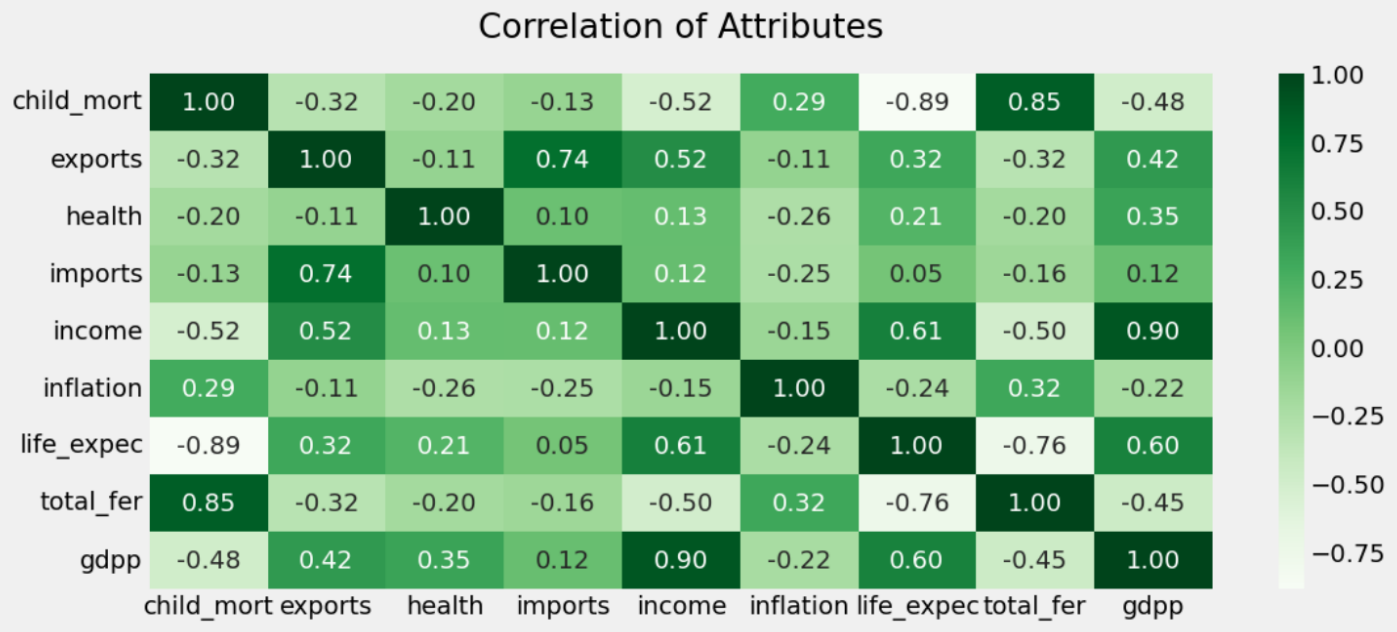
Elahe Ghods

Medical Genetics and Biotechnology

✓ Data analysis

After initial data analysis, including the range of changes, type of features, missing values, etc., it was determined that the data does not have any missing values.

- ❖ Investigating the correlation between features:

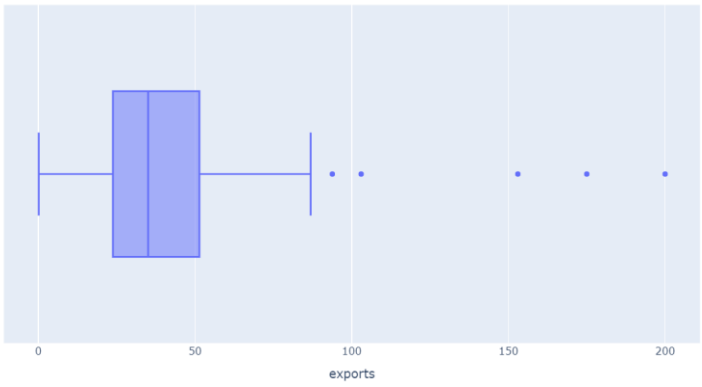


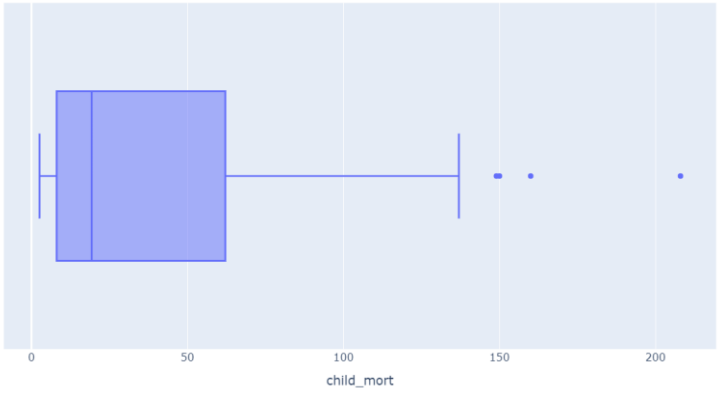
According to the above table, the most important correlations between features are as follows:

1. Child_mort has an inverse relationship with income (0.52) and life_expec (0.89), but a direct relationship with total_fer (0.85).
   In other words, the higher the child mortality rate under the age of 5, the higher the fertility rate, while income and life expectancy decrease.
2. Export has a direct relationship with import. (As exports increase, imports also increase.)
3. Income has a strong direct relationship with gdpp (0.90). (The higher the income, the higher the gross domestic product per capita (gdpp))
4. Life_expec has a direct relationship with gdpp and income, but an inverse relationship with total_fer. (Countries with higher income and, consequently, higher gross domestic product per capita have higher life expectancy, but the fertility rate decreases.)

❖ **Investigating outliers:**

Based on box plots, such as the one shown below, outliers were identified for each feature.





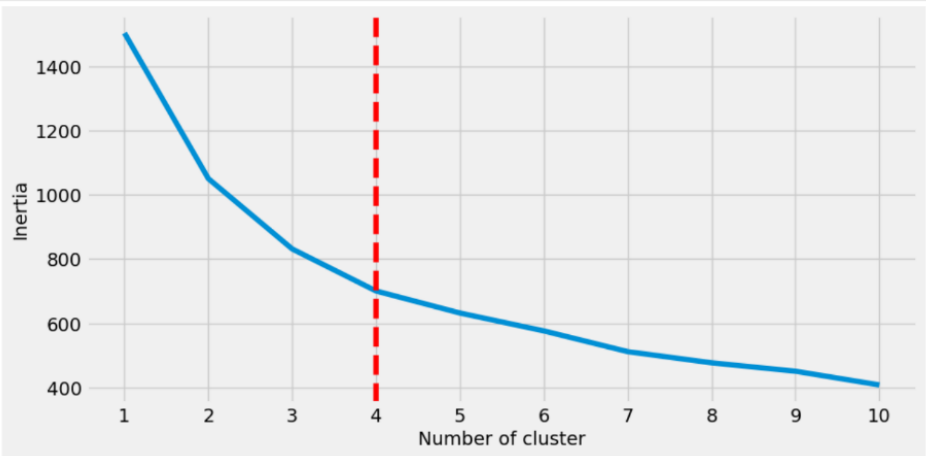
Make a dataframe from outliers:

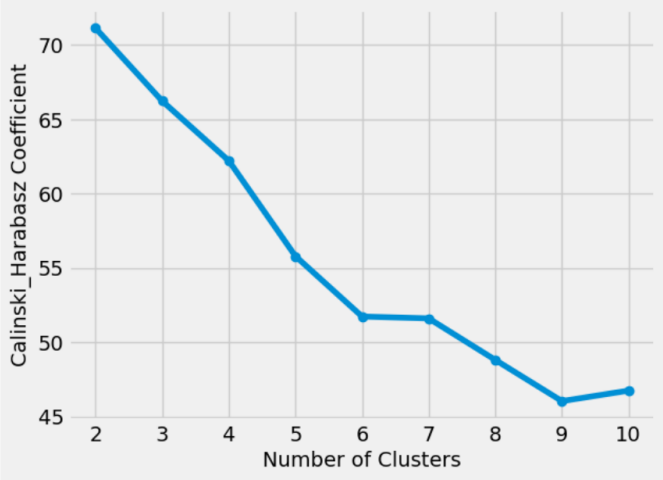
```
# these are outliers :

out1 = df[df['child_mort'] > 148]
out2 = df[df['exports'] > 90]
out3 = df[df['health'] > 14]
out4 = df[df['imports'] > 140]
out5 = df[df['income'] > 120000]
out6 = df[df['inflation'] > 30]
out7 = df[df['life_expec'] < 50]
out8 = df[df['total_fer'] > 7]
out9 = df[df['gdpp'] > 60000]


outliers = pd.concat([out1, out2, out3, out4, out5, out6, out7, out8, out9])     # make a dataframe from outliers
outliers
```

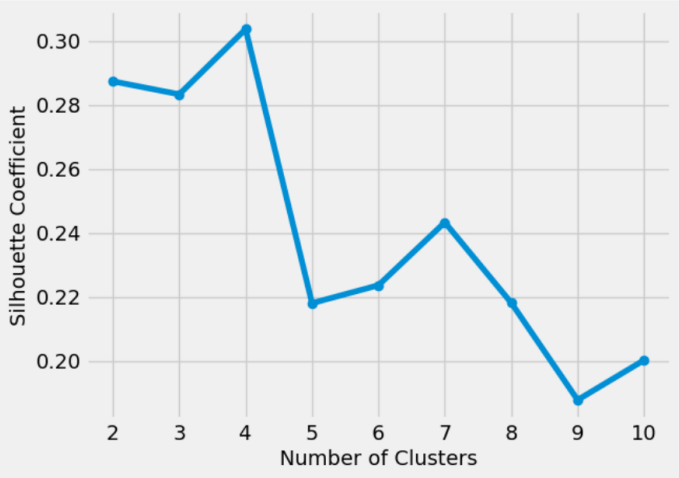
✓ Modeling:

First, the number of clusters were determined using k-means. (modeling was performed once with outliers and once without them.)

❖ **Modeling with outliers:**

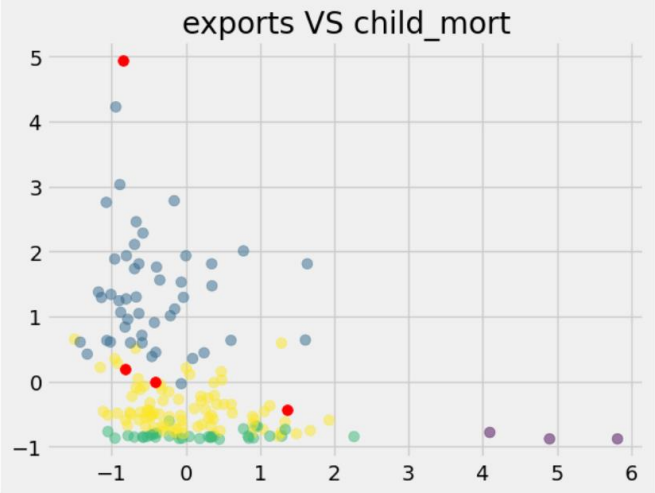


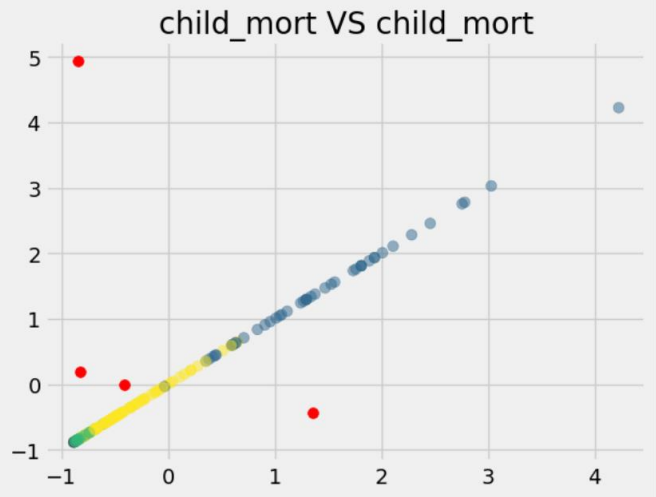


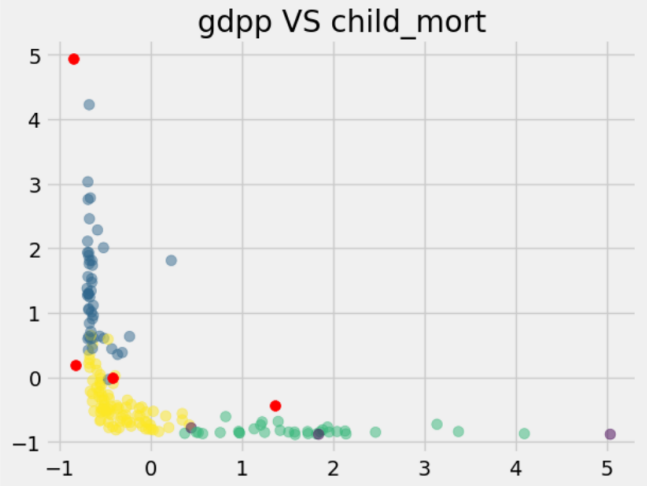


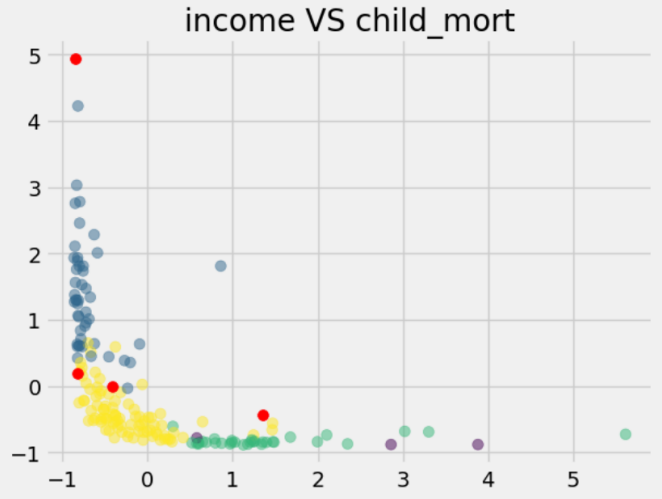
After modeling and evaluating using silhouette_score and calinski_harabasz_score parameters, the number of clusters was determined to be 4.

❖ **Visualization** :Below are some of the scatter plots after clustering.









As evident from the plots, due to the presence of outliers, clustering has not been performed well, and it seems that 4 clusters are not suitable for this dataset. This has led a centroid to be placed in an outlier, which has very little data and is clearly out of the range.

Therefore, it is better to separate the outliers from the original data so that we do not encounter any problems in the modeling process and the modeling can be done more accurately.

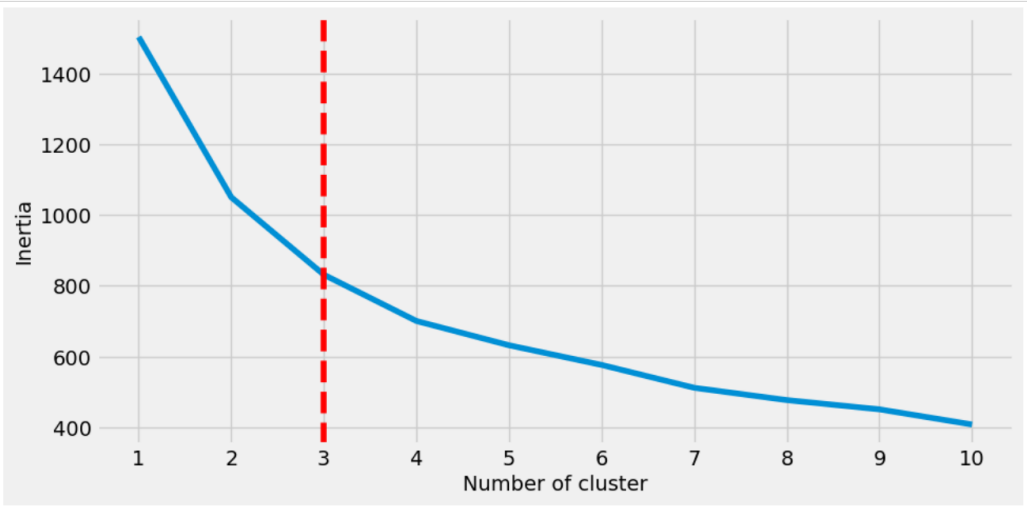
✓ **Modeling without outliers**:

```
df_without_out.drop(index=outliers.index, inplace=True)    # creat a dataframe without outliers
df_without_out.reset_index(inplace=True, drop=True)
df_without_out
```

	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
...	...	...	...	...	...	...	...	...	...
143	36.3	31.7	5.81	28.5	4240	16.50	68.8	2.34	1380
144	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
145	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
146	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
147	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

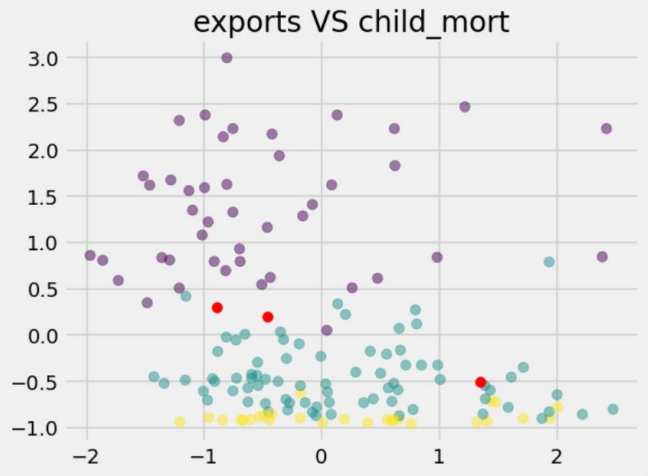
148 rows × 9 columns

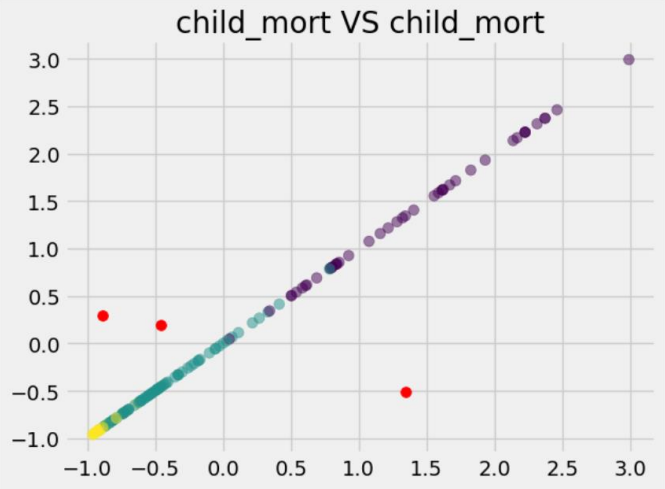
After removing the outliers from the original data, the modeling was performed again.

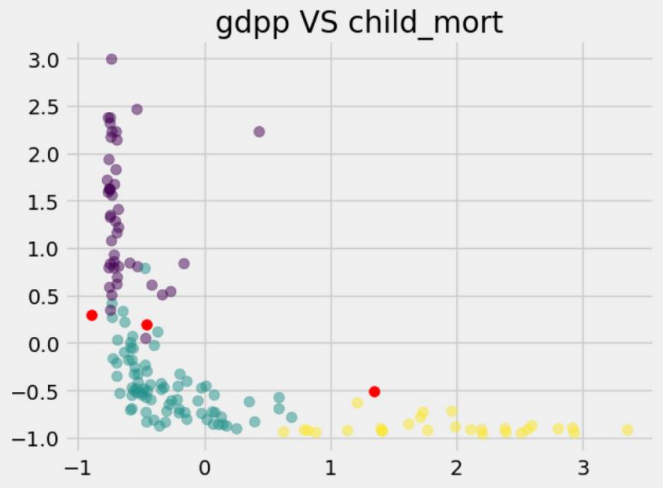


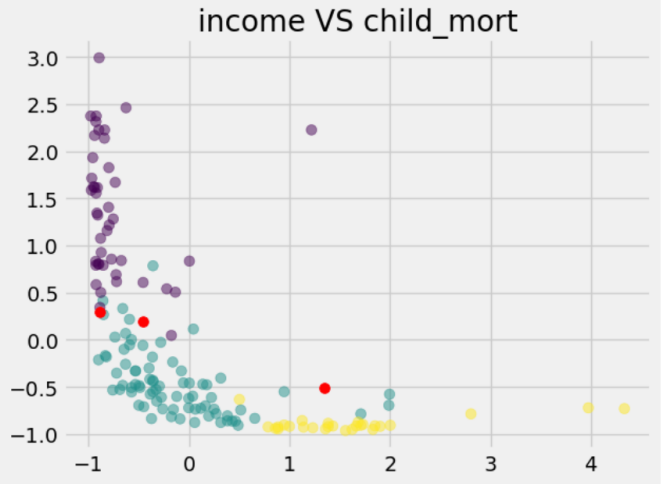
After modeling and evaluating using silhouette_score and calinski_harabasz_score parameters, the number of clusters was determined to be 3.

- **Visualization** :Below are some of the scatter plots after clustering.

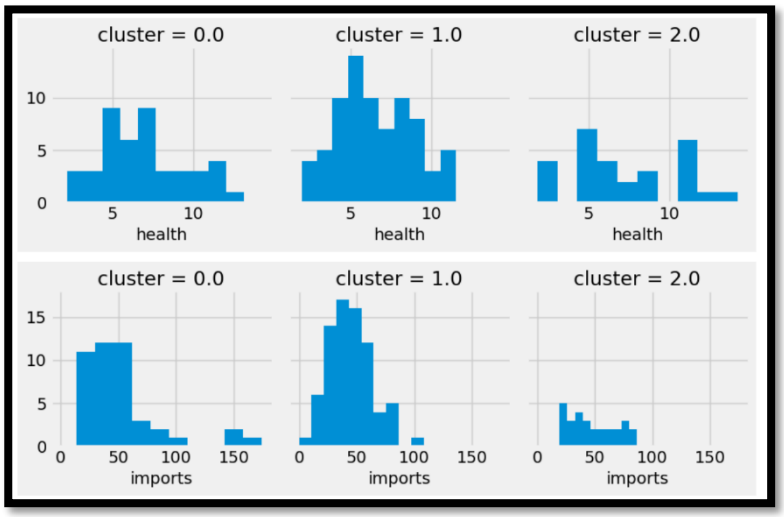


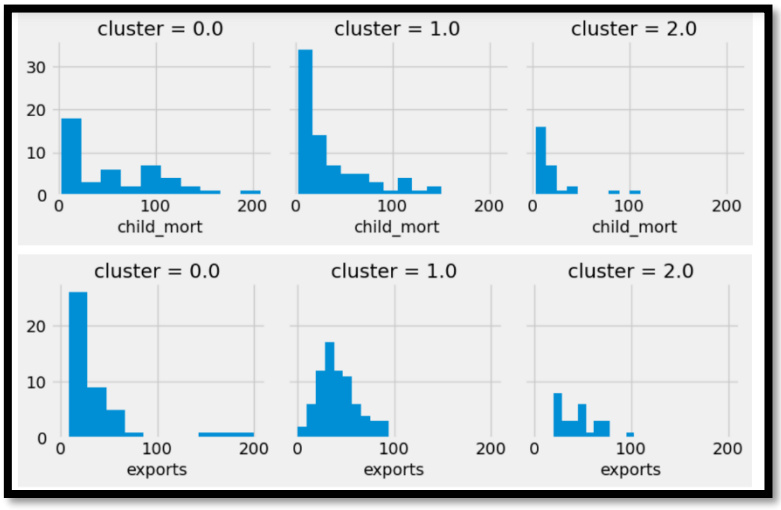


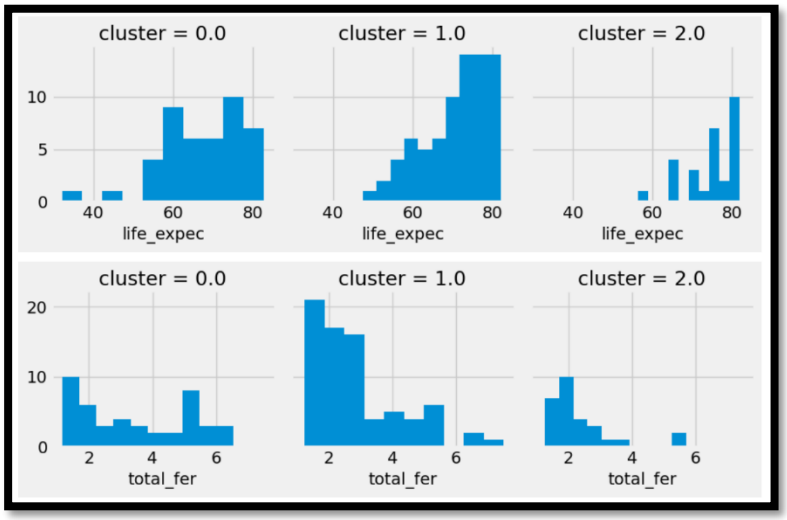


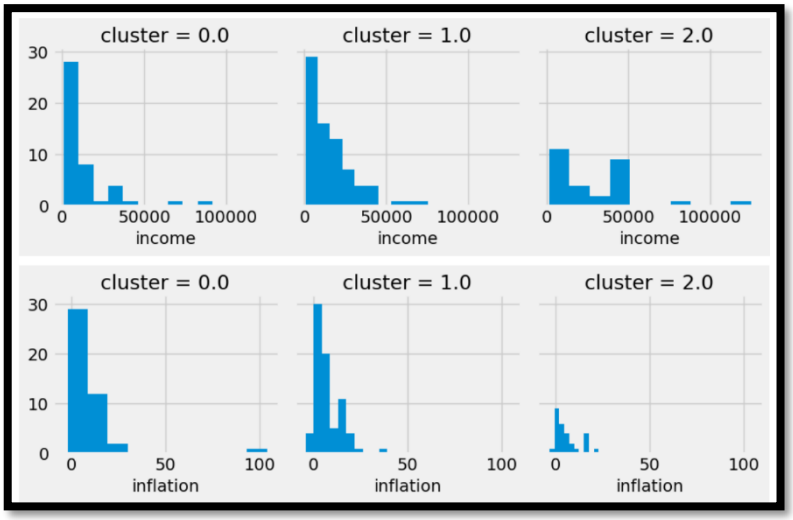


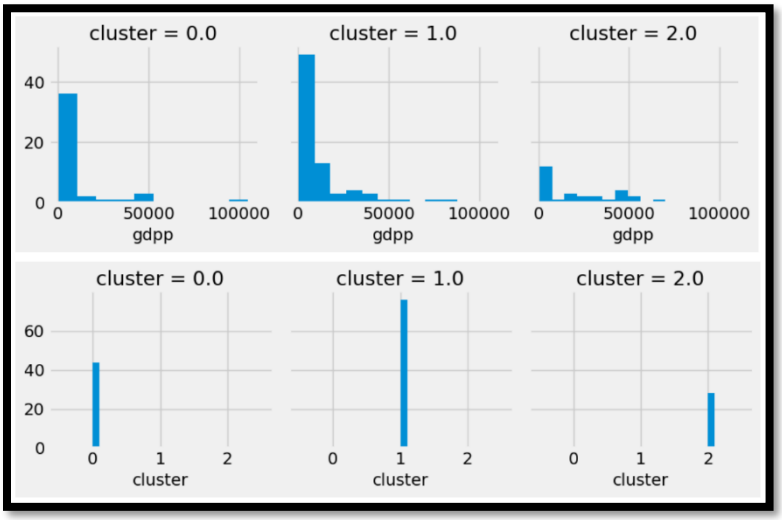
After analyzing the features separately for each cluster (the following plots), the conditions and characteristics of each cluster were determined.











Clustering is as follows:

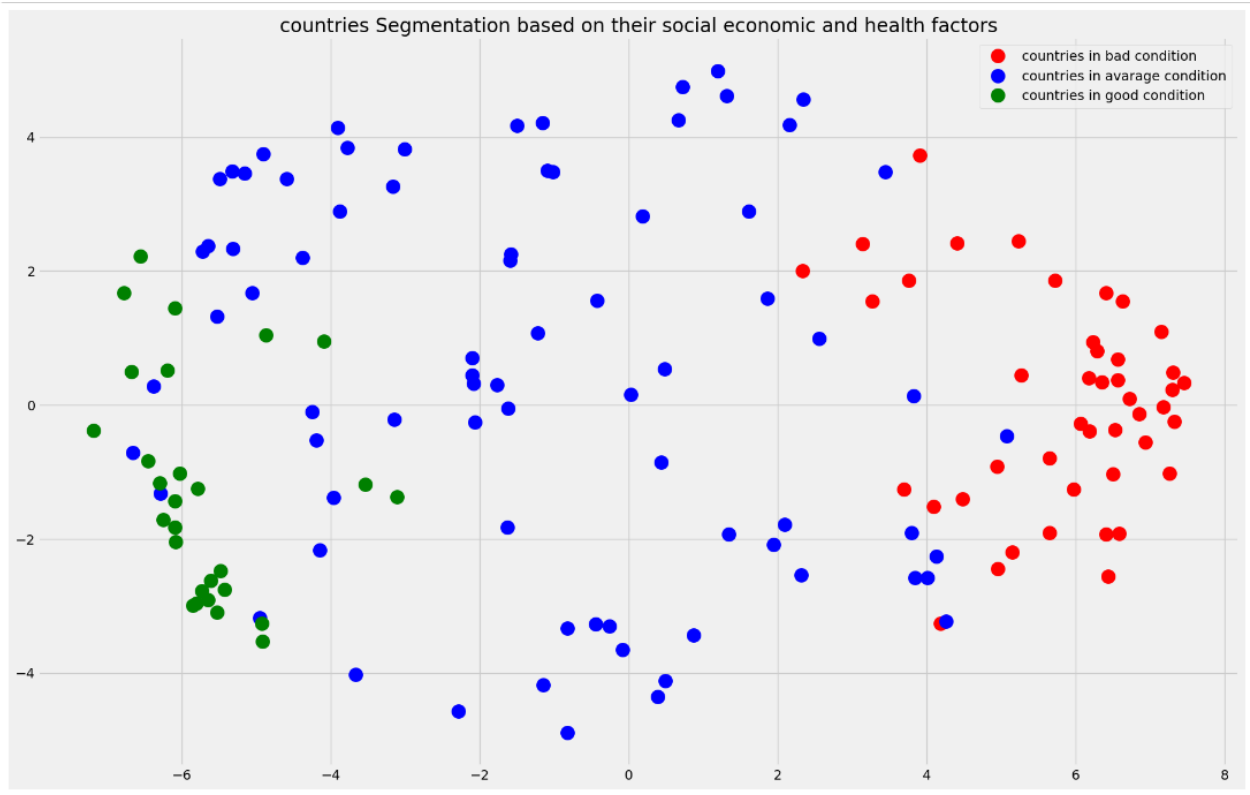
Cluster 0: Countries that are in worse conditions in terms of social, economic, and health factors. In these countries, according to the above plots, "child mortality rate" and "number of children born" are higher, but "income" and "gdpp" are lower.

Cluster 1: Countries that are in average conditions in terms of social, economic, and health factors compared to other countries. In these countries, according to the above plots, the rate of "exports" and "imports" is higher.

Cluster 2: Countries that are in better conditions in terms of social, economic, and health factors compared to other countries. In these countries, according to the above plots, "child mortality rate", "number of children born", "inflation", and "gdpp" are lower, but "income" and "life expectancy" are higher.

✓ PCA

Using PCA, we can display all countries with their features on a plot.



Based on the above cluster analysis and the conditions of the countries, the countries in Cluster 0 require the highest amount of budget and aid, followed by Cluster 1 and finally Cluster 2, which had better conditions.

The countries in Cluster 0 that require the most aid in comparison to others are:

```
clusters2[clusters2.cluster == 0].country.values
```

```
array(['Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso',
       'Burundi', 'Cameroon', 'China', 'Colombia', 'Comoros',
       'Congo, Rep.', 'Egypt', 'El Salvador', 'Finland', 'France',
       'Georgia', 'Grenada', 'Guatemala', 'Haiti', 'India', 'Jamaica',
       'Japan', 'Kenya', 'Kyrgyz Republic', 'Lesotho', 'Liberia',
       'Luxembourg', 'Macedonia, FYR', 'Mali', 'Malta', 'Mauritania',
       'Mauritius', 'Montenegro', 'Nigeria', 'Pakistan', 'Poland',
       'Rwanda', 'Senegal', 'Serbia', 'Sierra Leone', 'Singapore',
       'South Korea', 'Tajikistan', 'Tanzania'], dtype=object)
```

The countries in Cluster 1 that have average conditions and require less aid compared to Cluster 0 are:

```
array(['Albania', 'Algeria', 'Antigua and Barbuda', 'Argentina',
       'Armenia', 'Azerbaijan', 'Bangladesh', 'Barbados', 'Belarus',
       'Belize', 'Bhutan', 'Bolivia', 'Bosnia and Herzegovina', 'Brazil',
       'Bulgaria', 'Cambodia', 'Cape Verde', 'Central African Republic',
       'Chad', 'Chile', 'Congo, Dem. Rep.', 'Costa Rica',
       'Czech Republic', 'Denmark', 'Dominican Republic', 'Ecuador',
       'Equatorial Guinea', 'Eritrea', 'Gabon', 'Ghana', 'Greece',
       'Guinea', 'Guinea-Bissau', 'Hungary', 'Iceland', 'Iraq', 'Israel',
       'Italy', 'Kazakhstan', 'Kiribati', 'Kuwait', 'Lao', 'Latvia',
       'Lebanon', 'Libya', 'Lithuania', 'Madagascar', 'Malawi',
       'Malaysia', 'Maldives', 'Mongolia', 'Morocco', 'Mozambique',
       'Myanmar', 'Namibia', 'Nepal', 'New Zealand', 'Niger', 'Norway',
       'Oman', 'Panama', 'Paraguay', 'Philippines', 'Romania', 'Russia',
       'Samoa', 'Seychelles', 'Slovak Republic', 'Slovenia',
       'Solomon Islands', 'South Africa', 'Spain', 'Sudan', 'Suriname',
       'Sweden', 'Switzerland'], dtype=object)
```

The countries in Cluster 2 have better economic and health conditions compared to the previous two clusters and require less budget for aid. The countries in Cluster 2 are:

```
array(['Australia', 'Austria', 'Bahamas', 'Bahrain', 'Belgium', 'Brunei',
       'Canada', "Cote d'Ivoire", 'Croatia', 'Cyprus', 'Estonia', 'Fiji',
       'Gambia', 'Germany', 'Guyana', 'Indonesia', 'Iran', 'Ireland',
       'Jordan', 'Micronesia, Fed. Sts.', 'Moldova', 'Netherlands',
       'Peru', 'Portugal', 'Qatar', 'Saudi Arabia', 'Sri Lanka',
       'St. Vincent and the Grenadines'], dtype=object)
```