

# Penerapan Data Cleansing Pada Dataset List\_of\_orders



# Kenapa Data Cleansing Penting?

- Data mentah sering berisi duplikat, format tidak konsisten, dan informasi tidak relevan.
- Analogi: Data kotor ibarat beras yang belum dicuci sebelum dimasak 🍲.
- Goal: data rapi → analisis lebih akurat.





# Data Cleansing Steps



Mounted at /content/drive

## Import Data

Gunakan Pandas & Google Colab (akses dari Google Drive).

```
Order ID      object
Order Date    object
CustomerName  object
State         object
City          object
dtype: object
```

## Cek Struktur Data

Tampilkan info dataset & tipe data tiap kolom

	Order ID
0	B-25601
1	B-25602
2	B-25603
3	B-25604
4	B-25605

## Standarisasi Order ID

Format jadi B-xxxxx (5 digit).

	CustomerName	Nama Pelanggan Bersih
0	Bharat	Bharat
1	Pearl	Pearl
2	Jahan	Jahan
3	Divsha	Divsha
4	Kasheen	Kasheen

## Bersihkan Customer Name

Hapus gelar, angka, dan simbol pakai Regex.



# Data Cleansing Steps

```
/tmp/ipython-input-3118742467.py:5:  
standardized_date = pd.to_datetime
```

	Order Date	Order Date Standar
0	1/4/2018	04-01-2018
1	1/4/2018	04-01-2018
2	3/4/2018	04-03-2018
3	3/4/2018	04-03-2018
4	5/4/2018	04-05-2018

## Standarisasi Order

### Date

Ubah semua ke format  
dd-mm-YYYY

	Order ID	Order Date	CustomerName	State	City	Nama Pelanggan Bersih	Order Date Standar
0	B-25601	1/4/2018	Bharat	Gujarat	Ahmedabad	Bharat	04-01-2018
1	B-25602	1/4/2018	Pearl	Maharashtra	Pune	Pearl	04-01-2018
2	B-25603	3/4/2018	Jahan	Madhya Pradesh	Bhopal	Jahan	04-03-2018
3	B-25604	3/4/2018	Divsha	Rajasthan	Jaipur	Divsha	04-03-2018
4	B-25605	5/4/2018	Kasheen	West Bengal	Kolkata	Kasheen	04-05-2018

## Deduplicate Data

Hapus baris ganda  
berdasarkan kombinasi  
kolom

```
Data setelah dibersihkan:  
Order ID Order Date CustomerName State City  
0 B-25601 1/4/2018 Bharat Gujarat Ahmedabad  
1 B-25602 1/4/2018 Pearl Maharashtra Pune  
2 B-25603 3/4/2018 Jahan Madhya Pradesh Bhopal  
3 B-25604 3/4/2018 Divsha Rajasthan Jaipur  
4 B-25605 5/4/2018 Kasheen West Bengal Kolkata  
  
Nama Pelanggan Bersih Order Date Standar  
0 Bharat 04-01-2018  
1 Pearl 04-01-2018  
2 Jahan 04-03-2018  
3 Divsha 04-03-2018  
4 Kasheen 04-05-2018
```

## ✓ Hasil Akhir:

Dataset rapi, konsisten,  
bebas duplikasi → siap  
untuk analisis 🚀



# Outline

- Data Cleansing Standarisasi kolom Order ID
- Data Cleansing Standarisasi kolom Order Date
- Data Cleansing Standarisasi CustomerName
- Data Cleansing Deduplikasi Data

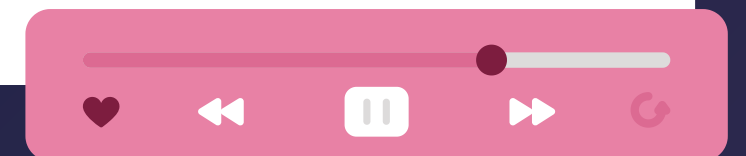




Order ID	Order Date	CustomerName	State	City
B-25601	1/4/2018	Bharat	Gujarat	Ahmedabad
B-25602	1/4/2018	Pearl	Maharashtra	Pune
B-25603	3/4/2018	Bapak Jahan	Madhya Pradesh	Bhopal
B-25604	3/4/2018	Bu Divsha	Rajasthan	Jaipur
B-25605	5/4/2018	Kasheen	West Bengal	Kolkata
B-25606	6/4/2018	Hazel	Karnataka	Bangalore
B-25607	6/4/2018	Ibu Sonakshi	Jammu and Kashmir	Kashmir
B-2560	8/4/2018	Aarushi	Tamil Nadu	Chennai
B-25609	9/4/2018	Jitesh	Uttar Pradesh	Lucknow
B-25610	9/4/2018	Yogesh	Bihar	Patna
B-25611	11/4/2018	Ibu Anita	Kerala	Thiruvananthapuram
B-25612	12/4/2018	Shrichand	Punjab	Chandigarh
B-25613	12/4/2018	Mukesh	Haryana	Chandigarh
B-2561	13-04-2018	Vandana	Himachal Pradesh	Simla
B-25615	15-04-2018	Bapak Bhavna	Sikkim	Gangtok
B-25616	15-04-2018	Kanak	Goa	Goa
B-25617	17-04-2018	Sagar	Nagaland	Kohima
B-25618	18-04-2018	Manju	Andhra Pradesh	Hyderabad

# Dataset Kotor



Dataset yang digunakan berasal dari kaggle.com dan sengaja disusun dalam kondisi kotor, yaitu terdapat perbedaan format antar data pada kolom yang sama serta adanya data yang hilang (missing values)




Order ID	Order Date	CustomerName	State	City
B-25601	1/4/2018	Bharat	Gujarat	Ahmedabad
B-25602	1/4/2018	Pearl	Maharashtra	Pune
B-25603	3/4/2018	Bapak Jahan	Madhya Pradesh	Bhopal
B-25604	3/4/2018	Bu Divsha	Rajasthan	Jaipur
B-25605	5/4/2018	Kasheen	West Bengal	Kolkata
B-25606	6/4/2018	Hazel	Karnataka	Bangalore
B-25607	6/4/2018	Ibu Sonakshi	Jammu and Kashmir	Kashmir
B-2560	8/4/2018	Aarushi	Tamil Nadu	Chennai
B-25609	9/4/2018	Jitesh	Uttar Pradesh	Lucknow
B-25610	9/4/2018	Yogesh	Bihar	Patna
B-25611	11/4/2018	Ibu Anita	Kerala	Thiruvananthapuram
B-25612	12/4/2018	Shrichand	Punjab	Chandigarh
B-25613	12/4/2018	Mukesh	Haryana	Chandigarh
B-2561	13-04-2018	Vandana	Himachal Pradesh	Simla
B-25615	15-04-2018	Bapak Bhavna	Sikkim	Gangtok
B-25616	15-04-2018	Kanak	Goa	Goa
B-25617	17-04-2018	Sagar	Nagaland	Kohima
B-25618	18-04-2018	Manju	Andhra Pradesh	Hyderabad

# Dataset Kotor


- **Order ID** : Merupakan kode unik dari setiap pesanan yang bersifat identifikasi tunggal (tidak ada duplikasi).
- **Order Date** : Tanggal saat pesanan dilakukan atau dicatat dalam sistem.
- **Customer Name** : Nama pelanggan yang melakukan pemesanan.
- **State** : Nama provinsi atau negara bagian tempat pelanggan berada.
- **City** : Nama kota tempat pelanggan melakukan pemesanan atau alamat tujuan.



Order ID
B-25601
B-25602
B-25603
B-25604
B-25605
B-25606
B-25607
B-2560
B-25609
B-25610
B-25611
B-25612
B-25613
B-2561
B-25615
B-25616
B-25617
B-25618



# Kolom Order ID

- Kolom Kode ID adalah kolom primary key, kolom pembeda antara baris data ini dengan baris data lainnya di dalam dataset pelanggan
  - Kolom primary key biasanya memiliki pola yang teratur, untuk dataset, pelanggan ini polanya adalah : a. Memiliki prefix atau awalan teks yang fix bernilai "B-". b. Memiliki suffix atau akhiran angka – dengan format lima digit angka. c. Karena pola yang fix tersebut, panjang total kolom tersebut adalah 7 karakter/digit.
  - Namun pada baris tertentu ada pola yang tidak sesuai, dimana jumlah angka digit di belakang "B-" hanya empat seperti terlihat pada screenshot.
  - Ada permasalahan inkonsistensi pola dengan panjang yang berbeda.
- 





#### CustomerName

Bharat

Pearl

Bapak Jahan

Bu Divsha

Kasheen

Hazel

Ibu Sonakshi

Aarushi

Jitesh

Yogesh

Ibu Anita

Shrichand

Mukesh

Vandana

Bapak Bhavna

Kanak

Sagar

Manju

# Kolom CustomerName

- Kolom CustomerName adalah kolom ketiga pada dataset
- Disini terlihat ada contoh penulisan panggilan untuk data "Bapak jahan, ibu sonakshi, ibu anita, dan bapak bhava". Ini pada sebagian perusahaan tidak menjadi masalah, namun untuk industri perbankan yang mengharuskan standarisasi nama berdasarkan regulasi OJK (Otoritas Jasa Keuangan), maka nama panggilan Ibu, bapak ini harus dihilangkan.



# Kolom Order Date

- Kolom Order Date adalah kolom yang berisi informasi tanggal pemesanan
- Disini sudah langsung terlihat masalahnya, yaitu ada beberapa pola yang penulisannya berbeda. Ada yang memiliki pemisah tanda minus (-) dan ada yang pakai garis miring (/)

Order Date
1/4/2018
1/4/2018
3/4/2018
3/4/2018
5/4/2018
6/4/2018
6/4/2018
8/4/2018
9/4/2018
9/4/2018
11/4/2018
12/4/2018
12/4/2018
13-04-2018
15-04-2018
15-04-2018
17-04-2018
18-04-2018

# Data Duplikat

A	B	C	D	E
Order ID	Order Date	CustomerName	State	City
B-25599	1/4/2018	Bharat	Gujarat	Ahmedabad
B-25600	1-4-1018	Bharat	Gujarat	Ahmedabad
B-25601	1/4/2018	Bharat	Gujarat	Ahmedabad
B-25602	1/4/2018	Pearl	Maharashtra	Pune
B-25603	3/4/2018	Bapak Jahan	Madhya Pradesh	Bhopal
B-25604	3/4/2018	Bu Divsha	Rajasthan	Jaipur
B-25605	5/4/2018	Kasheen	West Bengal	Kolkata
B-25606	6/4/2018	Hazel	Karnataka	Bangalore
B-25607	6/4/2018	Ibu Sonakshi	Jammu and Kashmir	Kashmir
B-2560	8/4/2018	Aarushi	Tamil Nadu	Chennai
B-25609	9/4/2018	Jitesh	Uttar Pradesh	Lucknow
B-25610	9/4/2018	Yogesh	Bihar	Patna
B-25611	11/4/2018	Ibu Anita	Kerala	Thiruvananthapuram
B-25612	12/4/2018	Shrichand	Punjab	Chandigarh
B-25613	12/4/2018	Mukesh	Haryana	Chandigarh
B-2561	13-04-2018	Vandana	Himachal Pradesh	Simla
B-25615	15-04-2018	Bapak Bhavna	Sikkim	Gangtok
B-25616	15-04-2018	Kanak	Goa	Goa

- Selain isi data yang tidak standar, ternyata dataset ini juga memiliki duplikat untuk pelanggan yang sama.
- Terlihat tiga baris data dengan nama Bharat ini sebenarnya sama terlihat dari isi data CustomerName, State, City dan Order Date. Hanya saja format penulisan Order Date berbeda.

# Data Bersih



Data setelah dibersihkan:

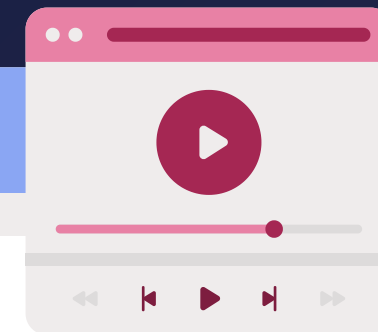
	Order ID	Order Date	CustomerName	State	City	\
0	B-25601	1/4/2018	Bharat	Gujarat	Ahmedabad	
1	B-25602	1/4/2018	Pearl	Maharashtra	Pune	
2	B-25603	3/4/2018	Jahan	Madhya Pradesh	Bhopal	
3	B-25604	3/4/2018	Divsha	Rajasthan	Jaipur	
4	B-25605	5/4/2018	Kasheen	West Bengal	Kolkata	

Nama Pelanggan Bersih Order Date Standar

0	Bharat	04-01-2018
1	Pearl	04-01-2018
2	Jahan	04-03-2018
3	Divsha	04-03-2018
4	Kasheen	04-05-2018

Data bersih (clean data) adalah dataset yang sudah melalui proses data cleansing, sehingga:

- Bebas dari duplikasi → tidak ada baris yang tercatat lebih dari sekali.
- Konsisten formatnya → misalnya tanggal seragam (dd-mm-yyyy), ID berurutan rapi, nama pelanggan tanpa simbol/angka.
- Lengkap dan relevan → hanya berisi data yang diperlukan untuk analisis.
- Minim error atau noise → salah ketik, karakter aneh, atau nilai kosong sudah diperbaiki/diatasi.



# Kesimpulan

- Data Cleansing adalah proses penting untuk memastikan dataset rapi, konsisten, dan akurat.
- Data bersih → menjadi fondasi analisis yang valid dan kredibel.
- Dengan data yang sudah bersih:
  - ✓ Analisis lebih cepat & efisien.
  - ✓ Insight lebih tepat sasaran.
  - ✓ Keputusan bisnis lebih terpercaya.
- 📌 Ingat: Garbage in, garbage out → kualitas data menentukan kualitas hasil analisis.





**Terima  
Kasih**

