



# **Pandas Dataframe - Introduction**



# Table of Content

## What will We Learn Today?

1. **Apa itu Pandas dan Mengapa Pandas?**
2. **List, Tupple, Numpy Array, Pandas Series**
3. **Apa itu dataframe?**
4. **Membuat dataframe**  
**Membaca dataframe**  
**Memilih dataframe**  
**Menulis dataframe**





# Profile




## Professional

- Senior Data Analyst – Kompas (2021 – Present)
- Data Scientist – Rukita (2020 – 2021)
- Research Assistant Analyst – Ensterna (2017 – 2019)

## Educational Background

- Nuclear Engineering – Universitas Gadjah Mada

## Connect with me

-  <https://dataimpact.medium.com/>
-  <https://www.linkedin.com/in/ariprabowo/>
-  <https://github.com/densaiko>



**Ari Sulistiyo Prabowo**



# **Pandas**

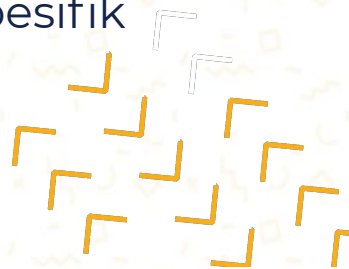


*Pandas adalah **alat analisis** dan **manipulasi data** yang handal, cepat, dan mudah digunakan yang dibangun di atas bahasa pemrograman Python*



# Mengapa Pandas digunakan?

- ✓ **Mudah** untuk melakukan analisis data dengan memanipulasi data yang sesuai kebutuhan
- ✓ **Dapat** membaca beberapa tipe file seperti CVS, Excel, database SQL, dan HDFS file
- ✓ **Melakukan** pivot table seperti yang ada di excel
- ✓ **Mudah** untuk memfilter data yang diinginkan untuk analisa lebih dalam dan spesifik





# List, Tuple, Numpy Array, Pandas Series

List [ ... ]	Tuple ( ... )	Array np[ ... ]	Series pd[ ... ]
Native Python	Native Python	Numpy	Pandas
mutable	immutable	mutable	mutable
Anggota list dapat diubah dan diganti	Anggota tuple tidak dapat diubah dan diganti	Anggota np.array dapat diubah dan diganti	Anggota pd.series dapat diubah dan diganti
indexed	indexed	indexed	indexed
Bisa memuat berbagai macam tipe data dalam 1 list	Bisa memuat berbagai macam tipe data dalam 1 tuple	Array hanya menyimpan tipe data yang sama dalam 1 array	Series hanya menyimpan tipe data yang sama dalam 1 series





# Dataframe

The diagram illustrates a Dataframe as a table with columns and rows. The columns are labeled: Name, Team, Number, Position, and Age. The rows are indexed from 0 to 6. A specific row (index 2) is highlighted with a purple box, and its data is labeled 'Data'. The 'Data' label is connected to the highlighted row by a purple line. The 'Rows' label is connected to the row indices by orange arrows. The 'Columns' label is connected to the column headers by blue arrows.

	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

**Dataframe** adalah struktur data yang berlabel 2 dimensi dengan kolom yang memiliki tipe data yang berbeda. Seperti halnya yang ada di Excel atau SQL

Dataframe memiliki **3 komponen**:

- Kolom
- Baris
- dan Data



# Membuat Dataframe



dari **Python List**

```
list1 = ["Joy", "Steward", "Nelly"]  
list2 = [23, 28, 25]  
list3 = ["Medan", "Jakarta", "Surabaya"]  
  
dataframe_list = pd.DataFrame(list(zip(list1, list2, list3)),  
                               columns=['Name', 'Age', 'Location'])  
dataframe_list
```

	Name	Age	Location
0	Joy	23	Medan
1	Steward	28	Jakarta
2	Nelly	25	Surabaya





# Membuat Dataframe



dari **Python Tuple**

```
tupple1 = ("Harry Potter", 8)
tupple2 = ("Jack Bordon", 5)

dataframe_tupple = pd.DataFrame([tupple1, tupple2], columns=["Name", "Rating"])
dataframe_tupple
```

	Name	Rating
0	Harry Potter	8
1	Jack Bordon	5



# Membuat Dataframe



dari **Python Numpy Array**

```
array = np.array([[ "Meetball",5,7.7], [ "Fried Rice",3,8.0],[ "Pizza",10.0,9.0]])

dataframe_array = pd.DataFrame(array, columns=[ "Food","Price (USD)","Rate" ])
dataframe_array
```

	Food	Price (USD)	Rate
0	Meetball	5	7.7
1	Fried Rice	3	8.0
2	Pizza	10.0	9.0



# Membuat Dataframe



dari **Python Pandas Series**

```
series1 = pd.Series(["House A", "House B", "House C"])
series2 = pd.Series([100000, 250000, 300000])

series_dataframe = pd.DataFrame({"Type of House":series1, "Price":series2})
series_dataframe
```

	Type of House	Price
0	House A	100000
1	House B	250000
2	House C	300000





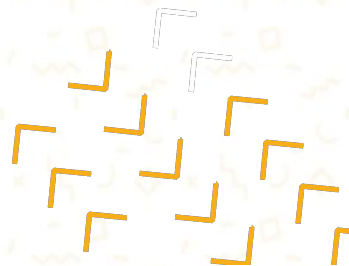
# Membaca file menggunakan Pandas



dari **CSV Format**

```
csv_file = pd.read_csv("insurance.csv")
csv_file.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520





# Membaca file menggunakan Pandas



dari **Excel Format**

```
excel_file = pd.read_excel("Startups Data (1).xlsx", sheet_name="Overview")
excel_file.head()
```

	ID	Name	Industry	Description	Year Founded	Employees	State	City	Metro Area
0	1	Over-Hex	Software	Provides a Web-based CRM tool that allows hosp...	2006	25	TN	Franklin	Nashville
1	2	Unimattax	IT Services	Helps law firms use Thomson Reuters Elite prac...	2009	36	PA	Newtown Square	Philadelphia
2	3	Lexila	Real Estate	Offers investment, construction, residential, ...	2013	38	IL	Tinley Park	Chicago
3	4	Greenfax	Retail	A Verizon Wireless premium retailer that offer...	2012	320	SC	Greenville	Newberry, SC
4	5	Saoace	Energy	An energy efficiency consulting firm that work...	2009	24	WI	New Holstein	Appleton, WI

**Thank  
YOU**

