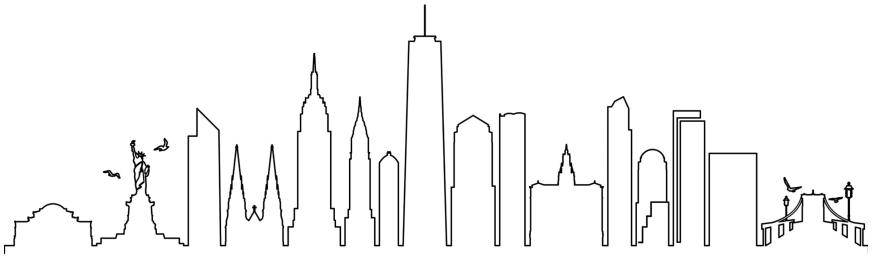
Econometric model: Airbnb prices in New York

Mariana Cerra
Thu Ha Mai
Juan Esteban Cruz

AMSE
2022-2023

Abstract

In the hospitality industry, Airbnb emerges as a housing platform that offers tourists an alternative to traditional hotels. At Airbnb, it is possible to find houses and apartments at a wide range of prices. Using data from Airbnb, this study proposes three models to account for non-linear relationships and to offer insights into the drivers of Airbnb prices in New York in 2019. The results indicate that the generalized additive model (GAM) with an interaction function performs better than the other models.



1 Standard parametric econometric model

1.1 Introduction

New York City is the most populous city in the United States, and according to a 2022 cost-of-living study conducted by Mercer[1], it is in the 7th place of most expensive cities in the world. Nevertheless, it is a top global destination for visitors drawn to museums, entertainment, restaurants, and commerce.

In this context, Airbnb appears as an alternative for travelers looking for more affordable options to stay. Airbnb is an online marketplace that lets property owners rent out their spaces to travelers, whether for a few days or months. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves. We can expect the type of rooms and the neighborhood to impact the price significantly. Therefore this project aims to identify the main drivers of NYC Airbnb prices during 2019.

This study proceeds as follows. Section 1 presents the data, a detailed description of our main variables, the descriptive statistics, and the standard parametric model. Then, section 2 introduces two generalized additive models in an effort to tackle the research question in a non-parametric way. Section 3 compares the previous results and proposes a parametric model that accounts for nonlinearities that were missed in the first approach.

1.2 Presentation of the data

The database comes from "Inside Airbnb,"[2] a mission-driven project that provides data and advocacy about Airbnb's impact on residential communities. The source of data behind the Inside Airbnb site is publicly available information from the Airbnb site, and from it, we used the following variables:

- *neighbourhood_group* is a categorical variable that contains four neighborhood groups: Manhattan, Brooklyn, Queens, and Bronx;
- *latitude* and *longitude* give us the coordinates of the location;
- *room_type* is a categorical variable that indicates the type of accommodation, which could be a shared room, a private room, or an entire apartment/house;
- *minimum_nights* is the minimum number of nights the property has to be booked for;
- *number_of_reviews* and *reviews_per_month* give information about the reviews of each property;
- *availability_365* tells how many days this property is available in a year;
- *price* per night in US dollars will be our target variable.

We started our sample with 48.895 observations. We proceed to take charge of the missing values. Then, by analyzing the minimum value of the variable price, we found out there were places with a cost of $0. Since this does not make sense in practice, we decided to include only the positive

[1] https://www.mercer.com/our-thinking/career/cost-of-living.htmlcollapse0

[2] http://insideairbnb.com/

values. With this procedure, we ended up with a sample of 48.884 observations. To guarantee the robustness of the models and obtain more interpretable results, we transformed the dependent variable price to its logarithmic form log price.

In the database, the variables we selected presented outliers, as can be inferred from Graph A.1 in the appendix. To identify them, we defined the bounds to consider if a value is an outlier or not by using the interquartile range (IQR) criterion. IQR criterion means that observations above the 75th or below the 25th percentile are considered potential outliers by a factor of 1.5 times the IQR. Then, our final sample has 31.575 observations, and the distribution of the variables is as in Graph A.2 in the appendix.

Given our variables are now well defined, we outline in Tables 1 and 2 the main descriptive statistics for the cleaned database.

Table 1: Descriptive statistics of numerical variables

Statistic	Min	1st Q.	Median	Mean	3rd Q	Max
logprice	2.303	4.248	4.700	4.755	5.193	9.210
latitude	40.58	40.69	40.72	40.73	40.76	40.87
longitude	−74.04	-73.98	−73.96	−73.96	-73.94	-73.87
minimum_nights	1	1	2	2.898	4	10
number_of_reviews	0	1	3	9.766	13	65
reviews_per_month	0	0.03	0.23	0.5572	0.87	2.93
availability_365	0	0	2	79.77	129	365

Table 2: Descriptive statistics of categorical variables

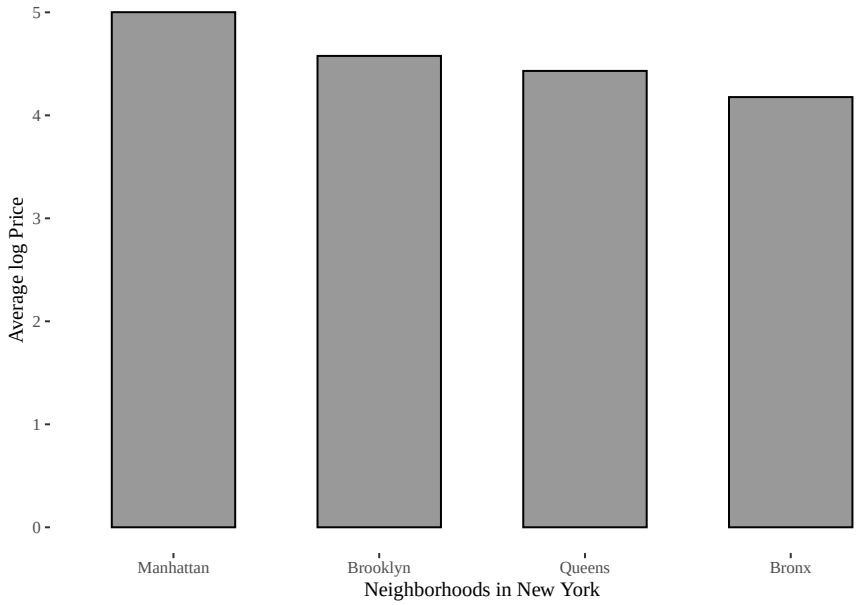
room_type	Frequency	Percentage
Entire home/apt	15955	0.505
Private room	14886	0.471
Shared room	734	0.023

neighbourhood_group	Frequency	Percentage
Bronx	406	0.012
Brooklyn	14406	0.456
Manhattan	14419	0.456
Queens	2344	0.074

Having depicted the procedure to get our final sample, we will continue with some insights into our main variables. We consider the following section fundamental to base our motivation for this study. This scheme makes us wonder: How important the neighborhood is when looking for the best prices? Does the type of room influence the cost of a night? What is the impact of the minimum number of nights to be booked? Or does the price have other drivers? These are the questions we will intend to answer in this study.

In figure 1, we plotted the average log price of rooms in each neighborhood group. The behavior of the log price tends to vary among neighborhoods, where Manhattan is the most expensive, and the Bronx is the cheapest, on average.

Figure 1: Average log price by neighborhood



Then, in figure 2, we have the average log price of each neighborhood group according to the type of accommodation. We witness the same behavior in the price of each type of room in all the neighborhoods; the cheapest is the shared room, while the most expensive is the entire apartment or house, on average. The first conclusion we could extract from this analysis is that, for a whole apartment in Manhattan, we could expect the price to be higher than for a shared room in the Bronx. Nevertheless, to better interpret and analyze the drivers of the price, it is recommended to await the econometric estimations.

Figure 2: Average log price by neighborhood and type of room

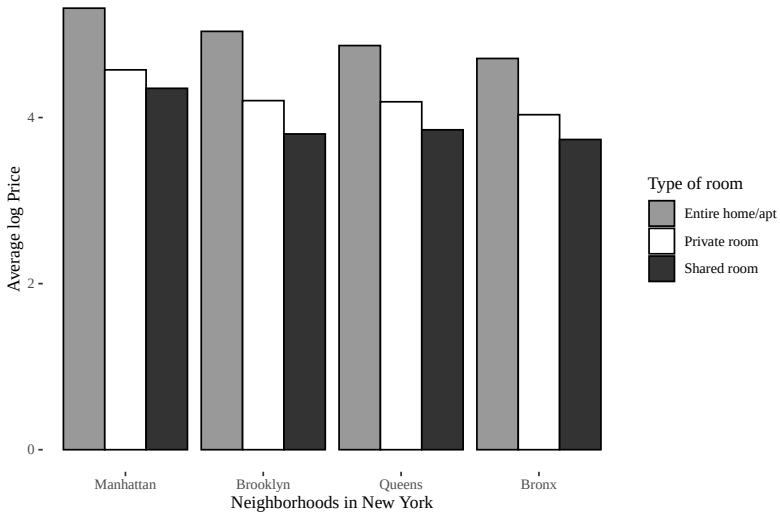
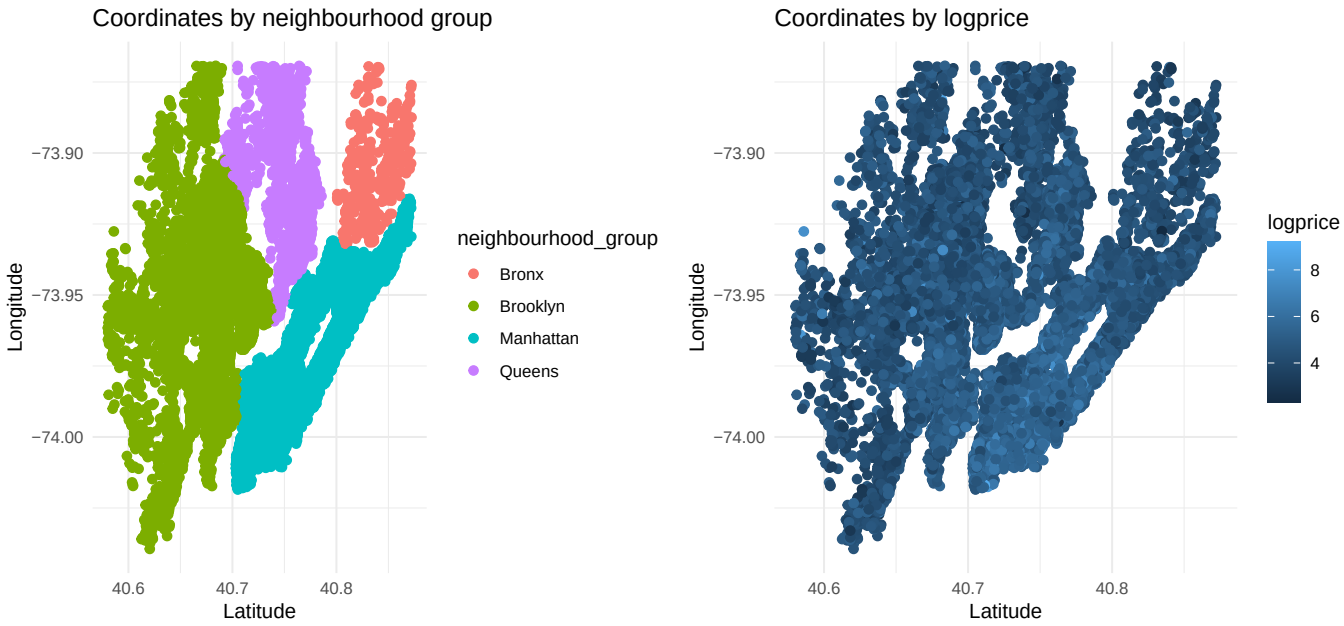


Figure 3 helps us to understand how the price behaves according to latitude and longitude coordinates. In the left-side graph, we have the coordinates by the four neighborhood groups, and in the right-side chart, we have the price on a scale where the darker the intensity of the color, the lower the log price of the house. We could infer from it that the prices in Manhattan are higher, given that the color is lighter for that region.

Figure 3: Coordinates by neighborhood group and by log price



1.3 Regression model

In this first part of the study, we will perform an Ordinary Least Squares estimation (OLS). The econometric model we have implemented is the following,

$$\begin{aligned} logprice_i = \beta_0 + \beta_1 \cdot neighbourhood\_group_i + \beta_2 \cdot latitude_i + \beta_3 \cdot longitude_i + \beta_4 \cdot room\_type_i + \\ \beta_5 \cdot minimum\_nights_i + \beta_6 \cdot number\_of\_reviews_i + \beta_7 \cdot reviews\_per\_month_i \\ + \beta_8 \cdot availability\_365_i + u_i \quad (1) \end{aligned}$$

where $logprice_i$ is the logarithmic price of property i, in New York City in 2019. The subscript i refers to a property. For the second OLS estimation, we added an interaction term between the room type and the neighborhood group, given the suspicion that these two variables combined might significantly affect the price. This second model is as follows,

$$\begin{aligned} logprice_i = \alpha_0 + \alpha_1 \cdot neighbourhood\_group_i + \alpha_2 \cdot latitude_i + \alpha_3 \cdot longitude_i + \alpha_4 \cdot room\_type_i + \\ \alpha_5 \cdot minimum\_nights_i + \alpha_6 \cdot number\_of\_reviews_i + \alpha_7 \cdot reviews\_per\_month_i + \alpha_8 \cdot availability\_365_i + \\ \alpha_9 \cdot (neighbourhood\_group_i \cdot room\_type_i) + \varepsilon_i \quad (2) \end{aligned}$$

1.4 Estimation results

In this section, we will expose in Table 3 our estimation results of the two models using Ordinary Least Squares (OLS). The first column shows the result of the first model and the second column for the second model. We will then discuss the results and draw some conclusions.

Table 3: OLS estimation results

	Dependent variable:	
	logprice	
	(1)	(2)
neighbourhood_groupBrooklyn	0.140***	0.187***
	(0.028)	(0.051)
neighbourhood_groupManhattan	0.282***	0.298***
	(0.026)	(0.050)
neighbourhood_groupQueens	0.209***	0.191***
	(0.028)	(0.052)
latitude	0.693***	0.559***
	(0.101)	(0.102)
longitude	−5.809***	−5.695***
	(0.145)	(0.145)
room_typePrivate room	−0.766***	−0.729***
	(0.006)	(0.057)
room_typeShared room	−1.142***	−1.031***
	(0.019)	(0.095)
minimum_nights	−0.024***	−0.024***
	(0.001)	(0.001)
number_of_reviews	−0.002***	−0.002***
	(0.0002)	(0.0002)
reviews_per_month	−0.032***	−0.032***
	(0.005)	(0.005)
availability_365	0.001***	0.001***
	(0.00002)	(0.00002)

	(1)	(2)
Brooklyn:Private room		−0.084
		(0.058)
Manhattan:Private room		−0.002
		(0.058)
Queens:Private room		0.030
		(0.061)
Brooklyn:Shared room		−0.287***
		(0.099)
Manhattan:Shared room		0.021
		(0.099)
Queens:Shared room		−0.063
		(0.109)
Constant	−452.882***	−439.032***
	(13.401)	(13.480)
Observations	31,575	31,575
R^2	0.500	0.502
Adjusted R^2	0.500	0.501
Residual Std. Error	0.494 (df = 31563)	0.493 (df = 31557)
F Statistic	2,870.327*** (df = 11; 31563)	1,868.980*** (df = 17; 31557)
Note:		*p<0.1; **p<0.05; ***p<0.01

Column 1 presents our primary model corresponding to equation (1) in section 1.3. From this first introductory regression, we get that all the variables in the model significantly impact the price of Airbnb places at the 1% level. In the case of the neighborhood group, with similar other characteristics, a property in Manhattan is expected to cost 28.2% more than one in the Bronx. The same applies to Brooklyn and Queens, which are expected to cost 14% and 20.9% more than a place in the Bronx.

When analyzing the type of room, we found that a private room would have a price 76.6% lower than an entire apartment/house, with the rest of the variables constant. In the case of a shared space, this would lower up to 114.2%. These results give us a first insight that our reasoning from the graphical representation of the data may be represented in our regression results.

Concerning the other variables, one additional night required to book the property would decrease the price by 2.4%, ceteris paribus, given that adding requirements to the renter should be compensated by a lower price. One extra review would lower the cost by 0.2%, one additional review in the month would decrease it by 3.2%, and one extra day that the property is available during the year will increase the price by 0.1% for a place with similar other characteristics.

Continuing with the OLS regression, column 2 shows the model's results when adding the interaction term between the neighborhood group and the room type. One change we notice concerning the model from column 1 is that now the expected cost of a property in Brooklyn is higher. Additionally, we do not observe a significant improvement in the adjusted R-squared.

2 Non-parametric econometric model

2.1 Regression model

The model presented in equation 3 is known as a generalized additive model (GAM), or partially linear model, in which we keep the parametric estimations for five features, namely the two dummy variables, neighborhood_group, and room_type, as well as three numerical ones, number_of_reviews, availability_365, and reviews _per_month. The reason for not applying the smoothing terms for the three aforementioned numerical features is that the 2Ds plots depicting the relationship between the target variable and these features do not show extreme non-linearity, despite the opposite result from the hypothesis testing extracted from the GAM regression. Another reason is for the sake of simplicity. Therefore, the smoothing terms are applied for the three latter features, latitude, longitude, and minimum_nights.

$$\begin{aligned} logprice_i = \gamma_0 + \gamma_1 \cdot neighbourhood\_group_i + \gamma_2 \cdot room\_type + \gamma_3 \cdot number\_of\_reviews_i + \\ \gamma_4 \cdot reviews\_per\_month_i + \gamma_5 \cdot availability\_365_i + m_1(latitude_i) + m_2(longitude_i) + \\ m_3(minimum\_nights_i) + \eta_i \end{aligned} \quad (3)$$

The main limitation of the additive model in equation (3) is that it can ignore pertinent interactions. For example, we have geographical variables in the model, which we assume to be highly nonlinear. Therefore, it might be advisable to propose a more flexible model where we consider the interaction between the latitude and the longitude. This way, the spatial dependence is specified fully nonparametrically. The GAM model in equation (4) introduces this interaction function.

$$\begin{aligned} logprice_i = \theta_0 + \theta_1 \cdot neighbourhood\_group_i + \theta_2 \cdot room\_type + \theta_3 \cdot number\_of\_reviews_i + \\ \theta_4 \cdot reviews\_per\_month_i + \theta_5 \cdot availability\_365_i + m_1(latitude_i \cdot longitude_i) + \\ m_2(minimum\_nights_i) + \nu_i \end{aligned} \quad (4)$$

2.2 Estimation results

We will proceed by presenting three models in Table 4, where the first column corresponds to equation (3), the second to equation (4), and the third column to equation (1) to make comparisons.

Table 4

	Dependent variable:		
	logprice		
	(1)	(2)	(3)
latitude			0.693***
			(0.101)
longitude			−5.809***
			(0.145)
neighbourhood_groupBrooklyn	0.063**	0.118**	0.140***
	(0.032)	(0.056)	(0.028)
neighbourhood_groupManhattan	0.100***	−0.004	0.282***
	(0.027)	(0.049)	(0.026)
neighbourhood_groupQueens	−0.138***	0.149***	0.209***
	(0.029)	(0.057)	(0.028)
room_typePrivate room	−0.733***	−0.724***	−0.766***
	(0.006)	(0.006)	(0.006)
room_typeShared room	−1.103***	−1.088***	−1.142***
	(0.018)	(0.018)	(0.019)
minimum_nights			−0.024***
			(0.001)
number_of_reviews	−0.002***	−0.002***	−0.002***
	(0.0002)	(0.0002)	(0.0002)
reviews_per_month	−0.027***	−0.027***	−0.032***
	(0.005)	(0.005)	(0.005)
availability_365	0.001***	0.001***	0.001***
	(0.00002)	(0.00002)	(0.00002)
Constant	5.003***	5.001***	−452.882***
	(0.029)	(0.050)	(13.401)

	Approximate significance of smooth terms: effective degrees of freedom (Ref. df)		
	(1)	(2)	(3)
s(latitude)	8.762***		
	(8.984)		
s(longitude)	8.931***		
	(8.999)		
s(latitude,longitude)		28.63***	
		(28.991)	
s(minimum_nights)	7.044***	6.79***	
	(7.847)	(7.635)	
Observations	31,575	31,575	31,575
Adjusted R^2	0.537	0.545	0.500
Log Likelihood	−21,313.760	−21,046.420	−22,517.750
UBRE	0.226	0.222	0.244
Note:		*p<0.1; **p<0.05; ***p<0.01	

In column (1), we have the additive model, where the impact of every feature on the log price can be studied while maintaining the rest of the variables constant.

We can make some standard interpretations, maintaining the ceteris paribus assumption. On average, a property in Brooklyn and Manhattan neighborhoods is expected to cost 6.3% and 10% more than in the Bronx, respectively. At the same time, in Queens, this comparison is 13.8% lower. A private and shared room in the Bronx is expected to cost 73.3% and 110.3% less than a whole apartment or a house in the same neighborhood and conditions. Accommodation with an extra review and an additional review per month is also expected to cost 0.2% and 2.7% less, respectively. It is worth mentioning that all the coefficients are significant at the 1% level, except for the neighborhood_group in Brooklyn which is statistically significant at the 5% level.

Moving to the non-parametric components, the p-values of the three features all show non-linear relationships with the target variable. The following part will give some information about these non-linearities in more detail.

Similarly, we could analyze the results from the second column, where we consider the interaction function. Under similar characteristics, a place in Brooklyn and Queens neighborhoods is expected to cost 11.8% and 14.9% more than in the Bronx, respectively, while that in Manhattan shows no statistically significant difference. By examining the impact of the room type, we obtain quite

similar results from those in column (1), as well as for the rest of the numerical variables.

Then, in the non-parametric components, the p-values of the feature minimum_nights and the interaction term between latitude and longitude all show non-linear relationships with the target variable.

The adjusted R-squared in the second model shows little improvement, compared to the first model, without the interaction term between latitude and longitude (54.5% vs. 53.7%).

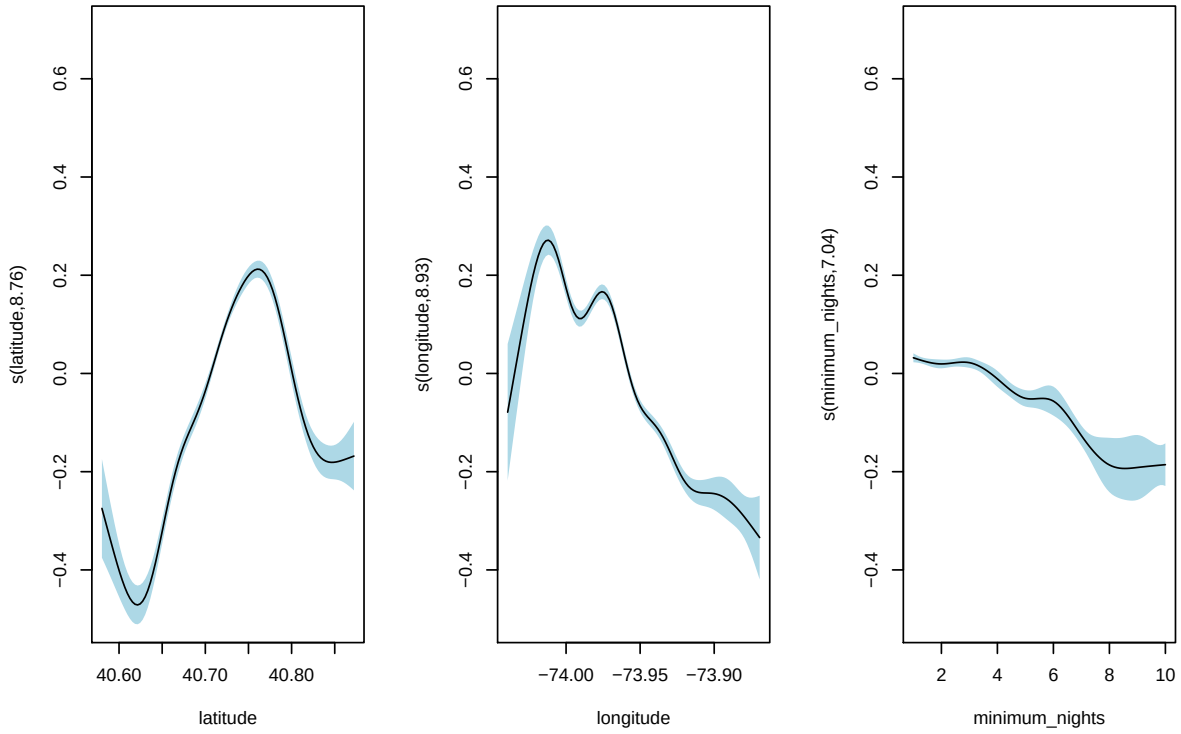
3 Results comparison

3.1 Nonlinearities in the parametric approach

The three component plots in figure (4) help examine the effect of different smoothing features on the target variable. First of all, the plots on latitude and longitude do not suggest simple parametric modeling, and the spatial dependence has to be specified nonparametrically. On the contrary, the component plot of minimum_nights suggests that a parametric relationship could capture this non-linearity.

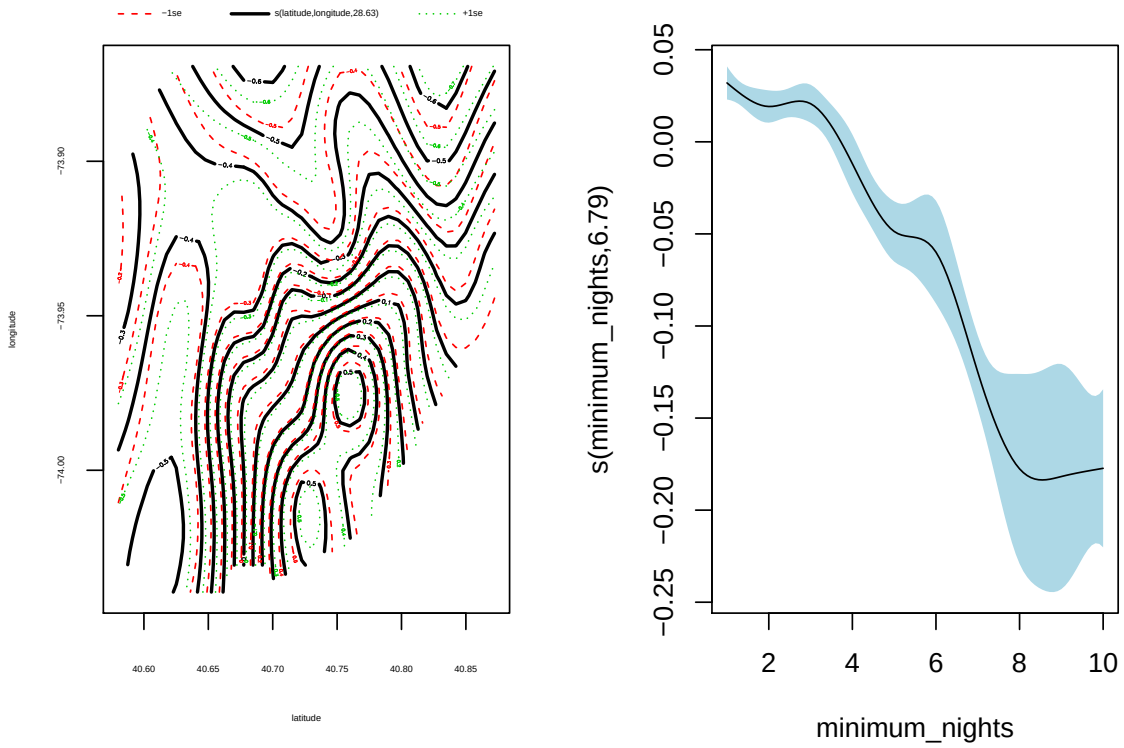
Looking closer at the component plot of minimum_nights, a piecewise linear model is recommended with the choice of two knots at around 3.8 and 8, as the curve seems to decrease slightly until the first knot, then continues with a sharp decrease, and from the second knot a subtle increase. This assumption leads to a proposal of another parametric model with the introduction of a piecewise linear for the minimum_nights feature.

Figure 4: Component plots



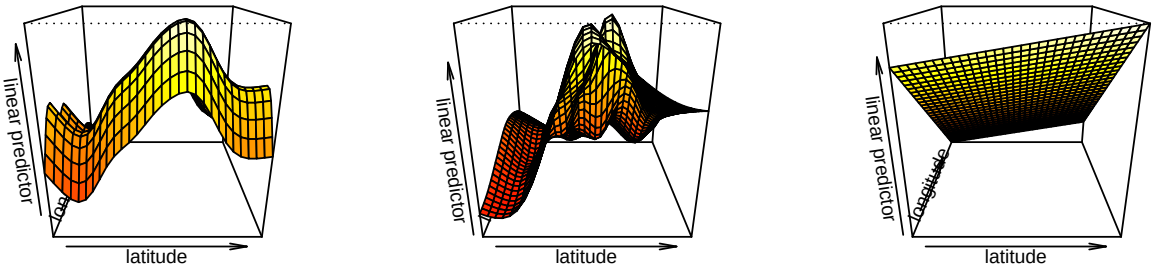
In figure 5, we plotted the two nonparametric components from the model with the interaction function. On the right-hand side, we have the variable minimum_nights. Here we could interpret that, for example, it first decreases with a low slope and then continues with a steeper slope to end up constant, while this kind of analysis could not be done with a fully nonparametric model. On the left-hand side, we have a contour plot of the interaction of latitude and longitude. To gain more insights, we will plot a 3D figure shown in graph 6.

Figure 5: Model with interaction function



With respect to figure 6, in the left-hand plot, the model for the latitude and longitude coordinates is additively decomposable. As a result, we are already able to capture some nonlinearities. This is modeled as $m_1(latitude) + m_2(longitude)$. In the middle plot, we have the model for coordinates specified nonparametrically with the interaction function. This model is more flexible than the previous one and allows us to identify more complex nonlinearities. We observe a strong nonlinearity, which confirms that these relationships would be difficult to capture with a parametric model. This is specified as $m_1(latitude, longitude)$. On the right-hand side, there is the plot for the linear model, defined as $\beta_1 latitude + \beta_2 longitude$.

Figure 6: 3D Figure for the three models



3.2 Another parametric model

As explained above, this new parametric model is proposed to capture some non-linearities missed in the first parametric approach. This model contains a linear effect of all features and a piecewise linear relationship with two knots at 3.8 and 8 for the minimum_nights variable. Their coefficients will indicate the direction and magnitude of the change.

$$\begin{aligned} logprice_i = \theta_0 + \theta_1 \cdot neighbourhood\_group_i + \theta_2 \cdot latitude_i + \theta_3 \cdot longitude_i + \theta_4 \cdot room\_type_i + \\ \theta_5 \cdot minimum\_nights_i + \theta_6 \cdot (minimum\_nights_i - 3.8)_+ + \theta_7 \cdot (minimum\_nights_i - 8)_+ \\ + \theta_8 \cdot number\_of\_reviews_i + \theta_9 \cdot reviews\_per\_month_i + \theta_{10} \cdot availability\_365_i + \upsilon_i \end{aligned} \quad (5)$$

The results in Table 5 show that all features are statistically significant at the level of 1%, except for the minimum_nights at 5%. As expected, the price for a private room and a shared room is lower than a whole apartment or house when other characteristics remain constant. Similarly to the previous OLS model, accommodation in Manhattan, Queens, and Brooklyn is expected to be more expensive than a property in the Bronx with the same conditions. Additionally, negative coefficients for number_of_reviews and reviews_per_month suggest a hypothesis that a larger part of reviews about an accommodation is bad ratings. By introducing a piecewise linear relationship for the minimum_nights covariate, surprisingly, there is a positive and significant sign in the coefficient

of minimum_nights3 where we put a knot at 8. This result shows that the introduction of the knot was necessary to account for the nonlinearity. When comparing the three models that attempt to account for the nonlinearities, the GAM with the interaction function from equation (4) is proven to work slightly better when its adjusted R-squared equals 54.5% compared to 53.7% in the GAM model from equation (3) and 50% in the second parametric model from equation (5).

Table 5

	Dependent variable:
	logprice
neighbourhood_groupBrooklyn	0.138^{***}
	(0.028)
neighbourhood_groupManhattan	0.282^{***}
	(0.026)
neighbourhood_groupQueens	0.209^{***}
	(0.028)
latitude	0.687^{***}
	(0.101)
longitude	-5.811^{***}
	(0.145)
room_typePrivate room	-0.761^{***}
	(0.006)
room_typeShared room	-1.132^{***}
	(0.019)
minimum_nights	-0.007^{**}
	(0.003)
minimum_nights2	-0.038^{***}
	(0.006)
minimum_nights3	0.041^{***}
	(0.015)

number_of_reviews	-0.002^{***}
	(0.0002)
reviews_per_month	-0.032^{***}
	(0.005)
availability_365	0.001^{***}
	(0.00002)
Constant	-452.862^{***}
	(13.394)
Observations	31,575
R^2	0.501
Adjusted R^2	0.500
Residual Std. Error	0.493 (df = 31561)
F Statistic	$2{,}434.242^{***}$ (df = 13; 31561)
Note:	*p<0.1; **p<0.05; ***p<0.01

4 References

https://m-clark.github.io/generalized-additive-models/application.html#visualization-1

https://noamross.github.io/gams-in-r-course/

https://online.stat.psu.edu/stat501/lesson/8/8.8

5 Appendix

Figure A1. Data distributions before removing outliers

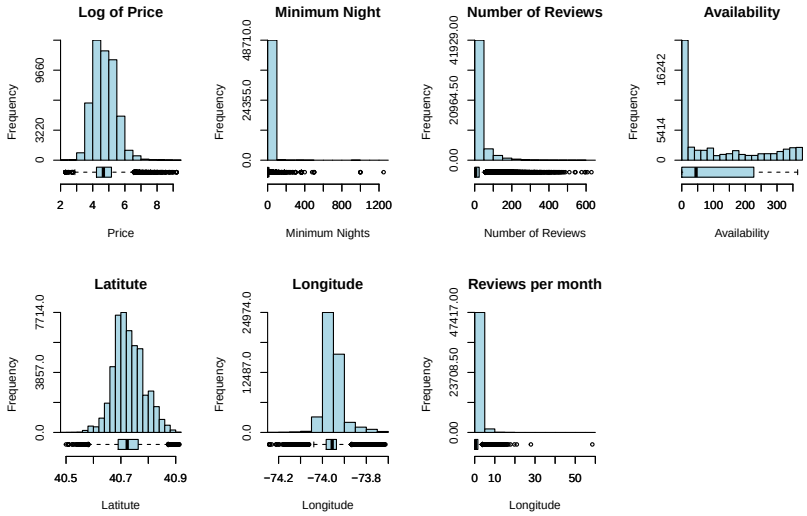


Figure A2. Data distributions after removing outliers