**FINAL PROJECT REPORT**

SANBERCODE PYTHON-DATA SCIENCE



Dibuat Oleh:

Mohammad Dwiantara Mahardhika

# **Formulasi Masalah**

## **Permasalahan**

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 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.

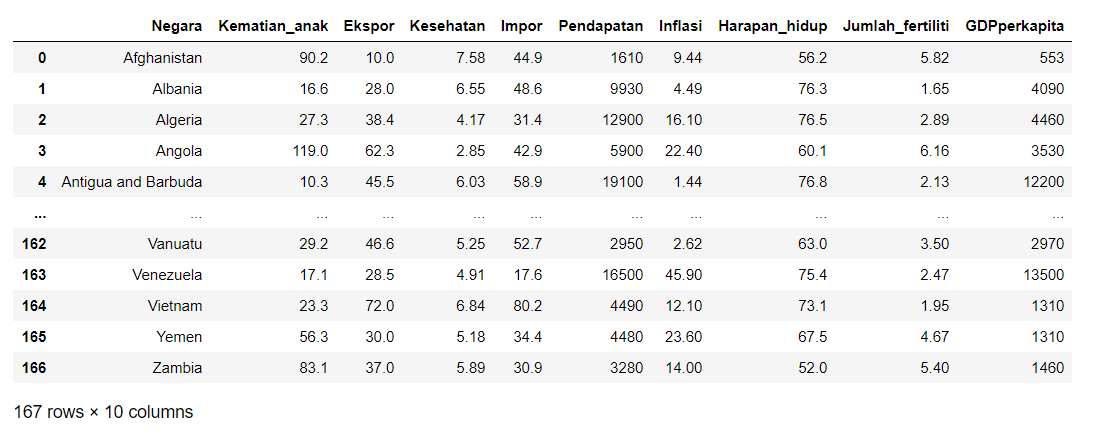
## **Tujuan**

Untuk mengkategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan.

# **Reading and Understanding Data**

* **Reading Data**

Data yang digunakan berasal dari file ‘Data\_Negara\_HELP.csv’ seperti berikut.



Dari pembacaan dataset diatas didapatkan bahwa data yang ada berjumlah 10 kolom dan 167 baris dengan keterangan kolom sebagai berikut.

|  |  |
| --- | --- |
| Kolom | Keterangan |
| 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. |

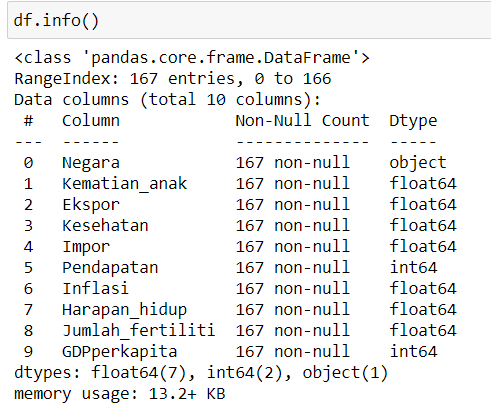
* **Describing Data**

Dilakukan analisa statistika deskriptif terhadap dataset yang digunakan dengan menggunakan library pandas. Didapatkan data-data statistik dari dataset seperti rata-rata nilai setiap kolom, nilai minimum dan maksimum dari setiap kolom dan lain sebagainya.



* **Data Info**

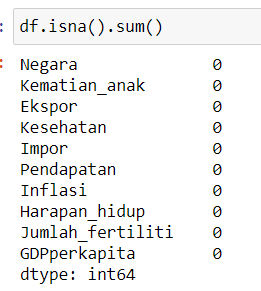
Dari analisa data info ini dapat diketahui tipe data setiap kolom dari dataset beserta jumlah data non-null pada setiaap kolom. Kolom negara bertipe objek tidap perlu diubah karena kolom negara hanya digunakan sebagai ‘index’. Untuk kolom yang lainnya merupakan kolom numerikal, semua kolom numerikal tersebut sudah bertipe float / int sehingga tidak perlu diubah.



# **Exploratory Data Analysis**

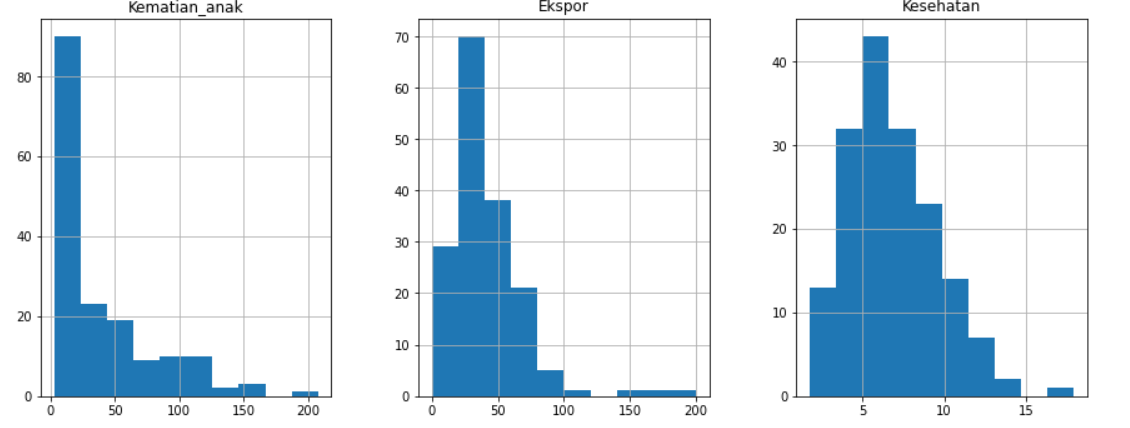
* **Cleaning Data**

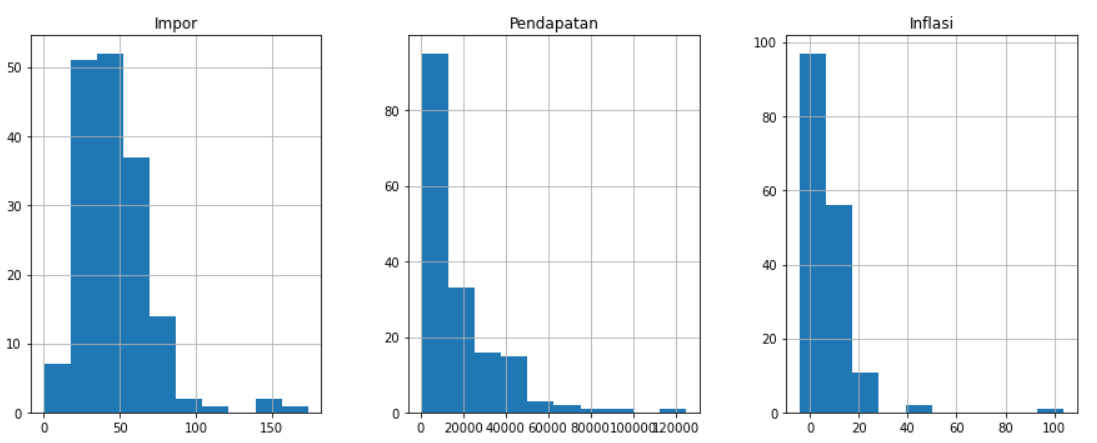
Setelah dilakukan analisa missing values pada dataset, didapatkan bahwa dataset yang digunakan tidak mempunyai missing values di semua kolom sehingga tidak dilakukan cleaning data.

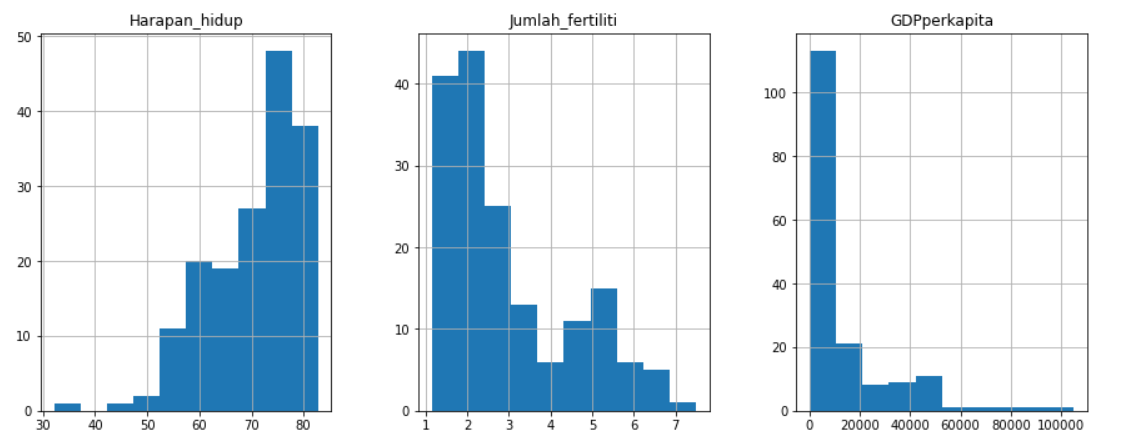


* **Univariate Analysis**

Dilakukan analisa univariate dengan menggunakan histogram untuk melihat persebaran data pada setiap kolom numerik di dataset.



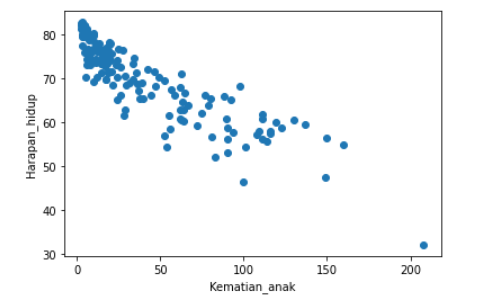




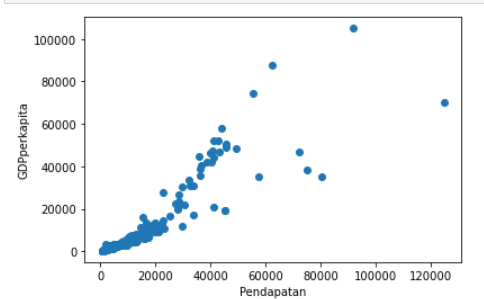
Dengan analisa univariate menggunakan histogram diatas kita dapat mengetahui persebaran data pada masing-masing kolom. Sebagai contoh pada kolom GDPperkapita dapat terlihat ternyata masih banyak negara yang mempunyai GDPperkapita rendah pada range 0-20000 yang lebih dari 100 negara begitu pula pada kolom Pendapatan, masih banyak negara yang mempunyai pendapatan rendah. Selain itu juga dari histogram diatas kita dapat mengetahui bahwa terdapat beberapa negara yang mempunyai kematian\_anak yang sangat tinggi dan juga terdapat negara yang mempunyai Harapan\_hidup yang sangat rendah.

* **Bivariate Analysis**

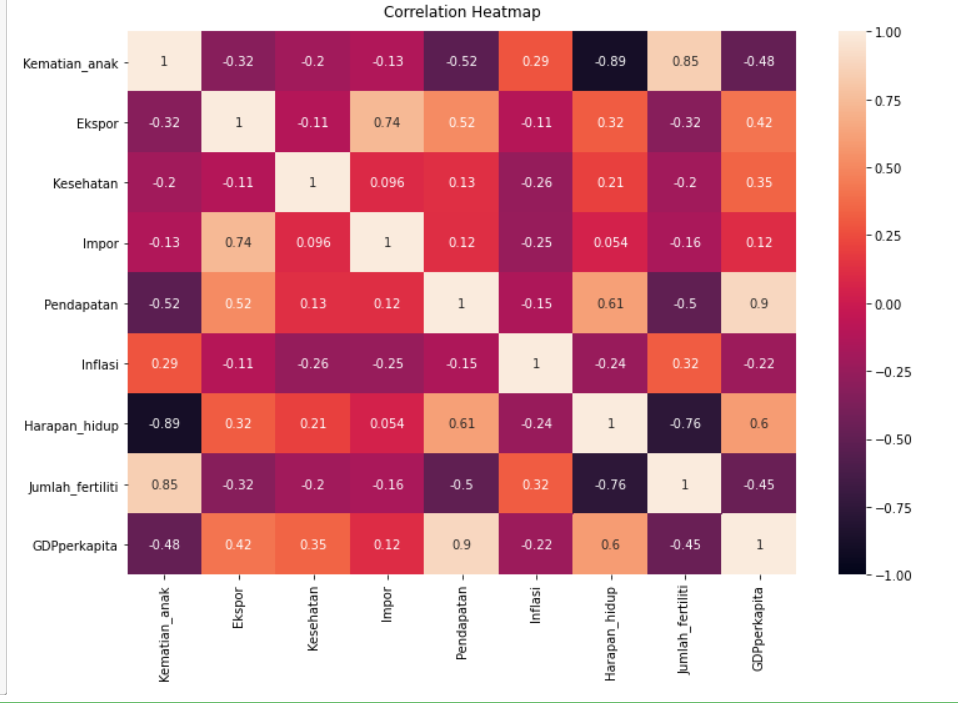
Dilakukan analisa bivariate terhadap variabel kematian\_anak dengan harapan\_hidup menggunakan scatter plot. Hasil visualisasi scatter menampilkan bahwa kedua variabel tersebut memiliki korelasi negatif karena semakin tinggi tingkat kematian\_anak di suatu negara maka harapan\_hidupnya semakin rendah



Dilakukan juga analisa terhadap hubungan pendapatan dengan GDPperkapita menggunakan scatter plot. Dari hasil yang didapat dapat disimpulkan bahwa kedua variabel tersebut memiliki korelasi positif karena semakin tinggi pendapatan perorang di suatu negara maka GDPperkapita di negara tersebut juga meningkat.

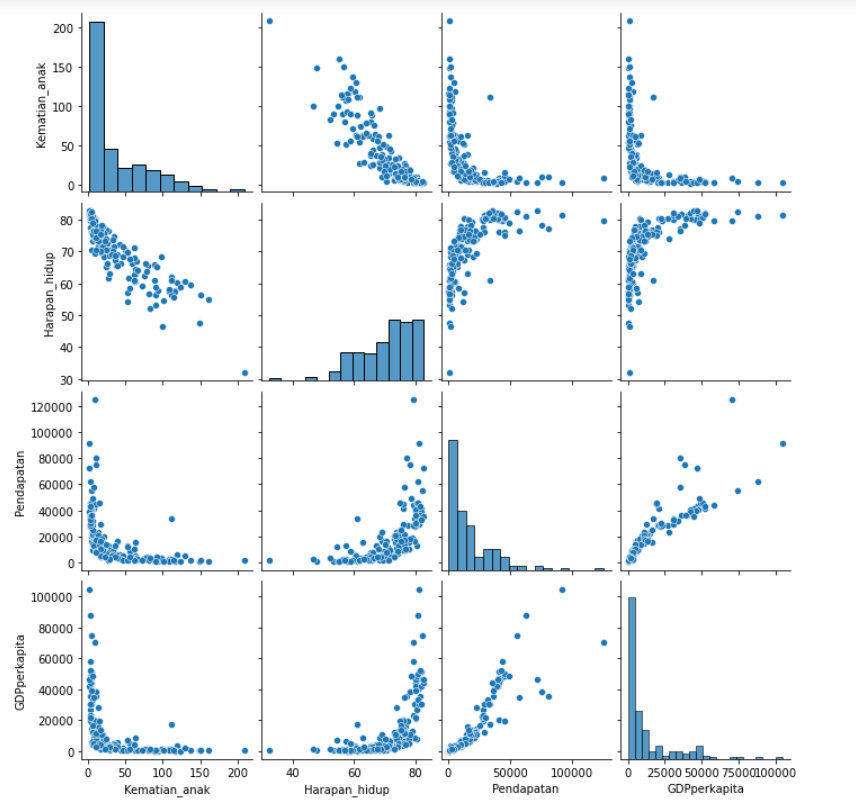


Dilakukan analisa korelasi keseluruhan variabel data dengan menggunakan heatmap. Dari hasil heatmap ini didapatkan bahwa korelasi antar Kematian\_anak dengan Harapan\_hidup (-0.89) dan pendapatan dengan GDPperkapita (0.9) mempunyai korelasi paling kuat.



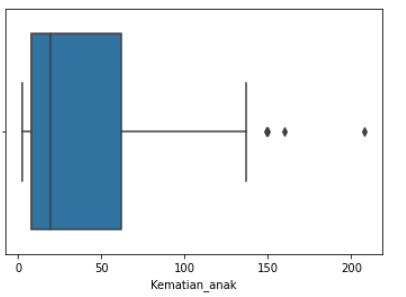
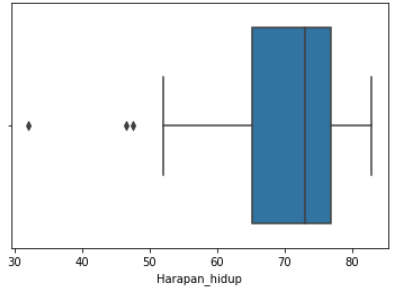
* **Multivariate Analysis**

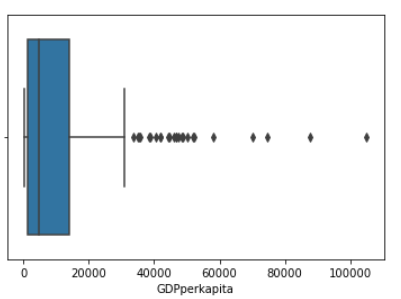
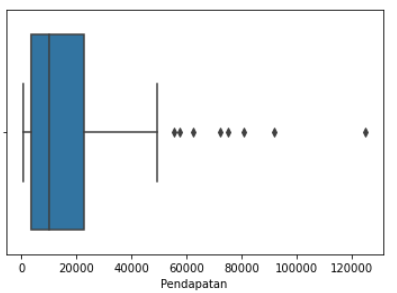
Dari hasil analisa univariate dan bivariate yang saya lakukan, saya memutuskan untuk menggunakan Kematian\_anak, Harapan\_hidup, Pendapatan, dan GDPperkapita untuk dianalisa lebih lanjut. Dari hasil pairplot untuk keempat variabel tersebut didapatkan hubungan dan persebaran data keempat data tersebut.



# **Outlier Treatment**

Setelah memilih variabel apa saja yang akan digunakan, kemudian dilakukan pengecekan terdapat data outlier pada kolom-kolom yang akan digunakan tersebut menggunakan boxplot.

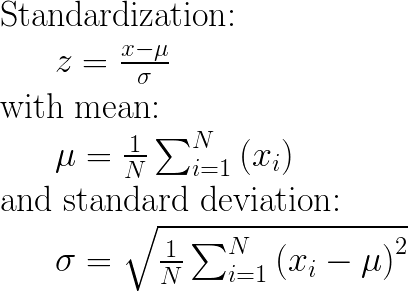
 

Dari hasil boxplot diatas dapat diketahui setiap kolom masih mempunyai outlier. Oleh karena itu dilakukan outlier handling menggunakan **Interquartile Range**masing-masing variabel.

# **Scaling Data**

Untuk mempermudah tahapa clustering dilakukan scaling data menggunakan StandardScaler.



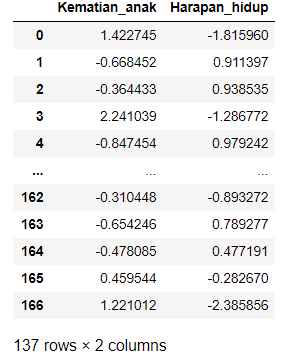


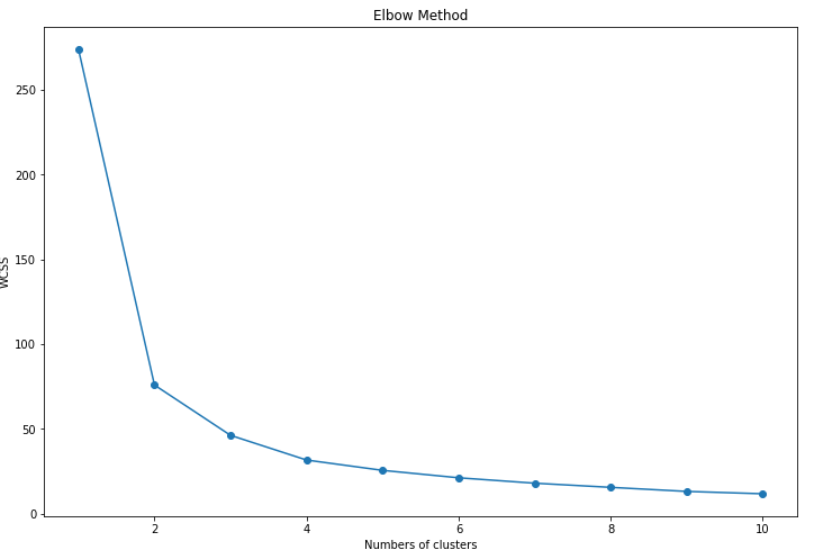
# **Clustering**

Pada analisa kali ini saya melakukan clustering untuk variabel Kematian\_anak dengan Harapan\_hidup dan Pendapatan dengan GDPperkapita.

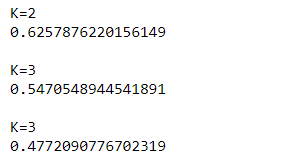
* **Kematian\_anak dengan Harapan\_hidup (Data Model 1)**

Data kematian\_anak dengan Harapan\_hidup.

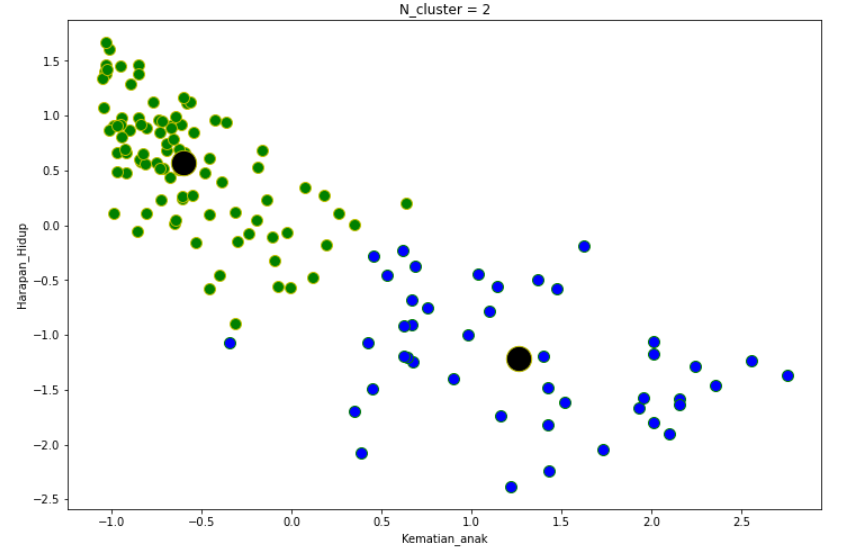




Dari hasil elbow method diatas dilakukan k yg optimal berada antara 2-4. Kemudian dilakukan validasi menggunakan Silhouette Score untuk nilai K = 2 sampai K = 4.



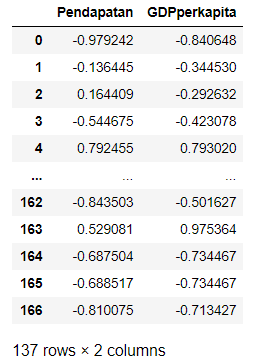
Dari hasil Silhouette Score didapatkan bahwa K=2 adalah mempunyai nilai terbaik. Oleh karena itu, clustering dilakukan menggunakan K=2 dengan hasil sebagai berikut.

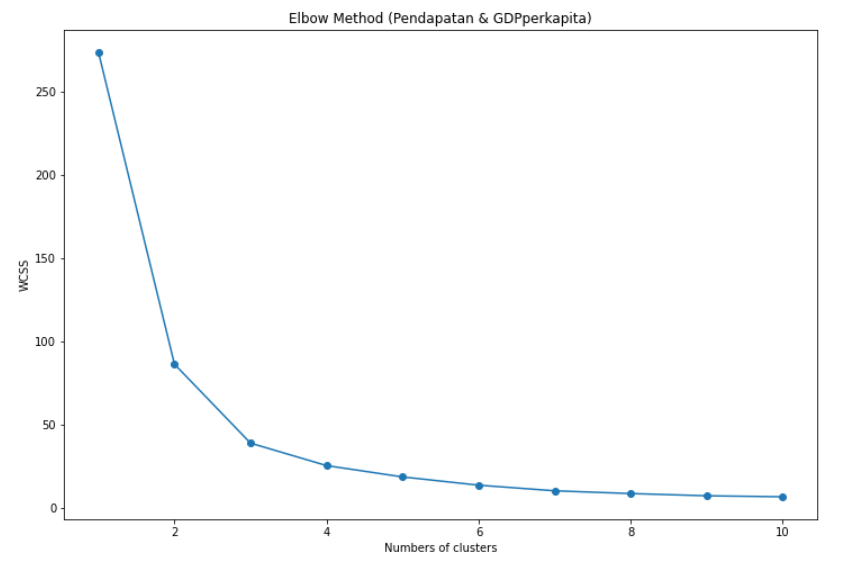


Dari hasil clustering diatas cluster 1 (Hijau) merupakan negara-negara yang mempunyai harapan hidup tinggi dan tingkat kematian\_anak rendah sedangkan cluster 2 (Biru) merupakan negara-negara yang mempunyai tingkat harapan\_hidup rendah dan tingkat kematian\_anak tinggi. Sehingga dapat diambil kesimpulan **Cluster 2 (Biru) adalah negara-negara kandidat fokus bantuan.**

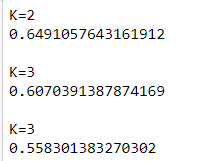
* **Pendapatan dengan GDPperkapita (Data Model 2)**

Data Pendapatan dengan GDPperkapita

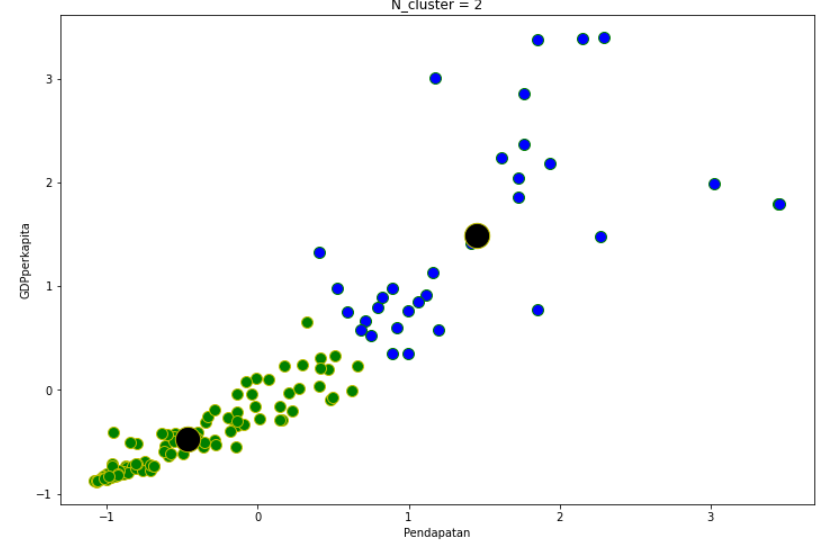




Dari hasil elbow method diatas dilakukan k yg optimal berada antara 2-4. Kemudian dilakukan validasi menggunakan Silhouette Score untuk nilai K = 2 sampai K = 4.



Dari hasil Silhouette Score didapatkan bahwa K=2 adalah mempunyai nilai terbaik. Oleh karena itu, clustering dilakukan menggunakan K=2 dengan hasil sebagai berikut.



Dari hasil clustering diatas cluster 1 (Hijau) merupakan negara-negara yang mempunyai GDPperkapita dan Pendapatan rendah sedangkan cluster 2 (Biru) merupakan negara-negara yang mempunyai GDPperkapitan dan Pendapatan tinggi. Sehingga dapat diambil kesimpulan **Cluster 1 (Hijau) adalah negara-negara kandidat fokus bantuan.**

# **Summary**

Terakhir dilakukan merge data hasil clustering data model 1 dengan clustering data model 2. Negara yang mempunyai Harapan\_hidup rendah dan kematian\_anak **(Cluster 2 Data Model 1)** tinggi serta mempunyai Pendapatan dan GDPperkapita rendah **(Cluster 1 Data Model 2)** adalah negara-negara yang perlu menjadi fokus bantuan. Berikut hasil merge kedua hasil clustering.