



DIGITAL  
TALENT  
SCHOLARSHIP

# Professional Academy Data Analyst (DAQ)

“Sharing Session”  
EXPLORATORI & DATA ANALYSIS



DΦLab

[www.dqlab.id](http://www.dqlab.id)



KOMINFO

#JADIJAGOANDIGITAL

Badan Penelitian dan Pengembangan Sumber Daya Manusia



IG : @Sandi\_Wanda

# HELLO!

**I am Sandi Wanda Harlan**

S1 - Mathematics (Andalas  
University)

**Region Credit Analyst (Fifgroup)**

**LinkedIn : Sandi Wanda Harlan**



IG : @Sandi\_Wanda

# HELLO!

I am Sandi Wanda Harlan

## Course & Certificate :

- ❑ LinkedLearning Became Data Analyst Certified (2020)
- ❑ IBM Data Science Professional Certified (2020)
- ❑ Datacamp SQL for Business Analyst Skill Track (2020)
- ❑ Datacamp Finance Fundamental Skill Track (2020)
- ❑ IBM Data Analyst Professional Certified (2021)
- ❑ Google Data Analytics Certified (2021)
- ❑ DQLab Pandora Box – Data Engineer Program (2021)
- ❑ Asisten Mentor DTS – Thematic Academy, Big Data for Social Science (2021)



# Materi hari ini

- Introuduction & Review
- Library Python yang digunakan dalam EDA
- Retrieving Data
- Cleaning Data
- EDA dengan menggunakan Statistik
- EDA dengan mengggunakan Visualisasi
- Hands on



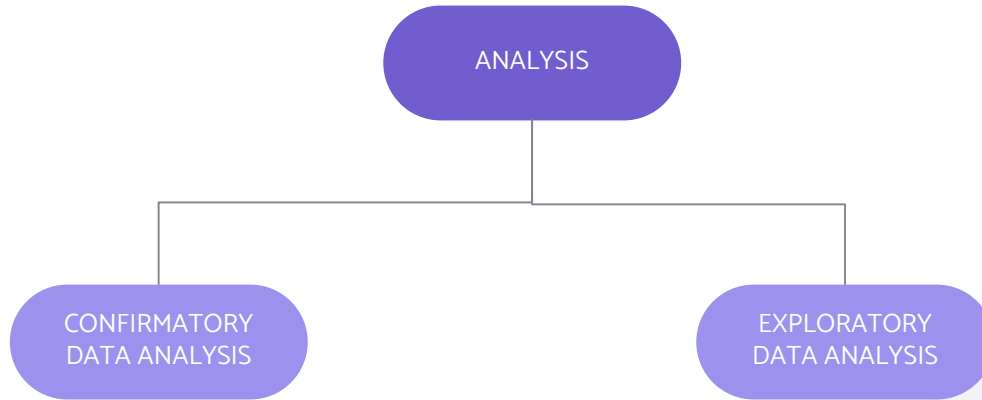
1

# Exploratory Data Analysis

Materi Presentasi

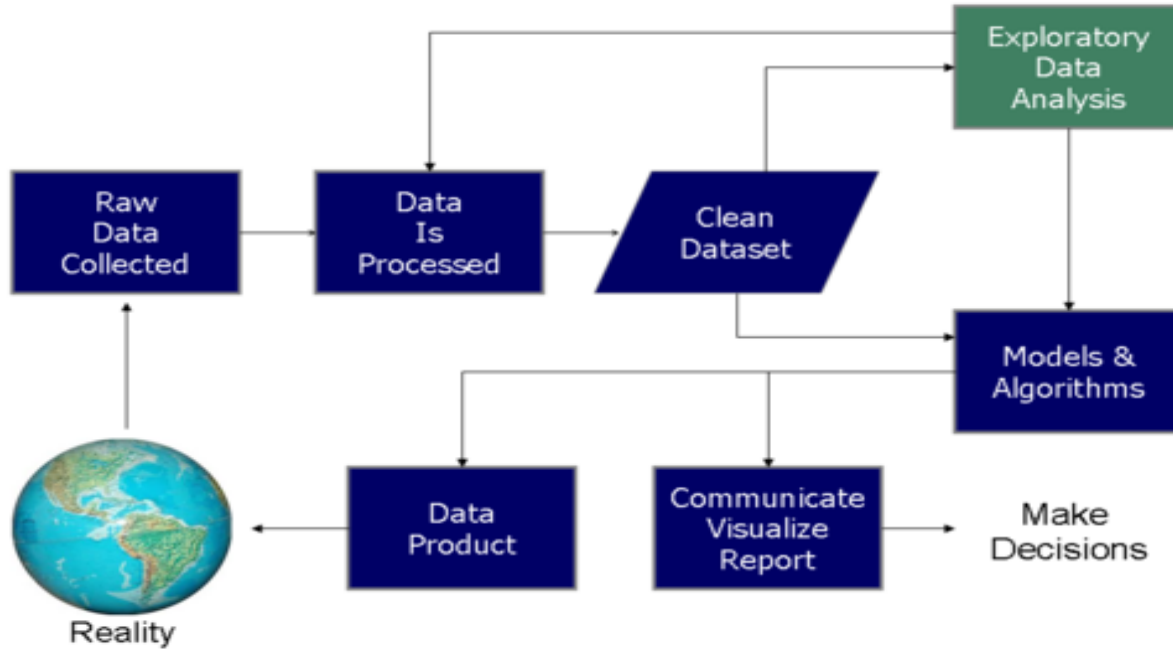
# Apa itu Exploratory data analysis ?

- Exploratory Data Analysis adalah proses kritis dalam melakukan investigasi awal pada data untuk menemukan pola, untuk menemukan anomali, untuk menguji hipotesis dan untuk memeriksa asumsi dengan bantuan statistik ringkasan dan representasi grafis





## Data Science Process





## Tujuan EDA :

- Memaksimalkan penyelidikan data
- Mencari struktur data yang tersembunyi
- Melihat Variabel penting
- Mendeteksi kelainan atau anomali
- Membangun Model





## Bagaimana jika skip EDA :

- Model tidak akurat
- Memilih variable yang salah untuk model
- Informasi kurang akurat (Analisa dangkal)



# Point penting dalam EDA:

## Statistik

- Measure of central Tendency
- Measure of Variation
- Hypothesis Testing
- ANOVA
- Regresi
- Correlation
- Etc.

## Visualisasi

- Histogram
- barplot
- Heatmap
- Boxplot
- Etc.



2

## **Library EDA**

Materi Presentasi



# Mengapa Python?

- Python diciptakan oleh Guido van Rossum dan pertama kali diperkenalkan pada tahun 1991 sebagai sebuah proyek open-source
- General-purpose programming
- High-level programming language.
- Lisensi dari Python bersifat open-source dari Python

# Library yang digunakan :

## Numpy

Numpy berasal dari kata '*Numerical Python*', sesuai namanya NumPy berfungsi sebagai library untuk melakukan proses komputasi numerik terutama dalam bentuk *array* multidimensional (1-Dimensi ataupun 2-Dimensi).

*Array* merupakan kumpulan dari variabel yang memiliki tipe data yang sama. NumPy menyimpan data dalam bentuk *arrays*.



<https://numpy.org/>

Bentuk 1D NumPy *array* dapat diilustrasikan sebagai berikut:



Bentuk 2D NumPy *array* dapat diilustrasikan sebagai berikut:



# Library yang digunakan :

## Pandas

Pandas merupakan library yang memudahkan dalam melakukan manipulasi, *cleansing* maupun analisis struktur data. Dengan menggunakan Pandas, dapat memanfaatkan lima fitur utama dalam pemrosesan dan analisis data, yaitu *load*, *prepare*, *manipulate*, *modelling*, dan *analysis* data.

Pandas menggunakan konsep array dari NumPy namun memberikan index kepada array tersebut, sehingga disebut *series* ataupun *data frame*. Sehingga bisa dikatakan Pandas menyimpan data dalam *dictionary-based* NumPy arrays. 1-Dimensi labelled array dinamakan sebagai *Series*. Sedangkan 2-Dimensi dinamakan sebagai *Data Frame*.

Bentuk dari *series* diilustrasikan sebagai berikut

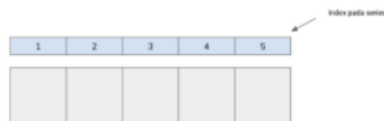


Diagram illustrating the structure of a Pandas Series. It shows a single row of 5 cells. Above the cells are indices 1, 2, 3, 4, and 5. An arrow points to the indices with the label "Index pada series".

1	2	3	4	5

Bentuk dari *data frame* diilustrasikan sebagai berikut:

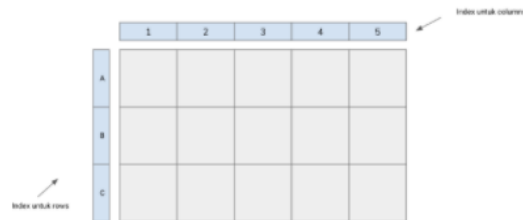


Diagram illustrating the structure of a Pandas Data Frame. It shows a grid of 3 rows and 5 columns. The rows are labeled A, B, and C on the left. The columns are labeled 1, 2, 3, 4, and 5 at the top. Arrows point to the row and column labels with the labels "Index untuk rows" and "Index untuk columns" respectively.

	1	2	3	4	5
A					
B					
C					



<https://pandas.pydata.org/>

# Library yang digunakan :

## SciPy

Scipy dibangun untuk bekerja dengan array NumPy dan menyediakan banyak komputasi numerik yang ramah pengguna dan efisien seperti rutinitas untuk integrasi, diferensiasi dan optimasi numerik. Baik NumPy maupun SciPy berjalan pada semua operating system, cepat untuk diinstall dan gratis. NumPy dan SciPy mudah digunakan, tetapi cukup kuat untuk diandalkan oleh beberapa data *scientist* dan *researcher* terkemuka dunia.



<https://www.scipy.org/>

# Library yang digunakan :

## Matplotlib & seaborn

Matplotlib merupakan library dari Python yang umum digunakan untuk visualisasi data. Matplotlib memiliki kapabilitas untuk membuat visualisasi data 2-dimensional. Contoh visualisasi yang dapat dibuat dengan menggunakan matplotlib diantaranya adalah

1. Line chart
2. Bar chart
3. Pie chart
4. Box plot chart
5. Violin chart
6. Errorbar chart
7. Scatter chart

Jenis-jenis *chart* lainnya juga dapat dibuat melalui library ini.

<https://matplotlib.org/>

<https://seaborn.pydata.org/>







# 3

## Retrieving data

Materi Presentasi

<https://archive.ics.uci.edu/ml/datasets/Automobile>

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	symboling	normalize	make	fuel-type	aspiration	num-of-d	body-style	drive-whe	engine-lo	wheel-ba	length	width	height
2	3 ?		alfa-rome	gas	std	two	convertib	rwd	front	88.6	168.8	64.1	48.8
3	3 ?		alfa-rome	gas	std	two	convertib	rwd	front	88.6	168.8	64.1	48.8
4	1 ?		alfa-rome	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4
5	2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3
6	2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3
7	2 ?		audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1
8	1	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7
9	1 ?		audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7
10	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9
11	0 ?		audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52
12	2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3
13	0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3
14	0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3
15	0	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3
16	1 ?		bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7
17	0 ?		bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7
18	0 ?		bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7
19	0 ?		bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56.3
20	2	131	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.2	52.2



# Retrieving Data

- ◆ Import Library

```
import pandas as pd
import numpy as np
```

- ◆ Baca File sebagai Data frame

```
# Membaca file CSV
[nama_variabel] = pd.read_csv("nama_file.csv")
```

```
# Membaca file Excel
[nama_variabel] = pd.read_excel("nama_file.xlsx")
```



# 3

## Cleaning data

Materi Presentasi



# Pembersihan Data

- Duplikat data

Solusi: Remove Duplicate

- Format yang tidak konsisten

Solusi: Ganti ke format yang sesuai

- Nilai NULL atau Missing Value

Solusi : hapus baris/kolom, mengisi nilai

- Outlier, Deteksi dengan menggunakan Histogram/ boxplot

Solusi : Hapus outlier, mengisi nilai, Normalisasi,  
Satandarisasi



3

## EDA Basic

Materi Presentasi



# EDA Basic

## Inspeksi struktur data frame

- melihat struktur data frame, (shape)

```
print([nama_dataframe].shape)
```

- melihat preview data dari dataframe tersebut (head)

```
# Menampilkan konten teratas dari [nama_dataframe]  
# untuk sejumlah bilangan bulat [jumlah_data]  
print([nama_dataframe].head([jumlah_data]))
```

- membuat summary data sederhana dari dataset(describe)

```
print([nama_dataframe].describe())
```



# EDA Basic

Measure of central Tendency

- Mean
- Modus
- Median
- Etc.





# EDA Basic

Measure of central Variation

- Variance
- SD
- Range
- Interkuartil
- Heatmap
- Etc.



# EDA Basic

Basic pengolahan data

- Value count
- Slicing
- Grouping
- Pivot
- Etc.



# EDA

## Visualisasi

- Histogram (sebaran data)
- Boxplot (deteksi outlier)
- Regplot (korelasi)
- Heatmap(korelasi)



2

# Hands On EDA

Dataset Automobile