Predictive Analysis of Price on Amsterdam Airbnb Listings Using Ordinary Least Squares

Marietta Papagrigoriou, Efklidis Katsaros and Michail Tsiaousis

Leiden University, The Netherlands

***Abstract*— Nowadays, rapid advancements in technology facilitated traveling around the globe, in the frame of which, rental websites emerged as intercessors between house owners and guests. Airbnb renders the most prominent one with more than 200 millions of users over 191 countries worldwide. In this paper, we use data compiled on April, 2017 specifically, Airbnb accommodation listings for Amsterdam. In the process of exploratory statistical analysis, we discard many variables through inferring, testing and validating interchangeably and we build a Linear Regression model to express the relationship between expected price and a number of predictors for an overnight occupation of a registered lodge. For this purpose, we utilize fundamental tools of Linear modeling and Algebra. After measuring our models in terms of various metrics using 10-fold Cross Validation, we opt for the robust in terms of outliers, Mean Absolute Error (MAE) and the model that predicts the price more accurately according to this one. Consequently, we investigate on the assumptions of linear models which were not satisfied. Thus, the estimators would be biased, not efficient and standard errors would be inflated, so response was rescaled using the logarithm transformation which settles the model in terms of constant variance, linearity and error's normality and allows for reliable predictions.**

I. INTRODUCTION

Airbnb is an American enterprise maintaining a lodging platform, enabling users to rent or share their properties in terms of a reliable, review-based, and active community. Using self-definition extracted from the website[1]; ”Founded in August of 2008 and based in San Francisco, California, Airbnb is a trusted community marketplace for people to list, discover, and book unique accommodations around the world online or from a mobile phone or tablet.” Users are benefited bidirectionally; evacuated lodges are rented to increase owners income, whereas renters grasp the opportunity to reside in cozy and beneficial, from a value-for-money perspective, accommodations.

Exact origins of the data set is the *Inside Airbnb*[2] website, which is sourced from publicly available information from the Airbnb website, on an -approximately- monthly basis. Concerning the project they work on, authors and analysts define it as ”an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.” They state that, ”By analyzing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so you can see how Airbnb is being used to compete with the residential housing market.”

Due to controversies and the public debate on the way that ”sharing economy” affects contemporary cities and their lifestyle, there are a few cleaned data versions for which we did not opt, as the goal was to gain hands-on-experience in each and every section of the analysis extracting and filtering the bulk of available information ourselves.

II. DATA WRANGLING

The specific dataset was compiled on April, 2017 and consists of 15181 observations and 95 variables, which are accessible to everyone using Airbnb's online platform. Our goal is to build a model so as to predict the price of an accommodation for an overnight occupation. Many variables are dropped out during the initial data wrangling procedure; the aim is to retain the ones that are able to provide information for the modeling process, by means of explaining proportion of price variance given certain predictors. The set of these explanatory variables forms the linear predictor function, whose parameters estimation is to formulate a Linear Regression modeling approach, taking into consideration that the dependent variable is a positive scalar one.

Subsequently, data filtering is the first phase of the analysis; most variables contain irrelevant information such as text chunks of `summary`, `description`, `id`, `country`, and many more columns carrying information for the host, the neighborhood etc. It is apparent that this kind of variables is not the one to contribute for this analysis. Instead, clearly specified qualitative and quantitative ones, answering to intuitive questions, should be the ones concerning prospective researchers. According to the latter criterion, we proceeded to an initial examination of 13 potential regressors, on which, 12811 observations would be utilized to model the expected value in terms of euro currency units, for an overnight occupancy of a listed accommodation.

However, each and every modeling orientated, statistical analysis that is to be considered, has to carefully assess the number of missing values, as well as the columns in which they appear. For this data set, there are only a few missing

The following analysis has been conducted in the framework of the course *Linear, Generalized Linear Models and Linear Algebra*, offered by Leiden University's Mathematical Institute for masters students of Data Science and Statistics using R statistical software. In terms of the aforementioned course, many topics are covered, constructing a strong tool set so as to not only deal with real life problems, but establish a thorough background to elaborate further, delving into the mathematical aspects as well.

[1] `www.airbnb.com`

[2] `www.insideairbnb.com`

TABLE I

FIRST TWO OBSERVATIONS OF THE DATA SET.

neighbourhood	De Baarsjes - Oud-West	De Baarsjes - Oud-West
property_type	Apartment	Bed & Breakfast
room_type	Entire home/apt	Private room
accommodates	6	4
bathrooms	1	1
bedrooms	2	1
beds	6	2
price	145	100
cleaning_fee	65	0
availability_30	14	4
number_of_reviews	4	63
review_scores_rating	95	90
reviews_per_month	0.68	1.39

values in integral variables, allowing for respective observations' omission. Simultaneously, NA entries in reviews' related variables exist, standing for, approximately, 15% of the total number of rows, which are eventually not selected for the subsequent part of the analysis. That is, NA values corresponding to reviews' related variables do not stimulate any further interest on either imputing or predicting them.

The quantitative variables consists of the number of accommodates, reviews, beds, bedrooms, bathrooms, reviews frequency and scores, as well as the availability for the next 30 days and the cleaning fee hosts require. Qualitative are the ones treated as factors in R statistical software, split to levels via the introduction of dummies, which form differences based on measurements between groups. In the case discussed, these are the neighborhood as well as the property and the room type. In Table I, an overview of the two first observations is depicted.

Additionally, variables cleaning_fee and price appear as class factor in R. So, in the initial data set, for the first observation, price would be ”€145”. Subsequently, string processing was necessary before proceeding to the next step. For this purpose, stringr package was utilized. Besides that, cleaning_fee contained plenty of missing values; this occurs due to the fact that when no cleaning fee is required, respective value is missing and, thus, perceived as NA. This would be misleading though, and would deduct from the variable any practical value, so NAs were set to zero. Finally, property_type consisted of 22 levels initially, but we decided to remove the observations belonging to groups that encompassed less than five. Consequently, we ended up with 25 rows less, and the number of levels decreased to just 12.

Attempting to predict Airbnb's nightly accommodation fares in Amsterdam through, one would expect to access straightforwardly informative variables, such as the square meters of a lodge. Although it does exist, 97 percent of rows appear to be missing, not allowing for its utilization. Relevant information could be extracted using the number of accommodates, bedrooms bathrooms and beds. A combination of these could potentially provide us with similar details forming measures in terms of new regressors regarding settlement's premises or even spaciousness. In addition to that, information about the variance of the price is provided in terms of community, expressed by reviews and the availability status for the 30 succeeding days. However, Airbnb's bidirectional review system between owners and guests affects critiques' scaling, as both sides attempt to receive positive feedback for their profiles. Therefore, reviews' distributions are extremely left-skewed. Further elaborated analysis is held in the upcoming sections of the paper. Finally, factorial neighborhood, property and room type are also considered for the initial phase, expected to contribute thoroughly in model's construction.

III. RELATED WORK

Few works have been published regarding price prediction and recommendation of vacation rental websites such as Airbnb. Wang and Nicolau [11] analyze a sample of 180533 accommodation offers in 33 cities, listed in Airbnb's website, using ordinary least squares and quantile regression analysis. Their aim is to investigate the determinants of price through the study of 25 variables that span five general categories, specifically, host attributes, size and property attributes, amenities and services, rental rules, and number of online review ratings.

Gibbs et al. [7] examine the impact of a variety of variables on the price of Airbnb listings, in five metropolitan areas in Canada. They found that physical attributes such as room type and size of accommodation, location, and host characteristics significantly impact the price. Additionally, more reviews were associated with a decrease in response variable.

In [9], the factors that affect the prices of Airbnb and TripAdvisor listings in Los Angeles, London, and Tokyo are studied. The framework of Multi-Scale Affinity Propagation (MSAP) is employed in order to cluster houses based on landmark and house facility information. Then, the Nash Equilibrium filter is used to remove outliers, and finally, Linear Regression with Normal Noise (LRNN) is utilized to predict prices of listings.

A General Linear Model (GLM), and a Geographically Weighted Regression model are applied in [12] in order to identify the key factors that affect Airbnb's listings of business units in Metro Nashville, Tennessee. The authors found that the GWR model, in which the estimated parameters vary across regions to accommodate for potential spatial dependencies [3], [12], outperforms the GLM model in terms of accuracy.

IV. EXPLORATORY DATA ANALYSIS

In the beginning of our analysis, it is essential to visualize the data and especially the variables we are interested in. For the response variable, price, the histogram of Fig. 1 provides us with information regarding its distributional properties. According to the plot, it is apparent that the distribution of price is not normal, but rather skewed to the right, leading to a right heavy tail.

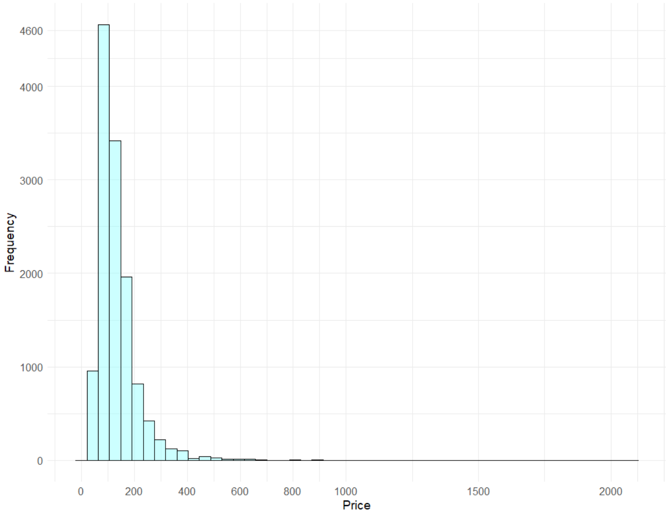

Fig. 1. Distribution of price.

TABLE II
DESCRIPTIVE STATISTICS.

	mean	sd	median	min	max
accommodates	2.91	1.39	2	1	17
bedrooms	1.40	0.88	1	0	10
beds	1.87	1.43	1	1	16
price	132.54	77.88	115	19	2100
cleaning_fee	28.23	22.24	30	0	200
availability_30	2.21	4.26	0	0	30
number_of_reviews	20.90	36.75	9	1	496

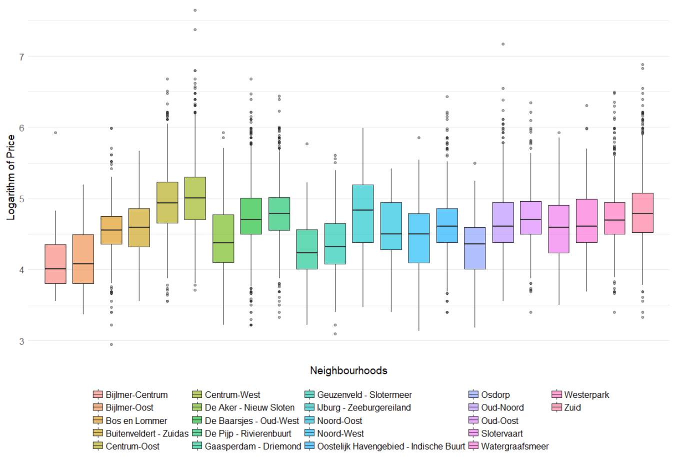

Fig. 2. Logarithm of price, by neighborhood.

TABLE III
MEAN PRICES BY ROOM TYPE.

room_type	price
Entire home/apt	145.66
Private room	87.03
Shared room	89.27

Table II illustrates descriptive statistics for the variables of interest. Statistics such as the mean, median, standard deviation, minimum and maximum values are shown and could proved useful, in addition to the plots, to acquire insights from the data. As shown, the mean value of price is 133€ per night with quite large variability, considering the fact that the lowest price for an airbnb room is 19€ and the highest is 2100€ in a data set of 12811 observations.

Proceeding to the visualization part of the analysis, plots were constructed according to the type of each variable. In order to explore the differences in each level of the categorical ones, we created three boxplots. This kind of plots is informative as they provide a graphical representation of the distribution's summary. More precisely, boxplots visualize the minimum and maximum values, the first, second and third quartile, and finally outliers should any exist. Moreover, the length of the box is a measure of spread of the data, indicating their range. Before presenting the boxplots, it should be remarked that the price is transformed on the logarithm scale in order to compress the data and make them more distinguishable.

Fig. 2 represents side-by-side boxplots for each neighborhood, whereas Fig. 3 illustrates comparative boxplots for property type and room type respectively. We observe that there are differences in prices among different types of property, rooms, and different neighborhoods in the city of Amsterdam. As one would assume, highest prices are accumulated in listings around the city center. For property and room types, guest houses and entire homes/apartments, consist on average, the most expensive choices. It is important to mention that there are many outliers in each plot, indicating large deviation from the expected behavior [2], [10] among the different levels of each qualitative variable.

Table III consists of the mean values of price in each room type. Prices among private and shared rooms do not differ considerably. On the contrary, we observe that price increases dramatically when it comes to Entire home/apt.

It would be convenient to obtain an idea of the relationship between price and the other quantitative variables. A common way to achieve this, is to create scatterplots for each quantitative variable versus the price so as to examine linearity of relationships and check for variability.

Fig. 4 illustrates the scatterplots of the logarithm of price versus physical characteristics, namely, accommodates, bathrooms, bedrooms and beds. There is a positive curvilinear relationship, which is reasonable since we would expect listings with a large number of accommodates, bedrooms, bathrooms, and beds, to be more expensive; still, the derivative of the curve gets smaller reducing its angle, as the value of the independent variable (x-axis) increases. Furthermore, Fig. 5 illustrates the relationship between the logarithm of price and cleaning_fee. As in the case of physical characteristics, the relationship is positive curvilinear, where larger cleaning fees imply that the listings are more expensive. Generally,

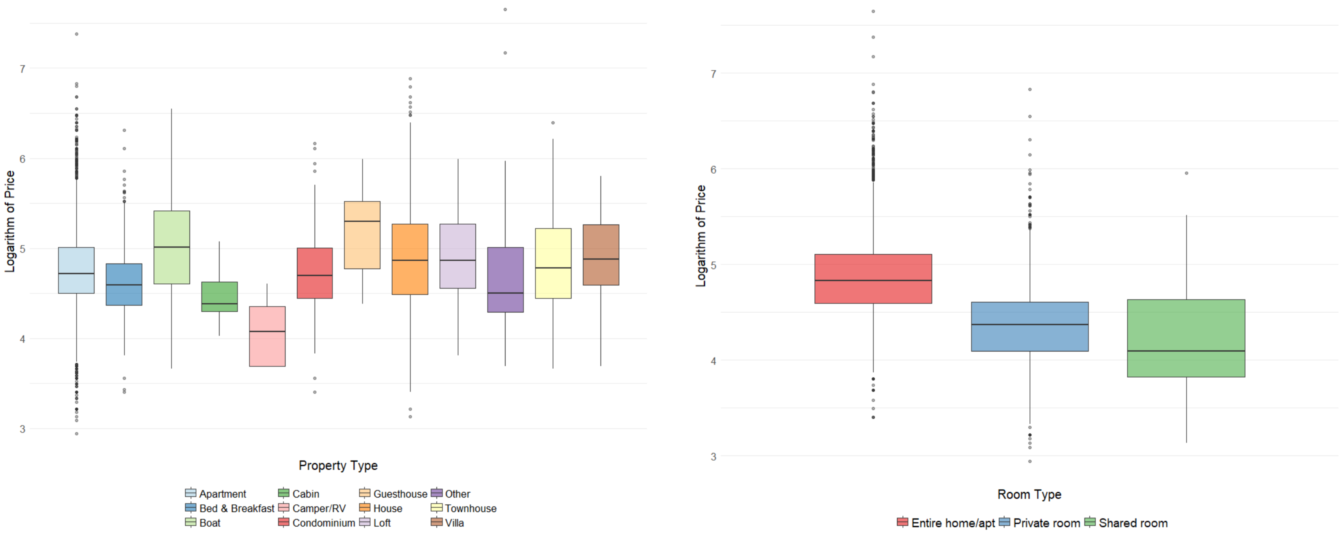

Fig. 3. Left: Logarithm of price, by property type. Right: Logarithm of price, by room type.

TABLE IV
Result of `drop1()` for variable `number_of_reviews`.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
12767	36529060				
12768	36530253	-1	-1193	0.42	0.52

taking into account the previous plots, transformation of the data is a key factor for the analysis. Rescaling the data is helpful to approximate the assumptions of symmetry, linearity and equal variance, guiding us to form an explanatory model.

V. Variable Selection

Moving forward to the the model construction, it is crucial to examine which regressors are important for the model in terms of main effects tests. Using the `drop1` function in R, we will implement partial F-tests comparing the full model, which is the one containing all the regressors, with a different reduced model each time. More specifically, the idea behind `drop1` function (selecting test = F), is that it compares the full model with a reduced model in which the independent variables are removed one at a time, with respect to the principle of marginality. That is, `drop1` conducts F-tests to examine the significance of a regressor, only when the latter does not contain a higher-order term. The null hypothesis is that $H_0 : \beta_i = 0$, indicating no effect of the i^{th} regressor. This procedure yields that all variables are statistically significant, except `number_of_reviews`. Table III illustrates the result of `drop1()` for the latter variable, which is excluded from the model.

Having determined the explanatory variables that will be utilized for the model, it is important to examine whether each on of them provides information overlapping, in an extent, with knowledge acquired by others. This is defined as multicollinearity. In linear regression, multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated, thus, proportion of variance explained by one regressor, is partly explained by another one, leading to variance inflation of the coefficient estimates. Subsequently, methods that detect the existence of multicollinearity and limit its effects are to be implemented. Before proceeding to diagnostics that will be used in order to identify multicollinearity issues, it would be useful to visualize the correlation between all the quantitative variables. Fig. 6 depicts Pearson's correlation of variables when examined via pairwise comparisons. Particularly, red color stands for negative correlation, with scaling intensity dependent on its exact value, whereas blue hues represent positive correlation. Sample Pearson's correlation coefficient is given by

$$r_{xy} = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})}{S_x} \frac{(y_i - \bar{y})}{S_y}$$

The largest correlation is observed between `accommodates` and `beds`, followed by the pairs `bedrooms`, `beds` and `bedrooms`, `accommodates`. The respective correlation values are 0.82, 0.73 and 0.72, respectively.

After examining the correlogram, we investigate multicollinearity in terms of the Variance Inflation Factor (VIF) criterion for the estimated coefficients of each regressor. The standard errors, and hence the variance, of the estimated coefficients are inflated when multicollinearity exists. VIF consists a quantifier of the aforementioned issue. The idea is to detect collinearity due to each explanatory variable, by forming a model with a particular regressor as the response variable and all other independent variables as the explanatory ones. Procedure is repeated for all explanatory variables and demonstrates, in what extent, one predictor is explained by the others. As a result, if one of the calculated models

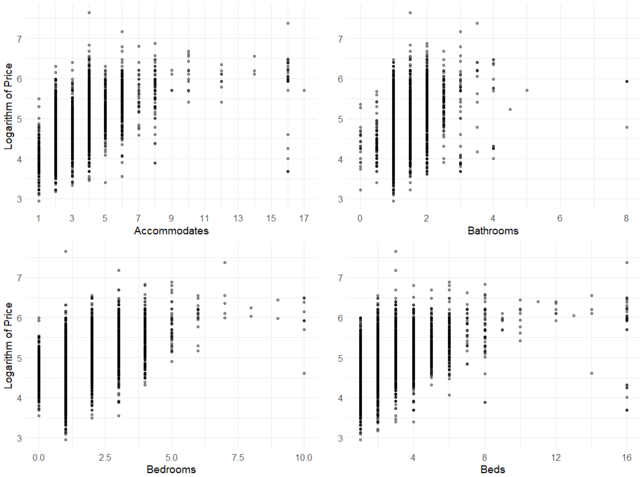

Fig. 4. Logarithm of price with respect to accommodates, bathrooms, bedrooms, and beds respectively.

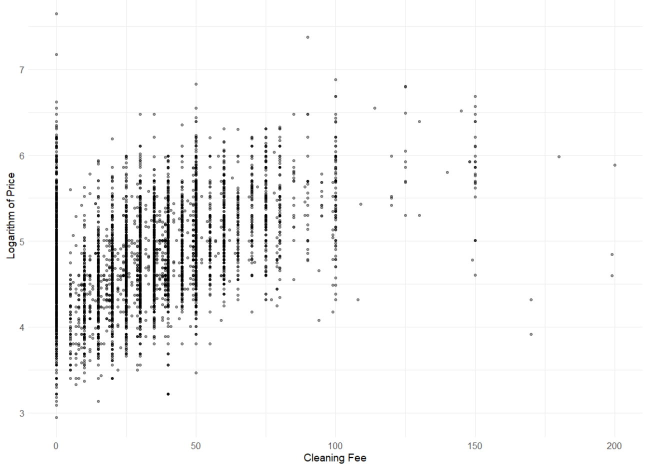

Fig. 5. Logarithm of price with respect to cleaning fee.

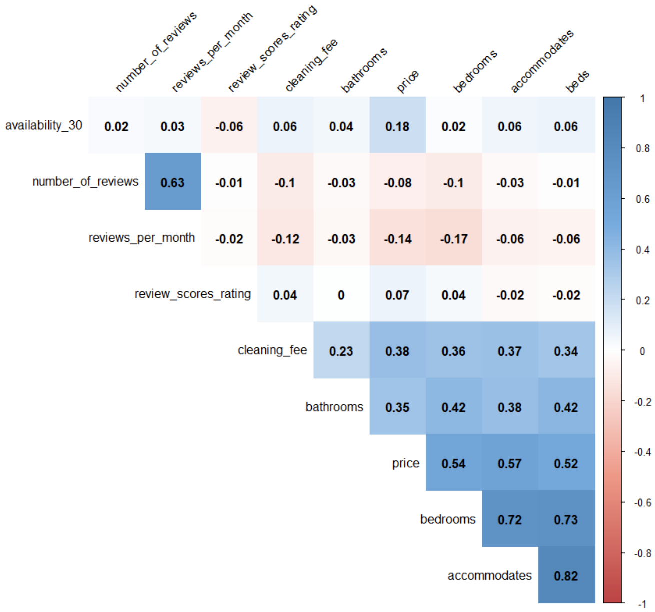

Fig. 6. Pairwise correlations between variables.

TABLE V
VIF VALUES.

Variables	VIF
`accomodates`	3.45
`bathrooms`	1.26
`bedrooms`	2.57
`beds`	3.58
`cleaning_fee`	1.2
`availability_30`	1.01
`review_scores_rating`	1.01
`reviews_per_month`	1.04

presents a high VIF value (empirically, many researchers use ten as a threshold [8]), corresponding response variable (which serves as predictor for the main model) is explained by the others. Table V depicts the values of VIF for each variable, extracted from the variance of each slope calculated from the formula

$$Var(b_k) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)} \times \frac{1}{1 - R_k^2}$$

where,

$$VIF_k = \frac{1}{1 - R_k^2}$$

As represented in Table V, values of Variance Inflation Factor fluctuate between 1 and 2 for most predictors. When it comes to beds and accommodates values are slightly bigger, reaching 3.58 and 3.45 respectively. Intuitively, number of beds are related to the number of guests; still, its values do not emerge any further concerns. Therefore, selected regressors will not cause collinearity issues.

Our aim is to create a model that balances between parsimony and performance, that is, it is easily interpretable because it leads to easier explanations on how each variable affects the response, and simultaneously explains a large variation in the data. In order to achieve this, we use the `leaps()` function which employs an exhaustive search to find the model that yields the greatest adjusted R^2. The function works only for quantitative variables. Thus, `leaps()`, builds models with every combination of variables, reports the adj. R^2, and we can choose a parsimonious model able to explain a large proportion of variance in the data. The resulting model contains variables `accommodates`, `bedrooms`, `cleaning_fee` and `availability_30`, with adj. R^2 equal to 0.41. Regarding categorical variables, all three are statistically significant as indicated from the partial F-tests conducted using `drop1()`. In order to decide

TABLE VI

PREDICTORS TO BE USED IN THE LINEAR MODEL.

VARIABLES
accommodates
bedrooms
cleaning_fee
availability_30
neighborhood
room_type
neighborhood:cleaning_fee
neighborhood:availability_30

which of those we are going to include in our model, we build full and reduced models and we examine the decrease in adj. R^2. In other words, we examine how well the model fits the data when we exclude the categorical variables, one at a time. Building a full model that contains the four quantitative and three qualitative variables respectively, we acquire a model with 0.514 in adj. R^2. In the first reduced model, we remove variable room_type, which resulted in a decrease of 0.02. Furthermore, removing property_type and neighborhood yields a decrease of 0.006 and 0.072, respectively. We decided to include variables room_type and neighborhood.

Forward selection, a stepwise procedure based on hypothesis tests is used in order to examine the presence of interactions between quantitative and qualitative variables. Using the full model described above, we update it by including different interactions one at a time, and we select the one that provides the greatest reduction in Residual Sum of Squares (RSS), and subsequently the smallest p-value, based on F-tests. We update the model and we conduct the F-test using the update(), and anova() function, respectively. The anova() function constructs sequential F-tests by adding one variable at a time starting from the null model (intercept only), until the full model is constructed. Since the interaction is added last, the result for the latter is equivalent to conducting a partial F-test. The interaction neighborhood:cleaning_fee provided the largest decrease in RSS, followed by neighborhood:availability_30. The interaction of bedrooms with the two qualitative variables resulted in a rank deficient model matrix due to linear combinations of the columns, and therefore, we have not included those interactions in the model. As we have mentioned, our focus lies on creating a model that is relatively simple with good prediction accuracy. That being said, we decided to add only two interactions in total. The chosen predictors after the inclusion of interactions can be found in Table VI.

VI. MODEL SELECTION

The model containing the regressors of Table VI can explain 51.7% of the variance in the data. Fig. 7 represents the Quantile-Quantile (Q-Q) plot of residuals of the original and log transformed response, respectively. The residuals of the original response indicate a large deviation from normality. Although the residuals of the transformed response behave more normally, deviation from normality is still apparent. For the remaining part of this analysis, we use the logarithm transformed response.

Transformation of variables is not recommended if not inevitable, as it makes interpretation of the model more complex. Nonetheless, when it comes to real data, simultaneous holding of assumptions as studied in Statistics' books is utopian and researchers' interference is crucial. Due to the fact that the response variable is skewed to the right, a ”down-the-ladder” transformation [5] is utilitarian. This could be the logarithm or some power of roots; Even though both work smoothly, logarithm performs slightly better. Still, it is interpretable and the inverse function is easily computed. That is, it is bijective, and consequently invertible, reverting transformed values to the original ones.

Variables accommodates and bedrooms have a curvilinear relationship with the logarithm of price, as can be seen in Fig. 4. This indicates that a polynomial factor could be introduced. Another option would be to use the logarithm transformation for accommodates and bedrooms, in order to make the relationship more linear. Furthermore, it could be the case that a model with no or only one interaction provides better predictions on unseen (new) data compared to a model with two interactions. A procedure called Cross Validation (CV) can be used to choose between different models with respect to the accuracy they provide on new data. In k-fold CV, the data are split into k folds. The first fold serves as a test set, whereas the rest $k-1$ folds indicate the train set. The model is fitted on the train set, and it is evaluated on the test set using metrics such as the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), or the Median Absolute Error (MedAE). Then, the second fold serves as a test set, and the rest correspond to the train set. This process is repeated until all folds have been used as a test set once. The average error over the k test sets provides an estimation of the model's prediction error on unseen data. Applying CV for different models, we can choose the one with the lowest prediction error.

In total, we have considered 21 models containing different combinations of interactions, polynomial terms of second order, and log transformations, which can be found in the Appendix. We perform 10-fold CV, repeated 20 times using the set.seed() function, that is, repeated for different samples of train and test folds. In each fold, we calculate three metrics, specifically the RMSE, MAE, and MedAE. In each repeat, we average the error over the 10 folds, and then we average the errors over the 20 repetitions. For each metric, a different model was found as the best one. Table VII illustrates the results. The RMSE is the standard deviation of the differences between the predicted and observed values. A disadvantage of RMSE is that it is sensitive to outliers, because the deviation of each predicted and observed value is squared. Thus, few outliers with large deviation can inflate the value of RMSE. The model containing the two interactions and a polynomial term for accommodates resulted in the lowest RMSE of 52.89€.

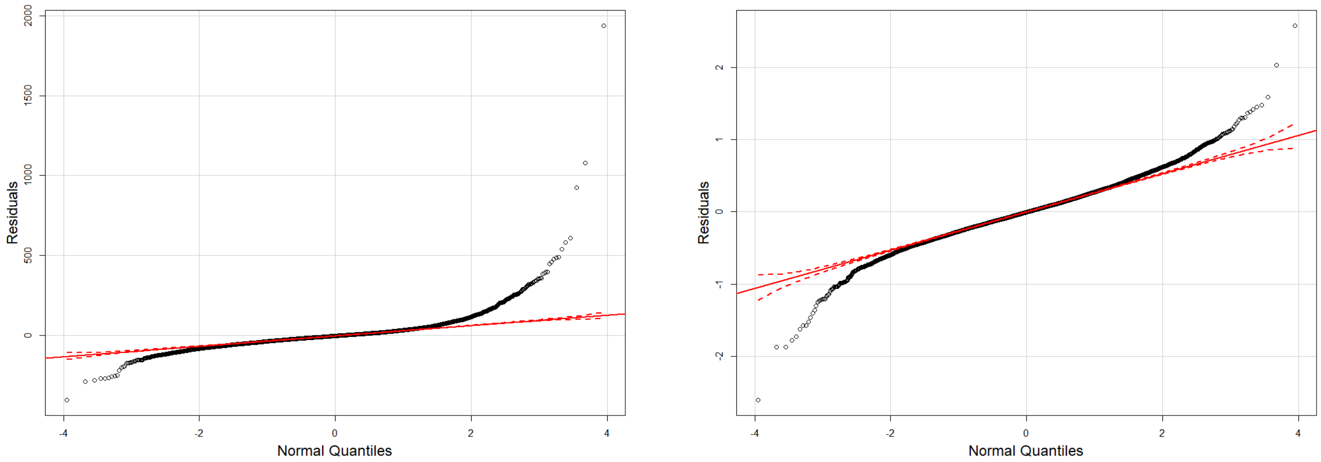

Fig. 7. Left: Q-Q plot of model's residuals using the original response variable. Right: Respective Q-Q plot using the log transformed response.

The MAE corresponds to the absolute difference between predicted and observed values. It is more robust to outliers compared to RMSE because of the calculation of absolute (instead of squared) differences. The model containing an interaction between `neighborhood` and `cleaning_fee`, and a polynomial term for `accommodates` yielded the lowest MAE of 31.05€. Finally, MedAE is a measure of statistical dispersion and represents the median of the absolute difference between predicted values and the median of the observed values. The lowest MedAE (19.86€) resulted from the most parsimonious model, that is the one with no interactions and no polynomial terms.

VII. ASSUMPTIONS CHECK

Now, that the most efficient -in price prediction- model has been selected, a rigorous examination of its adherence to the linear assumptions renders the next step; discussing assumptions has nothing to do with the response variable itself, even though it is one of the most common mistakes researchers undergo. Instead, residuals deriving from the model consist the topic of elaboration. That is, model's residuals should be independent and identically distributed, complying with homoscedasticity, while their expected value should be zero. Mathematically, that is

$$\varepsilon_i \sim N(0, \sigma_{\varepsilon}^2)$$

Linear assumptions' regnancy guarantees validity of least square estimators. In concrete words, when the sample is large enough, errors are perceived as normally distributed, due to the Central Limit Theorem. In reality, however, efficiency is achieved only when they are indeed, approximately, normally distributed. Should the errors deviate from normality, least square estimators' robustness is debatable and the latter ones tend to be biasedly estimated, contradicting Gauss-Markov Theorem. Even though the Central Limit Theorem compensates for potentially not normally distributed residuals, this is not the case for heteroscedasticity. According to the latter, constant variance assumption is violated and plotted residuals against fitted values demonstrate a specific pattern which cannot be explained by randomness. In other words, the systematic part of the model is incapable of explaining the proportion of variance it is supposed to, as there could be omitted explanatory variables, an erroneous data sampling procedure, heavy influence of outliers or a transformation for either the dependent or some of the independent variables is needed. Finally, residuals should be independently distributed, otherwise researchers employ mixed models that consider ”the correlation between relatives” [4]. In the latter case, dependency of response on regressors should be linear and the expected value of the former conditional to the latter ones should be zero.

Concerning this case study, we assume linearity and independence of residuals; pricing of lodges have no correlation in between them. Intuitively, house owners declare prices for their properties regardless of the ones asked by others. Checking for normality of the errors, it is quite common to use Q-Q plots instead of plotting a histogram; that is, models' residuals plotted against the theoretical normal line. Fox incorporates function `qqPlot` in `car` package [6] which draws 95% confidence intervals using the standard error of the order statistic X_i, computed from the respective distribution as

$$SE(X_i) = \frac{\hat{\sigma}}{p(z_i)} \sqrt{\frac{P_i(1-P_i)}{n}},$$

where $p(z_i)$ is the probability density function corresponding to the Cumulative Distribution Function (CDF) $P(z_i)$, n is the sample size and $\hat{\sigma}$ is the residual standard error.

The Q-Q plot of the fitted model's residuals is shown in Fig. 9. Apparently, residuals distribution is quite heavy-tailed. Thus we cannot declare normality of errors. Besides that, Fig. 10 illustrates that variance does not increase with

TABLE VII

BEST MODELS WITH RESPECT TO RMSE, MAE, AND MEDAE, AS SELECTED FROM 10-FOLD CV, REPEATED 20 TIMES.

RMSE (52.89)	MAE (31.05)	MedAE (19.86)
`poly(accommodates, 2)`	`poly(accommodates, 2)`	`accommodates`
`bedrooms`	`bedrooms`	`bedrooms`
`cleaning_fee`	`cleaning_fee`	`cleaning_fee`
`availability_30`	`availability_30`	`availability_30`
`neighborhood`	`neighborhood`	`neighborhood`
`room_type`	`room_type`	`room_type`
`neigh/hood:clean._fee`	`neigh/hood:clean._fee`	
`neigh/hood:avail._30`		

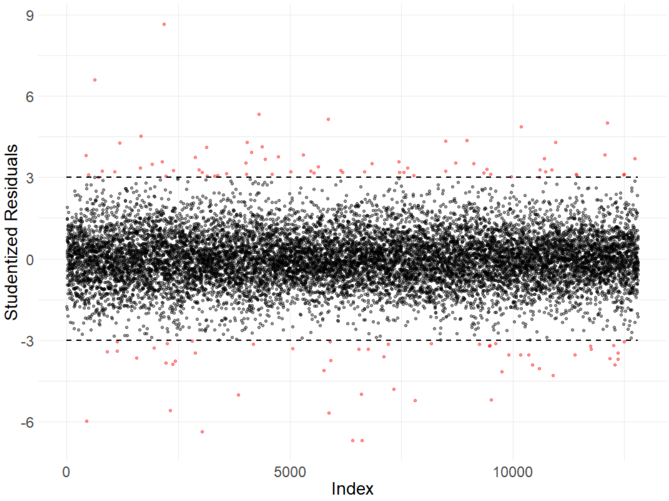

Fig. 8. Studentized residuals with absolute value greater than three are indicated with red color.

the mean, therefore, non-constant variance does not exist. This leads to unbiased standard error estimation, which in turn results in decrease of the probability of making a type II error, being itself complementary with power.

A. Leverage and Influence

Having examined the aforementioned plots, assessment of observations causing high leverage or influence, triggering outliers, is necessary. One of the most common methods for detecting such points is Cooks' Distance. An empirical threshold suggested in [1] is $\frac{4}{n}$, where n is the number of observations in the data set. Observations that surpass this threshold are considered to be outliers. A total of 732 observations exceeds the aforementioned threshold. However, they account for 6% of the total data set, so excluding them would lead to us losing great amount of information. Instead, allowing for a bigger threshold which would exclude only a few observations is a more concrete approach. Taking this into consideration, excluding 121 observations whose studentized residuals are greater than the absolute value of three, seemed more rational. Fig. 8 represents the studentized residuals. Those with absolute value greater than three are depicted with red color.

B. Assumptions reassessment

After deleting the 121 observations based on the residuals they generate, the model was refit and assumptions were examined again. This time, Q-Q plot is more steady in terms of normality, as depicted in Fig. 11. Fig. 12 illustrates the residuals against the fitted values. Homoscedasticity is still present, as there is no obvious pattern witnessed, that is, residuals form a cloud-like figure. That being said, we can evidently state that the linear assumptions are met. Subsequently, the ordinary least squares estimators are the most efficient, unbiased and consistent of all estimators minimizing the sum of squared residuals via the formula

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

VIII. MODEL VALIDATION

We use 10-fold CV in order to validate our model and obtain an estimation on how well it predicts prices of new listings. The problem we face is that we need to incorporate the process of training a model using a set that contains no outliers, but at the same time, include the outliers in the test set so we can obtain a valid result regarding model's performance. In order to solve this problem, first, we exclude all the studentized residuals with absolute value greater than three, as calculated by fitting the model in the whole data set. We create the train and test folds using the data set by excluding the outliers. In each of the 10 repetitions, we train the model using the train set. Then, we acquire a sample of the outliers and we add it in the test set. The percentage of outliers added in the test set is equal to the percentage of outliers found in the original data set. In this way, we estimate the prediction error of a model trained in a data set without outliers, by evaluating it in a test set that represents reality, that is, it contains both normal and outlier observations. The absolute difference between predicted and observed values (MAE) is 31.11 €. The RMSE is 52.64 €, whereas MedAE is 19.88 €.

IX. CONCLUSION

To sum up, in the framework of this paper, we analyzed Airbnb listings from Amsterdam, extracted from `www.insideairbnb.com`. Objective was to implement Linear or Generalized Linear Models, so as to approximate prices of listings conditional to explanatory variables which would provide the optimal equilibrium between performance and

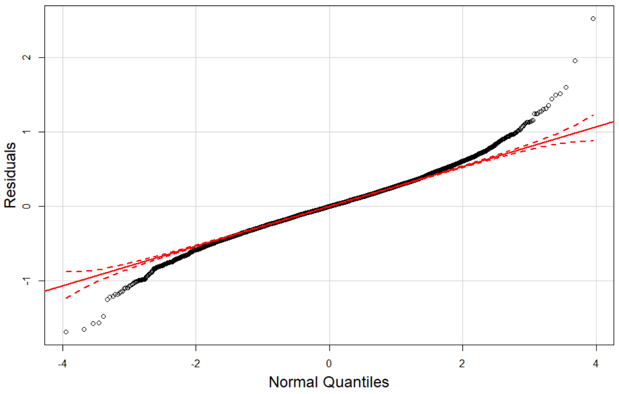

Fig. 9. Q-Q plot of residuals from model containing the outliers.

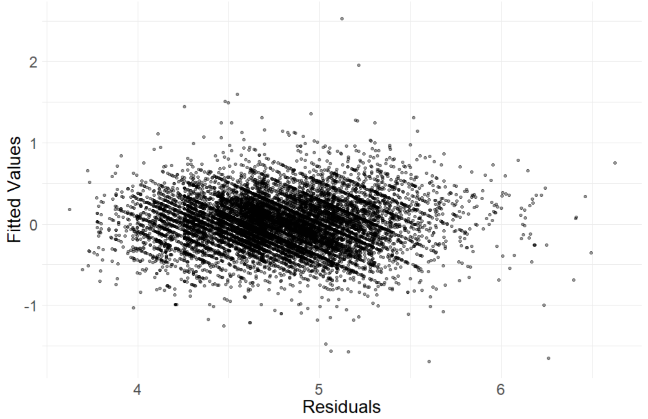

Fig. 10. Residuals vs fitted values from model containing the outliers.

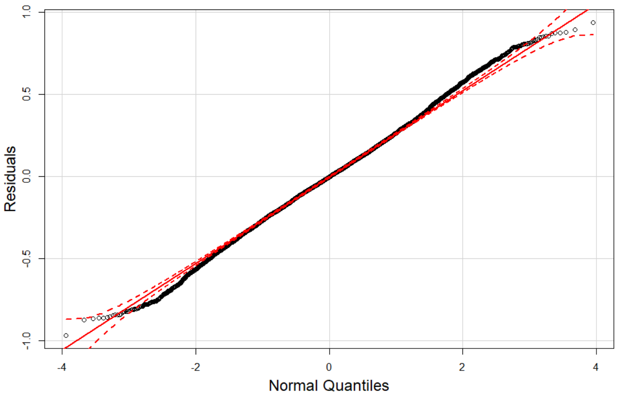

Fig. 11. Q-Q plot of residuals from model without outliers.

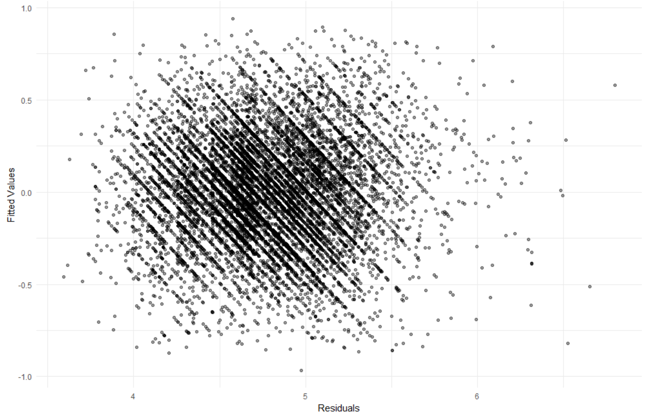

Fig. 12. Residuals vs fitted values from model without outliers.

parsimony. Towards this orientation, we cleaned and formulated the data so they would be apt for data analysis, being an indispensable section in every inquiry itself. Next step was to plot the data; plotting enables us to distinguish relationships between response and predictors and obtain a general idea of the individual variables' distributions. We checked for multicollinearity, we used incremental sums of squares and compared models based on both p-values and proportion of variance explained. Subsequently, 10-fold cross validation was employed 20 times to estimate model's prediction error. This was measured in terms of a variety of metrics, out of whom, the easily interpretable and robust to outliers MAE metric was chosen to indicate the best model. This procedure was conducted using the natural logarithm transformation for the response variable, in order to normalize residuals and homogenize variance, so as final model is the one whose estimators are the most efficient, unbiased and consistent.

REFERENCES

[1] K. A. Bollen and R. W. Jackman. Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods & Research*, 13(4):510–542, 1985.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[3] Y.-C. Chiou, R.-C. Jou, and C.-H. Yang. Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice*, 78:161 – 177, 2015.

[4] R. A. Fisher. Xv.the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

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

[6] J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011.

[7] C. Gibbs, D. Guttentag, U. Gretzel, J. Morton, and A. Goodwill. Pricing in the sharing economy: a hedonic pricing model applied to airbnb listings. *Journal of Travel & Tourism Marketing*, 35(1):46–56, 2018.

[8] M. H. Kutner, C. Nachtsheim, and J. Neter. *Applied linear regression models*. McGraw-Hill/Irwin, 2004.

[9] Y. Li, S. Wang, T. Yang, Q. Pan, and J. Tang. *Price Recommendation on Vacation Rental Websites*, pages 399–407.

[10] K. Narita and H. Kitagawa. Outlier detection for transaction databases using association rules. In *2008 The Ninth International Conference on Web-Age Information Management*, pages 373–380, July 2008.

[11] D. Wang and J. L. Nicolau. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb.com. *International Journal of Hospitality Management*, 62:120 – 131, 2017.

[12] Z. Zhang, R. J. C. Chen, L. D. Han, and L. Yang. Key factors affecting the price of airbnb listings: A geographically weighted approach. *Sustainability*, 9(9), 2017.