

Analisis

Step by step:

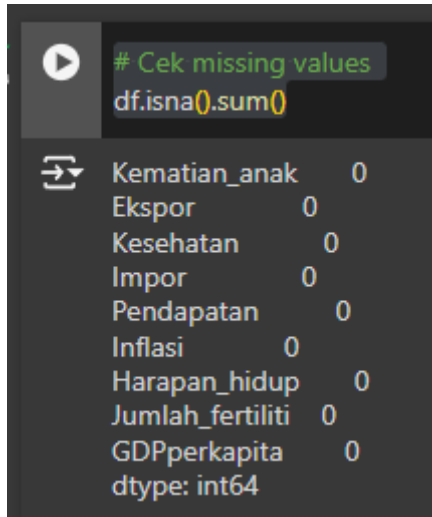
1. Reading Data
2. Exploratory Data Analysis (EDA)
3. Handling Outliers
4. Scaling Data
5. Clustering Kmeans and Visualisation
6. Filtering Negara

READING DATA

```
Kematian_anak Ekspor Kesehatan Impor Pendapatan \
Negara
Afghanistan      90.2  10.0    7.58 44.9   1610.0
Albania          16.6  28.0    6.55 48.6  9930.0
Algeria          27.3  38.4    4.17 31.4 12900.0
Angola           119.0  62.3    2.85 42.9  5900.0
Antigua and Barbuda 10.3  45.5    6.03 58.9 19100.0

Inflasi Harapan_hidup Jumlah_fertiliti GDPperkapita
Negara
Afghanistan      9.44    56.2    5.82   553.0
Albania          4.49    76.3    1.65  4090.0
Algeria          16.10   76.5    2.89  4460.0
Angola           22.40   60.1    6.16  3530.0
Antigua and Barbuda 1.44    76.8    2.13 12200.0
<class 'pandas.core.frame.DataFrame'>
Index: 167 entries, Afghanistan to Zambia
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Kematian_anak    167 non-null   float64
1   Ekspor           167 non-null   float64
2   Kesehatan        167 non-null   float64
3   Impor            167 non-null   float64
4   Pendapatan       167 non-null   float64
5   Inflasi          167 non-null   float64
6   Harapan_hidup    167 non-null   float64
7   Jumlah_fertiliti 167 non-null   float64
8   GDPperkapita     167 non-null   float64
dtypes: float64(9)
memory usage: 13.0+ KB
```

Mengimport library yang diperlukan untuk analisis data dan membaca, serta memproses dataset. Selanjutnya, menampilkan informasi awal untuk memahami karakteristik dataset sebelum melakukan analisis lebih lanjut.

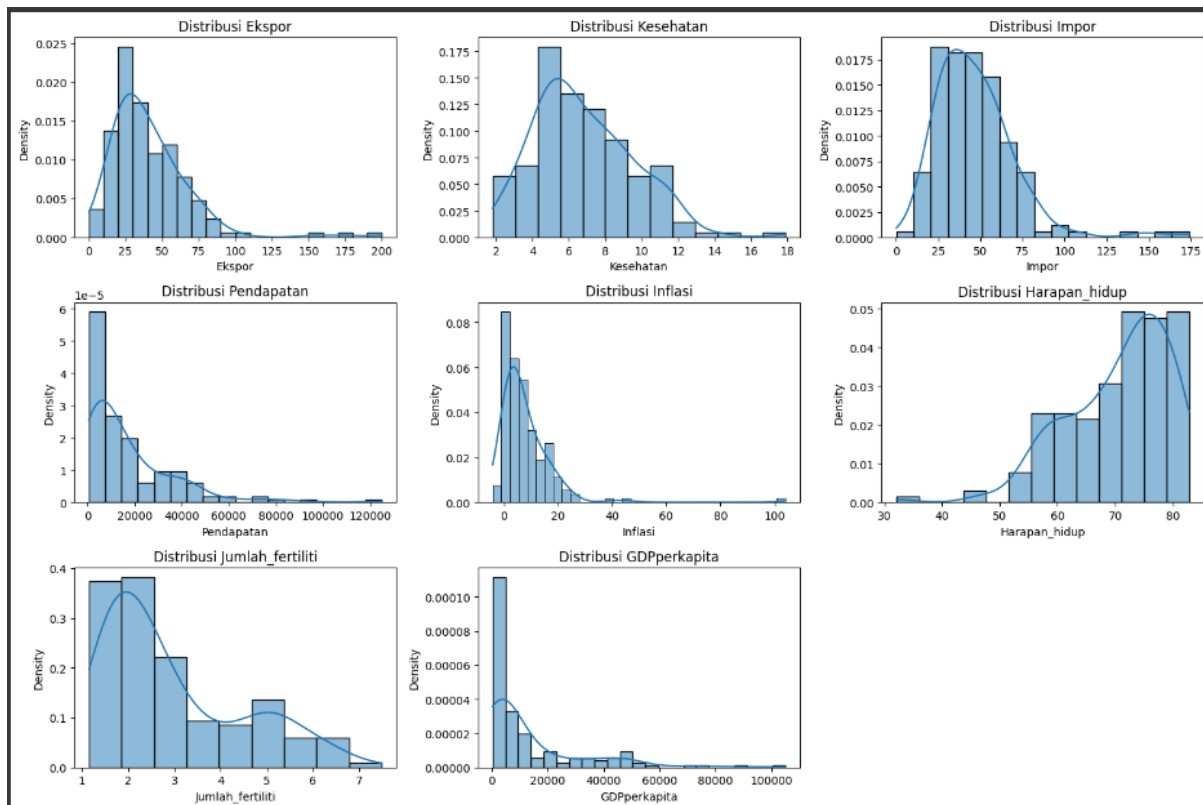


```
# Cek missing values
df.isna().sum()
```

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

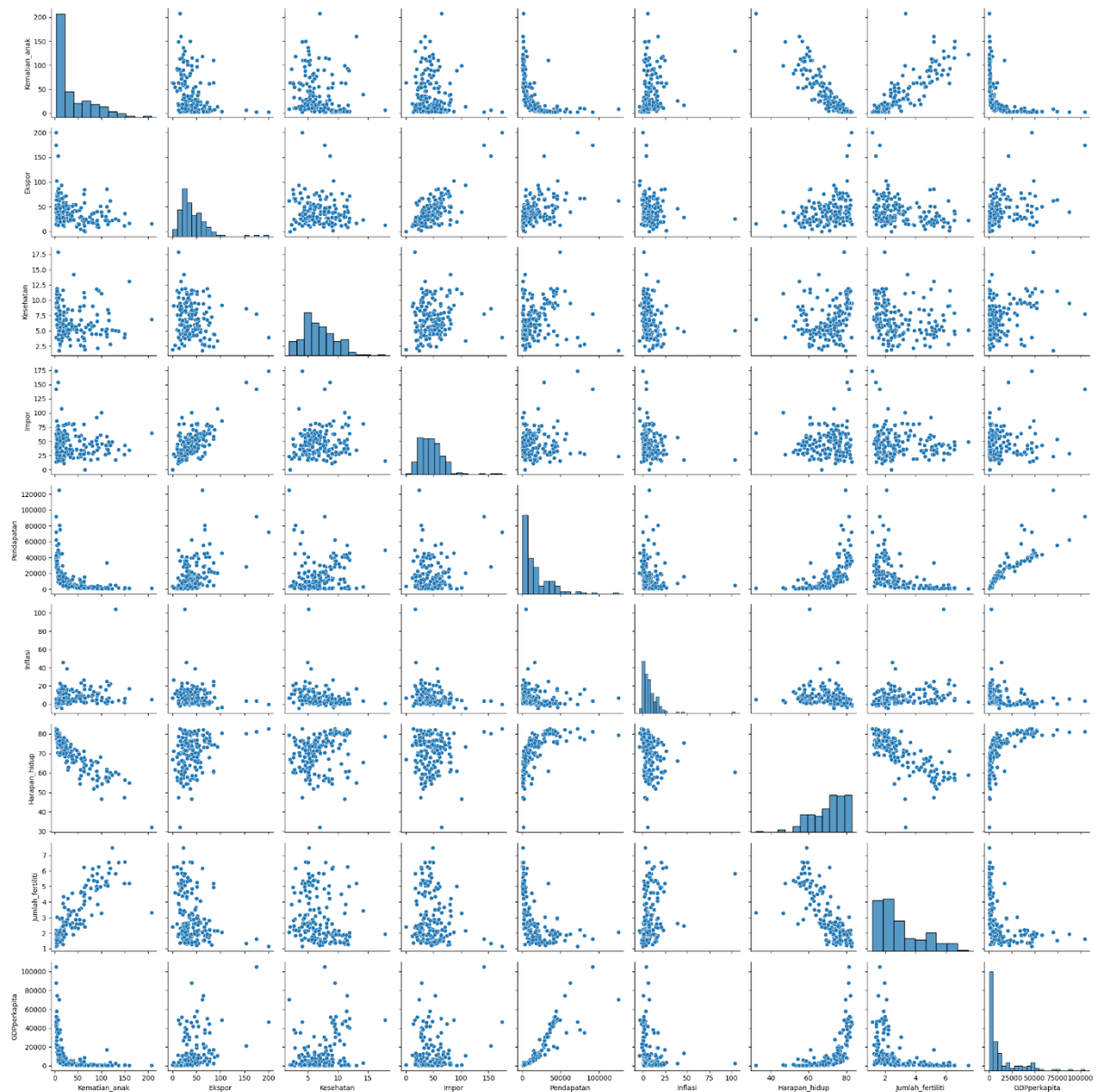
Mengidentifikasi jumlah nilai yang hilang di setiap kolom untuk mengetahui apakah diperlukan langkah tambahan dalam pra-proses data, seperti imputasi atau penghapusan nilai yang hilang. Dari sini value menunjukkan nilai 0 pada semua kolom, hal ini berarti tidak terdapat data yang hilang pada setiap kolom

EXPLORATORY DATA ANALYSIS (EDA)

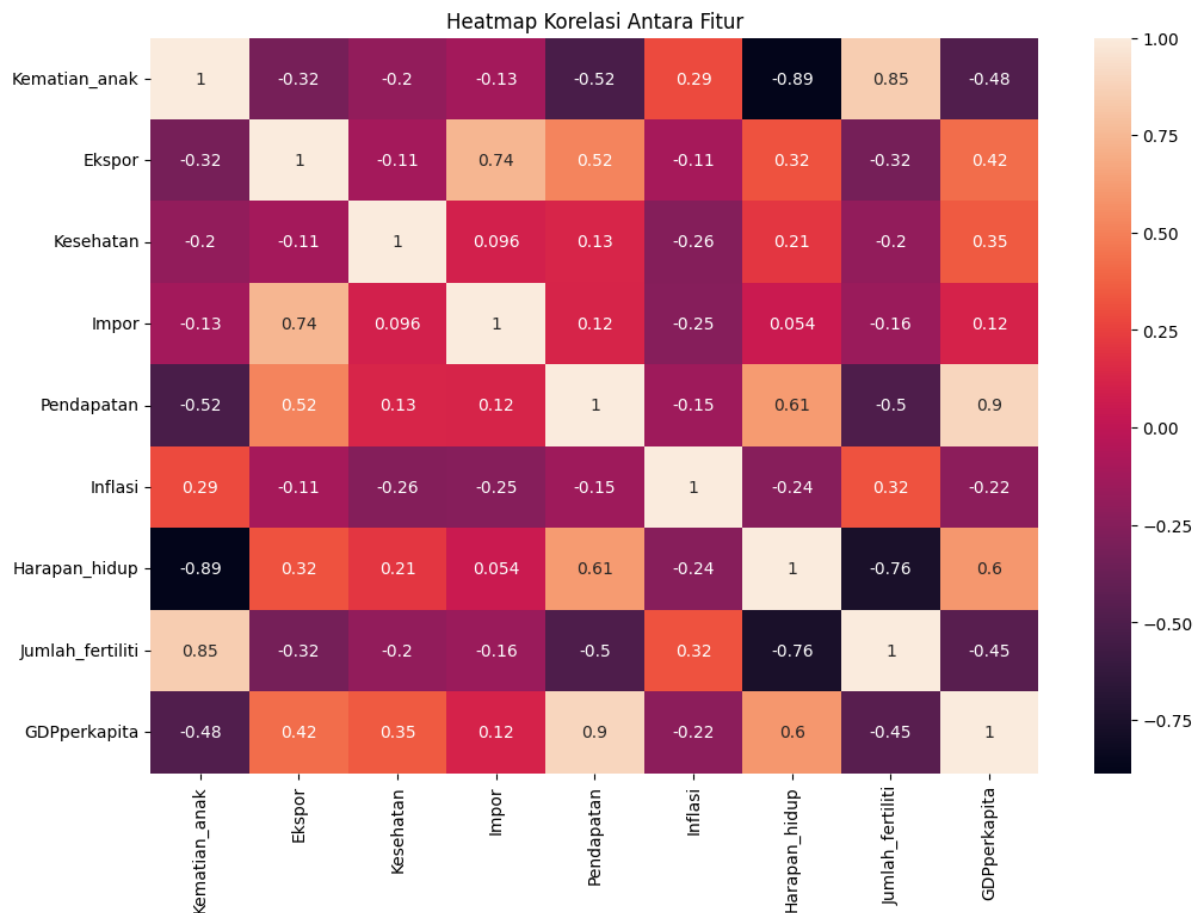


Dari histogram, kita dapat melihat apakah data dari setiap variabel berdistribusi normal, memiliki skewness, atau terdapat outliers. Dengan menambahkan estimasi kepadatan kernel ($kde=True$), visualisasi menjadi lebih informatif, menunjukkan bagaimana data tersebar dan membantu mengidentifikasi pola distribusi dengan lebih jelas.

Setiap histogram memberikan informasi tentang frekuensi atau kerapatan nilai dalam setiap variabel. Misalnya, pada histogram Harapan_Hidup menunjukkan kalau sebagian besar negara memiliki harapan hidup yang tinggi. Tetapi, pada sektor ekonomi sebagian besar negara berada di sebelah kiri, baik itu variable GDPperkapita, pendapatan, atau inflasi sekalipun.



Dari sini kita dapat matriks pair plot yang memvisualisasikan hubungan antara setiap pasangan variabel dalam dataset "Data_Negara_HELP". Kita dapat melihat korelasi antara fitur-fitur seperti pendapatan per kapita, harapan hidup, dan angka kematian anak.

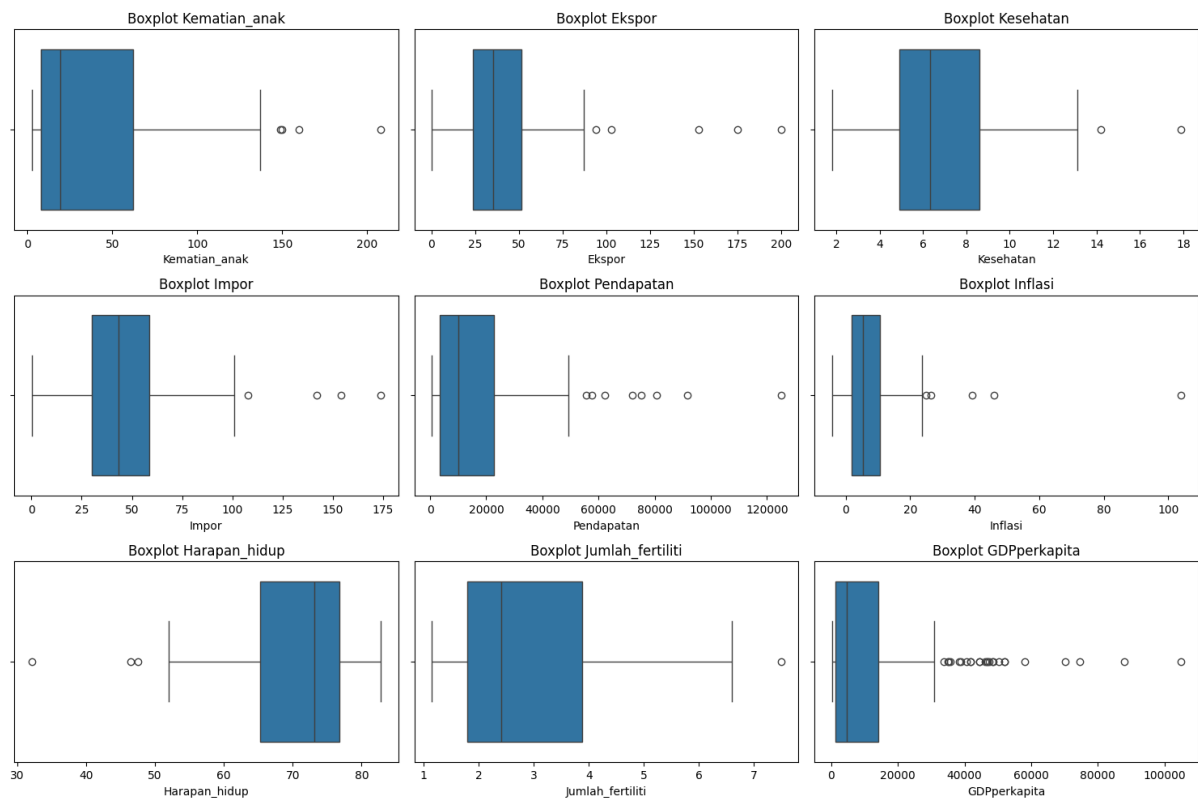


Kemudian, di dalam heatmap ini:

- Ekspor dan Impor memiliki korelasi positif, artinya negara – negara aktif melakukan perdagangan di sektor ekspor maupun impor nya.
- Pendapatan memiliki korelasi positif dengan GDPperkapita. Jadi, semakin tinggi GDP dalam suatu negara, semakin tinggi juga pendapatan di negara tersebut.
- Jumlah_fertiliti memiliki korelasi negatif yang cukup tinggi dengan variabel Harapan hidup dan pendapatan.
- Kemudian, jumlah_fertiliti memiliki korelasi yang paling kuat dengan kematian_anak. Maka dari itu, semakin tinggi tingkat kelahiran anak di suatu negara, semakin tinggi pula tingkat kematian anak di negara tersebut.
- Variabel Kematian_anak memiliki korelasi negatif yang tinggi dengan Harapan_hidup. Artinya, semakin rendah harapan hidup di suatu negara, maka tingkat kematian anaknya semakin tinggi.

- Variabel kesehatan tidak terlalu mengindikasikan penting dengan seluruh variabel. Artinya, masyarakat di negara maju atau berkembang sekalipun belum tentu mengeluarkan biaya yang besar untuk kesehatan.

HANDLING OUTLIERS



Dari hasil boxplot diatas, dapat dilihat bahwa setiap variabelnya memiliki outliers yang cukup banyak. Tetapi, dari sini kita fokus terhadap outliers dari variabel pendapatan, GDPperkapita, dan harapan_hidup saja karena outliers pada variabel ini dianggap sebagai negara yang sudah maju dan tidak memerlukan bantuan keuangan lagi.

```

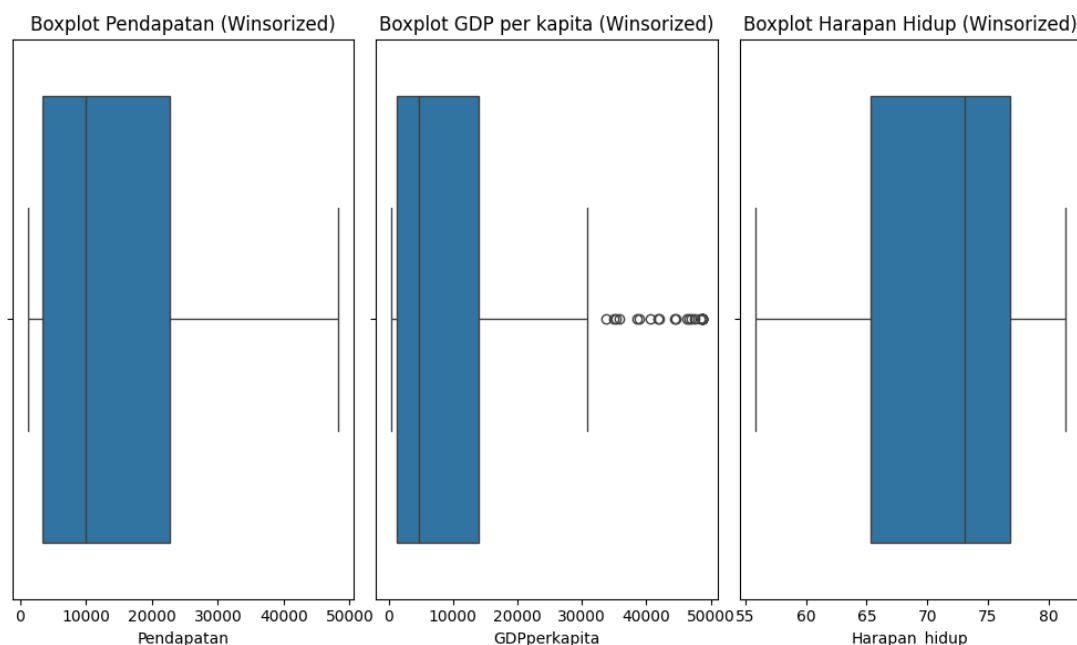
import numpy as np
# Handling with winsorization
# Tentukan batas winsorization
q1_pendapatan = df['Pendapatan'].quantile(0.05)
q3_pendapatan = df['Pendapatan'].quantile(0.95)
q1_gdp = df['GDPperkapita'].quantile(0.05)
q3_gdp = df['GDPperkapita'].quantile(0.95)
q1_harapan_hidup = df['Harapan_hidup'].quantile(0.05)
q3_harapan_hidup = df['Harapan_hidup'].quantile(0.95)

# Terapkan winsorization
df['Pendapatan'] = np.where(df['Pendapatan'] < q1_pendapatan, q1_pendapatan, df['Pendapatan'])
df['Pendapatan'] = np.where(df['Pendapatan'] > q3_pendapatan, q3_pendapatan, df['Pendapatan'])
df['GDPperkapita'] = np.where(df['GDPperkapita'] < q1_gdp, q1_gdp, df['GDPperkapita'])
df['GDPperkapita'] = np.where(df['GDPperkapita'] > q3_gdp, q3_gdp, df['GDPperkapita'])
df['Harapan_hidup'] = np.where(df['Harapan_hidup'] < q1_harapan_hidup, q1_harapan_hidup, df['Harapan_hidup'])
df['Harapan_hidup'] = np.where(df['Harapan_hidup'] > q3_harapan_hidup, q3_harapan_hidup, df['Harapan_hidup'])

```

Berikut adalah cara kita menangani outliers dengan cara menggunakan metode **“Winsorization”**. Winsorization adalah teknik untuk menangani outliers dengan mengubah nilai outlier ke nilai yang lebih dekat dengan batas persentil tertentu (dalam hal ini, persentil ke-5 dan ke-95). Dengan demikian, outliers dalam ketiga kolom ini dikendalikan, yang dapat memberikan distribusi data yang lebih normal dan membuat analisis berikutnya, seperti clustering, menjadi lebih stabil dan akurat.

Selanjutnya, ini adalah visualisasi boxplot setelah dilakukan penghapusan outlier:



SCALING DATA

```
# Inialisasi objek StandardScaler
scaler = StandardScaler()

# Melakukan scaling data
scaled_data = scaler.fit_transform(df)

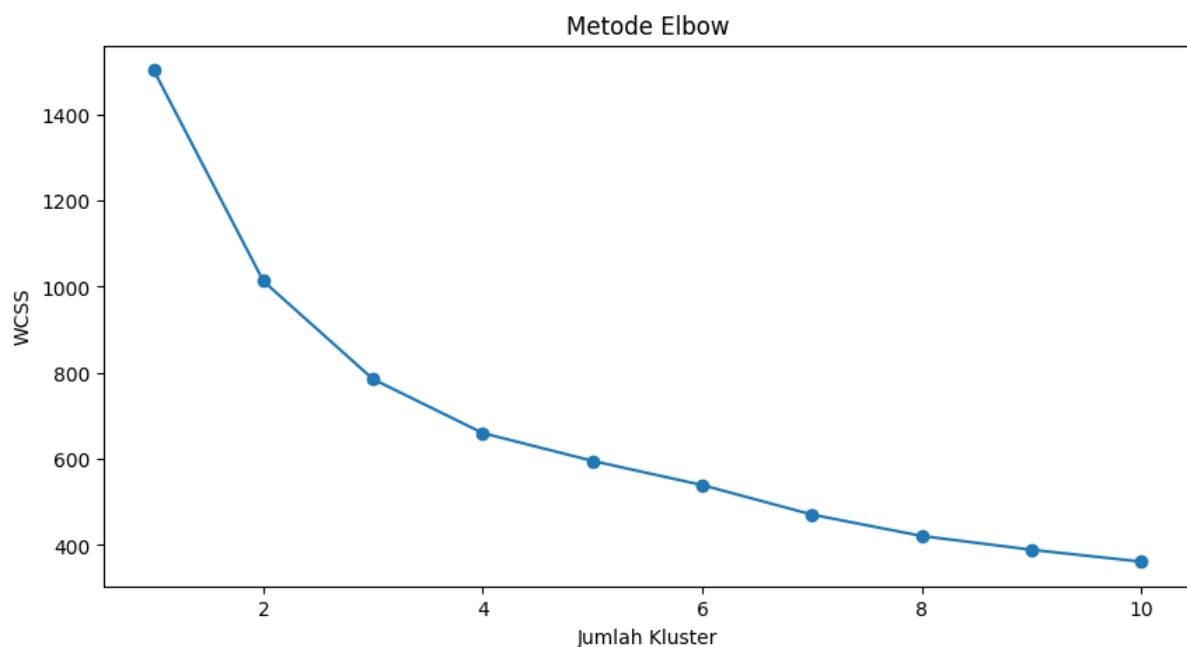
# Membuat DataFrame baru dari data yang telah discaling
df_scaled = pd.DataFrame(scaled_data, columns=df.columns)
df_scaled
```

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	Cluster
0	1.291532	-1.138280	0.279088	-0.082455	-0.960575	0.157336	-1.825310	1.902882	-0.757874	-1.394921
1	-0.538949	-0.479658	-0.097016	0.070837	-0.395590	-0.312347	0.682454	-0.859973	-0.523775	0.946303
2	-0.272833	-0.099122	-0.966073	-0.641762	-0.193907	0.789274	0.707406	-0.038404	-0.499286	0.946303
3	2.007808	0.775381	-1.448071	-0.165315	-0.669255	1.387054	-1.338729	2.128151	-0.560839	-1.394921
4	-0.695634	0.160668	-0.286894	0.497568	0.227115	-0.601749	0.744836	-0.541946	0.012991	0.946303
...

Scaling data dilakukan untuk membuat pengelompokan lebih akurat, kami menstandarisasi data dengan menskalakan ulang menggunakan scaler standar yang disediakan oleh scikit-learn.

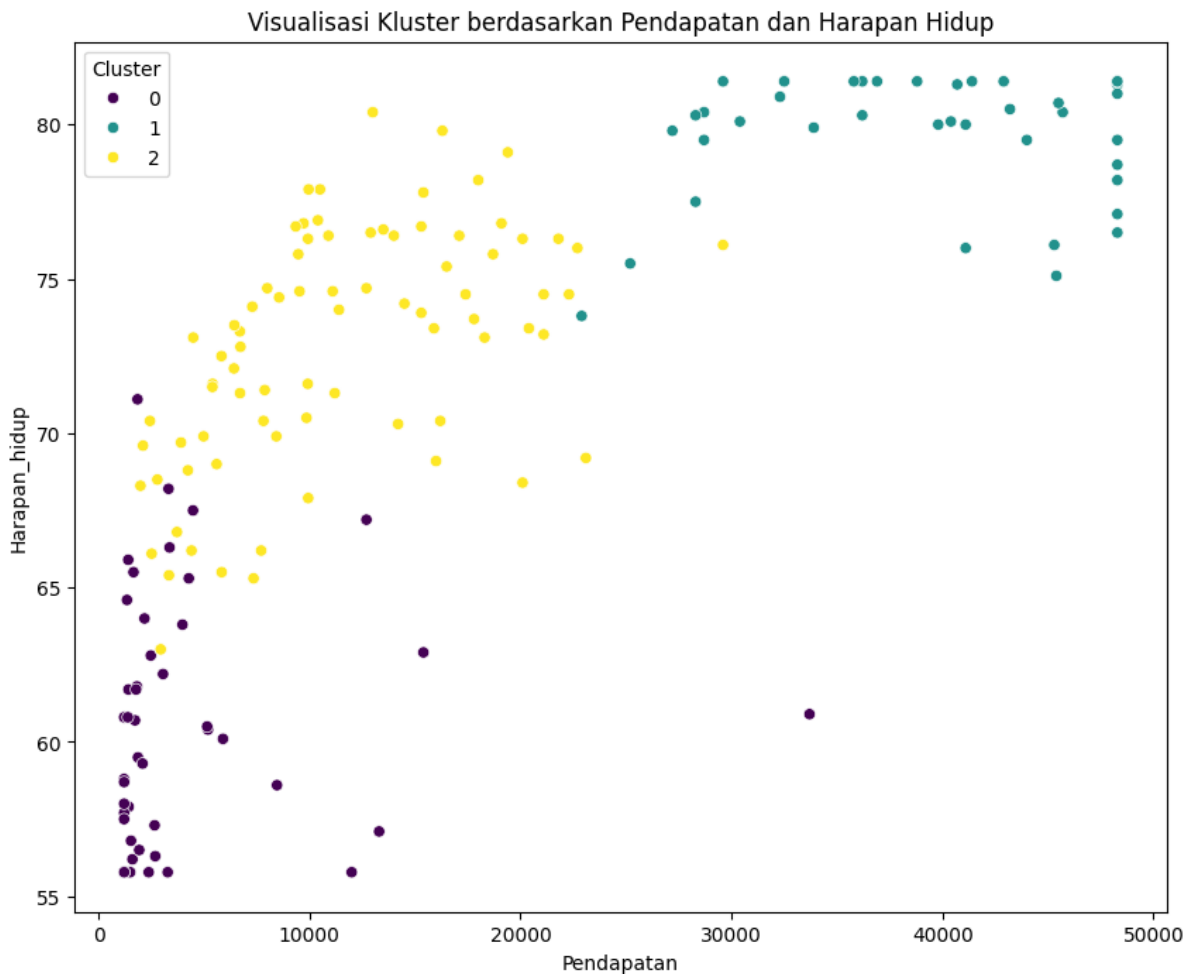
KMEANS CLUSTERING AND VISUALIZING CLUSTERS FORMED

- Menentukan Jumlah Cluster dengan Elbow Method



Berdasarkan perhitungan menggunakan metode Elbow, jumlah cluster yang optimal untuk digunakan dalam clustering negara adalah 3. Nilai ini diperoleh dari interpretasi grafik Elbow, di mana garis mulai membentuk siku pada angka 3, sehingga jumlah cluster yang efektif adalah 3 cluster.

- **KMeans Clustering**



Dari hasil grafik atau gambar diatas, dapat disimpulkan negara – negara dibagi menjadi 3 cluster berdasarkan harapan_hidup dan pendapatan masing – masing negaranya, yaitu:

- Cluster 0 (Ungu): Negara dengan IPM rendah
- Cluster 1 (Hijau): Negara dengan Indeks Pembangunan Manusia tinggi.
- Cluster 2 (Kuning): Negara dengan IPM sedang.

Dari grafik diatas juga dapat disimpulkan bahwa negara yang layak untuk mendapatkan bantuan adalah negara dengan **cluster 0**. Cluster 0 ini adalah kumpulan negara dengan IPM (Indeks Pembangunan Rendah) rendah. Selanjutnya, akan dianalisis top 10 negara terendah untuk memudahkan dalam pendistribusian bantuan.

```
# Menampilkan negara dalam setiap kluster dari yang terendah ke tertinggi
bottom_countries_cluster = {}
for cluster in range(3):
    bottom_countries_cluster[cluster] = df[df['Cluster'] == cluster].sort_values(by='Pendapatan', ascending=True).head(10).index.values

for cluster, countries in bottom_countries_cluster.items():
    print(f"\n10 Negara Terendah dalam Kluster {cluster}:")
    print(countries)
```

10 Negara Terendah dalam Kluster 0:
['Liberia' 'Togo' 'Burundi' 'Central African Republic' 'Niger' 'Guinea'
'Congo, Dem. Rep.' 'Mozambique' 'Malawi' 'Sierra Leone']

10 Negara Terendah dalam Kluster 1:
['Bahamas' 'Slovak Republic' 'Portugal' 'Malta' 'Czech Republic' 'Greece'
'Slovenia' 'Israel' 'South Korea' 'New Zealand']

10 Negara Terendah dalam Kluster 2:
['Nepal' 'Tajikistan' 'Bangladesh' 'Cambodia' 'Kyrgyz Republic' 'Vanuatu'
'Micronesia, Fed. Sts.' 'Myanmar' 'Moldova' 'Uzbekistan']

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertilitas	GDPperkapita	Cluster
Negara										
Liberia	89.3	19.100	11.80	92.6000	1213.0	5.470	60.80	5.02	465.9	0
Togo	90.3	40.200	7.65	57.3000	1213.0	1.180	58.70	4.87	488.0	0
Burundi	93.6	8.920	11.60	39.2000	1213.0	12.300	57.70	6.26	465.9	0
Central African Republic	149.0	11.800	3.98	26.5000	1213.0	2.010	55.78	5.21	465.9	0
Niger	123.0	22.200	5.16	49.1000	1213.0	2.550	58.80	7.49	465.9	0
Guinea	109.0	30.300	4.93	43.2000	1213.0	16.100	58.00	5.34	648.0	0
Congo, Dem. Rep.	116.0	41.100	7.91	49.6000	1213.0	20.800	57.50	6.54	465.9	0
Mozambique	101.0	31.500	5.21	46.2000	1213.0	7.640	55.78	5.56	465.9	0
Malawi	90.5	22.800	6.59	34.9000	1213.0	12.100	55.78	5.31	465.9	0
Sierra Leone	160.0	16.800	13.10	34.5000	1220.0	17.200	55.78	5.20	465.9	0
Bahamas	13.8	35.000	7.89	43.7000	22900.0	-0.393	73.80	1.86	28000.0	1

Dari gambar diatas dapat disimpulkan, ada 10 negara terendah dalam cluster 0 (kota merah) yang layak mendapatkan distribusi bantuan berupa uang, diantaranya, yaitu:

1. Liberia
2. Togo
3. Burundi
4. Central African Republik
5. Niger
6. Guinea
7. Congo, Dem. Rep
8. Mozambique
9. Malawi
10. Sierra Leone