Airbnb Price Prediction

Team 1: Veronika Junková, Filip Rott, Lucie Pinterová, Daniel Borner

Course: Data X

1 Data Understanding

For predicting Airbnb prices in Prague, we uses the listings dataset.

Listings.csv.gz

This dataset contains records of individual Airbnb listings in Prague. It includes the price, ratings, maximum occupancy, and many other accommodation-related details.

The dataset contains 8366 rows and 75 columns. It encompasses text, numerical, categorical, boolean, date, ID, json and null columns. Among these, 26 columns (34,6%) contain some null values, while 3 columns (4 %) solely consist of null values. There are in total 7,82 % missing cells in Listings dataset.

Text Columns: listing_url, name, neighborhood_overview, picture_url, host_url, host_name, host_location, host_about, host_thumbnail_url, host_picture_url, host_neighbourhood, neighbourhood, property_type, bathrooms_text, description

Numeric Columns: host_response_rate, host_acceptance_rate, host_listings_count, host_total_listings_count, latitude, longitude, accommodates, beds, price, minimum_nights, maximum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm, availability_30, availability_60, availability_90, 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, calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, reviews_per_month, bathrooms, bedrooms

Categorical Columns: source, host_response_time, host_verifications, neighbourhood_cleansed, room_type

Null colums: neighbourhood_group_cleansed, calendar_updated, license

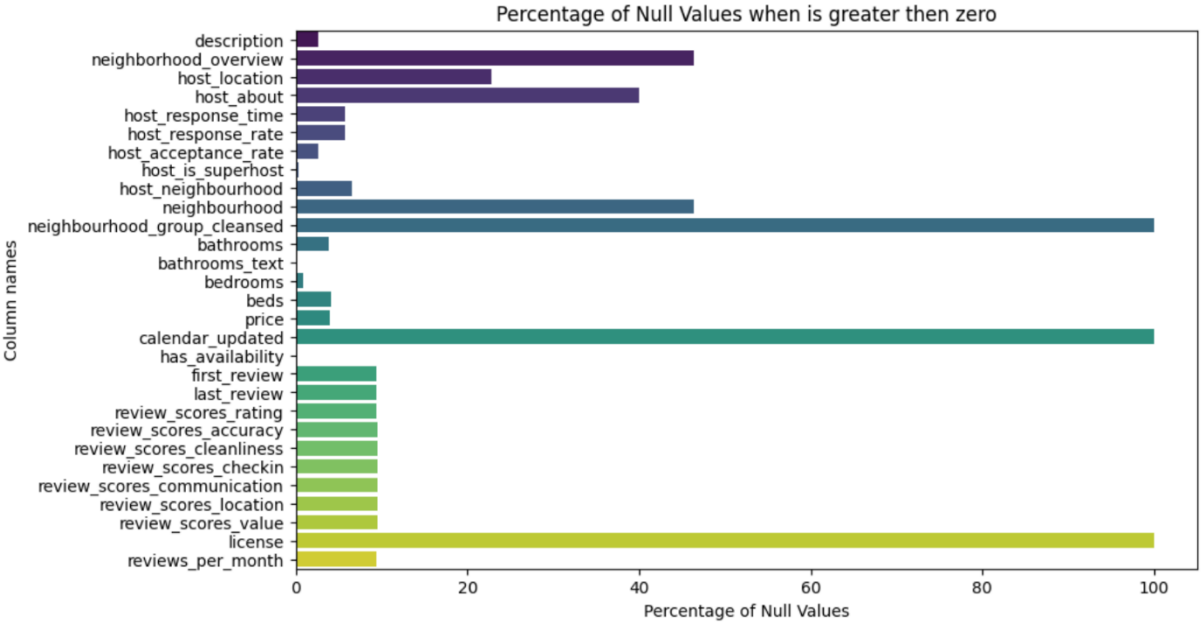
Datetime columns: last_scraped, host_since, calendar_last_scraped, first_review, last_review

Id columns: id, host_id, scrape_id

Boolean columns: host_is_superhost, host_has_profile_pic, host_identity_verified, has_availability, instant_bookable

JSON columns: amenites

The chart above illustrates the percentages of null values present in each column.

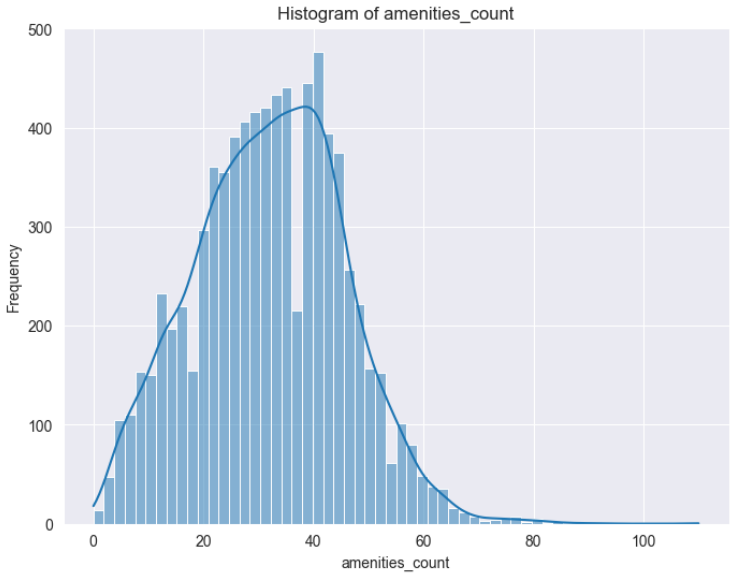


2 Data Preparation

Added Columns:

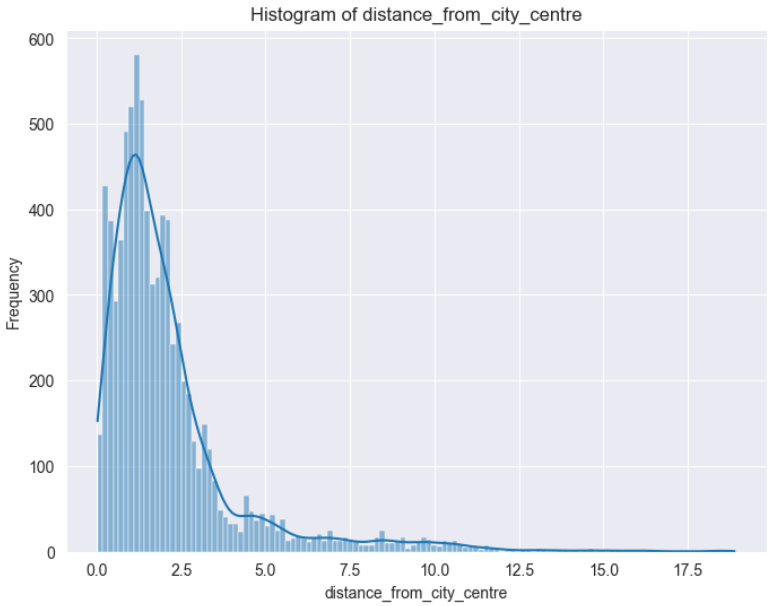
We created some other columns, which we thought would be useful for the modelling.

amenities_count – after parsing the amenities column we added the column amenities_count, which represents how many amenities the Airbnb has



count_verifications – instead of having one column (list of verifications), we decided to have a number of how many verifications host has in another column

distance_from_city_centre – as we all know, when booking an apartment, we all consider how far is the Airbnb from the city centre, so we took the location of Old Town Square as a city centre and calculated thanks to longitude and longitude of the apartment distance in kilometres



season, seasonal_availability – thanks to availability_30, availability_60, availability_90, availability_365 we added a column which represents the seasonal availability

minimal_rating, maximum_rating

Modified Columns:

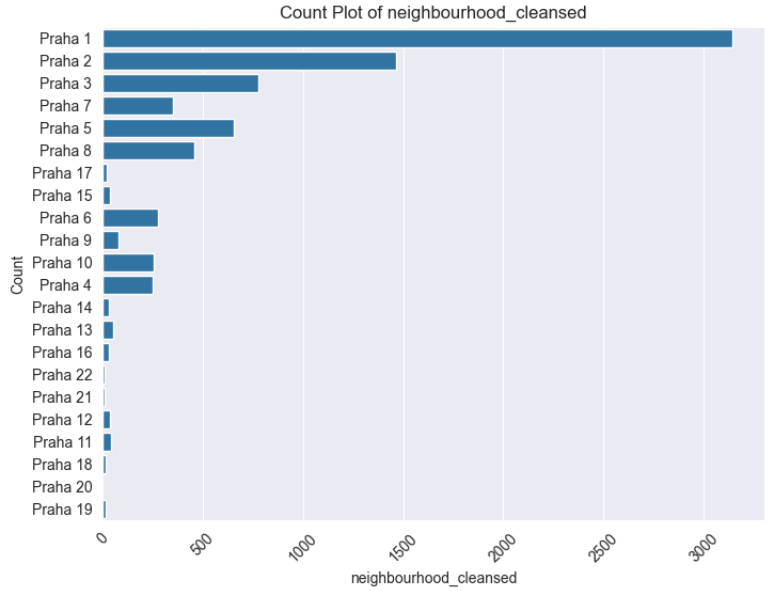
bathrooms_text: text descriptions of bathrooms were extracted into numeric values for easier analysis and aggregation (2 baths into 2)

has_availability, host_has_profile_pic, host_identity_verified, host_is_superhost, instant_bookable: boolean values were converted to 0 and 1, which allows simpler logical operations and integration into mathematical models

host_acceptance_rate, host_response_rate: Percentage values were extracted from the text format and converted to a numeric format

price: We removed the currency symbols and commas, converted it into a numeric format

neighbourhood_cleansed: we unified the neighbourhoods



Removed Columns:

scrape_id, calendar_last_scraped, last_scraped: columns about data collection, not useful for our analytical models

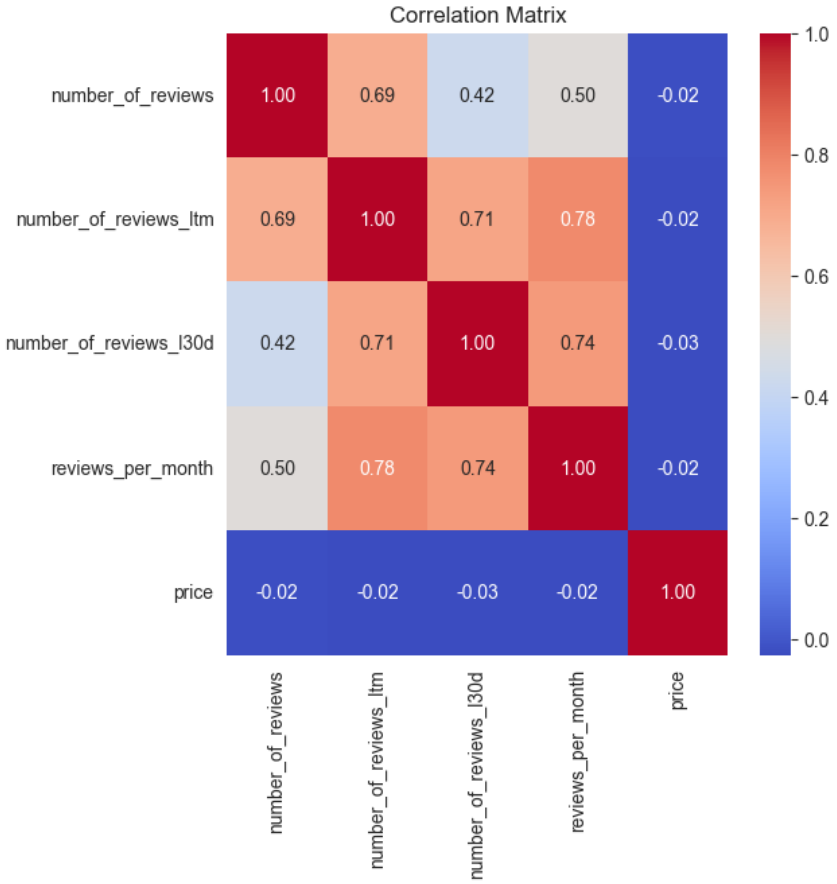
latitude, longitude, neighbourhood, neighbourhood_group_cleansed: we decided to use the information from these columns into new columns neighbourhood_cleansed, distance_from_city_centre, which are more useful for modelling, so we decided to remove the additional columns

description, neighbourhood_overview, amenities: these textual columns were not useful for other modelling, thus they were removed

availability_30, availability_60, availability_90, availability_365: instead of using these columns, column season and seasonal_availability replaced them

For outlier detection, we used the Z-score method to identify unusual data points. These outliers were visualized using box plots and removed.

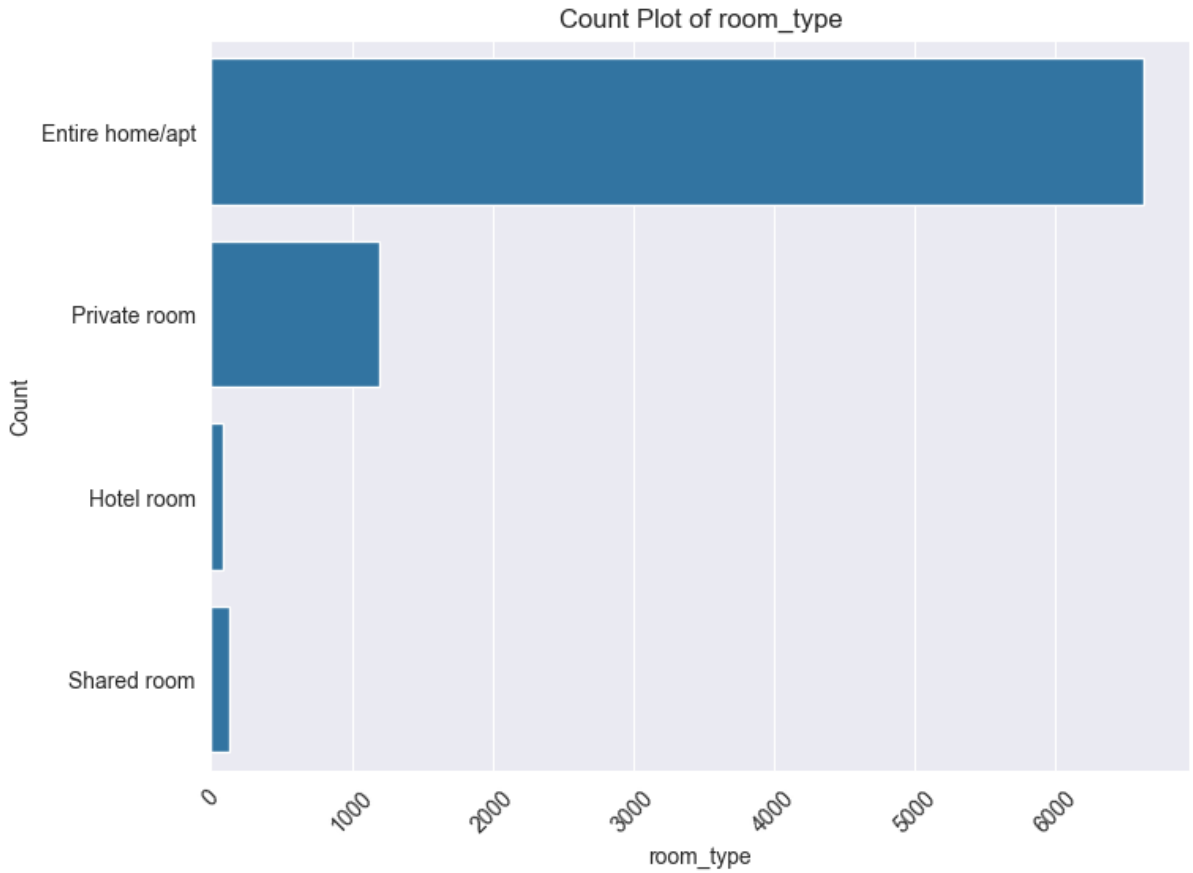
number_of_reviews, number_of_reviews_ltm, review_scores_rating – since we decided to just use minimal_rating and maximum_rating and don't do a sentiment analysis these columns were dropped



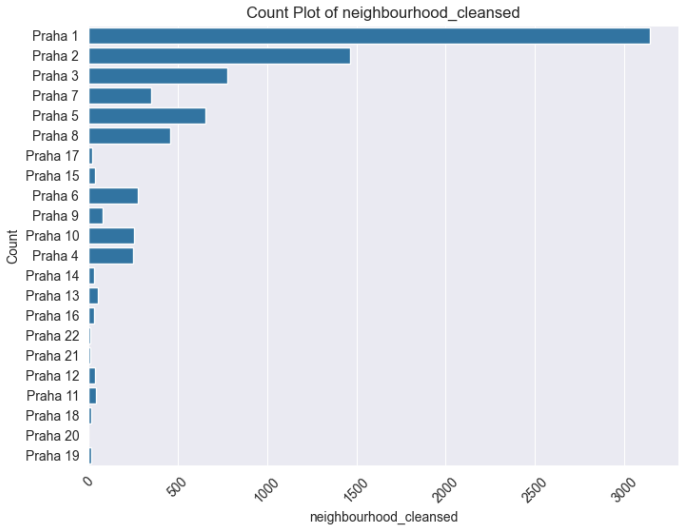
minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm – we thought that enough information is included in minimum_nights and maximum_nights so the others were removed

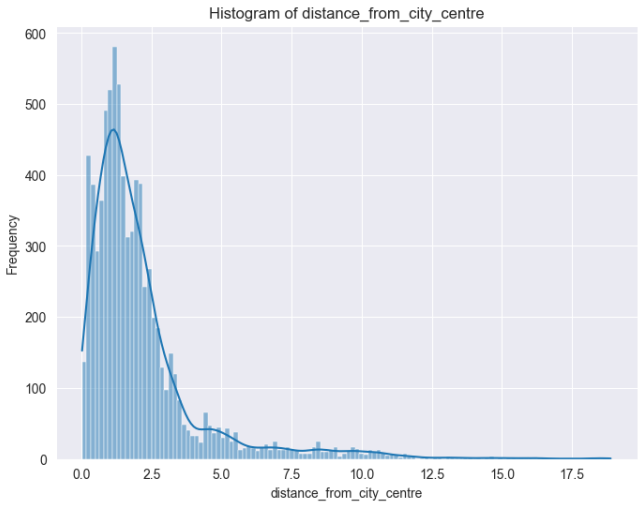
3 Data Visualization

Here we see that most of the Airbnbs are Entire homes/apartments.

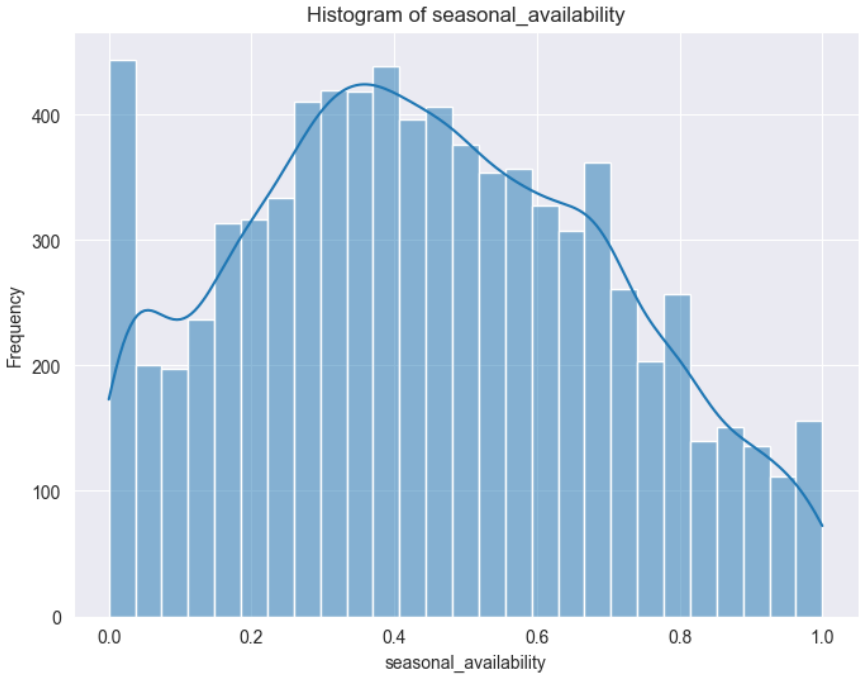


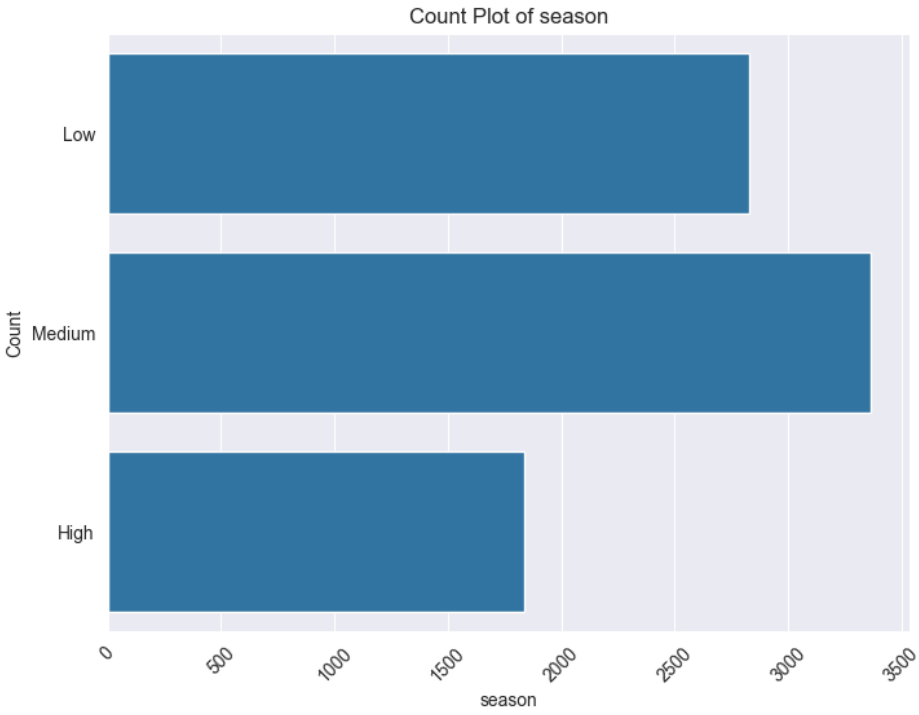
Most of the being in the city centre in Prague1, Prague 2 or Prague 3.



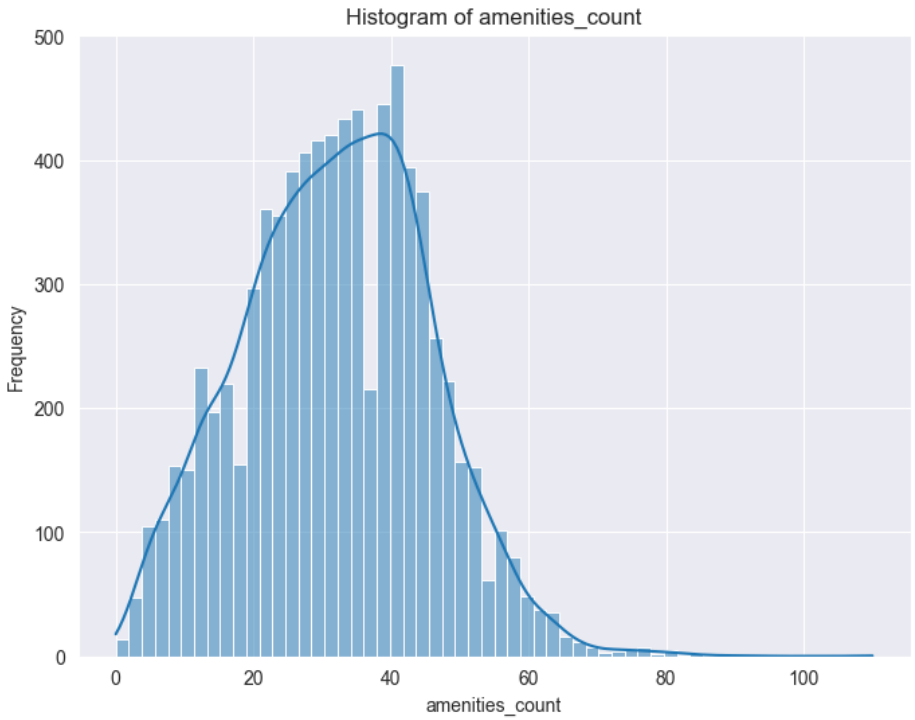


Most of the Airbnb is only available between 40 % and 80 % of the time. Most of the Airbnbs are medium-low available during the time.

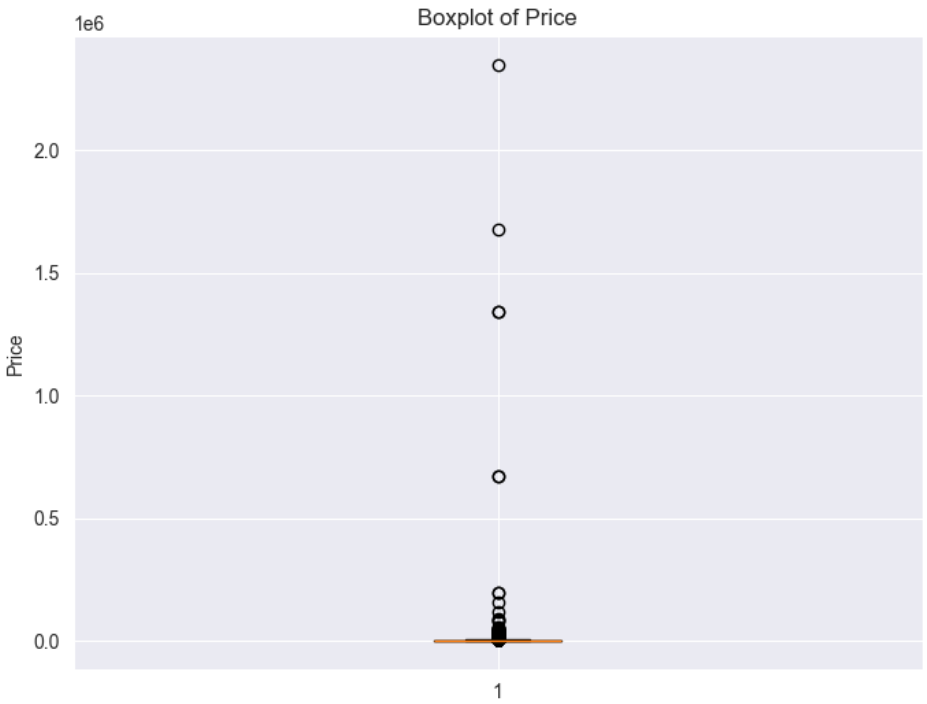




The number of amenities listed.



The median price is very low compared to the maximum, most prices are in the lower range. The outliers may affect the modelling.



4 Modelling

Setting train and test set

In the first step of modelling, we split our data into two groups: a test set and a training set. We divided the data into 20% for the test set and 80% for the training set. The training set is used for training the model, and the test set is used for comparing the models with each other. All models are trained on the same dataset. We used seed 30.

Model selection

For predicting Airbnb prices, we tried using the models Linear Regression, Random Forest, CatBoost, and XGBoost. For Random Forest and XGBoost, we tuned the hyperparameters using cross-validation. We call these improved models Optimized Random Forest and Optimized XGBoost.

Metrics

To compare the models, we chose the R-Squared value. This value, ranging from 0 to 1, shows how well the data fit the regression model.

Linear Regression

R-squared = 0.018

Linear regression with its R-squared value performed clearly the worst.

XGBoost

R-squared = 0.23

XGBoost performed slightly better than linear regression. Even though we tried different hyperparameter settings, we couldn't achieve an optimal model configuration.

CatBoost

R-squared = 0.36

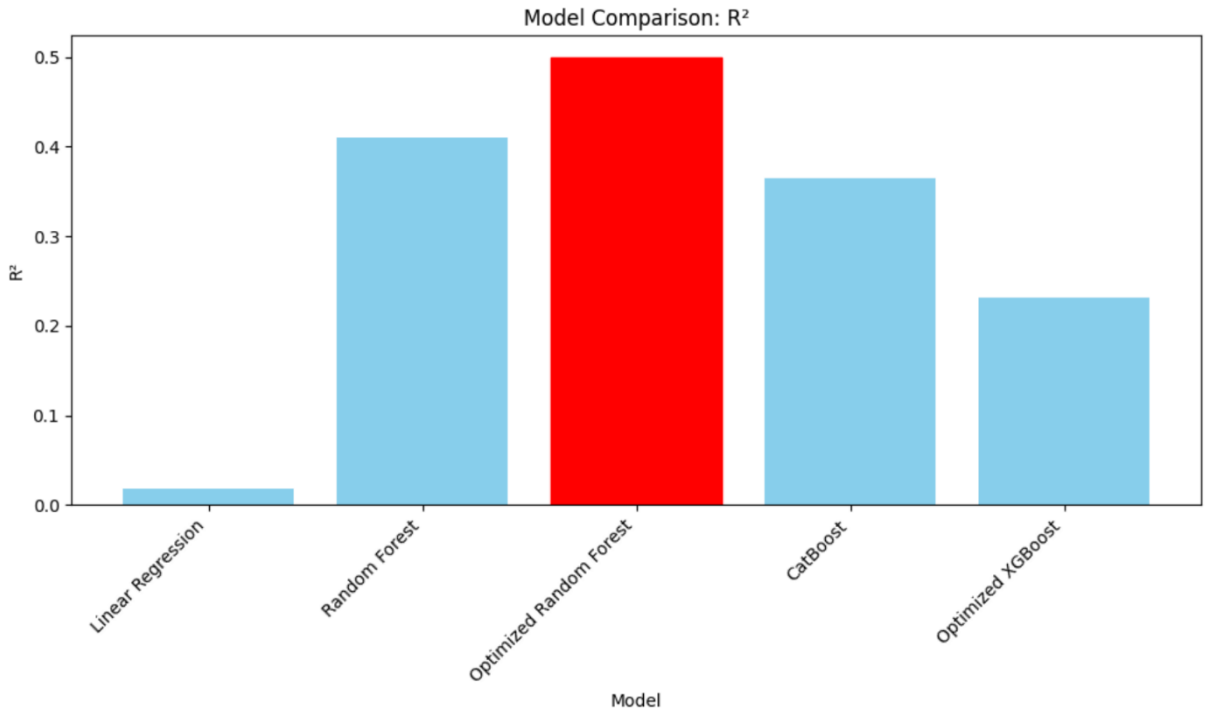
CatBoost performed better compared to XGBoost and linear regression, but its R-squared results were still not the highest.

Random Forest

Default random forest = 0.41

Optimized random forest = 0.50

The model with the highest R-squared is Random Forest. We first tried this model in its basic form, where it surprisingly performed the best. After tuning the hyperparameters, we managed to achieve an R-squared of 0.50. For selecting the best hyperparameters, we chose the GridSearchCV method, which, with the help of cross-validation, selected their optimal values. We selected hyperparameters within a reasonable range based on our knowledge from the lectures.



The image shows a comparison of the different models.

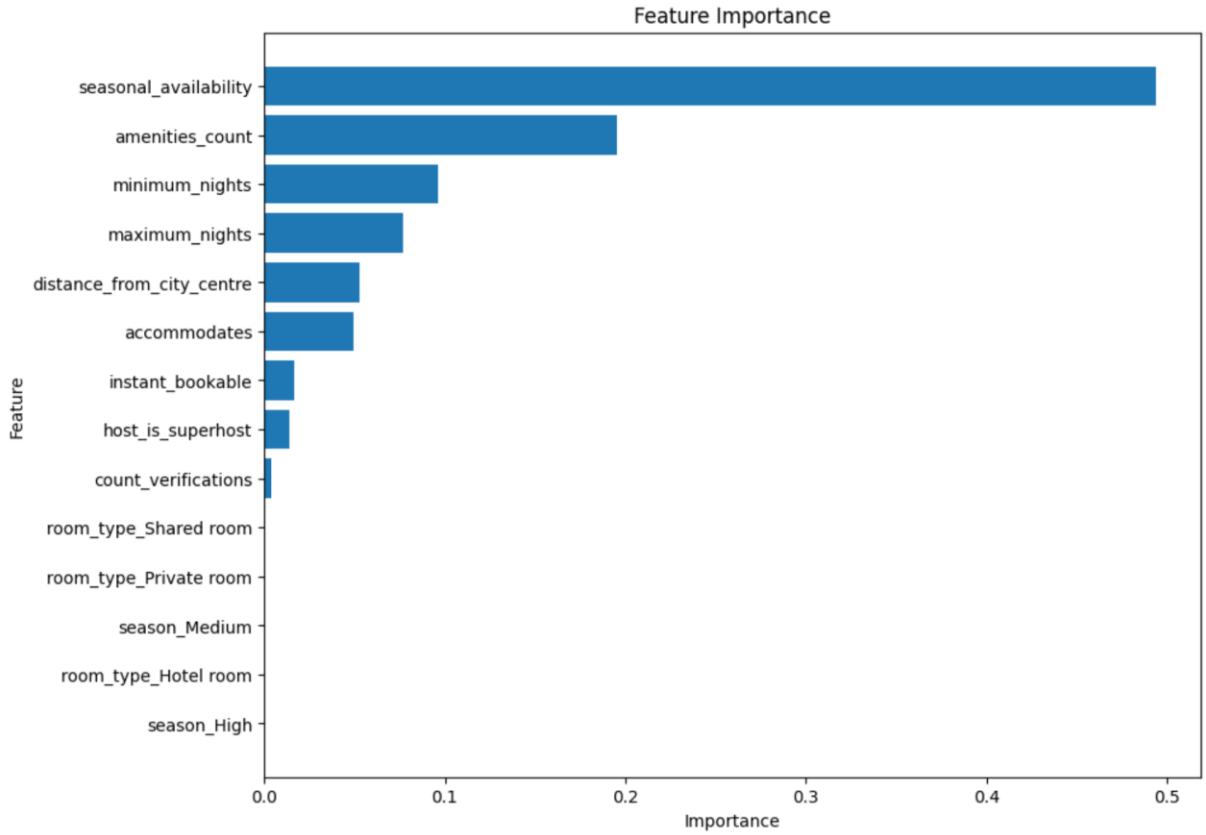
Model limitation

A significant limitation of the models we tested is data processing. Although we spent a considerable amount of time on data manipulation and cleaning, we believe that adding additional information and utilizing current dataset data more comprehensively could lead to better results. The second limitation is the choice of hyperparameters. With greater computational power, we could explore a wider range of hyperparameters and thus refine the results of our model.

5 Model interpretation

Feature importance

The top predictors of property prices in the Optimized Random Forest model are seasonal availability, distance from the city center, and amenities count. While seasonal availability indicates significant price fluctuations, amenities count suggests that more amenities tend to lead to higher prices, alongside proximity to the city center.



Feature importance of Optimized random forest model.