Explaining the prices of Airbnb listings in Amsterdam

From a linear model perspective

Deniz Sen
s1486438

Robin Labrujere
s1504843

Stephan van der Putten
s1528459

January 19, 2020

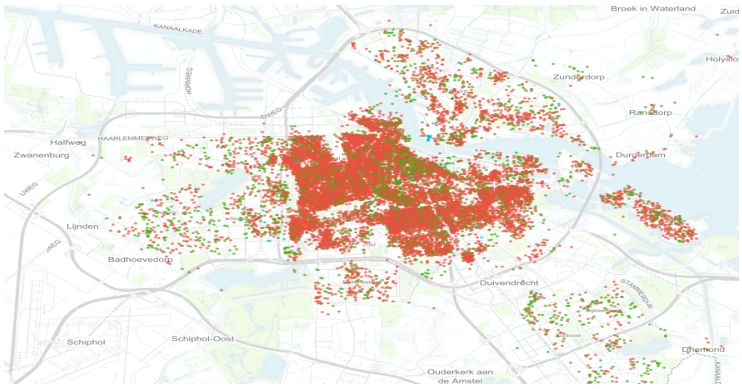


Figure 1: Listings in Amsterdam. The colour Red represent homes or apartments, green are private rooms and blue are shared rooms[1]

1 Introduction

In recent years there has been a surge of online peer-to-peer marketplaces. The most prominent examples are Uber, Fiver and Airbnb. These businesses aim to match sellers who are willing to share underutilized goods or services with buyers who need them. Airbnb is a primary example of such a marketplace often labelled a "shared-economy", since it connects hosts who have spare rooms with guests who require an accommodation. Founded in 2008, it now has more than two million listings in over 191 countries and an approximate number of 60 million guests who have used this service. Airbnb has grown into one of the most successful start-ups, challenging the hotel industry. With the aim of understanding the Airbnb marketplace, Airbnb listing data from 2019 in Amsterdam was examined. Since a key principle of such "shared economies" has been theorised to be sociability, the listing data was examined to determine whether this is true. That is, do host characteristics and social indicators in the Airbnb data explain housing prices or are housing prices solely explained by accommodation characteristics and location?

```
## # A tibble: 17,600 x 18
##    price neighbourhood_c… bedrooms review_scores_r… host_response_t…
##    <int> <fct>            <fct>               <int> <fct>
##  1    59 Oost             1                      98 within an hour
##  2    80 Centrum          1                      88 within a few ho…
##  3   125 Centrum          1                     100 within an hour
##  4   155 Centrum          1                      99 within an hour
##  5    75 Centrum          1                      97 within an hour
##  6    55 Centrum          1                      95 within an hour
##  7   219 Zuid             3                      95 N/A
##  8   159 Centrum          1                      98 within an hour
##  9   210 Centrum          2                      97 N/A
## 10   100 Zuid             1                      80 within a few ho…
## # … with 17,590 more rows, and 13 more variables:
## #   host_identity_verified <lgl>, property_type <fct>, room_type <fct>,
## #   accommodates <fct>, bathrooms <fct>, beds <fct>, bed_type <fct>,
## #   security_deposit <int>, cleaning_fee <int>, guests_included <fct>,
## #   number_of_reviews <int>, cancellation_policy <fct>,
## #   reviews_per_month <dbl>
```

Figure 2: Example of the dataset

2 Dataset & Problem

This section will take a closer look at the dataset. Furthermore we introduce our research question.

The dataset was scraped from insideAirbnb.com[1]. This website provides data from the Airbnb website for listings available in cities all over the world. The data is webscraped on a monthly basis to provide up-to-date data on the listings in various cities. For our research we used the dataset from 2019 November in Amsterdam, which was scraped by insideAirbnb in December 2019.

This raw dataset consists of 20239 observations and 106 variables, which are all entries with reviews from customers. The dataset includes information about the hosts, the listings, the location of each listing, and the review from customers. Although the magnitude of this dataset allows for analyses on a wide variety of topics, a smaller subset of variables was selected for the analysis. This meant, that redundant variables, e.g. 'id' were removed. After removal 18 variables were left in the dataset, which were subsequently used for analysis. NA's were removed. An example of our dataset is shown in Figure 2, which is the output in R.

To answer out research question, price was used as the response, see Figure 2. The remaining 17 variables are explanatory variables. A linear model was used to examine the relationship between price and how the different explanatory variables influence price as outcome. In short, we examine whether social factors can be used to predict price or whether they are mostly determined by location and accommodation characteristics. A linear model was used to approach this problem.

Our research question is as follows: **Research question**: Do social indicators have an impact on Amsterdam listing prices in Airbnb's social marketplace or are property characteristics more important?

[1] www.insideAirbnb.com

3 Approach

This section introduces the approach for dataset. We perform the research in R[2] that provides built-in functionality for linear regression and statistical methods.

A short overview of our steps through the dataset:

1. Preprocessing
2. Data Exploration
3. Explanatory Variable Selection
4. Model Selection

3.1 Preprocessing

Our first step is preprocessing. We disregard discussing the technical details as it is not the focus of this research and our examples in following sections are clear enough. The total scraped original dataset includes 20239 observations ($n = 20239$), 106 variables ($p = 106$). In order to address the research questions, we need to reduce the dataset to a subset. The subset should include host-characteristics as social indicator, accommodations characteristics and overall user rating. We reduced the neighbourhoods to include more general neighbourhoods, instead of specific neighbourhoods places. Missing data was removed and variables that did not relate to the research questions were excluded. The dataset used for the analysis included 18 variables ($p = 18$) and 17600 observations ($n = 17600$).

The dataset contained a lot of factorial variables, which often contained a large number of levels. We reduced the number of levels by fusing several levels into a larger group. For example, the variable neighbourhood initially had 22 groups with all neighbourhoods, which we reduced to 7 by grouping them in "Centrum", "Nieuw West", "Noord", "Oost", "West", "Zuid", and "Zuid Oost. Additionally, we deleted groups of factors which contained < 10 observations. Finally, several numerical variables ,like number of beds, were made into factorials because they contained a lot of repetitions on a small amount of real numbers.

[2] https://www.r-project.org/

3.2 Data exploration

Linear regression usually begins by assuming three fundamental assumptions about the errors that need to be examined. The assumption are:

1. normality
2. common error variance
3. linearity

These assumption can be compactly represented as followed:

$$\epsilon \sim N_n\left(0, \sigma_e^2 I_n\right) \tag{1}$$

We do speak of a fourth assumptions that is independence. Independence relates to the method of data collection and experimental design. For this dataset there is no reason to assume otherwise, so we assume independence and do not research this further.

Furthermore we check for collinearity between the response variables, as this would lead to coefficients not correctly being represented and also being redundant. They explain the same variance, and thus no valid results can be given for any of the individual predictors.

3.2.1 Normality of Error

In order to examine whether the residuals of the linear model are normally distributed we use a quantile comparison plot (QQ plot). In a QQ plot the theoretical quantile is on the x-axis and the residuals are on the y-axis. In Figure 3 our QQ plot is shown. If the errors follow a theoretical normal distribution, the dots in the graph would align on the straight line that is drawn from the lower left to the upper right corner.

Visibly the errors had a slight positive skew, which is further visualised in Figure 4a, furthermore revealing long tails on both sides. After applying a log10 transformation to the price as seen in Figure 4b, the distribution was no longer skewed and follows a normal distribution more closely. However, tails remain long on both sides. These are of smaller importance, as in heavy tailed distributions the mean is still a proper estimate of the values.

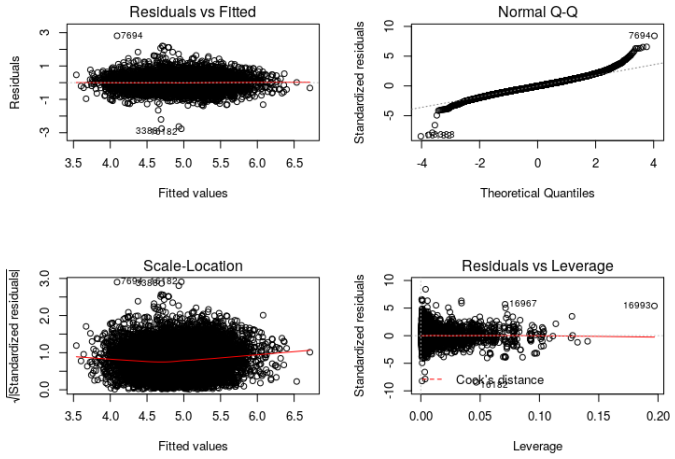
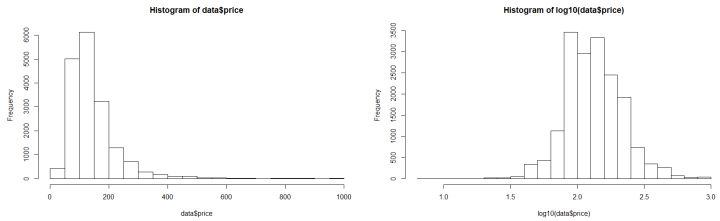


Figure 3: Four plots about residual assumptions.



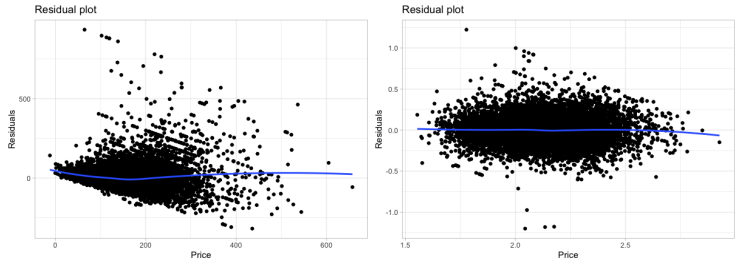
(a) (b)

Figure 4: (A) The original price skewness, (B) after the log10

3.2.2 Constant Error Variance

In order to check for constant error variance, fitted values were plotted against the residuals, see Figure 5. Under constant variances, the error variance is expected to be the same regardless of the X value. That is variance should not decreases or increase when moving up or down the values of X and an equal spread without a pattern should be visible.

In our original data, the points are cone shaped, as shown in Figure 5a, which indicates that the error variance increases with the value of the X's. This problem often occurs when data has a lower bound, as with a price, and can be fixed by stretching the data with transformation. After log10 transformation of price, the data points are now heterogeneously distributed over the values of X's, see Figure 5b.

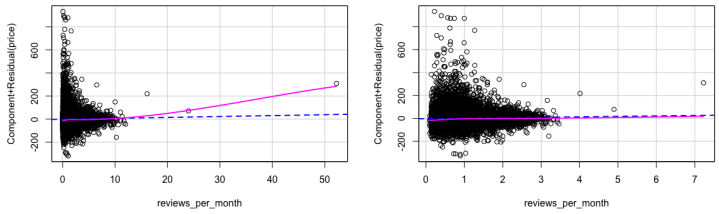


(a) (b)

Figure 5: (A) The original spread of errors, (B) after the log10

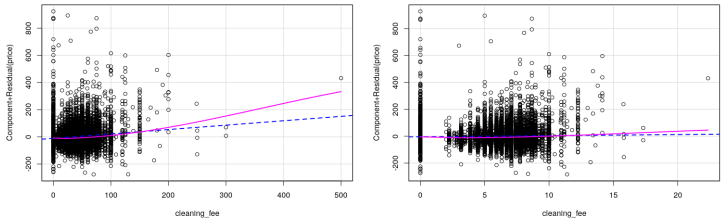
3.2.3 Linearity

In order to examine linearity, component plus residual plots were created for each response variable in Figure 6 and Figure 7. Component plus residual plots effectively examine the partial relationship of the outcome variable and one regressor controlling for all other regressors. Variables that showed a simple monotone nonlinear relationship were transformed according the Tukey and Mosteller's bulging rule to linearize the relationship [1]. Only two variables had a non-linear relationship with the outcome variable, namely "reviews per month" and "cleaning fee. After taking the square root both relationships with the outcome were linearized.



(a) (b)

Figure 6: (A) shows the partial relationship between price and reviews per month before transformation, (B) after a square root transformation



(a) (b)

Figure 7: (A) shows the partial relationship between price and cleaning fee before transformation, (B) after a square root transformation

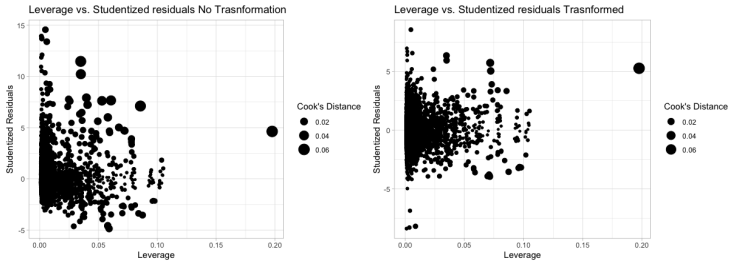
3.2.4 Influential Data

A common method of assessing influential data points is by using the Cook's distance, which is defined as followed [1]:

$$D_i = \frac{E^{\text{`}2}}{k+1} \times \frac{h_i}{1-h_i} \tag{2}$$

The first term is a measure of discrepancy and the second term a measure of leverage. Visibly, Cook's distance is a function of $discrepancy \times leverage$. An observations with a high studentized residuals (discrepancy), high leverage or both will have a high cooks distance.

The Cook's distance is examined by plotting the leverage against the studentized residuals, as seen in the bottom right of Figure 3. Observations are scaled by their Cook's distance. Observations with a large influence will have a bigger volume. The bubble plot in which the studentized residuals are plotted against the leverage can be seen in Figure 8. Plot (A), before correcting for linear model assumption, shows that a number of observations have extremely high studentized residuals > 10 on the standardized scale. Leverage values are however small, resulting in a small Cook's distance (small size). Furthermore, after correcting for model assumptions, the range of studentized residuals becomes smaller, while leverage values remain small. There is no indications for influential points.



(a) (b)

Figure 8: (A) Studentized residuals vs, (B) after the log10

3.2.5 VIF

The Variance Inflation Factor (VIF) is used to detect collinearity between regressors and is defined below as [1]:

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{3}$$

Collinearity between certain regressors in a linear model is a cause of increased sampling variances of the least-squares coefficients, until a point where the least-squares estimates are impaired. The VIF indicates the impact of collinearity on the precision of the estimate. More precisely, the square root of the VIF shows how much more the standard error increases compared to when no collinearity would occur at all.

If certain regressors are above the given threshold of 10 these regressors show collinearity. The variable with the highest VIF value is typically removed and the VIF is recalculated and the steps are repeated until the threshold is no longer exceeded.

The VIF before and after removal of the highest VIF can be examined in Table 1. It is visible that beds had the highest VIF of 42.30. We would expect that the collinearity in the variables, accommodates, bedrooms and beds to be collinear. One would be able to derive that big accommodates contain more beds and beds and vise versa. After removal of beds all VIFs are below 10, therefore collinearity is corrected for.

3.2.6 ANOVA

Analysis of variance (ANOVA) states the regression sum of squares, mean square error, F score and P score for F. In our research we are focused on the F value. If we look at Table 2, we see $F_{n,resdf} = F : 0.01$ with $Pr(> F) = 0.93$ for host identity verified and removed it based on this F statistic not being significant to our model. Furthermore almost all host identities are verified which gives the variable little added value. After we remove the variable we calculate the ANOVA. In Table 3 we see that all variables are relevant to our model.

Table 1: Variance inflation factor, before and after correction for collinearity

	VIF	New VIF
host_response_time	1.37	1.37
host_identity_verified	1.07	1.07
neighbourhood_cleansed	1.36	1.35
property_type	6.11	5.9
room_type	4.39	4.38
accommodates	**33.10**	8.80
bathrooms	1.66	1.62
bedrooms	**14.72**	7.98
beds	**42.30**	–
bed_type	1.03	1.03
security_deposit	1.19	1.19
cleaning_fee	1.48	1.48
guests_included	2.16	2.07
number_of_reviews	1.95	1.94
review_scores_rating	1.05	1.05
cancellation_policy	1.24	1.23
reviews_per_month	2.30	2.30

Table 2: Anova table, with 1 nonsignificant value in bold

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
host_response_time	4	62.18	15.55	138.14	0.0000
host_identity_verified	1	**0.00**	**0.00**	**0.01**	**0.9268**
neighbourhood_cleansed	6	285.61	47.60	422.97	0.0000
property_type	18	161.05	8.95	79.50	0.0000
room_type	3	759.46	253.15	2249.46	0.0000
accommodates	8	669.66	83.71	743.81	0.0000
bathrooms	7	53.05	7.58	67.34	0.0000
bedrooms	6	52.32	8.72	77.48	0.0000
bed_type	3	2.67	0.89	7.91	0.0000
security_deposit	1	11.87	11.87	105.49	0.0000
cleaning_fee	1	23.85	23.85	211.92	0.0000
guests_included	6	25.38	4.23	37.59	0.0000
number_of_reviews	1	8.29	8.29	73.64	0.0000
review_scores_rating	1	21.97	21.97	195.20	0.0000
cancellation_policy	4	2.99	0.75	6.64	0.0000
reviews_per_month	1	0.56	0.56	4.97	0.0257
Residuals	17528	1972.59	0.11		

3.2.7 Occam's razor

The theory of Occam's razor or law of parsimony was applied in model selection, which states that no more assumptions should be made then are necessary. More

Table 3: Anova table, all values are within the F score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
host_response_time	4	62.18	15.55	138.15	0.0000
neighbourhood_cleansed	6	285.55	47.59	422.92	0.0000
property_type	18	161.09	8.95	79.53	0.0000
room_type	3	758.94	252.98	2248.06	0.0000
accommodates	8	670.16	83.77	744.40	0.0000
bathrooms	7	53.07	7.58	67.37	0.0000
bedrooms	6	52.33	8.72	77.50	0.0000
bed_type	3	2.67	0.89	7.91	0.0000
security_deposit	1	11.86	11.86	105.40	0.0000
cleaning_fee	1	23.86	23.86	212.03	0.0000
guests_included	6	25.38	4.23	37.60	0.0000
number_of_reviews	1	8.12	8.12	72.13	0.0000
review_scores_rating	1	22.14	22.14	196.71	0.0000
cancellation_policy	4	2.99	0.75	6.64	0.0000
reviews_per_month	1	0.56	0.56	4.96	0.0259
Residuals	17529	1972.59	0.11		

simply put, more obvious outcomes are preferred over more complex ones if there is no clear substantiation to adopt the complex ones [1]. With this principle in mind we continue with our next step.

3.3 Variable Selection

After dummy coding the ANOVA model included 80 regressors. Due to this high number of variables included in the model, it is likely that a subset of these 80 variables will suffice. In order to examine this, forward subset selection was implemented. Forward eliminations starts with a null model, which includes one variable and then adds one variable in turn. In each turn goodness of fit statistics are then computed and based on the score model selection is made. Specifically for this study, we focused on the Bayesian information criterion (BIC) to examine whether there is a subset of variables that can be used to represent the number of variables included in the ANOVA.

Bayesian information criterion (BIC), unlike the R^2 smaller values indicate a better-fitting model. Furthermore, unlike the R^2 which never declines when another variable is added, the BIC penalises the fit for the number of parameters included in the model As such the BIC prefers parsimonious, smaller models [2].

In our Figure 9, a converges starts around 30 variables.

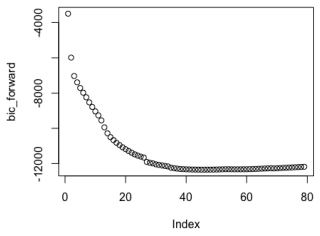


Figure 9: Forward selection showing the BIC as the number of variables increases

4 Results

The final results included 80 variables after dummy coding for categorical variables. This section describes these results, which are shown in Table 4, 5. The variables are stated in italic this section.

Under host characteristics we gathered the variables *host identity verified* and host response time. *Host identity verified* was removed after running an ANOVA on the preliminary model. *Host response time* was examined, and did not show any significant differences between the groups ($p > 0.05$). The other variables in our model are shared under accommodation characteristics, and are shortly discussed below.

Property type had Apartment/Hotel as the reference group. All property types are significant ($p < 0.05$), except the serviced apartment and boutique hotel, indicating that the prices of Apartment/Hotel, boutique hotel and serviced apartment are similar. *Accommodates* had 1 accommodate as the reference group, and having more than 1 accommodate has a significant difference in the price outcome ($p < 0.001$). *Bathrooms* had zero bathrooms as the reference group. No difference between the zero bathrooms and a shared bathroom was found, but from 1 bathroom on the price is significantly different from the reference ($p < 0.005$). *Bedrooms* had zero bedrooms as the reference group. Bedrooms between 1 and 5 were significantly different ($p < 0.001$), but the group that had more than 5 bedrooms was not. *Security deposit* and *cleaning fee* are both continuous variables, and are both significant ($p < 0.001$). *Guests included* had one guest as the reference group. Significant differences were found between 1 and 2, 1 and 4, 1 and 6 and 1 and > 6 ($p < 0.001$), while no significant differences were found between 1 and 3 and between 1 and 5 ($p > 0.05$). *Cancellation policy* had a flexible policy as the reference group. Moderate and Strict with a 14 day grace period both differ significantly from the reference group ($p < 0.001$), while the two groups of super strict policies did not show significant differences ($p > 0.05$) *Number of reviews*, *review score rating* and *reviews per month* were all significant ($p < 0.05$).

The location of the listings was analysed using the variable *neighbourhood*, which was divided into 7 groups. All of the groups were significantly different from the reference group "Centrum" regarding prices ($p < 0.001$)

Table 4: Summary output of the fitted model (part 1)

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8513	0.0588	31.46	0.0000
host_response_timeN/A	-0.0281	0.0091	-3.10	0.0020
host_response_timewithin a day	0.0158	0.0094	1.68	0.0927
host_response_timewithin a few hours	0.0114	0.0094	1.22	0.2233
host_response_timewithin an hour	0.0089	0.0092	0.96	0.3354
neighbourhood_cleansedNieuw_West	-0.1166	0.0042	-28.06	0.0000
neighbourhood_cleansedNoord	-0.1791	0.0054	-33.09	0.0000
neighbourhood_cleansedOost	-0.1246	0.0038	-33.15	0.0000
neighbourhood_cleansedWest	-0.1052	0.0034	-30.55	0.0000
neighbourhood_cleansedZuid	-0.0729	0.0036	-20.54	0.0000
neighbourhood_cleansedZuid_Oost	-0.2283	0.0089	-25.79	0.0000
property_typeApartment	-0.1950	0.0223	-8.75	0.0000
property_typeBed and breakfast	-0.1534	0.0205	-7.47	0.0000
property_typeBoat	-0.1302	0.0238	-5.48	0.0000
property_typeBoutique hotel	-0.0433	0.0320	-1.35	0.1759
property_typeCabin	-0.1533	0.0494	-3.10	0.0019
property_typeCondominium	-0.1950	0.0237	-8.21	0.0000
property_typeGuest suite	-0.1480	0.0256	-5.77	0.0000
property_typeGuesthouse	-0.1110	0.0308	-3.60	0.0003
property_typeHostel	-0.1124	0.0435	-2.59	0.0097
property_typeHotel	0.2133	0.0444	4.81	0.0000
property_typeHouse	-0.2027	0.0226	-8.97	0.0000
property_typeHouseboat	-0.0944	0.0239	-3.94	0.0001
property_typeLoft	-0.1188	0.0236	-5.03	0.0000
property_typeOther	-0.2137	0.0367	-5.82	0.0000
property_typeServiced apartment	-0.0140	0.0276	-0.51	0.6110
property_typeTiny house	-0.2138	0.0512	-4.17	0.0000
property_typeTownhouse	-0.1933	0.0231	-8.38	0.0000
property_typeVilla	-0.1651	0.0344	-4.80	0.0000
room_typeHotel room	-0.0900	0.0128	-7.01	0.0000
room_typePrivate room	-0.1362	0.0035	-38.67	0.0000
room_typeShared room	-0.2281	0.0231	-9.86	0.0000
accommodates2	0.1320	0.0083	15.96	0.0000
accommodates3	0.1667	0.0093	17.92	0.0000
accommodates4	0.2393	0.0091	26.43	0.0000
accommodates5	0.2507	0.0126	19.92	0.0000
accommodates6	0.3087	0.0128	24.19	0.0000
accommodates7	0.3382	0.0262	12.88	0.0000
accommodates8	0.3478	0.0233	14.92	0.0000
accommodatesbig	0.4590	0.0282	16.29	0.0000

5 Discussion

This section will discuss points for improvement of our research and future work.

Table 5: Summary output of the fitted model (part 2)

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8513	0.0588	31.46	0.0000
bathrooms0.5	0.0079	0.0356	0.22	0.8253
bathrooms1	0.0987	0.0310	3.19	0.0014
bathrooms1.5	0.1158	0.0311	3.73	0.0002
bathrooms2	0.1484	0.0315	4.71	0.0000
bathrooms2.5	0.1733	0.0327	5.30	0.0000
bathrooms3	0.1529	0.0354	4.31	0.0000
bathroomsbig	0.1619	0.0380	4.26	0.0000
bedrooms1	0.0348	0.0050	6.91	0.0000
bedrooms2	0.0817	0.0060	13.55	0.0000
bedrooms3	0.1013	0.0074	13.67	0.0000
bedrooms4	0.1350	0.0104	12.95	0.0000
bedrooms5	0.1269	0.0209	6.08	0.0000
bedroomsbig	0.0806	0.0366	2.20	0.0275
bed_typeFuton	-0.0535	0.0475	-1.13	0.2603
bed_typePull-out Sofa	-0.0536	0.0429	-1.25	0.2121
bed_typeReal Bed	0.0003	0.0406	0.01	0.9950
security_deposit	0.0000	0.0000	5.07	0.0000
cleaning_fee	0.0007	0.0001	13.94	0.0000
guests_included2	-0.0252	0.0027	-9.24	0.0000
guests_included3	-0.0020	0.0087	-0.23	0.8160
guests_included4	0.0444	0.0057	7.74	0.0000
guests_included5	0.0375	0.0192	1.95	0.0514
guests_included6	0.1021	0.0216	4.72	0.0000
guests_includedbig	0.1014	0.0262	3.87	0.0001
number_of_reviews	-0.0003	0.0000	-8.89	0.0000
review_scores_rating	0.0024	0.0002	14.07	0.0000
cancellation_policymoderate	0.0083	0.0031	2.66	0.0078
cancellation_policystrict_14_with_grace_period	0.0158	0.0031	5.03	0.0000
cancellation_policysuper_strict_30	-0.0135	0.0357	-0.38	0.7045
cancellation_policysuper_strict_60	0.0296	0.0298	0.99	0.3198
reviews_per_month	0.0026	0.0012	2.23	0.0259

In our preprocessing step we removed a lot of variables based on containing NA's. A variable that intuitively would be interesting to research for linear models is square feet. Given that the dataset actually contains enough observations, one could also include specifically all the observations that have square feet values and still be left with probably hundreds or thousands of observations. Enough to perform some qualitative research instead of quantitative research. As we only used the month November, including more months would make all variables possible to observe. We would ,however, expect this data aggregation step to introduce unknown problems. In short, to keep this research focused on exploring linear models we chose to simply exclude NA's and those variables that are incomplete.

We performed an anova on our preliminary model, and excluded one variable based on the results of this anova. The proper method would be to first perform model selection, and follow this up by an anova. This variable was nonsignificant in the full model based on summary and anova. We could argue that this variable would either be excluded by the model selection, or it would be excluded after because after the model selection it would still be nonsignificant.

5.1 Future work

A number of limitation should be discussed with respect to the provided study. A comparison between cities was not made. Whether our results could be generalised to other cities remains to be investigated. For it would be able to examine potential reasons for differences between cities. It would be possible to research the effects of different cultures, wealth, religion, and many more variables on the prices of Airbnb listings worldwide.

In this study we focused on explaining as much information as possible, and in doing so we included quite a lot of variables. Most of these variables also showed to be of importance in the model. This could be due to the large size of the data set. Even though we determined that around 30 out of 80 variables should be enough to predict listing prices using forward search, whether these 30 variables explain the highest variance among the 80 variables was not examine. Due to the computational limitation, forward selection was implemented to determine this subset, which has a number of drawbacks. In forward selection we start with a small model and built it up. Features that were once selected or eliminated cannot be later discarded re-selected. As such not all possible combinations of the 80 variables were examined. Exhaustive subset selection would be advised in future work. We suggest to run the analysis on a server cluster with parallel processing on multiple cores. An alternative approach to find a subset of variables to avoid the computational load is a Principal Component Analysis (PCA).

Furthermore, in this study we focused on the data from November 2019. The influence of holiday season, special events or other factors on prices was therefore not included. For example during vacations hosts could increase their price, as the demand for rooms would be higher than outside of vacation periods. By including more months we would be able to get a better generalisation of the data. Moreover, by taking data from each month in 2019, we could include the time of the year in predicting the prices. A longitudinal linear model could examine this.

6 Conclusion

Linear regression usually starts by assuming three fundamental assumptions about the errors distribution. The assumption are 1) normality, 2) common error variance and 3) linearity. Originally the errors did not follow a normal distribution, had uncommon error variance and a number of aggressors had a

slight curvilinear relationship with the outcome variable price. These assumptions have been checked and corrected for via power transformation. To correct for normality and non-constant error variance a log transformation was applied to the outcome variable price. Component-plus residual plots were examined for potential non-linear relationship between outcome and regressors. Two variables were found to have a non-linear relationship and the appropriate square root transformation as suggested by Tukey and Mostellers Bulging Rule was applied to linearize it. With respect to collinearity, the VIF was computed. Beds and number of accommodates and bedrooms showed a high VIF (VIF $>$ 10). This is to be expected, since the number of accommodates should reflect the number of beds available in the listing. Beds was excluded from the model to correct for presence of collinearity as it had the highest VIF. Studentized Residuals were plotted against leverage value scaled by their Cook's distance. After transformation no influential observations marketed further examination. No influential cases were present.

The final model included 80 variables after dummy coding for categorical variables. With respect to our research question, we observe that social indicators have no impact on price prediction. So regardless of social behaviour such as communication and host identity, price is solely determined by location and listing characteristics.

Contributions of the team members

All members contributed equally to the paper and were equally involved in the creation of this research.

References

[1] John Fox. *Applied regression analysis and generalized linear models.* Sage Publications, 2015.

[2] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.