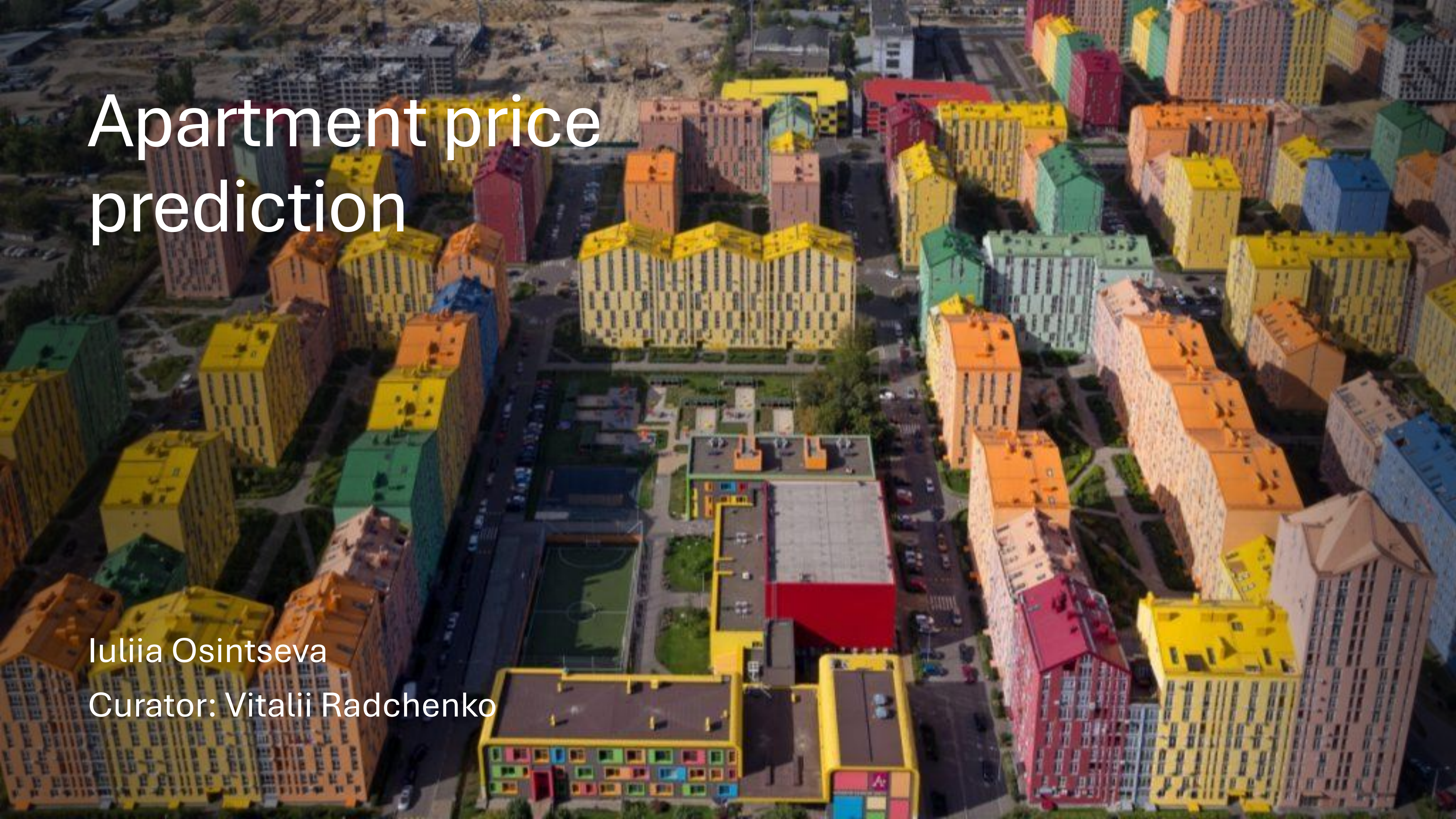


Apartment price prediction

Iuliia Osintseva

Curator: Vitalii Radchenko

Apartment price prediction

01

Task:

Build a prediction model to estimate an apartment price as precise as possible

*regression task

02

Goal:

Estimate the prices for the secondary market apartment. Luxury properties are not considered.

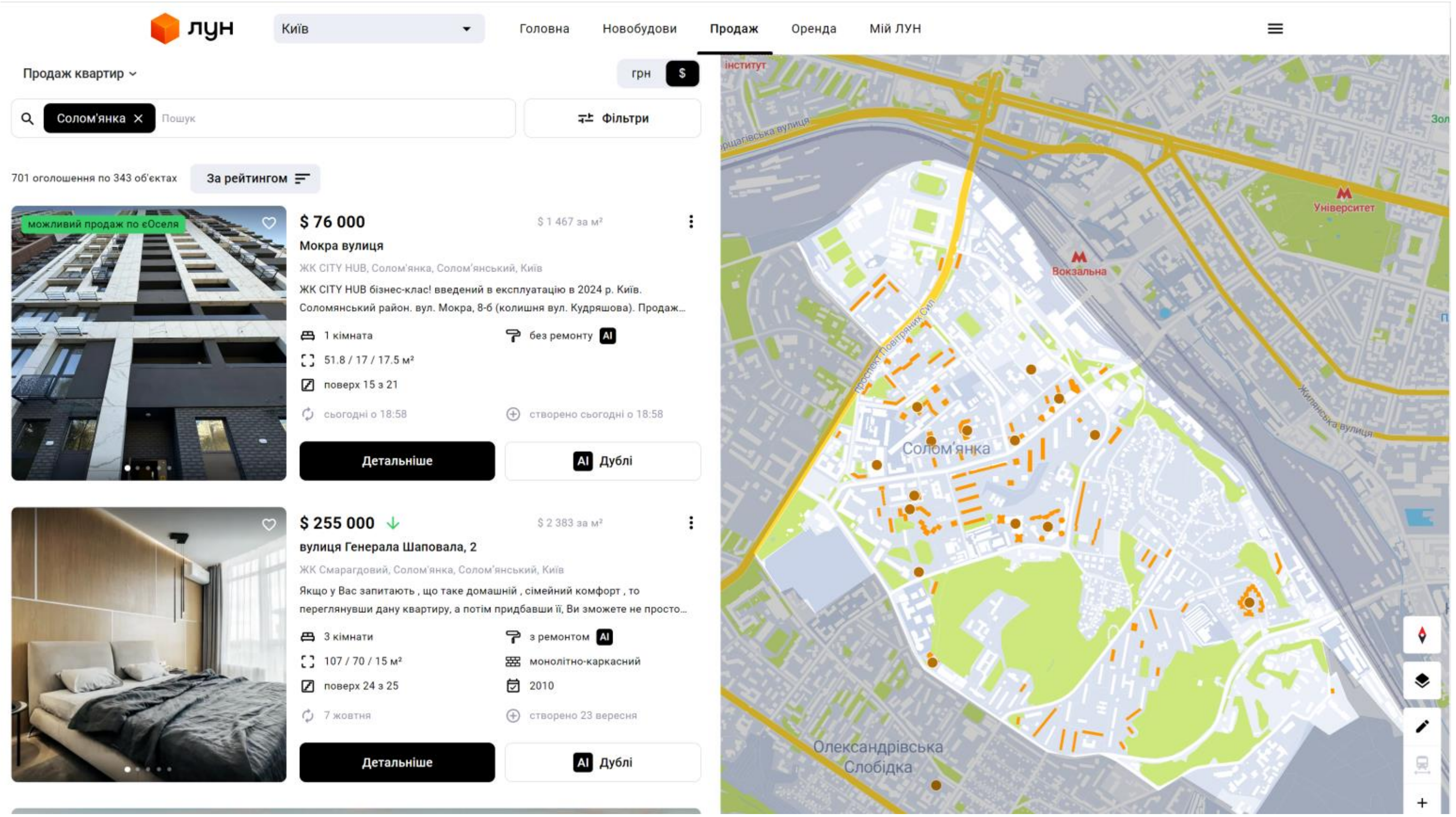
03

Metrics:

Excluding luxurious apartments allows using MAE as metrics.

Data collection and preparation

- Data was scraped from lun.ua
- Scraping happend in one iteration, so historical data on price development is unavailable
- 13.835 ads
- Available data:
  - Price, Price/m2
  - District ID
  - Rooms
  - Total Area
  - Kitchen Area
  - Floor (+ total floors)
  - Coordinates
  - Renovation (True/False)
  - Type of Construction
  - Year of Construction



701 оголошення по 343 обʼєктах

$ 76 000 — $ 1 467 за м²

Мокра вулиця

ЖК CITY HUB, Соломʼянка, Соломʼянський, Київ

ЖК CITY HUB бізнес-клас! введений в експлуатацію в 2024 р. Київ. Соломянський район. вул. Мокра, 8-б (колишня вул. Кудряшова). Продаж...

1 кімната — без ремонту AI
51.8 / 17 / 17.5 м²
поверх 15 з 21
сьогодні о 18:58 — створено сьогодні о 18:58

$ 255 000 — $ 2 383 за м²

вулиця Генерала Шаповала, 2

ЖК Смарагдовий, Соломʼянка, Соломʼянський, Київ

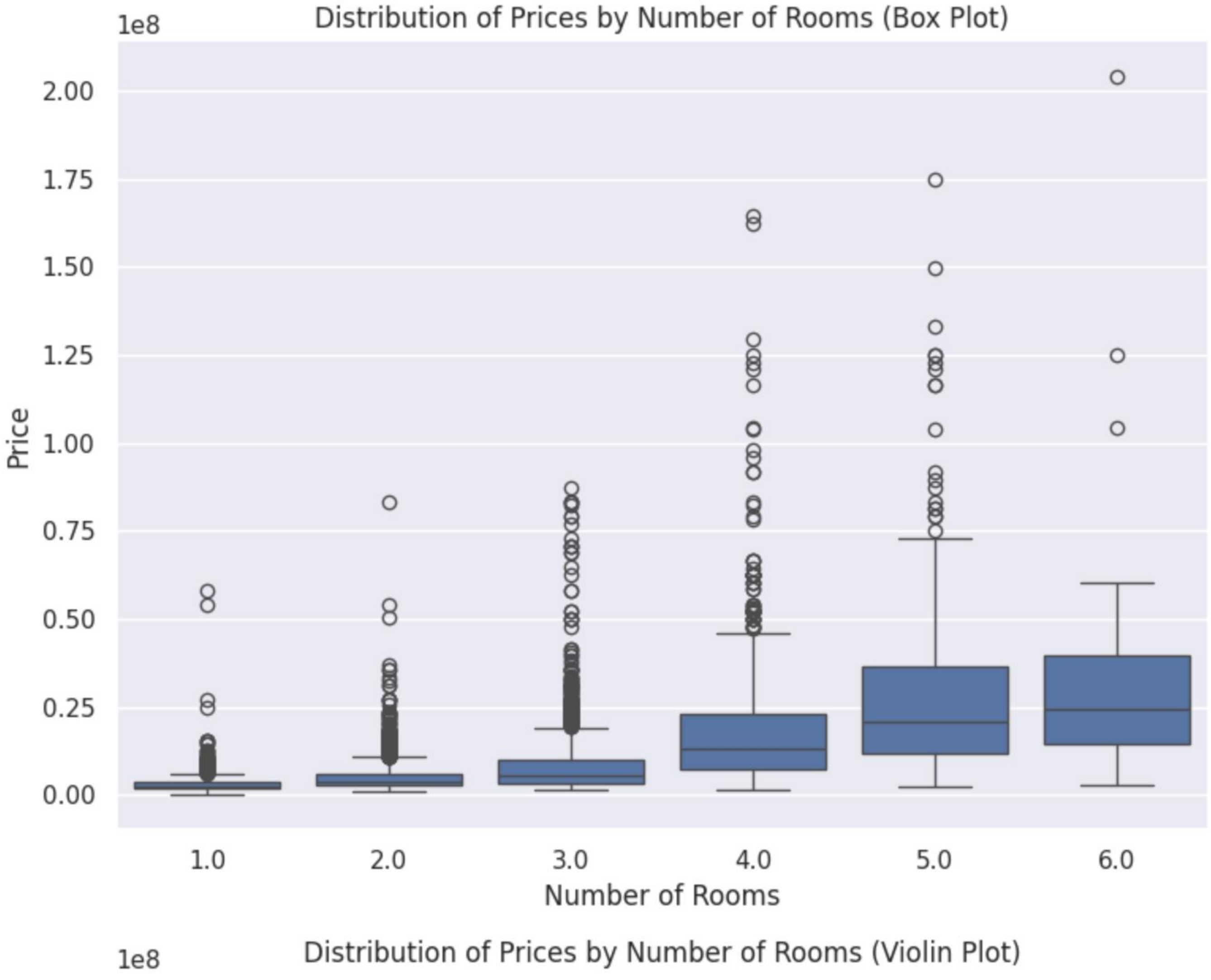
Якщо у Вас запитають , що таке домашній , сімейний комфорт , то переглянувши дану квартиру, а потім придбавши її, Ви зможете не просто...

3 кімнати — з ремонтом AI
107 / 70 / 15 м² — монолітно-каркасний
поверх 24 з 25 — 2010
7 жовтня — створено 23 вересня

EDA

*see Apartment_Prices_Cleanup.ipynb

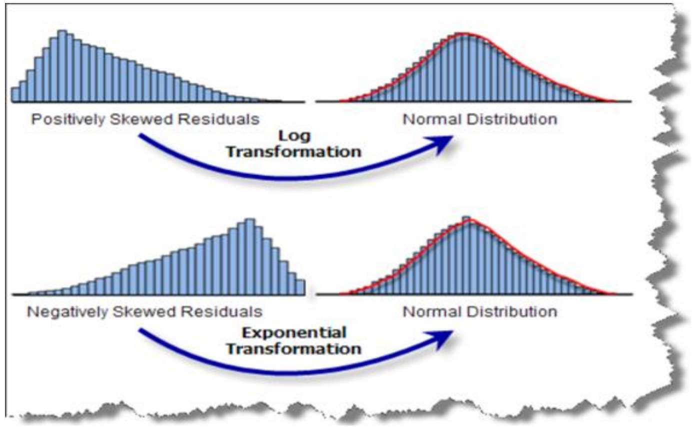
- Obviously, the average price is dependent from the nr of rooms



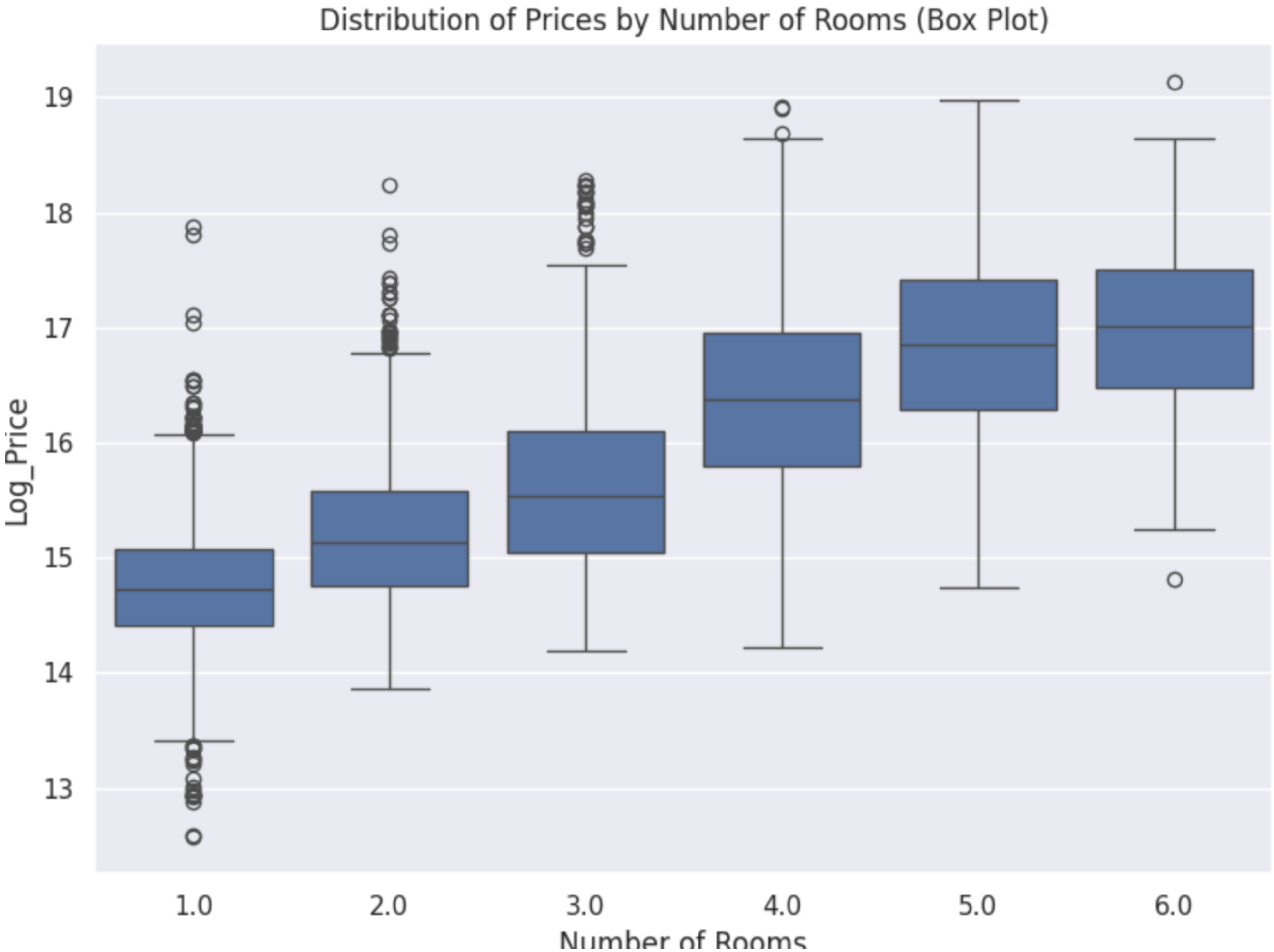
EDA

- Obviously, the average price is dependent from the nr of rooms

- And as well that log of the price normalizes the distribution

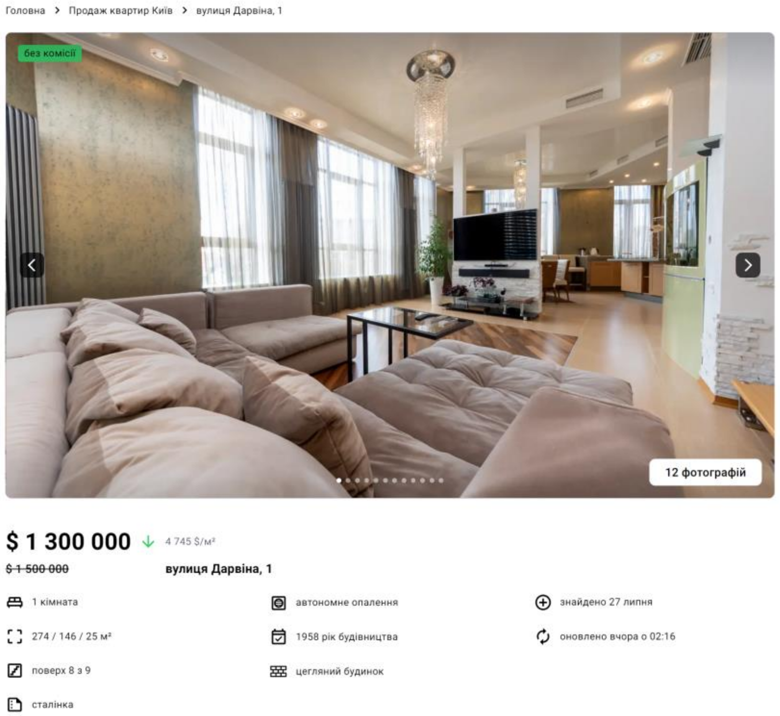

Positively Skewed Residuals
Normal Distribution
Log Transformation
Negatively Skewed Residuals
Normal Distribution
Exponential Transformation

https://discuss.boardinfinity.com/t/how-to-transform-the-data-to-make-normally-distributed/6716



EDA

- If we check the outliers in each group (grouped by rooms), we‘ll discover, that all of the apartments have a huge area. In general, over 50m2 for a 1r apartment is not typtical, and we have a certain amount of 1r apartments over 100m2. We can get rid of these.



Головна › Продаж квартир Київ › вулиця Дарвіна, 1

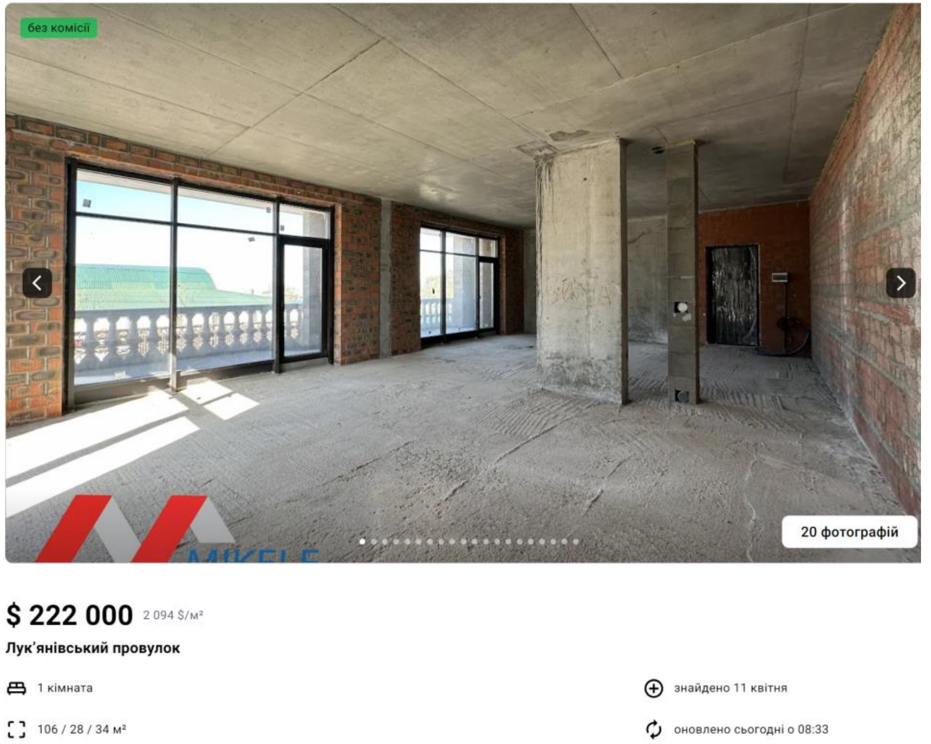
$ 1 300 000 ↓ 4 745 $/м²

~~$ 1 500 000~~ вулиця Дарвіна, 1

1 кімната
274 / 146 / 25 м²
поверх 8 з 9
сталінка

автономне опалення
1958 рік будівництва
цегляний будинок

знайдено 27 липня
оновлено вчора о 02:16



$ 222 000 2 094 $/м²

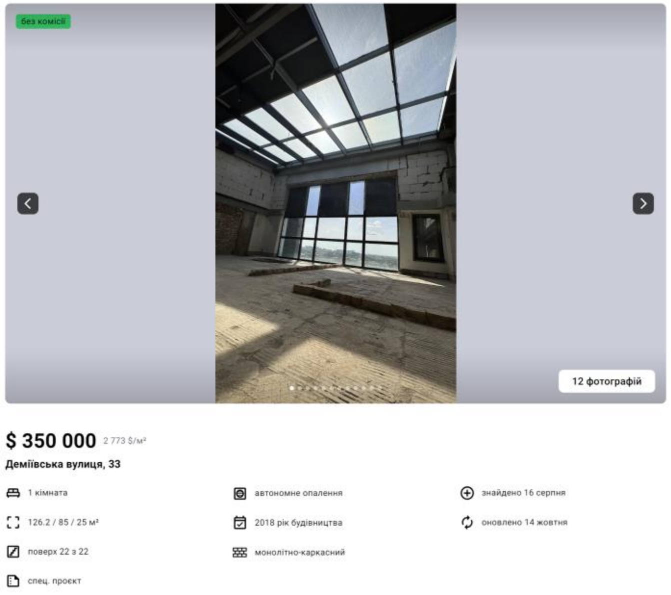
Лук’янівський провулок

1 кімната
106 / 28 / 34 м²
поверх 2 з 5

знайдено 11 квітня
оновлено сьогодні о 08:33

Опис

ЖК Liberty Residence - клубний будинок елітного
заміського будинку але в центрі міста, який об’



$ 350 000 2 773 $/м²

Деміївська вулиця, 33

1 кімната
126.2 / 85 / 25 м²
поверх 22 з 22
спец. проєкт

автономне опалення
2018 рік будівництва
монолітно-каркасний

знайдено 16 серпня
оновлено 14 жовтня

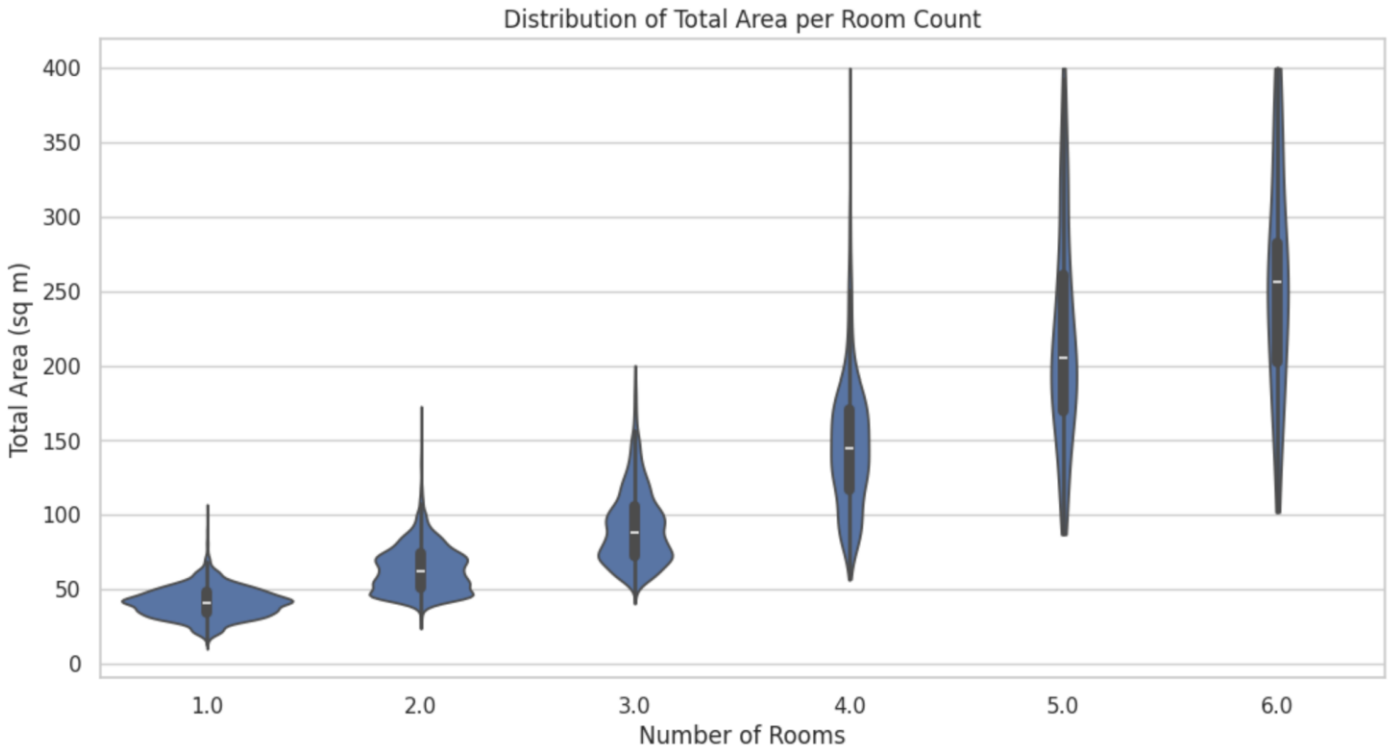
EDA

- Cleaning the data from outliers

```
price_outlier = df_num['Price per sqm'] > mean + 3*std
room1_outlier = (df_num['Rooms'] == 1) & (df_num['Area_total'] > 120)
room2_outlier = (df_num['Rooms'] == 2) & (df_num['Area_total'] > 180)
room3_outlier = (df_num['Rooms'] == 3) & (df_num['Area_total'] > 200)
room4_outlier = (df_num['Rooms'] == 4) & (df_num['Area_total'] > 200)
room4_outlier = (df_num['Area_total'] > 400)
```

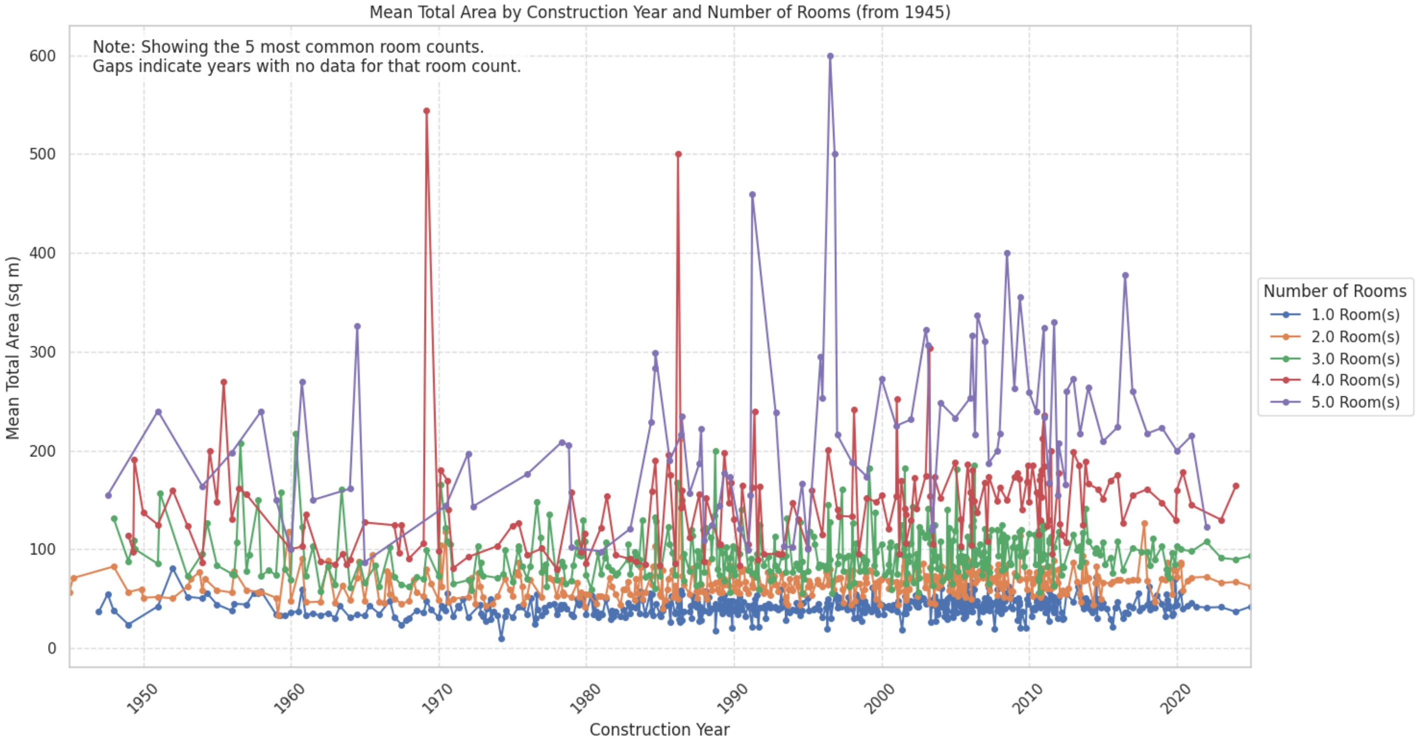
EDA

- Distribution of apartment‘s area depending on the rooms count



EDA

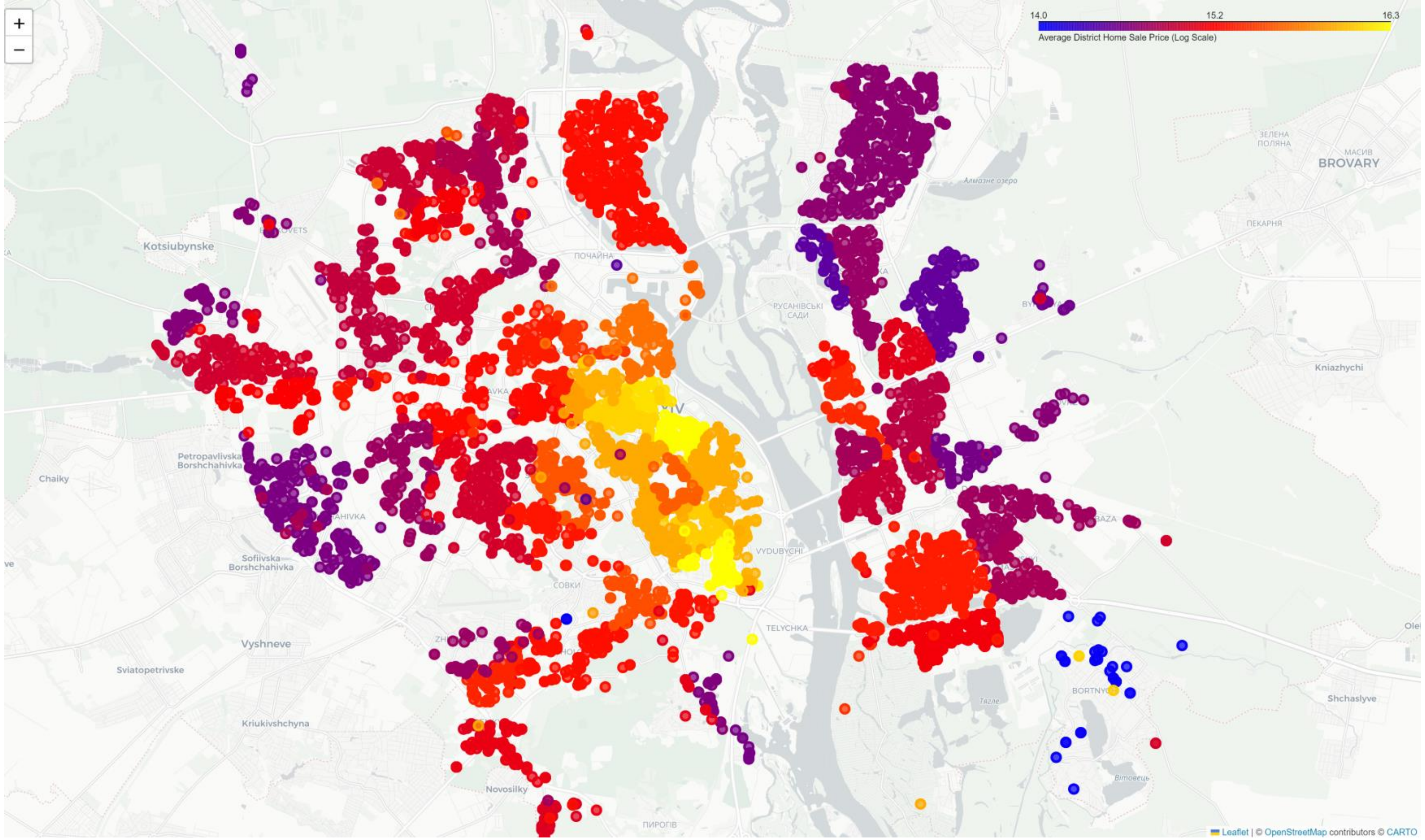
- Another stage of EDA was checking, if the age of the building influenced the apartments‘s area. It is obvious that 1 and 2r apartments remained about the same, however apartments with more rooms tend to get bigger in the recent years.



EDA

*див. Apartment_Prices_Geo_Features_Analysis.ipynb

- The check of the locations‘s influence over the apartment‘s price showed a clear dependency of the average price based on the district. It also clearly reveals where the center is.

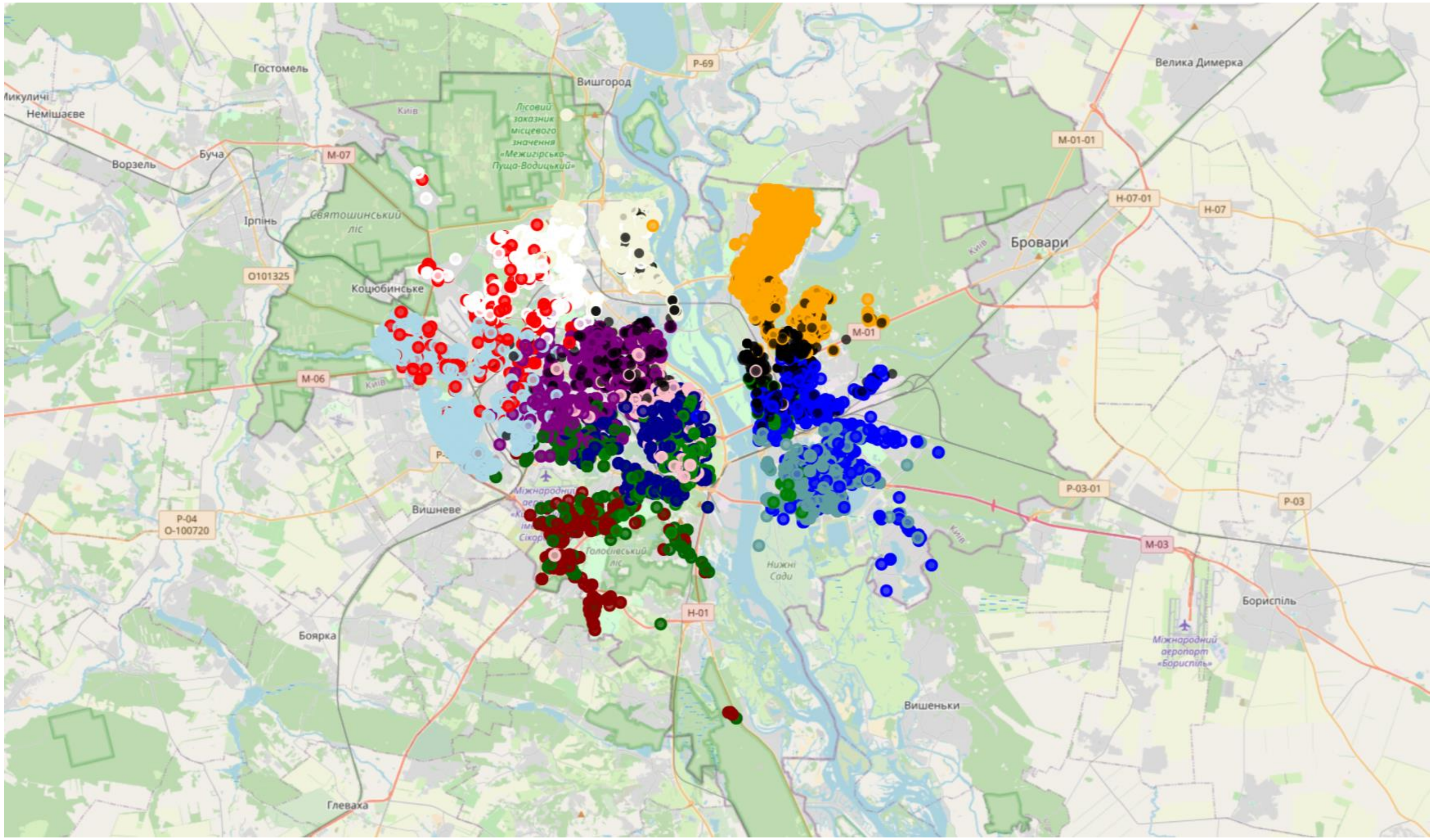


EDA

*див. Apartment_Prices_Geo_Features_Analysis.ipynb

- Unfortunately, clustering the data did not give good restults.

*probably, results could still be fine-tuneed though, looks like the nr of clusters is not meaningful. Maybe different algorithms for clusterization should be applied.



Feature Engineering

*see gpt_text_to_properties.py & run_gpt_features.py

- The ads normally contained the descrpition field where the advantages of the object are described. It was therefore interesting to try out to get extra features out of this text and turn them to categories or numbers

- Using chatgpt api, the following features from the ads were extracted:
  - Panoramic view (bool)
  - Balcony (bool)
  - Bathrooms (int)
  - Wardrobe room
  - Furnished
  - Appliences
  - Floor Heating
  - AC
  - Security
  - Renovation quality (int from 1 to 5)

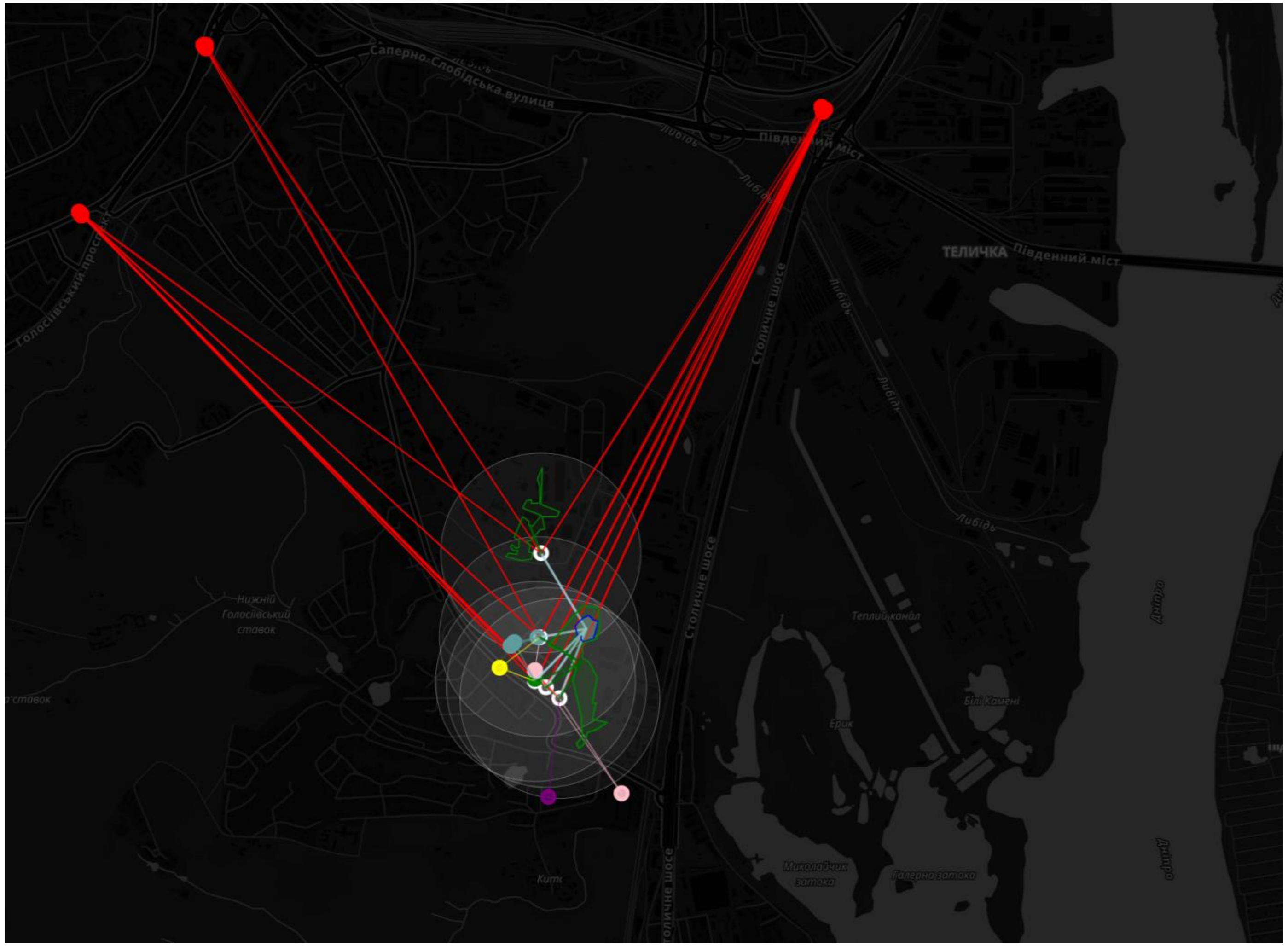
Feature Engineering

*див. Amenities_script.py & process_amenities.py

- Based on geo-location, the features of „proximity to the center“ and as well „proximity of the district to the center“ were extracted.

- Using openstreetmap api, following extra features were extracted:

  - Proximity to the closest subway
  - Sports amenities in r = 1km
  - Supermarkets in r = 300m
  - Schools in r = 300m
  - Kindergardens in r = 300m
  - Gastronomy in r = 300m
  - Public transport stops in r = 300m
  - Green features in r = 1km
  - Water features in r = 1km



Перевірка правильності зібраних геофіч на конкретних прикладах

Baseline & Experiments

*див. Experiments1.ipynb

- *Metrics: MAE*
- *Validation: cross-validation on 5 folds because the data set is quite small*

- The best baseline without using ML was grouping the apartments by districts and rooms count and assigning the average price per group.
  MAE = 1.456 mln hrn – average price 5.7 mln

- Linear Regression experiments. Starting with the main features only (the scraped ones), (тільки заскрейплені з сайту), however adding the proximity to the center as well as regularization, as well as log – did not help to overcome baseline.

- First score improvement happend after adding text features.
  MAE = 1.400 mln hrn

- Next imporvement happend after adding the geo-based features.
  MAE = 1.300 mln hrn

- Decision tree improved the score a little, but insignificantly

Baseline & Experiments

*див. Experiments2.ipynb

- *Метрика: MAE*
- *Валідація: кросс-валідація на 5 фолдах, тому що досить мало даних*

- Considering that adding new features did not improve the score enough, I did another iteration of the data cleansing: IQR-based outlier detection.
- After that, **baseline** showed: **MAE = 975 000 hrn** (avrg price = 4371 000грн).
- Linear Regression: Train MAE = 892 000 , **Val MAE = 901 000**
- Random Forest: Train MAE = 442 000, **Val MAE = 706 000**

*see. rf_r_seach & rf_final.py

- Light gbm boosting: Train MAE = 311 000, **Val MAE = 642 000**

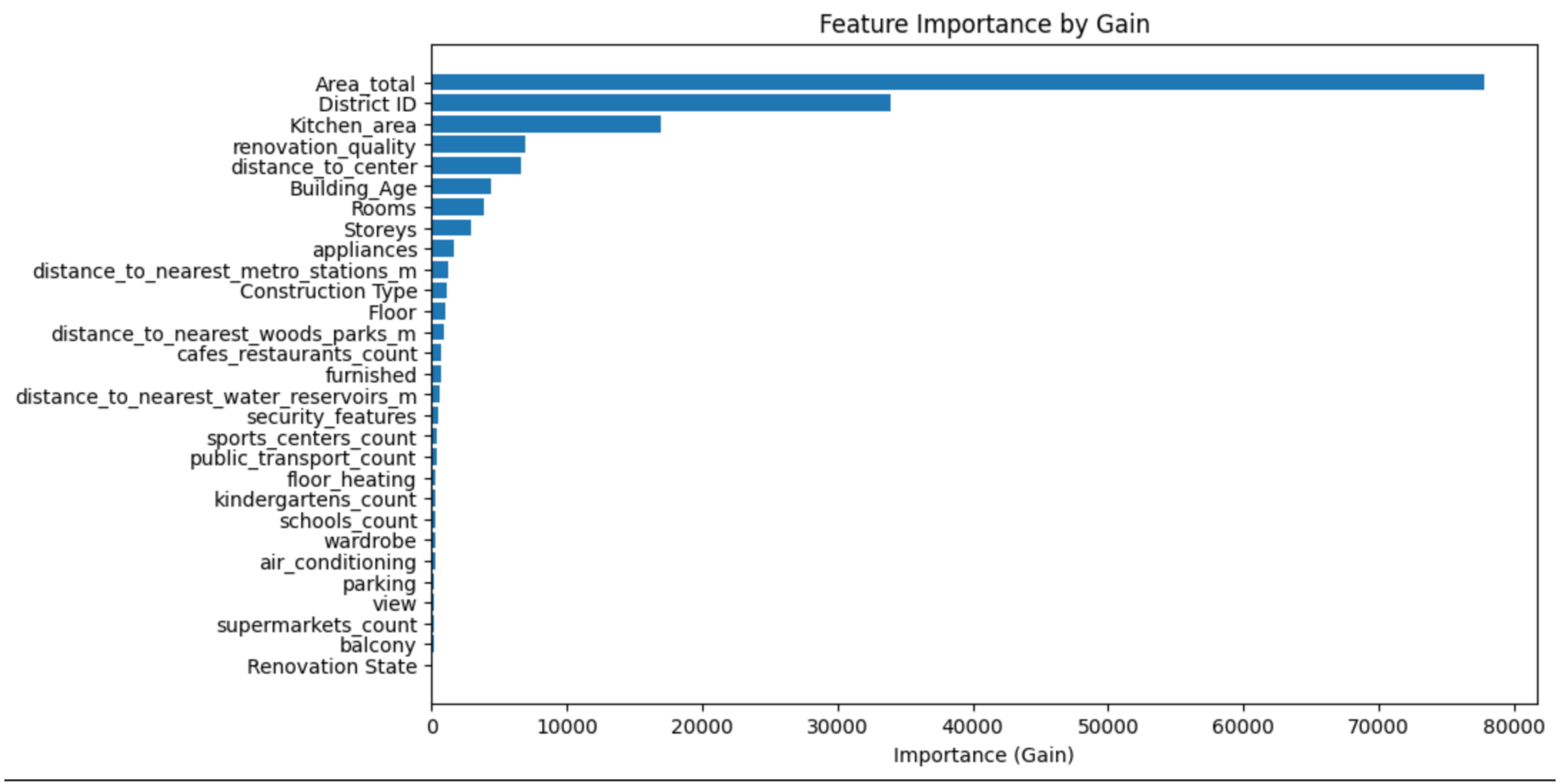
*see. Boosting & analysis.inypb

Lets try to find out where the model is being mistaken and why

Feature Importance

*див. Experiments2.ipynb

- Feature importance (for the light gbm) seems meaningful, although it was surprising, how little importance such features as access to infrastrure or balcony have)

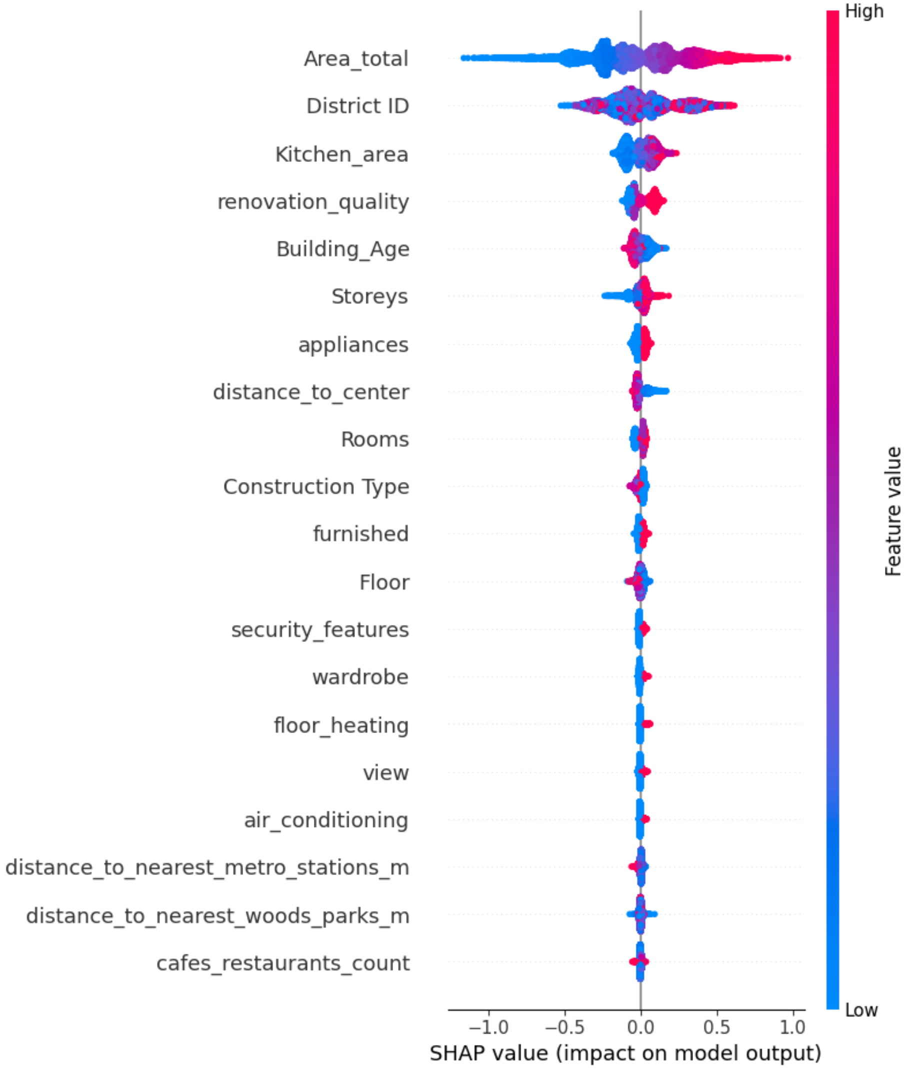


Feature importance by light GBM

Feature Importance

*див. Experiments2.ipynb

- Shaply analysis gives slightly different results, though the split into important and unimportant features is quite similar.

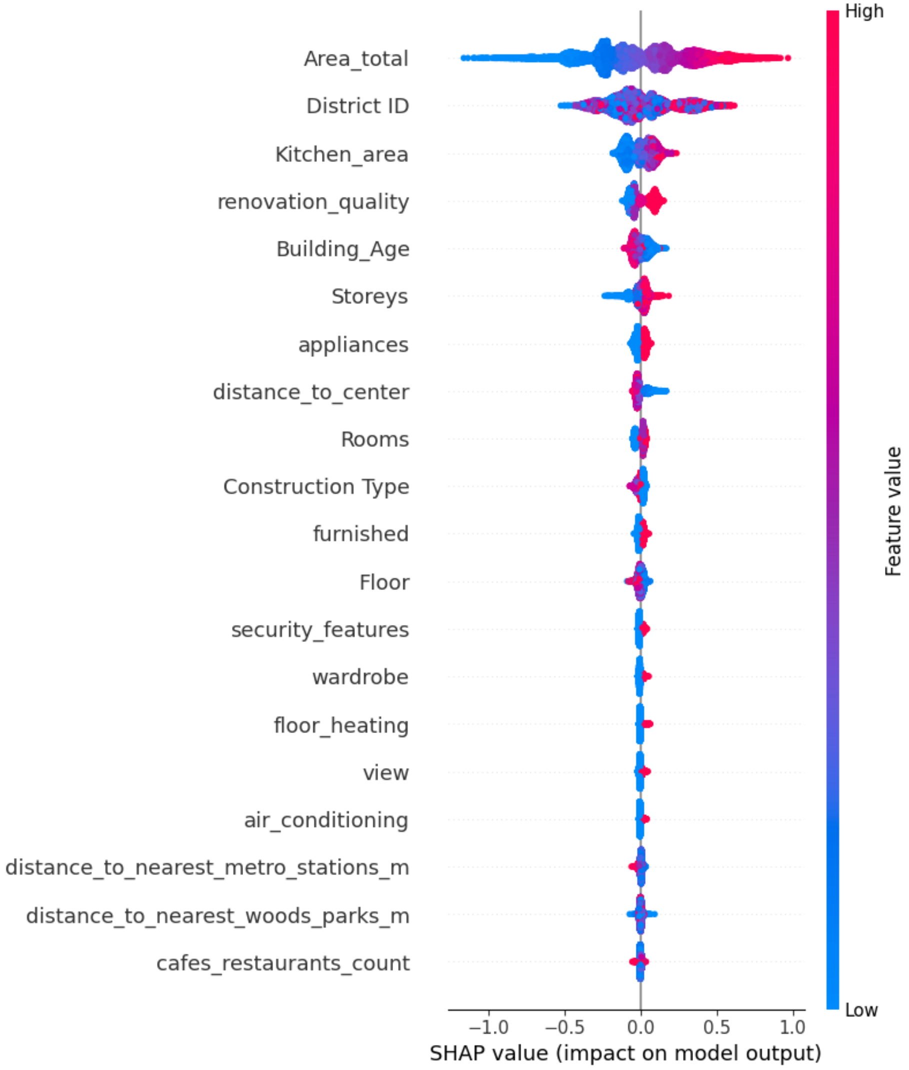


Feature importance with SHAP

Feature Importance

*див. Boosting & analysis.ipynb

- After these two analysis I deleted some of the least important features, reducing the model‘s complexity, and this helped improving result from 642 000 to 636 000.

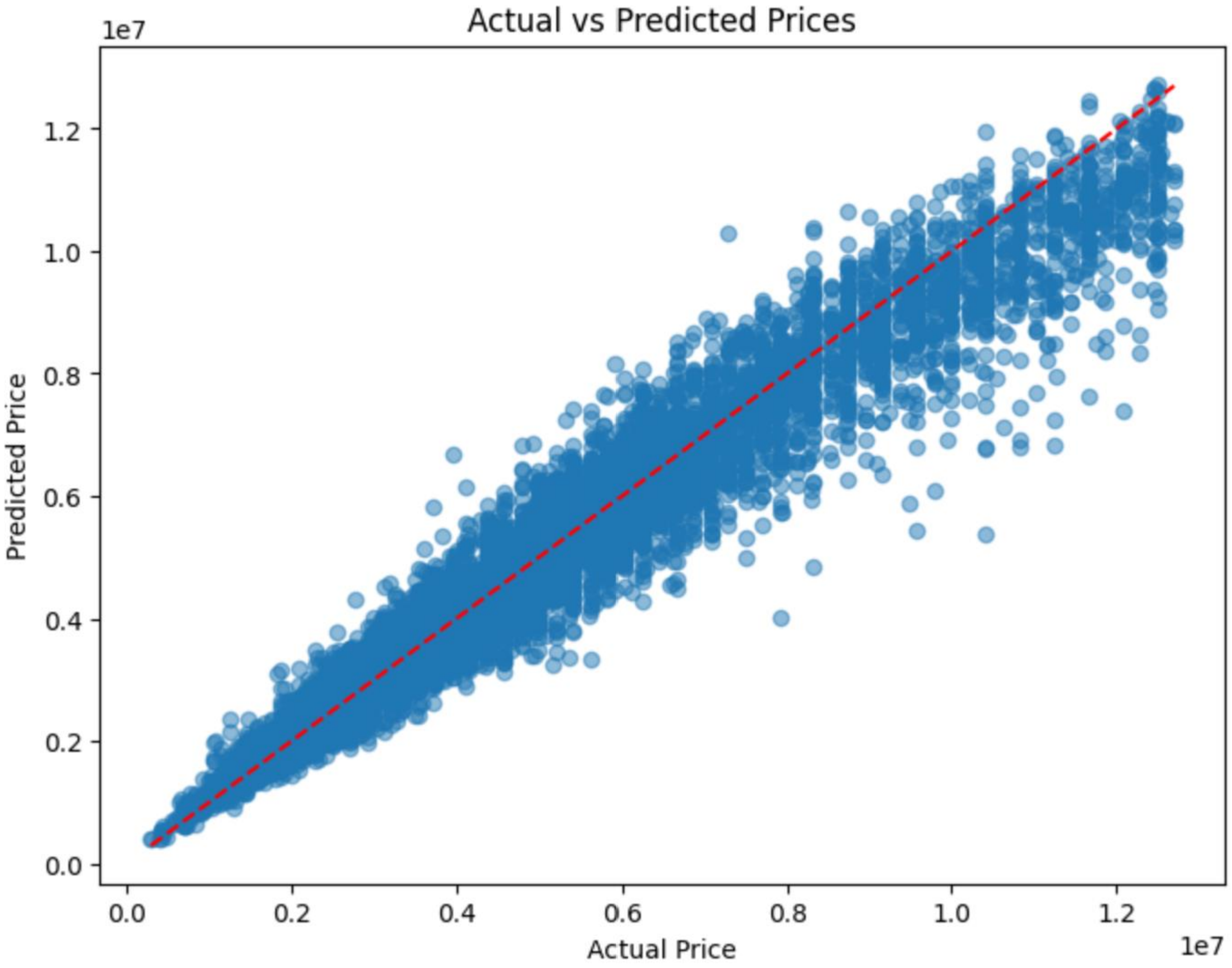


Feature importance with SHAP

Performance Analysis

*див. Boosting & analysis.ipynb

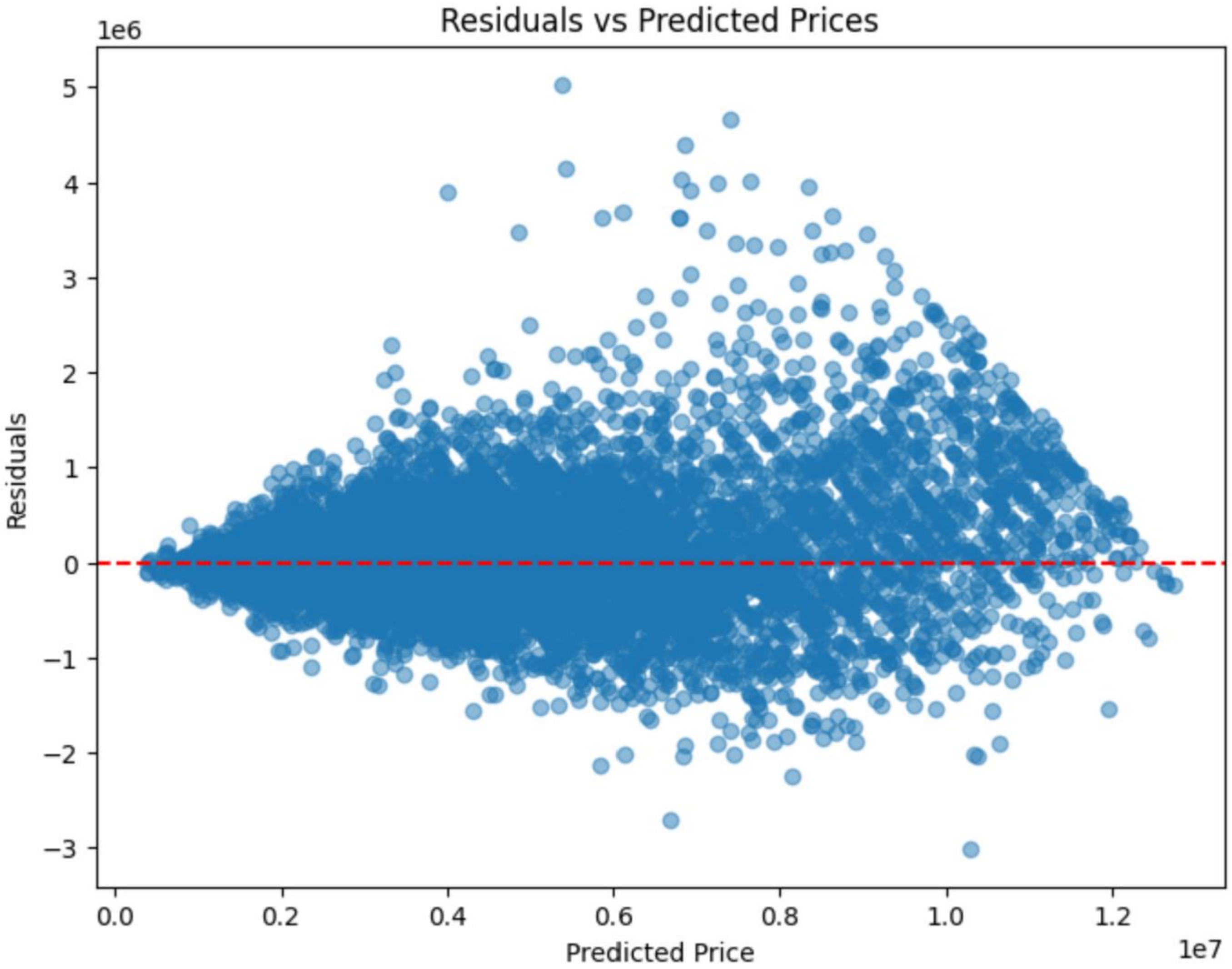
- If we compare the predicted vs real prices, we can see, that the model reduces the prediction for more expensive apartments. It might be so due the minority of such apartments in the dataset, and the model could not learn well enough to esimate these.



Performance Analysis

*див. Boosting & analysis.ipynb

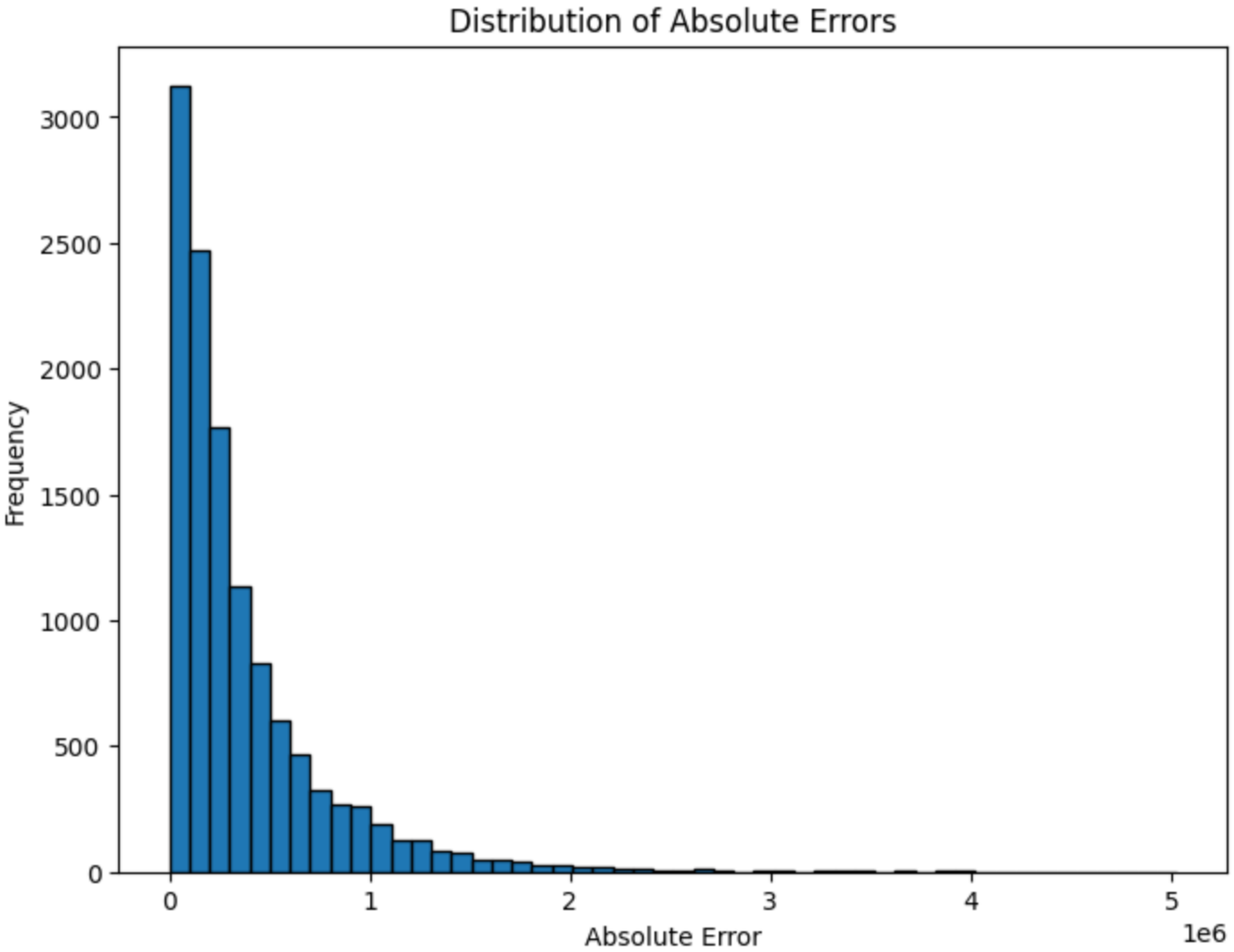
- The residuals plot also confirms the previous conclusion. Residuals are about equally distributed around 0, but positive residuals are a little further along „y“ axis and are more numerous.



Performance Analysis

*див. Boosting & analysis.ipynb

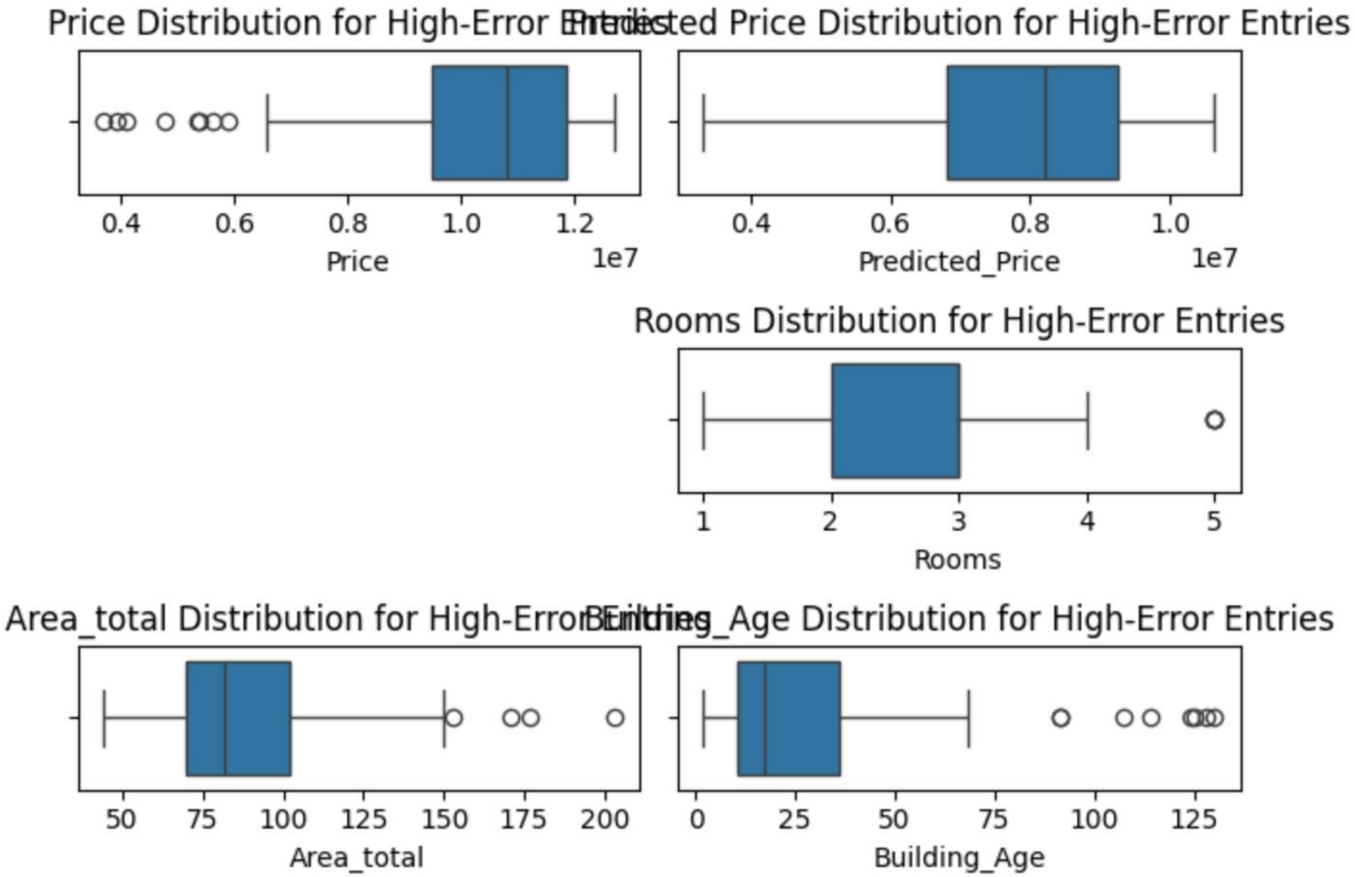
- Distirubtion of the absolute errors shows the greater amount of small errors, but some huge errors are present as well. Let us look at mistakes over 2 mln hrn.



Performance Analysis

*див. Boosting & analysis.ipynb

- Here we can see that the main large mistakes happend for generally more expensive apartments.
- We can see, how „skewed“ is the error for the expensive apartments. (price distributions vs. Predicted price distribution)
- Most of these expensive apartments are 2-3rooms (here is potentially the reason for the model‘s errors – as the 2 and 3r apartments should not generally be the most expensive)
- It is also obvíous that most of these expensive apartments are in the new developments.



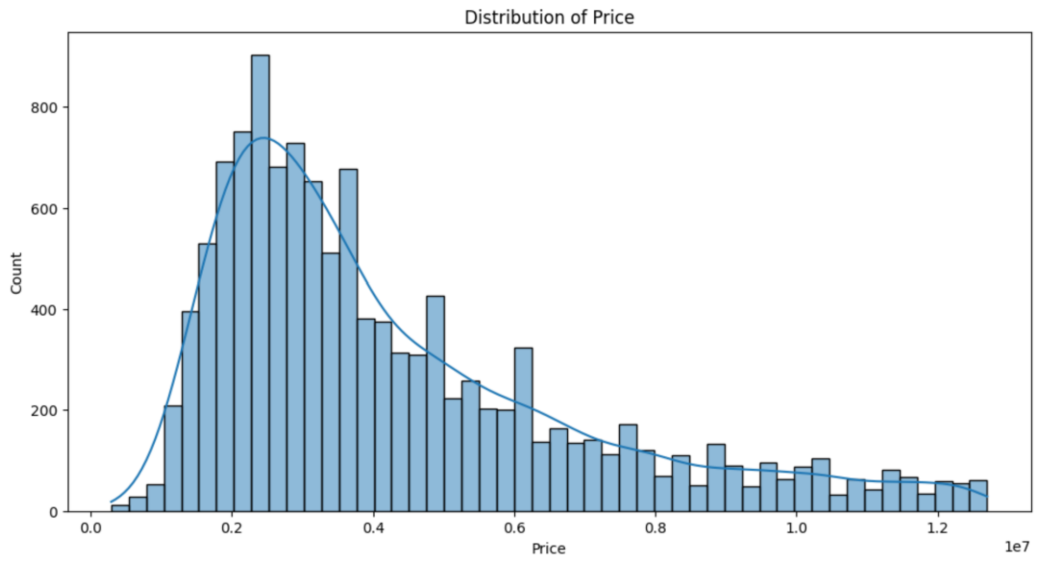
Conclusions

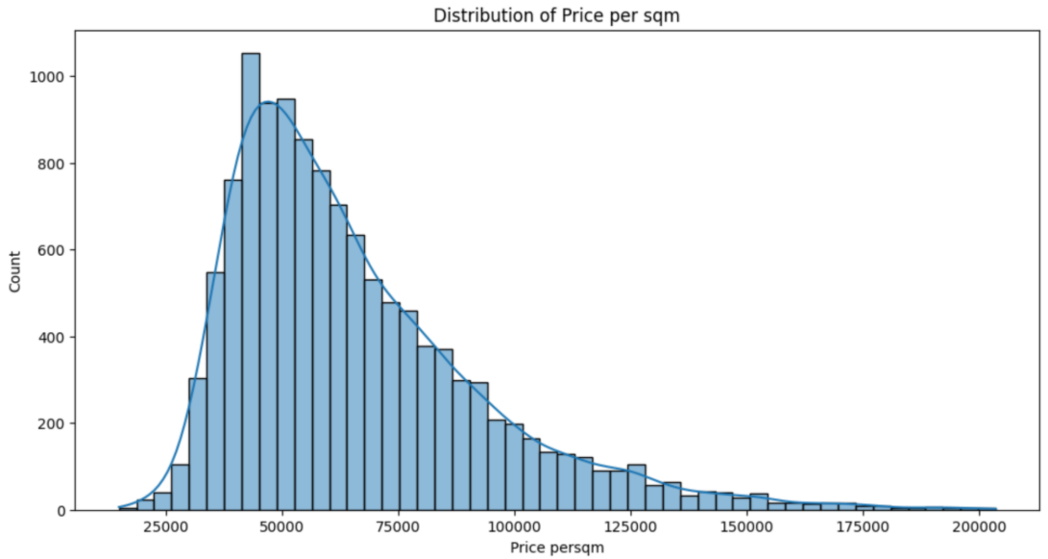
*see. Boosting & analysis.ipynb

- For further improvement, we definitely need to look at 2 and 3k apartments in the more expensive segment and try to understand why the model is wrong in them.
- Or at least try to exclude such data and check the model error then.
- From the impact of the features, it is obvious that such a feature as renovation(True/False) does not affect the result, but the quality of the renovation - from the text description - already has a certain significance.
- It is a bit strange that the distance to the subway did not seem to have a significant impact on the results. It is also possible to more carefully investigate the methods of clustering, because in such a problem it seems that this has the potential to help the model.
- Also applying ensembles might be beneficial.

Update after feedback

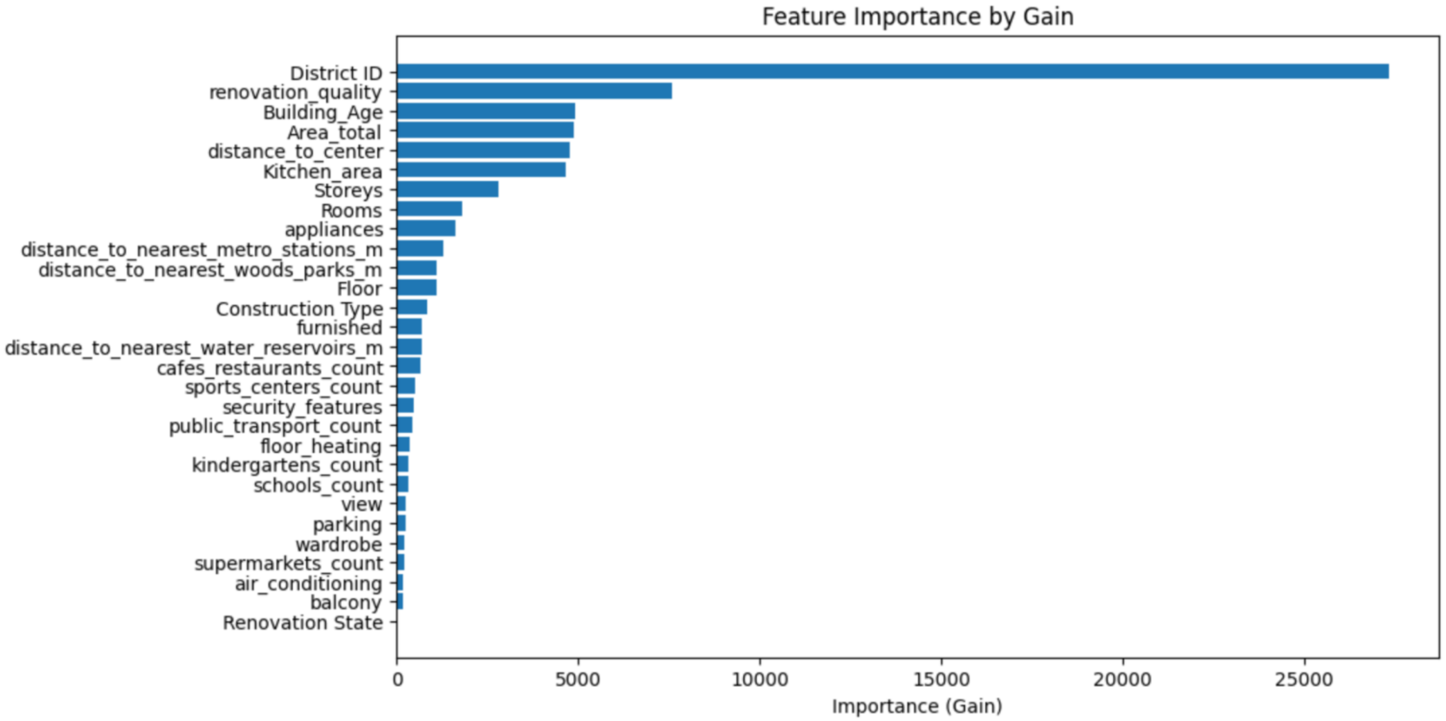
- On the feature importance graphics it was obvious, that the m2 is by far the most important feature. In order to make it a bit less outstanding, we can use price/m2 instead of price to help the model meet more accurate estimates. Obviously, distribution with this new features looks more normalized already.



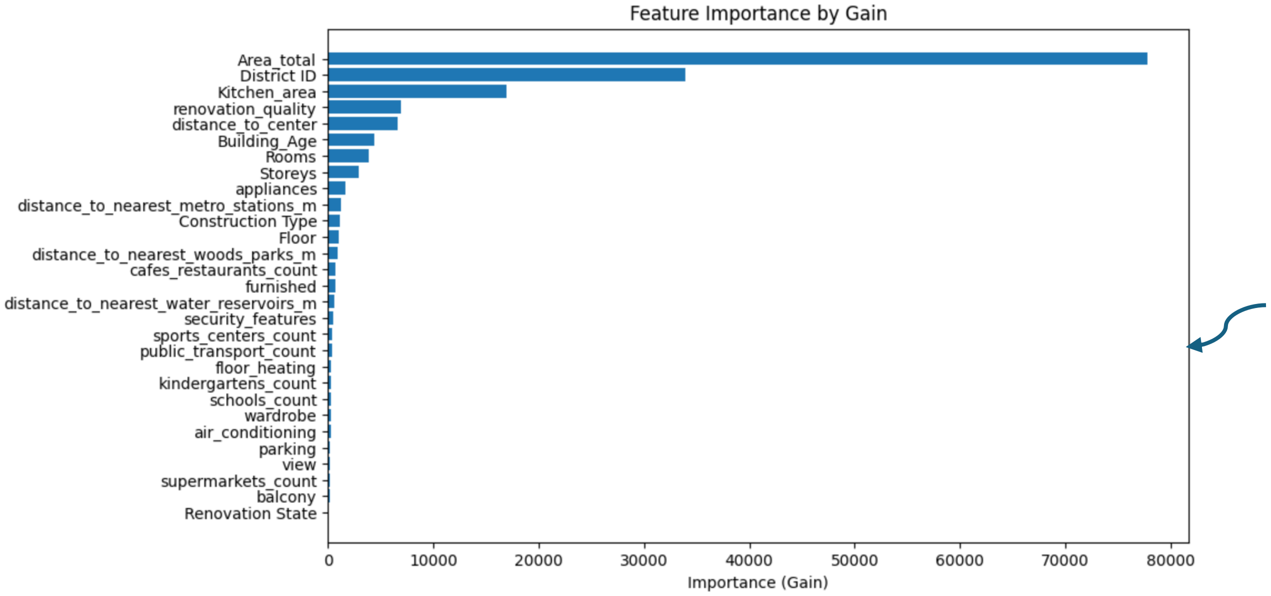


Update

- Stratification did not improve the result
- Feature importance seems better now.



Новий

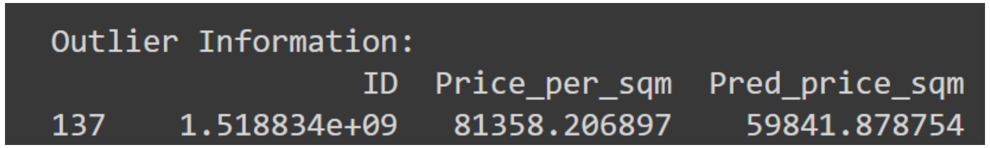


старий

High errors

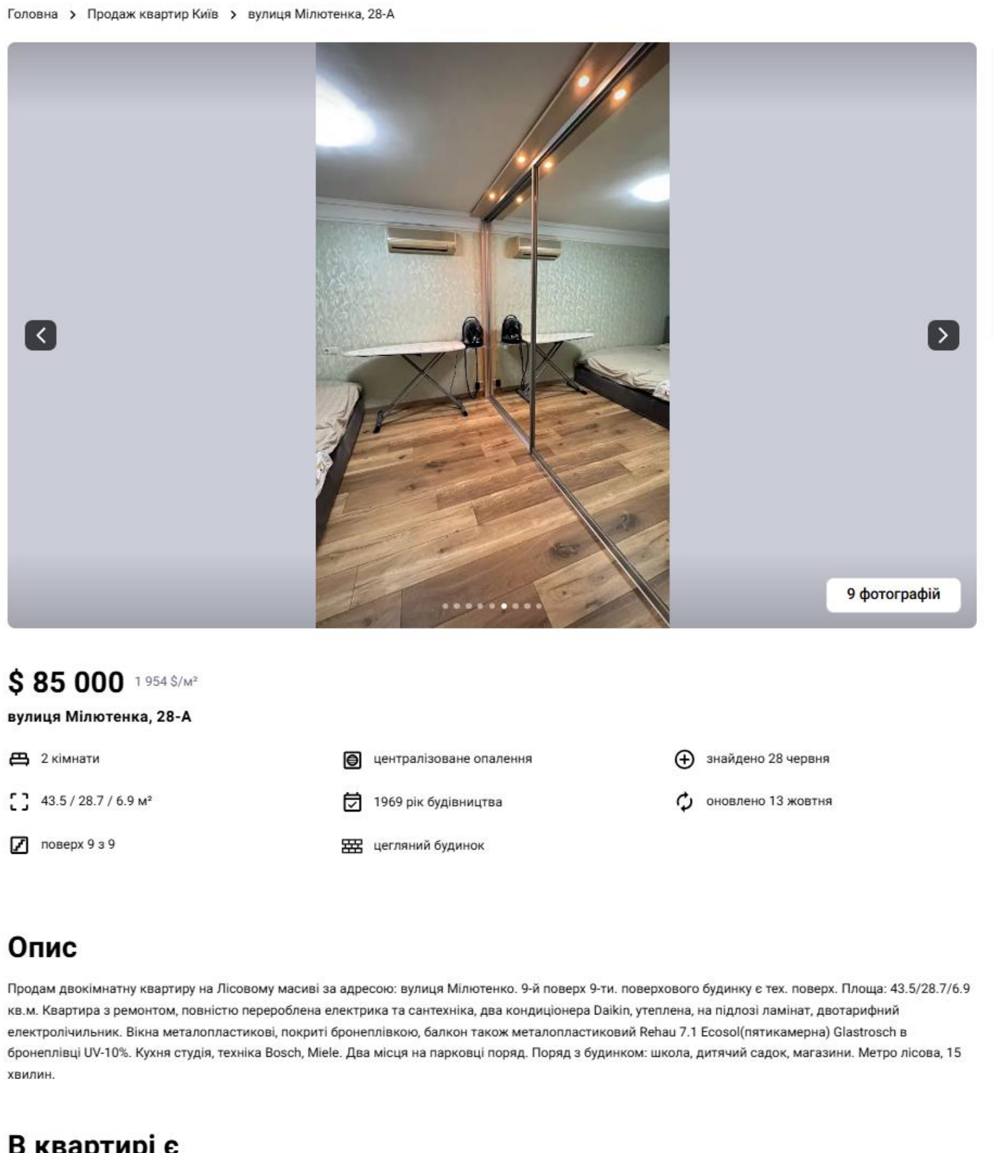
*див. Boosting & analysis.ipynb

- Lets have a close-up on apartments with the largest errors.



```
Outlier Information:
                ID  Price_per_sqm  Pred_price_sqm
137   1.518834e+09   81358.206897    59841.878754
```

- Avrg price for 10 closest neighbors with the same m2 and same rooms cound is 43651.
- 10 closest neighbours- 44800.
- In general, the prediction of the model (higher than average because of renovation quality and appliences) seems logical.



Головна › Продаж квартир Київ › вулиця Мілютенка, 28-А

9 фотографій

$ 85 000 1 954 $/м²

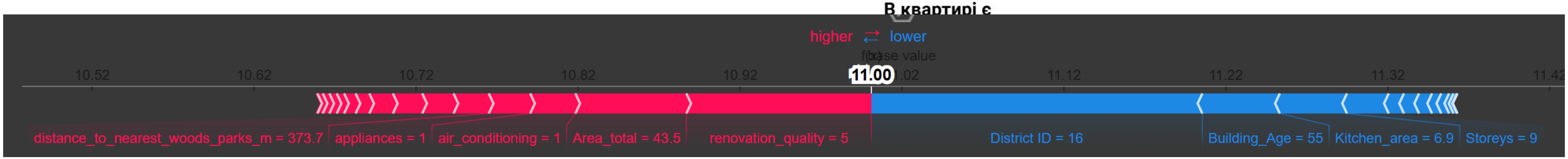
вулиця Мілютенка, 28-А

2 кімнати
43.5 / 28.7 / 6.9 м²
поверх 9 з 9
централізоване опалення
1969 рік будівництва
цегляний будинок
знайдено 28 червня
оновлено 13 жовтня

Опис

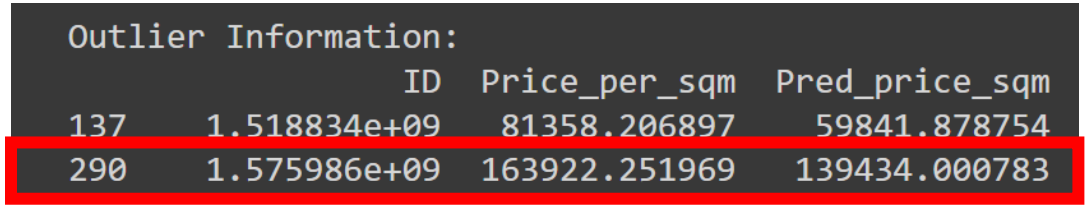
Продам двокімнатну квартиру на Лісовому масиві за адресою: вулиця Мілютенко. 9-й поверх 9-ти. поверхового будинку є тех. поверх. Площа: 43.5/28.7/6.9 кв.м. Квартира з ремонтом, повністю переоблена електрика та сантехніка, два кондиціонера Daikin, утеплена, на підлозі ламінат, двотарифний електролічильник. Вікна металопластикові, покриті бронеплівкою, балкон також металопластиковий Rehau 7.1 Ecosol(пятикамерна) Glastrosch в бронеплівці UV-10%. Кухня студія, техніка Bosch, Miele. Два місця на парковці поряд. Поряд з будинком: школа, дитячий садок, магазини. Метро лісова, 15 хвилин.

В квартирі є



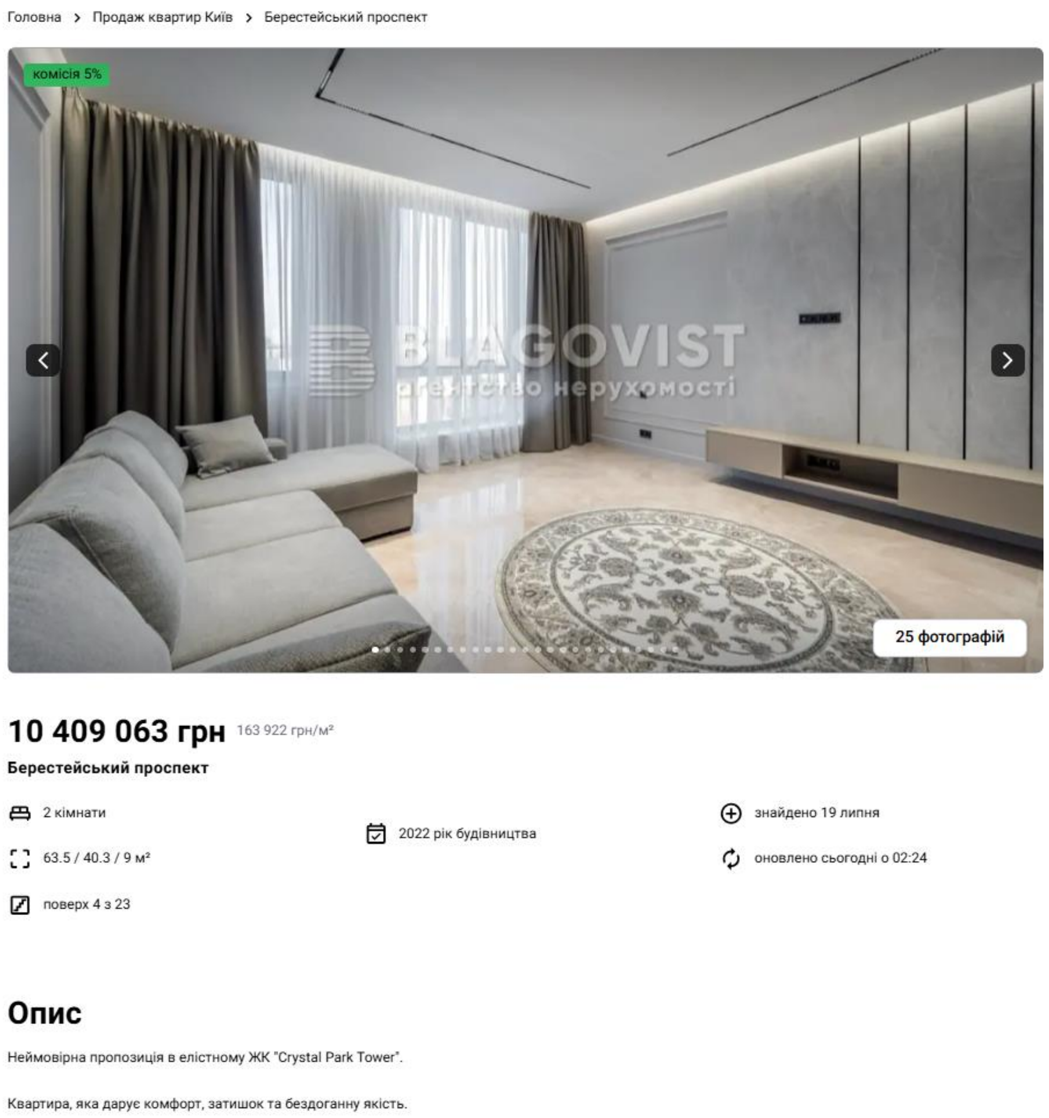
High errors

*див. Boosting & analysis.ipynb



```
  Outlier Information:
                  ID  Price_per_sqm  Pred_price_sqm
  137   1.518834e+09   81358.206897    59841.878754
  290   1.575986e+09  163922.251969   139434.000783
```

- Avrg price for 10 closest neighbors with the same m2 and same rooms cound is – 79000.
- 10 closest neighbours -- 54723.
- Shap analysis demonstrates why the price for this particualr apartment is already higher than the avrg neighbour with same parameters.



Головна › Продаж квартир Київ › Берестейський проспект

10 409 063 грн 163 922 грн/м²

Берестейський проспект

2 кімнати

63.5 / 40.3 / 9 м²

поверх 4 з 23

2022 рік будівництва

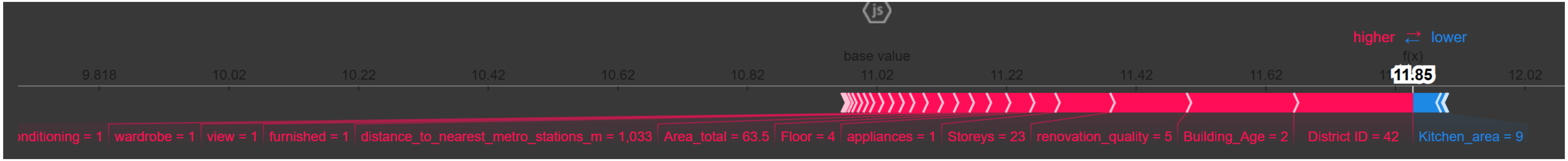
знайдено 19 липня

оновлено сьогодні о 02:24

Опис

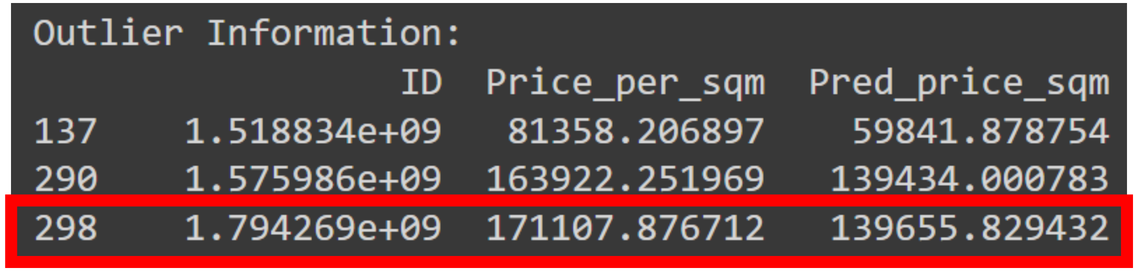
Неймовірна пропозиція в елістному ЖК "Crystal Park Tower".

Квартира, яка дарує комфорт, затишок та бездоганну якість.



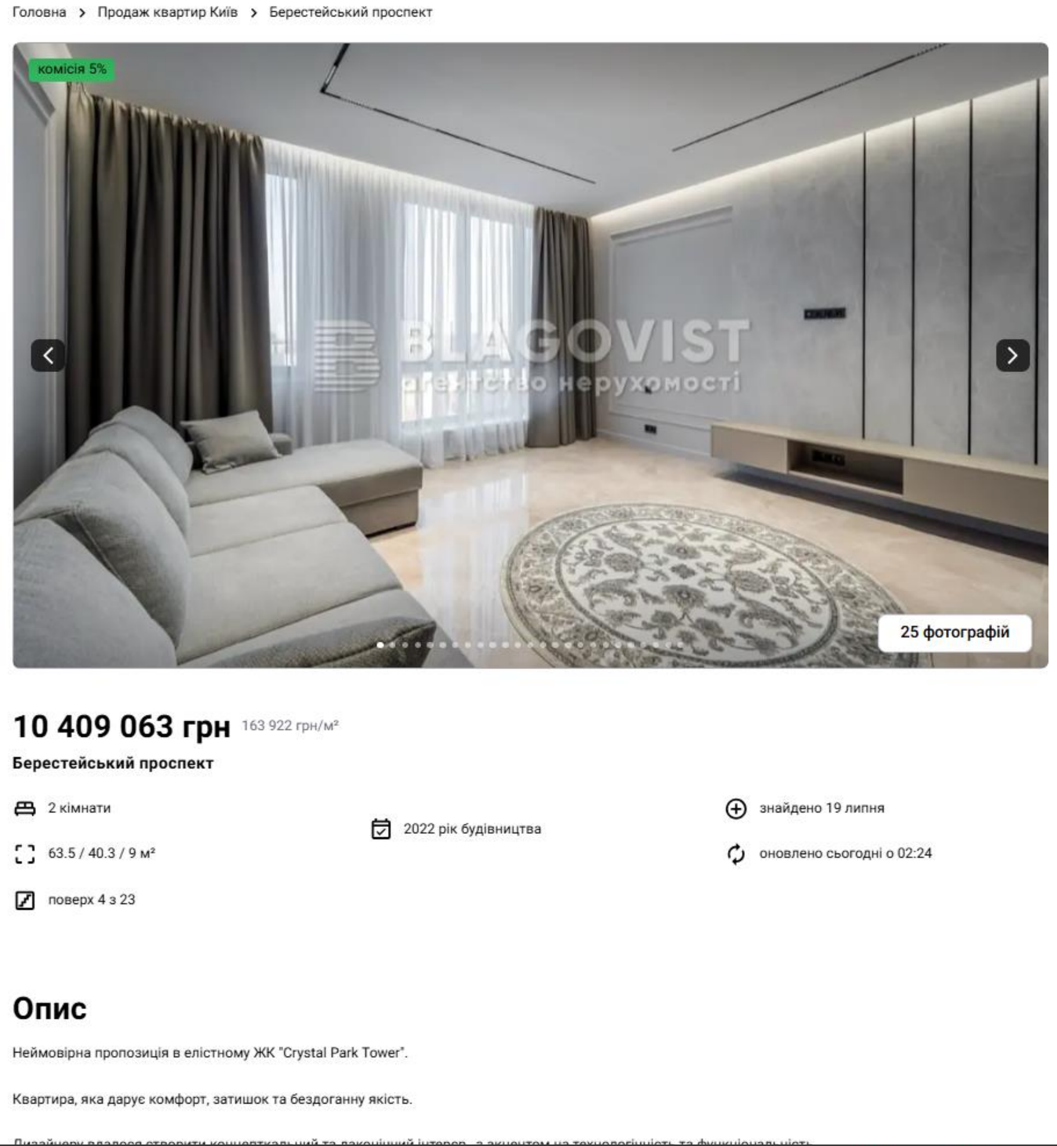
High errors

*див. Boosting & analysis.ipynb



```
Outlier Information:
                   ID   Price_per_sqm   Pred_price_sqm
137      1.518834e+09    81358.206897     59841.878754
290      1.575986e+09   163922.251969    139434.000783
298      1.794269e+09   171107.876712    139655.829432
```

- Avrg price for 10 closest neighbors with the same m2 and same rooms cound is – 68728.
- Shap analysis demonstrates why the price for this particualr apartment is already higher than the avrg neighbour with same parameters.
- *Last two apartments are from a gated community, it can be that this is also a feature increasing the price. It would be good to include this feature into the model. Probably also grouping for the same developments would help.
- *Certain features – as constant energy supply, or reinforced windows – might be significant as well, but this hypothesis needs testing.



Головна › Продаж квартир Київ › Берестейський проспект
комісія 5%
BLAGOVIST агентство нерухомості
25 фотографій

10 409 063 грн 163 922 грн/м²

Берестейський проспект

2 кімнати

63.5 / 40.3 / 9 м²

поверх 4 з 23

2022 рік будівництва

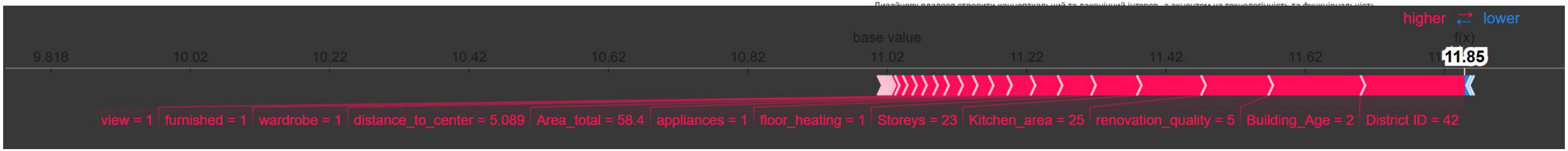
знайдено 19 липня

оновлено сьогодні о 02:24

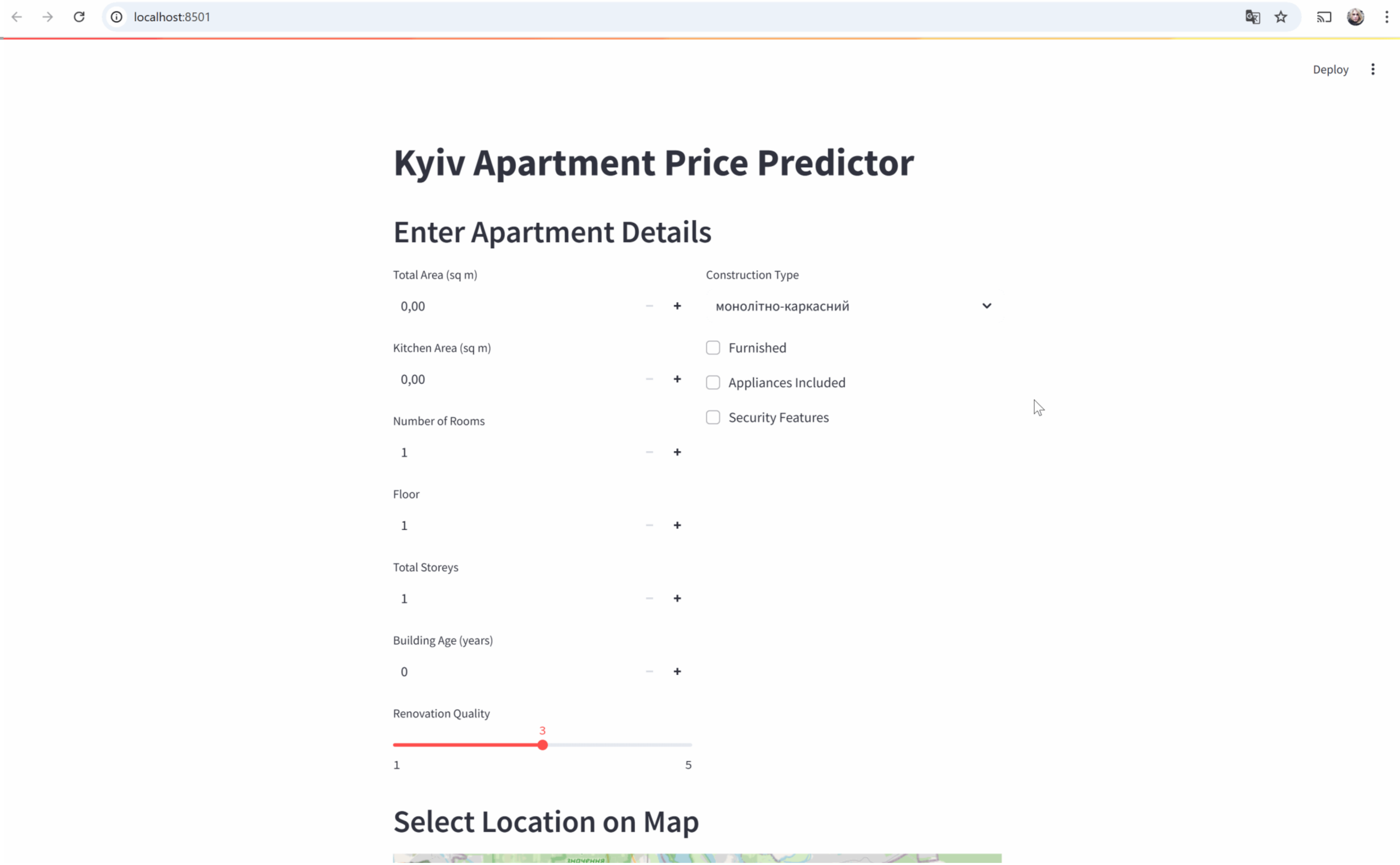
Опис

Неймовірна пропозиція в елістному ЖК "Crystal Park Tower".

Квартира, яка дарує комфорт, затишок та бездоганну якість.



Demo



Kyiv Apartment Price Predictor

Enter Apartment Details

Total Area (sq m)
0,00 − +

Kitchen Area (sq m)
0,00 − +

Number of Rooms
1 − +

Floor
1 − +

Total Storeys
1 − +

Building Age (years)
0 − +

Renovation Quality
3
1 5

Construction Type
монолітно-каркасний

□ Furnished

□ Appliances Included

□ Security Features

Select Location on Map

GIT HUB LINK: https://github.com/osintseva/ApartmentPrices/

THE END

Dear Rehubibies, if you‘d like an overview of different models and the way they work, that‘s a topic for another TA ;)