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
Гуща И.В.
Вариант 3
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.

Ход работы

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 numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler

file_path = '/content/iris.csv'

# 1. Загрузка  и проверка
df = pd.read_csv(file_path)
print("Пропущенных значений :")
print(df.isnull().sum())
```

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

2. Выведите количество образцов каждого вида ириса.

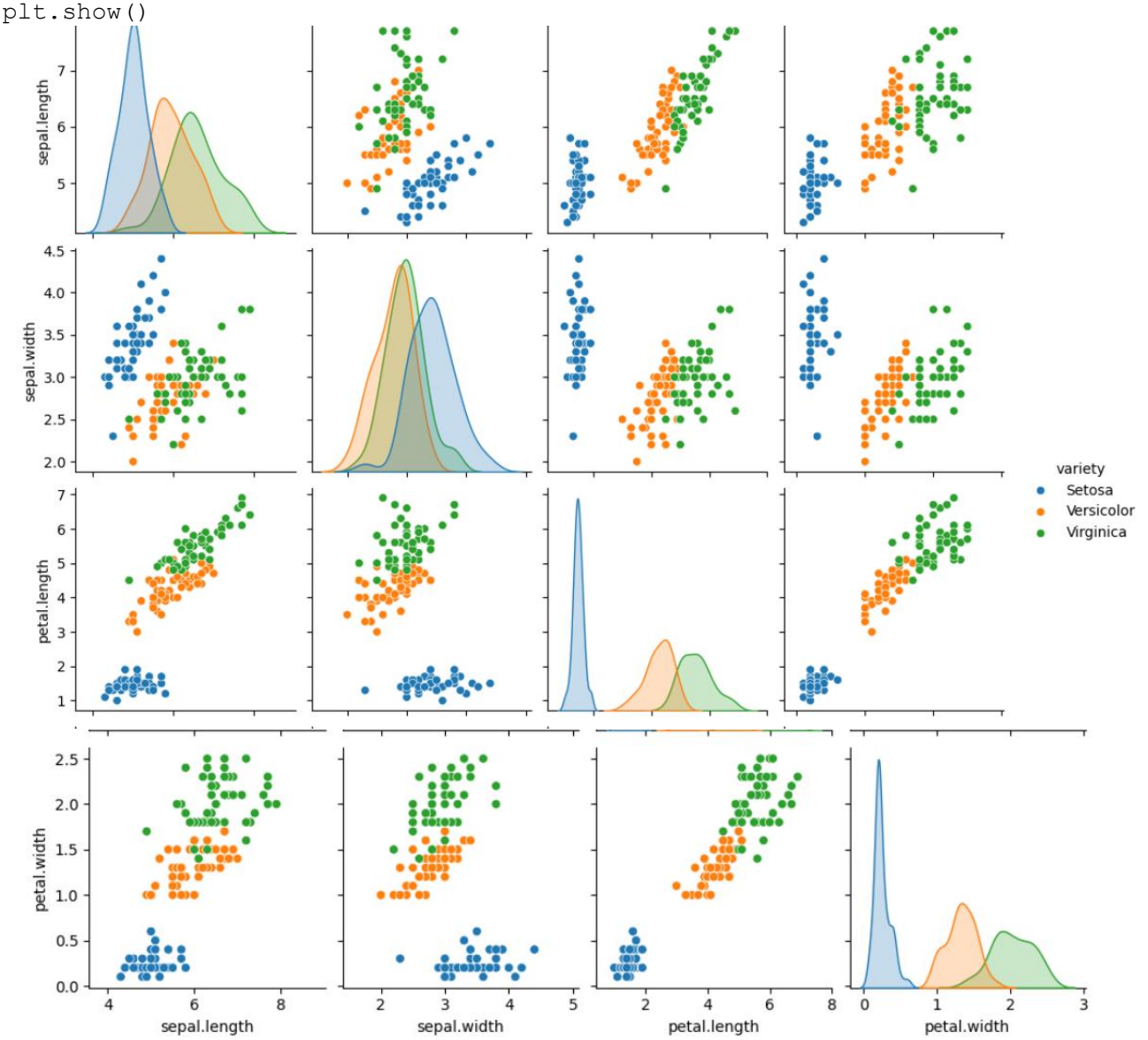
```
# 2. Количество образцов каждого вида
print("Количество образцов каждого вида:")
counts = {}  # пустой словарь для подсчёта

for item in df['variety']:
    if item in counts:
        counts[item] += 1
    else:
        counts[item] = 1
for key, value in counts.items():
    print(f"{key}: {value}")
```

```
Количество образцов каждого вида:
Setosa: 50
Versicolor: 50
Virginica: 50
```

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

```
# 3. Парные диаграммы рассеяния
sns.pairplot(df, hue='variety')
plt.show()
```



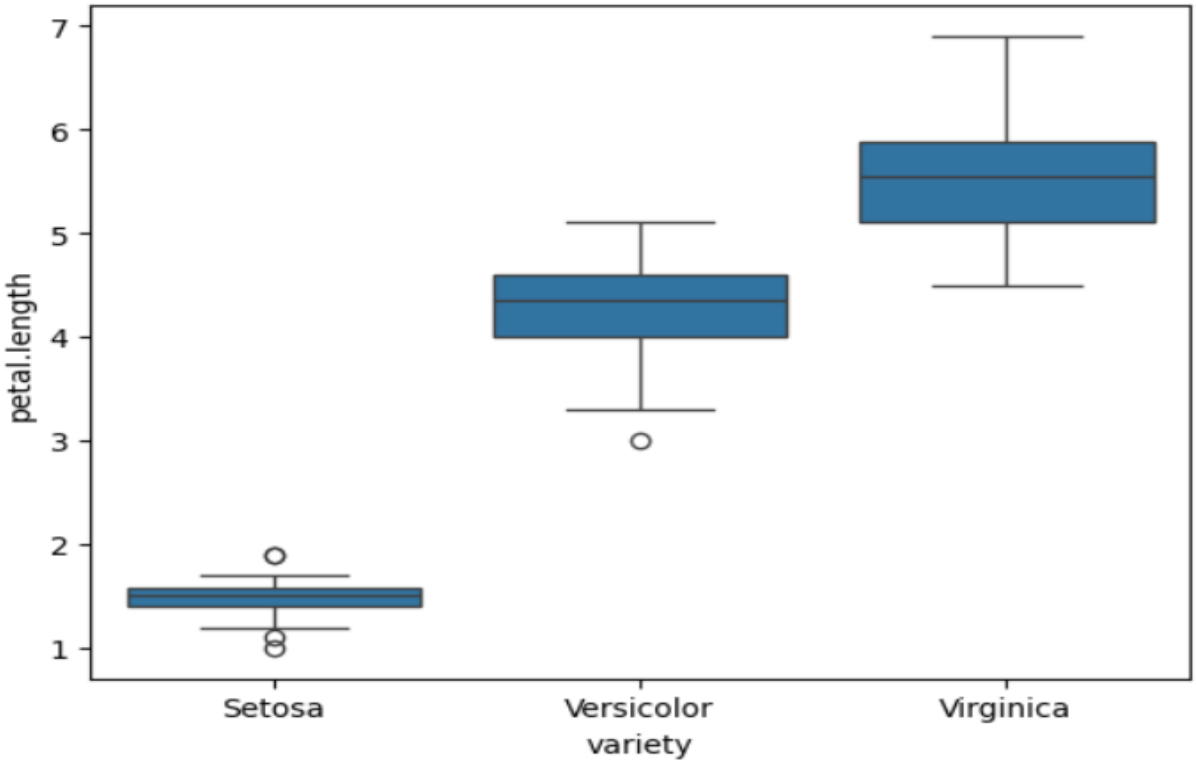
4. Для каждого вида ириса рассчитайте среднее значение по каждому из четырех признаков.

```
# 4. Средние значения
mean= df.groupby('variety').mean()
print(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.

```
# 5. Ящик с усами
sns.boxplot(x='variety', y='petal.length', data=df)
plt.show()
```



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

```
# 6. Стандартизация данных
features = ['sepal.length', 'sepal.width', 'petal.length', 'petal.width']
scaler = StandardScaler()
df_scaled = df.copy()
df_scaled[features] = scaler.fit_transform(df[features])
print(df_scaled.to_string())
```

	sepal.length	sepal.width	petal.length	petal.width	variety
0	-0.900681	1.019004	-1.340227	-1.315444	Setosa
1	-1.143017	-0.131979	-1.340227	-1.315444	Setosa
2	-1.385353	0.328414	-1.397064	-1.315444	Setosa
3	-1.506521	0.098217	-1.283389	-1.315444	Setosa
4	-1.021849	1.249201	-1.340227	-1.315444	Setosa
..	...	...	...	...	...
145	1.038005	-0.131979	0.819596	1.448832	Virginica
146	0.553333	-1.282963	0.705921	0.922303	Virginica
147	0.795669	-0.131979	0.819596	1.053935	Virginica
148	0.432165	0.788808	0.933271	1.448832	Virginica
149	0.068662	-0.131979	0.762758	0.790671	Virginica

Вывод: мы приобрели практические знания по работе с Pandas, Matplotlib, а также научились анализировать датасеты для дальнейшего обучения моделей на их основе.