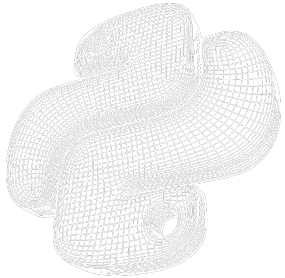
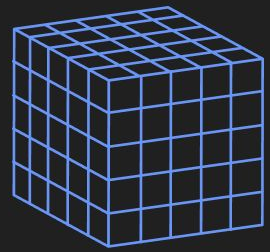
Python для Data Science



Заняття 5.
Exploratory Data Analysis(EDA)
та очищення даних

План заняття

- Визначення EDA й опис його можливих компонент
- Знайомство з бібліотекою візуалізації Seaborn
- EDA за допомогою Pandas
- Бібліотека Sweetwiz



Exploratory Data Analysis

Exploratory Data Analysis (EDA) належить до критично важливого процесу виконання початкового дослідження даних з метою виявлення закономірностей та аномалій, тестування гіпотез і перевірки припущень за допомогою зведеної статистики та графічних зображень.

Рекомендовано спочатку зрозуміти дані та спробувати зібрати з них якомога більше ідей. EDA полягає в тому, щоб зрозуміти наявні дані, перш ніж розпочинати моделювання.

EDA допомагає

- очистити набір даних
- зрозуміти, які моделі можна використовувати з цими даними та як підготувати дані для навчання
- розуміти, які додаткові змінні (ознаки, features) можуть бути згенеровані
- знаходити та усувати аномалії
- знаходити та видаляти пропущені значення
- розуміти характерні особливості
- знайти залежність у змінних

More motivation

Data in academia & research

Data in the real world

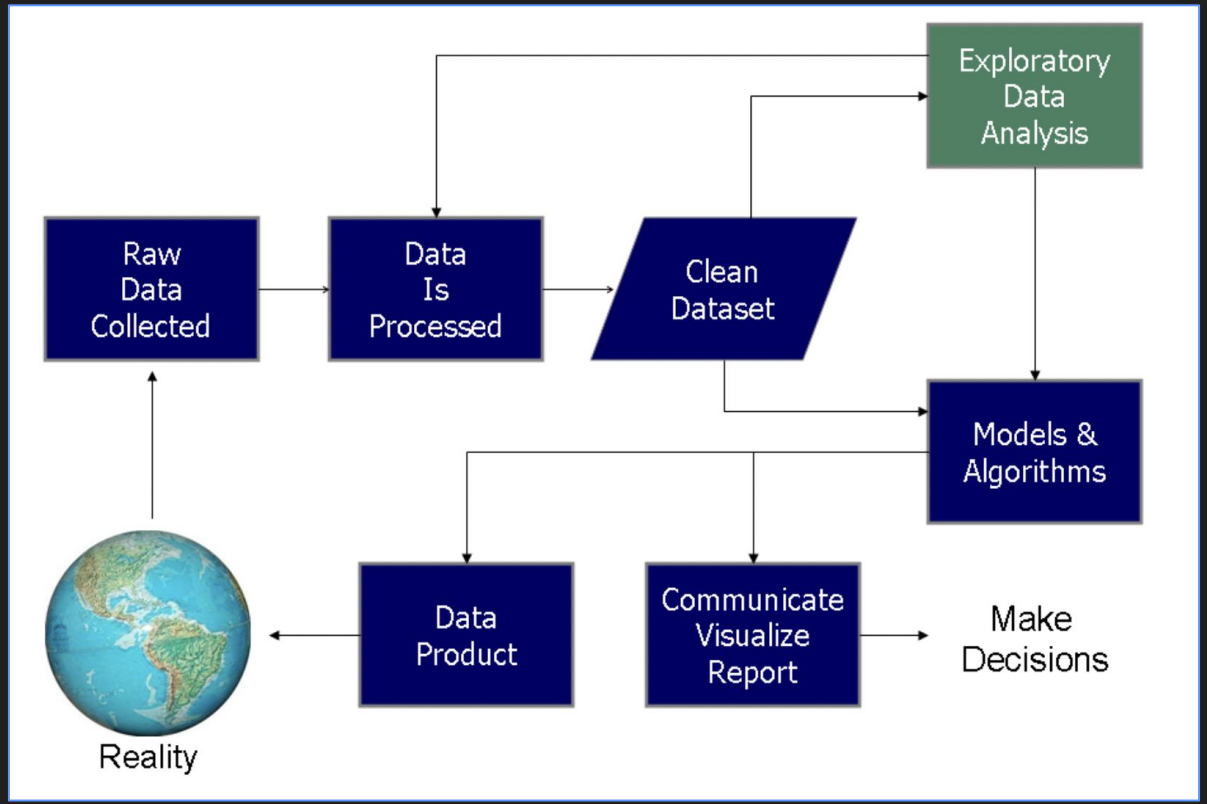
EDA steps

Загалом підхід EDA є дуже ітеративним. Наприкінці вашого дослідження ви можете виявити щось, що вимагатиме від вас переробити аналіз ще раз. Це нормально! Але щоб надати хоча б трохи структури, я пропоную такі категорії для ваших EDA:

- **Дослідження структури** — вивчення загальної форми набору даних, а також типів даних.
- **Дослідження якості** — проаналізуйте загальну якість набору даних: чи є дублікати, відсутні значення і небажані записи.
- **Дослідження змісту** — зрозумівши структуру та якість набору даних, ми можемо продовжити та виконати більш поглиблене дослідження значень ознак (features) і подивитися, як різні ознаки співвідносяться одна з одною.

EDA можна робити на різних етапах

Data Science Process



Live coding / Практика

???

Q&A