Predicting the Price of Austin Airbnbs

Linearity Legends - Barron Brothers, Jack Steel, Mert Bildirici, Kethan Poduri

2023-12-01

Introduction and Data

Airbnbs have historically been known as a cheaper (and better) alternative to hotels. As the Airbnb market has evolved, Airbnbs have typically been getting more expensive for solo travelers and couples. If many people rent an Airbnb for long periods of time, though, it could be very economical.

When people are travelling, it is important for them to get the best blend of price and experience possible. These two factors can be difficult to balance, though. As an Airbnb is more highly rated, the owner could decide to raise the price due to higher demand. People can also rate on multiple factors, such as the availability, the size of the room rented, and/or the location.

The research question we will focus on is as follows: For Airbnbs in Austin, TX, **What factors most influence the price on an Airbnb, and how does that compare to the customer's priorities?**

We predict that as availability decreases, the number of reviews a listing has increases, and as room type becomes larger and more private, and customer ratings become more positive, price will increase.

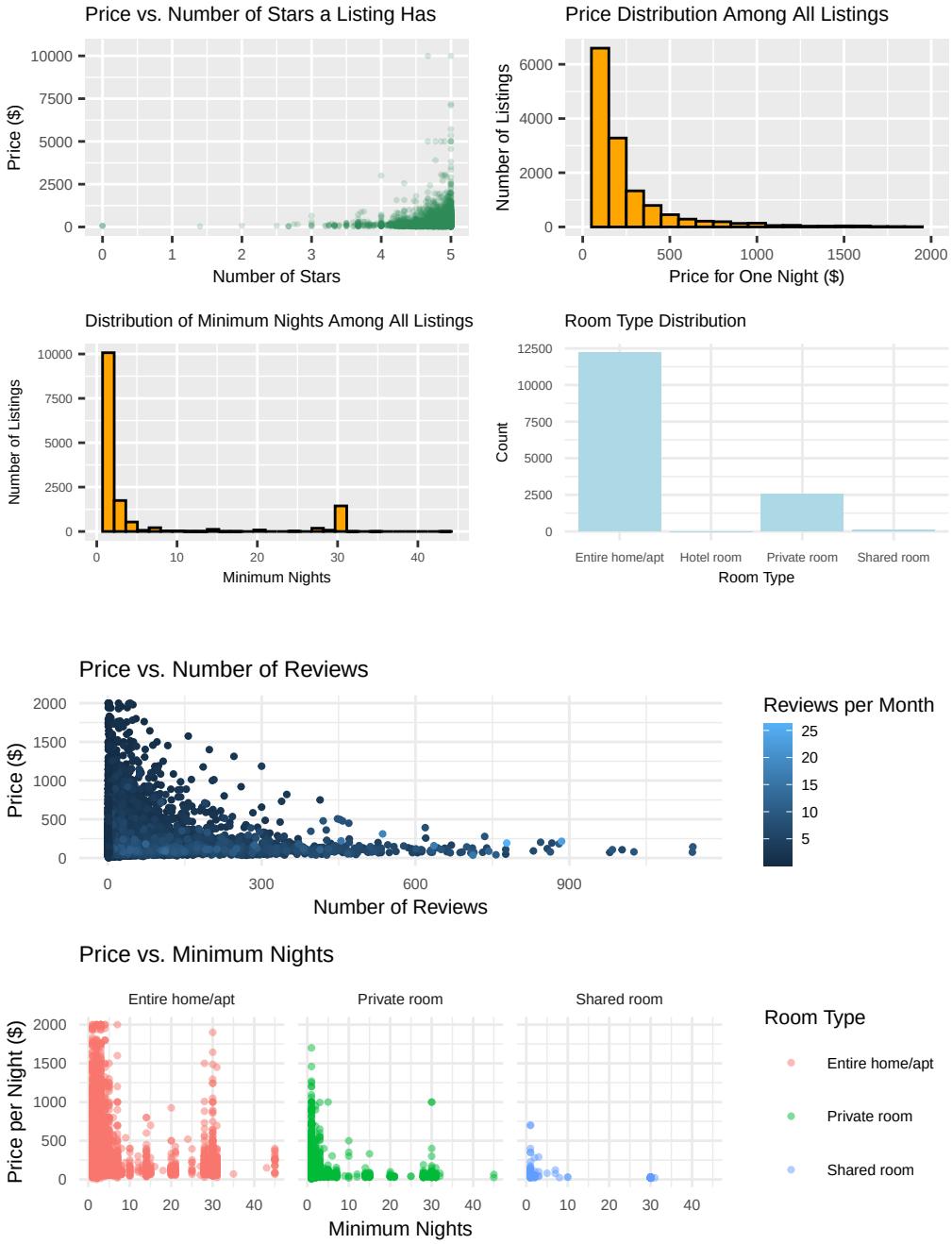
Data

The source of the dataset is here. While the data itself is displayed on Kaggle, the data is derived from a clear external source: InsideAirbnb. Our team has cross-validated the data to make sure that it is accurate. We found, from the InsideAirbnb website, that the data contains public information about the listing, such as the number of reviews a listing has, its rating, its location within the city, and its availability. Furthermore, we found that InsideAirbnb also further verifies the accuracy of the data and cleanses it themselves. The dataset itself is updated frequently, with the last update taking place on 07 Oct. 2023. The data have been collected since 2020, but prior dates are included.

There are 14,681 observations in the dataset, 4,381 of which have `number_of_reviews` ≥ 30. Generally, many of the useful variables have to do with the ratings of an observation (one Airbnb), the location of the observation, or describing the observation itself (i.e. `price`, `name`, … etc.) The data columns can be described as follows:

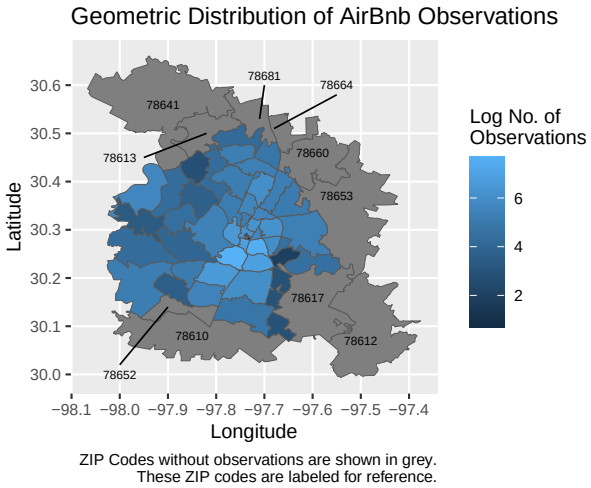
- `id`: The unique identifier for each Airbnb
- `name` contains multiple variables, listed in order:
  - Type of home rented
  - Location (We are only using Austin)
  - Rating (i.e. " 4.84")—renamed `stars`
  - No. of bedrooms
  - No. of beds
  - No. of baths
- `host_id`: The unique identifier for each host
- `neighbourhood`: The ZIP code that the Airbnb is in
- `room_type`: What was rented (i.e. "Entire home/apt", "Private room", … etc.)
- `price`: The daily price in local currency. Since we are only using Airbnbs in Austin, TX, this value will be in $.
- `minimum_nights`: The minimum amount of nights that a person can rent out the Airbnb.
- `availability_365`: The number of days per year that the host lists the Airbnb as available for rental.
- `number_of_reviews`: The number of reviews for the Airbnb.
- `reviews_per_month`: The number of reviews per month for an Airbnb.
- `income_level`: The income level of the neighborhood based on Median House Income for the Zipcode (feature-engineered)

EDA

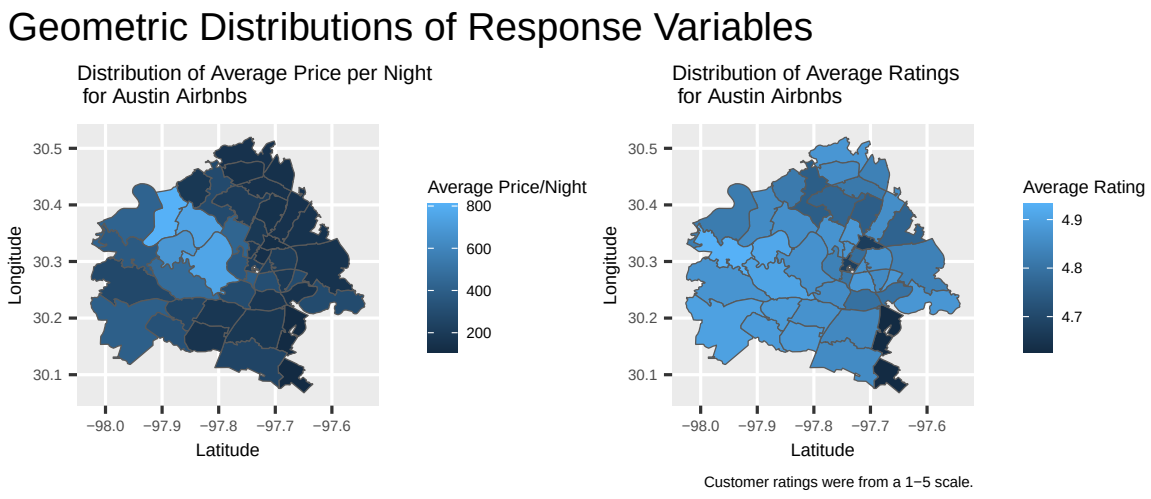


The price is right-skewed, with most listings at \$0-200. Also the room type is unimodal, with most rooms as entire home/apartments and some as private rooms. There are not many

observations on shared rooms and hotel rooms. Also, the number of stars is left-skewed. This is understandable as most people who leave stars either hated or loved their stay. The host has extra incentive to aim for the latter since a couple of bad reviews can make an Airbnb with relatively few ratings stand out.



Most of the listings are concentrated in Central and Eastern Austin. Log transform was used to improve clarity.



From the maps above we can see that the average price per night is the greatest in Airbnbs located in the center and northwest sectors of the city. Additionally, the maps indicate that Airbnbs are rated relatively highly, with an average rating of around 4.7 stars, throughout the city. However, it is important to note that two ZIP codes in the SE have the worst-rated Airbnbs in the city, with an average rating of approximately 4.6 stars. It is worth clarifying, though, that these two ZIP codes have relatively few observations.

Methodology

The goal of this model is to find the best estimate of an Airbnb's price based on the neighborhood, room type, reviews per month, available days per year, the number of stars customers gave, how many listings the host has, minimum nights you could stay at the listing, number of reviews, and number of reviews in the last twelve months. A linear model is appropriate since the predicted outcome is a numerical value (`price`), and not binary. Different room types might attract people from different backgrounds and might be in different price ranges. Also, the neighborhood that the listing is in—whether in an area with a high or low median household income—might affect the price, as Airbnbs in areas where people have more disposable income are expected to be more expensive. Fewer available days per year could result in a decrease in price due to a lack of customer repute about the location. An increased number of listings the host has could indicate a more experienced host, plausibly leading to a higher price. An increased number of reviews could show how popular the listing is, plausibly attracting more customers and thus increasing the price. The ratings themselves are also crucial—a higher rating could indicate higher demand for the listing, resulting in an increase in price. A listing's minimum night policy could be correlated with its yearly availability as listings with long stays limits the availability of said listings. An interaction effect exploring this could potentially be useful. The number of reviews in the last twelve months could also be correlated to the total number of reviews, so precautions should be taken to avoid potential multicolinearity.

Even without the neighborhoods in the Austin metro area that have no observations, the dataset still includes 42 neighborhoods, so the model could potentially overfit to the specific characteristics of each neighborhood, which would sacrifice generalizability to other cities. To solve this, rather than using each neighborhood as a predictor, the model includes faceted ZIP code levels based on average regional income: High-income ($106,001+), Upper-midddle-income ($80,001-106,000), Lower-middle-income ($60,001-80,000), and Low-income ($0-60,000) based on Austin's geographic socioeconomic data. This was achieved using `step_mutate()`. Subsequently, `neighbourhood` was removed with `step_rm()`, and the faceted levels were saved to `income_level()`. For this model, we assume that the incomes in the faceted ZIP codes are relatively consistent; i.e. within a ZIP code that's classified as lower-income, there isn't a higher-income neighborhood.

For the project, we decided to focus on the listings in Austin for two reasons. Out of 14861 listings, 14650 of them are in Austin, so the dataset lacks significant observations from other

locations. Additionally, `price` is listed in local currencies, which would add another layer of unnecessary complexity since currency exchange rates can and do change. For example, the Turkish lira constantly fluctuates against the dollar, which makes it difficult to convert the price on listings in Turkey to dollars.

To improve model accuracy, we also filtered out any observation in `price` > 500, which was approximately the outlier boundary of 1.5*IQR. (See Additional EDA Plots)

We first created a standard full model with `income_level`, `room_type`, `reviews_per_month`, `availability_365`, `stars`, `calculated_host_listings_count`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_ltm`. Seeing a pattern in the residuals, that violated linearity, we opted for a log-transformed linear model, which corrected this. (See the residual plots below.) Every predictor was well within the margin for statistical significance ($p_{x_i} \approx 0 \ll \alpha$) except for `number_of_reviews` and `number_of_reviews_ltm`, which had $p = 0.0724 > .05$ and $p = 0.7594 \gg .05$, respectively. To assess the validity of these predictors, we conducted a drop-in-deviance test, where:

$$H_0 : \beta_{\text{number of reviews}}, \beta_{\text{reviews in last 12 mo.}} = 0$$

$$H_A : \beta_{\text{number of reviews}} \text{ OR } \beta_{\text{reviews in last 12 mo.}} = 0$$

The results of the drop-in-deviance test are as follows:

Table 1.1: Drop-in-Deviance Test Results for Red. vs. Full Models

Reduced deviance	Full deviance	Test statistic	P-value
1613.424	1609.443	3.98074	0.13664

Below are some additional helpful comparative statistics between the two models:

Table 1.2: Red. vs. Full Comparitive Model Statistics

Model	AIC	BIC	R.squared	Adj.R.Squared	RMSE
Reduced	9452.197	9533.717	0.36441	0.36344	6579
Full	9439.918	9535.024	0.36598	0.36482	6577

Since $p = .1366 > \alpha$, we do not have enough evidence to reject H_0. In other words, we do not have enough evidence that the terms $\beta_{\text{number of reviews}}, \beta_{\text{reviews in last 12 mo.}}$ are useful to our model. Therefore, we will proceed with the reduced model.

The comparative model statistics are not cohesive. The BIC favors the reduced model over the full, and the AIC favors the opposite. This makes sense as BIC tends to favor more

streamlined models, and heavily penalizes models with more than 8 predictor terms. The Adj. R^2 decreases slightly and the $RMSE$ increases slightly, indicating that the reduced model has slightly more error and explains less of the variability in the data. To favor parsimony corresponding with the BIC and the drop-in-deviance test, the reduced model is chosen.

We also attempted to build an interaction-fit model. We started with all of the terms interacting with each other, which resulted in many terms having high VIF values ($\sim 10^2$ magnitude) that violated multicolinearity (see Table 3.1 in appendix). To attempt to resolve this, these terms were eliminated in successive iterations. By the second iteration, the R^2 of 0.3364 (see Table 3.2 in appendix) was significantly less than the R^2 of the reduced model (0.3644), which had no multicolinearity issues. Therefore, our final model chosen was the reduced model.

Check Conditions

There is no pattern in the residuals suggesting that linearity is violated, so the linearity condition is satisfied. While the fitted values in the residual plot are uneven, there is no obvious fan-shape in the data, so constant variance is satisfied. The distribution of residual values appears to be normal with a center of 0 and a peak at approximately 1000 observations, so normality is satisfied. See residual plots for reference.

Table 2.2: VIF Values for Reduced Model

Variables	Reduced Model
reviews_per_month	1.087242
availability_365	1.174996
stars	1.094667
calculated_host_listings_count	1.277296
minimum_nights	1.019223
room_type_Private.room	1.051847
room_type_Shared.room	1.070688
income_level_Upper.Mid.Income	2.019041
income_level_Lower.Mid.Income	2.180941
income_level_Low.Income	2.066820

For each variable in both models, the $VIF < 10$, so there is no significant correlation between the variables listed. (See the VIF breakdown for the full model in Table 2.1)

The data was collected spatially, but `income_level` was an attempt to account for this. To double-check, a spatial residual plot of the average residual per area was plotted with respect to the observations in that area to give a sense of expected variability. (Areas with a lower number of observations could have a higher variability.) There is a notable outlier at about (-97.4, 30.36), ZIP #78757, which could be investigated further, but isn't any discernable

pattern with the residuals in general. Other than spatial income level, we have no reason to believe that independence was violated with the data collection, so therefore independence is satisfied.

Model Testing

To test how our model performs on new data, we will use the testing data with the following results:

Table 4: Model Evaluation Test Statistics

R.squared	Adj.R.squared	RMSE
0.3504428	0.3493685	0.4991885

As seen above, the R^2 and Adj. R^2 statistics of the model on the testing data are similar to that of the model fit on the training data.

The average magnitude of the errors in the predictions, when transformed to the original scale, will be influenced by a factor of $e^{.499} = 1.647$. This is the factor by which the predicted prices, on average, deviate from the actual prices.

Results

The final reduced model is as follows:

Table 5: Reduced Model Log Terms

Term	Estimate	Std. Error	T-statistic	P-value
(Intercept)	5.25410	0.01591	330.31561	0.00000
reviews_per_month	-0.04277	0.00349	-12.25532	0.00000
availability_365	0.00036	0.00005	7.03791	0.00000
stars	0.27215	0.02576	10.56405	0.00000
calculated_host_listings_count	-0.00126	0.00033	-3.78404	0.00016
minimum_nights	-0.00131	0.00022	-5.99754	0.00000
room_type_Private.room	-0.85280	0.01778	-47.96178	0.00000
room_type_Shared.room	-1.72382	0.07002	-24.61939	0.00000
income_level_Upper.Mid.Income	-0.25020	0.02000	-12.51234	0.00000
income_level_Lower.Mid.Income	-0.24694	0.01909	-12.93781	0.00000
income_level_Low.Income	-0.35728	0.01991	-17.94503	0.00000

The equation for our final model is as follows:

$$\widehat{\log(price)} = 5.254 - 0.043 \times \text{reviews per month} + 3.56 * 10^{-4} \times \text{yearly availability}$$
$$+ 0.272 \times \text{stars} - 1.259 * 10^{-3} \times \text{number of listings host has} - 1.306 * 10^{-3} \times \text{minimum nights}$$
$$- 0.3572 \times \text{Income Level::Low Income} - 0.2469 \times \text{Income Level::Lower Mid Income}$$
$$- 0.2501 \times \text{Income Level::Upper Mid Income} - 0.8528 \times \text{Room Type::Private Room}$$
$$- 1.724 \times \text{Room Type::Shared Room}$$

The intercept of the model portrays a typical Airbnb setting: one with the average number of reviews per month (1.568, so 1-2 effectively), the average yearly availability (about 151 days), the average rating (4.84 stars), the average number of listings of the host (about 11 listings), the minimum number of nights (about a week), in a high-income area of Austin and a rental of the full property space. The average nightly price for this Airbnb is $e^{5.496}$ = \$191.33… depending on the occasion, it might just be better to get a hotel. However, there are some ways around this, as we see with our model.

An Airbnb's price is sensitive to customer ratings. If the customer is willing to risk staying at an Airbnb with a worse rating, then the price decreases significantly. With each star the Airbnb is rated less, the price multiplies by, on average, a factor of $e^{-.272} = .762$, holding all else constant. For the average Airbnb, this would result in a price drop to \$145.77/night.

Unsurprisingly, area is also a key factor here. However, customers might not want to stay in lower-income neighborhoods due to a desire for the potentially more upscale experience associated with living in areas with a greater household income, but, if the customer were to diverge from the typical high-income area Airbnbs are associated with, the price drops dramatically. Compared to a high-income area, an Airbnb in an upper-middle income area in Austin, TX, is estimated to be, on average, $e^{-.2501} = .7787$ times the cost, holding all else constant. Interestingly, this factor is nearly the same for lower-middle income areas: $e^{-.2469} = .7812$. For the same average Airbnb, this would translate to \$148.99/night and \$149.47/night, respectively. It is surprising that the price difference between high-income and upper-middle income neighborhoods is drastic while the price difference between upper-middle income and lower-middle income neighborhoods is negligible.

As we expected, a higher number of nights on a minimum night policy generally results in a cheaper Airbnb. However, the model above suggests that it does so by a surprisingly minimal amount—for each extra minimum night, the price of the Airbnb multiplies by, on average, a factor of $e^{-1.306*10^{-3}} = .9987$. This surprisingly minimal (yet statistically significant) effect shows that a customer might want to look for other options to reduce their Airbnb bill.

Conclusion, Limitations, and Further Discussion

Due to its considerable statistical significance, the log-transformed reduced model including reviews per month, availability out of 365 days (`availability_365`), rating in stars (`stars`),

number of listings the host has (`calculated_host_listings`), minimum nights required (`minimum_nights`), rental room type (`room_type`), and median income level of the ZIP code the Airbnb is in (`income_level`) as predictors was selected. Our model has identified several numerical and categorical factors that influence the price of Airbnb in Austin, Texas. One such example is seen with `income_level`—Airbnbs that fall in notably wealthier regions of Austin tended to be more expensive than Airbnbs in less wealthy neighborhoods. In other words, the coefficients for the income level tended to decrease as the median income level for the neighborhood progressively decreased (from the baseline of high-income neighborhoods), leading to a decrease in price for the Airbnb. Another significant categorical variable is `room_type`, as the average price of a shared room tends to be less than a private room which tends to be less than renting an entire proparty. This may indicate that smaller or shared spaces tend to be associated with smaller rental costs in regards to Airbnbs in Austin.

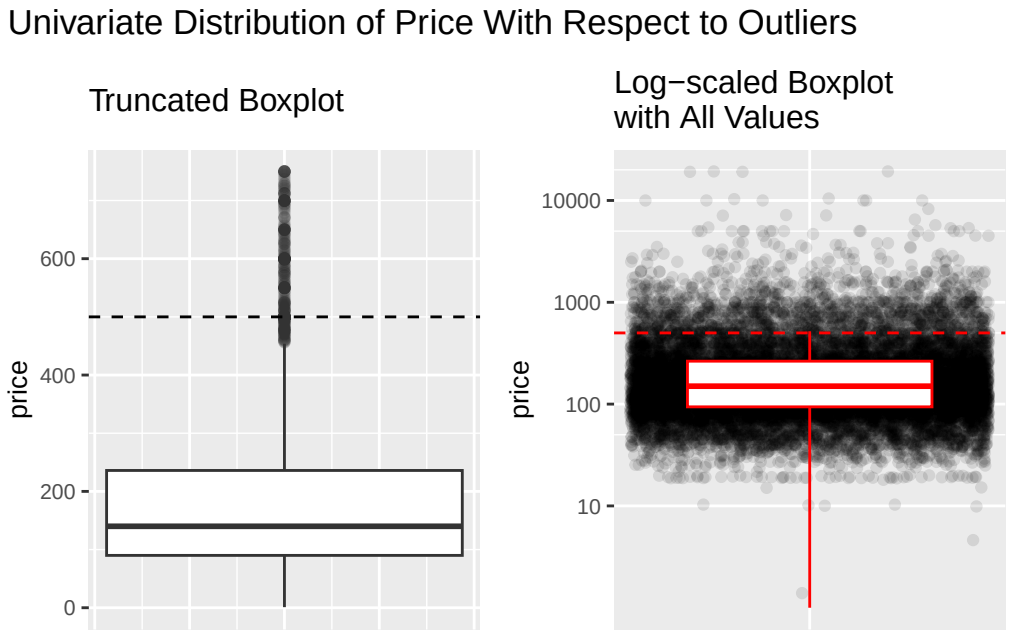
With these results, though, it is important to keep the limitations of our model in mind. One such limitation was the dataframe having a considerably low number of observations for Airbnbs outside of Austin, as well as for Airbnb that were classified as hotels. Since the data frame only had 205 observations outside Austin, Texas (and 14656 observations inside Austin), we decided to remove these 205 observations and generalize our data to the boundaries of the city. Likewise, there were only 6 observations for a room type of "hotel room", which were removed. This limits our ability to generalize both to all Airbnb destinations and other rental types but prevented potential skew in the data. We were unable to distinguish whether the Airbnbs classified as "Entire home/apt" were an entire house, an apartment, or a condo. Being able to decipher between these relatively different housing arrangements would have added complexity to the model that could have improved its predictive power. Extracting additional identifiers from the Airbnb title, such as number of beds and baths, could have allowed assessment on the how the structure impacts price. Our testing R^2 and Adj. R^2 were consistently about .35, indicating that only 35% of the variability in new data is explained by our model. More variables could be added—such as more detailed structure and property details—that would improve the model's predictive power.

For the most part, the model could be applied to other Airbnb datasets in different cities. Splitting the ZIP codes into equal, quartile ranges, as defined in the methodology instead of set boundaries limits this ability but was necessary since more data on other cities' median income levels would have been needed.

Further analysis on amenities and specific room type could be lucrative. Since reviews are relative and variable to the user, exploring the language used in each review using Natural Language Processing techniques could yield insight onto this relativity and its influence on the Airbnb prices as a whole. Furthermore, having information on the size of the Airbnb (square footage) as well as the number of bedrooms and bathrooms could also be key factors in determining the price of Airbnb's.

Appendix

Additional EDA Plots

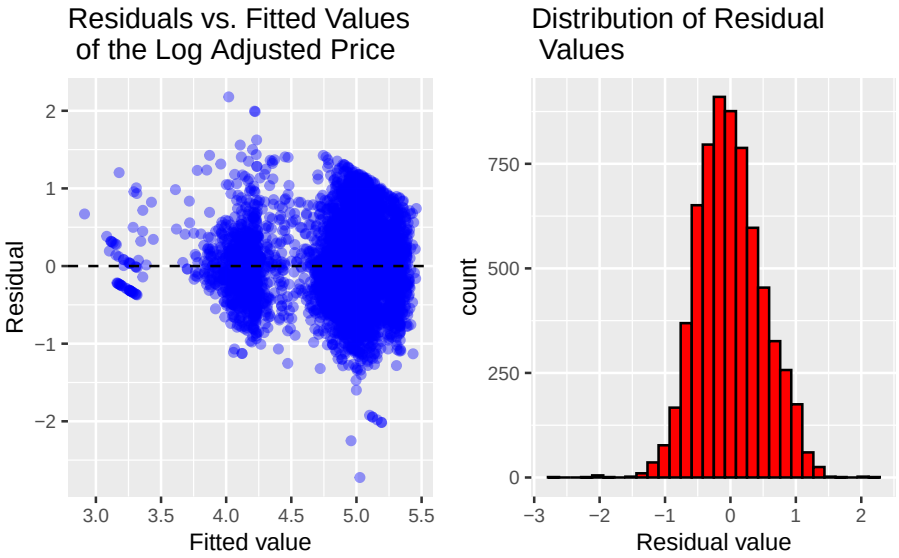


VIF

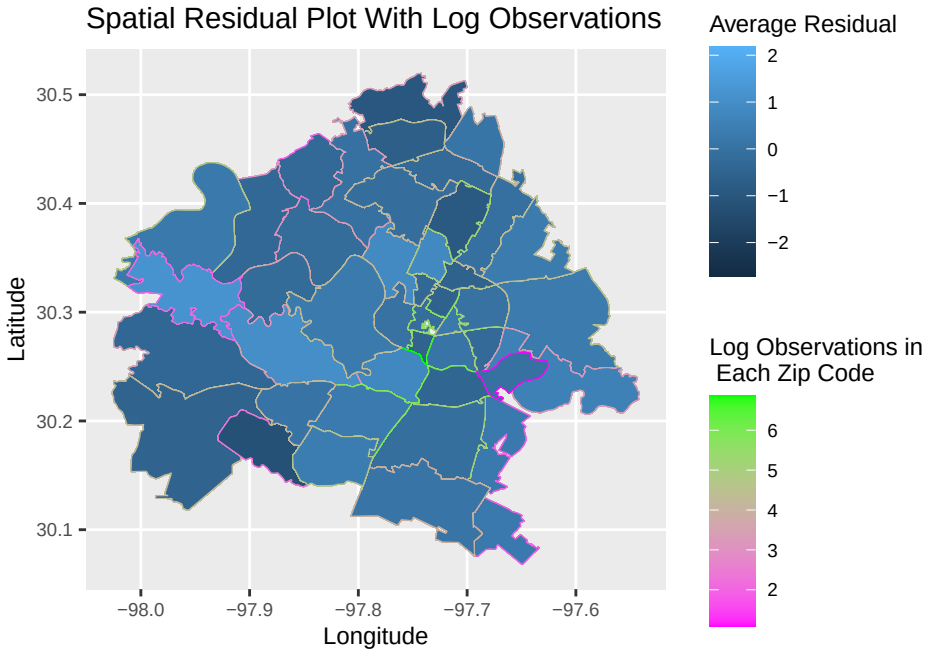
Table 2.1: VIF Values for Full Model

Variables	Full Model
reviews_per_month	3.749387
availability_365	1.182403
stars	1.096806
calculated_host_listings_count	1.286472
minimum_nights	1.020782
number_of_reviews	1.665116
number_of_reviews_ltm	3.982724
room_type_Private.room	1.052944
room_type_Shared.room	1.071225
income_level_Upper.Mid.Income	2.019066
income_level_Lower.Mid.Income	2.180952
income_level_Low.Income	2.077496

Residual Plots



Spatial Residual Plot



Interaction Model Statistics

Table 3.1: VIF Scores of Sucsessive Interaction Models

Terms	Original VIF	Iteration 1	Iteration 2
Single Terms			
reviews_per_month	957.586	NA	NA
availability_365	427.737	NA	NA
stars	13.642	10.984	3.200
host_listings	216.492	178.902	175.187
minimum_nights	2412.053	NA	NA
Private.room	353.870	327.734	313.024
Shared.room	954.029	766.253	757.155
Upper.mid.income	1154.728	1129.494	NA
Lower.mid.income	1298.502	1276.552	NA
Low.income	1101.694	1048.729	NA
Interaction Terms			
reviews_per_month_x_availability_365	5.210	NA	NA
reviews_per_month_x_stars	958.690	NA	NA
reviews_per_month_x_host_listings	5.107	NA	NA
reviews_per_month_x_minimum_nights	7.669	NA	NA
reviews_per_month_x_Private.room	1.954	NA	NA
reviews_per_month_x_Shared.room	2.818	NA	NA
reviews_per_month_x_Upper.mid.income	4.368	NA	NA
reviews_per_month_x_Lower.mid.income	5.358	NA	NA
reviews_per_month_x_Low.income	4.767	NA	NA
availability_365_x_stars	414.438	NA	NA
availability_365_x_host_listings	14.087	NA	NA
availability_365_x_minimum_nights	6.052	NA	NA
availability_365_x_Private.room	2.182	NA	NA
availability_365_x_Shared.room	16.545	NA	NA
availability_365_x_Upper.mid.income	6.365	NA	NA
availability_365_x_Lower.mid.income	7.243	NA	NA
availability_365_x_Low.income	6.410	NA	NA
stars_x_host_listings	194.464	160.178	173.129
stars_x_minimum_nights	2271.913	NA	NA
stars_x_Private.room	337.929	312.552	309.603
stars_x_Shared.room	821.508	744.772	719.487
stars_x_Upper.mid.income	1145.383	1119.691	NA
stars_x_Lower.mid.income	1285.813	1265.464	NA
stars_x_Low.income	1079.966	1027.089	NA
host_listings_x_minimum_nights	10.536	NA	NA

host_listings_x_Private.room	3.496	1.374	1.520
host_listings_x_Shared.room	319.846	10.331	11.357
host_listings_x_Upper.mid.income	2.761	2.400	NA
host_listings_x_Lower.mid.income	4.303	3.762	NA
host_listings_x_Low.income	4.023	3.509	NA
minimum_nights_x_Private.room	6.675	NA	NA
minimum_nights_x_Shared.room	468.126	NA	NA
minimum_nights_x_Upper.mid.income	8.146	NA	NA
minimum_nights_x_Lower.mid.income	1.750	NA	NA
minimum_nights_x_Low.income	4.041	NA	NA
Private.room_x_Upper.mid.income	3.915	3.765	NA
Private.room_x_Lower.mid.income	6.037	5.841	NA
Private.room_x_Low.income	6.054	5.855	NA
Shared.room_x_Upper.mid.income	11.784	7.209	NA
Shared.room_x_Lower.mid.income	43.121	23.913	NA
Shared.room_x_Low.income	11.369	6.527	NA

Note: Some terms in above table were truncated due to formatting issues.

Table 3.2: Comparative Interaction and Reduced R-squared Values

Original	Iteration 1	Iteration 2	Reduced Model
0.3890995	0.3580108	0.3364097	0.3644078