Langkah Awal Menjadi Data Analyst

Menguasai Statistika, Spreadsheet, SQL, Programming dan Visualisasi untuk Pemula

Aep Hidayatuloh

20 Oct 2024

Table of contents

# Pengantar

Data Analyst dan Data Scientist menjadi pekerjaan yang diinginkan banyak orang saat ini, terutama bagi lulusan baru dari universitas. Di Jobstreet tersedia **lebih dari 9.000 lowongan pekerjaan dengan kata kunci “data”**, termasuk data analyst, data entry, data scientist dan lain-lain dengan tingakat magang (*internship*), junior, senior hingga manager. Lowongan kerja dengan kata kunci “data analyst” sendiri ada lebih dari 1.000 lowongan pekerjaan.

Mereka yang memang menempuh pendidikan yang berkaitan erat dengan data seperti Statistika dan Ilmu Komputer, atau yang tidak hubungannya sama sekali seperti lulusan Ilmu Hukum, Perikanan dan Kelautan, dan lain-lain yang ingin beralih karir menjadi data analyst atau data scientist mulai banyak bermunculan.

Karena banyaknya keinginan dan kebutuhan akan pendidikan tentang data, banyak universitas atau kampus mulai mempertimbangkan atau bahkan sudah mulai untuk mengampu pendidikan yang berkaitan dengan bidang ini. Bahkan banyak lembaga penyedia pendidikan, training atau kursus yang menydiakan kesempatan bagi mereka yang ingin belajar untuk menjadi Data Analyst atau Data Scientist. Buku ini akan mengajak kamu untuk mulai memahami apa itu data analyst, kemampuan yang dibutuhkan dan disertai dengan contoh kasus dan teknis pengerjaannya yang akan menambah pemahaman dan kemampuan kamu untuk menjadi data analyst.

Alat atau *Tools* yang akan sangat berguna dan biasanya digunakan dalam mengerjakan analisis data adalah Excel, SQL, dan terkadang juga menggunakan bahasa pemrograman. Bahasa Pemrograman yang biasanya digunakan dalam kegiatan analisis data adalah R atau Python. Oleh karena itu, buku ini akan mengajak kamu untuk mengenal lebih jauh tentang data analytics dan *tools* yang biasanya digunakan untuk menunjang aktivitas analisis data.

Tentu saja penulis menyadari bahwa tulisan ini hanya berdasarkan pengetahuan dan pengalaman yang diketahui oleh penulis, shingga masih banyak kekurangan dan perlu perbaikan dalam berbagai hal. Jika kamu ingin berkontribusi pada tulisan ini, kamu dapat menjadi kontributor pada repository GitHub <https://github.com/aephidayatuloh/data-analyst>.

# Pola Pikir Data Analyst

Jangan menganalisis data!

Seiring dengan perkembangan teknologi dan digitalisasi, kebutuhan akan data analyst semakin meningkat. Seorang data analyst memiliki peran penting dalam menganalisis data untuk membantu organisasi membuat keputusan yang lebih baik. Namun, menjadi seorang data analyst yang efektif tidak hanya memerlukan keterampilan teknis tetapi juga pola pikir yang tepat. Bab ini akan membahas pola pikir yang harus dimiliki seorang data analyst untuk berhasil dalam profesinya.

Mejadi seorang data analyst atau data scientist memang perlu untuk menguasai berbagai macam hal, salah satunya yang penting adalah kemampuan untuk dapat menggali informasi tentang masalah apa yang akan dianalisis. Jadi jangan langsung menganalisis data tanpa tahu konteks dan tujuannya.

Beberapa hal yang perlu dilakukan ketika akan memulai kegiatan atau project data analytics adalah membuat pertanyaan tentang apa masalah atau project tersebut? Nilai atau *value* apa yang akan diperoleh setelah mendapatkan hasil? Bagaimana hasil dari analisis data ini akan digunakan nantinya?

## Berpikir Kritis

Bukan pikiran kamu yang dalam kondisi kritis!

Berpikir kritis adalah kemampuan untuk menganalisis dan mengevaluasi informasi secara mendalam sebelum mencapai suatu kesimpulan. Ketika seorang data analyst akan bekerja maka yang pertama kali harus dia lakukan adalah membuat pertanyaan atau hipotesis. Hal ini penting agar langkah yang diambil menjadi terarah dan bukan hanya pekerjaan *receh* yang nantinya akan gagal di tengah jalan atau bahkan hasil analisisnya tidak penting sama sekali. Bagi seorang data analyst, berpikir kritis melibatkan beberapa aspek:

### Mempertanyakan Masalah dan Data

Seorang data analyst harus selalu mempertanyakan sumber data, kualitas data, dan relevansi data dengan masalah yang sedang dianalisis. Mempertanyakan sumber data berarti memahami dari mana data berasal, apakah itu dari survei, sensor, transaksi, atau sumber lainnya. Sumber data yang berbeda memiliki tingkat keandalan yang berbeda pula. Dengan mempertanyakan sumber, seorang data analyst dapat memastikan bahwa data yang digunakan sah dan dapat dipercaya.

Lebih dari itu, kualitas data juga harus dipertanyakan. Kualitas data mencakup aspek seperti akurasi, kelengkapan, dan konsistensi data. Data yang tidak akurat atau tidak lengkap dapat menghasilkan kesimpulan yang salah. Misalnya, jika data penjualan memiliki banyak nilai yang hilang, analisis tren penjualan mungkin tidak akurat. Oleh karena itu, data analyst harus selalu melakukan pemeriksaan kualitas data sebelum melakukan analisis.

Relevansi data juga merupakan aspek penting yang harus dipertanyakan. Tidak semua data yang tersedia relevan untuk analisis tertentu. Seorang data analyst harus bisa menyaring data yang relevan dari data yang tidak relevan. Misalnya, data cuaca mungkin tidak relevan untuk analisis penjualan kecuali ada bukti bahwa cuaca mempengaruhi penjualan. Dengan demikian, mempertanyakan relevansi data membantu dalam fokus pada data yang benar-benar penting untuk analisis.

Mempertanyakan data juga melibatkan penilaian apakah data tersebut sudah cukup untuk menjawab pertanyaan yang diajukan. Terkadang, data yang ada mungkin tidak cukup untuk memberikan wawasan yang diperlukan. Dalam kasus seperti ini, data analyst mungkin perlu mencari sumber data tambahan atau menggunakan metode analisis yang lebih canggih untuk mengisi kekosongan data. Hal ini menunjukkan pentingnya fleksibilitas dan kreativitas dalam berpikir kritis.

Beberapa pertanyaan lain yang perlu diajukan diantaranya adalah:

1. Apa masalah yang ingin diselesaikan atau dianalisis dan apakah masalah ini penting untuk diselesaikan?
2. Siapa yang berkepentingan dan memiliki data untuk masalah ini?
3. Darimana dan bagaimana saya bisa mendapatkan data?
4. Kapan analisis data ini harus selesai?
5. Bagaimana hasil analisis data ini akan digunakan nantinya?

Terakhir, mempertanyakan masalah yang dikerjakan dan data yang digunakan juga berarti memahami batasan-batasan data. Setiap data memiliki keterbatasan, seperti cakupan waktu yang terbatas, sampel yang tidak representatif, atau bias dalam pengumpulan data. Memahami batasan-batasan ini membantu data analyst dalam menginterpretasikan hasil analisis dengan lebih hati-hati dan membuat rekomendasi yang lebih akurat.

### Mengidentifikasi Bias

Penting untuk menyadari bias yang mungkin ada dalam data dan bagaimana bias tersebut dapat mempengaruhi hasil analisis. Bias dalam data bisa datang dari berbagai sumber, seperti metode pengumpulan data, desain survei, atau bahkan dari cara data diinterpretasikan. Seorang data analyst harus mampu mengidentifikasi dan mengurangi bias ini untuk memastikan hasil analisis yang lebih objektif dan akurat.

Misalnya, bias seleksi dapat terjadi ketika sampel data tidak mewakili populasi yang lebih besar. Jika sebuah survei hanya mengumpulkan responden dari kota-kota besar, hasilnya mungkin tidak berlaku untuk daerah pedesaan. Data analyst perlu menyadari potensi bias ini dan, jika mungkin, memperbaikinya dengan menyeimbangkan sampel atau menggunakan teknik statistik yang tepat.

Bias dalam pengumpulan data juga bisa muncul dari cara pertanyaan disusun dalam survei atau dari alat yang digunakan untuk mengumpulkan data. Pertanyaan yang leading atau tidak jelas bisa mempengaruhi jawaban responden. Sebagai contoh, pertanyaan “Seberapa sering kamu menikmati bekerja di lingkungan yang dinamis?” sudah menyiratkan bahwa lingkungan yang dinamis adalah sesuatu yang positif, sehingga bisa mengarahkan responden untuk memberikan jawaban yang lebih positif. Mengidentifikasi bias semacam ini sangat penting dalam menjaga integritas data.

Bias interpretasi adalah jenis bias yang muncul saat data analyst menafsirkan hasil analisis. Setiap data analyst membawa perspektif dan asumsi mereka sendiri ke dalam proses analisis, yang bisa mempengaruhi cara mereka melihat data. Oleh karena itu, penting untuk selalu kritis terhadap interpretasi hasil dan, jika memungkinkan, mendapatkan opini kedua atau melakukan peer review untuk mengurangi bias interpretasi.

Mengatasi bias dalam data juga melibatkan penggunaan teknik statistik yang tepat, seperti penyesuaian untuk faktor pengganggu atau pengacakan dalam pengumpulan data. Teknik-teknik ini dapat membantu mengurangi efek bias dan memberikan hasil yang lebih andal. Data analyst harus terus belajar dan menerapkan metode terbaru untuk mengidentifikasi dan mengurangi bias dalam analisis mereka.

Akhirnya, transparansi dalam proses analisis adalah kunci dalam mengidentifikasi bias. Mendokumentasikan setiap langkah yang diambil dan alasan di balik setiap keputusan analisis membantu dalam mengidentifikasi potensi bias. Transparansi ini juga memudahkan orang lain untuk meninjau dan memahami analisis, serta memberikan umpan balik yang konstruktif.

Menghindari Asumsi yang Tidak Berdasar Setiap analisis harus didasarkan pada data yang kuat dan valid, bukan pada asumsi atau intuisi semata. Menghindari asumsi yang tidak berdasar berarti selalu mencari bukti yang mendukung setiap hipotesis atau kesimpulan yang diambil. Seorang data analyst harus mampu membedakan antara asumsi yang valid dan yang tidak.

Misalnya, asumsi bahwa peningkatan anggaran pemasaran akan selalu meningkatkan penjualan mungkin tidak selalu benar. Ada banyak faktor lain yang bisa mempengaruhi penjualan, seperti kondisi ekonomi, perubahan dalam preferensi konsumen, atau aktivitas pesaing. Data analyst harus menguji asumsi ini dengan data historis atau melalui eksperimen untuk memastikan kebenarannya.

Menghindari asumsi yang tidak berdasar juga melibatkan verifikasi data dari berbagai sumber. Misalnya, jika data penjualan menunjukkan peningkatan tajam, data analyst harus memverifikasi informasi tersebut dengan data dari sumber lain, seperti laporan keuangan atau data dari pihak ketiga. Ini membantu memastikan bahwa peningkatan tersebut bukan hasil dari kesalahan atau anomali dalam data.

Asumsi yang tidak berdasar juga bisa muncul dari generalisasi berlebihan. Sebagai contoh, hanya karena sebuah kampanye pemasaran berhasil di satu wilayah, tidak berarti kampanye yang sama akan berhasil di wilayah lain. Data analyst harus selalu menguji generalisasi ini dengan data tambahan dan mempertimbangkan konteks yang berbeda sebelum membuat kesimpulan.

Menghindari asumsi yang tidak berdasar juga berarti selalu menggunakan metode statistik yang tepat untuk menguji hipotesis. Misalnya, menggunakan uji signifikan untuk menentukan apakah perbedaan antara dua kelompok adalah signifikan atau hanya kebetulan. Data analyst harus memiliki pemahaman yang kuat tentang statistik dan metode analisis untuk membuat keputusan yang didasarkan pada data.

Akhirnya, penting untuk selalu siap untuk merevisi asumsi dan hipotesis berdasarkan temuan baru. Data analysis adalah proses iteratif, dan data analyst harus selalu terbuka untuk perubahan dan pembaruan dalam asumsi mereka. Dengan demikian, mereka dapat memastikan bahwa analisis mereka selalu didasarkan pada data yang paling akurat dan relevan.

## Berorientasi pada Nilai dan Detail

Tak jarang data analyst lebih berfokus pada metodologi atau teknologi. Metodologi terkini apa yang akan digunakan? Teknologi terbaru apa yang akan diimplementasikan? Dashboard seperti apa yang akan dibuat?

Agar pekerjaan analisis data yang kita lakukan memiliki manfaat maka dari awal kita harus selalu berorientasi pada nilai (*value*) yang akan dihasilkan nantinya. *Value* ini bisa kita lihat berdasarkan pertanyaan-pertanyaan di atas. Jika kita mendapatkan jawaban yang “memuaskan” untuk kita lakukan maka analisis data akan menjadi cara yang efektif untuk mendapatkan hasil.

Seorang data analyst juga harus memiliki perhatian yang tinggi terhadap detail. Kesalahan kecil dalam data atau analisis dapat menghasilkan kesimpulan yang salah dan keputusan yang buruk. Beberapa praktik yang mendukung orientasi pada detail meliputi:

### Validasi Data

Memastikan bahwa data yang digunakan akurat dan lengkap melalui berbagai metode validasi. Validasi data adalah langkah pertama yang sangat penting dalam setiap analisis data. Data yang tidak valid dapat mengarah pada kesimpulan yang salah dan keputusan yang merugikan. Oleh karena itu, data analyst harus selalu memeriksa dan membersihkan data sebelum melakukan analisis lebih lanjut.

Validasi data melibatkan beberapa langkah, termasuk pengecekan duplikasi, pengecekan konsistensi, dan pengecekan anomali. Misalnya, jika sebuah dataset berisi data transaksi, data analyst harus memeriksa apakah ada transaksi yang tercatat lebih dari sekali atau apakah ada transaksi yang hilang. Selain itu, data yang tidak konsisten, seperti tanggal yang tidak sesuai atau nilai yang tidak logis, harus diperbaiki atau dihapus.

Selain pengecekan internal, validasi data juga dapat melibatkan verifikasi terhadap sumber eksternal. Misalnya, data penjualan dapat divalidasi dengan laporan keuangan perusahaan atau data dari pihak ketiga. Verifikasi eksternal membantu memastikan bahwa data yang digunakan akurat dan dapat dipercaya. Dalam beberapa kasus, validasi data juga dapat melibatkan interaksi dengan tim lain atau ahli domain untuk memastikan bahwa data yang digunakan sesuai dengan konteks dan tujuan analisis.

Automasi dalam validasi data juga sangat membantu. Menggunakan alat dan skrip untuk otomatisasi pengecekan data dapat menghemat waktu dan mengurangi kesalahan manusia. Misalnya, menggunakan alat ETL (Extract, Transform, Load) untuk memvalidasi dan membersihkan data secara otomatis sebelum dimasukkan ke dalam sistem analisis. Dengan demikian, data analyst dapat fokus pada analisis data yang lebih mendalam daripada tugas-tugas administratif.

Penting juga untuk mendokumentasikan proses validasi data. Dokumentasi ini tidak hanya membantu dalam memastikan bahwa proses validasi dilakukan dengan benar, tetapi juga menyediakan catatan yang dapat ditinjau oleh pihak lain. Dokumentasi yang baik juga memudahkan untuk mereplikasi analisis di masa depan atau melakukan audit terhadap proses analisis.

### Pengecekan Ganda

Selalu melakukan pengecekan ganda pada setiap langkah analisis untuk menghindari kesalahan. Kesalahan kecil dalam pengolahan data dapat memiliki dampak besar pada hasil akhir. Oleh karena itu, data analyst harus selalu memeriksa ulang setiap langkah dalam proses analisis mereka. Ini termasuk pengecekan kode, pengecekan formula dalam spreadsheet, dan pengecekan hasil visualisasi.

Pengecekan ganda bisa melibatkan berbagai teknik. Salah satunya adalah cross-validation, di mana hasil dari satu metode analisis dibandingkan dengan hasil dari metode lain. Misalnya, jika menggunakan dua metode statistik yang berbeda untuk menganalisis data, hasil yang konsisten dari kedua metode tersebut memberikan keyakinan lebih tinggi bahwa hasilnya akurat. Jika ada perbedaan signifikan antara hasil dari dua metode, data analyst harus menyelidiki lebih lanjut untuk menemukan penyebabnya.

Peer review juga merupakan bagian penting dari pengecekan ganda. Meminta rekan kerja atau anggota tim lain untuk meninjau pekerjaan dapat membantu mengidentifikasi kesalahan yang mungkin terlewatkan. Peer review juga memberikan perspektif baru yang mungkin tidak terpikirkan sebelumnya, membantu memastikan bahwa analisis lebih komprehensif dan bebas dari bias.

Automasi juga bisa membantu dalam pengecekan ganda. Misalnya, menggunakan alat otomatis untuk menjalankan tes unit pada kode analisis atau memeriksa konsistensi data. Automasi ini tidak hanya meningkatkan efisiensi tetapi juga mengurangi risiko kesalahan manusia. Dengan adanya alat otomatis, data analyst dapat dengan cepat mendeteksi dan memperbaiki kesalahan sebelum melanjutkan ke tahap analisis berikutnya.

Akhirnya, penting untuk selalu mengingat bahwa pengecekan ganda bukan hanya tentang mencari kesalahan, tetapi juga tentang memastikan kualitas dan integritas analisis. Proses ini membantu memastikan bahwa setiap langkah analisis dilakukan dengan benar dan hasilnya dapat dipercaya. Dengan demikian, data analyst dapat memberikan hasil analisis yang lebih akurat dan berguna untuk pengambilan keputusan.

### Dokumentasi

Mendokumentasikan setiap langkah proses analisis untuk memastikan transparansi dan kemudahan pengulangan. Dokumentasi adalah aspek penting dari setiap proses analisis data. Dengan mendokumentasikan setiap langkah, data analyst dapat memastikan bahwa proses mereka transparan dan dapat diulang di masa depan. Dokumentasi yang baik juga memudahkan orang lain untuk memahami dan meninjau analisis yang dilakukan.

Dokumentasi harus mencakup berbagai elemen, termasuk sumber data yang digunakan, metode analisis yang diterapkan, asumsi yang dibuat, dan hasil yang diperoleh. Misalnya, jika data analyst menggunakan data penjualan dari tahun-tahun sebelumnya, dokumentasi harus mencatat sumber data tersebut, periode waktu yang dicakup, dan setiap penyesuaian atau pembersihan data yang dilakukan. Ini membantu memastikan bahwa analisis dapat diverifikasi dan direplikasi oleh orang lain.

Selain itu, dokumentasi harus mencakup rincian teknis tentang alat dan teknik yang digunakan dalam analisis. Misalnya, jika menggunakan software statistik tertentu atau algoritma machine learning, dokumentasi harus mencakup versi software yang digunakan, parameter yang diterapkan, dan alasan di balik pemilihan metode tersebut. Ini membantu memastikan bahwa orang lain yang ingin mengulangi atau memverifikasi analisis memiliki semua informasi yang diperlukan.

Dokumentasi yang baik juga mencakup interpretasi hasil dan implikasi dari analisis tersebut. Misalnya, jika analisis menunjukkan bahwa ada hubungan signifikan antara dua variabel, dokumentasi harus menjelaskan apa arti hubungan tersebut dan bagaimana hal itu dapat mempengaruhi keputusan bisnis. Ini membantu memastikan bahwa hasil analisis dapat dipahami dan digunakan dengan benar oleh pihak yang berkepentingan.

Automasi dalam dokumentasi juga sangat membantu. Menggunakan alat otomatis untuk menghasilkan laporan analisis atau mendokumentasikan proses analisis dapat menghemat waktu dan mengurangi risiko kesalahan. Misalnya, menggunakan notebook interaktif seperti Jupyter untuk mendokumentasikan dan menjalankan kode analisis secara bersamaan. Dengan demikian, data analyst dapat dengan mudah membagikan dan mereplikasi analisis mereka.

Akhirnya, penting untuk selalu memperbarui dokumentasi sesuai dengan perubahan atau pembaruan dalam proses analisis. Dokumentasi yang tidak diperbarui dapat menjadi usang dan menyesatkan. Dengan menjaga dokumentasi yang selalu mutakhir, data analyst dapat memastikan bahwa proses analisis mereka tetap relevan dan akurat.

## Analytical Thinking

*Analytical thinking* adalah cara untuk mengidentifikasi dan mendefinisikan sebuah masalah dan menyelesaikannya menggunakan data dengan terorganisir dan tahap demi tahap. *Analytical thinking* atau pemikiran analitis adalah keterampilan krusial yang membedakan data analyst yang efektif dari yang kurang efektif. Dalam konteks pekerjaan seorang data analyst, pemikiran analitis melibatkan kemampuan untuk memecahkan masalah kompleks, menginterpretasikan data, dan membuat keputusan berbasis data yang solid. Berikut adalah beberapa alasan mengapa analytical thinking sangat penting bagi seorang data analyst:

### Memecahkan Masalah Kompleks

*Analytical thinking* memungkinkan data analyst untuk memecahkan masalah kompleks dengan pendekatan yang terstruktur dan logis. Dalam dunia data, masalah sering kali tidak memiliki jawaban yang jelas atau langsung, sehingga memerlukan analisis mendalam untuk menemukan solusi. Kemampuan untuk mengidentifikasi inti masalah, membagi masalah menjadi komponen-komponen yang lebih kecil, dan mengembangkan solusi yang efektif adalah kunci untuk mencapai hasil yang bermanfaat.

Identifikasi Masalah: Data analyst harus mampu mengidentifikasi masalah yang mendasari data yang ada. Ini melibatkan pemahaman konteks bisnis dan mengenali pola atau anomali yang mungkin menunjukkan adanya masalah. Penyelesaian Masalah: Setelah masalah diidentifikasi, analytical thinking membantu dalam merancang dan menerapkan strategi untuk menyelesaikan masalah tersebut. Ini sering melibatkan penggunaan teknik statistik, machine learning, dan metode analisis lainnya.

### Menginterpretasikan Data dengan Akurat

Kemampuan untuk menginterpretasikan data secara akurat adalah elemen penting dari *analytical thinking*. Data yang besar dan kompleks dapat memiliki berbagai makna tergantung pada konteksnya. Seorang data analyst harus mampu menguraikan data dan memberikan wawasan yang relevan berdasarkan interpretasi yang tepat.

Analisis Pola dan Tren: Data analyst harus mampu menemukan pola, tren, dan hubungan dalam data. Ini melibatkan penggunaan teknik eksplorasi data untuk mengidentifikasi informasi yang signifikan dan relevan.

Evaluasi Hasil: *Analytical thinking* memungkinkan data analyst untuk mengevaluasi hasil analisis dan memverifikasi akurasi temuan. Ini membantu dalam memastikan bahwa keputusan yang diambil didasarkan pada data yang valid dan dapat diandalkan.

### Membuat Keputusan Berbasis Data

Salah satu peran utama seorang data analyst adalah memberikan rekomendasi yang berdasarkan pada data. *Analytical thinking* memungkinkan data analyst untuk membuat keputusan yang terinformasi dengan menggunakan data sebagai dasar utama.

Evaluasi Opsi: Data analyst harus dapat mengevaluasi berbagai opsi dan strategi berdasarkan data yang tersedia. Ini melibatkan analisis pro dan kontra dari setiap opsi serta dampaknya terhadap tujuan bisnis.

Pengambilan Keputusan: Dengan menggunakan data untuk mendukung keputusan, data analyst dapat membantu organisasi dalam membuat keputusan yang lebih baik dan lebih terukur. *Analytical thinking* memastikan bahwa keputusan yang diambil bukan hanya berdasarkan intuisi, tetapi juga pada analisis data yang mendalam.

### Meningkatkan Proses dan Efisiensi

*Analytical thinking* juga berperan dalam meningkatkan proses dan efisiensi operasional. Dengan menggunakan pendekatan analitis, data analyst dapat mengidentifikasi area untuk perbaikan dan mengembangkan solusi yang meningkatkan efisiensi.

Optimalisasi Proses: Data analyst dapat menganalisis proses yang ada dan mengidentifikasi kekurangan atau hambatan. Dengan pemikiran analitis, mereka dapat merancang perbaikan yang dapat meningkatkan efisiensi dan produktivitas. Evaluasi Kinerja: Kemampuan untuk menganalisis data kinerja membantu dalam mengevaluasi efektivitas strategi dan inisiatif yang ada. Ini memungkinkan data analyst untuk memberikan rekomendasi berbasis data untuk peningkatan berkelanjutan.

### Menghadapi Ketidakpastian dan Risiko

Dalam dunia yang tidak pasti, analytical thinking membantu data analyst dalam menghadapi ketidakpastian dan risiko. Dengan menggunakan metode analisis yang tepat, mereka dapat mengukur risiko dan membuat keputusan yang lebih baik dalam situasi yang tidak pasti.

Penilaian Risiko: Data analyst harus mampu menilai risiko yang terkait dengan keputusan dan tindakan tertentu. Ini melibatkan penggunaan teknik statistik dan model prediktif untuk memahami potensi risiko dan dampaknya. Strategi Mitigasi: Analytical thinking membantu dalam merancang strategi mitigasi untuk mengurangi risiko dan menangani ketidakpastian. Dengan memahami data dan potensi risiko, data analyst dapat mengembangkan rencana yang efektif untuk mengelola risiko.

### Beradaptasi dengan Perubahan

Dalam lingkungan yang cepat berubah, analytical thinking memungkinkan data analyst untuk beradaptasi dengan perubahan dan tetap relevan. Kemampuan untuk menganalisis dan memahami data yang baru dan berubah membantu dalam menyesuaikan strategi dan keputusan sesuai dengan kebutuhan yang berkembang.

Responsif terhadap Perubahan: Data analyst harus cepat dalam menganalisis data baru dan mengidentifikasi perubahan yang terjadi. Ini memungkinkan mereka untuk memberikan wawasan yang up-to-date dan relevan. Inovasi dan Adaptasi: Analytical thinking mendukung inovasi dengan memungkinkan data analyst untuk mengeksplorasi metode baru dan ide-ide segar berdasarkan analisis data terbaru.

Dengan kemampuan analytical thinking yang kuat, seorang data analyst dapat mengatasi tantangan yang kompleks, memberikan wawasan yang berharga, dan mendukung pengambilan keputusan berbasis data yang efektif. Ini adalah keterampilan yang mendasar untuk mencapai keberhasilan dalam peran data analyst dan memberikan kontribusi yang signifikan terhadap pencapaian tujuan organisasi.

## Curiosity dan Keinginan Belajar yang Tinggi

Dunia data analytics terus berkembang dengan cepat, dengan alat, teknik, dan metode baru yang muncul secara reguler. Seorang data analyst yang baik harus memiliki rasa ingin tahu yang tinggi dan keinginan untuk terus belajar. Ini termasuk:

### Mengikuti Perkembangan Teknologi

Selalu up-to-date dengan perkembangan terbaru dalam bidang data analytics. Teknologi dan alat analisis data terus berkembang, dan seorang data analyst harus selalu mengikuti perkembangan ini untuk tetap relevan dan efektif dalam pekerjaannya. Mengikuti perkembangan teknologi berarti memahami dan menguasai alat dan teknik baru yang muncul di pasar.

Salah satu cara untuk tetap up-to-date adalah dengan mengikuti kursus online, seminar, atau konferensi yang berfokus pada data analytics. Platform pembelajaran online seperti Coursera, Udacity, dan edX menawarkan kursus yang mencakup berbagai topik dalam analisis data, mulai dari dasar-dasar statistik hingga teknik machine learning yang canggih. Mengikuti kursus ini tidak hanya membantu dalam memperbarui pengetahuan tetapi juga dalam mendapatkan sertifikasi yang diakui di industri.

Membaca jurnal ilmiah dan publikasi industri juga merupakan cara yang efektif untuk mengikuti perkembangan teknologi. Jurnal seperti Journal of Data Science, Journal of Machine Learning Research, dan publikasi dari perusahaan teknologi besar seperti Google dan IBM sering kali mempublikasikan penelitian terbaru dan inovasi dalam analisis data. Dengan membaca jurnal ini, data analyst dapat tetap terinformasi tentang tren dan teknik terbaru.

Mengikuti komunitas online dan forum diskusi juga sangat membantu. Platform seperti Stack Overflow, Reddit, dan LinkedIn memiliki komunitas aktif dari data analyst yang sering berbagi pengetahuan, pengalaman, dan solusi untuk masalah analisis data. Bergabung dengan komunitas ini memungkinkan data analyst untuk belajar dari orang lain, mendapatkan umpan balik, dan tetap terhubung dengan perkembangan terbaru di industri.

Selain itu, mengembangkan proyek pribadi atau kontribusi ke proyek open-source adalah cara lain untuk mengikuti perkembangan teknologi. Dengan mengerjakan proyek nyata, data analyst dapat menguji dan menerapkan alat dan teknik baru, serta mendapatkan pengalaman praktis yang berharga. Proyek open-source juga memberikan kesempatan untuk berkolaborasi dengan data analyst lain dari seluruh dunia dan belajar dari kontribusi mereka.

Akhirnya, penting untuk selalu memiliki mindset pembelajar yang aktif dan terbuka terhadap perubahan. Teknologi terus berkembang, dan data analyst harus siap untuk terus belajar dan beradaptasi dengan alat dan teknik baru. Dengan memiliki keinginan yang kuat untuk belajar dan berkembang, data analyst dapat memastikan bahwa mereka selalu berada di garis depan dalam industri yang cepat berubah ini.

### Eksperimen

Tidak takut untuk mencoba pendekatan baru dan bereksperimen dengan data untuk menemukan wawasan baru. Eksperimen adalah inti dari analisis data yang efektif. Dengan mencoba pendekatan baru, data analyst dapat menemukan wawasan yang tidak terduga dan mengembangkan solusi inovatif untuk masalah yang kompleks. Eksperimen juga membantu dalam menguji hipotesis dan memvalidasi asumsi yang dibuat selama analisis.

Eksperimen dalam analisis data melibatkan penggunaan berbagai teknik dan alat untuk menganalisis data dari berbagai perspektif. Misalnya, data analyst dapat menggunakan teknik machine learning untuk memprediksi tren masa depan atau menggunakan analisis klaster untuk mengidentifikasi pola dalam data. Dengan mencoba berbagai teknik, data analyst dapat menemukan metode yang paling efektif untuk menganalisis data tertentu.

Eksperimen juga melibatkan penggunaan data dari berbagai sumber dan format. Data analyst harus tidak takut untuk menggabungkan data dari berbagai sumber, seperti data internal perusahaan, data eksternal dari pihak ketiga, atau data publik dari internet. Menggabungkan data dari berbagai sumber dapat memberikan wawasan yang lebih komprehensif dan mendalam. Misalnya, menggabungkan data penjualan dengan data demografis dan data cuaca dapat memberikan wawasan tentang bagaimana faktor-faktor ini mempengaruhi penjualan.

Mengembangkan dan menguji model prediktif adalah bagian penting dari eksperimen dalam analisis data. Data analyst harus tidak takut untuk membangun dan menguji berbagai model prediktif untuk menemukan model yang paling akurat dan andal. Ini melibatkan penggunaan teknik seperti regresi, decision tree, random forest, dan neural network. Dengan menguji berbagai model, data analyst dapat memilih model yang paling sesuai dengan data dan tujuan analisis.

Eksperimen juga melibatkan penggunaan teknik visualisasi data untuk menemukan pola dan tren dalam data. Data analyst dapat menggunakan alat visualisasi seperti Tableau, Power BI, atau D3.js untuk membuat visualisasi yang menarik dan informatif. Visualisasi ini tidak hanya membantu dalam memahami data tetapi juga dalam menyampaikan hasil analisis kepada pihak yang berkepentingan. Dengan mencoba berbagai teknik visualisasi, data analyst dapat menemukan cara terbaik untuk menyajikan data.

Akhirnya, penting untuk selalu mendokumentasikan dan mengevaluasi hasil eksperimen. Dokumentasi membantu dalam melacak metode yang digunakan, hasil yang diperoleh, dan pembelajaran dari eksperimen. Evaluasi hasil eksperimen membantu dalam memahami apa yang berhasil dan apa yang tidak, serta dalam mengidentifikasi area yang perlu diperbaiki. Dengan mendokumentasikan dan mengevaluasi eksperimen, data analyst dapat terus belajar dan berkembang dalam analisis data.

### Belajar dari Kesalahan

Menganggap kesalahan sebagai kesempatan untuk belajar dan berkembang. Kesalahan adalah bagian tak terhindarkan dari proses analisis data. Data analyst harus menganggap kesalahan sebagai kesempatan untuk belajar dan berkembang, bukan sebagai kegagalan. Dengan memahami dan menganalisis kesalahan, data analyst dapat mengidentifikasi area yang perlu diperbaiki dan menghindari kesalahan yang sama di masa depan.

Belajar dari kesalahan melibatkan refleksi diri dan evaluasi kritis terhadap proses analisis. Data analyst harus selalu bertanya pada diri sendiri apa yang salah, mengapa itu terjadi, dan bagaimana mencegahnya di masa depan. Misalnya, jika hasil analisis menunjukkan anomali yang tidak terduga, data analyst harus menyelidiki apakah itu disebabkan oleh kesalahan dalam pengumpulan data, pembersihan data, atau metode analisis yang digunakan. Dengan memahami penyebab kesalahan, data analyst dapat memperbaiki proses mereka dan meningkatkan kualitas analisis di masa depan.

Mendokumentasikan kesalahan juga sangat penting. Dokumentasi kesalahan membantu dalam melacak masalah yang terjadi dan solusi yang diterapkan. Ini tidak hanya membantu data analyst untuk belajar dari kesalahan mereka sendiri tetapi juga memungkinkan orang lain untuk belajar dari kesalahan tersebut. Misalnya, jika ada kesalahan dalam kode analisis, dokumentasi tentang kesalahan dan perbaikan yang diterapkan dapat membantu anggota tim lain dalam menghindari kesalahan yang sama.

Belajar dari kesalahan juga melibatkan mendapatkan umpan balik dari rekan kerja atau mentor. Umpan balik yang konstruktif membantu data analyst untuk melihat kesalahan dari perspektif yang berbeda dan mengidentifikasi area yang perlu diperbaiki. Misalnya, meminta rekan kerja untuk meninjau hasil analisis atau kode dapat membantu dalam mengidentifikasi kesalahan yang mungkin terlewatkan. Umpan balik juga membantu dalam memahami bagaimana memperbaiki kesalahan dan meningkatkan proses analisis.

Selain itu, penting untuk mengembangkan mindset yang tahan terhadap kegagalan. Kegagalan adalah bagian dari proses pembelajaran, dan data analyst harus tidak takut untuk gagal. Dengan mengembangkan mindset yang tahan terhadap kegagalan, data analyst dapat lebih berani dalam mencoba pendekatan baru dan bereksperimen dengan data. Ini membantu dalam menemukan solusi yang lebih inovatif dan efektif.

Akhirnya, penting untuk selalu merayakan pembelajaran dari kesalahan. Setiap kesalahan yang diatasi adalah langkah maju dalam proses pembelajaran dan perkembangan. Dengan menganggap kesalahan sebagai kesempatan untuk belajar dan berkembang, data analyst dapat terus meningkatkan keterampilan dan kemampuan mereka dalam analisis data.

## Kemampuan Komunikasi yang Efektif

Mampu menganalisis data dengan baik tidak akan berguna jika hasil analisis tersebut tidak dapat dikomunikasikan dengan jelas kepada pihak yang berkepentingan. Kemampuan komunikasi yang efektif adalah salah satu aspek kunci yang harus dimiliki oleh seorang data analyst. Tanpa kemampuan untuk menyampaikan hasil analisis dengan jelas dan persuasif, wawasan yang berharga bisa tidak dimanfaatkan sepenuhnya oleh pemangku kepentingan. Berikut adalah beberapa elemen penting dari komunikasi yang efektif.

### Menyampaikan Hasil dengan Jelas

Seorang data analyst harus mampu menyampaikan hasil analisis dalam bahasa yang mudah dimengerti oleh orang-orang yang mungkin tidak memiliki latar belakang teknis. Ini berarti menggunakan istilah yang sederhana dan menghindari jargon teknis yang dapat membingungkan. Misalnya, alih-alih mengatakan “model regresi linier menunjukkan koefisien beta sebesar 0,85 dengan R-squared 0,75,” lebih baik mengatakan “analisis kami menunjukkan bahwa ada hubungan kuat antara anggaran pemasaran dan peningkatan penjualan.”

Untuk memastikan pesan tersampaikan dengan jelas, data analyst juga harus mengorganisasi informasi dengan cara yang logis dan sistematis. Ini bisa melibatkan penyusunan laporan dengan pengantar yang menjelaskan konteks, metodologi yang digunakan, hasil yang diperoleh, dan kesimpulan yang diambil. Struktur yang jelas ini membantu audiens mengikuti alur pemikiran dan memahami argumen yang disampaikan.

Visualisasi data adalah alat penting dalam menyampaikan hasil analisis. Grafik, diagram, dan peta dapat membantu mengilustrasikan temuan dengan lebih efektif dibandingkan teks saja. Data analyst harus mahir dalam menggunakan alat visualisasi data seperti Tableau, Power BI, atau bahkan tools yang lebih sederhana seperti Excel untuk membuat presentasi data yang menarik dan mudah dipahami. Misalnya, menggunakan grafik batang untuk menunjukkan perbandingan penjualan antar wilayah atau peta panas untuk menunjukkan distribusi pelanggan di suatu area.

### Mendengarkan Aktif

Kemampuan komunikasi yang efektif tidak hanya tentang berbicara atau menulis dengan jelas tetapi juga tentang mendengarkan secara aktif. Data analyst harus mendengarkan kebutuhan dan masalah dari pemangku kepentingan untuk memahami konteks dan tujuan dari analisis yang diminta. Dengan mendengarkan secara aktif, data analyst dapat mengidentifikasi pertanyaan kunci yang perlu dijawab dan mengumpulkan informasi yang relevan untuk analisis.

Mendengarkan aktif melibatkan memberikan perhatian penuh pada pembicara, mengajukan pertanyaan klarifikasi, dan memberikan umpan balik yang tepat. Misalnya, ketika berinteraksi dengan tim pemasaran, data analyst harus mendengarkan dengan seksama untuk memahami tantangan yang mereka hadapi dan tujuan yang ingin dicapai. Mengajukan pertanyaan seperti “Bisakah kamu menjelaskan lebih lanjut tentang bagaimana kamu mengukur keberhasilan kampanye pemasaran?” dapat membantu mengumpulkan informasi yang lebih mendalam dan relevan untuk analisis.

Mendengarkan aktif juga membantu dalam membangun hubungan kerja yang baik dengan pemangku kepentingan. Dengan menunjukkan minat yang tulus pada masalah dan kebutuhan mereka, data analyst dapat membangun kepercayaan dan kolaborasi yang lebih baik. Ini pada gilirannya membantu dalam mendapatkan dukungan dan partisipasi yang lebih besar dalam proses analisis data.

### Menyesuaikan Pesan untuk Berbagai Audiens

Setiap audiens memiliki kebutuhan dan tingkat pemahaman yang berbeda. Data analyst harus mampu menyesuaikan pesan mereka agar sesuai dengan audiens yang berbeda. Misalnya, presentasi untuk tim manajemen mungkin memerlukan fokus pada implikasi bisnis dan strategi, sementara presentasi untuk tim teknis mungkin memerlukan penjelasan lebih rinci tentang metodologi dan teknik analisis yang digunakan.

Menyesuaikan pesan juga melibatkan memilih format komunikasi yang tepat. Misalnya, laporan tertulis mungkin lebih cocok untuk audiens yang membutuhkan dokumentasi yang rinci dan referensi yang dapat diakses di kemudian hari. Sementara itu, presentasi lisan atau visual mungkin lebih efektif untuk audiens yang lebih menyukai penjelasan langsung dan interaktif. Data analyst harus fleksibel dan kreatif dalam memilih format komunikasi yang paling sesuai untuk audiens mereka.

Selain itu, penting untuk selalu mempertimbangkan konteks budaya dan organisasi dari audiens. Misalnya, dalam organisasi dengan budaya formal, laporan yang rinci dan terstruktur mungkin lebih dihargai. Sebaliknya, dalam organisasi yang lebih dinamis dan informal, presentasi yang singkat dan to the point mungkin lebih efektif. Memahami konteks ini membantu data analyst dalam menyampaikan pesan dengan cara yang paling efektif.

### Mengkomunikasikan Keterbatasan dan Asumsi

Setiap analisis data memiliki keterbatasan dan asumsi yang mendasarinya. Data analyst harus transparan tentang keterbatasan ini dan mengkomunikasikannya dengan jelas kepada pemangku kepentingan. Ini membantu dalam mengelola ekspektasi dan memastikan bahwa keputusan yang diambil berdasarkan hasil analisis didasarkan pada pemahaman yang benar tentang data dan metode yang digunakan.

Mengkomunikasikan keterbatasan dan asumsi juga membantu dalam meningkatkan kredibilitas dan kepercayaan terhadap analisis. Misalnya, jika analisis penjualan menunjukkan peningkatan yang signifikan tetapi data hanya mencakup periode waktu tertentu atau wilayah tertentu, data analyst harus menjelaskan keterbatasan ini. Dengan demikian, pemangku kepentingan dapat membuat keputusan yang lebih informasional dan tidak terkejut jika hasil tidak sesuai dengan ekspektasi di masa mendatang.

Selain itu, penting untuk mengkomunikasikan potensi dampak dari keterbatasan dan asumsi tersebut. Misalnya, jika asumsi tertentu tidak terpenuhi, bagaimana hal itu bisa mempengaruhi hasil analisis? Dengan menjelaskan potensi dampak ini, data analyst membantu pemangku kepentingan dalam memahami risiko dan ketidakpastian yang terkait dengan hasil analisis.

### Berkolaborasi dengan Tim Lain

Kemampuan komunikasi yang efektif juga melibatkan kolaborasi dengan tim lain dalam organisasi. Data analyst sering kali perlu bekerja sama dengan tim pemasaran, keuangan, operasi, dan lain-lain untuk mengumpulkan data, memahami konteks bisnis, dan menyampaikan hasil analisis. Kolaborasi yang baik memerlukan komunikasi yang terbuka dan transparan serta kemampuan untuk bekerja dalam tim lintas fungsi.

Kolaborasi dengan tim lain juga membantu dalam mengidentifikasi kebutuhan dan prioritas yang berbeda dari setiap tim. Misalnya, tim pemasaran mungkin lebih tertarik pada analisis yang dapat membantu mereka meningkatkan kampanye pemasaran, sementara tim keuangan mungkin lebih fokus pada analisis yang dapat membantu mereka mengelola anggaran dan biaya. Dengan berkomunikasi dan berkolaborasi dengan tim lain, data analyst dapat memastikan bahwa analisis yang dilakukan relevan dan bermanfaat bagi seluruh organisasi.

Selain itu, kolaborasi dengan tim lain juga membantu dalam mengembangkan solusi yang lebih komprehensif dan efektif. Misalnya, bekerja sama dengan tim IT dapat membantu dalam mengembangkan sistem pengumpulan data yang lebih efisien, sementara kolaborasi dengan tim operasional dapat membantu dalam mengidentifikasi masalah dan peluang dalam proses bisnis. Dengan bekerja sama dan berkomunikasi dengan tim lain, data analyst dapat memberikan kontribusi yang lebih besar terhadap kesuksesan organisasi.

## Pemecahan Masalah dan Pengambilan Keputusan

Kemampuan untuk memecahkan masalah dan membuat keputusan berdasarkan data adalah keterampilan yang sangat penting bagi seorang data analyst. Ini melibatkan beberapa aspek:

### Identifikasi Masalah

Langkah pertama dalam pemecahan masalah adalah mengidentifikasi masalah dengan jelas. Ini melibatkan pemahaman yang mendalam tentang konteks bisnis dan tujuan yang ingin dicapai. Data analyst harus mampu mengidentifikasi masalah utama yang perlu dipecahkan dan merumuskan pertanyaan yang tepat untuk analisis.

Mengidentifikasi masalah memerlukan keterampilan observasi dan analisis yang kuat. Data analyst harus mampu melihat pola dan anomali dalam data yang mungkin menunjukkan adanya masalah. Misalnya, jika penjualan tiba-tiba menurun, data analyst harus menyelidiki data penjualan, data pelanggan, dan faktor-faktor eksternal untuk mengidentifikasi penyebab penurunan tersebut.

Selain itu, penting untuk melibatkan pemangku kepentingan dalam proses identifikasi masalah. Mendengarkan kebutuhan dan masalah dari pemangku kepentingan membantu dalam memahami konteks bisnis dan merumuskan masalah dengan lebih akurat. Dengan demikian, data analyst dapat memastikan bahwa analisis yang dilakukan relevan dan bermanfaat bagi pemangku kepentingan.

### Pengumpulan Data yang Relevan

Setelah masalah diidentifikasi, langkah berikutnya adalah mengumpulkan data yang relevan untuk analisis. Pengumpulan data yang tepat sangat penting untuk memastikan bahwa analisis yang dilakukan akurat dan andal. Data analyst harus mampu mengidentifikasi sumber data yang tepat dan memastikan bahwa data yang dikumpulkan lengkap dan akurat.

Pengumpulan data melibatkan beberapa langkah, termasuk menentukan jenis data yang dibutuhkan, mengidentifikasi sumber data, dan mengumpulkan data dari sumber tersebut. Misalnya, jika masalahnya adalah penurunan penjualan, data yang relevan mungkin mencakup data penjualan historis, data pelanggan, data pemasaran, dan data eksternal seperti data ekonomi dan data pesaing.

Selain itu, penting untuk memastikan bahwa data yang dikumpulkan berkualitas tinggi. Data analyst harus memeriksa data untuk memastikan tidak ada duplikasi, kesalahan, atau nilai yang hilang. Data yang tidak berkualitas tinggi dapat menghasilkan analisis yang tidak akurat dan kesimpulan yang salah.

Penggunaan alat dan teknik yang tepat juga sangat penting dalam pengumpulan data. Data analyst harus mahir dalam menggunakan alat pengumpulan data seperti SQL, alat ETL (Extract, Transform, Load), dan alat pengumpulan data web untuk mengumpulkan data dari berbagai sumber. Dengan menggunakan alat dan teknik yang tepat, data analyst dapat memastikan bahwa data yang dikumpulkan lengkap, akurat, dan relevan untuk analisis.

### Analisis Data

Setelah data dikumpulkan, langkah berikutnya adalah menganalisis data untuk menemukan pola dan wawasan yang relevan. Analisis data melibatkan penggunaan berbagai teknik statistik dan analisis untuk mengevaluasi data dan mengidentifikasi pola yang signifikan. Data analyst harus mahir dalam menggunakan alat analisis data seperti R, Python, dan software statistik lainnya untuk menganalisis data.

Analisis data melibatkan beberapa langkah, termasuk eksplorasi data, pembersihan data, analisis statistik, dan visualisasi data. Data analyst harus mampu mengidentifikasi dan mengatasi masalah data seperti outlier, nilai yang hilang, dan kesalahan data. Pembersihan data yang baik sangat penting untuk memastikan bahwa analisis yang dilakukan akurat dan andal.

Eksplorasi data melibatkan penggunaan teknik visualisasi data seperti grafik, diagram, dan peta untuk memahami distribusi data, pola, dan anomali. Data analyst harus mahir dalam menggunakan alat visualisasi data seperti Tableau, Power BI, dan matplotlib untuk membuat visualisasi yang menarik dan informatif. Visualisasi data membantu dalam memahami data dengan lebih baik dan mengkomunikasikan hasil analisis kepada pemangku kepentingan.

Selain itu, data analyst harus mampu menggunakan teknik statistik dan analisis yang tepat untuk menganalisis data. Ini melibatkan penggunaan teknik seperti regresi, analisis klaster, analisis faktor, dan teknik machine learning untuk menganalisis data dan menemukan pola yang signifikan. Dengan menggunakan teknik yang tepat, data analyst dapat menemukan wawasan yang berharga dan memberikan rekomendasi yang informasional kepada pemangku kepentingan.

### Pengambilan Keputusan Berbasis Data

Setelah analisis data dilakukan, langkah berikutnya adalah membuat keputusan berdasarkan wawasan yang diperoleh dari analisis. Pengambilan keputusan berbasis data melibatkan penggunaan data dan analisis untuk membuat keputusan yang informasional dan efektif. Data analyst harus mampu mengkomunikasikan hasil analisis dan rekomendasi kepada pemangku kepentingan untuk mendukung pengambilan keputusan yang berbasis data.

Pengambilan keputusan berbasis data melibatkan beberapa langkah, termasuk mengidentifikasi opsi keputusan, mengevaluasi opsi berdasarkan data, dan memilih opsi yang paling informasional. Data analyst harus mampu menggunakan data dan analisis untuk mengevaluasi pro dan kontra dari setiap opsi dan memberikan rekomendasi yang berbasis data kepada pemangku kepentingan.

Selain itu, penting untuk mengkomunikasikan hasil analisis dan rekomendasi dengan jelas dan persuasif. Data analyst harus mampu menyajikan hasil analisis dalam format yang mudah dipahami oleh pemangku kepentingan dan menjelaskan implikasi dari hasil analisis. Dengan mengkomunikasikan hasil analisis dengan jelas dan persuasif, data analyst dapat membantu pemangku kepentingan dalam membuat keputusan yang berbasis data dan efektif.

Pengambilan keputusan berbasis data juga melibatkan evaluasi hasil keputusan dan pembelajaran dari pengalaman. Data analyst harus mampu mengevaluasi hasil dari keputusan yang diambil berdasarkan data dan mengidentifikasi area yang perlu diperbaiki. Dengan mengevaluasi hasil keputusan dan pembelajaran dari pengalaman, data analyst dapat terus meningkatkan kemampuan mereka dalam pengambilan keputusan berbasis data.

### Iterasi dan Perbaikan Berkelanjutan

Proses pemecahan masalah dan pengambilan keputusan tidak berhenti pada satu analisis. Data analyst harus selalu siap untuk melakukan iterasi dan perbaikan berkelanjutan. Ini berarti mengulangi proses analisis dengan data baru atau metode yang lebih baik, dan terus memperbaiki pendekatan berdasarkan umpan balik dan hasil sebelumnya.

Iterasi dan perbaikan berkelanjutan melibatkan penggunaan pendekatan agile dalam analisis data. Data analyst harus mampu bekerja dalam siklus iteratif, di mana mereka terus-menerus mengumpulkan data baru, menganalisis data, dan membuat keputusan berdasarkan hasil analisis. Dengan menggunakan pendekatan iteratif, data analyst dapat terus meningkatkan kualitas analisis dan hasil yang diperoleh.

Selain itu, penting untuk selalu mencari umpan balik dan berkolaborasi dengan pemangku kepentingan dalam proses iterasi. Umpan balik dari pemangku kepentingan membantu dalam memahami kebutuhan dan masalah yang berubah, serta mengidentifikasi area yang perlu diperbaiki. Dengan berkolaborasi dengan pemangku kepentingan, data analyst dapat memastikan bahwa analisis yang dilakukan relevan dan bermanfaat bagi seluruh organisasi.

Iterasi dan perbaikan berkelanjutan juga melibatkan pembelajaran dari pengalaman dan mengevaluasi hasil keputusan sebelumnya. Data analyst harus mampu mengevaluasi hasil dari keputusan yang diambil berdasarkan data dan mengidentifikasi area yang perlu diperbaiki. Dengan mengevaluasi hasil keputusan dan pembelajaran dari pengalaman, data analyst dapat terus meningkatkan kemampuan mereka dalam analisis data dan pengambilan keputusan berbasis data.

## Etika dan Integritas

Dalam era di mana data menjadi semakin penting, etika dan integritas memainkan peran yang sangat penting dalam pekerjaan seorang data analyst. Beberapa aspek kunci dari etika dan integritas termasuk:

### Menghormati Privasi Data

Data analyst sering kali bekerja dengan data yang sensitif dan pribadi. Oleh karena itu, mereka harus selalu menghormati privasi data dan memastikan bahwa data tersebut digunakan dan disimpan dengan cara yang aman dan etis. Ini termasuk mematuhi peraturan dan standar privasi data seperti GDPR (General Data Protection Regulation) dan CCPA (California Consumer Privacy Act).

Menghormati privasi data melibatkan beberapa langkah, termasuk mengumpulkan data hanya untuk tujuan yang sah dan menginformasikan individu tentang bagaimana data mereka akan digunakan. Data analyst harus memastikan bahwa data yang dikumpulkan hanya digunakan untuk tujuan yang dinyatakan dan tidak disalahgunakan. Misalnya, jika data dikumpulkan untuk analisis penjualan, data tersebut tidak boleh digunakan untuk tujuan lain tanpa izin dari individu yang bersangkutan.

Selain itu, penting untuk melindungi data dari akses yang tidak sah. Data analyst harus menggunakan teknik keamanan data seperti enkripsi dan anonimisasi untuk melindungi data yang sensitif. Enkripsi membantu dalam melindungi data selama transmisi dan penyimpanan, sementara anonimisasi membantu dalam melindungi identitas individu dengan menghapus atau menyembunyikan informasi pribadi.

Menghormati privasi data juga melibatkan mematuhi kebijakan dan standar privasi yang ditetapkan oleh organisasi. Data analyst harus selalu mematuhi kebijakan privasi organisasi dan memastikan bahwa data yang digunakan dalam analisis sesuai dengan kebijakan tersebut. Dengan mematuhi kebijakan privasi, data analyst dapat memastikan bahwa data yang digunakan dalam analisis digunakan dengan cara yang etis dan sah.

### Transparansi

Transparansi adalah aspek penting dari etika dalam analisis data. Data analyst harus selalu transparan tentang metode yang digunakan, sumber data, dan asumsi yang dibuat selama proses analisis. Transparansi membantu dalam membangun kepercayaan dengan pemangku kepentingan dan memastikan bahwa hasil analisis dapat dipercaya dan diverifikasi.

Transparansi melibatkan mendokumentasikan setiap langkah dalam proses analisis dan mengkomunikasikan informasi ini kepada pemangku kepentingan. Data analyst harus mencatat sumber data yang digunakan, metode analisis yang diterapkan, dan asumsi yang dibuat. Dokumentasi yang baik membantu dalam memastikan bahwa proses analisis dapat diverifikasi dan diulang oleh orang lain.

Selain itu, penting untuk mengkomunikasikan hasil analisis dengan cara yang jelas dan transparan. Data analyst harus mampu menjelaskan metodologi yang digunakan dan asumsi yang dibuat dalam bahasa yang mudah dimengerti oleh pemangku kepentingan. Misalnya, jika analisis menggunakan model prediktif, data analyst harus menjelaskan bagaimana model tersebut dibangun, data apa yang digunakan, dan asumsi apa yang dibuat.

Transparansi juga melibatkan mengakui keterbatasan dan ketidakpastian dalam analisis. Data analyst harus selalu jujur tentang keterbatasan data dan metode yang digunakan, serta potensi dampak dari keterbatasan tersebut pada hasil analisis. Dengan mengakui keterbatasan ini, data analyst dapat membantu pemangku kepentingan dalam membuat keputusan yang lebih informasional dan realistis.

### Menghindari Bias

Bias dalam analisis data dapat menghasilkan kesimpulan yang tidak akurat dan keputusan yang salah. Data analyst harus selalu berusaha untuk menghindari bias dalam proses analisis dan memastikan bahwa hasil yang diperoleh seobjektif mungkin. Ini melibatkan beberapa langkah, termasuk penggunaan teknik pengumpulan data yang representatif dan penggunaan metode analisis yang tidak bias.

Menghindari bias dalam pengumpulan data melibatkan memastikan bahwa data yang dikumpulkan representatif dari populasi yang sedang dianalisis. Misalnya, jika menganalisis data pelanggan, data analyst harus memastikan bahwa data yang dikumpulkan mencakup berbagai segmen pelanggan dan tidak hanya fokus pada satu segmen tertentu. Dengan mengumpulkan data yang representatif, data analyst dapat memastikan bahwa hasil analisis mencerminkan kenyataan yang sebenarnya.

Menghindari bias dalam analisis melibatkan penggunaan metode analisis yang tepat dan tidak bias. Data analyst harus menggunakan teknik seperti random sampling, cross-validation, dan regularisasi untuk mengurangi bias dalam model analisis. Selain itu, penting untuk menguji dan mengevaluasi model analisis untuk memastikan bahwa model tersebut tidak bias terhadap variabel tertentu.

Menghindari bias juga melibatkan kesadaran akan bias pribadi dan usaha untuk meminimalkannya. Data analyst harus selalu sadar akan potensi bias pribadi mereka dan berusaha untuk meminimalkan pengaruhnya dalam proses analisis. Misalnya, jika data analyst memiliki preferensi tertentu terhadap hasil analisis, mereka harus berusaha untuk menghilangkan preferensi tersebut dan mengevaluasi data secara objektif.

### Keadilan

Keadilan adalah aspek penting dari etika dalam analisis data. Data analyst harus selalu berusaha untuk memastikan bahwa hasil analisis digunakan dengan cara yang adil dan tidak diskriminatif. Ini melibatkan memastikan bahwa analisis tidak menghasilkan kesimpulan yang merugikan atau mendiskriminasi kelompok tertentu.

Keadilan dalam analisis data melibatkan beberapa langkah, termasuk memastikan bahwa data yang digunakan tidak bias terhadap kelompok tertentu dan bahwa metode analisis yang digunakan adil. Misalnya, jika menganalisis data demografi, data analyst harus memastikan bahwa data yang digunakan mencakup berbagai kelompok demografi dan tidak hanya fokus pada satu kelompok tertentu.

Selain itu, penting untuk memastikan bahwa hasil analisis digunakan dengan cara yang adil dan tidak diskriminatif. Data analyst harus selalu mempertimbangkan dampak dari hasil analisis pada berbagai kelompok dan memastikan bahwa hasil tersebut tidak digunakan untuk tujuan yang merugikan atau mendiskriminasi. Misalnya, jika hasil analisis digunakan untuk pengambilan keputusan dalam perekrutan, data analyst harus memastikan bahwa hasil tersebut tidak digunakan untuk mendiskriminasi calon berdasarkan ras, gender, atau karakteristik lainnya.

Keadilan juga melibatkan kesadaran akan potensi dampak dari analisis data pada kelompok yang rentan. Data analyst harus selalu mempertimbangkan dampak dari hasil analisis pada kelompok yang rentan dan berusaha untuk meminimalkan dampak negatifnya. Misalnya, jika analisis data menunjukkan bahwa kelompok tertentu memiliki risiko yang lebih tinggi untuk mengalami masalah kesehatan, data analyst harus memastikan bahwa hasil analisis digunakan untuk membantu kelompok tersebut dan bukan untuk mendiskriminasinya.

### Tanggung Jawab Sosial

Sebagai profesional yang bekerja dengan data, data analyst memiliki tanggung jawab sosial untuk memastikan bahwa data digunakan dengan cara yang etis dan bermanfaat bagi masyarakat. Ini melibatkan beberapa langkah, termasuk memastikan bahwa data digunakan untuk tujuan yang sah dan bermanfaat, serta mempertimbangkan dampak sosial dari analisis data.

Tanggung jawab sosial melibatkan memastikan bahwa data digunakan untuk tujuan yang sah dan bermanfaat. Data analyst harus selalu mematuhi peraturan dan standar etika dalam penggunaan data dan memastikan bahwa data tidak digunakan untuk tujuan yang merugikan atau tidak sah. Misalnya, data analyst harus memastikan bahwa data pelanggan tidak digunakan untuk tujuan yang tidak sah atau merugikan pelanggan.

Selain itu, penting untuk mempertimbangkan dampak sosial dari analisis data. Data analyst harus selalu mempertimbangkan dampak dari hasil analisis pada masyarakat dan berusaha untuk meminimalkan dampak negatifnya. Misalnya, jika analisis data menunjukkan bahwa kelompok tertentu memiliki risiko yang lebih tinggi untuk mengalami masalah kesehatan, data analyst harus memastikan bahwa hasil analisis digunakan untuk membantu kelompok tersebut dan bukan untuk mendiskriminasinya.

Tanggung jawab sosial juga melibatkan berpartisipasi dalam inisiatif yang bermanfaat bagi masyarakat. Data analyst dapat berpartisipasi dalam proyek-proyek yang bertujuan untuk meningkatkan kesejahteraan masyarakat, seperti analisis data untuk proyek kesehatan masyarakat atau proyek lingkungan. Dengan berpartisipasi dalam inisiatif ini, data analyst dapat memberikan kontribusi yang positif bagi masyarakat dan memastikan bahwa data digunakan untuk tujuan yang bermanfaat.

## Kolaborasi dan Kerja Tim

Kolaborasi dan kerja tim adalah keterampilan penting bagi seorang data analyst. Dalam dunia bisnis yang kompleks dan dinamis, data analyst sering kali harus bekerja dalam tim lintas fungsi untuk mengumpulkan data, menganalisis data, dan mengkomunikasikan hasil analisis. Berikut adalah beberapa elemen penting dari kolaborasi dan kerja tim:

### Berbagi Pengetahuan

Kolaborasi yang efektif melibatkan berbagi pengetahuan dan keterampilan dengan anggota tim lainnya. Data analyst harus bersedia untuk berbagi pengetahuan mereka tentang teknik analisis data, alat analisis data, dan metodologi dengan anggota tim lainnya. Ini membantu dalam meningkatkan keterampilan dan kemampuan seluruh tim, serta memastikan bahwa semua anggota tim dapat berkontribusi secara efektif dalam proses analisis.

Berbagi pengetahuan melibatkan memberikan pelatihan dan dukungan kepada anggota tim lainnya. Data analyst dapat memberikan pelatihan tentang teknik analisis data dan alat analisis data kepada anggota tim yang mungkin tidak memiliki latar belakang teknis. Misalnya, data analyst dapat memberikan pelatihan tentang penggunaan alat visualisasi data seperti Tableau atau Power BI kepada anggota tim pemasaran atau keuangan.

Selain itu, penting untuk menciptakan budaya berbagi pengetahuan dalam tim. Data analyst harus mendorong anggota tim untuk berbagi pengetahuan dan keterampilan mereka dengan orang lain, serta menciptakan lingkungan di mana semua anggota tim merasa nyaman untuk berbagi ide dan informasi. Dengan menciptakan budaya berbagi pengetahuan, data analyst dapat memastikan bahwa seluruh tim dapat berkontribusi secara efektif dalam proses analisis data.

Berbagi pengetahuan juga melibatkan berkolaborasi dengan tim lain dalam organisasi. Data analyst sering kali harus bekerja sama dengan tim pemasaran, keuangan, operasi, dan lain-lain untuk mengumpulkan data, memahami konteks bisnis, dan menyampaikan hasil analisis. Kolaborasi dengan tim lain membantu dalam mengidentifikasi kebutuhan dan prioritas yang berbeda dari setiap tim, serta memastikan bahwa analisis yang dilakukan relevan dan bermanfaat bagi seluruh organisasi.

### Menyelaraskan Tujuan

Kolaborasi yang efektif melibatkan menyelaraskan tujuan dan prioritas tim. Data analyst harus bekerja sama dengan anggota tim lainnya untuk mengidentifikasi tujuan dan prioritas yang sama, serta memastikan bahwa semua anggota tim bekerja menuju tujuan yang sama. Menyelaraskan tujuan membantu dalam menciptakan kesatuan dalam tim dan memastikan bahwa semua anggota tim dapat berkontribusi secara efektif dalam proses analisis.

Menyelaraskan tujuan melibatkan komunikasi yang terbuka dan transparan. Data analyst harus berkomunikasi dengan anggota tim lainnya untuk memahami kebutuhan dan prioritas mereka, serta memastikan bahwa semua anggota tim memiliki pemahaman yang sama tentang tujuan dan prioritas tim. Misalnya, data analyst dapat mengadakan pertemuan tim untuk mendiskusikan tujuan dan prioritas proyek, serta memastikan bahwa semua anggota tim memiliki pemahaman yang sama tentang tujuan dan prioritas tersebut.

Selain itu, penting untuk menetapkan tujuan yang jelas dan terukur. Data analyst harus bekerja sama dengan anggota tim lainnya untuk menetapkan tujuan yang jelas dan terukur, serta memastikan bahwa semua anggota tim memiliki pemahaman yang sama tentang tujuan tersebut. Dengan menetapkan tujuan yang jelas dan terukur, data analyst dapat memastikan bahwa semua anggota tim bekerja menuju tujuan yang sama dan dapat mengukur kemajuan mereka terhadap tujuan tersebut.

Menyelaraskan tujuan juga melibatkan menyesuaikan prioritas berdasarkan kebutuhan dan tuntutan yang berubah. Data analyst harus fleksibel dan siap untuk menyesuaikan prioritas mereka berdasarkan perubahan kebutuhan dan tuntutan dari pemangku kepentingan. Misalnya, jika ada perubahan dalam prioritas bisnis, data analyst harus siap untuk menyesuaikan tujuan dan prioritas mereka untuk memastikan bahwa analisis yang dilakukan tetap relevan dan bermanfaat.

### Pembagian Tugas yang Efisien

Kolaborasi yang efektif melibatkan pembagian tugas yang efisien dan pemanfaatan keterampilan dan kekuatan masing-masing anggota tim. Data analyst harus bekerja sama dengan anggota tim lainnya untuk mengidentifikasi tugas yang perlu dilakukan dan memastikan bahwa tugas tersebut dibagi secara efisien berdasarkan keterampilan dan kekuatan masing-masing anggota tim.

Pembagian tugas yang efisien melibatkan pemahaman yang mendalam tentang keterampilan dan kekuatan masing-masing anggota tim. Data analyst harus mengenali keterampilan dan kekuatan anggota tim lainnya, serta memastikan bahwa tugas-tugas yang diberikan sesuai dengan keterampilan dan kekuatan tersebut. Misalnya, jika ada anggota tim yang memiliki keterampilan analisis statistik yang kuat, data analyst dapat memberikan tugas analisis statistik kepada anggota tim tersebut.

Selain itu, penting untuk memastikan bahwa tugas-tugas yang diberikan dikelola dengan baik dan diselesaikan tepat waktu. Data analyst harus bekerja sama dengan anggota tim lainnya untuk memastikan bahwa tugas-tugas yang diberikan memiliki jadwal yang jelas dan realistis, serta memastikan bahwa semua anggota tim memiliki sumber daya yang diperlukan untuk menyelesaikan tugas tersebut. Dengan mengelola tugas dengan baik, data analyst dapat memastikan bahwa proyek analisis data berjalan lancar dan efisien.

Pembagian tugas yang efisien juga melibatkan kolaborasi dan dukungan antar anggota tim. Data analyst harus mendorong anggota tim untuk bekerja sama dan mendukung satu sama lain dalam menyelesaikan tugas-tugas mereka. Misalnya, jika ada anggota tim yang menghadapi kesulitan dalam menyelesaikan tugas tertentu, anggota tim lainnya harus siap untuk memberikan bantuan dan dukungan. Dengan menciptakan budaya kolaborasi dan dukungan, data analyst dapat memastikan bahwa semua anggota tim dapat bekerja secara efisien dan efektif.

### Membangun Kepercayaan

Kepercayaan adalah elemen penting dalam kolaborasi dan kerja tim. Data analyst harus bekerja untuk membangun kepercayaan dengan anggota tim lainnya, serta memastikan bahwa semua anggota tim merasa nyaman untuk berbagi ide, informasi, dan umpan balik. Membangun kepercayaan membantu dalam menciptakan lingkungan kerja yang positif dan produktif, serta memastikan bahwa semua anggota tim dapat berkontribusi secara efektif dalam proses analisis.

Membangun kepercayaan melibatkan beberapa langkah, termasuk berkomunikasi dengan jujur dan transparan, menghargai kontribusi anggota tim lainnya, dan memberikan umpan balik yang konstruktif. Data analyst harus selalu berkomunikasi dengan jujur dan transparan tentang proses analisis, hasil yang diperoleh, dan tantangan yang dihadapi. Misalnya, jika ada kesulitan dalam mengumpulkan data atau masalah dengan kualitas data, data analyst harus berkomunikasi dengan jujur tentang masalah tersebut dan bekerja sama dengan anggota tim lainnya untuk menemukan solusi.

Selain itu, penting untuk menghargai kontribusi anggota tim lainnya. Data analyst harus selalu menghargai upaya dan kontribusi yang diberikan oleh anggota tim lainnya, serta memberikan pengakuan dan apresiasi yang pantas.

# Fondasi Data Analytics

Data analytics, atau analisis data, telah menjadi salah satu komponen paling krusial dalam pengambilan keputusan modern. Faktor yang paling krusial ketika Anda menjadi seorang data analyst tentu saja **Data**. Di era informasi saat ini, di mana data dihasilkan dalam jumlah yang sangat besar setiap hari, kemampuan untuk mengolah dan menganalisis data menjadi sangat penting. Data analytics mengacu pada proses pengumpulan, pembersihan, dan analisis data untuk mengidentifikasi pola, tren, dan wawasan yang dapat membantu dalam membuat keputusan yang lebih baik. Dengan semakin banyaknya perusahaan yang mengandalkan data untuk strategi dan operasi mereka, data analytics tidak hanya penting bagi para profesional di bidang data tetapi juga bagi hampir semua individu yang terlibat dalam pengambilan keputusan bisnis dan organisasi.

## Konsep Dasar Data Analytics

Untuk memahami data analytics dengan lebih baik, penting untuk membedakan antara data dan informasi. Data adalah kumpulan fakta atau angka mentah yang belum diproses atau dianalisis. Data bisa berupa angka, teks, gambar, atau jenis informasi lain yang dikumpulkan dari berbagai sumber. Di sisi lain, informasi adalah hasil dari pengolahan dan analisis data. Informasi adalah data yang telah diproses sehingga menjadi sesuatu yang bermakna dan dapat digunakan untuk membuat keputusan.

Data dapat dikategorikan menjadi dua jenis utama: kuantitatif dan kualitatif. Data kuantitatif terdiri dari angka dan dapat diukur secara objektif, seperti hasil survei yang menunjukkan skor kepuasan pelanggan atau data penjualan bulanan. Data kualitatif, di sisi lain, bersifat deskriptif dan lebih sulit diukur, seperti ulasan pelanggan yang berisi komentar tentang pengalaman mereka dengan produk atau layanan.

Selain itu, data juga dapat diklasifikasikan sebagai data terstruktur dan tidak terstruktur. Data terstruktur adalah data yang telah diorganisasi dengan cara tertentu, biasanya dalam format tabel atau database yang dapat dengan mudah diquery menggunakan alat analisis. Contoh data terstruktur termasuk data dalam spreadsheet Excel atau tabel database relasional. Sebaliknya, data tidak terstruktur tidak memiliki format yang teratur, seperti teks bebas dari email, posting media sosial, atau gambar. Meskipun data tidak terstruktur dapat menjadi sumber informasi yang kaya, analisisnya sering kali memerlukan teknik yang lebih kompleks.

## Proses Data Analytics

Proses data analytics terdiri dari beberapa tahap yang saling terkait, dimulai dari pengumpulan data hingga interpretasi hasil analisis. Tahap pertama dalam proses ini adalah pengumpulan data. Data dapat dikumpulkan dari berbagai sumber, seperti survei, sistem transaksi, data sensor, atau sumber data eksternal seperti media sosial. Pengumpulan data yang efektif memerlukan pemilihan metode yang sesuai dengan tujuan analisis dan memastikan bahwa data yang dikumpulkan relevan dan berkualitas.

Setelah data dikumpulkan, tahap berikutnya adalah pembersihan data. Pembersihan data adalah proses yang penting untuk memastikan bahwa data yang digunakan dalam analisis adalah akurat dan bebas dari kesalahan. Data yang tidak bersih atau tidak konsisten dapat mengarah pada hasil analisis yang salah atau menyesatkan. Proses pembersihan data melibatkan identifikasi dan perbaikan data yang hilang, duplikat, atau tidak konsisten. Teknik ini termasuk pengisian nilai yang hilang, menghapus duplikat, dan mengoreksi kesalahan format.

Eksplorasi data adalah langkah selanjutnya setelah pembersihan. Pada tahap ini, analisis awal dilakukan untuk memahami pola dan tren dalam data. Teknik eksplorasi data melibatkan statistik deskriptif seperti mean, median, dan modus, serta visualisasi awal seperti histogram dan diagram pencar. Tujuan eksplorasi data adalah untuk mendapatkan wawasan awal tentang data dan menentukan apakah ada pola atau anomali yang perlu diteliti lebih lanjut.

Tahap berikutnya adalah analisis data, di mana teknik analisis yang lebih canggih diterapkan untuk mengidentifikasi hubungan dan pola yang lebih kompleks dalam data. Metode analisis data dapat mencakup teknik statistik seperti regresi, clustering, atau analisis korelasi. Misalnya, regresi digunakan untuk memahami hubungan antara variabel dan memprediksi nilai variabel dependen berdasarkan variabel independen. Clustering, di sisi lain, digunakan untuk mengelompokkan data ke dalam grup yang memiliki kemiripan tertentu. Teknik ini dapat digunakan untuk segmentasi pelanggan atau analisis pola pembelian.

Interpretasi hasil adalah tahap terakhir dalam proses data analytics. Setelah analisis data dilakukan, penting untuk menginterpretasikan hasilnya dengan benar dan mengkomunikasikan temuan kepada pemangku kepentingan. Interpretasi hasil melibatkan pemahaman tentang apa arti hasil analisis dalam konteks bisnis atau organisasi dan bagaimana hasil tersebut dapat digunakan untuk mengambil keputusan. Presentasi hasil analisis harus jelas dan mudah dipahami oleh audiens non-teknis, sering kali menggunakan visualisasi data seperti grafik dan diagram untuk menyampaikan temuan dengan lebih efektif.

## Alat dan Teknologi Data Analytics

Dalam dunia data analytics, berbagai alat dan teknologi tersedia untuk membantu dalam setiap tahap proses analisis. Salah satu alat yang paling umum digunakan adalah Microsoft Excel. Excel menawarkan berbagai fitur untuk analisis data, termasuk tabel pivot, grafik, dan formula statistik. Meskipun Excel adalah alat yang kuat untuk analisis data dasar, alat yang lebih canggih sering kali diperlukan untuk analisis data yang lebih kompleks.

R dan Python adalah dua bahasa pemrograman yang sangat populer dalam data analytics. R adalah bahasa pemrograman yang dirancang khusus untuk analisis statistik dan visualisasi data. R menawarkan berbagai paket dan pustaka yang memudahkan analisis data dan pembuatan grafik. Python, di sisi lain, adalah bahasa pemrograman umum yang juga sangat populer dalam data analytics karena kemampuannya dalam pemrosesan data dan analisis. Python memiliki pustaka seperti Pandas, NumPy, dan SciPy yang sangat berguna untuk analisis data, serta pustaka visualisasi seperti Matplotlib dan Seaborn.

Alat visualisasi data seperti Tableau dan Power BI juga memainkan peran penting dalam data analytics. Tableau adalah alat visualisasi yang memungkinkan pengguna untuk membuat dashboard interaktif dan grafik yang memudahkan pemahaman data. Power BI, yang dikembangkan oleh Microsoft, menawarkan fitur serupa dan integrasi dengan berbagai sumber data. Kedua alat ini membantu dalam menyajikan hasil analisis data dengan cara yang lebih intuitif dan mudah dipahami.

Dalam konteks data besar, teknologi big data seperti Hadoop dan Apache Spark menjadi semakin relevan. Hadoop adalah framework sumber terbuka yang memungkinkan penyimpanan dan pemrosesan data dalam skala besar dengan menggunakan model pemrograman MapReduce. Apache Spark, di sisi lain, adalah framework pemrosesan data yang lebih cepat dan lebih efisien dibandingkan dengan Hadoop. Spark mendukung pemrosesan batch dan streaming, serta analisis data real-time, yang memungkinkan analisis data besar dengan waktu respons yang lebih cepat.

## Keterampilan dan Kompetensi

Untuk menjadi seorang data analyst yang efektif, diperlukan berbagai keterampilan dan kompetensi. Salah satu keterampilan utama adalah pemahaman statistik. Statistik dasar seperti mean, median, standar deviasi, dan distribusi probabilitas adalah alat penting dalam analisis data. Pemahaman statistik memungkinkan analisis data yang lebih mendalam dan membantu dalam mengidentifikasi pola dan tren dalam data.

Kemampuan pemrograman juga sangat penting dalam data analytics. Bahasa pemrograman seperti R dan Python digunakan secara luas dalam analisis data karena kemampuan mereka untuk memproses dan menganalisis data dengan cepat. Keterampilan pemrograman memungkinkan analisis data yang lebih kompleks dan pengembangan model prediktif. Selain itu, pemahaman tentang SQL (Structured Query Language) sangat berguna untuk mengquery dan memanipulasi data dalam database relasional.

Selain keterampilan teknis, kemampuan komunikasi juga sangat penting. Seorang data analyst harus dapat menyampaikan hasil analisis kepada audiens non-teknis dengan cara yang jelas dan mudah dipahami. Kemampuan untuk menjelaskan temuan analisis dan memberikan rekomendasi yang dapat ditindaklanjuti adalah keterampilan yang sangat berharga. Ini sering kali melibatkan pembuatan laporan, presentasi, dan visualisasi data yang efektif.

## Etika dan Privasi Data

Etika dan privasi data adalah aspek penting dalam data analytics yang tidak boleh diabaikan. Dalam pengumpulan dan penggunaan data, penting untuk mematuhi regulasi privasi data seperti General Data Protection Regulation (GDPR) di Uni Eropa dan Health Insurance Portability and Accountability Act (HIPAA) di Amerika Serikat. Regulasi ini mengatur bagaimana data pribadi harus dikumpulkan, disimpan, dan digunakan, serta memberikan hak kepada individu atas data mereka.

Praktik etis dalam penggunaan data juga harus diperhatikan. Data harus dikumpulkan dan digunakan dengan cara yang transparan dan tidak merugikan individu. Ini termasuk mendapatkan izin yang jelas dari individu sebelum mengumpulkan data mereka dan memastikan bahwa data tidak disalahgunakan atau dibagikan tanpa izin.

## Studi Kasus dan Contoh

Untuk memberikan gambaran yang lebih konkret tentang penerapan data analytics, studi kasus nyata dapat menjadi sumber pembelajaran yang berharga. Misalnya, sebuah perusahaan ritel dapat menggunakan data analytics untuk memahami pola pembelian pelanggan dan mengidentifikasi produk yang paling laris. Dengan menganalisis data transaksi, perusahaan dapat melakukan segmentasi pelanggan dan mengembangkan strategi pemasaran yang lebih efektif.

Proyek praktis juga dapat membantu pembaca untuk mulai menerapkan konsep data analytics. Misalnya, proyek sederhana seperti menganalisis data penjualan untuk mengidentifikasi tren musiman atau menganalisis data survei untuk memahami kepuasan pelanggan dapat memberikan pengalaman langsung dalam proses data analytics. Panduan langkah demi langkah untuk proyek ini dapat mencakup pengumpulan data, pembersihan, eksplorasi, analisis, dan presentasi hasil.

# Keuntungan Menjadi Data Analyst

Menjadi seorang data analyst adalah salah satu profesi yang semakin diminati di era digital saat ini. Peran seorang data analyst tidak hanya terbatas pada perusahaan teknologi, tetapi meluas ke berbagai industri seperti keuangan, kesehatan, retail, hingga pemerintahan. Dengan kemampuan untuk memahami, menganalisis, dan menginterpretasikan data, seorang data analyst menjadi aset berharga bagi perusahaan dalam mengambil keputusan strategis. Namun, selain memberikan manfaat bagi perusahaan, profesi ini juga menawarkan banyak keuntungan bagi individu yang memilih jalur karier ini.

Salah satu keuntungan utama menjadi seorang data analyst adalah kesempatan untuk bekerja secara fleksibel, termasuk bekerja secara remote. Di era pasca-pandemi ini, banyak perusahaan yang telah mengadopsi model kerja jarak jauh. Sebagai seorang data analyst, Anda tidak terikat pada satu lokasi fisik untuk menyelesaikan pekerjaan. Dengan koneksi internet yang stabil dan perangkat yang memadai, seorang data analyst dapat bekerja dari mana saja, baik itu di rumah, kafe, atau bahkan dari negara lain. Fleksibilitas ini memberikan keleluasaan yang besar, terutama bagi mereka yang mengutamakan work-life balance. Tentu saja hal iini sangat bergantung dengan kebijakan perusahaan tempat data analyst tersebut bekerja.

Selain fleksibilitas, potensi penghasilan seorang data analyst juga sangat menarik. Data menjadi aset yang sangat berharga bagi perusahaan, dan kemampuan untuk mengolah data dengan baik sering kali dihargai dengan gaji yang kompetitif. Menurut berbagai survei gaji global, seorang data analyst dapat menikmati penghasilan yang lebih tinggi dibandingkan profesi lain dengan tingkat pendidikan dan pengalaman yang serupa. Ditambah lagi, seiring bertambahnya pengalaman dan keterampilan yang dimiliki, gaji seorang data analyst bisa meningkat secara signifikan, terutama jika mereka juga menguasai keterampilan lanjutan seperti machine learning, pemrograman, dan penggunaan alat analitik modern.

Tidak hanya soal fleksibilitas dan penghasilan, menjadi seorang data analyst juga menawarkan kesempatan untuk terus berkembang. Dunia data terus berubah dengan cepat. Setiap tahun, muncul teknologi dan metode baru yang bisa mempermudah proses analisis dan membuka wawasan yang lebih mendalam. Oleh karena itu, seorang data analyst dituntut untuk terus belajar dan meningkatkan keterampilannya. Bagi banyak orang, kesempatan untuk selalu belajar dan berkembang ini merupakan salah satu hal yang membuat pekerjaan ini menarik. Menjadi data analyst adalah tentang rasa ingin tahu dan kemampuan untuk terus beradaptasi dalam menghadapi tantangan baru.

# Metrik Bisnis

Salah satu tugas utama seorang data analyst adalah membantu perusahaan dalam mengukur performa bisnis melalui data yang dihasilkan. Untuk mencapai tujuan ini, data analyst menggunakan berbagai metrik bisnis yang relevan dengan industri dan kebutuhan perusahaan. Metrik bisnis adalah alat penting untuk memahami apakah suatu perusahaan sedang bergerak ke arah yang benar atau perlu melakukan penyesuaian. Setiap industri memiliki metrik-metrik yang spesifik, namun konsep dasarnya tetap sama: mengukur dan mengevaluasi kinerja untuk pengambilan keputusan yang lebih baik.

Dalam dunia retail, misalnya, metrik seperti Customer Lifetime Value (CLV) dan Customer Acquisition Cost (CAC) menjadi sangat penting. CLV mengukur total nilai yang bisa dihasilkan dari seorang pelanggan selama masa hidup pelanggan tersebut dalam suatu bisnis. Sebaliknya, CAC menghitung berapa biaya yang dikeluarkan untuk mendapatkan pelanggan baru. Dengan membandingkan kedua metrik ini, seorang data analyst bisa memberikan rekomendasi apakah strategi marketing yang digunakan perusahaan efektif atau perlu diubah. Misalnya, jika biaya untuk mendapatkan pelanggan lebih besar daripada nilai yang dihasilkan pelanggan, perusahaan mungkin harus meninjau kembali strategi pemasaran mereka atau meningkatkan upaya retensi pelanggan.

Di sektor kesehatan, metrik yang sering digunakan adalah Patient Satisfaction Score dan Average Treatment Cost. Patient Satisfaction Score mengukur tingkat kepuasan pasien terhadap pelayanan yang diterima di rumah sakit atau klinik, yang sangat penting untuk menjaga reputasi dan kualitas pelayanan. Seorang data analyst di bidang ini akan menganalisis tren data terkait kepuasan pasien dan mengidentifikasi area mana saja yang memerlukan perbaikan. Di sisi lain, Average Treatment Cost membantu rumah sakit dalam mengelola biaya operasional dan menyeimbangkan antara kualitas pelayanan dan efisiensi biaya. Dengan metrik ini, seorang data analyst dapat mengusulkan strategi untuk menekan biaya tanpa mengorbankan kualitas layanan kesehatan.

Dalam industri teknologi, metrik seperti Monthly Active Users (MAU) dan Churn Rate sering kali menjadi pusat perhatian. MAU mengukur jumlah pengguna aktif yang menggunakan suatu aplikasi atau platform setiap bulan, yang memberikan gambaran langsung tentang seberapa banyak interaksi yang terjadi. Sementara itu, Churn Rate menghitung persentase pelanggan atau pengguna yang berhenti menggunakan layanan dalam periode tertentu. Data analyst akan memantau metrik ini untuk memastikan bahwa perusahaan dapat menjaga dan meningkatkan tingkat keterlibatan pengguna, serta meminimalkan kehilangan pelanggan. Misalnya, jika Churn Rate meningkat, perusahaan mungkin harus melihat faktor apa yang menyebabkan pengguna berhenti menggunakan layanan mereka, seperti masalah dalam pengalaman pengguna atau kurangnya fitur yang menarik.

Metrik-metrik ini hanyalah sebagian kecil dari alat yang digunakan data analyst untuk membantu perusahaan dalam berbagai industri. Pada akhirnya, dengan menganalisis metrik-metrik bisnis ini, seorang data analyst berperan penting dalam memastikan bahwa perusahaan dapat mengambil keputusan yang lebih baik, meningkatkan efisiensi, dan menciptakan nilai jangka panjang. Bab ini akan membahas lebih lanjut bagaimana metrik bisnis diterapkan di industri-industri lain seperti keuangan, manufaktur, dan media, serta bagaimana data analyst dapat membuat perubahan nyata di dalam perusahaan melalui pemahaman yang mendalam terhadap metrik ini.

Berikut adalah daftar metrik bisnis yang sering digunakan oleh data analyst di berbagai industri:

## Retail & E-commerce

1. ***Customer Lifetime Value*** (CLV) – Nilai total yang dihasilkan dari seorang pelanggan selama masa hidupnya.
2. ***Customer Acquisition Cost*** (CAC) – Biaya yang dikeluarkan untuk memperoleh satu pelanggan baru.
3. ***Conversion Rate*** – Persentase pengunjung yang berubah menjadi pelanggan.
4. ***Average Order Value*** (AOV) – Rata-rata nilai transaksi per pesanan.
5. ***Cart Abandonment Rate*** – Persentase pelanggan yang meninggalkan keranjang belanja sebelum menyelesaikan pembelian.
6. ***Repeat Purchase Rate*** – Persentase pelanggan yang melakukan pembelian berulang.
7. ***Gross Margin*** – Selisih antara pendapatan penjualan dan biaya produksi, dinyatakan dalam persentase.

## Sektor Kesehatan

1. ***Patient Satisfaction Score*** – Tingkat kepuasan pasien terhadap pelayanan yang diterima.
2. ***Average Treatment Cost*** – Rata-rata biaya perawatan atau pengobatan.
3. ***Hospital Readmission Rate*** – Persentase pasien yang kembali dirawat dalam waktu tertentu setelah keluar dari rumah sakit.
4. ***Patient Retention Rate*** – Persentase pasien yang tetap menggunakan layanan dalam periode waktu tertentu.
5. ***Operational Efficiency*** – Ukuran efisiensi operasional dalam hal biaya dan waktu.

## Teknologi & Aplikasi Digital

1. ***Monthly Active Users*** (MAU) – Jumlah pengguna yang aktif menggunakan aplikasi/platform setiap bulan.
2. ***Daily Active Users*** (DAU) – Jumlah pengguna yang aktif setiap hari.
3. ***Churn Rate*** – Persentase pengguna atau pelanggan yang berhenti menggunakan layanan dalam periode waktu tertentu.
4. ***Retention Rate*** – Persentase pengguna yang tetap menggunakan layanan setelah periode tertentu.
5. ***Average Revenue Per User*** (ARPU) – Pendapatan rata-rata yang dihasilkan dari setiap pengguna dalam periode tertentu.

## Keuangan

1. ***Return on Investment (ROI)*** – Ukuran pengembalian investasi relatif terhadap biaya yang dikeluarkan.
2. ***Net Profit Margin*** – Persentase laba bersih terhadap total pendapatan.
3. ***Operating Cash Flow*** – Jumlah uang tunai yang dihasilkan oleh operasi perusahaan.
4. ***Debt-to-Equity Ratio*** – Rasio antara total utang dan ekuitas pemegang saham.
5. ***Current Ratio*** – Rasio likuiditas yang mengukur kemampuan perusahaan untuk membayar kewajiban jangka pendek.

## Manufaktur

1. ***Overall Equipment Effectiveness*** (OEE) – Ukuran efektivitas peralatan produksi, menggabungkan ketersediaan, kinerja, dan kualitas.
2. ***Production Downtime*** – Waktu yang hilang akibat mesin tidak berfungsi atau gangguan produksi lainnya.
3. ***Cost per Unit*** – Biaya untuk memproduksi satu unit produk.
4. ***Inventory Turnover*** – Frekuensi rata-rata persediaan diperbarui dalam satu periode waktu.
5. ***Defect Rate*** – Persentase produk cacat dalam produksi.

## Media & Periklanan

1. ***Cost per Click*** (CPC) – Biaya yang dibayar untuk setiap klik iklan.
2. ***Cost per Acquisition*** (CPA) – Biaya yang dibayar untuk setiap tindakan atau konversi yang diinginkan.
3. ***Click-Through Rate*** (CTR) – Persentase orang yang mengeklik iklan setelah melihatnya.
4. ***Impressions*** – Jumlah total tampilan iklan.
5. ***Engagement Rate*** – Persentase interaksi dengan konten atau iklan, seperti like, share, dan komentar.

## SaaS (Software as a Service)

1. ***Annual Recurring Revenue*** (ARR) – Pendapatan tahunan yang dapat diprediksi dari pelanggan berlangganan.
2. ***Customer Retention Rate*** – Persentase pelanggan yang tetap berlangganan setelah periode tertentu.
3. ***Customer Churn Rate*** – Persentase pelanggan yang berhenti berlangganan dalam periode tertentu.
4. ***Lifetime Value to Customer Acquisition Cost*** (LTV) Ratio – Perbandingan antara CLV dan CAC.
5. ***Net Promoter Score*** (NPS) – Ukuran kesetiaan pelanggan berdasarkan rekomendasi.

# Mengenal Tipe Data

Ketika kita berbicara tentang data, hal pertama yang harus kita pahami adalah bahwa tidak semua data diciptakan sama. Dalam dunia statistik dan analisis data, tipe data memiliki peranan penting karena menentukan bagaimana kita bisa memproses dan menganalisis informasi tersebut. Misalnya, data yang berupa angka akan diperlakukan berbeda dengan data yang berupa teks atau kategori. Jika kita salah memahami tipe data, maka hasil analisis yang kita lakukan bisa saja menjadi tidak akurat, atau bahkan lebih buruk, tidak relevan.

Setiap jenis data memiliki ciri khas dan metode analisis yang berbeda. Ini mengapa kita perlu memahami tipe data dengan baik. Sebagai contoh, ketika kita berbicara tentang tinggi badan seseorang, kita tentu berurusan dengan angka (data kuantitatif), berbeda ketika kita membicarakan warna favorit, yang merupakan data kategori (data kualitatif). Kesalahan dalam mengidentifikasi tipe data ini bisa mengarah pada kesalahan dalam memilih metode statistik yang tepat untuk digunakan. Oleh karena itu, pemahaman tentang tipe data bukan hanya dasar, tetapi juga fundamental dalam analisis data.

Ada dua kategori utama dalam tipe data, yaitu data **kuantitatif** dan data **kualitatif**. Data kuantitatif mengacu pada data yang bisa diukur dan dihitung, sedangkan data kualitatif berkaitan dengan deskripsi atau kualitas dari sesuatu yang tidak bisa diukur secara numerik. Di antara dua kategori besar ini, ada beberapa subkategori yang lebih spesifik, yang akan kita bahas lebih dalam pada bagian selanjutnya. Memahami perbedaan antara keduanya membantu kita dalam menentukan metode analisis yang sesuai, apakah kita harus menggunakan statistik deskriptif sederhana atau justru model statistik yang lebih kompleks.

## Data Kuantitatif

Data kuantitatif adalah tipe data yang berfokus pada jumlah atau angka. Tipe data ini dibagi lagi menjadi dua subkategori, yaitu data diskrit dan data kontinu. Data diskrit adalah data yang terdiri dari nilai-nilai terbatas yang dapat dihitung, seperti jumlah anak dalam keluarga atau jumlah mobil yang dimiliki seseorang. Sebaliknya, data kontinu adalah data yang bisa mengambil nilai dalam rentang tertentu, seperti tinggi badan, berat badan, atau suhu. Dalam hal ini, kita bisa memiliki banyak sekali nilai yang mungkin, tergantung pada seberapa detail kita mengukurnya. Sebagai contoh, suhu badan bisa diukur hingga dua desimal, seperti 36,75°C.

Secara matematis, kita bisa membedakan data diskrit dan kontinu dengan cara mendefinisikan himpunan nilai yang mungkin. Misalnya, jika kita mengatakan adalah jumlah mobil yang dimiliki oleh sebuah keluarga, maka , yang merupakan contoh dari data diskrit. Sedangkan untuk data kontinu, misalnya tinggi badan seseorang , kita bisa menyatakan cm, di mana nilainya bisa berupa bilangan real dalam rentang tersebut.

## Data Kualitatif

Berbeda dengan data kuantitatif, data kualitatif berhubungan dengan deskripsi atau kategori. Tipe data ini juga sering disebut sebagai data kategori, karena mencakup hal-hal yang tidak dapat diukur dengan angka. Contoh dari data kualitatif adalah warna rambut, jenis kelamin, atau status pernikahan. Data kualitatif sendiri terbagi menjadi dua jenis, yaitu **nominal** dan **ordinal**.

Data nominal adalah data kategori yang tidak memiliki urutan atau hierarki. Contoh data nominal adalah jenis kelamin (pria atau wanita) atau warna favorit (merah, biru, hijau). Sebaliknya, data ordinal adalah data kategori yang memiliki urutan atau hierarki tertentu, seperti tingkat pendidikan (SMA, S1, S2) atau tingkat kepuasan pelanggan (sangat puas, puas, tidak puas). Namun, meskipun data ordinal memiliki urutan, jarak antara kategori tersebut tidak dapat diukur secara kuantitatif.

Sebagai contoh, jika kita memiliki data tingkat kepuasan pelanggan dengan nilai 1 hingga 5 (dari sangat tidak puas hingga sangat puas), meskipun ada urutan dalam skala tersebut, kita tidak bisa mengatakan bahwa perbedaan antara nilai 1 dan 2 sama dengan perbedaan antara nilai 4 dan 5. Ini karena data ordinal tidak memiliki jarak tetap di antara nilainya.

## Contoh Kasus

Mari kita ambil sebuah contoh sederhana dalam dunia bisnis. Sebuah perusahaan e-commerce ingin menganalisis data pelanggan mereka untuk melihat pola pembelian. Mereka mengumpulkan data berikut: usia pelanggan, jenis kelamin, jumlah produk yang dibeli, dan tingkat kepuasan. Dalam hal ini, usia pelanggan adalah data kuantitatif kontinu, jenis kelamin adalah data kualitatif nominal, jumlah produk yang dibeli adalah data kuantitatif diskrit, dan tingkat kepuasan adalah data kualitatif ordinal. Dengan memahami tipe data yang berbeda ini, perusahaan bisa menentukan metode analisis yang tepat. Misalnya, untuk data kuantitatif seperti usia, mereka bisa menggunakan rata-rata, sedangkan untuk data ordinal seperti tingkat kepuasan, mereka bisa menggunakan median, modus atau membuat tabel frekuensi.

## Penutup

Memahami tipe data adalah langkah pertama yang krusial dalam melakukan analisis statistik yang akurat. Tanpa pemahaman yang benar tentang tipe data, kita bisa salah dalam memilih metode analisis yang tepat, yang pada akhirnya bisa menghasilkan kesimpulan yang keliru. Data bukan hanya sekedar angka atau kategori, tapi juga memiliki karakteristik yang berbeda, dan pemahaman mendalam tentang tipe data ini akan mempermudah kita dalam melakukan analisis yang lebih efektif dan tepat sasaran.

# Statistika Dasar

Statistika adalah fondasi dari data analytics yang memungkinkan data analyst untuk dapat memahami, menginterpretasikan, dan membuat keputusan berdasarkan data. Tanpa statistika, data hanyalah angka-angka yang tidak berarti. Statistika memberikan alat dan metode yang memungkinkan kita untuk menarik kesimpulan yang valid dan relevan dari data yang tersedia. Di dalam dunia yang semakin data-driven ini, kemampuan untuk memahami dan menerapkan konsep-konsep statistika dasar menjadi sangat penting bagi seorang data analyst.

Statistika dasar mencakup berbagai teknik dan metode yang memungkinkan kita untuk menganalisis data secara efektif. Ini termasuk konsep seperti ukuran pemusatan data, ukuran penyebaran, distribusi probabilitas, dan inferensi statistik. Dengan menguasai statistika dasar, kita dapat mengevaluasi dan menyajikan data dengan cara yang memberi nilai tambah dan mendukung pengambilan keputusan yang berbasis data. Selain itu, statistika dasar juga membekali kita dengan kemampuan untuk mengidentifikasi pola, membuat prediksi, dan mengkomunikasikan temuan dengan jelas dan akurat.

Dalam bab ini, kita akan menjelajahi berbagai konsep statistika dasar secara mendalam. Kita akan memulai dengan pengertian populasi dan contoh (sampel), ukuran pemusatan data, dan ukuran penyebaran data. Kemudian, kita akan membahas distribusi probabilitas dan bagaimana mereka dapat digunakan untuk memahami perilaku data. Terakhir, kita akan mempelajari dasar-dasar inferensi statistik, termasuk uji hipotesis dan interval kepercayaan, yang memungkinkan kita untuk membuat kesimpulan tentang populasi berdasarkan sampel.

Salah satu konsep fundamental dalam statistika dasar adalah ukuran pemusatan data. Ukuran pemusatan memberikan gambaran tentang titik pusat atau nilai representatif dari sebuah dataset. Mean atau rata-rata adalah salah satu ukuran pemusatan yang paling umum digunakan. Mean dihitung dengan menjumlahkan semua nilai dalam dataset dan membaginya dengan jumlah data. Namun, mean bisa dipengaruhi oleh nilai ekstrim atau outlier, yang membuatnya kadang tidak mencerminkan representasi yang akurat dari data. Dalam kasus ini, median menjadi ukuran pemusatan yang lebih baik karena tidak terpengaruh oleh nilai ekstrem. Median adalah nilai tengah dari dataset ketika data diurutkan. Jika jumlah data genap, median adalah rata-rata dari dua nilai tengah. Mode, ukuran pemusatan lainnya, adalah nilai yang paling sering muncul dalam dataset. Meskipun mode tidak selalu ada atau mungkin lebih dari satu, ia memberikan wawasan tentang nilai yang paling umum dalam data.

Ukuran dispersi data juga merupakan bagian penting dari statistika dasar. Ukuran dispersi mengukur sejauh mana nilai-nilai dalam dataset menyebar atau bervariasi dari nilai pemusatan. Rentang adalah ukuran dispersi yang paling sederhana, dihitung dengan mengurangkan nilai minimum dari nilai maksimum dalam dataset. Namun, rentang tidak mempertimbangkan bagaimana nilai-nilai di antara nilai ekstrem tersebar. Untuk mengatasi hal ini, kita menggunakan varians dan deviasi standar. Varians adalah rata-rata kuadrat selisih antara setiap nilai dan mean dataset. Semakin besar varians, semakin besar penyebaran data dari mean. Deviasi standar, yang merupakan akar kuadrat dari varians, memberikan ukuran penyebaran yang lebih mudah diinterpretasikan karena berada dalam satuan yang sama dengan data asli.

Distribusi probabilitas adalah konsep statistika dasar yang penting untuk memahami bagaimana data tersebar dan berperilaku. Distribusi probabilitas menggambarkan kemungkinan terjadinya berbagai nilai dalam dataset. Dua distribusi probabilitas yang sering digunakan adalah distribusi normal dan distribusi binomial. Distribusi normal, yang sering disebut sebagai distribusi Gaussian, berbentuk lonceng dan simetris di sekitar mean. Distribusi ini sering digunakan karena banyak fenomena alami mengikuti pola distribusi normal. Di sisi lain, distribusi binomial digunakan untuk memodelkan jumlah keberhasilan dalam sejumlah percobaan yang independen, seperti jumlah heads yang muncul dalam serangkaian lemparan koin.

Inferensi statistik memungkinkan kita untuk menarik kesimpulan tentang populasi berdasarkan sampel data. Dua konsep utama dalam inferensi statistik adalah uji hipotesis dan interval kepercayaan. Uji hipotesis digunakan untuk menguji pernyataan atau klaim tentang populasi dengan menggunakan data sampel. Proses ini melibatkan penentuan hipotesis nol () dan hipotesis alternatif (: “H 1”), serta menggunakan data sampel untuk menentukan apakah ada cukup bukti untuk menolak hipotesis nol. Interval kepercayaan memberikan rentang nilai yang, dengan tingkat kepercayaan tertentu, diyakini mencakup parameter populasi yang tidak diketahui. Interval kepercayaan membantu kita untuk memahami seberapa akurat estimasi parameter dari sampel kita.

Memahami dan menerapkan statistika dasar adalah keterampilan yang sangat berharga bagi seorang data analyst. Dengan konsep-konsep ini, kita dapat menganalisis data dengan lebih efektif, mengidentifikasi pola, dan membuat keputusan yang didasarkan pada bukti yang kuat. Statistika dasar juga menyediakan alat yang diperlukan untuk mengkomunikasikan temuan dengan jelas dan akurat, yang merupakan aspek penting dari analisis data yang sukses.

## Populasi dan Sampel dalam Statistika

Dalam statistika, istilah populasi dan contoh (sample, sampel) merujuk pada dua konsep penting yang membantu kita dalam pengumpulan dan analisis data. Memahami perbedaan antara populasi dan sampel sangat penting dalam desain penelitian dan interpretasi hasil analisis data.

**Populasi** adalah keseluruhan kelompok yang menjadi objek perhatian atau penelitian. Populasi mencakup semua unit, individu, atau item yang memenuhi kriteria tertentu dan kita ingin membuat kesimpulan atau generalisasi. Populasi bisa sangat besar atau sangat kecil, tergantung pada konteks penelitian. Misalnya, jika kita tertarik untuk mempelajari kebiasaan makan masyarakat di suatu negara, maka seluruh populasi yang dimaksud adalah seluruh warga negara tersebut. Dalam studi tersebut, populasi mencakup setiap individu yang tinggal di negara itu.

Contoh lain dari populasi adalah dalam studi kesehatan. Jika peneliti ingin meneliti prevalensi diabetes di antara semua orang berusia 60 tahun ke atas di kota tertentu, maka populasi adalah semua orang berusia 60 tahun ke atas di kota tersebut. Populasi ini mencakup semua individu dalam kelompok umur yang relevan, yang memungkinkan peneliti untuk mendapatkan gambaran yang menyeluruh tentang prevalensi diabetes di seluruh kelompok tersebut.

Di Indonesia pengumpulan data populasi ini dilakukan dengan sensus seperti yang dilakukan oleh Badan Pusat Statistik (BPS). Misalnya sensus penduduk, sensus ekonomi dan sensus pertanian. Karena biaya dan tenaga yang dibutuhkan untuk melakukan sensus tidak sedikit, maka sensus dilakukan dalam rentang waktu yang cukup lama dari sensus ke sensus berikutnya. Contohnya sensus penduduk dilakukan 10 tahun sekali pada setiap tahun yang berakhiran angka 0.

Namun, dalam praktiknya, sering kali tidak mungkin atau tidak praktis untuk mengumpulkan data dari seluruh populasi. Di sinilah konsep contoh (sampel) menjadi penting. Contoh (sampel) adalah subset dari populasi yang dipilih untuk mewakili populasi secara keseluruhan. Penggunaan sampel memungkinkan peneliti untuk melakukan analisis yang dapat digunakan untuk membuat inferensi tentang populasi tanpa harus memeriksa setiap individu dalam populasi tersebut. Sampel diambil dengan berbagai metode, termasuk sampling acak, stratified sampling, dan lain-lain, untuk memastikan bahwa sampel tersebut representatif dari populasi yang lebih besar.

Contoh:

Kembali ke contoh studi kebiasaan makan masyarakat di suatu negara. Alih-alih mengumpulkan data dari semua warga negara, peneliti mungkin memutuskan untuk mengambil sampel. Misalnya, mereka bisa memilih 1.000 orang dari berbagai wilayah di negara tersebut, dengan proporsi yang mencerminkan distribusi demografis negara tersebut. Sampel ini diharapkan dapat memberikan gambaran yang cukup representatif tentang kebiasaan makan warga negara secara keseluruhan.

Dalam studi kesehatan tentang prevalensi diabetes, peneliti mungkin memilih sampel dari 500 orang yang berusia 60 tahun ke atas untuk melakukan survei. Sampel ini dipilih secara acak atau berdasarkan kriteria tertentu untuk memastikan bahwa sampel tersebut mewakili populasi berusia 60 tahun ke atas dengan cara yang adil. Hasil survei pada sampel ini kemudian digunakan untuk membuat estimasi tentang prevalensi diabetes di seluruh populasi.

|  |
| --- |
| Dalam Matematika dan Statistika |
| Berbagai notasi untuk populasi umumnya menggunakan huruf kapital, misalnya , , (dibaca: sigma), (dibaca: myu) dan lain-lain, sedangkan untuk contoh (sampel) umumnya menggunakan notasi huruf kecil. Misalnya , , , (dibaca: “X bar”). |

## Ukuran Pemusatan Data

Ukuran pemusatan dan penyebaran data adalah dua konsep fundamental yang membentuk dasar dari analisis statistik. Untuk memahami data secara menyeluruh, tidak cukup hanya dengan mengetahui nilai rata-rata (mean atau average) atau median; penting juga untuk memahami bagaimana data tersebar di sekitar nilai pusat tersebut.

Ukuran pemusatan data adalah statistik yang digunakan untuk menentukan nilai representatif dari suatu dataset. Ukuran ini memberikan gambaran tentang pusat distribusi data, yang dapat membantu dalam memahami pola dan tren dalam data. Tiga ukuran pemusatan utama yang sering digunakan adalah mean, median, dan modus (mode). Masing-masing memiliki kegunaan dan aplikasi tertentu tergantung pada sifat data yang dianalisis.

### Rata-Rata

Rata-rata, atau mean, adalah ukuran pemusatan yang paling umum dan sering digunakan. Mean dihitung dengan menjumlahkan semua nilai dalam dataset dan membaginya dengan jumlah data. Mean memberikan gambaran umum tentang pusat distribusi data.

**Contoh**:

Misalkan kita memiliki dataset yang berisi nilai-nilai berikut: 3, 7, 5, 9, dan 11. Untuk menghitung rata-rata:

1. Jumlahkan semua nilai: 3 + 7 + 5 + 9 + 11 = 35
2. Hitung jumlah data: 5
3. Bagi total jumlah dengan jumlah data: 35 / 5 = 7

Jadi, rata-rata dari dataset ini adalah 7. Ini berarti bahwa, secara umum, nilai-nilai dalam dataset cenderung berada di sekitar angka 7.

|  |
| --- |
| Berhati-hatilah |
| Jika ada nilai ekstrem (outlier) yang sangat tinggi atau rendah, rata-rata dapat terpengaruh. Misalnya, jika kita menambahkan nilai 100 ke dataset di atas (sehingga menjadi 3, 7, 5, 9, 11, dan 100), rata-ratanya akan berubah:   1. Jumlahkan semua nilai baru: 3 + 7 + 5 + 9 + 11 + 100 = 135 2. Hitung jumlah data: 6 3. Bagi total jumlah dengan jumlah data: 135 / 6 = 22.5   Rata-rata pada data baru ini adalah 22.5, yang jauh lebih tinggi dibandingkan sebelumnya karena nilai ekstrem 100. |

Dari contoh sederhana di atas kita dapat memahami bahwa untuk mendapatkan nilai rata-rata adalah dengan cara menjumlahkan semua nilai dan dibagi dengan banyaknya data.

|  |
| --- |
| Dalam Matematika dan Statistika |
| Penjumlahan dinotasikan dengan simbol Sigma (), masing-masing nilai pada data dinotasikan dengan (dibaca: “X i”), banyaknya data adalah , dan nilai rata-rata dinotasikan dengan (dibaca: “X bar”). Formula untuk menghitung rata-rata adalah:  Penulisan disederhanakan menjadi  Pada contoh di atas nilai . Secara umum, jika ada sebanyak data maka akan ada . Oleh karena itu penulisannya akan menjadi  Setelah itu hasil penjumlahan di atas dibagi sesuai dengan banyaknya data, yaitu .  atau |

### Nilai Tengah

Nilai tengah, atau median, adalah nilai tengah dari dataset setelah data diurutkan. Jika jumlah data genap, median adalah rata-rata dari dua nilai tengah.

**Contoh 1 (Jumlah Data Ganjil)**:

Misalkan kita memiliki dataset berikut: 2, 8, 5, 3, 9. Untuk menghitung median:

1. Urutkan data: 2, 3, 5, 8, 9
2. Temukan nilai tengah: Nilai tengah adalah 5, karena 5 adalah nilai yang berada di tengah dataset yang telah diurutkan.
3. Jadi, median dari dataset ini adalah 5.

**Contoh 2 (Jumlah Data Genap)**:

Misalkan kita memiliki dataset berikut: 3, 2, 9, 8. Untuk menghitung median:

1. Urutkan data: 2, 3, 8, 9
2. Temukan dua nilai tengah: 3 dan 8
3. Hitung rata-rata dari dua nilai tengah: (3 + 8) / 2 = 5.5

Jadi, median dari dataset ini adalah 5.5.

|  |
| --- |
| Catatan |
| Median tidak terpengaruh oleh nilai ekstrem, sehingga memberikan gambaran yang lebih stabil tentang pusat data dalam distribusi yang tidak simetris atau mengandung outlier. |

### Modus

Modus, atau mode, adalah nilai yang paling sering muncul dalam dataset. Dataset bisa memiliki lebih dari satu modus atau tidak memiliki modus sama sekali jika tidak ada nilai yang berulang.

**Contoh 1 (Satu Modus)**:

Misalkan kita memiliki dataset berikut: 4, 1, 2, 4, 5, 4. Untuk menghitung modus:

Nilai 4 muncul tiga kali, lebih sering daripada nilai lainnya. Jadi, modus dari dataset ini adalah 4.

**Contoh 2 (Beberapa Modus)**:

Misalkan kita memiliki dataset berikut: 1, 2, 2, 3, 3, 4. Untuk menghitung modus:

Nilai 2 dan 3 masing-masing muncul dua kali, yang lebih sering daripada nilai lainnya. Jadi, modus dari dataset ini adalah 2 dan 3, sehingga dataset ini memiliki dua modus.

**Contoh 3 (Tidak Ada Modus)**:

Misalkan kita memiliki dataset berikut: 1, 2, 3, 4, 5. Untuk menghitung modus:

Semua nilai muncul hanya sekali, jadi tidak ada nilai yang lebih sering muncul. Jadi, dataset ini tidak memiliki modus.

Selain mean, median, dan modus, ada beberapa ukuran pemusatan lain yang mungkin berguna tergantung pada konteks data. Misalnya, trimmed mean adalah jenis rata-rata yang dihitung setelah menghilangkan persentase terendah dan tertinggi dari data. Ini berguna ketika data mengandung outlier yang signifikan yang dapat mempengaruhi rata-rata secara tidak proporsional.

**Trimmed Mean** adalah salah satu variasi dari mean yang digunakan untuk mengatasi masalah yang disebabkan oleh nilai ekstrem atau outlier dalam dataset. Dengan memangkas sebagian dari data di ujung distribusi, trimmed mean memberikan gambaran yang lebih stabil tentang pusat data, mengurangi dampak nilai ekstrem yang mungkin mengganggu hasil analisis.

*Trimmed mean* dihitung dengan menghilangkan persentase tertentu dari nilai-nilai tertinggi dan terendah dalam dataset sebelum menghitung rata-rata. Biasanya, proses ini dilakukan dengan menghapus nilai ekstrem dari kedua ujung distribusi (atas dan bawah) untuk meminimalkan pengaruh outlier.

Misalnya, dalam sebuah dataset dengan 100 nilai, kita dapat menghapus 5% nilai terendah dan 5% nilai tertinggi, dan kemudian menghitung mean dari nilai-nilai yang tersisa.

**Langkah-Langkah Menghitung Trimmed Mean**

1. Urutkan data dari nilai terendah ke tertinggi.
2. Pilih persentase nilai yang akan dipangkas dari kedua ujung distribusi. Misalnya, jika kita menggunakan 10% pemangkasan, kita akan menghapus 10% nilai terendah dan 10% nilai tertinggi.
3. Hapus nilai dari kedua ujung sesuai dengan persentase yang telah ditentukan.
4. Hitung rata-rata dari nilai-nilai yang tersisa setelah pemangkasan.

**Contoh**

Misalkan kita memiliki dataset dengan 6 nilai berikut: 2, 3, 4, 8, 12, 100.

1. Data yang sudah diurutkan: 2, 3, 4, 8, 12, 100
2. Misalkan kita menggunakan 20% pemangkasan.  
   Dalam dataset ini yang berisi 6 nilai, 20% dari 6 adalah 1.2. Karena kita harus menggunakan angka bulat, kita memangkas 1 nilai terendah dan 1 nilai tertinggi.
3. Hapus nilai 2 dan 100. Dataset yang tersisa adalah: 3, 4, 8, 12
4. Hitung Mean dari Sisa Data:  
   Jumlahkan nilai-nilai yang tersisa: 3 + 4 + 8 + 12 = 27  
   Hitung jumlah data yang tersisa: 4  
   Hitung mean: 27 / 4 = 6.75

Jadi, trimmed mean dari dataset ini dengan 20% pemangkasan adalah 6.75.

**Keuntungan dan Kekurangan**

Keuntungan:

1. **Mengurangi Pengaruh Outlier**  
   Trimmed mean mengurangi dampak nilai ekstrem yang bisa mempengaruhi rata-rata aritmetika secara signifikan.
2. **Memberikan Gambaran Lebih Stabil**  
   Trimmed mean memberikan gambaran yang lebih stabil tentang pusat distribusi data, terutama dalam dataset yang mengandung pencilan atau outlier.

Kekurangan:

1. **Menghilangkan Informasi**  
   Dengan memangkas data, kita menghilangkan sebagian informasi yang mungkin masih relevan.
2. **Penentuan Persentase Pemangkasan**  
   Menentukan persentase pemangkasan yang tepat bisa bersifat subjektif dan tergantung pada konteks data.

*Trimmed mean* sering digunakan dalam situasi di mana data mengandung outlier yang dapat mempengaruhi hasil analisis. Misalnya, dalam analisis pendapatan rumah tangga, di mana beberapa individu mungkin memiliki pendapatan yang sangat tinggi atau sangat rendah dibandingkan dengan sebagian besar populasi, trimmed mean dapat memberikan gambaran yang lebih akurat tentang pendapatan tipikal.

Dalam penelitian ilmiah, *trimmed mean* juga digunakan untuk mengurangi pengaruh data yang tidak konsisten atau anomali yang dapat mengganggu hasil eksperimen. Dengan menggunakan *trimmed mean*, peneliti dapat memastikan bahwa analisis mereka lebih representatif dari data yang mayoritas, dan lebih sedikit terpengaruh oleh nilai ekstrem yang mungkin tidak mencerminkan pola umum.

## Ukuran Penyebaran Data

Statistik lain seperti rentang (*range*), ragam (*variance*), dan deviasi standar (*standard deviation*) memberikan gambaran tentang penyebaran data. Namun, terkadang sulit untuk menginterpretasikan ragam karena satuannya adalah kuadrat dari unit data. Di sinilah deviasi standar menjadi lebih berguna karena berada dalam unit yang sama dengan data asli. Deviasi standar memberikan wawasan yang lebih intuitif tentang penyebaran data.

### Rentang

Rentang, atau range, adalah ukuran penyebaran yang paling sederhana. Range dihitung dengan mengurangkan nilai minimum dari nilai maksimum dalam dataset. Range memberikan gambaran kasar tentang seberapa jauh nilai-nilai dalam dataset tersebar, tetapi tidak memberikan informasi tentang distribusi nilai di antara nilai ekstrem. Misalnya, jika kita memiliki dataset nilai tes: 55, 67, 78, 85, dan 92, rentangnya adalah 92 - 55 = 37. Range ini menunjukkan bahwa ada perbedaan 37 poin antara nilai terendah dan tertinggi, tetapi tidak menggambarkan bagaimana nilai-nilai lain tersebar di antara keduanya.

### Ragam

Ragam, atau variance, adalah ukuran penyebaran yang lebih komprehensif dibandingkan range. Ragam dihitung dengan menghitung rata-rata kuadrat selisih antara setiap nilai dan rata-rata dataset. Ragam memberikan informasi tentang seberapa besar variasi data dari nilai rata-rata, tetapi satuannya adalah kuadrat dari unit data asli, yang dapat membuat interpretasinya menjadi kurang intuitif. Untuk menghitung ragam, pertama-tama kita menghitung rata-rata dari dataset, lalu kita hitung selisih antara setiap nilai dan rata-rata, kuadratkan selisih tersebut, dan ambil rata-ratanya.

Sebagai contoh, jika dataset nilai ujian adalah 55, 67, 78, 85, dan 92, pertama-tama kita hitung rata-rata (75.4). Selisih setiap nilai dari rata-rata adalah 55-75.4 = -20.4, 67-75.4 = -8.4, dan seterusnya hingga 92-75.4 = 16.6. Kuadratkan selisih tersebut dan rata-ratakan hasilnya untuk mendapatkan varians.

|  |
| --- |
| Dalam Matematika dan Statistika |
| Kita memiliki nili ujian dari 5 siswa adalah , . Pertama, kita hitung nilai rata-rata, , terlebih dahulu.  Selanjutnya hitung selisih masing-masing nilai terhadap rata-rata dan dikuadratkan.  Jika diperhatikan, kita dapat menyederhanakan penulisan di atas menjadi notasi matematika  Terakhir, kita rata-ratakan dengan membagi nilai tersebut dengan banyaknya data , sedangkan untuk data contoh (sample) maka pembaginya adalah .  Untuk populasi kita gunakan  atau  sedangkan untuk sampel kita gunakan  atau |

Deviasi standar adalah ukuran dispersi yang merupakan akar kuadrat dari varians. Deviasi standar memberikan ukuran variasi data dalam satuan yang sama dengan data asli, membuatnya lebih mudah diinterpretasikan dibandingkan varians. Jika deviasi standar kecil, berarti data cenderung berkumpul di sekitar mean, sedangkan deviasi standar besar menunjukkan variasi yang lebih besar. Menggunakan contoh yang sama, setelah menghitung varians, kita menghitung akar kuadrat dari varians untuk mendapatkan deviasi standar. Dalam kasus dataset nilai tes sebelumnya, jika variansnya adalah 142.16, deviasi standarnya adalah √142.16 ≈ 11.94. Ini menunjukkan seberapa jauh, rata-rata, nilai-nilai dalam dataset menyimpang dari mean.

Ketiga ukuran ini—rentang, varians, dan deviasi standar—menyediakan informasi penting tentang bagaimana data tersebar dan bervariasi. Rentang memberikan gambaran umum, tetapi tidak mendetail. Varians memberikan informasi lebih mendalam namun dalam unit kuadrat, sedangkan deviasi standar, sebagai akar kuadrat dari varians, menawarkan ukuran variasi dalam satuan yang lebih intuitif dan mudah dipahami. Dalam analisis data, memilih ukuran dispersi yang tepat membantu kita untuk mendapatkan gambaran yang jelas tentang distribusi dan variasi data, yang penting untuk pengambilan keputusan dan interpretasi data yang akurat.

Untuk data yang sangat besar atau kompleks, konsep seperti koefisien variasi juga dapat digunakan. Koefisien variasi adalah rasio deviasi standar terhadap mean, sering dinyatakan sebagai persentase. Ini memungkinkan perbandingan variasi antara dataset yang memiliki unit yang berbeda atau mean yang berbeda.

## Distribusi Probabilitas

Distribusi probabilitas adalah konsep yang menghubungkan probabilitas dengan nilai data yang mungkin. Dalam statistika, distribusi probabilitas digunakan untuk memahami dan memodelkan bagaimana data terdistribusi dan seberapa mungkin nilai-nilai tertentu terjadi.

Distribusi normal, atau distribusi Gaussian, adalah salah satu distribusi yang paling sering digunakan dalam statistika. Distribusi ini memiliki bentuk lonceng simetris dan ditentukan oleh dua parameter: mean dan deviasi standar. Banyak fenomena alami, seperti tinggi badan manusia dan hasil ujian, cenderung mengikuti distribusi normal.

Namun, tidak semua data mengikuti distribusi normal. Dalam kasus seperti itu, distribusi lain mungkin lebih sesuai. Misalnya, distribusi binomial digunakan untuk model probabilitas dari sejumlah percobaan independen dengan dua hasil yang mungkin, seperti lemparan koin. Distribusi Poisson digunakan untuk menghitung probabilitas kejadian yang jarang terjadi dalam interval waktu tertentu.

Selain distribusi-distribusi tersebut, distribusi eksponensial, chi-square, dan t-student juga memiliki aplikasi penting dalam statistika, terutama dalam analisis inferensial dan uji hipotesis.

## Inferensi Statistik

Inferensi statistik adalah proses menarik kesimpulan tentang populasi berdasarkan data sampel. Ini adalah salah satu aspek terpenting dari analisis data karena memungkinkan kita untuk membuat keputusan dan prediksi yang informatif dengan hanya menggunakan subset data yang lebih kecil daripada populasi keseluruhan.

Uji hipotesis adalah metode yang digunakan untuk menguji klaim atau pernyataan tentang populasi berdasarkan sampel. Proses ini melibatkan penyusunan hipotesis nol dan hipotesis alternatif. Hipotesis nol adalah pernyataan yang ingin kita uji dan sering kali merupakan asumsi bahwa tidak ada efek atau perbedaan. Hipotesis alternatif adalah pernyataan yang menunjukkan adanya efek atau perbedaan. Setelah menetapkan hipotesis, analisis dilakukan untuk menentukan apakah data mendukung hipotesis alternatif atau tidak.

Selain uji hipotesis, interval kepercayaan memberikan rentang nilai yang diharapkan mencakup parameter populasi dengan tingkat kepercayaan tertentu. Misalnya, interval kepercayaan 95% berarti bahwa jika kita melakukan analisis yang sama berulang kali, 95% dari interval kepercayaan yang dihasilkan akan mencakup parameter populasi yang sebenarnya. Interval kepercayaan membantu kita untuk memahami sejauh mana estimasi sampel dapat dipercaya.

## Aplikasi dalam Data Analytics

Konsep-konsep statistika dasar ini sangat penting dalam praktik data analytics. Dalam proses analisis data, statistika membantu kita untuk mengidentifikasi pola, tren, dan hubungan yang signifikan. Misalnya, ukuran pemusatan dan dispersi data memungkinkan kita untuk memahami distribusi data dan mengidentifikasi outlier yang mungkin memerlukan perhatian khusus.

Distribusi probabilitas membantu kita untuk memodelkan dan memahami ketidakpastian dalam data, sedangkan inferensi statistik memungkinkan kita untuk membuat keputusan berdasarkan data sampel. Semua ini adalah keterampilan yang sangat berharga dalam dunia data yang semakin kompleks dan berbasis bukti.

Sebagai contoh, dalam analisis data penjualan, kita dapat menggunakan statistika dasar untuk mengidentifikasi produk yang paling populer, mengevaluasi tren penjualan dari waktu ke waktu, dan membuat prediksi tentang penjualan masa depan. Dengan memahami ukuran pemusatan dan dispersi, kita dapat mengidentifikasi produk yang memiliki penjualan tinggi secara konsisten dan produk yang menunjukkan variasi besar dalam penjualan.

Distribusi probabilitas dapat digunakan untuk memodelkan kemungkinan permintaan produk di masa depan, dan uji hipotesis dapat membantu kita untuk menguji apakah strategi pemasaran baru memiliki dampak yang signifikan terhadap penjualan.

# Eksplorasi dan Visualisasi

Ketika berbicara tentang data, langkah pertama yang sangat penting adalah memahami data tersebut dengan baik. Sebelum kita mulai menarik kesimpulan atau membangun model, kita harus melihat data secara mendalam—memahami distribusi, tren, pola, atau bahkan anomali yang ada di dalamnya. Di sinilah eksplorasi data berperan. Eksplorasi data atau *Exploratory Data Analysis* (EDA) adalah proses awal dalam analisis data yang berfokus pada pemahaman karakteristik dasar dari dataset. Dengan teknik ini, kita bisa lebih peka terhadap pola-pola yang mungkin tersembunyi, serta menilai asumsi-asumsi awal yang bisa kita gunakan dalam analisis lebih lanjut.

Eksplorasi data tidak hanya berhenti pada angka-angka. Kadang, angka saja tidak cukup untuk memberikan gambaran lengkap tentang data. Oleh karena itu, visualisasi memainkan peran krusial. Visualisasi data memungkinkan kita untuk mengubah angka menjadi gambar yang lebih mudah dipahami dan ditangkap oleh otak manusia. Misalnya, dengan menggunakan grafik seperti *histogram*, *boxplot*, atau *scatter plot*, kita dapat melihat distribusi dan hubungan antar variabel dengan lebih jelas. Visualisasi yang baik bisa membantu kita mengidentifikasi tren atau korelasi yang mungkin terlewat jika hanya berfokus pada angka.

Sebagai seorang data analyst, kemampuan untuk melakukan eksplorasi data dan menciptakan visualisasi yang informatif sangat penting. Visualisasi yang tepat bukan hanya untuk analisis pribadi, tetapi juga sangat berguna ketika kita ingin menyampaikan hasil analisis kepada audiens yang lebih luas, termasuk mereka yang mungkin tidak memiliki latar belakang teknis. Seorang data analyst yang baik adalah yang mampu menceritakan cerita di balik data melalui visualisasi yang sederhana namun kuat.

## Teknik Eksplorasi Data

Eksplorasi data melibatkan beberapa teknik yang cukup sederhana namun efektif. Teknik-teknik ini biasanya dimulai dengan perhitungan statistik dasar seperti mean, median, mode, varians, dan standar deviasi untuk mendapatkan gambaran awal tentang distribusi data. Misalnya, jika kita memiliki dataset yang berisi tinggi badan sekelompok orang, langkah pertama yang bisa dilakukan adalah menghitung rata-rata dan standar deviasinya.

Dari sini, kita bisa mengetahui apakah tinggi badan rata-rata kelompok tersebut tinggi atau rendah, serta seberapa besar penyebaran data tersebut. Selain itu, kita juga bisa mencari nilai ekstrim atau pencilan (*outlier*) yang mungkin ada dalam data. *Outlier* ini bisa menjadi tanda adanya kesalahan dalam pengumpulan data atau justru merupakan indikasi fenomena khusus yang perlu ditelusuri lebih lanjut.

Teknik eksplorasi data adalah metode yang digunakan untuk memahami dan menganalisis data secara awal, sebelum menerapkan model statistik atau machine learning. Berikut adalah beberapa jenis teknik eksplorasi data yang umum digunakan:

1. Statistik Deskriptif  
   Teknik ini melibatkan perhitungan nilai-nilai statistik seperti rata-rata, median, modus, standar deviasi, dan berbagai ukuran dispersi lainnya. Statistik deskriptif membantu merangkum dan memahami distribusi data, menyajikan informasi tentang karakteristik penting data secara numerik.
2. Visualisasi Data  
   Visualisasi digunakan untuk menampilkan data dalam bentuk grafis, sehingga pola dan tren bisa lebih mudah dipahami. Beberapa visualisasi umum termasuk histogram, boxplot, scatter plot, heatmap, dan bar chart. Teknik ini sangat efektif untuk menemukan outlier, distribusi data, dan hubungan antar variabel.
3. Tabulasi Silang (*Crosstabulation*)  
   Teknik ini digunakan untuk menganalisis hubungan antara dua variabel kategorikal dengan menampilkan data dalam bentuk tabel kontingensi. Ini membantu dalam memahami interaksi atau pola yang ada antara variabel-variabel tersebut.
4. Analisis Korelasi  
   Korelasi digunakan untuk menentukan hubungan linier antara dua variabel numerik. Koefisien korelasi Pearson adalah salah satu metode yang paling umum digunakan untuk mengukur seberapa kuat hubungan antar variabel. Korelasi dapat membantu mengidentifikasi variabel yang saling mempengaruhi.
5. Deteksi Outlier  
   Teknik ini digunakan untuk mengidentifikasi data yang jauh menyimpang dari mayoritas data lainnya. Outlier dapat memberikan informasi penting tentang kemungkinan kesalahan data, peristiwa langka, atau aspek yang perlu dianalisis lebih lanjut.
6. Analisis Dimensi dan Reduksi Dimensi  
   Ketika bekerja dengan data yang memiliki banyak fitur, reduksi dimensi seperti Principal Component Analysis (PCA) dapat membantu mengurangi kompleksitas data sambil mempertahankan informasi yang paling penting. Teknik ini berguna untuk mengidentifikasi pola dalam data berdimensi tinggi.
7. Grouping atau *Clustering*  
   Clustering seperti K-Means atau DBSCAN adalah teknik yang digunakan untuk mengelompokkan data ke dalam beberapa grup berdasarkan kesamaan. Ini berguna dalam eksplorasi data yang tidak berlabel untuk menemukan kelompok atau segmen alami dalam data.

## Jenis-jenis Visualisasi Statistik

Ada berbagai jenis visualisasi yang bisa digunakan tergantung dari tipe data dan tujuan analisis yang ingin dicapai. Beberapa visualisasi dasar yang sering digunakan dalam eksplorasi data antara lain:

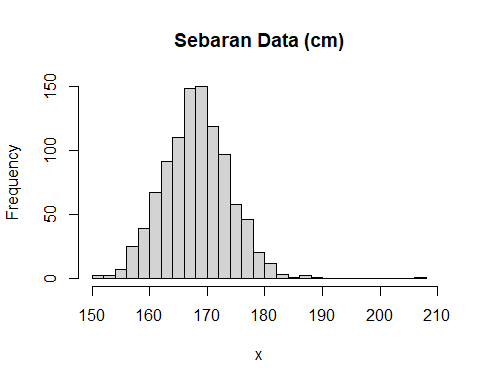
### Histogram

Histogram Sangat berguna untuk melihat distribusi data kontinu. Histogram memecah data menjadi beberapa “bin” atau kelompok interval, dan memberikan gambaran tentang frekuensi data pada masing-masing kelompok. Misalnya, jika kita ingin mengetahui bagaimana distribusi penghasilan bulanan di suatu perusahaan, histogram dapat membantu kita melihat rentang penghasilan yang paling umum.

|  |
| --- |
| Ingat! |
| Histogram hanya bisa digunakan untuk memvisualisasikan data numerik, tidak bisa untuk data kategorik. Untuk memvisualisasikan data kategorik kamu dapat menggunakan Bar chart (diagram batang) atau pie chart (diagram lingkaran) |

Informasi atau insight yang dapat diperoleh dari histogram antara lain:

1. Mengetahui bentuk sebaran dan ukuran pemusatan data kontinu dengan cepat. Bentuk sebaran data dapat berupa simetrik, seragam, bimodal, menjulur ke kanan atau ke kiri.
2. Mengetahui rentang nilai data dengan mudah karena kita dapat langsung melihat nilai terendah dan tertinggi. Batang histogram menunjukkan banyaknya data yang berada pada rentang tersebut. Semakin tinggi batangnya maka semakin banyak data pada rentang tersebut.
3. Mengidentifikasi outlier pada data. Jika ada batang yang berada jauh atau terpisah dari batang histogram yang lain maka dapat dicurigai bahwa data tersebut adalah outlier.



Dari contoh histogram di atas kita dapat mengetahui bahwa sebaran data tersebut hampir simetrik. Tetapi karena ada nilai data yang jauh lebih besar dari nilai pada umumnya, sehingga menyebabkan sebarannya sedikit menjulur ke kanan(*skewed to the right*). Data yang jauh ini kita dapat mencurigainya sebagai pencilan/outlier. Outlier tersebut bernilai antara 205 hingga 210 cm.

Kita perlu mengetahui rentang nilai yang dimiliki oleh masing-masing batang. Diantara nilai 160 hingga 170 terdapat 5 batang, sehingga rentang batang adalah (170-160)/5 = 2 cm. Lebar rentang ini sama untuk semua batang. Jadi untuk rentang batang dari 160 sampai 170 cm adalah [160, 162], [162, 164], [164, 166], [166, 168], dan [168, 170]. Oleh karena itu kita dapat menyimpulkan bahwa rentang nilai terbanyak pada data tersebut adalah [168, 170] cm karena batangnya yang paling tinggi. Jika kita mengabaikan outlier, maka nilai rata-rata dan median dari data ini sekitar 168 cm.

Contoh pembuatan histogram berikut ini akan membantu kamu dalam memahami bagaimana histogram bekerja dan interpretasinya. Tentu saja ketika kamu bekerja sebagai data analyst tidak perlu menghitung manual. kamu dapat menggunakan software untuk mempercepat pekerjaan kamu.

Misalnya, kita memiliki data tinggi badan dari 20 orang (dalam cm):

180.7, 167, 166.9, 153.5, 164.8, 167.2, 174.3, 164.4, 162.7, 158.8, 169.8, 177.5, 164.4, 170.7, 176.2, 168.6, 178.8, 162, 160.2, 171.2.

Maka kita tahu bahwa .

1. Tentukan rentang kelas  
   Rentang () kelas dapat diperoleh dengan cara mengurangi nilai terbesar dan terkecil. Pada contoh data ini nilai terbesar adalah 180.7 dan nilai terkecil adalah 153.5.
2. Menentukan jumlah kelas (binning)  
   Penentuan jumlah kelas () dalam pembuatan histogram dapat dipilih berdasarkan keinginan pembuat histogram atau menggunakan aturan yang sudah dibuat. Penentuan jumlah kelas berdasarkan keinginan pembuatnya misalnya dengan menentukan sebanyak 5, 6, 7, dan seterunya. Penentuan jumlah kelas yang digunakan di software seperti R atau Python secara default menggunakan aturan yang sudah ada seperti yang diusulkan oleh H.A Sturges, DP Doane, Freedman-Diaconis, dan lain-lain. Untuk contoh ini kita akan menggunakan aturan Sturges dengan formula

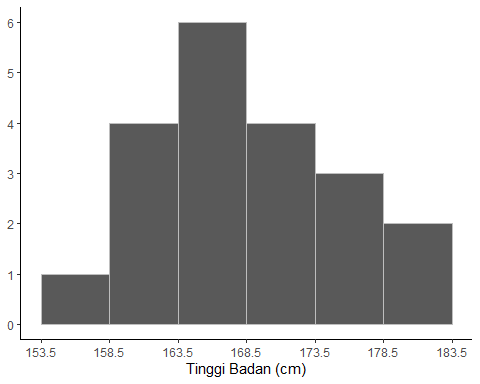
* Maksud dari simbol adalah hasil perhitungan dibulatkan ke atas (ceiling). Karena kita memiliki maka
* kita dapat menggunakan $ k = 6$ untuk pembuatan histogram ini. Artinya akan ada 6 batang yang terbentuk.

1. Menghitung lebar kelas  
   Kita dapat menghitung lebar kelas atau lebar batang histogram ()dengan
2. Membuat batas kelas Batas kelas adalah rentang nilai yang nantinya akan mengandung nilai data dari masing-masing kelas. Karena nilai terkecil pada data di atas adalah min = 153.5 dan lebar kelas adalah L = 5. maka batas kelas pertama adalah = . Kita dapat melakukan cara yang sama untuk batas kelas selanjutnya hingga terbentuk 5 kelas.
3. Membuat tabel frekuensi  
   Tabel frekuensi adalah sebuah tabel yang memberikan informasi banyaknya kemunculan dari setiap nilai atau kelompok nilai.

| Kelas | Batas Kelas | frekuensi (fi) |
| --- | --- | --- |
| 1 | [153.5, 158.5) | 1 |
| 2 | [158.5, 163.5) | 4 |
| 3 | [163.5, 168.5) | 6 |
| 4 | [168.5, 173.5) | 4 |
| 5 | [173.5, 178.5) | 3 |
| 6 | [178.5, 183.5] | 2 |

* Total dari frekuensi tersebut jika dijumlahkan adalah 20 dan sesuai dengan di awal, semua data masuk ke setiap kelas dan tidak ada nilai yang terhimpit diantara dua kelas.

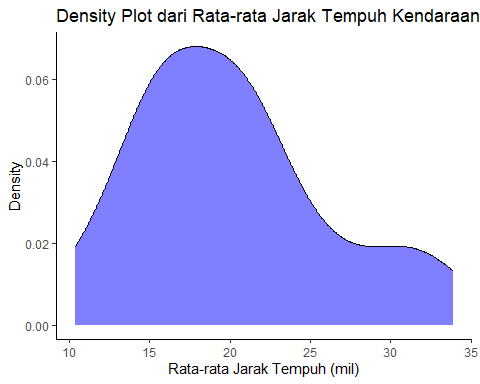
1. Menggambar histogram  
   Selanjutnya membuat batang histogram dengan batas kelas berada di sumbu X (horizontal) dan frekuensi di sumbu Y (vertikal). Tinggi batang menyesuaikan dengan nilai frekuensi pada setiap kelas. Histogram akan terlihat seperti berikut ini.



### Density

Density plot merupakan salah satu metode yang sering digunakan untuk visualisasi distribusi data numerik. Berbeda dengan histogram yang memecah data ke dalam interval atau bin, density plot menampilkan estimasi *smooth* (halus) dari sebaran peluang variabel numerik. Estimasi ini biasanya diperoleh menggunakan metode *Kernel Density Estimation* (KDE). Bentuk sebaran data dengan histogram sangat bergantung pada banyaknya kelas dalam membentuk sebaran data. Dengan density plot, kita bisa lebih mudah melihat pola sebaran data dan membandingkan sebaran antar kelompok yang berbeda dalam data.

Density plot sangat berguna dalam tahap eksplorasi data, terutama ketika kita ingin mengetahui bentuk sebaran (apakah normal, skewed, atau bimodal), dan mengidentifikasi kemungkinan *outliers* (nilai yang berada jauh dari mayoritas data). Dengan tampilan yang lebih halus daripada histogram, density plot memudahkan kita untuk mengenali pola sebaran dengan lebih jelas. Berikut ini contoh visualisasi density plot dari sebuah pengukuran jarak yang dapat ditempuh oleh 32 jenis kendaraan sehingga dapat diperoleh hasil rata-rata jarak tempuh dalam mil per galon.



|  |
| --- |
| Dalam Matematika dan Statistika |
| Untuk membuat sebuah density plot dari data empiris kamu memerlukan beberapa langkah yang bisa dibilang cukup rumit bagi orang yang tidak memiliki latar belakang pendidikan matematika/statsitika apalagi yang tidak suka matematika.   1. Data Sampel Pertama, kita harus memiliki data sampel terlebih dahulu. Misalnya, sebagai contoh sederhana, kita memiliki data {2, 3, 5, 6, 8}. 2. Memilih Kernel Kernel adalah fungsi yang digunakan untuk memperkirakan distribusi data di sekitar titik data. *Kernel Density Estimation* (KDE) yang biasanya digunakan adalah kernel **Gaussian**. Gaussian Kernel memiliki rumus   Mirip dengan fungsi sebaran Normal? Ya! Rumus di atas adalah fungsi sebaran Normal dengan dan .   1. Memilih Bandwidth **Bandwidth** (**h**) adalah parameter penting dalam KDE yang menentukan lebar kernel akan menyebar di sekitar tiitk setiap titik data. Bandwidth bisa dipilih secara manual atau menggunakan aturan standar, seperti \***Silverman’s rule**   dengan adalah standar deviasi sampel, dan n adalah banyaknya data. Bandwidth menentukan tingkat *smoothness* dari sebuah density plot. Untuk contoh ini kita gunakan agar sederhana.   1. Menghitung KDE Density di setiap titik dalam rentang yang kita ingin hitung dengan rumus berikut:   dengan \* adalah titik data ke- \* adalah bandwidth \* adalah kernel (Gaussian)  Kita akan menghitung estimasi KDE secara manual pada beberapa titik untuk data sampel kita, misalnya . Untuk kita hitung KDE dengan rumus:  Sekarang kita hitung masing-masing komponen Gaussian untuk setiap titik data.   * Untuk * Untuk * Untuk * Untuk * Untuk   Sekarang kita hitung dengan menjumlahkan semuanya   1. Visualisasi Jika kita ulangi langkah ini untuk berbagai nilai dalam rentang data kita, kita bisa mendapatkan kurva *smooth* yang merepresentasikan density plot secara keseluruhan. Density plot ini akan menunjukkan sebaran dari sampel data kita dengan representasi *smooth* berdasarkan KDE. |

### Boxplot

Boxplot atau diagram kotak garis adalah salah satu visualisasi paling populer untuk melihat distribusi data, terutama dalam mengidentifikasi median, kuartil, dan outlier. Dengan boxplot, kita bisa langsung melihat apakah ada data yang jauh lebih besar atau kecil dari yang lain, serta bagaimana rentang data pada 50% tengahnya. Jenis visualisasi ini juga sangat berguna untuk menggambarkan sebaran data dan membandingkan beberapa kelompok data.

|  |
| --- |
| Ingat! |
| Boxplot juga hanya bisa untuk data numerik seperti pada histogram, data kategorik pada boxplot hanya dapat digunakan untuk pengelompokan boxplot. |

Dengan menggunakan data pada contoh histogram sebelumnya kita akan membuat boxplot.

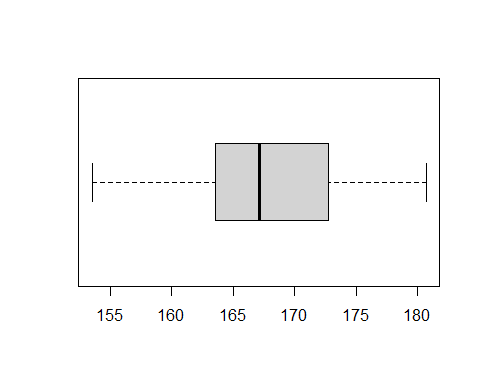
1. Hitung nilai , dan   
   Untuk mendapatkan nilai , dan kita harus mengurutkan data dari terkecil ke terbesar terlebih dahulu. Data yang sudah diurutkan akan terlihat seperti berikut ini.

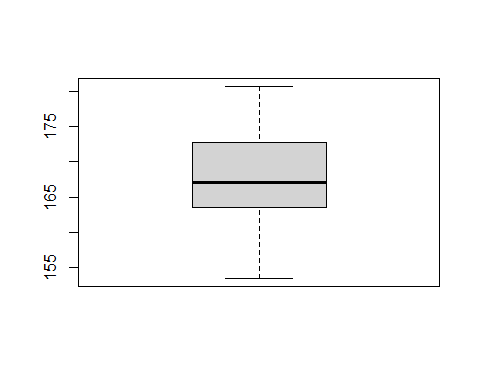
* 153.5, 158.8, 160.2, 162, 162.7, 164.4, 164.4, 164.8, 166.9, 167, 167.2, 168.6, 169.8, 170.7, 171.2, 174.3, 176.2, 177.5, 178.8, 180.7
  + adalah data pada urutan ke 25%.
  + adalah data ke 50% atau tengah-tengah, disebut juga Median atau Nilai Tengah.
  + adalah data pada uturan ke 75%.
* Langkah pertama kita hitung karena lebih mudah. Setelah data diurutkan kita bagi dua data sama banyaknya, yaitu
* 180.7 167 166.9 153.5 164.8 167.2 174.3 164.4 162.7 158.8, dan  
  169.8 177.5 164.4 170.7 176.2 168.6 178.8 162 160.2 171.2.
* Karena tidak ada data yang tepat berada di tengah kedua bagian tersebut maka adalah rata-rata dari dua data yang paling dekat dengan tengah pembagi tersebut.
* Artinya nilai ada di antara data ke-10 dan ke-11. Oleh karena itu kita perlu menghitung nilai rata-rata dari kedua data tersebut. Data ke-10 adalah 167.0 dan data ke-11 adalah 167.2. Maka = (167.0+167.2)/2 = 167.1.
* Kita sebut 50% dari data yang nilainya lebih kecil dari adalah data sebelah kiri, sedangkan data yang lebih besar dari adalah data sebelah kanan. Untuk menghitung nilai kita bisa lakukan hal yang sama seperti pada perhitungan tapi hanya pada data sebelah kiri. Karena data sebelah kiri ada n\_1 = 10 maka
* Kita juga bisa menghitung dari data sebelah kiri, ada data.
* Dari kedua perhitungan tersebut kita dapat menyimpulkan bahwa nilai ada di antara data ke-5 dan ke-6, yaitu 162.7 dan 164.4. Rata-rata dari kedua data tersebut adalah = 163.55.
* Selanjutnya menghitung adalah data pada urutan ke-75% .
* Atau data pada urutan ke 50% dari data di sebelah kanan.
* Jadi nilai adalah rata-rata dari data urutan ke-15 dan ke-16 dari seluruh data yang sudah diurutkan atau data ke-5 dan ke-6 dari data sebelah kanan. Kedua data tersebut adalah 170.7 dan 174.8, sehingga = (170.7+174.8)/2 = 172.75.
* Tahap pertama ini cukup panjang karena kita harus menghitung tiga nilai, yaitu , dan . Namun jika data yang kita miliki hanya sedikit maka kita dapat secara mudah hanya dengan melihat pada data yang sudah diurutkan.

1. Menghitung Inter-Quartil Range (IQR)  
   Inter-Quartil Range (IQR) diperoleh dengan menghitung selisih antara dengan , sehingga
2. Menghitung batas bawah dan batas atas  
   Batas bawah dan batas atas digunakan untuk membuat batasan penentuan suatu data dianggap sebagai pencilan atau outlier.

* Data yang lebih kecil dari batas bawah atau lebih besar dari batas atas dianggap pencilan/outlier. Pada contoh ini semua data berada dalam rentang batas bawah dan batas atas sehingga tidak ada data pencilan/outlier.

1. Membuat visualisasi boxplot  
   Pada ilustrasi boxplot di atas, boxplot terbentuk dari kotak dan garis. Kotak tersebut dibentuk dari nilai dan , sedangkan garis di tengah kotak dalah . Garis di sebelah kiri kotak dibentuk dengan menarik garis dari ujung kotak sebelah kiri () ke nilai batas bawah. Garis di sebelah kanan kotak dibentuk dengan menarik garis dari ujung kotak sebelah kanan () ke nilai batas atas. Jika ada data di luar batas bawah atau batas atasImaka ditandai dengan titik. Boxplot dapat dibuat dengan posisi horizontal atau vertikal. Berikut ini visualisasi boxplot yang dibuat dalam posisi horizontal dan vertikal.



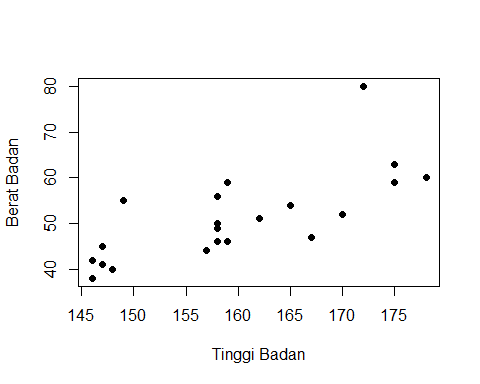


### Scatter Plot

Scatter plot atau diagram pencar atau plot tebaran, digunakan untuk melihat hubungan antar dua variabel numerik. Jika kita ingin mengecek apakah ada hubungan antara tinggi badan dan berat badan, scatter plot dapat menunjukkan apakah ada pola tertentu. Bentuk visual dari scatter plot berupa titik-titik berdasarkan koordinat nilai berpasangan antara variabel satu dan variabel kedua. Tujuan dari scatter plot adalah untuk melihat pola hubungan antar variabel. Jika dua variabel tersebut memiliki hubungan sebab-akibat atau sejenisnya, maka umumnya variabel yang menjadi penyebab ditempatkan pada sumbu horizontal dan variabel yang menjadi sebab ditempatkan pada sumbu vertikal.

|  |
| --- |
| Ingat! |
| Scatter plot hanya menunjukkan pola hubungan antara dua variabel numerik, tidak secara tegas menunjukkan sebab-akibat antara dua variabel. |

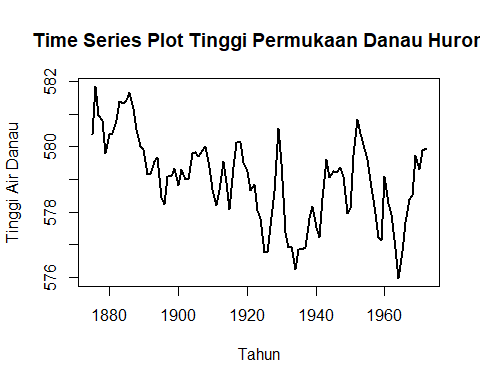
Berikut ini contoh visualisasi antara tinggi badan dan berat badan.



Dalam beberapa kasus, kita juga bisa menggunakan visualisasi yang lebih kompleks seperti heatmap untuk melihat korelasi antar banyak variabel atau pairplot untuk membandingkan distribusi dan hubungan antar beberapa variabel sekaligus.

### Line Chart

Line chart atau diagram garis adalah salah satu jenis visualisasi data yang paling umum dan digunakan untuk menampilkan perubahan atau tren dari data secara kontinu sepanjang waktu atau urutan. Biasanya, line chart menampilkan hubungan antara dua variabel, di mana sumbu horizontal (sumbu X) mewakili variabel waktu atau urutan, sedangkan sumbu vertikal (sumbu Y) mewakili nilai dari variabel yang diukur. Berikut ini contoh visualisasi ketinggian muka air danau Huron yang diukur sekali setiap tahun. Variabel Tahun menjadi urutan waktu dan ditempatkan pada sumbu X, variabel Tinggi Air Danau menjadi variabel yang diukur dan ditempatkan pada sumbu Y.



Beberapa poin penting tentang line chart:

1. ***Menggambarkan Tren Waktu***: Line chart sangat cocok untuk menampilkan data yang terorganisir berdasarkan waktu (seperti harian, mingguan, bulanan, atau tahunan). Misalnya, tren penjualan tahunan atau perubahan suhu selama beberapa hari.
2. ***Perbandingan Antara Beberapa Seri Data***: kamu juga dapat menggunakan line chart untuk membandingkan beberapa set data dalam satu grafik. Misalnya, membandingkan tren penjualan dua produk yang berbeda dalam satu tahun.
3. ***Sederhana dan Efektif***: Line chart mudah dibaca dan memberikan gambaran visual yang jelas tentang bagaimana nilai suatu variabel berubah dari waktu ke waktu. Ini menjadikannya alat yang kuat untuk memahami pola atau fluktuasi data.

Contohnya, jika kamu memiliki data penjualan bulanan selama satu tahun, kamu dapat memplotkan waktu pada sumbu X (bulan) dan nilai penjualan pada sumbu Y. Garis yang menghubungkan titik-titik data akan menunjukkan bagaimana penjualan meningkat atau menurun selama waktu tersebut. Diagram garis sering digunakan dalam laporan bisnis, penelitian ilmiah, dan berbagai aplikasi data analitik untuk membantu melihat perubahan dan membuat keputusan berdasarkan pola yang terlihat.

## Contoh Kasus Eksplorasi dan Visualisasi

Mari kita lihat sebuah contoh kasus sederhana. Misalkan kita bekerja dengan data penjualan sebuah toko online, yang mencakup jumlah produk terjual per hari, harga per produk, dan total pendapatan harian. Langkah pertama adalah melakukan eksplorasi untuk melihat tren penjualan secara keseluruhan. Dengan menggunakan statistik dasar seperti rata-rata penjualan harian, kita bisa mengetahui apakah ada hari-hari tertentu di mana penjualan lebih tinggi dari biasanya. Kemudian, kita bisa membuat visualisasi seperti line plot untuk menggambarkan tren penjualan dari waktu ke waktu, serta scatter plot untuk melihat apakah ada korelasi antara harga produk dan jumlah penjualan.Praktik eksplorasi data retail akan dibahas di [Bab 23](#sec-r-explore) dan [Bab 26](#sec-proyek1).

Dengan langkah-langkah ini, kita tidak hanya memahami data secara mendalam, tetapi juga mampu menyajikan informasi tersebut dengan cara yang mudah dipahami oleh pihak lain. Eksplorasi dan visualisasi adalah dua komponen yang tidak terpisahkan dalam dunia analisis data, keduanya membantu kita dalam mengambil keputusan yang lebih baik berdasarkan data yang ada.

# Sebaran Peubah Acak

Dalam statistik, peubah acak adalah variabel yang nilainya dihasilkan oleh proses acak. Misalnya, ketika melempar sebuah dadu, hasil yang muncul (angka 1 hingga 6) merupakan contoh dari peubah acak diskrit, karena hasilnya berupa bilangan bulat yang tetap. Sebaliknya, tinggi badan seseorang bisa dianggap sebagai peubah acak kontinu, karena bisa memiliki nilai dalam rentang tertentu, bukan hanya angka bulat.

Secara formal, peubah acak dibagi menjadi dua jenis: peubah acak diskrit dan peubah acak kontinu. Peubah acak diskrit memiliki sejumlah kemungkinan hasil yang terbatas dan bisa dihitung, seperti jumlah kendaraan yang lewat di suatu jalan pada hari tertentu. Sedangkan peubah acak kontinu memiliki rentang nilai yang tak terhingga, seperti berat badan atau suhu udara.

## Sebaran Peluang

Setiap peubah acak memiliki pola sebaran atau sebaran peluang, yang menunjukkan bagaimana peluang disebarankan di antara nilai-nilai yang mungkin. Untuk peubah acak diskrit, sebaran peluang bisa dijelaskan dengan Fungsi Massa Peluang (FMP), sedangkan untuk peubah acak kontinu, kita menggunakan Fungsi Kepekatan Peluang (FKP) atau Probability Density Function (PDF).

Sebagai contoh, untuk peubah acak diskrit seperti hasil lemparan koin, FMP memberikan peluang untuk masing-masing hasil (misalnya, peluang munculnya angka atau gambar). Sedangkan untuk peubah acak kontinu seperti tinggi badan, FKP memberikan kepadatan peluang untuk berbagai interval nilai, dan peluang untuk suatu rentang dihitung dengan mengintegrasikan FKP di sepanjang rentang tersebut.

## Contoh Sebaran Peubah Acak

Beberapa contoh sebaran yang sering digunakan dalam dunia data analysis meliputi:

1. **Sebaran Binomial**: Sebaran ini digunakan untuk memodelkan jumlah keberhasilan dalam serangkaian percobaan yang independen. Misalnya, jika kita ingin menghitung peluang munculnya kepala dalam 10 lemparan koin, kita bisa menggunakan sebaran binomial.
2. **Sebaran Normal**: Sebaran ini adalah salah satu yang paling umum digunakan dalam analisis data. sebaran normal digunakan untuk memodelkan banyak fenomena alam seperti tinggi badan, berat badan, dan IQ. Ciri khas dari sebaran ini adalah bentuknya yang simetris dengan mean, median, dan modus yang sama.
3. **Sebaran t-Student**: Sebaran ini adalah salah satu yang paling umum digunakan dalam analisis data. sebaran normal digunakan untuk memodelkan banyak fenomena alam seperti tinggi badan, berat badan, dan IQ. Ciri khas dari sebaran ini adalah bentuknya yang simetris dengan mean, median, dan modus yang sama.
4. **Sebaran Poisson** (dibaca [pwasɔ̃]: poassong): Sebaran ini digunakan untuk memodelkan jumlah kejadian dalam interval waktu atau ruang tertentu, misalnya, jumlah pelanggan yang datang ke sebuah restoran dalam satu jam.

Pemahaman yang mendalam mengenai sebaran peubah acak sangat penting bagi seorang data analyst. Sebagai seorang data analyst, Anda akan sering berhadapan dengan data yang berasal dari berbagai jenis sebaran peluang. Memahami karakteristik setiap sebaran akan membantu Anda dalam membuat inferensi yang lebih akurat, memodelkan ketidakpastian, dan pada akhirnya, menghasilkan analisis data yang lebih valid.

Dengan pengetahuan ini, seorang data analyst bisa lebih percaya diri dalam menavigasi proses pengambilan keputusan berbasis data, karena mereka memahami pola sebaran yang mendasari data yang mereka analisis.

## Sebaran Binomial

**Sebaran Binomial** adalah salah satu sebaran peluang yang paling umum digunakan dalam statistik, khususnya dalam memodelkan situasi di mana hasil percobaan bisa dibagi menjadi dua kemungkinan hasil: **sukses** dan **gagal**. Sebaran ini sangat berguna dalam berbagai situasi di dunia nyata, seperti menghitung peluang keberhasilan dalam serangkaian percobaan independen dengan hasil biner (dua hasil).

### Karakteristik Sebaran Binomial

Beberapa karakteristik utama dari sebaran binomial adalah:

1. **Percobaan Berulang**: Sebaran ini mengasumsikan ada serangkaian percobaan yang identik, dan setiap percobaan hanya memiliki dua hasil yang mungkin.
2. **Independensi**: Hasil dari setiap percobaan tidak mempengaruhi hasil percobaan lainnya. Artinya, setiap percobaan berdiri sendiri dan tidak ada ketergantungan antar percobaan.
3. **Peluang Tetap**: Peluang sukses tetap sama untuk setiap percobaan, begitu juga dengan peluang gagal .

### Fungsi Massa Peluang Binomial

Fungsi massa peluang (FMP) atau Probability Mass Function (PMF) dari sebaran binomial memberikan peluang bahwa kita akan mendapatkan sejumlah sukses tertentu dalam percobaan. Jika adalah peubah acak yang mengikuti sebaran binomial dengan parameter (jumlah percobaan) dan (peluang sukses), maka peluang , yaitu mendapatkan sukses dari percobaan, diberikan oleh rumus:

di mana:

* adalah jumlah percobaan,
* adalah jumlah sukses yang kita hitung,
* adalah peluang sukses dalam satu percobaan,

adalah koefisien binomial atau kombinasi, yang dihitung sebagai:

dengan dan .

Rumus ini menghitung berapa banyak cara kita bisa mendapatkan sukses dari percobaan, dikalikan dengan peluang bahwa tepat dari percobaan akan berhasil.

Misalnya dan , maka

### Contoh Aplikasi Sebaran Binomial

Bayangkan kita melakukan eksperimen dengan melempar koin yang seimbang (fair coin) sebanyak 10 kali. Kita ingin mengetahui peluang mendapatkan tepat 6 kali kepala (sukses) dari 10 lemparan.

Dalam hal ini:

* (jumlah lemparan),
* (peluang munculnya kepala dalam satu lemparan, karena koinnya seimbang),
* (jumlah kepala yang diinginkan).

Dengan menggunakan rumus di atas:

Perhitungan ini menghasilkan peluang bahwa kita akan mendapatkan tepat 6 kali kepala dari 10 lemparan sebesar atau sekitar .

### Parameter Sebaran Binomial

Sebaran binomial memiliki dua parameter yang penting, yaitu:

* **Mean** (rata-rata): . Rata-rata jumlah sukses dalam percobaan adalah , yaitu jumlah percobaan dikalikan dengan peluang sukses dalam satu percobaan.
* **Varian**: . Varian menunjukkan seberapa banyak penyebaran hasil dari rata-rata. Semakin besar atau , semakin besar juga varian sebaran.

### Kegunaan Sebaran Binomial

Sebaran binomial sangat berguna dalam banyak situasi nyata, terutama ketika kita ingin memodelkan kejadian yang hanya memiliki dua kemungkinan hasil, seperti:

* **Survei dan Eksperimen**: Sebaran binomial dapat digunakan untuk menghitung peluang seseorang memberikan jawaban tertentu dalam survei (misalnya, “ya” atau “tidak”), atau peluang keberhasilan suatu eksperimen.
* **Uji Hipotesis**: Dalam uji hipotesis, sebaran binomial bisa digunakan untuk menguji apakah suatu hasil eksperimen berbeda secara signifikan dari yang diharapkan berdasarkan peluang sukses yang diketahui.
* **Analisis Risiko**: Banyak model risiko menggunakan sebaran binomial untuk menghitung peluang dari berbagai kejadian biner yang mungkin terjadi, seperti kebangkrutan perusahaan atau suksesnya suatu proyek.

## Sebaran Normal

**Sebaran Normal** adalah salah satu sebaran peluang yang paling penting dan umum dalam statistika. Sebaran ini sering digunakan untuk memodelkan fenomena di alam dan dalam berbagai bidang, termasuk psikologi, ekonomi, ilmu sosial, dan banyak lagi. Sebaran normal juga dikenal sebagai sebaran Gaussian, dinamai setelah matematikawan Carl Friedrich Gauss yang pertama kali memperkenalkan konsep ini.

### Karakteristik Sebaran Normal

1. **Bentuk Kurva**: Sebaran normal memiliki bentuk kurva lonceng yang simetris di sekitar mean (rata-rata). Kurva ini ditandai dengan dua parameter utama: mean (, dibaca myu) dan deviasi standar (, dibaca sigma).
   * **Mean** (): Titik pusat sebaran, di mana nilai-nilai tertinggi dari peluang terletak.
   * **Deviasi Standar** (): Mengukur sebaran data di sekitar mean. Deviasi standar yang lebih kecil berarti data lebih terpusat di sekitar mean, sementara deviasi standar yang lebih besar berarti data lebih tersebar.
2. **Simetri**: Sebaran normal adalah simetris. Hal ini berarti bahwa peluang untuk mendapatkan nilai yang lebih besar atau lebih kecil dari mean adalah sama.
3. **Aturan Empiris (68-95-99.7)**:
   * Sekitar 68% dari data dalam sebaran normal terletak dalam satu deviasi standar ().
   * Sekitar 95% dari data terletak dalam dua deviasi standar ().
   * Sekitar 99.7% dari data terletak dalam tiga deviasi standar ().

### Fungsi Kepekatan Peluang (FKP)

Fungsi kepekatan peluang (FKP) untuk sebaran normal diberikan oleh rumus berikut:

di mana:

* adalah nilai kepadatan peluang pada titik ,
* adalah mean,
* adalah deviasi standar,
* adalah bilangan Euler (sekitar 2.71828),
* adalah bilangan pi (sekitar 3.14159).

### Transformasi Z

Dalam analisis statistik, sering kali lebih mudah untuk bekerja dengan sebaran normal standar, yang memiliki mean 0 dan deviasi standar 1. Untuk mentransformasikan nilai ke dalam sebaran normal standar, kita menggunakan rumus berikut:

di mana:

* adalah nilai standar,
* adalah nilai asli,
* adalah mean,
* adalah deviasi standar.

### Contoh Aplikasi Sebaran Normal

1. **Analisis Data**: Banyak fenomena alami, seperti tinggi badan, berat badan, dan nilai ujian, mengikuti sebaran normal. Misalnya, jika tinggi badan orang dewasa di suatu populasi memiliki mean 170 cm dengan deviasi standar 10 cm, kita dapat menggunakan sebaran normal untuk menghitung peluang seseorang memiliki tinggi badan antara 160 cm dan 180 cm.
2. **Statistika Inferensial**: Sebaran normal digunakan dalam pengujian hipotesis dan interval kepercayaan. Banyak metode statistik, seperti analisis regresi dan ANOVA, mengasumsikan bahwa data mengikuti sebaran normal.
3. **Kontrol Kualitas**: Dalam industri, sebaran normal digunakan untuk mengontrol kualitas produk. Misalnya, ukuran produk yang diproduksi dapat dianalisis untuk memastikan bahwa sebagian besar produk berada dalam toleransi yang ditentukan.

## Sebaran t-Student

**Sebaran t-Student**, atau sering disebut sebagai **sebaran t**, adalah sebran peluang yang digunakan dalam statistik inferensial, terutama ketika ukuran sampel kecil dan varians populasi tidak diketahui. Sebaran ini diperkenalkan oleh **William Sealy Gosset**, yang menggunakan nama samaran “Student” untuk publikasinya, sehingga distribusi ini sering disebut sebagai sebaran **t-Student**.

Sebaran t sangat mirip dengan sebaran normal, namun memiliki ekor yang lebih tebal (lebih lebar). Artinya, dalam sebaran t, peluang kejadian ekstrem (nilai yang jauh dari rata-rata) lebih tinggi dibandingkan dengan sebaran normal. Sebaran t digunakan terutama untuk mengestimasi mean suatu populasi kecil dan untuk pengujian hipotesis, seperti uji t (t-test).

### Karakteristik Sebaran t

1. **Ekor Lebih Tebal**: Sebaran t memiliki ekor yang lebih tebal dibandingkan dengan sebaran normal. Ini berarti probabilitas observasi ekstrem (nilai yang jauh dari mean) lebih besar dalam sebaran t, yang penting ketika bekerja dengan sampel kecil di mana ada lebih banyak ketidakpastian.
2. **Derajat Kebebasan** (*Degrees of Freedom*, ): Parameter kunci dari sebaran t adalah derajat kebebasan. Semakin besar ukuran sampel (semakin besar ), semakin mirip sebaran t dengan sebaran normal. Sebaliknya, dengan yang lebih kecil (misalnya, sampel yang sangat kecil), ekor sebaran t akan semakin tebal. Derajat kebebasan biasanya dihitung sebagai , di mana adalah ukuran sampel.
3. **Mean dan Varians**:
   * Mean (rata-rata) dari sebaran t adalah 0, sama seperti pada sebaran normal standar.
   * Variansnya lebih besar daripada sebaran normal, terutama untuk nilai kecil, yang mencerminkan ketidakpastian yang lebih besar dalam estimasi ketika sampel kecil.

### Fungsi Kepekatan Peluang (FKP) Sebaran t

Fungsi kepekatan peluang (FKP) untuk sebaran t-Student adalah:

di mana:

* adalah variabel acak yang mengikuti sebaran t,
* df adalah derajat kebebasan,
* adalah fungsi gamma (*generalization of factorial*).

Meskipun rumus ini tampak rumit, dalam praktiknya, penggunaan sebaran t lebih sering dilakukan dengan bantuan tabel t atau perangkat lunak statistik seperti R, Python, SAS, SPSS, dan lainnya, yang secara otomatis menghitung nilai probabilitas terkait dengan nilai t.

### Perbandingan Sebaran t dengan Sebaran Normal

Sebaran t sangat mirip dengan sebaran normal, tetapi ada beberapa perbedaan kunci:

1. **Ukuran Sampel**: Sebaran t digunakan ketika ukuran sampel kecil () dan varians populasi tidak diketahui. Sebaran normal digunakan ketika ukuran sampel besar atau varians populasi diketahui.
2. **Bentuk Kurva**: Sebaran t lebih lebar dan memiliki ekor yang lebih tebal dibandingkan dengan sebaran normal. Ini mengakomodasi variabilitas yang lebih besar karena sampel kecil.
3. **Derajat Kebebasan**: Sebaran t bergantung pada derajat kebebasan. Ketika meningkat (misalnya, ukuran sampel meningkat), sebaran t mendekati sebaran normal. Sebagai contoh, untuk sangat besar (misalnya ), sebaran t hampir identik dengan sebaran normal standar.

### Aplikasi Sebaran t

Sebaran t terutama digunakan dalam statistik inferensial, khususnya dalam pengujian hipotesis dan analisis interval kepercayaan. Berikut adalah beberapa aplikasi umum:

1. **Uji t Satu Sampel** (*One-Sample t-Test*): Uji ini digunakan untuk menguji apakah rata-rata sampel secara signifikan berbeda dari nilai tertentu ketika varians populasi tidak diketahui dan ukuran sampel kecil. Sebagai contoh, Anda mungkin ingin menguji apakah rata-rata tinggi badan suatu populasi berbeda dari 170 cm berdasarkan sampel kecil.
2. **Uji t Dua Sampel** (*Two-Sample t-Test*): Uji ini digunakan untuk membandingkan rata-rata dua kelompok yang berbeda. Misalnya, membandingkan rata-rata gaji pria dan wanita dalam suatu perusahaan berdasarkan sampel kecil.
3. **Uji t Berpasangan** (*Paired t-Test*): Digunakan untuk membandingkan rata-rata dua pengukuran yang berhubungan (misalnya, sebelum dan sesudah perlakuan dalam eksperimen).
4. **Interval Kepercayaan**: Sebaran t digunakan untuk menghitung interval kepercayaan untuk rata-rata populasi ketika varians populasi tidak diketahui dan ukuran sampel kecil. Interval kepercayaan memberi rentang di mana rata-rata populasi mungkin terletak dengan tingkat kepercayaan tertentu (misalnya 95%).

## Sebaran Poisson

**Sebaran Poisson** adalah salah satu sebaran peluang yang digunakan untuk memodelkan jumlah kejadian dalam suatu interval waktu atau ruang yang tetap, di mana kejadian tersebut bersifat acak dan terjadi secara independen satu sama lain. Sebaran ini sangat berguna dalam situasi di mana kita menghitung jumlah kejadian tertentu, seperti jumlah pelanggan yang datang ke restoran dalam satu jam, jumlah kecelakaan di suatu jalan raya dalam sehari, atau jumlah panggilan telepon yang diterima oleh sebuah pusat layanan dalam satu jam.

### Karakteristik Sebaran Poisson

1. **Jumlah Kejadian dalam Interval Tetap**: Sebaran Poisson memodelkan jumlah kejadian yang terjadi dalam interval waktu atau ruang yang tetap, seperti jumlah kendaraan yang lewat di jalan raya dalam satu jam, atau jumlah bintang dalam suatu area tertentu di langit.
2. **Independensi Kejadian**: Kejadian-kejadian yang dihitung haruslah bersifat independen. Ini berarti bahwa kejadian satu tidak memengaruhi kemungkinan terjadinya kejadian yang lain.
3. **Tingkat Kejadian Tetap** (*Rate of Occurrence*): Tingkat kejadian, atau rata-rata jumlah kejadian dalam interval yang ditetapkan, harus tetap. Tingkat kejadian ini sering disebut sebagai (dibaca lambda) , yang merupakan parameter utama dari distribusi Poisson.
4. **Kejadian Tunggal**: Dalam sebaran Poisson, hanya satu kejadian yang dapat terjadi pada waktu atau ruang yang sangat kecil (tidak ada dua kejadian yang terjadi pada saat yang bersamaan dalam periode yang sangat pendek).

### Fungsi Massa Peluang Poisson

Fungsi Massa Peluang (FMP) dari sebaran Poisson memberikan peluang bahwa akan ada kejadian dalam suatu interval, ketika rata-rata tingkat kejadian adalah . Rumus FMP Poisson adalah:

​

di mana:

* adalah probabilitas terjadinya kejadian dalam interval,
* adalah rata-rata jumlah kejadian per interval waktu atau ruang,
* adalah bilangan Euler (),
* adalah jumlah kejadian yang terjadi (bilangan bulat non-negatif),
* adalah faktorial dari .

### Parameter Sebaran Poisson

Sebaran Poisson memiliki hanya satu parameter, yaitu , yang merupakan rata-rata atau tingkat kejadian per unit interval. Rata-rata dan varian dari sebaran Poisson keduanya sama dengan , yaitu:

* **Mean** (Rata-rata):
* **Varian**:

Ini menunjukkan bahwa dalam sebaran Poisson, deviasi standar (akar kuadrat dari varian) juga bergantung pada .

### Contoh Aplikasi Sebaran Poisson

1. **Jumlah Panggilan Telepon**: Misalkan sebuah pusat panggilan menerima rata-rata 5 panggilan per menit (). Kita bisa menggunakan sebaran Poisson untuk menghitung probabilitas bahwa mereka akan menerima tepat 3 panggilan dalam satu menit.

Menggunakan rumus FMP:

​

Ini akan memberikan peluang untuk menerima 3 panggilan dalam satu menit.

1. **Kecelakaan di Jalan Raya**: Jika diketahui bahwa di suatu jalan raya rata-rata terjadi 2 kecelakaan setiap minggu (), kita bisa menggunakan sebaran Poisson untuk menghitung peluang bahwa minggu ini tidak akan ada kecelakaan () atau bahwa akan terjadi lebih dari 3 kecelakaan ().
2. **Jumlah Pelanggan yang Datang ke Restoran**: Jika sebuah restoran menerima rata-rata 10 pelanggan per jam, sebaran Poisson bisa digunakan untuk memodelkan jumlah pelanggan yang datang dalam satu jam, dan menghitung peluang berbagai skenario, seperti kedatangan tepat 15 pelanggan atau kurang dari 5 pelanggan dalam satu jam.

### Hubungan dengan Distribusi Lain

Sebaran Poisson sering kali muncul sebagai limit dari sebaran binomial ketika jumlah percobaan () sangat besar dan peluang sukses () sangat kecil, tetapi nilai rata-rata () tetap konstan. Dalam hal ini, sebaran binomial dapat diaproksimasi dengan sebaran Poisson, di mana .

### Kapan Menggunakan Sebaran Poisson?

Sebaran Poisson digunakan ketika:

* Kita menghitung jumlah kejadian dalam suatu interval waktu atau ruang (misalnya jumlah pelanggan, kecelakaan, atau panggilan telepon).
* Tingkat kejadian per unit waktu atau ruang bersifat tetap dan diketahui ().
* Kejadian-kejadian tersebut terjadi secara acak dan independen satu sama lain.
* Tidak ada batas maksimum untuk jumlah kejadian yang dapat terjadi (meskipun secara praktis, nilai yang sangat besar mungkin jarang terjadi).

### Contoh Kasus

Misalkan di sebuah bank, rata-rata ada 7 nasabah yang datang setiap jam (). Kita ingin mengetahui peluang bahwa dalam satu jam, ada tepat 5 nasabah yang datang. Dengan menggunakan rumus Poisson:

Jadi, peluang bahwa tepat 5 nasabah datang dalam satu jam adalah sekitar 12.8%.

# Uji Hipotesis

Dalam dunia data, uji hipotesis adalah salah satu teknik yang sangat sering digunakan untuk menentukan apakah data yang kita kumpulkan cukup untuk mendukung atau menolak suatu dugaan. Sama seperti seorang detektif yang mencari bukti di tempat kejadian, kita juga menggunakan uji hipotesis untuk “membaca” data dan mencari tahu apakah hipotesis yang kita buat benar atau tidak.

Pada bagian ini, kita akan membahas beberapa jenis uji hipotesis yang sering digunakan dengan sebaran yang berbeda: binomial, normal, t-student, dan Poisson. Setiap distribusi memiliki kegunaan tertentu dalam situasi yang berbeda, tergantung pada jenis data yang kita miliki.

## Uji Hipotesis dengan Sebaran Binomial

Sebaran binomial digunakan ketika kita berurusan dengan data yang bersifat diskrit, biasanya untuk eksperimen yang hanya memiliki dua hasil, seperti **sukses** atau **gagal**.

**Contoh kasus**:  
Katakanlah kamu bekerja di perusahaan e-commerce dan ingin menguji apakah tingkat keberhasilan pengiriman barang tetap pada 95%. Hipotesis nolnya () adalah bahwa tingkat keberhasilan pengiriman sama dengan 95%, sedangkan hipotesis alternatifnya () adalah tingkat keberhasilan pengiriman berbeda dari 95%.

|  |
| --- |
| Dalam Matematika dan Statistika |
| Notasi hipotesis adalah dan . biasanya untuk hipotesis yang mengandung nilai “sama dengan” (=, , ), sedangkan adalah kebalikan atau negasinya (, , ).  Hipotesis dua arah:  Hipotesis satu arah: |

Misalnya, kamu mengambil sampel dari 100 () pengiriman dan mendapati bahwa 92 pengiriman berhasil. Kita bisa melakukan uji binomial untuk melihat apakah ada cukup bukti bahwa tingkat keberhasilan pengiriman telah berubah. Jika nilai p-value dari uji binomial lebih kecil dari tingkat signifikansi (misalnya 0.05), kita dapat menolak hipotesis nol dan menyimpulkan bahwa ada perubahan pada tingkat keberhasilan pengiriman.

Dengan menghitung sebaran binomial kumulatif (*Cumulative Distribution Function*, CDF) untuk , dan , kita dapat menuliskannya menjadi

Ingat bahwa

maka

Jika kita menggunakan taraf nyata (*significant level*, ) sebesar 5% atau , maka kita gagal menolak karena p-value > . Artinya kamu masih bisa mengatakan secara statistik bahwa tingkat keberhasilan pengiriman barang tetap pada 95%.

## Uji Hipotesis dengan Sebaran Normal

Distribusi normal adalah distribusi kontinu yang paling sering digunakan ketika data menyebar secara simetris di sekitar mean. Uji hipotesis yang menggunakan distribusi normal sering digunakan ketika ukuran sampel besar, dan kita ingin menguji apakah suatu rata-rata populasi berbeda dari nilai tertentu.

**Contoh kasus**:

Bayangkan kamu bekerja di sebuah perusahaan manufaktur yang memproduksi bola lampu. Hipotesis nol () menyatakan bahwa rata-rata umur bola lampu adalah 800 jam. Kamu mengambil sampel 50 bola lampu dan mendapati rata-rata umur sampel adalah 780 jam dengan standar deviasi 30 jam. Untuk melihat apakah ada perbedaan signifikan, kamu bisa menggunakan uji z (uji hipotesis untuk rata-rata dengan distribusi normal). Jika p-value lebih kecil dari tingkat signifikansi, kita dapat menyimpulkan bahwa umur rata-rata bola lampu mungkin berbeda dari 800 jam.

## Uji Hipotesis dengan Sebaran t-Student

Distribusi t-Student digunakan ketika ukuran sampel kecil dan varians populasi tidak diketahui. Biasanya digunakan untuk menguji rata-rata ketika data mengikuti distribusi normal atau mendekati normal, tetapi ukuran sampelnya terlalu kecil untuk menggunakan distribusi normal.

**Contoh kasus**:

Misalkan kamu bekerja di sebuah restoran cepat saji, dan kamu ingin tahu apakah waktu rata-rata penyajian burger lebih dari 5 menit. Kamu mengambil sampel kecil (misalnya 10 pesanan) dan mendapati bahwa rata-rata waktu penyajian adalah 5,3 menit dengan standar deviasi 0,4 menit. Karena ukuran sampel kecil, kita bisa menggunakan uji t-Student untuk menguji apakah rata-rata penyajian lebih lama dari 5 menit. Dengan menghitung statistik uji t, kita dapat membandingkan hasilnya dengan nilai kritis dari tabel distribusi t, dan membuat keputusan berdasarkan hasil tersebut.

## Uji Hipotesis dengan Sebaran Poisson

Distribusi Poisson digunakan ketika kita berurusan dengan data yang mewakili jumlah kejadian dalam suatu interval waktu atau ruang. Distribusi ini sering digunakan ketika data jarang terjadi, seperti jumlah kecelakaan lalu lintas per bulan atau jumlah keluhan pelanggan dalam satu hari.

**Contoh kasus**:

Kamu bekerja di sebuah perusahaan transportasi dan ingin mengetahui apakah rata-rata kecelakaan dalam sebulan telah meningkat. Misalkan hipotesis nol (H₀) adalah bahwa rata-rata kecelakaan per bulan adalah 3. Kamu mencatat bahwa dalam bulan terakhir, terjadi 5 kecelakaan. Dengan menggunakan uji Poisson, kita dapat menghitung apakah kejadian 5 kecelakaan ini cukup signifikan untuk menyimpulkan bahwa ada peningkatan kecelakaan. Jika hasil uji menunjukkan bahwa 5 kecelakaan jauh lebih tinggi dari yang diharapkan, kita dapat menolak hipotesis nol dan menyimpulkan bahwa ada peningkatan dalam rata-rata kecelakaan.

### Contoh Penggunaan Sebaran t

Contoh 1: Uji t Satu Sampel

Misalkan Anda ingin menguji apakah rata-rata waktu tidur mahasiswa berbeda dari 8 jam per malam. Anda mengambil sampel 15 mahasiswa dan mendapatkan rata-rata 7.5 jam tidur dengan standar deviasi sampel 0.8 jam. Dengan ukuran sampel kecil dan varians populasi yang tidak diketahui, Anda menggunakan uji t.

Langkah-langkah uji t:

1. Hipotesis nol (): Rata-rata tidur sama dengan 8 jam.
2. Hipotesis satu/alternatif ( atau ): Rata-rata tidur tidak sama dengan 8 jam.
3. Hitung statistik t:

Bandingkan nilai t dengan nilai kritis dari tabel t untuk 𝑑 𝑓 = 𝑛 − 1 = 14 df=n−1=14.

Contoh 2: Interval Kepercayaan Jika Anda ingin menghitung interval kepercayaan 95% untuk rata-rata tidur berdasarkan sampel 15 mahasiswa, Anda bisa menggunakan sebaran t untuk menghitung margin of error:

𝐶 𝐼 = 𝑥 ˉ ± 𝑡 𝛼 / 2 , 𝑑 𝑓 ⋅ 𝑠 𝑛 CI= x ˉ ±t α/2,df ​ ⋅ n ​

s ​

Dengan nilai 𝑡 𝛼 / 2 , 𝑑 𝑓 = 14 t α/2,df=14 ​ dari tabel t dan menggunakan nilai sampel, Anda dapat menghitung interval tersebut.

# Pengenalan Spreadsheet

# Tabel dan Formula

# Pivot Tabel

# Visualisasi Chart

# Database

# SELECT Clause

# Filtering

# Sorting

# Summary

In summary, this book has no content whatsoever.

1 + 1

[1] 2

# Join Tabel

# R Programming

# Eksplorasi Data

# Panduan Visualisasi

# Storytelling

# Contoh Use Case 1

# Contoh Use Case 2

# Membangun Portofolio

# Tips Mencari Pekerjaan

# Pustaka

# Tentang Penulis

**Aep Hidayatuloh**

Sejak lulus dari Program Sarjana Statistika di Institut Pertanian Bogor (IPB University), dia menempati beberapa posisi sebagai Data Analyst mulai dari level internship, junior, senior hingga Head of Data Analytics di beberapa perusahaan dan konsultan analisis data science. Sepanjang karirnya dia memiliki pengalaman luas terlibat di berbagai proyek analisis data di perbankan, telekomunikasi, instansi pemerintahan dan retail. Bahasa pemrograman R menjadi pilihan utama, Python sebagai bahasa pemrograman kedua dan juga kemampuan SAS. Dia mengajar di berbagai pelatihan, menjadi dosen tamu di beberapa universitas dan menjadi narasumber di berbagai webinar.

# Penutup

Buku ini disusun sedemikian rupa sehingga memudahkan bagi Anda yang ingin mulai mempelajari tentang Data Analytics dan menjadi seorang Data Analyst. Buku ini akan menjadi panduan yang lengkap.

Di awal buku ini Anda akan diajak untuk memahami cara berpikir seorang data analyst. Tentu saja Anda tidak bisa lepas dari kewajiban untuk memahami tentang Statistika sebagai fondasi mutlak untuk menjadi seorang data analyst. Tapi tenang saja! Pembahasan Statistika dibuat sederhana agar mudah dipahami dan tetap dapat digunakan sebagai bekal Anda menjadi data analyst.

Anda juga akan belajar menggunakan Spreadsheet sebagai alat bantu paling fundamental ketika bekerja dengan data, memahami SQL sebagai alat untuk berkomunikasi dengan data di dalam database, menggunakan bahasa pemrograman sebagai alat yang lebih advance untuk data analysis dan cara membuat visualisasi untuk mengomunikasikan hasil analisis data yang Anda lakukan.