HATLAN WORK PROJECT

Honest Albert Tèmu
Andrea Corradini
Tommaso Premoli
Luca Codeluppi
Antonio De Patto
Niccolo’ Cibei

1 - Introduction

Our work is based entirely on the dataset provided by the World Happiness Report. Published more than 10 years ago by the Earth Institute, the first report strives to focus on happiness and the absence of misery as useful criteria for improving government policy. And so, to get at people's well-being, it would come naturally to ask them how satisfied they are with their lives. It would be simple, therefore, to ask the level of life satisfaction on a scale of 0 to 10. This allows people to assess their happiness without making assumptions about the causes. However, it makes even more sense to ask what habits, institutions and material conditions generate a society in which people have higher well-being.

How do people acquire the skills to promote their long-term well-being? The World Reports on Happiness have studied these types of questions each year, in part by comparing average life satisfaction in different countries and testing which population characteristics explain these differences. The aspects that World Reports want to focus on are the ethical ones and the institutional ones. So, it is necessary to ask whether people are trustworthy, generous and supportive of each other; at the same time, are people free to make important decisions about their lives? And to this one cannot but add material living conditions, such as income and health. Why would we ever want to realize unsupervised and supervised learning models? Starting from the unsupervised learning model, the decision to use clustering and PCA (Principal Component Analysis) on a topic as complex as world happiness is motivated by the desire to explore and understand the intricate relationships between various dimensions of well-being and differences between countries. But what are the main objectives of this analysis?

Through clustering, we intend to identify homogeneous patterns and groups of countries with similarities in the variables considered. This will allow us to identify common characteristics of countries with high levels of happiness or low levels of misery, as well as to identify any regional or cultural differences. It is interesting to see if countries that are part of the same geographic area (e.g., Scandinavian countries, Arab countries, Central African countries) show the same trends. Is it possible that neighboring countries show opposite trends? All of this can help to understand the distinguishing characteristics of happy and unhappy countries, highlighting socioeconomic, cultural or political differences that might influence the well-being of populations. Segmenting countries into clusters can be useful for targeted public policy formulation, allowing the specific needs of different clusters of countries to be identified and tailored measures to address their unique challenges. Once clusters of happy and unhappy countries have been identified, predictive models can be used to predict potential changes in happiness levels based on changes in the variables considered. This can be valuable for planning long-term interventions and policies to improve the well-being of populations.

PCA will enable us to reduce the complexity of the dataset and identify the main factors that contribute to the variation in happiness levels. This will help identify the most influential variables in determining the overall well-being of a nation and identify the main dimensions of well-being that emerge from the data. In a dataset as complex as the world happiness dataset, with many variables that may be interrelated, reducing dimensionality helps simplify the

understanding of the dataset. By reducing the number of variables to consider, it is easier to identify significant patterns, relationships and trends in the data. PCA identifies the major components or factors that contribute to variation in the data. These factors can be interpreted as latent dimensions or abstract concepts that influence happiness levels in different countries. Identifying these factors can provide valuable insights into the multidimensional nature of human well-being.

Turning to the supervised learning part, making a linear regression model can be important in the analysis of world happiness for several reasons. First, it allows us to analyze the relationships between the dependent variable (the level of happiness) and the independent variables (such as GDP, social support, generosity, etc.). This helps to understand how different variables influence countries' level of happiness and to identify which factors are most significant in determining well-being. Once the relationships between the variables are identified, the regression model can be used to predict happiness levels for countries where data are not available or to project future changes in happiness levels in response to changes in the independent variables. In addition to providing a clear framework for communicating the results of the analysis, linear regression can be used to assess the effectiveness of public policies in influencing the level of happiness of citizens.

1.1 - Dataset and variables

The dataset consists of 136 countries and showcases a first variable (Ladder score) that reports the national average level of happiness on a scale of 0 to 10. The measure comes from the January 20, 2023 Gallup World Poll publication covering the years 2005 to 2022. Next, here are the 6 variables designed to define an explanation of the level of happiness:

1. GDP: amount of product of each country, divided by the number of inhabitants of the country. GDP per capita provides information on the size of the economy and its performance. It is defined in terms of purchasing power parity (PPP) adjusted to constant 2017 international dollars, taken from the World Bank's World Development Indicators (WDI).
2. Social support: it consists of having someone to rely on in times of trouble. The question asked of respondents, then, is "If you were in trouble, do you have relatives or friends you can count on to help you when you need it, or not?" The variable is proposed to generate a national average of binary responses, where 0 corresponds to "No" and 1 to "Yes."
3. Healthy life expectancy: what is your level of physical and mental health? Mental health is a key component of subjective well-being and is also a risk factor for future physical health and longevity. Mental health influences and guides a range of individual choices, behaviors, and outcomes. Healthy life expectancies at birth are based on data extracted from the World Health Organization's (WHO) Global Health Observatory data archive.
4. Freedom to make life choices: "Are you satisfied or dissatisfied with your freedom to choose what to do with your life?" Thus, it is the national average of binary responses to that GWP question.

5. Generosity: "Have you donated money to a charity in the last month?" A clear indicator of a sense of positive community engagement and a central way humans connect with each other. Research shows that in all cultures, starting in early childhood, people are attracted to behaviors that benefit other people. Thus, it is the residual of the regression of the national average of GWP responses to the question about donating on the log of GDP per capita.
6. Corruption perception: "Is corruption widespread or not within the government?" and "Is corruption widespread or not within business?" The measure is the national average of responses to these two GWP questions. Basically, we ask whether people trust the government and the benevolence of others. The overall perception is just the average of the two 0-o-1 responses. The perception of corruption at the national level is just the average response of the overall perception at the individual level.

1.2 - Ladder score

The Gallup World Poll, which remains the main data source for this report, asks respondents to rate their current life as a whole using the image of a scale, with the best possible life for them as a 10 and the worst possible as a 0. Normally, about 1000 responses are collected each year for each country. The weights are used to construct nationally representative population averages for each year in each country. What the report does is project the usual happiness rankings onto a three-year average of these life ratings, since the larger sample size allows for more precise estimates.

Tab.1.1: an overview of the world's happiest and unhappiest countries

No.	Country name	Ladder score	No.	Country name	Ladder score
1	Finland	7,804	15	United States	6,894
2	Denmark	7,586	25	Singapore	6,587
3	Iceland	7,530	28	Uruguay	6,494
4	Israel	7,473	33	Italy	6,405
5	Netherlands	7,403	59	Mauritius	5,902
6	Sweden	7,395	64	China	5,818
7	Norway	7,315	85	South Africa	5,275
8	Switzerland	7,240	126	India	4,036
9	Luxembourg	7,228	136	Lebanon	2,392
10	New Zealand	7,123	137	Afghanistan	1,859

Finland occupies the first place with a significantly higher score than all other countries. Second place is held by Denmark, with a confidence region bounded by 2nd and 4th place. Iceland is in 3rd place and, with a smaller sample size, has a confidence region ranging from 2nd to 7th place. Israel is in 4th place, up five places from the previous year's data, with a confidence interval of 2nd to 8th. The fifth to tenth positions are occupied by the Netherlands, Sweden, Norway, Switzerland, Luxembourg and New Zealand. As can be seen, there is a wide gap between countries in the top and bottom positions, with the top grouped more closely than the

bottom. Within the top group, the national life assessment scores have a gap of 0.40 between the first and fifth positions and another 0.28 between the fifth and tenth positions. So, there is a difference of less than 0.7 points between the first and tenth positions. If we look at the last 10 countries, however, the gap is much wider, with a range of about 2.1 points. The table shows Afghanistan in last position, and Lebanon in second to last, with scores significantly different from each other and from all the top countries. In addition, to summarize a dataset consisting of 137 countries, we wanted to prioritize countries that are part of continents other than Europe. Singapore, which occupies the twenty-fifth position, is the first Asian country on the list; Uruguay, on the other hand, while being the twenty-eighth country on the list, is the first among South American countries. The situation regarding the African continent, on the other hand, was easily predicted: the happiest country (Mauritius) does not reach the sufficiency, and the second (South Africa) is even 0.7 points behind the first, occupying the eighty fifth position in the ranking. Two world superpowers such as the United States and China have a gap of just over 1 point and are stationed in 15th and 64th position, respectively. And finally Italy, which, with a more than adequate score, is the thirty-third happiest country in the world.

2 - Principal Component Analysis (PCA)

PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while explaining as much possible variation in the data. PCA is scale-insensitive, therefore data normalization is not necessary. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. The principal components are eigenvectors of the data's covariance matrix.

```
> str(happiness_dataset)
tibble [137 × 19] (S3: tbl_df/tbl/data.frame)
 $ Country name                          : chr [1:137] "Finland" "Denmark" "Iceland" "Israel" ...
 $ Ladder score                          : num [1:137] 7.8 7.59 7.53 7.47 7.4 ...
 $ Standard error of ladder score        : num [1:137] 0.0362 0.041 0.0486 0.0316 0.0293 ...
 $ upperwhisker                          : num [1:137] 7.88 7.67 7.62 7.53 7.46 ...
 $ lowerwhisker                          : num [1:137] 7.73 7.51 7.43 7.41 7.35 ...
 $ Logged GDP per capita                 : num [1:137] 10.8 11 10.9 10.6 10.9 ...
 $ Social support                        : num [1:137] 0.969 0.954 0.983 0.943 0.93 ...
 $ Healthy life expectancy               : num [1:137] 71.1 71.3 72.1 72.7 71.6 ...
 $ Freedom to make life choices          : num [1:137] 0.961 0.934 0.936 0.809 0.887 ...
 $ Generosity                            : num [1:137] -0.0188 0.1342 0.211 -0.0231 0.2127 ...
 $ Perceptions of corruption             : num [1:137] 0.182 0.196 0.668 0.708 0.379 ...
 $ Ladder score in Dystopia              : num [1:137] 1.78 1.78 1.78 1.78 1.78 ...
 $ Explained by: Log GDP per capita      : num [1:137] 1.89 1.95 1.93 1.83 1.94 ...
 $ Explained by: Social support          : num [1:137] 1.58 1.55 1.62 1.52 1.49 ...
 $ Explained by: Healthy life expectancy : num [1:137] 0.535 0.537 0.559 0.577 0.545 ...
 $ Explained by: Freedom to make life choices: num [1:137] 0.772 0.734 0.738 0.569 0.672 ...
 $ Explained by: Generosity              : num [1:137] 0.126 0.208 0.25 0.124 0.251 ...
 $ Explained by: Perceptions of corruption : num [1:137] 0.535 0.525 0.187 0.158 0.394 ...
 $ Dystopia + residual                   : num [1:137] 2.36 2.08 2.25 2.69 2.11 ...
```

After observing the variables inside our dataset, their class, and the first few observations of each. In fact, the dataset has 137 observations and 19 variables. Some of the variable names carry less significance to the analysis we are aiming for, thus we decided to remove them completely (all variables starting with “Explained by together with whisker low and whisker high variables from my dataset because these variables give only the lower and upper confidence interval of happiness score and there is no need to use them for visualization and prediction).

2.2 - Applying PCA

```
> # Summary of PCA
> summary(pca_result)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.0958 1.1074 0.86972 0.79876 0.69285 0.47620 0.39663 0.35047
Proportion of Variance 0.5491 0.1533 0.09455 0.07975 0.06001 0.02835 0.01966 0.01535
Cumulative Proportion  0.5491 0.7023 0.79688 0.87663 0.93664 0.96498 0.98465 1.00000
```

Now, all the resources are available to conduct the PCA analysis. First, the princomp() computes the PCA, and summary() function shows the result.
From the above image, we notice that eight principal components have been generated (PC1 to PC8), which also correspond to the number of variables in the data.
Each component explains a percentage of the total variance in the data set. In the Cumulative Proportion section, the first principal component explains almost 55% of the total variance. This

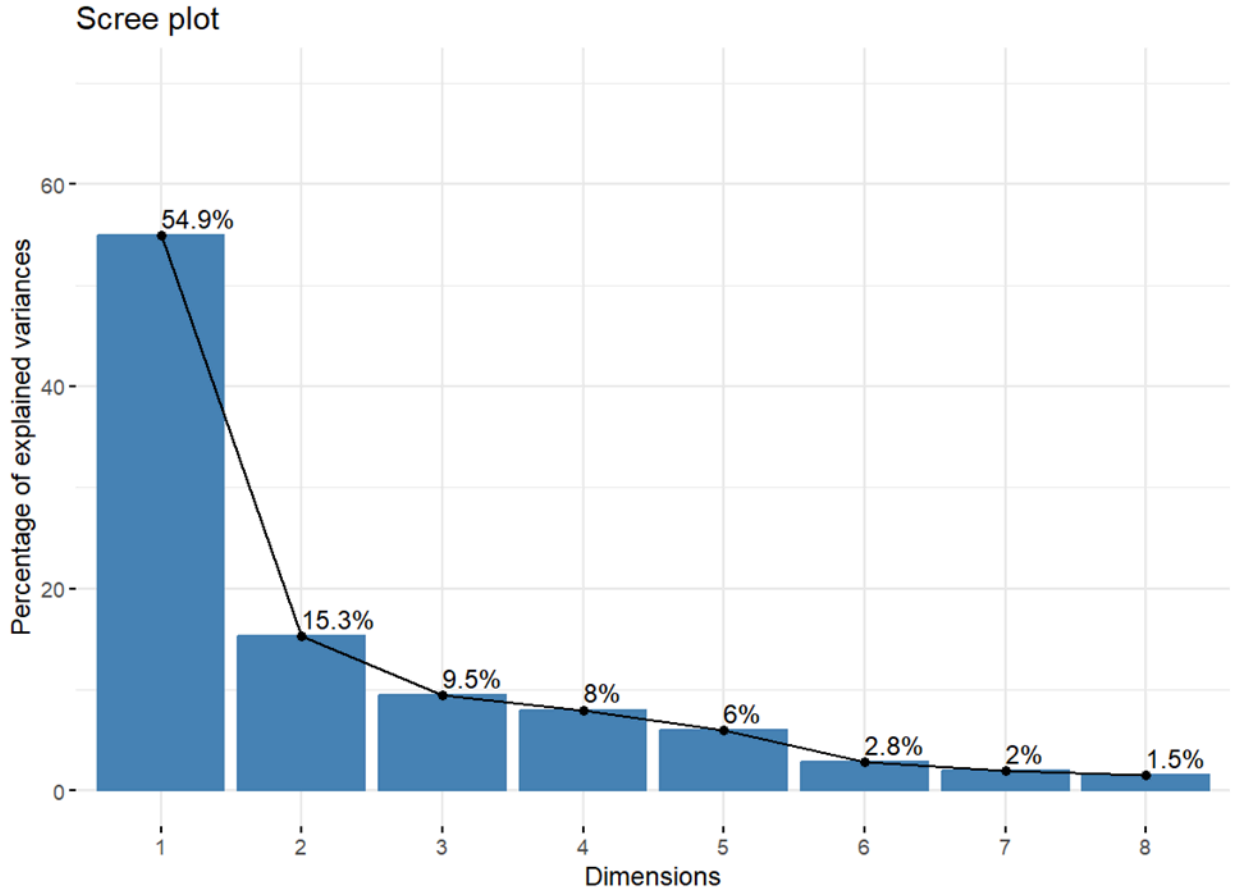
implies that more than half of the data in the set of 8 variables can be represented by just the first principal component. The second one explains 15.32% of the total variance.
The cumulative proportion PC1 and PC2 explains 70% of the total variance. This means that the first two principal components can accurately represent the data.

1.3 - Visualization of the principal components

There are a couple of standard visualization strategies that can help the user glean insight into the data, and this section aims to cover some of those approaches, starting with the scree plot.

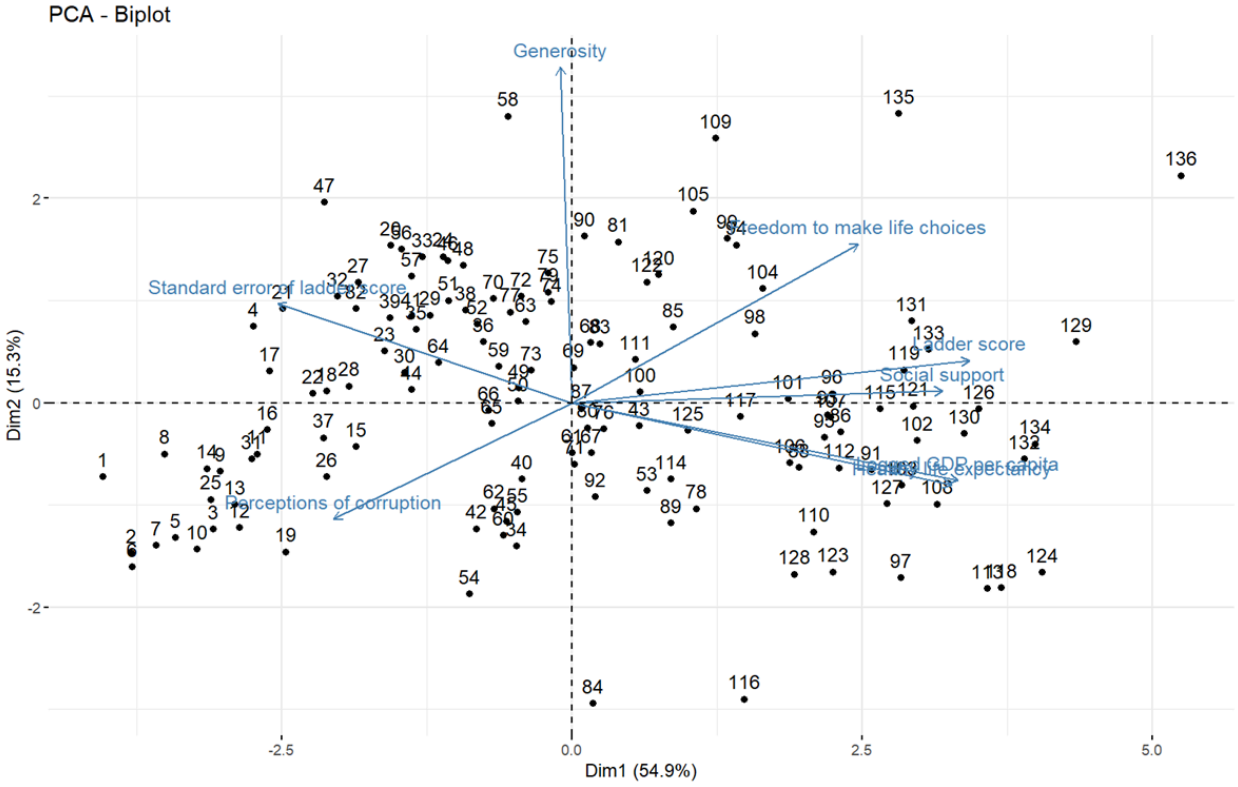
1.3.1 - The Scree Plot

The first approach of the list is the scree plot. It is used to visualize the importance of each principal component and can be used to determine the number of principal components to retain. The scree plot can be generated using the fviz_eig() function.



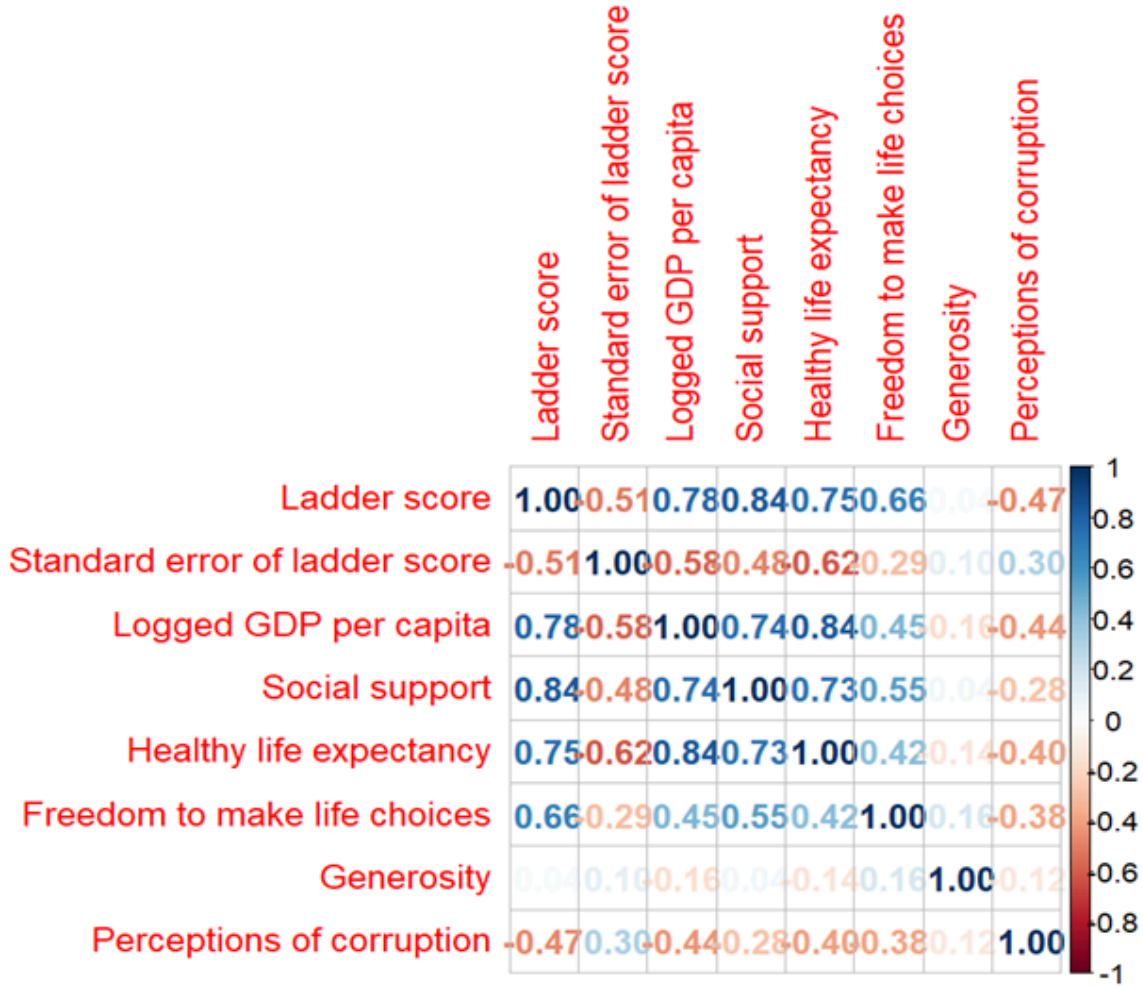
This plot shows the eigenvalues in a downward curve, from highest to lowest. The first two components can be considered to be the most significant since they contain almost 70.2% of the total information of the data. Applying the ”elbow rule” it can be seen that one can optimally retain 2 components.

1.3.2 - The Biplot



We see that for instance happiness score and social support are positively correlated to each other. Same thing can be said with the GDP per capita and Healthy life expectancy.
We also see that PC1 is positively correlated with happiness score, social support and freedom to make life choices, GDP per capita and Healthy life expectancy while being negatively correlated with corruption and generosity.
PC2 however is highly positively correlated with generosity while being negatively correlated with the perceptions of corruption as well as GDP per capita and Healthy life expectancy contrary to the first principle component.

1.3.3 - Correlation plot



From the above chart, we can obtain the following conclusions:
The Happiness score is highly related to the GDP per Capita, Social Support, Healthy Life Expectancy together with the freedom to make life choices.
The Happiness score however was found NOT to be related at all with the Generosity Variable, which can be a surprise to a lot of people as well as having a negative relation to perceptions of corruption which tends to make more sense.

We can see that in countries with High Level of GDP, they will have a very High Life Expectancy and Social Support. Thus, confirming that rich countries tend to be the happiest, live longer and have more family support.

1.4 - Correlation between PC's and given variables.

From the above chart between PC's and our given variables we can observe that PC1 has a high-direct correlation with the Happiness score, GDP, Social Support, Healthy Life Expectancy and Freedom of choice. Negatively with Corruption. So, due to the correlation of this principal component with the Happiness Score, we will say that this principal component will refer to the Happiest Countries.
Same as for the principal component 2 with Generosity we can conclude that this principal component will refer to the most generosity countries.
All the above seem to match exactly as the conclusions obtained from the biplot

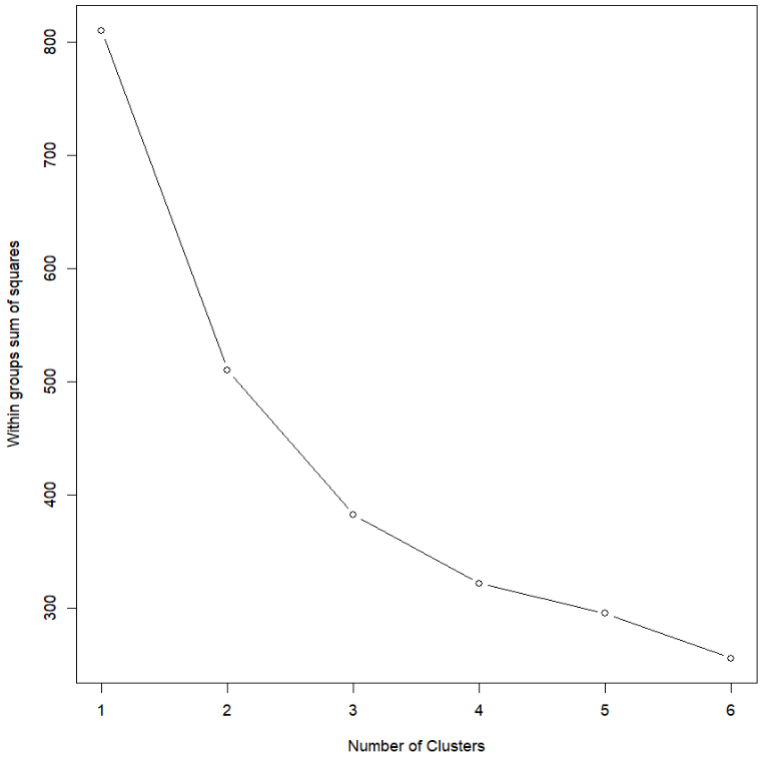
3 - k-means clustering

Having provided a comprehensive description of the dataset, we can now proceed with the clustering process. This involves employing two unsupervised learning methodologies, namely hierarchical and k-means clustering. These methodologies group the input data into subsets known as clusters, leveraging the similarity of their characteristics as a basis for classification. The first technique is based on the partition of the observations on the basis of a pre-specified number of clusters while the second allows us to obtain a dendogram, which is a representation of the information in a tree-shaped graph which allows us to visualize the clusters obtained.We will now look at cluster analysis using the k-means method, a non-hierarchical classification method. This algorithm allows the statistical units to be classified into n distinct groups, with n fixed a priori, through an iterative procedure based on the following steps:

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
    - For each of the K clusters, compute the cluster centroid. The k-th cluster centroid is the vector of the p feature means for the observations in the k-th cluster.
    - Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance)

The basic idea of k-means is to define clusters so that variation within the cluster is minimized. To quantify this variance, we calculate something called the within-cluster sum of squares (WCSS):

$$WCSS = \sum_{C_k}^{C_n} (\sum_{d_i in C_i}^{d_m} distance(d_i, C_k)^2)$$



where C are the centroids and d are the points in each cluster. As a first step, therefore, we will have to deal with the number of clusters to consider. This aspect can be analyzed with two distinct techniques, the first relating to the wss plot (within sum of squares plot) and the second relating to the silhouette plot. Within-Cluster Sum of Squares (WSS) is a measure of how far away each centroid is from their respective instances or points. The larger the WSS, the more dispersed the cluster values are from the centroid. The objective of

this metric is to find the “elbow” of the WSS curve in order to determine the smallest number of clusters that captures the most amount of signal in your data. As can be seen from the wss plot, there is no sharp drop in the within variance and approximately considering 3 or 4 groups would be the best choice in our case. We can therefore proceed by considering both options and subsequently verify which subdivision into groups is best. In any case we can rely on the second method, called the silhouette method. The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation) in a range of [-1, 1]. Silhouette coefficients near 1 indicate that the sample is far away from the neighboring clusters, so a score of 1 denotes that the data point is very compact within the cluster to which it belongs and far away from the other clusters.
A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters and we could have two or more clusters overlapping. Finally, negative values indicate that those samples might have been assigned to the wrong cluster.
The Silhouette Value *s(i)* for each data point *i* is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

s(i) is defined to be equal to zero if *i* is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters. Here, *a(i)* is the measure of similarity of the point *i* to its own cluster. It is measured as the average distance of *i* from other points in the cluster.

For each data point $i \in C_i$ (data point i in the cluster C_i), let

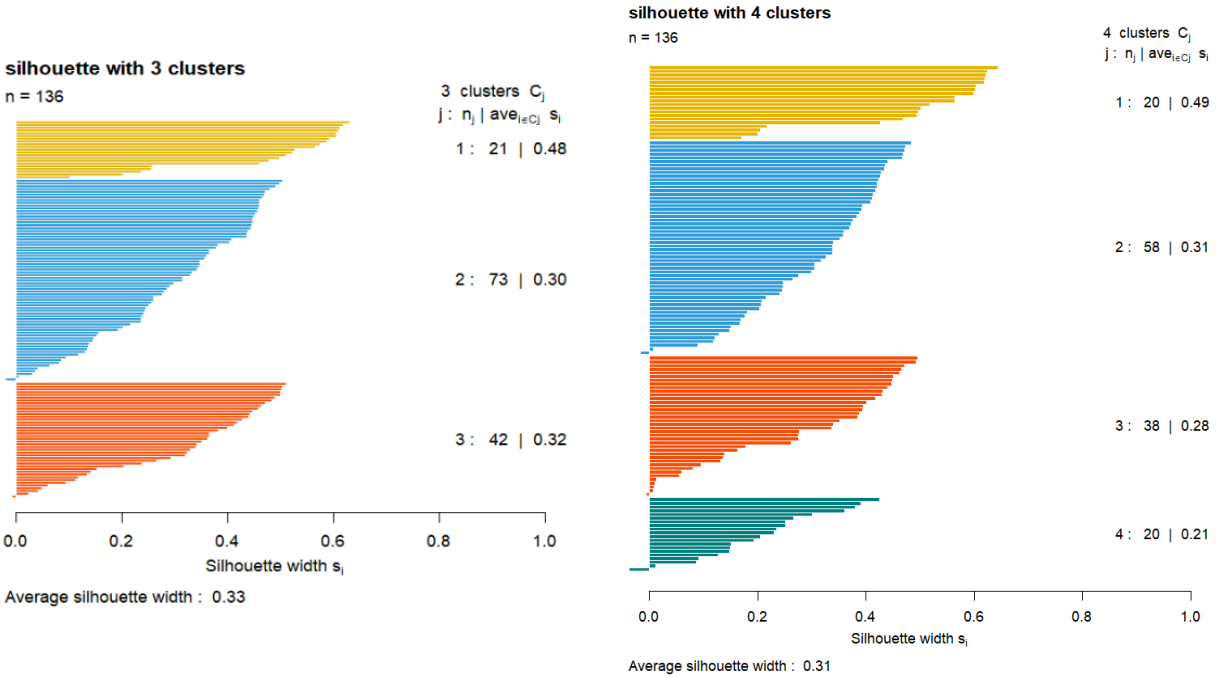
$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Similarly, *b(i)* is the measure of dissimilarity of *i* from points in other clusters. *b(i)* is the *minimum average distance* from i to all clusters to which i does not belong. *d(i, j)* is the distance between points *i* and *j*. Generally, Euclidean Distance is used as the distance metric.

For each data point $i \in C_i$, we now define

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

Therefore, the graph of the silhouette with 3 clusters has a higher average value than the graph with 4 clusters, indicating that on average each cluster better represents the distribution of the points and the separation between clusters is clearer. In any case, the difference between the two graphs is not so marked since considering 3 clusters we have an average value of 0.33 versus a value of 0.31 if we consider 4 clusters. We can therefore consider both choices on the number of clusters valid.

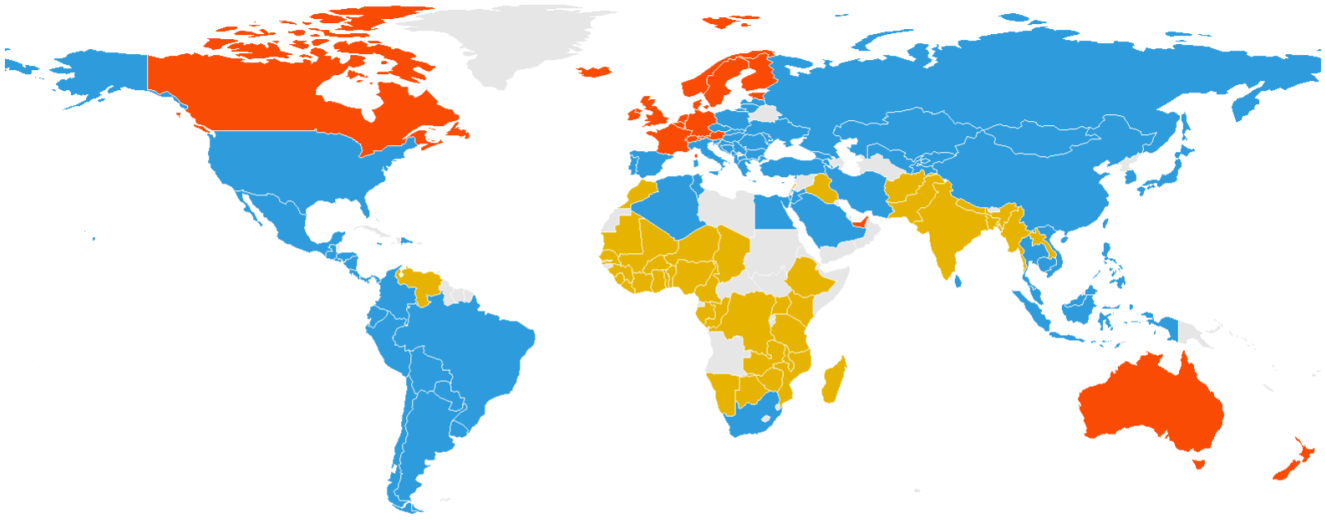


Therefore, considering 3 clusters and looking at the map of the states we can see how the richest states are grouped together in the first cluster, indicated in orange. The states in question, which include Canada, France, Germany, the United Kingdom, Australia and the Scandinavian countries, are known for their high levels of GDP and welfare policies which lead to greater levels of happiness for the population. Considering the second cluster however, the most representative states are the countries of Eastern and Southern Europe, China, the United States and most of South America. Finally, the poorest countries fall into the last cluster, characterized by high levels of corruption, low life expectancy, limited growth and investment opportunities which in general lead these countries to be characterized as the unhappiest. We find, for example, countries such as Venezuela, most of the central African states, Pakistan and India. Looking at the map it therefore comes naturally to think that the richest countries can generally also be defined as the happiest.

Looking at the variables present in the summary tables of each cluster, we note how the gdp gradually decreases between the various clusters, indicating how happiness is somehow correlated to wealth. The same can be said for the variable linked to social support. In this case there is a sharp difference between the second and third clusters in which the value of the variable goes from 0.8546 to 0.6437. In relation to life expectancy, we note that this gradually decreases between one cluster and another and the same can be said of the freedom variable in which there are no sudden jumps between one cluster and another but a gradual decrease in values. In relation to generosity, it can be stated that the most generous countries are also the richest and happiest ones, however the level of generosity is greater in the third cluster, relating to less happy countries, compared to the second, relating to averagely happy countries. This is therefore a surprising aspect for our analyzes and with the opposite trend to all the remaining variables. Let's now move on to the last variable, that is, the one linked to corruption. This is the variable that varies most on average between the first and second clusters, going from an average value of 0.384 in the first cluster to a value of 0.7899 in the second. There is therefore

a very marked variation if we compare the second and third clusters as the difference between the two values is really small

Fig. 2.2: world map with the states classified into clusters using k-means method



Tab. 2.3: summary of the first cluster

gdp	social_supp	life_exp	freedom	generosity	corruption
Min. :10.54	Min. :0.817	Min. :66.24	Min. :0.6870	Min. :-0.10000	Min. :0.146
1st Qu.:10.79	1st Qu.:0.888	1st Qu.:71.05	1st Qu.:0.8550	1st Qu.: 0.02700	1st Qu.:0.271
Median :10.90	Median :0.920	Median :71.30	Median :0.8870	Median : 0.09600	Median :0.385
Mean :10.97	Mean :0.914	Mean :71.49	Mean :0.8846	Mean : 0.09357	Mean :0.384
3rd Qu.:11.09	3rd Qu.:0.943	3rd Qu.:72.05	3rd Qu.:0.9340	3rd Qu.: 0.16500	3rd Qu.:0.496
Max. :11.66	Max. :0.983	Max. :77.28	Max. :0.9610	Max. : 0.25300	Max. :0.668

Tab. 2.4: summary of the second cluster

gdp	social_supp	life_exp	freedom	generosity	corruption
Min. : 8.237	Min. :0.7160	Min. :56.99	Min. :0.4750	Min. :-0.25400	Min. :0.5220
1st Qu.: 9.367	1st Qu.:0.8110	1st Qu.:65.30	1st Qu.:0.7690	1st Qu.:-0.09900	1st Qu.:0.7210
Median : 9.811	Median :0.8670	Median :67.00	Median :0.8090	Median :-0.05700	Median :0.8080
Mean : 9.816	Mean :0.8546	Mean :67.06	Mean :0.8085	Mean :-0.01501	Mean :0.7889
3rd Qu.:10.353	3rd Qu.:0.9060	3rd Qu.:69.00	3rd Qu.:0.8770	3rd Qu.: 0.04000	3rd Qu.:0.8660
Max. :11.048	Max. :0.9530	Max. :74.35	Max. :0.9580	Max. : 0.53100	Max. :0.9290

Tab. 2.5: summary of the third cluster

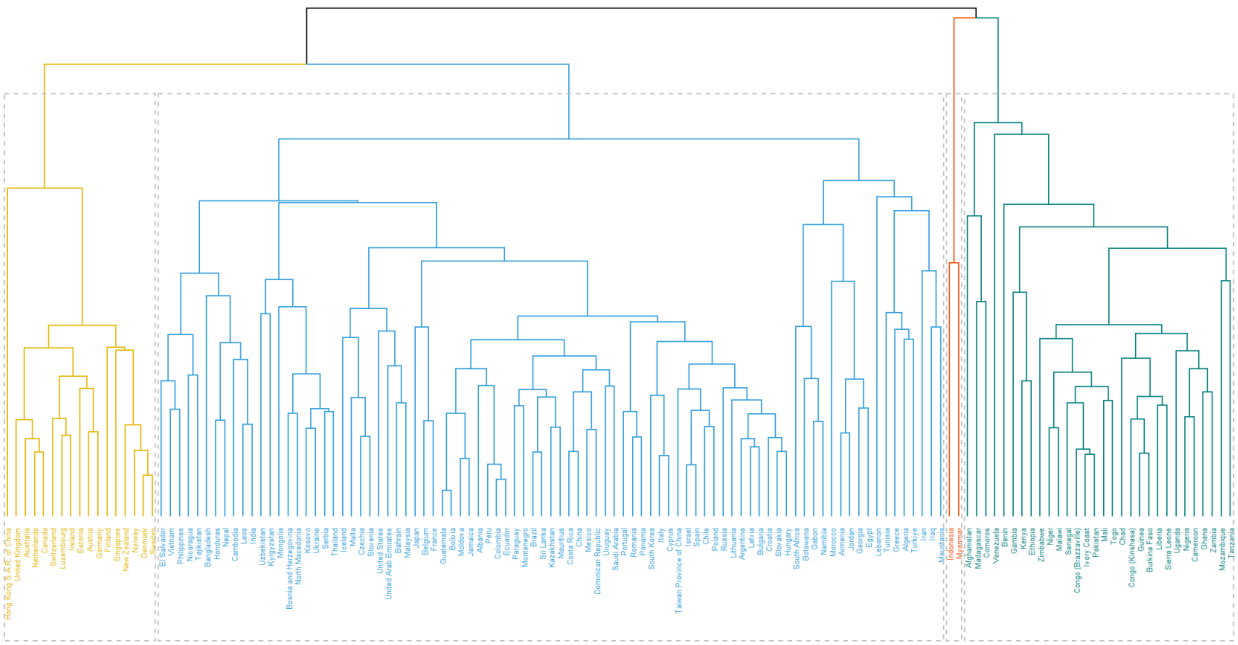
gdp	social_supp	life_exp	freedom	generosity	corruption
Min. :5.527	Min. :0.3410	Min. :51.53	Min. :0.3820	Min. :-0.23100	Min. :0.5540
1st Qu.:7.643	1st Qu.:0.5907	1st Qu.:55.42	1st Qu.:0.6590	1st Qu.:-0.01100	1st Qu.:0.7405
Median :8.085	Median :0.6495	Median :57.67	Median :0.7155	Median : 0.03950	Median :0.7880
Mean :8.070	Mean :0.6437	Mean :58.07	Mean :0.7043	Mean : 0.05562	Mean :0.7831
3rd Qu.:8.587	3rd Qu.:0.7120	3rd Qu.:60.36	3rd Qu.:0.7705	3rd Qu.: 0.12225	3rd Qu.:0.8460
Max. :9.629	Max. :0.8390	Max. :66.15	Max. :0.9190	Max. : 0.49100	Max. :0.9110

4 - Hierarchical clustering

The hierarchical classification model generates a family of partitions of the n units starting from the one in which all the units are distinct, subsequently obtaining those with n-1, n-2... to reach up to the one in which all the units are reunited in a single group. The first step is to create a distance matrix, that is, a double entry matrix in which for each pair of observations the distance between them is calculated by standardizing the data. This matrix is calculated to observe the similarity or dissimilarity between observations. The distance matrix is used to determine the dendrogram, which represents the cluster hierarchy. We will consider three different types to measure the distance between groups, namely the average bond method, the complete method and the Ward method. In the average bond method the distance between two groups is defined as the arithmetic mean of the n1, n2 distances between each of the units of one group and each of the units of the other group:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{r=1}^{m} \sum_{s>r}^{m} d_{rs}$$

Fig. 3.1: hierarchical cluster with Average linkage

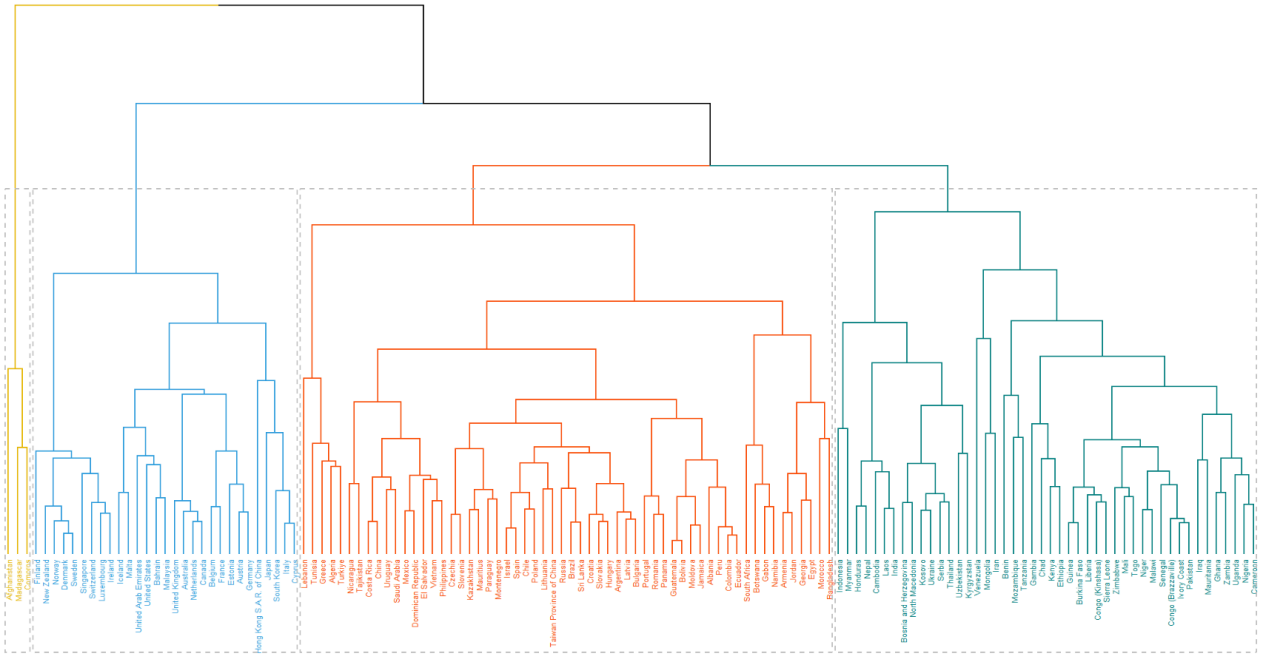


The graph does not have a uniform distribution as the four clusters have 17, 87, 2 and 30 observations respectively. The grouping on the left is made up of countries that have higher values such as GDP per capita. The blue group has a significantly larger observation number than the others.A further method that was used in the analysis is the calculation of hierarchical clustering with complete linkage mode. In this approach we rely on the maximum distance between individual points in the different clusters. The method results in shape groups that resemble a sphere. In this dendrogram the distance between the most distant pairs of observations is maximum.

$$d(C_1, C_2) = \max(d_{rs})$$

The distribution of this graph is more uniform than the previous one. This more homogeneous distribution is highlighted by the quantity of observation of the different clusters. Specifically, the four graphs are composed of 29, 58, 46 and 3 observations, respectively.

Fig. 3.2: hierarchical cluster with Complete linkage



Finally, to complete the hierarchical clustering procedure, the last distance taken into consideration is the Ward distance. With Ward's method, the distances associated with all possible groupings are calculated and the aggregation with minimum deviance is carried out. The distance between the two different groups is calculated through the difference between the overall deviance and the sum of the internal deviances of the two groups. Therefore, at each step those groups for which the deviance within the groups increases to a lesser extent are aggregated together.

$$d(C_1, C_2) = \frac{n_1 n_2}{n_1 + n_2} \|\overline{x_1} - \overline{x_2}\|^2$$

The graphical representation of the dendrogram returns a more homogeneous distribution of observations compared to previously used methods. In fact, the clusters are formed by 16, 68, 35 and 17 observations respectively. The dendrogram is shown below.

Fig. 3.3: hierarchical cluster with Ward linkage

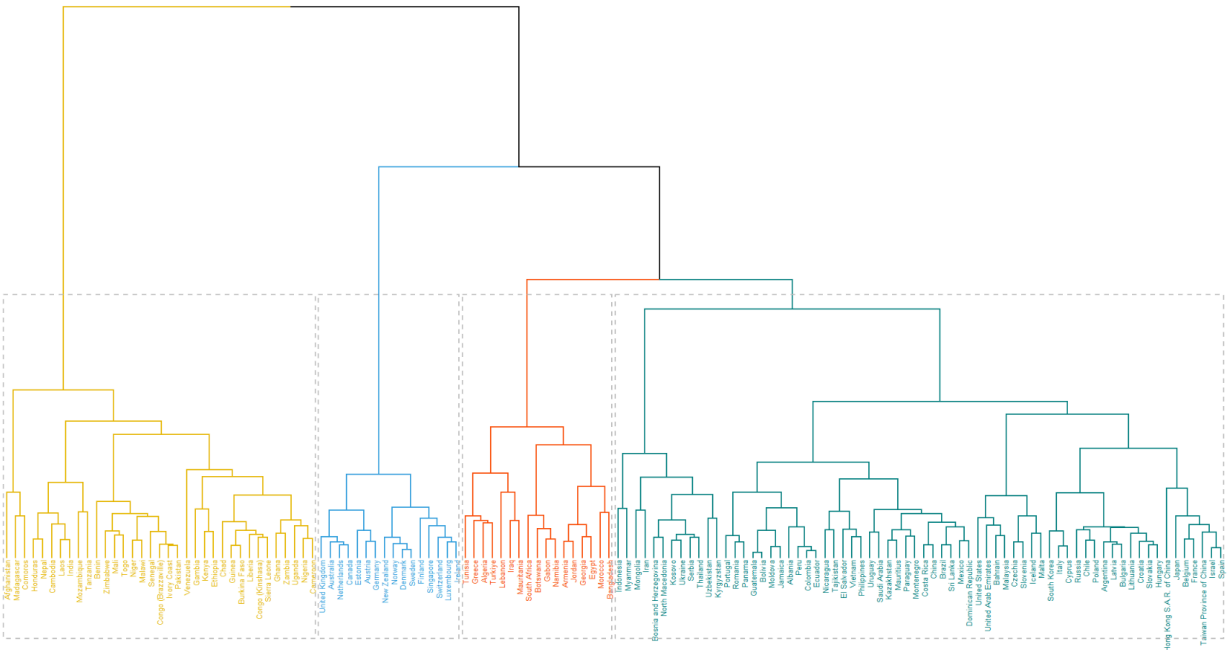
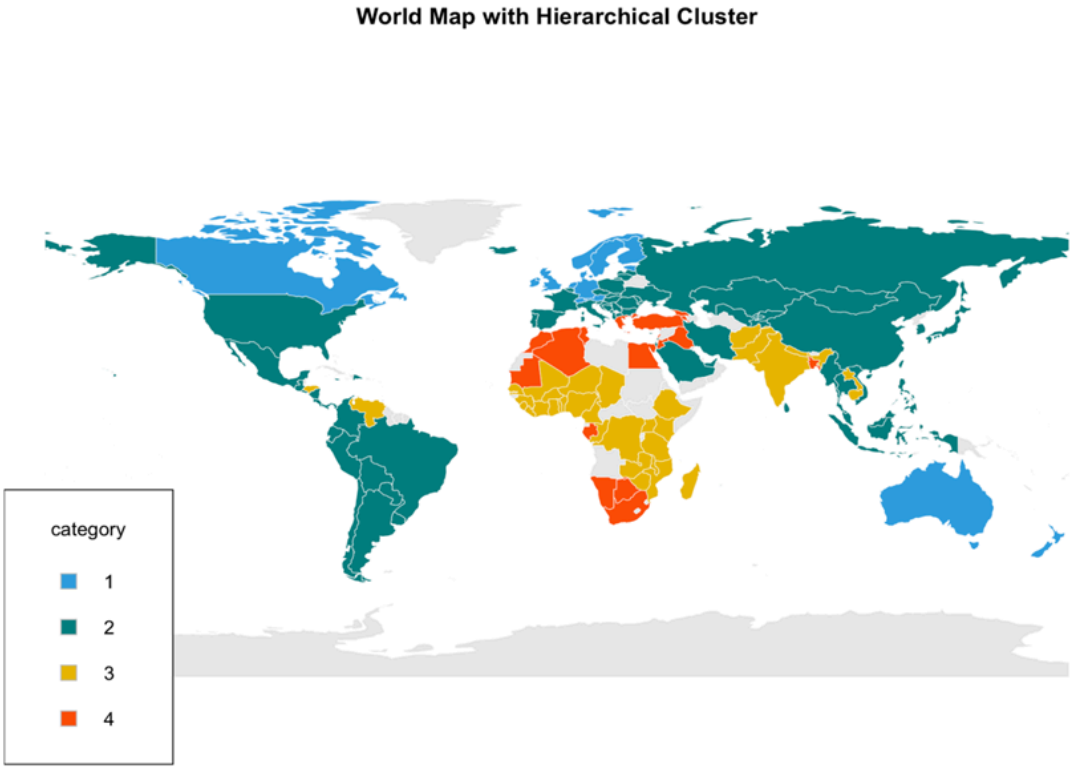


Fig. 3.4: world map with the states classified into clusters using hierarchical clustering method



The graphical display presents a notable improvement, featuring a more uniform distribution of observations. Consequently, for the continuation of the hierarchical cluster analysis, the groupings formed using Ward's method will be taken into account. For a better graphic view to better understand the distribution of the clusters, a geographical map of the world is shown with the respective clusters divided by color.

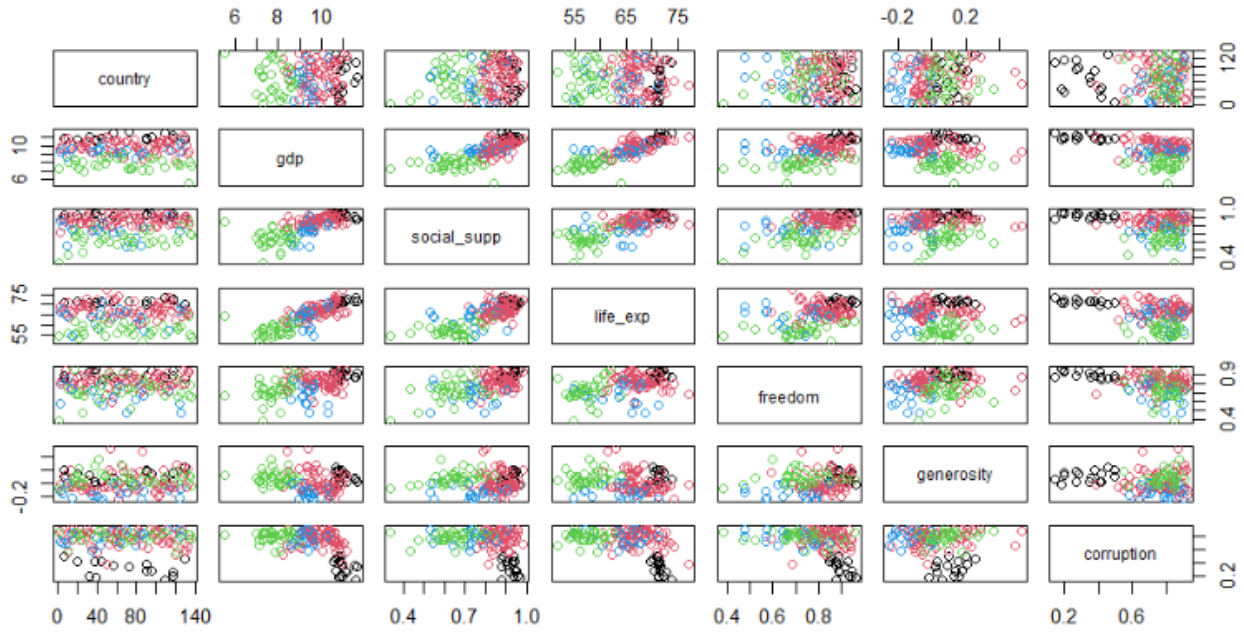
From the list of the first cluster it can be observed that they are mainly Northern European countries and some non-European ones including New Zealand, Australia, Canada and Singapore. The geographical similarity could reflect some similarities for the analysis of happiness across countries. Furthermore, these countries have on average a significantly higher GDP per capita than others belonging to other groupings. The same thing can be said for the variables regarding social support, the level of healthy life expectancy, freedom of making life choices and the level of generosity. On the contrary, the level of corruption in the countries is significantly lower than in the countries of the other clusters. Cluster number 2 instead includes a mix of countries geographically located throughout the world. Subsequently, cluster number 3 includes countries that are located in the southern part of the world such as Africa, Latin America and Asia. Finally, the last cluster (colored yellow) includes countries located mainly in Africa and South Asia. These countries have significantly low values of the variables taken into consideration.

It is clear from the table below that the level of Gross Domestic Product (GDP) per capita is significantly higher in the first cluster, followed by the second, the fourth and finally the third cluster. This pattern is also found for other variables such as social support and life expectancy, where the first cluster shows higher values than the others. It is noteworthy that the first cluster is the most prosperous and shows the highest values for the variables considered. Next, the second cluster is observed. However, it is interesting to note that the values of the variables are reversed for the third and fourth clusters.

Group.1	gdp	social_supp	life_exp	freedom	generosity	corruption
1	10.991750	0.9215000	71.40469	0.8980000	0.11143750	0.3327500
2	9.942706	0.8680294	67.58069	0.8257500	0.01298529	0.7728235
3	7.841343	0.6424286	57.57506	0.7208286	0.08365714	0.7872571
4	9.381588	0.7270000	63.67700	0.6724118	-0.14052941	0.7714118

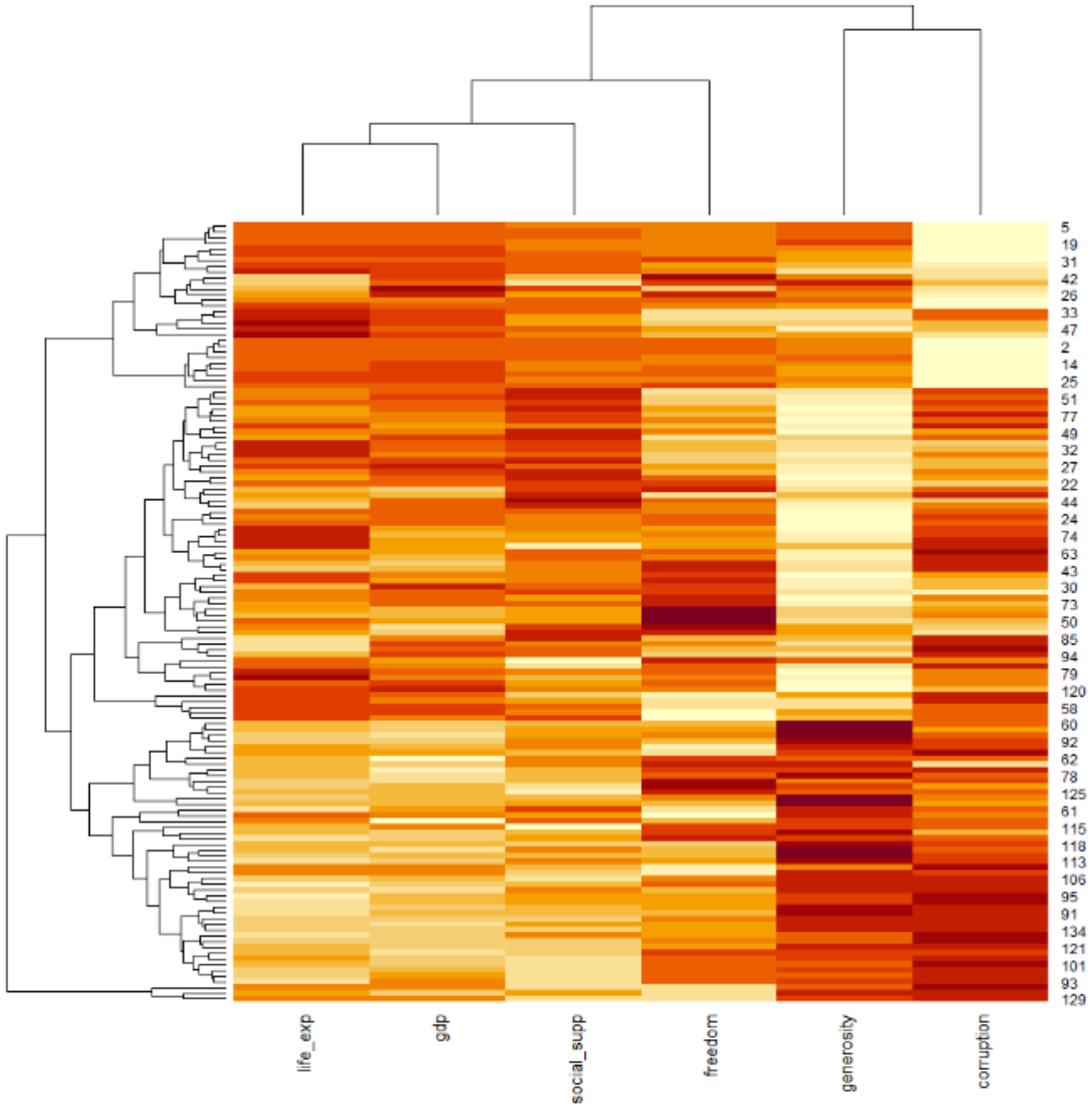
Subsequently, to deepen the analysis, a scatter plot is created. On the main diagonal you can observe the variables taken into consideration for the study. Inside each box, the scatter plot for the pairs of variables is observed to more quickly observe the relationships between them.

Fig. 3.5: scatterplot using Ward linkage



To optimize the clarity of the visualization, the observations in the scatterplot are differentiated using colors based on the cluster assignment, obtained through the Ward method in the hierarchical clustering process. The use of colors allows you to easily identify observations belonging to the same cluster, making the proximity between them within each group evident. This phenomenon is indicative of high cohesion within the clusters, suggesting that the Ward distance metric effectively grouped observations based on their similarities.

Fig. 3.6: heatmap

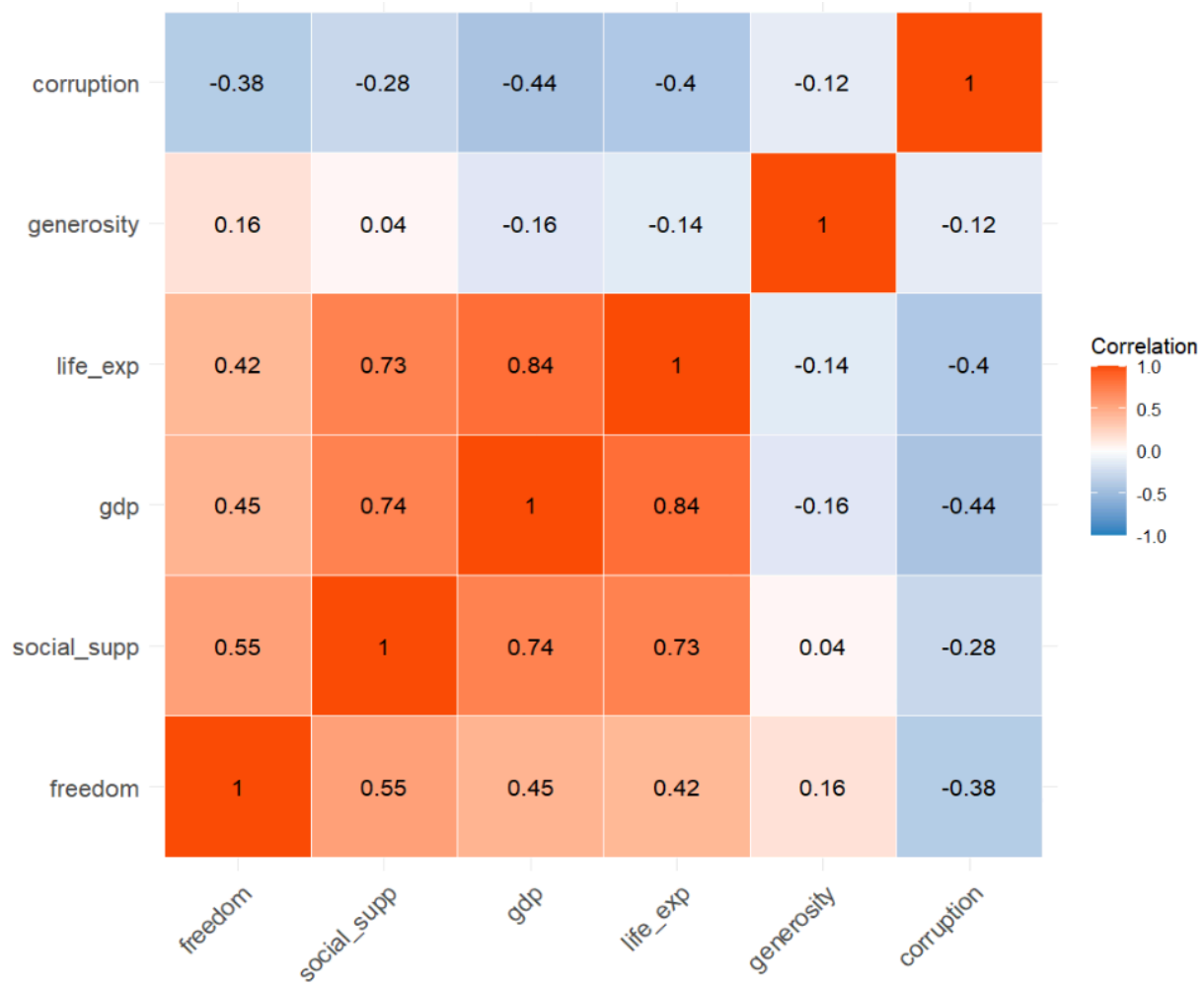


Next, a heatmap of the standardized data is generated. In this graphical representation, each row corresponds to a country, while each column represents a specific variable. The intensity of the color in each rectangle reflects the degree to which the associated value is above or below the average. It follows that a rectangle with a more intense color indicates a value significantly higher than average. For the first four variables, a similar trend is observed: countries with a higher level of happiness show higher values in these variables. On the contrary, regarding the level of corruption, more intense colors are highlighted for countries characterized by a lower level of happiness. A particular case emerges for the generosity variable, in which happier countries present a moderate value of generosity, while those with a lower level of happiness show higher values of generosity. This observation represents an exception compared to the general trend of the other variables. The graph is shown below.

4 - Linear Regression

4.1 - Expectation

Looking at the cluster analysis as well as at the k-means we noticed how the states were divided into 4 clusters and we can see how the richest states are included in the first cluster. Therefore it is reasonable to think that the gdp variable positively affects happiness. The same thing will happen with the variables 'social support' and 'life expectancy'. Then looking at the correlation plot we notice how the gdp variable is strongly correlated with many variables, and to avoid multicollinearity problems we proceed to remove it from our regression model. Similarly we remove the life expectancy variable which is strongly correlated with the social support variable



4.2 - Analysis

The analysis is continued by working out a multiple linear regression model, taking into consideration the level of happiness as the dependent variable, defined as y, which is influenced

by a number of independent factors. The independent variables are the gdp per capita, social support, the life expectancy index, the level of freedom, of generosity, of corruption and finally the clustering that was carried out in the first part of this study by using Ward's method. The initial output obtained for multiple linear regression is as follows. As we can see and as we have previously mentioned, the presence of highly correlated variables such as 'gdp' and 'life expectancy' makes the model not very significant. Therefore we can proceed with removing these last two variables and see how the model comes out

```
Call:
lm(formula = happiness ~ gdp + life_exp + social_supp + freedom +
    generosity + corruption + cluster, data = happiness_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30510 -0.19680  0.01897  0.26772  0.95222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.34921    1.15239  -2.039  0.04359 *
gdp          0.24241    0.07554   3.209  0.00169 **
life_exp     0.02139    0.01449   1.477  0.14219
social_supp  3.98488    0.58413   6.822 3.36e-10 ***
freedom      1.99937    0.48606   4.113 6.99e-05 ***
generosity  -0.33621    0.34960  -0.962  0.33805
corruption  -0.46592    0.40837  -1.141  0.25607
cluster2    -0.23392    0.23178  -1.009  0.31480
cluster3    -0.04205    0.33169  -0.127  0.89932
cluster4    -0.58541    0.30160  -1.941  0.05449 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4706 on 126 degrees of freedom
Multiple R-squared:  0.8417,    Adjusted R-squared:  0.8304
F-statistic: 74.45 on 9 and 126 DF,  p-value: < 2.2e-16
```

The first variable to be eliminated is the one with the highest p-value, namely ‘gdp’.

```
Call:
lm(formula = happiness ~ life_exp + social_supp + freedom + generosity +
    corruption + cluster, data = happiness_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30251 -0.21926  0.00313  0.29014  1.01767

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.29531    1.14432  -1.132  0.25979
life_exp     0.03939    0.01384   2.847  0.00515 **
social_supp  4.33856    0.59427   7.301 2.77e-11 ***
freedom      2.05031    0.50327   4.074 8.09e-05 ***
generosity  -0.54808    0.35566  -1.541  0.12580
corruption  -0.53406    0.42248  -1.264  0.20851
cluster2    -0.38767    0.23493  -1.650  0.10138
cluster3    -0.42400    0.32073  -1.322  0.18854
cluster4    -0.77985    0.30607  -2.548  0.01203 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4876 on 127 degrees of freedom
Multiple R-squared:  0.8288,    Adjusted R-squared:  0.818
F-statistic: 76.84 on 8 and 127 DF,  p-value: < 2.2e-16
```

Looking at the result obtained we notice how the variables are more statistically significant, as we have eliminated one of the variables most correlated with the others. We can therefore proceed and eliminate another variable that is strongly correlated, namely 'life expectation'.

```
Call:
lm(formula = happiness ~ social_supp + freedom + generosity +
    corruption + cluster, data = happiness_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4627 -0.2332  0.0491  0.3005  1.0558

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3870     0.6673   2.078  0.03966 *
social_supp   4.7417     0.5930   7.997 6.57e-13 ***
freedom       1.8825     0.5135   3.666  0.00036 ***
generosity   -0.7460     0.3583  -2.082  0.03935 *
corruption   -0.7390     0.4277  -1.728  0.08644 .
cluster2     -0.4582     0.2400  -1.909  0.05850 .
cluster3     -0.7984     0.3005  -2.656  0.00890 **
cluster4     -1.0037     0.3039  -3.303  0.00124 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5009 on 128 degrees of freedom
Multiple R-squared:  0.8179,    Adjusted R-squared:  0.8079
F-statistic:  82.1 on 7 and 128 DF,  p-value: < 2.2e-16
```

Proceeding in this way we obtain the final output. The model, therefore, tries to explain the variation in happiness using only the significant variables (‘social_supp’, 'freedom’, 'generosity’, ‘corruption’ and the cluster variable). Cluster number 1 consists of the countries located mainly in Northern Europe, especially the Scandinavian countries. The other clusters, on the other hand, have a more dispersed geolocation than the first one.

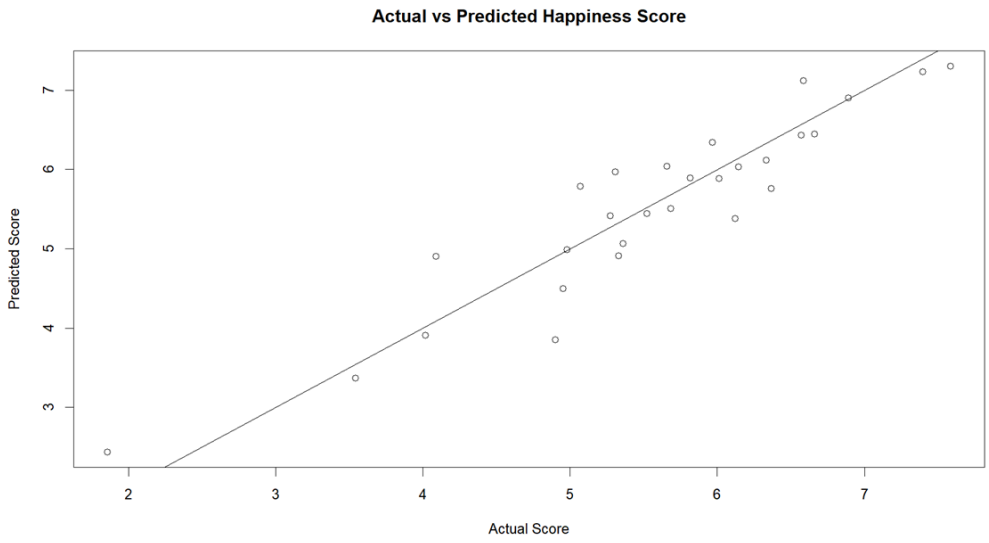
The method used to conduct multiple linear regression is known as stepwise regression. This technique is used to optimize the selection of predictor variables through a stepwise process of eliminating the less significant ones. By progressively eliminating the variable with the highest p-value, an increase in the coefficients of the remaining variables can be observed. This occurs because the remaining variables must now explain more variation in the model, as the less significant ones have been removed. In essence, stepwise regression helps to simplify the model by focusing only on those variables that contribute significantly to explaining the variation in the dependent variable.

Using these results, evaluations can be made. For example an increase of one unit in social support is associated with an increase of 4.74 points in the happiness ratio on average and keeping the other variables constant. After that, it can be stated that a one-unit increase in freedom increases the happiness value by 1.88, on average and keeping all the variables as constant. An increase of one unit in the level of generosity is associated with a decrease of approximately 0.74 on average in the independent variable, ceteris paribus. Also, looking at the

variable ‘ corruption’ we can say that when corruption increases, happiness decreases by 0.74 points on average.

Finally, looking at the final cluster we can say that cluster 1 is taken as the reference variable. These coefficients represent the differences in happiness levels between each cluster and the reference cluster, controlling for the effects of all other variables in the model. The second cluster has a lower happiness index of about 0.45 compared to the reference cluster, the third cluster has a lower happiness index of about 0.79 and, finally, the last cluster has a lower happiness index of about 1.00. These coefficients provide information on the differences in happiness between the different country clusters compared to the reference one. Looking at the regression, therefore, we can state that the blue cluster, i.e. the one attributed to the richest and most socially developed countries, is happier than the remaining clusters. Obviously this aspect was largely predictable. It is important to note that the level of happiness is much greater for these clusters, as the value of happiness is measured on a scale between 0 and 1, a difference of 0.45 compared to the second cluster is economically significant. To conclude, let's consider the last cluster, i.e. the one referring to poor countries but which we do not classify as third world countries. In this case the difference with the first cluster is very high, because it has a value of -0.655, i.e. the countries belonging to this cluster have on average 1.00 points of happiness less than the happiest countries of the first cluster under equal conditions. In this case there is a statistically significant value and it can therefore be concluded that the difference between states belonging to different clusters is very marked. Furthermore, we can affirm that the social policies that each state proposes is the most important one, since the variable relating to social support is the one with the greatest magnitude. Next we find freedom in choices for each individual which has a very high coefficient of 1.88, to express how freedom is another aspect that must be highly considered to improve the happiness of a population. Furthermore, it can be said that the model explains about 83.48% of the variation in happiness by the multiple R-squared.

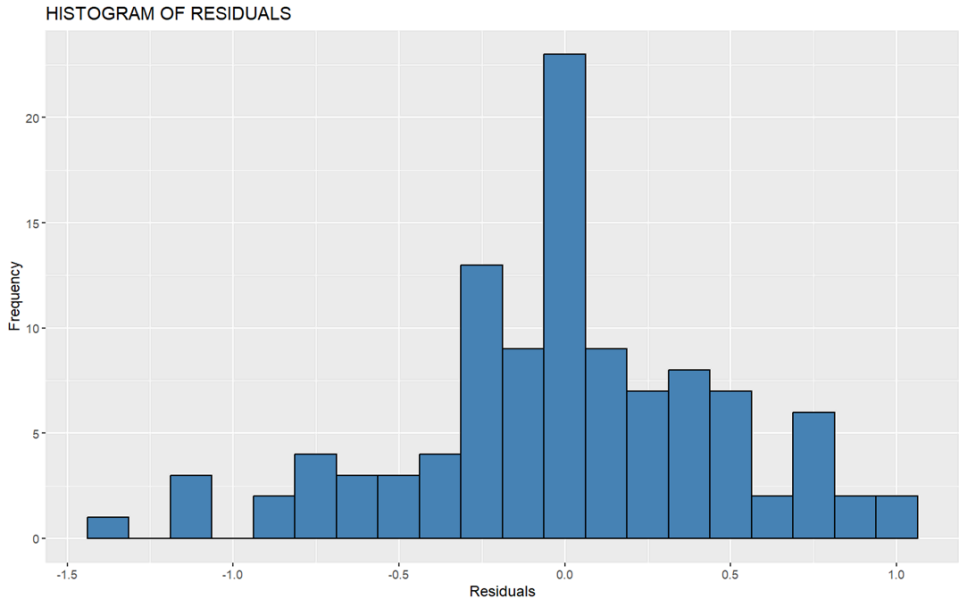
4.3 - Prediction



In order to evaluate whether your regression model is performing good or bad, a predicted vs Actual score plot can help us visualize the performance of a regression model more straight forward compared to other methods. For our case the x-axis will represent the actual happiness score while the y-axis will represent the predicted happiness score. Ideally, if the predictions are perfect, the points will lie along a straight line with a slope of 1.

Residual Plot

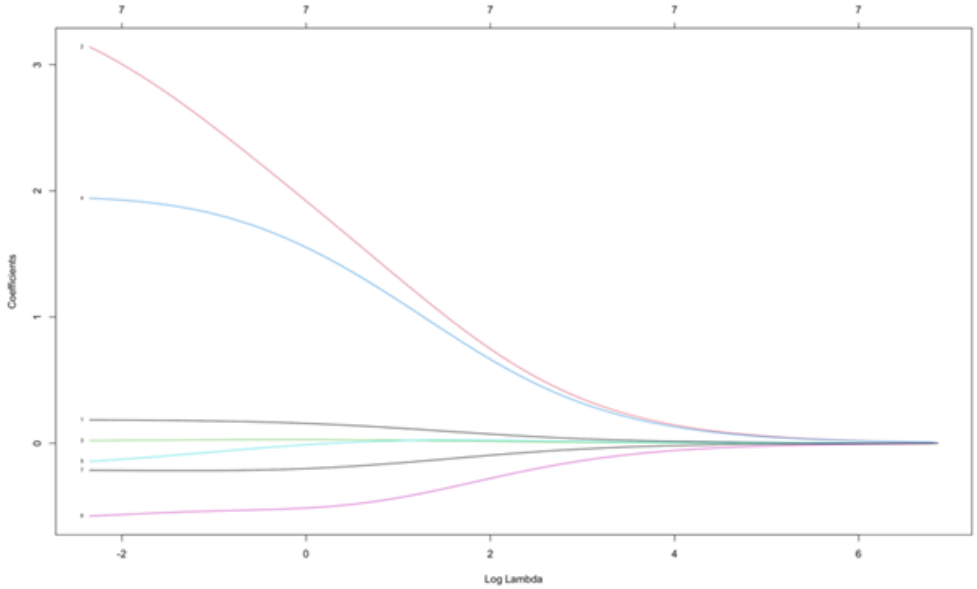
One of the main assumptions of an ideal linear regression is that the residuals are normally distributed with a mean of 0. The histogram below shows a roughly symmetric distribution centered around 0, indicating that the residuals are normally distributed.



RIDGE AND LASSO REGRESSION

Standard OLS linear regression is based on assumptions such as minimising the sum of squares of the errors between predicted and observed values. There are two more advanced regression techniques that are based on regularisation and penalty to OLS coefficients: Ridge and Lasso regression. These two modes limit the complexity of the model by affecting the coefficients of the variables.

In Ridge regularisation, a penalty term proportional to the square of the absolute values of the model coefficients is added to keep the coefficients small but non-zero. This is done to make the model more stable and to prevent overfitting. In our case, the optimal regularisation parameter λ (lambda) is 0.1261445. The more λ tends to 0, the more the coefficients are the same as in linear regression. Conversely, the more λ increases, the more the coefficients tend towards 0. This can be observed from the graph below.



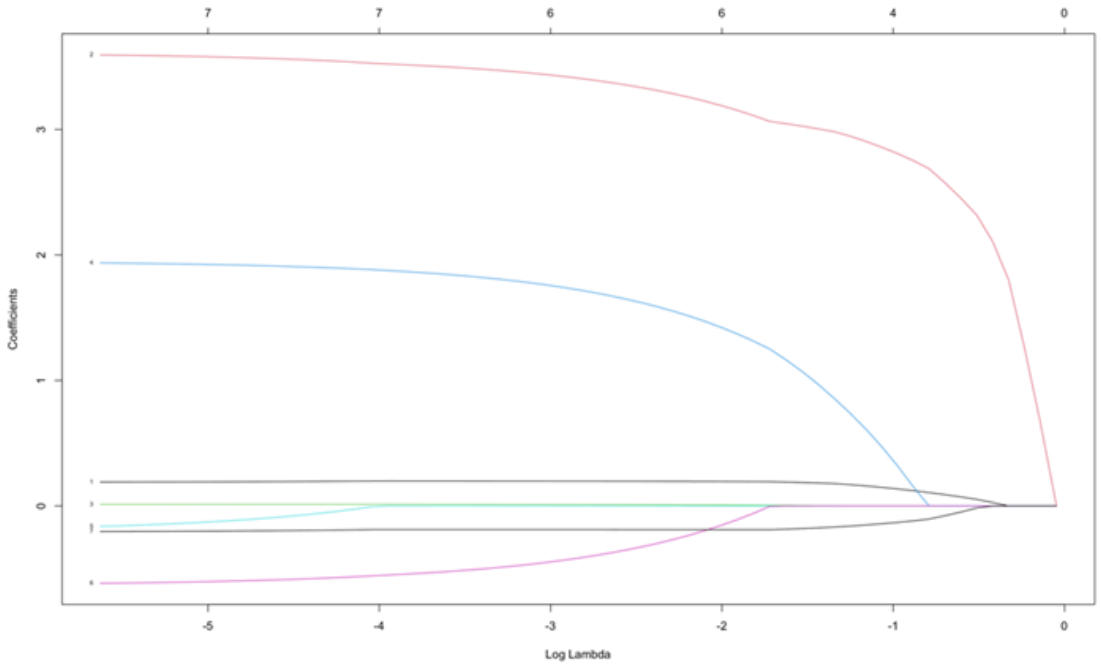
For the Ridge regression, the coefficients are as shown below.

```
8 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -0.58601553
gdp          0.18336608
social_supp  3.03212164
life_exp     0.02132172
freedom      1.93052120
generosity  -0.13211830
corruption  -0.56810017
cluster     -0.21623226
```

A one-unit increase in GDP per capita is associated with an increase of about 0.18336608 in the happiness index, holding the other variables constant. A one-unit increase in social support is associated with an increase of about 3.03212164 in the happiness index, holding other variables constant. a one-unit increase in life expectancy is associated with an increase of about 0.02132172 in the happiness index,

holding other variables constant. a one-unit increase in freedom is associated with an increase of about 1. 93052120 in the happiness index, holding the other variables constant. a one-unit increase in generosity is associated with a decrease of about 0.13211830 in the happiness index, holding the other variables constant. a one-unit increase in corruption is associated with a decrease of about 0.56810017 in the happiness index, holding the other variables constant. For clusters, the coefficient is approximately -0.21623226. This represents the contribution of the cluster to the variability of the happiness index, holding the other variables in the model constant.

In contrast, a penalty term proportional to the absolute value of the model's coefficients is added for Lasso regression. Unlike Ridge regularisation, Lasso regularisation can cause some coefficients to become exactly zero, making the model more interpretable and automatically selecting a subset of the variables. In our case, the value of λ is 0.02534538. This value is relatively low, thus having coefficients subject to strong regularisation. Thus, the coefficients tend to be reduced to 0 compared to the Ridge. The graph to observe λ is as follows.



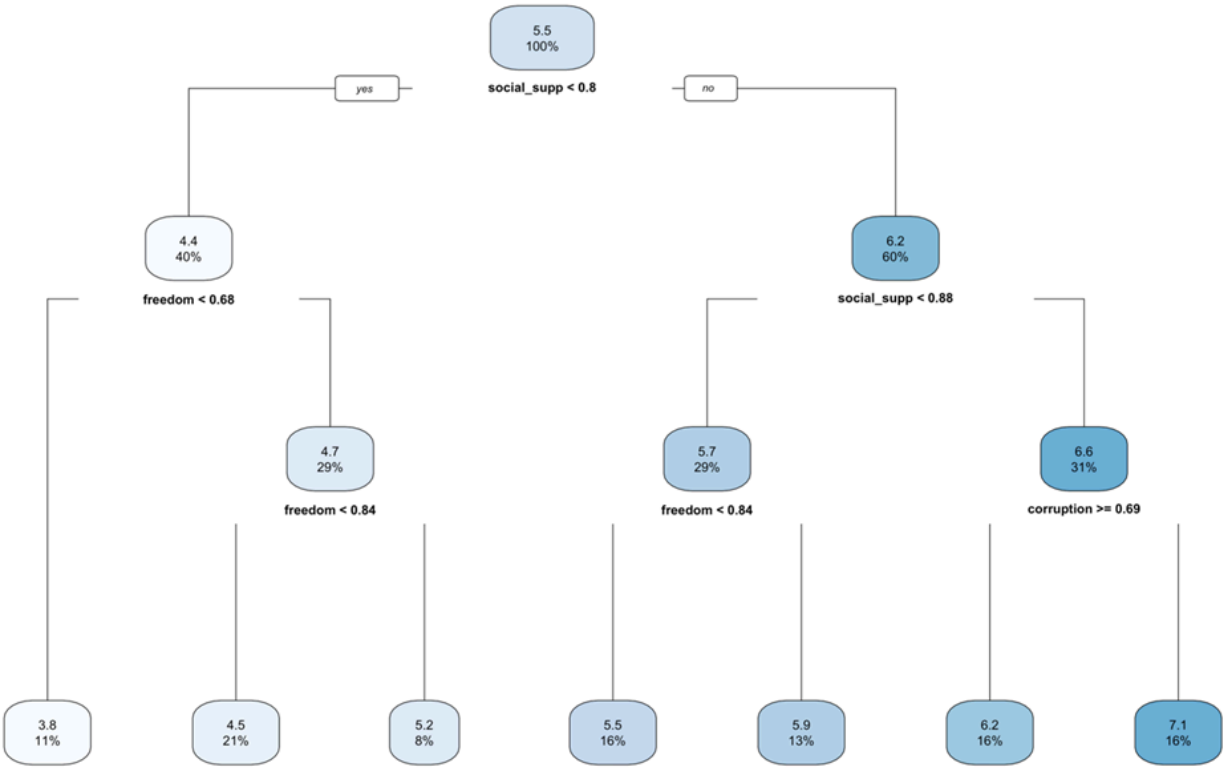

```
8 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -0.59956374
gdp          0.19871121
social_supp  3.50201462
life_exp     0.01296036
freedom      1.85286613
generosity   .
corruption  -0.52939785
cluster     -0.18842764
```

Holding all other variables constant, a one-unit increase in gdp is associated with an increase of approximately 0.199 in happiness expectancy. Holding all other variables fixed, a one-unit increase in 'social_supp' is associated with an increase of about 3.502 in happiness expectancy. holding all other variables fixed, a one-unit

increase in life expectancy is associated with an increase of about 0.013 in happiness expectancy. holding all other variables fixed, a one-unit increase in level of freedom is associated with an increase of about 1.853 in happiness expectancy. The coefficient for generosity is ".", which indicates that the coefficient value for this variable was adjusted to zero by the Lasso model. Therefore, this variable did not contribute significantly to the prediction of happiness and was excluded from the model. Holding all other variables fixed, an increase of one corruption unit is associated with a decrease of approximately 0.529 in happiness expectancy. In contrast, the coefficient of the cluster variable indicates that holding all other variables fixed, a one-unit increase in the cluster variable is associated with a decrease of about 0.188 in the expectation of happiness.

4.4 - Decision Trees

The objective of this methodology is to obtain a hierarchical segmentation of a set of units, sometimes very large, through the identification of 'rules' that exploit the relationship existing between the class to which they belong and the variables detected for each unit. The graphic output of the procedure consists of a tree structure - with nodes, branches and leaves - which has points of contact with the dendrogram of cluster analysis. The application of decision trees, on the other hand, requires the a priori knowledge of the class to which each unit belongs: the aim of the technique is in fact to identify the optimal decision rule, i.e. the rule that, given a certain set of surveyed variables, best predicts the class to which each unit belongs. To achieve this objective, the classes must obviously be known a priori. What then is the advantage of segmenting a set of units whose group they already belong to? The advantage lies in the fact that the segmentation 'rules' identified in this way can also be easily applied to different units than those that make up the starting data set and for which the group they belong to is instead unknown. Decision trees therefore belong to the class of so-called supervised classification techniques, as segmentation can take advantage of additional information on the membership group that is known for a small number of units. The set of units on which the tree is determined, and for which the membership group is given, is also called the training set or training sample.



The analysis proceeds with decision tree analysis, a supervised predictive model that is very easy to interpret and can be used for regression cases. The output that is obtained goes on to

predict the continuous dependent variable, happiness. The data is split according to certain criteria. The graph consists of nodes and leaves. The nodes are the places where the data is split and is associated with an input value. Leaves, on the other hand, are the intermediate or final results and are associated with an output value. The leaves display the data once it has been split. In addition, a training subset consisting of 90 countries was used to create the graph. The output we obtain is as follows.

Looking at the regression tree graph, it is clear that the significant and relevant variables to explain the level of happiness of countries are 'social support', 'freedom' and 'corruption'. This indicates that the degree of happiness of a country may depend on the level of social support, freedom and corruption. Each leaf of the tree represents the predictive value of the response variable, i.e. the level of happiness, together with the percentage of total observations in the dataset that fall into that specific leaf and fulfil the specified conditions.

Starting from the left side of the tree, we find about 10 countries with a 'social support' of less than 0.8 and a 'freedom' level of less than 0.68, which have an average happiness of about 3.8, representing about 11% of the observations in the training dataset. Proceeding to the right in the regression tree, we see that about 19 countries have a happiness level of 4.5. These countries share a 'social support' of less than 0.8 and a 'freedom' level between 0.68 and 0.84. Furthermore, about 7 countries have an average happiness of 5.2, with a 'social support' of less than 0.8 and a 'freedom' level of more than 0.84. Next, some 14 countries with a 'social support' between 0.8 and 0.88 and a 'freedom' level below 0.84 show an average happiness of 5.5. Similarly, with the same level of 'social support' but a level of 'freedom' greater than 0.84, there are 11 countries with an average happiness of 5.9. With 'social support' greater than 0.88, but 'corruption' greater than or equal to 0.69, we find some 14 countries with an average happiness of 6.2. Finally, the highest happiness level of 7.1 is observed in another 14 or so countries with a 'social support' greater than 0.88 and 'corruption' less than 0.69.