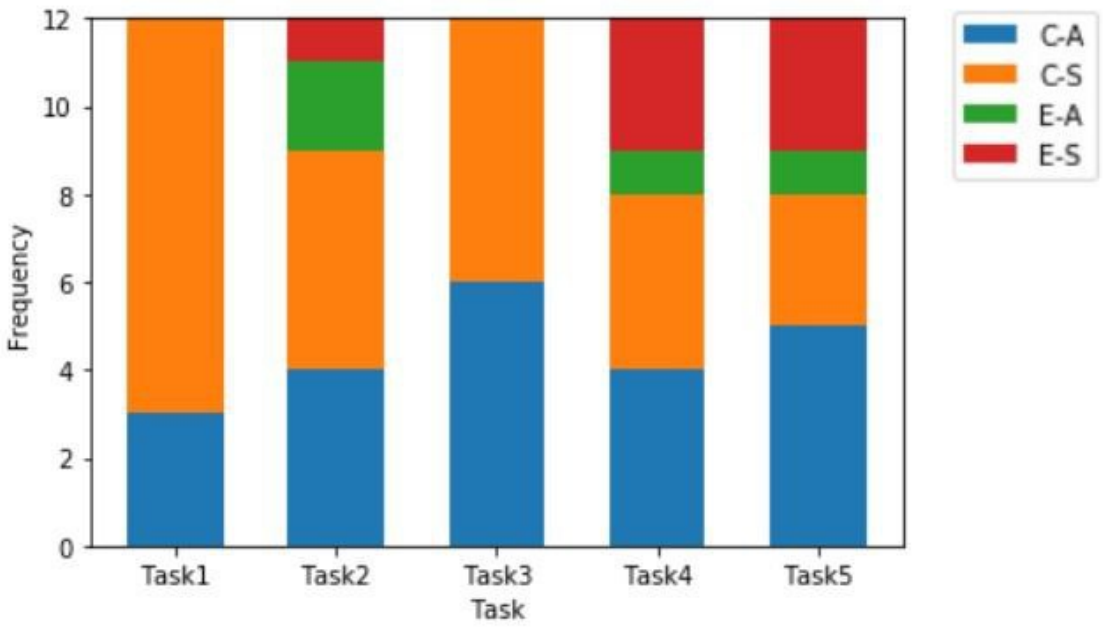

Figure 7.5: Users' performances on the five tasks

As can be seen from the results in Figure 7.5, tasks 1 and 3 did not get wrong answers and most of them were performed by users without assistance. This makes us understand that the Choropleth maps are easy to read and understand by all users, as well as the comparison of the bar charts relating

to the reviews.
A third of the answers to tasks 4 and 5 are incorrect, indicating that some users had difficulty reading the segment bar chart relating to the price ranges and the butterfly chart. For task 4, some users needed a little more time to locate the graph within the infographic allowing them to answer the question, secondly they had confusion reading the graph and understanding that the segments represented different areas and that their size was proportional to the number of apartments.
Also with regard to task 5 some users had difficulty in identifying quickly which was the graph concerning the question, later some had difficulty reading the graph perhaps due to the fact that it is not a graph that is seen every day.
Task 2 got a quarter of wrong answers, this is probably due to the fact that, it is not easy for everyone to identify which dumbbells have the opposite color of the extremes compared to most to understand which services are offered more by non-superhosts than by superhosts.
In general, therefore, the tasks obtained a positive result since more than half of the answers are correct, but on the other hand we noticed that the graphs containing the price ranges on the ordinates, were a bit confusing for less experienced users but still they remained easy to read for slightly more experienced users

7.4. Psychometric questionnaire

For the realization of the psychometric questionnaire we decided to adopt the Cabitza-Locoro scale, and then to conclude it with a final request for an overall evaluation of the whole infographic. This scale allows the assessment of the quality of the infographic on a scale of 1 to 6 for the following four fields:

- Utility
- Clarity
- Informativeness
- Beauty

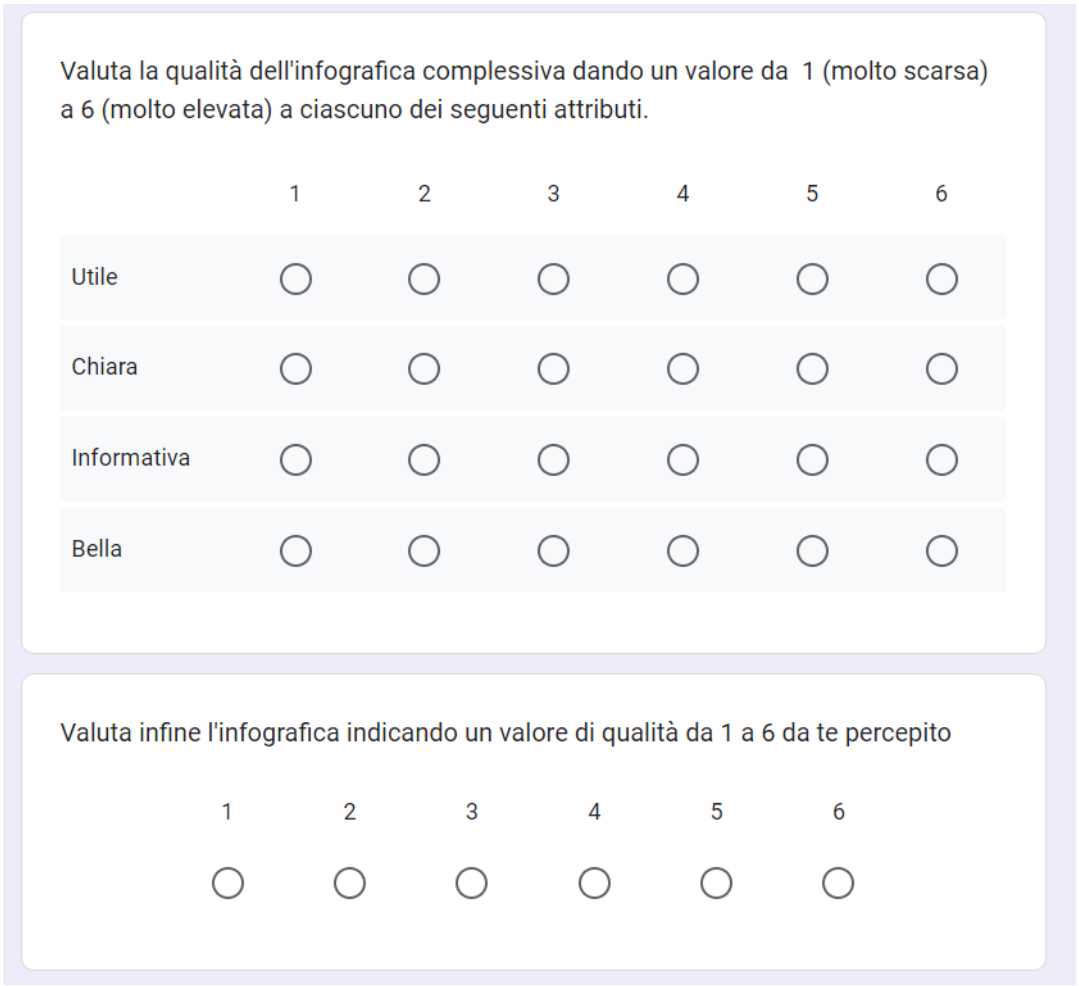

Figure 7.6: The Cabitza-Locoro scale in our google form

We provided the psychometric questionnaire with the Cabitza-Locoro scale to 24 users, we now report the results.

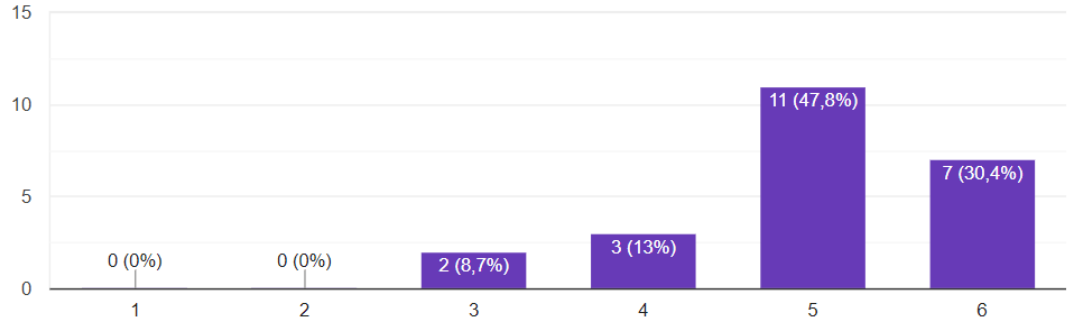

Figure 7.7: Result to the questionnaire to evaluate the infographic indicating a perceived quality value from 1 to 6

Almost half of the people rated our infographic with a score of 4 out of 5, while 30.4% gave a rating of 5 out of 5, a result that leaves us pleasantly impressed and satisfied. 13% gave an overall rating of 3 out of 5, while only 8.7% gave a rating of 2. One person refrained from answering, probably due to a careless error.

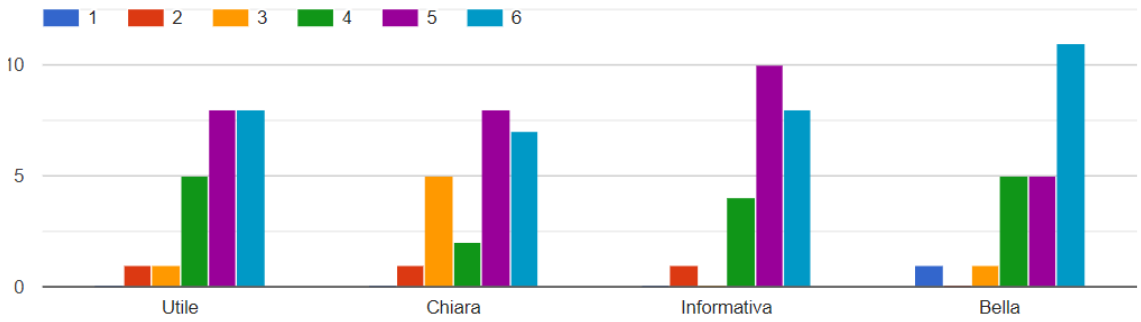

Figure 7.8: Attribute assessment questionnaire result

Also with regard to the questionnaire relating to attributes, positive results were obtained since most of the people gave scores between 4 and 5 and only one person gave low scores.

8. Conclusions and future developments

In this project, an infographic has been developed that can display the data referring to the apartments on the AirBnB website in 2018.
Through the use of Tableau functions, dashboards and different types of graphs have been built which have allowed us to implement statistical analyzes. The main objective was to initially analyze the distribution of the apartments in the various areas of Milan, comparing prices and reviews based on the host category (superhost or non-superhost). Subsequently, the attributes that characterize and differentiate normal hosts from superhosts were explored.
The conclusions obtained are therefore that in zone 1 there are the highest number of super-hosts, which certainly have a higher average review score in all categories in comparison with superhosts. They also offer a greater number of services and therefore have well-stocked apartments with the exception of the possibility of accommodating pets and smoking inside.

A possible future development of our analysis would certainly be to obtain an updated dataset to see if there have been any variations over the years. It would also be interesting to carry out the same type of analysis for the apartments on the AirBnB website in the years of the Coronavirus to see how drastic the changes were.
It would certainly be interesting to also obtain a broader dataset containing the information of all Italian apartments to make comparisons between regions or states if we are talking about Europe, or make comparisons between the various seasons of the year if it were possible to obtain a dataset with data quarterly.

The analysis could also be deepened by trying to understand which are the factors that most influence customer reviews through a study of the correlation with the variables of interest. To do this, however, a larger and more detailed dataset would be needed that also contains the text of the reviews in order to obtain very precise analyses. This would definitely help hosts get

higher reviews and thus achieve superhost status.

In general, through the development of this project, we have learned new data visualization techniques that we did not know before, and we have also learned to use Tableau at a rather advanced level that will certainly be useful in a business context. Moreover, we learnt how to choose the best graph in all circumstances, aware that the first idea is not always the best to represent our information.