Estate rental market in Prague – price prediction and trend analysis

Tomasz Bialy

01/09/2021

Business problem

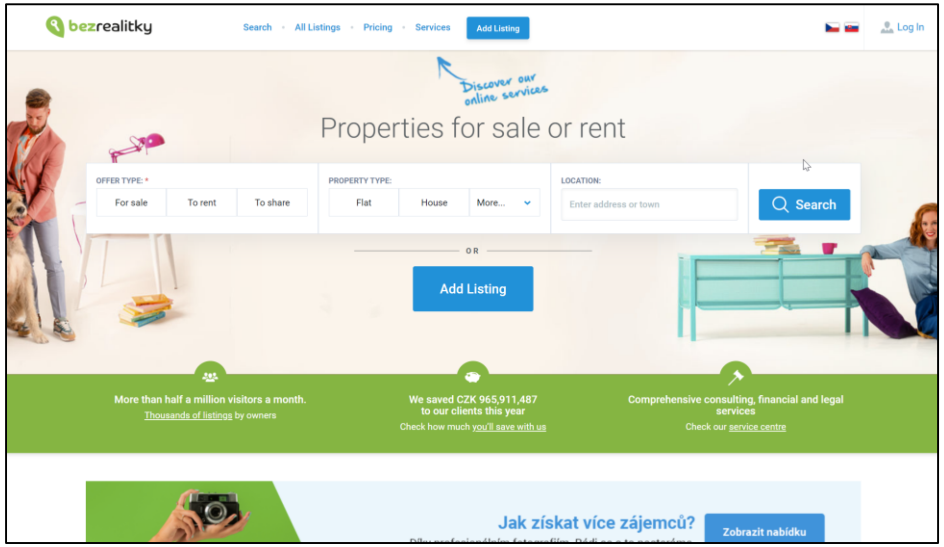
Renting an apartment is always a difficult and stressing process that often seems impulsive or risky. If the renting itself is already hard, doing it in Prague doesn’t make it any easier. With increasing city regulations, growing tourism and population, renting an apartment in Prague leaves many desperate for the first opportunity available, becoming vulnerable to scams and overpriced contracts.

In this project, I will go through a process of gathering the data, data cleaning, data exploration, feature engineering and finally applying a machine learning model that predicts rental prices of city’s apartments based on chosen features. At the end, I will visualize the apartments price change over some period of time, which could help to understand the market trends.

Prague

Gathering the data

The dataset used for the project was extracted from *Bezrealitky.cz* - the largest webpage in Czech Republic that host the ads for housing rental directly from the owners without agencies. I used the same portal to find the apartment where I currently live. For that purpose I used my own script written in Python that is able to extract 23 features from every ad listed in the webpage. A total 2605 apartment offerings have been collected during the data collection.

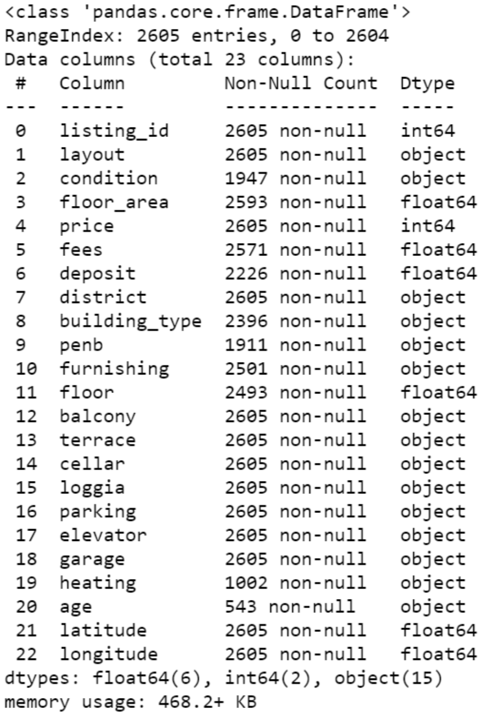


More than half a million visitors a month.
Thousands of listings by owners

We saved CZK 965,911,487 to our clients this year
Check how much you'll save with us

Comprehensive consulting, financial and legal services
Check our service centre

Bezrealitky.cz portal



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2605 entries, 0 to 2604
Data columns (total 23 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   listing_id     2605 non-null   int64  
 1   layout         2605 non-null   object 
 2   condition      1947 non-null   object 
 3   floor_area     2593 non-null   float64
 4   price          2605 non-null   int64  
 5   fees           2571 non-null   float64
 6   deposit        2226 non-null   float64
 7   district       2605 non-null   object 
 8   building_type  2396 non-null   object 
 9   penb           1911 non-null   object 
 10  furnishing     2501 non-null   object 
 11  floor          2493 non-null   float64
 12  balcony        2605 non-null   object 
 13  terrace        2605 non-null   object 
 14  cellar         2605 non-null   object 
 15  loggia         2605 non-null   object 
 16  parking        2605 non-null   object 
 17  elevator       2605 non-null   object 
 18  garage         2605 non-null   object 
 19  heating        1002 non-null   object 
 20  age            543 non-null    object 
 21  latitude       2605 non-null   float64
 22  longitude      2605 non-null   float64
dtypes: float64(6), int64(2), object(15)
memory usage: 468.2+ KB
```

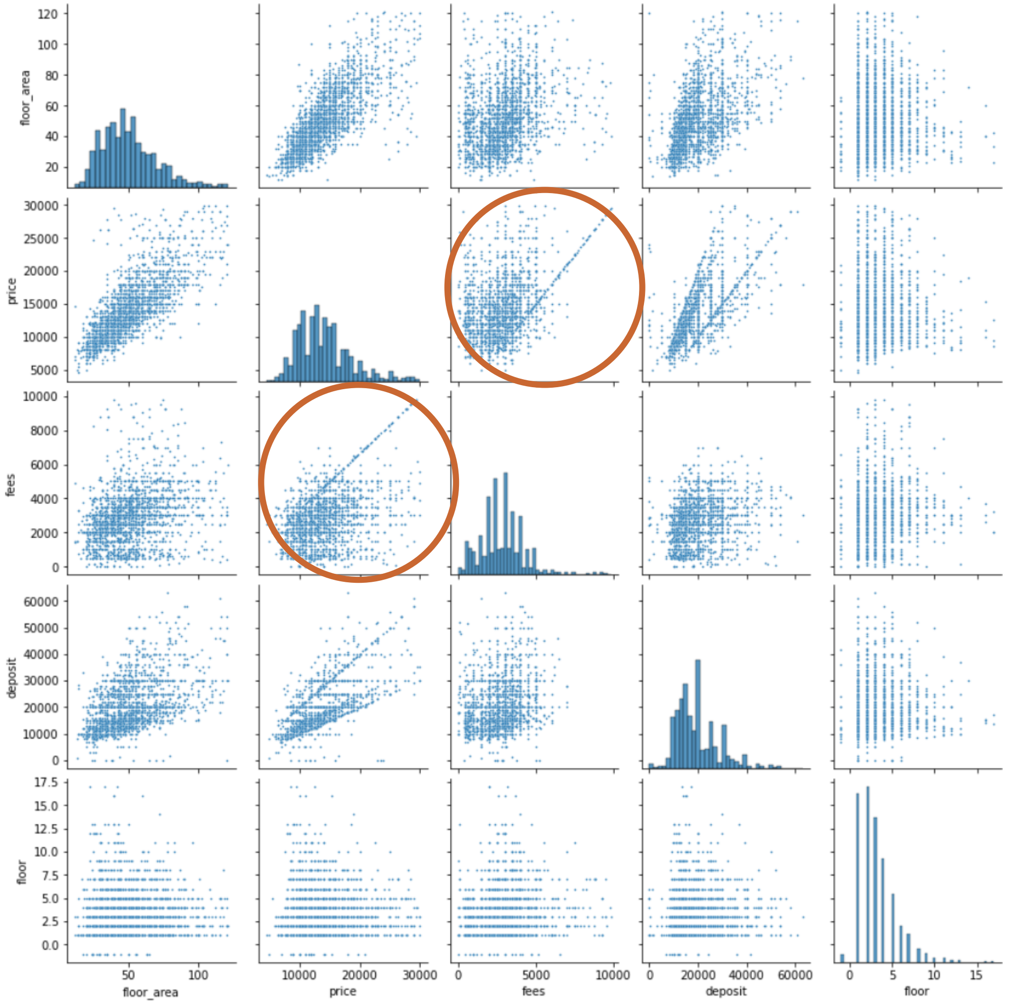
Obtained data features

Data cleaning

At the very beginning I am dropping the outliers or apartments with values that that seems unreal:

- Apartments below 10m^2
- Apartments over 125m^2
- Apartments cheaper than 3,000 CZK
- Apartments more expensive than 30,000 CZK
- Apartments with fees below 500 CZK

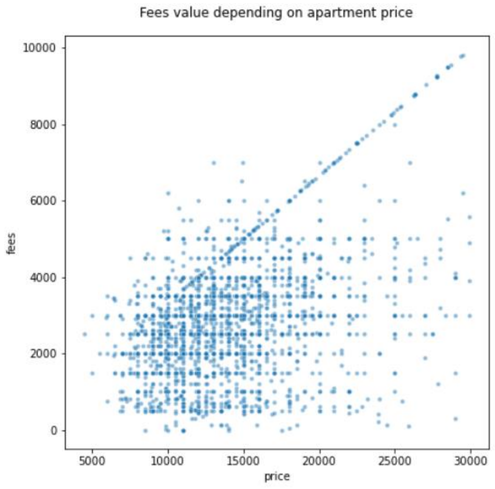
Data cleaning



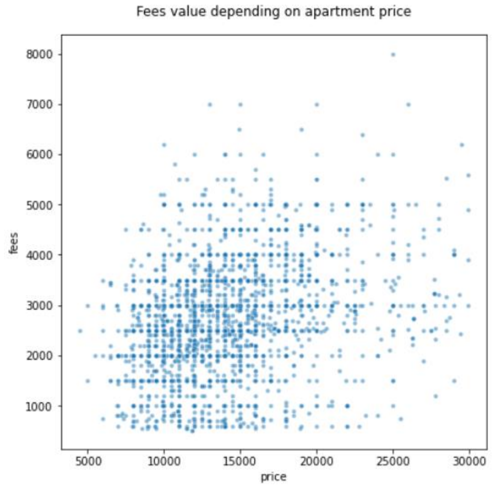
Pairplots

While plotting the pairplots, I have noticed strange linear dependency between price and the fees of the apartment. It seems that the apartment listing portal is putting the fees values equal 1/3 of the apartment price in case the user did not listed it.

I decided to replace that wrongly estimated fees values with a my own polynomial regression model which is based on the size of the apartment (floor area).

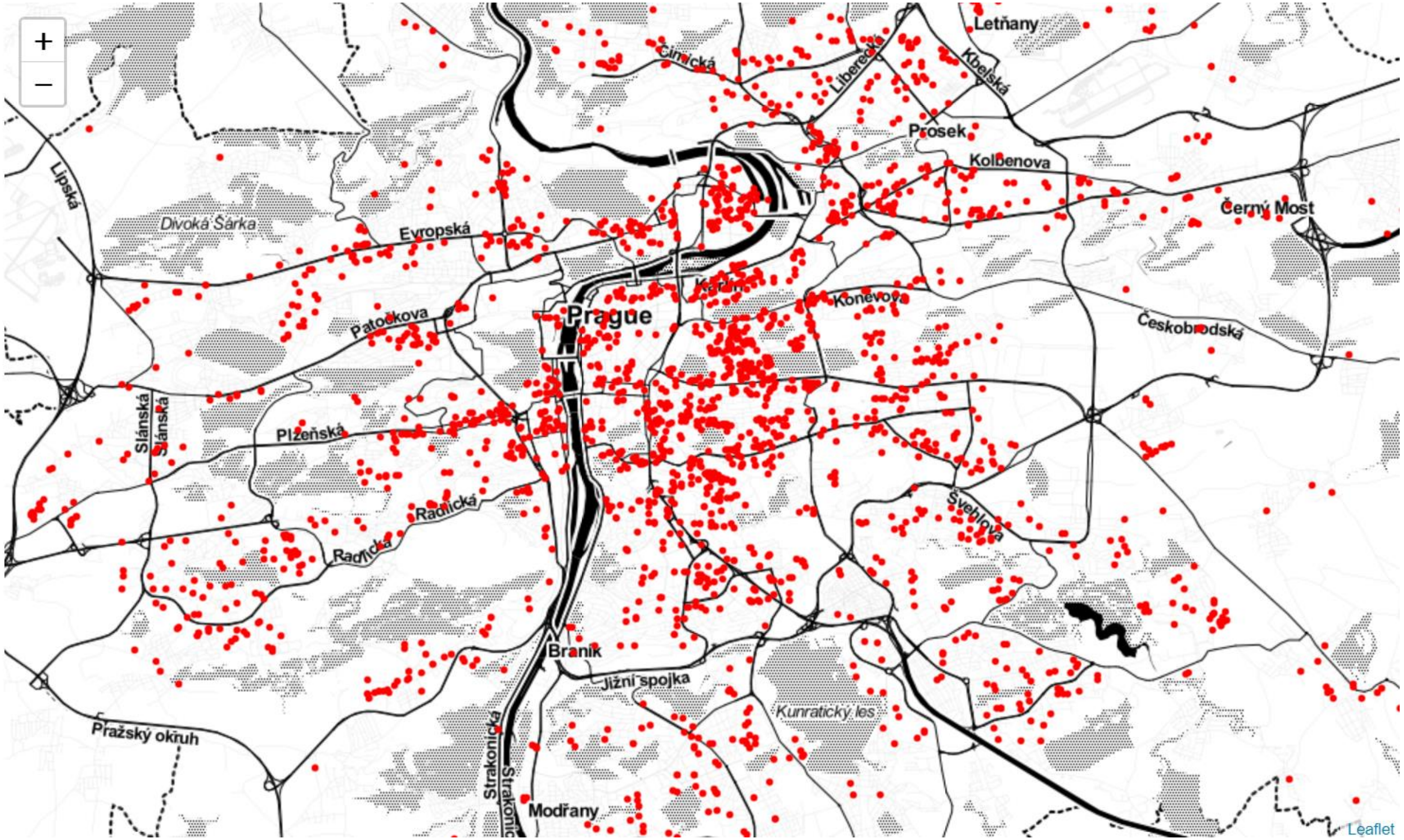


Price-Fees dependency - BEFORE



Price-Fees dependency - AFTER

Data exploration



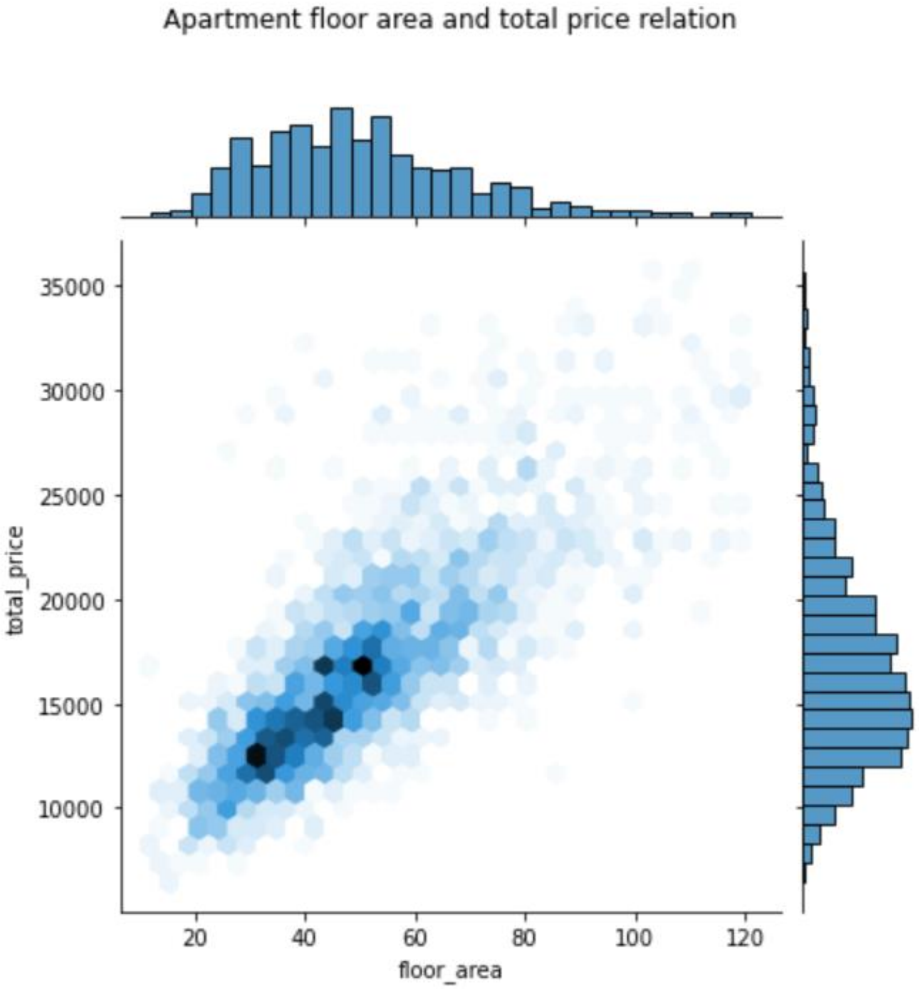
Location of listed apartments

Average fees for listed apartments are: **2755 CZK**

Average deposit values are **1.4x** of the apartment rent price.

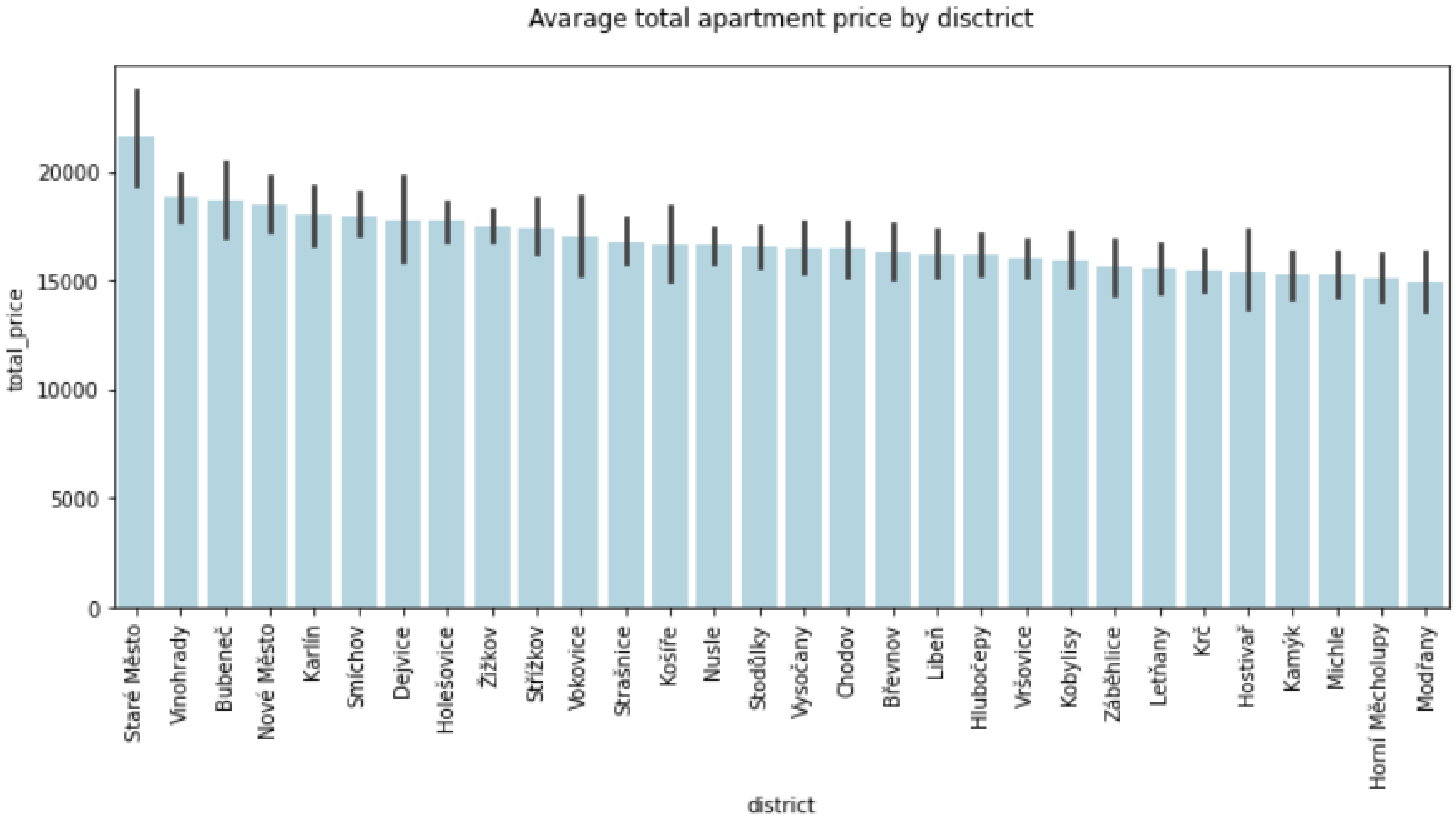
Data exploration

Apartment size and apartment price has visible correlation. It is the strongest price indicator of the apartment.



Data exploration

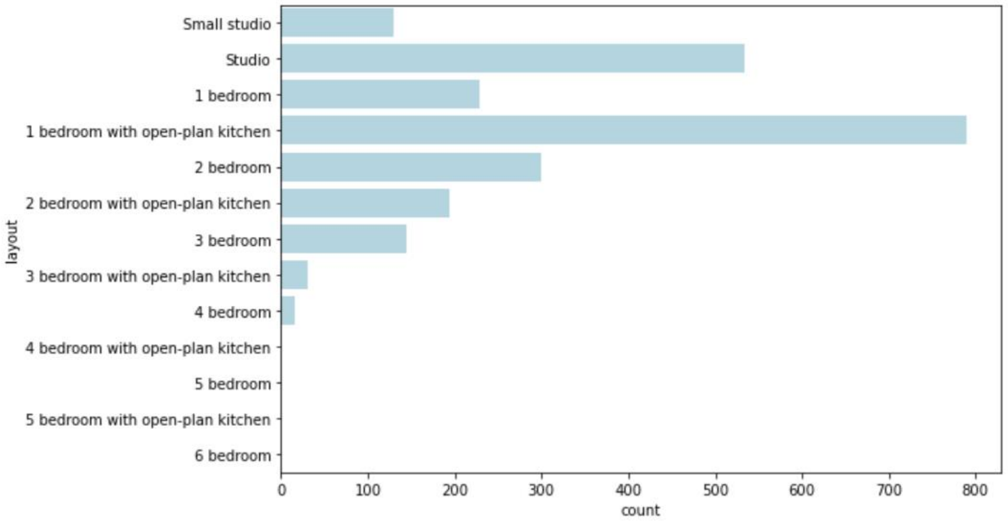
Another strong price factor is the location of the estate. The average apartment price for most popular districts are plotted below. *Stare Mesto* is the most expensive district to rent a flat. Since it is historic center of the Prague, it was an expected result.

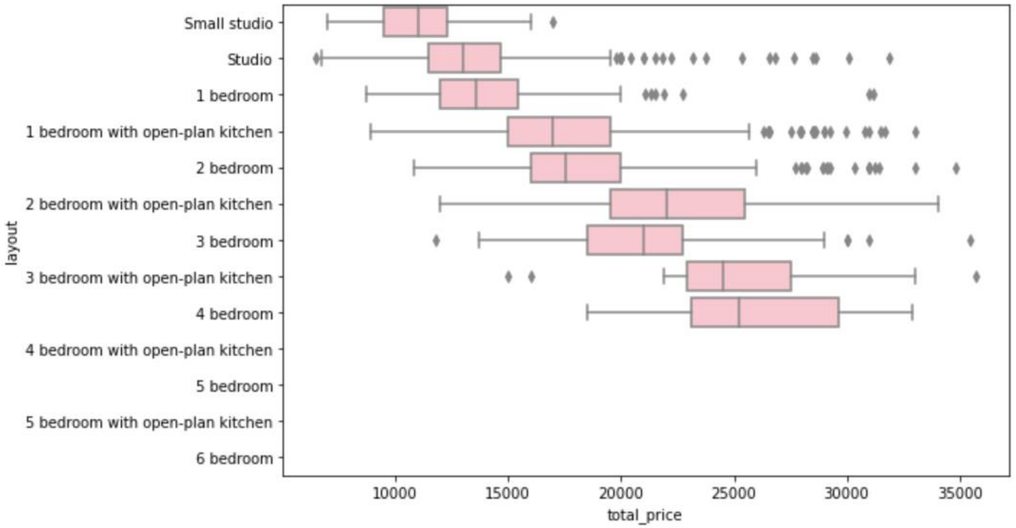


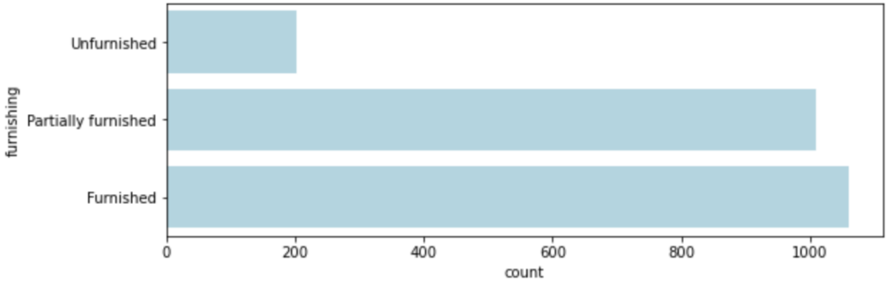
Avarage total apartment price by disctrict

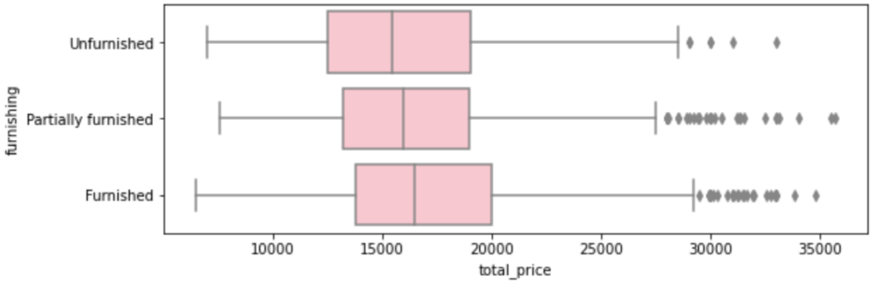
Data exploration

The most popular listings have one room with kitchen layout. It is noticeable that the larger number of the rooms in the apartment, the higher the prices get. The apartments comes with or without the furniture, which has reflection on the price. The unfurnished apartments are minority.


Apartment layout type count plot

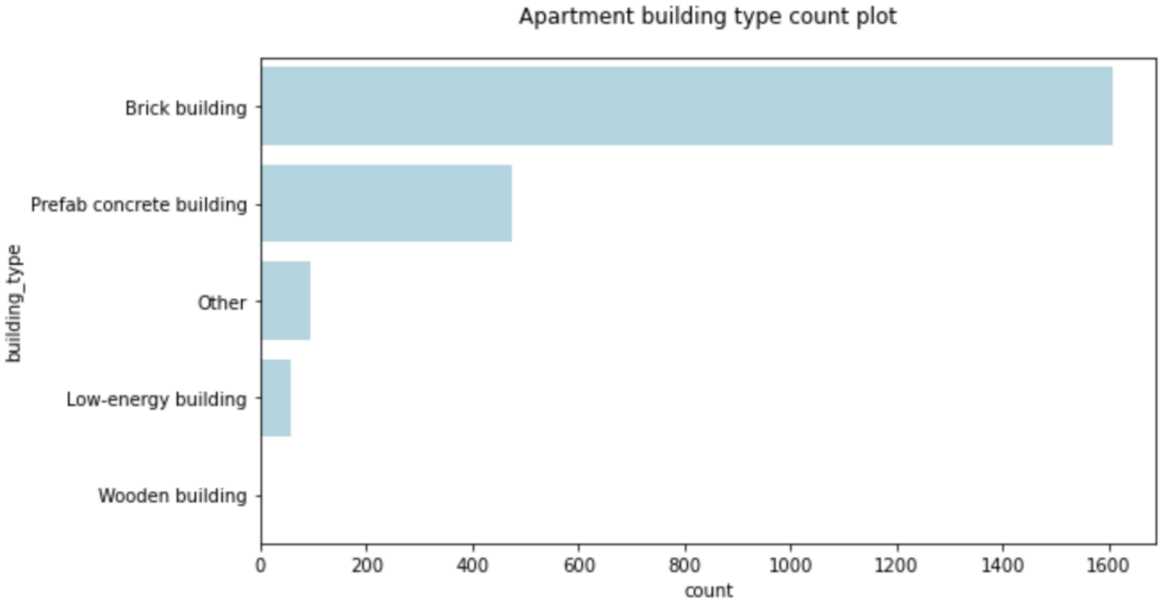

Apartment layout and price relation

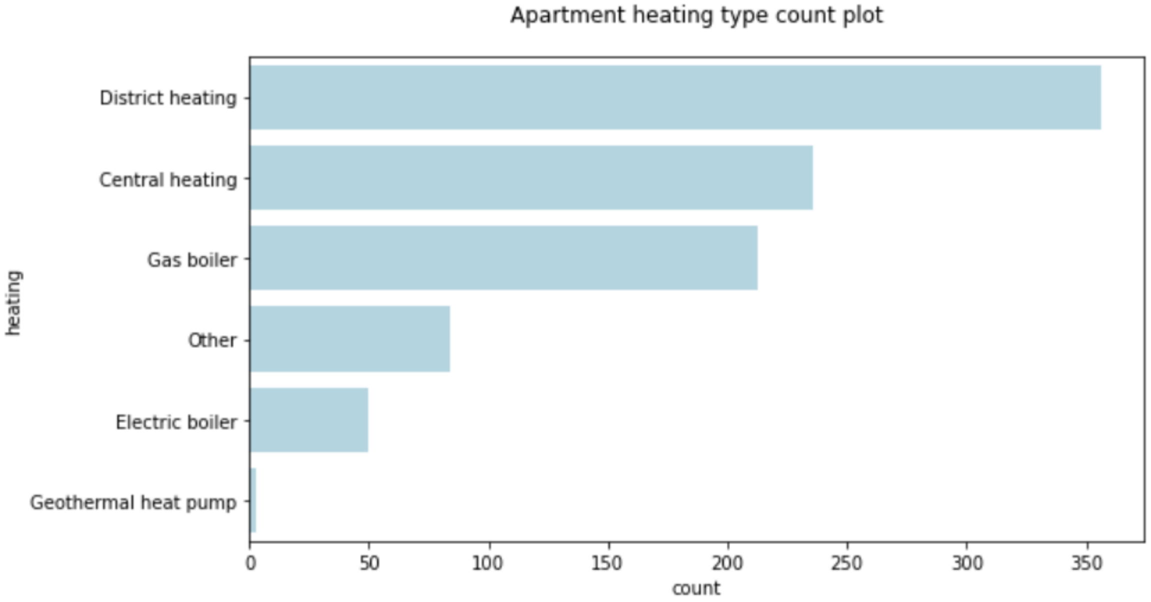

Apartment furnishment count plot


Apartment furnishment price

Data exploration

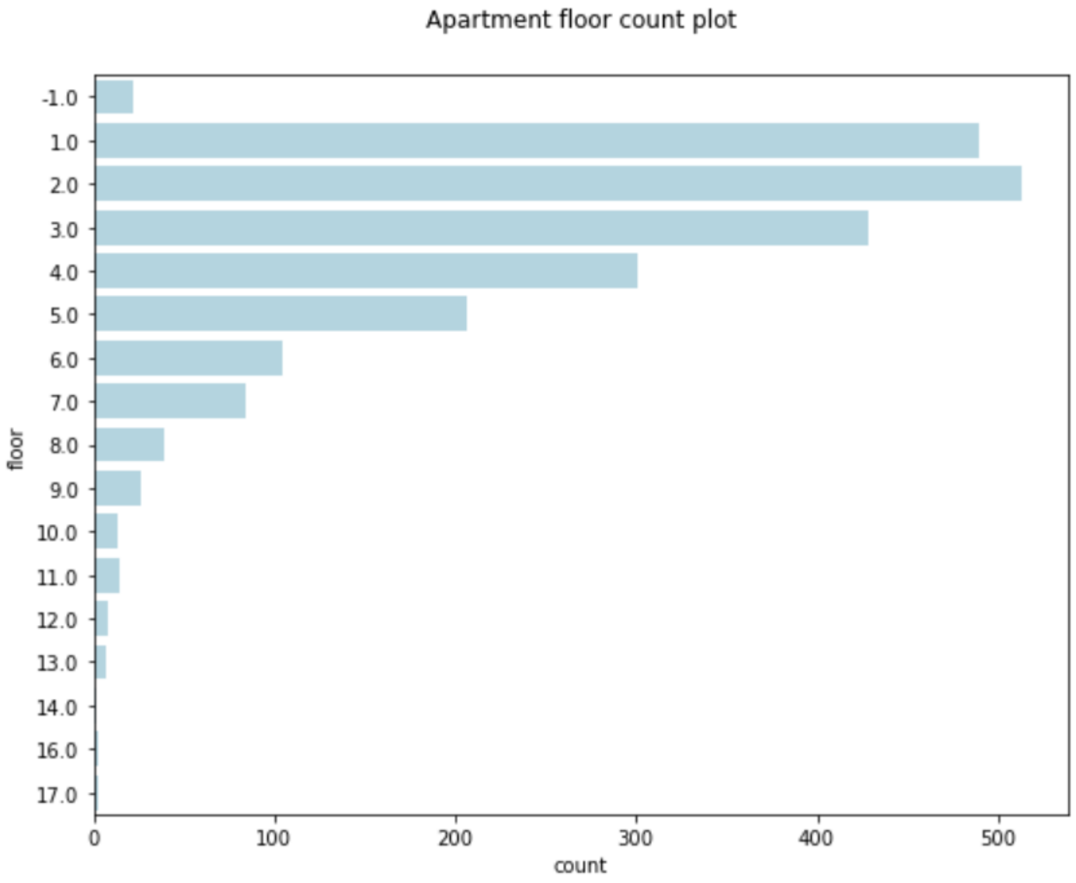
Most of the apartments that are currently for rent in Prague are made of brick. Also the majority comes with district or central heating.

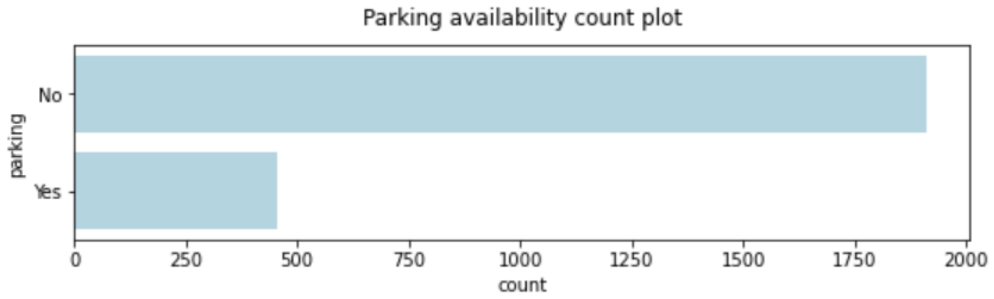


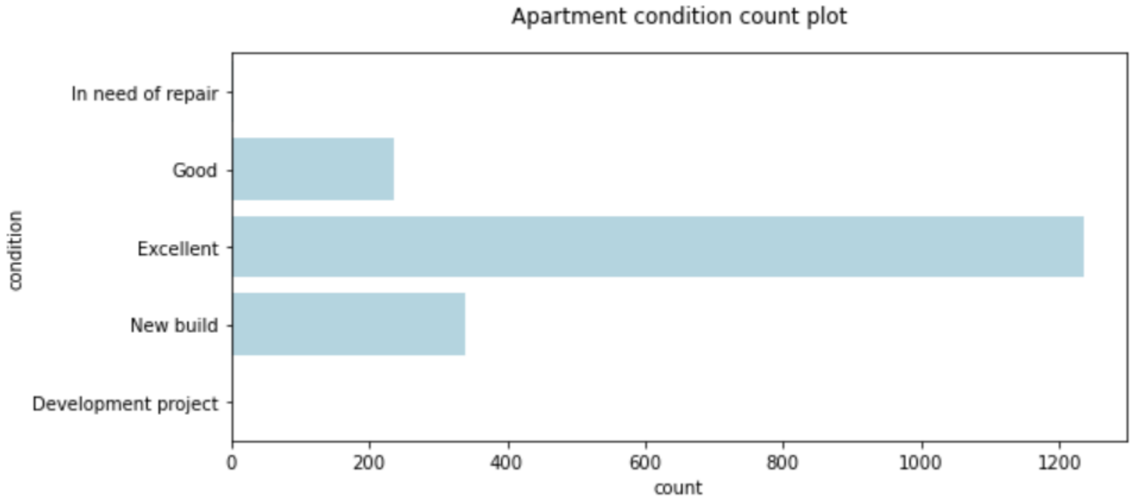


Data exploration

The other features that can be explored are: apartment floor, apartment condition and availability of the parking for a car.







Feature engineering

In order to make the machine learning model understand the data correctly, the ordinal and binary values of the features will be converted to numerical. Additionally, the categorical district features have been one-hot encoded.

Furnishing:
Unfurnished - 1
Partially furnished – 2
Furnished- 3

Age:
1 to 10 years - 1
10 to 30 years - 2
30 to 50 years - 3
over 50 years – 4

Condition:
Good - 1
Excellent – 2
New Build – 3

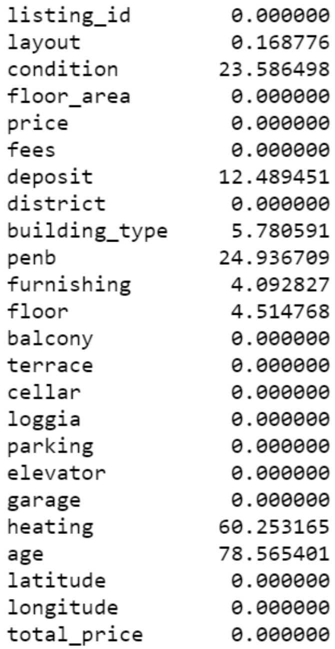
Layout:
Small studio - 0.5
Studio - 0.75
1 bedroom – 1
1 bedroom with open-plan kitchen - 1.5
2 bedroom - 2
2 bedroom with open-plan kitchen - 2.5
3 bedroom - 3
3 bedroom with open-plan kitchen - 3.5
4 bedroom - 4
4 bedroom with open-plan kitchen - 4.5
5 bedroom - 5
5 bedroom with open-plan kitchen - 5.5
6 bedroom - 6.5

Energy performance certificate (PENB):
A – 1
B – 2
C – 3
D – 4
E – 5
F – 6
G - 7

Balcony:	**Terrace:**	**Cellar:**	**Loggia:**	**Parking:**	**Elevator:**	**Garage:**
Yes – 1	Yes – 1	Yes – 1	Yes – 1	Yes – 1	Yes – 1	Yes – 1
No – 0	No – 0	No – 0	No – 0	No – 0	No – 0	No – 0

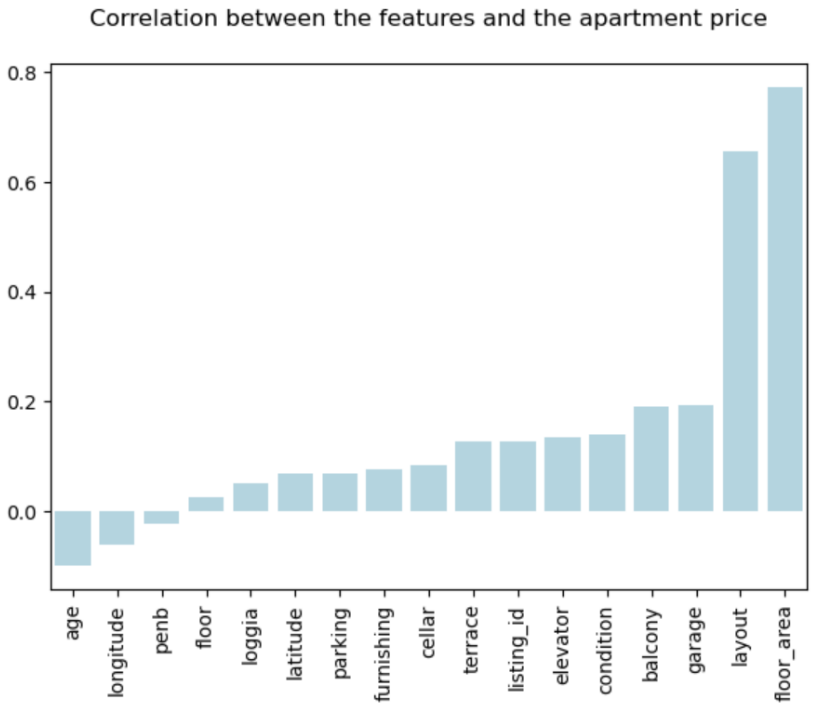
Feature engineering

The missing data is shown below as a percentage of missing values for each feature in the dataset. In order to decide which columns/rows to keep and which to drop, the correlation of each feature to the apartment price is plotted.



```
listing_id        0.000000
layout            0.168776
condition        23.586498
floor_area        0.000000
price             0.000000
fees              0.000000
deposit          12.489451
district          0.000000
building_type     5.780591
penb             24.936709
furnishing        4.092827
floor             4.514768
balcony           0.000000
terrace           0.000000
cellar            0.000000
loggia            0.000000
parking           0.000000
elevator          0.000000
garage            0.000000
heating          60.253165
age              78.565401
latitude          0.000000
longitude         0.000000
total_price       0.000000
```

Percent of missing values



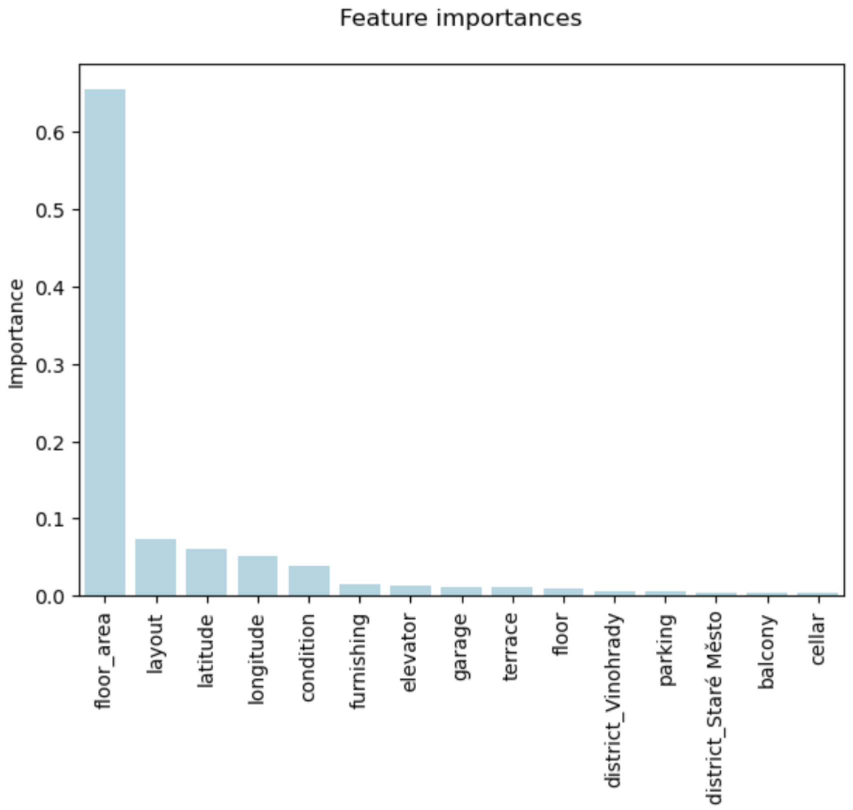
The *heating* and *age* features are dropped since they miss majority of data. *PENB* feature will be also dropped since it do not have strong correlation to the price. *Fees*, *deposit* and *price* are also dropped because they will be not used in prediction model. The columns with missing *condition* will be dropped because it has quite strong correlation to the total price. The rest of rows with missing data are removed as well since they are small percentage.

Modeling and evaluation

The purpose of the model is to predict the total price of the apartment based on 57 input variables. The data set is spitted into training and testing sets with 85/15 ratio. The training data is used to train the machine learning model and perform hyper parameter tuning with cross validation sampling. The remaining test data is to validate the model on the “unseen” data.

The Gradient Boosting Trees machine learning model has been chosen for the problem, since it outperformed other regression models. In the figure on the side, it is possible to see the importance of the most important features that are used by the model to predict the apartment price.

After running a test on the test data the results are:
Mean absolute error: **1892 CZK**
Mean absolute error: **11.0 %**
Meaning that we are able to estimate the price of the apartment with average 1892 kc error.

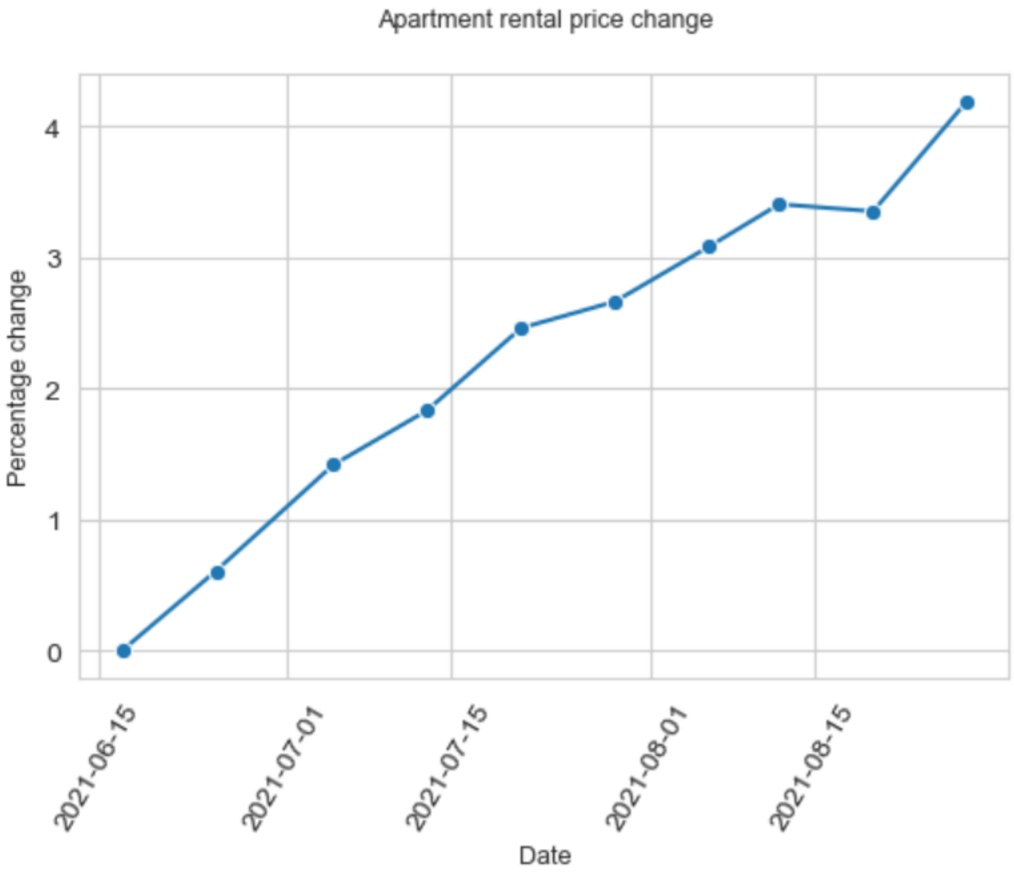


Price trend

The same machine learning model was trained multiple times on the data that was collected on the weekly basis.

Each model was given the same input dataset to predict the prices of the same apartments. Since the models were trained with the data collected in a different periods, their price estimation slightly differs from each other.

Averages of the relative price changes are plotted on the right. It is visible that post-pandemic rental market in Prague is rising with the prices growing around 4% over the period of 2.5 months.



Thank you

Tomasz Bialy
01/09/2021