FINAL PROJECT SANBERCODE

Muhammad Firman Alamsyah

PRESENTATION OUTLINE:

- ✓Background
- ✓Objective
- ✓Methods
- ✓Result & Analysis
- ✓Conclusion

Background

Background

- **About Organization**:

HELP International is an international humanitarian NGO *commited to fighting poverty and providing basic facilities to people in underdeveloped countries* during disasters and natural disasters.

- **Problem**:

HELP International has raised $10 million and need to make a *decision to select the countries that need the most help.*

Objective

Project’s Objective

- ✓To chategorize countries by socio-economic & health aspects that determine the overall development of the country.

- ✓To determine the countries that this organization should focus on based on the clustering result.

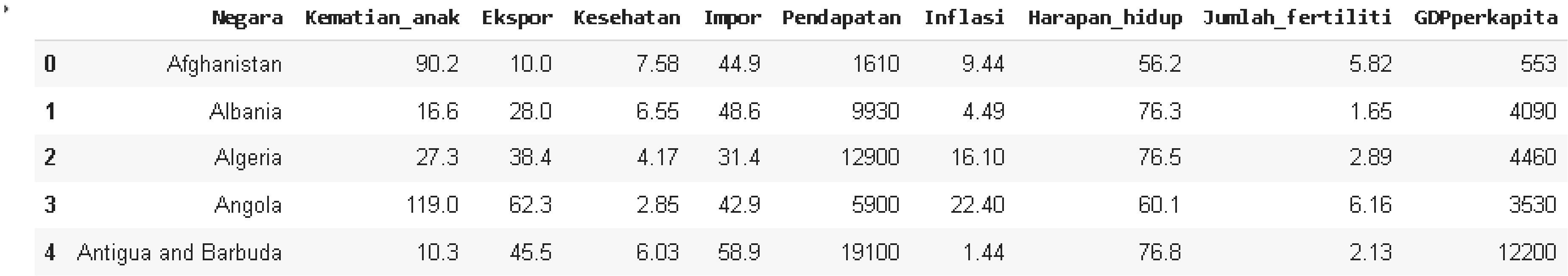
Methods

Methods

1. Exploratory Data Analysis
   - Reading and Understanding data
   - Data Cleaning
   - Univariate Analysis
   - Bivariate Analysis
   - Multivariate Analysis
2. Outliers Treatment
3. Data Clustering

Exploratory Data Analysis

Reading and Understanding data



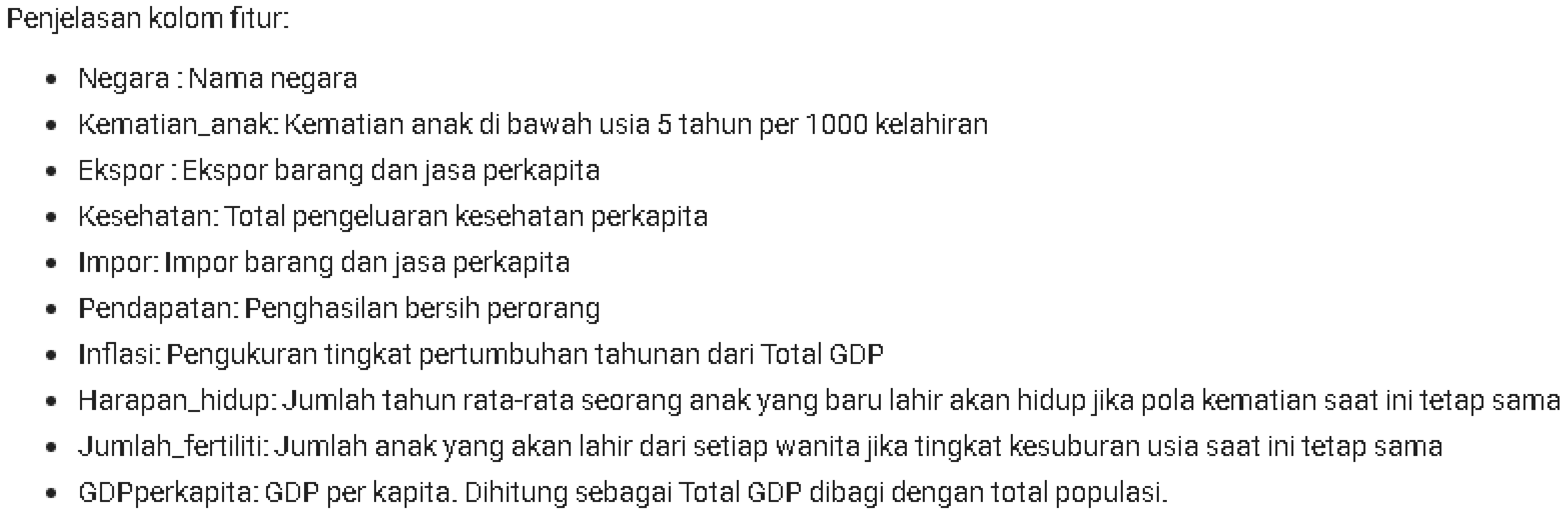
	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Data_Negara_HELP.csv

Data has 167 rows and 10 columns

Exploratory Data Analysis

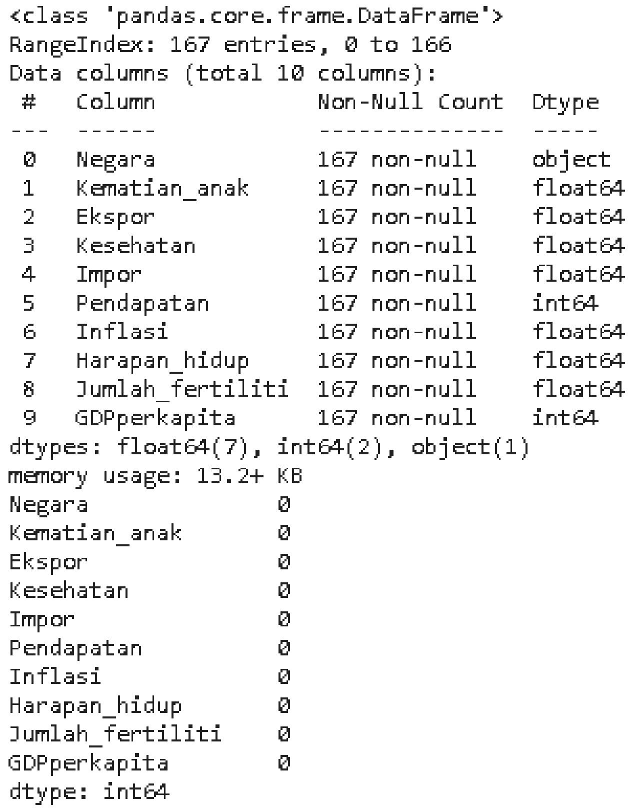
Reading and Understanding data



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.

Description of each features

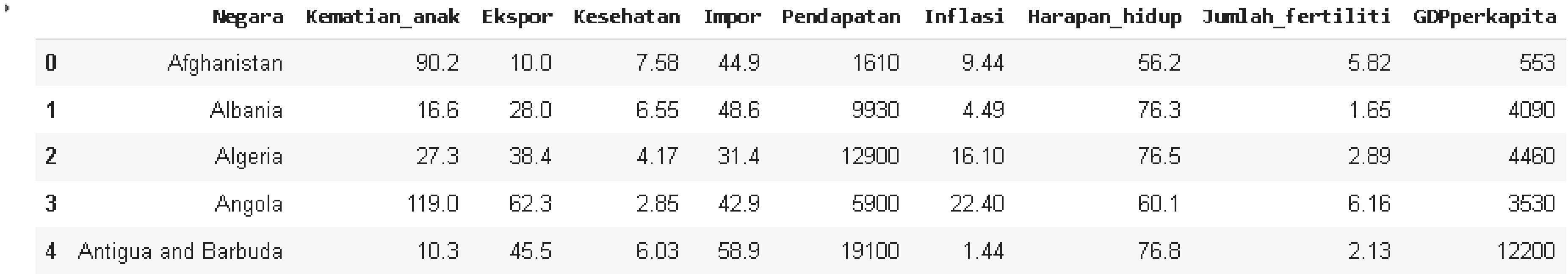


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Negara            167 non-null    object 
 1   Kematian_anak     167 non-null    float64
 2   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
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
```

1 object, 7 float, and 2 integer data types

Exploratory Data Analysis

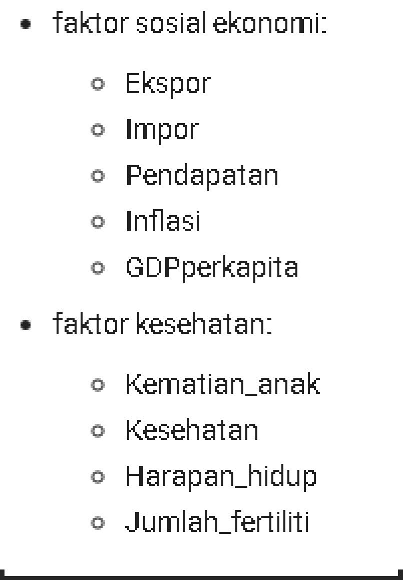
Reading and Understanding data



	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	1.44	76.8	2.13	12200

There are two feature classified:

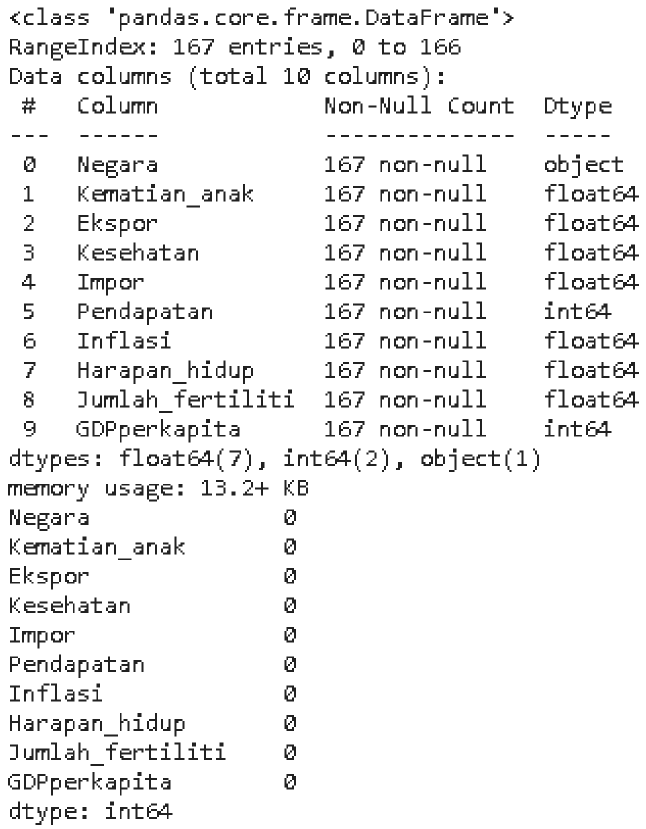
- Socio-Economic Aspect
- Health Aspect



- faktor sosial ekonomi:
  - Ekspor
  - Impor
  - Pendapatan
  - Inflasi
  - GDPperkapita
- faktor kesehatan:
  - Kematian_anak
  - Kesehatan
  - Harapan_hidup
  - Jumlah_fertiliti

Exploratory Data Analysis

Data Cleaning

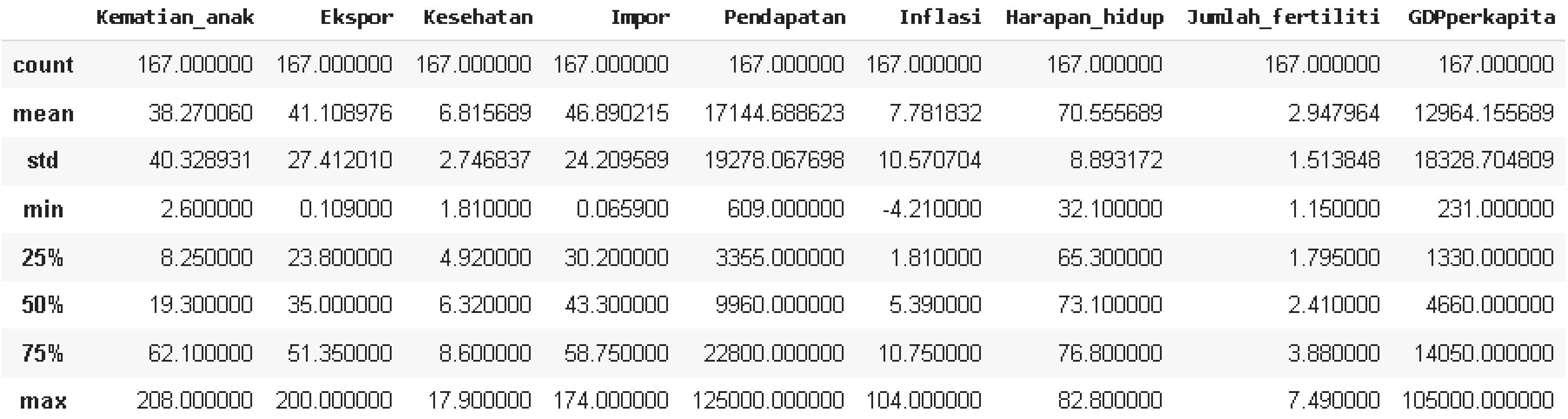


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Negara            167 non-null    object
 1   Kematian_anak     167 non-null    float64
 2   Ekspor            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
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
```

There are no missing value for each feature. So, Handling missing value is not conducted

Exploratory Data Analysis

Univariate Analysis

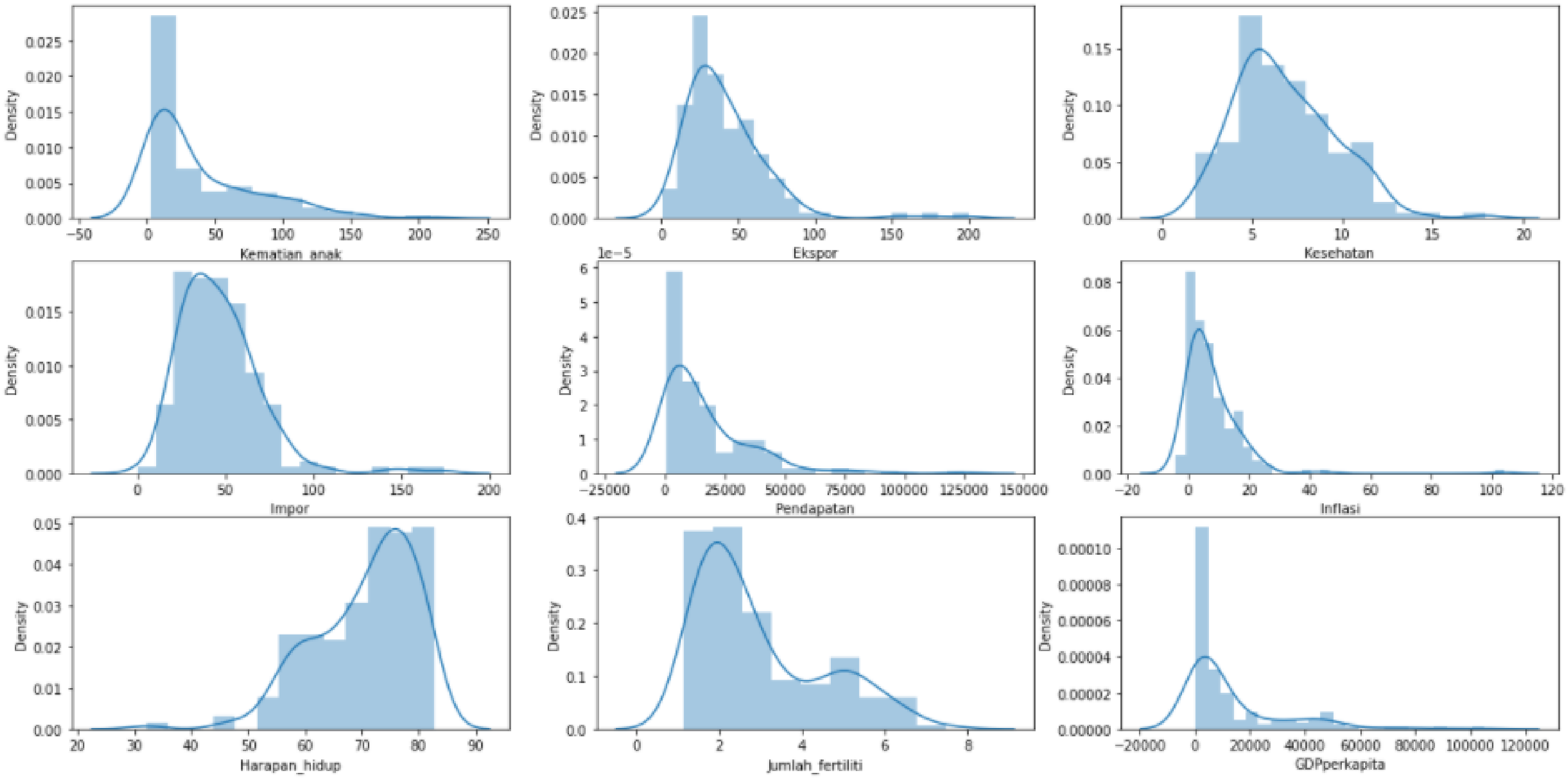


	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
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

These are the descriptive statistics value for each parameter, there are mean, deviation standard, minimum, Q1, Q2, Q3, and maximum value for each parameter.

Exploratory Data Analysis

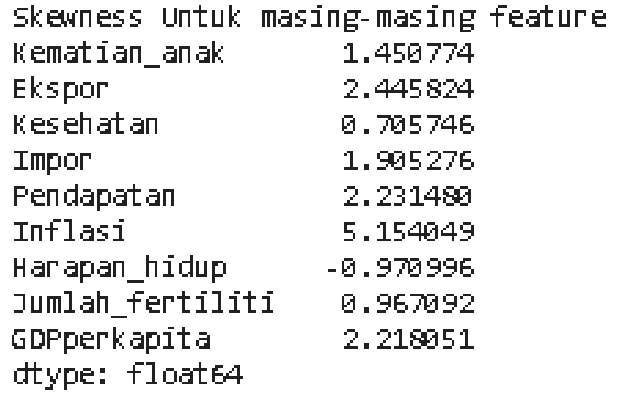
Univariate Analysis



Plot each parameter on histogram to **determine the distribution of the data (skewness and kurtosis)**

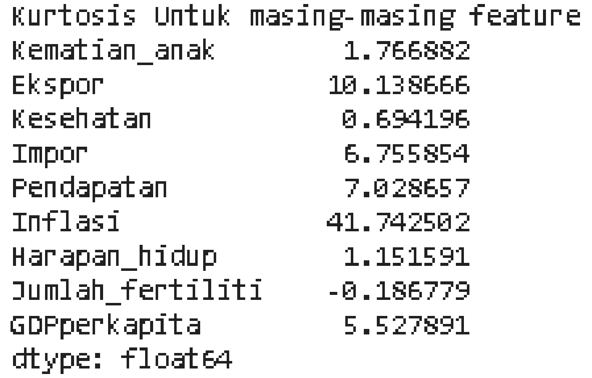
Exploratory Data Analysis

Univariate Analysis

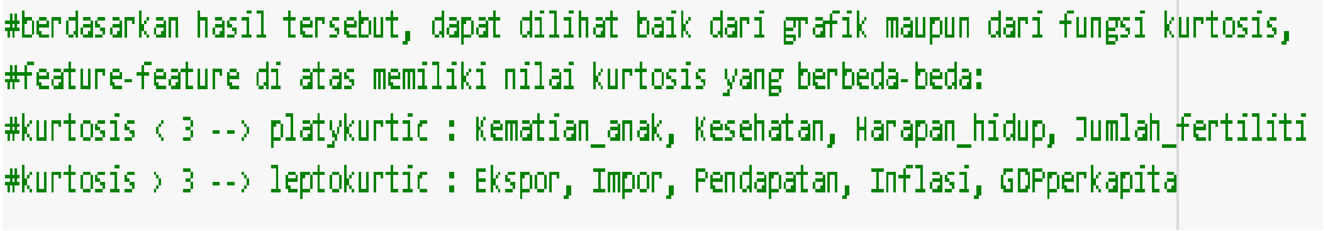


```
Skewness Untuk masing-masing feature
Kematian_anak       1.450774
Ekspor              2.445824
Kesehatan           0.705746
Impor               1.905276
Pendapatan          2.231480
Inflasi             5.154049
Harapan_hidup      -0.970996
Jumlah_fertiliti    0.967092
GDPperkapita        2.218051
dtype: float64
```

Based on the graph and skewness calculation, **each features have a positively skew except “Harapan_hidup” that has a negative value for skewness**



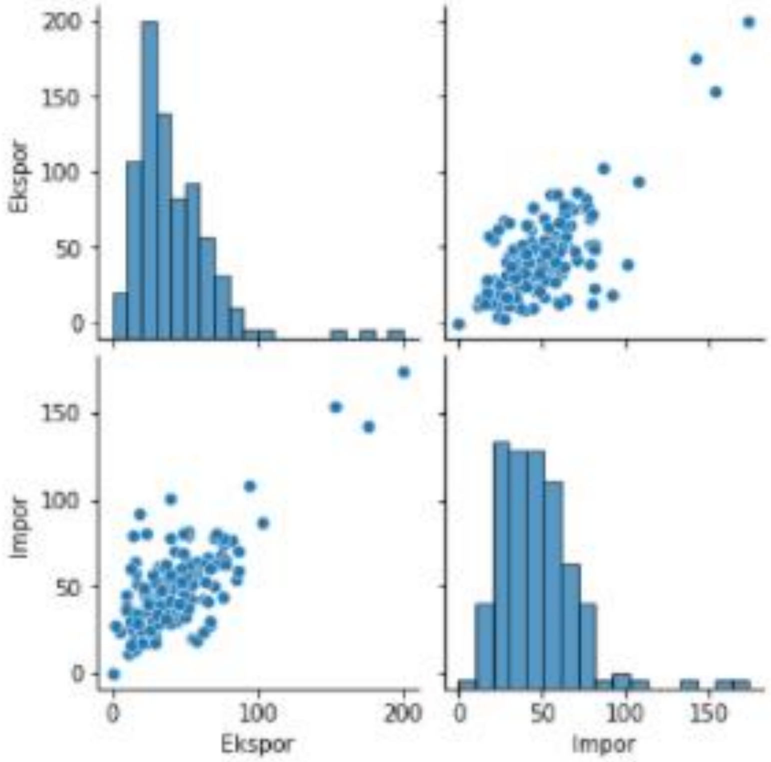
```
Kurtosis Untuk masing-masing feature
Kematian_anak        1.766882
Ekspor              10.138666
Kesehatan            0.694196
Impor                6.755854
Pendapatan           7.028657
Inflasi             41.742502
Harapan_hidup        1.151591
Jumlah_fertiliti    -0.186779
GDPperkapita         5.527891
dtype: float64
```



```
#berdasarkan hasil tersebut, dapat dilihat baik dari grafik maupun dari fungsi kurtosis,
#feature-feature di atas memiliki nilai kurtosis yang berbeda-beda:
#kurtosis < 3 --> platykurtic : Kematian_anak, Kesehatan, Harapan_hidup, Jumlah_fertiliti
#kurtosis > 3 --> leptokurtic : Ekspor, Impor, Pendapatan, Inflasi, GDPperkapita
```

Exploratory Data Analysis

Bivariate Analysis

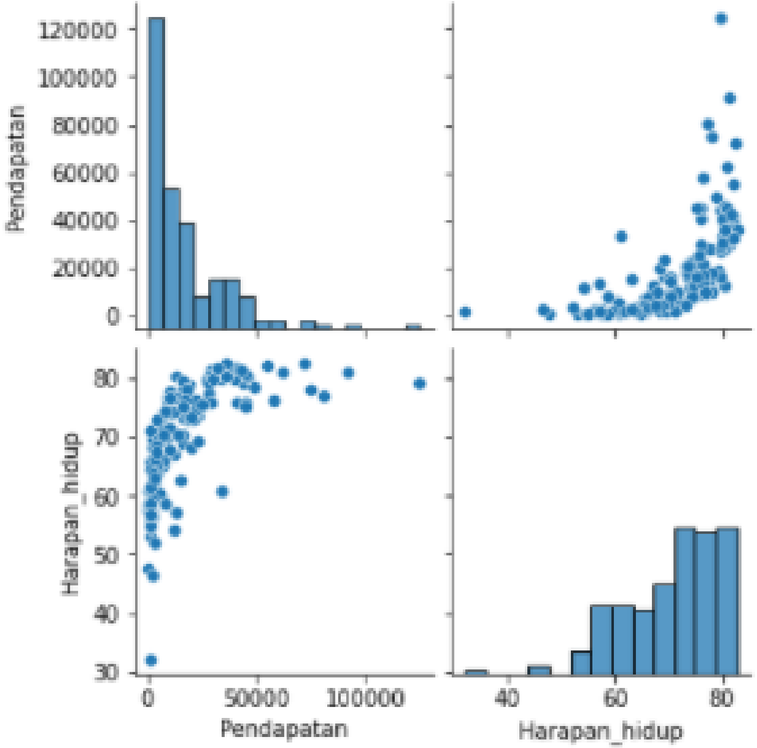


For bivariate analysis, take a sample with choose ‘Impor’ and ‘Ekspor’ to be analyzed

As we see from the scatter plot, **Impor and Ekspor variable have a linear correlation.**

Exploratory Data Analysis

Bivariate Analysis

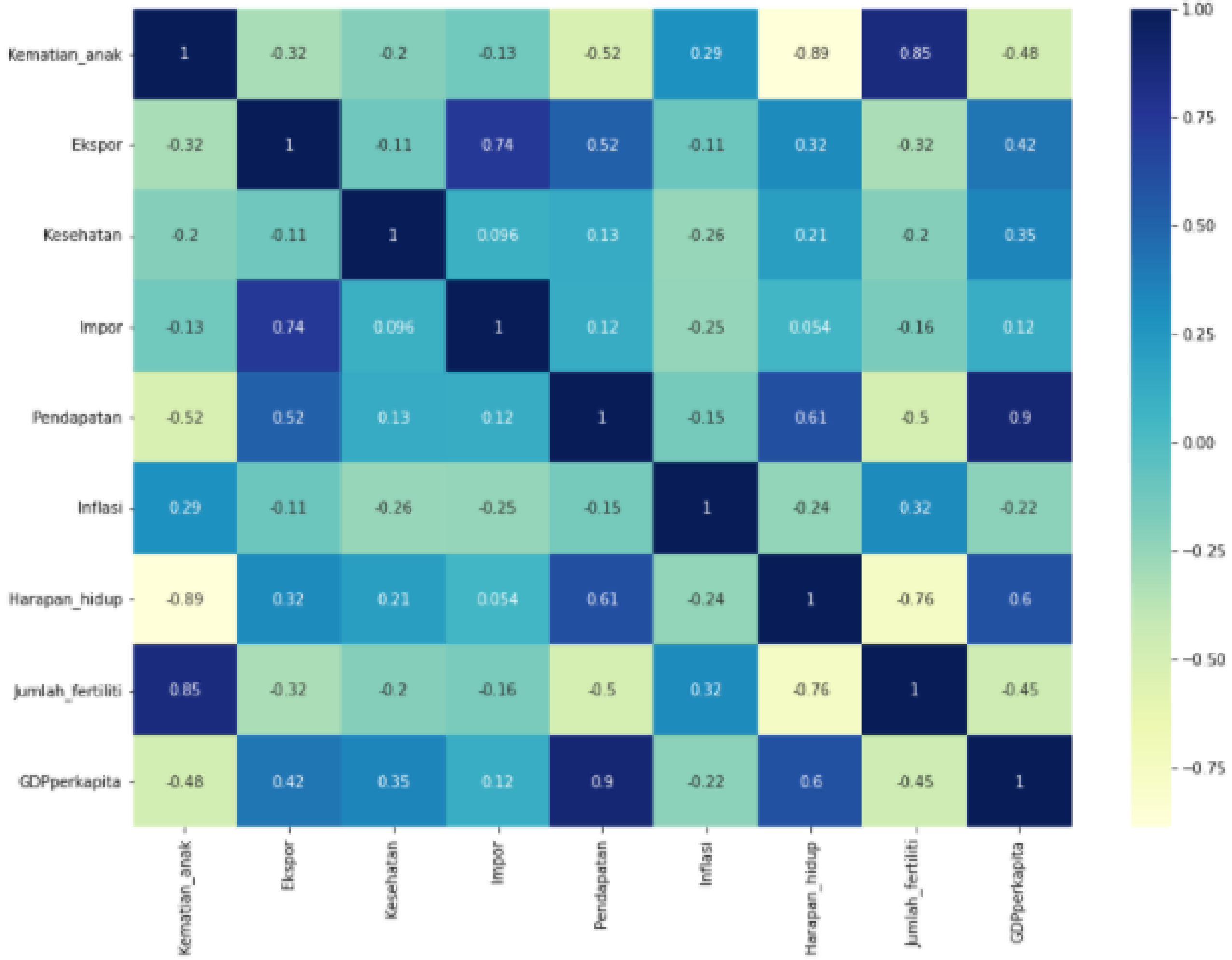


Another sample with choose ‘Pendapatan’ and ‘Harapan_hidup’ variables to be analyzed

As we see from the scatter plot, **‘Pendapatan’ and ‘Harapan_hidup’ variable have a medium linear correlation. And also has asimptotik value for Harapan_hidup.**

Exploratory Data Analysis

Multivariate Analysis

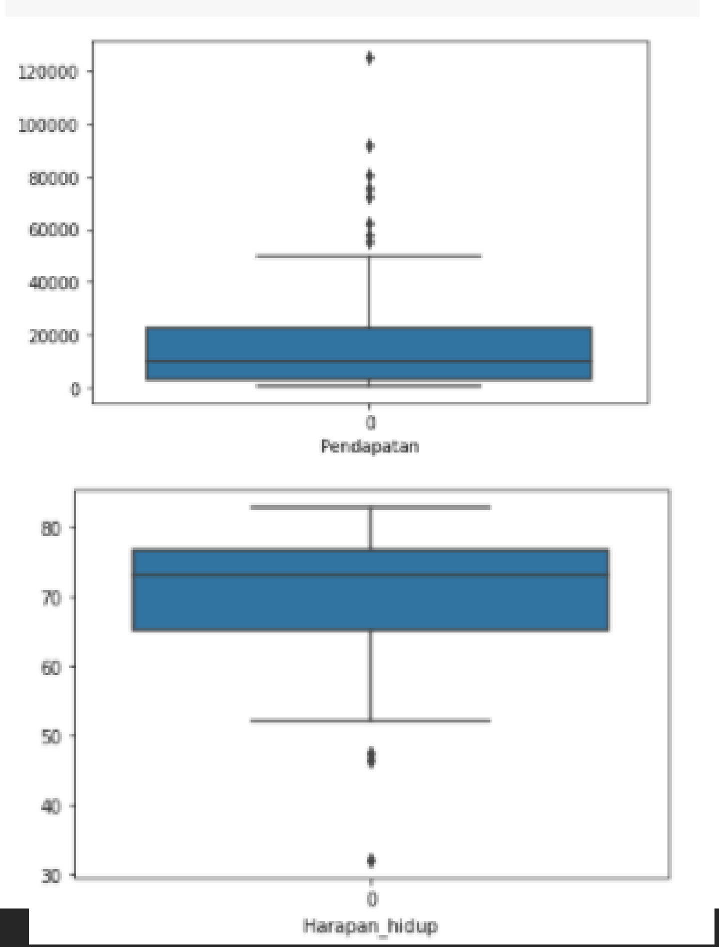


From Multivariate Analysis we can see linear correlation for each parameter and its value. From the heatmap we can choose the variable that we want to use in clustering proccess. Variable that has a good value for linear correlation is:

- Pendapatan & Jumlah_Fertiliti (-0.5)
- Kematian_anak & pendapatan(-0.57)
- Harapan_hidup & GDPperkapita(0.6)
- **Harapan_hidup & Pendapatan(0.61)**

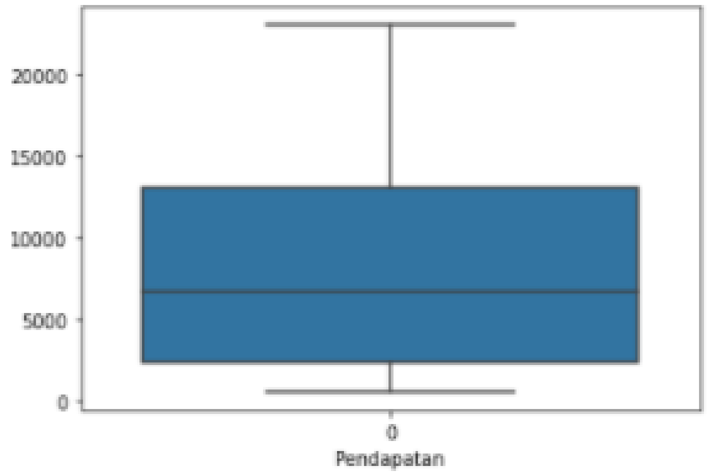
We choose Harapan_hidup & Pendapatan with a highest linear correlation

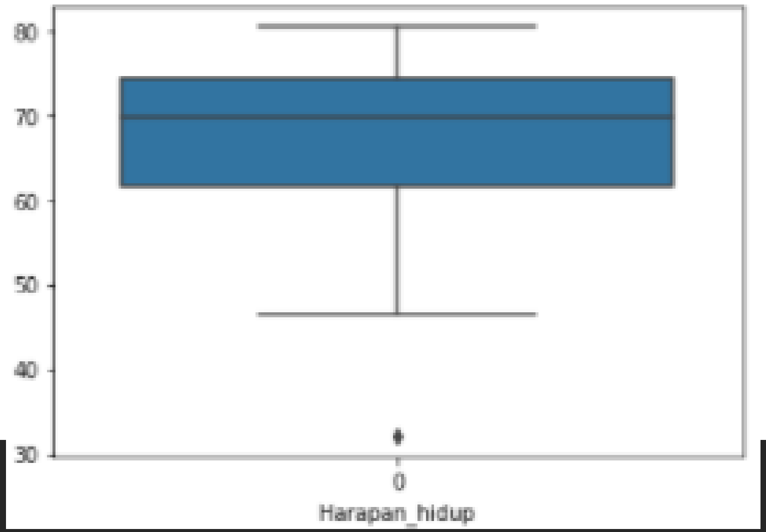
Outliers Treatment



From boxplot we can see there are outliers in ‘Pendapatan’ variable in higher than upper bound. And also in the Harapan_hidup variable there are outliers in lower than lower bound. **So we need to replace the outlier with the upper and lower band value to optimize the clustering process.**

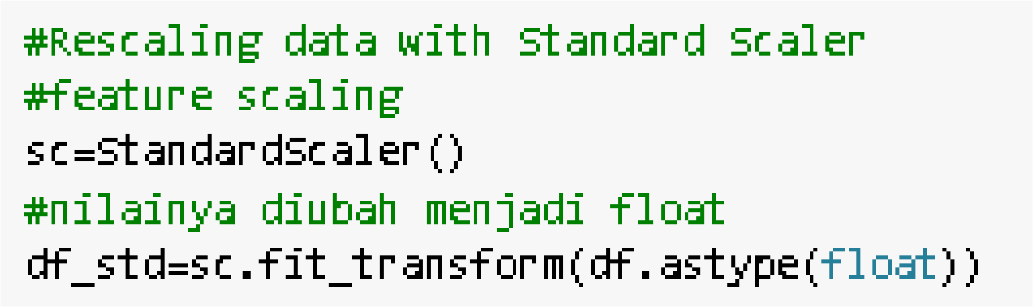
Outliers Treatment





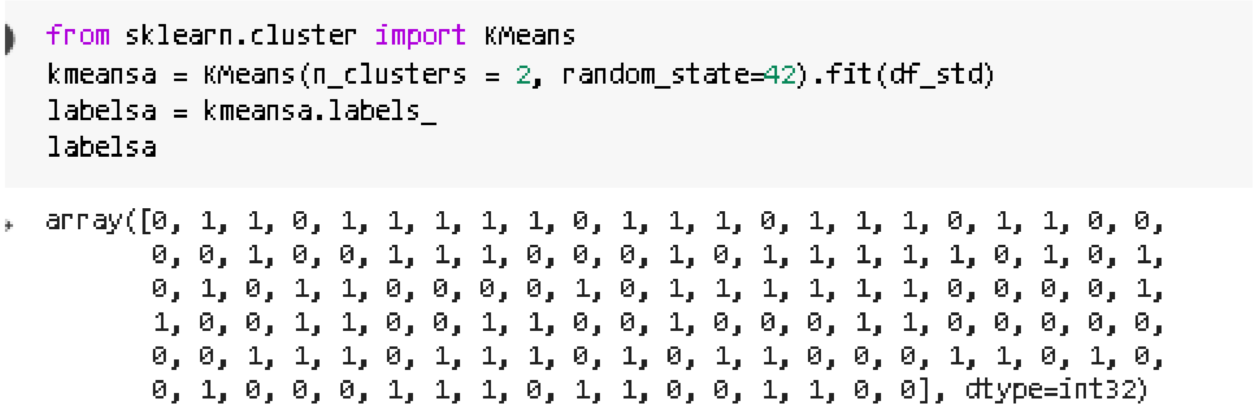
After handling outliers with inter quartile method, ‘pendapatan’ variable has been removed the outliers. But Harapan_hidup still has outliers from the boxplot. We can ignore the outliers from the Harapan_hidup boxplot, **because the outliers doesn’t too significant from lower bound value.**

Data Clustering



```
#Rescaling data with Standard Scaler
#feature scaling
sc=StandardScaler()
#nilainya diubah menjadi float
df_std=sc.fit_transform(df.astype(float))
```

Rescalling data to create variable value to be a z score and calculate the distance from center point.

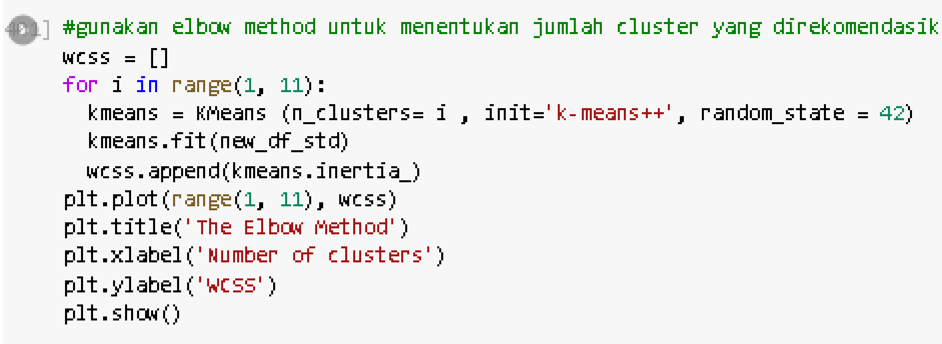


```
from sklearn.cluster import KMeans
kmeansa = KMeans(n_clusters = 2, random_state=42).fit(df_std)
labelsa = kmeansa.labels_
labelsa
```

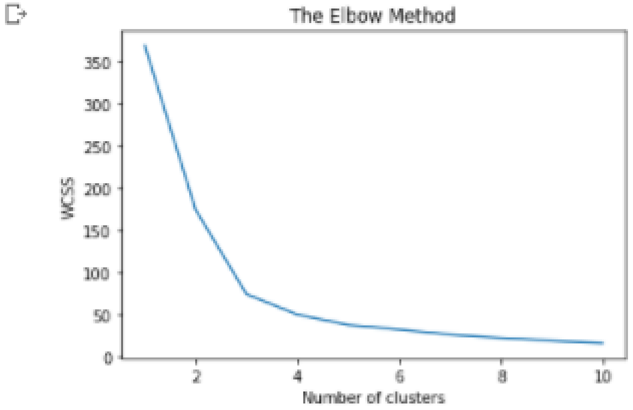
```
array([0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0,
       0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1,
       1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0,
       0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0], dtype=int32)
```

Clustering with Kmeans and inverse it to get a real value from each feature. Use 2 cluster as a default cluster

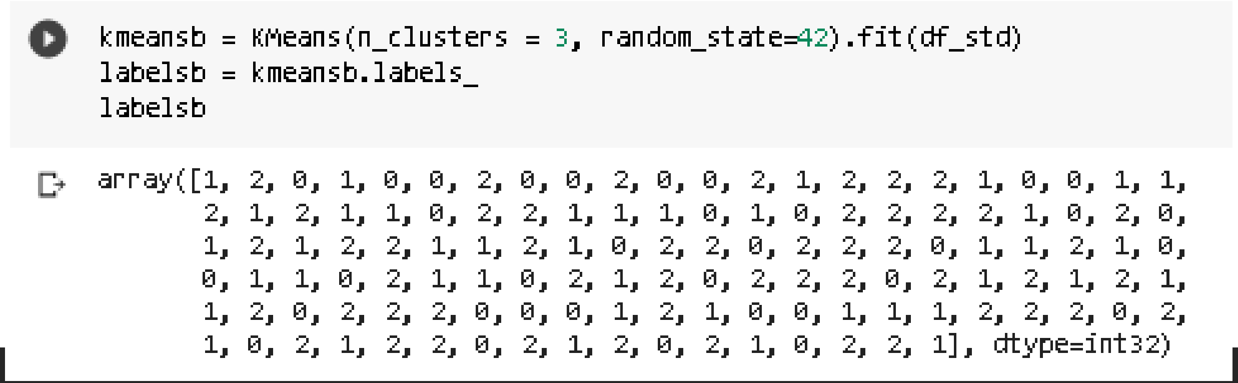
Data Clustering



```
#gunakan elbow method untuk menentukan jumlah cluster yang direkomendasik
wcss = []
for i in range(1, 11):
  kmeans = KMeans (n_clusters= i , init='k-means++', random_state = 42)
  kmeans.fit(new_df_std)
  wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



Use elbow method to determine the best number of cluster that we use



```
kmeansb = KMeans(n_clusters = 3, random_state=42).fit(df_std)
labelsb = kmeansb.labels_
labelsb
```

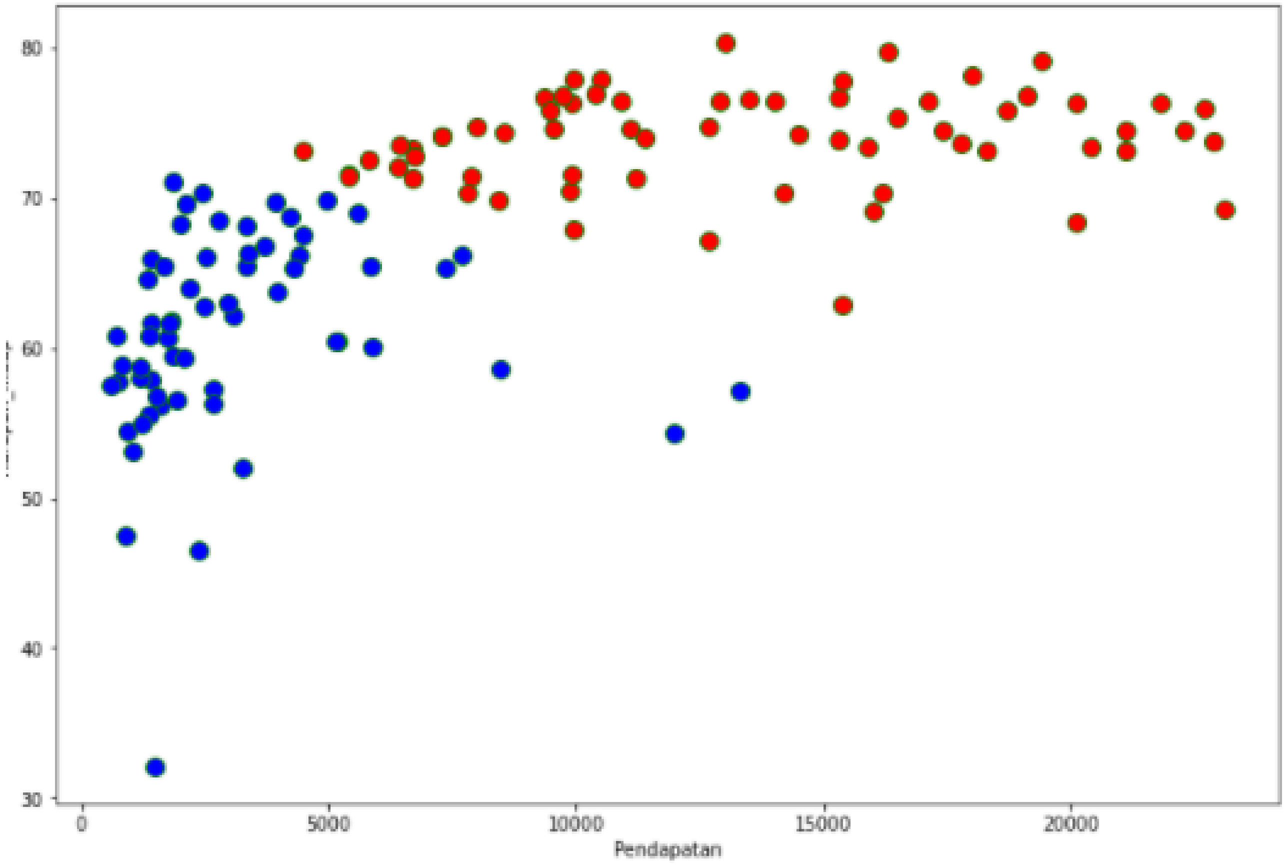
```
array([1, 2, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0, 2, 1, 2, 2, 2, 1, 0, 0, 1, 1,
       2, 1, 2, 1, 1, 0, 2, 2, 1, 1, 1, 0, 1, 0, 2, 2, 2, 2, 1, 0, 2, 0,
       1, 2, 1, 2, 2, 1, 1, 2, 1, 0, 2, 2, 0, 2, 2, 2, 0, 1, 1, 2, 1, 0,
       0, 1, 1, 0, 2, 1, 1, 0, 2, 1, 2, 0, 2, 2, 2, 0, 2, 1, 2, 1, 2, 1,
       1, 2, 0, 2, 2, 2, 0, 0, 0, 1, 2, 1, 0, 0, 1, 1, 1, 2, 2, 2, 0, 2,
       1, 0, 2, 1, 2, 2, 0, 2, 1, 2, 0, 2, 1, 0, 2, 2, 1], dtype=int32)
```

Clustering with Kmeans and inverse it to get a real value from each feature. Use 3 cluster as recommendation from elbow method

Results & Analysis

Results & Analysis

2 number cluster for Kmeans



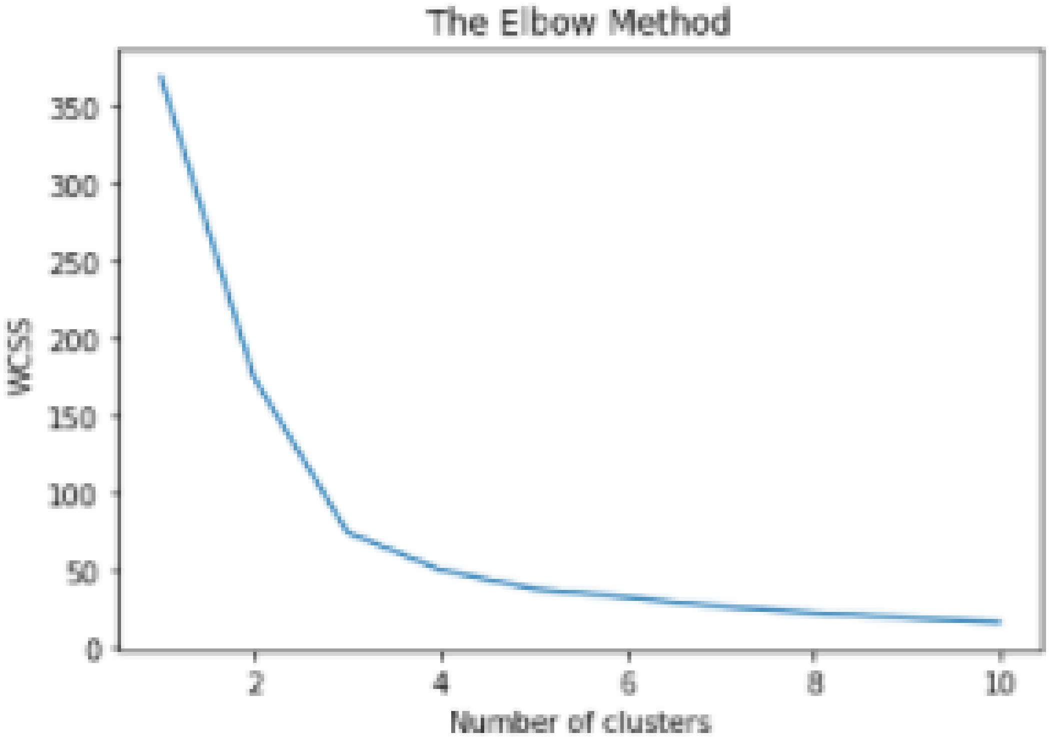
There are two classification country from this result:

-high value for pendapatan and Harapan_hidup variable

-low value for pendapatan and Harapan_hidup variable

Results & Analysis

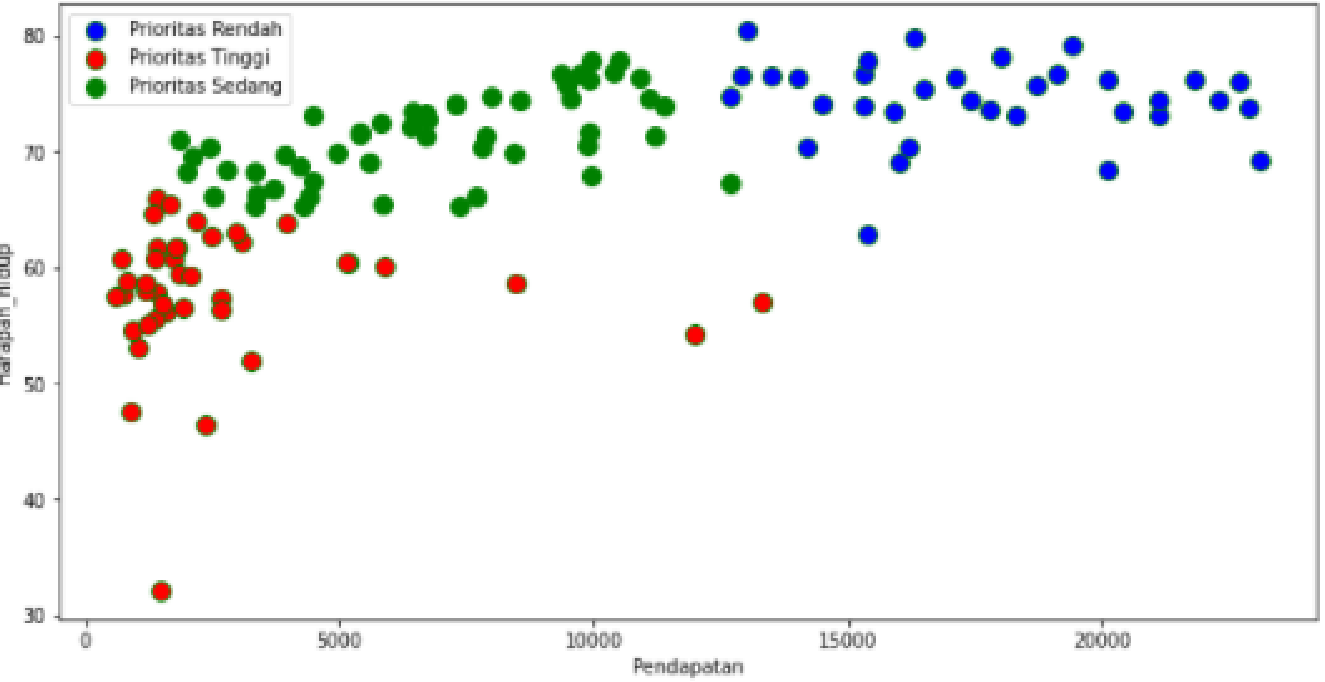
Elbow Method



From the elbow method, the optimum number cluster for these variable is 3 cluster. So we need to recluster to get a better clustering result for these variabel

Results & Analysis

3 number cluster for Kmeans

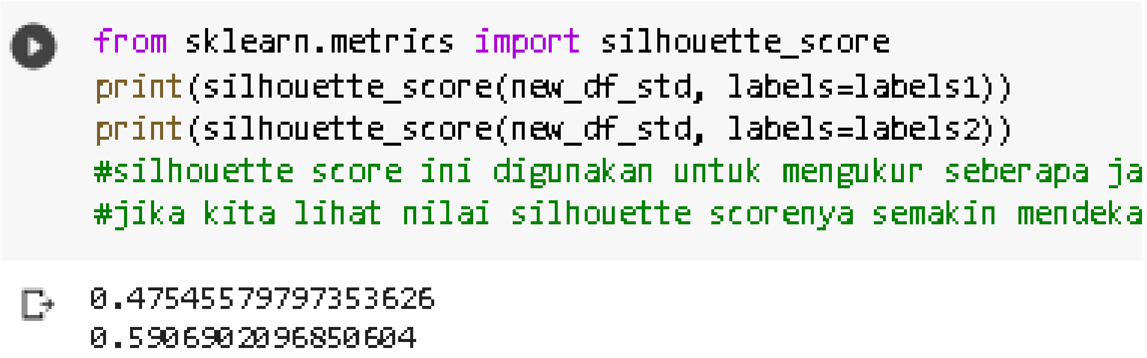


From the graph, there are three classification country from this result:

- **high value for pendapatan and Harapan_hidup variable(Prioritas Rendah)**
- **Medium value for pendapatan and Harapan_hidup variable(Prioritas Rendah)**
- **Low value for pendapatan and Harapan_hidup variable(Prioritas Tinggi)**

Results & Analysis

Silhouette Score



```
from sklearn.metrics import silhouette_score
print(silhouette_score(new_df_std, labels=labels1))
print(silhouette_score(new_df_std, labels=labels2))
#silhouette score ini digunakan untuk mengukur seberapa ja
#jika kita lihat nilai silhouette scorenya semakin mendeka
```

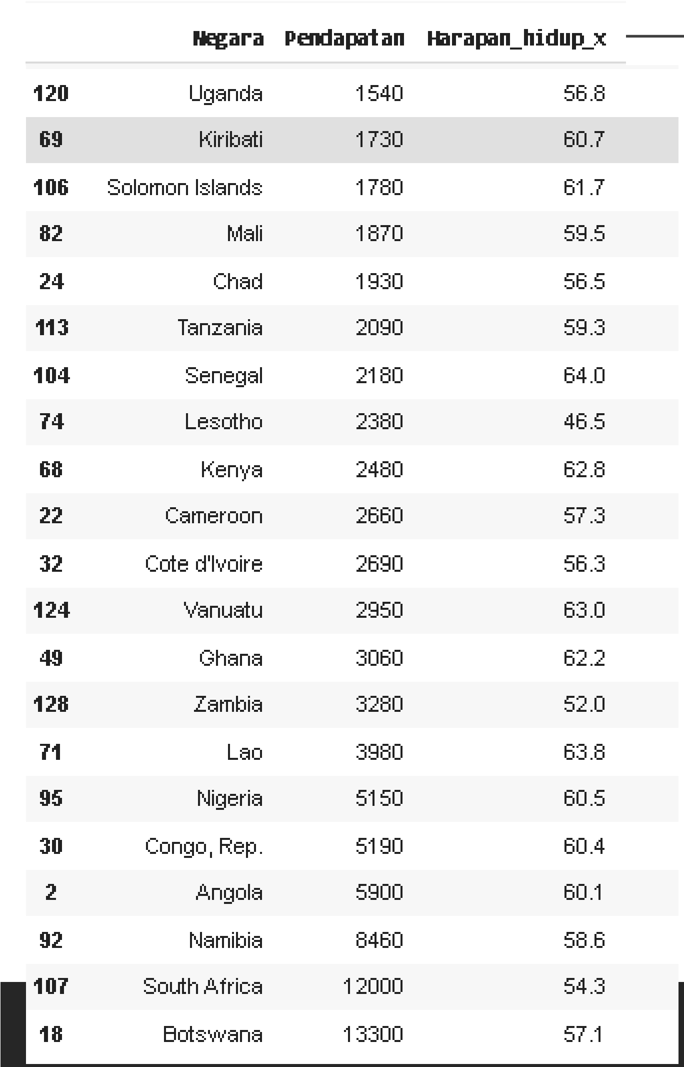
```
0.47545579797353626
0.5906902096850604
```

From the silhouette score, we can see that:

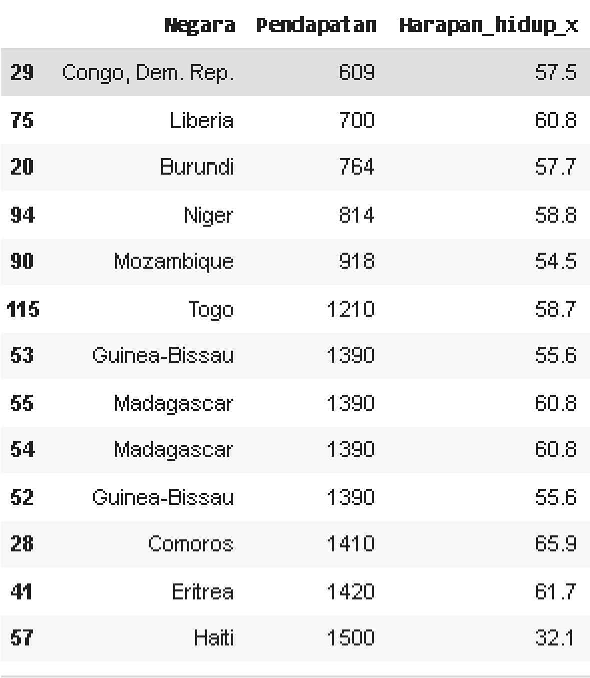
3 clustering is better than 2 clustering for this variable. **It indicated from the silhoute score that getting closer to the value of 1 that means the clusters are well separated**

Results & Analysis

Silhouette Score



	Negara	Pendapatan	Harapan_hidup_x
120	Uganda	1540	56.8
69	Kiribati	1730	60.7
106	Solomon Islands	1780	61.7
82	Mali	1870	59.5
24	Chad	1930	56.5
113	Tanzania	2090	59.3
104	Senegal	2180	64.0
74	Lesotho	2380	46.5
68	Kenya	2480	62.8
22	Cameroon	2660	57.3
32	Cote d'Ivoire	2690	56.3
124	Vanuatu	2950	63.0
49	Ghana	3060	62.2
128	Zambia	3280	52.0
71	Lao	3980	63.8
95	Nigeria	5150	60.5
30	Congo, Rep.	5190	60.4
2	Angola	5900	60.1
92	Namibia	8460	58.6
107	South Africa	12000	54.3
18	Botswana	13300	57.1



	Negara	Pendapatan	Harapan_hidup_x
29	Congo, Dem. Rep.	609	57.5
75	Liberia	700	60.8
20	Burundi	764	57.7
94	Niger	814	58.8
90	Mozambique	918	54.5
115	Togo	1210	58.7
53	Guinea-Bissau	1390	55.6
55	Madagascar	1390	60.8
54	Madagascar	1390	60.8
52	Guinea-Bissau	1390	55.6
28	Comoros	1410	65.9
41	Eritrea	1420	61.7
57	Haiti	1500	32.1

There are 34 countries that classified as countries with high priority to help **because their ‘Pendapatan’ and ‘Harapan_hidup’ value is the lowest than the other cluster.**

Conclusion

Conclusion

✓From the socio-economic(‘pendapatan’) & health (‘Harapan_hidup’) aspects, there are three classification country:

- high value for pendapatan and Harapan_hidup variable(Prioritas Rendah)
- Medium value for pendapatan and Harapan_hidup variable(Prioritas Rendah)
- Low value for pendapatan and Harapan_hidup variable(Prioritas Tinggi)

✓Countries that need to be focused on for assistance are countries that are in the high priority cluster, the cluster with the countries that have the lowest 'income' and 'life_expectations' values than the other 2 clusters.

“Thank You”