CLUSTERING THE COUNTRIES BY USING K-MEANS FOR HELP INTERNATIONAL

HELMY NAUFAL AZIZ

OVERVIEW

- Profil Organisasi:

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.

- Tujuan:

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

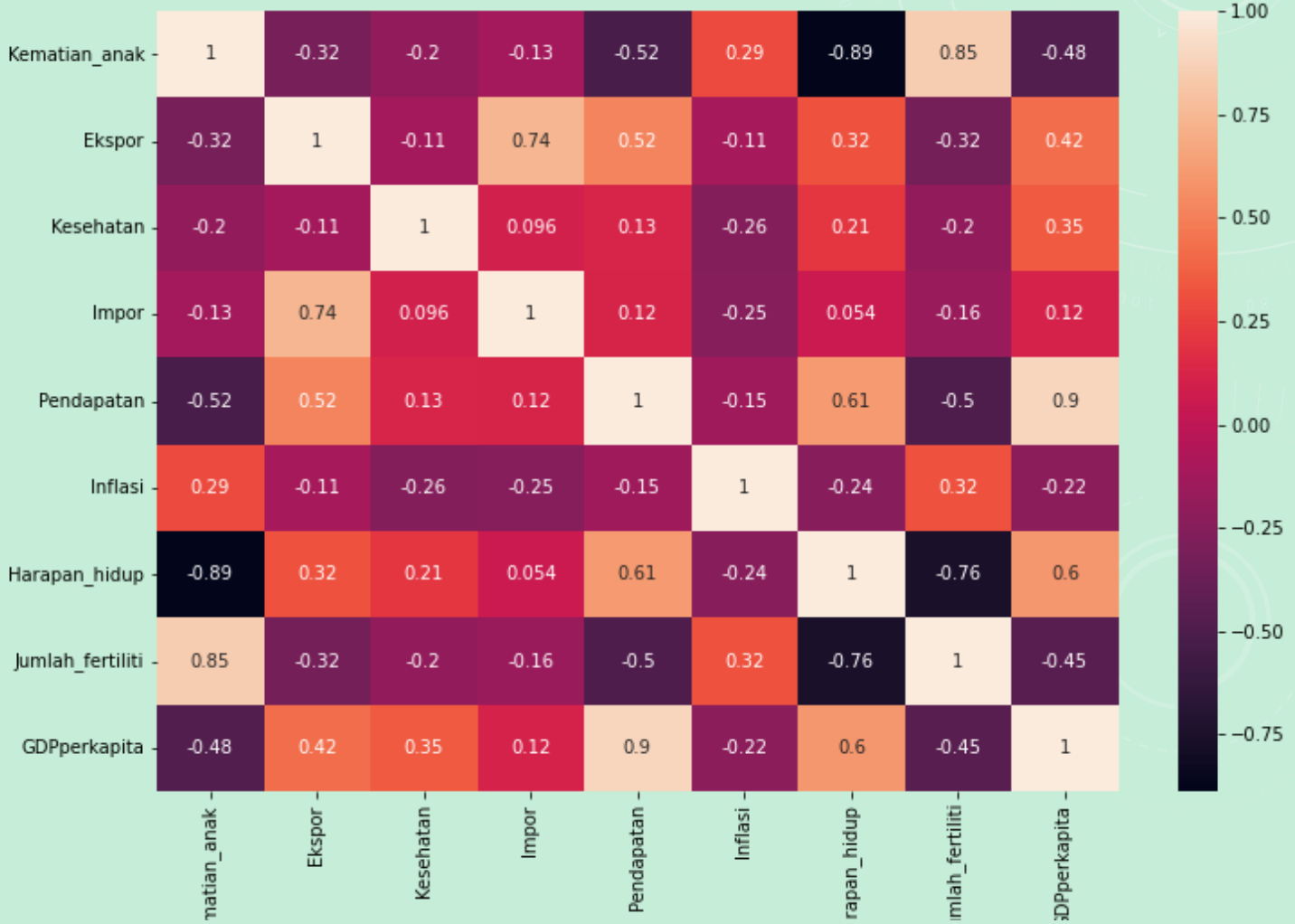
Data pendukung yang digunakan dalam project ini memiliki 167 baris dan 10 kolom dengan indikator sebagai berikut:

	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

MULTIVARIATE ANALISIS

- Untuk mengetahui korelasi dari setiap indikator yang dimiliki dapat dilakukan multivariate analisis dan diperoleh data seperti di samping.
- Dari data di samping dapat dilihat bahwa indikator GDPperkapita dan Pendapatan memiliki nilai koefisien positif tertinggi yakni 0.9. Kemudian Kematian_anak dan Harapan hidup memiliki nilai koefisien negatif tertinggi yakni –0.89. Selain itu, Kematian_anak juga memiliki koefisien negative yang tinggi terhadap GDPperkapita yang bisa dikatakan sebagai indikator kesejahteraan suatu negara.
- Sehingga indikator Pendapatan dapat mewakili indikator dari aspek Sosial Ekonomi serta indikator Kematian_anak mewakili aspek Kesehatan.



DATA CLEANING

1. Missing Value

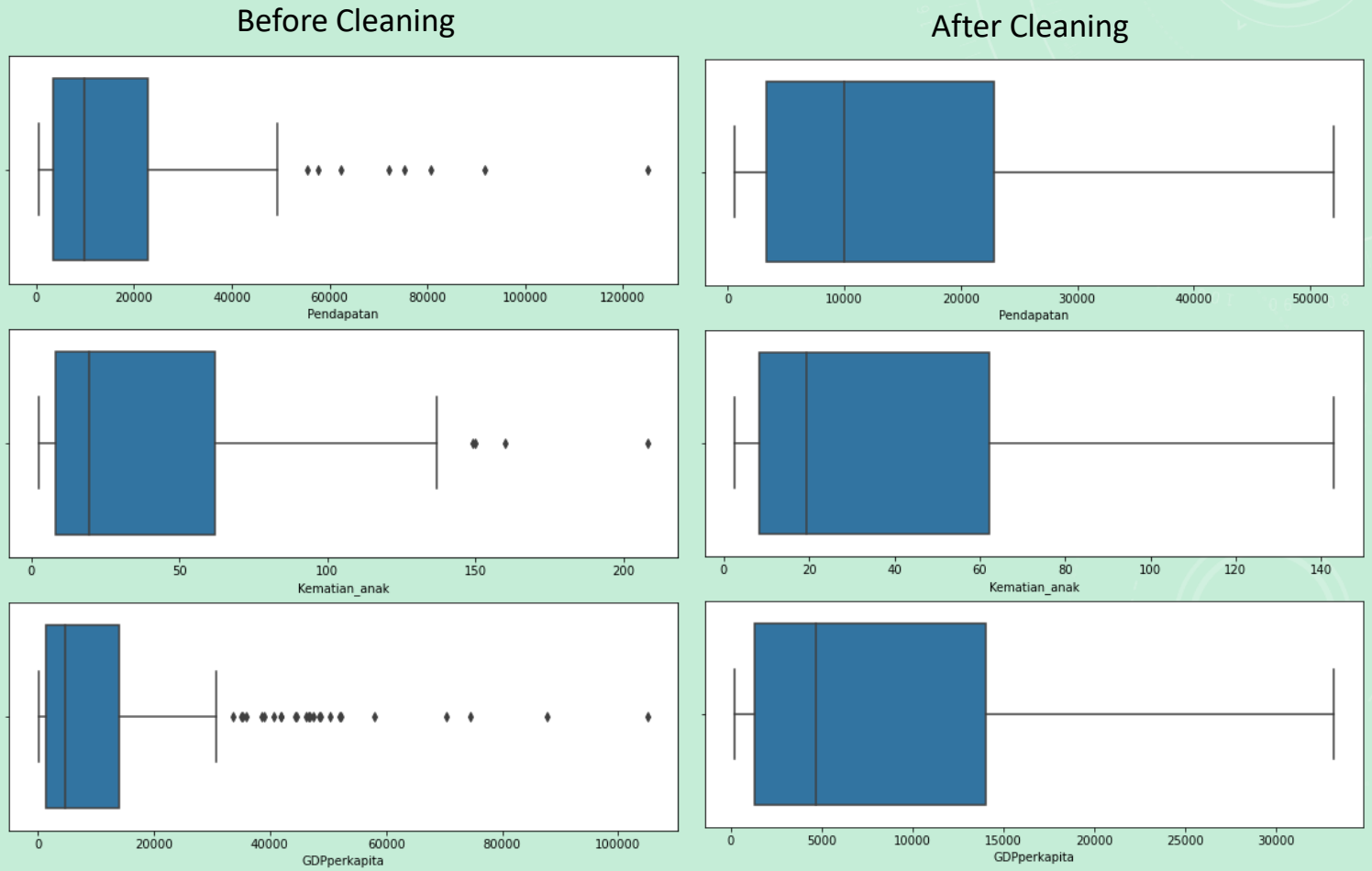
- Untuk menghindari missing value dapat diketahui dengan mengecek informasi data yang kita punya menggunakan method .info()
- Dapat dilihat bahwa range index yang dimiliki sebanyak 167 dan dari setiap kolom juga memiliki jumlah Non-null kolom sebanyak 167. Sehingga dapat disimpulkan data ini tidak memiliki missing value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Negara          167 non-null    object
 1   Pendapatan      167 non-null    int64
 2   Kematian_anak   167 non-null    float64
 3   GDPperkapita    167 non-null    int64
dtypes: float64(1), int64(2), object(1)
memory usage: 5.3+ KB
```

DATA CLEANING

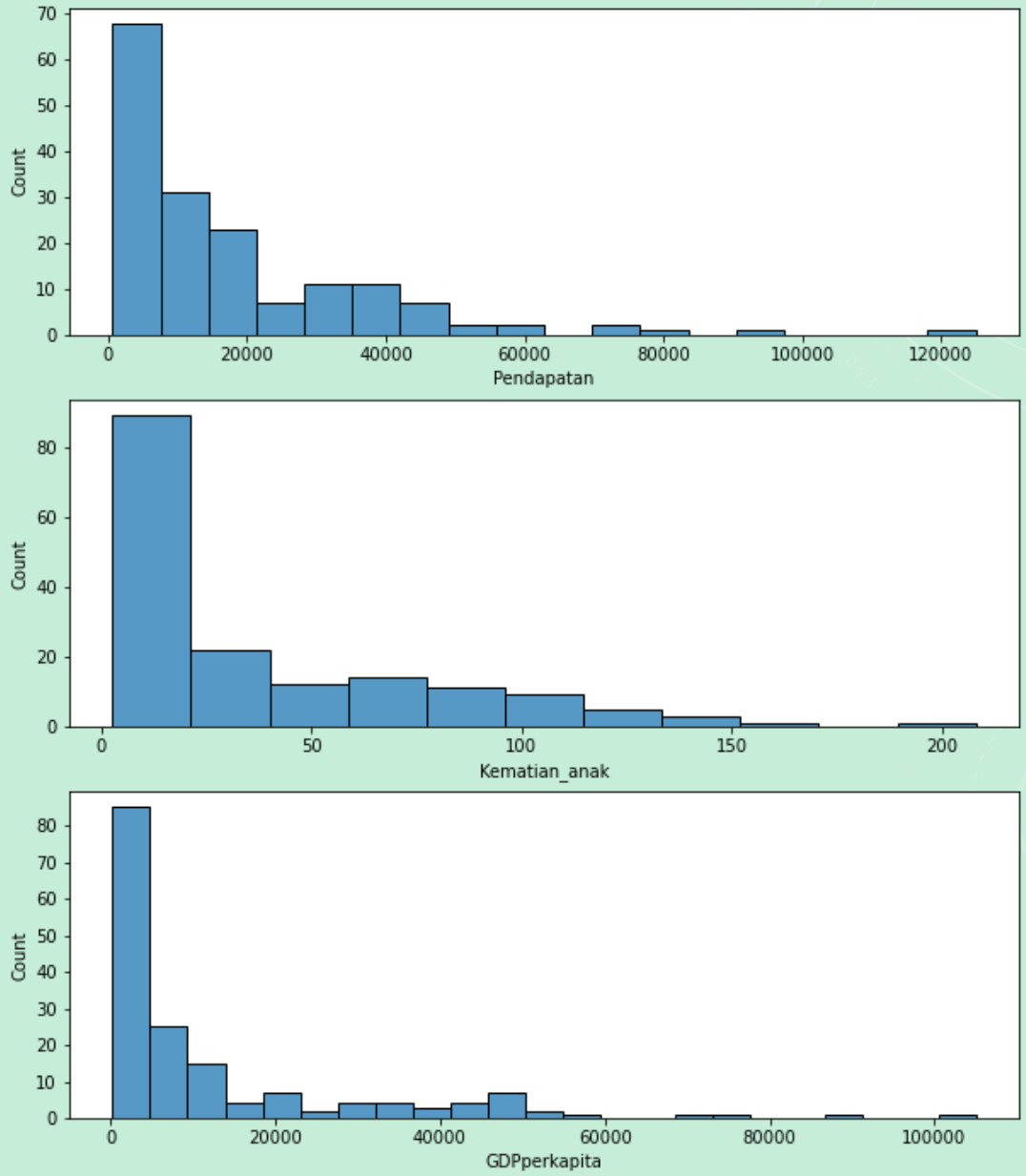
2. Outliers

- Untuk menghilangkan outlier dapat mengganti data outlier dengan batas bawah atau batas atas sehingga tidak mempengaruhi urutan sebaran data secara keseluruhan.
- Dari gambar di samping dapat dilihat bahwa sebelum dilakukan cleaning masih terdapat beberapa data pencilan dan setelah cleaning data tersebut tidak muncul.



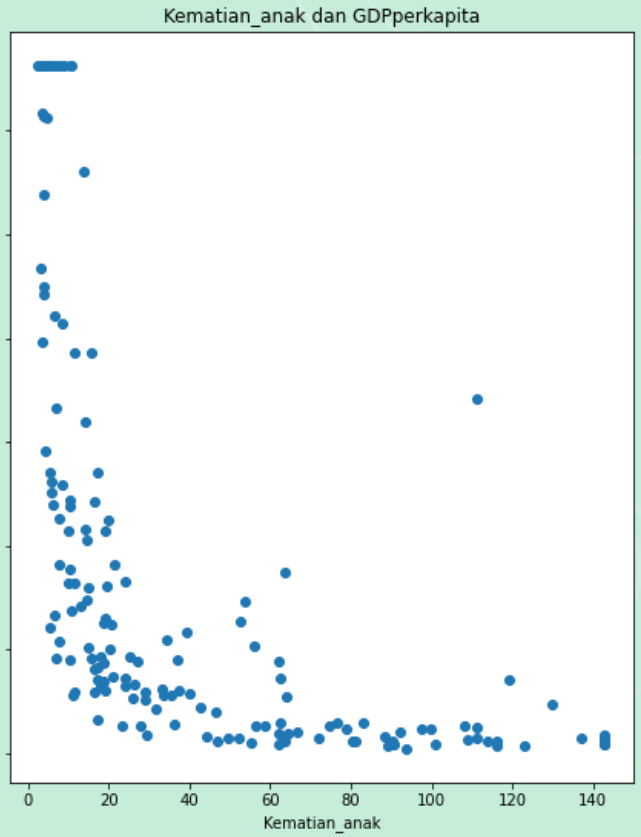
UNIVARIATE ANALYSIS

- Untuk melakukan univariate analysis dapat dilakukan menggunakan histogram untuk mengetahui sebaran data dari suatu variabel.
- Dari gambar di samping dapat dilihat negara terdapat banyak negara yang memiliki pendapatan dan GDP yang rendah sehingga bisa dikatakan bahwa bantuan yang akan diberikan akan sangat berguna bagi negara tersebut.



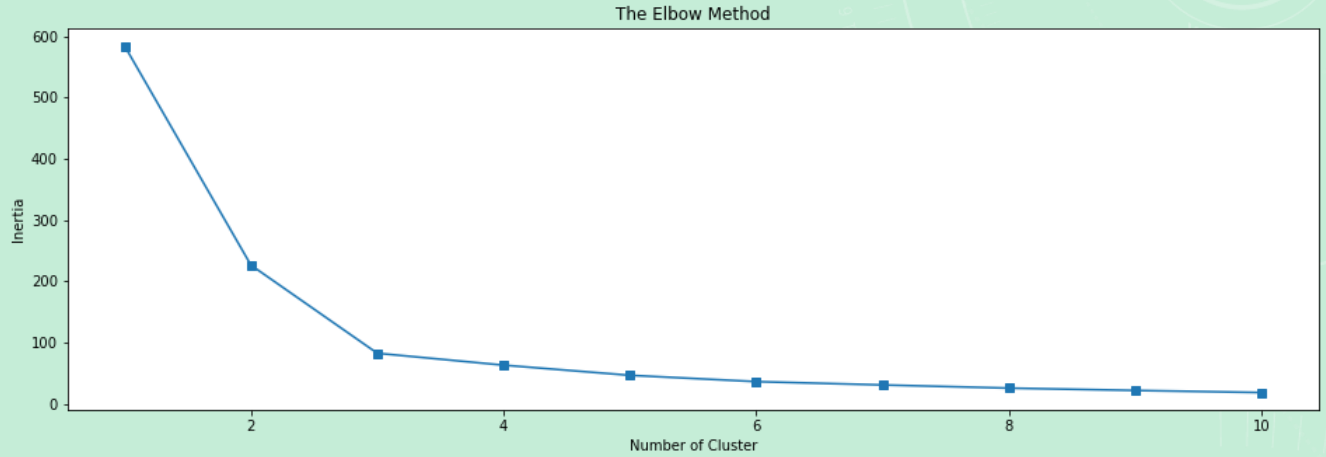
BIVARIATE ANALYSIS

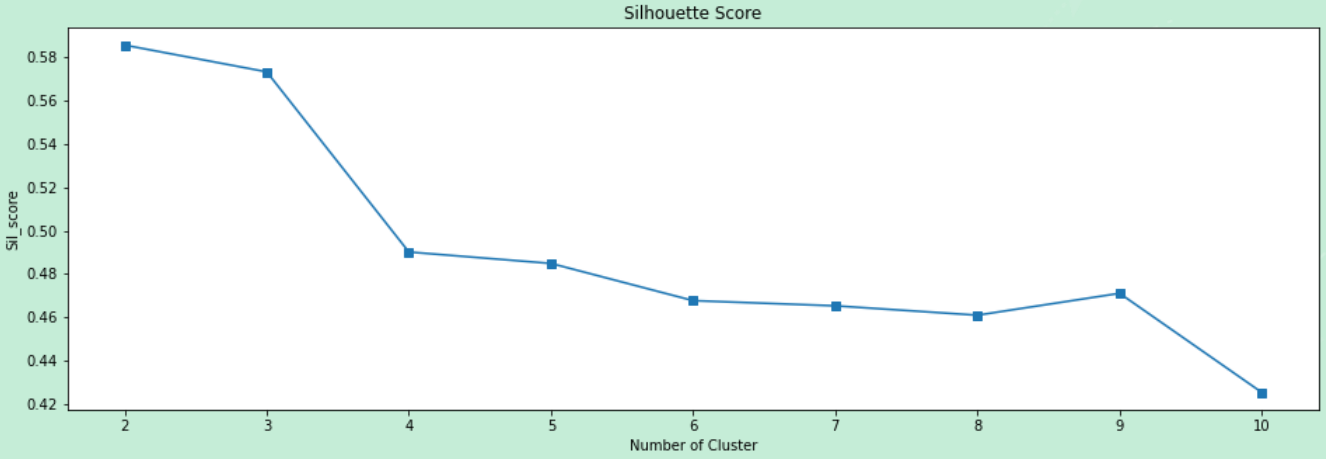
- Bivariate analisis dapat dilakukan menggunakan scatter plot untuk mengetahui keterkaitan anatara dua variable.
- Dari data di samping dapat dilihat bahwa Pendapatan dan GDP memiliki data yang linear dimana negara dengan pendapatan tinggi maka GDP akan tinggi
- Sebaliknya pada Kematian anak dan GDP memilki data yang terbalik dimana negara dengan kematian anak rendah akan memilki GDP yang tinggi



CLUSTERING DATA

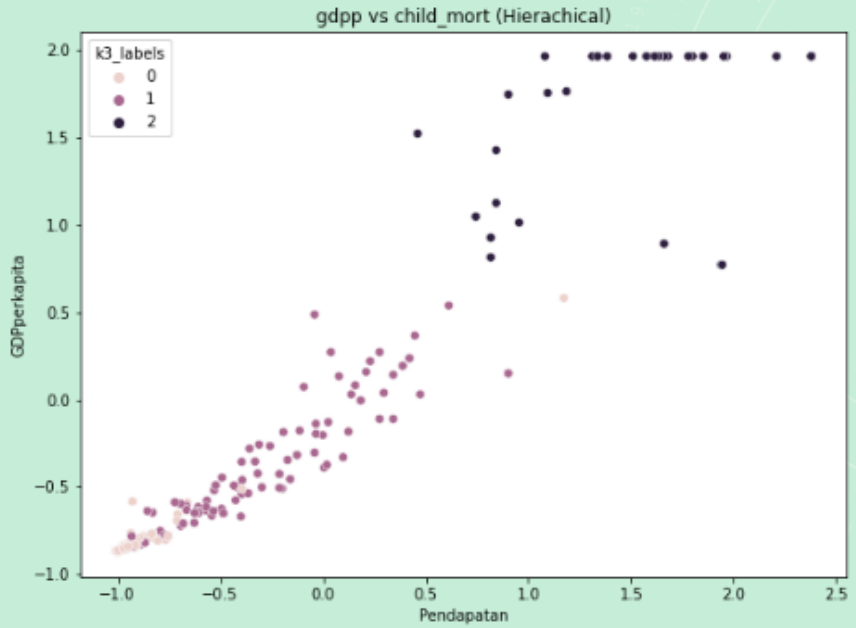
- Sebelum melakukan clustering data, kita perlu mengetahui berapa banyak cluster yang perlu kita buat. Jumlah ini dapat diketahui dengan menggunakan Elbow Method atau Silhouette Score.
- Dari data di samping dapat dilihat dengan menggunakan Elbow Method didapatkan nilai optimalnya yaitu 3. Kemudian pada Silhouette Score nilai tertinggi adalah 2 dan 3.
- Dapat disimpulkan dari 2 metode ini nilai 3 dapat menjadi jumlah cluster yang akan dibuat.

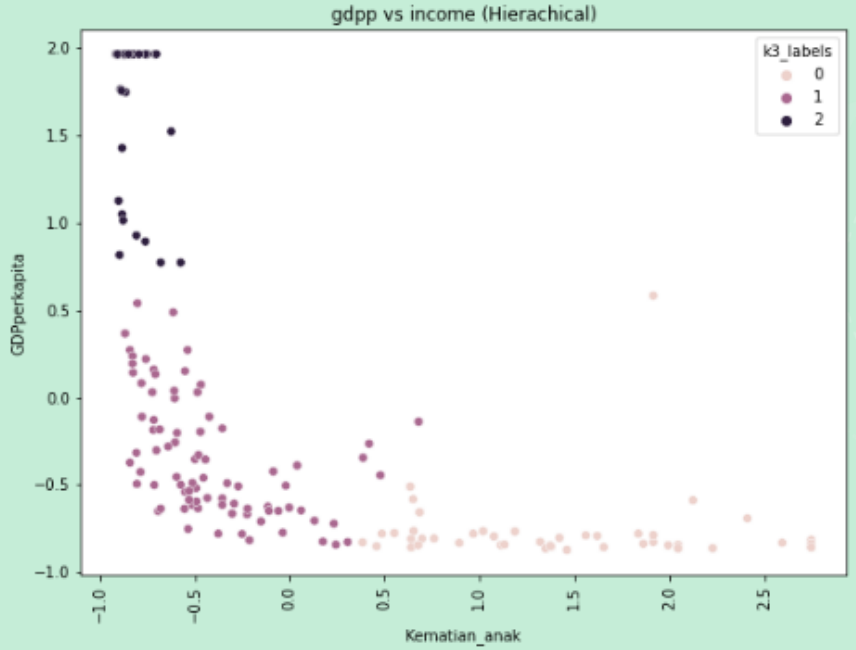




CLUSTERING DATA

- Setelah melakukan clustering data dengan jumlah cluster 3, data tersebut dapat ditampilkan menggunakan scatter plot dengan membagi 3 warna untuk mengetahui label dari cluster yang dibuat.
- Dari gambar di samping dapat dilihat bahwa cluster 0 merupakan negara dengan pendapatan yang rendah dan kematian anak yang tinggi. Sehingga cluster ini dapat menjadi acuan untuk memilih negara yang akan menerima bantuan yang akan diberikan.





RECOMMENDATION

- Setelah dilakukan clustering data dan diketahui bahwa cluster 0 merupakan negara yang berhak menerima bantuan, didapatkan bahwa di dalam terdapat 46 negara yang terdapat pada cluster 0.
- Apabila jumlah uang yang akan di donasikan dibagikan kepada 46 negara maka nilai tersebut akan terlalu kecil, mengingat negara pada cluster 0 memiliki pendapatan dan GDP yang rendah. Sehingga sangat membutuhkan bantuan ekonomi.
- Oleh karenanya akan dilakukan sorting data dan diberikan kepada 10 negara teratas dari sorting data tersebut.

	Negara	Pendapatan	Kematian_anak	GDPperkapita	cluster
0	Afghanistan	1610.0	90.200	553	0
3	Angola	5900.0	119.000	3530	0
17	Benin	1820.0	111.000	758	0
25	Burkina Faso	1430.0	116.000	575	0
26	Burundi	764.0	93.600	231	0
28	Cameroon	2660.0	108.000	1310	0
31	Central African Republic	888.0	142.875	446	0
32	Chad	1930.0	142.875	897	0
36	Comoros	1410.0	88.200	769	0

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 46 entries, 0 to 166
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Negara          46 non-null     object
 1   Pendapatan      46 non-null     float64
 2   Kematian_anak   46 non-null     float64
 3   GDPperkapita    46 non-null     int64
 4   cluster         46 non-null     int32
dtypes: float64(2), int32(1), int64(1), object(1)
memory usage: 2.0+ KB
```

RECOMMENDATION

Setelah dilakukan sorting data berdasarkan GDP per kapita, Pendapatan dan Kematian anak diperoleh 10 daftar negara sebagai berikut:

	index	Negara	cluster
0	26	Burundi	0
1	88	Liberia	0
2	37	Congo, Dem. Rep.	0
3	112	Niger	0
4	132	Sierra Leone	0
5	93	Madagascar	0
6	106	Mozambique	0
7	31	Central African Republic	0
8	94	Malawi	0
9	50	Eritrea	0