

Eksplorasi Data

PERTEMUAN 2



Apa itu Eksplorasi Data?

Eksplorasi data merupakan langkah untuk memahami data sebelum dilakukan praproses. Tujuan dari ekplorasi data adalah menyeleksi teknik pemrosesan dan analisis data yang sesuai dengan dataset yang dimiliki.



Sumber: <https://lp2m.uma.ac.id/wp-content/uploads/2022/06/data.png>

Apa itu Eksplorasi Data?

Exploratory Data Analysis (EDA) adalah suatu pendekatan analisis data dengan menggunakan berbagai teknik untuk mendapatkan pengetahuan atau wawasan tentang data.

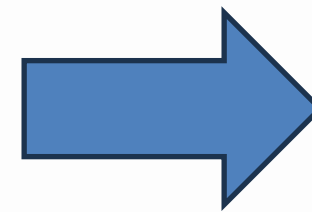
Langkah-langkah
dasar dalam
analisis data
eksplorasi:

- Pembersihan dan pra-pemrosesan
- Analisis statistik
- Visualisasi untuk analisis tren, deteksi anomali, deteksi outlier (dan penghapusan).

Ilustrasi



<https://www.qoala.app/id/blog/wp-content/uploads/2020/12/16-Tips-Beli-HP-Bekas-Tetap-Berkualitas-Walau-Ponsel-Second.jpg>



EDA dapat digambarkan ketika seseorang hendak membeli HP. Tentunya, hal pertama yang akan dilakukan adalah mencari tahu terkait spesifikasi HP yang akan dibeli baik dengan membaca artikel maupun dengan menonton video unboxing.

Tujuan: agar pembeli mengetahui dan memahami apa saja spesifikasi dari HP tersebut. Misalnya RAM 4GB, Kamera 50MP, battery 5000 mAh, dan seterusnya. Proses untuk mengetahui dan memahami spesifikasi dari HP tersebutlah yang dikenal dengan EDA.

Mengapa perlu dilakukan EDA?



Meningkatkan pemahaman tentang variabel dengan mengekstraksi nilai rata-rata, mean, minimum, dan maksimum, dll.

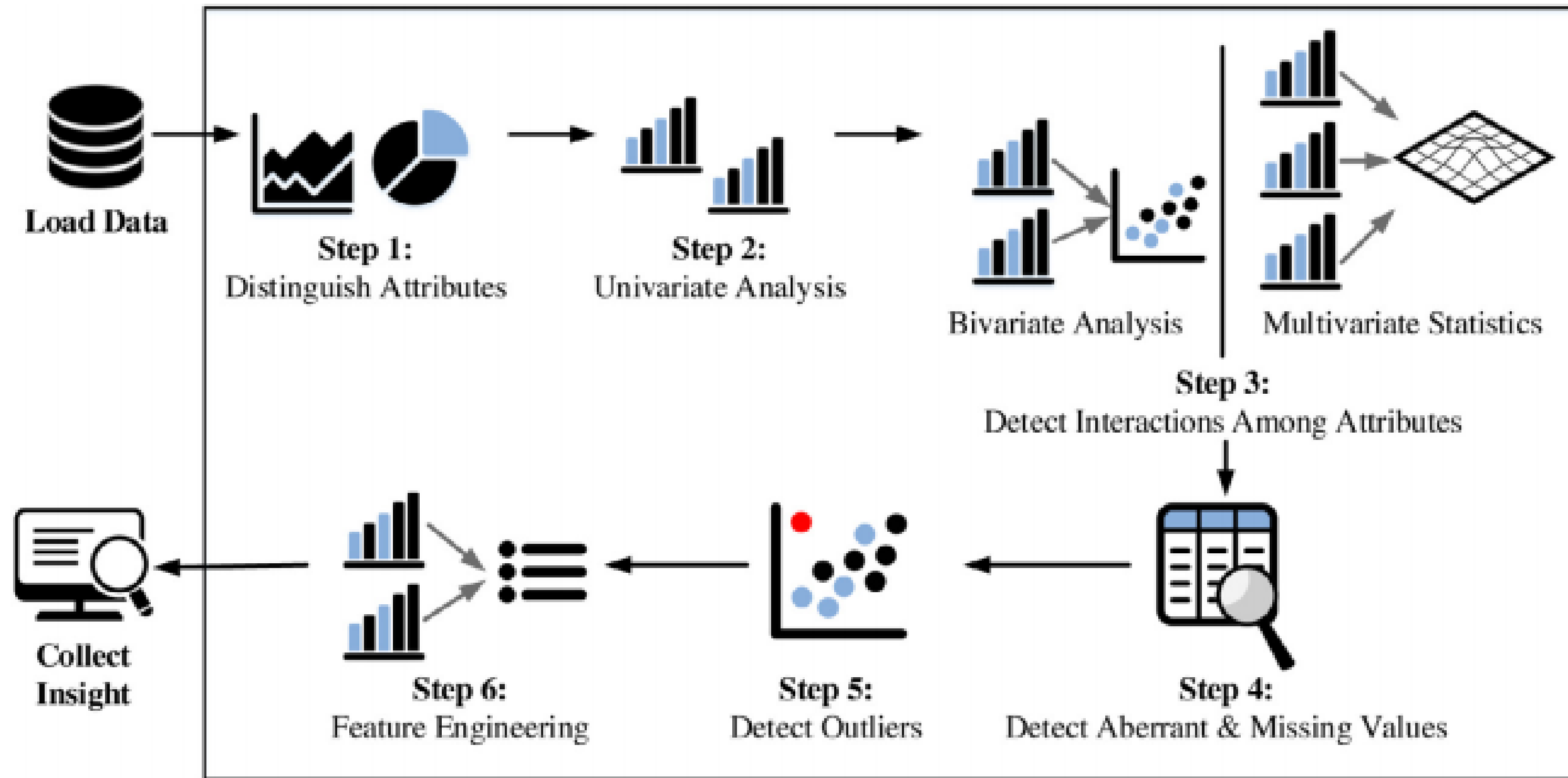


Temukan kesalahan (errors), outlier, dan missing values pada data.



Mengidentifikasi pola dengan memvisualisasikan data dalam grafik seperti grafik batang, plot sebar, peta panas, dan histogram.

Proses EDA?



Apa itu Data

- ❑ Data: Koleksi fakta yang terdiri dari objek dan atribut
- ❑ Atribut (*attributes*) merupakan karakteristik dari sebuah objek
 - ❑ Contoh: warna mata, berat badan, temperature, dan lainnya.
 - ❑ *Attributes* juga dikenal sebagai *variable*, *field*, *characteristic*, atau *feature*.
 - ❑ *Attributes* biasanya direpresentasikan sebagai kolom dalam format tabel.
- ❑ Kumpulan beberapa *attributes* menjelaskan sebuah objek.
 - ❑ Objects juga sering disebut dengan istilah *record*, *point*, *case*, *sample*, *entity*, or *instance*.
 - ❑ Objek biasanya direpresentasikan sebagai baris dalam format tabel.

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

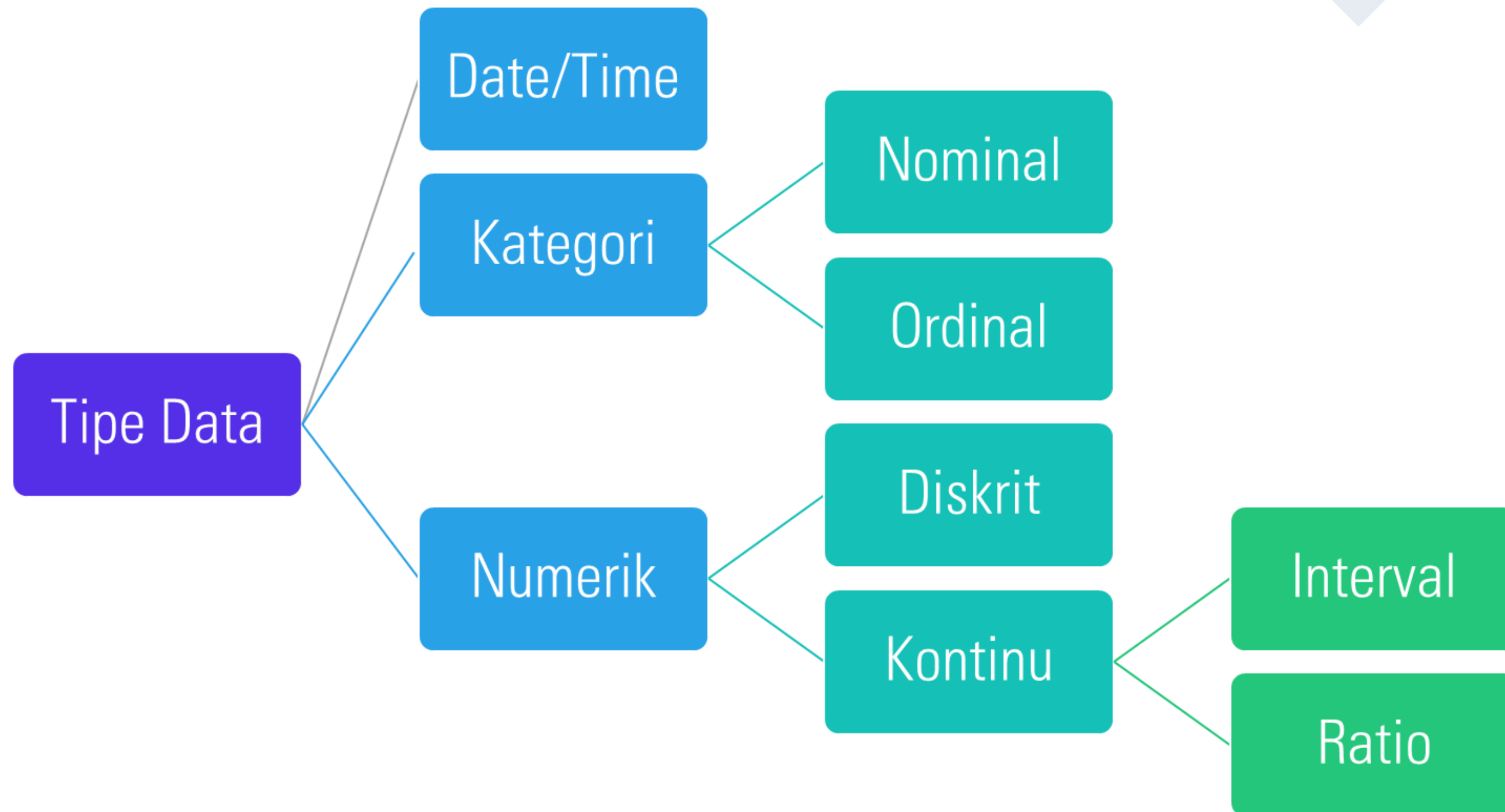
Himpunan Data (Dataset)

- Jenis dataset ada dua: **Private** dan **Public**
- **Private Dataset**: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
 - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- **Public Dataset**: data set dapat diambil dari repositori publik yang disepakati oleh para peneliti data mining
 - **UCI Repository** (<http://www.ics.uci.edu/~mlearn/MLRepository.html>)
 - **ACM KDD Cup** (<http://www.sigkdd.org/kddcup/>)
 - **PredictionIO** (<http://docs.prediction.io/datacollection/sample/>)
 - **Kaggle** (<https://www.kaggle.com/>)

7 Cara Mendapatkan Data

1. **Created data** : Survei pasar, focus groups biasanya berbentuk data terstruktur atau semi-structured, dan bisa berasal dari internal atau external.
2. **Provoked data** : tidak akan ada data, jika tidak mengundang orang untuk mengekspresikan pandangan mereka. Misalnya like bintang 5 pada system rating, biasanya terstruktur atau semi terstruktur, bisa internal atau external.
3. **Transaction data** : dihasilkan setiap waktu ketika seseorang melakukan pembelian / transaksi, biasanya data internal terstruktur.
4. **Compiled data** : dari database yang besar, mengkompilasi sejumlah data dari berbagai sumber, biasanya data external terstruktur
5. **Experimental data** : hybrid data yang dibuat dan data transaksi, kumpulan pelanggan yang berbeda menerima perlakuan pemasaran yang berbeda (dibuat) dan mengamati hasilnya di dunia nyata (transaksi), biasanya terstruktur atau semi-terstruktur, bisa eksternal / internal.
6. **Captured data** : GPS data, sensors, IoT, biasanya tidak terstruktur dan bisa internal/eksternal.
7. **User-Generated data** : individu dan perusahaan menghasilkan secara sadar/tdk, biasanya tidak terstruktur dan dapat bersifat internal atau eksternal.

Tipe Data



* Jumlah data diskrit < 10, bisa masuk ke dalam data kategori

Data Numerik

Pembagian pertama:

- **Diskret (*discrete*)**
 - **Ada karena dihitung (*counted*) → bisa dihitung menggunakan jari**
 - Tidak memiliki koma (*integer*)
 - Contoh:
 - Lemparan 2 buah dadu (2,3,4,5,6,7,8,9,10,11,12)
 - Jumlah orang (2 orang, 5 orang, 10 orang, dan seterusnya)
- **Kontinu (*continuous*)**
 - **Ada karena diukur (*measured*)**
 - Memiliki koma, dan nilainya tak terbatas (*infinite*)
 - Biasanya memiliki satuan ukur (misal: meter, liter, kilogram)
 - Contoh:
 - Jarak (12,5 km, 100,0 km, 197,65 km)
 - Gaya (75,4 N, 120,09 N, 0,87 N)
 - Tinggi badan, berat badan, volume air, dan lain-lain.

Data Numerik (2)

Pembagian kedua:

- **Interval**

- **Tidak memiliki arti angka nol (*no true zero*). Nilai nol bukan berarti tidak ada data (tidak ada derajat ukurnya).**
- Bisa bernilai *negative* dan bisa dijumlahkan/dikurangi
- Contoh: temperature (*celcius*), waktu (pukul 22.00, pukul 00.00, pukul 02.00)
- Kita tidak bisa mengatakan suhu 40°C adalah 2x dari 20°C
 - Pukul 00.00 artinya tengah malam, dan bukan berarti waktunya kosong (nol).

- **Ratio**

- **Memiliki arti angka nol (*true zero*) → 0 = *an absence of the thing being measured***
- Bisa dijumlah, tambah, bagi, dan kali
- Contoh: jarak, kadar obat, usia, *response time*, durasi, dll.
- Kita bisa mengatakan 200 km adalah setengah dari 100 km
 - Jarak 0 km, memang ada maknanya, kita tidak bergerak kemana-mana.

Data Nominal

- Lebih bersifat kualitatif
- Bagian dari beberapa kelompok
- Tidak ada perbedaan 'peringkat' antara kategori satu dengan lainnya.
- Contoh:
 - Data biner (*binary*) yaitu data dengan 2 kategori. Misal cacat/tidak cacat, yes/no, baik/buruk, dan seterusnya.
 - Data ras , misalnya jawa, sunda, bugis, minang, dan seterusnya.
 - Data warna rambut, misalnya hitam, coklat, pirang, merah, dan seterusnya
- Angka numerik dari data kategori ini tidak memiliki nilai matematis

Data Ordinal

- Data kategori yang memiliki peringkat (adanya urutan di mana nilai lebih besar memiliki arti lebih penting/lebih berbobot).
- Merupakan campuran antara data numerik dan data kategori
- Angka yang dimiliki oleh data *ordinal* memiliki arti matematis
- Nilai yang dimungkinkan dalam data *ordinal* ini biasanya memiliki rentang tertentu.
- Contoh:
 - Rating yang kita berikan saat mengisi data kuesioner. Biasanya rentang antara 1 sampai 5 atau bahkan 1 sampai 10.
 - Rating yang kita berikan di aplikasi toko online apakah penjualnya memiliki rating buruk (1) atau sangat baik (5). Biasanya dalam bentuk bintang.
 - Pertanyaan kuesioner, "Seberapa Lelah Anda?" di mana jawabannya adalah "sangat lelah", "cukup lelah", "tidak lelah".
- Nilai ordinal berupa teks bisa dirubah menjadi angka untuk menunjukkan peringkatnya (level)

Data Waktu

- Tipe data ini bisa terdiri dari: *date* (tanggal), *time* (waktu), gabungan *date & time*
- Contoh:
 - Tanggal lahir ('29-10-1987', '17-01-2021')
 - Tanggal kejadian ('2017-Mar', '2014-Jun')
 - Waktu kejadian ('13:24:50')
 - Waktu pembayaran ('11-01-2021 15:20:32')

Data Campuran

- Terkadang sebuah variabel adalah kombinasi dari beberapa jenis data (angka dan kategori/huruf).
- Contoh:
 - Cabin (Titanic): A15, B18, ...
 - Tiket (Titanic): A103349, ...
 - Plat nomor: B 1203 ZZ, ...
 - Kode pos (luar negeri): SE102, ...
- Data campuran ini dapat diekstrak (dipisahkan antara angka dan huruf) untuk membentuk variabel baru yang nantinya bisa dianalisis lebih lanjut.

Tiga Jenis Data Set Berdasarkan Golongannya

1. Record

- ✓ Matrik data
- ✓ Data transaksi
- ✓ Data dokumen

2. Graph

- ✓ Word Wide Web
- ✓ Struktur molekul

3. Ordered Data

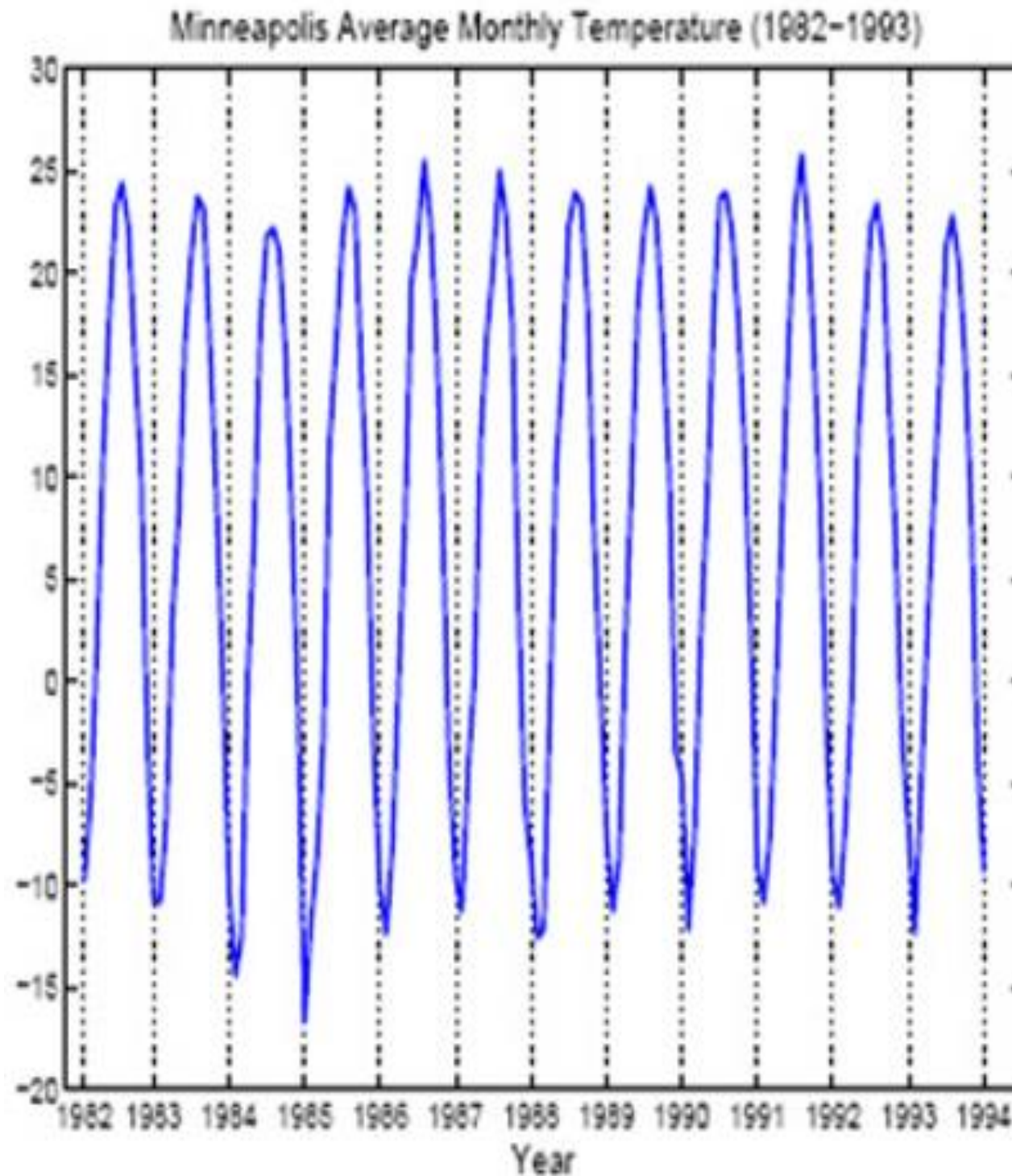
- ✓ Data spasial
- ✓ Data temporal
- ✓ Data sekuensial
- ✓ Data urutan genetic (*genetic sequence*)

Data Transaksi

- Setiap *record* (transaksi) melibatkan satu set item yang biasanya menyertakan nomor identitas transaksi.
 - Misalnya;
kumpulan produk yang dibeli oleh seorang *customer* selama satu kali belanja dan dikategorikan satu kali transaksi atau disebut juga dengan *market basket*.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Data Temporal



Data temporal adalah data yang objeknya memiliki atribut yang mewakili pengukuran yang diambil dari waktu ke waktu.

Misalnya, kumpulan data keuangan adalah deret waktu yang memberi harga harian berbagai saham.

Seri waktu adalah urutan pengukuran beberapa atribut

Misalnya, harga saham atau curah hujan, diambil pada (biasanya reguler) pada waktunya.)

Kualitas Data

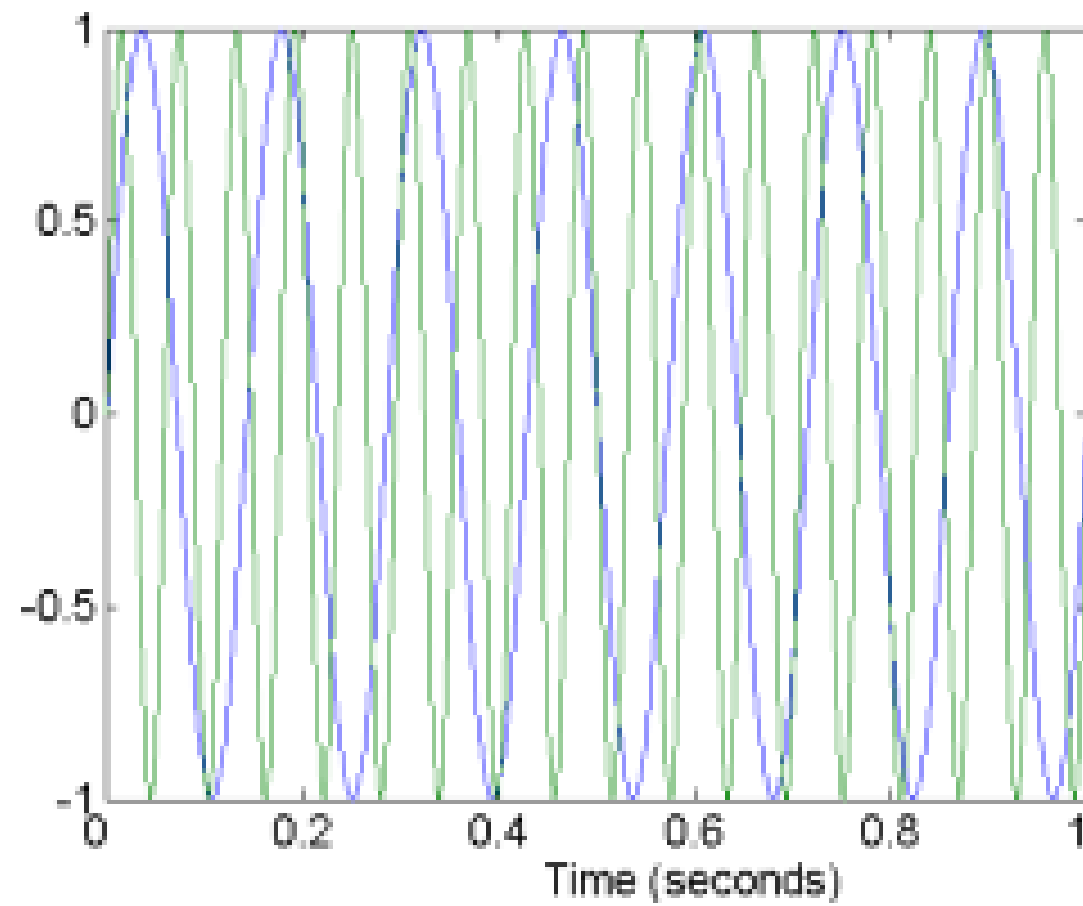
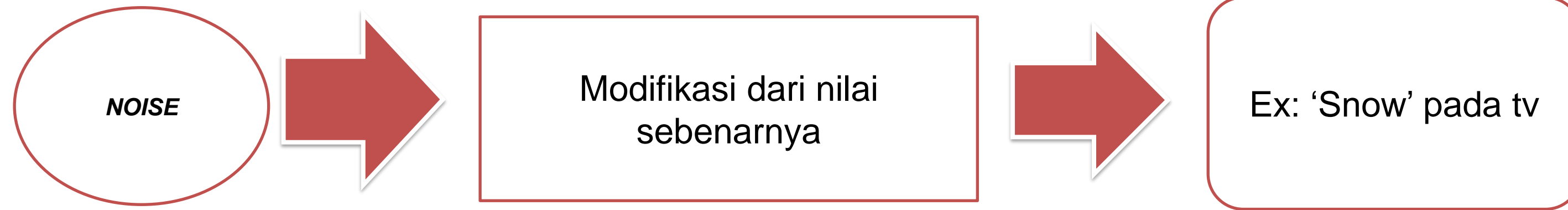
Noise

Outliers

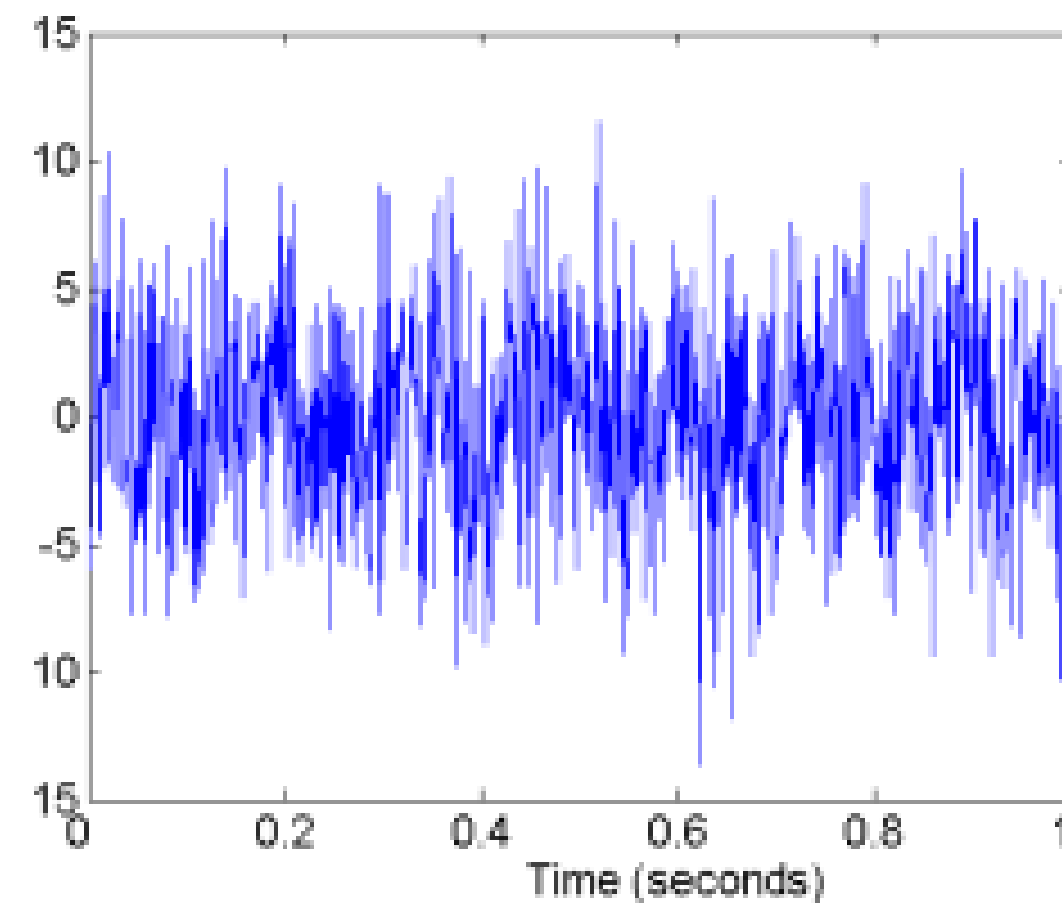
Missing Value

Duplicate

Kualitas Data



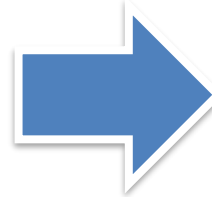
Two Sine Waves



Two Sine Waves + Noise

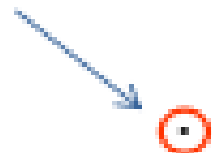
Kualitas Data

OUTLIER



Objek data yang memiliki karakteristik berbeda dengan data lainnya

Outlier



Outlier



Outlier



Outlier dapat dipandang sebagai *noise* tetapi berguna dalam *fraud detection*, *rare event analysis*

Kualitas Data



Kualitas Data

Duplicate data

Masalah utama ketika
menggabungkan data dari
berbagai sumber



Data Cleaning
Menghilangkan noise dan data
yang tidak konsisten



VISUALISASI DATA

Visualisasi data adalah salah satu teknik dalam eksplorasi data.

Manfaat visualisasi data:

- Dapat mendeteksi general pola dan trends
- Dapat mendeteksi *outlier* atau *unusual* trends



Boxplots



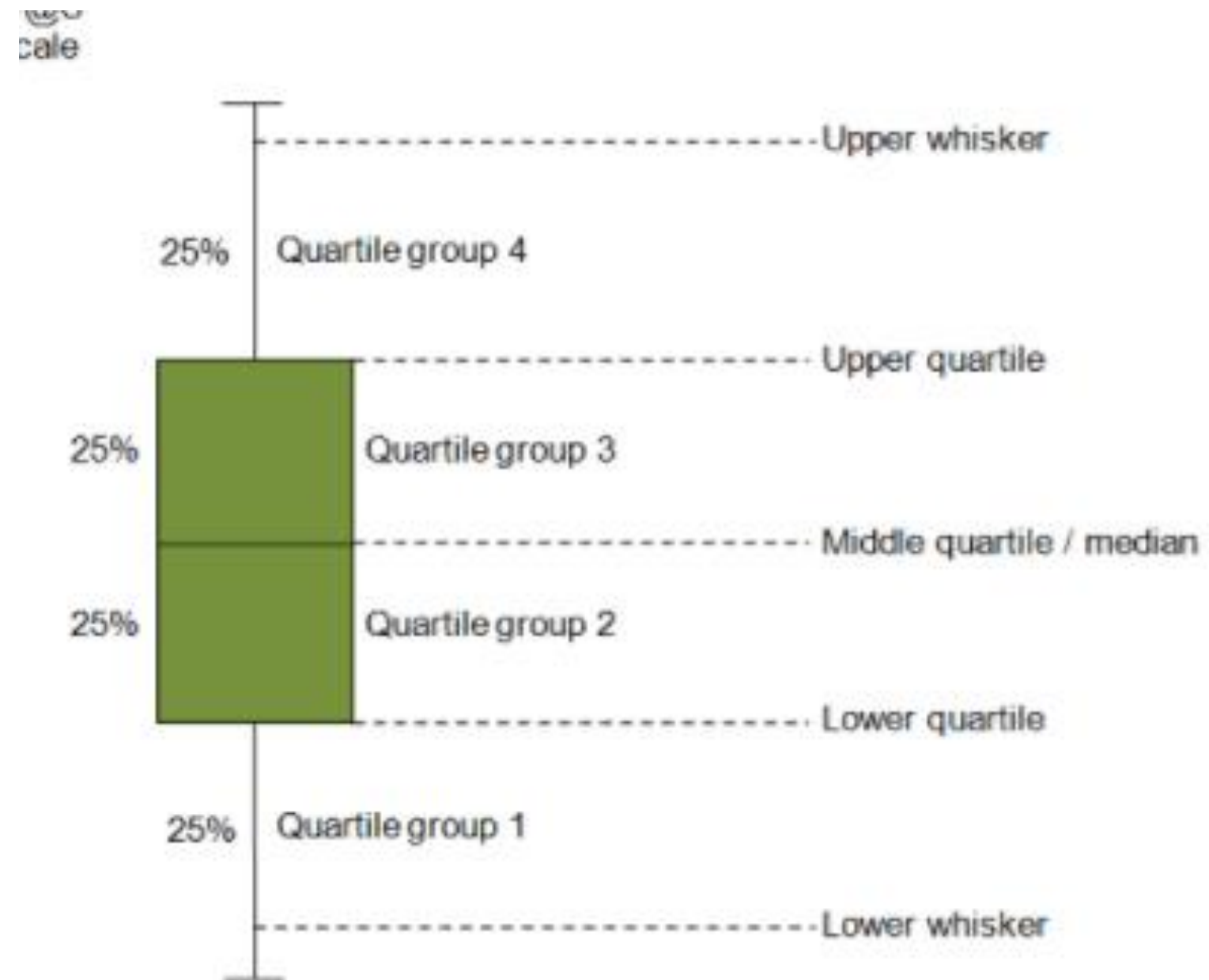
Histogram



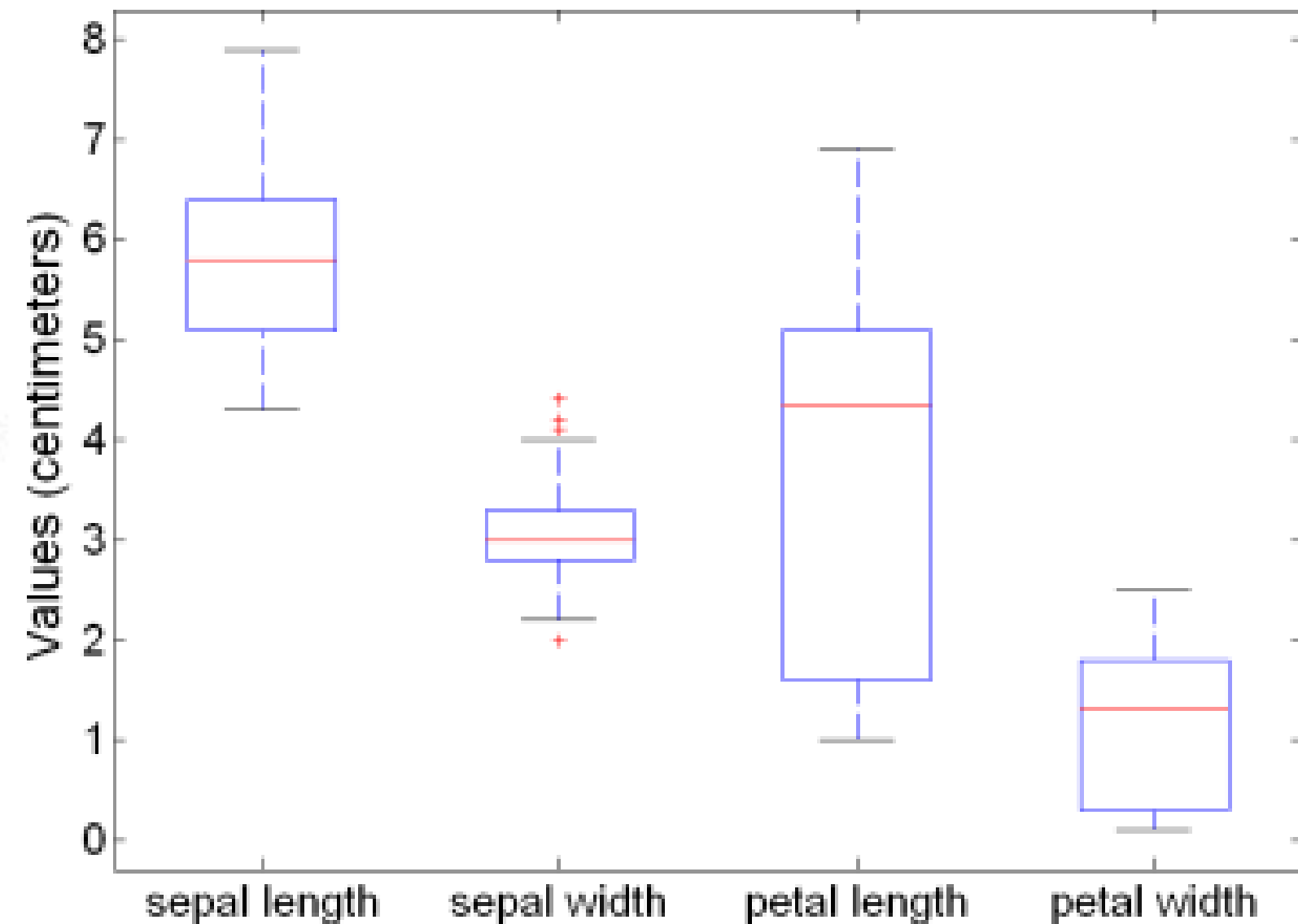
Scatter
plot

VISUALISASI DATA

- BOXPLOTS



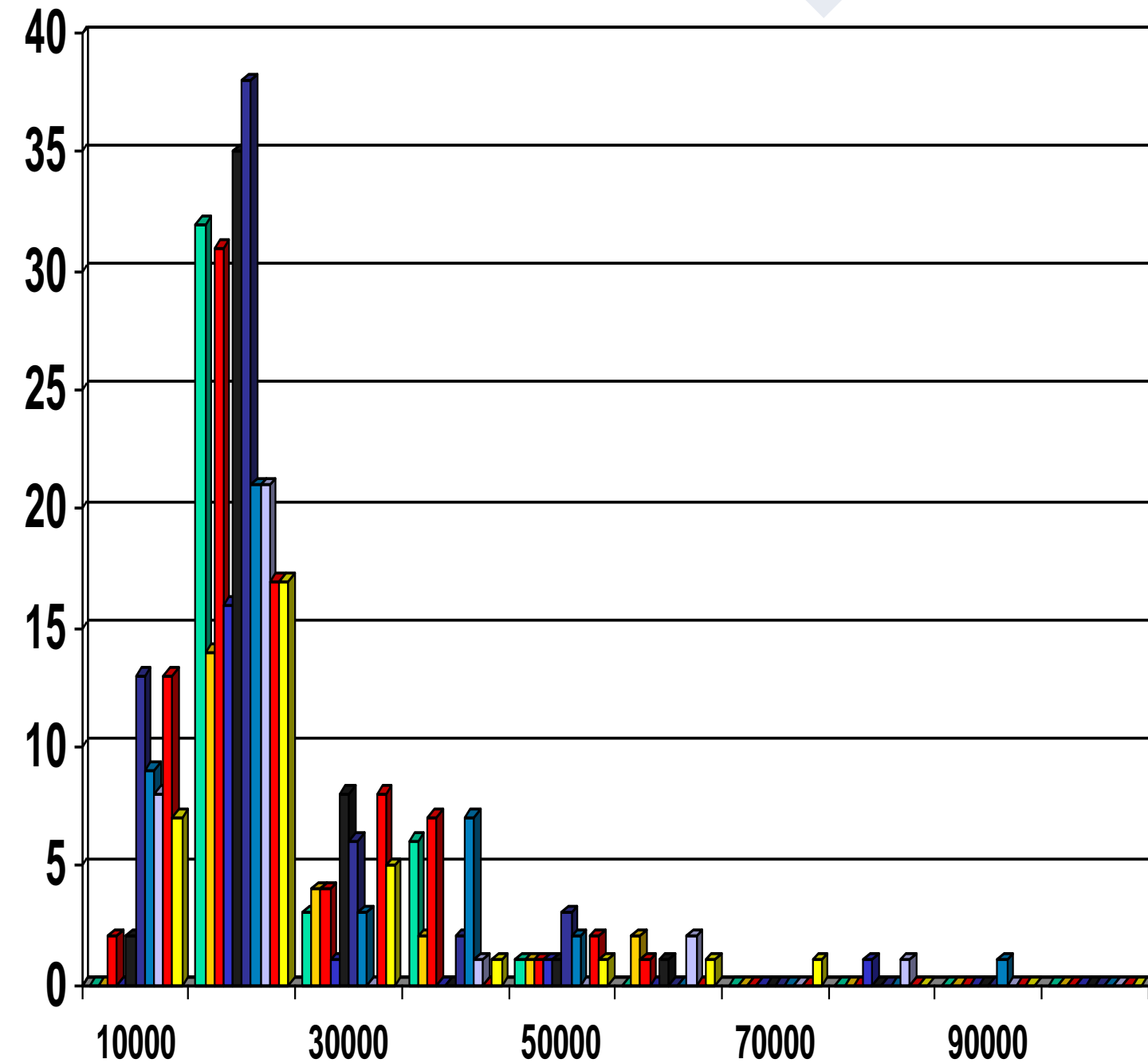
Bagian-bagian dari boxplots



Boxplots dapat digunakan untuk membandingkan atribut

- **HISTOGRAM**

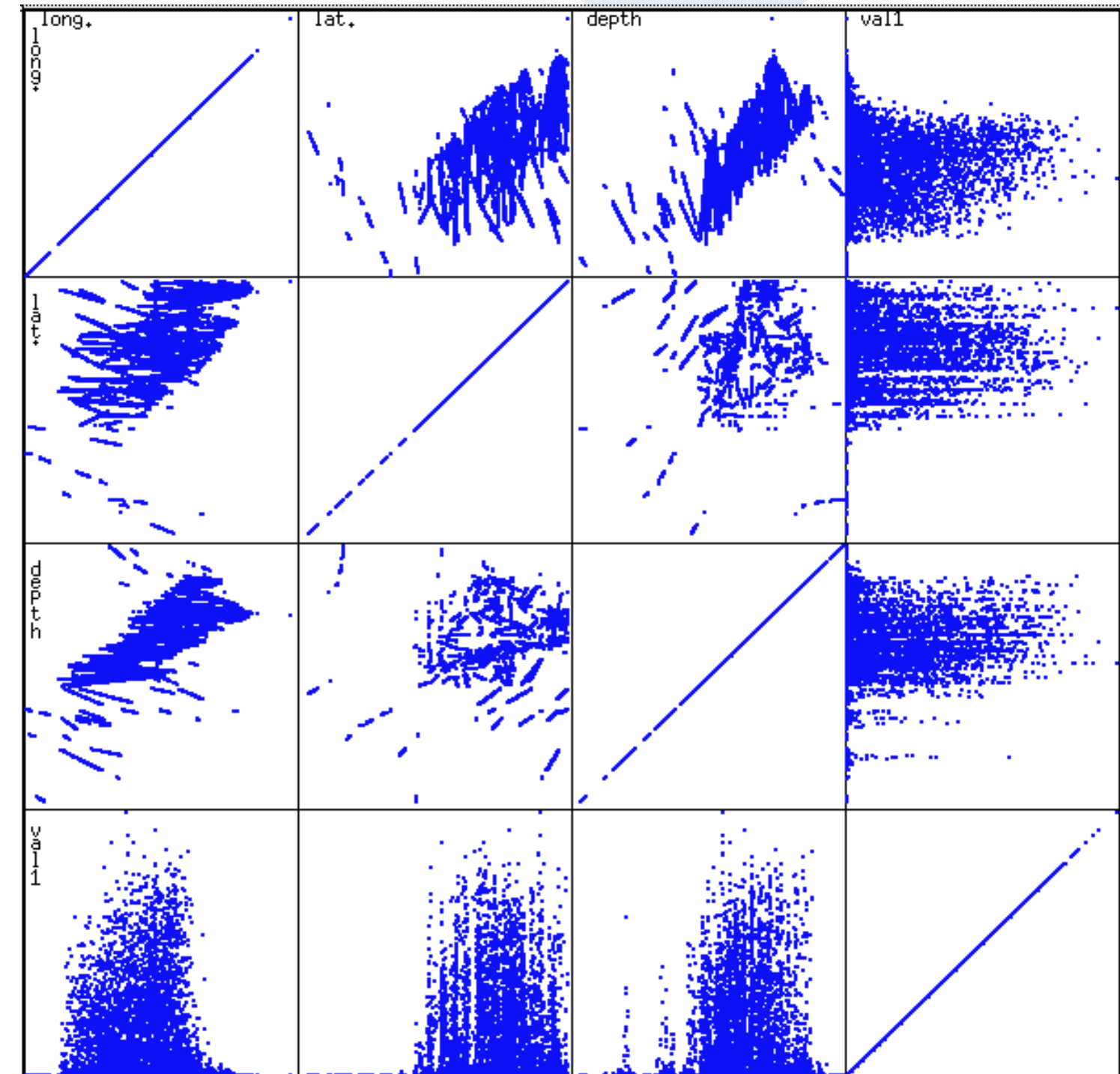
- ✓ Histogram mendistribusikan nilai-nilai suatu atribut
- ✓ Membagi nilai menjadi bin-bin dan barplot menunjukkan jumlah objek pada setiap bin
- ✓ Tinggi dari setiap bar mengidentifikasikan jumlah objek



VISUALISASI DATA

- **SCATTER PLOT**

- Nilai atribut menentukan posisi
- Atribut ditunjukkan dengan warna maupun bentuk yang berbeda dengan atribut lainnya
- Dapat melihat hubungan beberapa pasangan atribut



EDA dengan Python

Import data into workplace(Jupyter notebook, Google colab, Python IDE)

Descriptive statistics

Removal of nulls

Visualization

1. Packages and data import

- Step 1 : Import packages to the workplace.
 - Import pandas as pd *#untuk mengubah dimensi data, membuat tabel, memeriksa data, membaca data dan lain sebagainya.*
 - Import numpy as np *#untuk memudahkan operasi perhitungan tipe data numerik seperti penjumlahan, perkalian, pengurangan, dan operasi aritmatika lainnya.*
 - Import seaborn as sns *#untuk menampilkan visualisasi data*
 - Import matplotlib.pyplot as plt *#untuk membuat chart atau grafik*
 - %matplotlib inline
- Step 2 : Read data/dataset into Pandas dataframe. Different input formats include:
 - Excel : pd.read_excel
 - CSV: pd.read_csv
 - JSON: pd.read_json
 - HTML and many more
 - df = pd.read_csv *#untuk membaca dataset yang sudah disediakan dalam format CSV*
 - df.head() *#untuk menampilkan isi dataset 5 teratas. Jika ingin menampilkan 5 data terbawah ganti head dengan tail. Jika ingin menampilkan sebanyak 10 data teratas , isi nilai di dalam kurung menjadi (10).*
 - df.dtypes *#untuk memeriksa tipe data yang digunakan setiap fitur*

1. Packages and data import

- Step 3 : Renaming the columns.
 - `df = df.rename(columns={"Preferred_Browser": "Browser", "Preferred_Search_Engine": "Search_Engine"})` *# untuk mengganti nama fitur atau kolom sesuai dengan keinginan kita. Dalam kasus ini, mengubah Preferred_Browser menjadi Browser dan Preferred_Search_Engine menjadi Search Engine.*
 - `df.head()`
- Step 4 : Dropping the duplicate rows
 - `df.shape` *#untuk melihat jumlah baris dan kolom yang dimiliki dataset.*
 - `duplicate_rows_df = df[df.duplicated()]`, *# untuk melihat jumlah baris yang memiliki duplikat dengan diwakili oleh variabel duplicate_rows_df. Pada saat data duplikat ini akan dihapus, maka variabel tersebutlah yang akan dihapus.*
 - `Print("baris yang memiliki duplikat: ", duplicate_rows_df.shape)`, *# untuk menampilkan data yang memiliki duplikat di dalam dataset.*
 - `df.count()` *# untuk menghitung jumlah seluruh baris dari tiap kolom yang ada sebelum data duplikatnya dihapus.*
 - `df = df.drop_duplicates()` *# untuk menghilangkan data duplikat yang terdapat di dalam dataset.*
 - `df.head()`
 - `df.count()`

2. Descriptive Stats (Pandas)

- Digunakan untuk membuat penilaian awal tentang distribusi populasi variabel.
- Statistik yang umum digunakan:

1. Central tendency :

- Mean – Nilai rata-rata semua titik data. : `df.mean()`
- Median – Nilai tengah ketika semua titik data dimasukkan ke dalam daftar terurut: `df.median()`
- Mode – Titik data yang paling banyak muncul dalam kumpulan data :`df.mode()`

2. Spread : Merupakan ukuran seberapa jauh jarak titik data dari mean atau median

- Varians – Varians adalah rata-rata kuadrat dari masing-masing deviasi: `df.var()`
- Standard deviation – Standard deviation adalah akar kuadrat dari varians:`df.std()`

3. Skewness: Ini adalah ukuran asimetri: `df.skew()`

2. Descriptive Stats (Lanjutan..)

- Metode lain untuk melihat data dengan cepat:
- `Describe()` : Meringkas kecenderungan sentral, penyebaran, dan bentuk distribusi kumpulan data, tidak termasuk nilai NaN.

Syntax: `pandas.dataframe.describe()`

- `Info()` : Mencetak ringkasan singkat dari kerangka data. Metode ini mencetak informasi tentang kerangka data termasuk indeks dtype dan kolom, nilai bukan nol, dan penggunaan memori.

Syntax: `pandas.dataframe.info()`

3. Nilai Kosong (*Null values*)

Detecting

Detecting Null-values:

- `IsNull()`: It is used as an alias for `dataframe.isna()`. This function returns the dataframe with boolean values indicating missing values.
- Syntax : `dataframe.isnull()`

Handling

Handling null values:

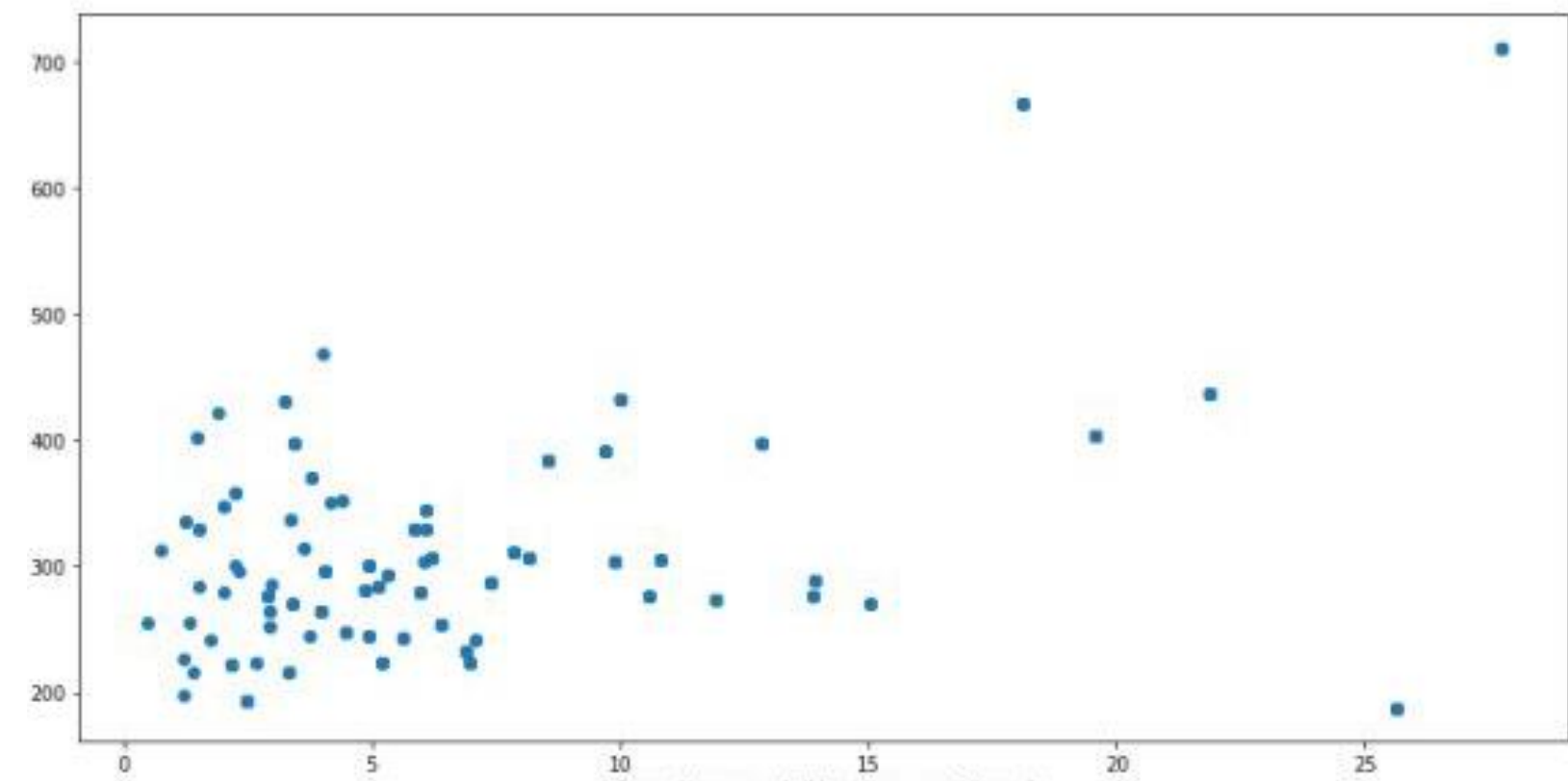
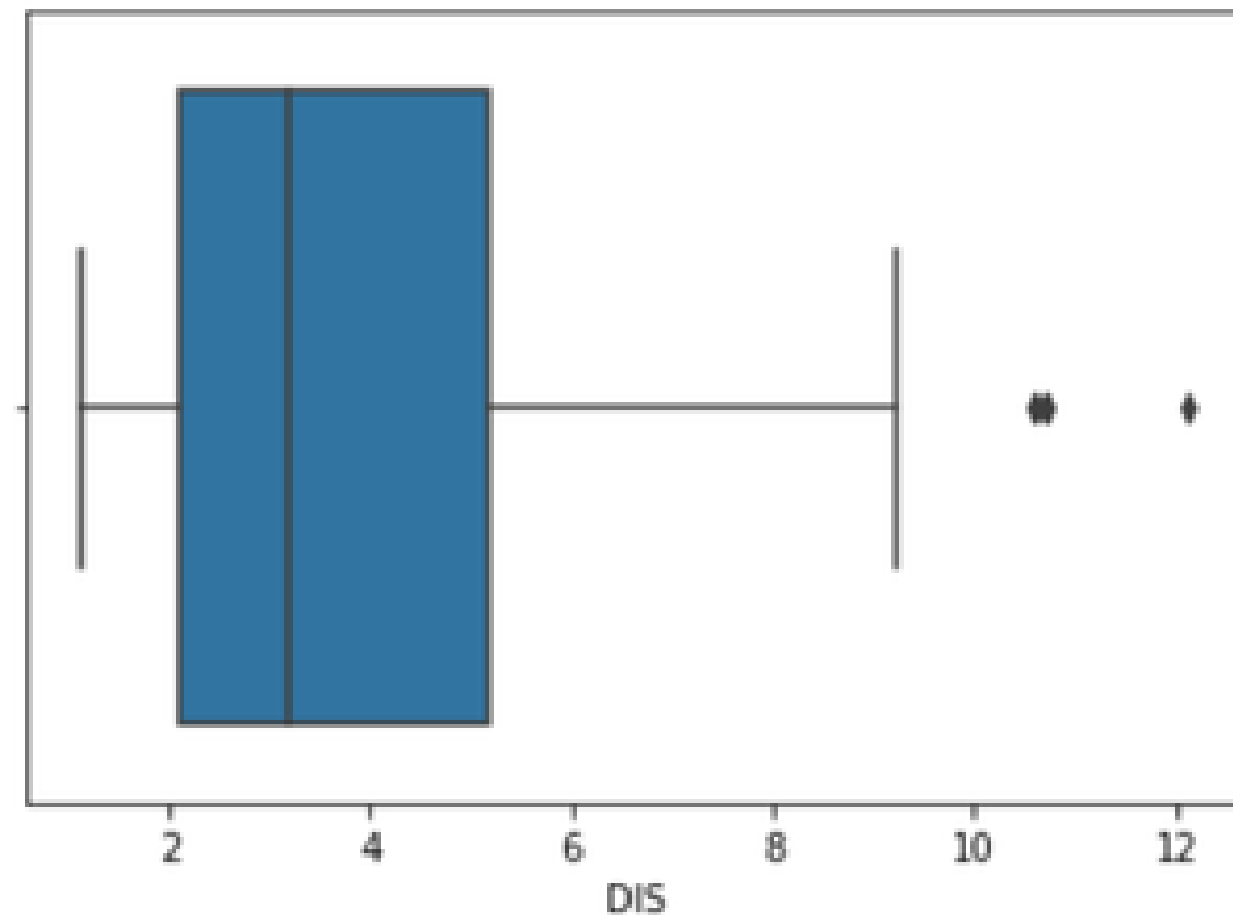
- Dropping the rows with null values: `dropna()` function is used to delete rows or columns with null values.
- Replacing missing values: `fillna()` function can fill the missing values with a special value value like mean or median.

- Dropping The Missing or Null Value

- `print(df.isnull().sum())` *#untuk menampilkan seluruh baris (fitur) yang didalamnya terdapat missing value.*
- `df = df.dropna()` *#untuk menghapus seluruh missing value yang terdapat di dalam dataset.*
- `df.count()` *#untuk menghitung kembali total keseluruhan baris (fitur) yang terdapat di dalam dataset.*
- `print(df.isnull().sum())` *#untuk memastikan kembali missing value yang terdapat pada dataset yang kita gunakan.*

4. Deteksi outlier

- Outlier adalah titik atau kumpulan titik data yang terletak jauh dari nilai data lainnya dalam kumpulan data tersebut.
- Pencilan mudah diidentifikasi dengan memvisualisasikan data.
- Misalnya.
- Dalam plot kotak, titik data yang berada di luar batas atas dan bawah dapat dianggap sebagai outlier
- Dalam plot sebar, titik data yang berada di luar kelompok titik data dapat dianggap sebagai outlier



Outlier removal

Hitung IQRnya sebagai berikut:

- Hitung kuartil pertama dan ketiga ($Q1$ dan $Q3$)
- Hitung rentang interkuartil, $IQR = Q3 - Q1$
- Temukan batas bawahnya yaitu $Q1 * 1.5$
- Temukan batas atas yaitu $Q3 * 1.5$
- Ganti titik data yang berada di luar rentang ini.
- Mereka dapat diganti dengan mean atau median.

a. Dengan menggunakan Boxplot

- `f = plt.figure(figsize=(12,4))`
- `f.add_subplot(1,2,1)`
- `Df['Hours_Per_Day'].plot(kind='kde')`
- `f.add_subplot(1,2,2)`
- `Plt.boxplot(df['Hours_Per_Days'])`
- `Plt.show`

Atau

- `sns.boxplot(x=df['Hours_Per_Day'])`

5. Visualization

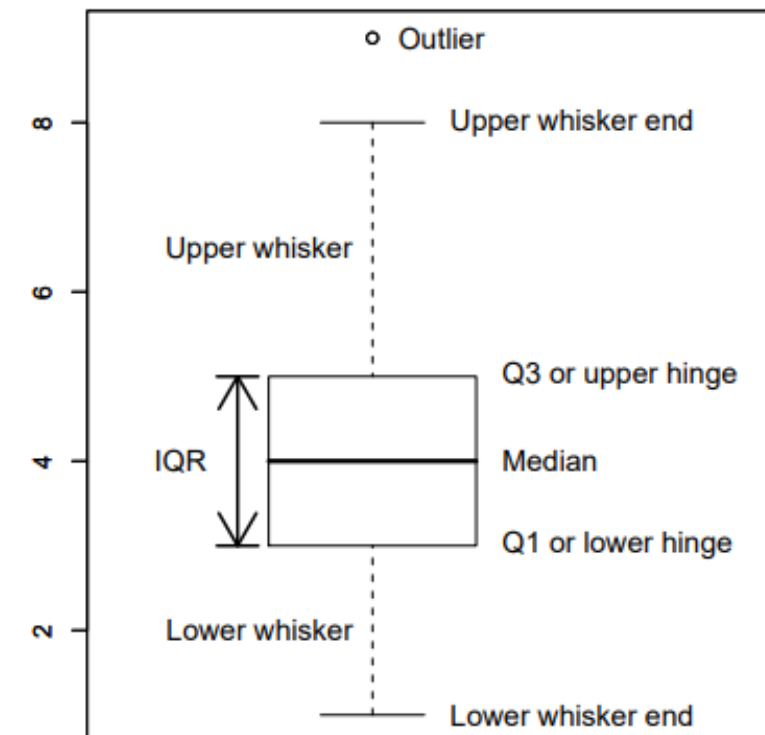
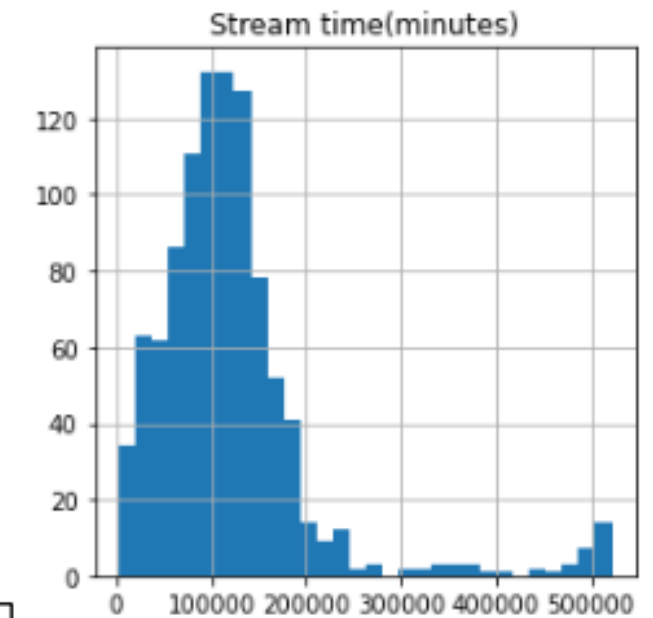
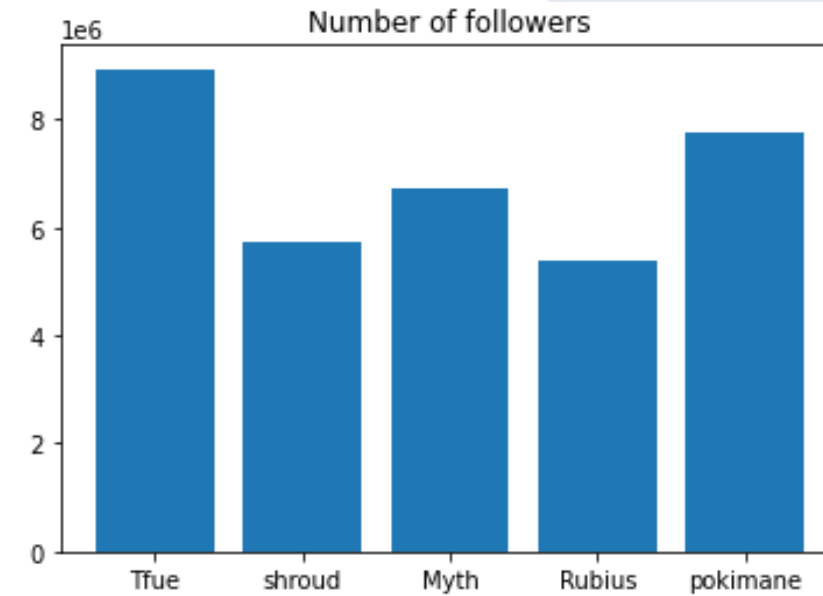
- Univariate: Melihat satu variabel/kolom pada satu waktu
 - Bar-graph
 - Histograms
 - Boxplot
- Multivariate : Melihat hubungan antara dua variabel atau lebih
 - Scatter plots
 - Pie plots
 - Heatmaps(seaborn)

Bar-Graph, Histogram and Boxplot

- Bar graph: plot yang menyajikan data dalam bentuk batang persegi panjang yang panjangnya sebanding dengan nilai yang diwakilinya.
- Boxplot : Menggambarkan data numerik secara grafis melalui kuartilnya. Kotak tersebut terbentang dari nilai kuartil data $Q1$ hingga $Q3$, dengan garis di median ($Q2$).
- Histogram: Representasi sebaran data.

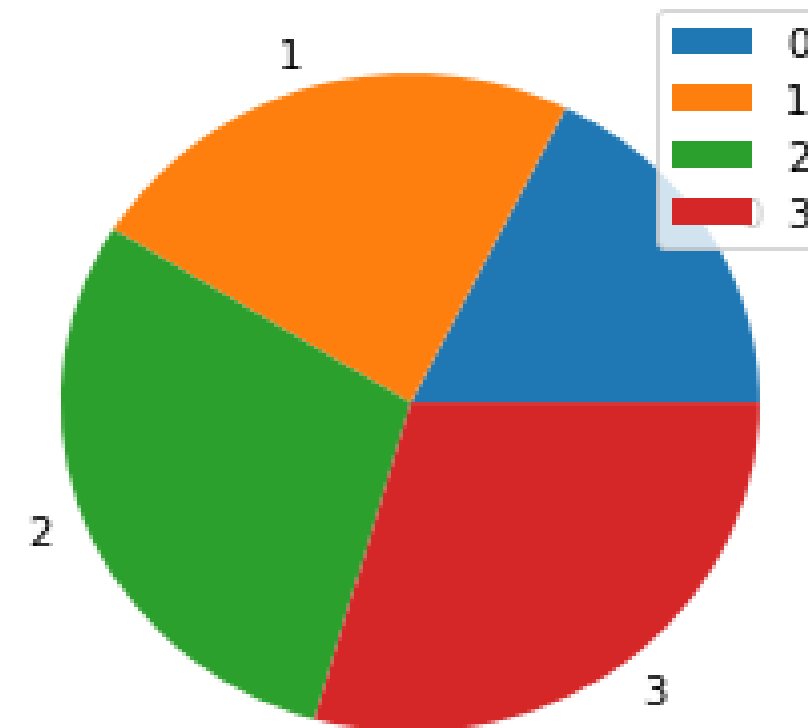
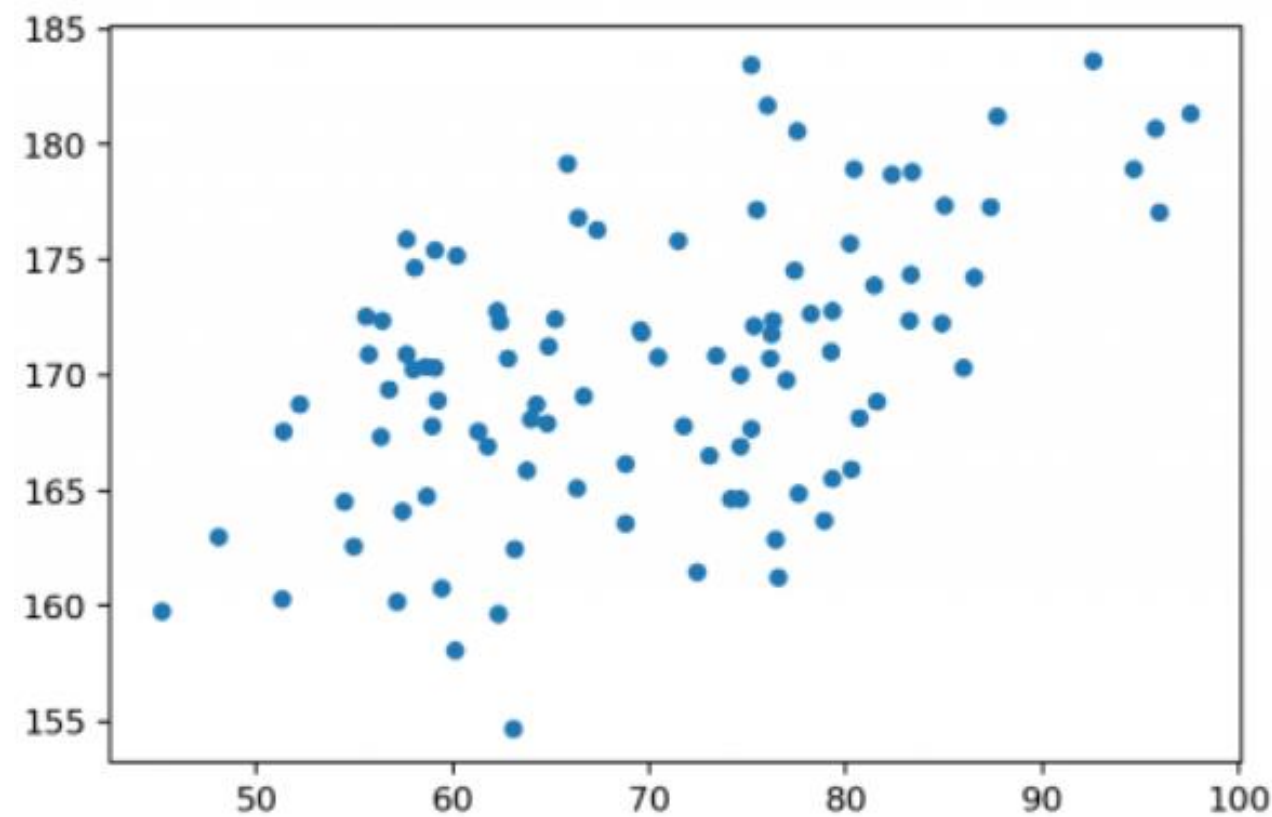
Bar-Graph, Histogram and Boxplot

- Bar graph: plot yang menyajikan data dalam bentuk batang persegi panjang yang panjangnya sebanding dengan nilai yang diwakilinya.
- Boxplot : Menggambarkan data numerik secara grafis melalui kuartilnya. Kotak tersebut terbentang dari nilai kuartil data Q1 hingga Q3, dengan garis di median (Q2).
- Histogram: Representasi sebaran data.



Scatterplot, Pieplot

- Scatterplot : Menampilkan data sebagai kumpulan poin.
Syntax: `dataframe.plot.scatter(x = 'x_column_name', y = 'y_columnn_name')`
- Pie plot : Representasi proporsional dari data numerik dalam kolom.
Syntax: `dataframe.plot.pie(y='column_name')`




Referensi

- Informasi lebih lanjut tentang alat EDA dan Pandas dapat ditemukan di tautan di bawah ini:
 - https://pandas.pydata.org/docs/user_guide/index.html
 - https://pandas.pydata.org/docs/user_guide/missing_data.html
 - https://pandas.pydata.org/docs/user_guide/visualization.html

Latihan

1. Data dapat ditemukan pada link berikut: <http://becomingvisual.com/python4data/tv.csv>
 - a) Buatlah Statistika Deskriptif dari data tersebut dengan menggunakan pandas.
 - b) Buatlah scatter plot yang menampilkan hubungan antara peringkat rata-rata (x) dan jumlah musim (y) dengan menggunakan pandas.



TERIMA KASIH