Segmentasi Pelanggan: dari spreadsheet penjualan menjadi kelompok yang bisa ditindaklanjuti

Studi kasus analisis data (Python). Periode transaksi 2025-11-01 s.d. 2026-10-31; tanggal acuan 2026-11-01.

Data sintetis, bukan data klien. Seluruh data pada studi kasus ini dibangkitkan untuk portofolio dan sengaja dibuat berantakan (duplikat, tanggal campur format, nama produk tidak konsisten). Angka yang tampil tidak mewakili toko nyata mana pun, dan hasil pada data nyata dapat berbeda.

Ringkasan temuan

Setelah dibersihkan, data berisi **1.248 pelanggan**, **21.112 pesanan** dan **86.356 baris produk** bernilai total **Rp 4.260,7 jt** dalam 12 bulan. Pelanggan dikelompokkan dengan k-means pada enam ukuran perilaku: keterkinian (recency), frekuensi, rata-rata nilai keranjang, dan tiga porsi belanja (pokok, jajan, perawatan) sebagai ukuran afinitas produk. Jumlah kelompok dipilih **k = 4** berdasarkan silhouette tertinggi (0,362) di antara kandidat yang stabil.

1. Kelompok **Keluarga Setia** hanya 36% dari pelanggan tetapi menyumbang **72%** pendapatan; ini kelompok yang paling perlu dijaga.

2. Pelanggan Tidur berjumlah 376 orang (30%), median 230 hari sejak belanja terakhir, namun hanya 3% pendapatan: layak diajak kembali dengan biaya rendah.

3. Pembeli Jajan Harian sangat sering datang (median 38 pesanan setahun) dengan keranjang kecil (Rp 34.940); **Pemborong Berkala** sebaliknya jarang datang (median 5 pesanan) dengan keranjang Rp 386.250.

Pembersihan data

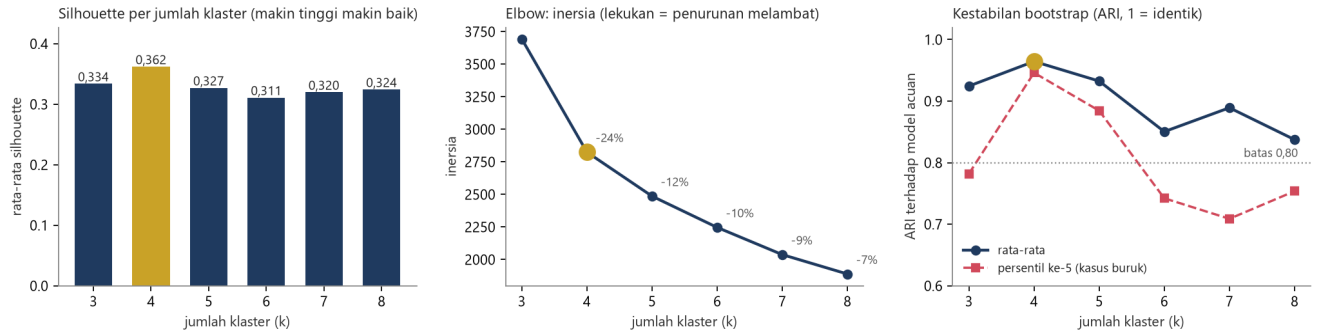
Langkah	Jumlah baris
Baris penjualan mentah	91.682
Variasi penulisan nama produk, dirapikan menjadi	1.099 menjadi 40 produk baku
Baris duplikat persis dibuang	2.843
Format tanggal campur (ISO, hari/bulan/tahun, nama bulan) diseragamkan	seluruh baris terbaca
Tanpa ID pelanggan / ID tak dikenal dibuang	1.338 / 475
Retur, jumlah nol, dan jumlah tak wajar (>100) dikeluarkan	537 + 133
Harga kosong diisi median / harga salah ketik dikoreksi	879 / 425
Baris transaksi bersih akhir	86.356

Catatan: baris tanpa ID tidak bisa dikaitkan ke pelanggan sehingga dibuang; sebagian pesanan karenanya kehilangan beberapa baris produk, dan ini sedikit menurunkan nilai keranjang yang terhitung. Tanggal bergaya angka dibaca hari/bulan/tahun.

Memilih jumlah klaster dan menguji kestabilannya

Aturan ditetapkan sebelum melihat hasil: kandidat k = 3 sampai 8; syaratnya kestabilan bootstrap (ARI rata-rata) minimal 0,80 dan klaster terkecil minimal 5% pelanggan; dari yang memenuhi syarat dipilih silhouette tertinggi. Hasilnya k = 4: silhouette 0,362 (tertinggi), Davies-Bouldin 1,07 (terendah), Calinski-Harabasz 685 (tertinggi), dan kurva elbow melandai setelah k = 4 (penurunan inersia 24% pada k = 4, lalu 12% pada k = 5). Skor silhouette sekitar 0,36 tergolong sedang: pelanggan nyata membentuk kontinum, bukan gugus yang terpisah tegas.

Pemilihan jumlah klaster: k = 4 (ditandai emas) pada data sintetis 1.248 pelanggan



k	Silhouette	Davies-Bouldin	Calinski-Harabasz	Penurunan inersia	ARI bootstrap (rata-rata / p5)	Klaster terkecil
3	0,334	1,25	639	-	0,924 / 0,782	20,7%
4	0,362	1,07	685	23,5%	0,965 / 0,946	13,5%
5	0,327	1,18	625	12,0%	0,932 / 0,884	11,8%
6	0,311	1,26	580	9,6%	0,850 / 0,742	10,9%
7	0,320	1,20	553	9,3%	0,889 / 0,709	4,9%
8	0,324	1,22	525	7,3%	0,837 / 0,754	3,7%

Uji kestabilan model akhir

Ulang dengan seed lain: 30 inisialisasi acak (satu kali jalan tiap seed, tanpa pemilihan terbaik) menghasilkan ARI rata-rata 0,981 terhadap model acuan; satu seed terburuk hanya 0,53, yaitu terjebak di solusi lokal. Karena itu model akhir memakai 50 inisialisasi dan memilih yang inersianya terendah. **Bootstrap:** 200 sampel ulang pelanggan; model dilatih di tiap sampel lalu dipakai menilai semua pelanggan. ARI rata-rata 0,966, persentil ke-5 0,940. Kemiripan Jaccard terbaik per segmen: Pembeli Jajan Harian 0,99; Pemborong Berkala 0,96; Pelanggan Tidur 0,97; Keluarga Setia 0,98. Semua di atas 0,75 sehingga segmen dianggap stabil.

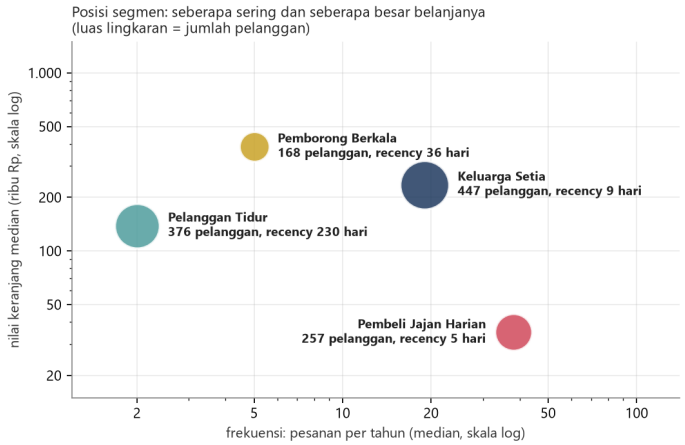
Penerapan ulang: model tersimpan menghasilkan label yang sama persis untuk 100% pelanggan. Bila dinilai dengan data 30 hari lebih awal lalu dibandingkan dengan sekarang, 94,7% dari 1.192 pelanggan tetap berada di segmen yang sama; sisanya berpindah karena perilaku mereka memang berubah.

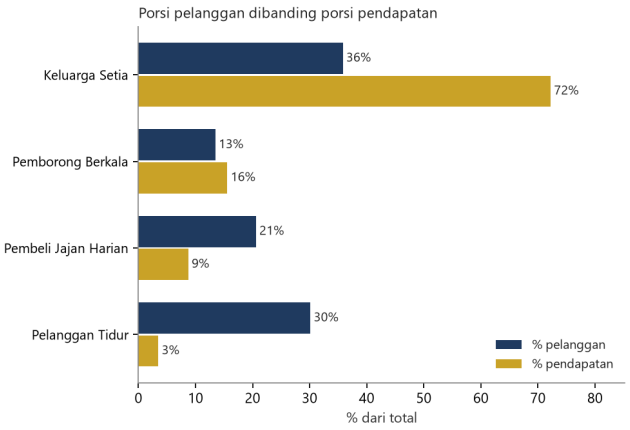
Pembanding dengan kunci jawaban pembangkit data

Pembangkit data menyisipkan 6 kelompok perilaku. Kunci ini **tidak dipakai** untuk memilih fitur, k, maupun aturan nama; ia dibuka hanya di tahap akhir. Hasil: ARI 0,56, NMI 0,64, kemurnian 66%. Model menemukan 4 segmen karena beberapa kelompok asli memang mirip perilakunya: misalnya kelompok asli A dan B sama-sama berada terutama di Keluarga Setia (100% dan 82%). Kelemahan yang terlihat: sekitar 16% pelanggan tidur asli (kelompok E) masuk ke segmen lain karena keranjang dan porsi belanjanya mirip segmen itu walau sudah lama tidak belanja; di model, 56 anggota Pemborong Berkala berstatus lama tak belanja (lebih dari 150 hari) dan diberi penanda di lembar Pelanggan. Angka ini berasal dari data sintetis yang relatif bersih dan **bukan janji hasil pada data nyata**.

Profil dan ciri tiap segmen

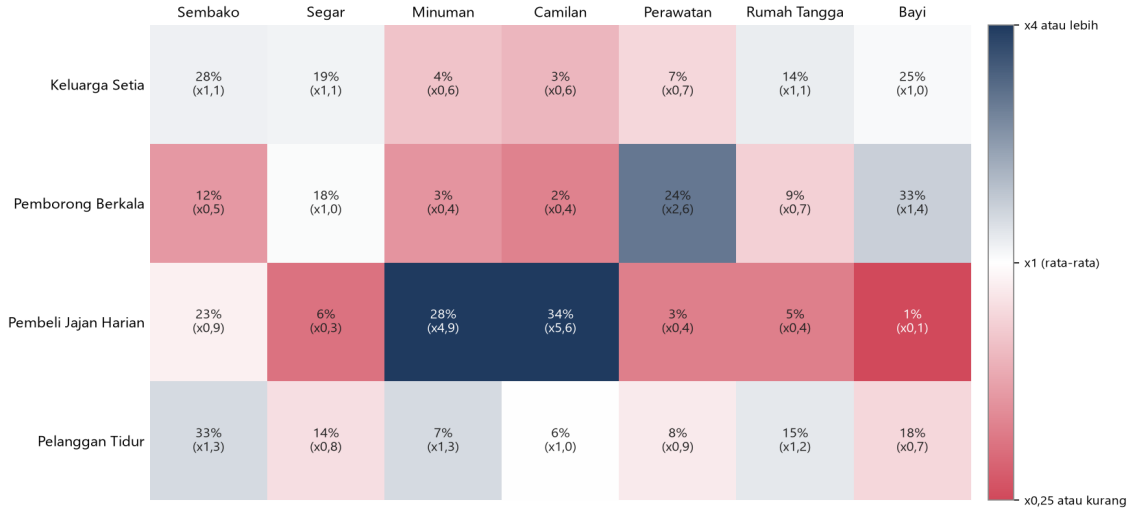
Profil segmen pelanggan (data sintetis, 12 bulan terakhir)





Segmen	Pelanggan	Pendapatan	Recency median	Pesanan /tahun	Jeda antar pesanan	Keranjang median	Baris berdiskon	Belanja rata-rata
Keluarga Setia	447 (36%)	72%	9 hari	19	17 hari	Rp 235.800	32%	Rp 6,9 jt
Pemborong Berkala	168 (13%)	16%	36 hari	5	33 hari	Rp 386.250	18%	Rp 3,9 jt
Pembeli Jajan Harian	257 (21%)	9%	5 hari	38	9 hari	Rp 34.940	9%	Rp 1,4 jt
Pelanggan Tidur	376 (30%)	3%	230 hari	2	23 hari	Rp 138.838	12%	Rp 0,4 jt

Peta panas afinitas produk: porsi belanja per kategori dan indeks terhadap rata-rata toko



Sel: % nilai belanja segmen pada kategori tersebut (x = indeks: 1 = sama dengan rata-rata seluruh pelanggan; warna dipotong di x0,25 dan x4). Data sintetis.

Satu catatan: segmen **Keluarga Setia** adalah yang terbesar dan masih beragam. 43% anggotanya sangat peka diskon (30% baris atau lebih berdiskon) sedangkan 53% hampir tak memakainya. Pemecahan menjadi k = 5 tidak dipilih karena silhouette (0,327) dan kestabilannya turun, jadi perbedaan itu ditangani lewat filter % baris diskon pada lembar Pelanggan di workbook.

Rekomendasi promo dan stok per segmen

Segmen	Promo	Stok
Keluarga Setia 447 pelanggan	Mempertahankan 447 pelanggan yang menyumbang 72% pendapatan dengan belanja rutin (median 19 pesanan setahun). Poin member dan harga langganan untuk produk andalan (Susu Formula 400g, Beras Premium 5kg, Deterjen Bubuk 1.8kg); tawarkan produk rumah tangga/bayi sebagai silang jual. 32% baris belanja mereka sudah berdiskon, jadi batasi diskon tambahan agar margin tidak tergerus. Segmen ini masih beragam: 43% pelanggannya sangat peka diskon (30% baris atau lebih berdiskon) sedangkan 53% hampir tak memakainya, jadi kirim promo hanya ke kelompok pertama (filter kolom % baris diskon di lembar Pelanggan).	Stok penyangga terbesar ada di sini: pokok dan rumah tangga. Produk teratas: Susu Formula 400g, Beras Premium 5kg, Deterjen Bubuk 1.8kg. Tinjau kebutuhan mingguan di lembar Stok_per_Segmen.
Pemborong Berkala 168 pelanggan	Memperpendek jeda belanja 168 pelanggan berkeranjang besar (Rp 386.250 per pesanan) yang datang tiap ~33 hari. Diskon bertingkat untuk keranjang di atas Rp 386.250, paket belanja bulanan, dan pengingat sebelum jeda biasanya berakhir. Tonjolkan kategori Perawatan dan Bayi serta Serum Wajah 30ml, Lotion Tubuh 200ml. Catatan: 56 anggota segmen ini sudah lebih dari 150 hari tidak belanja (lihat kolom Status di lembar Pelanggan); perlakukan mereka dengan kampanye win-back, bukan promo rutin.	Jaga ketersediaan item bernilai tinggi dan kategori Perawatan dan Bayi; sediakan layanan pesan-dulu agar stok tidak habis saat mereka datang.
Pembeli Jajan Harian 257 pelanggan	Menjaga kebiasaan 257 pembeli kecil tetapi sangat sering (median 38 pesanan setahun) dan menaikkan nilai keranjang dari Rp 34.940. Paket bundel harga tetap minuman + camilan (Keripik Kentang 70g, Kopi Sachet 10pcs, Susu UHT 1L) dan program stempel: setiap kunjungan ke-10 mendapat bonus. Hindari diskon dalam per item karena margin tipis.	Minuman dan camilan paling cepat berputar: isi ulang rak tiap 2-3 hari. Produk andalan: Keripik Kentang 70g, Kopi Sachet 10pcs, Susu UHT 1L.
Pelanggan Tidur 376 pelanggan	Menarik kembali 376 pelanggan yang median-nya sudah 230 hari tidak belanja. Kampanye win-back berbatas waktu (14 hari) lewat pesan langsung: voucher belanja ulang bernilai tetap, bukan diskon luas. Angkat produk yang dulu mereka beli (Beras Premium 5kg, Susu Formula 400g, Deterjen Bubuk 1.8kg). Ukur dengan persentase yang kembali belanja dalam 30 hari.	Tidak perlu stok khusus. Pastikan produk pemicu kembali (kategori Minuman dan Sembako) selalu tersedia selama kampanye.

Rekomendasi disusun dari pola pada data sintetis; uji dulu pada kelompok kecil (A/B) dengan ukuran keberhasilan yang jelas sebelum dipakai luas.

Menjalankan ulang bila ada transaksi baru

Skrip **segmentasi.py** (disertakan) membersihkan data, menghitung fitur, melatih model, dan menerapkan model tersimpan. Perintahnya:

python segmentasi.py latih --penjualan P.csv --pelanggan K.csv --master M.csv --keluar-model model_segmen.json --keluar segmen.csv
python segmentasi.py terapkan --penjualan P_baru.csv --pelanggan K.csv --master M.csv --model model_segmen.json --keluar segmen_baru.csv

Gunakan **terapkan** tiap bulan agar nama segmen konsisten dan pelanggan baru langsung terkelompok; gunakan **latih** tiap kuartal atau saat komposisi bergeser jauh, lalu periksa kembali silhouette dan kestabilan. Fitur dihitung dari 365 hari terakhir sebelum tanggal acuan. Nama segmen berasal dari aturan bisnis pada fungsi beri_nama (recency, frekuensi, nilai keranjang, porsi kategori) dan dapat disesuaikan. Panduan yang sama ada di lembar Panduan pada workbook.

Batas hasil

1. Data sintetis: pola dibuat oleh pembangkit data, sehingga segmen terlihat lebih rapi daripada data nyata. **2.** Tidak ada data biaya atau margin, jadi rekomendasi bersifat arah, bukan perhitungan laba. **3.** Satu tahun transaksi tidak menangkap musim (misalnya Lebaran) berulang. **4.** Segmentasi menggambarkan perilaku lampau; ia bukan prediksi pelanggan akan membeli.

Berkas pendukung: segmentasi_pelanggan.xlsx (12 lembar: ringkasan, profil, rekomendasi, afinitas, stok, pelanggan, pesanan bersih, pemilihan k, kestabilan, evaluasi, log, panduan), segmen_pelanggan.csv, model_segmen.json, segmentasi.py, dan tiga grafik PNG.