Final Project

Clustering the Countries by using K-Means for HELP International

Reporter: RETNO W

TENTANG ORGANISASI

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

Dana Bantuan yang Terkumpul oleh HELP

$10.000.000

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.

Fitur Data

Menentukan fitur data menjadi dua kategori yakni **sosial ekonomi** dan **kesehatan**

Hasil

List negara-negara yang layak mendapatkan bantuan dari HELP

Data

Data yang digunakan merupakan data dari HELP Internasional.
Data_Negara_HELP.cs
v

Perangkat

Mengolah data menggunakan
Python 3.12.3

TAHAPAN

1. Membaca data file csv, dan menampilkan informasi isi dari file dengan df.info()
2. Pengecekan apakah terdapat data kosong atau missing value dan data duplikat
3. Pengecekan korelasi dari setiap data. Menentukan data berkorelasi tinggi, berhubungan dengan **sosial ekonomi** serta **kesehatan**
4. Melakukan pengecekan apakah terdapat data outlier dengan boxplot
5. Melakukan handling outlier
6. Clustering data dengan K-Means

1

Membaca data file csv, dan menampilkan informasi isi dari file dengan df.info()

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100				200

Hasil dari df.info()

```
df.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)
memory usage: 13.2+ KB
```

2

Pengecekan apakah terdapat data kosong atau missing value dan data duplikat

Pengecekan jumlah data kosong atau NAN bisa dilakukan dengan **df.isna().sum()**. Sedangkan pengecekan data duplikat bisa dilakukan dengan **df.duplicated().sum()**

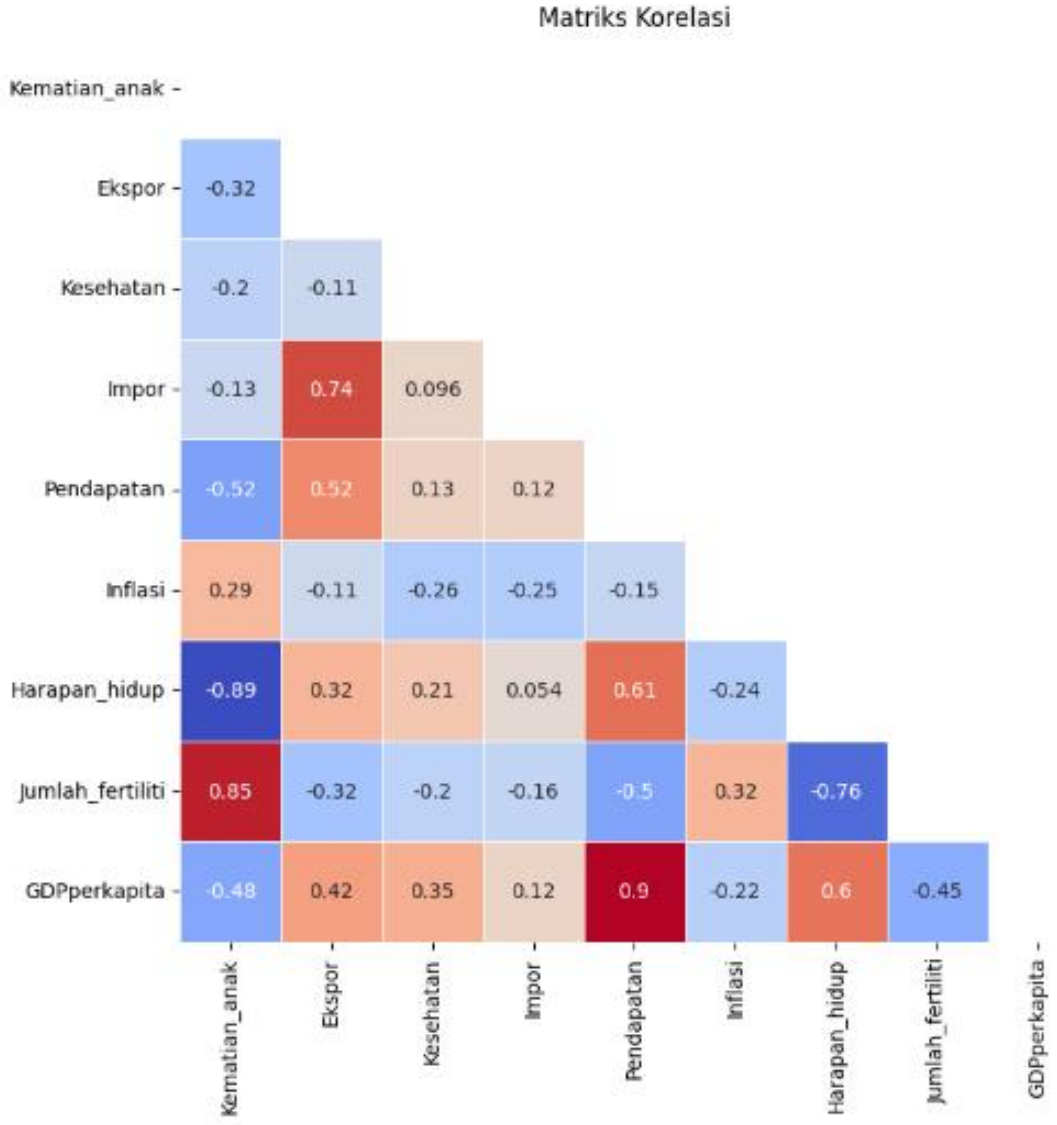
Dari hasil tersebut jumlah data kosong ataupun data duplikat adalah 0

```
print(df.isna().sum())
print('Jumlah data duplikat :', df.duplicated().sum())
```

```
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
Jumlah data duplikat : 0
```

3

Pengecekan korelasi dari setiap data. Menentukan data berkorelasi tinggi, berhubungan dengan **sosial ekonomi** serta **kesehatan**



Dari gambar disampig dapat dilihat bahwa korelasi positif tersinggi adalah

1. Pendapatan dengan GDPperkapita dengan korelasi 0,9
2. Kematian_Anak dan Jumlah_fertiliti dengan korelasi 0,85
3. Ekspor dan Impor dengan korelasi 0,74
4. Pendapatan dan Harapan_hidup dengan korelasi 0,61

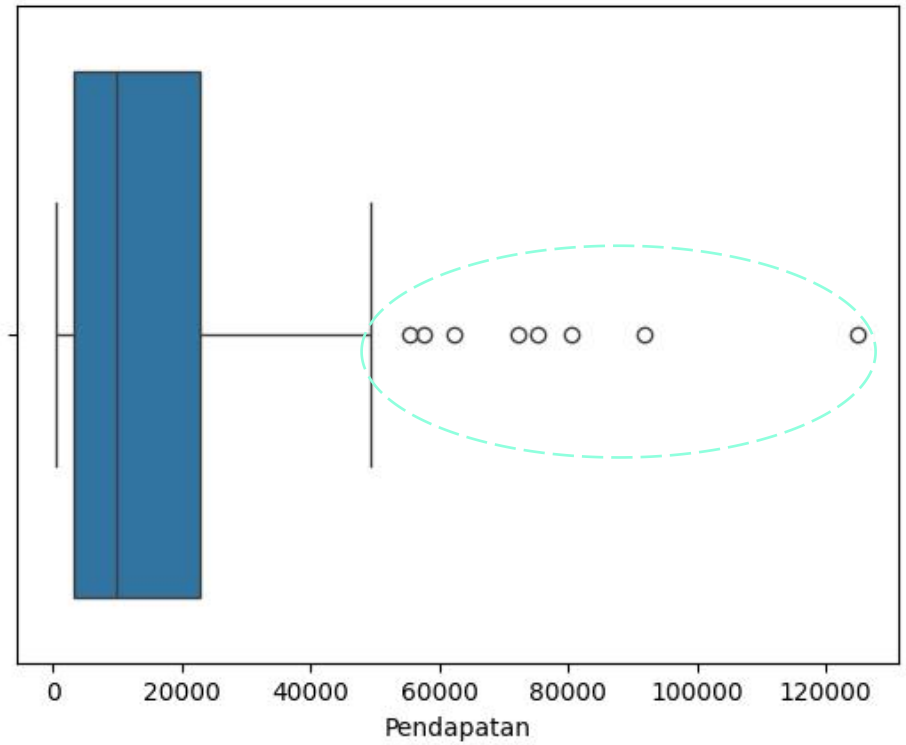
Dari data tersebut korelasi keempat yang digunakan. Dengan ketentuan Pendapatan mewakili Sosial ekonomi dan Harapan_hidup mewakili Kesehatan

4

Melakukan pengecekan apakah terdapat data outlier dengan boxplot

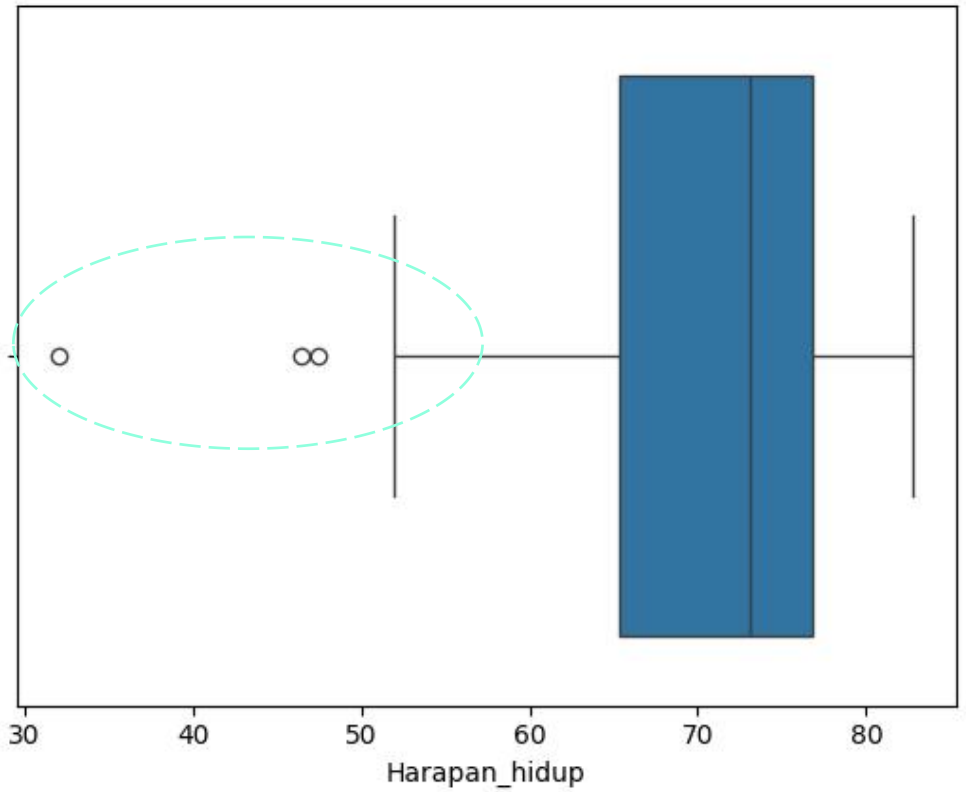
```
sns.boxplot(x='Pendapatan', data=df)
```

```
<Axes: xlabel='Pendapatan'>
```




```
sns.boxplot(x='Harapan_hidup', data=df)
```

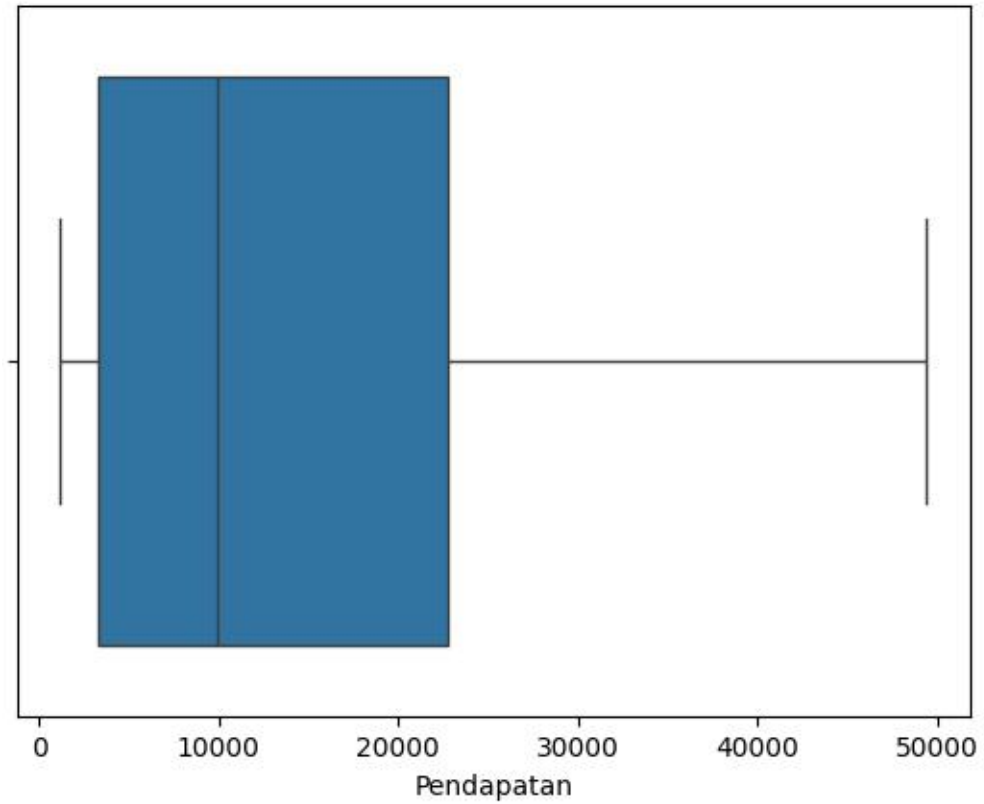
```
<Axes: xlabel='Harapan_hidup'>
```

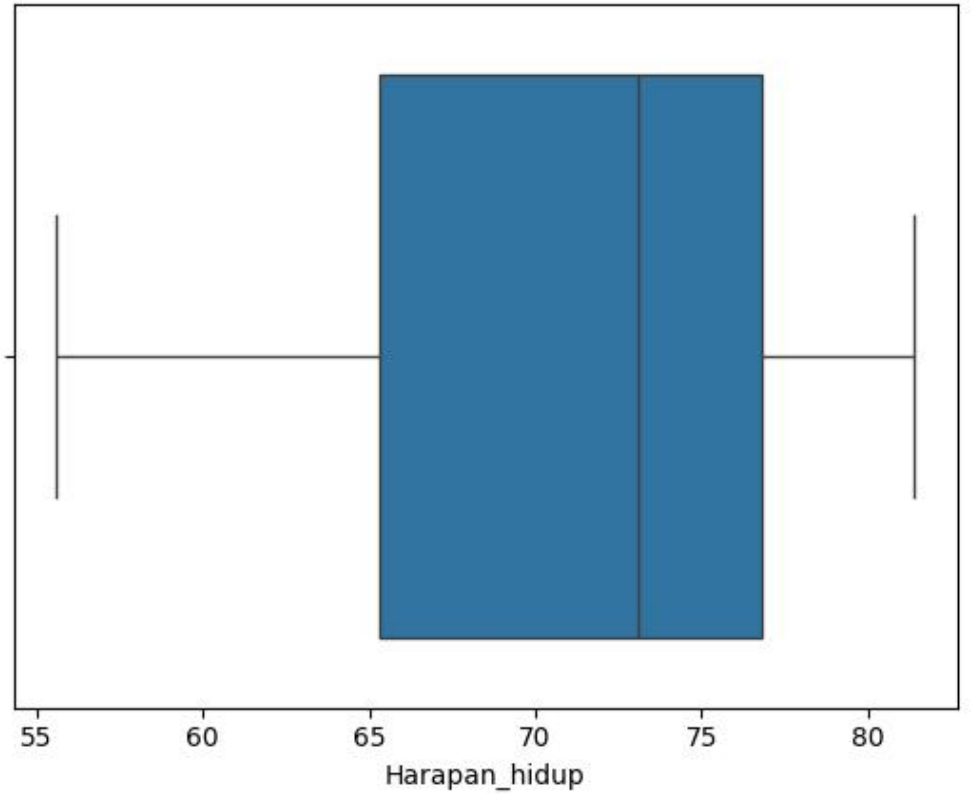


Terdapat data outlier dengan tanda warna hijau

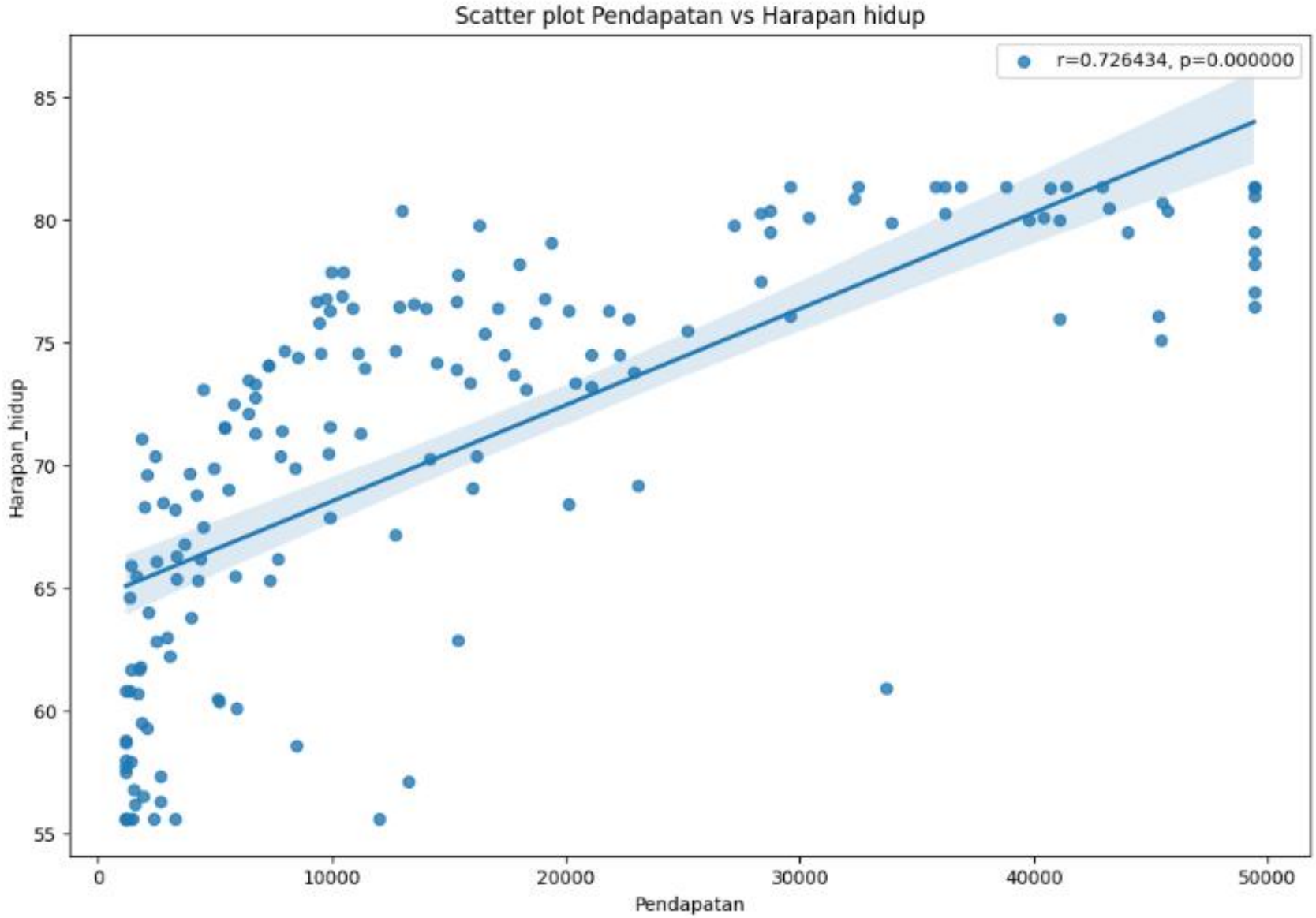
5 Melakukan handling outlier

Setelah dilakukan Winsorizing dimana data outlier tidak dihapus karena data sekecil apapun berpengaruh dengan hasil. Sehingga transformasi statistik dengan membatasi nilai ekstrim dalam data statistik untuk mengurangi efek kemungkinan data outlier dilakukan.



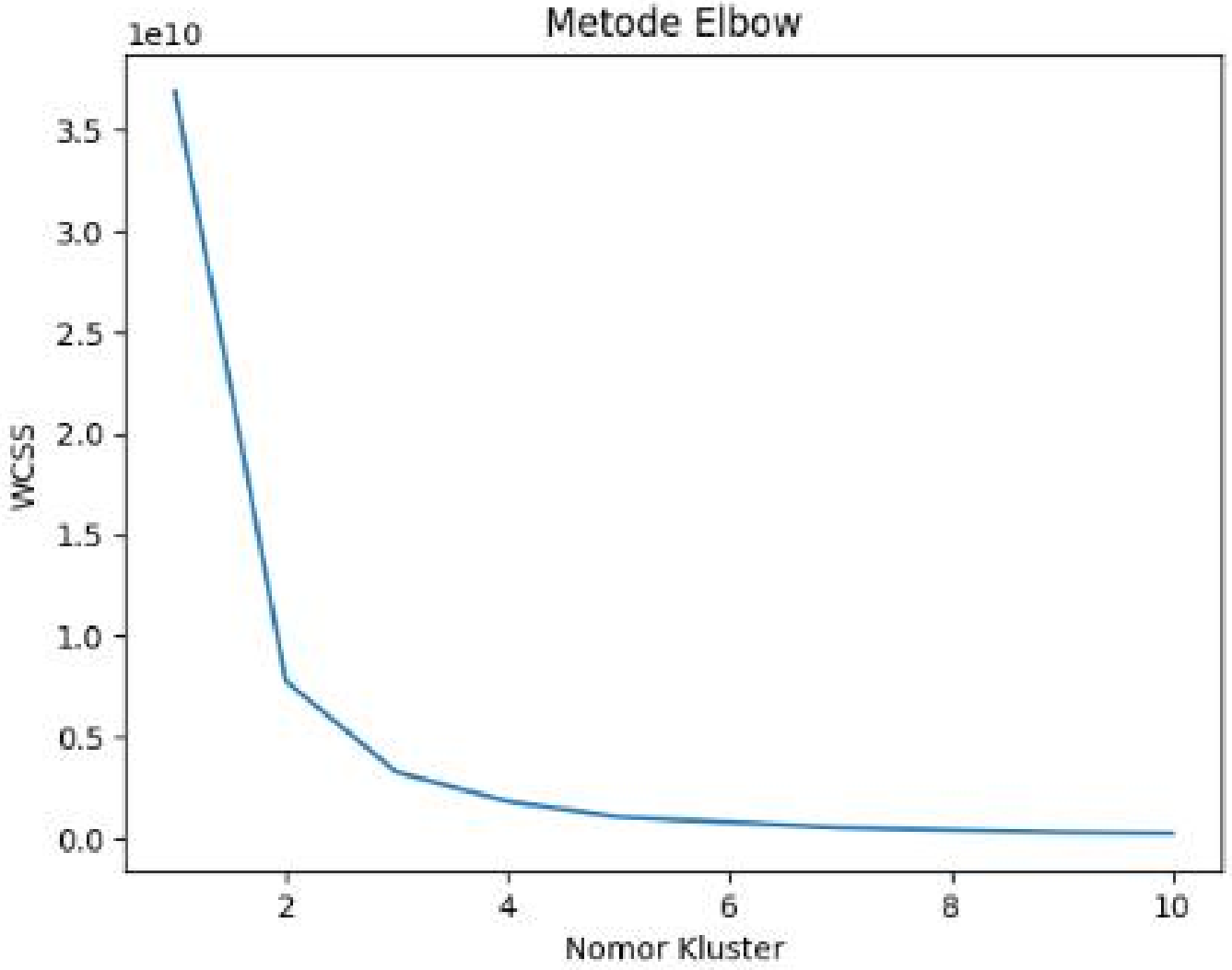


Hasil Scatter Plot dengan nilai r=0,726434 yang menunjukkan korelasi Pendapatan dan Harapan_hidup tinggi.

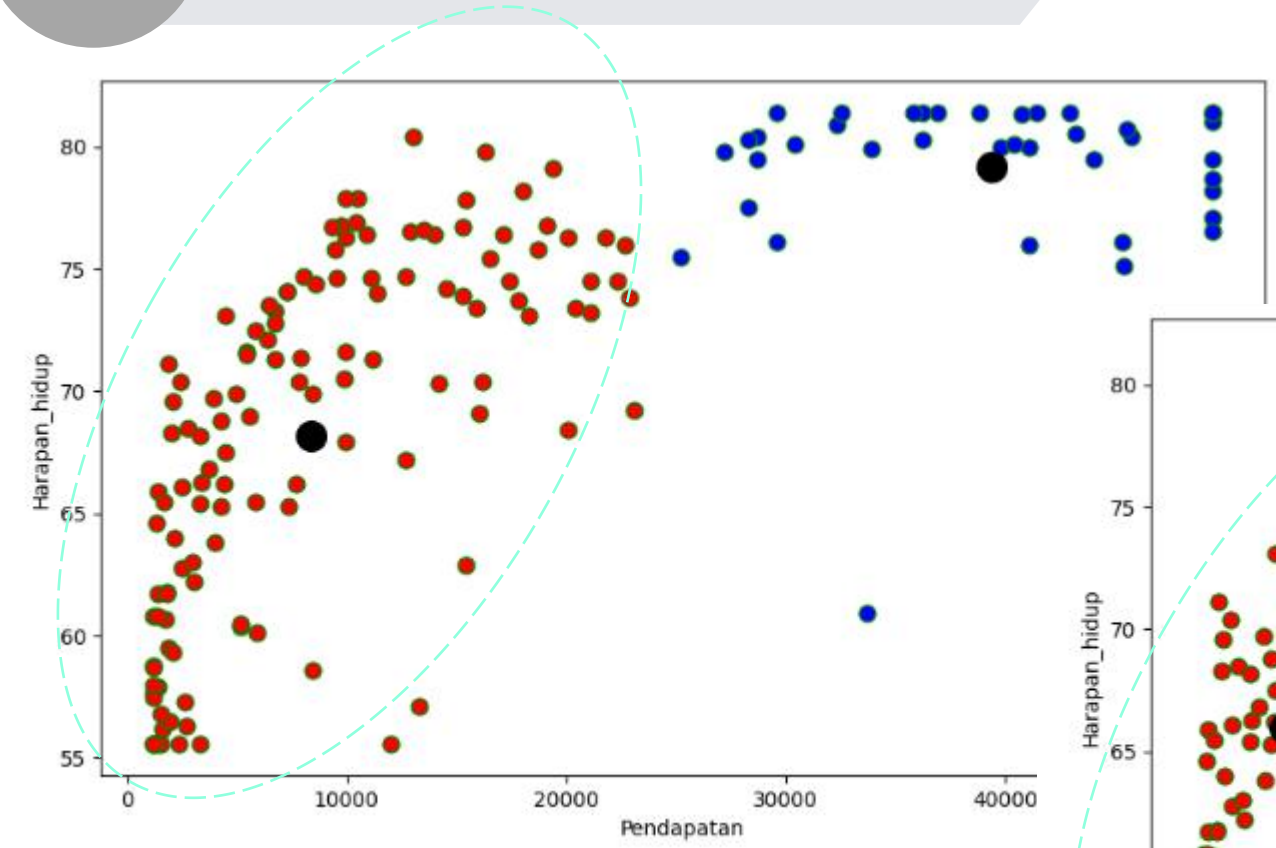


6 Clustering data dengan K-Means

Menentukan Nilai K dengan metode Elbow, di mana dari data di samping nilai K bisa di tentukan 2 ataupun 3. Disini saya melihat hasil dari kedua nilai K

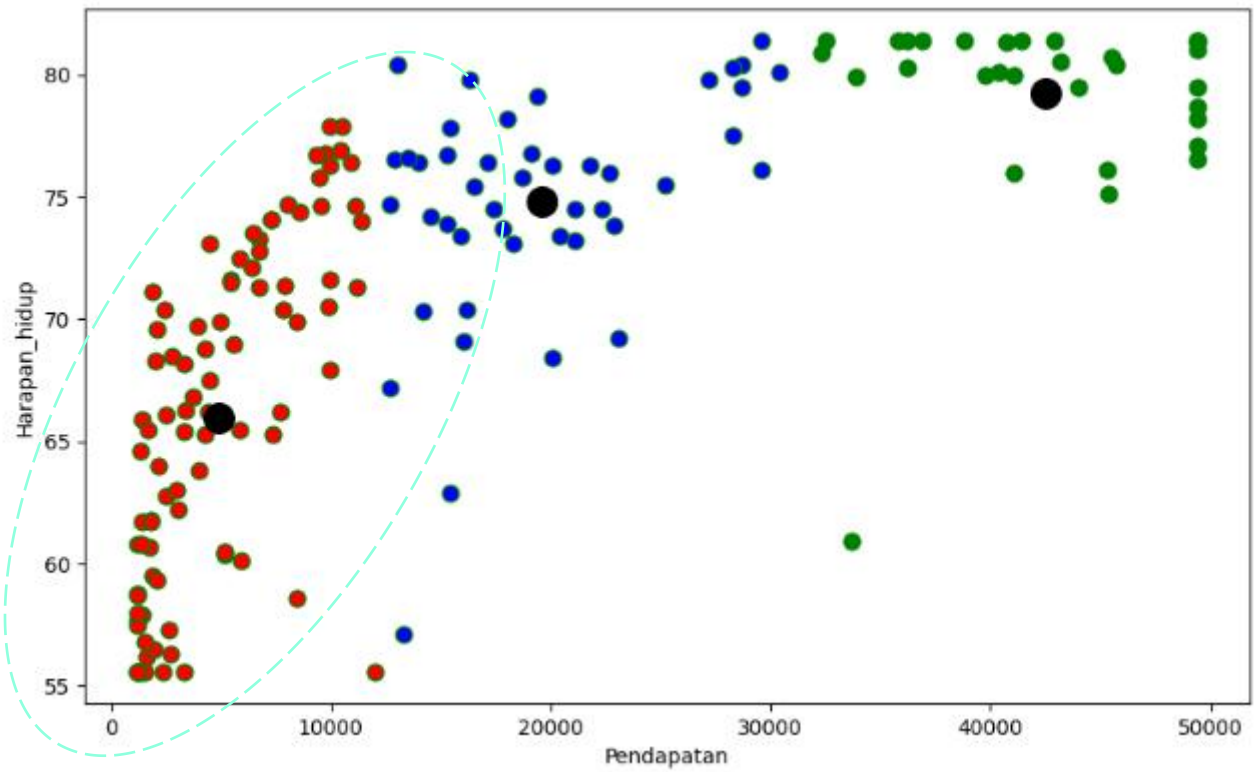


6 Clustering data dengan K-Means



dengan K=2

dengan K=3



6 Clustering data dengan K-Means

Dari hasil diatas didapatkan data negara-negara yang memiliki Pendapatan dan Harapan_hidup rendah hingga yang tinggi. Lima negara terendah yaitu Central African Republic, Malawi, Mozambique, Congo, Dem. Rep. dan Burundi

Negara	Pendapatan	Harapan_hidup
Central African Republic	1210.0	55.6
Malawi	1210.0	55.6
Mozambique	1210.0	55.6
Congo, Dem. Rep.	1210.0	57.5
Burundi	1210.0	57.7

HASIL DATA

Dari hasil diatas didapatkan data negara-negara yang memiliki Pendapatan dan Harapan_hidup rendah hingga yang tinggi. Lima negara terendah yaitu Central African Republic, Malawi, Mozambique, Congo, Dem. Rep. dan Burundi

Negara	Pendapatan	Harapan_hidup
Central African Republic	1210.0	55.6
Malawi	1210.0	55.6
Mozambique	1210.0	55.6
Congo, Dem. Rep.	1210.0	57.5
Burundi	1210.0	57.7

Final Project

TERIMAKASIH

Reporter: RETNO W