



# Pengantar Data Sains

#5 Meeting

Missing & Conversion Data

Ferdian Bangkit Wijaya, S.Stat., M.Si

NIP. 199005202024061001

# Missing Data



Apa itu Data Hilang (Missing Data)?

Terjadi ketika tidak ada nilai data yang tersimpan untuk sebuah variabel dalam suatu observasi.

Mengapa ini menjadi Masalah Besar?

- Mengurangi Kekuatan Statistik: Lebih sedikit data berarti presisi analisis menurun.
- Dapat Menyebabkan Bias: Jika data tidak hilang secara acak, kesimpulan yang kita ambil bisa sepenuhnya salah.
- Merusak Algoritma: Sebagian besar algoritma machine learning tidak dapat bekerja dengan data yang hilang dan akan menghasilkan error.

# Alasan Data menjadi Hilang

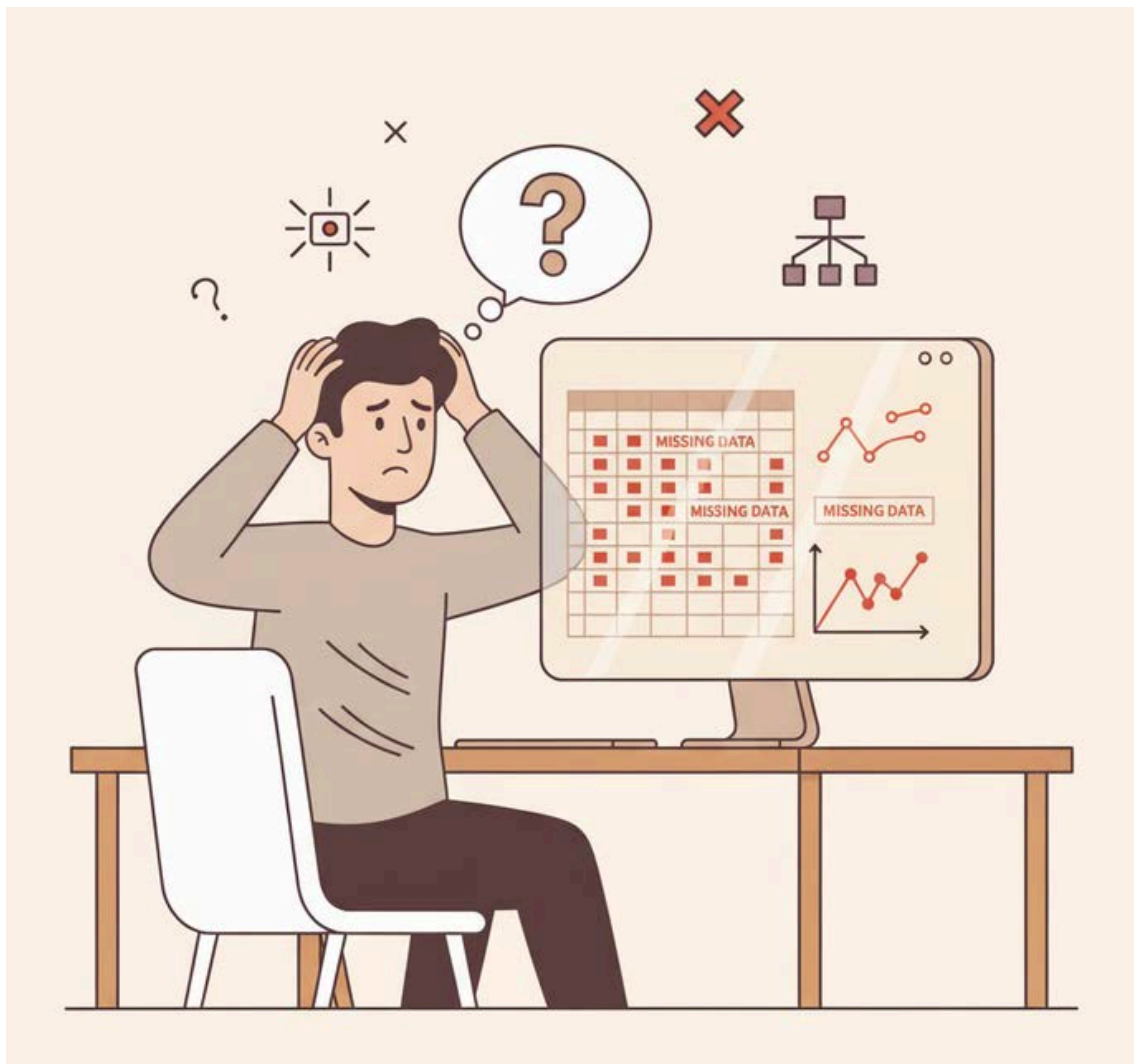


Data bisa hilang karena berbagai alasan, baik teknis maupun non-teknis:

- Kesalahan Manusia (Human Error): Kesalahan saat memasukkan data, lupa mengisi formulir.
- Masalah Teknis: Kegagalan sensor, error saat transfer data, atau masalah pada software.
- Privasi: Responden sengaja tidak mengisi data sensitif seperti pendapatan atau penyakit.
- Struktur Pengumpulan Data: Pertanyaan survei yang bersifat kondisional (misalnya, pertanyaan "bagaimana pengalaman melahirkan?" hanya berlaku untuk wanita yang punya anak).
- Atrisi: Dalam studi jangka panjang, partisipan mungkin keluar dari studi (drop out).



# Tipe Missing Data



Ada 3 mekanisme utama mengapa data bisa hilang. Memahaminya membantu kita memilih strategi yang tepat.

1. MCAR (Missing Completely at Random):

- Penyebab data hilang sama sekali tidak berhubungan dengan data apa pun, baik yang ada maupun yang hilang.

2. MAR (Missing at Random):

- Penyebab data hilang berhubungan dengan data lain yang ada (terobservasi), tetapi tidak dengan data yang hilang itu sendiri.

3. MNAR (Missing Not at Random):

- Penyebab data hilang berhubungan dengan nilai dari data yang hilang itu sendiri. Ini yang paling problematik.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

# MCAR (Missing Completely at Random)

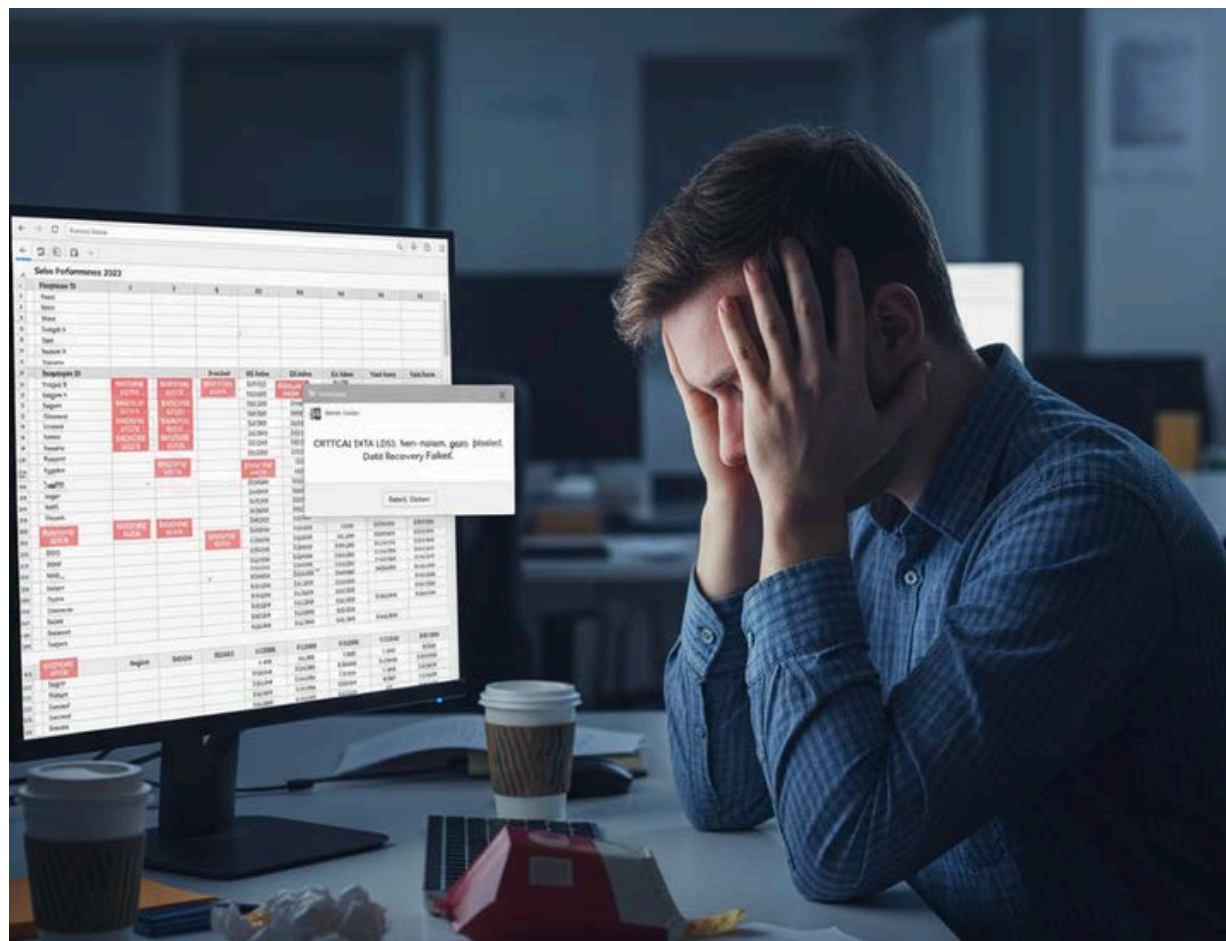
Sales Performance 2023

Employee ID	1	3	6	02	03	08	04	01
West					MISSING DATA		MISSING DATA	
West								
West				MISSING DATA				
Widget A								MISSING DATA
West								
Brdget A								
Vianton		MISSING DATA						
Employee ID			Product	Q1 Sales	Q2 Sales	Q1 Sales	Total Sales	Total Sales
Widget B			MISSING DATA	\$35,000	MISSING DATA	\$50,000		
Widget A				\$50,000				
Region				\$30,000	\$30,000			
Watume				\$10,000	\$50,000	\$50,000		
Fropiet				\$10,000	\$50,000	\$50,000		
Sinton				\$15,000	\$10,000	\$30,000	MISSING DATA	
Eogrials				\$10,000	\$30,000	\$60,000		
Pispigel				\$10,000	\$32,000	\$67,000		
Potalen		MISSING DATA		MISSING DATA	\$5,000	MISSING DATA		
Region				\$10,000	\$10,000	\$50,000		
Regior				\$10,000	\$50,000	\$40,000		
Realt			MISSING DATA	\$50,000	\$50,000	\$50,000		
Prctesk				\$33,000	\$30,000			
MISSING DATA		MISSING DATA		\$35,000	\$20,000	5,000	\$150,000	\$100,000
E123			MISSING DATA	\$10,000	\$50,000	50,000	\$150,000	\$150,000
Efsgil				\$10,000	\$10,000	\$50,000	\$150,000	\$100,000
Floyit	MISSING DATA			\$10,000	MISSING DATA	\$10,000	\$160,000	\$150,000
Velton				\$13,000	MISSING DATA	\$19,000	\$150,000	\$150,000
Regiar				\$10,000	\$10,000	\$50,000	\$140,000	\$100,000
Feyles				\$10,000	\$10,000	\$50,000	MISSING DATA	\$190,000
Resetmer				\$15,000	\$10,000	\$10,000	\$160,000	\$150,000
Geale				\$10,000	\$25,000	\$56,000	\$160,000	\$150,000
Setomat				\$19,000	\$10,000	\$10,000		
Eefgne				\$15,000	\$19,000	\$30,000	\$160,000	

- Artinya: Kehilangan data bersifat murni acak, seperti sebuah undian. Tidak ada pola sama sekali.
- Analogi: Seorang responden survei secara tidak sengaja melewati satu pertanyaan karena halamannya menempel.
- Contoh: Dalam dataset pasien, beberapa nilai tekanan darah hilang karena sampel darahnya tidak sengaja terjatuh di lab. Kejadian ini tidak ada hubungannya dengan tekanan darah pasien itu sendiri atau data lainnya.
- Implikasi: Jika data benar-benar MCAR, maka data yang tersisa masih merupakan sampel yang representatif. Metode penghapusan data bisa aman digunakan jika jumlah data yang hilang sedikit.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.

# MAR (Missing at Random)

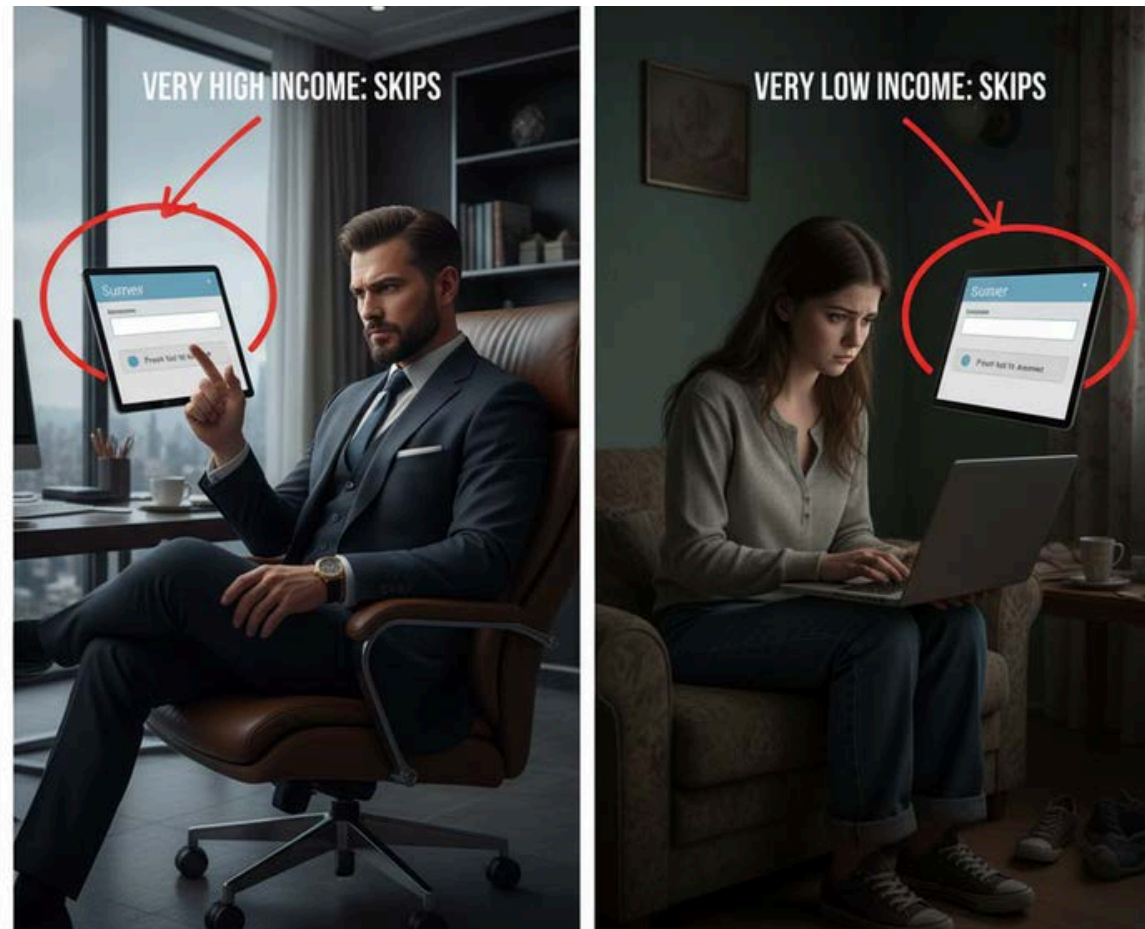


- Ada pola, tapi bisa dijelaskan oleh variabel lain.
- Contoh: Pria lebih jarang mengisi survei tentang depresi dibandingkan wanita. Jadi, data skor depresi yang hilang bergantung pada kolom jenis kelamin (yang terisi), bukan pada seberapa depresi orang itu.
- Untuk MAR: "Kita bisa 'menebak' skor depresi yang hilang dengan melihat data lain, seperti jenis kelamin."
- Implikasi: Metode penghapusan akan menghasilkan bias. Metode imputasi yang lebih canggih (KNN, MICE) sangat efektif di sini.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.



# MNAR (Missing Not at Random)



- Pola kehilangan data terkait dengan nilai itu sendiri.
- Contoh: Orang dengan pendapatan sangat tinggi atau sangat rendah cenderung tidak mengisi kolom pendapatan.
- Untuk MNAR: "Kita tidak bisa menebak pendapatan yang hilang, karena alasan data itu hilang adalah justru karena nilainya sendiri yang ekstrim."
- Implikasi: Ini adalah skenario terburuk. Semua metode sederhana akan menghasilkan bias. Penanganannya memerlukan pemodelan statistik yang kompleks atau pengetahuan domain yang mendalam.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.

# Penanganan Missing Data : Deletion

## Metode Penghapusan (Deletion)

- Pendekatan Paling Sederhana: "Jika ada yang rusak, buang saja."
- Tujuan: Mendapatkan dataset lengkap dengan cepat, meskipun harus mengorbankan sebagian data.

### Dua Jenis Utama:

- Listwise Deletion (Hapus Baris): Menghapus seluruh baris jika ada satu saja nilai yang hilang.
- Pairwise Deletion (Hapus Pasangan): Hanya digunakan dalam perhitungan statistik tertentu.

### Kapan Boleh Digunakan?

- Hanya jika yakin data hilang secara MCAR.
- Dan jika proporsi data yang hilang sangat kecil (aturan praktis  $< 5\%$ ).

Ini adalah "opsi nuklir". Cepat dan mudah, tetapi berpotensi menghancurkan banyak informasi berharga dan bisa membuat hasil analisis bias jika asumsi MCAR tidak terpenuhi.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.



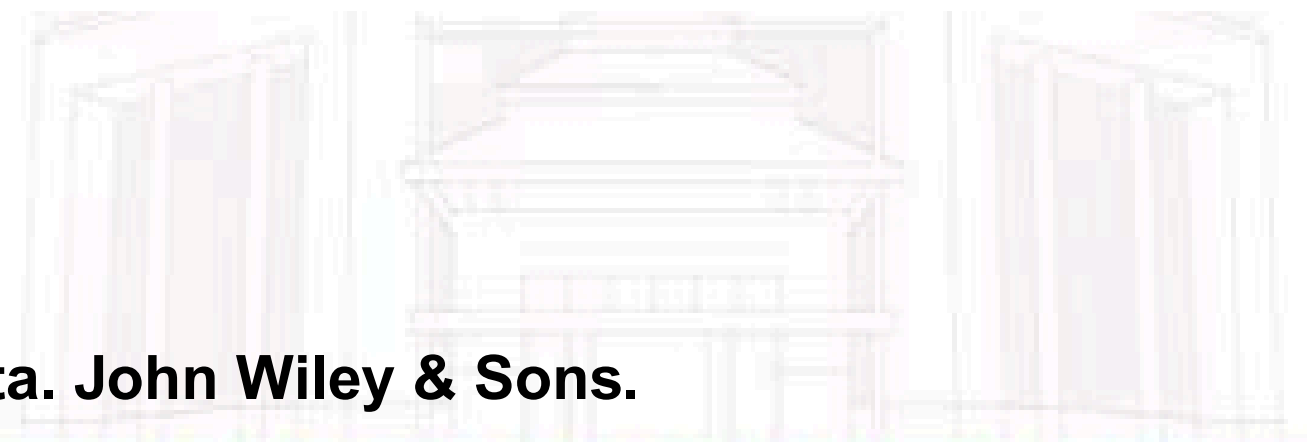


# Penanganan Missing Data : Deletion



Usia	Pendapatan	Rating
25	50 Juta	4
30	(hilang)	5
(hilang)	60 Juta	3
45	70 Juta	(hilang)

- **Listwise Deletion:**
  - Hanya baris pertama yang akan tersisa. Kita kehilangan 75% data!
- **Pairwise Deletion:**
  - Saat menghitung korelasi Usia & Pendapatan, hanya baris 1 & 4 yang dipakai.
  - Saat menghitung korelasi Usia & Rating, hanya baris 1 & 2 yang dipakai.
  - Masalah: Perhitungan yang berbeda menggunakan jumlah data yang berbeda, bisa menyebabkan masalah statistik (misal: matriks korelasi yang tidak konsisten).

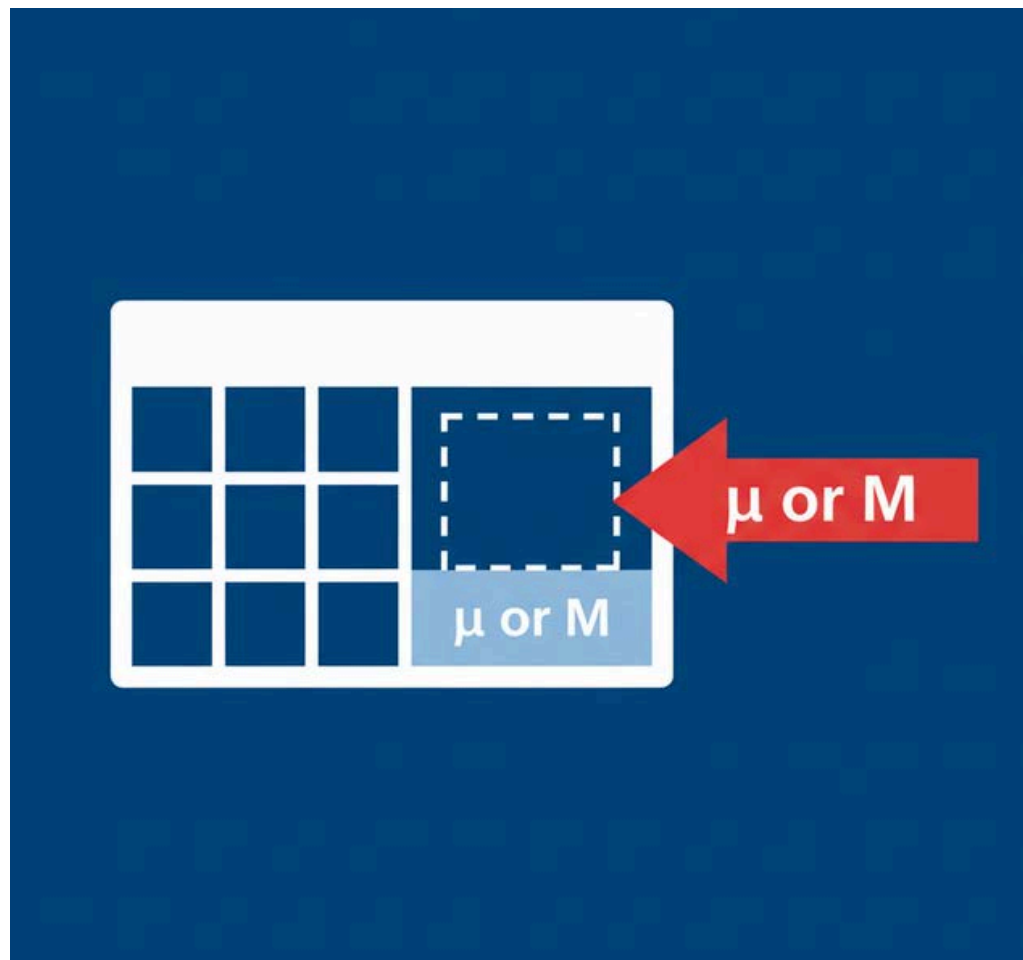


Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.

# Penanganan Missing Data : Imputasi

## Metode Imputasi Sederhana

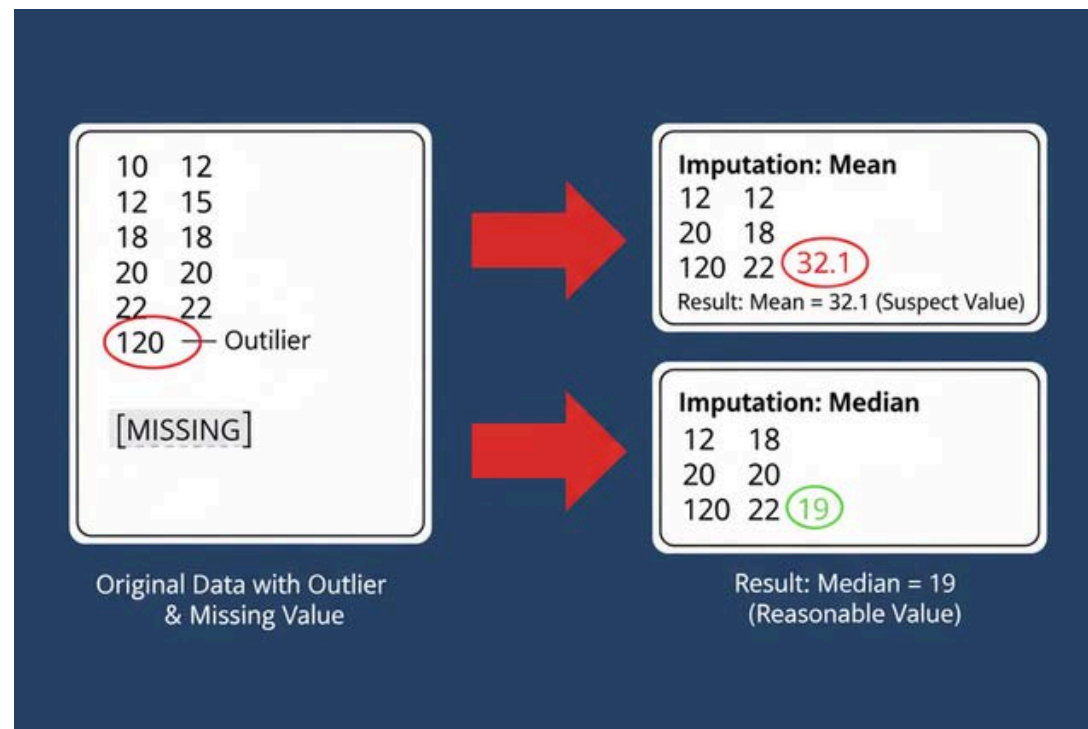
- Pendekatan yang Lebih Baik: Daripada membuang, mari kita "isi bagian yang kosong" dengan tebakan yang masuk akal.
- Imputasi: Proses mengganti data yang hilang dengan nilai pengganti.
- Tujuan: Mempertahankan ukuran sampel data sehingga tidak kehilangan kekuatan statistik.
- Metode Sederhana: Menggunakan ringkasan statistik dari kolom itu sendiri untuk mengisi nilai yang hilang.



Seorang restorator tidak akan membuang lukisan bersejarah hanya karena ada sedikit lubang. Mereka akan dengan hati-hati menambal lubang itu menggunakan bahan yang semirip mungkin dengan aslinya. Itulah yang kita lakukan dengan imputasi.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

# Penanganan Missing Data : Imputasi



- **Imputasi Mean (Rata-rata):** Mengganti nilai hilang dengan mean dari kolom.
- Contoh: [10, 20, 30, ?, 1000]. Mean = 265. Nilai hilang diisi 265.
- Kelemahan: Sangat sensitif terhadap outlier (pencilan) seperti nilai 1000.
- **Imputasi Median:** Mengganti nilai hilang dengan median (nilai tengah) dari kolom.
- Contoh: [10, 20, 30, ?, 1000]. Median = 25. Nilai hilang diisi 25.
- Kelebihan: Jauh lebih robust (tahan) terhadap outlier. Seringkali pilihan yang lebih aman.
- **Imputasi Modus:** Mengganti dengan modus (nilai paling sering muncul).
- Cocok untuk: Angka diskrit, seperti jumlah anak atau jumlah kamar.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.



# Penanganan Missing Data : Imputasi



- **Imputasi Modus (Most Frequent Value):**
  - Pilihan paling umum dan logis. Mengganti nilai hilang dengan kategori yang paling sering muncul.
  - Contoh: Di kolom Kota, jika "Jakarta" muncul paling banyak, maka semua Kota yang hilang akan diisi dengan "Jakarta".
- **Imputasi Kategori Konstan:**
  - Membuat kategori baru untuk nilai yang hilang.
  - Contoh: Mengisi semua Kota yang hilang dengan nilai "Tidak Diketahui".
  - Kapan Berguna? Jika fakta bahwa data tersebut hilang adalah sebuah informasi penting itu sendiri. Mungkin ada alasan mengapa data kota dari sekelompok orang ini tidak tercatat, dan kita ingin model kita mempelajari pola tersebut.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.

# Penanganan Missing Data : Imputasi

Meskipun berguna, imputasi sederhana memiliki kelemahan signifikan:



1. Mengabaikan Hubungan Antar Variabel:
  - Imputasi mean/median untuk usia tidak mempertimbangkan informasi dari kolom lain (misalnya, jabatan atau lama bekerja).
2. Mengurangi Varians (Keragaman Data):
  - Dengan mengisi banyak nilai hilang dengan satu angka yang sama (misal, median), kita secara artifisial mengurangi keragaman alami dalam data.
3. Menghasilkan Estimasi yang Bias:
  - Karena dua poin di atas, hasil standar deviasi, korelasi, dan koefisien regresi bisa menjadi tidak akurat.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

# Imputasi Lanjut - KNN Imputation

## K-Nearest Neighbors (KNN) Imputation:

- Ide Utama: "Beritahu saya siapa teman-temanmu, maka akan kuberitahu siapa kamu."


### Konsep:

1. Untuk sebuah baris dengan nilai usia yang hilang, cari 'k' baris lain (misal,  $k=5$ ) yang paling mirip dengannya berdasarkan semua kolom lain yang lengkap (seperti pendapatan, lama bekerja).
2. Nilai usia yang hilang diisi dengan rata-rata atau median dari nilai usia kelima "tetangga terdekat" tersebut.

Kelebihan: Jauh lebih akurat karena memanfaatkan hubungan antar variabel.

**THE NEAREST NEIGHBORS CONCEPT**

Original Room: Too Much Variation



Age = ?

**X** Average Age of Everyone

Problem: Guessing age based EVERYONE'S average is inaccurate.

Finding Neighbors: A Better Guess



Age = ?

Solution: Find 5 MOST SIMILAR people...

Age Guess = Average Age of 5 Neighbors

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.



# Imputasi Lanjut - Iterative Imputation

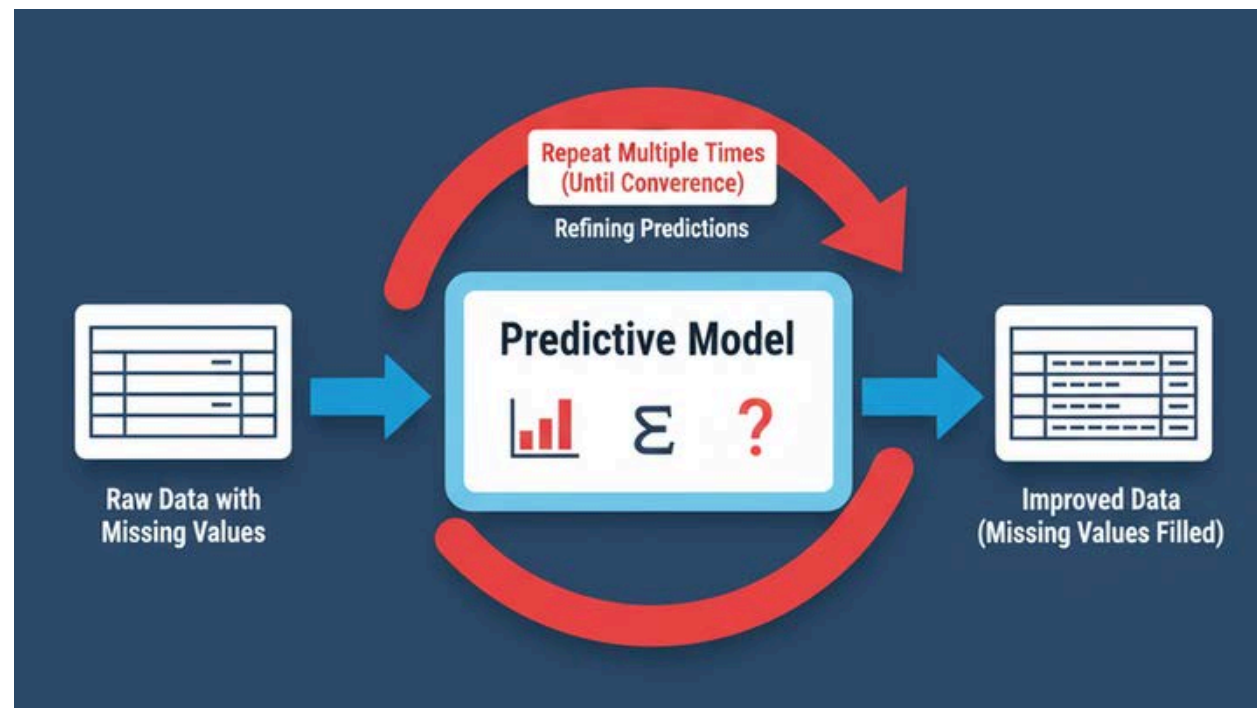
## MICE (Multivariate Imputation by Chained Equations):

- Ide Utama: Menggunakan machine learning untuk memprediksi nilai yang hilang.

### Konsep:

1. Isi semua nilai hilang dengan imputasi sederhana (misal, median).
2. Ambil satu kolom (misal, usia) dan kembalikan nilai yang diimputasi menjadi hilang lagi.
3. Latih sebuah model regresi di mana usia adalah target (Y) dan semua kolom lain adalah fitur (X).
4. Gunakan model ini untuk memprediksi nilai usia yang hilang.
5. Ulangi proses ini untuk setiap kolom yang memiliki nilai hilang, dan lakukan beberapa putaran (iterasi) hingga semua prediksi stabil.

Kelebihan: Sering dianggap sebagai gold standard untuk imputasi karena sangat akurat dan menjaga struktur data.



Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.



# Study Case : Data Titanic



Running di R





# Study Case : Data Titanic



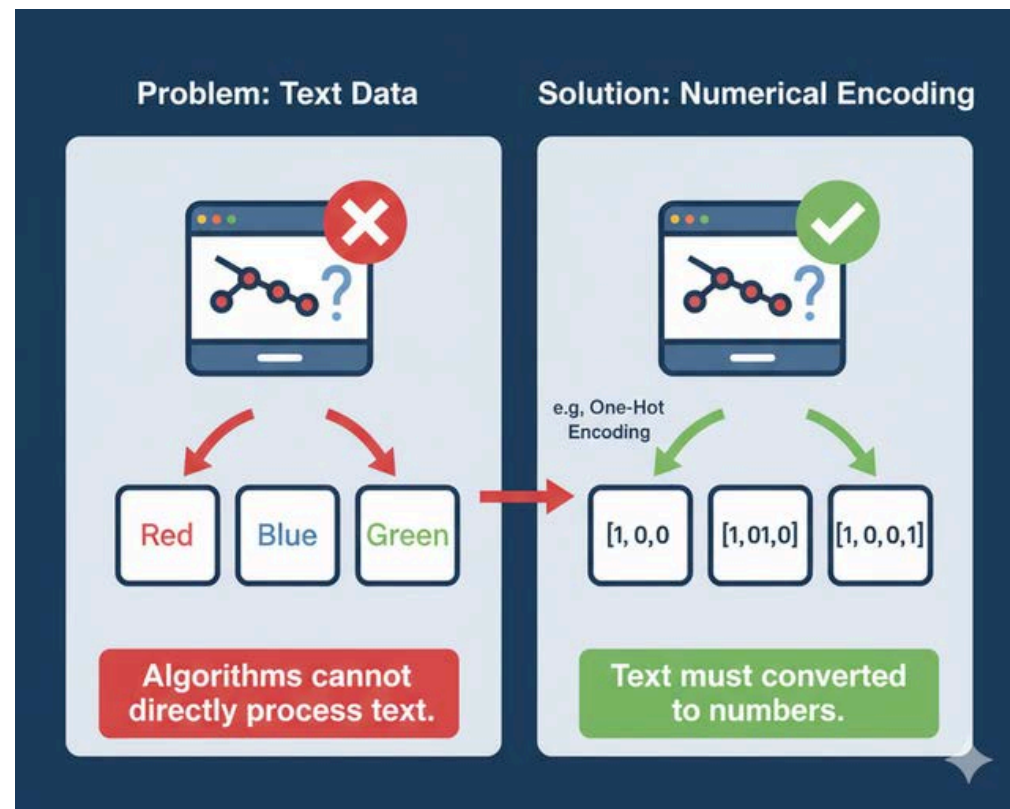
Running di Python





# Data Konversi

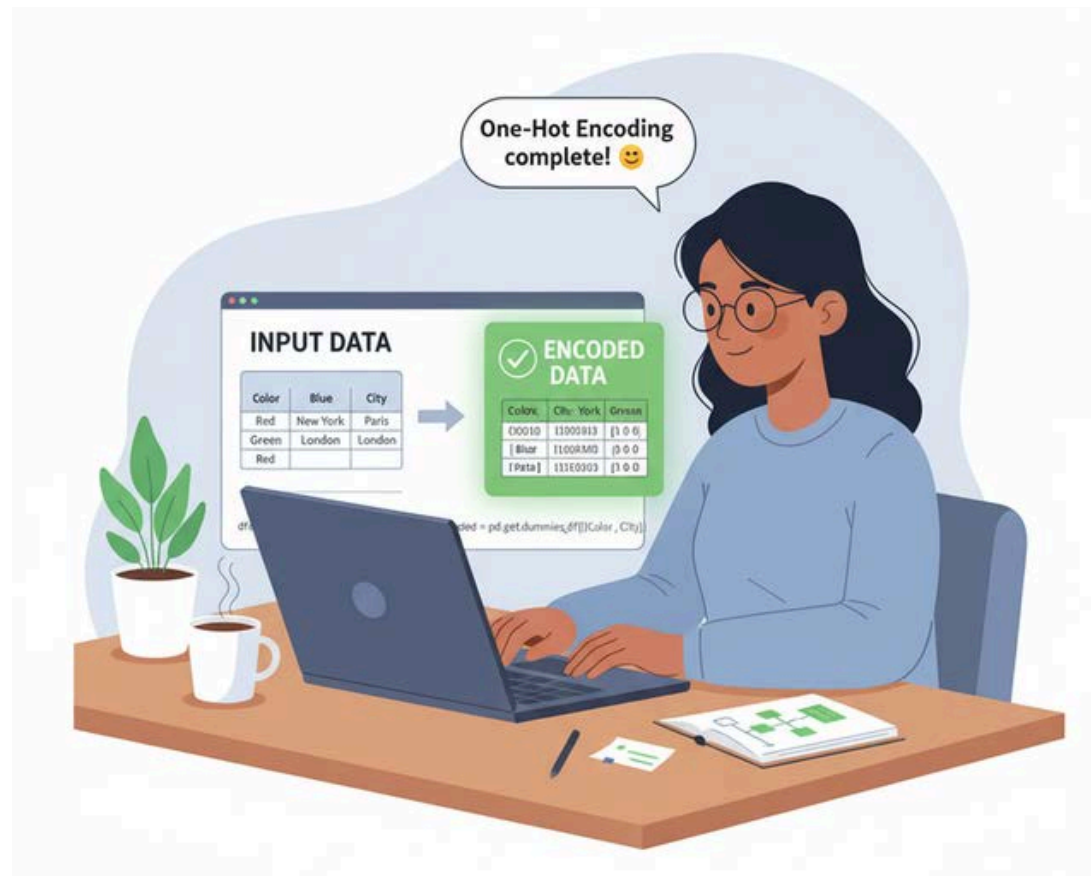
## Mengapa Kita Perlu Mengonversi Data?



- Sebagian besar algoritma machine learning didasarkan pada operasi matematika (penjumlahan, perkalian, perhitungan jarak, gradien, dll).
- Algoritma tidak bisa memproses data teks/kategori secara langsung.
  - Regresi Linear tidak bisa mengalikan usia dengan “Pria”.
  - K-Nearest Neighbors tidak bisa menghitung "jarak" antara “Jakarta” dan “Surabaya”.
- Tujuan Konversi: Mengubah semua fitur (variabel independen) dan target (variabel dependen) menjadi representasi numerik tanpa menghilangkan informasi penting atau menambahkan informasi yang salah.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

# Konversi Tipe Data Dasar



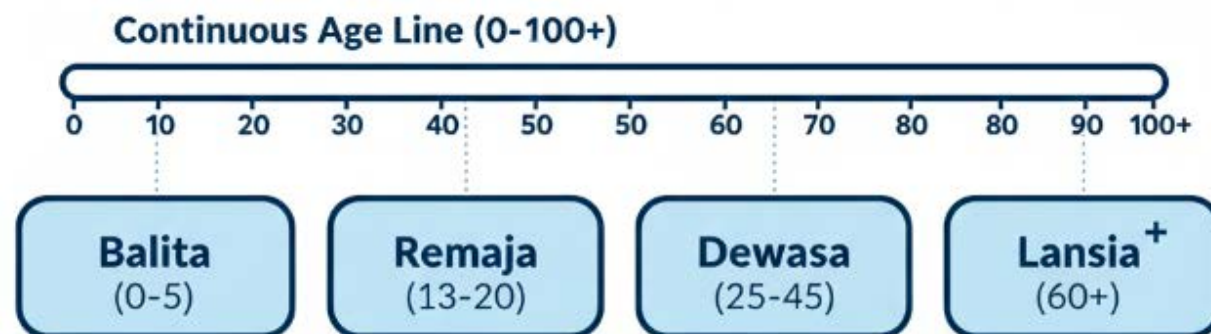
- Ini adalah langkah pertama yang paling sering dilakukan dalam data wrangling.
- Mengubah tipe data sebuah kolom secara eksplisit agar sesuai.
- Contoh Umum:
- String ke Numerik: Kolom harga ("150,000") diubah menjadi integer (150000).
- Object/String ke Datetime: Kolom tanggal\_transaksi ("2025-09-21") diubah menjadi tipe datetime untuk analisis waktu.
- Numerik ke String/Object: Kolom ID unik (misal: Kode Pos 14021) diubah menjadi teks agar tidak diperlakukan sebagai angka.

Jika tanggal masih berupa teks, kita tidak bisa menghitung selisih hari. Jika harga masih berupa teks, kita tidak bisa menghitung total pendapatan. Tipe data yang benar adalah fondasi analisis.

McKinney, W. (2017). Python for Data Analysis, 2nd Edition. O'Reilly Media.

# Binning atau Discretization

- Konsep: Mengubah variabel numerik kontinu menjadi variabel kategorikal diskrit (kelompok/bins).
- Contoh: Variabel Usia (nilai dari 0-100) diubah menjadi Kelompok Usia ("Balita", "Remaja", "Dewasa", "Lansia").



- Mengapa Melakukan Ini?
- Menyederhanakan Model: Terkadang hubungan antara fitur dan target tidak linear. Binning dapat membantu model menangkap pola kelompok dengan lebih mudah.
- Menangani Outlier: Mengelompokkan nilai ekstrim ke dalam satu bin dapat mengurangi dampaknya pada model.

Bagi sebuah perusahaan asuransi, mungkin tidak ada perbedaan risiko yang signifikan antara seseorang berusia 31 atau 32 tahun. Tapi ada perbedaan besar antara kelompok usia 20-30an dan 60-70an. Binning membantu model fokus pada perbedaan yang lebih signifikan ini.

Kuhn, M., & Johnson, K. (2013). **Applied Predictive Modeling**. Springer.



# Variabel Kategorikal

- Ini adalah inti dari masalah konversi data.
- Variabel Kategorikal: Variabel yang nilainya berasal dari sekumpulan kategori yang terbatas.

Dua Jenis Utama:

- Nominal: Kategori yang tidak memiliki tingkatan atau urutan.
- Contoh: Kota, Jenis Kelamin, Warna.
- Ordinal: Kategori yang memiliki tingkatan atau urutan yang jelas.
- Contoh: Tingkat Pendidikan, Ukuran Baju (S, M, L), Rating Kepuasan.
- Memilih teknik encoding yang salah bisa merusak model.

Kesalahan paling umum adalah memperlakukan variabel nominal seolah-olah ia memiliki urutan. Kita akan lihat mengapa ini berbahaya.

Stevens, S. S. (1946). "On the theory of scales of measurement." Science.

# Label Encoding

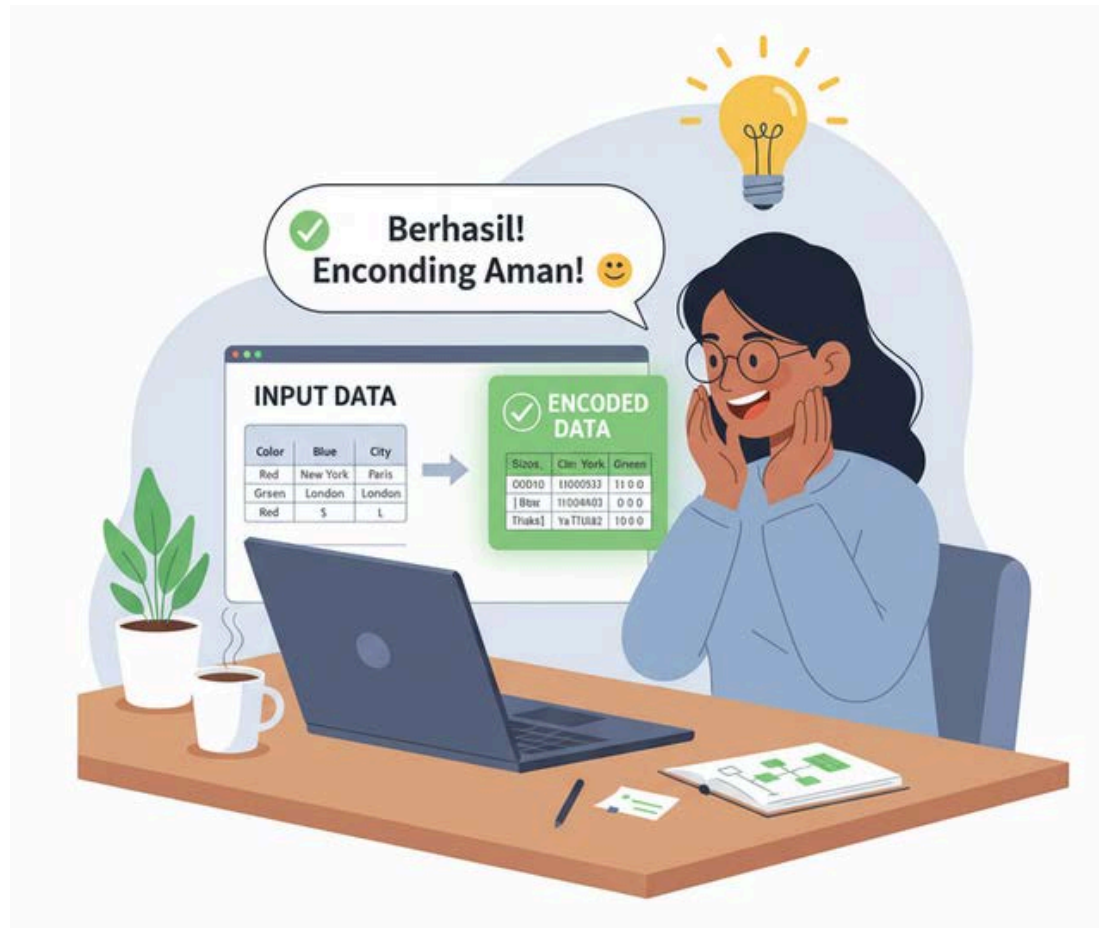
- Konsep: Menetapkan setiap kategori unik ke sebuah angka integer.
- Merah -> 0
- Biru -> 1
- Hijau -> 2
- Masalah Besar: Metode ini secara implisit menciptakan urutan artifisial.
- Algoritma akan "berpikir" bahwa Hijau (2) > Biru (1), yang secara logis salah. Ini dapat memasukkan informasi palsu ke dalam model.
- Kesimpulan: SANGAT BERBAHAYA untuk digunakan pada variabel fitur yang bersifat NOMINAL.



Bagi algoritma, angka 2 lebih besar dari 1. Jika kita menggunakan Label Encoding pada warna, kita secara tidak sengaja memberitahu model bahwa 'Hijau' memiliki nilai lebih tinggi daripada 'Biru', yang sama sekali tidak masuk akal."

Stevens, S. S. (1946). "On the theory of scales of measurement." Science.

# Label Encoding



- Label Encoding tetap sangat berguna dalam dua skenario spesifik:
- Untuk Variabel Fitur yang Bersifat Ordinal:
- Jika variabelnya memang memiliki tingkatan, urutan numerik justru menyimpan informasi yang benar.
- Contoh Ukuran Baju: S -> 0, M -> 1, L -> 2, XL -> 3. Di sini, XL (3) > M (1) adalah pernyataan yang valid.
- Untuk Variabel Target (Y) dalam Masalah Klasifikasi:
- Ini adalah penggunaan yang paling umum dan wajib.
- Model perlu memprediksi angka, bukan teks.
- Contoh Churn: No -> 0, Yes -> 1.

Stevens, S. S. (1946). "On the theory of scales of measurement." Science.



# One-Hot Encoding (OHE)



- Solusi standar untuk Variabel Nominal.
- Konsep: Membuat kolom biner (0 atau 1) baru untuk setiap kategori unik. Kolom baru ini disebut dummy variable.
- Contoh Warna:
  - | Warna | | Warna\_Merah | Warna\_Biru | Warna\_Hijau |
  - | Merah | ➡ | 1 | 0 | 0 |
  - | Biru | ➡ | 0 | 1 | 0 |
  - | Hijau | ➡ | 0 | 0 | 1 |
- Hasil: Tidak ada urutan. Setiap kategori direpresentasikan sebagai vektor on/off yang unik.

Scikit-learn Developers. User Guide: Encoding categorical features. [scikit-learn.org](https://scikit-learn.org).  
Wickham, H., et al. The tidymodels framework. [tidymodels.org](https://tidymodels.org).





# Study Case : Adult Census Income



Running di Python





# Study Case : Adult Census Income



Running di R





# SEE YOU NEXT WEEK !

**Ferdian Bangkit Wijaya, S.Stat., M.Si**

**NIP. 199005202024061001**

**[ferdian.bangkit@untirta.ac.id](mailto:ferdian.bangkit@untirta.ac.id)**