Airbnb Pricing Analysis

Mengqin Gong
June 2019

Problem

Airbnb provides travellers with the opportunity to stay in a unique home and to deeply experience the local life.

This analysis focuses on one of the most important aspects of an Airbnb listing - the price.

I’m going to build a model that can predict the price of a listing in **San Francisco** based on its location, amenities, reviews and any other information that may apply.

Data

San Francisco Airbnb Open Data:

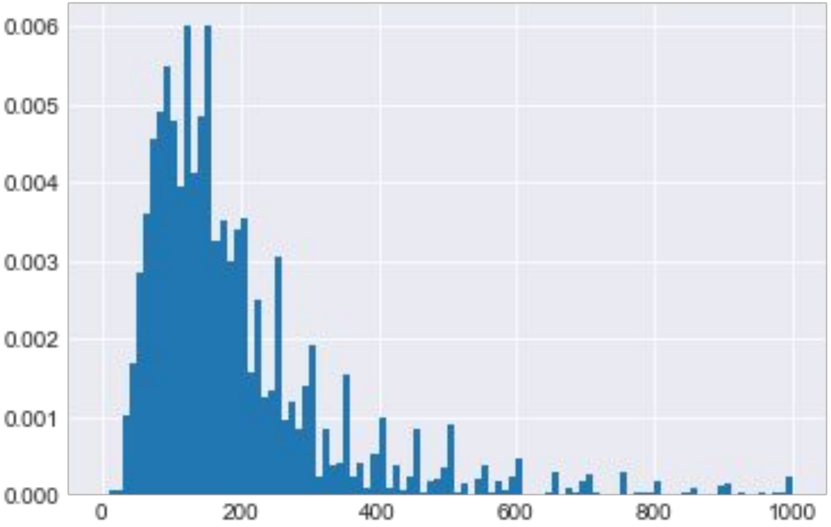
http://insideairbnb.com/get-the-data.html

- 7151 listings in San Francisco
- 106 columns
  - Listing price
  - Host info
  - Amenities
  - Property info
  - …

Exploratory Data Analysis

Price

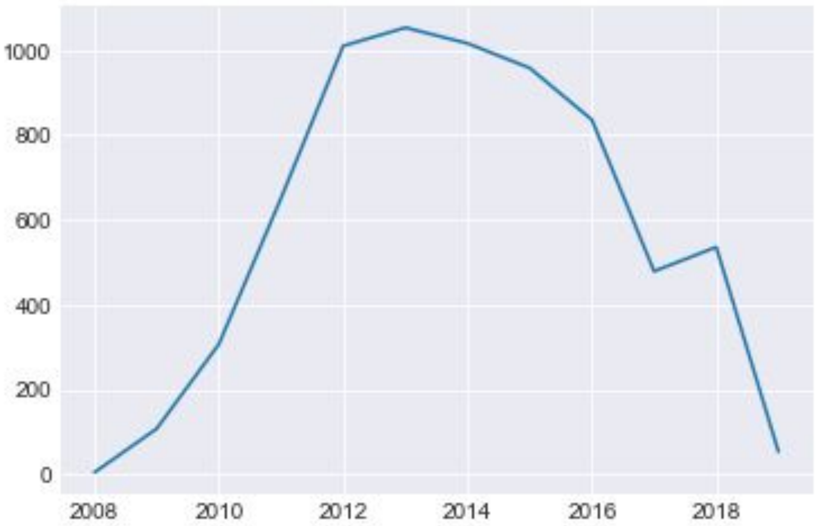
Histogram of listing price:



The distribution of price is right skewed with most listings priced under 200 dollars.

Host start year

Number of hosts joined per year:

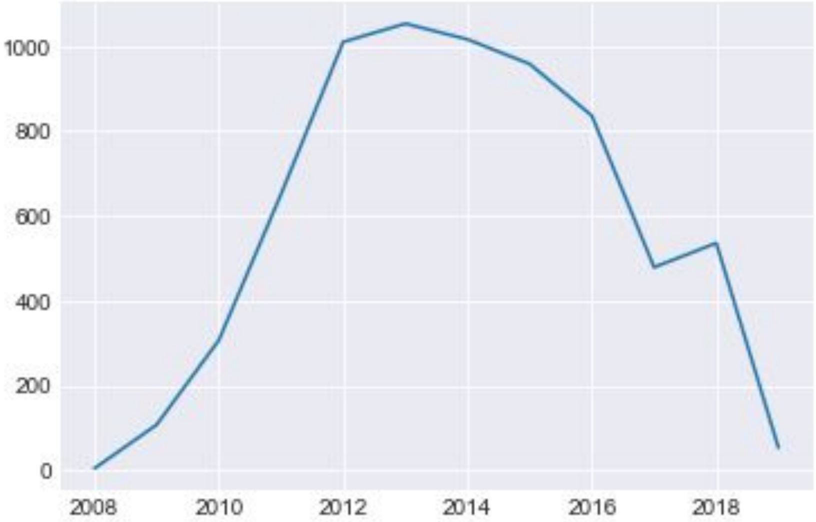


In 2008-2012 there was a fast increase in new hosts, and 2013 had the most newly-joined hosts. After 2013, number of new hosts dropped until 2018.

The trend seems to match the home price trend, in 2008-2013 there was a continuous decline in home price after the market crash in 2008, and starting from 2013, the home market in San Francisco recovered with price increasing rapidly.

Host start year

Number of hosts joined per year:



This may indicate that the number of new hosts in Airbnb market is negatively related to the home price index, when home price is increasing, owners may tend to sell their extra properties instead of renting them out.

Another possible explanation is that, after Airbnb was founded in 2008, the company experienced an increase in new hosts for 5 years till 2013, and the market calmed down and number of new hosts dropped after 2013.

Superhost

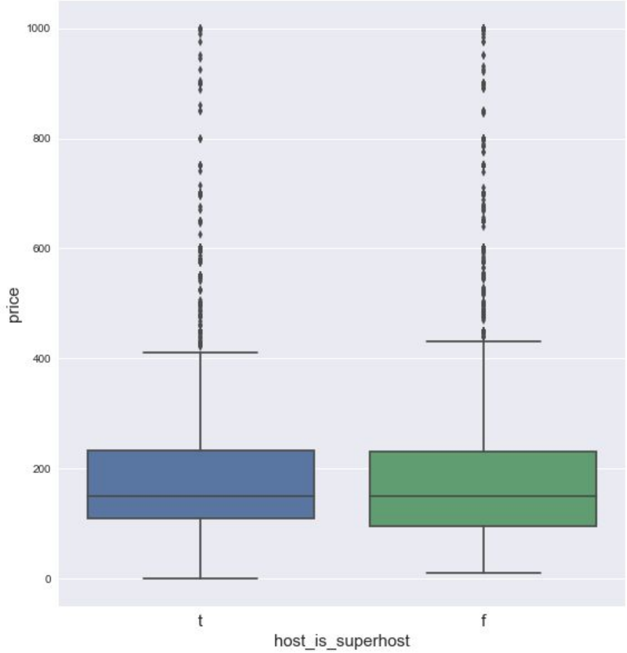
Superhosts are hosts who meet certain criteria such as

- 4.8+ overall rating
- 10+ stays during the past year
- etc.

Airbnb states that superhosts earn up to **22%** more than other hosts.

Superhost

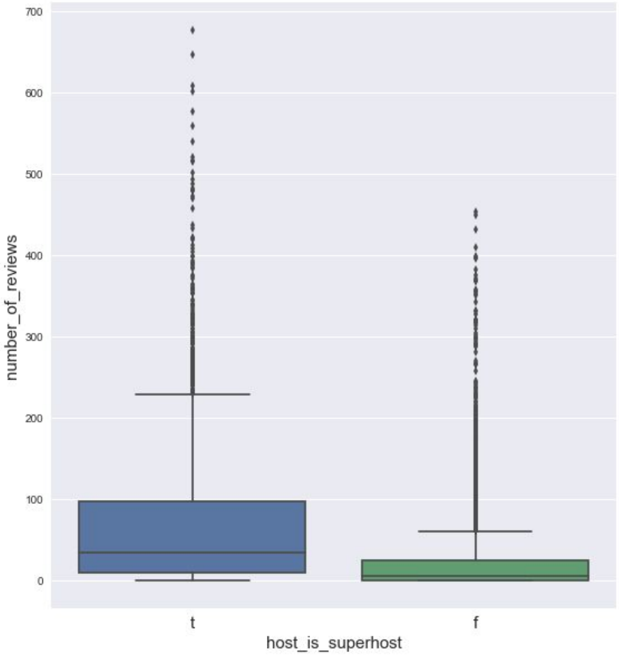
Listing price vs. host_is_superhost (t: superhost, f: non-superhost)



The distribution of price is quite similar regardless of superhost or not.

Superhost

Number of reviews vs. host_is_superhost (t: superhost, f: non-superhost)

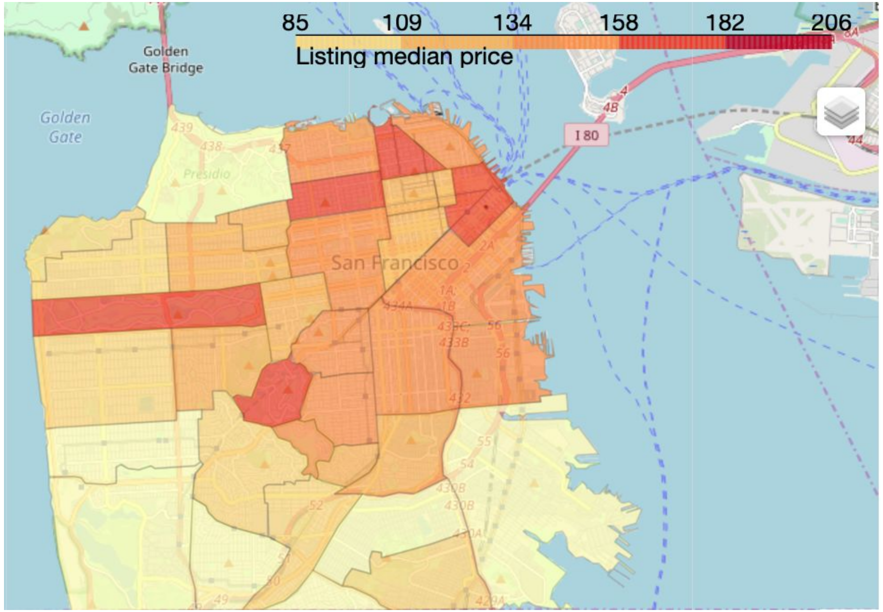


Superhosts have a higher median number of reviews (which is related to the number of bookings) than non-superhosts.

Therefore superhost may not be related to higher price or lower price, but may be related to more reviews (more bookings).

Neighborhood

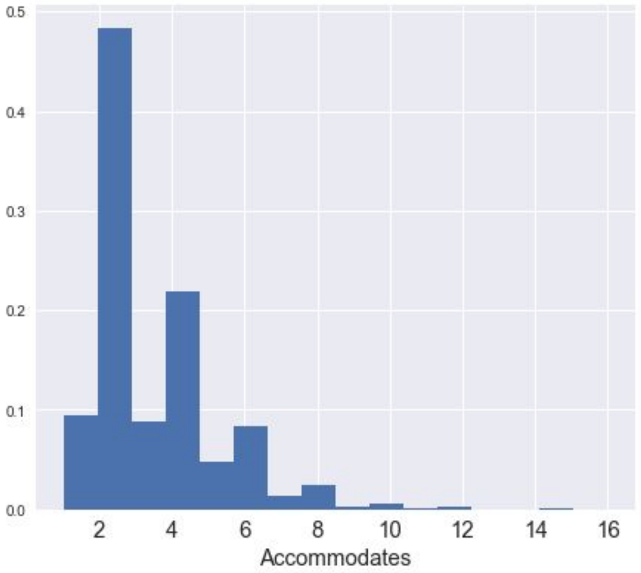
Listing median price by neighbourhoods



Listings in the northeastern part of SF have a relatively higher median price, and southern part has the lower median listing price.

Number of people can accommodate

Histogram:

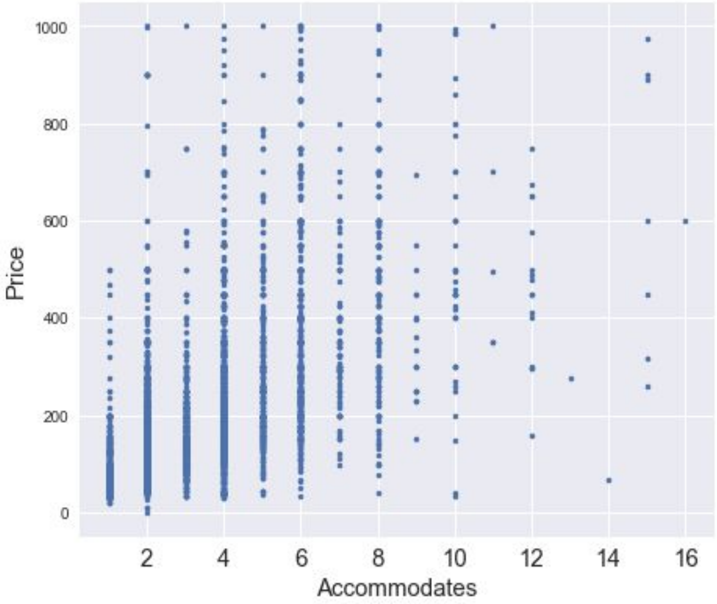


The number of people a listing can accommodate reflects the size of the listing.

Nearly half of the listings can accommodate only 2 people, and around 80% of the listings can accommodate 4 or less people.

Number of people can accommodate

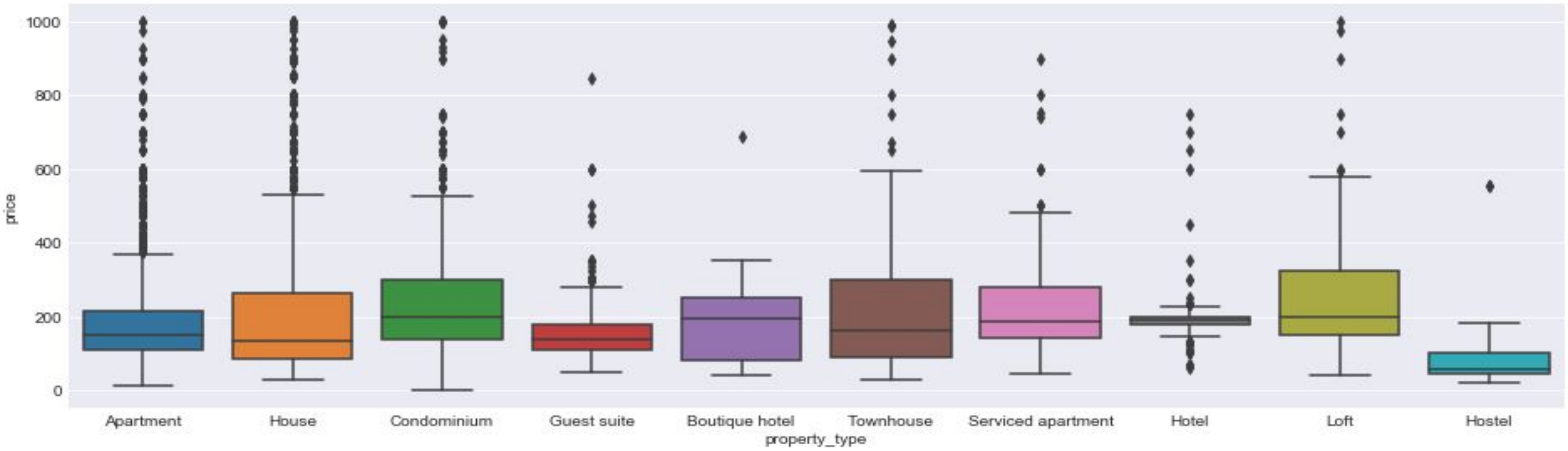
Price vs. accommodates:



There is a positive relationship between price and accommodates.

The relationship between price and accommodates is intuitive: the more people can a listing accommodate, the higher the price tends to be.

Property type

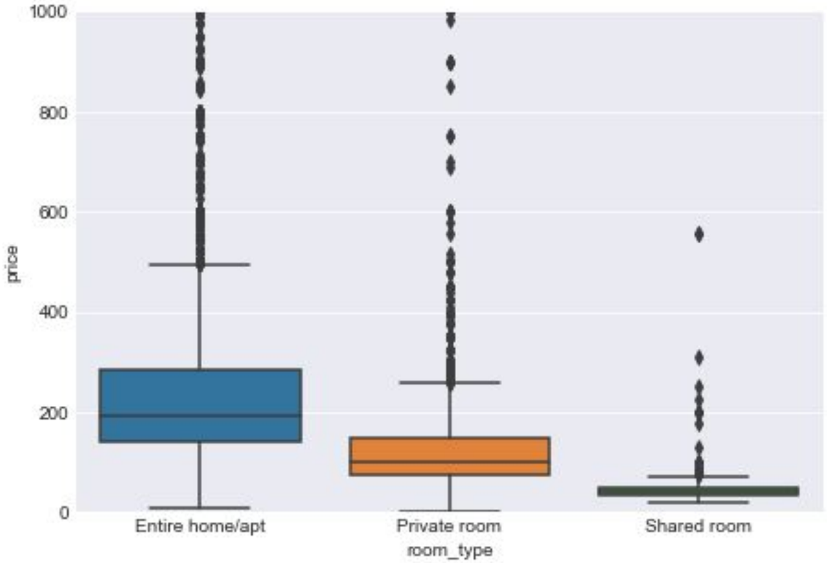


The boxplot shows the distribution of listing prices by property type.

We can see that condominiums, boutique hotels, and loft have relatively higher median price, and hostels are usually lower priced.

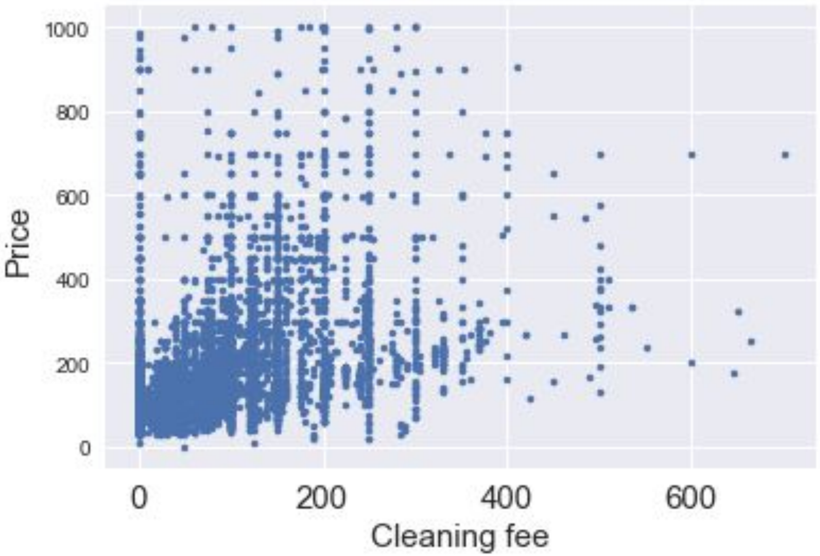
Room type

Price by room type:



The price of shared rooms seems to be really nice - the median price is about 1/4 of that of entire home/apt!

Cleaning fee

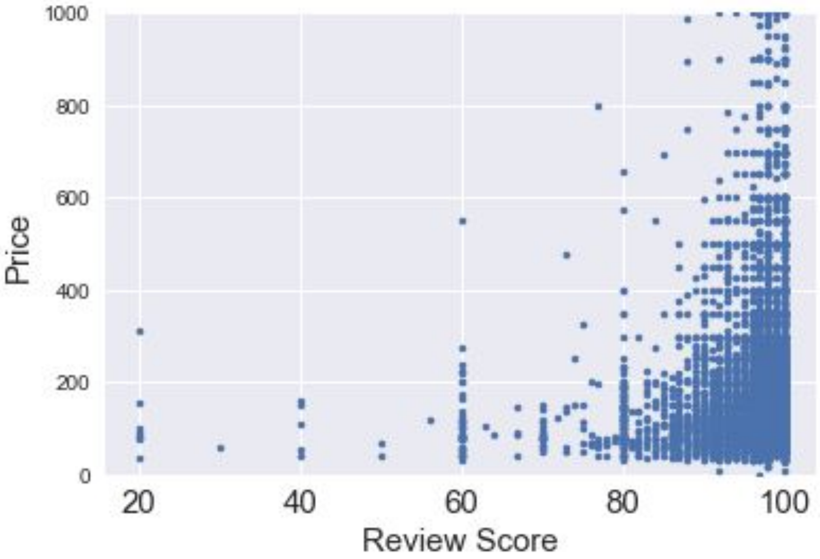


Cleaning fee is an extra cost for Airbnb customers.

The scatterplot shows that there is a positive correlation between cleaning fee and price.

It is intuitive as the more cleaning fee it charges, the more labor it requires to clean the listing, and usually it indicates that the listing is either large or complicated.

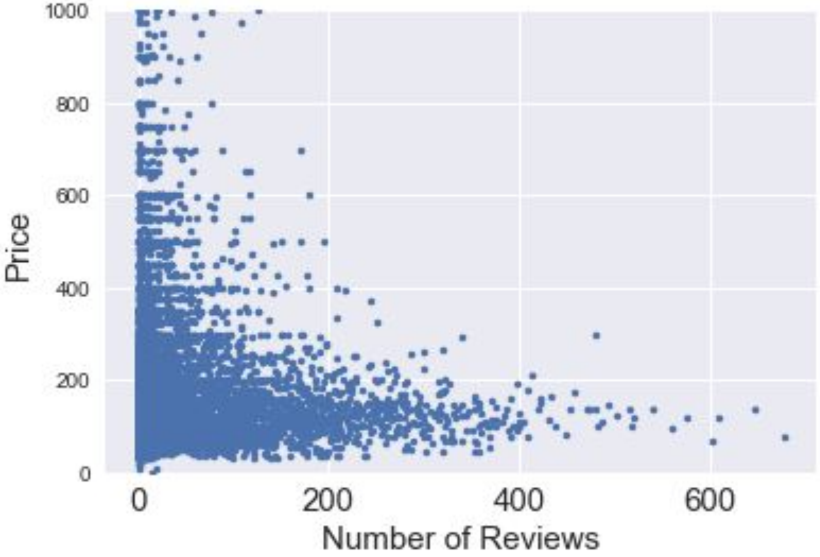
Review score



Review score indicates how customers like the listing.

The scatter plot shows a positive correlation between them, that is, a higher rated listing is likely to have a higher price.

Number of reviews



Number of reviews can somewhat represent the popularity of the listing.

The scatterplot shows that price is negatively correlated with number of reviews, which indicates that people prefer booking listings with lower price.

Modeling

Process

Data processing:

- Training set (70%)
- Testing set (30%)
- K-Fold Cross Validation is applied on training set

Feature selection/reduction:

- Backward elimination
- Principal Component Analysis (PCA)

Modeling:

- Linear Regression
- Random Forest Regression

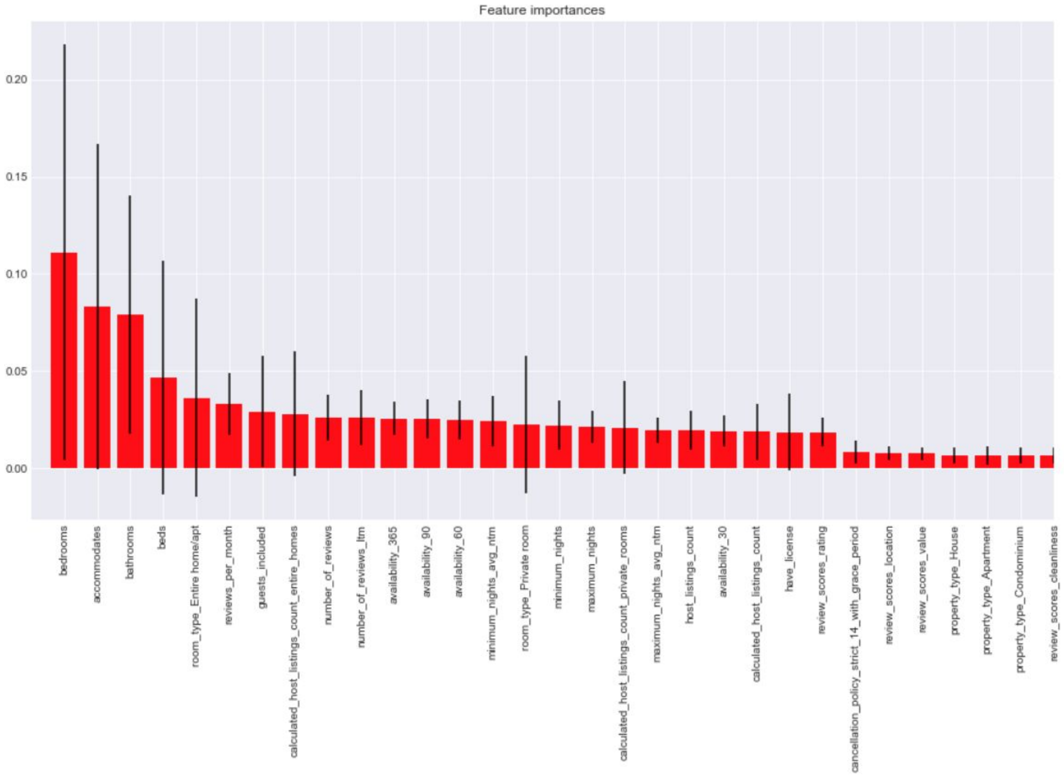
Model Evaluation and Selection

Metrics

Model	Details	R^2 training	R^2 testing	MSE training	MSE testing
1	Linear Regression (LR) on numeric features	0.4791	0.4981	11120.12	10540.95
2	LR on Standard scaled numeric data	0.4791	0.4981	11119.31	10542.13
3	LR on MinMax scaled numeric data	0.4791	0.4981	11119.31	10542.13
4	LR on Robust scaled numeric data	0.4791	0.4981	11120.12	10540.95
5	LR on normalized numeric data	0.3089	0.3392	14699.55	10629.94
6	LR on non-collinearity numeric data	0.4751	0.4939	11200.50	10629.94
7	LR on numeric data after PCA	0.0914	0.0986	19259.13	18931.74
8	LR on data from model 6 plus 2 categorical features	0.5715	0.5940	9553.42	8526.51
9	Random Forest Regression on numeric and categorical features	0.9472	0.6315	1111.65	7739.23

Conclusion

Feature importances



Feature ranking (top 10):

1. bedrooms (0.111054)

2. accommodates (0.083135)

3. bathrooms (0.079152)

4. beds (0.046686)

5. room_type_Entire home/apt (0.036040)

6. reviews_per_month (0.032954)

7. guests_included (0.029107)

8. calculated_host_listings_count_entire_homes (0.027959)

9. number_of_reviews (0.025851)

10. number_of_reviews_ltm (0.025715)

Feature importances

The top 5 features are:

- number of bedrooms
- number of people the listing can accommodate
- number of bathrooms
- number of beds
- and type of the room

So we can conclude that the size and capacity of a listing play the most important role in pricing, other features such as number of reviews, number of listings a host has (may be an indicator of if the host is experienced) and number of guests included also contribute to the price of a listing.

Limitations and Next Steps

Limitation

There is an **overfitting problem** in the Random Forest Regression model:

- R^2 on training set is 0.95 vs. that on testing set is 0.63
- MSE of prediction on testing set is much greater than that on training set.

This is mainly because the minimum number of samples required at each leaf node is lowered to 1. Increasing the minimum_samples_leaf can reduce the overfitting problem, however, the prediction R^2 drops and MSE increases at the same time.

Next steps

Following are some more approaches to try in the future:

- Creating a model with top features so that overfitting problem may be reduced
- Grouping categorical levels based on some specific rules may help us find the underlying relationship
- Categorizing price into several levels and turn the problem into a classification problem

Thank you.