

# Laporan Analisis Data Help International

SEPTEMBER 2021

Oleh Wafika Samsea

# 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.



# Step Penyelesaian 1

## EXPLORATORY DATA ANALYSIS

- Visualisasi data
- Cek outlier (data pencilan)
- Cek missing value (baris / kolom data yang datanya gaada)
- Cek korelasi antar fitur

## DATA PREPROCESSING

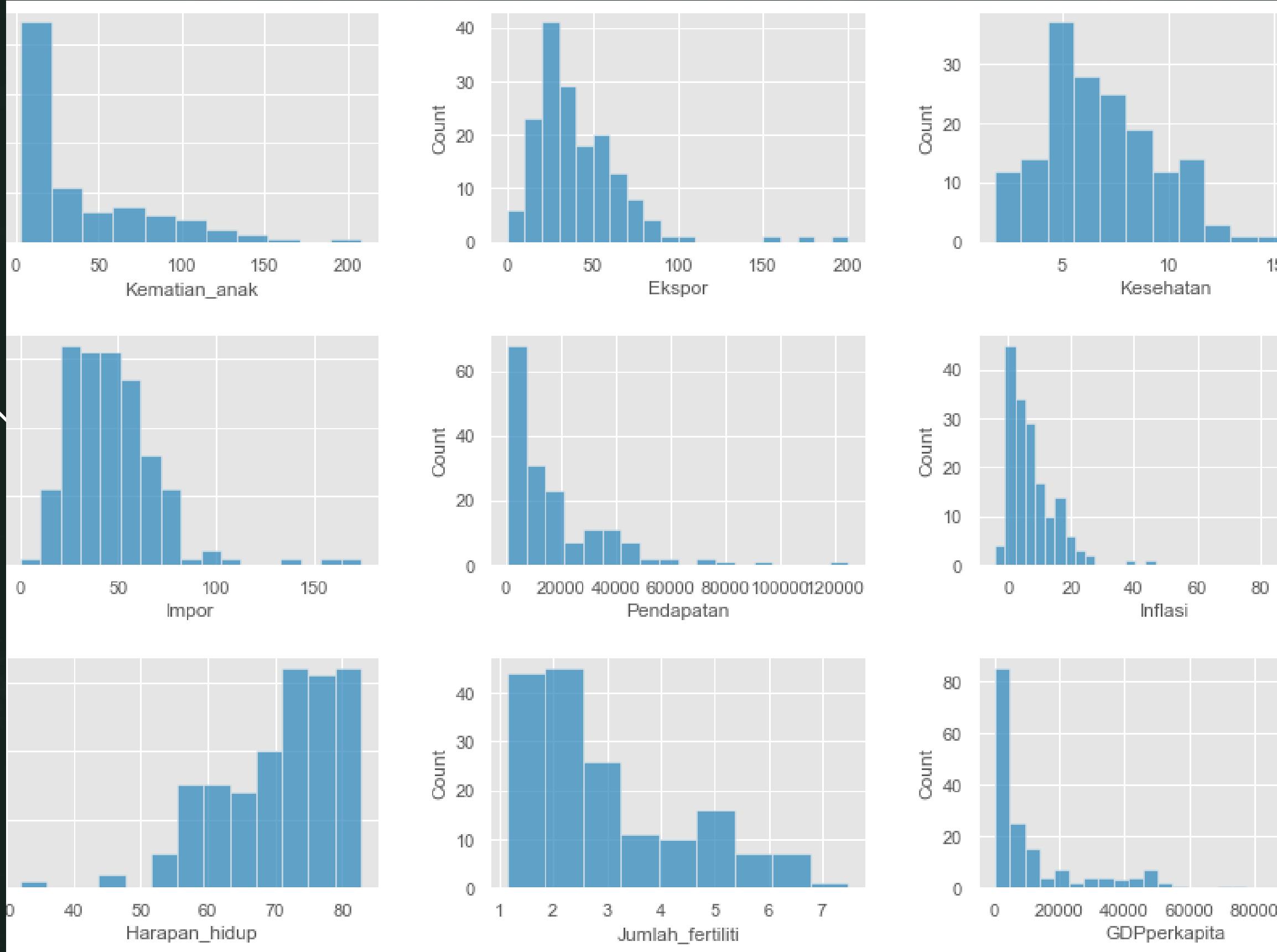
- Menghilangkan baris yang memiliki missing value
- Melakukan standarisasi data
- Menghapus outliers
- Menghapus/Mengganti data kosong

## CLUSTERING & DECISION MAKING

- Mencari n clusters yang tepat
- Mengurutkan features bagi cluster terpilih
- Membuat kesimpulan terhadap keputusan akhir

# EXPLORATORY DATA ANALYSIS

2



**GAMBAR 1.0**

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

fig = plt.figure(figsize=(15, 10))
fig.subplots_adjust(hspace=0.4, wspace=0.3)

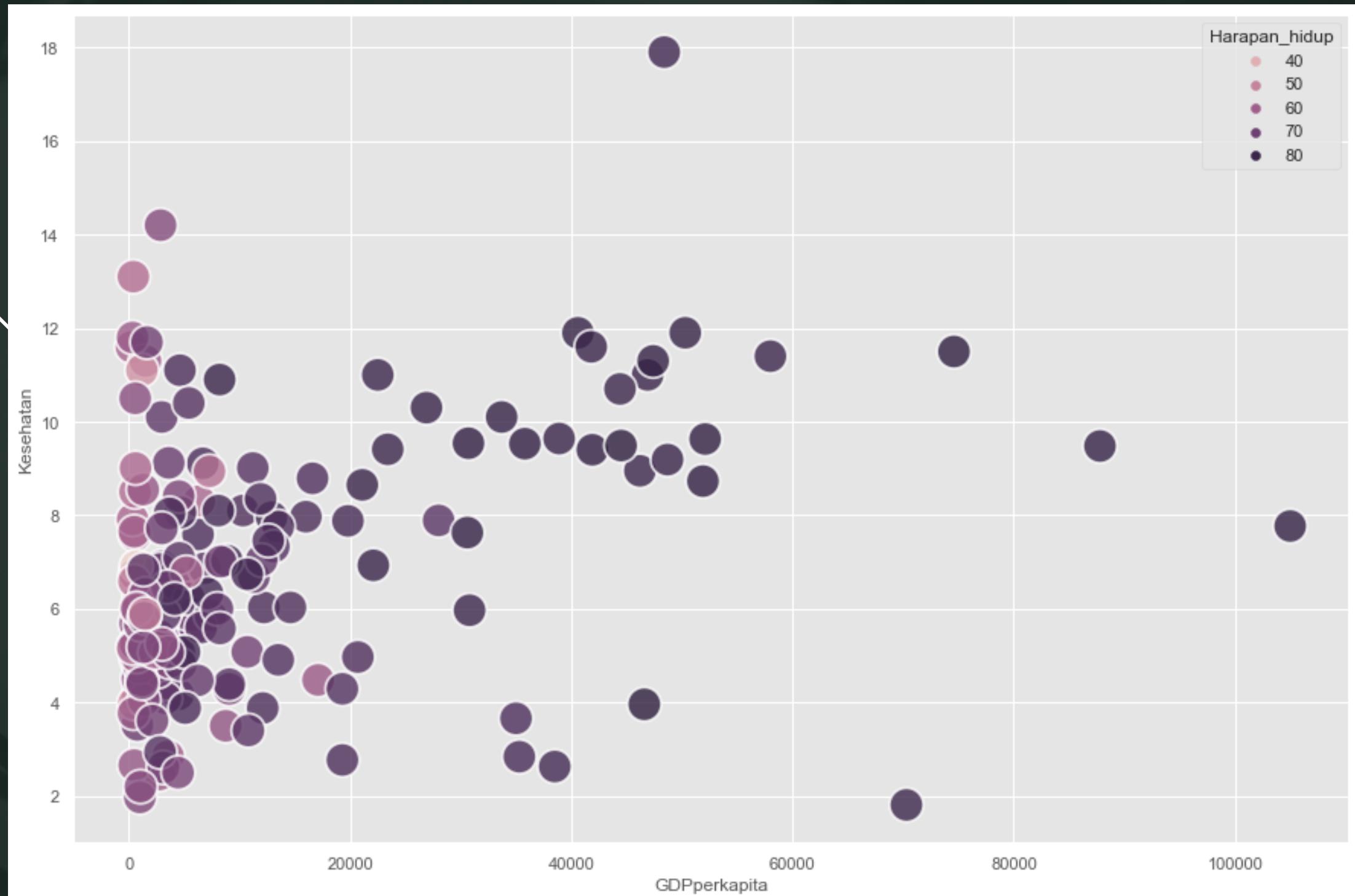
for idx, columns in zip(range(len(df_uni.columns)), df_uni.columns):
    ax = fig.add_subplot(3, 3, idx+1)
    sns.histplot(df[columns])

✓ 1.8s
```

## 1. Univariate Analysis

- Dalam analisis ini, saya menggunakan Count plot untuk visualisasi. Plotting ini digunakan untuk menganalisa sebaran data
- Pada Gambar 1.0 dapat dilihat bahwa 8 dari 9 data membentuk skew ke kanan, dan 1 data skew ke kiri
- Hal ini menandakan bahwa semakin besar value data, persebarannya semakin sedikit

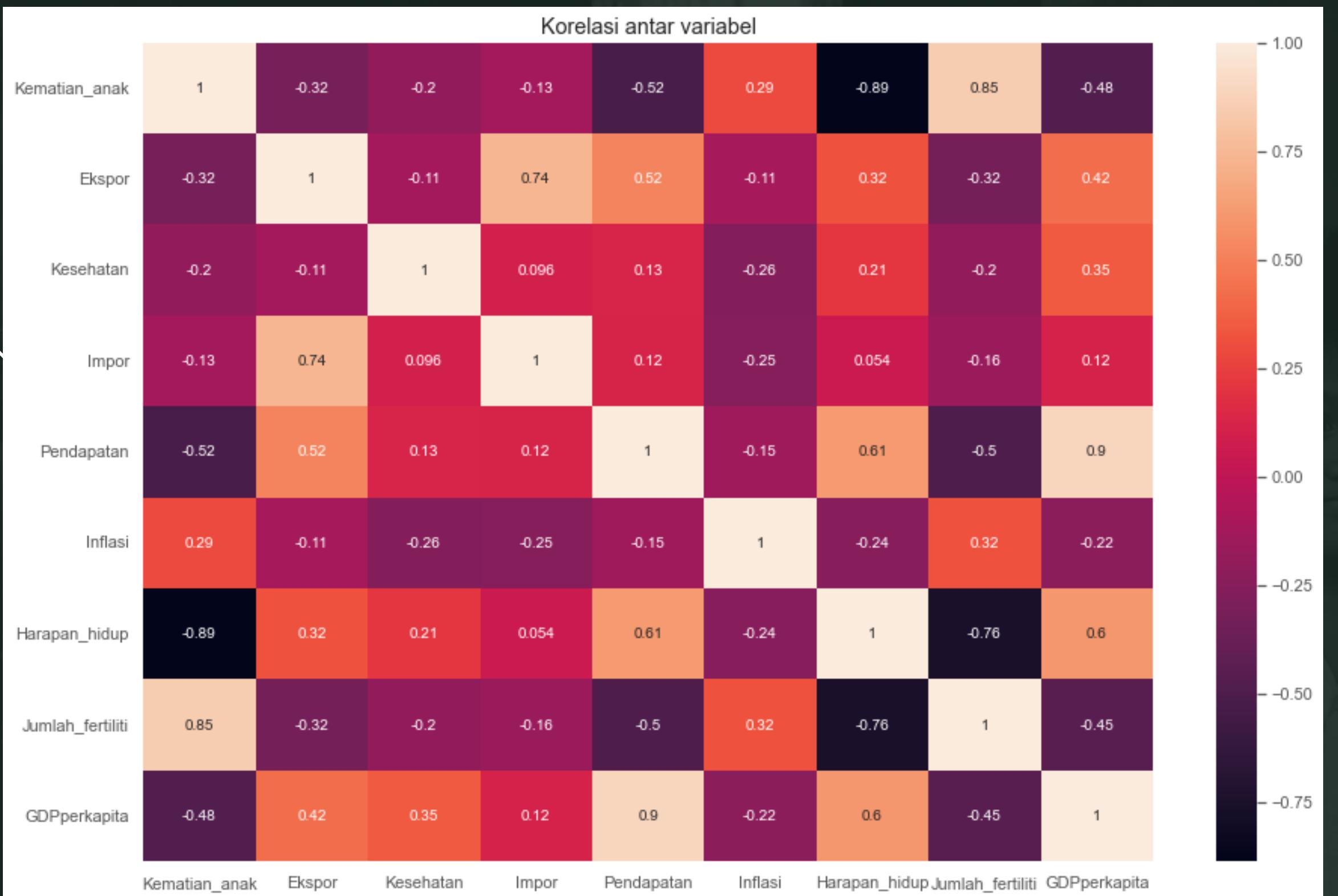
```
sns.scatterplot(x = 'GDPperkapita', y = 'Kesehatan', hue= 'Harapan_hidup',data=df , s= 500, alpha=0.8)
```



GAMBAR 2.0

## 2. Bivariate Analysis

- Dalam analisis ini, saya menggunakan Scatter plot untuk visualisasi.
- Analisis ini digunakan untuk mengetahui hubungan (GDPperkapita dan Kesehatan) dengan indikasi warna yang merupakan variabel Harapan hidup
- Terlihat variabel kesehatan tidak terlalu memiliki keselarasan dengan GDPperkapita, namun angka harapan\_hidup cenderung selaras dengan GDPperkapita



```
sns.heatmap(df.corr(), annot=True)
plt.title('Korelasi antar variabel')
plt.show()

✓ 0.5s
```

### 3. Multivariate Analysis

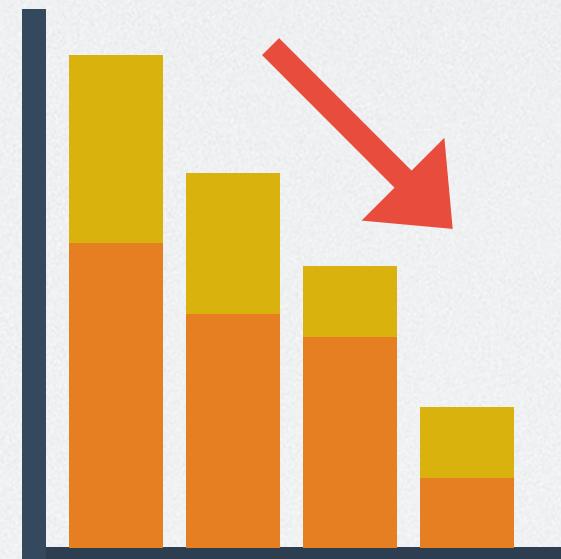
- Dalam analisis ini, saya menggunakan Heatmap dengan parameter dataframe correlation untuk visualisasi.
- Analisis ini diperlukan agar diketahui data dengan korelasi yang tinggi dan berhubungan dengan yang lain sehingga nanti dapat dipakai untuk analisa

**GAMBAR 3.0**

# DATA PREPROCESSING

3

# Menghapus beberapa features yang tidak diperlukan



Impor



Ekspor



Inflasi

Penghapusan 3 features tersebut karena memiliki korelasi yang rendah terhadap variabel lain dan tidak berhubungan dengan tujuan analisis.

# Outliers Treatment

- Remove outliers yang berada di atas upper bound untuk feature :
  1. GDPperkapita
  2. Harapan Hidup
  3. Kesehatan
- Remove outliers yang berada di bawah lowerbound untuk feature :
  1. Kematian Anak

## Kenapa memilih 4 feature itu untuk di remove ?

Keempat feature tersebut sudah mewakili 3 parameter, yakni faktor Ekonomi, Sosial, dan Kesehatan, sehingga negara yang menjadi outliers sudah tidak menjadi kandidat untuk mendapatkan bantuan.



# Lampiran code handling outliers

```
def upper_outliers(x):
    Q1 = df[x].quantile(0.25)
    Q3 = df[x].quantile(0.75)
    IQR = Q3-Q1
    upper_bound = Q3 + 1.5*IQR
    x_drop = df[df[x] > upper_bound]
    df.drop(df[df[x]>upper_bound].index,inplace=True)
    print("upper bound:",upper_bound)
    print("outliers:\n",x_drop['Negara'])
```

```
def lower_outliers(x):
    Q1 = df[x].quantile(0.25)
    Q3 = df[x].quantile(0.75)
    IQR = Q3-Q1
    lower = Q1 - 1.5*IQR
    x_drop = df[df[x] < lower]
    df.drop(df[df[x]< lower].index,inplace=True)
    print("upper bound:",lower)
    print("outliers:\n",x_drop['Negara'])
```

✓ 0.4s

```
upper_outliers('GDPperkapita')
upper_outliers('Kesehatan')
upper_outliers('Harapan_hidup')
lower_outliers(['Kematian_anak'])
```

✓ 0.3s

# **CLUSTERING & MAKING DECISION**

4

# Data Scaling

## MACHINE LEARNING

```
from sklearn import preprocessing
scaled = preprocessing.StandardScaler().
fit_transform(df.drop('Negara', axis=1))
df_scaled = pd.DataFrame(scaled, columns=
['Kematian_anak', 'Kesehatan',
'Pendapatan', 'Harapan_hidup',
'Jumlah_fertiliti','GDPperkapita[]])
```

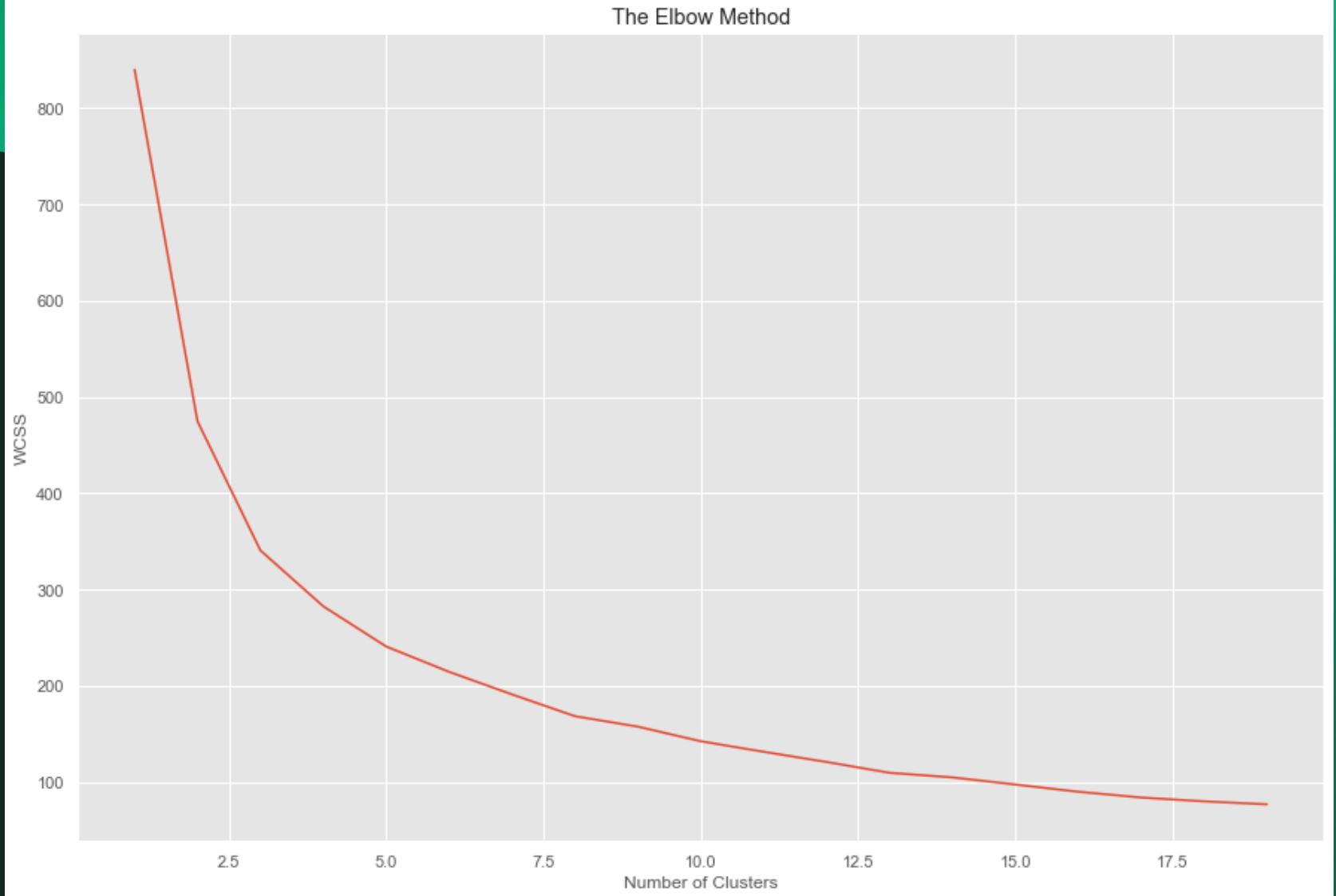
✓ 0.3s

Python

- Scaling data dengan cara import preprocessing dari scikit learn
- Tujuan dari langkah ini adalah menyederhanakan value dari tiap features, dari nilai 0 sampai 1.

```
from sklearn.cluster import KMeans
wcss= []
for i in range (1,20):
    kmeans = KMeans(n_clusters=i, init='k-means++',random_state = 42)
    kmeans.fit(df_scaled)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,20),wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```

✓ 1.5s



# Data Clustering

## ELBOW METHOD

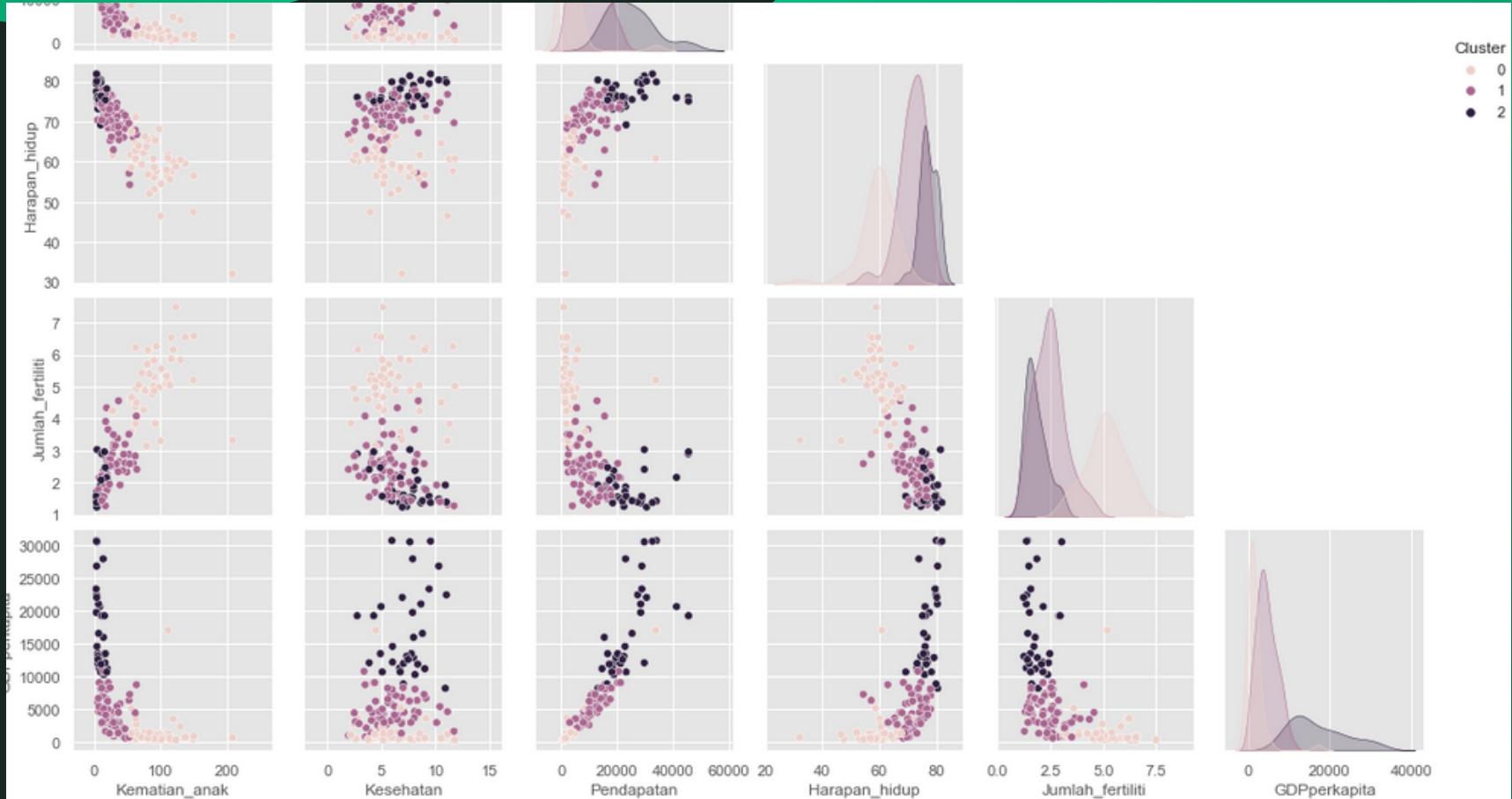
- Tujuan dari penggunaan metode ini adalah untuk mengetahui mana n cluster yang paling optimal dalam clustering data
- Didapat n = 3 dengan perpotongan sudut yang lebih besar daripada yg lain

```
kmeans = KMeans(n_clusters=3)
kmeans.fit(df_scaled)
df['Cluster'] = kmeans.labels_
df
```

✓ 0.1s

	Negara	Kematian_anak	Kesehatan	Pendapatan	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	Cluster
0	Afghanistan	90.2	7.58	1610	56.2	5.82	553	0
1	Albania	16.6	6.55	9930	76.3	1.65	4090	1
2	Algeria	27.3	4.17	12900	76.5	2.89	4460	1
3	Angola	119.0	2.85	5900	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	6.03	19100	76.8	2.13	12200	2
...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	5.25	2950	63.0	3.50	2970	1
163	Venezuela	17.1	4.91	16500	75.4	2.47	13500	2
164	Vietnam	23.3	6.84	4490	73.1	1.95	1310	1
165	Yemen	56.3	5.18	4480	67.5	4.67	1310	0
166	Zambia	83.1	5.89	3280	52.0	5.40	1460	0

140 rows × 8 columns



# Data Clustering

NUMBER OF CLUSTERS = 3

- Cluster target kita adalah cluster 0, yang mana cocok dengan data kita yang mencari negara yang membutuhkan bantuan

# 5 Negara yang membutuhkan bantuan \$ 10 Juta

		Negara	Kematian_anak	Kesehatan	Pendapatan	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	Cluster
37		Congo, Dem. Rep.	116.0	7.91	609	57.5	6.54	334	0
88		Liberia	89.3	11.80	700	60.8	5.02	327	0
26		Burundi	93.6	11.60	764	57.7	6.26	231	0
112		Niger	123.0	5.16	814	58.8	7.49	348	0
31		Central African Republic	149.0	3.98	888	47.5	5.21	446	0

- Penarikan kesimpulan berdasarkan data dengan feature "Cluster" bernilai 0
- Data ini di sorting berdasarkan value pendapatan dari yang terkecil