Nama: Febrian Aji Pangestu

Clustering the Countries by using K-Means for HELP International

Objective:

Untuk mengkategorikan negara menggunakan factor sosial ekonomi dan Kesehatan yang menentukan pembanganunan negara secara keseluruhan.

Tentang 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.

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.

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.

Metodologi

- Sebagai permulaan, dilakukan import library yang dibutuhkan dan import dataset
- Melakukan business understanding terhadap data yang ada
- Melakukan data preprocessing untuk melihat karakteristiknya, kemudian dibersihkan, dan diolah sebelum dilanjutkan ke proses berikutnya
- Modelling data menggunakan K Means clustering
- Melakukan analisis terhadap hasil clustering

Data Collection

Langkah awal yang dilakukan yaitu dengan mengimport datanya kedalam python seperti pada gambar berikut:

Data Collection

```
df = pd.read_csv('Data_Negara_HELP.csv')
df
```

	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

Gambar 1. Dataset

Data Preprocessing

1. Data Profilling

Data profiling merupakan adalah proses memeriksa data yang tersedia dari sumber informasi yang ada dan mengumpulkan statistik atau ringkasan informatif tentang data itu. Disini saya menggunakan shape, info dan describe untuk mengetahui karakteristik dari dataset yang digunakan.

```
 #   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
```

```
df.describe()
```

	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

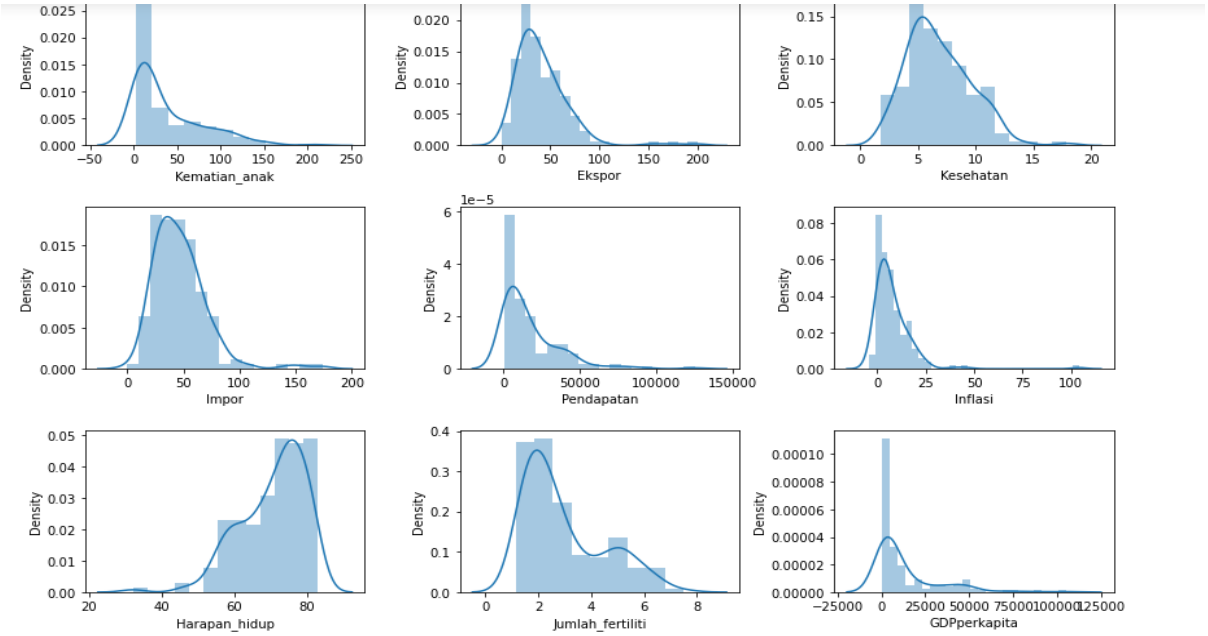
Gambar 2. Data Profilling

Berikut gambar dari data profiling. Untuk next step seharusnya dilakukan data cleaning, tetapi karena tidak ada missing values, maka proses ini di skip dan dilanjutkan ke proses Eksploratory Data Analysis atau biasa disingkat EDA

2. EDA

2.1 Univariate Analysis

Dengan menggunakan displot, dilakukan univariate analysis untuk melihat distribusi data dari setiap variable/column dan grafik skew nya.



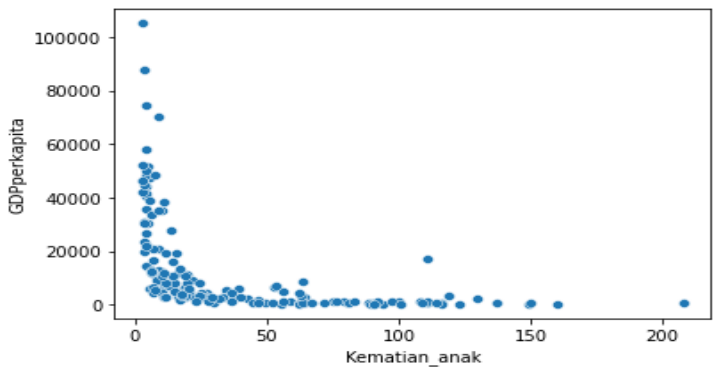
Gambar 3. Distribusi data tiap variabel

Dilihat dari gambar berikut, hanya variable Harapan_hidup yang skew ke kiri, sedangkan sisanya skew ke kanan

2.2 Bivariate Analysis

Pada bivariate analysis, saya mengamati variable Kematian_anak dan GDPperkapita untuk memperoleh nilai relationship antar kedua variable tersebut. Berikut gambarnya

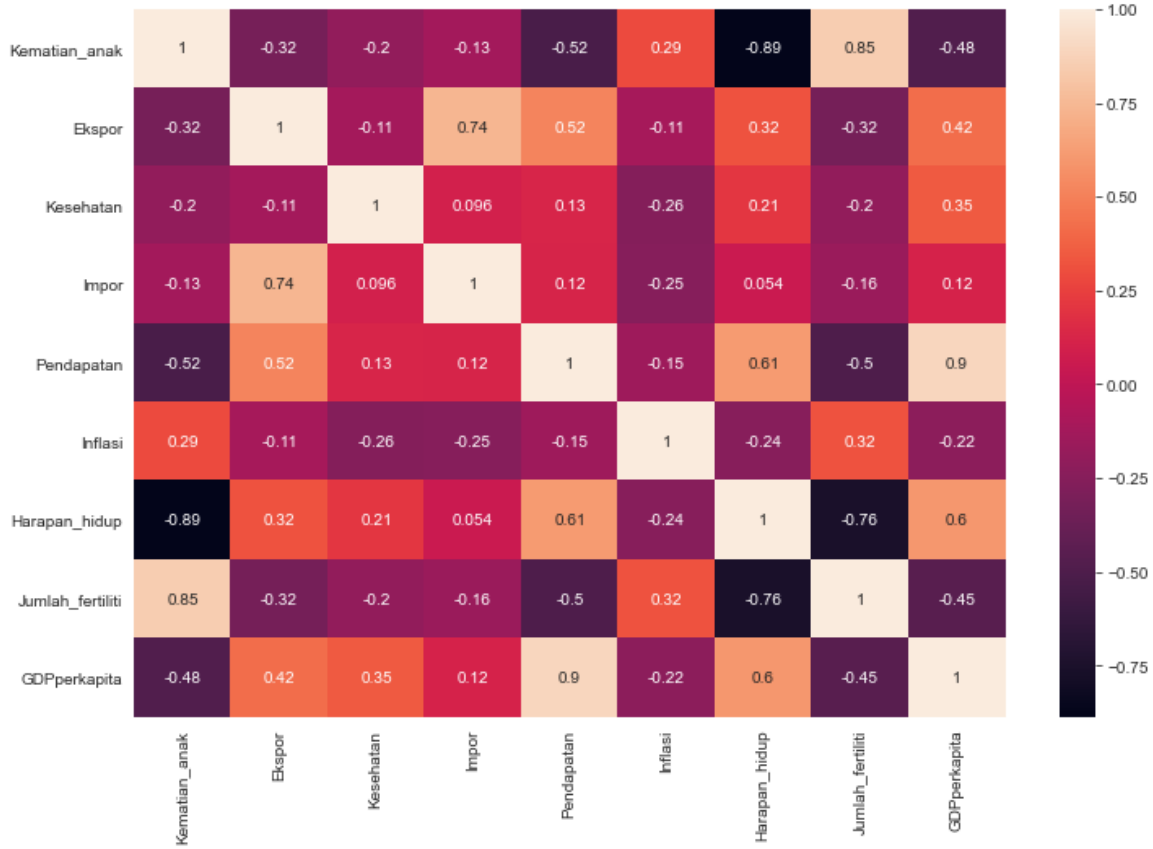
```
sns.scatterplot(data=df, x='Kematian_anak', y='GDPperkapita')
plt.show()
```



Gambar 4. Bivariate Analysis

2.3 Multivariate Analysis

Pada multivariate, dilakukan mapping mengenai korelasi antar tiap variable. Jika nilainya dekat dengan angka 1 maka korelasinya high positive, sedangkan jika mendekati -1 maka korelasinya high negative.



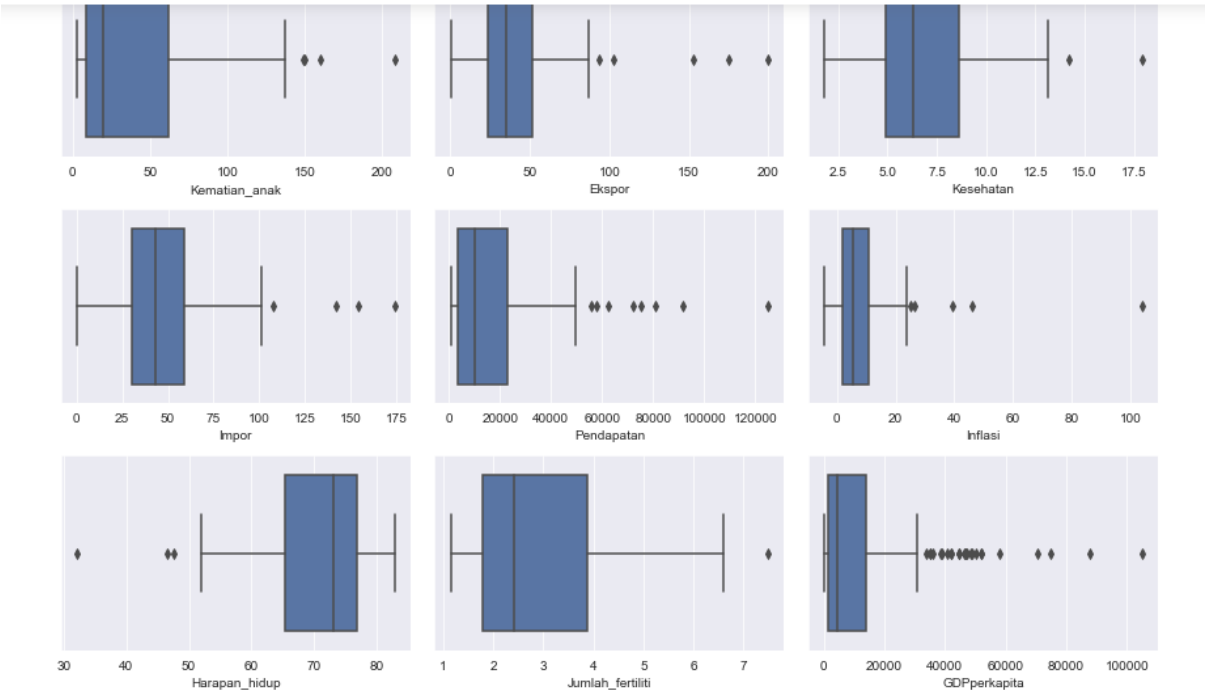
Gambar 5. Heatmap

Insight yang didapat pada heatmap tersebut adalah sebagai berikut:

1. Ekspor dan Impor memiliki korelasi positive yang cukup tinggi
2. kematian anak memiliki korelasi negative dengan harapan hidup
3. kematian anak memiliki korelasi positive dengan jumlah fertiliti
4. GDPperkapita memiliki korelasi positive yang tinggi dengan Pendapatan
5. GDPperkapita memiliki korelasi positive yang cukup tinggi dengan Ekspor

3. Check Outliers

Dari hasil visualisasi menggunakan boxplot, terlihat memang bahwa setiap variable memiliki outliers, Tetapi saya tidak akan menghandling outliers ini karena setiap negara pasti memiliki kondisi berbeda-beda. Menghapus outliers bukanlah solusi mengingat jumlah baris dataset yang sedikit.



Gambar 6. Outliers

Adanya outliers pada variabel kematian anak dan harapan hidup memberikan kita informasi bahwa outliers ini merupakan indikator bahwa negara tersebut yang membutuhkan bantuan.

4. Feature Selection

Berdasarkan pada heatmap pada gambar 5, berikut adalah list variabel yang memiliki korelasi tinggi

- Ekspor <-> Impor
- Kematian_anak <-> Harapan_hidup
- Kematian_anak <-> Jumlah_fertiliti
- GDPperkapita <-> Pendapatan
- GDPperkapita <-> Ekspor

Dari list tersebut terdapat 2 variabel saling berpasangan oleh karena itu saya harus meremove salah satu variabel yang memiliki korelasi tinggi. Kenapa? karena jika 2 variabel memiliki korelasi yang tinggi maka 2 variabel tersebut menunjukan informasi yang sama. Sehingga ditakutkan hasil dari k means akan ketarik oleh variabel2 yang memiliki korelasi tinggi.

Namun dibanding meremove variabel, saya memutuskan untuk menggunakan variabel Kematian_anak serta GDPperkapita

5. Scaling

Karena saya memutuskan untuk tidak menghandle outliers, maka saya menggunakan robust scaler yang diklaim mampu bekerja dengan baik jika ada outlier dibanding standard scaler & minmax scaler yang performanya buruk jika ada outlier di datasetnya serta menyebabkan mislead

Modelling

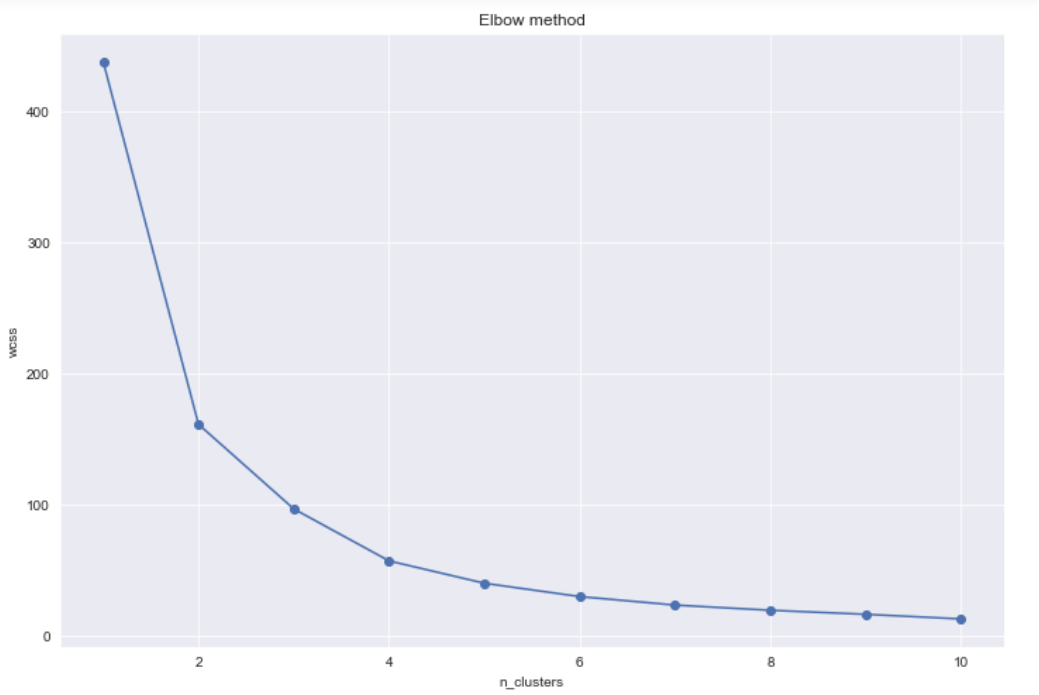
Dengan menggunakan K Means, dipilih angka 2 sebagai jumlah cluster, angka 2 dipilih sebagai initial modelling. Berikut hasil labeling dengan jumlah cluster 2

```
#clustering with KMeans

from sklearn.cluster import KMeans

kmeans1 = KMeans(n_clusters = 2, random_state=42).fit(df_std)
labels1 = kmeans1.labels_
labels1
```

```
array([1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1])
```

Gambar 7. Labeling 2 Cluster

Kemudian dilakukan elbow method untuk melihat jumlah cluster yang cocok untuk dipilih. Berdasarkan grafiknya angka 3 dipilih karena grafiknya mulai flatten.



Gambar 8. Elbow Method

Lalu dilakukan clustering lagi dengan jumlah cluster 3

```
kmeans2 = KMeans(n_clusters = 3, random_state=42).fit(df_std)
labels2 = kmeans2.labels_
labels2
```

```
array([2, 0, 0, 2, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 2, 0, 0, 0, 0,
       0, 1, 0, 2, 2, 0, 2, 1, 0, 2, 2, 0, 0, 0, 2, 2, 2, 0, 2, 0, 1, 0,
       1, 0, 0, 0, 0, 2, 2, 0, 0, 1, 1, 2, 2, 0, 1, 2, 0, 0, 0, 2, 2, 0,
       2, 0, 1, 2, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 2, 2, 1, 0, 2, 0, 0, 2,
       2, 0, 0, 1, 0, 2, 2, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0,
       1, 1, 2, 2, 1, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0,
       2, 1, 0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 1, 1, 2, 2, 0, 2, 2, 0, 0, 0,
       2, 2, 0, 1, 1, 1, 0, 0, 0, 0, 0, 2, 2])
```

Gambar 9. Labeling 3 Cluster

Dari hasil ini, dicheck performa modelling menggunakan silhouette score, berikut hasilnya:

Silhouette Score

```
from sklearn.metrics import silhouette_score

print(silhouette_score(df_std, labels= labels1)) # 2 cluster
print(silhouette_score(df_std, labels= labels2)) # 3 cluster
```

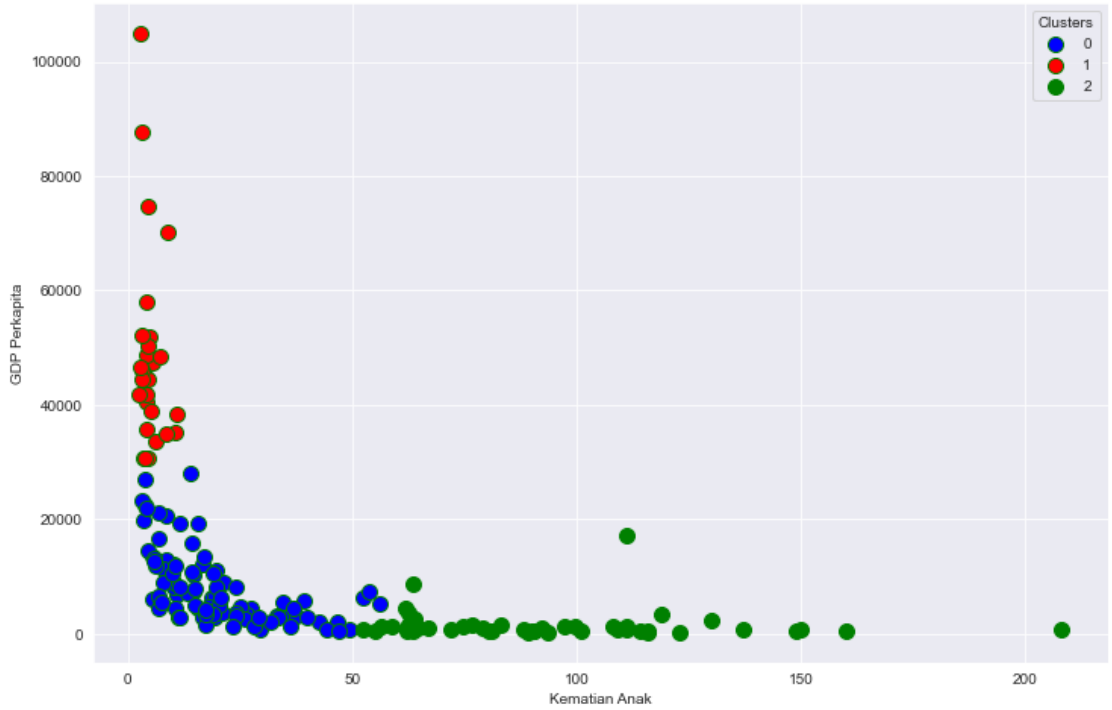
```
0.6696658626312174
0.5411808389626492
```

Gambar 10. Silhouette Score

Terlihat bahwa 3 cluster memiliki nilai yang lebih baik disbanding 2 cluster karena semakin mendekati 0 semakin baik

Analisa Hasil Clustering

Berikut hasil clustering yang telah dilakukan, terlihat bahwa terdapat 3 cluster. Pada cluster 1 memiliki GDP yang tinggi dan Kematian_anak yang kecil, sedangkan pada cluster 2 memiliki GDP yang rendah dan kematian anak yang tinggi. Pada cluster 0 memiliki GDP yang rendah dan kematian anak yang cukup rendah



Gambar 11. Hasil clustering denga variable Kematian_anak dan GDPperkapita

Dari sini kemudian variable dari dataset dibagi menjadi 2 bagian yaitu health dan economic, sehingga didapat kesimpulan bahwa:

- Label 0 merupakan kelompok negara **Low-Mid Economic, Mid-High Sanitation**

- Label 1 merupakan kelompok negara **High Economic, High Sanitation**

- Label 2 merupakan kelompok negara **Low Economic, Low Sanitation**

Selanjutnya akan dilakukan mapping untuk memberikan kategori dari setiap label untuk memudahkan dalam pembacaan berikut listnya:

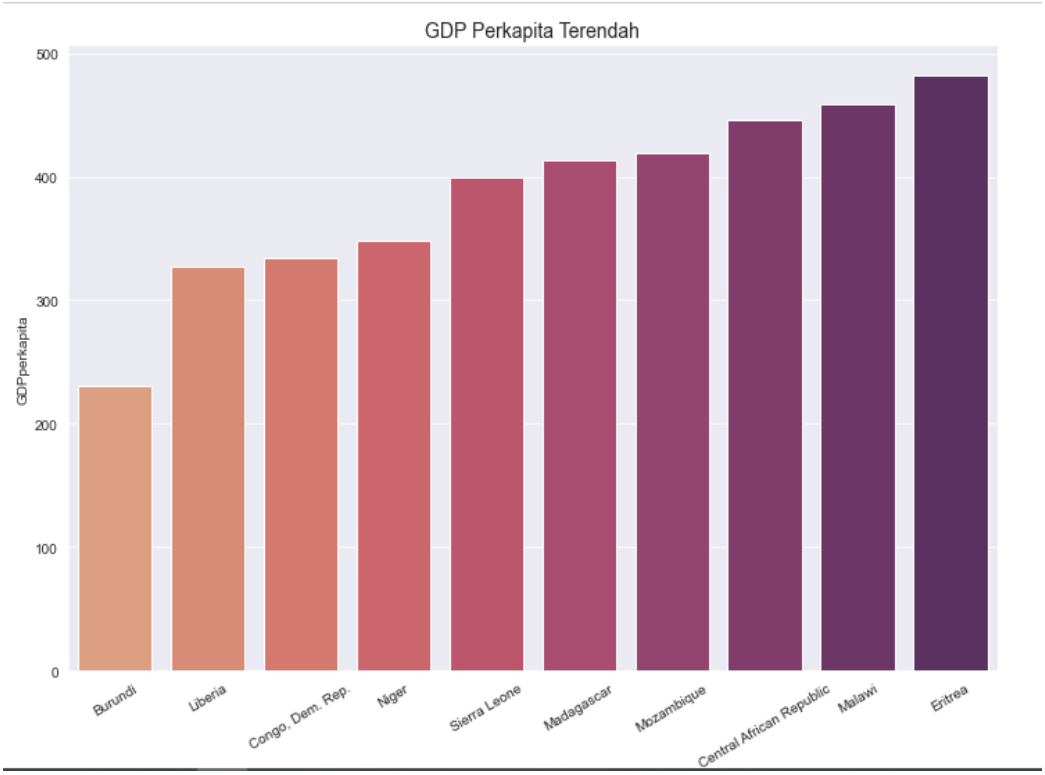
- Label 1 termasuk negara dengan kategori 'Safe' (aman)

- Label 0 termasuk negara dengan kategori 'Warning' (waspada)

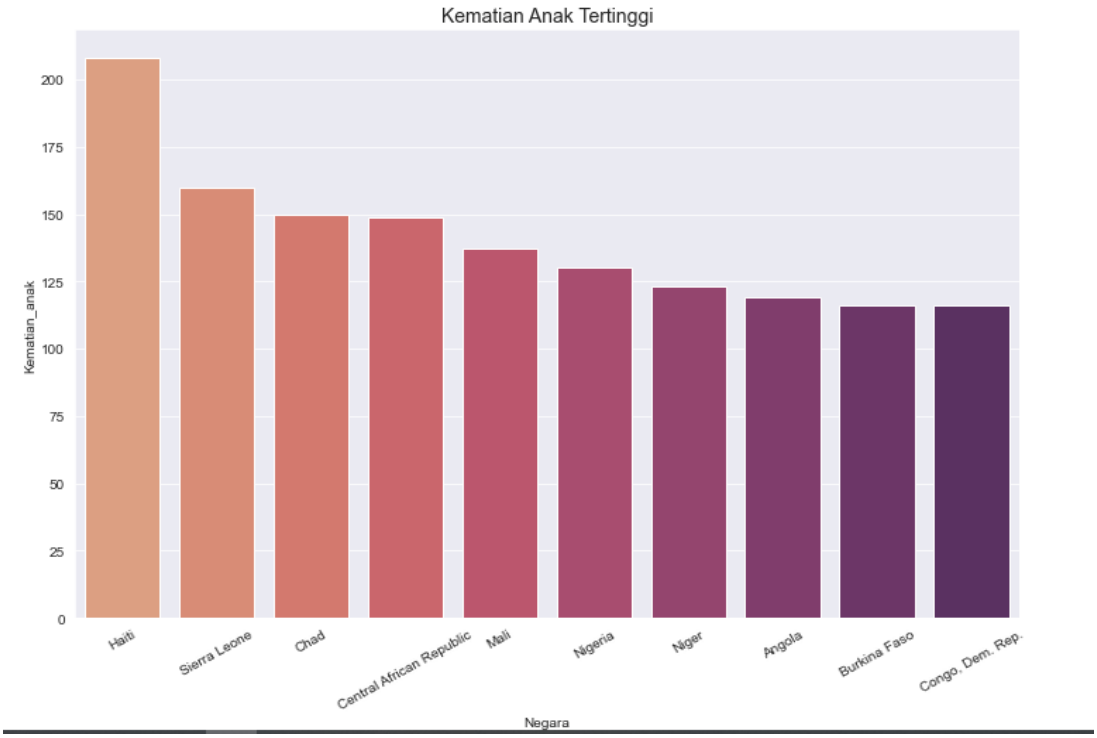
- Label 2 termasuk negara dengan kategori 'Danger' (bahaya)

Conclusion

Dari hasil yang diperoleh, ada 47 negara yang masuk kategori Danger dan membutuhkan bantuan, kemudian diurutkan lagi berdasarkan GDP terendah dan Kematian anak tertinggi.



Gambar 12. Negara Kategori Danger dengan GDP terendah



Gambar 13. Negara Kategori Danger dengan Kematian anak tertinggi

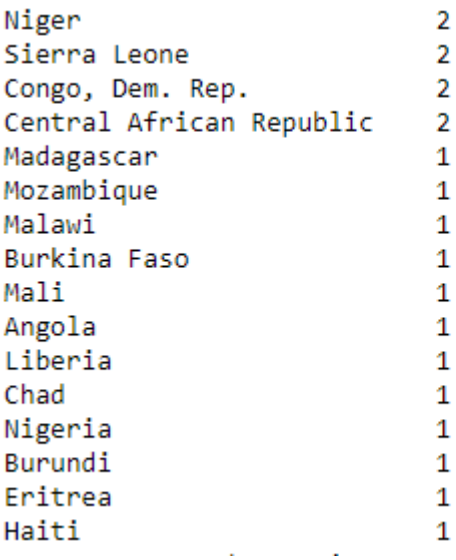
Dari hasil ini terdapat 4 negara yang masuk kedalam negara dengan GDP terendah dan kematian anak tertinggi yaitu:

- Central African Republic

- Congo, Dem. Rep

- Sierra Leone

- Niger



Niger	2
Sierra Leone	2
Congo, Dem. Rep.	2
Central African Republic	2
Madagascar	1
Mozambique	1
Malawi	1
Burkina Faso	1
Mali	1
Angola	1
Liberia	1
Chad	1
Nigeria	1
Burundi	1
Eritrea	1
Haiti	1

Gambar 14. Data negara yang masuk GDP terendah dan Kematian Anak tertinggi

Angka 2 menunjukan bahwa negara tersebut masuk kedalam negara dengan GDP terendah dan Kematian anak tertinggi, sedangkan angka 1 berarti masuk salah satunya.

Keempat negara ini masuk sebagai prioritas utama untuk mendapatkan bantuan, disusul oleh negara yang masuk kedalam GDP terendah atau kematian anak tertinggi. Lalu dilanjutkan oleh negara yang masuk kedalam kategori Danger dan belum mendapatkan bantuan.