matplotlib

2025

MODUL PRAKTIKUM 5

VISUALIZATION BIG DATA USING MATPLOTLIB

OPTIMIZED VERSION

Disusun Oleh:
Ryan F F Hakim
Silvi Nurinsan

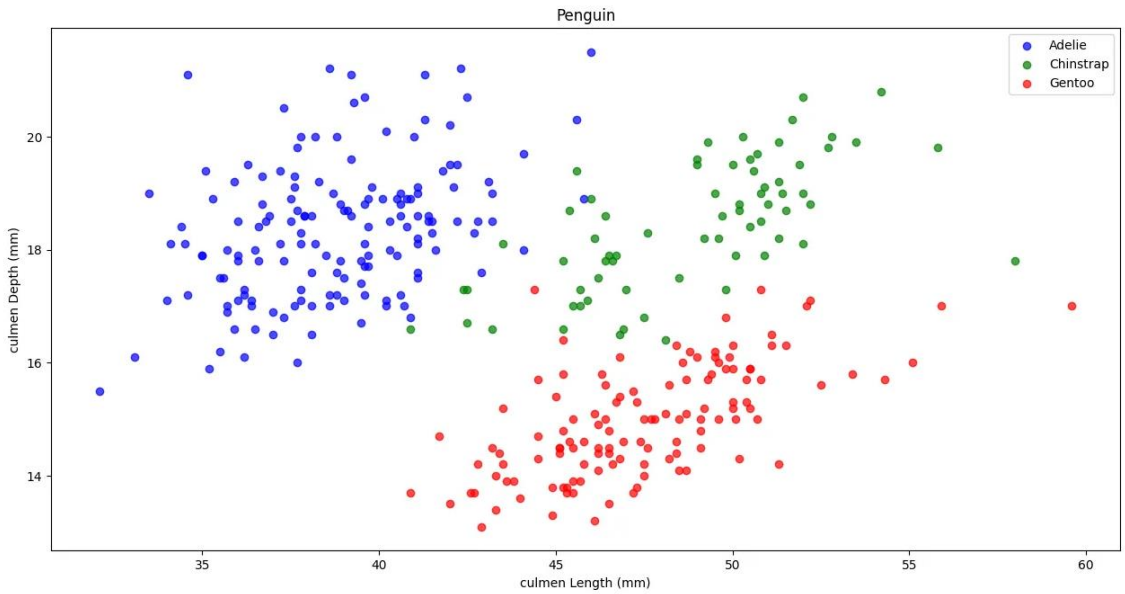
Dipersembahkan Untuk:
Sains Data

Mata Kuliah
Analisis Big Data

MENGAPA VISUALISASI DATA PENTING

Visualisasi data adalah komponen penting dari analisis data modern, yang memungkinkan transformasi kumpulan data kompleks menjadi wawasan yang dapat dimengerti dan ditindaklanjuti. Di era big data, kemampuan untuk menafsirkan sejumlah besar informasi dengan cepat dan akurat adalah hal yang terpenting. Visualisasi data berfungsi sebagai jembatan antara data mentah dan kognisi manusia, memungkinkan analis untuk mengungkap tren, anomali, dan hubungan yang mungkin tersembunyi.

Manusia memproses informasi visual 60.000 kali lebih cepat daripada teks, yang menggarisbawahi efisiensi visualisasi dalam menafsirkan data. Alat visual seperti bagan dan grafik membantu dalam mengenali pola, korelasi, dan pencilan. Visualisasi menyederhanakan data multidimensi, mengurangi beban kognitif, dan dengan menyajikan data secara intuitif, mendukung pengambilan keputusan yang lebih cepat dan lebih tepat. Penelitian menunjukkan bahwa alat bantu visual meningkatkan retensi data hingga 80% dibandingkan dengan data tekstual saja. Selain itu, representasi visual data mengarah pada keputusan yang lebih cepat dan lebih akurat, dengan partisipan dalam kelompok yang dibantu visualisasi menunjukkan peningkatan akurasi sebesar 25%.



PERAN VISUALISASI DATA DALAM SIKLUS ANALISIS BIG DATA

Visualisasi data memainkan peran penting dalam setiap tahapan siklus analisis big data. Siklus ini umumnya melibatkan pengumpulan data, pemrosesan data, pembersihan data, analisis data, dan akhirnya komunikasi hasil.

Pada tahap awal pengumpulan dan pemrosesan data, visualisasi dapat membantu dalam memahami struktur data mentah dan mengidentifikasi potensi masalah kualitas data. Selama pembersihan data, visualisasi dapat digunakan untuk mendeteksi pencilan atau nilai yang hilang yang mungkin memerlukan perhatian.

Tahap analisis data sangat bergantung pada visualisasi. Analis menggunakan berbagai jenis plot untuk mengeksplorasi hubungan antar variabel, mengidentifikasi tren, pola, dan anomali dalam kumpulan data yang besar dan kompleks. Misalnya, scatter plot dapat mengungkapkan korelasi antara dua variabel numerik, sementara histogram dapat menunjukkan distribusi variabel tunggal. Kemampuan untuk memvisualisasikan data secara interaktif memungkinkan analis untuk menggali lebih dalam dan mengajukan pertanyaan yang lebih bernuansa tentang data.

Terakhir, dalam tahap komunikasi hasil, visualisasi adalah alat yang sangat diperlukan. Menyajikan temuan analisis data dalam format visual yang jelas dan ringkas jauh lebih efektif daripada menyajikan tabel angka mentah. Visualisasi yang dirancang dengan baik dapat menceritakan sebuah kisah, menyoroti wawasan utama, dan memfasilitasi pemahaman di antara audiens yang beragam, termasuk pemangku kepentingan non-teknis. Hal ini memungkinkan pengambilan keputusan berbasis data yang lebih baik di berbagai domain.

APA ITU MATPLOTLIB

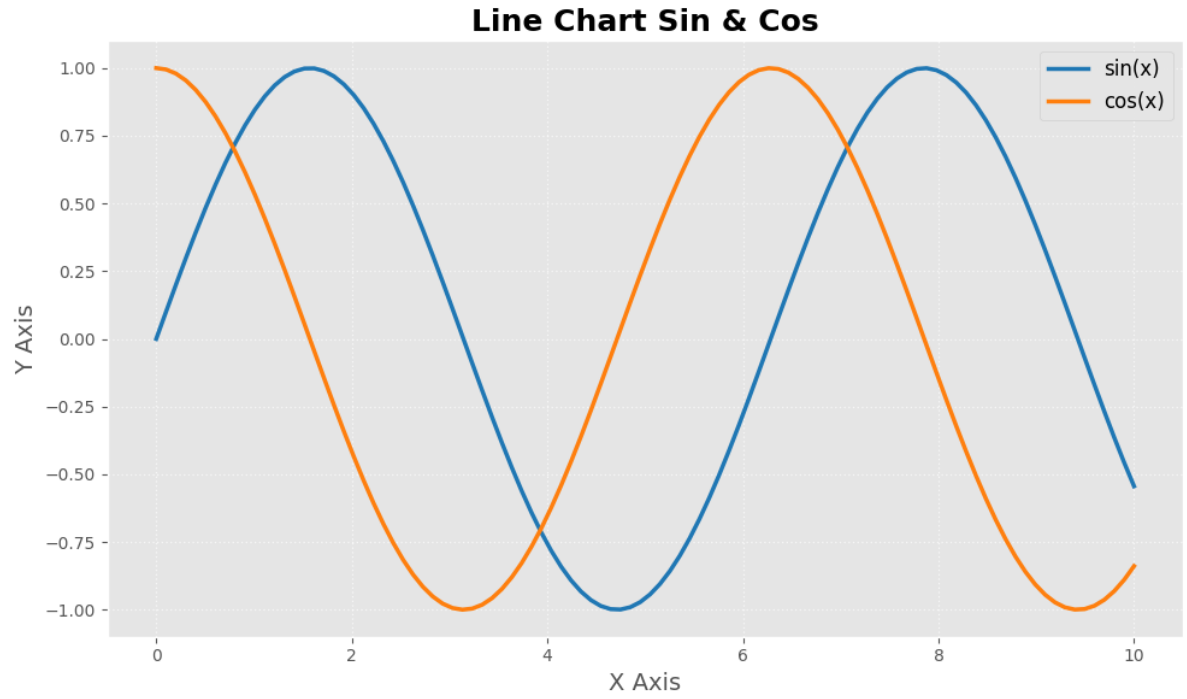
Matplotlib adalah pustaka plotting 2D portabel dan paket pencitraan yang terutama ditujukan untuk visualisasi data ilmiah, teknik, dan keuangan dalam bahasa pemrograman Python. Dikembangkan oleh John D. Hunter pada tahun 2003, Matplotlib memungkinkan pengguna untuk merepresentasikan data secara grafis, memfasilitasi analisis dan pemahaman yang lebih mudah. Pustaka ini dirancang dengan filosofi bahwa pengguna harus dapat membuat plot sederhana hanya dengan beberapa perintah. Matplotlib dapat digunakan secara interaktif dari shell Python, dipanggil dari skrip Python, atau disematkan dalam aplikasi GUI.

Matplotlib menawarkan berbagai jenis plot, termasuk grafik garis, diagram batang, histogram, scatter plot, diagram lingkaran, dan bahkan plot 3D. Fitur utamanya meliputi kemampuan untuk membuat plot berkualitas publikasi, membuat gambar interaktif yang dapat diperbesar, digeser, dan diperbarui, menyesuaikan gaya visual dan tata letak, mengekspor ke berbagai format file, dan menyematkan dalam JupyterLab serta Antarmuka Pengguna Grafis (GUI). Pustaka ini juga terintegrasi dengan baik dengan pustaka ilmu data Python lainnya seperti NumPy dan Pandas.

Komponen utama dari plot Matplotlib meliputi *Figure*, yang merupakan wadah menyeluruh yang menampung semua elemen plot; *Axes*, yaitu area di dalam *figure* tempat data diplot; *Axis*, yang merepresentasikan sumbu x dan y; serta elemen seperti garis, penanda, judul, dan label. Modul **pyplot** dalam Matplotlib menyediakan antarmuka seperti MATLAB untuk membuat plot dengan lebih sederhana.

LINE CHART

Grafik garis adalah jenis grafik yang menampilkan informasi sebagai serangkaian titik data yang disebut 'penanda' (markers) yang dihubungkan oleh segmen garis lurus. Titik-titik data ini biasanya diurutkan berdasarkan sumbu horizontal (seringkali waktu atau kategori berurutan). Sumbu vertikal mewakili nilai kuantitatif. Grafik garis sangat efektif dalam menunjukkan tren atau perubahan data dari waktu ke waktu atau melalui urutan tertentu.



KAPAN DIGUNAKAN

1. Menunjukkan Tren Seiring Waktu: Ini adalah penggunaan paling umum. Misalnya, melacak perubahan harga saham selama setahun, pertumbuhan suhu bulanan, atau jumlah pengunjung situs web harian.

2. Membandingkan Beberapa Seri Data: Anda dapat memplot beberapa garis pada grafik yang sama untuk membandingkan tren antara beberapa kelompok atau kategori. Misalnya, membandingkan penjualan tiga produk berbeda selama periode yang sama.

3. Data Berkelanjutan (Continuous Data): Sangat cocok untuk data yang nilainya dapat berubah secara terus menerus, seperti suhu, berat, atau waktu.

4. Mengidentifikasi Pola dan Perubahan Cepat: Perubahan curam atau landai pada garis dapat dengan mudah menunjukkan percepatan, perlambatan, atau volatilitas dalam data.

5. Ketika Urutan Penting: Jika urutan titik data memiliki makna (misalnya, langkah-langkah dalam suatu proses, tahapan perkembangan), grafik garis dapat menggambarkannya dengan baik.

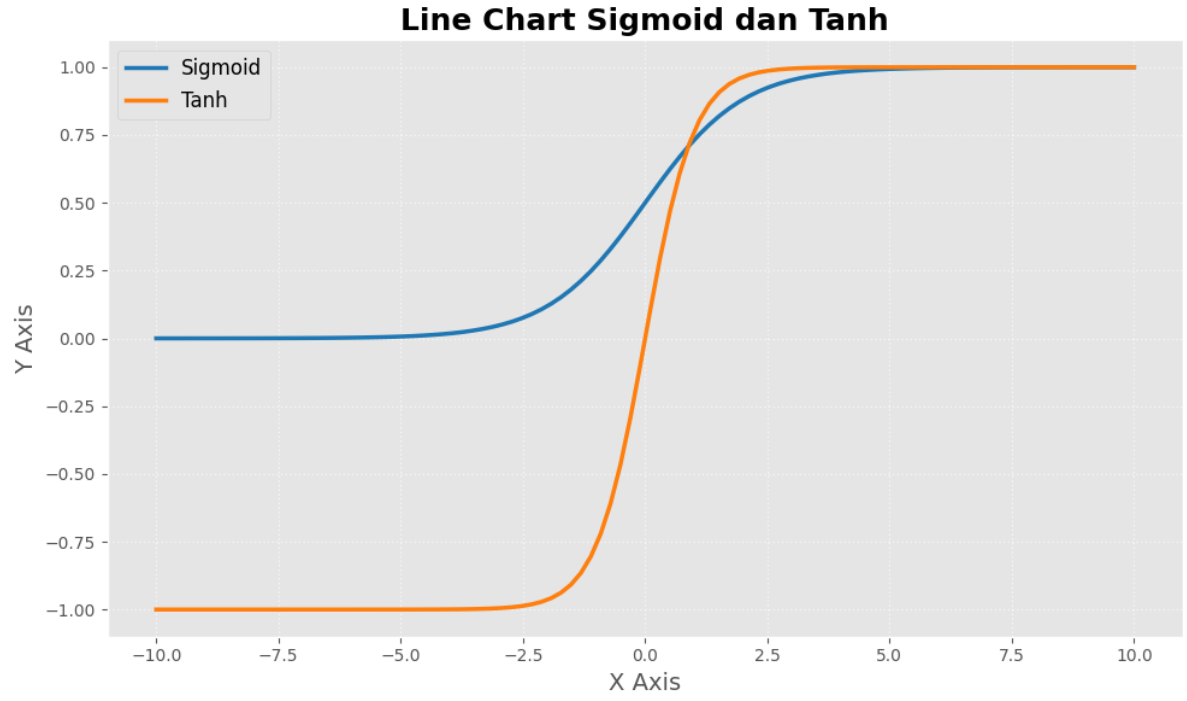
Interpolasi vs. Ekstrapolasi: Garis di antara titik-titik data menyiratkan adanya nilai di antara titik-titik tersebut (interpolasi). Namun, berhati-hatilah saat mencoba memprediksi nilai di luar rentang data yang ada (ekstrapolasi), karena tren mungkin tidak berlanjut.

KAPAN SEBAIKNYA DIHINDARI

- Data Kategorikal Tanpa Urutan Alami: Jika sumbu X adalah kategori yang tidak memiliki urutan logis (misalnya, jenis buah, nama negara), menghubungkannya dengan garis bisa menyesatkan karena menyiratkan kesinambungan atau urutan yang tidak ada. Diagram batang lebih cocok dalam kasus ini.
- Terlalu Banyak Garis: Jika memplot terlalu banyak seri data (misalnya, lebih dari 5-7 garis), grafik bisa menjadi berantakan dan sulit dibaca. Pertimbangkan untuk memecahnya menjadi beberapa grafik atau menggunakan teknik lain.

PENTINGNYA RASIO ASPEK

Rasio antara tinggi dan lebar grafik dapat memengaruhi persepsi kemiringan garis. Rasio aspek yang sangat berbeda dapat membuat tren terlihat lebih curam atau lebih landai dari yang sebenarnya.



GRAFIK GARIS SUMBU GANDA (DUAL-AXIS)

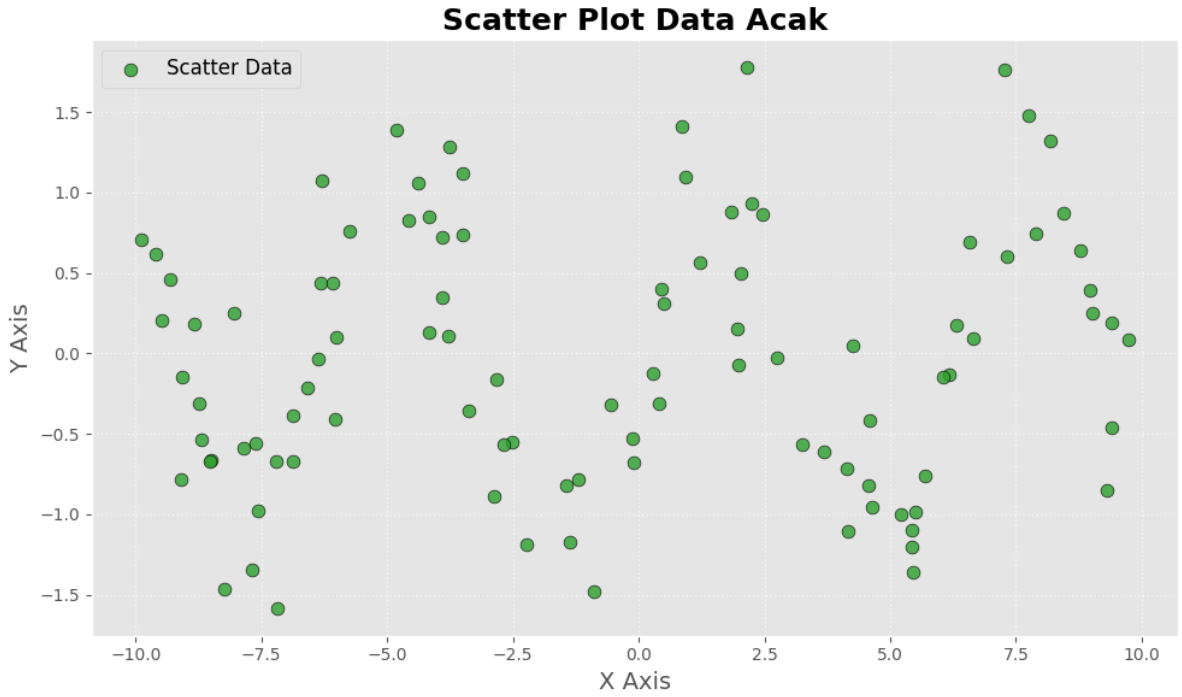
Terkadang digunakan untuk memplot dua seri data dengan skala yang sangat berbeda pada grafik yang sama menggunakan sumbu Y kiri dan kanan. Namun, ini harus digunakan dengan hati-hati karena dapat menyesatkan jika skala tidak dipilih dengan cermat, dan dapat menyiratkan hubungan yang sebenarnya tidak ada.

MENYOROTI TITIK DATA PENTING

Dapat menggunakan penanda (markers) yang lebih besar atau warna berbeda untuk menyoroti titik data tertentu yang signifikan, seperti puncak, lembah, atau anomali.

SCATTER PLOT

Diagram pencar menggunakan titik-titik untuk merepresentasikan nilai dari dua variabel numerik yang berbeda. Posisi setiap titik pada sumbu horizontal (X) dan sumbu vertikal (Y) menunjukkan nilai untuk satu titik data individu. Diagram ini digunakan untuk mengamati dan menunjukkan hubungan atau korelasi antara dua variabel tersebut.



KAPAN DIGUNAKAN

1. Mengidentifikasi Hubungan/Korelasi: Untuk melihat apakah ada hubungan antara dua variabel.
    - Korelasi Positif: Ketika satu variabel meningkat, variabel lainnya juga cenderung meningkat (titik-titik membentuk pola menanjak).
    - Korelasi Negatif: Ketika satu variabel meningkat, variabel lainnya cenderung menurun (titik-titik membentuk pola menurun).
    - Tidak Ada Korelasi: Titik-titik tersebar secara acak tanpa pola yang jelas.
2. Melihat Pola Distribusi Data: Bagaimana data tersebar atau mengelompok. Ini bisa menunjukkan adanya cluster atau kelompok data tertentu.
3. Mengidentifikasi Outlier (Pencilan): Titik data yang jauh berbeda dari pola umum dapat dengan mudah terlihat.

4. Analisis Regresi: Sering digunakan sebagai langkah awal sebelum melakukan analisis regresi untuk memodelkan hubungan antar variabel.

5. Ketika Memiliki Banyak Titik Data: Efektif untuk menampilkan sejumlah besar pasangan data dan melihat gambaran besarnya.

KORELASI BUKAN KAUSALITAS

Ini adalah poin krusial. Diagram pencar dapat menunjukkan adanya hubungan (korelasi) antara dua variabel, tetapi tidak secara otomatis membuktikan bahwa satu variabel menyebabkan perubahan pada variabel lainnya. Mungkin ada variabel ketiga yang memengaruhi keduanya, atau hubungannya bisa jadi kebetulan.

JENIS KORELASI: SELAIN POSITIF, NEGATIF, ATAU TIDAK ADA KORELASI, POLA BISA LEBIH KOMPLEKS

- Linear: Titik-titik cenderung membentuk garis lurus.
- Non-linear (Curvilinear): Titik-titik membentuk kurva (misalnya, parabola, eksponensial).

MENAMBAHKAN GARIS TREN (TREND LINE)

Seringkali berguna untuk menambahkan garis tren (misalnya, garis regresi linear) ke diagram pencar untuk membantu memvisualisasikan kekuatan dan arah hubungan secara keseluruhan.

MASALAH OVERPLOTTING

Jika memiliki banyak sekali titik data, titik-titik tersebut dapat tumpang tindih sehingga sulit untuk melihat kepadatan atau pola sebenarnya. Solusi meliputi:

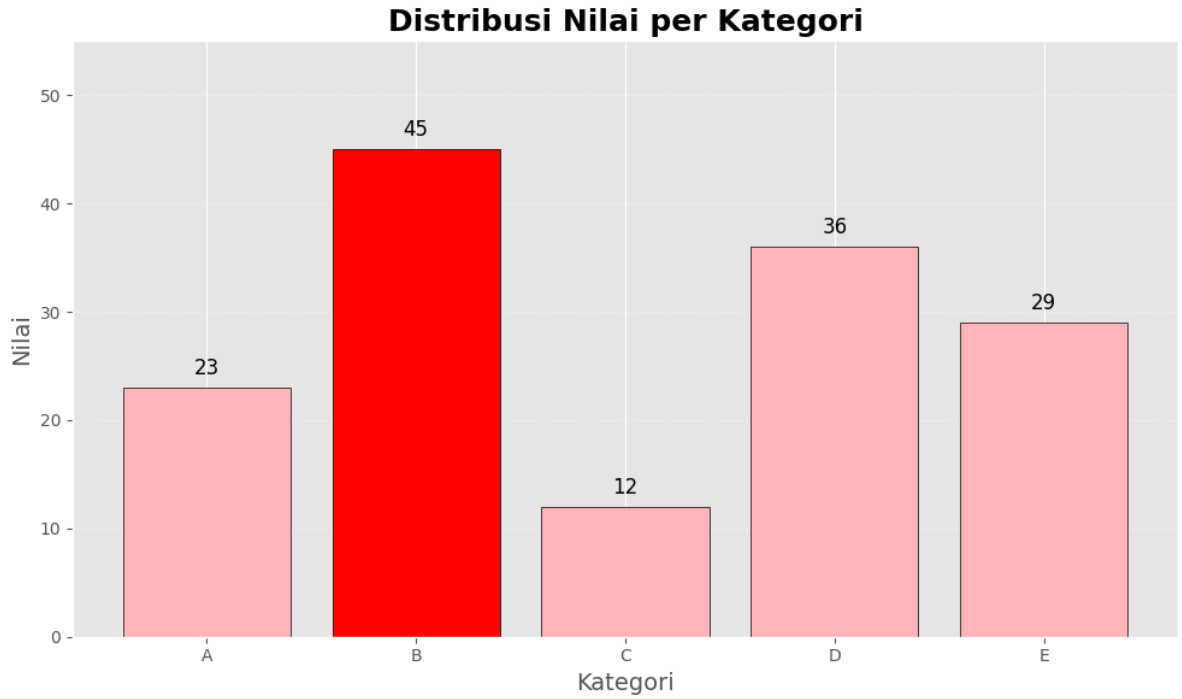
- Transparansi (Alpha Blending): Membuat titik semi-transparan sehingga area padat terlihat lebih gelap.
- Jittering: Menambahkan sedikit "noise" acak ke posisi titik untuk memisahkannya.
- Sampling: Hanya memplot sebagian kecil dari data jika dataset sangat besar.
- Heatmap atau Density Plot: Menggunakan warna untuk menunjukkan kepadatan titik di area tertentu.

VARIABEL KETIGA

Dapat menambahkan dimensi ketiga ke diagram pencar dengan menggunakan ukuran atau warna titik untuk mewakili variabel ketiga. Ini disebut bubble chart jika ukuran digunakan.

BAR CHART

Diagram batang merepresentasikan data kategorikal dengan batang persegi panjang. Panjang atau tinggi batang proporsional dengan nilai yang diwakilinya. Batang dapat diplot secara vertikal atau horizontal. Setiap batang mewakili kategori tertentu, dan ada jarak antar batang untuk menekankan bahwa kategori tersebut diskrit.



KAPAN DIGUNAKAN:

1. Membandingkan Nilai Antar Kategori: Sangat baik untuk membandingkan besaran atau frekuensi dari kategori yang berbeda. Misalnya, membandingkan jumlah penjualan berdasarkan produk, populasi berdasarkan negara, atau jumlah suara untuk kandidat yang berbeda.

2. Menampilkan Peringkat: Mudah untuk melihat kategori mana yang memiliki nilai tertinggi atau terendah.

3. Data Diskrit: Cocok untuk data yang dikelompokkan ke dalam kategori yang terpisah dan tidak tumpang tindih.

4. Ketika Label Kategori Panjang: Diagram batang horizontal lebih disukai jika label kategori panjang, karena memberikan lebih banyak ruang untuk teks.

5. Menunjukkan Perubahan Seiring Waktu (dengan Batasan): Meskipun grafik garis lebih umum untuk tren waktu, diagram batang dapat digunakan untuk membandingkan nilai pada titik waktu diskrit (misalnya, penjualan per kuartal).

DIAGRAM BATANG BERTUMPUK (STACKED) VS. BERKELOMPOK (GROUPED)

- Bertumpuk: Setiap batang dibagi menjadi beberapa segmen yang mewakili sub-kategori. Baik untuk menunjukkan total dan proporsi sub-kategori dalam setiap kategori utama.

- Berkelompok: Batang untuk sub-kategori ditempatkan berdampingan dalam setiap kategori utama. Lebih baik untuk membandingkan nilai sub-kategori secara langsung antar kategori utama.

PENTINGNYA MEMULAI SUMBU Y DARI NOL

Untuk perbandingan yang akurat, sumbu nilai (biasanya sumbu Y untuk batang vertikal) harus selalu dimulai dari nol. Jika tidak, perbedaan antar batang dapat terlihat dilebih-lebihkan atau diremehkan.

HORIZONTAL VS. VERTIKAL:

- Vertikal: Umum digunakan, terutama jika label kategori pendek.
- Horizontal: Lebih baik jika label kategori panjang, karena memberikan lebih banyak ruang dan keterbacaan. Juga baik untuk menampilkan peringkat.

KETERBATASAN DALAM MENUNJUKKAN DISTRIBUSI

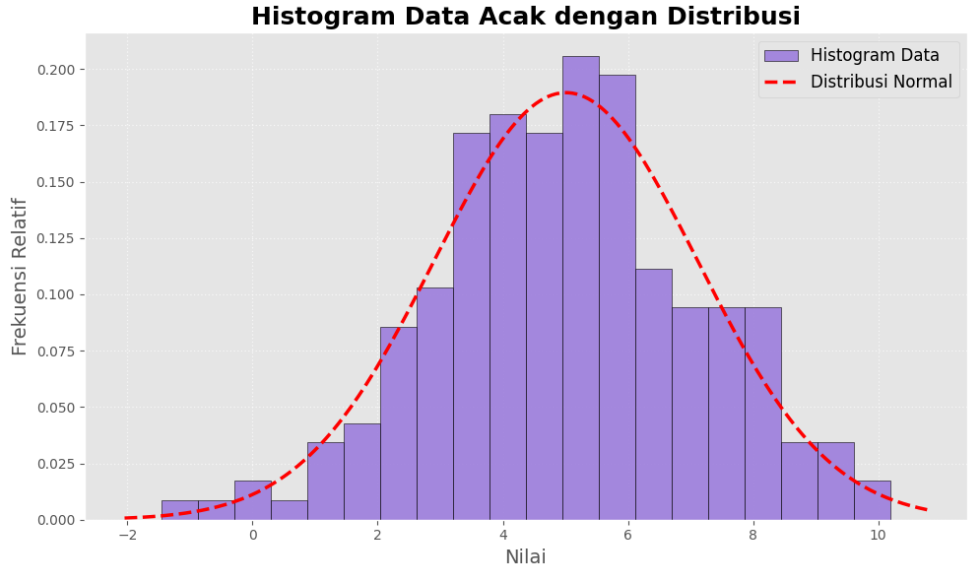
Diagram batang menunjukkan nilai tunggal (misalnya, rata-rata, total) untuk setiap kategori. Ia tidak menunjukkan bagaimana data terdistribusi di dalam kategori tersebut. Jika Anda perlu menunjukkan distribusi, histogram atau box plot mungkin lebih sesuai.

URUTAN KATEGORI

Mengurutkan batang berdasarkan nilai (misalnya, dari tertinggi ke terendah) seringkali dapat membuat grafik lebih mudah dibaca dan dipahami, kecuali jika ada urutan alami untuk kategori tersebut (misalnya, bulan dalam setahun).

HISTOGRAM

Histogram adalah representasi grafis dari distribusi frekuensi data numerik yang berkelanjutan. Berbeda dengan diagram batang yang membandingkan kategori, histogram menunjukkan seberapa sering nilai dalam rentang tertentu (disebut 'bin' atau 'interval') muncul. Batang-batang dalam histogram saling bersentuhan (kecuali jika ada bin tanpa data), menunjukkan sifat berkelanjutan dari data yang diukur. Lebar batang mewakili interval, dan tinggi batang mewakili frekuensi atau jumlah data dalam interval tersebut.



KAPAN DIGUNAKAN

1. Memahami Distribusi Data: Untuk melihat bentuk distribusi data numerik:
    - Normal (Simetris/Lonceng): Sebagian besar data terpusat di tengah.
    - Miring ke Kanan (Positively Skewed): Ekor panjang di sisi kanan, sebagian besar data di sisi kiri.
    - Miring ke Kiri (Negatively Skewed): Ekor panjang di sisi kiri, sebagian besar data di sisi kanan.
    - Bimodal: Memiliki dua puncak.
    - Uniform: Semua interval memiliki frekuensi yang kurang lebih sama.
2. Melihat Frekuensi dalam Interval Tertentu: Mengetahui berapa banyak titik data yang jatuh ke dalam setiap rentang nilai.
3. Mengidentifikasi Pusat, Sebaran, dan Bentuk Data: Memberikan gambaran tentang tendensi sentral (misalnya, di mana data paling banyak terkumpul) dan variabilitas data.
4. Ketika Bekerja dengan Data Numerik Berkelanjutan: Seperti tinggi badan, berat badan, suhu, waktu, skor ujian.
5. Menentukan Jumlah Bin yang Tepat: Jumlah bin dapat memengaruhi interpretasi, jadi penting untuk memilih jumlah yang sesuai untuk mengungkapkan pola dalam data.

PERBEDAAN KUNCI DENGAN DIAGRAM BATANG

1. Data: Histogram untuk data numerik berkelanjutan; diagram batang untuk data kategorikal.
2. Sumbu X: Sumbu X histogram adalah skala numerik yang dibagi menjadi interval (bin); sumbu X diagram batang adalah kategori diskrit.
3. Spasi Antar Batang: Batang histogram biasanya bersentuhan (menunjukkan kesinambungan); batang pada diagram batang memiliki spasi.

KRITISNYA PEMILIHAN LEBAR BIN (INTERVAL)

1. Terlalu Lebar: Dapat menyembunyikan detail penting dalam distribusi.
2. Terlalu Sempit: Dapat menghasilkan banyak noise dan membuat pola sulit dilihat.
3. Tidak ada aturan baku, seringkali memerlukan eksperimen. Beberapa perangkat lunak memiliki algoritma default (misalnya, aturan Freedman-Diaconis, aturan Sturges).

HUBUNGAN DENGAN FUNGSI KEPADATAN PROBABILITAS (PDF)

Histogram dapat dianggap sebagai perkiraan empiris dari PDF yang mendasari data. Jika Anda menormalkan area total histogram menjadi 1, bentuknya akan mendekati PDF.

BUKAN UNTUK PERBANDINGAN LANGSUNG ANTAR DATASET YANG BERBEDA UKURAN

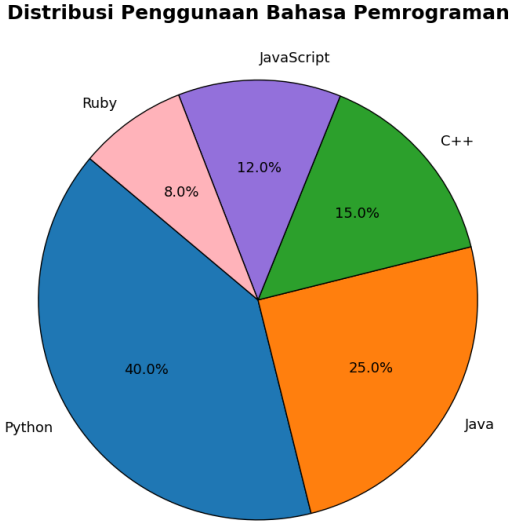
Jika Anda ingin membandingkan distribusi dua dataset dengan jumlah observasi yang berbeda menggunakan histogram, lebih baik menggunakan frekuensi relatif (persentase) daripada frekuensi absolut pada sumbu Y, atau menormalkan histogram.

BENTUK DISTRIBUSI MEMBERI WAWASAN

1. Unimodal: Satu puncak, menunjukkan nilai yang paling umum.
2. Bimodal/Multimodal: Dua atau lebih puncak, mungkin menunjukkan adanya sub-populasi yang berbeda dalam data.
3. Skewness (Kemiringan): Memberi tahu apakah data cenderung mengumpul di satu sisi dan memiliki ekor yang panjang di sisi lain.
4. Kurtosis: Mengukur "kepuncakan" atau "kerataan" distribusi dibandingkan dengan distribusi normal.

PIE CHART

Diagram lingkaran adalah grafik statistik melingkar, yang dibagi menjadi irisan-irisan untuk menggambarkan proporsi numerik. Dalam diagram lingkaran, panjang busur setiap irisan (dan akibatnya sudut pusat dan luasnya) sebanding dengan kuantitas yang diwakilinya. Seluruh lingkaran mewakili 100% atau keseluruhan dari data.



Kapan Digunakan:

1. Menunjukkan Komposisi atau Bagian dari Keseluruhan: Paling efektif ketika Anda ingin menunjukkan bagaimana satu kesatuan dibagi menjadi beberapa bagian. Misalnya, persentase pangsa pasar berbagai perusahaan, komposisi anggaran, atau jenis respons dalam survei.

2. Menampilkan Persentase: Sangat intuitif untuk menampilkan proporsi dalam bentuk persentase.

3. Ketika Jumlah Kategori Sedikit: Idealnya digunakan untuk 2 hingga 5-7 kategori. Jika terlalu banyak irisan, diagram menjadi sulit dibaca dan dibandingkan.

4. Ketika Total dari Semua Bagian adalah 100%: Penting bahwa semua irisan jika dijumlahkan membentuk keseluruhan yang bermakna.

5. Menekankan Proporsi yang Signifikan: Dapat dengan cepat menyoroti segmen terbesar atau terkecil dari keseluruhan.

Peringatan untuk Pie Chart: Meskipun populer, diagram lingkaran sering dikritik karena sulit bagi mata manusia untuk secara akurat membandingkan ukuran sudut atau luas area, terutama ketika perbedaannya kecil atau ada banyak irisan. Diagram batang seringkali menjadi alternatif yang lebih baik untuk perbandingan yang akurat.

LATIHAN PRAKTIKUM VISUALISASI DATA

Untuk latihan praktikum visualisasi data dapat diakses pada link berikut:

https://colab.research.google.com/drive/1-wfohZ-VI9PVwFt7F8JwM_BpNpRPAT3q?usp=sharing