Google Developer Student Clubs

# Data Cleaning & Data Transform with Pandas

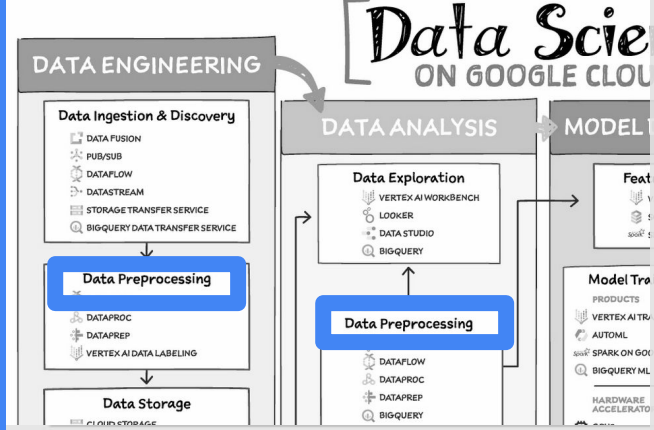

**Ayu Sudi Dwijayanti**
GDSC Lead Binus University Online.
Instructor at Hacktiv8 Indonesia.

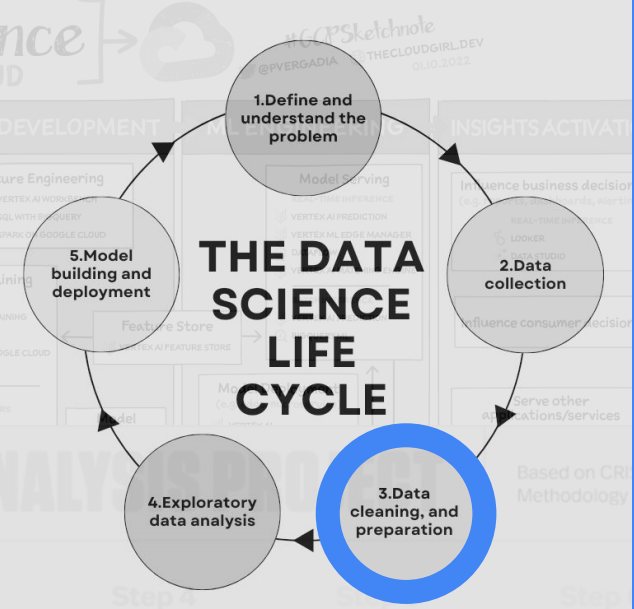
# Agenda:

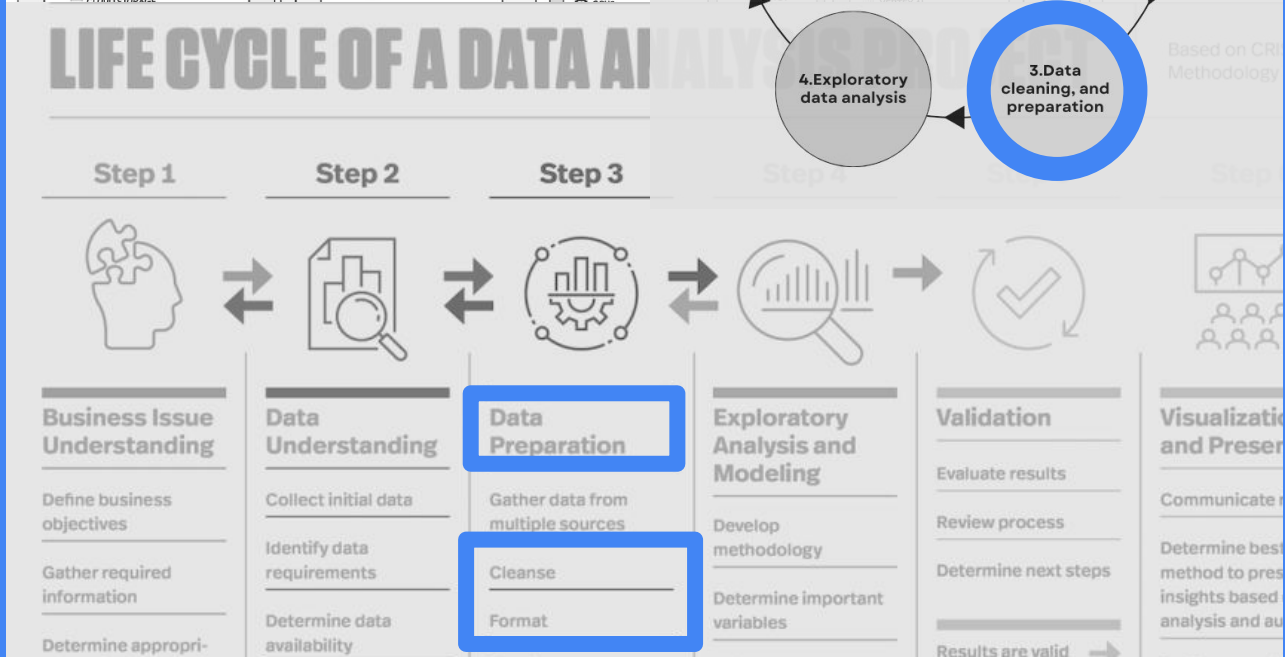
- **Teori**
  - Apa Itu Data Cleaning & Data Transform ?
  - Data Cleaning : Cara Mengatasi Missing Value
  - Data Cleaning : Cara Mengatasi Data Duplikasi & Outlier
  - Data Transform : Pengubahan Yang Dilakukan
- **Pengenalan Tools & Tech Stack**
- **Demo**
- **Assist Session**

**Objective :** Experience data cleaning & data transform process

# WHY ???




Google Developer Student Clubs

Google Developer Student Clubs

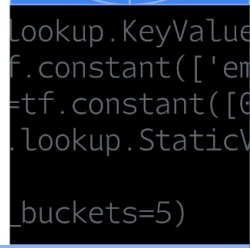
# WHAT IS . . .

- **Data Cleaning**

  Menangani nilai kosong (missing value), nilai duplikat dan nilai ekstrim (outlier) untuk memastikan konsistensi data.

- **Data Transform**

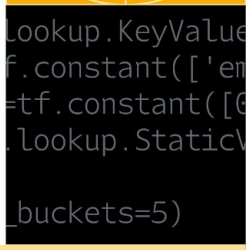
  Mengubah data type, penskalaan atau penyesuaian nilai (normalizing numerical values), membuat column baru berdasarkan fitur atau kondisi yang kita butuhkan (generating derived features).



DATA CLEANING

## Cara Mengatasi Missing Value :

- **Menghapus Row / Column**
  Kapan?? Data yang kosong tidak berpengaruh ke analisis. Data tersebut bisa dipertimbangkan untuk dihapus.
- **Imputation**
  Kapan?? Data yang kosong ada pengaruhnya namun tidak terlalu berpengaruh, sehingga imputasi dengan angka mean atau median atau bahkan bisa dengan angka modus.
- **Advanced Techniques**
  Kapan?? Data yang kosong merupakan data yang significant (dibutuhkan) dan tidak random kekosongannya seperti ada patternnya, sehingga metode advance yang dimaksud bisa menggunakan multiple imputation, maximum likelihood estimation, atau metode imputasi berbasis machine learning.

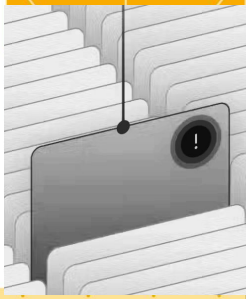



Google Developer Student Clubs

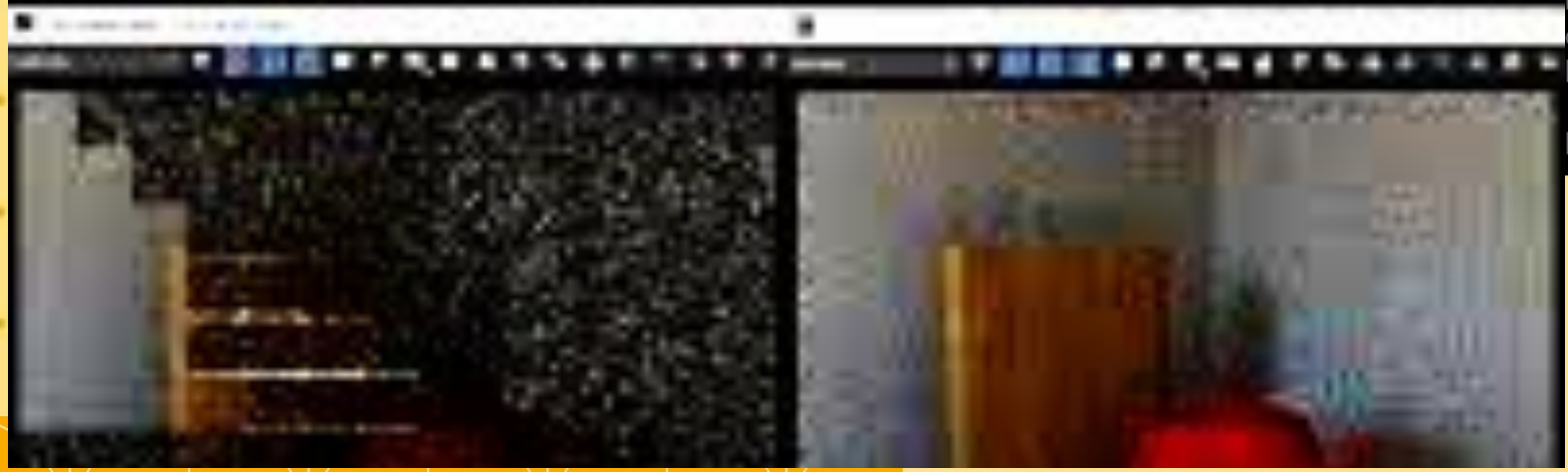
DATA CLEANING

## Mengatasi Data Duplikasi & Outlier :

1. **Pengecekan**
   - Pengecekan data duplikasi bisa disesuaikan dengan kasusnya.
   - Pengecekan outlier bisa dengan box plot.
2. **Pertimbangkan**
   Pertimbangkan seberapa penting dan pengaruhnya data tersebut jika dihapus. Example : pengaruh pada angka mean (rata-rata).
3. **Penghapusan**
   Dilakukan setelah sudah mendapatkan jawaban dari hasil pertimbangan. Kembali ke issue "but, it depends on the case".



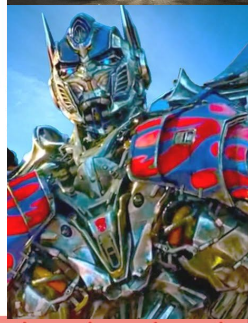
# BUT, IT DEPENDS ON THE CASE



# DATA TRANSFORM

**Pengubahan / konversi yang dilakukan pada data transform :**

- **Rename** : Mengubah nama column
- **Cast** : Mengubah data type
- **Join** : Menggabungkan data dari 2 atau lebih column
- **Enrich** : Memperbanyak/menambahkan column



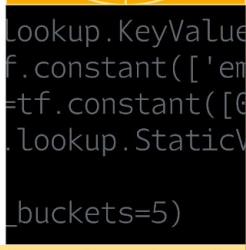
Google Developer Student Clubs

# TOOLS FOR OUR DEMO

**PYTHON** : PROGRAMMING LANGUAGE

**PANDAS** : LIBRARY DI PYTHON UNTUK MEMPROSES DATA DAN MEMVISUALISASIKAN DATA

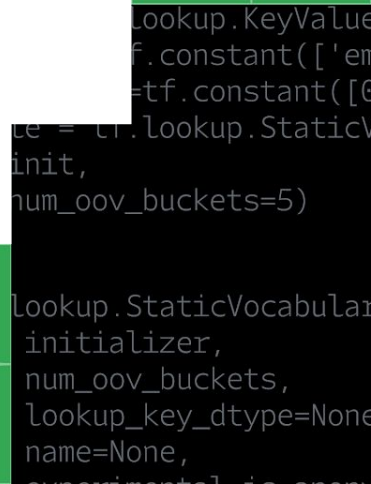
**GOOGLE COLAB :** PRODUCT GOOGLE YANG DIGUNAKAN UNTUK RESEARCH & CODING PYTHON DI CLOUD COMPUTING.



Google Developer Student Clubs

# DEMO

**File Data :**
https://raw.githubusercontent.com/ayusudi/womenland_class_demo/main/data_penjualan.csv
**File Data Preview :**
https://github.com/ayusudi/womenland_class_demo/blob/main/data_penjualan.csv
**File Demo :** (setelah akan di push ke Github)
https://github.com/ayusudi/womenland_class_demo



Google Developer Student Clubs

# Thank you!

## Connect with me

Ayu Sudi Dwijayanti

@ayusudii

@ayusudii

**JOIN ASSIST SESSION FOR DISCUSSION OR HELP**


**Ayu Sudi Dwijayanti**
GDSC Lead Binus University Online.
Instructor at Hacktiv8 Indonesia.