

# **FINAL PROJECT**

**SANBERCODE BASIC PYTHON FOR DATA SCIENCE #51**

**JANE TAMARA SETIADI**





# CONTENT

- 
- |    |                                |
|----|--------------------------------|
| 01 | BUSINESS/PROJECT UNDERSTANDING |
| 02 | GOALS AND OBJECTIVES           |
| 03 | PROJECT TIMELINE               |
| 04 | PRIORITY                       |
| 05 | ANALYTICAL PROCESS             |
| 06 | DECISION MAKING                |
| 07 | RECOMMENDATION                 |

# BUSINESS/PROJECT UNDERSTANDING



Mengategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan Pembangunan negara secara keseluruhan.



HELP International adalah LSM kemanusiaan internasional yang berkomitmen untuk memerangi kemiskinan dan menyediakan fasilitas dan bantuan dasar bagi Masyarakat di negara-negara terbelakang saat terjadi bencana.





# BUSINESS/PROJECT UNDERSTANDING



HELP Internasional telah berhasil mengumpulkan sekitar \$ 10 juta. Saat ini, CEO LSM perlu memutuskan bagaimana menggunakan uang ini secara strategis dan efektif. Jadi, CEO harus mengambil keputusan untuk memilih negara yang paling membutuhkan bantuan. Hal yang diharapkan dari proyek ini adalah saran berupa negara mana saja yang paling perlu menjadi fokus CEO.



# GOALS AND OBJECTIVES

## ANALISIS DATA

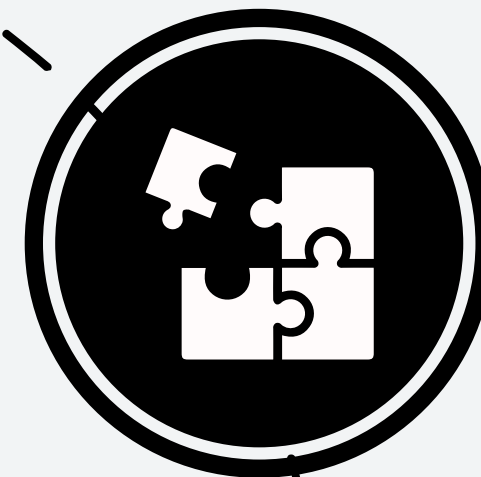
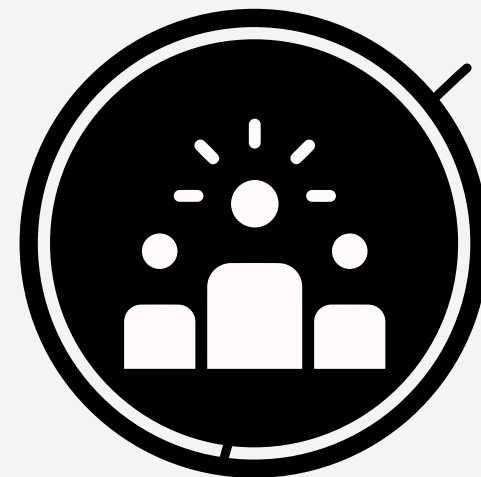
Melakukan penelitian terhadap data mengenai kondisi setiap negara berdasarkan karakter yang diukur dalam dataset

## DATA VISUALISASI

Untuk mempermudah pengambilan keputusan maka perlu dibuat beberapa visualisasi data

## REKOMENDASI

Output yang diharapkan dari proses analisis dan visualisasi data adalah informasi negara terpilih sebagai saran untuk CEO dalam mengambil keputusan





# PROJECT TIMELINE





01

---

**DATA CLEANING**

Melakukan identifikasi dan memperbaiki ketidakakuratan dalam dataset, untuk meningkatkan kualitas data

02

**EDA PART 1**

Melakukan eksplorasi data dengan melakukan *multivariate analysis*, untuk melihat hubungan antar variabel data dan menentukan variabel untuk clustering

03


**EDA PART 2**

Melakukan *univariate analysis* dan *bivariate analysis* untuk memahami karakteristik variabel terpilih, serta melihat hubungan antar keduanya

04

**CLUSTERING**

Melakukan data scaling, menentukan K-means dengan Elbow Method, dan melakukan pengelompokan data untuk pengambilan keputusan





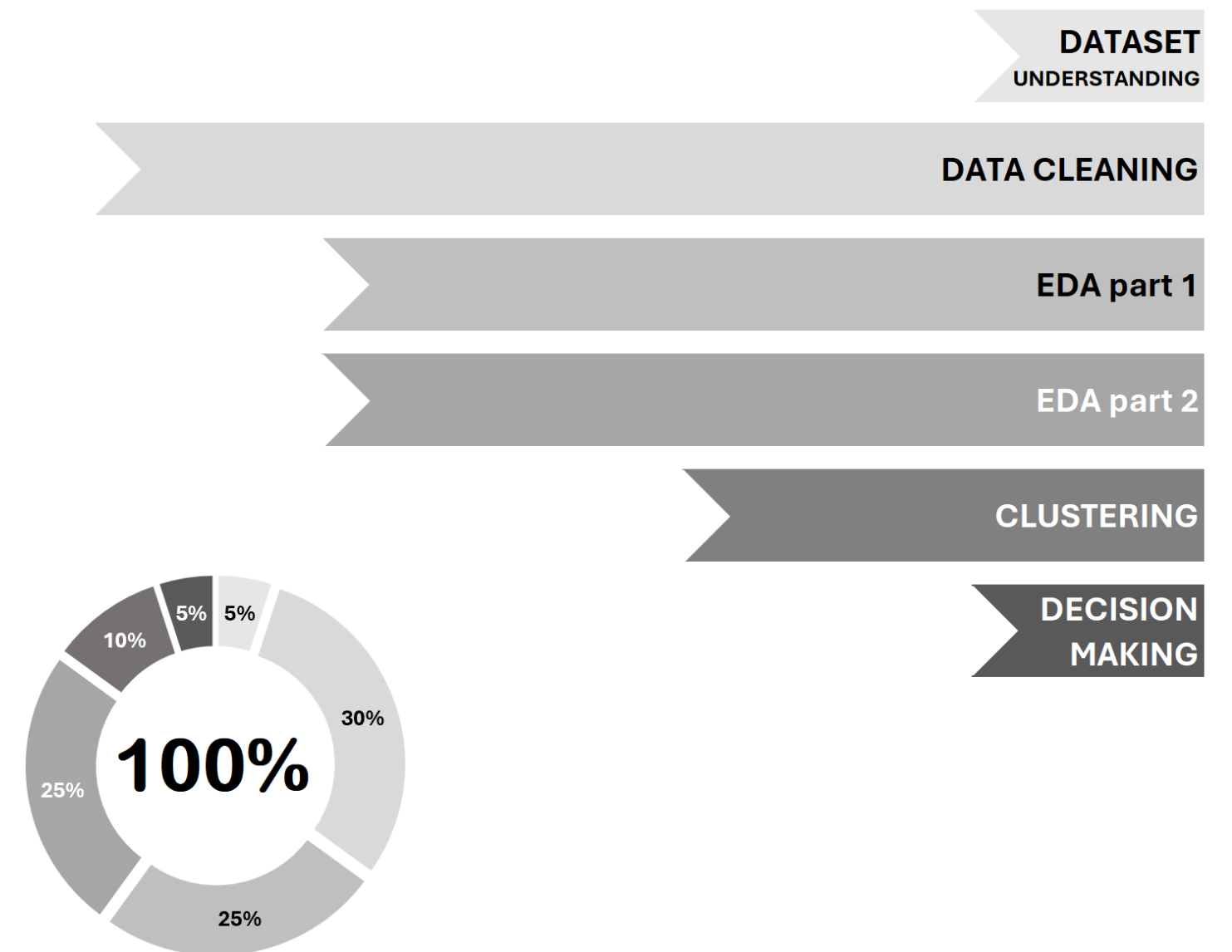
# PRIORITY

Berdasarkan tujuan dari proyek ini yaitu untuk memberi bantuan kepada negara yang bermasalah dalam hal financial, maka ditentukan bahwa faktor yang dapat mempengaruhi pembangunan suatu negara adalah faktor social ekonomi dan kesehatan. Dengan demikian penelitian ini difokuskan dengan variabel yang mengacu pada kedua factor tersebut.





# ANALYTICAL PROCESS



# DATA UNDERSTANDING

Source data : Data\_Negara\_HELP.csv

```
df=pd.read_csv('Data_Negara_HELP.csv')
df
```

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   Negara                167 non-null   object  
 1   Kematian_anak         167 non-null   float64 
 2   Ekspor                167 non-null   float64 
 3   Kesehatan             167 non-null   float64 
 4   Impor                 167 non-null   float64 
 5   Pendapatan            167 non-null   int64   
 6   Inflasi               167 non-null   float64 
 7   Harapan_hidup         167 non-null   float64 
 8   Jumlah_fertiliti      167 non-null   float64 
 9   GDPperkapita          167 non-null   int64   
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

Dataset terdiri dari 167 baris dalam 10 kolom variabel. Dari info yang diperoleh maka dapat dilihat bahwa 90% data berupa numerik (angka).

# DATA UNDERSTANDING

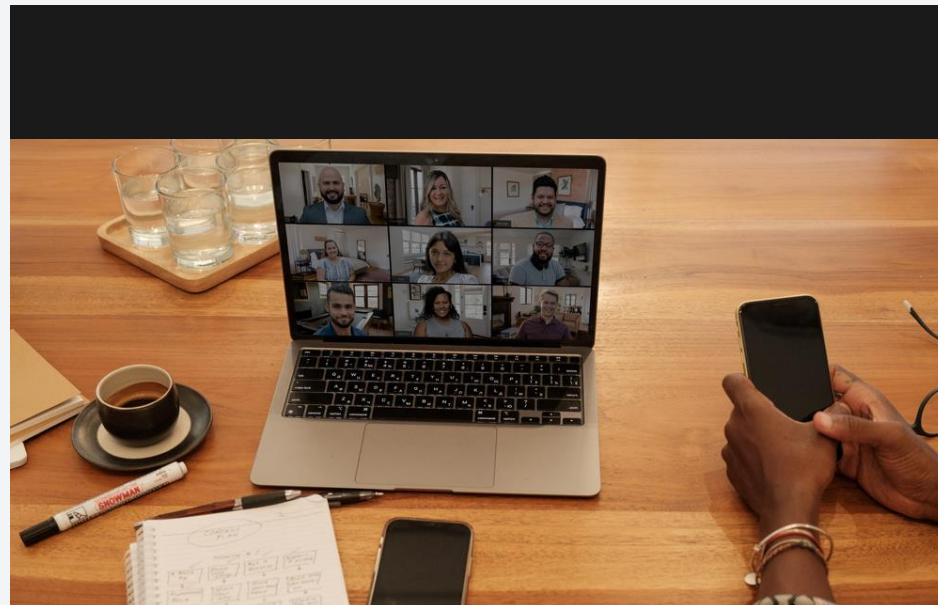
Untuk melihat karakteristik dari dataset maka perlu dilakukan perhitungan statistika deskriptif terhadap dataset :

```
df.describe()
```

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
<b>count</b>	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
<b>mean</b>	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
<b>std</b>	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
<b>min</b>	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
<b>25%</b>	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
<b>50%</b>	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
<b>75%</b>	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
<b>max</b>	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

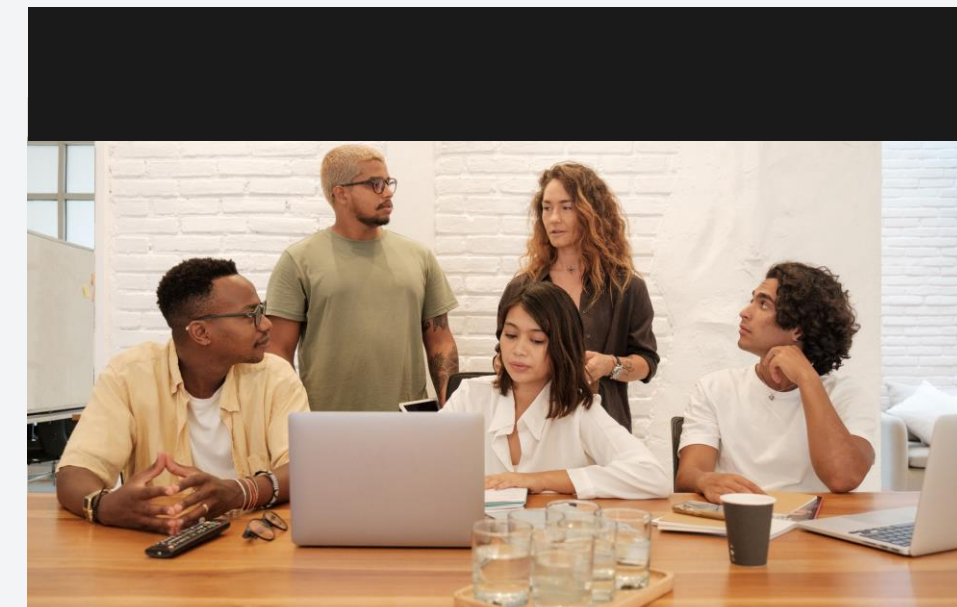


# THE DATA VARIABLES



- Negara : nama negara
- Kematian anak : kematian anak usia di bawah 5 tahun per 1000 kelahiran
- Ekspor : nilai ekspor barang dan jasa perkapita
- Kesehatan : total pengeluaran Kesehatan perkapita
- Impor : nilai impor barang dan jasa perkapita
- Pendapatan : penghasilan bersih per orang

- Inflasi : pengukuran tingkat pertumbuhan tahunan dari Total GDP
- Harapan\_hidup : jumlah tahun rata-rata seorang anak yang baru lahir akan hidup jika pola kematian saat ini tetap sama
- Jumlah\_fertility : jumlah anak yang akan lahir dari setiap wanita jika tingkat kesuburan usia saat ini tetap sama
- GDPperkapita : total GDP dibagi total populasi



# DATA CLEANING

Melakukan pemeriksaan terhadap data untuk memastikan bahwa data akurat untuk dilakukan proses Analisa, sehingga output yang dihasilkan merupakan hasil yang akurat.

Melakukan pemeriksaan terhadap data jika ada nilai data yang kosong ('null') atau nilai yang tidak sesuai dengan type data variabel

**MISSING VALUE**

Melakukan pemeriksaan jika terdapat duplikasi data pada data yang akan dianalisis

**DUPLICATED**

Melakukan pemeriksaan jika terdapat data pencilan, untuk memastikan bahwa data tidak terdistorsi dari nilai yang asli

**OUTLIERS**

# MISSING VALUE

Melihat apakah ada nilai 'null' dalam dataset :

```
df.isnull().sum()
```

```
Negara          0
Kematian_anak    0
Ekspor           0
Kesehatan        0
Impor            0
Pendapatan       0
Inflasi          0
Harapan_hidup    0
Jumlah_fertiliti 0
GDPperkapita     0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Negara                167 non-null   object
1   Kematian_anak         167 non-null   float64
2   Ekspor                167 non-null   float64
3   Kesehatan             167 non-null   float64
4   Impor                167 non-null   float64
5   Pendapatan            167 non-null   int64
6   Inflasi               167 non-null   float64
7   Harapan_hidup         167 non-null   float64
8   Jumlah_fertiliti     167 non-null   float64
9   GDPperkapita          167 non-null   int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

Berdasarkan pemeriksaan .isnull() dan .info() di atas maka dapat dilihat bahwa dataset tidak mengandung nilai 'null', oleh karenanya dapat disimpulkan bahwa tidak ada missing value pada dataset.



## DUPLICATED

Melihat apakah terdapat duplikasi data dalam dataset :

```
df[df.duplicated(keep=False)]
```

Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
--------	---------------	--------	-----------	-------	------------	---------	---------------	------------------	--------------

Berdasarkan hasil pemeriksaan di atas maka diperoleh informasi bahwa tidak ada duplikasi data dalam dataset.

## OUTLIERS

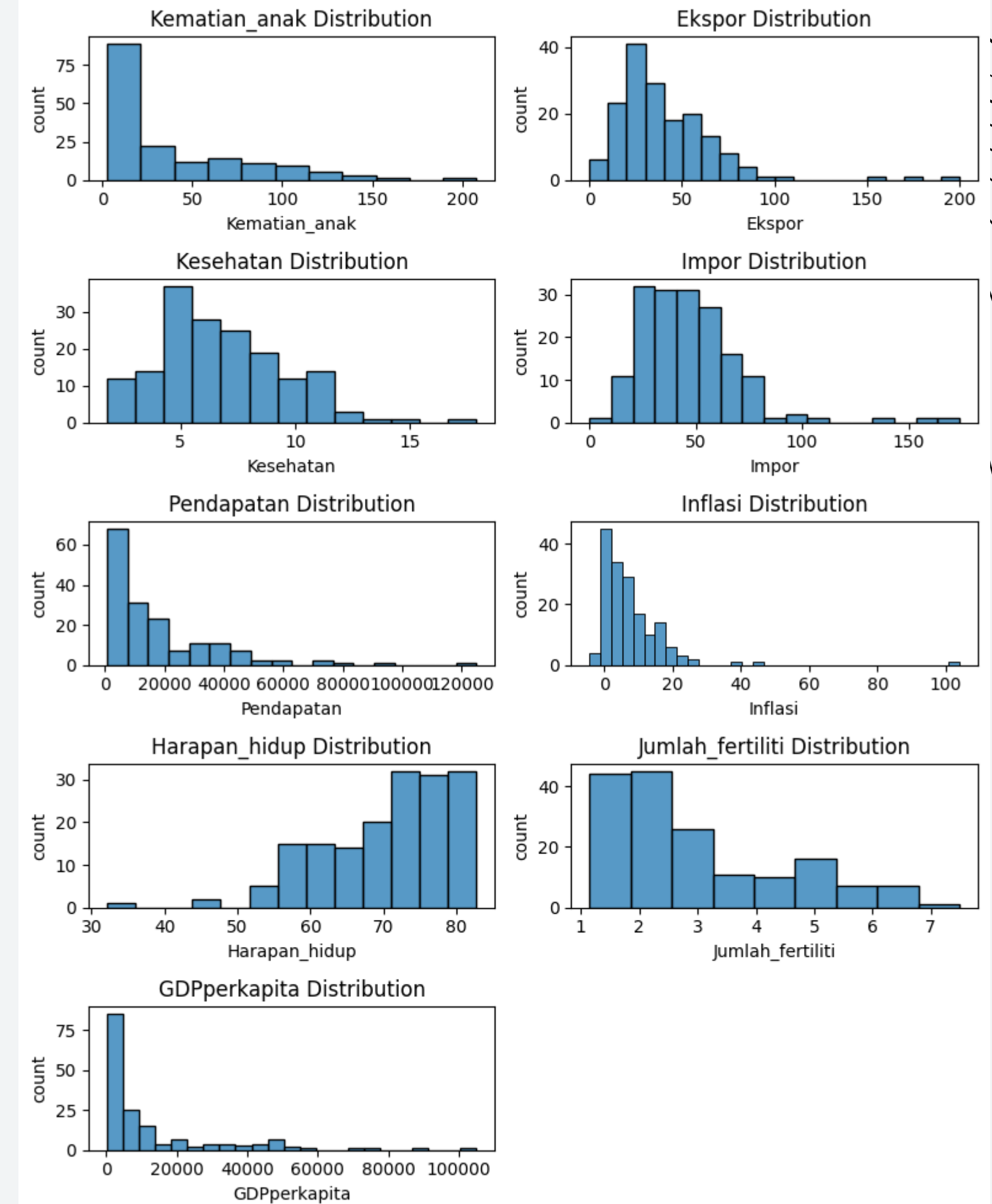
Melakukan pencarian jika ada nilai pencilan dalam dataset yang dapat menyebabkan distorsi data. Metode yang digunakan adalah dengan membuat histogram dan boxplot untuk setiap nilai variabel, untuk melihat jenis distribusi dan sebaran data dari setiap variabel.

```
# memvisualisasikan distribusi dari setiap variabel data
```

```
plt.figure(figsize=(8,10))  
for i, j in enumerate(df.describe().columns):  
    plt.subplot(5,2, i+1)  
    sns.histplot(x=df[j])  
    plt.xlabel(j)  
    plt.ylabel('count')  
    plt.title('{} Distribution'.format(j))  
    plt.subplots_adjust(wspace=.2, hspace=.5)  
    plt.tight_layout()  
plt.show()
```

Dari histogram dapat diambil kesimpulan sebagai berikut :

- Kematian\_anak: right-skewed distribution.
- Ekspor: right-skewed distribution.
- Kesehatan: right-skewed distribution.
- Impor: right-skewed distribution.
- Pendapatan: right-skewed distribution.
- Inflasi: right-skewed distribution.
- Harapan\_hidup: left-skewed distribution.
- Jumlah\_fertiliti: right-skewed distribution.
- GDPperkapita: right-skewed distribution.



```
# membuat diagram boxplot
```

```
plt.figure(figsize=(10,8))
```

```
for i, j in enumerate(df.describe().columns):
```

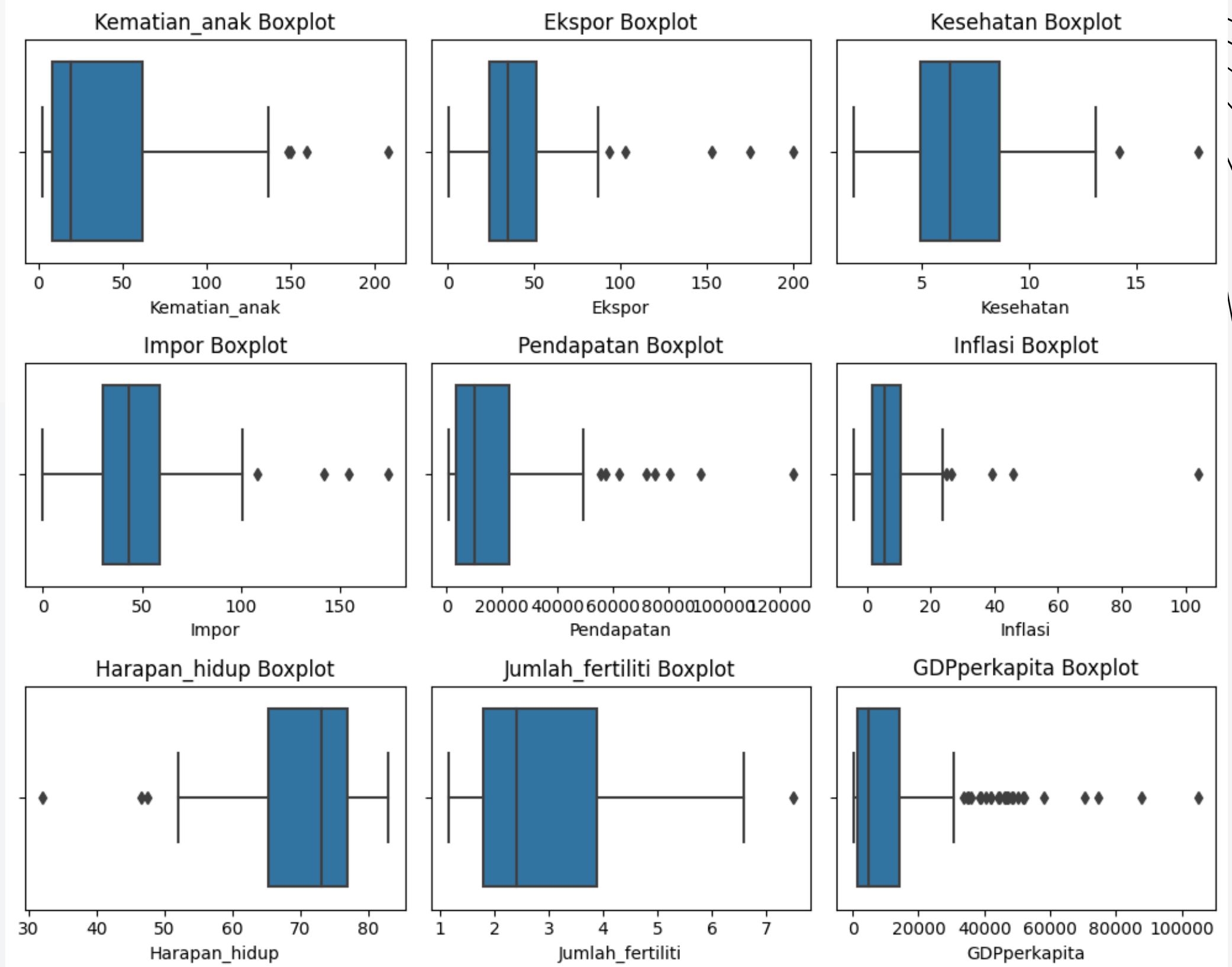
```
    plt.subplot(3,3, i+1)
```

```
    sns.boxplot(x=df[j])
```

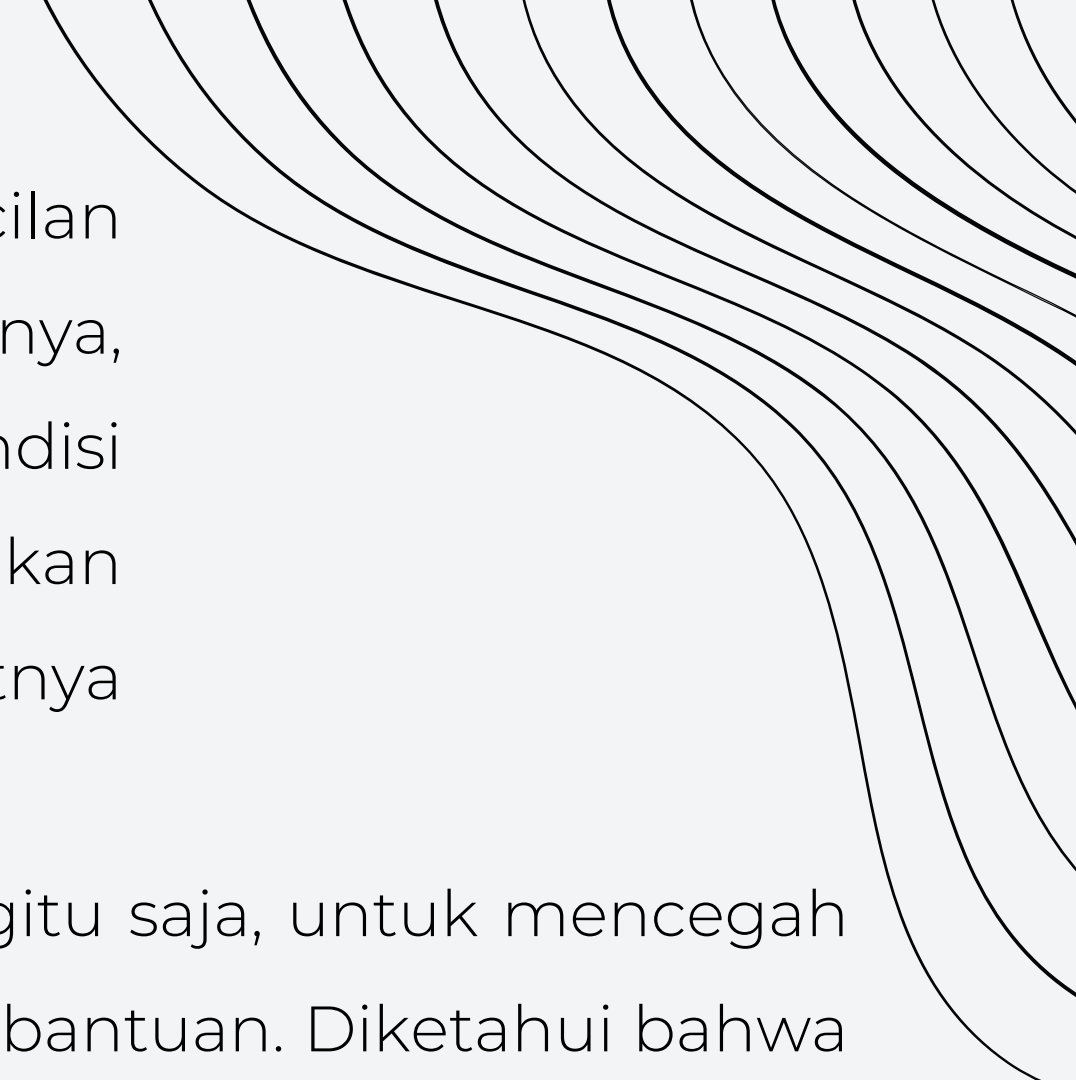
```
    plt.title('{} Boxplot'.format(j))
```

```
    plt.tight_layout()
```

```
plt.show()
```





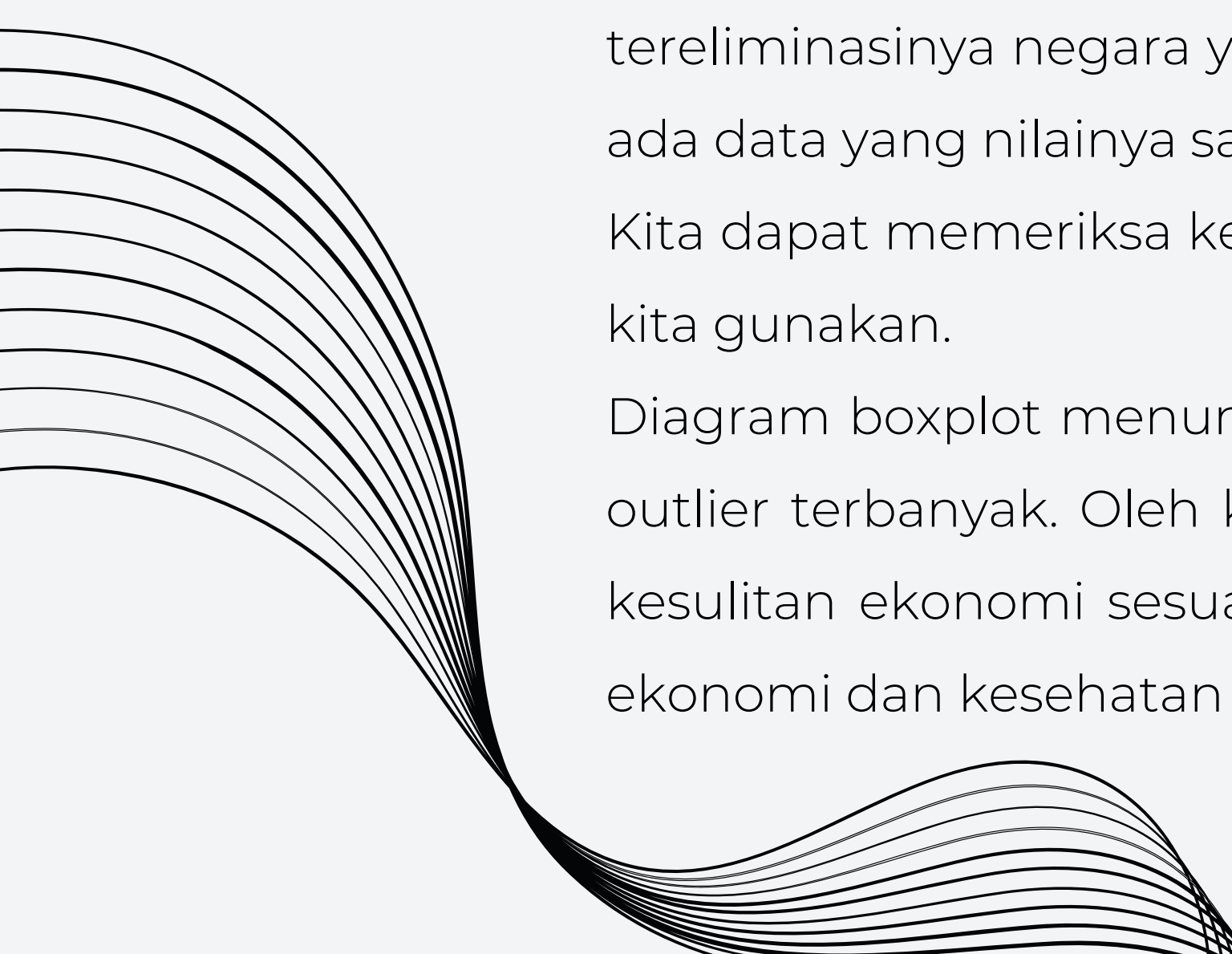


Dari histogram dan boxplot dapat diketahui bahwa terdapat cukup banyak pencilan data dalam dataset, dan kita perlu mempertimbangkan cara untuk mengatasinya, agar dataset dapat dikatakan akurat untuk dianalisis. Setiap negara memiliki kondisi yang pasti berbeda satu dengan yang lain, karena semua tergantung pada kebijakan pemerintahan masing-masing. Oleh karenanya data yang terkumpul ini juga sifatnya sangat variatif.

Dengan demikian kita tidak dapat menghilangkan outlier begitu saja, untuk mencegah tereliminasinya negara yang sebetulnya berhak mendapatkan bantuan. Diketahui bahwa ada data yang nilainya sangat tinggi tapi ada juga yang nilainya sangat rendah.

Kita dapat memeriksa kembali outlier setelah kita memutuskan variabel mana yang akan kita gunakan.

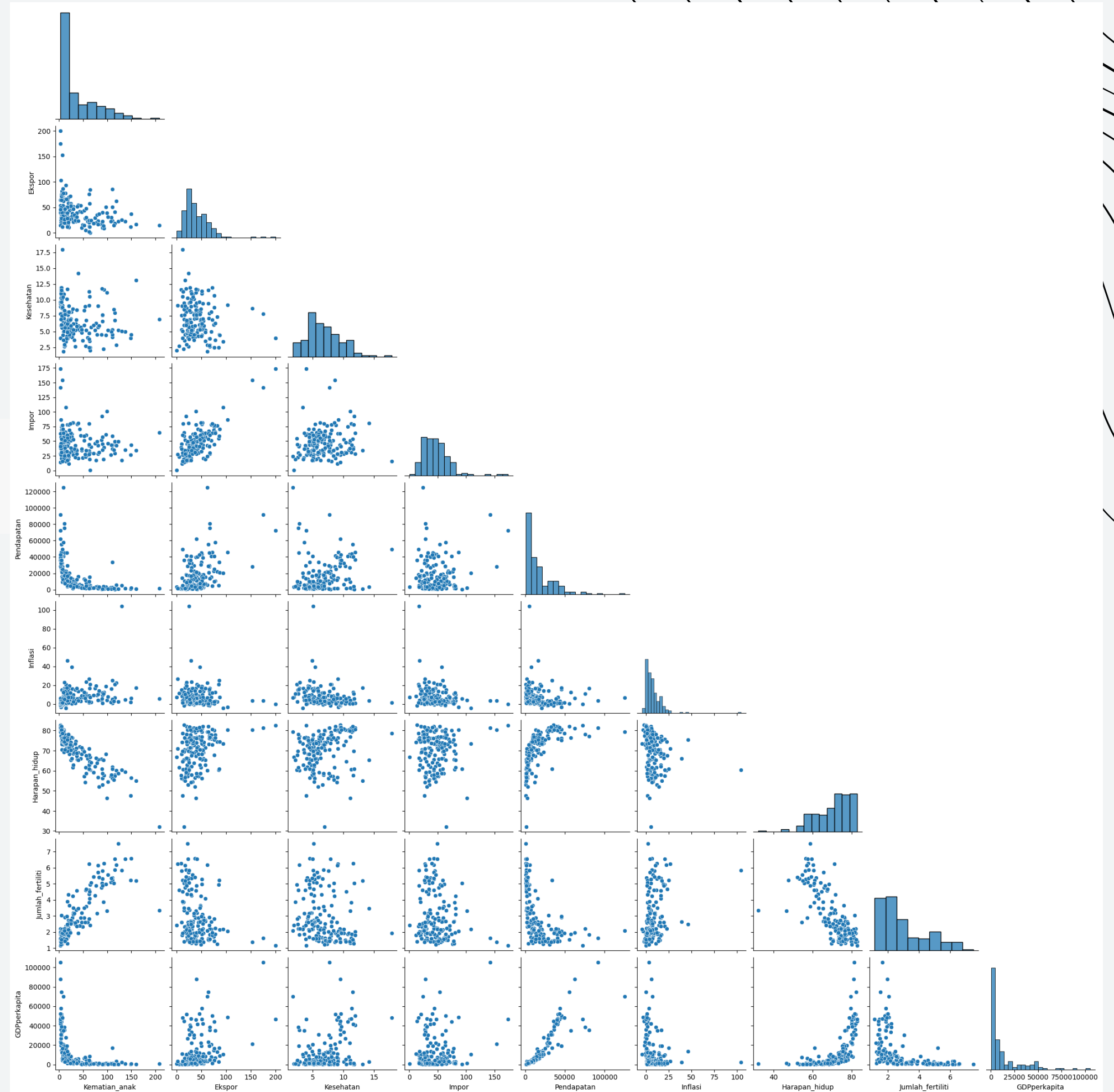
Diagram boxplot menunjukkan bahwa variabel GDPperkapita dan Pendapatan memiliki outlier terbanyak. Oleh karenanya kita perlu mengevaluasi negara mana yang memiliki kesulitan ekonomi sesuai tujuan kasus proyek ini, dengan mempertimbangkan tingkat ekonomi dan kesehatan masing-masing negara.



# EDA PART 1

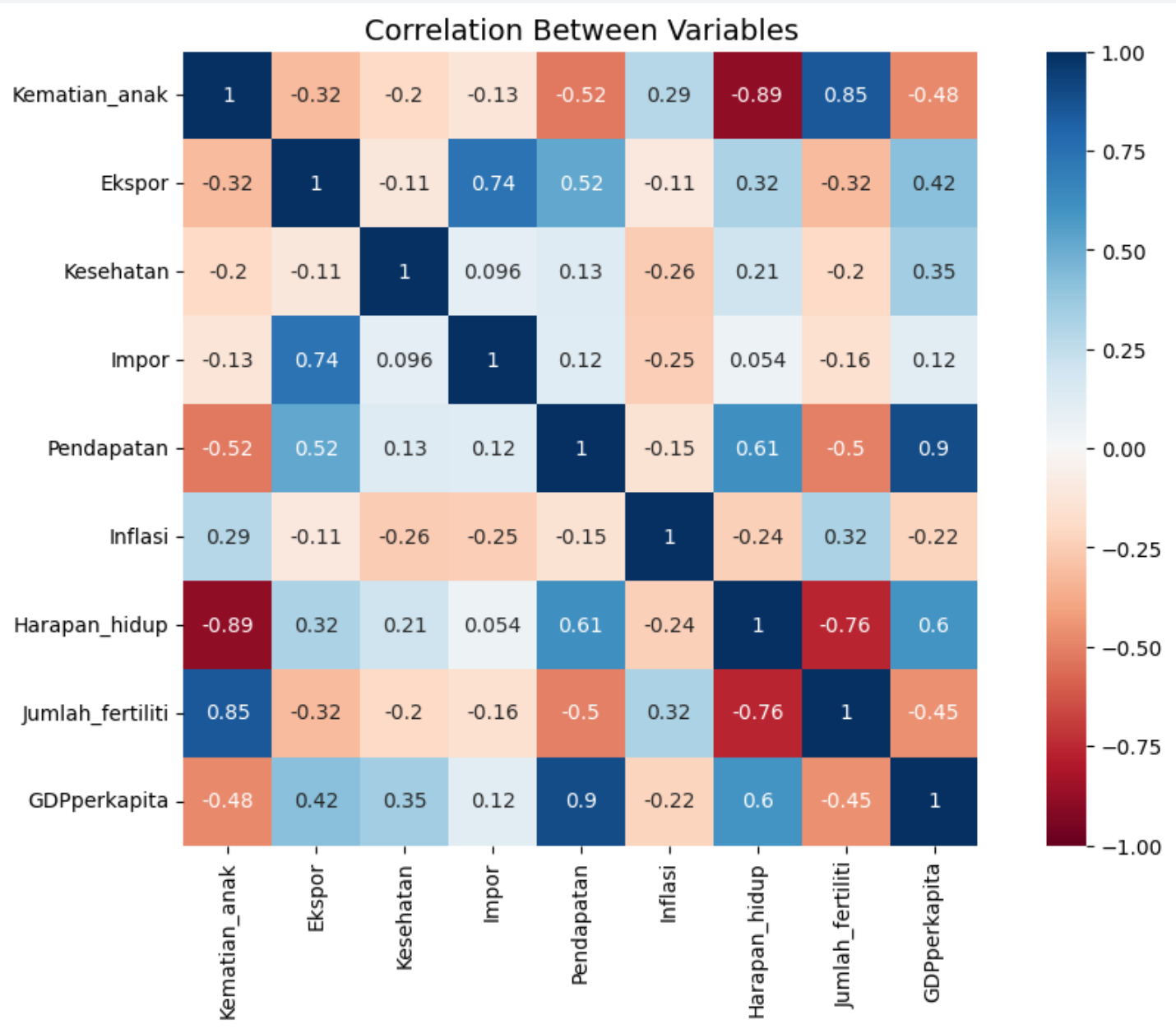
Kita akan memulai proses EDA dengan melakukan multivariate analysis. Analisis ini sangat berguna untuk melihat korelasi antar variabel data.

```
pairplot = sns.pairplot(df, corner=True)  
plt.show(pairplot)
```



Selain itu juga dapat dibuat diagram heatmap untuk melihat nilai koefisien korelasi antar variabel.

```
correlation_metrics=df.corr()
fig = plt.figure(figsize=(12,7))
sns.heatmap(correlation_metrics,square=True, annot=True, vmax=1, vmin=-1,
            cmap='RdBu')
plt.title('Correlation Between Variables', size=14)
plt.show()
```



Nilai koefisien korelasi dapat diinterpretasikan sebagai berikut :

- -1 hingga -0.91 ATAU 0.91 hingga 1 : sangat kuat
- -0.90 hingga -0.71 ATAU 0.71 hingga 0.90 : kuat
- -0.70 hingga -0.51 ATAU 0.51 hingga 0.70 : sedang
- -0.50 hingga -0.31 ATAU 0.31 hingga 0.50 : lemah
- -0.30 hingga -0.01 ATAU 0.01 hingga 0.30 : sangat lemah
- 0 : tidak berhubungan

Dari diagram heatmap dan nilai korelasi di atas, dapat kita peroleh kesimpulan sebagai berikut :

- Pendapatan & GDPperkapita : korelasi sangat kuat positif
- Harapan\_hidup & Jumlah\_fertiliti : korelasi kuat negatif
- Jumlah\_fertiliti & Kematian\_anak : korelasi kuat positif
- Harapan\_hidup & Kematian\_anak : korelasi kuat negatif
- Ekspor & Impor : korelasi kuat positif
- Harapan\_hidup & GDPperkapita : korelasi sedang positif
- Harapan\_hidup & Pendapatan : korelasi sedang positif
- Pendapatan & Kematian\_anak : korelasi sedang negatif
- Pendapatan & Ekspor : korelasi sedang positif



# PEMILIHAN VARIABEL

Untuk menentukan variabel yang tepat, kita perlu meninjau kembali tujuan dari proyek ini. Proyek ini bertujuan untuk mengkategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan.

Variabel Pendapatan dan GDPperkapita adalah variabel yang tepat untuk mewakili faktor sosial-ekonomi suatu negara. Tapi dengan mempertimbangkan korelasi antara kedua variabel tersebut yang merupakan korelasi kuat positif, maka kita dapat hanya menggunakan variabel Pendapatan saja. Hal ini untuk mencegah hasil analisa merujuk pada negara yang memiliki Pendapatan yang tinggi.

Selanjutnya kita perlu menentukan pasangan variabel yang tepat, untuk mewakili faktor kesehatan suatu negara, dan memiliki hubungan dengan Pendapatan. Dari korelasi heatmap, variabel Pendapatan memiliki hubungan (sedang) dengan Kematian\_anak, Ekspor, dan Harapan \_hidup. Dari pilihan variabel ini kita dapat menentukan Kematian \_anak sebagai wakil faktor kesehatan suatu negara.

Dengan demikian untuk selanjutnya kita akan menggunakan variabel Pendapatan dan Kematian\_anak sebagai fitur pengelompokkan.

# HANDLING OUTLIER

Sebelum kita lanjutkan ke proses pengelompokkan, kita perlu melakukan 'handling outlier' seperti yang sudah disebutkan sebelumnya.

Untuk memperkecil kelompok negara yang akan kita pilih, kita dapat memfilter negara dengan Pendapatan di bawah nilai median Pendapatan, untuk memastikan bahwa perusahaan memberikan bantuan keuangan pada negara yang termasuk memiliki tingkat pendapatan rendah.

Nilai yang digunakan sebagai acuan adalah nilai median karena distribusi data variabel Pendapatan berbentuk skew ke kanan, sehingga akan lebih tepat jika kita menggunakan nilai median daripada nilai mean untuk memfilter berdasarkan Pendapatan.

```
# filter negara dengan pendapatan di bawah nilai median
df_filter_pendapatan = df[df.Pendapatan < df.Pendapatan.median()]
df_filter_pendapatan
```

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
6	Armenia	18.1	20.8	4.40	45.3	6700	7.77	73.3	1.69	3220
12	Bangladesh	49.4	16.0	3.52	21.8	2440	7.14	70.4	2.33	758
...	...	...	...	...	...	...	...	...	...	...
161	Uzbekistan	36.3	31.7	5.81	28.5	4240	16.50	68.8	2.34	1380
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

83 rows × 10 columns

Dari hasil filter tersebut diperoleh jumlah negara yang dapat dipilih sekitar 50% dari jumlah awal. Kita bisa periksa kembali histogram dan boxplot untuk melihat jika masih terdapat outlier pada data.



```
# membuat histogram dan boxplot baru untuk Pendapatan dan Kematian_anak
```

```
df_baru=df_filter_pendapatan
```

```
fig = plt.figure(figsize=(8,6))
```

```
plt.subplot(2,2,1)
```

```
sns.boxplot(x=df_baru['Pendapatan'])
```

```
plt.title('Pendapatan Boxplot New')
```

```
plt.tight_layout()
```

```
plt.subplot(2,2,2)
```

```
sns.boxplot(x=df_baru['GDPperkapita'])
```

```
plt.title('GDPperkapita Boxplot New')
```

```
plt.tight_layout()
```

```
plt.subplot(2,2,3)
```

```
sns.histplot(x=df_baru['Pendapatan'])
```

```
plt.title('Pendapatan Histogram New')
```

```
plt.tight_layout()
```

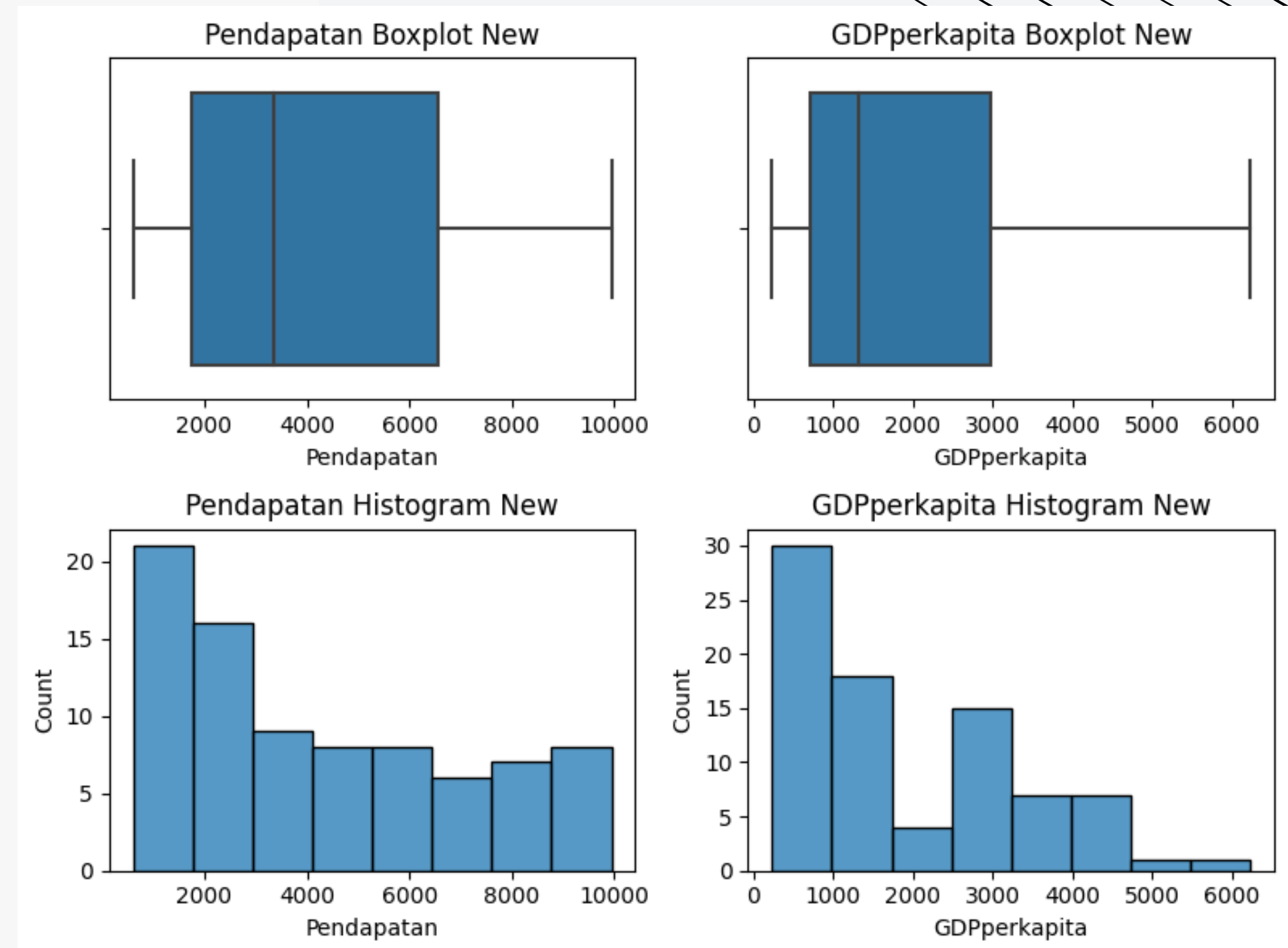
```
plt.subplot(2,2,4)
```

```
sns.histplot(x=df_baru["GDPperkapita"])
```

```
plt.title('GDPperkapita Histogram New')
```

```
plt.tight_layout()
```

```
plt.show()
```



Dari diagram di atas, kita dapat menyimpulkan bahwa data baru setelah filter Pendapatan lebih "clean" dari data awal, di mana tidak terdapat nilai outlier pada data baru. Dengan demikian data baru ini sudah dapat dikatakan akurat untuk diolah dalam pengelompokan.

# EDA PART 2

## 1. Univariate analysis

Untuk melakukan Univariate Analysis pada kedua variabel, kita akan membuat Histogram & Poligon. Histogram akan menunjukkan distribusi frekuensi suatu data, sedangkan poligon akan menghubungkan titik tengah atas setiap batang histogram.

## 2. Bivariate analysis

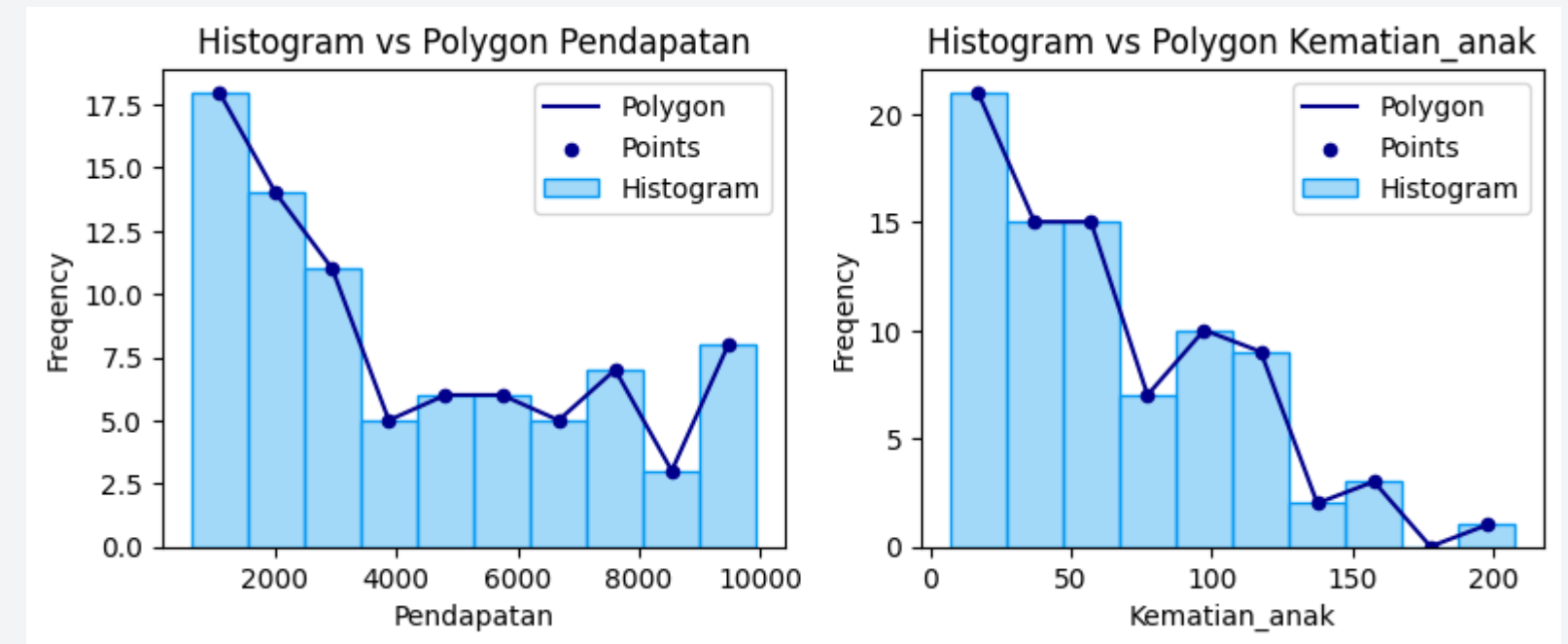
Sebelum kita lakukan clustering, kita perlu melihat korelasi antar variabel yang telah kita pilih dengan menggunakan scatterplot dan fungsi `.hexbin()` untuk melihat konsentrasi nilai variabel tersebut.

# UNIVARIATE ANALYSIS

```
# membuat histogram & polygon untuk Pendapatan dan Kematian_anak
fig = plt.figure(figsize=(8,6))
plt.subplot(2,2,1)
sns.histplot(df_baru['Pendapatan'], bins=10, alpha=0.7, color='#7BC8F6',
             edgecolor='#069AF3', label='Histogram')
hist, edges = np.histogram(df_baru['Pendapatan'], bins=10)
midpoints = (edges[1:] + edges[:-1])/2
plt.plot(midpoints, hist, color='#00008B', label='Polygon')
plt.scatter(midpoints, hist, color='#00008B', marker='o', s=20, label='Points')
plt.xlabel('Pendapatan')
plt.ylabel('Frequency')
plt.title('Histogram vs Polygon Pendapatan')
plt.legend()
plt.tight_layout()

plt.subplot(2,2,2)
sns.histplot(df_baru['Kematian_anak'], bins=10, alpha=0.7, color='#7BC8F6',
             edgecolor='#069AF3', label='Histogram')
hist, edges = np.histogram(df_baru['Kematian_anak'], bins=10)
midpoints = (edges[1:] + edges[:-1])/2
plt.plot(midpoints, hist, color='#00008B', label='Polygon')
plt.scatter(midpoints, hist, color='#00008B', marker='o', s=20, label='Points')
plt.xlabel('Kematian_anak')
plt.ylabel('Frequency')
plt.title('Histogram vs Polygon Kematian_anak')
plt.legend()
plt.tight_layout()

plt.show()
```





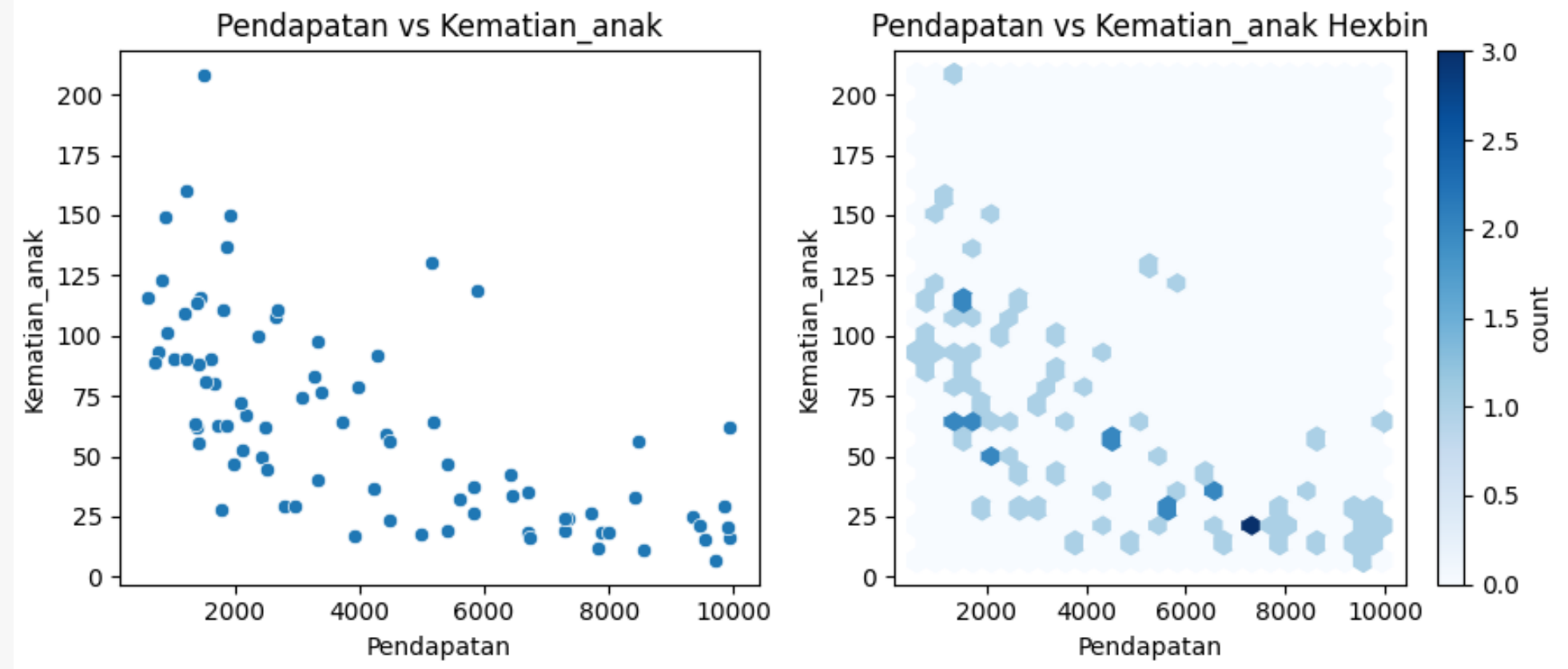
# BIVARIATE ANALYSIS

```
# membuat scatterplot dan hexagonal bin
fig = plt.figure(figsize=(9,4))

plt.subplot(1,2,1)
sns.scatterplot(x=df_baru['Pendapatan'],
               y=df_baru['Kematian_anak'])
plt.title('Pendapatan vs Kematian_anak')
plt.tight_layout()

plt.subplot(1,2,2)
hexplot=plt.hexbin(x=df_baru['Pendapatan'],
                  y=df_baru['Kematian_anak'],
                  gridsize = 25, cmap='Blues')
plt.title('Pendapatan vs Kematian_anak Hexbin')
plt.xlabel('Pendapatan')
plt.ylabel('Kematian_anak')
plt.colorbar(hexplot, label='count')
plt.tight_layout()

plt.show()
```



Dari kedua diagram di atas dapat dilihat bahwa variabel Pendapatan dan Kematian\_anak memiliki hubungan sedang negatif. dan dari diagram hexbin menunjukkan bahwa ada hubungan kuat antara kedua variabel tersebut pada beberapa negara.

# CLUSTERING

## 1. Scaling Process

```
# membuat dataframe baru untuk scaling process
```

```
df_cluster = df_baru[['Negara', 'Pendapatan', 'Kematian_anak']].reset_index()
df_cluster.drop('index', inplace=True, axis=1)
display(df_cluster)
```

	Negara	Pendapatan	Kematian_anak
0	Afghanistan	1610	90.2
1	Albania	9930	16.6
2	Angola	5900	119.0
3	Armenia	6700	18.1
4	Bangladesh	2440	49.4
...	...	...	...
78	Uzbekistan	4240	36.3
79	Vanuatu	2950	29.2
80	Vietnam	4490	23.3
81	Yemen	4480	56.3
82	Zambia	3280	83.1

83 rows × 3 columns

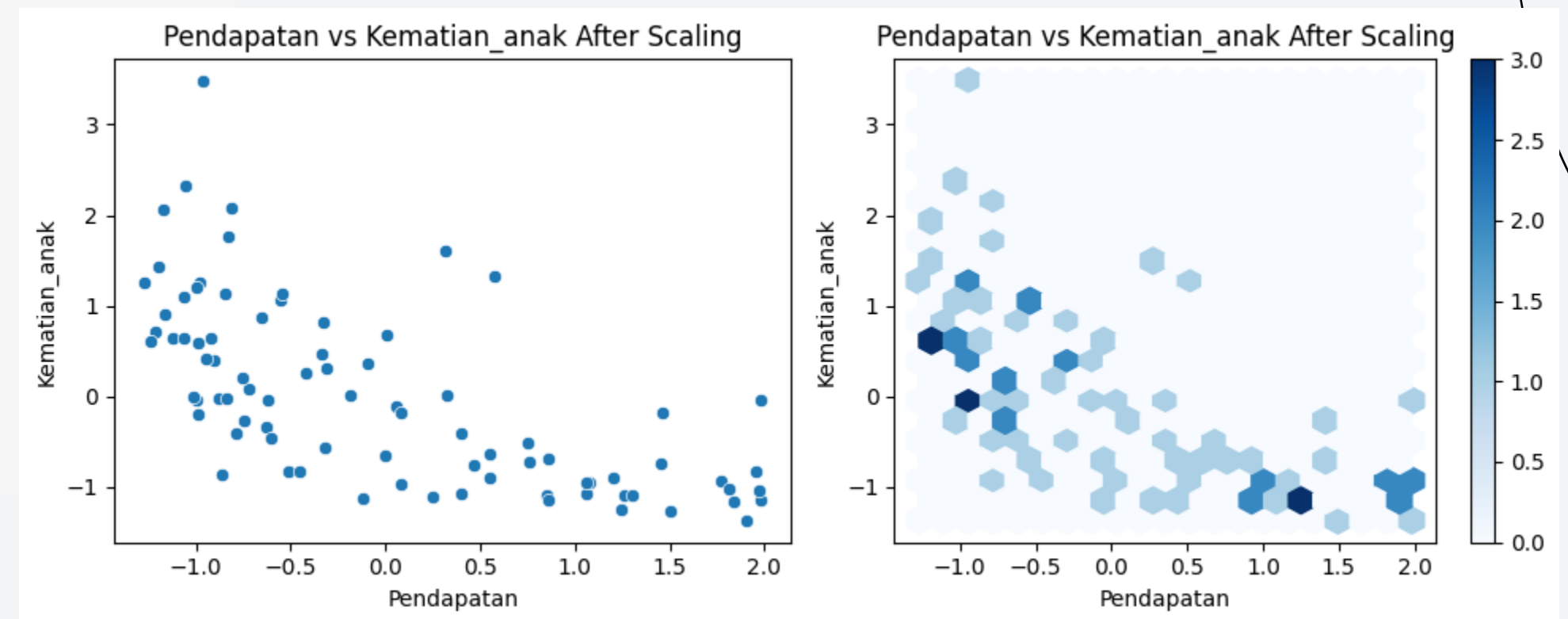
```
#scaling process
```

```
sc = skp.StandardScaler()
data_scale = np.array(df_cluster[['Pendapatan', 'Kematian_anak']])
scaled = sc.fit_transform(data_scale.astype(float))
df_scaled = pd.DataFrame(scaled, columns=['Pendapatan', 'Kematian_anak'])
display(df_scaled)
```

	Pendapatan	Kematian_anak
0	-0.920666	0.640089
1	1.977480	-1.132268
2	0.573691	1.333620
3	0.852359	-1.096147
4	-0.631548	-0.342414
...	...	...
78	-0.004545	-0.657874
79	-0.453897	-0.828848
80	0.082539	-0.970926
81	0.079055	-0.176255
82	-0.338947	0.469114

83 rows × 2 columns

```
# do bivariate analysis untuk melihat kondisi data setelah scaling
fig = plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
sns.scatterplot(x=df_scaled['Pendapatan'],
                y=df_scaled['Kematian_anak'])
plt.title('Pendapatan vs Kematian_anak After Scaling')
plt.tight_layout()
plt.subplot(1,2,2)
hb = plt.hexbin(x=df_scaled['Pendapatan'],
                y=df_scaled['Kematian_anak'],
                gridsize = 20, cmap = 'Blues')
cb = plt.colorbar(hb)
plt.title('Pendapatan vs Kematian_anak After Scaling')
plt.xlabel('Pendapatan')
plt.ylabel('Kematian_anak')
plt.tight_layout()
plt.show()
```



Dari hasil perbandingan bivariate analysis, tidak ada perbedaan yang signifikan. Dengan demikian dapat disimpulkan bahwa data scaling sudah tepat.

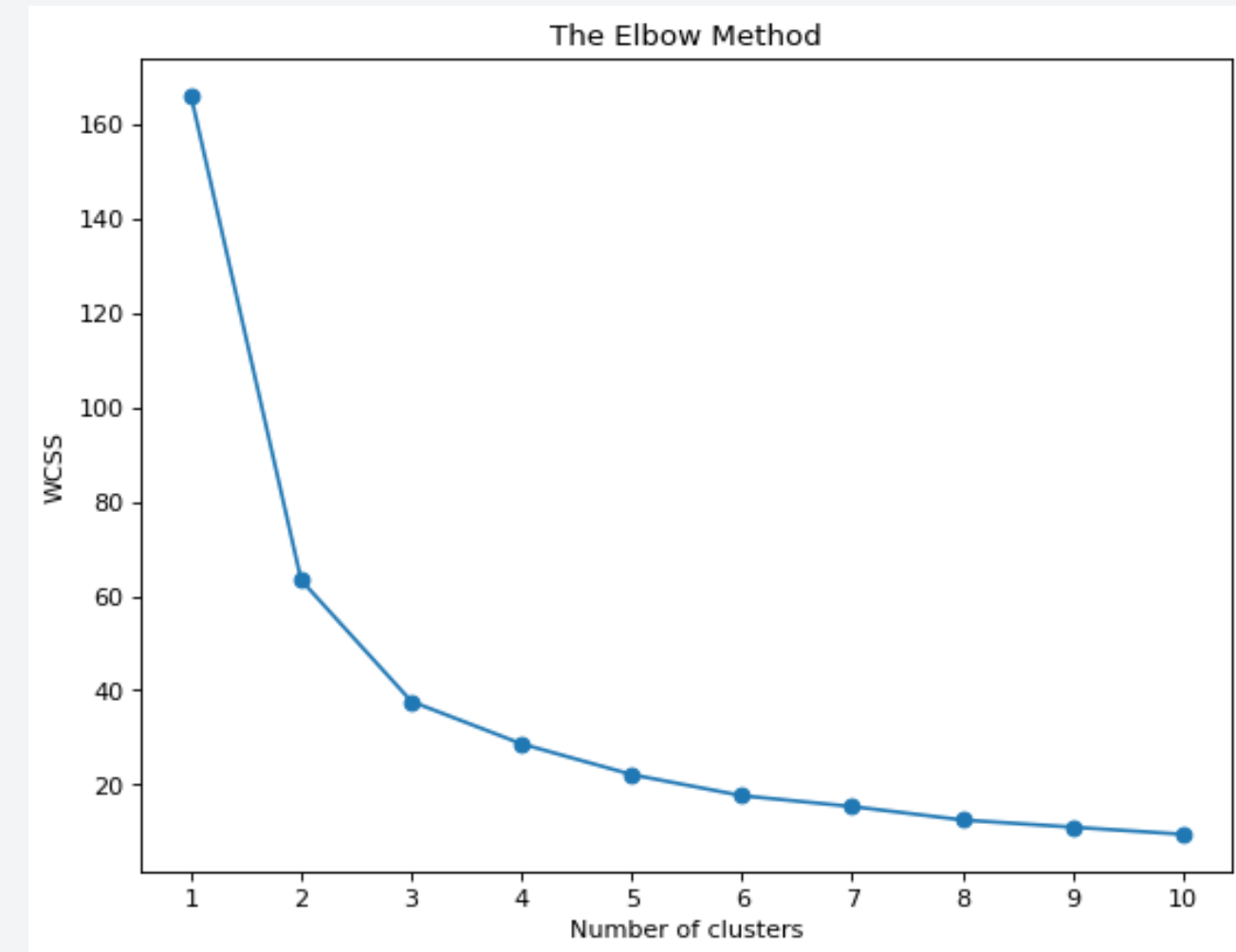
## 2. Menentukan K-Means untuk parameter clustering

Metode yang digunakan adalah Elbow Method

```
# elbox method untuk menentukan k
wcss=[]
k_range = range(1,11)
for i in k_range:
    kmeans = skc.KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(df_scaled)
    wcss.append(kmeans.inertia_)
fig, ax = plt.subplots(figsize=(8, 6), dpi=80)
plt.plot(k_range, wcss, marker='o')

plt.xticks(k_range)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')

plt.show()
```



Dari hasil perhitungan Elbow Method diperoleh grafik di atas, dan dapat dilihat bahwa nilai k yang sesuai adalah 3. Dengan demikian kita akan melakukan clustering dengan jumlah kelas = 3.



### 3. Do Clustering

```
# Clustering K Means, K=3
kmeans_3 = skc.KMeans(n_clusters=3,random_state=42)
kmeans_3.fit(df_scaled)
kmeans_3.labels_

array([2, 1, 2, 1, 0, 1, 2, 1, 0, 1, 2, 2, 0, 2, 1, 2, 2, 1, 2, 2, 0, 2,
       1, 1, 1, 0, 1, 0, 1, 0, 1, 2, 2, 1, 2, 0, 1, 1, 1, 0, 0, 0, 0, 2,
       2, 0, 2, 2, 2, 0, 0, 1, 1, 2, 0, 1, 0, 2, 2, 0, 1, 1, 0, 1, 0, 2,
       0, 1, 1, 0, 0, 0, 0, 2, 1, 1, 0, 1, 0, 0, 0, 0, 0], dtype=int32)
```

```
# menggabungkan hasil clustering ke dalam setiap negara dalam dataframe

df_cluster['cluster_id'] = kmeans_3.labels_
display(df_cluster)
```

	Negara	Pendapatan	Kematian_anak	cluster_id
0	Afghanistan	1610	90.2	2
1	Albania	9930	16.6	1
2	Angola	5900	119.0	2
3	Armenia	6700	18.1	1
4	Bangladesh	2440	49.4	0
...	...	...	...	...
78	Uzbekistan	4240	36.3	0
79	Vanuatu	2950	29.2	0
80	Vietnam	4490	23.3	0
81	Yemen	4480	56.3	0
82	Zambia	3280	83.1	0

83 rows × 4 columns

```
# menghitung jumlah negara dalam setiap cluster
```

```
df_cluster.cluster_id.value_counts(ascending=True)
```

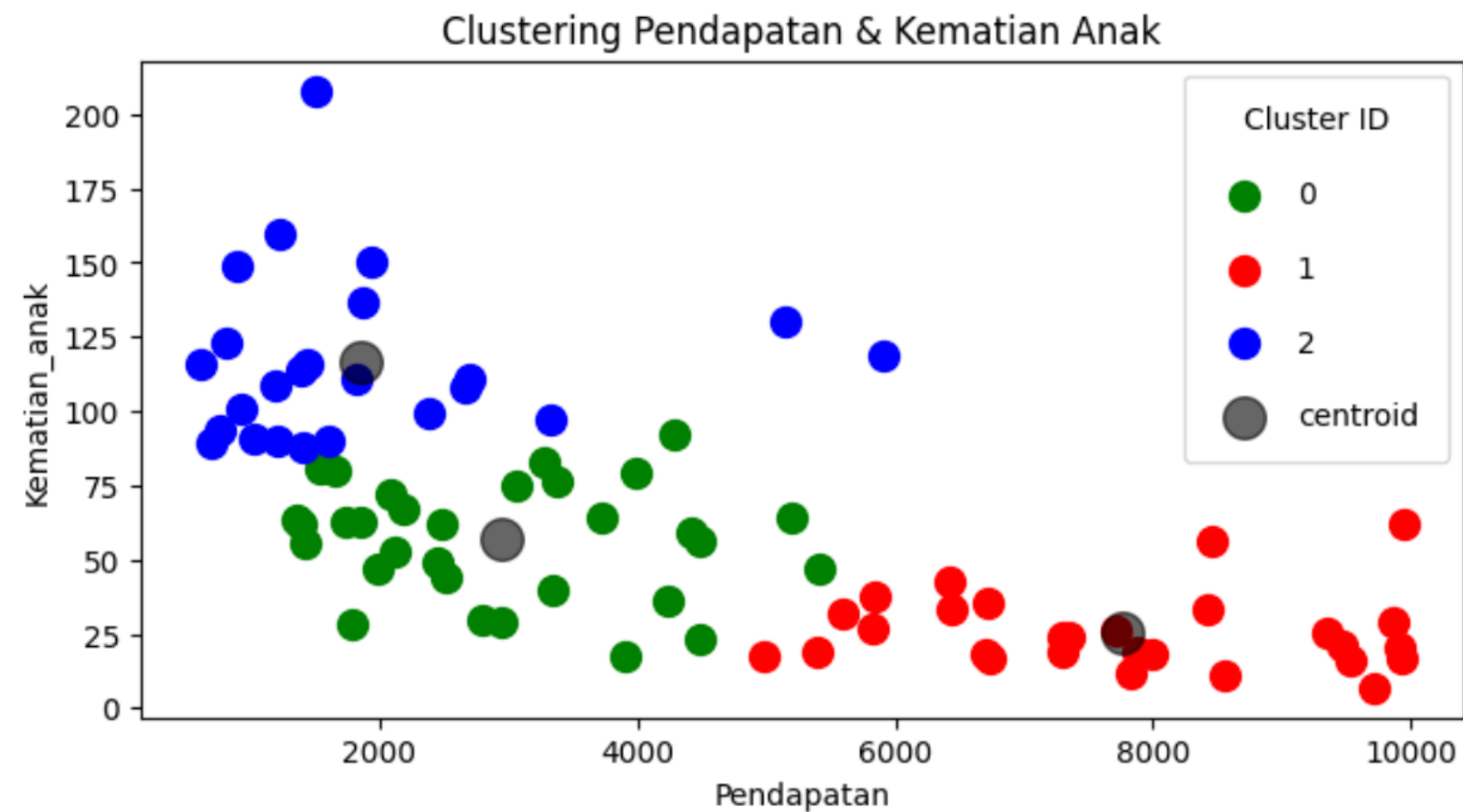
```
2      24
1      28
0      31
Name: cluster_id, dtype: int64
```

## 4. Membuat Grafik Hasil Clustering

```
# Centroid Inverse Scaling
centroids_ori_scale = sc.inverse_transform(kmeans_3.cluster_centers_)

# Plot hasil clustering
fig, ax = plt.subplots(figsize=(8, 4), dpi=100)
plt.scatter(df_cluster.Pendapatan[df_cluster.cluster_id == 0], df_cluster['Kematian_anak'][df_cluster.cluster_id == 0], color = 'green', s=100, label= '0')
plt.scatter(df_cluster.Pendapatan[df_cluster.cluster_id == 1], df_cluster['Kematian_anak'][df_cluster.cluster_id == 1], color = 'red', s=100, label= '1')
plt.scatter(df_cluster.Pendapatan[df_cluster.cluster_id == 2], df_cluster['Kematian_anak'][df_cluster.cluster_id == 2], color = 'blue', s=100, label= '2')
ax.scatter(centroids_ori_scale[:, 0], centroids_ori_scale[:,1], c='black', s=200, marker='o', alpha=0.6, label = 'centroid')
plt.legend(title= "Cluster ID", labelspace=1.5, borderpad=1)
plt.xlabel('Pendapatan')
plt.ylabel('Kematian_anak')
plt.title("Clustering Pendapatan & Kematian Anak")

plt.show()
```



# DECISION MAKING



Choose which country  
cluster to focus on

**STRATEGY N°1**



Show which countries are  
in the cluster

**STRATEGY N°2**



Choose which countries are  
eligible to received the  
USD 10Mill fund  
(recommendation for the  
CEO of HELP International)

**STRATEGY N°3**





# CHOOSE WHICH COUNTRY CLUSTER TO FOCUS


```
# melihat karakter tiap cluster

grouped_df=df_cluster.groupby('cluster_id').agg({'Pendapatan':['max','min','mean'],
                                                'Kematian_anak':['max','min','mean']})

grouped_df
```

cluster_id	Pendapatan			Kematian_anak		
	max	min	mean	max	min	mean
0	5410	1350	2949.354839	92.1	17.2	56.803226
1	9940	4980	7756.071429	62.0	6.9	25.653571
2	5900	609	1850.125000	208.0	88.2	116.716667



Country to focus : cluster 2 (berwarna biru) merupakan kelompok negara yang paling berhak menerima bantuan ekonomi, dimana rata-rata pendapatan kelompok ini adalah yang terendah, dan memiliki nilai rata-rata kematian anak yang tertinggi.

## SHOW THE COUNTRIES IN CLUSTER 2

Untuk melihat negara mana saja yang termasuk dalam cluster 2, maka kita perlu memasukkan informasi variabel cluster ke dalam data.

```
# menunjukkan isi negara di cluster 2
df_cluster2 = df_cluster[df_cluster.cluster_id == 2]
display(df_cluster2)
```

	Negara	Pendapatan	Kematian_anak	cluster_id
0	Afghanistan	1610	90.2	2
2	Angola	5900	119.0	2
6	Benin	1820	111.0	2
10	Burkina Faso	1430	116.0	2
11	Burundi	764	93.6	2
13	Cameroon	2660	108.0	2
15	Central African Republic	888	149.0	2
16	Chad	1930	150.0	2
18	Comoros	1410	88.2	2
19	Congo, Dem. Rep.	609	116.0	2
21	Cote d'Ivoire	2690	111.0	2
31	Guinea	1190	109.0	2
32	Guinea-Bissau	1390	114.0	2
34	Haiti	1500	208.0	2
43	Lesotho	2380	99.7	2
44	Liberia	700	89.3	2
46	Malawi	1030	90.5	2
47	Mali	1870	137.0	2
48	Mauritania	3320	97.4	2
53	Mozambique	918	101.0	2
57	Niger	814	123.0	2
58	Nigeria	5150	130.0	2
65	Sierra Leone	1220	160.0	2
73	Togo	1210	90.3	2

## CHOOSE THE COUNTRY TO RECEIVED USD 10Mill

Untuk memilih negara yang berhak menerima bantuan maka diperlukan analisis lanjutan untuk negara yang berada di cluster 2. Dengan berdasarkan pada variabel yang menjadi fokus dalam proyek ini, maka kita dapat mengurutkan peringkat negara cluster 2 untuk mendapatkan negara yang memiliki Pendapatan terendah dan Kematian\_anak tertinggi.

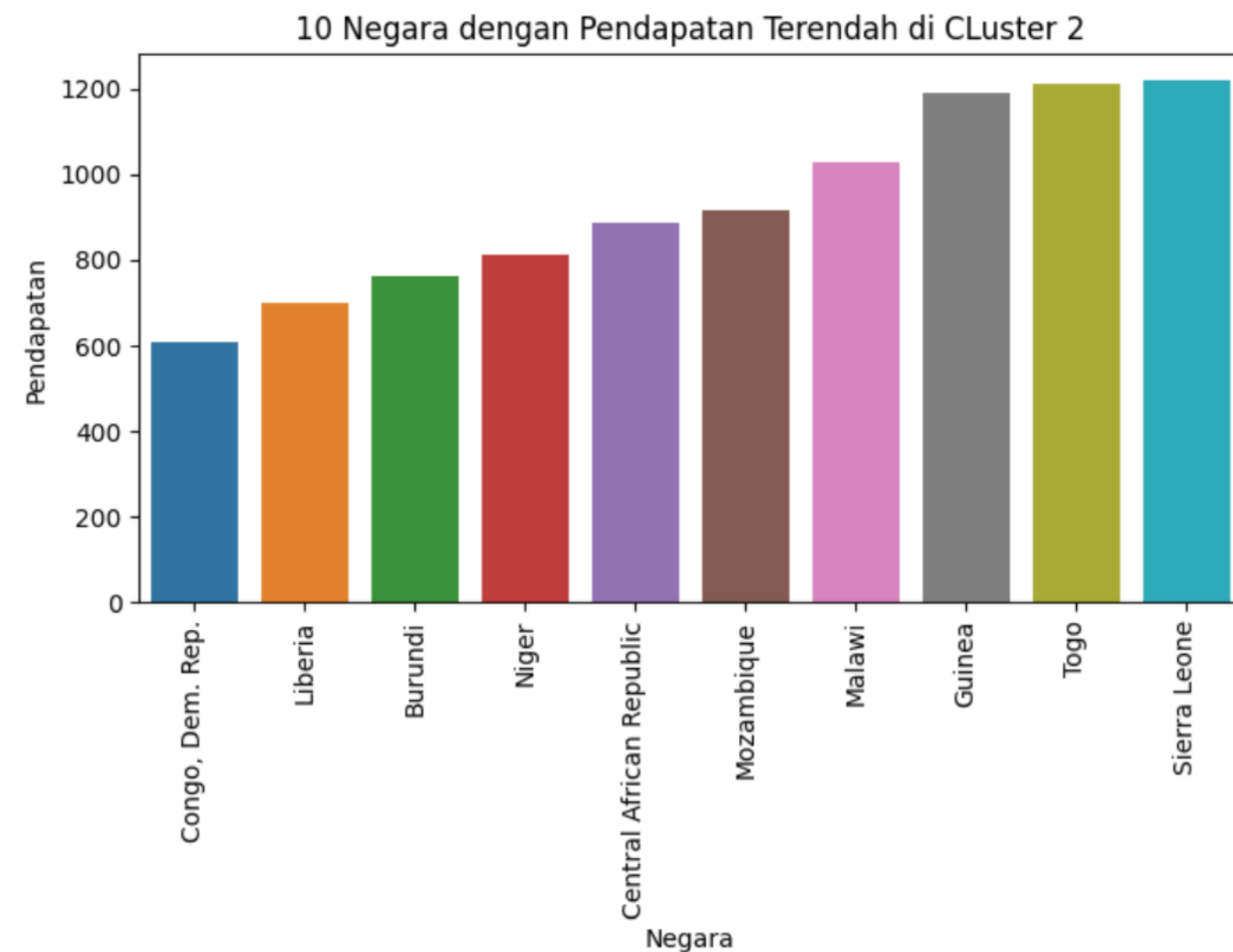
```
# mengurutkan data berdasarkan pendapatan terendah (asc)
df_cluster2_sort_pendapatan = df_cluster2.sort_values('Pendapatan').head(10).reset_index().drop('index', axis=1, inplace=False)
display(df_cluster2_sort_pendapatan)

# Visualisasi data setelah diurutkan
x = df_cluster2_sort_pendapatan.Negara.tolist()
y = df_cluster2_sort_pendapatan.Pendapatan.tolist()
fig, ax = plt.subplots(figsize=(8, 4), dpi=100)
sns.barplot(x=df_cluster2_sort_pendapatan.Negara,
            y=df_cluster2_sort_pendapatan.Pendapatan)
ax.set_xticklabels(df_cluster2_sort_pendapatan.Negara,
                  rotation = 90)

plt.title('10 Negara dengan Pendapatan Terendah di CLuster 2')

plt.show()
```

	Negara	Pendapatan	Kematian_anak	cluster_id
0	Congo, Dem. Rep.	609	116.0	2
1	Liberia	700	89.3	2
2	Burundi	764	93.6	2
3	Niger	814	123.0	2
4	Central African Republic	888	149.0	2
5	Mozambique	918	101.0	2
6	Malawi	1030	90.5	2
7	Guinea	1190	109.0	2
8	Togo	1210	90.3	2
9	Sierra Leone	1220	160.0	2



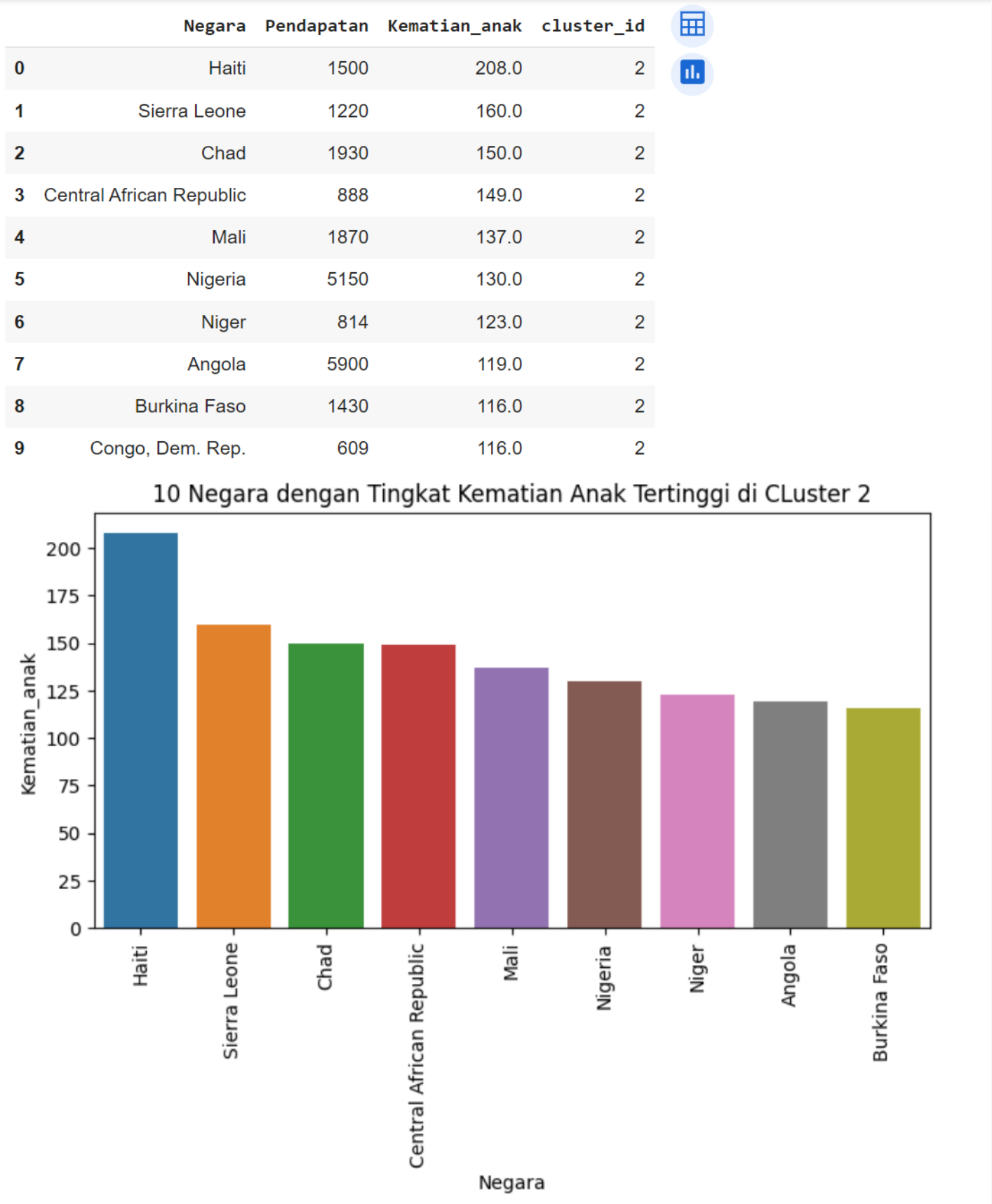


```
# mengurutkan data berdasarkan tingkat kematian anak (desc)
df_cluster2_sort_kematian = df_cluster2.sort_values('Kematian_anak', ascending=False).head(10).reset_index().drop('index', axis=1, inplace=False)
display(df_cluster2_sort_kematian)

# Visualisasi data setelah diurutkan
fig, ax = plt.subplots(figsize=(8, 4), dpi=100)
df_cluster2_sort_kematian = df_cluster2.sort_values('Kematian_anak',
                                                    ascending=False).head(9)

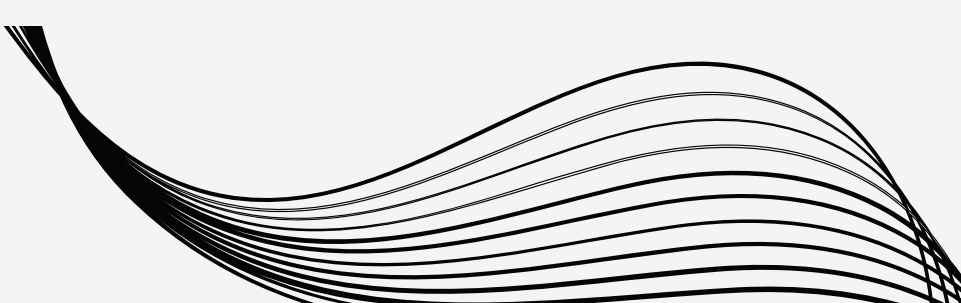
sns.barplot(x=df_cluster2_sort_kematian.Negara,
            y=df_cluster2_sort_kematian.Kematian_anak)
ax.set_xticklabels(df_cluster2_sort_kematian.Negara,
                  rotation = 90)
plt.title('10 Negara dengan Tingkat Kematian Anak Tertinggi di CLuster 2')

plt.show()
```





Setelah pengurutan data, dapat kita peroleh informasi sebagai berikut :

- Negara pendapatan terendah : Congo
  - Negara dengan tingkat kematian anak tertinggi : Haiti
  - Niger merupakan negara dengan pendapatan k-4 terendah dan memiliki tingkat kematian anak ke-7 tertinggi
  - Central African Republik merupakan negara dengan pendapatan k-5 terendah dan memiliki tingkat kematian k-4 tertinggi
  - Sierra Leone merupakan negara dengan pendapatan ke-10 terendah dan memiliki tingkat kematian anak k-2 tertinggi
  - Negara lainnya merupakan negara yang termasuk di salah satu urutan peringkat pendapatan terendah atau tingkat kematian tertinggi (tidak termasuk pada 2 kategori variabel secara bersamaan)
- 

# RECCOMMENDATION

Sesuai dengan tujuan proyek untuk memberikan bantuan finansial kepada negara yang membutuhkan berdasarkan faktor sosial ekonomi dan kesehatan, maka kita memprioritaskan pada faktor pendapatan terendah dan tingkat kematian anak tertinggi. Adapun setelah diurutkan sesuai faktor pendapatan terendah, negara Congo ada di peringkat terendah. Sedangkan dari faktor tingkat kematian tertinggi, maka Haiti menduduki peringkat pertama.

Selanjutnya kita juga dapat mengambil negara yang termasuk dalam kategori 10 negara pendapatan terendah dan 10 negara dengan tingkat kematian anak tertinggi. Maka kita dapat merekomendasikan negara sbb:

- Niger
- Central African Republik
- Sierra Leone

# CREATOR



Jane Tamara Setiadi

*Data Analysis Enthusiast*

[www.linkedin.com/in/jane-tamara](https://www.linkedin.com/in/jane-tamara)



**THANK'S FOR  
WATCHING**

