Predicting Airbnb Prices Using Multiple Regression Analysis

Authors: Amaya McNealey, Nhu Nguyen, Emily Hasler, Mahya Qorbani, Nicole (Yuge) Hu

Introduction

The aim of this study is to develop a model to predict how the overall Airbnb prices are related to various predictors for location, size, amenities, and more. The analysis was performed on a dataset published by Gyódi and Nawaro [1], which contains Airbnb prices in 10 popular European cities, including Amsterdam, Athens, Barcelona, and others. The developed model is a contribution to the growing field of data-driven insights and decision-making in the hospitality industry, providing valuable information and recommendations to both hosts and guests. The tool is intended to help both the hosts and travellers on Airbnb hosts to improve their service and price their accommodations. Prospective guests will be able to use this tool to evaluate their options, understand the market rates, and better prepare the financials for their trips.

Keywords: Regression, Airbnb, House Prices, Multiple Linear Regression

Problem Statement

The project aims to predict the price of Airbnb listings and identify the key predictors that affect the listing prices the most. The multiple regression models and statisticaly analyses were examined in this study to identify most important predictors on overall Airbnb prices, to evaluate models quality and determine techniques to improve model fits. The identified important predictor variables are: cleanliness rating, room type, guest satisfaction score, number of bedrooms, distance from the city center, host designation, and maximum guest capacity. The dependent variable is the total price of the Airbnb listing for two people and two nights in EUR (variable: realSum).

There are no specific constraints to the problem, and various regression techniques and models will be used to analyze the data and identify the key factors that impact the price of Airbnb listings. Ultimately, the goal is to provide insights and recommendations to both Airbnb hosts and prospective guests.

Data Description

The dataset used in the project was published by Gyódi, K., & Nawaro, Ł. (2021) [1] and obtained from Kaggle. The data was collected in 2019 through web automation to query Airbnb prices. A feature indicates whether the price is for a weekday or weekend accommodation. Each of the 10 cities in the dataset has between 1000 to 4600 data points, all with the same set of predicting and response variables. For instance, there are 3129 data points for Paris on weekdays alone. The full dataset has ~ 50k data points.

The dependent variable to be predicted is realSum in the dataset, which refers to the total Airbnb listing price for 2 nights and 2 people. The subset of key predictors from the final model are listed in Table 1. A comprehensive list of all predictors can be found in Appendix Table A1.

Table 1: a subset of predictors in the dataset

Variable	Variable Type	Description
room_type	Qualitative (categorical)	The type of room being offered (shared room, private room, entire home/appt)
room_shared	Qualitative (boolean)	Whether the room is shared or not
room_private	Qualititative (boolean)	Whether the room is private or not (is not always the opposite of room_shared in the database)
person_capacity	Quantitative (1-6 people)	The maximum number of people that can stay in the room
host_is_ superhost	Qualitative (boolean)	Whether the host of the listing is a superhost

Analysis

Data Preprocessing

Preprocessing steps were performed to prepare the dataset for use in the model: preliminary First, the distribution of all predictors are inspected in box plots and scatter plots to examine for trends indicating relationships among predicting variables and between predicting and response variables as shown in Appendix A1. As an example, Figure 1 shows the house prices for weekdays and weekends in each city, where there is no significance difference in prices between weekday and weekend bookings in most cities, with the exception of Amsterdam. Amsterdam also has the highest average price of any city, and it is evident that there is a huge difference in prices for all cities.

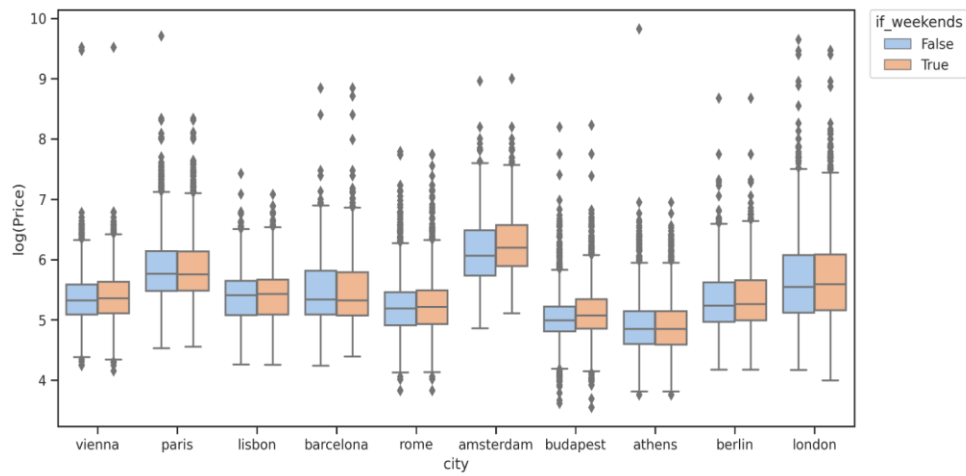


Figure 1: Comparison of house prices for 2 nights and 2 people in 10 major cities on weekdays or weekends

The next preprocessing step was to perform a correlation analysis to determine highly linearly correlated predicting variables. The correlation matrix is shown in Figure 2 for all variables. Strong correlation was observed between the cleanliness rating and the guest satisfaction rating, as it is usual for a cleaner environment to lead to higher overall satisfaction. Other variables such as “room_shared”/“room_type_Shared room” and “room_shared”/ “room_type_Private room” are highly correlated, which were expected, due to redundancy. Latitude and longitude are manually removed because the geographical information cannot be captured by a linear model, and the dataset has a predictor, dist, to capture the distance between the listing to city center. From the correlation analysis, predicting variables with correlations greater than 0.75 were excluded, leaving 23 predictors to build the regression models.

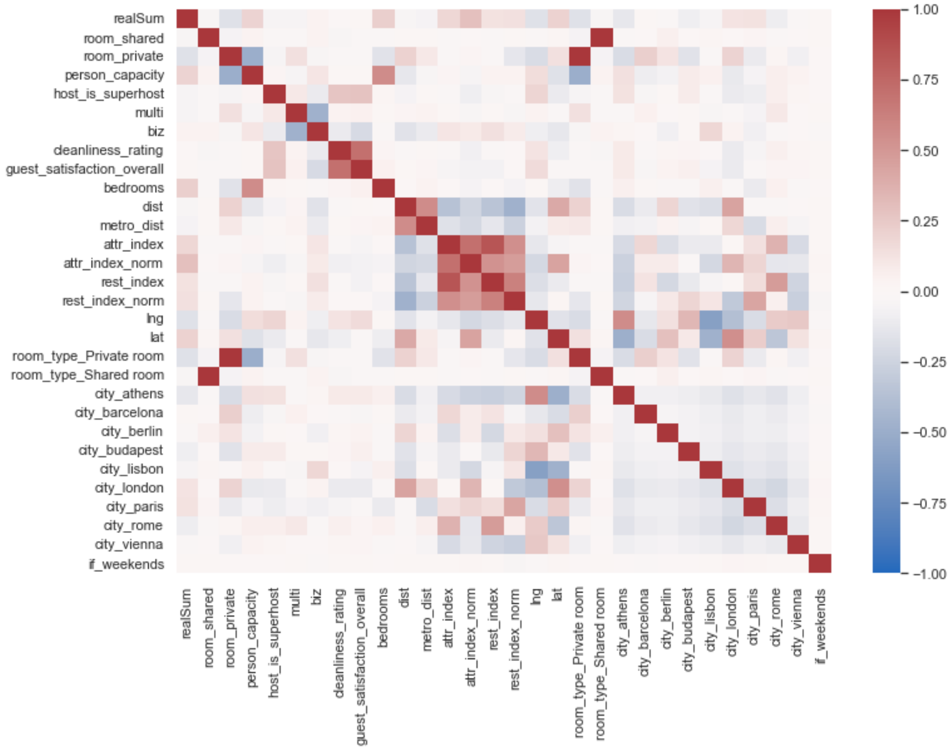


Figure 2: Heat map of correlation matrix on all predictors for the Full 50k dataset

Next, quantitative variables are scaled using standardization to address the impact of large values in the dataset. This is important because these large values may have a significant influence on the analysis and could potentially obscure the patterns in other variables. Due to an overly large dataset, cook's distance on the preliminary multiple linear regression was ineffective to determine outliers for the full 50k dataset. As a result, data points that were outside the three sigma range of the response variable were instead treated as outliers and excluded for the preliminary model, with the results shown in Figure A2 (appendix). Removal of outliers by Cook's distances was performed for the model trained on 500 data points randomly selected with a uniform distribution from the original 50k full dataset (Section. 500 Datapoints Model).

Preliminary Model

The preliminary model utilized only 5 cities and any predictors with a correlation larger than 0.75 were removed. A log transformation was performed on the price variable to produce a better R^2 value than the untransformed price variable. As can be seen in the validated model output below in Figure A3 (appendix), the model is significant as all the predictors used in the model are significant with a p-value close to 0. Amsterdam was used as the level for the city categorial variable. Thus, since all of the coefficents for the city predictors are negative, Amsterdam is the most expensive city on average which follows what was seen when plotting the data in Figure 1.

As seen in Figure A4 (appendix), the histogram of residuals generally fits the normal distribution without outliers. The residuals vs fitted values plot shows a general fit to the linearity assumption though is unclear due to a large number of data points. The normal Q-Q plot does have a mostly linear relationship though there is a skew in the upper quartile. As seen in Figure A5 (appendix), there are no VIF values larger than max[10,1/(1-R_j^2)] =10, therefore no multicollinearity is present in the model.

Full Dataset Model

The next step was to develop the model using data from all of 10 cities (non-transformed full dataset model). Figure A6 (appendix) shows the model output for the untransformed price variable with all predictors. The p-values of the model and all estimators ≈ 0 indicate that the model is significant but the low adjusted R^2 value of 0.24 indicated that model was not very useful in explaining the total variability. From the residual analysis of the model on the original response variable in Figure A7 (appendix), the normality and constant-variance assumptions are violated because the residuals do not scatter uniformly against fitted values in Figure A7a and obvious outliers are observed in the QQ plot (Figure A7b). The violations suggests that a transformation is needed. No obvious multicolinearity is observed given the VIF analysis in Figure A8.

To improve the explanatory power of the model (R^2), a log transformation was performed on the response variable price and a multilinear regression was fitted. The residual analysis for the model with transformation is shown in Figure 3. The scatter plot (Figure 3a) and the histogram (Figure 3c) show better compliance to the normality and constant-variance assumptions compared to non-transformed model while the transformed model still shows some signs of non-constant variance and non-normality. The QQ plot (Figure 3b) shows deviation from the normal distribution at the upper quantiles (> 2). Figure 3 shows the positive effect of the log transformation of the price on the residual values compared to the non-transformed model in Figure A7. Overall, the log transformation on the predicting variable boosted the R^2 to 0.6587 and the adjusted R^2 to 0.6585 and improved the error distributions. VIF values were also examined after the log transformation, with none larger than 10 (critical value), confirming no multicollinearity.

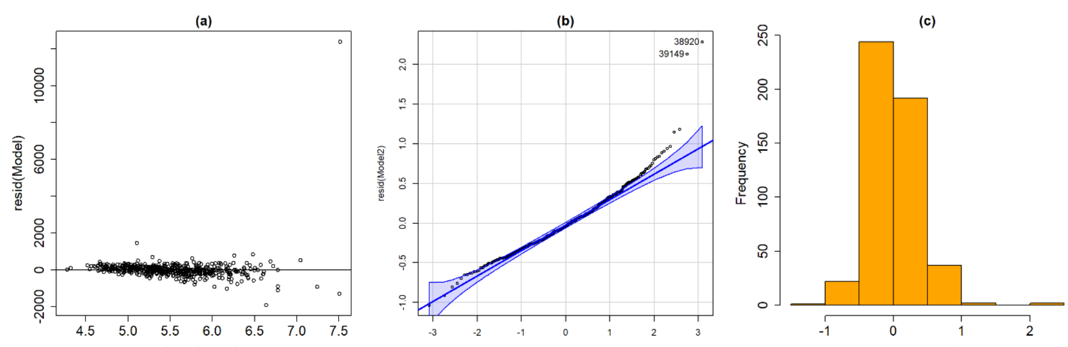


Figure 3: Residual analysis for log transformation on the y variable

A forward feature selection algorithm was performed to reduced the number of predictors – a crucial step due to large number of predictors having the potential to inflate the R^2 values. Figure 4 shows the output of the model developed by this algorithm. This reduced model is also statistically significant with a model p-value ≈ 0 and with all remaining predictors being statistically significant. The reduced model from forward selection resutled in a slightly lower R^2 value of 0.6571 due to reduced number of predictors. As seen in Table 2, this model does not contain multicolinearity, indicated by all VIF values being smaller than 10 (critical value).

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.418983   0.001841 2943.17   <2e-16 ***
room_private       -0.196957   0.002334  -84.38   <2e-16 ***
person_capacity     0.118213   0.002626   45.02   <2e-16 ***
bedrooms            0.091390   0.002301   39.71   <2e-16 ***
attr_index_norm     0.201996   0.002469   81.81   <2e-16 ***
city_athens        -0.395796   0.003515 -112.60   <2e-16 ***
city_paris         -0.180176   0.003611  -49.90   <2e-16 ***
city_budapest      -0.358015   0.003114 -114.95   <2e-16 ***
city_rome          -0.377124   0.003937  -95.78   <2e-16 ***
room_shared        -0.067770   0.001858  -36.47   <2e-16 ***
city_vienna        -0.217445   0.002998  -72.53   <2e-16 ***
city_lisbon        -0.242342   0.003540  -68.46   <2e-16 ***
city_berlin        -0.169598   0.002682  -63.23   <2e-16 ***
city_london        -0.250757   0.004094  -61.25   <2e-16 ***
city_barcelona     -0.140712   0.002803  -50.21   <2e-16 ***
cleanliness_rating  0.043949   0.001877   23.41   <2e-16 ***
biz                 0.040742   0.001959   20.79   <2e-16 ***
---

    Residual standard error: 0.3503 on 36177 degrees of freedom
    Multiple R-squared:  0.6571,    Adjusted R-squared:  0.6569
    F-statistic:  4333 on 16 and 36177 DF,  p-value: < 2.2e-16
```

Figure 4: Model output for forward feature selection algorithm reduced model

Table 2: Multicolinearity analysis for the reduced model on full dataset

Predictors	**VIF Value**	**Predictors**	**VIF Value**
room_private	1.607317	city_paris	3.845776
person_capacity	2.033804	city_budapest	2.861128
room_shared	1.018339	city_rome	4.573042
cleaniness_rating	1.039256	city_barcelona	2.316918
biz	1.132594	city_vienna	2.651416
attr_index_norm	1.798444	city_lisbon	3.696726
bedrooms	1.562132	city_berlin	2.121857
city_athens	3.644370	city_london	4.944549

The reduced model from forward selection algorithm showed that Amsterdam (baseline case) continues to be the most expensive city as all of the city predictor coefficients are negative. The model also informs the most important factors on the overall price. For every one-unit increase in the normalized attraction index, there is a 22.4% increase in price. There is a 17.9% decrease in price if the listing is for a private room, compared to the level of the listing for the whole apartment. If the person capacity of a listing increases by one unit, the price increases by

12.5%. There is another 9.6% increase in price for a one-unit increase in the number of bedrooms at the listing. A one unit increase in cleanliness rating leads to a 4.5% increase in price.

The results of the partial F-test between the full model and the reudced model with forward selection in Table A2 suggest that the difference is insignificant, and it's safe to remove the predictors from forward selection and use the reduced model to predict the prices while maintaing the predicitve power.

Feature selection with L1 regularization LASSO and backward selection were also performed but both did not lead to effective reduction of the number of predicing variables. Refer to Appendix. Model feature selection for more details.

500 Datapoints Model

Similar analysis were performed on a smaller dataset of 500 data points to investigate the inflation effect on p-values due to a large amount of data points. A total of 500 data points were randomly selected from the >50k full dataset of all 10 cities. The resulted scatter plot matrix is shown in Figure A9. Most predictors do not have a linear relationship with the response variable, given the amount of categorical predictors contained in the dataset. Reducing the dataset to 500 datapoints still resulted in similar R^2 value of 0.6478 and error distribution (comparing Figures A6 and A10 and Figures 4 and 9), suggesting the robustness of the full model.

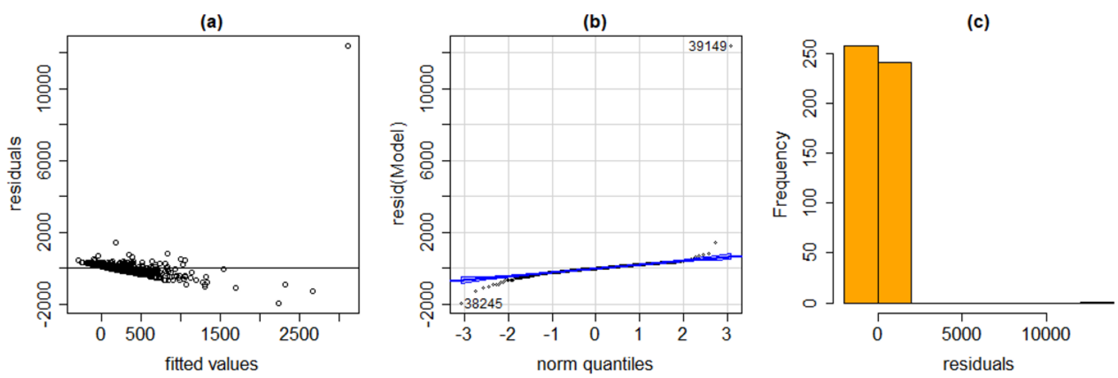


Figure 6: (a) Residual plots for 500 datapoints model, (b) QQ plot and (c) distribution of residuals for the untransformed regression model

Figure 6 shows the residual analysis following similar trend to results shown in Figure A7 for the larger dataset on untransformed response variable. With a log transformation on the response variable, Figure 7c shows the residuals following a normal distribution, suggesting the effectiveness of the transformation. Similarly, with forward feature selection, both models with the full predictor set and with the reduced predictor set are not significantly diffrent from each other as seen in ANOVA results presented in Figure 10. Multicollinearity is also not a problem in this model, indicated by the low VIF values shown in Table 3.

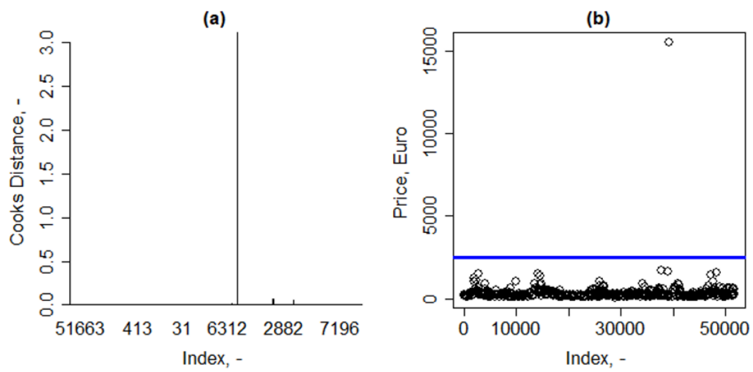


Figure 7: (a) Cook's distance and (b) 3-sigma plot for untransformed model

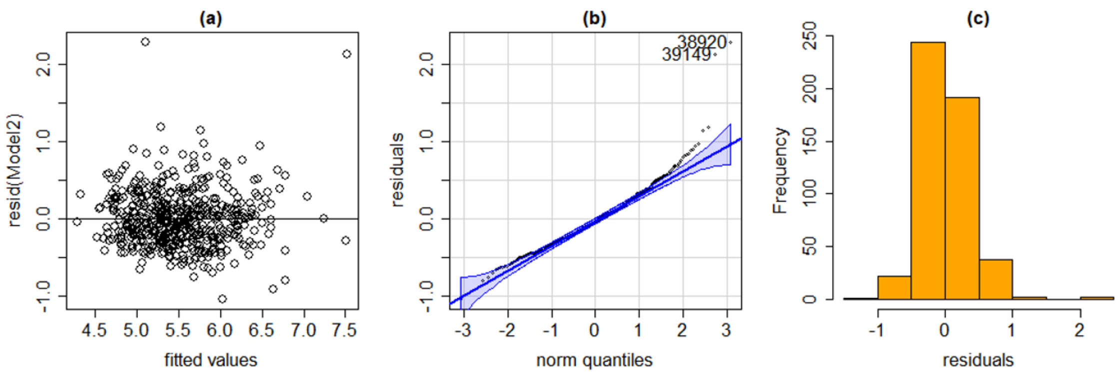


Figure 8: (a) Residual plots for 500 datapoints model, (b) QQ plot and (c) distribution of residuals for the post-transformed (log(y)) regression model

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.88146    0.26236  18.606  < 2e-16 ***
attr_index_norm    0.22088    0.04949   4.464 1.01e-05 ***
room_private      -0.45754    0.04466 -10.244  < 2e-16 ***
bedrooms           0.18422    0.03404   5.412 9.84e-08 ***
city_budapest     -1.28560    0.09830 -13.079  < 2e-16 ***
city_athens       -1.44290    0.11324 -12.742  < 2e-16 ***
city_rome         -1.00370    0.08663 -11.586  < 2e-16 ***
city_vienna       -0.94515    0.10239  -9.231  < 2e-16 ***
city_berlin       -0.74918    0.11574  -6.473 2.37e-10 ***
room_shared       -1.05772    0.28270  -3.742 0.000205 ***
person_capacity    0.06777    0.01936   3.501 0.000507 ***
city_lisbon       -0.85350    0.10190  -8.376 6.05e-16 ***
dist              -0.02315    0.01344  -1.722 0.085695 .
city_barcelona    -0.60093    0.10118  -5.939 5.49e-09 ***
city_paris        -0.50219    0.09140  -5.495 6.36e-08 ***
city_london       -0.49208    0.10090  -4.877 1.47e-06 ***
biz                0.15091    0.04398   3.432 0.000652 ***
cleanliness_rating 0.06718    0.02126   3.160 0.001675 **
multi              0.05044    0.04488   1.124 0.261680
---

  Residual standard error: 0.3792 on 481 degrees of freedom
  Multiple R-squared:  0.6478,    Adjusted R-squared:  0.6346
  F-statistic: 49.14 on 18 and 481 DF,  p-value: < 2.2e-16
```

Figure 9: Regression model output for 500 data points forward feature selection reduced model

```
Analysis of Variance Table

Model 1: log(realSum) ~ attr_index_norm + room_private + bedrooms + city_budapest +
    city_athens + city_rome + city_vienna + city_berlin + room_shared +
    person_capacity + city_lisbon + dist + city_barcelona + city_paris +
    city_london + biz + cleanliness_rating + multi
Model 2: log(realSum) ~ room_shared + room_private + person_capacity +
    host_is_superhost + multi + biz + cleanliness_rating + guest_satisfaction_overall +
    bedrooms + dist + metro_dist + attr_index_norm + city_athens +
    city_barcelona + city_berlin + city_budapest + city_lisbon +
    city_london + city_paris + city_rome + city_vienna + if_weekends
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    481 69.177
2    477 68.880  4   0.29642 0.5132 0.7261
```

Figure 10: ANOVA test output for model comparison (full vs. reduced) on 500 data points

Next, outliers were removed by computing the corresponding Cook's distances in Figure A11. The resulted model after removing these outliers is shown in Figure 12. This model has the best fit for the normality, the constant variance, and the independence assumptions compared to other models discussed previously in the study, with the highest R^2 value of 0.72.

Table 4: VIF values of predicting variables for 500 datapoints reduced model

Predictors	VIF Value	Predictors	VIF Value
room_private	1.551829	city_paris	3.845776
person_capacity	2.022411	city_budapest	2.415707
room_shared	1.106931	city_rome	3.740080
cleaniness_rating	1.081233	city_barcelona	2.316918
biz	1.525603	city_vienna	2.246752
attr_index_norm	4.465055	city_lisbon	3.191167
bedrooms	1.443818	city_berlin	2.212112
city_athens	3.5779953	city_london	4.944549
multi	1.399469	dist	3.49583

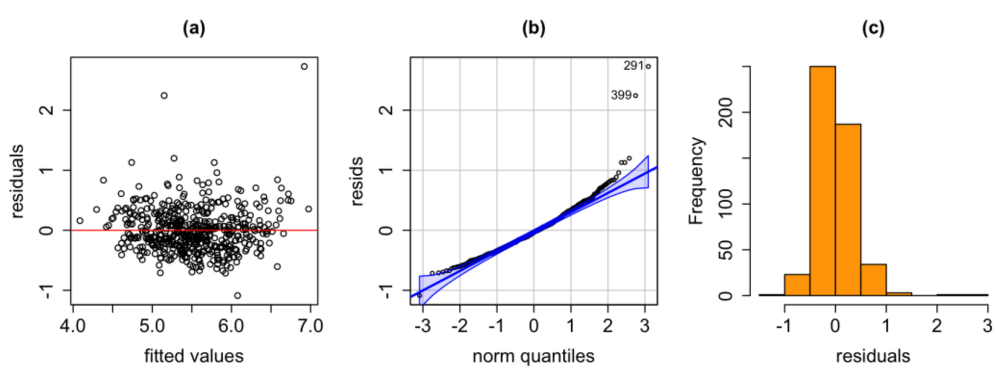


Figure 11: (a) Residual plots for 500 datapoints model, (b) QQ plot and (c) distribution of residuals for the reduced (forward selection method) and post-transformed (log(y))regression model

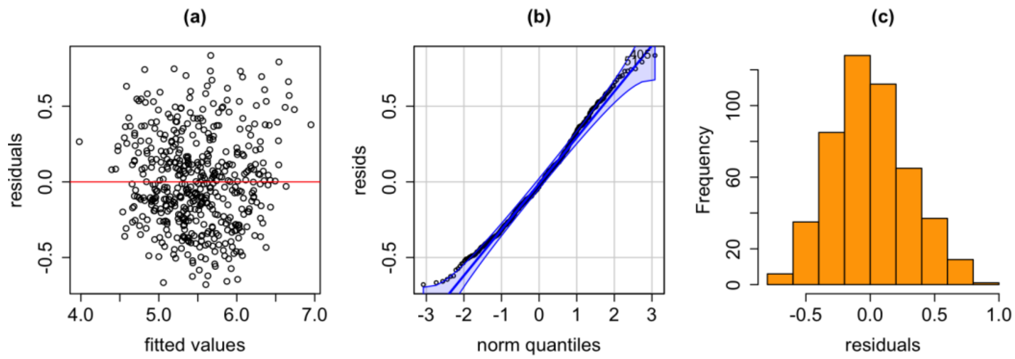


Figure 12: (a) Residual plots for 500 datapoints model, (b) QQ plot and (c) distribution of residuals for the reduced (forward selection method) and post-transformed (log(y))regression model after removing Cook's Distance outliers

Conclusions and Recommendations:

This project aimed to identify the key predictors for Airbnb listing prices and to provide an estimation of Airbnb listing prices to better guide both the Airbnb hosts and prospective guests. Based on the analysis conducted, the study identified that the most significant predictors were the attraction index in the neighborhood of the listing, whether the listing is for a private room, a shared room or a whole apartment listing, the overall person capacity, cleaniless level. The study found that a 22.4% increase in price was observed for every one-unit increase in the normalized attraction index, a 17.9% decrease in price for a private room listing, and a 5% increase in price for every one-unit increase in person capacity. Athens was identified as the cheapest city, while Amsterdam being the most expensive city. Weekends vs weekdays were not significant for most predictors in determining the best time to visit.

Table 4: Model Comparison

Model	R^2	Adjusted R^2	MSPE	MAE	MAPE	PM
Untransformed full dataset	0.222	0.221	69856.16	100.792	0.413	0.738
Transformed, Reduced full dataset	0.657	0.657	67107.38	79.297	0.259	0.709
Transformed, Reduced, outliers removed, 500 datapoints	0.726	0.716	13068.14	74.817	0.301	0.385

The error metrics of important models are presented in Table 4, showing the model that best fits the data is the one that employed a transformed response variable, removed outliers, and eliminated highly correlated features on a randomly subsampled 500 data points. To validate and ensure the models' robustness, all models developed in this study underwent a 70/30 train/test split, and the results of the test data were reported in Table 4. The robustness models was also verified by comparing the whole dataset vs subset of 500 data points, which yielded similar regression results.

This project encountered several challenges that had the potential to impact the validity and reliability of our findings. Table 5 provides a summary of the challenges in this projcet and the strategies have been employed to address them. This project was able to successfully navigate these obstacles and produce meaningful and reliable insights that contribute to the existing body of knowledge in this area.

Table 5: Challenges and mitigation

Challenge	Mitigation
Data set is too large to use Cook's Distance	Use a smaller subset of data
Large data set leading to a higher level of significance than what is true	70/30 Train/Test split for validation
We are unable to predict the higher quartile prices because our model accounts for the averages rather than extraneous listings.	Include more variables that can be correlated with higher prices
Geographic predictors highly correlated within each cities	Cluster graphical predictors for each cities and create a new predictor to encapsulate this information

Overall, the findings of this project provide valuable insights into the factors and predictors that affect Airbnb pricing, which include the location of the listing (city and attraction index), the inherent property of the listing (whole apartment vs. Private/shared room, person capaity, number of bedrooms). This study could be useful for hosts to optimize and reference back for their listing prices and for guests to make informed decisions when booking their stays. It is hoped that this research will contribute to the growth and sustainability of the short-term rental market.

Appendix

Data Description

Table A1: All of the predictors in the dataset

Variable	Variable Type	Description
room_type	Qualitative (categorical)	The type of room being offered (shared room, private room, entire home/appt)
room_shared	Qualitative (boolean)	Whether the room is shared or not
room_private	Qualilititative (boolean)	Whether the room is private or not (is not always the opposite of room_shared in the database)
person_capacity	Quantitative (1-6 people)	The maximum number of people that can stay in the room
host_is_ superhost	Qualitative (boolean)	Whether the host of the listing is a superhost
multi	Qualitative (boolean)	Whether the listing is for multiple rooms or not
biz	Qualitative (boolean)	Whether the listing is for business purposes or not
cleanliness_ rating	Quantitative (2-10 score)	The cleanliness rating of the listing
guest_satisfaction_overall	Quantitative (20-100 score)	The overall guest satisfaction rating of the listing
bedrooms	Quantitative (0-8 bedrooms)	The number of bedrooms in the listing
dist	Quantitative (km)	The distance from the city centre
metro_dist	Quantitative (km)	The distance from the nearest metro station
attr_index	Quantitative (score)	
attr_index_norm	Quantitative (0-100)	Normalized version of the attraction index
rest_index	Quantitative (score)	
rest_index_norm	Quantitative (0-100)	Normalized version of the restaurant index
lng	Quantitative	The longitude coordinate of the listing
lat	Quantitative	The latitude coordinate of the listing
if_weekends	Qualitative (boolean)	Whether the booking is for the weekend or the weekdays (1 if it is a weekend, 0 otherwise)
City	Qualitative (categorical)	The city the listing is in

Analysis

Data Preprocessing

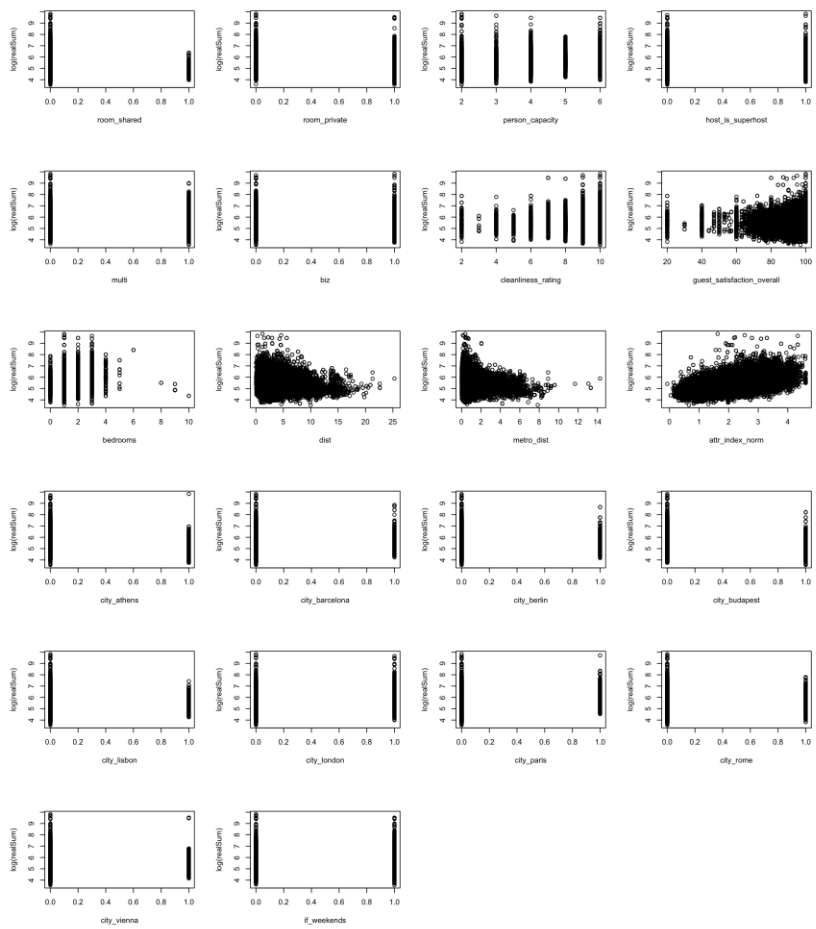


Figure A1: Scatter plots of transformed response variable, log(realSum), and various predicting variables.

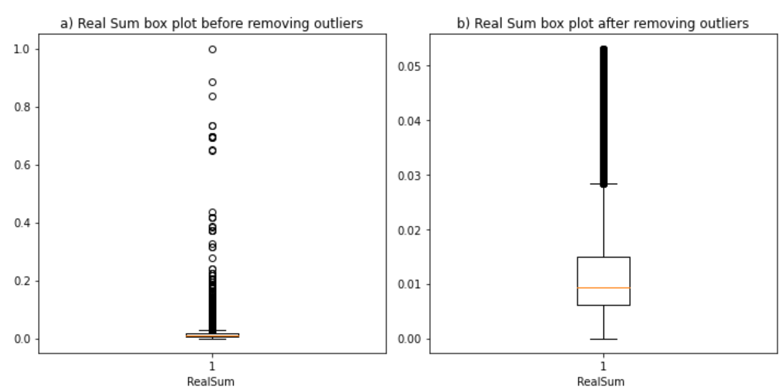


Figure A2: Outlier detection analysis, 3a) box plot of real sum before removing the outliers. 3b) box plot of real sum afterremoving the outliers.

Preliminary Model

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.7525060  0.0261134 220.290  < 2e-16 ***
room_shared               -0.9379834  0.0238891 -39.264  < 2e-16 ***
room_private              -0.4862183  0.0058876 -82.584  < 2e-16 ***
person_capacity            0.1316436  0.0026710  49.286  < 2e-16 ***
host_is_superhost          0.0580817  0.0064257   9.039  < 2e-16 ***
multi                      0.0167746  0.0054159   3.097  0.00196 **
guest_satisfaction_overall 0.0025232  0.0002522  10.005  < 2e-16 ***
city_barcelona            -0.5336117  0.0108288 -49.277  < 2e-16 ***
city_berlin               -0.5575751  0.0112446 -49.586  < 2e-16 ***
city_london               -0.3130948  0.0093252 -33.575  < 2e-16 ***
city_paris                -0.4090277  0.0094438 -43.312  < 2e-16 ***
bedrooms                   0.1536893  0.0048342  31.792  < 2e-16 ***
dist                      -0.0615857  0.0010020 -61.460  < 2e-16 ***

Residual standard error: 0.3642 on 23928 degrees of freedom
Multiple R-squared:  0.6218,    Adjusted R-squared:  0.6216
F-statistic:  3278 on 12 and 23928 DF,  p-value: < 2.2e-16
```

Figure A3: Preliminary model using 5 cities

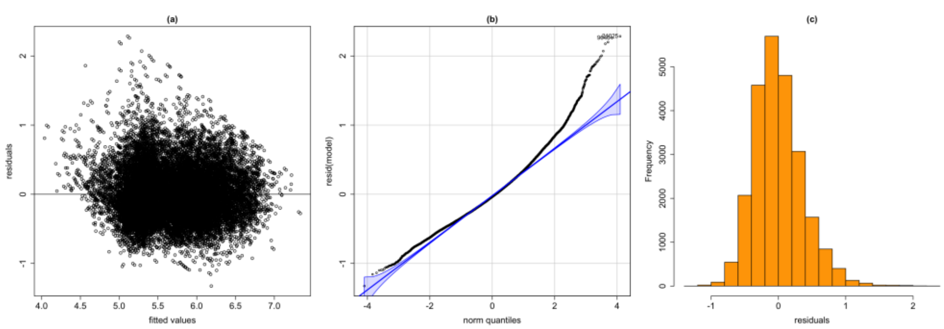


Figure A4: Residual analysis plots for Preliminary model

Predictors	VIF Value	Predictors	VIF Value
room_private	1.564179	city_paris	3.234365
person_capacity	1.851472	city_london	3.809554
room_shared	1.018204	city_berlin	2.118040
host_is_superhost	1.085298	city_barcelona	2.203160
multi	1.051239	bedrooms	1.517538
guest_satisfcation_overall	1.093823	dist	1.335472

Figure A5: VIF values to test for multicolinearity in the Preliminary model

Full Model

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                280.318      1.560 179.739  < 2e-16 ***
room_shared                -16.974      1.575 -10.779  < 2e-16 ***
room_private               -54.303      2.002 -27.123  < 2e-16 ***
person_capacity             33.510      2.230  15.027  < 2e-16 ***
host_is_superhost           -1.124      1.668  -0.674  0.50054    
multi                        6.145      1.827   3.364  0.00077 ***
biz                         16.436      1.937   8.484  < 2e-16 ***
cleanliness_rating           6.162      2.266   2.719  0.00655 ** 
guest_satisfaction_overall   6.482      2.311   2.805  0.00504 ** 
bedrooms                    55.287      1.953  28.315  < 2e-16 ***
dist                        -8.097      3.188  -2.540  0.01110 *  
metro_dist                  -6.345      2.014  -3.151  0.00163 ** 
attr_index_norm             52.070      3.125  16.662  < 2e-16 ***
city_athens               -132.249      3.233 -40.909  < 2e-16 ***
city_barcelona             -58.582      2.393 -24.483  < 2e-16 ***
city_berlin                -63.496      2.443 -25.991  < 2e-16 ***
city_budapest             -122.807      2.673 -45.947  < 2e-16 ***
city_lisbon               -105.892      3.152 -33.592  < 2e-16 ***
city_london                -86.156      3.973 -21.686  < 2e-16 ***
city_paris                 -73.765      3.199 -23.062  < 2e-16 ***
city_rome                 -143.705      3.369 -42.653  < 2e-16 ***
city_vienna                -84.676      2.574 -32.897  < 2e-16 ***
if_weekends                  4.017      1.563   2.570  0.01016 *  
---

Residual standard error: 296.7 on 36171 degrees of freedom
Multiple R-squared:  0.2218,    Adjusted R-squared:  0.2213 
F-statistic: 468.5 on 22 and 36171 DF,  p-value: < 2.2e-16
```

Figure A6: Full dataset multiple linear regression model

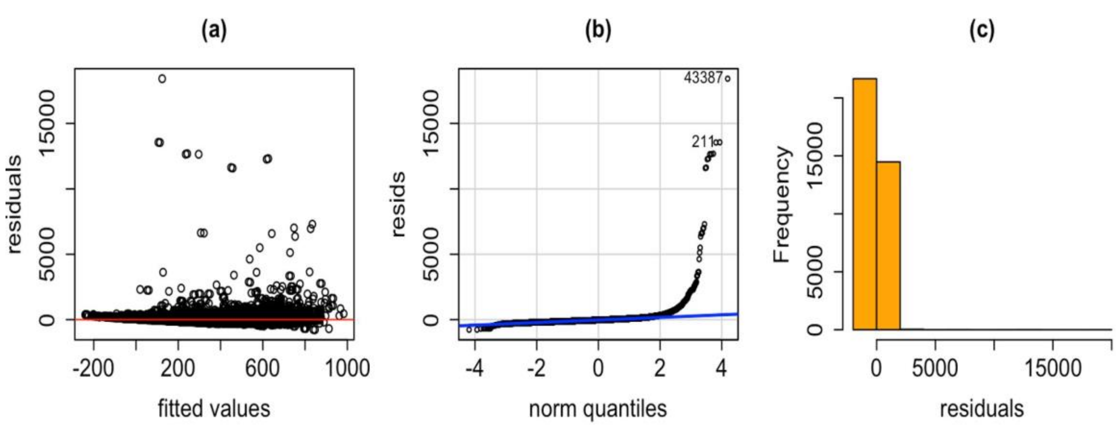


Figure A7: Residual analysis with untransformed y variable (price)

Predictors	VIF Value	Predictors	VIF Value
room_private	1.647888	city_paris	4.206064
person_capacity	2.044600	city_budapest	2.936975
room_shared	1.019548	city_rome	4.666644
cleaniness_rating	2.111131	city_barcelona	2.353873
biz	1.543194	city_vienna	2.723809
attr_index_norm	4.015323	city_lisbon	4.085261
bedrooms	1.562132	city_berlin	2.453651
city_athens	4.296603	city_london	6.489250
dist	4.179046	guest_satisfcation_overall	2.196060
metro_dist	1.667465	if_weekends	1.003876

Figure A8: VIF values to test for multicolinearity in the full dataset model

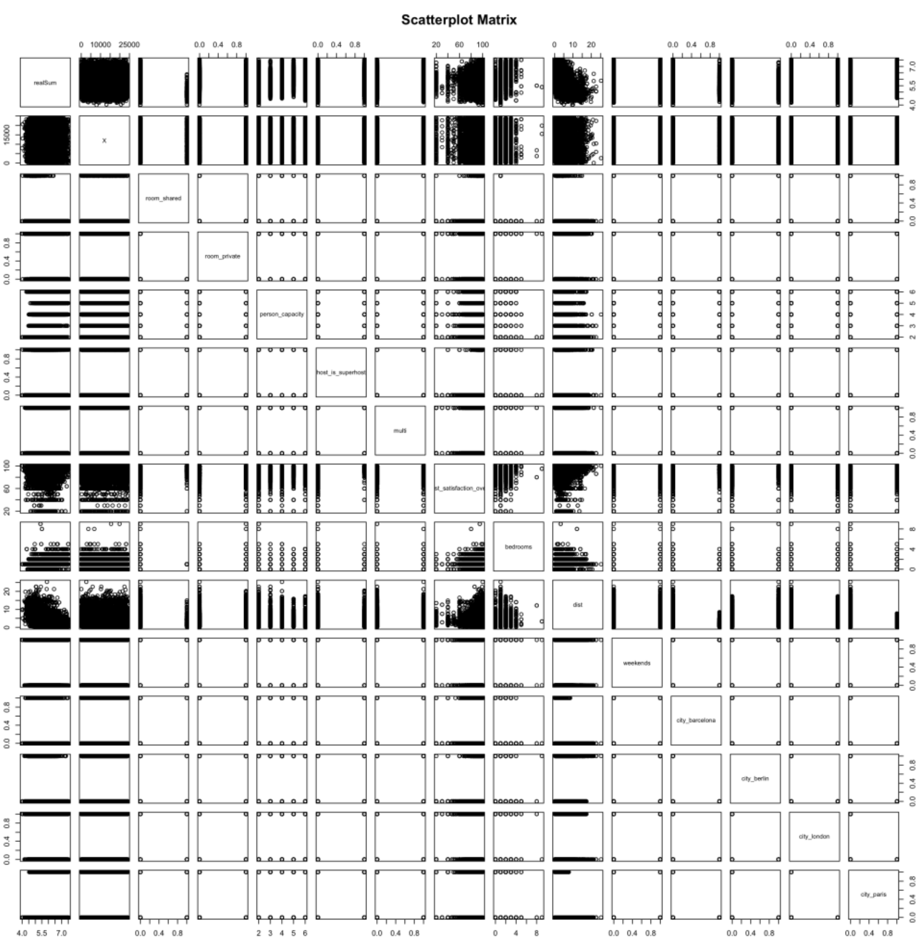


Figure A9: 500 data points all predictors Scatterplot Matrix

Table A2: Results of Partial F-test

Model	R^2	Adjusted R^2	MSPE	MAE	MAPE	PM
MLR-10-full (Log transformed)	0.6597	0.6595	0.1204	0.2580	0.0474	1.610e-06
MLR-10-reduced-FW (Log transformed)	0.6571	0.6569	0.1209	0.2587	0.0475	1.617e-06

500 datapoints model

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.838697   0.301604  16.043  < 2e-16 ***
room_shared                 -1.066633   0.283740  -3.759 0.000192 ***
room_private                -0.456128   0.045303 -10.068  < 2e-16 ***
person_capacity              0.069338   0.019525   3.551 0.000422 ***
host_is_superhost            0.025945   0.042207   0.615 0.539031
multi                        0.051489   0.045075   1.142 0.253900
biz                          0.156355   0.044683   3.499 0.000510 ***
cleanliness_rating           0.051630   0.030228   1.708 0.088283 .
guest_satisfaction_overall   0.001617   0.003415   0.473 0.636163
bedrooms                     0.183384   0.034299   5.347 1.39e-07 ***
dist                        -0.018313   0.015751  -1.163 0.245560
metro_dist                  -0.013908   0.026832  -0.518 0.604458
attr_index_norm              0.224571   0.050488   4.448 1.08e-05 ***
city_athens                 -1.446775   0.114362 -12.651  < 2e-16 ***
city_barcelona              -0.603378   0.102572  -5.882 7.61e-09 ***
city_berlin                 -0.768395   0.119876  -6.410 3.50e-10 ***
city_budapest               -1.292050   0.098698 -13.091  < 2e-16 ***
city_lisbon                 -0.845853   0.103713  -8.156 3.10e-15 ***
city_london                 -0.507462   0.103698  -4.894 1.36e-06 ***
city_paris                  -0.511159   0.094169  -5.428 9.08e-08 ***
city_rome                   -1.009976   0.087192 -11.583  < 2e-16 ***
city_vienna                 -0.956480   0.103166  -9.271  < 2e-16 ***
if_weekends                  0.034716   0.034668   1.001 0.317160

Residual standard error: 0.38 on 477 degrees of freedom
Multiple R-squared:  0.6493,	Adjusted R-squared:  0.6331
F-statistic: 40.14 on 22 and 477 DF,  p-value: < 2.2e-16
```

Figure A10: 500 data points all predictors model

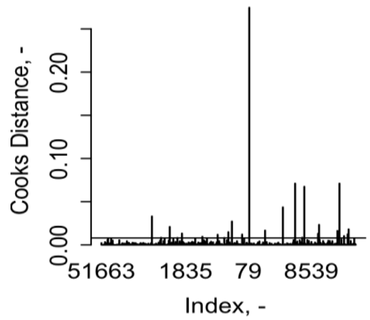
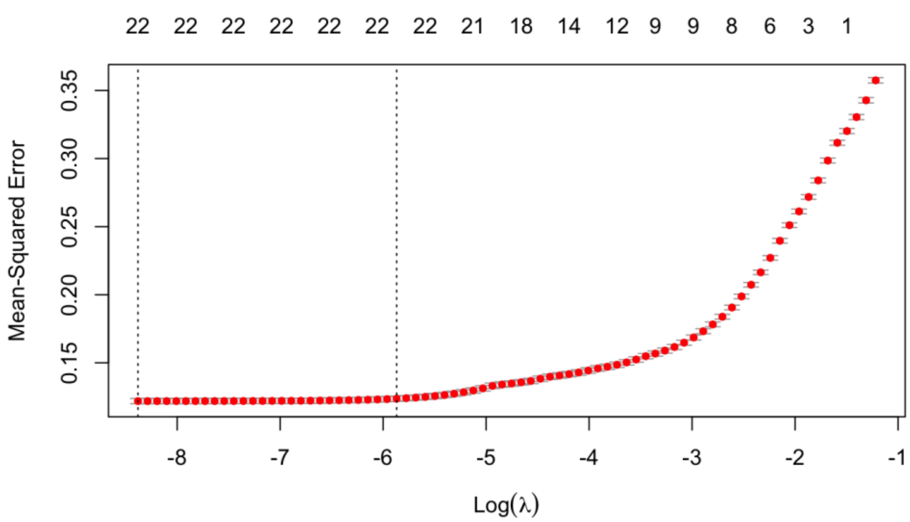


Figure A11: Cook's Distance for 500 datapoint reduced model

Model feature selection

A. LASSO (best lambda = 0.00023) with cross-validation on 100 lambda values



LASSO with the optimal lambda value also yields the same model as using all the predictors on the full dataset, therefore, no further analysis was performed.

B. Backward selection

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	dist	0.6597	0.6595	21.3338	26532.1221	0.3490

Backward selection removed the predictor, dist. As the forward selection removed this predictor as well, the reduced model from the forward selection was chosen for further analysis.

Reference

[1] Gyódi, K., & Nawaro, Ł. (2021). Determinants of Airbnb prices in European cities: A spatial econometrics approach. Tourism Management, 86, 104319.