FINAL PROJECT REPORT

SANBERCODE PYTHON-DATA SCIENCE

Dibuat Oleh:

Mohammad Dwiantara Mahardhika

1. Formulasi Masalah

- **Permasalahan**

  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. 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.

- **Tujuan**

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

2. Reading and Understanding Data

- **Reading Data**

  Data yang digunakan berasal dari file 'Data_Negara_HELP.csv' seperti berikut.

	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 pembacaan dataset diatas didapatkan bahwa data yang ada berjumlah 10 kolom dan 167 baris dengan keterangan kolom sebagai berikut.

Kolom	Keterangan
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.

- **Describing Data**

  Dilakukan analisa statistika deskriptif terhadap dataset yang digunakan dengan menggunakan library pandas. Didapatkan data-data statistik dari dataset seperti rata-rata nilai setiap kolom, nilai minimum dan maksimum dari setiap kolom dan lain sebagainya.

```
df.describe()
```

	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

- **Data Info**

  Dari analisa data info ini dapat diketahui tipe data setiap kolom dari dataset beserta jumlah data non-null pada setiaap kolom. Kolom negara bertipe objek tidap perlu diubah karena kolom negara hanya digunakan sebagai ‘index’. Untuk kolom yang lainnya merupakan kolom numerikal, semua kolom numerikal tersebut sudah bertipe float / int sehingga tidak perlu diubah.

```
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
```

3. Exploratory Data Analysis

- **Cleaning Data**

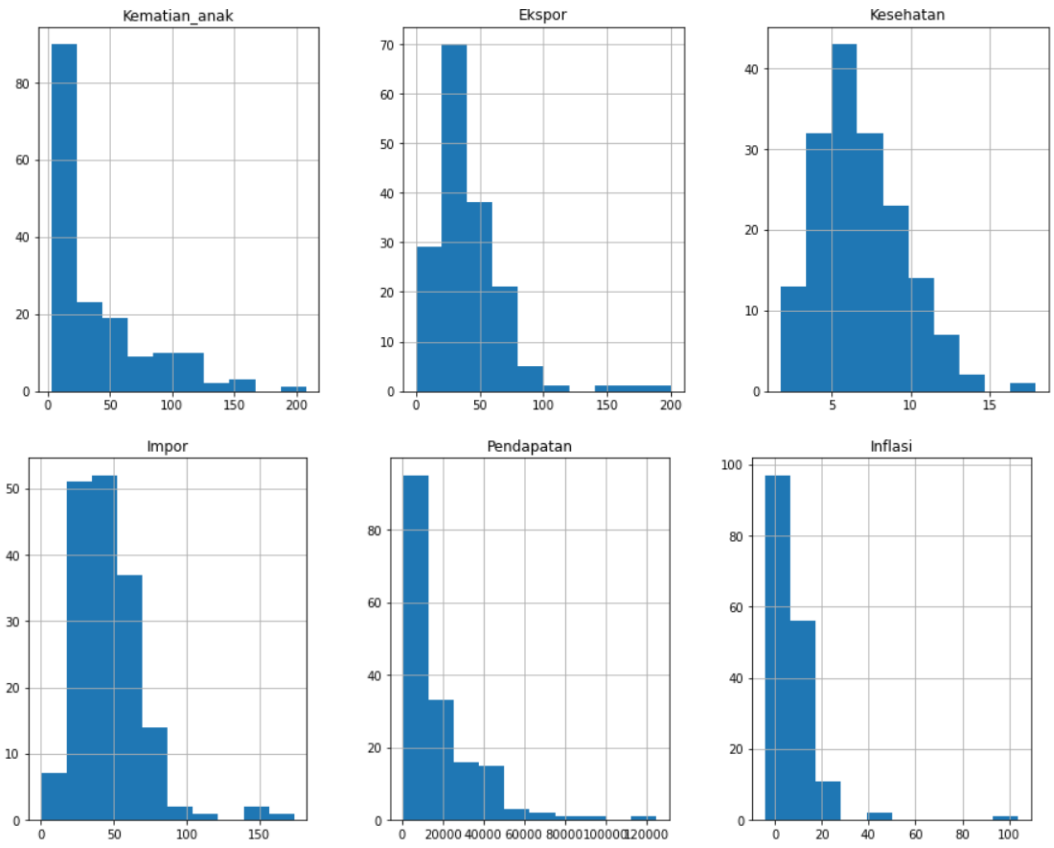
  Setelah dilakukan analisa missing values pada dataset, didapatkan bahwa dataset yang digunakan tidak mempunyai missing values di semua kolom sehingga tidak dilakukan cleaning data.

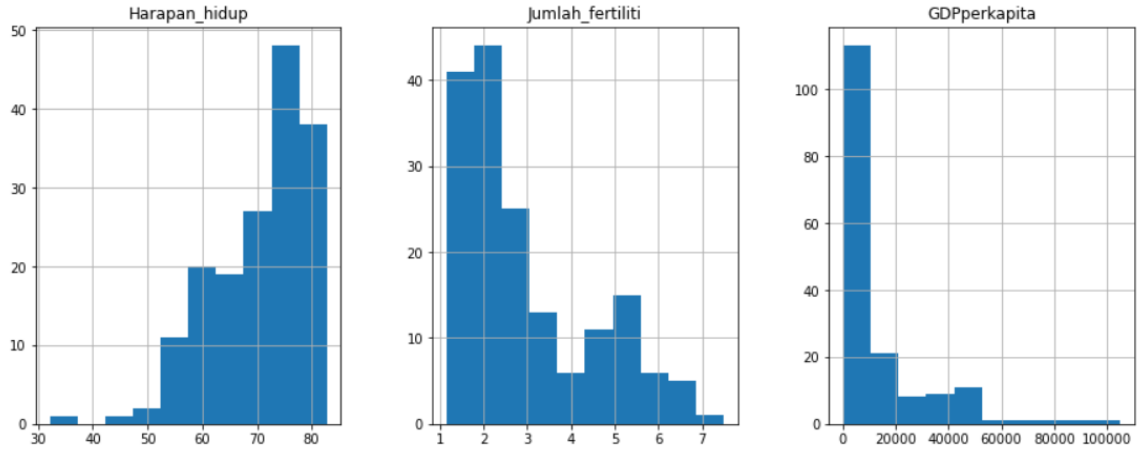
```
: df.isna().sum()

: 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
```

- **Univariate Analysis**

  Dilakukan analisa univariate dengan menggunakan histogram untuk melihat persebaran data pada setiap kolom numerik di dataset.

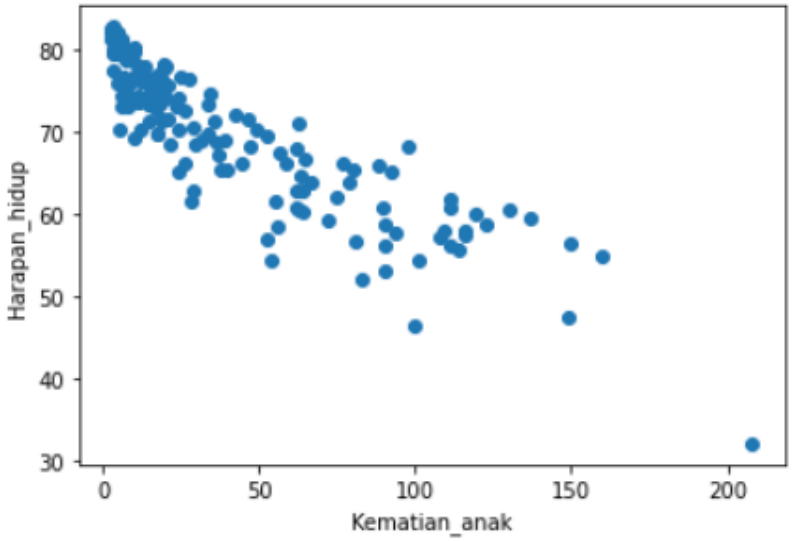




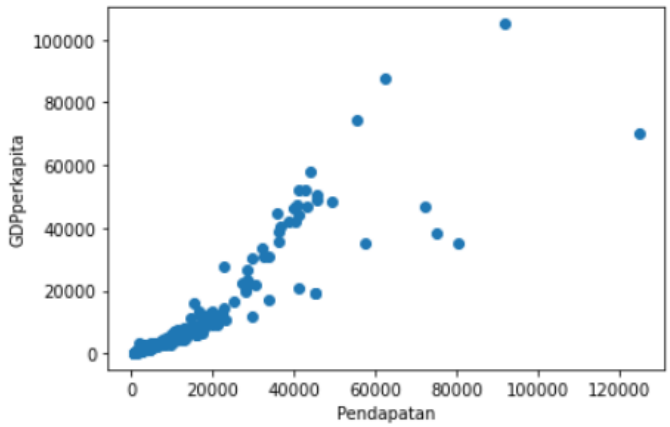
Dengan analisa univariate menggunakan histogram diatas kita dapat mengetahui persebaran data pada masing-masing kolom. Sebagai contoh pada kolom GDPperkapita dapat terlihat ternyata masih banyak negara yang mempunyai GDPperkapita rendah pada range 0-20000 yang lebih dari 100 negara begitu pula pada kolom Pendapatan, masih banyak negara yang mempunyai pendapatan rendah. Selain itu juga dari histogram diatas kita dapat mengetahui bahwa terdapat beberapa negara yang mempunyai kematian_anak yang sangat tinggi dan juga terdapat negara yang mempunyai Harapan_hidup yang sangat rendah.

- **Bivariate Analysis**

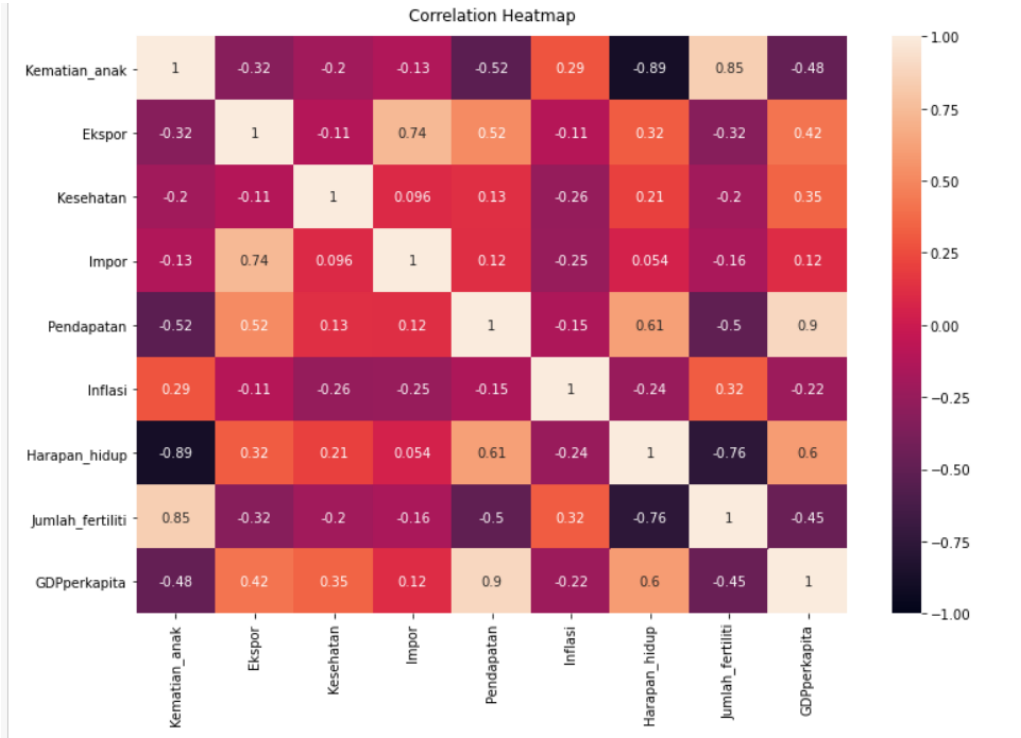
  Dilakukan analisa bivariate terhadap variabel kematian_anak dengan harapan_hidup menggunakan scatter plot. Hasil visualisasi scatter menampilkan bahwa kedua variabel tersebut memiliki korelasi negatif karena semakin tinggi tingkat kematian_anak di suatu negara maka harapan_hidupnya semakin rendah



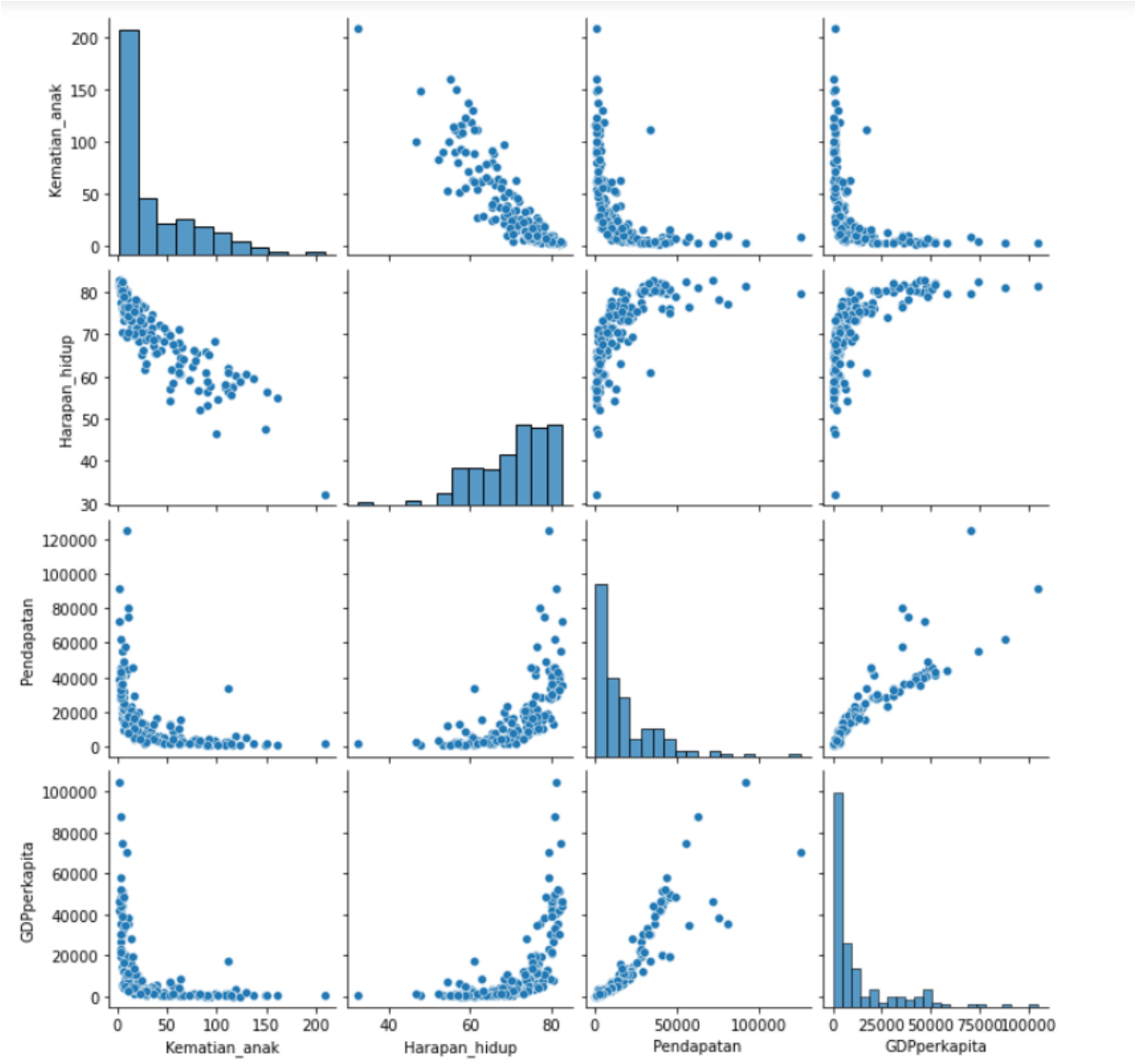
Dilakukan juga analisa terhadap hubungan pendapatan dengan GDPperkapita menggunakan scatter plot. Dari hasil yang didapat dapat disimpulkan bahwa kedua variabel tersebut memiliki korelasi positif karena semakin tinggi pendapatan perorang di suatu negara maka GDPperkapita di negara tersebut juga meningkat.



Dilakukan analisa korelasi keseluruhan variabel data dengan menggunakan heatmap. Dari hasil heatmap ini didapatkan bahwa korelasi antar Kematian_anak dengan Harapan_hidup (-0.89) dan pendapatan dengan GDPperkapita (0.9) mempunyai korelasi paling kuat.

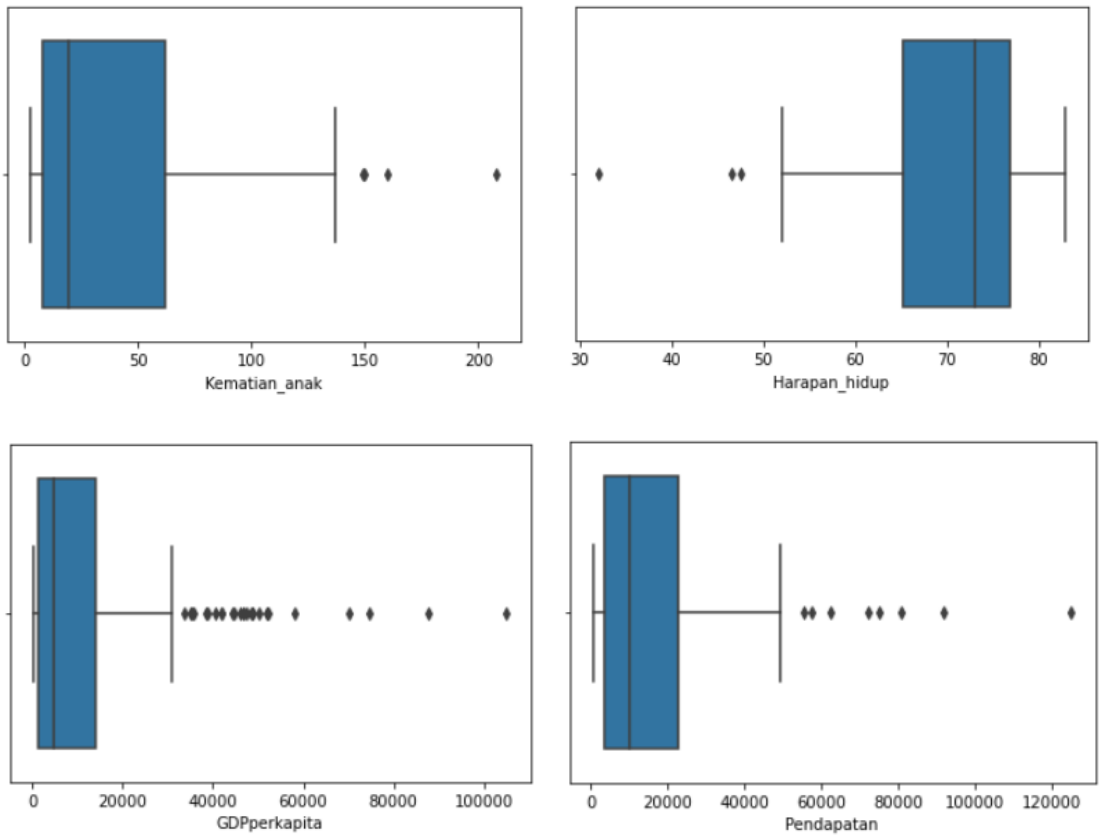


- **Multivariate Analysis**
  Dari hasil analisa univariate dan bivariate yang saya lakukan, saya memutuskan untuk menggunakan Kematian_anak, Harapan_hidup, Pendapatan, dan GDPperkapita untuk dianalisa lebih lanjut. Dari hasil pairplot untuk keempat variabel tersebut didapatkan hubungan dan persebaran data keempat data tersebut.



4. Outlier Treatment

Setelah memilih variabel apa saja yang akan digunakan, kemudian dilakukan pengecekan terdapat data outlier pada kolom-kolom yang akan digunakan tersebut menggunakan boxplot.



Dari hasil boxplot diatas dapat diketahui setiap kolom masih mempunyai outlier. Oleh karena itu dilakukan outlier handling menggunakan **Interquartile Range** masing-masing variabel.

5. **Scaling Data**

Untuk mempermudah tahapa clustering dilakukan scaling data menggunakan StandardScaler.

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

	Negara	Kematian_anak	Harapan_hidup	Pendapatan	GDPperkapita
0	Afghanistan	1.422745	-1.815960	-0.979242	-0.840648
1	Albania	-0.668452	0.911397	-0.136445	-0.344530
2	Algeria	-0.364433	0.938535	0.164409	-0.292632
3	Angola	2.241039	-1.286772	-0.544675	-0.423078
4	Antigua and Barbuda	-0.847454	0.979242	0.792455	0.793020
...	...	...	...	...	...
162	Vanuatu	-0.310448	-0.893272	-0.843503	-0.501627
163	Venezuela	-0.654246	0.789277	0.529081	0.975364
164	Vietnam	-0.478085	0.477191	-0.687504	-0.734467
165	Yemen	0.459544	-0.282670	-0.688517	-0.734467
166	Zambia	1.221012	-2.385856	-0.810075	-0.713427

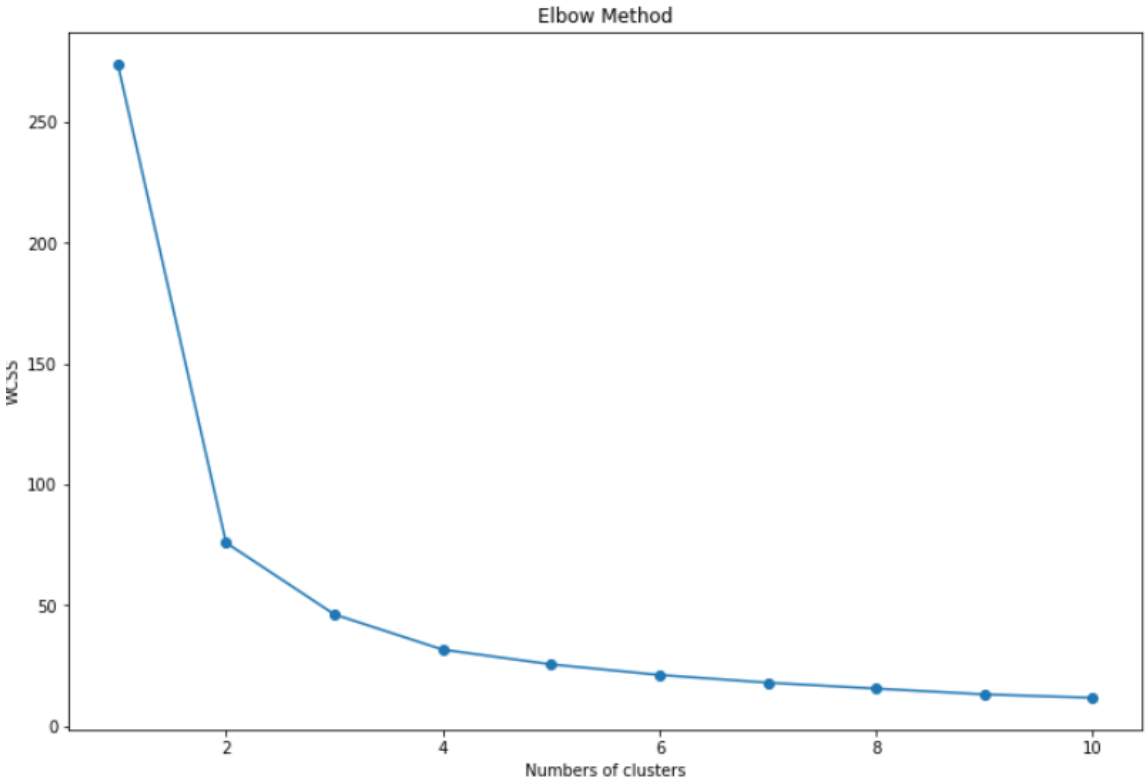
6. **Clustering**

Pada analisa kali ini saya melakukan clustering untuk variabel Kematian_anak dengan Harapan_hidup dan Pendapatan dengan GDPperkapita.

- **Kematian_anak dengan Harapan_hidup (Data Model 1)**
  Data kematian_anak dengan Harapan_hidup.

	Kematian_anak	Harapan_hidup
0	1.422745	-1.815960
1	-0.668452	0.911397
2	-0.364433	0.938535
3	2.241039	-1.286772
4	-0.847454	0.979242
...	...	...
162	-0.310448	-0.893272
163	-0.654246	0.789277
164	-0.478085	0.477191
165	0.459544	-0.282670
166	1.221012	-2.385856

137 rows × 2 columns



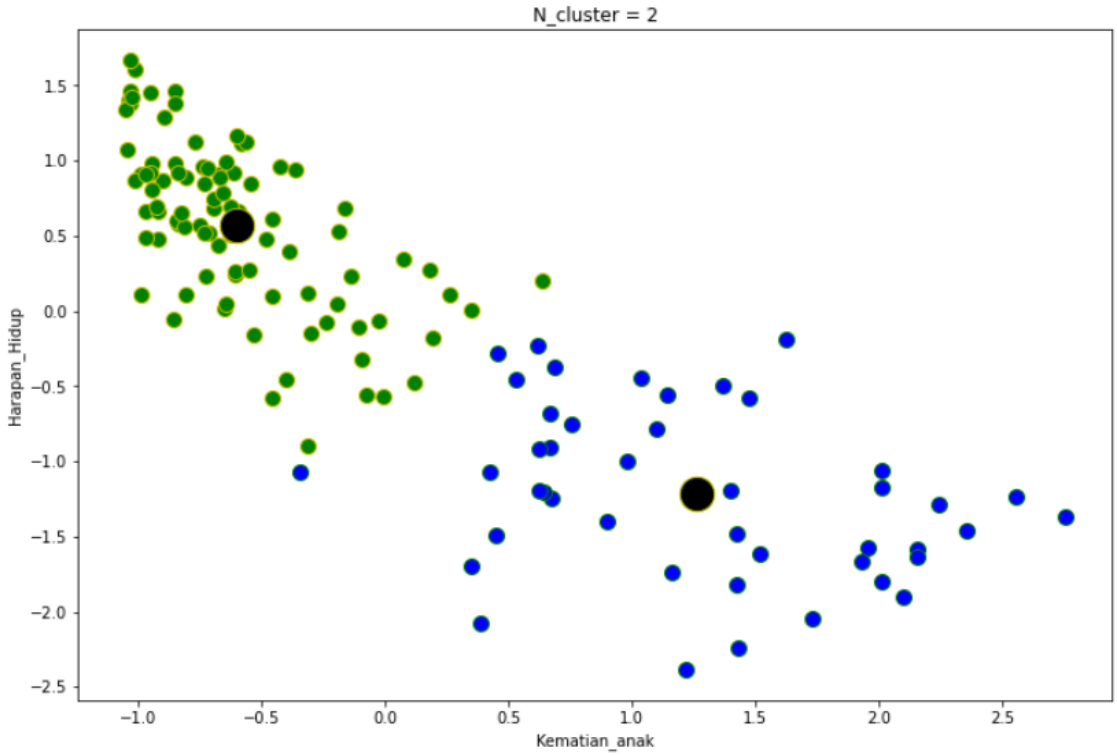
Dari hasil elbow method diatas dilakukan k yg optimal berada antara 2-4. Kemudian dilakukan validasi menggunakan Silhouette Score untuk nilai K = 2 sampai K = 4.

```
K=2
0.6257876220156149

K=3
0.5470548944541891

K=3
0.4772090776702319
```

Dari hasil Silhouette Score didapatkan bahwa K=2 adalah mempunyai nilai terbaik. Oleh karena itu, clustering dilakukan menggunakan K=2 dengan hasil sebagai berikut.



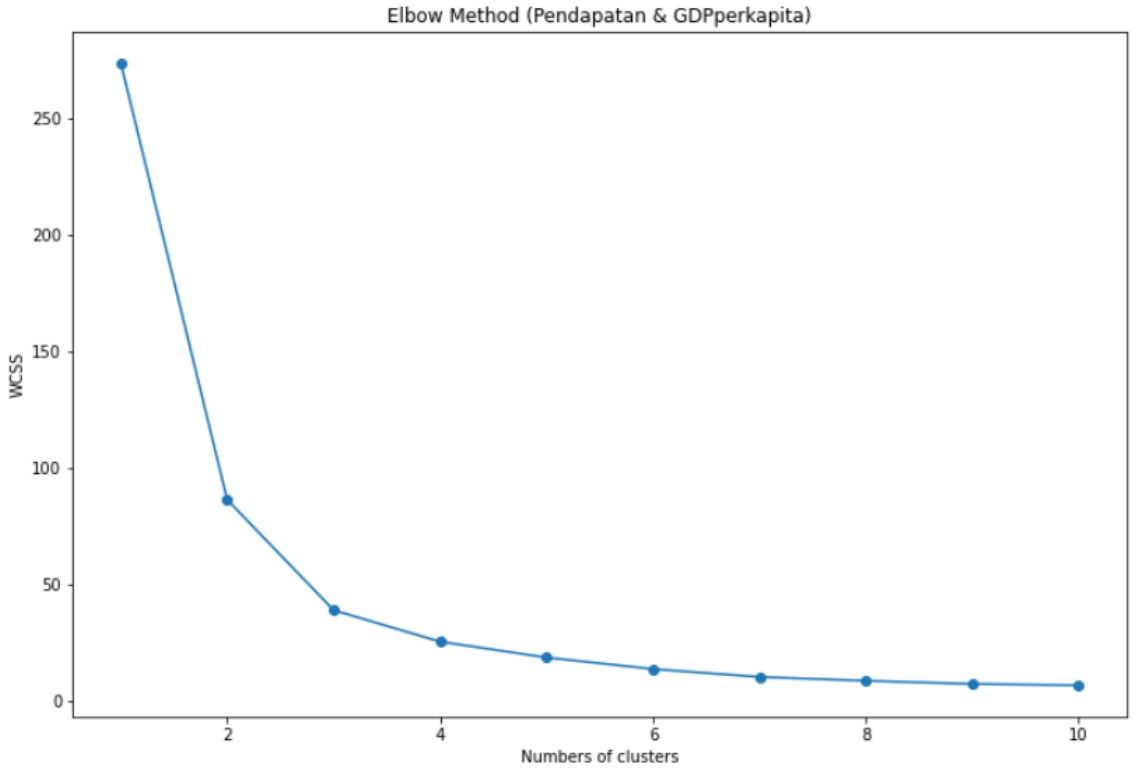
Dari hasil clustering diatas cluster 1 (Hijau) merupakan negara-negara yang mempunyai harapan hidup tinggi dan tingkat kematian_anak rendah sedangkan cluster 2 (Biru) merupakan negara-negara yang mempunyai tingkat harapan_hidup rendah dan tingkat kematian_anak tinggi. Sehingga dapat diambil kesimpulan **Cluster 2 (Biru) adalah negara-negara kandidat fokus bantuan.**

- **Pendapatan dengan GDPperkapita (Data Model 2)**

  Data Pendapatan dengan GDPperkapita

	Pendapatan	GDPperkapita
0	-0.979242	-0.840648
1	-0.136445	-0.344530
2	0.164409	-0.292632
3	-0.544675	-0.423078
4	0.792455	0.793020
...	...	...
162	-0.843503	-0.501627
163	0.529081	0.975364
164	-0.687504	-0.734467
165	-0.688517	-0.734467
166	-0.810075	-0.713427

137 rows × 2 columns



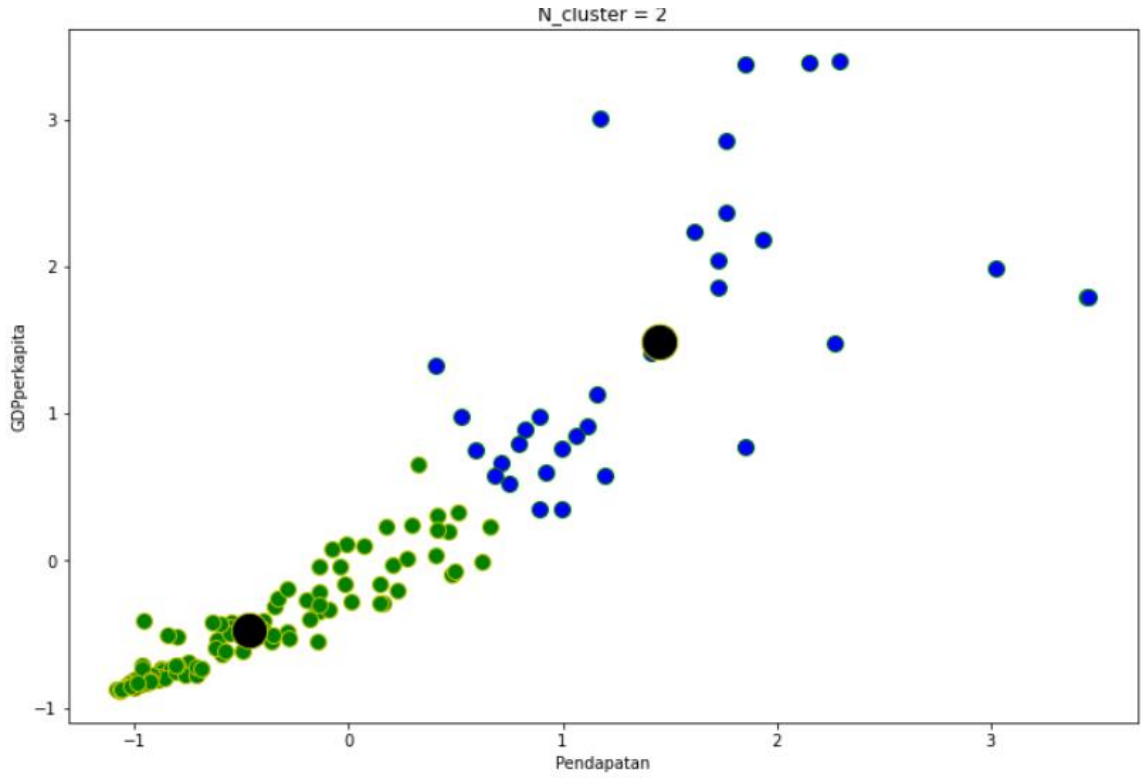
Dari hasil elbow method diatas dilakukan k yg optimal berada antara 2-4. Kemudian dilakukan validasi menggunakan Silhouette Score untuk nilai K = 2 sampai K = 4.

```
K=2
0.6491057643161912

K=3
0.6070391387874169

K=3
0.558301383270302
```

Dari hasil Silhouette Score didapatkan bahwa K=2 adalah mempunyai nilai terbaik. Oleh karena itu, clustering dilakukan menggunakan K=2 dengan hasil sebagai berikut.



Dari hasil clustering diatas cluster 1 (Hijau) merupakan negara-negara yang mempunyai GDPperkapita dan Pendapatan rendah sedangkan cluster 2 (Biru) merupakan negara-negara yang mempunyai GDPperkapitan dan Pendapatan tinggi. Sehingga dapat diambil kesimpulan **Cluster 1 (Hijau) adalah negara-negara kandidat fokus bantuan.**

7. Summary

Terakhir dilakukan merge data hasil clustering data model 1 dengan clustering data model 2. Negara yang mempunyai Harapan_hidup rendah dan kematian_anak **(Cluster 2 Data Model 1)** tinggi serta mempunyai Pendapatan dan GDPperkapita rendah **(Cluster 1 Data Model 2)** adalah negara-negara yang perlu menjadi fokus bantuan. Berikut hasil merge kedua hasil clustering.

	Negara	K
0	Afghanistan	
1	Angola	
2	Benin	
3	Botswana	
4	Burkina Faso	
5	Burundi	
6	Cameroon	
7	Comoros	
8	Congo, Dem. Rep.	
9	Congo, Rep.	
10	Cote d'Ivoire	
11	Eritrea	
12	Gabon	
13	Gambia	
14	Ghana	
15	Guinea	
16	Guinea-Bissau	
17	India	
18	Kenya	
19	Kiribati	
20	Lao	
21	Liberia	
22	Madagascar	
23	Malawi	
24	Mali	
25	Mauritania	
26	Mozambique	
27	Myanmar	
28	Namibia	
29	Niger	
30	Nigeria	
31	Pakistan	
32	Rwanda	
33	Senegal	
34	Solomon Islands	
35	South Africa	
36	Sudan	
37	Tanzania	
38	Togo	
39	Turkmenistan	
40	Uganda	
41	Yemen	
42	Zambia	