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-66
Янчук А.Ю.
Проверил:
Крощенко А.А.

Брест 2025

Цель: 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 13

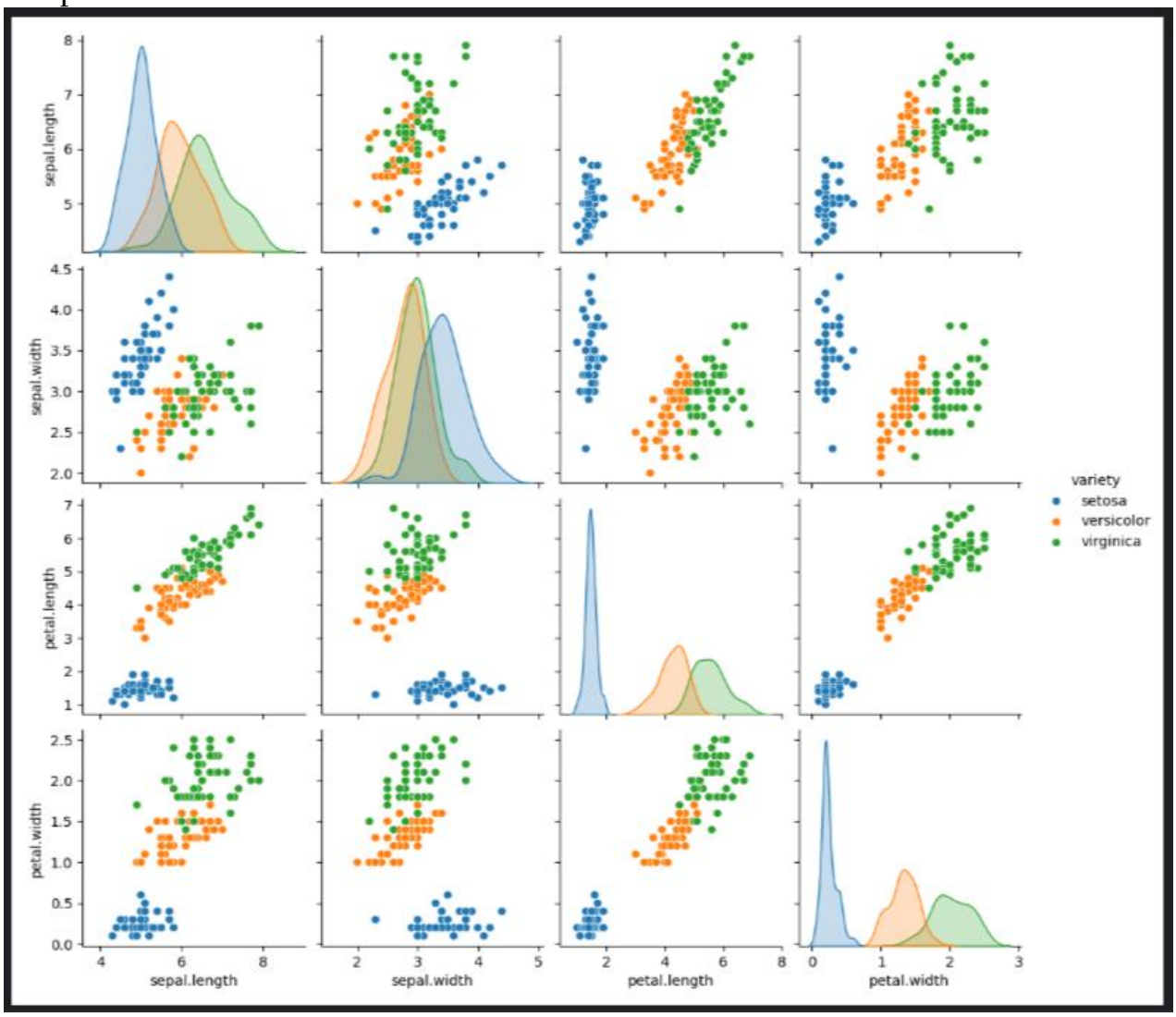
Выборка Iris. Классический набор данных для классификации, содержащий измерения длины и ширины чашелистиков и лепестков для трех видов ирисов. Задачи: 1. значения. 2. 3. Загрузите данные и проверьте, есть ли в них пропущенные Выведите количество образцов каждого вида ириса. Постройте парные диаграммы рассеяния (pair plot) для всех признаков, чтобы визуально оценить их разделимость. 4. Для каждого вида ириса рассчитайте среднее значение по каждому из четырех признаков. 5. Создайте "ящик с усами" (box plot) для признака Petal Length (cm), чтобы сравнить его распределение по разным видам ирисов. 6. Стандартизируйте данные (приведите к нулевому среднему и единичному стандартному отклонению).

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pathlib import Path

iris = sns.load_dataset("iris")

iris.rename(columns={
    "sepal_length":  "sepal.length",
    "sepal_width":   "sepal.width",
    "petal_length":  "petal.length",
    "petal_width":   "petal.width",
    "species":       "variety"
}, inplace=True)

print("Shape:", iris.shape)
print("Columns:", iris.columns)
print(iris.head())

missing = iris.isnull().sum()
counts  = iris["variety"].value_counts()
means   = iris.groupby("variety").mean(numeric_only=True)

features = ["sepal.length", "sepal.width", "petal.length", "petal.width"]
iris_scaled_df = iris.copy()
iris_scaled_df[features] = (iris[features] - iris[features].mean()) /
iris[features].std()

iris_encoded = pd.get_dummies(iris_scaled_df, columns=["variety"])

report_path = Path(__file__).parent / "iris_report.txt"
with report_path.open("w", encoding="utf-8") as f:
    f.write("Исходные данные (все строки):\n")
    f.write(iris.to_string(index=False) + "\n\n")
    f.write("Проверка пропущенных значений:\n")
    f.write(missing.to_string() + "\n\n")
    f.write("Количество образцов по каждому виду:\n")
    f.write(counts.to_string() + "\n\n")
```

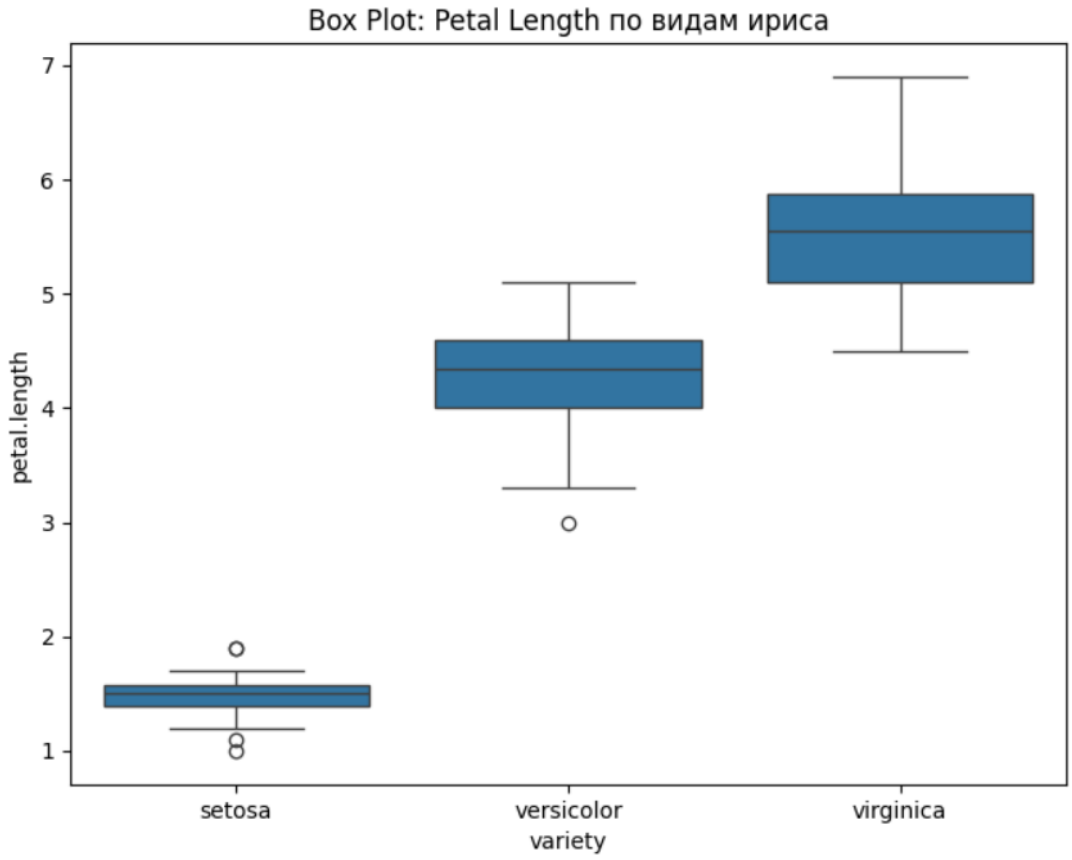
```
    f.write("Средние значения признаков по каждому виду:\n")
    f.write(means.to_string() + "\n\n")
    f.write("Стандартизованные данные (первые 5 строк):\n")
    f.write(iris_scaled_df.head().to_string(index=False) + "\n\n")
    f.write("One-Hot Encoding (первые 5 строк):\n")
    f.write(iris_encoded.head().to_string(index=False) + "\n\n")
print(f"Отчёт сохранён в {report_path.name}")

sns.pairplot(iris, hue="variety", diag_kind="kde")
plt.suptitle("Pair Plot признаков Iris", y=1.02)
plt.savefig("pairplot.png")
plt.close()

plt.figure(figsize=(8, 6))
sns.boxplot(x="variety", y="petal.length", data=iris)
plt.title("Box Plot: Petal Length по видам ириса")
plt.savefig("boxplot.png")
plt.close()

print("Графики сохранены в pairplot.png и boxplot.png")
```

```
Shape: (150, 5)
Columns: Index(['sepal.length', 'sepal.width', 'petal.length', 'petal.width',
       'variety'],
      dtype='object')
   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
Отчёт сохранён в iris_report.txt
Графики сохранены в pairplot.png и boxplot.png
```

Графики:
Pairplot:



Boxplot:


Box Plot: Petal Length по видам ириса
petal.length
7
6
5
4
3
2
1
setosa
versicolor
virginica
variety

```
Проверка пропущенных значений:
sepal.length    0
sepal.width     0
petal.length    0
petal.width     0
variety         0

Количество образцов по каждому виду:
variety
setosa        50
versicolor    50
virginica     50

Средние значения признаков по каждому виду:
            sepal.length  sepal.width  petal.length  petal.width
variety
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

Стандартизованные данные (первые 5 строк):
 sepal.length  sepal.width  petal.length  petal.width variety
    -0.897674     1.015602     -1.335752    -1.311052  setosa
    -1.139200    -0.131539     -1.335752    -1.311052  setosa
    -1.380727     0.327318     -1.392399    -1.311052  setosa
    -1.501490     0.097889     -1.279104    -1.311052  setosa
    -1.018437     1.245030     -1.335752    -1.311052  setosa

One-Hot Encoding (первые 5 строк):
 sepal.length  sepal.width  petal.length  petal.width  variety_setosa  variety_versicolor  variety_virginica
    -0.897674     1.015602     -1.335752    -1.311052            True               False              False
    -1.139200    -0.131539     -1.335752    -1.311052            True               False              False
    -1.380727     0.327318     -1.392399    -1.311052            True               False              False
    -1.501490     0.097889     -1.279104    -1.311052            True               False              False
    -1.018437     1.245030     -1.335752    -1.311052            True               False              False
```

```
Исходные данные (все строки):
    sepal.length  sepal.width  petal.length  petal.width     variety
              5.1          3.5           1.4          0.2      setosa
              4.9          3.0           1.4          0.2      setosa
              4.7          3.2           1.3          0.2      setosa
              4.6          3.1           1.5          0.2      setosa
              5.0          3.6           1.4          0.2      setosa
              5.4          3.9           1.7          0.4      setosa
              4.6          3.4           1.4          0.3      setosa
              5.0          3.4           1.5          0.2      setosa
              4.4          2.9           1.4          0.2      setosa
              4.9          3.1           1.5          0.1      setosa
              5.4          3.7           1.5          0.2      setosa
              4.8          3.4           1.6          0.2      setosa
              4.8          3.0           1.4          0.1      setosa
              4.3          3.0           1.1          0.1      setosa
              5.8          4.0           1.2          0.2      setosa
              5.7          4.4           1.5          0.4      setosa
              5.4          3.9           1.3          0.4      setosa
              5.1          3.5           1.4          0.3      setosa
              5.7          3.8           1.7          0.3      setosa
              5.1          3.8           1.5          0.3      setosa
              5.4          3.4           1.7          0.2      setosa
```

Вывод: в результате выполнения данной лабораторной работы 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.