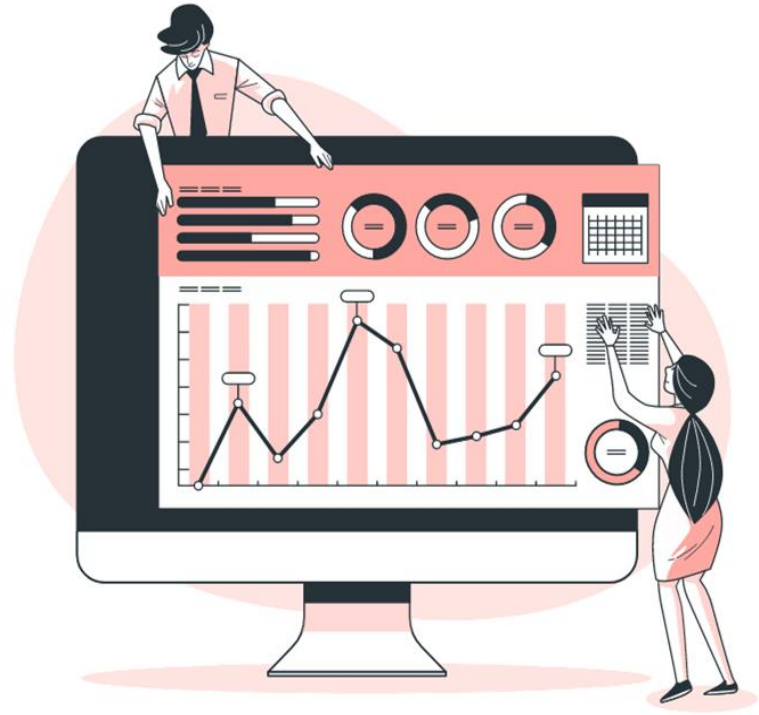


Clustering the Countries by using K-Means for HELP International

Python Data Science Final Project - Batch 29

-

Muhammad Alif Faddy Respatyadi
aliffaddly680@gmail.com

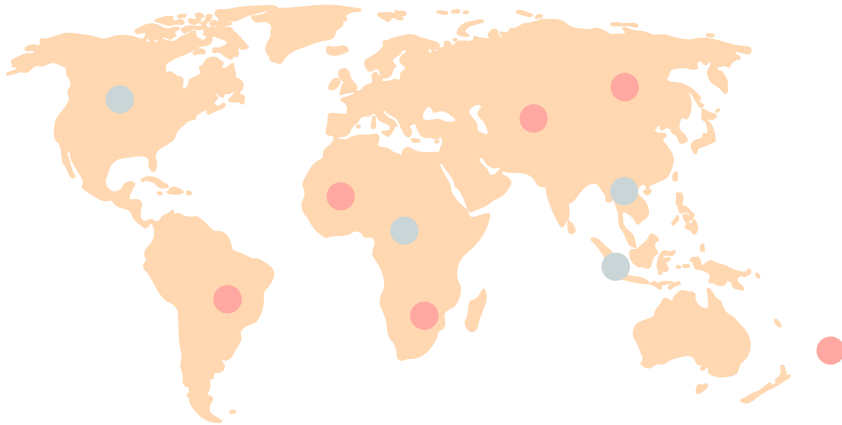


Objective



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

Tentang 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



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.

Segi Ekonomi



01



Reading and Understanding Data

Dataset

	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

Dataset terdiri dari 167 negara dengan faktor sosial, ekonomi, dan Kesehatan yang berbeda. Penjelasan masing-masing kolom adalah sebagai berikut:

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.

Missing Value

```
In [5]: 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
```

Berdasarkan hasil diatas, dapat kita simpulkan bahwa dataset tidak memiliki missing value

Statistik Deskriptif

In [7]: `df.describe()`

Out[7]:

	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

Hasil diatas menunjukkan statistik deskriptif untuk masing-masing fitur pada dataset

02

Exploratory Data Analysis (EDA)



Bivariate Analysis

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
26	Burundi	93.6	8.92	11.60	39.2	764	12.300	57.7	6.26	231
88	Liberia	89.3	19.10	11.80	92.6	700	5.470	60.8	5.02	327
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.800	57.5	6.54	334
112	Niger	123.0	22.20	5.16	49.1	814	2.550	58.8	7.49	348
132	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.200	55.0	5.20	399
...
44	Denmark	4.1	50.50	11.40	43.6	44000	3.220	79.5	1.87	58000
123	Qatar	9.0	62.30	1.81	23.8	125000	6.980	79.5	2.07	70300
145	Switzerland	4.5	64.00	11.50	53.3	55500	0.317	82.2	1.52	74600
114	Norway	3.2	39.70	9.48	28.5	62300	5.950	81.0	1.95	87800
91	Luxembourg	2.8	175.00	7.77	142.0	91700	3.620	81.3	1.63	105000

167 rows x 10 columns

Negara berdasarkan nilai GDP Perkapita terendah

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.800	57.5	6.54	334
88	Liberia	89.3	19.10	11.80	92.6	700	5.470	60.8	5.02	327
26	Burundi	93.6	8.92	11.60	39.2	764	12.300	57.7	6.26	231
112	Niger	123.0	22.20	5.16	49.1	814	2.550	58.8	7.49	348
31	Central African Republic	149.0	11.80	3.98	26.5	888	2.010	47.5	5.21	446
...
133	Singapore	2.8	200.00	3.96	174.0	72100	-0.046	82.7	1.15	46600
82	Kuwait	10.8	66.70	2.63	30.4	75200	11.200	78.2	2.21	38500
23	Brunei	10.5	67.40	2.84	28.0	80600	16.700	77.1	1.84	35300
91	Luxembourg	2.8	175.00	7.77	142.0	91700	3.620	81.3	1.63	105000
123	Qatar	9.0	62.30	1.81	23.8	125000	6.980	79.5	2.07	70300

167 rows x 10 columns

Negara berdasarkan nilai GDP Perkapita terendah

Negara dengan pendapatan seseorang dan GDP perkapita terendah adalah negara yang kebanyakan berada di benua Afrika

Bivariate Analysis

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
66	Haiti	208.0	15.3	6.91	64.7	1500	5.450	32.1	3.33	662
87	Lesotho	99.7	39.4	11.10	101.0	2380	4.150	46.5	3.30	1170
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.010	47.5	5.21	446
166	Zambia	83.1	37.0	5.89	30.9	3280	14.000	52.0	5.40	1460
94	Malawi	90.5	22.8	6.59	34.9	1030	12.100	53.1	5.31	459
...
68	Iceland	2.6	53.4	9.40	43.3	38800	5.470	82.0	2.20	41900
7	Australia	4.8	19.8	8.73	20.9	41400	1.160	82.0	1.93	51900
145	Switzerland	4.5	64.0	11.50	53.3	55500	0.317	82.2	1.52	74600
133	Singapore	2.8	200.0	3.96	174.0	72100	-0.046	82.7	1.15	46600
77	Japan	3.2	15.0	9.49	13.6	35800	-1.900	82.8	1.39	44500

167 rows x 10 columns

Negara berdasarkan nilai harapan hidup terendah

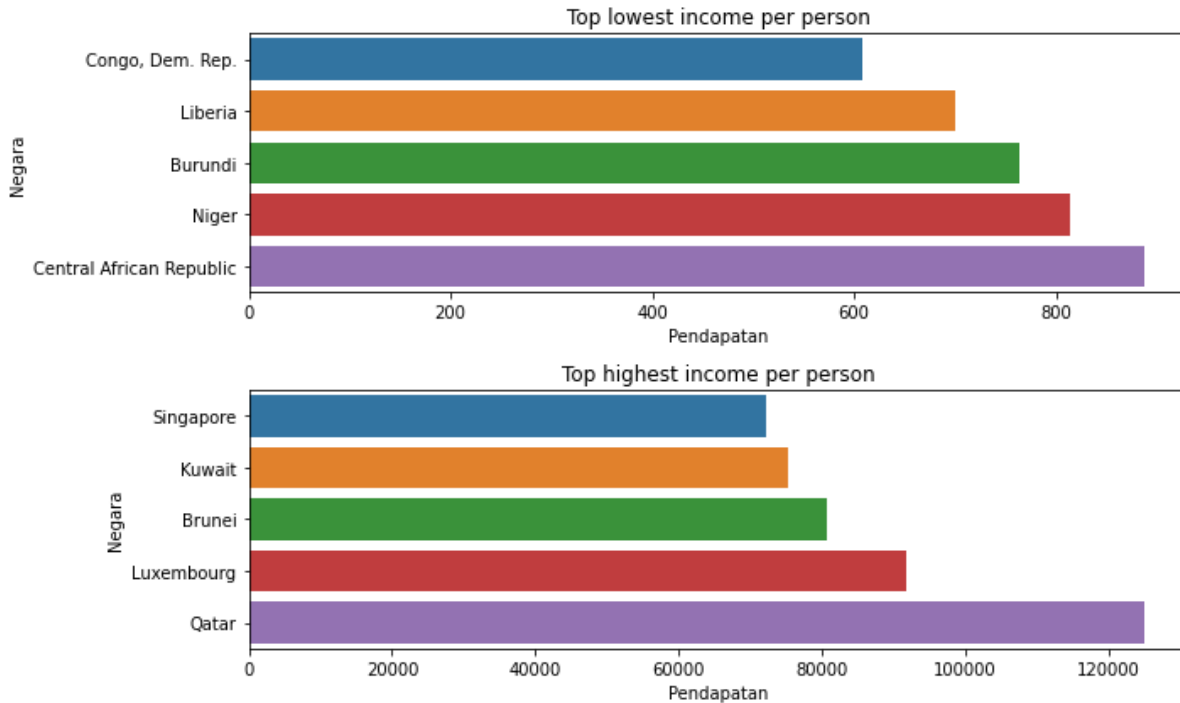
	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
66	Haiti	208.0	15.3	6.91	64.7	1500	5.450	32.1	3.33	662
132	Sierra Leone	160.0	16.8	13.10	34.5	1220	17.200	55.0	5.20	399
32	Chad	150.0	36.8	4.53	43.5	1930	6.390	56.5	6.59	897
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.010	47.5	5.21	446
97	Mali	137.0	22.8	4.98	35.1	1870	4.370	59.5	6.55	708
...
53	Finland	3.0	38.7	8.95	37.4	39800	0.351	80.0	1.87	46200
144	Sweden	3.0	46.2	9.63	40.7	42900	0.991	81.5	1.98	52100
133	Singapore	2.8	200.0	3.96	174.0	72100	-0.046	82.7	1.15	46600
91	Luxembourg	2.8	175.0	7.77	142.0	91700	3.620	81.3	1.63	105000
68	Iceland	2.6	53.4	9.40	43.3	38800	5.470	82.0	2.20	41900

167 rows x 10 columns

Negara berdasarkan nilai kematian anak tertinggi

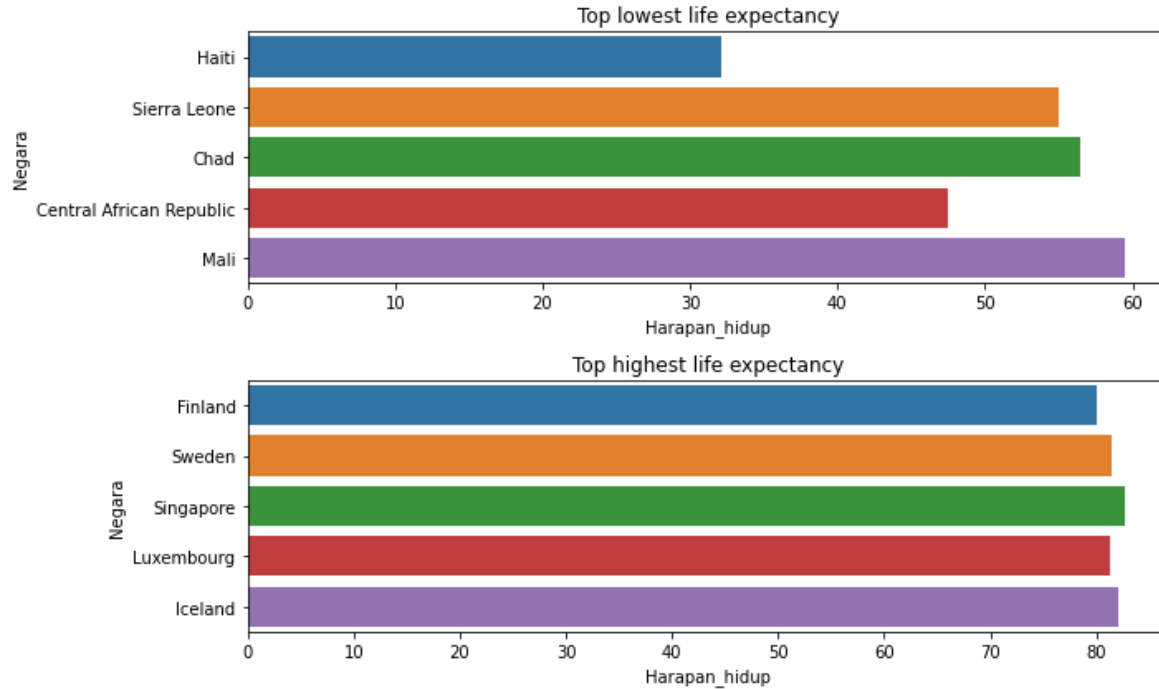
Negara di benua Afrika menjadi negara mayoritas yang memiliki harapan hidup rendah dan angka kematian anak tertinggi.

Bivariate Analysis



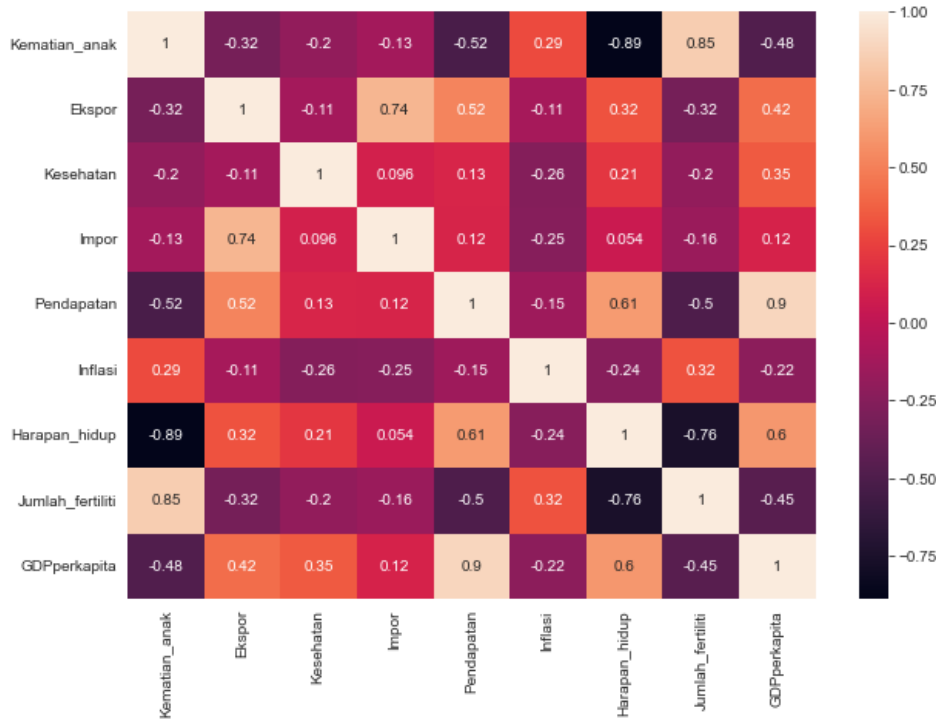
Grafik perbandingan 5 negara dengan pendapatan per orang terendah vs tertinggi

Bivariate Analysis



Grafik perbandingan 5 negara dengan indikator harapan hidup terendah vs tertinggi

Multivariate Analysis



Korelasi positif terkuat adalah antara Pendapatan dan GDPperkapita dengan nilai korelasi 0.9

Yang menandakan Ketika pendapatan meningkat, maka GDP perkapita juga meningkat

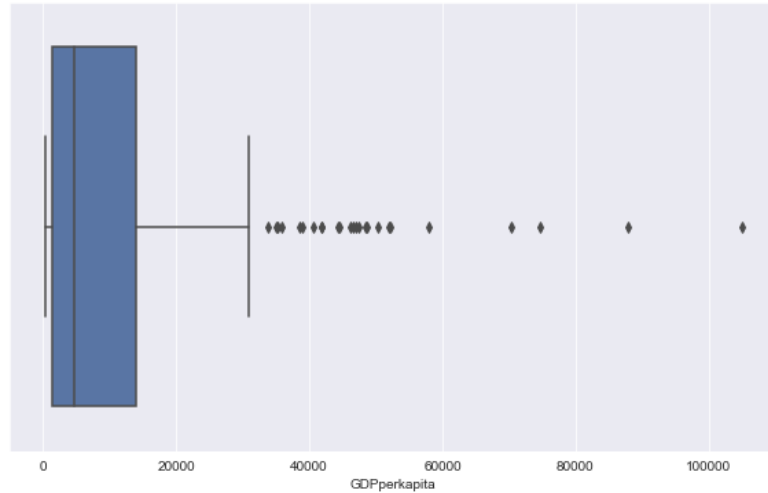
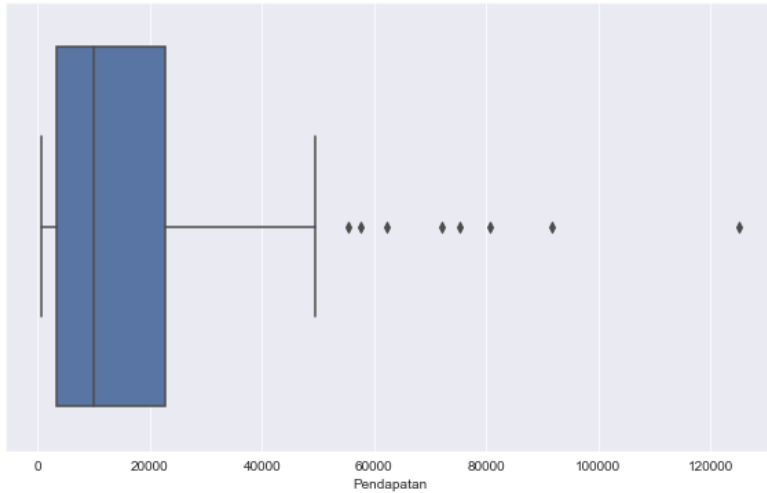
Selain itu, kematian_anak & Jumlah_fertility dan Impor & Ekspor juga memiliki korelasi positif yang kuat

03

Outlier Treatment

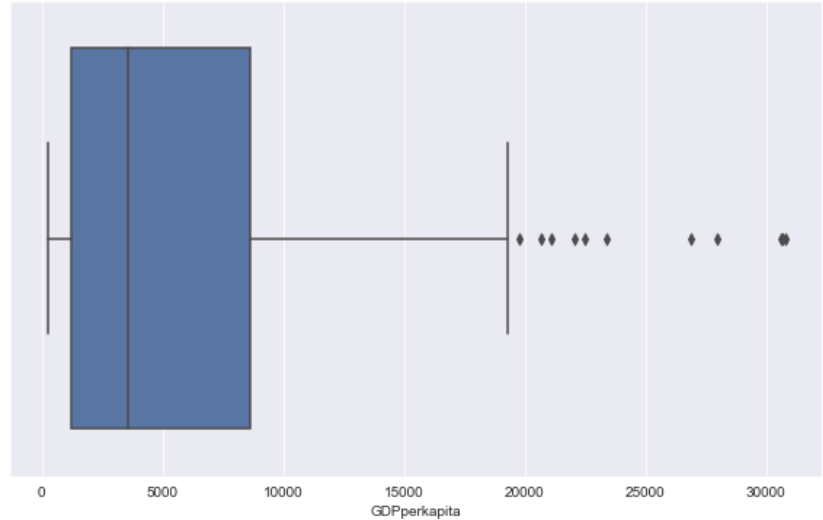
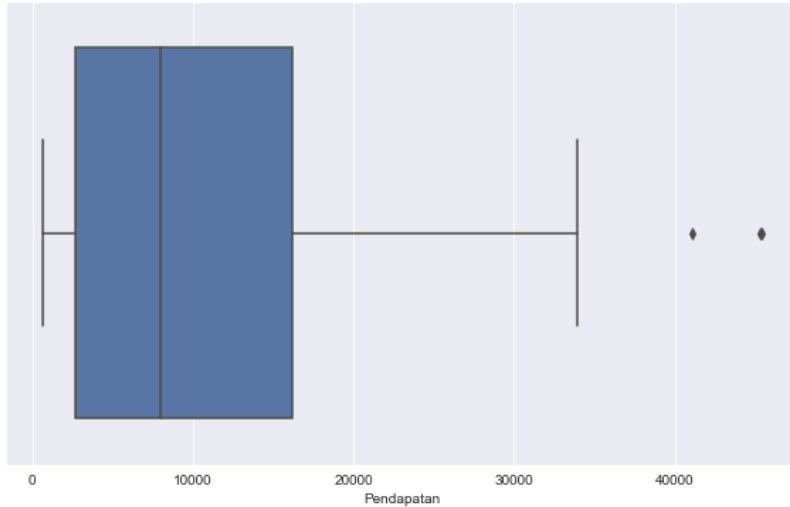


Outlier

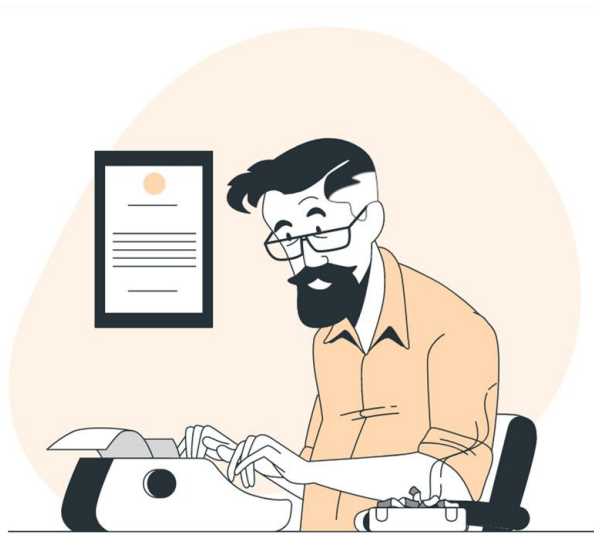


Terdapat beberapa outlier pada data pendapatan dan gdp perkapita, dimana outlier pendapatan terdapat di atas 50000 dan outlier GDP perkapita berada diatas 30000

Handling Outlier



Setelah dilakukan handling outlier, outlier pada kedua data berkurang seperti gambar diatas



04

Scaling Data & Plot

Scaling Data

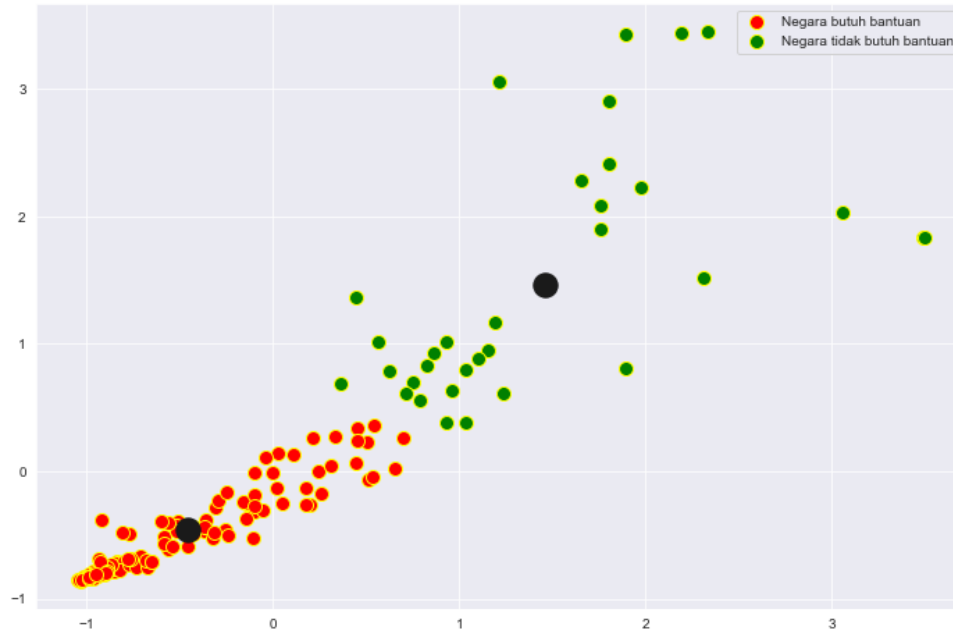
	Pendapatan	GDPperkapita	labels1
0	-0.945783	-0.816938	0
1	-0.101991	-0.317709	0
2	0.199219	-0.265485	0
3	-0.510703	-0.396750	0
4	0.828006	0.826975	1
...
137	-0.809884	-0.475791	0
138	0.564321	1.010463	1
139	-0.653701	-0.710091	0
140	-0.654715	-0.710091	0
141	-0.776416	-0.688919	0

142 rows × 3 columns

Scaling dilakukan menggunakan standard scaler dari sklearn sehingga data berdistribusi normal, yaitu memiliki rentan nilai antara -1 dan 1.

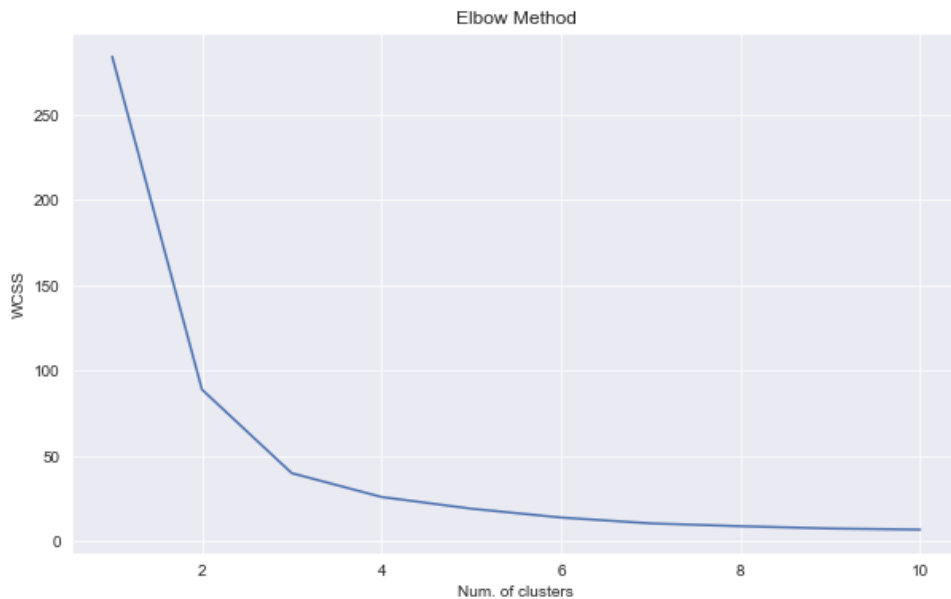
Negara yang membutuhkan bantuan di beri label 0 dan negara yang tidak membutuhkan bantuan diberi label 1

Plot



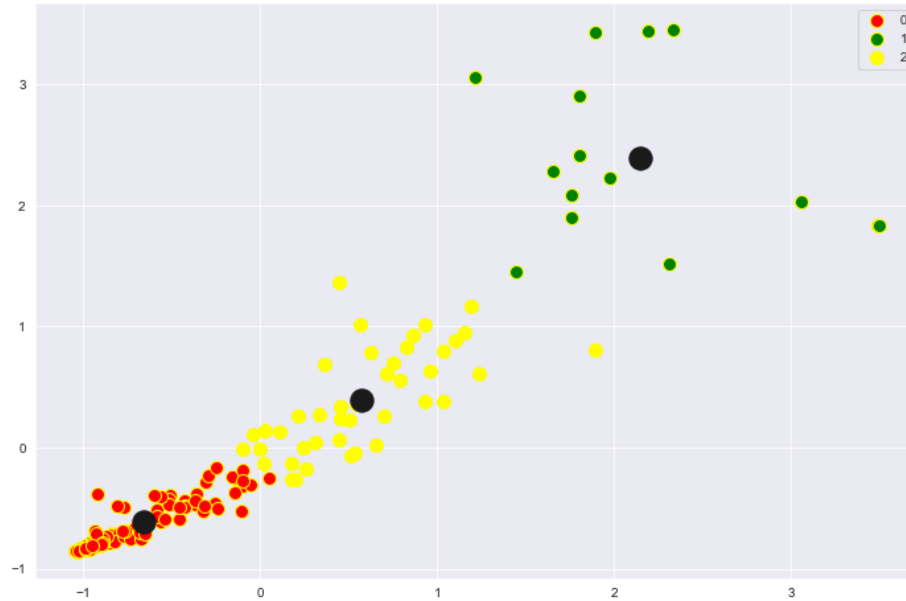
Negara yang membutuhkan bantuan berada pada rentan -1 hingga 1 dengan warna merah, Plot ini menggunakan metode Kmeans dengan n_clusters berjumlah 2

Elbow Method



Agar tau berapa jumlah `n_clusters` yang optimal bagi model, digunakanlah metode elbow. Disini saya coba untuk ambil `n_clusters=3`

Plot



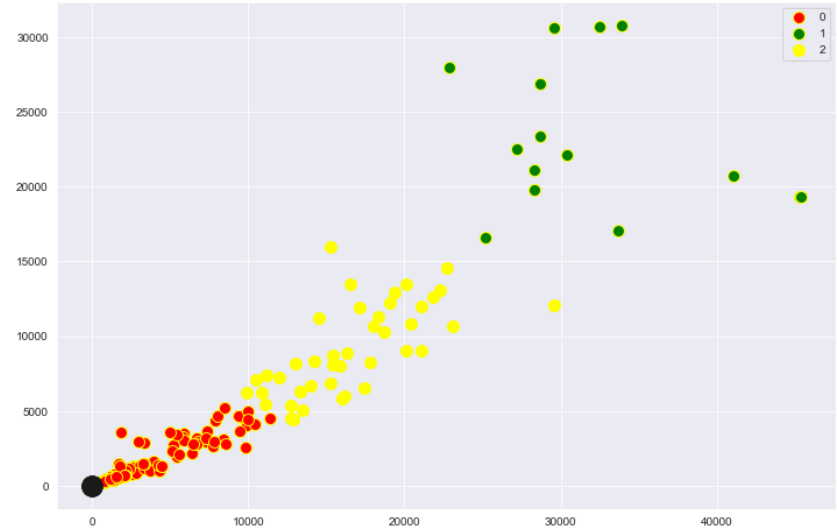
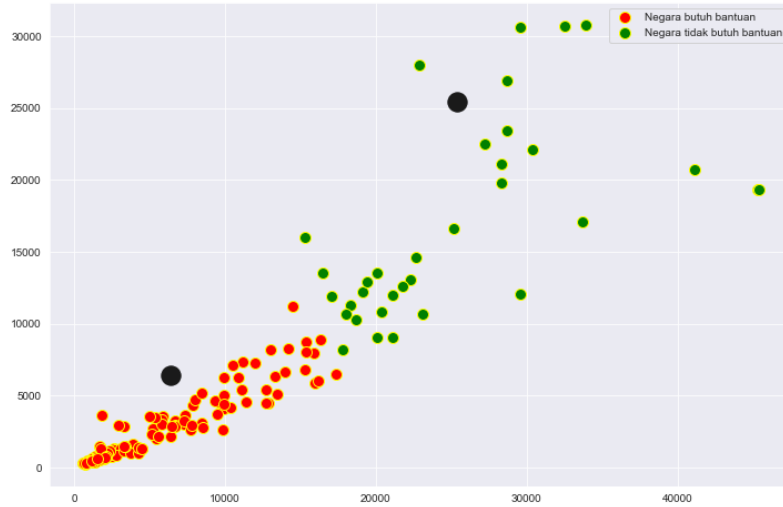
Setelah menggunakan `n_clusters=3` didapatkan klusterisasi data sebagai diatas, negara yang membutuhkan bantuan diberi warna merah.

Silhoutte Score

```
Score n_clusters = 2: 0.6547621766577557  
Score n_clusters = 3: 0.6156187054467405
```

Setelah melakukan uji coba, didapati silhouette score untuk nilai n clusters = 2 lebih baik dibanding dengan n_clusters = 3, maka akan digunakan n_clusters = 2

Plot

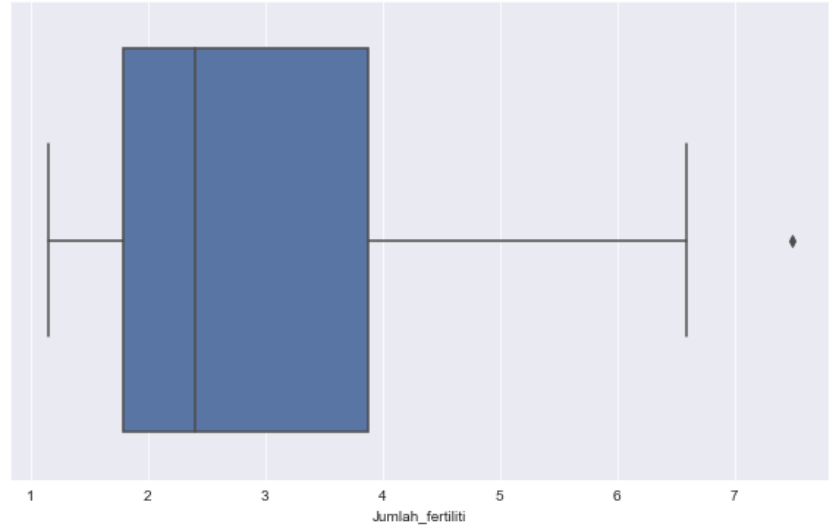
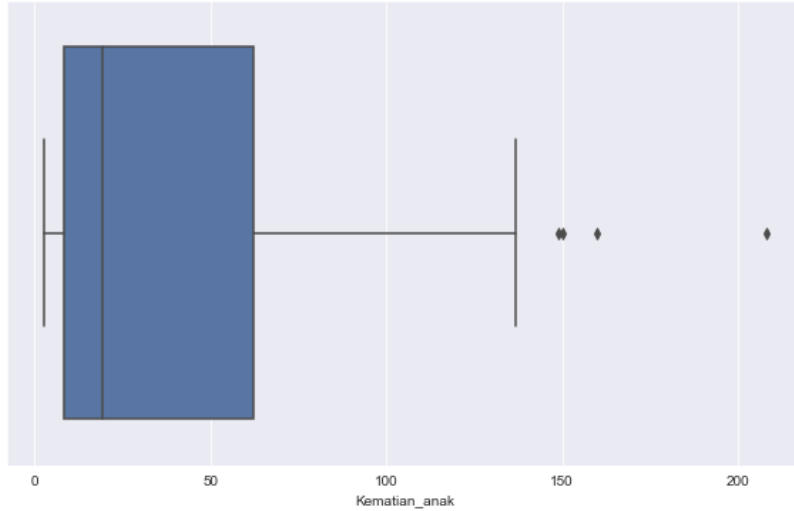


Plot diatas adalah plot setelah dilakukan inverse transform

Segi Sosial Kesehatan

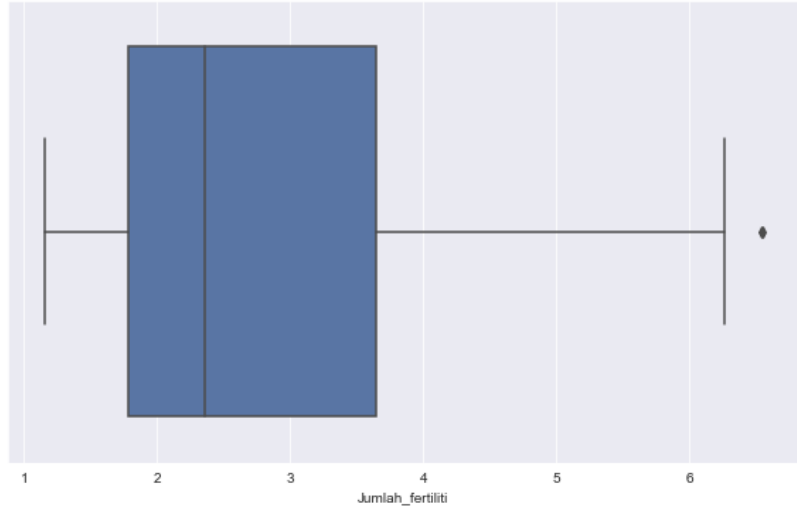
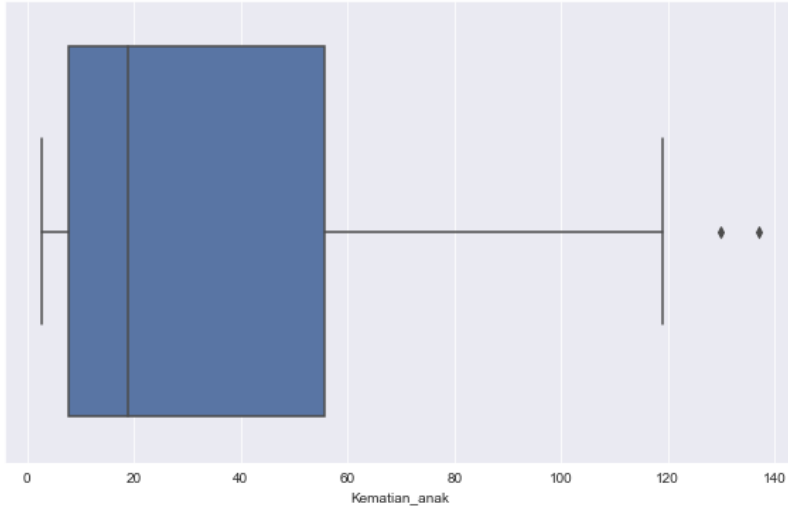


Outlier



Terdapat beberapa outlier untuk data kematian anak dan jumlah fertiliti

Handling Outlier



Outlier berkurang setelah dilakukan handling outlier

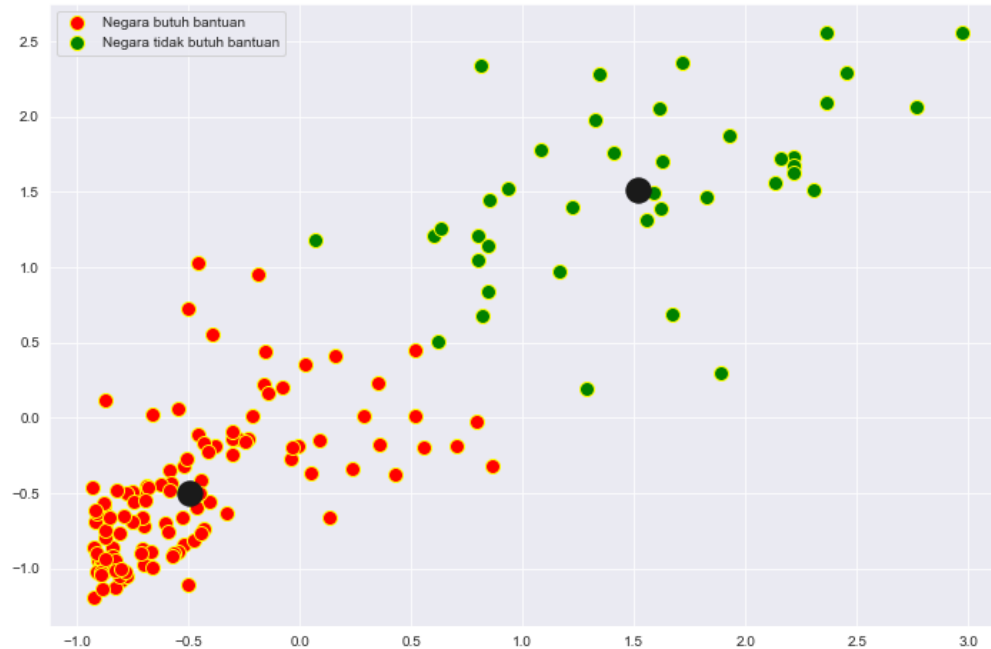
Handling Outlier

	Kematian_anak	Jumlah_fertiliti	labels1_kes
0	1.614829	2.052926	1
1	-0.521814	-0.846277	0
2	-0.211188	0.015836	0
3	2.450907	2.289312	1
4	-0.704706	-0.512556	0
...
157	-0.156030	0.439940	0
158	-0.507299	-0.276170	0
159	-0.327310	-0.637701	0
160	0.630696	1.253386	1
161	1.408713	1.760920	1

162 rows × 3 columns

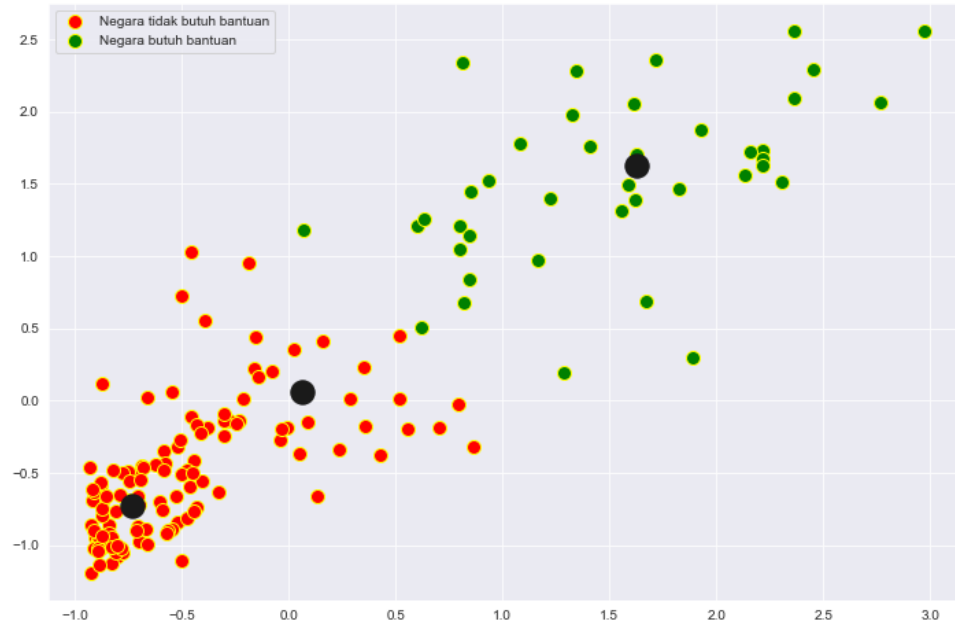
Dilakukan standarisasi dengan standard scaler, label 0 untuk negara yang membutuhkan bantuan dan label 1 untuk negara yang tidak membutuhkan bantuan

Plot



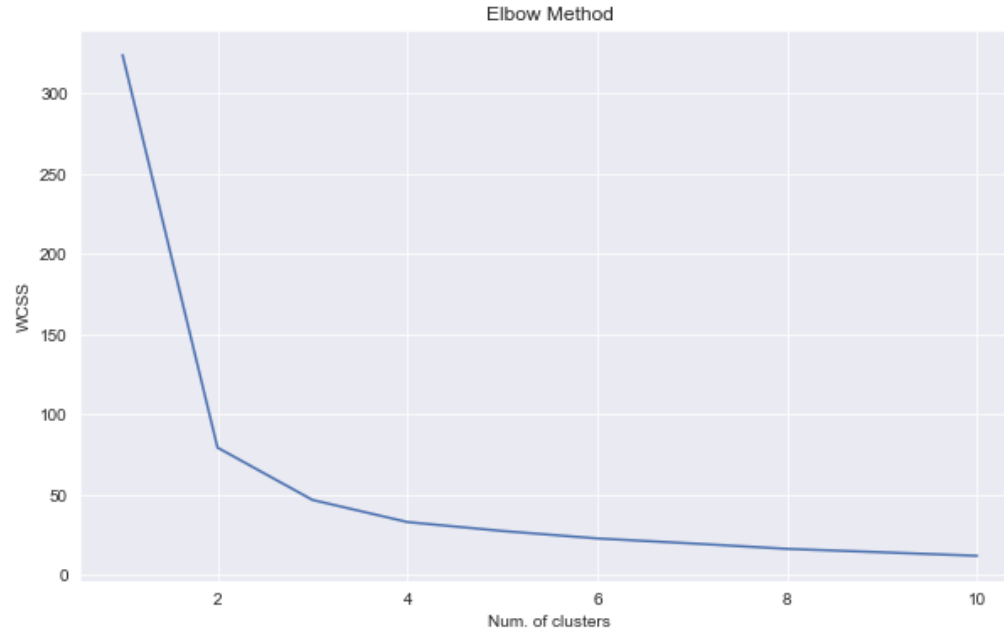
Negara yang membutuhkan bantuan diberi warna hijau, dan tersebar antara nilai 1 hingga 3. Hal ini menunjukkan bahwa jika kematian anak yang tinggi, menandakan jumlah fertilitas juga tinggi

Plot



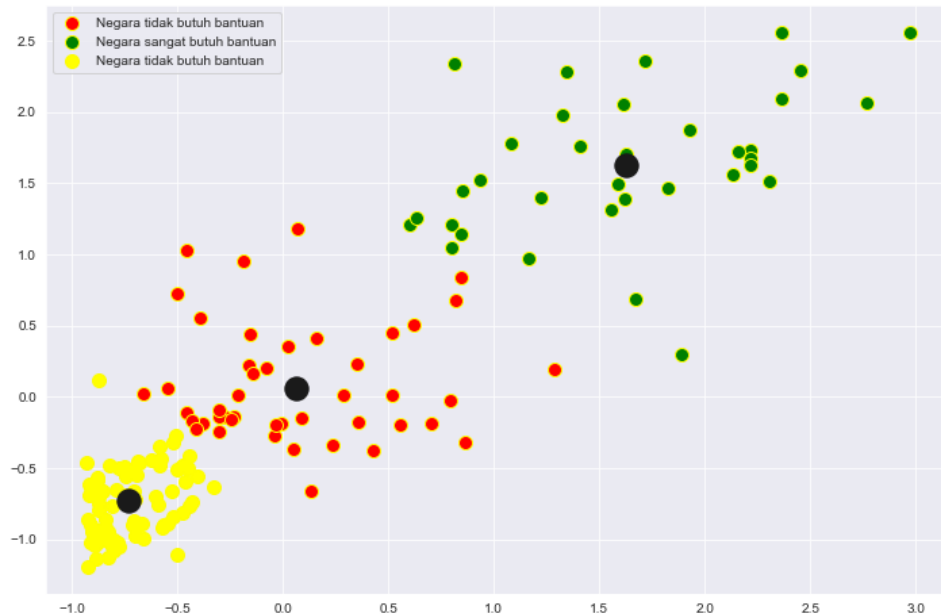
Negara yang membutuhkan bantuan diberi warna hijau, dan tersebar antara nilai 1 hingga 3. Hal ini menunjukkan bahwa jika kematian anak yang tinggi, menandakan jumlah fertilitas juga tinggi

Metode Elbow



Agar tau berapa jumlah `n_clusters` yang optimal bagi model, digunakanlah metode elbow. Disini saya coba untuk ambil `n_clusters=3`

Plot



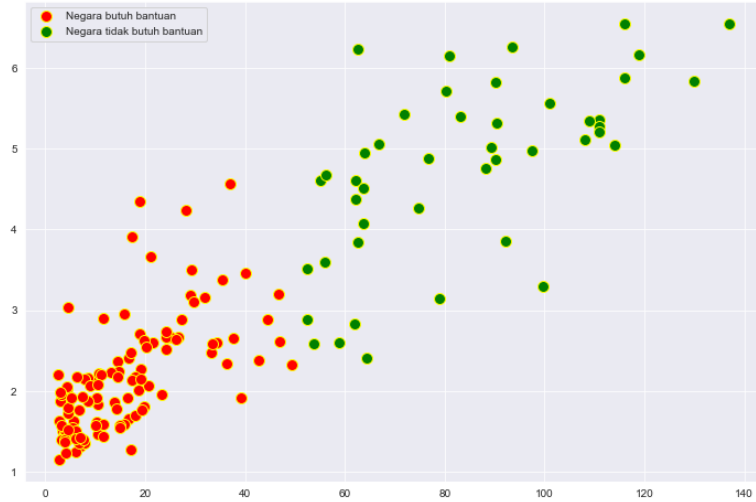
Terdapat 3 cluster seperti gambar diatas, negara yang sangat membutuhkan bantuan, membutuhkan bantuan, dan tidak membutuhkan bantuan

Silhoutte score kesehatan

```
Score n_cluster = 2 kesehatan: 0.6866507776692582  
Score n_cluster = 3 kesehatan: 0.5469473064927436
```

Didapatkan score tertinggi diberikan oleh nilai $n_cluster = 2$. maka dari itu pada kasus ini akan dibuat 2 cluster saja. Yaitu negara yang membutuhkan bantuan dan yang tidak membutuhkan bantuan

Plot inverse transform



Plot setelah dilakukan inverse transform pada data

Penggabungan Dataframe

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	labels1	labels2	labels1_kes	lab
0	Afghanistan	90.2	10.0	7.58	44.9	1610.0	9.440	56.2	5.82	553.0	0.0	0.0	1.0	
1	Albania	16.6	28.0	6.55	48.6	9930.0	4.490	76.3	1.65	4090.0	0.0	0.0	0.0	
2	Algeria	27.3	38.4	4.17	31.4	12900.0	16.100	76.5	2.89	4460.0	0.0	2.0	0.0	
3	Angola	119.0	62.3	2.85	42.9	5900.0	22.400	60.1	6.16	3530.0	0.0	0.0	1.0	
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100.0	1.440	76.8	2.13	12200.0	1.0	2.0	0.0	
5	Argentina	14.5	18.9	8.10	16.0	18700.0	20.900	75.8	2.37	10300.0	1.0	2.0	0.0	
6	Armenia	18.1	20.8	4.40	45.3	6700.0	7.770	73.3	1.69	3220.0	0.0	0.0	0.0	
7	Australia	4.8	19.8	8.73	20.9	41400.0	1.160	82.0	1.93	51900.0	NaN	NaN	NaN	
8	Austria	4.3	51.3	11.00	47.8	43200.0	0.873	80.5	1.44	46900.0	NaN	NaN	NaN	
9	Azerbaijan	39.2	54.3	5.88	20.7	16000.0	13.800	69.1	1.92	5840.0	0.0	2.0	0.0	

Berikut merupakan 10 data teratas dari hasil penggabungan dataframe ekonomi dan social kesehatan

Summary



Filtering data

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	labels1	labels2	labels1_ke
0	Afghanistan	90.2	10.000	7.58	44.9000	1610.0	9.440	56.2	5.82	553.0	0.0	0.0	1
3	Angola	119.0	62.300	2.85	42.9000	5900.0	22.400	60.1	6.16	3530.0	0.0	0.0	1
17	Benin	111.0	23.800	4.10	37.2000	1820.0	0.885	61.8	5.36	758.0	0.0	0.0	1
21	Botswana	52.5	43.600	8.30	51.3000	13300.0	8.920	57.1	2.88	6350.0	0.0	2.0	1
26	Burundi	93.6	8.920	11.60	39.2000	764.0	12.300	57.7	6.26	231.0	0.0	0.0	1
28	Cameroon	108.0	22.200	5.13	27.0000	2660.0	1.910	57.3	5.11	1310.0	0.0	0.0	1
36	Comoros	88.2	16.500	4.51	51.7000	1410.0	3.870	65.9	4.75	769.0	0.0	0.0	1
37	Congo, Dem. Rep.	116.0	41.100	7.91	49.6000	609.0	20.800	57.5	6.54	334.0	0.0	0.0	1
38	Congo, Rep.	63.9	85.100	2.46	54.7000	5190.0	20.700	60.4	4.95	2740.0	0.0	0.0	1
40	Cote d'Ivoire	111.0	50.600	5.30	43.3000	2690.0	5.390	56.3	5.27	1220.0	0.0	0.0	1
50	Eritrea	55.2	4.790	2.66	23.3000	1420.0	11.600	61.7	4.61	482.0	0.0	0.0	1
55	Gabon	63.7	57.700	3.50	18.9000	15400.0	16.600	62.9	4.08	8750.0	0.0	2.0	1
56	Gambia	80.3	23.800	5.69	42.7000	1660.0	4.300	65.5	5.71	562.0	0.0	0.0	1

Dilakukan filtering data dengan nilai label ekonomi (labels1) = 0 dan label social Kesehatan (labels1_kes) = 1 untuk mendapatkan negara mana saja yang membutuhkan bantuan

Kluster Negara yang membutuhkan bantuan – Faktor GDP Perkapita

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	labels1	labels2	labels1_kes
26	Burundi	93.6	8.92	11.60	39.2	764.0	12.30	57.7	6.26	231.0	0.0	0.0	1.0
88	Liberia	89.3	19.10	11.80	92.6	700.0	5.47	60.8	5.02	327.0	0.0	0.0	1.0
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609.0	20.80	57.5	6.54	334.0	0.0	0.0	1.0
93	Madagascar	62.2	25.00	3.77	43.0	1390.0	8.79	60.8	4.60	413.0	0.0	0.0	1.0
106	Mozambique	101.0	31.50	5.21	46.2	918.0	7.64	54.5	5.56	419.0	0.0	0.0	1.0
94	Malawi	90.5	22.80	6.59	34.9	1030.0	12.10	53.1	5.31	459.0	0.0	0.0	1.0
50	Eritrea	55.2	4.79	2.66	23.3	1420.0	11.60	61.7	4.61	482.0	0.0	0.0	1.0
150	Togo	90.3	40.20	7.65	57.3	1210.0	1.18	58.7	4.87	488.0	0.0	0.0	1.0
64	Guinea- Bissau	114.0	14.90	8.50	35.2	1390.0	2.97	55.6	5.05	547.0	0.0	0.0	1.0
0	Afghanistan	90.2	10.00	7.58	44.9	1610.0	9.44	56.2	5.82	553.0	0.0	0.0	1.0

Berikut merupakan 10 negara yang membutuhkan bantuan diurut berdasarkan GDP perkapita terendah. Negara Burundi menjadi negara dengan GDP perkapita terendah, Berdasarkan lokasi, mayoritas negara yang membutuhkan bantuan adalah negara di benua Afrika. Hal ini didukung oleh beberapa factor seperti factor Kesehatan, harapan hidup, ekspor, dan pendapatan yang rendah.

Kluster Negara yang Membutuhkan Bantuan – Faktor Kesehatan

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	labels1	labels2	labels1_kes
107	Myanmar	64.4	0.109	1.97	0.0659	3720.0	7.040	66.8	2.41	988.0	0.0	0.0	1.0
116	Pakistan	92.1	13.500	2.20	19.4000	4280.0	10.900	65.3	3.85	1040.0	0.0	0.0	1.0
38	Congo, Rep.	63.9	85.100	2.46	54.7000	5190.0	20.700	60.4	4.95	2740.0	0.0	0.0	1.0
154	Turkmenistan	62.0	76.300	2.50	44.5000	9940.0	2.310	67.9	2.83	4440.0	0.0	0.0	1.0
50	Eritrea	55.2	4.790	2.66	23.3000	1420.0	11.600	61.7	4.61	482.0	0.0	0.0	1.0
3	Angola	119.0	62.300	2.85	42.9000	5900.0	22.400	60.1	6.16	3530.0	0.0	0.0	1.0
55	Gabon	63.7	57.700	3.50	18.9000	15400.0	16.600	62.9	4.08	8750.0	0.0	2.0	1.0
93	Madagascar	62.2	25.000	3.77	43.0000	1390.0	8.790	60.8	4.60	413.0	0.0	0.0	1.0
69	India	58.8	22.600	4.05	27.1000	4410.0	8.980	66.2	2.60	1350.0	0.0	0.0	1.0
17	Benin	111.0	23.800	4.10	37.2000	1820.0	0.885	61.8	5.36	758.0	0.0	0.0	1.0

Berikut merupakan 10 negara yang membutuhkan bantuan diurut berdasarkan factor kesehatan terendah. Negara Myanmar menjadi negara dengan Kesehatan terendah, Berdasarkan lokasi, mayoritas negara yang membutuhkan bantuan adalah negara di benua Afrika dan Asia.



Thanks!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Storyset