Bab 03B — Eksplorasi, Visualisasi, dan Kualitas Data

Cara membaca bab ini

Bab ini sengaja ditempatkan setelah Bab 3 dan sebelum Bab 4. Alasannya: sebelum pembaca masuk ke matematika AI, probabilitas, ML, dan model, pembaca harus kuat dulu dalam membaca data. Di lapangan, kegagalan proyek AI jarang hanya karena algoritma kurang canggih. Lebih sering karena data kotor, tipe data keliru, visualisasi menipu, split data bocor, atau model dipilih tanpa memahami bentuk data.

Bab ini adalah bab khusus data. Formatnya mengikuti Bab 7: berbasis subbab, detail, dan praktis. Setiap subbab punya konsep, contoh, cara membaca rumus jika ada, kesalahan umum, dan tes cepat. Tujuannya agar pembaca tidak hanya bisa menjalankan kode, tetapi bisa bertanya: “Data ini jenis apa? Grafik apa yang pantas? Apa yang harus dibersihkan? Apa yang bisa bocor? Model apa yang masuk akal?”

Subbab 1 — Mengapa eksplorasi data adalah fondasi AI

Inti subbab: model AI hanya sebaik data, definisi masalah, dan proses evaluasinya.

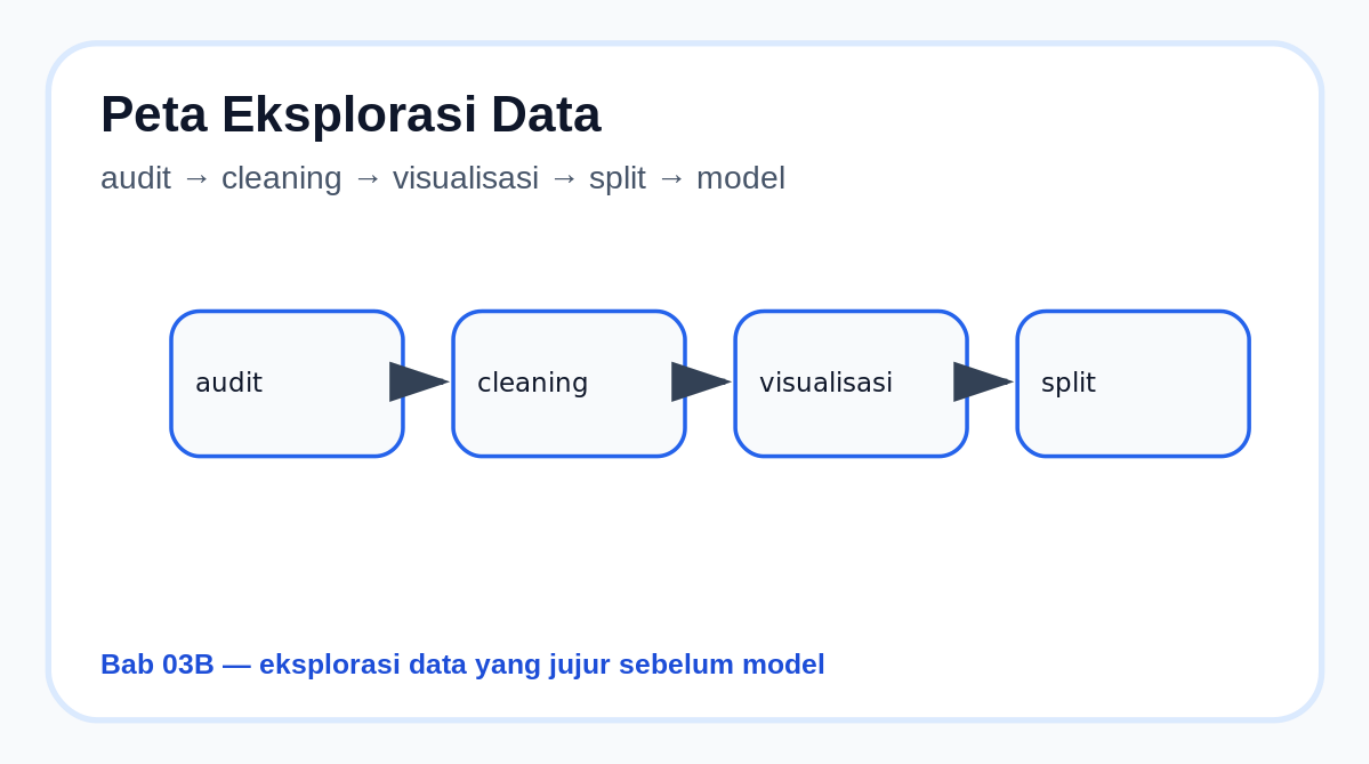
Bayangkan seseorang ingin membuat model prediksi penjualan warung. Ia langsung mengambil algoritma canggih, menjalankan training, lalu mendapat akurasi bagus. Setelah dicek, ternyata data training berisi kolom `total_penjualan_besok` yang seharusnya belum diketahui saat prediksi. Model terlihat pintar karena bocor. Ini bukan kecerdasan; ini kesalahan data.

Eksplorasi data atau EDA (*exploratory data analysis*) adalah proses memahami data sebelum membuat model. EDA bukan aktivitas sampingan, melainkan fondasi. Kita memeriksa bentuk data, tipe kolom, rentang nilai, missing value, duplikasi, distribusi, outlier, hubungan antarfitur, bias, dan potensi leakage.

Pipeline aman:

pertanyaan bisnis → audit data → cleaning → visualisasi → insight → preprocessing → split → model → evaluasi

Cara membaca pipeline: panah berarti urutan berpikir. Model datang setelah kita memahami data, bukan sebelum.



Peta eksplorasi data

Contoh pertanyaan yang harus dijawab sebelum model:

Apa satu baris data merepresentasikan apa?
Apa target yang ingin diprediksi?
Kapan setiap kolom tersedia?
Apakah data mewakili kondisi lapangan?
Apakah ada kolom yang tidak boleh dipakai?
Apa konsekuensi jika model salah?

Kesalahan umum: menganggap EDA hanya membuat grafik cantik. Grafik adalah alat berpikir. Jika grafik tidak menjawab pertanyaan, ia hanya dekorasi.

Tes cepat subbab 1

1. Mengapa model bagus bisa gagal jika data salah?

2. Sebutkan tiga hal yang harus dicek sebelum training model.

3. Apa beda grafik sebagai dekorasi dan grafik sebagai alat analisis?

Subbab 2 — Unit observasi: satu baris data mewakili apa?

Inti subbab: sebelum membaca kolom, tentukan dulu arti satu baris.

Unit observasi adalah “benda” yang direkam oleh satu baris data. Satu baris bisa berarti satu pelanggan, satu transaksi, satu kunjungan website, satu sensor per menit, satu dokumen, satu gambar, atau satu pasien. Banyak kekacauan analisis muncul karena unit observasi tidak jelas.

Contoh:

Dataset	Satu baris berarti	Risiko jika salah paham
Transaksi toko	satu transaksi	pelanggan yang sering transaksi dihitung terlalu dominan
Pelanggan	satu pelanggan	riwayat waktu bisa hilang
Sensor mesin	satu timestamp sensor	harus hati-hati dengan urutan waktu

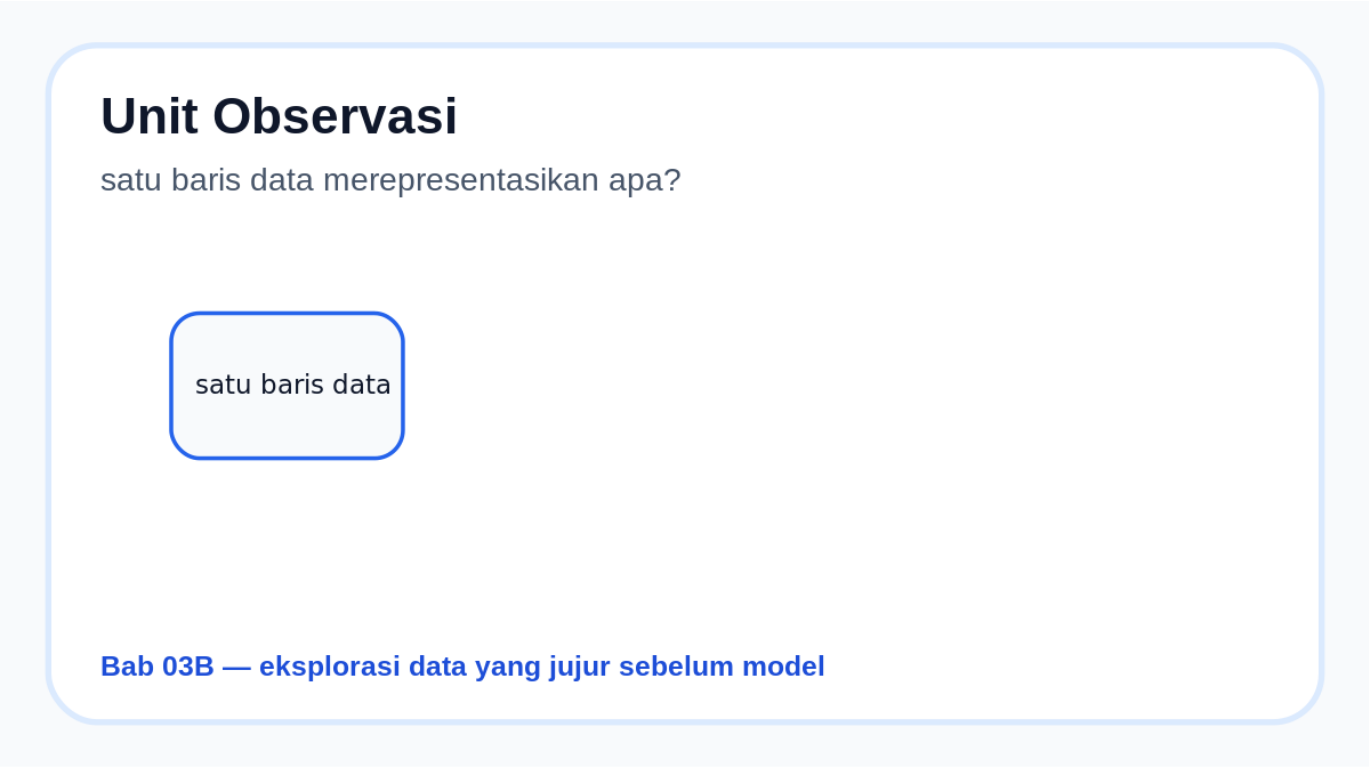
Dataset	Satu baris berarti	Risiko jika salah paham
Komentar marketplace	satu teks ulasan	butuh preprocessing teks

Jika target adalah “pelanggan churn atau tidak”, satu baris sebaiknya merepresentasikan pelanggan pada periode tertentu, bukan transaksi acak. Jika target adalah “penjualan besok”, satu baris bisa merepresentasikan satu hari toko.

Rumus ringkas jumlah baris dan fitur

$$X \in R^{(n \times d)}$$

Cara membaca rumus: `X` adalah matriks data. `n` adalah jumlah baris atau observasi. `d` adalah jumlah fitur atau kolom input. Jika ada 1.000 pelanggan dan 12 fitur, maka $X \in R^{(1000 \times 12)}$.



Unit observasi

Contoh hitung: dataset berisi 30 hari penjualan, masing-masing punya fitur cuaca, promo, harga, dan stok. Maka:

$$n = 30$$
$$d = 4$$
$$X \in R^{(30 \times 4)}$$

Tes cepat subbab 2

1. Apa unit observasi untuk dataset komentar produk?

2. Mengapa dataset transaksi dan dataset pelanggan tidak sama?

3. Jika ada 500 baris dan 8 fitur, tulis ukuran matriks `X`.

Subbab 3 — Tipe data tingkat besar: terstruktur, semi terstruktur, tidak terstruktur

Inti subbab: bentuk data menentukan cara menyimpan, membersihkan, memvisualisasikan, dan memilih model.

Tiga keluarga besar data:

1. Data terstruktur: tabel rapi, kolom jelas, tipe data relatif konsisten. Contoh: CSV penjualan, database pelanggan, spreadsheet nilai siswa.

2. Data semi terstruktur: punya struktur tetapi fleksibel. Contoh: JSON log aplikasi, XML, data API, event tracking.

3. Data tidak terstruktur: tidak berbentuk tabel langsung. Contoh: teks, gambar, audio, video, PDF scan.

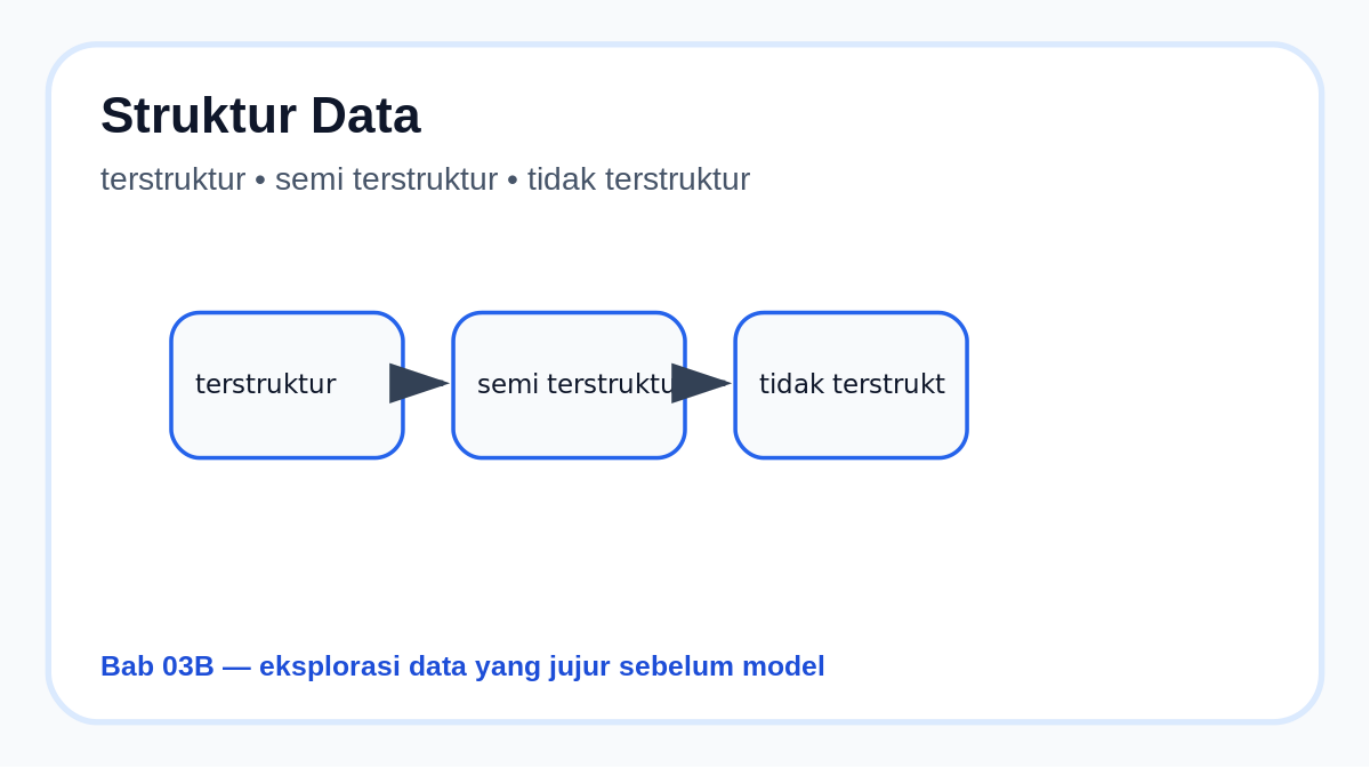
Data tabular paling sering muncul di bisnis. Namun AI modern banyak memakai data tidak terstruktur seperti teks dan gambar. Kuncinya: data tidak terstruktur sering harus diubah menjadi representasi terstruktur atau embedding sebelum masuk model.

Contoh JSON semi terstruktur:

```
{"user":"Ayu", "event":"click", "metadata":{"page":"promo", "device":"mobile"}}
```

Data ini punya key, tetapi `metadata` bisa berbeda antar event. Perlu flattening:

```
user | event | page  | device
Ayu  | click | promo | mobile
```



Tipe struktur data

Kesalahan umum: memaksa semua data menjadi tabel tanpa memikirkan makna. Teks panjang yang dipotong sembarangan bisa kehilangan konteks. Gambar yang hanya diubah menjadi rata-rata warna bisa kehilangan objek penting.

Tes cepat subbab 3

1. Beri contoh data semi terstruktur.

2. Mengapa teks disebut tidak terstruktur?

3. Apa risiko mengubah data tidak terstruktur menjadi fitur terlalu sederhana?

Subbab 4 — Tipe kolom: numerik, kategorikal, biner, ordinal, tanggal, teks

Inti subbab: grafik dan model yang cocok bergantung pada tipe kolom.

Tipe kolom umum:

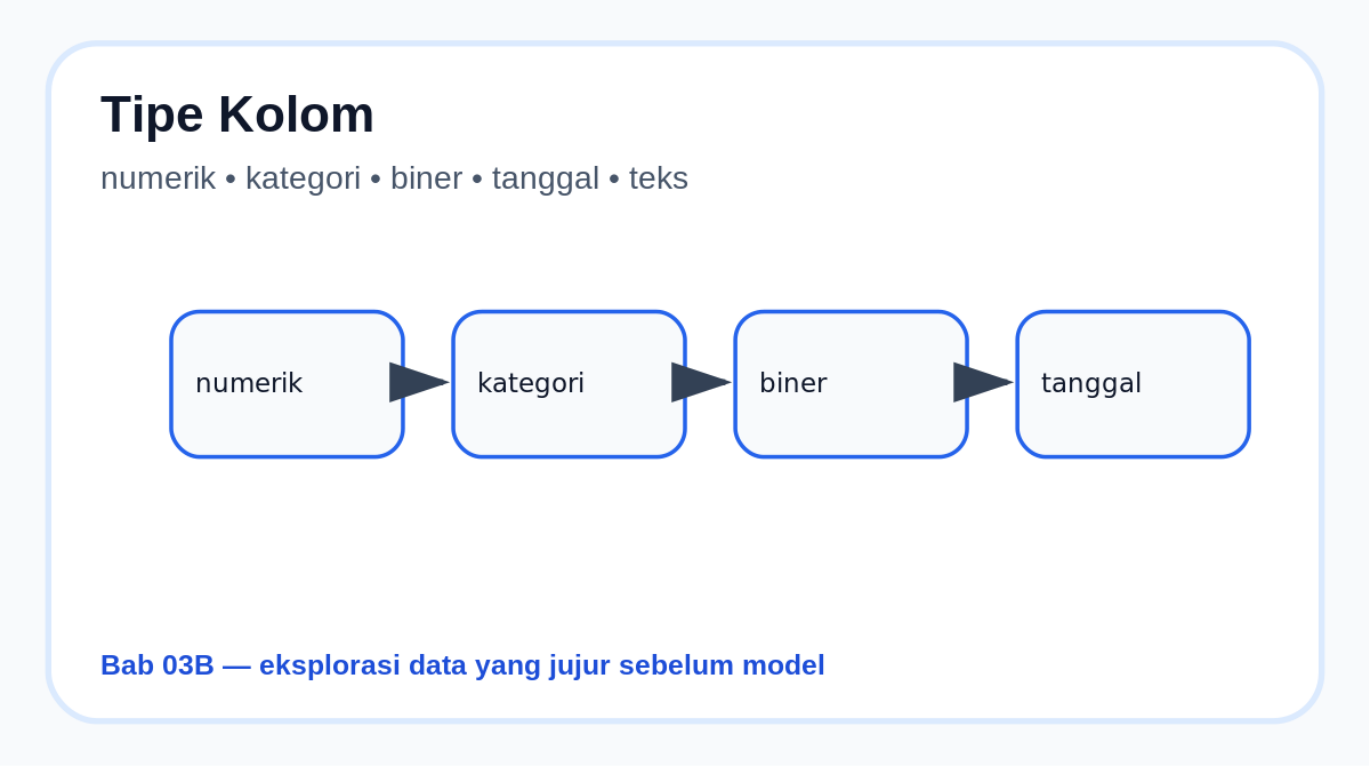
Tipe	Contoh	Operasi yang masuk akal
Numerik kontinu	tinggi, harga, durasi	mean, median, histogram, regresi
Numerik diskrit	jumlah anak, jumlah klik	bar/histogram, count, rate
Kategorikal nominal	kota, metode bayar	count, proporsi, bar plot
Ordinal	rendah/sedang/tinggi	urutan, median ordinal, bar ordered
Biner	ya/tidak, fraud/tidak	proporsi, confusion matrix
Tanggal/waktu	timestamp transaksi	line plot, resampling, lag
Teks	ulasan, komplain	panjang teks, kata kunci, embedding

Tipe data di file belum tentu sama dengan tipe makna. Kode pos terlihat angka, tetapi tidak boleh dirata-ratakan. Nomor pelanggan terlihat angka, tetapi itu ID. Rating 1–5 angka, tetapi maknanya ordinal.

Contoh kesalahan:

```
mean(kode_pos) = 55283
```

Angka ini tidak bermakna. Kode pos adalah kategori/ID geografis, bukan nilai kuantitatif.



Tipe kolom

Cara membaca tipe: tanyakan “apakah selisih antarangka bermakna?” Selisih harga 20.000 dan 10.000 bermakna. Selisih kode pos 55283 dan 55284 tidak bermakna.

Tes cepat subbab 4

1. Apakah nomor pelanggan termasuk numerik bermakna?

2. Apa beda nominal dan ordinal?

3. Mengapa tanggal perlu perlakuan khusus?

Subbab 5 — Statistik deskriptif: ringkasan sebelum visualisasi

Inti subbab: statistik deskriptif memberi ringkasan cepat, tetapi harus dibaca bersama distribusi.

Ukuran dasar:

```
mean = Σxᵢ / n
median = nilai tengah setelah diurutkan
range = max - min
variance = Σ(xᵢ - μ)² / n
standard deviation = sqrt(variance)
```

Cara membaca rumus mean: jumlahkan semua nilai x■, lalu bagi jumlah data n.

Cara membaca variance: setiap nilai dikurangi rata-rata, selisihnya dikuadratkan, dijumlahkan, lalu dirata-ratakan. Variance besar berarti data menyebar jauh dari rata-rata.

Contoh data belanja:

```
[20, 25, 25, 30, 100]
```

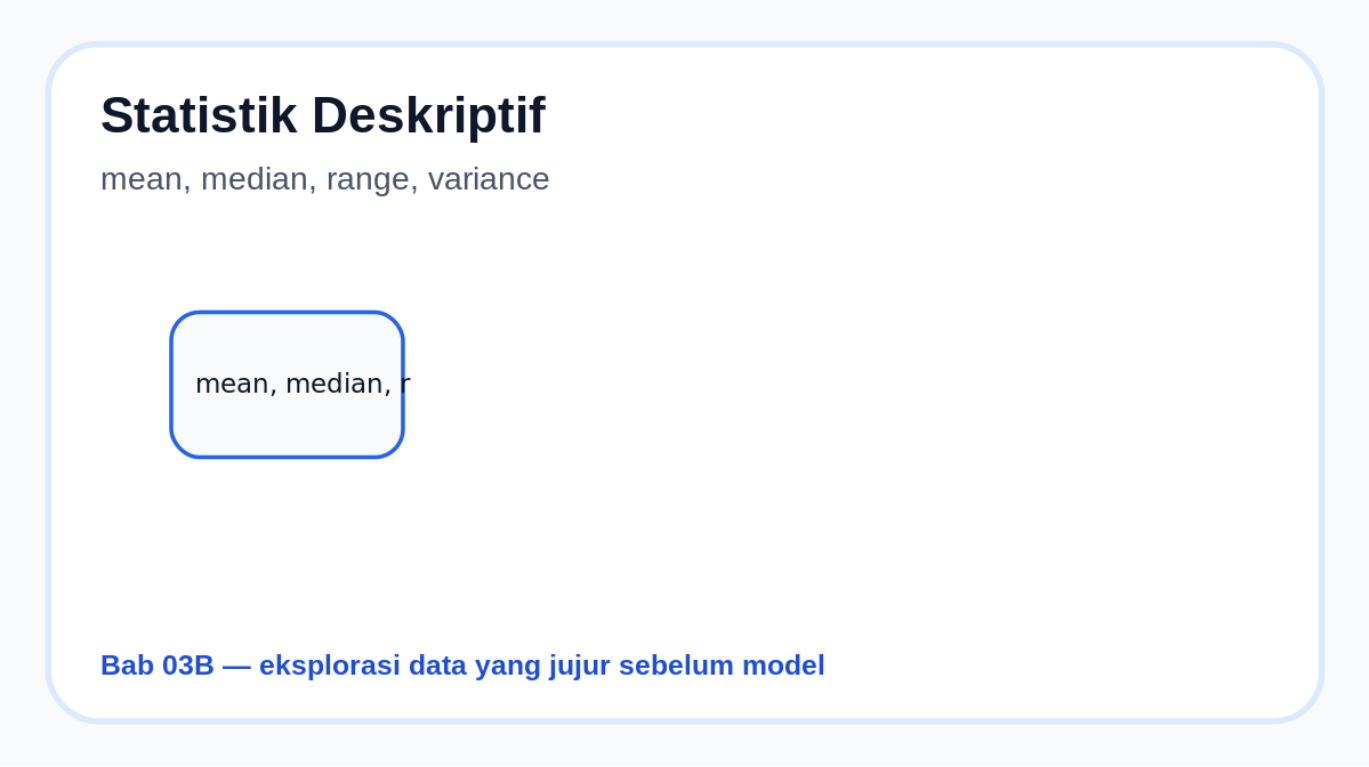
Mean:

```
(20+25+25+30+100)/5 = 200/5 = 40
```

Median:

```
nilai tengah = 25
```

Mean 40 tampak lebih tinggi karena ada outlier 100. Median 25 lebih mewakili transaksi umum.



Statistik deskriptif

Aturan praktis: jika mean jauh dari median, cek distribusi dan outlier. Jangan langsung memakai mean sebagai “nilai biasa”.

Tes cepat subbab 5

1. Hitung mean dari `[2,4,6,8]`.

2. Mengapa median tahan terhadap outlier?

3. Apa arti standar deviasi besar?

Subbab 6 — Data cleaning: missing value, duplikasi, tipe salah, rentang tidak logis

Inti subbab: cleaning adalah proses membuat data cukup jujur untuk dianalisis, bukan memoles data agar terlihat bagus.

Checklist cleaning:

missing value
baris duplikat
kolom duplikat
format tanggal salah
tipe numerik terbaca teks
nilai negatif yang mustahil
kategori typo: "Jakarta", "jakarta", "JKT"
unit campur: rupiah vs ribu rupiah

Missing rate

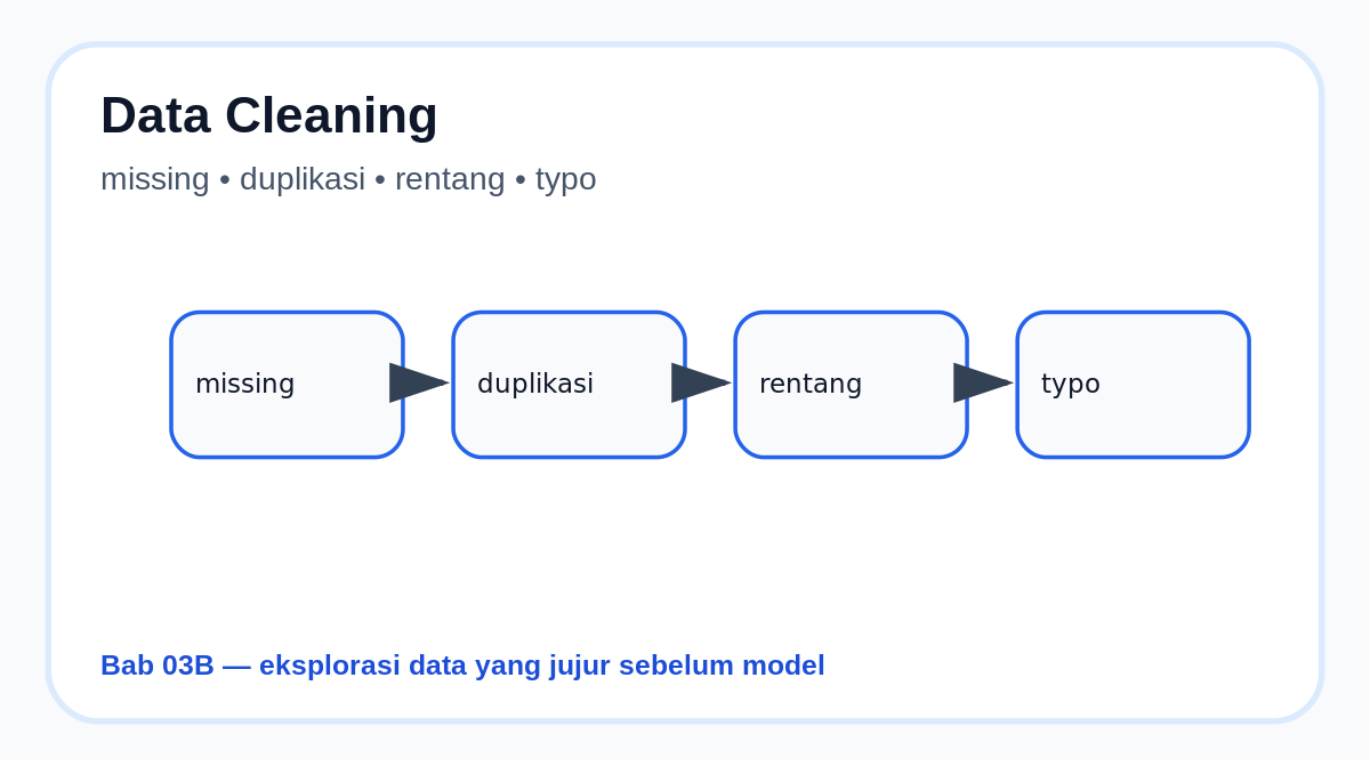
missing_rate = jumlah_nilai_kosong / jumlah_sel

Cara membaca rumus: pembilang adalah jumlah sel kosong. Penyebut adalah total sel yang dicek. Jika 12 dari 300 sel kosong:

missing_rate = 12/300 = 0,04 = 4%

Pilihan menangani missing:

Situasi	Strategi awal
missing sedikit dan acak	imputasi median/mode
missing berarti sesuatu	buat indikator missing
missing terlalu banyak	pertimbangkan buang kolom
missing pada target	biasanya baris tidak bisa dipakai training supervised



Data cleaning

Kesalahan umum: menghapus semua baris missing tanpa cek pola. Jika data missing lebih sering pada kelompok tertentu, penghapusan bisa membuat bias.

Tes cepat subbab 6

1. Hitung missing rate jika 5 dari 100 baris kosong pada satu kolom.

2. Mengapa kategori typo perlu disatukan?

3. Kapan kolom dengan missing banyak sebaiknya dibuang?

Subbab 7 — Preprocessing: encoding, scaling, parsing tanggal, dan feature extraction

Inti subbab: preprocessing mengubah data bersih menjadi bentuk yang bisa dipakai model.

Operasi preprocessing:

scaling numerik
encoding kategori
parsing tanggal
membuat fitur turunan
normalisasi teks
resize gambar
resampling time series

Encoding kategori:

• One-hot encoding cocok untuk kategori nominal: metode bayar = cash, QRIS, kartu.

• Ordinal encoding cocok jika urutan bermakna: rendah < sedang < tinggi.

• Target encoding bisa kuat tetapi rawan leakage jika tidak dilakukan dengan hati-hati.

Scaling:

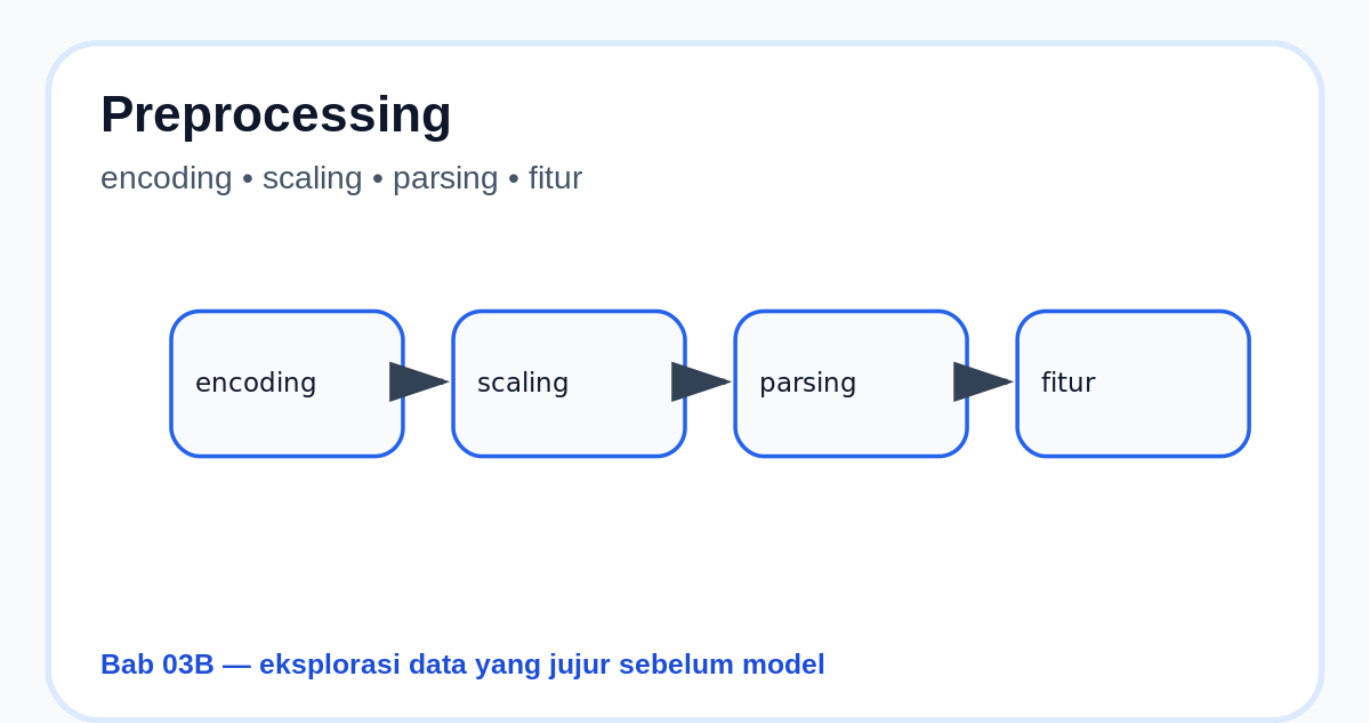
$$z = (x - \mu) / \sigma$$

Cara membaca rumus: nilai x dikurangi rata-rata μ, lalu dibagi standar deviasi σ. Hasilnya menunjukkan posisi relatif terhadap distribusi.

Contoh:

$x=80, \mu=50, \sigma=10$
$z=(80-50)/10=3$

Nilai 80 berada 3 standar deviasi di atas rata-rata.



Preprocessing data

Aturan penting: preprocessing yang belajar dari data, seperti mean, standar deviasi, median, atau vocabulary, harus dipelajari dari training set saja, lalu diterapkan ke validation/test.

Tes cepat subbab 7

1. Kapan one-hot encoding cocok?

2. Hitung z-score untuk $x=70$, $\mu=50$, $\sigma=10$.

3. Mengapa scaling tidak boleh fit pada test set?

Subbab 8 — Visualisasi data: grafik adalah argumen, bukan hiasan

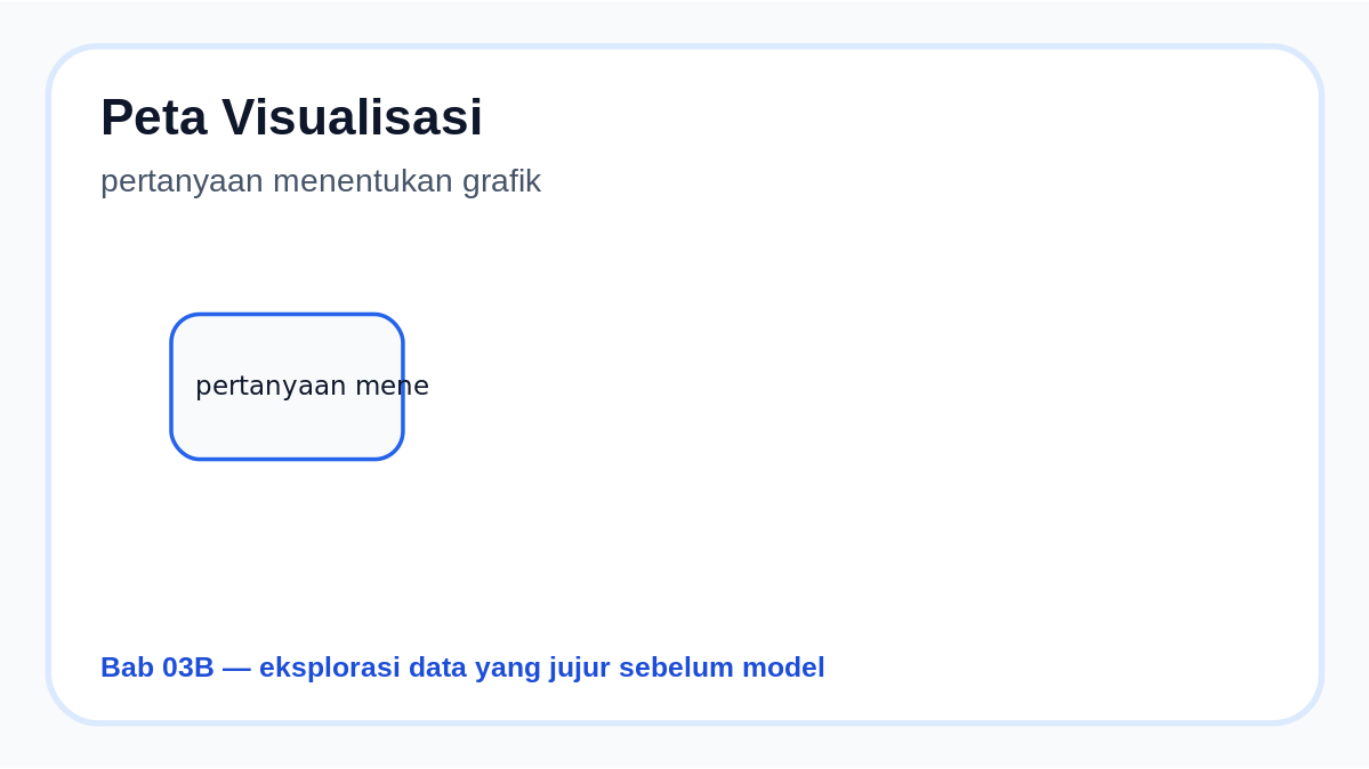
Inti subbab: grafik yang baik menjawab pertanyaan tertentu dengan tipe data yang tepat.

Sebelum membuat grafik, tulis pertanyaan:

> Saya ingin membandingkan apa?
> Saya ingin melihat distribusi apa?
> Saya ingin melihat hubungan antara variabel apa?
> Saya ingin melihat perubahan waktu apa?
> Saya ingin menemukan outlier apa?

Peta awal:

Tujuan	Grafik cocok
Membandingkan jumlah kategori	bar plot
Melihat proporsi sedikit kategori	pie chart atau bar proporsi
Melihat distribusi numerik	histogram, KDE, box plot
Melihat hubungan dua numerik	scatter plot
Melihat tren waktu	line plot
Melihat outlier	box plot, scatter, residual plot
Melihat korelasi banyak fitur	heatmap



Peta visualisasi

Kesalahan umum: memilih grafik karena terlihat keren, bukan karena cocok. Pie chart untuk 15 kategori sering sulit dibaca. Scatter plot untuk kategori nominal tidak memberi makna jika sumbu kategori tidak terurut.

Tes cepat subbab 8

1. Grafik apa untuk distribusi umur?

2. Grafik apa untuk hubungan durasi belajar dan nilai?

3. Mengapa grafik harus dimulai dari pertanyaan?

Subbab 9 — Bar plot: membandingkan kategori

Inti subbab: bar plot cocok untuk membandingkan jumlah atau nilai agregat antar kategori.

Contoh data metode pembayaran:

metode	jumlah transaksi
QRIS	120
Cash	80
Kartu	40

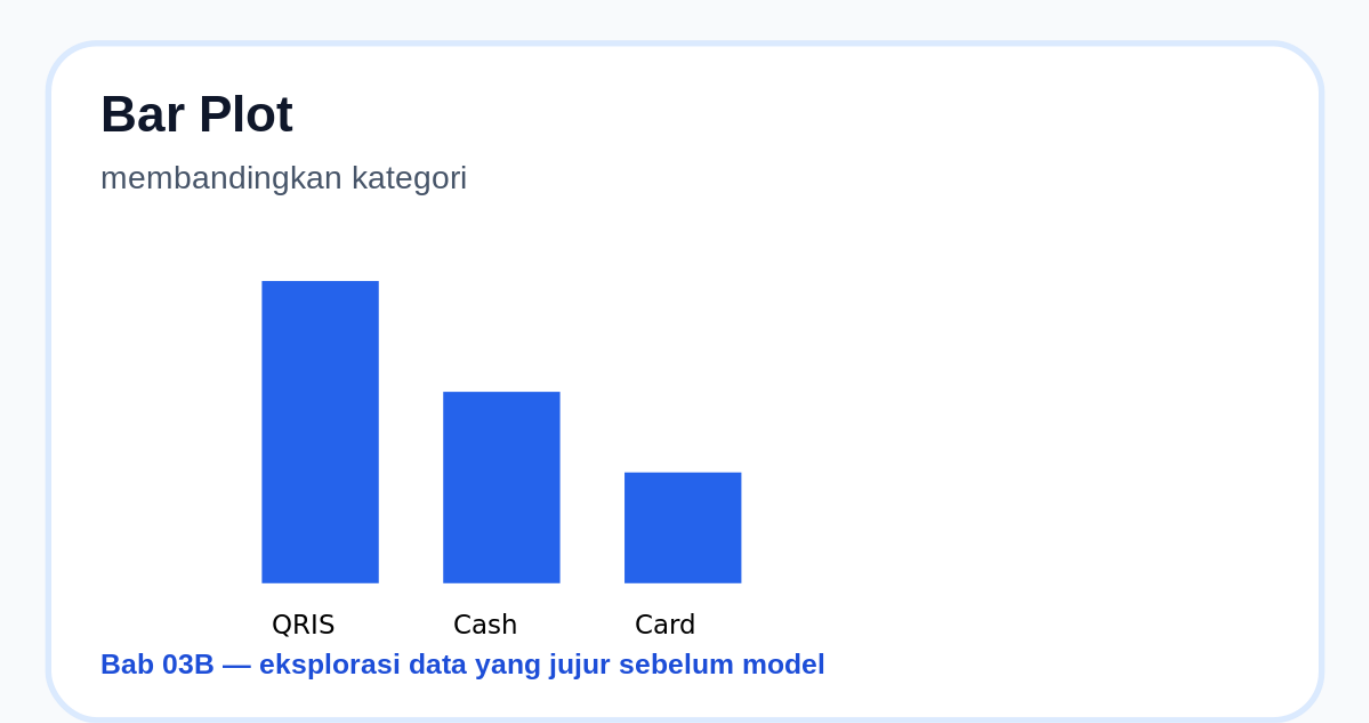
Bar plot cocok karena metode pembayaran adalah kategori nominal. Panjang bar menyatakan nilai. Urutan bisa berdasarkan jumlah terbesar agar mudah dibaca.

Rumus proporsi kategori

```
proporsi_kategori = count_kategori / total_count
```

Cara membaca rumus: jumlah item pada satu kategori dibagi total item. Untuk QRIS:

```
120 / (120+80+40) = 120/240 = 0,5 = 50%
```



Bar plot

Kesalahan umum: memakai line plot untuk kategori nominal. Garis memberi kesan ada urutan kontinu antara QRIS, Cash, dan Kartu, padahal tidak.

Tes cepat subbab 9

1. Mengapa bar plot cocok untuk kategori?

2. Hitung proporsi Cash dari data di atas.

3. Kapan bar plot lebih baik daripada pie chart?

Subbab 10 — Pie chart: proporsi sederhana, bukan semua komposisi

Inti subbab: pie chart cocok jika kategori sedikit dan tujuannya menunjukkan bagian dari keseluruhan.

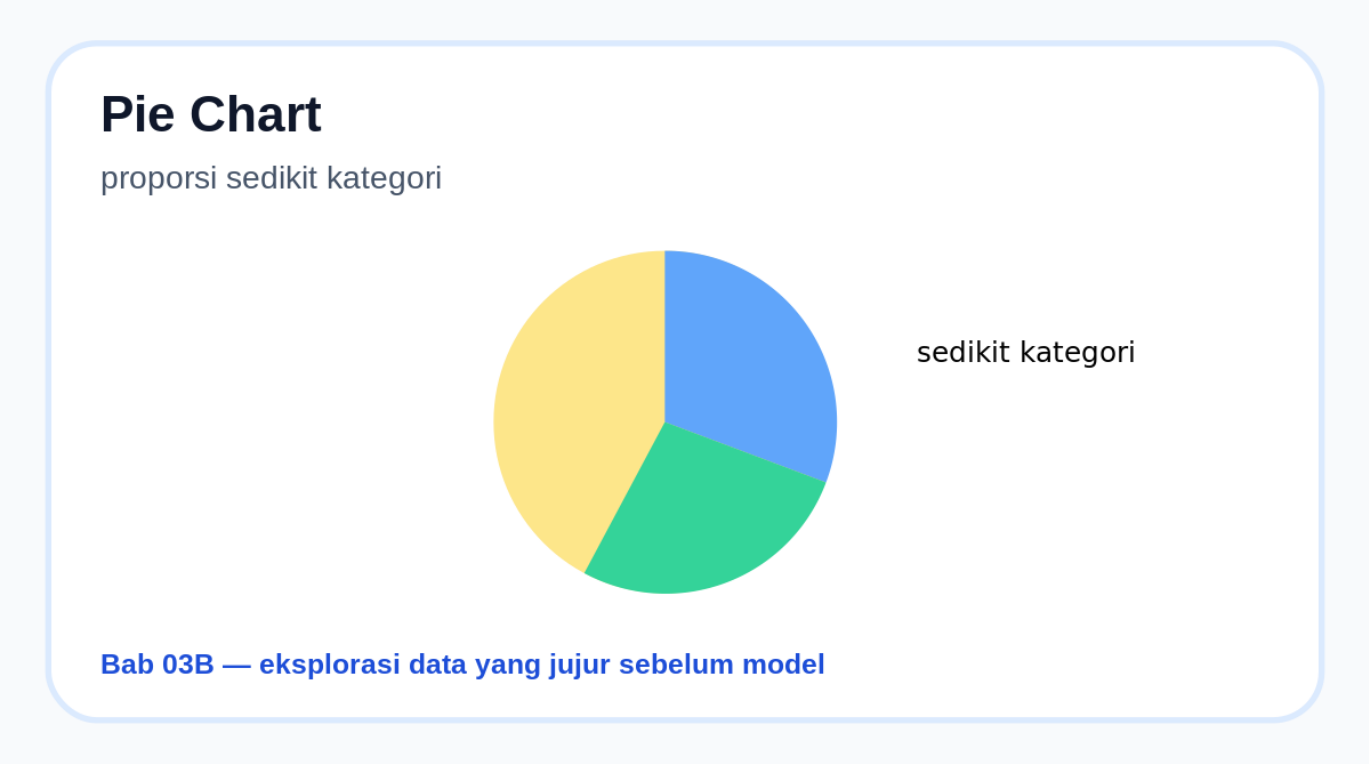
Pie chart bisa berguna untuk 2–4 kategori yang jelas, misalnya komposisi perangkat pengguna:

> Mobile 70%, Desktop 25%, Tablet 5%

Namun pie chart buruk jika kategori banyak, nilai mirip, atau pembaca perlu membandingkan perbedaan kecil. Mata manusia lebih mudah membandingkan panjang bar daripada sudut irisan.

Kapan pie chart cocok?

> Kategori sedikit
> Jumlah total bermakna 100%
> Perbedaan proporsi cukup jelas
> Tidak perlu membaca angka sangat presisi



Pie chart

Contoh kesalahan: membuat pie chart untuk 12 provinsi dengan proporsi mirip. Grafik akan penuh warna, legenda panjang, dan sulit dibaca. Gunakan bar plot terurut.

Tes cepat subbab 10

1. Kapan pie chart masih masuk akal?

2. Mengapa pie chart buruk untuk banyak kategori?

3. Untuk 10 kategori produk, pilih pie atau bar? Mengapa?

Subbab 11 — Histogram dan KDE: membaca distribusi numerik

Inti subbab: histogram menunjukkan bentuk distribusi nilai numerik.

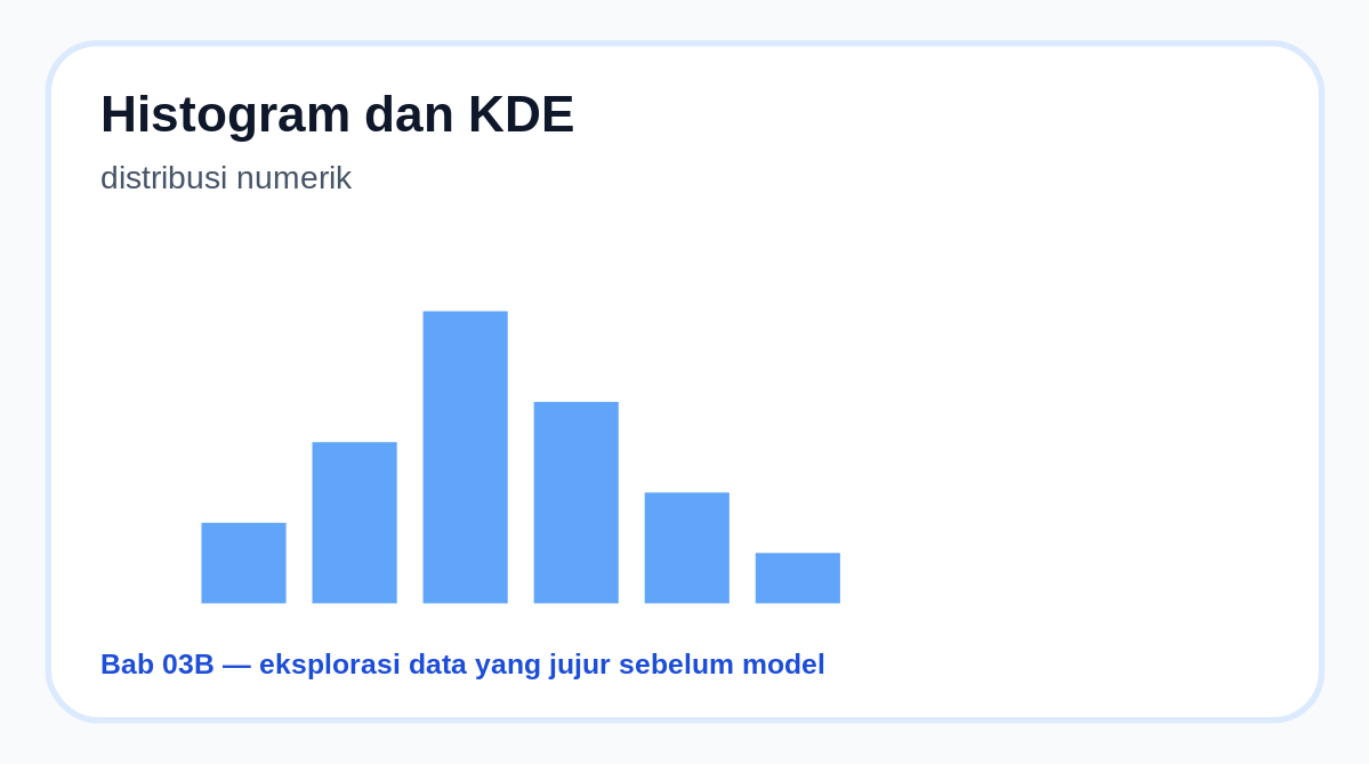
Histogram membagi rentang nilai menjadi bins, lalu menghitung jumlah data di setiap bin. Cocok untuk melihat umur, harga, durasi, penjualan, atau error model.

Contoh data durasi belajar:

```
[10, 12, 15, 20, 22, 60]
```

Histogram akan menunjukkan sebagian besar durasi di bawah 25, dengan satu nilai 60 yang jauh.

KDE (*kernel density estimate*) adalah kurva halus perkiraan distribusi. KDE membantu melihat bentuk, tetapi bisa menipu jika data sedikit atau bandwidth tidak cocok.



Histogram KDE

Cara membaca histogram: sumbu-X adalah rentang nilai. Sumbu-Y adalah jumlah/frekuensi. Bar tinggi berarti banyak data berada di rentang itu.

Kesalahan umum: mengubah jumlah bins sampai grafik sesuai cerita yang diinginkan. Bins terlalu sedikit menyembunyikan pola; bins terlalu banyak membuat grafik berisik.

Tes cepat subbab 11

1. Untuk data numerik kontinu, grafik apa yang cocok untuk distribusi?

2. Apa arti bar tinggi pada histogram?

3. Mengapa jumlah bins penting?

Subbab 12 — Scatter plot: hubungan dua variabel numerik

Inti subbab: scatter plot cocok untuk melihat hubungan dua fitur numerik dan menemukan cluster atau outlier.

Scatter plot memakai dua sumbu numerik. Setiap titik adalah observasi. Cocok untuk:

belanja vs kunjungan
luas rumah vs harga
lama belajar vs nilai
umur mesin vs biaya perawatan

Pola yang mungkin terlihat:

• naik: korelasi positif,

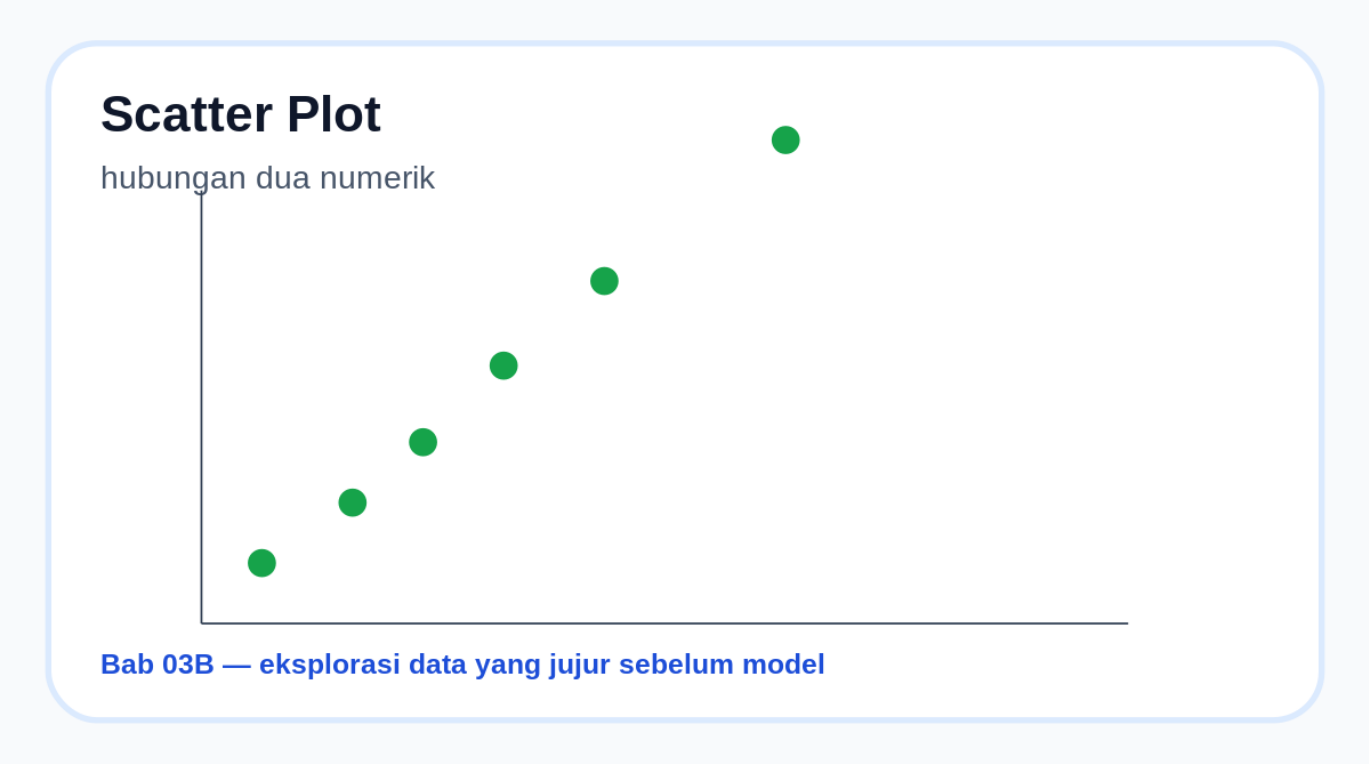
• turun: korelasi negatif,

• awan acak: hubungan lemah,

• kurva: hubungan non-linear,

• kelompok titik: cluster,

• titik jauh: outlier.



Scatter plot

Rumus korelasi Pearson

$$r = \Sigma(x_i-\bar{x})(y_i-\bar{y}) / \text{sqrt}(\Sigma(x_i-\bar{x})^2\ \Sigma(y_i-\bar{y})^2)$$

Cara membaca rumus: pembilang mengukur apakah x dan y bergerak bersama. Penyebut menormalkan agar nilai r berada antara -1 dan 1.

Catatan penting: korelasi bukan kausalitas. Es krim dan jumlah orang berenang bisa naik bersama karena cuaca panas, bukan karena es krim menyebabkan orang berenang.

Tes cepat subbab 12

1. Kapan scatter plot cocok?

2. Apa arti pola naik pada scatter plot?

3. Mengapa korelasi tidak otomatis berarti sebab-akibat?

Subbab 13 — Box plot dan violin plot: median, kuartil, dan outlier

Inti subbab: box plot cocok untuk membandingkan distribusi numerik antar kategori dan melihat outlier.

Box plot menunjukkan:

median
Q1 dan Q3
IQR = Q3 - Q1
whisker
kandidat outlier

Batas outlier umum:

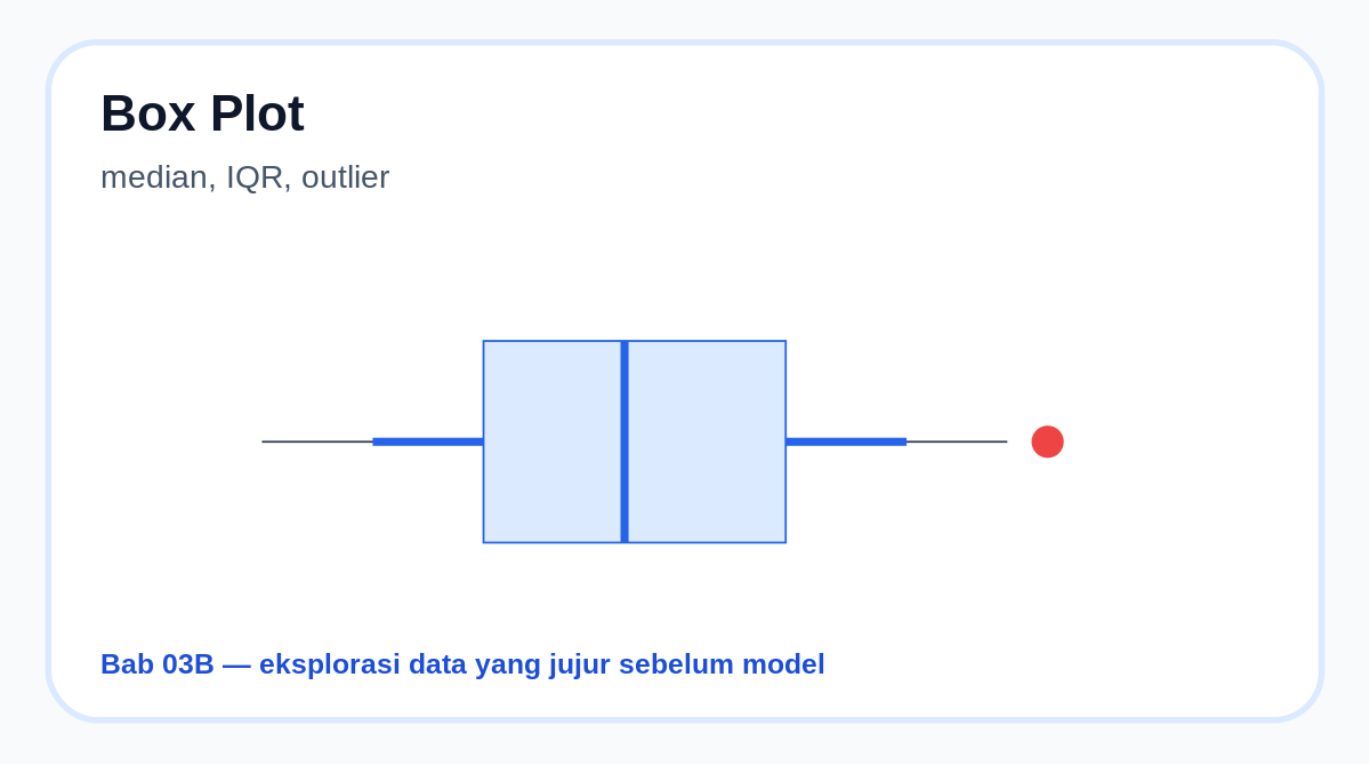
```
batas_bawah = Q1 - 1,5×IQR
batas_atas  = Q3 + 1,5×IQR
```

Cara membaca rumus: IQR adalah lebar kotak tengah. Nilai jauh di luar kotak dianggap kandidat outlier.

Contoh:

Q1=20, Q3=40
IQR=20
batas_atas=40+1,5×20=70

Nilai 90 menjadi kandidat outlier.



Box plot

Violin plot menambahkan bentuk distribusi. Ia berguna jika pembaca sudah cukup nyaman, tetapi untuk pemula box plot biasanya lebih mudah.

Tes cepat subbab 13

1. Apa itu IQR?

2. Hitung batas atas jika `Q1=10`, `Q3=30`.

3. Mengapa box plot cocok untuk outlier?

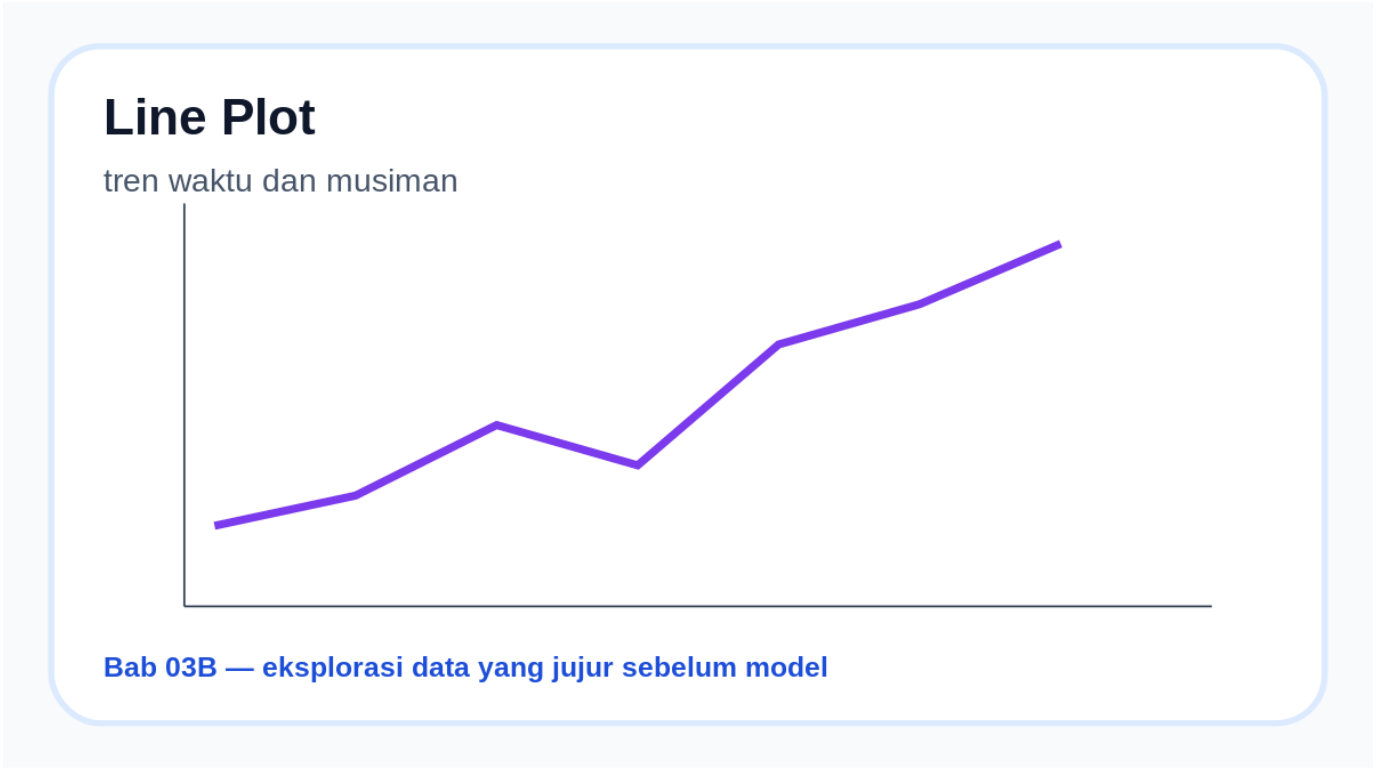
Subbab 14 — Line plot dan time series: data yang punya urutan waktu

Inti subbab: jika data punya waktu, urutan tidak boleh diacak sembarangan.

Line plot cocok untuk tren penjualan harian, suhu per jam, traffic website per menit, atau metrik model per epoch. Sumbu-X biasanya waktu. Sumbu-Y adalah nilai.

Pertanyaan penting:

Apakah ada tren naik/turun?
Apakah ada musiman?
Apakah ada lonjakan aneh?
Apakah ada perubahan setelah event tertentu?



Line plot time series

Rolling mean

rolling_mean_t = rata-rata nilai beberapa periode terakhir

Cara membaca rumus: untuk setiap titik waktu, hitung rata-rata dari jendela waktu tertentu. Rolling mean membantu mengurangi noise.

Split time series: jangan random split. Data masa depan tidak boleh masuk training masa lalu. Gunakan urutan waktu:

train: bulan 1-8
validation: bulan 9-10
test: bulan 11-12

Tes cepat subbab 14

1. Mengapa time series tidak boleh diacak sembarangan?

2. Apa fungsi rolling mean?

3. Grafik apa yang cocok untuk penjualan harian?

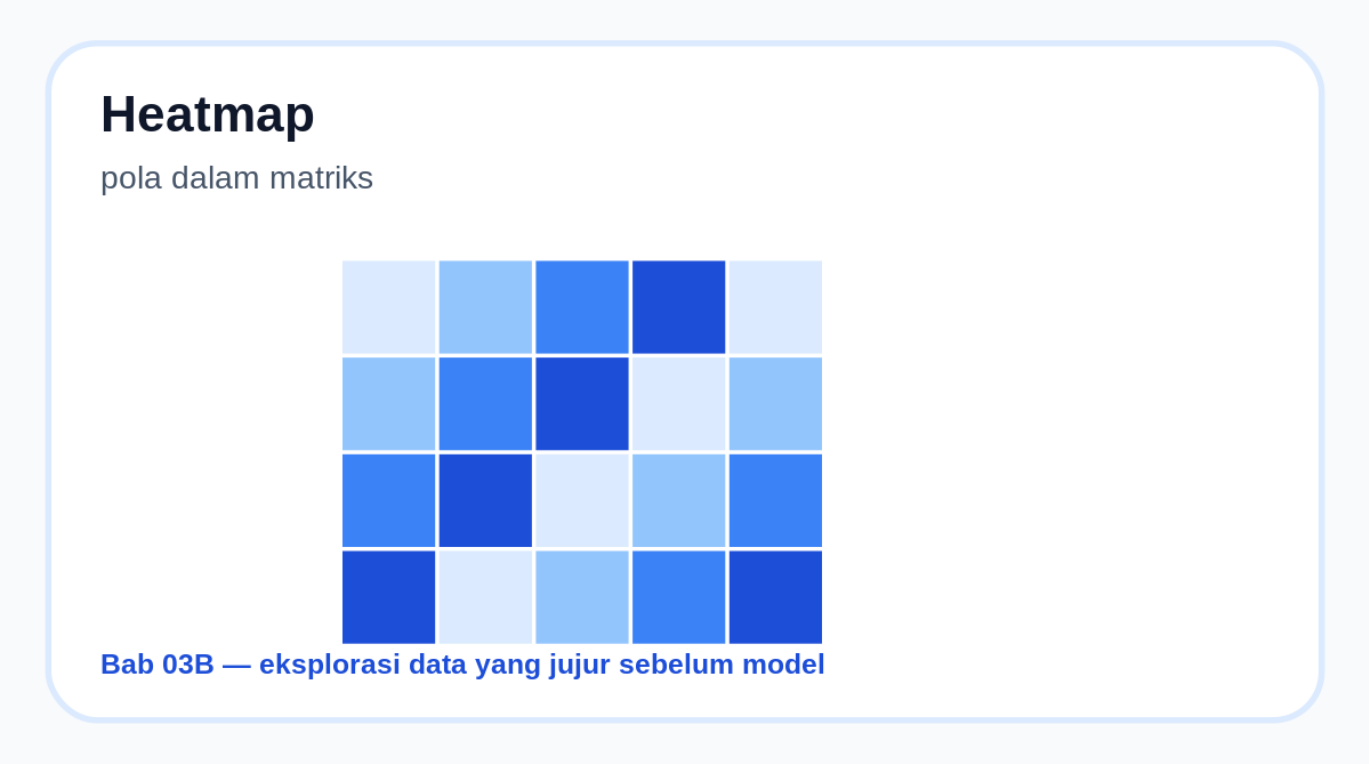
Subbab 15 — Heatmap: korelasi, missing value, dan matriks kebingungan

Inti subbab: heatmap cocok untuk melihat pola dalam matriks.

Heatmap memakai warna untuk menunjukkan nilai. Cocok untuk:

matriks korelasi fitur numerik
pola missing value
confusion matrix
kemiripan antar item
aktivitas per jam dan hari

Matriks korelasi membantu menemukan fitur yang bergerak bersama. Jika dua fitur sangat berkorelasi, mungkin salah satunya redundant.



Heatmap

Kesalahan umum: memakai palet warna yang menipu. Warna harus konsisten, punya legenda, dan tidak membuat perbedaan kecil terlihat dramatis.

Cara membaca heatmap: cari blok warna kuat, pola diagonal, baris/kolom yang berbeda, dan nilai ekstrem. Jangan hanya terpaku pada warna tercerah.

Tes cepat subbab 15

1. Sebutkan dua penggunaan heatmap.

2. Mengapa legenda warna penting?

3. Apa risiko palet warna yang buruk?

Subbab 16 — Visualisasi outlier: tidak semua titik jauh harus dibuang

Inti subbab: outlier harus dipahami, bukan otomatis dihapus.

Outlier bisa muncul karena:

kesalahan input
sensor rusak
kejadian langka tetapi valid
fraud
pelanggan VIP
perubahan sistem

Visualisasi outlier:

• box plot untuk satu fitur,

• scatter plot untuk dua fitur,

• residual plot untuk regresi,

• time series plot untuk lonjakan waktu,

• z-score plot untuk nilai ekstrem relatif distribusi.

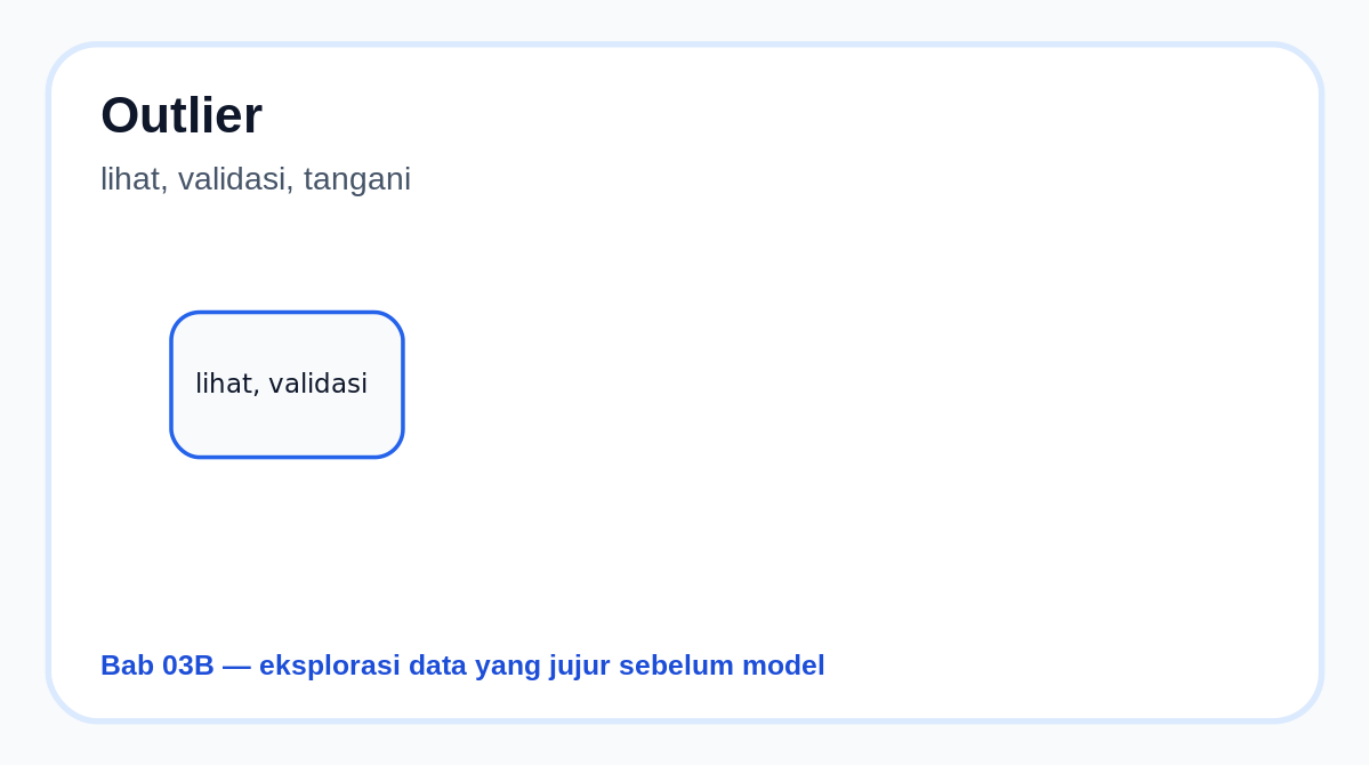
Z-score

$$z = (x - \mu) / \sigma$$

Cara membaca rumus: nilai dikurangi rata-rata lalu dibagi standar deviasi. Jika $|z|$ besar, nilai jauh dari pusat distribusi.

Contoh:

x=170, μ=100, σ=20
z=(170-100)/20=3,5



Visualisasi outlier

Strategi mengatasi outlier: validasi sumber, cap/winsorize, transformasi log, robust scaler, model robust, atau pisahkan sebagai kasus khusus.

Tes cepat subbab 16

1. Mengapa outlier tidak otomatis salah?

2. Hitung z-score untuk $x=130$, $\mu=100$, $\sigma=10$.

3. Grafik apa yang cocok melihat outlier dua fitur?

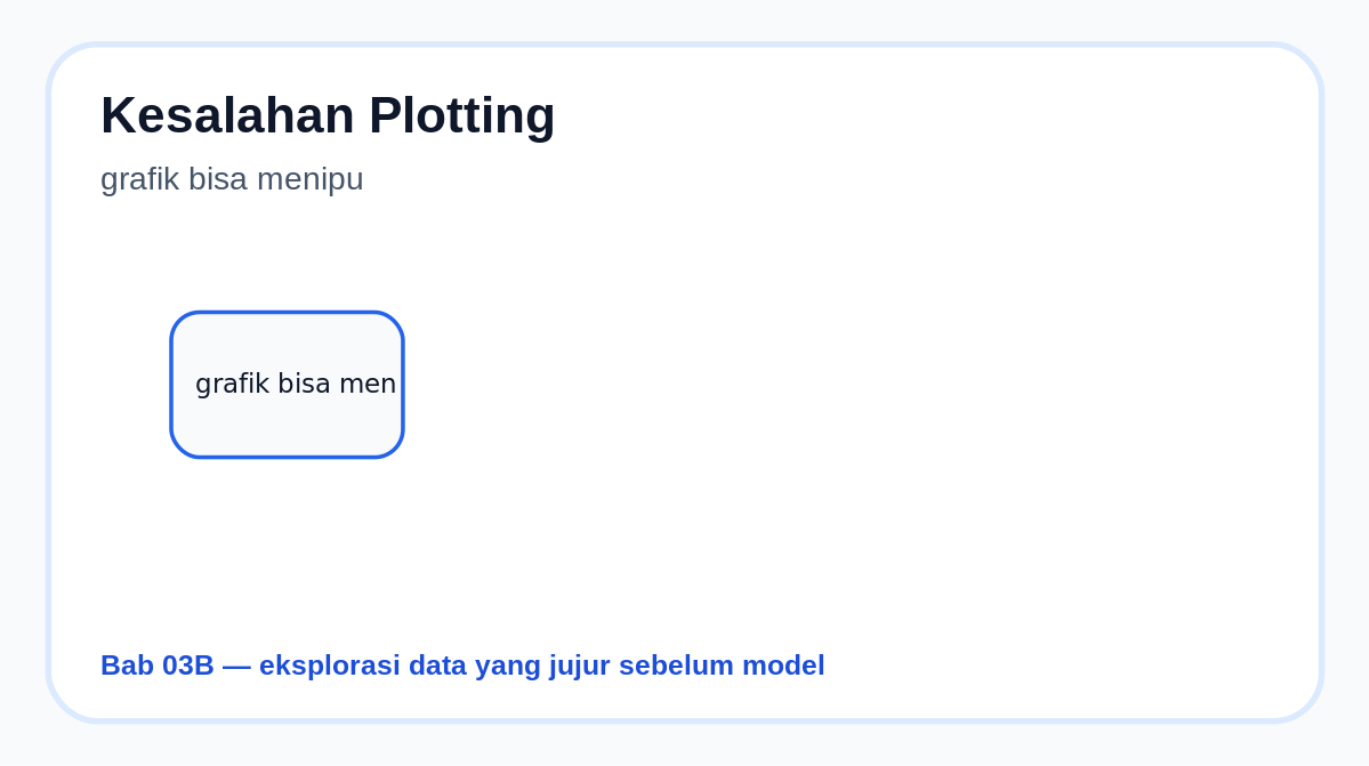
Subbab 17 — Kesalahan plotting yang sering menipu pembaca

Inti subbab: visualisasi bisa membantu, tetapi juga bisa menyesatkan.

Kesalahan plotting umum:

sumbu-Y dipotong tanpa alasan
pie chart terlalu banyak kategori
line chart untuk kategori nominal
warna terlalu banyak tanpa makna
scatter plot tanpa transparansi saat titik bertumpuk
rata-rata ditampilkan tanpa sebaran
urutan kategori tidak logis
skala log dipakai tanpa penjelasan

Contoh: bar plot penjualan dengan sumbu-Y dimulai dari 95 padahal nilai 96 dan 100. Perbedaan kecil terlihat dramatis. Jika sumbu dipotong, tulis jelas.



Kesalahan plotting

Aturan jujur: grafik harus membuat pembaca lebih paham, bukan lebih mudah dipengaruhi. Sertakan judul, label sumbu, satuan, legenda, dan sumber data.

Tes cepat subbab 17

1. Mengapa sumbu-Y terpotong bisa menipu?

2. Apa masalah line chart untuk kategori nominal?

3. Mengapa rata-rata sebaiknya ditemani sebaran?

Subbab 18 — Matplotlib, seaborn, dan pilihan alat visualisasi

Inti subbab: matplotlib memberi kontrol dasar; seaborn memudahkan grafik statistik; pandas membantu eksplorasi tabel.

Contoh matplotlib:

```
import matplotlib.pyplot as plt
plt.hist(df["belanja"])
plt.xlabel("Belanja")
plt.ylabel("Jumlah")
plt.show()
```

Contoh seaborn:

```
import seaborn as sns
sns.boxplot(data=df, x="metode_bayar", y="belanja")
```

Matplotlib cocok ketika butuh kontrol detail. Seaborn cocok untuk EDA statistik cepat seperti boxplot, pairplot, heatmap, dan categorical plot. Pandas cocok untuk ringkasan data tabular.

Matplotlib seaborn

Catatan praktik: di beberapa environment, versi numpy/matplotlib bisa konflik. Karena itu lab bab ini menyediakan fallback SVG standard library. Tetapi pembaca tetap diperkenalkan ke matplotlib/seaborn karena keduanya sangat umum di dunia data.

Tes cepat subbab 18

1. Apa kelebihan matplotlib?

2. Apa kelebihan seaborn?

3. Mengapa fallback plot berguna untuk pembelajaran?

Subbab 19 — Data sintetis: kapan dipakai dan kapan berbahaya

Inti subbab: data sintetis adalah data buatan yang meniru pola tertentu, berguna untuk belajar dan testing, tetapi tidak otomatis mewakili dunia nyata.

Alasan memakai data sintetis:

melatih konsep tanpa data sensitif
menguji pipeline
membuat kasus langka
menyeimbangkan kelas
simulasi sebelum data asli tersedia

Metode sederhana:

• sampling dari distribusi normal/uniform,

• aturan bisnis buatan,

• bootstrap/resampling,

• augmentasi gambar/teks,

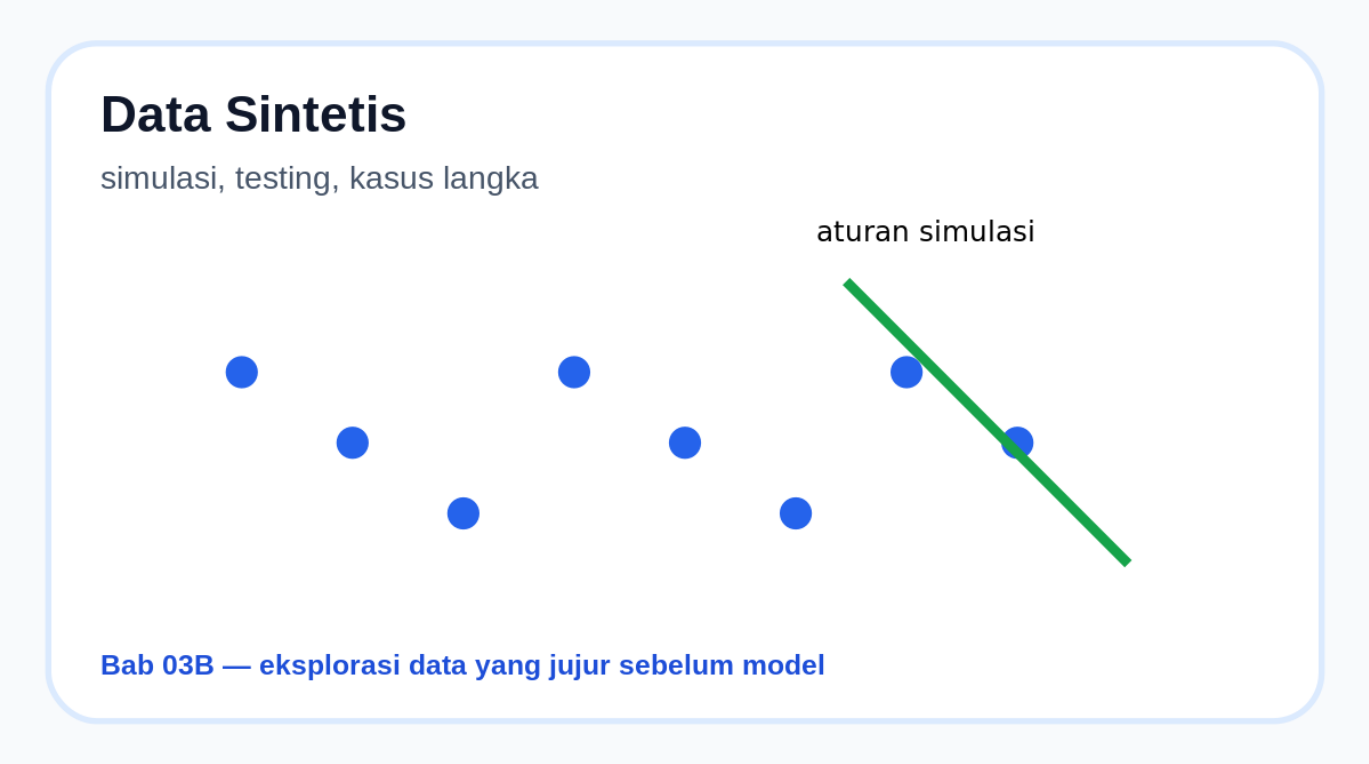
• SMOTE untuk tabular imbalance,

• generative model untuk data kompleks.

Contoh normal distribution:

$$x \sim \text{Normal}(\mu, \sigma^2)$$

Cara membaca rumus: x diambil dari distribusi normal dengan rata-rata μ dan variance σ^2.



Data sintetis

Bahaya: data sintetis bisa terlalu rapi. Model yang bagus di data sintetis bisa gagal di data asli karena noise, bias, missing value, dan perilaku manusia tidak ikut tersimulasikan.

Tes cepat subbab 19

1. Mengapa data sintetis berguna untuk belajar?
2. Sebutkan satu risiko data sintetis.
3. Apa beda data sintetis untuk testing pipeline dan untuk training final?

Subbab 20 — Split train, validation, test: evaluasi yang jujur

Inti subbab: split data menentukan apakah evaluasi model jujur atau bocor.

Split umum:

```
train      → melatih model
validation → memilih hyperparameter/model
test       → evaluasi akhir yang jarang disentuh
```

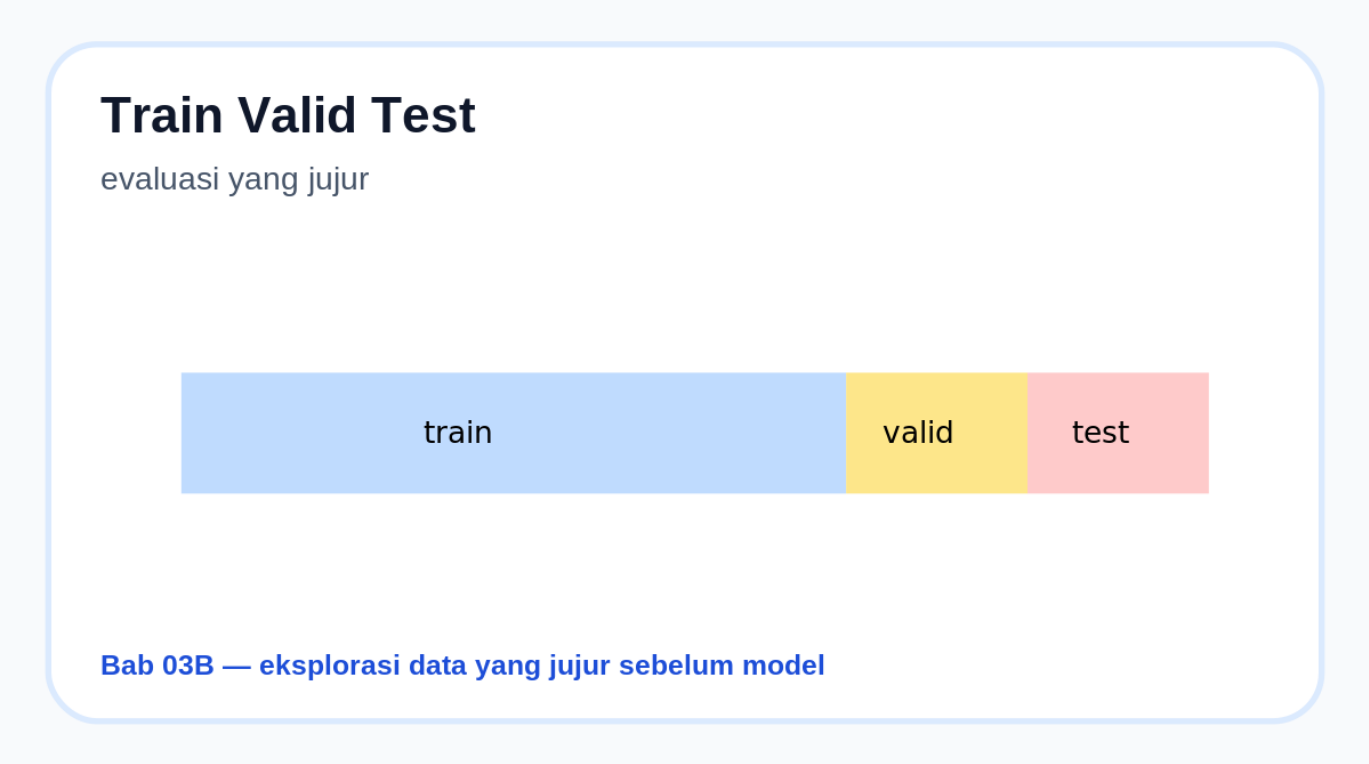
Contoh proporsi:

70% train, 15% validation, 15% test

Jika ada 1.000 baris:

```
train = 700
validation = 150
test = 150
```

Cara membaca split: training adalah tempat belajar. Validation adalah tempat memilih. Test adalah ujian akhir.



Train validation test

Jenis split:

• random split untuk data iid,

• stratified split untuk klasifikasi imbalance,

• group split jika banyak baris dari orang yang sama,

• time-based split untuk data waktu.

Tes cepat subbab 20

1. Apa fungsi validation set?

2. Hitung 80/10/10 untuk 500 data.

3. Mengapa time series memakai split waktu?

Subbab 21 — Data leakage: musuh diam-diam model AI

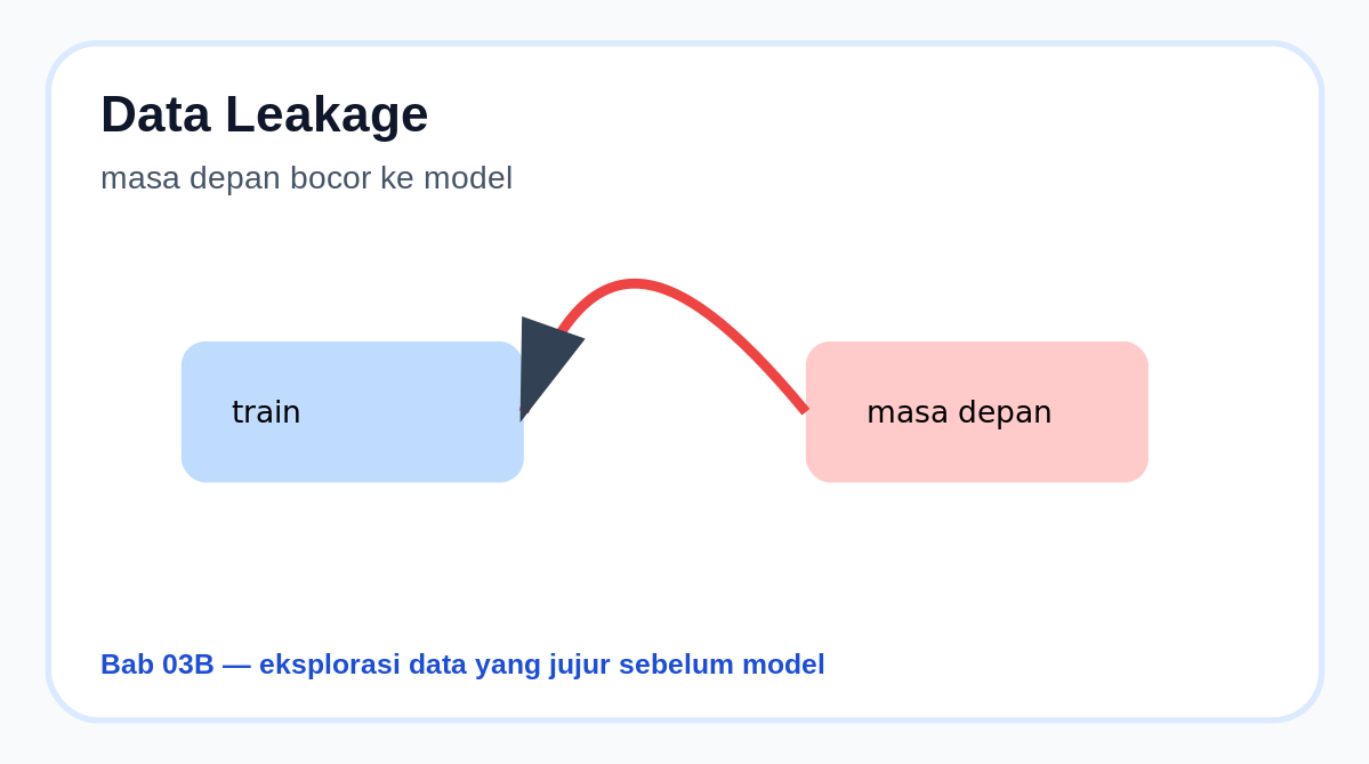
Inti subbab: leakage terjadi ketika informasi yang seharusnya tidak tersedia saat prediksi masuk ke training/evaluasi.

Contoh leakage:

```
kolom hasil masa depan masuk fitur
preprocessing fit pada seluruh data sebelum split
baris pelanggan yang sama muncul di train dan test
fitur dibuat memakai target
oversampling dilakukan sebelum split
```

Leakage membuat model terlihat sangat bagus tetapi gagal di dunia nyata.

Contoh: memprediksi apakah pelanggan akan churn, tetapi fitur berisi `tanggal_penutupan_akun`. Kolom ini baru diketahui setelah churn, sehingga bocor.



Data leakage

Checklist anti-leakage:

Kapan fitur tersedia?
Apakah fitur memakai masa depan?
Apakah preprocessing fit hanya di train?
Apakah entitas sama bocor ke test?
Apakah target tersirat dalam fitur?

Tes cepat subbab 21

1. Mengapa leakage berbahaya?

2. Beri contoh fitur masa depan.

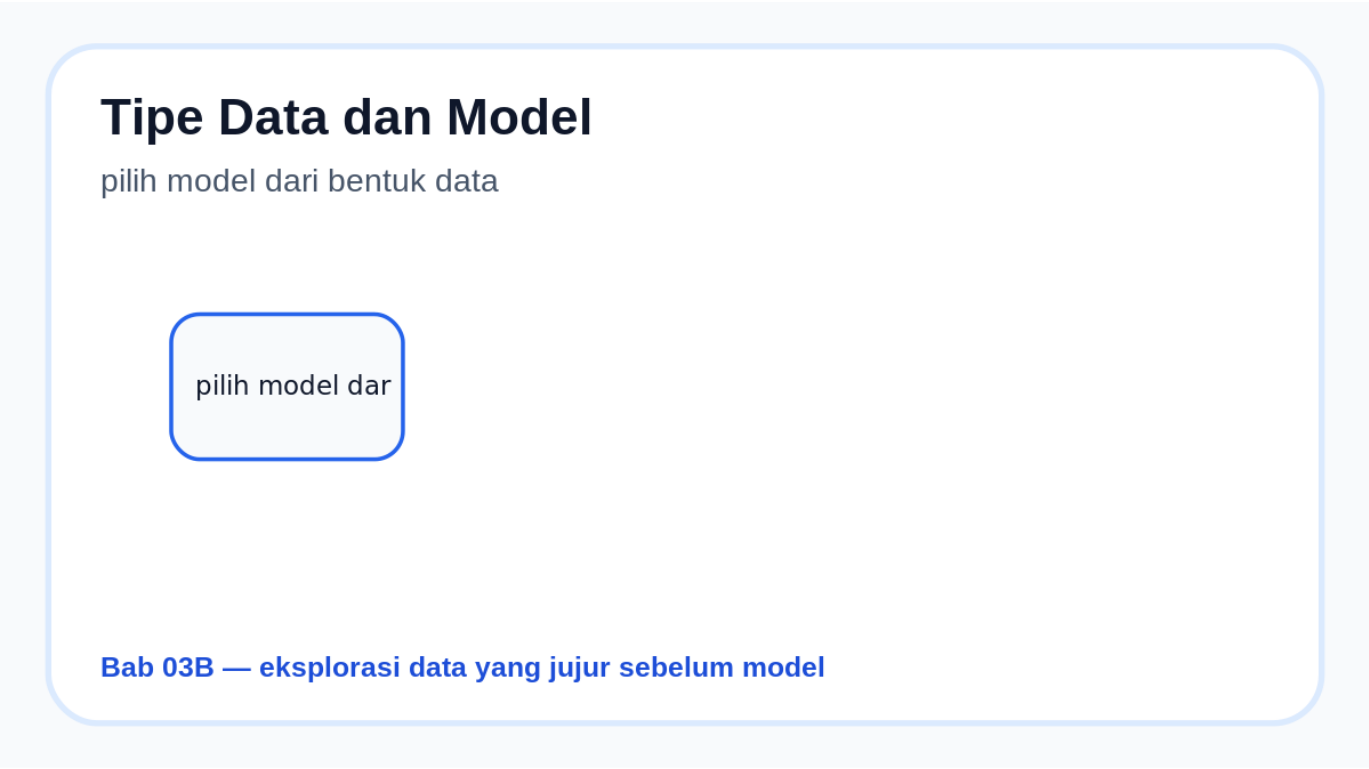
3. Mengapa scaling harus fit di train saja?

Subbab 22 — Tipe data dan model yang cocok

Inti subbab: model dipilih setelah memahami tipe data, ukuran data, kebutuhan interpretasi, dan metrik.

Peta praktis:

Data	Masalah	Model awal yang cocok
Tabular numerik/kategori	klasifikasi/regresi	baseline, linear/logistic, tree, random forest, boosting
Teks pendek	klasifikasi/topik	TF-IDF + linear, embedding + classifier
Gambar	klasifikasi/deteksi	CNN/transfer learning
Time series	forecasting	naive seasonal, ARIMA/Prophet, tree with lag, neural sequence
Graph	relasi entitas	graph features, GNN jika perlu
Data tanpa label	segmentasi/anomali	clustering, PCA, anomaly detection



Tipe data dan model

Aturan emas: mulai dari baseline. Jangan langsung model paling rumit. Untuk tabular bisnis, model tree/boosting sering kuat. Untuk teks/gambar, embedding dan deep learning sering relevan.

Tes cepat subbab 22

1. Model awal apa untuk tabular kecil?

2. Mengapa teks butuh representasi khusus?

3. Kapan clustering dipakai?

Subbab 23 — Data tabular secara mendalam: skema, key, join, agregasi

Inti subbab: data tabular terlihat sederhana, tetapi banyak risiko tersembunyi.

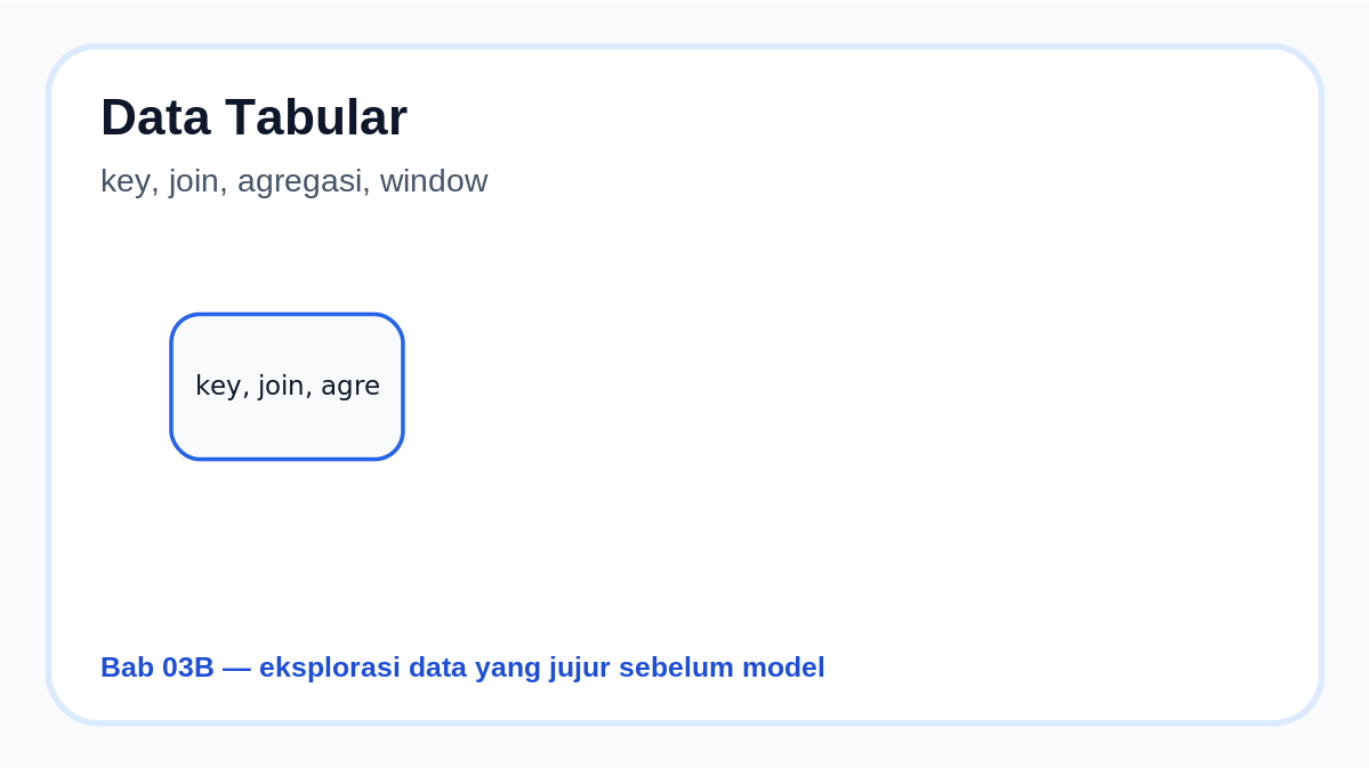
Konsep penting:

```
primary key: identitas unik baris
foreign key: penghubung ke tabel lain
join: menggabungkan tabel
aggregation: merangkum banyak baris menjadi satu entitas
window: periode waktu untuk menghitung fitur
```

Contoh: tabel transaksi perlu diagregasi menjadi fitur pelanggan:

```
jumlah_transaksi_30_hari
rata_belanja_30_hari
hari_sejak_transaksi_terakhir
```

Risiko: jika target adalah churn bulan depan, fitur harus dihitung hanya dari data sebelum tanggal prediksi.



Data tabular

Kesalahan umum: join menggandakan baris. Jika satu pelanggan punya banyak transaksi dan banyak tiket komplain, join langsung bisa membuat kombinasi berlipat. Selalu cek jumlah baris sebelum dan sesudah join.

Tes cepat subbab 23

1. Apa itu primary key?

2. Mengapa join bisa menggandakan baris?

3. Apa itu fitur agregasi 30 hari?

Subbab 24 — Data tidak terstruktur: teks, gambar, audio, video

Inti subbab: data tidak terstruktur perlu diubah menjadi fitur atau embedding sebelum dianalisis dengan banyak model.

Teks:

cleaning → tokenisasi → TF-IDF/embedding → model

Gambar:

resize → normalisasi pixel → augmentasi → CNN/embedding

Audio:

sampling → spectrogram/MFCC → model

Video:

frame sampling → fitur visual + temporal → model

Data tidak terstruktur

Kesalahan umum: memperlakukan teks seperti kategori biasa. Kalimat punya urutan dan konteks. Begitu juga gambar punya struktur spasial.

Tes cepat subbab 24

1. Mengapa teks perlu tokenisasi atau embedding?

2. Apa contoh preprocessing gambar?

3. Mengapa video lebih kompleks daripada gambar tunggal?

Subbab 25 — Data semi terstruktur: JSON, log, event, dan API

Inti subbab: data semi terstruktur perlu dinormalisasi tanpa kehilangan konteks.

Log aplikasi sering berbentuk event:

```
{"user_id":7,"event":"add_to_cart","time":"2026-01-01T10:00:00","metadata":{"item":"kopi","price":18000}}
```

Pertanyaan EDA:

Event apa paling sering?
User mana paling aktif?
Urutan event apa yang umum?
Metadata mana yang sering kosong?
Apakah format berubah antar versi aplikasi?

Data semi terstruktur

Flattening membuat kolom dari key bersarang, tetapi jangan kehilangan urutan event. Untuk analisis perilaku, sequence sering penting.

Tes cepat subbab 25

1. Mengapa JSON disebut semi terstruktur?

2. Apa risiko flattening berlebihan?

3. Sebutkan satu pertanyaan EDA untuk event log.

Subbab 26 — Praktikum terpadu: audit, bersihkan, plot, split, dan laporkan

Inti subbab: pembaca mempraktikkan siklus data sebelum model.

Praktikum menggunakan file:

```
chapters/03b-eksplorasi-visualisasi-data/code/data_exploration_lab.py
chapters/03b-eksplorasi-visualisasi-data/code/data_exploration_lab.ipynb
```

Output yang harus dihasilkan:

```
outputs/data_audit_report.md
outputs/cleaned_customers.csv
outputs/bar_payment.svg
outputs/pie_device.svg
outputs/hist_spend.svg
outputs/scatter_visit_spend.svg
outputs/box_spend_by_segment.svg
outputs/line_daily_sales.svg
outputs/outlier_zscore.svg
outputs/split_manifest.json
```

Laporan akhir minimal berisi:

1. unit observasi,

2. tipe setiap kolom,

3. masalah data mentah,

4. keputusan cleaning,

5. grafik dan insight,

6. outlier dan tindakan,

7. split train/validation/test,

8. potensi leakage,

9. model awal yang disarankan.

Praktikum data

Tes cepat subbab 26

1. Apa output terpenting dari praktikum ini?

2. Mengapa split manifest perlu disimpan?

3. Apa isi minimal laporan data?

Pendalaman praktik — workflow data lapangan dari awal sampai laporan

Bagian ini menyatukan semua subbab menjadi workflow kerja. Pembaca dapat memakai urutan ini sebagai checklist ketika menerima dataset baru dari klien, dosen, tim bisnis, organisasi, atau proyek pribadi.

1. Baca konteks sebelum membaca file

Jangan langsung membuka CSV dan membuat model. Tanyakan dulu:

Siapa pemilik data?

Bagaimana data dikumpulkan?
Apa tujuan analisis?
Apa definisi sukses?
Apa konsekuensi keputusan?
Apa batasan privasi dan etika?

Contoh: data penjualan warung bukan sekadar angka transaksi. Ada promosi, cuaca, hari libur, stok, jam buka, lokasi, dan perilaku pelanggan. Tanpa konteks, grafik bisa benar secara teknis tetapi salah tafsir.

2. Buat data dictionary

Data dictionary adalah tabel kecil yang menjelaskan kolom:

kolom	tipe teknis	tipe makna	contoh	boleh untuk model?	catatan
customer_id	string	ID	C001	tidak	hanya identitas
spend	float	numerik kontinu	58	ya	cek outlier
payment	string	kategori nominal	QRIS	ya	perlu encoding
returned	int	biner	0/1	tergantung target	jangan bocor

Data dictionary mencegah kesalahan seperti merata-ratakan ID, menganggap ordinal sebagai nominal, atau memasukkan kolom target bocor ke fitur.

3. Audit kualitas data dengan angka dan contoh baris

Audit minimal:

jumlah baris
jumlah kolom
missing per kolom
duplikasi
rentang numerik
kategori unik
contoh nilai aneh
tipe data aktual vs tipe makna

Jangan hanya menulis “ada missing”. Tulis angka dan dampaknya. Misalnya: “Kolom spend punya missing 8%, tersebar terutama pada device tablet.” Pernyataan ini lebih berguna daripada “data kurang bersih”.

4. Visualisasi bukan satu grafik, tetapi rangkaian pertanyaan

Untuk satu dataset tabular, urutan visualisasi yang sehat:

bar plot kategori utama
histogram numerik utama
box plot numerik per kategori
scatter dua numerik penting
line plot jika ada waktu
heatmap jika banyak fitur numerik
outlier plot untuk nilai ekstrem

Setiap grafik harus menghasilkan insight tertulis. Jika tidak ada insight, grafik belum selesai. Insight harus berbentuk kalimat:

Mayoritas pembayaran memakai QRIS, sehingga campaign pembayaran digital relevan.
Belanja memiliki ekor kanan panjang, sehingga median lebih aman daripada mean.
Pelanggan VIP membuat outlier belanja; jangan dihapus sebelum validasi bisnis.

5. Putuskan tindakan outlier berdasarkan sebab

Outlier tidak punya satu solusi universal.

Sebab outlier	Tindakan awal
typo input	koreksi jika sumber jelas, atau buang
kejadian valid langka	pertahankan dan beri indikator
fraud	pisahkan untuk investigasi
pelanggan VIP	jangan hapus; mungkin segmen penting
sensor rusak	buang atau imputasi dengan catatan
perubahan sistem	analisis sebelum/sesudah perubahan

Jika outlier dihapus, laporan harus menjawab: berapa baris dihapus, aturan apa yang dipakai, dan apakah distribusi berubah drastis?

6. Data sintetis harus punya tujuan jelas

Data sintetis untuk belajar sangat baik karena aman dan mudah dikontrol. Data sintetis untuk training final perlu sangat hati-hati. Pertanyaan wajib:

Pola apa yang disimulasikan?
Noise apa yang ditambahkan?
Apakah distribusi mirip data asli?
Apakah data sintetis memperkuat bias?
Apakah metrik diuji pada data asli?

Contoh penggunaan aman: membuat dataset kecil untuk menguji apakah pipeline cleaning berjalan. Contoh penggunaan berisiko: mengganti data pelanggan asli sepenuhnya dengan data sintetis lalu mengklaim model siap produksi.

7. Split data harus mengikuti cerita data

Random split tidak selalu salah, tetapi tidak selalu benar. Jika data memiliki waktu, gunakan split waktu. Jika banyak baris berasal dari pelanggan yang sama, gunakan group split. Jika kelas target langka, gunakan stratified split. Jika data berasal dari beberapa toko, pertimbangkan apakah test harus berisi toko yang belum pernah dilihat model.

Pertanyaan anti-leakage:

Apakah baris test punya entitas yang sama dengan train?
Apakah fitur dibuat memakai data masa depan?
Apakah statistik preprocessing dihitung dari seluruh data?
Apakah oversampling dilakukan sebelum split?
Apakah ada kolom yang secara tidak langsung menyimpan target?

8. Pilih model setelah baseline dan EDA

Model awal bukan model paling keren, tetapi model yang menjawab pertanyaan dengan risiko rendah. Untuk tabular kecil, baseline mean/mode dan model linear/tree sering cukup untuk tahap awal. Untuk teks, mulai dari panjang teks, keyword, TF-IDF, lalu embedding. Untuk gambar, mulai dari inspeksi label dan baseline transfer learning. Untuk time series, mulai dari naive baseline seperti “prediksi sama dengan kemarin” sebelum model kompleks.

9. Laporan data harus bisa dibaca non-programmer

Laporan data yang baik berisi:

ringkasan tujuan
unit observasi
kualitas data
keputusan cleaning
visualisasi utama
insight
risiko bias/leakage
split strategy
rekomendasi model awal

pertanyaan lanjutan

Tujuan akhirnya bukan membuat notebook panjang, tetapi membuat keputusan lebih jujur. Jika pembaca bisa menjelaskan dataset kepada orang non-teknis tanpa menyembunyikan risiko, berarti eksplorasi data sudah matang.

Kamus cara membaca rumus Bab 03B

Rumus	Cara membaca	Makna
`X ∈ R^(n×d)`	X adalah matriks dengan n baris dan d fitur.	Bentuk dataset.
`mean = Σx■/n`	Jumlah semua nilai dibagi banyak data.	Nilai rata-rata.
`variance = Σ(x■-μ)²/n`	Rata-rata jarak kuadrat dari mean.	Sebaran data.
`missing_rate = missing/total`	Sel kosong dibagi total sel.	Tingkat data hilang.
`z=(x-μ)/σ`	x dikurangi mean lalu dibagi standar deviasi.	Posisi relatif nilai.
`IQR=Q3-Q1`	Kuartil atas dikurangi kuartil bawah.	Lebar tengah distribusi.
`r=...`	Pergerakan bersama x dan y dinormalisasi.	Korelasi Pearson.
`■=wx+b`	Prediksi y dari garis dengan slope w dan intercept b.	Regresi linear.
`e=y-■`	Aktual dikurangi prediksi.	Residual/outlier tren.

Ringkasan Bab 03B

• Data exploration adalah fondasi AI yang jujur.

• Unit observasi harus jelas sebelum membaca kolom.

• Data bisa terstruktur, semi terstruktur, atau tidak terstruktur.

• Tipe kolom menentukan statistik, visualisasi, preprocessing, dan model.

• Data cleaning mencakup missing value, duplikasi, tipe salah, rentang tidak logis, dan kategori typo.

• Visualisasi harus dipilih berdasarkan pertanyaan dan tipe data.

• Bar plot cocok untuk kategori; histogram untuk distribusi numerik; scatter untuk dua numerik; box plot untuk distribusi dan outlier; line plot untuk waktu; heatmap untuk matriks.

• Pie chart hanya cocok untuk sedikit kategori dan proporsi yang jelas.

• Outlier perlu dipahami, bukan otomatis dibuang.

• Data sintetis berguna untuk belajar, testing pipeline, dan kasus langka, tetapi tidak otomatis mewakili dunia nyata.

• Split train/validation/test harus disesuaikan dengan struktur data.

• Data leakage harus dicegah sejak tahap preprocessing.

• Pemilihan model harus mengikuti tipe data, bukan tren algoritma.

Referensi utama bab

• Tukey, J. W. *Exploratory Data Analysis*.

• McKinney, W. *Python for Data Analysis*.

• VanderPlas, J. *Python Data Science Handbook*.

• Wickham, H., & Grolemund, G. *R for Data Science*.

• Cleveland, W. S. *The Elements of Graphing Data*.

• scikit-learn documentation: preprocessing, model selection, data leakage, train_test_split.

• pandas documentation: data cleaning and exploratory analysis.

• matplotlib and seaborn documentation: plotting APIs.

Catatan validasi internal v0.1

• Bab ditempatkan sebagai Bab 03B karena menjadi jembatan antara Python dasar Bab 3 dan matematika/modeling Bab 4–8.

• Struktur mengikuti pola Bab 7: subbab, tes cepat per subbab, rumus dengan cara membaca, dan praktikum.

• Topik mencakup tipe data, data cleaning, visualisasi, plotting, outlier, data sintetis, splitting, leakage, data tabular, semi terstruktur, tidak terstruktur, dan peta model.