

Tugas Data Cleaning

Berikut dokumentasi/screenshot langkah-langkah dalam proses data Data Cleaning yang telah saya kerjakan: menggunakan bahasa pemrograman python dengan tools google colab

Source code projek di akun github saya: https://github.com/maxwellmassie/Tugas-Data-Cleaning-Sains-Data/blob/main/Tugas_Data_Cleaning.ipynb

▼ Tugas Data Science...Data Cleaning

Nama: Gold Stein Maxwell Massie
Kelas: 2IA20
Npm: 50422625

Berikut Tahapan Proses data cleaning

1. **Define:** pada tahap ini, kita akan membuat rancangan tahapan serta metode pembersihan data berdasarkan masalah yang ditemukan dalam proses assessing data. Hal ini dapat dijadikan sebagai dokumentasi untuk memastikan orang lain memahami setiap tahapan dalam pembersihan data yang akan kita lakukan.
2. **Code:** setelah membuat rancangan pembersihan data, tahap selanjutnya ialah mengonversi hal tersebut menjadi sebuah kode program yang dapat dijalankan.
3. **Test:** setelah menjalankan kode program untuk membersihkan data, kita perlu memeriksa kembali data yang telah dibersihkan tersebut. Hal ini untuk memastikan proses pembersihan data dilakukan sesuai ekspektasi kita.

Untuk mengaplikasikan Tahapan poses data cleaning diatas, berikut metode yang akan digunakan

- Teknik untuk Mengatasi Missing Value dengan cara Dropping, Imputation, Interpolation
- Teknik untuk Mengatasi Outlier dengan cara drop, Imputation
- Teknik untuk Mengatasi Duplicate Data dengan cara drop_duplicates()

Sebelum membersihkan data(data cleaning) kita perlu masuk ke tahap assesing data atau tahap analysis data dimana masalah umum dapat di temukan seperti Missing value, Invalid value, Duplicate data, Inaccurate value, Inconsistent value, Outlier

Berikut penerapan Data Cleaning sesuai dataset yang tersedia yaitu

▼ movie_sample_dataset.csv

Mengimport library yang dibutuhkan

```
✓ 0.0 [1] #import library
import pandas as pd
```

Membuat dataframe baru dengan memuat dataset

```
✓ 0.0 movie_df = pd.read_csv("movie_sample_dataset.csv")
movie_df.head() #menampilkan dataset
```

	color	director_name	duration	gross	genres	movie_title	title_year	language	country	budget	imdb_score	actors	movie_facebook_likes
0	Color	Martin Scorsese	240	116866727.0	Biography Comedy Crime Drama	The Wolf of Wall Street	2013	English	USA	100000000.0	8.2	Leonardo DiCaprio,Matthew McConaughey,Jon Favreau	138000
1	Color	Shane Black	195	408992272.0	Action Adventure Sci-Fi	Iron Man 3	2013	English	USA	200000000.0	7.2	Robert Downey Jr.,Jon Favreau,Don Cheadle	95000
2	color	Quentin Tarantino	187	54116191.0	Crime Drama Mystery Thriller Western	The Hateful Eight	2015	English	USA	44000000.0	7.9	Craig Stark,Jennifer Jason Leigh,Zoë Bell	114000
3	Color	Kenneth Lonergan	186	46495.0	Drama	Margaret	2011	English	usa	14000000.0	6.5	Matt Damon,Kieran Culkin,John Gallagher Jr.	0
4	Color	Peter Jackson	186	258355354.0	Adventure Fantasy	The Hobbit: The Desolation of Smaug	2013	English	USA	225000000.0	7.9	Aidan Turner,Adam Brown,James Nesbitt	83000

▼ Tahap Assesing Data

Menilai dataset untuk melihat informasi typedata dan jumlah data dari setiap kolom yang tidak bernilai Nan(Missing Value)

```
movie_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  --
 0   color                 88 non-null    object  
 1   director_name         88 non-null    object  
 2   duration              99 non-null    int64   
 3   gross                 91 non-null    float64  
 4   genres                98 non-null    object  
 5   movie_title           99 non-null    object  
 6   title_year            99 non-null    int64   
 7   language              99 non-null    object  
 8   country               99 non-null    object  
 9   budget               95 non-null    float64  
10   imdb_score            99 non-null    float64  
11   actors               99 non-null    object  
12   movie_facebook_likes  99 non-null    int64   
dtypes: float64(3), int64(3), object(7)
memory usage: 10.2+ KB
```

memeriksa dan menampilkan informasi terkait jumlah missing value pada dataset

```
movie_df.isna().sum()

color                 11
director_name         11
duration              0
gross                 8
genres                1
movie_title           0
title_year            0
language              0
country               0
budget               4
imdb_score            0
actors               0
movie_facebook_likes  0
dtype: int64
```

dari hasil pemeriksaan diatas, terdapat **11 missing value pada kolom color**, **11 missing value pada kolom director_name**, **8 missing value pada kolom gross**, **1 missing value pada kolom genres**, dan **4 missing value pada kolom budget**

selanjutnya, memeriksa duplikasi data dan menampilkan ringkasan parameter statistik (mean, median, dll.) dari kolom numerik pada sebuah DataFrame dengan method describe()

```
print("Jumlah duplikasi: ",movie_df.duplicated().sum())
movie_df.describe()
```

	duration	gross	title_year	budget	imdb_score	movie_facebook_likes
count	99.000000	9.100000e+01	99.000000	9.500000e+01	99.000000	99.000000
mean	155.494949	1.541914e+08	1976.444444	1.048570e+08	6.892929	66045.707071
std	72.797927	1.399503e+08	255.880601	7.703169e+07	1.925514	58108.860365
min	-50.000000	4.122900e+04	202.000000	1.735000e+04	-7.500000	0.000000
25%	138.500000	4.720632e+07	2012.000000	4.000000e+07	6.550000	25000.000000
50%	143.000000	1.156040e+08	2013.000000	8.000000e+07	7.200000	54000.000000
75%	155.000000	2.374894e+08	2014.000000	1.740000e+08	7.850000	85500.000000
max	650.000000	6.232795e+08	2016.000000	2.500000e+08	8.800000	349000.000000

dari hasil pemeriksaaan diatas, terdapat jumlah duplikasi data sebanyak 6 dan tidak terdapat outlier data

▼ Cleaning Data

Setelah melakukan analisis data dan diperoleh masalah pada data yang harus di bersihkan/diperbaiki

Mengatasi missing value

1. terdapat 11 missing value pada kolom color
2. terdapat 11 missing value pada kolom director_name
3. terdapat 8 missing value pada kolom gross.
4. terdapat 1 missing value pada kolom genres
5. terdapat 4 missing value pada kolom budget

Menghapus Duplikasi Data

1. Terdapat 6 duplikasi data

berikut perbaikan missing value dari kolom-kolom diatas dan penghapusan duplikasi data

1. Perbaikan 11 missing value pada kolom color

menampilkan 11 missing value pada kolom color
movie_df[movie_df.color.isna()]

	color	director_name	duration	gross	genres	movie_title	title_year	language	country	budget	imdb_score	actors	movie_facebook_likes
5	NaN	NaN	183	330249062.0	Action Adventure Sci-Fi	Batman v Superman: Dawn of Justice	202	English	USA	250000000.0	6.9	Henry Cavill,Lauren Cohan,Alan D. Purwin	197000
10	NaN	Tom Tykwer	172	27098580.0	Drama Sci-Fi	Cloud Atlas	2012	English	Germany	102000000.0	-7.5	Tom Hanks,Jim Sturgess,Jim Broadbent	124000
15	NaN	Richard Linklater	165	25359200.0	Drama	Boyhood	2014	English	USA	4000000.0	8.0	Ellar Coltrane,Lorelei Linklater,Libby Villari	92000
18	NaN	Christopher Nolan	164	448130642.0	Action Thriller	The Dark Knight Rises	2012	English	USA	250000000.0	8.5	Tom Hardy,Christian Bale,Joseph Gordon-Levitt	164000
56	NaN	NaN	143	NaN	Drama Horror Thriller	The Ridges	2011	English	USA	17350.0	3.0	Robbie Barnes,Alana Kaniewski,Brandon Landers	33
65	NaN	Oliver Stone	141	47307550.0	Crime Drama Thriller	Savages	2012	English	USA	45000000.0	6.5	Demian Bichir,Shea Whigham,Gary Stretch	28000
74	NaN	Terrence Malick	139	13303319.0	Drama Fantasy	The Tree of Life	2011	English	USA	32000000.0	6.7	Brad Pitt,Tye Sheridan,Fiona Shaw	39000
76	NaN	Robert Zemeckis	138	93749203.0	Drama Thriller	Flight	2012	English	USA	31000000.0	7.3	Denzel Washington,Bruce Greenwood,Nadine Velazquez	64000
80	NaN	James Mangold	138	132550960.0	Action Adventure Sci-Fi Thriller	The Wolverine	2013	English	USA	120000000.0	6.7	Hugh Jackman,Tao Okamoto,Rila Fukushima	68000
83	NaN	Walter Salles	137	717753.0	Adventure Drama	On the Road	2012	English	France	25000000.0	6.1	Kristen Stewart,Viggo Mortensen,Kirsten Dunst	27000
87	NaN	Seth MacFarlane	136	42615685.0	Comedy Western	A Million Ways to Die in the West	2014	English	USA	40000000.0	6.1	Liam Neeson,Charlize Theron,Seth MacFarlane	24000

[7] movie_df.color.value_counts() #mengidentifikasi nilai yang dominan pada kolom color

color 86
color 1
Black and white 1
Name: color, dtype: int64

Berdasarkan hasil di atas, dapat diketahui bahwa nilai yang paling dominan dalam kolom gender ialah "Color ". Nilai inilah yang selanjutnya akan kita gunakan sebagai pengganti missing value. Proses penggantian ini dapat dilakukan menggunakan method fillna()seperti contoh berikut.

movie_df.color.fillna(value="Color", inplace=True)

2. Perbaikan 11 missing value pada kolom director_name

menampilkan 11 missing value pada kolom director_name
movie_df[movie_df.director_name.isna()]

	color	director_name	duration	gross	genres	movie_title	title_year	language	country	budget	imdb_score	actors	movie_facebook_likes
5	Color	NaN	183	330249062.0	Action Adventure Sci-Fi	Batman v Superman: Dawn of Justice	202	English	USA	250000000.0	6.9	Henry Cavill,Lauren Cohan,Alan D. Purwin	197000
24	Color	NaN	156	183635922.0	Adventure Drama Thriller Western	The Revenant	2015	English	USA	135000000.0	8.1	Leonardo DiCaprio,Tom Hardy,Lukas Haas	190000
32	Color	NaN	150	182204440.0	Biography Drama History War	Lincoln	2012	English	USA	65000000.0	7.4	Joseph Gordon-Levitt,Hal Holbrook,Bruce McGill	71000
41	Color	NaN	147	407197282.0	Action Adventure Sci-Fi	Captain America: Civil War	2016	English	USA	250000000.0	8.2	Robert Downey Jr.,Scarlett Johansson,Chris Evans	72000
56	Color	NaN	143	NaN	Drama Horror Thriller	The Ridges	2011	English	USA	17350.0	3.0	Robbie Barnes,Alana Kaniewski,Brandon Landers	33
59	Color	NaN	142	407999255.0	Adventure Drama Sci-Fi Thriller	The Hunger Games	2012	English	USA	78000000.0	7.3	Jennifer Lawrence,Josh Hutcherson,Anthony Reyn...	140000
71	Color	NaN	139	150832203.0	Adventure Mystery Sci-Fi	Divergent	2014	English	USA	85000000.0	6.7	Kate Winslet,Theo James,Mekhi Phifer	49000
75	Color	NaN	138	150117807.0	Crime Drama	American Hustle	2013	English	USA	40000000.0	7.3	Jennifer Lawrence,Christian Bale,Bradley Cooper	63000
82	Color	NaN	137	37304950.0	Biography Crime Drama	J. Edgar	2011	English	USA	35000000.0	6.6	Leonardo DiCaprio,Naomi Watts,Kaitlyn Dever	16000
84	Color	NaN	137	281666058.0	Adventure Sci-Fi	The Hunger Games: Mockingjay - Part 2	2015	English	USA	160000000.0	6.6	Jennifer Lawrence,Philip Seymour Hoffman,Josh ...	38000
91	Color	NaN	136	52474616.0	Drama	Wall Street: Money Never Sleeps	2010	English	USA	70000000.0	6.3	Frank Langella,Austin Pendleton,John Buffalo M...	13000

```
✓ [10] movie_df.director_name.value_counts() #mengidentifikasi nilai yang ada pada kolom director_name
```

```
0.0
Ridley Scott      4
Timur Bekmambetov 3
Peter Jackson    3
Joss Whedon       3
Sam Mendes        3
..
Adam McKay        1
Zack Snyder        1
Edward Hall        1
Kenneth Lonergan   1
Clint Eastwood     1
Name: director_name, Length: 63, dtype: int64
```

Berdasarkan hasil di atas, kita akan menginput missing value(NaN) dengan "No Name". Nilai inilah yang selanjutnya akan kita gunakan sebagai pengganti missing value. Proses penggantian ini dapat dilakukan menggunakan method fillna()seperti contoh berikut.

```
✓ [11] movie_df.director_name.fillna(value="No Name", inplace=True)
```

3. Perbaiki 8 missing value pada kolom gross.

```
✓ [12] # menampilkan 8 missing value pada kolom gross
movie_df[movie_df.gross.isna()]
```

	color	director_name	duration	gross	genres	movie_title	title_year	language	country	budget	imdb_score	actors	movie_facebook_likes
7	Color	Edward Hall	180	NaN	Drama/Romance	Restless	2012	English	UK	NaN	7.2	Rufus Sewell,Hayley Atwell,Charlotte Rampling	434
27	Color	Gnana Rajasekaran	153	NaN	Biography/Drama/History	Ramanujan	2014	English	India	NaN	7.0	Bharathi,Michael Lieber,Kevin McGowan	58
37	Color	Jay Oliva	148	NaN	Action/Animation/Crime/Sci-Fi/Thriller	Batman: The Dark Knight Returns, Part 2	2013	English	USA	3500000.0	8.4	Michael Emerson,Mark Valley,Grey Griffin	5000
56	Color	No Name	143	NaN	Drama/Horror/Thriller	The Ridges	2011	English	USA	17350.0	3.0	Robbie Barnes,Alana Kaniewski,Brandon Landers	33
61	Color	Timur Bekmambetov	141	NaN	Adventure/Drama/History	Ben-Hur	2016	English	USA	100000000.0	6.1	Morgan Freeman,Ayelet Zurer,Moises Arias	0
62	Color	Timur Bekmambetov	141	NaN	Adventure/Drama/History	Ben-Hur	2016	English	USA	100000000.0	6.0	Morgan Freeman,Ayelet Zurer,Moises Arias	0
63	Color	Timur Bekmambetov	141	NaN	Adventure/Drama/History	Ben-Hur	2016	English	USA	100000000.0	6.1	Morgan Freeman,Ayelet Zurer,Moises Arias	0
92	Color	Sadyk Sher-Niyaz	135	NaN	Action/Biography/Drama/History	Queen of the Mountains	2014	English	Kyrgyzstan	1400000.0	8.7	Elina Abai Kyzy,Aziz Muradilayev,Irliran Abdul...	0

```
✓ [13] movie_df.gross.describe() #mengidentifikasi nilai rata-rata/mean pada kolom gross
```

```
0.0
count    9.100000e+01
mean     1.541914e+08
std       1.399503e+08
min       4.122900e+04
25%       4.720632e+07
50%       1.156040e+08
75%       2.374894e+08
max       6.232795e+08
Name: gross, dtype: float64
```

Berdasarkan hasil di atas, kita akan menginput missing value(NaN) dengan nilai rata-rata/mean dari gross yaitu "154191431.2747253". Nilai inilah yang selanjutnya akan kita gunakan sebagai pengganti missing value. Proses penggantian ini dapat dilakukan menggunakan method fillna()seperti contoh berikut.

```
✓ [14] movie_df.gross.fillna(value="154191431.2747253", inplace=True)
```

4. Perbaiki 1 missing value pada kolom genres

```
✓ [15] # menampilkan 1 missing value pada kolom genres
movie_df[movie_df.genres.isna()]
```

	color	director_name	duration	gross	genres	movie_title	title_year	language	country	budget	imdb_score	actors	movie_facebook_likes
12	Color	Christopher Spencer	170	59696176.0	NaN	Son of God	2014	English	USA	22000000.0	5.6	Roma Downey,Amber Rose Revah,Darwin Shaw	15000

Berdasarkan hasil di atas, kita akan menginput missing value(NaN) dengan "No Input". Nilai inilah yang selanjutnya akan kita gunakan sebagai pengganti missing value. Proses penggantian ini dapat dilakukan menggunakan method fillna()seperti contoh berikut.

```
✓ [16] movie_df.genres.fillna(value="No Input", inplace=True)
```

5. Perbaikan 4 missing value pada kolom budget

```
[17] # menampilkan 4 missing value pada kolom budget
movie_df[movie_df.budget.isna()]
```

	color	director_name	duration	gross	genres	movie_title	title_year	language	country	budget	imdb_score	actors	movie_facebook_likes
7	Color	Edward Hall	180	154191431.2747253	Drama Romance	Restless	2012	English	UK	NaN	7.2	Rufus Sewell, Hayley Atwell, Charlotte Rampling	434
27	Color	Gnana Rajasekaran	153	154191431.2747253	Biography Drama History	Ramanujan	2014	English	India	NaN	7.0	Mani Bharathi, Michael Lieber, Kevin McGowan	58
33	Color	Mike Leigh	150	3958500.0	Biography Drama History	Mr. Turner	2014	English	UK	NaN	6.8	Lesley Manville, Ruth Sheen, Karl Johnson	0
95	Color	Richard J. Lewis	134	7501404.0	Comedy Drama	Barney's Version	2010	English	Canada	NaN	7.3	Mark Addy, Atom Egoyan, Paul Gross	0

```
[18] movie_df.budget.describe() #mengidentifikasi nilai rata-rata pada kolom budget
```

```
count    9.500000e+01
mean     1.048570e+08
std       7.703169e+07
min      1.735000e+04
25%      4.000000e+07
50%      8.000000e+07
75%     1.740000e+08
max      2.500000e+08
Name: budget, dtype: float64
```

Berdasarkan hasil di atas, kita akan menginput missing value(NaN) dengan nilai rata-rata/mean dari gross yaitu "104857024.73684211". Nilai inilah yang selanjutnya akan kita gunakan sebagai pengganti missing value. Proses penggantian ini dapat dilakukan menggunakan method fillna()seperti contoh berikut.

```
[19] movie_df.budget.fillna(value="104857024.73684211", inplace=True)
```

Setelah kita memperbaiki missing value dari setiap kolom, kita akan memeriksa/testing kembali apakah perbaikan missing value berhasil pada semua kolom yang memiliki missing value.

```
[20] movie_df.isna().sum()
```

```
color                0
director_name        0
duration             0
gross                0
genres               0
movie_title          0
title_year           0
language             0
country              0
budget              0
imdb_score           0
actors               0
movie_facebook_likes 0
dtype: int64
```

dari hasil diatas, dapat disimpulkan bahwa proses perbaikan missing value telah berhasil sempurna karena seluruh kolom tidak memiliki missing value (berjumlah 0)

Menghapus Duplikasi Data

Berdasarkan hasil assending data/analysis data terdapat 6 duplikasi data dan berikut pembersihan duplikasi data

```
[21] movie_df.drop_duplicates(inplace=True)
```

memeriksa/testing kembali apakah penghapusan data duplicate telah berhasil

```
[22] print("Jumlah duplikasi: ", movie_df.duplicated().sum())
```

```
Jumlah duplikasi: 0
```

dari hasil diatas, proses penghapusan data duplicate berhasil karena jumlah data duplicate adalah 0

▼ export/menyimpan data yg telah dibersihkan dan menampilkan dataset yang telah dibersihkan

- Sebelum mengexport/menyimpan data pastikan data cleaning telah berjalan sesuai metode


```
movie_df.to_csv("cleaned_movie_sample_dataset.csv", index=False)
```

+ Kode

+ Teks

0 d selesai pada 22.02

Hasil export data yang telah clean dengan nama cleaned_movie_sample_dataset.csv

 cleaned_movie_sample_dataset.csv