FINAL PROJECT CLUSTERING THE COUNTRIES BY USING K-MEANS FOR HELP INTERNATIONAL

Dibuat oleh :
Nickolaus Satria Bagaskara

OBJECTIVE

Untuk mengkategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan.

PERMASALAHAN:

HELP International 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. Oleh karena itu, Tugas teman-teman adalah mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Kemudian kalian perlu menyarankan negara mana saja yang paling perlu menjadi fokus CEO.

DATA YANG DIGUNAKAN

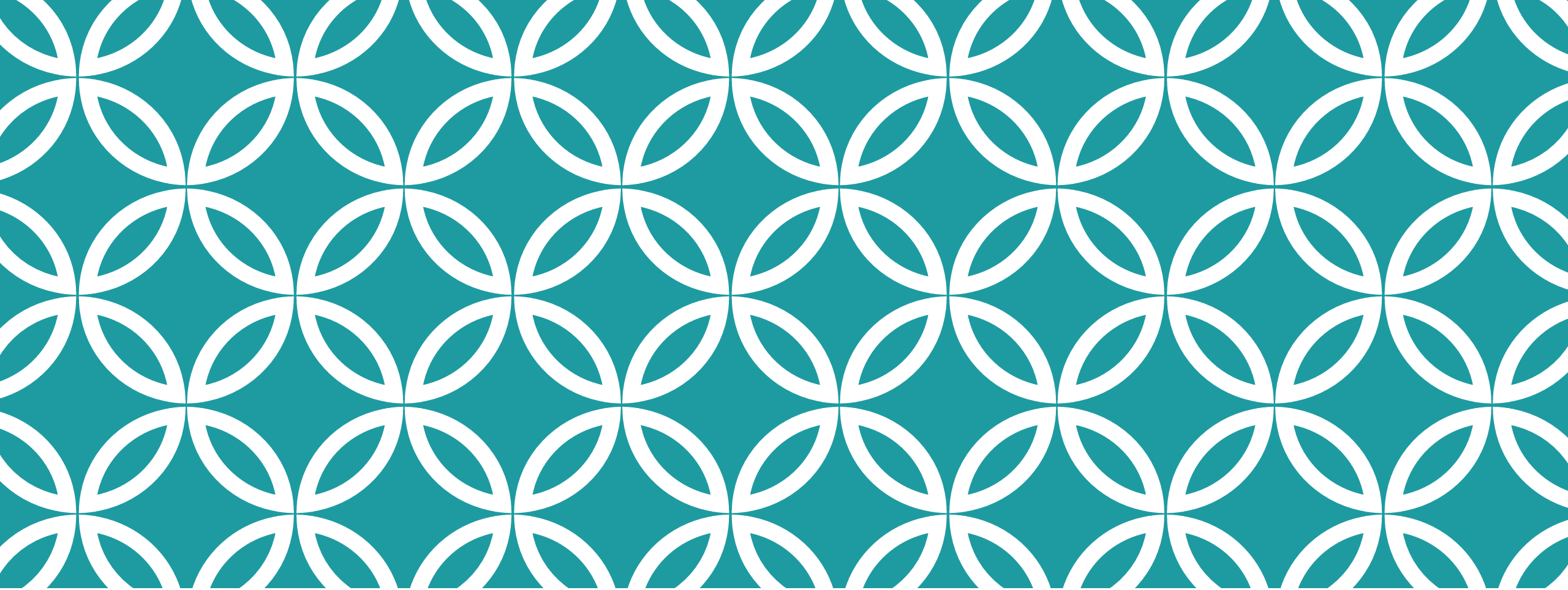
Data yang digunakan ialah data yang memiliki komponen :

- Negara : Nama negara
- Kematian_anak: Kematian anak di bawah usia 5 tahun per 1000 kelahiran
- Ekspor : Ekspor barang dan jasa perkapita
- Kesehatan: Total pengeluaran kesehatan perkapita
- Impor: Impor barang dan jasa perkapita
- Pendapatan: Penghasilan bersih perorang
- 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_fertiliti: Jumlah anak yang akan lahir dari setiap wanita jika tingkat kesuburan usia saat ini tetap sama
- GDPperkapita: GDP per kapita. Dihitung sebagai Total GDP dibagi dengan total populasi.

KETERANGAN DATA

- Data yang digunakan memiliki ukuran 167 x 10
- Data Negara memiliki jenis data object
- Data Kematian anak, Ekspor, Kesehatan, Import, Inflasi, Harapan hidup, dan Jumlah fertiliti memiliki jenis data float
- Data Pendapatan dan GDP perkapita memiliki jenis data integer
- Memori yang digunakan dari data ini ialah 13,2+ Kb

```
(167, 10)
<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
None
```



ANALYSIS

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
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

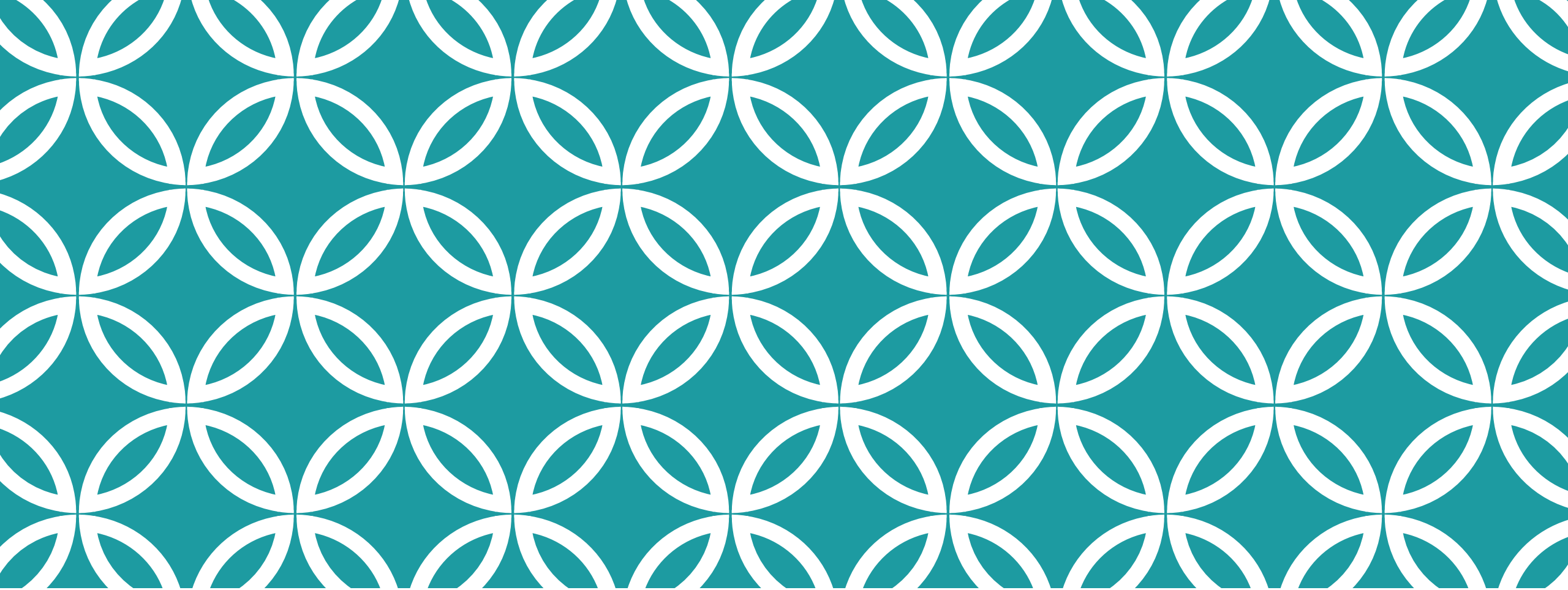
SUMMARY STATISTIK PADA DATA YANG DIGUNAKAN

Dari gambar diatas dapat dilihat nilai jumlah data (count), rata-rata data (mean), nilai standarisasi data (std), nilai paling rendah pada data (min), nilai paling tinggi pada data (max), nilai dari quartil bawah data (25%), nilai dari quartil tengah data (50%) , dan nilai dari quartil atas data (75%)

PENGECEKAN DATA KOSONG

Setelah di priksa dapat dilihat bahwa ada data kosong atau data yang bertumpuk pada data yang digunakan

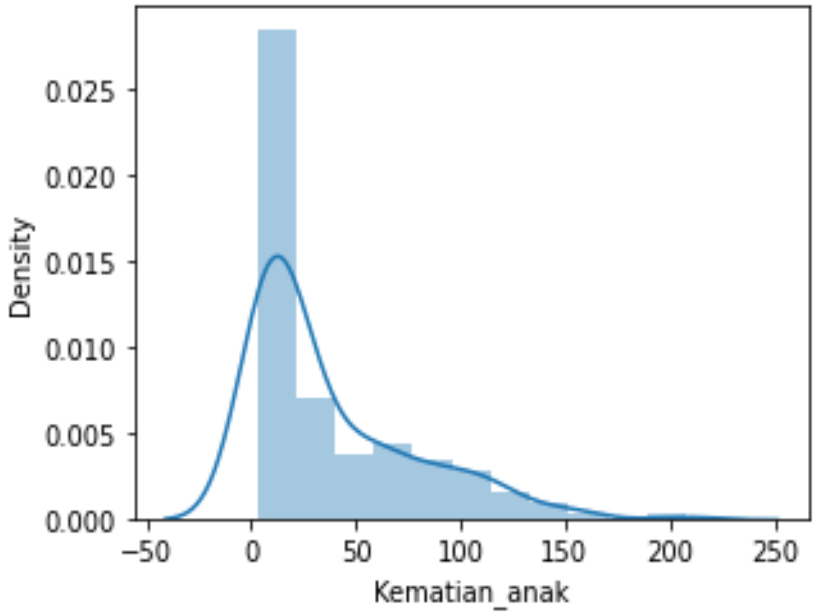
```
Negara              0
Kematian_anak       0
Ekspor              0
Kesehatan           0
Impor               0
Pendapatan          0
Inflasi             0
Harapan_hidup       0
Jumlah_fertiliti    0
GDPperkapita        0
dtype: int64
```

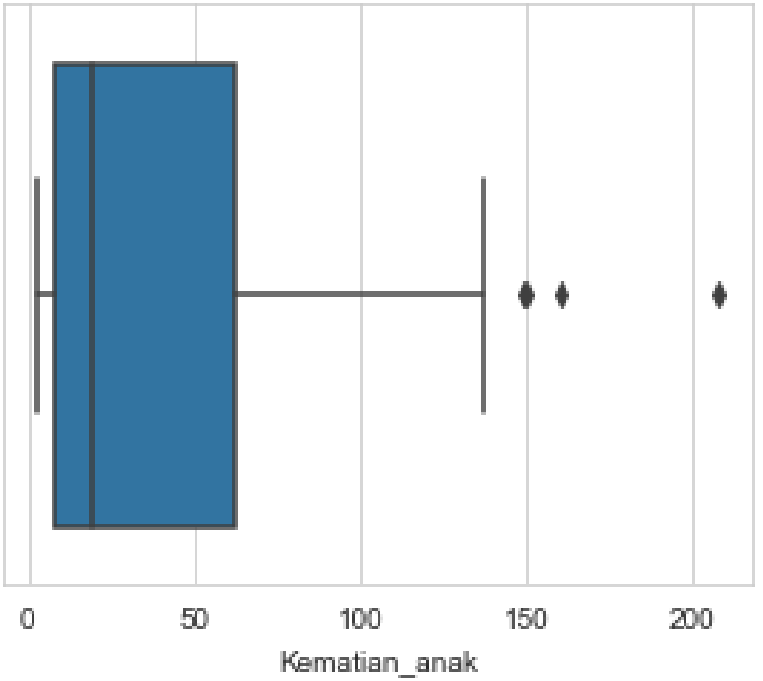


EDA UNIVARIATE ANALYSIS

ANALISIS DATA KEMATIAN ANAK

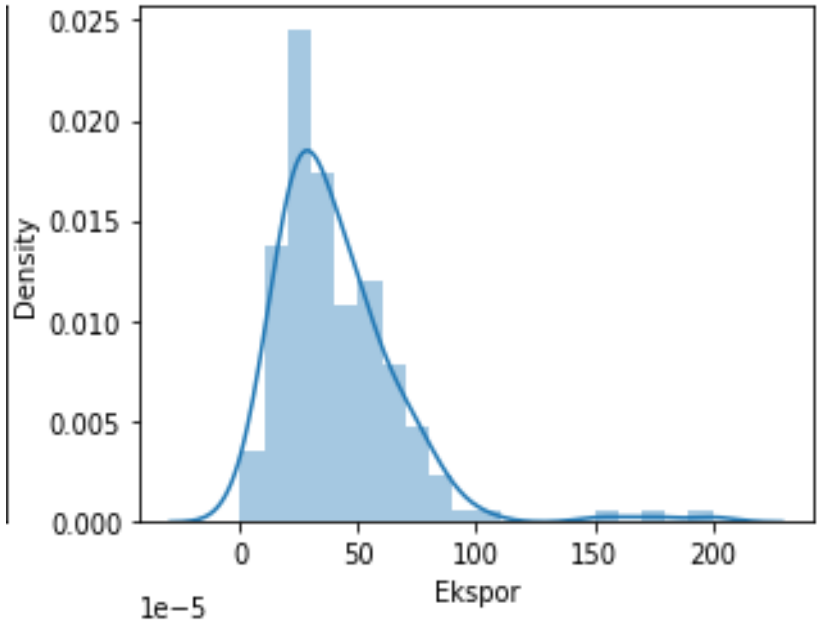
- Data terbanyak terdapat pada rentang 0 – 25
- Rentang data antara 0 - 225
- Terdapat sedikit data outlier pada data ini dan bernilai lebih tinggi dari nilai batas quartil atas

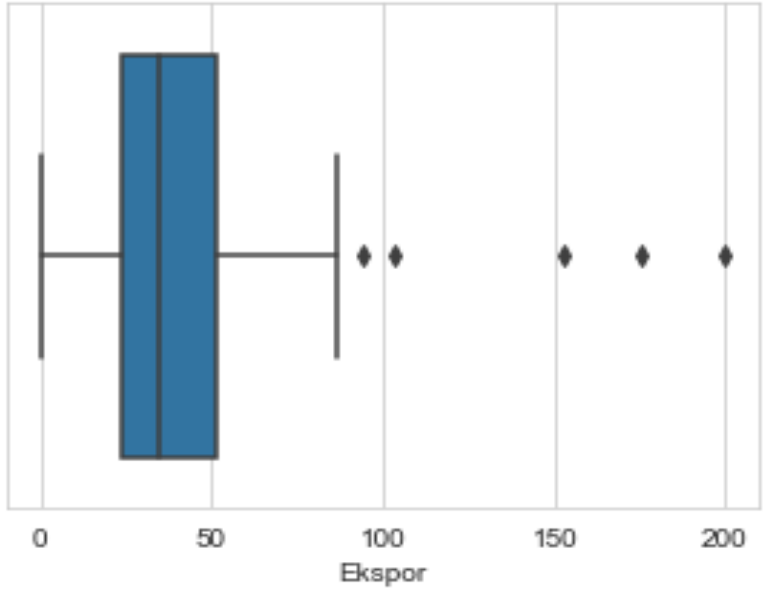




EKSPOR

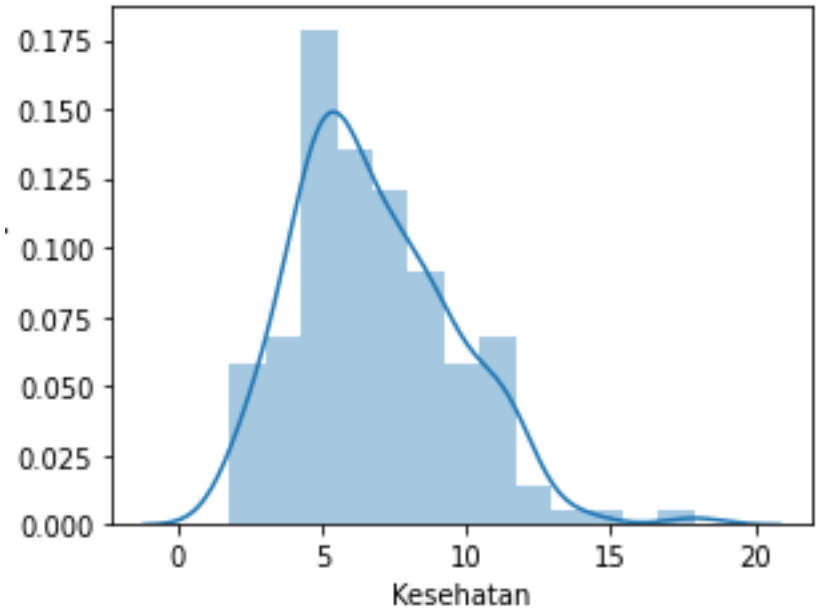
- Data terbanyak terdapat pada rentang 25-50
- Rentang data bernilai 0 - 200
- Terdapat beberapa outlier pada data ini dan bernilai lebih tinggi dari nilai batas quartil atas

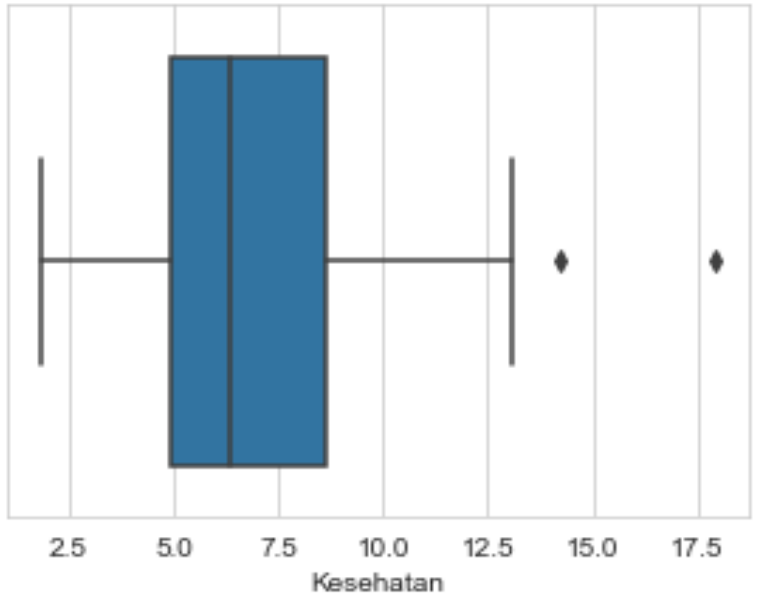




KESEHATAN

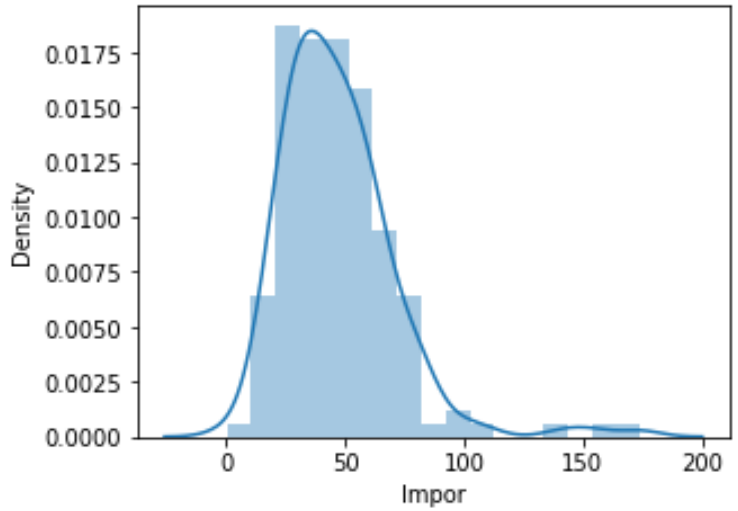
- Data terbanyak terdapat pada rentang *5 - 7,5*
- Rentang data bernilai 2 - 18
- Terdapat 2 outlier pada data ini dan bernilai lebih tinggi dari nilai batas quartil atas

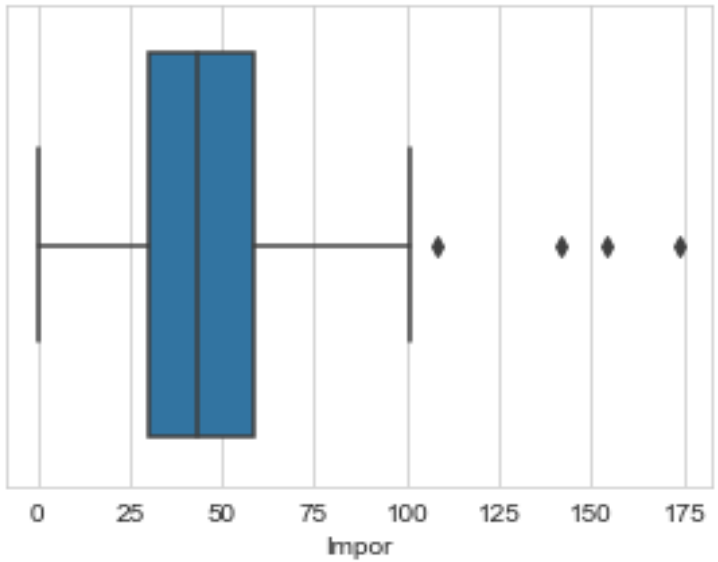




IMPOR

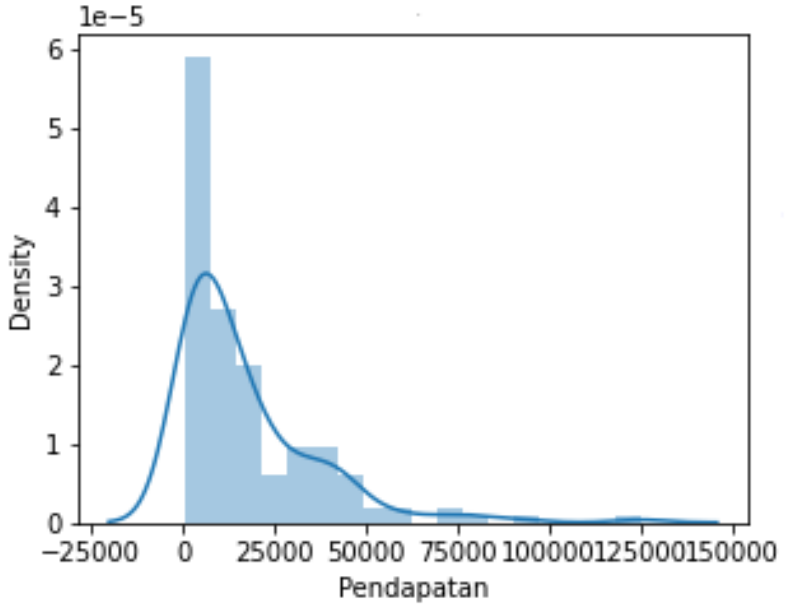
- Data terbanyak terdapat pada rentang 25 - 66
- Rentang data bernilai 0 - 175
- Terdapat beberapa outlier pada data ini dan bernilai lebih tinggi dari nilai batas atas

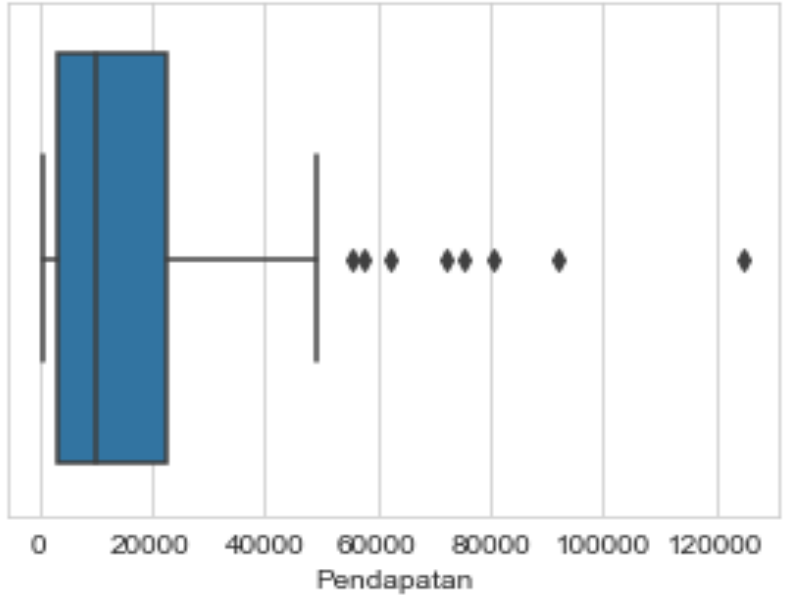




PENDAPATAN

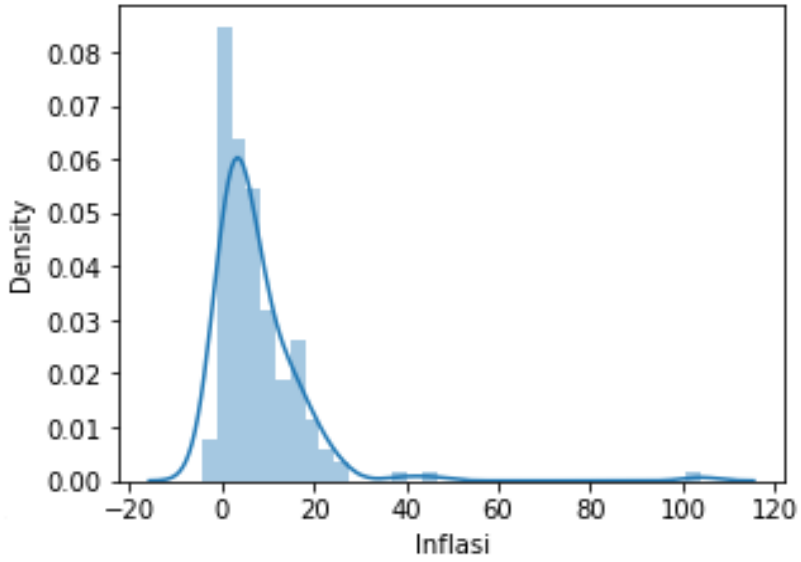
- Data terbanyak terdapat pada rentang 0 - 6500
- Rentang data bernilai 0 - 125000
- Terdapat beberapa outlier pada data ini dan bernilai lebih tinggi dari nilai batas kuartil atas

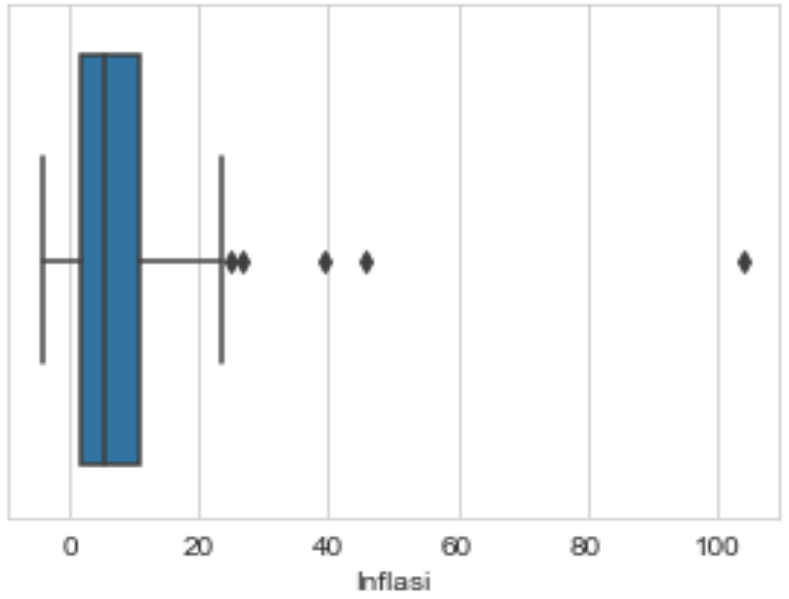




INFLASI

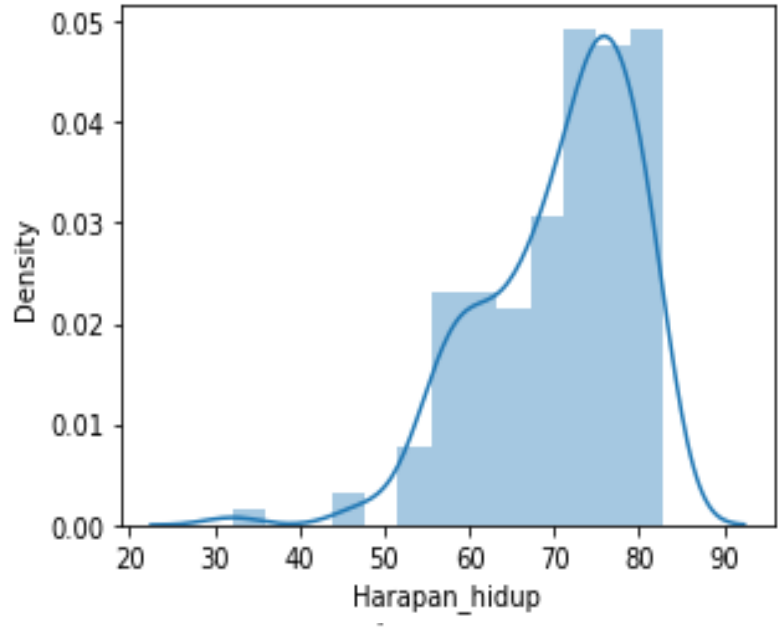
- Data terbanyak terdapat pada rentang 0 - 10
- Rentang data bernilai 0 - 110
- Terdapat beberapa outlier pada data ini dan bernilai lebih tinggi dari nilai batas quartil atas

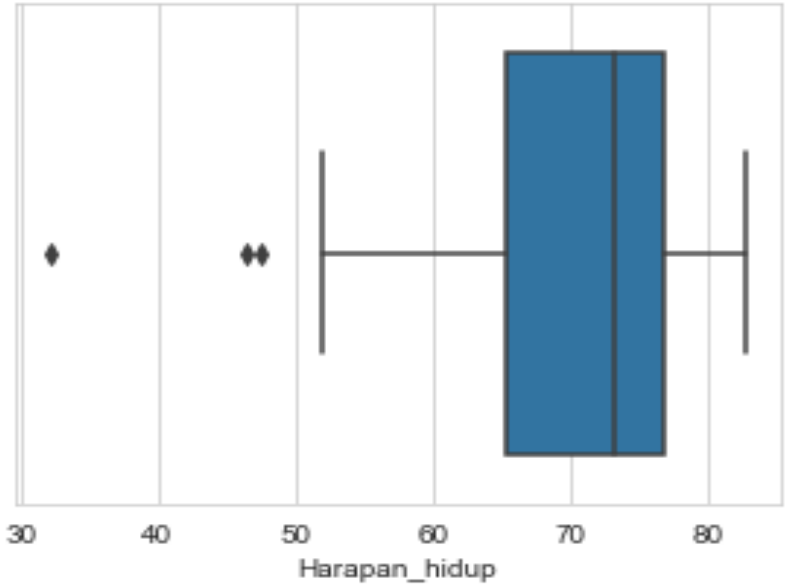




HARAPAN HIDUP

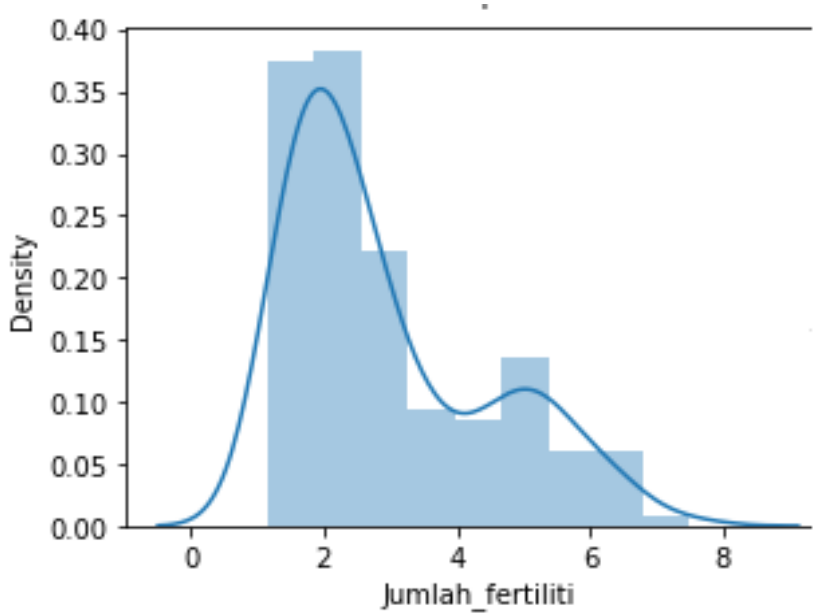
- Data terbanyak terdapat pada rentang 70 - 80
- Rentang data bernilai 30 - 85
- Terdapat beberapa outlier pada data ini dan bernilai lebih rendah dari nilai batas quartil bawah

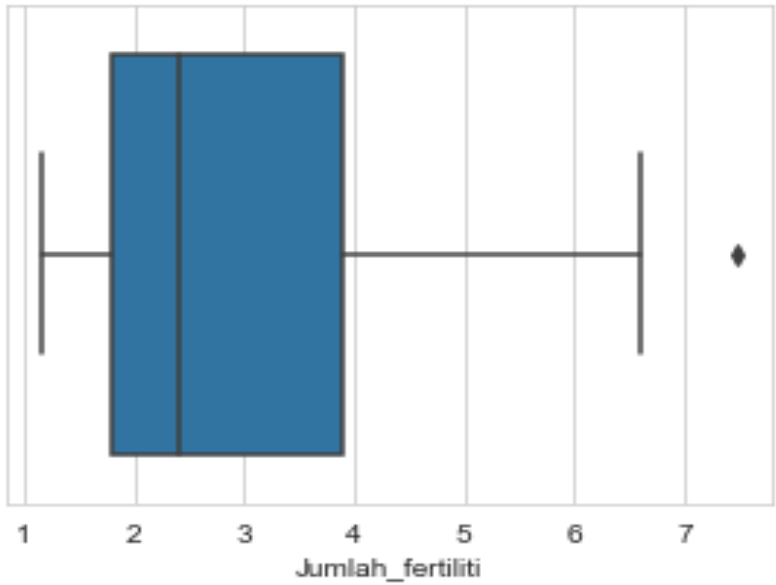




JUMLAH FERTILITI

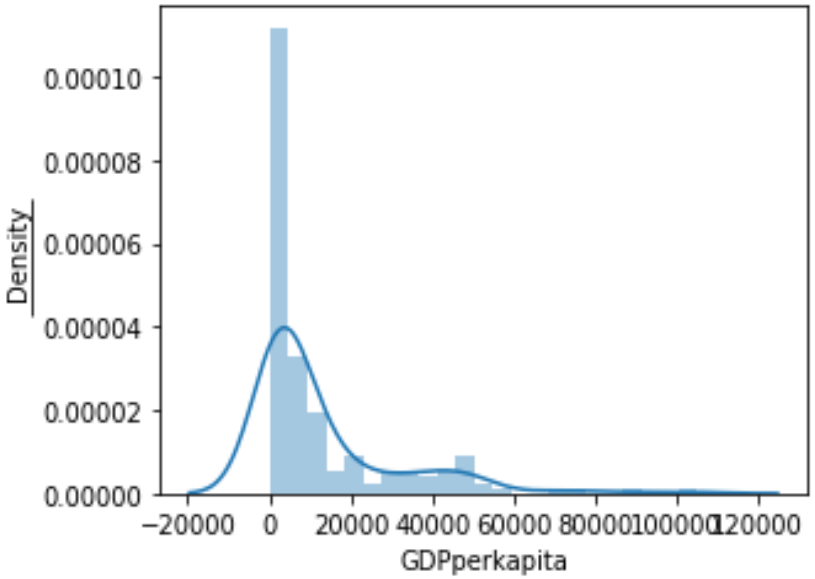
- Data terbanyak terdapat pada rentang 1 - 2
- Rentang data bernilai 1 - 18
- Terdapat sebuah outlier pada data ini dan bernilai lebih tinggi dari nilai batas quartil atas

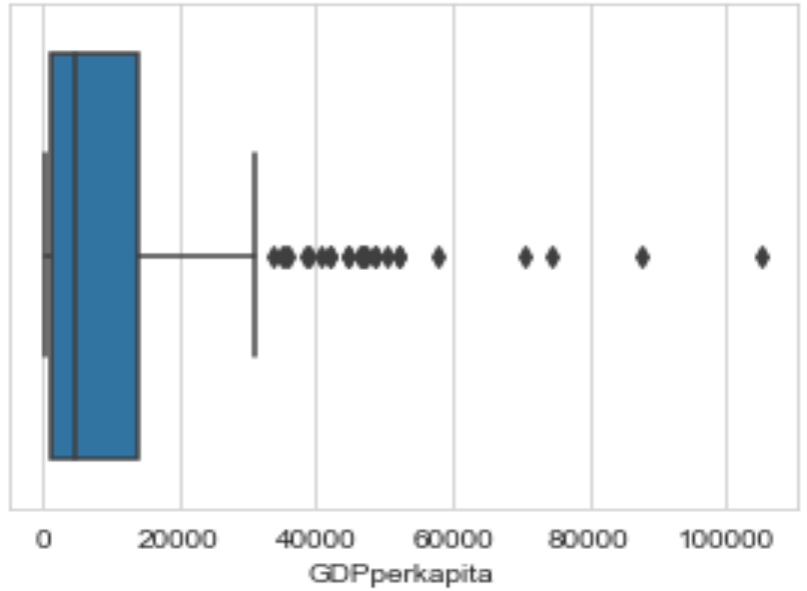


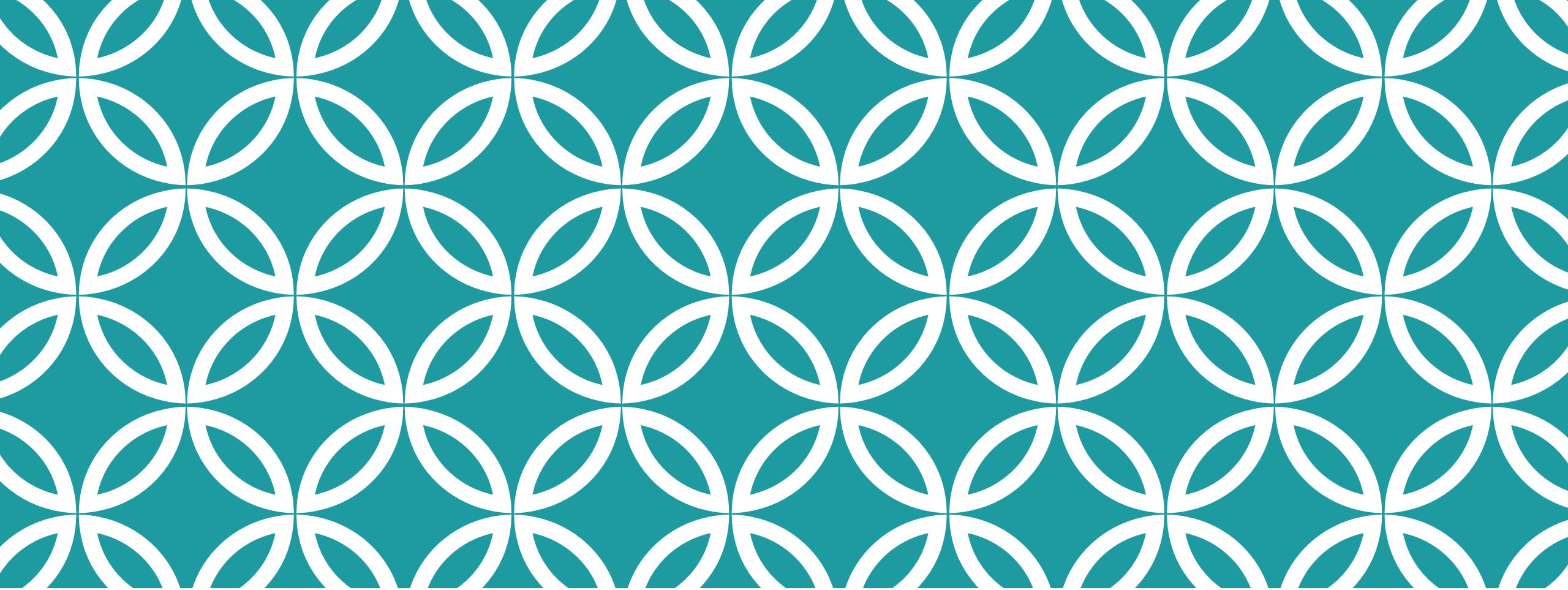


GDP PERKAPITA

- Data terbanyak terdapat pada rentang 0 - 4000
- Rentang data bernilai 0 - 120000
- Terdapat banyak data outlier pada data ini dan bernilai lebih tinggi dari nilai batas quartil atas





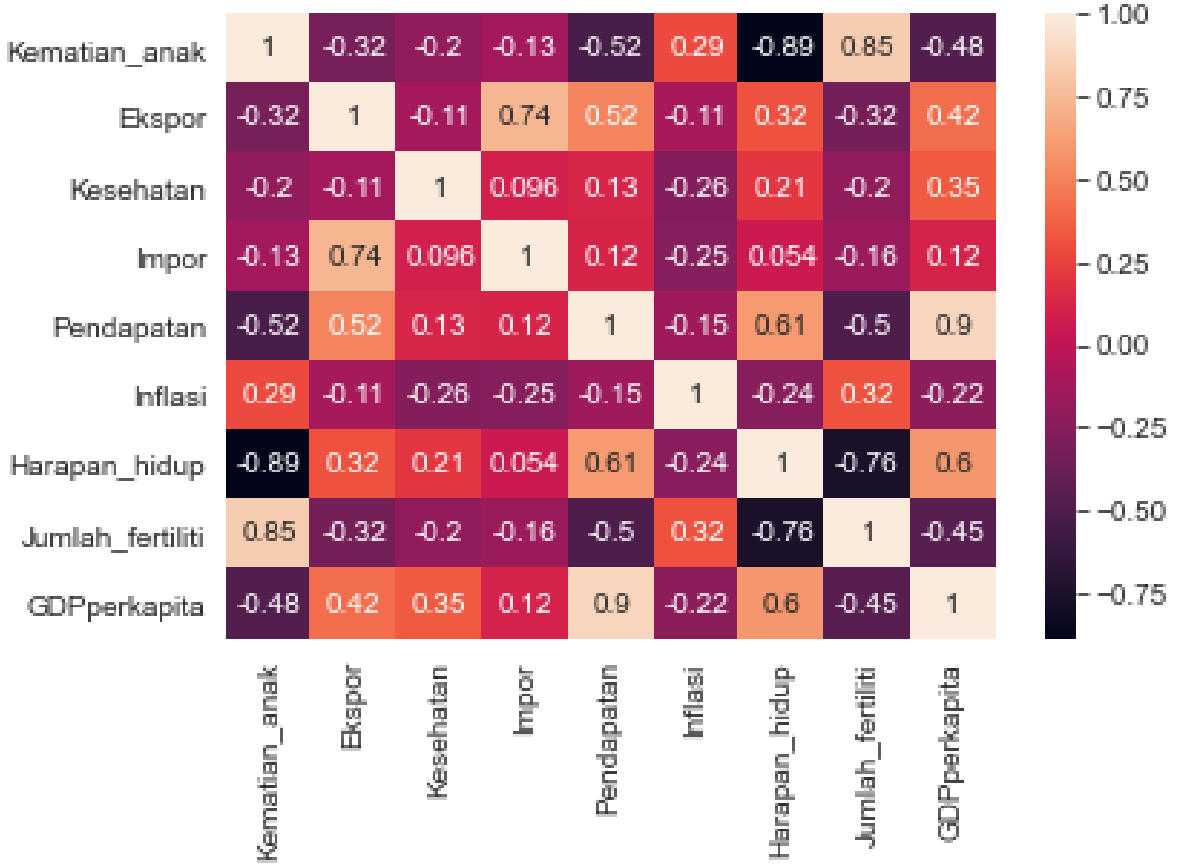


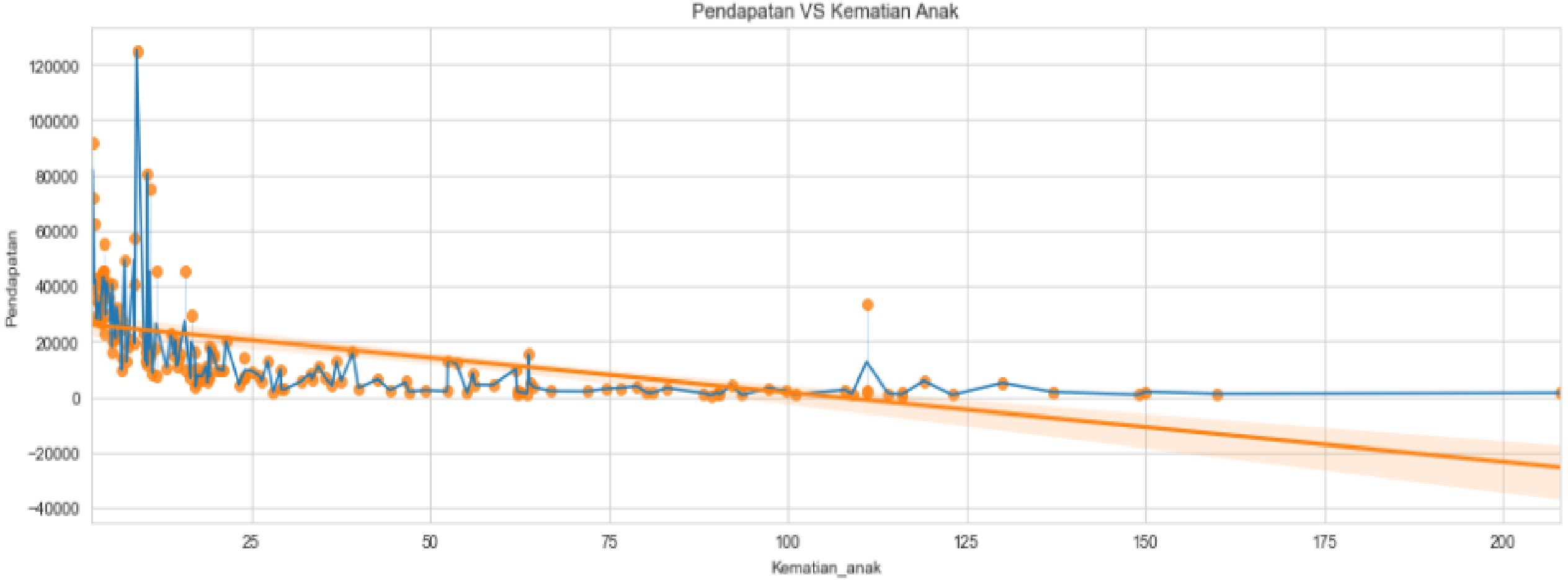
EDA MULTIVARIATE & BIVARIATE ANALYSIS

MULTIVARIATE

Analisis multivariate yang digunakan menggunakan metode heat map dengan menganalisa korelasi antara variabel-variabel tersebut. Hasil dari analisa ini digunakan untuk menjadi dasar dari analisa bivariate. Dari grafik disamping dapat diambil kesimpulan bahwa :

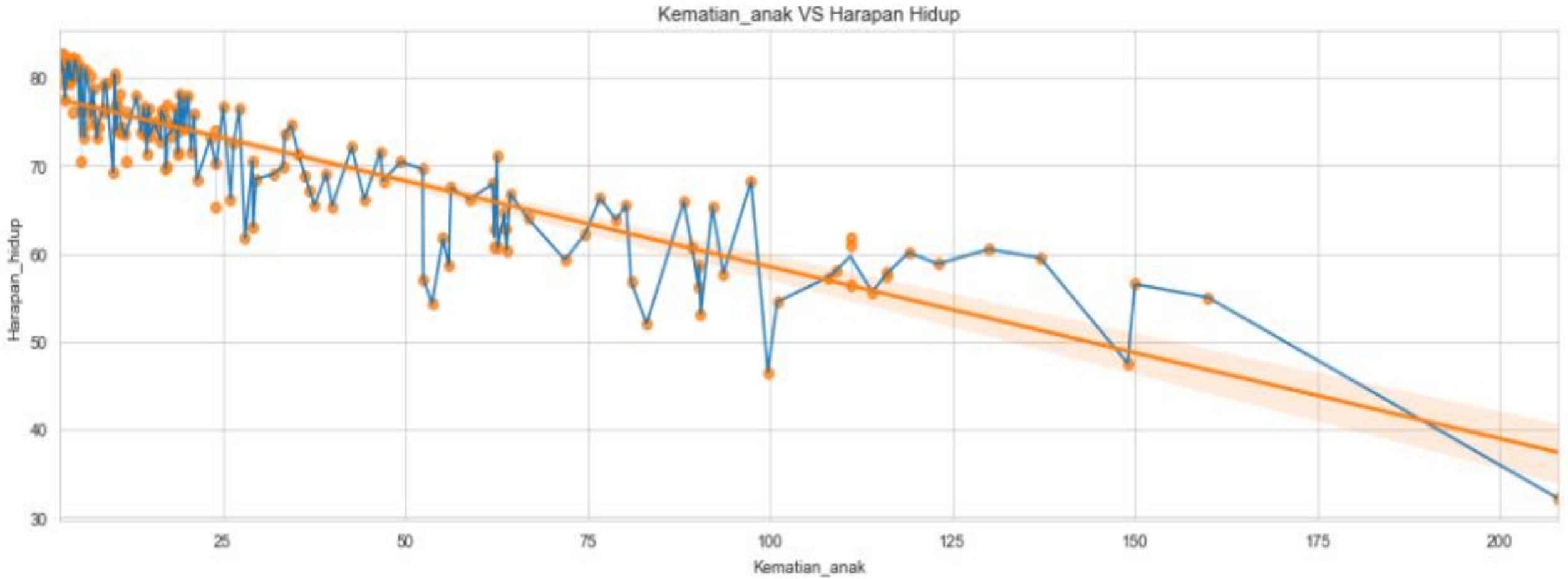
1. Kematian anak berbanding lurus dengan jumlah fertiliti dengan nilai 0,85
2. Kematian anak sangat berbanding terbalik dengan harapan hidup dengan nilai -0,89
3. Kematian anak cukup berbanding terbalik dengan pendapatan dengan nilai -0,52
4. Ekspor sangat berbanding lurus dengan Impor dengan nilai 0,74
5. Ekspor cukup berbanding lurus dengan pendapatan dengan nilai 0,52
6. Kesehatan tidak memiliki korelasi dengan faktor lainnya
7. Pendapatan sangat berbanding lurus dengan GDPperkapita dengan nilai 0,9
8. Pendapatan cukup berbanding lurus dengan harapan hidup dengan nilai 0,61
9. Pendapatan cukup berbanding terbalik dengan jumlah fertiliti dengan nilai -0,5
10. Inflasi sangat tidak memilki korelasi dengan faktor lainnya
11. Harapan hidup cekup berbanding lurus dengan GDP perkapita dengan nilai 0,6
12. Harapan hidup sangat berbanding terbalik dengan jumlah fertiliti dengan nilai -0,76





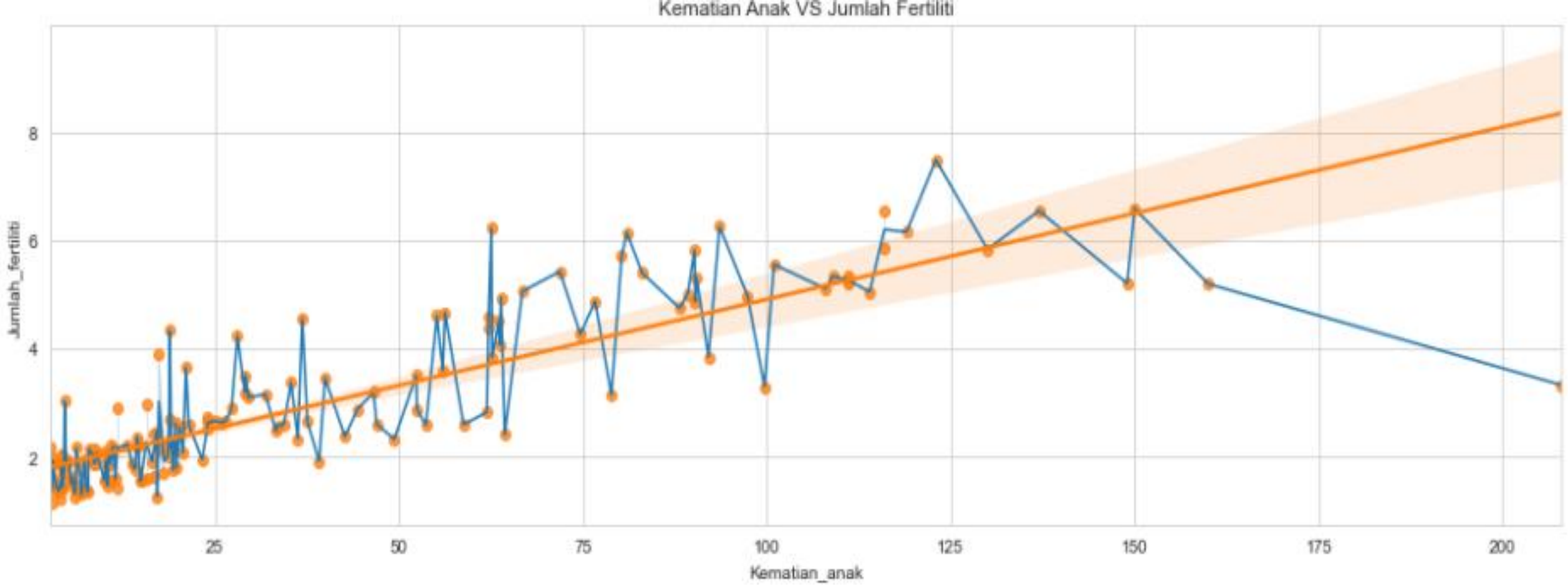
PENDAPATAN VS KEMATIAN ANAK

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa pendapatan berbanding terbalik dengan kematian anak. Sehingga, negara yang akan menerima bantuan hendaknya merupakan negara yang memiliki angka kematian yang tinggi dan memiliki pendapatan yang rendah



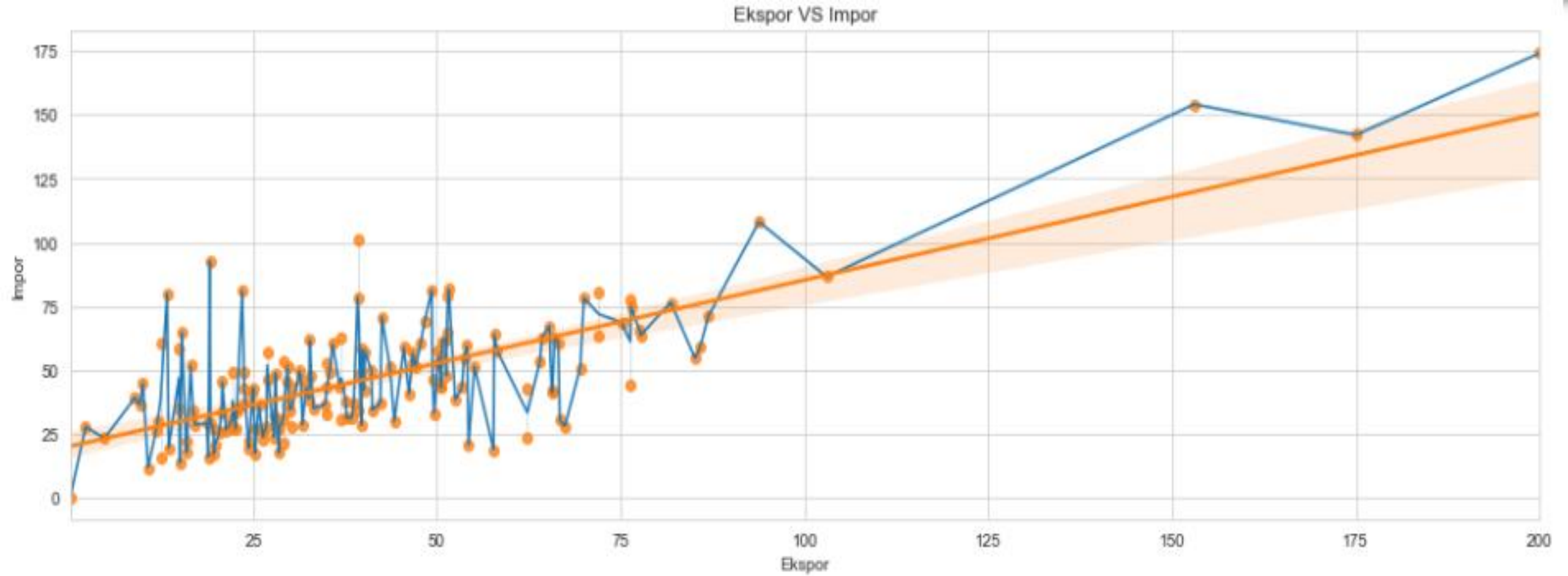
KEMATIAN ANAK VS HARAPAN HIDUP

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa harapan hidup berbanding terbalik dengan kematian anak. Sehingga, negara yang akan menerima bantuan merupakan negara dengan angka kematian yang tinggi dan harapan hidup yang rendah.



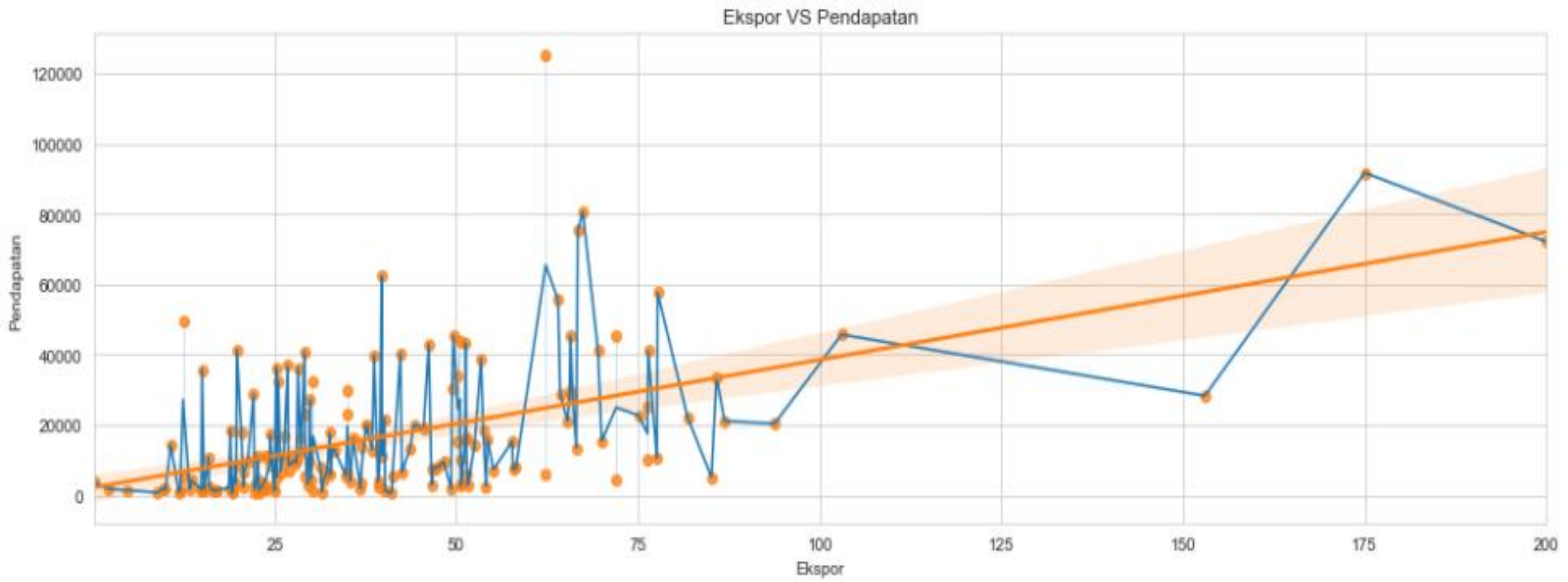
KEMATIAN ANAK VS JUMLAH FERTILITI

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa jumlah fertiliti berbanding lurus dengan kematian anak. Sehingga negara yang akan menerima bantuan merupakan negara dengan nilai kematian anak yang tinggi dan jumlah fertiliti yang tinggi



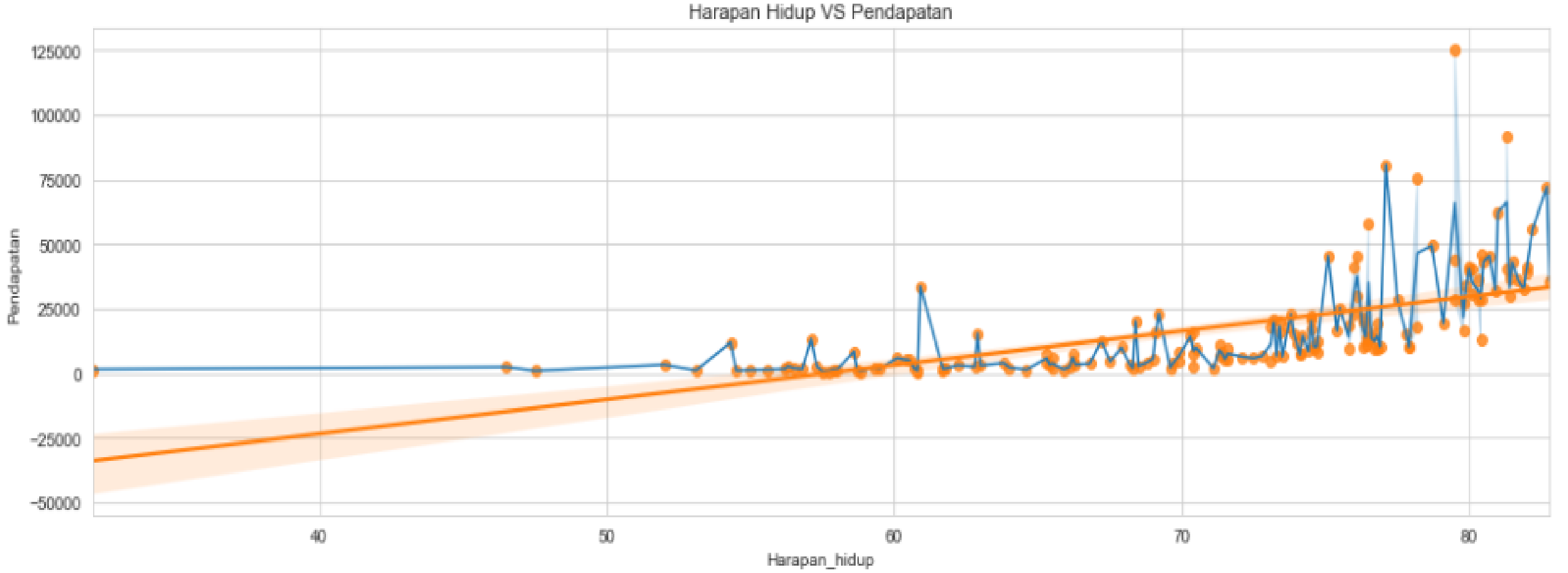
EKSPOR VS IMPOR

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa Ekspor berbanding lurus dengan impor. Sehingga, negara yang akan menerima bantuan merupakan negara dengan ekspor dan impor yang rendah



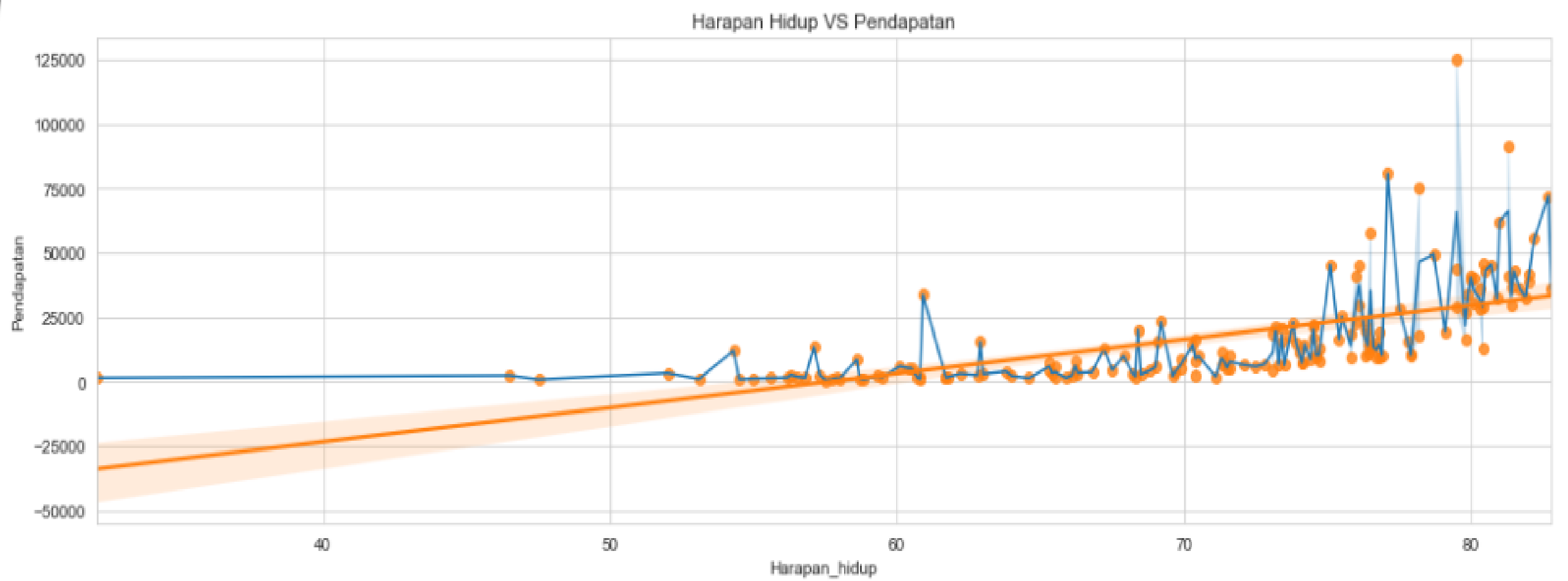
EKSPOR VS PENDAPATAN

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa ekspor berbanding lurus dengan pendapatan. Sehingga, negara yang akan mendapat bantuan merupakan negara dengan nilai ekspor dan pendapatan yang rendah



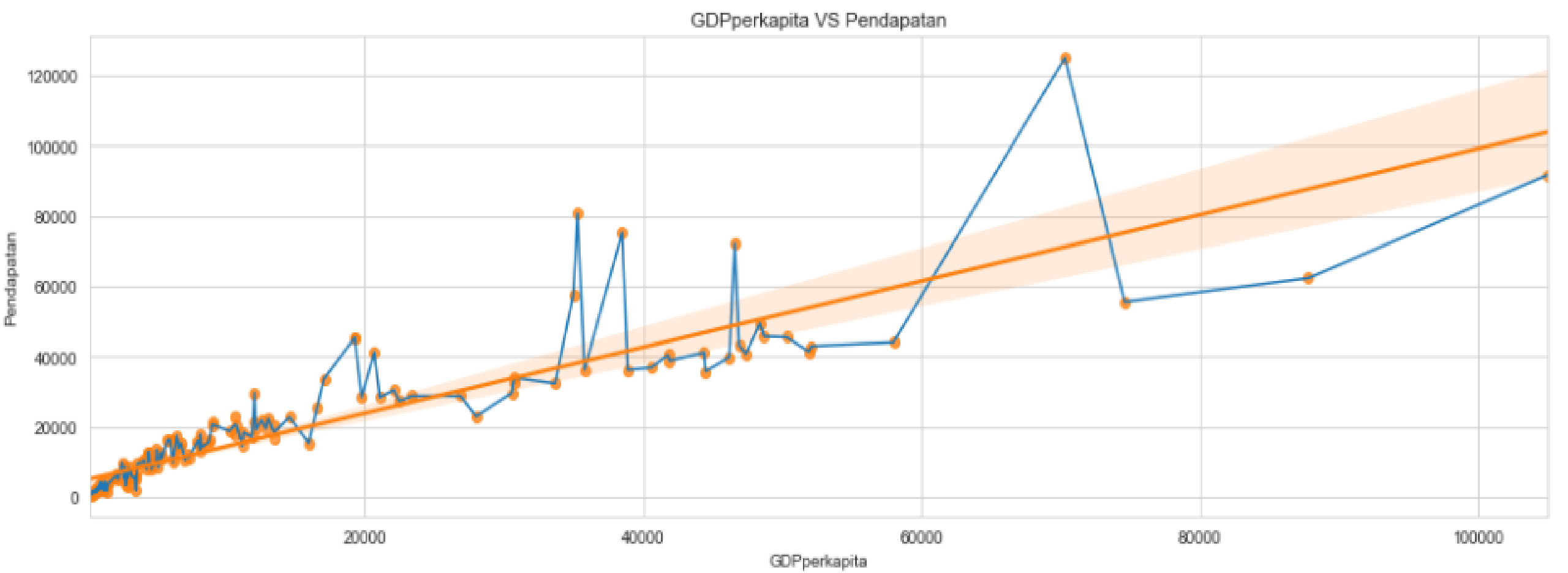
HARAPAN HIDUP VS PENDAPATAN

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa harapan hidup berbanding lurus dengan pendapatan. Sehingga, negara yang hendaknya mendapatkan bantuan merupakan negara dengan harapan hidup dan pendapatan yang rendah



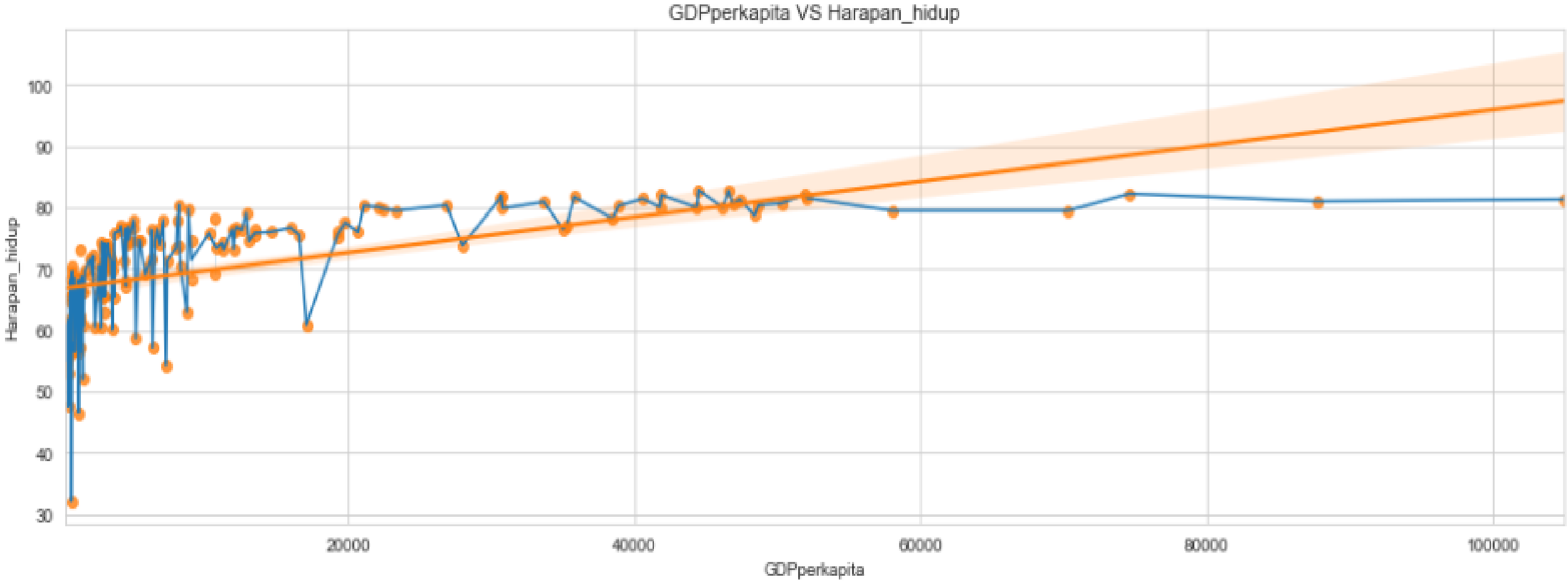
HARAPAN HIDUP DAN PENDAPATAN

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa harapan hidup berbanding lurus dengan pendapatan. Sehingga, negara yang hendaknya mendapatkan bantuan ialah negara dengan nilai harapan hidup dan pendapatan yang rendah



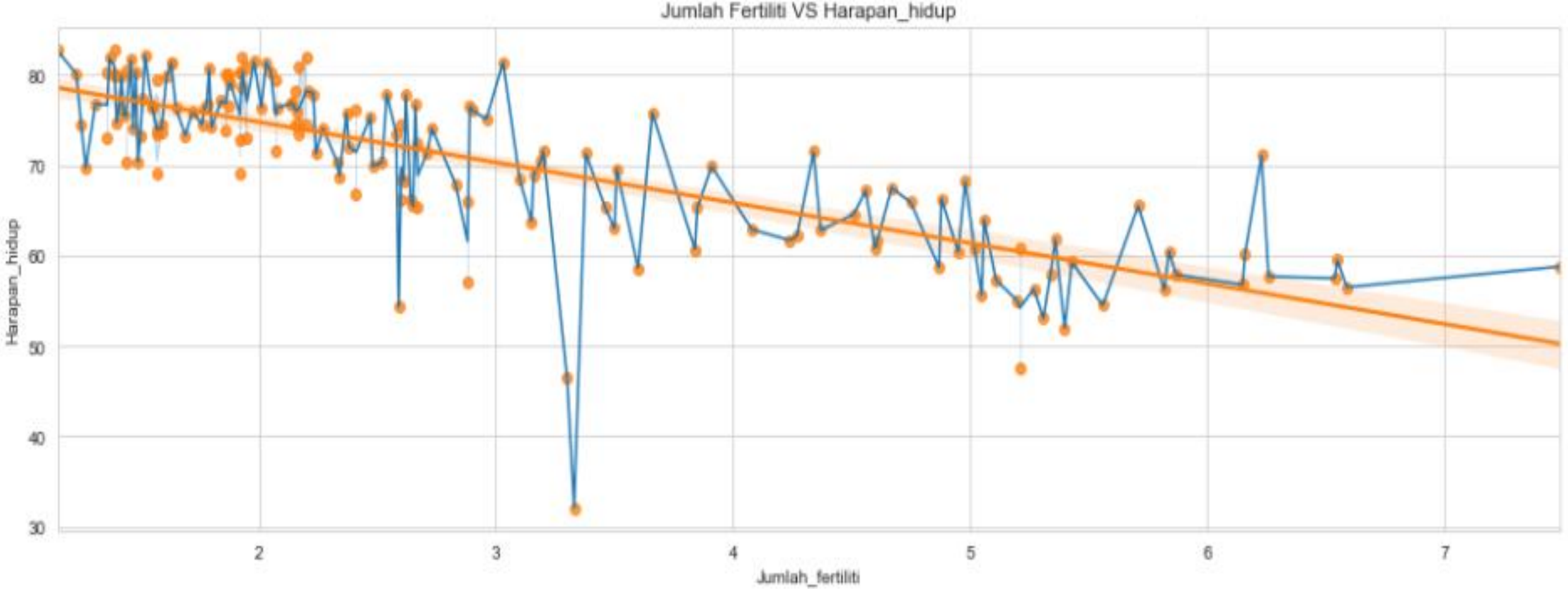
GDP PERKAPITA VS PENDAPATAN

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa GDP perkapita berbanding lurus dengan pendapatan. Sehingga, negara yang hendaknya memperoleh bantuan ialah negara dengan GDP perkapita dan pendapatan yang rendah



GDP PERKAPITA VS HARAPAN HIDUP

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa harapan hidup berbanding lurus dengan GDPperkapita. Sehingga, negara yang hendaknya memperoleh bantuan ialah negara dengan GDP perkapita yang rendah dan harapan hidup yang rendah

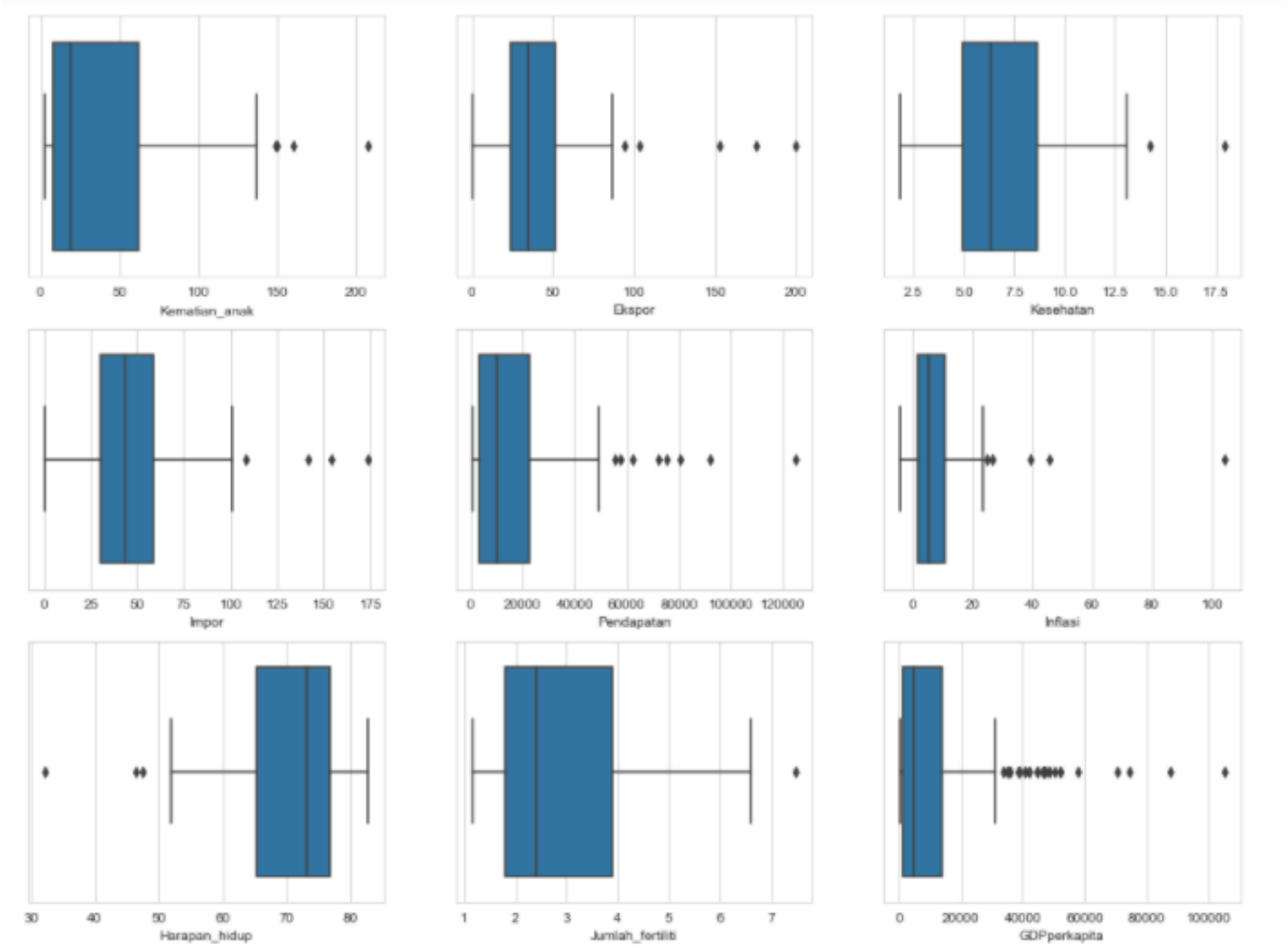


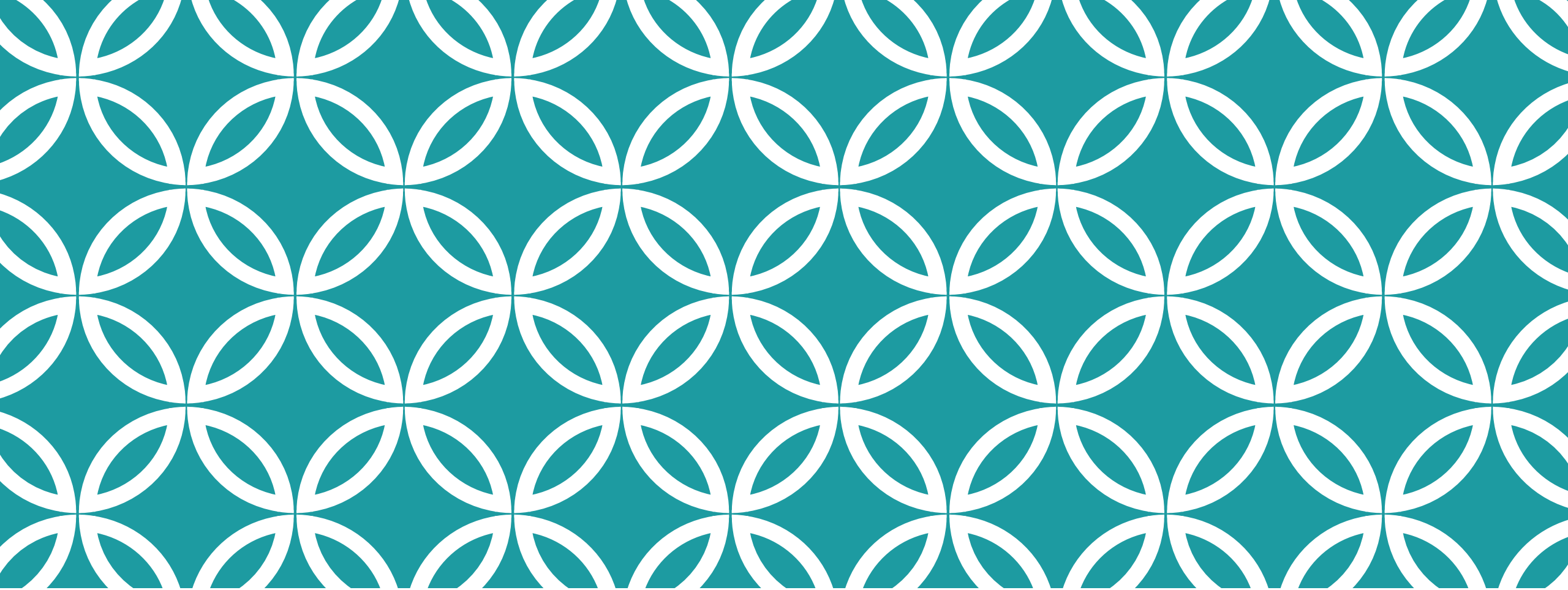
JUMLAH FERTILITI DAN HARAPAN HIDUP

Jika ditarik garis lurus secara linier dapat dinyatakan bahwa jumlah fertiliti berbanding terbalik dengan harapan hidup. Sehingga, negara yang hendaknya memperoleh bantuan ialah negara yang memiliki jumlah fertiliti yang tinggi dan memiliki harapan hidup yang rendah

OUTLIER THREATMENT

Pada pengaturan outlier kali ini penguji tidak menghilangkan outlier hal ini disebabkan agar tidak ada negara yang hilang dalam perhitungan walaupun memungkinkan mengurangi keakuratan data pada clustering





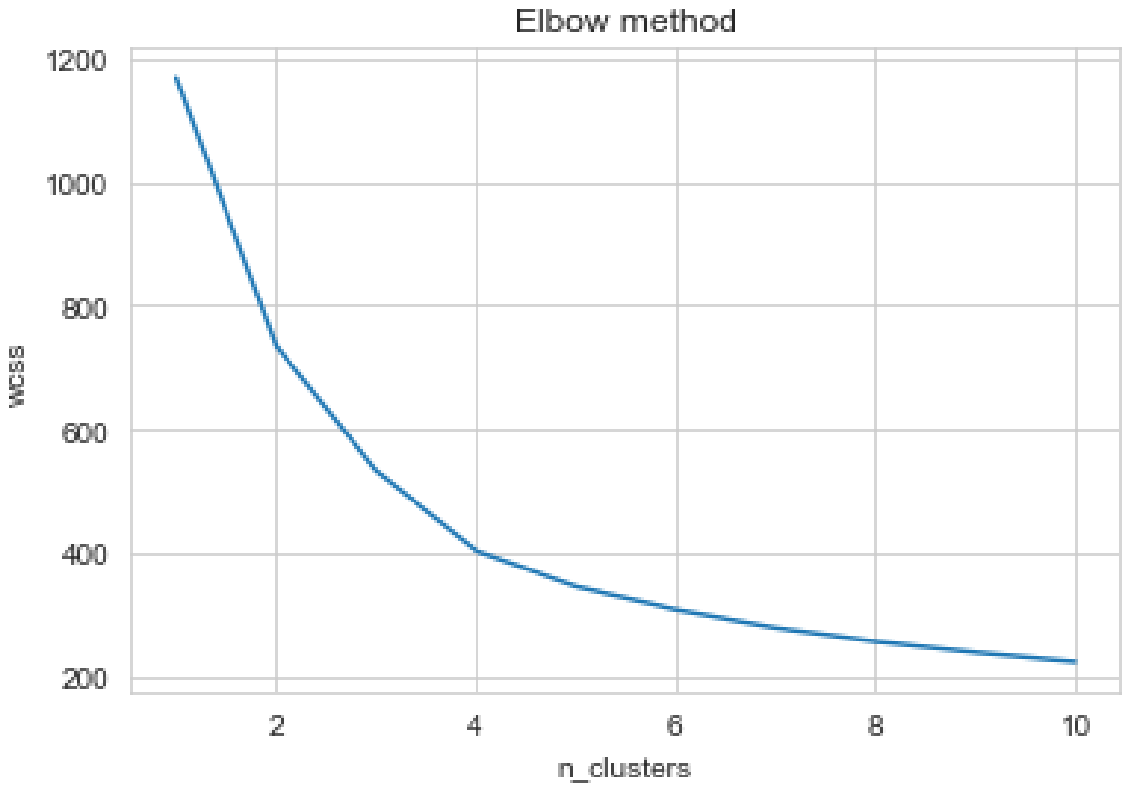
CLUSTERING

TAHAP CLUSTERING

1. Standarisasi data
   - Standarisai data yang digunakan ialah metode StandardScaler
2. Menentukan jumlah kelas clustering
   - Metode yang digunakan dalam mementukan jumlah kelas clustering ialah metode elbow
3. Membuat clustering
   - Membuat clustering berdasarkan faktor-faktor pada analisis bivariet
4. Membuat kesimpulan dari clustering yang telah dibuat

MENENTUKAN JUMLAH CLUSTERING

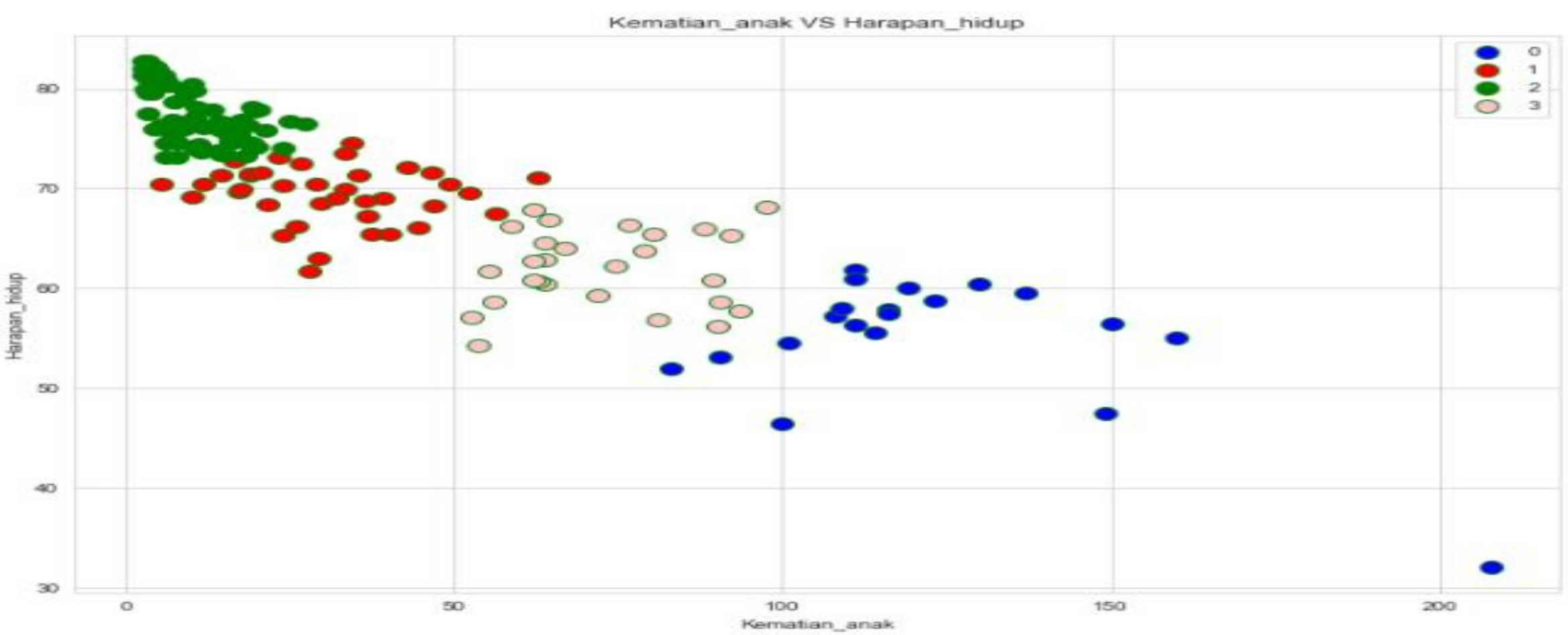
Jika dilihat dari grafik di samping nilai kelas yang paling optimum ialah 4



0.5092340679384362

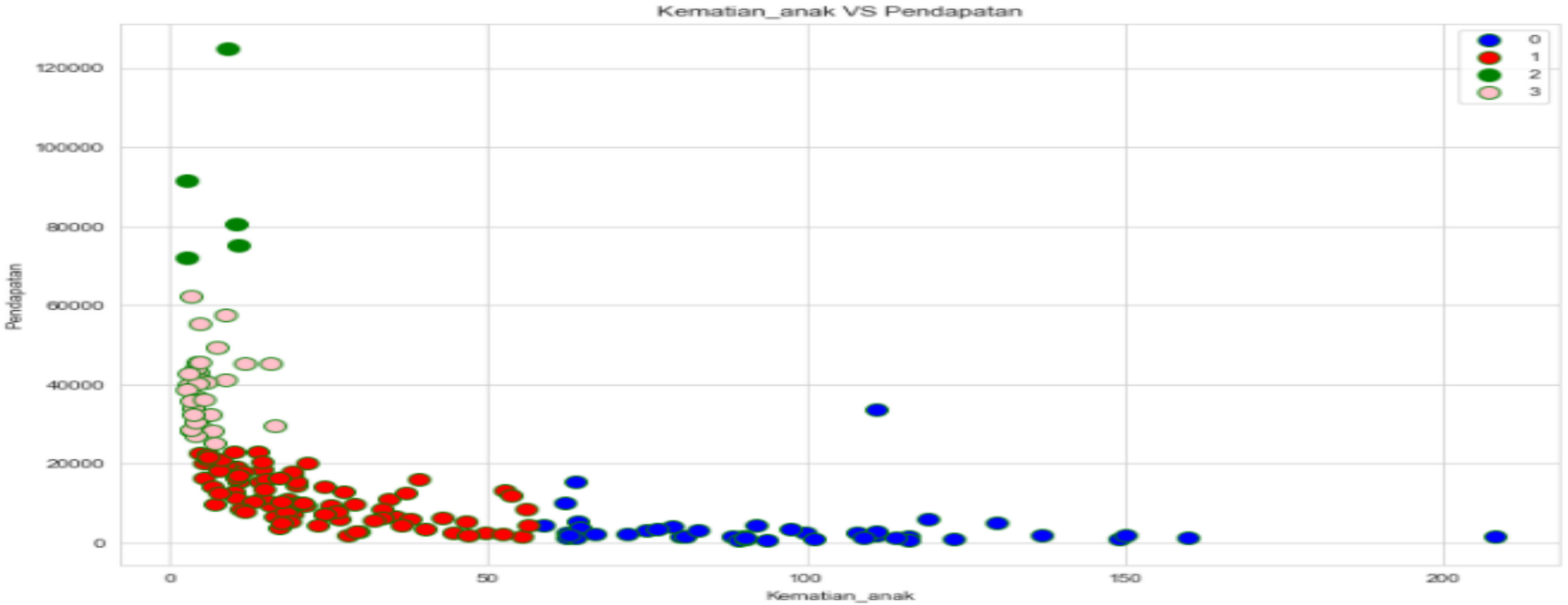
CLUSTERING KEMATIAN ANAK VS JUMLAH FERTILITI

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai kematian anak yang tinggi dan jumlah fertiliti yang tinggi



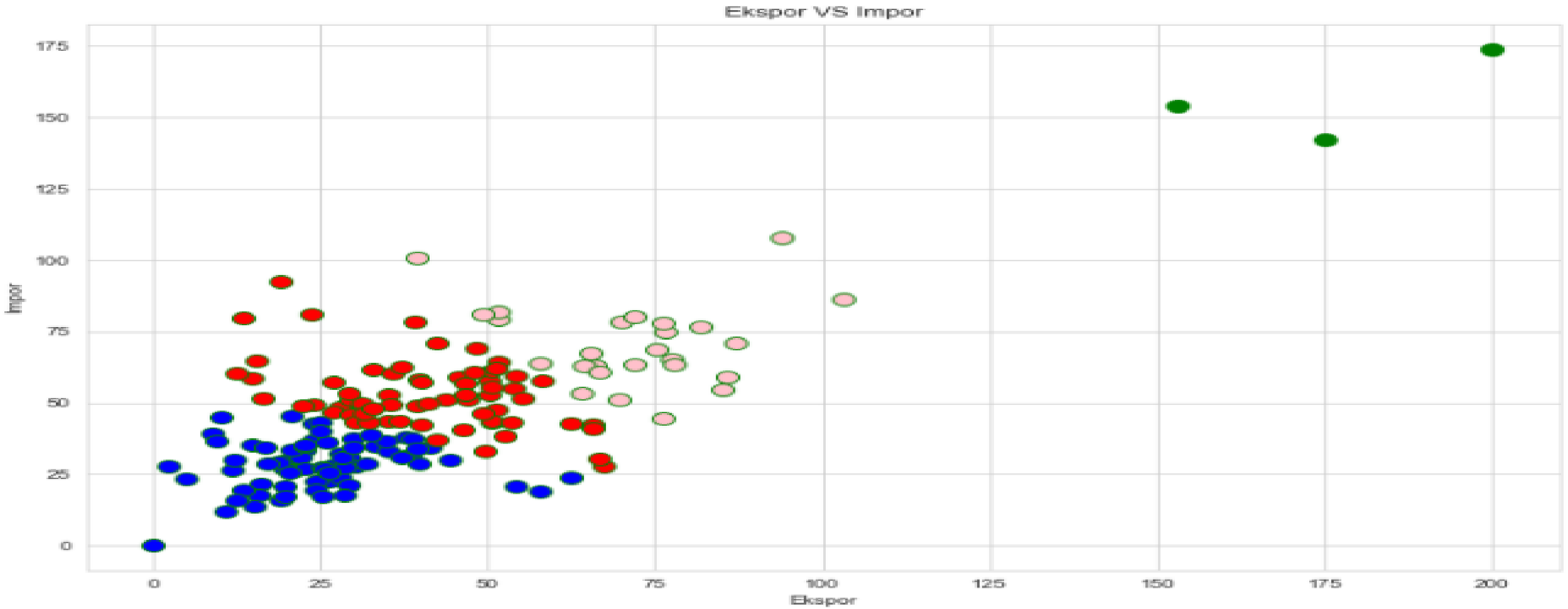
CLUSTERING KEMATIAN ANAK VS HARAPAN HIDUP

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai kematian anak yang tinggi dan nilai harapan hidup yang rendah



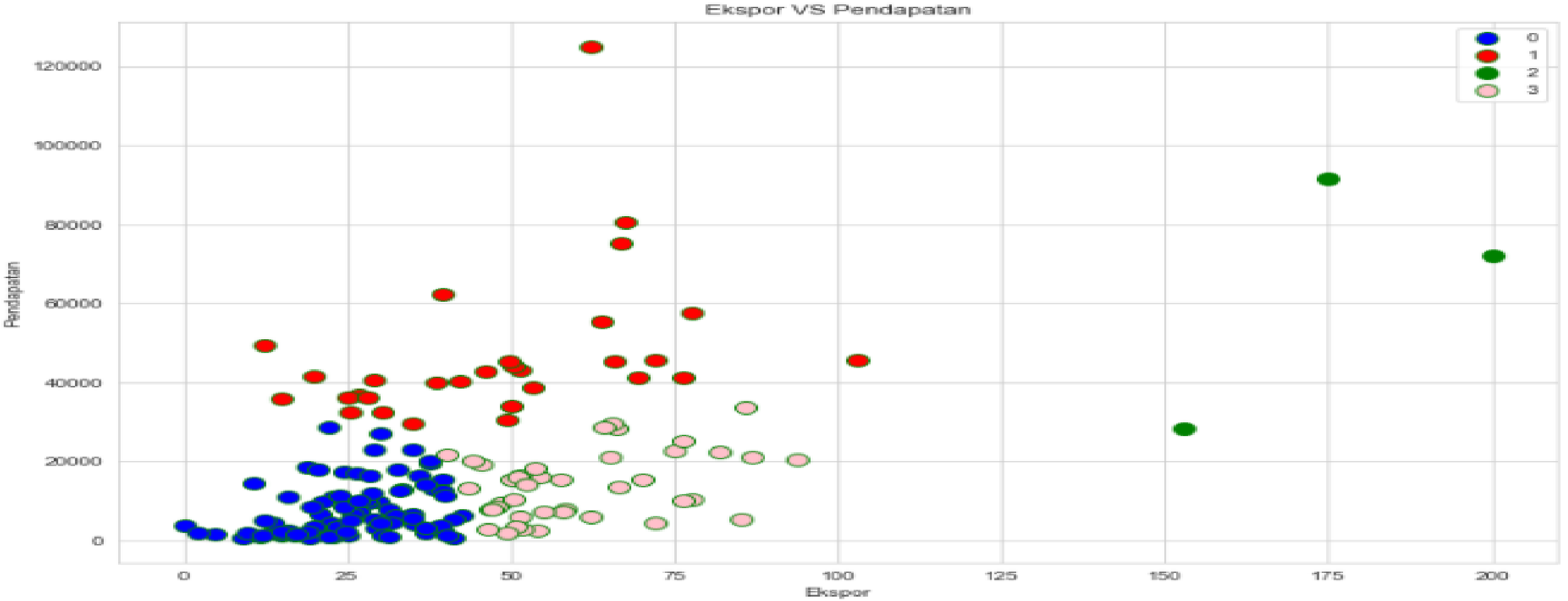
CLUSTERING KEMATIAN ANAK VS PENDAPATAN

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai kematian anak yang tinggi dan nilai pendapatan yang rendah



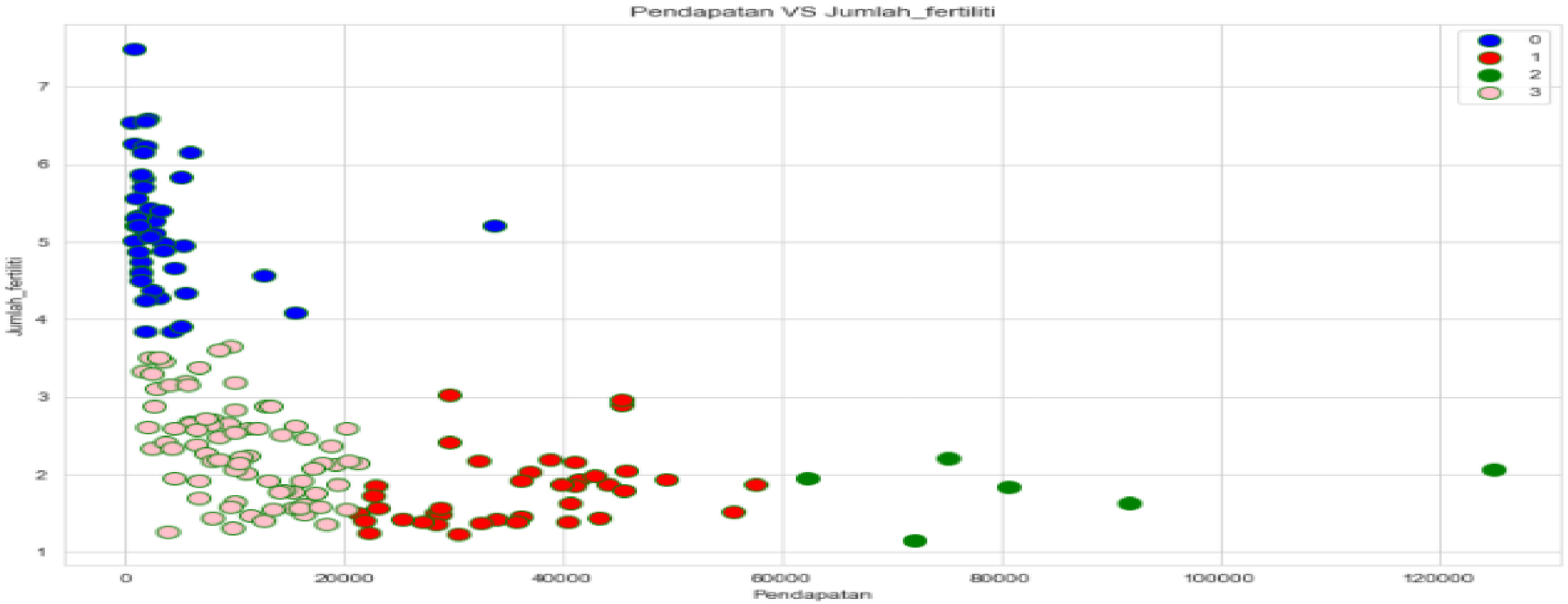
CLUSTERING EKSPOR VS IMPOR

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai impor dan ekspor yang rendah.



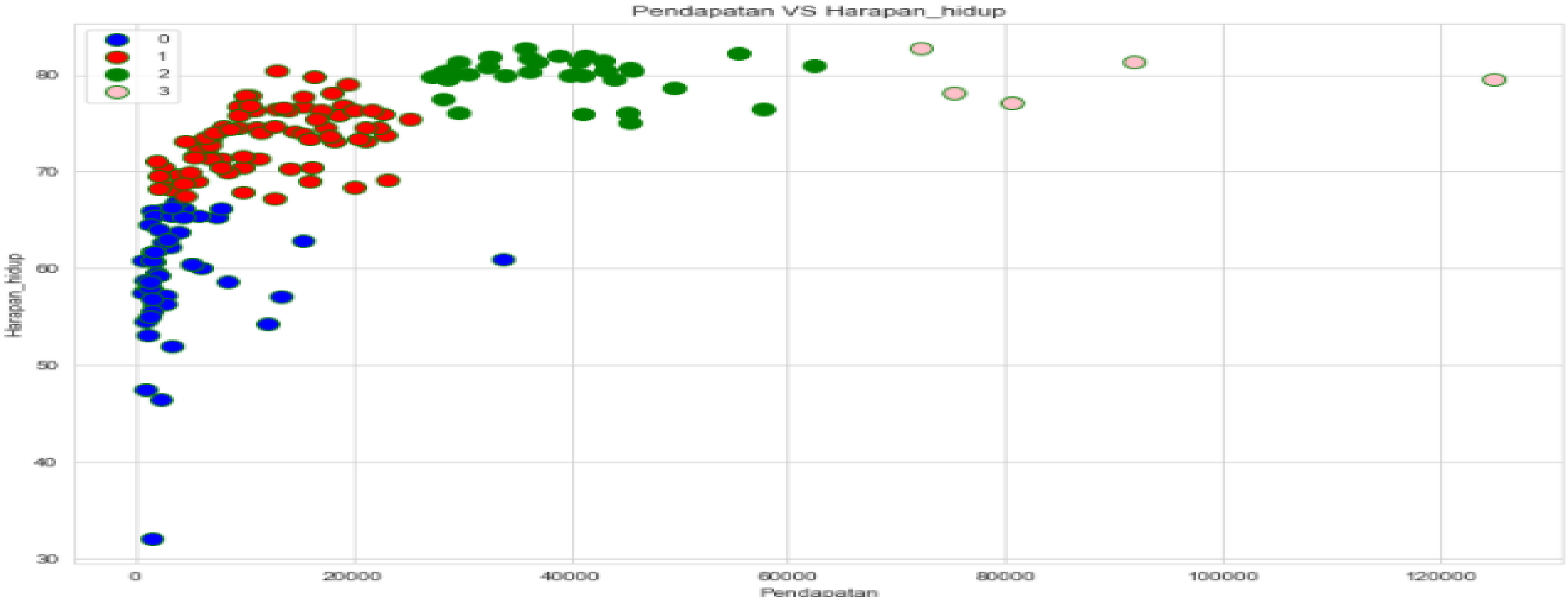
CLUSTERING EKSPOR VS PENDAPATAN

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai pendapatan dan ekspor yang rendah.



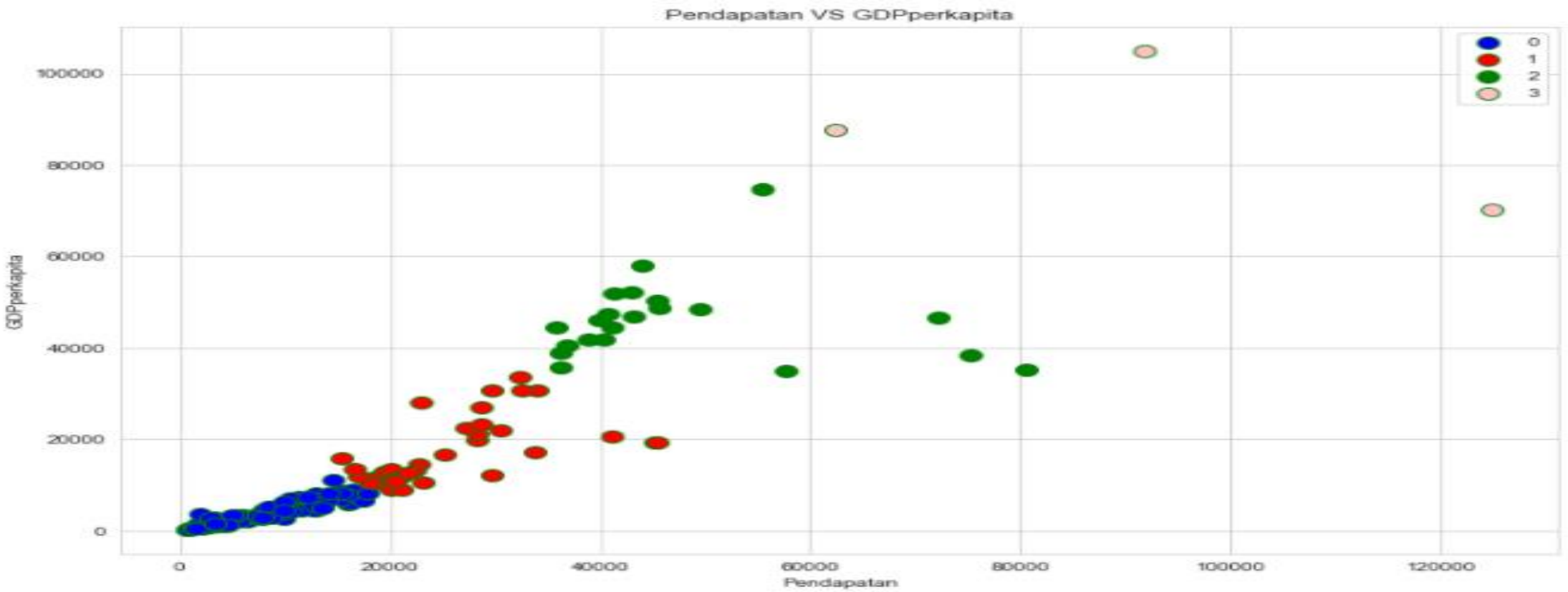
CLUSTERING PENDAPATAN VS JUMLAH FERTILITI

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai pendapatan yang rendah dan memiliki jumlah fertiliti yang tinggi



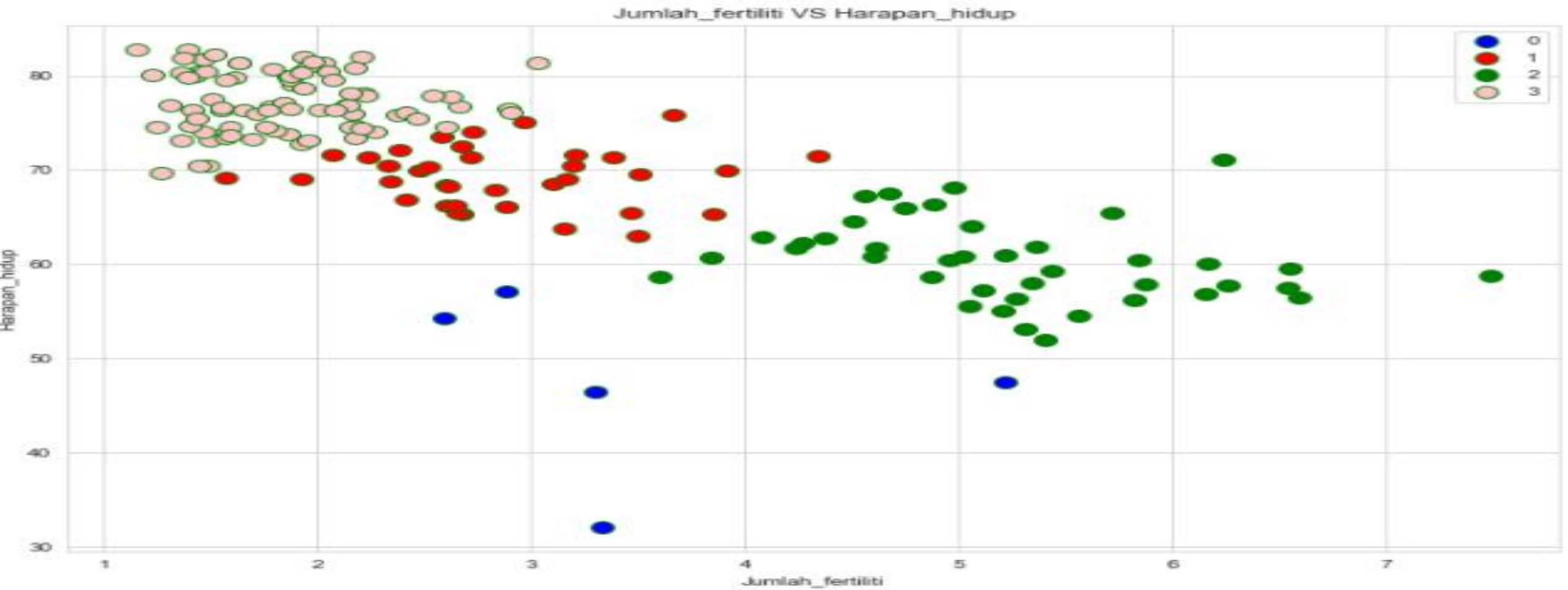
CLUSTERING PENDAPATAN VS HARAPAN HIDUP

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai pendapatan dan nilai harapan hidup yang rendah



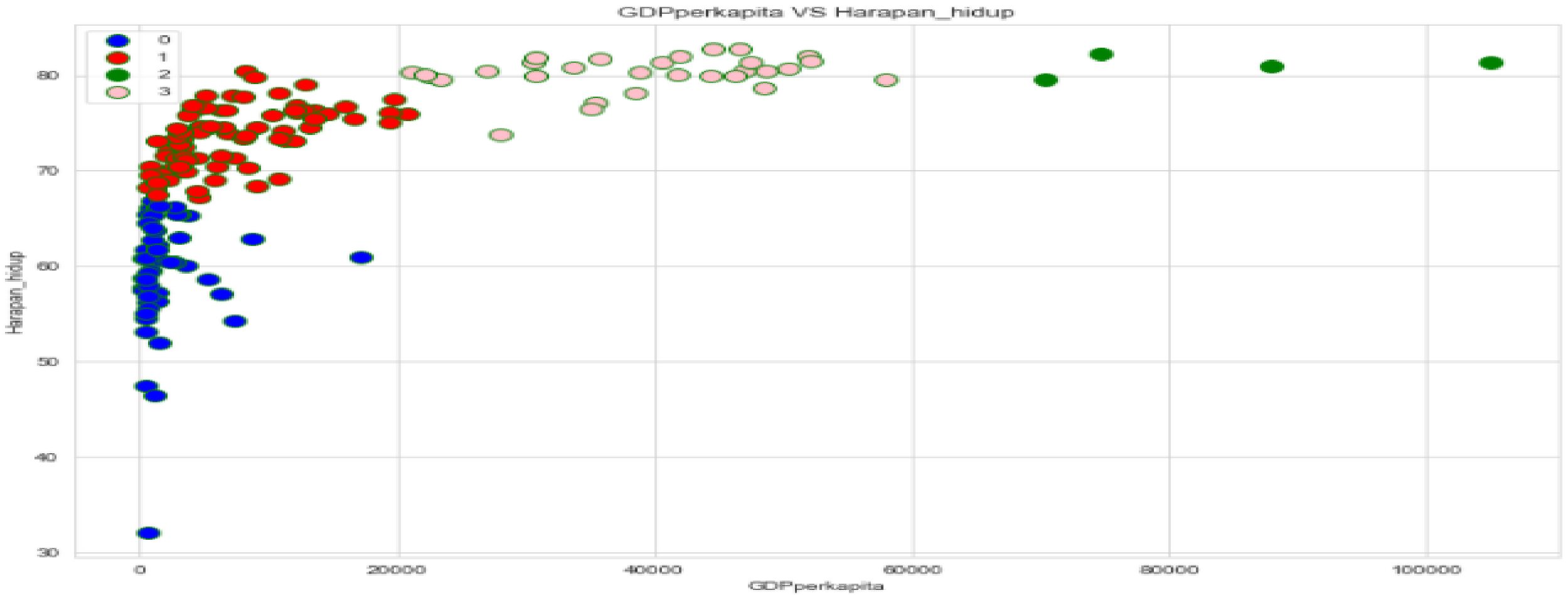
CLUSTERING PENDAPATAN VS GDP PERKAPITA

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai pendapatan dan nilai GDP perkapita yang rendah



CLUSTERING JUMLAH FERTILITI VS HARAPAN HIDUP

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai harapan hidup yang rendah dan memiliki jumlah fertiliti yang tinggi



CLUSTERING GDP PERKAPITA VS HARAPAN HIDUP

Negara yang paling membutuhkan bantuan ialah negara dengan warna biru karena negara tersebut memiliki nilai GDP perkapita dan nilai harapan hidup yang rendah

SUMMARY CLUSTERING

Setelah semua faktor telah diclustering kemudian hasil nilai 0 dijumlahkan serta disortir berdasarkan jumlah 0 yang terbanyak sebab ketika clustering sudah diatur agar yang bernilai 0 merupakan negara paling membutuhkan bantuan. Kemudian data tersebut disortir bedasarkan nilai 0 yang terbanyak. Data sortir tersebut diambil indexnya untuk memperoleh negara yang membutuhkan bantuan

	kmeans1	kmeans2	kmeans3	kmeans4	kmeans5	kmeans6	kmeans7	kmeans8	kmeans9	kmeans10	R
31	0	0	0	0	0	0	0	0	0	0	10
28	0	0	0	0	0	0	0	0	2	0	9
25	0	0	0	0	0	0	0	0	2	0	9
132	0	0	0	0	0	0	0	0	2	0	9
64	0	0	0	0	0	0	0	0	2	0	9
113	0	0	0	0	0	0	0	0	2	0	9
17	0	0	0	0	0	0	0	0	2	0	9
97	0	0	0	0	0	0	0	0	2	0	9
26	0	3	0	0	0	0	0	0	2	0	8
32	0	0	0	1	0	0	0	0	2	0	8

KESIMPULAN

	Negara
0	Central African Republic
1	Cameroon
2	Burkina Faso
3	Sierra Leone
4	Guinea-Bissau
5	Nigeria
6	Benin
7	Mali
8	Burundi
9	Chad

- Negara yang membutuhkan bantuan atau fokus dari grup HELP ialah nama – nama negara yang tertera pada tabel disamping.
- Faktor yang mempengaruhi dan menjadi dasar analisis pada data ialah angka kematian anak, jumlah fertiliti, nilai harapan hidup, nilai pendapatan, nilai ekspor, nilai impor, dan GDP perkapita