Final Project Report
Data Science Clustering
SanberCode

by : Muhammad Aziz Pratama

date: April 2021

Daftar Isi

Pengantar

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 dana dan Saat ini, CEO LSM perlu memutuskan bagaimana menggunakan uang ini secara strategis dan efektif.

Permasalahan dan Tujuan

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, perlu untuk mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Kemudian kita perlu menyarankan negara mana saja yang paling perlu menjadi fokus CEO.

Reading and Understanding Data

Telah disediakan dataset Data_Negara_HELP.csv yang berisi tentang data-data negara yang telah dikumpulkan oleh perusahaan Help International. Pada dataset ini terdapat beberapa kolom feature, yaitu :

- **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.

Info data

Dataset ini formatnya merupakan dataframe dengan data sebanyak 167 baris yang berbeda-beda untuk tiap negaranya.

```
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)
```

- Feature Negara merupakan feature yang bertype object karena feature ini berisi data string nama negara
- Feature kematian_anak, ekspor, Kesehatan, impor, inflasi, harapan_hidup, jumlah_fertiliti merupakan data yang bertype float
- Sedangkan feature pendapatan dan GDPperkapita merupakan data bertype int

Deskripsi Data

	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

Dapat kita liat pada deskripsi data diatas, tersedia banyak data, nilai rata-rata, standar deviasi, nilai minimum dan maksimum serta quantile tiap data per-feature. Pada tiap feature tidak ada keanehan atau memiliki anomaly data yang tidak sesuai

```
Negara              0
Kematian_anak       0
Ekspor              0
Kesehatan           0
Impor               0
Pendapatan          0
Inflasi             0
Harapan_hidup       0
Jumlah_fertiliti    0
GDPperkapita        0
```

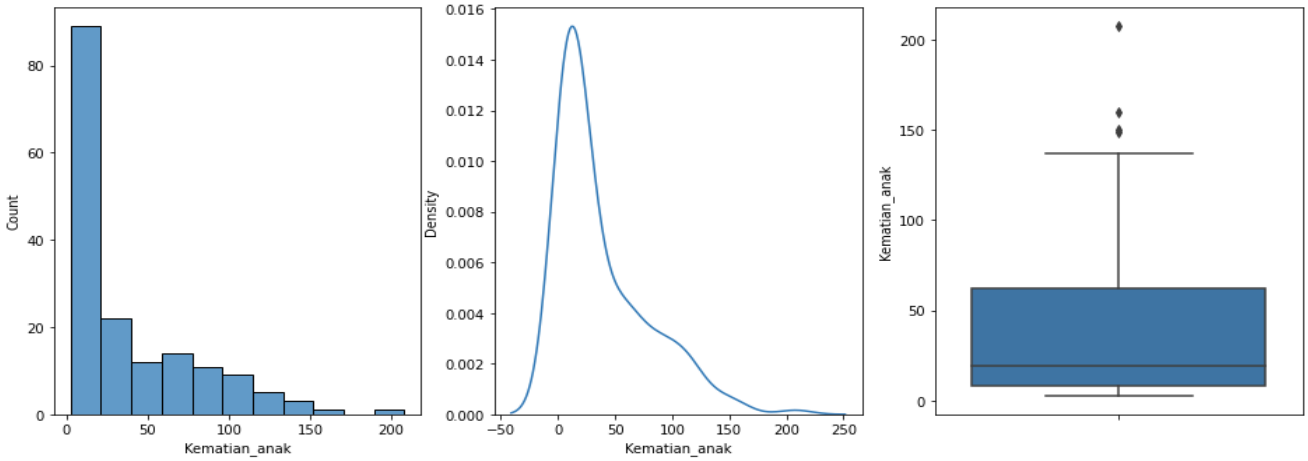
Pada dataset ini ditiap featurenya tidak ada yang memiliki data Nan atau null.

Exploratory Data Analysis

Univariate Analysis

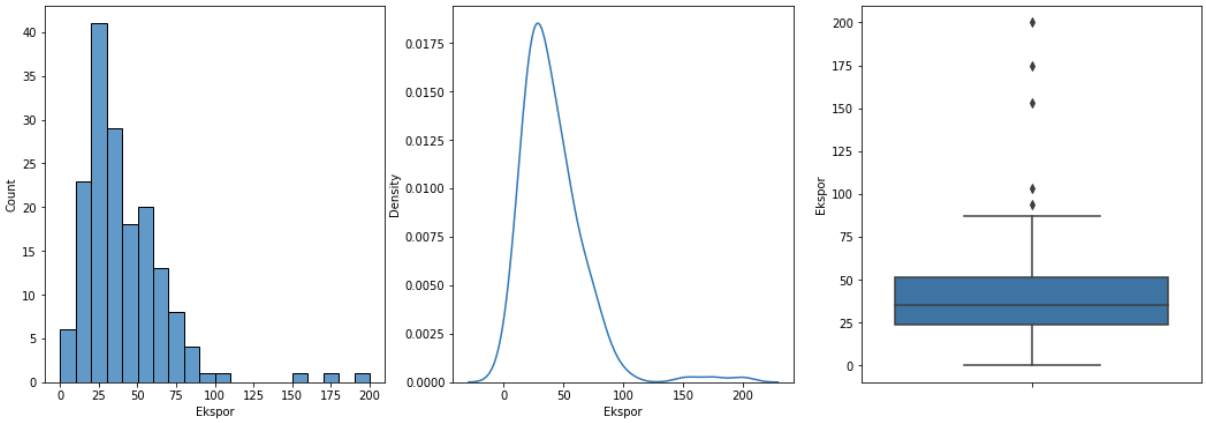
a. Feature Kematian_anak

Pada feature ini, angka kematian anak dari dataset negara banyak terdapat pada sebelum 50 dan pada data ini terdapat 3 data outlier diatas upperbound.



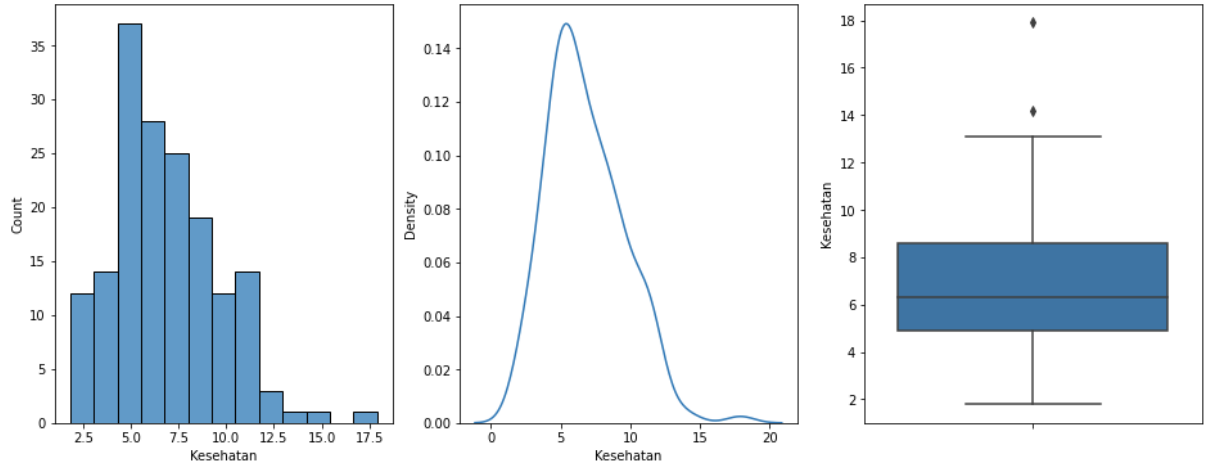
b. Feature Ekspor

Pada feature ini, angka ekspor dari dataset negara banyak terdapat pada sebelum 40 dan pada data ini terdapat 5 data outlier diatas upperbound.



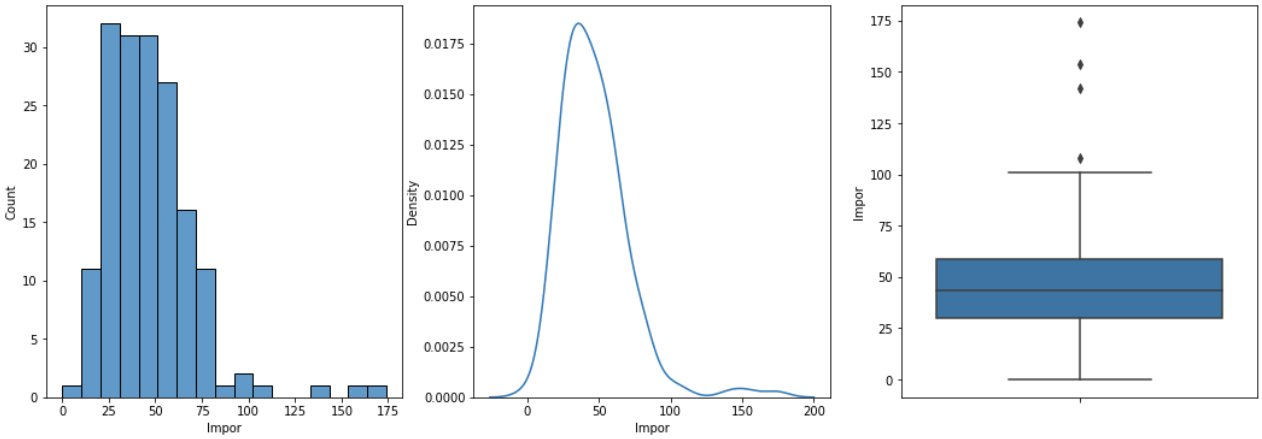
c. Feature Kesehatan

Pada feature ini, angka kesehatan dari dataset negara banyak terdapat pada angka sebelum 10 dan pada data ini terdapat 2 data outlier diatas upperbound.



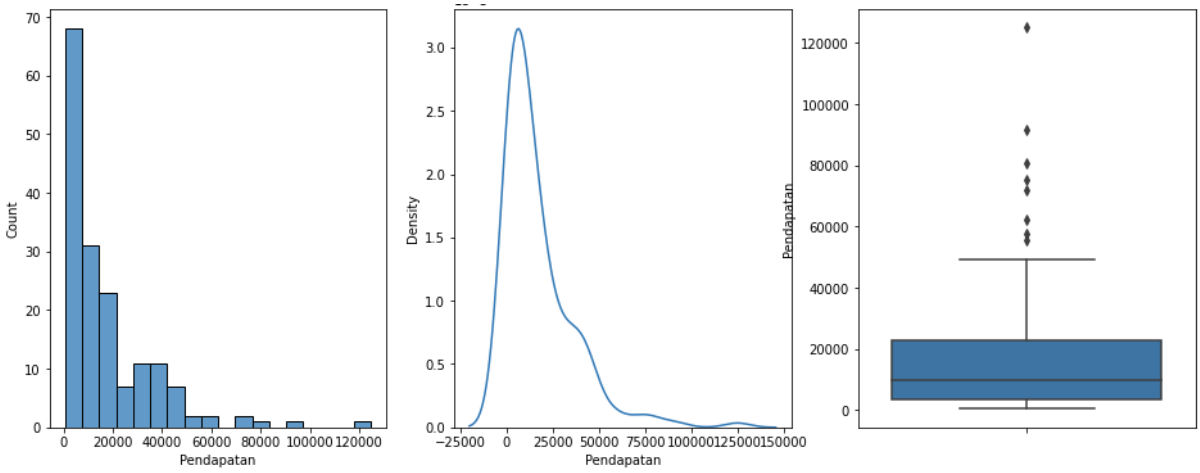
d. Feature Impor

Pada feature ini, angka impor dari dataset negara banyak terdapat pada angka sebelum 75 dan pada data ini terdapat 4 data outlier diatas upperbound.



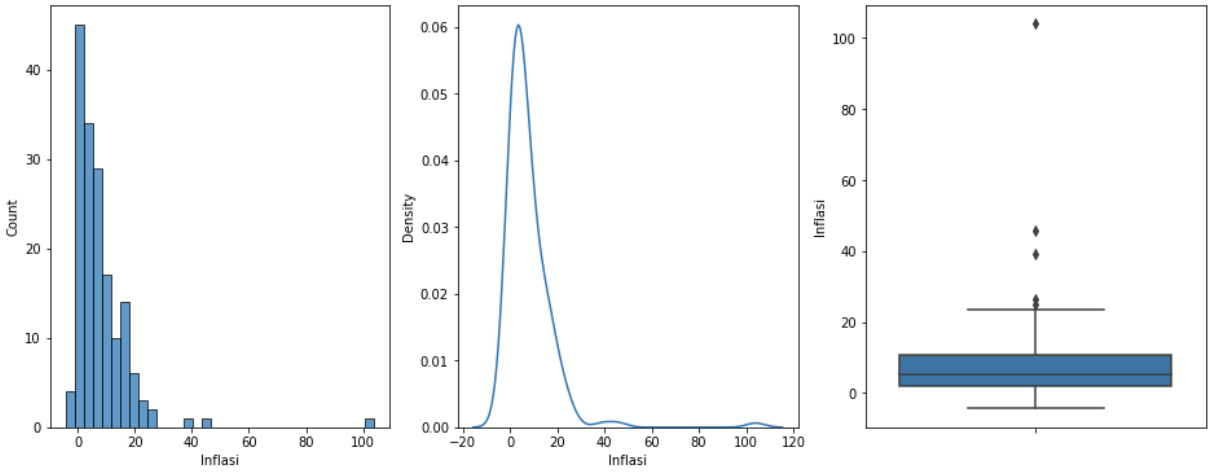
e. Feature Pendapatan

Pada feature ini, angka pendapatan dari dataset negara banyak terdapat pada angka sebelum 20.000 dan pada data ini terdapat 8 data outlier diatas upperbound.



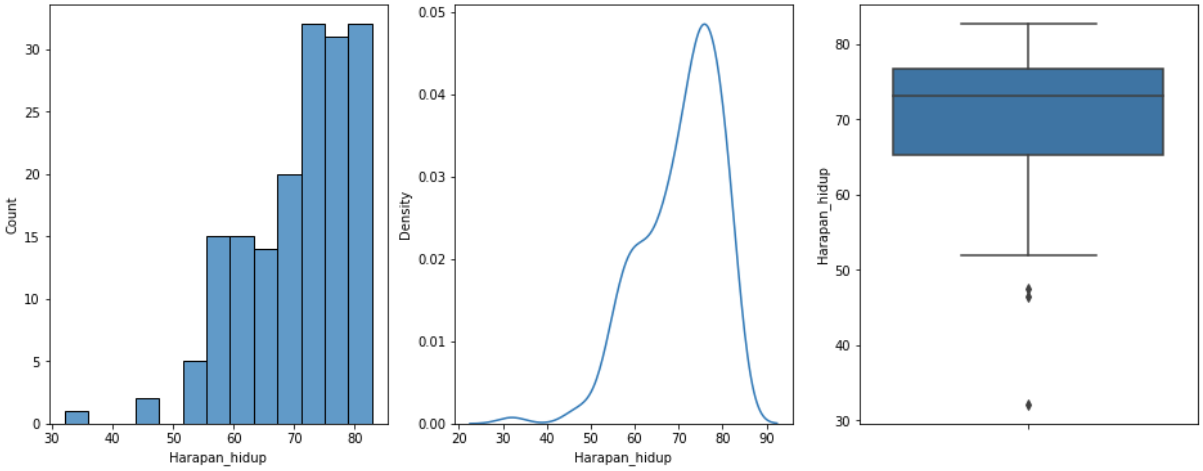
f. Feature Inflasi

Pada feature ini, angka inflasi dari dataset negara banyak terdapat pada angka sebelum 20 dan pada data ini terdapat 5 data outlier diatas upperbound.



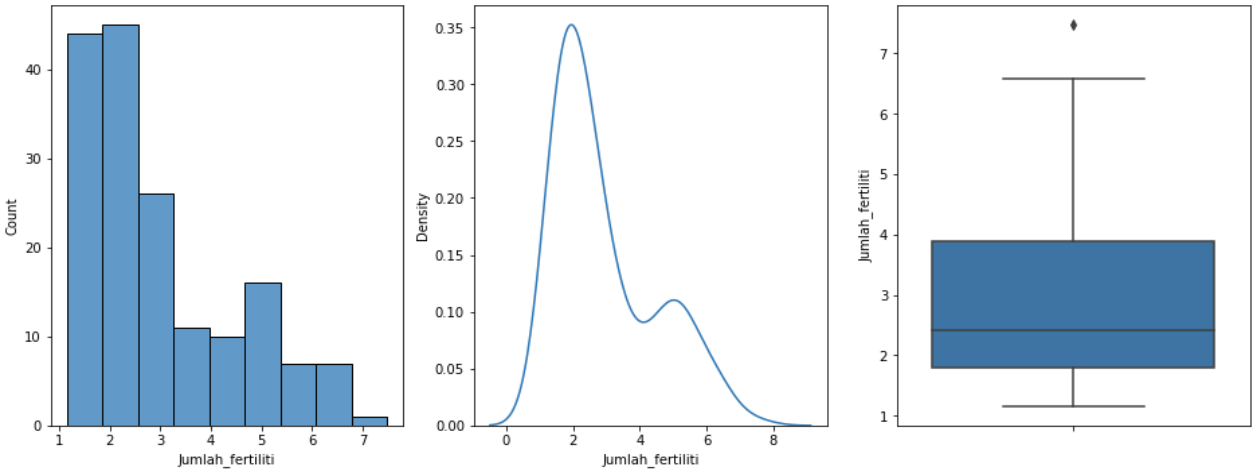
g. Feature Harapan_hidup

Pada feature ini, angka harapan hidup dari dataset negara banyak terdapat pada angka lebih dari 60 dan pada data ini terdapat 3 data outlier dibawah lowerbound.



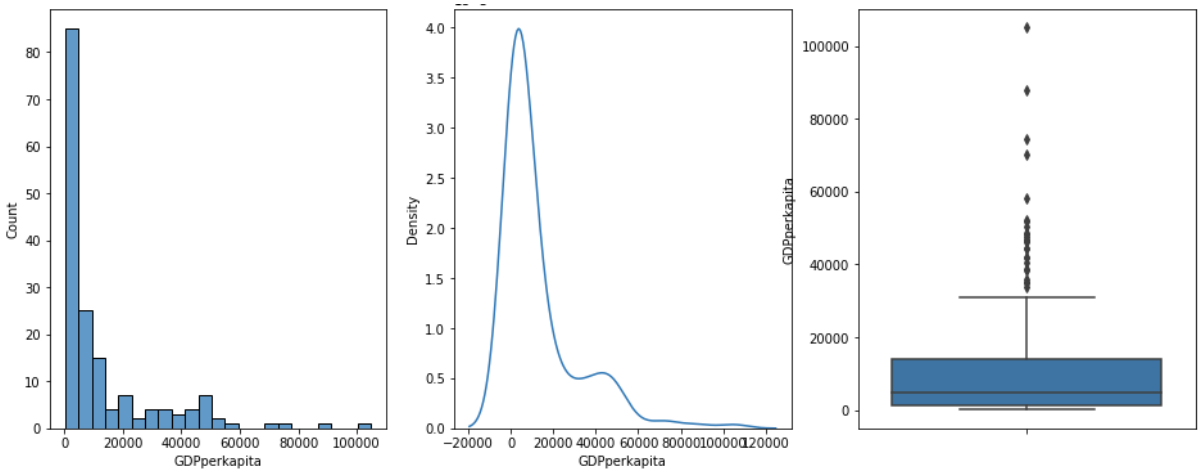
h. Feature Jumlah_fertiliti

Pada feature ini, angka jumlah fertiliti tiap negara banyak terdapat pada angka sebelum 4 dan pada data ini terdapat 1 data outlier diatas upperbound.



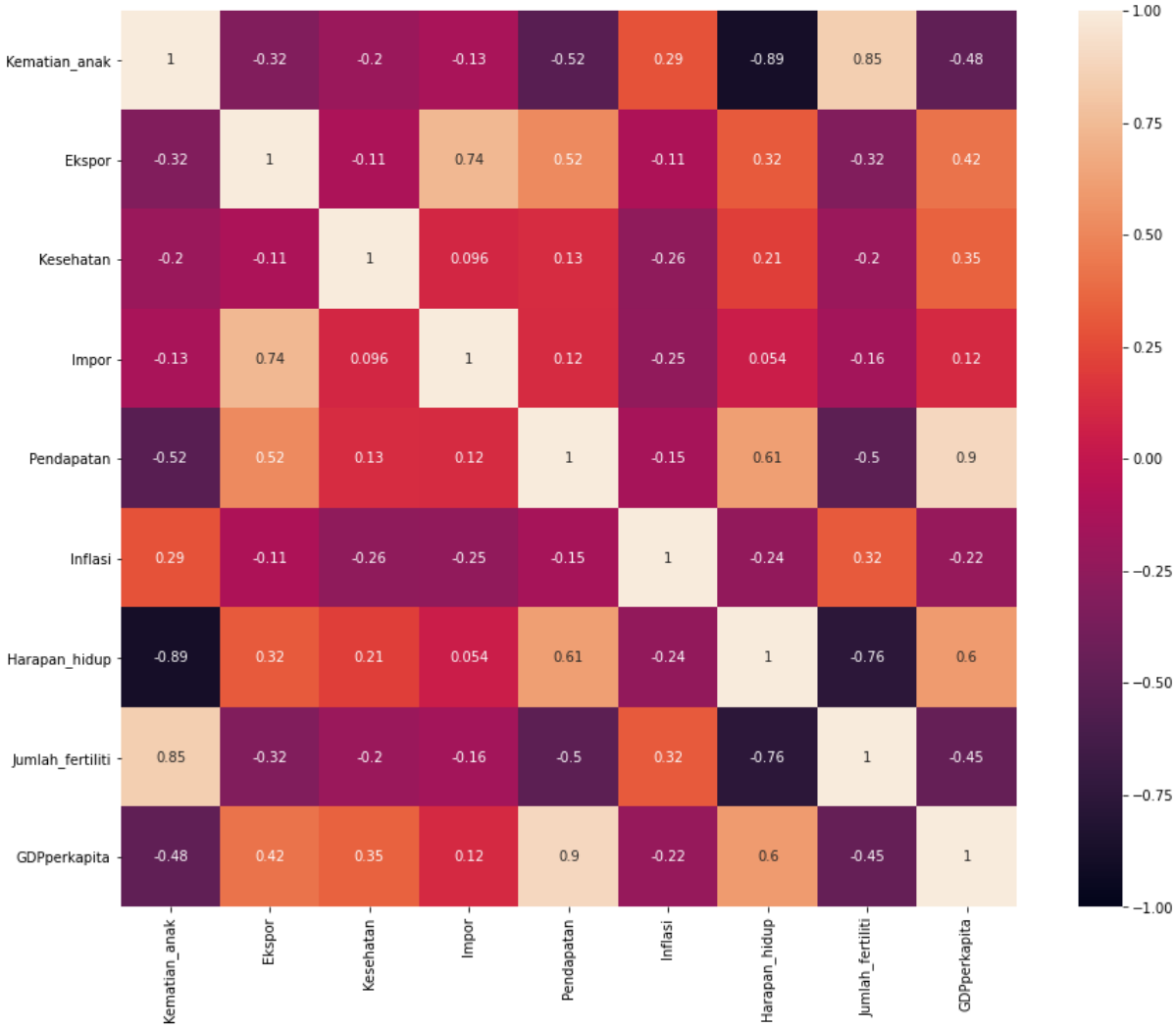
i. Feature GDPperkapita

Pada feature ini, angka gdp perkapita tiap negara banyak terdapat pada angka sebelum 30.000 dan pada data ini terdapat lumayan banyak data outlier diatas upperbound.



Bivariate Analysis

a) Heatmaps

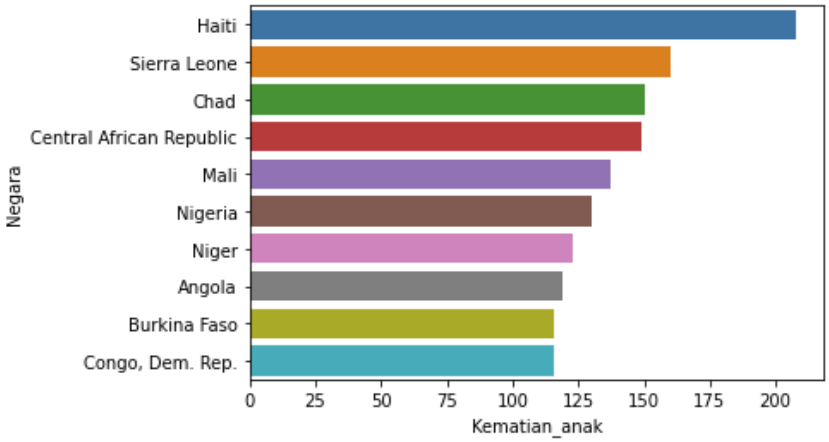


dari heatmap ini, dapat kita lihat hubungan dari feature negara bahwa :

- Korelasi antara pendapatan dan gdp perkapita sangat kuat, semakin tinggi angka pendapatan maka akan semakin tinggi angka gdp perkapita.
- Semakin tinggi angka jumlah_fertiliti, maka akan semakin tinggi angka kematian_anak.
- Semakin tinggi angka jumlah_fertiliti, maka angka semakin kecil angka harapan_hidup.
- Semakin tinggi angka harapan_hidup, maka makin rendah angka kematian_anak.
- Semakin tinggi angka impor, maka makin tinggi angka ekspor-nya
- Makin banyak angka pendapatan, angka ekspor juga tinggi.
- Semakin tinggi angka pendapatan maka angka harapan_hidup juga makin tinggi.

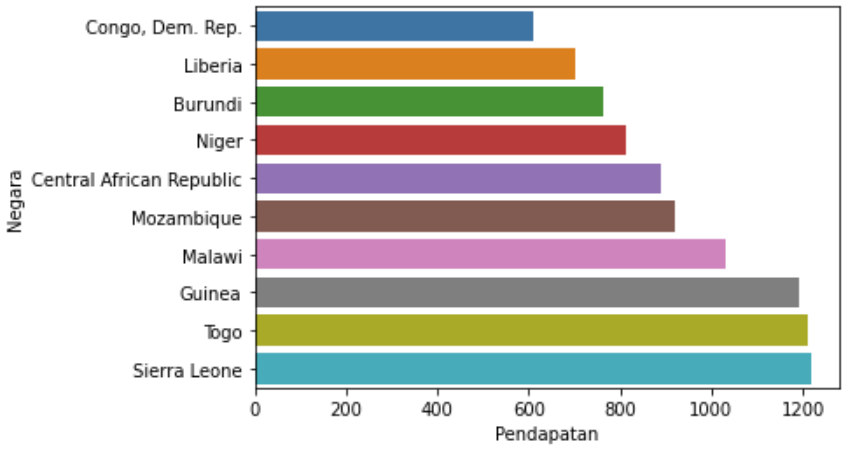
b) Barplot

- **Feature Kematian_anak**



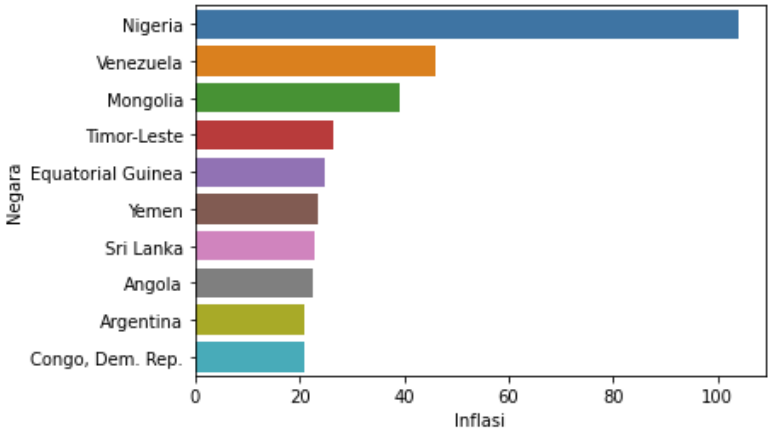
Barplot diatas, menunjukkan 10 dari 167 negara dengan angka kematian_anak paling tinggi dimana Haiti merupakan negara dengan angka kematian_anak tertinggi.

- **Feature Pendapatan**



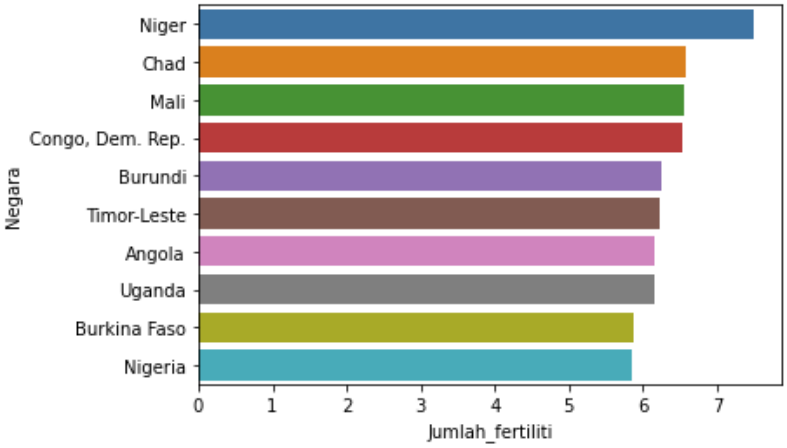
Barplot diatas, menunjukkan 10 dari 167 negara dengan angka pendapatan paling rendah dimana Congo diikuti Liberia, Burundi merupakan negara dengan angka pendapatan ter-rendah.

- **Feature Inflasi**



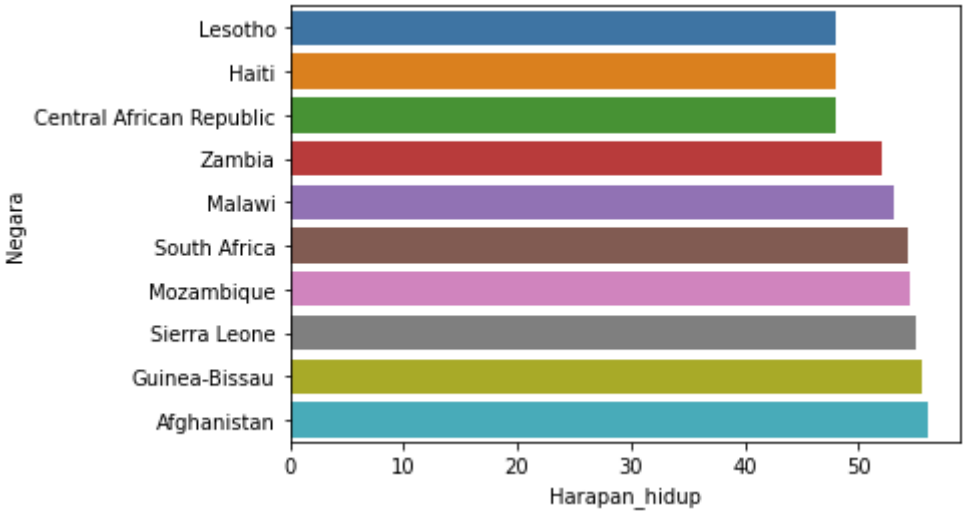
Barplot diatas, menunjukkan 10 dari 167 negara dengan angka inflasi paling tinggi yang menyebabkan penurunan nilai mata uang dimana Nigeria merupakan negara dengan angka inflasi tertinggi.

- **Feature Jumlah_fertiliti**



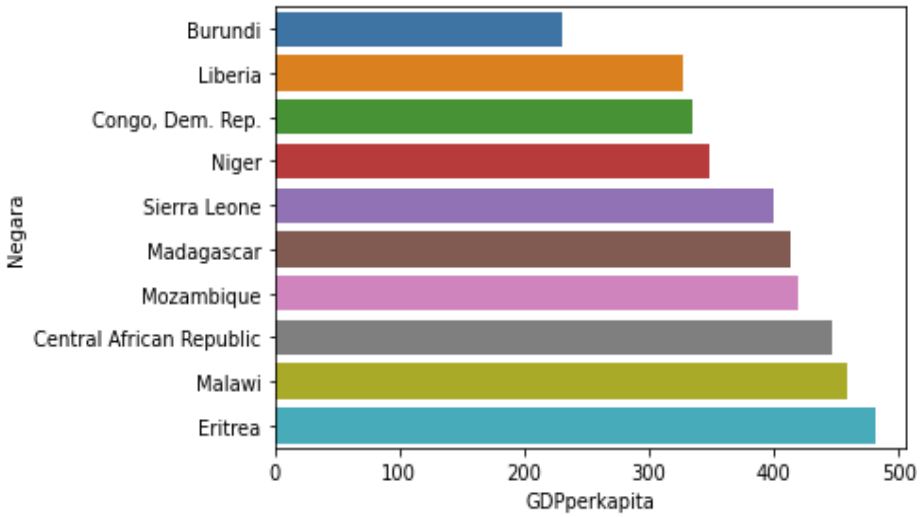
Barplot diatas, menunjukkan 10 dari 167 negara dengan angka jumlah_fertiliti paling tinggi dimana Niger merupakan negara dengan angka jumlah_fertiliti tertinggi.

- **Feature Harapan_hidup**



Barplot diatas, menunjukkan 10 dari 167 negara dengan angka Harapan_hidup paling rendah dimana Lesotho, Haiti, Central African Republic merupakan negara dengan angka Harapan_hidup terendah.

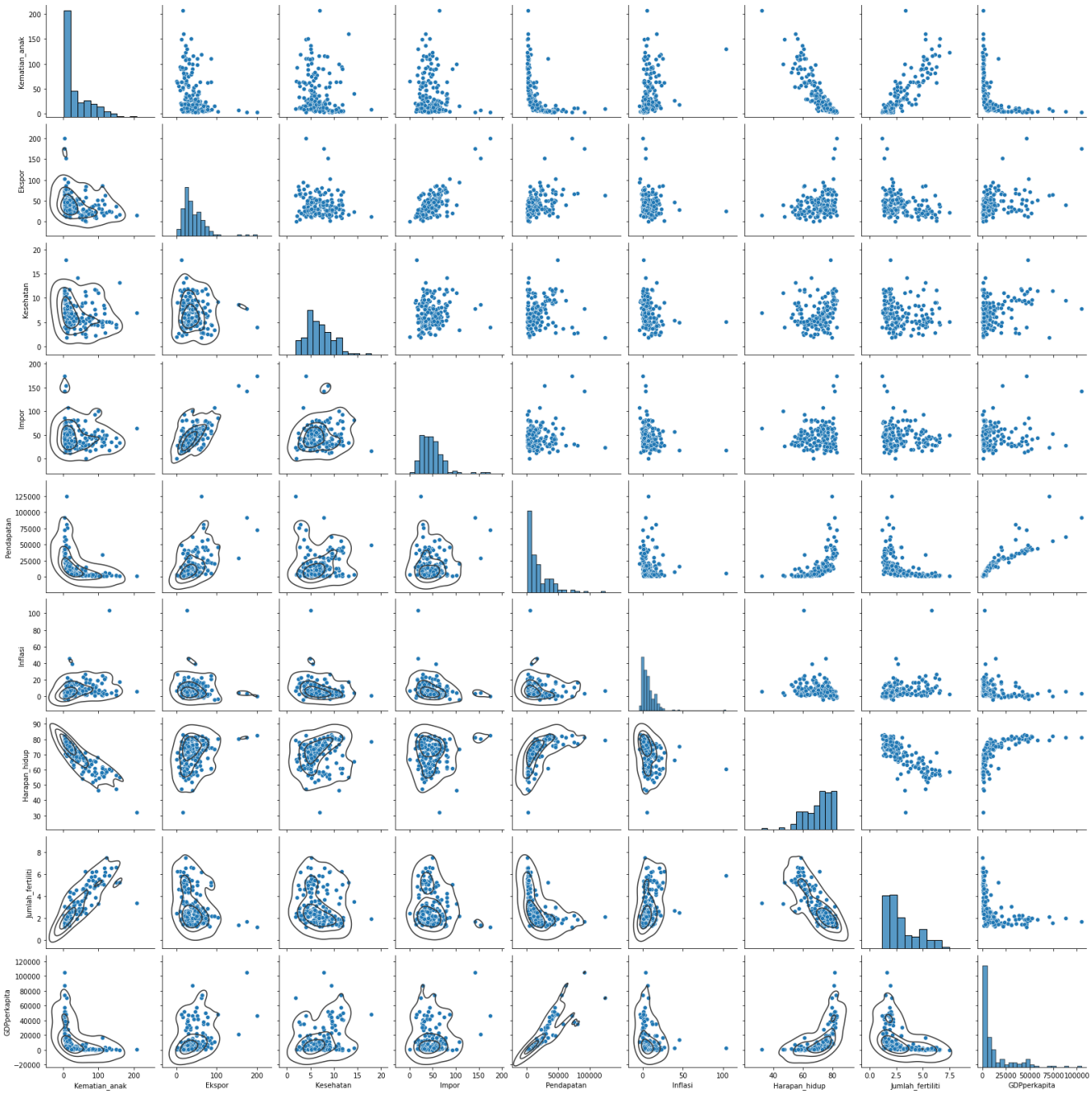
- **Feature GDPperkapita**



Barplot diatas, menunjukkan 10 dari 167 negara dengan angka GDP perkapita paling rendah dimana Burundi merupakan negara dengan angka GDP perkapita yang paling rendah.

Multivariate

- **Pairplot**

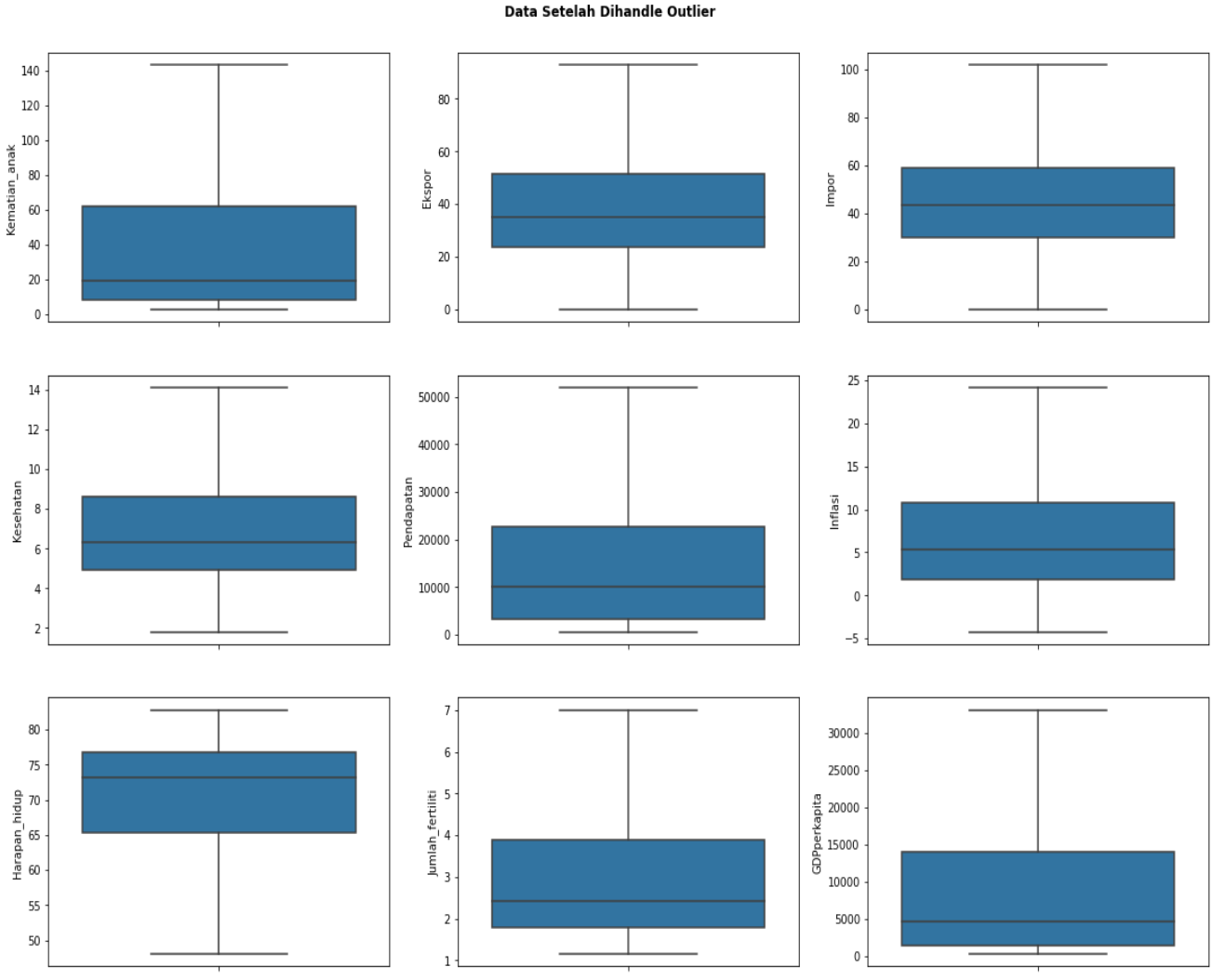


Untuk pairplot ini sendiri, insight yang dapat kita ambil dapat sama seperti heatmaps yang telah dijelaskan sebelumnya, hanya saja disini kita dapat melihat titik dari persebaran datanya.

Outlier Treatment

Dari analysis data per-feature sebelumnya, pada setiap kolom feature terdapat data pencilan atau outlier. Untuk meng-handle data outlier ini saya pun mereplace data outlier tersebut dengan data upperbound atau lowerbound dari feature tersebut. Apabila data outlier terdapat diatas upperbound, maka akan direplace dengan nilai upperbound dan apabila data outlier terdapat dibawah lowerbound, maka akan direplace dengan nilai lowerbound.

Data yang dihandle outlier nya merupakan feature yang bertipe float atau int, sehingga kecuali feature Negara yang berisi nama negara, 9 feature lainnya dilakukan handle outlier.



Scaling Data

Scaling data dilakukan menggunakan MinMax Scaling dari library sklearn. Scaling data ini berguna untuk mempercepat proses perhitungan pada kmeans dan membuat range nilai menjadi sama yaitu di rentang [0,1]. Scaling ini dilakukan pada feature bertipe float dan int. berikut data dari hasil MinMax Scaling :

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	0.624488	0.106853	0.468725	0.441676	0.019490	0.481142	0.234532	0.797268	0.009788
1	0.099804	0.301309	0.385053	0.478126	0.181489	0.306662	0.812950	0.085361	0.117298
2	0.176083	0.413662	0.191714	0.308683	0.239318	0.715897	0.818705	0.297055	0.128545
3	0.829799	0.671856	0.084484	0.421973	0.103021	0.937963	0.346763	0.855314	0.100277
4	0.054892	0.490364	0.342811	0.579594	0.360038	0.199154	0.827338	0.167307	0.363810

Clustering Using Kmeans

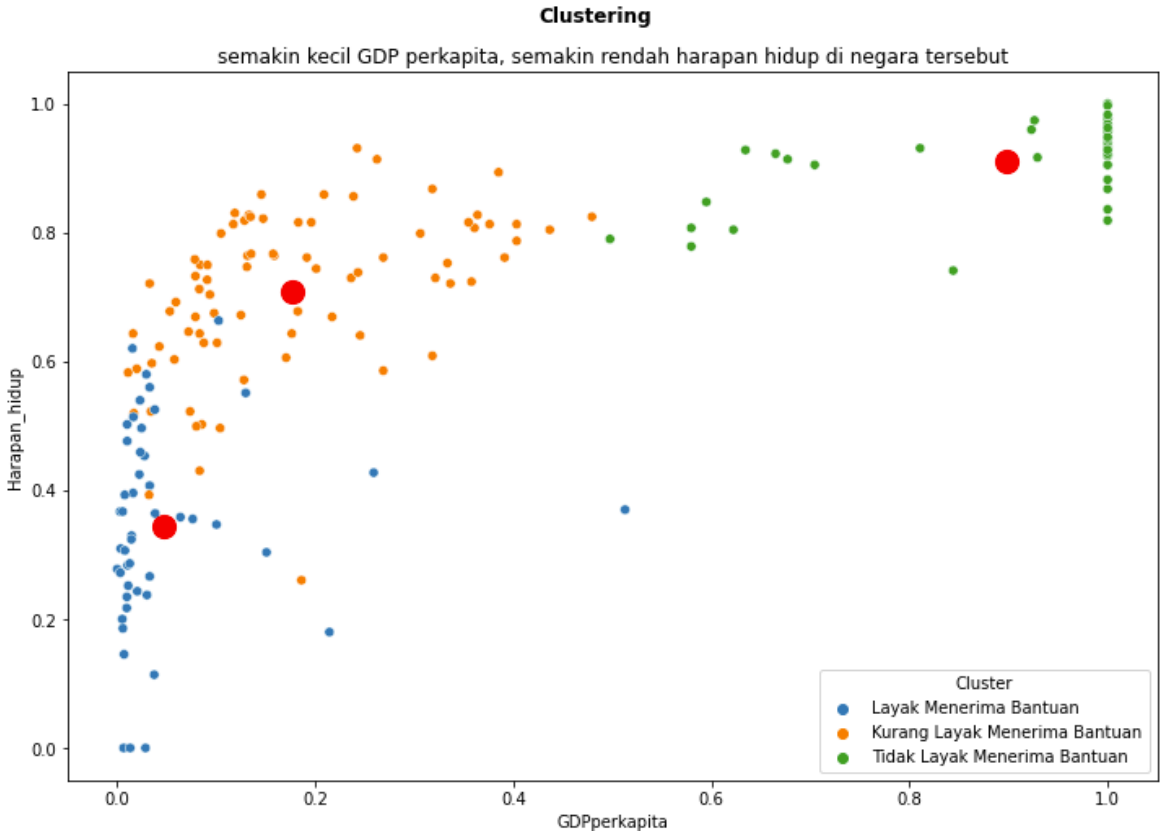
Kmeans digunakan untuk mengkelompokkan negara berdasarkan cluster. Hasil dari clustering Kmeans ini adalah kumpulan negara-negara yang telah dikelompokkan berdasarkan yang berhak menerima bantuan Help International. Cara kerja Kmeans ini mengumpulkan data yang jaraknya dekat atau mirip menjadi satu cluster. Untuk mendapatkan banyak cluster (K) yang efektif, kita dapat menentukannya dengan Elbow method.

Eksperimen Clustering

Sebelum melakukan clustering, ada beberapa parameter yang saya tetapkan terlebih dahulu, yaitu:

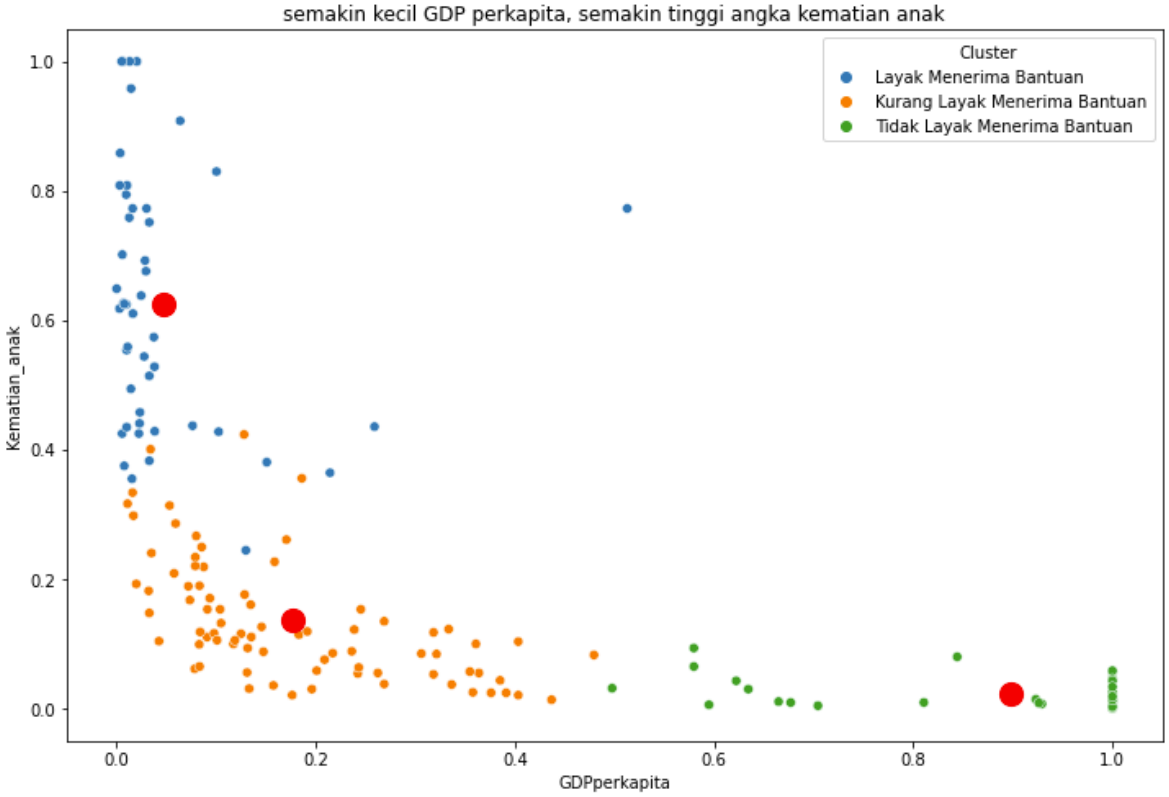
1. n_cluster = 3
2. random_state = 100

Hasil dari clustering adalah sebagai berikut :



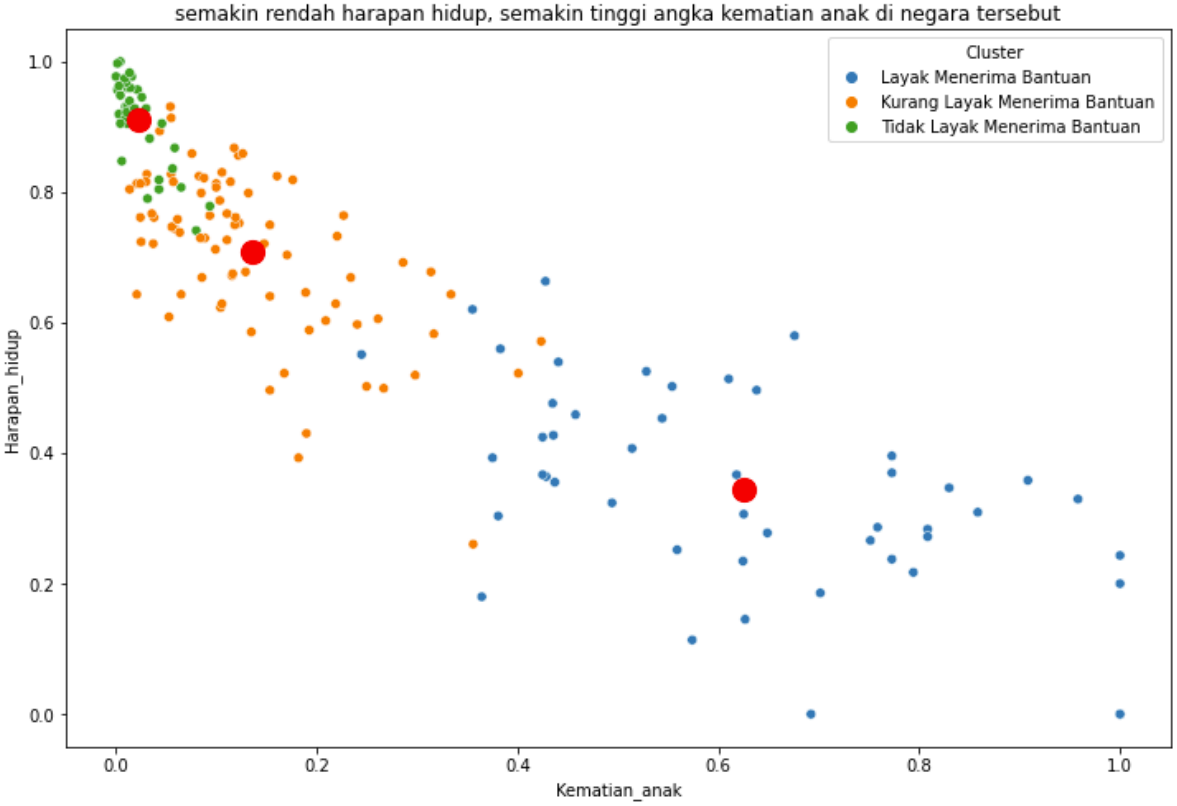
Clustering

semakin kecil GDP perkapita, semakin tinggi angka kematian anak



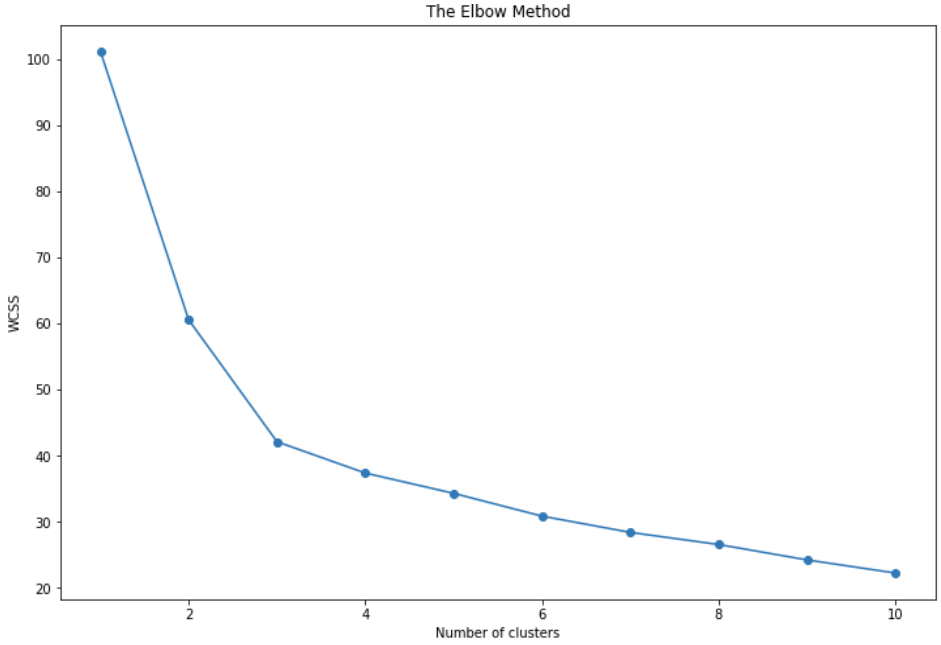
Clustering

semakin rendah harapan hidup, semakin tinggi angka kematian anak di negara tersebut



Elbow Method

Untuk mengevaluasi Kembali model kmeans yang telah digunakan, saya memakai elbow method untuk mengetahui K berapa yang efektif untuk dataset ini. Hasilnya adalah K = 3 adalah yang efektif untuk dataset ini.



Insight dan Kesimpulan

Dari hasil pemahaman data dan clustering yang telah dilakukan, maka dapat kita ambil insight sebagai berikut:

a) GDP perkapita berpengaruh dalam baik buruk nya kondisi ekonomi dan Kesehatan dalam suatu negara
b) Angka kematian anak yang tinggi berkaitan dengan angka harapan hidup rendah yang dipengaruh juga oleh ekonomi yang kurang baik.
c) Hasil dari clustering menunjukkan, negara yang layak diberikan bantuan adalah negara yang memiliki GDP perkapita yang kecil dan harapan hidup yang rendah
d) Hasil dari clustering menunjukkan, negara yang layak diberikan bantuan adalah negara yang memiliki GDP perkapita yang kecil dan kematian anak yang tinggi
e) Hasil dari clustering menunjukkan, negara yang layak diberikan bantuan adalah negara yang memiliki harapan hidup yang rendah dan kematian anak yang tinggi

Rekomendasi Negara Penerima Bantuan

Setelah Clustering, data negara telah dikelompokkan dan selanjutnya negara disortir berdasarkan GDP perkapita terendah atau kematian anak yang tinggi dan dapat menjadi rekomendasi negara yang dibantu oleh HELP International.

a. Rekomendasi 10 negara berdasarkan GDP terendah karena ekonomi yang kurang baik dan mungkin dapat memengaruhi bidang Kesehatan.

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	Cluster
26	Burundi	93.600	8.92	11.60	39.2	764.0	12.30	57.70	6.2600	231.0	Layak Menerima Bantuan
88	Liberia	89.300	19.10	11.80	92.6	700.0	5.47	60.80	5.0200	327.0	Layak Menerima Bantuan
37	Congo, Dem. Rep.	116.000	41.10	7.91	49.6	609.0	20.80	57.50	6.5400	334.0	Layak Menerima Bantuan
112	Niger	123.000	22.20	5.16	49.1	814.0	2.55	58.80	7.0075	348.0	Layak Menerima Bantuan
132	Sierra Leone	142.875	16.80	13.10	34.5	1220.0	17.20	55.00	5.2000	399.0	Layak Menerima Bantuan
93	Madagascar	62.200	25.00	3.77	43.0	1390.0	8.79	60.80	4.6000	413.0	Layak Menerima Bantuan
106	Mozambique	101.000	31.50	5.21	46.2	918.0	7.64	54.50	5.5600	419.0	Layak Menerima Bantuan
31	Central African Republic	142.875	11.80	3.98	26.5	888.0	2.01	48.05	5.2100	446.0	Layak Menerima Bantuan
94	Malawi	90.500	22.80	6.59	34.9	1030.0	12.10	53.10	5.3100	459.0	Layak Menerima Bantuan
50	Eritrea	55.200	4.79	2.66	23.3	1420.0	11.60	61.70	4.6100	482.0	Layak Menerima Bantuan

b. Rekomendasi 10 negara berdasarkan kematian anak yang tinggi karena bidang Kesehatan yang kurang baik maupun karena ekonomi kurang baik.

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	Cluster
66	Haiti	142.875	15.3	6.91	64.7	1500.0	5.45	48.05	3.3300	662.0	Layak Menerima Bantuan
31	Central African Republic	142.875	11.8	3.98	26.5	888.0	2.01	48.05	5.2100	446.0	Layak Menerima Bantuan
32	Chad	142.875	36.8	4.53	43.5	1930.0	6.39	56.50	6.5900	897.0	Layak Menerima Bantuan
132	Sierra Leone	142.875	16.8	13.10	34.5	1220.0	17.20	55.00	5.2000	399.0	Layak Menerima Bantuan
97	Mali	137.000	22.8	4.98	35.1	1870.0	4.37	59.50	6.5500	708.0	Layak Menerima Bantuan
113	Nigeria	130.000	25.3	5.07	17.4	5150.0	24.16	60.50	5.8400	2330.0	Layak Menerima Bantuan
112	Niger	123.000	22.2	5.16	49.1	814.0	2.55	58.80	7.0075	348.0	Layak Menerima Bantuan
3	Angola	119.000	62.3	2.85	42.9	5900.0	22.40	60.10	6.1600	3530.0	Layak Menerima Bantuan
25	Burkina Faso	116.000	19.2	6.74	29.6	1430.0	6.81	57.90	5.8700	575.0	Layak Menerima Bantuan
37	Congo, Dem. Rep.	116.000	41.1	7.91	49.6	609.0	20.80	57.50	6.5400	334.0	Layak Menerima Bantuan