Final Report

Tejaswi Bhavaraju, Akshat Talreja

Dec 19, 2021

Contents


```
data = read.csv("../data/clean/cleaned_listings.csv")
```

The code to reproduce this report is available on Github.

1 Executive Summary

Problem

Problem. While the general impact of Airbnb and its short-term rentals on residential communities is substantial, the impact at the neighbourhood or city level can be seen to be much more varying. People belonging to and visiting a particular city have a unique combination of factors that affects their affinity for short term housigrng, which in turn determines the price of an Airbnb listing. And in certain cities such as Asheville (North Carolina), the state of short-term rentals is so crucial to tourism that studying their pricing can be very important. For our final project, we are looking at the Airbnb listing data of Asheville and studying the relationship between amenities, duration of stay, number of customers, several other factors, and the price of an Airbnb listing. Considering these factors, we hoped to extract a way of explaining and predicting the price of an Airbnb listing so that stakeholders could have a clear understanding of how pricing of Airbnbs in this particular community could be strategized.

Data. We obtained data from the website of "Inside Airbnb", a mission driven activist project which aims to quantify the impact of short-term rentals on housing and residential communities. The dataset included the price of Airbnb listings and a number of features related to them such as room type, duration of stay, number of views of the listing, and several more. Our explanatory variables could be divided into three categories: host details (e.g. host response time, host acceptance rate) , room characteristics (e.g. room type, bathroom type), and reviews (review scores rating, number of reviews). As mentioned above, our primary response variable was the daily price in local currency (labelled as "price").

Analysis The main aim of our analysis is to better understand what listing attributes contribute to prices. After the cleaning process, we had 1,882 listings which we split into training and testing sets. Our exploratory data analysis studied the variation of our response variable and its relationship with features from the three main feature categories, namely host, room, and review related features. The main aim of our exploratory analysis was to uncover some preliminary trends on important attributes that can influence prices, and to get a better understanding of the distribution of our response variable, the price, and the correlation between different features. We further predicively modelled price using a range of statistical machine learning methodologies including Ordinary Least Squares, Ridge Regression, Lasso Regression, and Random Forest. For each model, we extracted important contributing features to better understand features that have predictive potential when it comes to prices. Lastly, we utilised these to come up with recommendations and takeaways for current and new airbnb hosts on pricing their properties and to potentially improve the value of their listings on the platform.

2 (Akshat) Conclusions paragraph

3 Introduction

Background. Asheville has been Airbnb's most hospitable city in the US for the past few years. Local Airbnb hosts earned $19.8 million in 2017, which is more than the collective revenue of the next four biggest cities in North Carolina.[1] Additionally, the 'Convention and Visitors Bureau of Asheville' has concluded that tourism is vital for the economy of the city because of visitor spending at local businesses and taxes generated by tourism.[2] Considering Airbnb's crucial position in tourism, the ability to optimally determine the price of Airbnb's listings can help utilize the platform's success and generate the maximum revenue possible from short-term rentals. This in turn would significantly contribute to the city's prosperity. Additionally, with an increasing number of people using Airbnb, and being hosts, a better way to price would increase transparency on the market and better equip new hosts to get their journey kickstarted. Additionally, a

[1] Airbnb Prevalence in Asheville: https://www.citizen-times.com/story/news/local/2018/08/30/airbnb-asheville-3-percent-highest-used-rate-united-states-housing/1143541002/

[2] Importance of Tourism in Asheville: https://www.ashevillecvb.com/tourism-builds-community/

better understand of listing related factors driving prices can further allow hosts to increase the value of their listings and ultimately, earn better passive income through the platform.

Analysis goals. The main aim of our analysis is to better understand what listing attributes contribute to prices, and use this to develop reommendations on factors that airbnb hosts can consider when pricing their properties and those that they can potentially use to increase the value and appeal of their properties on the platform. To do so, we performed a series of exploratory analysis to uncover the relationships between key listing and host attributes and prices, and further performed predictive modelling to better understand features that significantly contribute to price. These features were split across three main categories: host, room/acommodation, and reviews, and detailed description on each feature used can be found within the Appendix. While our main aim was explaining the reasoning behind prices on the platform, we nevertheless set out to build a robust predictive model. Thus, we choose root mean squared error as the metric of choice to evaluate the model, gives its high interpretability in the context of data.

Significance. The significance of this analysis can be traced back to Airbnb's explosive growth and the rise in the number of people using this as a source of passive income. A better understand of factors driving prices can allow Airbnb hosts to more successfully operate on the platform. Additionally, a better understanding of features driving higher prices can further allow hosts to increase the value of their property by optimizing things within their control. Airbnb is a lucrative channel for passive income, and as we come out of the COVID pandemic with a likely pent-up demand for travel, potential hosts looking to earn through this channel can take away insights on factors driving prices an more strategically operate their listings.

4 Data

4.1 Data sources

We obtained our dataset from the website of "Inside Airbnb", a mission driven activist project which aims to quantify the impact of short-term rentals on housing and residential communities.[3] The project collects data on Airbnb listings of about 80 cities around the World using a scraping algorithm. The project was founded by Murray Cox, an Australian-American community activist who started scraping information from the Airbnb website in 2014 and compared it to public data release from the company. Considering that Asheville is the most hospitable city in the US with reference to Airbnb, the website has extensive listing data for the city.[4]

4.2 Data cleaning

We started with removing some unnecessary columns that were in the raw dataset. We then created features for the length of the description of listing and the amount of days a host had been on airbnb. Each host can have a number of verifications. To abstract away the complexity of handling each verification separately, we used the number of verifications as a proxy for the potential credibility of a listing to eventually investigate if this would have any impact on the price. We had to convert several columns into different formats for them to be useful for analysis, such as the price column into a double by removing characters from it, percentage variables into doubles, and all the character data types into factors for modelling. Finally, we dropped all the listings which had any NA values. Having started with 2,563 listings, our cleaned dataset had 1,882 listings.

4.3 Data description

The cleaned dataset had 1,882 entries which each represent a unique listing on Airbnb within Asheville. Each entry shows a specific listing. For each listing, there are a total of 35 features including room types, host

[3]Inside Airbnb Project: http://insideairbnb.com/about.html
[4]Airbnb Asheville Dataset: http://insideairbnb.com/get-the-data.html

status, reviews, among others. A detailed description of these variables can be found within the Appendix

4.3.1 Response Variable

The response variable is the Price of the Airbnb listing, which is a continuous variable. In studying the distribution of the response variable "Price", the price is right skewed which reflects the fact that there are several listings which are priced over $250 even though the median is $150. There is a long tail to the right reflecting outlier listings with extremely high pricing, which makes it interesting to predict potential factors that enable high pricing. The histogram below further shows this distribution

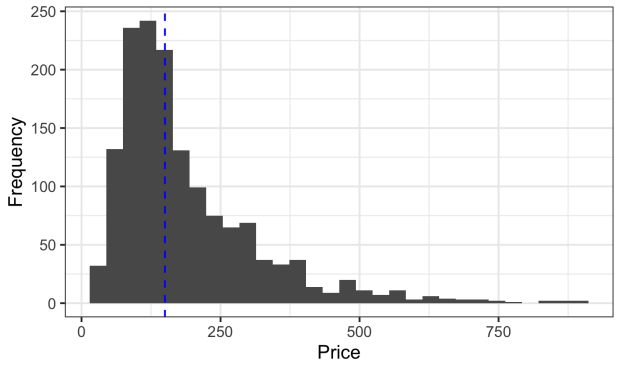

Figure 1: Distribution of prices

4.4 Data allocation

We used an 80-20 split for this project such that 80% of the 1,882 clean entries were used for training and 20% of the entries were used for testing. Since we had categorical variables, the split was performed when training each model, however a fixed seed was used to make sure that the split is constant, and a fair judgement on the performance of models explored later can be studied.

4.5 Data exploration

4.5.1 Response

Considering the Top 10 listings by Price, it can be seen that the hosts for these listings operate at top quality standards. All hosts of the 10 most expensive Airbnb listings respond to every message within an hour and 9 out of them are superhosts.

4.5.2 Features

4.5.2.1 Correlation Map of All Variables To begin with, we analysed the correlation between our numeric explanatory features to better understand their relationships with each other. This can be seen in Figure <>

As we see, most features are not strongly correlated with each other, however, some clusters of features do showcase a positive relationship. For instance, we observe a positive correlation between number of people accomodated, number of bedrooms, and number of beds, which are features that are intuitively closely linked to one another. In addition, we see a close relationship between availabilities at different dates. Lastly, we also see positive correlations between different review scores, which we would once again expect. Here, we

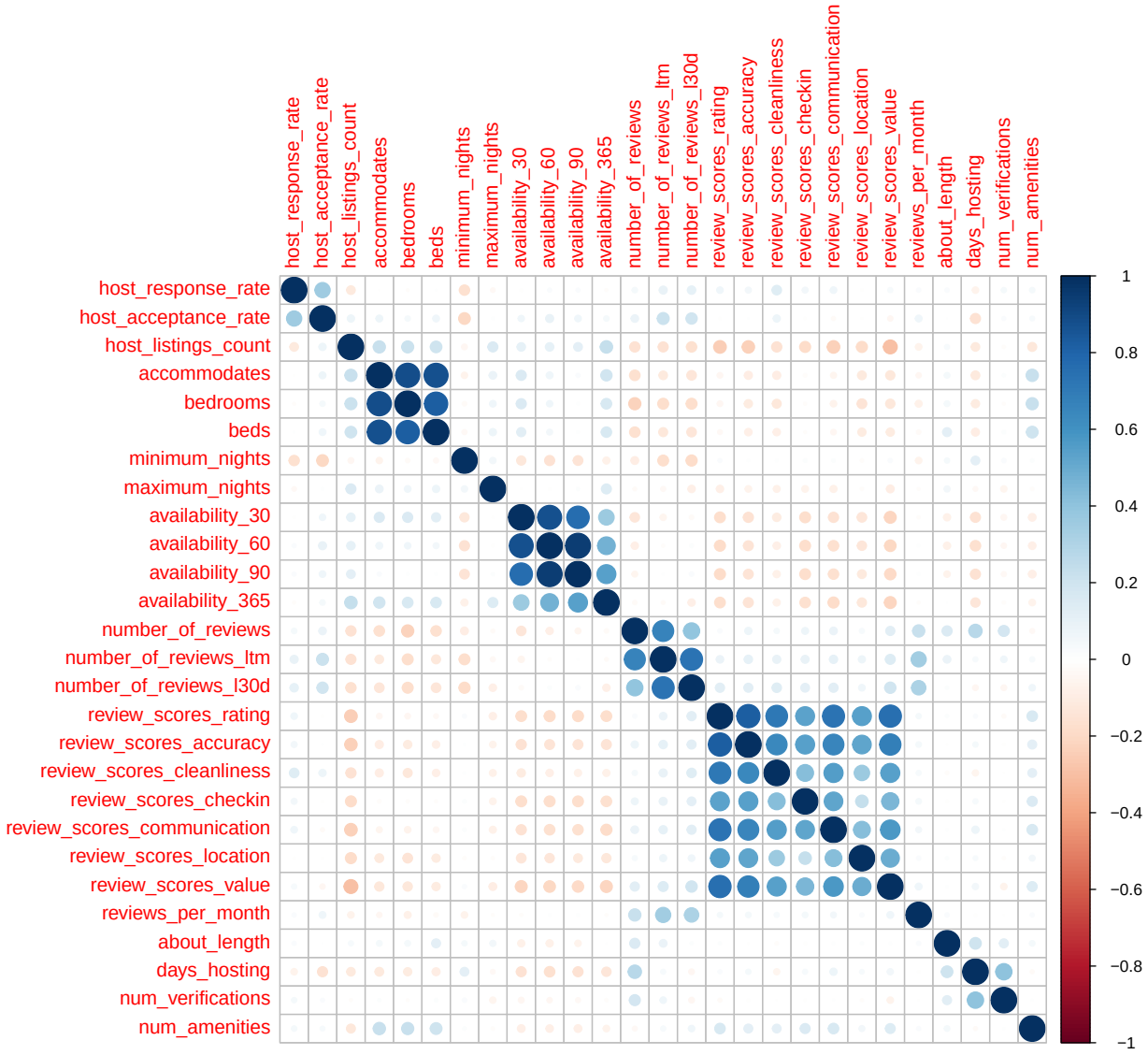

Figure 2: Correlation between explanatory features

Table 1: Top 10 Listings by Price

Price	Response Time	Response Rate	Is Host a Superhost?
893	within an hour	1.00	t
893	within an hour	1.00	f
865	within an hour	0.99	t
861	within an hour	0.99	t
846	within an hour	1.00	t
825	within an hour	0.99	t
766	within an hour	1.00	t
755	within an hour	1.00	t
734	within an hour	1.00	t
727	within an hour	1.00	t

made a decision to leave all features within the explanatory set, however, another possible method would be to drop highly correlated features, which we can further experiment with in future iterations of this analysis.

4.5.2.2 Analysing Relationship between price & key variables In addition to the above, we analysed the relationship between price and some key variables to uncover potential attributes of houses. The results of this analysis are as follows

Price by Room Type

We created a boxplot graphic to study the relationship between the type of the room within the listing (Entire home, Hotel Room, Private Room, and Shared Room). As we expected, hotel rooms were the most expensive kind of listing in distribution. While entire homes or apartments were second in terms of pricing, they had a large number of outliers which had rent upwards of $500. On the other hand, private or shared rooms tend to have lower prices.

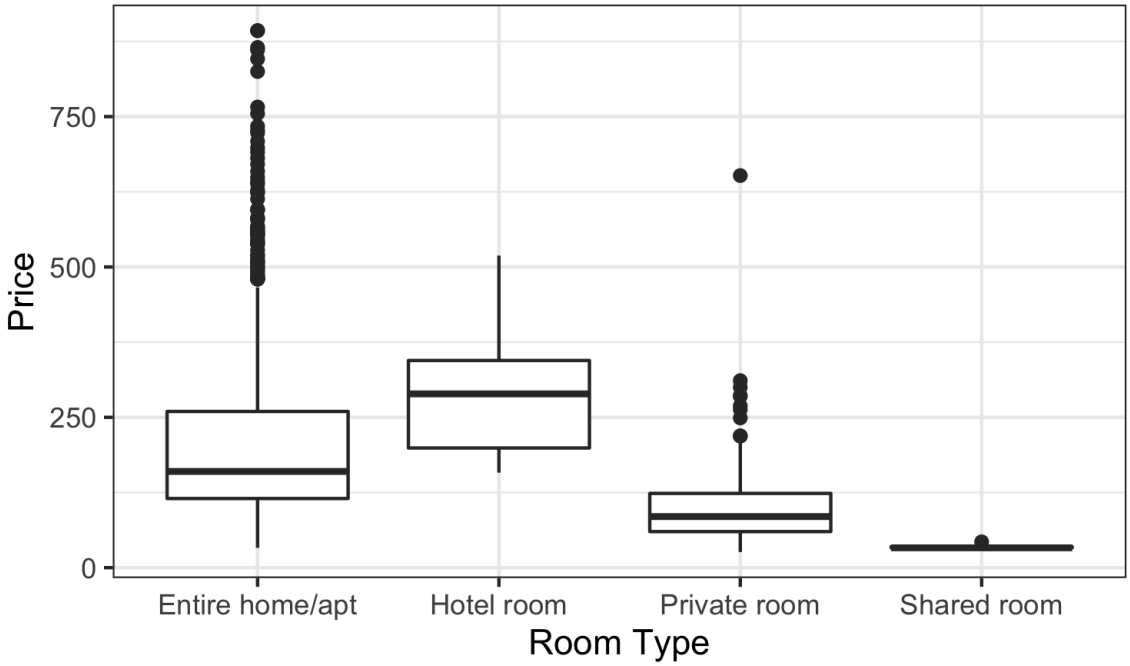

Figure 3: Prices by Room Type

Price and Number of Reviews

A key feature which belongs to the category of reviews was the number of total reviews provided on a listing. Our initial understanding was that a listing which has high reviews would transfer its popularity in its pricing. However, as the scatterplot shows, generally, at high number of reviews, the listings tended to have lower price. This is an interesting revelation and a potential reason for this could be that the clientele for lower priced listings is more likely to be supportive of their hosts, among others

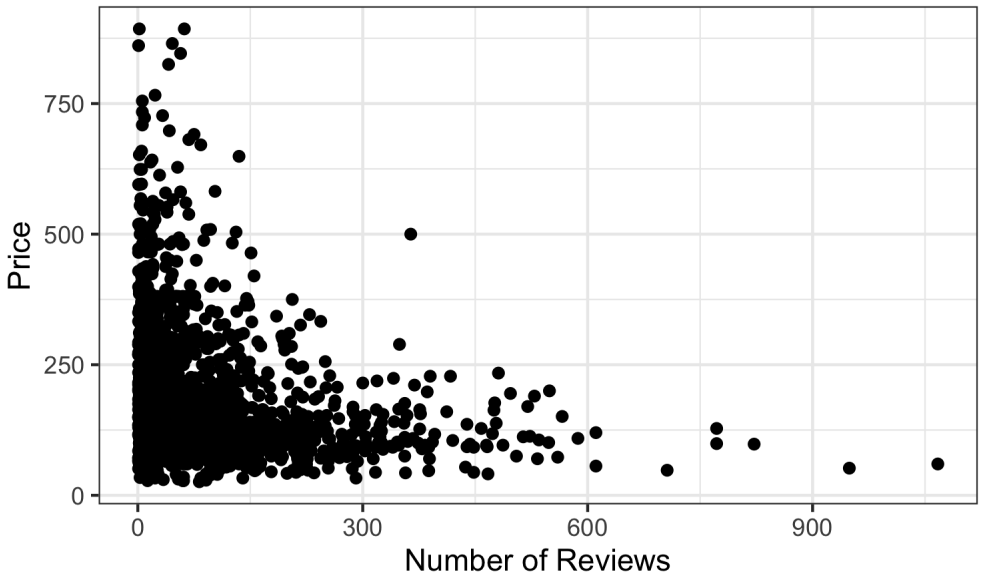

Figure 4: Number of reviews and prices

Price and Superhost Status

We also wanted to test whether most listings which are created by superhosts are priced higher than otherwise. However, we found that the median price of a listing for superhosts is the same as those listings whose hosts are not superhosts. An additional aspect to note though is that superhosts had much more outliers in terms of listings whose prices are considerably higher than the median.

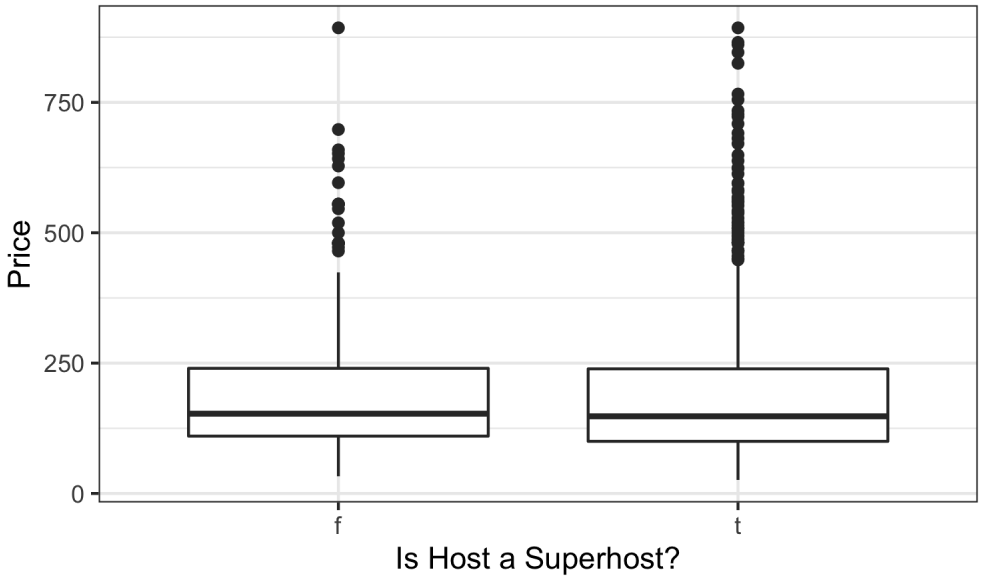

Figure 5: Superhost status and prices

Price & Verification

Additionally, we analysed how a host's identity being verified can influence prices. Based on the plot below, we see that while listings with verified host identidities have a higher median price, the difference is likely minimal and both have outliers tending to higher prices. This was an interesting finding especially considering our hypothesis that verification would result in a boost in prices due to higher credibility in the hands of a verified host.

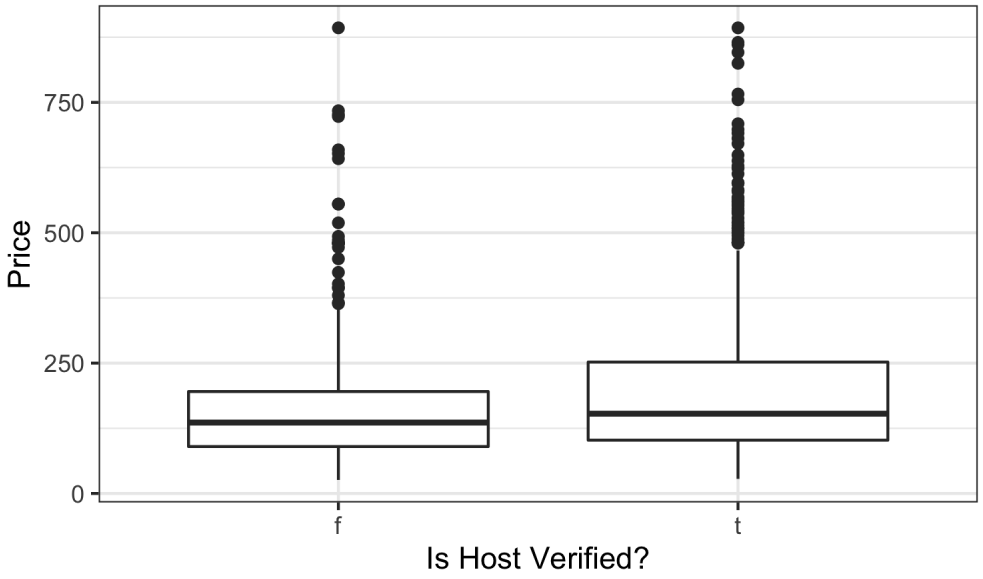

Figure 6: Host verification and prices

Price & Time taken by host to respond

We further analysed how a host's response time affects price. As was our hypothesis, we see that quicker responses yield a higher median price, and furthermore, listings with a host that responds quicker have more outliers tending to higher prices, which indicates that being responsive can be a way to boost value of a property. What is interesting is that the jump in median prices prices between hosts that respond within a few days and more andd those that respond within a day is much higher compared to subsequent jumps, indicating that taking more than a day in responding can be especially disadvantageous.

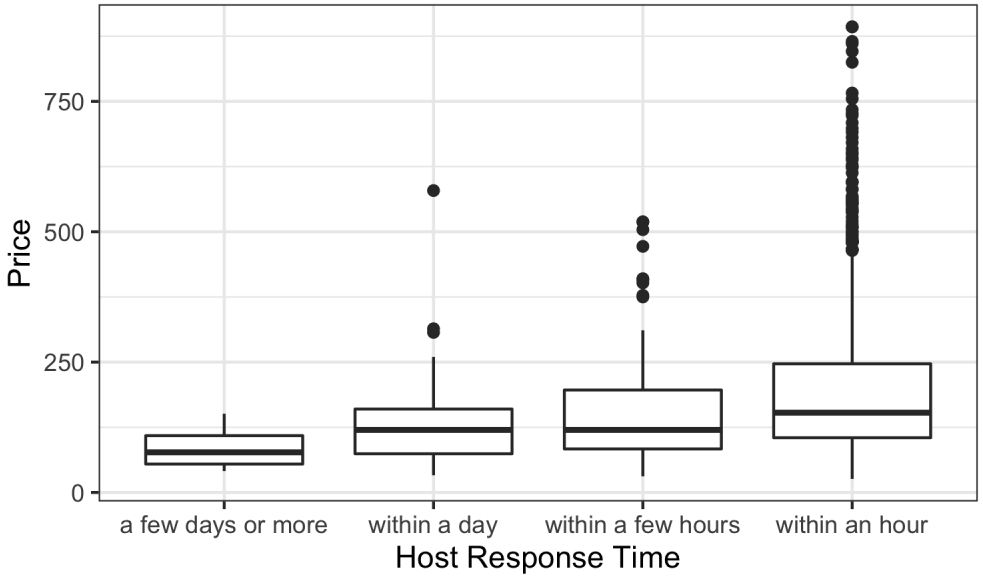

Figure 7: Host response time and prices

5 Modeling

5.1 Regression-based methods

5.1.1 Ordinary least squares

To begin our analysis, we ran an ordinary least squares regression on price against the 35 potentially explanatory variables. To understand drivers behind prices, we further studied the variables with a p-value lesser than a significance level of 0.05. The top 10 variables by significance were:

The multiple least squares further indicated that the variables captured 67.3% of the variation which, while not perfect, gives us a fairly solid proof of concept. Additionally, since our objective with this investigation is to yield factors contributing to a high price, we can sacrifice some levels of metric optimization for the purpose of explaining the impact of different listing attributes on price.

5.1.2 Penalized regression

While ordinary least squares worked fairly well, we further decided to experiment with penalized regression models since ordinary least squares models might suffer from overfitting and high variance, and regularization can allow us to get to a more robust model.

To begin with, we ran a cross validated ridge regression model on the price against the 35 explanatory variables. Figure <> showcases the top 10 features selected first by the ridge model. Notably, we see the following 10 features:

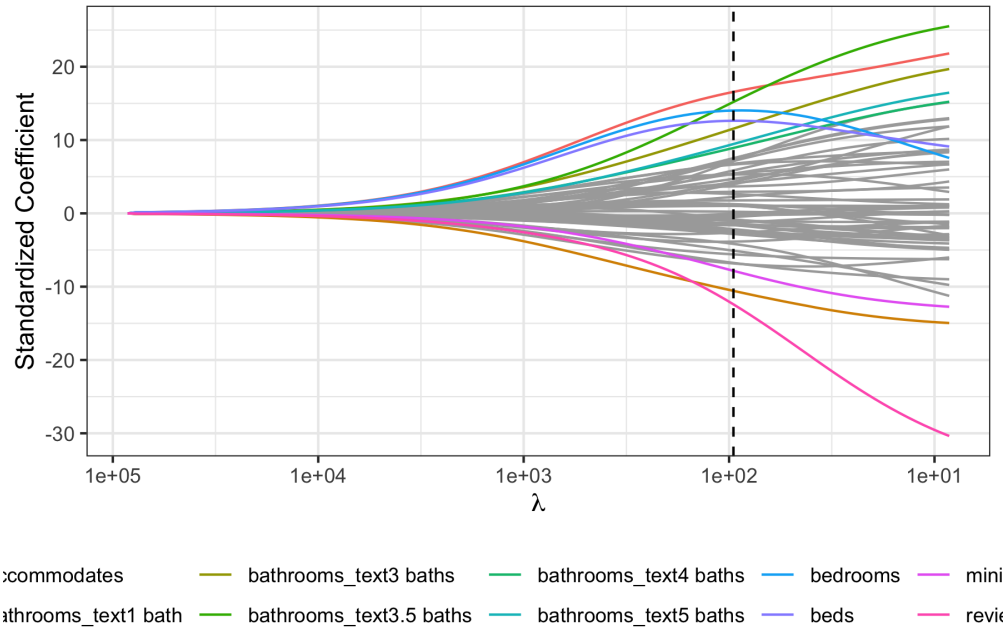

Figure 8: Top features in ridge regression

Furthermore, we also implemented a cross validated lasso regression model. The model ended up selecting all features, potentially indicating that ridge regression might have been a better choice. Nevertheless, the first 10 features selected by the model were:

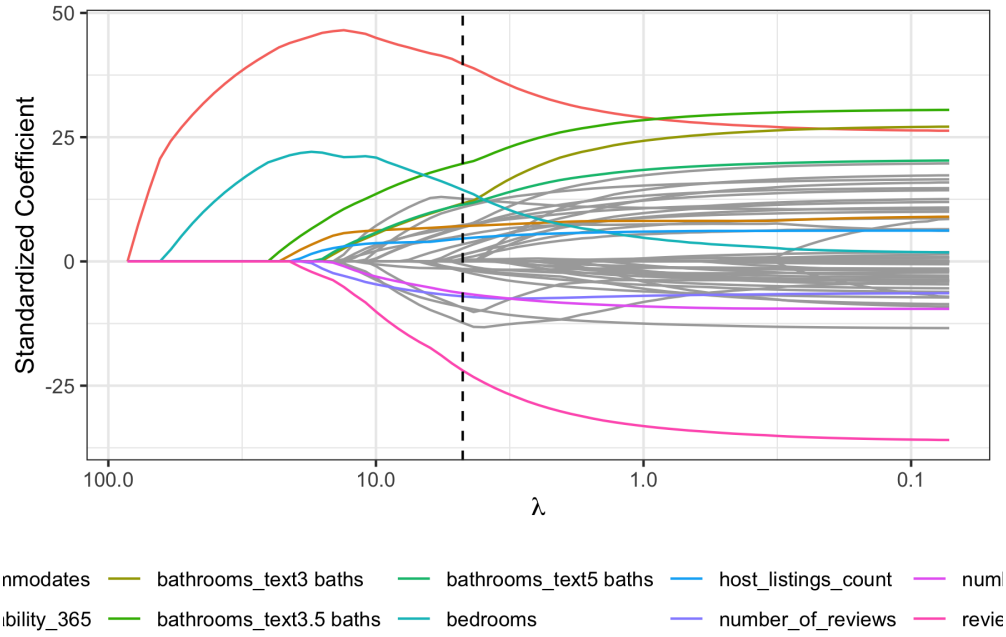

Figure 9: Top features in lasso regression

Furthermore, from the lasso model, we extracted the coefficients to analyse the top 10 contributing parameters within the cross validated model that contributed to price. This was done by selecting the top 10 features based on the highest absolute coefficient values. These are demonstrated in Table

Table 2: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
accommodates	39.71
review_scores_value	-21.94
bathrooms_text3.5 baths	19.67
bedrooms	14.39
review_scores_cleanliness	12.61
room_typePrivate room	-12.34
bathrooms_text3 baths	11.67
bathrooms_text5 baths	11.42
review_scores_location	10.89
bathrooms_text1 bath	-9.36

5.2 Tree-based methods

5.2.1 Random forest

Beyond regression models, we further experimented with tree based methodologies. In particular, we created a random forest model to predict price based on the aforementioned 35 features. When training the model,

we further tuned the number of parameters considered at each split. We tried values ranging from 1 to 30, and the out-of-bag errors for each of these is demonstrated below

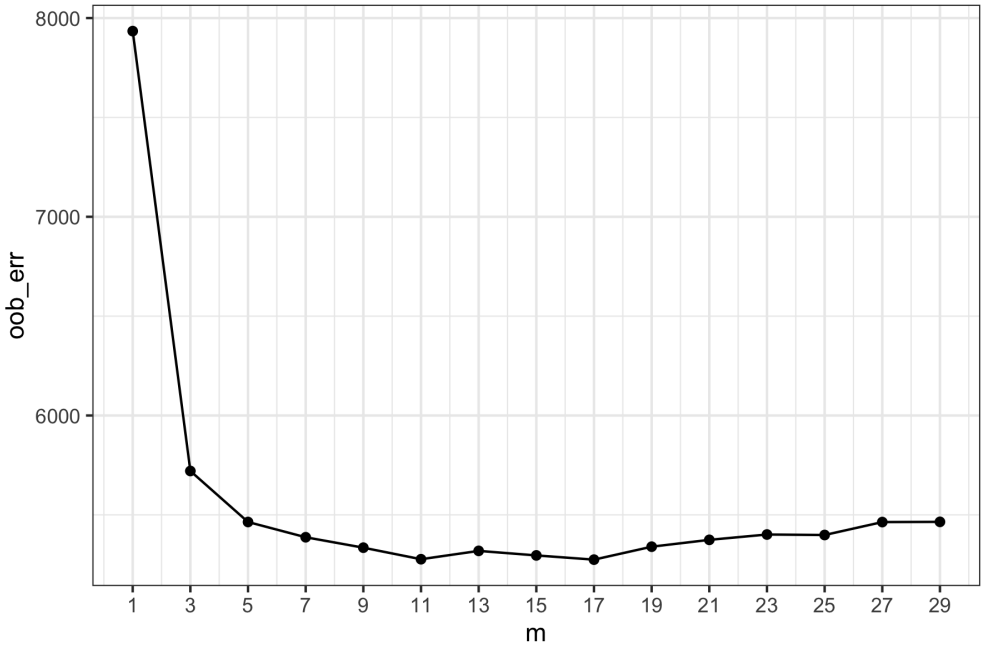

Figure 10: OOB Errors and number of features at each split

We see above that the out of bag error is minimized when at 17 features for each split. We used this tuned value of the parameter to construct our final model. Based on this, we analysed the variable importance using both OOB and purity variable importance to better understand features that are strong predictors towards price. The results of this are demonstrated in Figure <>

As we see above,

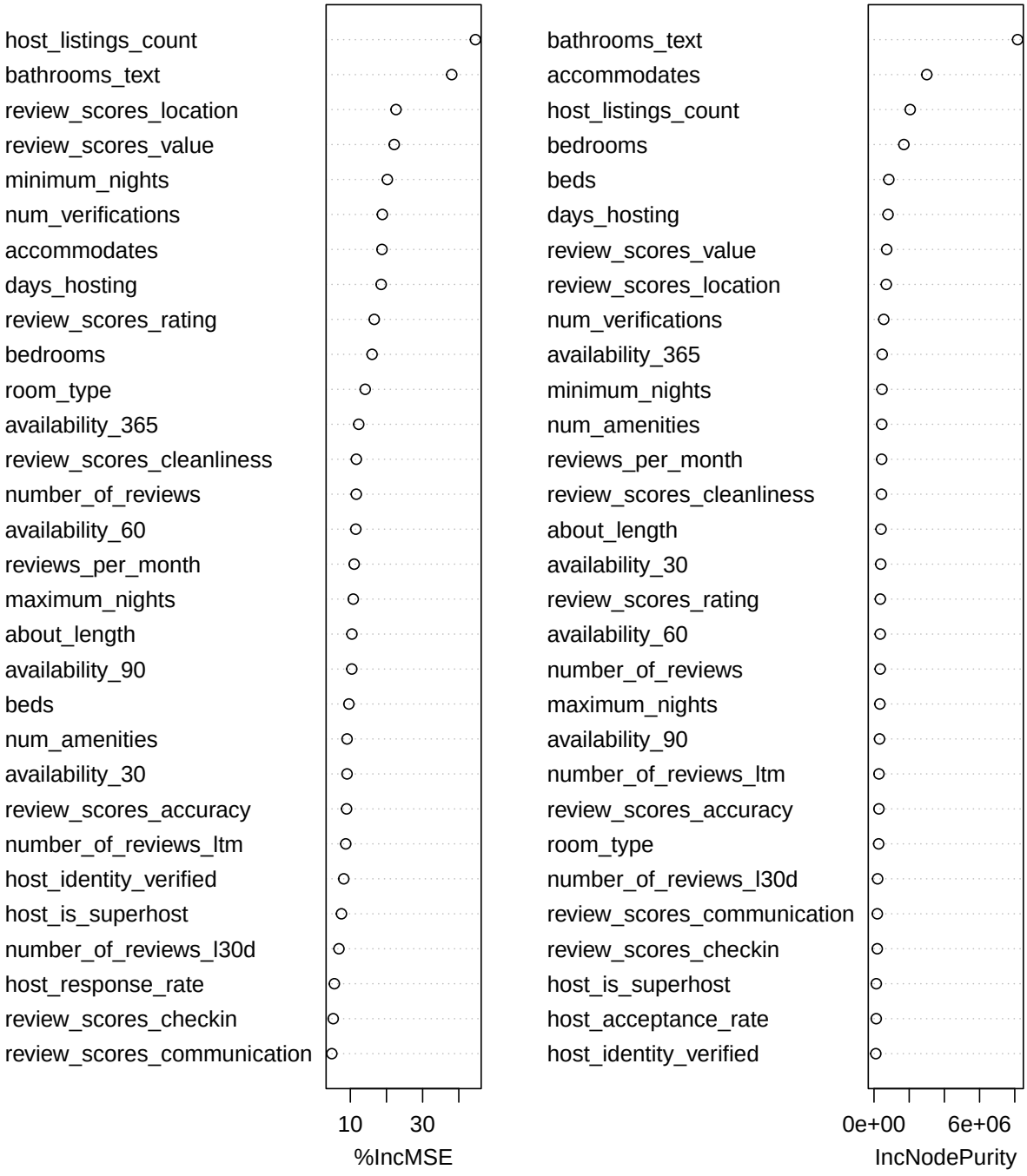

Figure 11: Random Forest Variable Importance

6 Conclusions

6.1 Method comparison

Overall, we experimented with 4 distinct modelling methodologies, namely OLS, ridge regression, lasso regression, and random forests. To analyse each of them, we considered the root mean squared errors by analysing their predictions on the test dataset. In addition to the above variables, we further included an intercept only model, which predicts the average price from the training dataset for each listing, as a benchmark to better understand the predictive potential of our models. The results of this investigation are summarized in Table <>

Table 3: Root-mean-squared prediction errors for lasso and ridge regressions.

Model	RMSE
Intercept Only	136.6
Linear Model	93.5
Ridge	93.1
Lasso	93.0
Random Forest	77.2

As seen above, each of the models experimented with perform better than the intercept only model, indicating that there is some predictive potential for each. Clearly, the random forest model outperforms the others with an RMSE of ~ <...>. While considering that our mean price is ~ <>, we can see that our RMSE values are moderately high, for the purposes of our investigation our models nevertheless give us interesting insights on predictive factors towards optimal listing prices.

6.2 Takeaways

6.3 Limitations

6.3.1 Dataset limitations

The dataset in consideration did present some challenges that could potentially be addressed going forward. Firstly, the data was reliant on scraping performed by a third party as opposed to Airbnb itself, increasing the likelihood of potential deviations from the true picture. In addition, we found that the listing prices were those quoted by the hosts as opposed to ones actually paid. While the fact that most listings had previously been reviewed and consumed is indicative that these prices have worked, a dataset on deals completed could be better to analyse optimal ways of pricing an Airbnb property. Additionally, the dataset used was only for a specific geographical region, namely Asheville, and a wider group of regions could further give a better picture of the reality on a macro level, which could increase the scope and applicability of this analysis. Additionally, there were missing values in some instances, which had to be dropped, limiting our data size. Regardless of these limitations, we were able to determine interesting relationships between listing attributes and prices, and going forward, we hope to address some of these challenges to increase the robustness of our analysis.

6.3.2 Analysis limitations

Within our analysis, we used proxies for certain variables such as amenities and verifications, whereas considering the presence of each amenity of verification as a categorical variable could potentially yield better insights. In addition, a number of textual features were dropped within the scope of this analysis

such as neighbourhood overview and host information. Factoring this within our predictive models could have improved the relevance of our analysis particularly considering that authentic and catchy descriptions generally lend credibility to a listing and may allow hosts to charge higher prices if they are able to better sell their listings using such tactics. On a more subtle implementation related note, using a different seed when splitting the data might further lead to different results, which could be another area to expand to going forward.

6.4 Follow-ups

We truly believe that a transparent idea for pricing Airbnb listings could be extremely beneficial given the number of people using these as a source of passive income and the rising prevelance of the model. To this end, there are a number of ways in which this analysis can be taken forward particularly to address the aforementioned dataset and analysis limitations. Firstly, we can try collected price data on completed deals to get an idea of prices that work instead of using quoted prices as a proxy. Further, we can perform the analysis on a wider dataset considering multiple geographical regions. Additionally, we can analyse textual features using tfidf, word2vec, and other feature extraction methods to incorporate their effect towards price. Moreover, it would further be beneficial to conduct the analysis with the missing data collected, as that could give more robust interpretations on even the current datasets.

A Appendix: Descriptions of features

Below are the 35 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are continuous.

Host features

- *Host Response*
  - Host Response Rate (`host_response_rate`): Measure of how soon the host responds to messages of guests.
- *Host Miscellaneous*
  - Host Acceptance Rate (`host_acceptance_rate`): That rate at which a host accepts booking requests.
  - Host Superhost Status (`host_is_superhost`): Boolean variable showing whether a host is a superhost or not.
  - Host Listings Count (`host_total_listings_count`): The number of listings the host has (per Airbnb calculations)
  - Host Profile Pic Status (`host_has_profile_pic`): Boolean showing whether host has a profile picture or not.
  - Host Identity Verification Status (`host_identity_verified`): Boolean showing whether host's identity has been verified or not.
  - Limited access to healthy foods (`num_verifications`): Number of different possible methods of verification for listing, such as email, telephone, etc

Room/Accomodation Features::

- *Amenities*
  - Room Type (`room_type`): Type of room, being one of the four: entire place, private rooms, hotel rooms, and shared rooms.
  - Listing Capacity (`accommodates`): The maximum capacity of the listing

  - Number of Bathrooms (`bathrooms_text`): The number of bathrooms in the listing.
  - Number of Bedrooms (bedrooms): The number of bedrooms
  - Number of Beds (`beds`): The number of bed(s)
  - Minimum Number of Nights (`minimum_nights`): Minimum number of night stay for the listing (calendar rules may be different)
  - Minimum Number of Nights(`maximum_nights`): Maximum number of night stay for the listing (calendar rules may be different)
  - Days Hosted (`days_hosting`): Number of Days hosted.
  - Number of Amenities (`num_amenities`): The number of amenities available for listing.

- *Booking Features*

  - Availability after 30 Days (`availability_30`): The availability of the listing 30 days in the future as determined by the calendar.
  - Availability after 60 Days (`availability_60`): The availability of the listing 60 days in the future as determined by the calendar.
  - Availability after 90 Days (`availability_90`): The availability of the listing 90 days in the future as determined by the calendar.
  - Availability after 365 Days (`availability_365`): The availability of the listing 365 days in the future as determined by the calendar.
  - About Length (`about_length`): Length of the about description of the host.
  - Instant Bookable (`instant_bookable`): Boolean variable showing whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing.

Review Features number_of_reviews; number_of_reviews_ltm; number_of_reviews_l30d; review_scores_rating; review_scores_accuracy; review_scores_cleanliness; review_scores_checkin; review_scores_communication; review_scores_location; review_scores_value; instant_bookable; reviews_per_month

- *Frequency*

  - Number of Reviews (`number_of_reviews`): The number of reviews the listing has.
  - Number of Reviews in Last Year (`number_of_reviews_ltm`): The number of reviews the listing has (in the last 12 months).
  - Number of Reviews in last 30 Days (`number_of_reviews_l30d`): The number of reviews the listing has (in the last 30 days).
  - Number of Reviews per month (`reviews_per_month`): The number of reviews the listing has over the lifetime of the listing.

- *Review Qualitative Features*

  - Average Review Rating (`review_scores_rating`): Average review rating given to a listing.
  - Review Scores Accuracy (`review_scores_accuracy`):
  - Review Scores Cleanliness (`review_scores_cleanliness`):
  - Review Scores Checking (`review_scores_checkin`):
  - Review Scores Communication (`review_scores_communication`):
  - Review Scores Location (`review_scores_location`):
  - Review Scores Value (`review_scores_value`):