A Statistical Analysis of Airbnb Data

Math 533 Statistical Learning

Seth Arreola

2022-11-25

Section 1

Back-ground

Back-Ground

Airbnb - "Air Bed and Breakfast," is a service that lets property owners rent out their spaces to travelers looking for a place to stay. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves.

Back-Ground

What makes Airbnb interesting to study?

Communities of people have shared the use of assets for thousands of years, but the advent of the Internet and its use of big data has made it easier for asset owners and those seeking to use those assets to find each other. This sort of dynamic can also be referred to as the **shareconomy**, or a sharing economy.

- physical assets $\implies$ shared as services
- ride-sharing, short-term rentals, coworking, grocery delivery services
- Many arguments for and against this type of market

Back-Ground

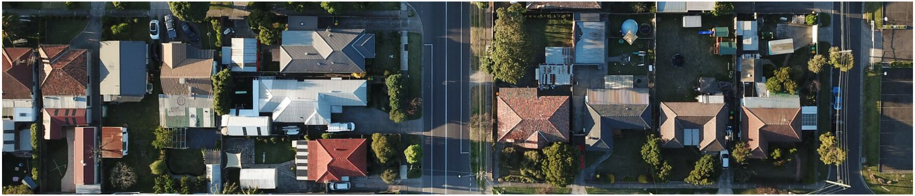
Due to the relative “newness” of many sharing-economy companies, many things are not ironed-out.

- Recent research argues that almost all hosts fail to maximize their potential profit due to poorly pricing their listing (Gibbs et al., 2018).

Airbnb recommends:

- “Do a little market research”, “Consider your location and hospitality”, “Stand out with great pricing”.
- Use the **Smart Pricing Tool**, however is requires a minimum price be set by user.

Our Goal



In this analysis we aim to:

- Develop a statistical model to predict the price of a given Airbnb listing
- Identify, if any, listing features that contribute to price
- We will focus on the San Francisco Airbnb market

Litature Review

- Gibbs C., Guttentag D., Gretzel U., Morton J. and Goodwill,A. (2018), "Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings", Journal of Travel & Tourism Marketing
- Daniel J. Stekhoven, Peter Buhlmann (2011), "MissForest-non-parametric missing value imputation for mixed-type data", Journal of Bioinformatics
- Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, Hoormazd Rezaei, "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis", Stanford University

Section 2

The Data

The Data

The data was retrieved from Inside-Airbnb
http://insideairbnb.com/get-the-data/

Inside Airbnb
Adding data to the debate

Data About Support Organise Donate!

San Francisco, California, United States

Explore the San Francisco data.

Date Compiled	Country/City	File Name	Description
07 September, 2022	San Francisco	listings.csv.gz	Detailed Listings data
07 September, 2022	San Francisco	calendar.csv.gz	Detailed Calendar Data
07 September, 2022	San Francisco	reviews.csv.gz	Detailed Review Data
07 September, 2022	San Francisco	listings.csv	Summary information and metrics for listings in San Francisco (good for visualisations).
07 September, 2022	San Francisco	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	San Francisco	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	San Francisco	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

show archived data
(generally quarterly data for the last 12 months. For additional data, make an archived data request.)

The Data

The Names of our Features in the Data		
First 25 Features	Second 25 Features	Last 25 Features
id	host_has_profile_pic	has_availability
listing_url	host_identity_verified	availability_30
scrape_id	neighbourhood	availability_60
last_scraped	neighbourhood_cleansed	availability_90
source	neighbourhood_group_cleansed	availability_365
name	latitude	calendar_last_scraped
description	longitude	number_of_reviews
neighborhood_overview	property_type	number_of_reviews_ltm
picture_url	room_type	number_of_reviews_l30d
host_id	accommodates	first_review
host_url	bathrooms	last_review
host_name	bathrooms_text	review_scores_rating
host_since	bedrooms	review_scores_accuracy
host_location	beds	review_scores_cleanliness
host_about	amenities	review_scores_checkin
host_response_time	price	review_scores_communication
host_response_rate	minimum_nights	review_scores_location
host_acceptance_rate	maximum_nights	review_scores_value
host_is_superhost	minimum_minimum_nights	license
host_thumbnail_url	maximum_minimum_nights	instant_bookable
host_picture_url	minimum_maximum_nights	calculated_host_listings_count
host_neighbourhood	maximum_maximum_nights	calculated_host_listings_count_entire_homes
host_listings_count	minimum_nights_avg_ntm	calculated_host_listings_count_private_rooms
host_total_listings_count	maximum_nights_avg_ntm	calculated_host_listings_count_shared_rooms
host_verifications	calendar_updated	reviews_per_month

The Data

The dimensions of the data:

With respect to	Dimention
Rows	6629
Cols	75

Variable types:

Number of Chr	Number of dbl	Number of lgl	Number of date
25	37	8	5

Section 3

Cleaning and Wrangling

Cleaning ans Wrangling

Many columns had redundant information, and were removed

First 10 Removed	Second 10 Removed	Third 10 Removed	Last 3 Removed
id	host_verifications	host_total_listings_count	availability_30
listing_url	neighbourhood_group_cleansed	calculated_host_listings_count_entire_homes	availability_60
scrape_id	bathrooms	calculated_host_listings_count_private_rooms	availability_90
source	calendar_updated	calculated_host_listings_count_shared_rooms	
picture_url	license	minimum_minimum_nights	
host_id	last_scraped	maximum_minimum_nights	
host_url	neighbourhood	minimum_maximum_nights	
host_location	calendar_last_scraped	maximum_maximum_nights	
host_thumbnail_url	host_neighbourhood	minimum_nights_avg_ntm	
host_picture_url	host_listings_count	maximum_nights_avg_ntm	

Resulting in

With respect to	Dimention
Rows	6629
Cols	42

Cleaning and Wrangling

Multiple variables which should be numeric were given as strings, for example

price	host_response_rate
$1,149.00	100%

Cleaning and Wrangling

The following date variables: calendar_last_scraped, last_review, host_since, first_review, Were used to create:

- months_since_last_review
- months_being_host
- months_till_first_review

Cleaning and Wrangling

Some text variables required far more cleaning.

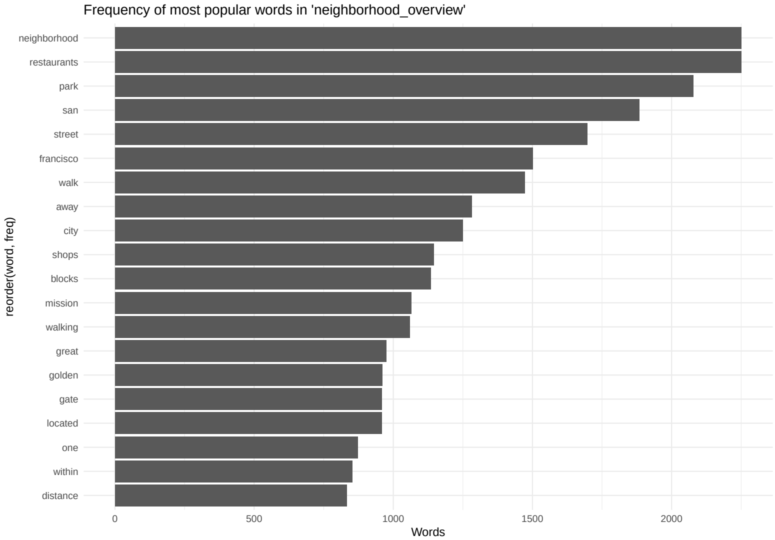
- description
- neighborhood_overview
- host_about
- amenities

For example, one observation of “neighborhood_overview”:

- “Quiet cul de sac in friendly neighborhoodSteps away from grassy park with 2 playgrounds and Recreational CenterVery family-friendly neighborhoodQuaint shops, grocery stores and restaurants all within a 5-10 minute walk”

Cleaning and Wrangling: neighborhood_overview

Cleaning and Wrangling: neighborhood_overview



Cleaning and Wrangling: neighborhood_overview

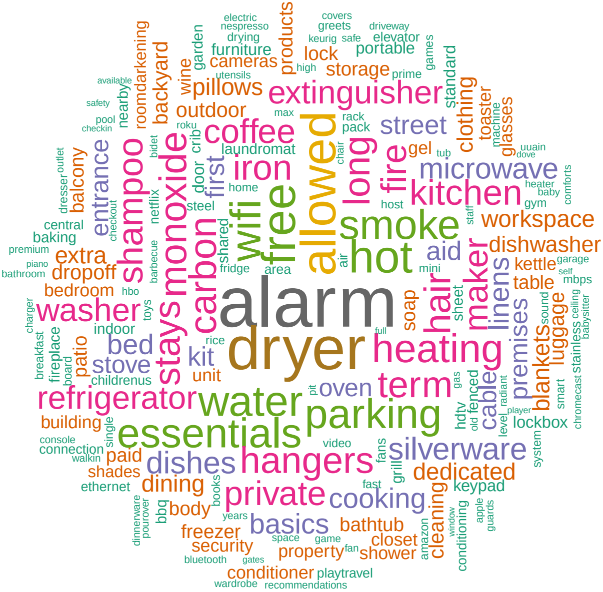
From some of the most common words identified, logical variables were created which identify if a particular word was included in the neighborhood_overview feature.

From neighbourhood_overview, we created:

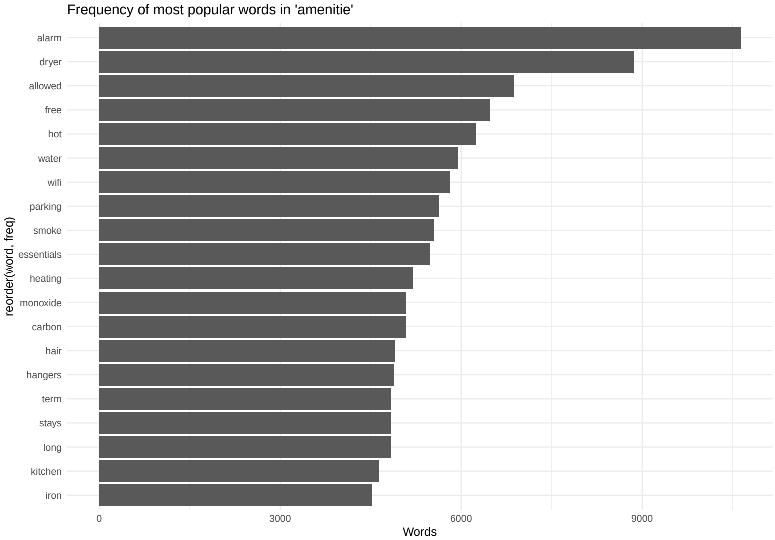
- restaurants_mentioned
- park_mentioned
- walk_mentioned
- shops_mentioned
- quiet_mentioned

Then neighbourhood_overview was removed.

Cleaning and Wrangling: amenities



Cleaning and Wrangling: amenities



Cleaning and Wrangling: amenities

From some of the most common words identified, logical variables were created which identify if a particular word was included in the amenities feature.

From amenities, we created:

- alarm_amenitie
- dryer_amenitie
- wifi_amenitie
- smoke_amenitie
- washer_amenitie

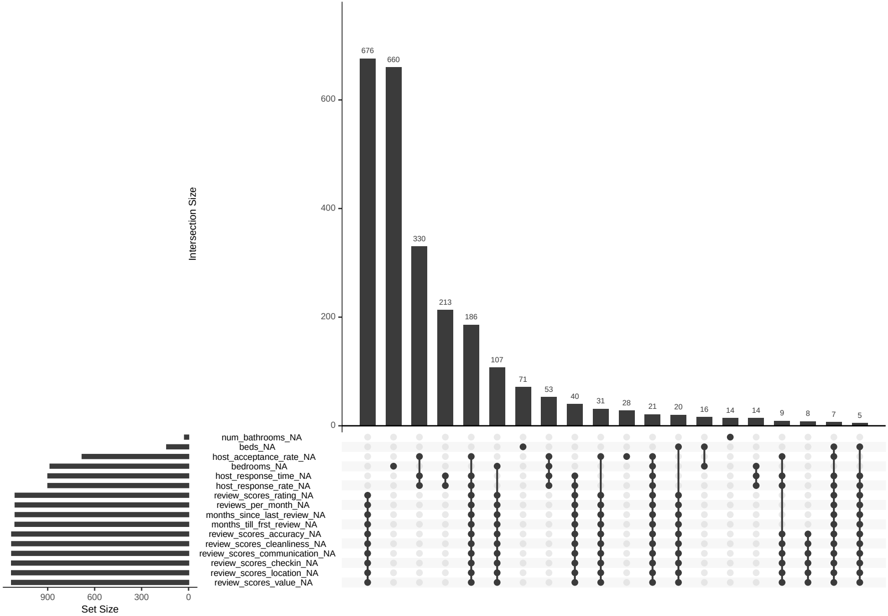
Then amenities was removed.

Cleaning and Wrangling

The other two text variables were not as interesting.

- From “host_about”, a logical feature was created which states if the host mentioned they have a family.
- “description” mainly contained redundant information found in other variables.

Cleaning and Wrangling: Missing values



Cleaning and Wrangling: Missing values

A model based imputation method is selected: MissForest.

- MissForest is a random forest imputation algorithm for missing data.
- It initially imputes all missing data using the mean/mode, then for each variable with missing values, MissForest fits a random forest on the observed part and then predicts the missing part.
- This process of training and predicting repeats in an iterative process until a stopping criterion is met, or a maximum number of user-specified iterations is reached.

The original paper, Stekhoven & Buhlmann (2011), MissForest out-performed many other algorithms, in some cases reducing the imputation error by more than 50%. The primary downside is imputation time.

Cleaning and Wrangling

To summarize, cleaning the data amounted to:

- Coercing character features into numeric and categorical features
- Create Boolean features from text variables
- Impute missing data via random-forest

The resulting data contained 6,629 observations and 53 variables

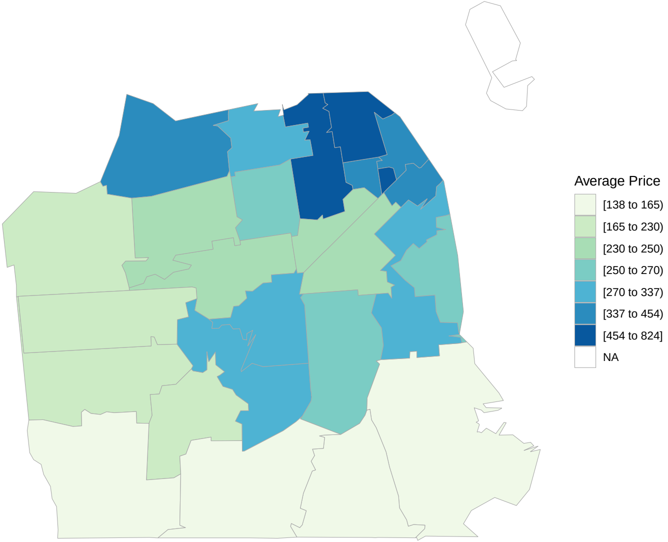
Section 4

Exploritory Data Anlaysis

EDA

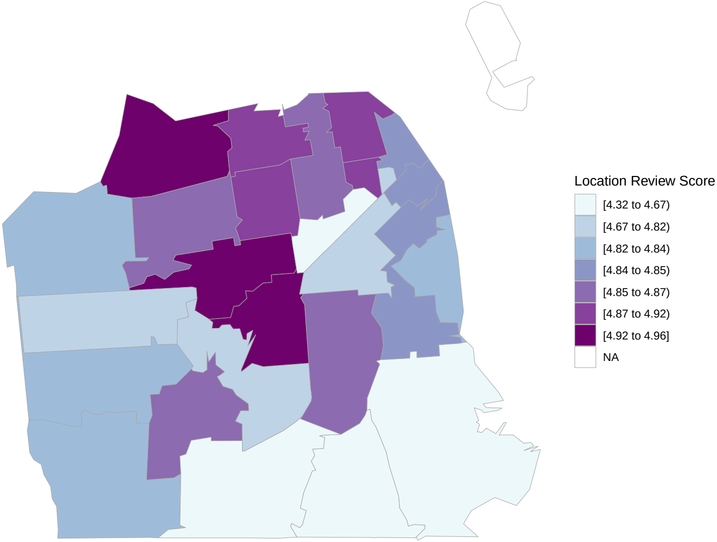
Which area is expensive?

Map showing Average Location Review by Area

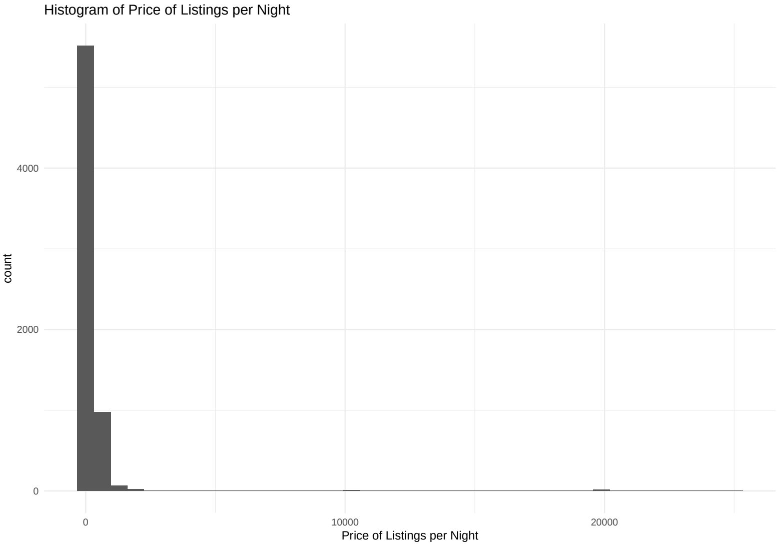


Which area is the best?

Map showing Average Location Score by Area



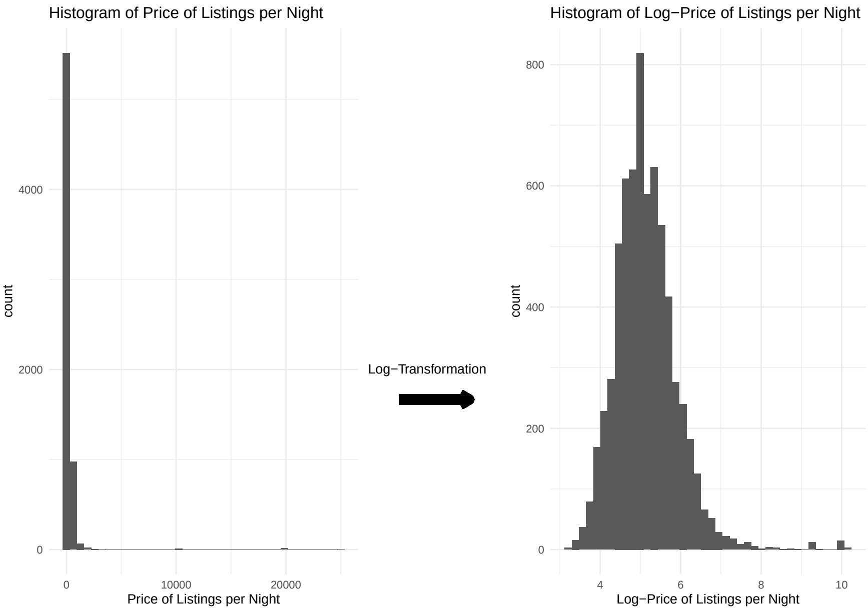
EDA


Histogram of Price of Listings per Night
count
4000
2000
0
0
10000
20000
Price of Listings per Night

EDA

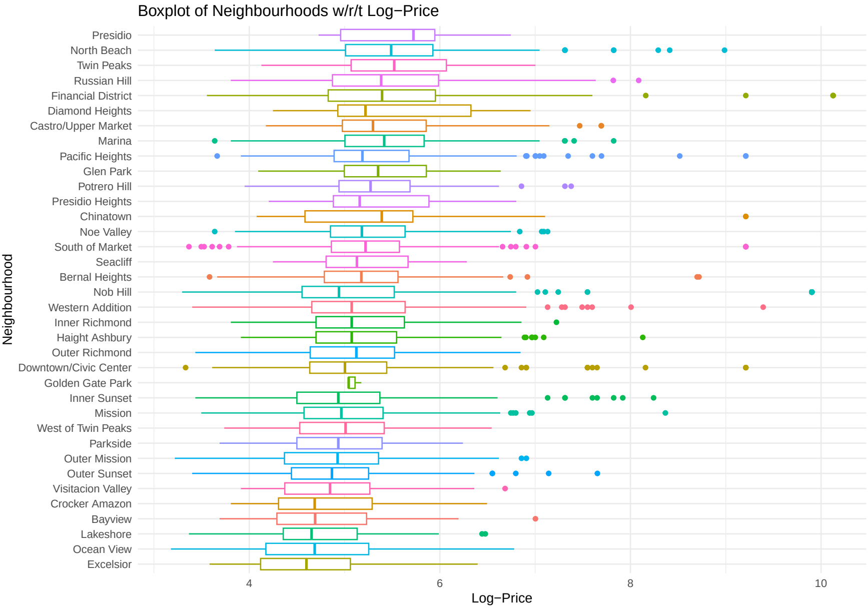

Histogram of Price of Listings per Night
4000
2000
0
y
Very Skewed
A few obs. w/ very high price
0
10000
20000
Price of Listings per Night

EDA



Histogram of Price of Listings per Night
count
4000
2000
0
0
10000
20000
Price of Listings per Night
Log−Transformation
Histogram of Log−Price of Listings per Night
count
800
600
400
200
0
4
6
8
10
Log−Price of Listings per Night

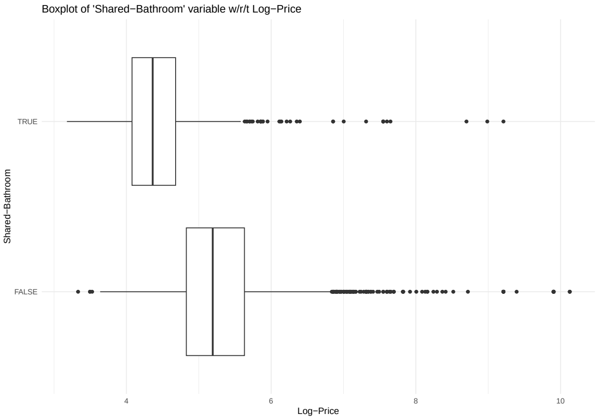
EDA


Boxplot of Neighbourhoods w/r/t Log−Price
Neighbourhood
Presidio
North Beach
Twin Peaks
Russian Hill
Financial District
Diamond Heights
Castro/Upper Market
Marina
Pacific Heights
Glen Park
Potrero Hill
Presidio Heights
Chinatown
Noe Valley
South of Market
Seacliff
Bernal Heights
Nob Hill
Western Addition
Inner Richmond
Haight Ashbury
Outer Richmond
Downtown/Civic Center
Golden Gate Park
Inner Sunset
Mission
West of Twin Peaks
Parkside
Outer Mission
Outer Sunset
Visitacion Valley
Crocker Amazon
Bayview
Lakeshore
Ocean View
Excelsior
4
6
8
10
Log−Price

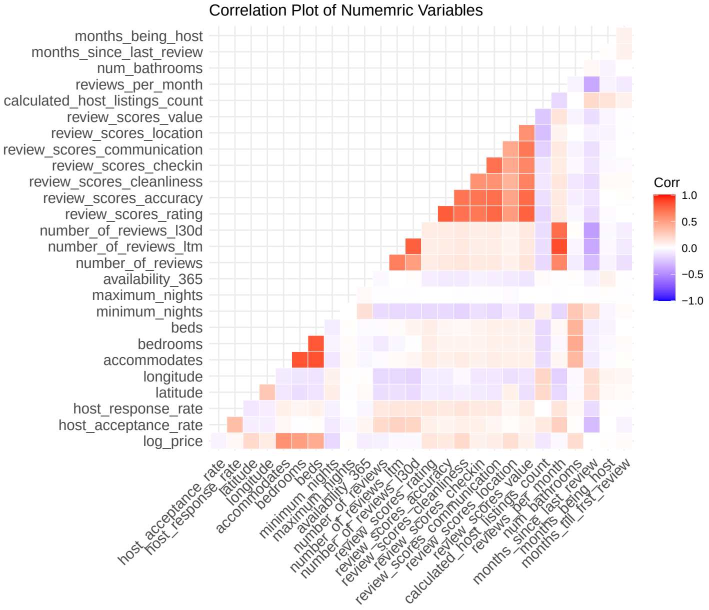
EDA

Most of the categorical variables do not seem to have obvious relations with price/log-price.

- Before we move on to the numeric variables, one categorical variable does seem interesting.

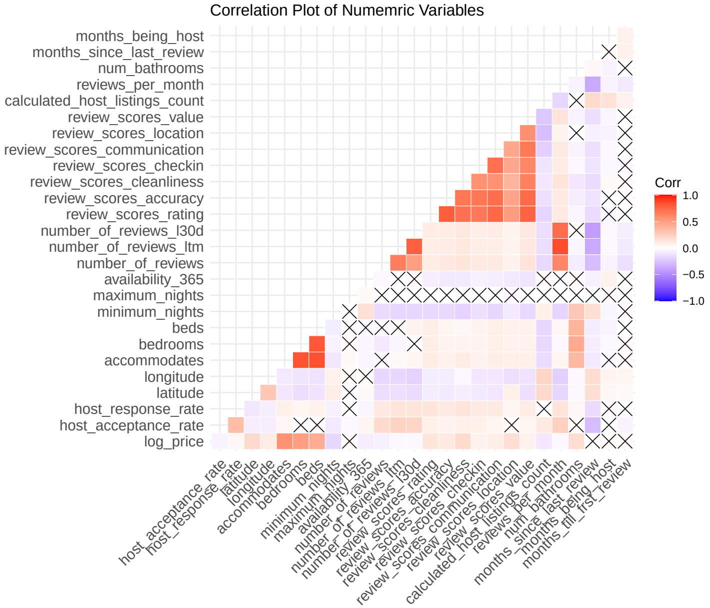


EDA



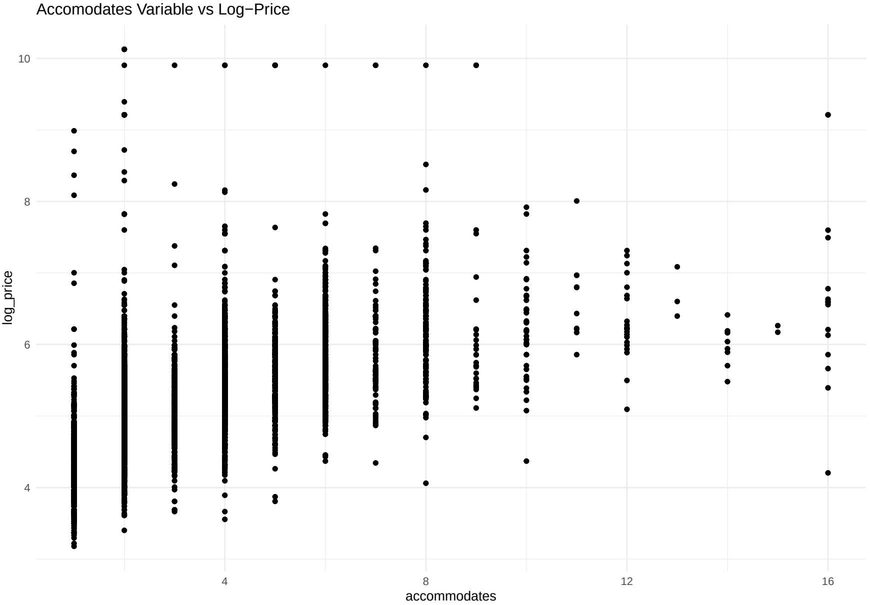
Correlation Plot of Numemric Variables
months_being_host
months_since_last_review
num_bathrooms
reviews_per_month
calculated_host_listings_count
review_scores_value
review_scores_location
review_scores_communication
review_scores_checkin
review_scores_cleanliness
review_scores_accuracy
review_scores_rating
number_of_reviews_l30d
number_of_reviews_ltm
number_of_reviews
availability_365
maximum_nights
minimum_nights
beds
bedrooms
accommodates
longitude
latitude
host_response_rate
host_acceptance_rate
log_price
host_acceptance_rate
host_response_rate
latitude
longitude
accommodates
bedrooms
beds
minimum_nights
maximum_nights
availability_365
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
calculated_host_listings_count
reviews_per_month
num_bathrooms
months_since_last_review
months_being_host
months_till_frst_review
Corr
1.0
0.5
0.0
−0.5
−1.0

EDA

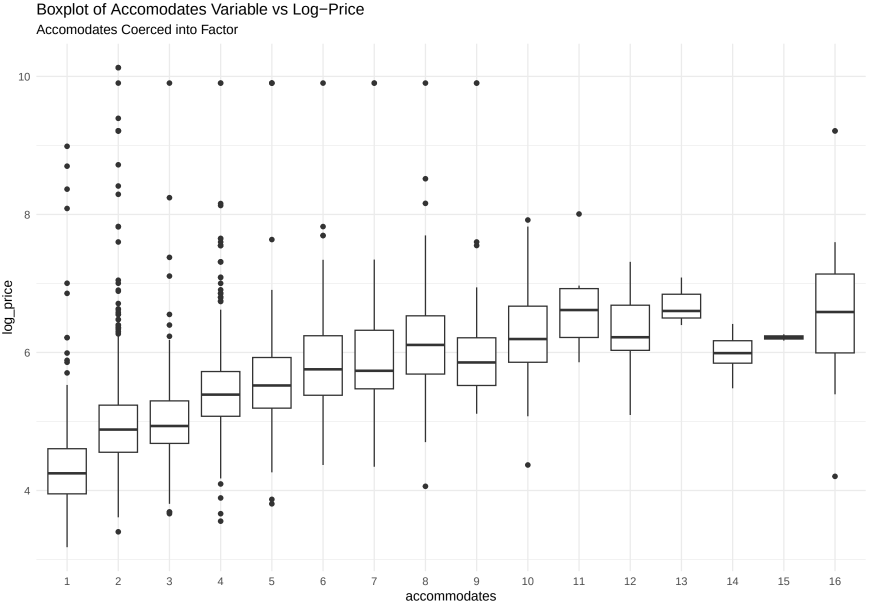


Correlation Plot of Numemric Variables
months_being_host
months_since_last_review
num_bathrooms
reviews_per_month
calculated_host_listings_count
review_scores_value
review_scores_location
review_scores_communication
review_scores_checkin
review_scores_cleanliness
review_scores_accuracy
review_scores_rating
number_of_reviews_l30d
number_of_reviews_ltm
number_of_reviews
availability_365
maximum_nights
minimum_nights
beds
bedrooms
accommodates
longitude
latitude
host_response_rate
host_acceptance_rate
log_price
host_acceptance_rate
host_response_rate
latitude
longitude
accommodates
bedrooms
beds
minimum_nights
maximum_nights
availability_365
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
calculated_host_listings_count
reviews_per_month
num_bathrooms
months_since_last_review
months_being_host
months_till_frst_review
Corr
1.0
0.5
0.0
−0.5
−1.0

EDA

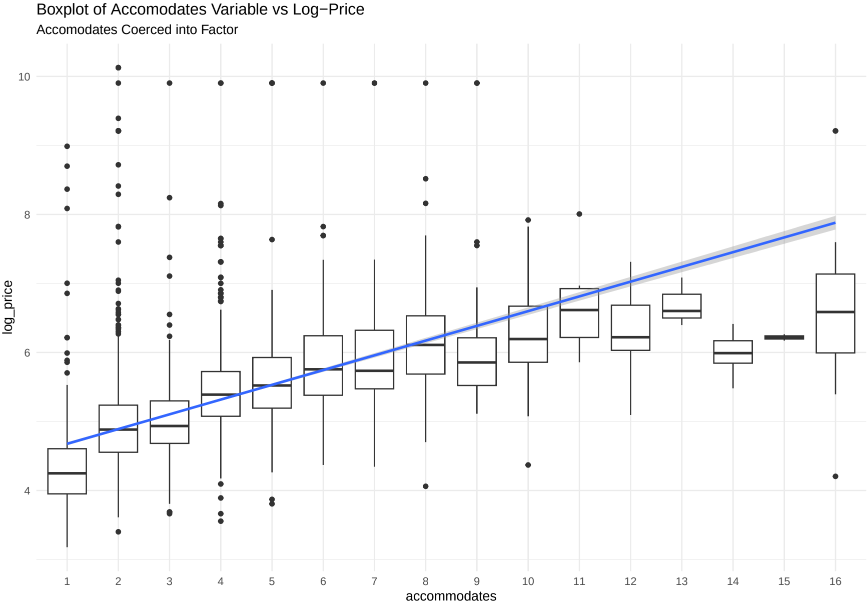

Accomodates Variable vs Log−Price
log_price
10
8
6
4
accommodates
4
8
12
16

EDA



Boxplot of Accomodates Variable vs Log−Price
Accomodates Coerced into Factor
log_price
10
8
6
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
accommodates

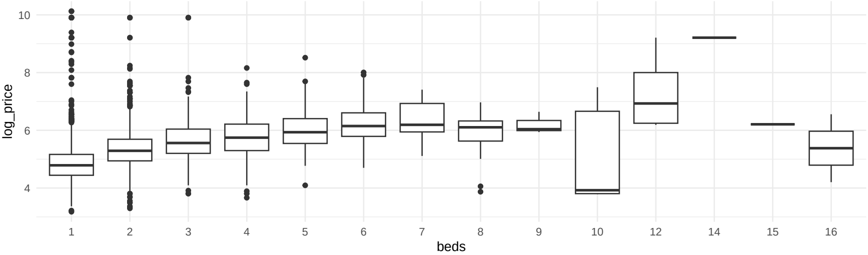
EDA



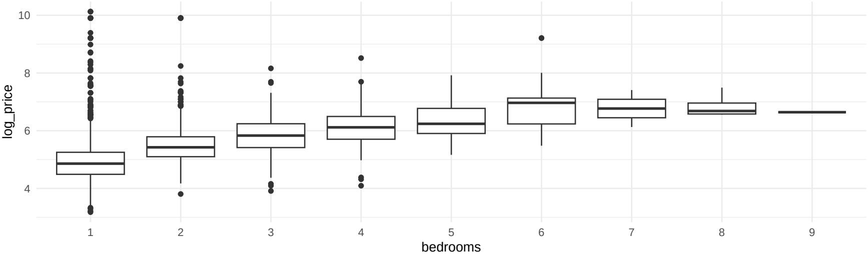
Boxplot of Accomodates Variable vs Log−Price
Accomodates Coerced into Factor
log_price
10
8
6
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
accommodates

EDA

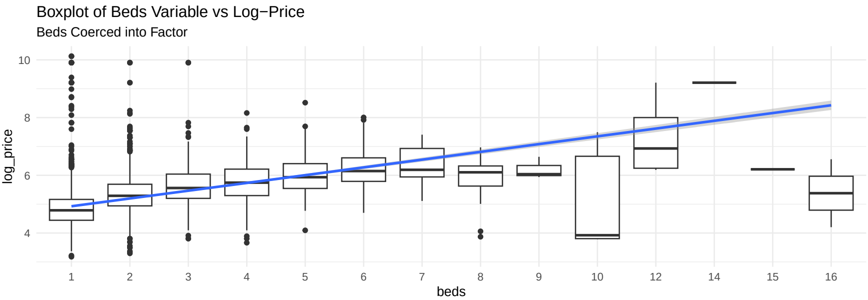
Boxplot of Beds Variable vs Log−Price
Beds Coerced into Factor

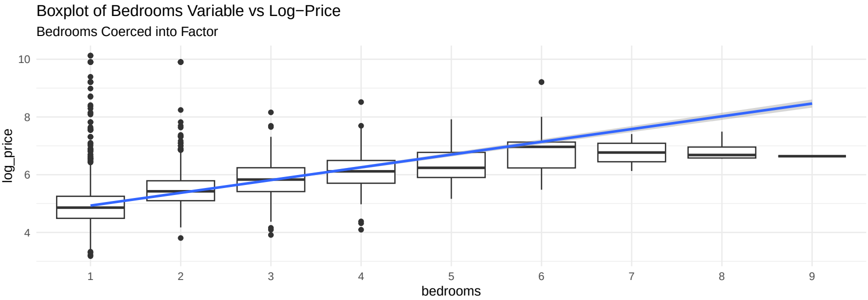


Boxplot of Bedrooms Variable vs Log−Price
Bedrooms Coerced into Factor

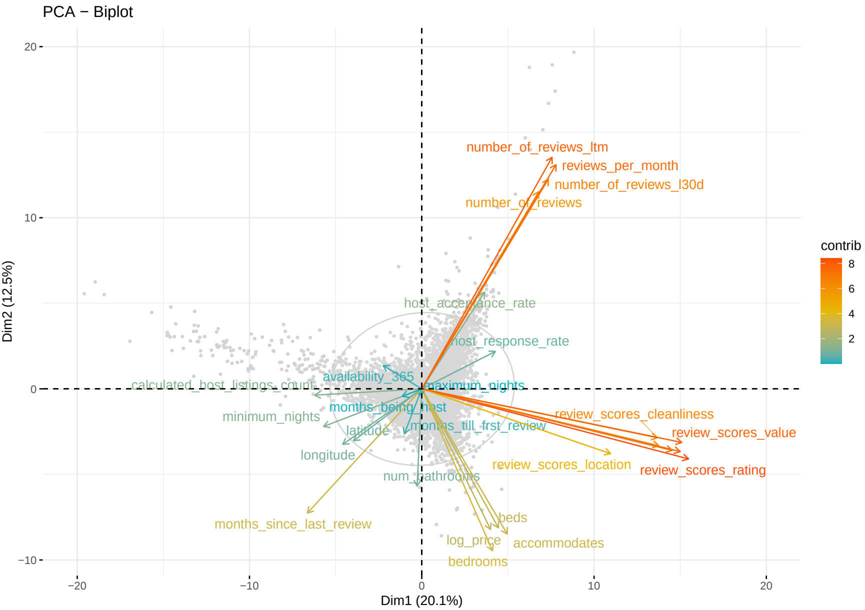


EDA

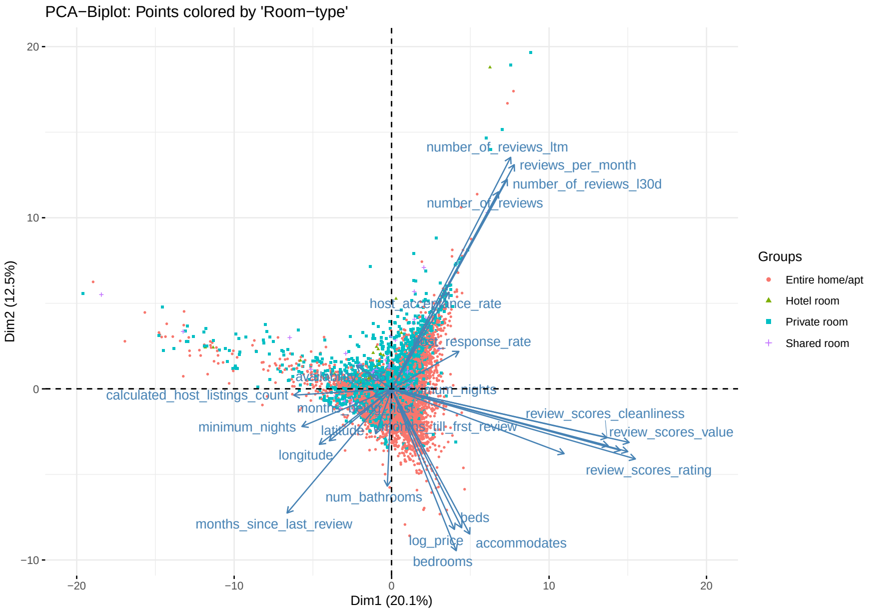

Boxplot of Beds Variable vs Log−Price
Beds Coerced into Factor
log_price
10
8
6
4
1
2
3
4
5
6
7
8
9
10
12
14
15
16
beds


Boxplot of Bedrooms Variable vs Log−Price
Bedrooms Coerced into Factor
log_price
10
8
6
4
1
2
3
4
5
6
7
8
9
bedrooms

EDA


PCA − Biplot
Dim2 (12.5%)
Dim1 (20.1%)
20
10
0
−10
−20
number_of_reviews_ltm
reviews_per_month
number_of_reviews_l30d
number_of_reviews
host_acceptance_rate
host_response_rate
availability_365
calculated_host_listings_count
maximum_nights
months_being_host
minimum_nights
latitude
months_till_frst_review
longitude
num_bathrooms
review_scores_cleanliness
review_scores_value
review_scores_location
review_scores_rating
months_since_last_review
beds
log_price
accommodates
bedrooms
contrib
8
6
4
2

EDA


PCA−Biplot: Points colored by 'Room−type'
Dim2 (12.5%)
Dim1 (20.1%)
20
10
0
−10
−20
−10
0
10
20
number_of_reviews_ltm
reviews_per_month
number_of_reviews_l30d
number_of_reviews
host_acceptance_rate
host_response_rate
calculated_host_listings_count
minimum_nights
latitude
longitude
num_bathrooms
months_since_last_review
review_scores_cleanliness
review_scores_value
review_scores_rating
beds
accommodates
log_price
bedrooms
Groups
Entire home/apt
Hotel room
Private room
Shared room

EDA

So what did we learn?

- We have some pockets of colinearity in the data, which needs to be accounted for during modeling
- We can only identify a handful of features that seem to explain the price of listings
- The data seems to be quite noisy with a non-linear structure

A quick note: Clustering the data seemed to only identify the neighborhoods that exist, which maybe interesting in future work.

Methods

The data was modeled with five different models

- Multiple Regression
- Lasso Regression
- Random Forest
- XGBoost
- Neural Net

Methods

Tuning Models: For each model tuning parameters are selected via 10-fold cross-validation grid search.

Methods

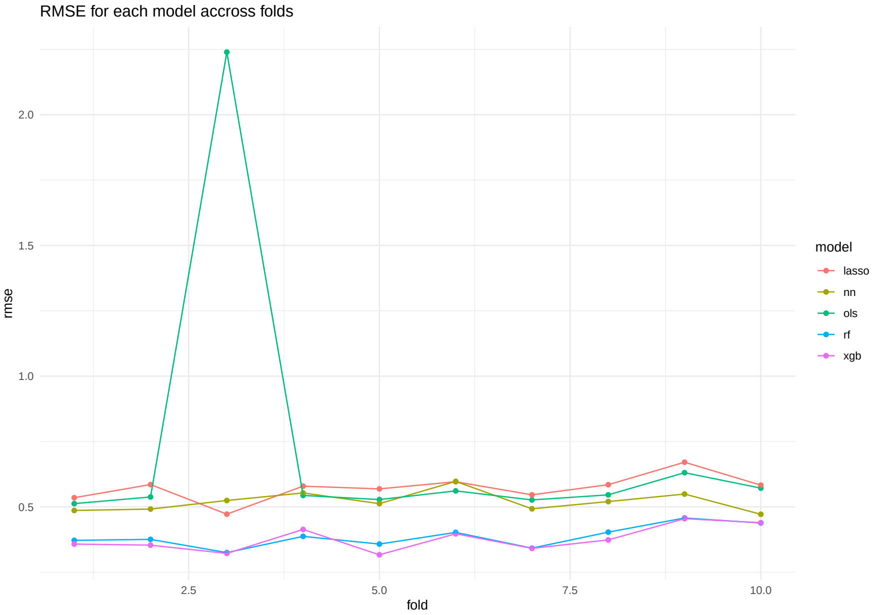
Tuning Parameters:

- Lasso: 50 different possible values of lambda penalty
- Random Forest: 80 x 2 grid of possible values
  - mtry
  - min_n
- XGBoost: 40 x 6 grid of possible values
  - tree depth
  - min n
  - loss reduction
  - sample size
  - mtry
  - learning rate
- Neural Net: Two neural nets were tests
  - One hidden layer with 60 units
  - Three hidden layers with units 70, 40, 10 respectivly

Methods

Cross-Validation between models: Once models are tuned 10-fold cross validation is performed between models (at the same time) i.e. sample training and testing splits.

Methods


RMSE for each model accross folds
rmse
2.0
1.5
1.0
0.5
2.5
5.0
7.5
10.0
fold
model
lasso
nn
ols
rf
xgb

Methods

Random Forest and XGBoost perform the best and comparably

Random Forest RMSE	XGBoost RMSE	Neural Net RMSE	Lasso RMSE	OLS Regression RMSE
0.3860604	0.3768912	0.5198146	0.5720703	0.7196674

The model was then validated on a hold-out set untouched during training and model selection:

RMSE
0.37

Results

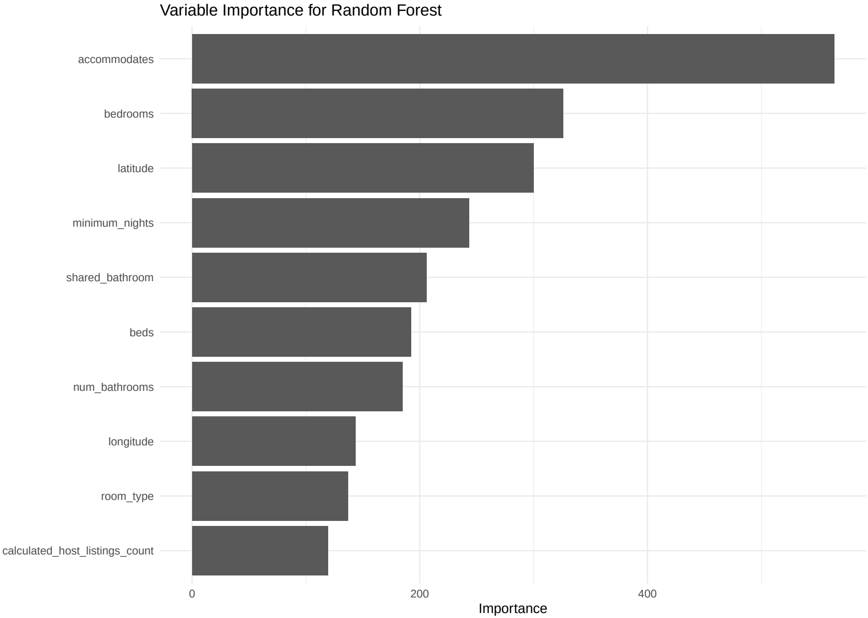
Its important to point out that the signal to noise ratio is somewhat low is this data.

- The information in the features don't explain the variation in price to a high degree, which makes sense.

Amount of variation in price explained by the regression model

R-Squared
0.7496

Results

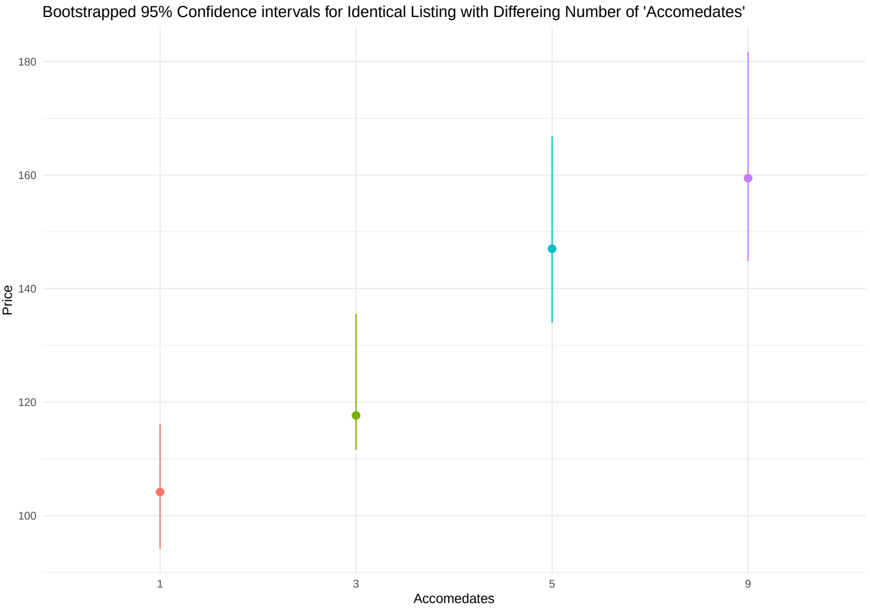


Variable Importance for Random Forest
accommodates
bedrooms
latitude
minimum_nights
shared_bathroom
beds
num_bathrooms
longitude
room_type
calculated_host_listings_count
0
200
400
Importance

Results

Measured by total decrease in node impurities from splitting on the variable, averaged over all trees. For regression, it is measured by residual sum of squares.

Results


Bootstrapped 95% Confidence intervals for Identical Listing with Differeing Number of 'Accomedates'
Price
180
160
140
120
100
1
3
5
9
Accomedates

Conclusion

We were able to model the price of a given listing using information provided about the listing on Airbnb's website. The number of people a listing can host seems to be the most important feature of the property in predicting the price. However, there still is a lot of variability on the price of listing and thus the model itself, thus it should only be used to get an "idea of price".

Future Work

- Perform a text analysis on “neighborhood overview” and “amenities” and use the results as predictive features for the price.
- Perform a more intelligent imputation method “Multiple-Imputation” to not rely on one “estimate” for imputed values.
- Extend the scope of the research beyond San Francisco.
- Further tuning of models.

Thank You!