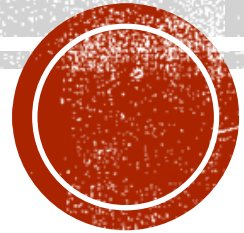
MEMAHAMI DATA DARI KAGGLE

Wicaksono Yuli Sulistyo



POIN PEMBELAJARAN

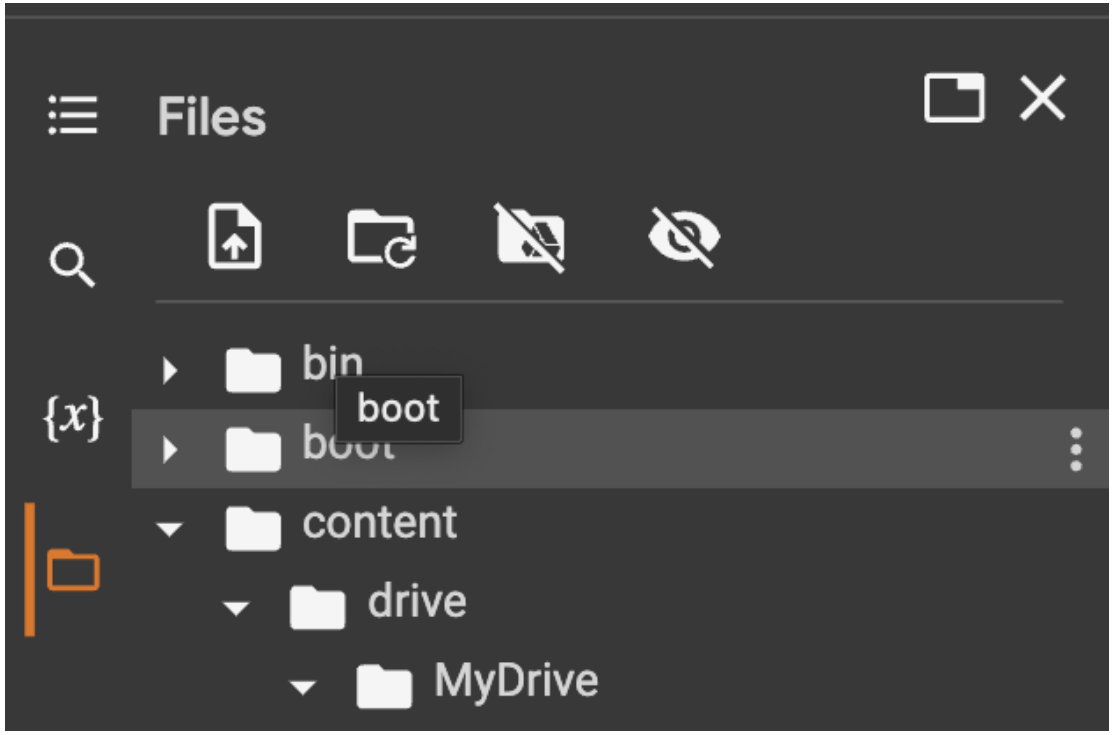
- Import
- Display Data
- Indexing

Tools

https://www.kaggle.com/datasets/ummisalma/mcu-movies-and-tv-shows

Menghubungkan Google Drive ke Google Collaboratory

1. Menjalankan kode di bawah ini **from google.colab import drive**
   drive.mount('/content/gdrive')
1. Pop up izin akses ke google drive akan muncul, lalu pilih izinkan.



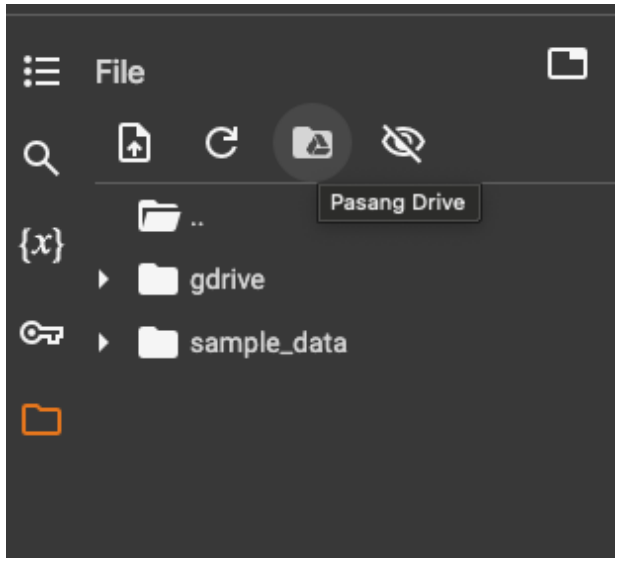
```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

Salah satu tanda jika sudah berhasil menghubungkan google drive dengan google collab

Atau dengan cara ke 2, di klik bagian folder

Klik revoke jika belum terhubung dengan drive



IMPORT MODULE

Module merupakan sekumpulan *function* yang dibangun dengan tujuan tertentu untuk membantu proses dalam melakukan pemrograman

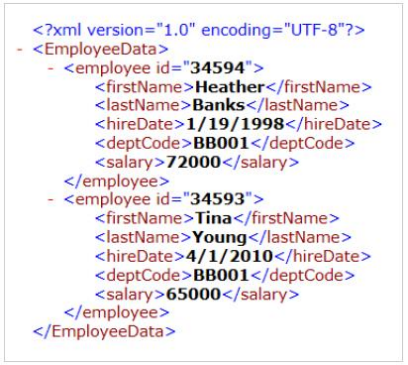
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os

os.chdir("/content/drive/MyDrive/{lokasi folder}
```

Jalankan program di atas agar Google Collab terhubung dengan folder tempat kita ingin bekerja

FILE FORMAT

Beberapa format file yang sering digunakan dalam analisis data

1. Excel
2. CSV (Comma separated value)
3. JSON (Javascript Object Notation)
4. XML

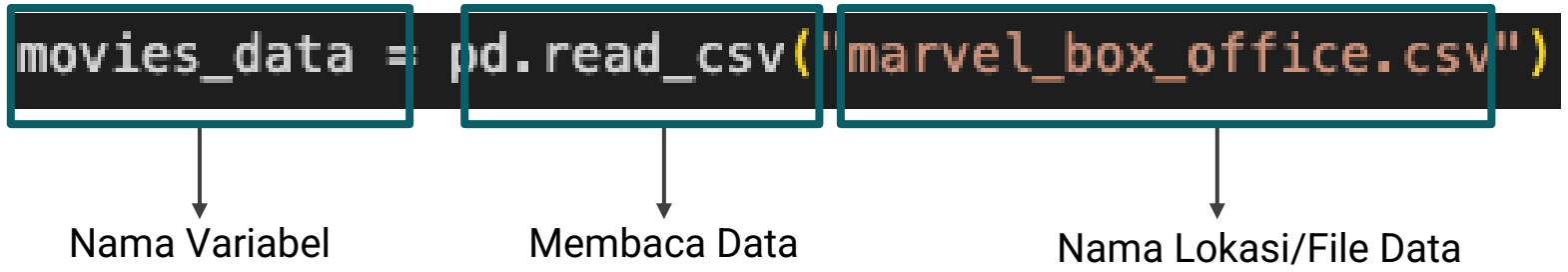


XML

CSV

JSON

IMPORT DATA


movies_data = pd.read_csv("marvel_box_office.csv")
Nama Variabel
Membaca Data
Nama Lokasi/File Data

DATA INFO

Informasi yang bisa didapatkan:

1. Jumlah baris dan kolom
2. Nama kolom
3. Kelengkapan data pada setiap kolom
4. Tipe data setiap kolom
5. Penggunaan ruang penyimpanan

```
movies_data.info() #
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66 entries, 0 to 65
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Movie                           66 non-null     object
 1   Release Date                    66 non-null     object
 2   Release Month                   66 non-null     object
 3   Release Day                     66 non-null     int64
 4   Release Year                    66 non-null     int64
 5   Ownership                       66 non-null     object
 6   Domestic Box Office             66 non-null     int64
 7   Inflation Adjusted Domestic     66 non-null     int64
 8   International Box Office        66 non-null     int64
 9   Inflation Adjusted International  66 non-null   float64
 10  Worldwide Box Office            66 non-null     int64
 11  Inflation Adjusted Worldwide    66 non-null     float64
 12  Opening Weekend                 66 non-null     int64
 13  Inflation Adjusted Opening Weekend  66 non-null  int64
 14  Budget                          66 non-null     int64
 15  Inflation Adjusted Budget       66 non-null     int64
 16  IMDb Score                      66 non-null     float64
 17  Meta Score                      66 non-null     int64
 18  Tomatometer                     66 non-null     int64
 19  Rotten Tomato Audience Score    66 non-null     int64
 20  Run Time In Minutes             66 non-null     int64
 21  Phase                           33 non-null     object
 22  Director                        66 non-null     object
dtypes: float64(3), int64(14), object(6)
memory usage: 12.0+ KB
```

MENAMPILKAN DATA

```
movies_data.head(10)
```

	Movie	Release Date	Release Month	Release Day	Release Year	Ownership	Domestic Box Office	Inflation Adjusted Domestic
0	Iron Man	5/2/2008	May	2	2008	Marvel Studios	318604126	467231126
1	The Incredible Hulk	6/13/2008	June	13	2008	Marvel Studios	134806913	197704288
2	Iron Man 2	5/7/2010	May	7	2010	Marvel Studios	312433331	416973763
3	Thor	5/6/2011	May	6	2011	Marvel Studios	181030624	240384926

- Menampilkan *n* data pertama menggunakan method **.head(jumlah data)**
- Menampilkan *n* data terakhir menggunakan method **.tail(jumlah data)**
- Menampilkan *n* data secara acak menggunakan method **.sample(jumlah data)**

```
movies_data.sample(10)
```

	Movie	Release Date	Release Month	Release Day	Release Year	Ownership	Domestic Box Office	Inflation Adjusted Domestic
18	Avengers Infinity War	4/27/2018	April	27	2018	Marvel Studios	678815482	784624259
60	The Punisher	4/16/2004	April	16	2004	Lionsgate Films	33664370	57083056
27	Doctor Strange in the Multiverse of Madness	5/6/2022	May	6	2022	Marvel Studios	411331607	411331607
4	Captain America: The First Avenger	7/22/2011	July	22	2011	Marvel Studios	176654505	234574020

```
movies_data.tail(10)
```

	Movie	Release Date	Release Month	Release Day	Release Year	Ownership	Domestic Box Office	Inflation Adjusted Domestic
56	Fantastic Four: Rise of the Silver Surfer	6/15/2007	June	15	2007	20th Century Fox	131921738	201909286
57	Fantastic Four (2015)	8/7/2015	August	7	2015	20th Century Fox	56117548	70097010
58	Elektra	1/14/2005	January	14	2005	20th Century Fox	24409722	40098956
						20th		

DATA INDEX

.index
Menampilkan nama index atau nama baris yang ada pada data, secara *default*, ia akan menggunakan angka dimulai dari 0 - jumlah data - 1

.set_index()
Membuat nama indeks berdasarkan kolom tertentu, akan tetapi nama ini harus *unique*

.reset_index()
Melakukan reset dari index yang ada sekarang, bisa menggunakan parameter *drop = True* untuk menghapus

```
[29] movies_data.loc[2,"Movie"] #misal ingin liat judul di index ke 2

'Iron Man 2'

[30] movies_data.loc[:,"Movie"]#misal ingin liat semua judul

0                                  Iron Man
1                        The Incredible Hulk
2                                Iron Man 2
3                                      Thor
4          Captain America: The First Avenger
                          ...
61                       Punisher: Warzone
62                                   Blade
63                                Blade II
64                           Blade: Trinity
65                                 Morbius
Name: Movie, Length: 66, dtype: object
```