Airbnb Analysis

Matteo Biglioli

07/06/2021

Abstract

This paper will present a complete statistical analysis of Airbnb data from six major european cities. We start by presenting an exploratory analysis of our dataset, in which we try to identify both the most relevant components that drive prices and possible differences between the collected cities. We then evaluate different statistical learning models that predict the prices given different instrumental variables. At the end we generate different clusters both from the whole dataset and from subset related to single cities, to better understand the composition of the dataset.

We find that the most important drivers for prices are (apart from the actual city in which is located the property) the type of the room, weather is shared or private, and the number of guests that can stay at the property, the higher the number, the lower the price. These main variables are followed by others like the rating of the property, the number of bathrooms and the presence of air conditioning.

We estimated a simple pruned tree, a Random Forest and a Boosted tree model which achieve a Root Mean Squared Error from 80 to 73.

We finally try and cluster our data but we find that the clusters have no actual grographical interpretation.

1. Introduction

Airbnb is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. It was born in 2007 and, since then, it has grown to 4 million Hosts who have welcomed more than 900 million guest arrivals in almost every country across the globe. The app works as most of the other bookings apps: a client can just select a city he/she want to visit, the dates of the trip and he/she will get a list of possible locations.

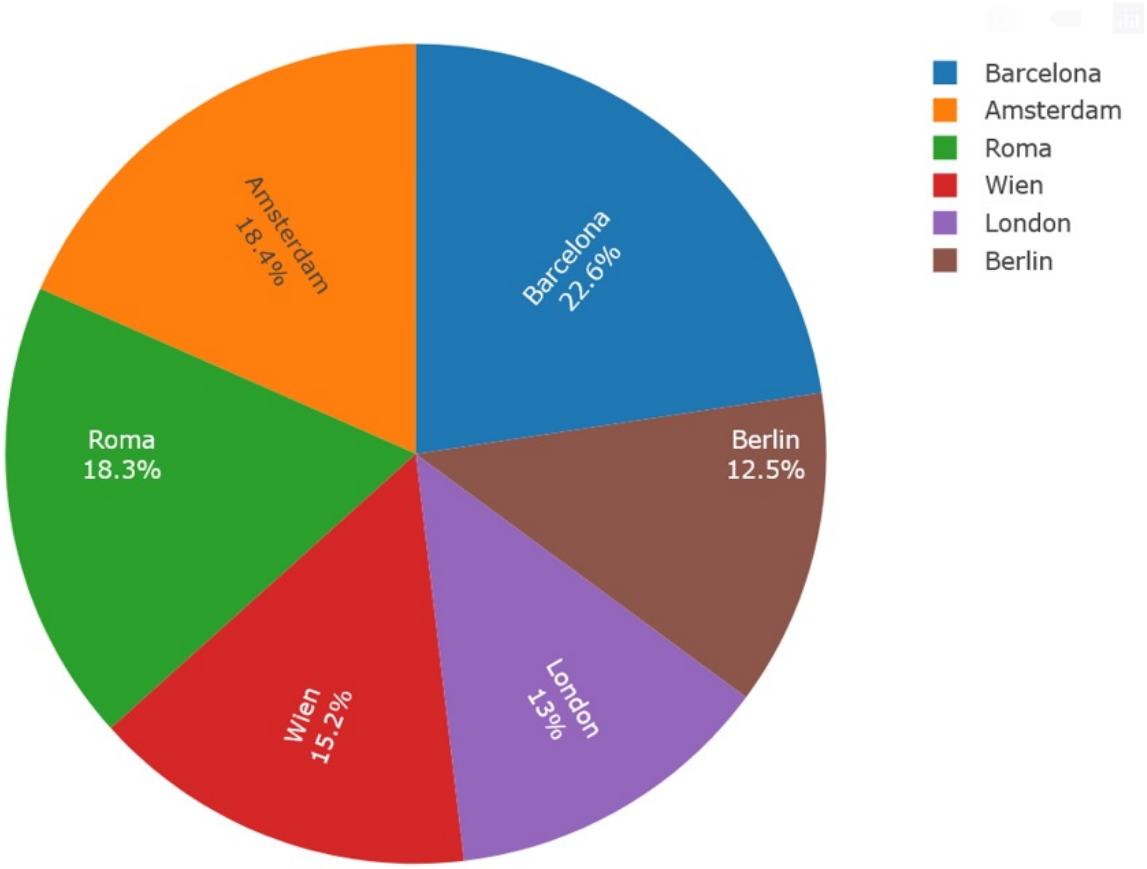
The main feature that makes Airbnb unique in its kind is that in the website you can find not only hotels and apartments, but also single rooms that are rented out for a few days from private Hosts.

In this kind of environment Hosts are pushed to compete in an almost-free market, where they have to set the right price for their properties in order to stay competitive and win guests over. We can assume that a good percentage of private Hosts has no experience in both Marketing and Real-Estate, hence the definition of the right price could become a entry barrier that stops most of them from ever trying to compete.

Our goal in this paper will be to understand which are the components that drive prices and to construct a model that can help both Hosts to define the right price for their properties and Guests to check weather a location is truthfully priced.

2. Data

We start our analysis by presenting our dataset. Before diving into the actual variables we present the six different cities and the related number of records:



For a better understanding we splitted the variables into five different groups that will be presented below.

2.1 Host-related variables

These variables are related to features of the hosts, we have:

- **Host.Since**: number of days since the Host is present on the platform.
- **Host.Response.Time**: average response time of the Host.
- **Host.Response.Rate**: average rate of responses of the Host.
- **Host.ProfilePic**: weather the Host has a profile pic.
- **Host.SuperHost**: weather the Host is a SuperHost.
- **Host.verified**: weather the Host is verified.

```
##    Host.Since            Host.Response.Time Host.Response.Rate Host.ProfilePic
##  Min.   :1597   a few days or more: 103     Min.   :  0.00     Mode :logical
##  1st Qu.:2170   within a day      :1420     1st Qu.:100.00     FALSE:16
##  Median :2561   within a few hours:2305     Median :100.00     TRUE :9709
##  Mean   :2624   within an hour    :5897     Mean   : 96.38
##  3rd Qu.:3040                               3rd Qu.:100.00
##  Max.   :4571                               Max.   :100.00
##  Host.SuperHost  Host.verified
##  Mode :logical   Mode :logical
##  FALSE:7603      FALSE:3226
##  TRUE :2122      TRUE :6499
##
##
##
```

2.2 Review-related variables

These variables are related to the different reviews, we have:

- **Number.of.Reviews**: number of reviews of the property.
- **Review.Scores.Rating**: overall rating (1-100).
- **Review.Scores.Accuracy**: rating related to the accuracy of the host.
- **Review.Scores.Cleanliness**: rating related to the cleanliness of the host.

- **Review.Scores.Checkin**: rating related to checkin.
- **Review.Scores.Communication**: rating related to the communication of the host.
- **Review.Scores.Location**: rating related to the location.
- **Review.Scores.Value**: overall rating (1-10).

```
##  Number.of.Reviews Review.Scores.Rating Review.Scores.Accuracy
##  Min.   : 11.00    Min.   : 54.00       Min.   : 5.000
##  1st Qu.: 17.00    1st Qu.: 90.00       1st Qu.: 9.000
##  Median : 29.00    Median : 94.00       Median :10.000
##  Mean   : 43.71    Mean   : 92.59       Mean   : 9.543
##  3rd Qu.: 54.00    3rd Qu.: 97.00       3rd Qu.:10.000
##  Max.   :417.00    Max.   :100.00       Max.   :10.000
##  Review.Scores.Cleanliness Review.Scores.Checkin Review.Scores.Communication
##  Min.   : 5.000            Min.   : 6.000        Min.   : 6.000
##  1st Qu.: 9.000            1st Qu.: 9.000        1st Qu.: 9.000
##  Median :10.000            Median :10.000        Median :10.000
##  Mean   : 9.407            Mean   : 9.698        Mean   : 9.707
##  3rd Qu.:10.000            3rd Qu.:10.000        3rd Qu.:10.000
##  Max.   :10.000            Max.   :10.000        Max.   :10.000
##  Review.Scores.Location Review.Scores.Value
##  Min.   : 6.000         Min.   : 5.000
##  1st Qu.: 9.000         1st Qu.: 9.000
##  Median :10.000         Median : 9.000
##  Mean   : 9.439         Mean   : 9.272
##  3rd Qu.:10.000         3rd Qu.:10.000
##  Max.   :10.000         Max.   :10.000
```

2.3 Price-related variables

These variables are related to prices and fees, we have:

- **Price**: price per night.
- **Security.Deposit**: eventual security deposit (if NA, is set to 0).
- **Cleaning.Fee**: eventual cleaning fee, una tantum (if NA, is set to 0).

```
##      Price        Security.Deposit  Cleaning.Fee
##  Min.   :  9.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 45.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 71.00   Median :  0.00   Median : 20.00
##  Mean   : 88.51   Mean   : 96.94   Mean   : 25.39
##  3rd Qu.:109.00   3rd Qu.:150.00   3rd Qu.: 40.00
##  Max.   :900.00   Max.   :999.00   Max.   :400.00
```

2.4 Services-related variables

These variables are related to services included in booking, we have:

- **Instant_Bookable**: weather a property is instantly bookable (without hosts' confirmation).
- **Kitchen**: weather a kitchen is avaiable.
- **Washer**: weather a washer is avaiable.
- **Breakfast**: weather breakfast is included.
- **Air_conditioning**: weather air conditioning is avaiable.
- **Experiences.Offered**: kind of experience (romantic, business, …).
- **Cancellation.Policy**: cancellation policy (fees, …).
- **Minimum.Nights**: minumin number of nights to spend at the property.

```
##  Instant_Bookable  Kitchen          Washer        Breakfast
##  Mode :logical    Mode :logical   Mode :logical   Mode :logical
##  FALSE:6282       FALSE:845       FALSE:2561      FALSE:8623
##  TRUE :3443       TRUE :8880      TRUE :7164      TRUE :1102
##
##
##  Air_conditioning Experiences.Offered      Cancellation.Policy
##  Mode :logical    business:  54       flexible       :1700
##  FALSE:7039       family  :  29       moderate       :3578
##  TRUE :2686       none    :9586       strict         :4441
##                   romantic:  16       super_strict_30:   5
##                   social  :  40       super_strict_60:   1
```

2.5 Accomodation-related variables

These variables are related to the actual location, we have:

- **Property.Type**: type of the property.
- **Room.Type**: type of the room.
- **Accommodates**: number of people that can stay at the property.
- **Bathrooms**: number of bathrooms.
- **Bedrooms**: number of bedrooms
- **Beds**: number of beds
- **Bed.Type**: type of the beds

```
##          Property.Type            Room.Type     Accommodates      Bathrooms
##  Apartment      :8201   Entire home/apt:6239   Min.   : 1.000   Min.   :0.000
##  House          : 572   Private room   :3403   1st Qu.: 2.000   1st Qu.:1.000
##  Bed & Breakfast: 396   Shared room    :  83   Median : 3.000   Median :1.000
##  Condominium    : 153                          Mean   : 3.382   Mean   :1.204
##  Loft           : 133                          3rd Qu.: 4.000   3rd Qu.:1.000
##  Boat           :  90                          Max.   :16.000   Max.   :8.000
##  (Other)        : 180
##     Bedrooms           Beds                Bed.Type    Minimum.Nights
##  Min.   : 0.000   Min.   : 1.000   Airbed       :   8   Min.   :  1.000
##  1st Qu.: 1.000   1st Qu.: 1.000   Couch        :  29   1st Qu.:  1.000
##  Median : 1.000   Median : 2.000   Futon        :  69   Median :  2.000
##  Mean   : 1.357   Mean   : 2.081   Pull-out Sofa: 202   Mean   :  2.749
##  3rd Qu.: 2.000   3rd Qu.: 2.000   Real Bed     :9417   3rd Qu.:  3.000
##  Max.   :10.000   Max.   :16.000                        Max.   :180.000
##
```

3. Definition of the dependent variable

The main goal of this analysis is to find what drive prices, the main issue we need to address before diving in is: **How do we define *Price*?**.

We observe that there are different variables related to prices and fees, that are:

- Total price per night
- Cleaning fee
- Security deposit

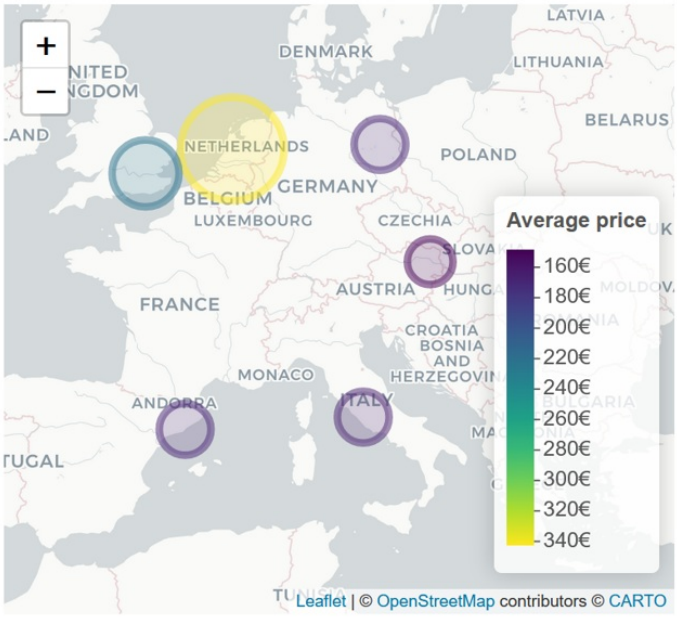
For the scope of this paper we decide that we will define price as the per-person price of a seven-night stay at the property, computed as:

```
# Compute new price and remove useless columns
df = df %>% mutate(Price = (7*Price + Cleaning.Fee)/Accommodates) %>% select(-Cleaning.Fee)
```

4. Exploratory analysis

We now present our dataset in a more sofisticated way, focusing on the relationships between the different variables that we collected and the prices.

We start by showing how prices differs by city:



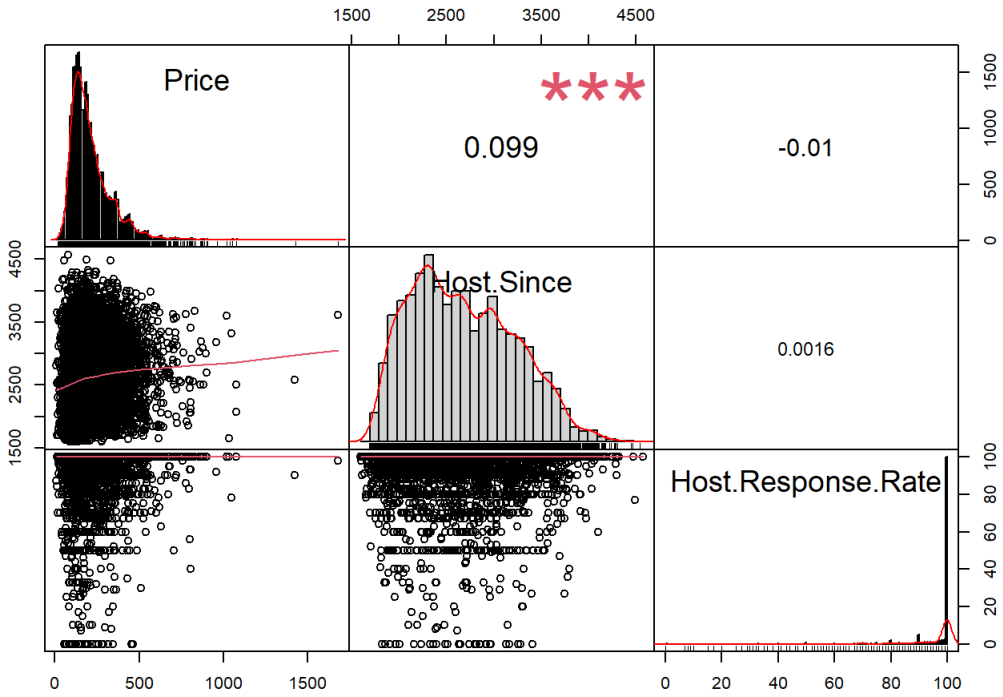
We can see from the two graphs above that 4 out of the 6 cities are quite similar (Barcelona, Rome, Wien and Berlin), with an average 7-nights per-person price of about 160€. The two “outliers” are London, with a value of 220€ and Amsterdam, which unexpectedly shows a price of almost 350€.

Because we collected a sample dataset of less than 10k observations, we know that our analysis cannot assess that Amsterdam is overrall the most expensive city; that is because in the investigation above we do not control for other factors, such as accomodation and service features.

Proceeding with the order we used in the **Data** section, we will now try to assess if the different groups of variables could be considered a driver for prices.

4.1 Host-related variables

We show the relationship between Host-related variables and our prices.



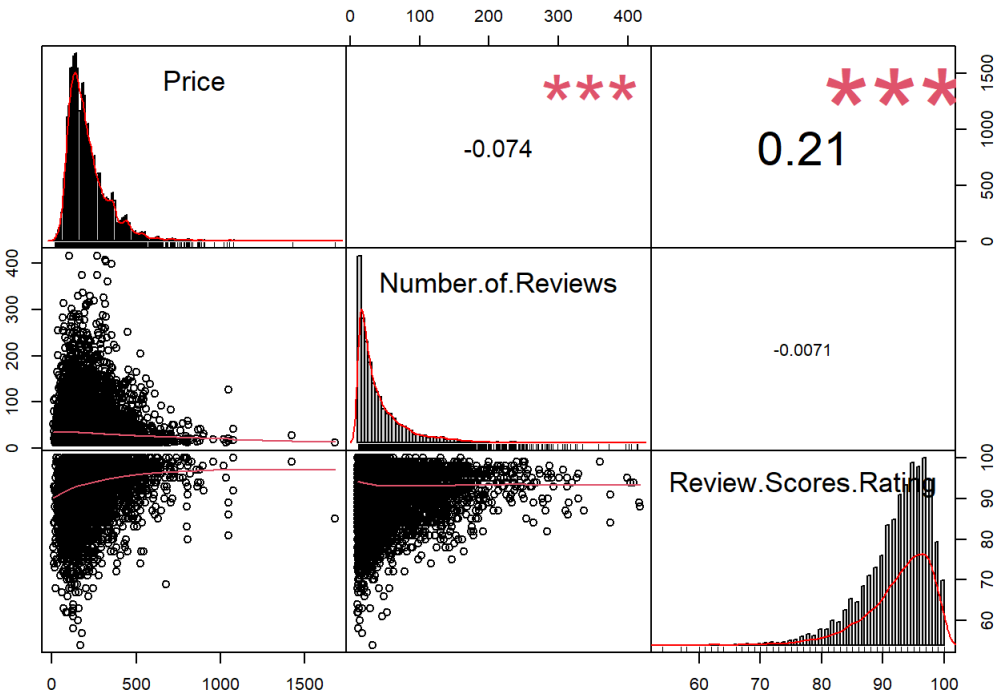
We observe that there is no strong correlation between this group of variables and the prices we computed.

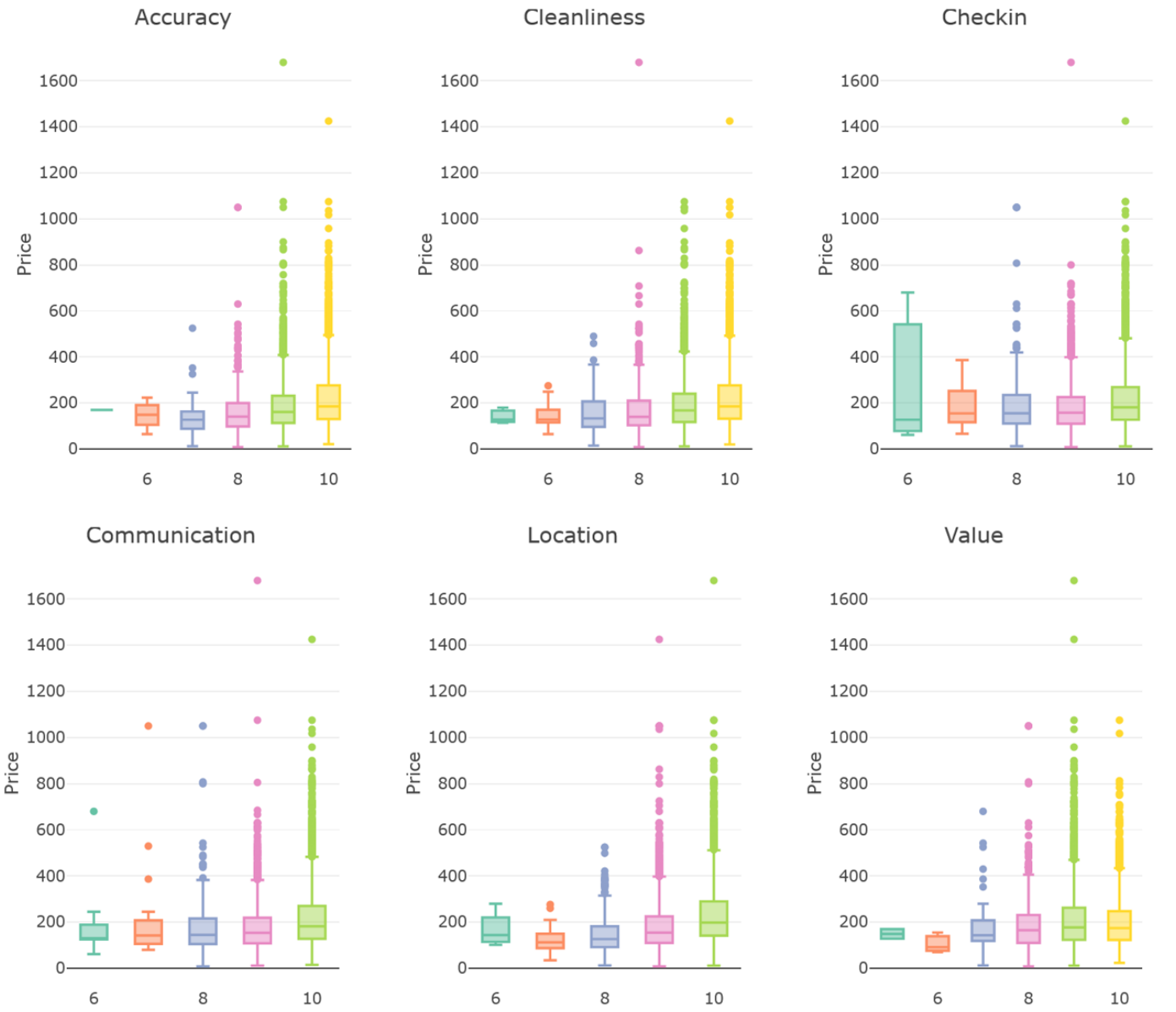
From these graphs we can extract two main findings:

- There is a slight but significant correlation between the number of days since the host became active on the platform and the price; this could be due to the fact that experienced host tend to increment the value of their properties over time, maybe adding feature and services that they found relevant in this market.
- It seems that higher prices are in some way correlated with faster response time (note that we cannot assess it firmly due to the fact that the confidence bands overlaps, but we can look at the outliers number of each group). This could be an example of a causation-correlation problem: we could (wrongly) suppose that faster response times drive higher prices, but more probably it is higher prices, related to more valuable accomodations, that justify faster response times.

4.2 Review-related variables

We show the relationship between review-related variables and our prices.





Before commenting the graphs we need to explain why different kind of reviews (Rating as opposed to the other types) are shown in different plots: that is because the Rating variables could take values in [1, 100] while the other ratings' domain is just [1, 10] (most of the times even [5, 10]), effectively making them qualitative variables instead of quantitative ones.

We observe that there is no strong correlation between the single-category ratings and our prices, this is shown in the boxplots above but we could easily expect this from the summary reported in the **Data** section, where whe showed that for all 6 variables, the 1st Quantile is 9, meaning that 75% of our data assume values between 9 and 10, making the added information of this variables useless.

We still observe small but significant correlations between Number.of.Reviews, Review.Scores.Rating and our prices, we can comment these results as follows:

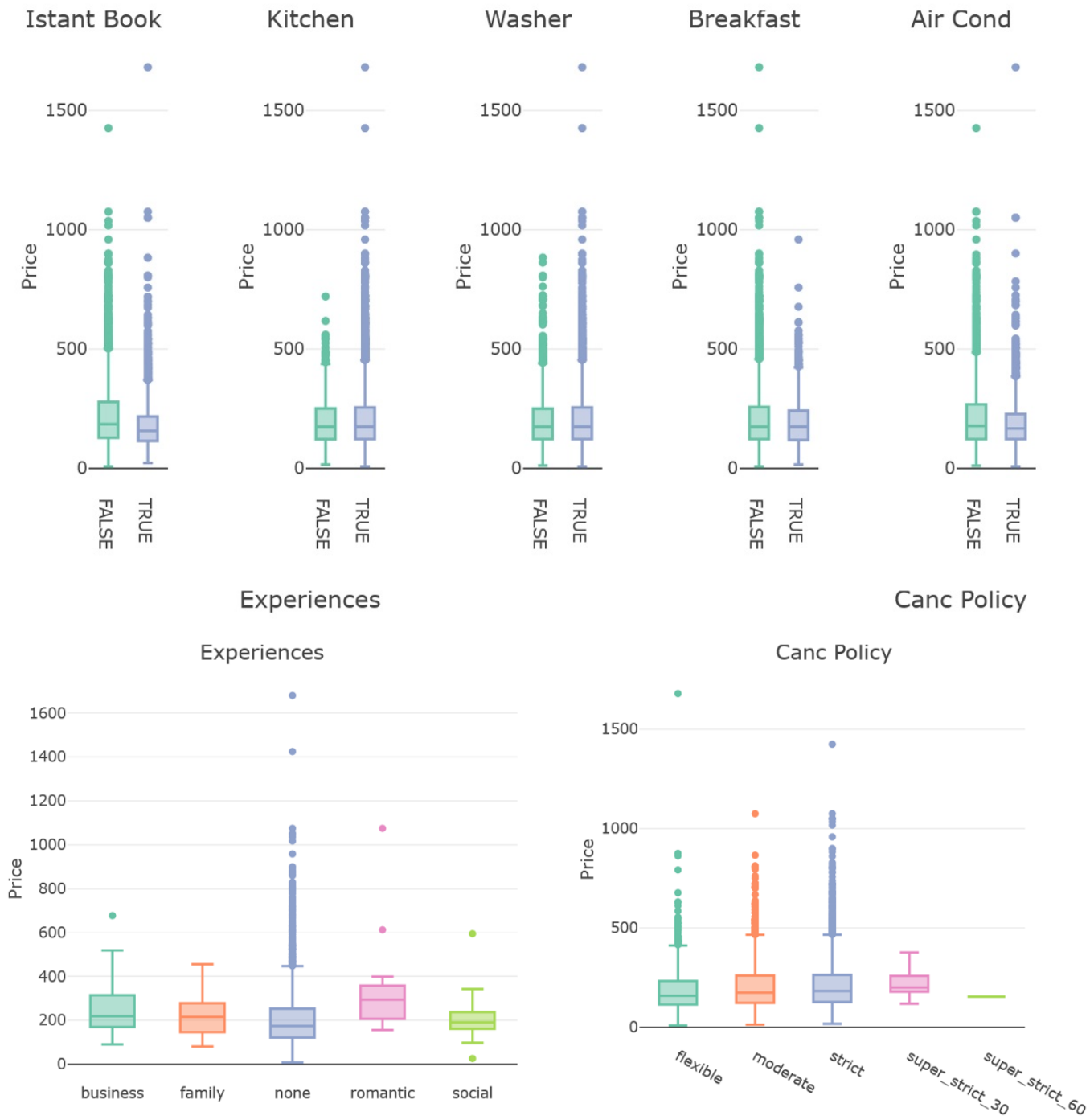
- The correlation between the number of reviews and the prices could be driven by the fact that most people leave only negative reviews, that is because the process of reviewing is long and most people follow it just to get some kind of "revenge" over their negative experience.
- The correlation between overrall rating and price is significant as expected, we could interpret this in two ways:
    - A property with better rating could easily increase its price knowing that guests would still be attracted by it.
    - A property with higher price (and consequently higher value) would be rated higher by Airbnb guest, this is due to the fact that the usual Airbnb guest may be used to low-price and low-value properties (given the fact that most of the properties on the platform are rented rooms).

Because we noticed, both here and in **Data** section, that most of the review variables do not actually add information to our dataset, we remove them as follow:

```
df = df %>% select(-Review.Scores.Accuracy, -Review.Scores.Cleanliness, -Review.Scores.Checkin,
                   -Review.Scores.Communication, -Review.Scores.Location, -Review.Scores.Value)
```

4.3 Services-related variables

We show the relationship between services-related variables and our prices.



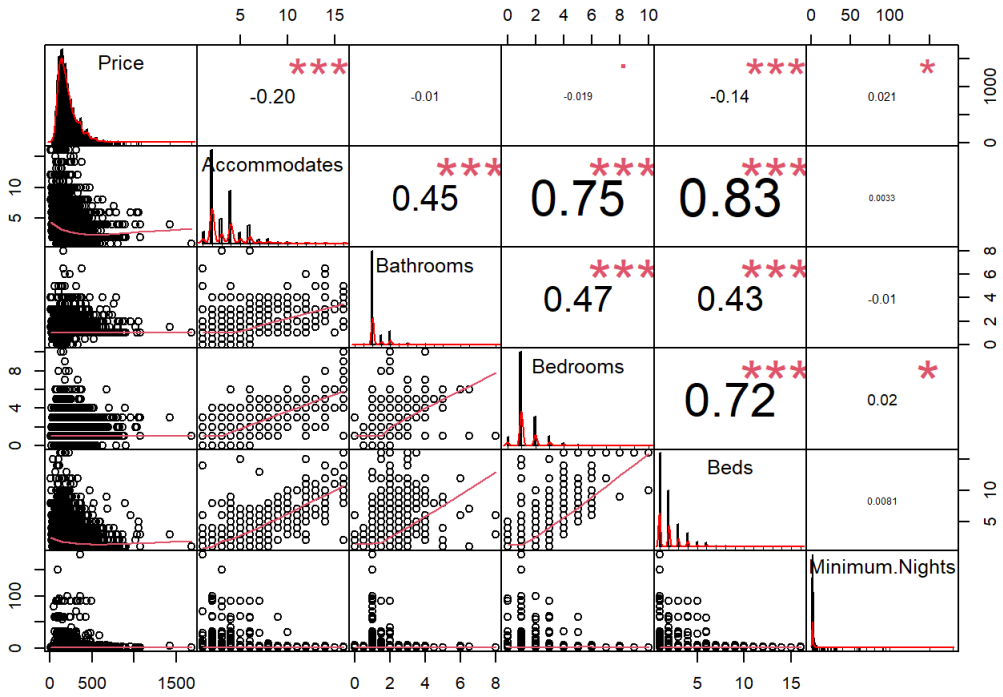
We start by commenting the basic services (T/F services): the boxplots show that there is no relevant correlation between the presence of one of the services addressed and an higher price; this could be due to the fact that while there are properties that cost more due to this additional services, there are also properties, like hotels, which have an high price even without this kind of services.

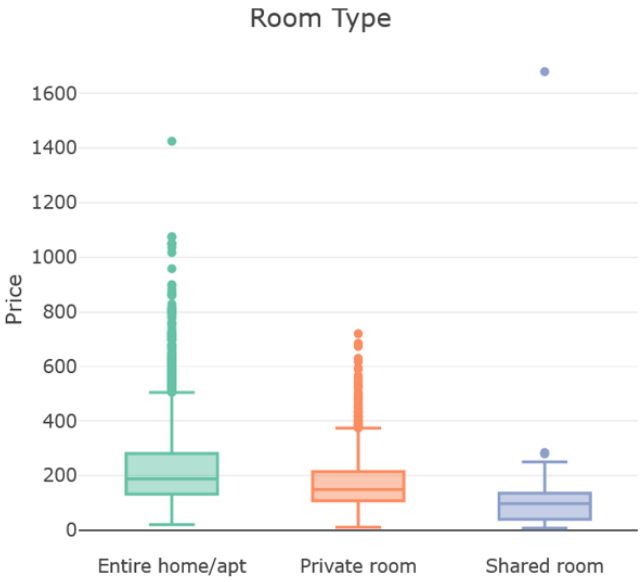
For what it concerns the other two variables we have that:

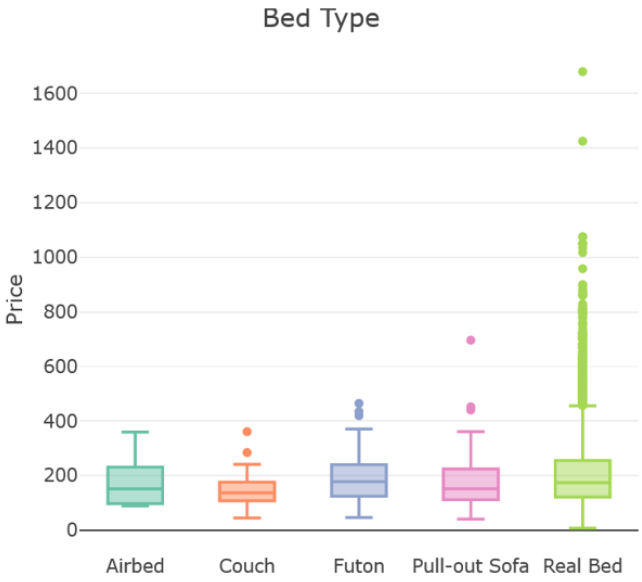
- Romantic experiences seems to have a positive impact on prices, this could be explained by the fact that this kind of properties targets the most ready-to-spend market while business, family and social trips tend to look for ceapher properties.
- The more strict the cancellation policy, the higher the price; this is another example of a causation-correlation problem, in fact here we could easily assess that hosts who own properties with higher value implements stricter cancellation policies in order not to lose high-spender guests.

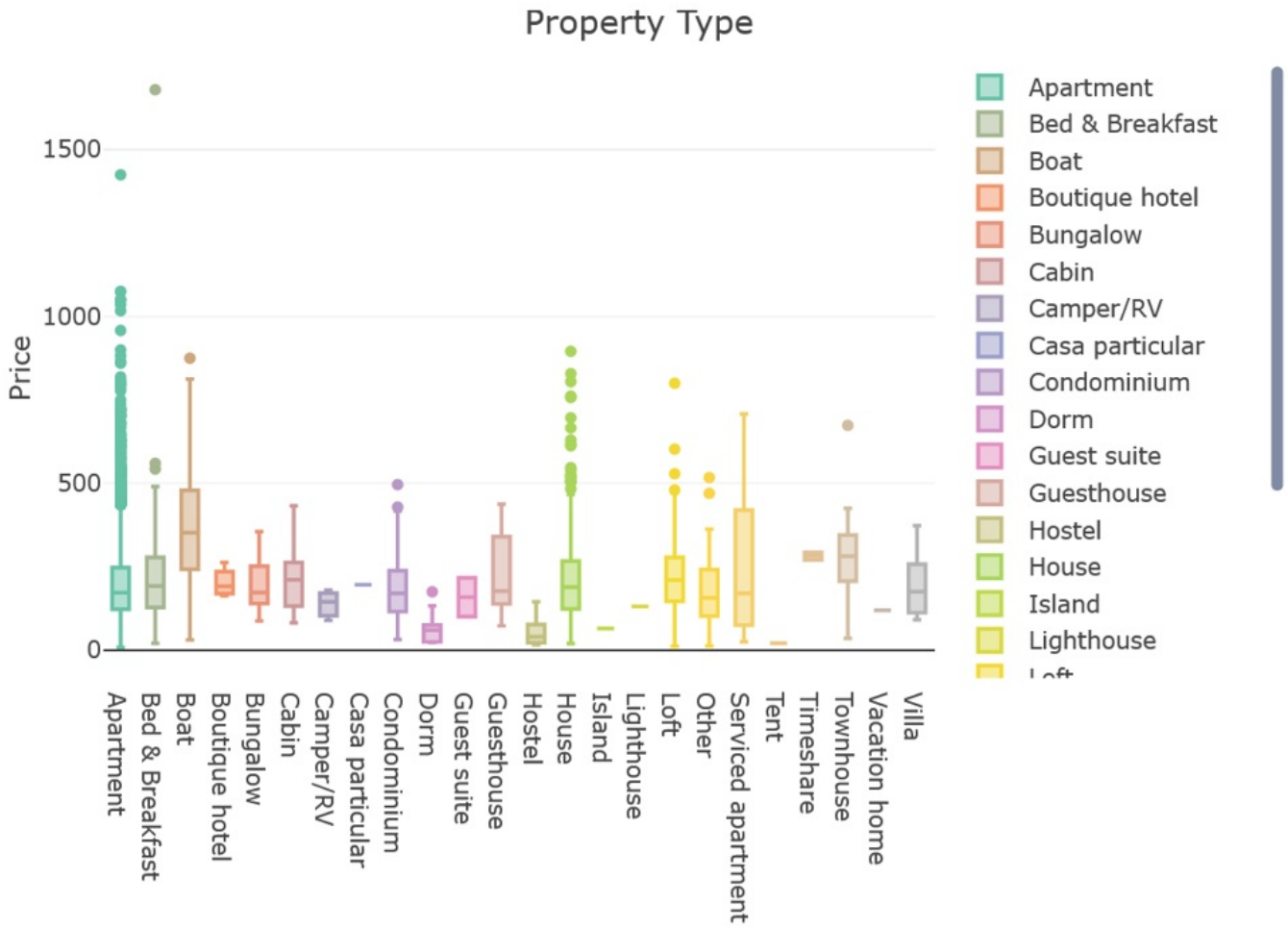
4.4 Accomodation-related variables

We show the relationship between accomodation-related variables and our prices.


Price
Accommodates
Bathrooms
Bedrooms
Beds
Minimum.Nights
-0.20
-0.01
-0.019
-0.14
0.021
0.45
0.75
0.83
0.0033
0.47
0.43
-0.01
0.72
0.02
0.0081


Room Type
Price
Entire home/apt
Private room
Shared room


Bed Type
Price
Airbed
Couch
Futon
Pull-out Sofa
Real Bed



As expected we have that higher-value room and bed type drive higher prices, this could be seen in the first two graphs which show that a real bed and an entire home are easily correlated with higher prices with respect to couches and shared rooms.

For what it concerns property type we have a variety of them, and, as expected, high-end properties like boats and serviced apartments are correlated to higher prices with respect to low-end ones, like dorms, hostels and tents.

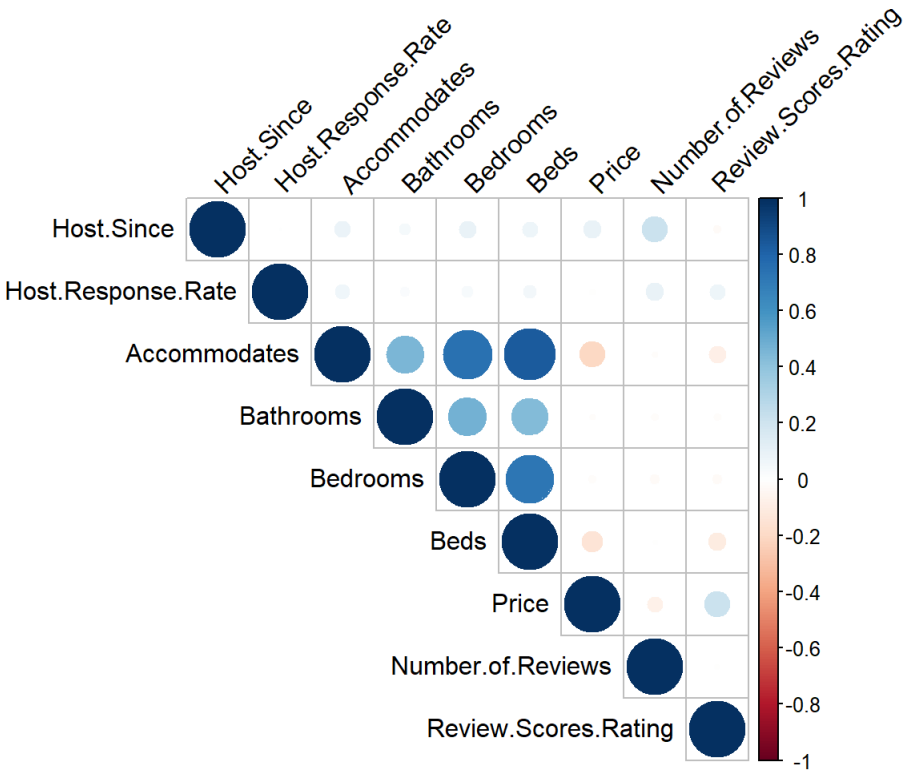
Finally we see that the features of the property, like number of beds/bedrooms, bathrooms and guests are highly correlated with each other as expected. We also find that Accomodates and Beds have a slightly negative but significant impact on price, that could be because properties that accept a lot of guests (like big family houses) leads to discounts on the price-per-person (this dynamic can be seen in all booking platforms, in which renting a property with a lot of people is cheaper than booking a single or double room).

We do find just a slightly relevant correlation with the minumum number of nights to spend at the property; for this reason and for the fact that 75% of our data has a minimum of 3 nights, we exclude the variable from our model as below:

```
df = df %>% select(-Minimum.Nights)
```

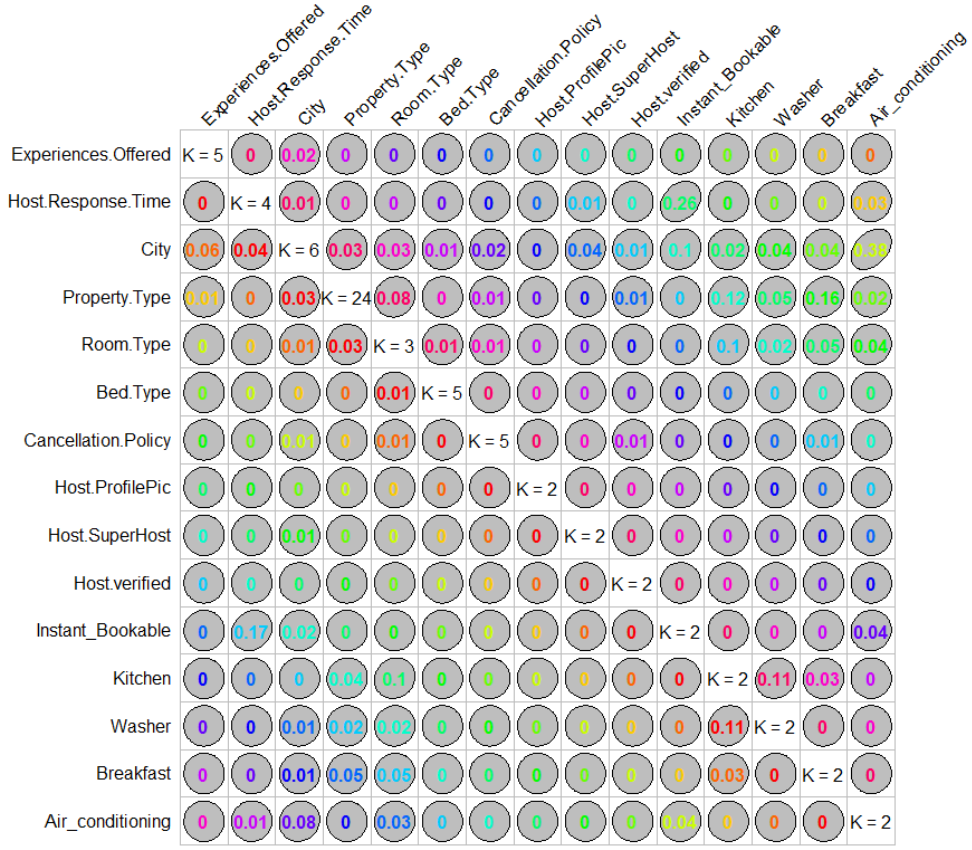
5. Correlation between variables

We will now show the **qualitative** and **quantitative** correlations between the variables of our dataset:



From the graph above we can see that there is an high correlation between the variables which define the actual property, this is basically due to the fact that a high number of beds leads to a lot of bathrooms for the guests which leads to high square footage etc. This first interpretation is pretty basic and we will not dive deeper into that.

As spotted before we can also see the small correlation between Host.Since and Number of reviews, this could be explained by the fact that a property listed for a long time gets more guests and consequently more reviews.



From the graph above we can start drawing some key findings:

- There is a correlation between instant_bookable and host response time which could be explained by the fact that highly active hosts implement the “instant bookable” feature because they know that they will easily accept guests even with a short notice.
- There is a correlation between City and Air conditioning; this could be easily explained by the fact that city with a hotter weather needs to specify the presence of air conditioning that could easily move the choice of the guests.
- There is a correlation between property type and breakfast, explained by the fact that high-end properties like hotels and entire apartments easily offer breakfast to their guests while low-end ones, like Hostels and Tents, know that there would be no market for

this kind of add-on between their customers.
- Finally we notice a correlation between the presence of a Kitchen and the presence of a Washer, easily reconducted to the fact that more “complete” properties like Entire Houses or Apartments usually present both.

6. Supervised learning

After this exploratory analysis, which helped us to better visualize and refine our dataset, we start with a supervised learning approach. The goal here would be to build a model that can accurately predict the price of a property given its features; this kind of model could be implemented in different use cases such as:

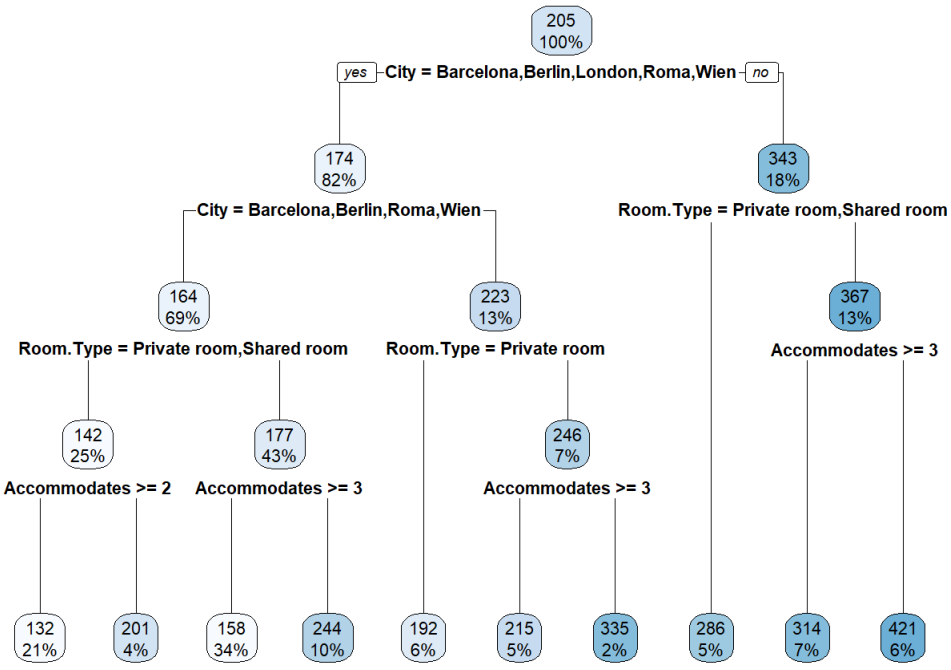
- **Airbnb price recomendation**: the company could help Hosts to evaluate their property in order to be competitive on the market.
- **Third party check for guests**: guests could take profit of a third-party model that could easily spot properties with extremely low/high prices with respect to the estimated value.

6.1 Basic models

Given the composition of our dataset, that present both quantitative and qualitative variables, we decide to implement a Tree-based algorithm. This kind of algorithm is definetly efficent in hybrid datasets and will provide us with great interpretability; in this way we can both get an efficent model and understand which are the main variables that drive Airbnb prices.

We start by estimating a tree over the whole dataset (80-20 train-test split).

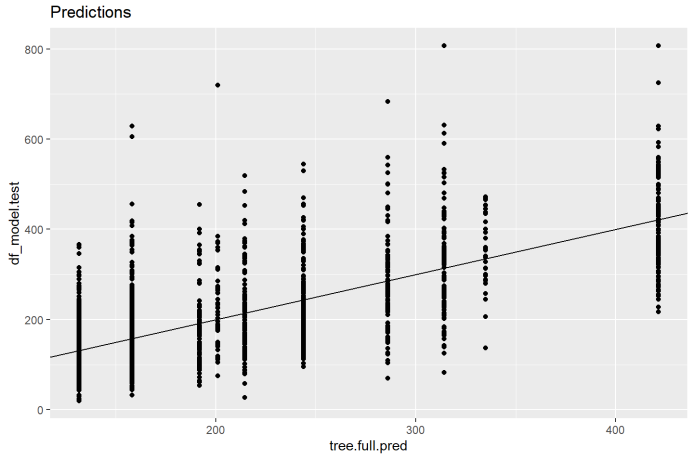
```
##
## Regression tree:
## rpart(formula = Price ~ ., data = df_model, subset = train.full)
##
## Variables actually used in tree construction:
## [1] Accommodates City         Room.Type
##
## Root node error: 113060445/7780 = 14532
##
## n= 7780
##
##         CP nsplit rel error  xerror     xstd
## 1 0.296703      0   1.00000 1.00030 0.037903
## 2 0.025776      1   0.70330 0.70400 0.032991
## 3 0.021319      4   0.62597 0.62826 0.031799
## 4 0.010646      6   0.58333 0.58750 0.031629
## 5 0.010373      7   0.57269 0.57253 0.031530
## 6 0.010000      9   0.55194 0.57041 0.031993
```

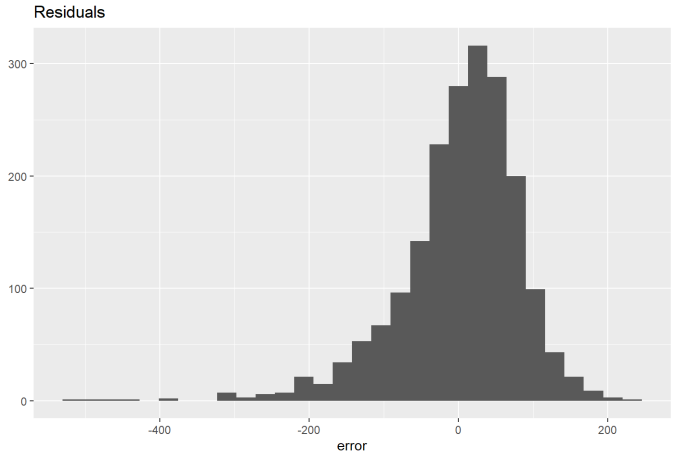


From the tree structure we can easily identify the main drivers for our price variables:

- **City**: we start by splitting Amsterdam from the other cities, this is due to the fact that, as shown above, this city present relatively higher price than the others. At the second step we also construct a different subtree for London which, as shown above, present prices that are lower than the ones in Amsterdam but still definetly higher than the other cities.
- **Room Type**: all the subtrees proceed by separating Private/Private&Shared rooms from entire apartments.
- **Accomodates**: at the end all trees proceed by splitting by number of guests that the property can hold, where, as discussed above, more people in the same property lead to lower prices.

We now proceed to predict the values for our test set and evaluate them.

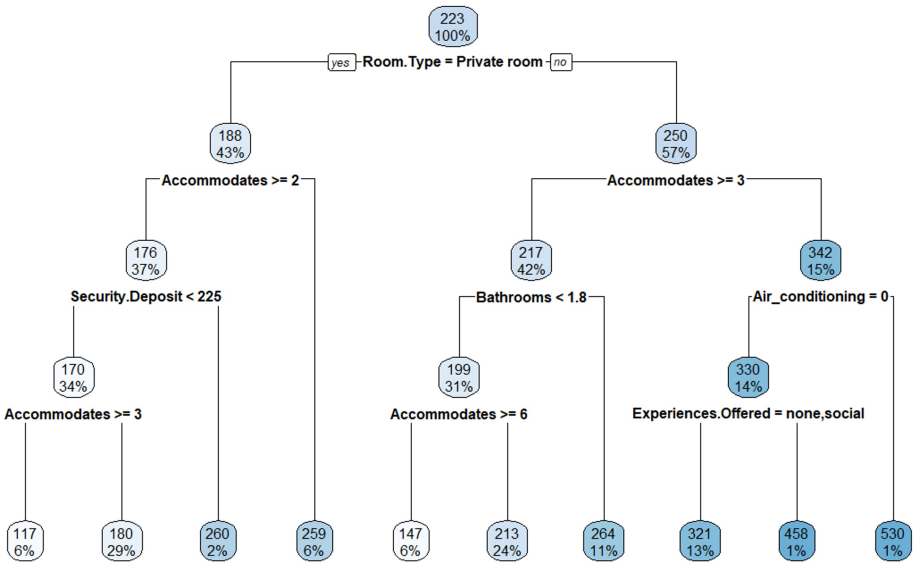


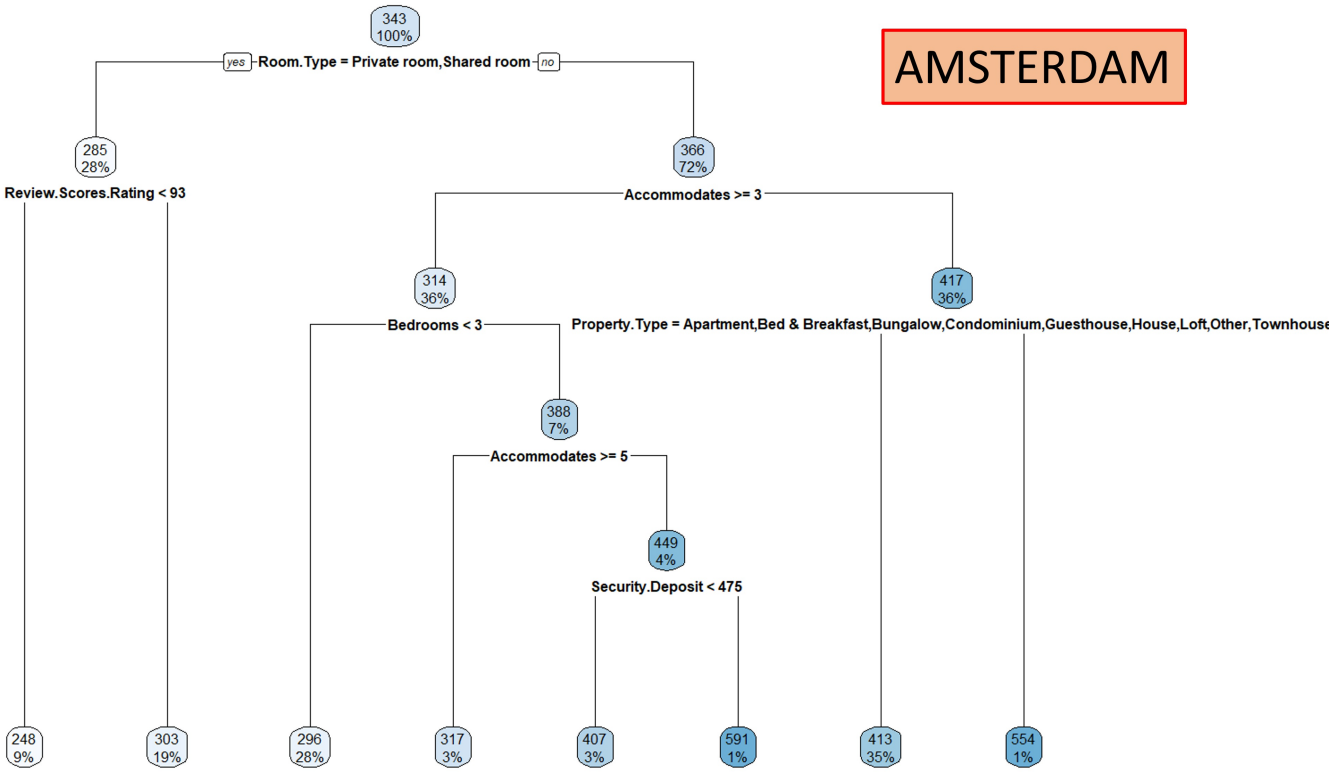


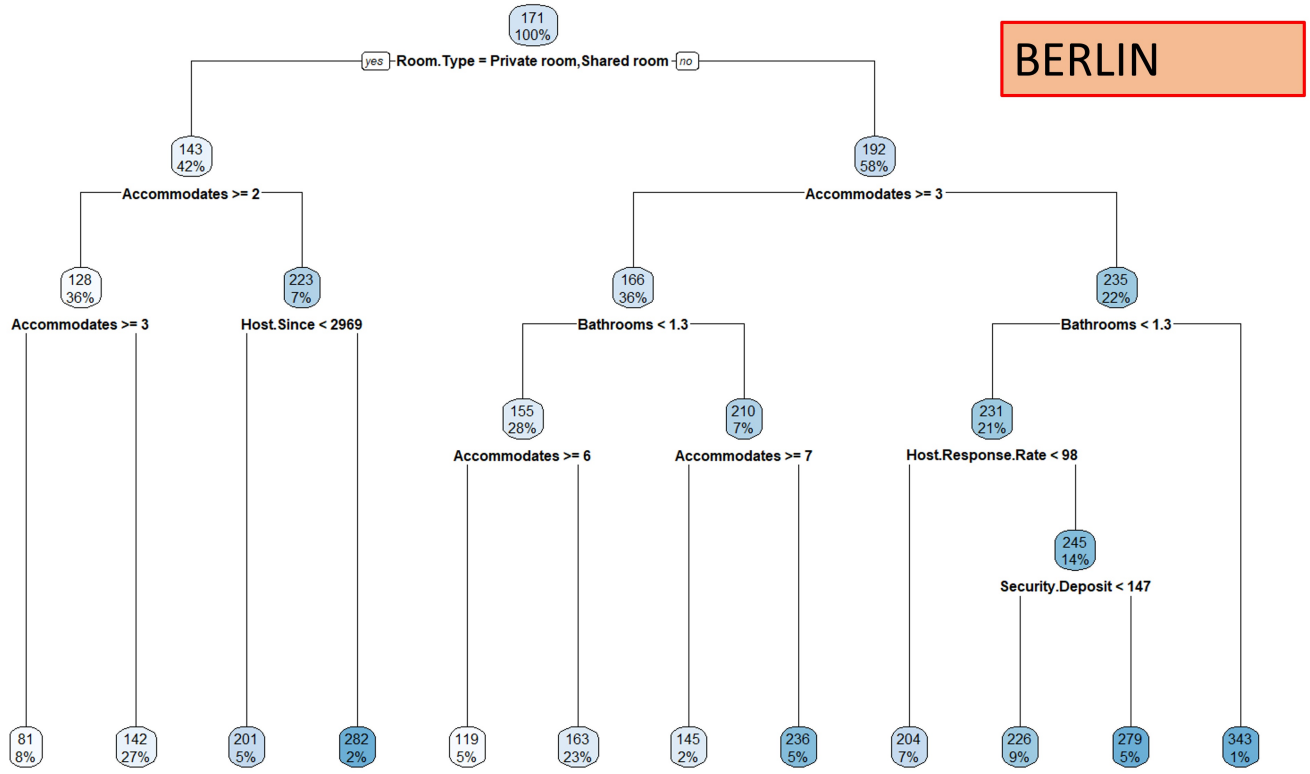

```
## [1] "Root Mean Squared Error:  80.8827730952067"
```

We now proceed by trying to estimate different trees for the different cities to see if there is any difference in the nodes:

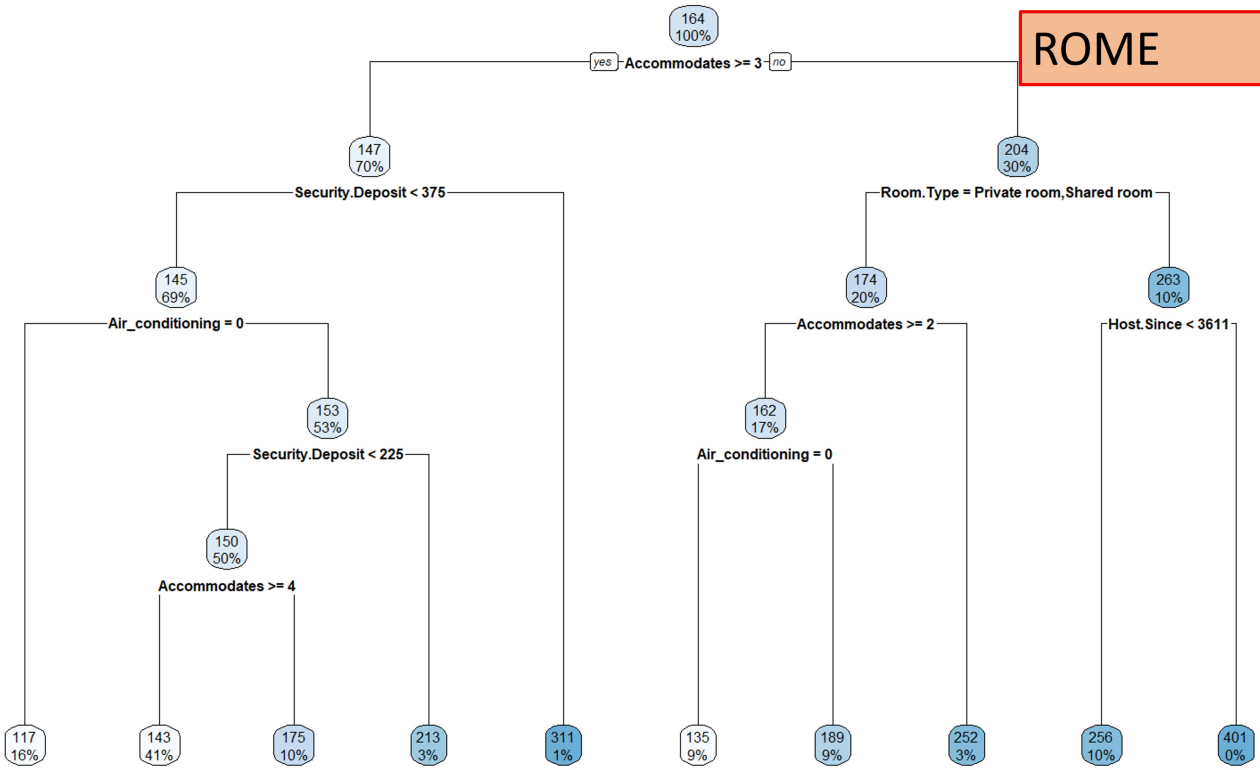
LONDON



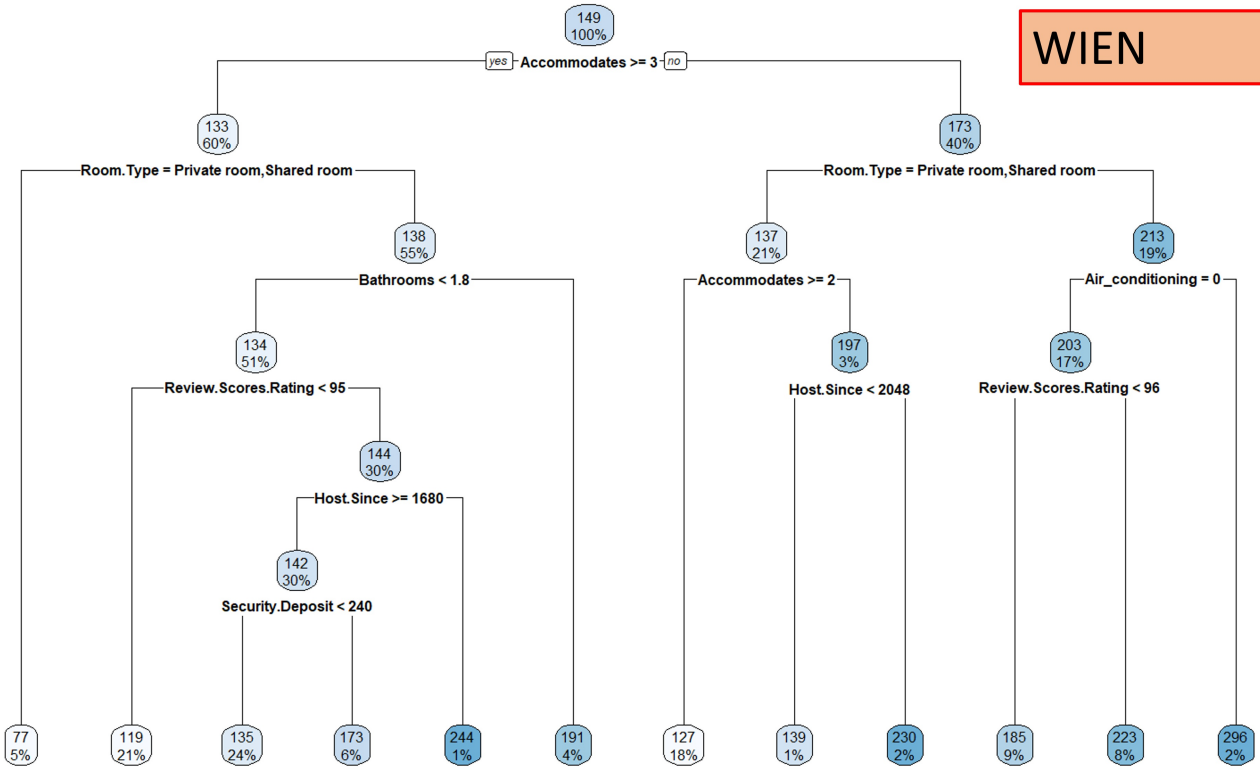

AMSTERDAM
343
100%
yes
Room.Type = Private room,Shared room
no
285
28%
366
72%
Review.Scores.Rating < 93
Accommodates >= 3
314
36%
417
36%
Bedrooms < 3
Property.Type = Apartment,Bed & Breakfast,Bungalow,Condominium,Guesthouse,House,Loft,Other,Townhouse
388
7%
Accommodates >= 5
449
4%
Security.Deposit < 475
248
9%
303
19%
296
28%
317
3%
407
3%
591
1%
413
35%
554
1%

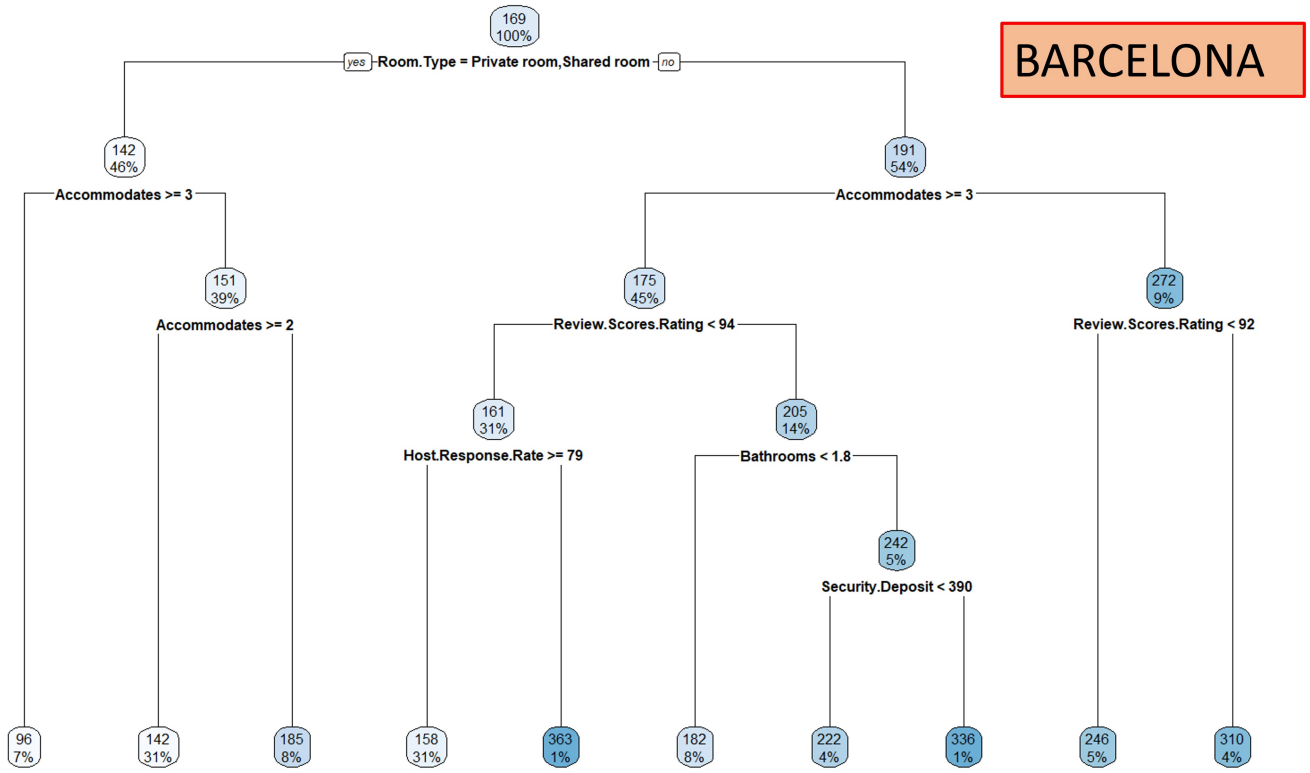

BERLIN
171
100%
yes
Room.Type = Private room,Shared room
no
143
42%
192
58%
Accommodates >= 2
Accommodates >= 3
128
36%
223
7%
166
36%
235
22%
Accommodates >= 3
Host.Since < 2969
Bathrooms < 1.3
Bathrooms < 1.3
155
28%
210
7%
231
21%
Accommodates >= 6
Accommodates >= 7
Host.Response.Rate < 98
245
14%
Security.Deposit < 147
81
8%
142
27%
201
5%
282
2%
119
5%
163
23%
145
2%
236
5%
204
7%
226
9%
279
5%
343
1%

ROME


164
100%
yes
Accommodates >= 3
no
147
70%
204
30%
Security.Deposit < 375
Room.Type = Private room,Shared room
145
69%
174
20%
263
10%
Air_conditioning = 0
Accommodates >= 2
Host.Since < 3611
153
53%
162
17%
Security.Deposit < 225
Air_conditioning = 0
150
50%
Accommodates >= 4
117
16%
143
41%
175
10%
213
3%
311
1%
135
9%
189
9%
252
3%
256
10%
401
0%

WIEN


149
100%
yes
Accommodates >= 3
no
133
60%
173
40%
Room.Type = Private room,Shared room
Room.Type = Private room,Shared room
138
55%
137
21%
213
19%
Bathrooms < 1.8
Accommodates >= 2
Air_conditioning = 0
134
51%
197
3%
203
17%
Review.Scores.Rating < 95
Host.Since < 2048
Review.Scores.Rating < 96
144
30%
Host.Since >= 1680
142
30%
Security.Deposit < 240
77
5%
119
21%
135
24%
173
6%
244
1%
191
4%
127
18%
139
1%
230
2%
185
9%
223
8%
296
2%



From the trees above we can see that the six different datasets construct trees that differs. The nodes that are always present in all six trees are obviously room type and accomodates, which, as seen also above, seems to be the main drivers for the price. The other variables that appears as expected are mainly:

- **Host since**: a more experienced host is always correlated with an higher price (as discussed above).
- **Air conditioning**: the presence of air conditioning increment the price (almost 50€!).
- **Security deposit**: a high security deposit is correlated to an high price, this may be due to the fact that properties that requires a security deposit are more valuable.
- **Rating**: as expected an higher rating is correlated to an higher price.
- **# Bathrooms**: having more than 1.5 bathrooms (at least 2) is correlated to an higher price.

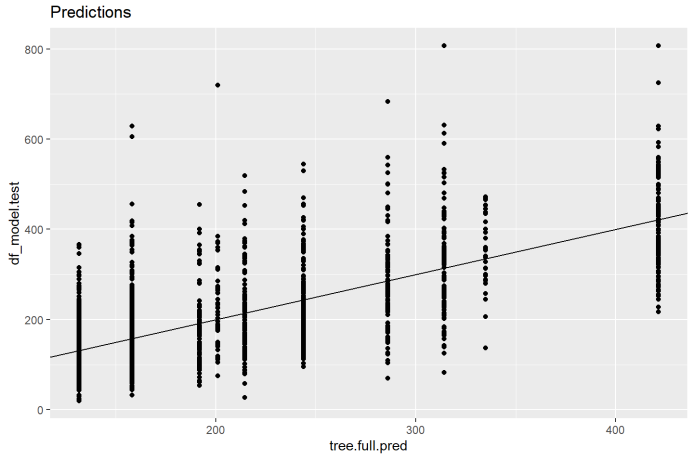
6.2 Advanced models

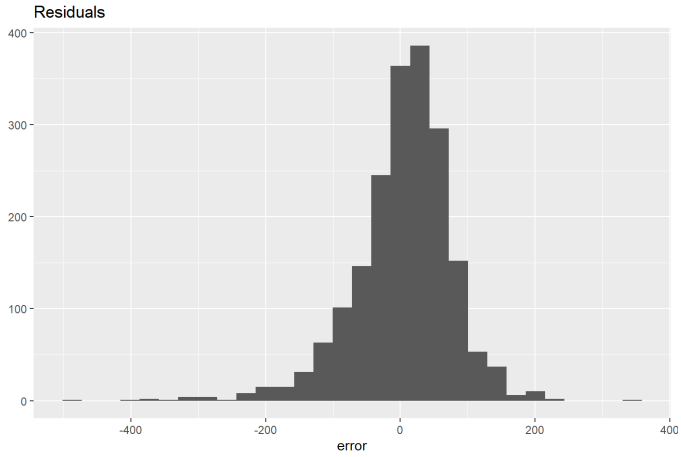
At last we try to construct a more robust model in which we try to achieve more accurate results at the expense of some interpretability.

6.2.1 Random Forest

We start by evaluating a Random Forest over our full dataset.

```
##
## Call:
##  randomForest(formula = Price ~ ., data = df_model, subset = train.full)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 8
##
##           Mean of squared residuals: 6300.264
##                     % Var explained: 56.65
```



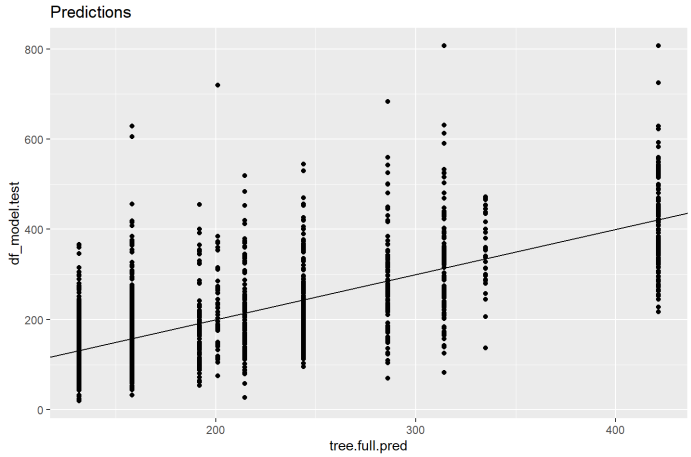


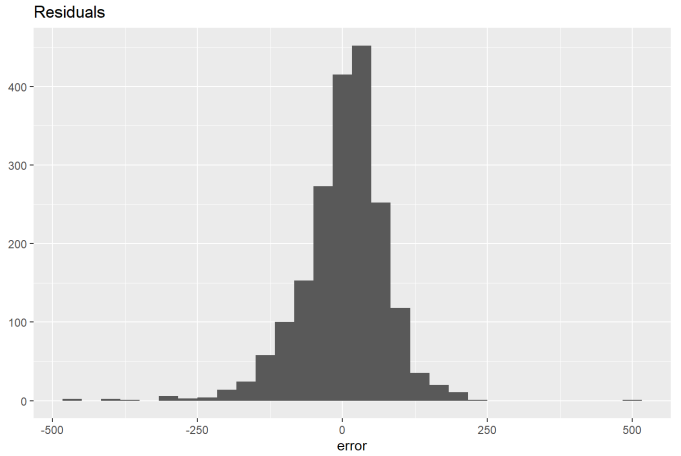

```
## [1] "Root Mean Squared Error:  73.1557067944934"
```

We can see that this kind of model explain just 56% of the variability of our dataset, and it returns a Root Mean Squared Error of 73.15 (a major improvement from the 80 of the single tree model).

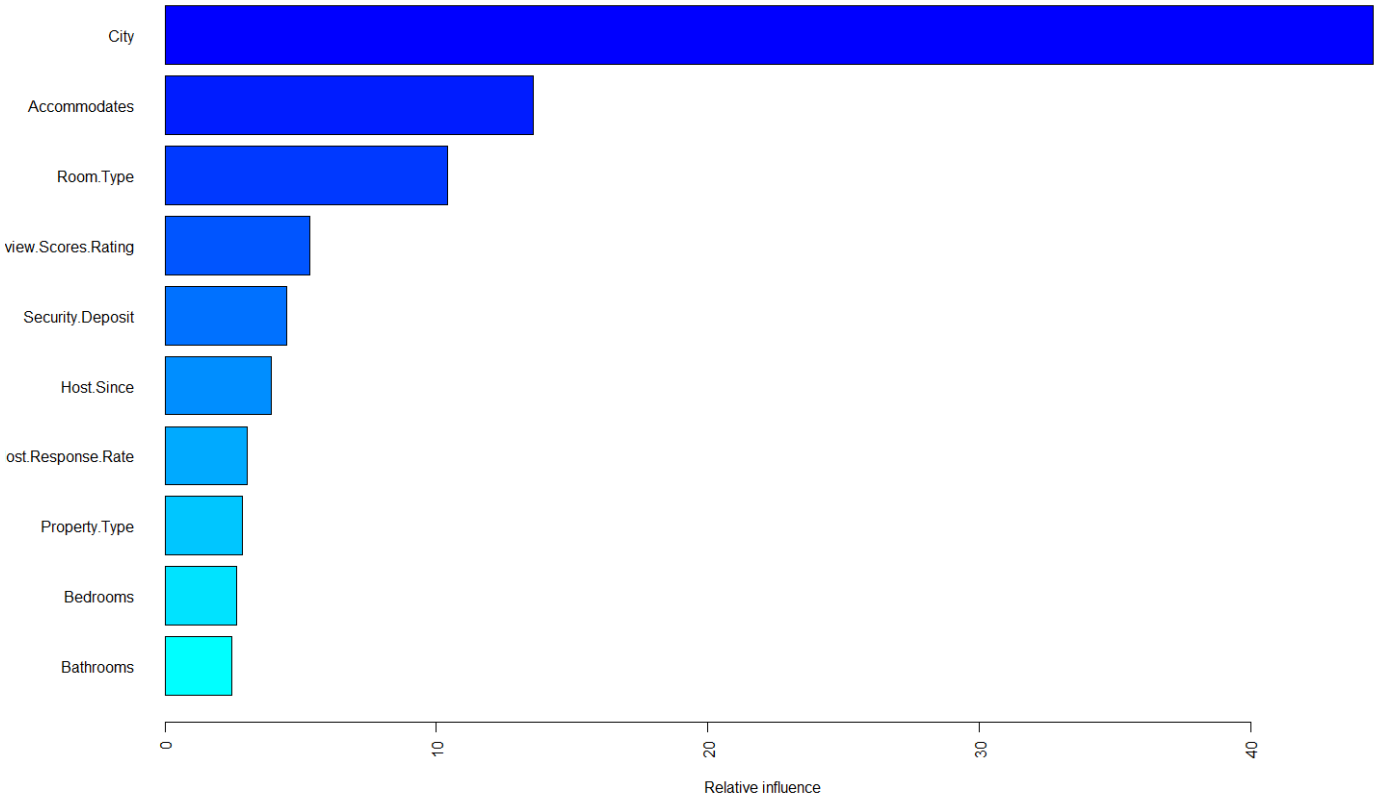
6.2.1 Boosting

We proceed by applying boosting.






```
## [1] "Root Mean Squared Error:  73.8598968588597"
```



We can see from the **Relative influence** plot shown above that the most relevant drivers for prices are the same we found in the single tree model. In this case, as in the RF model, the Root Mean Squared Error lowers, at 73.85.

6.2 Conclusions

We can conclude that, while this kind of models gave us a huge help in interpret our dataset, they still seem to slightly underperform. The reason for this could be the fact that our dataset is really heterogeneous, beign composed by cities that differs so much even in the mean price level. Another reason could be the fact that we still do not have all the variables we need to correctly construct a prediction model.

Some *Next steps* in order to improve the performance of our model and perform a more complete analysis could be:

- Add more relevant variables to the dataset, such as **Distance from city center**.
- Implement a **Convolutional Neural Network** that can evaluate the photos of the property and give a rate.
- Perform a more accurate analysis that could implement also a time span, as it is possible that **summer/winter season** directly affect the prices shown in the platform (sadly I did not find open data regarding this).

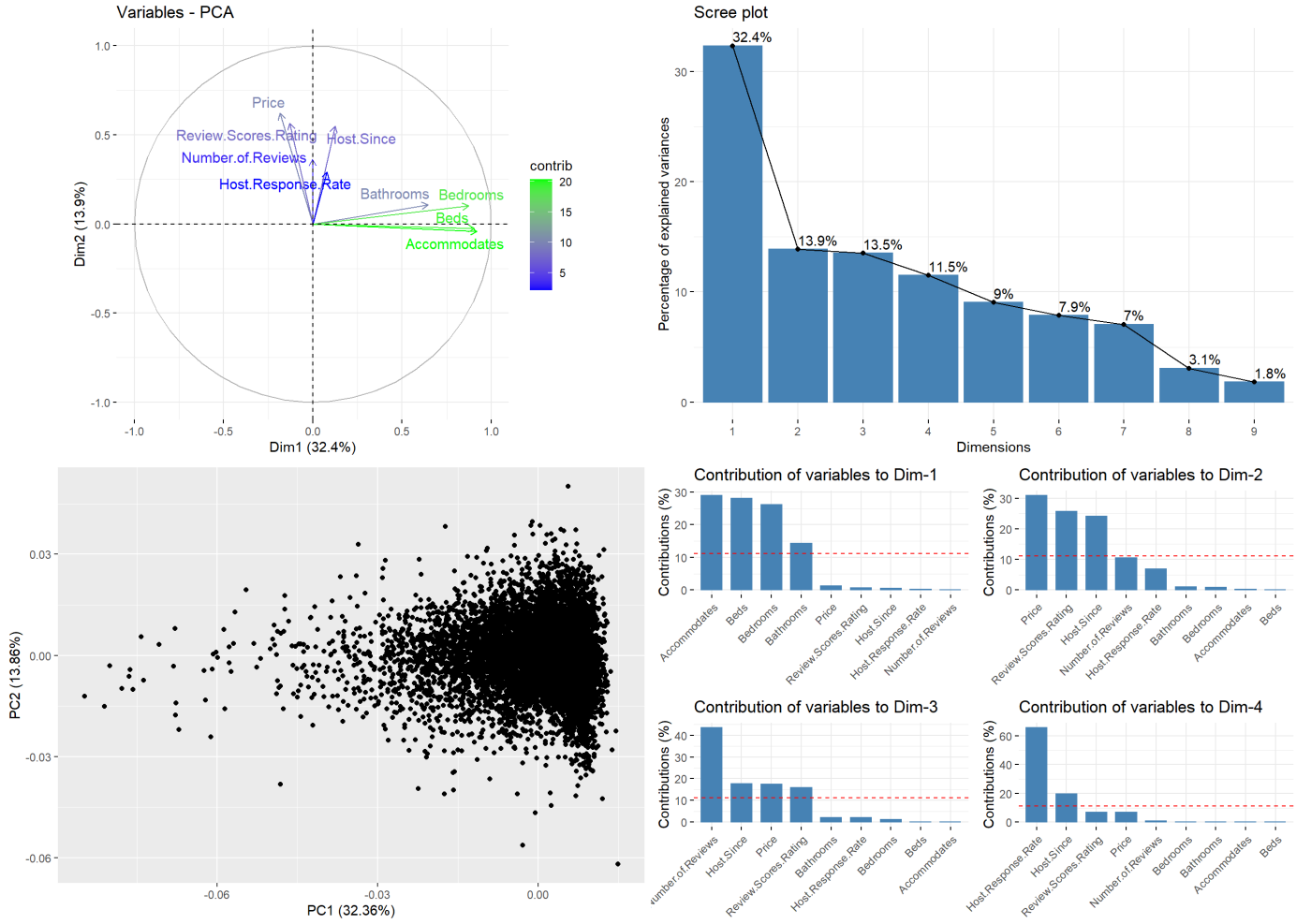
7. Unsupervised Learning

Now we will try and approach the problem with unsupervised learning models. This new approach could help us better understand our findings of the analysis above but also show us a different point of view when looking at the dataset.

7.1 Principal Component Analysis

We will start from the most famous unsupervised learning model, that can help us identify the principal components in which our data are spread out.

Before computing the model we recall that PCA can only be applied to quantitative variables; this means that even if we expect the results of this model be aligned with our previous findings, we cannot fully rely on this model in order to completely analyse our dataset.



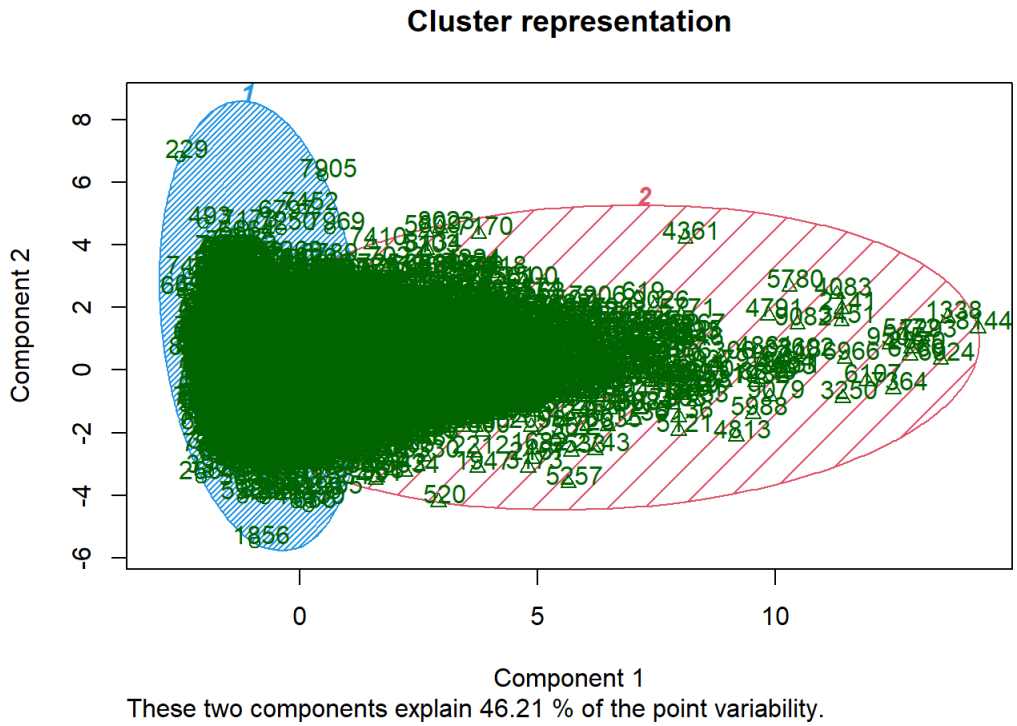
From the plots above we can see that there is clearly one dimension in which the data are spread out the most, and it is the one related to the actual property (accomodates, beds, bathrooms and bedrooms).

We then have a second and third dimension, both capturing less than a half of the variability of the first dimension and they are respectively a *value-of-property* dimension, given the fact that the main components are Price, Rating and Host.Since (recall that we interpred experienced host as an added value) and what it seems to be a *popularity* dimension, given the fact that its main component is given by the number of reviews.

This model is aligned with our first analysis, and it assess that the variability of the dataset can be reconducted by group of variables that we already presented as correlated.

Unfortunately we cannot see any clear cluster in the graph reporting the projected dimensions, this could be easily caused by the fact that all qualitative variables were left out from this model.

We shown below that the kmeans algorithm confirm our expectations: we are not able to identify clear clusters just by looking at qunatitative variables.



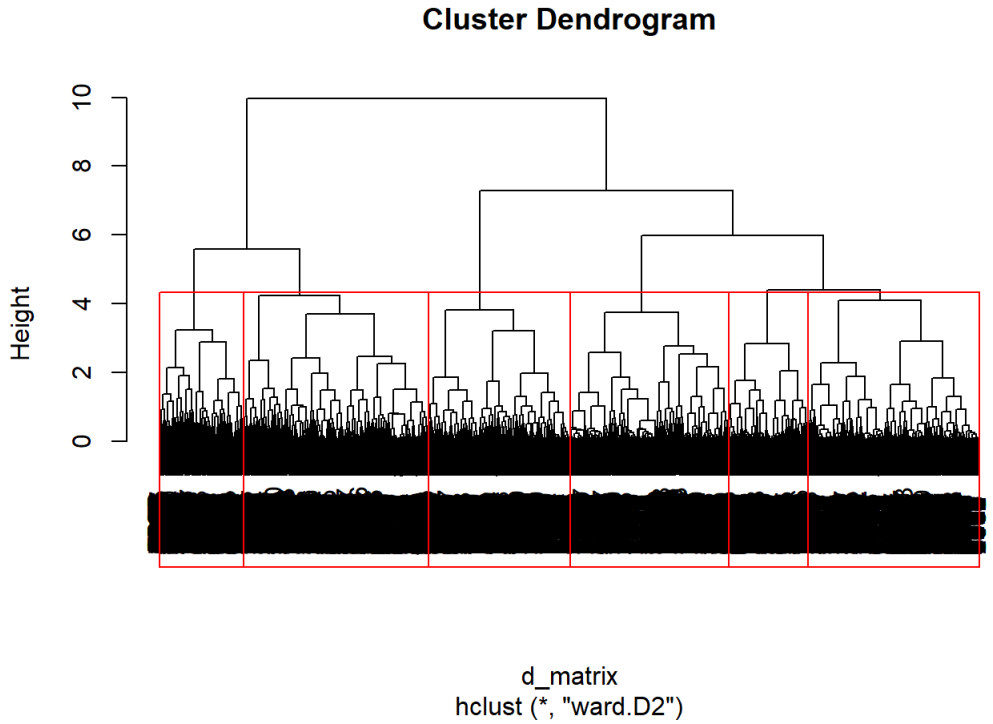
As expected the two clusters we get (that seems to try and separate *small* properties, with low number of beds and accomodetes from *large* ones) are completely overlapped.

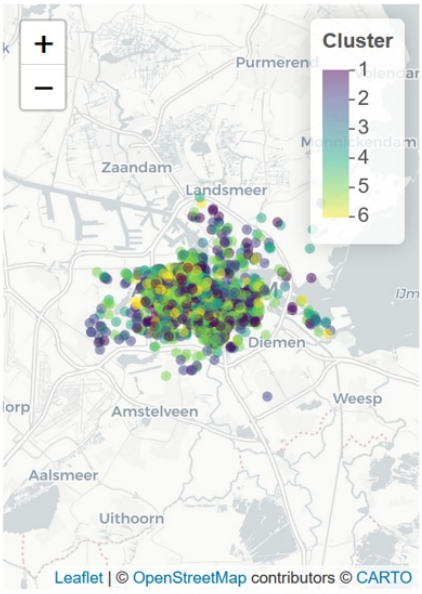
7.2 Hierarchical Clustering

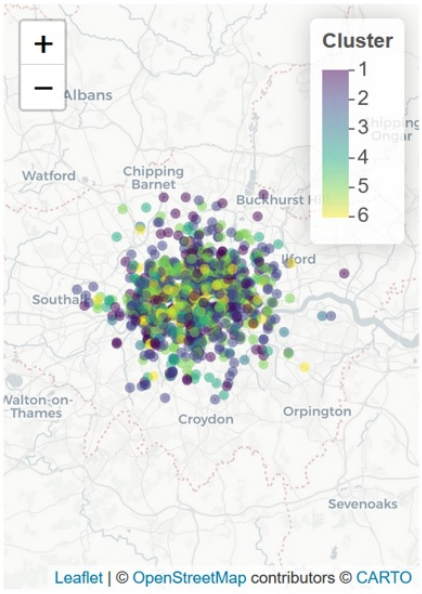
Given what we said aboce, we will now try to compute different cluster using hierarchical clustering; in this way we would be able to add also qualitative variables to our analysis, hopefully improving the model.

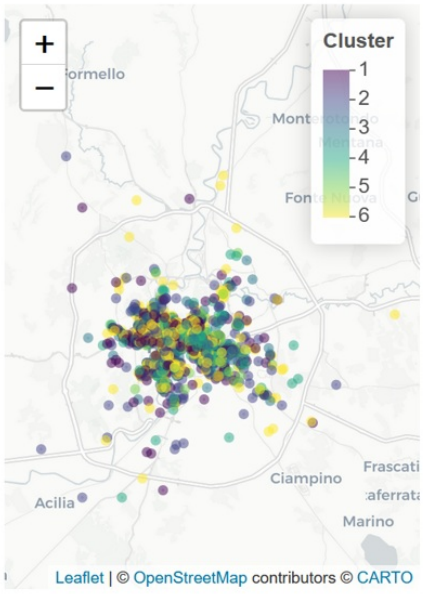
In order to evaluate the model with both quantitative and qualitative variables we use the Gower distance.

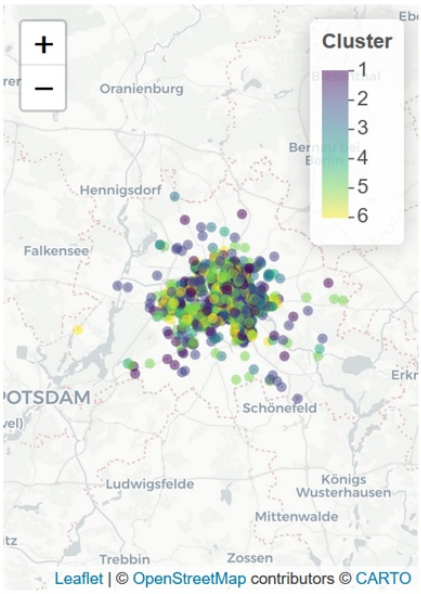
We start by trying to compute 6 clusters (that we hope to reconduct to our 6 cities). We will remove in this analysis the variable **City** in order to actually understand if there our cities actually differs.

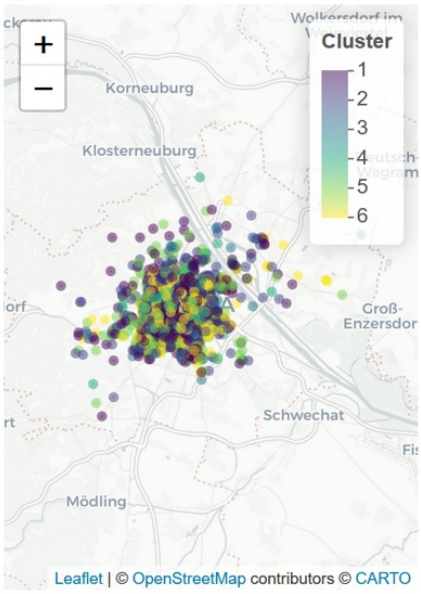


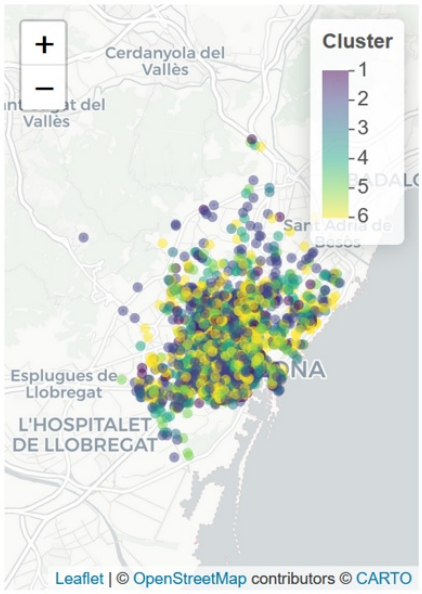






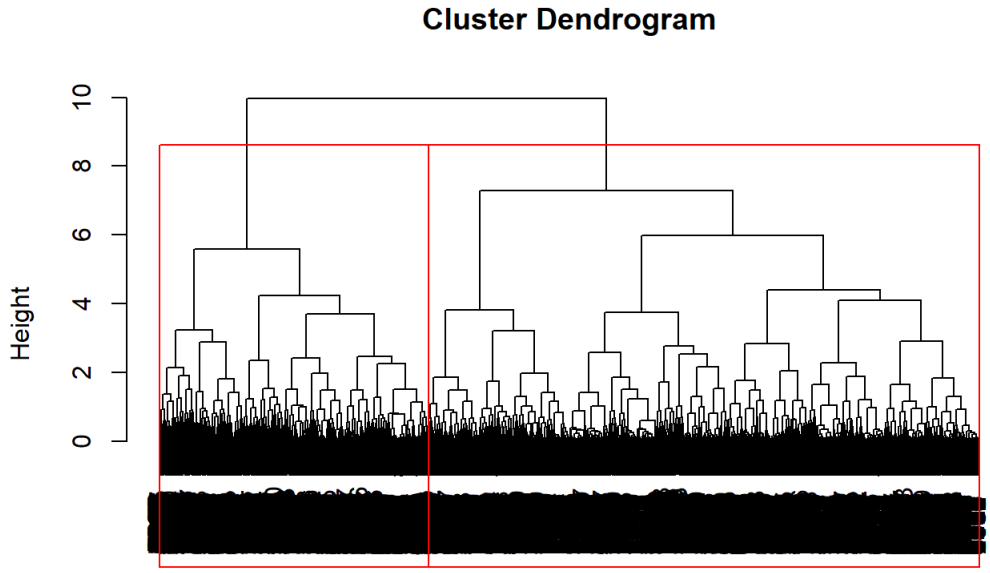


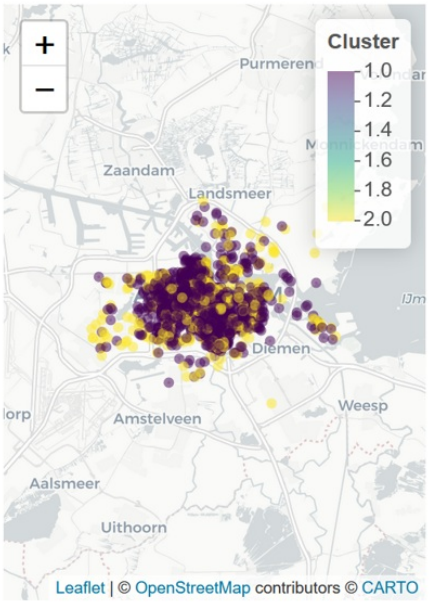


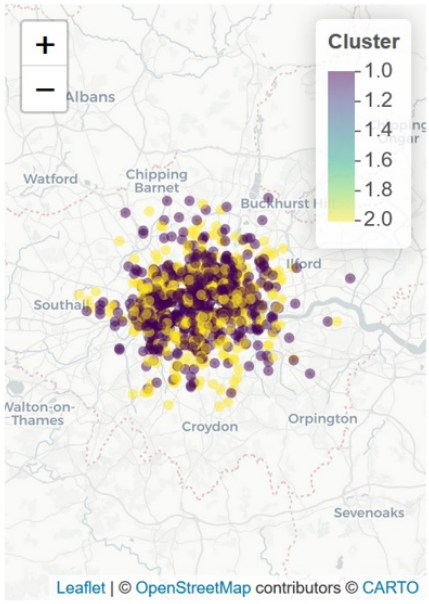


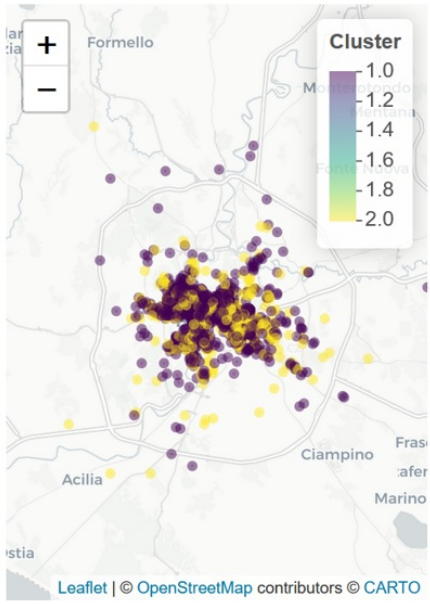
From the graphs above we can see that there is not actual difference and the 6 clusters seems to overlap in each of the cities.

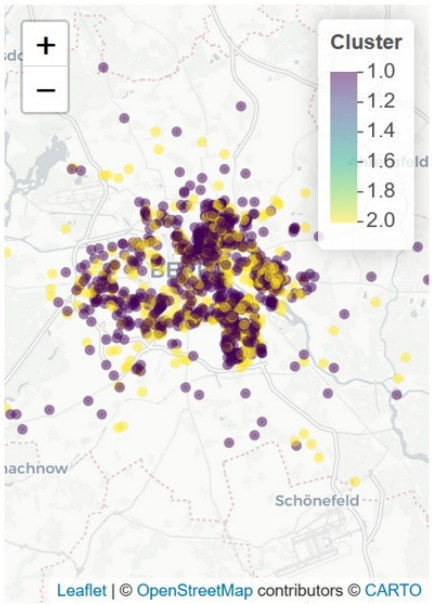
We will now try the same method but with a small number of cluster, in order to understand if there are some differences between properties that are not related to the city but maybe could be related by different locations in the cities

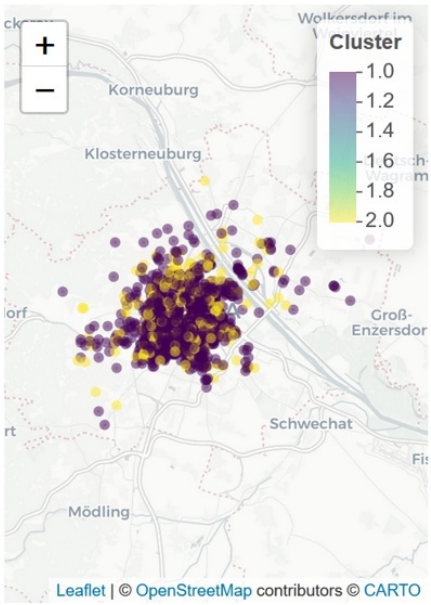


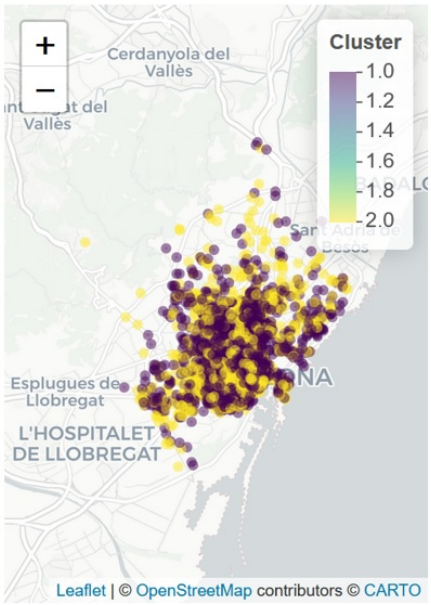






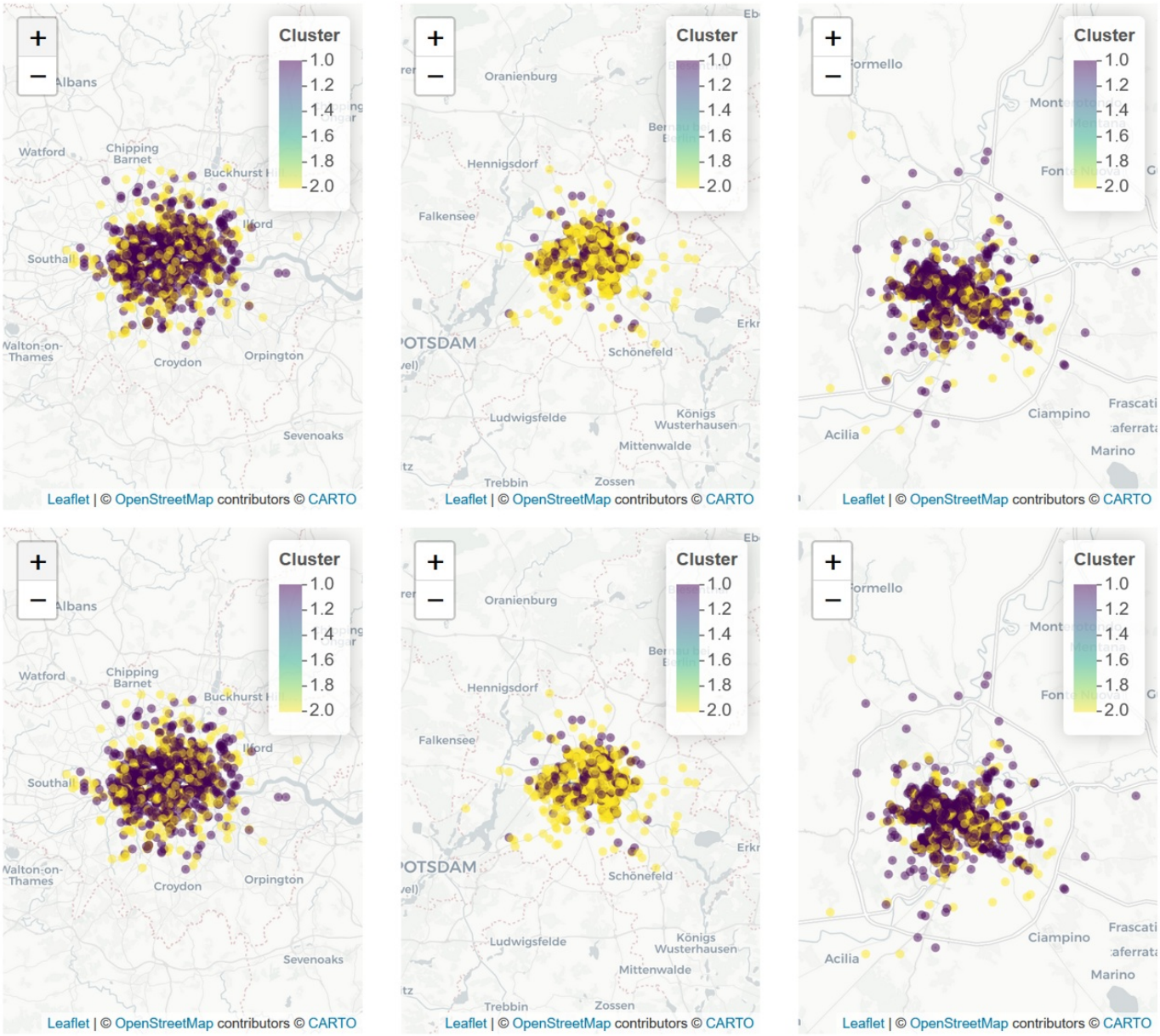






As can be see from the graph above we did not get the results that we expected, the clusters seems to overlap in all the cities without any geographical meaning.

We will now try and define different clustering models, one for each city, in order to understand if our data can be grouped in a way that has some geographical interpretation.



We can notice that in some cities (like London and Berlin) there seems to be a slightly geographical interpretation of the clusters, but they still seem to overlap in most part of the city.

7.3 Conclusions

This unsupervised approach definetly helped us for the first part. in identifying the principal components in which our data are spread out.

Unfortunately we did not obtain the result that we expected in the clustering analysis; this could be due to the fact that we have a lot of variables (like host-related) that are not correlated to the construction of geographical clusters.

Some *Next steps* for this kind of analysis would be:

- Same as before, adding some variables that could help us better cluster our data.
- Try and work with non-geographical clusters in order to assess if there is any other way in which uor data can be grouped.
- Try and work with lesser variables in order to better interpret and identify geographical clusters