/00

HELP International

Clustering the Countries by using K-Means for **HELP International**

Humanity Project

HELP International | 2022

Business/Project Understanding

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. HELP International telah mendapatkan $10 juta, untuk bantuan negara yang kesulitan. Client ingin melihat negara yang paling membutuhkan, agar bantuan dapat dipergunakan secara strategis dan efektif.

Tujuan

HELP International ingin mengkategorikan negara menggunakan faktor Sosial, Ekonomi dan Kesehatan. Kemudian menentukan pembangunan negara yang memiliki kesulitah tersebut secara keseluruhan

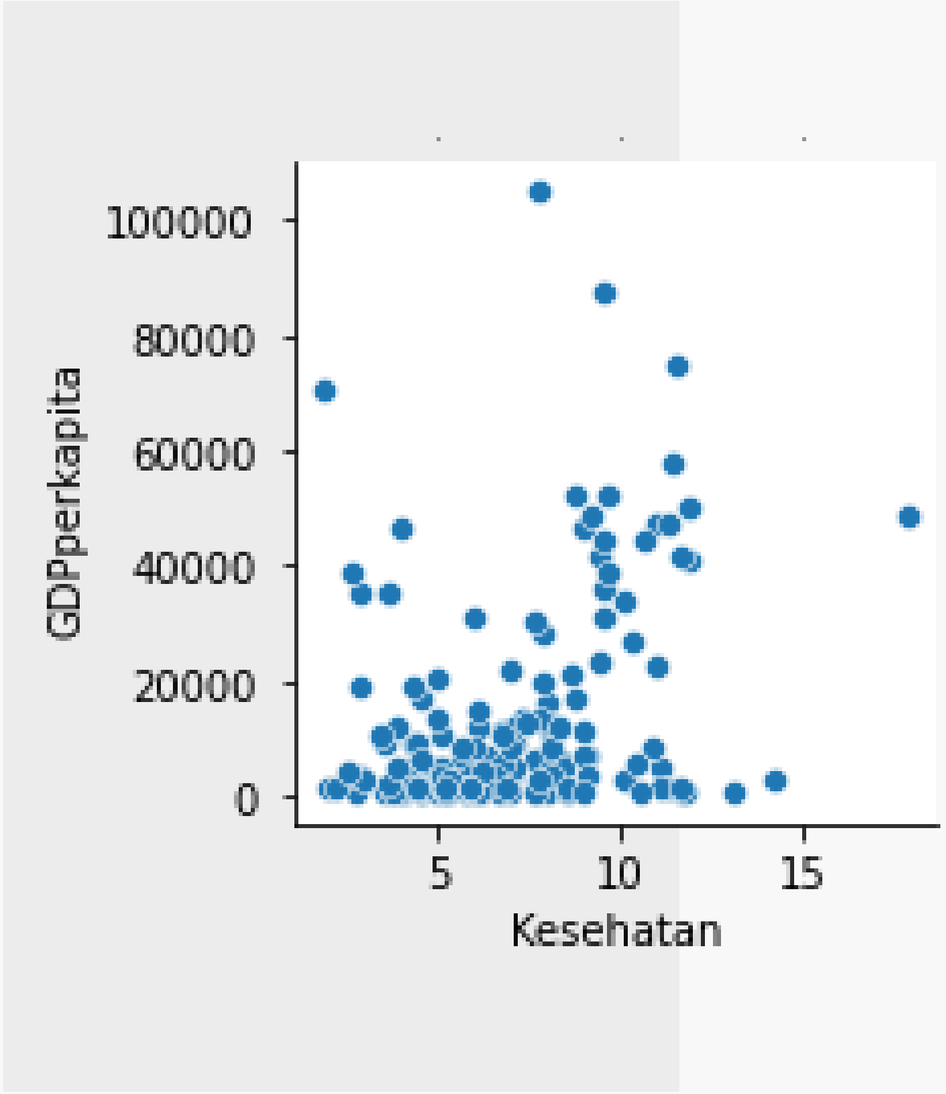
The Data

Dataset Understanding

Profil Data

Data yang tersedia memiliki 10 variabel, dengan 167 baris data. Variabelnya yaitu: Negara, Kematian_anak, Ekspor, Kesehatan, Impor, Pendapatan, Inflasi, Harapan_hidup, Jumlah_fertiliti, GDPperkapita.

/02



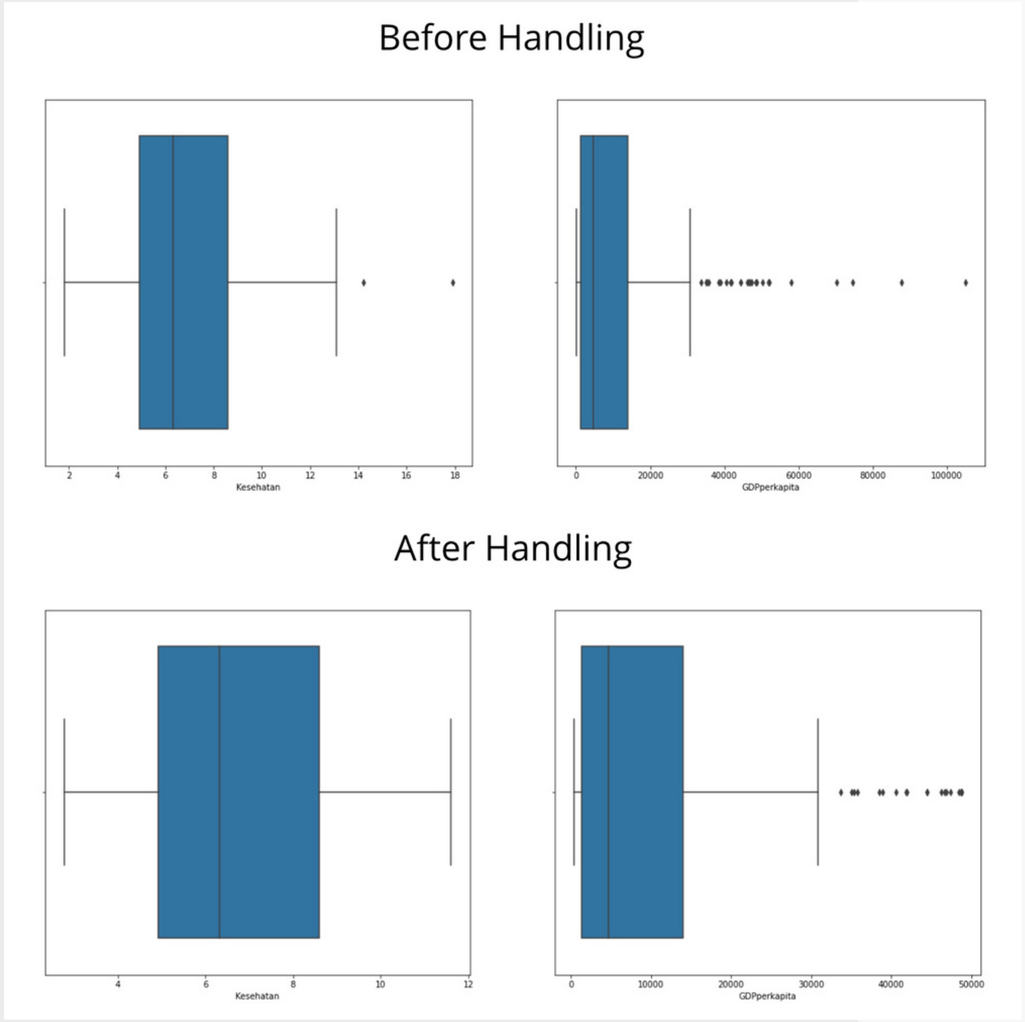
Pemilihan Variabel

Disini kami melakukan pairplot untuk menganalisis keterkaitan antar variabel. Selanjutnya kami memilih variabel Kesehatan dan GDP perkapita. Hal tersebut karena menurut kami pengeluaran yang digunakan untuk kepentingan kesehatan sangat menunjukkan keadaan suatu negara tersebut, baik digunakan untuk pengobatan maupun pencegahan. Selain itu, satuan kedua variabel tersebut sama, yaitu perkapita. Anda bisa melihat visualisasi pairpot kami disini

Data Cleaning

Disini kami melakukan data cleaning, tahap pertama yaitu mengatasi missing value. Setelah dilakukan pengecekan missing value, tidak ditemukan satupun missing value untuk data yang akan kami analisis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Negara         167 non-null    object
 1   Kesehatan      167 non-null    float64
 2   GDPperkapita   167 non-null    int64
dtypes: float64(1), int64(1), object(1)
memory usage: 4.0+ KB
Kesehatan           -0.6
GDPperkapita   -17750.0
dtype: float64
Kesehatan           14.12
GDPperkapita    33130.00
dtype: float64
```



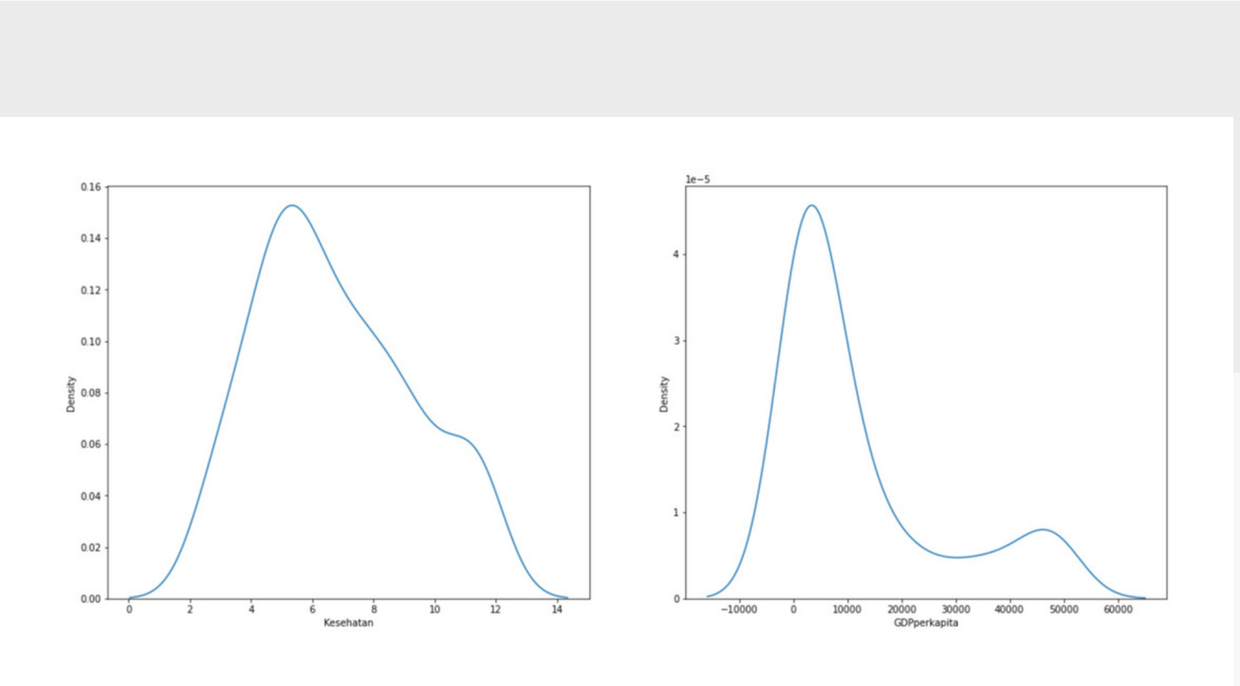
Handling Outliers

Selanjutnya kami mengatasi outliers yang ada pada dataset. Kami mengatasi outliers dengan cara membatasi nilai outliers sebesar 5% di lower dan upper bound, dan mengganti dengan nilai yang berada di 95% pada data. Pada proyek kali ini, client meminta kami untuk melakukan kategorisasi untuk negara yang berkembang dan kurang mampu. Sehingga melakukan handle lebih lanjut untuk data dengan nilai melebihi upper bound tidak diperlukan

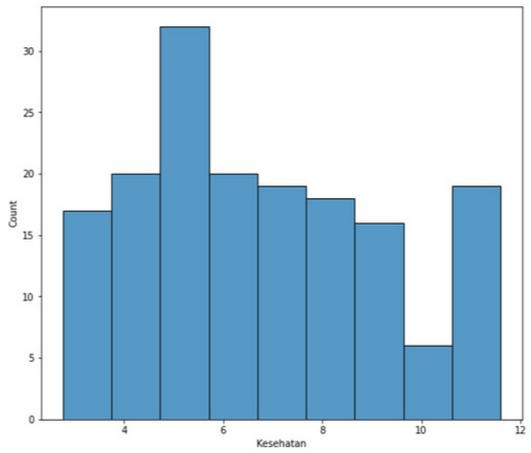
Exploratory data analysis

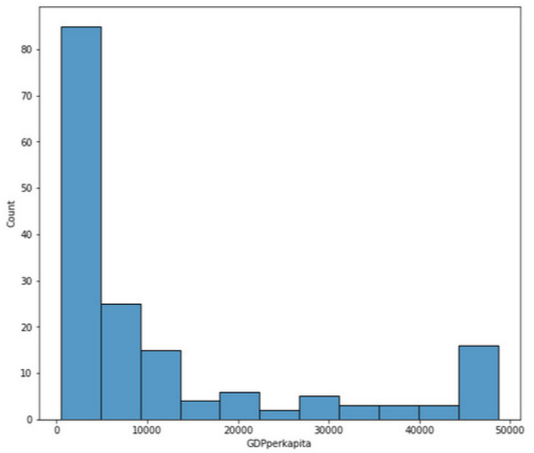
Univariate Analysis

Disini kami melakukan analisis univariate, yaitu density plot dan histogram. Density plot berguna untuk melihat distribusi data secara keseluruhan. Berdasarkan hasil analisis baik variabel 'Kesehatan' maupun 'GDPperkapita' memiliki distribusi data yang positif. Artinya bahwa nilai rata-rata dari variabel 'Kesehatan' dan 'GDPperkapita' lebih besar dari nilai modus dan median nya.

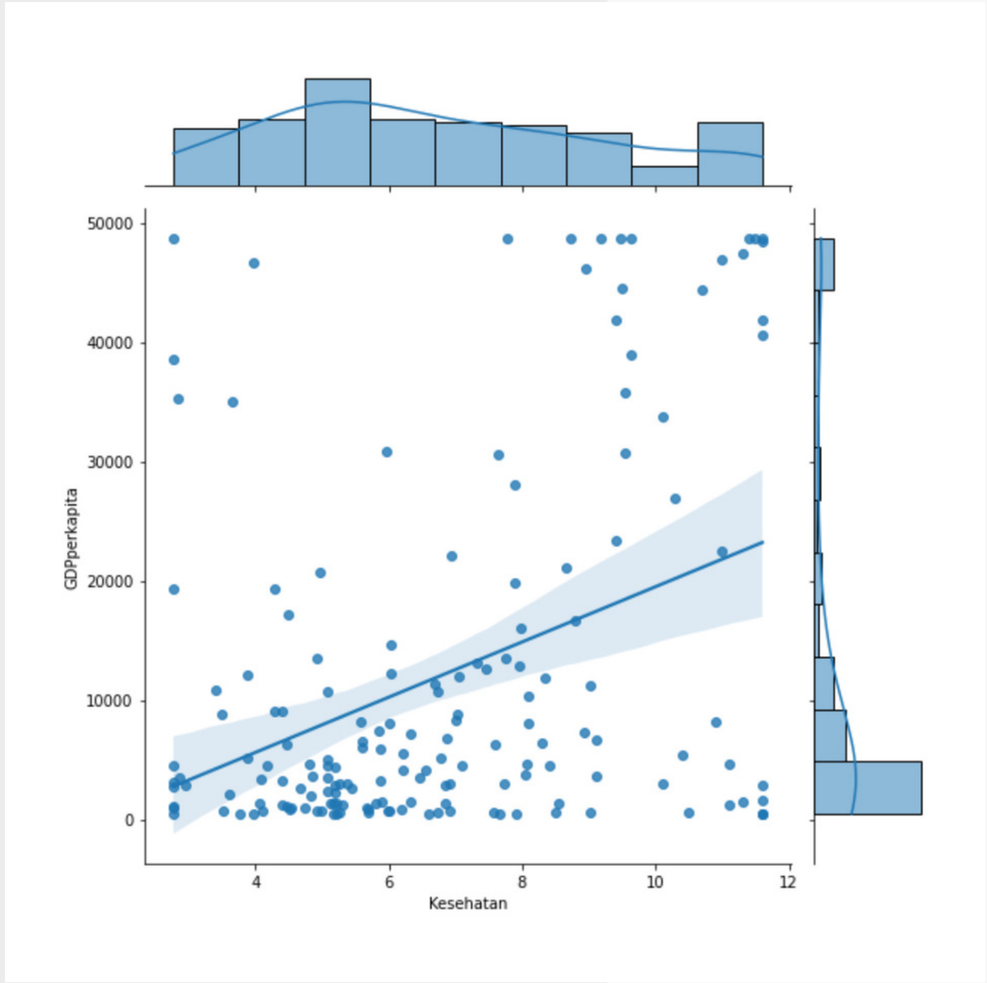


Exploratory data analysis





Selanjutnya kami melakukan analisis dengan menggunakan histogram. Dengan menggunakan histogram kita dapat melihat dengan lebih rinci distribusi data pada variabel 'Kesehatan' dan 'GDPperkapita'. Berdasarkan hasil analisis, dapat diketahui bahwa nilai modus untuk variabel 'Kesehatan' berkisar antara 4 - 6, sedangkan nilai modus untuk variabel 'GDPperkapita' yaitu berkisar antara 0 - 10.000.



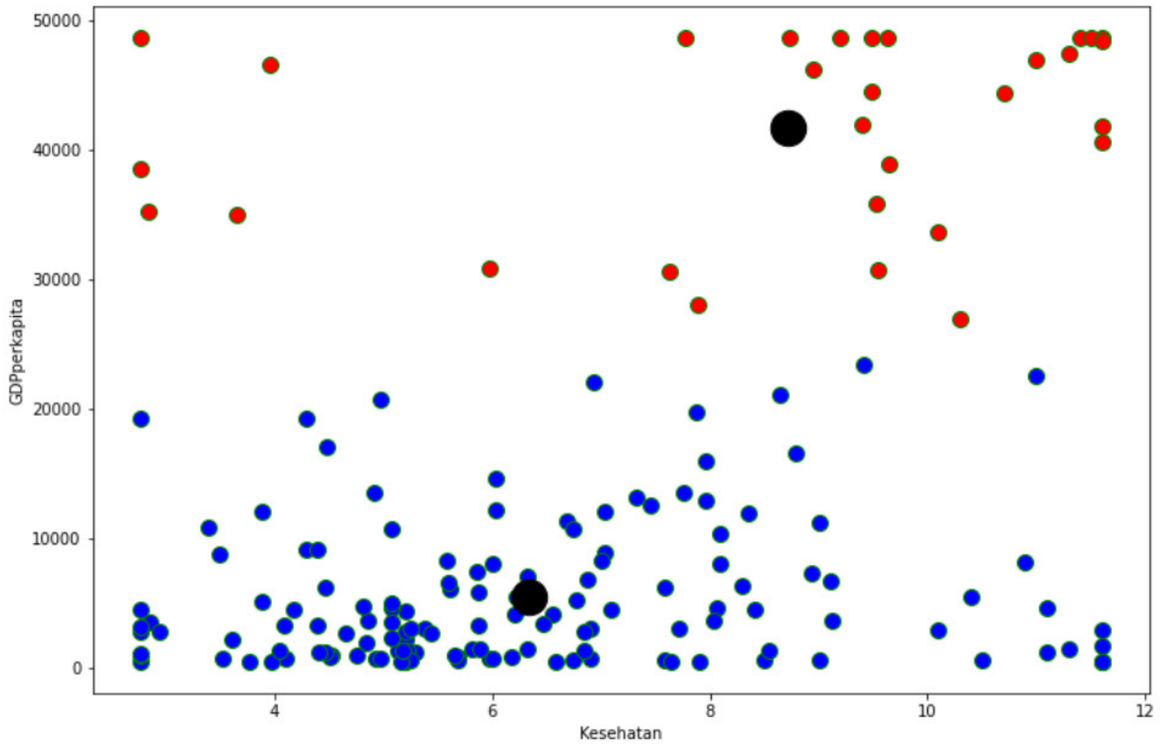
Bivariate Analysis

Kami menggunakan analisis regresi untuk bivariate analysis. Analisis regresi menjelaskan bagaimana suatu variabel mempengaruhi variabel lain. Berdasarkan hasil analisis regresi, variabel 'Kesehatan' memiliki pengaruh positif yang rendah bagi variabel 'GDPperkapita'. Hal tersebut ditunjukkan dari sebaran data yang mayoritas nya berada dibawah garis regresi.

Clustering

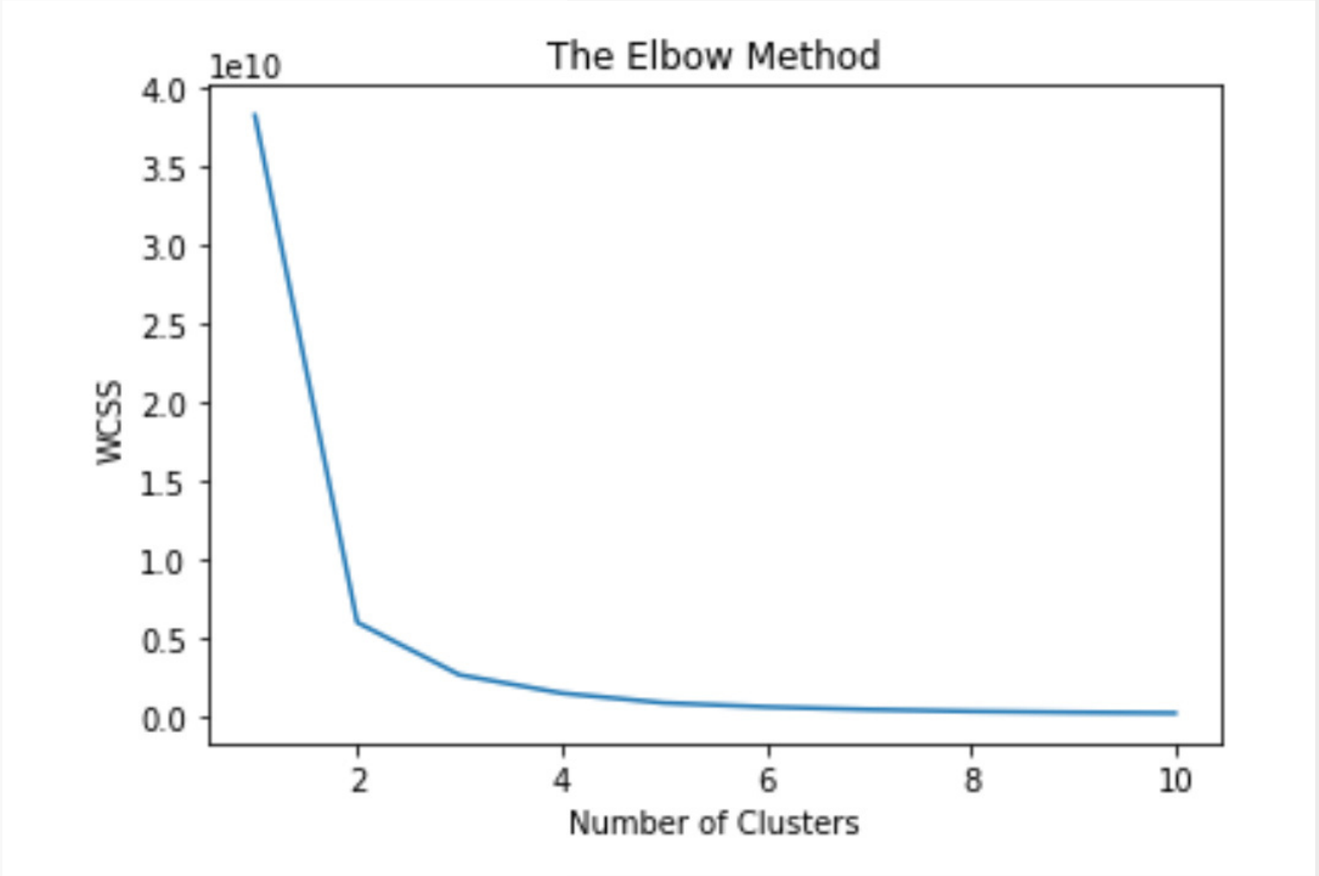
Clustering

Setelah dilakukan berbagai analisis, tahap terakhir dari proyek ini yaitu dilakukan clustering pada data. Disini kami mencoba membagi menjadi dua cluster. Dengan asumsi kedua cluster itu yaitu 'negara yang maju' dan ' negara yang berkembang' atau bisa dikatakan sebagai 'negara dengan GDPperkapita tinggi' dan 'negara dengan GDPperkapita rendah'



The Elbow Method

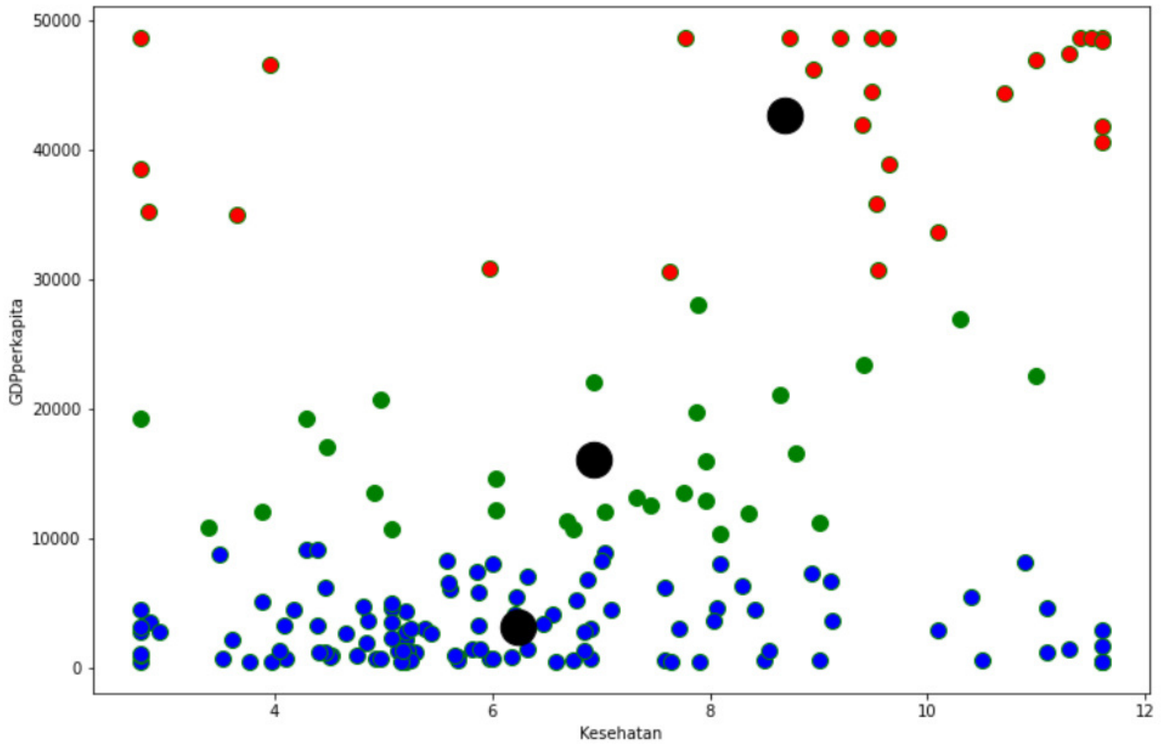
Selanjutnya kami melakukan elbow method untuk pengecekan apakah membagi menjadi dua cluster. Berdasarkan hasil analisis grafik nya mulai terlihat stabil di angka tiga. Maka, kami putuskan untuk melakukan clustering sekali lagi, untuk kemudian dilakukan silhouette score



Clustering 2

Setelah dilakukan clustering kedua, dan melakukan silhouette score kepada kedua cluster tadi, didapatkan bahwa nilai silhouette score untuk cluster 1 lebih besar dibandingkan dengan cluster 2. Maka, pembagian cluster terbaik adalah cluster 1.

```
Nilai silhouette score clustering 1 = 0.8030977186129541
Nilai silhouette score clustering 2 = 0.6992948786824138
```



Recommendation

Recommendation /04

	Negara	Kesehatan	GDPperkapita
93	Madagascar	3.77	459
31	Central African Republic	3.98	459
112	Niger	5.16	459
106	Mozambique	5.21	459
94	Malawi	6.59	459

Kami merekomendasikan dua negara teratas dengan tingkat 'Kesehatan' dan 'GDPperkapita' yang rendah, yaitu negara Madagarscar dan negara Central African Republic. Hal tersebut karena terbatasnya dana yang dimiliki oleh HELP International

Silahkan lihat proses analisis disini

Thank you!