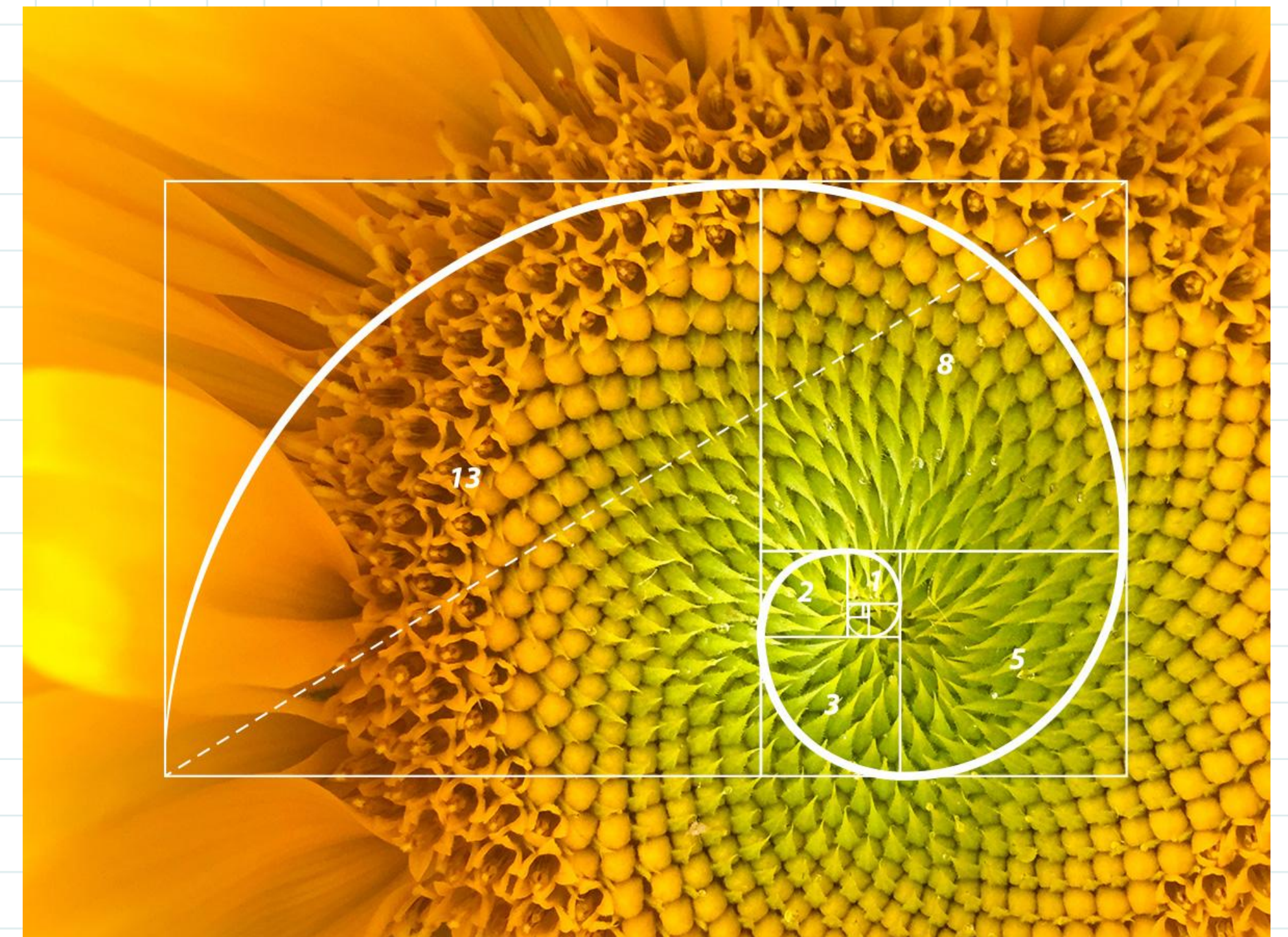


MACHINE LEARNING



Dasar Statistik & Probabilitas ML

Dr. Sirojul Munir, S.Si., M.Kom.
rojulman@nurulfikri.ac.id

ARTIFICIAL INTELLIGENCE – INFORMATICS STTNF



Pengantar Statistik Probabilitas

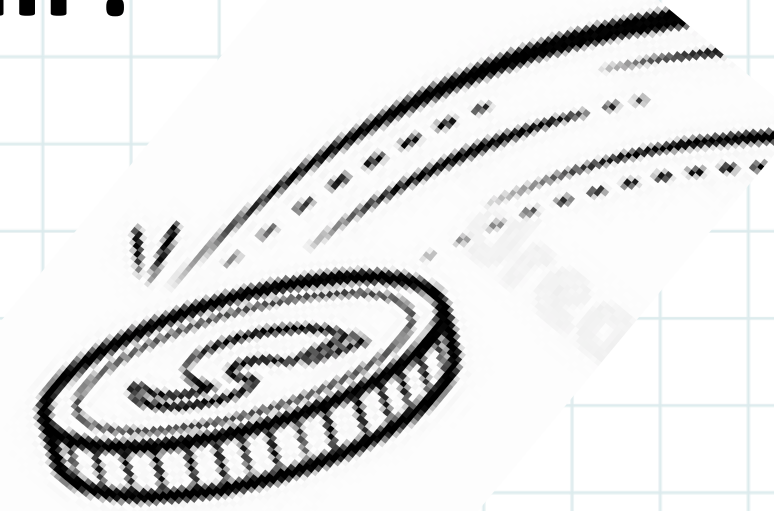
- **Statistik:** Ilmu yang mempelajari cara mengumpulkan, mengolah, menganalisis, menafsirkan, dan menyajikan **data** untuk menarik kesimpulan. Dengan fokus pada **data** hasil **pengamatan**, **survei** atau **eksperimen**
- **Probabilitas:** Ilmu yang mempelajari **peluang terjadinya suatu kejadian** berdasarkan teori dan model matematis. Dengan fokus pada **prediksi** atau **model ketidakpastian**
- ❖ **Statistik deskriptif:** merangkum data (mean, median, grafik)
- ❖ **Statistik Inferensial:** Generalisasi dari sampel ke populasi (estimasi, uji)
- ❖ **Bayesian** menyatukan keduanya : **data + Probabilitas (Prior) → Statistik (Posterior)**

Contoh Statistik Probabilitas

Dari 100 lemparan koin muncul 62 kepala, **Apakah koin fair?**



Estimasi $p \approx 0.62$



Lalu di uji hipotesisnya !, harusnya jika fair maka: $P(\text{Kepala}) = P(\text{Ekor}) = 0.5$

Ternyata: koin fisik bisa sedikit tidak fair karena berat/teknik melempar, setelah lakukan banyak lemparan dan uji, proporsi kepala berbeda signifikan dari 0,5 (uji binomial)

Koin tidak fair (bias): peluang tidak sama, mis. $P(\text{Kepala}) = 0.6$, $P(\text{Ekor}) = 0.4$.

Definisi Ruang Probabilitas - 1

- **Eksperimen** adalah suatu proses untuk mendapatkan hasil observasi dari suatu/beberapa fenomena. Sebuah kegiatan yang dilakukan pada suatu eksperimen disebut sebagai **percobaan** (*trial*), sedangkan hasil observasi dari percobaan tersebut dikenal sebagai **hasil** (*outcome*).
- Himpunan dari seluruh hasil yang mungkin muncul dari suatu eksperimen disebut sebagai **ruang sampel** (*sample space*) yang dinotasikan dengan S . Masing-masing anggota/elemen dari ruang sampel disebut sebagai **titik sampel** (*sample point*). Catat bahwa hanya terdapat satu hasil/satu titik sampel (dari seluruh kemungkinan hasil) yang muncul pada suatu percobaan dari eksperimen.



Definisi Ruang Probabilitas -2

- Apabila ruang sampel terdiri atas sejumlah titik sampel yang terhitung, maka ruang sampel tersebut dikatakan sebagai **ruang sampel terhitung** (*countable sample space*). Sebaliknya, apabila ruang sampel terdiri atas sejumlah titik sampel yang tak terhitung, maka ruang sampel tersebut dikatakan sebagai **ruang sampel kontinu** (*continuous sample space*).

Contoh:

Diketahui suatu eksperimen pelemparan dua buah koin logam, di mana masing-masing koin memiliki bagian muka (disebut M) dan bagian belakang (disebut B). Apabila kita ingin mengetahui seluruh kemungkinan hasil observasi dari eksperimen tersebut (disebut sebagai ruang sampel), maka kita peroleh:

$$S = \{MM, MB, BM, BB\}$$



Terlihat bahwa ruang sampel merupakan seluruh kemungkinan kombinasi antara notasi M (muka) dan B (belakang).



Latihan

1. Berikan ruang sampel hasil pertandingan sepakbola antara Persib vs Persija?
2. Berikan ruang sampel dari pelemparan sebuah dadu?



Diberikan sebuah eksperimen yang terdiri dari pelemparan sebuah dadu enam sisi. Hasil dari ruang sampel pada eksperimen tersebut merupakan himpunan $S = \{1, 2, 3, 4, 5, 6\}$. Masing-masing angka adalah titik sampel yang merepresentasikan mata dadu yang muncul ketika dadu tersebut dilemparkan. Catat bahwa titik sampel 1 dan 2 (atau lebih formal disebut sebagai $\{1\}$ dan $\{2\}$) merupakan salah satu contoh dari kejadian saling asing, karena kedua titik sampel tersebut tidak dapat muncul secara bersamaan dalam 1 kali pelemparan dadu. Kumpulan dari seluruh titik sampel 1 sampai 6 adalah kejadian lengkap untuk percobaan dari dadu karena dari salah satu titik pasti akan terjadi/muncul.

Latihan

Berikutnya, kita definisikan kejadian di dalam ruang sampel S sebagai berikut:

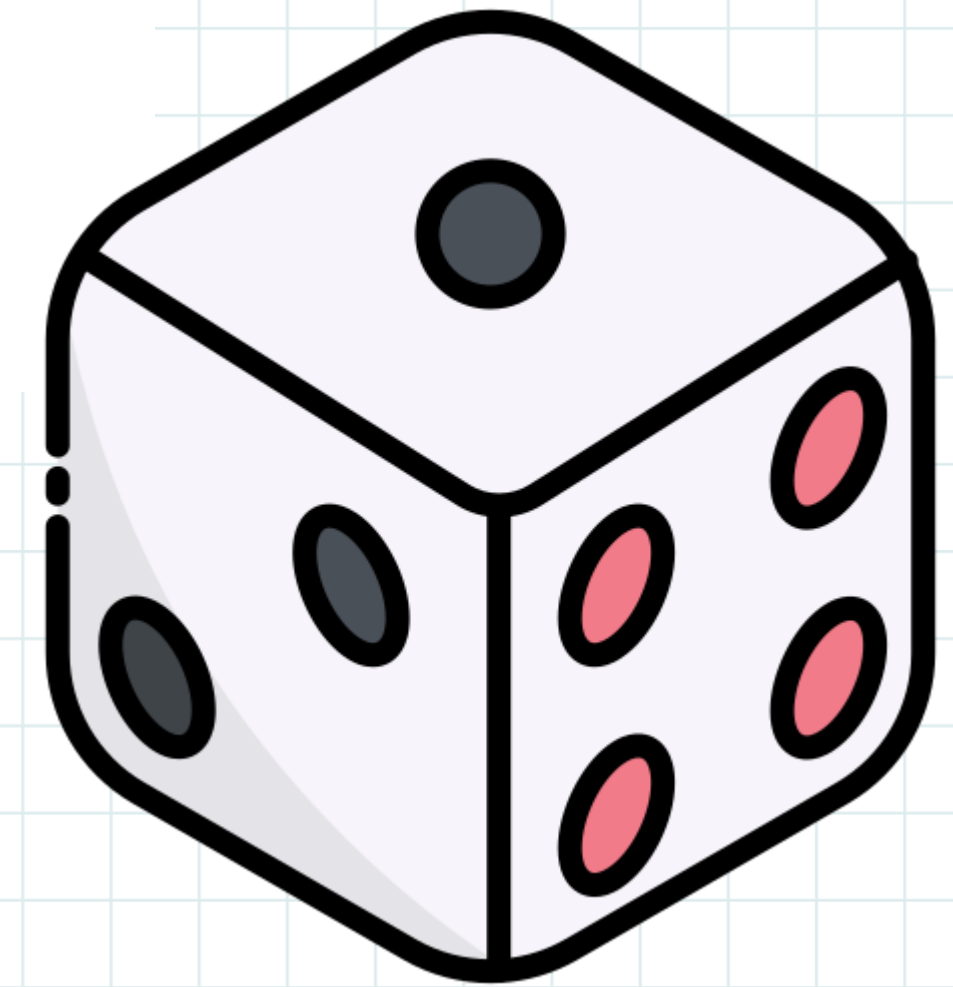
$A = \{1, 2, 3\}$ = “kejadian pelemparan mata dadu yang kurang dari 4”

$B = \{2, 4, 6\}$ = “kejadian pelemparan mata dadu genap”

$C = \{4\}$ = “kejadian pelemparan mata dadu 4”

$D = \{2\}$ = “kejadian pelemparan mata dadu 2”

$$S = \{1, 2, 3, 4, 5, 6\}$$



Dengan demikian dapat kita peroleh beberapa hasil sebagai berikut:

(i) $A \cup B = \{1, 2, 3, 4, 6\}$

(ii) $A \cap B = \{2\}$

(iii) A dan C adalah kejadian saling asing karena $A \cap C = \emptyset$

(iv) $D \subset B$

(v) $A' = \{4, 5, 6\}$ merupakan komplemen dari kejadian A

(vi) $B' = \{1, 3, 5\}$ merupakan komplemen dari kejadian B

(vii) $A \cup B = \{1, 2, 3, 4, 6\}$ sehingga $(A \cup B)' = \{5\} = A' \cap B'$ (Hukum DeMorgan)



Definisi Ruang Probabilitas -5

- Ruang sampel (S): semua kemungkinan.
- Kejadian (A): subset dari S .
- Peluang: $0 \leq P(A) \leq 1, P(S) = 1$.
- Operasi: komplemen A^c , irisan $A \cap B$, gabungan $A \cup B$
- Aturan:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

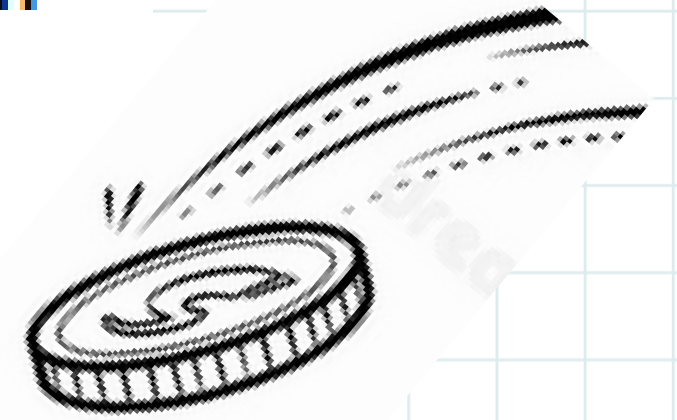


Contoh Statistik Probabilitas

Pada koin fair, hitung peluang 3 kepala dalam 5x lemparan !

Untuk koin fair ($p(\text{Kepala}) = 0.5$), peluang tepat 3 kepala dari 5 lemparan:

$$P(X = 3) = \binom{5}{3} (0.5)^3 (0.5)^2 = \binom{5}{3} (0.5)^5$$



Hitung:

- $\binom{5}{3} = \frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$
- $(0.5)^5 = \frac{1}{32} = 0.03125$

Jadi

$$P(X = 3) = 10 \times \frac{1}{32} = \frac{10}{32} = 0.3125 = \mathbf{31.25\%}.$$

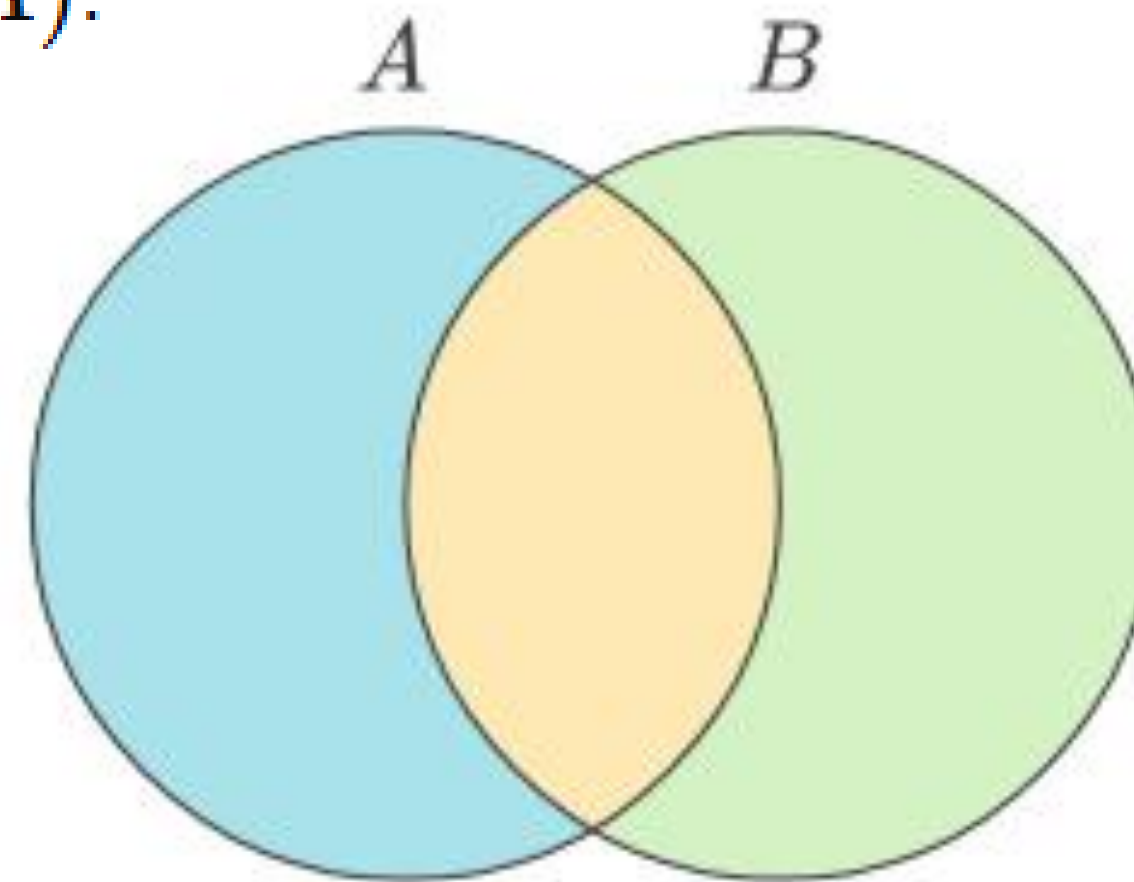
Definisi Ruang Probabilitas -6




Definisi:

- $P(A | B) = \frac{P(A \cap B)}{P(B)}$ (dengan $P(B) > 0$).

Aturan Perkalian:

$$P(A \cap B) = P(A | B) P(B) = P(B | A) P(A).$$



 $P(A)$
 $P(B)$
 $P(A \cap B)$

Conditional Probability Formula

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Probability that A occurs given that B has already occurred

Bayes Probabilitas

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$P(A)$ = The probability of A occurring

$P(B)$ = The probability of B occurring

$P(A|B)$ = The probability of A given B

$P(B|A)$ = The probability of B given A

Contoh Bayes Probabilitas

Skenario

Kita belum tahu koinnya fair atau bias. Bandingkan 2 hipotesis:

- H_f : koin fair $\rightarrow p(\text{Kepala}) = 0.5$
- H_b : koin bias $\rightarrow p(\text{Kepala}) = 0.7$

Ambil prior sama besar: $P(H_f) = P(H_b) = 0.5$.

Lakukan 5 lemparan dan amati data D : tepat 3 kepala.



Hitung (Langkah Bayes)

1. Likelihood (Binomial):

$$P(D | H_f) = \binom{5}{3} (0.5)^5 = \frac{10}{32} = 0.3125$$

$$P(D | H_b) = \binom{5}{3} (0.7)^3 (0.3)^2 = 10 \times 0.343 \times 0.09 = 0.3087$$

2. Evidence:

$$P(D) = P(D | H_f)P(H_f) + P(D | H_b)P(H_b) = 0.3125 \times 0.5 + 0.3087 \times 0.5 = 0.3106$$

3. Posterior:

$$P(H_f | D) = \frac{0.3125 \times 0.5}{0.3106} \approx \mathbf{0.503}, \quad P(H_b | D) \approx \mathbf{0.497}.$$

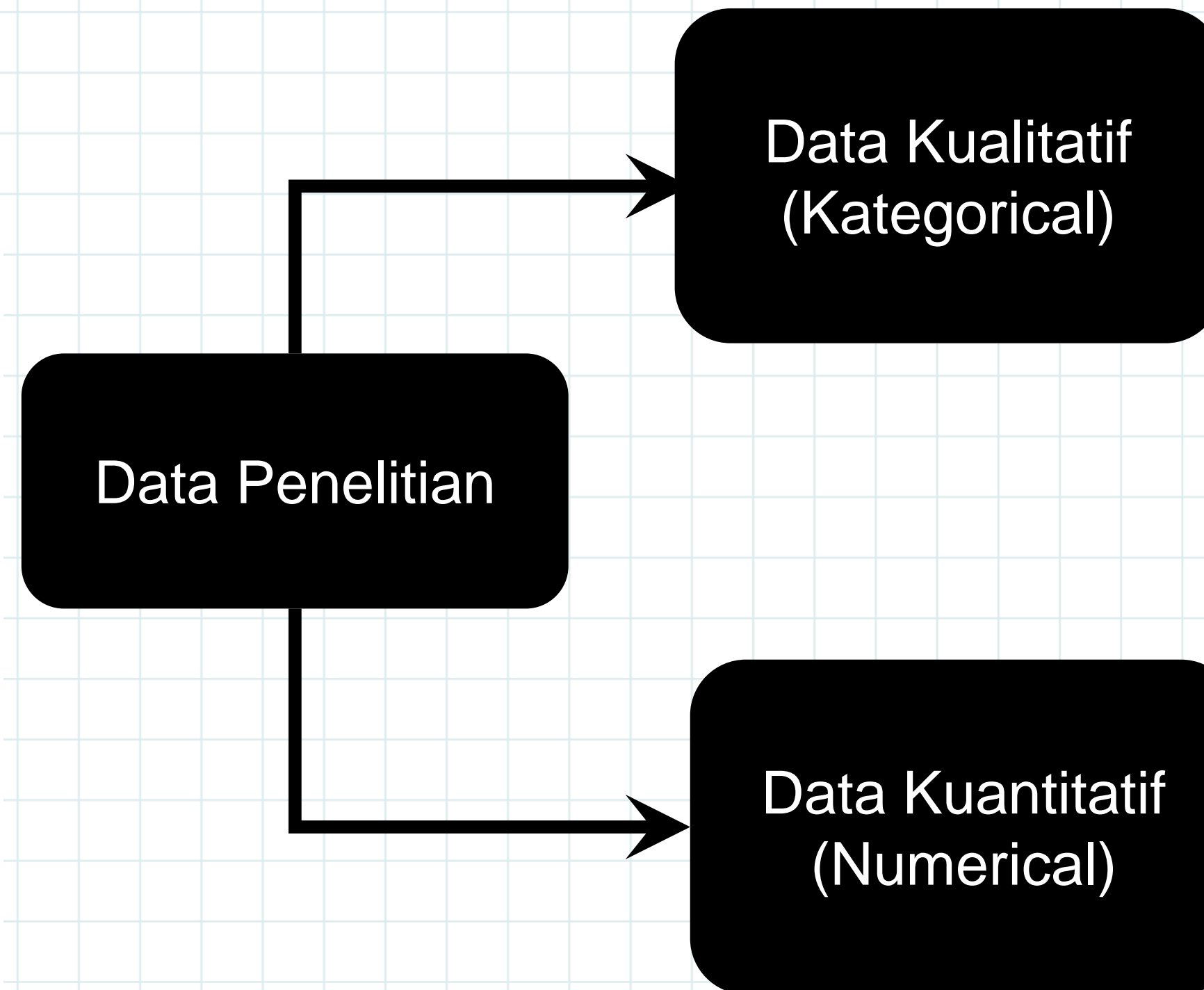
Makna: Dengan 3 kepala dari 5, bukti masih nyarisimbang—sedikit lebih mendukung “koin fair”.

(Opsional) Prediksi lemparan berikutnya

Bayesian model averaging:

$$P(\text{Kepala selanjutnya} | D) = 0.503 \times 0.5 + 0.497 \times 0.7 \approx \mathbf{0.599}.$$

Data Penelitian

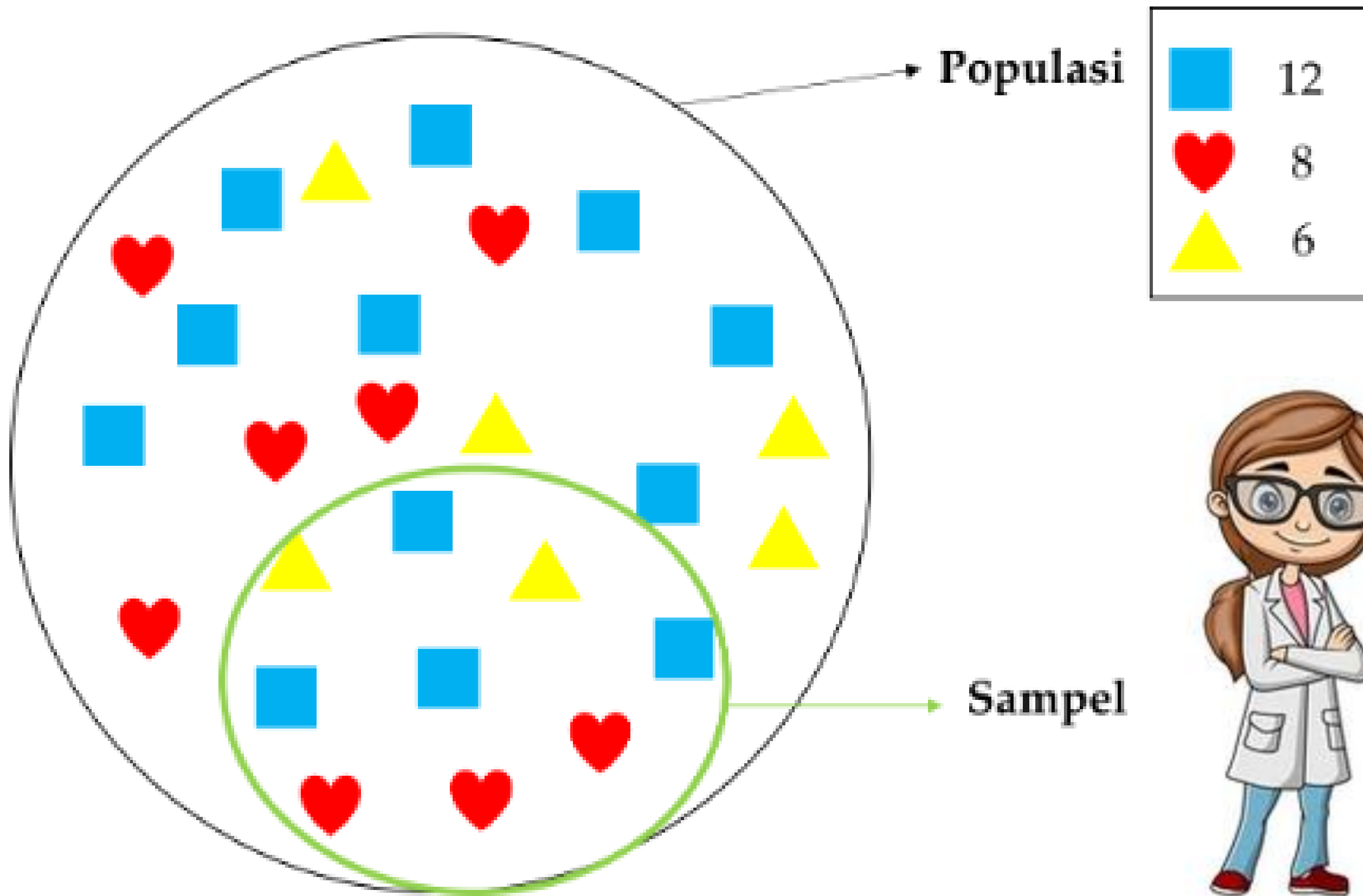


- **Nominal** (tanpa urutan; contoh: warna, jenis kelamin)
- **Ordinal** (ada urutan, jarak tak terdefinisi; contoh: tingkat kepuasan 1–5)

- **Diskrit** (bilangan cacah; contoh: jumlah cacat per batch)
- **Kontinu** (riil; contoh: berat, waktu)

Data adalah **output** atau **hasil** yang dikumpulkan dalam sebuah **eksperimen** untuk menguji hipotesis dan mencari hubungan sebab-akibat

Sampel Data



Perbandingan Populasi

4 : 2,67 : 2

Model 1 . kotak : hati : segitiga

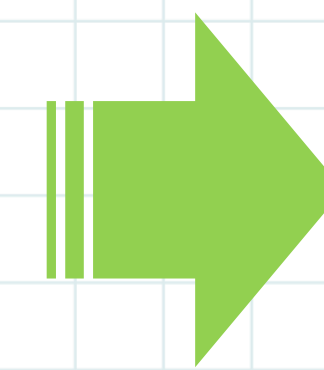
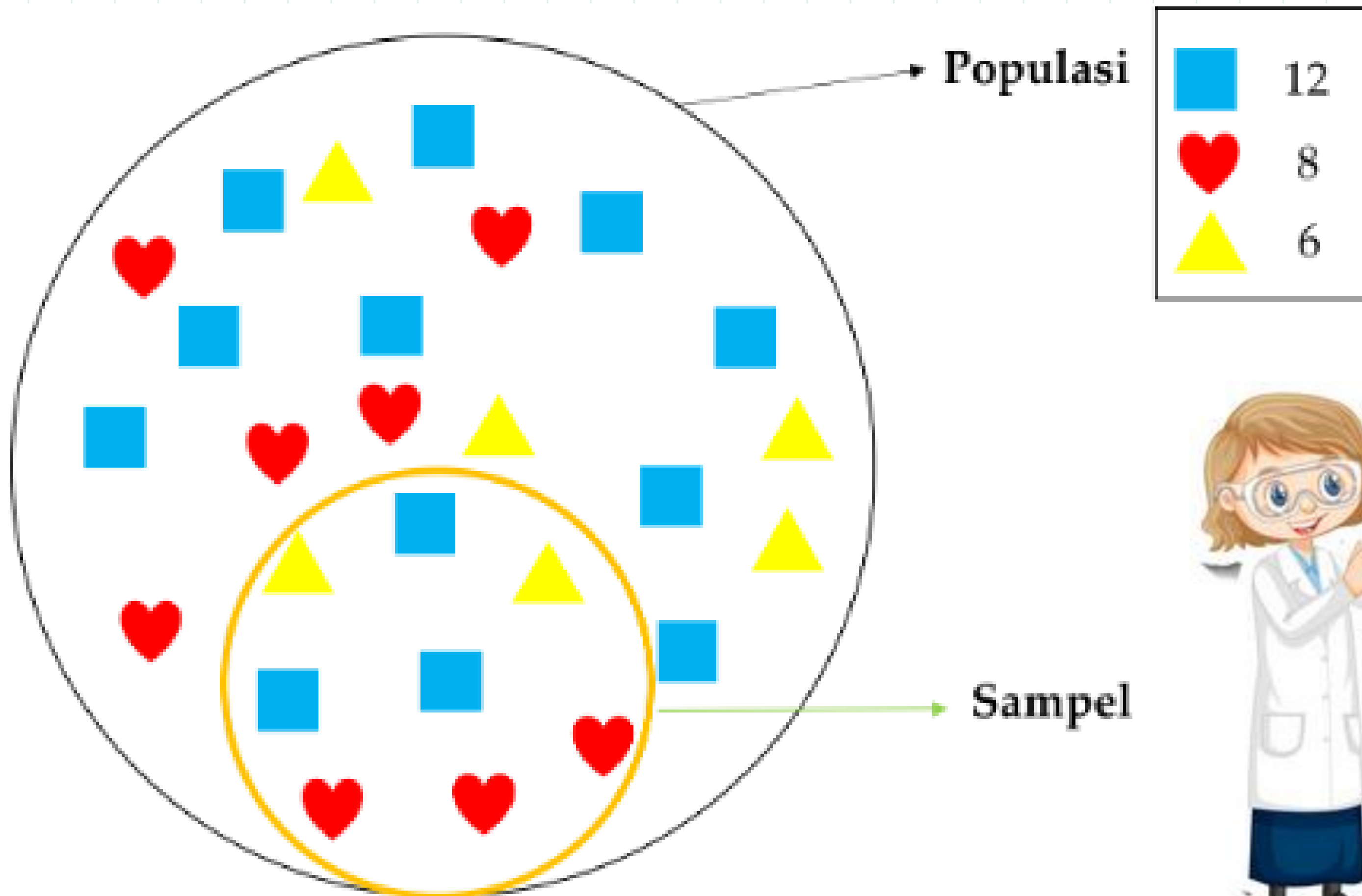
4 : 3 : 2

Analisis Sampel:

Model 1: Memiliki perbandingan yang hampir menyerupai kenyataan seluruh Populasi, dikatakan **kondisi ideal**



Sampel Data



Perbandingan Populasi

4 : 2,67 : 2



Model 2 . kotak : hati : segitiga

3 : 3 : 2

Analisis Sampel:

Model 2: Memiliki perbandingan yang melenceng dari populasi



Sampel Data

1. Pemilihan sampel data (disebut data training) adalah sangat penting, apabila **data training tidak merepresentasikan populasi**, maka model yang dihasilkan dari pembelajaran (training) tidak bagus.
2. Selain **data training**, perlunya **data validation** dan **data testing**
3. Data scientist berusaha mencari tahu populasi dengan cara menyelidiki fitur (**features** atau sifat-sifat) yang dimiliki sampel



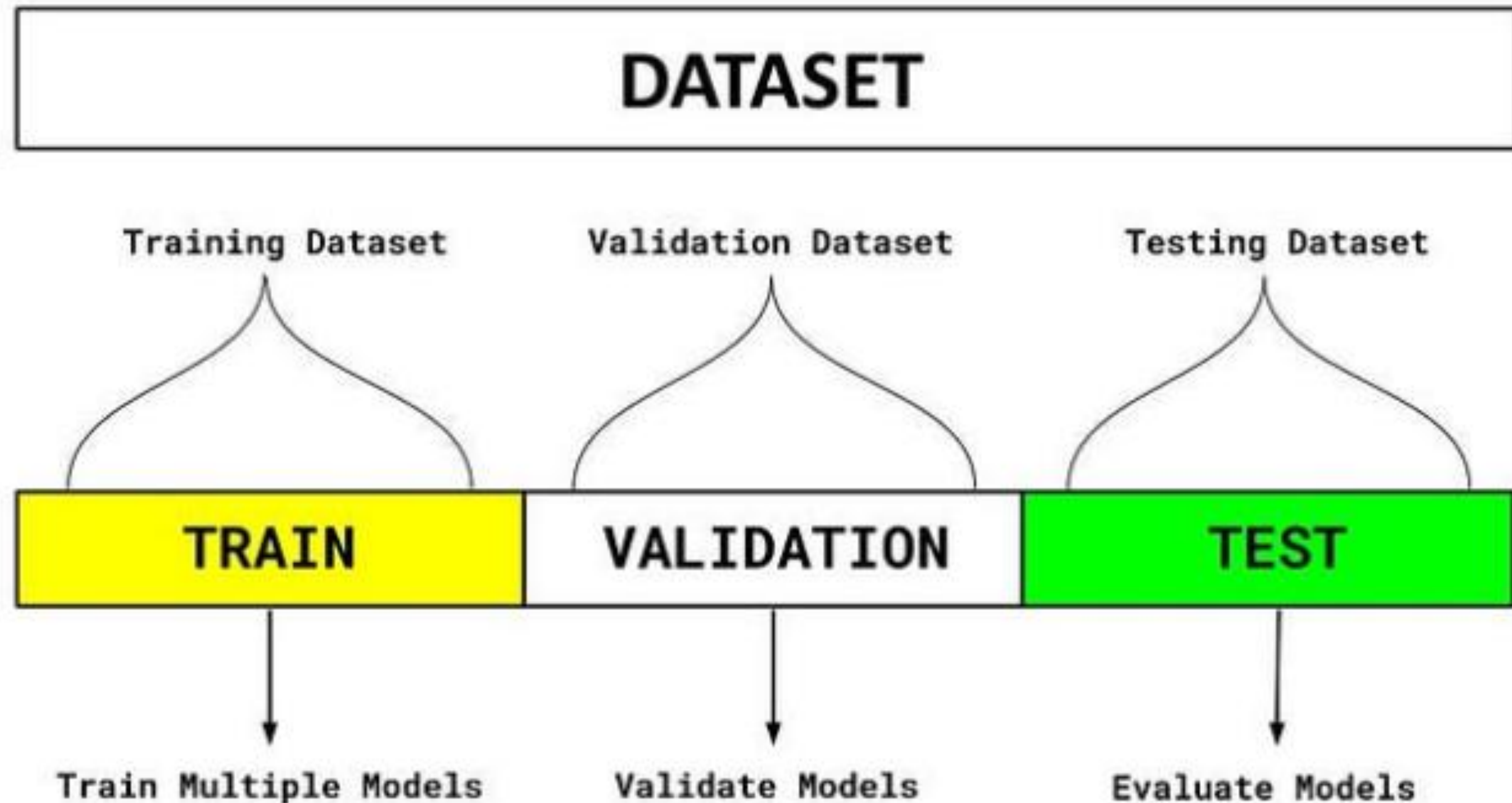
Data Training, Validation dan Testing



1. **Training Set:** himpunan data yang digunakan untuk melatih atau membangun model.
2. **Validation Set:** himpunan data yang digunakan untuk mengoptimisasi saat melatih model. Model dilatih menggunakan **training set** dan pada umumnya **kinerja** saat latihan **diuji** dengan **validation set**. Hal ini berguna untuk generalisasi (agar model mampu mengenali pola secara generik).
3. **Testing Set:** himpunan data yang digunakan untuk menguji model setelah proses latihan selesai. Testing set adalah **unseen data**. Artinya, model dan manusia tidak boleh melihat sampel ini saat proses Latihan (Testing set tidak diikutkan dalam proses training model).

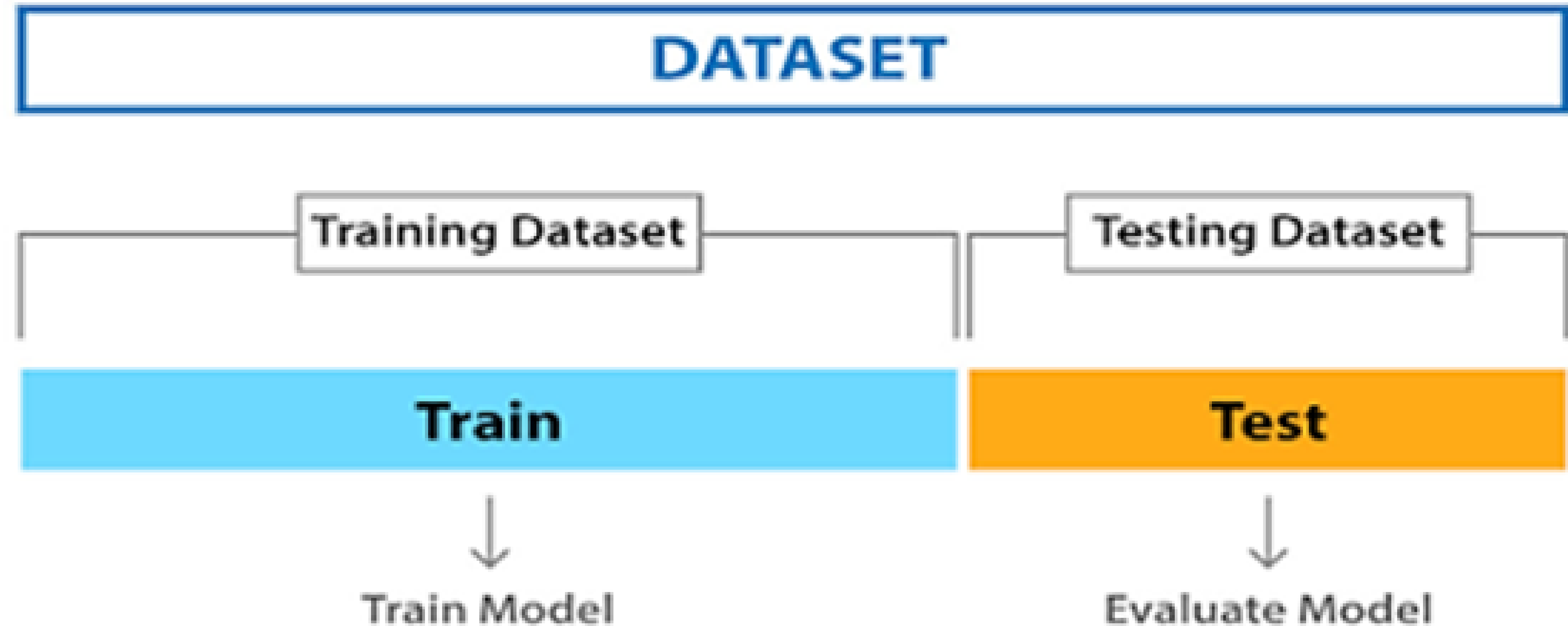
Rasio Pembagian Dataset

1. Pada umumnya, rasio pembagian dataset (**training: validation: testing**) adalah (80% : 10% : 10%) atau (90% : 5% : 5%).



Ratio Pembagian Dataset

2. **Validation Set** bisa tidak digunakan jika ukuran data kecil, sehingga dataset hanya dibagi dua : **data training** dan **data testing**, dengan rasio pembagian data: (90% : 10%), (80% : 20%), (70% : 30%), atau (50% : 50%).



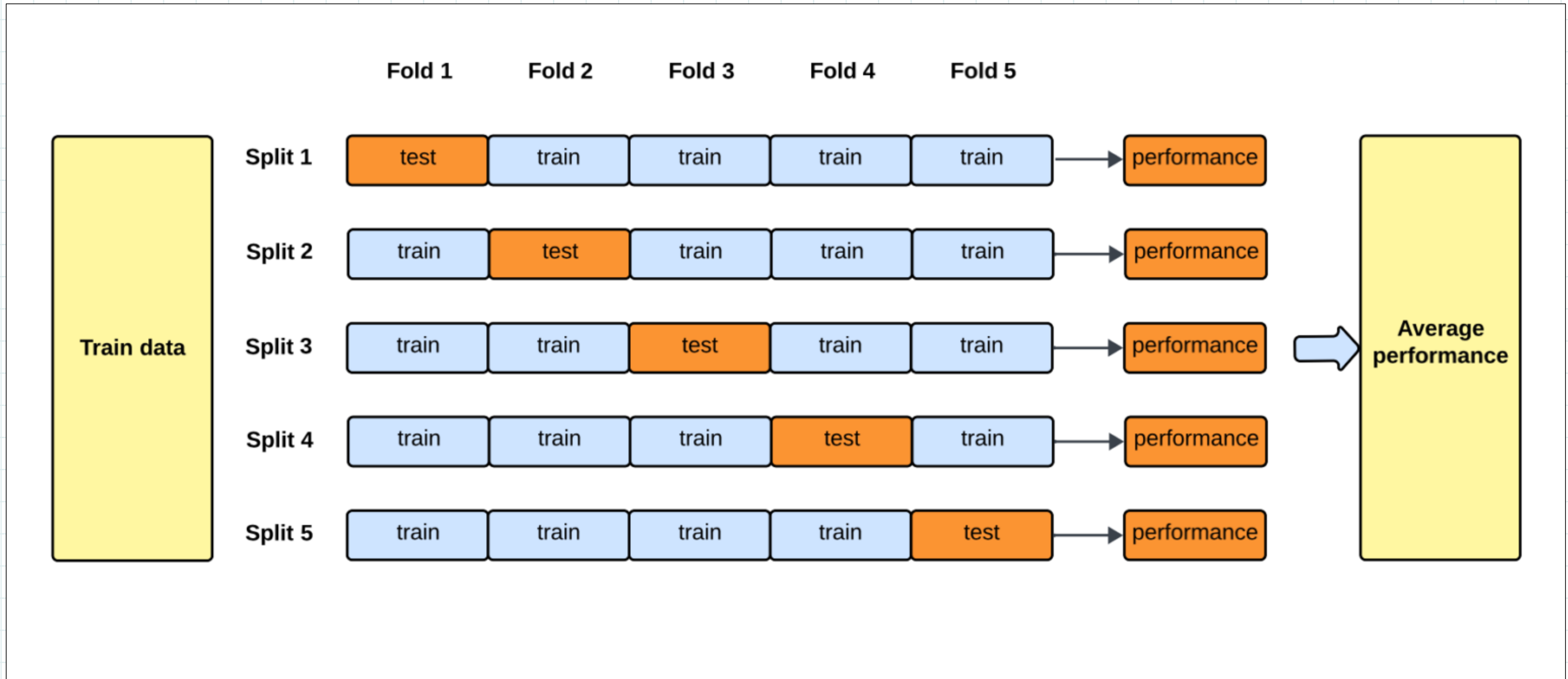
K – Cross Validation



1. Saat tidak menggunakan validation set, opsi evaluasi dengan **K-cross validation** menjadi pilihan.
2. Artinya, membagi training dataset (atau keseluruhan dataset) menjadi K bagian yang sama. menggunakan K – 1 bagian untuk training, kemudian menguji kinerja model saat latihan (validation) menggunakan satu bagian
3. Hal ini diulangi sebanyak K kali dimana sebuah bagian data digunakan sebagai testing set sebanyak sekali (bergilir).
4. Semua data akan merasakan menjadi data training dan data testing

K – Cross Validation

- K – Cross Validation, $K = 5$



Data Deskriptif

- **Motivasi:**
 - Lebih memahami data: kecenderungan sentral data, variasi dan sebaran
- **Karakteristik Sebaran Data:**
 - Mean, Median, Modus,
 - Max, Min, Quantiles, Outliers, Variance
- **Sebaran Dimensi Data Numerik**
 - Analisis sebaran data dengan rinci dan presisi
 - Analisis BoxPlot atau Quantile pada interval terurut

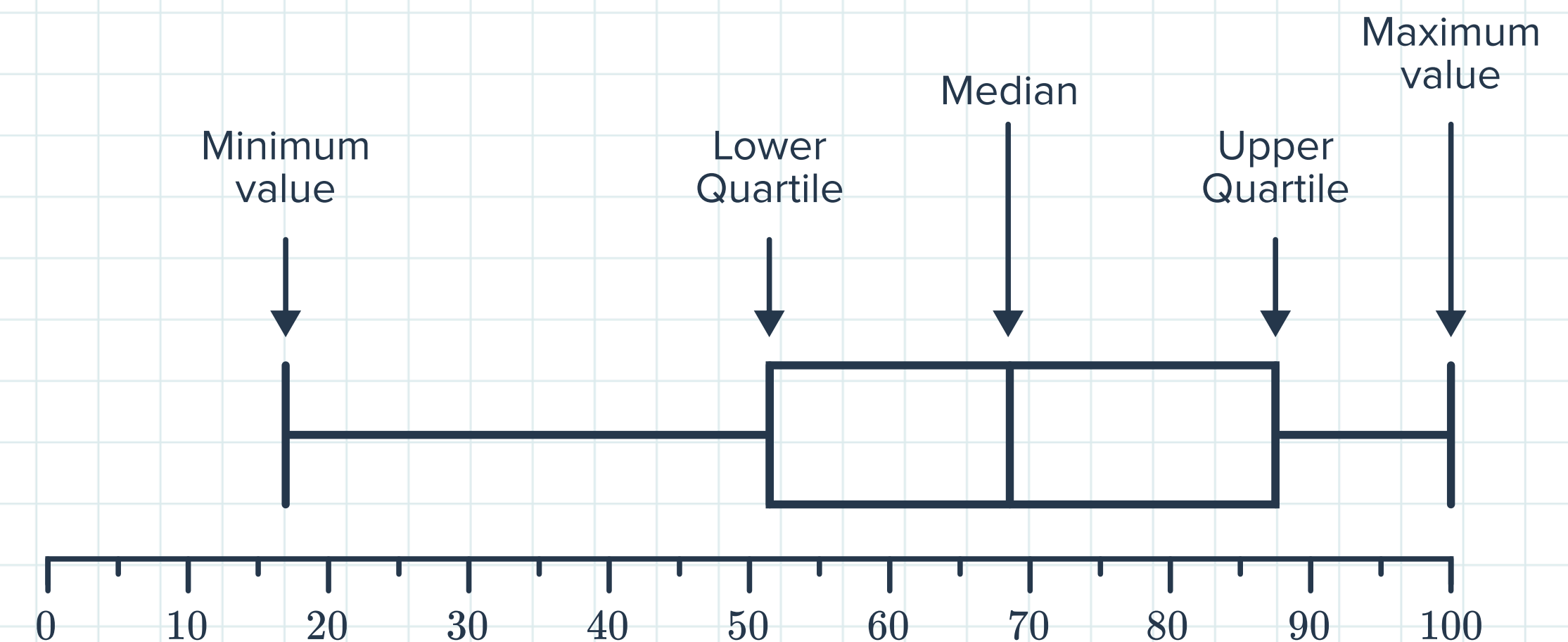
Statistics Formula

$$\text{Mean } \bar{x} = \frac{\sum x_i}{N}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{N}$$

$$\text{Median} = \begin{cases} \frac{(N+1)^{\text{th}}}{2} \text{ term; when } N \text{ is odd} \\ \frac{\frac{N}{2}^{\text{th}} \text{ term} + (\frac{N}{2} + 1)^{\text{th}} \text{ term}}{2}; \text{ when } N \text{ is even} \end{cases}$$

Mode = The value in the data set that occurs most frequently

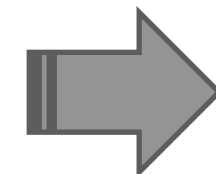


Pandas Dataframe : mean, median, modus

```
1 import pandas as pd
2 import numpy as np
3 # Importing and Exporting Data
4 df = pd.read_csv('500_Person_Gender_Height_Weight_Index.csv')
5 df
```

PRAKTIKUM

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
...
495	Female	150	153	5



```
1 df['Height'].mean()
```

169.944

```
1 df["Height"].median()
```

170.5

```
1 df["Height"].mode()
```

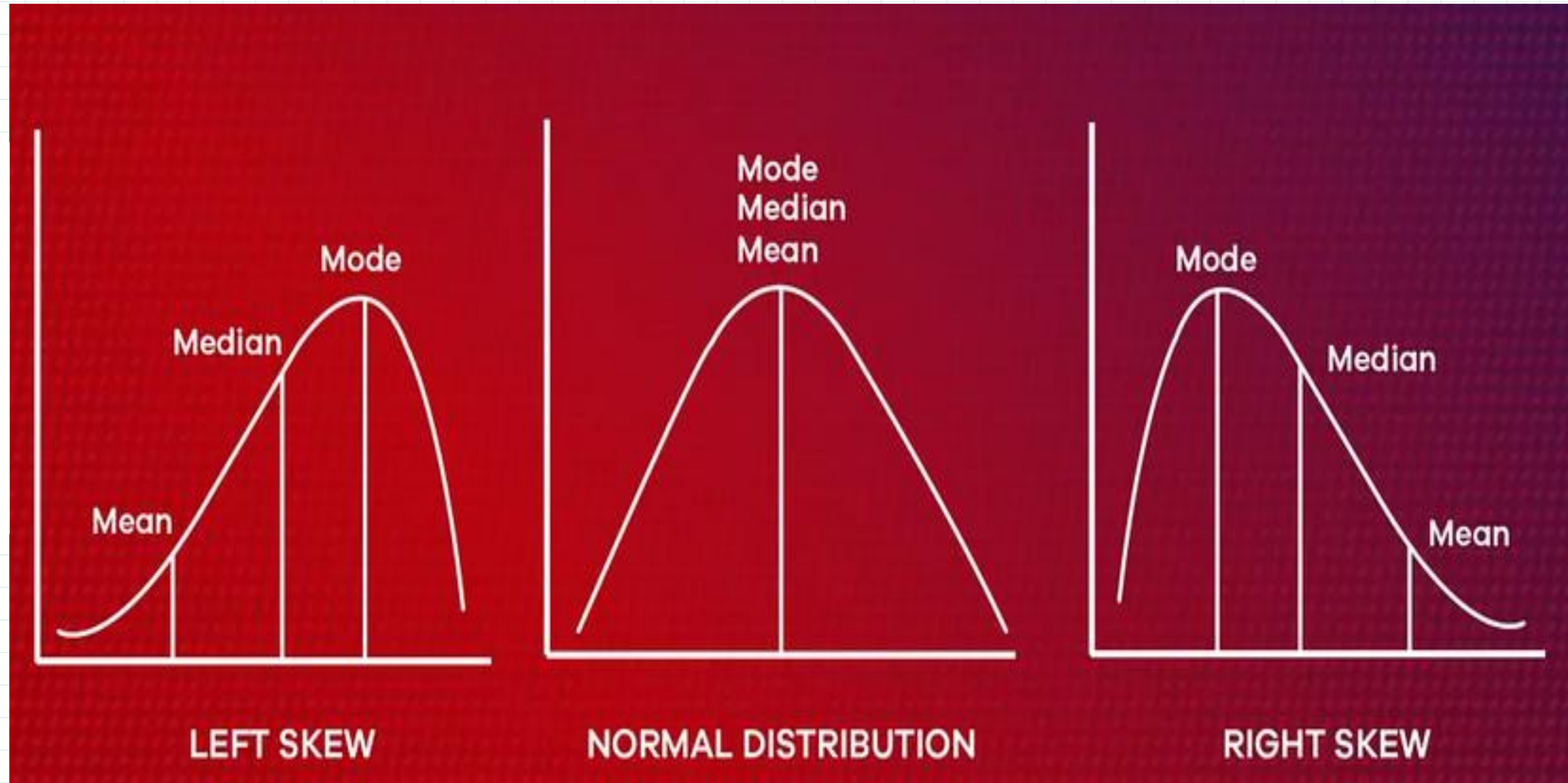
0 188

Kecenderungan Sentral Data

`df.mean()`

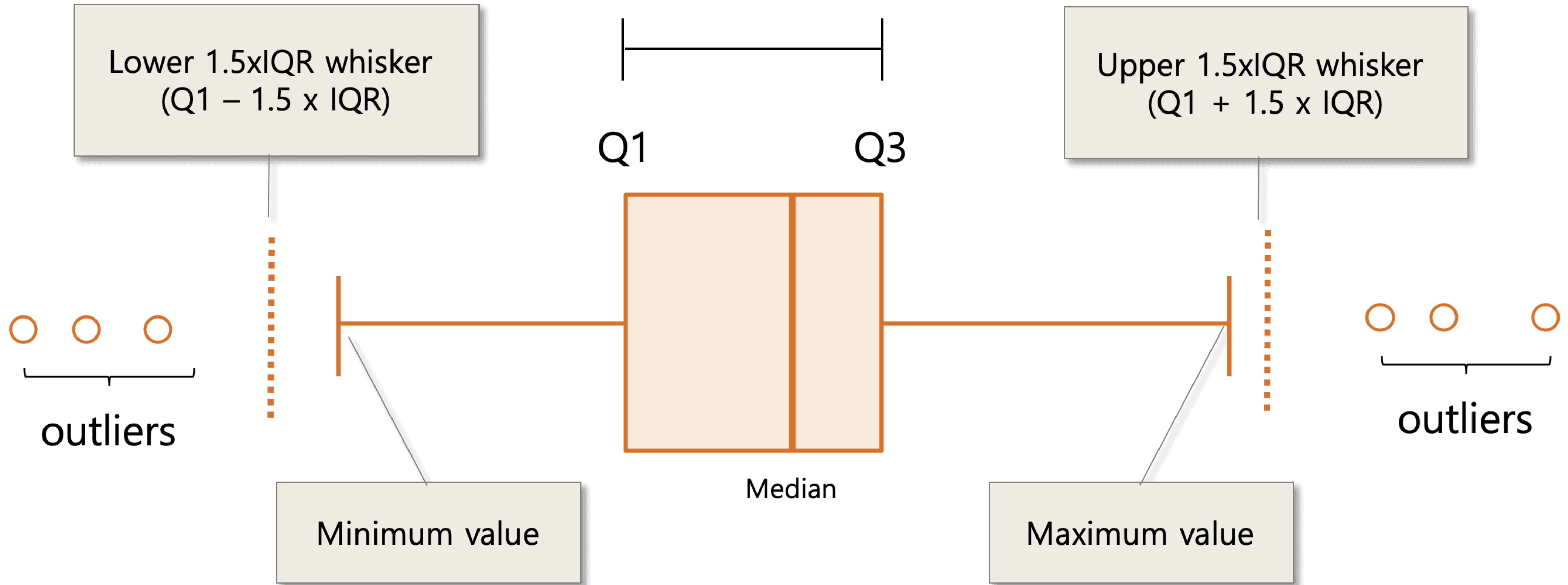
`df.median()`

`df.mode()`



Analisis Boxplot

Interquartile range (IQR)

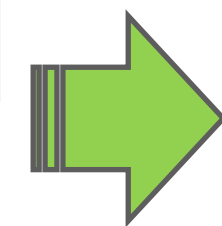


Pandas Dataframe : Quartiles

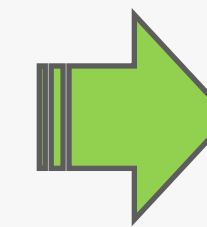
```
1 import pandas as pd
2 import numpy as np
3 # Importing and Exporting Data
4 df = pd.read_csv('500_Person_Gender_Height_Weight_Index.csv')
5 df
```

PRAKTIKUM

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
...
495	Female	150	153	5



```
1 # Hitung kuartil pertama (Q1)
2 q1 = df['Height'].quantile(0.25)
3 print("Q1 : ", q1)
4 # Hitung kuartil ketiga (Q3)
5 q3 = df['Height'].quantile(0.75)
6 print("Q3 : ", q3)
7 iqr = q3 - q1
8 print("IQR : ", iqr)
```



Q1 : 156.0
Q3 : 184.0
IQR : 28.0

```
1 df.describe()
```

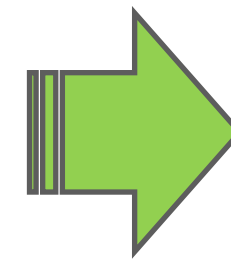
	Height	Weight	Index
count	500.000000	500.000000	500.000000
mean	169.944000	106.000000	3.748000
std	16.375261	32.382607	1.355053
min	140.000000	50.000000	0.000000
25%	156.000000	80.000000	3.000000
50%	170.500000	106.000000	4.000000
75%	184.000000	136.000000	5.000000
max	199.000000	160.000000	5.000000

Pandas Dataframe : Boxplot

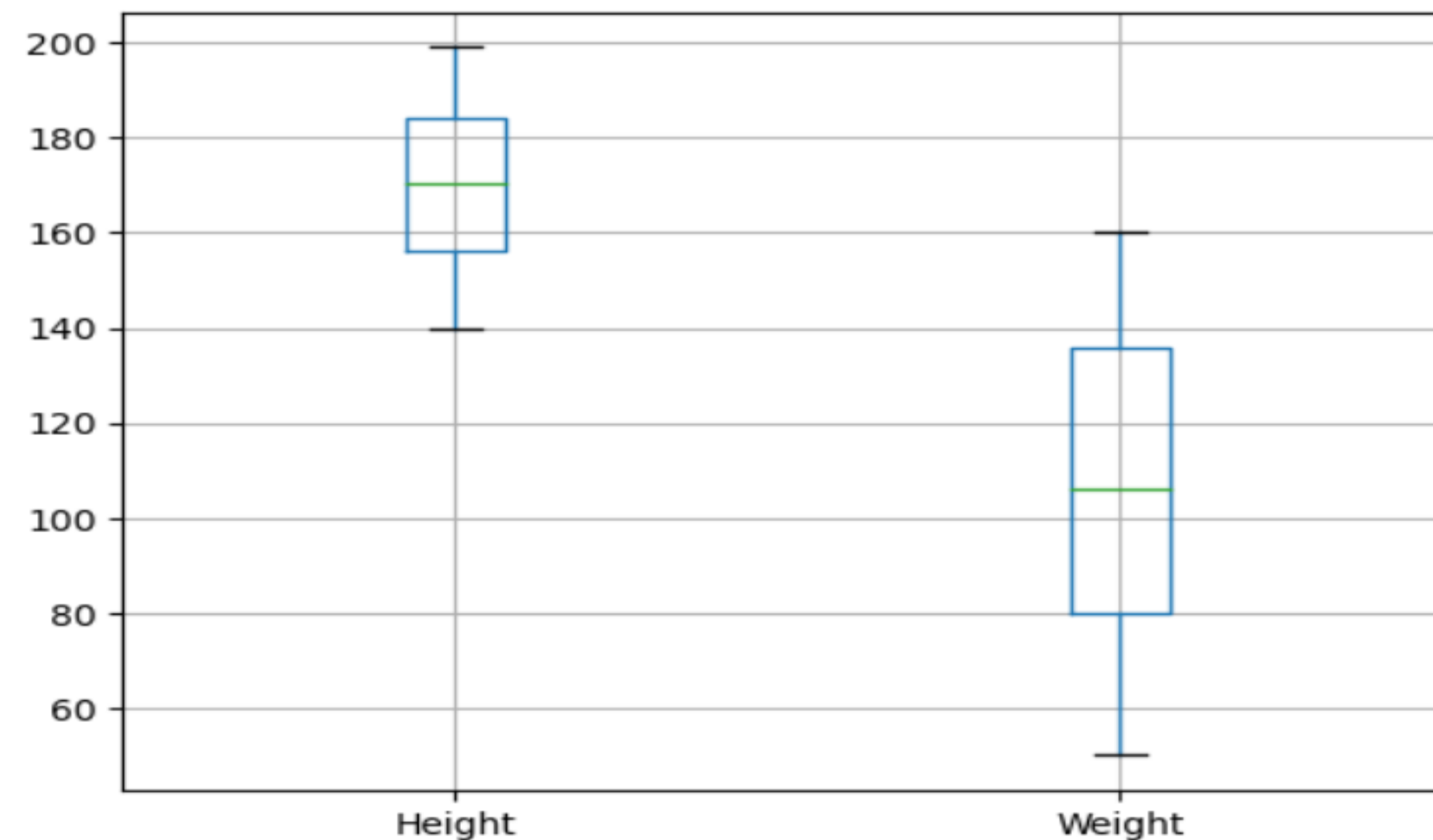
```
1 import pandas as pd
2 import numpy as np
3 # Importing and Exporting Data
4 df = pd.read_csv('500_Person_Gender_Height_Weight_Index.csv')
5 df
```

PRAKTIKUM

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
...
495	Female	150	153	5



```
1 df.boxplot(column=['Height', 'Weight'])
```

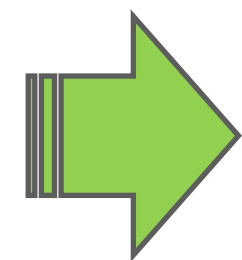


Pandas Dataframe : Outlier

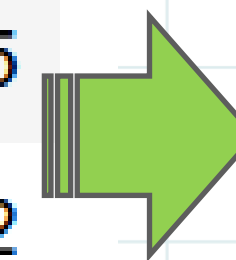
```
1 import pandas as pd
2 import numpy as np
3 # Importing and Exporting Data
4 df = pd.read_csv('500_Person_Gender_Height_Weight_Index.csv')
5 df
```

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
...
495	Female	150	153	5

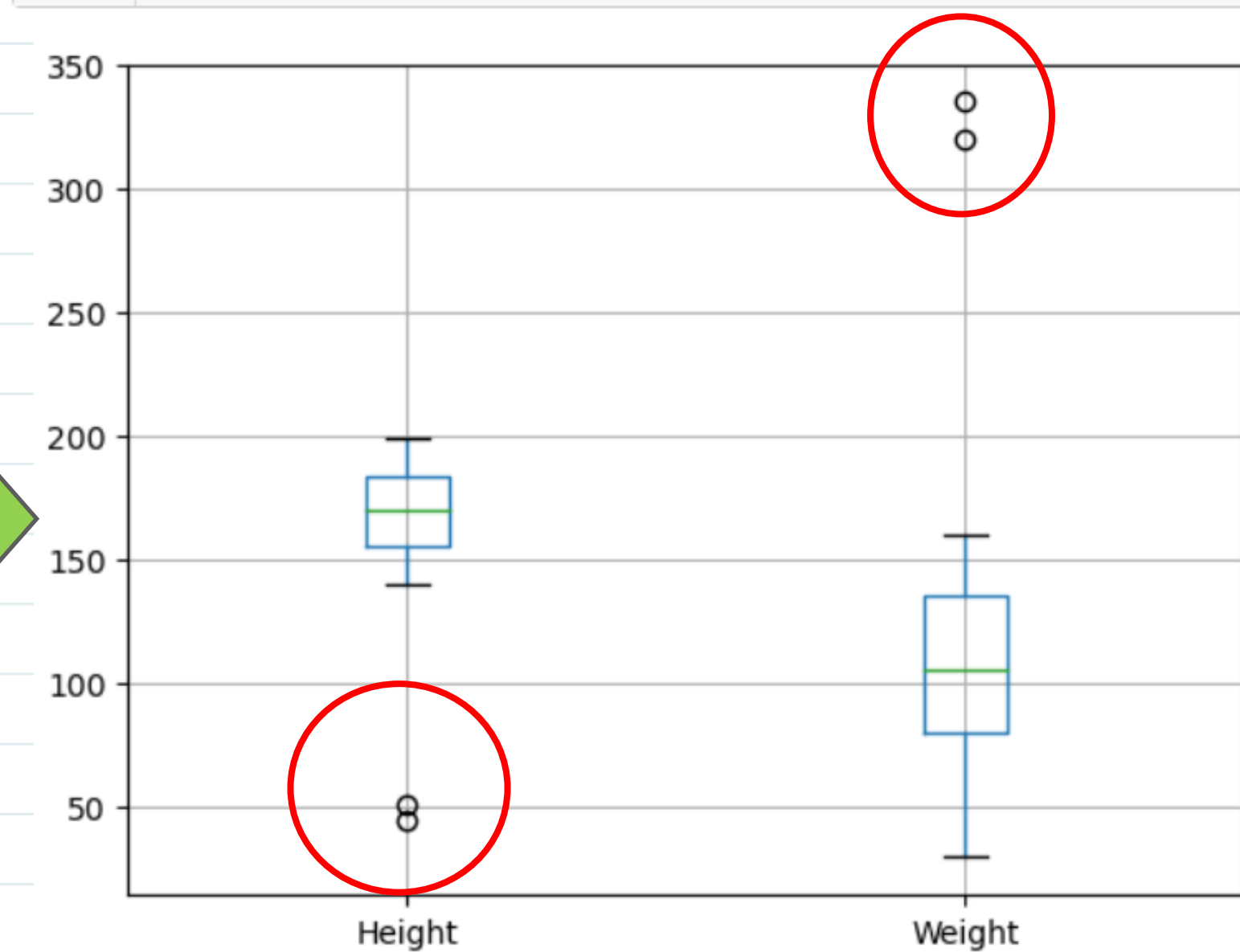
Misal: Tambahkan data outlier berikut ini



	Gender	Height	Weight	Index
501	Male	51	33	2
502	Female	155	320	5
503	Female	45	30	2
504	Male	160	335	5

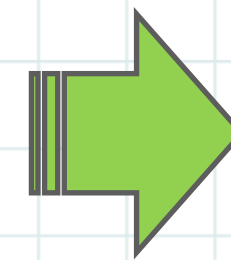


```
1 df.boxplot(column=['Height', 'Weight'])
```



Pandas Dataframe : Outlier

	Gender	Height	Weight	Index
501	Male	51	33	2
502	Female	155	320	5
503	Female	45	30	2
504	Male	160	335	5



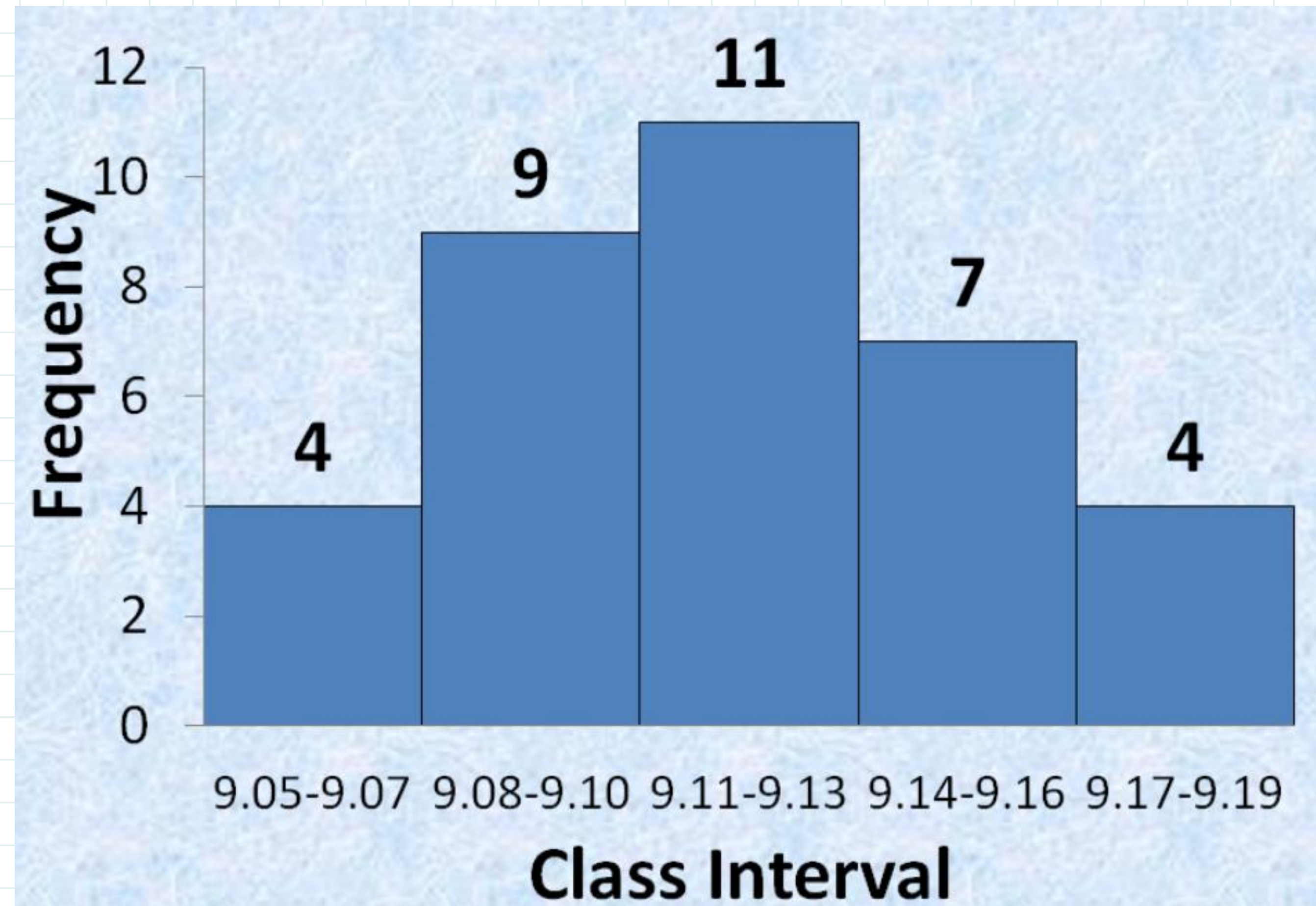
```
1 # Hitung kuartil pertama (Q1)
2 q1 = df['Height'].quantile(0.25)
3 # Hitung kuartil ketiga (Q3)
4 q3 = df['Height'].quantile(0.75)
5 iqr = q3 - q1
6 # Hitung batas bawah (Lower Bound) dan batas atas (Upper Bound) untuk outlier
7 lower_bound = q1 - 1.5 * iqr
8 upper_bound = q3 + 1.5 * iqr
9
10 # Temukan outlier dalam DataFrame
11 outliers = df[(df['Height'] < lower_bound) | (df['Height'] > upper_bound)]
12
13 print("Outlier Height:")
14 print(outliers)
```

Outlier Height:

	Gender	Height	Weight	Index
501	Male	51	33	2
503	Female	45	30	2

Analisis Histogram

- **Frequency Histogram**
 - Grafik yang menggambarkan analisis univariat
 - Terdiri dari sekumpulan persegi panjang yang mencerminkan hitungan atau frekuensi kelas yang ada dalam data yang diberikan



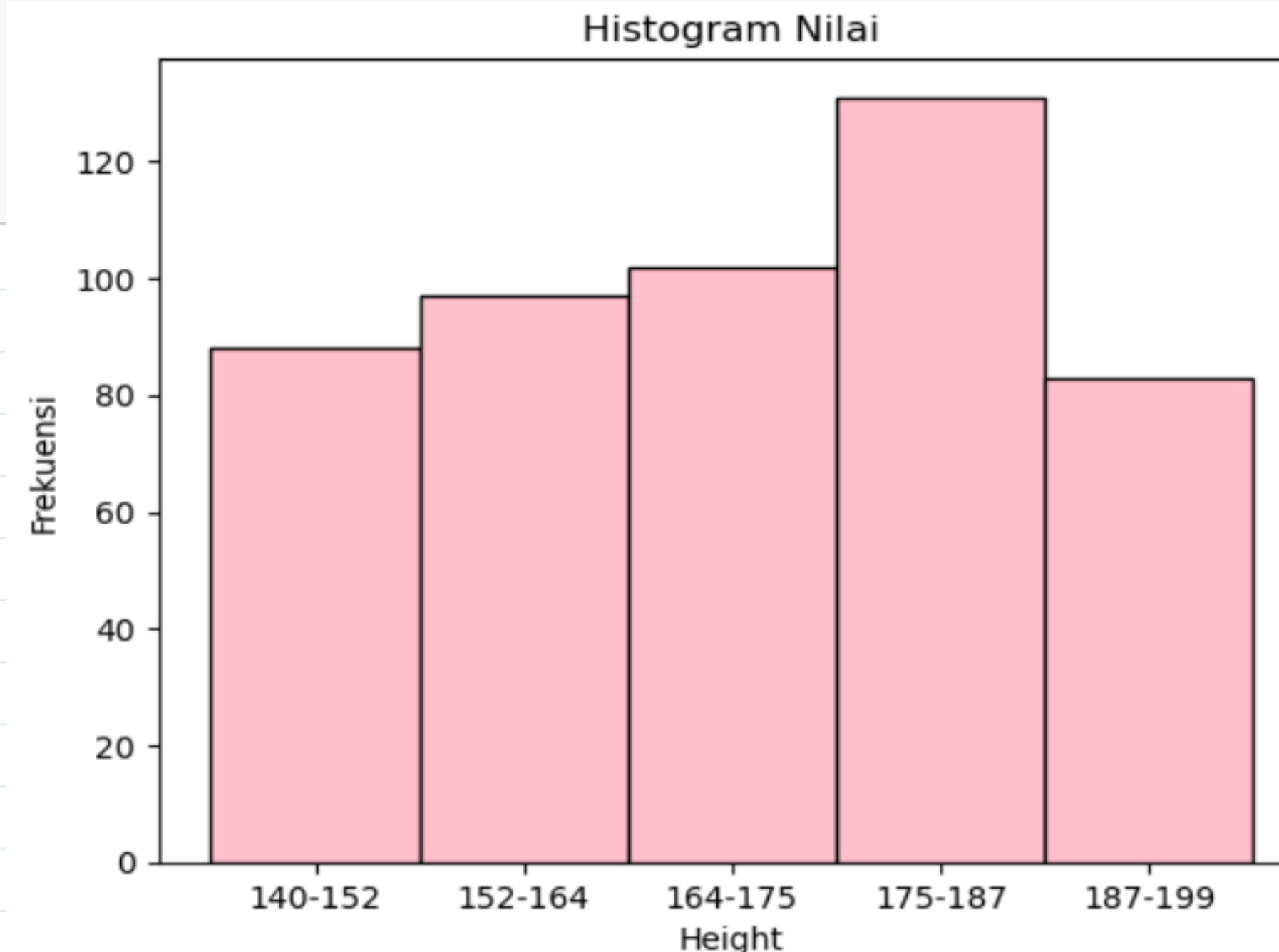
<https://techqualitypedia.com/histogram/>

Analisis Histogram

PRAKTIKUM

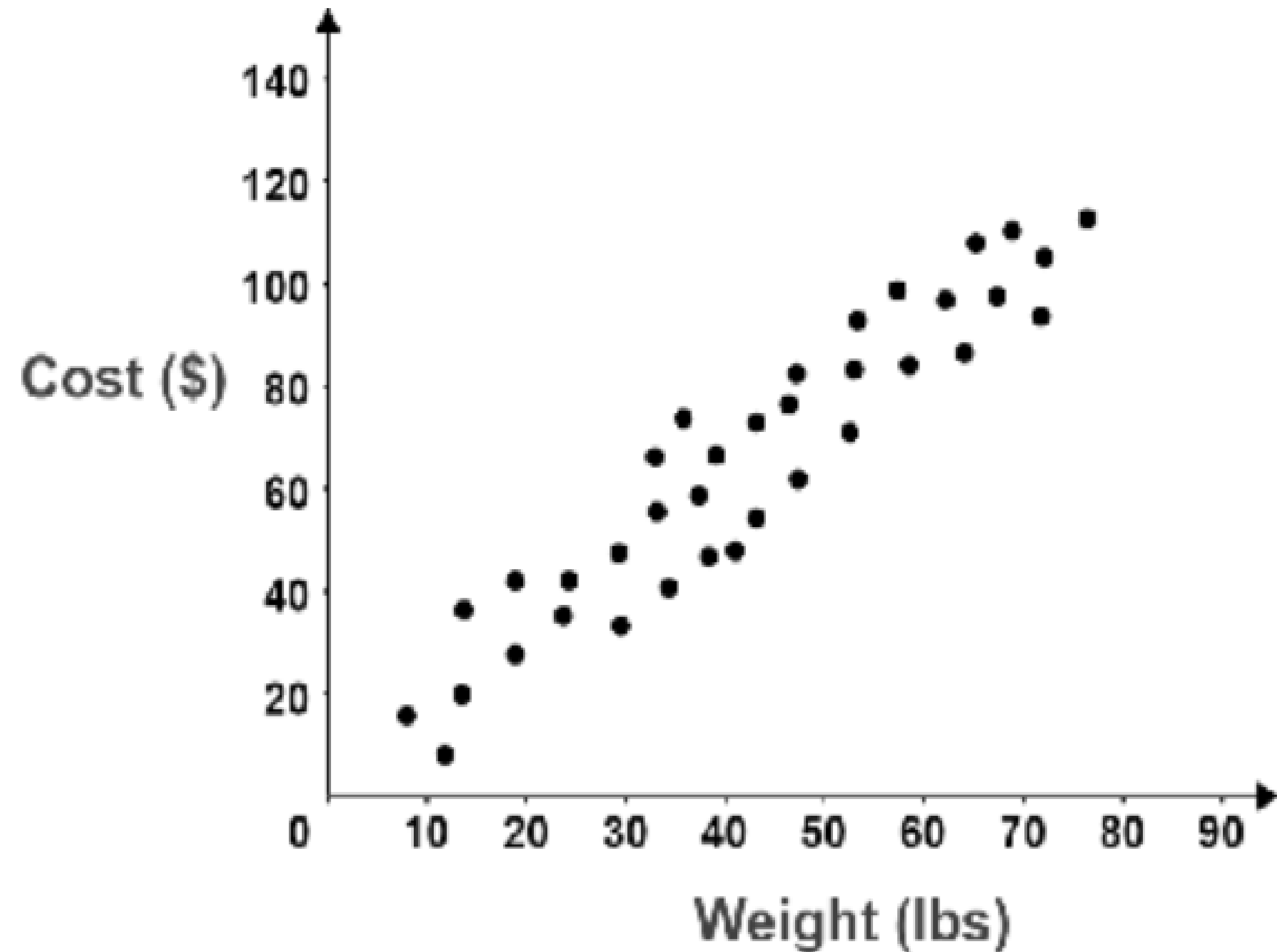


```
1 data_height = df["Height"]
2 # Buat histogram
3 n, bins, patches = plt.hist(data_height, bins=5, color='pink', edgecolor='black')
4
5 # Tambahkan Label
6 plt.title('Histogram Nilai')
7 plt.xlabel('Height')
8 plt.ylabel('Frekuensi')
9
10 # Tampilkan rentang frekuensi di sumbu x
11 bin_centers = 0.5 * (bins[:-1] + bins[1:])
12 plt.xticks(bin_centers, ['{:.0f}-{:.0f}'.format(bins[i], bins[i+1]) for i in range(len(bins)-1)])
13
14 # Tampilkan histogram
15 plt.show()
16
```



Scatter Plot

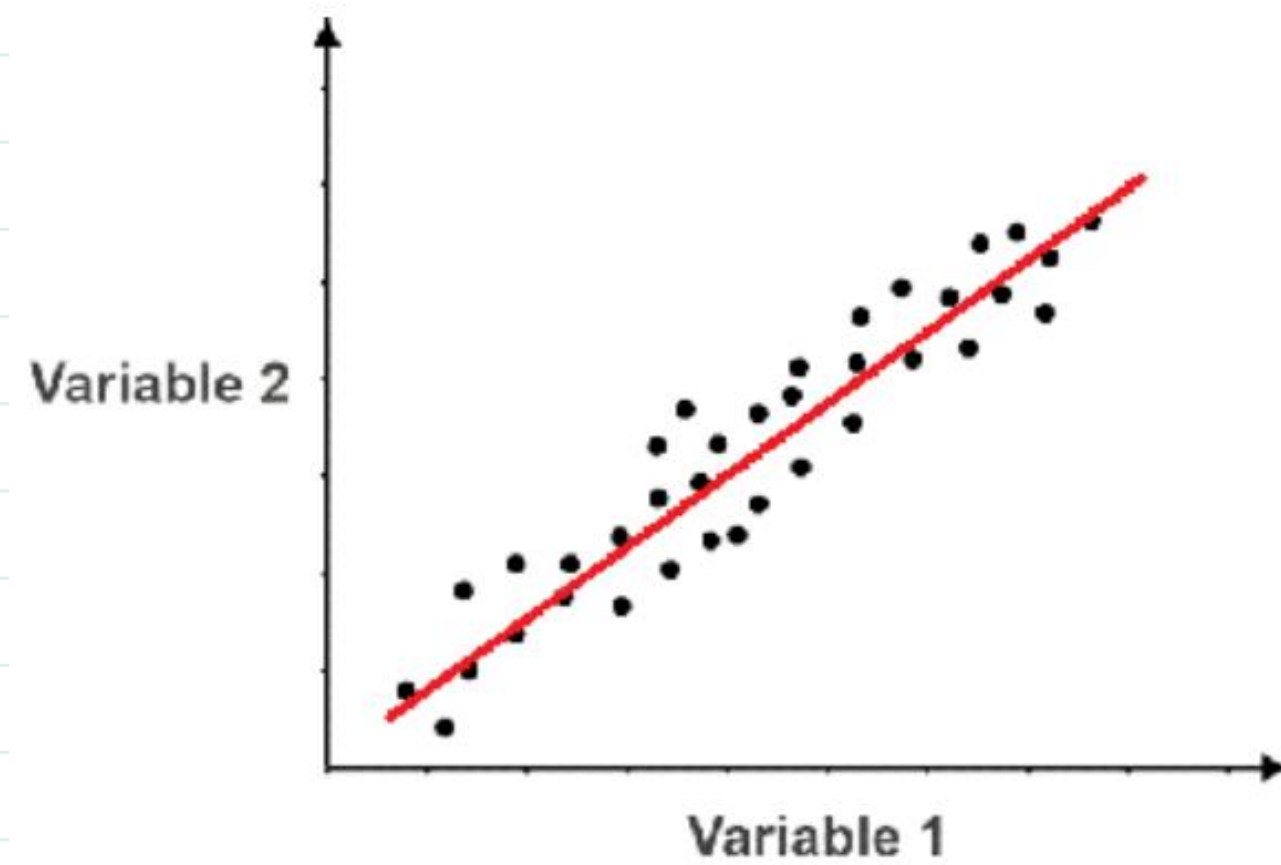
- Memberikan tampilan pertama pada data bivariat untuk melihat clusternya dari setiap titik dan outlier
- Setiap pasangan nilai diperlakukan sebagai pasangan koordinat dan diplot sebagai titik-titik pada bidang tersebut



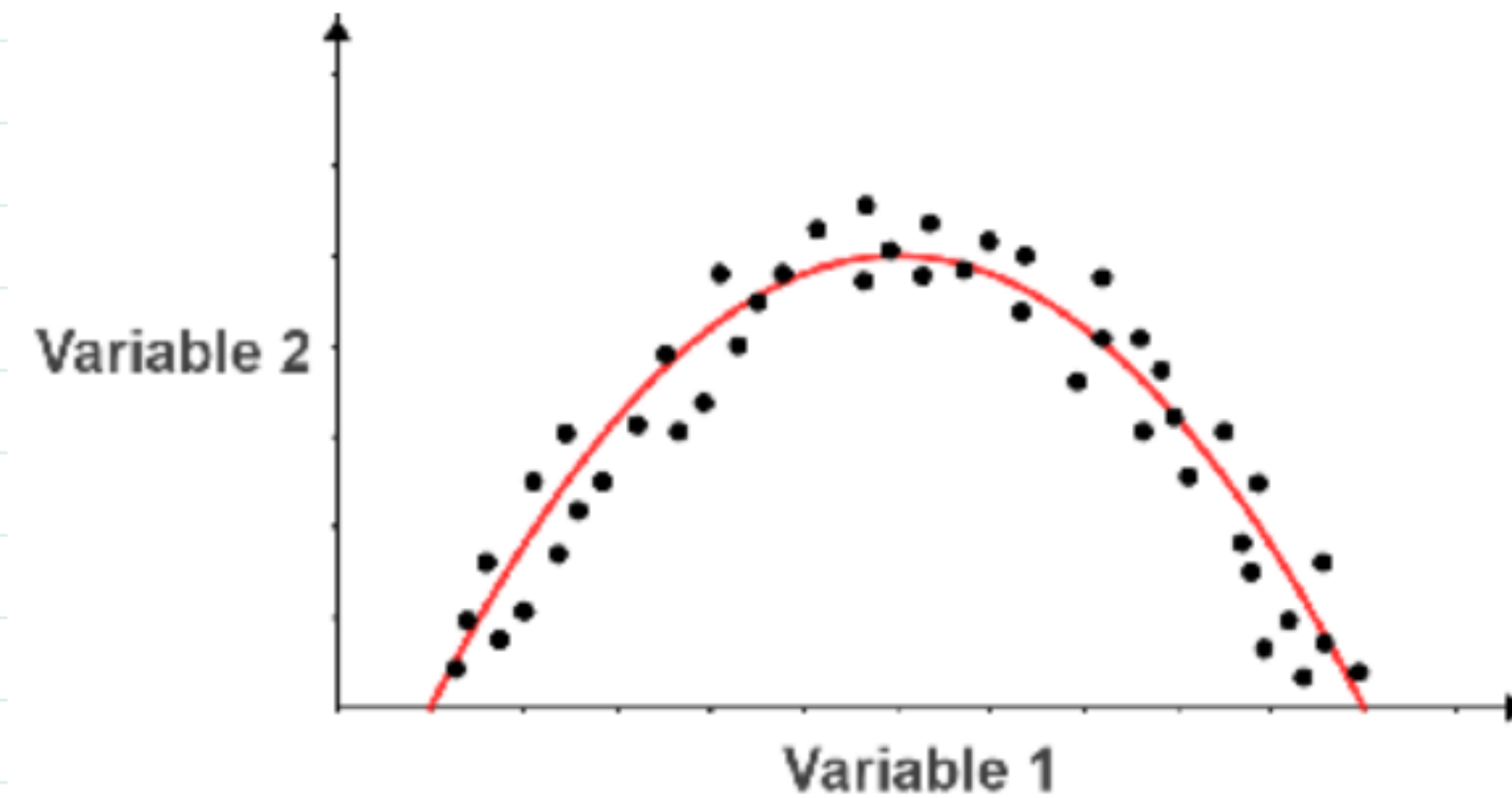
<https://www.math.net/scatter-plot>

Scatter Plot

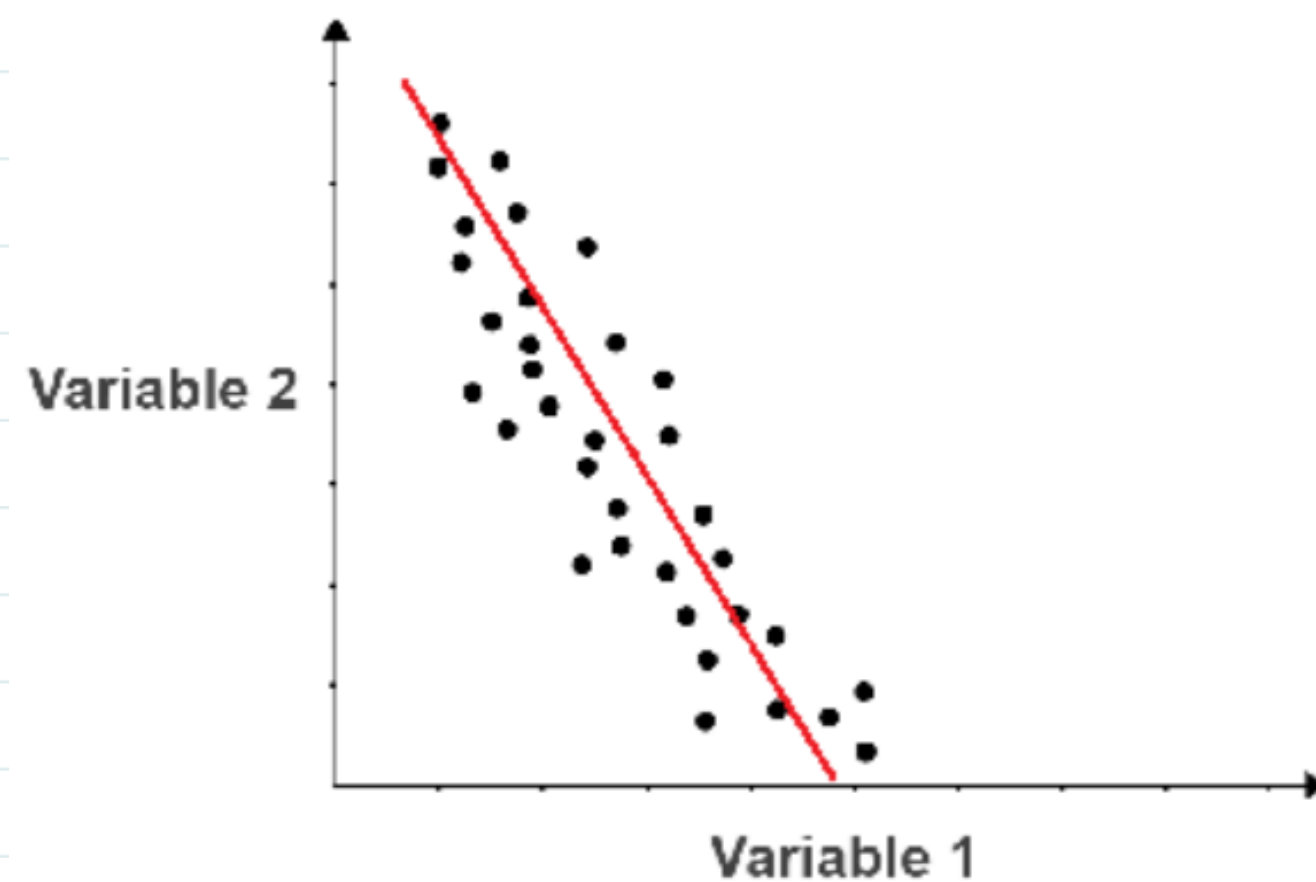
- Korelasi Positif



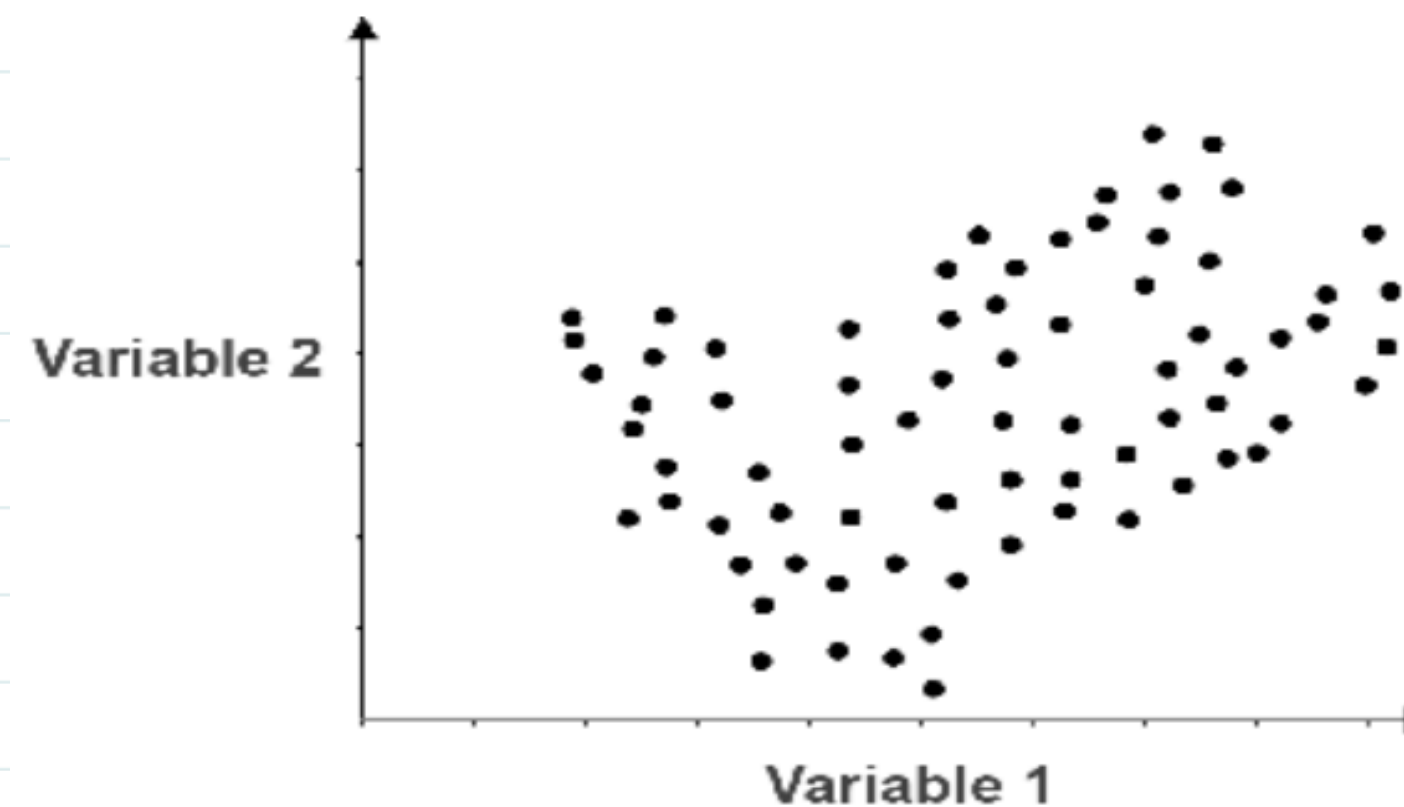
- Korelasi Non Linier



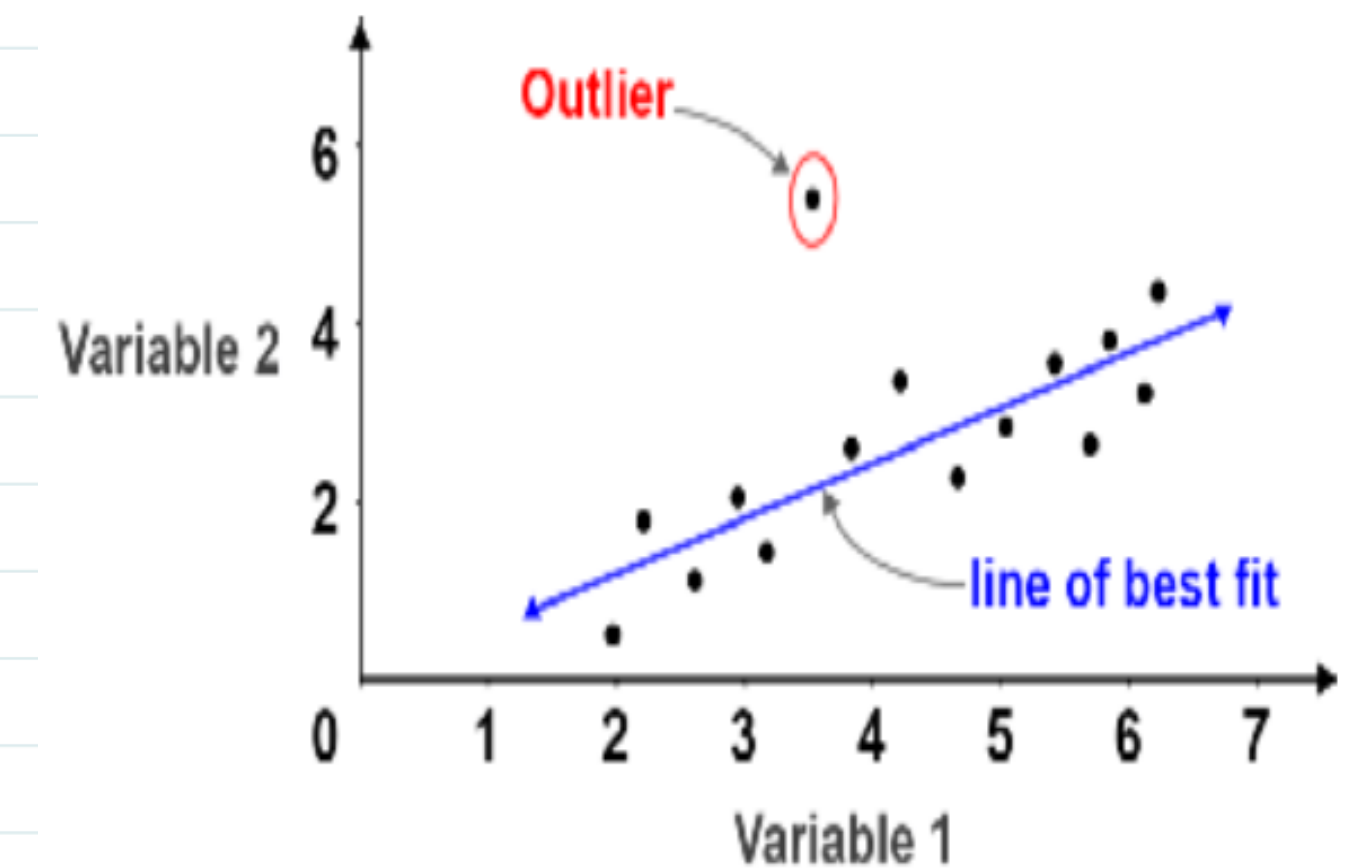
- Korelasi Negatif



- Tidak Ada Korelasi



- Outlier

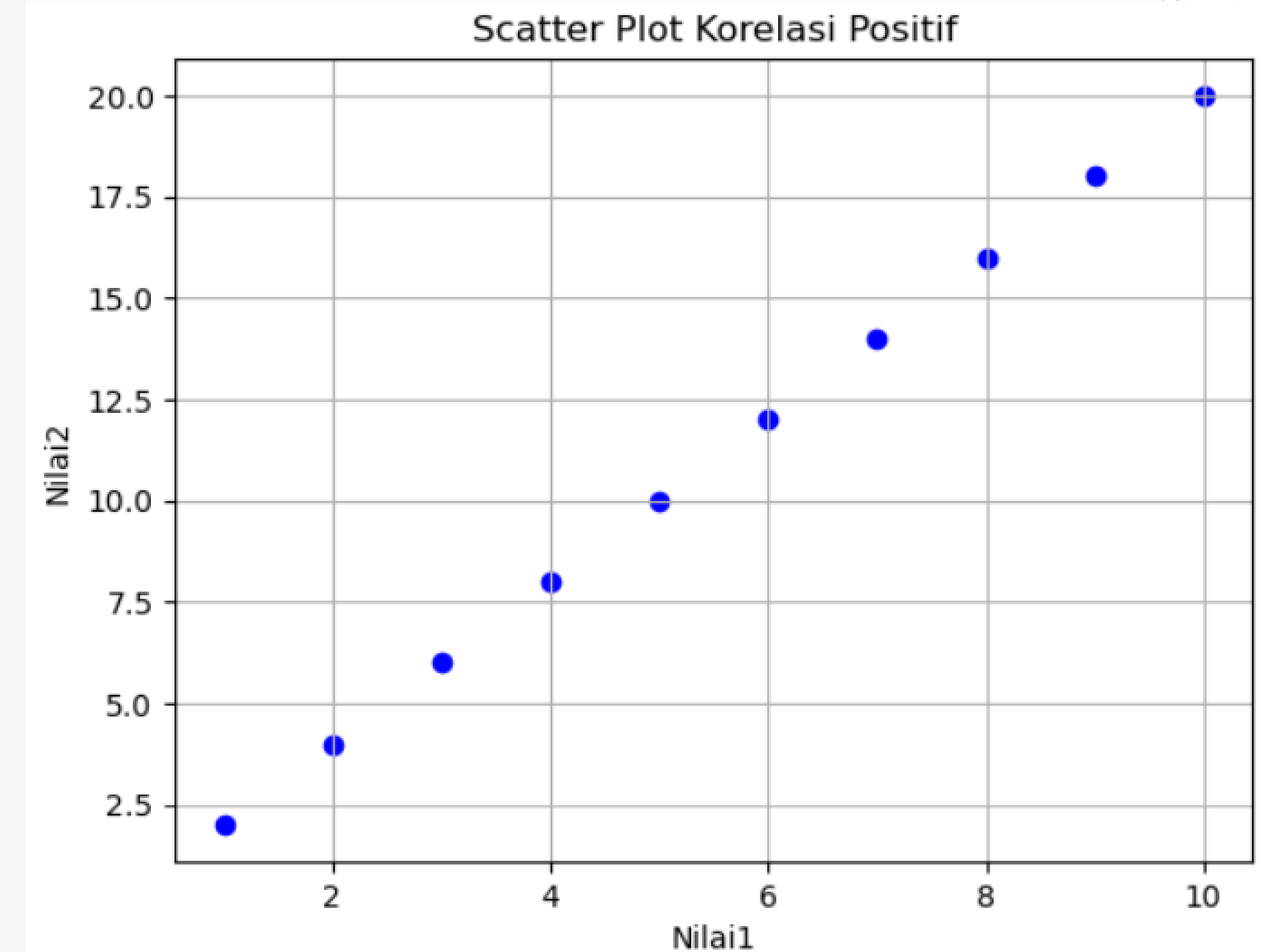


<https://www.math.net/scatter-plot>

Scatter Plot :: Korelasi Positif

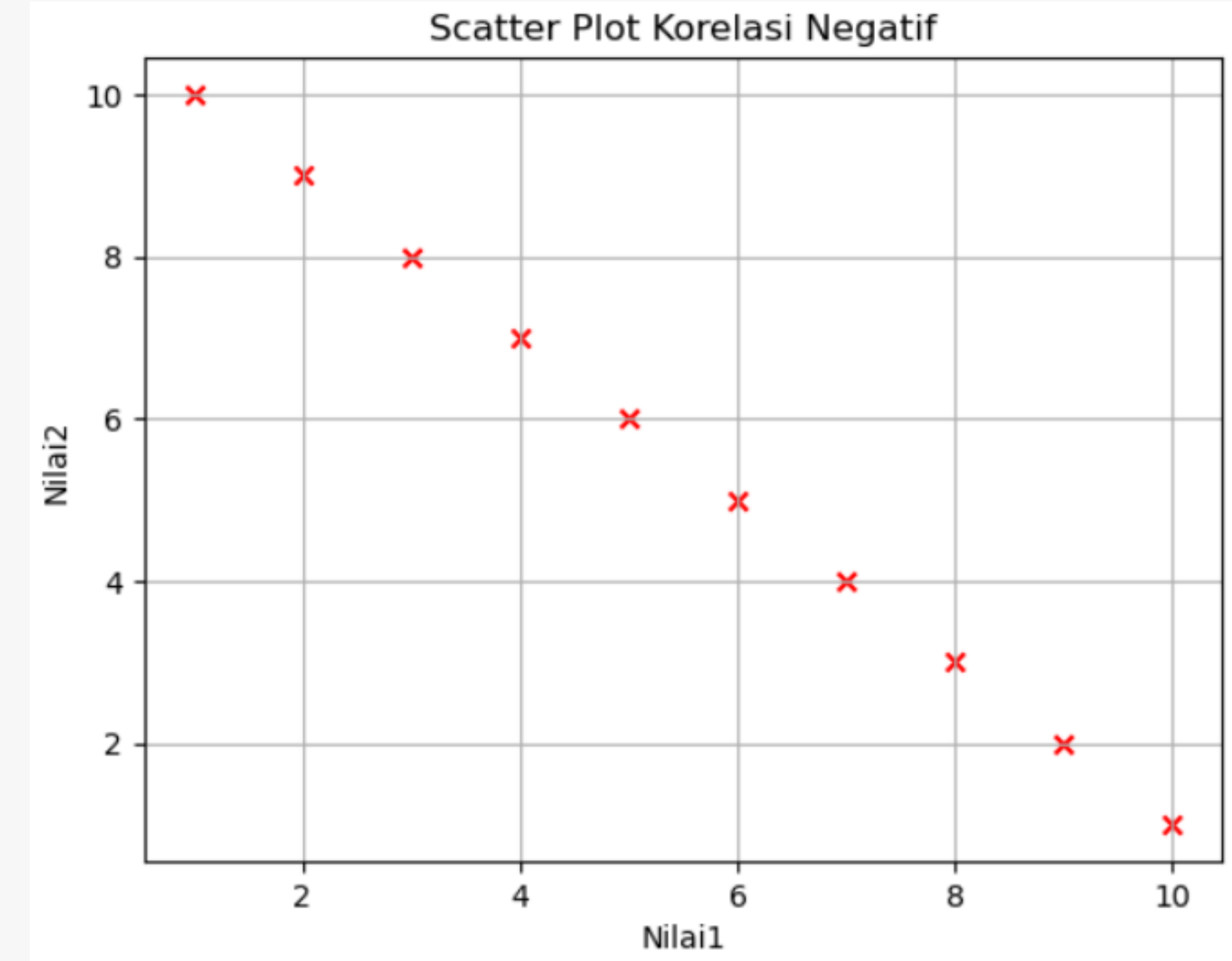
```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3
4 # Buat DataFrame contoh
5 data = {'Nilai1': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
6         'Nilai2': [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]}
7
8 df = pd.DataFrame(data)
9
10 # Buat scatter plot
11 plt.scatter(df['Nilai1'], df['Nilai2'], color='blue', marker='o')
12
13 # Tambahkan label
14 plt.title('Scatter Plot Korelasi Positif')
15 plt.xlabel('Nilai1')
16 plt.ylabel('Nilai2')
17
18 # Tampilkan plot
19 plt.grid(True)
20 plt.show()
```

PRAKTIKUM



Scatter Plot :: Korelasi Negatif

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3
4 # Buat DataFrame contoh
5 data = {'Nilai1': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
6         'Nilai2': [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]}
7
8 df = pd.DataFrame(data)
9
10 # Buat scatter plot
11 plt.scatter(df['Nilai1'], df['Nilai2'], color='red', marker='x')
12
13 # Tambahkan label
14 plt.title('Scatter Plot Korelasi Negatif')
15 plt.xlabel('Nilai1')
16 plt.ylabel('Nilai2')
17
18 # Tampilkan plot
19 plt.grid(True)
20 plt.show()
21
```



PRAKTIKUM

Praktikum

1. Praktikan kode program dalam slide ini, dikumpulkan sebagai praktikum Pekan 2 dikoordinir oleh ASDOS

2. Buat program untuk membagi dataset menjadi:

(a) Data Training: 80%

(b) Data Validation:
10% dari data training

(c) Data Testing: 20%

dan tampilkan datanya!

```
[8]: import pandas as pd

# Read the CSV file with a comma delimiter
df = pd.read_csv('../data/day.csv', sep=',')

# cetak header data (5 baris data) dari file
df.head()
```

Data diambil
dari praktikum sebelumnya

```
[8]:
```

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Referensi:



1. Sugiyono. (2013). Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif dan R&D). Bandung: Alfabeta.
2. Wawan Hafid Syaifudin, Achmad Choiruddin. 2021. Pengantar Teori Probabilitas dan Statistika. Elmarkazi Publisher.



Terima Kasih

<http://youtube.com/@rojulman>