Xccelerate

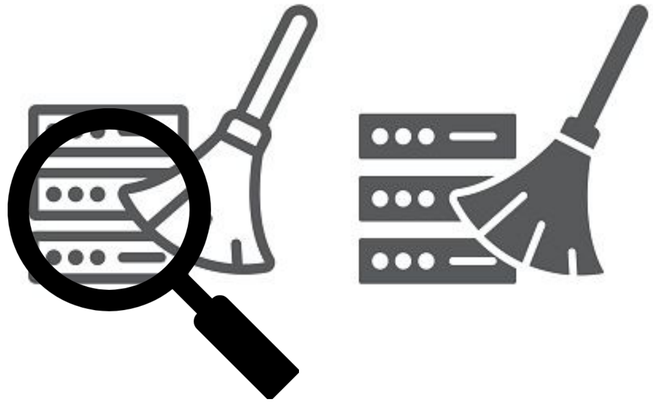
DATA CLEANING

Contents

- What is data cleaning?
- The need for data cleaning
- Ways to clean data
- Regular expressions
    - What is it?
    - Why use them?

What is Data Cleaning?

- Identifying a data error
  - null
  - syntax
  - inaccurate
  - irrelevant
- Solving the data error
  - replace
  - modify
  - delete



Why?

- Garbage in, garbage out (GIGO)

Ways to Clean Data

- Replace - to change the value of the data to a different value
- Modify - to change the current value of the data slightly
- Delete - to remove the data from the dataset entirely
- If null?
  - Delete
  - Replace
- If syntax?
  - Modify
  - Delete
- If inaccurate?
  - Replace
  - Delete
- If irrelevant?
  - Delete

REGULAR EXPRESSIONS

Regular Expressions (Regex)

- used to define search patterns in text
- can concisely define which part of the text we want
- Example:
- `Regex: pattern = "cat"`
- `String: "the cat sat on the mat"`
- `Result: "cat"`

REGULAR EXPRESSIONS

Xccelerate

DATA CLEANING