The World Report

Mikaela Pisani, Marcin Grzechowiak, Jereamy Riggs, Roger Valdez

Group 4

Multivariate Analysis

I. Introduction

The World Report aims to profile countries based on information relating to economic, social, financial, and technological factors.

Dataset:

The dataset is used from GapMinder World, which contains data from several sources. It presents economic, social, financial and technological indicators about countries. The variables considered for this study are Murder Total, Armed Forces Personnel Total, Cell Phones per 100 People, Children per Woman, Suicide Total, Life Expectancy, Corruption Perception Index, Internet Users (% of Population), Child Mortality, Income Per Person, Sex Ratio, Investment (as % of GDP) and Inequality Index (GINI). For more information about the indicators see the appendix of the report. All indicator values come from the year 2016. The only exception to this is Sex Ratio. It is assumed that the sex ratio did not change significantly between 2015 and 2016 and using data from 2015 is a good approximation of the year 2016 (Table 1 and Table 2 in Appendix).

Motivation:

The motivation is focused on comparing countries from the economic, social, and government perspectives. This study can be useful for instance, for a person that wants to move to another country, it might be important to consider these aspects to make the decision. In addition, it could be useful for the World Health Organization when determining which countries it should support in each field. Finally, firms that want to open an office in another country might be interested in this data as careful considerations are made when making such a large business decision.

Scholarly Citations:

One of the perspectives of the project is that of a business looking to expand infrastructure by investigating factors for potential new locations in different countries. The article, “Factors affecting location decisions in international operations – a Delphi study” focuses on factors that guide firms to choose locations while opening a branch in a new

country. The factors in this report can be considered in future decision making when firms are looking to expand into different countries.

On the other hand, country profiling might be useful to recognize places where global health resources are needed the most. The study "Financing of global health: tracking development assistance for health from 1990 to 2007," based their findings on factors that included income, HIV/AIDS, and other diseases. The factors presented here may be considered as complementary factors when the World Health Organization is deciding which countries to fund. As a result, 'The World Report' might be useful in all mentioned contexts.

II. Data Cleaning and Outlier Visualization

It is important to acknowledge the missing values in the dataset in order to perform accurate calculations. In the case of visualization, the ability to view the outliers gives the opportunity to account for their presence and to decide if removal is necessary.

The first step was importing the data from multiple csv files and merging the columns on a single column that each file shared, the country column. Once the merging process was complete, a function was found that would divide the countries by the continent they belonged to in order to create a new column called "continent." Unfortunately, the function did not divide the Americas into North, South, and Central so a *for loop* was created to iterate through the countries column in order to place them into a more specific part of the Americas. The reason for the new column "continent" was to account for missing values in the data. If an NA value was spotted it would insert the median value depending on the continent associated with the country that had the missing value. Since outliers have less of an effect on the median, it was chosen as the replacement for the NA values, rather than the mean.

Outliers are present in the data and boxplots were created in order to visualize which countries were outliers according to the indicator. Outliers were not removed because in this case the outlier values were not the result of an error, but instead were actual data points. If the outliers were removed, the potential to lose meaning and value from the data would increase greatly. Furthermore, so much of the data would have to be removed, causing

additional problems to arise in the future section of analysis. The reason is an outlier would be associated with a country and in order to remove the outlier we would have to delete an entire row. It is believed that those 'outliers' in our data in fact store very important information and should not be removed.

In order to get insights about outliers, three columns were presented on boxplots below (Figure 1). Please find all boxplots for each column in the Appendix.

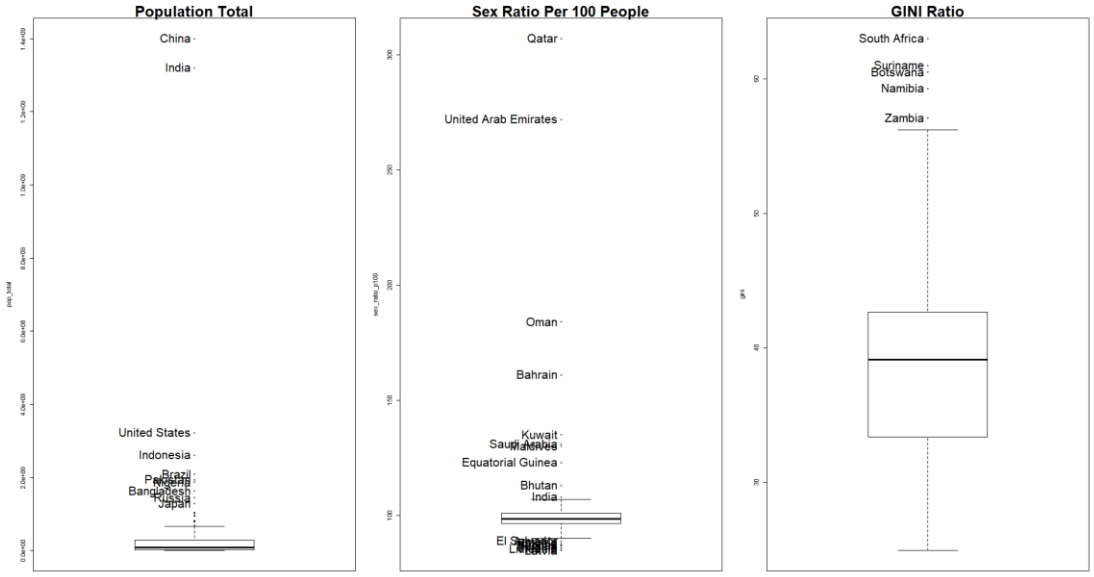


Figure 1. Example of outliers in the data. All graphs are available in the Appendix.

In addition to the univariate boxplots, bivariate boxplots were also created to assist in outlier analysis based on two variables. Income per person and GINI (inequality index) were compared first. It was observed that Qatar, Luxembourg, and Singapore were the highest in income per person and had average levels for GINI while South Africa, Suriname, and Botswana were the highest in GINI and had very low levels of income per person. Further insights can be gathered by looking at three of the bivariate boxplots in Figure 2.

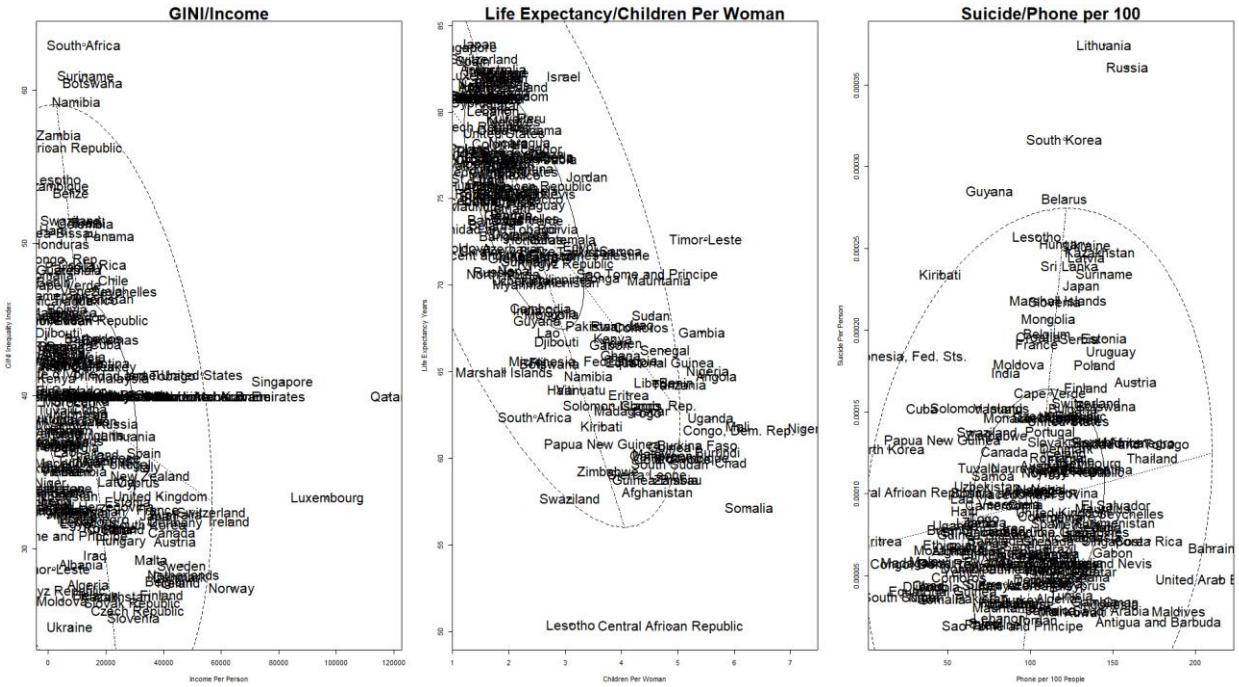


Figure 2. Example of outliers in the data via bivariate boxplots.

III. Dimension Reduction Analysis

The data set contains 14 variables. In order to easily present them in a 2 or 3-dimensional plot, dimension reduction analyses were computed. Two techniques were used: Multidimensional Scaling (MDS) and Principal Component Analysis (PCA).

a. Multidimensional Scaling (MDS)

In order to provide a general insight into the data, all countries were presented in 3-dimensional space. At first glance, clusters between continents can be observed. Countries which belong to the same continent present a similar profile. The most diverse continent is Asia with many outliers. Countries in Asia spread from Europe (the one end) to Africa (the second end). It can be seen that North America and South America are similar to each other (Figure 3).

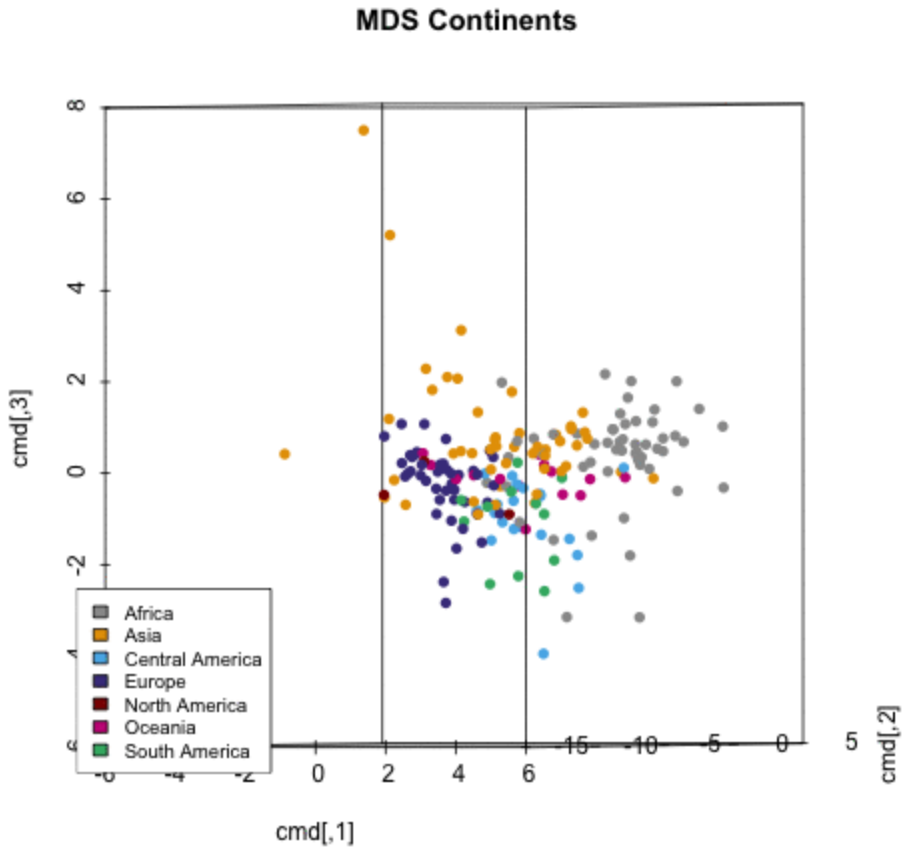


Figure 3. Continents presented in 3D space. (If not displayed correctly, please see a .gif here: https://bit.ly/2DMeYEt)

b. Principal Components Analysis (PCA)

The Principal Component Analysis (PCA) was used in order to provide more insights into the data and visualize it in two-dimensional plots. Three principal components were presented on the plots below and their cumulative proportion of variance is equal to 74%.

Interpretation of PC1, PC2, and PC3 is as follows:

PC1: is highly loaded in variables such as number of phones, life expectancy, corruption index, access to the internet and income.

PC2: is highly loaded in the number of suicides and sex ratio.

PC3: is especially meaningful in the context of inequality.

More information about components can be found in Figure 4 below.

```
Importance of components:
                          Comp.1    Comp.2     Comp.3
Standard deviation     2.2699037 1.1802863 0.93771762
Proportion of Variance 0.5152463 0.1393076 0.08793143
Cumulative Proportion  0.5152463 0.6545539 0.74248529


Loadings:
                Comp.1 Comp.2 Comp.3
PHONES           0.303  0.159  0.290
CHILDREN        -0.383  0.134 -0.221
LIFE EXP         0.398
SUICIDE          0.162 -0.525  0.225
SEX RATIO               0.723
LESS CORRUPTION  0.305
INTERNET         0.411
CHILD MORT.     -0.396
INCOME           0.344  0.317
INEQUALITY      -0.184  0.200  0.887
```

Figure 4. Summary Result for the first three Principal Components

In order to present the PCA results, two graphs are displayed below. Ellipses were added to the graphs to better showcase the concentration(s) of points. The size of each ellipse is largely influenced by outliers present within the data. For an example of this, observe the ellipse for the continent of Asia. There are outliers like Qatar and United Arab Emirates that cause the ellipse to be broadened.

Plot PC1 vs PC2

On the right of the plot with a high value of PC1, are what we consider highly-developed countries (Figure 5). Both Europe and North America can be spotted here. Those continents are above the average in the context of less corruption, life expectancy, internet access, number of phones and income per person.

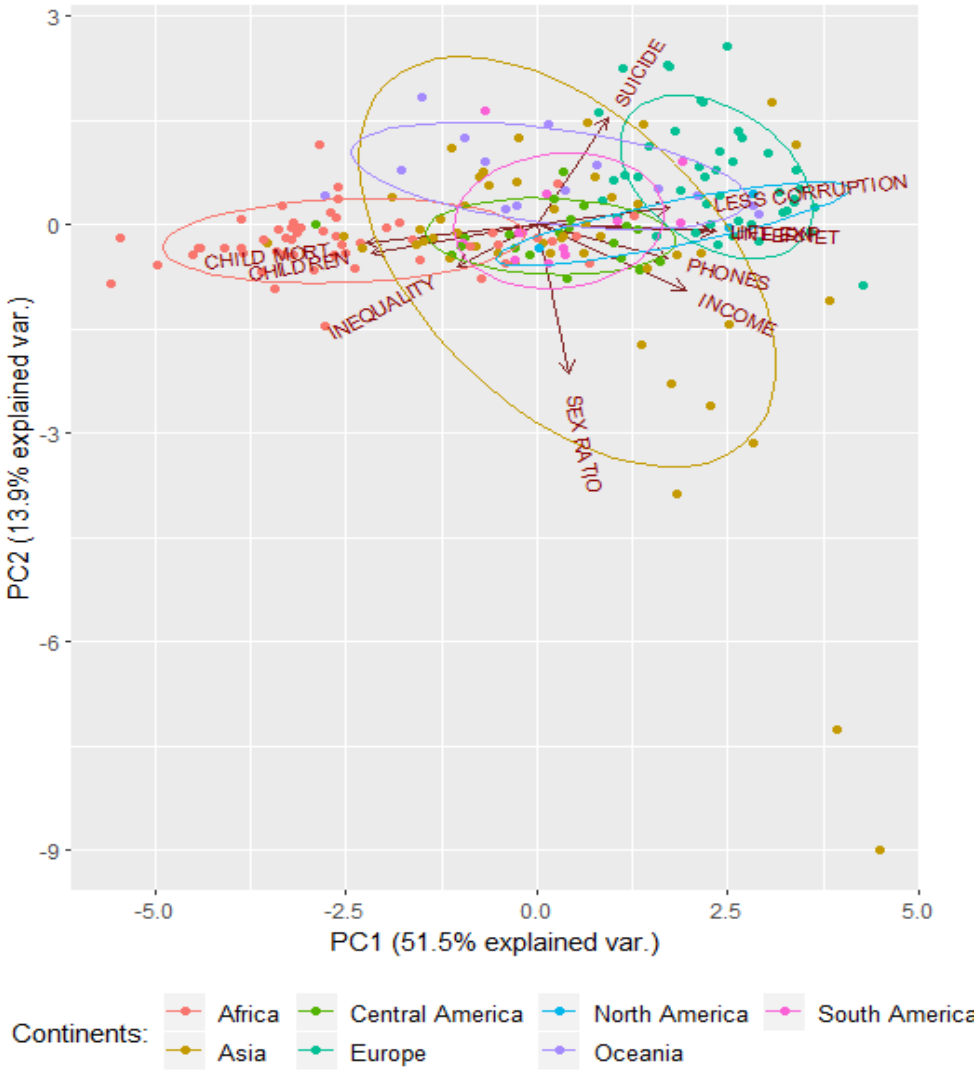


Figure 5. Relation between PC1 and PC2.

On the left side of the plot, with a low value of PC1 are the countries that are less developed. As somewhat expected, Africa can be spotted here. Those countries are above the average in the context of high child mortality, number of children per woman and inequality.

An interesting phenomenon can be seen when looking at Asia. The continent is the most diverse among all others in both directions for PC1 and PC2. Some countries in Asia are highly developed while others are rather poor (PC1). Think about the stark contrast between the economies of Afghanistan, Yemen, Nepal and those of China, Qatar, and Singapore. This helps create an image of how different the countries are within Asia as a whole. Within the

context of PC2, some countries have an 'extreme' value for sex ratio (men outnumber women significantly). The same countries that were mentioned to influence the concentration ellipses, Qatar and the UAE, are excellent examples of countries where the population of men vastly outweighs that of the women. This example proves how important roles play in outliers within this report. If those data were deleted at the beginning, we would lose this information. PC1 and PC2 do not provide many insights into Central America nor South America, since values for these continents are centered around the average or can be found in the middle of the plot.

Plot PC1 vs PC3

The second plot shows that very high inequality exists especially in South America and Africa (Figure 6). The below plots show also that there is a high correlation between variables: number of phones, less corruption, internet access, and income. An additional group of highly correlated variables are child mortality and number of children per woman.

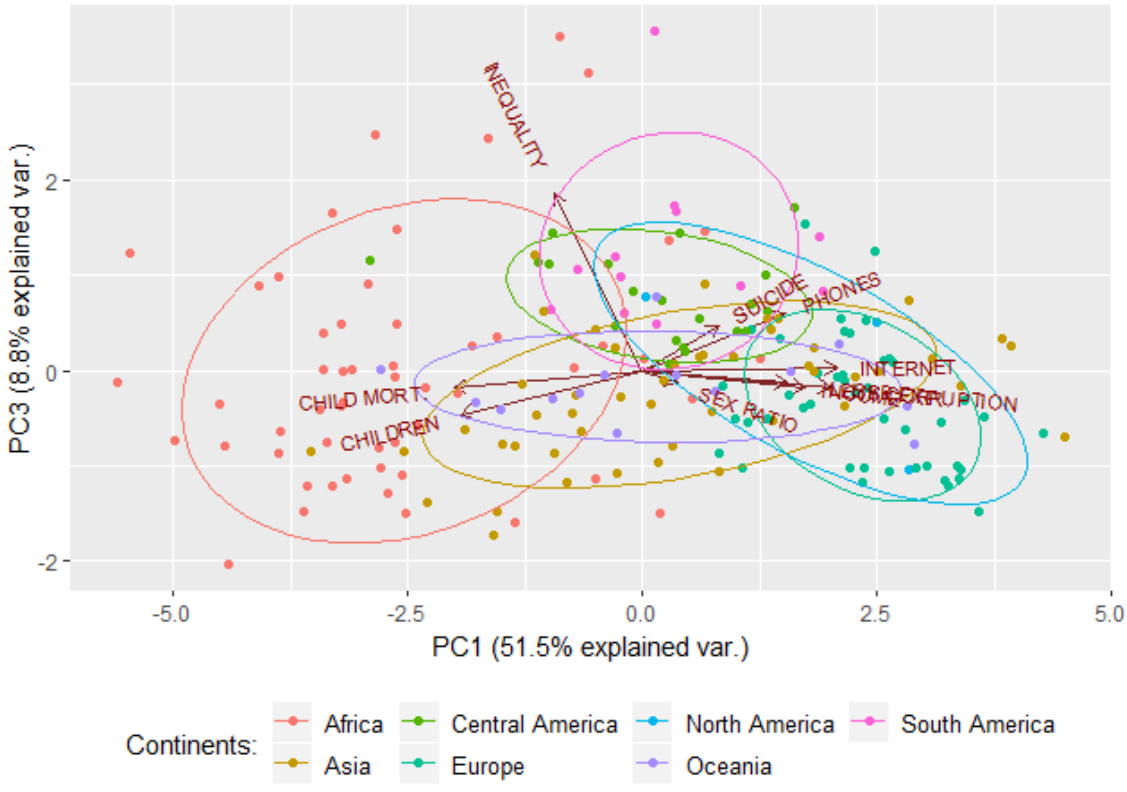


Figure 6. Relation between PC3 and PC1.

PCA on the World Map

In order to show which countries are the highest in each principal component, a World Map is presented. From each component (PC1, PC2, and PC3) the top 15 countries with the highest loadings in each of the group were chosen and plotted on the map (Figure 7). Please be aware that some highlighted countries are difficult to see as they are geographically small.

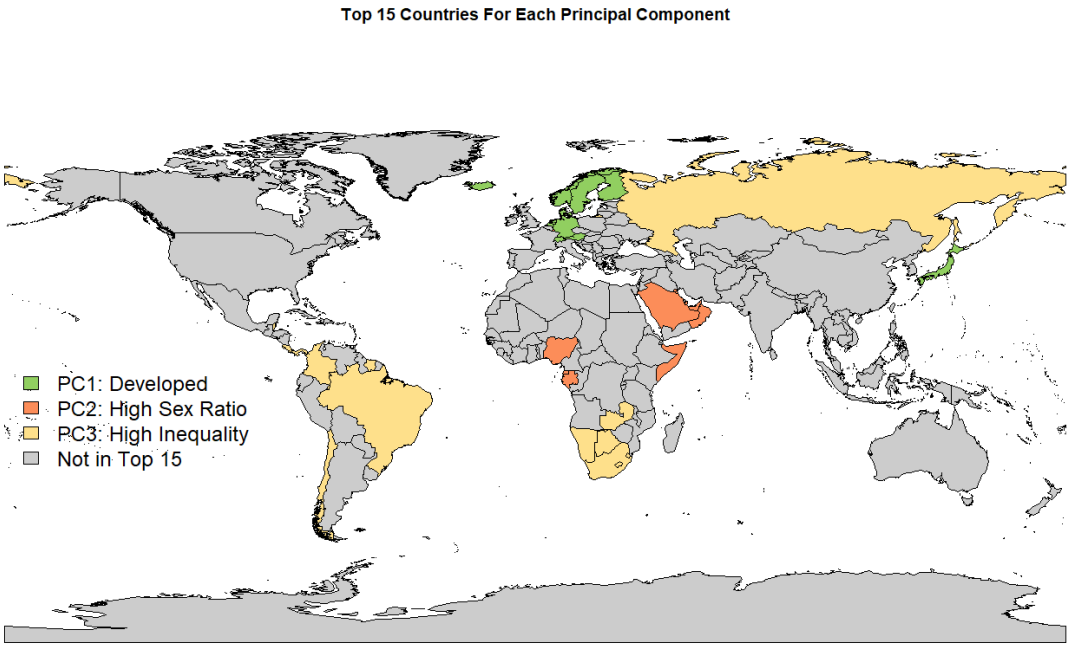


Figure 7. PCA results presented on the world map.

It can be seen that among highly-developed countries are Scandinavian Countries. High sex ratio index is presented in the Middle East and Africa, while countries with high inequality are for instance, South Africa, Brazil and Russia.

Note: from the analysis, columns such as number of murders, number of armed forces, and percentage of investments were excluded. These variables had a low correlation with the rest of the columns and much more dimensions would be needed to explain the data. As such in order to clearly explain the relation in data more dimension would need to be used, what would prevent to present variables on the two-dimensional plots.

IV. Cluster Analysis

For Cluster Analysis, the goal was to discover groups or clusters of observations that are homogeneous and separated from other groups. Three clustering methods were performed: Hierarchical, K-Means, and Model-Based. In the world map plots for K-Means and Model-Based, not all information for countries were available therefore these are displayed as 'No Data' on the plots.

a. Hierarchical Clustering between Continents

At the very beginning the cluster between continents were computed in order to grasp a big picture of the relation between the data. The hierarchical model below explains that the data has three main clusters (Figure 8). The first cluster includes North America, Europe, South America, Central America, and Asia while the second includes Oceania, and the third, Africa. The interesting thing is that Oceania is not clustered with the developed countries like North America and Europe. The reason this is interesting is the fact that Australia and New Zealand, which are part of the Oceania continent, and these countries are considered as highly developed. However, they are not grouped with the developed continents of North America and Europe. This is more than likely due to the reason that Oceania also has many small islands included in the continent that are not as developed.

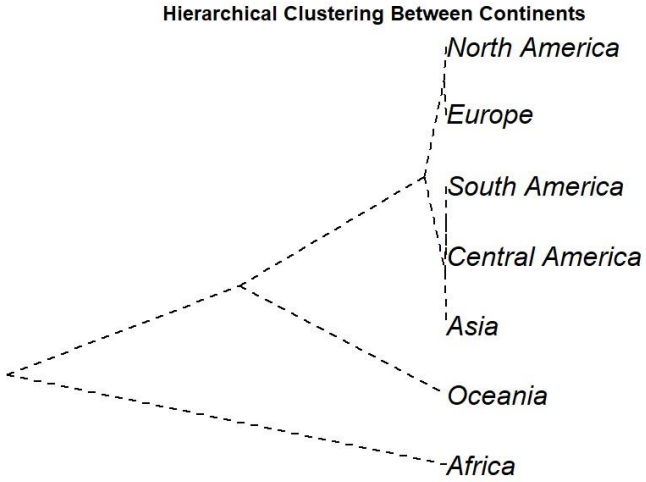


Figure 8. Hierarchical Clusters among columns

b. K-means and Model-Based clustering between Countries

In order to group countries two techniques were used and compared: K-means and Model-Based algorithm. In the case of K-means a scree plot suggests 5 groups. The plot of K-means result is presented below (Figure 9).

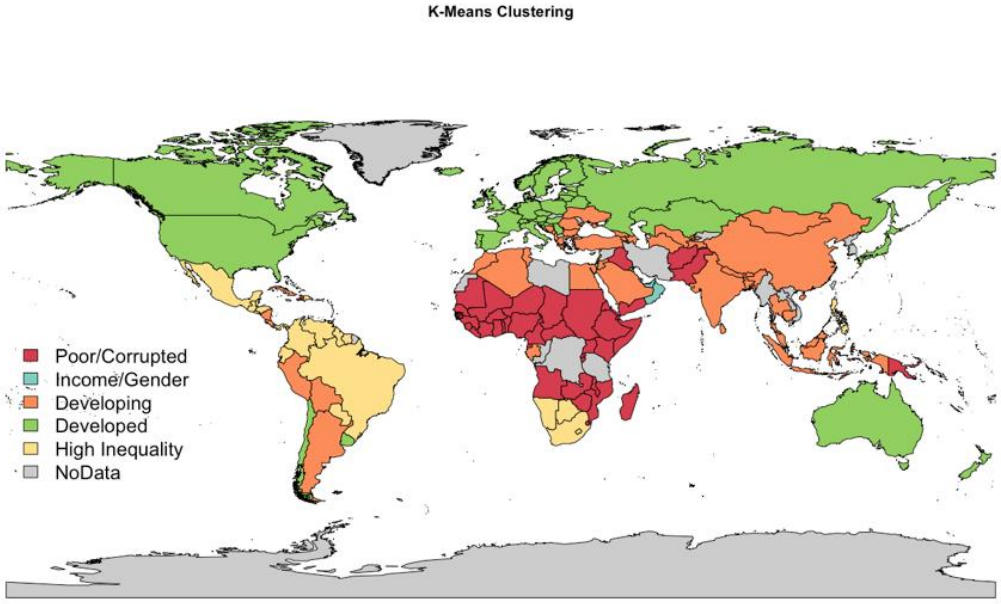


Figure 9. Result of K-means Clustering presented on the map.

The second clusters are based on Model-Based method. The algorithm decided the number of groups automatically, as such 6 groups were created (Figure 10).

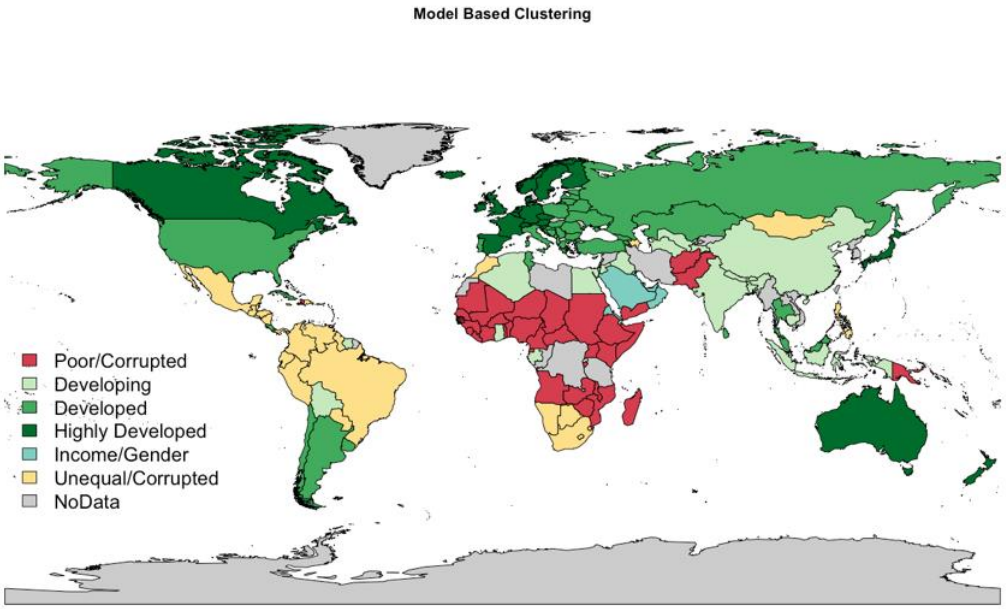


Figure 10. Result of Model-Based Clustering presented on the map.

The developed countries that were observed in the K-means model were split into three separate groups under the Model-based method, appropriately labeled developing, developed, and highly-developed. Highly-developed countries were those that had high principal component 1 loading values. Again, this component was loaded on characteristics like low child mortality, high income, high number of phones per person, and low inequality. The two other groupings involving developed countries were also loaded on these same variables, give or take one or two, however the degree to which they were loaded was less. As such, the interpretation for these groups suggested making tiers out of the development characteristic. Some African countries, China, and India joined the 'developing' group under Model-Based clustering.

The poorer, less-developed countries like those found in Africa remained consistent between the two clustering techniques. While Africa primarily has poor, corrupt, or countries with high inequality, some of the northern countries are observing some development, demonstrating diversity within this continent. The Model-based clusters of Europe showed an interesting characteristic that as you moved East, the level of development in the countries decreased. So, Western Europe was more developed than Eastern Europe.

Additionally, a comparison of the chi-squared test was performed to determine if a dependency between groups and continents existed. The results suggest that Model-based groups are more similar to continents. The chi-squared test is significant in both methods indicating dependency between the groups and continents. However, for K-Means method, X-squared is equal to 236.76, while for Model-Based method, X-squared is equal to 272.71.

V. Exploratory Factor Analysis

Exploratory Factor Analysis looks to link observable variables to unobservable variables via regression modeling. The discovery of the relationship(s) between variables without making any assumptions that such relationships may exist is crucial when defining what factors, and how many, may best describe the data.

In order to find the number of factors, EFA was performed starting with 1 factor, increasing the number of factors until getting a value for RMSE lower than 0.05. It was concluded that the optimal number of factors is four with an RMSE value of 0.04. Performing EFA with 4 factors, the loadings are:

```
Loadings:
                Factor1 Factor2 Factor3 Factor4
MURDER                   0.766
ARMED
PHONES           0.609
CHILDREN        -0.895
LIFE EXP.        0.902
SUICIDE                          0.977
SEX RATIO                                0.553
LESS CORRUPTION  0.493
INTERNET         0.825
CHILD MORT.     -0.926
INCOME           0.558                   0.778
INVESTMENT
INEQUALITY               0.664

               Factor1 Factor2 Factor3 Factor4
SS loadings      4.213   1.225   1.200   1.200
Proportion Var   0.324   0.094   0.092   0.092
Cumulative Var   0.324   0.418   0.511   0.603
```

Figure 11. Result of the EFA (loadings).

From the loadings it can be interpreted:

Factor 1 has high life expectancy, internet access, balanced income per person, it is low in child mortality and children per woman. For these reasons, Factor 1 is believed to represent the level of development of the country.

Factor 2 represents the level of inequality because of the high loadings in murder and GINI.

Factor 3 represents suicide.

Factor 4 represents the level of income related with the amount of men and women that the country has.

In order to visualize these four factors graphically, the scores of the top 15 were taken which allowed for the creation of four groups. Each group is a depiction of the factor with the most relevance, so for example group 1 was high in factor 1 which indicated developed countries.

The groups of countries are named as follows:

Factor 1: Developed

Factor 2: High Inequality

Factor 3: Suicide

Factor 4: Gender/Income

These can be visualized in the following graph below (Figure 12).

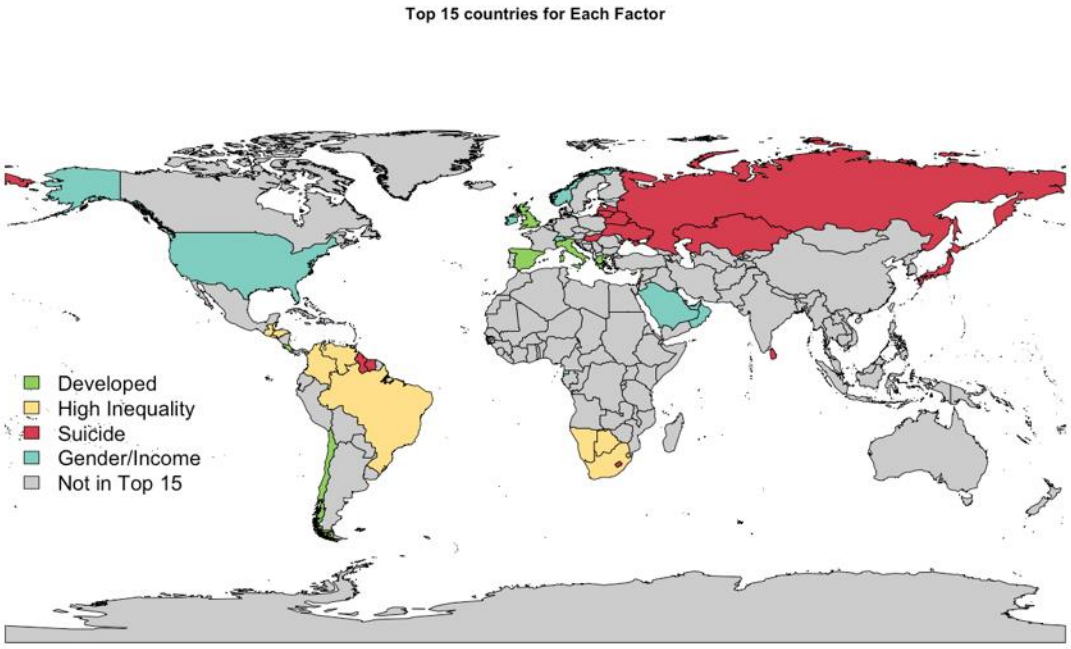


Figure 12. Result of the EFA presented on the Map.

Note: There are some countries such as Singapore or Qatar that are in the groups but are too small to show in the map.

Some notable countries in group 1 are: UK, Singapore, Liechtenstein, and Italy. In the case of group 2, South Africa, Venezuela, El Salvador, and Honduras are notable. For group 3, Japan, Russia, South Korea, and Latvia. Lastly, in group 4 the countries of Qatar, United Arab Emirates, and Norway were notable. Additionally, all list of countries can be found in the Appendix on Figure 1.

VI. Confirmatory Factor Analysis

CFA was performed with the intent of validating the EFA model in section V. Some attempts at confirmatory factor analysis were made using the 'sem' library, yet no results could be reached therefore the *lavaan* library was used.

During the first method the 'sem' library was used with 4 factors, however within the result, the following error was displayed: *"Coefficient covariances cannot be computed".*

The second attempt used the *lavaan* library, again with the 4 factors of the original data. The following error was displayed: "*some observed variances are (at least) a factor 1000 times larger than others; use varTable(fit)".*

Based on the error of using varTable(fit), this method was performed to investigate the problem and it could be seen that the variances for the variables were very high. It was decided that scaling the data might solve this issue.

The third attempt used a 'scale' of the data based on z-score, and the model did find a solution. The results using scaled data were: GFI: 0.799, AGFI: 0.651, SRMR: 0.070

It was decided that a different scaling method could be used so the log function was implemented to scale the data and a different set of results were produced. The results using logarithmic scaled data were: GFI: 0.820, AGFI: 0.688, SRMR: 0.069

Assuming an SRMR less than 0.05 is acceptable, and GFI/AGFI greater than 0.95, the data does not support the model (the model is NOT confirmed).

VII. Conclusion

Because of the big differences between countries dataset contains many outliers which might affect some of the result of analysis. However, if outliers were deleted, the analysis would not have made sense because it is believed that those outliers present meaningful information. Most of the analysis confirms what we know about countries and continents, however, some analysis provided very interesting information. One interesting result is that Oceania, which included well developed countries such as Australia and New

Zealand, were not classified in the same group as highly developed countries in North America and Europe. It might be assumed that small islands in Oceania pull down the entire continent. Another interesting result comes from comparing two grouping methods between each other: K-means and Model-Based method. The latter distinguishes developed countries more precisely and divides them into three groups, while K-means group developed countries into only one category.

The project was based on different indicators pulled from the GapMinder website. In order to improve analysis more economical, social, financial and technological factors might be added to the models presented in the report in order to profile countries more precisely. The project is available online, in order to contribute in the further analysis please visit: https://github.com/grzechowiak/Multivariate-Analysis-Project.

Based on the findings in 'The World Report' the World Health Organization might look towards providing aid to not only Africa, but also Oceania. From the business as well as personal perspective, the best place for a living location in a new country would be in Western Europe and Scandinavia due to the high development.

VIII. Appendix

Table 1. Indicator used in the analysis and their source.

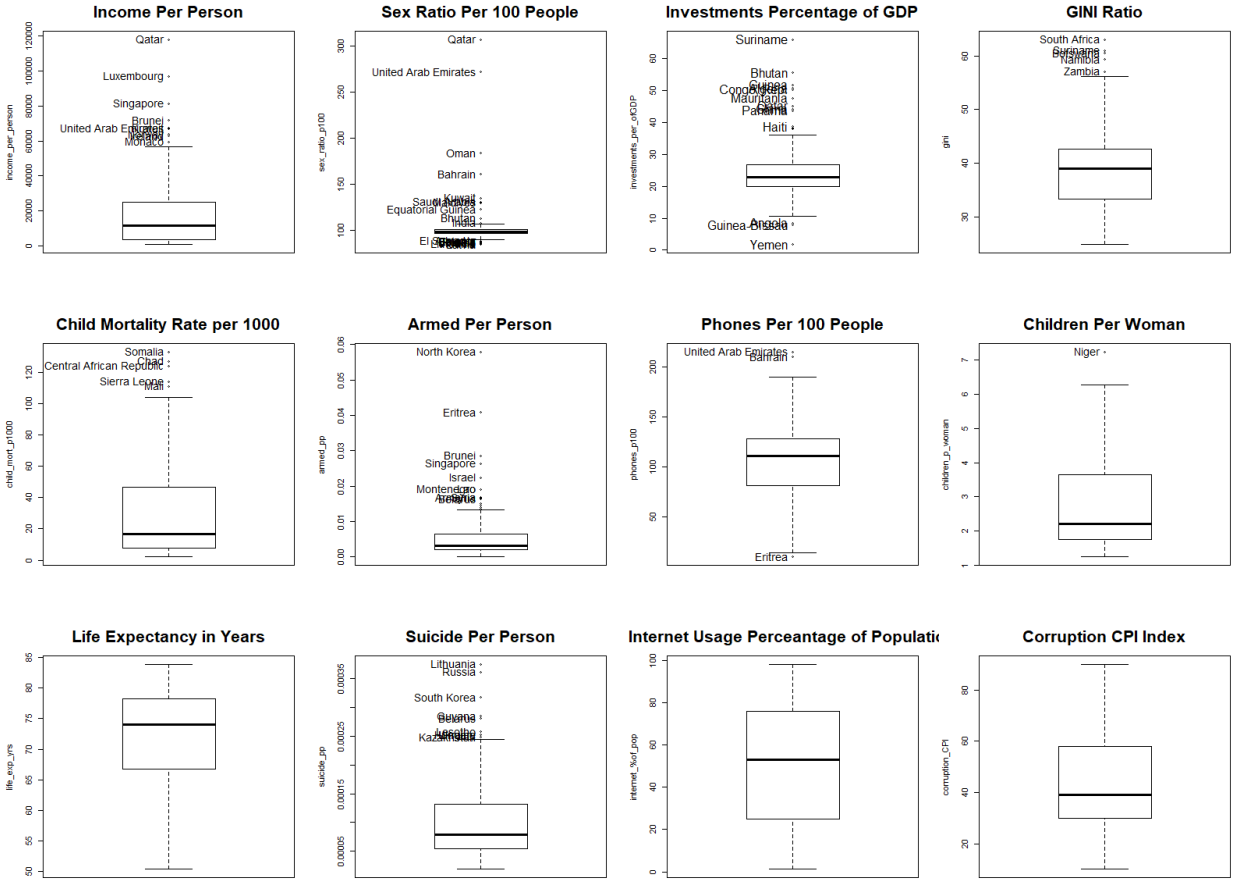
Indicator	Source
Murder total	WHO Global Burden of Disease
Armed forces personnel total	World Development Indicators
Cell Phones per 100 people	World Bank
Fertility rate total	Various Sources
Suicide total	WHO Global Burden of Disease
Life expectancy	Various Sources
Corruption Perception Index	Transparency International
Internet users (% of pop)	World Bank
Child Mortality	Various Sources
Income per person	Various Sources
Sex Ratio	UN Population Division
Investment (as % of GDP)	World Bank
Inequality Index (GINI)	World Bank

Table 2. Explanation of factors used in analysis.

Name	Shortcut	Meaning	Units
Murder per person	murder_pp	Calculated from the total divided by the population	per person
Armed forces personnel total	armed_pp	Calculated from the total divided by the population	total
Cell Phones per 100 people	phones_p100	Mobile cellular telephone subscription	100 people
Fertility rate total	children_p_woman	Children per woman total fertility	total
Suicide total	suicide_pp	Total number of estimated deaths from self-inflicted injury.	total
Life expectancy	life_exp_yrs	Average number of years a child would live.	year
Corruption Perception Index	corruption_CPI	International score of perception of corruption. Higher values indicates less corruption.	0-100
Internet users (% of pop)	internet_%of_pop	Percentage of individuals using internet	percentage
Child Mortality	child_mort_p1000	Death of children under 5 per 1000 born.	100 born
Income per person	income_per_person	GDP/Capita, inflation adjusted $	Per person
Sex Ratio*	sex_ratio_p100	Male/Female per 100 among all age groups	Per 100
Investment (as % of GDP)	investments_per_ofGDP	Gross capital formation. Includes fixed assets plus net changes.	percentage
Inequality Index (GINI)	gini	Gini shows income inequality in a society. Higher is more inequality	0-100

*Sex Ratio - is the only variable which data comes from 2015.

Table 3. All data presented on the boxplots.



Note: please zoom in the plot in order to investigate outliers.

Figure 1. Full list of Top 15 countries for EFA

Group 1: United Kingdom, Chile, Brazil, El Salvador, Liechtenstein, Italy, Malta, Lebanon, Andorra, Singapore, Costa Rica, Greece, Cyprus, Spain, Colombia

Group 2: Bahamas, Trinidad and Tobago, Jamaica, Belize, Botswana, Swaziland, Namibia, Honduras, Brazil, Guatemala, Colombia , Venezuela, Lesotho, El Salvador, South Africa

Group 3: Japan, Marshall Islands Sri Lanka, Suriname, Latvia, Ukraine, Hungary, Kazakhstan, Kiribati, Belarus, Guyana, Lesotho, South Korea, Russia, Lithuania

Group 4: Oman, Bahrain, Equatorial Guinea, Switzerland, United States, Monaco, Saudi Arabia, Norway, Ireland, Kuwait, Brunei, Singapore, United Arab, Emirates Luxembourg, Qatar

IX. References

B.L. MacCarthy, W. Atthirawong, (2003) "Factors affecting location decisions in international operations – a Delphi study", International Journal of Operations & Production Management, Vol. 23 Issue: 7, pp.794-818, https://doi.org/10.1108/01443570310481568

Nirmala Ravishankar,Paul Gubbins,Rebecca J Cooley,Katherine Leach-Kemon,Catherine M Michaud,Dean T Jamison,Christopher JL Murray, (2009) “Financing of global health: tracking development assistance for health from 1990 to 2007”, The Lancet, Vol. 373: Issue: 9681, pp. 2113-2124, https://www.sciencedirect.com/science/article/pii/S0140673609608813