Predicting The Price Of Airbnb Listings

Introduction:

As of 2020, Airbnb has 7M+ listings worldwide and 2M+ people staying on Airbnb per night. "We continue to see incredible growth in the Bay Area - even while tourism is essentially flat in the region - with 44 percent growth year-over-year in guest arrivals in the five counties surrounding San Francisco." (new.airbnb.com) Airbnb has become an important source of supplemental income and the company is only continuing to grow, leading to a healthy competition among Airbnb hosts.

Airbnb hosts make an average of $924 per month, according to research from Earnest (low-interest lender). It's a common side-hustle for many and the amount varies between area, renting frequency, quality of your home and amenities provided. Here, I will explore the most important factors that determine the price of an Airbnb rental and how you can maximize your time and money.

Issue:

San Francisco has over 7,800 listings and over 3,700 long term and hotel listings (as of Feb. 2019). How can a host stand out from the crowd and increase revenue? This analysis aims to determine the most important characteristics for a top listing and provide an ideal price range for rentals based on those factors. The increase in revenue for hosts will lead to increased revenue for Airbnb, happier hosts and an improved business model for pricing listings.

Data:

The data is provided on http://insideairbnb.com/

Deliverables:

- Code in Jupyter notebook
- Report that outlines steps taken to determine conclusion
- Slide deck for presentation

The Data

The data were sourced from http://insideairbnb.com/get-the-data.html which is an independent, non-commercial website that utilizes publicly available information about San Francisco's

Airbnbs. The raw dataset contained 17 columns and 8533 records. The data was compiled on December 4th, 2019. I selected the columns below for use in this analysis, dropping any rows that contained null values which came to ~23% of the data.

Selected Columns:

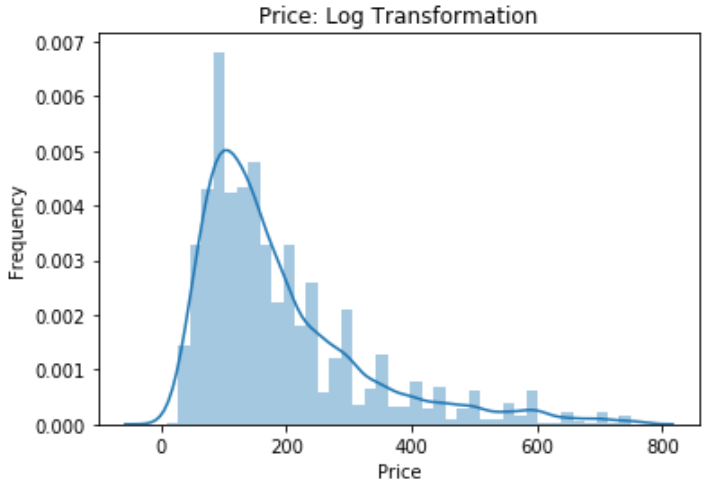
- id
- host_id
- zipcode
- property_type
- room_type
- accommodates
- bathrooms
- bedrooms
- beds
- bed_type
- amenities
- price
- minimum_nights
- availability_30
- number_of_reviews
- review_scores_rating
- calculated_host_listings_count

For a more detailed breakdown of the data wrangling steps, view my Data Wrangling Report.

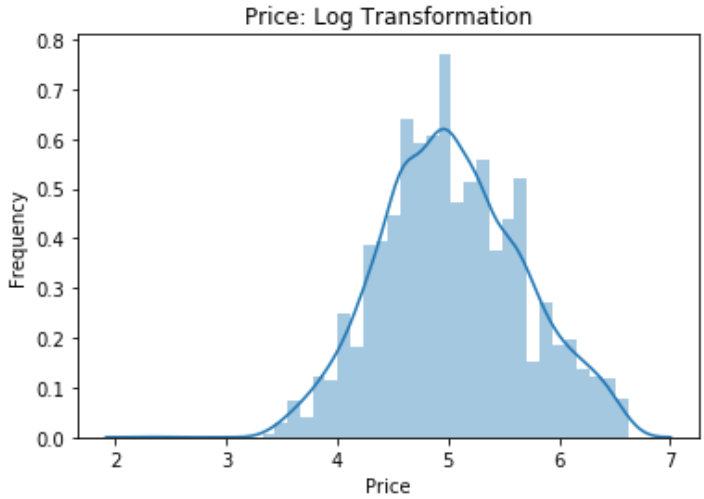
The total amount of null values were negligible, so these records were dropped from the dataframe.

Exploratory Data Analysis (EDA)

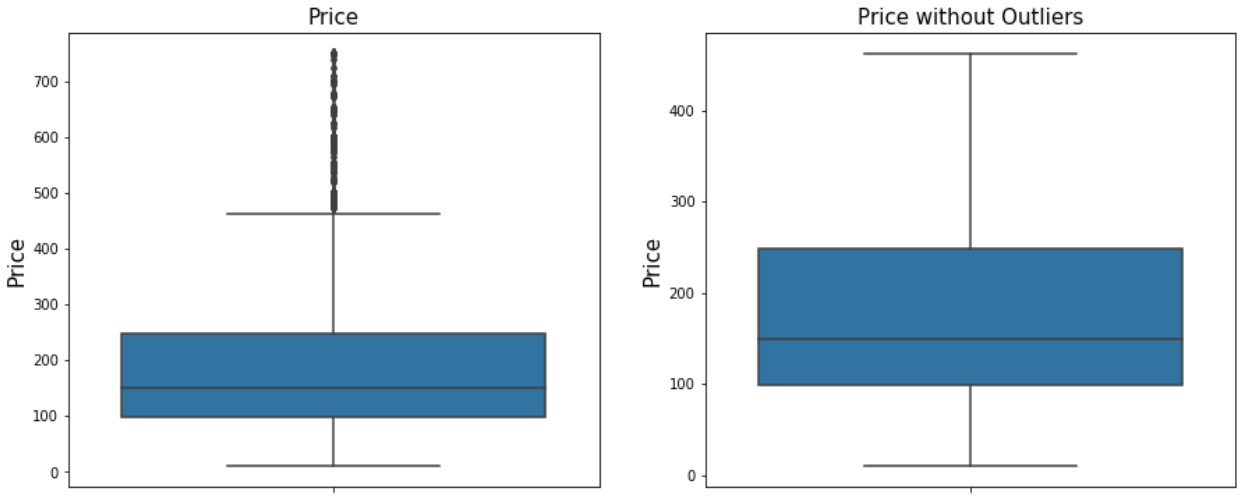
Price



The price histogram shows strongly right-skewed data with a long tail. The majority of prices are very low in comparison to the few high-priced Airbnbs which skew the data.

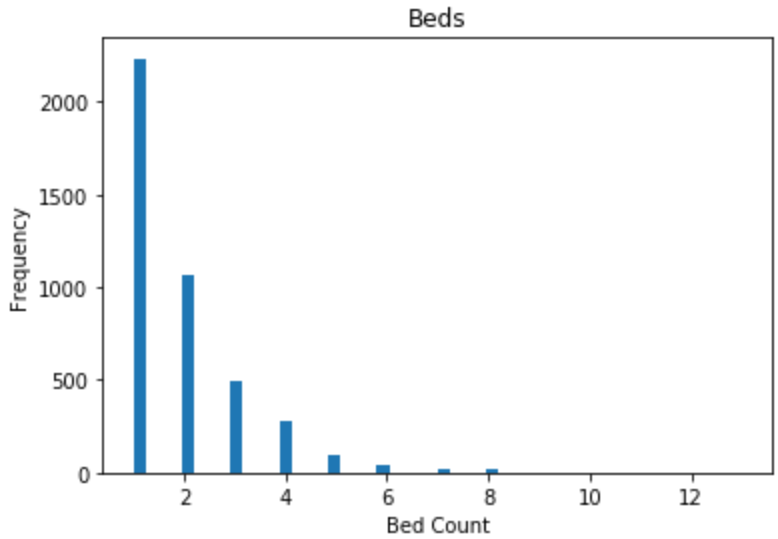


When the price is log-transformed, it becomes with a p-value of 1.7.

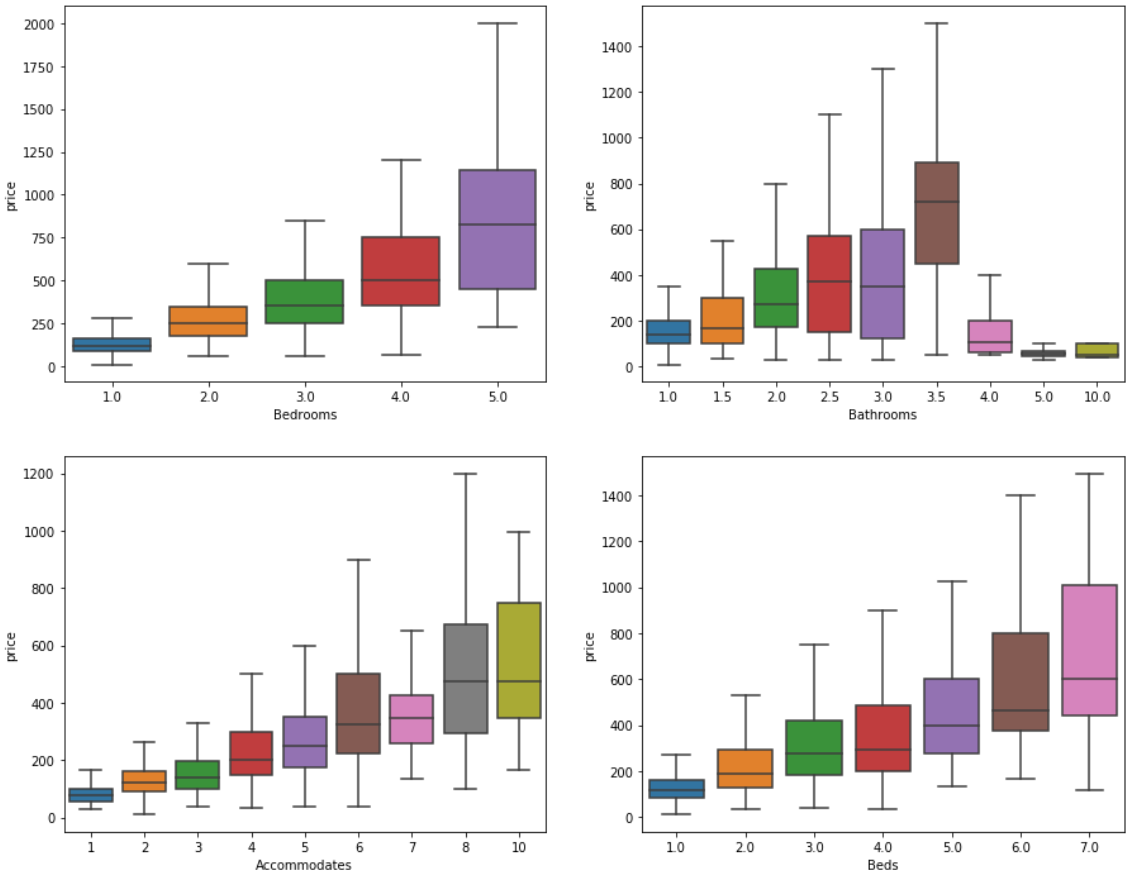


The boxplots above show how highly right-skewed the data is. The boxplot without outliers better depicts the price range between $100 and $250.

Accommodations

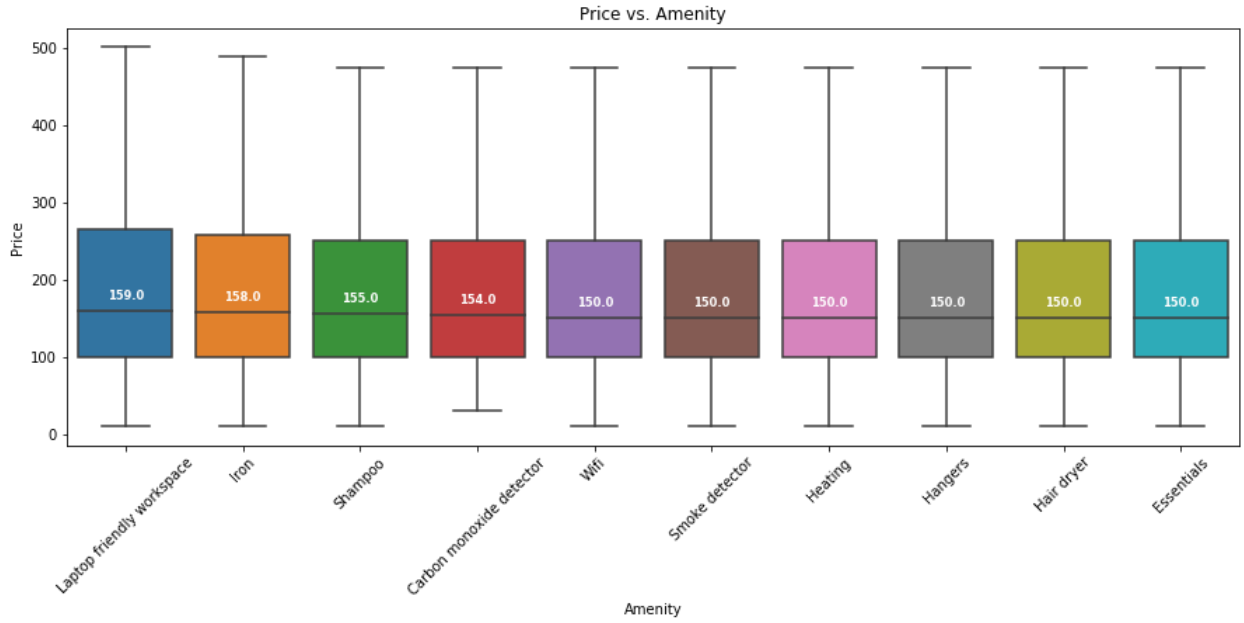


The vast majority listings offer one bed, again showing Airbnb strongly caters to solo and coupled guests. The amount of listings decreases as each bed count increases.



Bathrooms appear to be less correlated to price compared to bedrooms, accommodates and beds. All values with a sample size less than 20 were omitted from the boxplots. In addition, all features are significantly correlated with price based on the p-values obtained from the Point Biserial Correlation.

Amenities

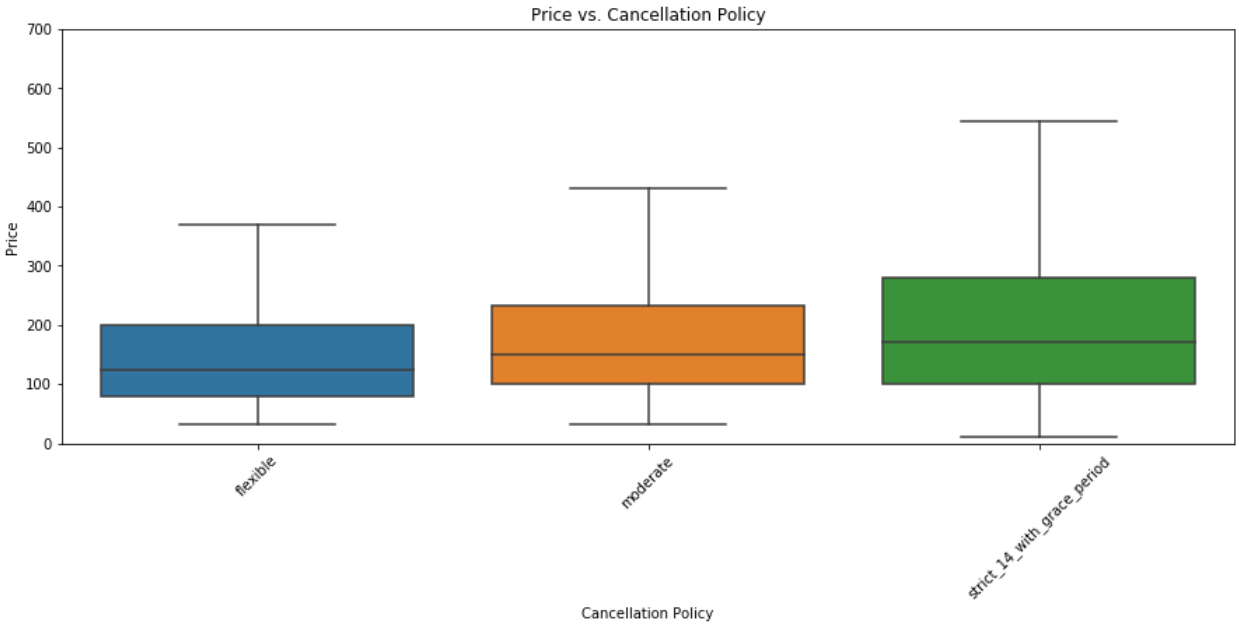


There is no visible difference between the top 10 amenities. All amenities except wifi and essentials are all significantly correlated with price based on the p-values obtained from the Point Biserial Correlation.

Room Type

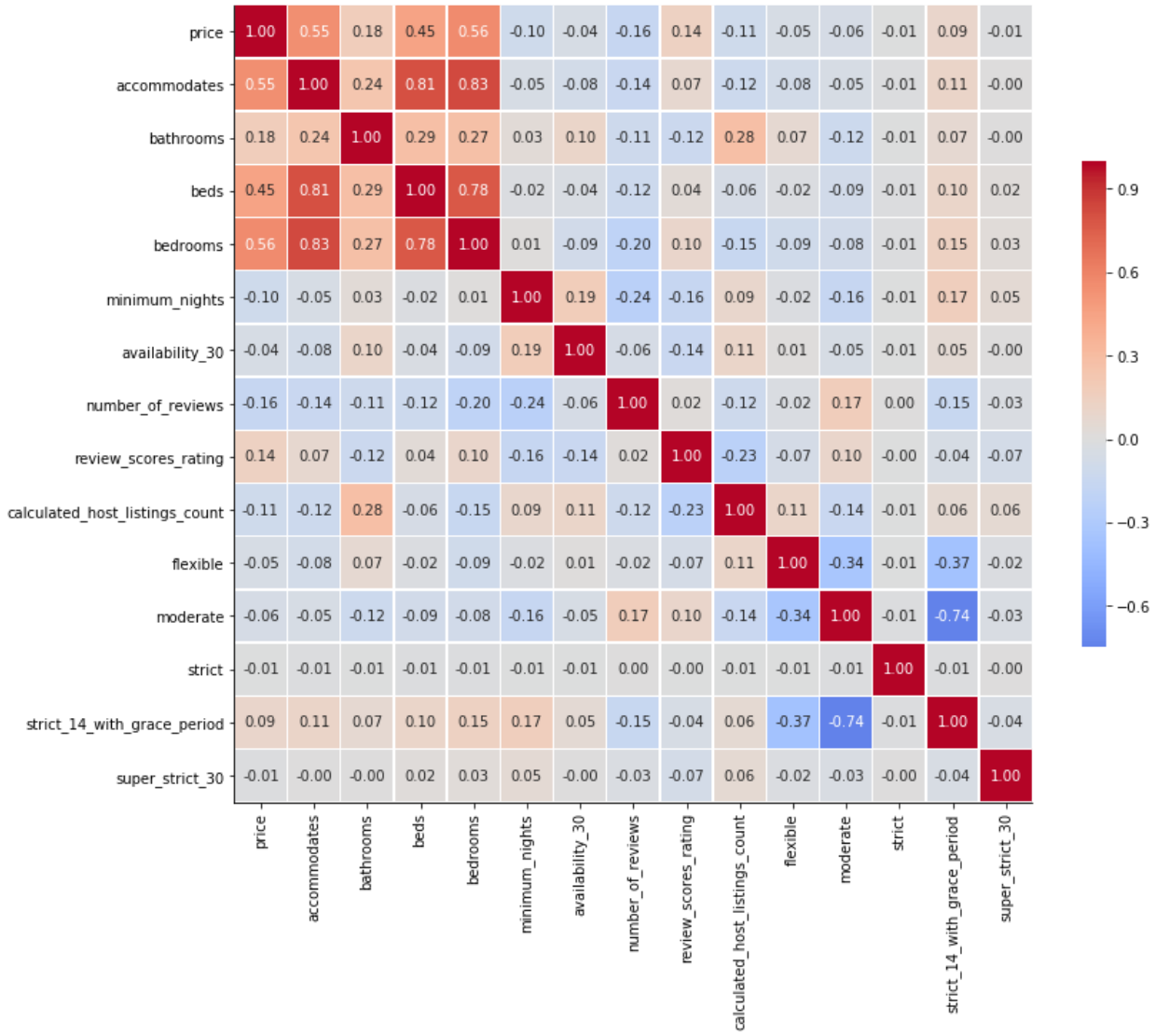
The size of room type appears to be highly correlated with price. Price increases as more privacy is given based on room type space.

Cancellation Policy



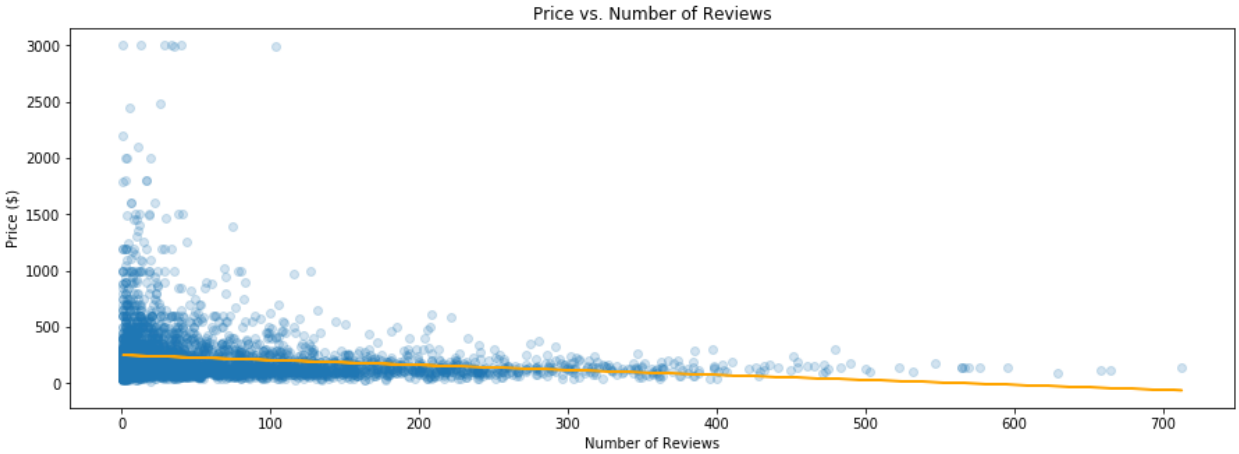
There appears to be a direct correlation between the price and level of cancellation policy, with prices increasing with a stricter policy. There is also a steadily increasing IQR as the cancellation policy becomes more strict.

Correlation Matrix between Price and Features:



Accommodates, bathrooms, beds and bedrooms are all highly correlated with each other which may be an issue in terms of multicollinearity. To combat this, I removed the "accommodates" column from the dataframe when testing machine learning models. The cancellation policies (flexible, moderate, strict, strict_14_with_grace_period and super_strict_30) are categorical data (hence correlation isn't technically correct), but included in the heatmap to get an overall sense of the differences between features and price.

Number of Reviews



The scatter plot above shows the number of reviews are highest for listings lowest in price.

Determining Feature Importance

After EDA, I used a linear regression in order to determine feature importance. First, however, I needed to examine the VIF scores to ensure the coefficients could be trusted to accurately represent the effects of the features on the target variable.

VIF

VIF is used to determine the reliability of the coefficients in a linear model, which will be essential in determining feature importance.

	VIF Factor	features
0	296.9	Intercept
1	1.1	availability_30
2	1.2	bathrooms
3	1.2	bedrooms
4	1.2	calculated_host_listings_count
5	1.2	flexible
6	1.1	minimum_nights
7	1.2	moderate
8	1.1	number_of_reviews
9	1.1	review_scores_rating

I decided to drop “beds” from the above analysis because it was highly correlated with “bedrooms”. Then, I ran an OLS regression using unscaled data and since the VIF scores are all below 10, the value of the coefficients should reasonably represent the effect of a one-unit increase of the feature on the target variable.

	features	estimatedCoefficients
0	bathrooms	1.008190
1	bedrooms	105.600163
2	minimum_nights	-1.146710
3	availability_30	0.070600
4	number_of_reviews	-0.142420
5	review_scores_rating	1.975249
6	calculated_host_listings_count	-0.462428
7	flexible	-19.347911
8	moderate	-5.986784

As we can see, for every new “bedroom” our model predicts a $106 increase in price, while a “flexible” cancellation policy is generally associated with a $19 decrease in the price of a rental.

Below is the OLS regression using scaled data.

	features	estimatedCoefficients
0	bathrooms	0.545889
1	bedrooms	103.906614
2	minimum_nights	-28.858344
3	availability_30	-0.280412
4	number_of_reviews	-12.878931
5	review_scores_rating	9.548635
6	calculated_host_listings_count	-1.226265
7	flexible	-17.967268
8	moderate	-3.842164

The coefficients have slightly changed from before. The model predicts every new “bedroom” to be a $104 increase in price and the “flexible” cancellation policy is now a $18 decrease in the price of a rental. The biggest feature coefficient changes appear to be “number of reviews”,

“review scores rating” and “moderate” with a negative $13, positive $10 and positive $4 per unit change, respectively.

Predictive Modeling

Next, I built a variety of predictive Machine Learning models and optimized them using Gridsearching, then compared the results. First, I completed an OLS regression and found bathrooms, availability_30 and moderate are not a statistically significant predictor of price.

In the beginning of machine learning, I explored the dataset and found bedrooms to be the most important feature, with a $106 increase in price for every unit increase. Beds appeared to be correlated to bedrooms, so I dropped this feature from the regression model. Then, I plotted the high leverage points and removed these from the dataframe. Next, I applied a Robust Scaler on the data and used this cleaned up dataframe to run multiple regression models.

Model Comparison:

	Model	Parameters	R^2 Score
0	Linear	N/A	0.427
1	Lasso	alpha = .01	0.402
2	Random Forest	in PDF	0.375
3	Ridge	alpha = 10	0.427
4	Elastic-Net	alpha = .01; l1_ratio = .4	0.426
5	KNN	n = 6	0.450

Model	Parameters	R^2 Score
Linear	N/A	.427
Lasso	alpha = .01	.402
Random Forest	Min_samples_split = 6; n_estimators = 150; max_features = sqrt	.375
Ridge	alpha = 10	.427
Elastic-Net	alpha = .01; l1_ratio = .4	.426
KNN	n = 6	.450

Conclusion

Ultimately, an Airbnb host will maximize the amount of money made by catering to a larger amount of guests in their home. Price depends heavily on the bedrooms offered and amount of beds available per home. One suggestion is offering the guest 2 beds in 1 room, which I have encountered in my own Airbnb experiences. Also, the price does not seem to depend heavily on the choice of cancellation policy, so Airbnb hosts should choose what fits best for them and can be rest assured it will not be a determining factor in how much they make for renting out their home.

In the modeling, the KNN regression performed the best with an R squared score of 45.0%, which was slightly increased from before scaling and grid-searching the features. In the future to improve this model, I would likely go back and explore different feature sets. In particular, I would use natural language procession to explore the text columns that were removed from my dataframe. It would be interesting to see the relationship between customer reviews, amenities notes and host reviews. I would also choose less features and explore the dataset in more depth. There were heavy outliers in my dataset, but the majority of the data was in a small price range. The top feature for Airbnb hosts is bedrooms, which seems obvious, so I would be interested in exploring the dataset more in the future to see if there are unexpected features that influence price as well.