Exploring Countries Development

Grigoris Ntoulaveris

Introduction

The objective of this analysis was to cluster 167 countries from the dataset Countries (https://www.kaggle.com/datasets) in order to categorize them using socio-economic and health factors that determine the overall development of the country. The dataset consisted of 9 features and 167 data points. Those features were:

1. Child_mortality: Death of children under 5 years of age per 1000 live births.
2. Exports: Exports of goods and services per capita. Given as %age of the GDP per capita.
3. Health: Total health spending per capita. Given as %age of GDP per capita.
4. Imports: Imports of goods and services per capita. Given as %age of the GDP per capita.
5. Income: Net income per person.
6. Inflation: The measurement of the annual growth rate of the Total GDP.
7. Life_expectancy: The average number of years a new born child would live if the current mortality patterns are to remain the same.
8. Total_fertility: The number of children that would be born to each woman if the current age-fertility rates remain the same.
9. GDPP: The GDP per capita (Calculated as the Total GDP divided by the total population).

Preprocessing of the data

The analysis began by exploring each feature of the dataset. Every feature was of type “double”, meaning numerical values with double precision. Below are presented some statistics of the data and the distribution of each feature.

The histograms were used to visualize the distribution of each feature and to search for unwanted linearities in them. No feature presented such behavior, so at this stage no feature was excluded from the analysis.

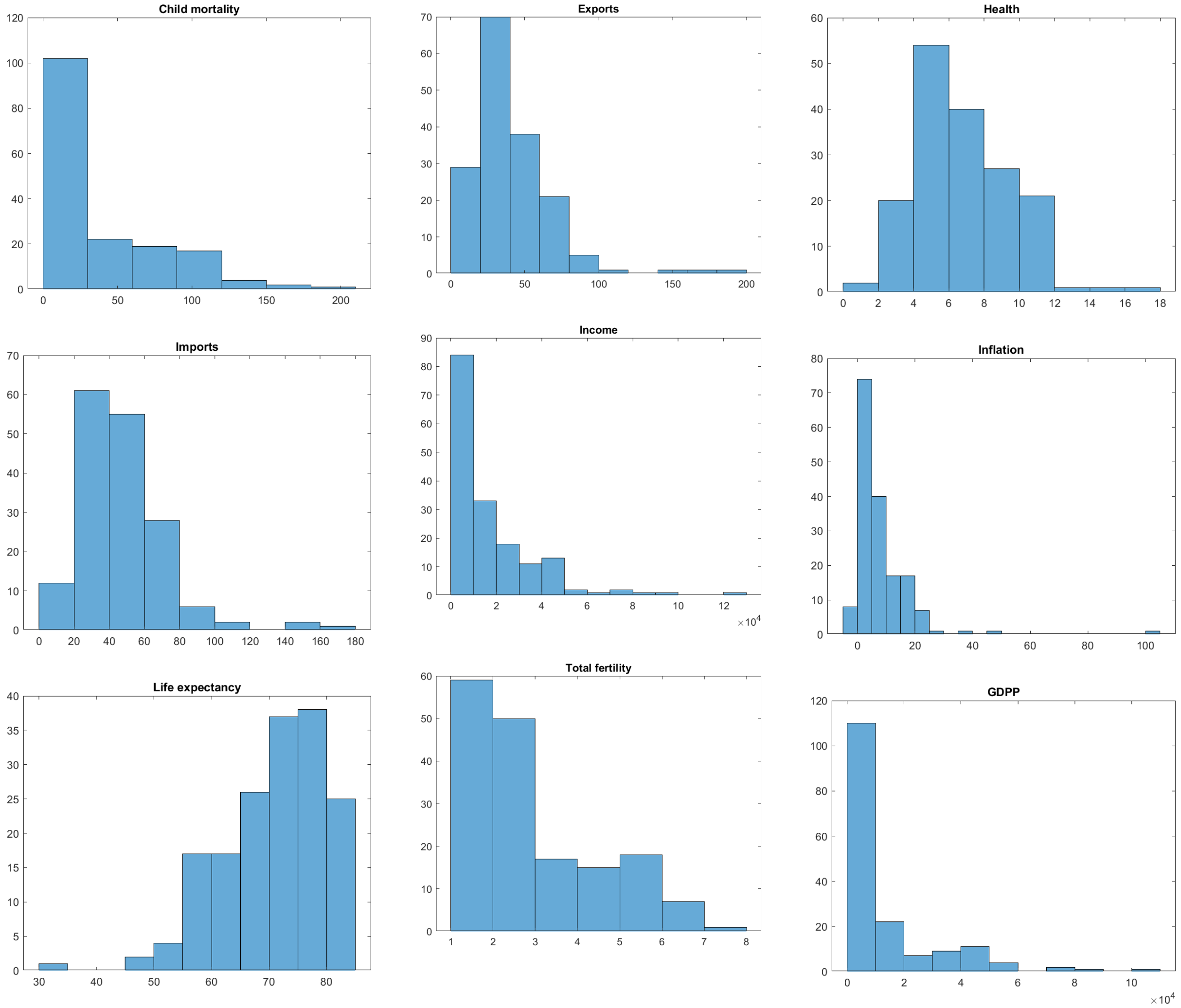


Figure 1. Histograms of every feature of the dataset.

Based on the minimum and maximum values of the features it is apparent that a normalization of the values is needed to accurately cluster the data, as the range of every feature is very different from all others. The features were normalized first by a standard score normalization and then by a min-max normalization. It should also be noted that the features *child mortality, exports* and *imports* presented a high standard deviation in regards to their range, meaning that those data presented high variability.

Feature	Minimum value
Child mortality	2.60
Exports	0.1090
Health	1.8100
Imports	0.0659
Income	609
Inflation	-4.2100
Life expectancy	32.100
Total fertility	1.1500
GDPP	231

Feature	Maximum value
Child mortality	208
Exports	200
Health	17.900
Imports	174
Income	125000
Inflation	104
Life expectancy	82.8000
Total fertility	7.4900
GDPP	105000

Table 1. The minimum value of every feature of the dataset Countrydata.

Feature	Mean value
Child mortality	38.2701
Exports	41.1090
Health	6.8157
Imports	46.8902
Income	17145e+04
Inflation	7.7818
Life expectancy	70.5557
Total fertility	2.9480
GDPP	1.2964e+04

Table 3. The mean value of every feature of the dataset Countrydata.

Table 2. The maximum value of every feature of the dataset Countrydata.

Feature	Standard deviation
Child mortality	40.3289
Exports	27.4120
Health	2.7468
Imports	24.2096
Income	1.9278e+04
Inflation	10.5707
Life expectancy	8.8932
Total fertility	1.5138
GDPP	1.8329e+04

Table 4. The standard deviation of every feature of the dataset Countrydata.

The boxplot for each feature was also created to evaluate the presence of outliers as well as the subsequent median values. The boxplots were created for the non-normalized features, for the standard score normalized features, for the min-max normalized features and finally for the features that were normalized with both standard score and min-max normalization.

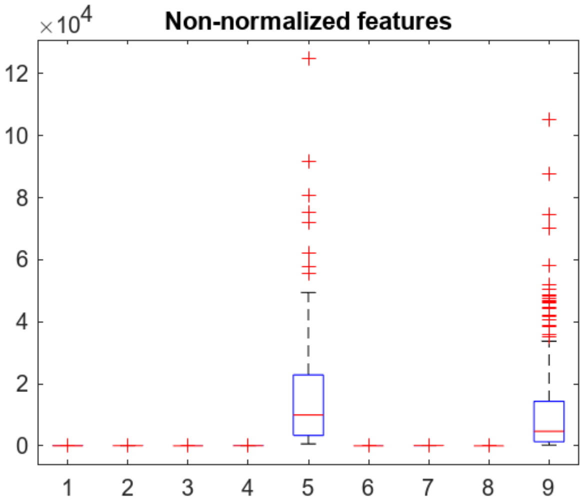


Figure 2. Boxplots of non-normalized features.

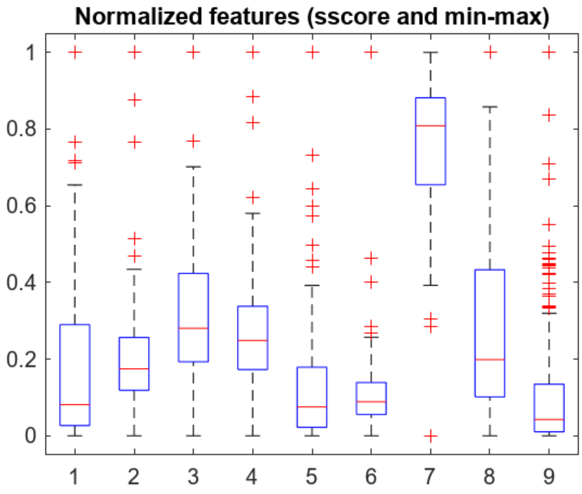


Figure 3. Boxplots of normalized features with both normalizations.

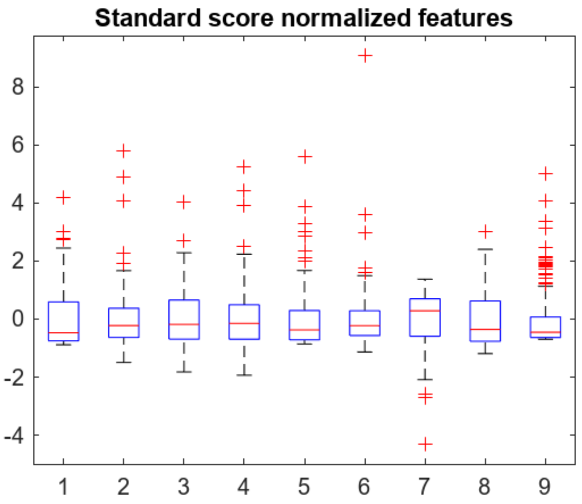


Figure 4. Boxplots of standard score normalized features.

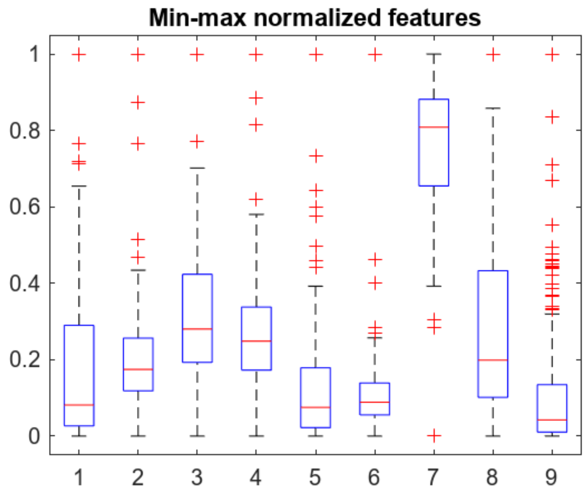
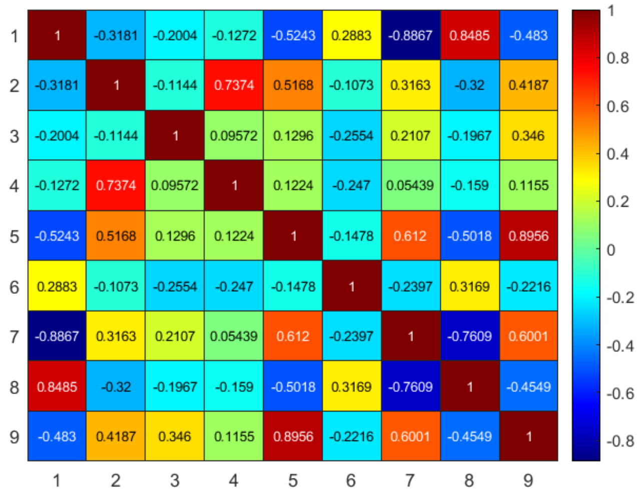


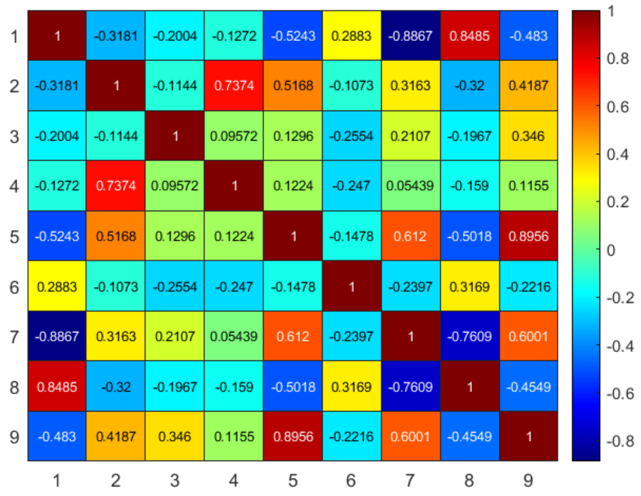
Figure 5. Boxplots of min-max normalized features.

Below the correlation coefficient matrix of all features is presented in order of appearance of the above tables. The first matrix corresponds to the correlation coefficient of the non-normalized features and the second to that of the normalized features (standard score and min-max normalization). Both matrices are identical which is to be expected. The size (range) of the features may have changed but their corresponding linear dependencies have not.



	1	2	3	4	5	6	7	8	9
1	1	-0.3181	-0.2004	-0.1272	-0.5243	0.2883	-0.8867	0.8485	-0.483
2	-0.3181	1	-0.1144	0.7374	0.5168	-0.1073	0.3163	-0.32	0.4187
3	-0.2004	-0.1144	1	0.09572	0.1296	-0.2554	0.2107	-0.1967	0.346
4	-0.1272	0.7374	0.09572	1	0.1224	-0.247	0.05439	-0.159	0.1155
5	-0.5243	0.5168	0.1296	0.1224	1	-0.1478	0.612	-0.5018	0.8956
6	0.2883	-0.1073	-0.2554	-0.247	-0.1478	1	-0.2397	0.3169	-0.2216
7	-0.8867	0.3163	0.2107	0.05439	0.612	-0.2397	1	-0.7609	0.6001
8	0.8485	-0.32	-0.1967	-0.159	-0.5018	0.3169	-0.7609	1	-0.4549
9	-0.483	0.4187	0.346	0.1155	0.8956	-0.2216	0.6001	-0.4549	1

Figure 6. The correlation coefficient matrix of all features of the dataset (1. Child mortality, 2. Exports, 3. Health, 4. Imports, 5. Income, 6. Inflation, 7. Life expectancy, 8. Total fertility, 9. GDPP)



	1	2	3	4	5	6	7	8	9
1	1	-0.3181	-0.2004	-0.1272	-0.5243	0.2883	-0.8867	0.8485	-0.483
2	-0.3181	1	-0.1144	0.7374	0.5168	-0.1073	0.3163	-0.32	0.4187
3	-0.2004	-0.1144	1	0.09572	0.1296	-0.2554	0.2107	-0.1967	0.346
4	-0.1272	0.7374	0.09572	1	0.1224	-0.247	0.05439	-0.159	0.1155
5	-0.5243	0.5168	0.1296	0.1224	1	-0.1478	0.612	-0.5018	0.8956
6	0.2883	-0.1073	-0.2554	-0.247	-0.1478	1	-0.2397	0.3169	-0.2216
7	-0.8867	0.3163	0.2107	0.05439	0.612	-0.2397	1	-0.7609	0.6001
8	0.8485	-0.32	-0.1967	-0.159	-0.5018	0.3169	-0.7609	1	-0.4549
9	-0.483	0.4187	0.346	0.1155	0.8956	-0.2216	0.6001	-0.4549	1

Figure 7. The correlation coefficient matrix of the normalized features of the dataset (1. Child mortality, 2. Exports, 3. Health, 4. Imports, 5. Income, 6. Inflation, 7. Life expectancy, 8. Total fertility, 9. GDPP)

Child mortality and *total fertility* present a high correlation to one another, as does *income* and *GDPP*. Even so, they were kept intact in the following analysis as they were deemed important for the nature of the clustering goal. Therefore, all features were used in a first analysis with a hard k-means algorithm to visualize the created clusters, which would help us determine if some of the features could be excluded from later experiments, due to an unclear separation of the data points to subsequent clusters.

Experiments

Experiment 1 (hard k-means clustering with all features)

For the previously established reasons an elbow method was used to determine the optimal number of clusters for the experiment. The “elbow” occurred at 3. Therefore the k-means algorithm was initialized with three representatives.

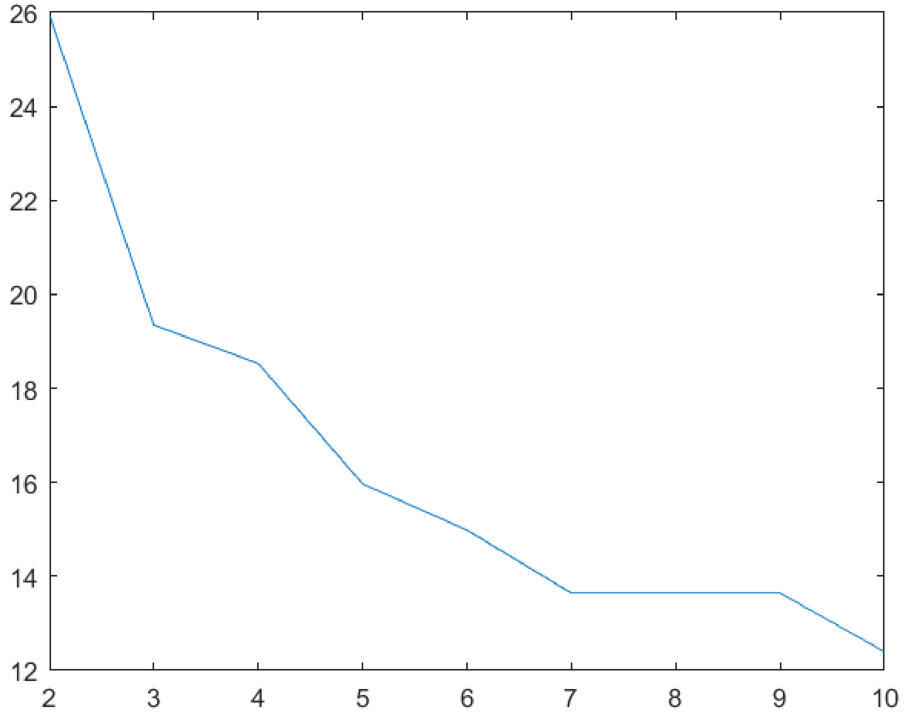
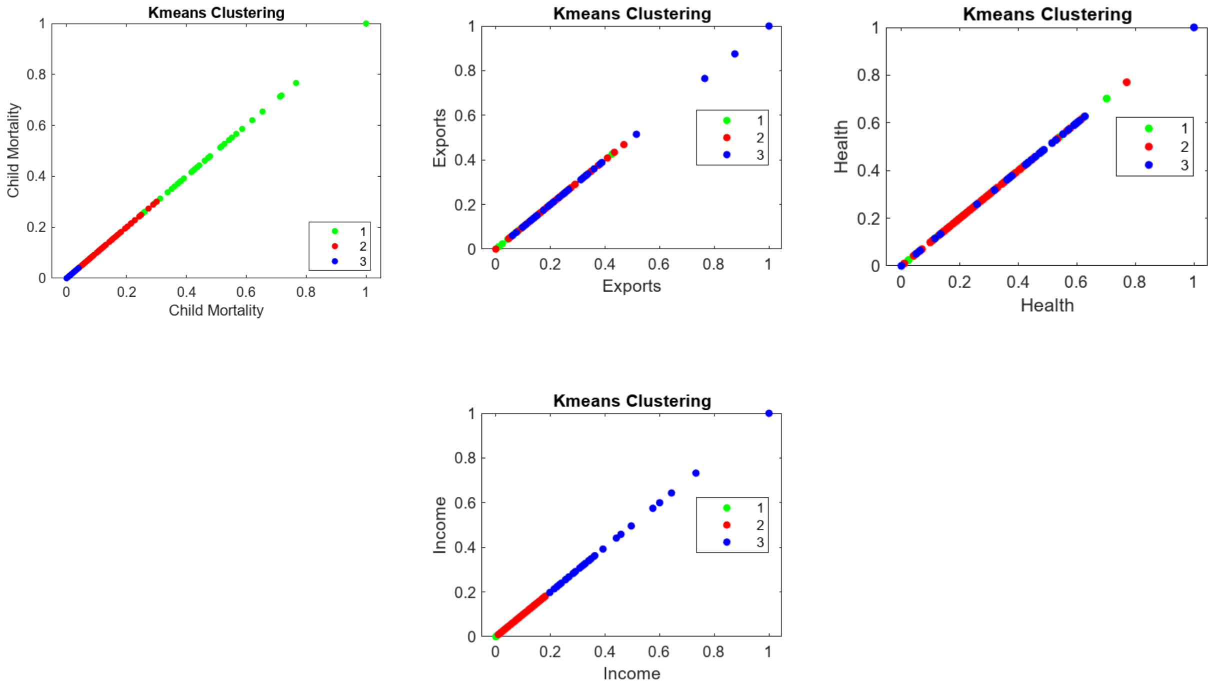


Figure 8. Elbow method of the first experiment (all features are included).

Each feature was plotted with itself in an attempt to discern features that didn't create well defined clusters. Those features could pose a problem in later clustering experiments by disfiguring the created clusters, thus making it hard to discern patterns in the data. As observed, the features *exports, health, imports* and *inflation* didn't create well defined clusters. Therefore they were excluded from the following experiments. The new dataset consisted of the five remaining features.



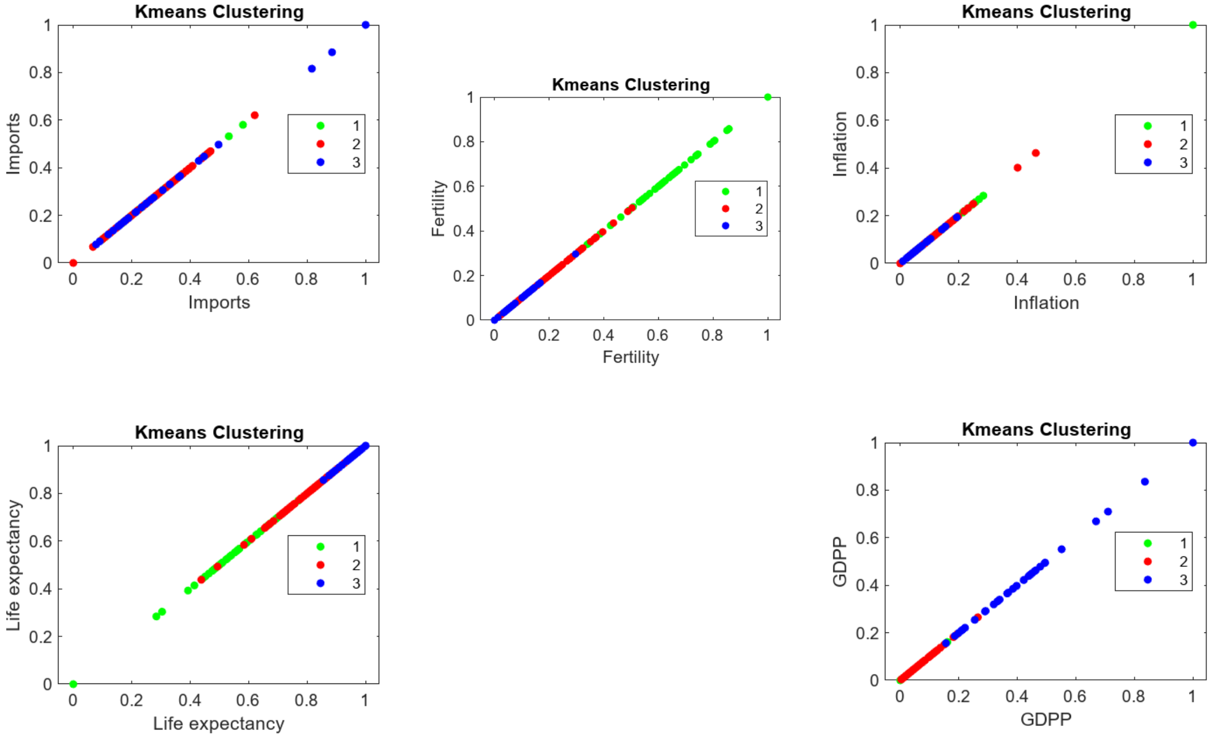


Figure 9. Clustering results of hard k-means of all features compared to themselves.

Experiment 2 (hard k-means with rand_data_init and selected features)

In this experiment a hard k-means algorithm was implemented on the remaining features. An elbow method was used to determine the optimum number of clusters, which was found to be once again 3. The representatives were initialized using rand_data_init. After that the analysis proceeded by plotting the clusters in two and three dimensions. In both cases all possible combinations of features were tested. Each country was classified in one of the three categories, Less Developed, Medium Developed, Highly Developed, according to the socio-economic and health reasons established by the features and their values for each cluster. For example countries of the blue cluster presented high GDPP and life expectancy among others, which were deemed as characteristics of a highly developed country. Based on similar intuition, the red cluster was characterized as countries that are medium developed and the countries of the green cluster were characterized as less developed.

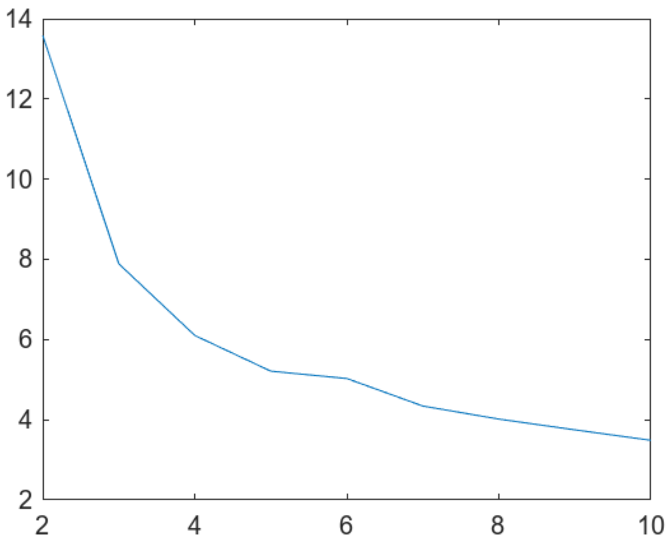
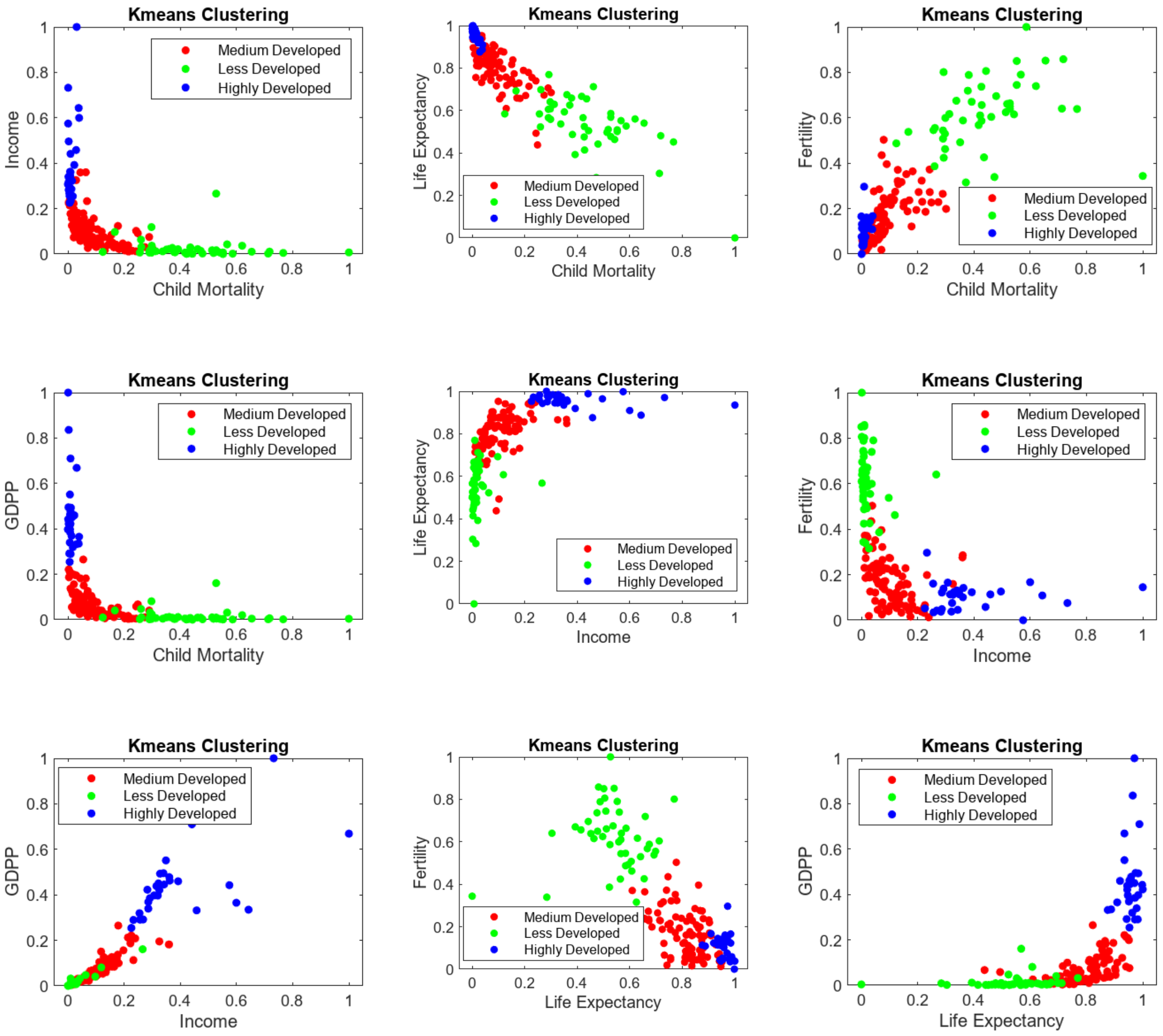


Figure 10. Elbow method for the hard k-means.



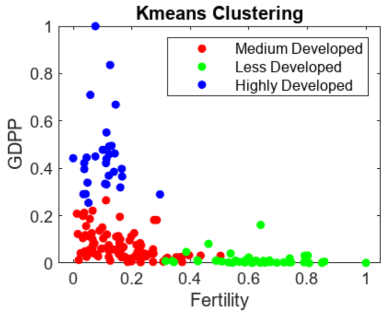


Figure 11. Clustering results of the hard k-means algorithm in two dimensions.

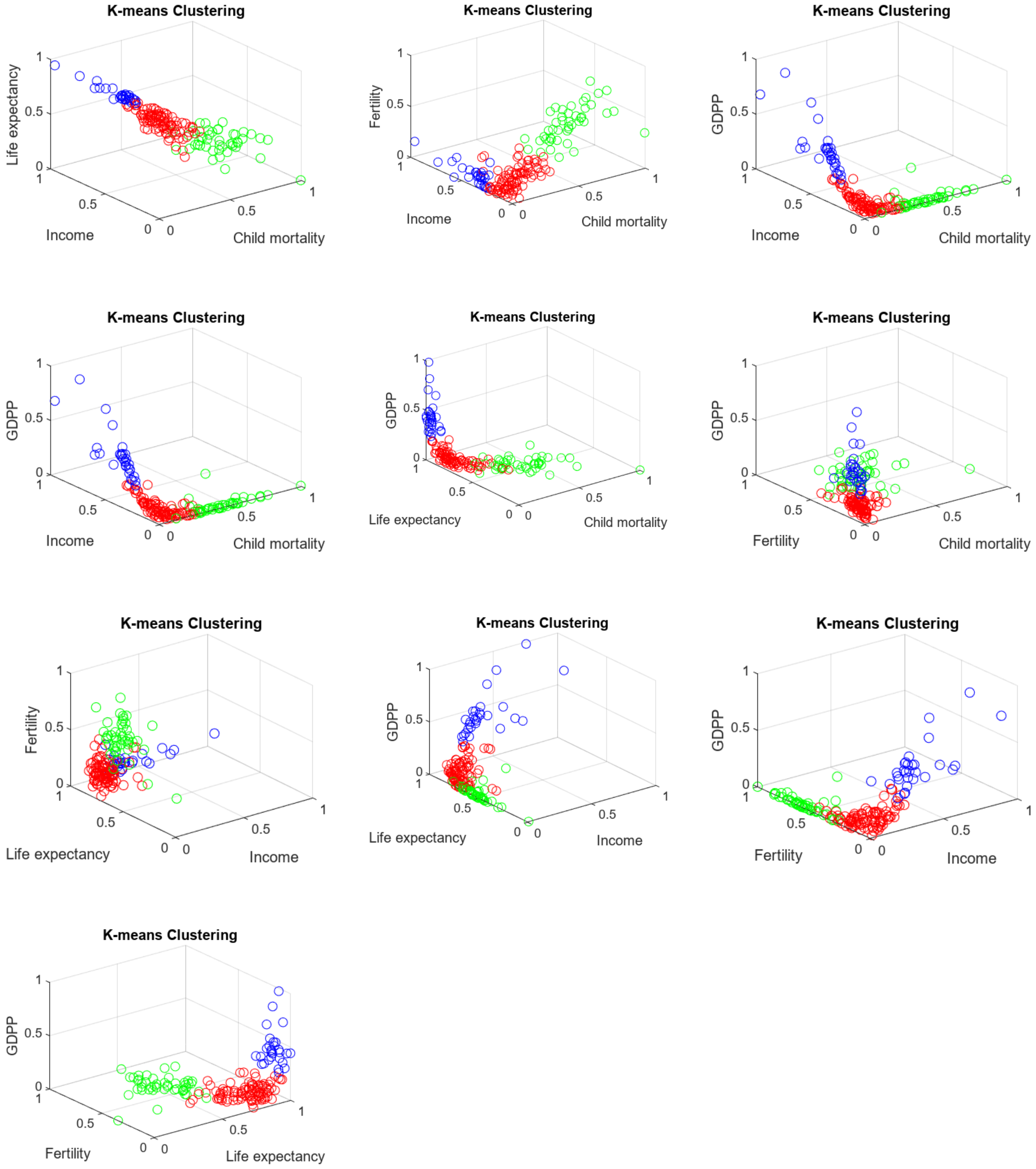


Figure 12. Clustering results of the hard k-means algorithm in two dimensions.

Experiment 3 (hard k-means with rand_init and selected features)

This experiment was done by implementing the hard k-means once more, but this time with a different initialization method, the rand_init. The objective of this experiment was to determine whether the formed clusters were similar to the first implementation of the k-means algorithm with the rand_data_init. Evidently the clusters that were formed were the same as the previous version of the algorithm, with slight differences of a few (1-4) data points.

It is noted that for seed = 1 only two clusters were formed, probably because the third representative was initialized far away from the data dense areas and therefore no point was assigned to it before the termination of the algorithm. Nonetheless, this was deemed as a wrong result both because no other k-means initialization presented such outcomes but also because it deters from the established objective of the problem.

Below are presented some of the clustering figures of this initialization.

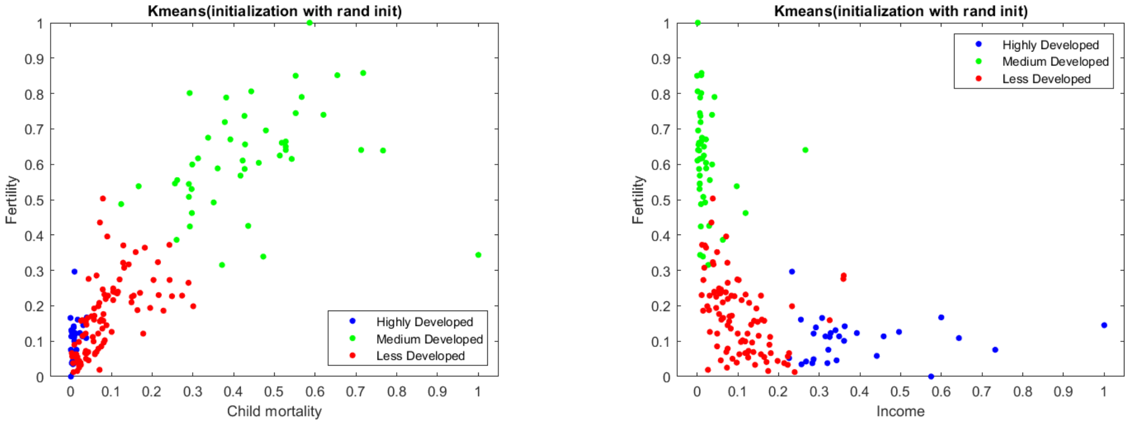


Figure 13. Clustering results of the hard k-means algorithm with dist_init.

Experiment 4 (hard k-means with dist_init and selected features)

Similarly to experiment 3, this experiment was also done to determine if similar clusters would once again be formed. This was also the case here and below are presented some of the clustering figures that are characterized by great similarity (1-4 points only differences) with the previous initializations and most importantly with the rand_data_init initialization which was chosen as the end implementation for the conclusive clustering.

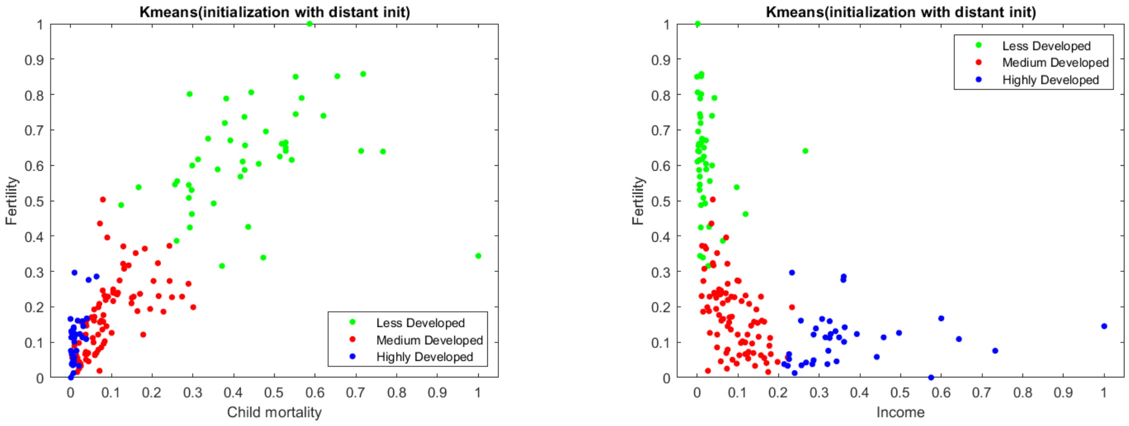


Figure 14. Clustering results of the hard k-means algorithm with dist_init.

Experiment 5 (hard k-medians with selected features)

In this experiment a hard k-medians algorithm was implemented on the remaining features. An elbow method was used to determine the optimum number of clusters, which was found to be once again 3. After that the analysis proceeded by plotting the clusters in two and three dimensions. However, in this case not all feature combinations were tested as it was observed that some of the clustering results of the k-medians algorithm didn't coincide with reality or at least weren't as accurate as the results of the hard k-means algorithm.

Mainly, the countries Portugal, Czech Republic, Slovenia and Bahrain were appointed in the Highly Developed countries cluster (blue color) whereas in the case of k-means they were appointed to the Medium Developed countries cluster (red color). The results of k-means were deemed more reality based given the health and socio-economic situation of those countries.

The difference in the results between the two algorithms could be attributed to the fact that k-medians is less influenced by outliers, which those countries apparently were, given the current features.

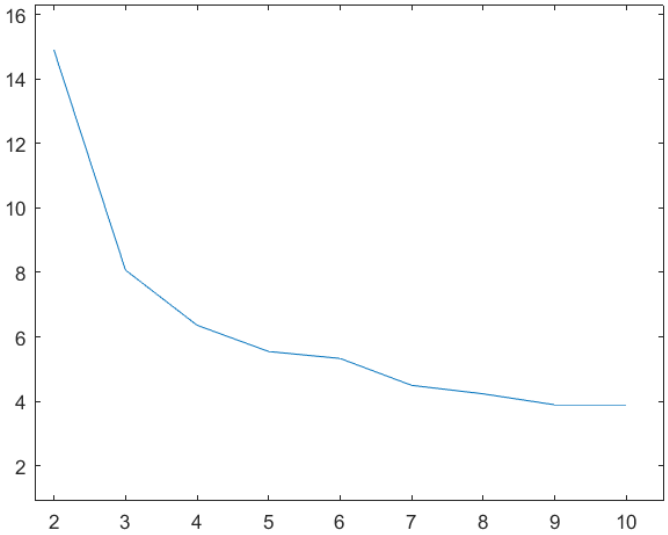


Figure 15. Elbow method for the hard k-medians.

.

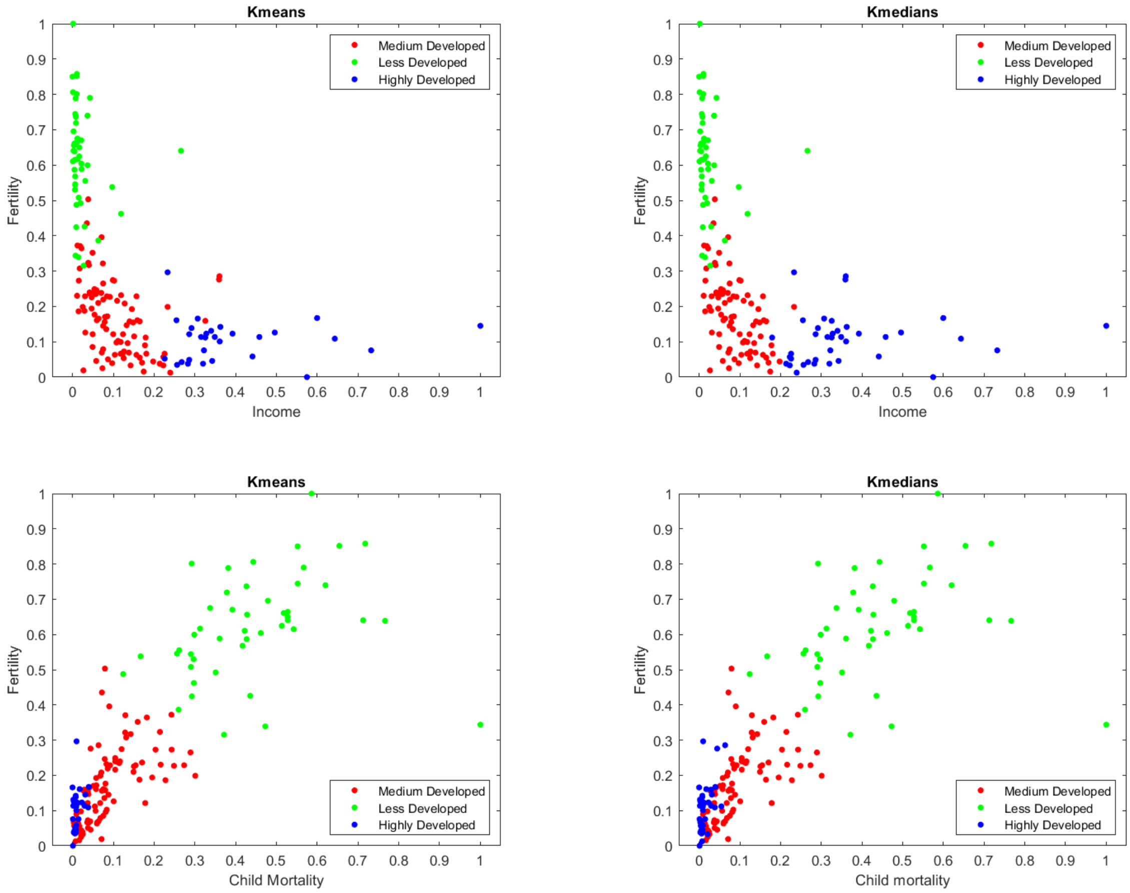


Figure 16. Slight differences are observed between the clustering results of the k-means (left column) and the k-medians (right column).

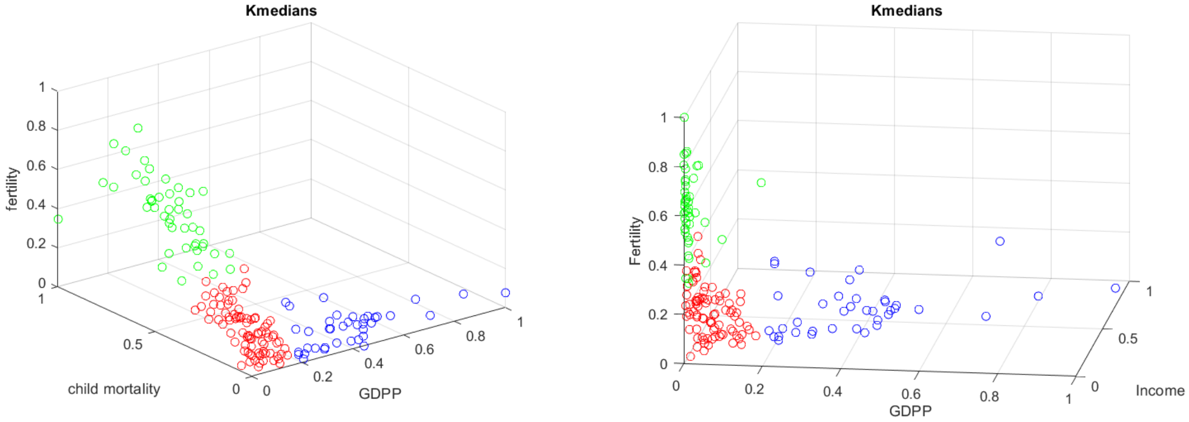


Figure 17. 3D plot of the clustering results of the k-medians algorithm in some of the features.

Experiment 6 (hard k-medoids with selected features)

In this experiment a hard k-medoids algorithm was implemented on the remaining features. As it was previously decided that the countries would be classified in three distinct categories, three initial clusters were used in this algorithm as well. After that, the analysis proceeded by plotting the clusters in two and three dimensions. In this case also not all feature combinations were tested as it was observed that some of the clustering results of the k-medians algorithm didn’t coincide with reality or at least weren’t as accurate as the results of the hard k-means algorithm.

Mainly Oman and Saudi Arabia were classified as less developed countries by k-medoids which seems intuitively wrong given their health and socio-economic status. K-means on the other hand classified them as medium developed countries, which once again seems to be closer to reality.

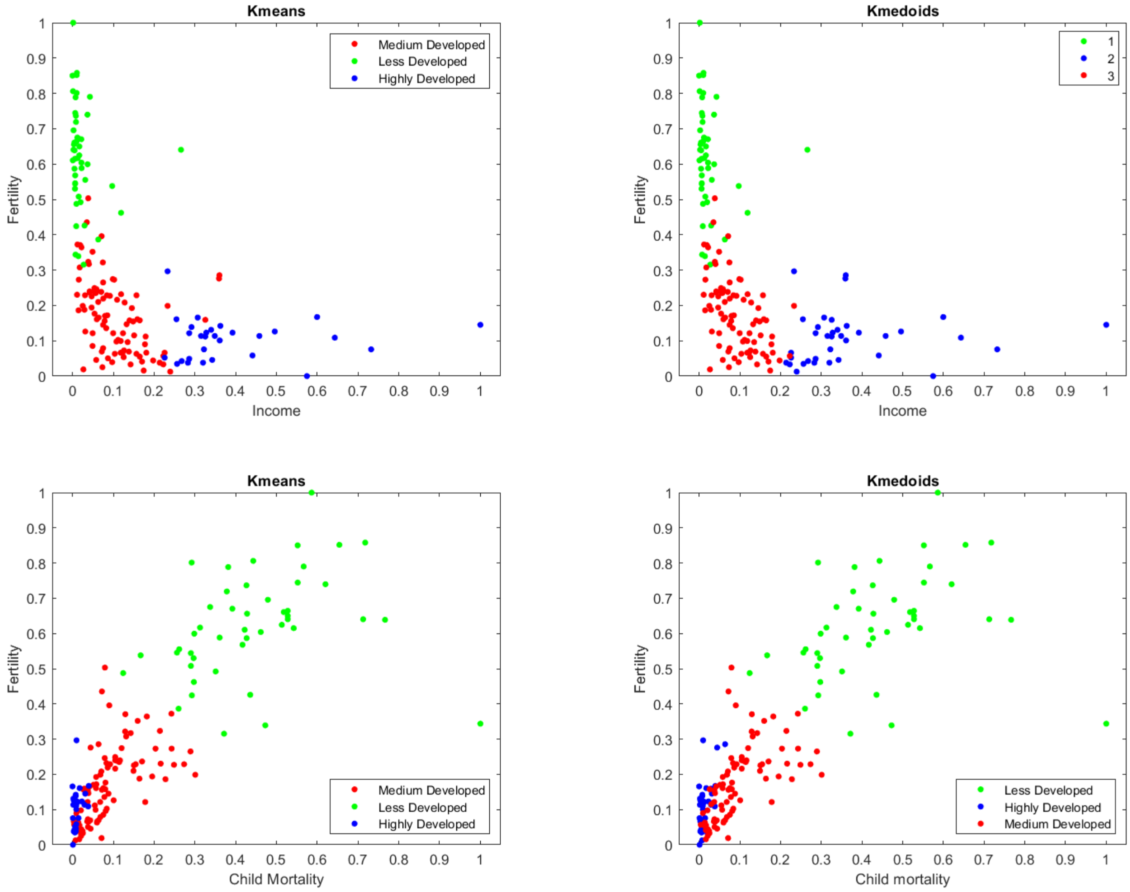


Figure 18. Slight differences are observed between the clustering results of the k-means (left column) and the k-medoids (right column).

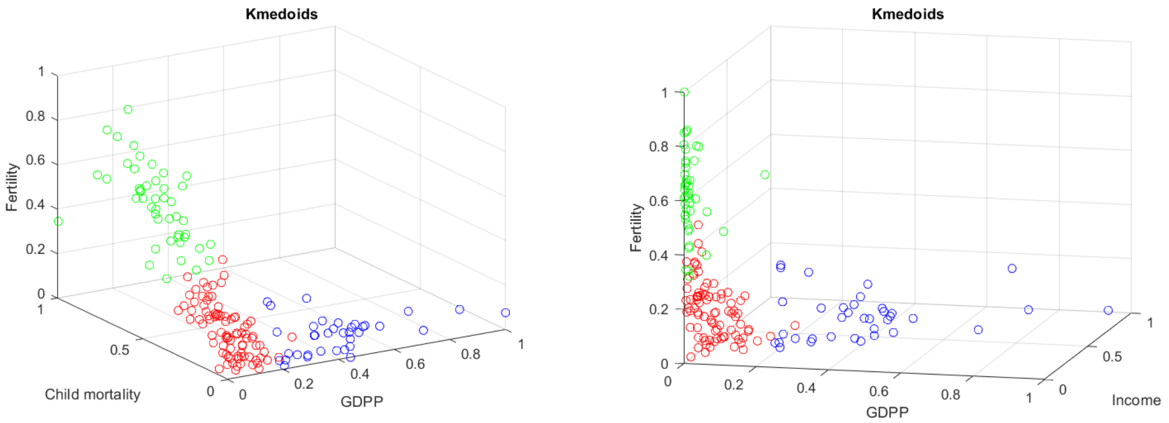


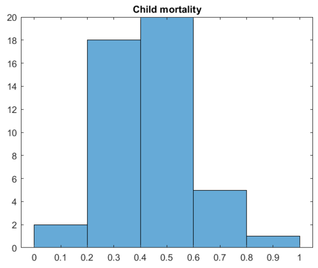
Figure 19. 3D plot of the clustering results of the k-medoids algorithm in some of the features.

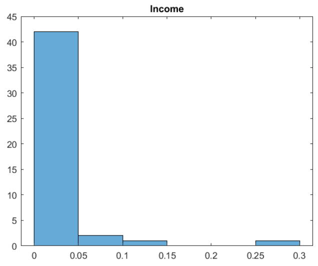
Conclusion

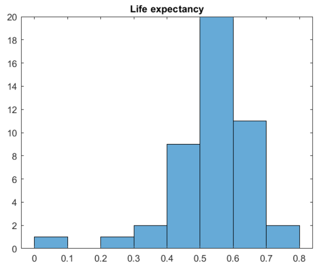
The algorithm of choice for the bulk of the experiments as well as for the final clustering results was the hard k-means algorithm with rand_data_init. That was for two reasons. The first one was based on intuition, because as established in the individual experiments, k-means generated clustering results that better coincided with the true economical, social and health status of some of the countries of the dataset. The second reason was that all in all the clustering results between the two algorithms were pretty similar (with the exception of the few aforementioned countries) so k-means was deemed as the best choice because it is faster than k-medians and k-medoids.

Below are presented the histograms with distributions of each feature for each cluster.

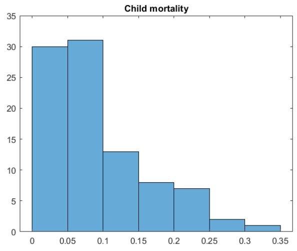
Less Developed

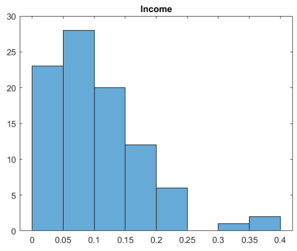


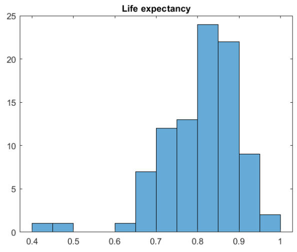




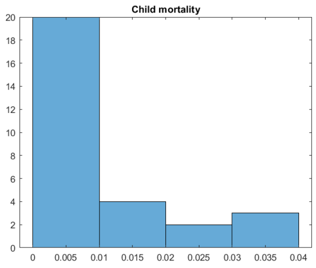
Medium Developed

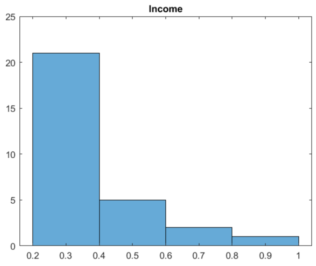


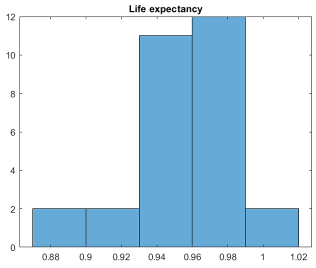




Highly Developed







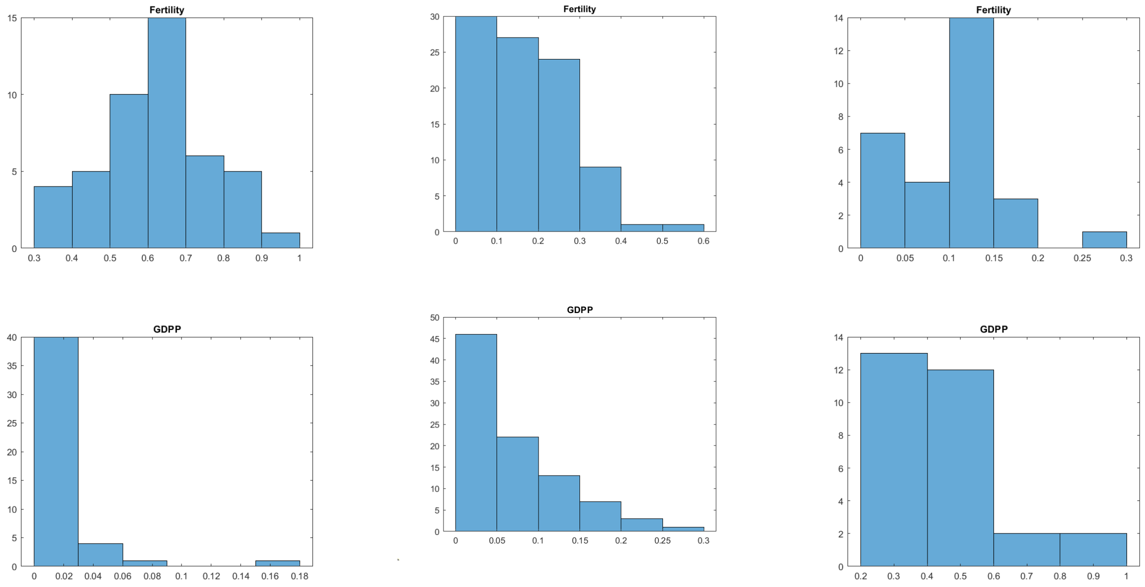


Figure 20. Histograms for each feature in each cluster.

What is interesting to be observed in the above histograms is the range of values especially in the economical features of the *GDPP* and *income* where the Less Developed countries are characterized by really low values, almost close to 0, and on the other end of the spectrum the Highly Developed countries have a range of values that begins almost where the values of the Medium Developed countries begin. Beyond the socio-economic features, the feature of *child mortality* should also be observed closely, as it gives high values for the Less Developed countries. The other clusters have once again characteristically distant values compared to the Less Developed countries.

Those results are further explored in Table 5, where the mean and standard deviation values for each feature in each cluster is presented. The number of countries in each cluster is also presented, giving an idea of the amount of countries in need of socio-economic and most importantly health aid.

	Less Developed	Medium Developed	Highly Developed
Number of countries	46	92	29
Mean values	[0.4415, 0.0234, 0.5383, 0.6215, 0.0140]	[0.0909, 0.1058, 0.8064, 0.1713, 0.0708]	[0.0113, 0.3929, 0.9556, 0.1039, 0.4532]
Standard deviation	[0.1659, 0.0433, 0.1277, 0.1483, 0.0267]	[0.0713, 0.0733, 0.0926, 0.1064, 0.0586]	[0.0109, 0.1725, 0.0299, 0.0586, 0.1666]

Table 5. The characteristics of each cluster. Each value of the mean and stdv vectors corresponds to one of the five selected features. In order of appearence from left to right: child mortality, income, life expectancy, fertility, GDPP.

As observed on Table 5, most of the countries of the dataset belong to the Medium Developed class, followed in numbers by the Less Developed and finally by the Highly Developed ones. The mean values further support what had been observed in the clustering figures. For example, it is evident that the mean value of the *GDPP* feature is a lot more prominent in the Highly Developed cluster compared to the other two, with the corresponding values for the other two clusters being closer. The reverse results were observed in the *fertility* feature, with the Less Developed cluster having more than 10 times greater values than the other two clusters, whose values for that feature are pretty close. Perhaps more important is the observation of the *child mortality* feature, which presents a very high mean value in the Less Developed countries, highlighting the need for support in that regard. The *life expectancy* is more balanced but still lower in the Less Developed countries and as expected from the GDPP, the *income* feature is also a whole digit lower in the Less Developed countries than it is on the others, with the Highly Developed countries having almost 4 times higher individual income than the Medium Developed countries.

All in all, those results and those of the histograms indicate a large imbalance in socio-economic and health prosperity between the less developed countries and the medium or highly developed ones, with some features indicating the need for medical assistance and advancement.

The clustering results of the k-means algorithm were used to create a world map plot of the 167 countries of the dataset to discern more easily their overall development. This plot is presented below.

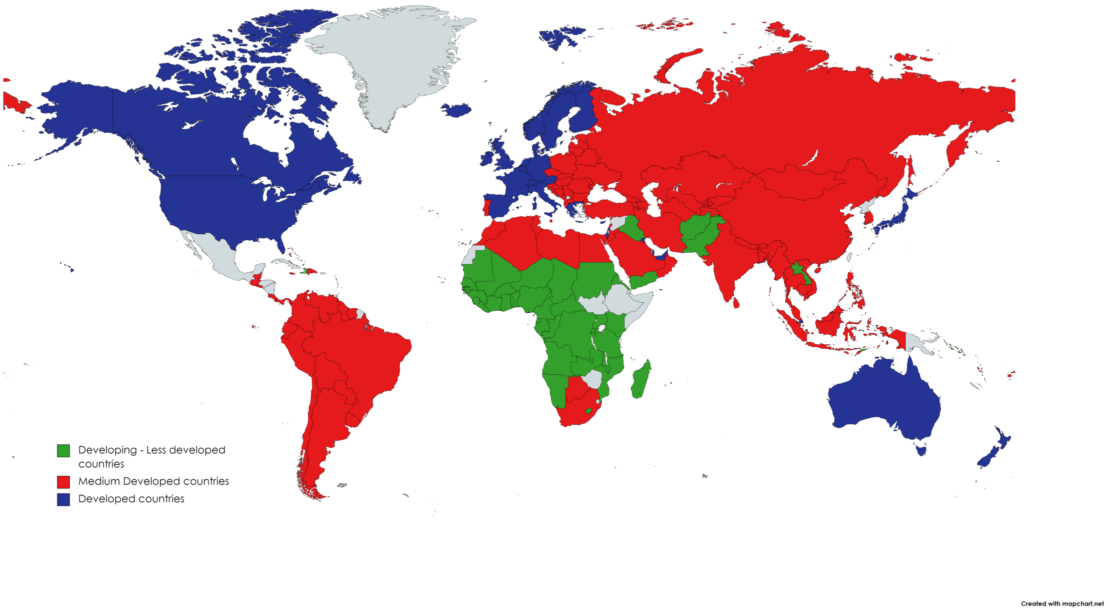

🟩 Developing - Less developed countries
🟥 Medium Developed countries
🟦 Developed countries
Created with mapchart.net

Figure 21. World map with colorings based on the clustering results. Highly developed countries are presented in blue, medium developed countries in red and less developed countries in green.

Appendix

```
load data_country

%extract every col (feature) in variables
%health factors
child_mort = Countrydata(:,1);
health = Countrydata(:,3);
life_exp = Countrydata(:,7);
fertility = Countrydata(:,8);

%socioeconomic factors
exports = Countrydata(:,2);
imports = Countrydata(:,4);
income = Countrydata(:,5);
inflation = Countrydata(:,6);
gdpp = Countrydata(:,9);

%Determine the type of every column
```

```
class(child_mort) %child mortality
class(health) %health
class(life_exp) %life expectancy
class(fertility) %total fertility

class(exports) %exports
class(imports) %imports
class(income) %income
class(inflation) %inflation
class(gdpp) %GDPP

%Determine the range of values for every feature
%range = max - min
mins = min(Countrydata);
maxs = max(Countrydata);
range = max(Countrydata) - min(Countrydata);

%See the statistics of the data (mean, stdv)
means = mean(Countrydata);
stdv = std(Countrydata);

%Create the hist for each feature to see its distribution
figure(1), histogram(Countrydata(:,1))
figure(2),histogram(Countrydata(:,2))
figure(3),histogram(Countrydata(:,3))
figure(4),histogram(Countrydata(:,4))
figure(5),histogram(Countrydata(:,5))
figure(6),histogram(Countrydata(:,6))
figure(7),histogram(Countrydata(:,7))
figure(8),histogram(Countrydata(:,8))
figure(9),histogram(Countrydata(:,9))

%Standarization
%Boxplots to see if there are outliers
boxplot(Countrydata)

%Compute corr coeff to see linear dependance of each feature
cor = corrcoef(Countrydata);
heatmap(cor)
colormap jet

%Standard score normalization
features = [child_mort,exports,health,imports,income,inflation,life_exp,fertility,gdpp];
zscored_features = zscore(features);

%Min-Max with standard score normalization
normalized_features = (zscored_features - min(zscored_features)) ./ (max(zscored_features) - min(zscored_features));

%Min Max without standard score normalization
minmax_only = (features - min(features)) ./ (max(features) - min(features));

%Frame with all features scaled with standard score and min max
new_corr_coeff = corrcoef(normalized_features);
boxplot(normalized_features)

%Heatmap for the normalized features
heatmap(new_corr_coeff)
colormap jet

%Box-plot features only with standard score
boxplot(zscored_features)

%Box-plots only with min max normalization
boxplot(minmax_only)
```

```
Section 2: Clustering experiments

%1. First Kmeans experiment (all features included)

%ELBOW to determine the number of clusters
final = transpose(normalized_features);
%elbow
% Compute J_m and plot J_m versus m
```

```
[l,N]= size(final);
nruns=10;
m_min=2;
m_max=10;
J_m=[];
for m=m_min:m_max
    J_temp_min=inf;
    for t=1:nruns
        rand('seed',100*t)
        theta_ini=rand(l,m);
        [theta,bel,J]= k_means(final,theta_ini);
        if(J_temp_min>J)
            J_temp_min=J;
        end
    end
    J_m=[J_m J_temp_min];
end
m=m_min:m_max;
figure(2), plot(m,J_m)

K-means (3 clusters)
k = 3;
% we can use also the functions 'distant_init' and 'rand_init'
% to initialize the thetas
theta_ini = rand_data_init(final, k);
[theta,bel,J]=k_means(final,theta_ini);

%Plot features combinations
figure
gscatter(final(1,:), final(1,:), bel)
xlabel('Child Mortality')
ylabel('Child Mortality')
title('Kmeans Clustering')

figure
gscatter(final(2,:), final(2,:), bel)
xlabel('Exports')
ylabel('Exports')
title('Kmeans Clustering')

figure
gscatter(final(3,:), final(3,:), bel)
xlabel('Health')
ylabel('Health')
title('Kmeans Clustering')

figure
gscatter(final(4,:), final(4,:), bel)
xlabel('Imports')
ylabel('Imports')
title('Kmeans Clustering')

figure
gscatter(final(5,:), final(5,:), bel)
xlabel('Income')
ylabel('Income')
title('Kmeans Clustering')

figure
gscatter(final(6,:), final(6,:), bel)
xlabel('Inflation')
ylabel('Inflation')
title('Kmeans Clustering')

figure
gscatter(final(7,:), final(7,:), bel)
xlabel('Life expectancy')
ylabel('Life expectancy')
title('Kmeans Clustering')

figure
gscatter(final(8,:), final(8,:), bel)
xlabel('Fertility')
ylabel('Fertility')
title('Kmeans Clustering')
```

```
figure
gscatter(final(9,:), final(9,:), bel)
xlabel('GDPP')
ylabel('GDPP')
title('Kmeans Clustering')
```

```
%2. Second Kmeans experiment (remaining features)

%{
Features used: - Child mortality
               - Income
               -  Life expectancy
               - Fertility
               - GDPP
%}

%New frame with the above 5 features
new_features= [normalized_features(:,1),normalized_features(:,5),normalized_features(:,7:9)];

%ELBOW to determine the number of clusters
final = transpose(new_features);
%elbow
% Compute J_m and plot J_m versus m
[l,N]= size(final);
nruns=10;
m_min=2;
m_max=10;
J_m=[];
for m=m_min:m_max
    J_temp_min=inf;
    for t=1:nruns
        rand('seed',100*t)
        theta_ini=rand(l,m);
        [theta,bel,J]= k_means(final,theta_ini);
        if(J_temp_min>J)
            J_temp_min=J;
        end
    end
    J_m=[J_m J_temp_min];
end
m=m_min:m_max;
figure(2), plot(m,J_m)

K-means (3 clusters)
k = 3;

% we can also use one of the functions 'rand_init' and 'distant_init'
% to initialize the thetas
theta_ini = rand_data_init(final, k);

%PLOT Results
%2D plots
colors = [0 0 1; 1 0 0; 0 1 0];

figure
gscatter(final(1,:), final(2,:), bel,colors)
xlabel('Child Mortality')
ylabel('Income')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed','Location','northeast')

figure
gscatter(final(1,:), final(3,:), bel,colors)
xlabel('Child Mortality')
ylabel('Life Expectancy')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed')

figure
gscatter(final(1,:), final(4,:), bel,colors)
xlabel('Child Mortality')
ylabel('Fertility')
```

```
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed')

figure
gscatter(final(1,:), final(5,:), bel,colors)
xlabel('Child Mortality')
ylabel('GDPP')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed','Location','northeast')

figure
gscatter(final(2,:), final(3,:), bel,colors)
xlabel('Income')
ylabel('Life Expectancy')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed','Location','southeast')

figure
gscatter(final(2,:), final(4,:), bel,colors)
xlabel('Income')
ylabel('Fertility')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed','Location','northeast')

figure
gscatter(final(2,:), final(5,:), bel,colors)
xlabel('Income')
ylabel('GDPP')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed')

figure
gscatter(final(3,:), final(4,:), bel,colors)
xlabel('Life Expectancy')
ylabel('Fertility')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed')

figure
gscatter(final(3,:), final(5,:), bel,colors)
xlabel('Life Expectancy')
ylabel('GDPP')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed','Location','northwest')

figure
gscatter(final(4,:), final(5,:), bel,colors)
xlabel('Fertility')
ylabel('GDPP')
title('Kmeans Clustering')
legend('Medium Developed','Less Developed','Highly Developed','Location','northeast')

%3D plots
colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(1,:), final(2,:),final(3,:),'CData',bel)
xlabel('Child mortality')
ylabel('Income')
zlabel('Life expectancy')
title('K-means Clustering')
%legend('Medium Developed', 'Less Developed', 'Highly Developed', 'Location', 'best')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(1,:), final(2,:),final(4,:),'CData',bel)
xlabel('Child mortality')
ylabel('Income')
zlabel('Fertility')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(1,:), final(2,:),final(5,:),'CData',bel)
xlabel('Child mortality')
ylabel('Income')
zlabel('GDPP')
title('K-means Clustering')
```

```
colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(1,:), final(3,:),final(4,:),'CData',bel)
xlabel('Child mortality')
ylabel('Life expectancy')
zlabel('Fertility')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(1,:), final(3,:),final(5,:),'CData',bel)
xlabel('Child mortality')
ylabel('Life expectancy')
zlabel('GDPP')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(1,:), final(4,:),final(5,:),'CData',bel)
xlabel('Child mortality')
ylabel('Fertility')
zlabel('GDPP')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(2,:), final(3,:),final(4,:),'CData',bel)
xlabel('Income')
ylabel('Life expectancy')
zlabel('Fertility')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(2,:), final(3,:),final(5,:),'CData',bel)
xlabel('Income')
ylabel('Life expectancy')
zlabel('GDPP')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(2,:), final(4,:),final(5,:),'CData',bel)
xlabel('Income')
ylabel('Fertility')
zlabel('GDPP')
title('K-means Clustering')

colormap([1 0 0; 0 1 0; 0 0 1;])
scatter3(final(3,:), final(4,:),final(5,:),'CData',bel)
xlabel('Life expectancy')
ylabel('Fertility')
zlabel('GDPP')
title('K-means Clustering')
```

```
%3. Kmedians experiment
%Features used: - Child mortality
%               - Income
%               -  Life expectancy
%               - Fertility
%               - GDPP

%%Elbow
% Compute J_m and plot J_m versus m
[l,N]= size(final);
nruns=10;
m_min=2;
m_max=10;
J_m=[];
for m=m_min:m_max
    J_temp_min=inf;
    for t=1:nruns
        rand('seed',100*t)
        theta_ini=rand(l,m);
        [theta,bel,J]= k_medians(final,theta_ini);
        if(J_temp_min>J)
            J_temp_min=J;
        end
    end
    J_m=[J_m J_temp_min];
end
m=m_min:m_max;
figure(2), plot(m,J_m)

%Kmedians
k = 3;
% we can also use the functions 'rand_init' and 'distant_init'
% to initialize the thetas
theta_ini = rand_data_init(final, k);
[theta,bel,J]=k_medians(final,theta_ini);

%Plot Results
1.
gscatter(final(1,:), final(4,:), bel)
xlabel('Child mortality')
ylabel('Fertility')
title('Kmedians')
legend('Medium Developed','Less Developed','Highly Developed','Location','southeast')

colormap([1 0 0; 0 1 0; 0 0 1])
scatter3(final(5,:), final(1,:),final(4,:),'CData',bel)
xlabel('GDPP')
ylabel('child mortality')
zlabel('fertility')
title('Kmedians')

2.
gscatter(final(2,:), final(4,:), bel)
xlabel('Income')
ylabel('Fertility')
title('Kmedians')
legend('Medium Developed','Less Developed','Highly Developed','Location','northeast')

colormap([1 0 0; 0 1 0; 0 0 1])
scatter3(final(5,:), final(2,:),final(4,:),'CData',bel)
xlabel('GDPP')
ylabel('Income')
zlabel('Fertility')
title('Kmedians')
```

```
%4. Kmedoids experiment
%{
Features used: - Child mortality
               - Income
               - Life expectancy
               - Fertility
               - GDPP
%}

%Elbow
% Compute J_m and plot J_m versus m
[l,N]= size(final);
nruns=5;
m_min=2;
m_max=7;
J_m=[];
for m=m_min:m_max
    J_temp_min=inf;
    for t=1:nruns
        rand('seed',100*t)
        theta_ini=rand(l,m);
        [bel,cost,w,a]= k_medoids(final,m,0);
        if(J_temp_min>J)
            J_temp_min=J;
        end
    end
    J_m=[J_m J_temp_min];
end
m=m_min:m_max;
figure(2), plot(m,J_m)

%K-medoids
```

```
k = 3;
[bel,cost,w,a]=k_medoids(final,k,0);

%Plot Results
1.
colors = [0 1 0; 0 0 1; 1 0 0];
gscatter(final(1,:), final(4,:), bel,colors)
xlabel('Child mortality')
ylabel('Fertility')
title('Kmedoids')
legend('Less Developed','Highly Developed','Medium Developed','Location','southeast')

colormap([0 1 0; 0 0 1; 1 0 0])
scatter3(final(5,:), final(1,:),final(4,:),'CData',bel)
xlabel('GDPP')
ylabel('Child mortality')
zlabel('Fertility')
title('Kmedoids')

2.
colors = [0 1 0; 0 0 1; 1 0 0];
gscatter(final(2,:), final(4,:), bel,colors)
xlabel('Income')
ylabel('Fertility')
title('Kmedoids')
legend('Less Developed','Highly Developed','Medium Developed','Location','northeast')

colormap([0 1 0; 0 0 1; 1 0 0])
scatter3(final(5,:), final(2,:),final(4,:),'CData',bel)
xlabel('GDPP')
ylabel('Income')
zlabel('Fertility')
title('Kmedoids')
```

```
% Cluster characteristic
% Find the unique cluster labels
clusters = unique(bel);

% Initialize an empty cell array to store the data for each cluster
cluster_data = cell(length(clusters), 1);

% Loop over the clusters
for i = 1:length(clusters)
    % Extract the data for the current cluster
    cluster_data{i} = final(:, bel == clusters(i));
end

cluster_data1 = transpose(cluster_data{1,1})
cluster_data2 = transpose(cluster_data{2,1})
cluster_data3 = transpose(cluster_data{3,1})

% mean of every cluster
mean(cluster_data1)
mean(cluster_data2)
mean(cluster_data3)

% standard deviation of every cluster
std(cluster_data1)
std(cluster_data2)
std(cluster_data3)

% Plot the histogramms for each feature of every cluster
figure(1),histogram(cluster_data{1,1}(1,:)), title('Child mortality')
figure(2),histogram(cluster_data{1,1}(2,:)),title('Income')
figure(3),histogram(cluster_data{1,1}(3,:)),title('Life expectancy')
figure(4),histogram(cluster_data{1,1}(4,:)),title('Fertility')
figure(5),histogram(cluster_data{1,1}(5,:)),title('GDPP')

figure(1),histogram(cluster_data{2,1}(1,:)), title('Child mortality')
figure(2),histogram(cluster_data{2,1}(2,:)),title('Income')
figure(3),histogram(cluster_data{2,1}(3,:)),title('Life expectancy')
figure(4),histogram(cluster_data{2,1}(4,:)),title('Fertility')
figure(5),histogram(cluster_data{2,1}(5,:)),title('GDPP')
figure(1),histogram(cluster_data{3,1}(1,:)), title('Child mortality')
figure(2),histogram(cluster_data{3,1}(2,:)),title('Income')
figure(3),histogram(cluster_data{3,1}(3,:)),title('Life expectancy')
figure(4),histogram(cluster_data{3,1}(4,:)),title('Fertility')
figure(5),histogram(cluster_data{3,1}(5,:)),title('GDPP')
```