01

FINAL PROJECT

CLUSTERING THE COUNTRIES BY USING K-MEANS FOR HELP INTERNATIONAL

BY: PATRICK AMADEUS IRAWAN

02

CONTENTS OF THE REPORT

SECTION 1

UNDERSTANDING AND OBJECTIVES

- Tujuan Report
- Organisasi HELP International
- Tujuan Organisasi

SECTION 2

ANALYSIS

- Tools
- Overview Data
- Proses Analisis Data

SECTION 3

IINSIGHTS

- Hasil Analisis Data
- Kesimpulan dan Objectives

SECTION 1

UNDERSTANDING OBJECTIVES OF HELP INT'L

01

GOAL OF THE REPORT

MENGKATEGORIKAN NEGARA MENGGUNAKAN FAKTOR SOSIAL EKONOMI DAN KESEHATAN YANG MENENTUKAN PEMBANGUNAN NEGARA SECARA KESELURUHAN.

MENENTUKAN NEGARA-NEGARA YANG WAJIB MENJADI FOKUS CEO HELP INTERNATIONAL

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.

06

THE OBJECTIVES

01

OBJECTIVE

Memberantas Kemiskinan dan Kelaparan di negara-negara terbeakang

OBJECTIVE

02

Mengurangi tingkat kesenjangan ekonomi dan sosial dalam lapis masyarakat di berbagai negara

OBJECTIVE 03

Secara proaktif membantu negara-negara terbelakang memerangi defisiensi material akibat bencana alam dan perperangan

SECTION 2

DATA ANALYSIS

02

DATA ANALYSIS TOOLS

Analisis data : Pandas
Visualisasi : Matplotlib , Seaborn
Modelling : Sklearn

1

2

3

Data Overview and Analysis

Gathering , Outliers and Missing Checking , visualization

1

DATA OVERVIEW

DATA SHAPE

```
df.shape
✓ 0.8s
(167, 10)
```

DATA HEAD

```
df.head()
```

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertil
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.

COLUMN EXPLANATION

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.

DATA COLUMN INFO

```
<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)
```

DATA STATS OVERVIEW

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

DATA COMPLETENESS

```
Negara              False
Kematian_anak       False
Ekspor              False
Kesehatan           False
Impor               False
Pendapatan          False
Inflasi             False
Harapan_hidup       False
Jumlah_fertiliti    False
GDPperkapita        False
dtype: bool
```

*Dapat dipastikan no missing data dan preprocessing difokuskan ke arah outliers handling

15

DATA PRE PROCESSING

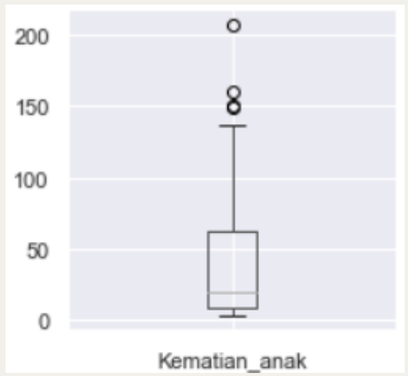
IQR FORMULA

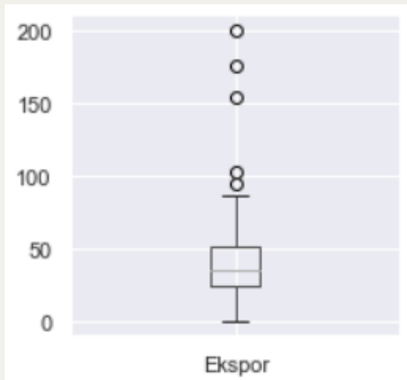
```
def batas(column,df):
    Q1,Q3 = df[column].quantile(0.25) , df[column].quantile(0.75)
    IQR = Q3-Q1
    return (Q1-1.5*IQR , Q3+1.5*IQR)
```

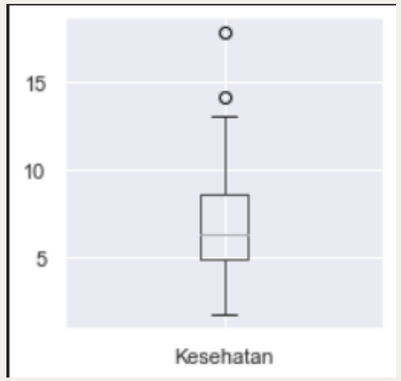
RANGE NORMAL UNTUK DATA DENGAN IQR FORMULA

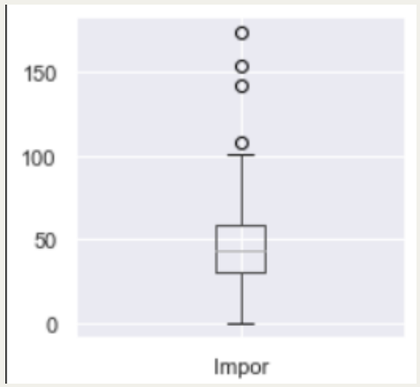
```
Kematian_anak (-72.525, 142.875)
Ekspor (-17.524999999999988, 92.67499999999998)
Kesehatan (-0.6000000000000023, 14.120000000000005)
Impor (-12.625000000000004, 101.575)
Pendapatan (-25812.5, 51967.5)
Inflasi (-11.6, 24.16)
Harapan_hidup (48.05, 94.05)
Jumlah_fertiliti (-1.3325, 7.0075)
GDPperkapita (-17750.0, 33130.0)
```

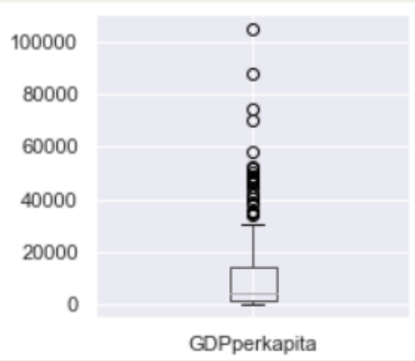
BOXPLOT VISUALIZATION PER FEATURES

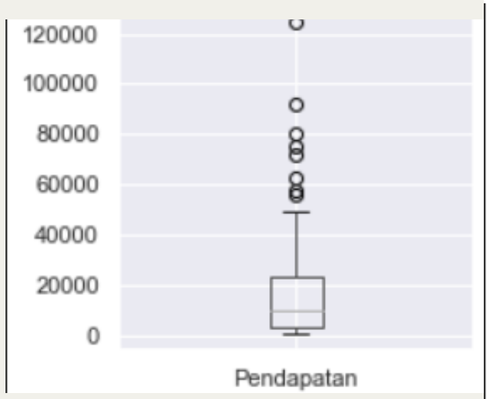


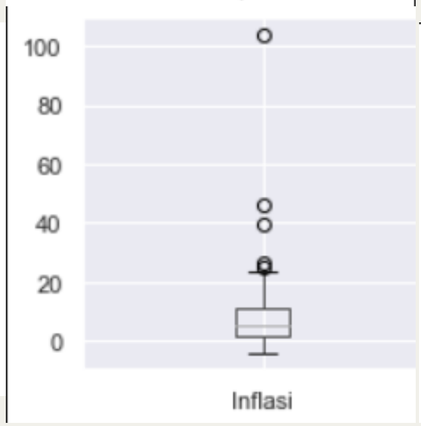


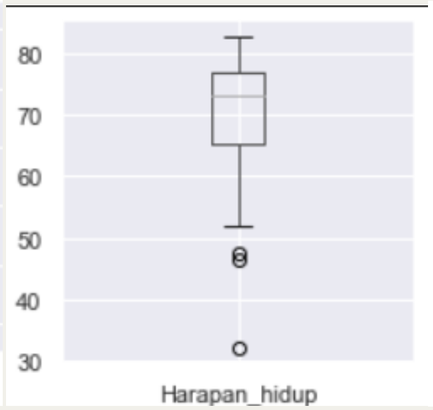


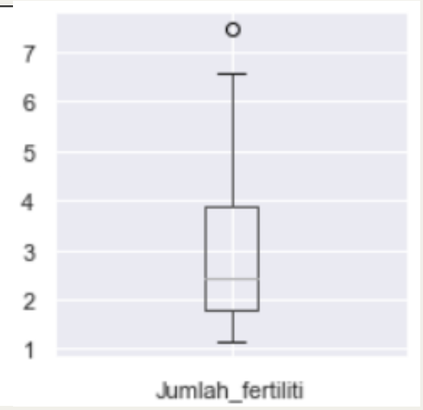






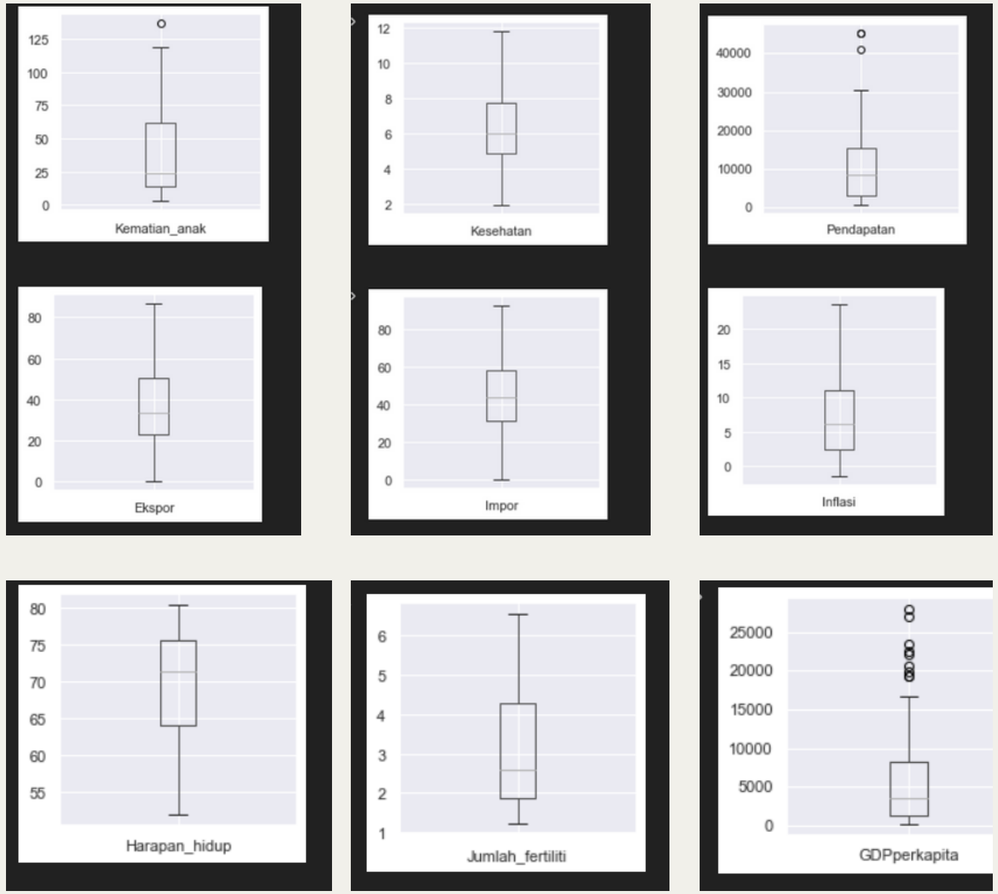






DATA INFORMATION AFTER POSSIBLE OUTLIERS HANDLING

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 125 entries, 0 to 166
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Negara            125 non-null    object
 1   Kematian_anak     125 non-null    float64
 2   Ekspor            125 non-null    float64
 3   Kesehatan         125 non-null    float64
 4   Impor             125 non-null    float64
 5   Pendapatan        125 non-null    int64
 6   Inflasi           125 non-null    float64
 7   Harapan_hidup     125 non-null    float64
 8   Jumlah_fertiliti  125 non-null    float64
 9   GDPperkapita      125 non-null    int64
dtypes: float64(7), int64(2), object(1)
```



SETELAH MELALUI BERBAGAI KONSIDERASI , OUTLIERS TIDAK AKAN DIHANDLING DI KASUS INI

MENGAPA OUTLIERS TIDAK PERLU DI HANDLING DI KASUS INI??

Sebanyak 25% data akan hilang akibat processing outliers ini, mengakibatkan banyaknya lost information baik untuk features dan labels

Berdasarkan research, sebagian besar dari terduga outliers ditemukan merupakan fakta (alias data memang memiliki variansi tinggi)

Data Scaling and Fitting

Modelling inertia , fitting models, Dimensionality Reduction

2

DATA SCALING

Penggunaan Algoritma KMeans memerlukan data scaling karena sistem algoritma ini berbasis nearest data dan data yang masih memiliki interval berbeda-beda butuh melalui tahap scaling dengan StandardScaler()

SCALED DATA

Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapit
Afghanistan	1.291532	-1.138280	0.279088	-0.082455	-0.808245	0.157336	-1.619092	1.902882	-0.67918
Albania	-0.538949	-0.479658	-0.097016	0.070837	-0.375369	-0.312347	0.647866	-0.859973	-0.48562
Algeria	-0.272833	-0.099122	-0.966073	-0.641762	-0.220844	0.789274	0.670423	-0.038404	-0.46537
Angola	2.007808	0.775381	-1.448071	-0.165315	-0.585043	1.387054	-1.179234	2.128151	-0.51626
Antigua and Barbuda	-0.695634	0.160668	-0.286894	0.497568	0.101732	-0.601749	0.704258	-0.541946	-0.04181

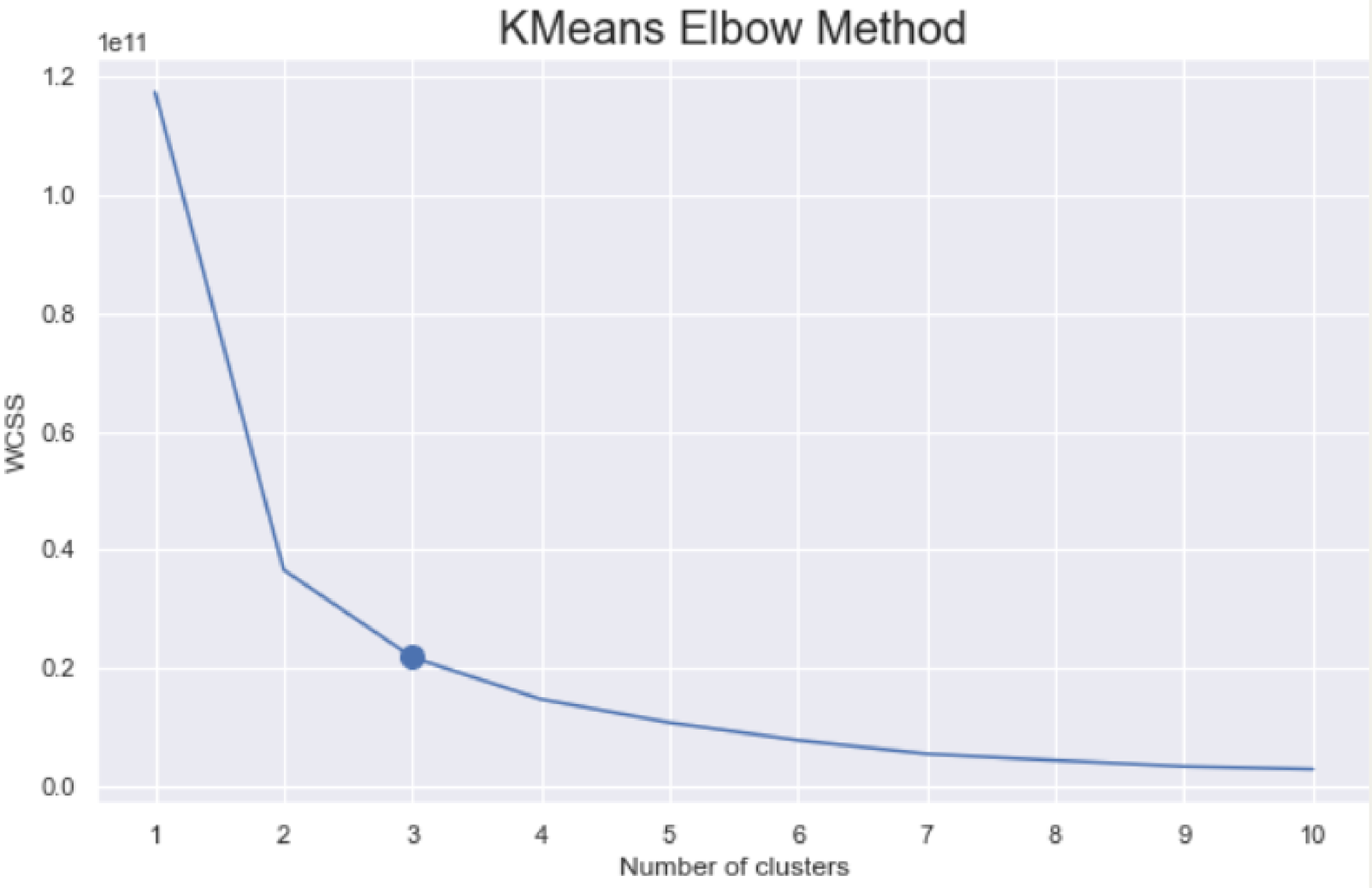
SKLEARN CLUSTER

Diperlukan analisis untuk menentukan n_clusters yang tepat, salah satu nya dengan metode elbow (siku)

```
clusters = []
for i in range(1,11):
    km = KMeans(n_clusters = i).fit(df_clean[df_clean.columns[1:]])
    clusters.append(km.inertia_)

f,ax = plt.subplots(figsize = (10,6))
plt.plot([i for i in range(1,11)] , clusters , markevery = [2] , marker = "o" , markersize = 10)
plt.title("KMeans Elbow Method" ,size= 20)
plt.xlabel("Number of clusters")
plt.ylabel("WCSS")
plt.xticks([i for i in range(1,11)])
plt.show()
```

ELBOW METHOD



MELALUI ANALISIS VISUALISASI ELBOW METHOD , DIDAPATKAN BAHWA ELBOW TERLETAK CENDERUNG PADA N = 3 (TITIK BIRU)

Model Visualization and Results

Visualize custering scatterplot , gathering categorical insights.

3

UNSUPERVISED MODEL

Setelah penentuan n_clusters yang optimal , model Machine Learning dapat dibentuk untuk klasterisasi secara otomatis disertai pemberian label untuk masing-masing negara

Fitting **Model** and Assigning **Labels**

```
km = KMeans(n_clusters = 3).fit(df_clean_scaled[df.columns[1:]])
df_clean_scaled['label'] = km.labels_
```

Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	label
Afghanistan	1.291532	-1.138280	0.279088	-0.082455	-0.808245	0.157336	-1.619092	1.902882	-0.679180	2
Albania	-0.538949	-0.479658	-0.097016	0.070837	-0.375369	-0.312347	0.647866	-0.859973	-0.485623	1
Algeria	-0.272833	-0.099122	-0.966073	-0.641762	-0.220844	0.789274	0.670423	-0.038404	-0.465376	1
Angola	2.007808	0.775381	-1.448071	-0.165315	-0.585043	1.387054	-1.179234	2.128151	-0.516268	2
tigua and Barbuda	-0.695634	0.160668	-0.286894	0.497568	0.101732	-0.601749	0.704258	-0.541946	-0.041817	1

PCA

Visualisasi dalam bentuk 2D mengharuskan kita melakukan deduction untuk features menggunakan Principal Component Analysis dengan statistical approach

```
reduced_data = PCA(n_components=2).fit_transform(df_clean_scaled[df_clean_scaled.columns[1:-1]])

df_final = pd.DataFrame(reduced_data,columns=['pca1','pca2'])
df_final['Negara'] = df_clean_scaled.Negara
df_final['Label'] = df_clean_scaled.label
df_final = df_final[['Negara','pca1','pca2','Label']]

df_final.head()
```

	Negara	pca1	pca2	Label
0	Afghanistan	-2.913025	0.095621	2
1	Albania	0.429911	-0.588156	1
2	Algeria	-0.285225	-0.455174	1
3	Angola	-2.932423	1.695555	2
4	Antigua and Barbuda	1.033576	0.136659	1

CLUSTER VISUALIZATION

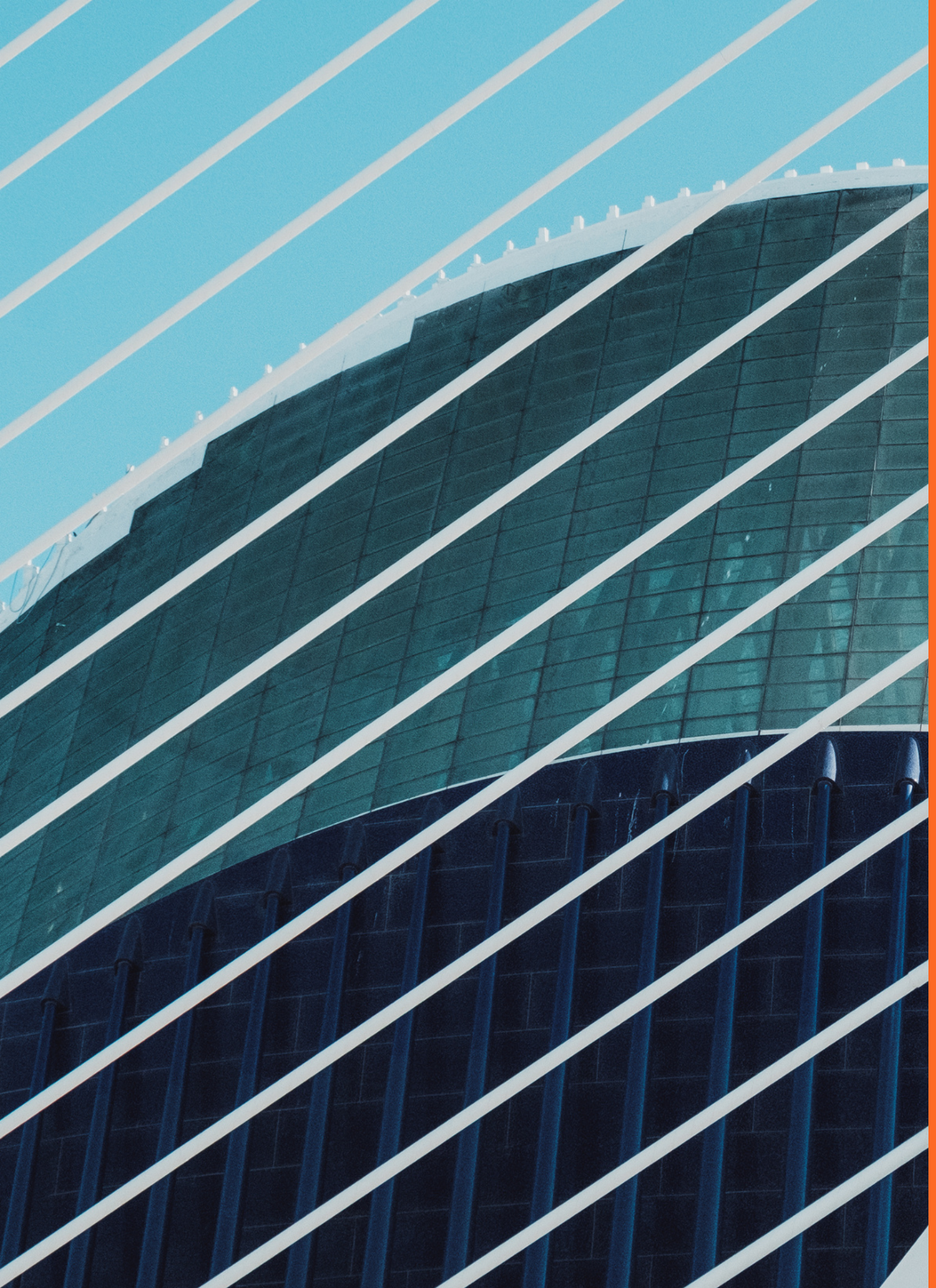
Changing Cluster Labels based on Features

```
df_final['Label'].replace([0,1,2],['Good','Better','Bad'],inplace=True)
```

Kode visualisasi

```
fig,ax = plt.subplots(figsize = (25,30))
sns.scatterplot(x="pca1", y="pca2", hue='Label', data=df_final , palette = ['r','y','g'],s = 120)
countries = df_final.Negara.values.tolist()
for i in df_final.values.tolist():
    plt.text(i[1], i[2], i[0] , size = 10)
plt.legend(loc = "upper center",ncol = 3 ,fontsize = 'xx-large')
plt.show()
```

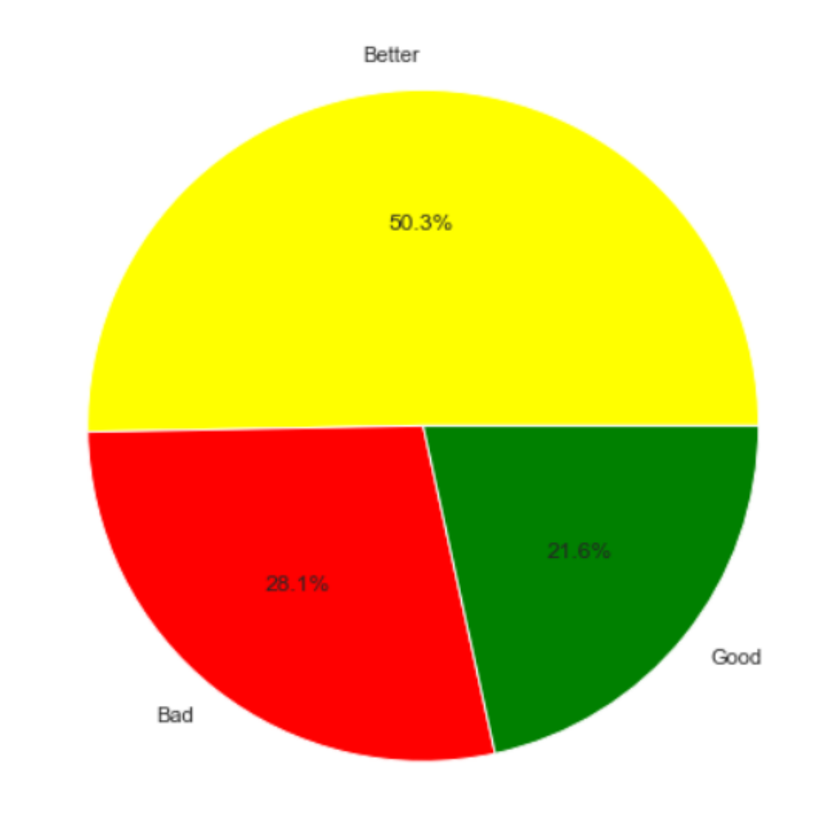
PARTIAL GRAPH (FULL VERSION IN .IPYNB)



SECTION 3

DECISION MAKING USING INSIGHTS

03



PERSEBARAN CLUSTER NEGARA

CLUSTER NEGARA

01

GOOD

Cluster negara dengan kondisi ekonomi , sosial , dan kesehatan yang baik dibandingkan dengan data lainnya

02

GOOD

Cluster negara dengan kondisi ekonomi , sosial , dan kesehatan yang sedang - baik dibandingkan dengan data lainnya

03

BAD

Cluster negara dengan kondisi ekonomi , sosial , dan kesehatan yang buruk dibandingkan dengan data lainnya

NEGARA DENGAN LABEL BAD

28.1%
DARI TOTAL DATA

47 NEGARA

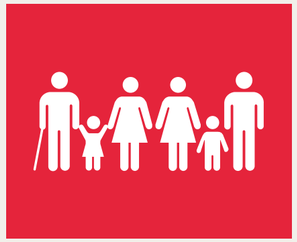
MAYORITAS BENUA AFRIKA

NEGARA DENGAN LABEL BAD

32

AFGHANISTAN ; ANGOLA ; BENIN ; BOTSWANA ; BURKINA FASO ; BURUNDI ; CAMEROON ; CENTRAL AFRICAN REPUBLIC ; CHAD ; COMOROS ; CONGO, DEM. REP. ; CONGO, REP. ; COTE D'IVOIRE ; EQUATORIAL GUINEA ; ERITREA ; GABON ; GAMBIA ; GHANA ; GUINEA ; GUINEA-BISSAU ; HAITI ; IRAQ ; KENYA ; KIRIBATI ; LAO ; LESOTHO ; LIBERIA ; MADAGASCAR ; MALAWI ; MALI ; MAURITANIA ; MOZAMBIQUE ; NAMIBIA ; NIGER ; NIGERIA ; PAKISTAN ; RWANDA ; SENEGAL ; SIERRA LEONE ; SOUTH AFRICA ; SUDAN ; TANZANIA ; TIMOR-LESTE ; TOGO ; UGANDA ; YEMEN ; ZAMBIA

NEXT STEPS



01

Elaborasi

mengelaborasi kesimpulan dengan CEO HELP international

02

Diskusi

Menentukan action step yang harus ditempuh untuk meningkatkan kualitas negara tertera

03

Aksi

Memastikan bantuan dan donasi sampai ke negara-negara yang membutuhkan

THANK YOU