**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

Аналитический обзор зависимости рыночной стоимости квартир от различных параметров

Выполнил(а)

_____
(подпись)

Проверил(а)

_____
(подпись)

Дата _____

МОСКВА 2022

## Оглавление

# 1 Постановка задачи

## 1.1 Описание бизнес-задачи и её актуальность

В рамках анализа данных о продаже квартир в городе Санкт-Петербурге и Ленинградской области стоит задача найти наиболее значимые параметры, влияющие на рыночную стоимость недвижимости.

Данный анализ полезен для формирования первоначальной стоимости квартиры при выставлении её на рынок. Основные показатели позволят спрогнозировать её стоимость и как следствие не позволит продешевить или же завысить этот показатель.

## 1.2 Метрики для проверки гипотез, источники информации для сбора данных

В качестве основных метрик для анализа данных гипотез и формирования новых были выбраны соответствующие показатели (а именно: этажность, площадь, количество комнат, отдалённость от центра) из датасета real_estate_data.csv.

# 2 Анализ исходных данных

## 2.1 Предварительный анализ данных и методы их обработки

Исходные данные были получены из сервиса Яндекс.Недвижимость и представляют собой данные о продаже квартир в Санкт-Петербурге и соседних населённых пунктах за последние несколько лет (см. ссылку в пункте [1] списка источников).

Файл, содержащий исходные данные, имеет 23699 объектов недвижимости, каждый из которых имеет 22 признака (на Рисунке 1.1 приведена информация об исходном датасете).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23699 entries, 0 to 23698
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   total_images          23699 non-null  int64  
 1   last_price            23699 non-null  float64
 2   total_area            23699 non-null  float64
 3   first_day_exposition  23699 non-null  object 
 4   rooms                 23699 non-null  int64  
 5   ceiling_height        14504 non-null  float64
 6   floors_total          23613 non-null  float64
 7   living_area           21796 non-null  float64
 8   floor                 23699 non-null  int64  
 9   is_apartment          2775 non-null   object 
 10  studio                23699 non-null  bool   
 11  open_plan             23699 non-null  bool   
 12  kitchen_area          21421 non-null  float64
 13  balcony               12180 non-null  float64
 14  locality_name         23650 non-null  object 
 15  airports_nearest      18157 non-null  float64
 16  cityCenters_nearest   18180 non-null  float64
 17  parks_around3000      18181 non-null  float64
 18  parks_nearest         8079 non-null   float64
 19  ponds_around3000      18181 non-null  float64
 20  ponds_nearest         9110 non-null   float64
 21  days_exposition       20518 non-null  float64
dtypes: bool(2), float64(14), int64(3), object(3)
memory usage: 3.7+ MB
```

Рисунок 1.1 – информация об исходных данных

По результатам внешнего анализа данных можем наблюдать, что в исходном датасете пропуски имеются в 14-ти полях, из них имеет смысл обработать ряд полей. Перечень полей, подлежащих заполнению пропусков:

– celiling_height;

– living_area;

– floors_total;

– balcony;

– kitchen_area;

– days_exposition.

Поле ceiling_height, отвечающее за высоту потолков содержит в себе некоторые аномальные значения, а именно: высота потолка 25 или же 35 метров. Подразумевается, что это была ошибка ввода пользователем и он имел ввиду 2.5 и 3.5 метров. Также в данном поле существуют слишком маленькие значения, такие как 1, 1.20, 1.75 и так далее, их стоит воспринимать как аномалии.

```
[ ]  df['ceiling_height'].value_counts().sort_index()

1.00       1
1.20       1
1.75       1
2.00      11
2.20       1
          ..
26.00      1
27.00      8
27.50      1
32.00      2
100.00     1
Name: ceiling_height, Length: 184, dtype: int64
```

Рисунок 1.2 – значения в поле ceiling_height

Поля living_area и kitchen_area, отвечающие за жилую площадь и площадь кухни соответственно, также имеют пропуски, которые целесообразно заменить на (при наличии) разницу между total_area и kitchen_area в случае с living_area и разницу между total_area и living_area в случае с kitchen_area. Пропуски в поле площади кухни в основном обусловлены тем, что в исходных данных доля квартир – студии и площадь их кухни не подлежит оценке.

Поле floors_total – отвечает за общее число этажей в доме. Логично оценить заполнение данных пропусков не представляется возможным, поэтому наиболее корректным решением будет удалить строки, содержащие пропуски (примерно 0.4% записей).

Поле balcony – число балконов. Потенциально пропуски здесь допущены у тех квартир, у которых нет балкона, поэтому вероятнее всего целесообразно будет заполнить их нулями.

```
[ ]  df['balcony'].value_counts()

1.0    4195
0.0    3758
2.0    3659
5.0     304
4.0     183
3.0      81
Name: balcony, dtype: int64
```

Рисунок 1.3 – уникальные значения в поле balcony

Поле days_exposition, отвечающее за длительность размещения объявления также обладает пропусками. Так как данное поле будет принимать участие в непосредственном анализе, то замену пропусков логично заменить на медиану.

```
df[df['days_exposition'].isna()]
```

| | total_images | last_price | total_area | first_day_exposition | rooms | ceiling_height | floors_total | living_area | floor | is_apartment | ... | kitchen_area | balcony | locality_name | airports_nearest | cityCenters_nearest | parks_around3000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 20 | 13000000.0 | 108.00 | 2019-03-07T00:00:00 | 3 | 2.70 | 16.0 | 51.0 | 8 | NaN | ... | 25.00 | NaN | Санкт-Петербург | 18863.0 | 16028.0 | 1.0 |
| 7 | 5 | 7915000.0 | 71.60 | 2019-04-18T00:00:00 | 2 | NaN | 24.0 | NaN | 22 | NaN | ... | 18.90 | 2.0 | Санкт-Петербург | 23982.0 | 11634.0 | 0.0 |
| 44 | 13 | 5350000.0 | 40.00 | 2018-11-18T00:00:00 | 1 | NaN | 22.0 | NaN | 3 | NaN | ... | NaN | 1.0 | Санкт-Петербург | 30471.0 | 11603.0 | 1.0 |
| 45 | 17 | 5200000.0 | 50.60 | 2018-12-02T00:00:00 | 2 | 2.65 | 9.0 | 30.3 | 7 | NaN | ... | 7.00 | NaN | Санкт-Петербург | 30011.0 | 12872.0 | 0.0 |
| 46 | 17 | 6600000.0 | 52.10 | 2019-01-31T00:00:00 | 2 | 2.60 | 24.0 | 29.7 | 9 | NaN | ... | 8.30 | 2.0 | Санкт-Петербург | 15114.0 | 12702.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 23684 | 20 | 21400000.0 | 145.00 | 2018-11-02T00:00:00 | 4 | 3.00 | 26.0 | 71.4 | 17 | NaN | ... | 15.60 | NaN | Санкт-Петербург | 11827.0 | 11459.0 | 0.0 |
| 23685 | 15 | 2490000.0 | 31.00 | 2019-01-24T00:00:00 | 1 | 2.50 | 5.0 | 17.3 | 5 | NaN | ... | 5.60 | 1.0 | Ломоносов | 48393.0 | 51818.0 | 0.0 |
| 23694 | 9 | 9700000.0 | 133.81 | 2017-03-21T00:00:00 | 3 | 3.70 | 5.0 | 73.3 | 3 | NaN | ... | 13.83 | NaN | Санкт-Петербург | 24665.0 | 4232.0 | 1.0 |
| 23696 | 18 | 2500000.0 | 56.70 | 2018-02-11T00:00:00 | 2 | NaN | 3.0 | 29.7 | 1 | NaN | ... | NaN | NaN | село Рождествено | NaN | NaN | NaN |
| 23698 | 4 | 1350000.0 | 32.30 | 2017-07-21T00:00:00 | 1 | 2.50 | 5.0 | 12.3 | 1 | NaN | ... | 9.00 | NaN | поселок Новый Учхоз | NaN | NaN | NaN |

3172 rows × 22 columns

Рисунок 1.4 – строки, в которых пропущено значение days_exposition

Также постфактум проведённого анализа было принято решение изменить типы данных некоторых столбцов (например, is_apartment на bool или же first_day_exposition на datetime) в силу нелогичности их хранения в исходном формате.

### 2.2 Алгоритмы и техники

При обработке пропусков, неверных типов данных и дубликатов разработка велась при помощи встроенных методов и возможностей библиотеки pandas. Также некоторые из параметров были обработаны с использованием самостоятельно разработанных функций, дабы разбить исходные данные на категории и провести

более осмысленный анализ. Код к вышеперечисленным методам будет приведён в главе 3.

Для визуализации данных использовалась библиотека matplotlib и seaborn. Функционал данных библиотек является более чем достаточным для решения поставленных задач.

Для формирования модели использовался модуль LinearRegression из библиотеки sklearn.

# 3 Методика решения

## 3.1 Выявление значимых столбцов исходных данных

Анализ данных показал, что наиболее статистически значимыми параметрами являются:

– living_area;

– kitchen_area;

– floor;

– last_price;

– total_area;

– rooms.

Данные показатели позволяют создать модель, способную предопределить рыночную стоимость недвижимости по заданным характеристикам.

## 3.2 Обработка пропусков, аномалий и дубликатов

Постфактум проведённых преобразований исходной информации получили более качественные данные, которые в дальнейшем будут пригодны для анализа. Исходный код обработки данных представлен по ссылке [2] в списке источников.

Результаты зависимости рыночной стоимости квартир от общей площади после очистки данных представлены на диаграмме на Рисунке 1.5

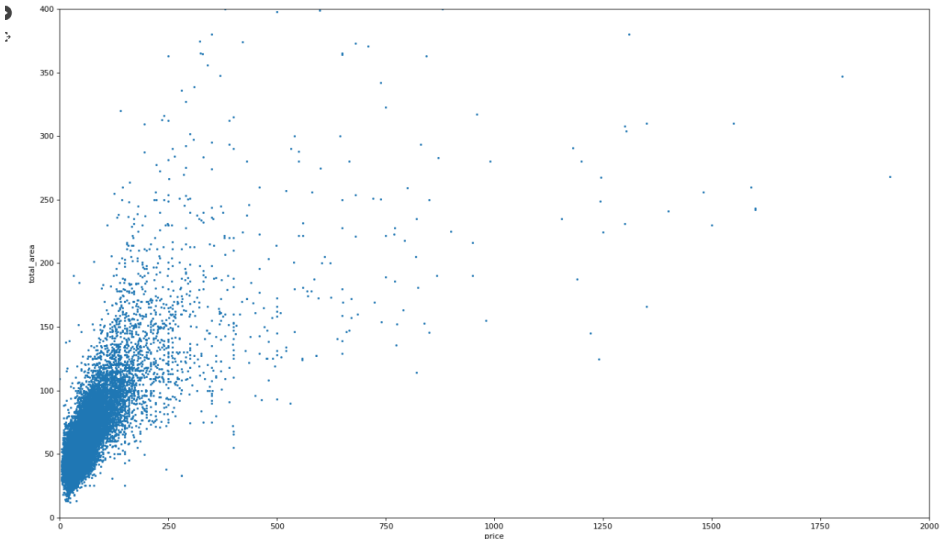


Рисунок 1.5 – диаграмма соотношения площади квартиры со стоимостью

Подытоживая полученные данные имеем, что «средняя квартира», выставленная на продажу, имеет следующие характеристики: это 2-комнатная квартира площадью 52 кв. м с потолками 2,65 м, проданная за 3 месяца (95 дней) по цене 4,7 млн рублей.

Аномальные характеристики квартир из представленного набора данных:

– 6-комнатная квартира и более;

– площадью более 114,3 кв. м;

– с потолками менее 2,25 м или более 3,05 м;

– по цене более 11,9 млн рублей;

### 3.3 Работа с моделью

Для анализа зависимости рыночной стоимости жилья от расстояния от центра города и иных параметров были написаны функции разбиения данных параметров по категориям (см. Рисунок 1.6) (см. ссылку [3] в списке ист.).

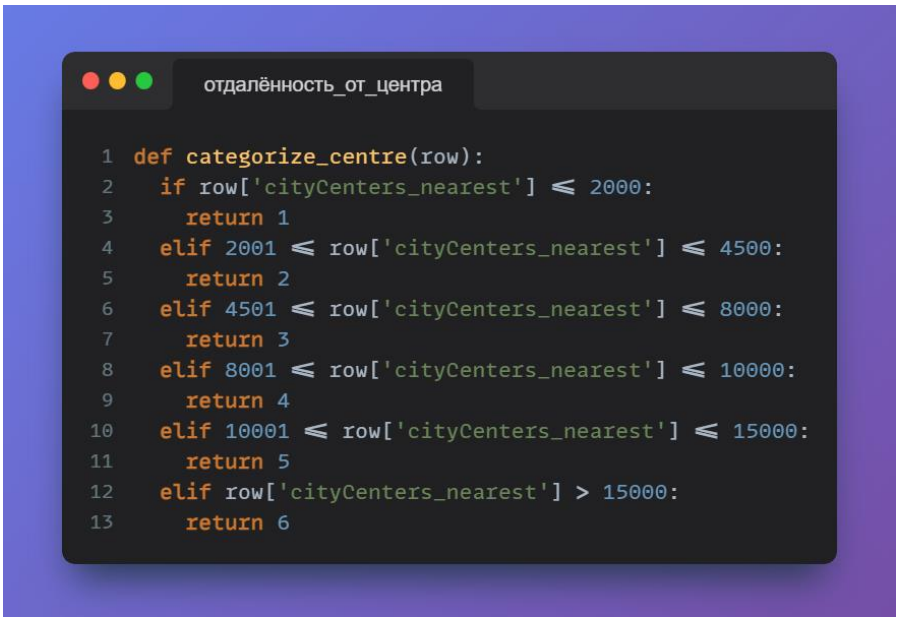


Рисунок 1.6 – функция разбиения отдалённости от центра города по категориям

Исходя из разбиения квартиры, отдалённые от центра не более чем на 8 километров, то есть категории 1, 2 и 3 можно считать близкими к центру и рассматривать как категорию «жильё в центре города».

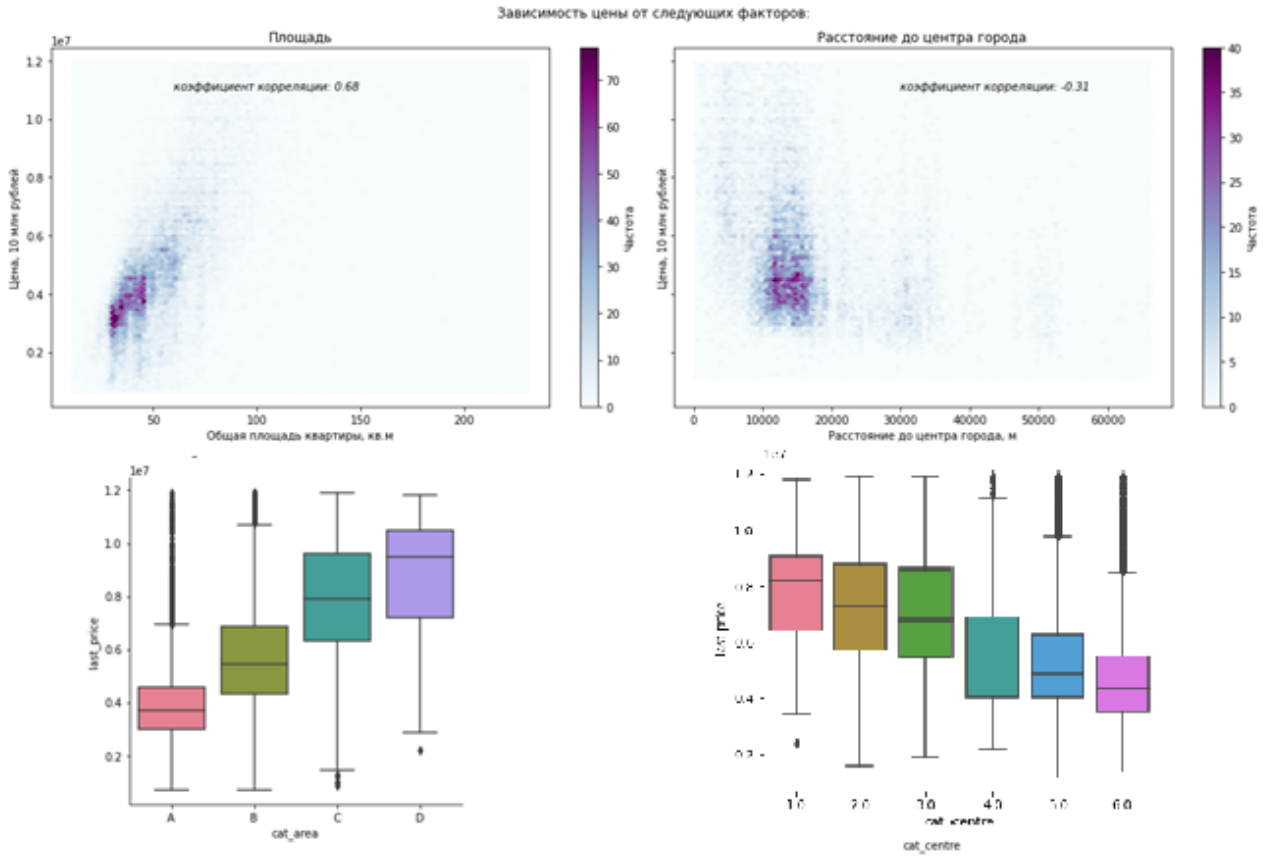


Рисунок 1.7 – визуализация зависимости рыночной стоимости от показателей площади и удалённости от центра города

По результатам проведённого анализа мы можем принять сформированную гипотезу о росте рыночной стоимости недвижимости относительно возрастания площади. Также постфактум отдаления от центра города на расстояние более чем 8 километров наблюдаем спад стоимости недвижимости, что также подтверждает выдвинутую гипотезу. (см. код по ссылке [3] в списке источников).

Далее был проведён анализ интересующих данных, а именно: зависимости между датой размещения объявления и формированием цены. Результаты показали, что взаимосвязь установить не является возможным, как следствие, дата не влияет на рыночную стоимость. (см. Рисунок 1.8)

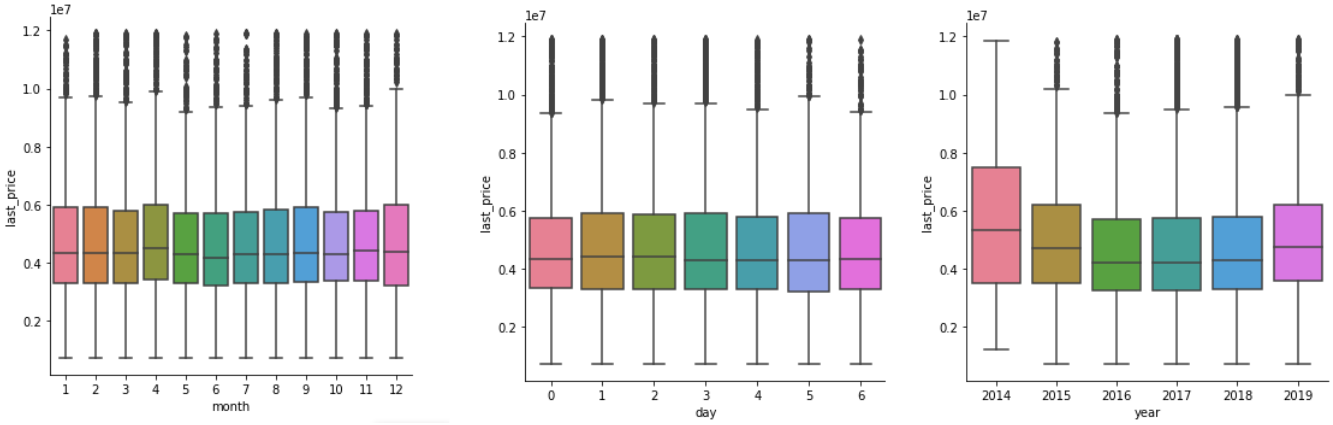


Рисунок 1.8 – показатели по различным дням, месяцам, годам

Зависимость между числом балконов и стоимостью квартир также не была обнаружена, квартиры, имеющие от нуля до двух балконов, входят в один ценовой диапазон, квартиры, имеющие от трёх балконов, в другой, более высокий, но в целом между собой несущественно различимы.

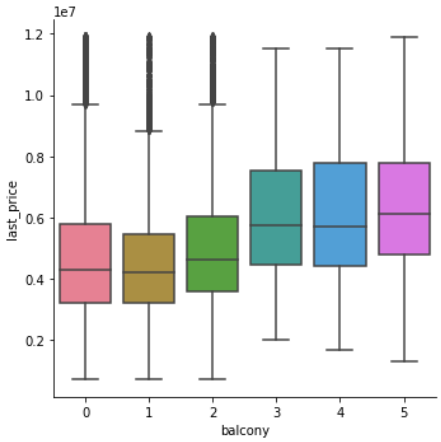


Рисунок 1.9 – показатели стоимости недвижимости с различным числом балконов

Затем была произведена оценка стоимости квартир от этажа, на котором она расположена, для этого был сформирован новый параметр floor_percent, который определяет степень того, насколько высоко расположена по этажам квартира относительно общего числа этажей в доме в процентах. Данный показатель был получен посредством деления параметра floor на floors_total. Затем была произведена категоризация данных значений, при этом квартиры, находящиеся на первом и последнем этажах, были вынесены в отдельные категории. (см. Рисунок 1.10)

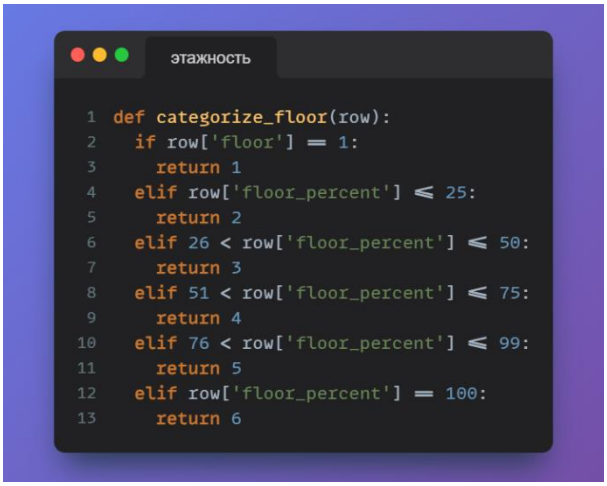


Рисунок 1.10 – функция разбиения квартир по этажному признаку

Анализ полученных результатов позволил сделать вывод, что различия в рыночной стоимости квартир не существенно различаются, чтобы сделать вывод о зависимости стоимости и этажа, но при этом явным образом выделяются квартиры, находящиеся на первом и последних этажах (категории 1 и 6 соответственно), их стоимость ниже, чем стоимость квартир на соседних этажах. (см Рисунок 1.11)

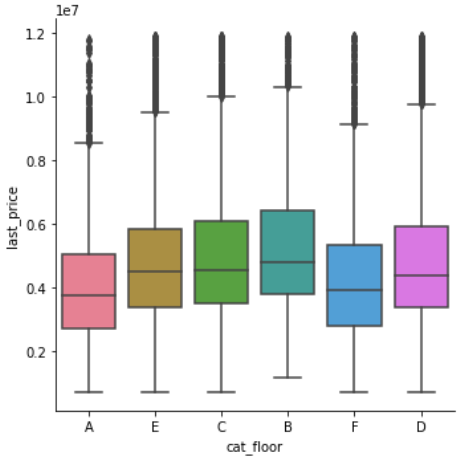


Рисунок 1.11 – показатели стоимости в зависимости от высоты этажа квартиры

Далее была обучена модель LinearRegression из библиотеки sklearn, которая по основным характеристикам жилья позволяет предсказать рыночную стоимость недвижимости. (см. Рисунок 1.12)

```model
X_df = df2[['cat_centre', 'cat_floor', 'rooms', 'total_area']]
y_df = df2['last_price']
X_train_df, X_test_df, y_train_df, y_test_df = train_test_split(X_df, y_df, test_size = 0.2)
model_df = LinearRegression()
model_df.fit(X_train_df, y_train_df)
y_pred_df = model_df.predict(X_test_df)
```

Рисунок 1.12 – формирование модели

В качестве основных характеристик на вход модели подаются: отдалённость от центра города (числовое значение в разбиении по категориям отдалённости), этажность квартиры (числовое значение в разбиение по категориям высоты этажа относительно общего числа этажей в доме), число комнат (числовое значение, определяющее количество комнат в квартире) и общая площадь (числовое значение, определяющее общую площадь квартиры).

Для проверки корректной работы модели был сформирован pandas.DataFrame с тестовыми значениями соответствующих параметров. Обученная ранее модель в итоге сформировала ориентировочную рыночную стоимость данного жилья.

```test
d = {'cat_centre': [1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6],
     'cat_floor': [1, 6, 3, 5, 1, 6, 2, 4, 3, 2, 5, 1],
     'rooms': [2, 2, 4, 1, 2, 2, 3, 2, 1, 2, 4, 2],
     'total_area': [50, 64, 92, 31, 46, 42, 58, 50, 29, 44, 100, 40]}
check_model = pd.DataFrame(data=d)
X_df_price = check_model[['cat_centre', 'cat_floor', 'rooms', 'total_area']]
check_model['last_price'] = model_df.predict(X_df_price)
```

Рисунок 1.14 – формирование тестовых данных для модели

| | cat_centre | cat_floor | rooms | total_area | last_price |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 2 | 50 | 6601839 |
| 1 | 1 | 6 | 2 | 64 | 8194039 |
| 2 | 2 | 3 | 4 | 92 | 9346651 |
| 3 | 2 | 5 | 1 | 31 | 4945801 |
| 4 | 3 | 1 | 2 | 46 | 5339186 |
| 5 | 3 | 6 | 2 | 42 | 5073702 |
| 6 | 4 | 2 | 3 | 58 | 5570292 |
| 7 | 4 | 4 | 2 | 50 | 5415489 |
| 8 | 5 | 3 | 1 | 29 | 3405707 |
| 9 | 5 | 2 | 2 | 44 | 4312410 |
| 10 | 6 | 5 | 4 | 100 | 8531556 |
| 11 | 6 | 1 | 2 | 40 | 3445207 |

Рисунок 1.15 – результат формирования цены недвижимости

## 4 Вывод

В результате проведённого анализа были сформированы гипотезы о взаимосвязи рыночной стоимости недвижимости от её параметров. Сформированы категориальные показатели, позволяющие наиболее точно определить рыночную стоимость недвижимости по её характеристикам.

Исследование показало, что:

- с ростом площади квартир увеличивается их стоимость;
- отдалённость от центра города свыше чем на 8 километров ведёт снижение в стоимости квартиры;
- день и месяц публикации никак не влияют на стоимость;
- этаж не является значимым показателем для стоимости квартиры, кроме первого и последнего этажа (на них стоимость ниже);
- число балконов никак не определяет стоимость.

Также было произведено разбиение исходных параметров по категориям и дальнейшее обучение модели определению рыночной стоимости недвижимости в зависимости от показателей категорий. Данная модель позволяет автоматизировать процесс формирования рыночной стоимости недвижимости, а также выявлять наиболее недооценённое жильё на рынке недвижимости.

## 5 Перечень источников


1. Файл, содержащий исходные данные: https://drive.google.com/file/d/1RIBuZ9DzJZD1tGutl4_shdYXFE9-05KX/view?usp=share_link
2. Предобработка данных https://colab.research.google.com/drive/1QGa5S7ymfhsMw12VbHAQuHAnauOaIo2o#scrollTo=D138JlSgDP4a
3. Проверка гипотез https://colab.research.google.com/drive/1QGa5S7ymfhsMw12VbHAQuHAnauOaIo2o#scrollTo=1w1EuQ524HVW&line=1&uniqifier=1