



# CLUSTERING THE COUNTRIES BY USING K-MEANS FOR HELP INTERNATIONAL

**TRIMITHA KARINA N.**

Oktober 2023

Data Science – Batch 50





# READ AND UNDERSTAND DATA

	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

Dari data tersebut terdapat 167 baris dan 10 kolom yang berisi variabel negara, kematian anak, ekspor, kesehatan, impor, pendapatan, inflasi, harapan hidup, jumlah fertiliti dan GDHperkapita.

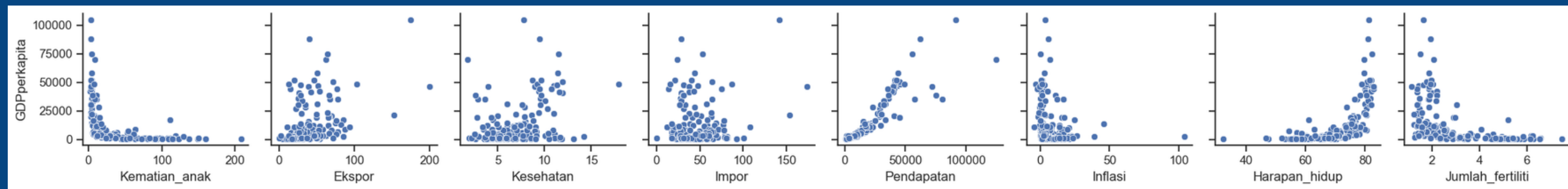
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Negara          167 non-null   object
1   Kematian_anak   167 non-null   float64
2   Ekspor          167 non-null   float64
3   Kesehatan       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
```

```
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
```

Dari 10 variabel yang ada, 7 diantaranya bertipe float, 2 bertipe integer dan 1 bertipe object. Dari kolom tersebut tidak ditemukan adanya *missing value*.

# EKSPLORASI DATA

## Univariate Analysis

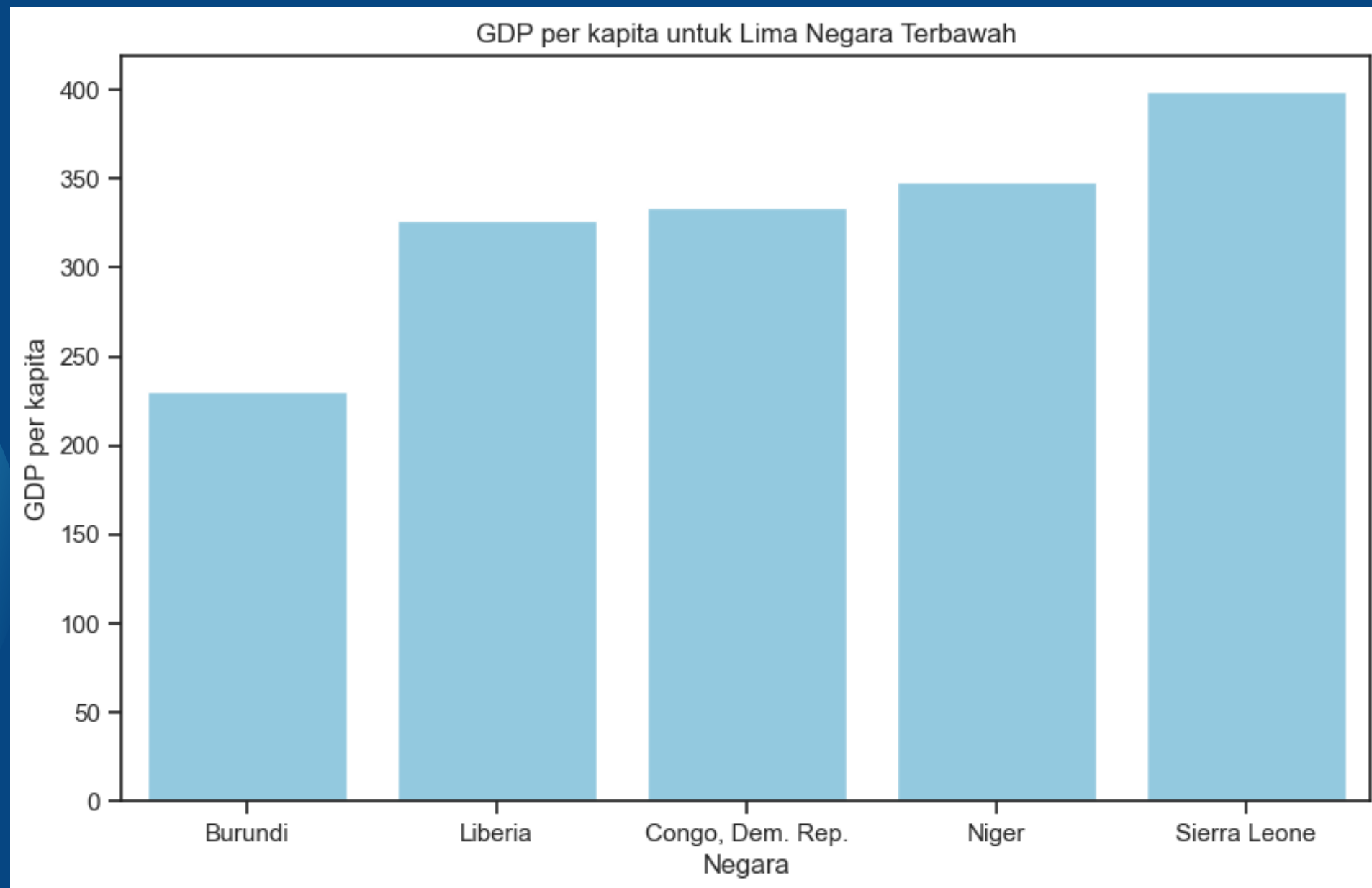


Dalam visualisasi ini kita dapat melihat hubungan antara variabel GDPperkapita dengan 8 variabel lainnya.

- Dalam variabel kematian anak, ekspor, kesehatan, impor, inflasi, harapan hidup dan jumlah fertiliti menunjukkan linear yang lemah antara beberapa variabel tersebut dan GDPperkapita.
- Dalam variabel pendapatan menunjukkan linear yang kuat antara variabel pendapatan dan GDPperkapita.

# EKSPLORASI DATA

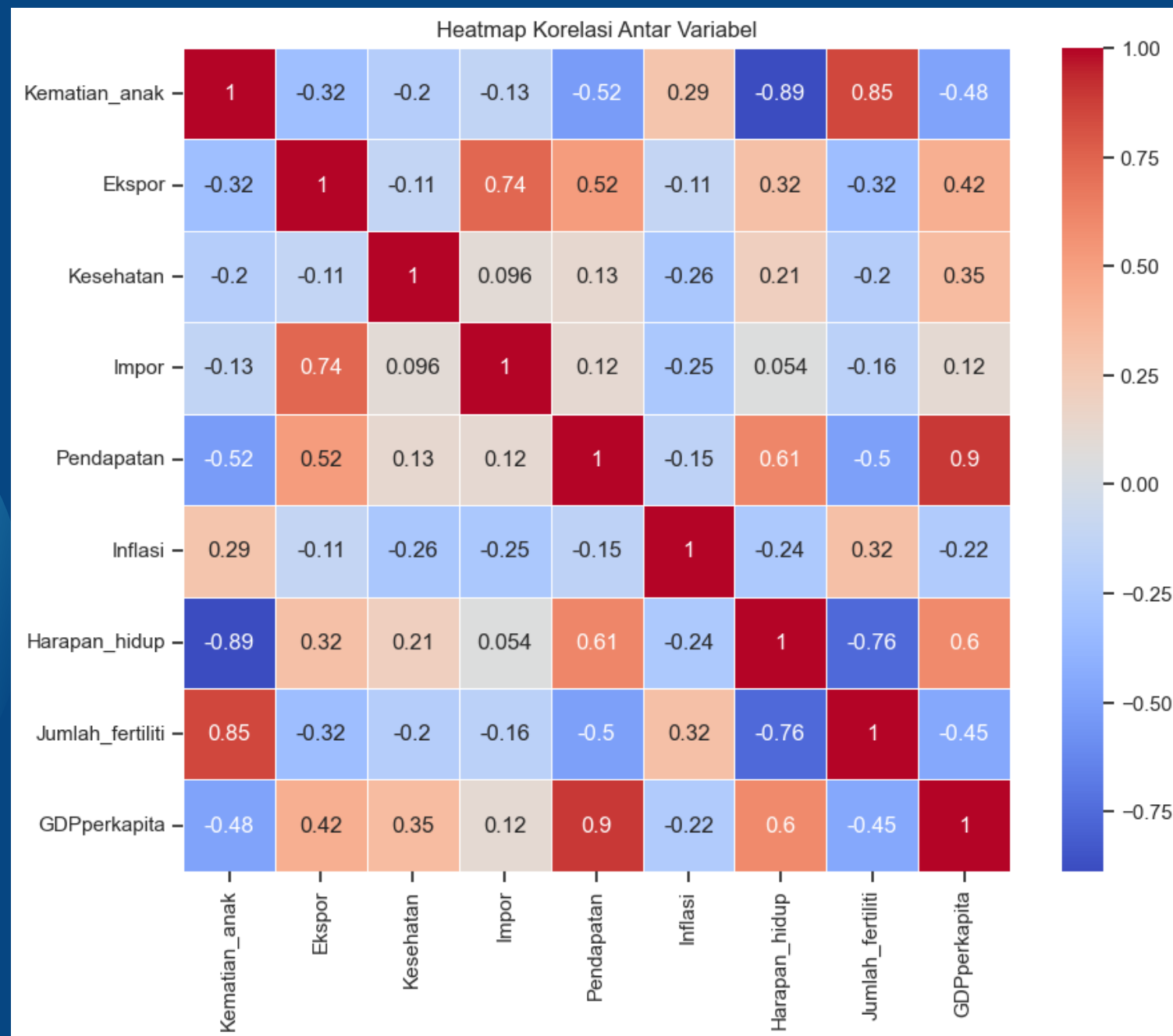
## Bivariate Analysis



Terlihat dalam nilai GDP per kapita untuk lima negara terbawah yaitu Burundi, Liberia, Congo, Dem.Rep, Niger dan Sierra Leone

# EKSPLORASI DATA

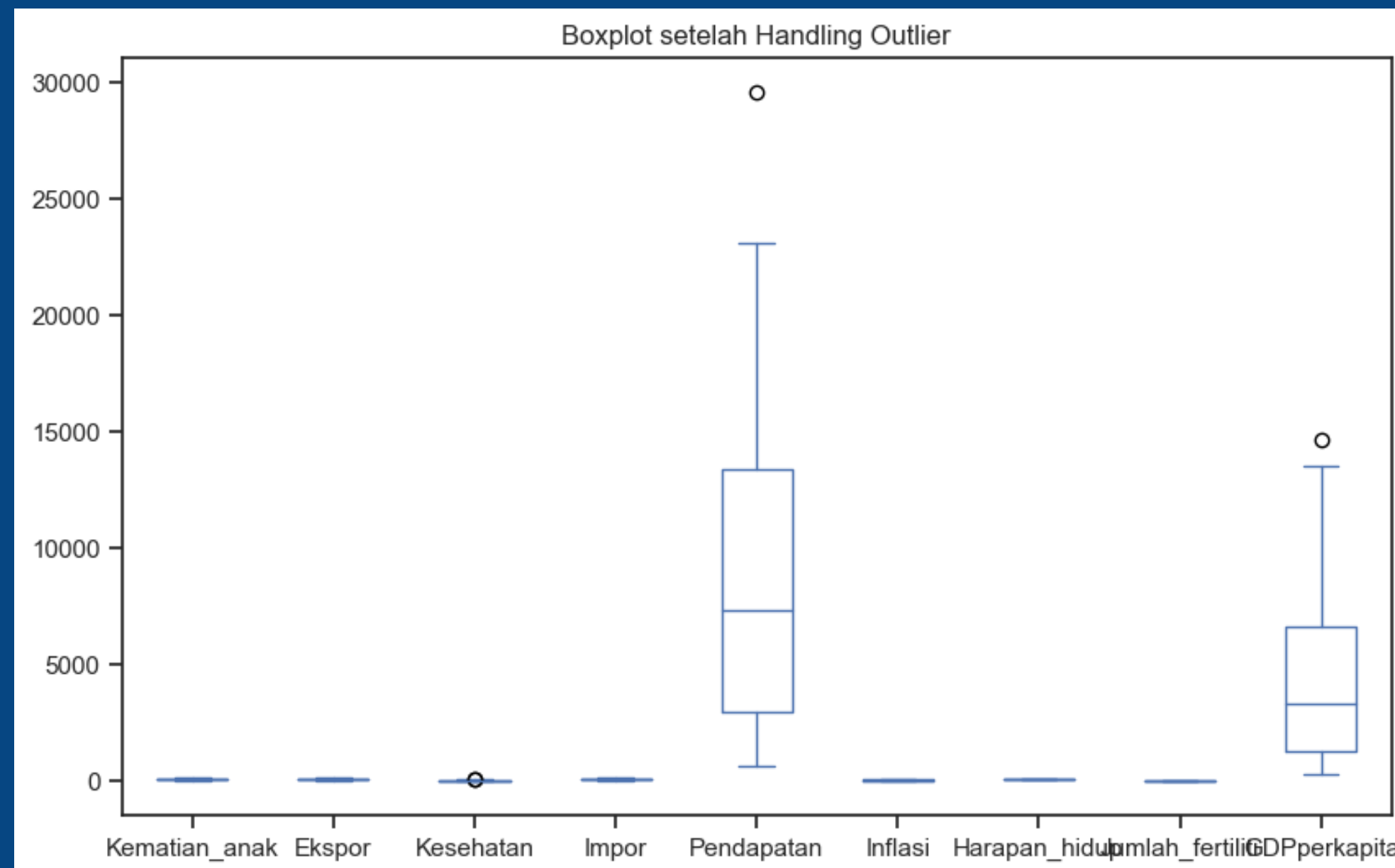
## Multivariate Analysis



- Dalam kematian anak menunjukkan korelasi yang kuat dengan variabel jumlah fertiliti.
- Jumlah ekspor memiliki korelasi yang kuat dengan variabel impor dan sebaliknya, variabel impor memiliki korelasi yang kuat dengan variabel ekspor.
- Pendapatan memiliki variabel yang kuat terhadap GDPperkapita dan sebaliknya.
- Jumlah terliti memiliki korelasi yang kuat dengan kematian anak.

# HANDLING OUTLIERS

Pada persebaran data yang melalui visualisasi univariate terlihat adanya outliers pada beberapa variabel, diperlukan adanya penanganan outliers.



# NORMALISASI DATA

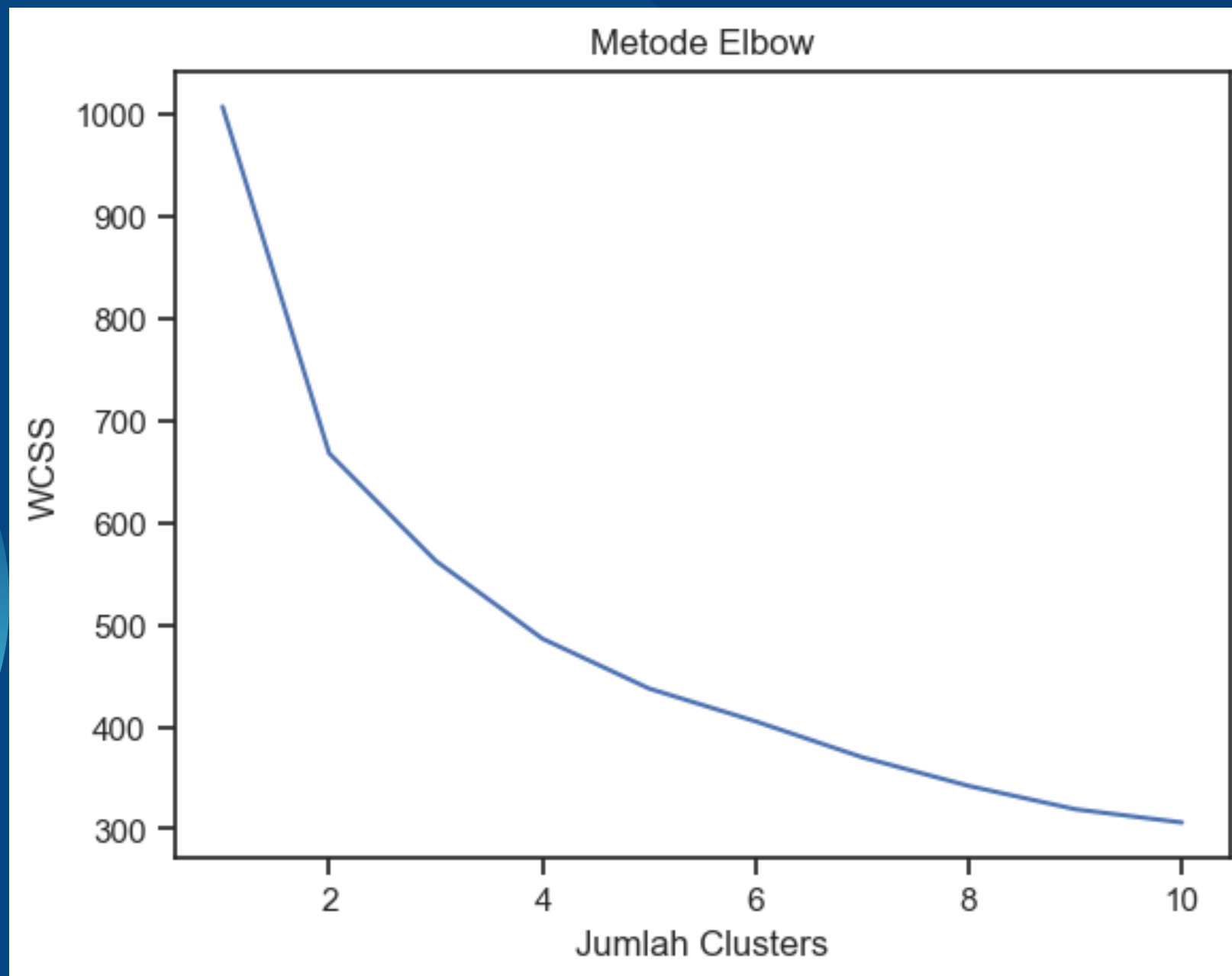
StandardScaler diperlukan untuk melakukan normalisasi data agar data yang digunakan tidak memiliki penyimpangan yang besar.

[illegible]

# CLUSTERING



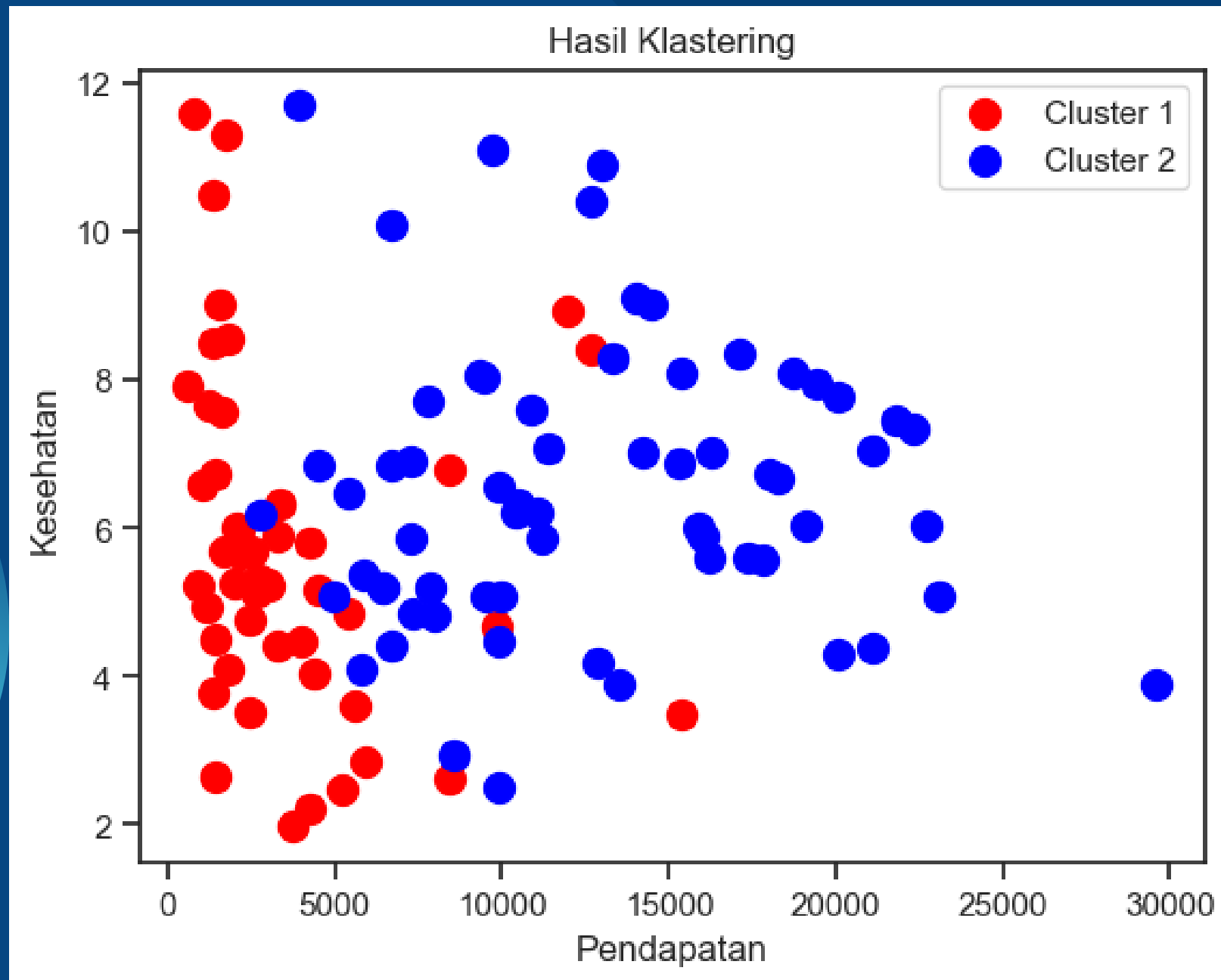
# METODE ELBOW



Terlihat bahwa jumlah cluster 2 memiliki distribusi data yang lebih seimbang antara klaster 0 dan klaster 1.

# HASIL CLUSTERING

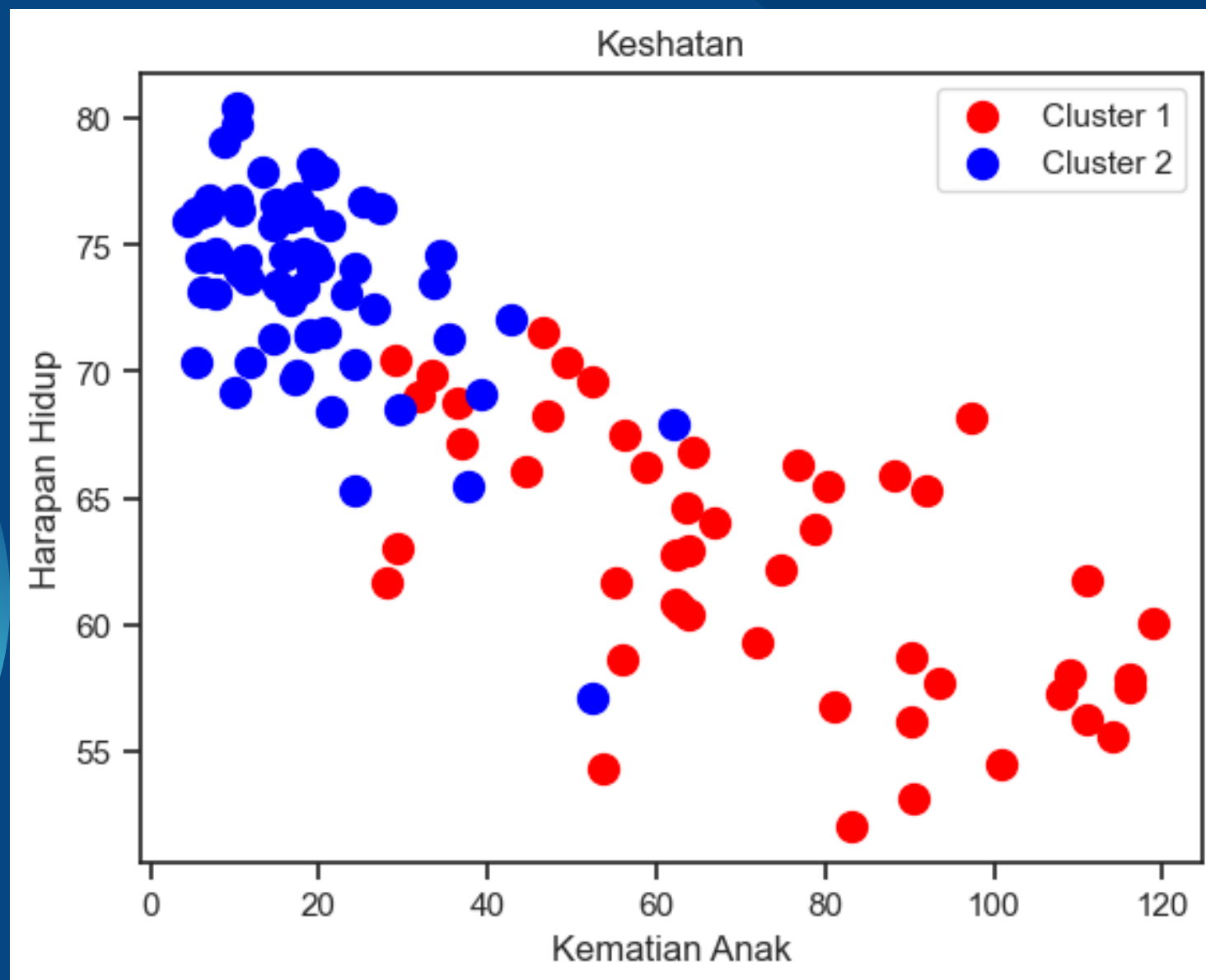
Mengidentifikasi Pendapatan dengan kesehatan



- Cluster 1 menunjukkan kombinasi nilai pendapatan yang rendah namun kesehatan yang tinggi. Hal ini dapat diinterpretasikan sebagai pola karakteristik yang menunjukkan adanya kemungkinan keberhasilan atau keunggulan dalam sistem kesehatan walaupun negara tersebut memiliki tingkat pendapatan yang rendah.
- Cluster 2 menggambarkan keragaman yang signifikan dalam hal kondisi sosial-ekonomi dan kesehatan antara negara-negara yang termasuk dalam klaster tersebut. Hal ini menunjukkan bahwa terdapat tantangan yang beragam yang perlu diatasi dalam upaya meningkatkan pembangunan dan kesejahteraan di negara-negara yang termasuk dalam klaster ini.

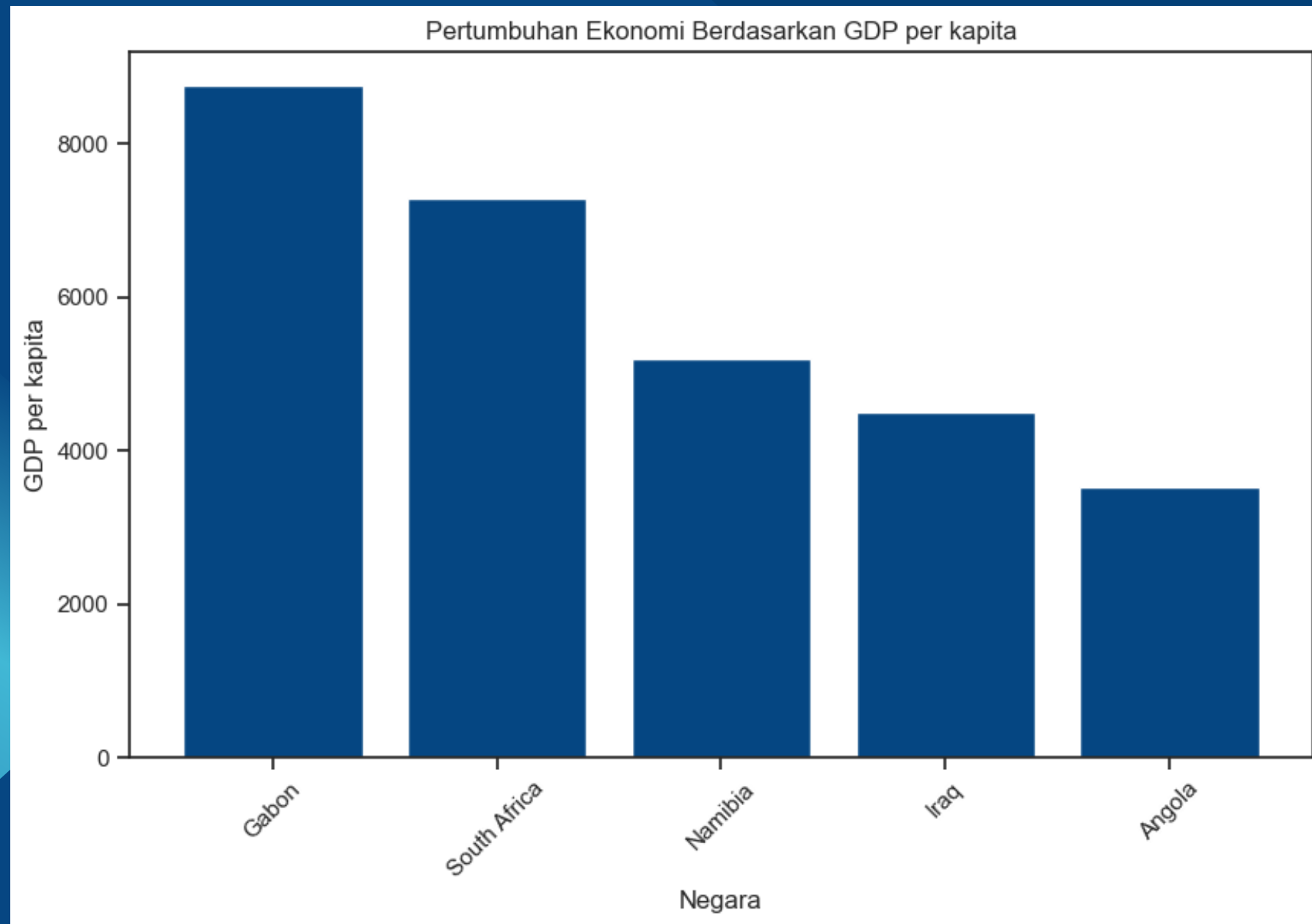
# HASIL CLUSTERING

Mengidentifikasi Harapan hidup dengan kematian anak



- Cluster 1 menunjukkan kombinasi nilai kematian anak yang tinggi maka menghasilkan angka harapan hidup yang rendah.
- Cluster 2 menggambarkan pola yang mana apabila jumlah kematian anak yang rendah maka nilai angka harapan hidup semakin tinggi

# KESIMPULAN



Grafik ini akan menampilkan kelima negara dengan nilai cluster 0 dan GDP per kapita tertinggi. Kelima negara tersebut, yaitu Gabon, Afrika Selatan, Namibia, Irak, dan Angola, membutuhkan perhatian khusus memiliki potensi untuk pertumbuhan ekonomi yang signifikan.

Dengan demikian, LSM tersebut dapat memberikan dampak yang lebih signifikan dalam membantu masyarakat yang kurang mampu di negara-negara tersebut.