Anudeep Kumar
Chu Nie
Aishwarya Sarkar
Yingjia Shang

Project Goals

Airbnb has revolutionized the travel industry by providing travelers with wonderfully simple and convenient places to stay during their travels. Nowadays, many people choose to stay at an Airbnb when traveling. However, the pricing for properties on Airbnb depends on various parameters, and tourists always have a hard time finding their optimal Airbnb for a reasonable price. In our project, we are exploring a dataset of Airbnb listings with features related to the amenities offered, type of rooms and apartment, host, location, ratings and reviews, and their price point. Our objective is to comprehend how each of these features affects price, find features common to the most expensive Airbnb, and build models to predict the price based on these features. All the features related to Airbnb form our predictor set, and the listed price is our target variable.

Problem Importance

The Airbnb economy is on the rise, and gradually people are leaning towards them as their preferred mode of stay during vacations or tours. Many people take it as a substitute for a hotel. Staying at an Airbnb would increase tourists' time spent at their destination, indirectly stimulating additional visits and bringing benefits to the local tourist industry. Investigating how certain features affect the price and attempting to predict price would yield a two-fold benefit for both buyers and sellers in the market.

Airbnb hosts look at numerous aspects when pricing their Airbnb property to achieve maximum profits. Listing properties properly and dynamically allows hosts to generate additional income. But with the abundance of Airbnb listings available today, hosts need to offer competitive pricing to attract customers. This analysis will help them understand what type of property and property features to invest in to attract more customers and fix an appropriate price point. In summary, it will allow an Airbnb host to make sure it has the necessary and preferred features so that it can charge a higher price without losing customers.

Many travelers prefer to stay at an Airbnb due to its flexibility and favorable price, so they pick their desired property more carefully. Every customer has certain preferences when it comes to booking stays. While the location is a priority for some, others pay more attention to the property features like the number of rooms, amenities, extra accommodations, or even recreational facilities provided on the property. Some travelers care about host reliability or property ratings, as these features would help them better evaluate the description of the property posted on the platform. If a traveler wants to find the cheapest listing available, along with certain features such as a flexible cancellation policy, he/she will know the factors to investigate and opt out of to get the lowest price possible.

Anudeep Kumar
Chu Nie
Aishwarya Sarkar
Yingjia Shang

Exploratory Analysis

Before formally cleaning the data for model fitting, we took a look at the dataset to explore the basic features and how we can fit the model on the data. The original dataset had 74,111 data points and 29 columns including the target variable, log_price, which we are trying to predict through different predictive modeling techniques. There were no duplicate rows present in the dataset. The median log_price is 4.605 and the median price (after taking the inverse transform of log) is 111. To explore the patterns and relationship between different features, we have broadly divided the features in the dataset into four categories: property description, host features, location, and ratings and reviews. Property description features include features such as property_type, room_type, amenities, or accommodations. For host features, there are features about the cancellation policy, whether the host has a profile picture, is the host's identity verified on the platform, the host's response rate, and when did he or she become a host. Location features include basic city, zip code, neighborhood, and geographic information like latitude and longitude. Rating and review features are straightforward; they are about when the first and last review is related to an Airbnb, how many total reviews there are, and the review score associated with the property. Although log_price was present in our dataset, to improve the interpretability of the EDA, 'price' has been used instead by taking the inverse logarithm of log_price.

We first started by exploring some basic feature distributions. By observing the distribution of the room_type column (Figure 1), we found out that only 3% of the listings are for shared rooms, which suggests that there is less number of hostel-type properties listed here. Looking at the bed_type feature (Figure 2), 97.2% of all properties have real beds. From the city distribution (Figure 3), we found that 73.9% of the listings are in NYC and LA. This also implies that, in this specific dataset, the emphasis is on Airbnb in NYC and LA; Airbnb properties in Boston, San Francisco, DC, and Chicago are under-represented in this dataset. Most hosts do respond to customers with a mean host response rate of 94.3% (Figure 4). Approximately 30% of all hosts have a flexible cancellation policy, while others have a moderate or stricter one (Figure 5). Most of the Airbnb properties have high review scores of greater than 85, which is surprising since this suggests that most of them are highly rated (Figure 6). Around 74% of the properties have a cleaning fee associated with them (Figure 7).

Then we examine some correlations between different variables. Looking at the pricing in different cities and neighborhoods, NYC has the highest number of expensive neighborhoods followed by LA. San Francisco has the highest average number of ratings per Airbnb, followed by Chicago. When we consider the column for the price, 6% of the values are considered outliers. (Outliers have been calculated by taking those data points which

are greater than the 75th percentile value by more than 1.5 times the interquartile range or less than the 25th percentile value by more than 1.5 times the interquartile range)
By exploring feature correlations and how they affect price, it is observed that the more expensive Airbnb also has a stricter cancellation policy. Surprisingly, host features seem to have a negligible effect on price. Also, SF has the highest average median price of Airbnb. Moreover, from the correlation heatmap (Figure 12), we see that variables that present higher correlations with price are accommodations, beds, bedrooms, and bathrooms.

Solutions and Insights

With our project goals of examining the important factors in predicting and determining Airbnb prices, we built four different predictive models to compare their accuracies and predicting power to support our statements. Before we started building the model, we preprocessed our features to aid the modeling process.

Feature Selection and Transformation

From our dataset, some features like id, names, description, latitude, and longitude, were not used since they were primary keys and did not add any value to our analysis. To better interpret the date type columns, such as host_since, and first/last_review, we transformed them to take the number of days elapsed since then. The amenities column in the dataset had a list of comma-separated amenities. The count of the number of amenities was taken and stored in a column named 'no_of_amenities'. The host_response_rate column was cleaned to get the percentage value as an integer from the original value present. The rows with null values were not too significantly large when compared to the dataset size, therefore they were dropped. The categorical variables were encoded using one-hot encoding before the data was fed to the model. Finally, we split our data into the training and test sets with an 80-20 split and scaled the predictor set.

Modeling

We explored 4 models to predict our target log_price: Linear regression, Decision tree regressor, Random Forest, and Gradient Boosting. We included and analyzed both parametric and non-parametric models. For the latter three models, we used hyperparameter tuning using GridsearchCV to find the best parameters. For the decision tree, we tuned the max_depth of the tree and found the optimal depth of 8. Similarly, for Random forest, after tuning the n_estimators and max_depth, we find the optimal values to be 300 and 80. For Gradient Boosting, we tuned the parameters and got the best case with 1000 in n_estimators and auto in max_features. Then, we

used the test set to examine the effectiveness of the models. Table 1 in the Appendix presents the summary of predictive model results on the test set. From the table, we see that the results were comparable to all the models. The R-squared values are almost the same for Linear Regression and Decision Tree models, and Gradient Boosting resulted in the best value of 68%. This means that 68% of the variance in log_price is explained by the independent features in the Boosting model. Additionally, Random Forest and Gradient Boosting models also yield the best MAPE of 6.3%.
To find the most relevant features in price prediction, we plotted feature importances for Random Forest (Figure 8), Decision tree (Figure 9), and Gradient Boosting (Figure 10). The top 10 features are identical for all 3 models: bedrooms, bathrooms, accommodates, days_since_last_review, Days_as_host, review_scores_rating, no_of_amenities, number_of_reviews, beds, and host_response_rate. The importance of the features drops dramatically after the top 10. In each of these plots, bedrooms, bathrooms, days_since_last_review, and days_as_host features are among the top 5 features with the highest importance. While it is common knowledge that the number of bedrooms and bathrooms always affects pricing, it is surprising that days_as_host is one of the most important features. This implies that customers pay heed to the host's tenure on the Airbnb platform as a means of understanding how reliable their listings might be.
From the summary of the linear regression model (Figure 11), we found that number_of_reviews, cancellation_policy_moderate, and cleaning fee have a p-value that is greater than 0.05, thus, we fail to reject the null hypothesis. While the 95% confidence interval is very likely to contain 0 in these variables, we would conclude that they are not statistically significant in predicting the target.

Insights
By fitting different predictive models, the feature importance analysis has provided us with evidence that our observations from EDA were accurate (Figure 12). According to the feature importance graphs, most features we hypothesized to have strong correlations with price were listed as the top 10 most important predictors. Therefore, we can confirm the importance of these features when predicting the price of Airbnbs.

Conclusion
After fitting four predictive models with the best model of Boosting, we have concluded that bedrooms, bathrooms, days_since_last_review, and days_as_host as our most important predictors with price prediction. For future reference, Airbnb hosts can adjust their property pricing based on these features, and travelers can use these features to evaluate whether the pricing level of the property they are looking for is appropriate.

Anudeep Kumar
Chu Nie
Aishwarya Sarkar
Yingjia Shang

Appendix

	Linear Regression	Decision Tree	Random Forest	Gradient Boosting
R-Squared	0.63	0.63	0.67	0.68
RMSE	0.4096	0.1702	0.3880	0.3844
MAPE	0.0672	0.0674	0.0630	0.0631

Table 1: Model Results Summary (Test set)

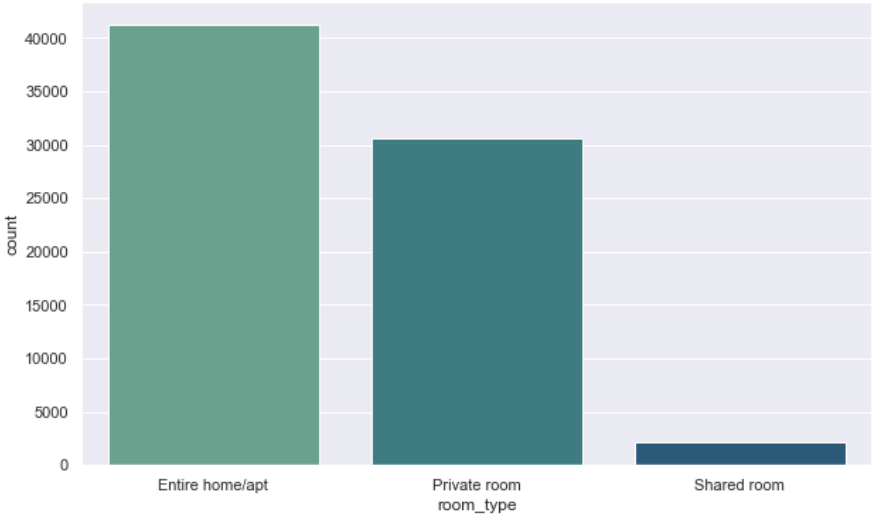

Figure 1: Room_type Distribution

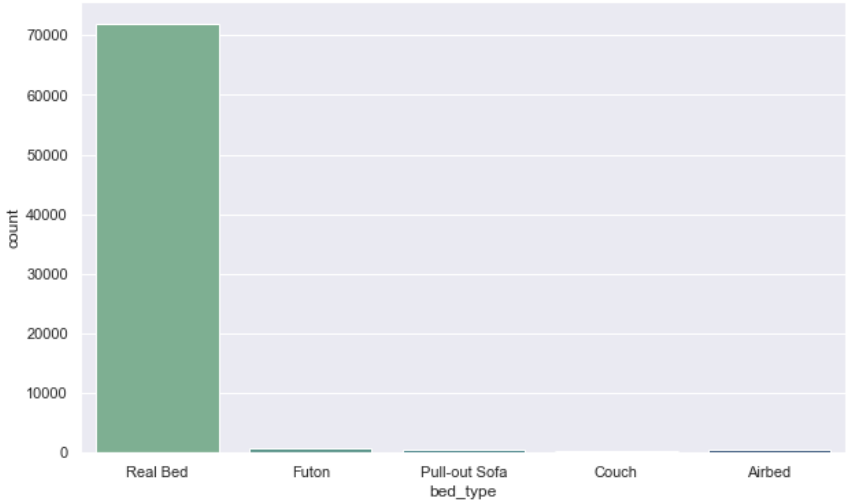

Figure 2: Bed_type Distribution

Anudeep Kumar
Chu Nie
Aishwarya Sarkar
Yingjia Shang

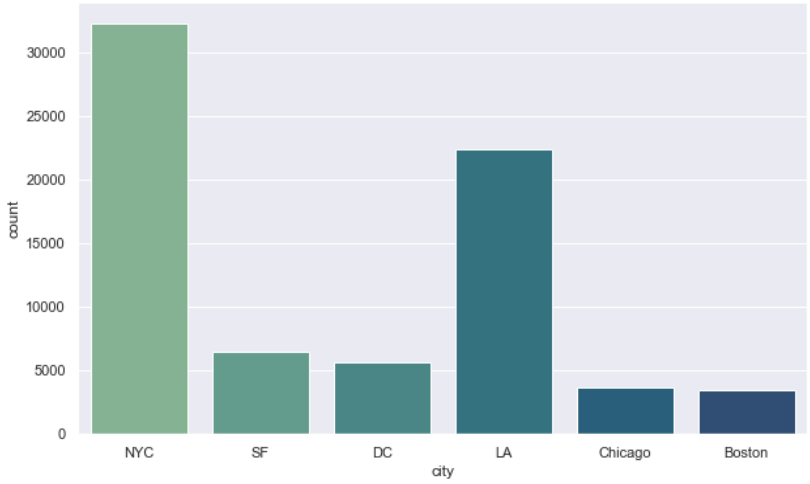

Figure 3: City Distribution

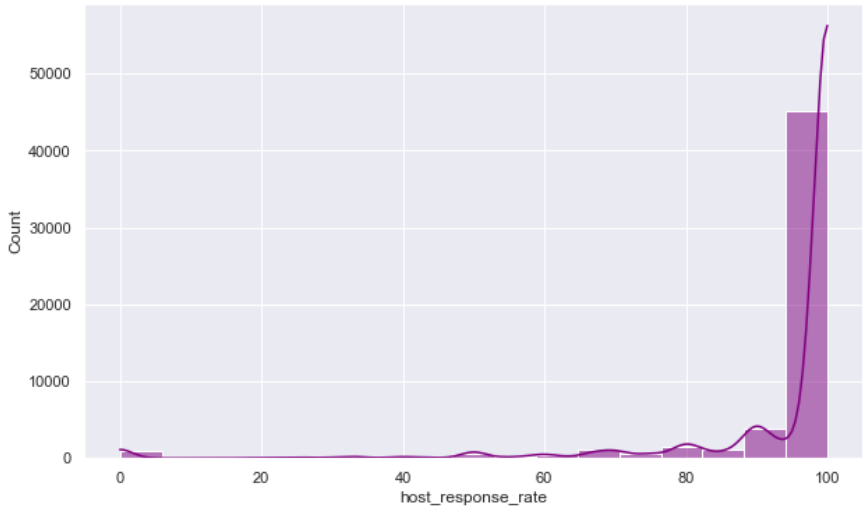

Figure 4: Host_response_rate distribution

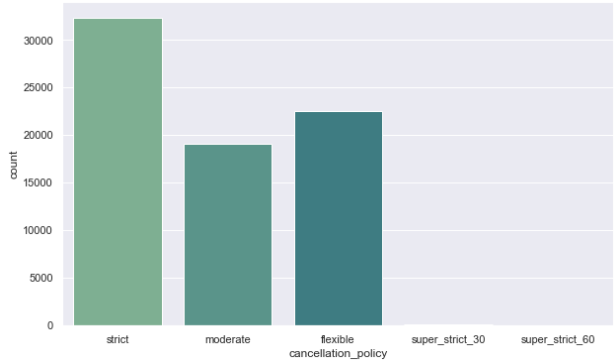

Figure 5: Cancellation_policy Distribution

Anudeep Kumar
Chu Nie
Aishwarya Sarkar
Yingjia Shang

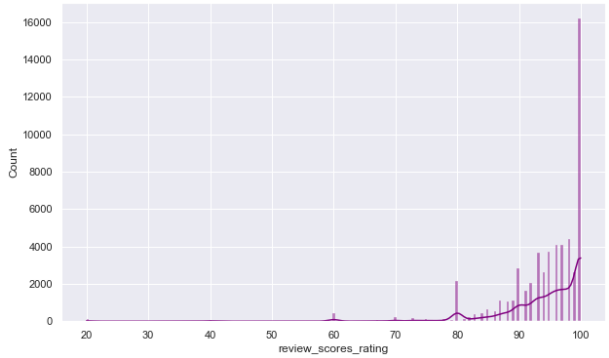

Figure 6: Review_score_rating Distribution

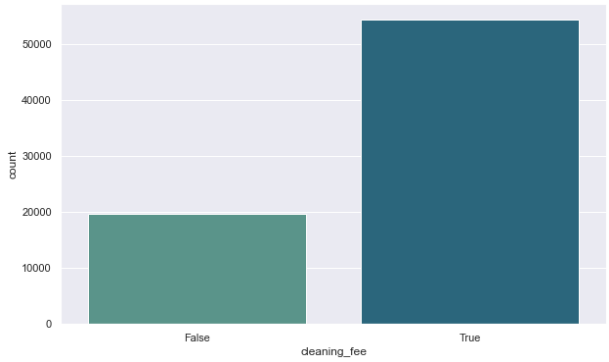

Figure 7: Cleaning_fee Distribution

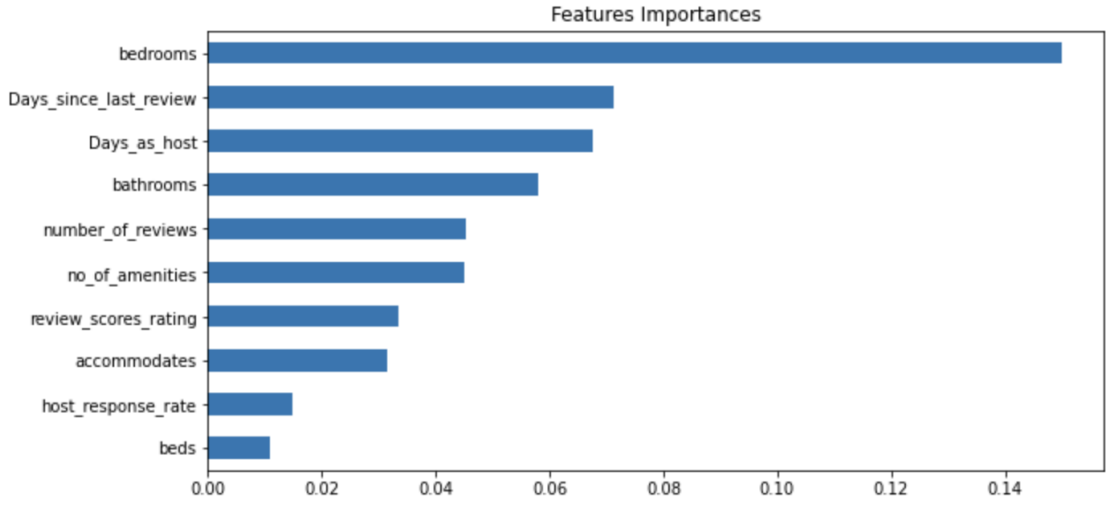

Figure 8: Features Importances of Random Forest

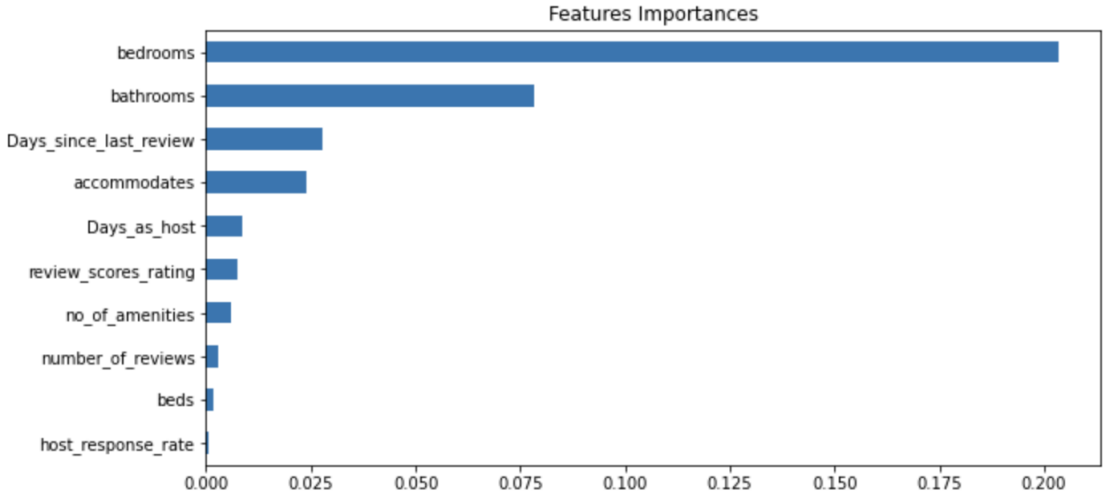

Figure 9: Features Importances of Decision Tree

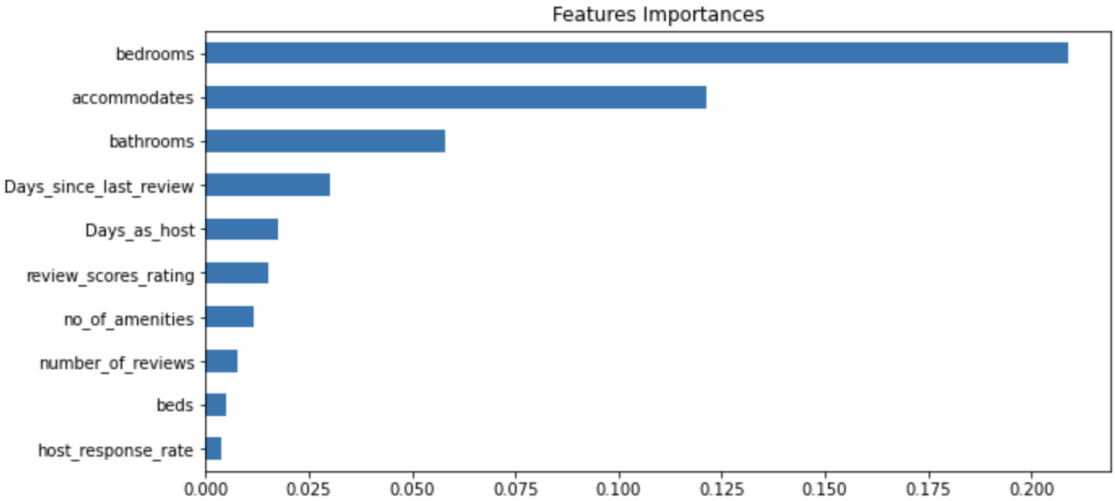

Figure 10: Features Importances of Gradient Boosting

Anudeep Kumar
Chu Nie
Aishwarya Sarkar
Yingjia Shang

	coef	std err	t	P>|t|	[0.025	0.975]
accommodates	0.0793	0.002	44.139	0.000	0.076	0.083
bathrooms	0.1476	0.004	34.697	0.000	0.139	0.156
host_response_rate	0.0008	0.000	5.425	0.000	0.001	0.001
number_of_reviews	-9.315e-05	4.81e-05	-1.935	0.053	-0.000	1.2e-06
review_scores_rating	0.0110	0.000	42.735	0.000	0.010	0.011
bedrooms	0.1376	0.004	37.491	0.000	0.130	0.145
beds	-0.0409	0.003	-14.734	0.000	-0.046	-0.035
Days_since_last_review	0.0005	1.23e-05	43.528	0.000	0.001	0.001
Days_as_host	5.048e-05	3.24e-06	15.600	0.000	4.41e-05	5.68e-05
no_of_amenities	0.0066	0.000	21.808	0.000	0.006	0.007
property_type_Bed & Breakfast	0.0911	0.023	3.889	0.000	0.045	0.137
property_type_Boat	0.2359	0.063	3.754	0.000	0.113	0.359
property_type_Boutique hotel	0.1288	0.069	1.855	0.064	-0.007	0.265
property_type_Bungalow	-0.0347	0.026	-1.360	0.174	-0.085	0.015
property_type_Cabin	-0.1244	0.055	-2.262	0.024	-0.232	-0.017
property_type_Camper/RV	-0.2338	0.052	-4.467	0.000	-0.336	-0.131
property_type_Castle	0.3486	0.117	2.988	0.003	0.120	0.577
property_type_Cave	0.2649	0.297	0.891	0.373	-0.318	0.848
property_type_Chalet	0.1015	0.188	0.540	0.590	-0.267	0.470
property_type_Condominium	0.0950	0.011	9.006	0.000	0.074	0.116
property_type_Dorm	-0.4148	0.042	-9.780	0.000	-0.498	-0.332
property_type_Earth House	0.0834	0.243	0.344	0.731	-0.392	0.559
property_type_Guest suite	-0.1229	0.042	-2.896	0.004	-0.206	-0.040
property_type_Guesthouse	-0.0645	0.021	-3.001	0.003	-0.107	-0.022
property_type_Hostel	-0.5097	0.058	-8.791	0.000	-0.623	-0.396
property_type_House	-0.0595	0.005	-11.396	0.000	-0.070	-0.049
property_type_Hut	-0.3741	0.159	-2.353	0.019	-0.686	-0.062
property_type_In-law	-0.2195	0.053	-4.119	0.000	-0.324	-0.115
property_type_Island	0.8016	0.420	1.907	0.057	-0.022	1.626
property_type_Loft	0.1478	0.015	10.169	0.000	0.119	0.176
property_type_Other	0.0318	0.021	1.478	0.139	-0.010	0.074
property_type_Serviced apartment	0.1899	0.109	1.749	0.080	-0.023	0.403
property_type_Tent	-0.2393	0.113	-2.125	0.034	-0.460	-0.019
property_type_Timeshare	0.4897	0.077	6.368	0.000	0.339	0.640
property_type_Tipi	0.6421	0.243	2.643	0.008	0.166	1.118
property_type_Townhouse	-0.0327	0.013	-2.597	0.009	-0.057	-0.008
property_type_Train	0.6677	0.297	2.246	0.025	0.085	1.251
property_type_Treehouse	0.3765	0.210	1.790	0.073	-0.036	0.789
property_type_Vacation home	0.3694	0.172	2.152	0.031	0.033	0.706
property_type_Villa	0.0476	0.038	1.252	0.211	-0.027	0.122
property_type_Yurt	0.1708	0.172	0.995	0.320	-0.166	0.507
room_type_Private room	-0.5956	0.005	-123.024	0.000	-0.605	-0.586
room_type_Shared room	-1.0389	0.013	-78.424	0.000	-1.065	-1.013
bed_type_Couch	0.5274	0.044	11.871	0.000	0.440	0.614
bed_type_Futon	0.4860	0.031	15.687	0.000	0.425	0.547
bed_type_Pull-out Sofa	0.5555	0.032	17.270	0.000	0.492	0.619
bed_type_Real Bed	0.5870	0.025	23.815	0.000	0.539	0.635
cancellation_policy_moderate	0.0076	0.006	1.296	0.195	-0.004	0.019
cancellation_policy_strict	0.0411	0.006	7.465	0.000	0.030	0.052
cancellation_policy_super_strict_30	0.2096	0.048	4.339	0.000	0.115	0.304
cancellation_policy_super_strict_60	0.7239	0.133	5.427	0.000	0.462	0.985
cleaning_fee_t	-0.0017	0.005	-0.322	0.748	-0.012	0.009
city_Chicago	-0.3475	0.012	-29.947	0.000	-0.370	-0.325
city_DC	-0.1398	0.011	-12.371	0.000	-0.162	-0.118
city_LA	-0.1792	0.010	-18.750	0.000	-0.198	-0.160
city_NYC	0.0453	0.009	4.992	0.000	0.027	0.063
city_SF	0.3028	0.011	27.905	0.000	0.282	0.324
host_has_profile_pic_t	1.5330	0.035	43.542	0.000	1.464	1.602
host_identity_verified_t	-0.0232	0.005	-4.980	0.000	-0.032	-0.014
instant_bookable_t	-0.0108	0.004	-2.451	0.014	-0.019	-0.002

Figure 11: Linear Regression Results Summary

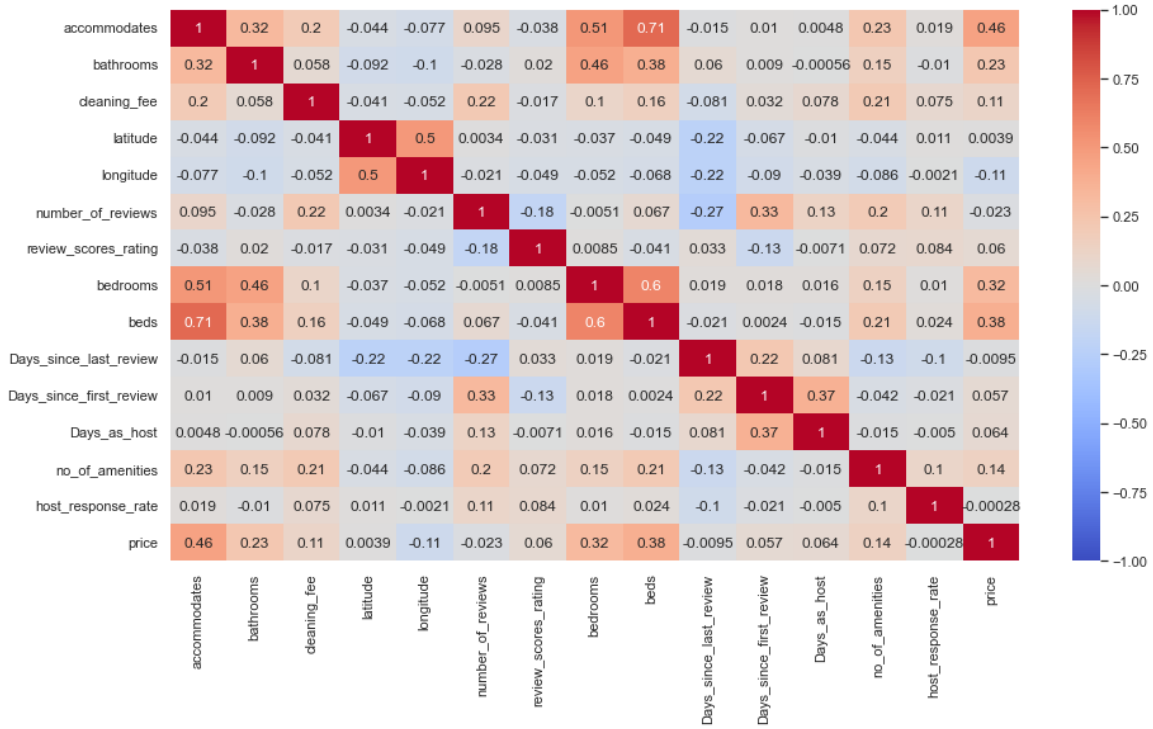

Figure 12: Correlation Heatmap