Designing the Priorities

Creator:

David Putra Yohast

Sanbercode Python - Data Science (Batch **25**)

Final Project

PART I

LEAST DEVELOPED COUNTRIES

HELP

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.

Ada apa dengan HELP?

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.

Negara terbelakang / *Least Developed Countries*

Negara-negara kurang berkembang (LDCs) adalah negara-negara berpenghasilan rendah menghadapi hambatan struktural yang parah untuk pembangunan berkelanjutan. Mereka sangat rentan terhadap guncangan ekonomi dan lingkungan dan memiliki tingkat aset manusia yang rendah (United Nations Department of Economic and Social Affairs).

PART II
DATASET

Dataset Negara HELP

Dataset ini berisikan list negara dan juga fitur-fiturnya yang berjumlah 9 fitur, antara lain:

- 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.

	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

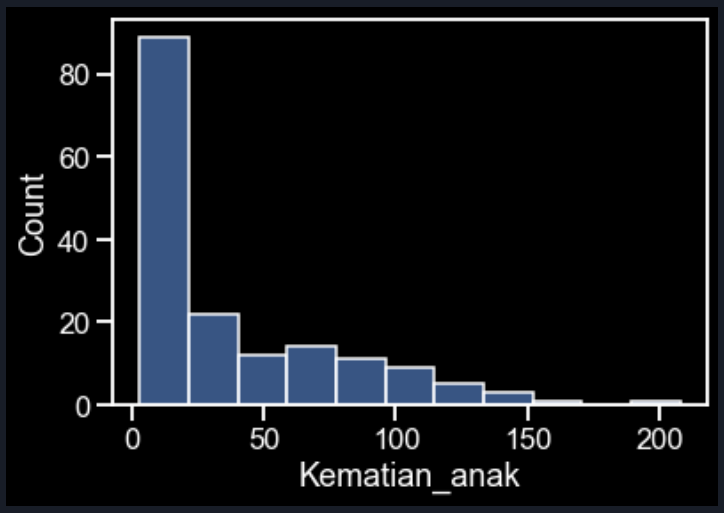
167 rows × 10 columns

Informasi Dataset

Dataset ini memiliki 167 baris, 10 kolom, dan tidak memiliki null value. Dan semua tipe data kolomnya adalah Float, kecuali untuk kolom Pendapatan dan GDP per kapita.

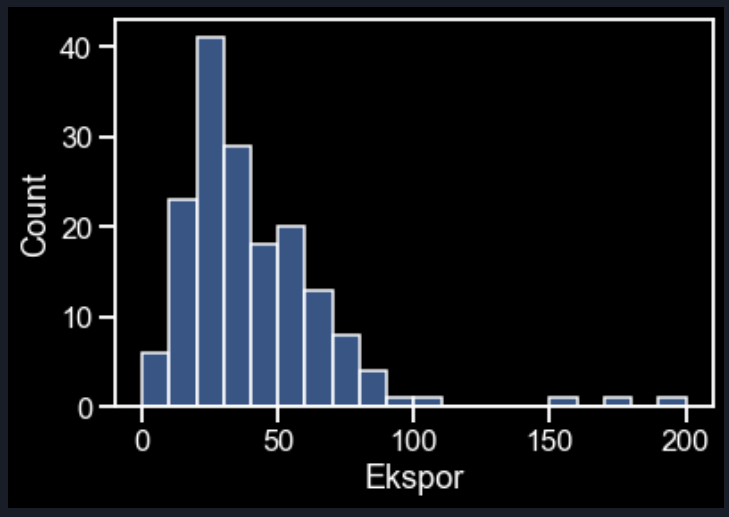
Fitur : Kematian Anak

Rata-rata kematian anak di bawah 5 tahun adalah 38.27 anak. Dengan nilai minimumnya 2.6 dan nilai maksimumnya 208. Interval paling umum adalah 0-20 kematian, dengan jumlah negara di interval tersebut lebih dari 90.



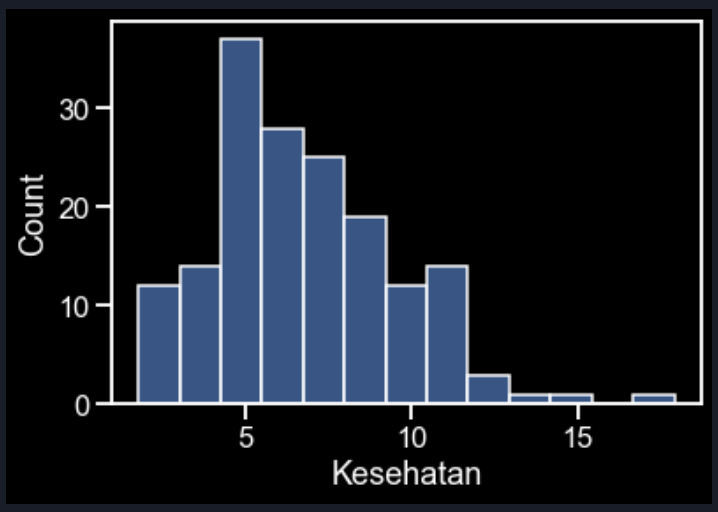
Fitur : Ekspor

Rata-rata ekspor adalah 41.108976. Dengan nilai minimumnya 0.109000 dan nilai maksimumnya 200. Interval paling umum adalah 20-30, dengan jumlah negara di interval tersebut lebih dari 40.



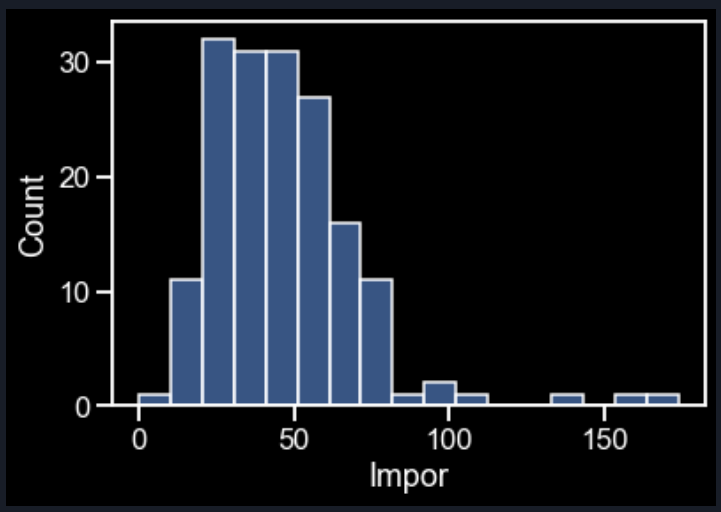
Fitur : Kesehatan

Rata-rata pengeluaran kesehatan adalah 6.815689. Dengan nilai minimumnya 1.81 dan nilai maksimumnya 17.9. Interval paling umum adalah 3-6, dengan jumlah negara di interval tersebut lebih dari 30.



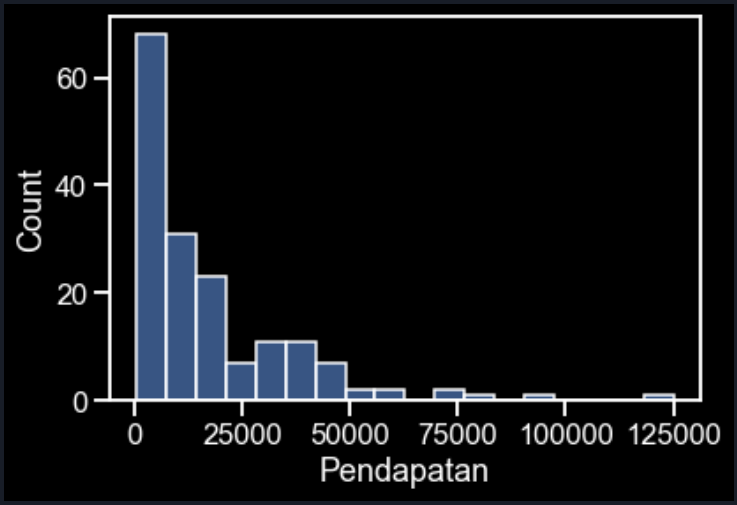
Fitur : Impor

Rata-rata Impor barang dan jasa adalah 46.890215. Dengan nilai minimumnya 0.065900 dan nilai maksimumnya 174. Interval paling umum adalah 20-50, dengan jumlah negara di interval tersebut lebih dari 90 (tiga interval tertinggi).



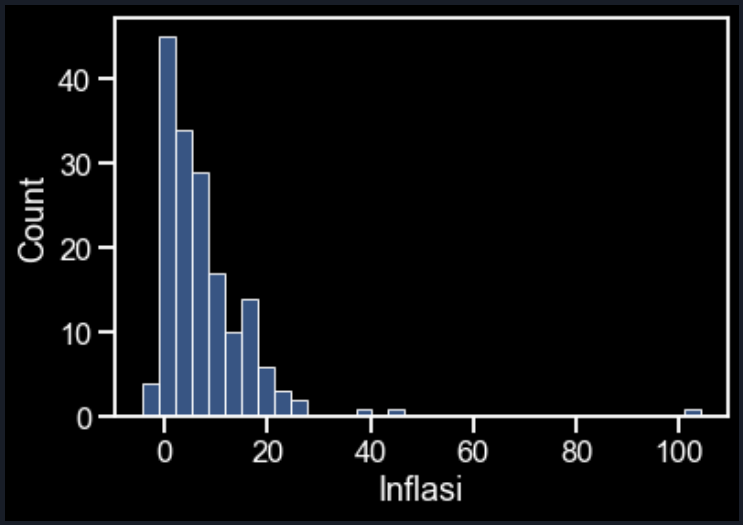
Fitur : Pendapatan

Rata-rata Pendapatan perorangan adalah 17144.688623. Dengan nilai minimumnya 609 dan nilai maksimumnya 125000 . Interval paling umum adalah 0-10000, dengan jumlah negara di interval tersebut lebih dari 60.



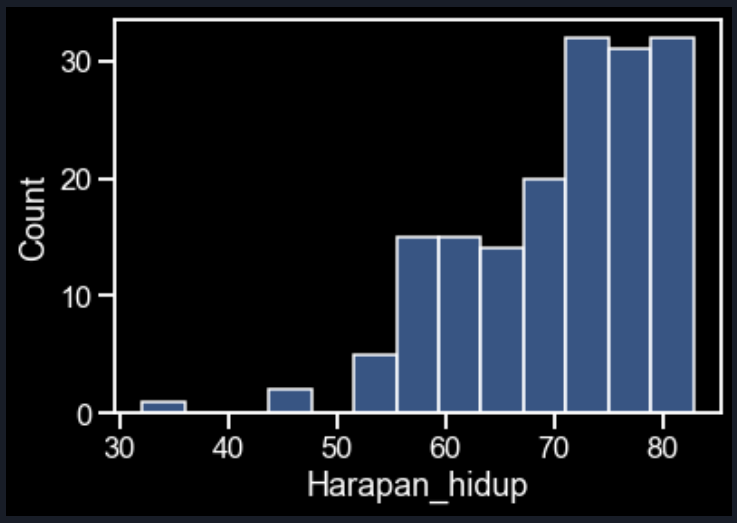
Fitur : Inflasi

Rata-rata inflasi adalah 7.781832. Dengan nilai minimumnya -4.21 dan nilai maksimumnya 104 . Interval paling umum adalah -1 – 2%, dengan jumlah negara di interval tersebut lebih dari 40.



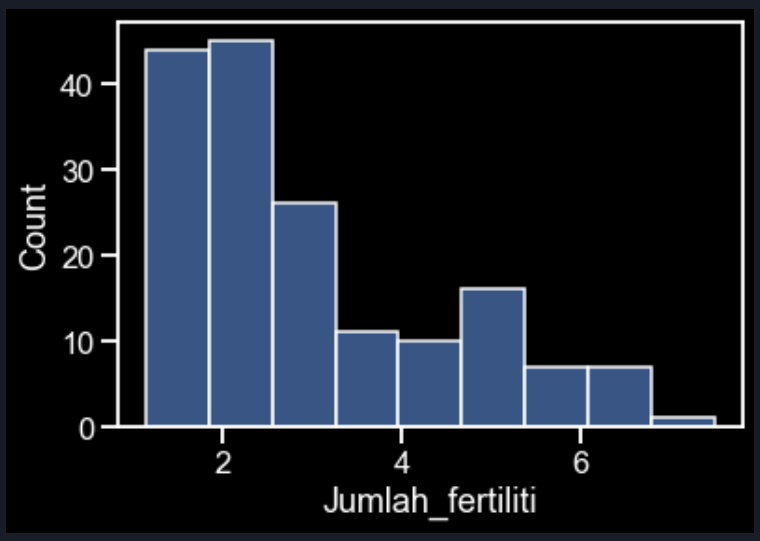
Fitur : Harapan Hidup

Rata-rata umur harapan hidup adalah 70.555689. Dengan nilai minimumnya 32.1 dan nilai maksimumnya 82.8 . Interval paling umum adalah 70-82, dengan jumlah negara di interval tersebut lebih dari 90 (gabungan tiga interval).



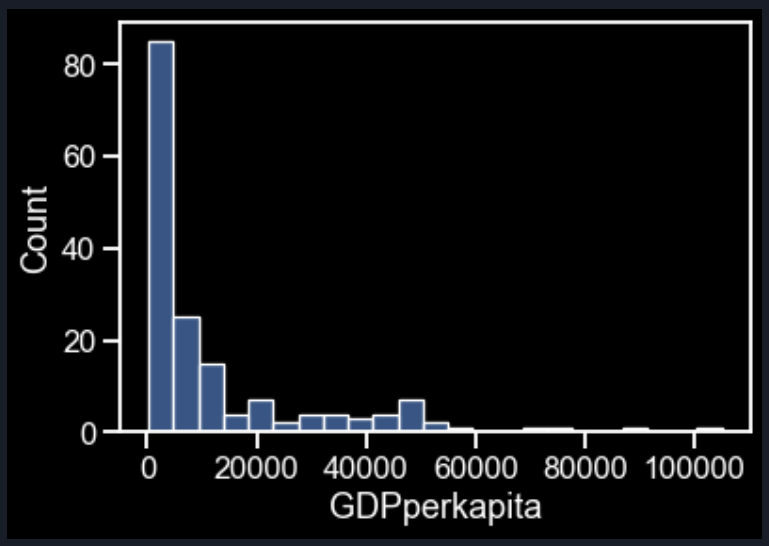
Fitur : Jumlah Fertiliti

Rata-rata jumlah fertiliti adalah 2.947964. Dengan nilai minimumnya 1.15 dan nilai maksimumnya 7.49. Interval paling umum adalah 0-2.5, dengan jumlah negara di interval tersebut lebih dari 80 (gabungan dua interval).



Fitur : GDP

Rata-rata GDP adalah 12964.155689. Dengan nilai minimumnya 231dan nilai maksimumnya 105000. Interval paling umum adalah 0-5000, dengan jumlah negara di interval tersebut lebih dari 80.

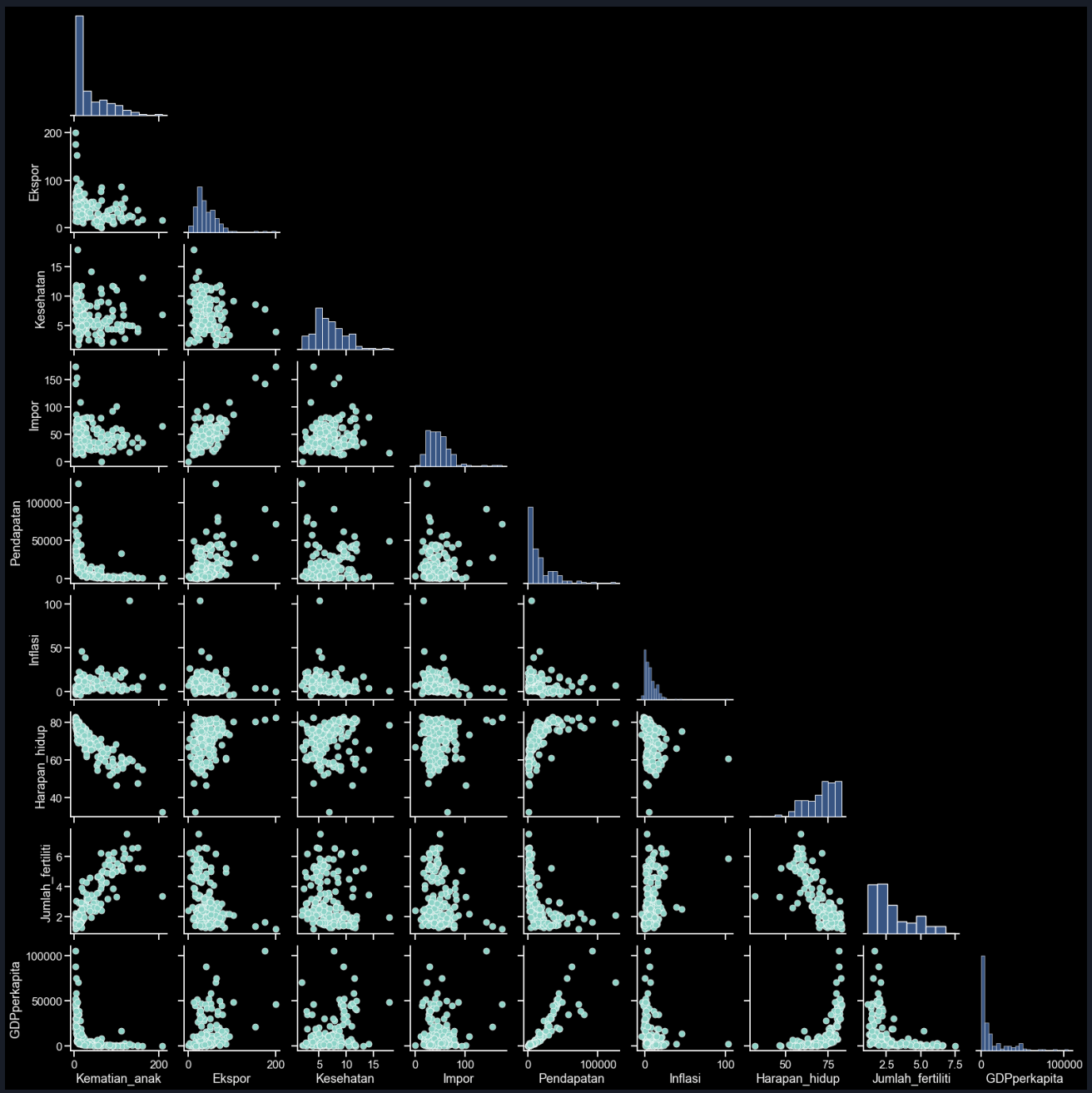


PART III

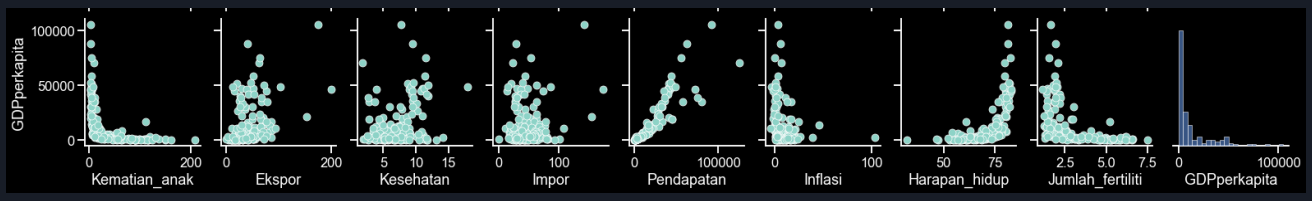
MACHINE LEARNING – K-Means Clustering

Pemilihan Variabel

Dari Sembilan variabel dalam dataset, dilakukan pair plot seperti berikut



Berdasarkan informasi dari UN General Assembly (GA) dan juga Economic and Social Council (ECOSOC), salah satu indikator yang menentukan Least Development Country adalah National Income (pendapatan nasional). Oleh karena itu, variabel pertama yang dipilih adalah GDP karena data GDP sudah menandakan pendapatan nasional dari negara tersebut.



Dari informasi grafik di atas, sepertinya jika GDP dipadukan dengan Pendapatan, akan kurang begitu bagus untuk dibuat cluster. Mengapa demikian? Hasil scatter plot menyatakan bahwa bentuk datanya menunjukkan adanya linearitas. Sehingga untuk Machine Learning, lebih tepat jika menggunakan Linear Regression untuk memprediksi rata-rata pendapatan orang jika diketahui variabel GDP nya atau sebaliknya.

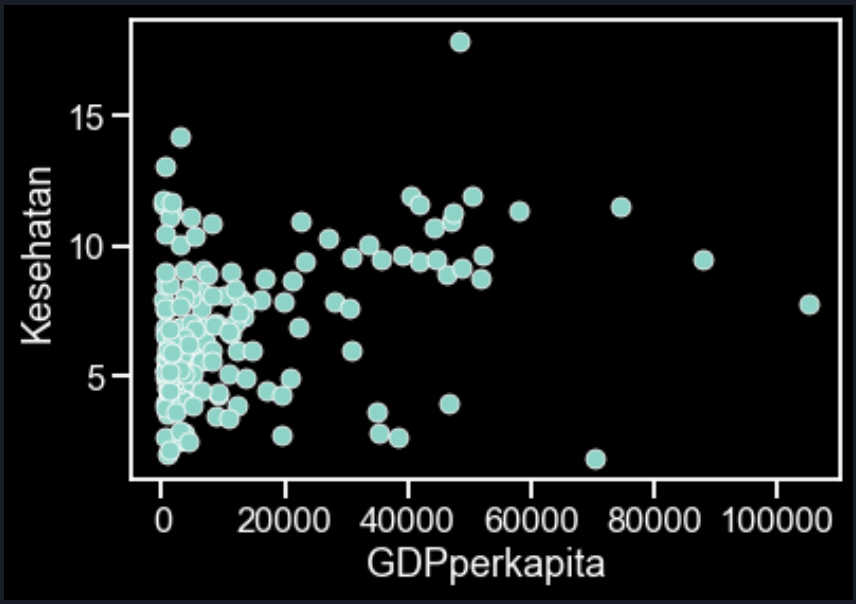
Selain itu, GDP dan pendapatan kurang begitu bagus dicari insightnya dalam konteks organisasi HELP. Hal ini dikarenakan jika suatu negara terkena bencana, dampak kepada nyawa lebih signifikan ketimbang dampak pendapatan. Pendapatan bisa pulih kembali seiring berjalannya waktu (seperti situasi COVID-19, meski banyak yang kehilangan pekerjaan, namun selalu ada cara alternatif dalam mendapatkan kembali pendapatan yang hilang seperti berjualan online/wfh/dll.)

Dampak yang paling signifikan adalah di bidang kesehatan. Negara terbelakang dilatarbelakangi oleh minimnya sumber daya manusia sehingga kesadaran pentingnya kesehatan masih minim. Hal seperti ini bahkan umum juga dijumpai negara berkembang. Oleh karena itu, negara yang masih minim investasi di bidang kesehatan bisa dikatakan kurang sadar akan pentingnya kesehatan. Ditambah dengan GDP nya yang rendah, hal ini bisa memperburuk keadaan ketika negara tersebut ditimpa bencana sewaktu-waktu, risiko kehilangan nyawa akan lebih besar ketimbang negara yang sadar akan pentingnya kesehatan.

Jadi, variabel yang dipilih adalah GDP dan Kesehatan.

GDP dan Kesehatan

Informasi lebih lengkap mengenai variabel GDP dan Kesehatan bisa ditemukan di bagian sebelumnya.



Dari scatter plot tersebut, terlihat bahwa GDP perkapita dan juga kesehatan memiliki outliers. Dengan menggunakan rumus batas bawah dan batas atas dari interquatil, diperoleh data sebagai berikut:

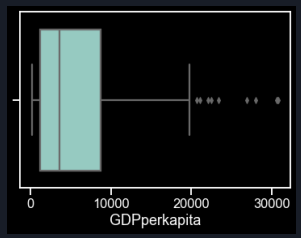
Variabel	Batas Bawah	Batas Atas	Outliers
GDPperkapita	-17750.0	33130.0	51900 46900 44400 35300 47400 58000 46200 40600 41800 41900 48700 35800 44500 38500 105000 50300 33700 87800 70300 46600 52100 74600 35000 38900 48400
Kesehatan	-0.6000000000000023	14.120000000000005	14.2 17.9

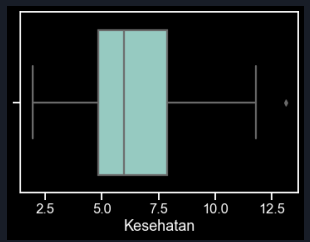
Tabel di atas menunjukan bahwa Outliers berada di atas batas atas (tidak ada yang di bawah batas bawah). Negara yang bisa dikecualikan adalah:

	Negara	GDPperkapita	Kesehatan
7	Australia	51900	8.73
8	Austria	46900	11.00
15	Belgium	44400	10.70
23	Brunei	35300	2.84
29	Canada	47400	11.30
44	Denmark	58000	11.40
53	Finland	46200	8.95
54	France	40600	11.90
58	Germany	41800	11.60
68	Iceland	41900	9.40
73	Ireland	48700	9.19
75	Italy	35800	9.53
77	Japan	44500	9.49
82	Kuwait	38500	2.63
91	Luxembourg	105000	7.77
101	Micronesia, Fed. Sts.	2860	14.20
110	Netherlands	50300	11.90
111	New Zealand	33700	10.10
114	Norway	87800	9.48
123	Qatar	70300	1.81
133	Singapore	46600	3.96
144	Sweden	52100	9.63
145	Switzerland	74600	11.50
157	United Arab Emirates	35000	3.66
158	United Kingdom	38900	9.64
159	United States	48400	17.90

Dalam kasus negara Micronesia, dia bisa dikecualikan karena meskipun GDP nya rendah, investasi akan kesehatannya cukup tinggi, sehingga kemungkinan besar, negara tersebut masyarakatnya sudah sadar akan kesehatan dan pemerintah setempat juga memprioritaskan pemulihan kesehatan jika ada bencana karena investasinya terhadap kesehatan yang cukup tinggi. Organisasi HELP lebih tepat memberikan bantuan yang bisa menaikkan stimulus kegiatan ekonomi negara tersebut untuk menaikkan GDP nya.

Dari seleksi outliers, dataset ini menghasilkan 141 negara (26 negara merupakan outliers). Untuk mengecek bahwa outliers benar-benar hilang, dilakukan boxplot.



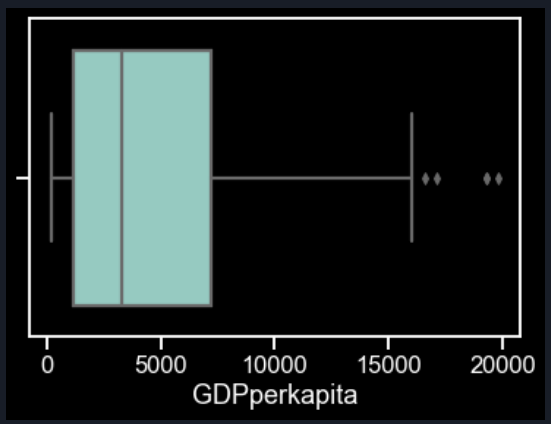


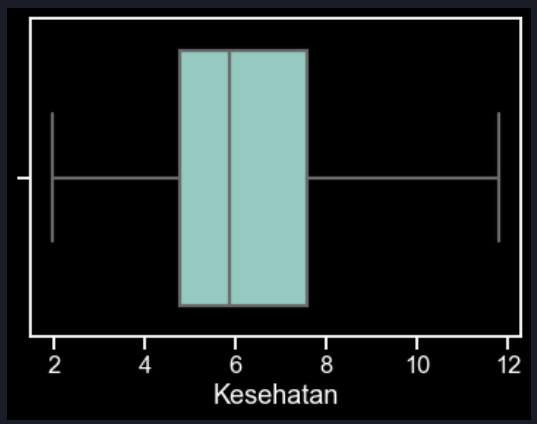
Dari boxplot tersebut, masih ditemukan outliers, hal ini tidak bisa dibiarkan karena metode Machine Learning ini sensitif terhadap outliers, maka data ini akan dihitung kembali untuk mencari outliers berikutnya. Berikut data outliers baru, dan negara yang bisa dikecualikan.

Variabel	Batas Bawah	Batas Atas	Outliers
GDPperkapita	-10125.0	20075.0	28000 20700 30800 26900 30600 21100 22500 23400 22100 30700
Kesehatan	0.33000000000000096	12.41	13.1

	Negara	GDPperkapita	Kesehatan
10	Bahamas	28000	7.89
11	Bahrain	20700	4.97
42	Cyprus	30800	5.97
60	Greece	26900	10.30
74	Israel	30600	7.63
98	Malta	21100	8.65
122	Portugal	22500	11.00
132	Sierra Leone	399	13.10
135	Slovenia	23400	9.41
138	South Korea	22100	6.93
139	Spain	30700	9.54

Perhitungan ini menyisakan 130 negara (11 negara merupakan outliers). Selanjutnya adalah boxplot untuk melihat apakah masih ada outliers



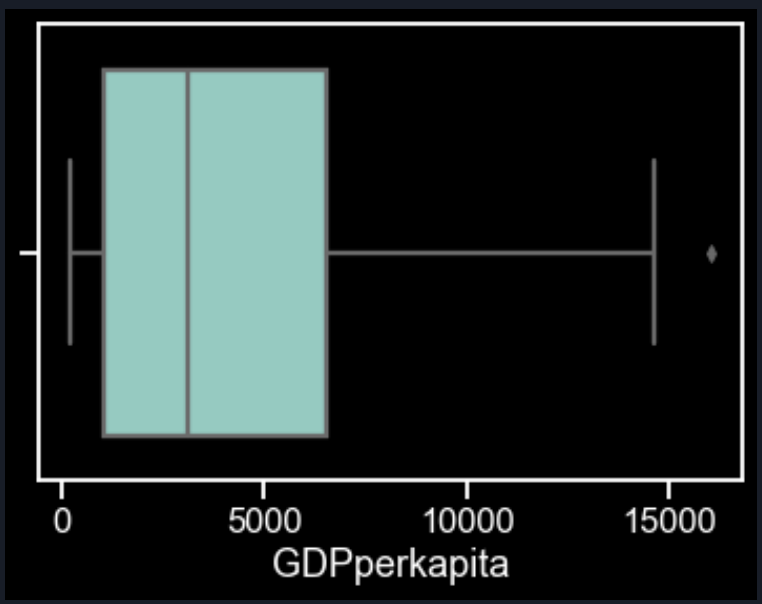


Boxplot menunjukan bahwa outliers sudah hilang di bagian kesehatan, namun masih ada di GDP. Oleh karena itu, GDP kembali dihitung untuk membuang outliers. Berikut data outliers baru, dan negara yang bisa dikecualikan.

Variabel	Batas Bawah	Batas Atas	Outliers
GDPperkapita	-7983.75	16366.25	19800 17100 19300 19300 16600

	Negara	GDPperkapita	Kesehatan
43	Czech Republic	19800	7.88
49	Equatorial Guinea	17100	4.48
115	Oman	19300	2.77
128	Saudi Arabia	19300	4.29
134	Slovak Republic	16600	8.79

Perhitungan ini menyisakan 125 negara (5 negara merupakan outliers). Dalam kasus dimana negara memiliki GDP yang tinggi, namun investasi kesehatannya rendah, maka yang dapat dilakukan HELP adalah memberikan edukasi untuk memberikan kesadaran pentingnya kesehatan ketimbang memberikan bantuan. Selanjutnya adalah boxplot untuk melihat apakah masih ada outliers.

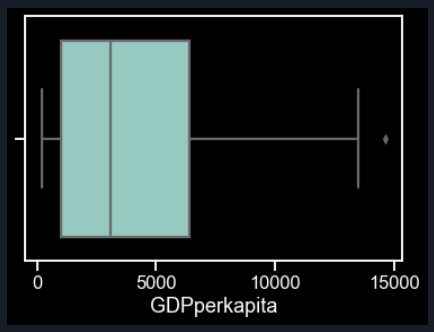


Boxplot menunjukan bahwa outliers sudah hilang di bagian kesehatan, namun masih ada di GDP. Oleh karena itu, GDP kembali dihitung untuk membuang outliers. Berikut data outliers baru, dan negara yang bisa dikecualikan.

Variabel	Batas Bawah	Batas Atas	Outliers
GDPperkapita	-7195.0	14765.0	16000

	Negara	GDPperkapita	Kesehatan
13	Barbados	16000	7.97

Perhitungan ini menyisakan 124 negara (1 negara merupakan outliers). Boxplot GDP adalah sebagai berikut.

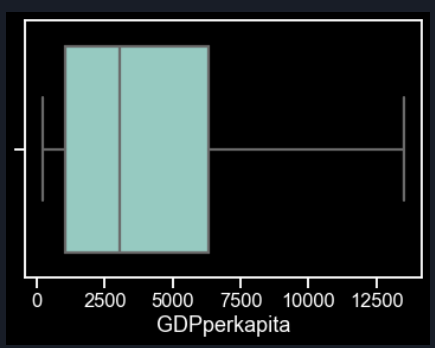


Boxplot menunjukan bahwa outliers sudah hilang di bagian kesehatan, namun masih ada di GDP. Oleh karena itu, GDP kembali dihitung untuk membuang outliers. Berikut data outliers baru, dan negara yang bisa dikecualikan.

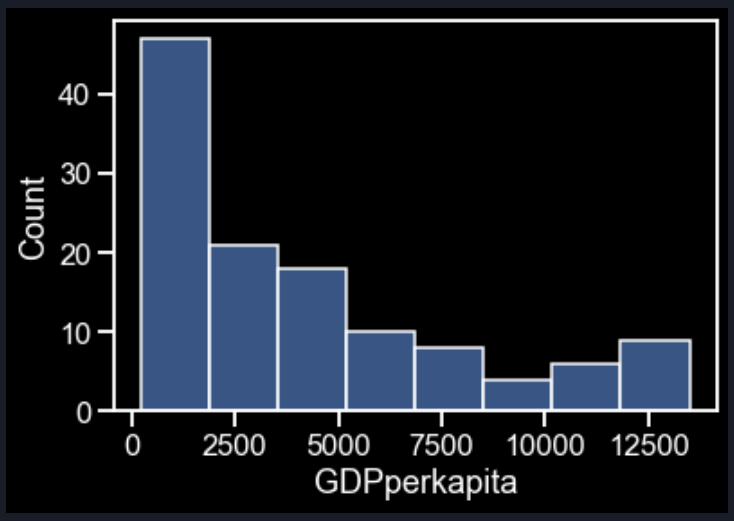
Variabel	Batas Bawah	Batas Atas	Outliers
GDPperkapita	-7017.5	14442.5	14600

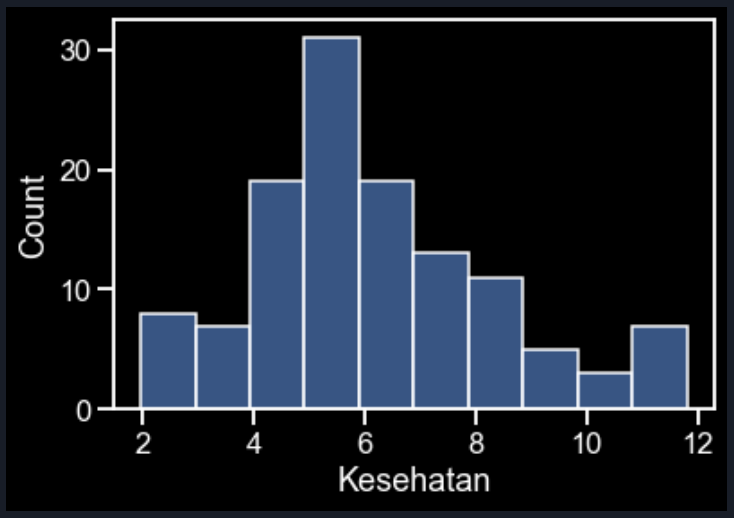
	Negara	GDPperkapita	Kesehatan
51	Estonia	14600	6.03

Perhitungan ini menyisakan 123 negara (1 negara merupakan outliers). Boxplot GDP adalah sebagai berikut.



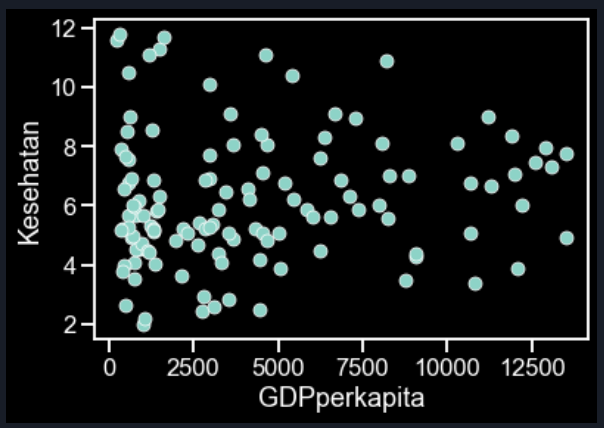
Dari grafik tersebut, outlier telah menghilang, dan berikut histogram GDP dan juga Kesehatan yang sudah dihilangkan outliernya.





Sekarang GDP memiliki rata-rata 4208.577236, dan interval terbanyak ada di 0-2000 dengan jumlah negara lebih dari 40. Sementara di bidang kesehatan rata-ratanya 6.182195, dan interval terbanyak ada di 5-6 dengan jumlah negara lebih dari 30.

Berikut adalah scatterplot dari GDP dengan Kesehatan tanpa adanya outliers.



Scaling

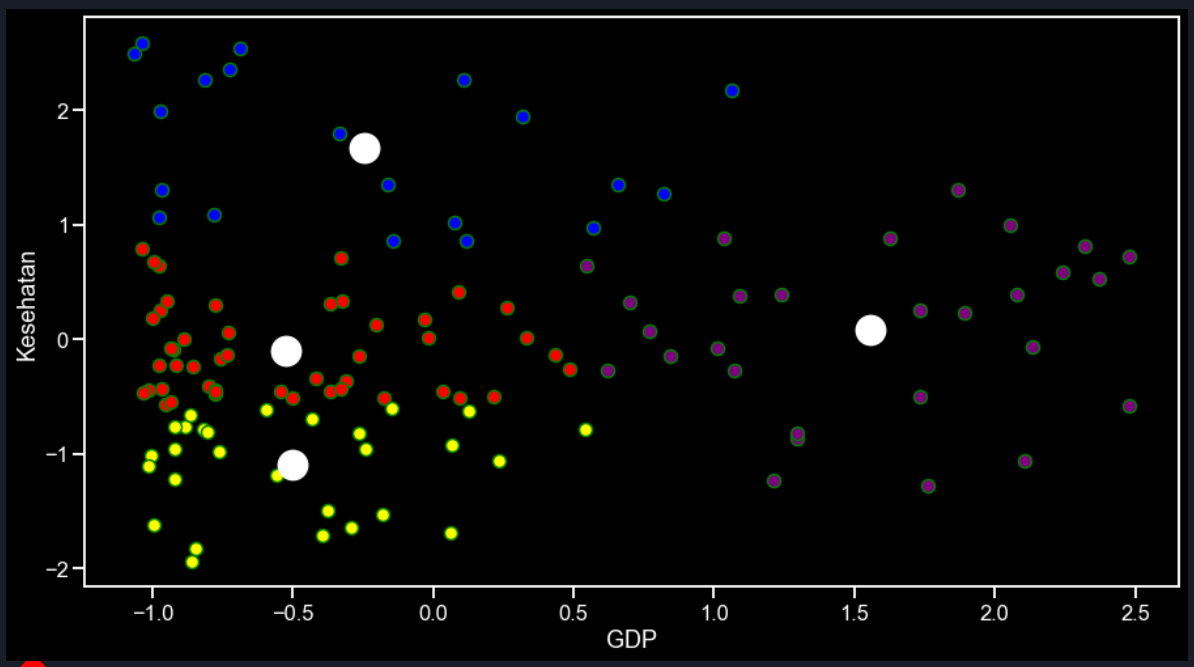
Karena rentang nilai GDP dan Kesehatan berbeda, maka nilai masing-masing fitur harus discaling agar sama, Dengan menggunakan standard scaler dari sklearn, GDP dan Kesehatan akan berubah rentang nilainya.

K-Means

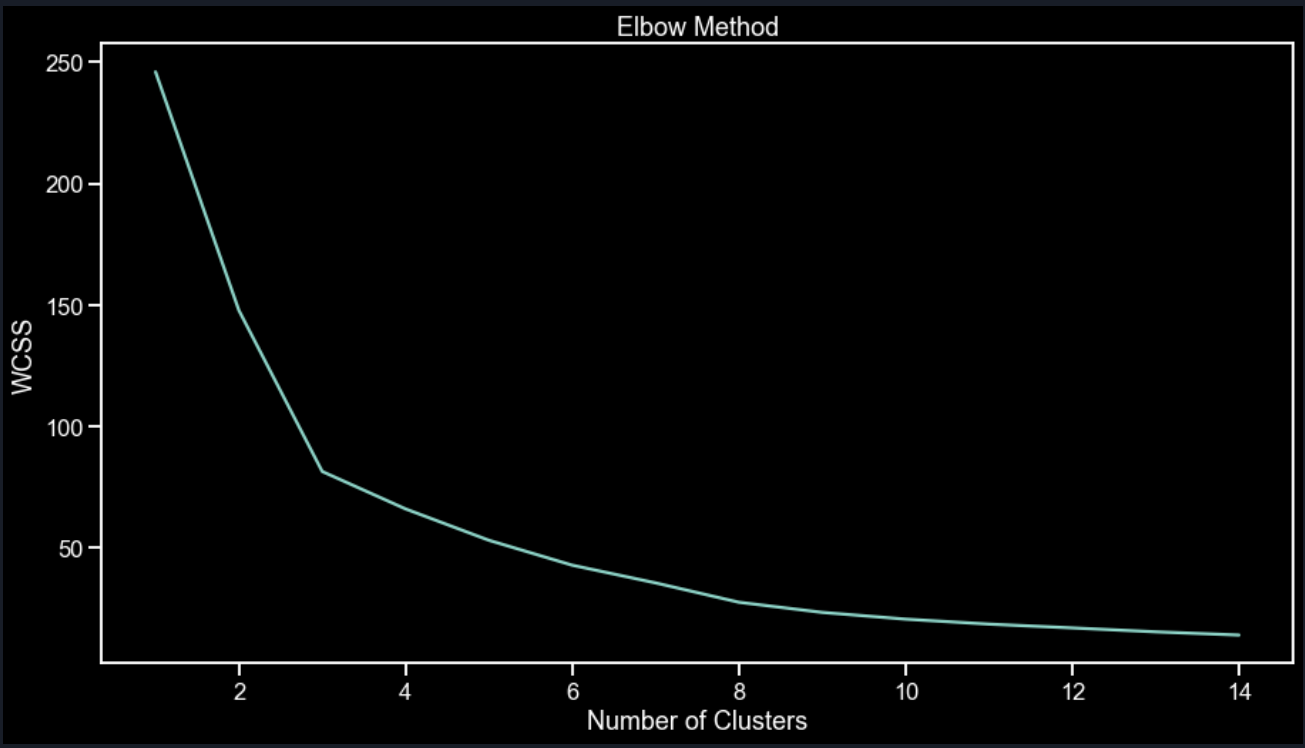
Sebagai pembukaan, saya menginginkan adanya 4 cluster dengan alasan:

- Kluster pertama untuk negara GDP rendah, Kesehatan rendah (yang akan saya masukan sebagai prioritas utama HELP untuk memberikan bantuan baik dana maupun bantuan kemanusiaan lainnya.)
- Kluster kedua untuk negara GDP rendah, Kesehatan tinggi (yang akan saya rekomendasikan pada HELP untuk memberikan bantuan berupa stimulus untuk menaikkan kegiatan ekonomi)
- Kluster ketiga untuk negara GDP tinggi, Kesehatan rendah (yang akan saya rekomendasikan pada HELP untuk memberikan edukasi mengenai kesadaran akan kesehatan)
- Kluster keempat untuk negara GDP tinggi, Kesehatan tinggi (yang tidak saya jadikan prioritas HELP)

Hasil K-Means menunjukan sebagai berikut

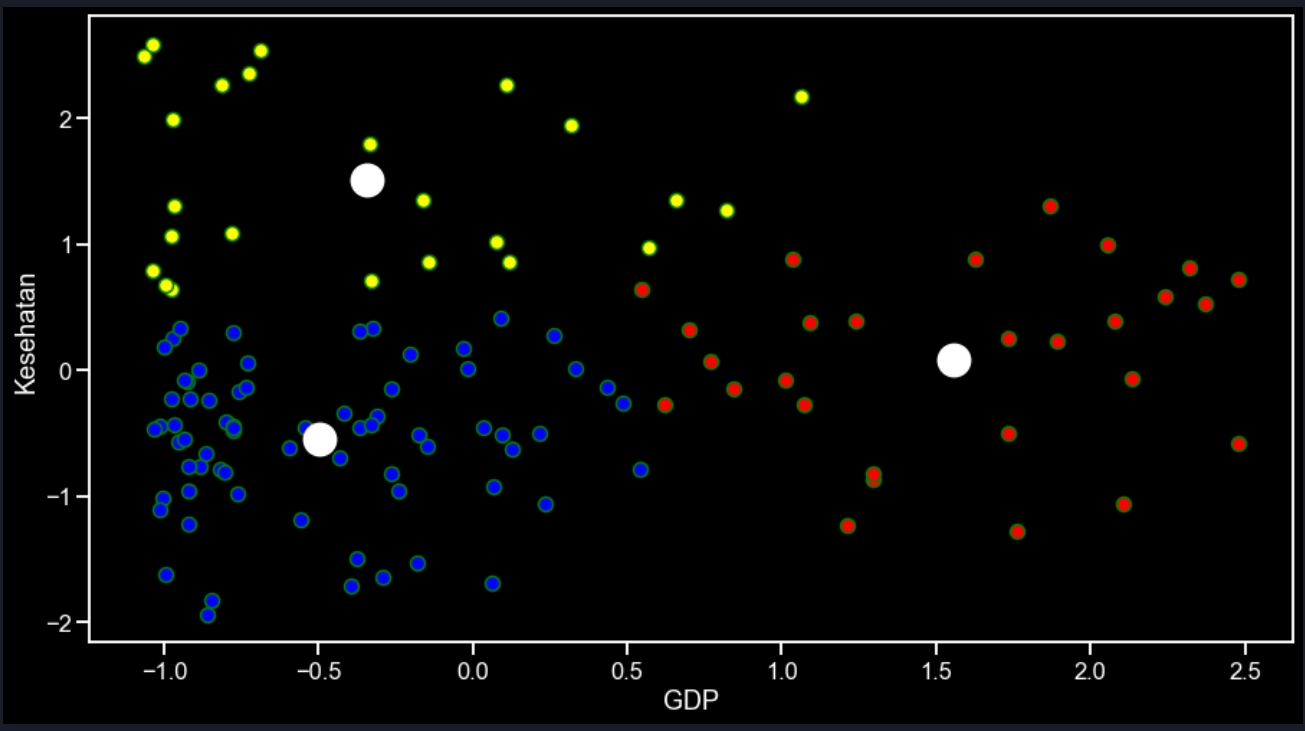


Namun, saya harus memperhitungkan jumlah kluster dengan hasil dari Elbow Method, yaitu sebagai berikut.



Menurut grafik Elbow Method, Jumlah cluster yang disarankan adalah 3 karena penurunan 3 ke 4 sudah mulai melandai.

Dibuat kluster ulang untuk memenuhi grafik Elbow Method, berikut adalah hasilnya.



Analisis

Menurut Machine Learning K-Means, data ini dapat dikelompokan menjadi 3 cluster, yaitu

- Kluster pertama (GDP tinggi / di atas rata-rata, dan Kesehatan bervariasi)
- Kluster kedua (GDP rendah / di bawah rata-rata, dan Kesehatan rendah / di bawah rata-rata)
- Kluster ketiga (GDP rendah / di bawah rata-rata, dan Kesehatan tinggi / di atas rata-rata)

Hasil Silhouette Score untuk 4 cluster adalah 0.36247252298806687, sementara skor untuk 3 cluster adalah 0.5532239930756935. Yang berarti, Data dengan tiga cluster memberikan insight yang lebih baik dalam segmentasi negara. Berikutnya akan dibahas negara mana saja yang harus menjadi prioritas HELP.

PART IV

Prioritized Countries for HELP

Tindakan HELP untuk masing-masing Cluster

Menurut hasil Kmeans sebelumnya, direkomendasikan tindakan untuk masing-masing kluster.

Kluster	GDP	Kesehatan	Action
1	High	Low to High	Bukan menjadi prioritas HELP untuk membantu negara-negara ini jika ada bencana.
2	Low	Low	Menjadi prioritas utama HELP untuk membantu negara-negara yang ada di kluster ini dengan jumlah yang se signifikan mungkin
3	Low	High	Menjadi prioritas kedua HELP untuk membantu negara-negara yang ada di kluster ini, namun lebih banyak ke bantuan yang memberikan stimulus untuk menggerakkan roda perekonomian

Kluster Prioritas (Kluster 2)

Kluster ini memiliki 71 negara sebagai berikut

	Negara	GDPperkapita	health
1	Albania	4090	6,55
2	Algeria	4460	4,17
3	Angola	3530	2,85
4	Armenia	3220	4,4
5	Azerbaijan	5840	5,88
6	Bangladesh	758	3,52
7	Belarus	6030	5,61
8	Belize	4340	5,2
9	Benin	758	4,1
10	Bhutan	2180	5,2
11	Bolivia	1980	4,84

12	Burkina Faso	575	6,74
13	Cambodia	786	5,68
14	Cameroon	1310	5,13
15	Cape Verde	3310	4,09
16	Central African Republic	446	3,98
17	Chad	897	4,53
18	China	4560	5,07
19	Comoros	769	4,51
20	Congo, Rep.	2740	2,46
21	Cote d'Ivoire	1220	5,3
22	Dominican Republic	5450	6,22
23	Egypt	2600	4,66
24	El Salvador	2990	6,91
25	Eritrea	482	2,66
26	Fiji	3650	4,86
27	Gambia	562	5,69
28	Ghana	1310	5,22
29	Guatemala	2830	6,85
30	Guinea	648	4,93
31	Guyana	3040	5,38
32	Haiti	662	6,91
33	India	1350	4,05
34	Indonesia	3110	2,61
35	Jamaica	4680	4,81
36	Kenya	967	4,75
37	Kyrgyz Republic	880	6,18
38	Lao	1140	4,47
39	Macedonia, FYR	4540	7,09
40	Madagascar	413	3,77
41	Malawi	459	6,59
42	Mali	708	4,98
43	Mauritania	1200	4,41
44	Mongolia	2650	5,44
45	Morocco	2830	5,2
46	Mozambique	419	5,21
47	Myanmar	988	1,97
48	Namibia	5190	6,78
49	Nepal	592	5,25
50	Niger	348	5,16
51	Nigeria	2330	5,07
52	Pakistan	1040	2,2
53	Paraguay	3230	5,87
54	Peru	5020	5,08
55	Philippines	2130	3,61
56	Samoa	3450	6,47

57	Senegal	1000	5,66
58	Sri Lanka	2810	2,94
59	St. Vincent and the Grenadines	6230	4,47
60	Sudan	1480	6,32
61	Tajikistan	738	5,98
62	Tanzania	702	6,01
63	Thailand	5080	3,88
64	Tonga	3550	5,07
65	Tunisia	4140	6,21
66	Turkmenistan	4440	2,5
67	Uzbekistan	1380	5,81
68	Vanuatu	2970	5,25
69	Vietnam	1310	6,84
70	Yemen	1310	5,18
71	Zambia	1460	5,89

Setelah mengurutkan berdasarkan GDP terendahnya, inilah sepuluh negara utama yang menjadi prioritas HELP memberikan bantuan.

	Negara	GDPperkapita	health
0	Niger	348	5.16
1	Madagascar	413	3.77
2	Mozambique	419	5.21
3	Central African Republic	446	3.98
4	Malawi	459	6.59
5	Eritrea	482	2.66
6	Gambia	562	5.69
7	Burkina Faso	575	6.74
8	Nepal	592	5.25
9	Guinea	648	4.93

Berdasarkan pengeluaran kesehatannya, HELP harus memberikan edukasi pentingnya kesehatan.

	Negara	GDPperkapita	health
0	Myanmar	988	1.97
1	Pakistan	1040	2.20
2	Congo, Rep.	2740	2.46
3	Turkmenistan	4440	2.50
4	Indonesia	3110	2.61
5	Eritrea	482	2.66
6	Angola	3530	2.85
7	Sri Lanka	2810	2.94
8	Bangladesh	758	3.52
9	Philippines	2130	3.61

Jika diperhatikan dengan seksama, Eritrea merupakan negara irisan yang ada dalam GDP terendah dan juga kesehatan rendah. Maka HELP wajib memberikan perhatian khusus pada negara Eritrea karena GDP nya termasuk dalam negara terendah, begitu juga dengan Kesehatannya. Direkomendasikan kepada HELP untuk tidak hanya memberikan bantuan pada saat bencana untuk negara Eritrea, namun sebisa mungkin Eritrea memberikan edukasi Kesehatan secara terus-menerus dan juga memberi bantuan untuk menggerakan ekonomi ke negara ini. Tidak menutup kemungkinan HELP juga harus memberikan edukasi di bidang-bidang lain bagi negara Eritrea untuk memerangi kemiskinan di negara tersebut dan juga meningkatkan SDMnya.