RAMA AULIA GEMILANG

FINAL PROJECT

BOOTCAMP PYTHON DATA SCIENCE BATCH 36

BOOTCAMP SANBERCODE

1

BUSINESS/PROJECT UNDERSTANDING

Latar belakang proyek

Menjelaskan latar belakang dan kebutuhan proyek.

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. Oleh karena itu, pada final project kali ini akan mencari atau mengkategorikan negara-negara yang menjadi fokus CEO.

2

THE DATA

Memrosesan Data

1. Pemahaman Dataset (Dataset Understanding)
2. EDA Part 1
3. Pemilihan Fitur (Feature Selection)
4. Pembersihan Data (Data Cleaning)
5. EDA Part 2

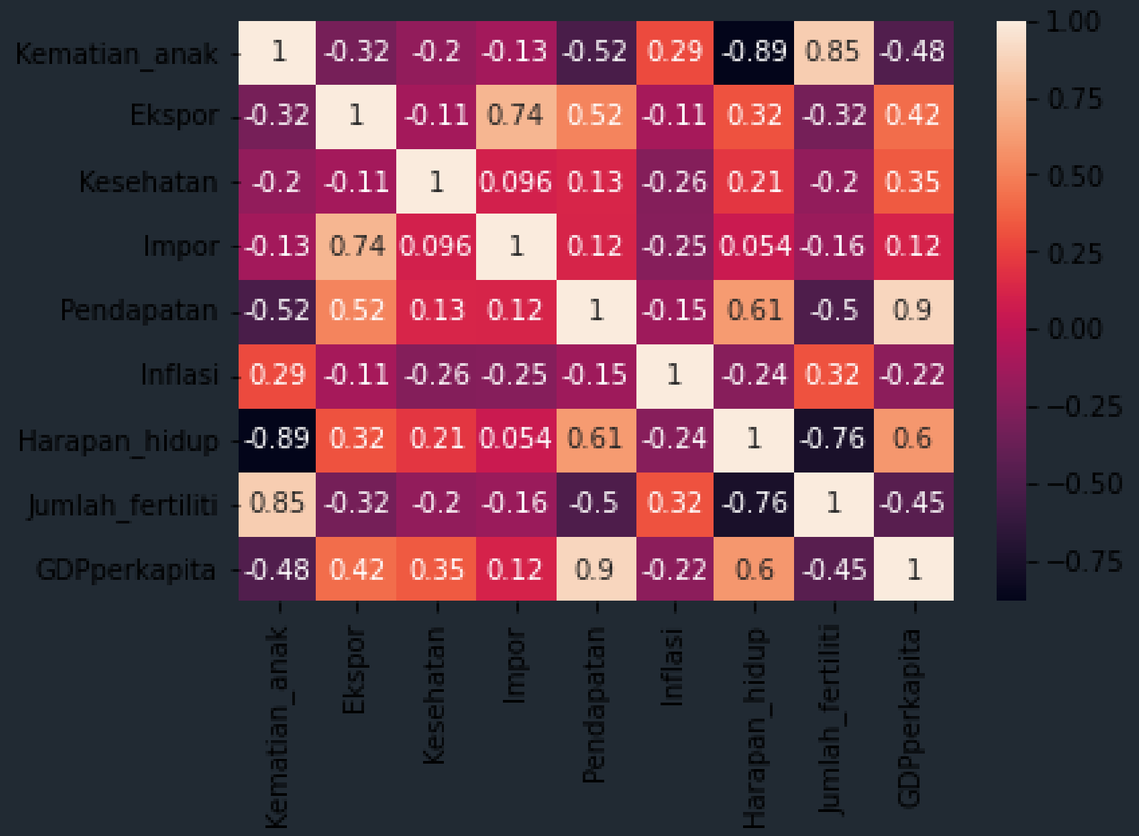
DATASET

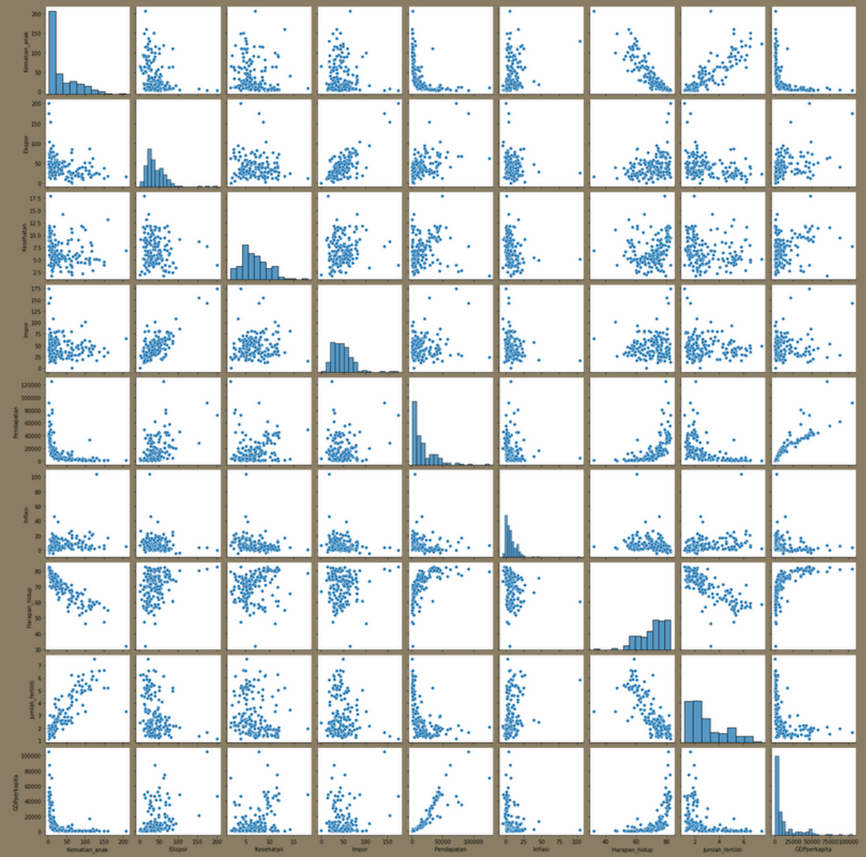
Dataset yang digunakan merupakan dengan kolom sebagai berikut :

- **Negara :** Nama negara (Dengan jumlah negara berkisar 167 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.

**167 rows x 10 columns*

MULTIVARIATE ANALYSIS (HEAT MAP & PAIR PLOT)





PEMILIHAN VARIABEL

Dari Heat Map dan Pair Plot sebelumnya maka akan digunakan hubungan antara GDP perkapita dengan Pendapatan karena korelasi hubungan tersebut (0.9) tergolong sangat kuat **(*High Positive Correlation*)** dibanding hubungan variabel yang lainnya. Hubungan antar kedua variabel tersebut dapat diartikan dengan semakin tinggi Pendapatan maka semakin tinggi GDP perkapita. Dalam meneyelesaikan final project ini diminta untuk mendapatkan negara yang membutuhkan bantuan, maka penyelesaiannya adalah dengan mendapatkan negara dengan Pendapatan terendah dengan GDP perkapita terendah.

```
df.info()
```

[8] ✓ 0.1s

```
<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
```

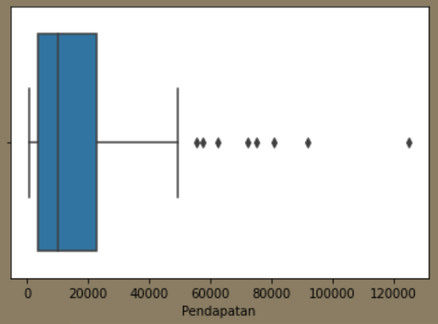
```
df.isnull().sum()
```

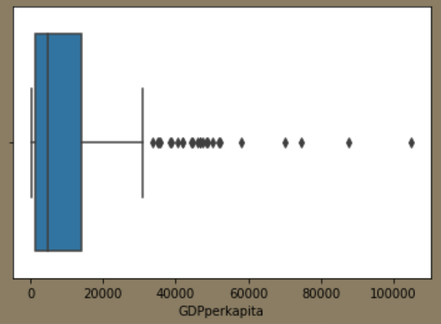
[9] ✓ 0.5s

```
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
```

MISSING VALUE

Dengan menggunakan atribut info() dan atribut isnull() dapat ditarik kesimpulan bahwa semua kolom tidak mempunyai missing value. Hal ini disebabkan karena deteksi nilai hilang (missing value) dengan menggunakan atribut isnull() memperoleh hasil 0 untuk semua kolom.






```
Output exceeds the size limit. Open the full output data in a text editor
23      80600
82      75200
91      91700
114     62300
123    125000
133     72100
145     55500
157     57600
Name: Pendapatan, dtype: int64
7       51900
8       46900
15      44400
23      35300
29      47400
44      58000
53      46200
54      40600
58      41800
68      41900
73      48700
75      35800
77      44500
82      38500
91     105000
110     50300
...
157     35000
158     38900
159     48400
Name: GDPperkapita, dtype: int64
```

CEK OUTLIERS

Dengan menggunakan uji boxplot dan uji lower bound & upper bound untuk variabel Pendapatan dan GDP perkapita ditemukan outliers. maka dari itu untuk menyelesaikan langkah selanjutnya terlebih dahulu menghapus outliers yang ada agar tidak menyebabkan distorsi terhadap nilai asli

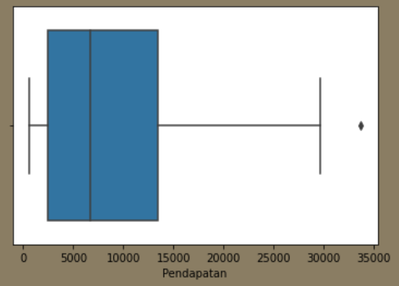
HAPUS OUTLIERS

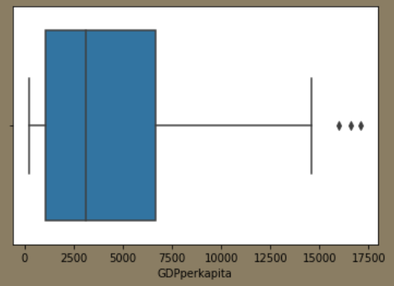
	Pendapatan	GDPperkapita
0	1610.0	553.0
1	9930.0	4090.0
2	12900.0	4460.0
3	5900.0	3530.0
4	19100.0	12200.0
...	...	...
162	2950.0	2970.0
163	16500.0	13500.0
164	4490.0	1310.0
165	4480.0	1310.0
166	3280.0	1460.0

142 rows × 2 columns

	Pendapatan	GDPperkapita
0	1610.0	553.0
1	9930.0	4090.0
2	12900.0	4460.0
3	5900.0	3530.0
4	19100.0	12200.0
...	...	...
162	2950.0	2970.0
163	16500.0	13500.0
164	4490.0	1310.0
165	4480.0	1310.0
166	3280.0	1460.0

129 rows × 2 columns

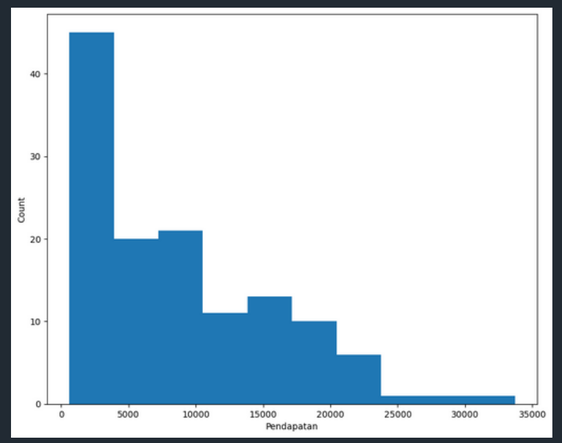




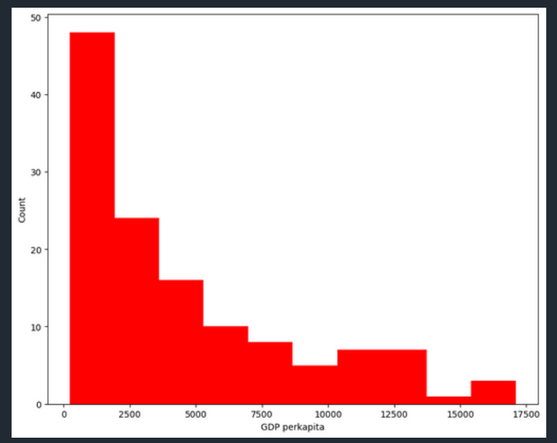
Dengan mengganti nilai outliers menggunakan nilai lower bound dan upper bound maka outliers yang ada pada kedua variabel dapat terhapus. Pada langkah ini dilakukan 2 kali penghapusan outliers karena pada penghapusan outliers pertama (gambar paling kiri atas) masih terdapat banyak outliers. Hasil penghapusan outliers untuk kedua kalinya terdapat pada gambar paling kanan atas dan gambar paling bawah (boxplot). meskipun masih ada beberapa outliers tetapi data tersebut tetap digunakan karena sudah tidak terlalu banyak outliers yang ada sehingga tidak terlalu mengganggu keakuratan proses selanjutnya.

UNIVARIATE ANALYSIS & BIVARIATE ANALYSIS

HISTOGRAM

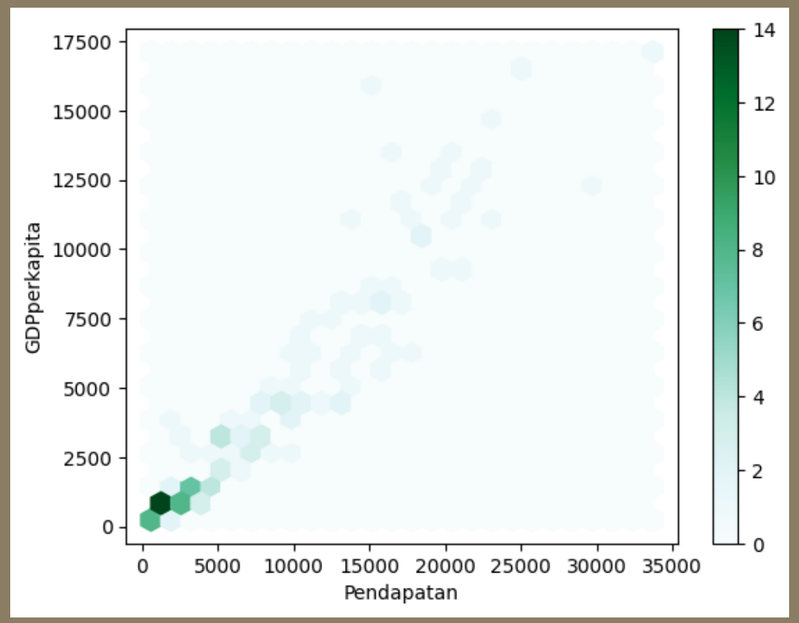


Pendapatan



GDP perkapita

HEXBIN PLOT



ANALISIS

UNIVARIATE ANALYSIS

Dari Histogram pada slide sebelumnya dapat dilihat bahwa pendapatan yang dimiliki suatu negara kebanyakan berkisar 1-4800 dan GDP perkapita yang dimiliki suatu negara kebanyakan berkisar 1-2300.

BIVARIATE ANALYSIS

Dari Hexbin Plot pada slide sebelumnya semakin memperkuat univariate analysis karena terlihat jelas bahwa Negara paling banyak berada dengan nilai pendapatan yang berkisar 0-5000 dan nilai GDP perkapita yang berkisar 0-2500.

3

CLUSTERING

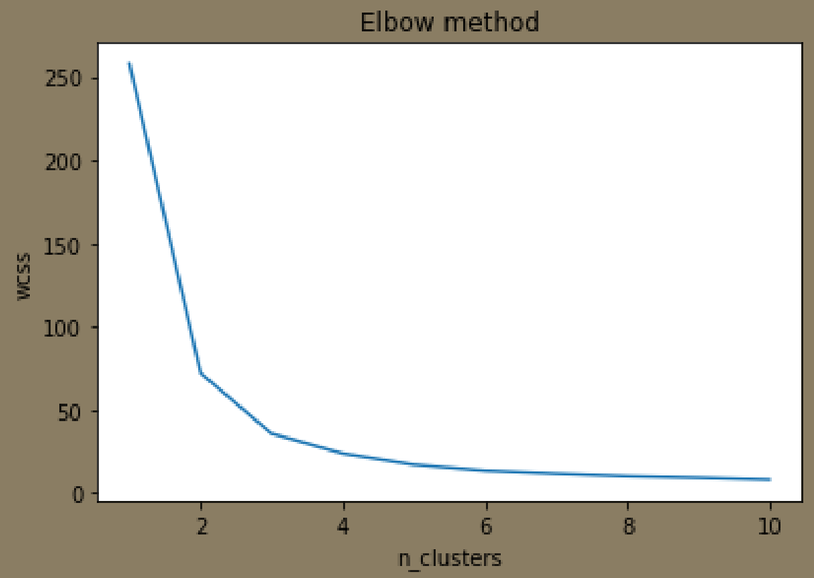
Clustering Data

1. Skala Data (Scale the Data)
2. Jumlah Cluster (Number of Cluster)
3. Clustering dengan jumlah cluster (Clustering)
4. Grafik Cluster (Cluster Graph)

```
Output exceeds the size limit. Open the full output data in a text editor
array([[-1.00340971e+00, -9.44550980e-01],
       [ 1.63532624e-01, -1.05881202e-01],
       [ 5.80097375e-01, -1.81492777e-02],
       [-4.01705067e-01, -2.38664656e-01],
       [ 1.44969382e+00,  1.81710774e+00],
       [ 1.39359083e+00,  1.36659245e+00],
       [-2.89499074e-01, -3.12169782e-01],
       [ 1.01489560e+00,  3.09067090e-01],
       [-8.86995988e-01, -8.95942751e-01],
       [ 9.16715355e-01,  2.71813832e+00],
       [ 1.04294710e+00,  3.54118619e-01],
       [-1.23995233e-01, -4.66028749e-02],
       [-9.73955633e-01, -8.95942751e-01],
       [-3.28771171e-01, -5.58767625e-01],
       [-4.70431238e-01, -6.06190287e-01],
       [ 1.34078551e-01,  1.74177188e-02],
       [ 6.36200371e-01,  4.29994878e-01],
       [ 8.04509361e-01,  1.57999443e+00],
       [ 9.16715355e-01,  5.46180400e-01],
       [-1.02865605e+00, -9.39334487e-01],
       [-1.12206754e+00, -1.02090147e+00],
       [-8.75775389e-01, -8.89303579e-01],
       [-8.56139340e-01, -7.65056204e-01],
       [-4.11523091e-01, -2.90829584e-01],
       [-1.10467562e+00, -9.69922104e-01],
...
       [-8.15464667e-01, -3.71448110e-01],
       [ 1.08502434e+00,  2.12535504e+00],
       [-5.99468130e-01, -7.65056204e-01],
       [-6.00870705e-01, -7.65056204e-01],
       [-7.69179695e-01, -7.29489208e-01]])
```

SCALING DATA

Dalam tahap ini dilakukan scaling data dengan tujuan untuk membuat numerical data pada dataset memiliki rentang nilai (scale) yang sama. Tidak ada lagi satu variabel data yang mendominasi variabel data lainnya. Scaling data pada tahap ini menggunakan StandarScaler yang bertujuan untuk membuat rata-rata 0 dan variansi 1.




```
print(silhouette_score(df_std, labels= labels1)) # n cluster = 2
print(silhouette_score(df_std, labels= labels2)) # n cluster = 3
```

```
0.6379518757336385
0.5890258762560227
```

JUMLAH CLUSTER

Dalam tahap menentukan jumlah cluster yang cocok digunakan 2 metode yaitu Elbow Method dan Silhouette Score. Dari metode tersebut dihasilkan hasil seperti gambar disamping. Meskipun silhouette score dari n cluster = 2 lebih besar daripada silhouette score n cluster = 3 tetapi scatter plot yang akan digunakan untuk proses selanjutnya adalah scatter plot dengan n cluster = 3 karena mengikuti grafik dari elbow method serta scatter plot n cluster = 3 lebih mudah untuk dianalisis.

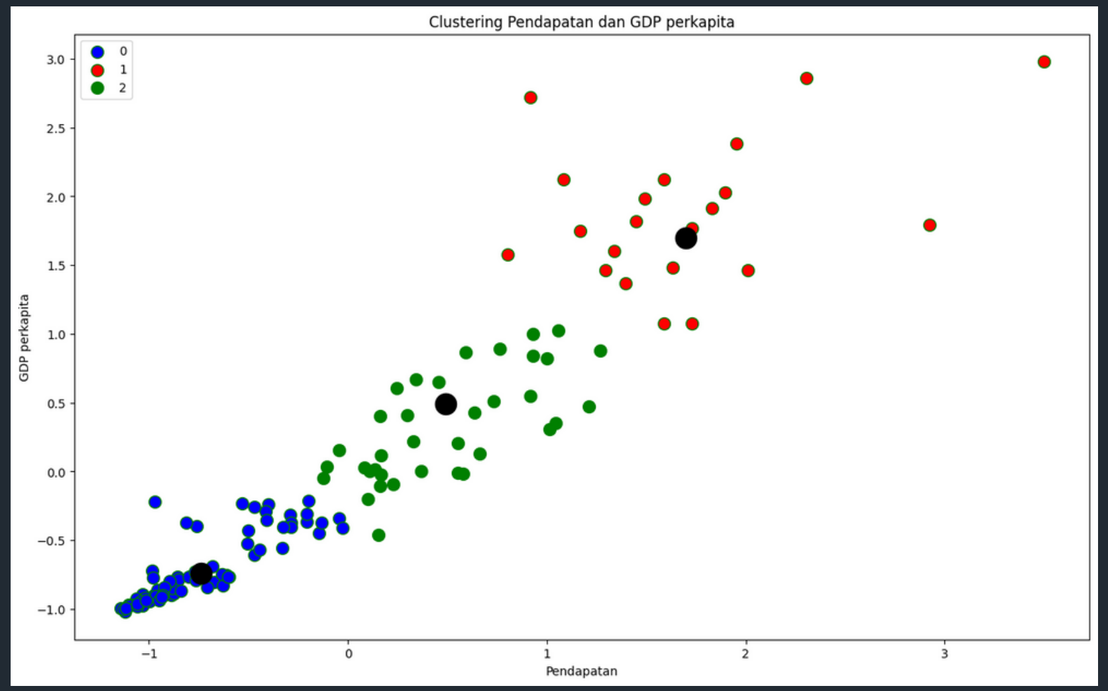
CLUSTERING DATA

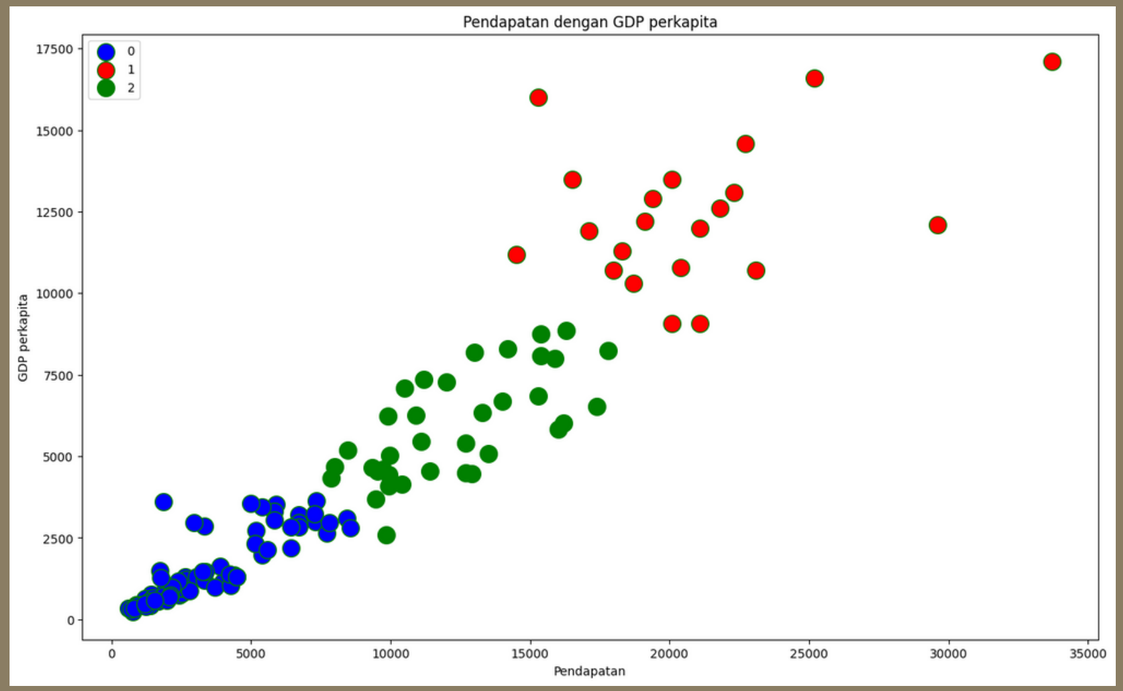
Dalam tahap ini dilakukan clustering data dengan n cluster = 3. Clustering data pada proses ini menggunakan Kmeans dengan tujuan untuk meminimalisasikan fungsi objective yang telah di set dalam proses clustering. Tujuan tersebut dilakukan dengan cara meminimalikan variasi data yang ada didalam cluster dan memaksimalikan variasi data yang ada di cluster lainnya.

```
array([0, 2, 2, 0, 1, 1, 0, 2, 0, 1, 2, 2, 0, 0, 0, 2, 2, 1, 2, 0, 0, 0,
       0, 0, 0, 0, 1, 2, 2, 0, 0, 0, 2, 0, 1, 2, 2, 2, 0, 1, 0, 1, 0, 2,
       0, 0, 0, 2, 0, 0, 0, 0, 0, 1, 0, 0, 2, 2, 2, 2, 1, 0, 0, 0, 0, 1,
       2, 0, 0, 1, 1, 2, 0, 0, 1, 2, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 2, 0,
       0, 0, 0, 2, 0, 2, 0, 1, 2, 1, 0, 0, 0, 2, 1, 0, 1, 0, 2, 0, 2, 0,
       2, 0, 0, 2, 0, 0, 0, 2, 1, 2, 0, 0, 1, 0, 0, 1, 0, 0, 0])
```

Label clustering dengan n cluster = 3

GRAFIK CLUSTERING





CLUSTER GRAPH

ANALISIS CLUSTERING

Dari grafik pada slide sebelumnya akan difokuskan pada label 0 karena pada label tersebut nilai Pendapatan dan GDP perkapita berada pada posisi paling rendah dibanding label yang lainnya. Sebelum mengambil data pada label 0 terlebih dahulu menggembalikan data pada bentuk awal dengan scaling data menggunakan metode Invers Transform. Metode ini berfungsi untuk mengembalikan skala data ke bentuk representasi asli. hasil dari clustering menggunakan n cluster = 3 dan scaling Invers Transform dapat dilihat pada slide sebelumnya grafik bagian kanan (grafik bagian kiri sama dengan bagian kanan, hanya saja bagian kiri menggunakan scaling StandarScaler).

TAMPILAN DATA HASIL CLUSTERING

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	label
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	0
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090	2
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460	2
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	1
...	...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970	NaN
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500	NaN
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310	NaN
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310	NaN
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460	NaN

167 rows × 11 columns

CLUSTERING GRAPH

4

RECOMMENDATION

Rekomendasi negara

Menampilkan negara yang cocok menerima bantuan dalam cluster yang telah dibuat

HASIL NEGARA YANG MENJADI FOKUS CEO

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita	label
88	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	0
112	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	0
31	Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446	0
94	Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459	0
50	Eritrea	55.2	4.79	2.66	23.3	1420	11.60	61.7	4.61	482	0
64	Guinea-Bissau	114.0	14.90	8.50	35.2	1390	2.97	55.6	5.05	547	0
0	Afghanistan	90.2	10.00	7.58	44.9	1610	9.44	56.2	5.82	553	0
126	Rwanda	63.6	12.00	10.50	30.0	1350	2.61	64.6	4.51	563	0
25	Burkina Faso	116.0	19.20	6.74	29.6	1430	6.81	57.9	5.87	575	0
109	Nepal	47.0	9.58	5.25	36.4	1990	15.10	68.3	2.61	592	0

Dari data yang telah dilakukan clustering maka untuk mendapatkan negara yang akan menjadi fokus ceo (menerima bantuan), maka akan diambil negara dengan label 0 (label berwarna biru) karena Dari data yang telah diberi label diatas akan diambil data dengan label 0 (label berwarna biru) karena pada label tersebut nilai Pendapatan dan nilai GDP perkapita berada pada posisi terendah dibanding label yang lainnya. Karena HELP International hanya memiliki budget $ 10 juta maka data diambil 10 data terkecil (negara paling miskin).

Sehingga negara yang menjadi fokus ceo demi mendapatkan bantuan adalah **Liberia, Niger, Central African Republic, Malawi, Eritra, Guinea-Bissau, Afghanistan, Rwanda, Burkina Faso, dan Nepal.**

RECOMMENDATION

TERIMA KASIH

FINAL PROJECT

SANBERCODE BOOTCAMP PYTHON DATA SCIENCE BATCH 36