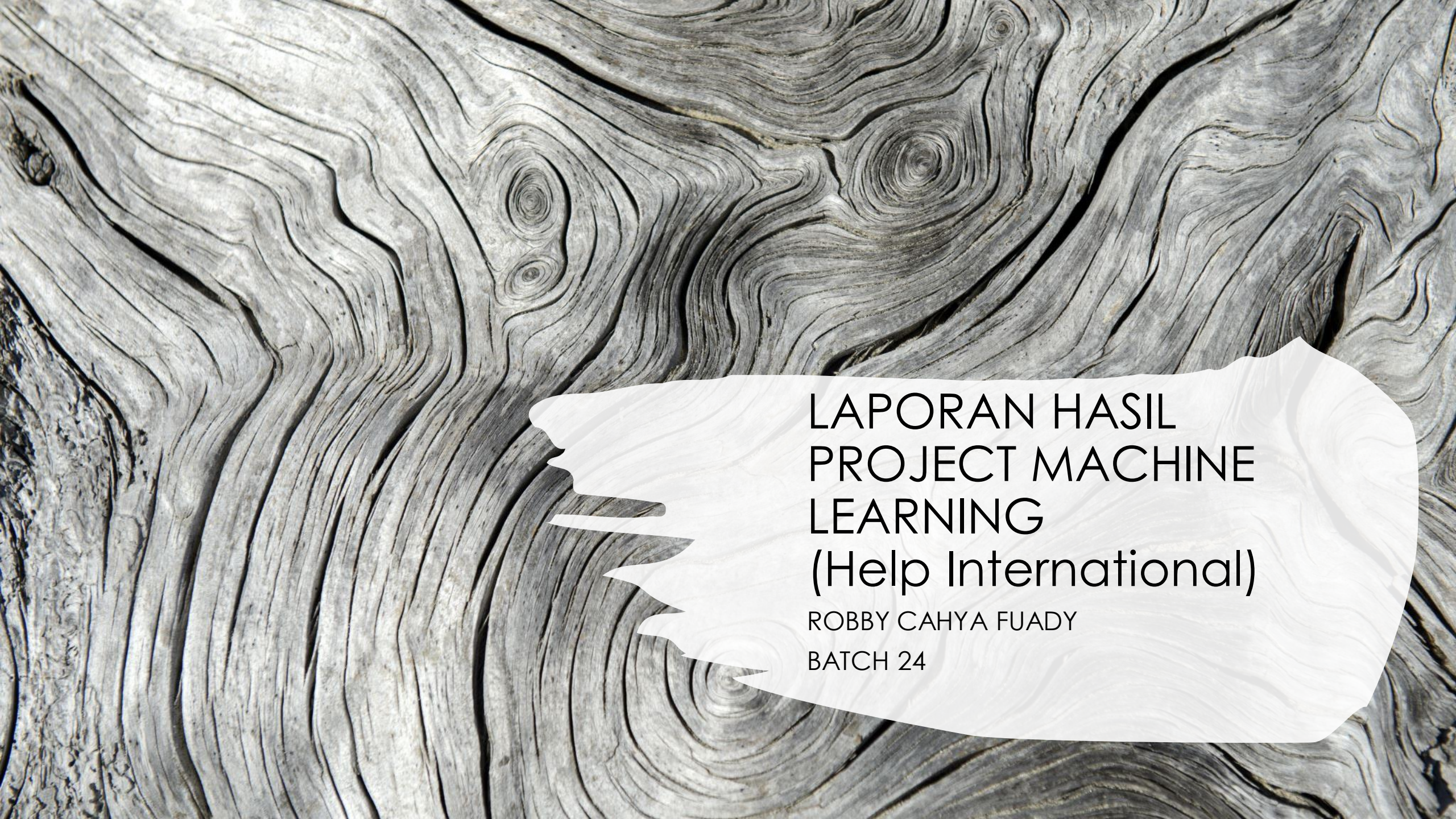


LAPORAN HASIL PROJECT MACHINE LEARNING (Help International)

ROBBY CAHYA FUADY

BATCH 24

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.

TUJUAN

- Mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan
- Dari hasil pengkategorian didapatan Negara-negara mana sajak yang paling perlu menjadi fokus CEO Help International.

METODE MACHINE LEARNING YANG AKAN DIGUNAKAN

- Metode K-Means Clustering berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain, algoritma dasarnya sebagai berikut:

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam cluster secara random
3. Hitung centroid/rata-rata dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan

HASIL DAN PEMBAHASAN

- Proses pembacaan dataset menggunakan library pandas dan didapatkan hasil 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

- Dataset tersebut memiliki 167 baris dan 10 kolom termasuk index default dari pandas yang berisikan data :

1. Negara
2. Kematian anak
3. Ekspor
4. Kesehatan
5. Impor
6. Pendapatan
7. Inflasi
8. Harapan hidup
9. Jumlah fertility
10. GDP perkapita

- Summary secara keseluruhan dari data tersebut yaitu sebagai berikut

	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

1.rata-rata nilai per kolom

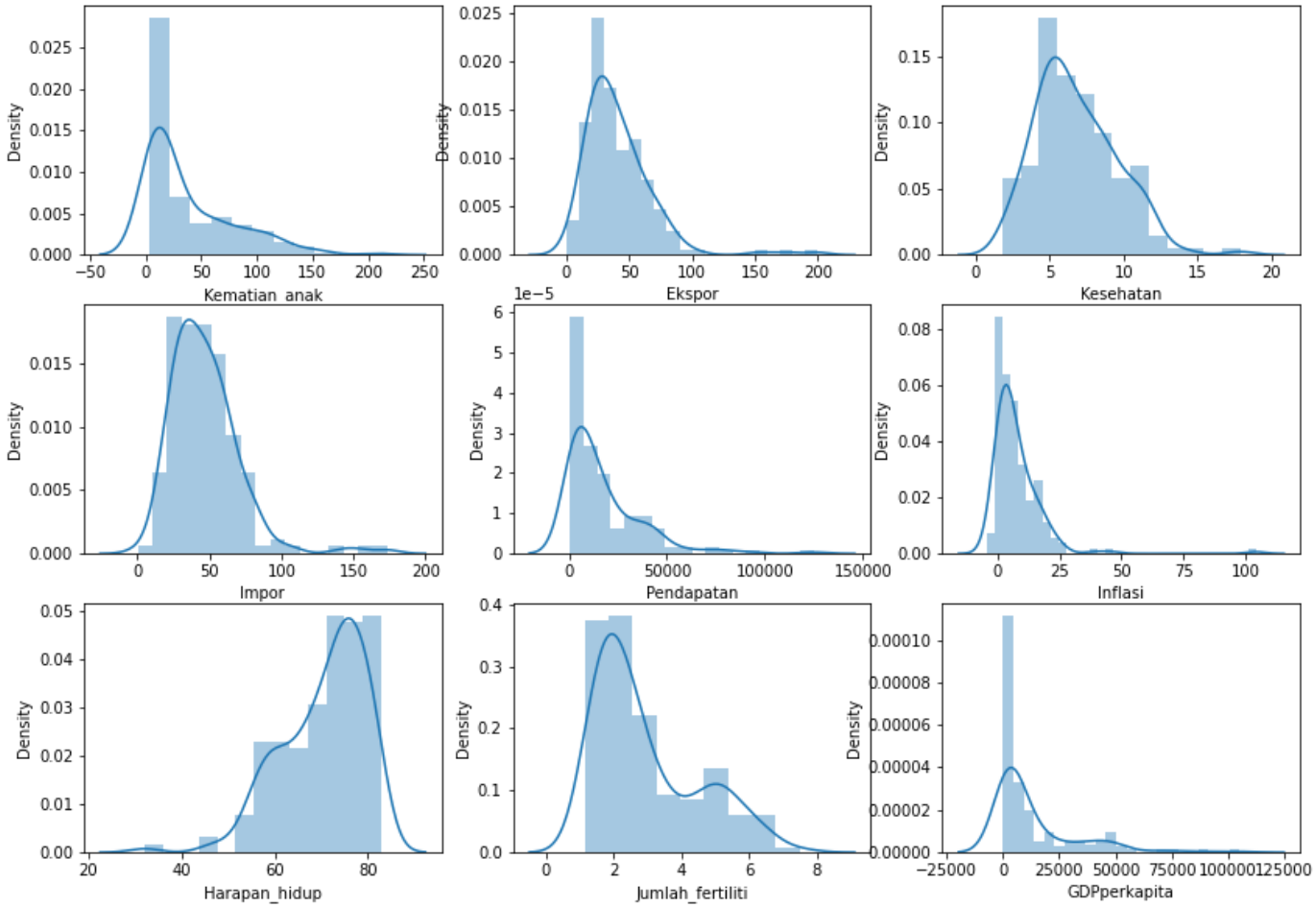
2.Standar deviasi

3.Nilai minimum

4.Nilai maximum

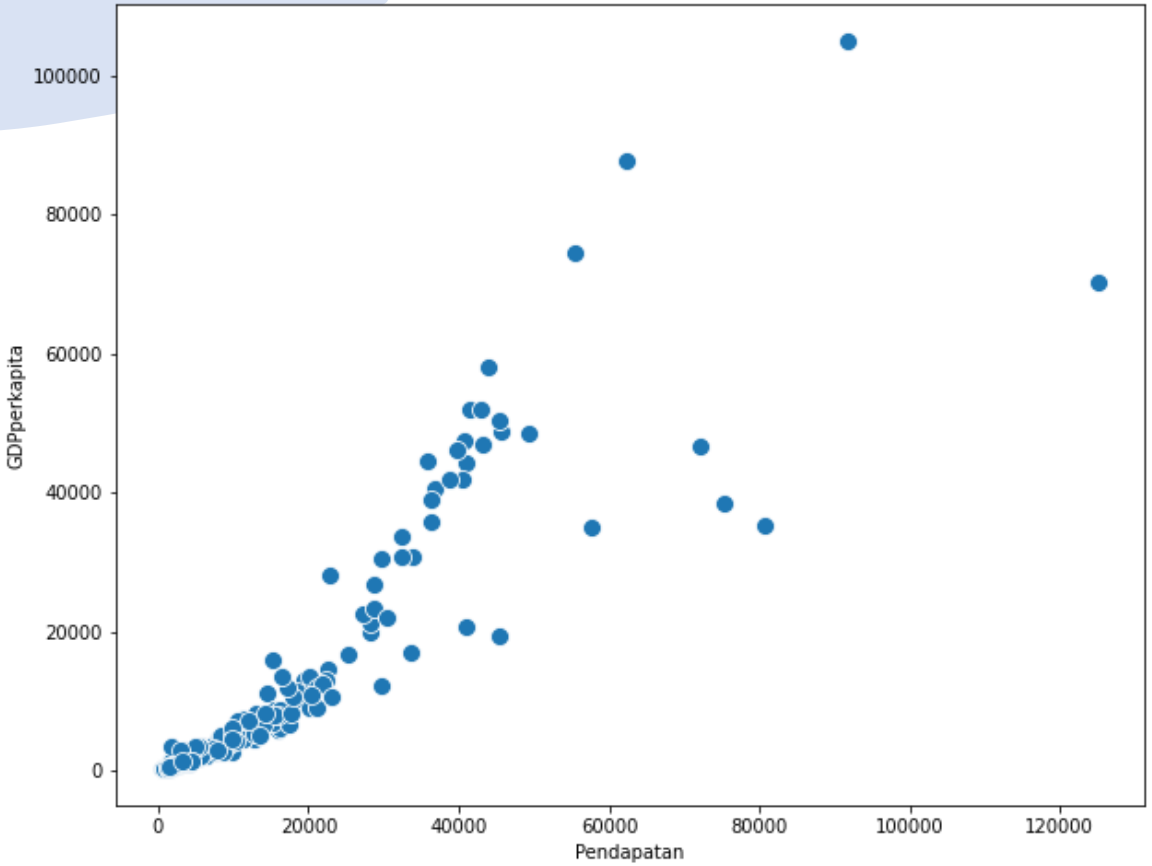
EXPLORATORY DATA ANALYSIS

- Analisa univariate



- Karena variable yang dipakai dalam projek ini hanya 4 yaitu kolom pendapatan, GDP perkapita, kematian anak, dan jumlah fertility maka dari plotting menggunakan distplot sebelumnya dapat dianalisa bahwa:
- Pada kolom kematian anak jumlah paling banyak terdapat pada skor 0-20
- Jumlah pendapatan tertinggi terdapat pada nilai 0-10000
- Kolom jumlah fertiliti memiliki jumlah tertinggi pada nilai 1 sampai 3
- Pada kolom GDP perkapita jumlah tertinggi terdapat pada nilai antara 0 sampai sekitar 5000

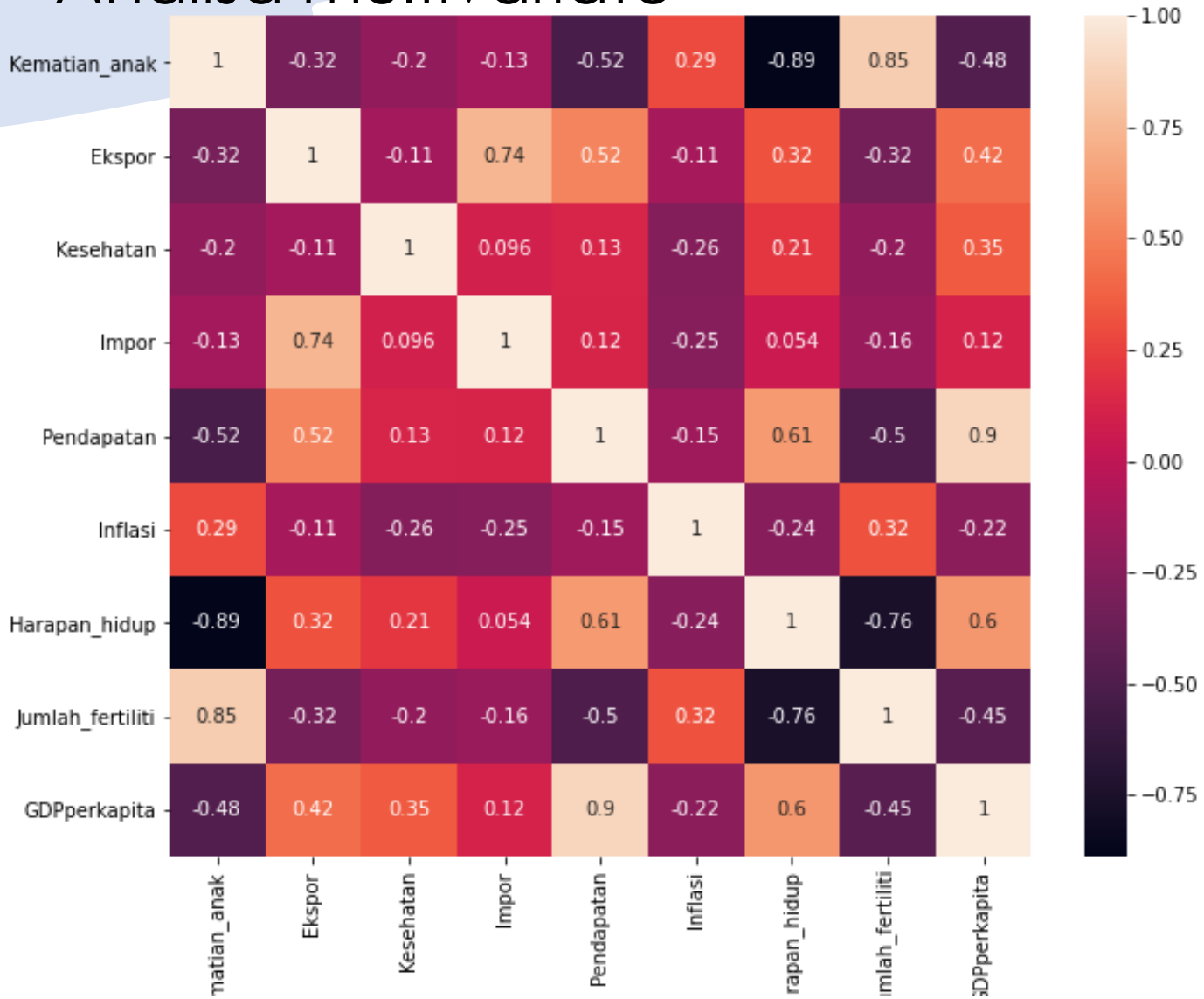
- Analisa bivariate



Analisa bivariate menggunakan scatterplot antara kolom pendapatan dan GDP perkapita

- Dari scatterplot didapatkan insight bahwa semakin tinggi pendapatan maka semakin tinggi pula GDP perkapita
- Dapat dilihat juga bahwa terdapat 2 data pencilan discatter plot

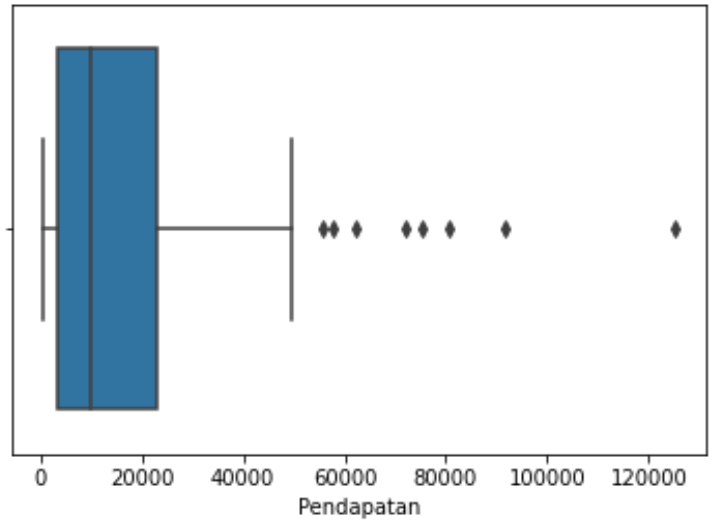
- Analisa multivariate

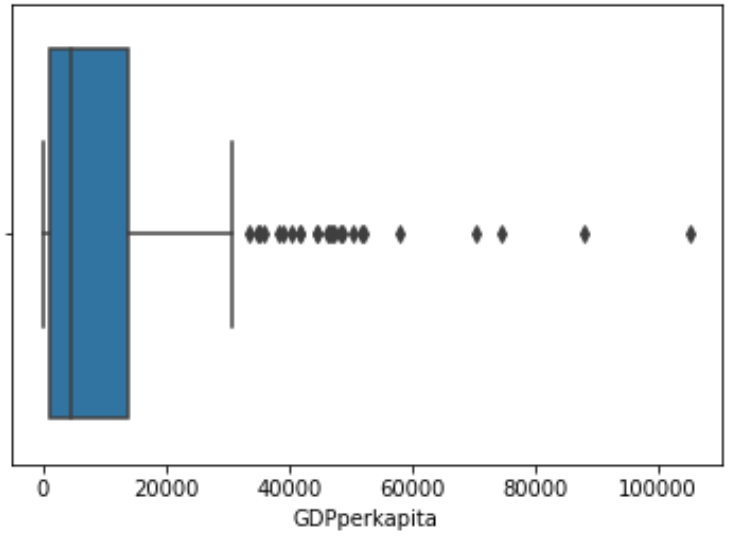


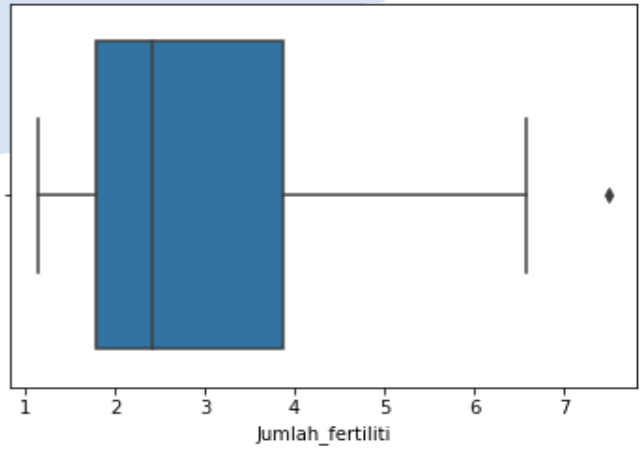
- Pada multivariate menggunakan plotting heatmap didapatkan nilai korelasi pada masing-masing kolom
- Dari nilai korelasi inilah yang menjadi dasar pemilihan variable yang akan dipilih untuk klustering
- Dari beberapa kolom nilai korelasi tersebut diambil nilai korelasi 0.9 antara kolom pendapatan dan GDP perkapita
- Nilai korelasi 0.85 juga diambil sebagai klustering kedua antara kolom jumlah fertility dan kematian anak

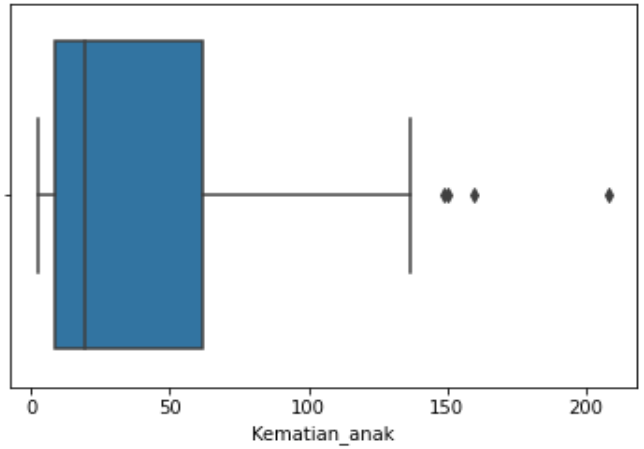
OUTLIER TREATMENT

- Dari 4 variable hasil pemilihin berdasarkan nilai korelasi dari Analisa multivariate masing-masing memiliki nilai outlier atau dinamakan data pencilan
- Berikut masing-masing plotting dari 4 variable tersebut menggunakan boxplot









- Dari 4 variable tersebut data pencilannya akan dihilangkan karena nantinya akan mempengaruhi plotting scatter dan pemberian label Kmeans nantinya

SCALING DATA DAN PEMBERIAN LABEL MENGGUNAKAN KMEANS CLUSTERING

- Scaling data perlu dilakukan karena ketika menggunakan metode klustering sangat sensitive terhadap jarak nilai antar data dikarenakan menggunakan Euclidean distance
- Maka dari 4 variable tersebut dilakukan scalling menggunakan StandardScaller tiap variable
- Setelah dilakukan scalling, selanjutnya diberikan jumlah cluster secara acak, disini diberi 2 cluster dan didapatkan hasil berikut

	Pendapatan	GDPperkapita	labels1_kmeans
0	-0.973550	-0.821294	0
1	-0.118393	-0.323232	0
2	0.186874	-0.271131	0
3	-0.532610	-0.402089	0
4	0.824131	0.818776	1
...	...	...	...
124	-0.703230	-0.704841	0
125	-0.835821	-0.480945	0
126	-0.677534	-0.714698	0
127	-0.678562	-0.714698	0
128	-0.801902	-0.693575	0

129 rows × 3 columns

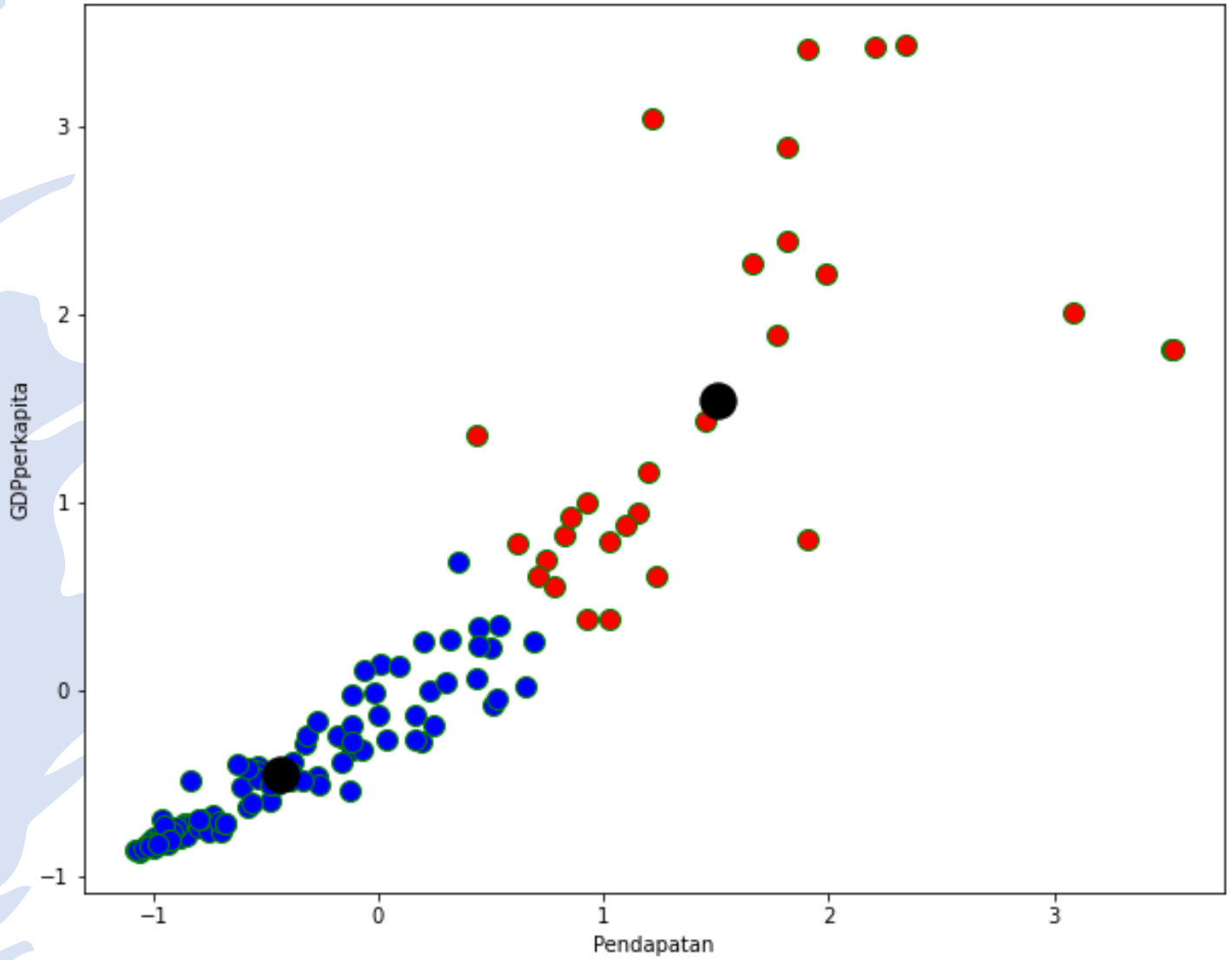
	Kematian_anak	Jumlah_fertiliti	labels2_kmeans
0	1.492609	1.922744	1
1	-0.664447	-0.943787	0
2	-0.350853	-0.091389	0
3	2.336675	2.156465	1
4	-0.849086	-0.613826	0
...	...	...	...
124	-0.087083	-0.469469	0
125	-0.295168	0.327935	0
126	-0.468085	-0.737561	0
127	0.499074	1.132214	1
128	1.284524	1.634028	1

129 rows × 3 columns

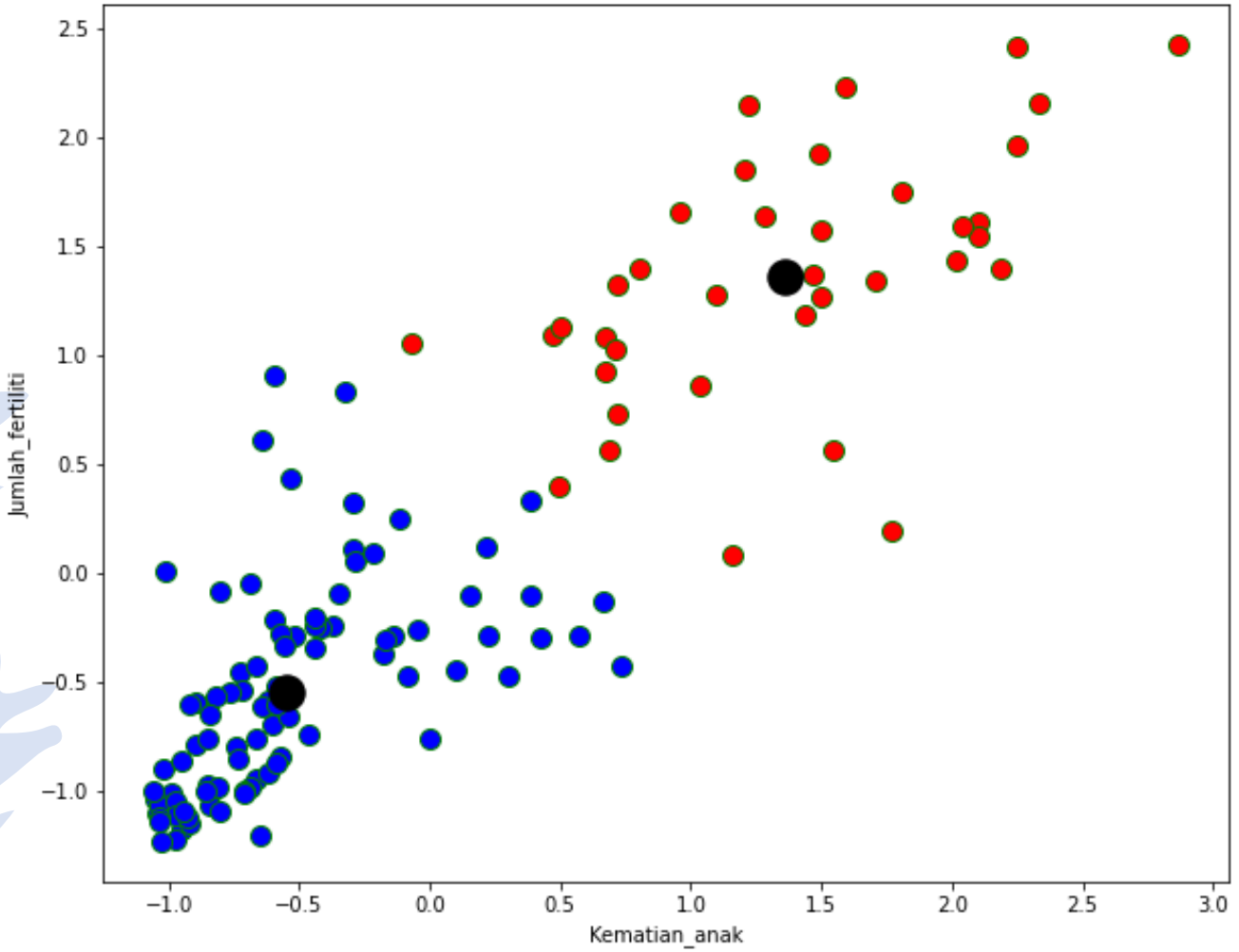
PLOTTING HASIL PENERAPAN KMEANS CLUSTERING

- Setelah penerapan Kmeans clustering, maka dilakukan plotting dari masing-masing 2 variable tersebut menggunakan scatte plot
- Hasilnya sebagai berikut:

- Dari hasil plotting disamping didapatkan insight pada klustering warna biru yaitu negara yang pendapatan dan GDP perkapitanta berada pada nilai -1 sampai 1 berada pada kluster 1, sedangkan negara yang berada pada kluster 2 yaitu pendapatan dan GDP perkapitanya dari nilai 1 sampai 3
- Didapat insight bahwa semakin tinggi pendapatan maka semakin tinggi pula GDP perkapitanya



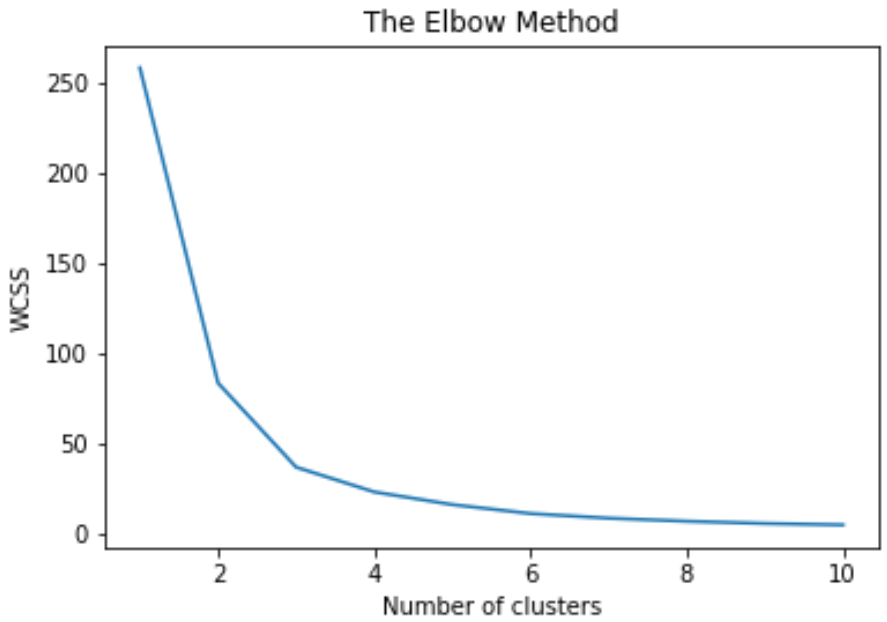
- Dari hasil plotting disamping didapatkan insight pada klustering warna biru yaitu negara yang kematian anak dan jumlah fertiliti berada pada nilai -1 sampai 1 berada pada kluster 1, sedangkan negara yang berada pada kluster 2 yaitu kematian anak dan jumlah fertiliti dari nilai 1 sampai 3
- Didapat insight bahwa semakin tinggi jumlah fertility maka semakin tinggi pula kematian anak, namun beberapa negara tidak mengalami kematian yang tinggi apabila jumlah fertility tinggi, sepertinya terjadi pada negara maju

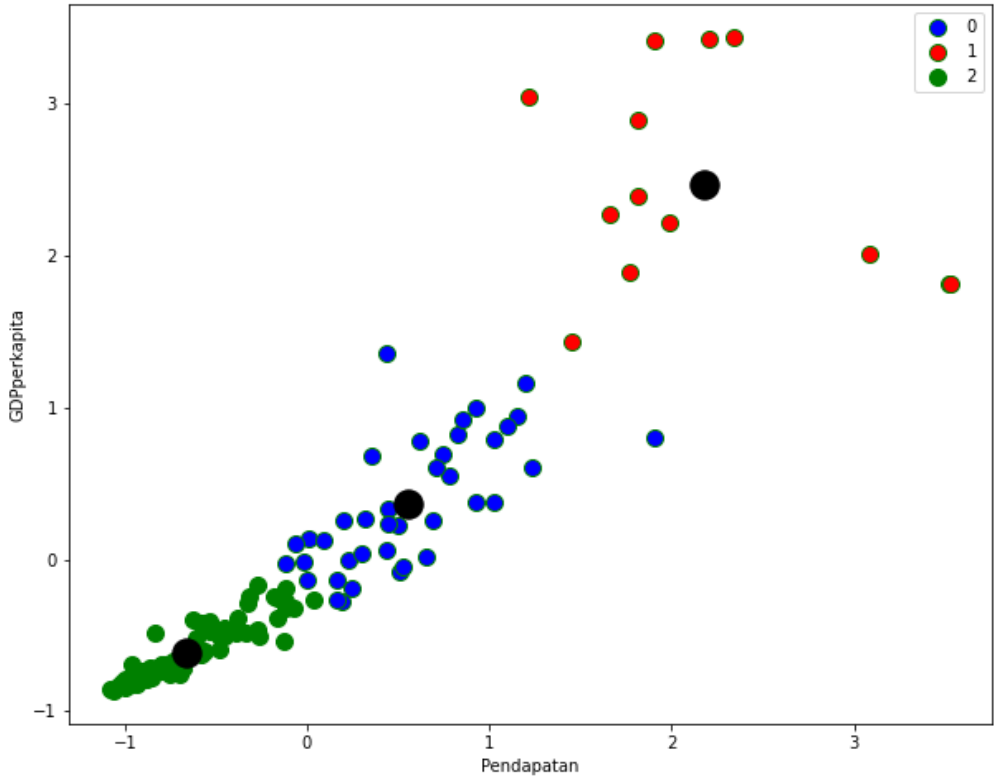


MENENTUKAN JUMLAH CLUSTER DENGAN ELBOW METHOD

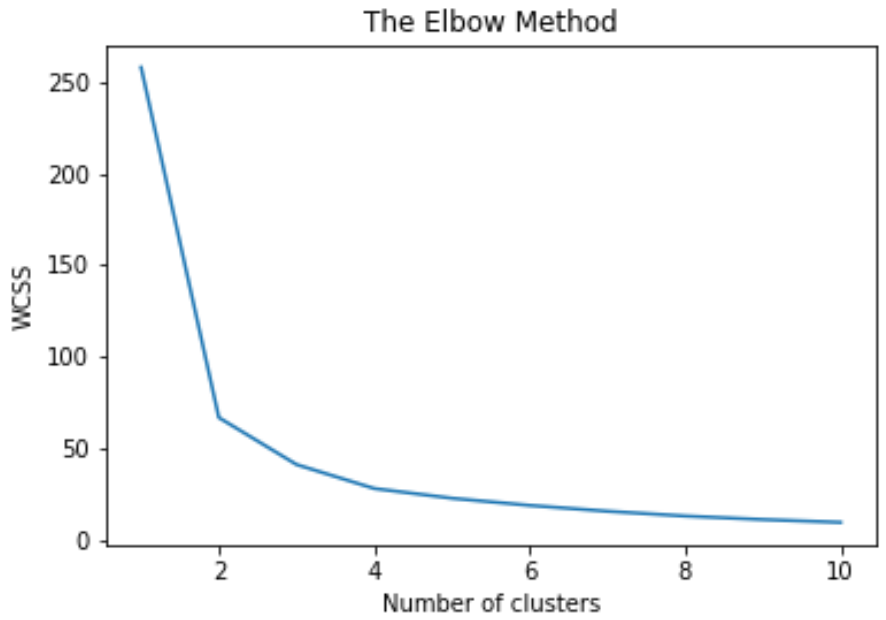
- Pada pemberian jumlah clustering sebelumnya ditentukan dengan jumlah acak yakni berjumlah 2 cluster
- Kali ini jumlah cluster ditentukan menggunakan elbow method, metode ini membandingkan jumlah kluster dengan inersianya.
- Berikut hasil scatter plot setelah menggunakan jumlah cluster hasil dari rekomendasi elbow method:

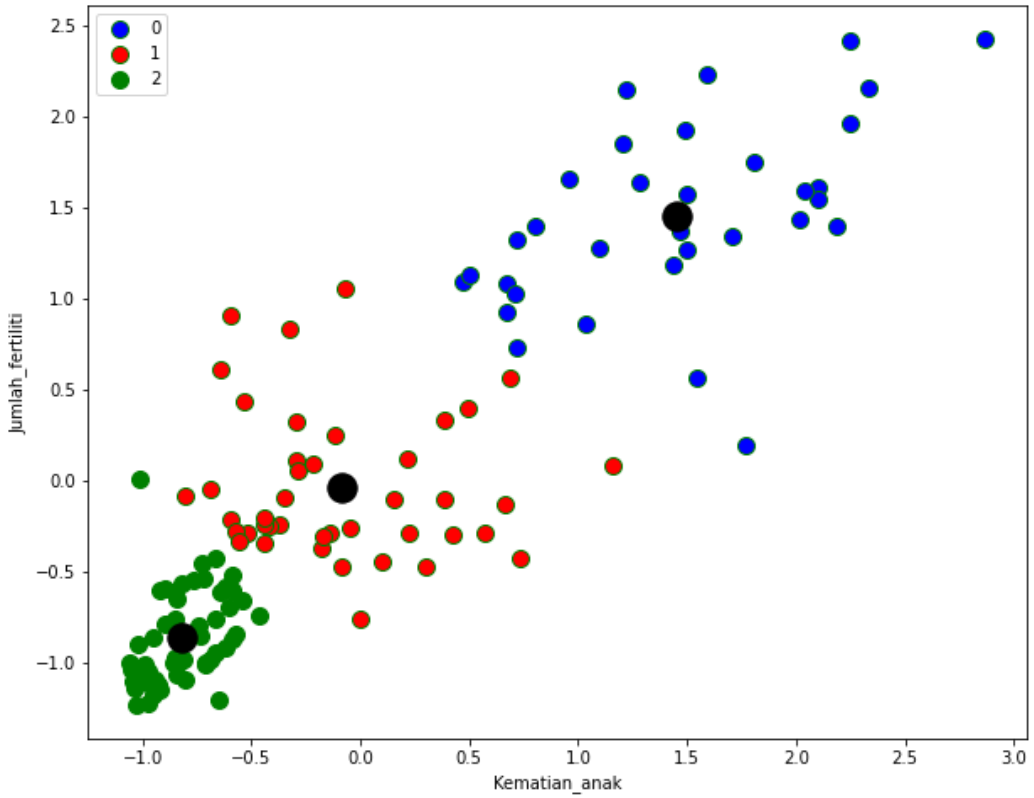
- Dari hasil penerapan elbow method antara kolom pendapatan dan GDP perkapita didapatkan jumlah cluster paling optimal yaitu berjumlah 3 cluster
- Cluster 0 berwarna biru yaitu negara yang berpendapatan antara skor 0 sampai 2
- Cluster 1 berwarna merah yaitu negara yang berpendapatan antara skor 2 sampai lebih dari 3
- Cluster 2 berwarna hijau yaitu negara yang berpendapatan antara skor -1 sampai 0
- Dari hasil clustering menggunakan elbow method diambil cluster 2 yang nantinya diolah lebih lanjut sebagai rekomendasi negara yang butuh bantuan mengacu pada rendahnya pendapatan dan GDP perkapita





- Juga diterapkan elbow method antara kolom jumlah fertiliti dan kematian anak didapat cluster paling optimal yaitu berjumlah 3 cluster
- Cluster 0 berwarna biru yaitu negara yang kematian anak antara skor 1 sampai 3
- Cluster 1 berwarna merah yaitu negara yang kematian anak antara skor -0.5 sampai 1
- Cluster 2 berwarna hijau yaitu negara yang kematian antara skor -0.5 sampai -1.0
- Diambil cluster 0 sebagai acuan semakin tinggi jumlah fertility semakin tinggi kematian anak menunjukkan buruknya Kesehatan di negara tersebut bagi anak





SILHOUETTE SCORE

- Silhouette score merupakan metode untuk menentukan jumlah cluster mana yang bisa optimal dalam memvisualisasikan data
- Nilai silhouette score antara jumlah cluster acak dan hasil dari elbow method dari kedua data yaitu sebagai berikut:
- Pendapatan dan GDP perkapita:

```
0.6499559730398784
0.6041024179853651
```

- Jumlah fertility dan kematian anak:

```
0.654615008086036
0.5095659613375952
```

- Dari nilai silhouette didapat kesimpulan bahwa penentuan jumlah cluster yaitu 2 pada masing-masing perbandingan variable merupakan jumlas kluster paling efektif dalam memvisulisasikan data dibanding menggunakan jumlah cluster dari elbow method
- Namun Ketika menggunakan jumlah cluster menggunakan elbow method data bisa divisualisasikan secara optimal dan memudahkan pembacaan hasil plotting data tersebut

PENAMBAHAN KOLOM BARU DARI DATA HASIL PENGHAPUSAN OUTLIER DENGAN KOLOM LABEL HASIL CLUSTERING

- Data hasil pendropan outlier ditambahkan dengan label klustering, dihasilkan table sebagai berikut:

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	K_means_labels	K_means_labels2
0	Afghanistan	90.2	10.0	7.58	44.9	1610.0	9.44	56.2	5.82	553.0	2	0
1	Albania	16.6	28.0	6.55	48.6	9930.0	4.49	76.3	1.65	4090.0	2	2
2	Algeria	27.3	38.4	4.17	31.4	12900.0	16.10	76.5	2.89	4460.0	0	1
3	Angola	119.0	62.3	2.85	42.9	5900.0	22.40	60.1	6.16	3530.0	2	0
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100.0	1.44	76.8	2.13	12200.0	0	2
...	...	...	...	...	...	...	...	...	...	...	...	...
161	Uzbekistan	36.3	31.7	5.81	28.5	4240.0	16.50	68.8	2.34	1380.0	2	1
162	Vanuatu	29.2	46.6	5.25	52.7	2950.0	2.62	63.0	3.50	2970.0	2	1
164	Vietnam	23.3	72.0	6.84	80.2	4490.0	12.10	73.1	1.95	1310.0	2	2
165	Yemen	56.3	30.0	5.18	34.4	4480.0	23.60	67.5	4.67	1310.0	2	0
166	Zambia	83.1	37.0	5.89	30.9	3280.0	14.00	52.0	5.40	1460.0	2	0

129 rows × 12 columns

FILTERING LABEL CLUSTER YANG DIPILIH

- Dari elbow method sebelumnya diambil cluster 2 dari antara pendapatan dan GDP perkapita, pada kolom jumlah fertility dan kematian anak dipilih cluster 0 sebagai representasi negara miskin banyak fertility namun kematiaan anak tinggi pula
- Berikut hasil table menggunakan filtering dan logika AND:

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	K_means_labels	K_means_labels2
0	Afghanistan	90.2	10.00	7.58	44.9	1610.0	9.440	56.2	5.82	553.0	2	0
3	Angola	119.0	62.30	2.85	42.9	5900.0	22.400	60.1	6.16	3530.0	2	0
17	Benin	111.0	23.80	4.10	37.2	1820.0	0.885	61.8	5.36	758.0	2	0
25	Burkina Faso	116.0	19.20	6.74	29.6	1430.0	6.810	57.9	5.87	575.0	2	0
26	Burundi	93.6	8.92	11.60	39.2	764.0	12.300	57.7	6.26	231.0	2	0
28	Cameroon	108.0	22.20	5.13	27.0	2660.0	1.910	57.3	5.11	1310.0	2	0
36	Comoros	88.2	16.50	4.51	51.7	1410.0	3.870	65.9	4.75	769.0	2	0
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609.0	20.800	57.5	6.54	334.0	2	0
38	Congo, Rep.	63.9	85.10	2.46	54.7	5190.0	20.700	60.4	4.95	2740.0	2	0
40	Cote d'Ivoire	111.0	50.60	5.30	43.3	2690.0	5.390	56.3	5.27	1220.0	2	0

SORTING DATA HASIL FILTERING

- Setelah melakukan filtering, Langkah terakhir yaitu melakukan sorting data dari yang terendah menggunakan sort.values
- Sorting didasarkan dari nilai pendapatan dan diurutkan dari yang terendah ke tinggi
- Sorting data dipilih dari 10 data teratas
- Berikut hasil sorting data:

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	K_means_labels	K_means_labels2
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609.0	20.80	57.5	6.54	334.0	2	0
88	Liberia	89.3	19.10	11.80	92.6	700.0	5.47	60.8	5.02	327.0	2	0
26	Burundi	93.6	8.92	11.60	39.2	764.0	12.30	57.7	6.26	231.0	2	0
106	Mozambique	101.0	31.50	5.21	46.2	918.0	7.64	54.5	5.56	419.0	2	0
94	Malawi	90.5	22.80	6.59	34.9	1030.0	12.10	53.1	5.31	459.0	2	0
63	Guinea	109.0	30.30	4.93	43.2	1190.0	16.10	58.0	5.34	648.0	2	0
150	Togo	90.3	40.20	7.65	57.3	1210.0	1.18	58.7	4.87	488.0	2	0
126	Rwanda	63.6	12.00	10.50	30.0	1350.0	2.61	64.6	4.51	563.0	2	0
93	Madagascar	62.2	25.00	3.77	43.0	1390.0	8.79	60.8	4.60	413.0	2	0
64	Guinea-Bissau	114.0	14.90	8.50	35.2	1390.0	2.97	55.6	5.05	547.0	2	0

KESIMPULAN

- Dari hasil tahap terakhir yaitu sorting data, didapatkan 10 data negara teratas yang bisa dijadikan rekomendasi untuk CEO HELP International untuk memprioritaskan pada 10 negara tersebut yakni:

1. Rep. demokratik congo
2. Liberia
3. Burundi
4. Mozambique
5. Malawi
6. guinea
7. togo
8. Rwanda
9. Madagaskar
10. Guinea-Bissau