Clustering the Countries by using K-Means for HELP International

Bagas Eko Tjahyono Putro

Latar Belakang

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.

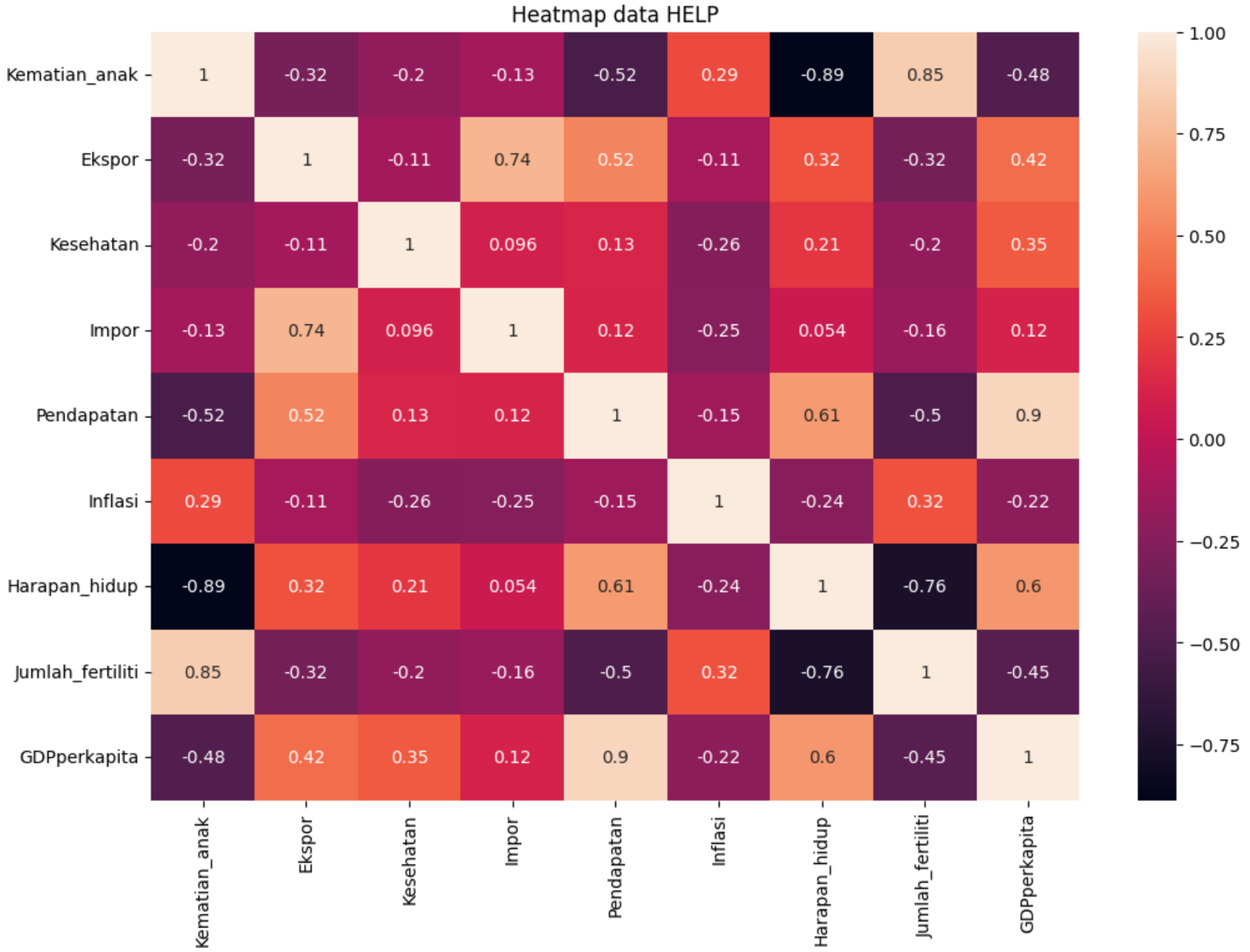
Dalam upaya untuk memutuskan alokasi dana yang sebesar $10 juta secara strategis dan efektif, CEO HELP International perlu mengambil keputusan yang berdasarkan analisis mendalam terhadap faktor-faktor sosial ekonomi dan kesehatan yang mempengaruhi perkembangan negara-negara secara keseluruhan. Tujuan utama dari keputusan ini adalah untuk memilih negara-negara yang paling membutuhkan bantuan dan memiliki potensi besar untuk mengalami kemajuan signifikan melalui intervensi kemanusiaan.

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.

Terdapat 167 baris dan 10 kolom

Visualisasikan data dengan *heatmap* untuk menunjukkan korelasi antar variabel



Feature Selection

Disini saya memilih fitur Pendapatan dan Kesehatan untuk menjadi dasar analisis dan clustering yang saya lakukan.

Alasan saya memilih dua fitur ini karena tingkat pendapatan dan tingkat kesehatan merupakan kunci utama dalam pembangunan di suatu negara dan juga dua fitur ini dapat menjadi indikator kuat untuk menentukan prioritas bantuan kepada negara tertentu, sehingga alokasi bantuan tepat sasaran.

Check dan Handling missing values

```
# cek data
df.isnull().sum()
```

✓ 0.0s

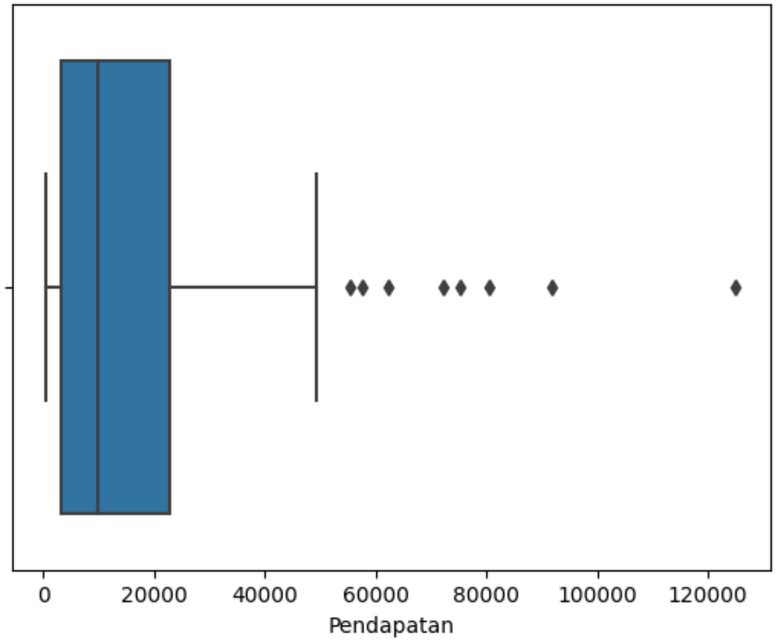
```
Kematian_anak        0
Ekspor               0
Kesehatan            0
Impor                0
Pendapatan           0
Inflasi              0
Harapan_hidup        0
Jumlah_fertiliti     0
GDPperkapita         0
dtype: int64
```

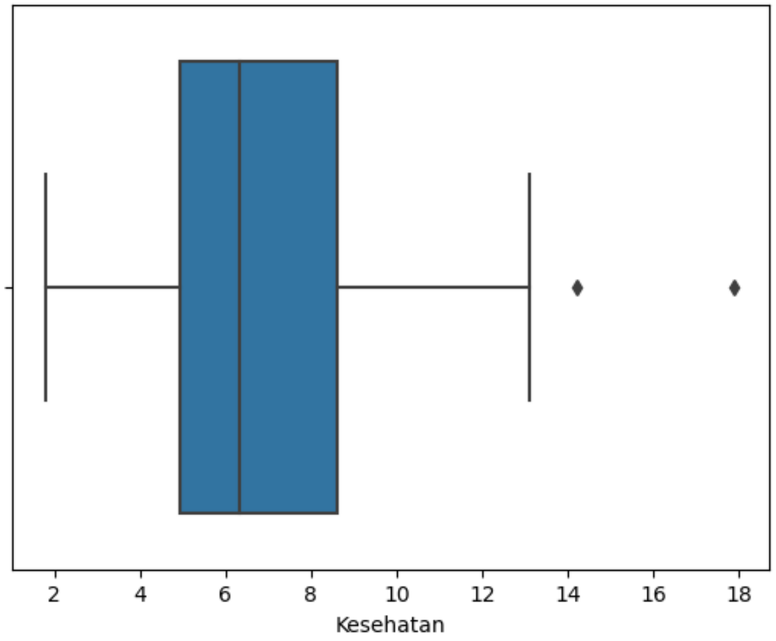
Hasil pengecekan missing value pada semua kolom menunjukan nilai 0, hal ini dapat diartikan bahwa tidak terdapat data yang hilang atau kosong pada setiap kolom.

Check dan Handling Outliers

```
# cek outlier
sns.boxplot(x = df_negara['Pendapatan'])
sns.boxplot(x = df_negara['Kesehatan'])
```

✓ 0.1s



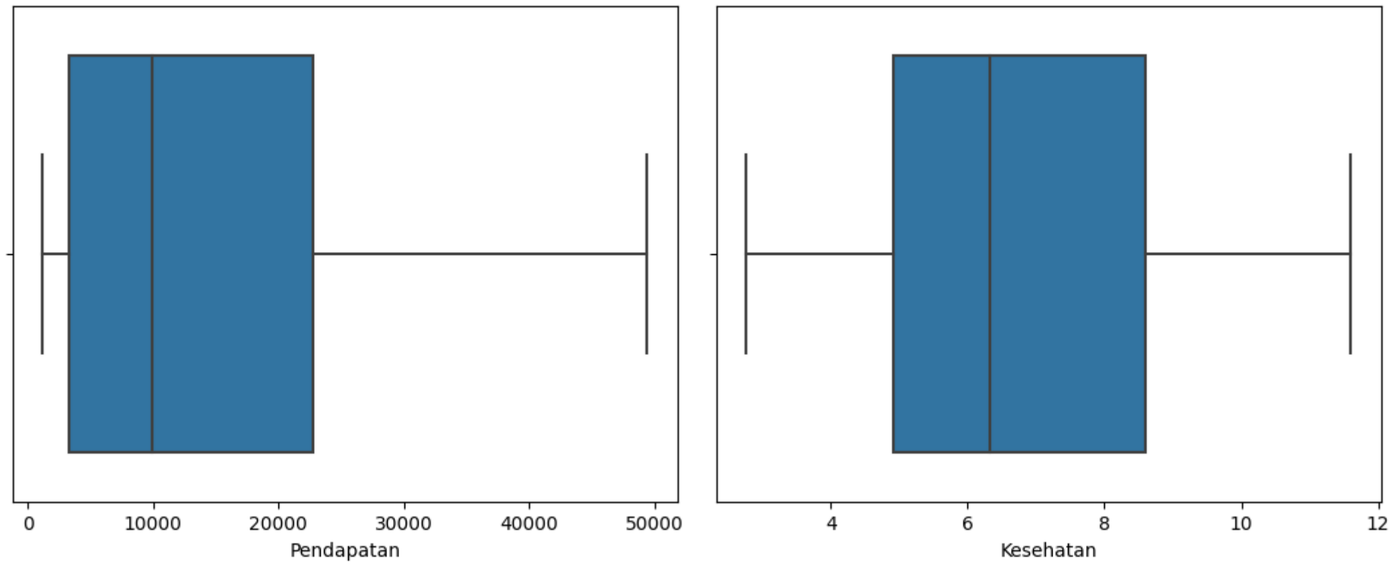


Untuk mengecheck ada atau tidaknya outliers, saya menggunakan boxplot karena boxplot dapat menvisualisasikan outliers dan juga agar lebih cepat dalam menganalisis data.

Check dan Handling Outliers

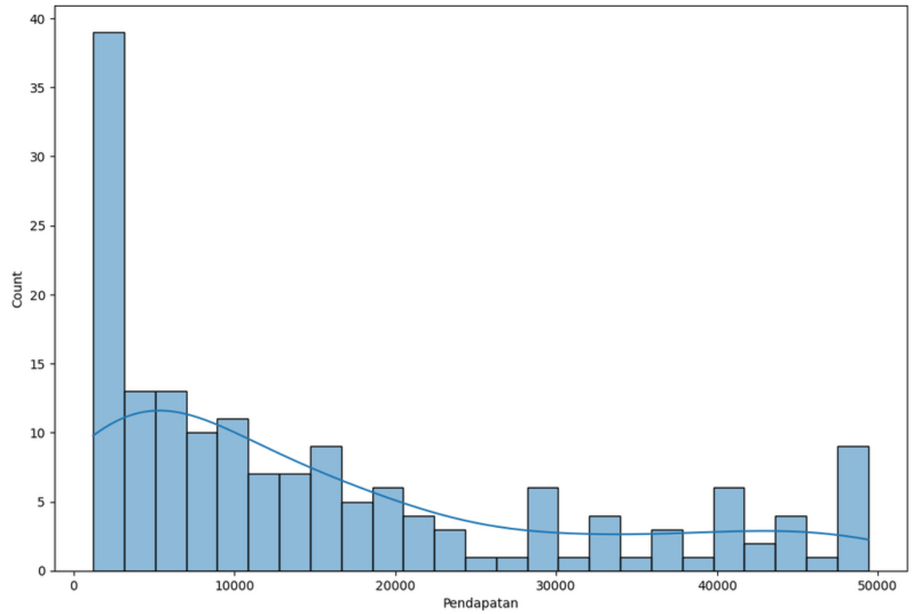
```
# Handling outliers dengan winsorize
from scipy.stats.mstats import winsorize

dfbaru =pd.DataFrame(data = df_negara, columns = ['Negara','Pendapatan', 'Kesehatan'])

dfbaru['Pendapatan'] = winsorize(dfbaru['Pendapatan'], limits =[0.05, 0.05])
dfbaru['Kesehatan'] = winsorize(dfbaru['Kesehatan'], limits = [0.05, 0.05])

country = dfbaru.set_index('Negara')
country
```

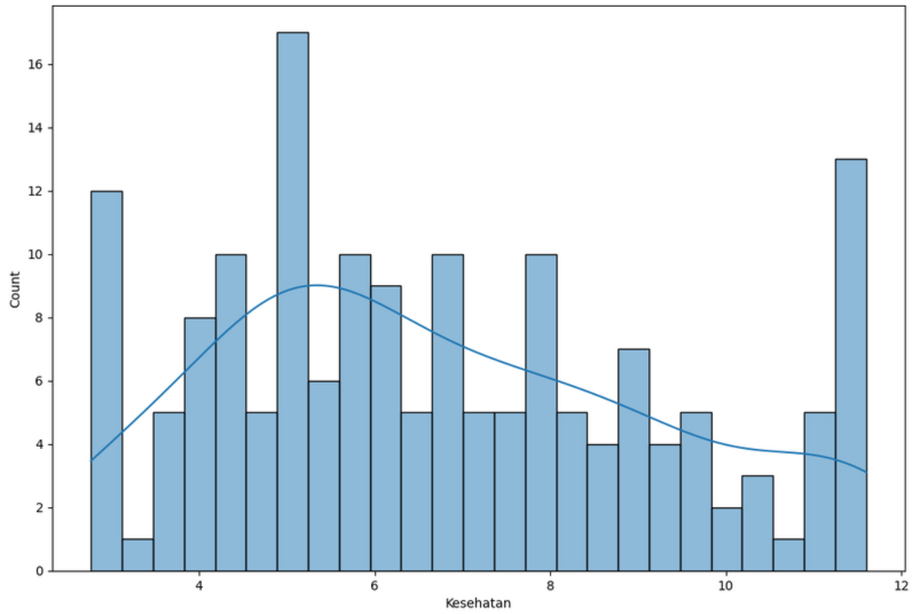
✓ 0.0s



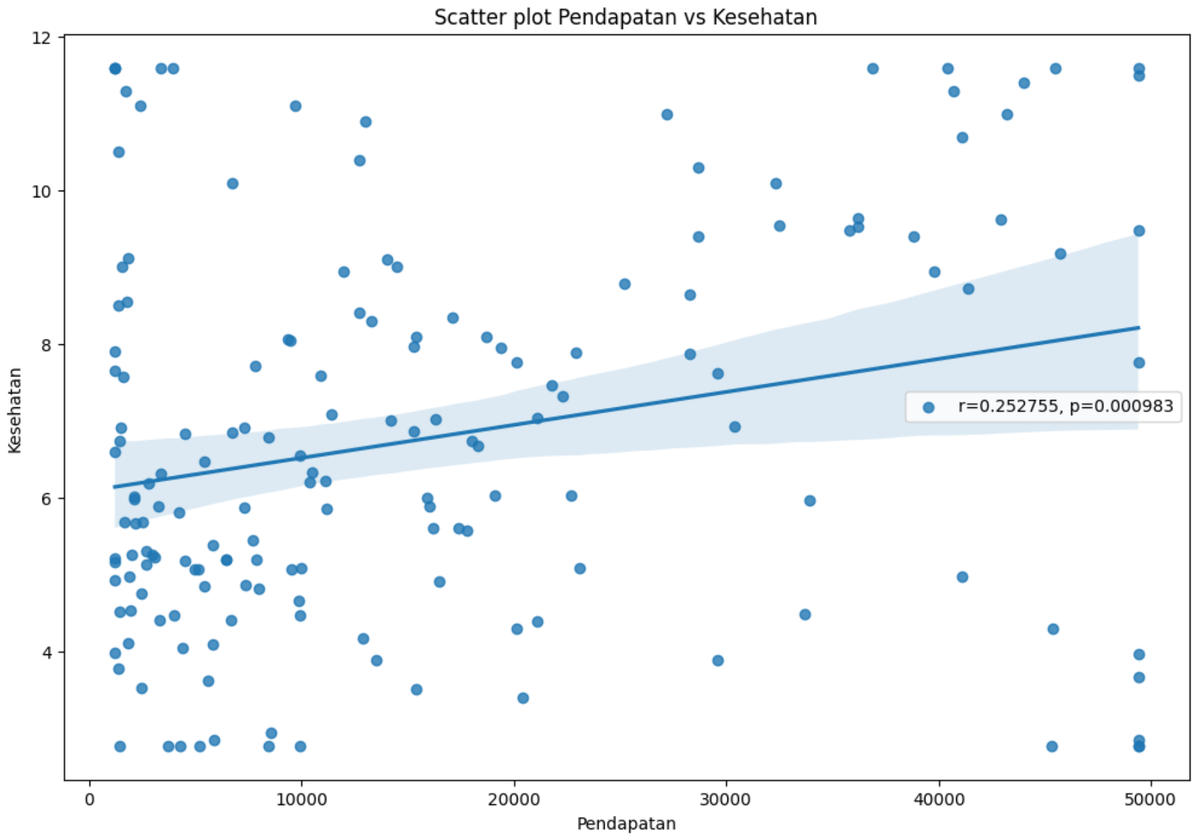
disini saya menggunakan metode winsorize untuk menghandle outliers, karena winsorization menggantikan nilai outliers dengan nilai-nilai yang berada pada batas tertentu (biasanya persentil tertentu) dari distribusi data, sehingga mengurangi dampak ekstrim dari outliers.

Univariate analysis



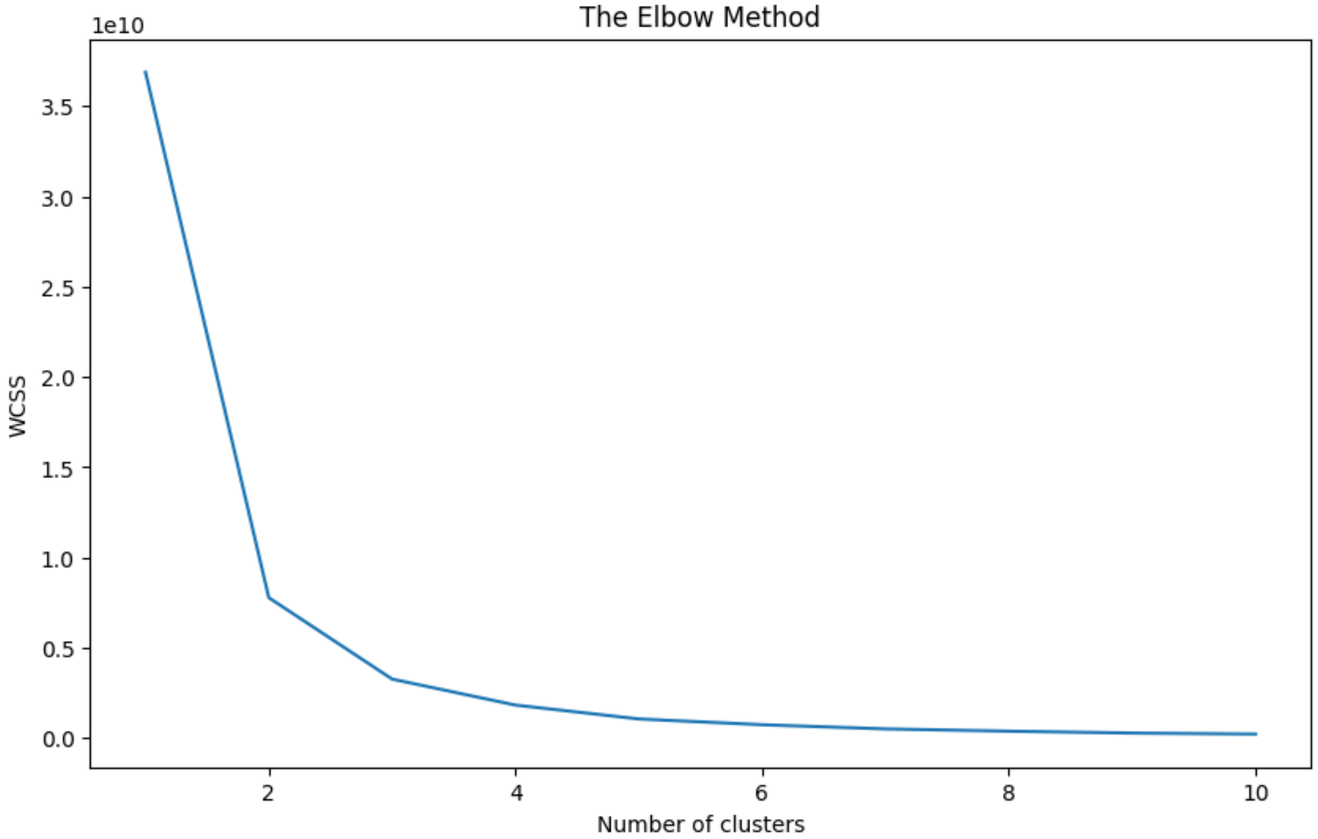


Bivariate analysis



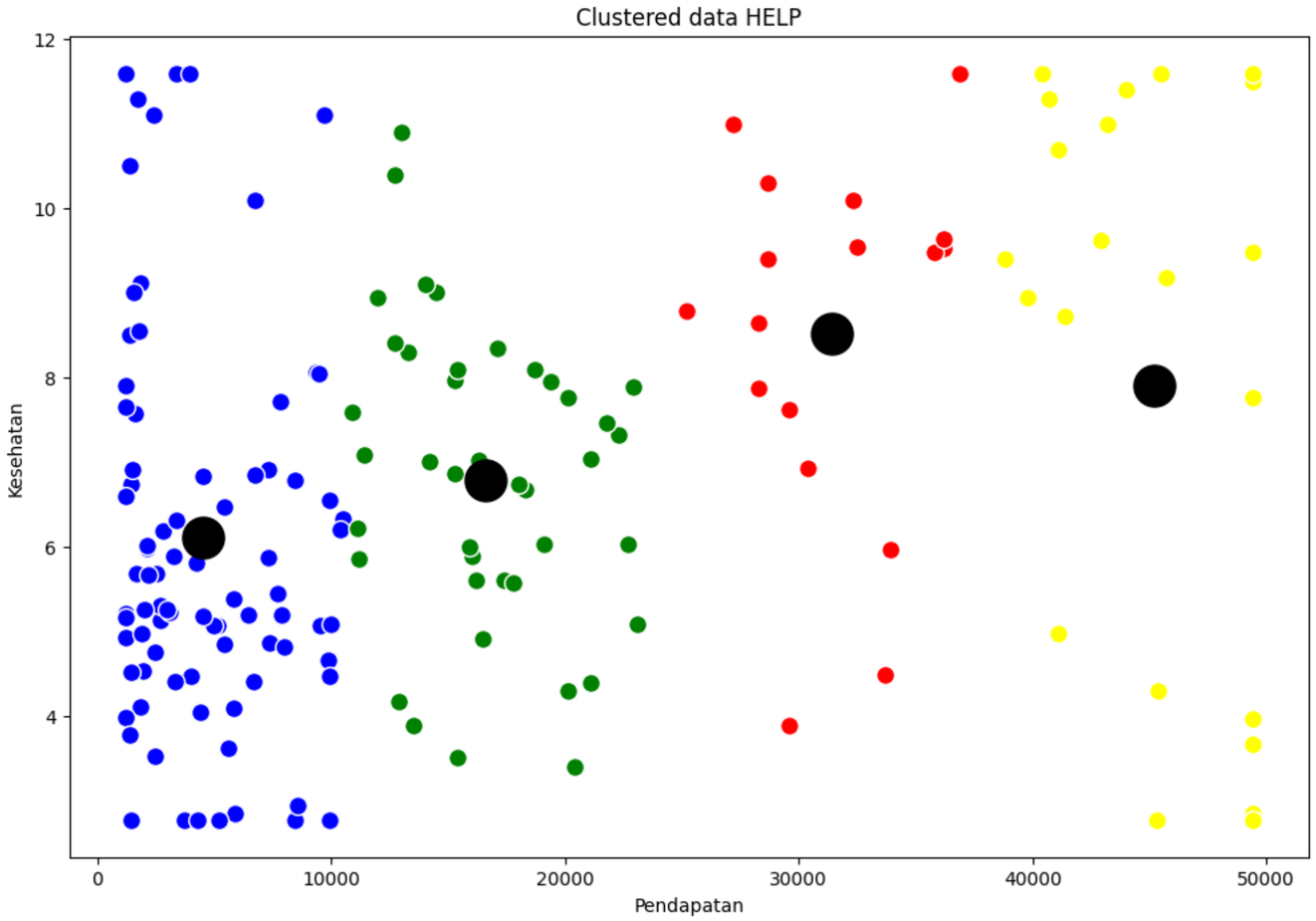
Hasil uji pearsonr menunjukkan nilai sebesar 0.252, nilai ini dapat diartikan bahwa variabel pendapatan dan kesehatan suatu negara memiliki korelasi yang cukup rendah.

Menentukan jumlah cluster dengan Elbow Method



Berdasarkan perhitungan dengan Elbow method, diketahui jumlah cluster yang baik untuk digunakan dalam clustering negara adalah 3. nilai 3 diperoleh dari interpretasi grafik dimana garis mulai menyiku pada angka 3 sehingga jumlah cluster yang efektif adalah 3 cluster.

K-Means Clustering



Berdasarkan grafik diatas, negara-negara dapat dibagi kedalam 3 cluster berdasarkan pendapatan dan harapan hidup masing-masing negara, yaitu :

Cluster 1 (biru) : Negara dengan indeks pembangunan rendah.

Cluster 2 (hijau) : Negara dengan indeks pembangunan sedang.

Cluster 3 (merah) : Negara dengan indeks pembangunan sedang.

Cluster 4 (kuning) : Negara dengan indeks pembangunan tinggi

Matriks yang butuh bantuan

Negara	Pendapatan_asli	Kesehatan_asli
Niger	814.0	5.16
entral African Republic	888.0	3.98
Mozambique	918.0	5.21
Guinea	1190.0	4.93
Madagascar	1390.0	3.77
Comoros	1410.0	4.51
Eritrea	1420.0	2.66
Gambia	1660.0	5.69
Benin	1820.0	4.10
Mali	1870.0	4.98

Data yang ditampilkan merupakan data yang telah dihitung berdasarkan rata-rata

Rekomendasi Negara

NIGER

Merupakan sebuah negara di Afrika Barat yang secara resmi dikenal sebagai Republik Niger. Ibu kota negara ini adalah Niamey

CENTRAL AFRICAN REPUBLIC

Republik Afrika Tengah (Central African Republic) adalah sebuah negara yang terletak di jantung Afrika.

MOZAMBIQUE

Merupakan sebuah negara di wilayah bagian tenggara Afrika. Ibu kota negara ini adalah Maputo