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:
Студенты 3-го курса
Группы АС-65
Осовец М. М.
Проверил:
Крощенко А. А.

Брест 2025

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 3

Выборка Iris. Классический набор данных для классификации, содержащий измерения длины и ширины чашелистиков и лепестков для трех видов ирисов.

Задачи:

1. Загрузите данные и проверьте, есть ли в них пропущенные

значения.

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler


# Загружаем CSV (обязательно должен лежать рядом в папке)

df = pd.read_csv('iris.csv')



# Первые 5 строк

df.head()
```

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

Проверка на пропуски

```
df.isnull().sum()
```

```
sepal.length    0
sepal.width     0
petal.length    0
petal.width     0
variety         0
dtype: int64
```

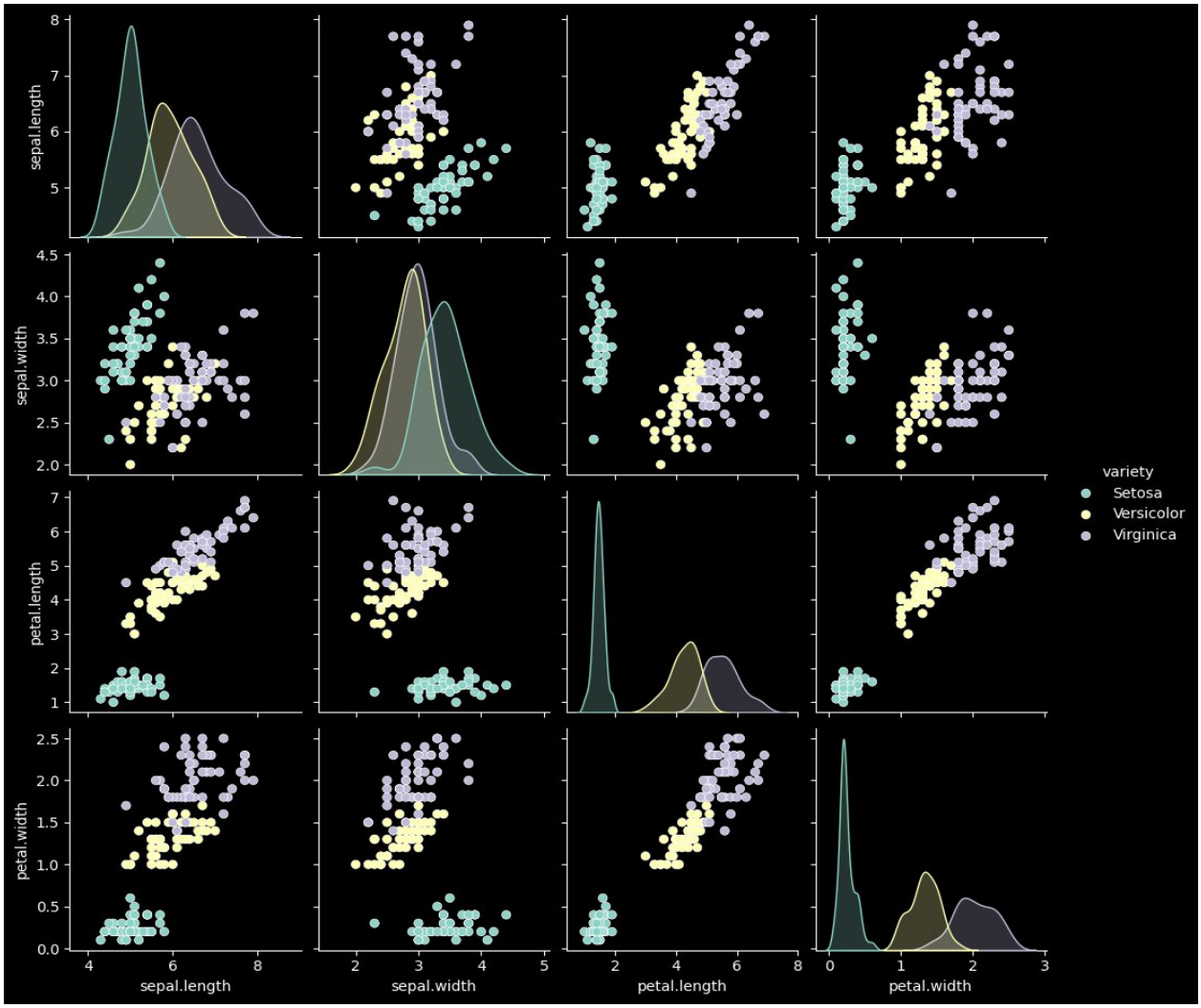
2. Выведите количество образцов каждого вида ириса.

```
df['variety'].value_counts()
```

```
variety
Setosa        50
Versicolor    50
Virginica     50
Name: count, dtype: int64
```

3. Постройте парные диаграммы рассеяния (pair plot) для всех признаков, чтобы визуально оценить их разделимость.

```
sns.pairplot(df, hue='variety')
plt.show()
```



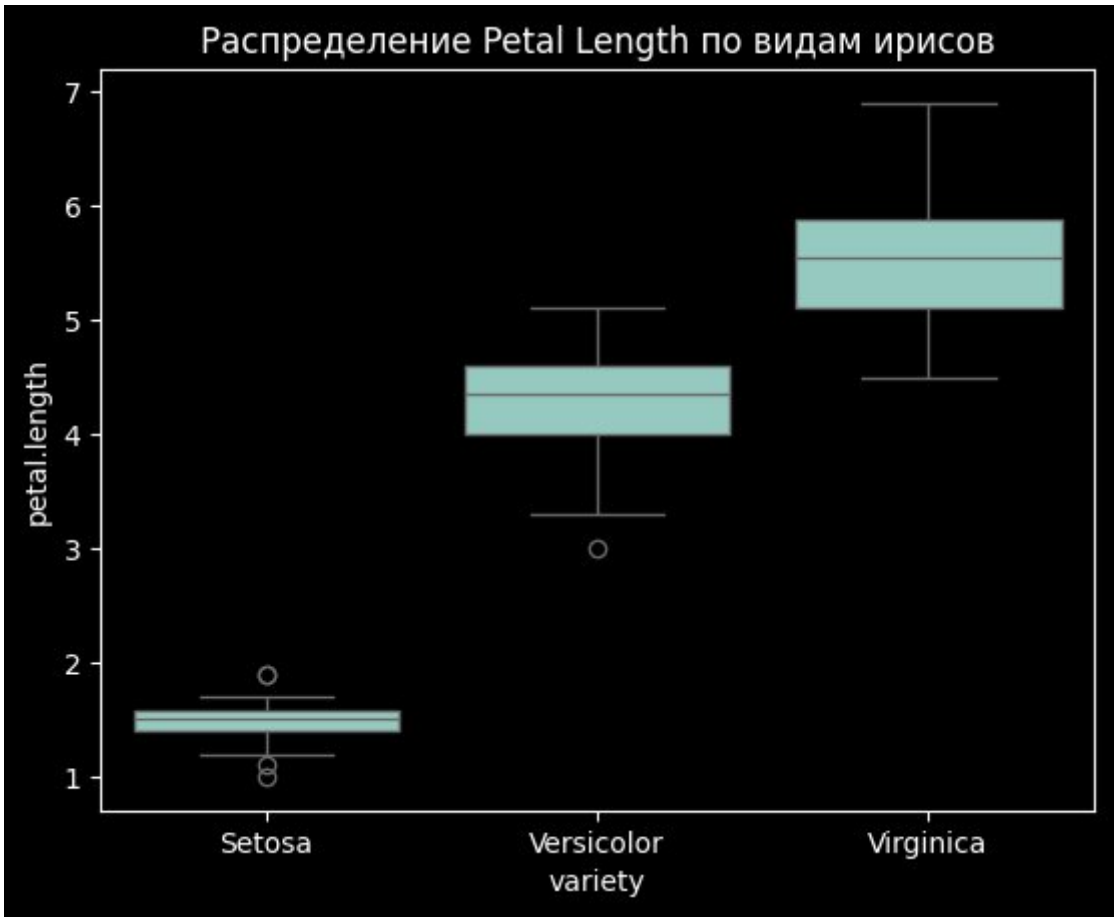
4. Для каждого вида ириса рассчитайте среднее значение по каждому из четырех признаков.

```
df.groupby('variety').mean()
```

variety	sepal.length	sepal.width	petal.length	petal.width
Setosa	5.006	3.428	1.462	0.246
Versicolor	5.936	2.770	4.260	1.326
Virginica	6.588	2.974	5.552	2.026

5. Создайте "ящик с усами" (box plot) для признака Petal Length (cm), чтобы сравнить его распределение по разным видам ирисов.

```
sns.boxplot(x='variety', y='petal.length', data=df)
plt.title('Распределение Petal Length по видам ирисов')
plt.show()
```



6. Стандартизируйте данные (приведите к нулевому среднему и единичному стандартному отклонению).

```
features = df.drop(columns=['variety'])
```

```
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

df_scaled = pd.DataFrame(scaled_features, columns=features.columns)
df_scaled.head()
```

	sepal.length	sepal.width	petal.length	petal.width
0	-0.900681	1.019004	-1.340227	-1.315444
1	-1.143017	-0.131979	-1.340227	-1.315444
2	-1.385353	0.328414	-1.397064	-1.315444
3	-1.506521	0.098217	-1.283389	-1.315444
4	-1.021849	1.249201	-1.340227	-1.315444

Вывод:

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