

Clustering the Countries by using K-Means

**Studi Kasus :
HELP Internasional**

**python notebook file :
bit.ly/UnsupervisedMLammryf**

Muammar Yusuf Fakhri

Pendahuluan

Latar Belakang

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

Permasalahan

Bagaimana cara mengkategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan ?

negara mana saja yang paling perlu menjadi fokus CEO. ?

Tujuan

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

untuk menentukan negara mana saja yang paling perlu menjadi fokus CEO

Langkah Penggeraan

1 Membaca dan memahami data

2 Data Analisis Eksplorasi
- Pengecekan data
- Univariat Analisis
- Bivariat Analisis
- Multivariat Analisis

3 Outliers Treatment

4 Scaling Data

5 Clustering K means dan Visualisasi

6 Laporan Hasil

Data

Review Data

Berikut data negara dari HELP Internasional

	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

Penjelasan kolom fitur:

- 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.

Informasi Data

RangelIndex: 167 entries, 0 to 166

Data columns (total 10 columns):

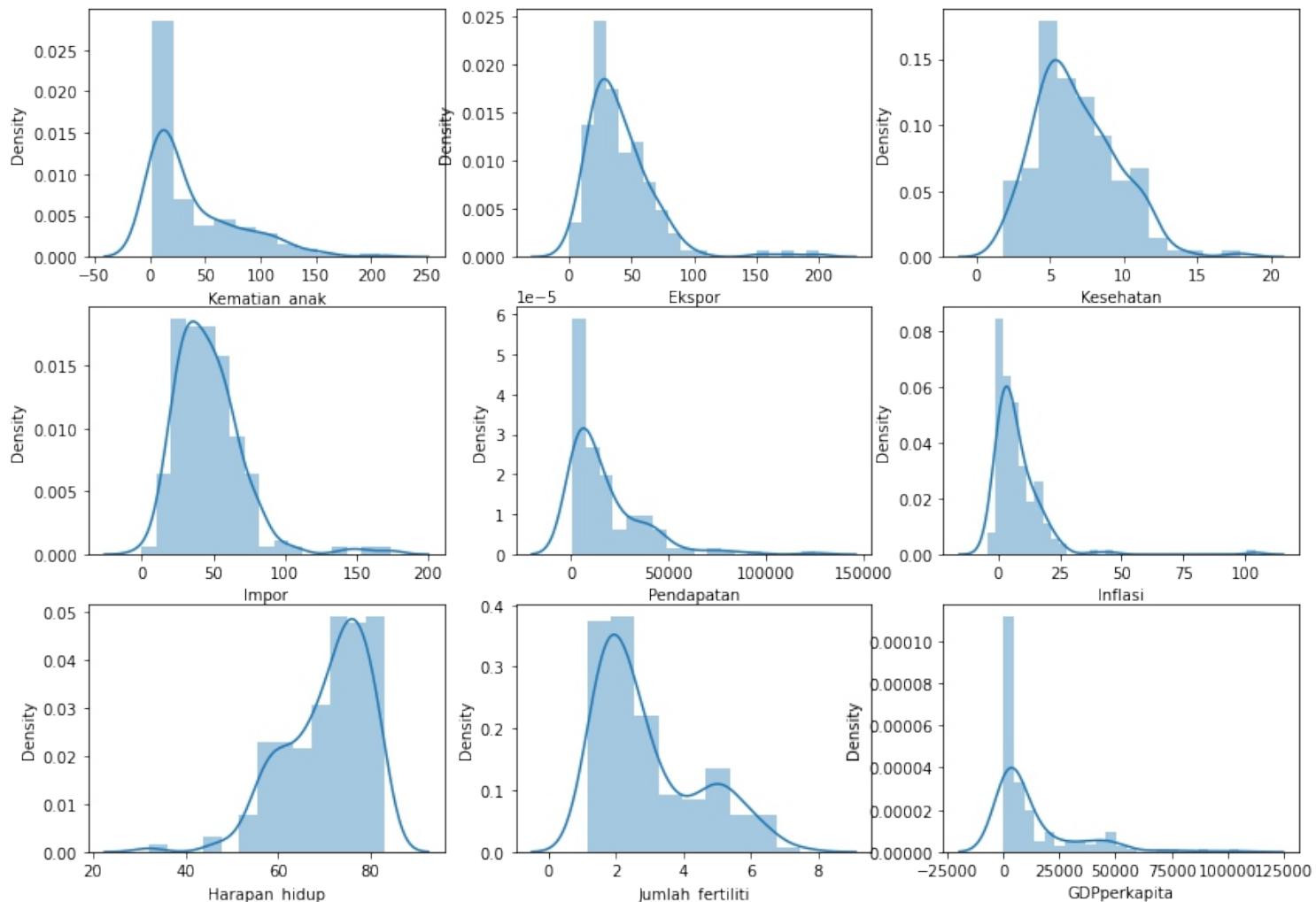
#	Column	Non-Null Count	Dtype
0	Negara	167 non-null	object
1	Kematian_anak	167 non-null	float64
2	Ekspor	167 non-null	float64
3	Kesehatan	167 non-null	float64
4	Impor	167 non-null	float64
5	Pendapatan	167 non-null	int64
6	Inflasi	167 non-null	float64
7	Harapan_hidup	167 non-null	float64
8	Jumlah_fertiliti	167 non-null	float64
9	GDPperkapita	167 non-null	int64

Statistika

	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 Analisis Eksplorasi

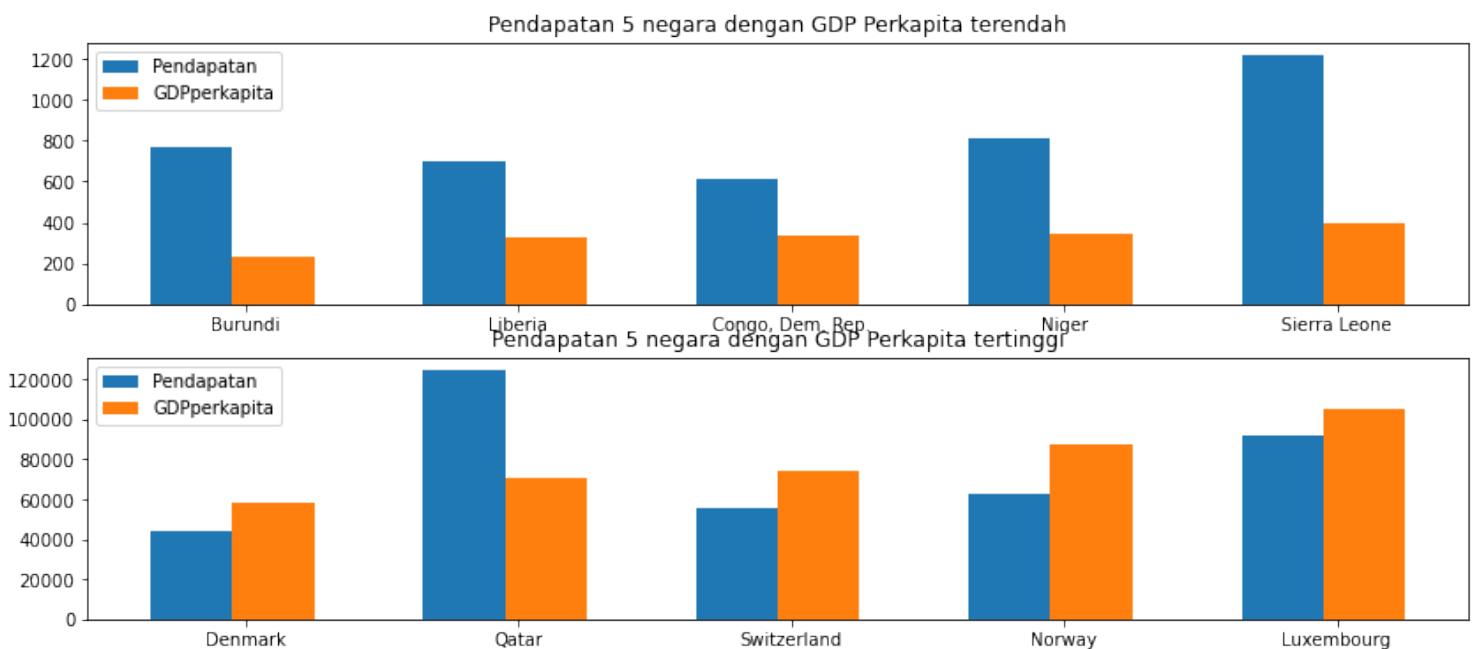
Univariat Analisis



Dari gambar grafik Univariat analisis untuk variabel Harapan_hidup kurva condong ke kiri atau disebut Negative skew. Kurva Negative Skew bernilai $\text{Mean} < \text{Median} < \text{Modus}$.

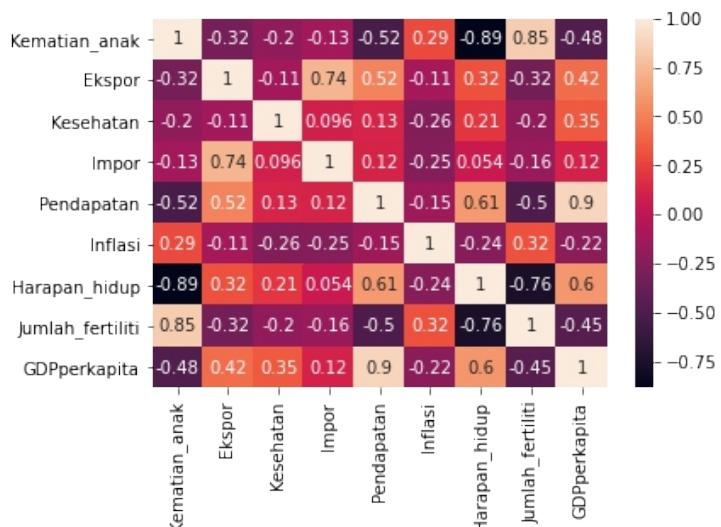
Sedangkan untuk variabel Kematian_anak, Ekspor, Kesehatan, Impor, Pendapatan, Inflasi, Jumlah_fertiliti, dan GDP perkapita kurva condong ke kanan atau disebut Positive skew. Kurva Positive Skew bernilai $\text{Mean} > \text{Median} > \text{Modus}$.

Bivariat Analisis



Variabel yang saya gunakan untuk Bivariat Analisis adalah Pendapatan dan GDP perkapita. Dari gambar grafik untuk 5 Negara dengan GDP perkapita terendah memiliki Pendapatan Lebih besar dari GDP perkapita. Negara dengan nilai Pendapatan terkecil adalah Burundi. Negara dengan nilai GDP perkapita terkecil adalah Congo, Dem Rep.

Multivariat Analisis



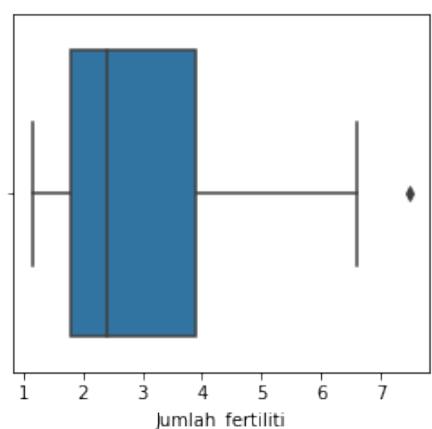
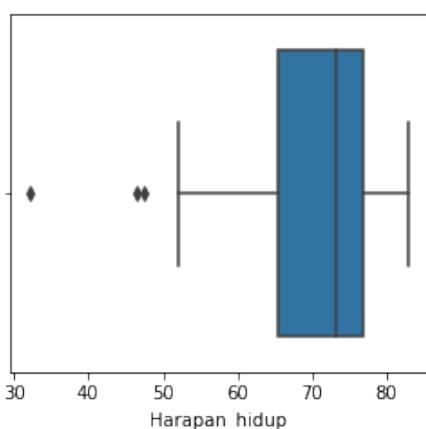
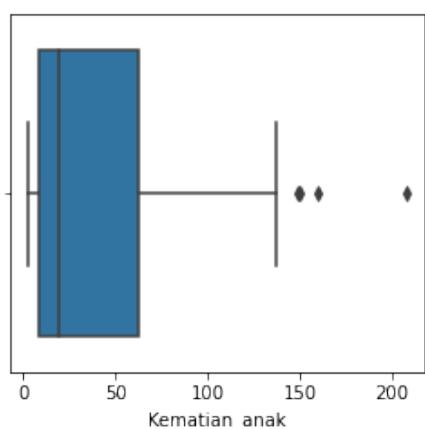
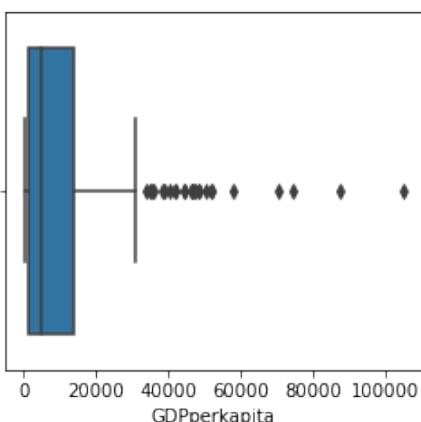
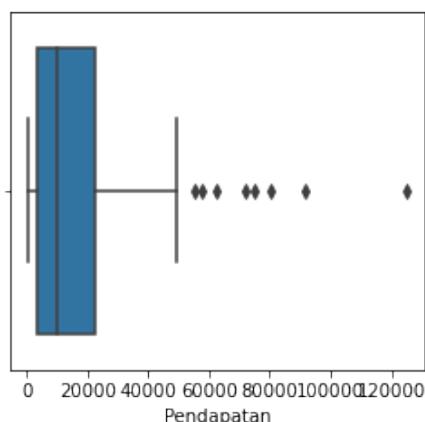
Nilai korelasi berkisar antara 1 sampai -1, nilai semakin mendekati 1 atau -1 berarti hubungan antara dua variabel semakin kuat.

Sebaliknya, jika nilai mendekati 0 berarti hubungan antara dua variabel semakin lemah. Nilai korelasi positif menunjukkan hubungan searah (Var1 naik, maka Var2 naik) sementara nilai korelasi negatif menunjukkan hubungan terbalik (Var1 naik, maka Var2 turun).

Dalam kasus ini saya mengambil 3 nilai korelasi yang kuat untuk menyelesaiannya yaitu

1. Variabel Pendapatan dengan Variabel GDP Perkapita bernilai korelasi 0.9
2. Variabel Kematian Anak dengan Variabel Harapan Hidup bernilai korelasi -0.89
3. Variabel Kematian Anak dengan Variabel Jumlah fertilit bernilai korelasi 0.85

Data Pencilan



Data pencilan Titik data yang berada di luar distribusi kumpulan data secara keseluruhan. Untuk memeriksa data pencilan saya menggunakan Metode Rentang Interkuartil pada python. Diatas adalah Grafik data pencilan. kemudian setelah menemukan data pencilan data ini akan di drop dari tabel sehingga tersisa 136 Negara. sebagai berikut

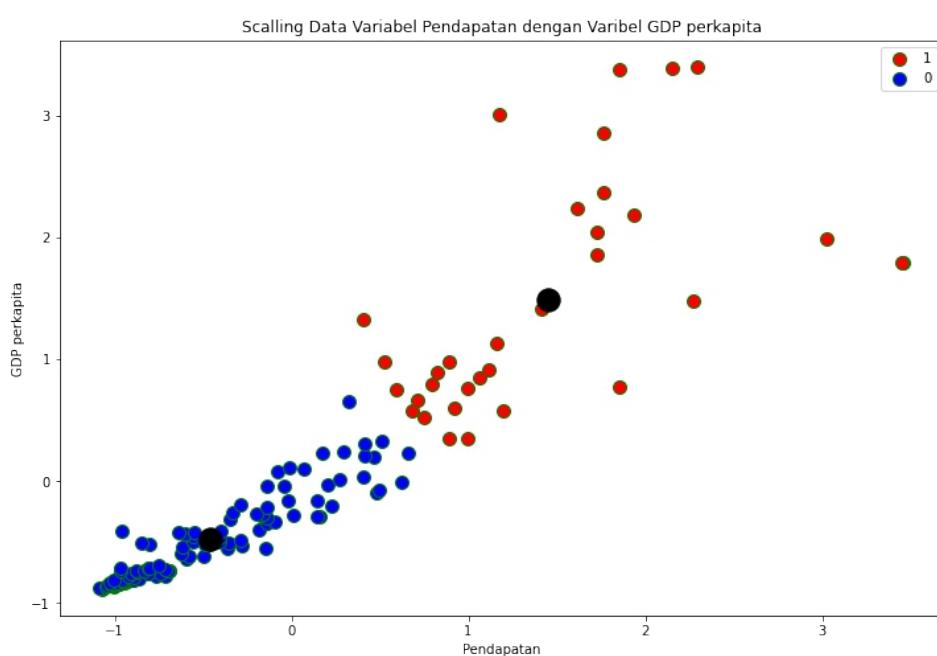
	Negara	Pendapatan	GDPperkapita	Kematian_anak	Jumlah_fertiliti	Harapan_hidup
0	Burundi	764.0	231.0	93.6	6.26	57.7
1	Liberia	700.0	327.0	89.3	5.02	60.8
2	Congo, Dem. Rep.	609.0	334.0	116.0	6.54	57.5
3	Madagascar	1390.0	413.0	62.2	4.60	60.8
4	Mozambique	918.0	419.0	101.0	5.56	54.5
...
132	Greece	28700.0	26900.0	3.9	1.48	80.4
133	Bahamas	22900.0	28000.0	13.8	1.86	73.8
134	Israel	29600.0	30600.0	4.6	3.03	81.4
135	Spain	32500.0	30700.0	3.8	1.37	81.9
136	Cyprus	33900.0	30800.0	3.6	1.42	79.9

K-Means

Dari tabel hasil data pencarian dilakukan K-means data atau mengelompokan (cluster) data untuk menentukan negara man saja yang mendapatkan dana.

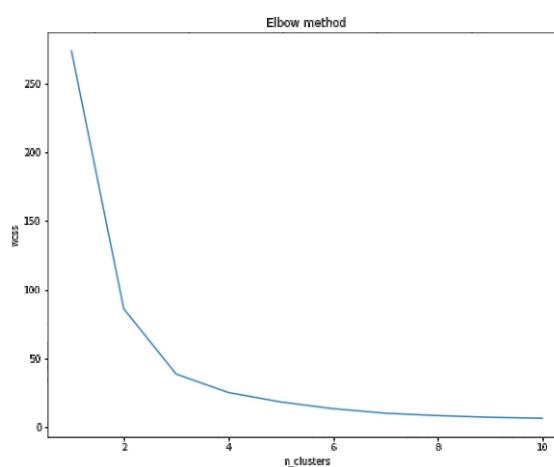
K-Means data ke-1

Berdasarkan analisis multivariat variabel pendapatan dengan variabel GDPperkapita memiliki nilai korelasi yang kuat pertama yaitu 0.9. Maka untuk Scalling data pertama menggunakan variabel pendapatan dan GDP per kapita.



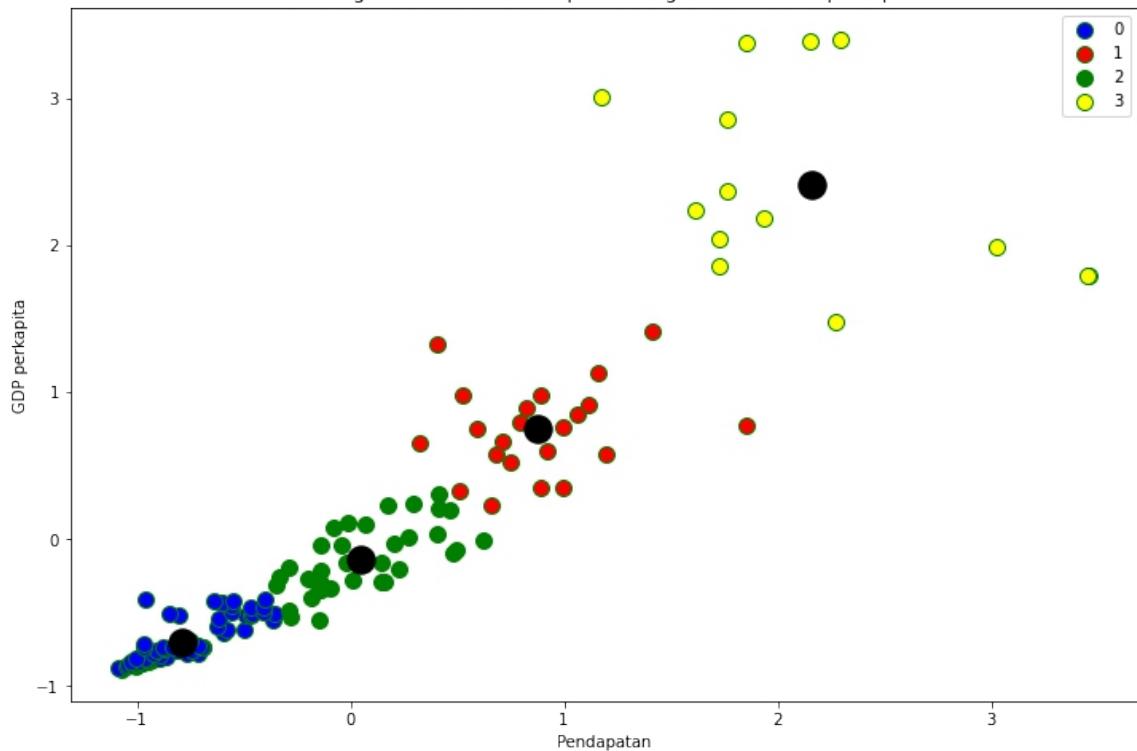
Dilakukan scaling data dengan Kmeans n=2 Karena masih sulit dipahami maka dilakukan elbow method pada langkah selanjutnya

2 |



Elbow method variabel Pendapatan dengan GDP perkapita, untuk mengetahui nilai n yang disarankan untuk scaling data. dari grafik nilai yang disarankan n=4

Scaling Data Variabel Pendapatan dengan Varibel GDP perkapita



Dilakukan clustering n=4 sesuai saran dari Elbow method dengan cluster :

1. cluster 0 (biru) negara dengan pendapatan rendah dan GDP perkapita rendah.
2. cluster 1 (merah) negara dengan pendapatan cukup tinggi dan GDP perkapita cukup tinggi.
3. cluster 2 (hijau) negara dengan pendapatan cukup rendah dan GDP perkapita cukup rendah.
4. cluster 3 (kuning) negara dengan pendapatan tinggi dan GDP perkapita tinggi

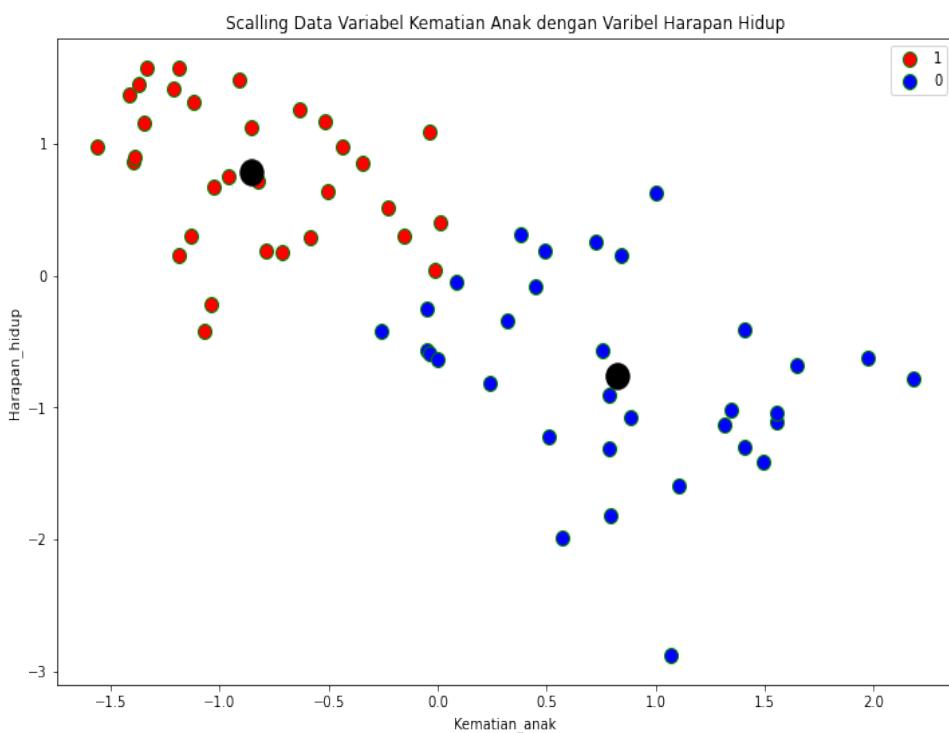
Selanjutnya saya akan memilih cluster 0 (biru) negara dengan pendapatan rendah dan GDP perkapita rendah untuk dilakukan scaling data kedua. dari scaling kesatu tersisa 65 Negara/berikut adalah gambaran data hasil scaling kesatu.

	Negara	Pendapatan	GDPperkapita	Kematian_anak	Jumlah_fertiliti	Harapan_hidup	K_means_labels
0	Burundi	764.0	231.0	93.6	6.26	57.7	0
1	Liberia	700.0	327.0	89.3	5.02	60.8	0
2	Congo, Dem. Rep.	609.0	334.0	116.0	6.54	57.5	0
3	Madagascar	1390.0	413.0	62.2	4.60	60.8	0
4	Mozambique	918.0	419.0	101.0	5.56	54.5	0
...
60	Samoa	5400.0	3450.0	18.9	4.34	71.5	0
61	Angola	5900.0	3530.0	119.0	6.16	60.1	0
62	Tonga	4980.0	3550.0	17.4	3.91	69.9	0
63	Timor-Leste	1850.0	3600.0	62.6	6.23	71.1	0
64	Fiji	7350.0	3650.0	24.1	2.67	65.3	0

K-Means data ke -2

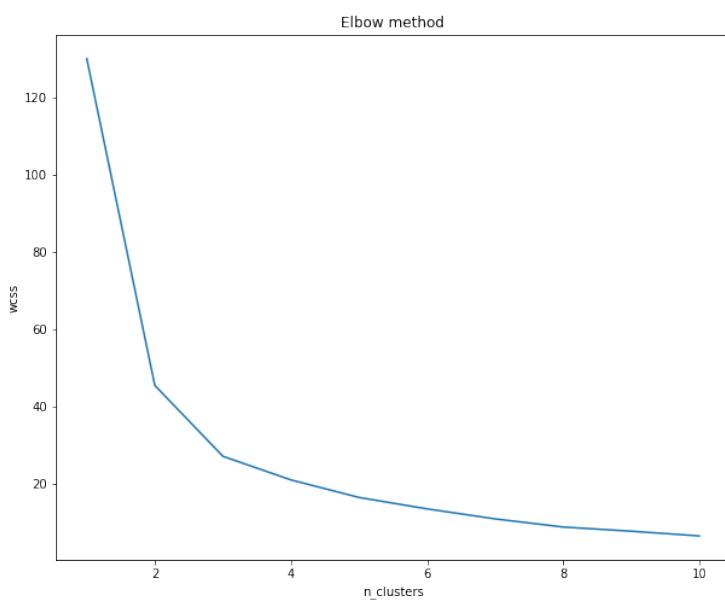
Berdasarkan analisis multivariat variabel Kematian anak dengan variabel Harapan Hidup memiliki nilai korelasi yang kuat pertama yaitu -0.89. Maka untuk Scalling data ke-2 menggunakan variabel Kematian anak dan Harapan hidup

1



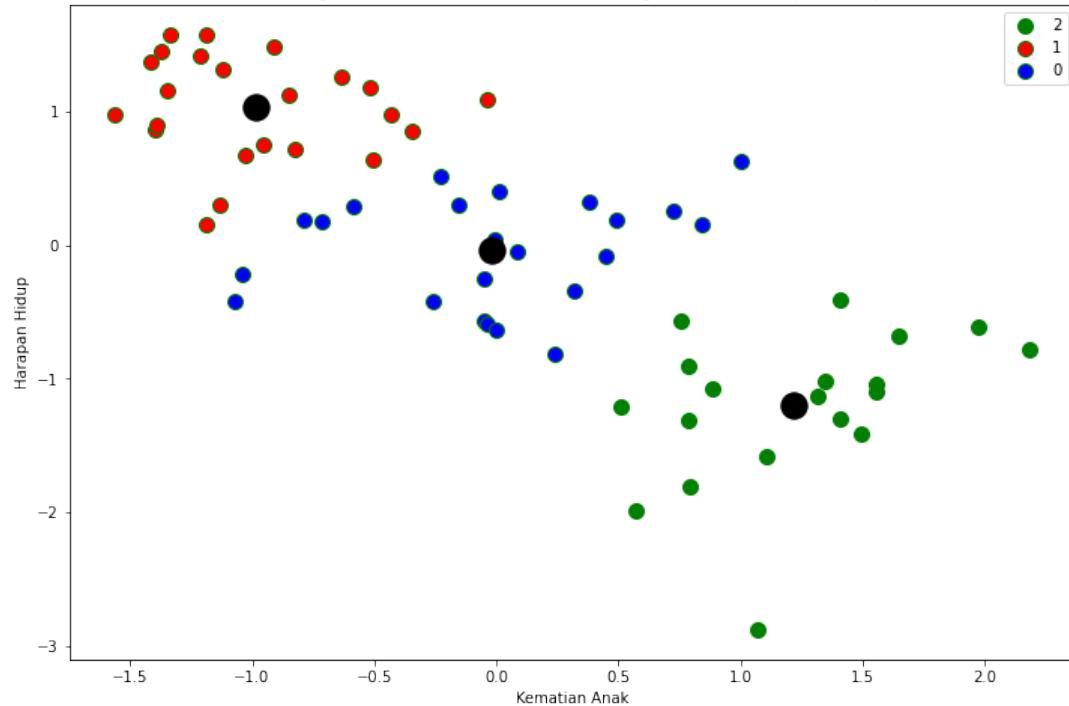
Dilakukan scalling data dengan Kmeans n=2. Karena masih sulit dipahami maka dilakukan elbow method pada kangkah selanjutnya

2



Elbow method variabel Kematian anak dan Harapan hidup, untuk mengetahui nilai n yang disarankan untuk scalling data. dari grafik nilai yang disarankan n=3

Scaling Data Variabel Kematian Anak dengan Varibel Harapan Hidup



Dilakukan clustering n=3 sesuai saran dari Elbow method dengan cluster :

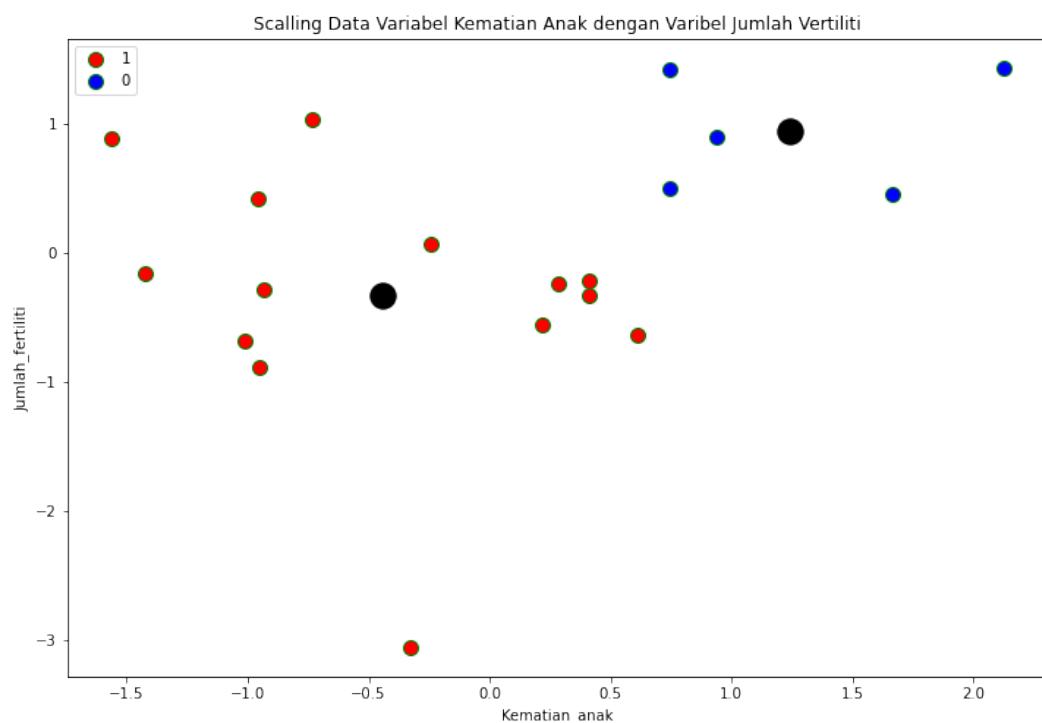
1. cluster 2 (biru) negara dengan Kematian Anak tinggi dan Harapan hidup rendah.
2. cluster 1 (merah) negara dengan Kematian Anak sedang dan Harapan hidup sedang.
3. cluster 0 (hijau) negara dengan Kematian Anak rendah dan Harapan hidup tinggi..

Selanjutnya saya akan memilih cluster 2 (biru) negara dengan Kematian Anak tinggi dan Harapan hidup rendah untuk dilakukan scaling data ketiga. dari scaling kedua tersisa 19 Negara.berikut adalah gambaran data hasil scaling kedua.

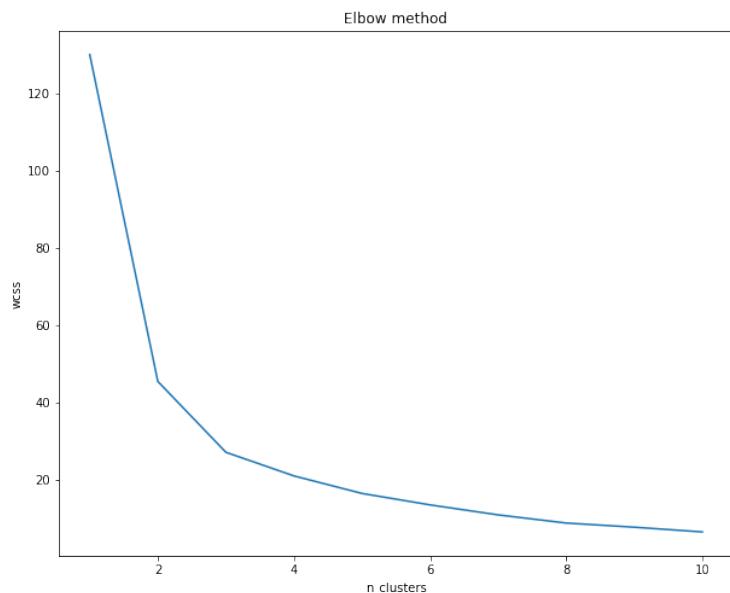
	Negara	Pendapatan	GDPperkapita	Kematian_anak	Jumlah_fertiliti	Harapan_hidup	K_means_labels	K_means_labels2
0	Burundi	764.0	231.0	93.6	6.26	57.7	0	2
1	Liberia	700.0	327.0	89.3	5.02	60.8	0	2
2	Congo, Dem. Rep.	609.0	334.0	116.0	6.54	57.5	0	2
3	Mozambique	918.0	419.0	101.0	5.56	54.5	0	2
4	Malawi	1030.0	459.0	90.5	5.31	53.1	0	2
5	Togo	1210.0	488.0	90.3	4.87	58.7	0	2
6	Guinea-Bissau	1390.0	547.0	114.0	5.05	55.6	0	2
7	Afghanistan	1610.0	553.0	90.2	5.82	56.2	0	2
8	Burkina Faso	1430.0	575.0	116.0	5.87	57.9	0	2
9	Uganda	1540.0	595.0	81.0	6.15	56.8	0	2
10	Guinea	1190.0	648.0	109.0	5.34	58.0	0	2
11	Mali	1870.0	708.0	137.0	6.55	59.5	0	2
12	Benin	1820.0	758.0	111.0	5.36	61.8	0	2
13	Lesotho	2380.0	1170.0	99.7	3.30	46.5	0	2
14	Cote d'Ivoire	2690.0	1220.0	111.0	5.27	56.3	0	2
15	Cameroon	2660.0	1310.0	108.0	5.11	57.3	0	2
16	Zambia	3280.0	1460.0	83.1	5.40	52.0	0	2
17	Nigeria	5150.0	2330.0	130.0	5.84	60.5	0	2
18	Angola	5900.0	3530.0	119.0	6.16	60.1	0	2

K-Means data ke -3

Berdasarkan analisis multivariat variabel Kematian anak dengan variabel Jumlah Fertiliti memiliki nilai korelasi yang kuat pertama yaitu 0.85. Maka untuk Scalling data ke-3 menggunakan variabel Kematian anak dan Jumlah fertiliti

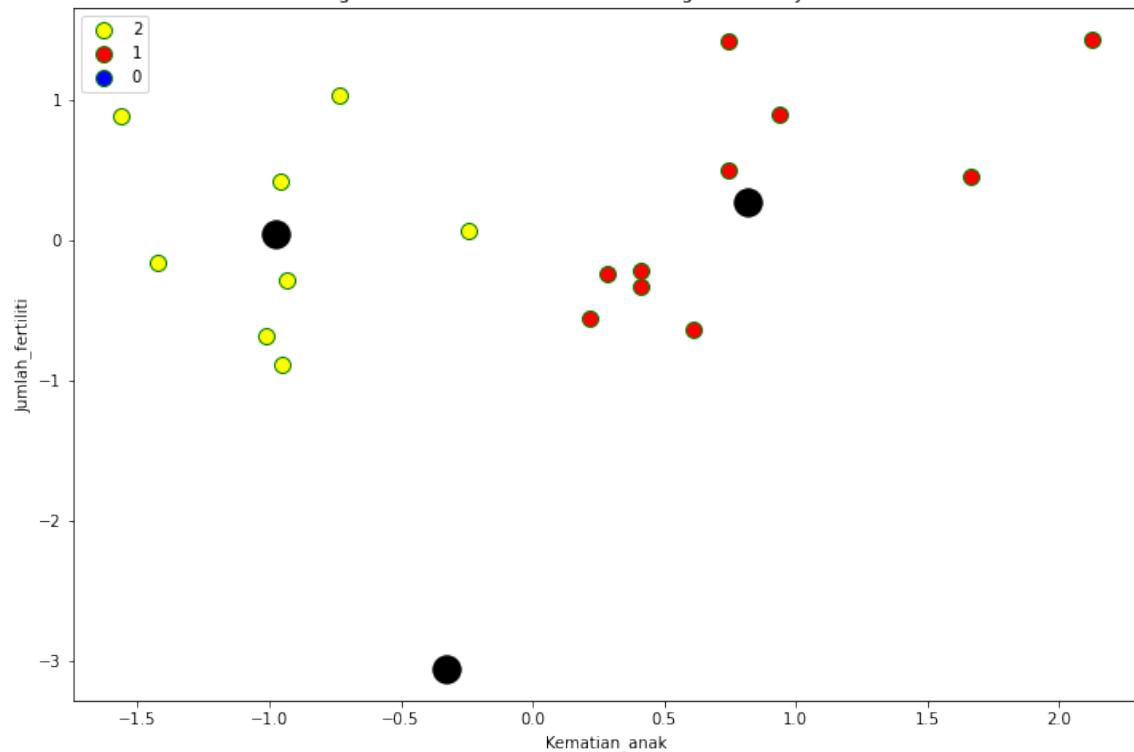


Dilakukan scalling data dengan Kmeans n=2. Karena masih sulit dipahami maka dilakukan elbow method pada kangkah selanjutnya



Elbow method variabel Kematian anak dan Jumlah Fertiliti, untuk mengetahui nilai n yang disarankan untuk scalling data. dari grafik nilai yang disarankan n=3

3



Dilakukan clustering n=3 sesuai saran dari Elbow method dengan cluster :

1. cluster 2 (kuning) negara dengan Kematian Anak rendah dan jumlah fertiliti tinggi.
2. cluster 1 (merah) negara dengan Kematian Anak tinggi dan jumlah fertiliti tinggi.
3. cluster 0 (biru) negara dengan Kematian Anak rendah dan Harapan hidup rendah.

Selanjutnya saya akan memilih cluster 1(merah) negara dengan Kematian Anak tinggi dan jumlah fertiliti tinggi.untuk yang medapatkan uang dari HELP Internasional 1 Negara berikut adalah

	Negara	Pendapatan	GDPperkapita	Kematian_anak	Jumlah_fertiliti	Harapan_hidup	K_means_labels3
0	Congo, Dem. Rep.	609.0	334.0	116.0	6.54	57.5	1
1	Guinea-Bissau	1390.0	547.0	114.0	5.05	55.6	1
2	Burkina Faso	1430.0	575.0	116.0	5.87	57.9	1
3	Guinea	1190.0	648.0	109.0	5.34	58.0	1
4	Mali	1870.0	708.0	137.0	6.55	59.5	1
5	Benin	1820.0	758.0	111.0	5.36	61.8	1
6	Cote d'Ivoire	2690.0	1220.0	111.0	5.27	56.3	1
7	Cameroon	2660.0	1310.0	108.0	5.11	57.3	1
8	Nigeria	5150.0	2330.0	130.0	5.84	60.5	1
9	Angola	5900.0	3530.0	119.0	6.16	60.1	1

Negara- negara yang mendapatkan dana dari HELP Internasional adalah

Negara

- 0 Congo, Dem. Rep.
- 1 Guinea-Bissau
- 2 Burkina Faso
- 3 Guinea
- 4 Mali
- 5 Benin
- 6 Cote d'Ivoire
- 7 Cameroon
- 8 Nigeria
- 9 Angola

Jika saat ini HELP Internasional telah berhasil mengumpulkan sekitar \$ 10 juta.
Maka masing - Masing negara mendapatkan \$1 juta