Final Project Report

Clustering the Countries by Using K-Means for HELP International

Fajrur Rahman Suprapto

Tentang Organisasi

HELP International adalah LSM kemanusiaan internasional yang berkomitmen untuk memerangi kemisikinan 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.

Analisis Data

Import Library

Lakukan import library yang diperlukan untuk melakukan analisa data.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import warnings

warnings.filterwarnings('ignore')

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

Analisis Data

Reading and Understanding Data

Terdapat 10 kolom dan 167 baris pada dataset ini.

```
<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
```

Analisis Data

Exploratory Data Analysis

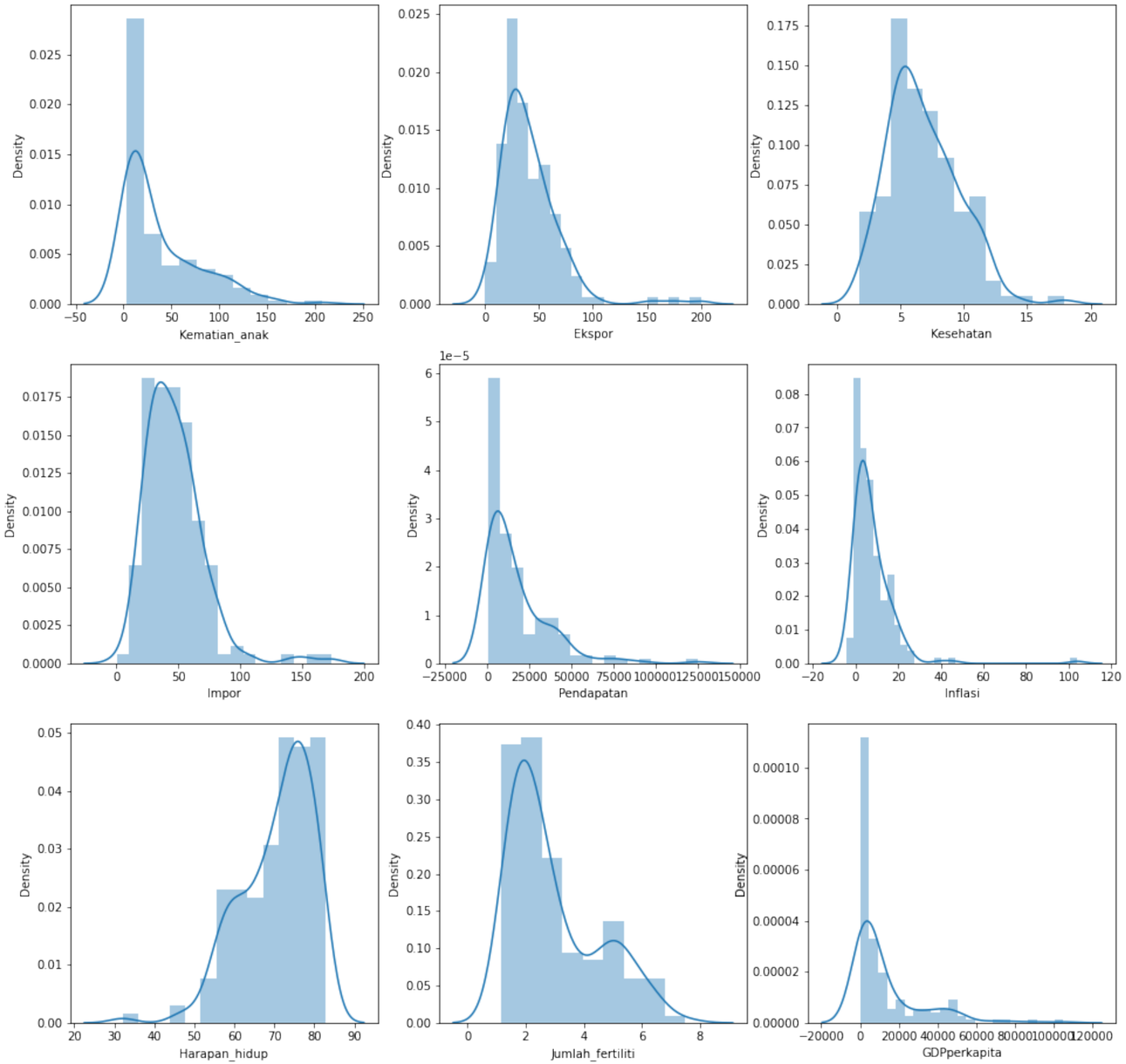
- Data Cleaning
  Setelah dilakukan pengecekan terhadap missing value, dataset ini tidak memiliki missing value.

```
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
```

Analisis Data

Exploratory Data Analysis

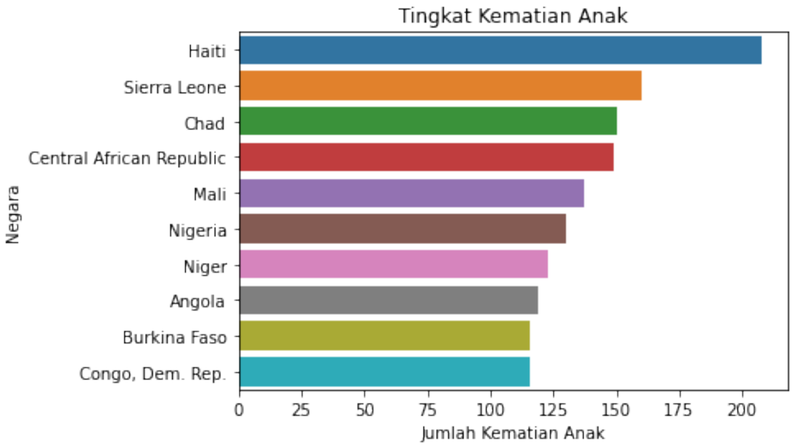
- Univariate Analysis
  Semua fitur mempunyai sebaran data yang positif, kecuali fitur Harapan_hidup.

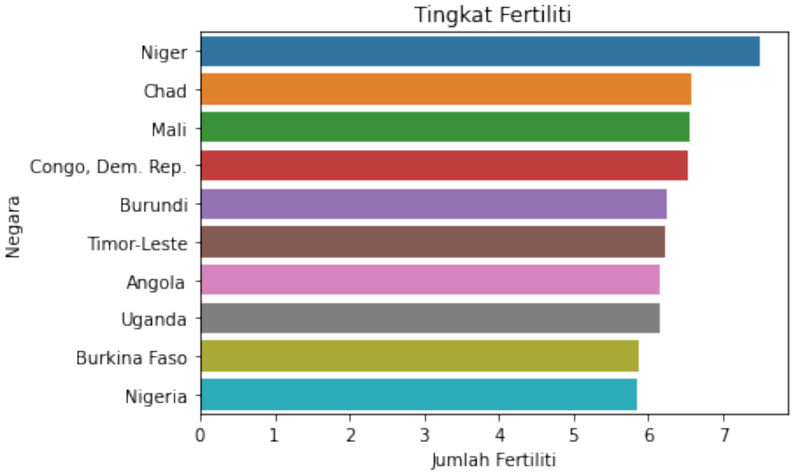


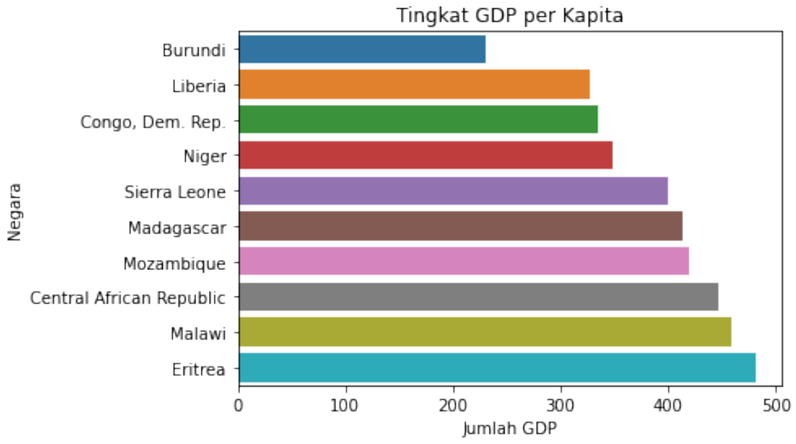
Analisis Data

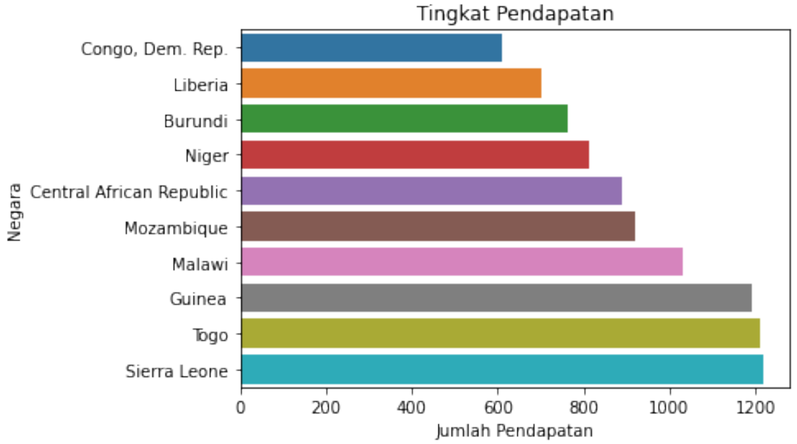
Exploratory Data Analysis

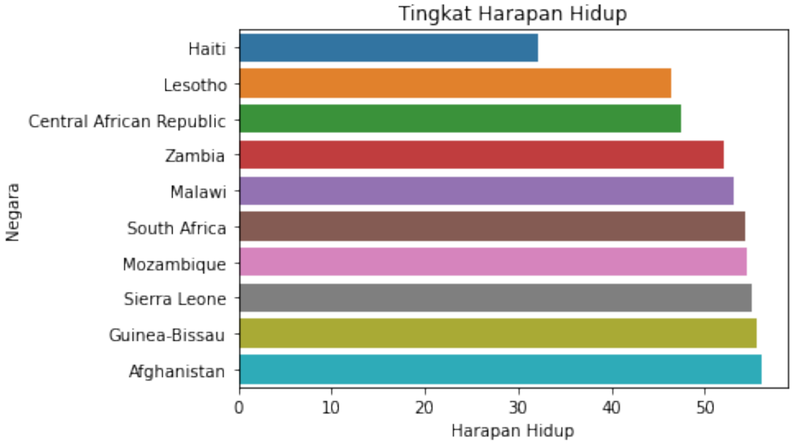
- Bivariate Analysis
  Fitur yang dibandingkan adalah Negara dengan Kematian_anak, Jumlah_Fertiliti, Harapan_hidup, Pendapatan, dan GDPperkapita.







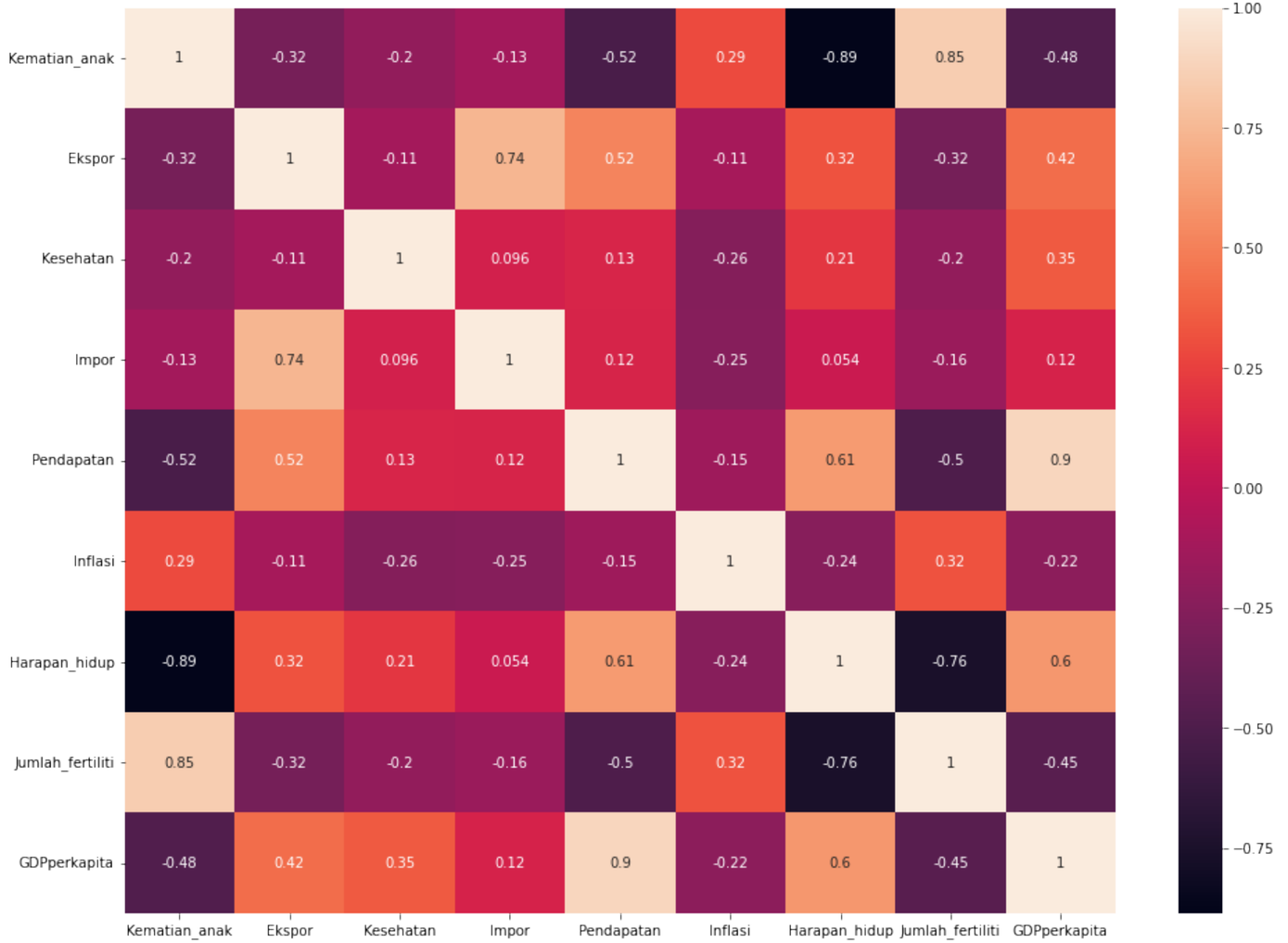




Analisis Data

Exploratory Data Analysis

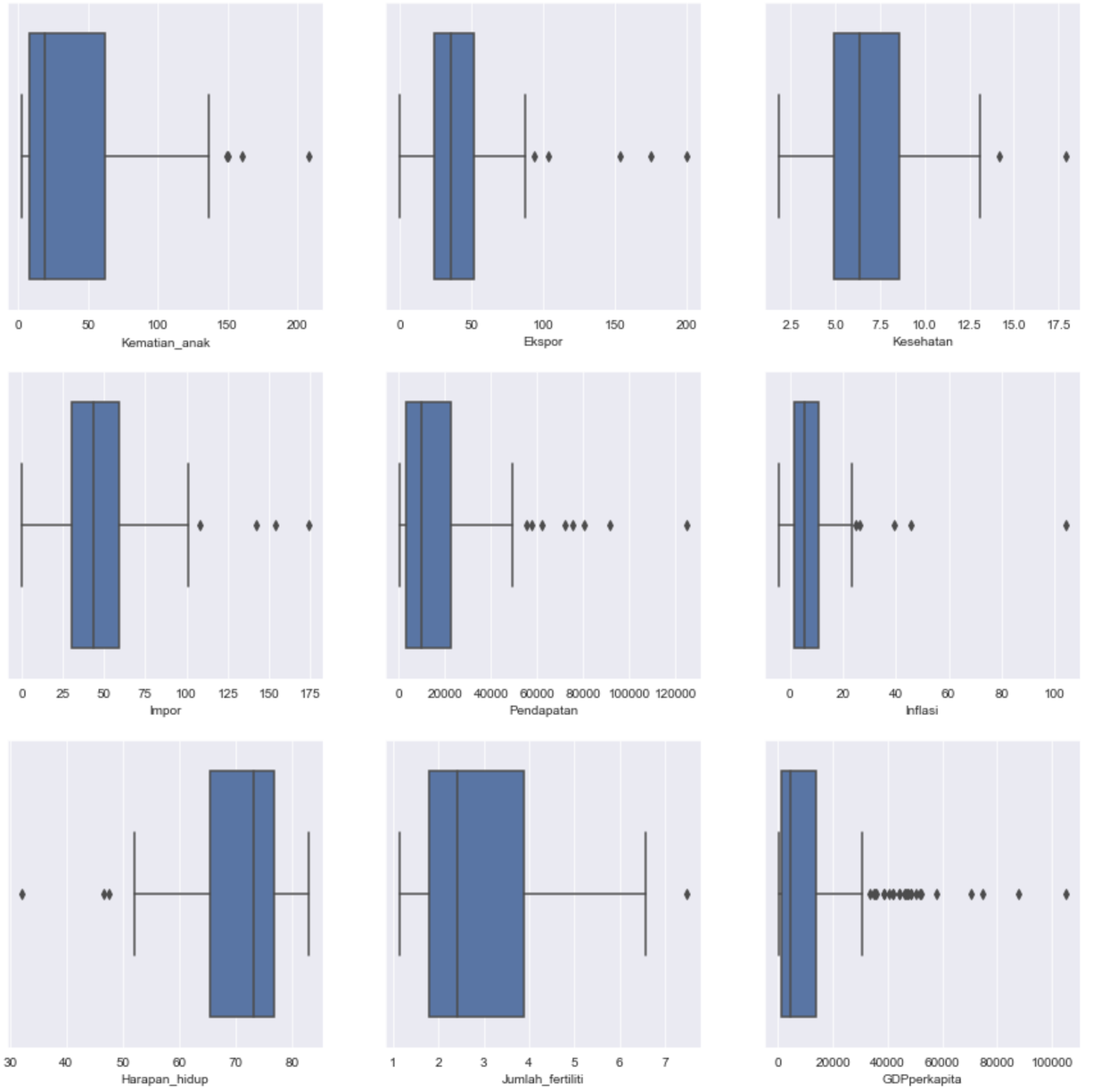
- Multivariate Analysis
  Berdasarkan heatmap di samping, fitur GDPperkapita dengan Pendapatan memiliki korelasi yang paling tinggi.



Analisis Data

Outliers Treatment

Terdapat *outliers* pada fitur-fitur yang bertipe float dan int. *Outliers* tersebut harus di-*handling* terlebih dahulu karena akan membuat hasil klasifikasi menjadi tidak akurat. Metode yang dilakukan adalah filtering data menggunakan Interquartile Range.



Analisis Data

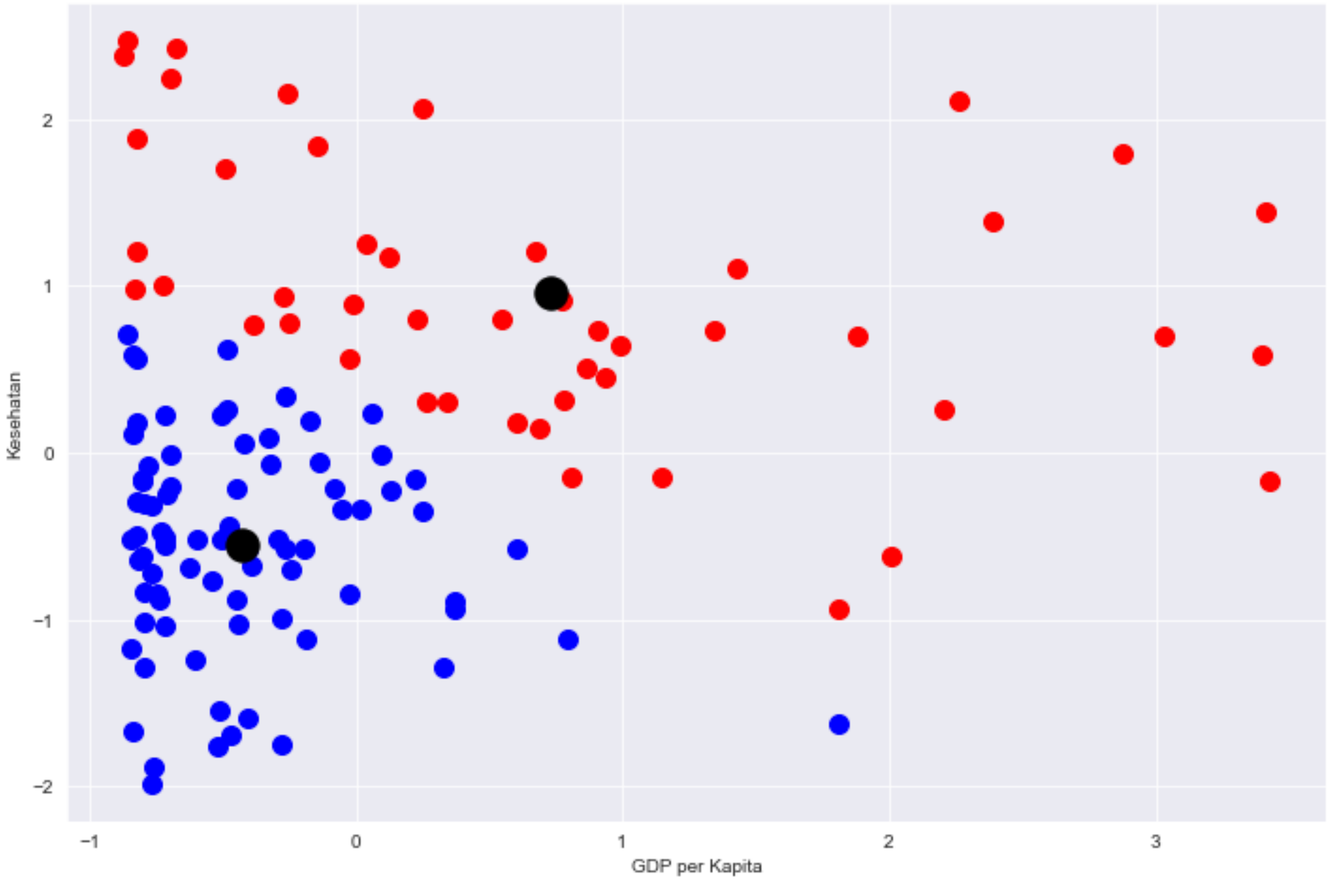
Scaling Data

Scaling Data dilakukan agar perbandingan persebaran data antar fitur mempunyai nilai yang sama. Metode yang dilakukan adalah menggunakan metode *Standard Scaling* dari *library sklearn*.

Analisis Data

Creating K-Means Clustering

K-Means digunakan untuk mengelompokkan negara. Grafik di samping merupakan hasil K-Means dengan 2 *clusters* dengan membandingkan fitur GDPperkapita dan Kesehatan.



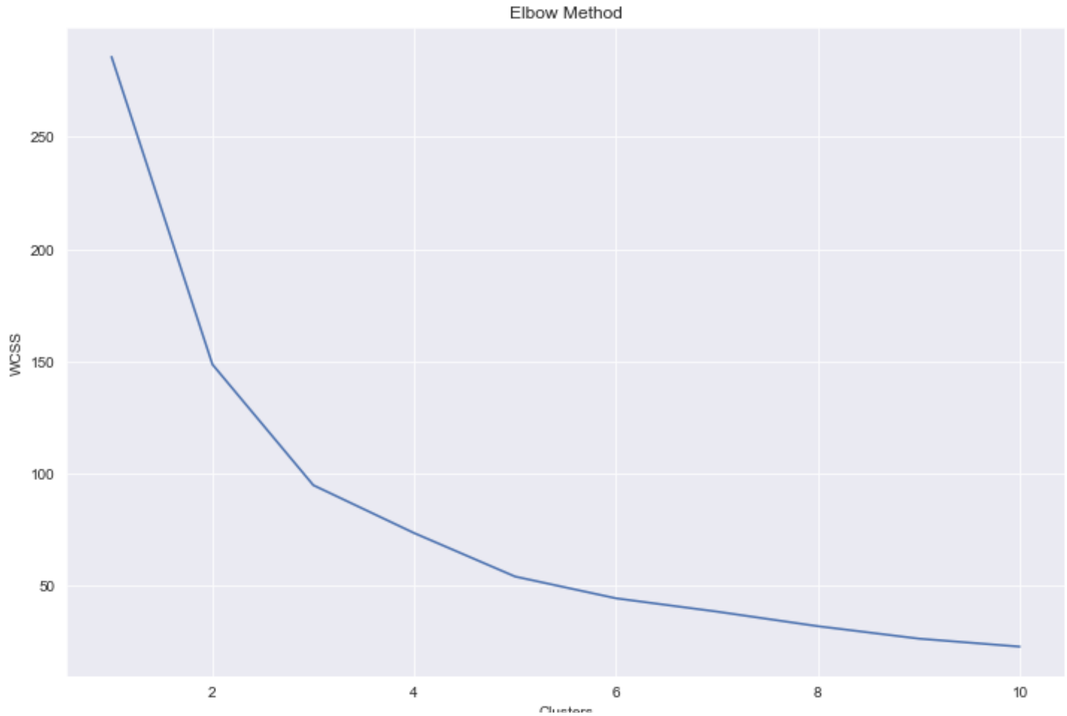
Analisis Data

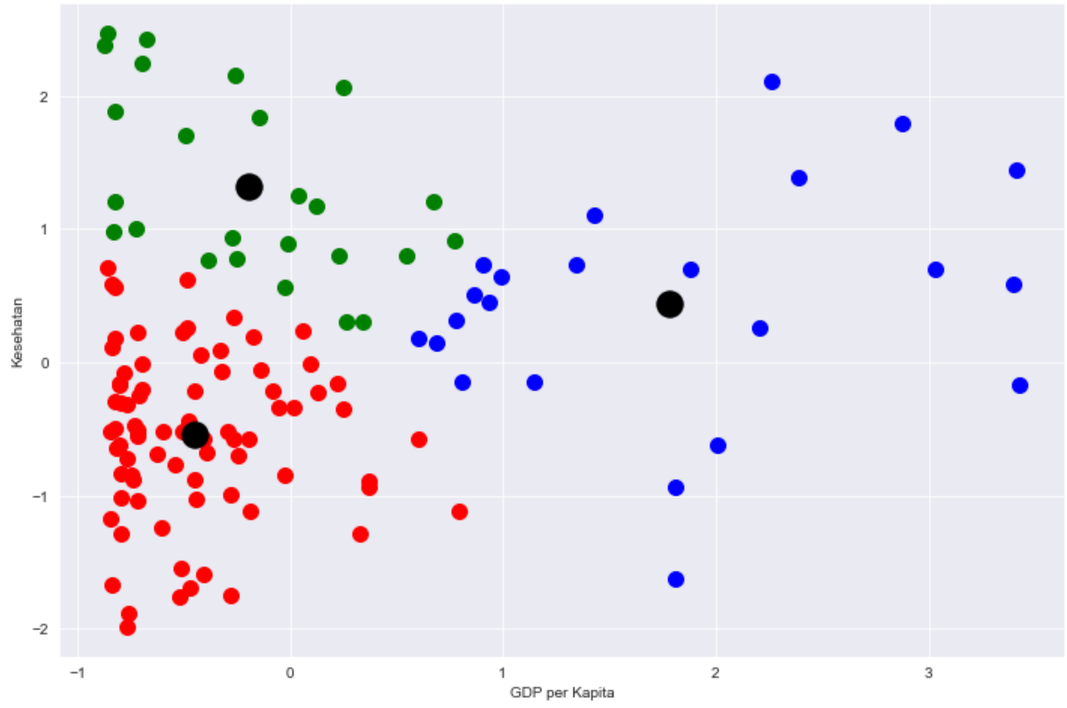
Finding Best Number of Cluster

Untuk mendapatkan jumlah *cluster* yang optimal, dapat digunakan *elbow method*. Pada grafik *elbow method*, didapatkan 3 *clusters* merupakan jumlah *cluster* yang optimal.

Grafik di samping merupakan hasil K-Means dengan menggunakan 3 *clusters*.

Cluster merah (1) menunjukkan negara berkembang, *cluster* hijau (2) menunjukkan negara menengah, dan *cluster* biru (0) menunjukkan negara maju.





Analisis Data

Report Countries

Tabel di samping merupakan 10 negara yang termasuk ke dalam *cluster* negara berkembang dan mempunyai tingkat GDP dan kesehatan yang rendah.

	Negara	GDPperkapita	Kesehatan	Cluster
0	Liberia	327	11.80	1.0
1	Congo, Dem. Rep.	334	7.91	1.0
2	Niger	348	5.16	1.0
3	Central African Republic	446	3.98	1.0
4	Eritrea	482	2.66	1.0
5	Afghanistan	553	7.58	1.0
6	Gambia	562	5.69	1.0
7	Rwanda	563	10.50	1.0
8	Burkina Faso	575	6.74	1.0
9	Nepal	592	5.25	1.0

Kesimpulan

Dari hasil *clustering* yang sudah dilakukan, maka dapat diambil kesimpulan bahwa negara-negara yang paling membutuhkan bantuan dari HELP International adalah Liberia, Congo, Niger, Central African, Eritrea, Afghanistan, Gambia, Rwanda, Burkina Faso, dan Nepal.