ЖИТОМИРСЬКА ПОЛІТЕХНІКА
ДЕРЖАВНИЙ УНІВЕРСИТЕТ

ПОПЕРЕДНЯ ОБРОБКА ТА ВІЗУАЛІЗАЦІЯ ДАНИХ

ЛЕКЦІЯ 2

План

1. Попередня обробка даних
2. Нормалізація і масштабування
3. Візуалізація даних

СТАНДАРТ CRISP

1. Business Understanding
розуміння цілей бізнесу

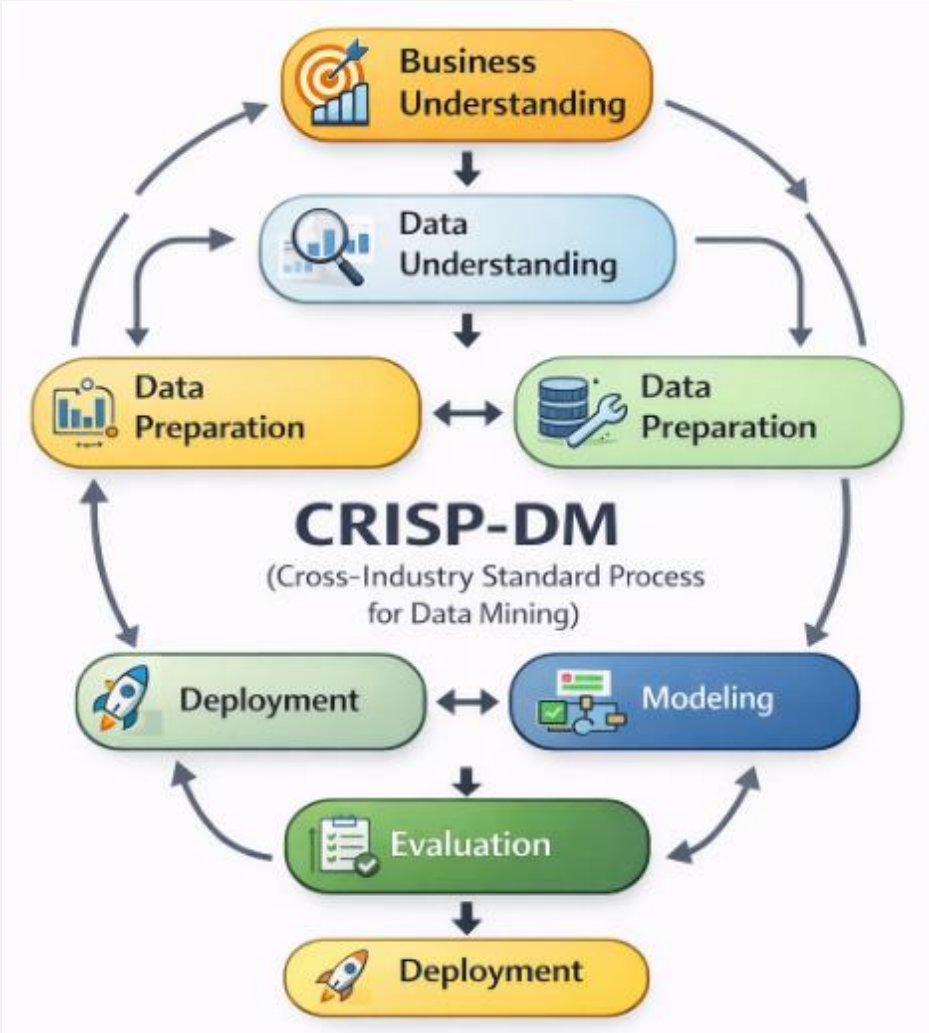
2. Data Understanding
збір та знайомство з даними

3. Data Preparation
підготовка даних (найбільш трудомісткий етап, до 80% часу)

4. Modeling
вибір та застосування алгоритмів

5. Evaluation
перевірка, чи відповідає модель бізнес-цілям

6. Deployment
впровадження моделі в реальну експлуатацію



ПОНЯТТЯ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ

Попередня обробка даних (Data Preprocessing) — це сукупність методів підготовки сирих даних до аналізу або моделювання.

Основні завдання попередньої обробки:

- виявлення та обробка пропущених значень;
- усунення помилок і викидів;
- перетворення типів даних;
- нормалізація та масштабування;
- створення нових ознак.

Попередня обробка є обов'язковою, оскільки реальні дані часто:

- неповні;
- багато шуму;
- неоднорідні;
- зібрані з різних джерел.

ПОНЯТТЯ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ

На першому етапі виконується ознайомлення з датасетом.

Основні інструменти Pandas:

- head() — перегляд перших рядків;
- tail() — перегляд останніх рядків;
- info() — структура датасету та використання пам’яті;
- describe() — базова статистика числових змінних.

На цьому етапі визначаються:

- кількість спостережень;
- типи змінних;
- наявність пропущених значень.

ПОПЕРЕДНЯ ОБРОБКА ДАНИХ АБО ОЧИЩЕННЯ ДАНИХ (DATA CLEANING)

Очищення даних - це фундамент, згідно з правилом GIGO якість моделі безпосередньо залежить від чистоти вхідних даних.

Правило *GIGO* (Garbage In, Garbage Out) — фундаментальний принцип аналізу даних та комп'ютерних наук, згідно з яким якість результатів обробки, моделювання або прогнозування безпосередньо залежить від якості вхідних даних. Використання зашумлених, неповних або некоректних даних неминуче призводить до отримання хибних або недостовірних результатів, незалежно від складності застосованих алгоритмів.

ВИДАЛЕННЯ ДУБЛІКАТІВ. ВИПРАВЛЕННЯ СТРУКТУРНИХ ПОМИЛОК

Видалення дублікатів. Повторювані записи можуть штучно завищувати значущість певних спостережень.

Виправлення структурних помилок. Неправильні назви категорій (наприклад, "Нью-Йорк" та "NY"), помилки в типах даних (число як рядок).

РОБОТА З ПРОПУЩЕНИМИ ЗНАЧЕННЯМИ

Пропущені значення (NaN) є однією з найпоширеніших проблем у даних.

Причини виникнення:
помилки збору даних;
відсутність відповіді респондента;
технічні збої.

Основні підходи:
видалення рядків або колонок;
заповнення середнім, медіаною, модою;
групове заповнення (наприклад, за регіоном).

ОБРОБКА ТИПІВ ДАНИХ ТА ОПТИМІЗАЦІЯ ПАМ’ЯТІ

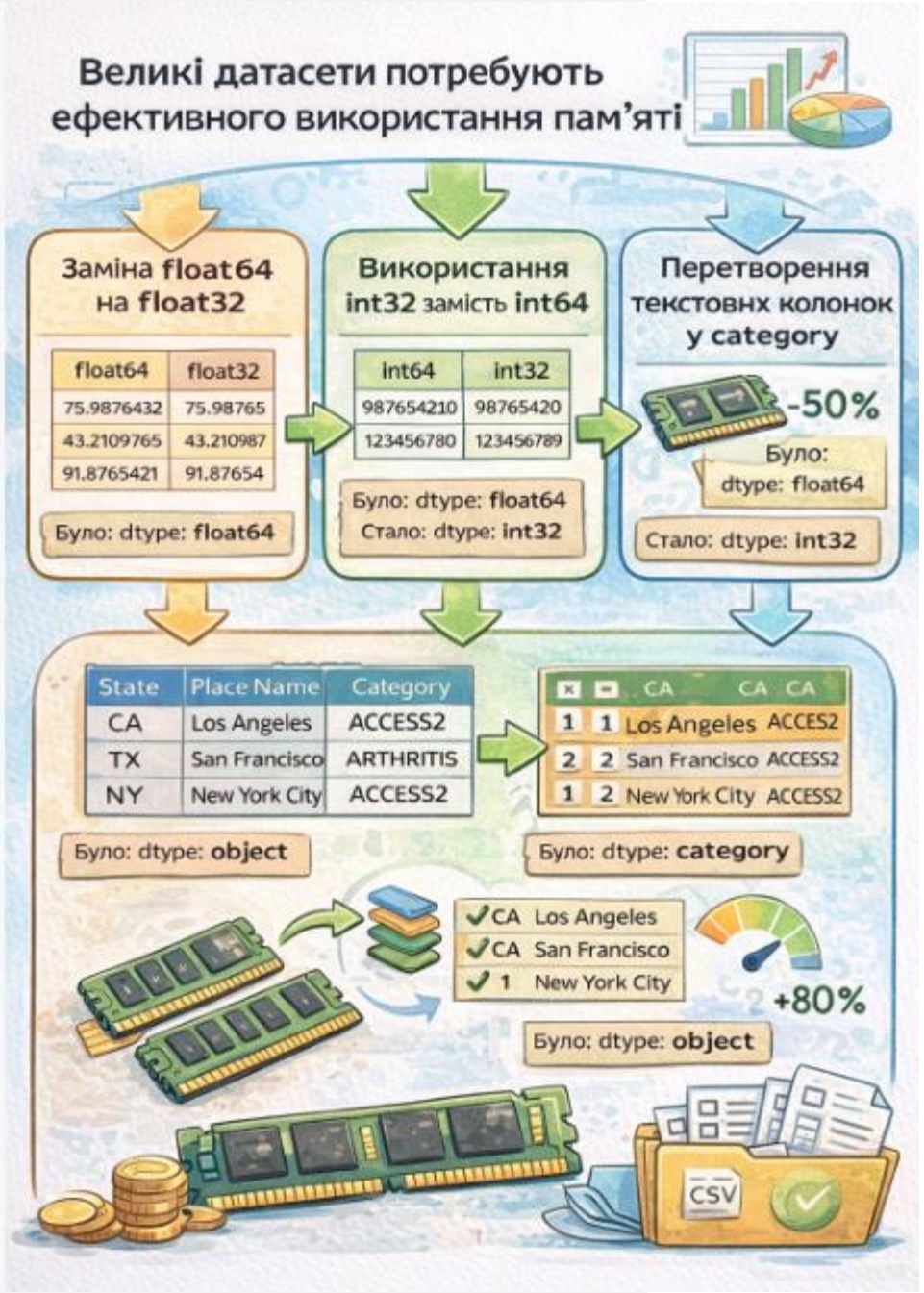
Великі датасети потребують ефективного використання пам’яті.

Основні підходи:

- заміна float64 на float32;
- використання int32 замість int64;
- перетворення текстових колонок у category.

Оптимізація типів дозволяє:

- зменшити використання пам’яті;
- прискорити обчислення;
- ефективніше працювати з великими даними.



ВИЯВЛЕННЯ ТА ОБРОБКА ВИКИДІВ

Викиди (outliers) — це значення, які суттєво відрізняються від більшості спостережень.

Методи виявлення:

- описова статистика;
- міжквартильний розмах (IQR);
- візуальний аналіз (boxplot).

Викиди можуть бути:

- помилками вимірювання;
- рідкісними, але реальними випадками.

Рішення щодо їх обробки приймається залежно від предметної області.

Методи: Використання Z-оцінки або міжквартильного розпаду (IQR).

ВИЯВЛЕННЯ ТА ОБРОБКА ВИКИДІВ

Z-оцінка (Z-score) *дозволяє визначити, наскільки значення відхиляється від середнього у вимірюваній кількості стандартних відхилень.*
Спостереження з $|Z| > 2,5$, як правило, розглядаються як потенційні викиди.

.

***Приклад*:**

Нехай маємо вибірку значень середнього часу відповіді сервера (мс):
120,125,130,128,127,126,129,400

Виявлення викидів за допомогою Z-оцінки
Обчислюємо середнє значення = 160,63
стандартне відхилення вибірки $\approx 90,4$
$Z = (400-160,63)/90,4=2,65$

$|Z| > 3$ (сильний викид)
$|Z| > 2,5$ (потенційний викід)

*Значення **400 мс** має Z-оцінку, що перевищує порогове значення, тому класифікується як статистичний викид і потребує окремого аналізу або видалення.*

ВИЯВЛЕННЯ ТА ОБРОБКА ВИКИДІВ

Міжквартильний розмах (Interquartile Range, IQR) *базується на квартилях розподілу та є стійким до екстремальних значень.*

Викидами вважаються спостереження, що виходять за межі інтервалу

$[Q1-1.5\cdot IQR,\ Q3+1.5\cdot IQR]$.

Виявлення викидів за допомогою IQR
Визначаємо перший (Q1) та третій (Q3) квартилі, обчислюємо міжквартильний розмах:
$IQR=Q3-Q1$

Нехай маємо вибірку значень середнього часу відповіді сервера (мс):
120,125,130,128,127,126,129,400

120, 125, 126, 127, 128, 129, 130, 400

Q1=(125+126)/2=125,5
Q3= (129+130)/2= 129,5

IQR= Q3 - Q1 = 129,5-125,5 = 4

Згідно з правилом Тьюкі:
Нижня межа: $Q1-1,5\cdot IQR=125,5-6=119,5$
Верхня межа: $Q3+1,5\cdot IQR=129,5+6=135,5$

Допустимий інтервал значень: [119,5; 135,5]
*Значення **400 мс** значно перевищує верхню межу*

СТВОРЕННЯ НОВИХ ОЗНАК

Feature engineering — процес створення нових змінних на основі існуючих.

Приклади:

- обчислення абсолютних значень (кількість людей замість відсотка);
- категоризація числових змінних;
- агрегування за групами.

Нові ознаки часто підвищують інформативність даних і якість аналізу.

НОРМАЛІЗАЦІЯ ТА МАСШТАБУВАННЯ

У практиці терміни часто плутають, тому варто чітко зафіксувати:

Масштабування (scaling) — приведення ознак до порівнюваних масштабів.
Нормалізація (normalization) — приведення до фіксованого діапазону або розподілу.

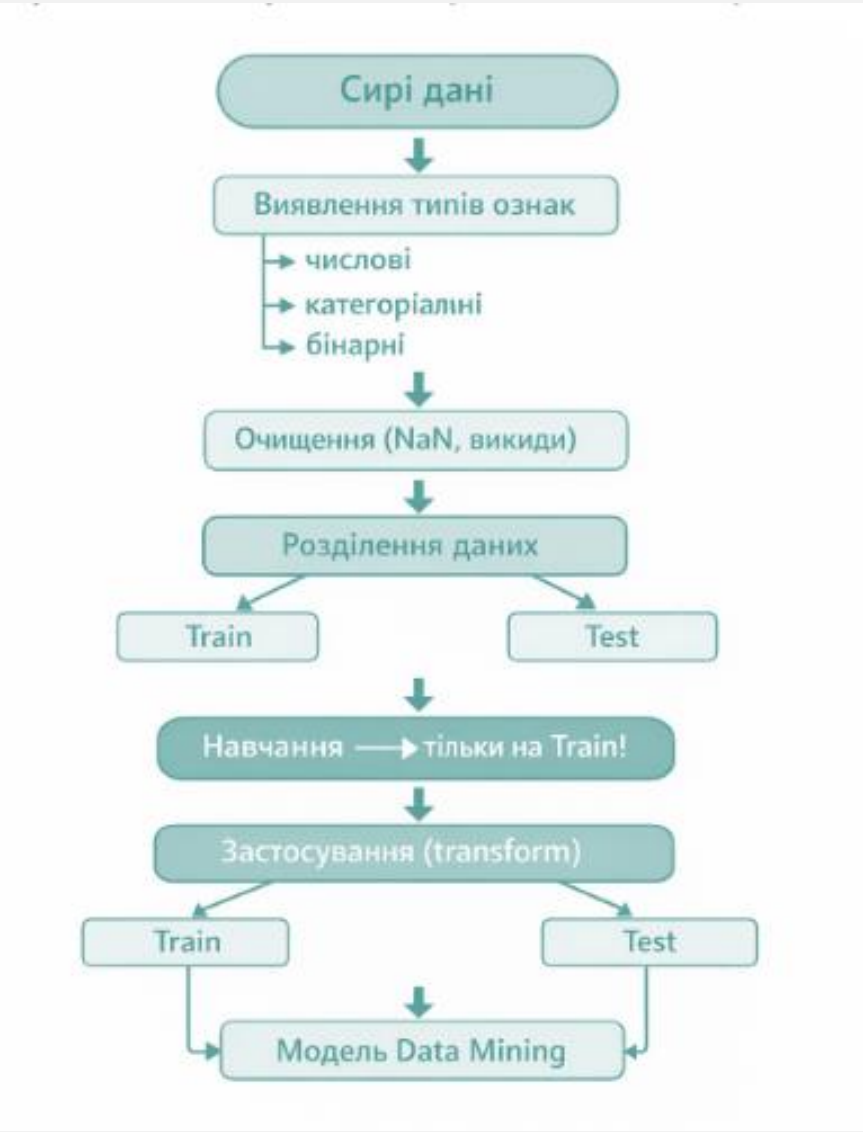
У бібліотеках Python обидва процеси реалізуються через **scalers**.

МАСШТАБУВАННЯ

Алгоритми, що базуються на відстані (k-NN, SVM) або градієнтному спуску, чутливі до масштабу ознак і можуть працювати некоректно.

Алгоритм	Чи потрібне масштабування
k-NN, k-means	✔ Обов’язково
SVM	✔
Логістична регресія	✔
Нейронні мережі	✔
Дерева рішень	✖
Random Forest	✖
XGBoost	✖

МАСШТАБУВАННЯ



Масштабування — це детермінований перетворення числових ознак, яке:

- **не змінює відносних залежностей між об’єктами;**
- **приводить всі ознаки до порівняного числового масштабу;**
- **впливає на метрику відстані та швидкість навчання моделей.**

📌 **Масштабування не створює нової інформації, а лише змінює спосіб її числового подання.**

МАСШТАБУВАННЯ

Метод	Опис	
Min-Max Scaling	Приводить дані до діапазону [0, 1]. Чутливий до викидів. **Доцільно для** нейронних мереж і алгоритмів з обмеженим діапазоном значень.	$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
Standardization (Z-score)	Центрує дані навколо 0 із стандартним відхиленням1. **Доцільно для** лінійних моделей, SVM, PCA.	$x' = \frac{x - \mu}{\sigma}$
Robust Scaling	Використовує медіану та IQR. Стійкий до викидів, добре працює з реальними “брудними” даними.	$x' = \frac{x - \text{median}}{IQR}$

ПОНЯТТЯ ТА РОЛЬ ВІЗУАЛІЗАЦІЇ ДАНИХ

Візуалізація даних — це графічне представлення інформації для полегшення її аналізу та інтерпретації.

Основні цілі:

- виявлення закономірностей;
- пошук аномалій;
- порівняння груп;
- комунікація результатів аналізу.

ОСНОВНІ ТИПИ ВІЗУАЛІЗАЦІЙ

Найпоширеніші графіки:

- гістограма — аналіз розподілу;
- стовпчикова діаграма — порівняння категорій;
- boxplot — виявлення викидів;
- scatter plot — аналіз взаємозв’язку змінних;
- лінійний графік — динаміка у часі.

Правильний вибір типу графіка залежить від:

- типу даних;
- мети аналізу;
- цільової аудиторії.

ВІЗУАЛІЗАЦІЯ ЯК ІНСТРУМЕНТ АНАЛІЗУ

Візуалізація використовується не лише для презентації результатів, але й для:

- перевірки гіпотез;
- виявлення помилок у даних;
- попереднього аналізу перед моделюванням;
- виявлення закономірностей;
- знаходження викидів та аномалій;
- інтерпретації результатів.

Поєднання візуального аналізу з описовою статистикою є ефективною практикою аналізу даних.

БІБЛІОТЕКИ PANDAS, MATPLOTLIB, SEABORN, PLOTLY

Pandas - маніпуляція даними. Основна бібліотека для роботи з табличними даними (DataFrames).

- **df.dropna(), df.fillna() — робота з пропусками.**
- **df.drop_duplicates() — видалення повторів.**
- **df.describe() — базова статистика.**

MATPLOTLIB, SEABORN

Matplotlib та Seaborn - статична візуалізація

Matplotlib - низькорівнева бібліотека, що надає повний контроль над кожним елементом графіка.

Seaborn - побудована на базі Matplotlib, спеціалізується на статистичних графіках з більш привабливим дизайном за замовчуванням.

PLOTLY

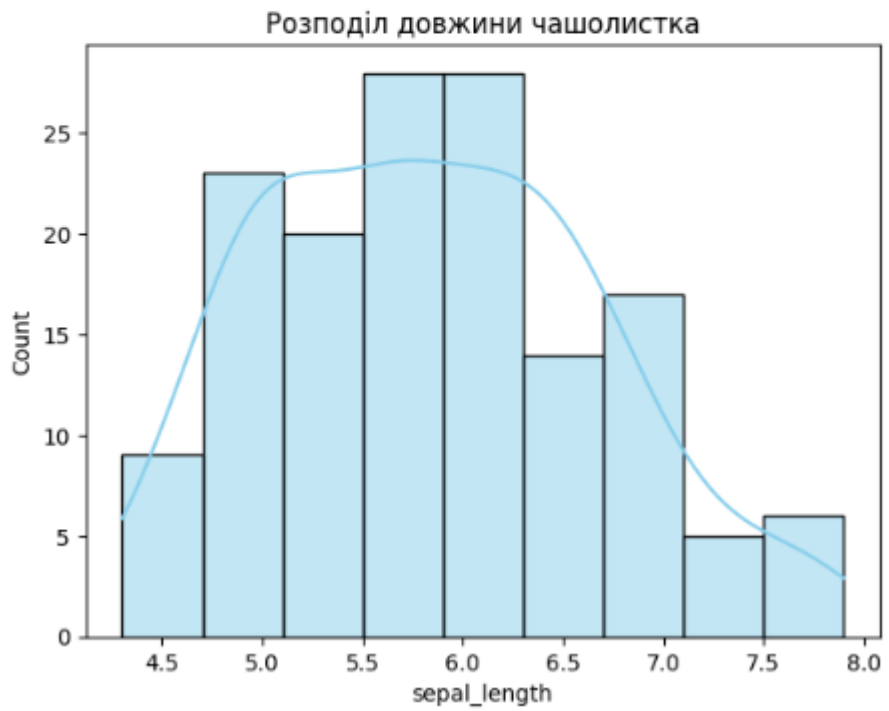
Plotly – інтерактивність

Ідеально підходить для веб-звітів та Dash-додатків.

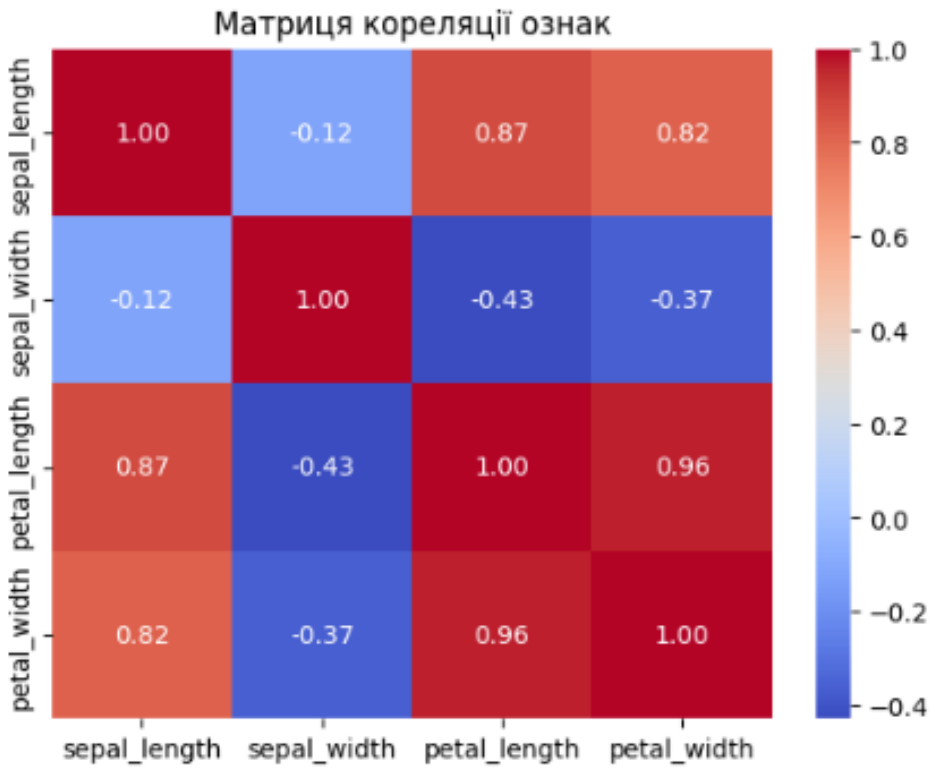
Дозволяє масштабувати графіки, переглядати значення при наведенні та створювати динамічні 3D-моделі.

ВИДИ ГРАФІКІВ

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

data = sns.load_dataset("iris")

sns.histplot(data['sepal_length'], kde=True, color='skyblue')
plt.title('Розподіл довжини чашолистка')
plt.show()
```

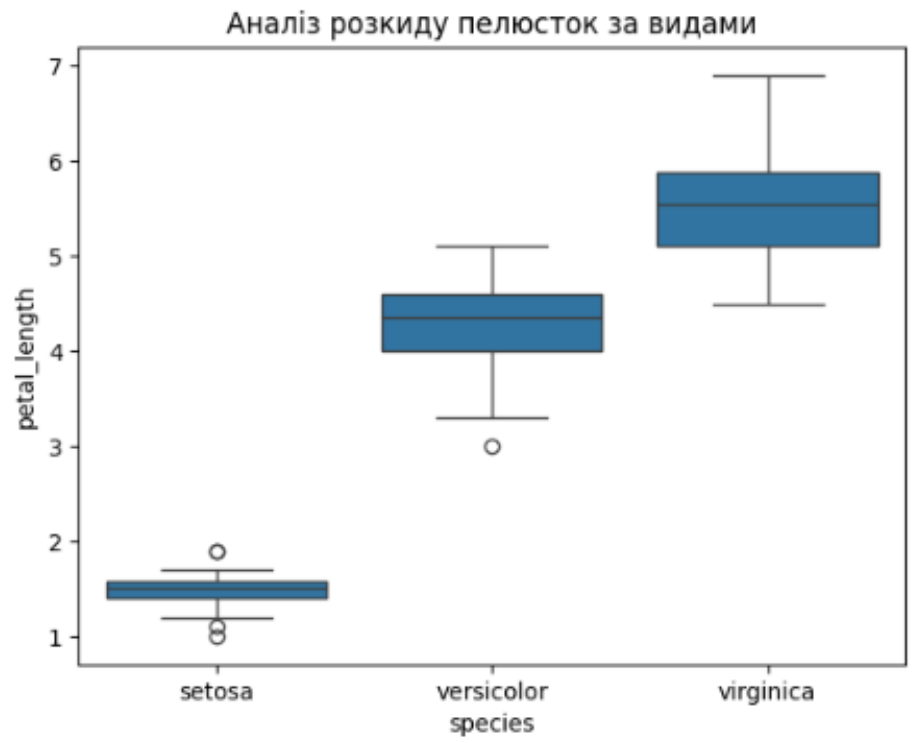



```python
corr = data.corr(numeric_only=True)

sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Матриця кореляції ознак')
plt.show()
```

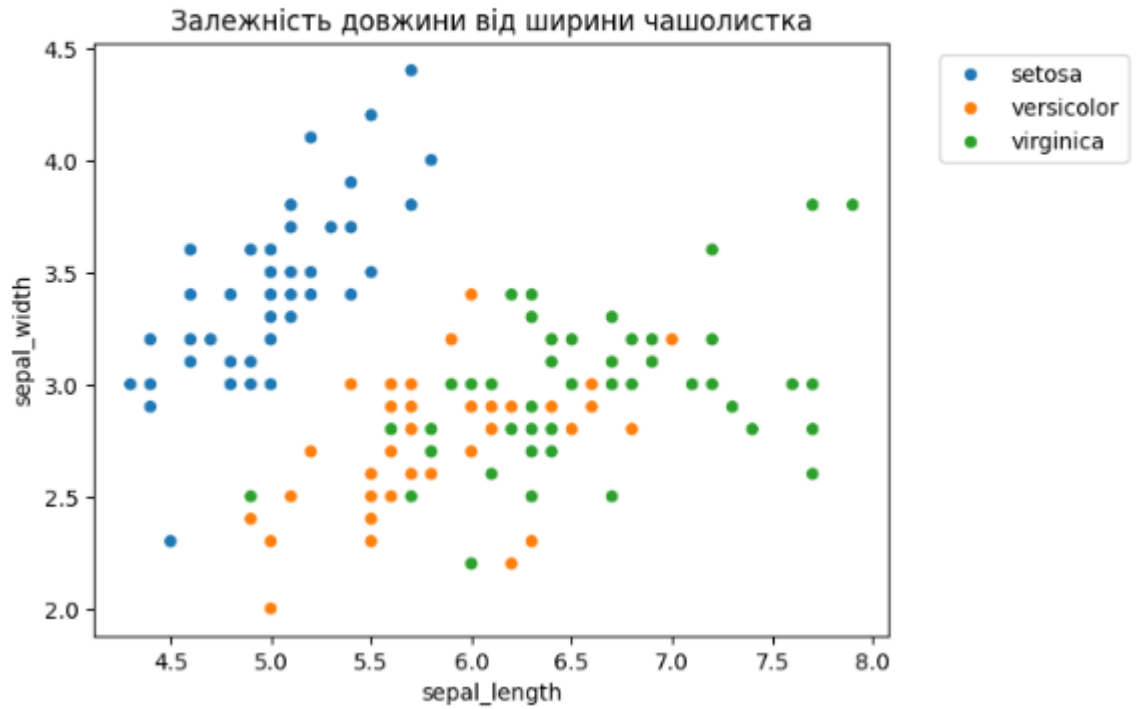


ВИДИ ГРАФІКІВ

```
sns.boxplot(x='species', y='petal_length', data=data)
plt.title('Аналіз розкиду пелюсток за видами')
plt.show()
```




```
sns.scatterplot(x='sepal_length', y='sepal_width', hue='species', data=data)
plt.title('Залежність довжини від ширини чашолистка')
plt.legend(bbox_to_anchor=(1.05, 1), loc=2)
plt.show()
```



САМОСТІЙНА РОБОТА

Розрахувати статистичні характеристики ряду (середнє, дисперсію, середнє квадратичне відхилення, моду, медіану) використовував Python модуль статистики, метод **describe**().

Прискорений розвідувальний аналіз даних з використанням бібліотеки **pandas-profiling**

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html