Laporan Kualitas Data – Modul 3

Dataset: Netflix TV Shows and Movies (titles.csv)
Penulis: Mochammad Delvin Farhan Akbar

1. Profiling Data Awal

Dataset *titles.csv* berisi daftar film dan TV show di Netflix dengan total 5.850 baris dan 15 kolom.
Kolom penting antara lain: `title`, `type`, `release_year`, `runtime`, `genres`, `production_countries`, `imdb_score`, dan `tmdb_score`.

Pemeriksaan awal dengan `df.info()` dan `df.describe()` menunjukkan:

- Beberapa kolom memiliki nilai kosong (missing values).
- Kolom `runtime` memiliki nilai minimum 0 menit, yang tidak valid untuk film/TV show.

```
import pandas as pd

df = pd.read_csv("titles.csv")

print("Shape dataset:", df.shape)
df.info()
df.describe(include="all").T.head(15)
```

```
Shape dataset: (5850, 15)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5850 entries, 0 to 5849
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    5850 non-null   object
 1   title                 5849 non-null   object
 2   type                  5850 non-null   object
 3   description           5832 non-null   object
 4   release_year          5850 non-null   int64
 5   age_certification     3231 non-null   object
 6   runtime               5850 non-null   int64
 7   genres                5850 non-null   object
 8   production_countries  5850 non-null   object
 9   seasons               2106 non-null   float64
 10  imdb_id               5447 non-null   object
 11  imdb_score            5368 non-null   float64
 12  imdb_votes            5352 non-null   float64
 13  tmdb_popularity       5759 non-null   float64
 14  tmdb_score            5539 non-null   float64
dtypes: float64(5), int64(2), object(8)
memory usage: 685.7+ KB
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
id	5850	5850	ts271048	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
title	5849	5798	Connected	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
type	5850	2	MOVIE	3744	NaN	NaN	NaN	NaN	NaN	NaN	NaN
description	5832	5829	Five families struggle with the ups and downs ...	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
release_year	5850.0	NaN	NaN	NaN	2016.417094	6.937726	1945.0	2016.0	2018.0	2020.0	2022.0
age_certification	3231	11	TV-MA	883	NaN	NaN	NaN	NaN	NaN	NaN	NaN
runtime	5850.0	NaN	NaN	NaN	76.888889	39.002509	0.0	44.0	83.0	104.0	240.0
genres	5850	1726	['comedy']	484	NaN	NaN	NaN	NaN	NaN	NaN	NaN
production_countries	5850	452	['US']	1959	NaN	NaN	NaN	NaN	NaN	NaN	NaN
seasons	2106.0	NaN	NaN	NaN	2.162868	2.689041	1.0	1.0	1.0	2.0	42.0
imdb_id	5447	5447	tt13711094	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
imdb_score	5368.0	NaN	NaN	NaN	6.510861	1.163826	1.5	5.8	6.6	7.3	9.6
imdb_votes	5352.0	NaN	NaN	NaN	23439.382474	95820.470909	5.0	516.75	2233.5	9494.0	2294231.0
tmdb_popularity	5759.0	NaN	NaN	NaN	22.637925	81.680263	0.009442	2.7285	6.821	16.59	2274.044

2. Menemukan Kesalahan Data

Dari pemeriksaan, ditemukan minimal tiga masalah utama:

1. **Missing Values**
   - `seasons`: 3.744 baris kosong.
   - `age_certification`: 2.619 baris kosong.
   - `imdb_score`: 482 kosong.
   - `imdb_votes`: 498 kosong.
   - `title`: 1 kosong.
2. **Outliers**
   - Kolom `runtime` memiliki nilai minimum 0 menit, tidak masuk akal untuk film/TV show.
3. **Tipe Data Salah**
   - `genres` dan `production_countries` disimpan dalam format string list, bukan array Python.
   - `seasons` bertipe float (contoh: 1.0, 2.0), padahal lebih logis integer.

```
# 4.1 Missing Values
print("Missing Values per Kolom:")
print(df.isnull().sum())

# 4.2 Outliers runtime
print("\nStatistik Runtime:")
print(df['runtime'].describe())

# 4.3 Duplikat data
print("\nJumlah duplikat:", df.duplicated().sum())

# 4.4 Contoh format genres & production_countries
print("\nContoh genres:", df['genres'].iloc[0])
print("Contoh production_countries:", df['production_countries'].iloc[0])
```

```
Missing Values per Kolom:
id                        0
title                     1
type                      0
description              18
release_year              0
age_certification      2619
runtime                   0
genres                    0
production_countries      0
seasons                3744
imdb_id                 403
imdb_score              482
imdb_votes              498
tmdb_popularity          91
tmdb_score              311
dtype: int64

Statistik Runtime:
count    5850.000000
mean       76.888889
std        39.002509
min         0.000000
25%        44.000000
50%        83.000000
75%       104.000000
max       240.000000
Name: runtime, dtype: float64

Jumlah duplikat: 0

Contoh genres: ['documentation']
Contoh production_countries: ['US']
```

3. Tindakan dan Justifikasi

Langkah perbaikan yang dilakukan:

- **Missing Values**
  - `seasons` kosong → diisi 0 (karena film tidak punya season).
  - `age_certification` kosong → diisi `"UNKNOWN"`.
  - `imdb_score` kosong → diganti dengan median skor IMDb.
- **Outliers**

- Menghapus baris dengan `runtime = 0`.

- **Tipe Data Salah**
  - Parsing `genres` dan `production_countries` menjadi list.
  - Mengubah `seasons` menjadi integer.

```
import ast

# Tangani missing values
df['seasons'] = df['seasons'].fillna(0).astype(int)
df['age_certification'] = df['age_certification'].fillna("UNKNOWN")
df['imdb_score'] = df['imdb_score'].fillna(df['imdb_score'].median())

# Tangani outliers runtime = 0
df = df[df['runtime'] > 0]

# Parsing kolom genres dan production_countries
df['genres'] = df['genres'].apply(lambda x: ast.literal_eval(x) if pd.notnull(x) else [])
df['production_countries'] = df['production_countries'].apply(lambda x: ast.literal_eval(x) if pd.notnull(x) else [])

# Simpan dataset bersih
df.to_csv("titles_clean.csv", index=False)
print("Dataset bersih disimpan sebagai titles_clean.csv")
```

```
Dataset bersih disimpan sebagai titles_clean.csv
```

4. Kesimpulan

Berdasarkan hasil pemeriksaan, dataset *titles.csv* memiliki beberapa masalah kualitas data seperti nilai kosong (missing values), outliers, dan tipe data yang tidak konsisten. Setelah dilakukan proses pembersihan, nilai kosong diisi dengan pendekatan yang sesuai, baris dengan runtime tidak valid dihapus, serta kolom yang semula dalam format string list berhasil diparsing menjadi list Python. Dengan langkah-langkah tersebut, dataset menjadi lebih konsisten, rapi, dan siap digunakan untuk analisis lebih lanjut. Dataset hasil akhir kemudian disimpan dalam file `titles_clean.csv` (https://drive.google.com/file/d/1e4PSHi7DtNclWLfbY1naxxMCo37ElJiY/view?usp=drive_link) agar dapat dimanfaatkan pada tahap berikutnya.