

Министерство образования Республики Беларусь
Учреждение образования
«Брестский государственный технический университет»
Кафедра ИИТ

Лабораторная работа №1
По дисциплине: «ОМО»
Тема: «Знакомство с анализом данных: предварительная обработка и
визуализация»

Выполнил:
Студент 3-го курса
Группы АС-65
Нестюк Н.С.
Проверил:
Крощенко А.А.

Цель работы: Получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Вариант 2

Выборка Boston Housing. Содержит информацию о жилье в разных районах Бостона, включая уровень преступности, количество комнат и медианную стоимость.

Задание 1. Для переданной в качестве параметра последовательности из N целых чисел написать утилиту с функционалом: 1) Для данной последовательности, выводит значение «равны» если все элементы последовательности равны и «не равны», в противном случае.

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

%matplotlib inline

sns.set_style("whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)

data_url = "http://lib.stat.cmu.edu/datasets/boston"
raw_df = pd.read_csv(data_url, sep=r"\s+", skiprows=22, header=None)

data = np.hstack([raw_df.values[::2, :], raw_df.values[1::2, :2]])
target = raw_df.values[1::2, 2]

feature_names = ['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE',
                  'DIS', 'RAD', 'TAX', 'PTRATIO', 'N', 'LSTAT']

df = pd.DataFrame(data, columns=feature_names)
df['MEDV'] = target

df.head()
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	N	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

Задание 2. Постройте матрицу корреляции и визуализируйте ее с помощью тепловой карты (heatmap).

Задание 3. Найдите признак, наиболее сильно коррелирующий с целевой переменной MEDV (медианная стоимость дома).

```
correlation_matrix = df.corr().round(2)

plt.figure(figsize=(9, 9))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)
plt.title("Матрица корреляции признаков Boston Housing")
plt.tight_layout()
plt.show()
```

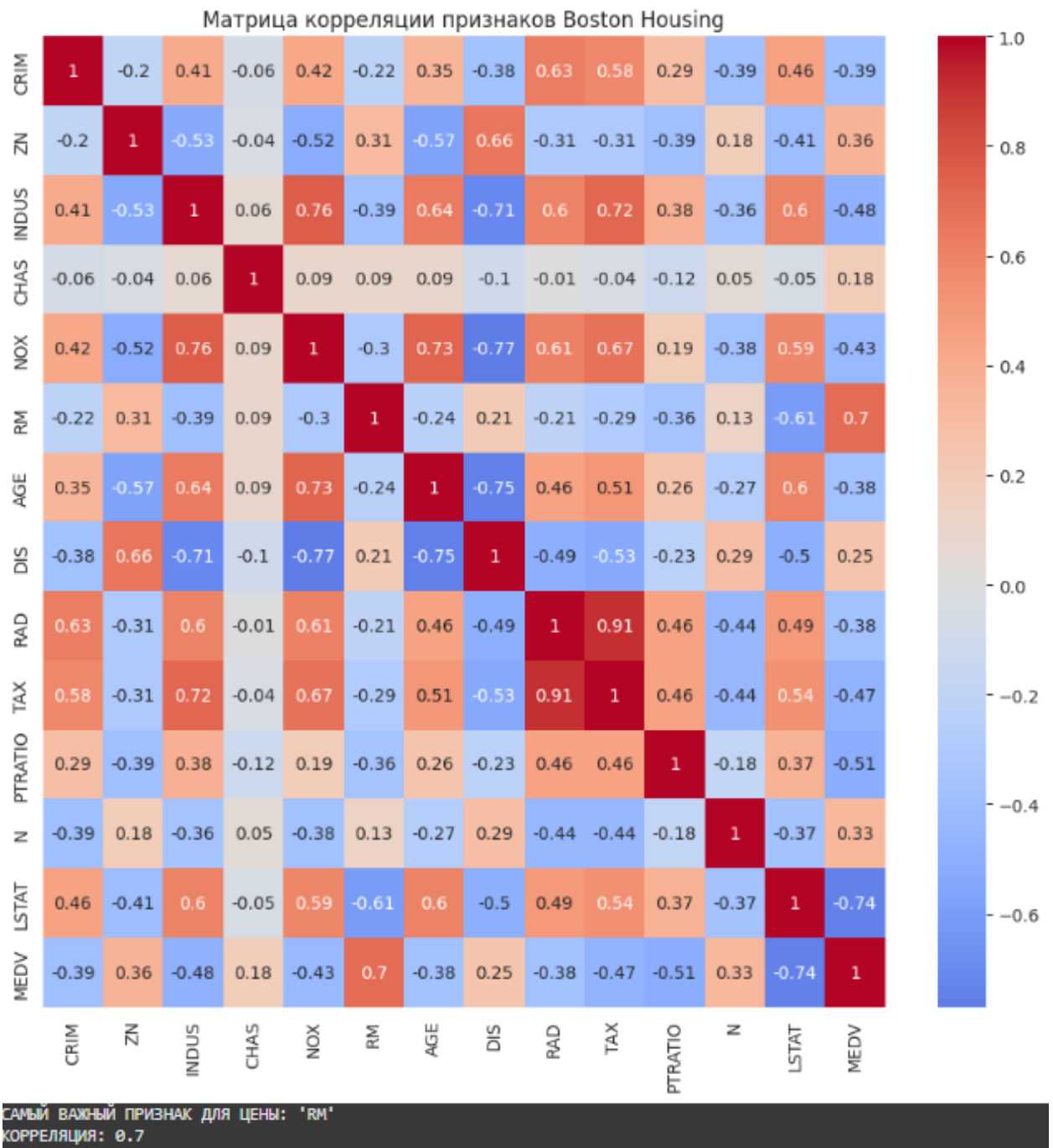
```

correlation_with_target = correlation_matrix['MEDV'].sort_values(ascending=False)

most_important_feature = correlation_with_target.index[1]
correlation_value = correlation_with_target.iloc[1]

print(f"САМЫЙ ВАЖНЫЙ ПРИЗНАК ДЛЯ ЦЕНЫ: '{most_important_feature}'")
print(f"КОРРЕЛЯЦИЯ: {correlation_value}")

```



Задание 4. Постройте диаграмму рассеяния (scatter plot) для этого признака и MEDV

```

correlation_with_target = df.corr()['MEDV'].sort_values(ascending=False)
most_correlated_feature = correlation_with_target.index[1]
highest_correlation = correlation_with_target.iloc[1]

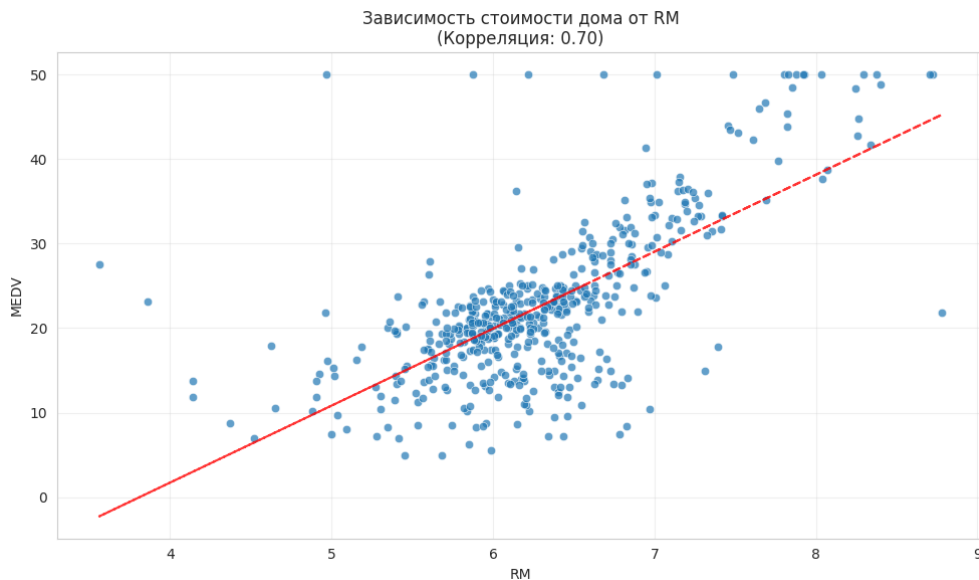
plt.figure(figsize=(10, 6))
plt.scatter(df[most_correlated_feature], df['MEDV'], alpha=0.7, edgecolors='w',
            linewidth=0.5)
plt.xlabel(most_correlated_feature)

```

```
plt.ylabel('MEDV')
plt.title(f'Зависимость стоимости дома от {most_correlated_feature}\n(Корреляция: {highest_correlation:.2f})')

z = np.polyfit(df[most_correlated_feature], df['MEDV'], 1)
p = np.poly1d(z)
plt.plot(df[most_correlated_feature], p(df[most_correlated_feature]), "r--",
alpha=0.8)

plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```



Задание 5. Нормализуйте все числовые признаки, приведя их к диапазону от 0 до 1.
`from sklearn.preprocessing import MinMaxScaler`

```
scaler = MinMaxScaler()

df_normalized = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)

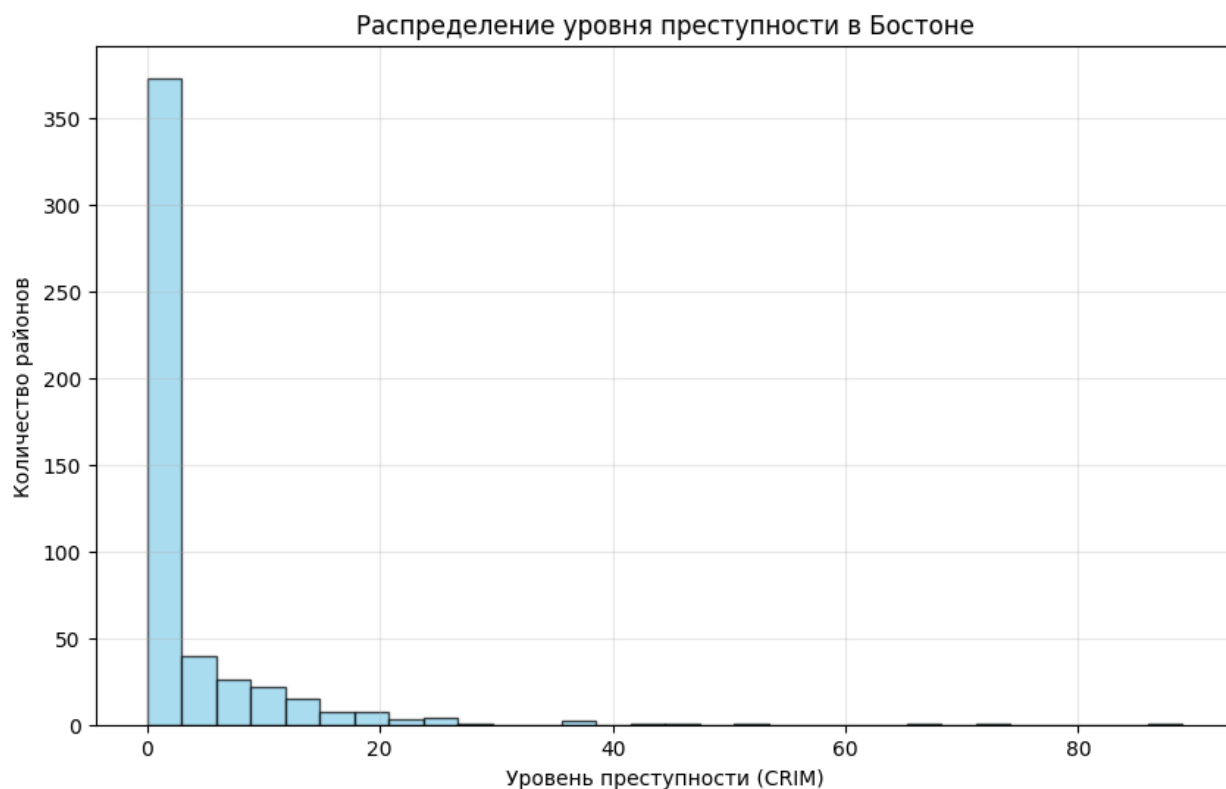
print(df_normalized.head())
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	\
0	0.000000	0.18	0.067815	0.0	0.314815	0.577505	0.641607	0.269203	
1	0.000236	0.00	0.242302	0.0	0.172840	0.547998	0.782698	0.348962	
2	0.000236	0.00	0.242302	0.0	0.172840	0.694386	0.599382	0.348962	
3	0.000293	0.00	0.063050	0.0	0.150206	0.658555	0.441813	0.448545	
4	0.000705	0.00	0.063050	0.0	0.150206	0.687105	0.528321	0.448545	
	RAD	TAX	PTRATIO	N	LSTAT	MEDV			
0	0.000000	0.208015	0.287234	1.000000	0.089680	0.422222			
1	0.043478	0.104962	0.553191	1.000000	0.204470	0.368889			
2	0.043478	0.104962	0.553191	0.989737	0.063466	0.660000			
3	0.086957	0.066794	0.648936	0.994276	0.033389	0.631111			
4	0.086957	0.066794	0.648936	1.000000	0.099338	0.693333			

Задание 6. Визуализируйте распределение уровня преступности (CRIM) с помощью гистограммы.

```
plt.figure(figsize=(10, 6))
plt.hist(df['CRIM'], bins=30, edgecolor='black', alpha=0.7, color='skyblue')
plt.xlabel('Уровень преступности (CRIM)')
plt.ylabel('Количество районов')
plt.title('Распределение уровня преступности в Бостоне')
plt.grid(True, alpha=0.3)
```

```
plt.show()
```



Вывод: Получил практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научился выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.