



ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана

# Предсказание стоимости квартиры в Москве

А.С. Кропотова



ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана

# Этапы выполнения работы:

1

Сбор данных

2

Первичный и визуальный анализ данных

3

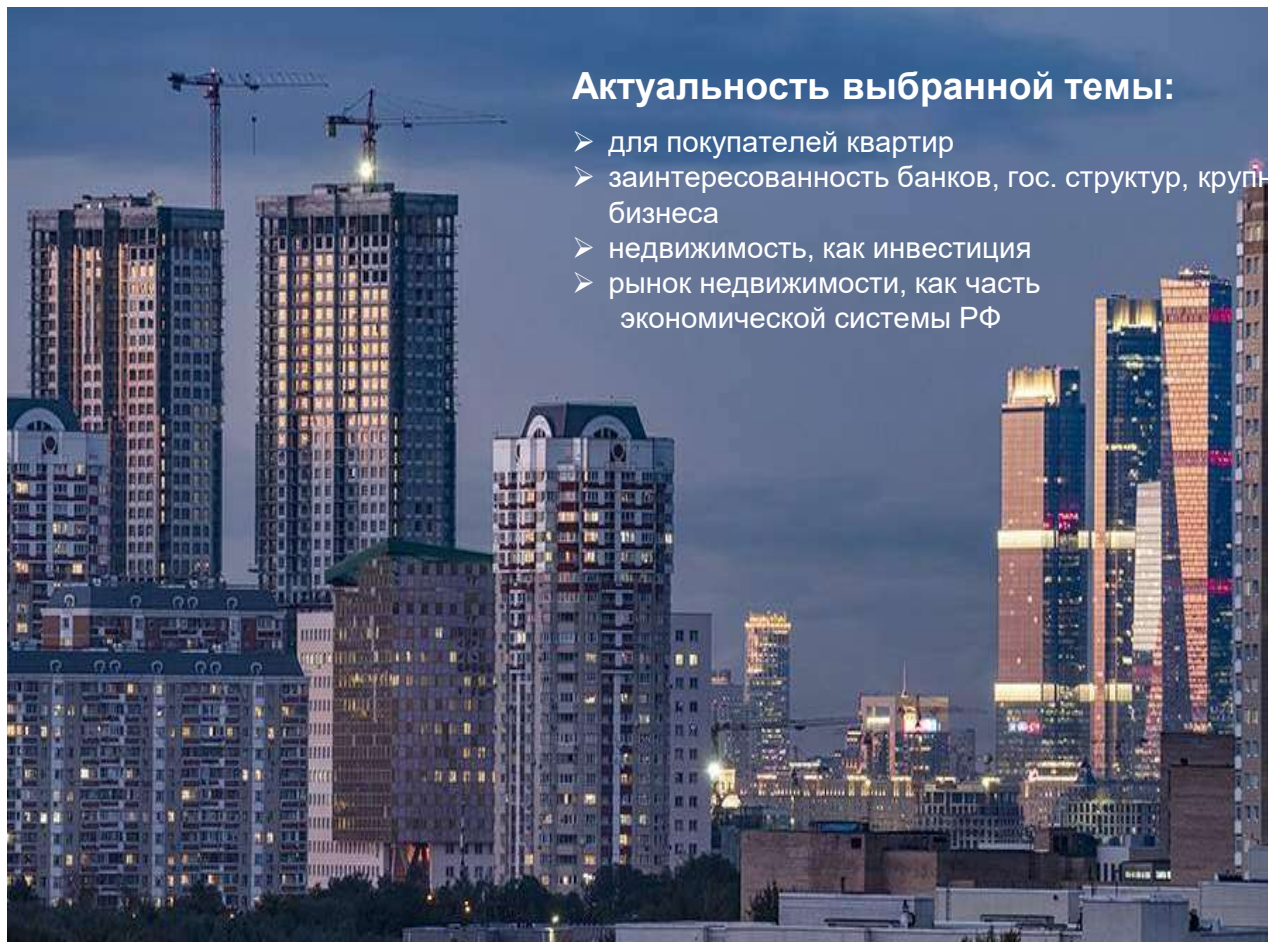
Предобработка данных

4

Обучение моделей и выбор лучшей

5

Создание приложения





ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана

# Использование библиотеки

## BeautifulSoup

Парсинг данных с сайта  
cian.ru

Ограничения:

28 объявлений на 1 странице

54 страницы для парсинга

Парсинг студий и 1-комн.квартир, новостройки, не апартаменты

```
data = []

for p in range(1, 55):
    print(p)
    url = (f'https://www.cian.ru/cat.php?deal_type=sale&engine_version=2&foot_min=45&object_type%5B%5D=2&offer_type=')
    r = requests.get(url)
    sleep(1)
    soup = BeautifulSoup(r.text, 'xml')
    flats = soup.findAll('article', class_='93444fe79c--container--Povoi_93444fe79c--cont--0zgVc')
    for flat in flats:
        try:
            link = flat.find('a', class_='93444fe79c--link--eoxce').get('href')
        except AttributeError:
            continue
        data.append([link])
```

```
for j in range(0, len(df1)):
    print(j)
    url = df1['link'][j]
    r = requests.get(url)
    #sleep(1)
    soup = BeautifulSoup(r.text, 'xml')
    try:
        df1['price'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--aside--qQIEI').find('span').text
        df1['min_to_metro'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
        df1['region_of_moscow'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
        for i in range(0, 6):
            if 'Общая площадь' in soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--text--epIqM').text:
                df1['total_area'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--text--epIqM').text
            else:
                df1['total_area'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
        for i in range(0, 4):
            if 'Жилая площадь' in soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text:
                df1['living_area'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
            else:
                df1['living_area'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
        for i in range(0, 4):
            if 'Этаж' in soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text:
                df1['floor'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
            else:
                df1['floor'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
        for i in range(0, 4):
            if 'Этаж' in soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text:
                df1['number_of_floors'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
            else:
                df1['number_of_floors'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
        for i in range(0, 4):
            if 'Год сдачи' in soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text:
                df1['construction_year'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
            else:
                df1['construction_year'][j] = soup.find('div', class_='a10a3f92e9--page--0Yngf').find('div', class_='a10a3f92e9--center--b3Pe0').text
    except:
        continue
    except AttributeError:
        continue
```





# Итоговый датасет

Размерность: (1937, 12)

```
#   Column      Non-Null Count  Dtype
---  -
0    price      1937 non-null      int64
1    min_to_metro 1908 non-null      float64
2    region_of_moscow 1919 non-null      object
3    total_area   1919 non-null      float64
4    living_area  1264 non-null      float64
5    floor        1904 non-null      float64
6    number_of_floors 1648 non-null      float64
7    construction_year 1245 non-null      float64
8    is_new       1937 non-null      int64
9    is_apartments 1937 non-null      int64
10   ceiling_height 631 non-null      float64
11   number_of_rooms 1937 non-null      int64
dtypes: float64(7), int64(4), object(1)
memory usage: 181.7+ KB
```

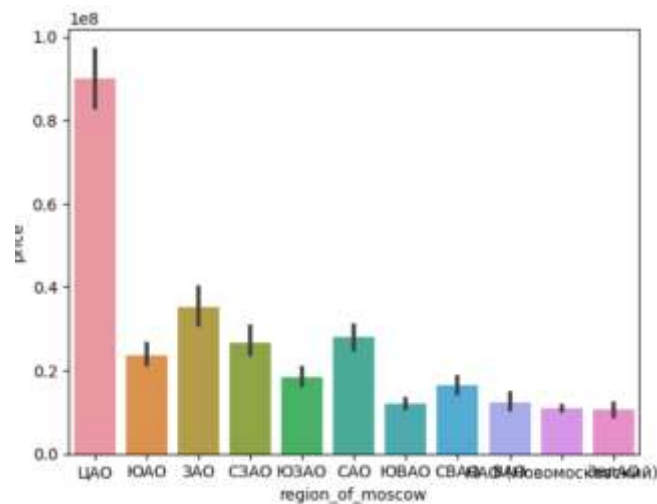
	price	min_to_metro	region_of_moscow	total_area	living_area	floor	number_of_floors	construction_year	is_new	is_apartments	ceiling_height	number_of_rooms
price	1.000000	-0.268555	-0.240612	0.782794	0.692338	0.173664	0.068358	0.078912	-0.097024	0.080637	0.363075	0.458756
min_to_metro	-0.268555	1.000000	0.173402	-0.271201	-0.215160	-0.141498	-0.190139	0.103728	0.127824	-0.110511	-0.147725	-0.129680
region_of_moscow	-0.240612	0.173402	1.000000	-0.203291	-0.176903	-0.128194	-0.134288	0.065288	0.118369	-0.138442	-0.166096	-0.124049
total_area	0.782794	-0.271201	-0.203291	1.000000	0.900997	0.288062	0.231385	0.037386	-0.148324	0.052037	0.358493	0.706081
living_area	0.692338	-0.215160	-0.176903	0.900997	1.000000	0.260786	0.245794	-0.038494	-0.236518	0.052046	0.287757	0.696982
floor	0.173664	-0.141498	-0.128194	0.288062	0.260786	1.000000	0.775258	0.237217	-0.029036	0.134029	0.197921	0.175736
number_of_floors	0.068358	-0.190139	-0.134288	0.231385	0.245794	0.775258	1.000000	0.320282	0.064298	0.081422	0.168462	0.094108
construction_year	0.078912	0.103728	0.065288	0.037386	-0.038494	0.237217	0.320282	1.000000	0.513752	0.266663	0.248801	-0.000076
is_new	-0.097024	0.127824	0.118369	-0.148324	-0.236518	-0.029036	0.064298	0.513752	1.000000	-0.066851	0.244369	-0.219900
is_apartments	0.080637	-0.110511	-0.138442	0.052037	0.052046	0.134029	0.081422	0.266663	-0.066851	1.000000	0.288919	0.083411
ceiling_height	0.363075	-0.147725	-0.166096	0.358493	0.287757	0.197921	0.168462	0.248801	0.244369	0.288919	1.000000	0.149365
number_of_rooms	0.458756	-0.129680	-0.124049	0.706081	0.696982	0.175736	0.094108	-0.000076	-0.219900	0.083411	0.149365	1.000000

Признак	Описание признака
price	Целевая переменная. Стоимость квартиры
min_to_metro	Количество минут до метро пешком
region_of_moscow	Адм. округа и регионы Москвы
total_area	Общее число кв. м квартиры
living_area	Число жилых кв. и квартиры
floor	Этаж, на котором располагается квартира
number_of_floors	Этажность дома
construction_year	Год постройки (сдачи) дома
is_new	Бинарный признак. 0 - вторичное жилье, 1 - новостройка
is_apartments	Бинарный признак. 0 - не апартаменты, 1 - апартаменты
ceiling_height	Высота потолка квартиры
number_of_rooms	Количество комнат в квартире
min_to_metro	Количество минут до метро пешком

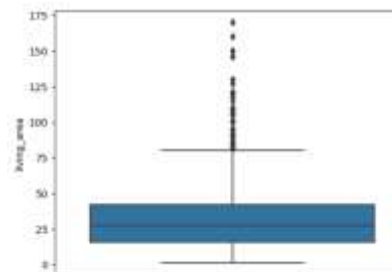
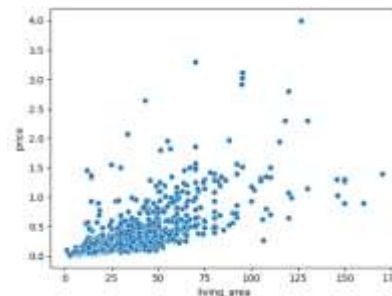
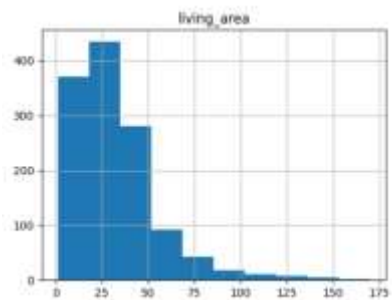


# Признаки

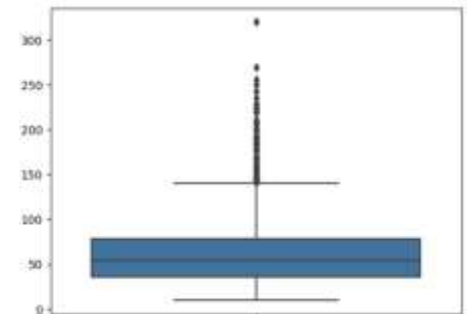
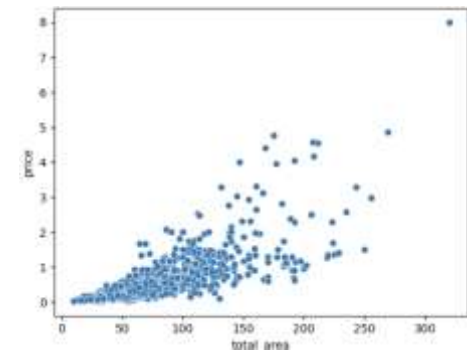
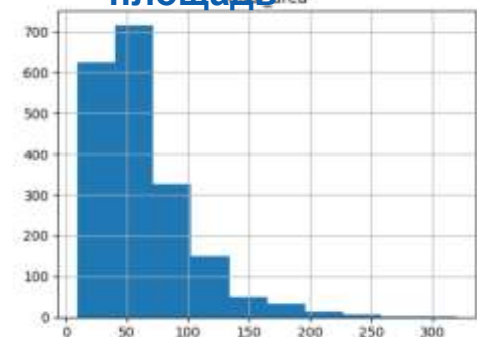
жилая площадь



Стоимость квартиры  
зависит от ее расположения



общая  
площадь

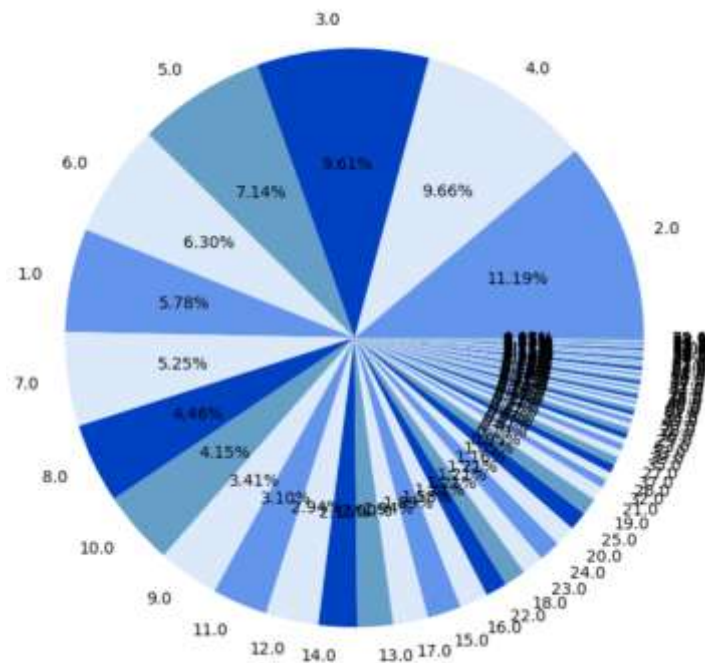




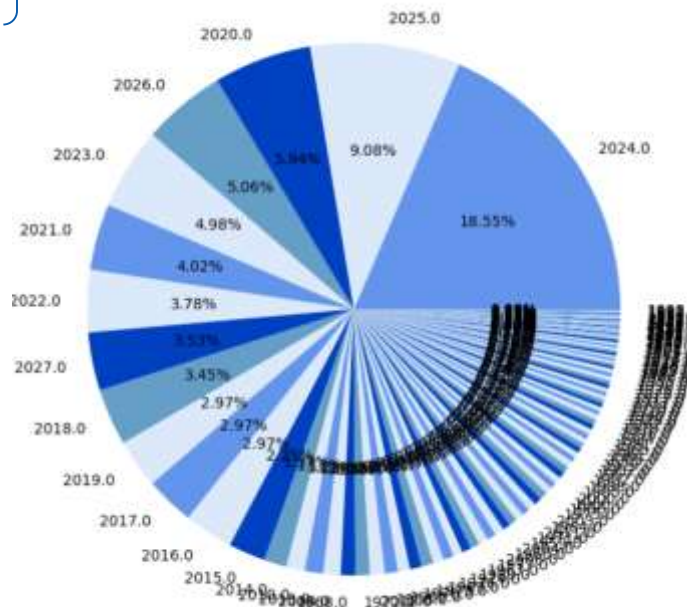
ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГУ им. Н.Э. Баумана

# Признаки

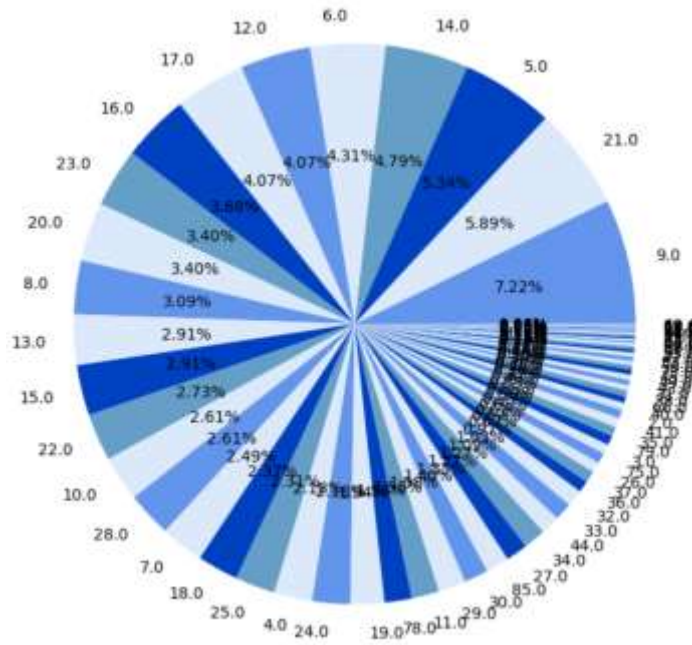
этаж



год сдачи дома



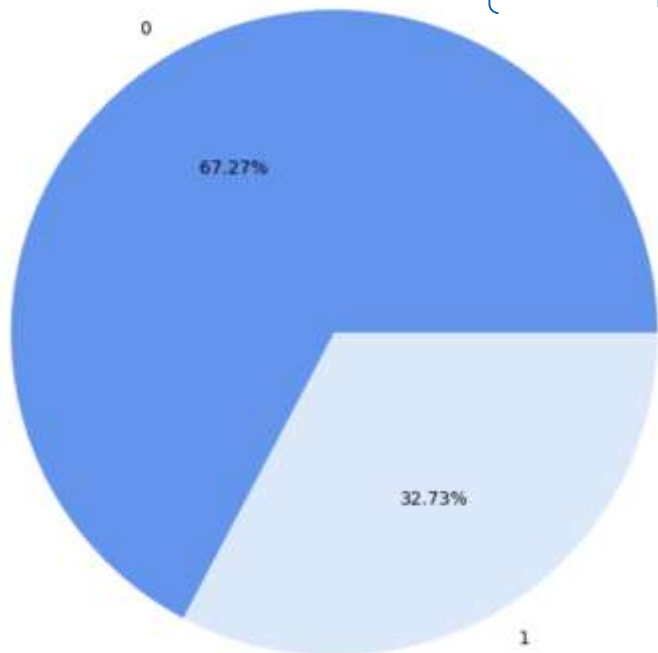
этажность дома



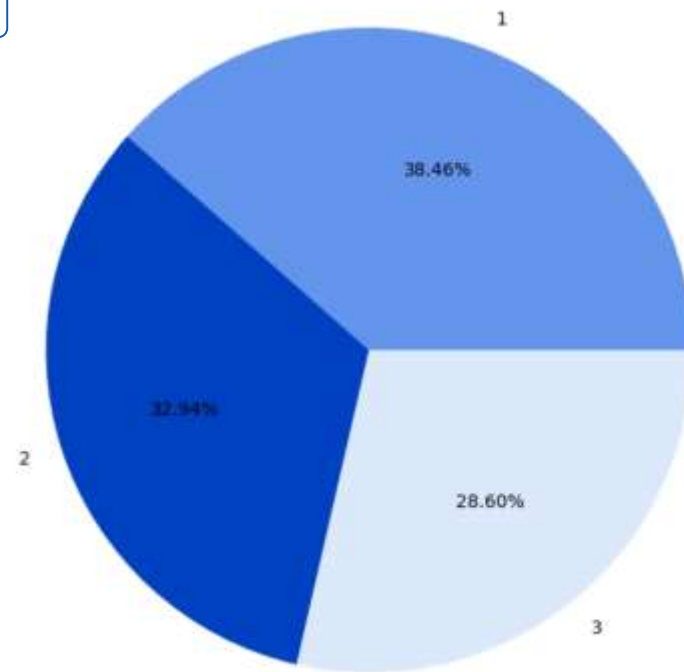


# Признаки

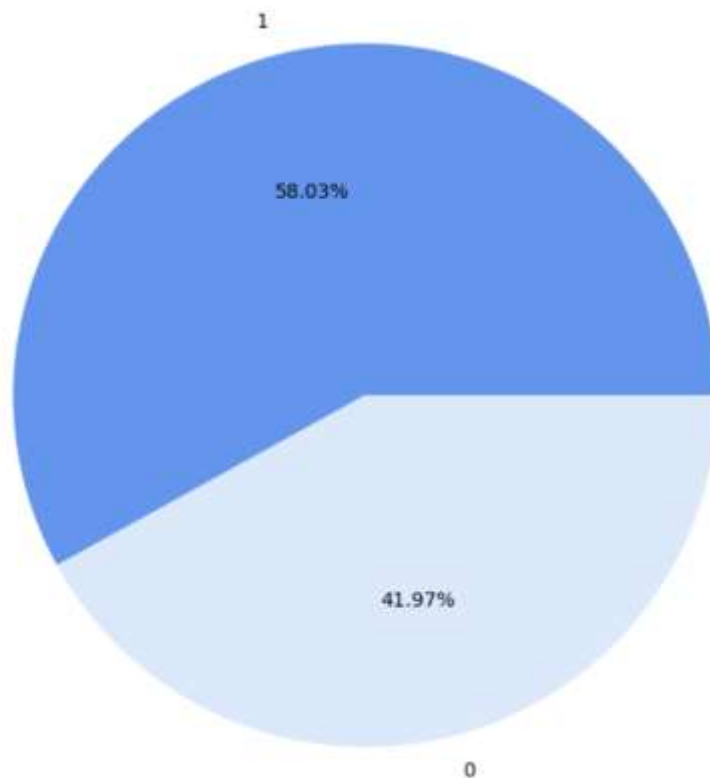
Новостройка или вторичка?



Количество комнат



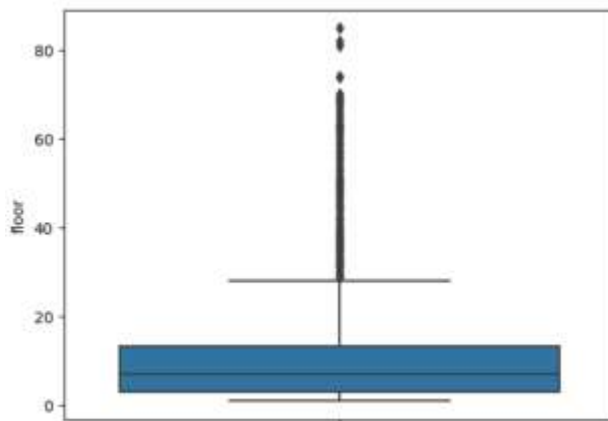
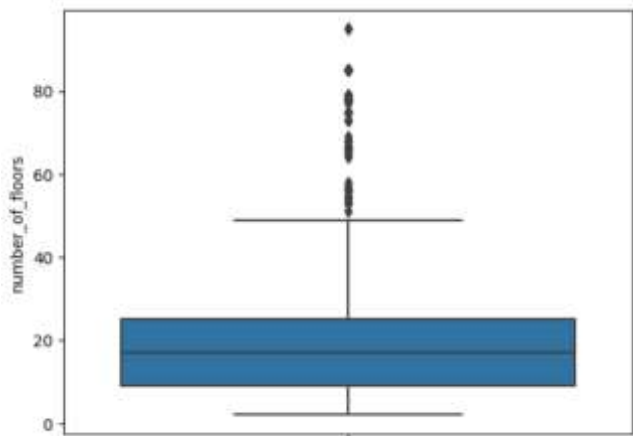
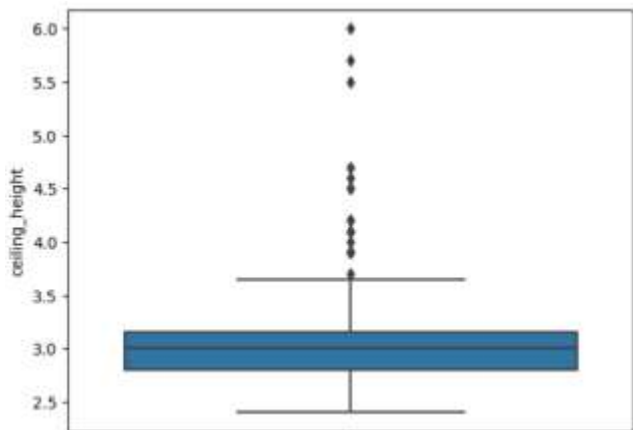
Апартаменты или квартира?





# Предобработка данных

## Работа с выбросами



## Замена пропущенных значений

( KNNImputer – для числовых признаков  
SimpleImputer (strategy='most\_frequent')  
– для категориальных признаков )

## Преобразование категориальных признаков

( OneHotEncoder  
LabelEncoder )

## Масштабирование данных

( StandartScaler )





# Выбор и обучение моделей

```
def get_score(X_train, X_test, y_train, y_test, model='LinearRegression', is_return=False):
    assert model in ['LinearRegression', 'DecisionTreeRegressor', 'BaggingRegressor', 'RandomForestRegressor', 'VotingRegressor', 'StackingRegressor'],
    if model == 'LinearRegression':
        model = LinearRegression()
    elif model == 'DecisionTreeRegressor':
        model = DecisionTreeRegressor(random_state=15, max_depth=8, min_samples_leaf=3)
    elif model == 'BaggingRegressor':
        model = BaggingRegressor(estimator=DecisionTreeRegressor(), random_state=15, n_estimators=100)
    elif model == 'RandomForestRegressor':
        model = RandomForestRegressor(random_state=15, n_estimators=100)
    elif model == 'VotingRegressor':
        model = VotingRegressor(estimators=[('rf', RandomForestRegressor(random_state=15, n_estimators=100)),
                                           ('bag', BaggingRegressor(estimator=DecisionTreeRegressor(), random_state=15, n_estimators=100)),
                                           ('tree', DecisionTreeRegressor(random_state=15, max_depth=10, min_samples_leaf=3))])
    elif model == 'StackingRegressor':
        model = StackingRegressor(estimators=[('rf', RandomForestRegressor(random_state=15, n_estimators=100)),
                                           ('bag', BaggingRegressor(estimator=DecisionTreeRegressor(), random_state=15, n_estimators=100)),
                                           ('tree', DecisionTreeRegressor(random_state=15, max_depth=10, min_samples_leaf=3))],
                                final_estimator=RandomForestRegressor(random_state=1))
    elif model == 'GradientBoostingRegressor':
        model = GradientBoostingRegressor(random_state=15)

    model.fit(X_train, y_train)
    y_train_pred = model.predict(X_train)
    y_test_pred = model.predict(X_test)
    RMSE_train = sqrt(mean_squared_error(y_train, y_train_pred))
    RMSE_test = sqrt(mean_squared_error(y_test, y_test_pred))
    R2_train = model.score(X_train, y_train)
    R2_test = model.score(X_test, y_test)

    if is_return:
        print(model)
        print('RMSE_train:', RMSE_train)
        print('RMSE_test:', RMSE_test)
        print('R2_train:', R2_train)
        print('R2_test:', R2_test)

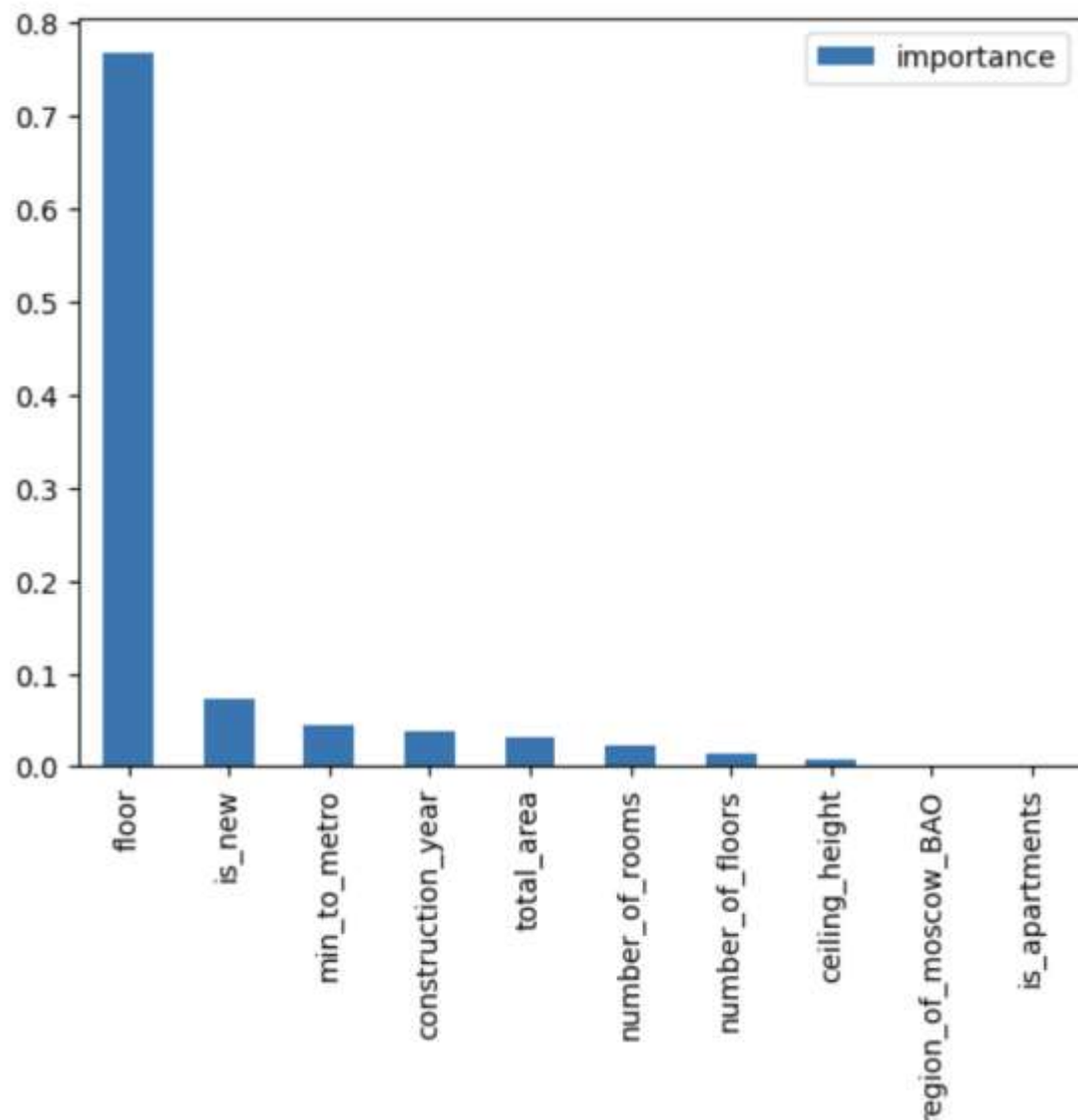
    return model
```

результаты  
оценки  
качества  
моделей

	model	RMSE_train	RMSE_test	R2_train	R2_test
0	LinearRegression	31 969 695	26 853 582	0.69	0.6
1	DecisionTreeRegressor	17 253 224	21 576 890	0.9	0.74
2	RandomForestRegressor	10 958 846	23 908 002	0.96	0.68
3	VotingRegressor	11 170 754	21 825 841	0.96	0.74
4	StackingRegressor	17 206 619	26 618 958	0.9	0.6
5	GradientBoostingRegressor	15 397 894	25 984 686	0.93	0.63
6	DecisionTreeRegressor_GS	22 945 650	27 655 807	0.84	0.58
7	GradientBoostingRegressor_RS	12 059 829	25 847 816	0.96	0.63



# Отбор признаков



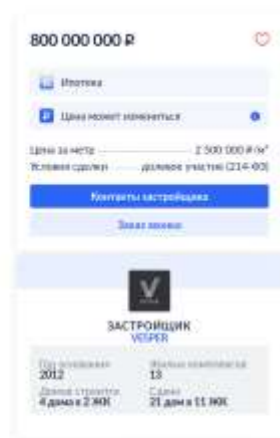
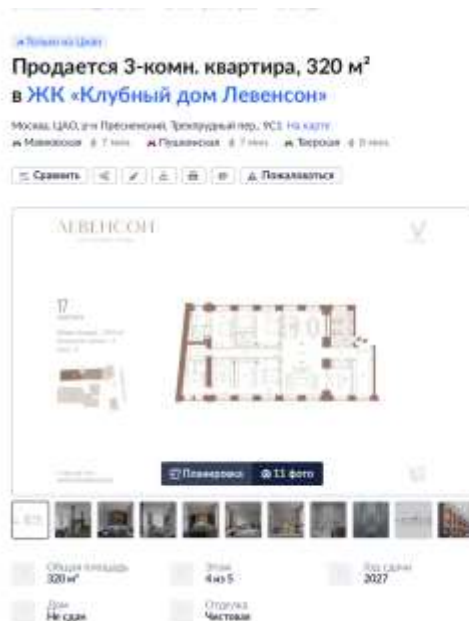
	feature_name	importance
2	floor	0.766339
5	is_new	0.072733
0	min_to_metro	0.044178
4	construction_year	0.038531
1	total_area	0.030422
8	number_of_rooms	0.022665
3	number_of_floors	0.013958
7	ceiling_height	0.008030
9	region_of_moscow_BAO	0.001891
6	is_apartments	0.001253



ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана

# Приложение

<http://127.0.0.1:5000/>



**Пожалуйста, введите данные для предсказания стоимости квартиры:**

регион Москвы

количество минут до метро пешком

общая площадь квартиры (кв.м)

этаж, на котором будет располагаться квартира

количество этажей дома

год постройки (сдачи квартиры)

новостройка(1) / вторичка (0)

апартаменты(1) / не апартаменты (0)

высота потолков (стандартно-2,5 м)

количество комнат

ПОЛУЧИТЬ ПРОГНОЗ ПО СТОИМОСТИ КВАРТИРЫ

Стоимость квартиры по заданным параметрам составит:

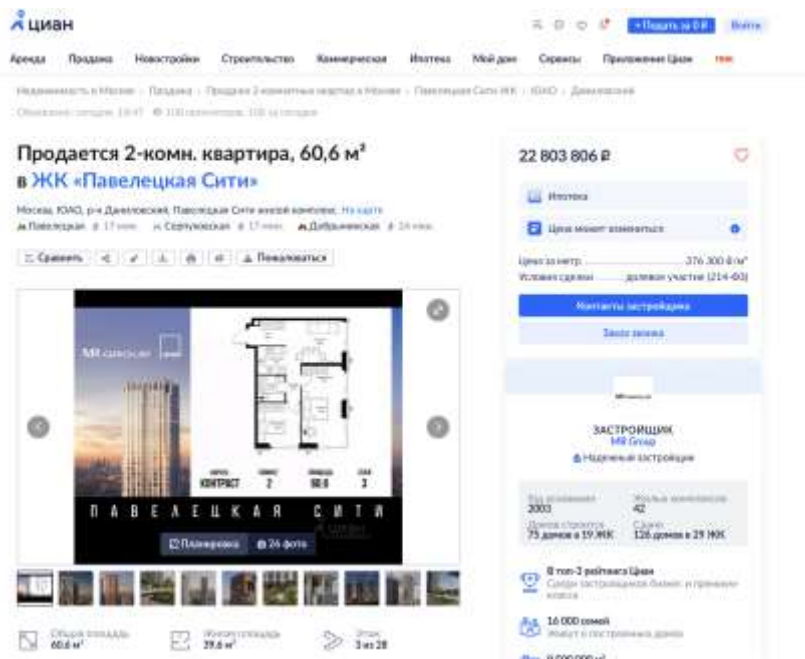
[6.43713105e+08]



ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана

# Приложение

<http://127.0.0.1:5000/>



<https://www.cian.ru/sale/flat/306614015/>

**Пожалуйста, введите данные для предсказания стоимости квартиры:**

регион Москвы

количество минут до метро пешком

общая площадь квартиры (кв.м)

этаж, на котором будет располагаться квартира

количество этажей дома

год постройки (сдачи квартиры)

новостройка(1) / вторичка (0)

апартаменты(1) / не апартаменты (0)

высота потолков (стандартно-2,5 м)

количество комнат

ПОЛУЧИТЬ ПРОГНОЗ ПО СТОИМОСТИ КВАРТИРЫ

Стоимость квартиры по заданным параметрам составит:

[23233872.67]





ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГУ им. Н.Э. Баумана

# Приложение

<http://127.0.0.1:5000/>

Хорошая цена

**Продается 2-комн. квартира, 42,4 м²**

Москва, м-р ЮЗАО (Новомосковский), д.б.б. Вольская, 7х1 на карте

Филиппов Лут. 7 мин. Новомосковская 5 мин. Пресненская 9 мин.

Калужское шоссе. 5 км от МКАД. Киевское шоссе. 9 км от МКАД.

Сравнить Показать фото



15 фото

Общая площадь: 42,4 м²

Этаж: 2 из 11

Жилая площадь: 16,7 м²

Площадь кухни: 15,7 м²

Подъезд: 2019

12 900 000 Р

Следить за изменениями цены

Предложить свою цену

Уточнение: 12 500 000

Выборная ипотека

Ипотека

Цена за метр: 304 245 Р/м²

Условия сделки: свободная продажа

Ипотека: возможно

Позвонить телефону

Написать


Собственник: Ю 116114838

Предоставил паспорт

Проверено Росреестром

ЖК «Саминский парк 2»

Жилой комплекс комфорт-класса с собственной инфраструктурой рядом с Юным лесопарком.



**Пожалуйста, введите данные для предсказания стоимости квартиры:**

регион Москвы

количество минут до метро пешком

общая площадь квартиры (кв.м)

этаж, на котором будет располагаться квартира

количество этажей дома

год постройки (сдачи квартиры)

новостройка(1) / вторичка (0)

апартаменты(1) / не апартаменты (0)

высота потолков (стандартно-2,5 м)

количество комнат

ПОЛУЧИТЬ ПРОГНОЗ ПО СТОИМОСТИ КВАРТИРЫ

Стоимость квартиры по заданным параметрам составит:

[11953620.46]



ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана

# Способы улучшения модели

1 Сбор большего количества разнообразных данных

2 Добавление новых признаков

3 Анализ рынка недвижимости и добавление новых признаков

4 Изучение текущей ситуации с программами: семейной, льготной, ипотеки, др. программ

5 Добавление даты и сезонного признака





ЦЕНТР  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
МГТУ им. Н.Э. Баумана



[do.bmstu.ru](https://do.bmstu.ru)