INTRODUCTION TO DATA MINING FOR BUSINESS INTELLIGENCE

WORLD HAPPINESS REPORT

Sofia Gervasoni, Anastasija Babarikina, Benedikt Gerber

(for the R Script please take a look at the R-files attached)

Introduction

The aim of our analysis is to find out how the factors of happiness and freedom of particular countries are affecting the well-being of inhabitants. In order to do so we started from two different data sets that were available on Kaggle.

World happiness Report (2016), our first data set, gives us information about how economic factors (GDP), health, family, trust towards governments, generosity and life expectancy influence the happiness score. In this data set we have information about 157 countries within 13 variables.

The human freedom index data set, our second data set, contains information from 2016 to 2018 about freedom of 167 countries within 127 variables. The human freedom can be distinguished from personal (rule of law, freedom of expression, religion, security and safety, freedom of movement, freedom of association, identity and relationship) and economic (legal, trade, money, government and regulation) freedom. The different variables are summarized in three main scores which are personal freedom score, economic freedom score and human freedom score with values from 0 for the lowest degree of freedom to 10 being the maximum degree of freedom.

We merged the two data sets in order to draw relations between happiness and freedom. We obtained a data set with 137 countries (the intersection of both data sets) and 13 variables.

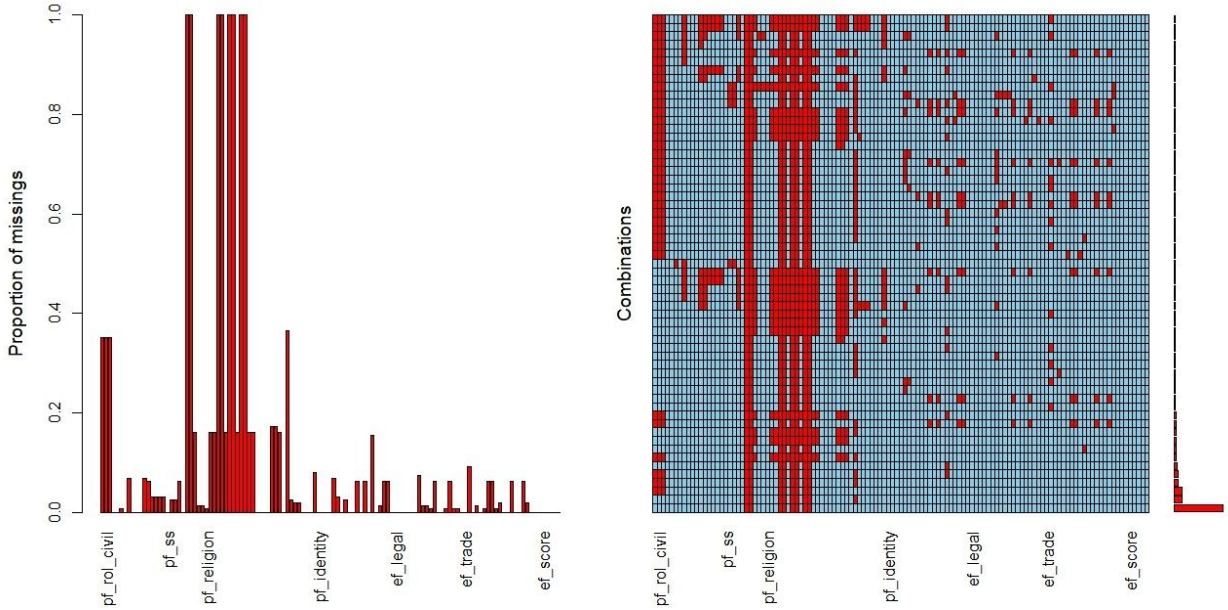
Why did we take only three variables from the data set "Freedom"?

As we have mentioned before, we checked all the variables from the given data set "Freedom" and we found out that all of them could be summarized as the three independed variables (pf_score, ef_score and hf_score), and here is a proof of this.

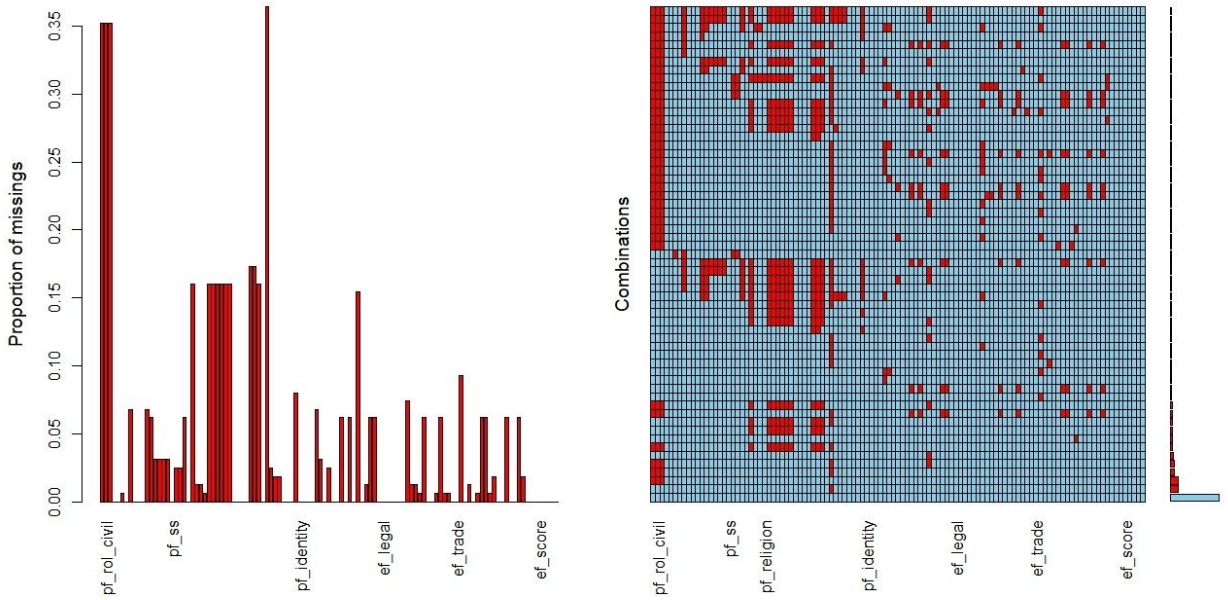
This data set had information about different years, but we were analyzing only the year of 2016, so that we kept only the first 162 rows by creating a new data set.

In order to make it possible to analyze, we had, firstly, to solve the problems with the missing values. Our team discovered that the given dataset had 2081 missing values and

most of them were summarized in 7 columns, so we dropped them off (they had only missing values).



After dropping these 7 columns, we still had 785 missing values located in different columns, distributed as shown in the following graph.



So we decided to use the command “knnImputation”, which helped us to easily attribute multiple missing values taking into consideration the correlation structure of the data.

Once the problem of missing data was solved, we started to study the correlation between the variables.

Particularly, we studied the correlation between the summary variables (the scores) and all the other variables, that are themself a summary of different subcategories.

First of all, we checked the correlation between “Personal Freedom” variables. They are grouped in 7 main categories, that are:

- “rol” - refers to the rule of law in the country (subcategories: procedural, civil and criminal)
- “ss” - refers to social security specifically for women and society overall (subcategories: disappearances, homicide, inheritance ...)
- “movement” - refers to the freedom to move (subcategories: domestic, foreign, women)
- “religion” - refers to freedom of religion (subcategories: estop, restrictions, harassment)
- “association” - refers to liberty of creating associations (subcategories: association, assembly, political, professional, sport)
- “expression” - refers to the possibility of free expression of opinion (subcategories: killed, jailed, influence, control, cable, newspapers, internet)
- “identity” - refers to to power to choose the social identity (subcategories: legal, marriage, divorce, male, female, identity divorce)

For each category we calculated the correlation between the variable that gives a summary for each category and the values for each subcategory. Our team obtained a highly positive correlation for all the variables, which means that each subcategory can be summarized in the variable related to the category. After that, we calculated the correlation between the “pf_score” (value that summarizes personal freedom) and the different categories. Also in this case we obtained a highly positive correlation, so the variable “pf_score” is enough to explain all the variables that refer to personal freedom.

We did the same with “Economic Freedom” (ef_score). In this case we have 5 variables:

- “government” - subcategories: consumption, transfers, enterprises, tax income and tax payroll
- “legal” - subcategories: judicial, court, protection, military, integrity, enforcement, restrictions, police, crime, gender
- “money” - subcategories: growth, sd, inflation, currency
- “trade” - subcategories: tariffs, non tariff,regulatory, black trade, foreign movements, capital movement, visit movement
- “regulation” - subcategories: credit, labor, business

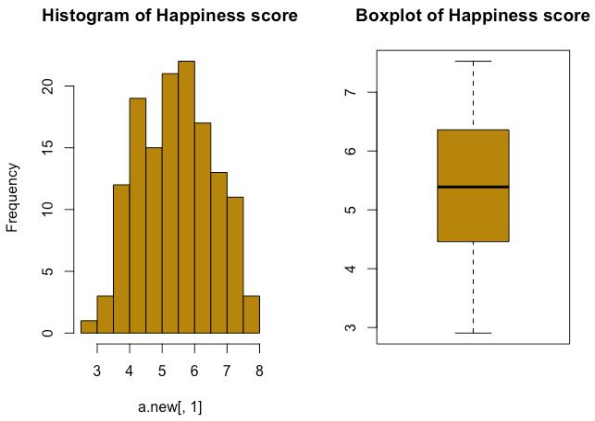
After all it was clear (thanks to the strong correlation) that all these categories/subcategories can be summarized in “ef_score”. Moreover, there is a strong correlation between “ef_score” & “hf_score” and “pf_score” & “hf_score”, so that both groups can be summarized in “hf_score” (although we prefered keep all the three variables).

Descriptive analysis

All the variables in the data set are numerical, there are no missing values.

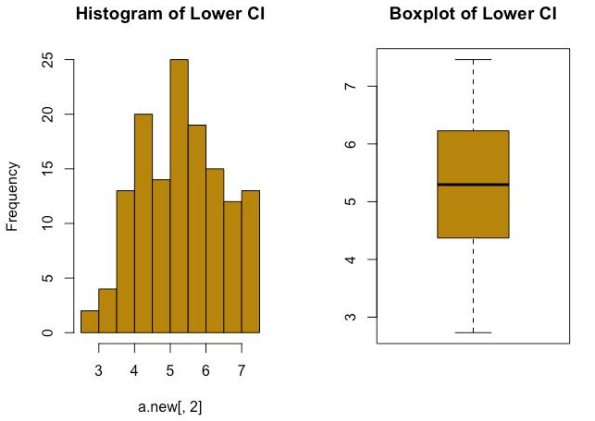
Happiness Score

The first variable “Happiness score” (A metric measured in 2016 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest") is symmetrically distributed as it is shown in the histogram (and boxplot) and which is also confirmed in the fact that the mean (5.43) and the median (5.39) are close to each other and are located between the first and third quantile. All the values are ranging from 2.91 to 7.53.



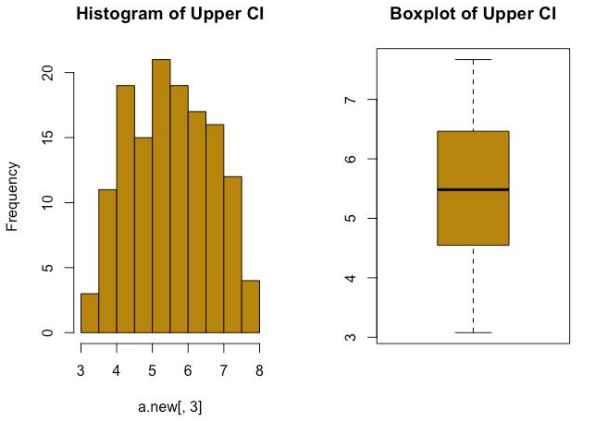
Lower Confidence Interval

The second variable “Lower Confidence Interval” (Lower Confidence Interval of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (5.33) and the median (5.29) are close to each other and are located between the first and third quantile. All the values are ranging from 2.73 to 7.46.



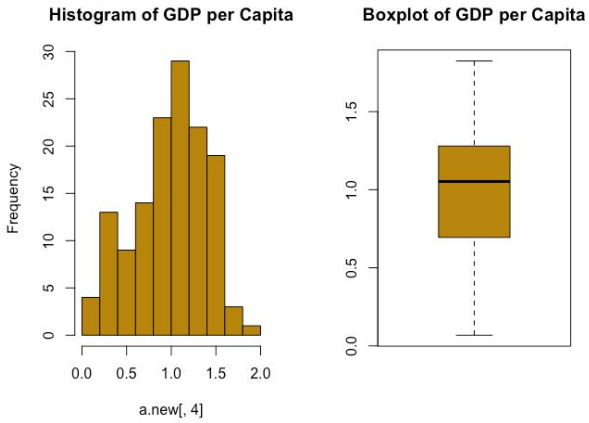
Upper Confidence Interval

The third variable “Upper Confidence Interval” (Upper Confidence Interval of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (5.52) and the median (5.48) are close to each other and are located between the first and third quantile. All the values are ranging from 3.08 to 7.67.



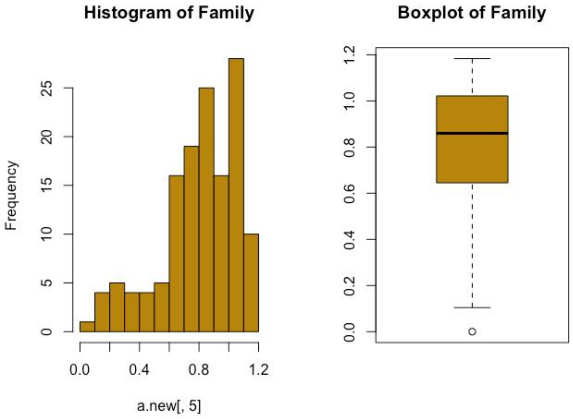
Economy (GDP per Capita)

The fourth variable “GDP per capita” (The extent to which GDP contributes to the calculation of the Happiness Score) is more or less symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (0.98) and the median (1.05) are close to each other and mean is closer to median than the first and third quantile. All the values are ranging from 0.07 to 1.82.



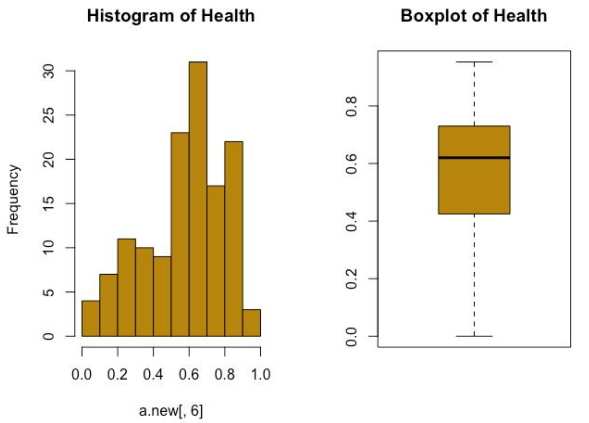
Family

The fifth variable “Family” (The extent to which Family contributes to the calculation of the Happiness Score), judging from the histogram seems to be not symmetrically distributed, but when it comes to mean (0.80) and median (0.86), which are very close to each other, we can see that the variable is actually symmetrically distributed. What is more, the mean is closer to median than the first and third quantile. All the values are ranging from 0 to 1.18.



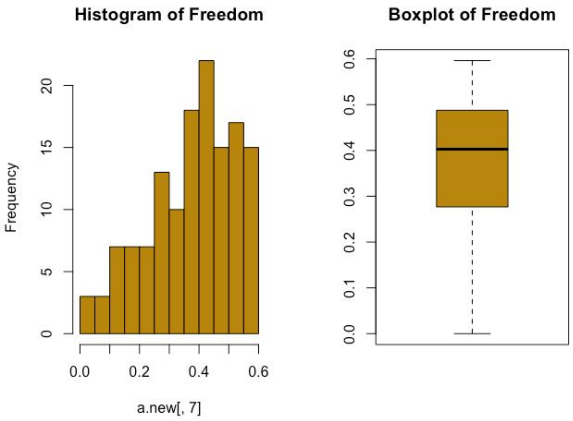
Health (Life Expectancy)

The sixth variable “Health (Life Expectancy)” (The extent to which Life expectancy contributed to the calculation of the Happiness Score), judging from the histogram seems to be not symmetrically distributed, but when it comes to mean (0.57) and median (0.62), which are very close to each other, we can see that the variable is actually symmetrically distributed. What is more, the mean is closer to median than the first and third quantile. All the values are ranging from 0 to 0.95.



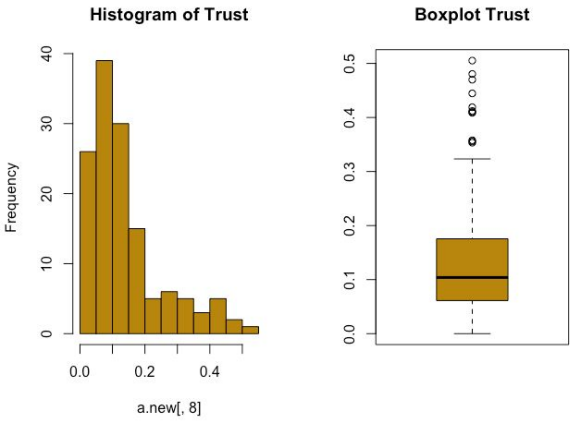
Freedom

The seventh variable “Freedom” (The extent to which Freedom contributed to the calculation of the Happiness Score), judging from the histogram seems to be not symmetrically distributed, but when it comes to mean (0.38) and median (0.4), which are very close to each other, we can see that the variable is actually symmetrically distributed. What is more, the mean is closer to median than the first and third quantile. All the values are ranging from 0 to 0.6.



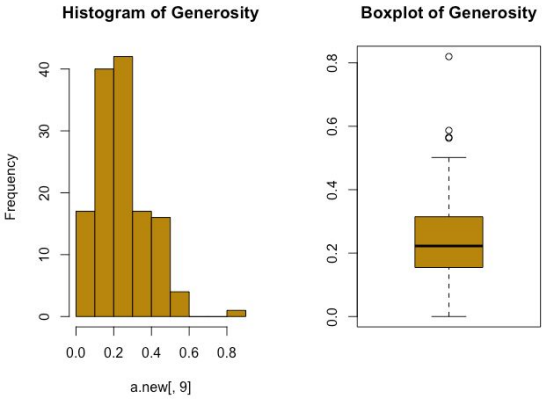
Trust (Government Corruption)

The eight variable “Trust(Government Corruption)” (The extent to which Perception of Corruption contributes to Happiness Score) seems to be not symmetrically distributed as it is observed in the histogram, but in terms of mean (0.14) and median (0.11) it is symmetrically distributed, because they are close to each other and are located between the first and third quantile. All the values are ranging from 0 to 0.51.



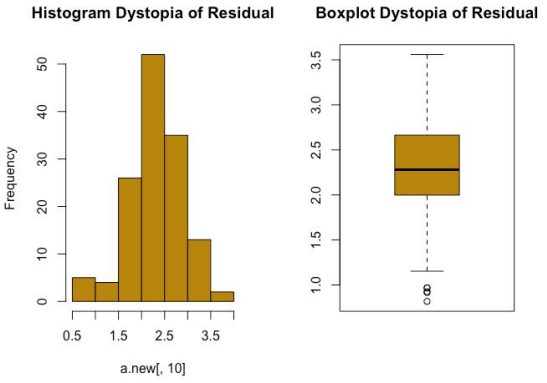
Generosity

The ninth variable “Generosity” (The extent to which Generosity contributed to the calculation of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (0.24) and the median (0.22) are close to each other and are located between the first and third quantile. All the values are ranging from 0 to 0.82.



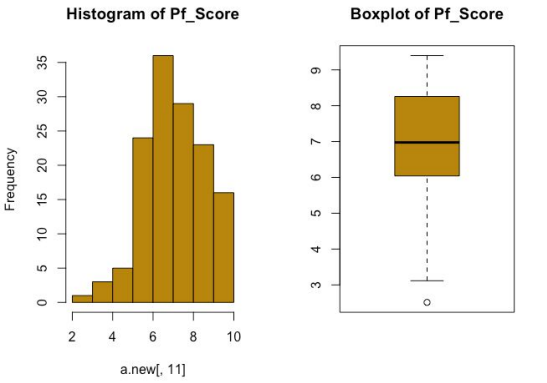
Dystopia Residual

The tenth variable “Dystopia Residual” (The extent to which Dystopia Residual contributed to the calculation of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (2.31) and the median (2.28) are close to each other and are located between the first and third quantile. All the values are ranging from 0.82 to 3.56.



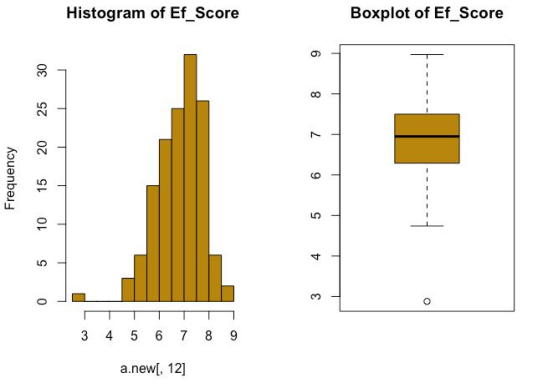
pf_score

The eleventh variable “pf_score (Personal Freedom index)” is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (7.04) and the median (6.96) are close to each other and are located between the first and third quantile. All the values are ranging from 2.51 to 9.4.



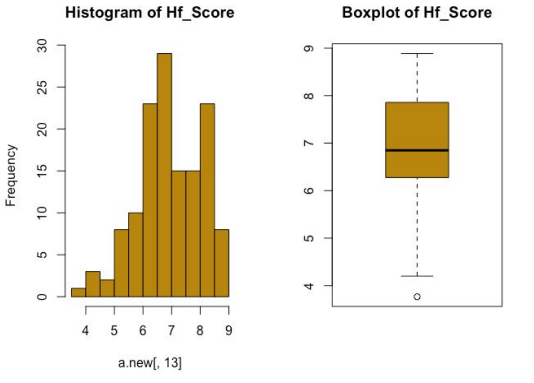
ef_score

The twelfth variable “ef_score (Economic freedom index)” is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (6.85) and the median (6.95) are close to each other and are located between the first and third quantile. All the values are ranging from 2.88 to 8.97.



hf_score

The thirteenth variable “hf_score (Human freedom index)” is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (6.95) and the median (6.85) are close to each other and are located between the first and third quantile. All the values are ranging from 3.77 to 8.89.



Correlation

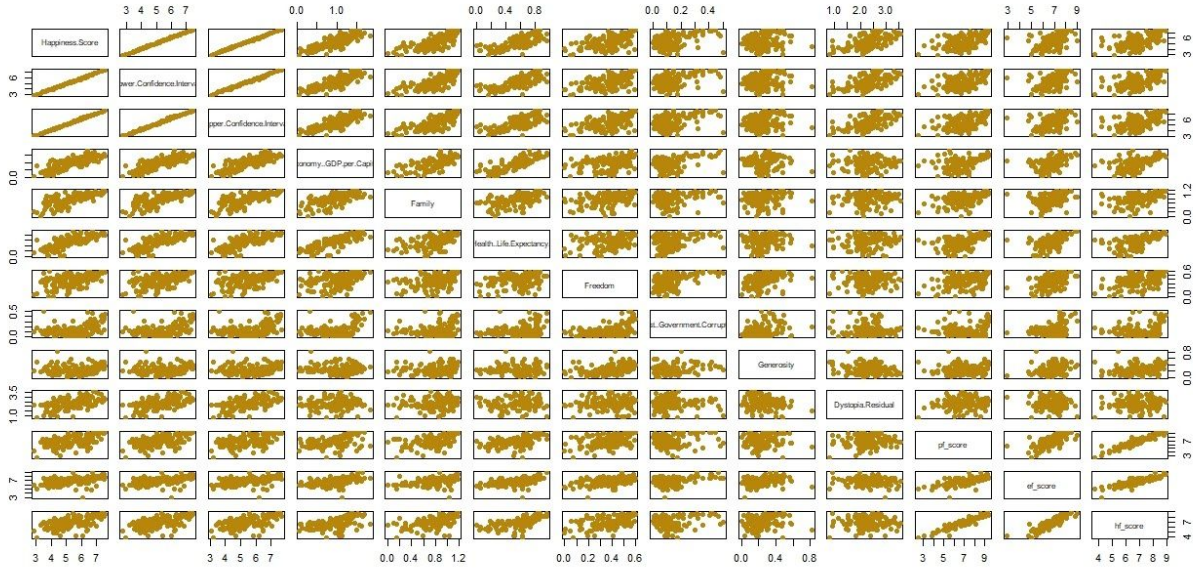
In general we can say that the correlation between the variables is mostly positive, even if the correlation between most of the variables is moderate. Although we can say that Happiness Score is mostly correlated with Economy (GDP per Capita) and Social aspects (like how families are involved in the life of the country and the life expectancy of the inhabitants). Actually, the correlation between Happiness score and:

- GDP is 0.80
- Family is 0.74
- Life expectancy is 0.76

The variables related to freedom do not seem to be strongly related to the happiness score, with the exception of the economic freedom (ef_score) that seems to be positively correlated to happiness.

So, in general, we could say that the happiness rate of a country is affected the most by the economic situation of the country (even if social factors still have their importance).

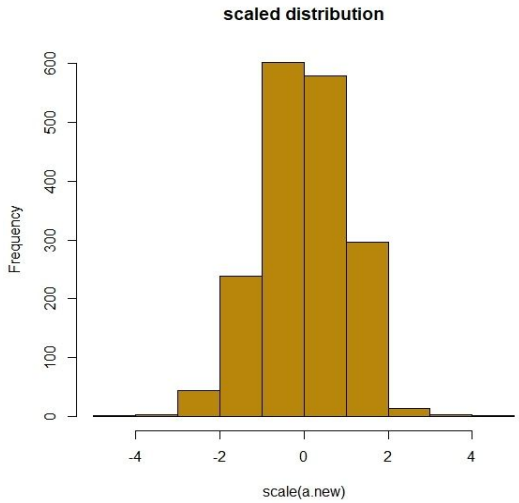
(There is a strong correlation between pf_score, ef_score and hf_score).



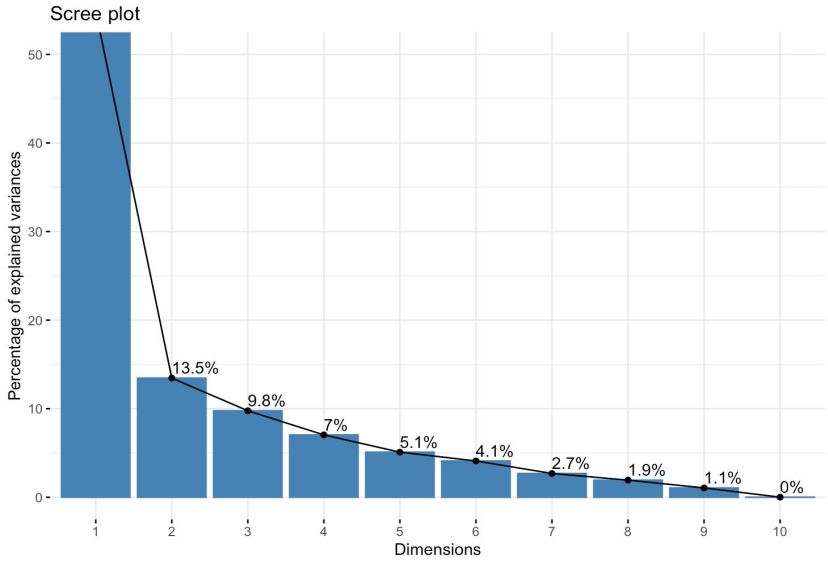
PCA

The aim of Principal component analysis is to visualize high dimensional projects in lower dimensions. While reducing the number of variables we have to make sure to principal components that maximize the explanation of the variance (in order to lose the lowest number of information as possible).

It is important to scale the datas before starting the analysis, so that it is possible to obtain a symmetric distribution (normalized).

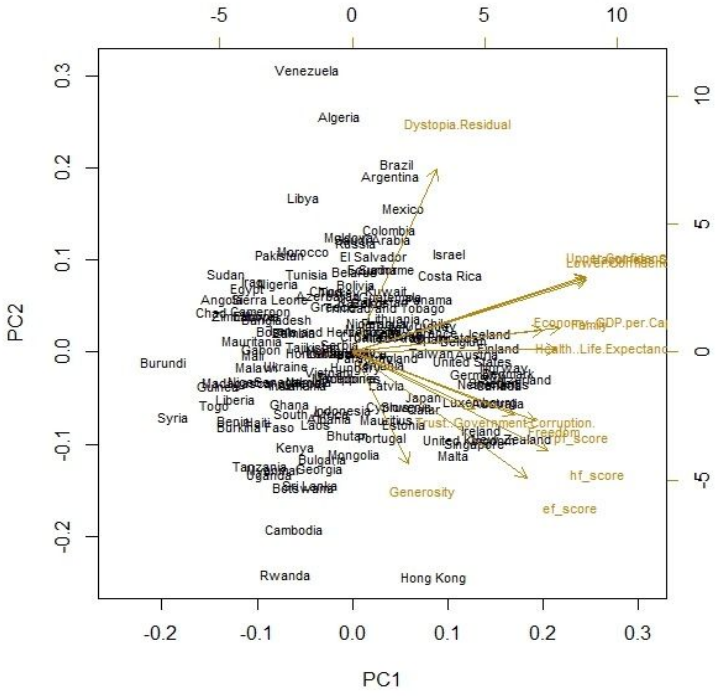


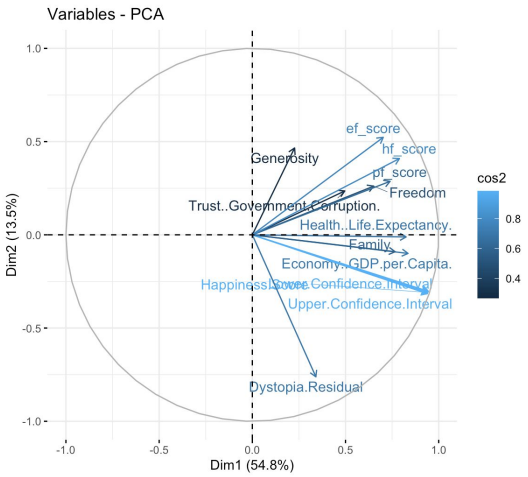
So we apply the command prcomp() in order to understand which are the principal components. From this command we obtained that the first principal component (PC1) is *Happiness Score* (and the Confidence interval related to it), whereas the second principal component (PC2) is the *Dystopia.residual*.



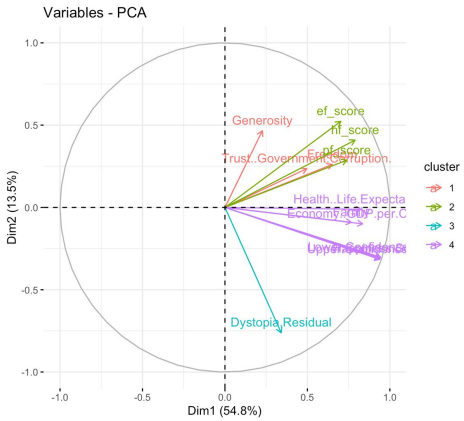
The first two variables explain 68.3% (a satisfying level of explanation) of the total variability, of which 54.8 % is explained by the first principal component that is “Happiness score”, whereas the second principal component, which is “Dystopia.Residual”, explains 13.5 % of the variability. As we can see from the Scree plot.

When it comes to analyse the biplot, we see that the two principal components are more or less ortogonal. Already from this graph it is possible to understand something about the level of happiness in the different countries. For example we could say that Syria is the country with the lower Happiness Score, because it is exactly opposite to the direction of the vector of Happiness score. We can also see that countries with positive values for GDP and health life expectancy (e.g. Iceland, Finland, Austria …) are also the ones that have the higher values for Happiness score. From this biplot moreover say that the level of freedom is not so important when it comes to explaining happiness.

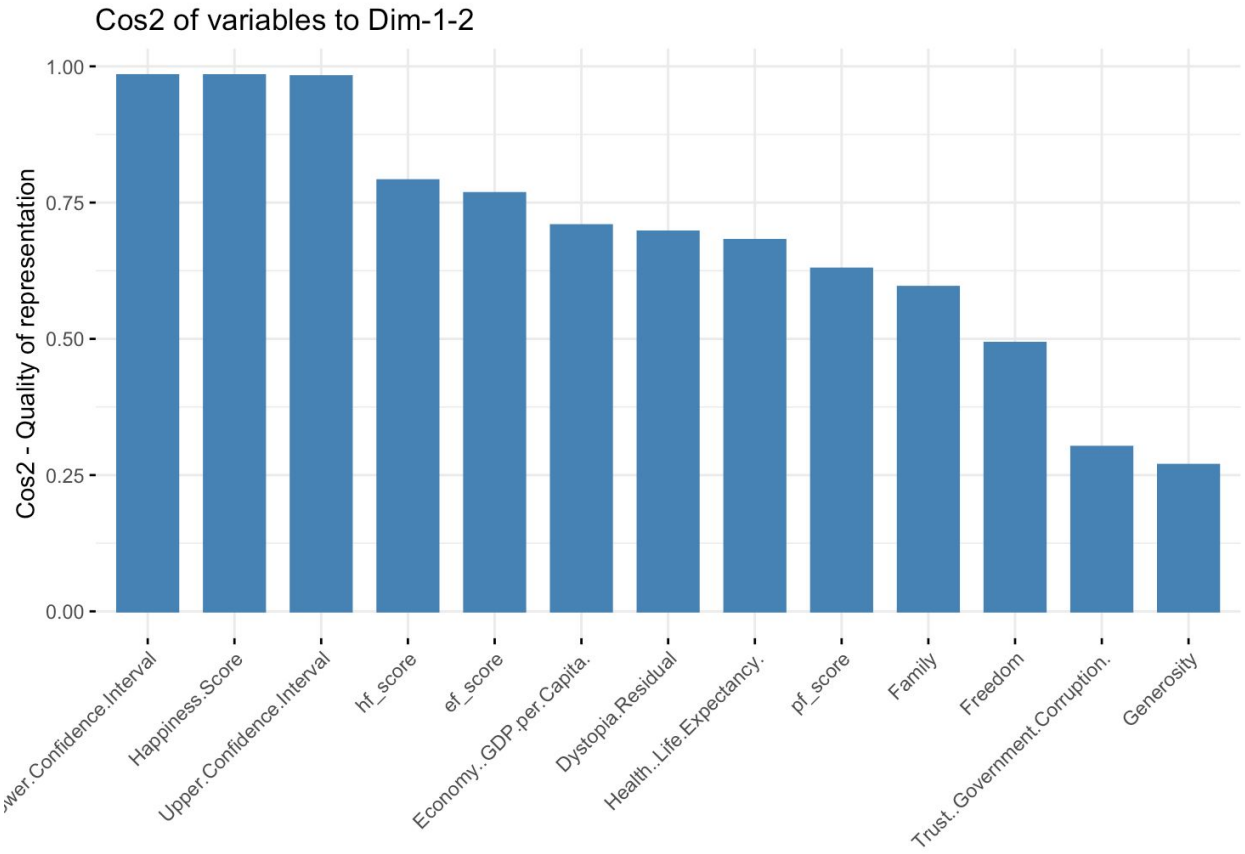




This graph shows the importance of a principal component for a given observation, light blue indicates high importance and dark blue indicates low importance. Thanks to this graph we have found out that “Happiness score” has the greatest importance of explaining the principal component whereas “Generosity” has the lowest importance.



In conclusion, thanks to this graph we have been able to group the variables in 4 clusters, and once again we have found out the variables that are mostly related to happiness are the ones related to Economy and Health/Life Expectancy.



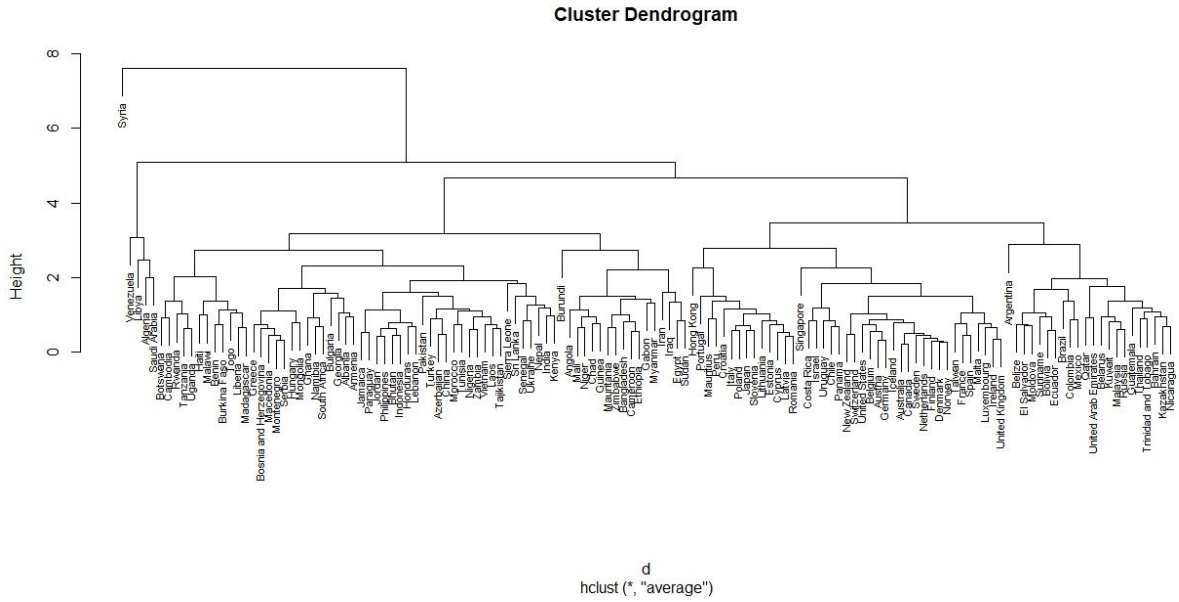
Cluster Analysis

As we have said in Descriptive Analysis we have no problems related to the distribution, because all the distributions are mostly symmetric. Therefore we did not conduct a transformation of the variables.

Hierarchical clustering

First we conducted cluster analysis with hierarchical method.

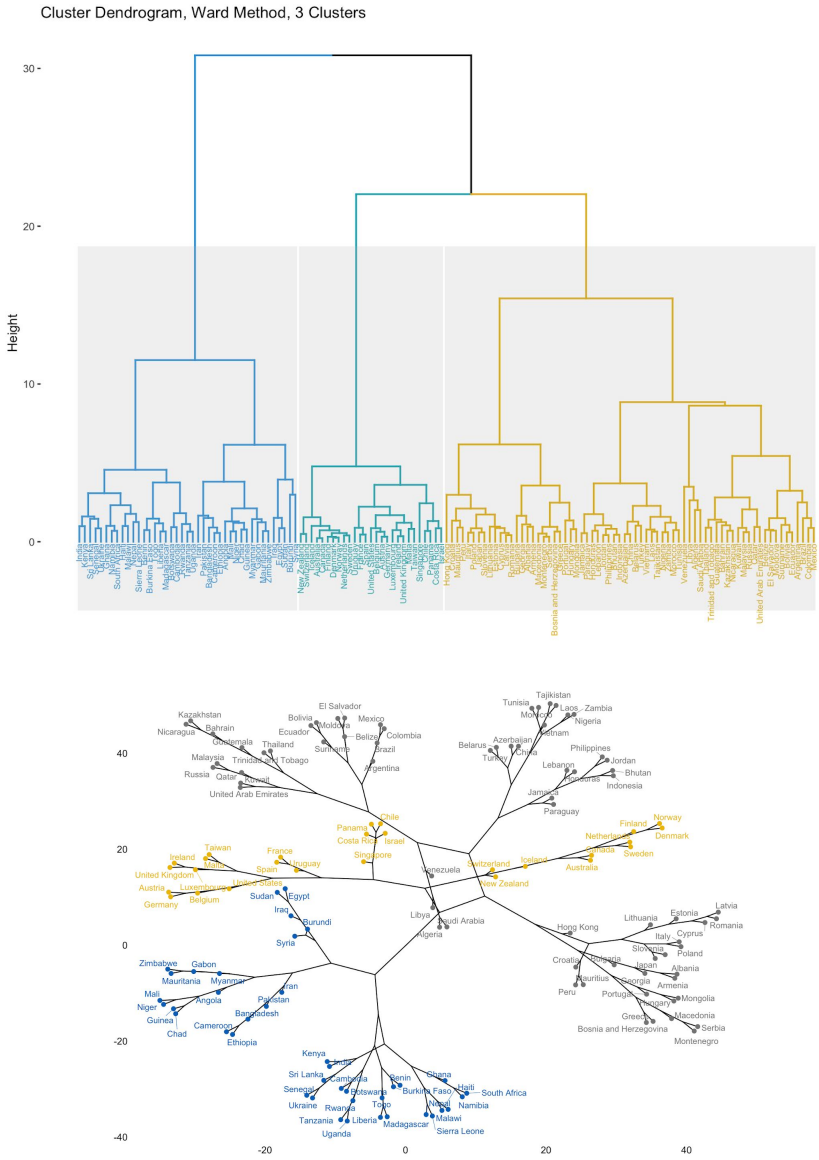
Before starting, we have conducted analysis to understand which is the best method (between ward, single, complete and average) to apply to hclust(). In order to do so, we applied the *cophenetic distance* and we found out that the higher correlation between the euclidean distance and cophenetic distance is obtained with the "average method". Although, applying the average method we obtained a chaining problem.



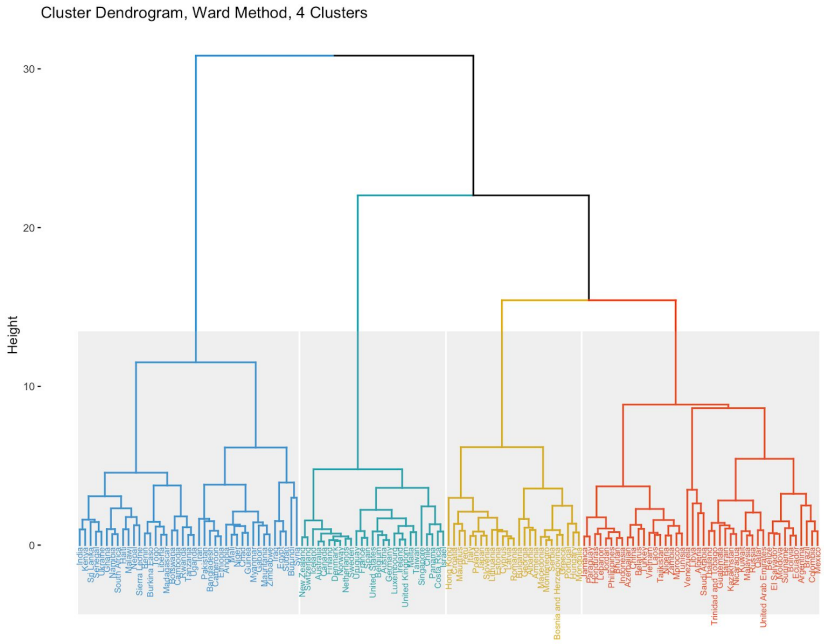
So we decided to apply the "*ward method*" that also showed a high correlation. With this graph it is easier to understand the clusters and we don't have any problems of "chaining". According to the dendrogram obtained, in our opinion, the ideal number of clusters is between 3 and 4.

The dendrogram with 3 clusters shows a division between countries that are underdeveloped (i.e. many countries in Africa and Asia), developed (i.e. countries in central europe, north america and oceania) and in development (i.e. south america, eastern europe and certain countries in asia).

- Underdeveloped → both for economy and rights
- Developed → both for economy and rights
- In development → some countries with strong economies but low rights and some countries with weak economies and more rights.



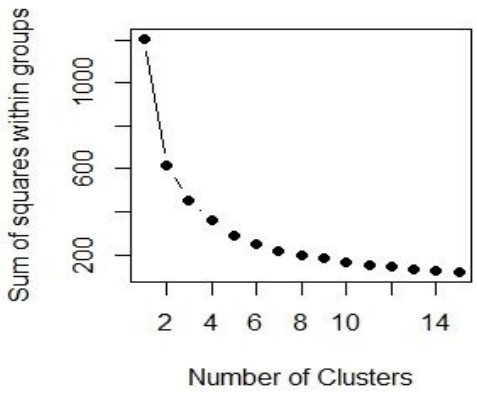
The main difference between having 3 or 4 clusters is the division of the “in development” countries located in East of Europe and the countries located in Asia and South America. So since it is just a geographical division we think that it is better to keep 3 clusters instead of 4.



K-Means

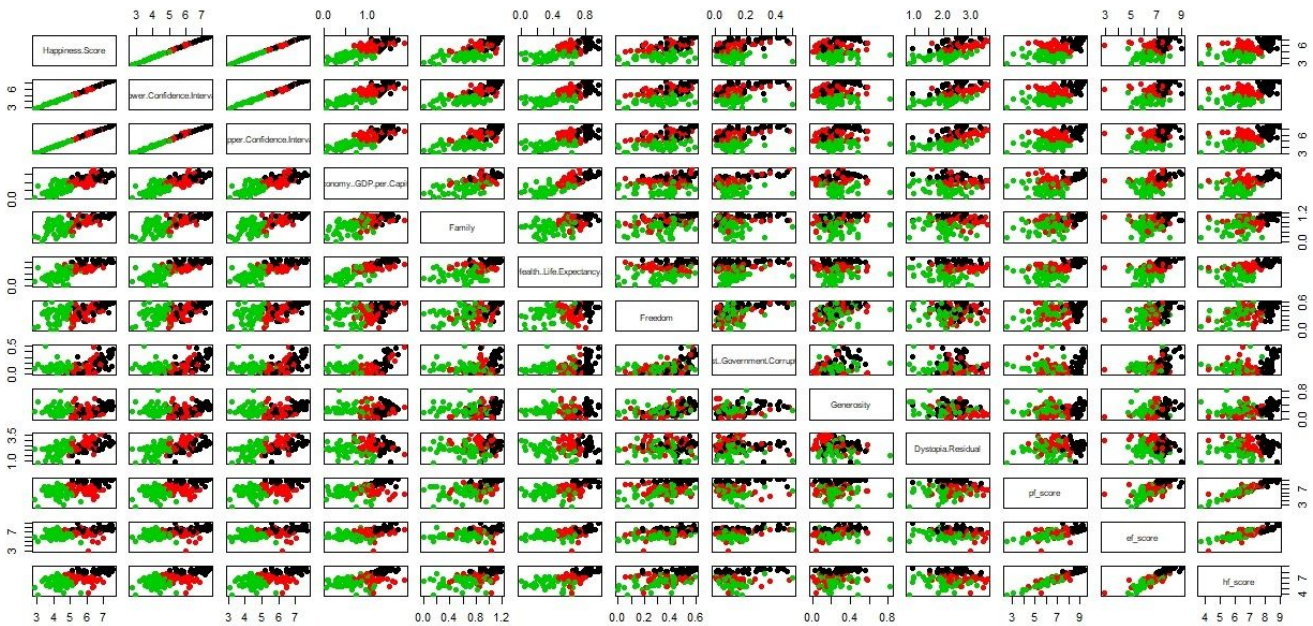
Now we move onto the K-Mean method.

By applying this method, we have to choose the number of clusters in advance, and in order to do so we use the “elbow chart”.

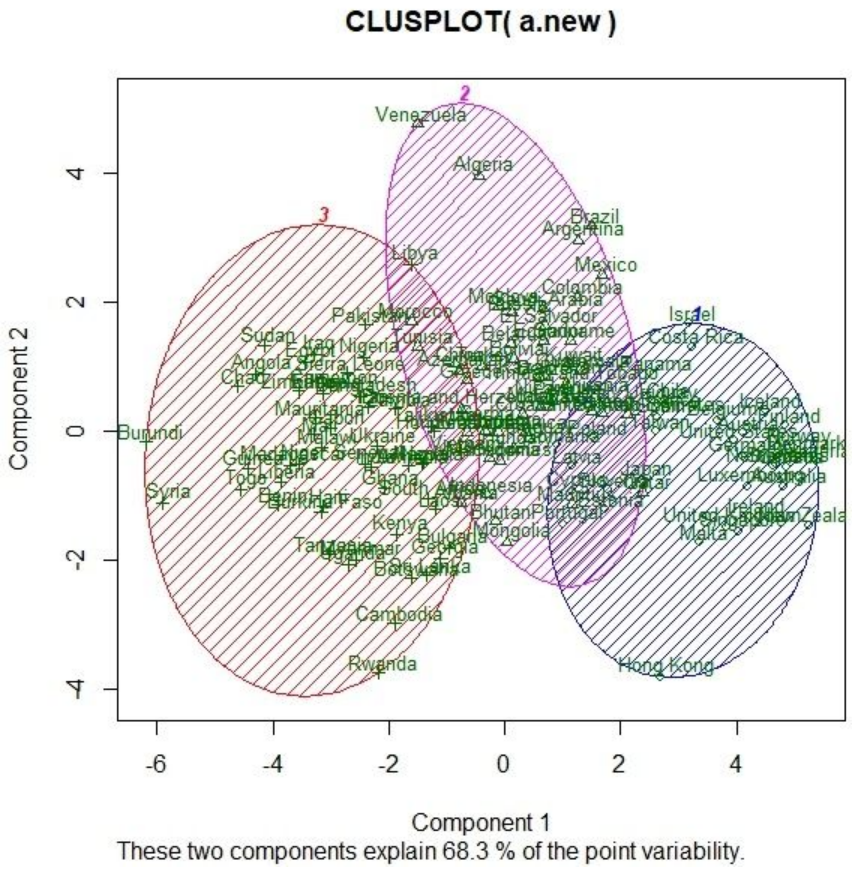


According to the previous graph the ideal number of clusters is between 3 and 5 (where the elbow is).

With 3 clusters:



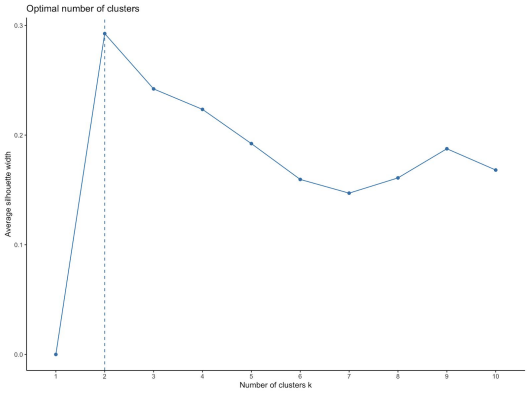
In general we can say that the correlation between the variables is mostly positive. Even if the correlation of most of the variables is moderate, although we can say that Happiness Score is mostly correlated with Economy (GDP per Capita).



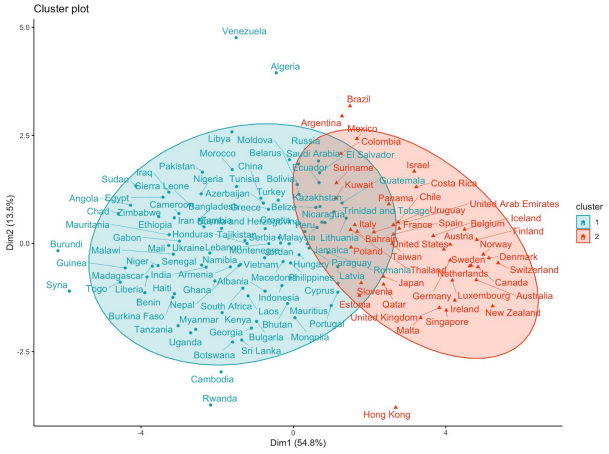
Our findings from the Clusplot seem to confirm what we have found out from our previous analysis with hierarchical method, except for a few cases. Actually, in group 1 we have countries that are developed in terms of economic as well as human rights issues, for these countries we have generally very high values for the Happiness Score. In group 2 we have countries with moderate values for the Happiness Score. In group 3 we have mostly underdeveloped countries both for economy and human rights and here the values for the Happiness Score are the lowest.

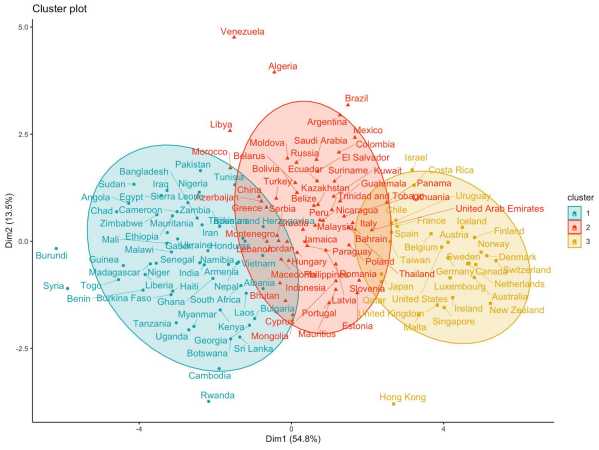
K-Medoids

To conclude, we did an analysis with the K-Medoids method. This was useful, because it is more robust to noise and outliers. The method is also called “PAM”. As it is observed from the following graph, the optimal amount of clusters is 2.



Although the suggested number of clusters is 2 comparing the two cluster plots we have obtained, we think it is better to divide the countries into three clusters because then the clusters can be better differentiated from each other and less overlap. Once again we are ready to confirm that it proves our earlier findings from the previous analysis.

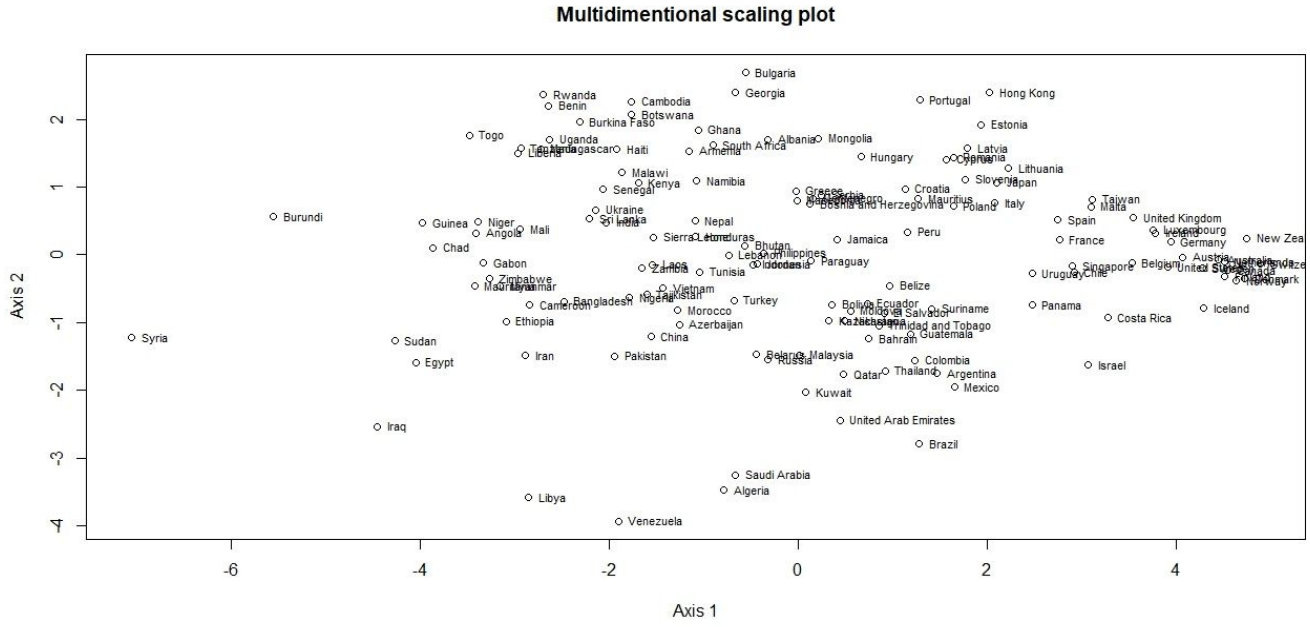




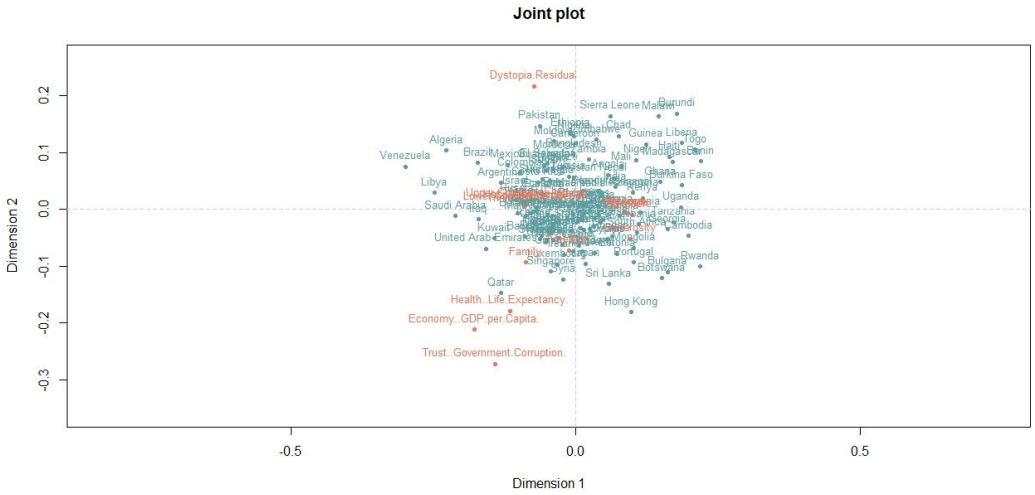
Multidimensional scaling

The Multidimensional scaling method gives information about if the variables are similar (or dissimilar) and in which way. Particularly, two different countries are more similar if the distance between the two is as short as possible (the closer they are, the more similar they are).

Countries that have similar characteristics (in terms of observed variables) are close to each other (shorter distance). On the left side of the plot we have what we can see so called "undeveloped countries" (mostly African countries), whereas on the right side of the plot we can see "developed countries" (mostly from Europe, Oceania and North America). Syria seems to be the only country with the lower/worse values in terms of happiness, while New Zealand is the country with the higher/better values. We can observe this difference, because these two countries are located so far from each other on the plot.



Perceptual map



The division into four quadrants gives us a possibility to group the observed countries by main characteristics:

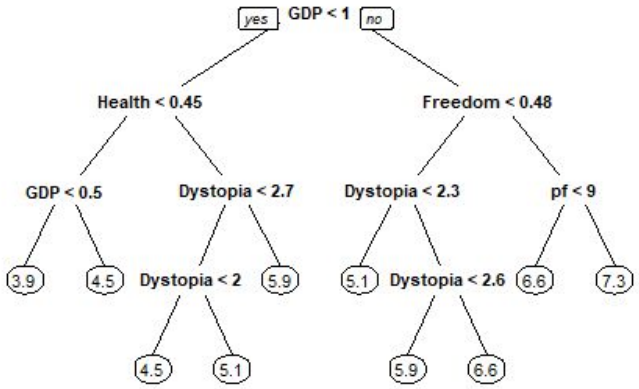
- Quadrant I: Distopia Residual and Happiness score
- Quadrant III: Economic, government, social and health factors
- Quadrant IV: Generosity and Human, Personal & Economic freedom

The closer a particular country is to these variables (that are shown in the joint plot), the more of those characteristics that country has.

It is interesting to analyze the plot, because we can clearly identify the countries that belong to the cluster, which we indicate as “developing countries” (is shown in our hierarchical cluster analysis). None of them are located in the Quadrant II (the countries located in this quadrant are “underdeveloped countries'', because they are far from all the variables that we analyze), some of them are highly valued in terms of Economy and Health, but they are so far from the variable “Dystopia Residual” (for example, Qatar or United Arabian Emirates). Whereas some of the observed countries have comparangly high (in a good way) values for the variable “Dystopia residual”, but have low values when it comes to Economy, Health and Government factors (for example, Algeria, Brazil, Mexico etc.).

In general, the “developed countries” are located in the middle of the plot, closer to the variable “Happiness score”, because they are the ones that have the higher and better values.

Classification with tree



Since we have only numerical variables we did classification with the regression tree.

First, we created a model where “Happiness score” is our response variable and all the other variables (less Upper and Lower confidence interval) are attributes.

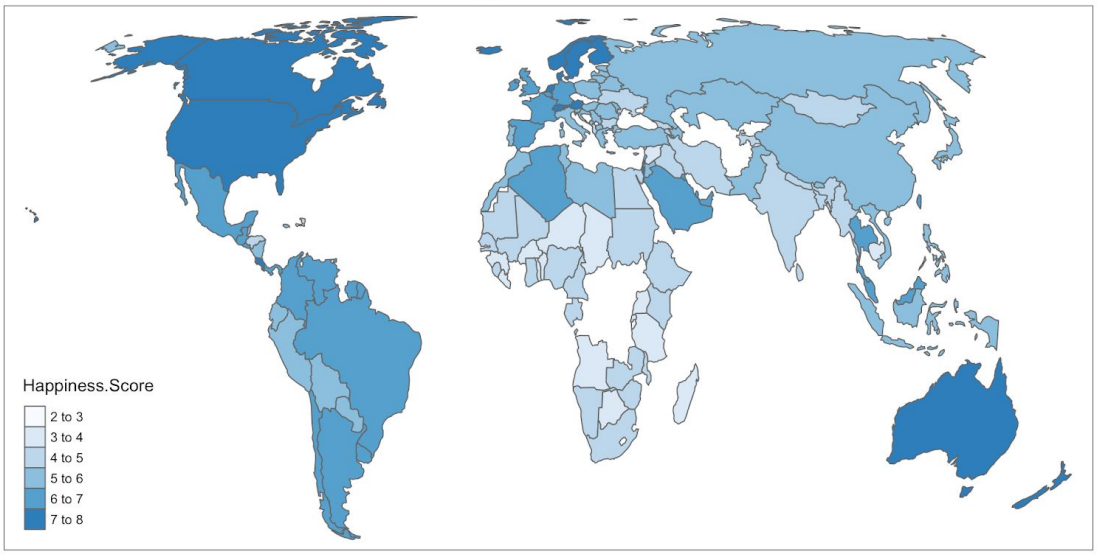
From this classification tree it is clear that the variables that determine the Happiness score of a country are GDP, Health, Freedom, Dystopia and Personal Freedom.

To understand how this tree works we can take as example Australia. The GDP Australia is not lower than 1, so we continue analysing the variable Freedom that in this case is not lower than 0.48, so that we continue analysing personal freedom (pf) that is not lower than 9 so the Happiness score for Australia is 7.3 .

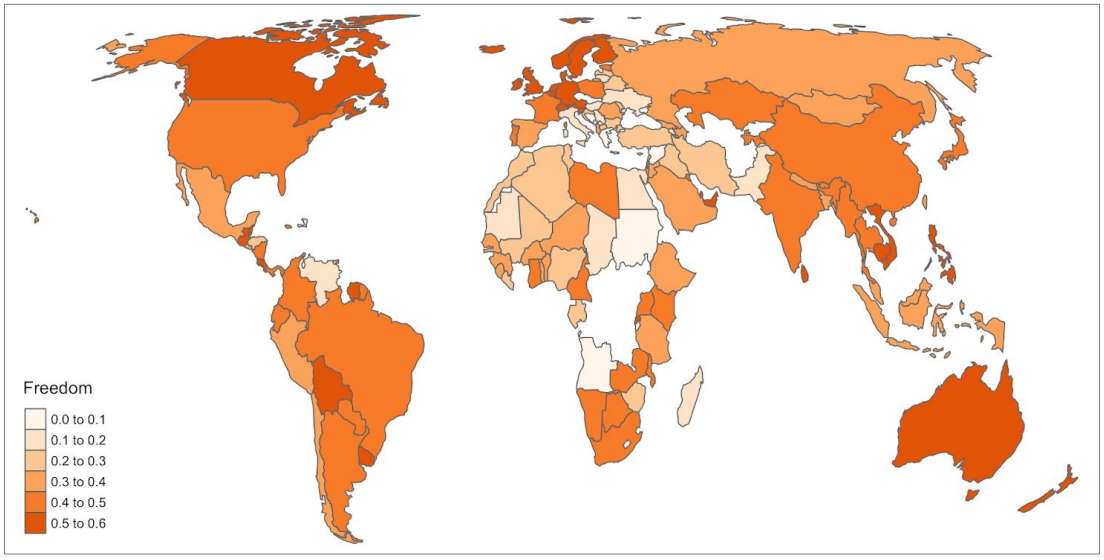
Excursus to geographical illustration of data

In the following, are shown two geographical maps, that illustrate either the corresponding values of the variables *Happiness Score* or *Freedom* for each country. The higher the score (or value), the darker the colour. In these two exemplifying illustrations we can clearly see that neighboring countries often share a similar level of happiness or freedom, as we can see in north America, Europe or Africa.

Happiness Score



Freedom



In order to generate these maps, we merged the dataset which we previously used with the "World" dataset from the R library (which contains geographical information for 177 countries). Since we do not have information for all the countries, some of them are uncoloured. For the plotting of the map data we used the library *tmap*.

Another interesting application for geographical illustration would be the cluster groups that we encountered earlier with cluster analysis. Since this is not within the scope of the requested report and presumably requires a lot more data wrangling we kept it with the illustration of the two variables as an example.