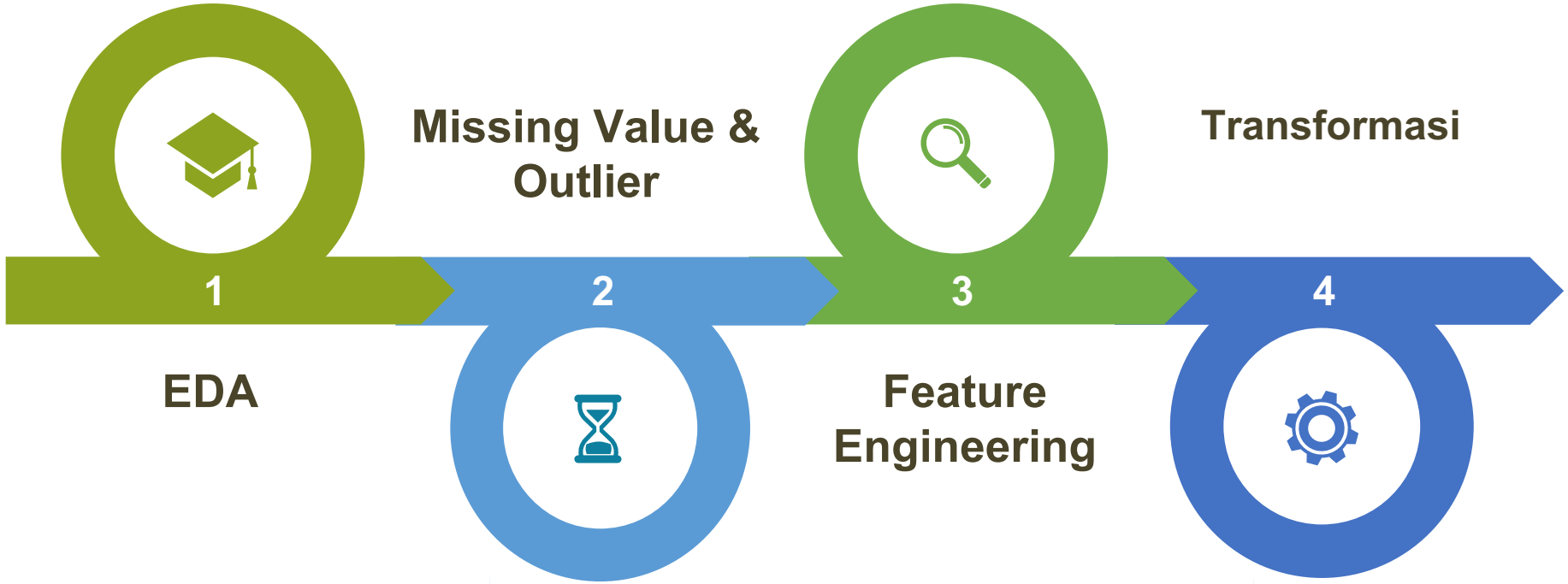
Data Preparation (2)

Rifqi F. dan Shofinurdin

Diklat Pengolahan Data dan Machine Learning Angkatan II 2024

OUTLINE


1
EDA
2
Missing Value & Outlier
3
Feature Engineering
4
Transformasi

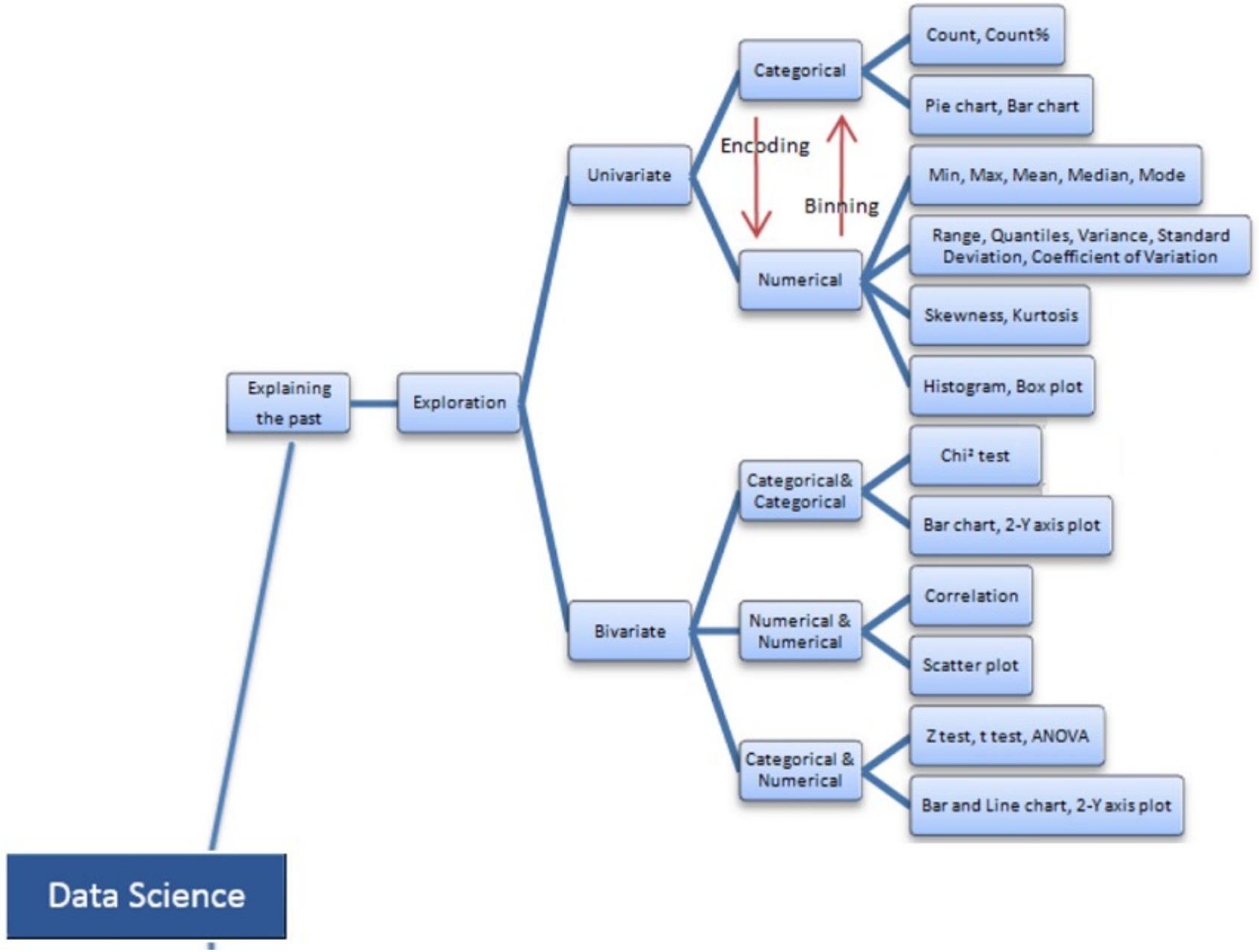
EDA

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to :

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

Manfaat EDA

- Memahami permasalahan yang ada pada data kita
- Merumuskan pertanyaan-pertanyaan bisnis lain
- Menghasilkan output berupa hasil analisis deskriptif yang dapat disajikan pada visualisasi data dan dashboard.


Data Science
Explaining the past
Exploration
Univariate
Bivariate
Encoding
Binning
Categorical
Numerical
Count, Count%
Pie chart, Bar chart
Min, Max, Mean, Median, Mode
Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
Skewness, Kurtosis
Histogram, Box plot
Categorical& Categorical
Numerical & Numerical
Categorical & Numerical
Chi² test
Bar chart, 2-Y axis plot
Correlation
Scatter plot
Z test, t test, ANOVA
Bar and Line chart, 2-Y axis plot

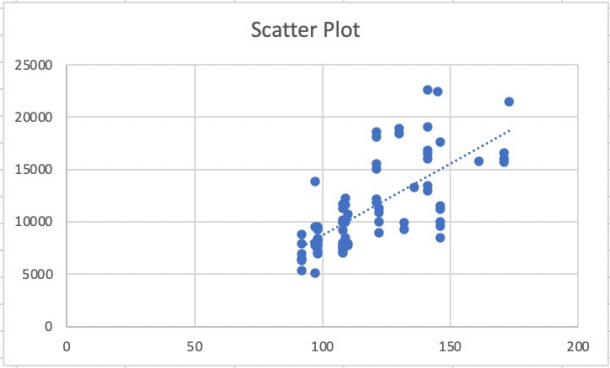
Tools

- Excel
- Python
- D Tale
- Bamboolib
- Etc.

Excel EDA

- Load dataset
- Rapikan dengan ‘text to colomn’
- Pilih kolom engine size dan price
- Ambil nilai korelasi
- Ambil nilai rata-rata masing kolom
- Buat scatterplot
- Tambahkan trendline

	A	B	C	D	E	F	G	H	I
1	engine-size	price							
2	130	13495							
3	130	16500							
4	152	16500							
5	109	13950							
6	136	17450		**korelasi**	**0,87233517**				
7	136	15250		**avg engine**	**126,907317**				
8	136	17710		**avg price**	**13207,1294**				
9	136	18920							
10	131	23875							
11	131	?							
12	108	16430							
13	108	16925							
14	164	20970							
15	164	21105							
16	164	24565							
17	209	30760							
18	209	41315							
19	209	36880							
20	61	5151							
21	90	6295							
22	90	6575							
23	90	5572							
24	90	6377							
25	98	7957							
26	90	6229							
27	90	6692							
28	90	7609							



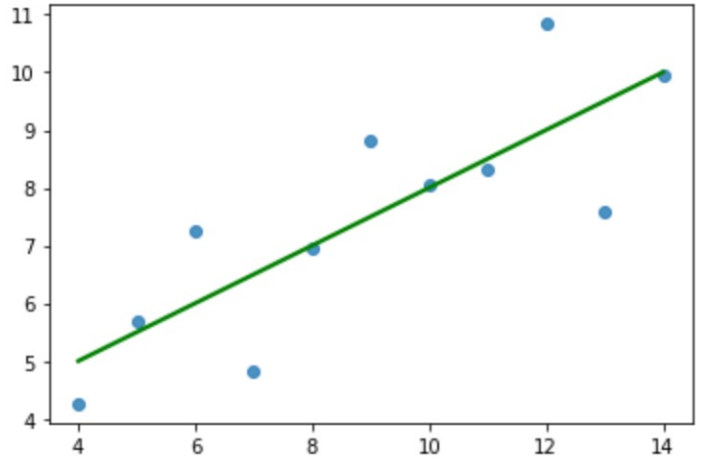
Python EDA

- Load dataset
- Statistic Describe
- Scatter plot

```
# membuat variable data
x = np.array([10.00, 8.00, 13.00, 9.00, 11.00, 14.00, 6.00, 4.00, 12.00, 7.00, 5.00])
y = np.array([8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68])
```

```
# # melihat statistik sederhana dari dataset
dataset.describe()
```

```
<AxesSubplot:>
```



D-Tale EDA

D-Tale is the combination of a Flask back-end and a React front-end to bring you an easy way to view & analyze Pandas data structures.

- Mudah digunakan.
- Terintegrasi dengan ipython notebooks dan python/ipython terminals.
- Sudah support DataFrame, Series, MultiIndex, DatetimeIndex & RangeIndex.

Step by step

- Install library D-Tale
- Import library
- Load dataset
- Show dataset

```
!pip3 install dtale
```

```
import dtale
```

```
import pandas as pd
df_titanic = pd.read_csv('https://raw.githubusercontent.com/CRMDSDIP/PJJDAS_IV/main/Data_Preparation/day2/d
```

```
dtale.show(df_titanic)
```

How to ?

- Describe, Histogram, QQ- Plot
- Cek outlier
- Cek korelasi antara kolom
- Cek missing value
- Charts

Bamboolib

GUI untuk pandas DataFrames yang memungkinkan siapa saja bekerja dengan Python di Jupyter Notebook atau JupyterLab

- Install bamboolib

```
!pip3 install bamboolib
```

- Import bamboolib
- Load dataset

```
import bamboolib as bam
df_titanic
```

executed in 10ms, finished 20:15:38 2022-10-01

Missing Value and Outlier

Missing values refer to the absence or lack of data for a variable or entry in a dataset. They occur when no data is recorded or available for a particular observation or variable. Missing values can arise due to various reasons, such as data collection errors, participant non-response, or data processing issues.

- Missing values are common occurrences in data.
- Unfortunately, most predictive modeling techniques cannot handle any missing values.
- Therefore, this problem must be addressed prior to modeling.

Missing Value Type

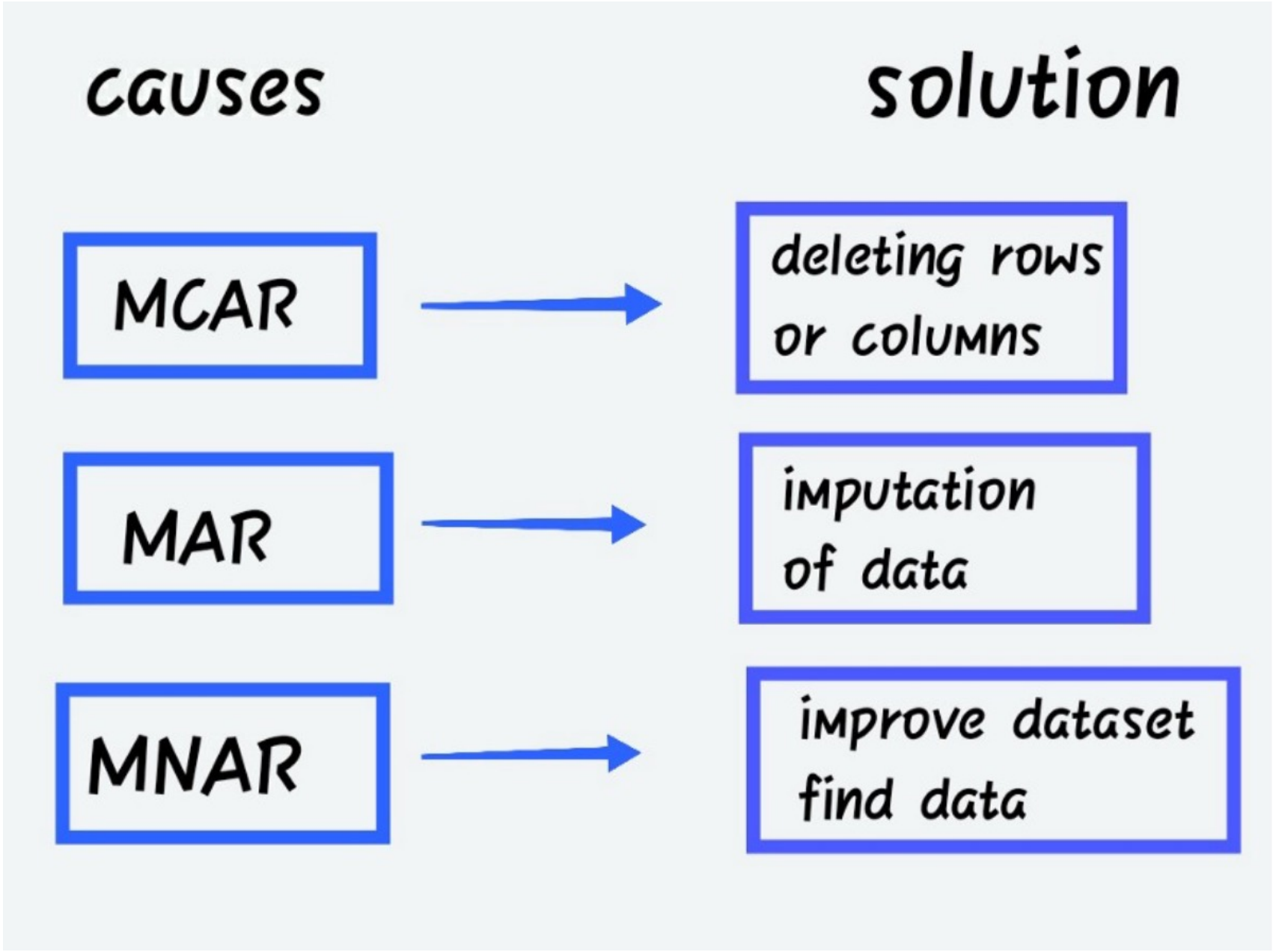
1. Missing completely at random (MCAR).
2. Missing at random (MAR).
3. Missing not random (MNAR).

Misal kita diminta untuk membuat model dari Age (Y) yang dipengaruhi oleh jenis kelamin (X) beberapa responden tidak memberikan jawaban atas pertanyaan berapa 'berat badan' mereka.

- MCAR terjadi jika tidak ada alasan yang jelas kenapa responden tidak memberikan jawaban.
- MAR terjadi jika orang yang berjenis kelamin perempuan akan cenderung tidak memberikan jawaban jika ditanya berat badan, jadi missing value dipengaruhi oleh X.
- MNAR terjadi jika orang dengan kelas penghasilan tertentu, memiliki kecenderungan tidak memberikan jawaban. Sehingga missing value dipengaruhi oleh nilai lain yang tidak teramati.

Beberapa fungsi utama yang digunakan:

1. `library missingno`
2. `df.isnull()`
3. `df.fillna()`
4. `df.groupby()`
5. `df.loc()`
6. `df.column.plot()`
7. `df.column.skew()`
8. `df.column.value_counts()`


causes
solution
MCAR
deleting rows or columns
MAR
imputation of data
MNAR
improve dataset find data

Handling Missing Value

Remove Rows / Columns

- remove column if n missing rows >> n rows
- remove row if n missing rows << n rows

Value Imputation

- mode / most frequent (categorical)
- mean / median (numerical)
- Random / defined value

Model based Imputation

- Use other features to predict missing rows

Missing Value Dataset Titanic

Hapus Baris yang NA

- `.dropna(how='any')` : hapus baris apabila memiliki **minimal 1 kolom** nilai missing value
- `.dropna(how='all')` : hapus baris apabila memiliki **semua kolom** nilai missing

Imputasi Kolom Age

Drop Kolom Cabin

Imputasi Kolom Embarked

Outliers

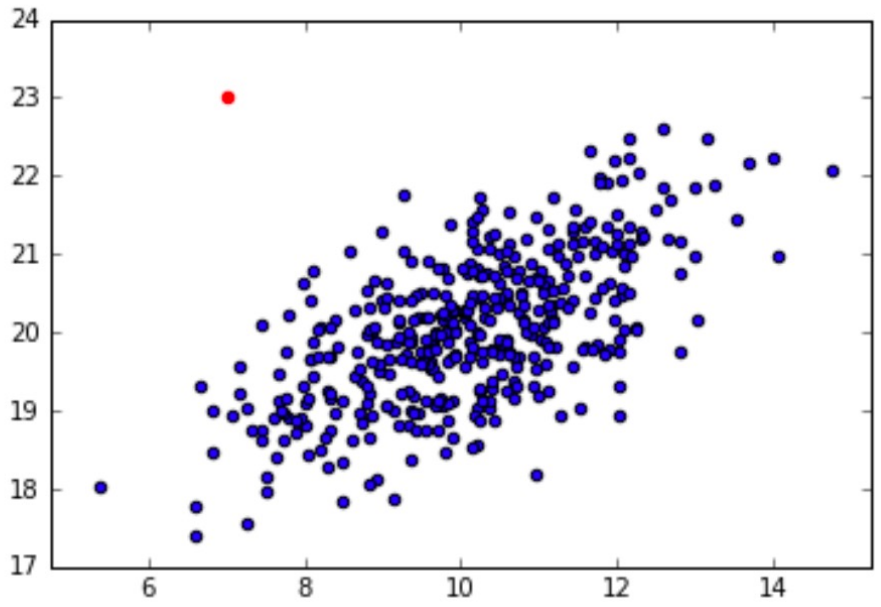
Merupakan suatu observasi yang berada di tempat yang jauh berbeda dibandingkan observasi lainnya dalam suatu populasi.

Contoh: hasil survei kekayaan

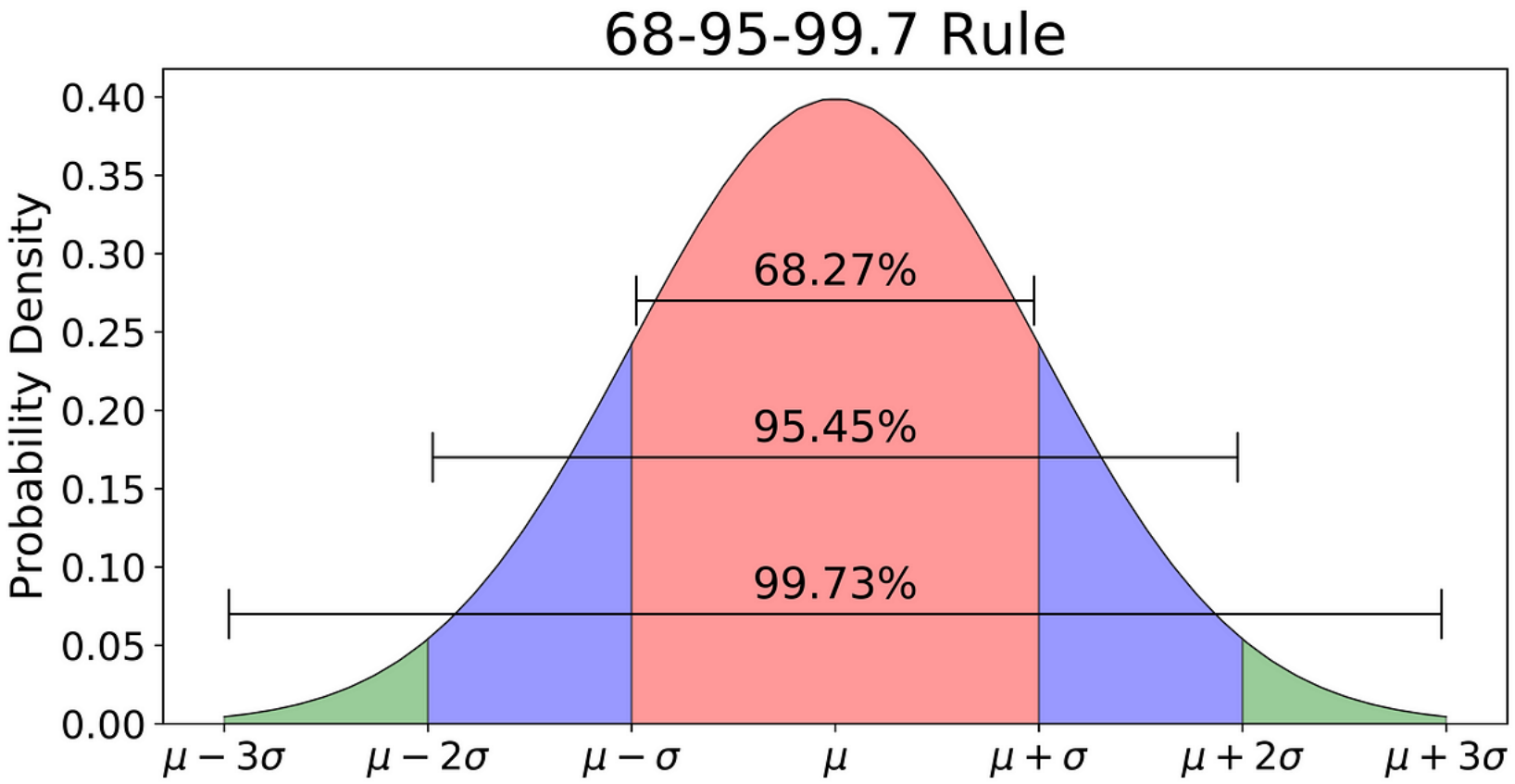
Definisi/Batasan outlier berbeda-beda, tergantung bagaimana analis menyikapi datanya.

Pengecekan outlier:

Memeriksa data menyeluruh melalui gambar/grafik dan/atau mencari data yang berbeda jauh dengan titik data secara umum



Three Sigma Rule



Hample Identifier

Kriteria Outlier untuk Hampel Identifier : 3xMADM

> Median Absolute Value from The Median (MADM)

$$MADM(x) = 1.4826 \times median\{|x_K - x^+|\}$$

Keterangan :

- x_K adalah data ke K
- x^+ adalah median dari data

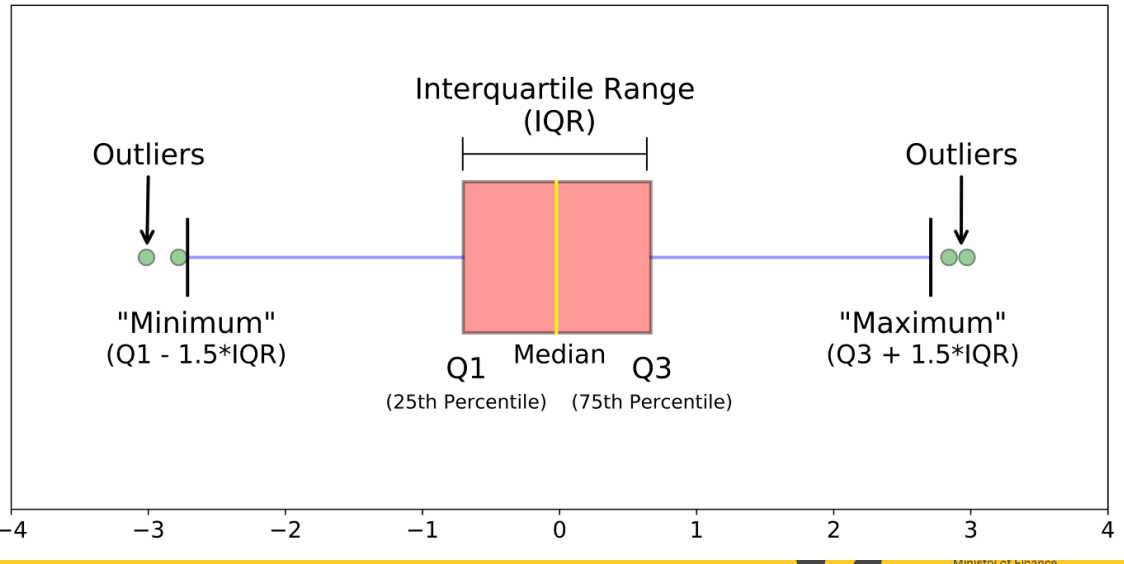
Boxplot

Kriteria Outlier untuk **Boxplot Outlier Rule** :

- $x_K > x_U + 1.5Q$
- $x_K > x_L - 1.5Q$

Keterangan :

- x_K adalah data ke-k,
- x_U adalah kuartil ke-1 atau disebut kuartil bawah (*lower quartile*),
- x_L adalah kuartil ke-3 atau disebut kuartil bawah (*upper quartile*),
- Q adalah jangkuan interkuartil (selisih kuartil bawah - kuartil atas)



Feature Enginering

feature engineering is focused on using the variables you already have to create additional features that are (*hopefully*) better at representing the underlying structure of your data

- Extraction
- Binning

Feature Transformation

- Transformasi fitur (feature transformation) adalah proses mengubah fitur (variabel) dalam dataset untuk meningkatkan kualitas analisis atau kinerja model yang akan dibangun.
- Tujuan dari transformasi fitur adalah untuk mengungkapkan informasi yang tersembunyi atau memperbaiki distribusi data yang tidak memenuhi asumsi tertentu

Feature Transformation:

- Encoding
- Normalization
- Extract Date/Time
- Feature form mathematics computation

Encoding

- Dictionary Encoding
- Label Encoding
- One Hot Encoding

Replacing datatypes with dictionary and indexes

Color
Red
Red
White
Black
Blue
Red
Blue
Black
Black

Dictionary Encoding

Color Id
0
0
1
2
3
0
3
2
2

Id	Color
0	Red
1	White
2	Black
3	Blue

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Kemenkeu
Corporate University

Terim Kasih

MoF-DAC
Ministry of Finance
Data Analytics Community