

РК1

ИУ5-61Б Холухоев Б. Р.

Вариант 16

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')
```

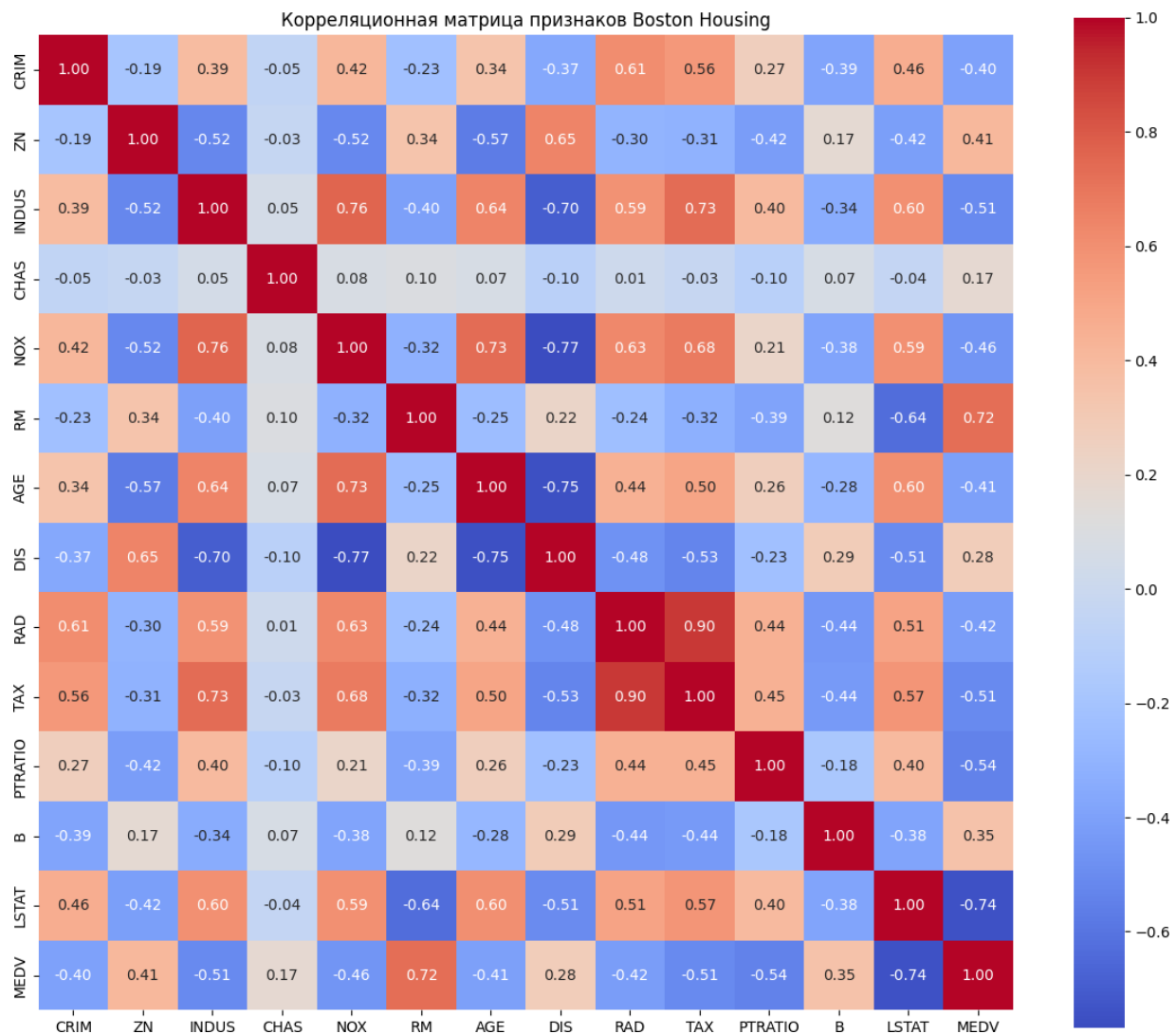
```
plt.title("Диаграмма рассеяния: LSTAT vs MEDV")
plt.xlabel("Процент малоимущих (LSTAT)")
plt.ylabel("Стоимость жилья (MEDV)")
plt.grid(True)
plt.tight_layout()
plt.show()
```

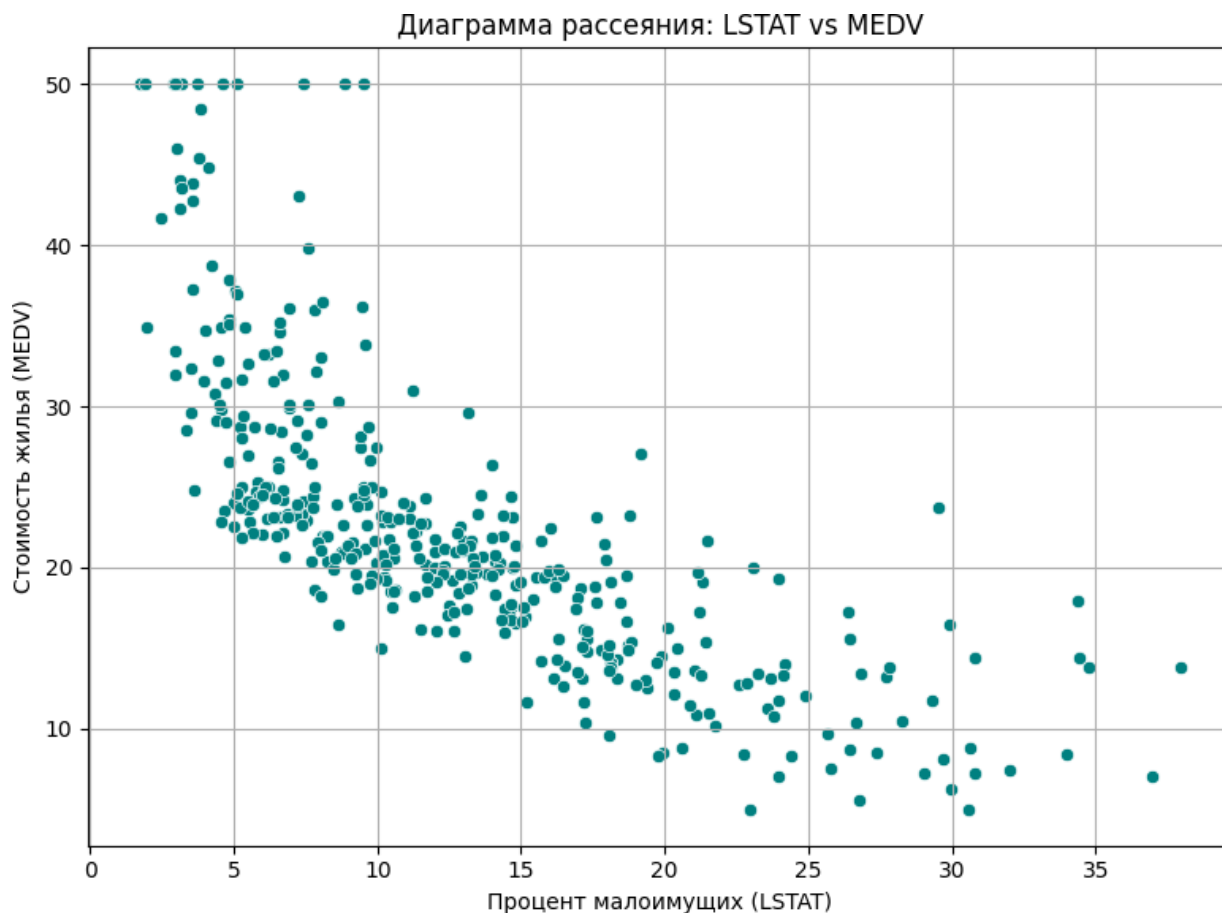
Количество пропусков по колонкам:

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)





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.