Unsupervised Learning

Greta Angolani

May 2024

WHICH ARE THE COUNTRIES TO WATCH OUT FOR?

1 Abstract

The aim of this paper is to retrieve the principal countries which have the lowest socio economic conditions applying unsupervised learning techniques like Principal component analysis and clustering (hierarchical and k-means)

2 Introduction

The Dataset was found on Kaggle Country-data and the aim is to categorise the countries using socio-economic and health factors that determine the overall development of the country.

Research questions:

Can be some cities create a cluster together or not? Are there any cities that does have some similarities considering the variables in the dataset? Among these cities that share some similarities, can we say which are the one that need more help than others?

The Dataset consists in 10 variables which are:

1. **Country**: 167 unique countries
2. Child Mortality: Death of children under 5 years of age per 1000 live births
3. **Exports**: Exports of goods and services per capita. Given as %age of the GDP per capita
4. **Health**: Total health spending per capita. Given as %age of GDP per capita
5. **Imports**: Imports of goods and services per capita. Given as %age of the GDP per capita
6. **Income**: Net income per person
7. **Inflation**: The measurement of the annual growth rate of the Total GDP
8. **Life Expectancy**: The average number of years a new born child would live if the current mortality patterns are to remain the same
9. **Total Fertility**: The number of children that would be born to each woman if the current age-fertility rates remain the same
10. **GDPP**: The GDP per capita. Calculated as the Total GDP divided by the total population

3 Data Pre-Processing

The aim of Data Pre-Processing techniques is to analyse the dataset, in order to better understand which are the variables and their correspondent values. We do this by checking the presence of missing or duplicate values, by plotting their distribution and by plotting the correspondent box-plots, in order to understand if there are any differences or any outliers.

In the dataset there were no missing values neither duplicates, so I have not remove any columns/rows. The absence of duplicates, especially in the country columns, had also allowed me to set the column as a row for better understand the result. Then I had dropped the country column and I had created a new dataset called "country_new". Proceeding plotting their graph and their correspondent box-plot, I have noticed first of all that the variables are in different scale, and second that the distribution of the variables is not uniform, highlighting the presence of some so called "outliers" as we can see in the figure[1]:

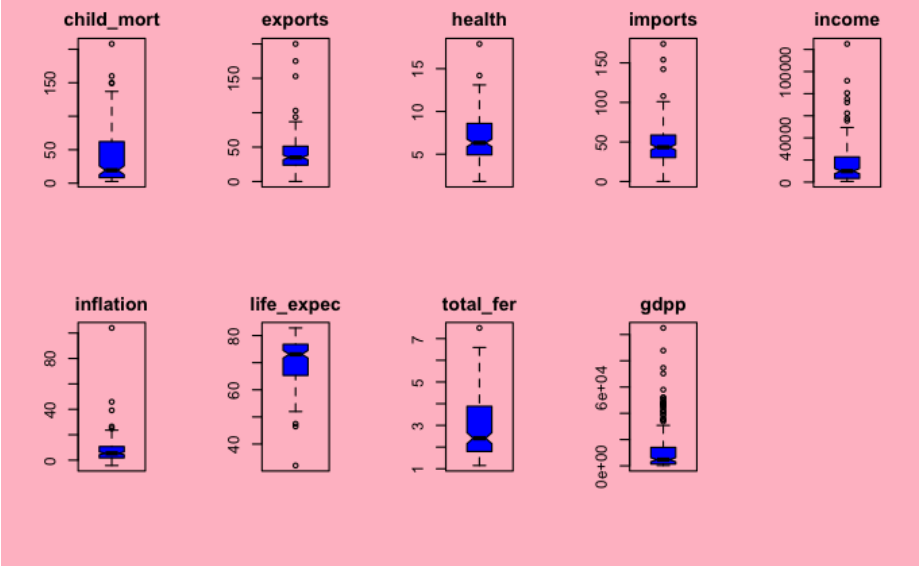


Figure 1: Boxplot

Even if some outliers are present in the dataset I have decided to keep them, cause otherwise the final clusters retrieved were not close to the real values, and some of the cities that actually do not need help were clustered with cities in which the socio-economic life conditions were not so good. Regarding the different distribution of the variables, since for PCA and clustering is important that the variables are distributed among the same scale, I had proceeded with the scaling process. In my dataset the variables are all numerical, so I have just applied the scale function in R. The very last step, before applying PCA,

is to check for correlation among the variables. As we can see in the correlation matrix there are some variables wich present an higher correlation index respect to the others. These variables are:

1. GDPP with Income, with a correlation index of 0.90
2. Total Fertility with Child Mortality, with a correlation index of 0.85
3. Life expectancy with Child Mortality, with a correlation index of -0.89
4. Imports with Exports, with a correlation index of 0.74
5. Life Expectancy and Total Fertility with a correlation index of -0.76

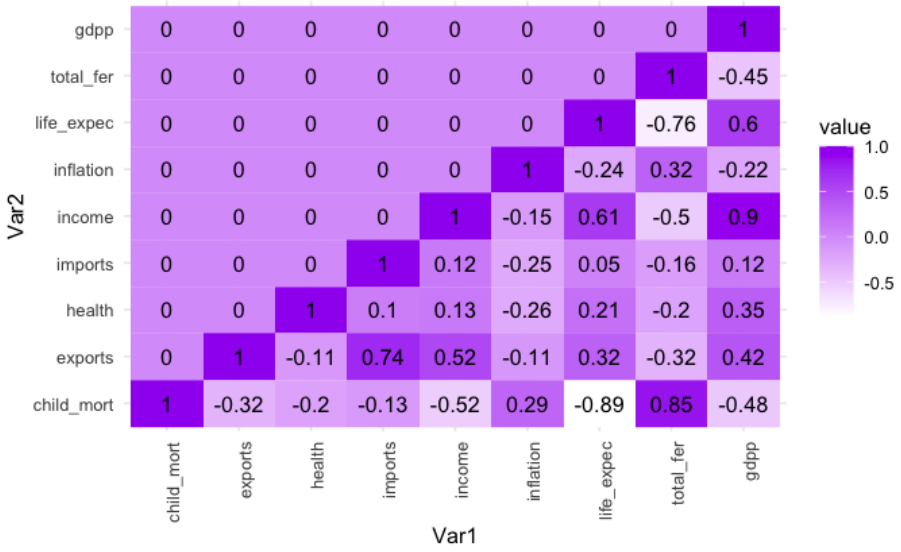


Figure 2: Correlation Matrix

At the beginning I have tried to remove them, but then I realized that the percentage of the total explained variance was lower than keeping all this variables, so I kept them inside the dataset.

4 PCA

The PCA step is very important, cause help us reducing the number of variables and to keep only the most informative ones, especially when we are dealing with very big datasets (which actually is not the case). I have applied the "prcomp" function in R and plotted the result:

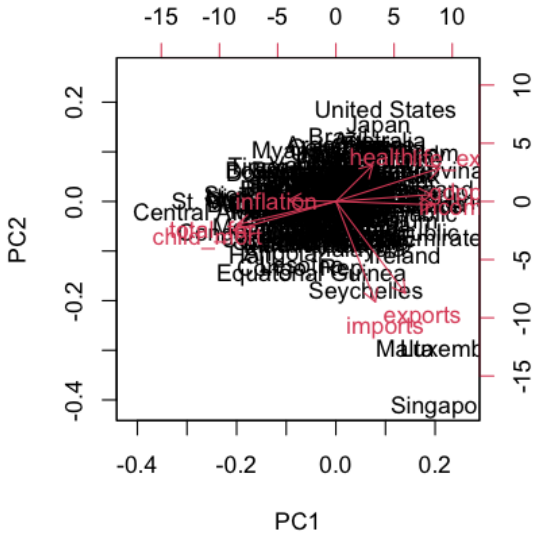


Figure 3: PCA

As we can already see from the graph there are 3 cities which are separated from the others, which are Singapore, Malta and Luxembourg to which correspond higher value of imports and exports.

After I have calculated the cumulative variance ratio: [1] 0.4595174 0.6313337 0.7613762 0.8719079 0.9453100 0.9701523 [7] 0.9827566 0.9925694 1.0000000

As we can see the first 4 principal components can explain the 87,19% of the total variance, which is a very good result considering the total number of components, which precisely is 9. Since we are able with 4 components to explain a great part of the cumulative variance, this means that we can reduce the overall number of dataset without losing any meaningful information for our analysis.

Other steps to understand how many components to keep were analyzed, for example the Kaiser criterion and the scree plot. The kaiser criterion suggest to keep 3 principal components, while the fourth one is very at the limit, while the scree plot suggest the same as the total variance, so 4/5 components.

At the end 4 components were chosen, and I proceed with the analysis using only this.

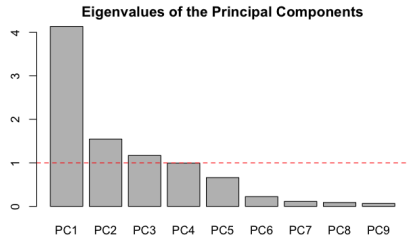


Figure 4: Kaiser Criterion

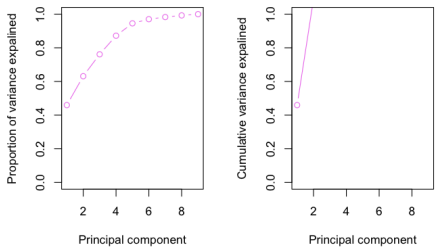
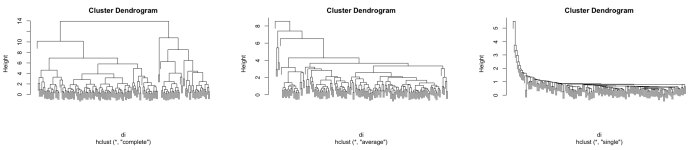


Figure 5: Scree Plot

5 Clustering

For the clustering, first of all I have checked if the variables have higher or lower tendency to cluster; the result, specifically the 99.50% indicate a very high tendency to cluster. This is a very good starting point. Then I had plotted the heatmap, to have a better look among the variables and the countries. After that I had proceed with the cluster and I had applied 3 methods: the complete cluster, the average cluster and the single cluster.



(a) Complete (b) Average (c) Single

Figure 6: Hierarchical Clustering

Then I had plotted the hierarchical cluster, using the euclidean method. The hierarchical cluster didn't ask for a specific number of cluster at the very beginning because builds a hierarchy of clusters by iteratively merging individual data points into clusters. I have then decided to select 4 clusters, and I have

splitted them as we can see in the figure:

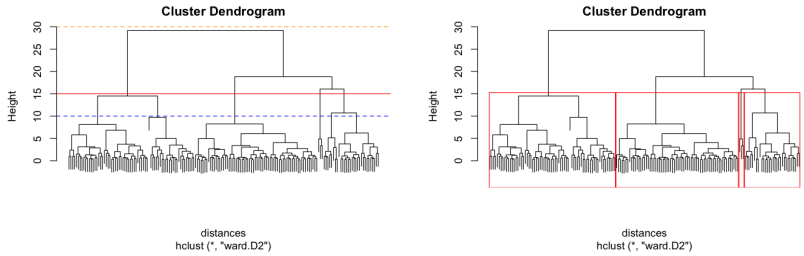


Figure 7: Hierarchical Clustering Dendogram

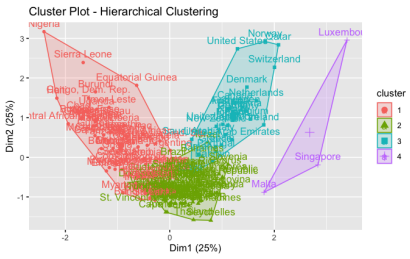
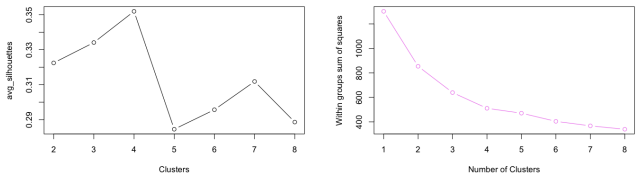


Figure 8: Hierarchical Clustering

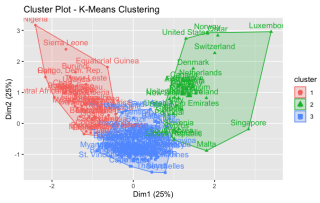
K-means Clustering Opposite to the hierarchical clustering, the K-means works only with a predefined number of cluster. So I have started the analysis by checking the correct number of cluster to use thanks to the wssplot function and, as we can see from the graph, the correct number seems to be 3/4 clusters. To better understand I had also checked the correct number of cluster using the silhouette method, and I had confirmed that the correct number of cluster to use is 4.



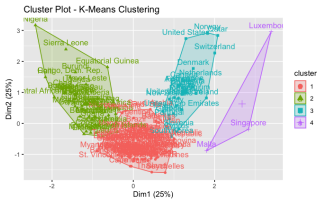
(a) Silhouette Method (b) Elbow Method

I have decided to plot clusters with 3 groups and 4 groups to better understand the difference among them, and appeared that the only difference with 4 clusters instead of 4 is that the 3 richest cities (Malta, Luxembourg and Singa-

pore) are in the 4th cluster instead to be incorporated in the 3rd one as we can see in the figures:



(a) 3 Clusters



(b) 4 Clusters

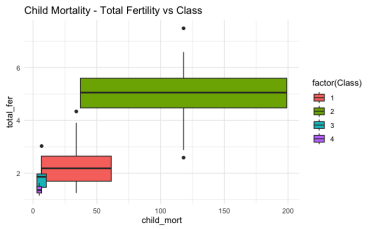
I have chosen to use 4 clusters for my analysis even if for the aim of the analysis 3 cluster would have been appropriate anyway.

At the end I had plotted the box plot considering some important features among the ones present in the dataset, and I had classified the countries belonging to clusters in 4 groups:

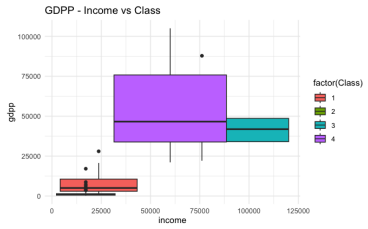
1. Need Help
2. Might Need Help
3. No Help
4. No help at all

Here are some boxplot analysis:

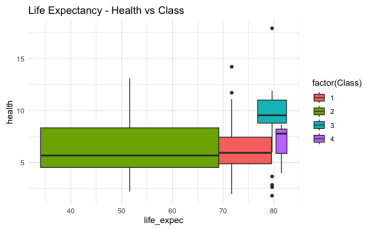
(a) Child Mortality vs Life Expectancy



(b) Child Mortality vs Total Fertility



(a) GDPP vs Income



(b) Life Expectancy vs Health

6 Conclusion

At the end of the analysis I was able to say that countries belonging to cluster 1 were the one that need more help, considering variables as "Life Expectancy" and "Child Mortality", as well as "Income" and the "Gdpp". Belonging to cluster 1 we have this countries:

Afghanistan, Angola, Benin, Botswana, 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, South Africa, Sudan, Tanzania, Timor-Leste, Togo, Uganda, Yemen, Zambia, So at the very end we can say that this analysis had produced the expected result, and now we are able to identify and provide more help to countries which are less developed in terms of socio-economic values.