Multivariate Analysis: Primer entregable

Marc Pastor - datexbio

2025-11-21

Table of contents

1 Preprocessing

For this assignment we have decided to work with data from the Inside Airbnb database. It was created originally by the activist Murray Cox in 2015 to show the impact of Airbnb in local communities, but nowadays it serves as a database for Airbnb in many cities in the world. Among them, we have selected Madrid, since it is a familiar place to all of the components of the group. Information was lastly updated on September the 14th 2025, so it is quite up to date.

The data "as is" contains 25000 observations of 79 variables, including categorical, quantitative, and binary data. Moreover, one of the variables (called `amenities`), holds a list of amenities available in the rental unit (such as WiFi, elevator, kitchen, etc.), which can be thought as binary variables as well (the rental uint has this amenity or not). The assignment required to use around 10-15 variables and around 100-1000 observations, so we have decided to:

1. Remove variables with high presence of NA values.
2. Remove every row containing any NA value (since we have so many there are enough left).
3. Decompose the `amenities` variable into binary columns.
4. Select the most interesting variables out of the remaining (so we can end up close to 15).
5. Sample the remaining observations until we keep 1000 of them.

After selection, we remain with 15 variables, which can be seen in Table 1.

Table 1: Selected variables from the data set: 10 quantitative variables, 3 binary, and 2 multiclass categorical.

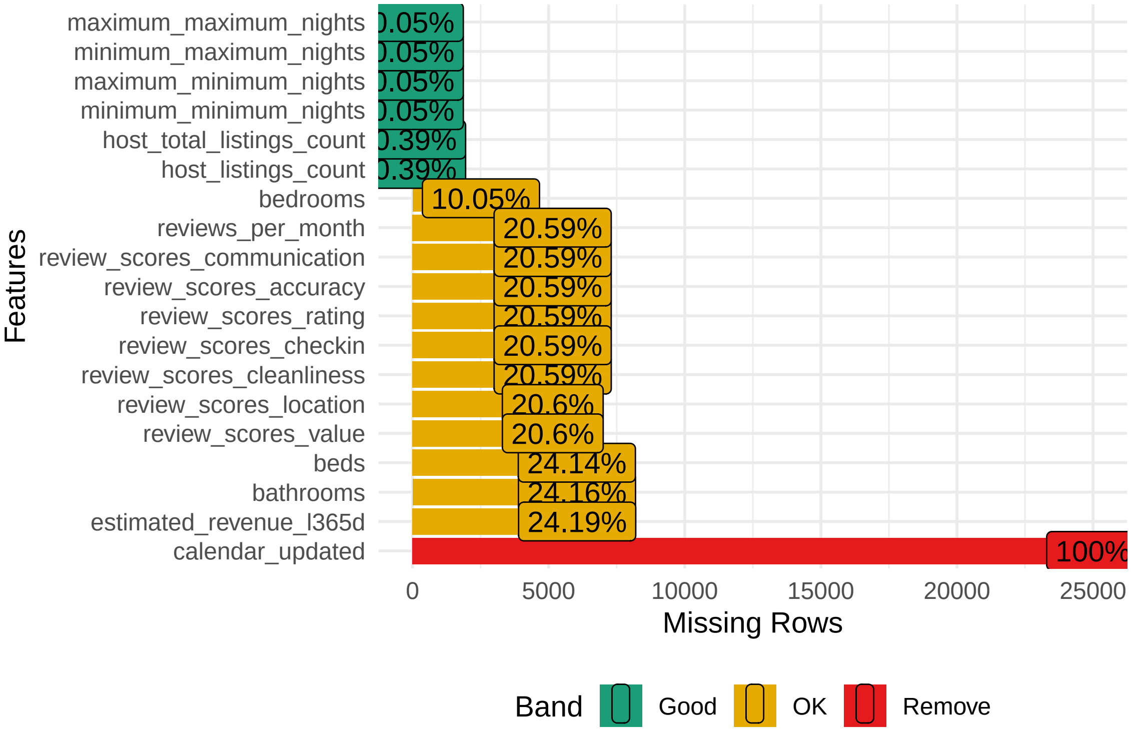
Variable	Type	Description
`accommodates`	Quantitative	The maximum capacity of the listing
`air_conditioning`	Binary	Is there an air conditioning in the lisitng?
`bathrooms`	Quantitative	The number of bathrooms in the listing
`bedrooms`	Quantitative	The number of bedrooms
`elevator`	Binary	Is there an elevator in the lisitng?

Variable	Type	Description
`estimated_occupancy_l365d`	Quantitative	Estimated number of days that the listing will be rented per year
`heating`	Binary	Is there a heating system in the listing?
`host_age_years`	Quantitative	Time (in years) that the owner of the listing has been with Airbnb
`host_total_listings_count`	Quantitative	The number of listings the host has (per Airbnb unknown calculations)
`minimum_nights`	Quantitative	Minimum number of night stay for the listing
`neighbourhood_group_cleansed`	Categorical	The neighbourhood group as defined by public digital shapefiles
`number_of_reviews`	Quantitative	The number of reviews the listing has
`price`	Quantitative	Daily price in local currency
`review_scores_value`	Quantitative	Averge of the reviews of the listing
`room_type`	Categorical	{ Entire place, Private room, Shared room, Entire place }

1.1 Selection criteria

1.1.1 Missing variables

Variables with a high percentage of missing values can be safely discarded. Let us take a look at the following plot.



We can see that the `calendar_updated` variable is almost entirely missing, so we discard it. Among the remaining variables with missing values, we can see that the worst case has around 25% values missing. If rows with missing variables do not match between them (in other words, if they are disjointed sets) we could run into trouble by discarding rows with any NA.

A better approach than removing the rows blindly is perhaps to first choose the variables we want to keep, which will minimize the possibility of finding one. In order to do so, we first need to decompose the `amenities`, as we see in the following section.

1.1.2 Frequent amenities

Among the possible amenities we have basic commodities that the listing is expected to have, such as WiFi or a kitchen, and others that are additions with less relevance, such as a dryer or a dedicated workspace. It is clear that the latter will have little impact on customers' decisions, whereas the former have a huge impact on the listing if they are missing. The criterion we have followed to select the amenities is to keep those that appear in more than 10% of the data.

Here below we show the list of amenities that survived the filter.

- `bed_linens`
- `cooking_basics`
- `wifi`
- `microwave`
- `coffee_maker`
- `air_conditioning`
- `kitchen`
- `refrigerator`
- `tv`
- `heating`
- `elevator`
- `oven`
- `washer`
- `pets_allowed`
- `dedicated_workspace`
- `dryer`
- `paid_parking_off_premises`
- `freezer`
- `dishwasher`
- `free_street_parking`

These are 20 new variables to consider with binary values (representing if it is present in the listing or not).

1.1.3 Selecting interesting variables

Since the number of variables still present in the data was too big (98), we decided to keep those that seemed the most interesting and relevant. We have already seen in Table 1 the ones we chose, where we tried to respect the original requirement of having at least two binary variables, two multiclass categoricals, and six quantitatives.

1.1.4 Removing rows with NA

Once we have our variables selected, we remove the rows with any missing values, which reduces the number of them from 25000 to 14970 (a 59.88% reduction).

1.1.5 Sampling the remaining rows

For the remaining rows, we wanted to make sure not to bias the sample by choosing to many samples from a single neighborhood, since the `Centro` neighborhood constitutes half of the data (6704 rows out of the $ $15000 remaining). To this end, we sample by maintaining the original proportion of listings per neighborhood. In the end, we remain with 1000 variables following said directive. Let us take a look at what we currently have in our data frame by using an univariate table.

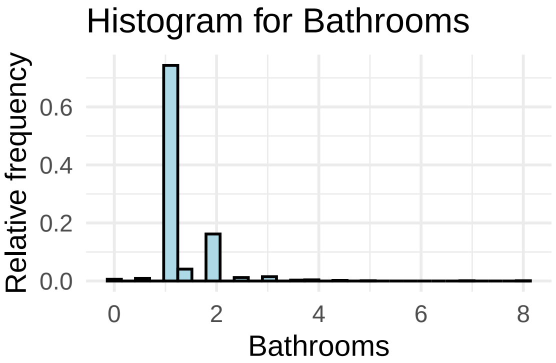
- Neighbourhood representation is preserved. The most important one is 'Centro', as we were mentioning before, with a representation of 45%.
- The most common type of listing is an entire apartment (75%), followed by a private room in someone's house (24%).

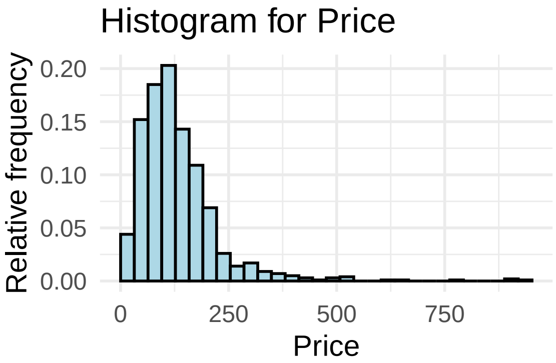
Univariate descriptive analysis of the sample

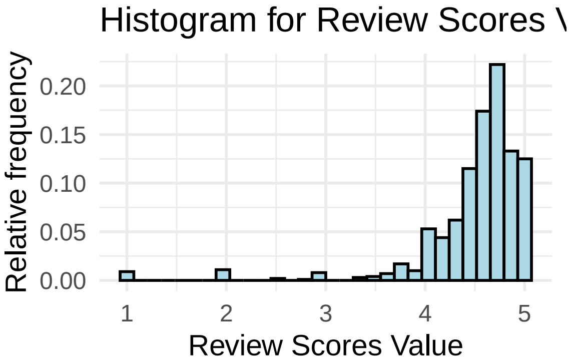
Variables	**N = 1,000**[1]
Host Total Listings Count	7.00 (38.00)
Neighbourhood Group Cleansed	
Arganzuela	51 / 1,000 (5.1%)
Barajas	6 / 1,000 (0.6%)
Carabanchel	35 / 1,000 (3.5%)
Centro	448 / 1,000 (45%)
Chamart	
'in	33 / 1,000 (3.3%)
Chamber	
'i	58 / 1,000 (5.8%)
Ciudad Lineal	32 / 1,000 (3.2%)
Fuencarral - El Pardo	12 / 1,000 (1.2%)
Hortaleza	21 / 1,000 (2.1%)
Latina	23 / 1,000 (2.3%)
Moncloa - Aravaca	22 / 1,000 (2.2%)
Moratalaz	5 / 1,000 (0.5%)
Puente de Vallecas	30 / 1,000 (3.0%)
Retiro	37 / 1,000 (3.7%)
Salamanca	64 / 1,000 (6.4%)
San Blas - Canillejas	20 / 1,000 (2.0%)
Tetu	
'an	66 / 1,000 (6.6%)
Usera	22 / 1,000 (2.2%)
Vic	
'alvaro	3 / 1,000 (0.3%)
Villa de Vallecas	4 / 1,000 (0.4%)
Villaverde	8 / 1,000 (0.8%)
Room Type	
Entire home/apt	751 / 1,000 (75%)
Hotel room	1 / 1,000 (0.1%)
Private room	243 / 1,000 (24%)
Shared room	5 / 1,000 (0.5%)
Accommodates	3.00 (2.00)
Bathrooms	1.00 (0.00)
Bedrooms	
0	65 / 1,000 (6.5%)
1	618 / 1,000 (62%)
2	227 / 1,000 (23%)
3	72 / 1,000 (7.2%)
4	8 / 1,000 (0.8%)
5	5 / 1,000 (0.5%)
6	2 / 1,000 (0.2%)
7	2 / 1,000 (0.2%)
8	1 / 1,000 (0.1%)
Price	112.00 (91.25)
Minimum Nights	2.00 (2.00)
Number Of Reviews	27.00 (77.00)
Estimated Occupancy L365d	96.00 (192.00)

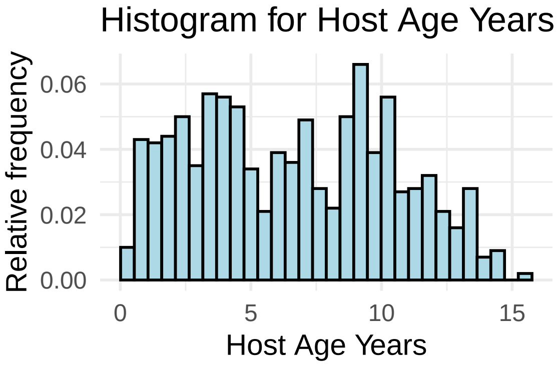
- Listings typically have only one bedroom (62%), and sometimes two (23%).
- The average `price` of listings per night is 112 €, and the average occupation per year is 96 days.

Let us take a look now to the quantitative data by using histograms. Among the available, we have chosen to take a look into the following ones.



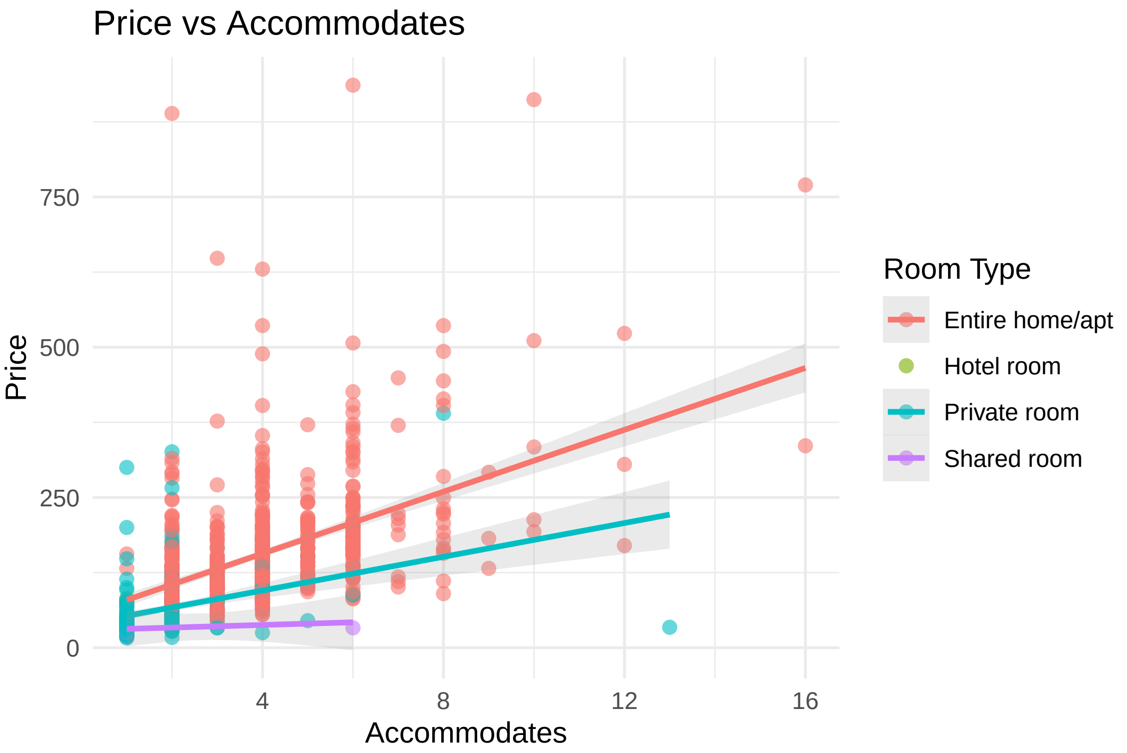




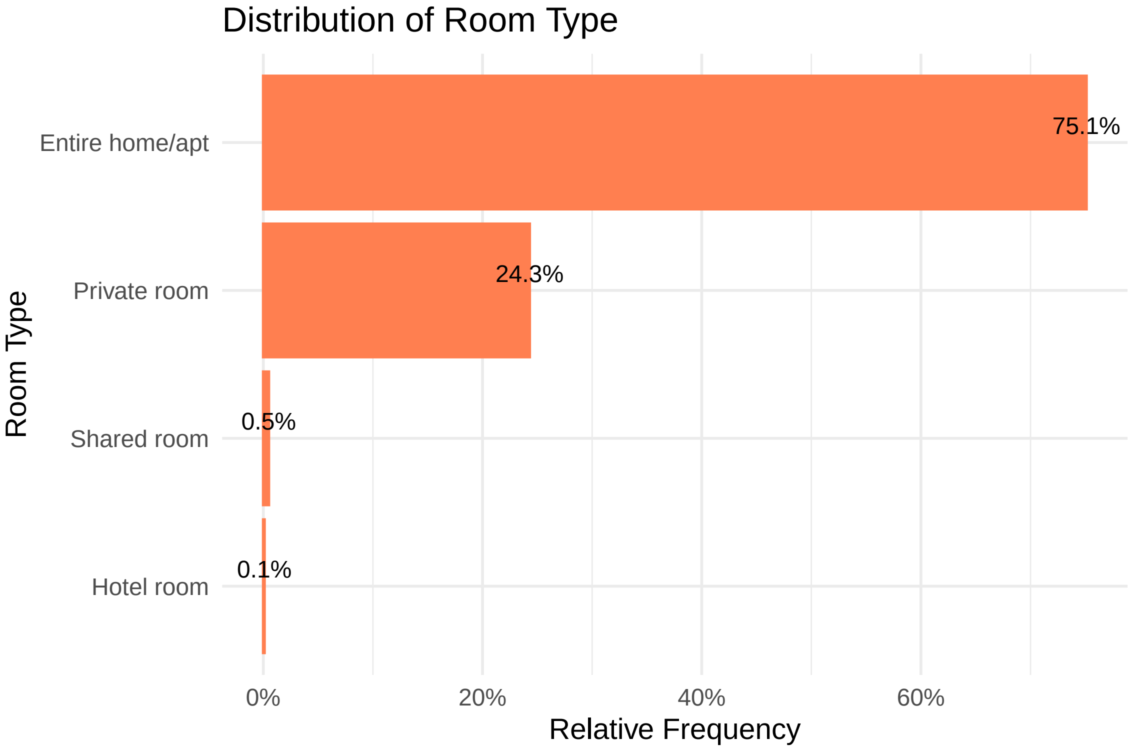


- Interestingly, we see that the bathrooms variable is not an integer but continuous. The reasonable explanation for this is that toilets (without a shower) are considered "half bathrooms". We can see, nevertheless, that listings typically have only one bathroom, sometimes two, and less frequently "one and a half".
- The price histogram is right-skewed because of some high-priced outliers. Average price, as we saw in the univariate table, is 112 € per night, with a standard deviation of ±91.25, which checks with what can be seen in the histogram.
- The review scores are left-skewed, which is to be expected. This types of situation arise frequently in review systems, since value 4 is considered as a good experience instead of 3 as naively we could think. The latter is usually considered as a "bad review".
- The number of years hosts have been with Airbnb does not follow a specific trend. Moreover, data is distributed along the 0-15 years, with the latter having slightly less representation.

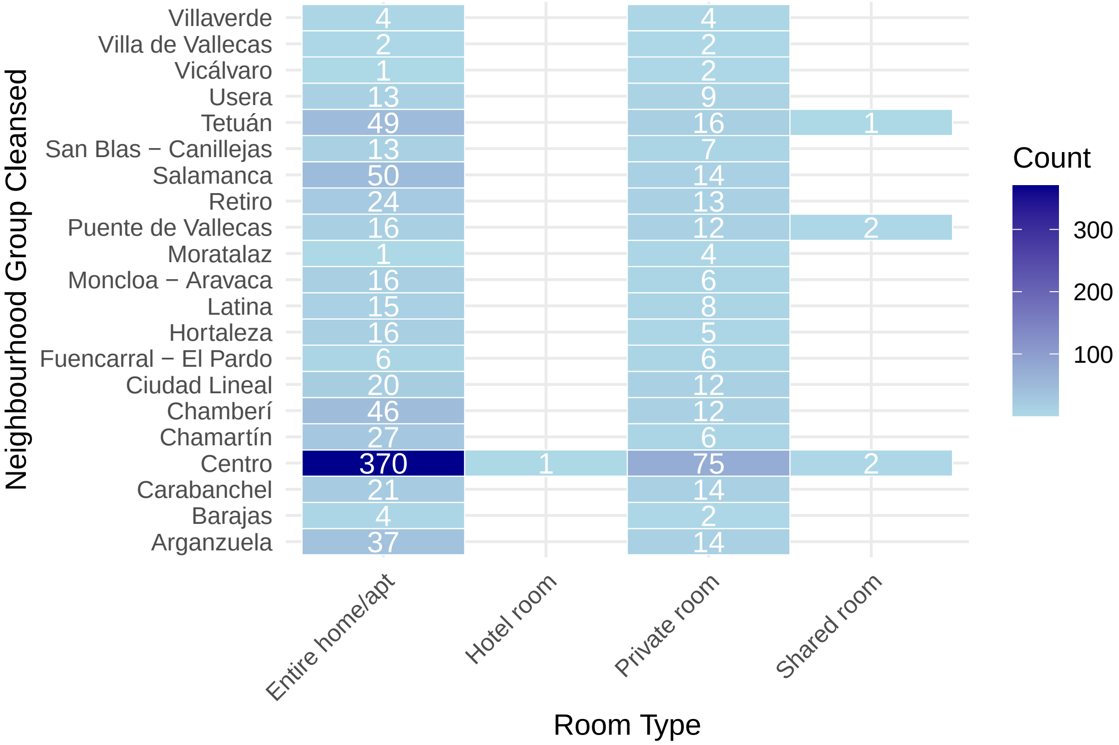
A very reasonable question to have is if the accommodates affects the price, as one could expect. In the following scatter plot we can see that this is the case.



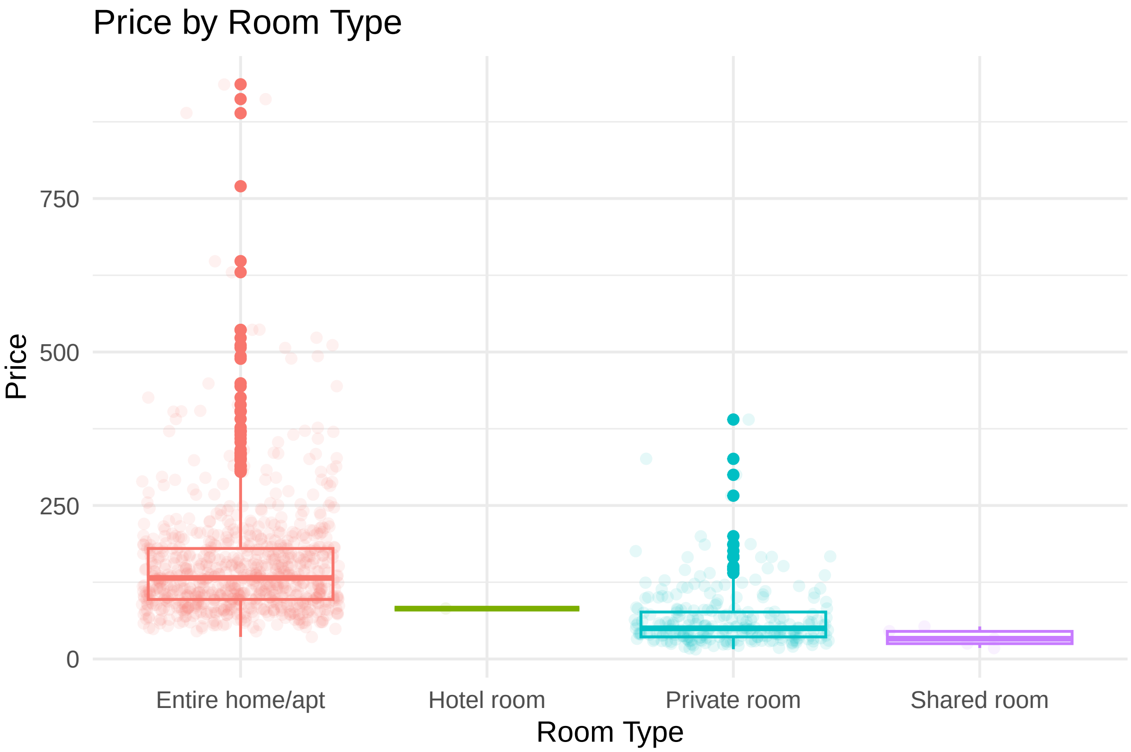
We can see that the tendency for prices is to increase more aggressively for entire apartments than for private rooms. Shared rooms, on the other hand, do not increase the price by the number of accommodates, most likely because there may be many beds in a single room, so there is little difference in having more or less accommodates. The following plot shows, moreover, that shared rooms are rare among the listings, dominated by entire apartments.



In fact, we can use a crosstab-heatmap to show just how clustered is the data around the city center, which is why we decided to respect the original proportions of listings per neighborhood.

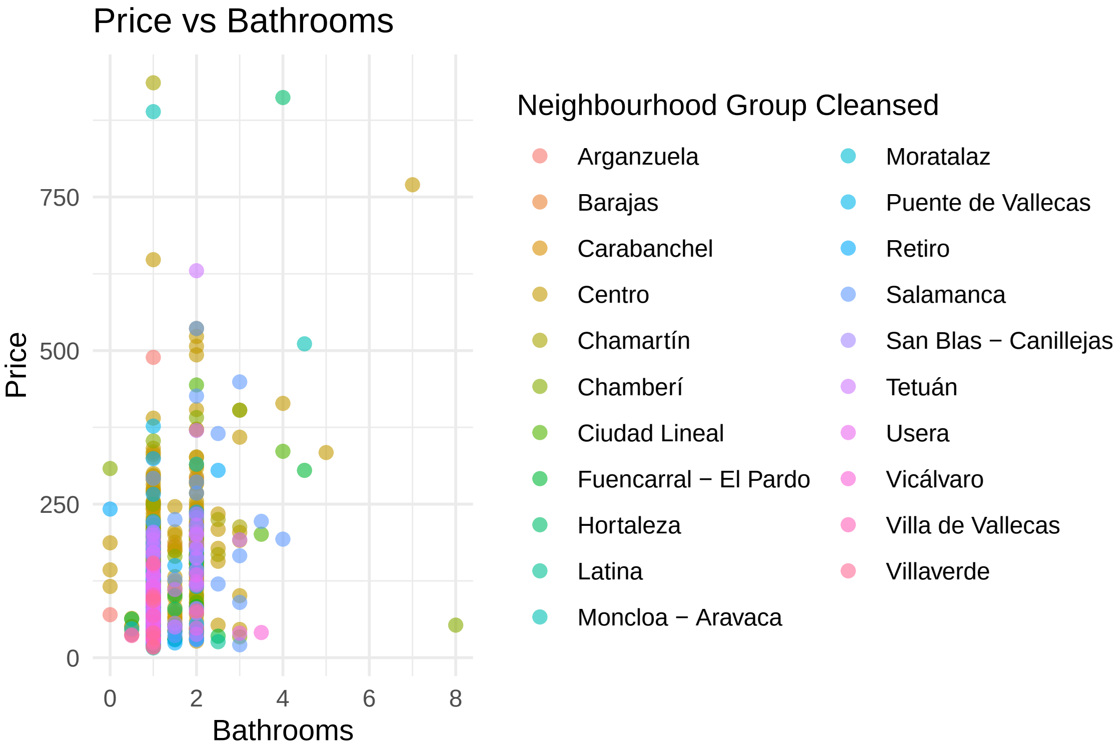


Note, as well, that more than 33% of the data comes from entire apartments in just the city center of Madrid. Let us be now more specific about the distribution of prices based on the type of listing. The following boxplot shows it.

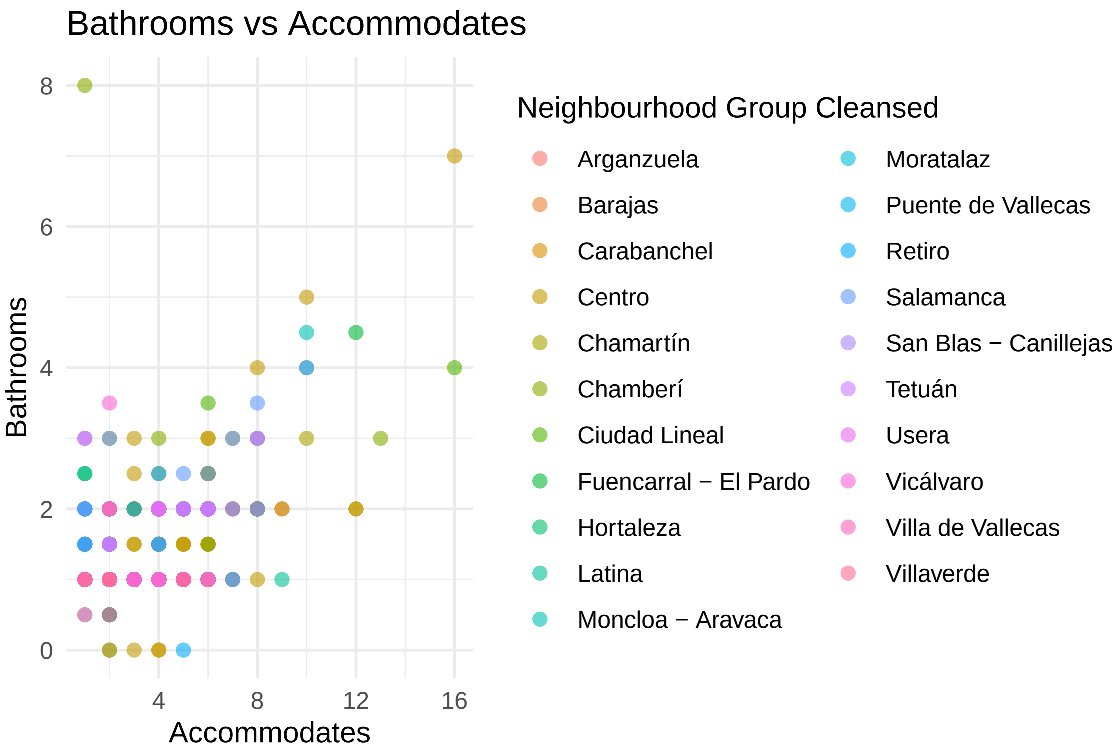


In general, entire apartments are more expensive than private rooms, most likely due to the privacy they provide. We can see as well that less clustering is present in the entire apartment class, with several outliers costing more than double or triple the average price. This effect is less exaggerated in the private room category.

Let us explore a different variable now, such as the number of bathrooms. Is there any tendency in the price according to them?

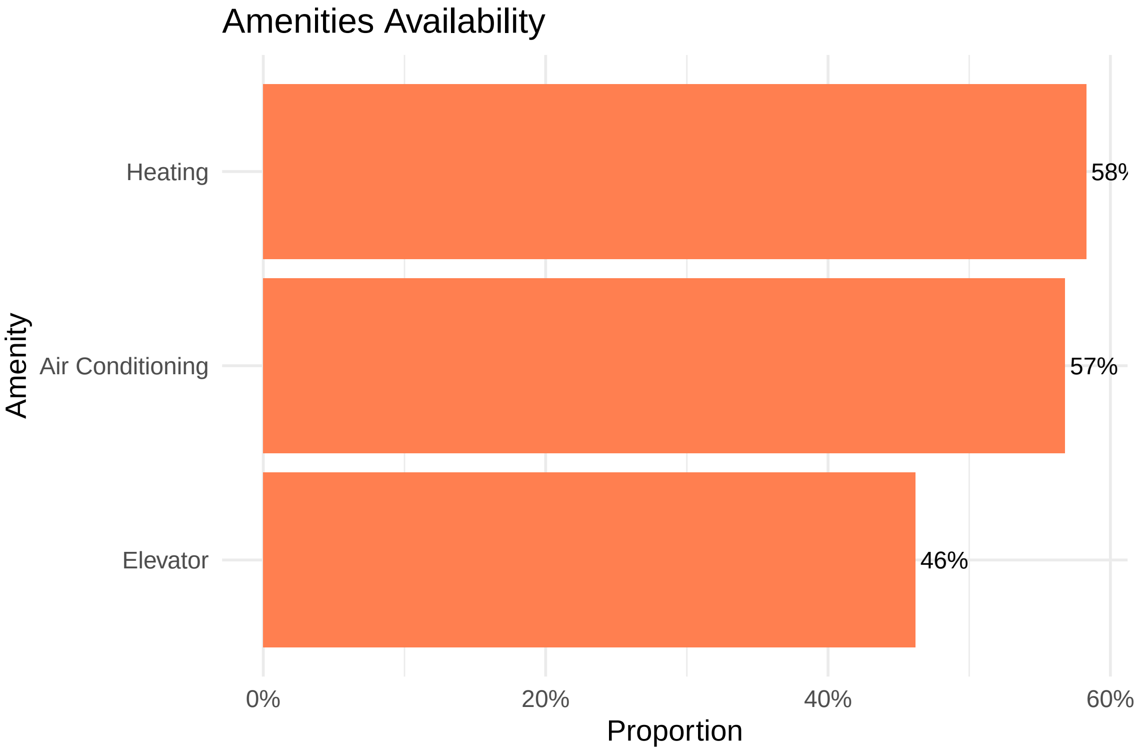


We can see that the data is clustered between 1.5 and 2.5 bathrooms, with no clear tendency. Perhaps it is worth taking a look to the distribution of bathrooms based on the number of accommodates of the listing.



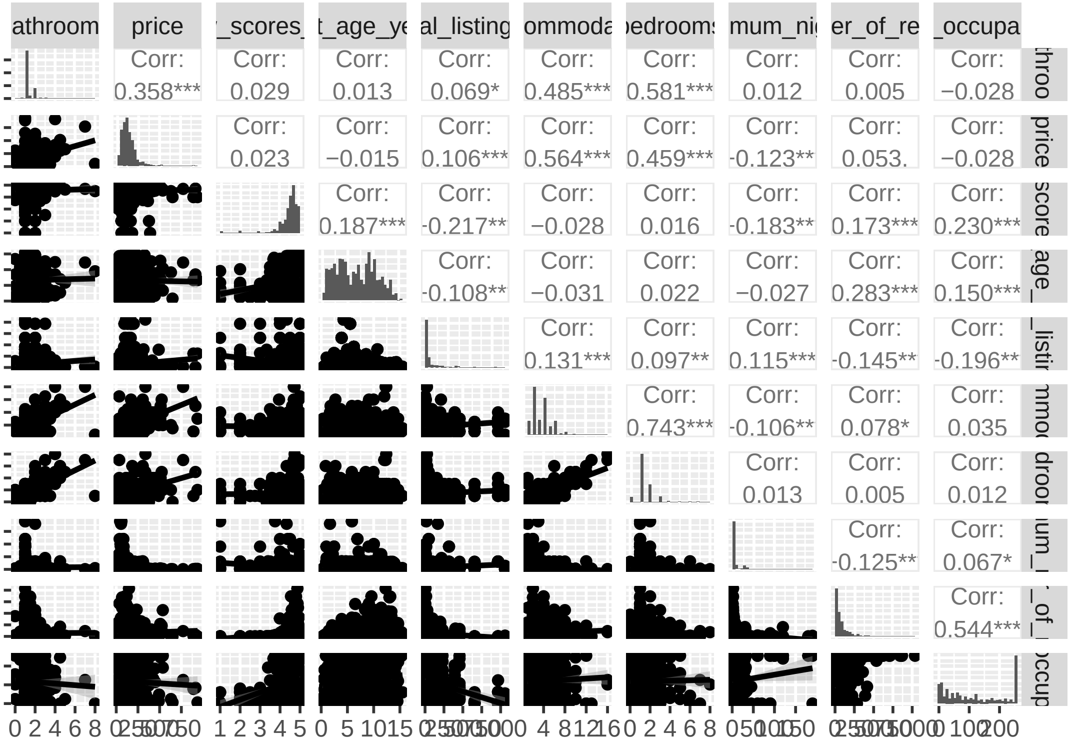
We can see that there is a trend, the number of bathrooms increases with the number of accommodates. Interestingly, we can see some unexpected outliers, such as one apartment in Chamberí with 8 bathrooms for a single accommodate. This is likely an error when inputting the data. We retained this outlier to demonstrate the robustness of PCA later, though we could as well choose to remove it.

For the sake of completion, we show below the distribution of the rest of the amenities.



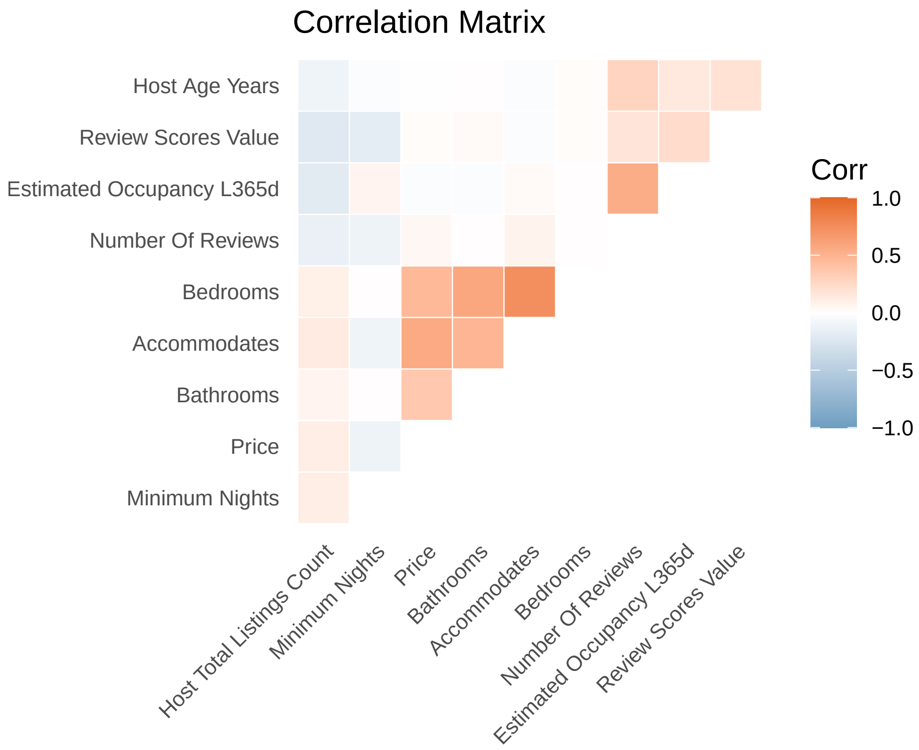
It is interesting to note that around 40% of the listings do not have a heating system, and 40% as well do not have air conditioning. As well, note than more than half of the listings do not posses elevator. The high density of listings in the city center could be an explanation for this, since old buildings are less likely to have elevator.

The last plot of this section shows a summary of what we have been presenting.



2 Correlation analysis

Let us take a look to a correlation matrix to see if there is any between the variables.



We can see that the price, the number of bathrooms, and the accommodates are highly correlated (the last two we had already seen it in the previous section). We can see as well that the occupation per year rate is correlated to number of reviews, which is to be expected.

Lastly, we present a detailed map of the occupation per neighborhood, as well as the average and median price for each of them.

We see that the most expensive district, on average, is Moncloa, followed by Centro and then by Salamanca. However, we can see as well that in Moncloa the median is somewhat far from the mean, suggesting there might be some notable outliers that are inflating the average.