# Airbnb Price Prediction

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

Course: Data X

# 1 Data Understanding

**Listings**

8366 lines and 82 rows

There are in total 7,82 % missing cells in Listings dataset.

Numeric variables: latitude, longitude, accommodates, bathrooms, bedrooms, beds, minimum_nights, maximum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm, calendar_updated, host_id, host_response_time, host_acceptance_rate, host_total_listings_count

Categorical variables: neighbourhood_cleansed, property_type, room_type, host_is_superhost, host_neighbourhood, host_verifications, host_has_profile_pic, host_identity_verified

Unstructured variables: name, description, picture_url, host_location, host_about, amenities

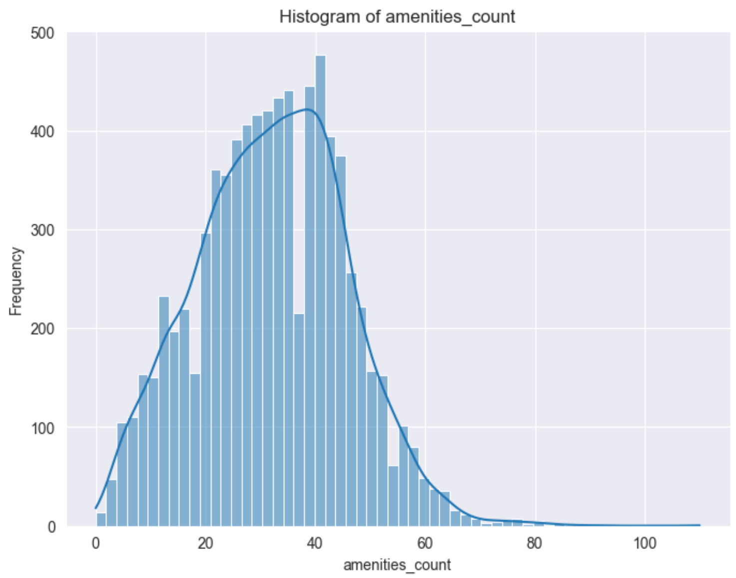
Datetime variables: host_since

# 2 Data Preparation

**Added Columns:**

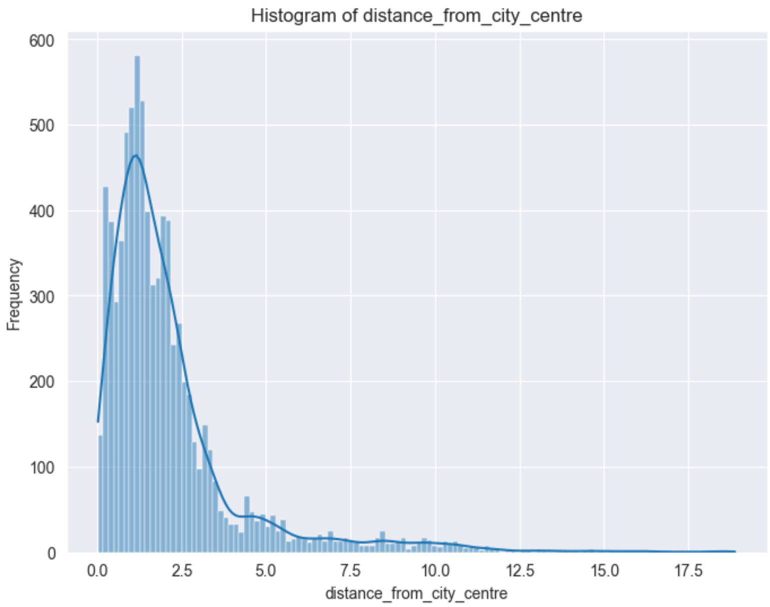
We created some other columns, which we thought would be useful for the modelling.

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

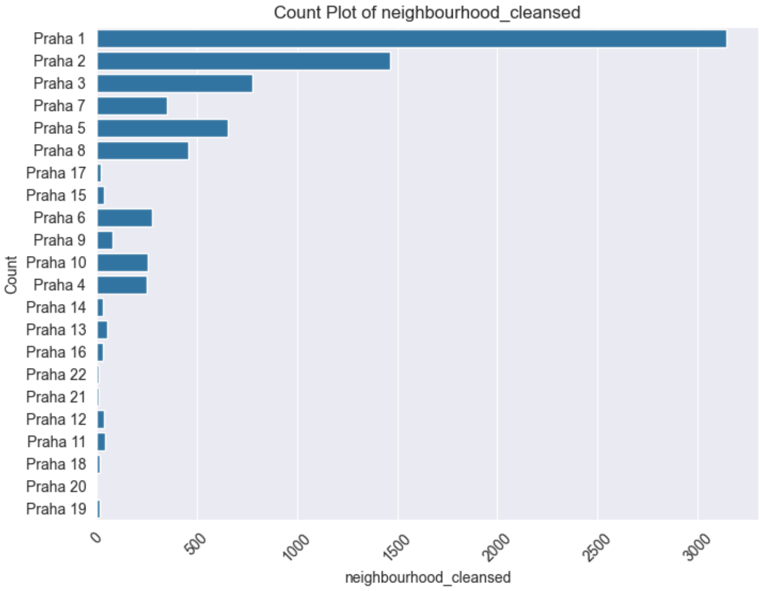
**Modified Columns:**

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



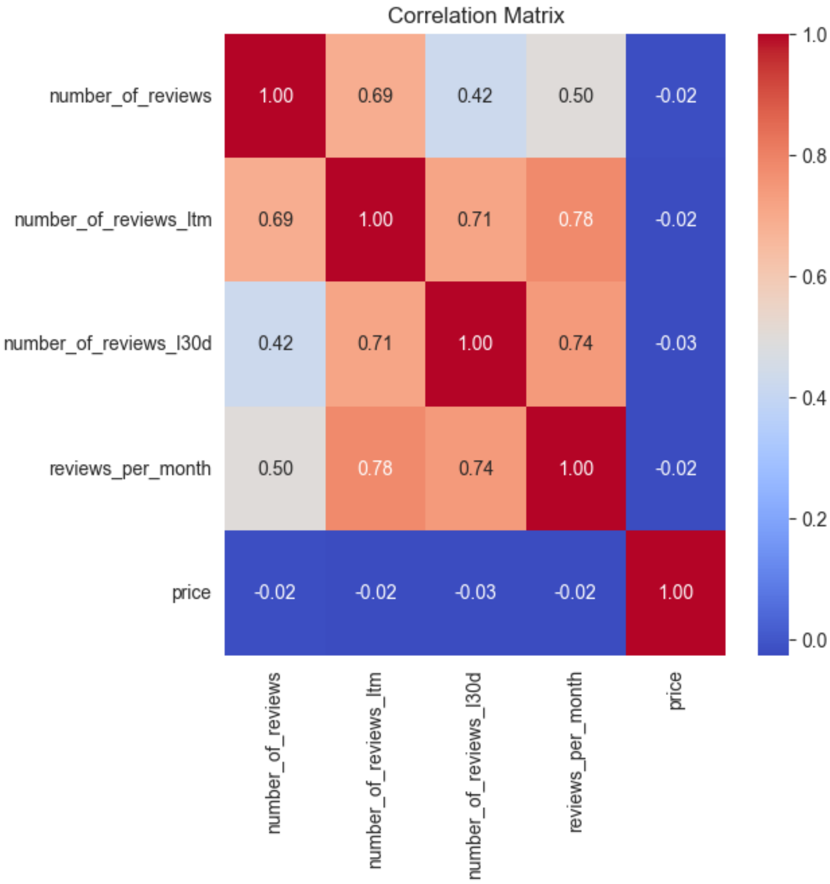

**Removed Columns:**

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

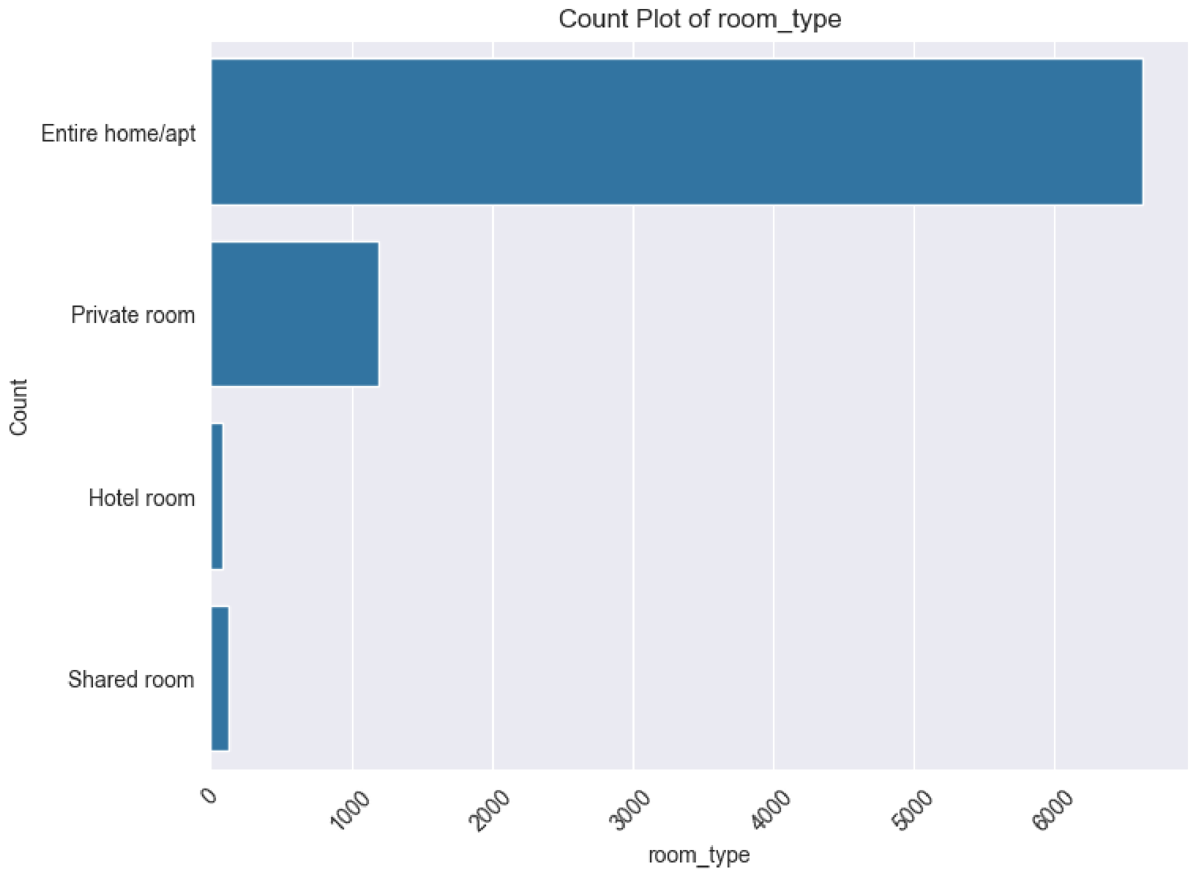
availability_30, availability_60, availability_90, availability_365: instead of using these columns, column season and seasonal_availability replaced them number_of_reviews, number_of_reviews_ltm, review_scores_rating – since we decided to just use minimal_rating and maximum_rating and don't do a sentiment analysis these columns were dropped



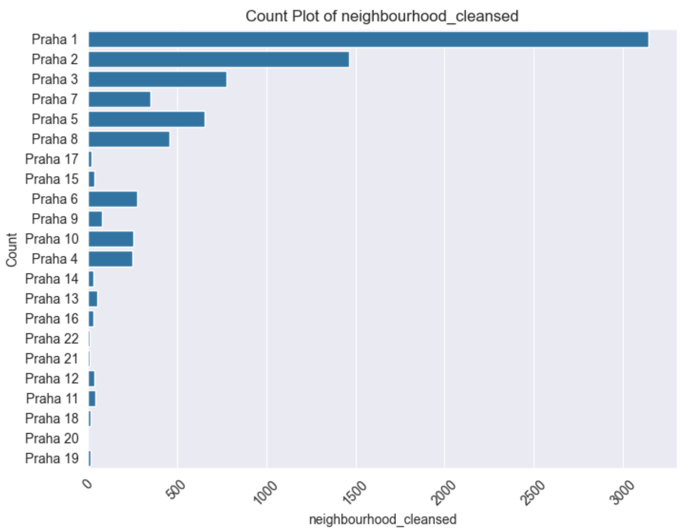

minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm – we thought that enough information is included in minimum_nights and maximum_nights so the others were removed

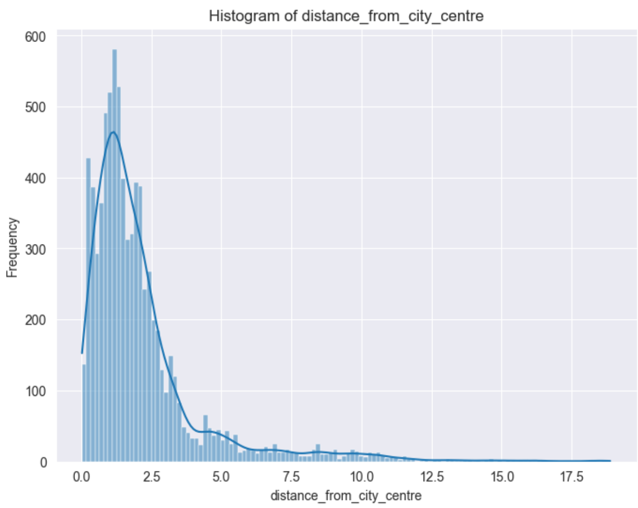
# 3 Data Visualization

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

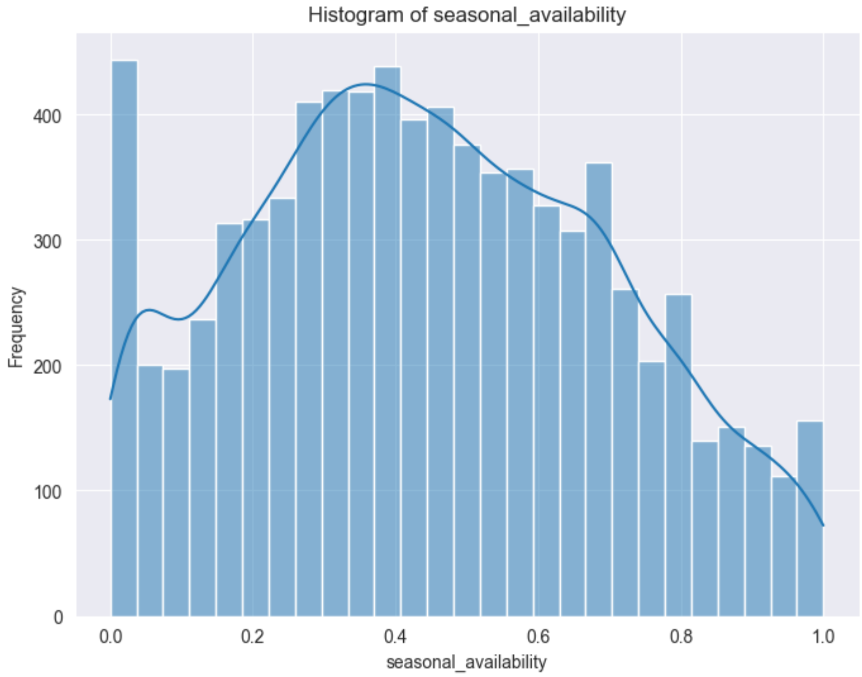


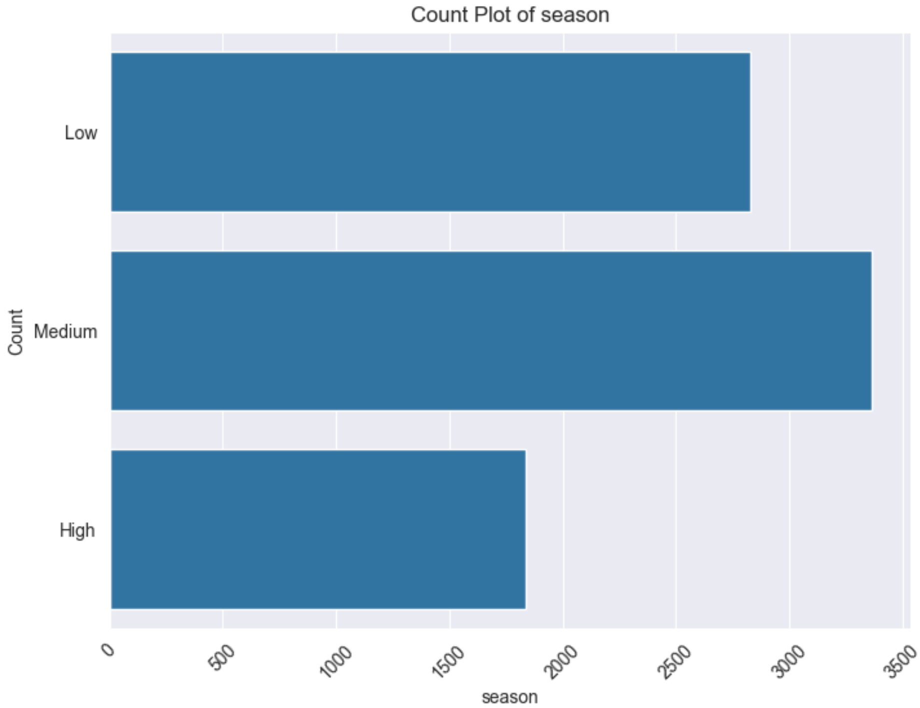

Most of the being in the city centre in Prague1, Prague 2 or Prague 3.



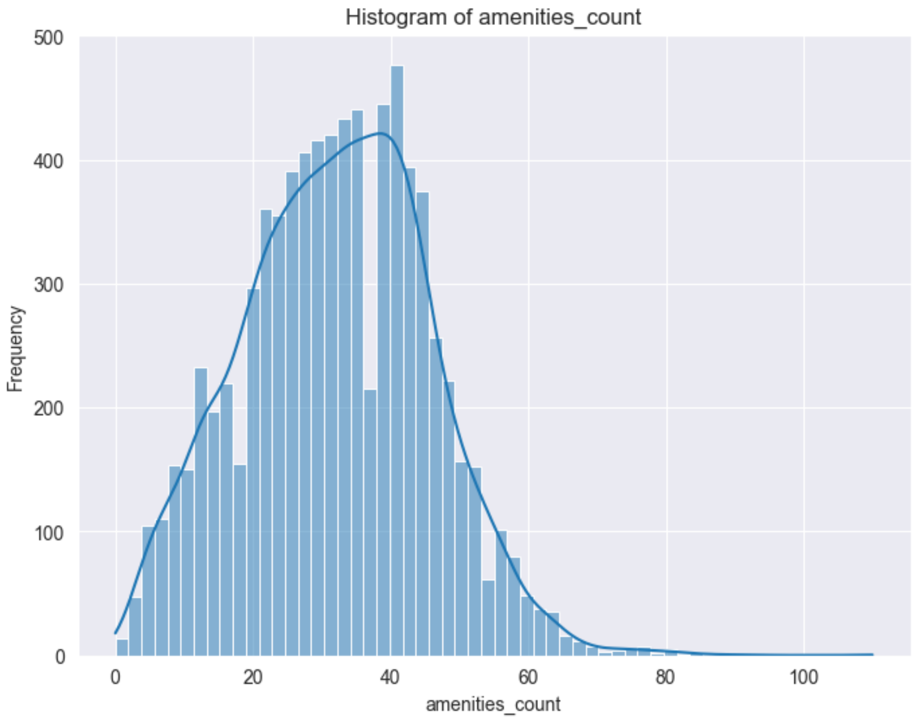



Most of the Airbnb is only available between 40 % and 80 % of the time. Most of the Airbnbs are medium-low available during the time.

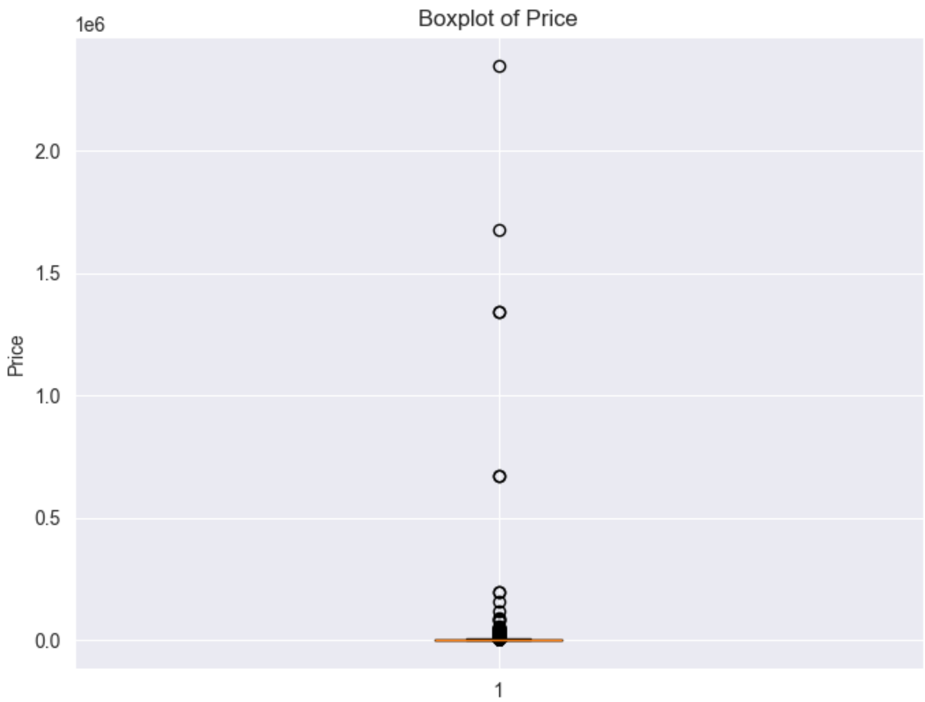


The number of amenities listed.




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



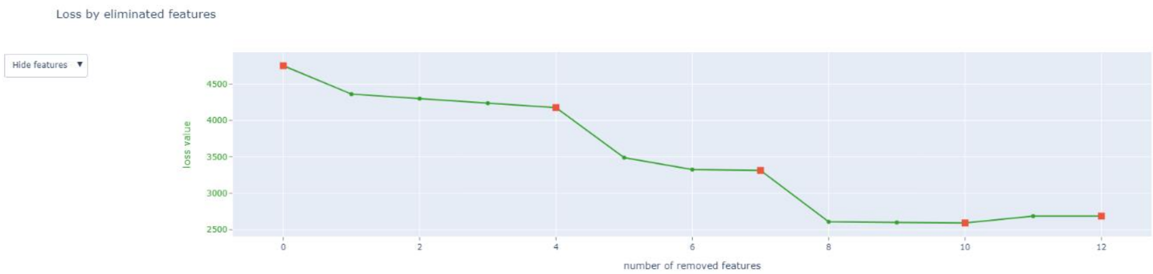
# 4 Modelling

**Model Selection**

In our analysis, we evaluated various predictive models to identify the one best suited for our dataset. Linear Regression, despite being a fundamental approach, yielded very poor performance with an R2 of only 3%, indicating a weak explanatory power for our data's variability. The Random Forest model performed significantly better, achieving a 34% R2 score, making it the second-best model in our trials. However, the CatBoost model outperformed all others with an initial R2 of 56%. Through the application of Recursive Feature Elimination (RFE), we were able to enhance its performance significantly, boosting the R2 to 89%. This substantial improvement highlights CatBoost's capability in handling categorical features and its robustness against overfitting.

**Feature Selection**

The model performed better as more features were removed




**Model limitations and considerations**

While it provides superior accuracy, its training time is considerably longer, which might be a constraint.