

Tugas akhir Python - Data Science

Clustering the Countries by using K-Means for HELP International

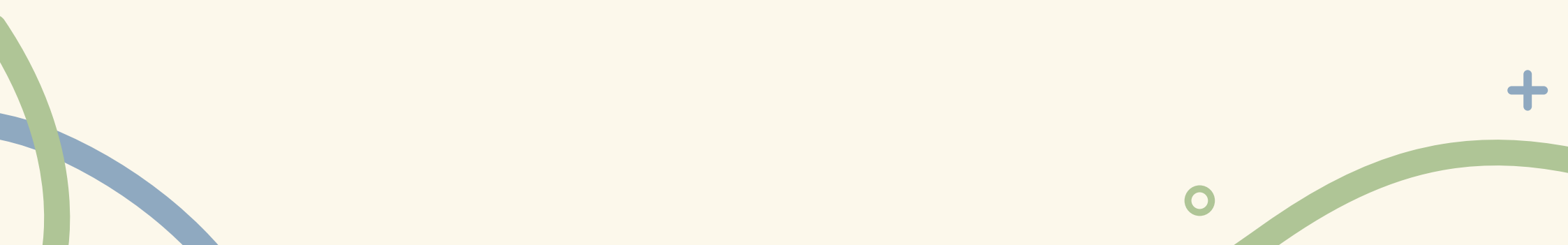


Oleh Aslam Fathin Rahmat



HELP International

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.





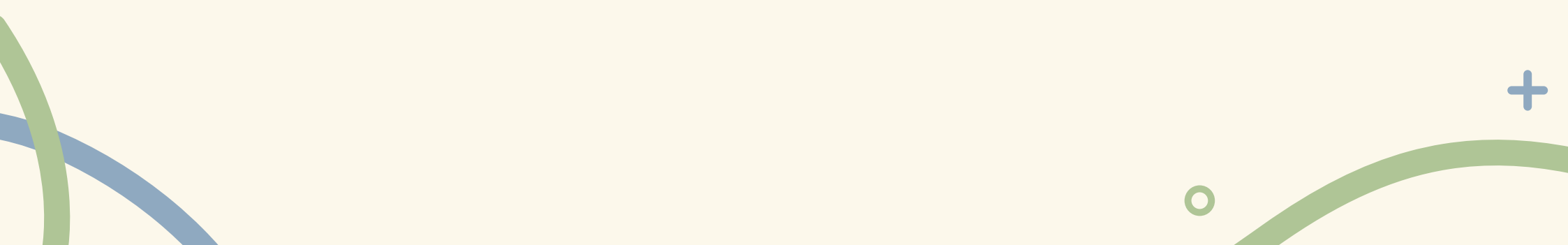
Permasalahan

HELP International memiliki dana sebesar \$10 juta untuk disumbangkan kepada negara yang membutuhkan. HELP Internasional memerlukan data negara yang paling membutuhkan bantuan.



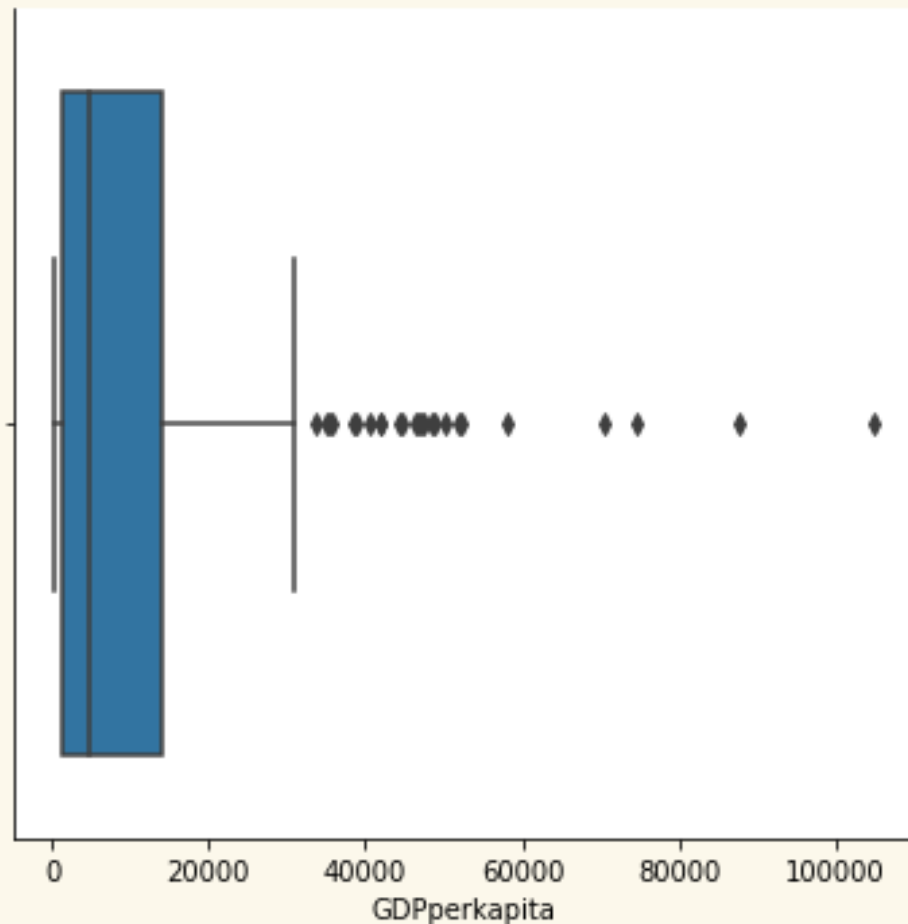
Data yang dimiliki

Untuk menentukan negara yang paling membutuhkan bantuan, digunakan data dari 167 negara yang terdiri dari :

- Nama negara
 - Jumlah kematian anak dibawah 5 tahun setiap 1000 kelahiran
 - Ekspor barang dan jasa per kapita
 - Total pengeluaran kesehatan perkapita
 - Impor barang dan jasa perkapita
 - Pendapatan bersih perorang
 - Nilai inflasi tahunan
 - Harapan hidup seorang anak yang baru lahir
 - Jumlah fertilitas anak yang akan lahir
 - GDP perkapita
- 

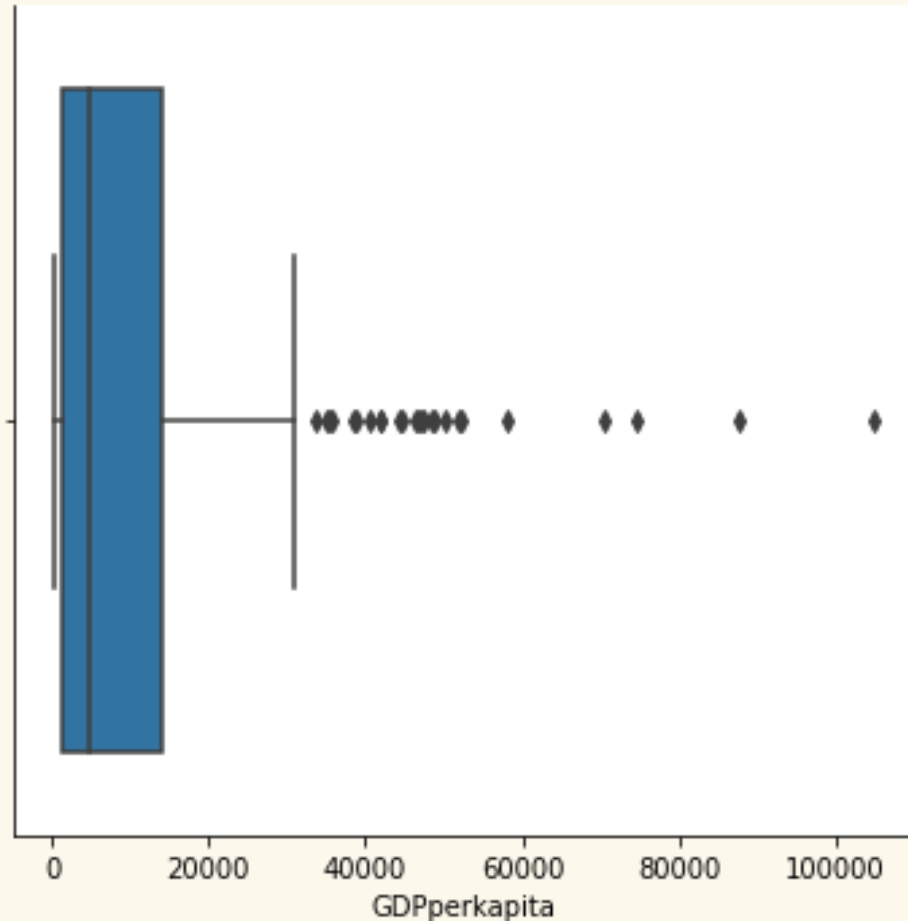
Permasalahan pada data

Pada data yang dimiliki terdapat *outliers* atau anomaly pada data. Salah satu contoh *outliers* terlihat pada *box plot* data GDP perkapita.



Box plot menggambarkan nilai maksimum, nilai minimum, median, quartil 1, quartil 3, dan nilai diluar batas maksimum atau minimum. Nilai diluar batas tersebut adalah *outlier* dilabelkan dengan titik dan dapat mempengaruhi analisa sehingga perlu dihilangkan

Permasalahan pada data

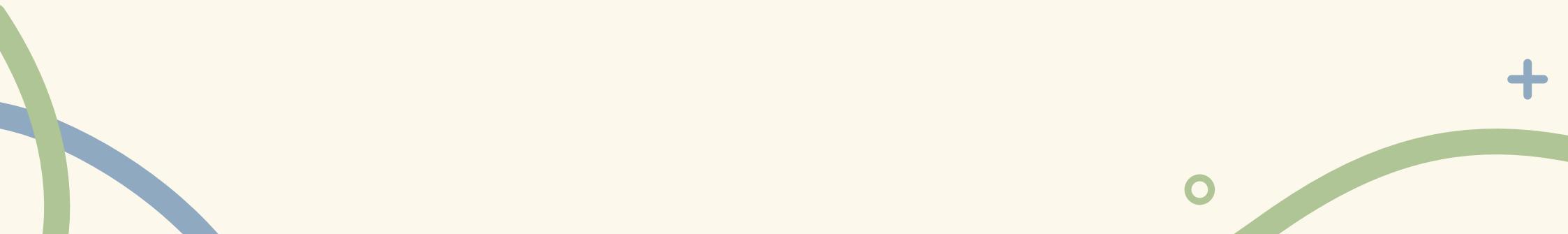


Outliers pada *box plot* GDP perkapita menandakan adanya negara dengan GDP perkapita yang sangat tinggi. Sudah pasti negara-negara tersebut tidak membutuhkan bantuan pendanaan. Oleh sebab itu, negara-negara tersebut harus dihilangkan dari data supaya tidak ikut teranalisa dan mempengaruhi hasil analisa.



Menghilangkan *outliers*

Untuk menghilangkan *outliers* dapat digunakan beberapa metode. Dalam analisa ini digunakan dua metode yaitu :

- Mengganti *outliers* dengan batas atas dan bawah
 - Menggunakan z-score
- 



Menghilangkan *outliers*

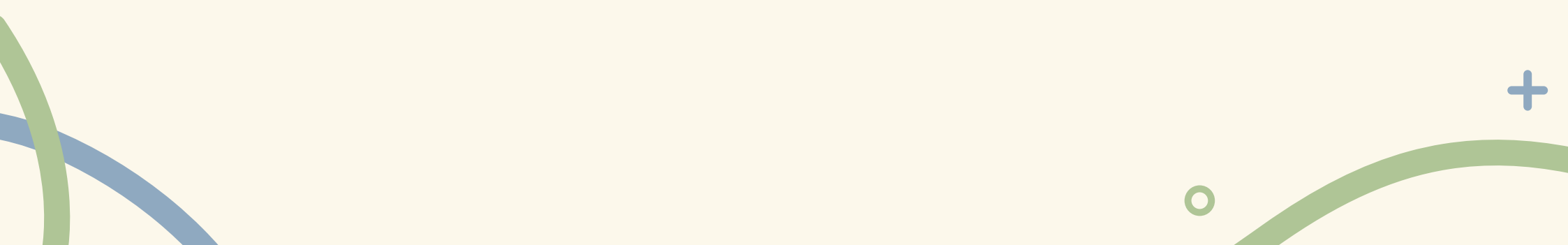
Tahap pertama dalam menghilangkan *outliers* dilakukan dengan mengganti *outliers* dengan batas atas dan batas bawah sesuai dengan persamaan berikut :

$$IQR = Q3 - Q1.$$

$$Lb = Q1 - (IQR \cdot 1.5)$$

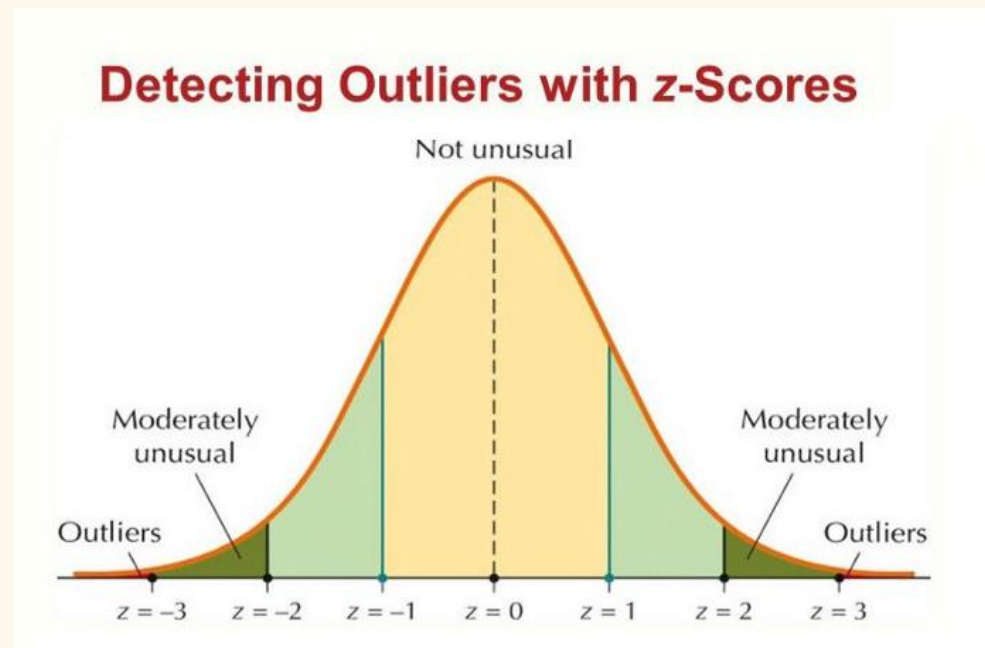
$$Ub = Q3 + (IQR \cdot 1.5)$$

Dengan metode ini *outliers* dapat dihilangkan, namun tidak seluruh *outliers* bisa dihilangkan karena nilai IQR yang terlalu besar mengakibatkan nilai batas atas atau bawah yang terlalu besar pula. Maka digunakan metode *z-score* untuk menghilangkan *outliers* hingga batas yang wajar.



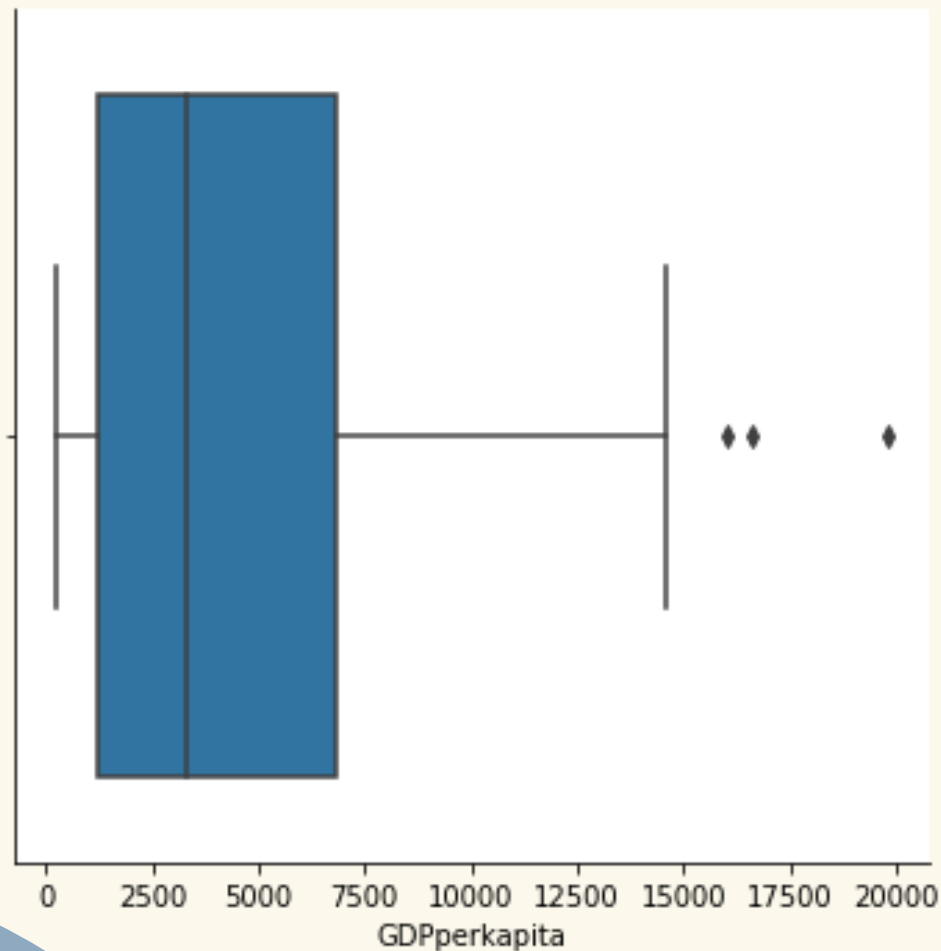
Menghilangkan *outliers*

Metode *z-score* menghilangkan *outliers* dengan melihat banyak standar deviasi sebuah data dari rata-rata. Misalkan sebuah data memiliki *z-score* = 2 berarti data tersebut berjarak 2 standar deviasi dari rata-rata. Pada umumnya *outliers* memiliki *z-score* lebih dari 3.



Menghilangkan *outliers*

Data yang telah dihilangkan *outliers* nya akan menyisakan data dengan persebaran yang lebih wajar seperti gambar di bawah.



Walaupun dalam *box plot* masih terdapat titik, namun data tersebut masih berada pada nilai yang wajar dan tidak terlalu jauh dari data yang lainnya. Setelah *outliers* dihilangkan terdapat 117 negara yang akan dianalisa.

Analisis data

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 117 entries, 0 to 166
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   Negara                117 non-null    object
 1   Kematian_anak          117 non-null    float64
 2   Ekspor                 117 non-null    float64
 3   Kesehatan              117 non-null    float64
 4   Impor                 117 non-null    float64
 5   Pendapatan            117 non-null    int64
 6   Inflasi               117 non-null    float64
 7   Harapan_hidup         117 non-null    float64
 8   Jumlah_fertiliti      117 non-null    float64
 9   GDPperkapita          117 non-null    int64
dtypes: float64(7), int64(2), object(1)
memory usage: 10.1+ KB
```

Dengan menggunakan fungsi `df.info()` didapatkan jumlah dan tipe data yang akan dianalisa. Terlihat seluruh fitur memiliki 117 non-null yang menandakan tidak ada nilai yang tidak valid atau NaN. Fitur yang akan dianalisis adalah :

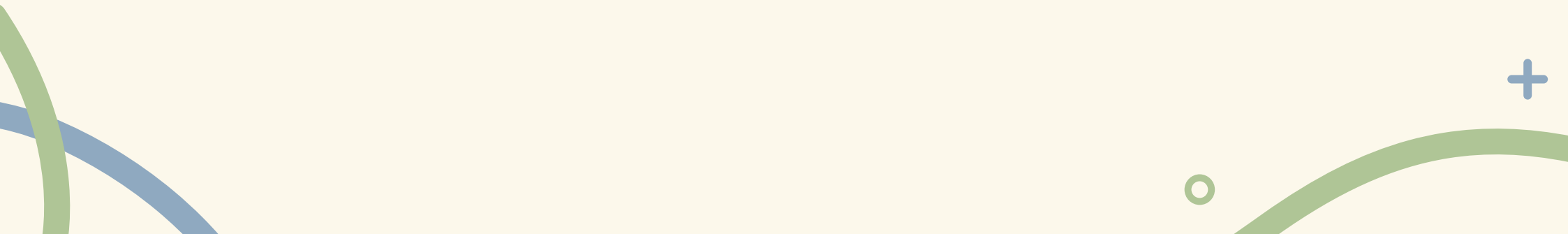
- Kematian anak
- Kesehatan
- Pendapatan.

Data akan divisualisasikan menggunakan *scatter plot* untuk melihat hubungan setiap data.

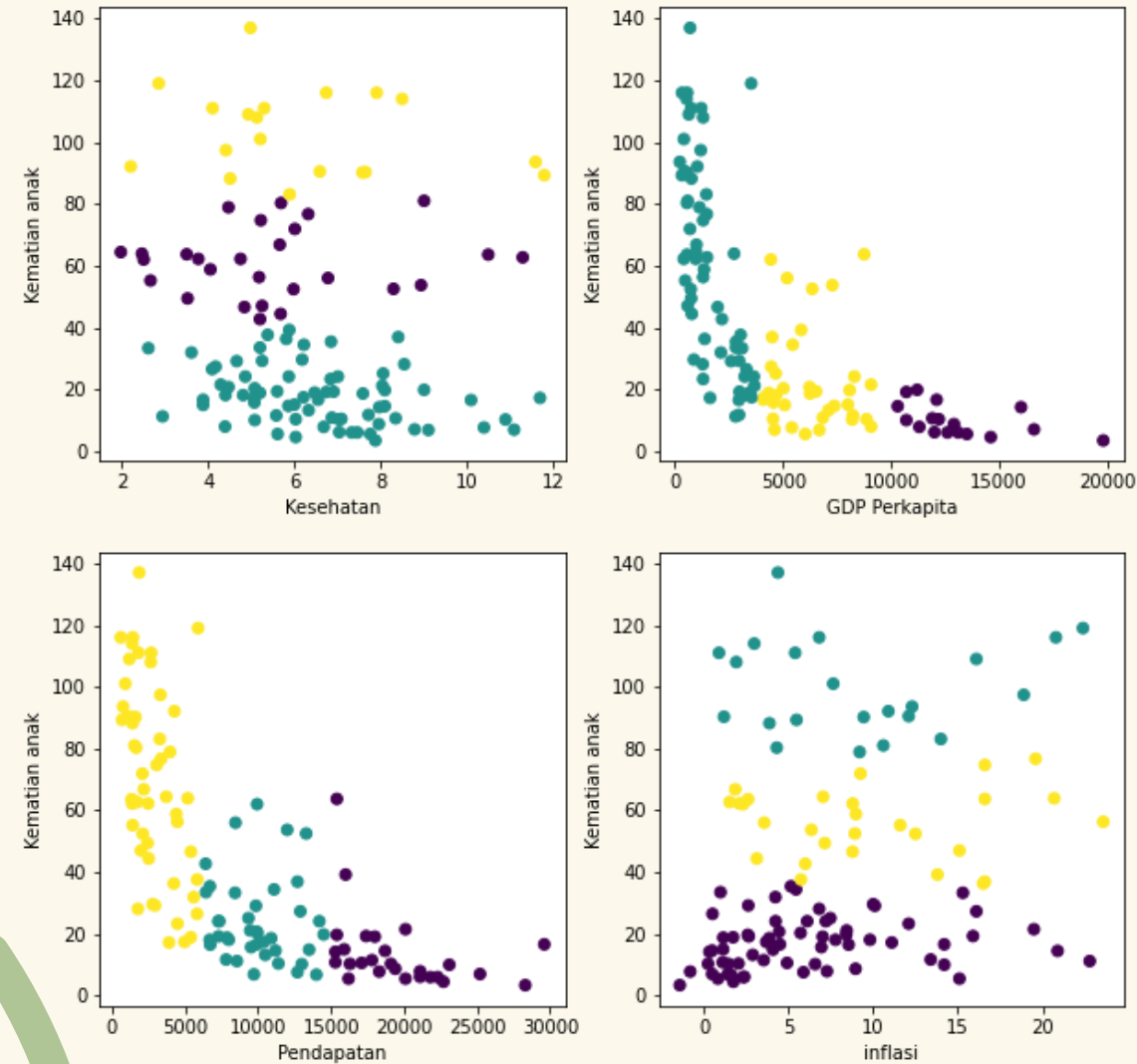


Analisis data kematian anak

Data kematian anak dikaitkan dengan 4 data lainnya yaitu kesehatan, GDP perkapita, pendapatan, dan inflasi. Dalam analisis digunakan K-Means *clustering* untuk membagi data menjadi 3 kelompok.



Analisis data kematian anak



Scatter plot disamping memperlihatkan bahwa :

1. Semakin tinggi GDP perkapita semakin rendah tingkat kematian anak
2. Semakin tinggi pendapatan bersih setiap orang semakin rendah tingkat kematian anak
3. Inflasi dan kesehatan tidak memberikan pengaruh pada tingkat kematian anak.

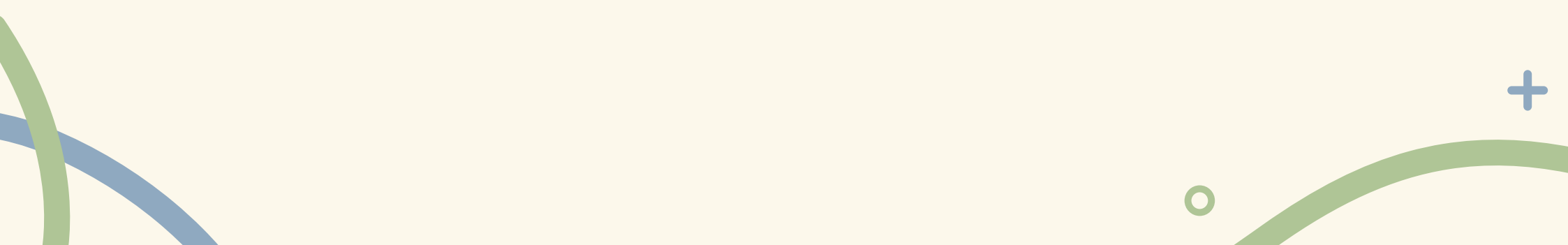


Analisis data kematian anak

Perbandingan kematian anak dengan kesehatan, GPD perkapita, pendapatan, dan inflasi menunjukkan bahwa kematian anak dipengaruhi oleh GPD perkapita negara tersebut dan pendapatan bersih perorang sedangkan inflasi tidak memberikan pengaruh apapun.

Uniknya, total pengeluaran kesehatan perkapita tidak mempengaruhi kematian anak. Hal ini menandakan bahwa besar-kecilnya pengeluaran perkapita untuk kesehatan tidak menjamin banyak-sedikitnya kematian seorang anak.

Oleh sebab itu, semakin Makmur sebuah negara semakin rendah kematian anak di negara tersebut. Hal ini dapat terjadi karena adanya faktor lain di luar data yang dimiliki seperti program kesehatan pemerintah.



Analisis data kematian anak

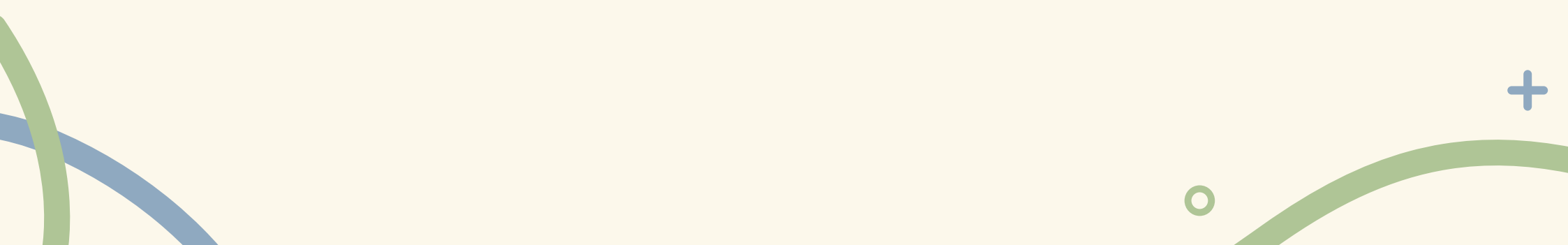
Negara	Mali	Negara	45.0
Kematian_anak	137.0	Kematian_anak	1.0
Ekspor	22.8	Ekspor	87.5
Kesehatan	4.98	Kesehatan	86.0
Impor	35.1	Impor	78.0
Pendapatan	1870	Pendapatan	98.0
Inflasi	4.37	Inflasi	72.0
Harapan_hidup	59.5	Harapan_hidup	100.0
Jumlah_fertiliti	6.55	Jumlah_fertiliti	1.0
GDPperkapita	708	GDPperkapita	100.0
Name: 97, dtype: object		Name: 97, dtype: float64	

Data di atas merupakan data negara dengan tingkat kematian anak tertinggi. Data di kiri merupakan nilai dari setiap fitur sedangkan data di kanan merupakan ranking setiap fitur. Terlihat Negara Mali memiliki tingkat kematian tertinggi dan juga jumlah fertility tertinggi. Selain itu, Mali memiliki harapan hidup yang rendah pada peringkat ke 100 dan GDP perkapita yang rendah pada peringkat ke 100.

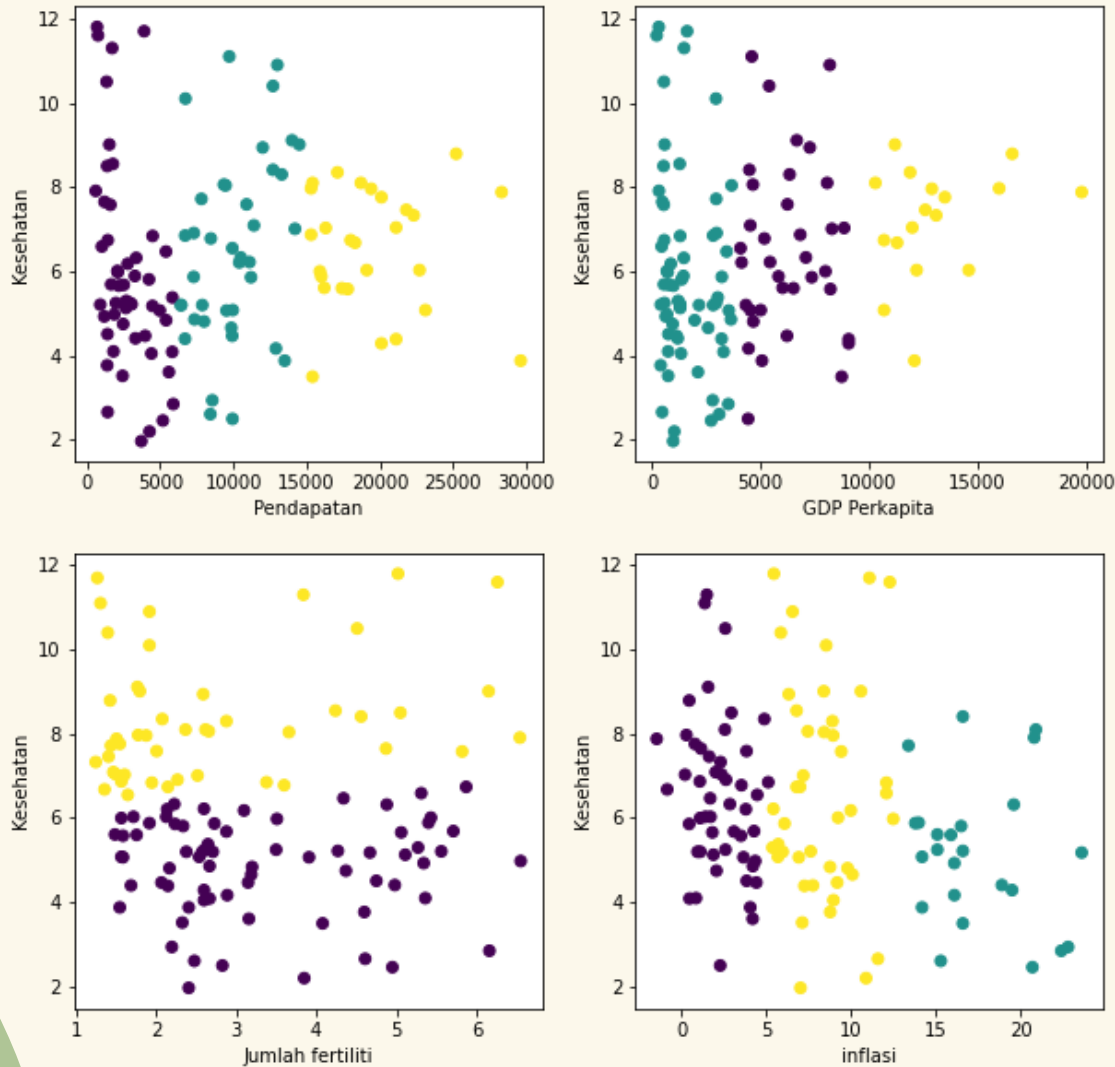


Analisis data kesehatan

Data kematian anak dikaitkan dengan 4 data lainnya yaitu GDP perkapita, pendapatan, jumlah fertiliti, dan inflasi. Dalam analisis digunakan K-Means *clustering* untuk membagi data menjadi 2 hingga 3 kelompok. Perlu diingat fitur kesehatan adalah pengeluaran untuk kesehatan perkapita bukan tingkat kesehatan di negara tersebut.



Analisis data kesehatan



Scatter plot disamping memperlihatkan bahwa tidak ada hubungan antara pengeluaran untuk kesehatan perkapita dengan pendapatan, GDP perkapita, jumlah fertility, dan inflasi.



Analisis data kesehatan

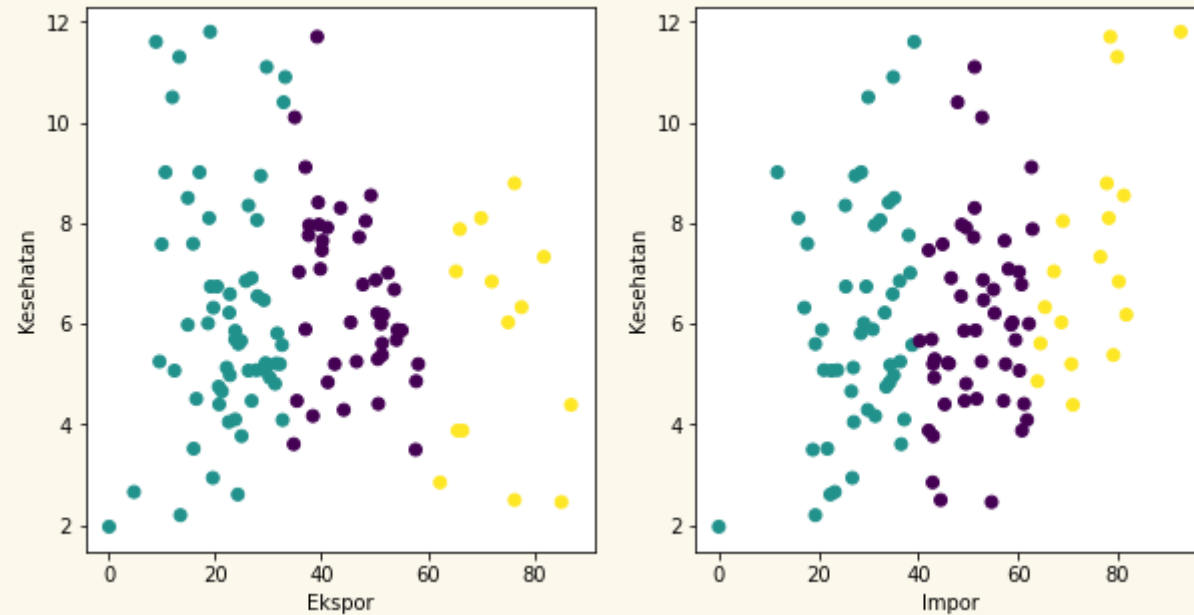
Dari *scatter plot* terlihat bahwa pengeluaran untuk kesehatan perkapita tidak dipengaruhi oleh pendapatan, GDP perkapita, jumlah fertility, dan inflasi. Data ekonomi yang disangka akan mempengaruhi pengeluaran untuk kesehatan justru tidak memiliki pengaruh sama sekali. Hal ini menandakan pengeluaran untuk kesehatan dilakukan sesuai kebutuhan perorang.

Analisis data kesehatan

Negara	Myanmar	Negara	38.0
Kematian_anak	64.4	Kematian_anak	27.0
Ekspor	0.109	Ekspor	117.0
Kesehatan	1.97	Kesehatan	117.0
Impor	0.0659	Impor	117.0
Pendapatan	3720	Pendapatan	82.0
Inflasi	7.04	Inflasi	53.0
Harapan_hidup	66.8	Harapan_hidup	76.0
Jumlah_fertiliti	2.41	Jumlah_fertiliti	68.5
GDPperkapita	988	GDPperkapita	92.0
Name: 107, dtype: object		Name: 107, dtype: float64	

Data di atas merupakan data negara dengan tingkat pengeluaran untuk kesehatan perkapita terendah. Data di kiri merupakan nilai dari setiap fitur sedangkan data di kanan merupakan ranking setiap fitur. Terlihat Negara Myanmar memiliki tingkat pengeluaran untuk kesehatan perkapita yang paling rendah pada peringkat ke 117. Terlihat pula Myanmar memiliki ekspor dan impor yang paling rendah.

Analisis data kesehatan



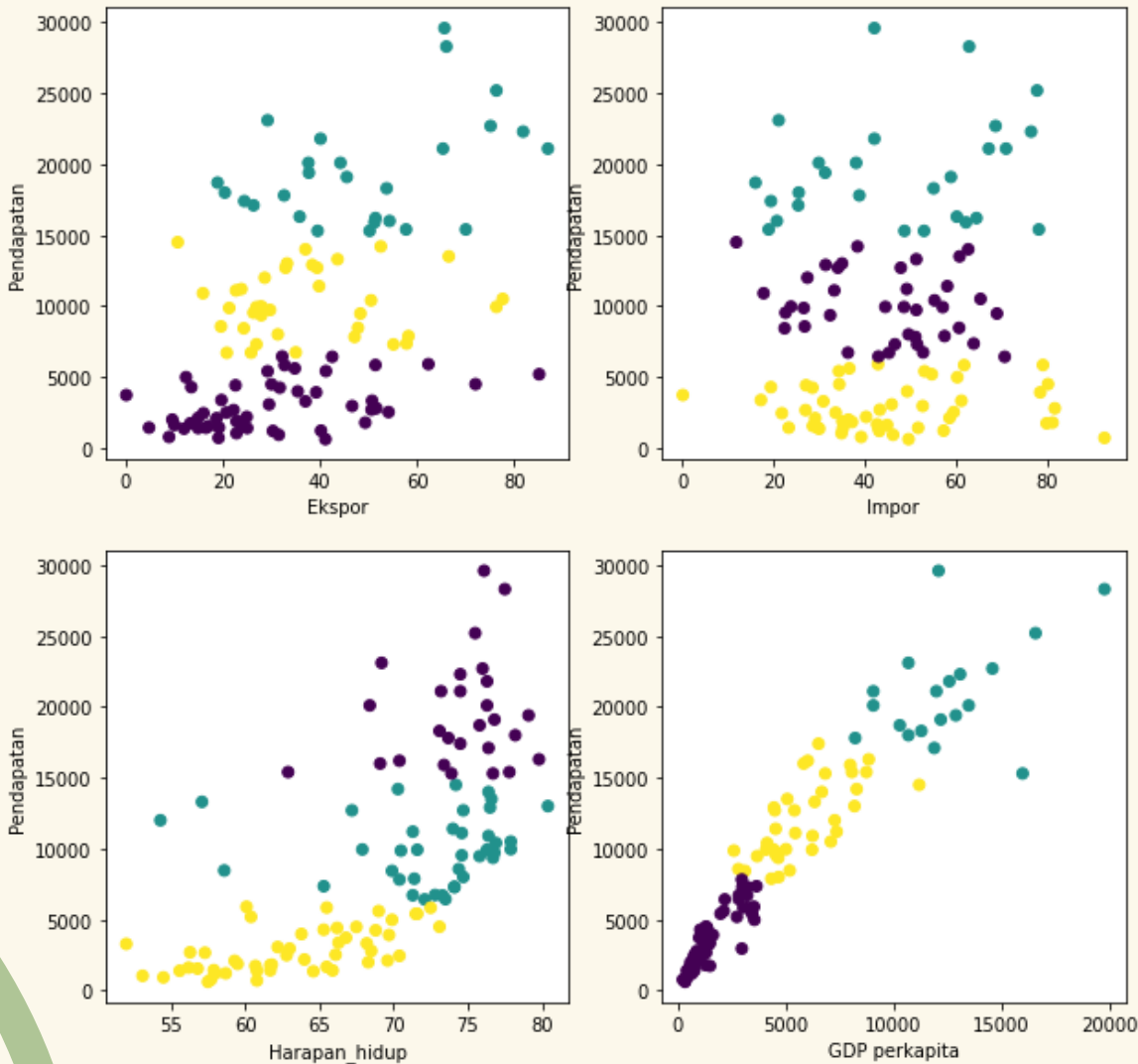
Namun, setelah data pengeluaran untuk kesehatan dibandingkan dengan impor dan ekspor dijabarkan pada *scatter plot*, terlihat tidak terdapat hubungannya. Hal Ini berarti bahwa Myanmar memiliki kebijakan ekonomi untuk tidak melakukan ekspor dan impor dalam jumlah besar.



Analisis data Pendapatan

Data kematian anak dikaitkan dengan 4 data lainnya yaitu GDP perkapita, ekspor, impor, dan harapan hidup. Dalam analisis digunakan K-Means *clustering* untuk membagi data menjadi 3 kelompok.

Analisis data pendapatan



Scatter plot disamping memperlihatkan bahwa :

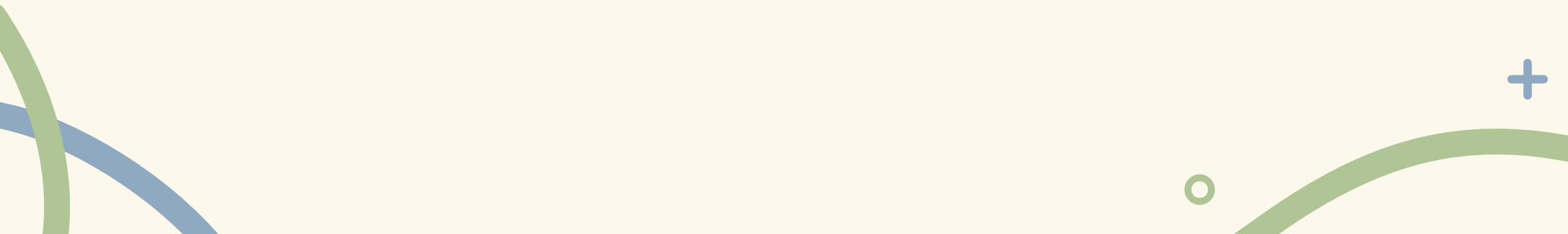
1. Semakin tinggi pendapatan semakin tinggi harapan hidup.
2. Semakin tinggi GDP perkapita semakin tinggi pendapatan.
3. Tidak ada hubungan pendapatan dengan ekspor dan impor.



Analisis data pendapatan

Pendapatan bersih perorang mempengaruhi harapan hidup. Semakin tinggi pendapatan bersih seseorang semakin tinggi harapan hidup anak diatas 5 tahun. Hal ini sesuai dengan analisis sebelumnya di mana semakin tinggi pendapatan semakin rendah tingkat kematian anak, begitu pula sebaliknya. Selain itu, pendapatan berbanding lurus dengan GDP perkapita bahkan hampir mendekati linear.

Terlihat pula ekspor dan impor tidak mempengaruhi pendapatan bersih perorang. Dapat diperkirakan bahwa pendapatan bersih perorang dipengaruhi oleh faktor lain seperti pekerjaan orang tersebut.



Analisis data kesehatan

```
Negara          Congo, Dem. Rep.  
Kematian_anak    116.0  
Ekspor           41.1  
Kesehatan        7.91  
Impor            49.6  
Pendapatan       609  
Inflasi          20.8  
Harapan_hidup    57.5  
Jumlah_fertiliti 6.54  
GDPperkapita     334  
Name: 37, dtype: object
```

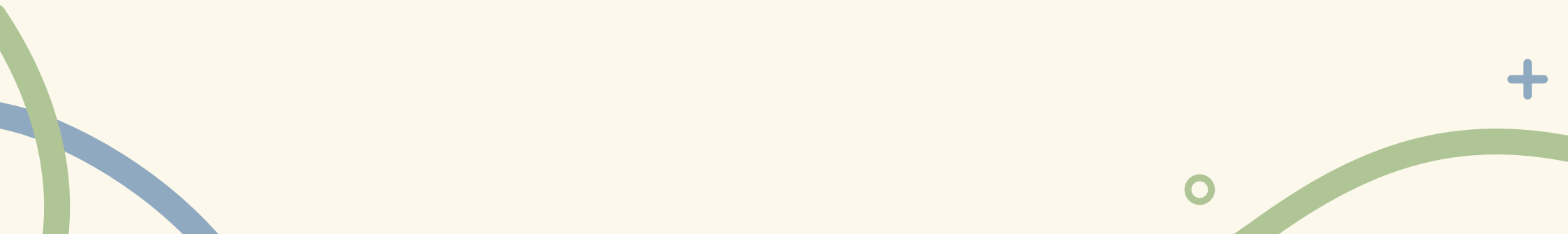
```
Negara          89.0  
Kematian_anak    3.5  
Ekspor           41.0  
Kesehatan        26.0  
Impor            47.5  
Pendapatan       117.0  
Inflasi          5.0  
Harapan_hidup    107.0  
Jumlah_fertiliti 2.0  
GDPperkapita     115.0  
Name: 37, dtype: float64
```

Data di atas merupakan data negara dengan tingkat pendapatan rendah. Data di kiri merupakan nilai dari setiap fitur sedangkan data di kanan merupakan ranking setiap fitur. Terlihat Negara Congo memiliki tingkat pendapatan yang paling rendah pada peringkat ke 117. Terlihat pula Congo memiliki nilai inflasi, kematian anak, dan jumlah fertilitas yang cukup tinggi.



Analisis lanjut

Setelah didapatkan 3 negara yang berpotensi mendapat bantuan dari 3 kategori, diperlukan analisis lebih lanjut untuk menentukan negara mana yang pantas mendapatkan bantuan dari HELP International.



Analisis lanjut

	Congo	Myanmar	Mali
Negara	89.0	38.0	45.0
Kematian_anak	3.5	27.0	1.0
Ekspor	41.0	117.0	87.5
Kesehatan	26.0	117.0	86.0
Impor	47.5	117.0	78.0
Pendapatan	117.0	82.0	98.0
Inflasi	5.0	53.0	72.0
Harapan_hidup	107.0	76.0	100.0
Jumlah_fertiliti	2.0	68.5	1.0
GDPperkapita	115.0	92.0	100.0
Keterangan	Pendapatan terendah	Kesehatan terendah	Kematian anak tertinggi

	Congo	Myanmar	Mali
Negara	Congo, Dem. Rep.	Myanmar	Mali
Kematian_anak	116.0	64.4	137.0
Ekspor	41.1	0.109	22.8
Kesehatan	7.91	1.97	4.98
Impor	49.6	0.0659	35.1
Pendapatan	609	3720	1870
Inflasi	20.8	7.04	4.37
Harapan_hidup	57.5	66.8	59.5
Jumlah_fertiliti	6.54	2.41	6.55
GDPperkapita	334	988	708
Keterangan	Pendapatan terendah	Kesehatan terendah	Kematian anak tertinggi

Tabel di atas merupakan data negara Congo, Myanmar, dan Mali. Tabel di kiri menunjukkan ranking setiap fitur sebuah negara sedangkan table di kanan menunjukkan nilai setiap fitur sebuah negara.



Analisis lanjut

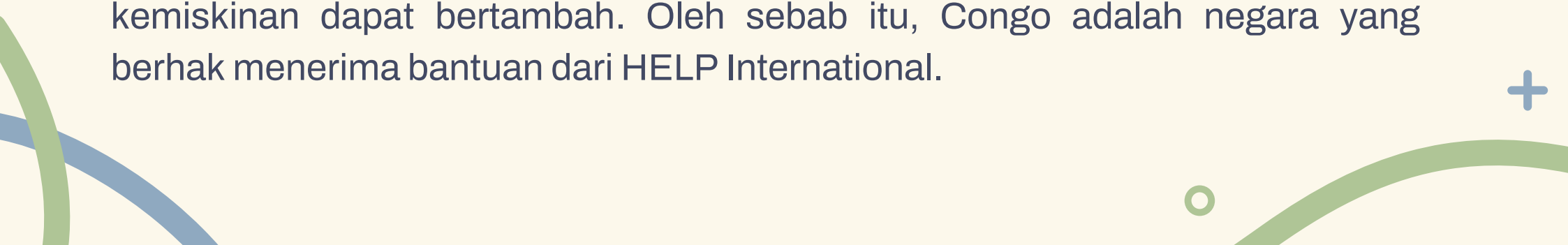
Dari kedua table sebelumnya terlihat bahwa :

1. Myanmar memiliki nilai keehatan, ekspor, dan impor terendah
2. Congo memiliki nilai pendapatan, GDP perkapita, dan harapan hidup terendah serta inflasi tertinggi.
3. Mali memiliki nilai kematian anak tertinggi.
4. Nilai kematian anak, harapan hidup, dan jumlah fertiliti Mali dan Congo hampir sama.



Kesimpulan

Dari analisis sebelumnya dapat disimpulkan bahwa negara yang pantas mendapatkan bantuan dari HELP International adalah Congo karena Congo memiliki GDP perkapita dan pendapatan bersih perorang yang paling rendah. GDP perkapita dan pendapatan bersih perorang sebuah negara yang rendah dapat terjadi akibat banyak hal. Dalam kasus ini kemungkinan besar rendahnya GDP perkapita dan pendapatan bersih perorang di Congo terjadi akibat inflasi yang tinggi hingga 20,4%. Tinggi nya inflasi berdampak pada banyak hal seperti harapan hidup yang rendah dan kematian anak yang tinggi. Selain itu, jumlah fertiliti di Congo termasuk tinggi dan hampir sama dengan Mali. Namun Congo memiliki masalah yang lebih besar diakibatkan oleh rendahnya pendapatan bersih perorang yang membuat pengeluaran untuk kesehatan menjadi rendah. Jika dibiarkan maka jumlah kematian anak dapat meningkat serta jumlah kemiskinan dapat bertambah. Oleh sebab itu, Congo adalah negara yang berhak menerima bantuan dari HELP International.





Sumber referensi :

1. <https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/>
2. <https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/>
3. Materi kelas Python – Data Science Sanbercode

Terima kasih!

CREDITS: This presentation template was created by **Slidesgo** and includes icons by **Flaticon**, infographics & images by **Freepik** and content by **Eliana Delacour**

