Міністерство освіти і науки України
Львівський національний університет імені Івана Франка
Факультет електроніки та комп’ютерних технологій
Кафедра системного проектування

Звіт
Про виконання лабораторної роботи №2
З курсу «Основи аналізу даних»
Описова статистика

Виконав:
Студент групи ФЕІ-42
Прізвище Ім’я

Перевірив:
асистент Азаров І.В

Львів 2024

Мета:
Засвоїти основні методи описової статистики для аналізу наборів даних. Навчитися обчислювати ключові статистичні показники та візуалізувати розподіл даних.
Хід роботи:
Завдання
1. Використайте набір "Iris"
Підключив потрібні бібліотеки та завантажив файл:

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

iris_data = pd.read_csv("iris_dataset.csv")
```

2. Обчислення середнього значення, медіани, моди:

- Обчисліть середнє значення для кожної числової колонки.

```
mean_values = iris_data.mean(numeric_only=True)
print("Середнє значення для кожної числової колонки:\n", mean_values)
```

```
[2] ✓ 0.0s
```

```
Середнє значення для кожної числової колонки:
 sepal length (cm)    5.843333
sepal width (cm)      3.054000
petal length (cm)     3.758667
petal width (cm)      1.198667
dtype: float64
```

- Обчисліть медіану для кожної числової колонки.

```
median_values = iris_data.median(numeric_only=True)
print("Медіана для кожної числової колонки:\n", median_values)
```

```
[5] ✓ 0.0s
```

```
Медіана для кожної числової колонки:
 sepal length (cm)    5.80
sepal width (cm)      3.00
petal length (cm)     4.35
petal width (cm)      1.30
dtype: float64
```

- Обчисліть моду для кожної числової колонки.

```
mode_values = iris_data.mode(numeric_only=True).iloc[0]
print("Мода для кожної числової колонки:\n", mode_values)
```

```
[21] ✓ 0.1s

Мода для кожної числової колонки:
 sepal length (cm)    5.0
sepal width (cm)      3.0
petal length (cm)     1.5
petal width (cm)      0.2
Name: 0, dtype: float64
```

3. Обчислення дисперсії та стандартного відхилення:

- Обчисліть дисперсію для кожної числової колонки.

```
variance_values = iris_data.var(numeric_only=True)
print("Дисперсія для кожної числової колонки:\n", variance_values)
```

```
[10] ✓ 0.0s

Дисперсія для кожної числової колонки:
 sepal length (cm)    0.685694
sepal width (cm)      0.188004
petal length (cm)     3.113179
petal width (cm)      0.582414
dtype: float64
```

- Обчисліть стандартне відхилення для кожної числової колонки.

```
std_dev__values = iris_data.std(numeric_only=True)
print("Стандартне відхилення для кожної числової колонки:\n", std_dev__values)
```

```
[12] ✓ 0.0s

Стандартне відхилення для кожної числової колонки:
 sepal length (cm)    0.828066
sepal width (cm)      0.433594
petal length (cm)     1.764420
petal width (cm)      0.763161
dtype: float64
```

- Обчисліть коефіцієнт асиметрії та ексцесу для числових змінних та проаналізуйте їх значення.

```
from scipy.stats import skew, kurtosis

skewness_values = iris_data.skew(numeric_only=True)
kurtosis_values = iris_data.kurtosis(numeric_only=True)
print("Коефіцієнт асиметрії:\n", skewness_values)
print("Коефіцієнт ексцесу:\n", kurtosis_values)
```

```
[13]  ✓ 16.0s

...  Коефіцієнт асиметрії:
      sepal length (cm)    0.314911
     sepal width (cm)      0.334053
     petal length (cm)    -0.274464
     petal width (cm)     -0.104997
     dtype: float64
     Коефіцієнт ексцесу:
      sepal length (cm)    -0.552064
     sepal width (cm)      0.290781
     petal length (cm)    -1.401921
     petal width (cm)     -1.339754
     dtype: float64
```

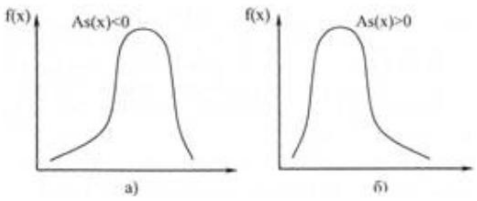
Пояснення:

Коефіцієнт асиметрії (skew)

Чим значення асиметрії ближче до 0 тим краще·

- **0**: Розподіл симетричний (наприклад, нормальний розподіл).
- · **Додатне значення**: Розподіл має **правосторонню (позитивну) асиметрію**. Це означає, що «хвіст» розподілу тягнеться більше вправо, тобто більші значення відхиляються більше від середнього.
- · **Від'ємне значення**: Розподіл має **лівосторонню (негативну) асиметрію**. «Хвіст» розподілу довший зліва.

Вигляд функції щільності розподілу ймовірності у випадках додатного та від'ємного коефіцієнтів асиметрії наведено на рис. 15.6.

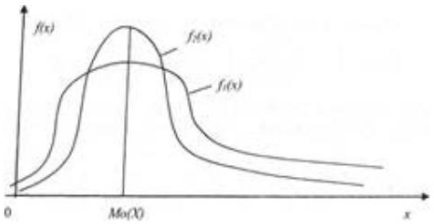


Коефіцієнт ексцесу (kurtosis)

- Показує, наскільки «гострим» або «плоским» є розподіл у порівнянні з нормальним розподілом.
- **Значення:**

- **0**: Розподіл відповідає нормальному, має середню "гостроту".
- **Додатне значення**: Розподіл має **гостру вершину** (розподіл з високим піком), що означає наявність великої кількості значень близьких до середнього.
- **Від'ємне значення**: Розподіл є **плоским**, тобто має ширший пік з меншою кількістю значень біля середнього.

Чим більше значення коефіцієнта ексцесу, тим більш гостровершинним на графіку буде розподіл випадкової величини, що характеризує аналізоване рішення (рис. 15.7).

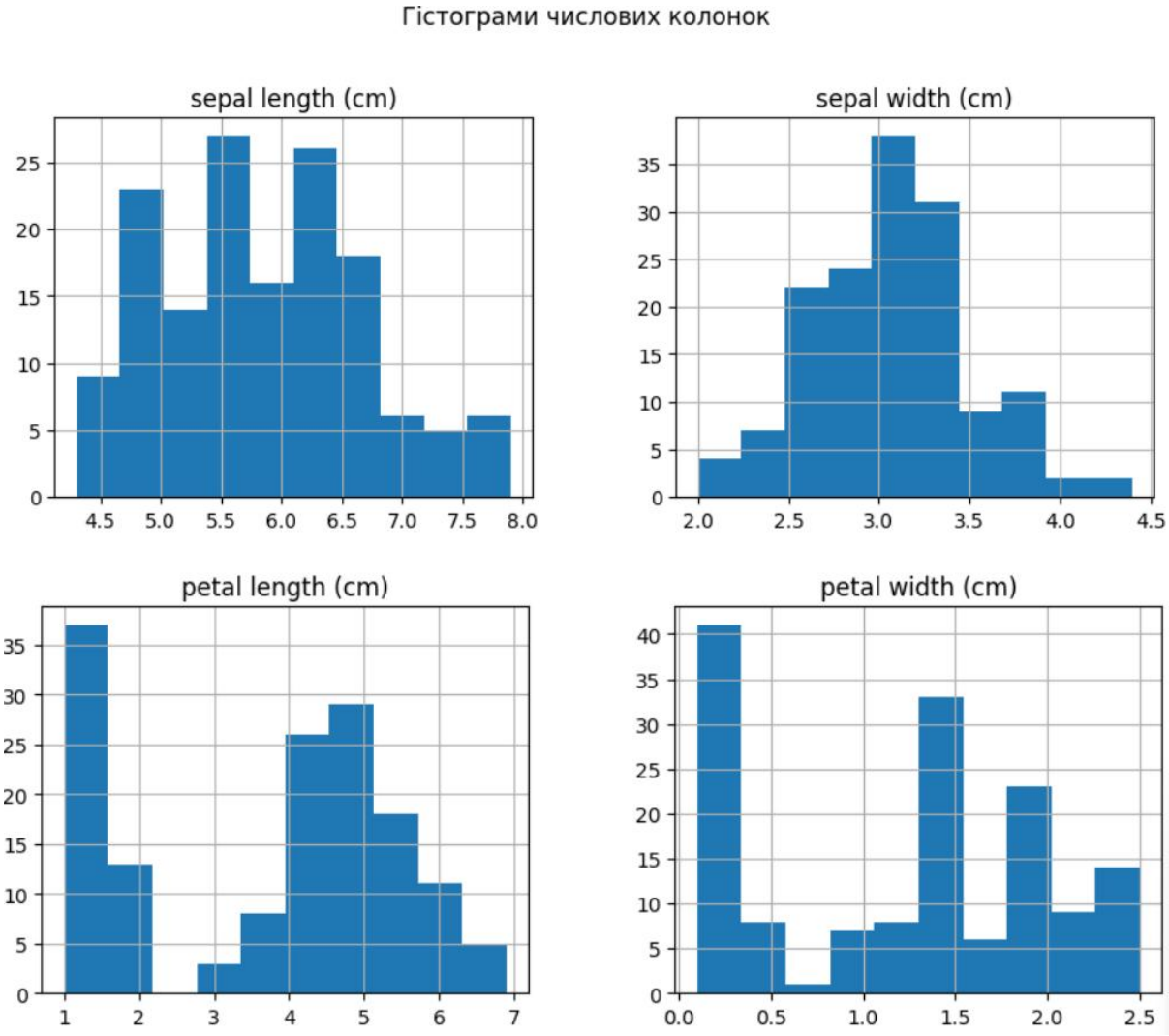


4. Візуалізація розподілу даних за допомогою гістограм та коробкових діаграм (box plots):

- Побудуйте гістограми для числових колонок.

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Гістограми для числових колонок
iris_data.hist(figsize=(10, 8))
plt.suptitle("Гістограми числових колонок")
plt.show()
```

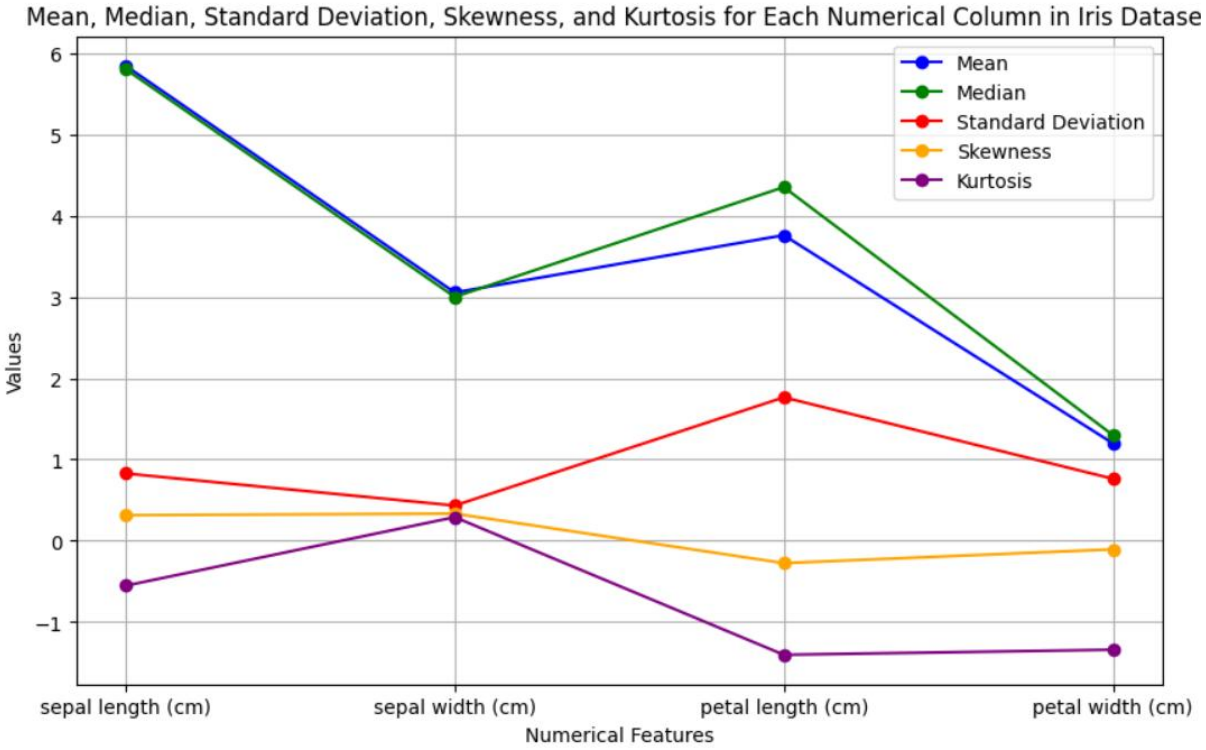


а. Побудуйте лінією три обчислені величини.

```python
summary_stats = pd.DataFrame({
    'Mean': mean_values,
    'Median': median_values,
    'Standard Deviation': std_dev__values,
    'Skewness': skewness_values,
    'Kurtosis': kurtosis_values
})

# Побудова графіків для кожної характеристики
plt.figure(figsize=(10, 6))
plt.plot(summary_stats.index, summary_stats['Mean'], marker='o', label='Mean', linestyle='-', color='blue')
plt.plot(summary_stats.index, summary_stats['Median'], marker='o', label='Median', linestyle='-', color='green')
plt.plot(summary_stats.index, summary_stats['Standard Deviation'], marker='o', label='Standard Deviation', linestyle='-', color='red'
plt.plot(summary_stats.index, summary_stats['Skewness'], marker='o', label='Skewness', linestyle='-', color='orange')
plt.plot(summary_stats.index, summary_stats['Kurtosis'], marker='o', label='Kurtosis', linestyle='-', color='purple')

# Оформлення графіка
plt.title('Mean, Median, Standard Deviation, Skewness, and Kurtosis for Each Numerical Column in Iris Dataset')
plt.xlabel('Numerical Features')
plt.ylabel('Values')
plt.legend()
plt.grid(True)
plt.show()
```

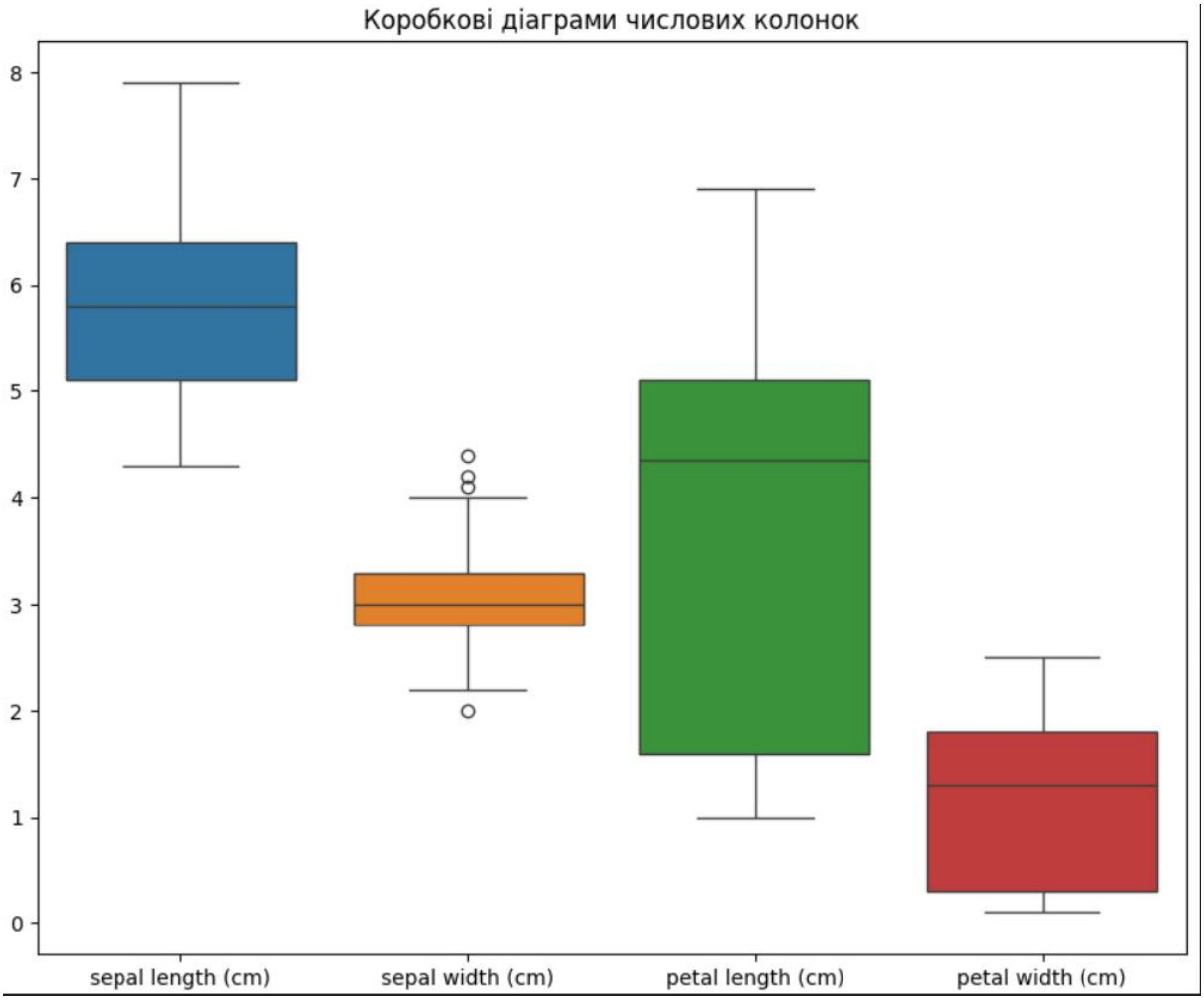


b. Вкажіть якому розподілу відповідають дані

- ТІльки для колонок зі значеннями sepal width та sepal lenght можна сказати, що дані віддалено відповідають нормальному розподілу(Гаусівському). Решта колонок (petal width, petal length), на жаль не мають явно вираженого нормального розподілу.

- Побудуйте коробкові діаграми (box plots) для числових колонок.

```python
# Коробкові діаграми (box plots)
plt.figure(figsize=(10, 8))
sns.boxplot(data=iris_data)
plt.title("Коробкові діаграми числових колонок")
plt.show()
```



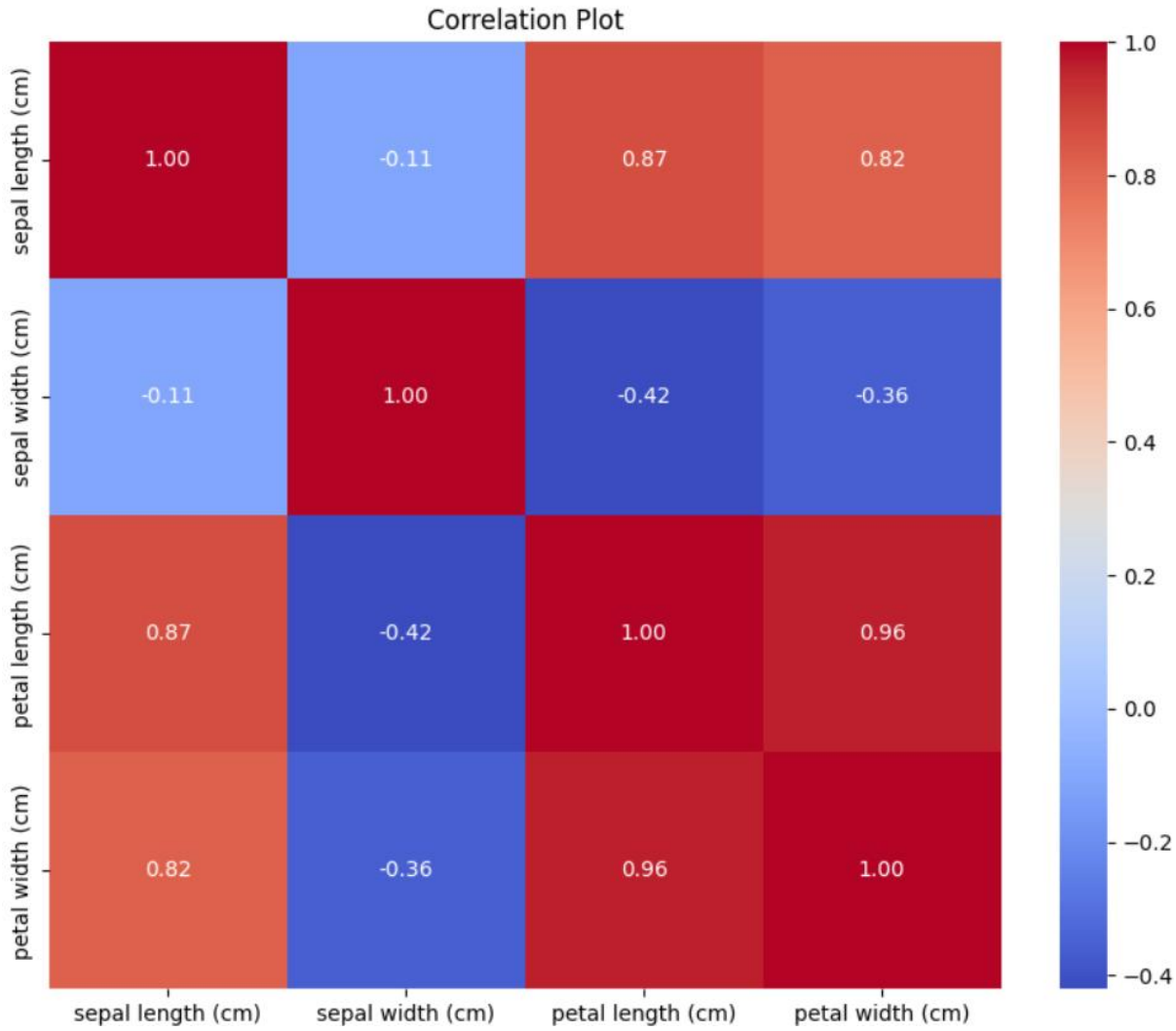
- Побудуйте графіки що дозволяють побачити взаємозв'язків між числовими змінними. Опишіть його.

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Видаляємо нечислові колонки, залишаючи тільки числові дані
iris_data = iris_data.select_dtypes(include=['float64'])

# Створюємо кореляційну матрицю
correlation_matrix = iris_data.corr()

# Побудова теплової карти кореляції
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Plot')
plt.show()
```



Для виявлення взаємозв’язків між числовими знінними - я побудував матрицю кореляції, яка показує на перетині двох значень число (в межах [-1,1])

Кореляція між двома величинами означає, що між ними існує певний статистичний зв'язок або залежність. Зокрема:

- **Позитивна кореляція**: Коли одна величина зростає, інша також має тенденцію зростати. Кореляційний коефіцієнт буде близьким до +1.
- **Негативна кореляція**: Коли одна величина зростає, інша зменшується. Кореляційний коефіцієнт буде близьким до -1.

З графіку видно, що майже усі значення добре корелють між собою, як позитивно(petal width I petal length), так і негативно(sepal width і petal length). На противагу, значення sepal width і sepal length погано корелють між собою, тобто залежності одного значення від іншого майже немає.

Висновок:

У ході виконання лабораторної роботи №2 я засвоїв основні методи описової статистики для аналізу даних та їх візуалізації. Використавши набір даних "Iris", я здійснив обчислення ключових статистичних показників, таких як середнє значення, медіана, мода, дисперсія, стандартне відхилення, а також коефіцієнти асиметрії та ексцесу для числових колонок. Також, більше дізнався про гістограми та коробкові діаграми, які допомогли візуалізувати розподіл числових змінних, а також виявити наявність викидів та особливості розподілу. Кореляційний аналіз дозволив оцінити зв'язки між змінними.