Course: Data Analysis 2
Author: Márton Nagy

Finding a good AirBnB deal for a family vacation in Oslo

1. Executive summary

I have built a regression model on price per night per guest versus distance to the city center for AirBnBs in Oslo at the end of June 2024. The model shows that the pattern of association is non-linear: a km difference in distance corresponds to a larger average percentage difference in prices (6.7%) for listings that are closer to the center than 4 km than for ones that are further away (4.2% average expected price difference). To capture the non-linear pattern, log-transformation of the price was needed, and a piecewise linear spline model was built. The model may be used to facilitate the selection of good deals for a family vacation in Oslo.

2. Introduction

2.1. Goal of the analysis

The goal of my analysis is to find the most underpriced AirBnBs with respect to their distance from the city center in Oslo for an imagined family vacation in July 2024.

2.2. Variables

My dependent variable is price per night per guest in an AirBnB (measured in EUR), and my independent variable is the AirBnB's calculated distance from the city center of Oslo (measured in kilometers). The population is all AirBnBs in the city of Oslo[1]. My initial sample was intended to provide complete coverage of the population on 29 June 2024 – however, as we do not precisely know the data collection process applied, we cannot be sure of this. So, we should rather think about the sample as one with relatively large coverage of AirBnBs in Oslo at the end of June 2024, though selection into the sample may not be random. The sample is relatively clean with no obvious duplicate entries and no entity resolution needed. Erroneous values for my variables were scarce (see Appx. A3.1). The key data quality issue is thus that of the missing values: data seems to be missing at random for price, but not for host response rate (for further details on this see Appx. A2 and A7). I also present the distribution of the variables in Appx. A3.1.

3. Data

3.1. Selecting observations

I narrowed down the sample to fit my analytical goals. I imagined travelling with a larger family, slightly on a budget, who do not mind splitting up and renting more than one place if needed, and who do not like commuting a lot on a vacation. This specification can be translated into the following constraints[2]: price per night per guest should be maximum 100 EUR, number of people accommodated should be maximum 6, and distance to center should be maximum 10 km. With

[1] Alternatively, we could rather think of a general pattern instead of a population. This general pattern would include all possible AirBnB listings in Oslo, and my sample would be a random realization of this general pattern at a given point in time.

[2] The constraints appr. correspond to the following percentiles in the unconstrained sample: below 97th for price, below 94th for number of people accommodated, and below 99th for distance.

these constraints, I included 8,080 observations out of the initial 10,099[3].

3.2. Summary statistics

We can now look at the key characteristics of the distributions of our dependent and independent variable in the filtered sample. These are summarized in Table 1 and visualized in Appx. A3.2.

Table 1: Summary statistics for our variables of interest

	Count	Mean	Std	Min	5%	25%	50%	75%	95%	Max
Price per night per guest (EUR)	8,080	41.95	16.43	5.85	20.69	29.63	39.49	51.70	74.58	99.96
Distance to center (km)	8,080	2.66	1.88	0.15	0.68	1.38	2.14	3.24	6.67	9.99

Our price variable ranges from appr. 6 to 100 EUR, with the average price per night per guest being 42 EUR. The distribution is skewed with a longer right tail (indicated by the median value of 39 EUR being slightly below the mean). The standard deviation of the distribution is 16 EUR. The distance to the center ranges between 150 meters and 10 kilometers. The average distance is appr. 2.7 km, and the standard deviation is 1.9 km. This distribution is also skewed with a long right tail, as indicated by the median value 2.1 km being lower than the mean.

3.3. Transformation of variables

Looking at the histograms in Appx. A3, we can see that both of our variables may be approximated by a log-normal distribution (longer right tail, but there is also an upward going part). This indicates (from a statistical point of view) that it might be sensible to log-transform both variables so that their distributions approximate a normal distribution (see the log histograms in Appx. A3.3). This makes sense if we want to *compress* the distribution so that extreme values in either of the variables do not distort our regression estimates. However, from a substantial point of view, log-transformation may have its caveats. For example, thinking of distances to the city center in terms of percentages may not be very intuitive, so this transformation hinders easy interpretation. On the other hand, prices are easy to think about in relative differences, and their log-transformation makes sense from a statistical point of view. So, for my main model, I have chosen to log-transform price per night per guest but leave the distance to center at level (although some other models have also been estimated with different transformations, as presented Appx. A5).

4. Model and evaluation

4.1 Estimating different models

During the analysis, I have estimated six different models: four simple linear models in level-level, level-log, log-level and log-log specifications; a piecewise linear spline model of the log-level set-up, and a quadratic model for the log-log case. From these models, the piecewise linear spline was chosen for the main reasons of fitting the non-parametric functional form (see lowess charts in Appx. A4) quite well and of ease of interpretation. The different models are presented in detail in Appx. A5 with visualizations and more detailed arguments backing the model selection.

4.2. Model choice

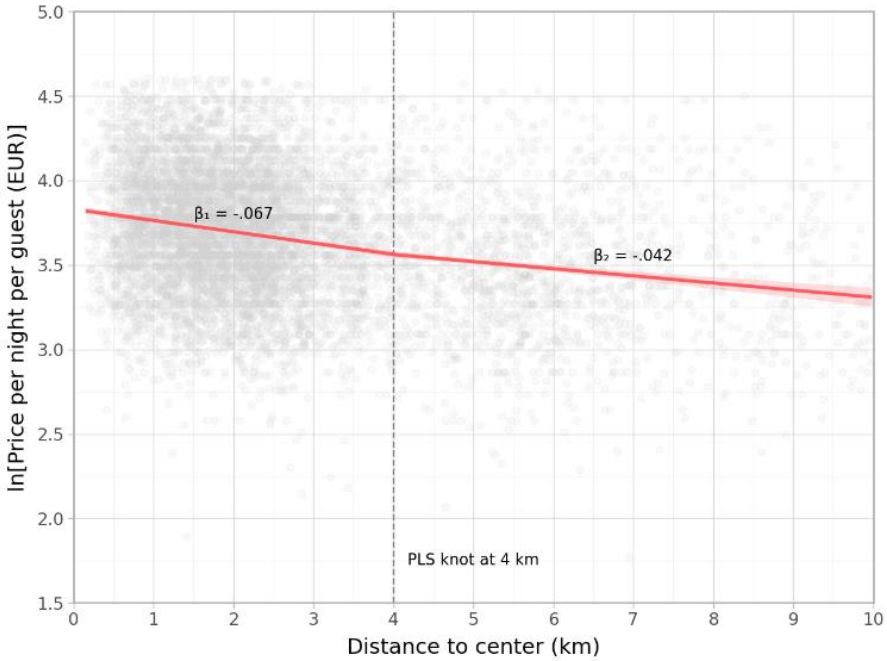
The model used for further analysis is a piecewise linear spline model of the logarithm of the price per guest in EUR versus the distance to the city center of Oslo in km, with a knot at 4 kilometers:

[3] Observations where host acceptance rate is missing are included, as excluding them leads to significantly different results – see in detail in Appx. A7.

$$ln(ppg)^E = \alpha_1 + \beta_1 d \cdot \mathbb{I}(d < 4) + (\alpha_2 + \beta_2 d) \cdot \mathbb{I}(d \geq 4),$$

where $ln(ppg)^E$ is the conditional expectation on the natural logarithm of price per guest in EUR, d is the distance to the city center of Oslo in km, $\mathbb{I}$ is the indicator function to fit a PLS model.

Figure 1: The estimated PLS model along with a scatterplot (N=8,080)



Note: 95% confidence interval on predicted values indicated by the shaded area.

The model, as illustrated in Figure 1 and presented in detail later in Table 1, suggests that for AirBnBs closer than 4 km to the center, listings that are 1 km further away are on average 6.7 percent less expensive. For places that are at least 4 km away, the model indicates that observations with 1 km higher distance from the center have a 4.2 percent lower price on average. Lastly, the model tells us that the average of log price per guest is 3.8 if the distance from the center would be zero.

4.3. Hypothesis testing

We can test whether our parameters are different from zero. Note, that I have chosen a heteroscedasticity corrected SE for estimating the model, as from the scatterplot the conditional variance of log prices on distance seemed not uniform, thus a homoscedastic SE may have been an underestimation. I will test the parameters' significance on the conventional 5% significance level.

For β_1, the hypotheses can be formulated as follows: $H_0: \beta_1 = 0 \; and \; H_A: \beta_1 \neq 0$. The heteroscedasticity corrected SE is 0.005 and the corresponding t-statistic is -14.60. As this falls outside the 95% confidence interval [-1.96, 1.96] for the t-statistic, we can reject the null and conclude that β_1 is different from zero at 5% significance.

For β_2, the hypotheses can be formulated as follows: $H_0: \beta_2 = 0 \; and \; H_A: \beta_2 \neq 0$. The heteroscedasticity corrected SE is 0.006 and the corresponding t-statistic is -7.68. As this falls outside the 95% confidence interval for the t-statistic, we can reject the null and conclude that β_2 is different from zero at 5% significance.

The whole estimated model is summarized in Table 2. Further interpretation is provided in the concluding section.

Table 2: Piecewise linear splines model on log price per guest versus distance to center for AirBnBs in Oslo

	Coef.	SE	t-stat.	p-val.	[.025	.975]
Constant	3.831	.010	372.801	.000	3.810	3.851
Distance to center < 4	-.067	.005	-14.598	.000	-.076	-.058
Distance to center ≥ 4	-.042	.006	-7.675	.000	-.053	-.031
No. of observations	8,080					
R^2	.072					

Note: Heteroscedasticity corrected standard errors (HC1). PLS knot at 4 kilometers.

4.4. Analysis of the residuals

We can use the above-described model to identify the most under and overpriced AirBnBs by selecting observations with the lowest and highest residuals, presented in Table 3. As price is log-transformed, we can identify pricing in relative terms: that is the least and most percentage differences between the actual and the predicted price levels.

Table 3: The least and most expensive AirBnBs as predicted by the PLS model specification

Category	ID	Price	Distance	Predicted log price	Model log residual	Predicted level price	Level price % diff.
Top 5 underpriced	833801927434015744	6.58	1.42	3.74	-1.85	44.98	-85.37
	6291412	5.85	6.95	3.44	-1.67	33.40	-82.49
	1173863950508313088	8.56	2.86	3.64	-1.49	40.84	-79.05
	32684127	7.90	4.65	3.53	-1.47	36.81	-78.55
	583855494264588928	8.82	3.44	3.60	-1.42	39.29	-77.56
Top 5 overpriced	833381244209057920	96.52	4.88	3.53	1.04	36.46	164.72
	42701431	87.75	7.63	3.41	1.07	32.45	170.36
	1000529416641035520	87.75	7.69	3.41	1.07	32.38	171.00
	1103424424967773696	87.75	7.89	3.40	1.08	32.11	173.28
	644075622945331584	90.29	8.39	3.38	1.13	31.43	187.27

Notes: Level prices (per night per guest) in EUR, level distance (to city center) in km. Predicted level price is calculated using the formula in Békés-Kézdi Ch. 14.3. Level price % difference means the percentage difference between the actual and the predicted prices. Table is ordered ascending by level price % difference.

We can see that the most underpriced listings are cheaper compared to their predicted price by 78-85 percent. Their distance to the center shows some variation so they are not situated in the same area. We can also see that the most overpriced AirBnBs are almost twice as costly as their predicted price. Interestingly, they are situated more on the outskirts of Oslo, as indicated by their distance to the center. The identified listings are visualized in Appx. A6.

5. Conclusion

To sum up, I have constructed a price model on distance to the center for AirBnBs in Oslo suitable for a family vacation. The model uses a PLS specification to uncover the pattern of association between distance to center and log-transformed price per night per guest. The model shows that for AirBnBs that are closer to the city center (below 4 km), listings that are 1 km further away are on average 6.7 percent cheaper. For places that are more than 4 km away, the average expected difference in prices is lower, only -4.2 percent per km. Using the model, I have identified listings that are the most under or overpriced in relative terms compared to their actual price.

Note that the model I used is relatively weak as it only explains 7.2 percent of the variance in prices. Including some further explanatory variables may help to build a more accurate model, though this is out of the scope of this assignment. Also, the external validity of the model is low, as the same pattern of association may not hold for other cities' listings or for other times in Oslo.

Appendix

A1. Data cleaning and manipulation

Before doing anything to the raw data file, I have performed some duplicate checking. It turned out that ID and the listing URL are completely unique. I have checked whether the other columns together have some duplicates, but there were none. Moreover, I checked together only the listing's name, description and host ID. This way, there were 28 duplicate entries. However, as I manually went through the listings' AirBnB pages, I noted that these are in fact unique observations, only the hosts usually do not bother with coming up with new names and descriptions for their properties if they are very similar. One other thing that caught my attention was that the listings' pictures were not unique, though I would have assumed that no picture is used more than once. However, as I also manually checked some duplicate pictures, I noticed that generally, hosts tend to re-use pictures of the outside of the building between listings if e.g. they are renting out two rooms in the same complex. So, all in all, I concluded that the data is generally free from duplicates.

After the duplicate checking, I filtered the dataset to the following variables: id, price, last_scraped, property_type, room_type, neighborhood_cleansed, accommodates, latitude, longitude, host_acceptance_rate, number_of_reviews, and review_scores_rating.

First, I had to deal with the appropriate formatting of the data. I converted last_scraped to a DateTime object (this was needed for converting between NOK and EUR at appropriate exchange rates).

Next, I converted the price string to a floating-point number, which I converted into EUR on the exchange rate of the scrape date, using the currency_converter Python package. Note, that if the exchange rate was not available on the scrape's exact date, I used the latest available rate relative to the scrape date.

Host acceptance rate has also been converted from a string to a numeric value, though I have left it in percentage format. The reason behind that was that if I were to use this variable as an independent variable in a regression, I would want the results to be interpretable as percentage point differences.

Then, I calculated the distance from the city center in kilometers for each listing's coordinates using the haversine formula. The city center of Oslo has been defined by the following coordinates: (59.9139, 10.7522).

The next variable I had to calculate was the price per night per guest. For this, I simply divided the price per night (now in EUR) by the number of people that can be accommodated in a certain AirBnB, and I set the value to NA if either of the variables were missing (in practice, only the price variable was missing sometimes).

Lastly, I grouped property types into broader categories as the ones provided by AirBnB seemed too narrow. The exact mapping I used is omitted from this description, as I ended up not using this variable in my analysis. In the end, I ended up with the following categories and value counts:

Table A1: Value counts for the created broader property type categories (N=10,099)

Property type category	Value count
Apt/Condo	7,872
Private Room - Apt/Condo	1,093
House/Townhouse	791

Property type category	Value count
Private Room – House	138
Guest Suite/House	59
Unique Stays	49
Shared Room	46
Private Room - B&B	29
Hotel-Like	22

After the data manipulation and analyzing the missing values (see in Appx. A2.), I narrowed down my sample to fit my analytical question, as described in Section 3.1.

A2. Missing values

Before narrowing down my sample, I noticed that some variables have a substantive number of missing values. The exact proportions per variable are presented in Table A2.

Table A2: Percentage of missing values in my initial sample (N=10,099)

Variable	Percentage of observations missing (%)
id	0.00
price_EUR	12.37
property_type	0.00
room_type	0.00
neighbourhood_cleansed	0.00
accommodates	0.00
host_acceptance_rate	12.96
number_of_reviews	0.00
review_scores_rating	26.87
distance_to_center	0.00
price_per_guest	12.37
property_type_grouped	0.00

It turns out that we should not worry too much about missing values in ratings, as these are missing if and only if the number of reviews is zero. Thus, these missing values are actually meaningful, and if we were to incorporate this variable into our analysis, we should include these in a meaningful way as well. However, missing values in price per night per guest (note that this is calculated from price per night (price_EUR), so the number of missing observations are the same) and in host acceptance rate might mean a problem. The key question we should check is whether these observations are missing at random. If they do, we should expect no difference in the distribution of other variables conditional on price or acceptance rate missing. To check this visually[4], I below present the conditional boxplots in question. For both variables with missing values, I present conditional distributions of other variables with no transformation and the most sensible transformation.

[4] Note, that while formally testing whether the distributions (or at least the means) are different is possible (and I did calculate the respective t- and Mann-Whitney-U-tests), such detailed statistical results are omitted from this report, as I think that the visual proof is much more telling and it indicates the same concerns as the formal tests.

Figure A1: Conditional distributions of quantitative variables on whether price per night per guest is missing with no transformation

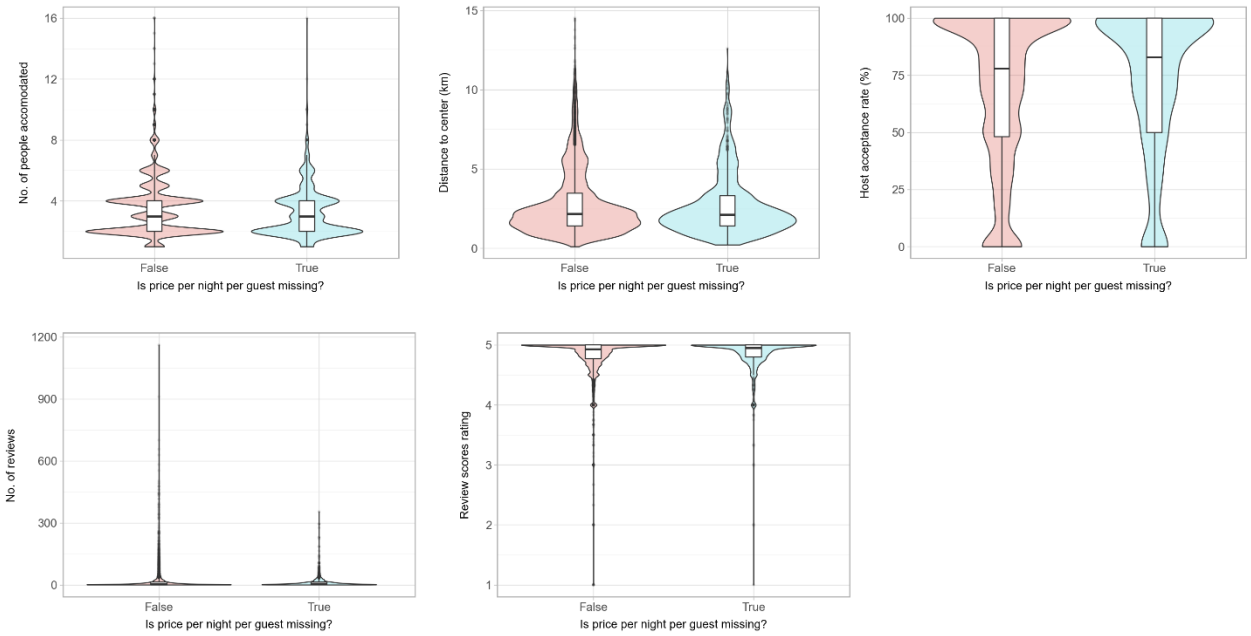
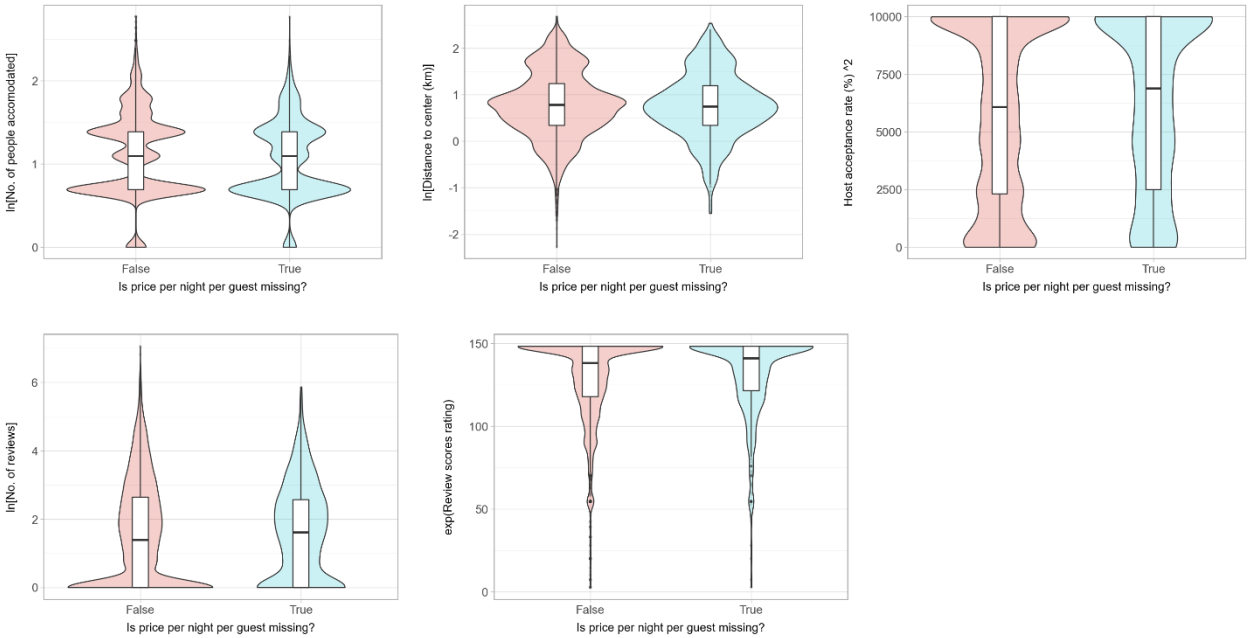


Figure A2: Conditional distributions of quantitative variables on whether price per night per guest is missing with the most sensible transformation to account for skewed distributions



Notes: For the host acceptance rate, a quadratic transformation has been applied. For the review scores rating, an exponential function has been applied. For all other variables, a simple log-transformation has been applied. Listings with zero reviews have been transformed to have one review before applying the log-transformation.

As we can see from the above figures, the conditional distributions (and the conditional means) on whether price per night per guest is missing are rather similar with only minor differences. This may suggest us that price per night per guest is indeed missing at random, so excluding observations with missing values may not distort our results.

Figure A3: Conditional distributions of quantitative variables on whether host acceptance rate is missing with no transformation

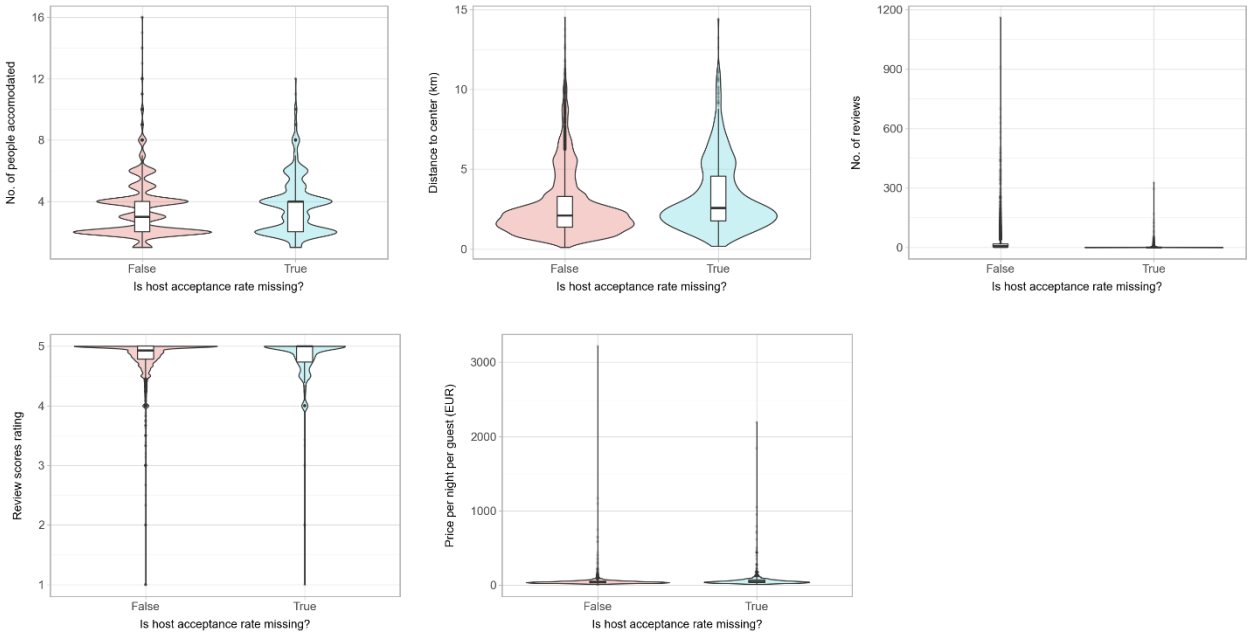
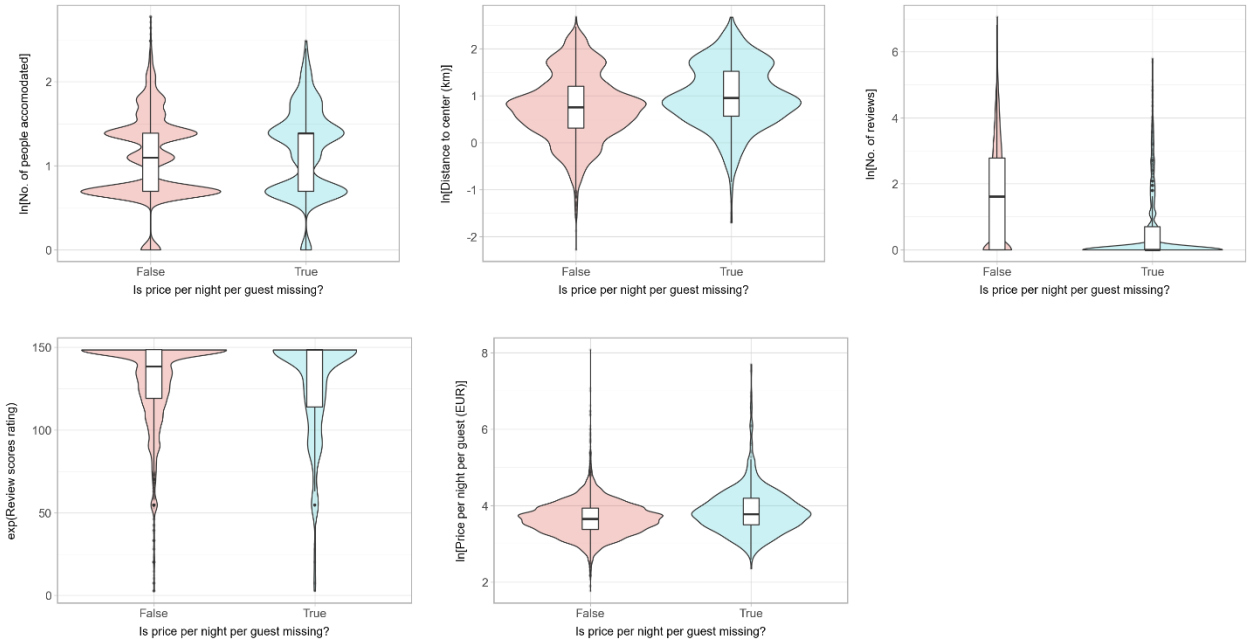


Figure A4: Conditional distributions of quantitative variables on whether host acceptance rate is missing with the most sensible transformation to account for skewed distributions



Notes: For the review scores rating, an exponential function has been applied. For all other variables, a simple log-transformation has been applied. Listings with zero reviews have been transformed to have one review before applying the log-transformation.

The above figures indicate that there is a slight difference in the conditional distributions (and conditional means) on whether host acceptance rate is missing in price per guest per night and distance to center. We can see some differences in the other variables which may further distort our results if we were to drop observations with missing host acceptance rate. To check if such distortions would indeed be present, I conduct a robustness check to this in Appx. A7.

A3. Summary statistics and distributions

A3.1. Full sample, non-transformed variables

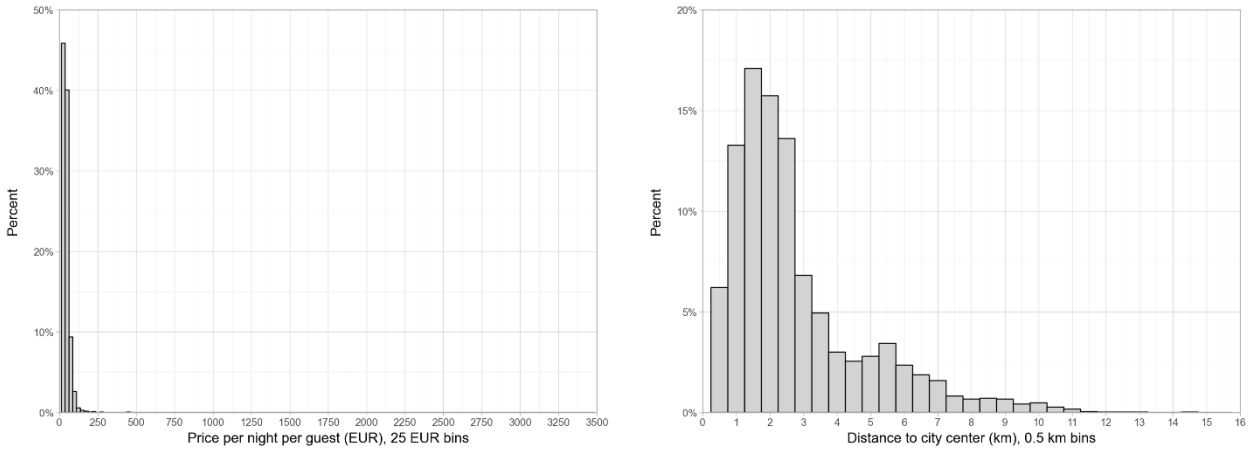
As Table A3 and Figure A5 shows, in our original, unfiltered sample both of our variables are skewed with a long right tail. This is taken to the extreme in the case of price per night per guest, as the 95th percentile is below a 100 EUR, whereas the maximum is over 3,000 EUR[5].

The mean of the prices is 46 EUR per night per guest, while the standard deviation is 60 EUR. The range of prices is between appr. 6 EUR and 3,217 EUR. The mean distance to the center is 2,8 km, while the standard deviation is 2,1 km. Distance to the center ranges from 100 meters to 14,5 km.

Table A3: Summary statistics of our variables of interest, full sample, no transformation

	Count	Mean	Std	Min	5%	25%	50%	75%	95%	Max
Price per night per guest (EUR)	8,850	46.05	60.27	5.85	20.18	29.25	39.49	52.65	83.36	3217.36
Distance to center (km)	10,099	2.79	2.06	0.10	0.69	1.41	2.17	3.43	7.04	14.52

Figure A5: Histograms of our variables of interest, full sample, no transformation (N_{price}=8,850 and $N_{distance}$=10,099)



A3.2. Filtered sample, non-transformed variables

For a detailed description, see Section 3.2.

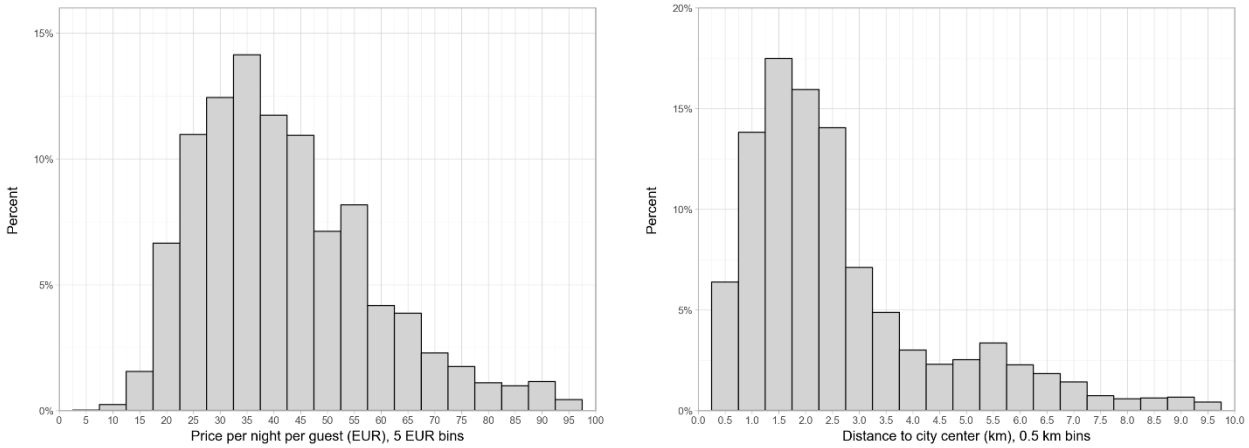
Table A4: Summary statistics of our variables of interest, filtered sample, no transformation

	Count	Mean	Std	Min	5%	25%	50%	75%	95%	Max
Price per night per guest (EUR)	8,080	41.95	16.43	5.85	20.69	29.63	39.49	51.70	74.58	99.96
Distance to center (km)	8,080	2.66	1.88	0.15	0.68	1.38	2.14	3.24	6.67	9.99

Note: same as Table 1.

[5] I have checked the listings with price per night per guest over 300 EUR and the price value seems to be either erroneous (as the currently available price for these places are well below what is listed in our sample) or simply very luxurious. Moreover, prices between 100 and 300 EUR can be also considered somewhat luxurious for our analytical question, thus they were dropped.

Figure A6: Histograms of our variables of interest, filtered sample, no transformation (N=8,080)



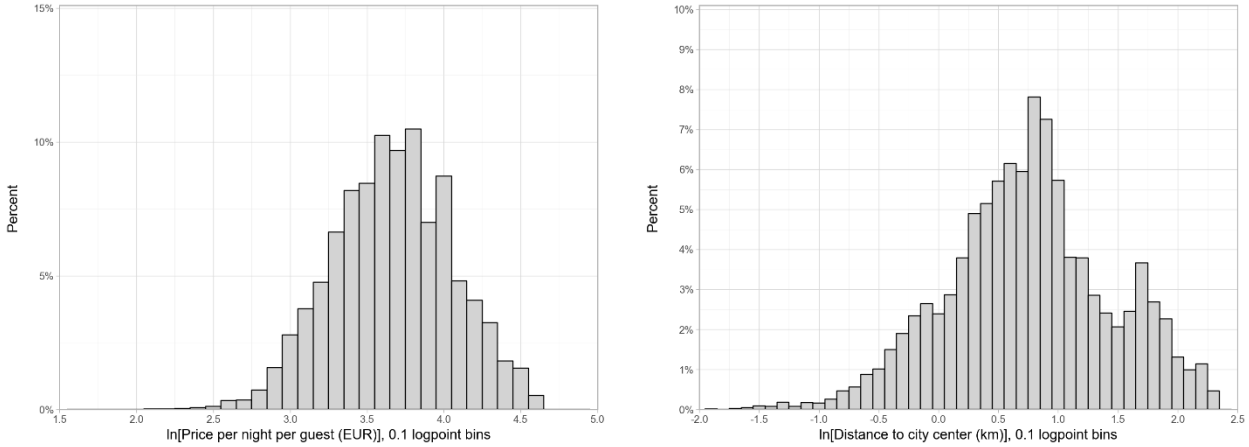
A3.3. Filtered sample, log-transformed variables

After applying a log-transformation for both variables[6], we can see that the distributions have been *compressed* and they now approximate a normal distribution quite well. Note, however, that ln(distance to the center) is bimodal.

Table A5: Summary statistics of our variables of interest, filtered sample, log-transformation

	Count	Mean	Std	Min	5%	25%	50%	75%	95%	Max
ln[Price per night per guest (EUR)]	8,080	3.66	0.39	1.77	3.03	3.39	3.68	3.95	4.31	4.60
ln[Distance to center (km)]	8,080	0.75	0.68	-1.88	-0.38	0.32	0.76	1.18	1.90	2.30

Figure A7: Histograms of our variables of interest, filtered sample, log-transformation (N=8,080)



A4. Lowess estimates for the functional form

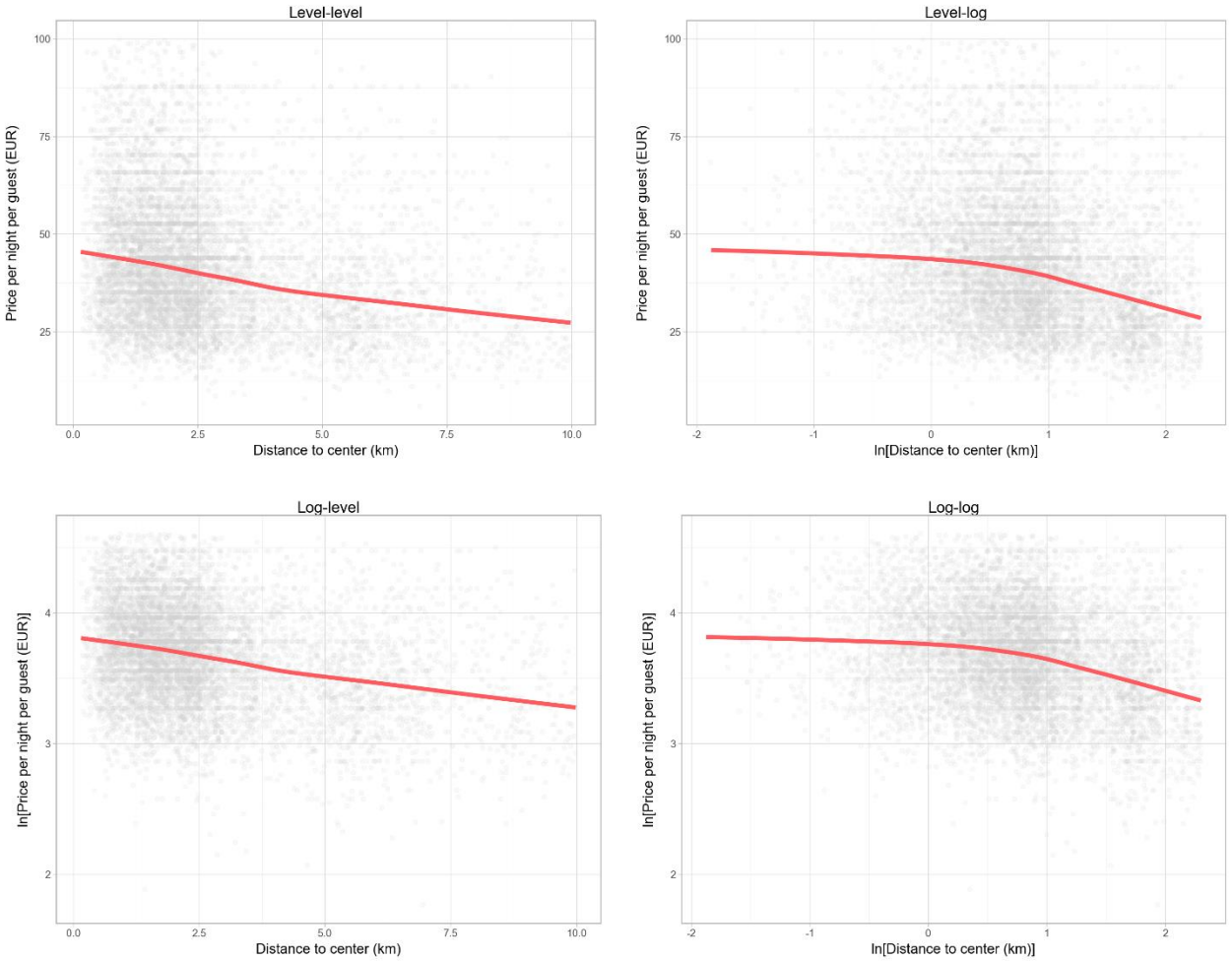
To get an idea on the functional form between our dependent variable and independent variable, I estimated lowess non-parametric regressions on every possible combination of transformation (that is level-level, level-log, log-level and log-log). From the charts in Figure A8, we can see that if distance to center is at level, the functional form is rather linear, with a slight break in the slope around 4 km. This might indicate that a PLS regression could be a good approximation of the

[6] Note, that as neither of the variables had zero or negative values, I could apply the log transformation in a straight-forward way.

functional form in these cases. If price per night per guest is at level, the functional form is rather a concave, very flat parabola, indicating a possible quadratic functional form.

Also, from the scatterplots below the lowess estimates, we can see that a log-transformation compresses the distributions – this might be a good thing if we aim for a better fit.

Figure A8: Lowess estimates with all possible combinations of log-transformation (N=8,080)



A5. Model comparison

To decide on which model to choose, I estimated six regressions: four simple linear models with level-level, level-log, log-level and log-log-transformations (regressions (1)-(4) in Table A6), a PLS model (5) with a knot at 4 km on the log-level transformed variables, and a quadratic model (6) on log-log-transformed variables. Table A6 presents the results of the regressions, while Figure A9 provides a graphical comparison for the different models along with the scatterplots.

Table A6: Comparison table for all estimated model specifications

	dependent variable: price per night per guest (EUR)	dependent variable: ln[price per night per guest (EUR)]				
	Level-level	Level-log	Log-level	Log-log	Log-level with PLS	Log-log Quadratic
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	47.401***	46.044***	3.809***	3.769***	3.831***	3.767***
	(0.314)	(0.280)	(0.007)	(0.006)	(0.010)	(0.006)

	dependent variable: price per night per guest (EUR)		dependent variable: ln[price per night per guest (EUR)]			
	Level-level	Level-log	Log-level	Log-log	Log-level with PLS	Log-log Quadratic
	(1)	(2)	(3)	(4)	(5)	(6)
Distance to center (km)	-2.049***		-0.055***			
	(0.093)		(0.002)			
Distance to center (km) < 4					-0.067***	
					(0.005)	
Distance to center (km) > 4					-0.042***	
					(0.006)	
ln[Distance to center (km)]		-5.446***		-0.143***		-0.056***
		(0.267)		(0.006)		(0.011)
ln[Distance to center (km)]²						-0.061***
						(0.007)
Observations	8,080	8,080	8,080	8,080	8,080	8,080
R^2	0.055	0.051	0.071	0.063	0.072	0.073
Adjusted R^2	0.055	0.051	0.071	0.062	0.072	0.073
Residual Std. Error	15.968 (df=8,078)	16.000 (df=8,078)	0.377 (df=8,078)	0.378 (df=8,078)	0.377 (df=8,077)	0.376 (df=8,077)
F Statistic	488.589*** (df=1; 8,078)	417.298*** (df=1; 8,078)	537.535*** (df=1; 8,078)	488.848*** (df=1; 8,078)	282.150*** (df=2; 8,077)	280.189*** (df=2; 8,077)

Notes: *p<0.1; **p<0.05; ***p<0.01. Heteroscedasticity corrected standard errors (HC1) in brackets.

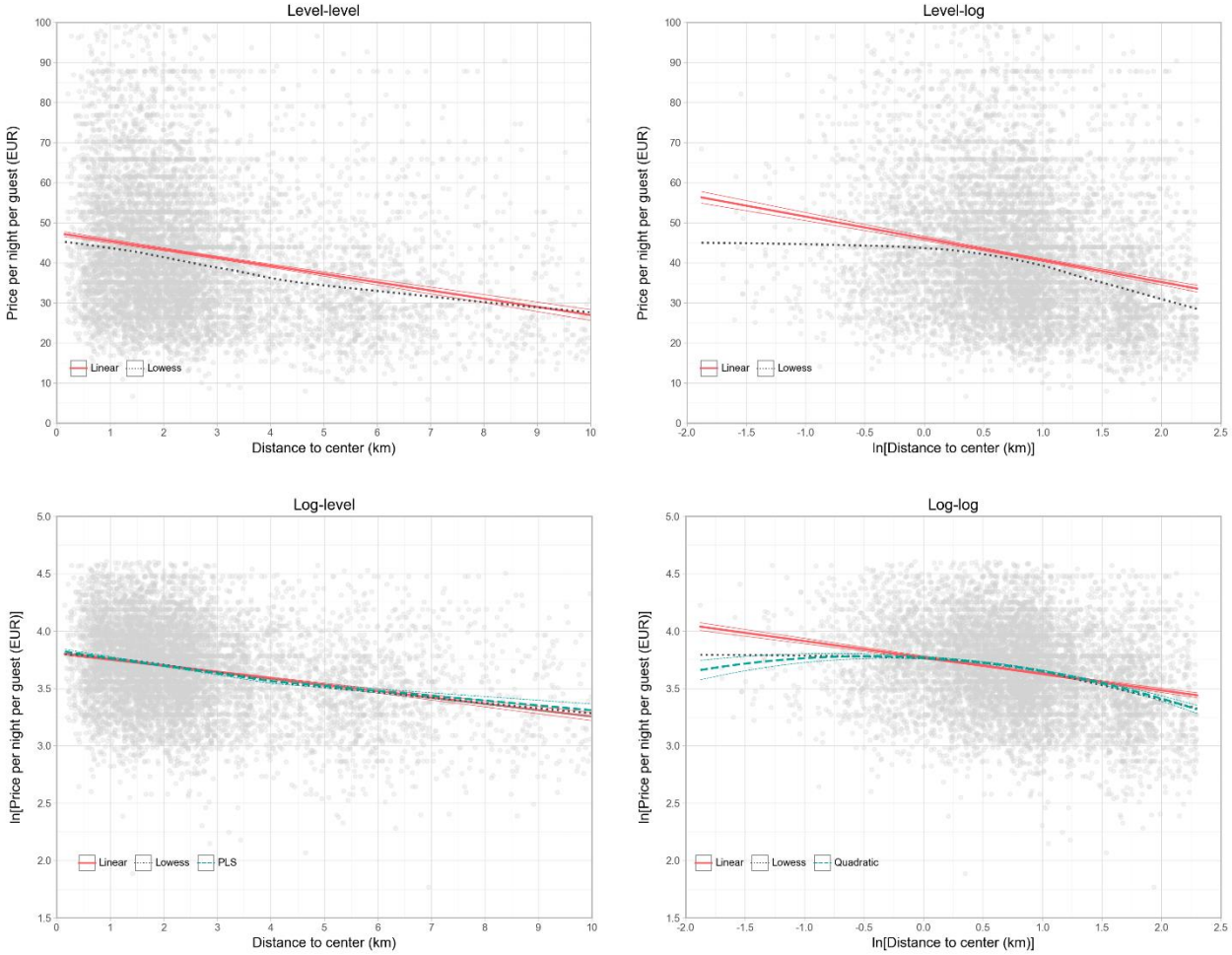
The linear level-level model suggests that listings with a 1 km higher distance to the center are on average 2 EUR cheaper per night per guest. The linear level-log model says that listings with a 10 percent higher distance to the center are on average 0.5 EUR cheaper per night per guest. Comparing the R^2 values, we can see from these two models, that the first one has a slightly larger explanatory power. This might be no surprise, as we have seen that for set-ups where distance is log-transformed, we can see a pattern resembling a quadratic one, so naturally fitting a simple line cannot capture that fully (this is also well illustrated in the level-log panel of Figure A9).

The linear log-level model suggests that listings with a 1 km higher distance to the center are on average 5.5 percent cheaper per night per guest. The linear log-log model says that listings with a 10 percent higher distance to the center are on average 1.4 percent cheaper per night per guest. The PLS log-level model suggests that for AirBnBs closer than 4 km to the center, listings that are 1 km further away are on average 6.7 percent less expensive per night per guest. For places that are at least 4 km away, the model suggests that observations with 1 km higher distance from the center have a 4.2 percent lower price per night per guest on average. Lastly, the quadratic log-log model indicates that listings with a 1 percent higher distance to the center are on average 0.056+0.122[ln(distance to center)] cheaper or more expensive per night per guest. For the average ln(distance to center), 0.75, this means a 0.15 percent lower price per night per guest.

We can also compare the R^2 values for models where the price per night per guest is log-transformed. As we can see, the best performing models are the PLS log-level and the quadratic log-log with the same R^2. I decided to work with the PLS model as it provides only a marginally worse fit than my best fitting quadratic model, while also being more straightforward to interpret. As I mentioned in Section 3.3., distances are rather hard to think about in percentage terms, so

keeping them at level can provide more understandable results.

Figure A9: Graphical comparison of the estimated regressions and lowess estimates (N=8,080)

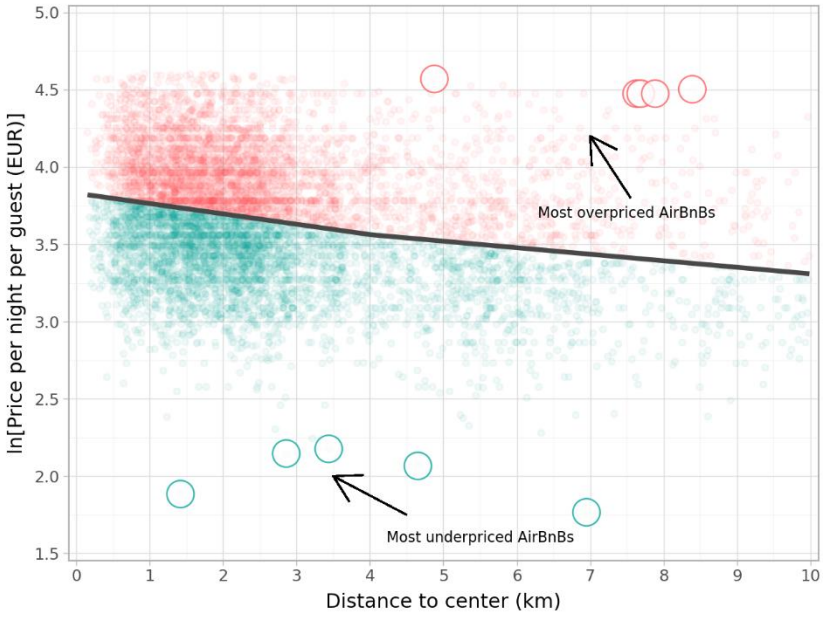


Note: 95% confidence intervals for the predicted values are indicated by narrower, same color lines. A confidence interval for the lowess estimates is not provided.

A6. Deal selection

Figure A10 provides a graphical illustration of the most under- and overpriced listings relative to our model. The figure highlights the top and bottom 5 listings with the highest and lowest model residuals. Note that as price is log-transformed, we can identify pricing in relative terms: that is the least and most percentage differences between the actual and the predicted price levels.

Figure A10: Graphical representation of under- and overpriced AirBnBs relative to our model (N=8,080)



A7. Robustness of results regarding missing values

To uncover whether dropping observations where host acceptance rate is missing would introduce some distortions to our model, I calculated models with both dropping and including these observations. Then, I compared the parameter estimates using t-tests. For calculating the standard errors for the parameter differences of the models, I used bootstrapping. I first resampled the unfiltered dataset with replacement 2,500 times, and then filtered the bootstrap samples for both cases. I fitted the appropriate regression lines to the filtered bootstrap samples and extracted the differences in each parameter between the two regression models. The standard deviation of these 2,500 bootstrap parameter differences became my bootstrap estimate for the SE of the difference in parameters: $SE[\beta_{i,\ with\ missing} - \beta_{i,\ without\ missing}]$.

Then, I calculated the relevant t-statistic for each parameter using the formula[7]:

$$t_i = \frac{\beta_{i,\ with\ missing} - \beta_{i,\ without\ missing}}{SE[\beta_{i,\ with\ missing} - \beta_{i,\ without\ missing}]}$$

The results are summarized in Table A7 below. For the interpretation, I used the conventional 5% significance level.

Table A7: Comparison of PLS model parameters for models excluding and including missing values in host acceptance rate

Parameter	PLS including missing values	PLS excluding missing values	t-stat. for difference in parameter	Interpretation
Constant	3.831	3.830	.228	No statistically significant difference.
Distance to center < 4	-.067	-.075	4.009	Statistically significant difference.

[7] Formally, I am testing the null of the difference between the parameters being zero, against the alternative of the difference not being zero.

Parameter	PLS including missing values	PLS excluding missing values	t-stat. for difference in parameter	Interpretation
Distance to center ≥ 4	-.042	-.046	1.506	No statistically significant difference.

From the results, we can see that the models are significantly different in their slope for the first spline. Thus, excluding observations with missing values in the host acceptance rate would have indeed distorted our results.