Using Machine Learning to Predict Airbnb Pricing in Austin, TX

By:
Luis Botello
Lorenzo Dube
Albert Joe

For:
STA 380.10: Mathematical Statistics for Applications
Prof. Jared Murray

December 5, 2022

Introduction

Airbnb was founded in 2007 by two roommates who could not afford rent in San Francisco and decided to host people attending a big design conference as a way to supplement their income. What started as a side hustle has grown into a platform with over 4 million hosts who have welcomed more than 1 billion guest arrivals [(2)].

Evidently, the Airbnb ecosystem has evolved significantly since its initial inception in 2007. What began as an opportunity for people to list unused rooms to supplement their income has evolved into a full time profession for many. We aim to provide hosts and potential investors insights into the greatest contributing factors to Airbnb pricing in Austin, TX to guide their business decisions. Furthermore, for those who already have properties they want to list, we built a prediction model that recommends a listing price based on the attributes of a host's property.

As expected, we found that the property location had the largest effect on listing prices. Some attributes with an unexpected large effect on listings prices were whether pets were allowed and whether or not a listing had a jacuzzi. Attributes that surprisingly had no effect included whether a host's identity was verified, whether a listing had parking, and whether a listing had self check-in.

Methods

Data Sources

The data were obtained from *Inside Airbnb*, a project that provides data and advocacy about Airbnb's impact on residential communities [(1)]. Although Airbnb is a dynamic marketplace where prices and availability fluctuate daily, our data are not. The data were scraped directly from *Airbnb's* website by *Inside Airbnb* in September 2022. The original dataset contained 18,337 listings. However, some records were deprecated during the cleaning process due to the prevalence of null values.

A list of amenities was cast into dummy variables, which include the following: *pool, fireplace, jacuzzi, self-check-in, parking on premises, lake access, sauna, pets allowed, gym, washer & dryer*. This list was guided by Aibnb's list of amenities that guests want the most when booking an airbnb. Furthermore, each listing was classified into one of the following categories: *private_room, shared_room, camper_rv_boat, outdoor, condo_apartment, house*. The cleaned data frame contains other basic information about each listing such as: *latitude, longitude*, *zip code*s, *neighborhood*, l*isting id*, number of *bedrooms*, number of *bathrooms*, and how many guests the listing can *accommodate*. The cleaned data set contains 16,339 records and 22 variables.

Exploratory Data Analysis

We first analyzed the distributions of some key numeric variables, shown in **Table A1**. All variables look skewed to the right, with the exception of *review_scores_rating*, which looks skewed to the left.

Table A2 shows the specific quantiles that can be observed from **Table A1**. We note that 95% of our data points have 3 or fewer *bathrooms* (max: 8 bathrooms), have 4 or fewer bedrooms (max 12 bedrooms), and have prices below $604 (max: $999). This indicates the presence of outliers, which could lead to bias in our regression results.

Next, we examined price against some variables of interest. As expected, *price* generally increases with an increase in *bedrooms*, *bathrooms*, and *accommodates* (refer to **Table A3**). We see that these relationships seem approximately linear. However, this relationship breaks down when listings have more than 7 bedrooms or more than 6 *bathrooms*. One possible reason is that there are few data points past these values, as can be seen from the scatterplot on **Table A3**. This may also indicate the presence of confounders.

Upon further examination, we found that some outliers were hosts trying to game the system. They would list the first two unbooked days for really low prices and then dramatically increase prices for all subsequent days. As a result, when users filter by price, that host's listing would come out first. However, when users clicked on the dates they actually wanted to book, the price would increase by as much as about 200% on weekdays and 1000% on weekends and holidays (refer to **Table A5**). Few listings did this so we chose to take these outliers out of our models.

Table A4 shows some other variables' relationship with *price*. We see that different values in *property_type_cleaned* and downtown seem to affect price. In this table, it is also interesting to see that *number_of_reivews* has a slight negative correlation with price, and many listings have 0 reviews. One possible explanation for this is that listings with high prices are booked less which results in lower numbers of reviews.

Analysis on the *price* variable led us to log transform this variable in our models. As seen in histograms and qqnorm plots in **Table A6**, taking the log transformation of *price* made the distribution of this variable approximately normal. Furthermore, when running multiple linear regression, using the log transformed variable consistently gave higher r-squared values by about 0.1.

Another variable we wanted to further analyze was *downtown*. **Table A7** shows listings that are downtown, downtown adjacent, and not downtown have different slopes when looking at the relationship between *bedrooms* and *price*. This finding led us to include interactions between *downtown* and *bedrooms* in some of our regression models.

<u>Simple Linear Regression and Numerical Value Analysis</u>

We performed simple linear regression of each continuous numerical variable on price. Each of the parameters was put through regression with linear, squared, cubic, and spline models. In all instances, the improvement of the fits using more complicated models did not warrant application of more complex models beyond the linear model.

Table A8 displays the correlation between each of the numerical parameters (including price), the distribution for each variable, and scatterplots with regression lines for each combination of variables. The number of people that a listing *accommodates*, number of *beds*, and number of *bedrooms* available are strongly correlated with each other and expectedly showed similar relationships with *price*. Because of this, we recognized these as potential confounding variables to consider later in multiple linear regression. They also showed a potential nonlinear trend in the residuals vs. predictor variable plot. However, summary statistics for nonlinear models did not improve R^2 or RSD enough to justify more complicated models as seen in **Table A9**.

Table **A10** shows the diagnostics between *accommodates* and *price*, the plot of residuals vs predictors shows a shift from more positive to more negative, indicating a potential non-linear trend. Additionally, the dated and lag plots of the residuals indicate a correlation within the residuals of (0.103), this is low, and 0 correlation cannot be expected, but it's still worth keeping in mind as another potential indicator of non-linearity. In the absolute value residuals vs. predictor, there is a clear increase in the mean of the residuals as the predictor increases - this is in conflict with our assumption of homoscedasticity in the linear model. Lastly, the residuals show a longer tail in the positive direction, in conflict with our assumption of gaussian residuals. This can be mitigated by calculating a confidence interval for coefficients using bootstrap. Considering these model diagnostics, we may better understand the relationships when considering other factors in a multiple regression.

Additionally, the *number_of_reviews* and *review_scores_rating* showed weak correlation(<0.05) to the price, however, the relationship showed a strong statistical significance (<0.001). This indicated that while these are poor individual predictors of price, they may still be important variables to consider in a larger model.

Multiple Linear Regression (MLR)

Based on our exploratory data analysis from the previous sections, we used 3 subsets of data and trained 4 models for each subset, totalling 12 models. As explained in the previous section, we regressed all models using the log transformation of price.

We created the first 2 subsets because the *review_scores_rating* variable had 2,953 NA values (2,953 listings did not have reviews). Subset 1 imputes the NA values of *review_scores_rating* with the mean, allowing us to keep all 16,339 data points. Subset 2 removes the 2,953 data points that do not have reviews, allowing us to better see the real effect of *review_score_ratings* on the log of price. Lastly, subset 3 includes only listings with reviews where the property types are houses, apartments, or condos. Most hosts would not be listing boats, RVs, ranches, tents, etc. so subset 3 allows us to focus on the main property types.

For each subset, we trained 4 models: a simple model, a complex model, an interactions model, and a polynomials model. The simple model includes only the

variables *accommodates*, *bedrooms*, *neighborhood*, and *bathrooms*. The complex model includes all the variables marked 'used in regression' in our variable list. The interactions model takes the complex model, but adds the interaction *bedrooms*neighborhood* (discussed in the previous sections). This model also adds the interaction *bedrooms*accommodates* because from the common sense perspective, listings that accommodate 12 people may not be appealing if it only has 2 bedrooms. The last model is the polynomial model in which we regress *accommodates*, *bedrooms*, and *bathrooms* as polynomials.

Model Selection

In building a predictive model for price, all models previously described were considered. In addition, a stepwise selection model was tested building up from the simple model to include all potential parameters in the interactions model, and then a step down model was built starting with the polynomials model. In evaluating the models, 10 randomized train-test iterations were performed at an 80%-20% split. Root mean square error was utilized to evaluate the performance of the models.

Results

Effects of Attributes on Airbnb Listing Prices

First, we note that our regression result estimates and standard errors (**Table B1**) had no significant differences to the results from our bootstrapped confidence intervals **Table B2**). Also note that we chose the Complex Model of data that had reviews (subset 2 from the MLR section) to demonstrate effects, because the r-squared value on this model (r-squared = 0.6464) is only slightly less than the Interactions Model used in our model selection while also providing more easily interpretable results. Also note that because we are regressing multiple variables on the log of price, the interpretation of the estimated coefficient is that (exp(coefficient) - 1)*100 represents the percentage increase in price for a one unit increase in the coefficient. Refer to **Table B1** for the specific results explained in subsequent paragraphs in this section.

On average, holding all other variables in our model constant, listings that were downtown_adjacent resulted in lower listing prices than downtown listings by about 29.5% while listings that were neither downtown nor downtown adjacent resulted in lower listing prices by about 48.8% than those downtown. One significant attribute that Airbnb hosts could invest in are jacuzzis, which resulted in an increase in listing price by about 19.9% (as compared to investing in a pool which is much more expensive and resulted in only an increase in listing price by about 11.2%). Also worthy of note is listings that allowed pets resulted in an increase in listing price by about 53.3%.

Also interesting is that the correlation between the number of reviews and the listing price was negative, and although statistically significant, had almost no real impact on the listing price. However, a one point increase in the rating resulted in higher

listing prices by about 5.8%. The takeaway from this is that hosts should probably not chase more reviews but focus on getting good reviews.

Lastly, we note that some attributes had no statistically significant effects on listing prices in our model: self check-in, parking availability, and host identity verification. This is likely because most listings in Austin, TX have these attributes so any benefits from these are difficult to detect.

Model Selection

The model with the lowest mean(115.0), median(114.8), and interquartile range(112.4-117.4) was the interactions model evaluated with the dataset that removed listings with *review_scores_rating* valuable. The dataset with an imputed *review_scores_rating* did not show a single low quartile below 121.2 rmse, and the dataset with all "shared room" listings removed did not show a predictive ability with a low quartile below 121.4.

This model seems to satisfy typical linear regression assumptions - the residuals stay steady around zero throughout the model and appear gaussian and homoscedastic as well as seen in **Table B3** along with variable importance. Bootstrapping our models confirms that the gaussian and homoscedastic assumptions were met as there were minimal changes to our estimates relative to the CLT based confidence intervals.

Discussion

Our model contains several limitations, some of them are discussed below. There is a practice among some hosts to offer a low daily rate while charging outrageous cleaning fees in order to attract potential customers by the seemingly low daily rate. For instance, Business Insider reported of a listing where the price for a room was $198 for two nights. However, once cleaning fees were included the total came out to be $413; meaning over 50% of the cost to rent that particular room came in the form of cleaning fees [(3)]. Unfortunately, the dataset obtained from *Inside Airbnb* does not contain any information on cleaning fees, only daily rate pricing is provided. It is possible that comparable listings in the dataset have significantly different price points merely because of this practice and our model is unable to detect this.

Another limitation from our analysis comes from the fact that the data obtained from *Inside Airbnb* is a static dataset scraped from Airbnb's website in September 2022. Airbnb is a marketplace where prices fluctuate on a daily basis due to a myriad of factors such as the time of year. Our analysis fails to capture the influence holidays, weekends or other special events might have on the price variable. We suspect holidays and weekends heavily influence prices. Analyzing the influence the time of year has on prices might be the logical next step for someone wanting to further the research.

Bibliography

1. Inside Airbnb: *About*. Available at: http://insideairbnb.com/about/ (Accessed December 4, 2022).
2. Airbnb: *About Us*. Available at: https://news.airbnb.com/about-us/ (Accessed December 4, 2022).
3. Business Insider: *Airbnb is getting roasted by travelers complaining about extensive cleaning fees and rules*. Available at: https://www.businessinsider.com/airbnb-trending-twitter-customers-complain-host-cleaning-fees-rules-price-2021-5?op=1 (Accessed December 4, 2022).

Appendix A: EDA

Table A1: Histogram of Frequencies

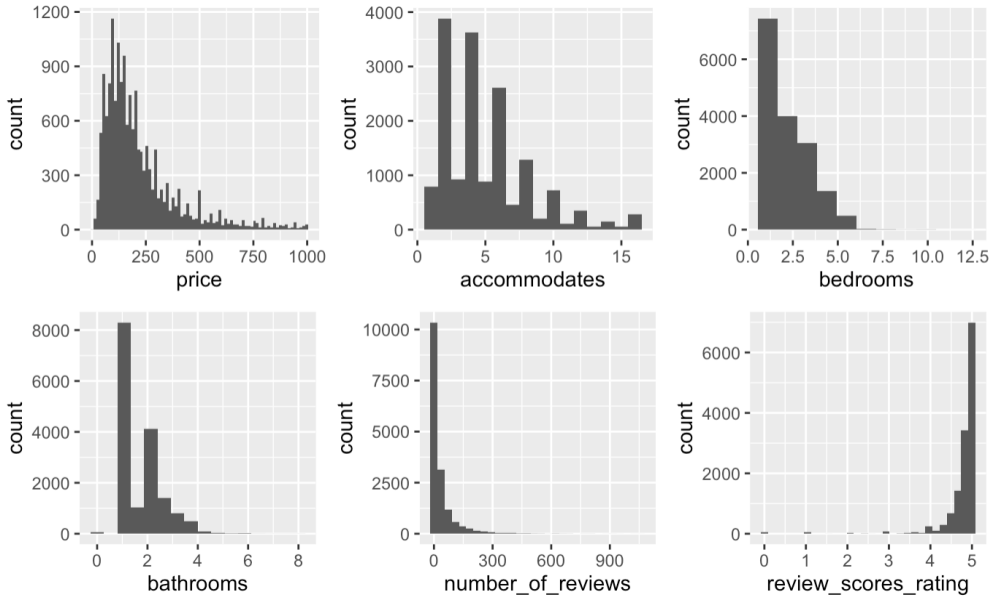


Table A2: Descriptive Statistics

var <chr>	min <dbl>	q25 <dbl>	median <dbl>	q75 <dbl>	q95 <dbl>	max <dbl>	mean <dbl>	sd <dbl>
accommodates	1	2	4	6	12	16	4.993718	3.1947879
bathrooms	0	1	1	2	3	8	1.632959	0.8153995
bedrooms	1	1	2	3	4	12	2.014943	1.1872901
price	1	100	165	275	604	999	220.315504	178.0958600

Table A3: Boxplots and Scatterplots of variables of interest on price

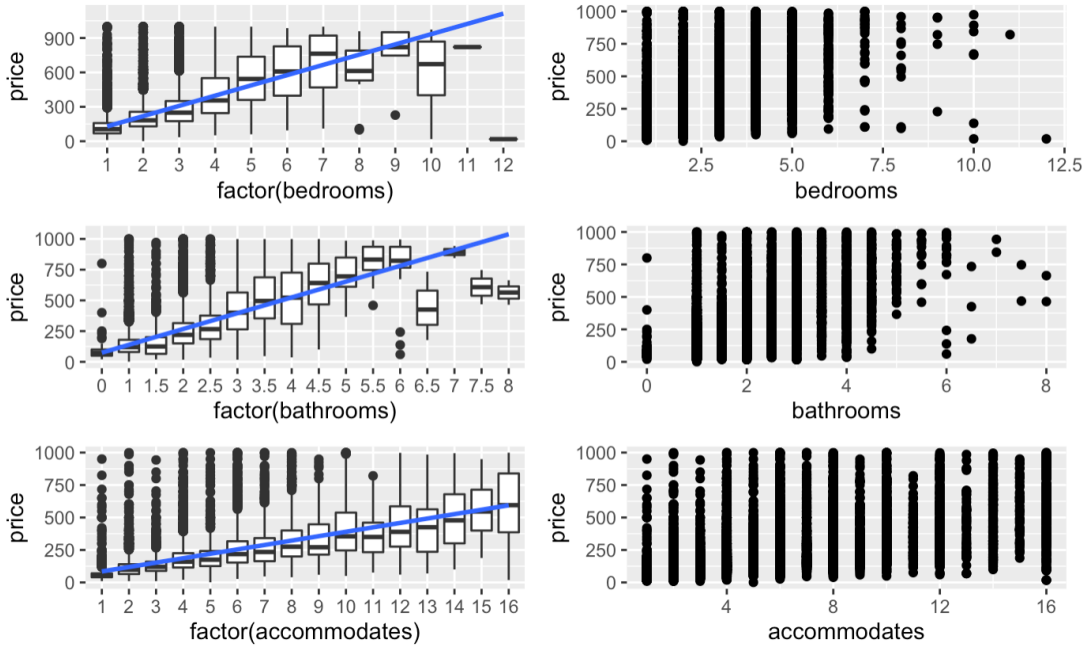


Table A4: Boxplots and Scatterplots of variables of interest on price

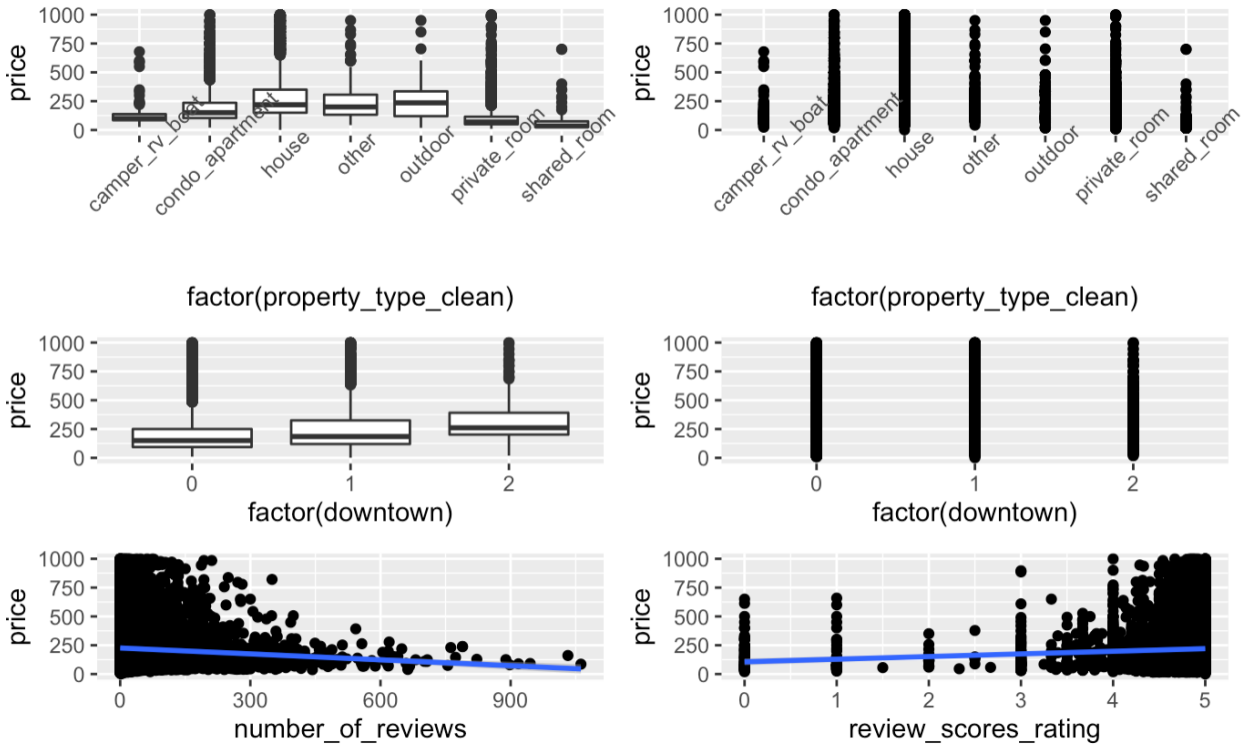


Table A5: Airbnb Gaming System Evidence

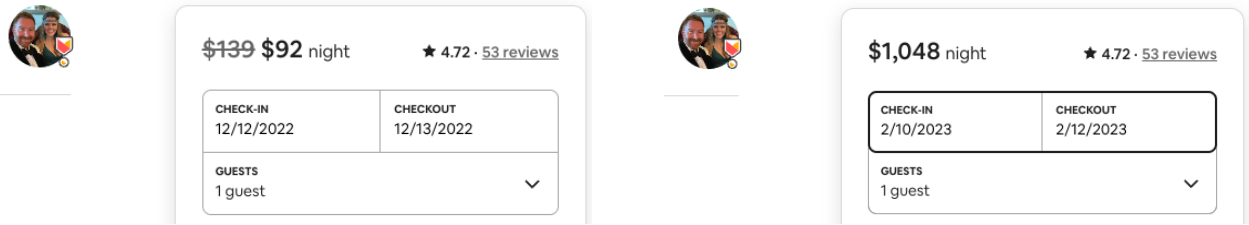


Table A6: Histogram and QQplots of price vs log(price)

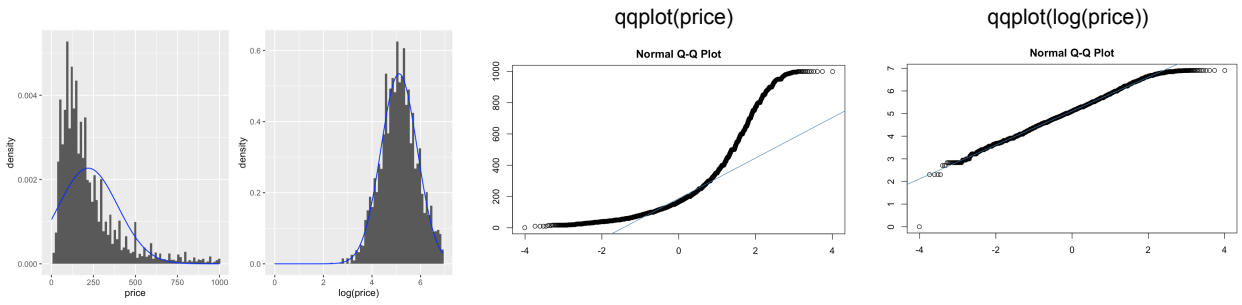


Table A7: Effect on slopes of bedrooms on price accounting for neighborhood

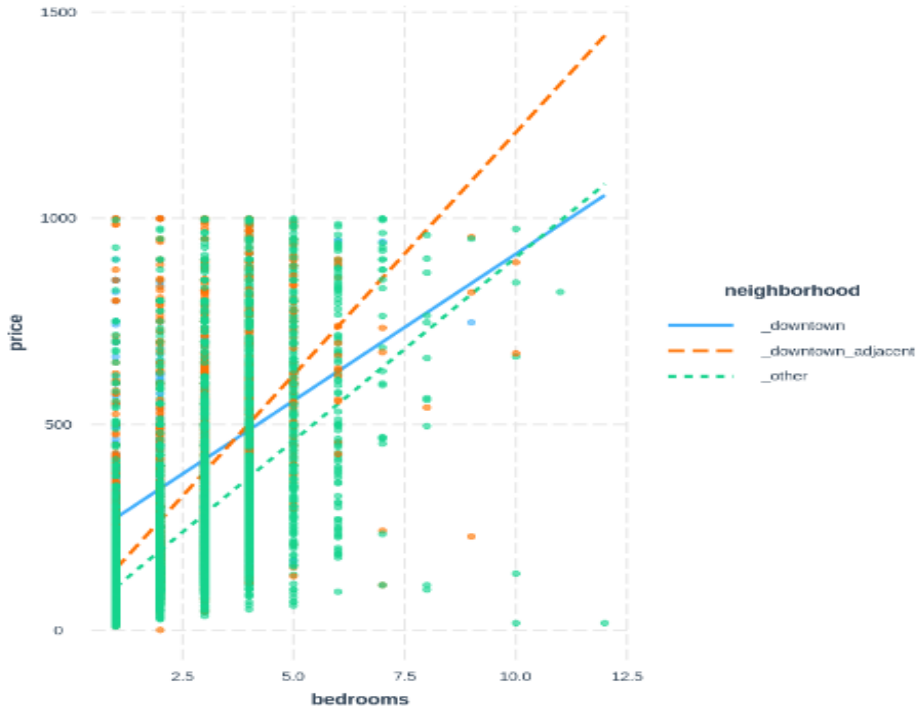


Table A8

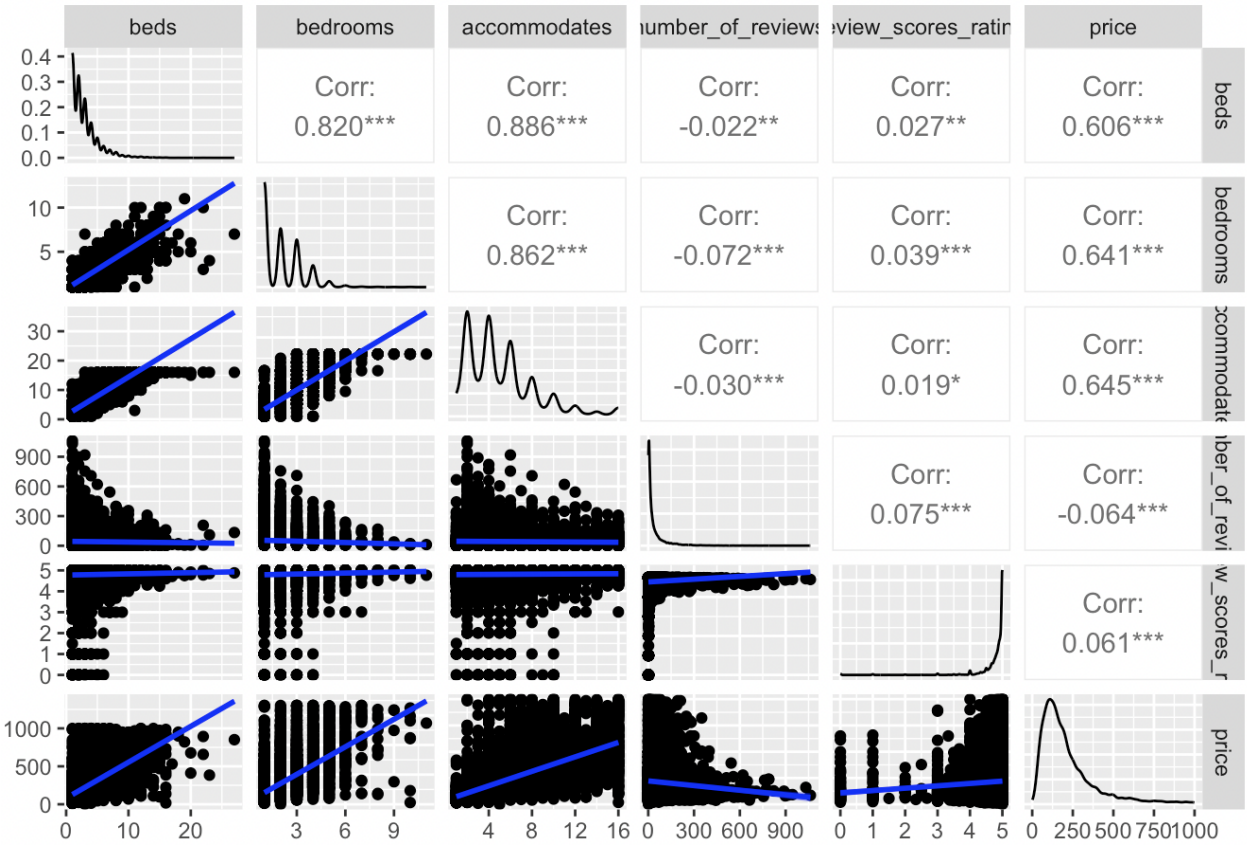
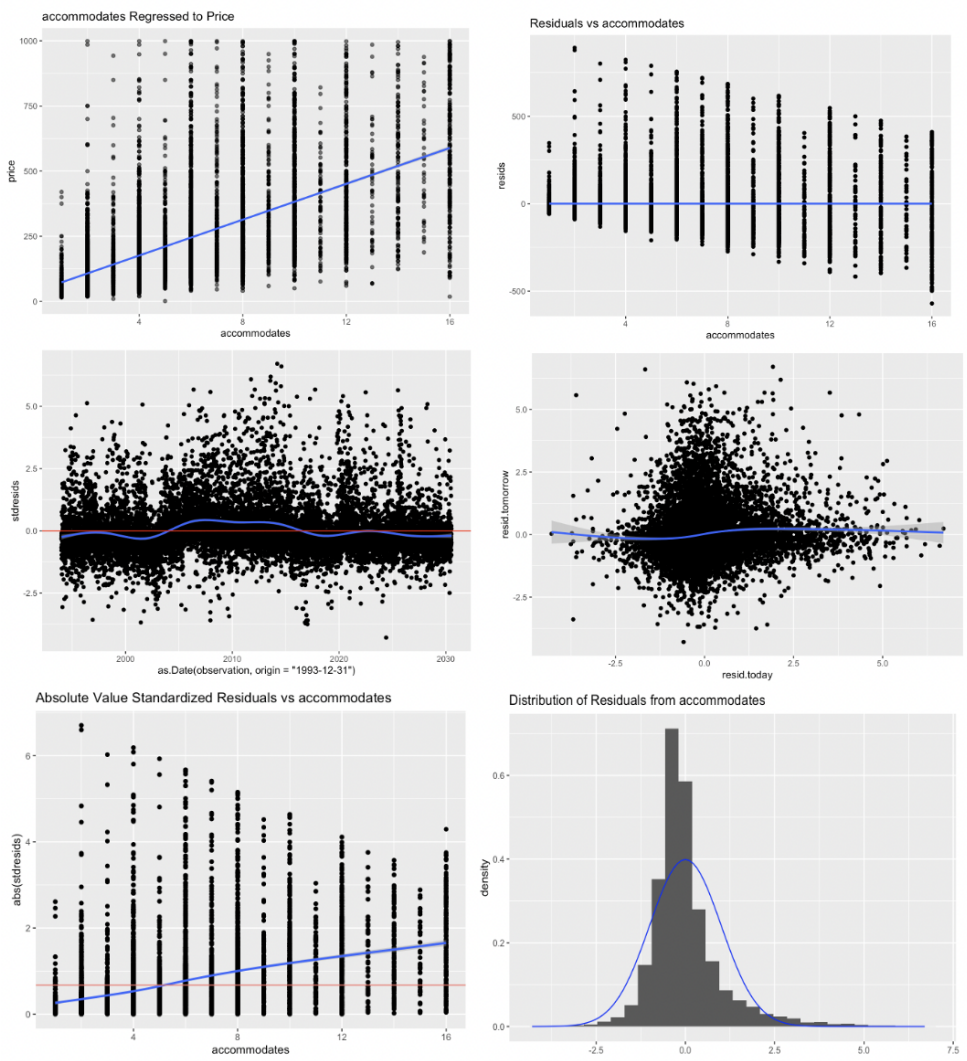


Table A9

```
[1] "Linear Model"

Call:
lm(formula = unlist(new.df[2]) ~ unlist(new.df[1]), data = new.df)

Residuals:
    Min      1Q  Median      3Q     Max
-570.77  -68.26  -23.83   38.78  891.78

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        38.4324     2.1582   17.81   <2e-16 ***
unlist(new.df[1])  34.3962     0.3533   97.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133 on 13342 degrees of freedom
Multiple R-squared:  0.4154, Adjusted R-squared:  0.4153
F-statistic:  9480 on 1 and 13342 DF,  p-value: < 2.2e-16

[1] "Cubic Model"

Call:
lm(formula = unlist(new.df[2]) ~ poly((unlist(new.df[1])), 3),
    data = new.df)

Residuals:
    Min      1Q  Median      3Q     Max
-562.29  -68.50  -23.14   37.86  893.86

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    216.161      1.151 187.751   <2e-16 ***
poly((unlist(new.df[1])), 3)1 12950.206    132.996  97.373   <2e-16 ***
poly((unlist(new.df[1])), 3)2  -269.424    132.996  -2.026   0.0428 *
poly((unlist(new.df[1])), 3)3    55.108    132.996   0.414   0.6786
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133 on 13340 degrees of freedom
Multiple R-squared:  0.4156, Adjusted R-squared:  0.4154
F-statistic:  3162 on 3 and 13340 DF,  p-value: < 2.2e-16

[1] "Squared Model"

Call:
lm(formula = unlist(new.df[2]) ~ poly((unlist(new.df[1])), 2),
    data = new.df)

Residuals:
    Min      1Q  Median      3Q     Max
-560.62  -68.76  -23.15   38.24  893.67

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    216.161      1.151 187.757   <2e-16 ***
poly((unlist(new.df[1])), 2)1 12950.206    132.992  97.376   <2e-16 ***
poly((unlist(new.df[1])), 2)2  -269.424    132.992  -2.026   0.0428 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133 on 13341 degrees of freedom
Multiple R-squared:  0.4156, Adjusted R-squared:  0.4155
F-statistic:  4743 on 2 and 13341 DF,  p-value: < 2.2e-16

[1] "Spline Model"

Family: gaussian
Link function: identity

Formula:
unlist(new.df[2]) ~ s(unlist(new.df[1]))

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   216.16       1.15   187.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                      edf Ref.df    F p-value
s(unlist(new.df[1])) 8.357   8.82 1079  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.416   Deviance explained = 41.7%
GCV =  17675  Scale est. = 17662     n = 13344
```

Table A10


accommodates Regressed to Price
price
accommodates
Residuals vs accommodates
resids
accommodates
stdresids
as.Date(observation, origin = "1993-12-31")
resid.tomorrow
resid.today
Absolute Value Standardized Residuals vs accommodates
abs(stdresids)
accommodates
Distribution of Residuals from accommodates
density

Appendix B: Results

Table B1: MLR of the Complex Model for listings that has reviews

```
Call:
lm(formula = log(price) ~ accommodates + bedrooms + neighborhood +
    bathrooms + number_of_reviews + review_scores_rating + host_identity_verified +
    shared_bath + pool + fireplace + jacuzzi + self_checkin +
    parking + lake_access + sauna + pets_allowed + washer_dryer +
    gym + property_type_clean, data = df_has_review)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2960 -0.2750 -0.0165  0.2558  2.4671

Coefficients:
                                   Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                      4.53727336  0.05748808  78.925 < 0.0000000000000002 ***
accommodates                     0.04273250  0.00246057  17.367 < 0.0000000000000002 ***
bedrooms                         0.12128488  0.00788876  15.374 < 0.0000000000000002 ***
neighborhood_downtown_adjacent  -0.35710712  0.02062477 -17.314 < 0.0000000000000002 ***
neighborhood_other              -0.67132067  0.01981916 -33.872 < 0.0000000000000002 ***
bathrooms                        0.18650140  0.00806513  23.124 < 0.0000000000000002 ***
number_of_reviews               -0.00046647  0.00005318  -8.771 < 0.0000000000000002 ***
review_scores_rating             0.05686657  0.00795520   7.148  0.00000000000092385 ***
host_identity_verified          -0.01604693  0.01041113  -1.541             0.123261
shared_bath                     -0.29272848  0.01783745 -16.411 < 0.0000000000000002 ***
pool                             0.10595847  0.00996596  10.632 < 0.0000000000000002 ***
fireplace                        0.02413472  0.00969384   2.490             0.012797 *
jacuzzi                          0.18158794  0.01329525  13.658 < 0.0000000000000002 ***
self_checkin                    -0.00672023  0.00846583  -0.794             0.427322
parking                          0.01710800  0.01517376   1.127             0.259563
lake_access                      0.14268656  0.01344789  10.610 < 0.0000000000000002 ***
sauna                            0.08008061  0.05080242   1.576             0.114977
pets_allowed                     0.42718702  0.06138839   6.959  0.00000000000359287 ***
washer_dryer                    -0.12293230  0.01180791 -10.411 < 0.0000000000000002 ***
gym                             -0.02554940  0.01386360  -1.843             0.065364 .
property_type_cleancondo_apartment  0.09617598  0.03723452   2.583             0.009806 **
property_type_cleanhouse         0.31152803  0.03649307   8.537 < 0.0000000000000002 ***
property_type_cleanother         0.40123174  0.05039656   7.961  0.00000000000000184 ***
property_type_cleanoutdoor       0.52925027  0.05582956   9.480 < 0.0000000000000002 ***
property_type_cleanprivate_room -0.13481737  0.03743998  -3.601             0.000318 ***
property_type_cleanshared_room  -0.73832173  0.05862485 -12.594 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4382 on 13360 degrees of freedom
Multiple R-squared:  0.6464,    Adjusted R-squared:  0.6458
F-statistic:   977 on 25 and 13360 DF,  p-value: < 0.00000000000000022
```

Table B2: Bootstrapped Confidence Intervals of the Complex Model that has reviews

name <chr>	lower <dbl>	upper <dbl>	level <dbl>	method <chr>	estimate <dbl>
Intercept	4.4009005900	4.661147881	0.95	percentile	4.5372733608
accommodates	0.0373135069	0.048046077	0.95	percentile	0.0427324973
bedrooms	0.1050842387	0.138048882	0.95	percentile	0.1212848837
neighborhood_downtown_adjacent	−0.3976778227	−0.313682081	0.95	percentile	−0.3571071173
neighborhood_other	−0.7093555098	−0.634079006	0.95	percentile	−0.6713206703
bathrooms	0.1704336590	0.203226589	0.95	percentile	0.1865014008
number_of_reviews	−0.0005630285	−0.000369941	0.95	percentile	−0.0004664678
review_scores_rating	0.0364604332	0.078618848	0.95	percentile	0.0568665674
host_identity_verified	−0.0369681432	0.005205518	0.95	percentile	−0.0160469334
shared_bath	−0.3290478602	−0.254300693	0.95	percentile	−0.2927284806

name <chr>	lower <dbl>	upper <dbl>	level <dbl>	method <chr>	estimate <dbl>
pool	0.0868024940	0.125059496	0.95	percentile	0.1059584707
fireplace	0.0058197620	0.042505497	0.95	percentile	0.0241347171
jacuzzi	0.1557784950	0.207985926	0.95	percentile	0.1815879374
self_checkin	−0.0221177444	0.008404321	0.95	percentile	−0.0067202299
parking	−0.0162172739	0.047320955	0.95	percentile	0.0171079968
lake_access	0.1184632194	0.167631381	0.95	percentile	0.1426865569
sauna	−0.0103650645	0.164289511	0.95	percentile	0.0800806063
pets_allowed	0.2647838040	0.607817806	0.95	percentile	0.4271870219
washer_dryer	−0.1463263984	−0.100150993	0.95	percentile	−0.1229323040
gym	−0.0517708091	0.003303913	0.95	percentile	−0.0255493961

name <chr>	lower <dbl>	upper <dbl>	level <dbl>	method <chr>	estimate <dbl>
property_type_cleancondo_apartment	0.0352304730	0.156911691	0.95	percentile	0.0961759844
property_type_cleanhouse	0.2529931073	0.372772281	0.95	percentile	0.3115280256
property_type_cleanother	0.3090140991	0.486334439	0.95	percentile	0.4012317366
property_type_cleanoutdoor	0.3815995095	0.676271353	0.95	percentile	0.5292502746
property_type_cleanprivate_room	−0.1963279379	−0.076181825	0.95	percentile	−0.1348173677
property_type_cleanshared_room	−0.8680597924	−0.602889810	0.95	percentile	−0.7383217318
sigma	0.4297113982	0.446296580	0.95	percentile	0.4382106332
r.squared	0.6342574811	0.659491191	0.95	percentile	0.6464225833
F	926.7372034225	1035.016088804	0.95	percentile	977.0087460790

Table B3: Diagnostics/Variable Importance for Interactions Model with Data that has Review Scores