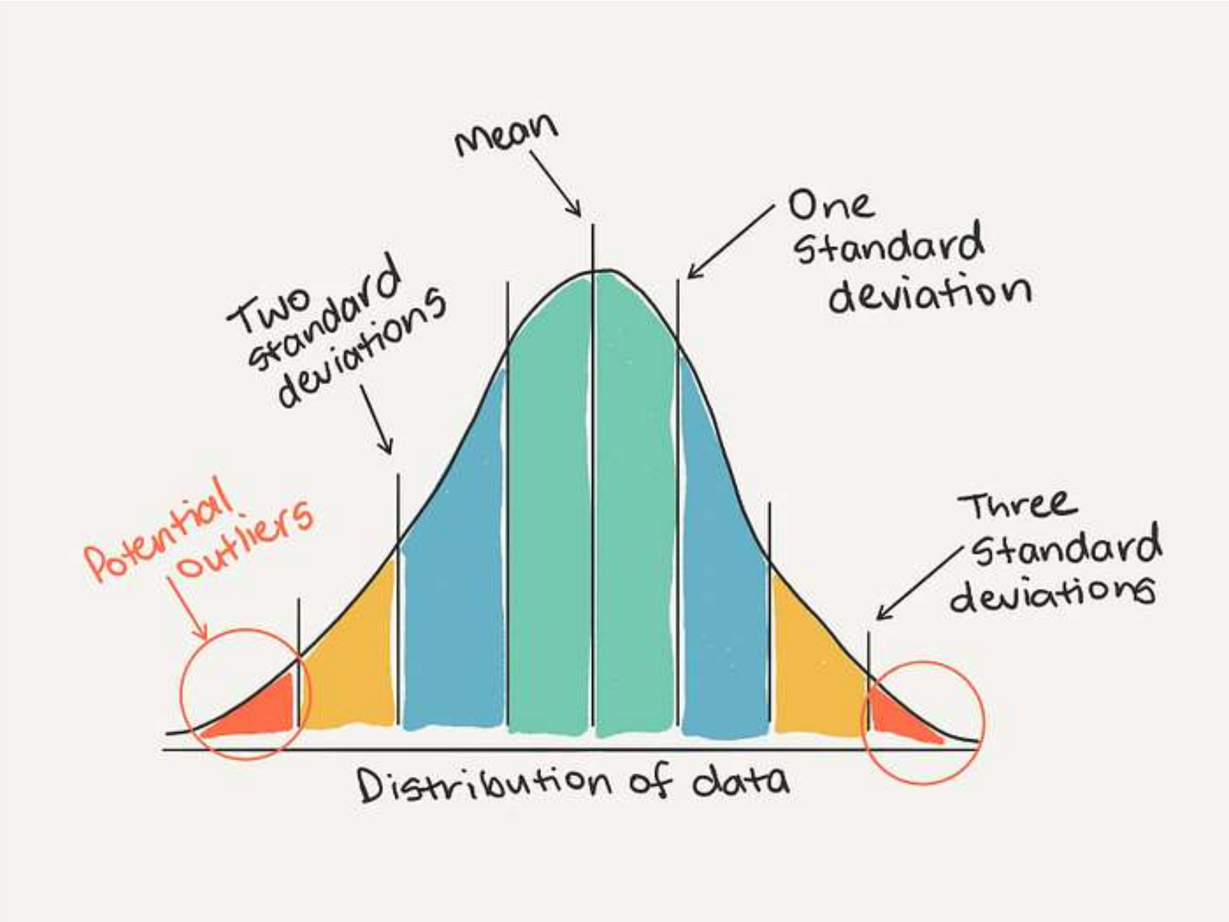
D A T A S C I E N C E
and
B U S I N E S S I N T E L L I G E N T

Author: Egi Safitri

Meeting 7

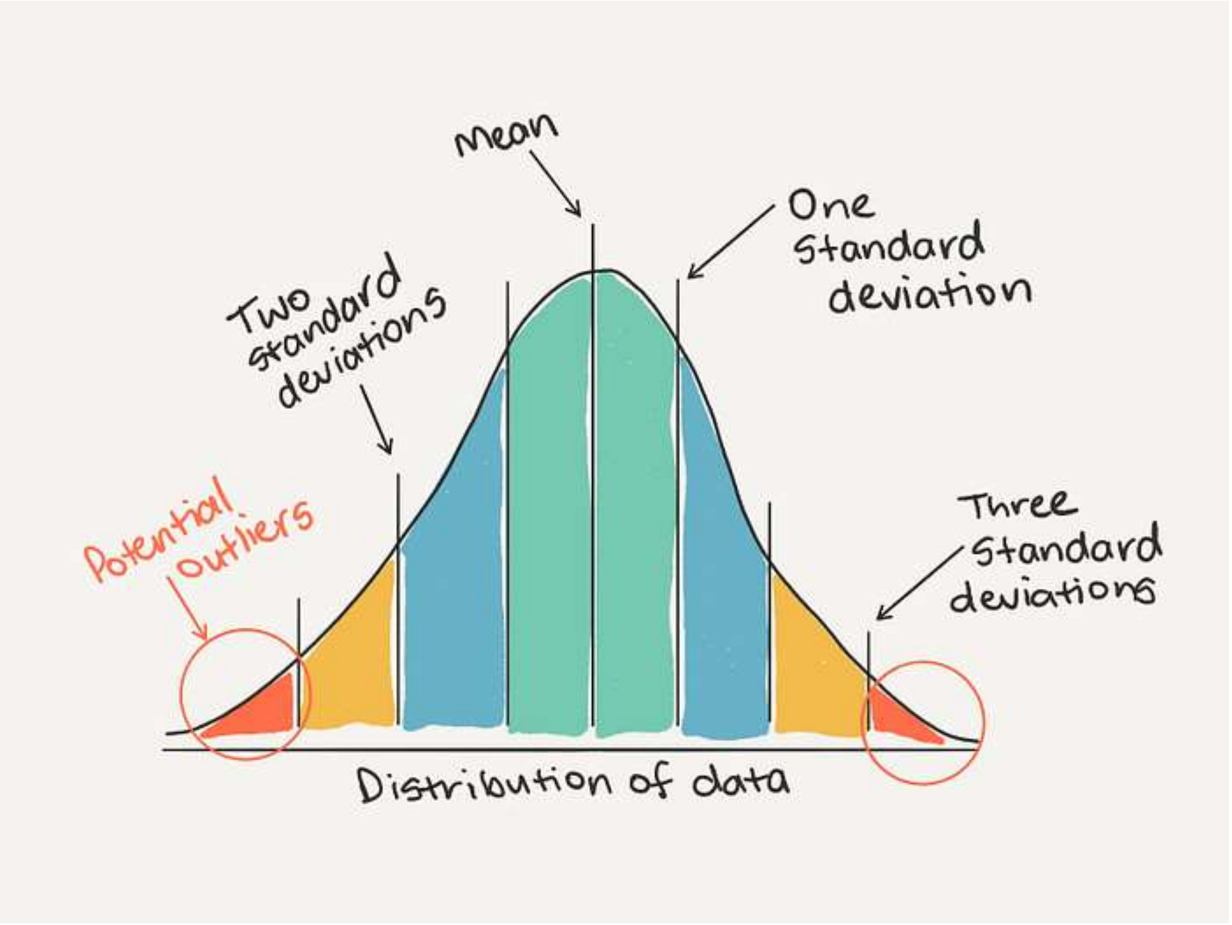
Apa itu EDA?

Secara sederhana Exploratory Data Analysis (EDA) adalah proses eksplorasi data yang bertujuan untuk memahami isi dan komponen penyusun data.



Apa itu EDA?

Eksplorasi data adalah cara untuk mengenal data sebelum mengolahnya. Melalui survei dan investigasi, kumpulan data berukuran besar disiapkan untuk analisis yang lebih mendalam dan terstruktur.

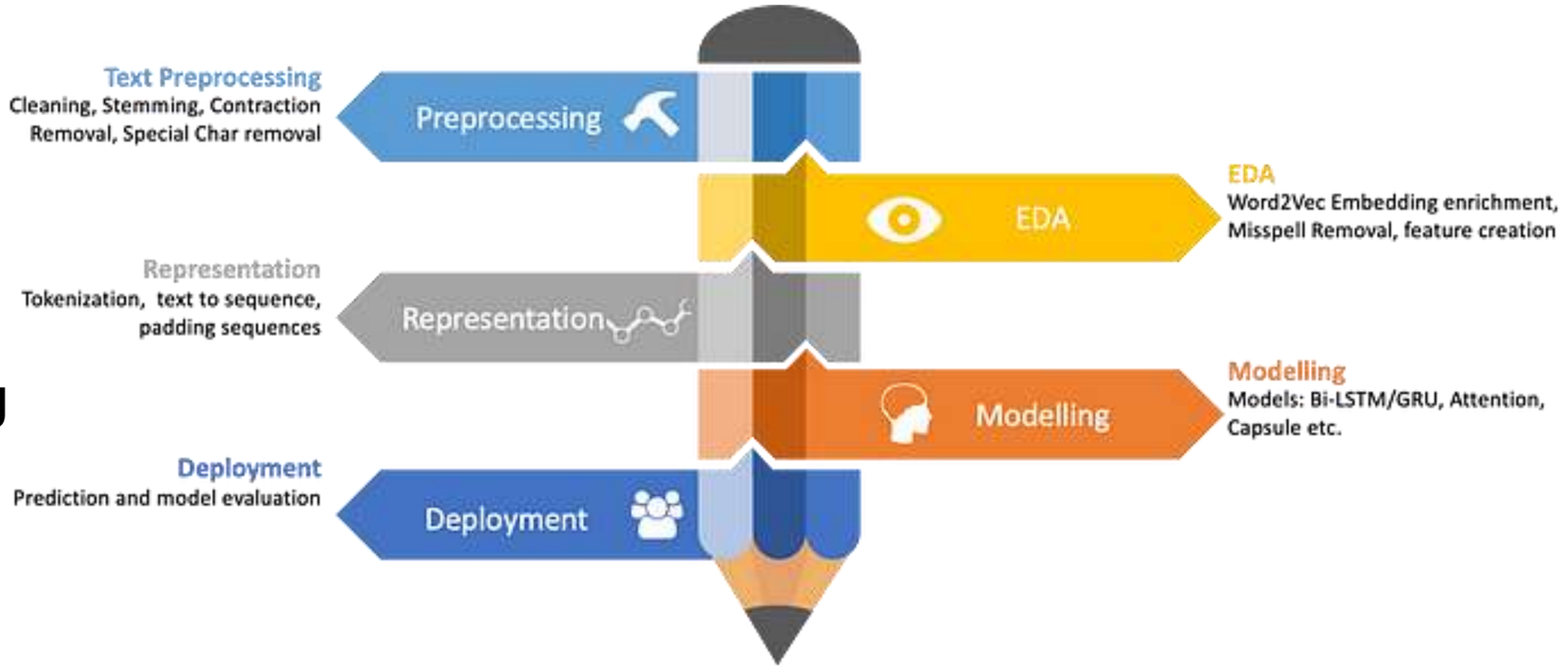


Apa itu EDA?

Sebagai gambaran, bayangkan anda bersama teman-teman berencana untuk berkumpul diwaktu libur, untuk menentukan tempat berkumpul, anda mulai untuk mendata beberapa cafe yang ada, kemudian anda akan membandingkan beberapa aspek dari setiap cafe yang ada untuk menentukan tempat yang paling cocok, mulai dari tempat mana yang memiliki pelayanan paling baik, harga yang terjangkau, kecepatan Wi-fi yang memadai dan memiliki spot foto terbaik, tidak berhenti disitu kemudian anda juga akan meminta beberapa saran dari teman-teman anda mengenai gambaran dari beberapa cafe yang sebelumnya anda tentukan. Setelah anda memiliki gambaran mengenai isi dan kualitas dari beberapa cafe anda akan lebih mudah untuk menentukan dimana anda dan bersama teman-teman anda berkumpul. Proses memahami aspek-aspek dari cafe tersebutlah yang disebut dengan EDA.

EDA dalam Data Science

Dalam kaitannya dengan data science, EDA dilakukan setelah tahap Preprocessing dan sebelum melakukan proses feature engineering dan modeling.



Jenis Eksplorasi Data

EDA dapat dibagi menjadi empat kategori berdasarkan grafis dan fokusnya. Berikut adalah 4 jenis Analisis Data Eksploratif.

1. Non-Grafis Univariat
2. Grafis Univariat
3. Non-Grafis Multivariat
4. Grafis Multivariat

Eksplorasi Data Vs Penambangan Data

Dalam ilmu data, ada dua metode utama untuk mengekstraksi data dari sumber berbeda, yakni eksplorasi data dan penambangan data.

Eksplorasi data adalah proses yang dilakukan oleh pebisnis untuk memahami tren dan pola data, dan dilakukan secara lebih luas.

Sementara penambangan data atau data mining yaitu proses yang lebih spesifik, biasanya dilakukan oleh para profesional data. Analis data membuat aturan dan parameter asosiasi untuk memilah kumpulan data yang sangat besar dan mengidentifikasi pola serta tren masa depan.

Mengutip situs *Techtarget*, biasanya eksplorasi data dilakukan terlebih dahulu untuk menilai hubungan antar variabel. Kemudian penambangan data dimulai. Melalui proses ini, model data dibuat untuk mengumpulkan wawasan tambahan dari data.

Cara kerja EDA

1. Memahami Variabel

Dasar setiap analisi data dimulai dengan memahami variabel. Melihat lebih detail katalog data, deskripsi bidang, dan metadata dapat memberikan wawasan tentang apa yang diwakili oleh setiap bidang dan membantu menemukan data yang hilang atau tidak lengkap.

Cara kerja EDA

2. Deteksi Outlier

Outlier atau anomali dapat menggagalkan analisis dan mendistorsi realitas kumpulan data, oleh karenanya penting utuk diidentifikasi sejak dini.Visualisasi data, metode numerik, rentang interkuartil, dan pengujian hipotesis adalah cara paling umum mendeteksi outlier. Plot kotak, histogram, atau plot sebar, misalnya, memudahkan untuk menemukan titik-titik yang jauh di luar rentang standar, sementara skor-z dapat menginformasikan seberapa jauh titik data dari rata-rata.Setelah ditemukan, seorang analis dapat menyelidiki, menyesuaikan, menghilangkan, atau mengabaikan outlier. Apa pun pilihannya, keputusan tersebut harus dicatat dalam analisis.

Cara kerja EDA

3. Periksa Pola dan Hubungan

Merencakan kumpulan data dalam berbagai cara memudahkan untuk mengidentifikasi dan memeriksa pola serta hubungan antar variabel.

Contoh EDA

Berikut ini contoh eksplorasi data dari sebuah penelitian pasar menggunakan data fiktif.

Tujuan Penelitian Pasar: Menganalisis preferensi konsumen terhadap tiga merek smartphone terkemuka di pasar.

Langkah 1: Pengumpulan Data

Pertama, data tentang preferensi konsumen terhadap tiga merek smartphone (A, B, dan C) dikumpulkan melalui survei. Responden diminta untuk memberikan peringkat 1 hingga 5 untuk setiap merek berdasarkan berbagai atribut seperti kualitas kamera, daya tahan baterai, harga, dan desain.

Contoh EDA

Langkah 2: Pembersihan Data

Data survei mungkin mengandung entri yang tidak valid atau hilang. Data ini perlu dibersihkan sebelum analisis. Misalnya, menghilangkan entri yang tidak lengkap atau tidak valid.

Contoh EDA

Langkah 3: Statistik Deskriptif

Setelah membersihkan data, kita dapat melakukan analisis statistik deskriptif untuk mendapatkan gambaran awal tentang preferensi konsumen. Contoh analisis statistik deskriptif dapat meliputi:Rata-rata peringkat untuk setiap merek pada setiap atribut.Grafik batang untuk menampilkan peringkat rata-rata berdasarkan merek.Perhitungan frekuensi peringkat tertentu pada atribut tertentu.

Contoh EDA

Langkah 4: Analisis Korelasi

Kita dapat melakukan analisis korelasi untuk melihat apakah terdapat korelasi antara atribut tertentu dan preferensi merek. Misalnya, kita dapat menggunakan korelasi Pearson untuk melihat apakah ada korelasi positif atau negatif antara harga dan preferensi merek.

Langkah 5: Analisis Segmentasi

Mungkin ada kelompok konsumen yang memiliki preferensi yang mirip. Analisis segmentasi dapat membantu mengidentifikasi kelompok-kelompok ini. Misalnya, kita dapat menggunakan analisis klaster untuk mengelompokkan responden berdasarkan preferensi mereka.

Contoh EDA

Langkah 6: Visualisasi Data

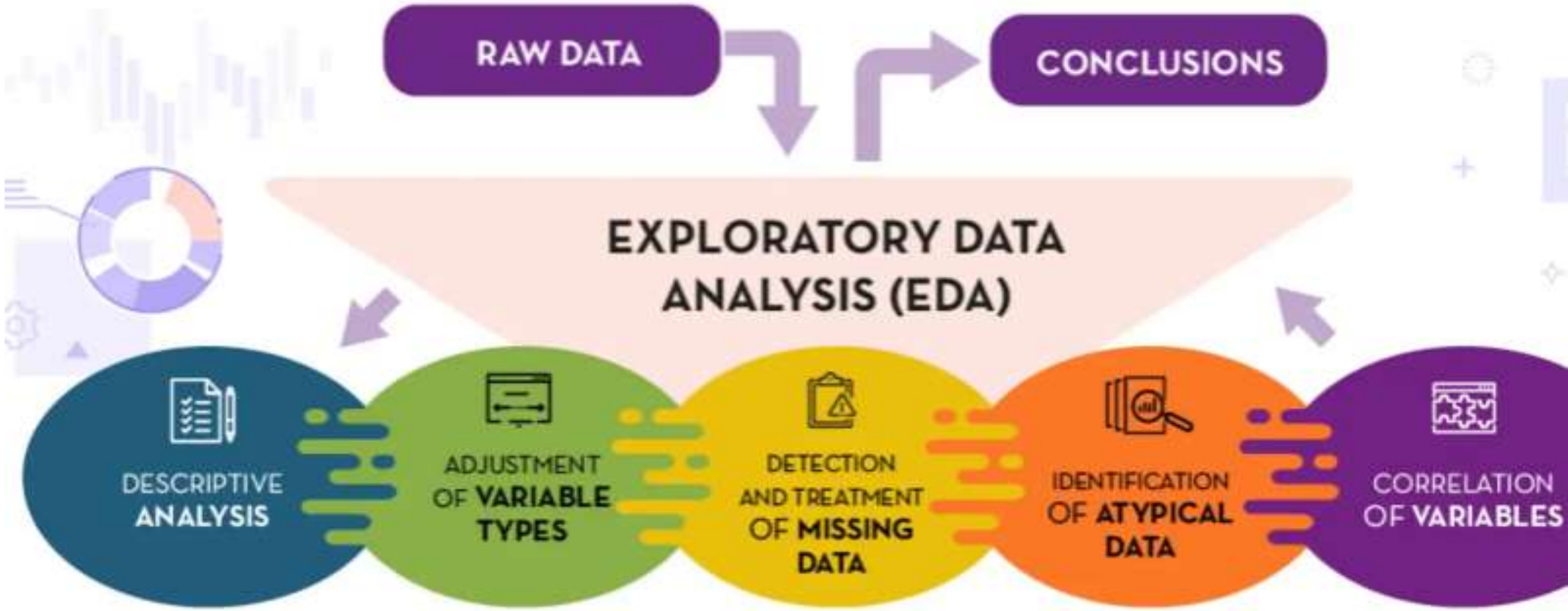
Visualisasi data merupakan alat yang sangat berguna untuk mengkomunikasikan hasil penelitian pasar. Beberapa contoh visualisasi data yang dapat digunakan adalah grafik batang, grafik lingkaran, atau peta panas untuk memvisualisasikan preferensi merek.

Langkah 7: Kesimpulan dan Rekomendasi

Berdasarkan analisis data, kita dapat membuat kesimpulan tentang preferensi konsumen terhadap merek smartphone dan memberikan rekomendasi kepada perusahaan berdasarkan temuan kita. Misalnya, merek A mungkin memiliki kualitas kamera yang lebih baik, sementara merek B mungkin lebih terjangkau dari segi harga.

Manfaat EDA

Analisis Data Eksploratif (ADE) berperan penting dalam proses analisis data, yang bermanfaat membantu analyst dan peneliti dalam mengungkap wawasan dari data.



Manfaat EDA

1. **Pemahaman Data yang Mendalam**

Analisis Data Eksploratif memungkinkan analis untuk mendapatkan pemahaman yang lebih baik tentang karakteristik data, termasuk distribusi, variabilitas, dan struktur dasar dataset. Dengan eksplorasi awal ini, analis dapat mengidentifikasi pola, anomali, atau inkonsistensi dalam data, yang penting untuk analisis lebih lanjut. Pemahaman yang mendalam ini membantu dalam merumuskan hipotesis yang tepat dan memilih teknik analisis data yang paling sesuai.

Manfaat EDA

2. Mengidentifikasi Hubungan Antar Variabel

Melalui EDA, analis dapat mengeksplorasi dan mengidentifikasi hubungan atau korelasi potensial antar variabel. Hal ini termasuk hubungan linier atau non-linier yang mungkin tidak terlihat tanpa pemeriksaan mendalam. Mengungkap hubungan ini sangat penting dalam menentukan variabel-variabel yang paling mempengaruhi variabel lain, yang berguna dalam membangun model prediktif atau analisis kausal.

Manfaat EDA

3. Deteksi Outlier dan Nilai Pengecualian

Salah satu aspek kunci dari Analisis Data Eksploratif adalah kemampuannya untuk mengidentifikasi outlier atau nilai anomali dalam dataset. Outlier dapat menunjukkan kesalahan pengumpulan data, kesalahan entri, atau variabilitas alami dalam populasi. Mendeteksi dan menangani outlier secara tepat sangat penting karena keberadaan mereka dapat mempengaruhi hasil analisis dan kesimpulan secara signifikan.

Manfaat EDA

4. Memilih Model dan Teknik Analisis yang Tepat

Dengan memahami struktur dan dinamika dataset melalui ADE, analis dapat membuat keputusan yang lebih tepat tentang model statistik atau teknik machine learning yang harus digunakan untuk analisis lebih lanjut. Analisis Data Eksploratif mengungkapkan apakah data memenuhi asumsi model tertentu dan membantu dalam memodifikasi data atau memilih transformasi yang diperlukan untuk memenuhi kriteria analisis.

Manfaat EDA

Namun dalam praktiknya tidak ada metode atau urutan baku yang digunakan dalam melakukan EDA, yang perlu diperhatikan adalah semakin banyak anda dapat mengumpulkan informasi dari data yang akan diolah, maka semakin baik pula pemahaman anda terhadapnya. Proses memahami data dapat dipermudah dengan menggunakan ilmu statistika.

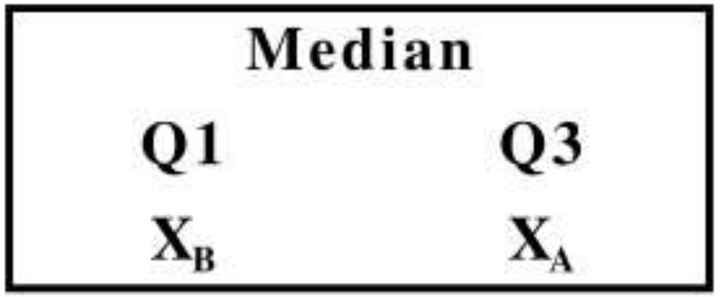
Statistika dalam EDA

Statistika adalah sebuah ilmu yang mempelajari bagaimana cara merencanakan, mengumpulkan, menganalisis, lalu menginterpretasikan, dan akhirnya mempresentasikan data. Singkatnya, statistika adalah ilmu yang bersangkutan dengan suatu data. Dalam pembahasan selanjutnya saya akan mengimplemantasikan statistika deskriptif pada data untuk melakukan proses EDA.

Ringkasan 5 Angka

Dalam analisis data eksploratif terdapat satu hal penting yang harus diketahui, yaitu ringkasan numerik. Ringkasan numerik merupakan ringkasan data yang terdiri dari ukuran-ukuran penting dari data yang menggambarkan gambaran karakteristik umum dari data. Karakteristik tersebut meliputi ukuran pusat data dan ukuran sebaran data.

Ringkasan numerik biasa disajikan dalam bentuk ringkasan lima angka. Disebut ringkasan lima angka karena memuat lima angka atau ukuran penting dari data. Ringkasan lima angka yang sering digunakan adalah Median, Kuartil 1, Kuartil 2, data terendah, dan data tertinggi. Kelima angka tersebut dapat disajikan dalam bentuk



Median merupakan nilai tengah data setelah data diurutkan. Karena median bersifat robust atau tidak terpengaruh oleh adanya data ekstrim, maka median ini digunakan untuk menunjukkan pusat data dalam ringkasan lima angka. Sedangkan empat angka yang lain, yaitu Q1 (Kuartil 1), Q3 (Kuartil 3), Xb (nilai data terendah), dan Xa (nilai data tertinggi) digunakan untuk menunjukkan sebaran data.


THANK
YOU