Problem Statement Definition

Target variable: log_price
Predictors:

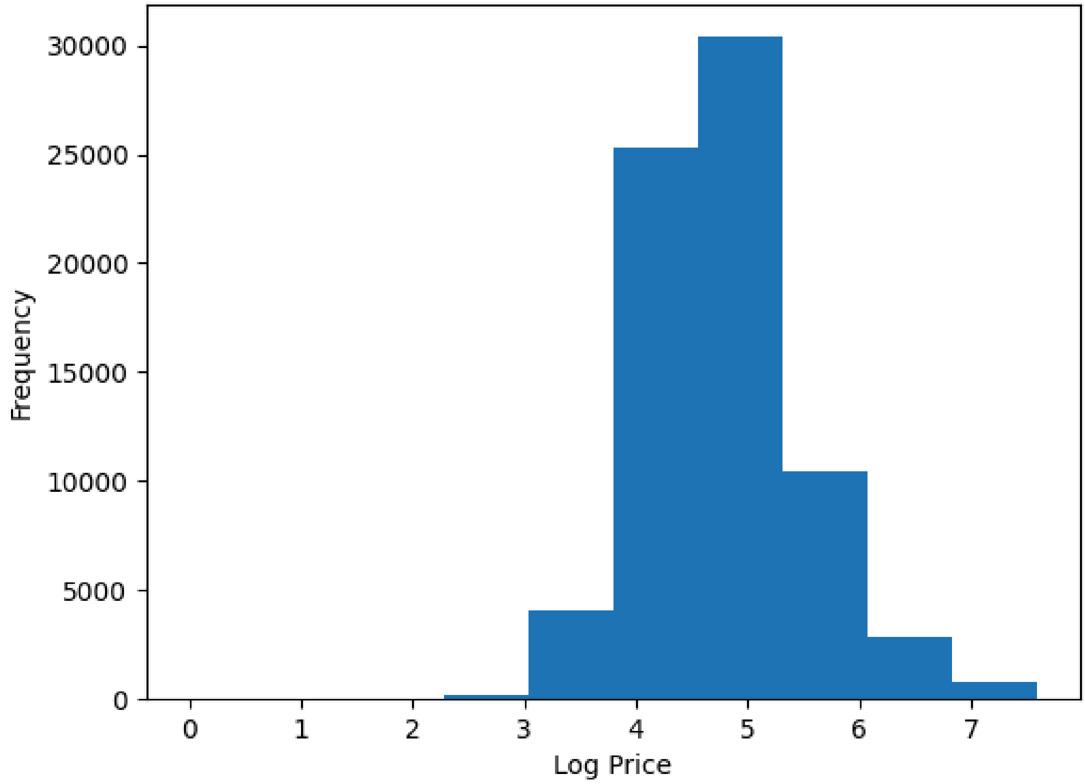
- id
- log_price
- property_type
- room_type
- amenities
- accommodates
- bathrooms
- bed_type
- cancellation_policy
- cleaning_fee
- city
- description
- first_review
- host_has_profile_pic
- host_identity_verified
- host_response_rate
- host_since
- instant_bookable
- last_review
- latitude
- longitude
- name
- neighbourhood
- number_of_reviews
- review_scores_rating
- thumbnail_url
- zipcode
- bedrooms
- beds

Algorithm

The target variable is continuous, so a linear regression algorithm will be used.

Target Variable Distribution

Using matplotlib, we can generate a histogram of the target variable (log_price) distribution:



The distribution appears close to a bell curve, so no further modifications need to be made to this variable.

Exploratory Data Analysis

There are 29 columns in this dataset:

Quantitative variables

- id
- log_price
- latitude
- longitude
- number_of_reviews
- review_scores_rating
- host_response_rate
- host_since
- last_review
- first_review

Qualitative variables

- property_type
- room_type

- amenities
- bed_type
- zipcode
- name
- neighbourhood
- thumbnail_url
- description
- city
- cancellation_policy

Categorical variables

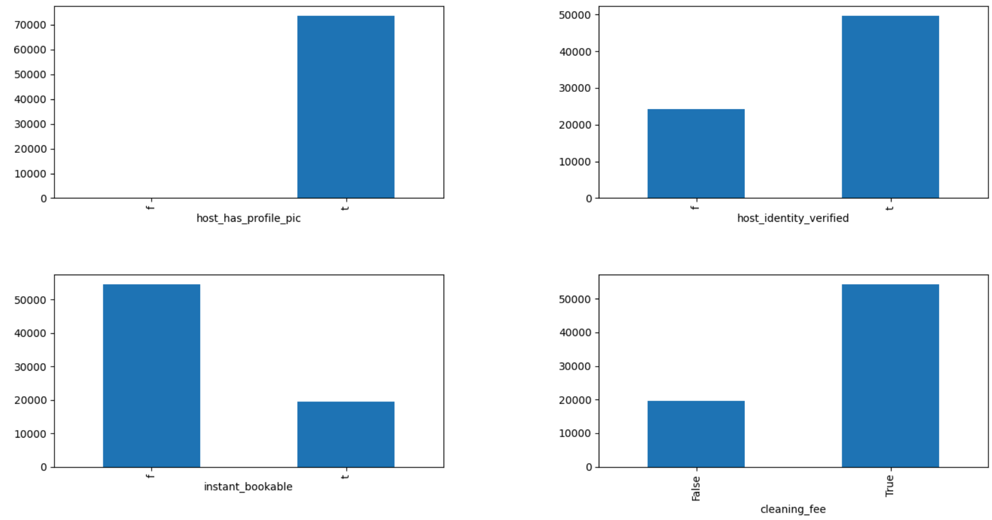
- bathrooms
- bedrooms
- beds
- accommodates
- host_has_profile_pic
- host_identity_verified
- instant_bookable
- cleaning_fee

10 variables are quantitative, 11 are qualitative, and 8 are categorical. Dates have been categorised as quantitative variables because they have many unique values and can easily be expressed as numerical values. Note that zipcode is listed as a qualitative variable because it does not serve the purpose of a numerical value, but is more similar to “city” or “neighbourhood”.

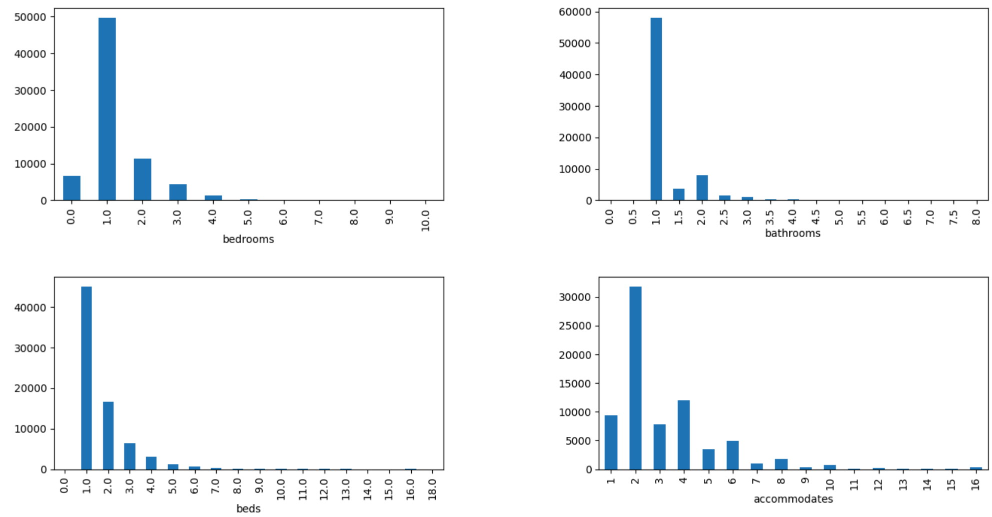
Unwanted columns

- id is metadata that is not related to any of the data in this set, so it should be removed.
- The qualitative variables property_type, room_type, amenities, bed_type, zipcode, name, neighbourhood, thumbnail_url, description, and city should be dropped as they cannot be easily converted into a useful numerical value. Cancellation_policy could be converted to a numerical value depending on the strictness of cancellation.

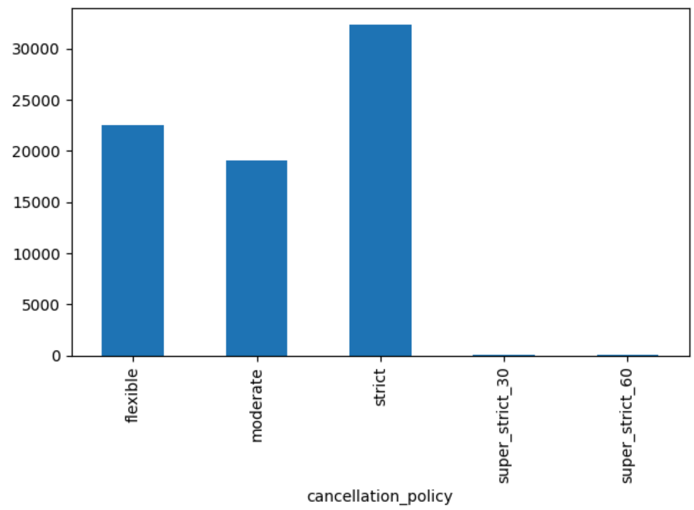
Visual Exploratory Analysis (Categorical)



The above figure shows the boolean variables. Most variables have a fairly even distribution of columns for each value. However, host_has_profile_pic has almost no columns that are false, so this variable needs to be dropped.

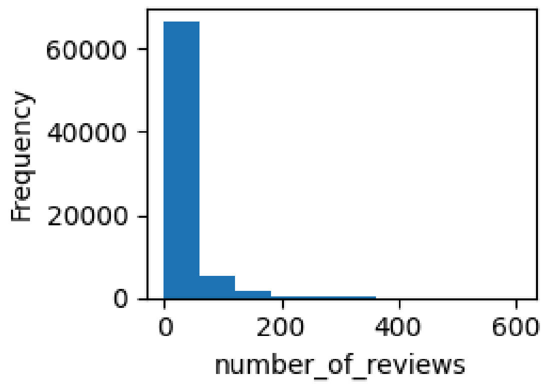


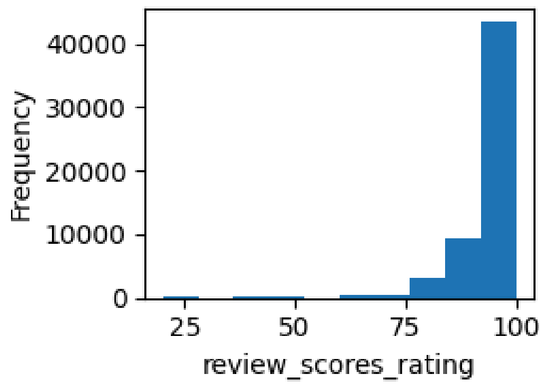
The above figure shows the numeric categorical variables. All of these generally follow a bell curve, but have a strong skew. Most of these variables appear to be useable, however the bathrooms variable has too high of a skew to be useful.

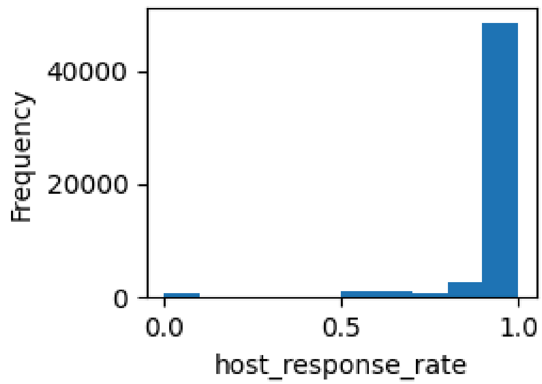


The above figure shows the values that correspond to the cancellation_policy variable. The values are evenly distributed among flexible, moderate and strict, however almost no rows have the values of “super_strict_30” and “super_strict_60”. This variable can still be useful by assigning “flexible” to 1, “moderate” to 2, and “strict”, “super_strict_30” and “super_strict_60” to 3.

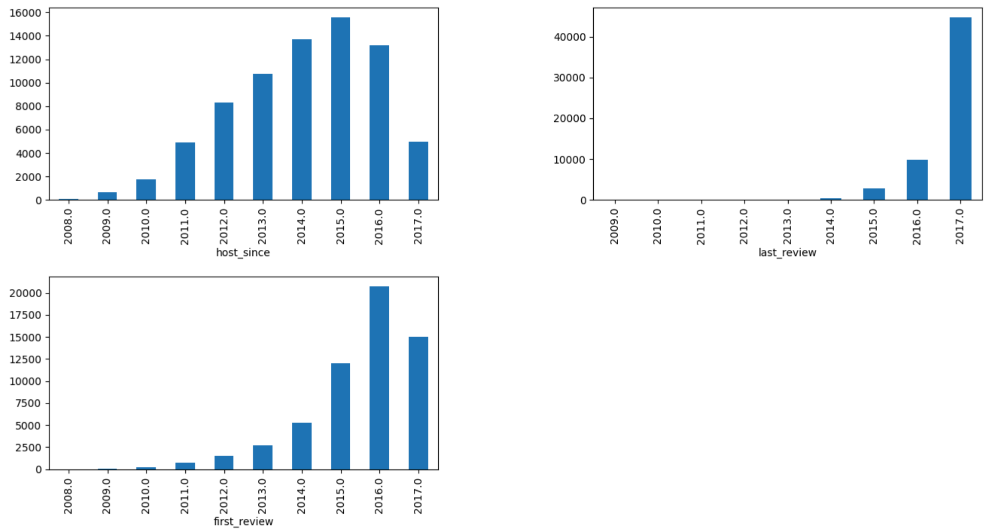
Visual Exploratory Analysis (Continous)



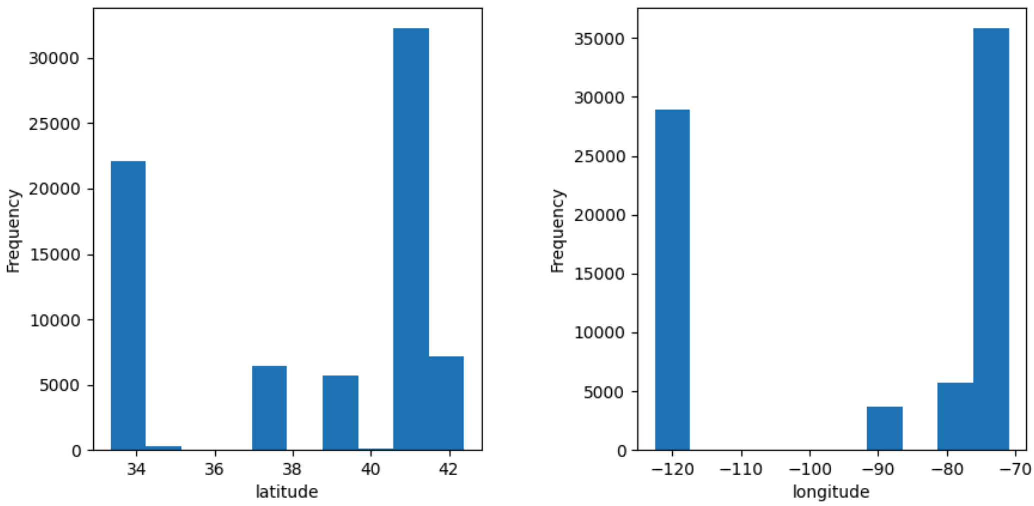




The above figure shows the distribution of variables number_of_reviews, review_scores_rating, and host_response_rate. Number_of_reviews and review_scores_rating are heavily skewed but still follow a bell curve. Host_response_rate is not only heavily skewed, but also does not appear to follow a bell curve.



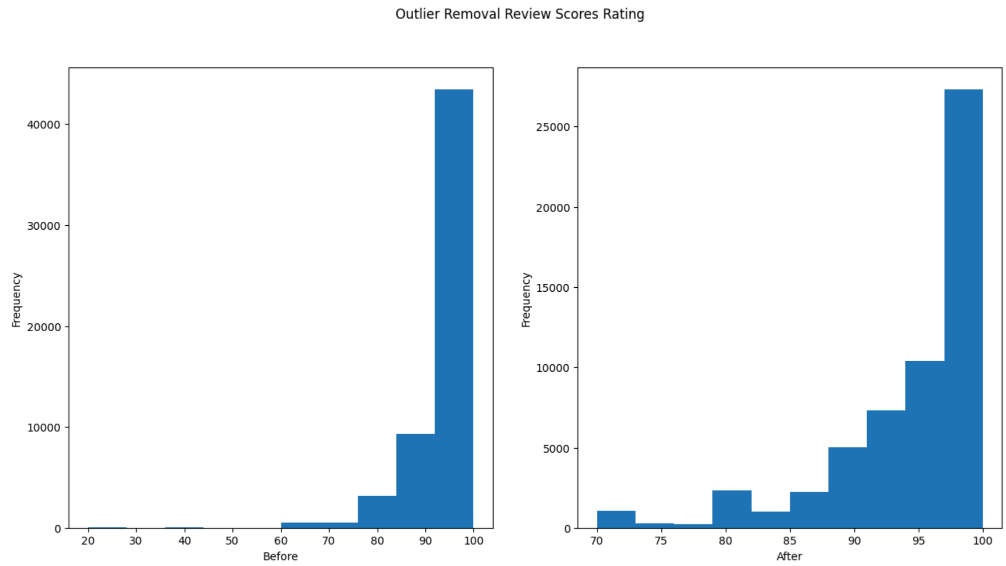
The above figure shows the distribution of the variables host_since, first_review, and last_review. The last_review variable is heavily skewed, but the host_since and first_review variables show nice bell curves. All of these variables will be kept for now, as the large skew on last_review isn't enough to warrant an immediate removal.

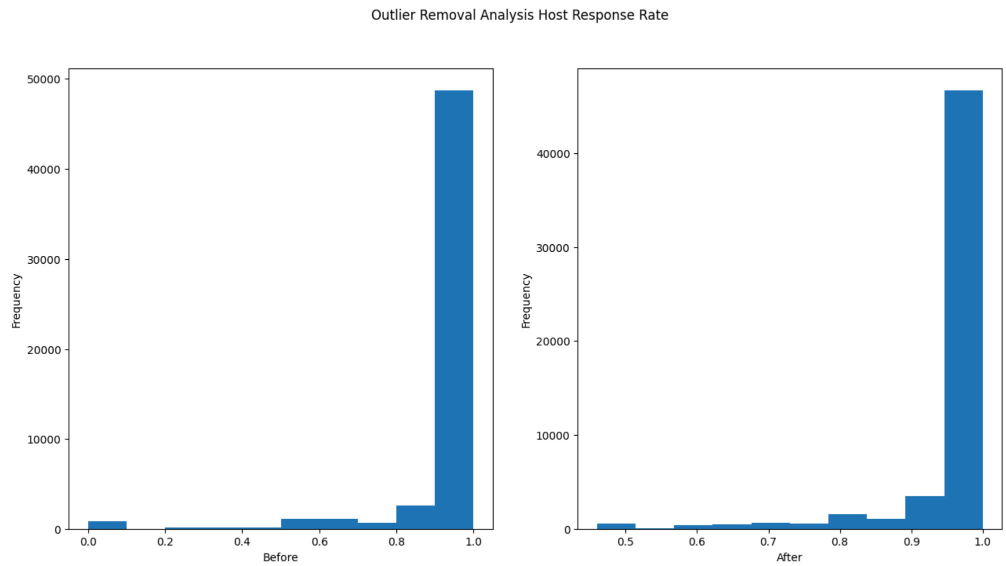


The above figure shows the distribution of latitude and longitude. These do not resemble bell curves at all, so these variables must be dropped.

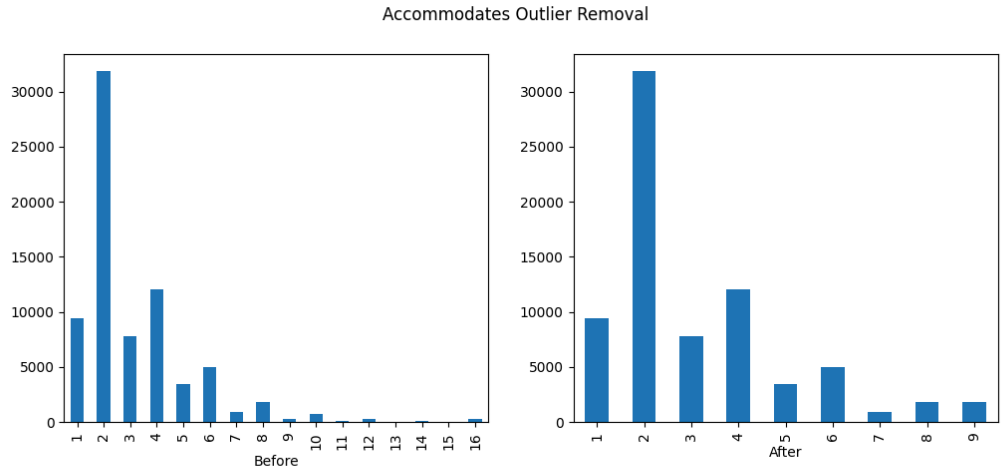
Outlier Analysis

From figure, it can be seen that the accommodates attribute has some outliers past 9, especially the relatively large rise at accommodates=16. The graphs of review_scores_rating and host_response_rate also have outliers that can be seen at the tails. I will use the winsorising method to handle outliers for all graphs, setting each outlier to a value that is within 3 standard deviations of the mean.

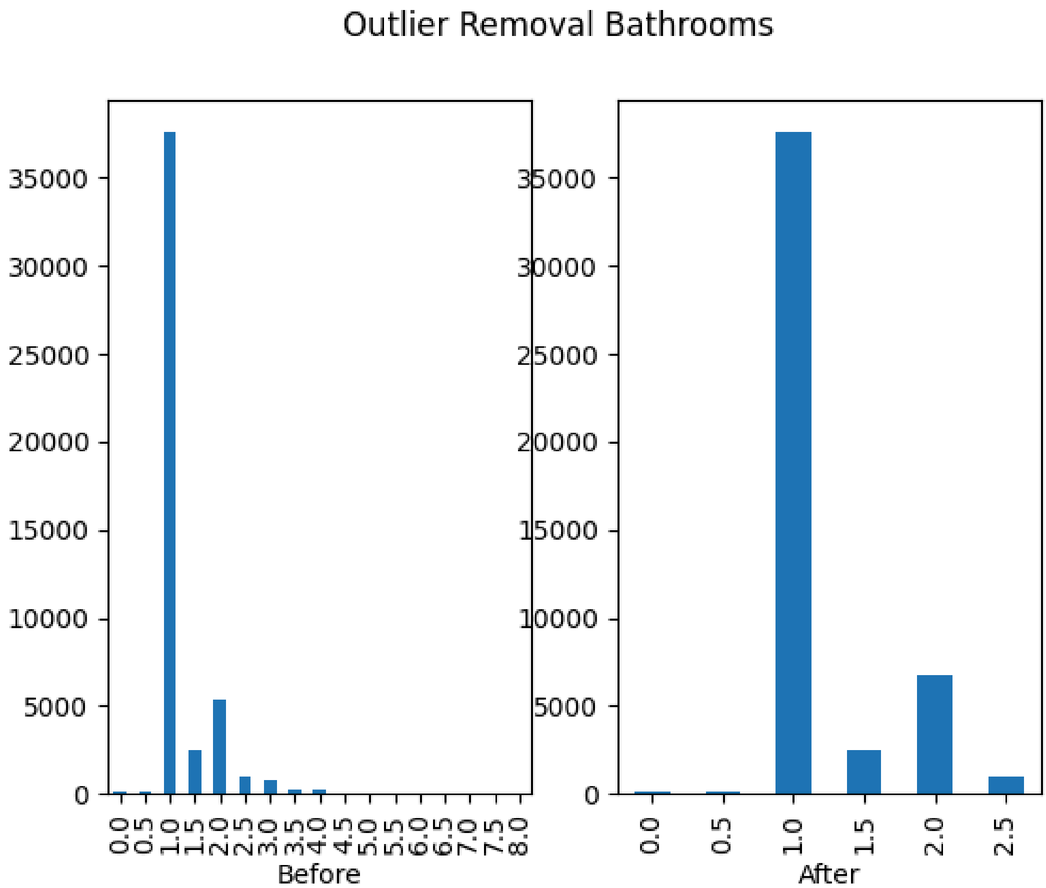


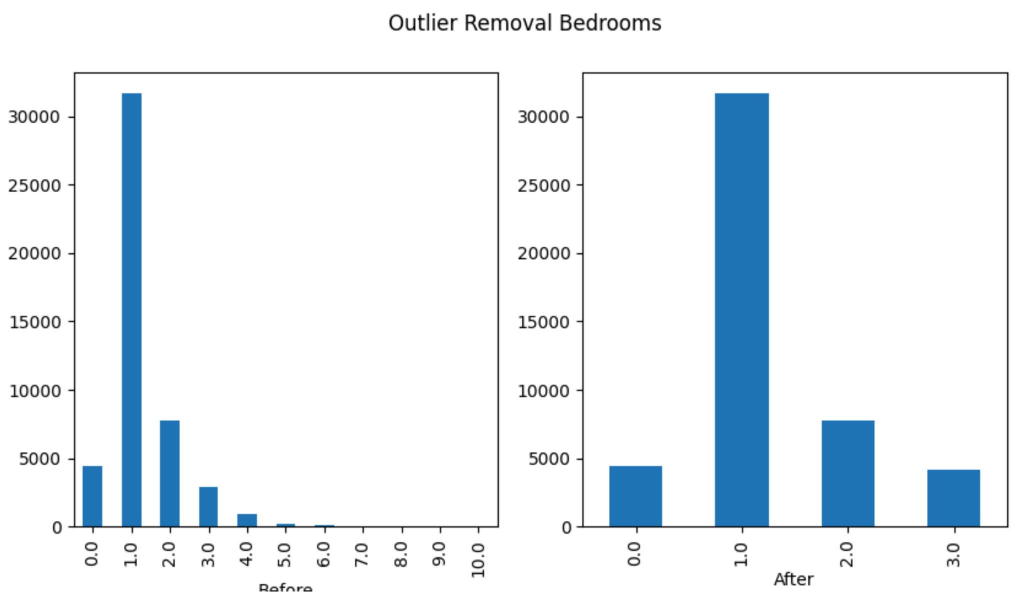


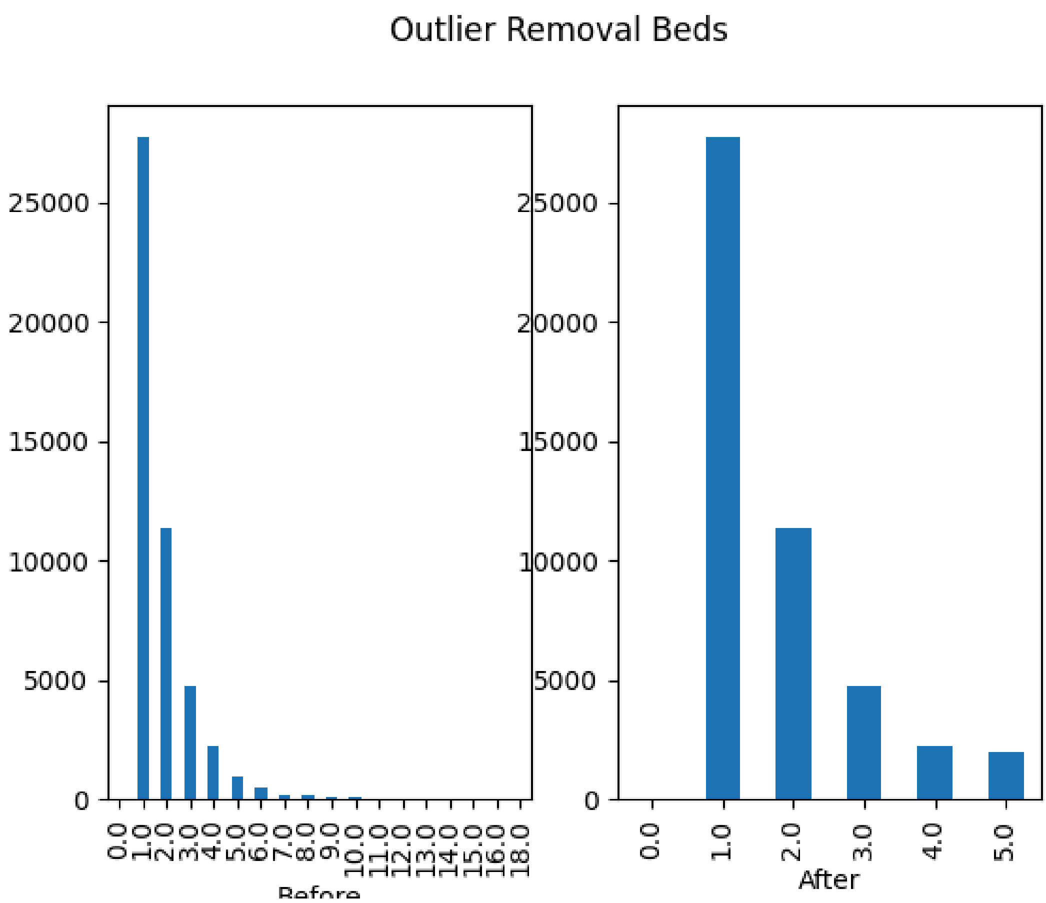
Outliers are below 0.45



Outliers are above 9







Missing Value Analysis

Column	Percentage that are null
accommodates	0.000000%
bathrooms	0.269865%

cancellation_policy	0.000000%
cleaning_fee	0.000000%
first_review	21.405729%
host_identity_verified	0.253674%
host_response_rate	24.691341%
host_since	0.253674%
instant_bookable	0.000000%
last_review	21.355804%
number_of_reviews	0.000000%
review_scores_rating	22.563452%
bedrooms	0.122789%
beds	0.176762%

The above table shows the percentage of values that are null for each attribute. Most have a small amount of null values, except for review_scores_rating, host_response_rate, first_review and last_review. first_review or and last_review being null suggests that these houses received no reviews. host_response_rate being null suggests that the host was never given anything to respond to. review_scores_rating being null suggests that the house received no reviews. If a home received no reviews, then its values for first_review and last_review would both be null. This can explain why the percentage of null values for first_review and last_review are very similar. Thus, dropping all null values in last_review would also drop most of the null values in first_review. This is also the same for reviews_scores_rating.

The null values of continuous attributes such as host_response_rate can be imputed with the median value. The null values of categorical variables such as bathrooms can be imputed with the mode.

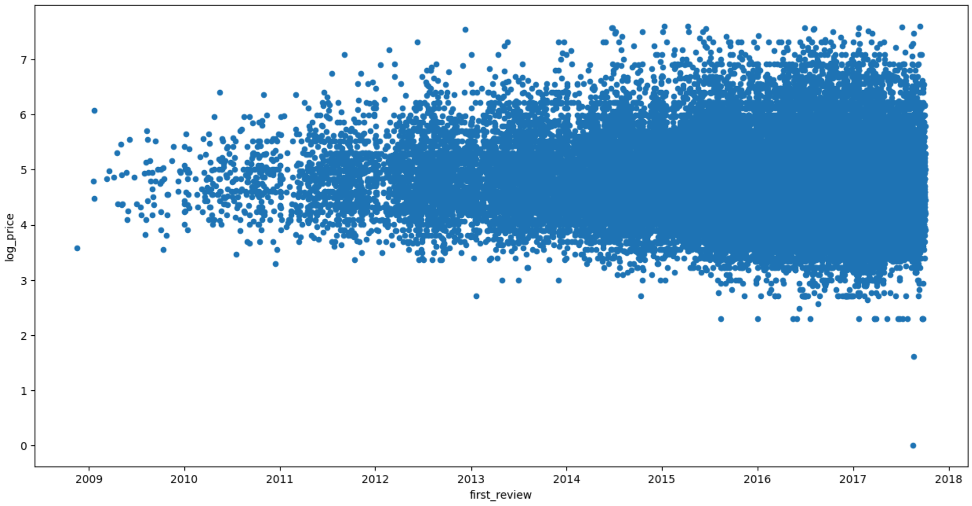
last_review first_review, and review_scores_rating carry a special meaning when they are null. It expresses that this house received no reviews, indicating a lack of interest in the house, which could affect the pricing of the house. This meaning may be lost if they are imputed with a value. Therefore, it is better to drop them entirely from the dataset rather than try to impute them.

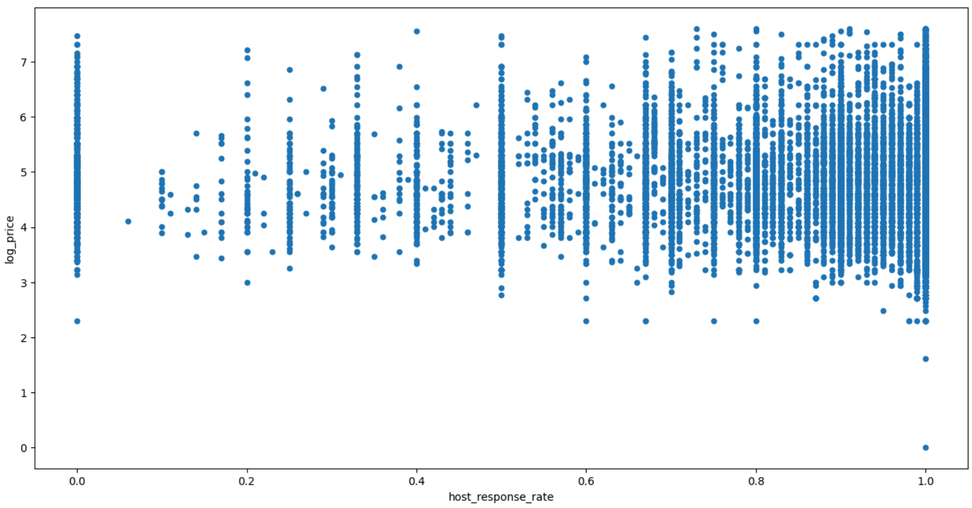
This brings the number of entries from 74111 to 48002, which is a 35% cut.

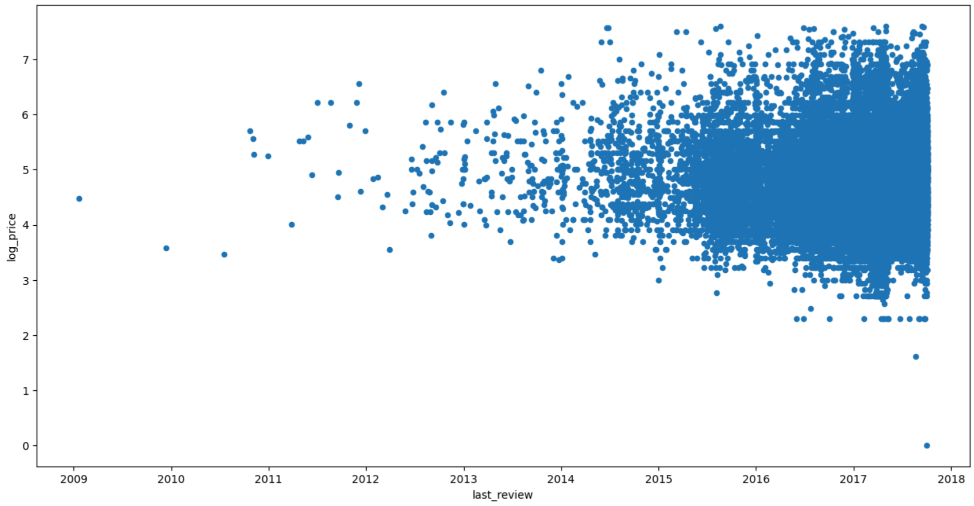
Continuous Correlation Analysis

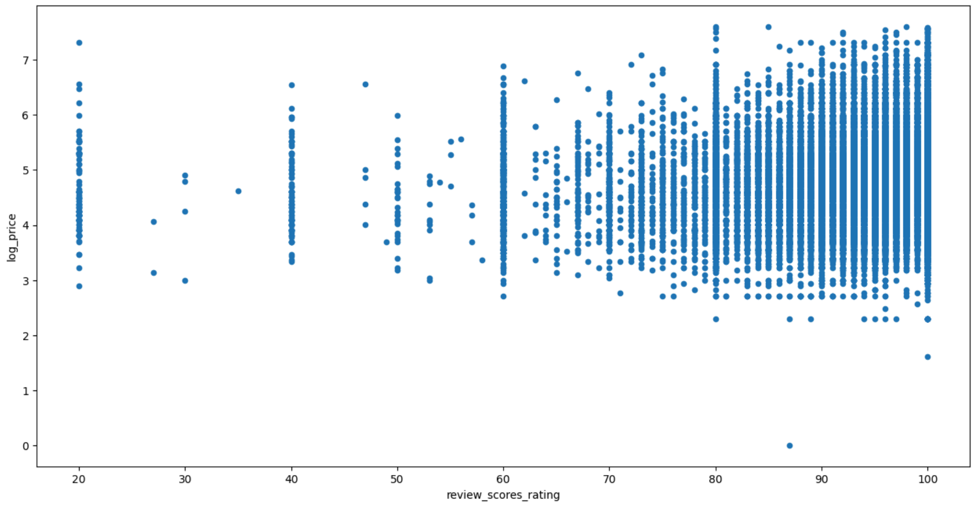
host_response_rate	-0.006777
first_review	-0.090415
last_review	-0.090108
review_scores_rating	0.091219
log_price	1.000000

None of these variables are correlated with the log_price. This means that the rows that were previously dropped due to a null value in one of these columns can be readded.


log_price
first_review
0
1
2
3
4
5
6
7
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

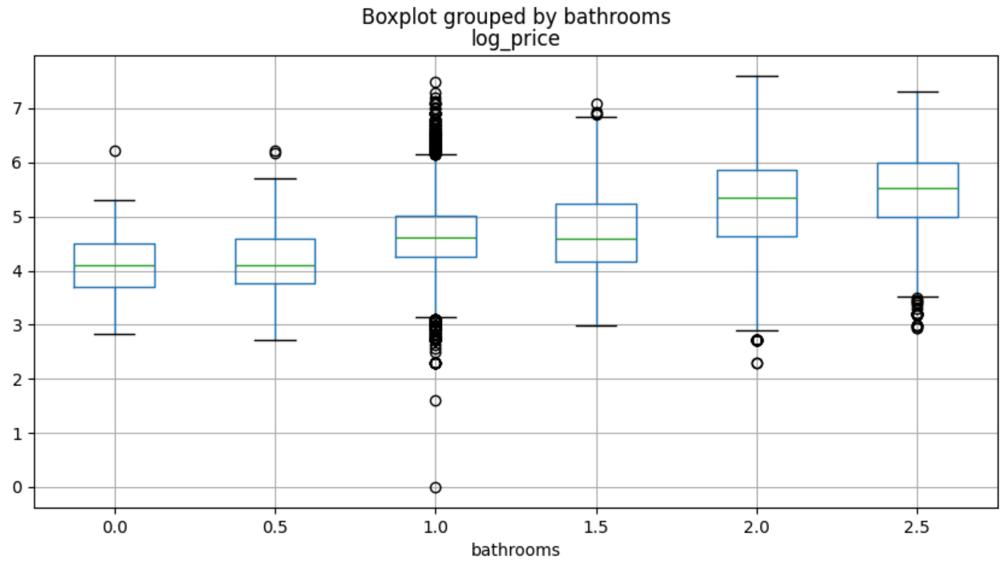

log_price
host_response_rate
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8
1.0

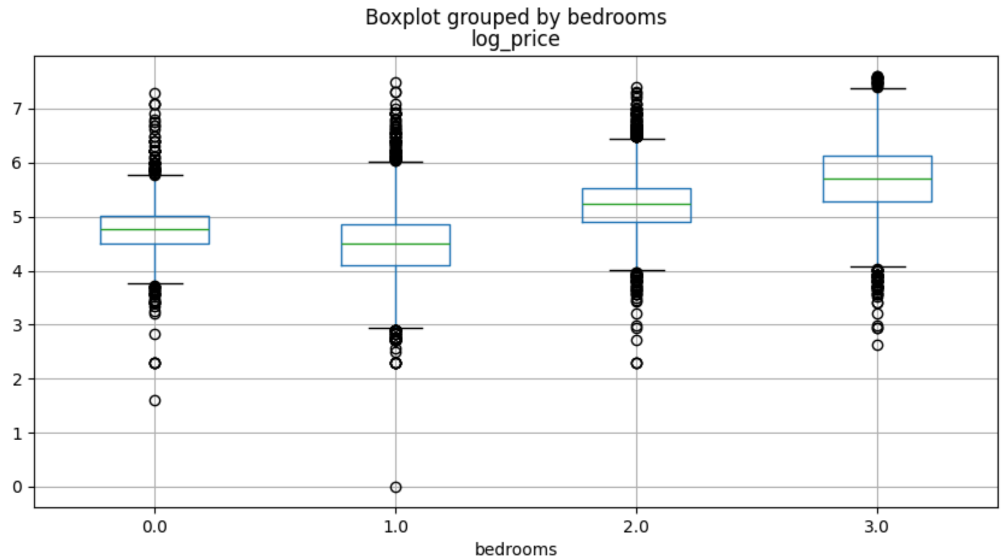


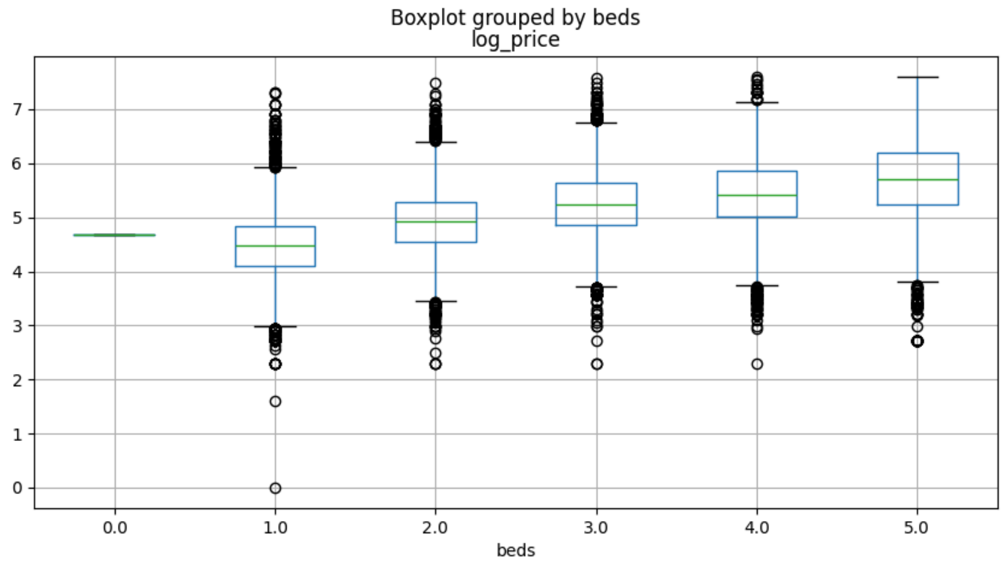


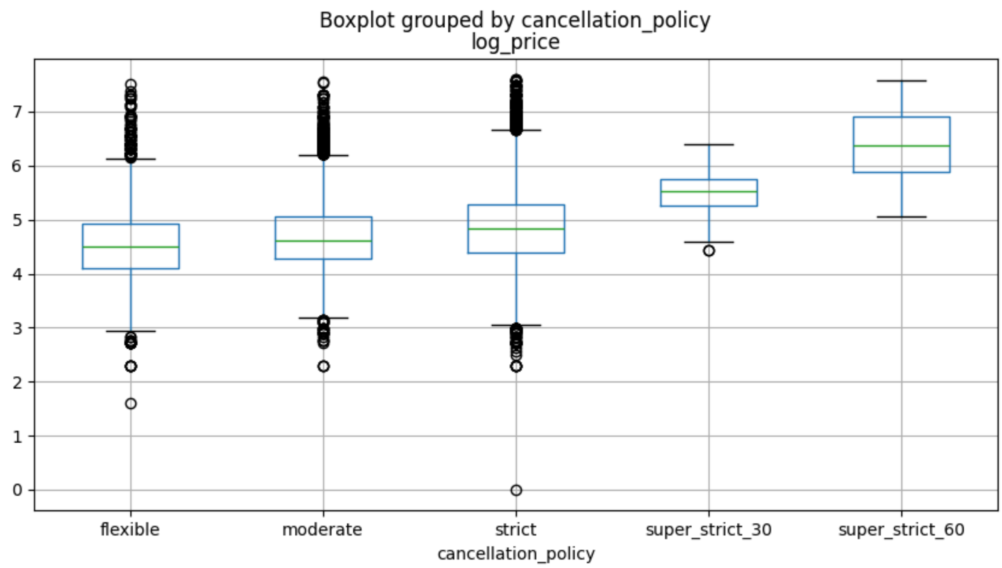
By looking at the above scatter plots, it is also evident that none of these variables are related to the target variable.

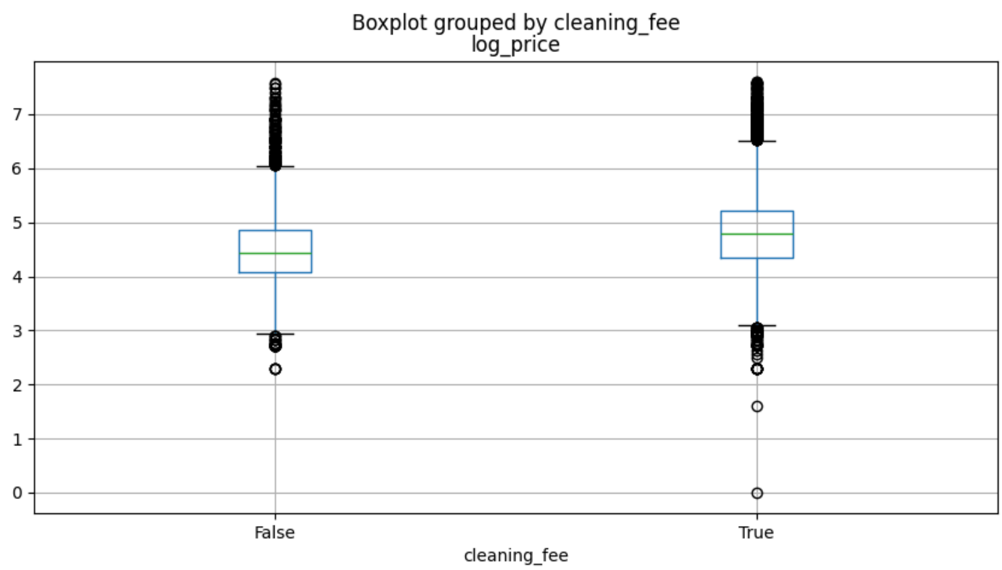
Categorical Correlation Analysis

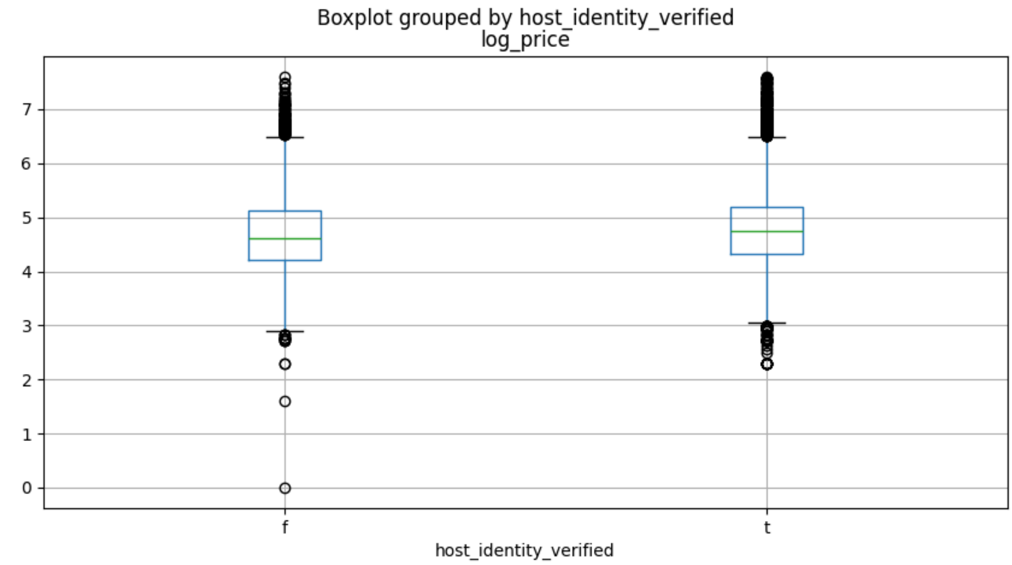


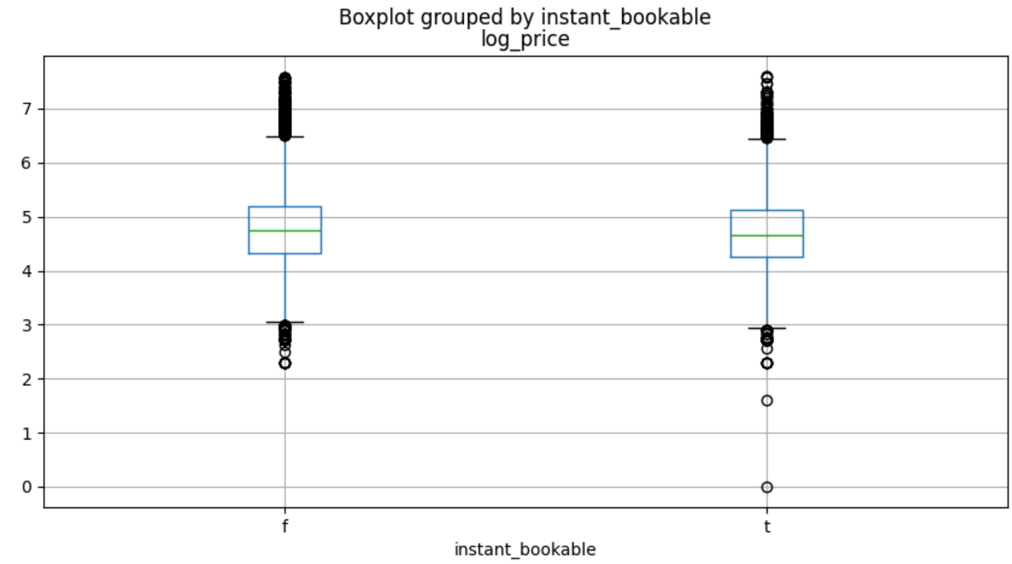


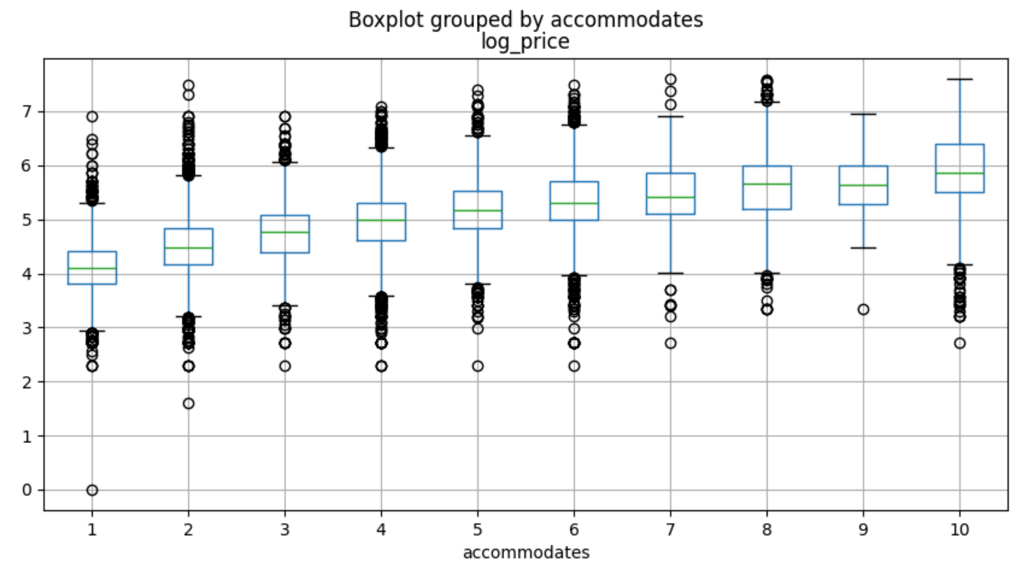

Boxplot grouped by beds
log_price
0
1
2
3
4
5
6
7
0.0
1.0
2.0
3.0
4.0
5.0
beds


Boxplot grouped by cancellation_policy
log_price
0
1
2
3
4
5
6
7
flexible
moderate
strict
super_strict_30
super_strict_60
cancellation_policy


Boxplot grouped by cleaning_fee
log_price
0
1
2
3
4
5
6
7
False
True
cleaning_fee


Boxplot grouped by host_identity_verified
log_price
0
1
2
3
4
5
6
7
f
t
host_identity_verified


Boxplot grouped by instant_bookable
log_price
0
1
2
3
4
5
6
7
f
t
instant_bookable


Boxplot grouped by accommodates
log_price
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
accommodates

All of these appear to show correlation, except for instant_bookable, host_identity_verified, and cleaning_fee. Although the strict_30 and strict_60 columns seem to differ from the rest, there is not enough data to work with these variables. Therefore, cancellation_policy seems to not have much correlation as well from the graph.

However, the anova test concludes that all of these are correlated with the target variable:
beds is correlated with log_price | P-Value: 0.0
bedrooms is correlated with log_price | P-Value: 0.0
bathrooms is correlated with log_price | P-Value: 0.0
accommodates is correlated with log_price | P-Value: 0.0
cancellation_policy is correlated with log_price | P-Value: 0.0
host_identity_verified is correlated with log_price | P-Value: 4.954484418718587e-11
instant_bookable is correlated with log_price | P-Value: 1.7637405373462655e-33
cleaning_fee is correlated with log_price | P-Value: 1.6697452663940516e-202

Feature Selection

The following features will be selected for machine learning

- beds
- bedrooms
- bathrooms
- accommodates
- cancellation_policy
- host_identity_verified
- instant_bookable
- cleaning_fee

Testing Multiple Regression Models

======== Testing linear regression model ========
Mean Accuracy on test data: 90.61709533449475
Median Accuracy on test data: 92.46944647403657

Accuracy values for 10-fold Cross Validation:
[90.63813916 90.64632411 90.59541243 90.39705603 90.45606519 90.60241999
90.64070368 90.56062081 90.67834837 90.44078487]

Final Average Accuracy of the model: 90.57

======== Testing Tree Regressor Model ========

Mean Accuracy on test data: 91.13075475987343

Median Accuracy on test data: 92.9522012732888

Accuracy values for 10-fold Cross Validation:
[91.16676279 91.1608484 91.07908259 90.9571824 91.03731476 91.16788612
91.12067008 91.10136785 91.28665554 90.9233445]

Final Average Accuracy of the model: 91.1

======== Random Forest Regressor ========

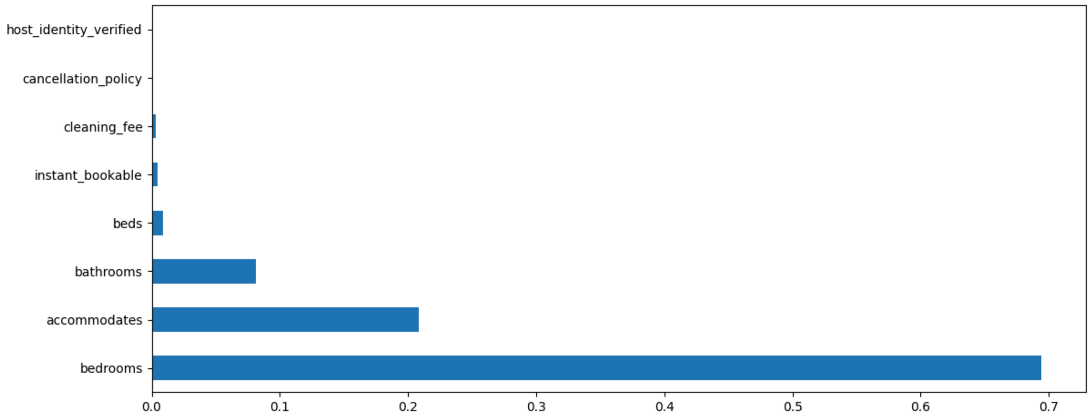
Mean Accuracy on test data: 91.09311096000181
Median Accuracy on test data: 93.04116555206596

Accuracy values for 10-fold Cross Validation:
[91.09737491 91.13384448 91.02980846 90.93610026 90.99830351 91.10403085
91.06871916 91.04921789 91.26410274 90.85259174]

Final Average Accuracy of the model: 91.05

The tree regressor model has the best accuracy with 91.1%.

Model Deployment



Graph of feature importance

The features that carry most of the importance are bathrooms, accommodates, and bedrooms, so the model will be cut down to use only these three features.

Output using three most important features:

Mean Accuracy on test data: 91.09305253401824
Median Accuracy on test data: 92.9251631580036

Accuracy values for 10-fold Cross Validation:
[91.12580588 91.14140728 91.04614247 90.93653806 91.01043136 91.11962987
91.11257291 91.07709595 91.27226457 90.88003075]

Final Average Accuracy of the model: 91.07

This is only 0.03% worse than with all of the unimportant predictors.