

Exploratory Data Analysis

| Bootcamp Data Science



Kurnia Anwar Ra'if

A Highly-motivated Data Scientist

Senior Data & AI Platform @ PT. Mastersystem Infotama

Data Scientist @ PT. KitaLulus International

Data Scientist @ PT. Sharing Vision– BRI Consultant

Software Engineering @ PT. AILIMA Geothermal

Mentor & Instructor DS/BI/AI ML @ dibimbing.id



Outline

Outline:

Exploratory Data Analysis (EDA)

- A. Introduction & Why EDA is important ?
- B. Statistical Descriptive : mean, median, mode, and standard deviation
- C. Technique Data Cleaning : Duplicate Data and Missing Data

A decorative graphic in the top-left corner consisting of a grid of colored squares: a yellow square, a grey square, a dark blue square, an orange square, a grey square, and a yellow square.

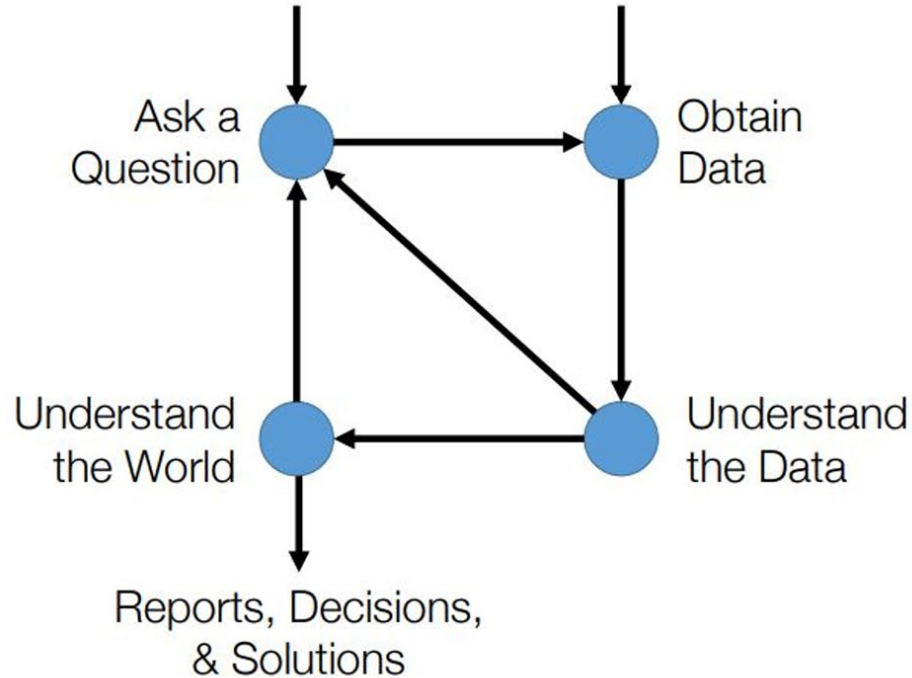
What is Exploratory Data Analysis (EDA)?

- An exercise to explore the data, to uncover insights from it
 - I.e. to really understand the data
- How exactly to do the “exploration”?
 - Essentially by looking at various angles of the data
- In today’s session, we will learn several **standard techniques** to perform EDA



Why EDA is important ?

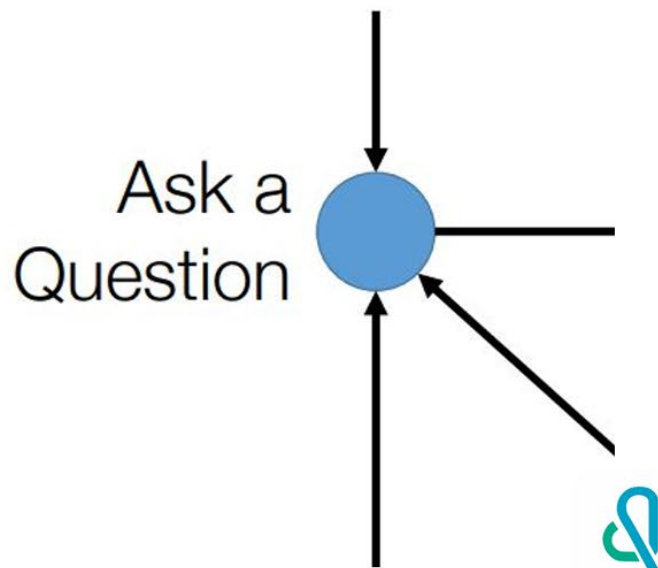
Data Science Process



Why EDA is important ?

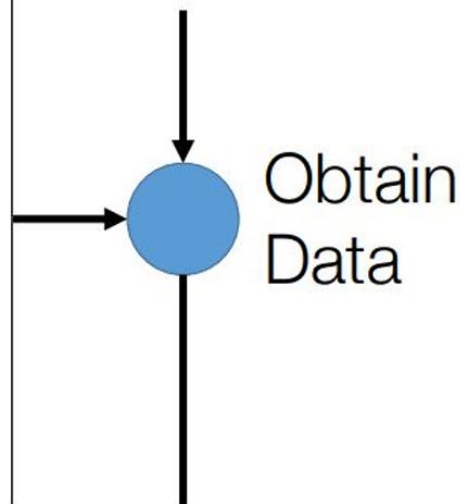
Question / Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics of success?



Why EDA is important ?

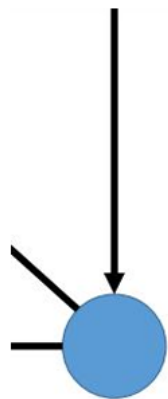
Data Acquisition and Cleaning



- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



Why EDA is important ?



Understand
the Data

- How is our data organized and what does it contain?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?

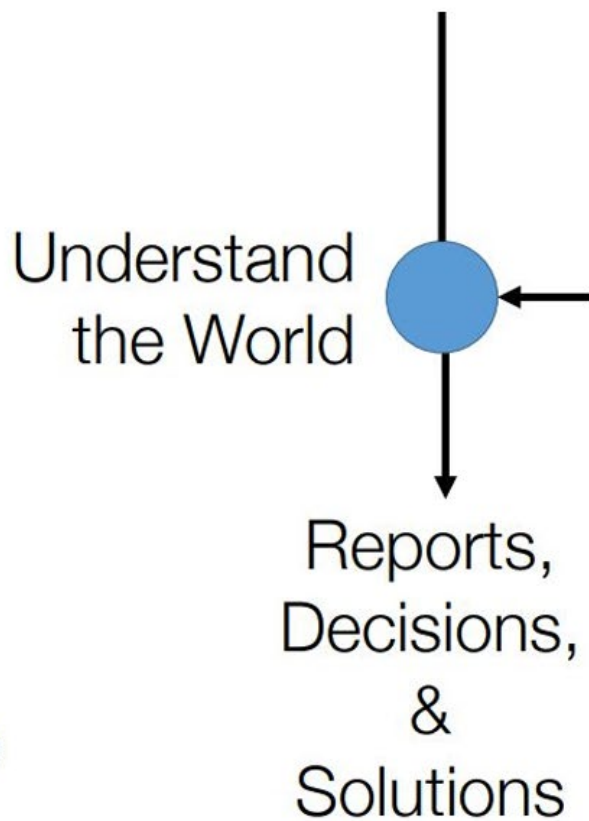
Exploratory Data Analysis & Visualization



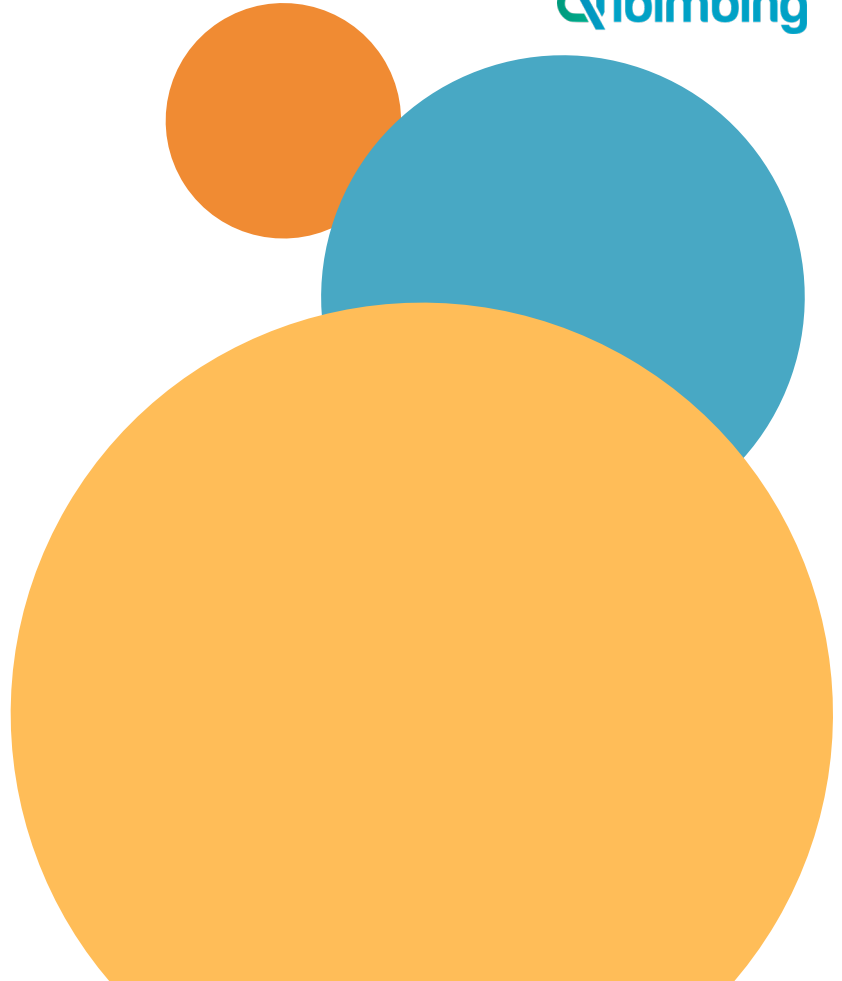
Why EDA is important ?

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?

Predictions and Inference



Statistical Descriptive :
mean, median, mode, and
standard deviation



Mean, Median, dan Modus

1 MEAN

Rata-rata adalah penjumlahan dari setiap nilai dibagi dengan banyaknya data.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$



Mean, Median, dan Modus

1 MEAN

Contoh:

Kita memiliki data nilai mahasiswa sebagai berikut

5,5,6,7,7,7,7,9,9,10

Carilah mean dari data berikut.

$$\begin{aligned}\text{mean} &= (5+5+6+7+7+7+7+9+9+10)/10 \\ &= 72 / 10 \\ &= 7.2\end{aligned}$$



Mean, Median, dan Modus

1 MEAN

Sifat Rata-rata (Mean)

Rata-rata sangat terpengaruh oleh nilai yang sangat besar sekali atau kecil sekali.

Nilai ini biasa disebut dengan outliers. **Tidak disarankan untuk menggunakan rata-rata apabila terdapat outliers pada data kita.**

Sifat ini dinamakan **not robust**



Mean, Median, dan Modus

2 MEDIAN

Median adalah nilai tengah data yang terletak di urutan ke-50% dari keseluruhan data jika data telah diurutkan dari yang terkecil ke terbesar

Median

n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

n is even,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2} + 1\right)^{th} \text{ observation}}{2}$$



Mean, Median, dan Modus

2 MEDIAN

Contoh:

Kita memiliki data nilai mahasiswa sebagai berikut

5,5,6,7,7,7,7,9,9,10

Carilah mean dari data berikut.

Banyaknya data = 10 (genap)

$$\begin{aligned}\text{Median} &= (x-5 + x-6)/2 \\ &= (7+7)/2 = 7\end{aligned}$$



Mean, Median, dan Modus

2 MEDIAN

Sifat Median

Median relatif **robust** terhadap outlier (tidak terpengaruh oleh nilai yang sangat tinggi atau rendah)

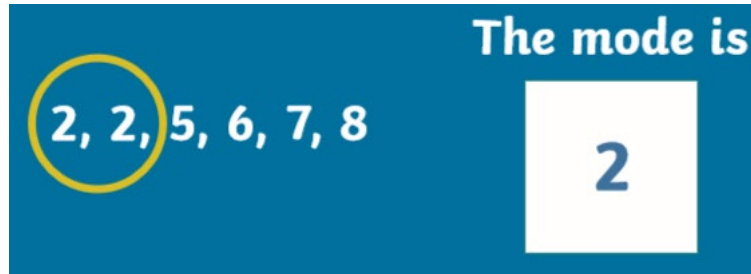
Biasanya digunakan untuk **distribusi skew** (menceng kiri/kanan)



Mean, Median, dan Modus

3 MODUS

Modus adalah data yang paling sering muncul/ memiliki frekuensi paling tinggi.



Mean, Median, dan Modus

2 MODUS

Sifat Modus

Modus biasa digunakan untuk **data yang bertipe kategorikal**
(bukan data numerik)

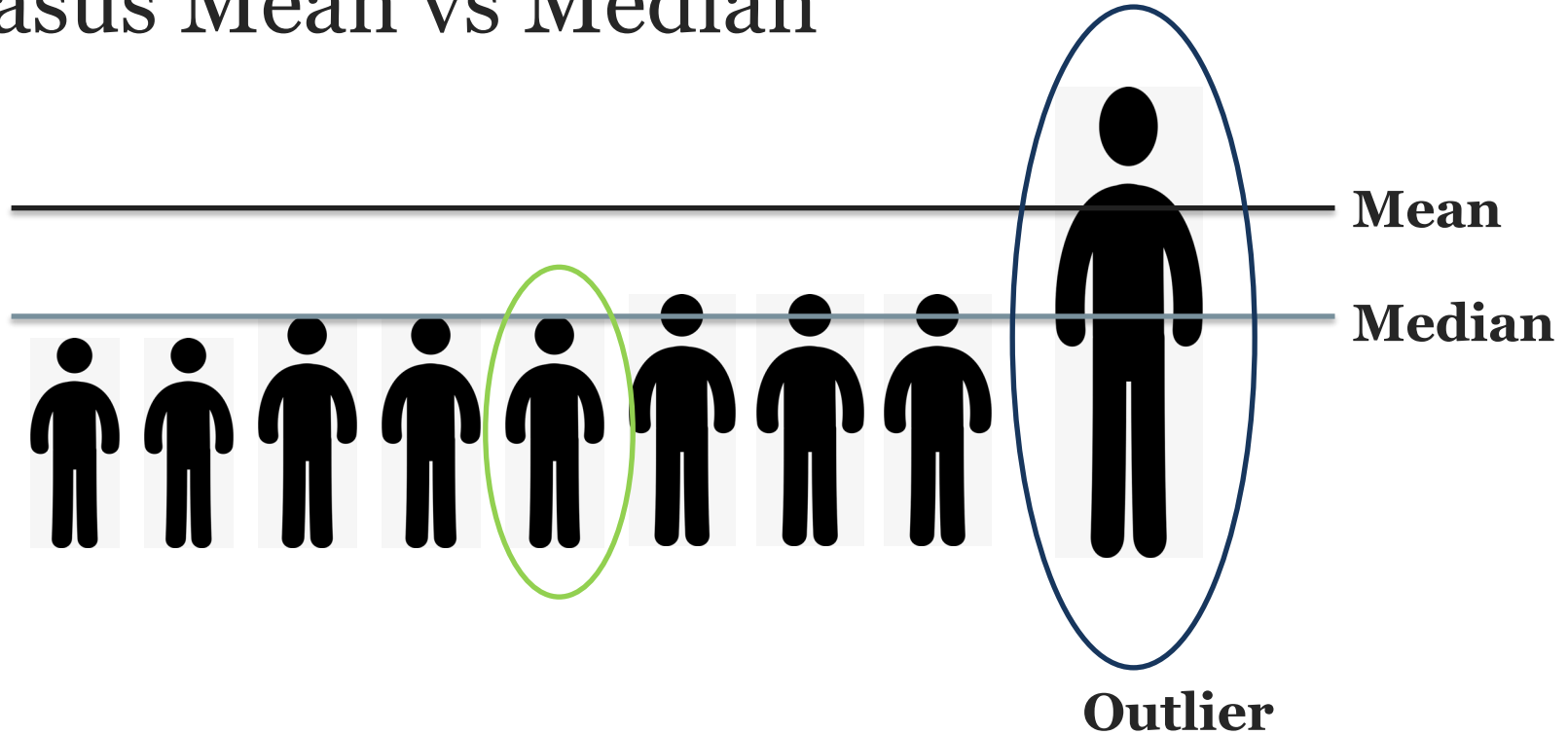
Contoh:


Gender = [F, F, M, M, M, F, M, F, F, F]

Modus = F



Kasus Mean vs Median





Variansi, Standar deviasi, Kovariansi, dan Korelasi



Variansi dan Standar deviasi

1

VARIANSI

- Rata-rata kuadrat selisih setiap data dari mean
- Mengukur persebaran dari sebuah data

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

2

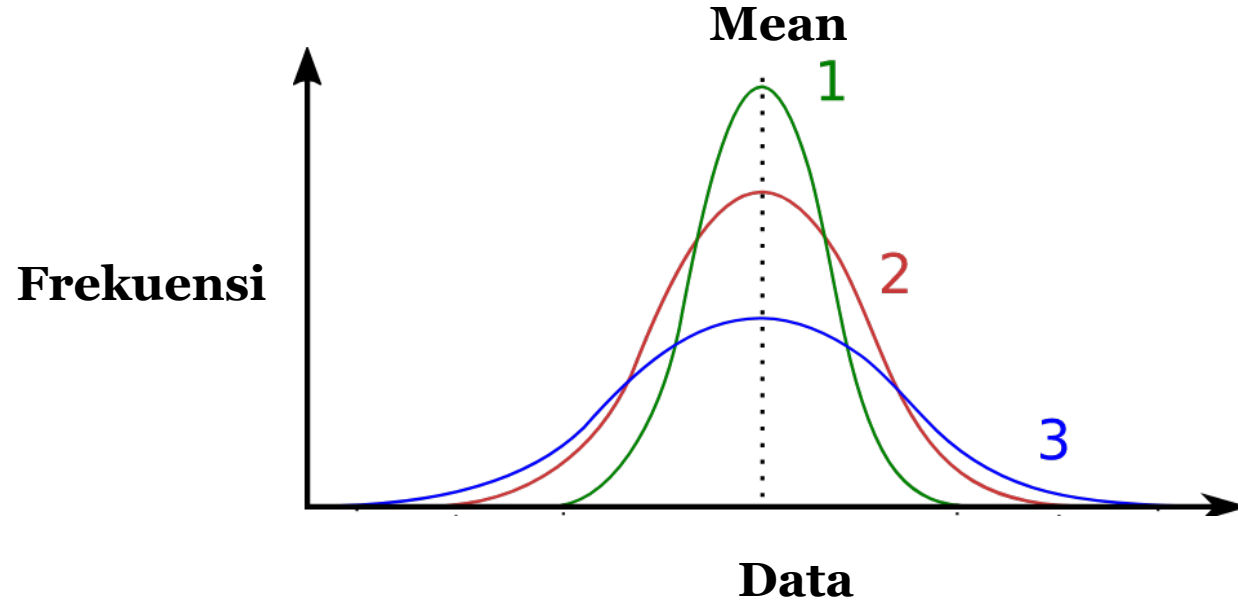
STANDAR DEVIASI

- Akar dari variansi (agar skalanya sebanding dengan data asal)
- Mengukur persebaran dari sebuah data

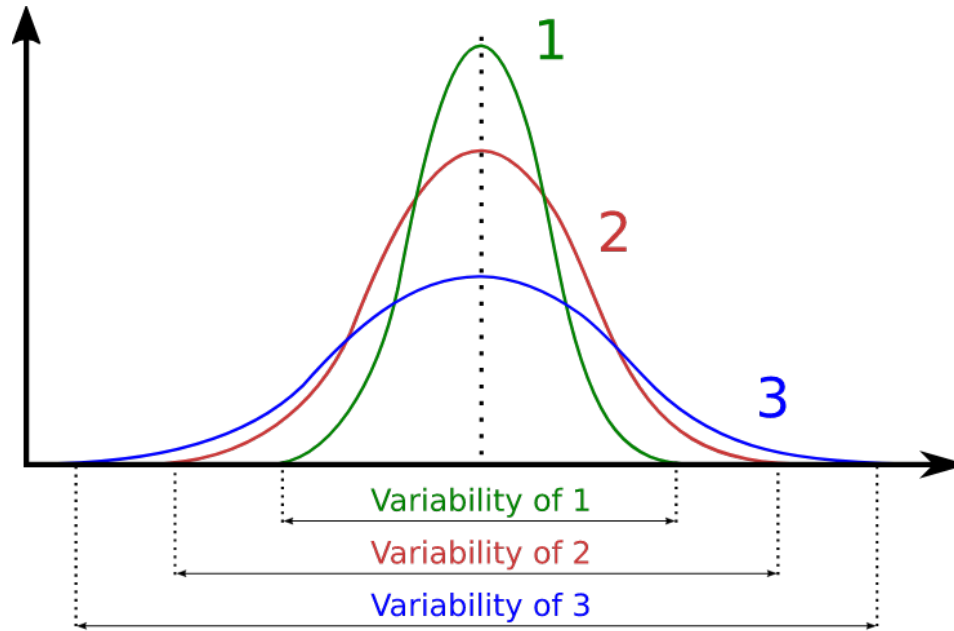
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$



Variansi dan Standar deviasi



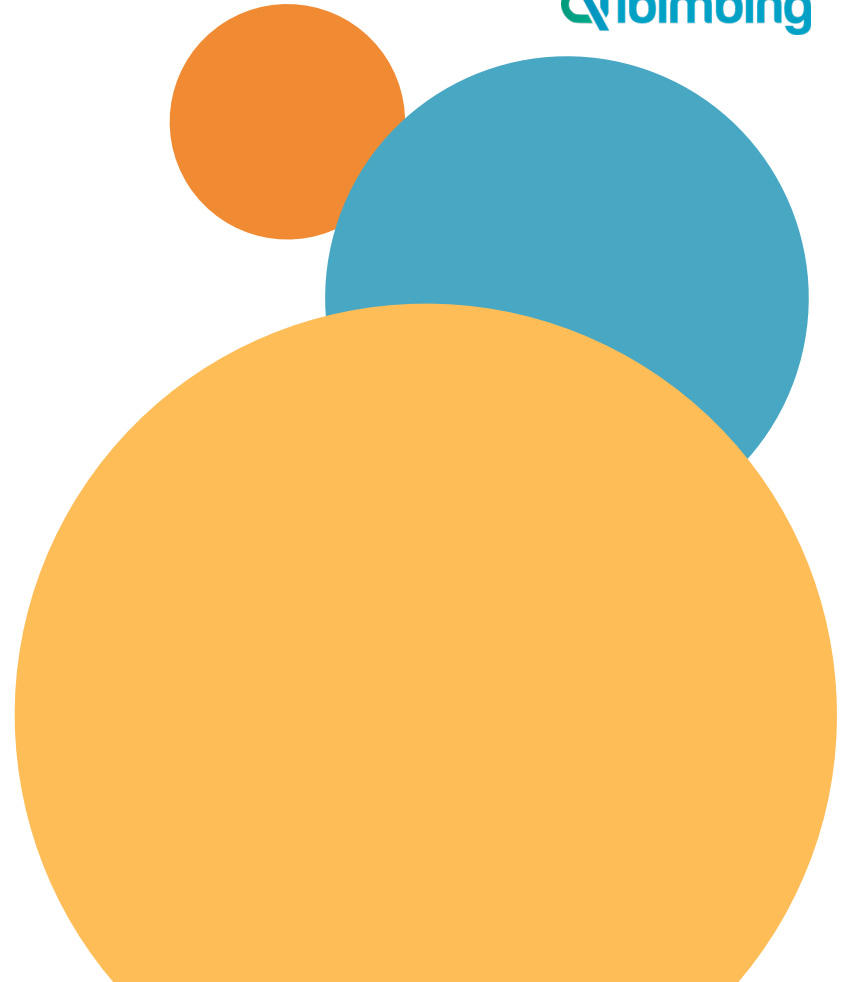
Variansi dan Standar deviasi



Technique Data Cleaning : Duplicate Data and Missing Data



Data Frame



THE PANDAS DATAFRAME

Pandas adalah library dasar yang digunakan untuk melakukan manipulasi pada data frame. Data Frame adalah tabel/data tabular dengan array dua dimensi yaitu baris dan kolom.

	0	1	2	3	4	
	Name	Age	Marks	Grade	Hobby	
0	S1	Joe	20	85.10	A	Swimming
1	S2	Nat	21	77.80	B	Reading
2	S3	Harry	19	91.54	A	Music
3	S4	Sam	20	88.78	A	Painting
4	S5	Monica	22	60.55	B	Dancing

Diagram illustrating a Pandas DataFrame structure with annotations:

- Column Label/ Header:** Points to the header row (Name, Age, Marks, Grade, Hobby).
- Index Label:** Points to the row index labels (S1, S2, S3, S4, S5).
- Column Index:** Points to the column headers (0, 1, 2, 3, 4).
- Row Index:** Points to the row index labels (0, 1, 2, 3, 4).
- Row:** Points to the entire row (S4).
- Column:** Points to the entire column (Marks).
- Element/ Value/ Entry:** Points to a specific cell (88.78).

Handle Duplikat

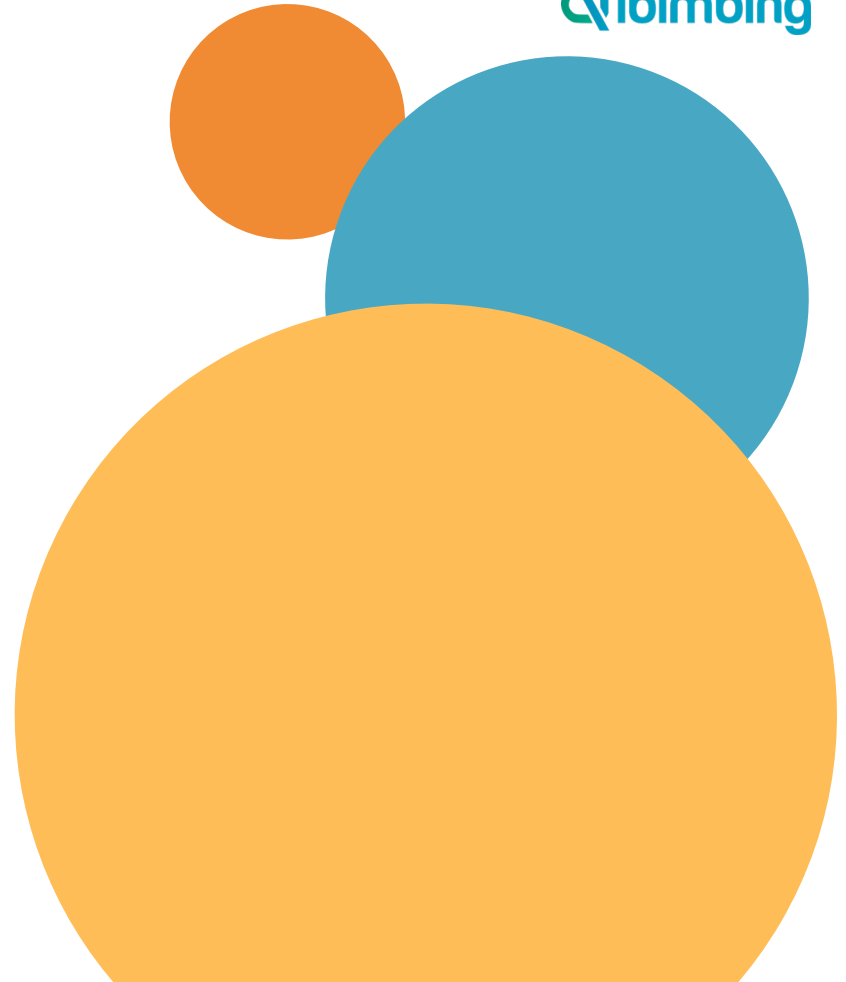
Machine Learning tidak bisa memproses data yang duplikat :

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
104	1	1	Eustis, Miss. Elizabeth Mussey	female	54.0	1	0	36947	78.2667	D20	C	4	NaN	Brookline, MA
349	1	1	Eustis, Miss. Elizabeth Mussey	female	54.0	1	0	36947	78.2667	D20	C	4	NaN	Brookline, MA

Maka harus dipilih salah satunya saja -> Handling Duplikat



Handling Missing Value





MISSING DATA

Terdapat berbagai cara untuk merepresentasikan data yang hilang atau nilai yang tidak sesuai didalam data.

```
df_company.sample(5)|
```

	Revenue	Size	Rating	Headquarters
512	2to5 billion (USD)	5001 to 10000 employees	3.5	Westminster, CO
521	10to25 million (USD)	51 to 200 employees	3.8	Arlington, VA
330	100to500 million (USD)	501 to 1000 employees	3.7	Lorton, VA
517	Unknown / Non-Applicable	51 to 200 employees	2.7	Chantilly, VA
129	2to5 billion (USD)	5001 to 10000 employees	NaN	Sunnyvale, CA

Missing values are treated as *np.NaN* when data is *not* read into Pandas

Missing values are treated as *np.NaN* when data is read into Pandas

Mengapa bisa ada missing value?



Free text

Or



Dropdowns

Data Entry Errors



Parsing Errors

IDENTIFYING MISSING DATA

Cara paling mudah untuk mengidentifikasi data yang hilang adalah dengan menggunakan :

- metode `.isna()`.
- Juga dapat menggunakan `.value_counts()`.

```
df_company.isna()
```

	Revenue	Size	Rating	Headquarters
0	False	False	True	False
1	False	False	False	False
2	False	False	False	False
3	False	False	False	False
4	False	False	False	False
5	False	False	False	False
6	False	False	False	False
7	False	False	True	False

This returns True for any missing values

```
df_company.isna().sum()
```

```
Revenue      0  
Size         0  
Rating       50  
Headquarters 0  
dtype: int64
```

Use `sum()` to return the missing values by column

```
df_company[df_company.isna().any(axis=1)]
```

	Revenue	Size	Rating	Headquarters
0	Unknown / Non-Applicable	1001 to 5000 employees	NaN	New York, NY
7	1 to 2 billion (USD)	1001 to 5000 employees	NaN	Bedford, MA
82	25 to 50 million (USD)	51 to 200 employees	NaN	Minneapolis, MN
84	25 to 50 million (USD)	51 to 200 employees	NaN	Austin, TX
92	1 to 2 billion (USD)	5001 to 10000 employees	NaN	Herndon, VA

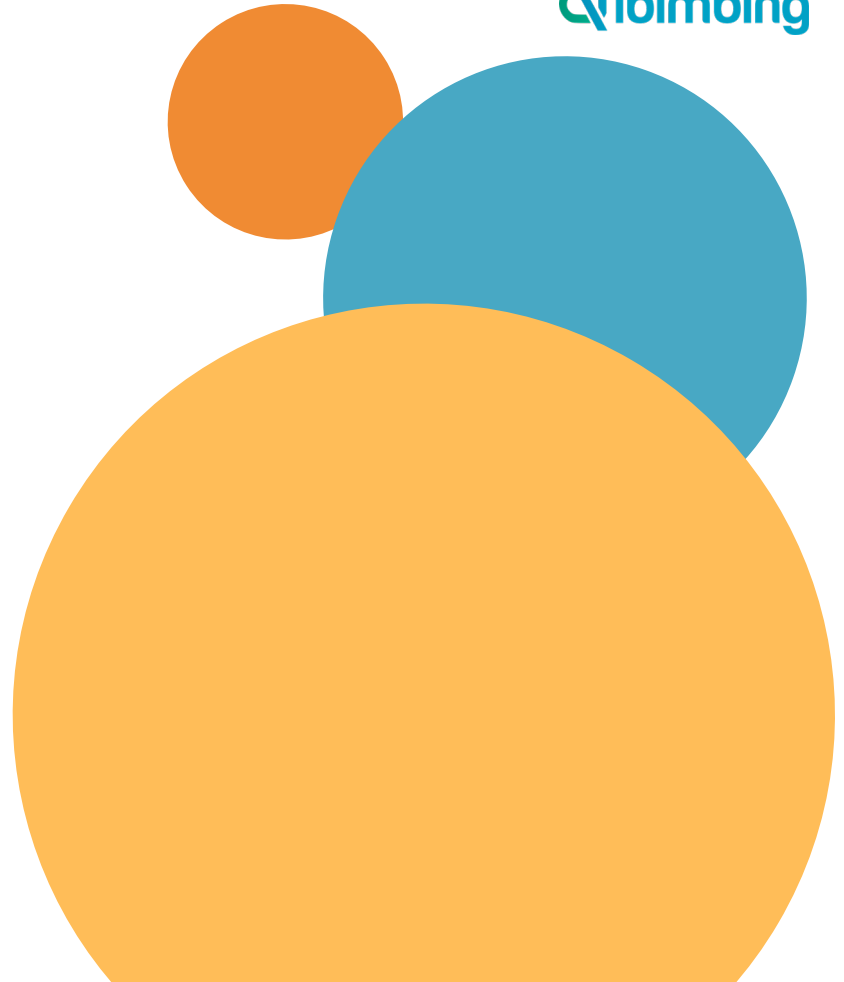
Or use `any(axis=1)` to select the rows with missing values

Flow Handling Missing Value

1. Cek terlebih dahulu, nilai unik di tiap kolom. Kemudian cek berapa persen missing value tiap kolom. **Secara Statistik jika > 20% maka drop kolomnya**, jika dibawah itu maka handling dengan cara :
2. Jika **numerikal** kolom maka handle menggunakan nilai **median** dari data train, lalu aplikasikan ke data train dan ke data test di kolom bersangkutan.
3. Jika kolom **kategorikal** maka handle menggunakan nilai **modus** dari data train, lalu aplikasikan ke data train dan ke data test di kolom bersangkutan.



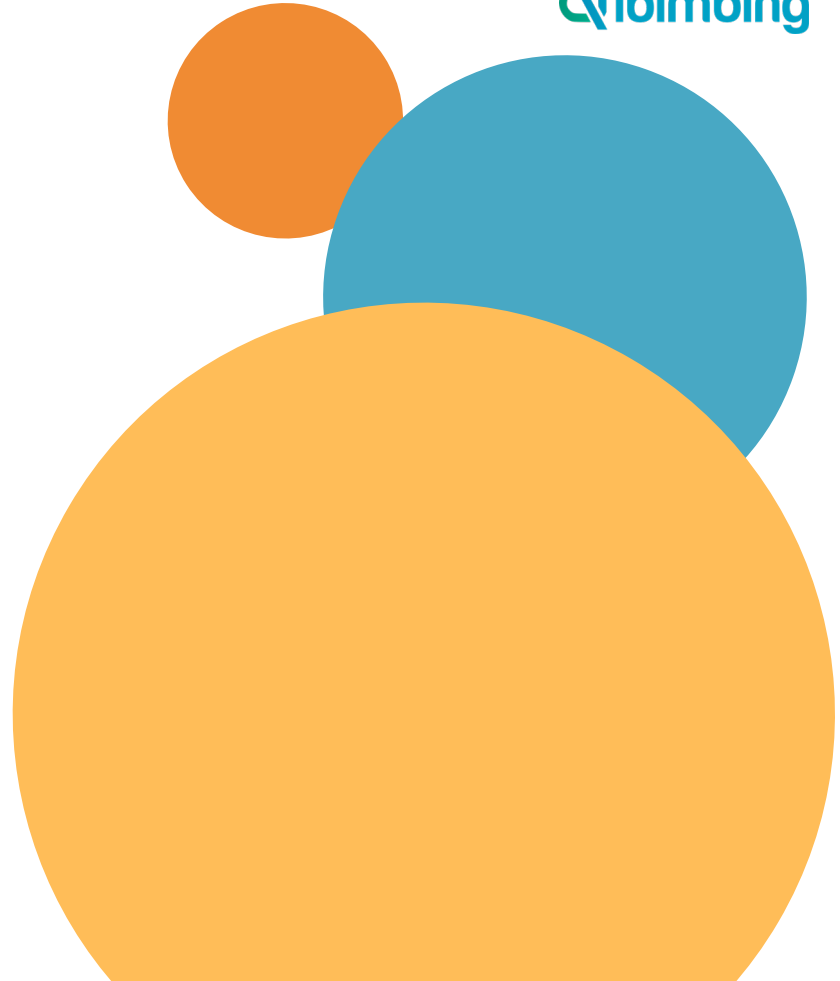
Hands on !



Assignment :

Jawablah pertanyaan pada template code dan cek link berikut :

<https://drive.google.com/drive/folders/1mkBZvoYfwn6RMm7MSIA3wl7n59JXNNjB?usp=sharing>



A large, stylized graphic on the left side of the slide. It consists of a blue outline of a person's head and shoulders. Inside the head is a series of concentric circles: a small yellow circle, followed by an orange ring, and then a larger yellow circle. The body is a large blue circle, and inside it is a large yellow circle with an orange ring in the center.

Thank you

 <https://www.linkedin.com/in/anwaraif/>

 kurniafreelancer@gmail.com