YOUrClass

Basic Data Science

Afifa Ayu Widhiyanthi

Chairperson at Data Science Indonesia

Experiences

- Data Analyst at INA Digital
- B.S in Computer Science Bina Nusantara University, Jakarta
- B.S in Statistics Bina Nusantara University, Jakarta
- President, Bina Nusantara Statistics Student Association 2019
- ex Tokopedia, ex Trevo

Afifa Widhiyanthi @afifadayu

YOUrClass

Data Science 101

Week 1

YOUrClass

Konten

YOUrClass

Pengenalan Data Science

1

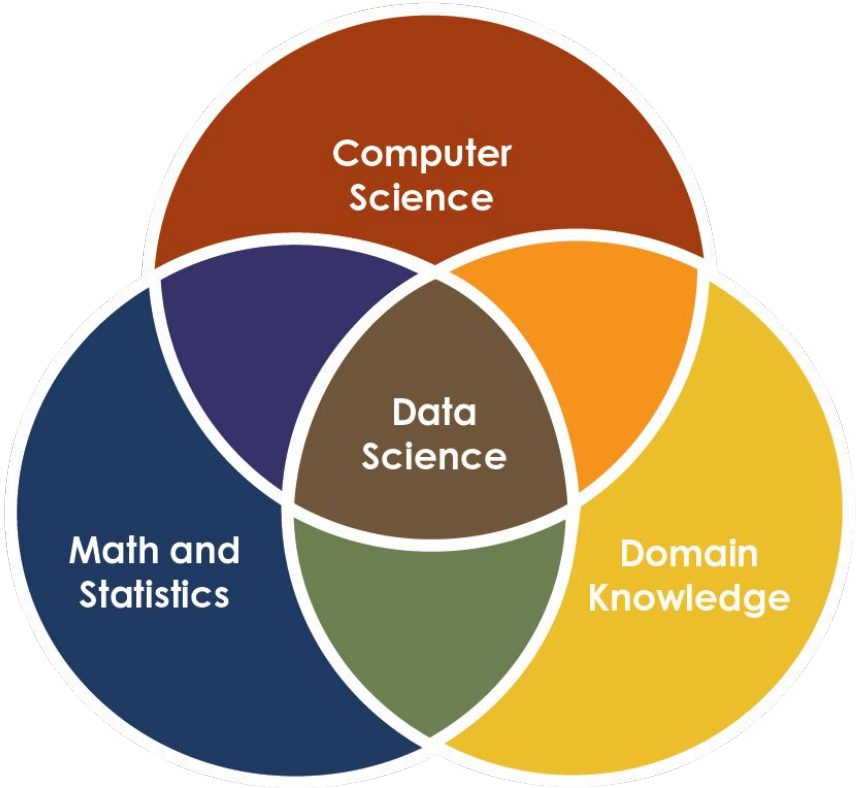
YOUrClass

Kumpulan fakta berisi angka, tulisan, atau pengukuran

Proses memahami dunia melalui observasi, eksperimen, dan logika

Proses dalam memahami sebuah masalah menggunakan data.

YOUrClass



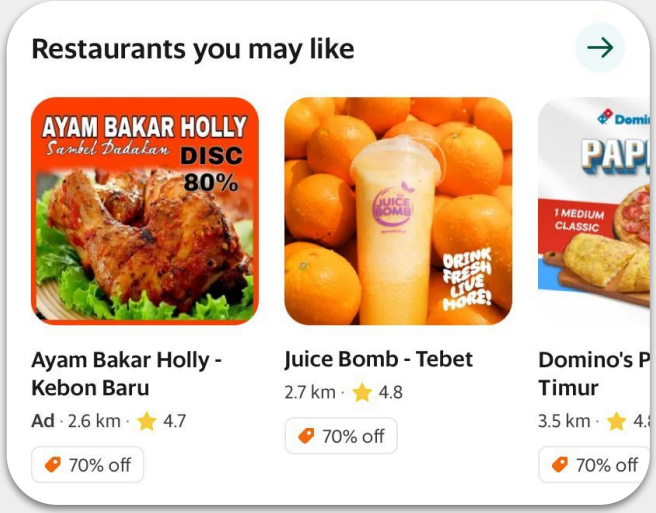
Source: datascience-pm

YOUrClass

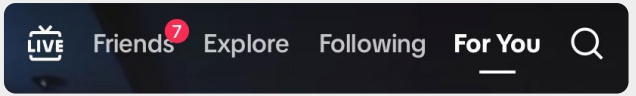
Dimana ada data?

- Sistem rekomendasi dalam *tech company* (*e-commerce*/*ride-hailing*), biasanya kamu lihat tulisan seperti ini "Anda mungkin suka"
- Pada media sosial, konten yang kamu **sukai** atau **komentari** rekomendasinya akan lebih sering muncul di "For You"

Rekomendasi pada aplikasi *online*



Rekomendasi pada *social media*



YOUrClass

Data tersebut menghasilkan

Big Data

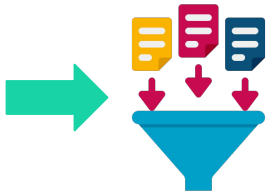
YOUrClass

Big Data

Ketika **volumenya sangat besar** yang diproses oleh mesin, alat, atau manusia. Sehingga, tidak bisa ditangani oleh penyimpanan tradisional.

Use case,

Menonton Netflix



Data dikumpulkan

Data disimpan

Maksimum 17 miliar sel
Baris = 1.048.576
Kolom = 6.384 kolom (XFD)

Maksimum 10 juta sel
Kolom = 18.278 kolom/tab

Source: kumparan, Freepik

YOUrClass

Ciri-ciri Big Data

disebut sebagai **5V**

1. **Volume**

   Data pengguna YouTube setiap hari.

2. **Velocity**
   (kecepatan)

   Transaksi *e-commerce* yang terjadi setiap detik.

3. **Variety**
   (variasi)

   Chat WhatsApp, story Instagram, lokasi GPS.

4. **Veracity**
   (keakuratan)

   Data sensor cuaca yang rusak.

5. **Value**
   (nilai)

   Data perilaku pelanggan (behaviour) untuk Ads/iklan.

YOUrClass

Langkah lanjutan

Menjadi **Storyteller**

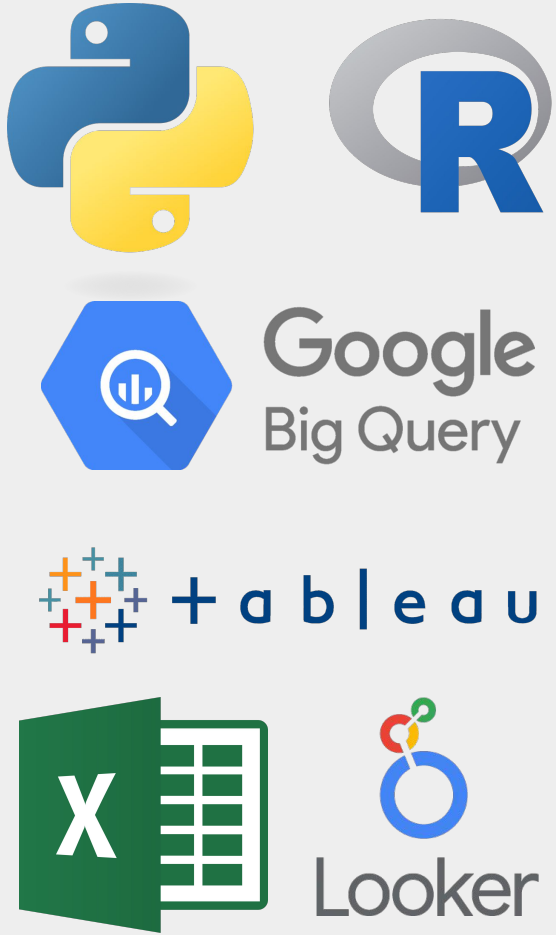
Tujuannya untuk **memberikan informasi** yang terdapat pada **data tersebut kepada orang lain**, sehingga seseorang dapat memahami dan mengetahui informasi yang kita ketahui.

YOUrClass

Tools

Tools/alat tersebut membantu kita melihat dan memahami tren dan pola dalam data.

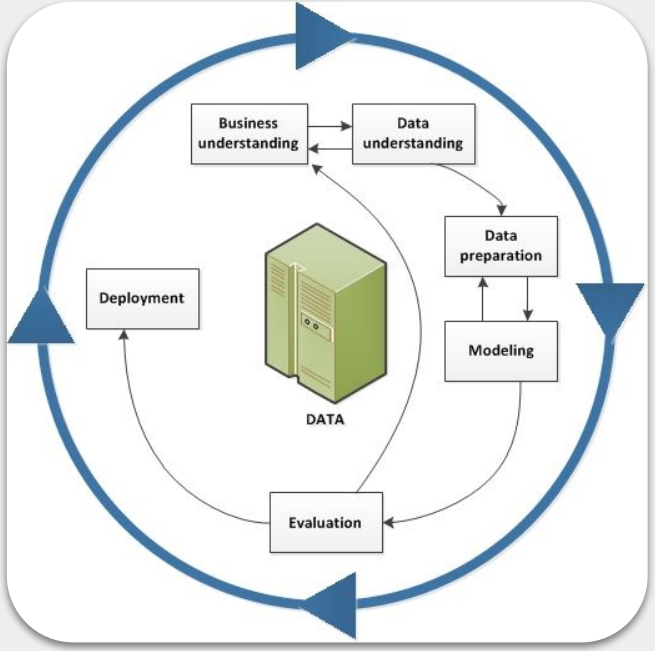
1. Python
2. R
3. SQL (Google BigQuery)
4. Microsoft Excel
5. Tableau
6. Google Looker



YOUrClass

Framework Data Science

Framework/rangka Data Science bertujuan agar pekerjaan menjadi lebih terstruktur, efisien, dan hasilnya relevan dengan kebutuhan **bisnis atau pengguna**.

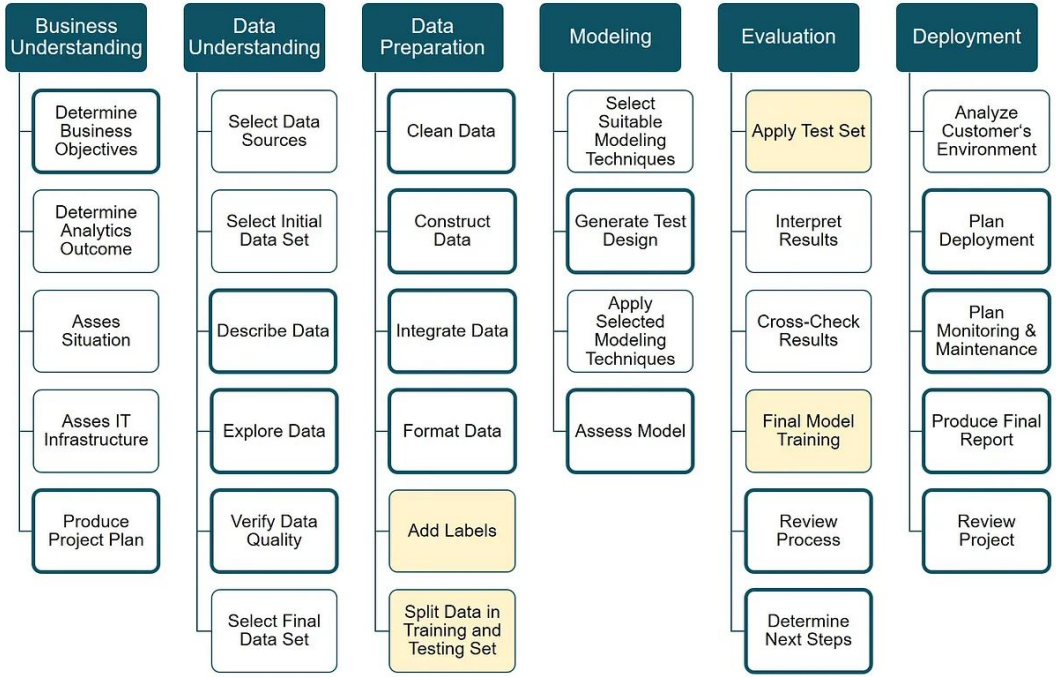


CRoss-Industry Standard Process for Data Mining [Proses Standar Lintas Industri untuk Data Mining]

- Sebuah proses yang memiliki enam fase untuk menggambarkan Data Science Life Cycle.

YOUrClass

CRISP-DM



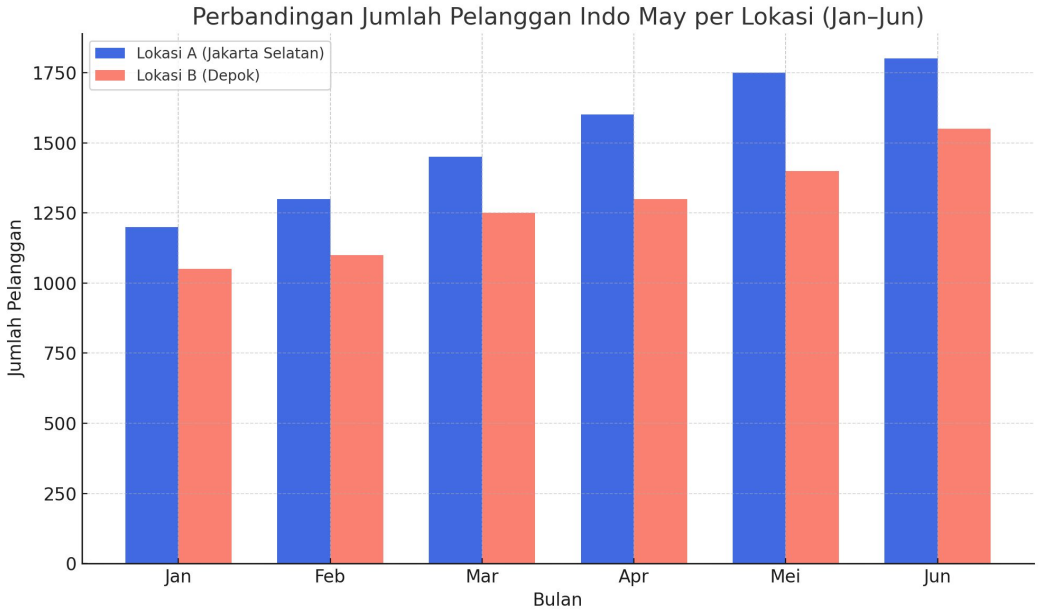
YOUrClass

Jumlah pelanggan Indo May pada 2 lokasi berbeda selama 6 bulan terakhir

Bulan	Lokasi A (Jakarta Selatan)	Lokasi B (Depok)
Januari	1.200	1.050
Februari	1.300	1.100
Maret	1.450	1.250
April	1.600	1.300
Mei	1.750	1.400
Juni	1.800	1.550

YOUrClass

Jumlah pelanggan Indo May pada 2 lokasi berbeda selama 6 bulan terakhir



YOUrClass

Data Science bukan tentang model paling keren, tapi tentang menyelesaikan **masalah nyata dengan cara paling bijak**

YOUrClass

Pengenalan Python

2

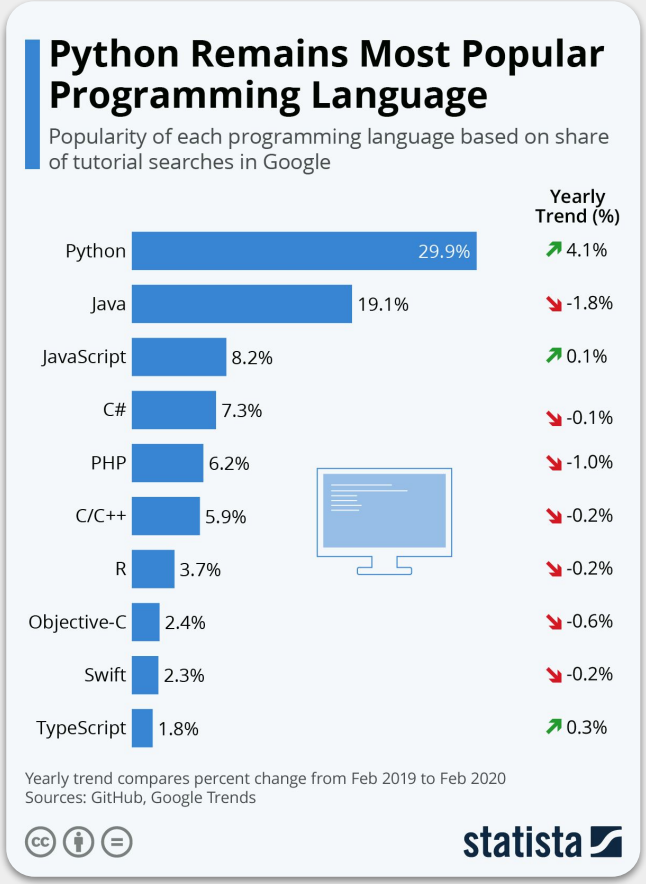
YOUrClass

Python

Bahasa pemrograman tingkat tinggi, dengan pengaplikasian dalam pemrograman web, penulisan skrip, komputasi ilmiah, dan kecerdasan buatan (AI).

Kenapa Python?

- Mudah dipelajari
- Komunitas yang besar
- Memiliki **library** untuk berbagai kebutuhan
- Kode yang mudah dibaca dan maintain.



Source: statista

YOUrClass

Library

Potongan kode yang dapat digunakan kembali. Library Python berisi kumpulan modul dan package terkait.

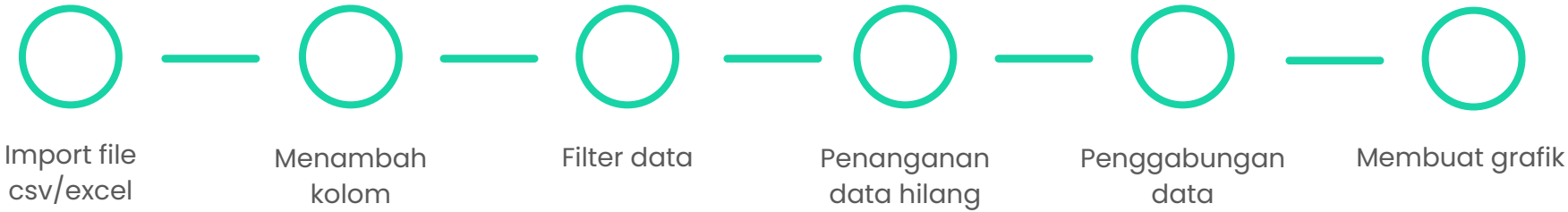
Source: DATARUNDOWN

YOUrClass

Pandas

Pandas berasal dari **Panel Data**. Panel Data terdiri dari pengamatan selama beberapa periode waktu untuk individu yang sama. Diimplementasikan pada tahun 2008 oleh Wes McKinney.

Pandas dapat melakukan, sebagai berikut:



YOUrClass

Fungsi dan Parameter

Fungsi adalah kode yang hanya berjalan saat dipanggil. Fungsi dapat menerima input saat dipanggil untuk menentukan sebuah perintah yang akan dijalankan yang disebut parameter.

Contoh:

Fungsi(Parameter)

Input : **print("hello world")**

Output : **hello world**

YOUrClass

Library

=Fungsi(Parameter)

F2 fx =SUMIF(C2:C12, "<"&F1, B2:B12)

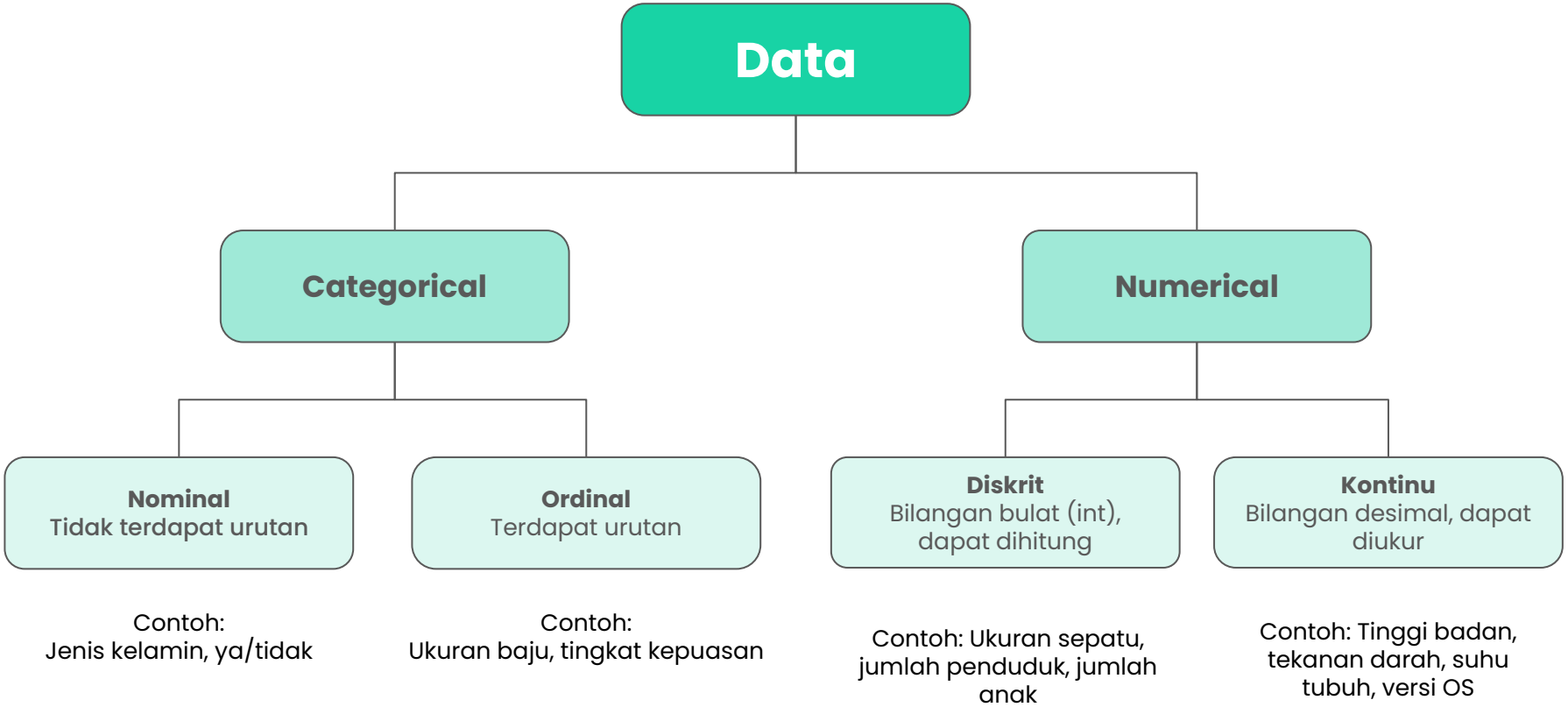
	A	B	C	D	E	F
1	**Item**	**Amount**	**Delivery date**		**Before**	4/11/2024
2	Apples	$250	4/10/2024		**Total**	$700
3	Bananas	$450	4/10/2024			
4	Oranges	$250	4/11/2024			

YOUrClass

Jenis-jenis Data

3

YOUrClass


Data
Categorical
Numerical
Nominal
Tidak terdapat urutan
Ordinal
Terdapat urutan
Diskrit
Bilangan bulat (int), dapat dihitung
Kontinu
Bilangan desimal, dapat diukur
Contoh:
Jenis kelamin, ya/tidak
Contoh:
Ukuran baju, tingkat kepuasan
Contoh: Ukuran sepatu, jumlah penduduk, jumlah anak
Contoh: Tinggi badan, tekanan darah, suhu tubuh, versi OS

YOUrClass

Latihan

Nama	Jumlah Buku Favorit	Tinggi Badan (cm)	Golongan Darah	Tingkat Minat Membaca	Waktu Tempuh ke Sekolah
Nina	3	158.2	A	Tinggi	25 menit
Bima	5	165.0	O	Sedang	15 menit
Sari	2	149.7	B	Rendah	40 menit

YOUrClass

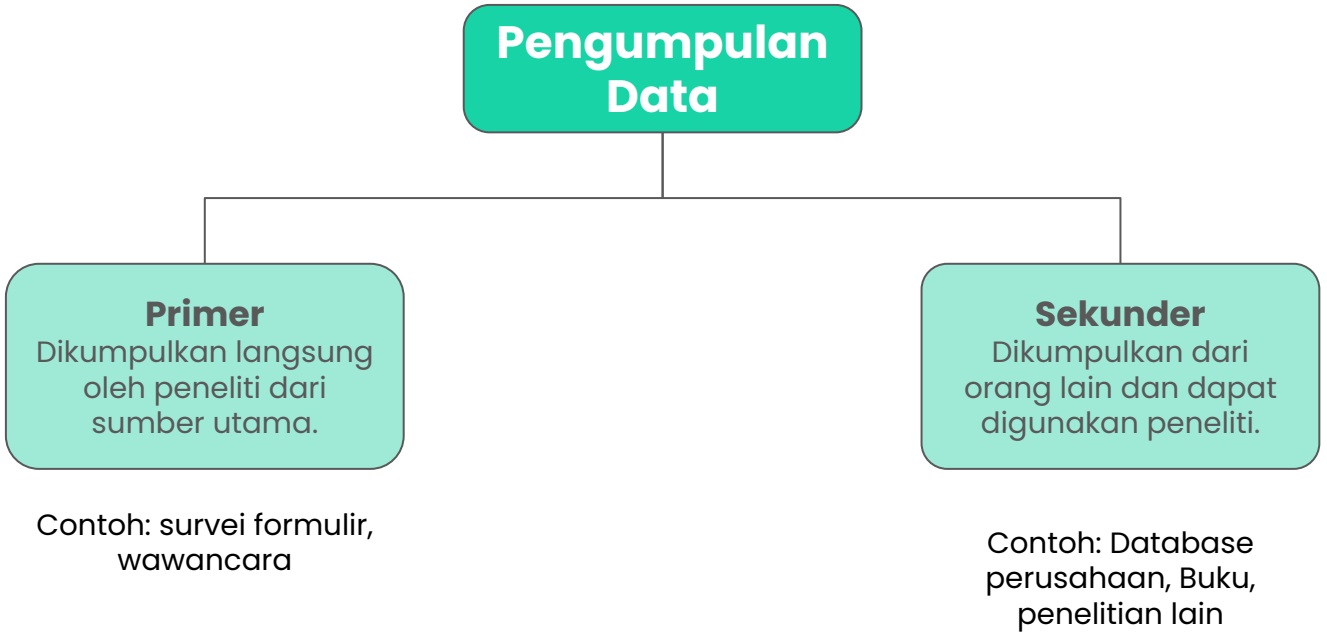
Latihan

Nama	Jumlah Buku Favorit	Tinggi Badan (cm)	Golongan Darah	Tingkat Minat Membaca	Waktu Tempuh ke Sekolah
Nina	3	158.2	A	Tinggi	25 menit
Bima	5	165.0	O	Sedang	15 menit
Sari	2	149.7	B	Rendah	40 menit

- Kontinu → Tinggi Badan (cm), Waktu Tempuh ke Sekolah
- Diskrit → Jumlah Buku Favorit
- Nominal → Nama, Golongan Darah
- Ordinal → Tingkat Minat Membaca

YOUrClass

Teknik pengumpulan data



YOUrClass

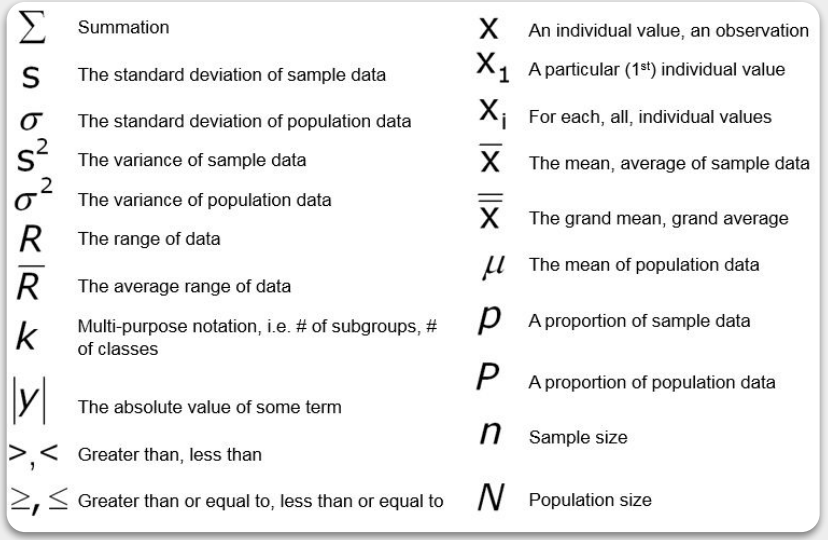
Statistika Dasar

3

YOUrClass

Statistika

Ilmu yang mempelajari **cara mengumpulkan, mengolah, menganalisis, menyajikan, dan menginterpretasi data** untuk membuat keputusan.

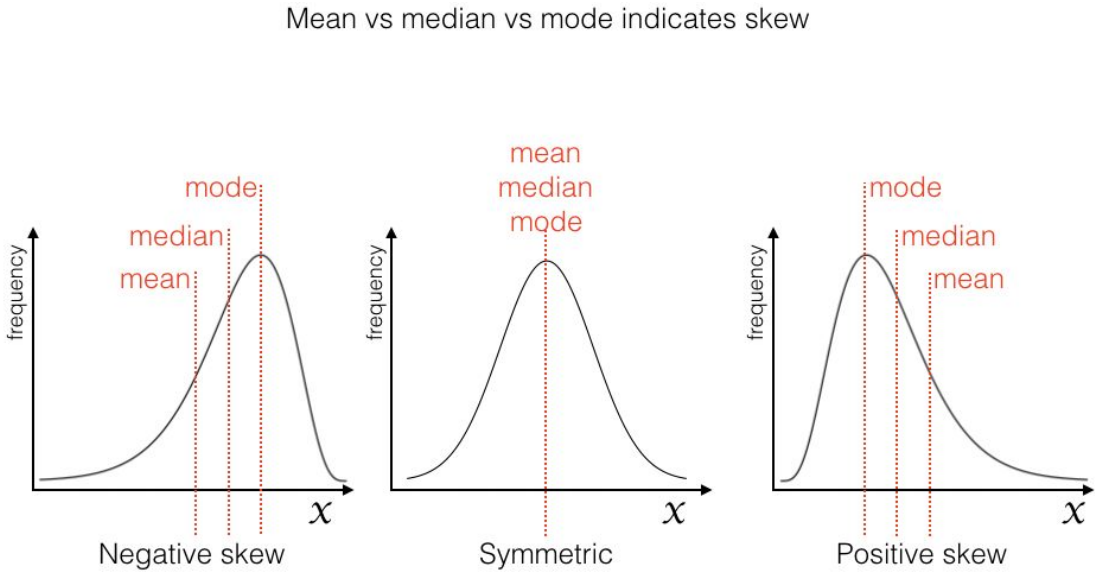


$\sum$	Summation	x	An individual value, an observation
s	The standard deviation of sample data	x_1	A particular (1st) individual value
σ	The standard deviation of population data	x_i	For each, all, individual values
s^2	The variance of sample data	$\bar{x}$	The mean, average of sample data
σ^2	The variance of population data	$\bar{\bar{x}}$	The grand mean, grand average
R	The range of data	μ	The mean of population data
$\bar{R}$	The average range of data	p	A proportion of sample data
k	Multi-purpose notation, i.e. # of subgroups, # of classes	P	A proportion of population data
$\lvert y \rvert$	The absolute value of some term	n	Sample size
$>, <$	Greater than, less than	N	Population size
$\geq, \leq$	Greater than or equal to, less than or equal to		

YOUrClass

Measure of Central Tendency

Sebuah teori pemusatan data yang terdiri dari: Mean, Median, Modus



YOUrClass

Measure of Variability

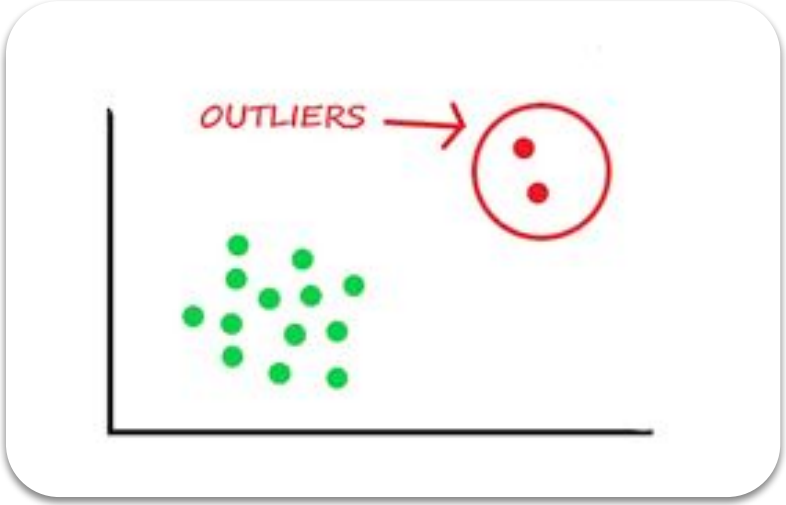
Sebuah teori penyebaran data yang terdiri dari: Range, Variance, Standar Deviasi

1. Range → Jarak antara data terbesar dan terkecil
2. Variance → Mengukur rata-rata perbedaan dari setiap poin
3. Standard Deviation → Menunjukkan **seberapa tersebarnya** data terhadap nilai rata-rata
    a. SD rendah : sebagian besar data berada di dekat nilai rata-rata
    b. SD tinggi : data lebih tersebar

YOUrClass

Outlier

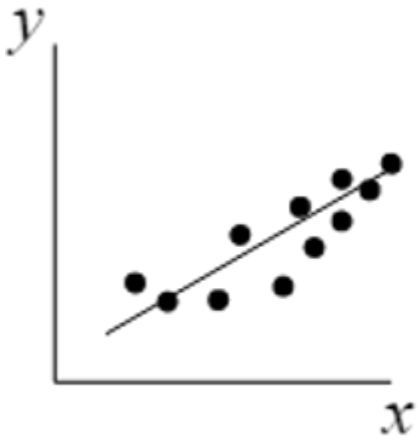
Data yang paling berbeda dari data lainnya, sehingga Outlier dapat memengaruhi nilai mean & standar deviasi.



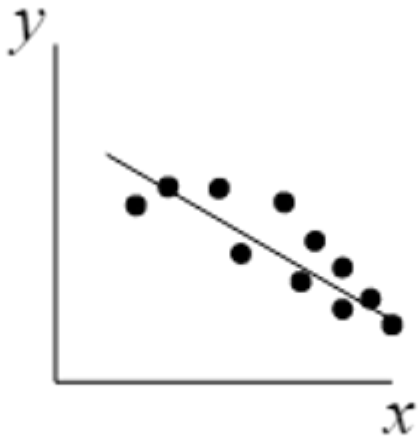
YOUrClass

Korelasi

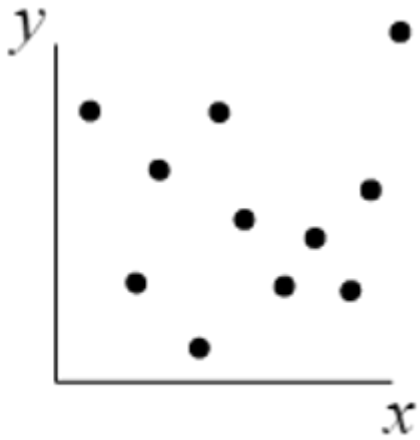
Mengukur hubungan antara dua variabel, Nilainya berisi rentang antara -1 hingga +1.



Positive



Negative



No correlation

YOUrClass

Hands On!

YOUrClass

Terima kasih

Afifa Widhiyanthi

@afifadayu