Rakamin Academy

Exploratory Data Analysis for Machine Learning

Rakamin Academy

Kenalan Yuk!

Hendra H Choiri
Data Scientist
gojek

gojek
Data Scientist

Past Experiences

Telkomsel
IT Analyst

bukalapak
Data Scientist Lead

Education Background

Institut Teknologi Bandung 1920
Teknik Informatika

TU Delft
Data Science & Technology

Saya

Universiteit Leiden

Master of Science
Leiden University (2017-2019)
Focused on application of theoretical machine learning & reinforcement learning

bukalapak

Senior Data Scientist
Bukalapak (2020 - present)

Pararawendy Indarjo

Email : pararawendy19@gmail.com
Linkedin : https://www.linkedin.com/in/pararawendy-indarjo/
Blog : medium.com/@pararawendy19

Saya

Pararawendy Indarjo

Linkedin : https://www.linkedin.com/in/pararawendy-indarjo/
Blog : medium.com/@pararawendy19

Universiteit Leiden

Master of Science
Leiden University (2017-2019)
Focused on application of theoretical machine learning & reinforcement learning

bukalapak

Senior Data Scientist
Bukalapak (2020 - July 2022)

allo fresh

Senior Data Manager
Allofresh (July 2022 - Now)

Rakamin Academy

Hafizh Adi P

Data Scientist

bukalapak

Halo!

東京工業大学
Tokyo Institute of Technology

Master of Eng. - Artificial Intelligence
(2017 - 2019)

Data Scientist (2019 - now)

Sesi ini:

1. Exploratory Data Analysis (EDA) itu apa?
2. Kenapa perlu melakukan EDA?
3. How to EDA
4. Hands-on: studi kasus botak
5. QnA

Rakamin Academy

Apa itu Exploratory Data Analysis (EDA)?

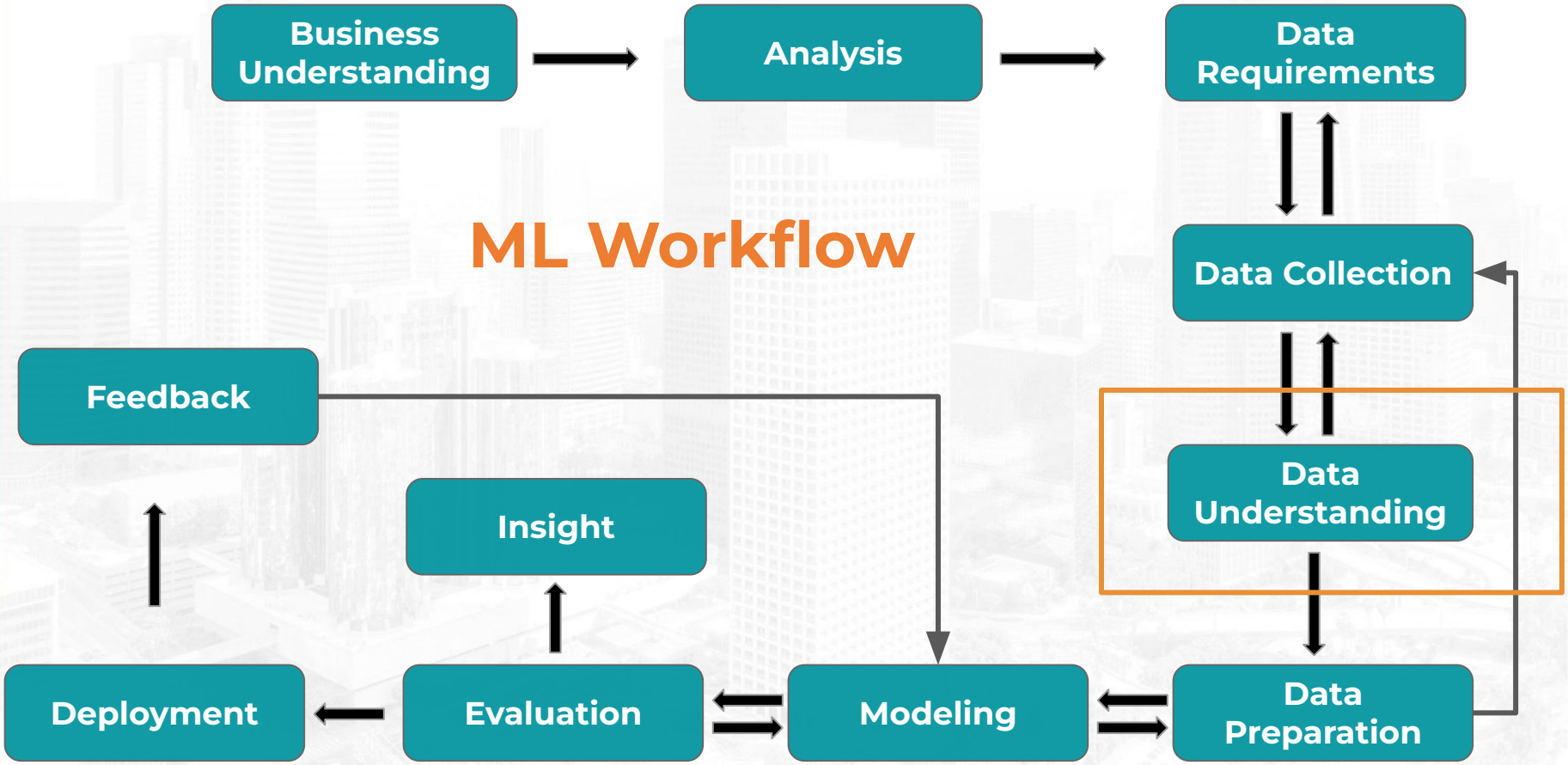
Data juga ingin dipahami

Rakamin Academy

“Exploratory Data Analysis (EDA) adalah proses analisis untuk memahami karakteristik data, dan hal-hal yang perlu kita lakukan agar data tersebut dapat digunakan untuk proses pembelajaran model ”

The Bigger Context

Rakamin Academy



Tahap	Masuk	Proses	Keluar
Data collection	-	• Survey/Labelling • ETL	• Data mentah
Data understanding	• **Data mentah**	• **Exploratory Data Analysis**	• **Data mentah** • **Insight/To do list**
Data preparation	• Data mentah • **To do list**	• Pre-processing • Feature processing	• Data *training* • Data *test/validation*
Modelling	• Data *training*	• Model training • Hyperparameter tuning	• ML Model
Evaluation	• Data *test/ validation*	• Validation	• Performance measure
Deployment	• Data baru	• Prediction	• Prediksi

Kenapa perlu EDA?

Rakamin Academy

Untuk menjawab pertanyaan berikut:

- Bagaimana sebaran nilai dalam variabel feature dan target kita?
- Apakah kira-kira feature yang kita miliki cukup baik untuk memprediksi target?
- **'Persiapan' macam apa** yang harus kita lakukan sebelum dataset kita dapat digunakan dalam proses pelatihan model ML?

Meningkatkan performa model yang kita bangun

Bagaimana cara melakukan EDA?

Dataset

Titanic

- **Deskripsi**:

  Memprediksi *survival* dari kecelakaan Titanic berdasarkan data-data penumpang.

- **Data**:

  Setiap baris mewakili penumpang, setiap kolom berisi atribut penumpang.

- **Link Kaggle**: https://www.kaggle.com/c/titanic/data

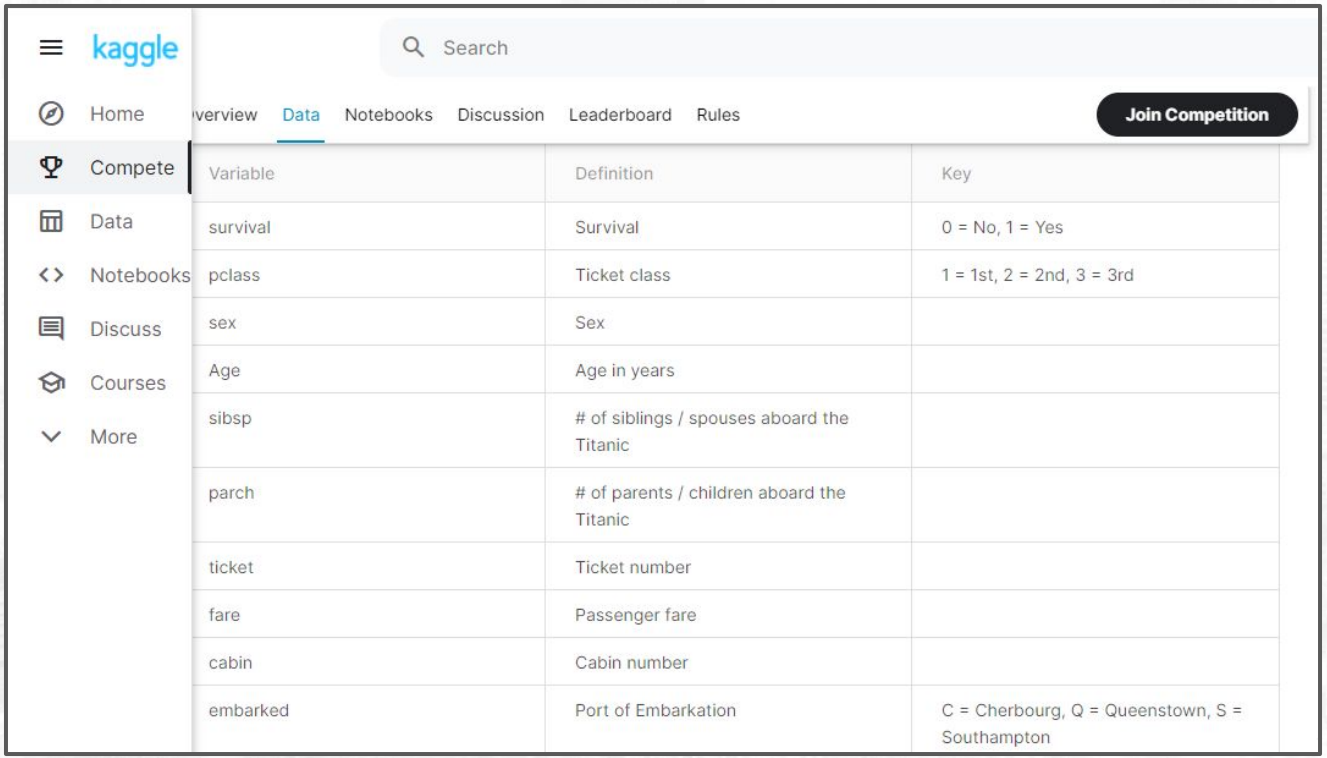
Import Libraries & Load Data

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 import seaborn as sns
```

```
1 df = pd.read_csv('train.csv')
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Data Dictionary (definisi setiap kolom)



Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Rakamin Academy

Bagaimana cara melakukan EDA?

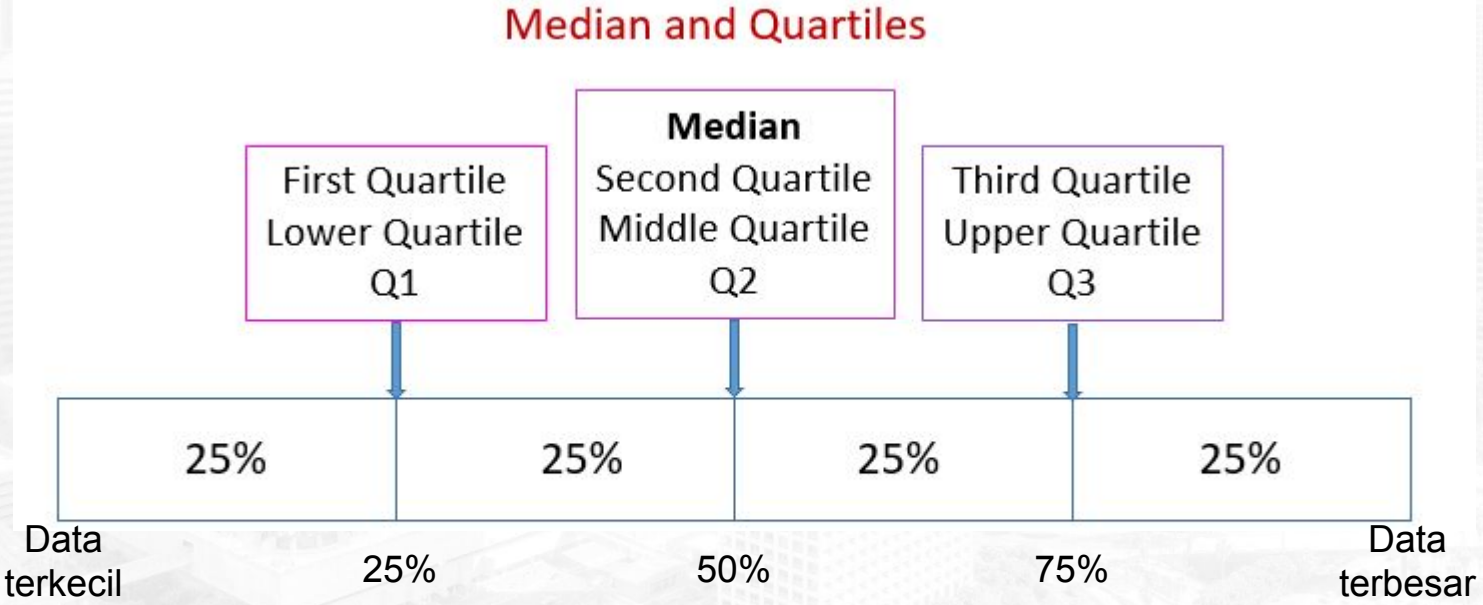
#1: Descriptive Statistics

***Descriptive Statistics*:**

Ringkasan statistik dari setiap kolom di dataset yang dapat memberikan gambaran besar keadaan data.

Rakamin Academy

Statistika Deskriptif



Pick + Separate Columns

```
1 numericals = ['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']
2 categoricals = ['Sex', 'Embarked']
```

Pisahkan kolom2 yang ingin dianalisis

Sample

```
1 df.sample(5)
```

`df.sample()`, `df.head()`, atau `df.tail()` akan menampilkan beberapa baris data secara langsung

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
464	465	0	3	Maisner, Mr. Simon	male	NaN	0	0	A/S 2816	8.0500	NaN	S
838	839	1	3	Chip, Mr. Chang	male	32.0	0	0	1601	56.4958	NaN	S
93	94	0	3	Dean, Mr. Bertram Frank	male	26.0	1	2	C.A. 2315	20.5750	NaN	S
442	443	0	3	Petterson, Mr. Johan Emil	male	25.0	1	0	347076	7.7750	NaN	S
386	387	0	3	Goodwin, Master. Sidney Leonard	male	1.0	5	2	CA 2144	46.9000	NaN	S

Yang perlu diperhatikan:

- **Apakah ada kolom dengan nilai yang tidak sesuai dengan nama kolom?**

Statistical Summary of Columns

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

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Yang perlu diperhatikan:

- **Apakah nilai yang tertera pada setiap kolom masuk akal?**
- **Apakah nilai maksimal/minimal masih berada di batas wajar?**
- **Jarak *mean vs median* bisa jadi indikasi kesimetrisan data**

Rakamin Academy

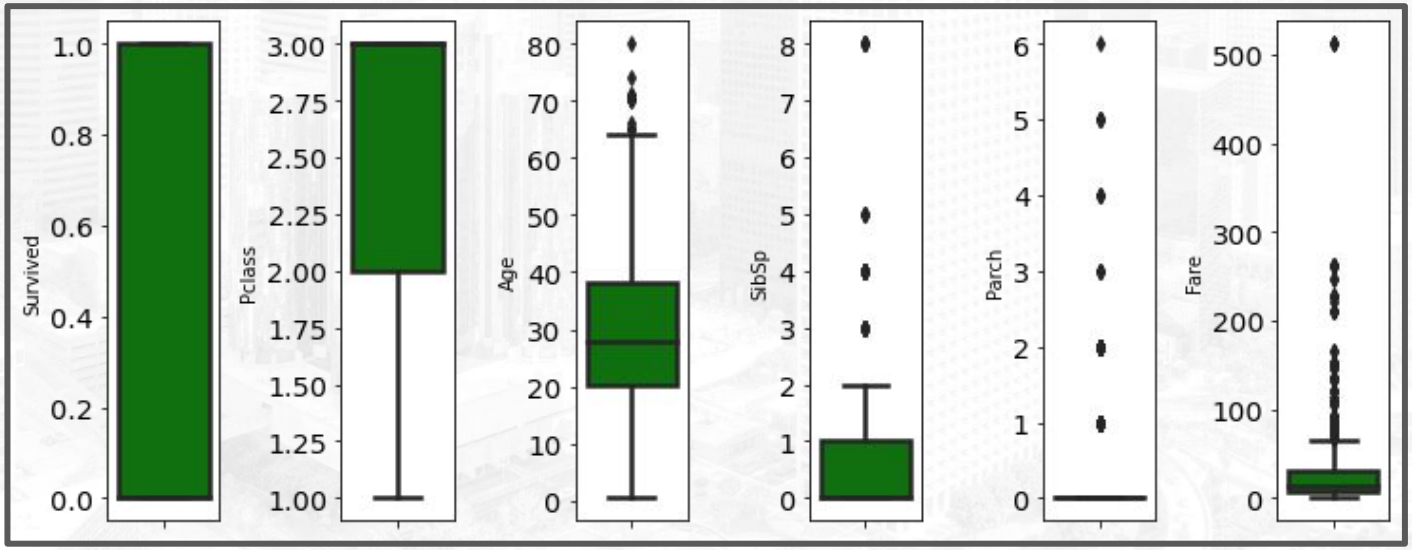
Bagaimana cara melakukan EDA?

#2: Univariate Analysis

Univariate Analysis:
Analisis setiap kolom secara terpisah, melihat distribusi nilainya secara detail

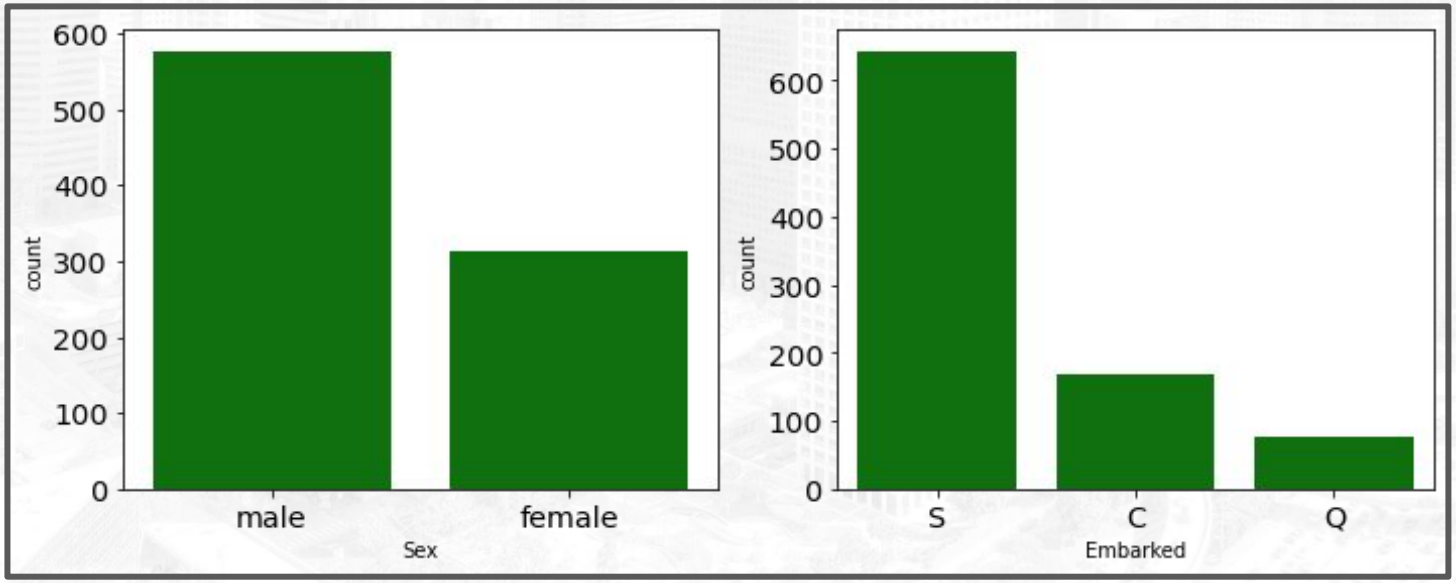
Individual Boxplots (Numeric)

```
1 plt.figure(figsize=(12, 4))
2 for i in range(0, len(features)):
3     plt.subplot(1, 7, i+1)
4     sns.boxplot(df[features[i]],color='green',orient='v')
5     plt.tight_layout()
```



Individual Countplot (Categorical)

```
features = categoricals
plt.figure(figsize=(10, 4))
for i in range(0, len(features)):
    plt.subplot(1, 2, i+1)
    sns.countplot(df[features[i]], color='green')
    plt.tight_layout()
```



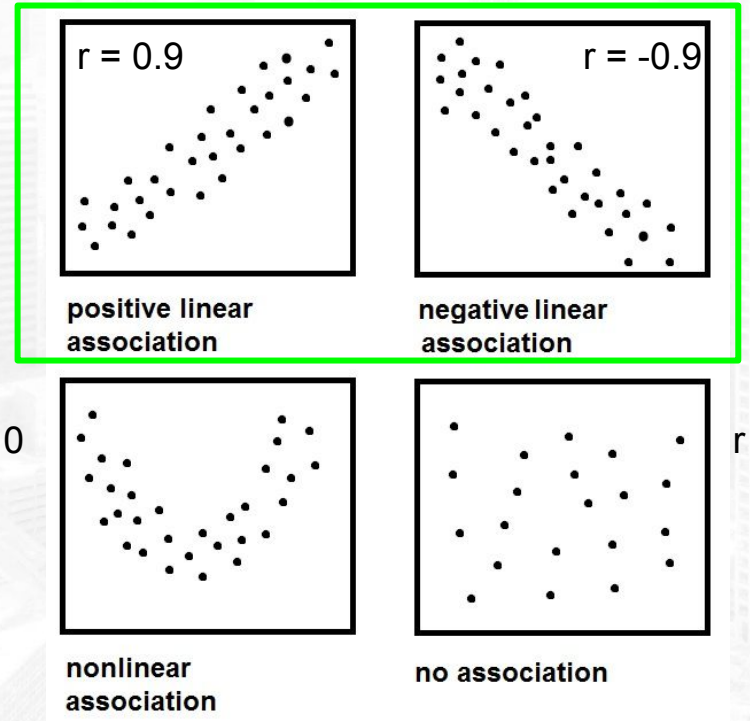
Rakamin Academy

Bagaimana cara melakukan EDA?

#3: Multivariate Analysis

Multivariate Analysis:
Analisis beberapa kolom sekaligus untuk mencari hubungan antar kolom

Korelasi Linear (Pearson correlation)

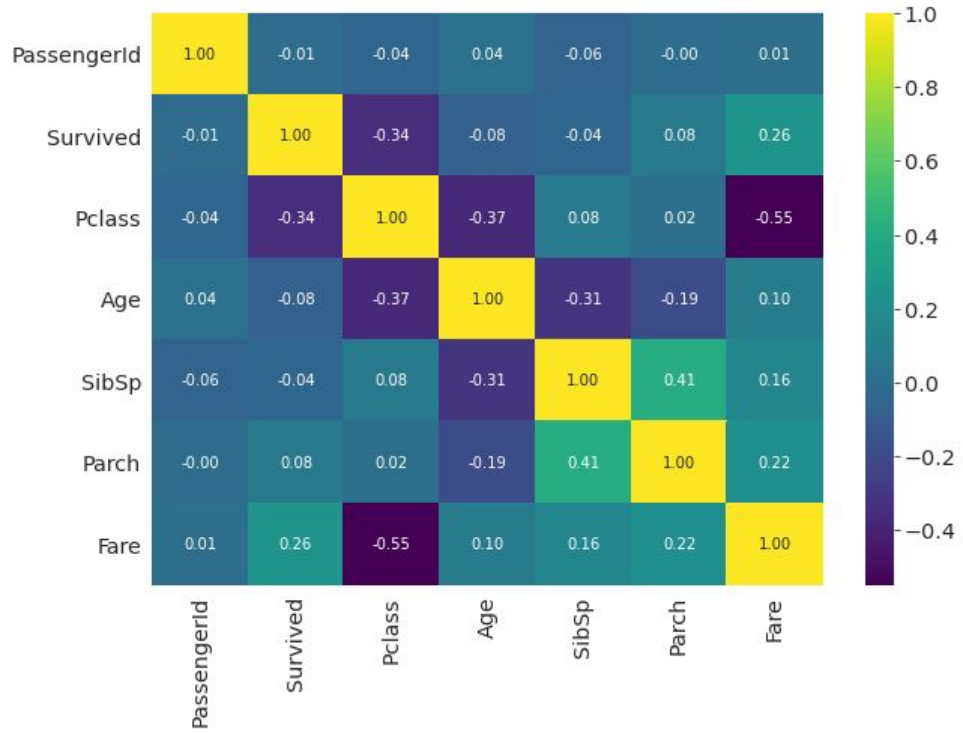


- **Pola hubungan antara X dan Y membentuk pola garis lurus**
  Semakin besar X, semakin besar pula Y ATAU semakin besar X, semakin kecil Y
  (kotak hijau di samping)

- **Nilai korelasi berkisar dari -1 s.d. 1**
  1: hubungan linear sempurna, searah
  -1 : hubungan linear sempurna namun berlawanan arah
  0: pola hubungan BUKAN linear

Correlation Heatmap (Numeric)

```
1 sns.heatmap(df.corr(), cmap='Blues', annot=True, fmt='.2f')
```



`df.corr()` akan mengembalikan matriks korelasi; `sns.heatmap()` membuat heatmap berdasarkan matriks

Yang perlu diperhatikan:

- **Apakah ada fitur-fitur yang berkorelasi kuat (>0.7)?** Bila ya, ada kemungkinan besar kedua feature tersebut redundan

Hands-on: EDA Prediksi Kebotakan

Hands-On Required :

Hands - On :

Exploratory Data Analysis Hands On.ipynb

Dataset :

1. **botak.csv**

Link: bit.ly/EDATrial

Dataset

Botak.csv

- **Deskripsi**:

  Dataset sintetik. Memprediksi peluang botaknya seseorang dari beberapa atribut mengenai orang tersebut.

- **Data**:

  Setiap baris mewakili satu orang, setiap kolom berisi atribut orang tersebut.

Sudah.

Sesi tanya-jawab