Topic

The assignment targets to implement K-Means and K-Medoid algorithms to cluster the dataset consists of socio-economic and health factors of countries and determine the overall development of the country Implementation

by

Kiran Kumar Dugana

1. Importation of all required Libraries:

I have important all libraries required as below

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import metrics
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import scipy.cluster.hierarchy as hcluster
from sklearn.cluster import AgglomerativeClustering
from sklearn.decomposition import PCA
from sklearn.preprocessing import MinMaxScaler
from sklearn_extra.cluster import KMedoids
```

2. Creation of data frame from given data:

I have used read_csv function to read data from given csv file

Top five rows of data

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
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

3. Data cleaning:

I checked null and NAN values presence

```
df_given_data.isna().sum().sum()
```

```
No of NAN values in given data
0
```

```
df_given_data.isnull().sum().sum()
```

```
No of NULL values in given data
0
```

I also checked if all names in country column are unique

```
print(df_given_data.country.nunique())
print(len(df_given_data.index))
```

```
Printing no of unique names in country coloumn and no of rows in data frame
If the both the counts are same then there no duplicates name of any country present in given data
167
167
```

I dropped the country column from given data to apply scaling on data

Top five rows of data with country column

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
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

Top five rows of data without country column

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

4. Scaling:

I scaled data to get data of all column in a same range

Top 4 rows of data before scaling

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Top 4 rows of data after scaling

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.426485	0.049482	0.358608	0.257765	0.008047	0.126144	0.475345	0.736593	0.003073
1	0.068160	0.139531	0.294593	0.279037	0.074933	0.080399	0.871795	0.078864	0.036833
2	0.120253	0.191559	0.146675	0.180149	0.098809	0.187691	0.875740	0.274448	0.040365
3	0.566699	0.311125	0.064636	0.246266	0.042535	0.245911	0.552268	0.790221	0.031488
4	0.037488	0.227079	0.262275	0.338255	0.148652	0.052213	0.881657	0.154574	0.114242

5. Dimension Reduction:

I reduced 9 dimensional data into 2 dimensional data using pca

Top 4 rows of data before pca

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.426485	0.049482	0.358608	0.257765	0.008047	0.126144	0.475345	0.736593	0.003073
1	0.068160	0.139531	0.294593	0.279037	0.074933	0.080399	0.871795	0.078864	0.036833
2	0.120253	0.191559	0.146675	0.180149	0.098809	0.187691	0.875740	0.274448	0.040365
3	0.566699	0.311125	0.064636	0.246266	0.042535	0.245911	0.552268	0.790221	0.031488
4	0.037488	0.227079	0.262275	0.338255	0.148652	0.052213	0.881657	0.154574	0.114242

Top 4 rows of data after pca

	x	y
0	-0.599078	0.095490
1	0.158474	-0.212092
2	0.003686	-0.135867
3	-0.650235	0.275975
4	0.200711	-0.064662

6. KMEANS CLUSTEERING:

I applied KMEANS scaled and dimensional reduced version of data. I assumed k value as 3

```
START OF KMEANS CLUSTERRING
Using kmeans model with k value 3
Printing the no of values in each cluster
2    83
1    46
0    38
dtype: int64
```

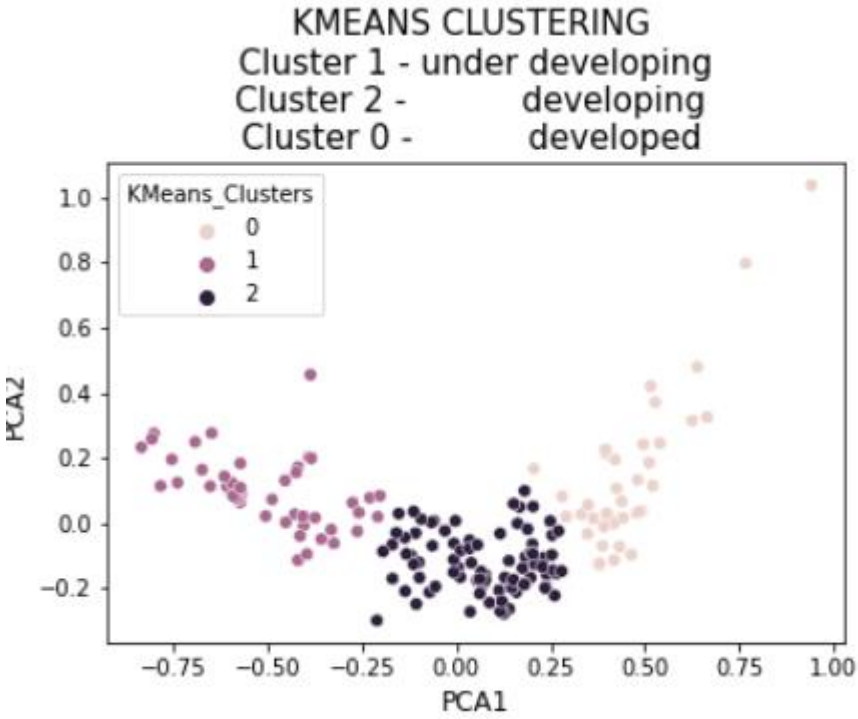
Top 4 rows of pca table with cluster labels added in each row

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	KMeans_Clusters
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	1
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	1
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	2

Top 4 rows of given table with cluster labels added in each row

	x	y	KMeans_Clusters
0	-0.599078	0.095490	1
1	0.158474	-0.212092	2
2	0.003686	-0.135867	2
3	-0.650235	0.275975	1
4	0.200711	-0.064662	2

7. Visualization:



8. Silhouette Coefficient Calculation:

I calculated to know how strong my clustering is

```
Printing shilhoutte scre
0.8096802162196737
```

9. Print countries cluster wise:

```
Printing list of under developing countries

['Afghanistan', 'Angola', 'Benin', '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', 'Solomon
Islands', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda',
'Yemen', 'Zambia']
```

```
Printing list of developed countries

['Australia', 'Austria', 'Bahrain', 'Belgium', 'Brunei',
'Canada', 'Cyprus', 'Czech Republic', 'Denmark', 'Estonia',
'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'Iceland',
'Ireland', 'Italy', 'Japan', 'Kuwait', 'Luxembourg', 'Malta',
'Netherlands', 'New Zealand', 'Norway', 'Portugal', 'Qatar',
'Seychelles', 'Singapore', 'Slovak Republic', 'Slovenia', 'South
Korea', 'Spain', 'Sweden', 'Switzerland', 'United Arab Emirates',
'United Kingdom', 'United States']


Printing list of developing countries

['Albania', 'Algeria', 'Antigua and Barbuda', 'Argentina',
'Armenia', 'Azerbaijan', 'Bahamas', 'Bangladesh', 'Barbados',
'Belarus', 'Belize', 'Bhutan', 'Bolivia', 'Bosnia and
Herzegovina', 'Botswana', 'Brazil', 'Bulgaria', 'Cambodia', 'Cape
Verde', 'Chile', 'China', 'Colombia', 'Costa Rica', 'Croatia',
'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Fiji',
'Georgia', 'Grenada', 'Guatemala', 'Guyana', 'India',
'Indonesia', 'Iran', 'Israel', 'Jamaica', 'Jordan', 'Kazakhstan',
'Kyrgyz Republic', 'Latvia', 'Lebanon', 'Libya', 'Lithuania',
'Macedonia, FYR', 'Malaysia', 'Maldives', 'Mauritius',
'Micronesia, Fed. Sts.', 'Moldova', 'Mongolia', 'Montenegro',
'Morocco', 'Myanmar', 'Nepal', 'Oman', 'Panama', 'Paraguay',
'Peru', 'Philippines', 'Poland', 'Romania', 'Russia', 'Samoa',
'Saudi Arabia', 'Serbia', 'South Africa', 'Sri Lanka', 'St.
Vincent and the Grenadines', 'Suriname', 'Tajikistan',
'Thailand', 'Tonga', 'Tunisia', 'Turkey', 'Turkmenistan',
'Ukraine', 'Uruguay', 'Uzbekistan', 'Vanuatu', 'Venezuela',
'Vietnam']
```