Университет ИТМО

Практическая работа №3

по дисциплине «Визуализация и моделирование»

Автор: Горбатовский Алексей Валерьевич
Поток: ВИМ 1.1
Группа: K3220
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Ссылка на датасет: https://www.kaggle.com/sameep98/housing-prices-in-mumbai

Краткое описание датасета:
Датасет содержит в себе информацию о домах в Мумбае, их ценой, площадью и другими характеристиками.
Предобработка для данного датасета как таковая не требуется, т.к. отсутствуют пустые значения и в столбцах,отражающих характеристики домов, уже представлены бинарные значение (0,1).
Была произведена проверка на наличие пустых значений и нетипичных значений в столбцах. Таких не было обнаружено

Проверка на наличие нетипичных значений в стоблцах

```
df_norm["No. of Bedrooms"].unique()
```

```
array([1, 4, 3, 2, 5, 6, 7], dtype=int64)
```

```
df_norm["New/Resale"].unique()
```

```
array([0, 1], dtype=int64)
```

```
df_norm["Gymnasium"].unique()
```

```
array([0, 1], dtype=int64)
```

```
df_norm["Lift Available"].unique()
```

```
array([1, 0], dtype=int64)
```

```
df_norm["Car Parking"].unique()
```

```
array([1, 0], dtype=int64)
```

```
df_norm["Maintenance Staff"].unique()
```

```
array([1, 0], dtype=int64)
```

```
df_norm["24x7 Security"].unique()
```

```
array([1, 0], dtype=int64)
```

```
df_norm["Children's Play Area"].unique()
```

```
array([0, 1], dtype=int64)
```

```
df_norm["Clubhouse"].unique()
```

```
array([0, 1], dtype=int64)
```

```
df_norm["Intercom"].unique()
```

```
array([0, 1], dtype=int64)
```

```
df_norm["Landscaped Gardens"].unique()
```

```
array([0, 1], dtype=int64)
```

Проверка наличия пустых значений в стобцах

```
df_na = {col: list(pd.isna(df[col])).count(True) for col in cols}
df_na
```

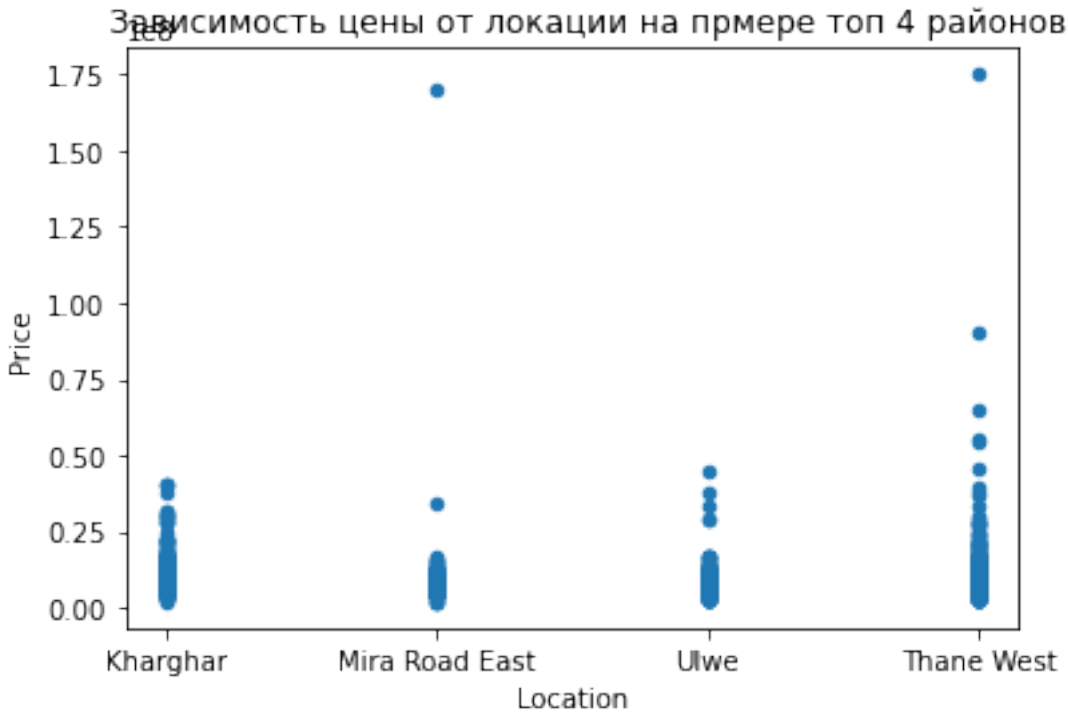
```
{'ID': 0,
 'Price': 0,
 'Area': 0,
 'Location': 0,
 'No. of Bedrooms': 0,
 'New/Resale': 0,
 'Gymnasium': 0,
 'Lift Available': 0,
 'Car Parking': 0,
 'Maintenance Staff': 0,
 '24x7 Security': 0,
 "Children's Play Area": 0,
 'Clubhouse': 0,
 'Intercom': 0,
 'Landscaped Gardens': 0,
 'Indoor Games': 0,
 'Gas Connection': 0,
 'Jogging Track': 0,
 'Swimming Pool': 0}
```

Название столбца	Даннные, хранящиеся в столбце	Тип данных	Шкала
ID	Id дома	Целое число	Номинальная
Price	Цена дома (USD)	Целое число	Относительная
Area	Площадь дома (кв. метры)	Целое число	Относительная
Location	Расположение	Строка	Номинальная
No. of Bedrooms	Количество спален	Целое число	Относительная
New/Resale	Новый/Перепроданный (1/0)	Бинарный	Дихотомическая
Gymnasium	Отсутствие/Наличие (0/1) школы рядом	Бинарный	Дихотомическая
Lift Available	Отсутствие/Наличие (0/1) лифта	Бинарный	Дихотомическая
Car Parking	Отсутствие/Наличие (0/1) парковки	Бинарный	Дихотомическая
Maintenance Staff	Отсутствие/Наличие (0/1) обслуживающего персонала	Бинарный	Дихотомическая
24x7 Security	Отсутствие/Наличие (0/1) охраны 24/7	Бинарный	Дихотомическая
Children's Play Area	Отсутствие/Наличие (0/1) детской площадки	Бинарный	Дихотомическая
Clubhouse	Отсутствие/Наличие (0/1) Ночного клуба	Бинраный	Дихотомическая
Intercom	Отсутствие/Наличие (0/1) внутренней системы коммуникаций	Бинарный	Дихотомическая
Landscaped Gardens	Отсутствие/Наличие (0/1) ландшафтных садов	Бинарный	Дихотомическая
Indoor Games	Отсутствие/Наличие (0/1) спортивной площадки	Бинарный	Дихотомическая
Gas Connection	Отсутствие/Наличие (0/1) газа	Бинарный	Дихотомическая
Jogging Track	Отсутствие/Наличие (0/1) пешеходной тропинки	Бинарный	Дихотомическая
Swimming Pool	Отсутствие/Наличие (0/1) бассейна	Бинарный	Дихотомическая

Гипотезы:

1. Предсказывание цены до по характеристикам (регрессиционная модель машинного обучения), чем разлиных характеристик больше, тем выше цеха, видно из графиков в предыдущей работе.

2. Есть зависимость цены от территории дома, что видно по графику, то есть можно найти оптимальную локацию

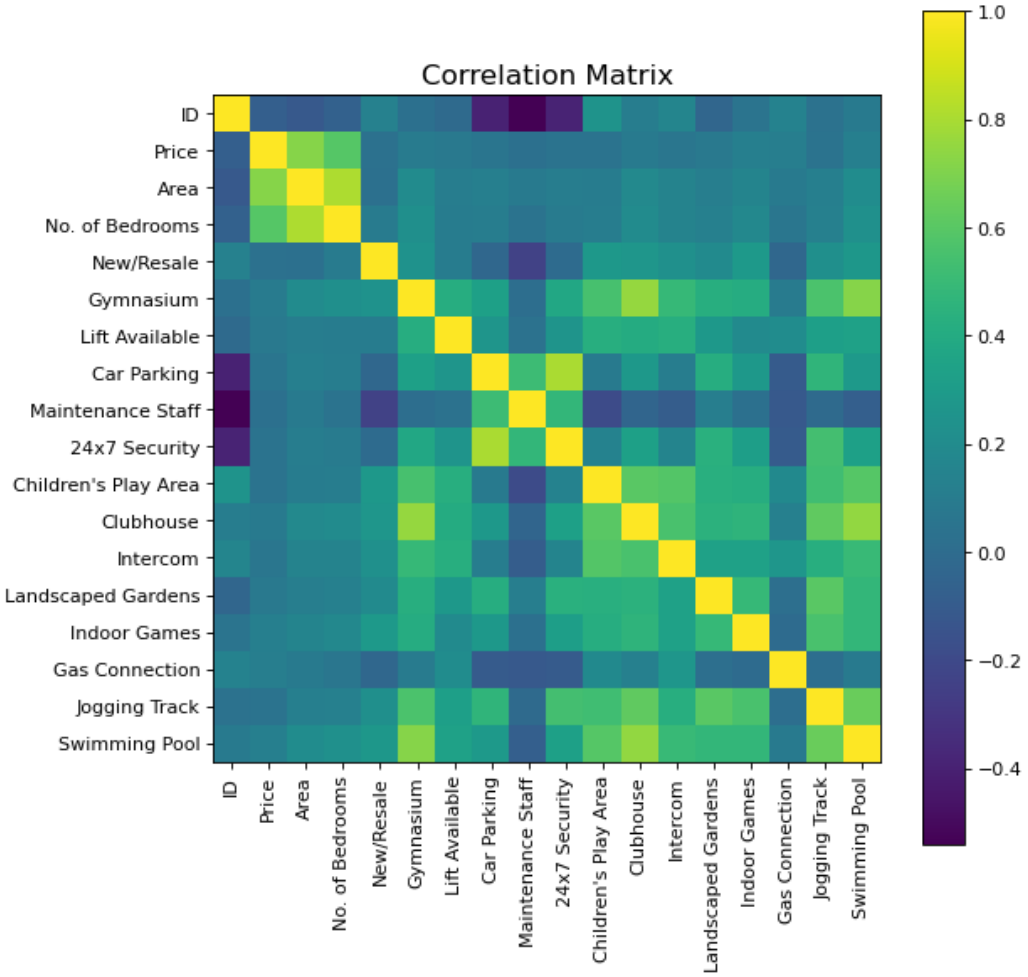


3. Поиск оптимального дома, т.е. наибольшее число характеристик и наибольшая площадь за наименьшую возможную цену

4. Зависимость перепродаваемости дома в зависимости от района и различных характеристик

5. Определенная зависимость между территорией и площадью дома

Доказательство гипотез придставленно корреляционной матрицей


Correlation Matrix
ID
Price
Area
No. of Bedrooms
New/Resale
Gymnasium
Lift Available
Car Parking
Maintenance Staff
24x7 Security
Children's Play Area
Clubhouse
Intercom
Landscaped Gardens
Indoor Games
Gas Connection
Jogging Track
Swimming Pool
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4