РК1

ИУ5-61Б Карпова Ксения

Вариант 8

Задача №1

Для заданного набора данных проведите корреляционный анализ.
В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски.
Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

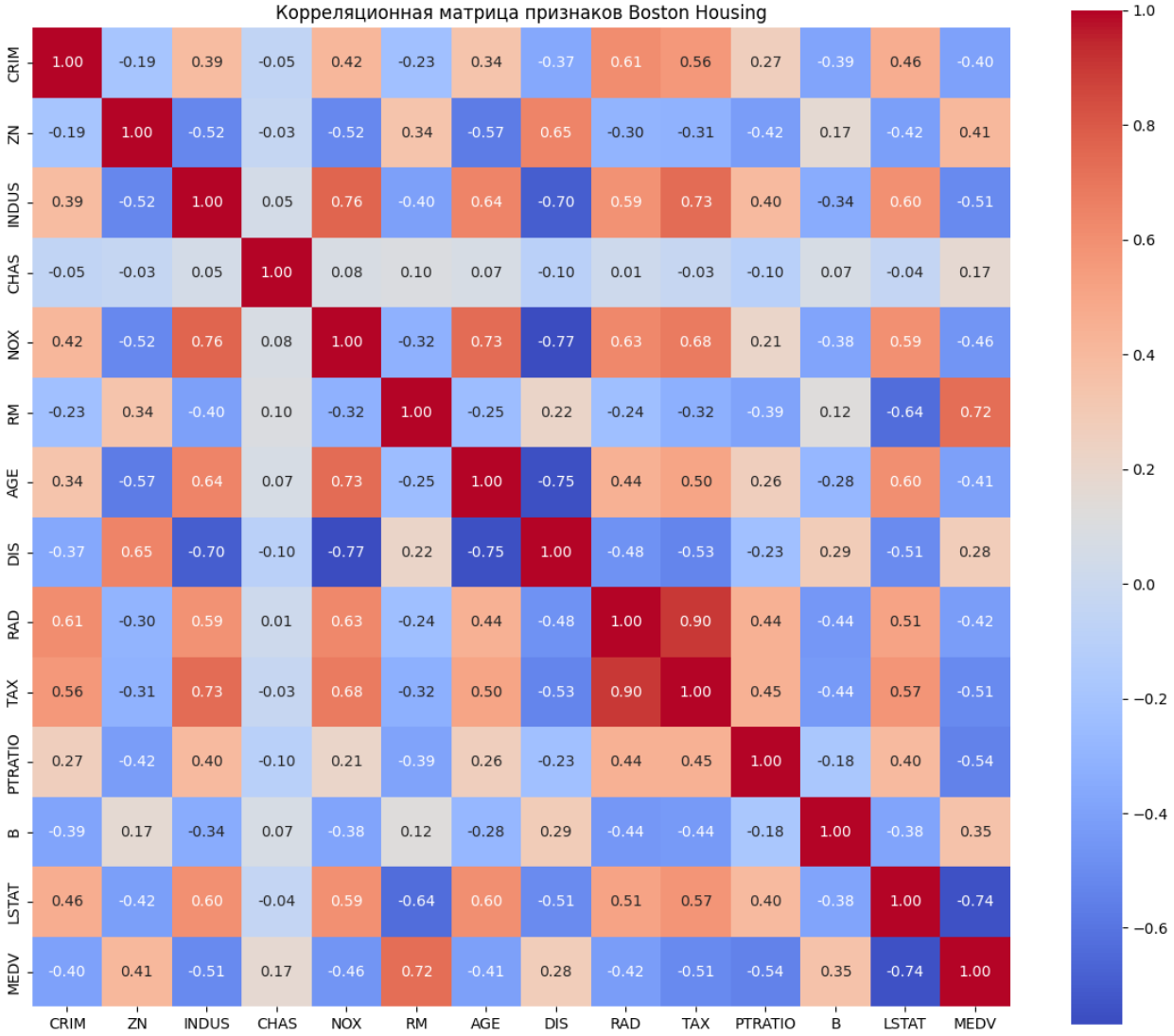
Для пары произвольных колонок данных построить график "Диаграмма рассеяния"

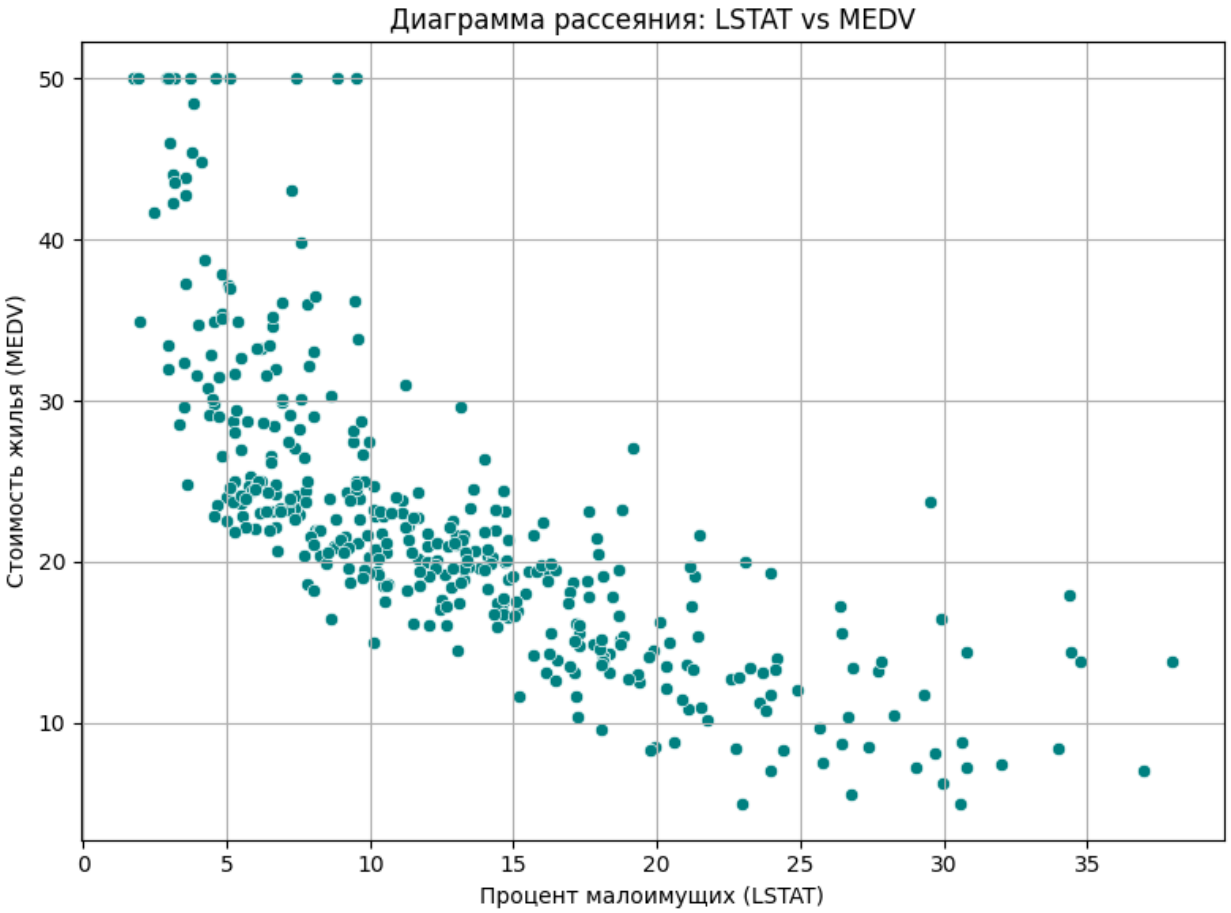
Ссылка на датасет

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# 1. Загрузка данных
df = pd.read_csv("HousingData.csv")

# 2. Проверка пропусков
print("Количество пропусков по колонкам:")
print(df.isnull().sum())

# 3. Удаление строк с пропущенными значениями
df_clean = df.dropna()
print("\nФорма набора после удаления пропусков:", df_clean.shape)

# 4. Корреляционный анализ
corr_matrix = df_clean.corr(numeric_only=True)

# 5. Визуализация корреляционной матрицы
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f",
square=True)
plt.title("Корреляционная матрица признаков Boston Housing")
plt.tight_layout()
plt.show()

# 6. Диаграмма рассеяния для пары признаков: LSTAT и MEDV
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df_clean, x='LSTAT', y='MEDV', color='teal')
```

```
plt.title("Диаграмма рассеяния: LSTAT vs MEDV")
plt.xlabel("Процент малоимущих (LSTAT)")
plt.ylabel("Стоимость жилья (MEDV)")
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
Количество пропусков по колонкам:
CRIM       20
ZN         20
INDUS      20
CHAS       20
NOX         0
RM          0
AGE        20
DIS         0
RAD         0
TAX         0
PTRATIO     0
B           0
LSTAT      20
MEDV        0
dtype: int64

Форма набора после удаления пропусков: (394, 14)
```


Корреляционная матрица признаков Boston Housing
CRIM 1.00 -0.19 0.39 -0.05 0.42 -0.23 0.34 -0.37 0.61 0.56 0.27 -0.39 0.46 -0.40
ZN -0.19 1.00 -0.52 -0.03 -0.52 0.34 -0.57 0.65 -0.30 -0.31 -0.42 0.17 -0.42 0.41
INDUS 0.39 -0.52 1.00 0.05 0.76 -0.40 0.64 -0.70 0.59 0.73 0.40 -0.34 0.60 -0.51
CHAS -0.05 -0.03 0.05 1.00 0.08 0.10 0.07 -0.10 0.01 -0.03 -0.10 0.07 -0.04 0.17
NOX 0.42 -0.52 0.76 0.08 1.00 -0.32 0.73 -0.77 0.63 0.68 0.21 -0.38 0.59 -0.46
RM -0.23 0.34 -0.40 0.10 -0.32 1.00 -0.25 0.22 -0.24 -0.32 -0.39 0.12 -0.64 0.72
AGE 0.34 -0.57 0.64 0.07 0.73 -0.25 1.00 -0.75 0.44 0.50 0.26 -0.28 0.60 -0.41
DIS -0.37 0.65 -0.70 -0.10 -0.77 0.22 -0.75 1.00 -0.48 -0.53 -0.23 0.29 -0.51 0.28
RAD 0.61 -0.30 0.59 0.01 0.63 -0.24 0.44 -0.48 1.00 0.90 0.44 -0.44 0.51 -0.42
TAX 0.56 -0.31 0.73 -0.03 0.68 -0.32 0.50 -0.53 0.90 1.00 0.45 -0.44 0.57 -0.51
PTRATIO 0.27 -0.42 0.40 -0.10 0.21 -0.39 0.26 -0.23 0.44 0.45 1.00 -0.18 0.40 -0.54
B -0.39 0.17 -0.34 0.07 -0.38 0.12 -0.28 0.29 -0.44 -0.44 -0.18 1.00 -0.38 0.35
LSTAT 0.46 -0.42 0.60 -0.04 0.59 -0.64 0.60 -0.51 0.51 0.57 0.40 -0.38 1.00 -0.74
MEDV -0.40 0.41 -0.51 0.17 -0.46 0.72 -0.41 0.28 -0.42 -0.51 -0.54 0.35 -0.74 1.00
CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT MEDV
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6



1. Возможность построения моделей машинного обучения

- Признаки в данных **демонстрируют значимую корреляцию** с целевой переменной `MEDV`, что позволяет использовать их для построения моделей регрессии или деревьев решений.
- Корреляции позволяют оценить **вклад признаков в модель**: чем выше по модулю корреляция с `MEDV`, тем выше потенциал объяснять колебания цены.

---

2. Наиболее важные признаки (по корреляции с `MEDV`)

Признак	Корреляция с `MEDV`	Влияние на модель
`RM`	**+0.72**	Сильное положительное влияние: больше комнат — выше цена
`LSTAT`	**-0.74**	Сильное отрицательное: больше малоимущих — ниже цена
`PTRATIO`	**-0.51**	Умеренное влияние: больше учеников на учителя — ниже цена

Признак	Корреляция с `MEDV`	Влияние на модель
`TAX`	**-0.51**	Высокие налоги — снижение цены жилья
`NOX`	**-0.46**	Загрязнение воздуха — негативно влияет
`INDUS`	**-0.51**	Промзоны снижают привлекательность жилья

3. Взаимные корреляции (мультиколлинеарность)

- **`TAX` и `RAD`**: очень высокая корреляция (**0.90**) — могут дублировать информацию → **рекомендуется выбрать только один** из них.
- `INDUS`, `NOX`, `AGE` — сильно коррелируют между собой → тоже стоит быть осторожным при включении всех сразу в модель.

Итоговый вывод:

Набор данных **подходит для построения регрессионных моделей**, а признаки `LSTAT`, `RM`, `PTRATIO` и `TAX` имеют наибольший вклад в целевую переменную.
При построении модели необходимо учитывать взаимные зависимости между признаками и избегать их дублирования.