FINAL PROJECT SANBERCODE PYTHON DATA SCIENCE BATCH 30

CLUSTERING NEGARA MENGGUNAKAN FAKTOR SOSIAL EKONOMI DAN KESEHATAN UNTUK PENDANAAN HELP INTERNATIONAL

by Dyaz Aerlangga

Tentang Organisasi

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 dan bencana alam.

Tujuan

Clustering negara ini ditujukan untuk menentukan negara mana yang membutuhkan bantuan dari HELP International dengan mengategorikan 167 negara menggunakan faktor sosial ekonomi dan kesehatan yang memengaruhi pembangunan negara secara keseluruhan.

Permasalahan

HELP International telah berhasil mengumpulkan sekitar $10 juta untuk pendanaan bantuan terhadap negara yang membutuhkan. CEO HELP Internasional perlu memutuskan bagaimana penggunaan uang tersebut agar strategis dan efektif.

READING AND UNDERSTANDING DATA

Syntax Pembacaan Dataset

```
[2]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.preprocessing import StandardScaler
     from sklearn.cluster import KMeans
```

```
[3]  df = pd.read_csv('Data_Negara_HELP.csv')
     df = df.rename(columns={'Kematian_anak':'Kematian Anak', 'Harapan_hidup':'Harapan Hidup','Jumlah_fertiliti':'Jumlah Fertilitas','GDPperkapita':'GDP pe
     display(df)
     df.info()
```

Result Data Tabular

	Negara	Kematian Anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan Hidup	Jumlah Fertilitas	GDP per Kapita
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

Terdapat 167 rows (list negara) dan 10 kolom yang berisikan Negara beserta aspek sosial ekonomi (Ekspor, Impor, Inflasi, Pendapatan, GDP per Kapita) dan Kesehatan (Kematian Anak, Kesehatan, Harapan Hidup, Jumlah Fertilitas).

Result 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 Fertilitas   167 non-null    float64
 9   GDP per Kapita      167 non-null    int64  
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

Dari hasil df.info() di samping, terlihat bahwa tidak terdapat missing value dari keseluruhan row dan column pada dataset tersebut.

Syntax dan Result df.describe()

```
[4]  df.describe()
```

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

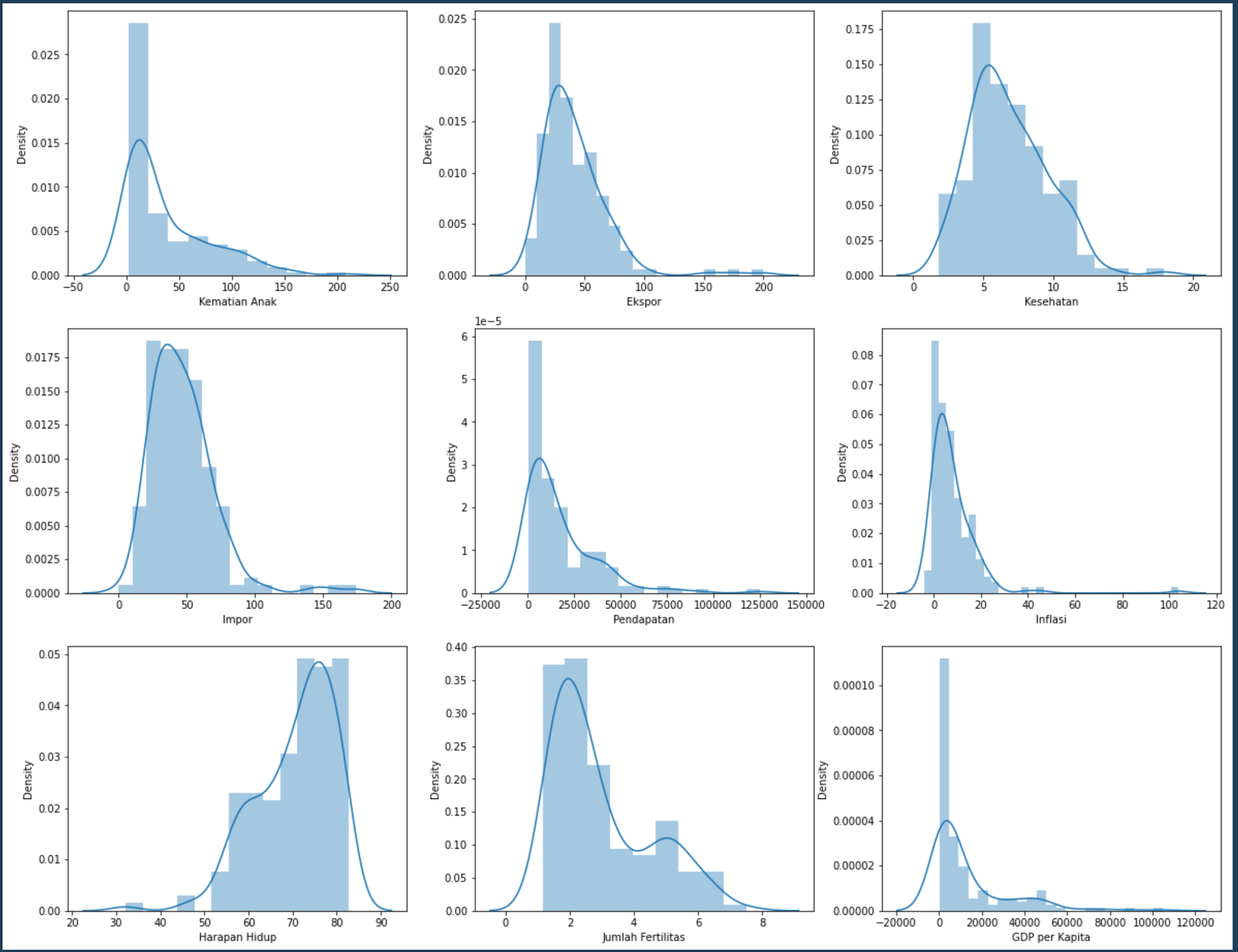
Dari hasil df.describe() di atas dapat dilihat rangkuman statisik per kolom yang memiliki tipe data float dan integer.

EXPLORATORY DATA ANALYSIS

Syntax Univariate Analysis

```
plt.figure(figsize=(20,16))

for i in enumerate(df.describe().columns):
  plt.subplot(3,3, i[0]+1)
  sns.distplot(df[i[1]])
plt.show()
```

Result Univariate Analysis



Hasil Univariate Analysis di samping menampilkan sebaran data dari kesembilan kolom dataset. Dapat dilihat bahwa tidak semua data simetris, dan masing-masing data tersebut memilki outliers yang harus dilakukan penanganan outliers.

Syntax dan Result variabel baru 'kematian'

```
kematian = df.sort_values('Kematian Anak', ascending=False)
kematian.head(10)
```

	Negara	Kematian Anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan Hidup	Jumlah Fertilitas	GDP per Kapita
66	Haiti	208.0	15.3	6.91	64.7	1500	5.45	32.1	3.33	662
132	Sierra Leone	160.0	16.8	13.10	34.5	1220	17.20	55.0	5.20	399
32	Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446
97	Mali	137.0	22.8	4.98	35.1	1870	4.37	59.5	6.55	708
113	Nigeria	130.0	25.3	5.07	17.4	5150	104.00	60.5	5.84	2330
112	Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
25	Burkina Faso	116.0	19.2	6.74	29.6	1430	6.81	57.9	5.87	575
37	Congo, Dem. Rep.	116.0	41.1	7.91	49.6	609	20.80	57.5	6.54	334

Variabel 'kematian' ini ditujukan untuk melihat negara mana yang memiliki angka tertinggi pada kolom 'Kematian Anak'.

Syntax dan Result variabel baru 'gdp'

```
gdp = df.sort_values('GDP per Kapita', ascending=True)
gdp.head(10)
```

	Negara	Kematian Anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan Hidup	Jumlah Fertilitas	GDP per Kapita
26	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231
88	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334
112	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348
132	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399
93	Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413
106	Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419
31	Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446
94	Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459
50	Eritrea	55.2	4.79	2.66	23.3	1420	11.60	61.7	4.61	482

Variabel 'gdp' ini ditujukan untuk melihat negara mana yang memiliki angka terendah pada kolom 'GDP per Kapita'.

Syntax dan Result Bivariate Analysis

```
anak_mati = df['Kematian Anak'] > 100
gdp_rendah = df['GDP per Kapita'] < 500

kematian_gdp = df[(anak_mati) & (gdp_rendah)].sort_values('Kematian Anak', ascending=False)
display(kematian_gdp)
```

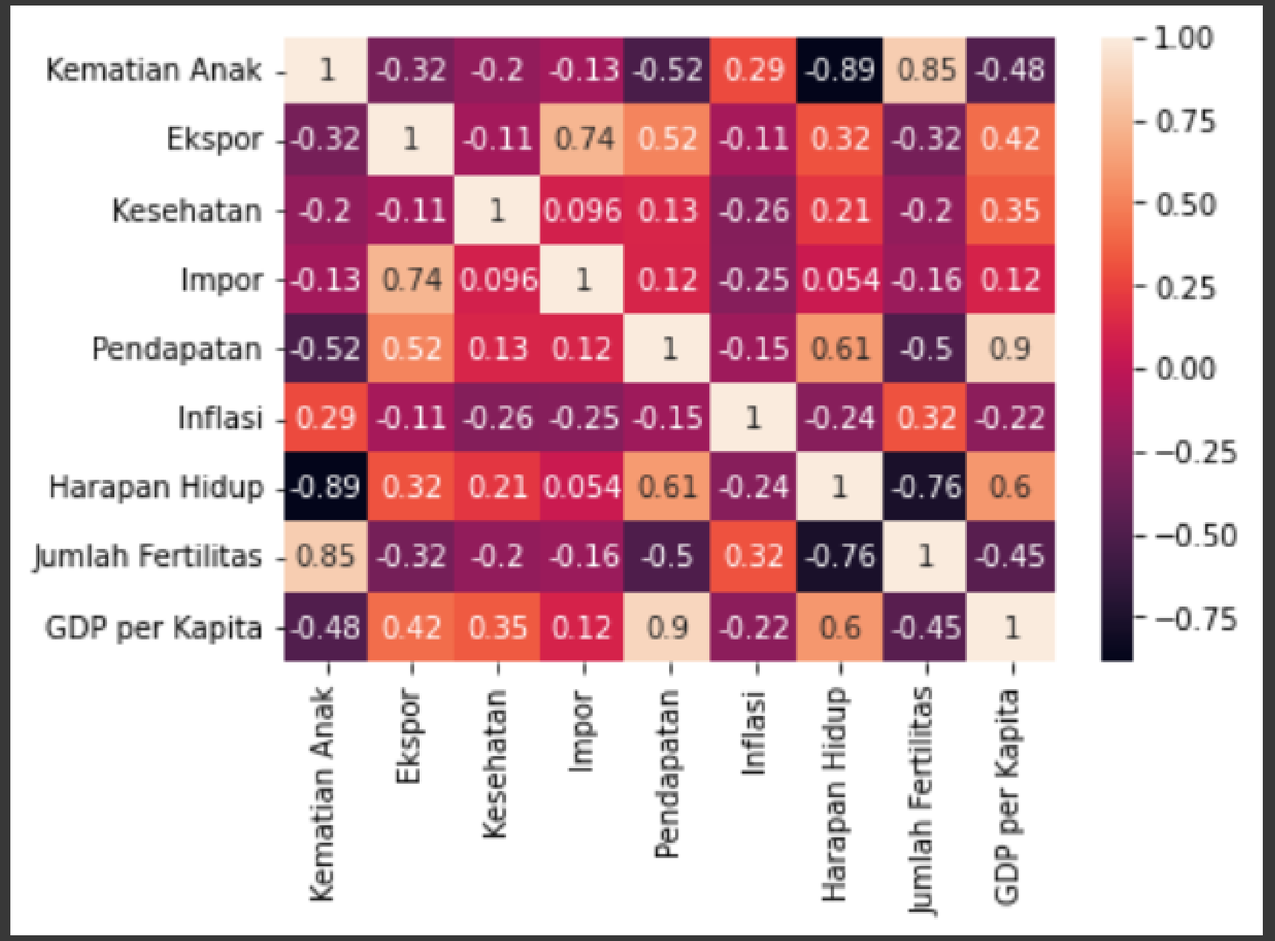
	Negara	Kematian Anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan Hidup	Jumlah Fertilitas	GDP per Kapita
132	Sierra Leone	160.0	16.8	13.10	34.5	1220	17.20	55.0	5.20	399
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446
112	Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348
37	Congo, Dem. Rep.	116.0	41.1	7.91	49.6	609	20.80	57.5	6.54	334
106	Mozambique	101.0	31.5	5.21	46.2	918	7.64	54.5	5.56	419

Bivariate Analysis ini diambil dari 2 variabel baru hasil sorting kolom 'Kematian Anak' yang memiliki nilai lebih dari 100 dan sorting kolom 'GDP per Kapita' yang memiliki nilai kurang dari 500.

Dengan gabungan kedua variabel baru tersebut ditemukan 4 negara yang dapat dikategorikan sebagai negara miskin sehingga berdampak pada tingginya kematian anak di negara tersebut.

Syntax dan Result Heatmap

```
sns.heatmap(df.corr(), annot =True, fmt='.2g');
```



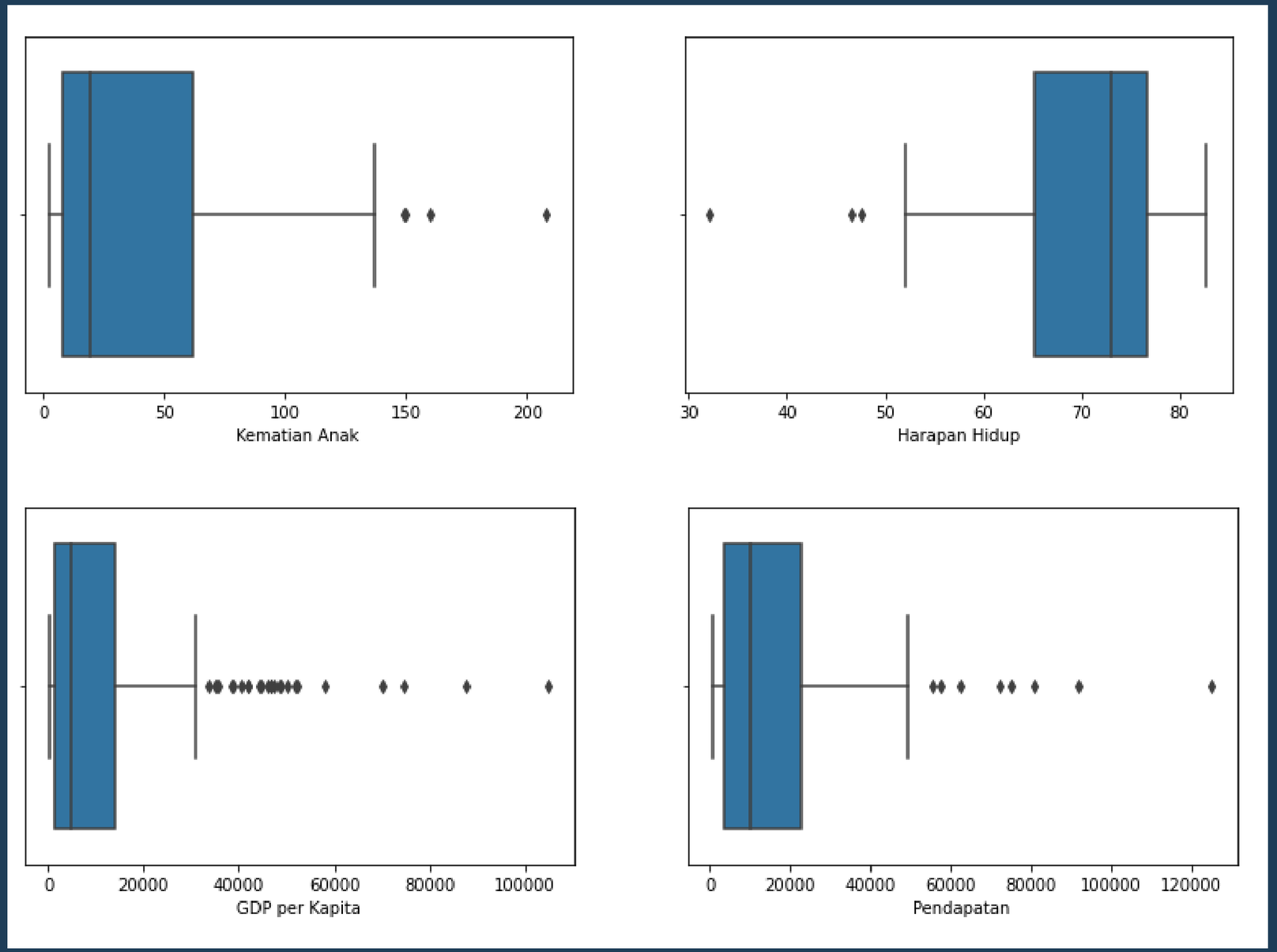
Heatmap disamping merupakan salah satu tools untuk melihat korelasi antar 2 variabel. Semakin cerah atau semakin mendekati angka 1, maka korelasi antar 2 variabel tersebut kuat. Semakin gelap atau semakin rendah angka tersebut maka menunjukkan korelas terbalik antar 2 variabel tersebut.

Dari heatmap pada slide sebelumnya, untuk aspek kesehatan didapat bahwa 'Kematian Anak' memiliki korelasi yang kuat dengan 'Jumlah Fertilitas' yaitu di angka 0.85, sedangkan memilki korelasi terbalik yang kuat dengan 'Harapan Hidup' di angka -0.89.

Untuk aspek ekonomi didapat bahwa ''Pendapatan' memiliki korelasi yang kuat dengan 'GDP per Kapita' yaitu di angka 0.9.

OUTLIERS

Boxplot Outliers



Dari hasil boxplot outliers pada slide sebelumnya, ditemukan outliers yang cukup banyak dan jauh dari batas atas dan batas bawah. Penanganan outliers diperlukan dan dalam kasus ini digunakan metode penggantian nilai outliers dengan nilai batas atas dan batas bawah tergantung pada outliers kolom tersebut.

Syntax dan Result Handling Outliers

```
[15] def remove_outlier(df):
       Q1 = df.quantile(0.25)
       Q3 = df.quantile(0.75)
       IQR = Q3-Q1
       df_final = df[~((df<(Q1-(1.5*IQR))) | (df>(Q3+(1.5*IQR))))]
       return df_final
```

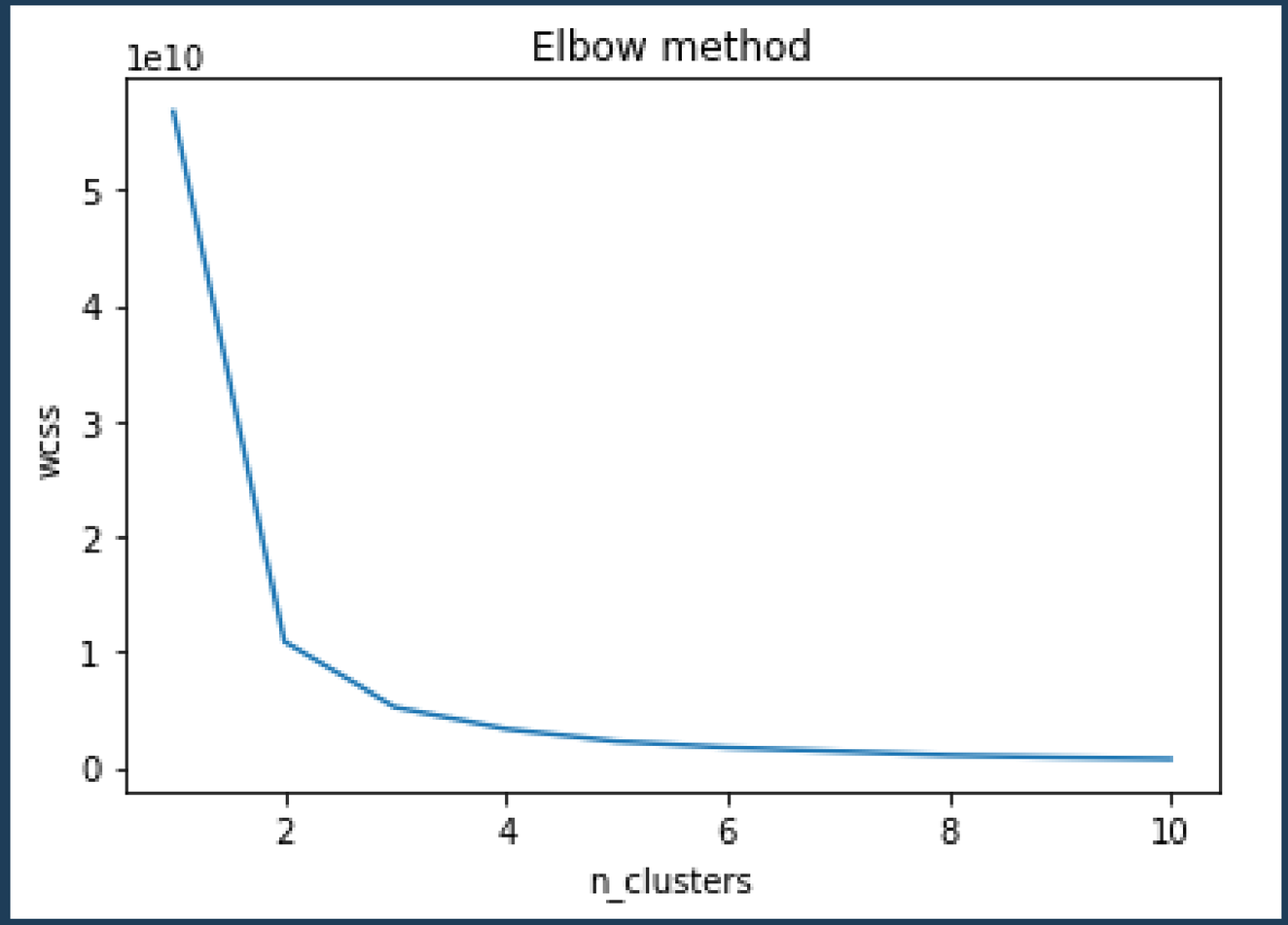
```
[16] df2 = remove_outlier(df[['Kematian Anak','Pendapatan','Harapan Hidup','Jumlah Fertilitas','GDP per Kapita']])
     df2.fillna(df2[['Kematian Anak','Pendapatan','Jumlah Fertilitas','GDP per Kapita']].max(),axis=0, inplace=True)
     df2.fillna(df2[['Harapan Hidup']].min(),axis=0, inplace=True)
```

```
[17] df2.shape

     (167, 5)
```

```
[18] print(finding_outlier(df2['Kematian Anak']))
     print(finding_outlier(df2['Pendapatan']))
     print(finding_outlier(df2['Harapan Hidup']))
     print(finding_outlier(df2['Jumlah Fertilitas']))
     print(finding_outlier(df2['GDP per Kapita']))

     Series([], Name: Kematian Anak, dtype: float64)
     Series([], Name: Pendapatan, dtype: float64)
     Series([], Name: Harapan Hidup, dtype: float64)
     Series([], Name: Jumlah Fertilitas, dtype: float64)
     Series([], Name: GDP per Kapita, dtype: float64)
```

CLUSTERING

Elbow Method



Elbow method disamping berguna untuk menujukkan berapa banyak cluster yang baik untuk digunakan dalam clustering. Dari hasil disamping ini ditemukan bahwa banyak cluster yang dapat digunakan untuk clustering adalah 2 - 3 cluster.

Syntax Clustering Aspek Kesehatan K = 2

```
[20] data_cluster = df2[['Kematian Anak','Harapan Hidup']]
```

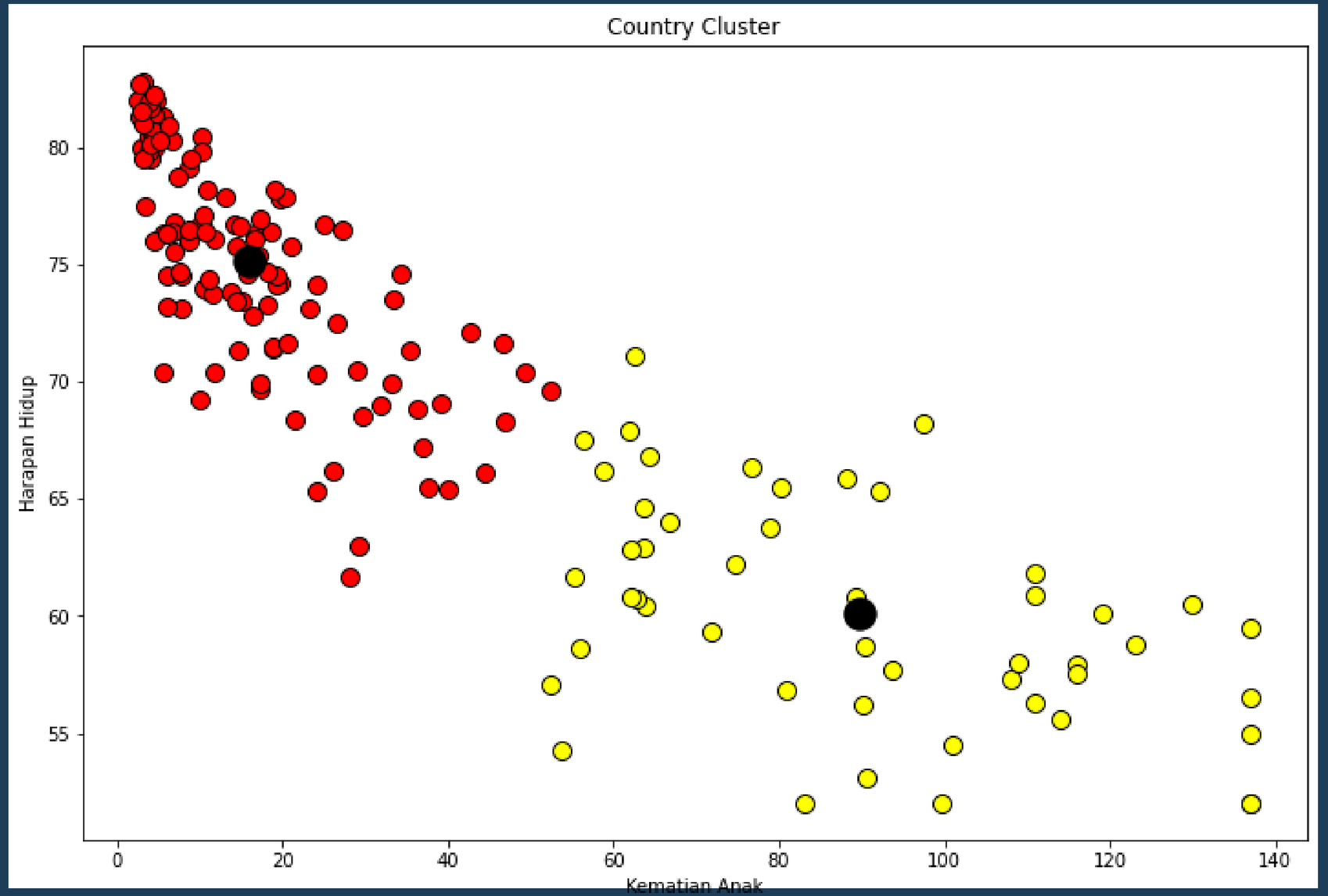
```
[21] from sklearn.cluster import KMeans

     kmeans1 = KMeans(n_clusters = 2, random_state=42).fit(data_cluster)
     labels1 = kmeans1.labels_
     labels1
     kmeans1
```

```
     KMeans(n_clusters=2, random_state=42)
```

```
[22] new_df = pd.DataFrame(data=data_cluster)
     new_df['label1_kmeans'] = labels1
     new_df
```

```
plt.figure(figsize=(12,8))

plt.scatter(new_df['Kematian Anak'][new_df.label1_kmeans == 0], new_df['Harapan Hidup'][new_df.label1_kmeans == 0], c='red', s=100, edgecolor='black')
plt.scatter(new_df['Kematian Anak'][new_df.label1_kmeans == 1], new_df['Harapan Hidup'][new_df.label1_kmeans == 1], c='yellow', s=100, edgecolor='bla

plt.scatter(kmeans1.cluster_centers_[:, 0],kmeans1.cluster_centers_[:, 1], c='k', s = 300)
plt.title('Country Cluster')
plt.xlabel('Kematian Anak')
plt.ylabel('Harapan Hidup')
plt.show()
```

Result Clustering Aspek Kesehatan K = 2



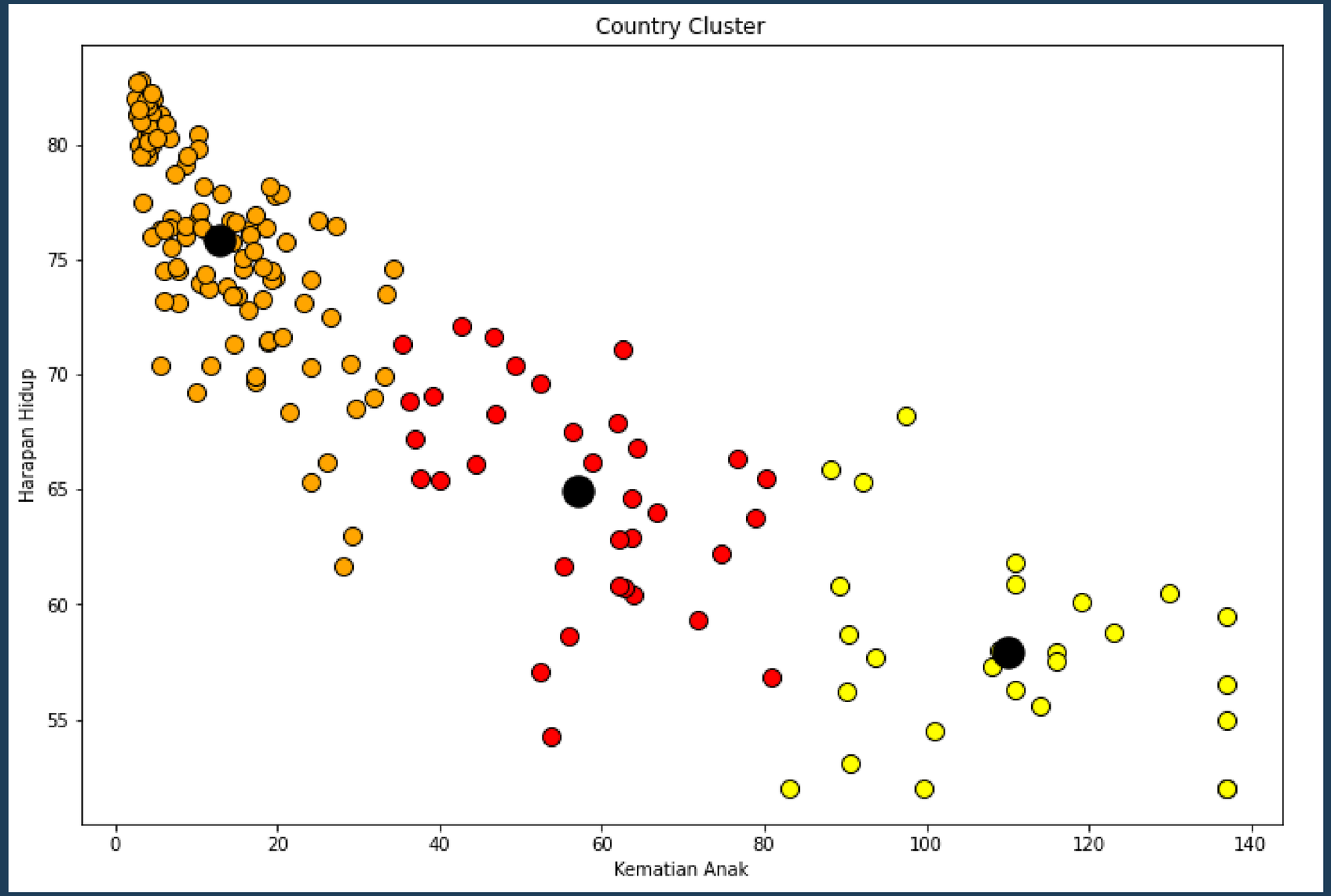
Syntax Clustering Aspek Kesehatan K = 3

```
[68] kmeans2 = KMeans(n_clusters = 3, init='k-means++', random_state=42).fit(data_cluster)
     labels2 = kmeans2.labels_
```

```
[88] new_df['label2_kmeans'] = labels2
```

```
[70] plt.figure(figsize=(12,8))

     plt.scatter(new_df['Kematian Anak'][new_df.label2_kmeans == 0], new_df['Harapan Hidup'][new_df.label2_kmeans == 0], c='red', s=100, edgecolor='black')
     plt.scatter(new_df['Kematian Anak'][new_df.label2_kmeans == 1], new_df['Harapan Hidup'][new_df.label2_kmeans == 1], c='yellow', s=100, edgecolor='bla
     plt.scatter(new_df['Kematian Anak'][new_df.label2_kmeans == 2], new_df['Harapan Hidup'][new_df.label2_kmeans == 2], c='orange', s=100, edgecolor='bla

     plt.scatter(kmeans2.cluster_centers_[:, 0],kmeans2.cluster_centers_[:, 1], c='k', s=300 )
     plt.title('Country Cluster')
     plt.xlabel('Kematian Anak')
     plt.ylabel('Harapan Hidup')
     plt.show()
```

Result Clustering Aspek Kesehatan K = 3



Dari hasil kedua clustering aspek kesehatan dengan mengambil variabel dari kolom 'Kematian Anak' dan 'Harapan Hidup', ditemukan bahwa semakin tinggi nya angka kematian anak maka semakin rendah harapan hidup di negara tersebut.

Syntax dan Result Silhoutte Score Aspek Kesehatan K = 2 dan K = 3

```
[116] from sklearn.metrics import silhouette_score
      print(silhouette_score(data_cluster, labels=labels1))
      print(silhouette_score(data_cluster, labels=labels2))
```

```
0.7089776790783209
0.6551138724013621
```

Dari hasil silhoutte score di atas, ditemukan bahwa jumlah cluster k = 2 memberikan nilai yang lebih akurat dimana semakin mendekati angka 1 maka clustering tersebut dinilai semakin baik.

Syntax Clustering Aspek Ekonomi K = 2

```
[123] data_cluster3 = df2[['GDP per Kapita','Pendapatan']]
```

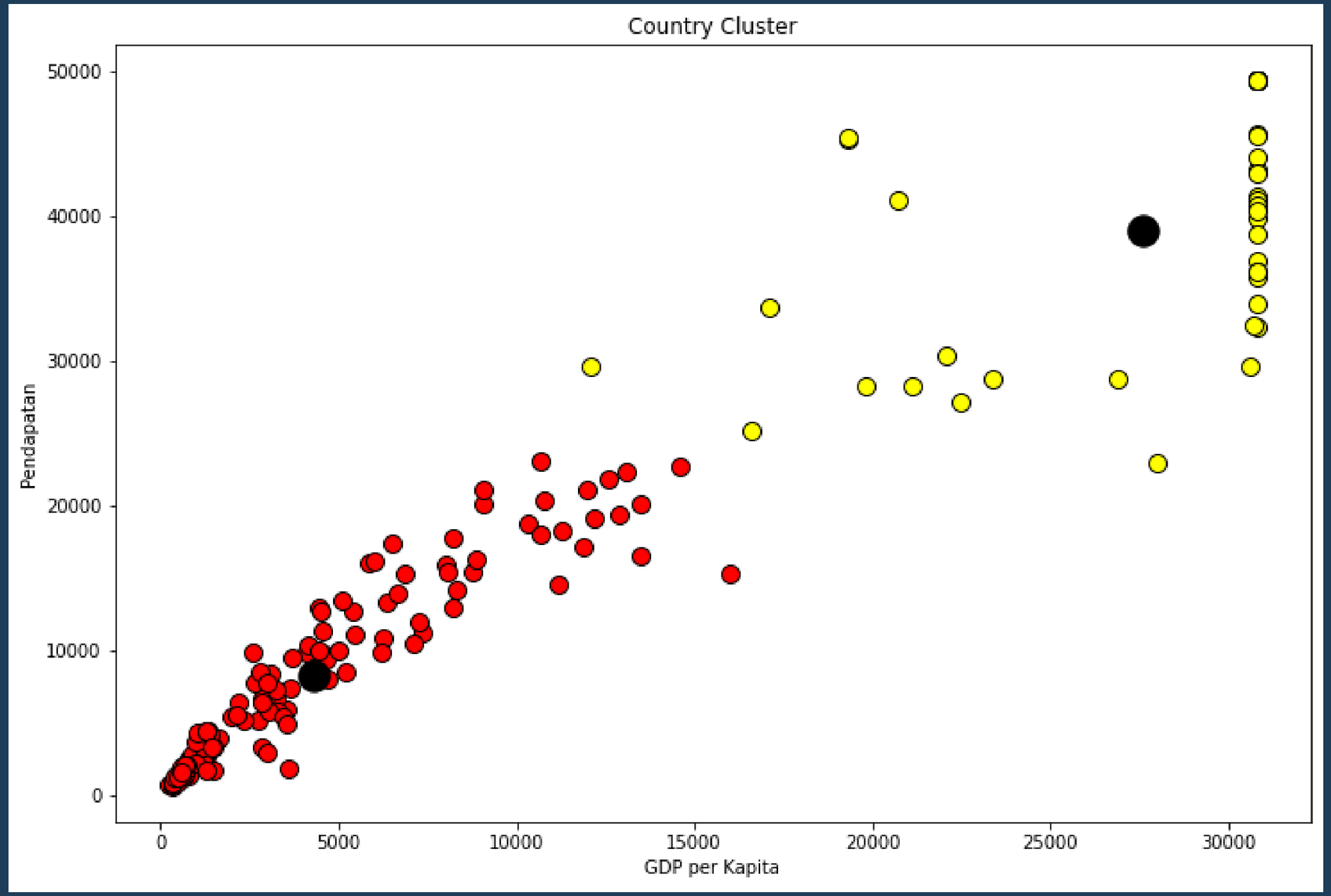
```
[124] kmeans4 = KMeans(n_clusters = 2, random_state=42).fit(data_cluster3)
      labels4 = kmeans4.labels_
      labels4
      kmeans4
```

```
      KMeans(n_clusters=2, random_state=42)
```

```
[125] new_df_eko = pd.DataFrame(data=data_cluster3)
      new_df_eko['label4_kmeans'] = labels4
      new_df_eko
```

```
[126] plt.figure(figsize=(12,8))

      plt.scatter(new_df_eko['GDP per Kapita'][new_df_eko.label4_kmeans == 0], new_df_eko['Pendapatan'][new_df_eko.label4_kmeans == 0], c='red', s=100, edg
      plt.scatter(new_df_eko['GDP per Kapita'][new_df_eko.label4_kmeans == 1], new_df_eko['Pendapatan'][new_df_eko.label4_kmeans == 1], c='yellow', s=100,

      plt.scatter(kmeans4.cluster_centers_[:, 0],kmeans4.cluster_centers_[:, 1], c='k', s = 300)
      plt.title('Country Cluster')
      plt.xlabel('GDP per Kapita')
      plt.ylabel('Pendapatan')
      plt.show()
```

Result Clustering Aspek Ekonomi K = 2



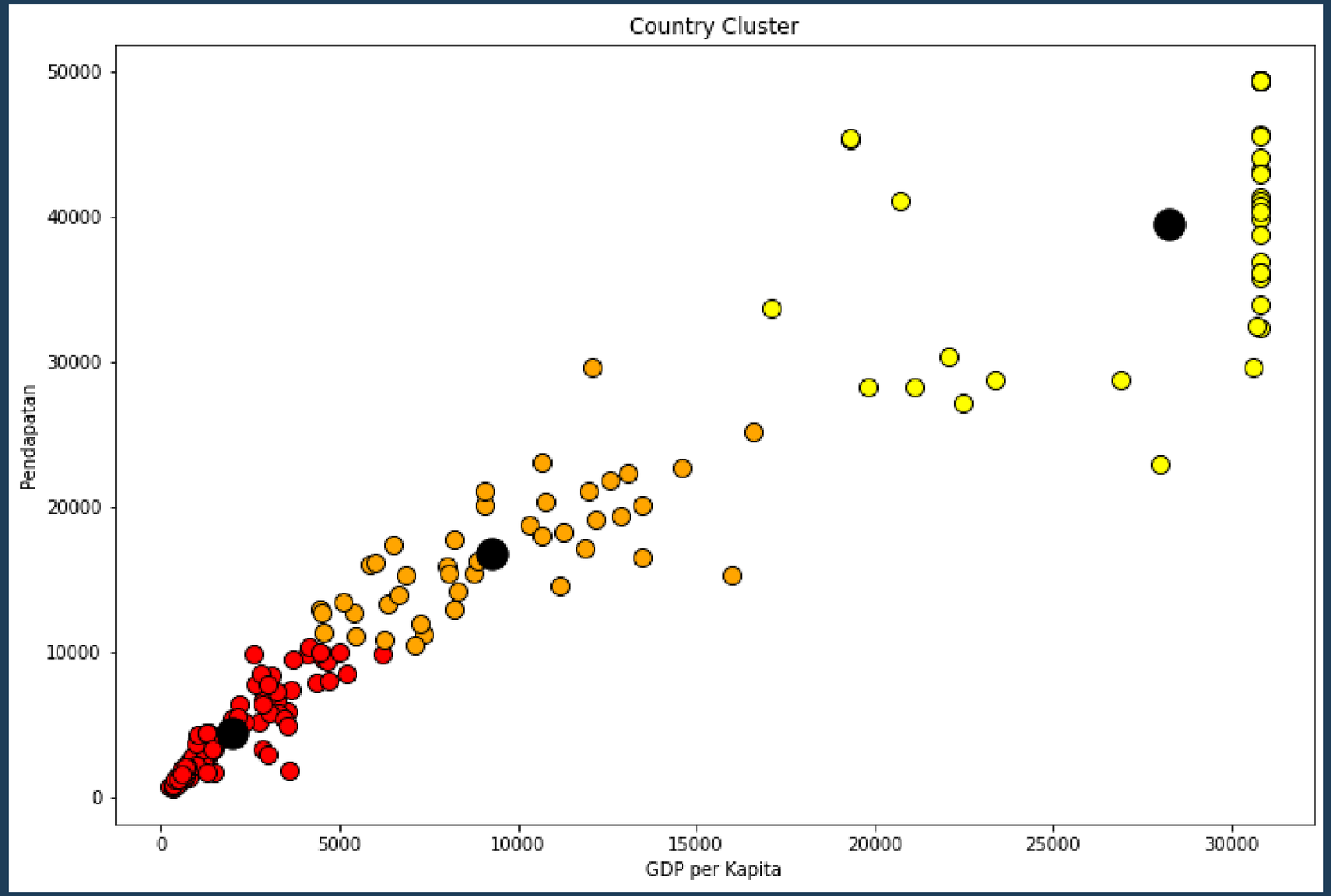
Syntax Clustering Aspek Ekonomi K = 3

```
kmeans5 = KMeans(n_clusters = 3, init='k-means++', random_state=42).fit(data_cluster3)
labels5 = kmeans5.labels_
```

```
[133] new_df_eko['label5_kmeans'] = labels5
```

```
[129] plt.figure(figsize=(12,8))

      plt.scatter(new_df_eko['GDP per Kapita'][new_df_eko.label5_kmeans == 0], new_df_eko['Pendapatan'][new_df_eko.label5_kmeans == 0], c='red', s=100, edg
      plt.scatter(new_df_eko['GDP per Kapita'][new_df_eko.label5_kmeans == 1], new_df_eko['Pendapatan'][new_df_eko.label5_kmeans == 1], c='yellow', s=100,
      plt.scatter(new_df_eko['GDP per Kapita'][new_df_eko.label5_kmeans == 2], new_df_eko['Pendapatan'][new_df_eko.label5_kmeans == 2], c='orange', s=100,

      plt.scatter(kmeans5.cluster_centers_[:, 0],kmeans5.cluster_centers_[:, 1], c='k', s=300 )
      plt.title('Country Cluster')
      plt.xlabel('GDP per Kapita')
      plt.ylabel('Pendapatan')
      plt.show()
```

Result Clustering Aspek Ekonomi K = 3



Dari hasil kedua clustering aspek ekonomi dengan mengambil variabel dari kolom 'GDP per Kapita' dan 'Pendapatan', ditemukan bahwa GDP per Kapita sangat berkaitan dengan Pendapatan yang dimana semakin kecil nya Pendapatan maka semakin kecil pula GDP per Kapita suatu negara.

Syntax dan Result Silhoutte Score Aspek Ekonomi K = 2 dan K = 3

```
[130] from sklearn.metrics import silhouette_score
      print(silhouette_score(data_cluster3, labels=labels4))
      print(silhouette_score(data_cluster3, labels=labels5))
```

```
0.7336172051216407
0.6234184889328376
```

Dari hasil silhoutte score di atas, ditemukan bahwa jumlah cluster k = 2 memberikan nilai yang lebih akurat yaitu di angka 0.73 dibandingkan dengan cluster k = 3 yang hanya berada di angka 0.62.

REPORT COUNTRIES

```
[132] kes = new_df[(cluster1) & (cluster2)].sort_values('Kematian Anak', ascending=False).head(10)
      eko = new_df_eko[(cluster4) & (cluster5)].sort_values('GDP per Kapita', ascending=True).head(10)

      overall = pd.merge(kes, eko, on='Negara', how='inner')
      overall = overall[['Negara','Kematian Anak','Harapan Hidup','GDP per Kapita','Pendapatan']]
      display(overall)
```

	Negara	Kematian Anak	Harapan Hidup	GDP per Kapita	Pendapatan
0	Sierra Leone	137.0	55.0	399.0	1220.0
1	Central African Republic	137.0	52.0	446.0	888.0
2	Niger	123.0	58.8	348.0	814.0
3	Congo, Dem. Rep.	116.0	57.5	334.0	609.0

Berdasarkan gabungan dari clustering aspek kesehatan dan ekonomi di atas, ditemukan ada 4 negara yang dapat direkomendasikan untuk mendapatkan pendanaan. Negara tersebut terpilih karena memiliki tingkat kematian anak yang tinggi dan juga merupakan negara miskin dengan tingkat GDP per Kapita yang rendah. 4 negara tersebut yaitu Sierra Leone, Central African Republic, Niger, dan Congo, Dem. Republic.