Nonparametric Statistics Project Report

AIRBNBs in Milan

How to become a smart user: host's and guest's perspectives

Arcardini Alice
Bianchi Chiara
Del Mul Edoardo
Dominoni Lorenzo

POLITECNICO
MILANO 1863

Table of Contents

1. Introduction

How we approach the problem: from the point of view of both Airbnb's hosts and users and what our goals are.
We found the Airbnb dataset on Kaggle. It contained information about more than 9000 Airbnbs in Milan including prices, facilities, host features, and rating scores. The data regarded only whole apartments, while on the real Airbnb website you can rent also single rooms or share a part of someone's house.
We were really interested into this dataset because we saw many possibilities of real world applications. This was the reason why we structured our analysis this way. We put ourselves into the shoes of the users of Airbnb: those who let their apartments for rent and those who rent them actually; and we considered our project as a "consulting" one.

In fact, from the host's point of view we searched for strategies to choose the best price for the apartments - whether the host was new to the business or he'd been around for some time. Moreover we wanted to discover if there were particular features or conditions under which he could raise or lower the price.

From the guest's point of view instead, we looked for optimal ways to find good offers and cheap location, avoiding to get cheated by a bad priced apartment or a too expensive one.

We used permutational tests to discover influent features in prices and ratings, so that we could draw conclusions for the host. For the same reason the regression model we built aimed at choosing a suitable starting price for the host. Finally, conformal prediction intervals proved useful for:

- fixing a price range in which the host could set the price;
- detemining whether the apartments represented good or bad offers for the guest.

The analysis we did was profitable because we found optimal strategies for both users, looking and analyzing data from two different perspectives.

Finally, our previous experience taught us real world data would hardly satisfy all the ideal assumptions they usually teach in Statistics 101. For this reason, the non-parametric approach and techniques studied in the course were really of great use and will be in the future beyond this project.

2. Data Exploration

Understanding variables and outliers.
The dataset we have provides the following information, divided in various categories:

- host info: if they are in Italy, their response rate and time, if they have been certified superhost (experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests), if they have a profile pic or their identity has been verified, the number of apartments they put on Airbnb;
- location info: municipalities of Milan, the zip code and their coordinates, longitude and latitude
- service info: guest included, minimum nights, cancellation policy, instant bookable, host greets you, paid parking on premises, luggage dropoff allowed, long terms stays allowed, doorman, pets allowed, smoking allowed, suitable for events, x24 check in;
- apartments info: room type, accommodates, bathrooms, bedrooms, beds, bed_type;
- apartments facilities: TV, wifi, air conditioning, wheelchair accessible, elevator, kitchen, breakfast, heating, washer, iron;
- reviews: number of reviews, total scores and partial scores for accuracy, cleanliness, checkin, communication, location and value;
- pricing: daily price, cleaning fee and security deposit, extra for guest;
- availability: if they were available in the next 30, 60, 90 and 365 days;
- other: require guest profile picture, require guest phone verification.

We noticed since the beginning we had to deal with some outliers, and so we removed some of them. There were some unrealistic data, such as an incredibly high number of bathrooms and data that were not the focus of our analysis. Even if not all the observations we have left out can be considered as outliers, these data lie outside our purpose of seeking for the best methods to follow to choose to rent an apartment. Moreover we removed those apartments where the host response rate was really low, since they were not trustworthy.

3. Data Visualization

Preliminary analysis through visualization, using non parametric tools such as bagplots.
To better understand the outliers we removed, we first went through data visualization to have an idea of how our data are distributed and which are the fundamental variables that influence the price and the ratings.

From the plot of the density of Airbnbs and their price range we understand that the geographical location can be an influential feature, since the highest prices seem to be concentrated in the center of the city as well as most of our Airbnbs.

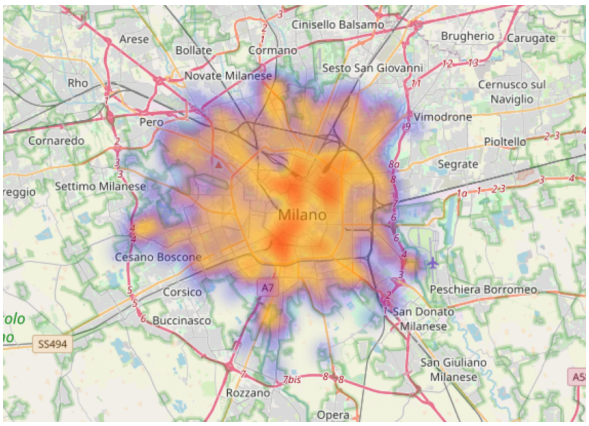


Fig. 1. Density of Airbnbs in Milan

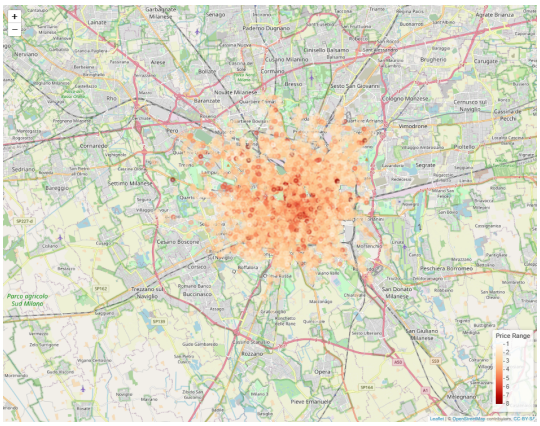


Fig. 2. Airbnbs' price range

To get a better understanding of the distributions of the outliers we went through a series of bagplots which are reported in the following.
We analyzed the data as a whole then we compared them when we have removed outliers.

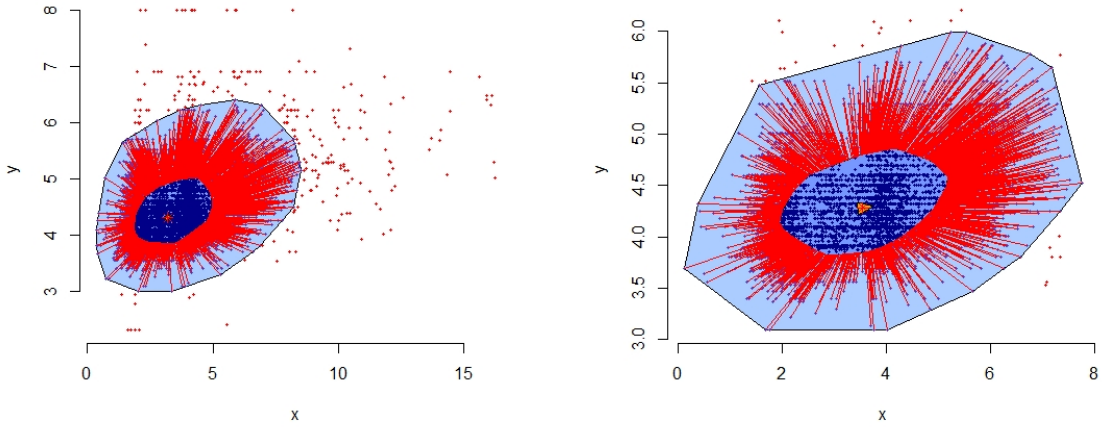


Fig. 3. Accommodates vs Log-price with (*left*) and without outliers (*right*)

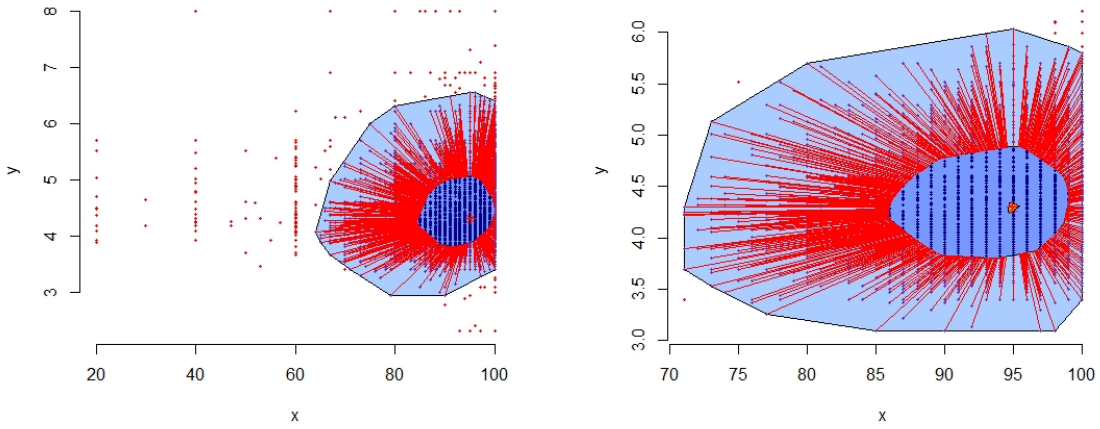


Fig. 4. Ratings vs Log-price with (*left*) and without outliers (*right*)

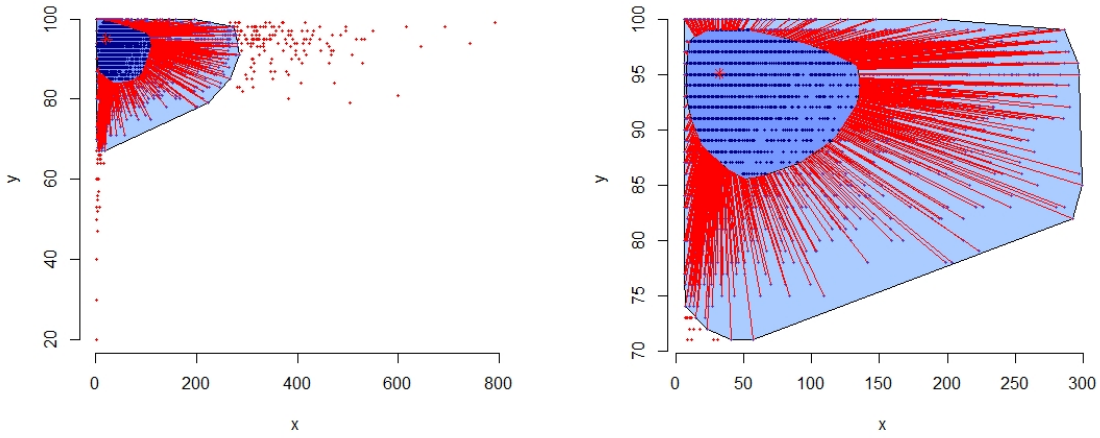


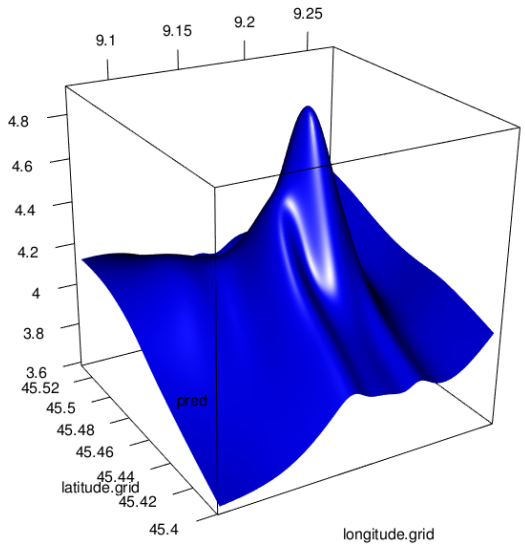
Fig. 5. Nr. of reviews vs Ratings with (*left*) and without outliers (*right*)

As we can immediately see most of the data falls inside the bag, but it must be noticed that the quantity outside is not to be neglected, since it appears to be heavy. As it will be shown later, data has a quiet fat right tail and this implies the previous behaviours of the bagplots. Most of this effect will be accounted through the Luxury variable, which clearly divides the expensive apartments from all the others.
Removing the outliers we get similar results, but as it is imaginable with softer tails.

4. Most influential variables and new variables

4.1 Distance

We headed directly towards the geographical coordinates, since, as common sense suggests, it might have a key role. With a smooth spline interpolation with prices we got the shape of the distributions of the prices over the coordinates and we noticed that the spike in the middle corresponds to the city center.



This hint suggested to us that distance is indeed a key variable.
Since working with the coordinates might be difficult we decided to re-elaborate them into a single variable that accounts for the distance in kilometers from significant places. The procedure will be explained in the following.

4.2 Cleaning fee and square footage

After all this preliminary considerations, we observed that our data was missing some very important features for determining the rent price:

- the square footage of the house;
- the distance from the most important attraction of Milan (ie. the Duomo);

Moreover there were discrepancies in the data between host who charged the cleaning fee on the daily price and host who accounted for the cleaning fee separately.
To handle the first problem, we tried to infer the square footage by exploiting the means of the size of the different rooms of a house.

We also found that by law there are some lower bounds for the size of each room, indeed we set:

- Kitchen: ca. 14 m^2;
- Single Bedroom: ca. 9 m^2;
- Double Bedroom: ca. 14 m^2;
- Bathroom: ca. 5 m^2.

The data was provided with the number of bathrooms, bedrooms and kitchen in every house, and through the information about the number of accommodates we computed the total number of different rooms and then an approximation of the square footage.

On the other hand, for the second issue, we exploited the given geographical coordinates to compute the distance from each important attraction of the city of Milan. To give more importance to the spots near the city center we decided to use a weighted average of the distances from all the different places of interest.

A similar approach was used to compute the distances from the metro stations, which we considered an influential feature over the price.

Finally, to obtain homogeneous data, we needed to compute the cleaning fee for all the host which accounted for that in the daily price (for these observations in the dataset we have cleaning fee = 1).
To tackle the problem, we first observed that in AirBnb the price for rent is a daily price whereas the cleaning fee is one off at the end of the stay. Secondly, we inferred a preliminary version of the cleaning fee through the square footage of the apartment and the average cleaning by square meter fee for each neighborhood. Thirdly we corrected daily price subtracting the computed cleaning fee taking into account that the average stay in an AirBnb is around 3.6 days.

4.3 Key variable: luxury

Why it is an important variable.
As a matter of fact, after building all the previous variables and removing the outliers, we realized that something was still missing and this was how luxury an apartment is. This variable is central in a pricing framework since two apartments with same square footage and facilities may have totally different daily prices depending on the quality of the apartment. In this scenario, we decided to infer it directly from the newly computed daily prices and the number of accommodates.
The apartments were divided in four categories depending on the price for accommodate by night:

1. BASIC: less than 20 € per night;
2. STANDARD: between 20 € and 30 € per night;

3. BUSINESS: between 30 € and 40 € per night;
4. SUPERIOR: between 40 € and 80 € per night;
5. LUXURY: more than 80 € per night.

The reader may think that this approach is not correct by a statistical point of view as we are trying to infer the price from a quantity that has been derived from it. Actually this is true, but we have to consider that this model will help future hosts in pricing their apartments and by that time it will be their duty to decide for which segment of the market to go after.
To test the validity of this assumption, in the following the model is going to be tested on new data which does not provide a priori a price but just in which category the apartment falls.

5. Permutational tests

Understanding categorical variables.
Before building our regression models, to make inference on the significance of categorical variables with respect to the price and rating, we performed several permutational tests of different types, due to the gaussianity assumption being highly violated. We chose those with respect to others due to their higher power, while computational complexity was not an issue. In this way we selected variables for the subsequent GAM and we discovered important characteristics of our data that can be useful for hosts.

First we did two independent univariate populations tests to discover if some binary variables related to the host (host is superhost, host has verified account...) and facilities (TV, WiFi...) were influential on prices and ratings. We used the difference between means of populations in every case as test statistics. We tried also with the medians because of the presence of particular distributions with heavy tails (for robustness), but we achieved practically the same results. We considered significant only very low p-values because of the high number of tests.

Apart from finding significant variables to use in later approaches (we are not going in detail here since there are 60 of those), we want to show two interesting results:

1. Generally tests on rating achieved lower p-values than on price, especially with host variables: this is explained due to the fact that being a "premium" host gets direct influence on your score, while your price is more due for example to location or square footage. So we highly suggest becoming a superhost and greet your guests if you want to get a good rating score, but it doesn't really matter that much if you want to make guests pay more

Variable	PRICES pvalue	RATING pvalue
host_is_superhost	0.7488	0
host_identity_verified	0.313	0
host_response_rate	0.8846	0
host_total_listings_count	0	0
host_greets_you	0.0712	0

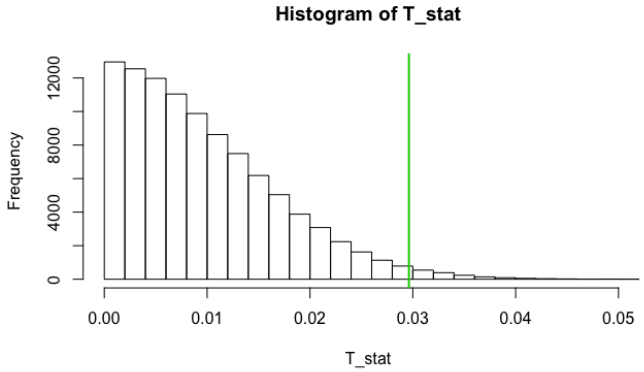


Fig. 6. Superhost vs Prices

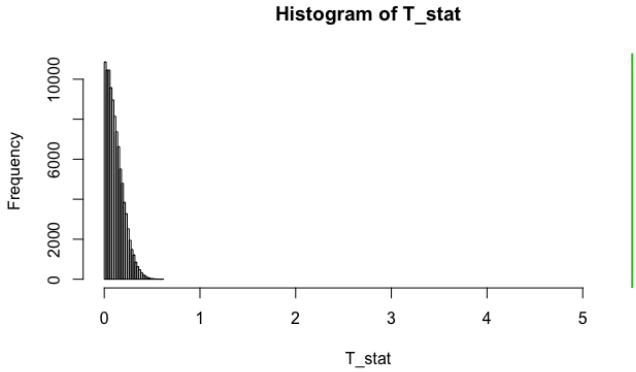


Fig. 7. Superhost vs Ratings

2. Tests agreed with our intuition on what is relevant for the majority of guests and what is not: for example for variable price this means that while having tv and air conditioning is very important, having an iron or allowing pets is important for a minority. Some exceptions to this rule were also found (kitchen not relevant, doorman relevant...): the list of all the p-values is in the R file. Here we report some interesting tests for price.

Variable	**pvalue**
TV	0
Air Conditioning	0
Iron	0.2002
Kitchen	0.472
Doorman	0
Pets	0.1691

Then we did ANOVA permutational tests using the F value statistics to discover if the location was really important. We used as categorical classes the different neighborhoods of Milan and as variable price. In this way we found that the p-value was 0, as expected, so the location is influential.
However the interesting fact is that repeating the same analysis without zone 1 (the center of Milan) p-value was not significant. We interpreted this result in this way: location is important only when considered as distance from the center and sightseeing, if you are at a specific distance from the Duomo, the exact part of the city doesn't really matter for prices.

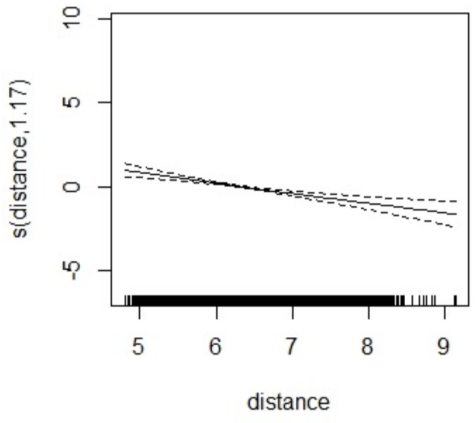
Variable	**pvalue**
neighbourhood	0
neighbourhood without zone 1	0.8583

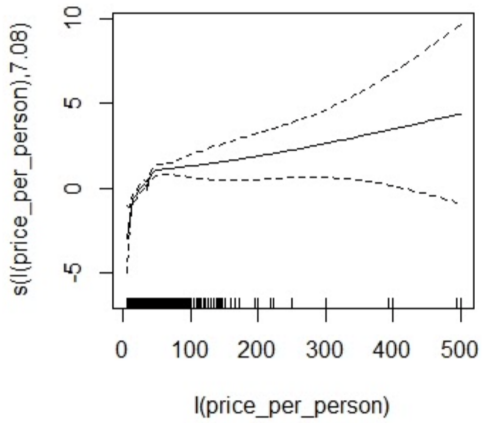
6. GAM on Ratings

Inference on ratings scores and useful strategies for hosts.
Let's start now with an analysis on the rating scores to pass then to our main goal about the prices. As we have anticipated before with the point of view of the host we want to understand how to improve the level of the offered service and so obtain higher rating scores.
In order to do that after having observed with the permutational tests which are the variables that seem to be more influent on the ratings we have also built an additive model with cubic splines for the continuous variables and linear terms for the categorical ones. We started with a model with a lot of different variables and we have then performed model selection using permutation tests to verify the significance of the terms since the residuals were not gaussian. The final model is not able to explain a lot of the variability of our data and is not very good for predictions, this is understandable since rating scores are a very subjective measure, but we can use it with an inferential point of view to extrapolate some good advice for Airbnb users.

From the results in fact we have seen that there is an high value for the intercept and then the covariates provide an increase or a decrease of the starting value. The variable that provides the highest increasing is "host_is_superhost" , then other features related to the host contribute with smaller values. As regards facilities, in general , as we can expect, having them seems to be appreciated by the clients; in particular those with higher coefficients are "Heating" and "Air condition". Moreover we have found negative coefficients for the two covariates "Pets allowed" and "Suitable for events".

As regards the continuous variables we can conclude that ratings are decreasing with the increasing of the distance from the city centre, but surely the host cannot do anything to improve this; and increasing with the price for each person, this makes sense since probably higher priced apartments are really better than cheaper ones.





7. GAM on Prices

A nonparametric regression model for the prices.
Let's focus now on the prices. After having created all the new support variables and isolated the relevant features from the permutational tests we built the regression model. We decided to use a General Additive Model because it represented the best choice since we have to deal with categorical variables such as the presence of Wifi, Elevator, etc. and continuous highly non linear variables such as the number of accommodates or the distance from the principal places of attraction of Milan.
At the beginning we implemented many models by trial and error. We started from essential regressors such as distance and accommodates, adding one at time the facilities which resulted significative from the permutational tests. Unfortunately the adjusted R2 squared of these models was really low.

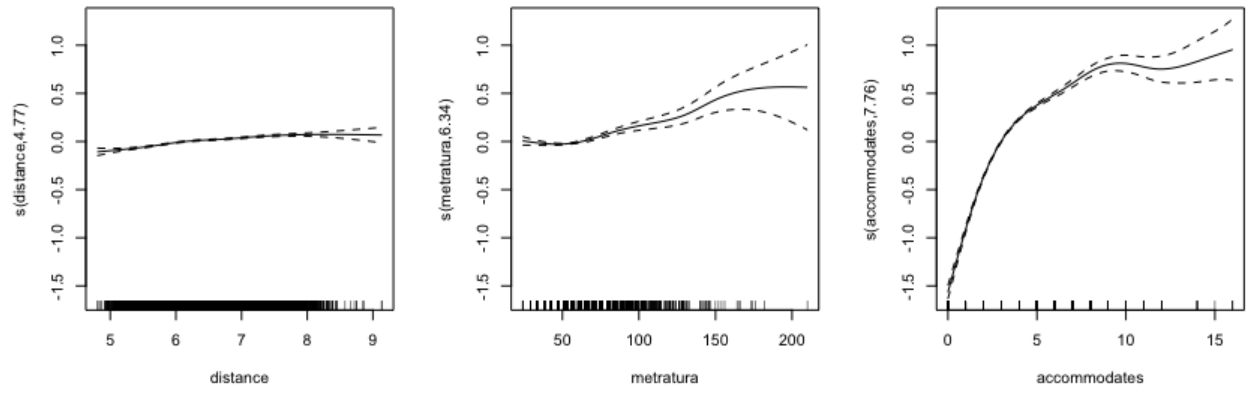
Our first aim was trying our best to improve our model. We succeed in this with the variables square footage and luxury. Using them as regressors made our R2 improving. This lead us to the choice of implementing two models for our data:

- a basic model with only essential variables
- a complete model which encloses all the fundamental regressors and the results of permutational tests on the facilities

The choice of regressors was made by permutational tests and it is discussed below.

7.1 Basic Model

It is thought for users who have little information about the apartment, or they have little time for a thorough analysis. This model presents the essential covariates: distance, square footage and number of guests. We wanted to see how these regressors were interpreted by the model.



In these graphs we can see the price is non-decreasing as all covariates increase, which makes sense.

7.2 Complete model

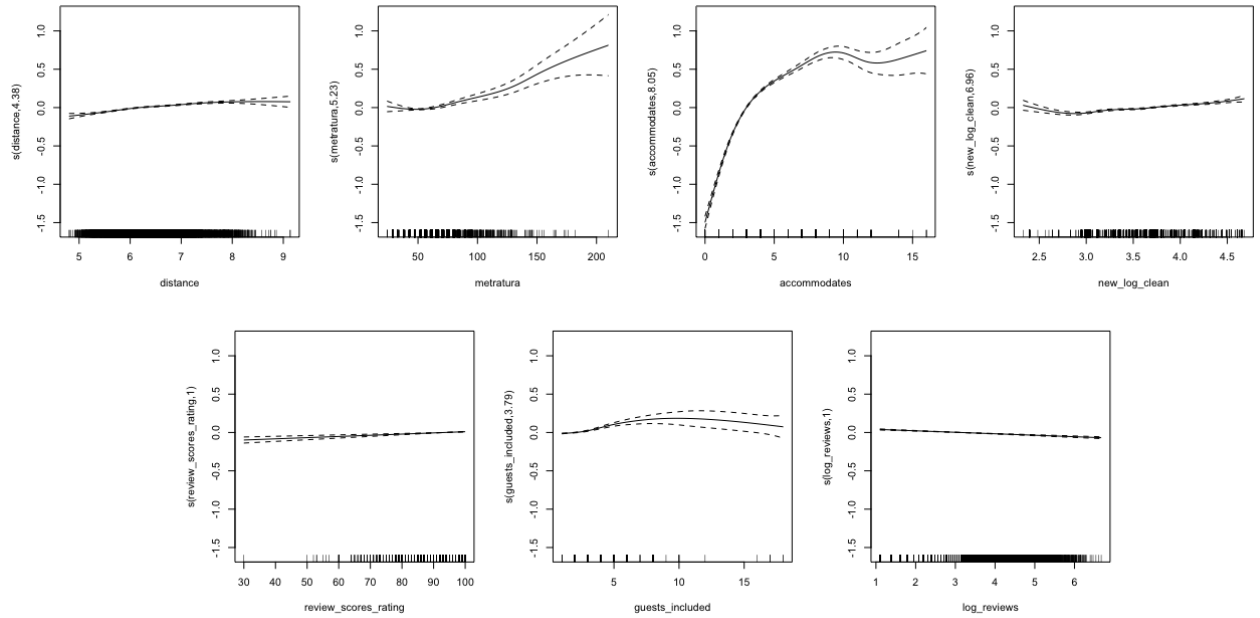
This model is much more complete than the one before. We can use this to predict new apartments' prices when we have all its specific characteristics. We added all the facilities which showed significance in the permutational tests run before and some numerical terms we haven't considered yet.
The facilities are:

- cancellation policy
- WiFi
- Air Conditioning
- Wheelchair_accessible
- Elevator
- Kitchen
- Luggage_dropoff_allowed
- Long_term_stays_allowed
- Doorman
- Smoking_allowed
- Suitable_for_events
- X24_hour_check_in

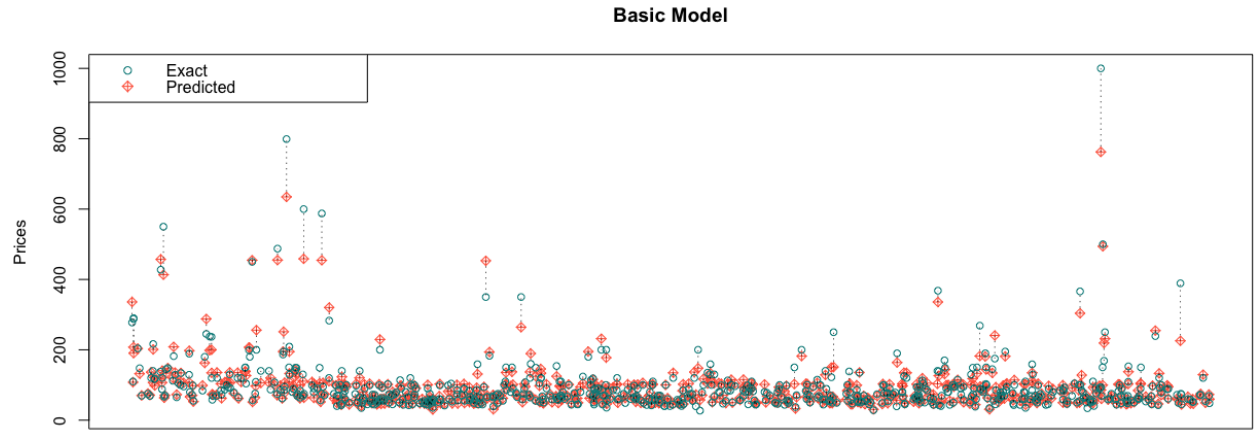
The numerical terms (implemented with smoothed splines) are:

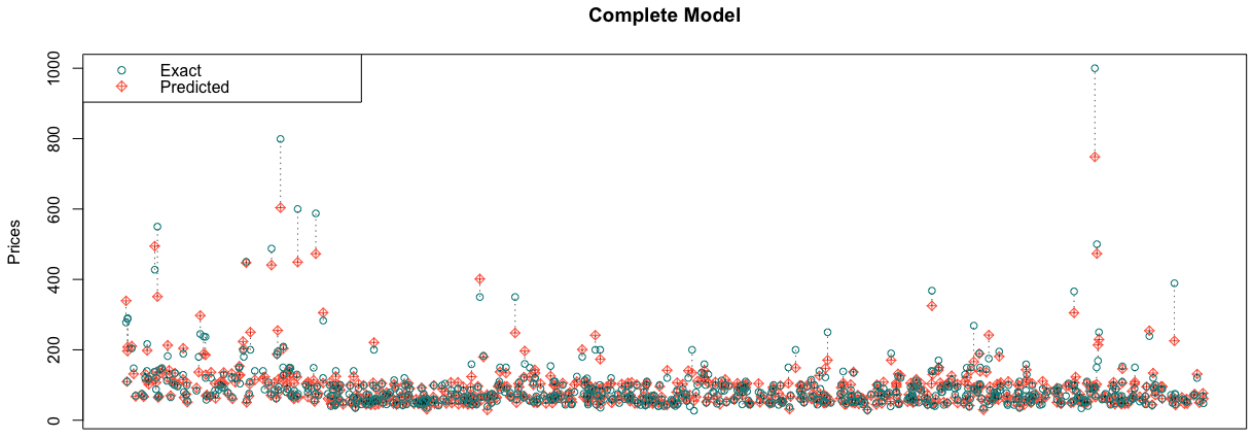
- log(new cleaning fee)
- reviews score rating
- guest included
- log(number of reviews)

Plotting the other spline terms we can see that all of them are coherent. As before many of them are increasing while the price is increasing: distance, square footage, accommodates, cleaning_fee and the review score rating.



We compared the validity and accuracy of the two models splitting our dataset into a train and a validation dataset and making point-wise prediction. We chose this technique over Cross Validation due to computational issues, moreover the dimension is large enough to allow us.

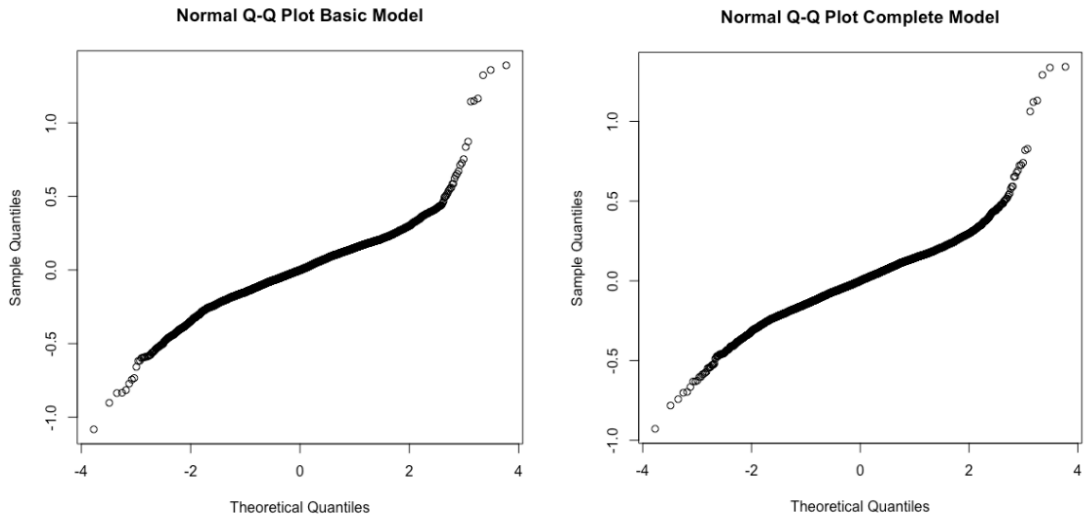




The models fitted really well. We noticed though that they tend to slightly underestimate the prices when the original price is high.
Our interpretation is that these models give a starting price that can be set higher.
From a host's point of view, the price can be increased depending on the luxury of the apartments. From a user's point of view the model gives the medium price corresponding to the feature he/she is looking for. Any price very lower than the predicted one can be a bargain.
Since the train/valid dataset were made randomly, we can produce many different pointwise prediction with respect to the ones reported above.

7.3 Permutational tests on GAM's coefficients

Understanding significative variables.
For every model we implemented we had to check on the gaussianity of the residuals. As we can see in the plots below, both the basic and complete model present highly non-gaussian residuals. The pvalue of the shapiro test for both model is low 1e-16.



This didn't allow us to rely on the summary of the models. We needed to perform a permutational test on the coefficients of the regressors in order to perform variable selection. We have implemented a piece of code suitable for different tests for both linear and smoothed terms.
First we tested the variables we thought were fundamental. As expected they are all significative. Then we proceeded to test the other numerical variables. We don't report all the pvalues since they are more than 25, but only some interesting results:

Variable	**pvalue**	**Variable**	**pvalue**
distance	0	dist_duomo	0.687
square distance	0	dist_metro	0.1
accommodates	0	availability_30	0.009
basic	0	security_deposit	0.74
standard	0	new_log_clean	0
business	0	log_reviews	0
superior	0	review_scores_rating	0
luxury	0		

We can see the significance of those variables we included in the complete model and some results that were not granted. In fact we can see that "distance from duomo" is not as significant as we thought, because its information is already stored in the variable sightseeing. We can explain similarly the non significance of the variable "metro", since many metro stops are close to places of interest (Porta Venezia, Duomo, etc) which are always embodied in the variable distance.

Regarding the *"availability"* regressor, even if they were quietly significant we decided to exclude them from the model because we don't think it depends on time and the R2 doesn't improve with them. Instead since the majority of the variables related to ratings score were significative we decided to add both the number of reviews (taken the logarithm) and the sum of the total scores, here *"review_scores_rating"*.
Even if some of the facilities were not so significant we decided anyway to keep them in our model to include the results given from the permutational tests at point 6.

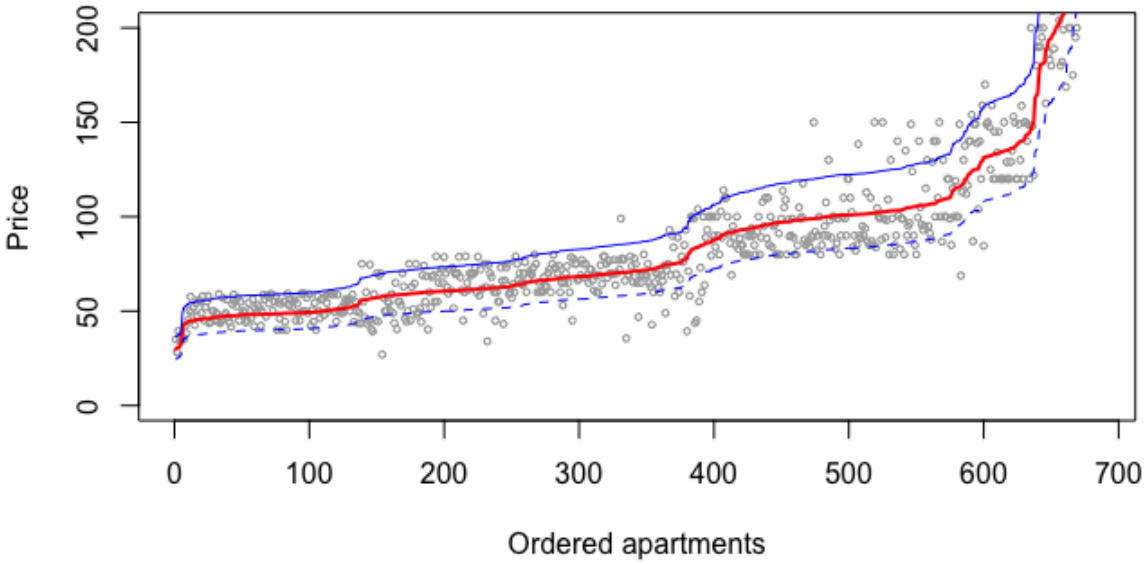
8. Price forecasting

Conformal intervals on prices.
Now we want to find prediction intervals for the price using conformal intervals. We have two motivations to this:

1. Find "normal" and "suspect" prices chosen by the Airbnb users given the features of their houses. In fact using prediction intervals of coverage 80%, we suppose that prices out of this range would be too high or too low, so that probably those would be either bad priced apartments or frauds. From the host perspective if your house is one of those we suggest to change price to a more desirable value to get either more guests or more income. From the guest one we suggest not to choose a house out of range because it is too expensive or too risky to get cheated.

2. Find good and bad offers in Airbnb. Knowing the intervals is easy to know how much over or under the mean value an apartment is, so that both guests and hosts can decide what to do accordingly. This choice would be more of a subjective one for the guest that has to decide whether it is risky or convenient to take an Airbnb, but prediction intervals can greatly improve his understanding in order to make a good choice of where to stay. For the host we suggest to start pricing from the pointwise prediction and then to adjust price around it in time, using progressively new infos of how much people are willing to stay there and their evaluation, but always remaining into those prediction intervals.

Practically from our division into training and validation set we could compute split conformal predictions on the training one and test performance on the validation set to see if the coverages were correct. We used split conformal for computational reasons.
We did it for the "basic" GAM obtaining very good coverages, shown in this plot. Houses are ordered by previsions in increasing order for better visualization and there are 80% prediction intervals (in red predictions, in grey real values, in blue intervals). This plot is the focus in the 0-200 euros a day price for better visualization.



Since the test with Airbnb apartments went well we concluded that our model was finally valid and ready to be used in real world applications.

9. What are our houses worth?

Pricing three different houses in Milan.
Unfortunately we don't own an Airbnb of our own to apply our results, but we were really interested to try the model on something we know well enough to grasp if it was concrete and well suited in general situations. So we used it for some of ours or our friends' houses to see how it performed in those situations. We have to take into account that a house for living is generally different from one on Airbnb (bigger, more facilities...), but making this consideration we tried with 3 very different houses to get the most variety possible.

Our model of choice was the basic one as we considered the case of a person who wants an approximate idea of the price without knowing everything. In fact we need only geographical coordinates, accommodates, square footage and the estimate of the luxury. Probably over 90% of the variability is already explained by this model with respect to the complete one, that can be used for refinement if we are really interested and ready to start a business on the site. However for one house we also tried that one to see the changes.

Here there are our houses of choice to show (prediction intervals in table below):

1. Piazza: Two-room apartment in Moscova for 2 people, used by 2 off-site students and so pretty essential.
2. Dominoni: 90 m^2 comfortable design apartment on Milan external ring road, family size for 4 people.
3. Camisa: Big and luxurious apartment near Montenapoleone, suited for 6 people.

	lower	**center**	**upper**
Piazza	43.72001	52.88767	63.97771
Dominoni	162.3885	196.4398	237.6313
Camisa	330.8879	400.2719	484.205

All the results are very accurate based on our knowledge and so we think this model can help in a great way people without experience in the Airbnb site. Moreover the intervals are important to have an idea of the possible maneuvering range.

With the complete model we predicted intervals for Dominoni's house, using all the variables and finding little difference: this prediction is more accurate, but it changes only slightly the previous one and is useful for refinements.

	lower	**center**	**upper**
Dominoni	156.6922	188.1357	225.8891

Here we want to make a point about the luxury variable: we created it ad hoc so it was "overestimating" performance on the training and validation sets, but since our goal is to generalize, we didn't really care about it, and it turned out to be very useful in this sense. By knowing a house, we have this variable, and it can radically change its price. Obtaining accurate results in new houses let us know that this variable was appropriately selected.

10. Conclusion

With our analysis we achieved all these final objectives that couldn't be obtained without non-parametric techniques:

1. Go in "depth" in understanding the Airbnb market in Milan. Using visualization tools, low dimensional splines, looking at existing variables and creating new ones that we thought could be useful, we found both reasonable and non-obvious results.

2. Make inference on the way all the variables can influence price and rating. We thought price and rating were the most interesting features about which we hoped to give advice to users. Leveraging various types of permutational tests we selected the meaningful and not meaningful things, using GAMs we predicted prices with accurate precision and constructing conformal prediction intervals we gave a reasonable range in which it is plausible to be able to adjust the correct pricing.

3. Offer useful models and suggestions to both guests and hosts. Our research could be directly applied to concrete pricing and evaluations of apartments basing business on a concrete statistical analysis, with the methods found by us. For example, we suggest as a host to provide the positive characteristics posed in evidence by tests when possible. Moreover, as both a host and a guest we give a pipeline to correctly estimate the right, reasonable or favourable price for a house when pricing or looking for it. We are sure that all this information would improve Airbnb users' business, particularly for beginners, but also for experts.

PS: If you are interested and want your house to be correctly priced for Airbnb, give us your information and we will provide you the best price to start!!!