DIGITAL SKILL FAIR 41.0 | FACULTY OF DATA

Exploratory Data Analysis

LIFE EXPECTANCY DATASET

MONIKA | SOFTWARE ENGINEERING TECHNOLOGY

DATA SCIENCE & DATA ANALYST

Introducing About Me

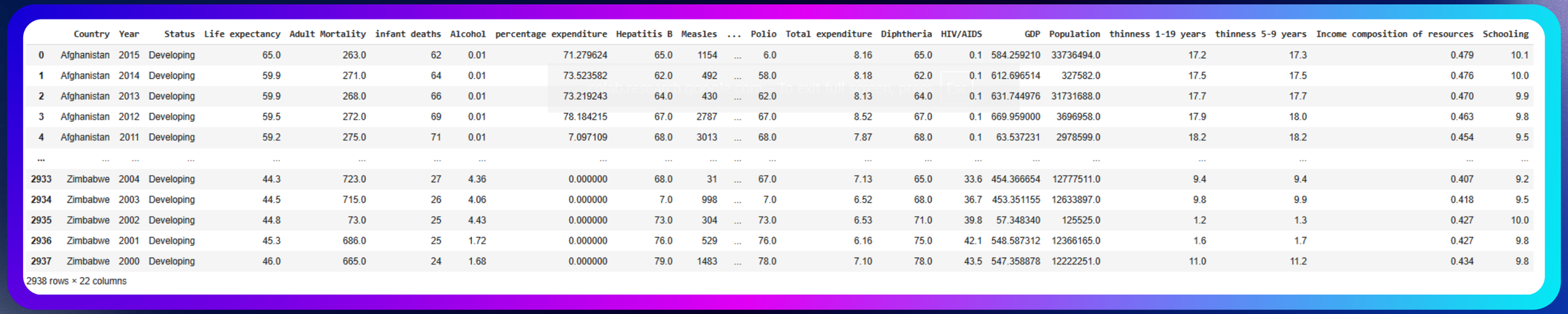
Halo, saya Monika, mahasiswa semester 4 Program Studi Teknologi Rekayasa Perangkat Lunak di Politeknik Manufaktur Negeri Bangka Belitung. Saya memiliki ketertarikan dalam bidang Data Science dan Data Analyst, meskipun sebelumnya saya telah mengeksplorasi UI/UX Design, Web Development, Artificial Intelligence, dan Machine Learning. Saat ini, saya mulai mendalami dasar-dasar Data Science untuk memperluas wawasan dan mempertimbangkan peluang karier di bidang tersebut.

Definition of Exploratory Data Analysis

Exploratory Data Analysis (EDA) merupakan proses awal dalam analisis data yang bertujuan untuk memahami struktur, pola, dan karakteristik data sebelum dilakukan pemodelan atau pengambilan keputusan lebih lanjut. Proses ini mencakup pemeriksaan ukuran dan struktur dataset, penanganan nilai yang hilang, identifikasi outlier, analisis distribusi variabel, dan pemahaman hubungan antar variabel. EDA juga dapat melibatkan visualisasi data untuk mempermudah interpretasi. Tujuan utamanya adalah memastikan kualitas dan kesiapan data untuk analisis lanjutan.

Case Study on Life Expectancy Analysis

Harapan hidup merupakan indikator utama untuk menilai kualitas kesehatan dan kesejahteraan suatu negara, yang dipengaruhi oleh berbagai faktor kesehatan dan sosial-ekonomi. Studi ini menggunakan dataset Life Expectancy Analysis dari Kaggle (https://www.kaggle.com/datasets/nailasrivastava/life-expectancy-analysis/data) yang memuat indikator-indikator yang memengaruhi harapan hidup di berbagai negara pada tahun 2000–2015. Sesuai dengan tahapan Exploratory Data Analysis (EDA), dilakukan proses awal berupa eksplorasi dan pembersihan data, seperti membaca dataset, menampilkan ringkasan data, menangani nilai yang hilang, serta menghapus data duplikat guna memastikan kualitas data sebelum analisis lebih lanjut.



	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259210	33736494.0	17.2	17.3	0.479	10.1
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696514	327582.0	17.5	17.5	0.476	10.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744976	31731688.0	17.7	17.7	0.470	9.9
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959000	3696958.0	17.9	18.0	0.463	9.8
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537231	2978599.0	18.2	18.2	0.454	9.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.000000	68.0	31	...	67.0	7.13	65.0	33.6	454.366654	12777511.0	9.4	9.4	0.407	9.2
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.000000	7.0	998	...	7.0	6.52	68.0	36.7	453.351155	12633897.0	9.8	9.9	0.418	9.5
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.000000	73.0	304	...	73.0	6.53	71.0	39.8	57.348340	125525.0	1.2	1.3	0.427	10.0
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.000000	76.0	529	...	76.0	6.16	75.0	42.1	548.587312	12366165.0	1.6	1.7	0.427	9.8
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.000000	79.0	1483	...	78.0	7.10	78.0	43.5	547.358878	12222251.0	11.0	11.2	0.434	9.8

2938 rows × 22 columns

Tahap Pertama

Membaca Dataset

Tahap pertama dalam proses EDA adalah membaca dataset. Berdasarkan hasil pembacaan, dataset Life Expectancy Analysis memiliki ukuran sebanyak 2938 baris dan 22 kolom, yang berarti terdapat 2938 entri data dengan 22 fitur. Informasi ini menjadi acuan awal dalam mengevaluasi jumlah data yang hilang (missing values) pada setiap kolom.

Tahap Kedua

Ringkasan Data

Setelah membaca dataset, ditampilkan informasi lebih rinci mengenai struktur dan isi data. Dari hasil tersebut, dapat diketahui bahwa beberapa fitur memiliki data hilang, yang ditandai dengan jumlah nilai non-null lebih sedikit dari total baris data, yaitu 2938. Sebagian besar fitur bertipe numerik, baik dalam bentuk float maupun integer. Sementara itu, hanya terdapat dua fitur yang bertipe data object, yang mengindikasikan bahwa keduanya termasuk dalam kategori data kategorik.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          2938 non-null   object 
 1   Year                             2938 non-null   int64  
 2   Status                           2938 non-null   object 
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64  
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64  
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64  
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years             2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

Life expectancy	10
Adult Mortality	10
Alcohol	194
Hepatitis B	553
BMI	34
Polio	19
Total expenditure	226
Diphtheria	19
GDP	448
Population	652
thinness 1-19 years	34
thinness 5-9 years	34
Income composition of resources	167
Schooling	163
dtype: int64	

Tahap Ketiga

Cek Jumlah Missing Value

Berdasarkan informasi pada tabel di samping, terdapat 14 fitur yang memiliki nilai hilang (missing value), antara lain **Life expectancy**, **Adult Mortality**, **Alcohol**, **Hepatitis B**, **BMI**, **Polio**, **Total expenditure**, **Diphtheria**, **GDP**, **Population**, **thinness 1-19 years**, **thinness 5-9 years**, **Income composition of resources**, dan **Schooling**. Jumlah missing value pada masing-masing fitur bervariasi, sehingga diperlukan visualisasi atau analisis menggunakan kode Python untuk melihat distribusi dan persentasenya secara lebih jelas.

Tahap Ketiga

Cek Persentase Missing Value

Berdasarkan informasi pada tabel di samping, fitur **Population** memiliki persentase missing value tertinggi, yaitu sebesar 22%. Meskipun demikian, seluruh persentase nilai yang hilang masih berada jauh di bawah ambang batas 90%, sehingga penanganan missing value dapat dilakukan dengan metode imputasi atau pengisian data, bukan dengan menghapus fitur terkait.

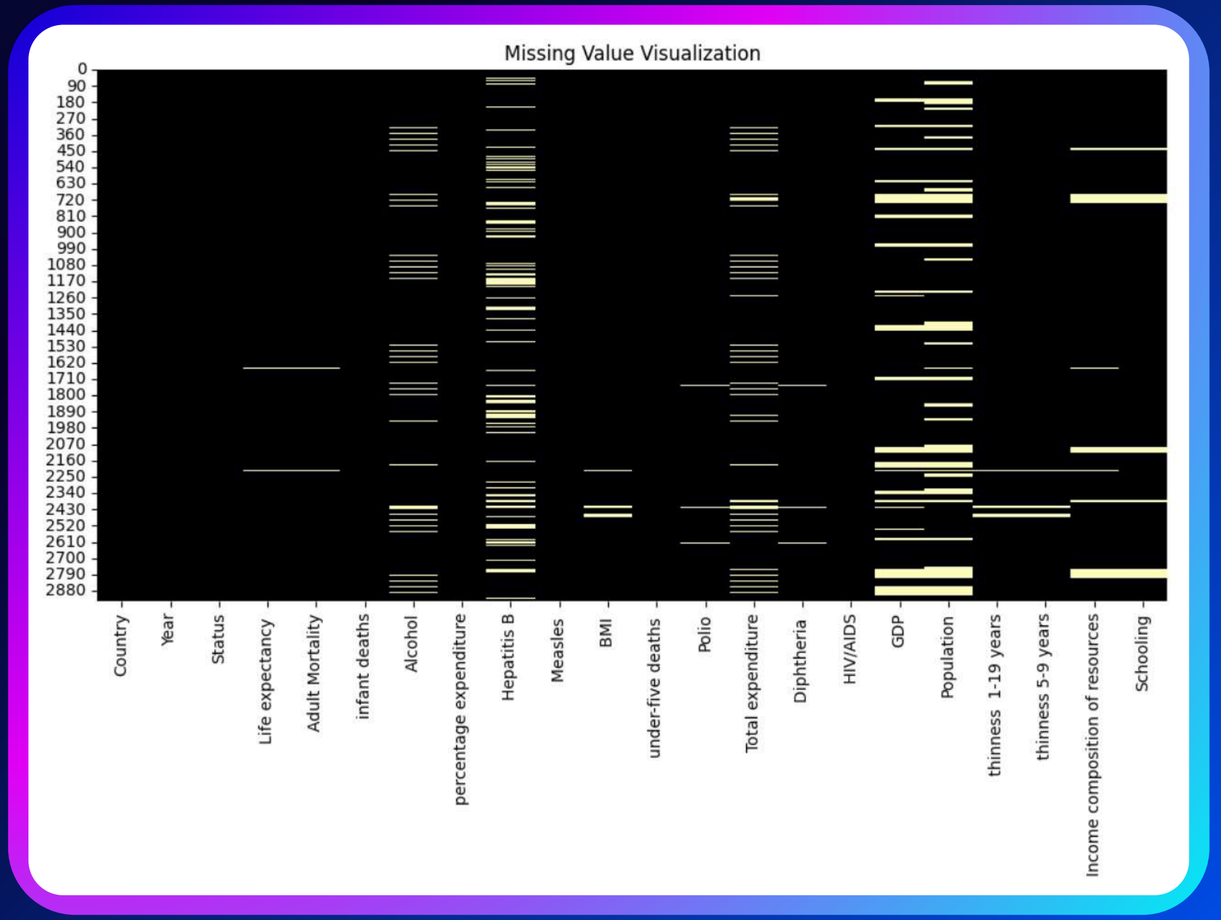
Life expectancy	0.340368
Adult Mortality	0.340368
Alcohol	6.603131
Hepatitis B	18.822328
BMI	1.157250
Polio	0.646698
Total expenditure	7.692308
Diphtheria	0.646698
GDP	15.248468
Population	22.191967
thinness 1-19 years	1.157250
thinness 5-9 years	1.157250
Income composition of resources	5.684139
Schooling	5.547992

dtype: float64

Tahap Ketiga

Visualisasi Missing Value

Berdasarkan hasil visualisasi di samping, dapat dilihat bahwa distribusi missing value tersebar di beberapa fitur. Beberapa fitur menunjukkan jumlah missing value yang cukup signifikan, seperti **Population**, **GDP**, dan **Hepatitis B**, yang ditandai dengan dominasi warna kuning pada kolom-kolom tersebut. Hal ini konsisten dengan perhitungan sebelumnya, di mana ketiga fitur tersebut memiliki persentase missing value tertinggi dibandingkan fitur lainnya.



Tahap Ketiga

Statistical Summary Data Numerik

Berdasarkan tabel di bawah ini, dapat dilihat ringkasan statistik (statistical summary) untuk data numerik yang mencakup nilai **count**, **mean**, **standard deviation (std)**, **min**, **25%**, **50% (median)**, **75%**, dan **max**. Informasi nilai mean dan median pada setiap fitur ini dapat digunakan untuk mengidentifikasi jenis distribusi data, apakah termasuk distribusi simetris (normal) atau distribusi skewed (tidak normal).

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	2938.000000	2919.000000	2712.00000	2919.000000	2938.000000	2490.000000	2.286000e+03	2904.000000	2904.000000	2771.000000	2775.000000
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240	38.321247	42.035739	82.550188	5.93819	82.324084	1.742103	7483.158469	1.275338e+07	4.839704	4.870317	0.627551	11.992793
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489	20.044034	160.445548	23.428046	2.49832	23.716912	5.077785	14270.169342	6.101210e+07	4.420195	4.508882	0.210904	3.358920
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.000000	3.000000	0.37000	2.000000	0.100000	1.681350	3.400000e+01	0.100000	0.100000	0.000000	0.000000
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000	19.300000	0.000000	78.000000	4.26000	78.000000	0.100000	463.935626	1.957932e+05	1.600000	1.500000	0.493000	10.100000
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.000000	93.000000	5.75500	93.000000	0.100000	1766.947595	1.386542e+06	3.300000	3.300000	0.677000	12.300000
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000	56.200000	28.000000	97.000000	7.49250	97.000000	0.800000	5910.806335	7.420359e+06	7.200000	7.200000	0.779000	14.300000
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.000000	99.000000	17.60000	99.000000	50.600000	119172.741800	1.293859e+09	27.700000	28.600000	0.948000	20.700000

	Country	Status
count	2938	2938
unique	193	2
top	Afghanistan	Developing
freq	16	2426

Tahap Ketiga

Statistical Summary Data Kategorik

Berdasarkan tabel di samping, dapat dilihat ringkasan statistik (statistical summary) untuk data kategorik yang mencakup nilai **count**, **unique**, **top**, dan **freq**. Dalam dataset ini, hanya terdapat dua fitur bertipe data object, yaitu **Country** dan **Status**. Kedua fitur tersebut memiliki nilai count sebesar 2938, yang menunjukkan bahwa tidak terdapat missing value pada fitur-fitur kategorik ini.

Tahap Ketiga

Distribusi Simetris/Skewed

Untuk mengetahui apakah suatu fitur memiliki distribusi simetris (normal) atau skewed (tidak normal), digunakan pendekatan selisih antara mean dan median. Jika selisihnya kurang dari 10% dari nilai mean, maka distribusinya dianggap normal, sedangkan jika lebih dari 10%, maka tergolong skewed. Berdasarkan hasil analisis, terdapat empat fitur dengan distribusi normal, yaitu **Life expectancy**, **Total expenditure**, **Income composition of resources**, dan **Schooling**, sementara sepuluh fitur lainnya terdistribusi skewed.

	Kolom	Status Distribusi
0	Life expectancy	Normal
1	Adult Mortality	Skewed
2	Alcohol	Skewed
3	Hepatitis B	Skewed
4	BMI	Skewed
5	Polio	Skewed
6	Total expenditure	Normal
7	Diphtheria	Skewed
8	GDP	Skewed
9	Population	Skewed
10	thinness 1-19 years	Skewed
11	thinness 5-9 years	Skewed
12	Income composition of resources	Normal
13	Schooling	Normal

```python
# Membersihkan spasi di awal dan akhir nama kolom
data.columns = data.columns.str.strip()

# Mengganti spasi ganda di tengah menjadi spasi satu
data.columns = data.columns.str.replace(r'\s+', ' ', regex=True)

# Variabel symmetric_columns menampung nama kolom yang distribusi simetris (normal)
symmetric_columns = ['Life expectancy', 'Total expenditure', 'Income composition of resources', 'Schooling']

# Pengisian missing value dengan nilai mean untuk kolom dalam variabel symmetric_columns
for column in symmetric_columns:
    data[column] = data[column].fillna(data[column].mean())

# Variabel skewed_columns menampung nama kolom yang distribusi skewed (tidak normal)
skewed_columns = ['Adult Mortality', 'Alcohol', 'Hepatitis B', 'BMI', 'Polio', 'Diphtheria', 'GDP', 'Population', 'thinness 1-19 years', 'thinness 5-9 years']

# Pengisian missing value dengan nilai median untuk kolom dalam variabel skewed_columns
for column in skewed_columns:
    data[column] = data[column].fillna(data[column].median())
```

Tahap Keempat

Handling Missing Value

Setelah diketahui fitur mana yang terdistribusi normal dan skewed, dapat disimpulkan bahwa dalam proses handling missing value, fitur dengan distribusi normal diisi menggunakan nilai mean, sedangkan fitur dengan distribusi skewed diisi menggunakan nilai median. Dengan demikian, empat fitur yang terdistribusi normal akan diisi menggunakan mean, dan sepuluh fitur lainnya yang tidak terdistribusi normal akan diisi menggunakan median.

	0
Country	0
Year	0
Status	0
Life expectancy	0
Adult Mortality	0
infant deaths	0
Alcohol	0
percentage expenditure	0
Hepatitis B	0
Measles	0
BMI	0
under-five deaths	0
Polio	0
Total expenditure	0
Diphtheria	0
HIV/AIDS	0
GDP	0
Population	0
thinness 1-19 years	0
thinness 5-9 years	0
Income composition of resources	0
Schooling	0

dtype: int64

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          2938 non-null   object
 1   Year                             2938 non-null   int64
 2   Status                           2938 non-null   object
 3   Life expectancy                  2938 non-null   float64
 4   Adult Mortality                  2938 non-null   float64
 5   infant deaths                    2938 non-null   int64
 6   Alcohol                          2938 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2938 non-null   float64
 9   Measles                          2938 non-null   int64
 10  BMI                              2938 non-null   float64
 11  under-five deaths                2938 non-null   int64
 12  Polio                            2938 non-null   float64
 13  Total expenditure                2938 non-null   float64
 14  Diphtheria                       2938 non-null   float64
 15  HIV/AIDS                         2938 non-null   float64
 16  GDP                              2938 non-null   float64
 17  Population                       2938 non-null   float64
 18  thinness 1-19 years              2938 non-null   float64
 19  thinness 5-9 years               2938 non-null   float64
 20  Income composition of resources  2938 non-null   float64
 21  Schooling                        2938 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

Tahap Keempat

Cek Kembali Ringkasan Data

Setelah proses handling missing value dilakukan, langkah selanjutnya adalah mengecek kembali struktur dan isi dataset. Jika seluruh fitur tidak lagi memiliki missing value dan jumlah data non-null pada setiap fitur sama dengan jumlah baris dataset (2938), maka dapat disimpulkan bahwa proses handling missing value telah berhasil.

Tahap Kelima

Handling Duplicate

Setelah menangani missing values, tahap selanjutnya adalah handling duplicate, yaitu menghapus data yang terduplikasi dalam dataset. Oleh karena itu, perlu dilakukan pengecekan jumlah data duplikat terlebih dahulu. Karena hasil pengecekan menunjukkan bahwa tidak terdapat data duplikat, maka tahap ini tidak memerlukan tindakan lanjutan.

```
# Mengecek apakah ada duplikat di seluruh kolom
check_duplicate = data.duplicated().sum()

print(f"Jumlah data duplikat: {check_duplicate}")
```

```
Jumlah data duplikat: 0
```

Kesimpulan

Beberapa fitur seperti **Hepatitis B**, **BMI**, **GDP**, **Polio**, dan **Schooling** memiliki nilai yang hilang (missing value), yang menunjukkan adanya ketidaklengkapan atau inkonsistensi dalam pelaporan data dari beberapa negara.

Berdasarkan statistical summary, beberapa fitur seperti **Adult Mortality**, **Alcohol**, dan **GDP** memiliki distribusi yang tidak simetris (skewed), sehingga perlu diperhatikan dalam menentukan metode handling missing value.

Fitur dengan distribusi normal (simetris) diisi menggunakan nilai rata-rata (mean), sedangkan fitur yang terdistribusi tidak normal (skewed) diisi menggunakan nilai tengah (median) untuk menjaga konsistensi dan keakuratan data.

Setelah dilakukan pemeriksaan terhadap data duplikat, tidak ditemukan adanya baris yang terduplikasi, sehingga tidak diperlukan langkah penanganan tambahan terkait duplikasi untuk dataset tersebut.

Thank You

Let’s Connect on Social Media!

 @Monikahung_

 in/monikahung

 monikahung580@gmail.com

 monikahung