## 



**Airbnb Data Analysis & Visualization**

**Group 3**

Rajasree Rajendran

Fernanda Ribeiro Bordin

Chaithanya Reddy Bogadi

Naga sindhu Korlapati

Megha Rajam Rao

Sindu Ravichandran

**Contents**

[**Introduction**](#_2et92p0) **2**

[**Motivation of study**](#_tyjcwt) **2**

[**Description of dataset**](#_3dy6vkm) **2**

[Design Principles:](#_1t3h5sf) 2

[Perceptual Properties:](#_4d34og8) 3

[**Data analysis**](#_2s8eyo1) **3**

[Location:](#_17dp8vu) 3

[Number of bedrooms:](#_lnxbz9) 5

[Amenities:](#_44sinio) 7

[Host status:](#_z337ya) 8

[Price acceptance:](#_4i7ojhp) 10

[**Insights gained**](#_49x2ik5) **11**

[**Key challenge faced**](#_147n2zr) **12**

[**Conclusion**](#_3o7alnk) **12**

# **Introduction**

For our final project, we have analyzed and visualized Airbnb Paris data from a corporate point of view. The strategic analysis and visualization have helped us identify the improvements that could be done for Airbnb to continue to excel in the hospitality industry.

# **Motivation of study**

Our main assumption is that hosts might be pricing their listing inadequately, most probably by underpricing them. We will identify the most relevant price-drivers, as well as analyze the price acceptance to verify this hypothesis. If proven correct, a price-recommendation feature could be implemented in Airbnb.

# **Description of dataset**

The Paris dataset obtained from “InsideAirbnb.com” was collected on 5th February 2019. The data points cover the year February 2019-February 2020 and are probably constituted of projected data based on their current reservations. All four csv files in the dataset were used and each file contains categorical, quantitative, text and date time data types.

· **Listings -** has 106 features for each of the 58067 listings. It contains various information ranging from description of the place provided by the host, location, amenities offered, review score of the place, all the way to resident’s neighborhood of the place.

· **Reviews -** has 6 features and 1097737 entries. It contains detailed description of the reviews provided by the users.

· **Calendar –** has 7 features with a total of 21195164 entries. It contains detailed information about the availability of listings and price

· **Neighbor -** has 2 features with 20 entries. It gives information about the neighborhoods’ details in the Paris region.

**Design process**

## Design Principles:

* Maintained graphical integrity, ensuring clear, detailed and thorough labeling in all graphs.
* Followed Tufte’s principles to achieve graphical excellence by maximizing data-ink ratio, removed annotations and used grey labels, as exemplified in figure 6.
* Avoided chart junk by removing redundant data, dark grid lines and background images. Maximized data-density ratio by maintaining figure size to get full glimpse of the graph.

## Perceptual Properties:

* Used variations in color, hue and size to represent categorical data. For example, in figure 8 and 9 we used color to represent booked and un booked data.
* According to Gestalt’s principles we maintained similarity, connectivity and continuity in all graphs, used sequential palette for fig 6 and diverging palette for fig 1 based on requirement for continuous and categorical data.
* In figure 8, we used bar chart and avoided individual pie-charts following Stephen Few’s principles.
* No time series were plotted, but we would have used a line chart to emphasize on patterns with respect to time.

# **Data analysis**

We will go through price drivers to assess their impact and distribution in pricing.

## Location:

Using the price and coordinates of the listings, it was possible to map the average district price (in euros). It clearly validates that there is a significant difference in average price depending on the address of the listing. Except for district 16th, all outside districts have lower average prices, ranging from €71 to €106.

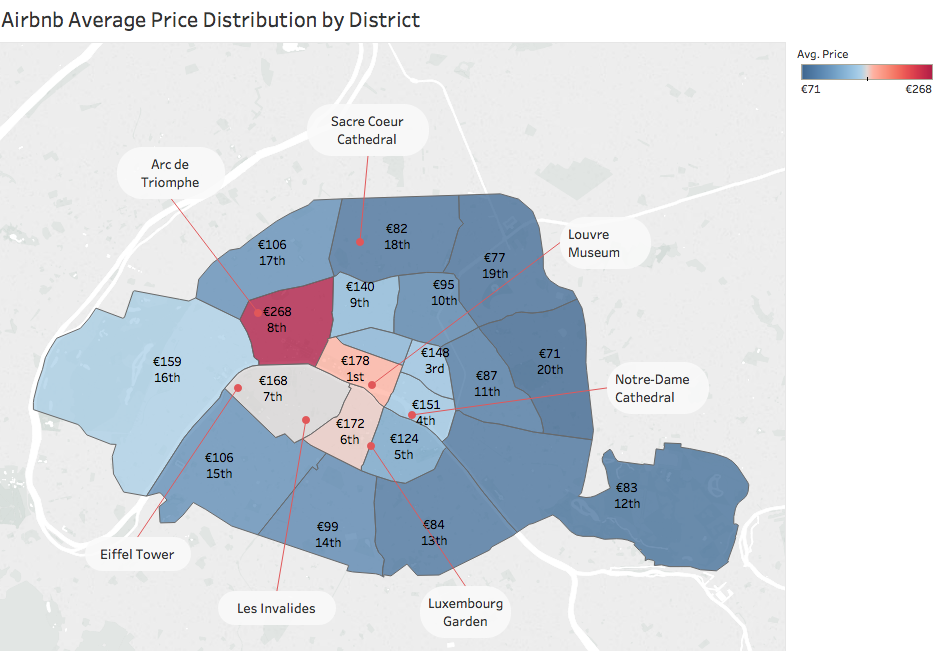


Figure 1

To investigate the reason for this distribution we analyzed the average household income, as well as the location of the most popular tourist attractions.

According to the study done by Grabar [Grabar, 2013], districts 1st, 6th, 7th, 8th and 16th have the highest income rates. The alignment with the price listing geographical distribution confirms that the district social status is impacting the pricing in a predictable way.

The location of Paris tourist attractions further supports the average price distribution, since districts with tourist points have a higher average price. Except for district 18th, that according to [Grabar, 2013] is a low-income area.

To try to determine the impact of having a tourist spot near, we plotted the price variation based on the distance from the Eiffel Tower. For this graphic we used the linear distance based on the coordinates only and disregarding the district or street layout.

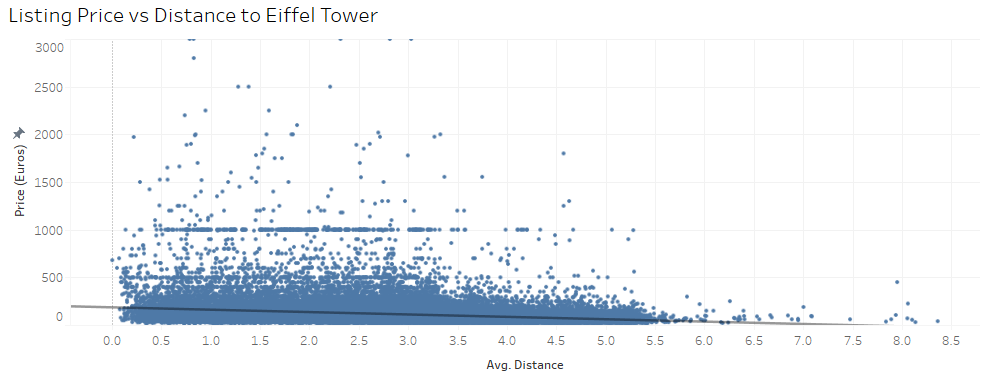


Figure 2

The graphic shows a slight price reduction as we get further from the Eiffel Tower. This effect combined with the large delta above the trend indicates that tourist point is not the main factor impacting price. The big price dispersion might signal overpricing.

## Number of bedrooms:

Our next analysis is done only on apartments. “Zero” bedrooms are equivalent to studios.

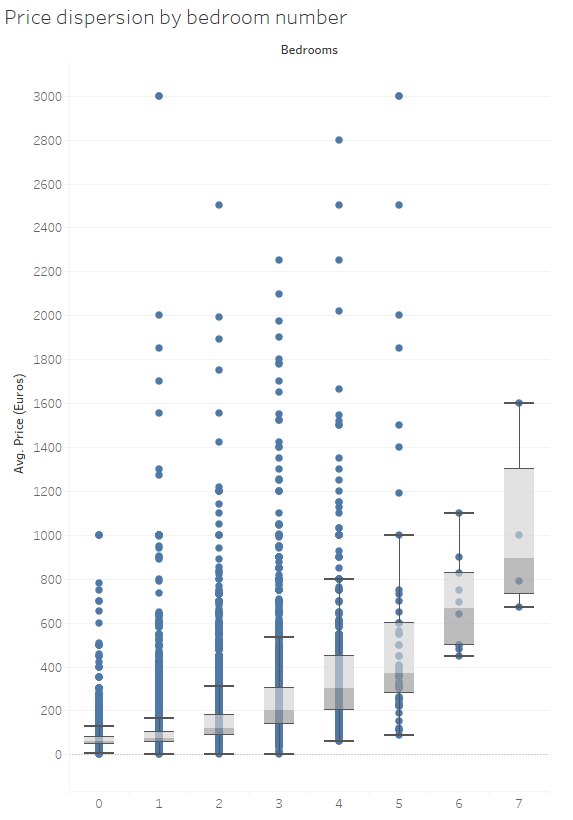


Figure 3

As expected, the number of bedrooms has a positive relationship with price, the almost exponential increase probably reflects the luxury level of the property. The big number of outliers signals other influencing factors and potentially an overpricing trend.

To exclude the hypothesis that location is driving this irregular price dispersion, we mapped the properties by bedroom and the lack of a pattern confirmed our supposition.

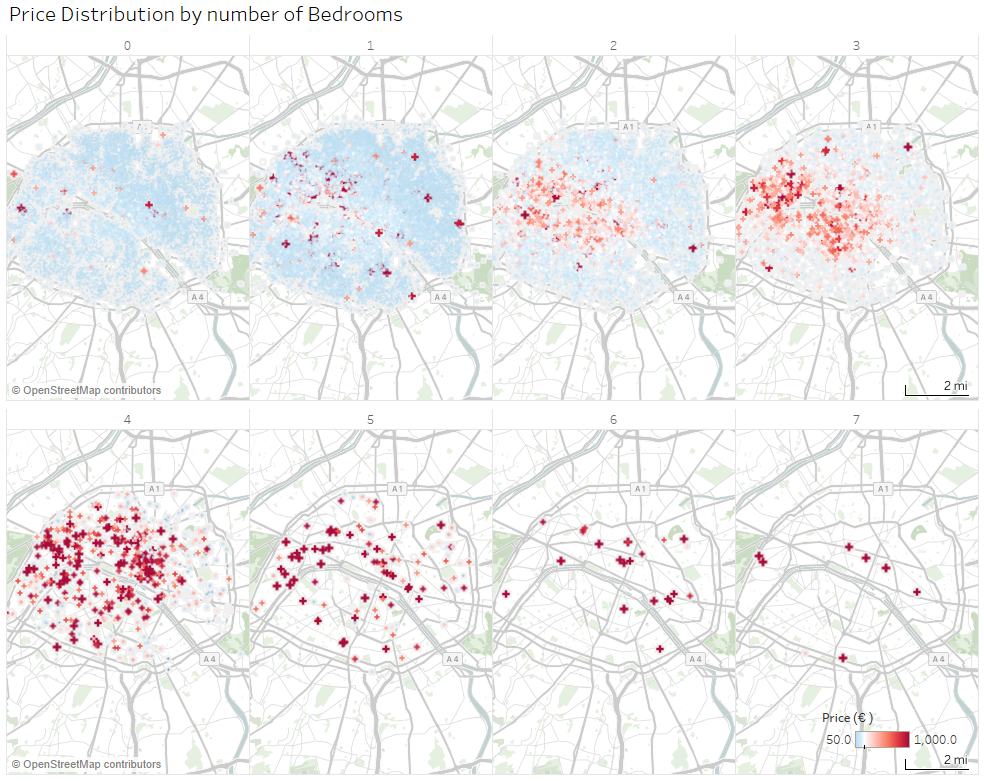


Figure 4

## Amenities:

To determine what amenities, add most value, we filtered the highest prices and plotted amenities. The most valued amenities are attached to luxury and bigger properties (swimming pool, tennis court, etc.) or support health issues (mobile-hoist, roll-in shower with chair, etc.).

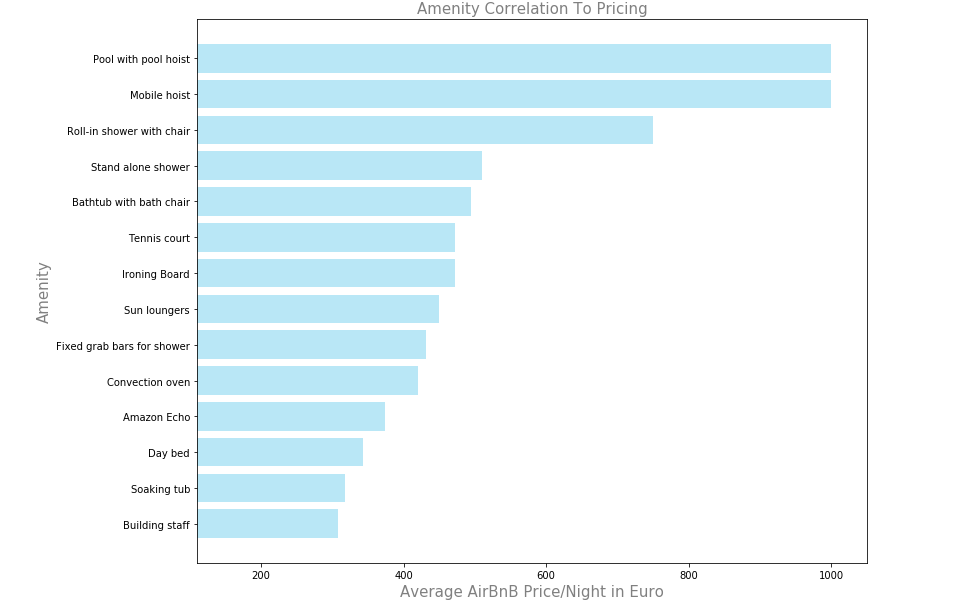


Figure 5

Whereas no causality can be inferred, a valid hypothesis to be explored in another study is whether by adding health support amenities (fixed grab bars, etc.), a host can indeed increase its listing price.

## Host status:

When comparing the average price of super-host listings with that of normal host listings, it appears that super-hosts charge €10 more (see figure below).

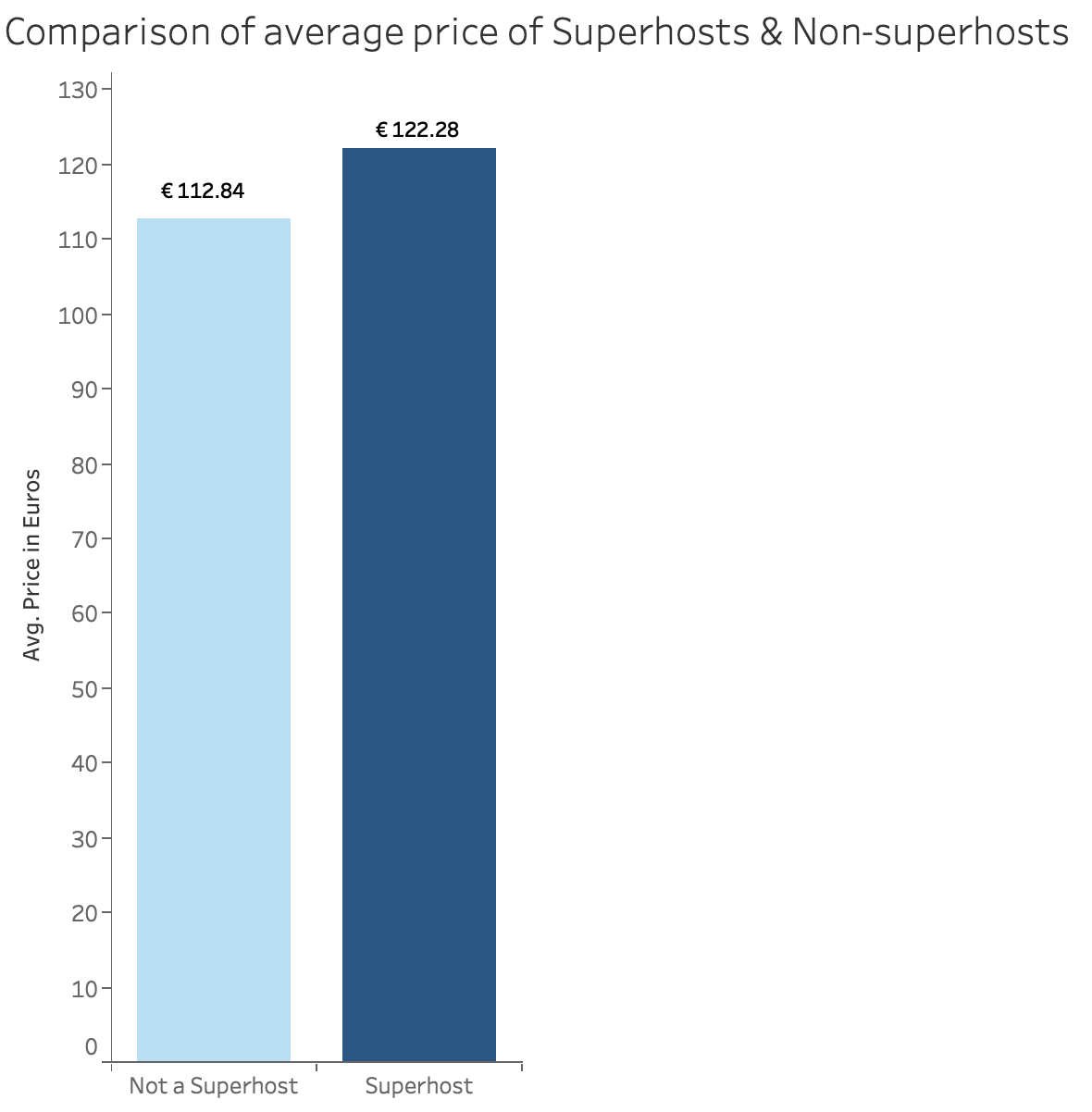


Figure 6

To confirm this hypothesis, we plotted the average price for top 10 neighborhoods (based on the listing price), for super-hosts and non-super-hosts.

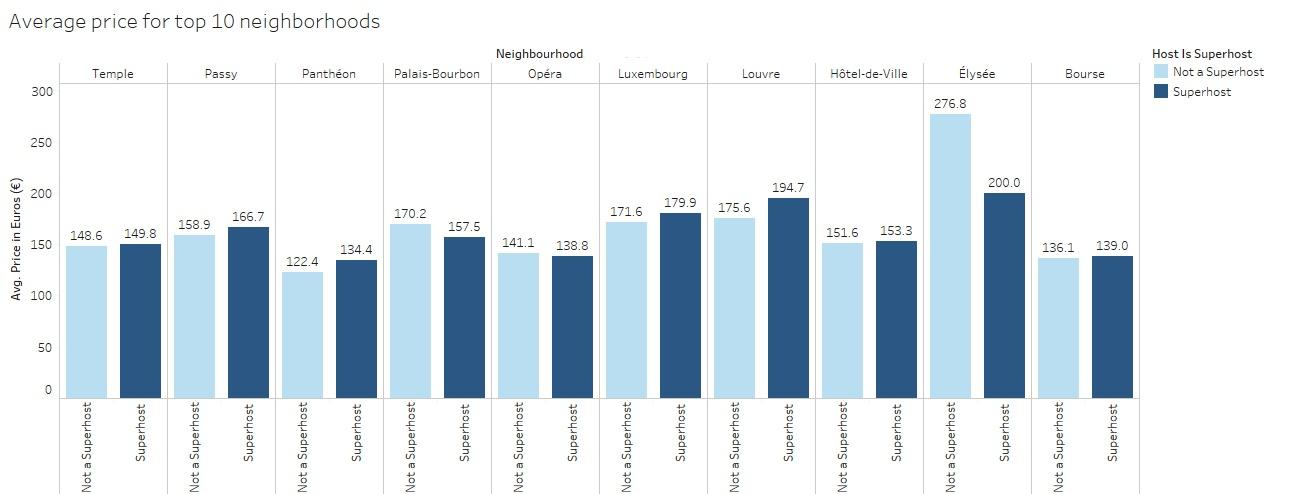


Figure 7

From this plot, it is clear that there is a trend of super-hosts charging more than normal hosts. Except for Elysee and Palais-Bourbon, super-hosts have higher prices than non-super-hosts. Elysee, being part of the 8th district is in one of the most exclusive and expensive area of Paris. Due to its relatively small size there are lesser listings, leading to a non-significant relationship.

## Price acceptance:

In several occasions during the analysis of price drivers and their price distribution, the data signaled that certain listings might be overpriced. We plotted the occupancy rate for the listings with price above €1000.



Figure 8

Clearly the booking rate for these listings are very low, being empty over 95% of the time. As a low-income neighborhood, Buttes-Chaumont only has 10 listings over €1000, leading to a misleading 76% occupancy.

Looking into moderate price listings (< €1000) this overpricing becomes more evident. The graphic below shows that in all the districts, these listings are booked more than 50%, thus confirming the tourists’ preference.

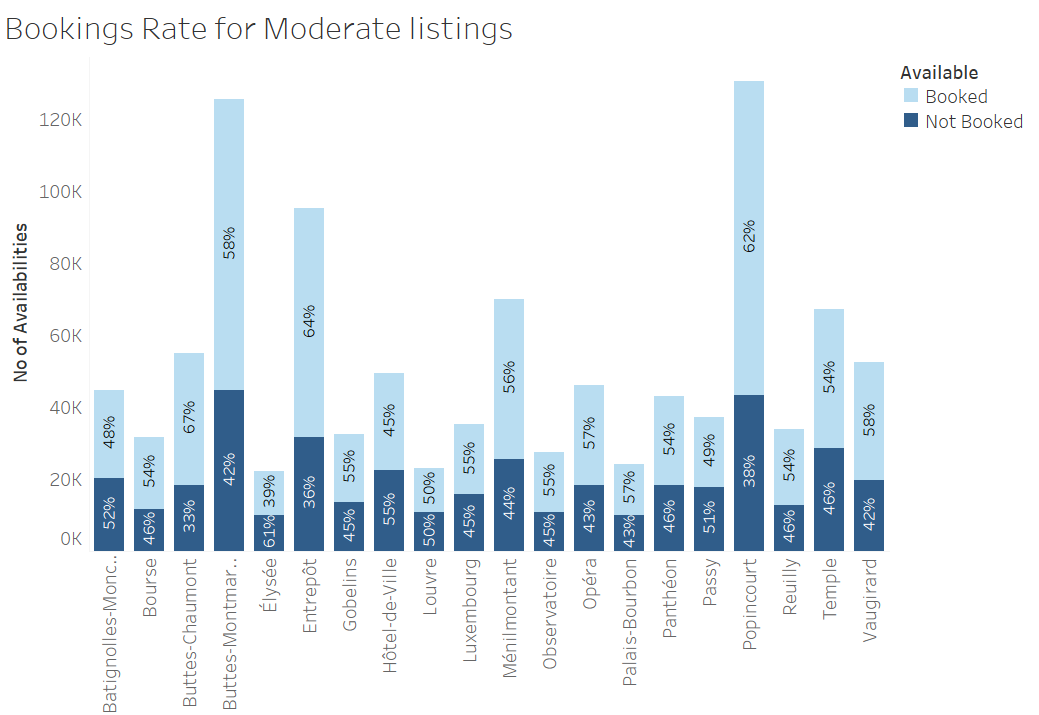


Figure 9

# **Insights gained**

After an extensive data analysis, we were able to confirm that listing pricing is influenced by all attributes examined, though none proved dominant.

* Location’s impact is mostly driven by household income distribution, with tourist spot having a small influence.
* The number of bedrooms, just as type of amenities, impact price mostly in relation with the property size and luxury.
* Adding health related amenities might allow for a higher asking price.
* Super hosts tend to charge slightly more than normal hosts.

No significant evidence of properties being underprice was found, indeed we uncovered evidence of overpriced listings. This was further confirmed, since properties priced over €1000 tend to be empty 95% of the time.

# **Key challenge faced**

One of the key challenges we faced during this project was cleaning the data. The dataset from “InsideAirbnb.com” required extensive cleaning in order for us to work with it, especially the data regarding neighborhoods. We cleaned the data in Python by performing exploratory data analysis.

# **Conclusion**

Our objective was to explore the possibility of implementing a price-recommendation system for Airbnb. Based on our findings, we can confirm that the attributes: location, number of bedrooms, amenities and host status are good attributes for the system.

It was understood from our research, that customers tend to choose moderately priced properties over luxurious ones. We could also deduce that amenities can add value to a property, which could imply that adding better amenities could act as a cost-driver. It appears that super-hosts have the advantage of listing higher prices due to their better ratings.

Given the trend of listing overpricing, implementing a price-recommendation system might be strategically interesting to support better occupancy rates. Nonetheless, given the probable high costs to implement such a tool and that mostly overpriced listings would benefit from it, we do not recommend the implementation of the system.