Final Project

Presented by Aisyah Humaira

About the Project

Objective:

Untuk mengkategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan 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, Project ini bertujuan mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Kemudian menyarankan negara mana saja yang paling perlu menjadi fokus CEO.

Project Overview

Impor Library yang digunakan

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score
from sklearn.cluster import KMeans
```

Kemudian Membaca dan Menampilkan Data

```
df =pd.read_csv('Data_Negara_HELP (1).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

Mengecek informasi dataset dan memastikan tidak ada missing value kemudian memeriksa statistik deskriptif untuk fitur numerik

```
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
```

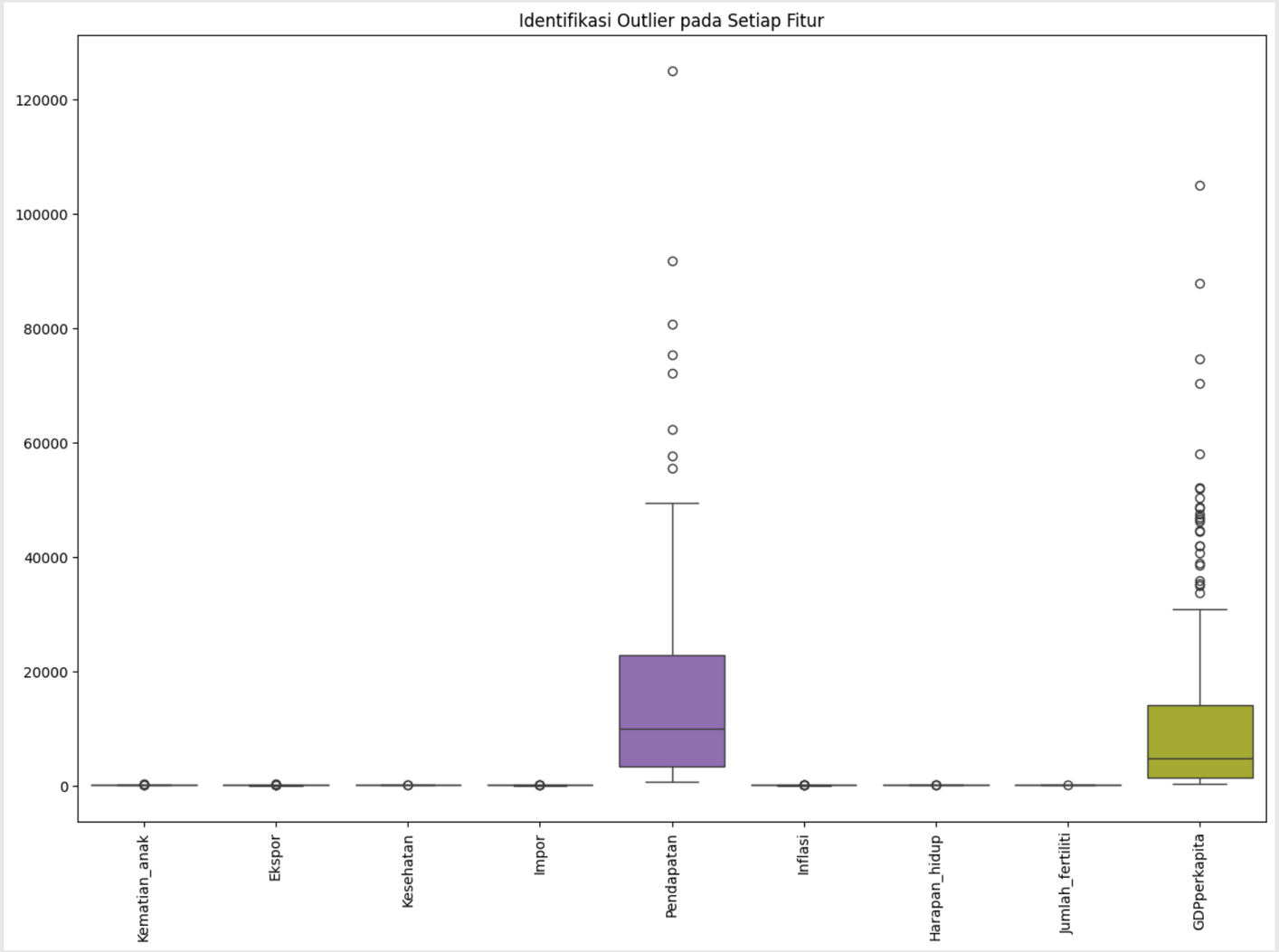
```
print(df.isnull().sum())
```

```
Negara              0
Kematian_anak       0
Ekspor              0
Kesehatan           0
Impor               0
Pendapatan          0
Inflasi             0
Harapan_hidup       0
Jumlah_fertiliti    0
GDPperkapita        0
dtype: int64
```

```
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

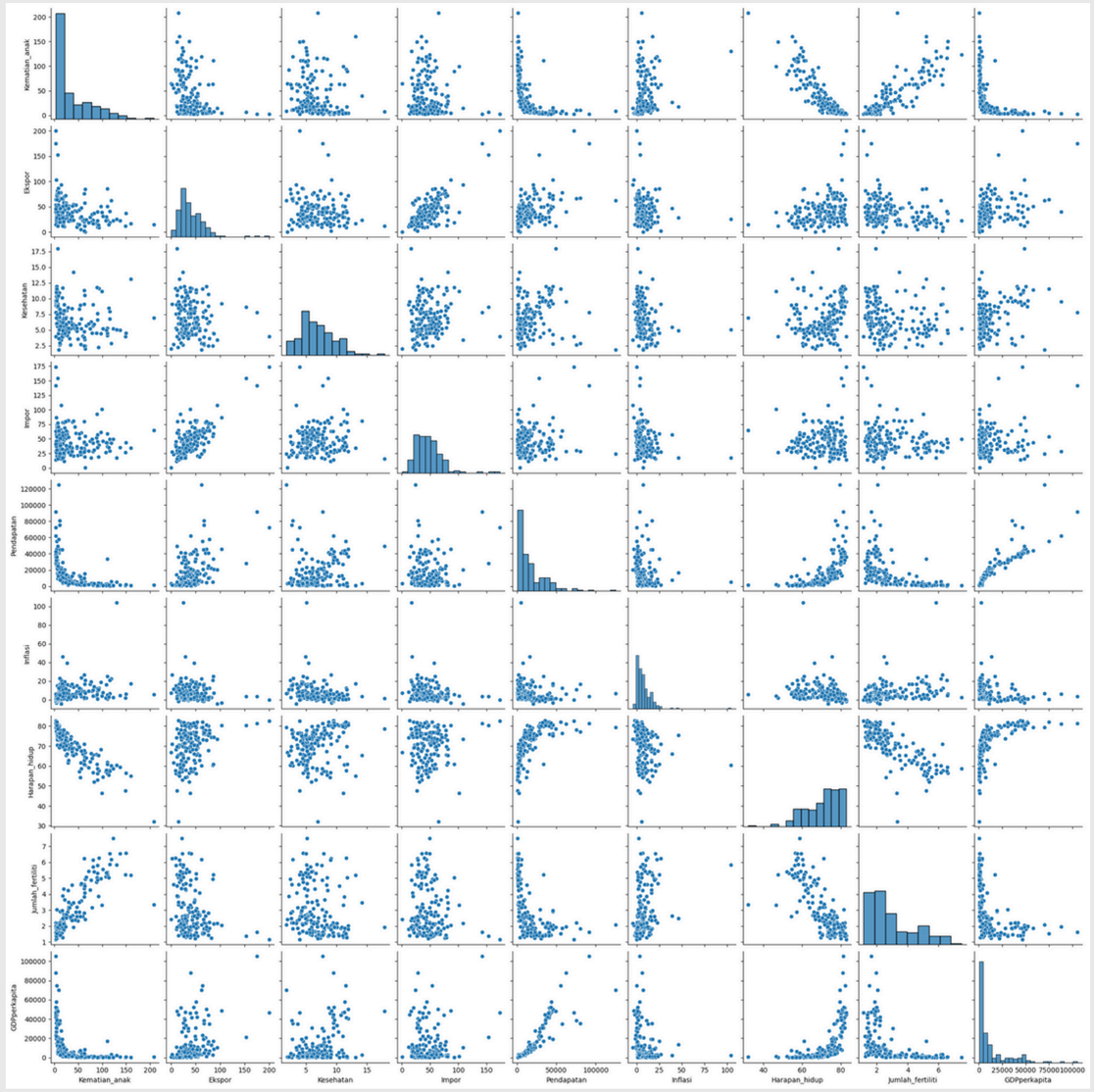
Identifikasi Outlier dengan Boxplot kecuali kolom Negara karena berisi data kategorikal



Dari visualisasi boxplot, terlihat bahwa beberapa fitur memiliki outliers yang cukup signifikan, terutama pada fitur 'Kematian_anak', 'Ekspor', 'Pendapatan', 'GDPperkapita', dan 'Inflasi'. Sehingga dilakukan penghapus outliers agar dapat membuat model lebih sederhana dan tidak terpengaruh oleh data ekstrem

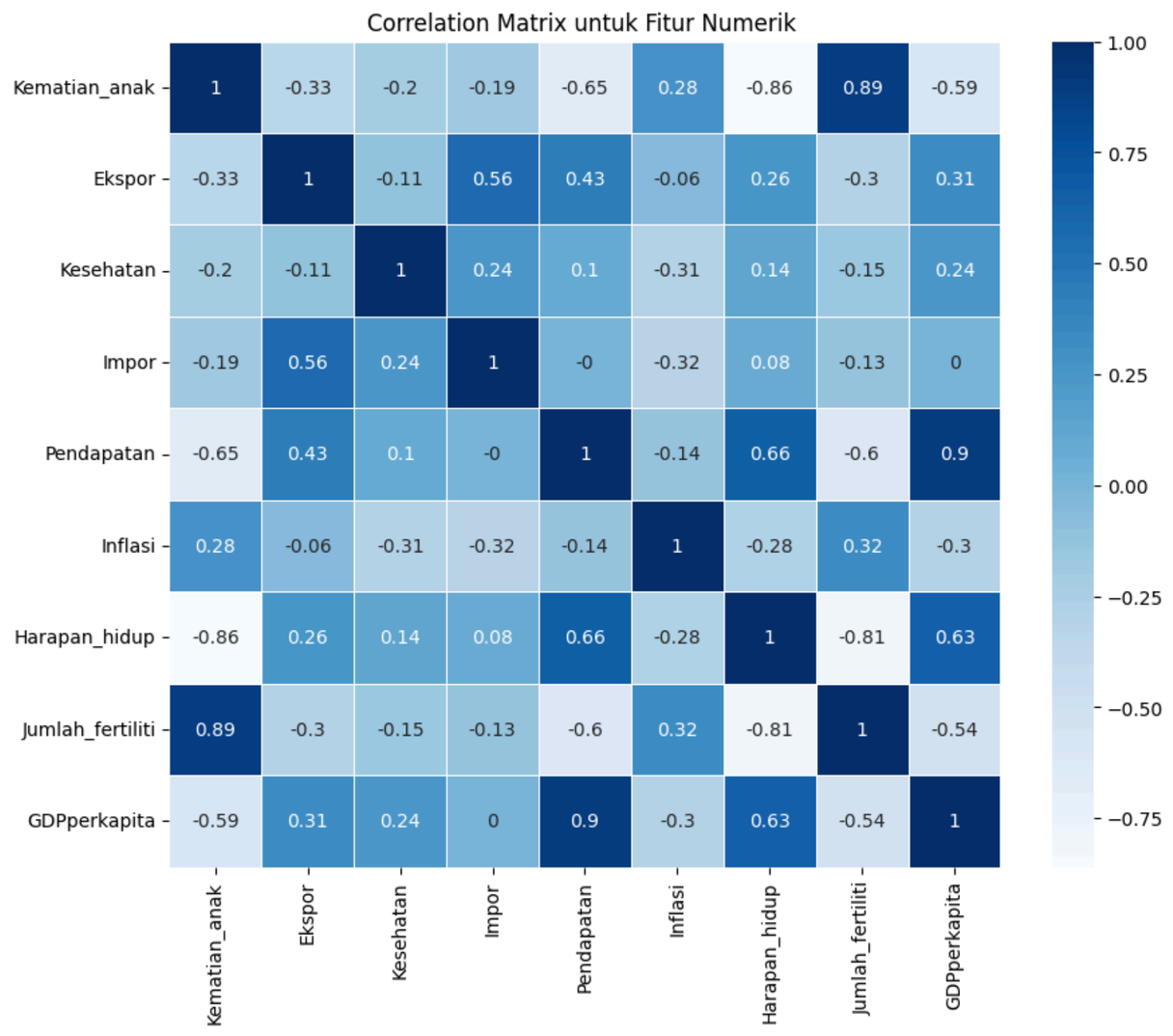
```
df_numeric = df.drop(columns=['Negara'])

Q1 = df_numeric.quantile(0.25)
Q3 = df_numeric.quantile(0.75)

IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

df_no_outliers = df[~((df_numeric < lower_bound) |
                      (df_numeric > upper_bound)).any(axis=1)]
```

Melakukan Multivariate Analysis untuk melihat korelasi antar fitur



- Beberapa pasangan variabel menunjukkan pola hubungan linear yang lemah atau tidak signifikan, misalnya, antara pendapatan dan populasi. Tidak banyak pola garis lurus yang terlihat pada scatter plot yang menandakan bahwa tidak ada hubungan linear yang kuat antara banyak variabel.
- Hubungan yang lebih jelas mungkin terlihat antara beberapa pasangan variabel tertentu. Ini bisa dilihat dari bentuk elips atau pola titik yang lebih terfokus.
- Titik-titik dalam scatter plot lebih terkonsentrasi di area tertentu, yang menunjukkan bahwa sebagian besar data berada dalam rentang nilai tertentu. Ada juga area yang jarang dengan titik-titik yang menyebar, menunjukkan outlier atau kasus yang jarang terjadi.
- Secara visual, dapat dilihat bahwa beberapa pasangan variabel yang mungkin memiliki korelasi positif atau negatif. Korelasi positif terlihat dari scatter plot yang miring ke arah atas, sedangkan korelasi negatif terlihat dari scatter plot yang miring ke arah bawah.

Untuk analisa lebih jauh, dilakukan visualisai heatmap untuk mengetahui derajat korelasi antar fitur

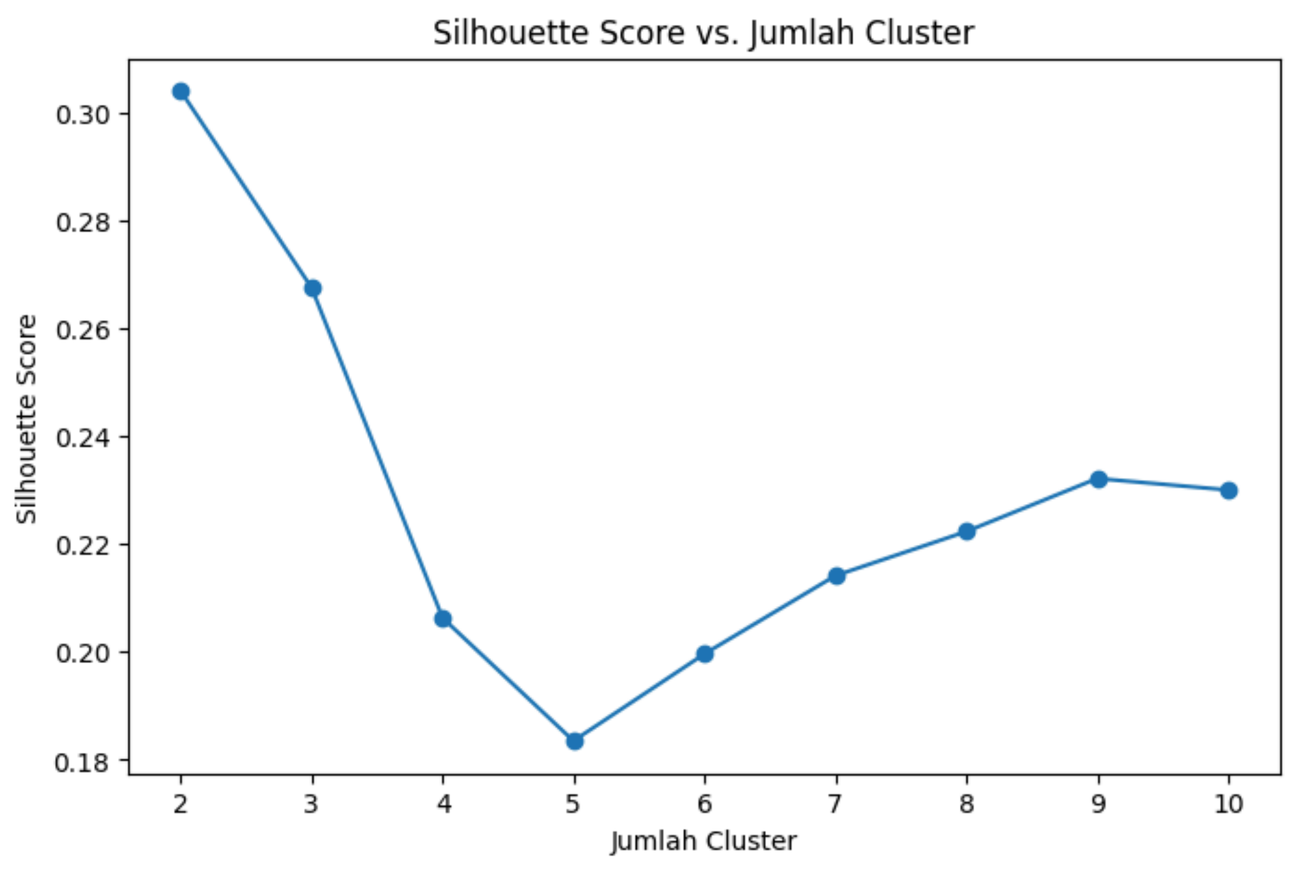


Dengan heatmap dapat memudahkan perbandingan antar kategori dengan memperlihatkan perbedaan dengan cepat melalui warna. Dari gambar tersebut, terlihat jika derajat korelasi yang paling tinggi terjadi antara fitur ‘Pendapatan’ dan ‘GDPperkapita’ sebesar 0.9 sehingga kedua fitur tersebut akan digunakan untuk perbandingan analisa selanjutnya.

Mempersiapan data dengan menentukan fitur yang akan digunakan untuk clustering, menstandarisasi fitur menggunakan StandardScaler, dan mengubah fitur menjadi skala yang sama.

```
features = ['Kematian_anak',
            'Ekspor',
            'Kesehatan',
            'Pendapatan',
            'Harapan_hidup',
            'Inflasi',
            'Jumlah_fertiliti',
            'GDPperkapita']

X = df_no_outliers[features]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

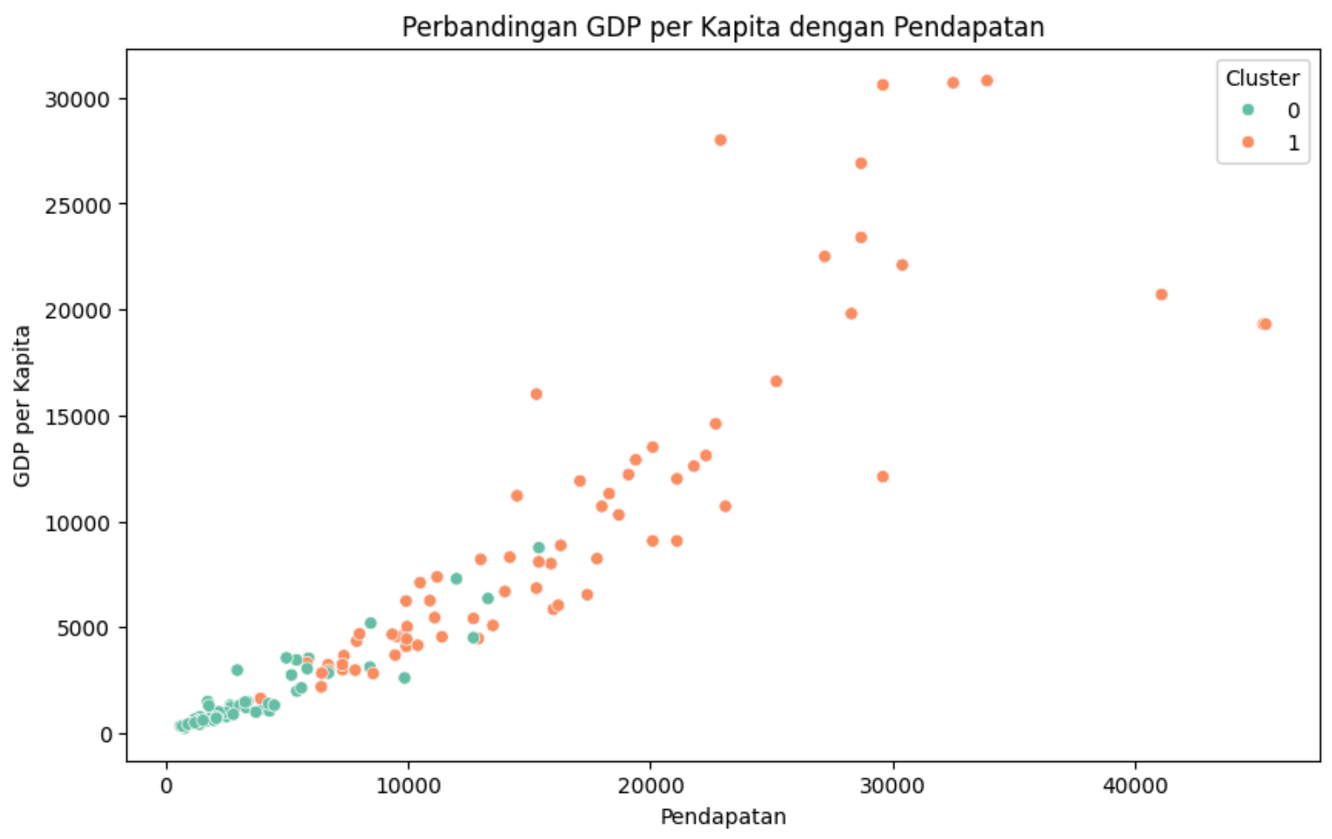
Menentukan jumlah cluster yang optimal dengan menggunakan Silhouette Method serta dilakukan pengecepak skor nilai pada tiap cluster




```
Jumlah Cluster: 2, Silhouette Score: 0.3039958616038549
Jumlah Cluster: 3, Silhouette Score: 0.2675192322736977
Jumlah Cluster: 4, Silhouette Score: 0.2063276077713101
Jumlah Cluster: 5, Silhouette Score: 0.18352940040845422
Jumlah Cluster: 6, Silhouette Score: 0.19964228930969977
Jumlah Cluster: 7, Silhouette Score: 0.21414165086084497
Jumlah Cluster: 8, Silhouette Score: 0.2223282106112131
Jumlah Cluster: 9, Silhouette Score: 0.23215980918229867
Jumlah Cluster: 10, Silhouette Score: 0.23003827519483633
```

dari pengecekan diatas didapatkan jika jumlah cluster berjumlah 2 memiliki Silhouette Score tertinggi sebesar 0.304. Ini menunjukkan bahwa dengan 2 cluster, data dapat terbagi dengan cukup baik, dengan cluster yang lebih terpisah dan homogen. Sehingga k-means akan menggunakan cluster dengan n = 2

K-Means Clustering dan Visualisasi



- Cluster 0 (hijau): Cluster ini terdiri dari negara-negara atau wilayah dengan GDP per kapita dan pendapatan yang relatif lebih rendah. Sebagian besar titik data dalam cluster ini berada di kisaran GDP per kapita dan pendapatan yang rendah hingga sedang.

- Cluster 1 (oranye): Cluster ini mencakup negara-negara atau wilayah dengan GDP per kapita dan pendapatan yang lebih tinggi. Cluster ini memiliki sebaran data yang lebih luas, menunjukkan bahwa beberapa negara atau wilayah di dalam cluster ini memiliki GDP per kapita dan pendapatan yang sangat tinggi.

Terdapat hubungan linear yang jelas antara GDP per kapita dan pendapatan. Negara-negara dengan pendapatan yang lebih tinggi cenderung memiliki GDP per kapita yang lebih tinggi juga, meskipun terdapat variasi dalam tingkat kenaikan.

Profil Cluster

Menghitung rata-rata nilai fitur dalam setiap cluster untuk memahami karakteristik masing-masing cluster.

Cluster 0:

- *Kematian Anak: 69.03*
- *Ekspor: 29.49*
- *Kesehatan: 5.87*
- *Pendapatan: 3875.73*
- *Harapan Hidup: 62.88*
- *Inflasi: 9.22*
- *Jumlah Fertiliti: 4.34*
- *GDP per Kapita: 1746.82*

Cluster 1:

- *Kematian Anak: 15.29*
- *Ekspor: 43.60*
- *Kesehatan: 6.72*
- *Pendapatan: 16807.50*
- *Harapan Hidup: 74.99*
- *Inflasi: 5.88*
- *Jumlah Fertiliti: 1.99*
- *GDP per Kapita: 10065.69*

Perbedaan Utama antara Cluster 0 dan Cluster 1:

- Kematian Anak: Cluster 0 memiliki tingkat kematian anak yang jauh lebih tinggi dibandingkan dengan Cluster 1.
- Pendapatan dan GDP per Kapita: Cluster 1 memiliki pendapatan dan GDP per Kapita yang jauh lebih tinggi daripada Cluster 0. Ini menunjukkan bahwa Cluster 1 mungkin terdiri dari negara-negara dengan ekonomi yang lebih maju.
- Ekspor dan Kesehatan: Cluster 1 memiliki nilai ekspor dan kesehatan yang lebih tinggi dibandingkan Cluster 0.
- Harapan Hidup: Harapan hidup di Cluster 1 lebih tinggi dibandingkan dengan Cluster 0.
- Inflasi dan Jumlah Fertiliti: Cluster 0 memiliki tingkat inflasi dan jumlah fertiliti yang lebih tinggi dibandingkan dengan Cluster 1.

Analisis Ekonomi dan Sosial:

- Ekonomi: Cluster 1 terdiri dari negara dengan ekonomi lebih kuat (lebih tinggi GDP per Kapita dan Pendapatan), yang mungkin juga tercermin dalam indikator kesehatan dan harapan hidup yang lebih baik.
- Kesehatan dan Sosial: Perbedaan dalam kematian anak dan kesehatan menunjukkan bahwa negara-negara di Cluster 1 mungkin memiliki sistem kesehatan yang lebih baik atau lebih maju.

dari hasil analisa yang telah dilakukan dapat ditarik kesimpulan yaitu

Negara dari Cluster 0 lebih membutuhkan bantuan karena:

- Indikator Kesehatan: Kematian anak yang tinggi dan indikator kesehatan lainnya yang rendah menunjukkan adanya kebutuhan mendesak untuk perbaikan dalam sistem kesehatan.
- Indikator Ekonomi: Pendapatan dan GDP per kapita yang rendah menunjukkan adanya kebutuhan untuk pengembangan ekonomi dan peningkatan standar hidup.
- Kondisi Sosial: Tingginya inflasi dan jumlah fertiliti mungkin menunjukkan tantangan tambahan yang memerlukan perhatian.

Sehingga negara - negara yang termasuk pada Cluster 0 untuk diberi bantuan antara lain yaitu

Afghanistan, Angola, Bangladesh, Benin, Bolivia, Botswana, Burkina Faso, Burundi, Cambodia, Cameroon, Comoros, the Democratic Republic of the Congo, Republic of the Congo, Cote d'Ivoire, Egypt, Eritrea, Gabon, Gambia, Ghana, Guatemala, Guinea, Guinea-Bissau, Guyana, India, Indonesia, Iraq, Kenya, Kiribati, Kyrgyz Republic, Lao, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Namibia, Nepal, Pakistan, Philippines, Rwanda, Samoa, Senegal, Solomon Islands, South Africa, Sudan, Tajikistan, Tanzania, Togo, Tonga, Uganda, Uzbekistan, Vanuatu, Yemen, Zambia

Thank you!