Аналитический отчет

Симбирцев Владимир Сергеевич

ИСП-21

Введение:

Цель: Собрать данные и провести разведочный исследовательский анализ данных (EDA) для построения модели, которая будет оценивать цену квадратного метра недвижимости в Московском регионе (Москва, Новая Москва, Московская область).

Основная часть:

Для начала аналитического отчета, я импортировал специальные библиотеки для сбора и работы с данными

```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

После этого я написал программу, которая будет создавать новый файл и сохранять туда очищенные данные

```
df = pd.read_csv(r'C:\Users\User\Desktop\VV\merged_file.csv')
print("Исходные данные:")
print(df.head())
```

```
df_cleaned.to_csv(r'cleaned_data.csv', index=False)
```

Анализируем, какие столбцы имеют больше всего пропусков и удаляем их

```
# Function to remove the "object_type" column
def remove_object_type_column(dataframe):
    return dataframe.drop(columns=['object_type'], errors='ignore')

# Function to remove the "house_material_type" column
def remove_house_material_type_column(dataframe):
    return dataframe.drop(columns=['house_material_type'], errors='ignore')

# Function to remove the "heating_type" column
def remove_heating_type_column(dataframe):
    return dataframe.drop(columns=['heating_type'], errors='ignore')

# Function to remove the "finish_type" column
def remove_finish_type_column(dataframe):
```

По окончанию очистки приступил к созданию диаграмм

После этого я сделал проверку на дубликаты, и проверку на пропущенные значения

```
missing_percentages = df_cleaned.isna().mean() * 100
print("Проценты пропущенных значений по столбцам:")
print(missing_percentages)
```

```
duplicate_count = df_cleaned.duplicated().sum()
print(f"Количество дубликатов после удаления: {duplicate_count}")
```

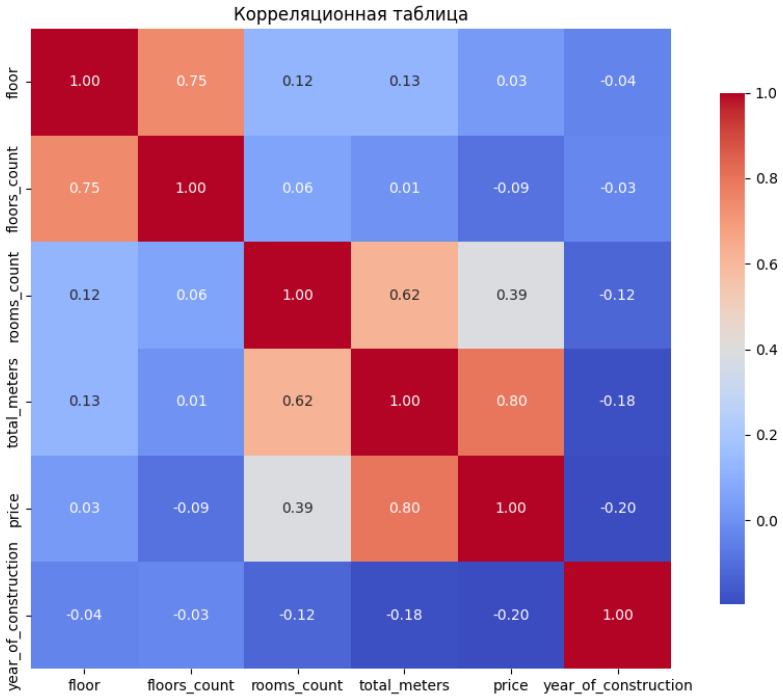
Вывод:

После анализа рынка недвижимости я пришел к выводу, что стоимость за квадратный метр в первую очередь зависит от:

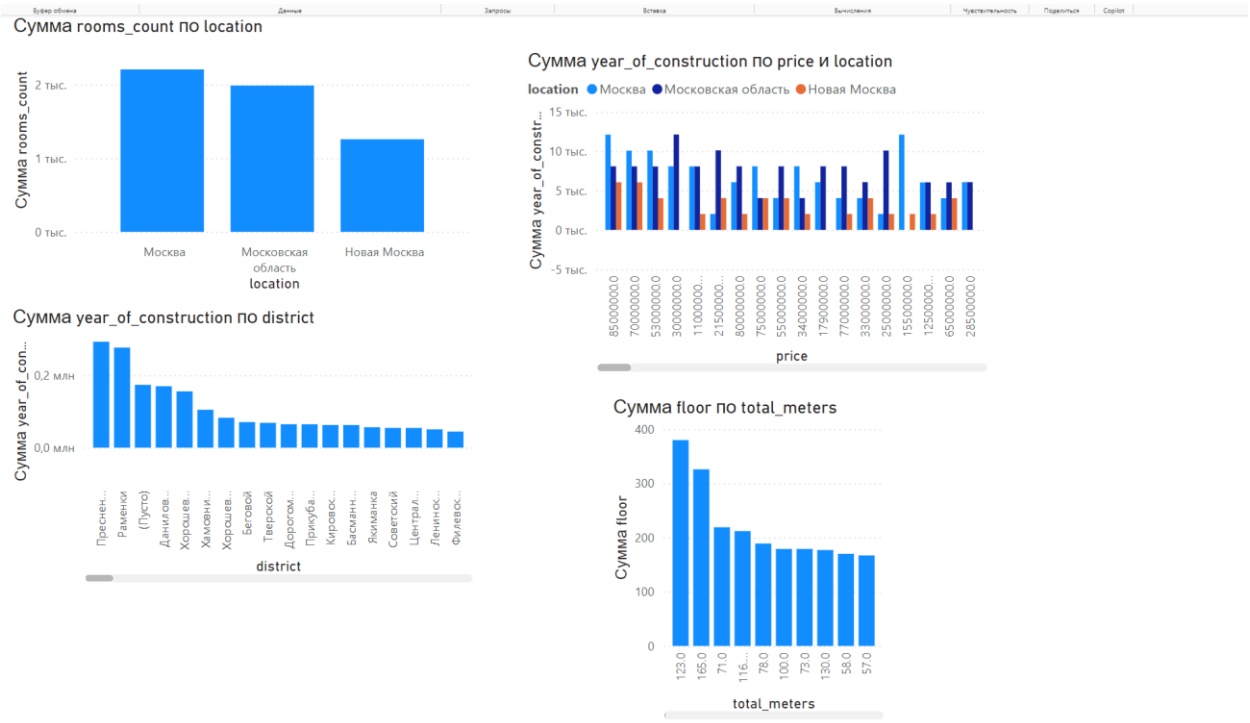
1.Общей площади

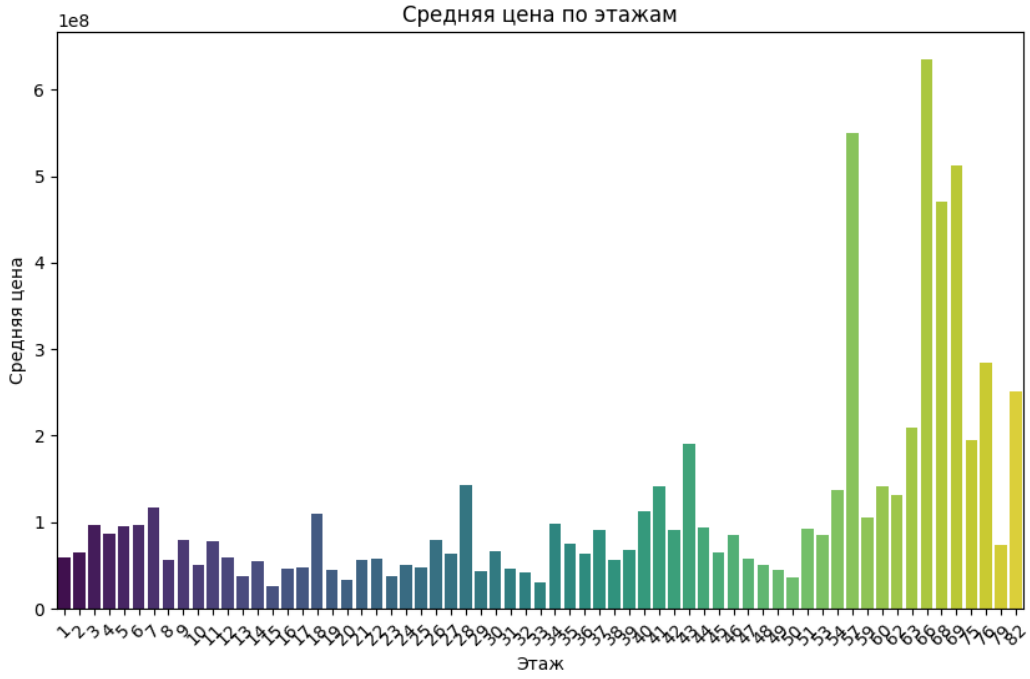
2.Колличества комнат

3.Локации



Так же был проведен анализ с помощью Power Bi





Рефлексия о проделанной работе

По окончанию работы я научился:

1. 1. Понимание данных
   Первым шагом в работе было знакомство с исходными данными.
2. Я изучил структуру DataFrame, определил, какие столбцы могут быть избыточными или несущественными для анализа. Это позволило мне осознать, как важно правильно выбирать данные, которые будут

использоваться в дальнейшем. Удаление ненужных столбцов, таких как 'author' и 'url', помогло сосредоточь

3. 2. Очистка данных
   Процесс очистки данных оказался ключевым этапом. Я реализовал функции для замены некорректных значений, таких как "Москква", и удаления строк с пропущенными значениями в критически важных столбцах. Это не только улучшило качество данных, но и дало мне понимание о том, как ошибки в данных могут повлиять на результаты анализа. Я также научился использовать методы для обработки отсутствующих значений, что является важным навыком в аналитике.

Ошибки при работе:

4. В ходе работы я столкнулся с некоторыми трудностями, которые мне пришлось решить
   1. При очистке данных от лишних столбцов, в самом начале, некоторые значения не удалялись и решение этой проблемы потребовало некоторое время.