Predicting the Cost of an Apartment using ML

The goal of this project is to develop a model that can accurately predict the price of an apartment based on various features such as location, size, and amenities. In this presentation, we will cover everything from preparing the data for selecting the best machine learning algorithm for the task.

Avetisov Arsenii 212
Stepanova Darina 213
Shafran Boris 213
Yurchenkov Nikita 213

Contribution to the project

1) Nikita Yurchenckov - data analysis
2) Boris Shafran - Linear regression
3) Avetisov Arsenii - Random Forest
4) Stepanova Darina - Random Forest

Understanding the Data

Our dataset is focused on apartment costs and contains 30 features and 29,044 objects.

ⓘ

9 are float64

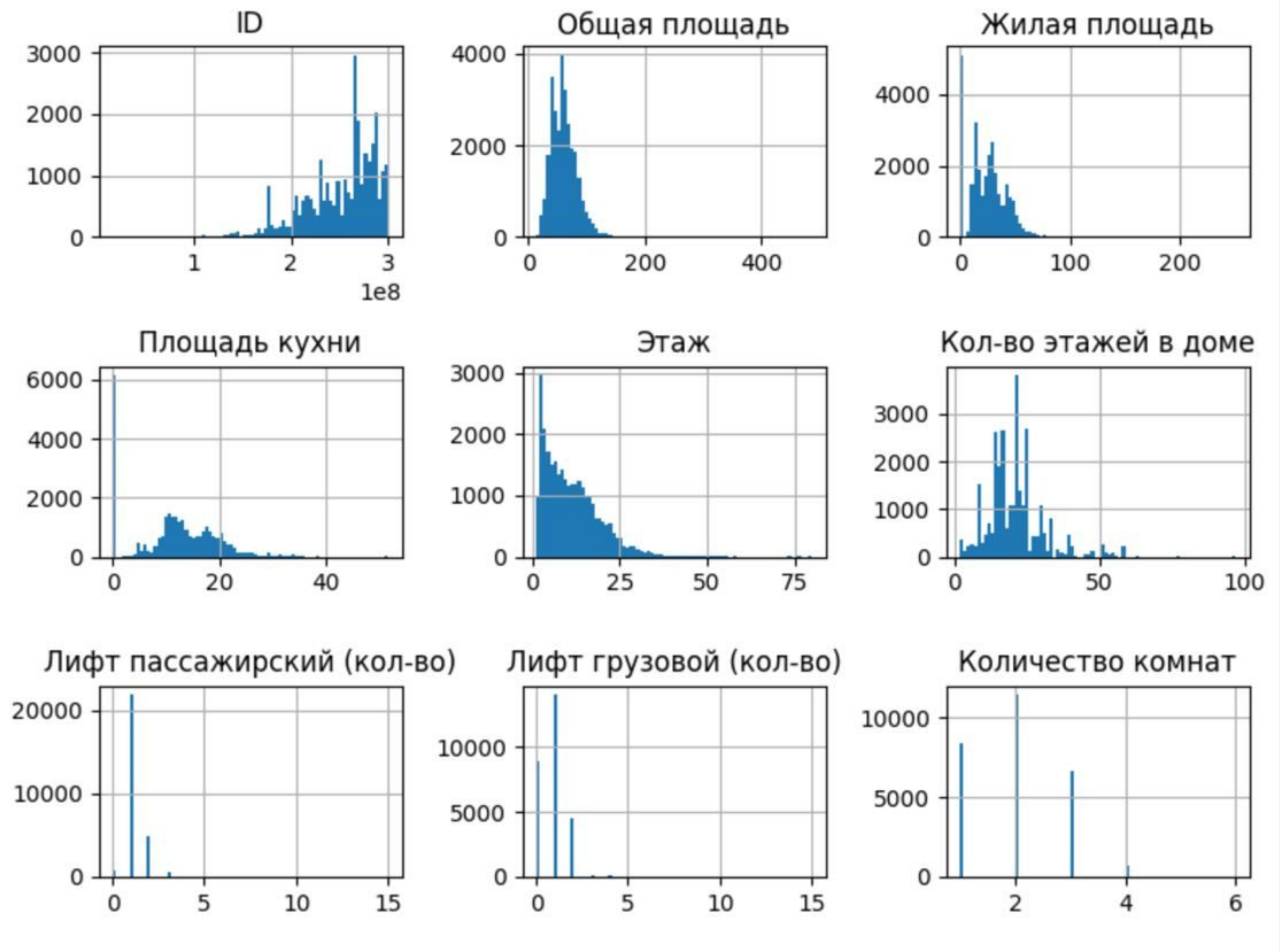
4 are int64

17 are object

Understanding the data is key to making informed decisions and accurate predictions.

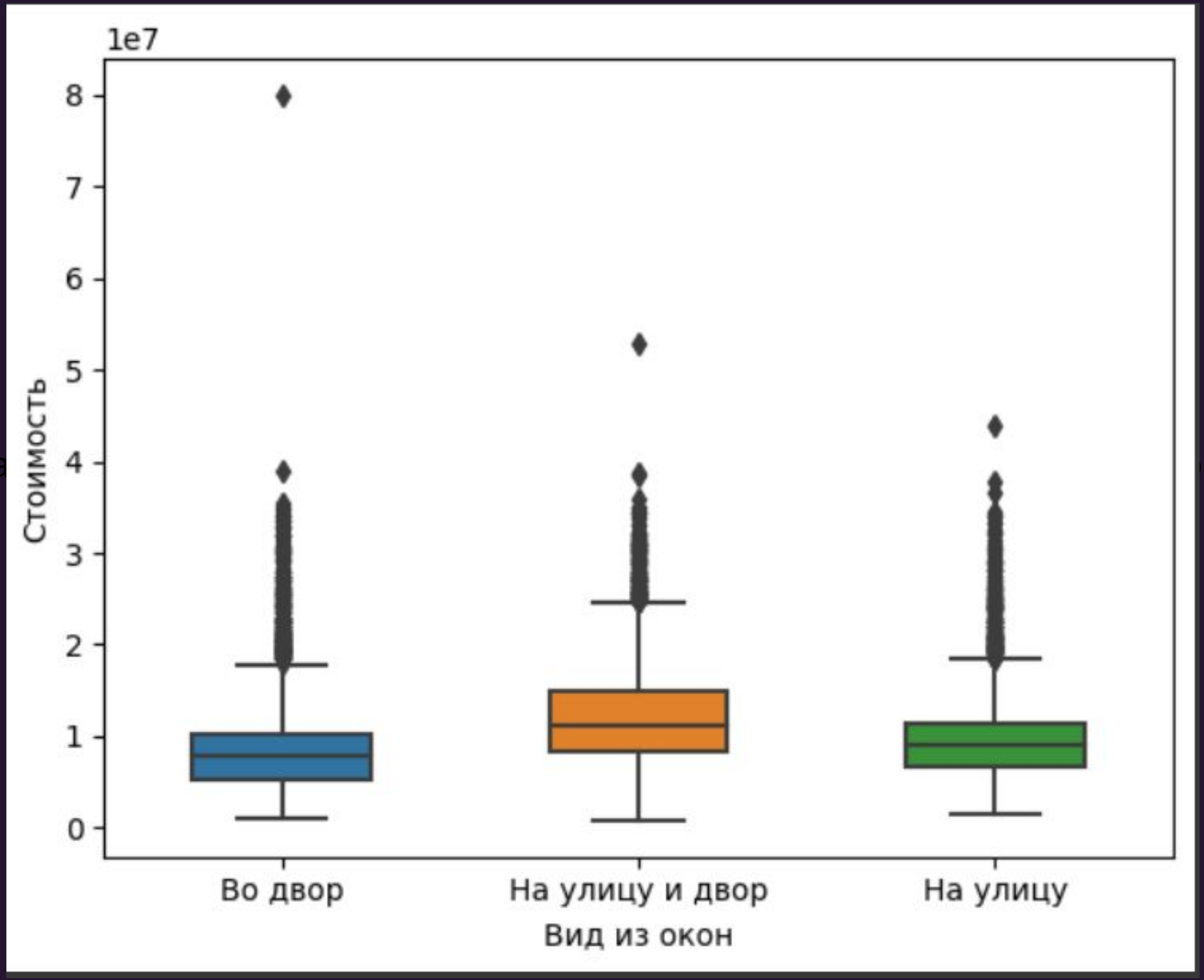
EDA Part

Our team provided EDA analysis to understand main patterns of the market and pay attention to redundant features.

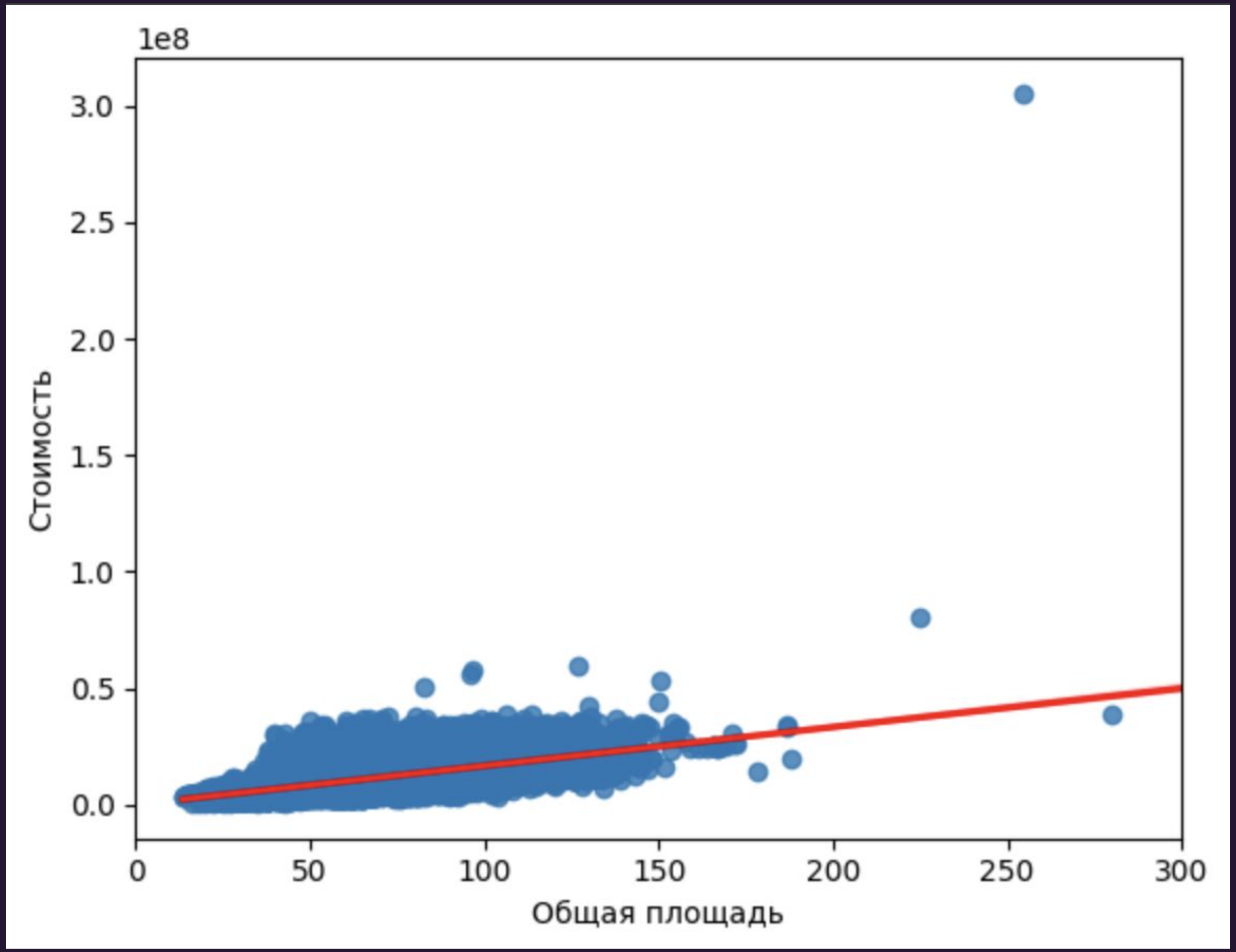


EDA Part

a gra



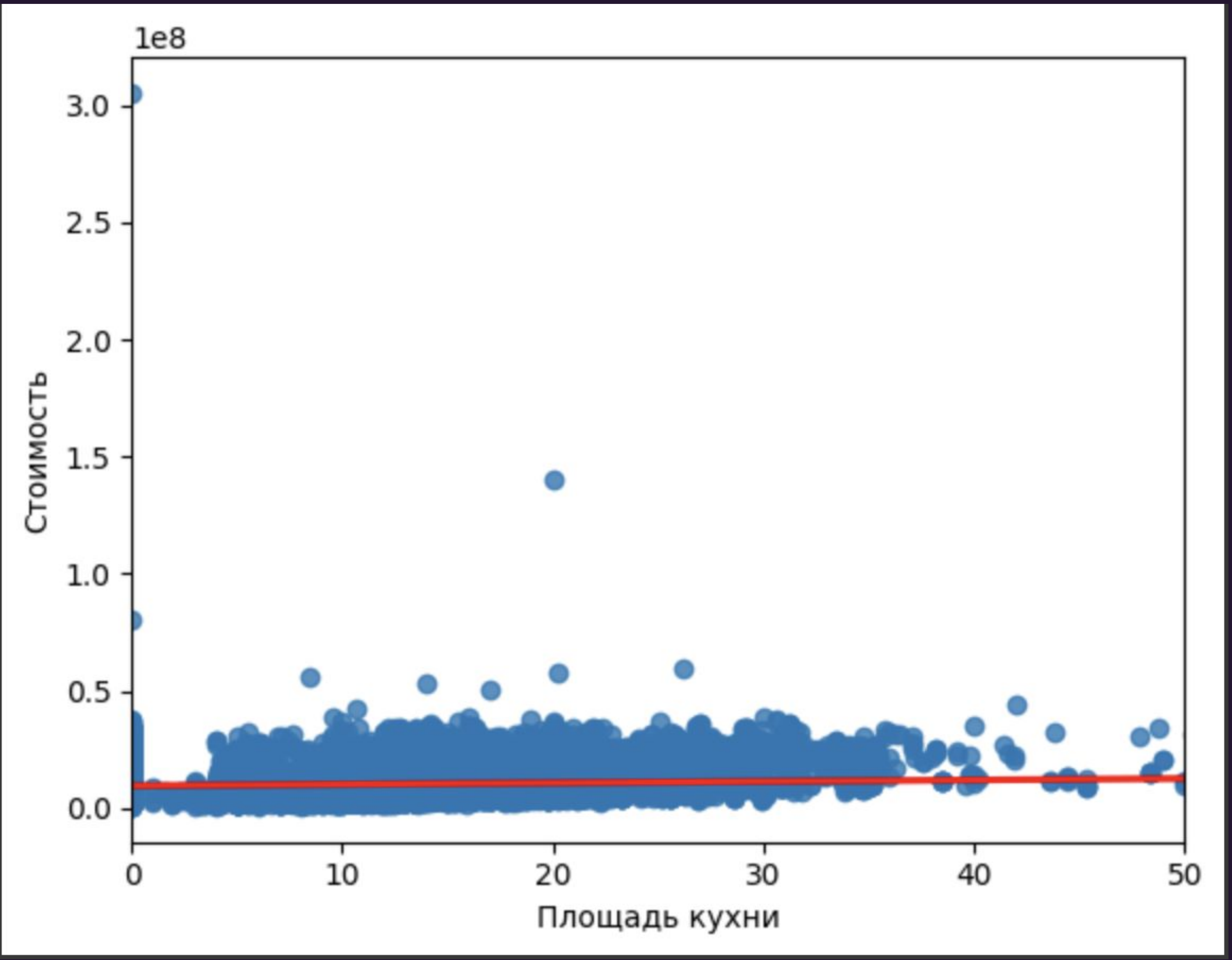
rice



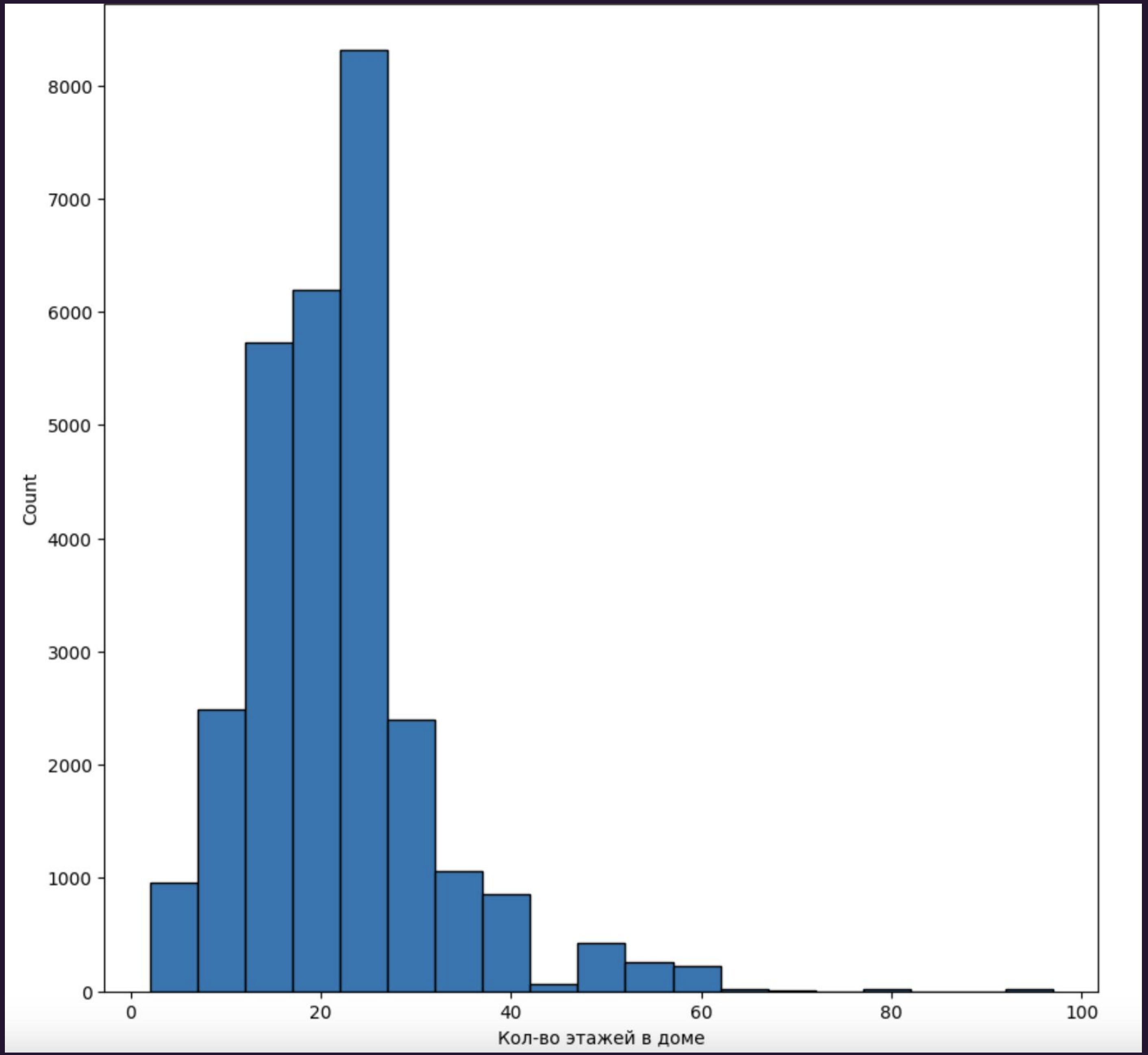
A graph showing the relationship between the price and the location of the windows (courtyard/street)

A graph showing the relationship between the total area of the house and the price

EDA Part



This graph shows the relationship between cost and kitchen area in the house

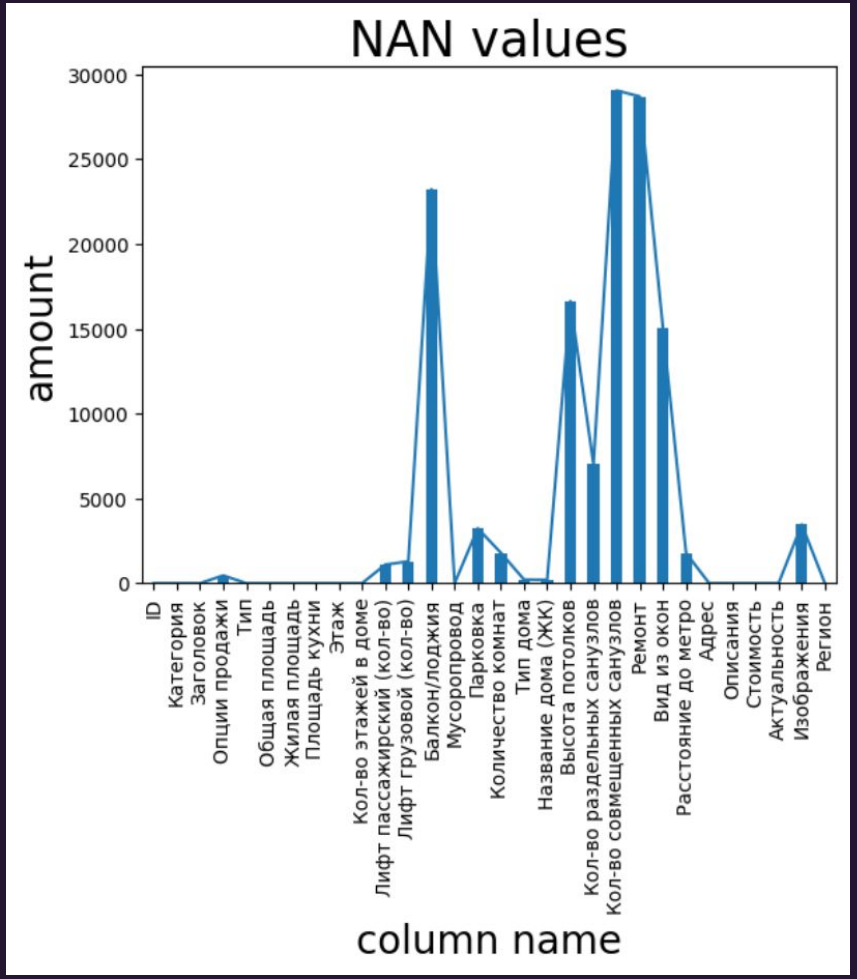


A simple graph showing how many houses have a certain number of floor levels

Cleaning Data from NaN Values

To clean the data, we first dropped columns with nonsense information. We then searched for and dropped columns with a high percentage of NaN values. Finally, we removed duplicates and the cost column.

Figure out columns with nonsense info



Then we have removed these columns

1. Категория
2. Мусоропровод
3. Заголовок
4. Адрес
5. Название дома(ЖК)
6. Описания
7. Изображения
8. Кол-во совмещенных санузлов
9. Ремонт
10. Балкон/лоджия
11. Высота потолка
12. Вид из окон

Linear Regression

Linear regression is a supervised learning algorithm used in machine learning to predict a continuous output variable based on one or more input variables. It assumes a linear relationship between the input and output variables and tries to find the best fit line that minimizes the difference between the predicted and actual values. It is commonly

✓ **Mean Squared Logarithmic Error**

0.3607976113852592

Cat boost regression advantages

CatBoost is a gradient boosting algorithm used in machine learning that can handle categorical features and missing values. It helps to improve the accuracy of the model by reducing overfitting, handling imbalanced data, and providing better generalization. It also has built-in feature selection and hyperparameter tuning capabilities.

- Handling Categorical Features CatBoost Regression models are better at processing categorical features, efficiently handling them and avoiding the pre-processing problem.

- Avoiding Overfitting CatBoost models have the ability to avoid overfitting by using random permutations in boosting.

- Handling Missing Values and Optimization CatBoost Regression models handle missing values intuitively by filling in the missing values with predictions from other features.

Cat boost regression

(the best as we will see later)

'iterations': 2500

'learning_rate': 0.03

'depth': 6

'l2_leaf_reg': 0

'loss_function': told by us

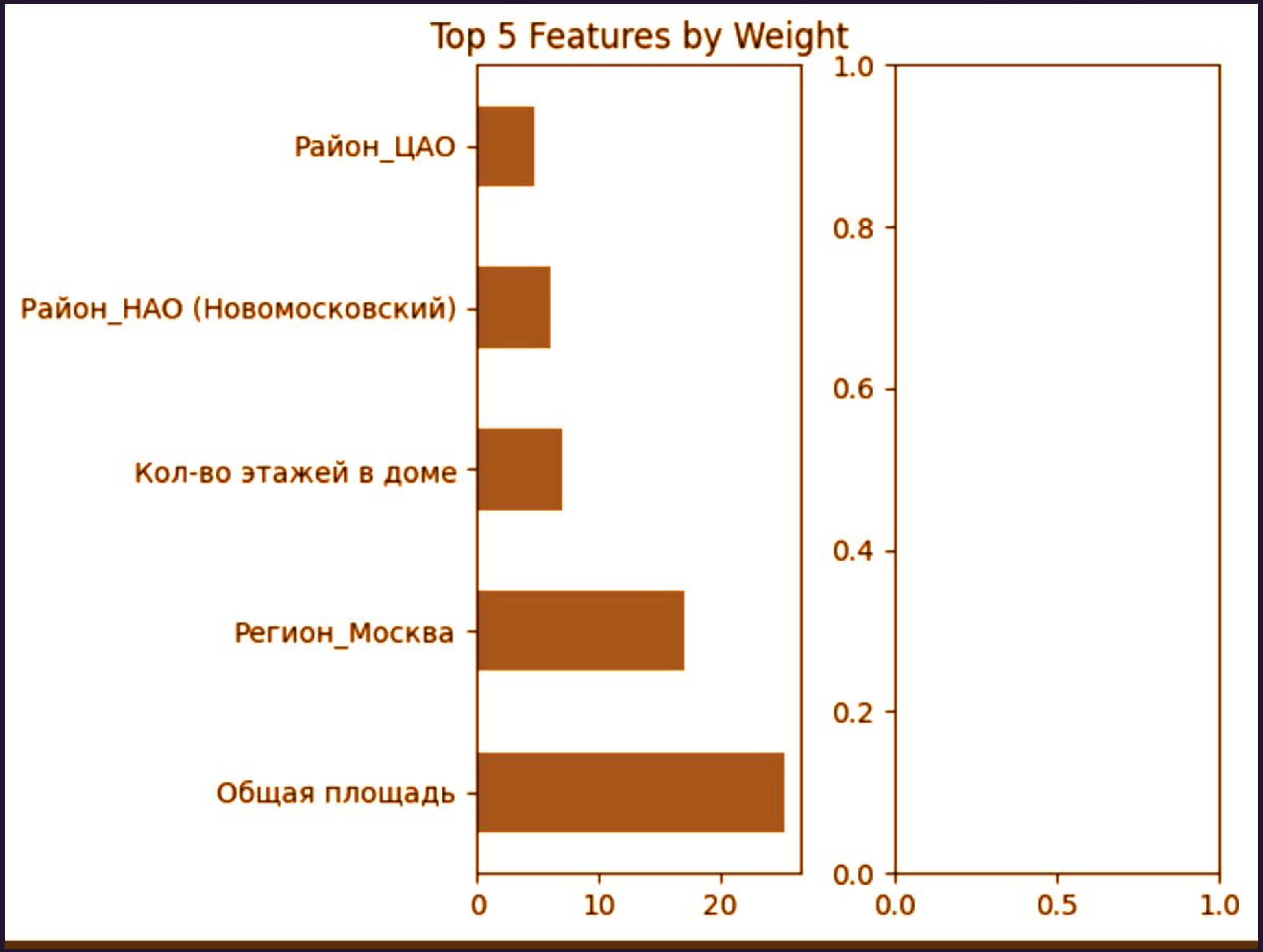
'eval_metric': told by us

'early_stopping_rounds': 100

Mean Squared Logarithmic error after CatBoost

0.08378167204733095

Feature importance (Top 5)



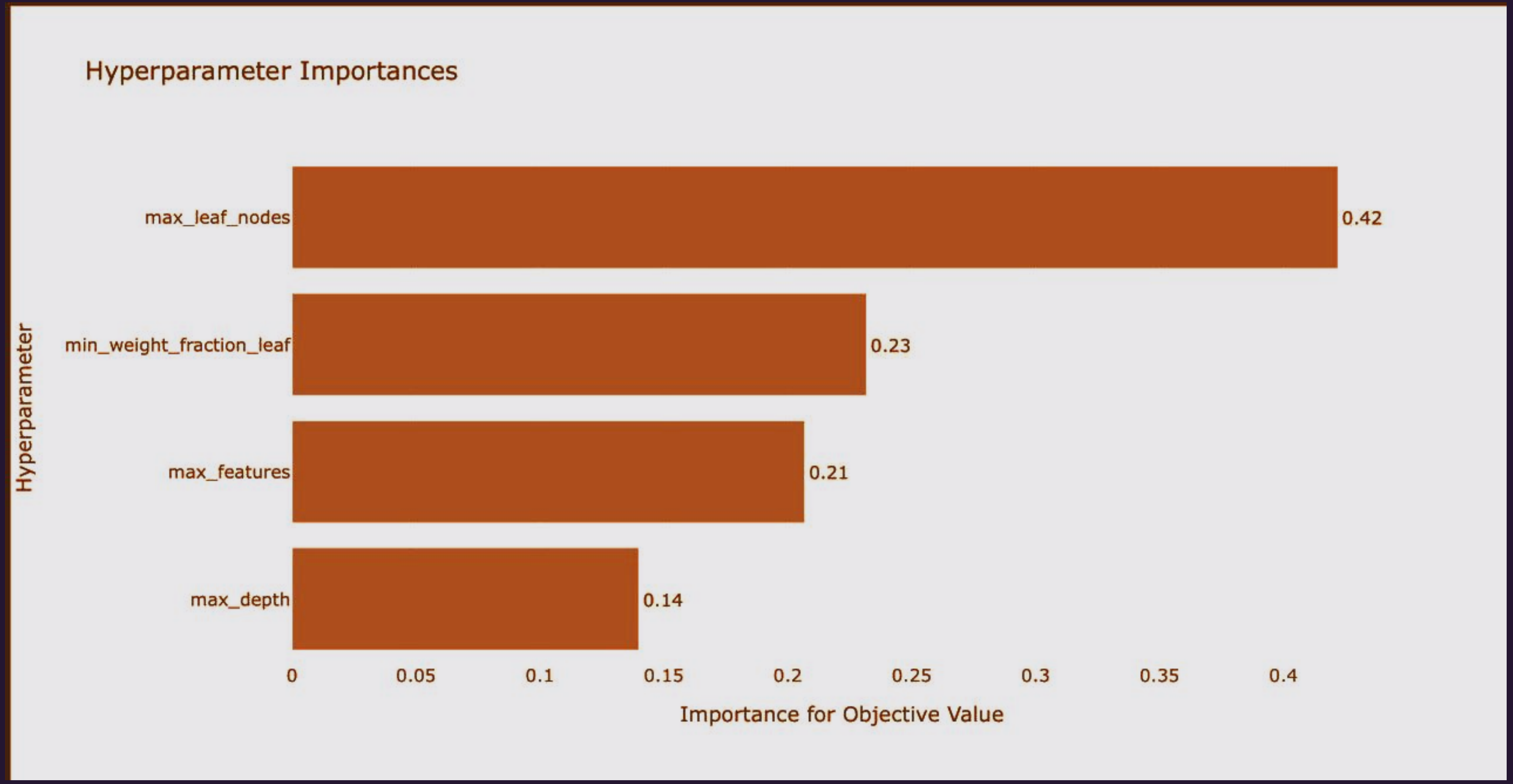
RandomForest

The RandomForest algorithm is a type of decision tree model used in machine learning. It works by creating a large number of decision trees and combining their results to make a final prediction.

✓ Mean Squared Logarithmic Error

0.4509251048440

Hyper par importance



Final Solution for Random Forest

However, there is still room for improvement in the model. One way to do this is through hyperparameter tuning using GridSearch. GridSearch is a method of searching through a range of hyperparameters to find the combination that results in the best model performance. By using GridSearch, we can optimize the RandomForest algorithm even further and potentially improve the accuracy of our predictions.

GridSearch Hyperparameters:

max_depth	280059
max_leaf_nodes	280059
min_weight_fraction_leaf	0.011441552100106757
max_features	0.1806166105643915
Mean Squared Logarithmic Error	0.4509251048440

QR Repository