6 кроків очищення даних, які повинен знати кожен аналітик даних

1. **Дослідження набору даних**

- Ідентифікуйте джерела даних.
- Розумійте типи даних (числові, категоріальні).
- Виявляйте потенційні проблеми якості даних.

Ключові дії:

- Перегляньте стовпці та діапазони значень.
- Визначте призначення набору даних.

2. **Обробка відсутніх даних**

- Відсутні значення можуть спотворювати аналіз і повинні бути ефективно оброблені.

Ключові дії:

- Ідентифікуйте відсутні значення (наприклад, порожні клітинки, NULL).
- Видаліть рядки або стовпці з багатьма відсутніми значеннями.
- Замініть відсутні значення середнім, медіаною або модою.

3. **Видалення дублікатів**

- Дублікати можуть призводити до спотворення аналізу та завищення результатів.

Ключові дії:

- Перевірте наявність однакових рядків або записів.
- Видаліть дублікати, зберігаючи унікальні дані.

4. **Обробка форматування**

- Невідповідності у форматуванні можуть призвести до помилок в аналізі.

Ключові дії:

- Стандартизуйте формати дат (наприклад, DD-MM-YYYY або MM/DD/YYYY).
- Узгодьте регістр тексту (наприклад, "Male" vs. "male").
- Використовуйте єдині одиниці вимірювання (наприклад, конвертація км у милі).

5. **Обробка викидів**

- Викиди можуть вказувати на помилки або містити значущу інформацію.

Ключові дії:

- Використовуйте статистичні методи (наприклад, Z-score, IQR) для виявлення викидів.
- Визначте, чи потрібно видалити, обмежити або додатково аналізувати викиди.

6. **Перевірка даних**

- Останній етап гарантує, що очищений набір даних є точним і готовим для аналізу.

Ключові дії:

- Перехресно перевірте очищені дані з вихідним джерелом.
- Переконайтеся, що дані відповідають визначеним правилам якості.
- Тестуйте дані на узгодженість і надійність в аналізі.

Ці кроки допоможуть отримати якісні та достовірні дані для подальшого аналізу.