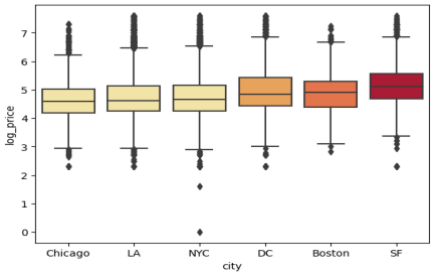
ANALYSIS OF AIRBNB DATASET:

General analysis Initially, the analysis examines the distribution of the target variable 'Log_price'. It is found that the data is concentrated in the lower price range and exhibits a right skewness indicated by a significant difference between the mean and the median. The general boxplot for 'Log_price' indicates the presence of a high number of upper outliers. The interquartile ranges are similar to one another, suggesting that most cities are comparable to one another with respect to the spread of the price.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.0    75.0   111.0   160.4   185.0  1999.0
```

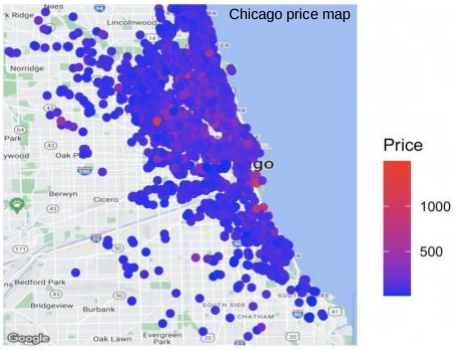


City and neighbourhood analysis:

Price distribution slightly differs among different cities due to several different factors as neighbourhoods, morphological reasons and so on.
We can notice how the price of accommodations within each city is very closely associated with neighbourhoods, as the median prices for each neighbourhood vary considerably across all cities. The reputation and luxury of the areas could be drivers for the price, but it is only speculation as there isn't any specific data in the dataset with respect to neighbourhood qualities.

Neighbourhoods Crime rate analysis

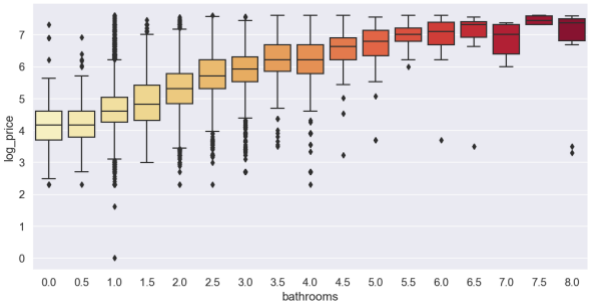
We took the liberty to do some research on the crime rates of the different neighbourhoods and we took notice of the fact that the cheapest neighbourhoods consistently had a remarkably high crime rate (safety score < 30) across all cities. However, the inverse could not be said for the most expensive neighbourhoods, as crime rate was much more inconsistent, varying from high to low. This could mean that crime rate is a driver of the price, but there could be more effective drivers of price that overpower the influence of high crime rate on price.



Property type analysis: If we instead analyse the median price for property type in all the cities, we can see that there is also a clear association between property type and the price. In most cities, we can notice how luxury is still one of the main drivers of price, since property types like "Boat", "Castle", "Loft" and "Serviced Apartment" commonly rank among the most expensive, whereas "Hostel", "Bed and Breakfast", and "Dorm" commonly rank among the cheapest. Our analysis about AirBnb price listings will continue by looking more in depth into the drivers that could potentially influence the target variable, Price.

Beds and bedrooms: Beds and Bedrooms behave very similarly due to their clear correlation, 0.7094158. Both variables show a considerable correlation with price: -Beds:0.4331619; -Bedrooms:0.4944369. Analysing the boxplot, when increasing the number of Beds or Bedrooms, the price increases at a decreasing rate, with every increment being less and less impactful. Data about properties with 0 beds seem to be faulty since there are only 4 observations, among which one has 'COZY PRIVATE Bedroom in 2Br Apt!' as its name.

Bathrooms: Bathrooms: a detailed overlook at the dataset highlights some faulty/wrongful data, in which the ratio bathrooms/accommodates is considerably above 1. For this reason, our analysis only focused on observation displaying maximum one bathroom for every accommodate. As expected, the boxplot shows that the number of bathrooms influences the Listing Price; properties with more bathrooms tend to have a higher price, with the increment reducing as the number of bathrooms increases, just like for Bedrooms and Beds.

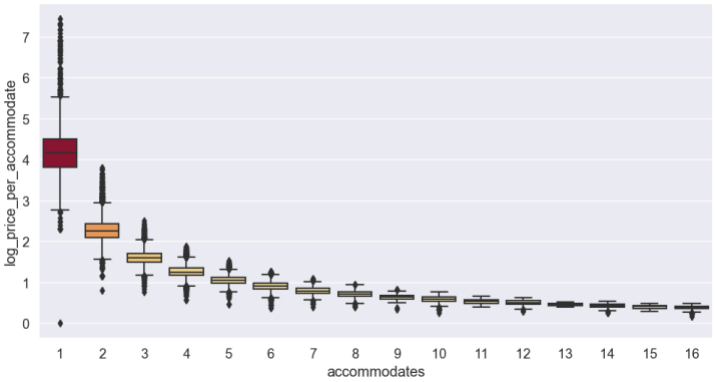


Cleaning fee: By analysing the "cleaning fee" variable, we found that accommodations with this feature are 73.41% of the total and are correlated with a higher median price of 120 versus a median of 95 for those without. In our analysis we discovered a correlation between variables, especially between accommodation size and cleaning fee. As accommodation size increases, the chance of having a cleaning fee increases until it reaches a certain threshold, after which the probability decreases. However, this could be due to inaccurate data because the number of accommodations decreases rapidly after a certain size. This shows how variables can confound each other.

Room types: The accommodations are divided into 3 categories: "Entire home/apt" (55.74% of the dataset with a median price of 160), "Private room" (41.34% of the dataset with a median price of 75), "Shared room" (2.92% of the dataset with a median price of 45); due to the scarcity of the latter, prices of shared rooms tend to be much less consistent

than the first two. However, data clearly show that ‘Private Room’ is the more expensive opposed to ‘Shared Room’ which represents the cheap option.

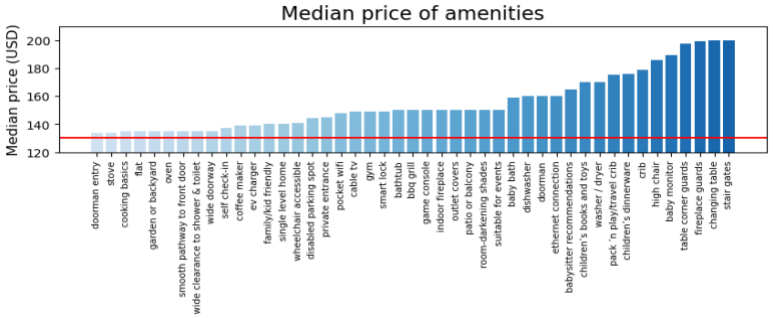
Accommodates: Missing the variable ‘rate of utilisation’, we assumed the number of accommodates to be equal to the maximum number of allowed guests. We adopted two different approaches to analyse the variable ‘Accommodates’:



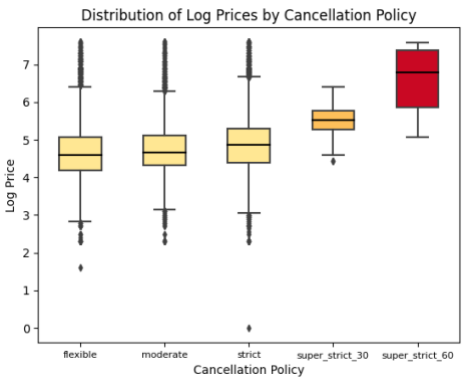
- One approach concerns the relation with ‘Price’, showing a positive correlation between ‘Price’ and ‘Accommodates’, 0.5193258, houses with higher number of allowed guests tend to have higher prices, following a similar decreasing rate of increment as Beds and Bedrooms. See Boxplot in R file.
- The second approach highlighted a negative correlation between ‘Price/Accommodate’ and ‘Accommodates’, following a clear trend (see graph above). Increasing the number of accommodates leads to a reduction of price per accommodate, with the total reduction increasing at a decreasing rate.

The two approaches offer significant insights for both property owners and prospective guests, enabling them to make well-informed decisions regarding pricing and reservations.

Amenities: We analysed individual amenities to identify those linked with higher prices. We calculated median prices for each and the overall median, only including amenities with at least 20 observations. Amenities above the overall median are labelled as "more expensive".



Reviews: We compared the prices of houses with high review scores but few reviews and houses with lower scores but more reviews. After filtering the data and dividing it into four classes, we found that the houses with the lowest scores and fewest reviews had the highest mean price, which was unexpected. However, this result makes sense because people consider price when giving reviews and have higher expectations for higher-priced properties.



Cancellation policies: In addition to our previous analysis, we also examined the "cancellation policy" variable, which contains 5 categories: Flexible (31% of total), Moderate (26% of total), Strict (43% of total), Strict_30 (almost 0% of total), and Strict_60 (almost 0% of total). The boxplots show that the stricter the policy, the higher the price; however, this doesn’t imply that ‘Cancellation Policy’ is a driver of Price. We further investigated the relationship between this variable and other variables and found a pattern similar to that of the "Cleaning Fee" variable. Specifically, we observed that as the size of the accommodation increases, the probability of having a “strict cancellation” policy also increases, therefore all the variables related to the size of the property could be potential confounders.

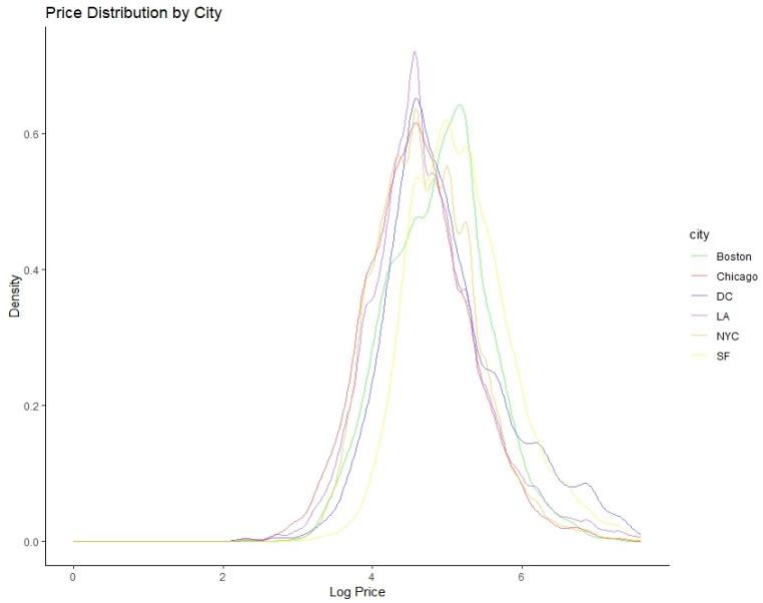
Instant bookable: Analysing the variable “instantly bookable," we found that only 26.25% of all accommodations in our dataset had this service, and their median price (105.0) was slightly lower than that of accommodations without this service (115). Despite the skewness of the price distribution, both mean and median are smaller in instant bookable properties. This could be explained by assuming that more expensive properties would require owner’s approval, since the risk of accepting a potential client is higher.

Conclusions and findings. In conclusion, the analysis of the AirBNB dataset has revealed several correlations and patterns that can provide valuable insights for both property owners and potential customers. The study highlights the influence of various factors such as location, neighborhood, property type, beds and bedrooms, bathrooms, accommodates, cleaning fees, and amenities on listing prices. Overall, these findings can assist property owners in making informed decisions regarding pricing, while potential customers can use them to make informed decisions about reservations.

***NOTE**: before analysing every variable, we cleaned out the dataset following simple guidelines. For most of the variables NA values added up to less than 0.001% of total observations, therefore we removed it.**

MISSING THINGS:

- Line graphs of 6 cities → no difference in the cities
- Maps of chicago+crime rate
- Nel caso riguardare le property types
- Reviews threshold



City analysis:
########## manca analisi Chicago ale###########

To gain further insights into the data, we examined the geographical distribution of prices across the city. Our analysis revealed a clear difference in price distribution between the northern and southern areas of the city.

While we examined the data for several other cities, our analysis did not reveal any notable differences from the trends observed in this report. As such, we will only discuss New York City in detail, as it presents distinctive features that merit further exploration.

###########stesura new york city ale#########

NOTE: before analysing every variable we cleaned out the dataset following simple guidelines. For most of the variables NA values added up to less than 0.001% of total observations, furthermore we removed it.

We analyzed individual amenities to identify those linked with higher prices. We calculated median prices for each and the overall median, only including amenities with at least 20 observations. Amenities above the overall median are labeled as "more expensive"