Multivariate Analysis Report

Diego Quintana

Marcel Pons

Manuel Breve

MVA Course
Master in Innovation and Research in Informatics

Universitat Politècnica de Catalunya

Facultat d'Informatica de Barcelona

- June 2020 -

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

FIB

Table of Contents

World Happiness Dataset

The World Happiness Report is a landmark survey of the state of global happiness that ranks, through a Happiness Score, 156 countries by how happy their citizens perceive themselves to be.

In this project, we analyze the World Happiness Report for the years 2018 and 2019 through multivariate techniques. First, we describe the 2018 dataset, then we analyze it with dimensionality reduction and clustering techniques, and finally, we use decision trees and random forest in order to create a model and predict the 2019 Happiness Score, that will be contrasted with the real scores.

Variables Description

The World Happiness Dataset is made of one (1) continuous response variable, six (6) continuous independent variables, and one (1) supplementary categorical variable.

Continuous Response Variable

1. **Score:** This Happiness Score is a subjective well-being perception based on the survey's answers, which ask people to evaluate different subjects about their life quality on a scale of 0 to 10.

Continuous Independent Variables:

Among the survey variables, we have selected the following six continuous variables in order to make the multivariate analysis:

1. **GDP per capita**: in terms of Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank. (Using natural log of GDP per capita, as this form fits the data significantly better than GDP per capita)
2. **Social support:** the national average of the binary response (either 0 or 1) to the Gallup World Poll (GWP) question "If you are in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?
3. **Healthy life expectancy at birth:** constructed from the World Health Organization (WHO) and WDI. Adjustment by applying the country-specific ratios to other years.
4. **Freedom to make life choices:** the national average of binary response to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
5. **Generosity:** residual of regressing the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.
6. **Perceptions of corruption:** the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption

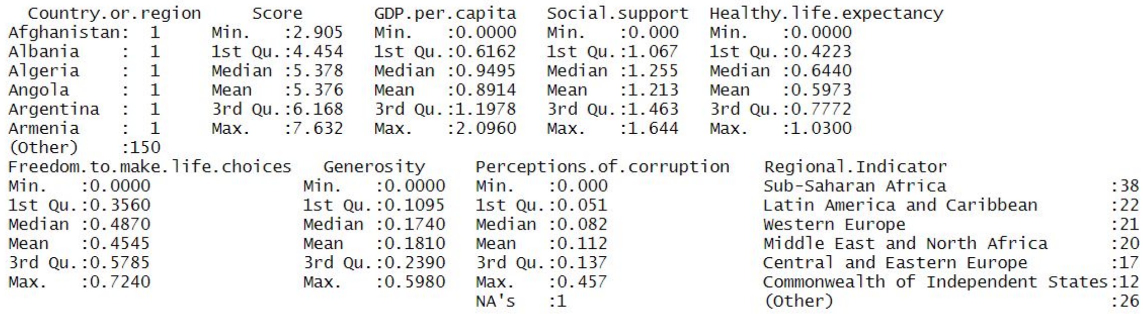
widespread within businesses or not?”. Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.

Supplementary Categorical Variable:

1. **Regional Indicator:** In order to enrich and improve our analysis, we’ve added an extra categorical variable for each of the 156 individuals, i.e. countries, which is its Regional Indicator, specifying their World Region. This variable has 10 levels, which are: Western Europe, Central and Eastern Europe, Sub-Saharan Africa, Middle East and North Africa, East Asia, Southeast Asia, South Asia, North America and ANZ, Latin America, Caribbean and Commonwealth of Independent States.

Dataset Description

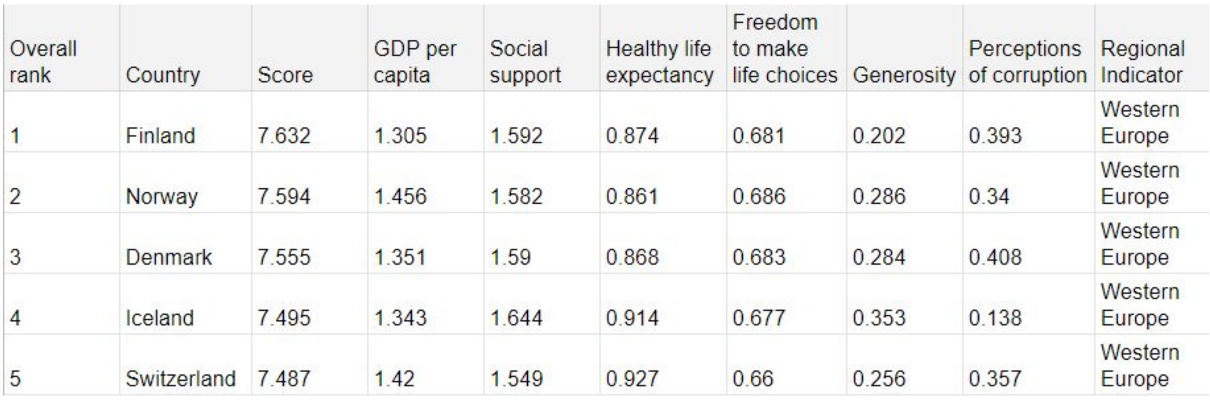
A basic description of the dataset is provided using the `summary` function in R



```
   Country.or.region      Score        GDP.per.capita    Social.support   Healthy.life.expectancy
 Afghanistan:  1     Min.   :2.905   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 Albania    :  1     1st Qu.:4.454   1st Qu.:0.6162   1st Qu.:1.067   1st Qu.:0.4223
 Algeria    :  1     Median :5.378   Median :0.9495   Median :1.255   Median :0.6440
 Angola     :  1     Mean   :5.376   Mean   :0.8914   Mean   :1.213   Mean   :0.5973
 Argentina  :  1     3rd Qu.:6.168   3rd Qu.:1.1978   3rd Qu.:1.463   3rd Qu.:0.7772
 Armenia    :  1     Max.   :7.632   Max.   :2.0960   Max.   :1.644   Max.   :1.0300
 (Other)    :150
 Freedom.to.make.life.choices   Generosity      Perceptions.of.corruption     Regional.Indicator
 Min.   :0.0000               Min.   :0.0000   Min.   :0.000             Sub-Saharan Africa              :38
 1st Qu.:0.3560               1st Qu.:0.1095   1st Qu.:0.051             Latin America and Caribbean     :22
 Median :0.4870               Median :0.1740   Median :0.082             Western Europe                  :21
 Mean   :0.4545               Mean   :0.1810   Mean   :0.112             Middle East and North Africa    :20
 3rd Qu.:0.5785               3rd Qu.:0.2390   3rd Qu.:0.137             Central and Eastern Europe      :17
 Max.   :0.7240               Max.   :0.5980   Max.   :0.457             Commonwealth of Independent States:12
                                               NA's   :1                 (Other)                         :26
```

Summary of the 2018 World Happiness Report Dataset

A sample of the 2018 World Happiness Dataset can be found next:

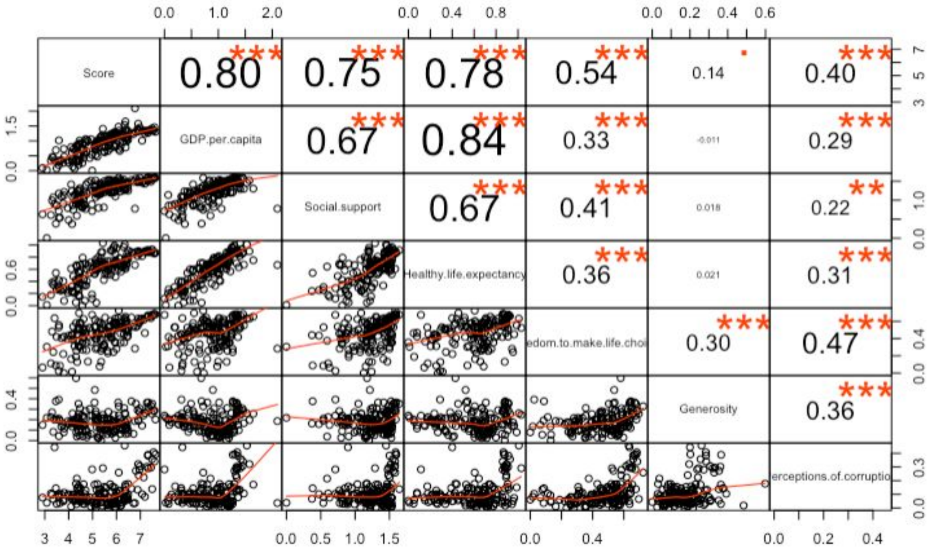


Overall rank	Country	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Regional Indicator
1	Finland	7.632	1.305	1.592	0.874	0.681	0.202	0.393	Western Europe
2	Norway	7.594	1.456	1.582	0.861	0.686	0.286	0.34	Western Europe
3	Denmark	7.555	1.351	1.59	0.868	0.683	0.284	0.408	Western Europe
4	Iceland	7.495	1.343	1.644	0.914	0.677	0.353	0.138	Western Europe
5	Switzerland	7.487	1.42	1.549	0.927	0.66	0.256	0.357	Western Europe

Sample of the 2018 World Happiness Report Dataset

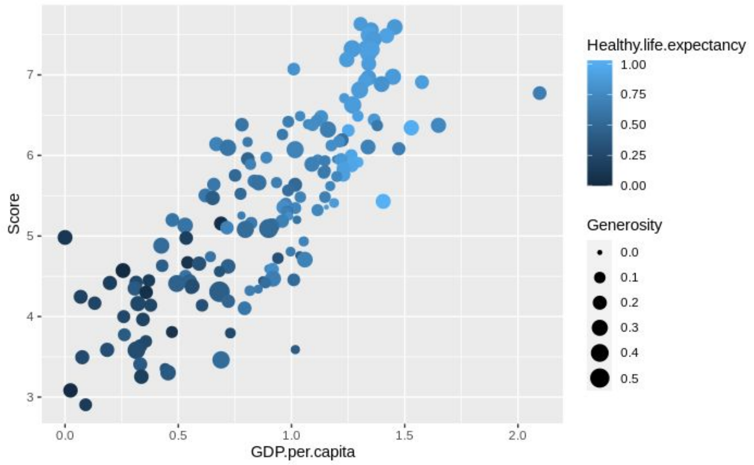
When looking for the correlations of the variables, we can observe that nearly all variables are correlated with the Happiness Score, being GDP per capita, Health life expectancy and Social support the most correlated ones.

Furthermore, most of the variables are kind of correlated with the others, except for the Generosity variable, which is not much correlated with any other variable.



Plot of correlations between the variables in the 2018 happiness report dataset

We observe that GPD is highly correlated with the Happiness Score obtained and the expectancy of healthy life. In other words, the happiest countries tend to be those with high GDP values and a higher life expectancy, however, this is not the case with the Generosity variable. We will see later how these variables behave using a Principal Component Analysis.



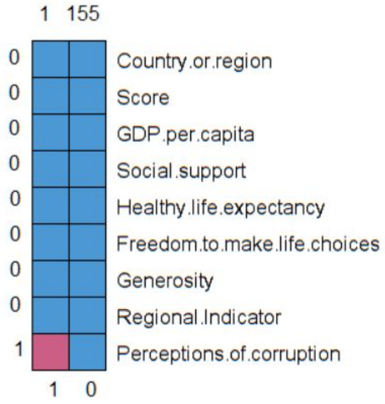
Plot of the GDP versus the happiness score from the 2018 report. Points size shows Generosity scores, whereas the blue hue in the point shows the expectancy of life score.

Preprocessing

In order to pre-process the dataset we will check for missing values, errors and outliers.

Missing Data

In order to look for missing values we're using the `mice`[1] package in R, from which we can observe there exists one missing value on the Perceptions of Corruption variable on the 2018 dataset in *United Arab Emirates*, the one was imputed through mice.

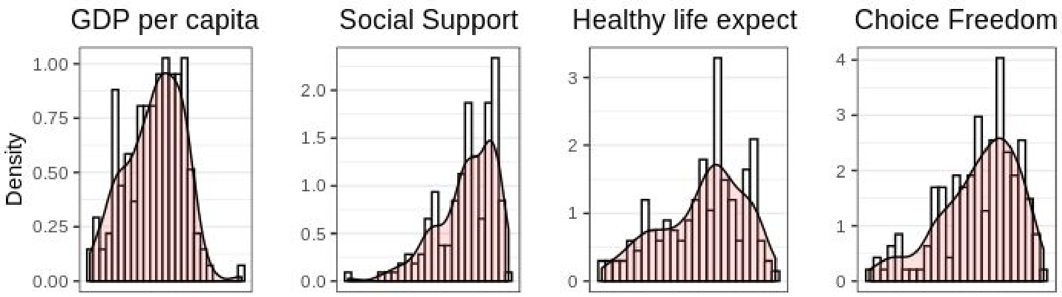


Missing Data Pattern: 155 Complete individuals and 1 country with missing value on Perceptions of Corruption on 2018 dataset

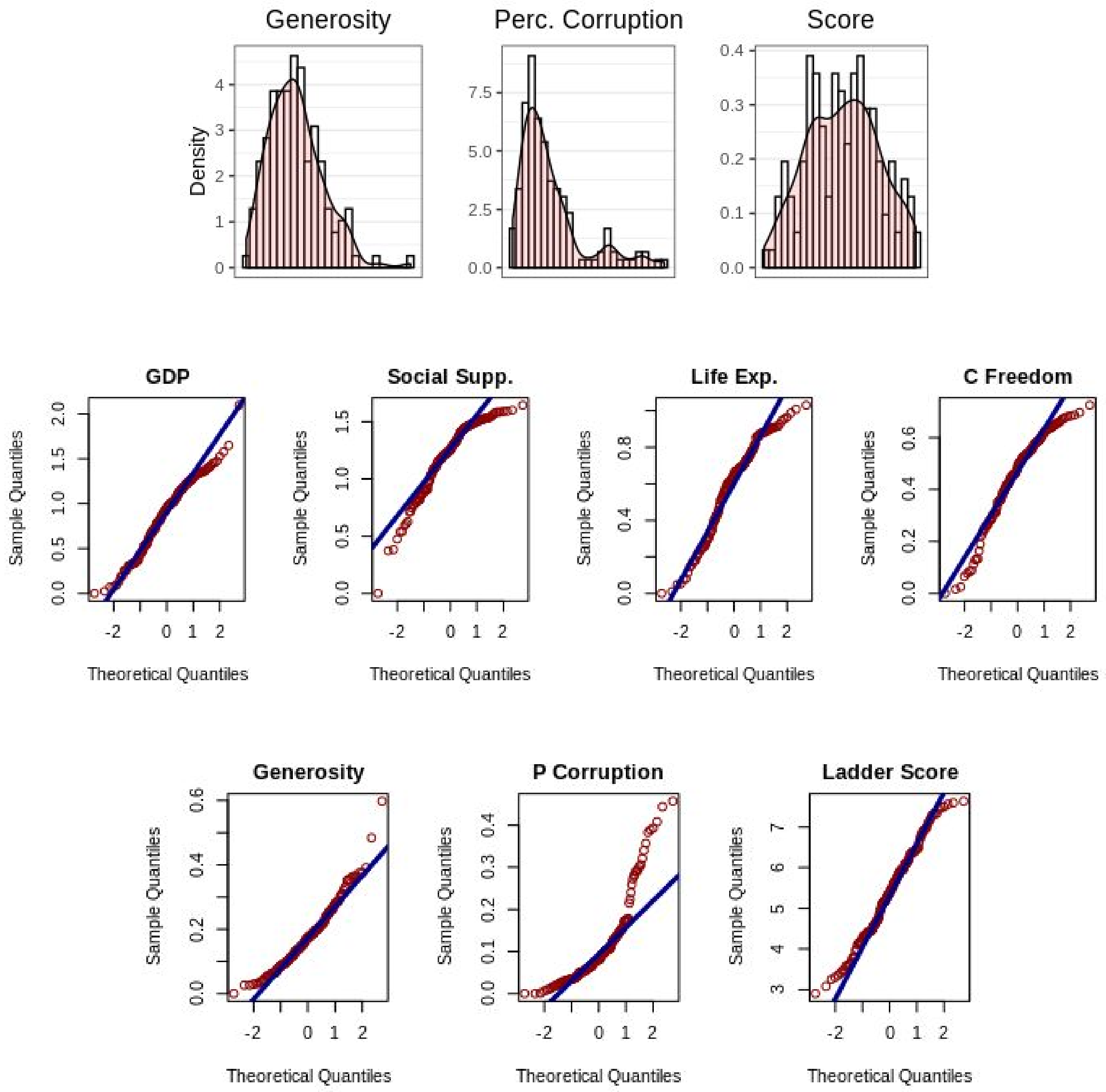
Assessing Normality

To continue with our analysis, we first assess if our data follow a multivariate normal distribution. If this assumption is fulfilled, we will be able to perform several hypothesis tests and modelling techniques that cannot be performed otherwise, such as assessing the independence among variables or the equality of last eigenvalues.

In order to follow a multivariate normal distribution, the data has to fulfill two conditions, the first one being that all the marginal distributions must follow a univariate normal distribution.



[1] https://cran.r-project.org/web/packages/mice/index.html



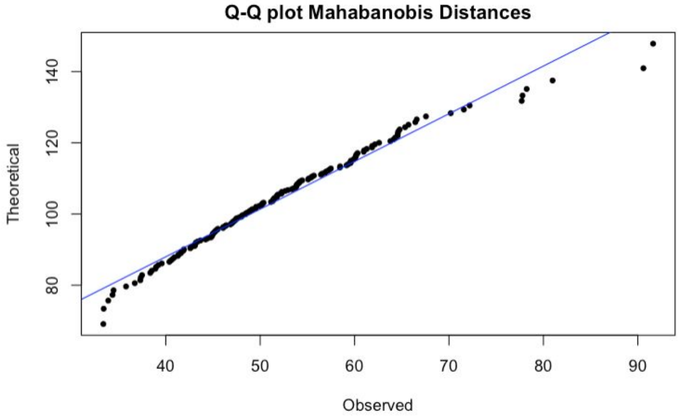
Density plots and QQplots to assess normality of data

From the visualization of the density distributions and the *qq plot* of each variable, we can say that the only variable that seems to be normally distributed is Score. We confirmed this by performing a Shapiro-Wilk normality test on all features.

From this point on, we can say that the data does not follow a normal distribution. However, the second condition that had to be met for it to be so, that all distances from the centroid must follow a chi squared distribution, is still important in determining whether we can use these distances to find multivariate outliers.

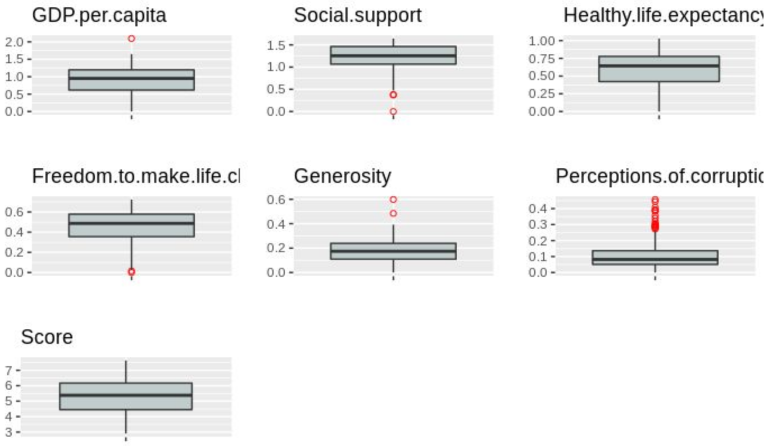
We perform a *qq plot* comparing the observed Mahalanobis distances of our data with respect to the theoretical distances that a Chi-Square distribution with 7 degrees of freedom should have. Additionally, we perform a statistical test by simulation for which we build a reference Chi-Square distribution which will be considered the null hypothesis. If the majority of our points fit with the reference distribution, the p-value will be greater than 0.05 and we could accept the null hypothesis that our data follow the same distribution.

From the *qq plot* we can see that almost all points fit with the reference line corresponding to the theoretical distribution, with only a few points that are further away from it. Moreover, this fact is corroborated with the acceptance of the null hypothesis of the statistical test with a p-value of 0.217. Therefore, the data does follow a Chi-Square distribution and we can use these distances in order to find for multivariate outliers.



Univariate Outliers

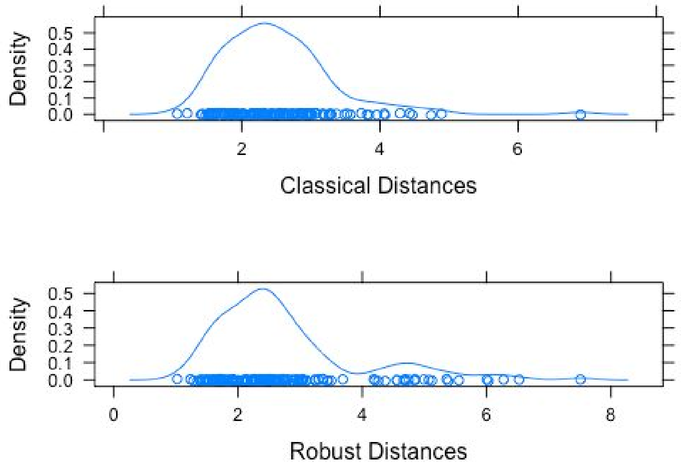
Before inspecting for multivariate outliers, we can search for univariate outliers. We recall that given a range of the form $(Q_1 - C \times IQR,\ Q_3 + C \times IQR)$, where $IQR = Q_3 - Q_1$ is called the *Interquartile Range*, and C is a constant. We say then that an observation X is an **extreme outlier** if it lies outside the range with C=3.



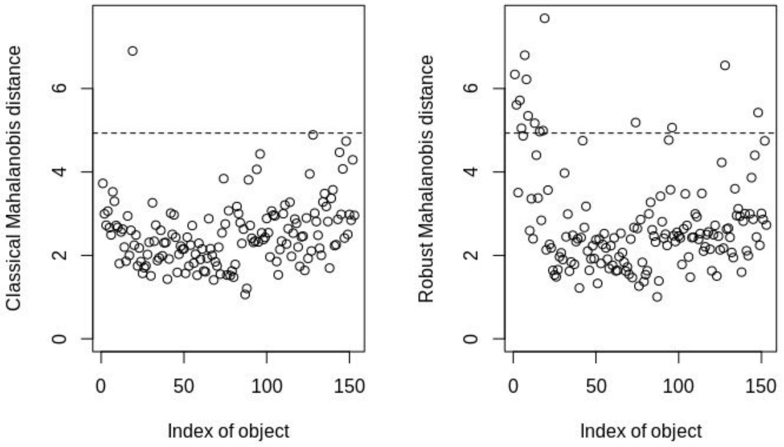
If there exists some mild univariate outliers in GDP.per.capita, Social.support, Freedom.to.make.life.choices and Generosity, we can only find extreme outliers in Perceptions.of.corruption variable. In particular there are 3 countries (*Denmark, Singapore and Rwanda*) considered as outliers, so we have decided to remove or use these instances as supplementary individuals in the next analysis.

Multivariate Outliers

Respect to multivariate outliers, we can obtain the *Mahalanobis Distances* from the World Happiness Report, and detect which countries have distances greater than the Chi-squared threshold. To do so, we use the `Moutlier` library from the `chemometrics` package[2] in R.



Assuming distances follow a Chi-squared distribution, we consider then the classical mahalanobis distances obtained to define as outliers those rows whose distances fall belong to the 1% farmost percentile in the distribution. By following this rule, only the *United Arab Emirates* is considered as an outlier. The Robust Mahalanobis distances considered outliers all happiest countries, so we consider that more than outliers, these countries represent another cluster in the dataset.



Summarizing, the complete list of outliers is: Denmark, Singapore, Rwanda, and United Arab Emirates.

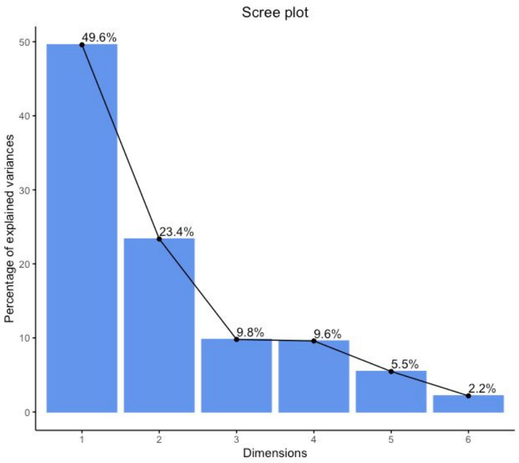
[2] https://cran.r-project.org/web/packages/chemometrics/index.html

Principal Component Analysis

The 2018 Happiness dataset will be subjected to a Principal Component Analysis (PCA) to extract the important information from the multivariate data and to express it as a set of few new artificial variables called principal components. These new variables correspond to a linear combination of the originals, being all of them orthogonal to each other and with a decreasing percentage of explained information (or inertia), where the first PC represents the highest percentage.

We apply the PCA function from `Factominer`[3], setting the Score variable as a quantitative supplementary variable and the countries labeled as outliers as *quantitative supplementary individuals*.

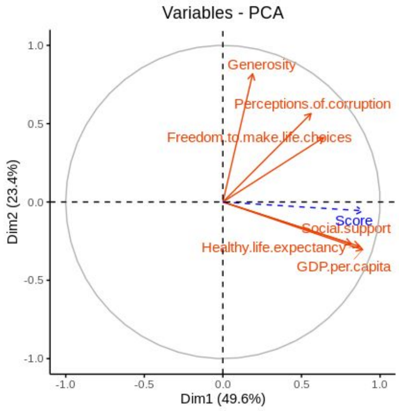
From the Scree Plot of the new Principal Components, we observe that the first principal component entails ~50% of the total variance in the linear transformation, whereas the second variable expresses ~23%. From the last elbow rule we can say that we have 5 significant components, retaining ~98% of the total inertia.



Variables Projection in $\mathbb{R}^n$

From the visualization of the variables (in R^n) we observe that the *Score* feature lies almost completely on the first dimension, and that it is correlated with *Social.Support*, *Healthy.life.expectancy* and *GDP.per.capita* variables, yet it is partially uncorrelated with *Generosity* variable, which is the best represented on the second principal component. *Freedom.to.make.life.choices* it's also partially correlated with *perceptions.of.corruption.*

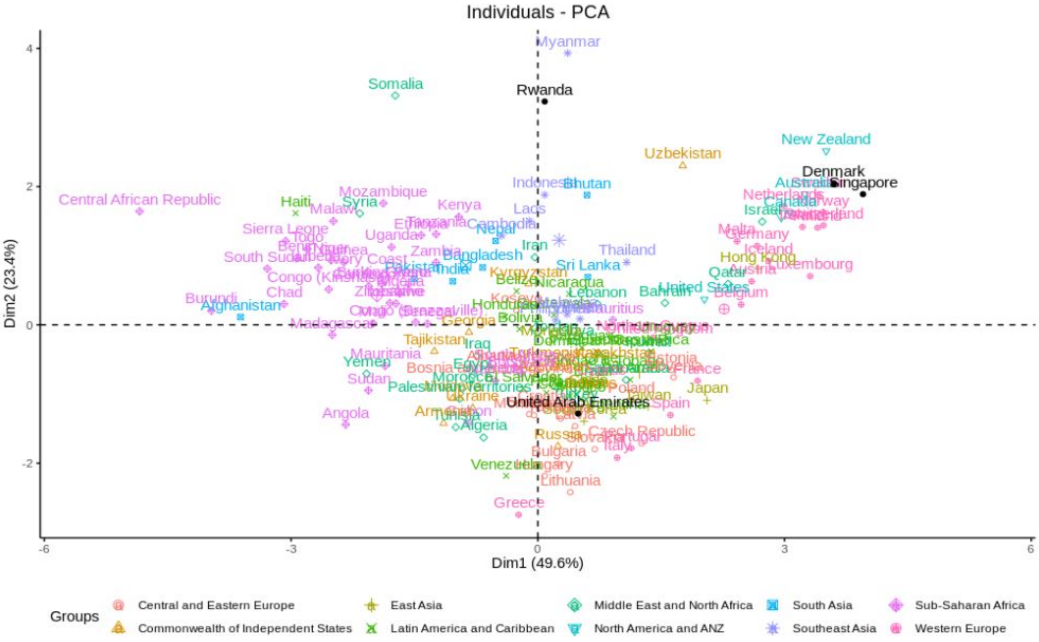
[3] https://www.rdocumentation.org/packages/FactoMineR



Plot of variables from the PCA output in R

Individuals Projection in $\mathbb{R}^p$

From the plot of individuals in R^p, we observe that they follow a very clear distribution by region along the first axis, being the countries from Sub-Saharan Africa at the leftmost part, almost all of Latin America and the Caribbean at the center, and Western Europe and North American Countries at the right part. As we can see from the plot, the *Guttman Effect* (Camiz 2004) is present on the PCA analysis, characterized by the shape of the distribution of the cloud of points (like a croissant). This is because of the extreme values of the individuals in some categories, i.e the Happiness Score. So the happiest countries are at the right side of the plot, while the less happiest countries are on the left-hand side of the plot.

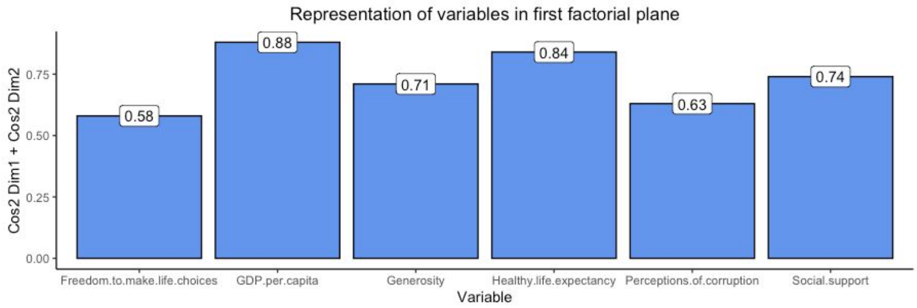


Plot of individuals in the PCA. The supplementary individuals (marked as outliers) are in black color

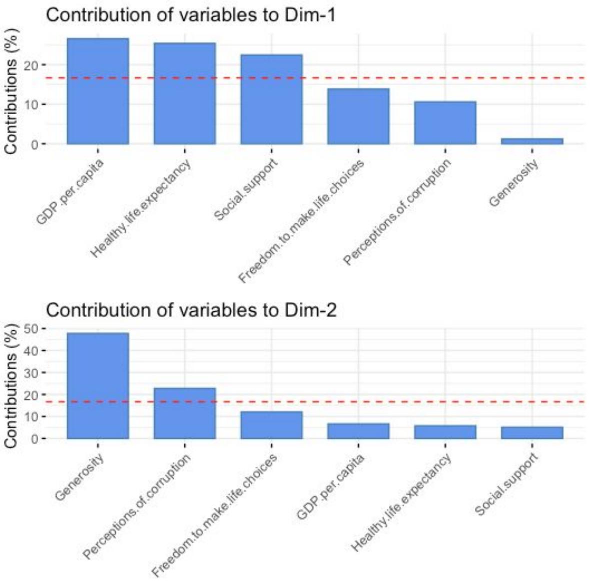
PCA Variables Interpretation

All top countries tend to have high values for all six of the key variables that have been found to support well-being: income, healthy life expectancy, social support, freedom, trust and generosity.

In order to extend our understanding of the Principal Components obtained, we analyze the different measures that reflect how they have been formed. With respect to the variables and by means of the cosine squared metric, we can observe that the most represented features on the first factorial plane are *GDP.per.capita* (0.88), *Healthy.life expectancy* (0.84) and *Social support* (0.74).



These three variables happen to be the ones that contribute more in the formation of the first Principal Component with equal ordinality than the quality assessed. The other three variables contribute more to the formation of the second Principal Component, with *Generosity* contribution almost at 50%. Due to this high contribution, we can see that *Generosity* also happens to be well represented on the first factorial with a cosine squared of 0.71.

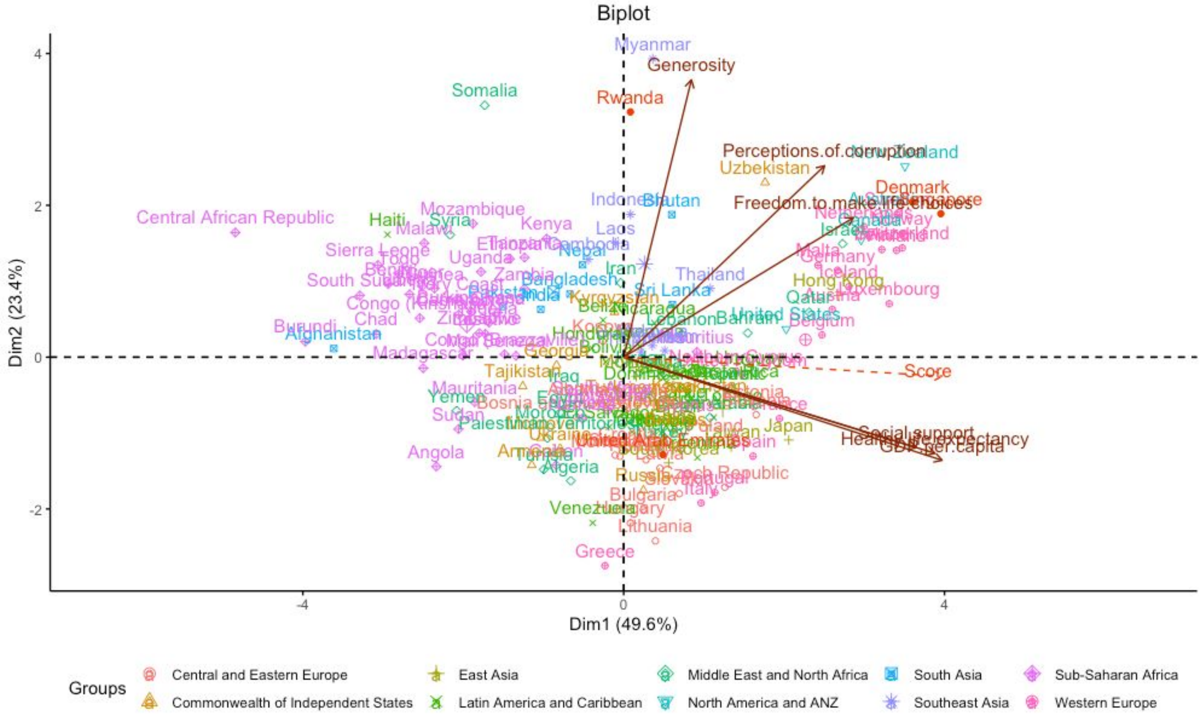


Contribution of the variables to the principal components

PCA Biplot

We can also plot and represent both the individuals and variables of a matrix of multivariate data on the same Biplot with the `factoextra`[4] package.

So far, we can get the same conclusions obtained from the previous PCA analysis, which is that most happiest countries are in the right-most part of the plot along with the projection of the continuous supplementary variable Happiness Score.



Varimax Rotation and Latent factors

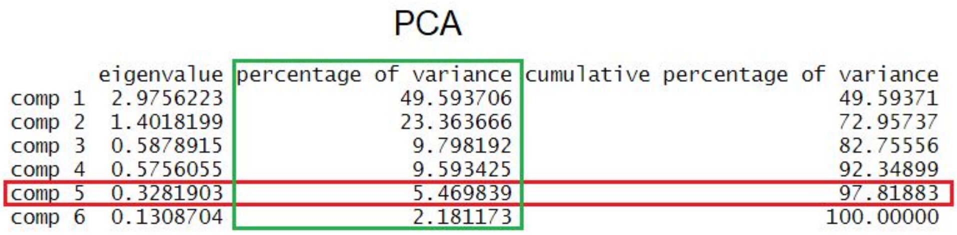
Factor rotation, like Varimax rotation, transforms the initial factors into new ones that are easier to interpret. This is an orthogonal rotation (factors remain orthogonal after the rotation), where the idea is to maximize the sum of the variance of the squared loadings (squared correlations between variables and factors).

Contrasting both PCA analysis and Varimax rotation we can observe that the total amount of variation explained by the 5 factors remains the same, however, there is a decrease in the amount of variation explained by the first factor. Varimax allows a better interpretation of the data on the cost of the variation explained by the first factor, which is distributed among the rest of the factors.

[4] https://cran.r-project.org/web/packages/factoextra/index.htm

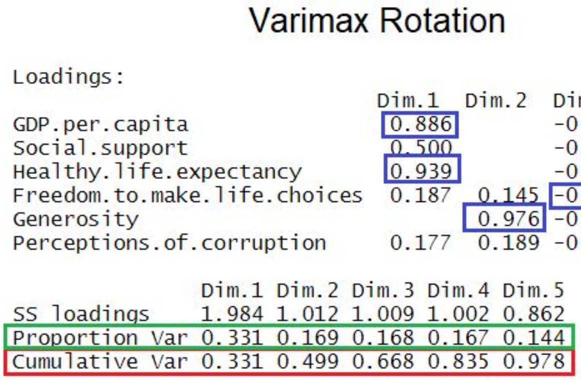
Respect to the Factors, we can see:

- First factor is primarily a measure of GDP.per.capita and Healthy.life.expectancy.
- Second Factor is about Generosity
- Third Factor is about Freedom.to.make.choices
- Fourth Factor is about Perceptions.of.corruption
- Fifth Factor is about Social.support



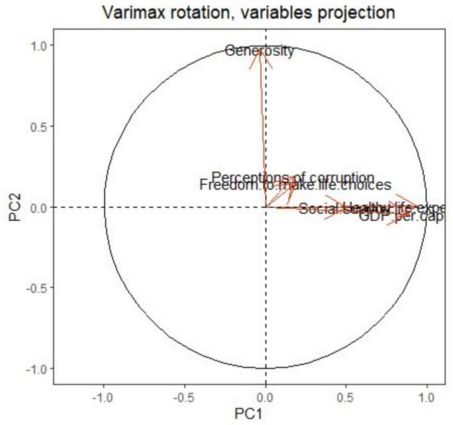
PCA

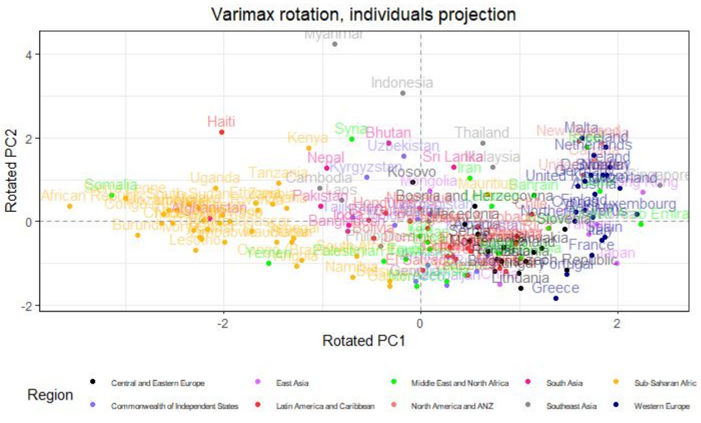
```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  2.9756223              49.593706                          49.59371
comp 2  1.4018199              23.363666                          72.95737
comp 3  0.5878915               9.798192                          82.75556
comp 4  0.5756055               9.593425                          92.34899
comp 5  0.3281903               5.469839                          97.81883
comp 6  0.1308704               2.181173                         100.00000
```



Varimax Rotation

```
Loadings:
                          Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
GDP.per.capita             0.886        -0.137 -0.157  0.306
Social.support             0.500        -0.188         0.842
Healthy.life.expectancy    0.939        -0.136 -0.112  0.186
Freedom.to.make.life.choices 0.187  0.145 -0.939 -0.201  0.148
Generosity                        0.976 -0.128 -0.170
Perceptions.of.corruption  0.177  0.189 -0.196 -0.944

               Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
SS loadings    1.984 1.012 1.009 1.002 0.862
Proportion Var 0.331 0.169 0.168 0.167 0.144
Cumulative Var 0.331 0.499 0.668 0.835 0.978
```



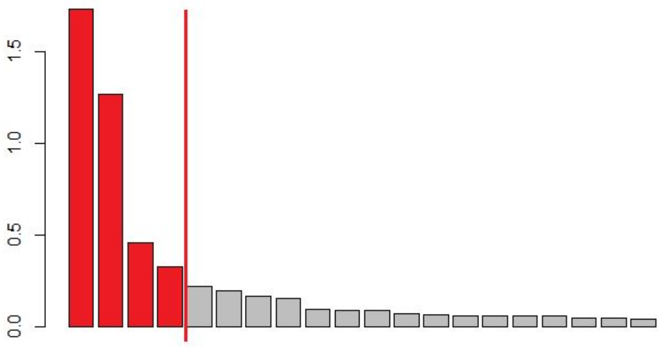


Clustering

After applying PCA we proceed to perform hierarchical clustering with consolidation from the significant components (5) with the new coordinates of the individuals (countries).

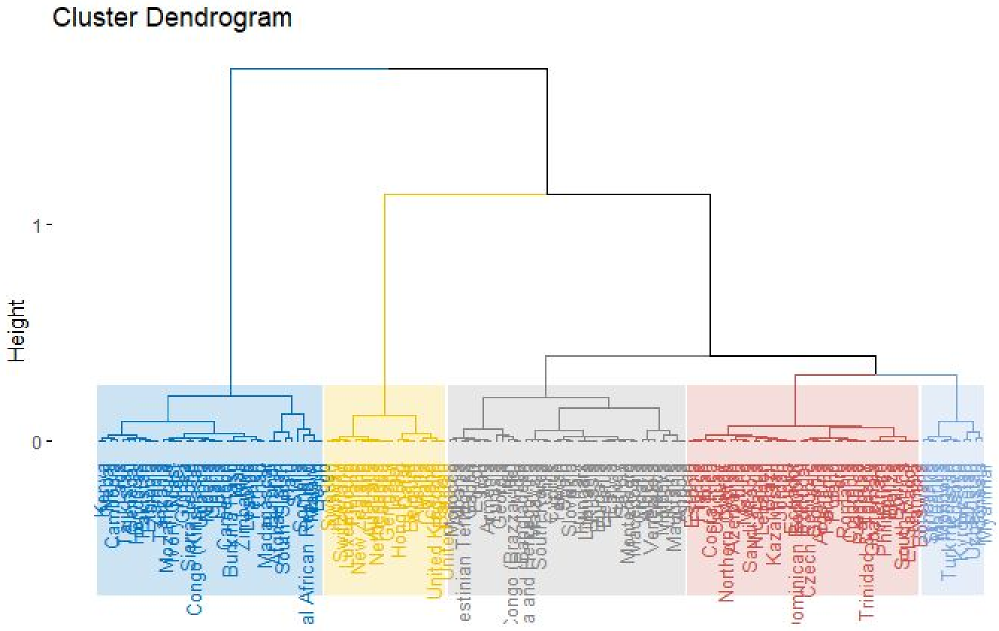
For this purpose, we have decided to use Ward's method given that it is quite similar, by its properties and efficiency, to K-means clustering, as both of them have the objective of minimizing within-cluster variance.

In hierarchical clustering by Ward method, every height value indicates the decrement of the between inertia in the current iteration. By plotting the height of the hierarchical clustering we can observe that h=4 is the last height where there is a significant loss of inertia, so we'll proceed by choosing 5 clusters.



Plot of dendrogram heights and criteria to pick the number of clusters

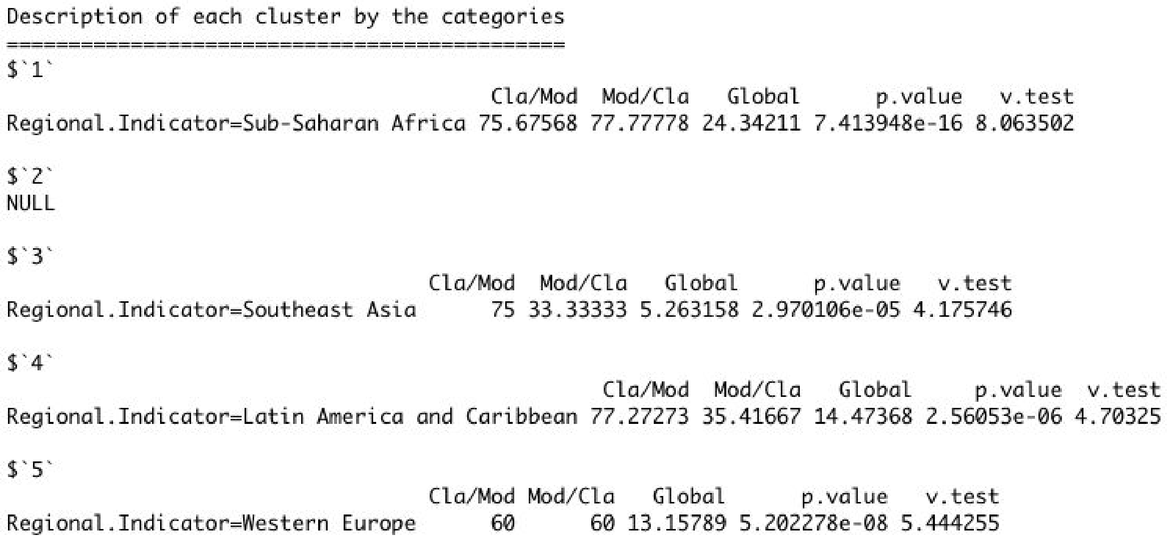
We can see the final distribution of the 5 clusters in a dendrogram. With this consolidation we have 5 clusters, where the biggest one has 48 countries and the smallest one has 18.



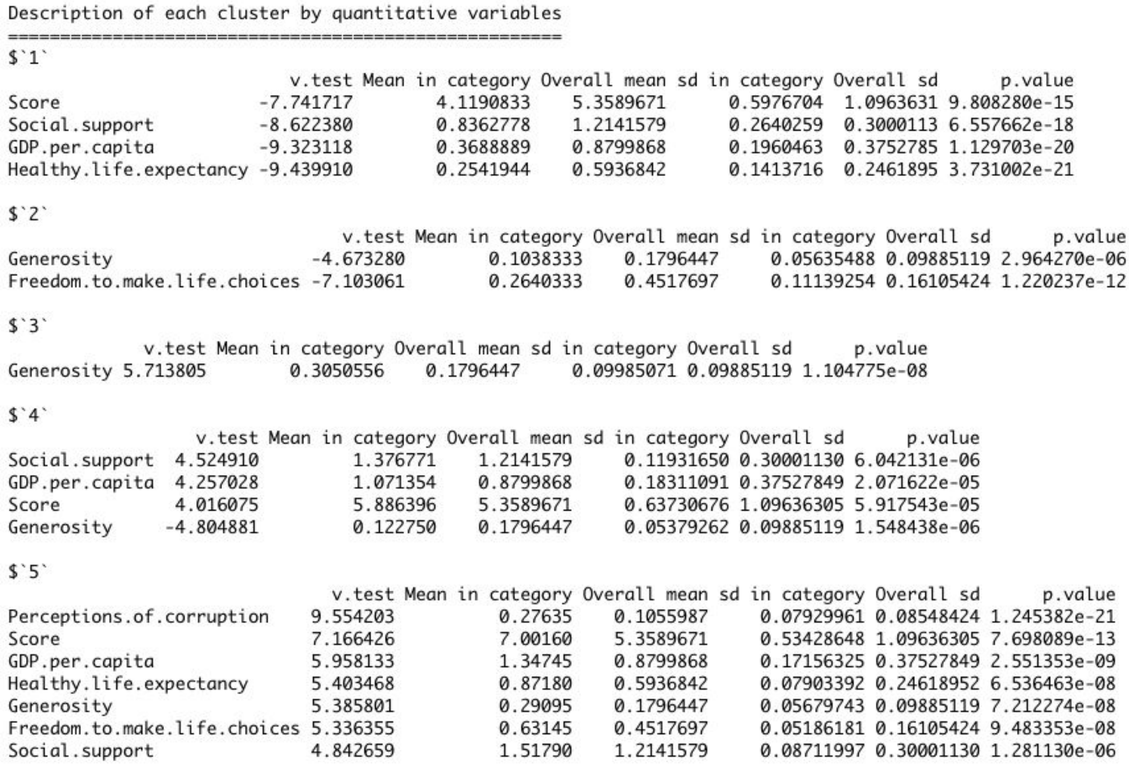
Dendrogram of the clusters after consolidation

Interpretation of the clusters

In order to analyze the clusters we used the `catdes` function.



```
Description of each cluster by the categories
=============================================
$`1`
                                       Cla/Mod  Mod/Cla   Global      p.value   v.test
Regional.Indicator=Sub-Saharan Africa 75.67568 77.77778 24.34211 7.413948e-16 8.063502

$`2`
NULL

$`3`
                                  Cla/Mod  Mod/Cla   Global      p.value   v.test
Regional.Indicator=Southeast Asia      75 33.33333 5.263158 2.970106e-05 4.175746

$`4`
                                                Cla/Mod  Mod/Cla   Global     p.value  v.test
Regional.Indicator=Latin America and Caribbean 77.27273 35.41667 14.47368 2.56053e-06 4.70325

$`5`
                                  Cla/Mod Mod/Cla   Global      p.value   v.test
Regional.Indicator=Western Europe      60      60 13.15789 5.202278e-08 5.444255
```



```
Description of each cluster by quantitative variables
=====================================================
$`1`
                           v.test Mean in category Overall mean sd in category Overall sd      p.value
Score                   -7.741717        4.1190833    5.3589671      0.5976704  1.0963631 9.808280e-15
Social.support          -8.622380        0.8362778    1.2141579      0.2640259  0.3000113 6.557662e-18
GDP.per.capita          -9.323118        0.3688889    0.8799868      0.1960463  0.3752785 1.129703e-20
Healthy.life.expectancy -9.439910        0.2541944    0.5936842      0.1413716  0.2461895 3.731002e-21

$`2`
                                v.test Mean in category Overall mean sd in category Overall sd      p.value
Generosity                   -4.673280        0.1038333    0.1796447     0.05635488 0.09885119 2.964270e-06
Freedom.to.make.life.choices -7.103061        0.2640333    0.4517697     0.11139254 0.16105424 1.220237e-12

$`3`
             v.test Mean in category Overall mean sd in category Overall sd      p.value
Generosity 5.713805        0.3050556    0.1796447     0.09985071 0.09885119 1.104775e-08

$`4`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Social.support  4.524910         1.376771    1.2141579     0.11931650 0.30001130 6.042131e-06
GDP.per.capita  4.257028         1.071354    0.8799868     0.18311091 0.37527849 2.071622e-05
Score           4.016075         5.886396    5.3589671     0.63730676 1.09636305 5.917543e-05
Generosity     -4.804881         0.122750    0.1796447     0.05379262 0.09885119 1.548438e-06

$`5`
                               v.test Mean in category Overall mean sd in category Overall sd      p.value
Perceptions.of.corruption    9.554203          0.27635    0.1055987     0.07929961 0.08548424 1.245382e-21
Score                        7.166426          7.00160    5.3589671     0.53428648 1.09636305 7.698089e-13
GDP.per.capita               5.958133          1.34745    0.8799868     0.17156325 0.37527849 2.551353e-09
Healthy.life.expectancy      5.403468          0.87180    0.5936842     0.07903392 0.24618952 6.536463e-08
Generosity                   5.385801          0.29095    0.1796447     0.05679743 0.09885119 7.212274e-08
Freedom.to.make.life.choices 5.336355          0.63145    0.4517697     0.05186181 0.16105424 9.483353e-08
Social.support               4.842659          1.51790    1.2141579     0.08711997 0.30001130 1.281130e-06
```

First Cluster - Least Developed Countries

- 36 Countries
- Paragons: Guinea, Niger, Cameroon, Ivory Coast, Liberia
- Characterized by countries located in Sub-Saharan Africa, which have a significantly smaller mean for Healthy life expectancy, GDP per capita, Social Support and Score than the rest of the countries. The score of these countries is significantly small in relation to the distribution of the data (v.test between ≈-7.7 and ≈-9.4)

Second Cluster

- 30 Countries
- Paragons: Montenegro, Palestinian Territories, Serbia, Armenia, Egypt
- Composed by countries of several regions, which are characterized by a smaller mean of Generosity and Freedom to make life choices.

Third Cluster

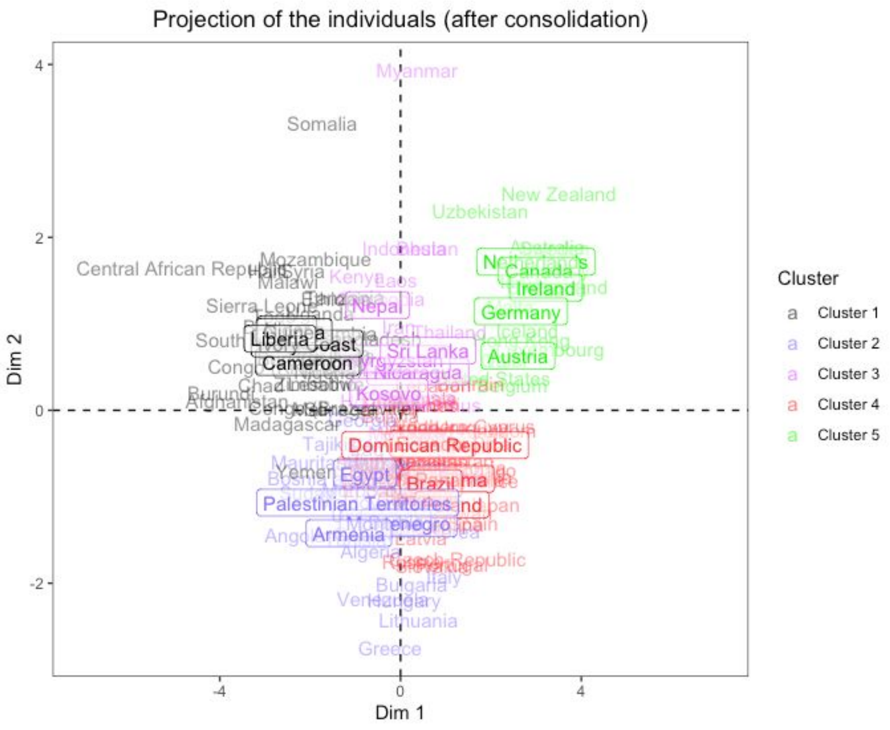
- 18 Countries
- Paragons: Nepal, Sri Lanka, Kyrgyzstan, Nicaragua, Kosovo
- Characterized by a greater generosity. The majority of countries within this cluster are located in Southeast Asia

Fourth Cluster

- 48 Countries
- Paragons: Brazil, Ecuador, Poland, Panama, Dominican Republic
- Characterized by a greater social support and Healthy life expectancy than the average. These countries also report significantly low generosity (v.test of -4.8)

Fifth Cluster

- 20 Countries
- Paragons: Germany, Canada, Netherlands, Ireland, Austria
- Characterized by the countries with the greatest Happiness score and all the six variables that support well-being.



Plot of clusterized individuals (5 clusters). The paragons of each cluster are shown with labels enclosed in a box.

Prediction Model

In order to create a model and create predictions of the Happiness Score, we've decided to use decision trees and random forests.

Validation Protocol

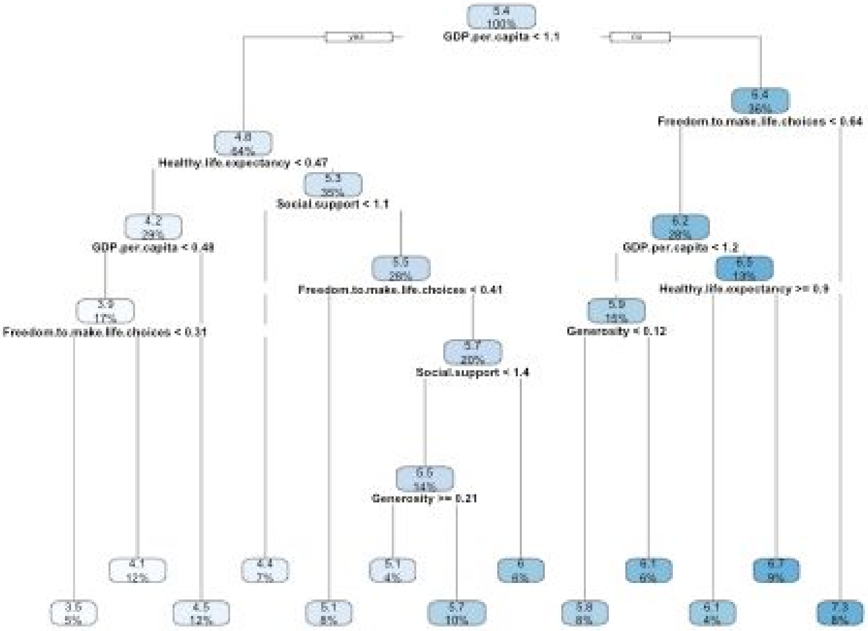
For the purposes of the analysis, all previous analysis have been done with the 2018 World Happiness Dataset. We're also using the 2018 dataset to train and build the models, the ones will be tested with the 2019 World Happiness Dataset.

Decision Trees

The decision tree was built with the 2018 (training) dataset and the `rpart`[5] function in order to predict the Score as a combination of all the response variables (except the Regional Indicator, which is the categorical one).

In order to construct a maximum tree, i.e, without pre-pruning, we call the function using different complexity parameters up to 0.001 with a ten-fold cross-validation.

As a measure of the quality of a node for continuous responses, i.e the node impurity, which is a measure of the homogeneity of the labels at the node, the `rpart` function considers the variance.

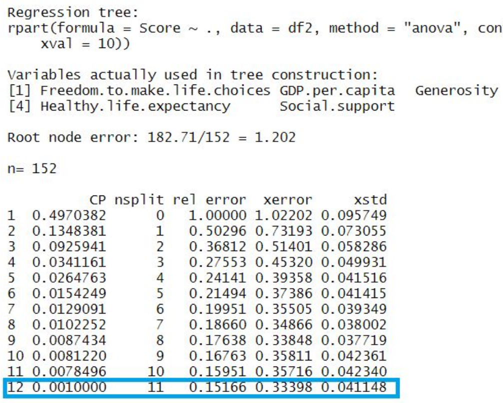


Decision tree of rpart with default parameters

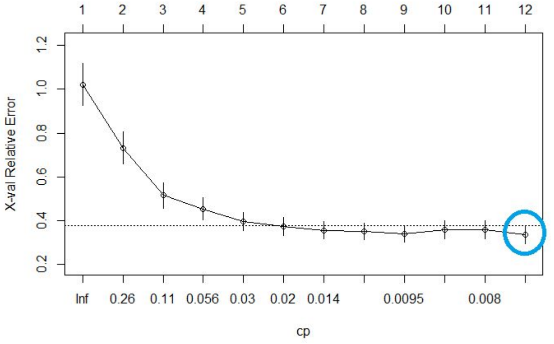
[5] https://www.rdocumentation.org/packages/rpart

Best parametrization and its generalization error

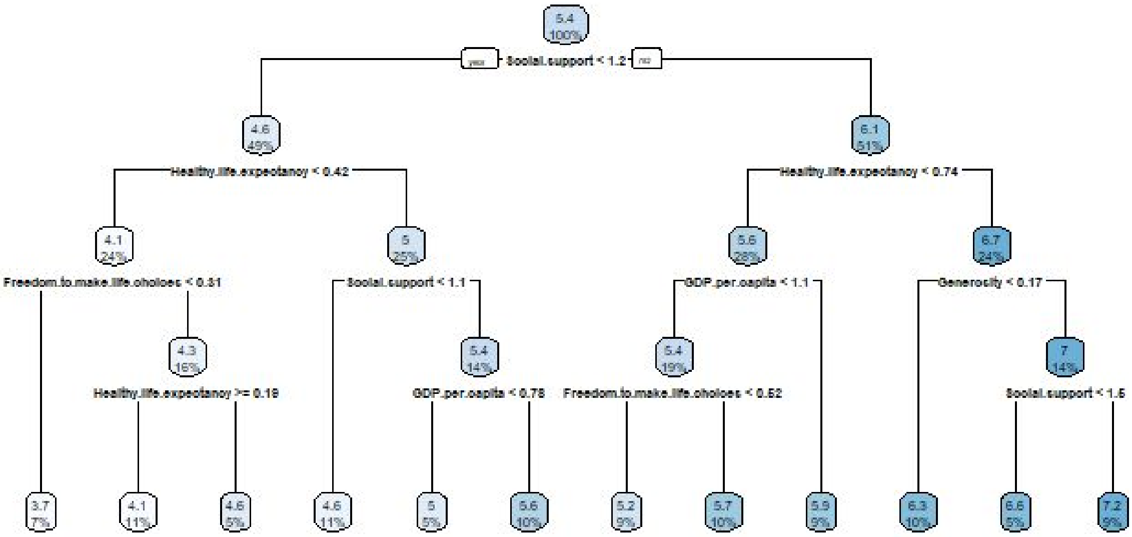
As a result, the maximum decision tree has tried different splits, and the one with the least cross-validated error (xerror) is the optimal value of the Complexity Parameter, in this case a cp=0.001



```
Regression tree:
rpart(formula = Score ~ ., data = df2, method = "anova", control = rpart.control(cp = 0.001,
    xval = 10))

Variables actually used in tree construction:
[1] Freedom.to.make.life.choices GDP.per.capita   Generosity
[4] Healthy.life.expectancy      Social.support

Root node error: 182.71/152 = 1.202

n= 152

          CP nsplit rel error  xerror     xstd
1  0.4970382      0   1.00000 1.02202 0.095749
2  0.1348381      1   0.50296 0.73193 0.073055
3  0.0925941      2   0.36812 0.51401 0.058286
4  0.0341161      3   0.27553 0.45320 0.049931
5  0.0264763      4   0.24141 0.39358 0.041516
6  0.0154249      5   0.21494 0.37386 0.041415
7  0.0129091      6   0.19951 0.35505 0.039349
8  0.0102252      7   0.18660 0.34866 0.038002
9  0.0087434      8   0.17638 0.33848 0.037719
10 0.0081220      9   0.16763 0.35811 0.042361
11 0.0078496     10   0.15951 0.35716 0.042340
12 0.0010000     11   0.15166 0.33398 0.041148
```



With this value we can *prune* the tree to obtain an optimized version and evaluate its performance on the test dataset (2019 World Happiness Dataset). We observe that the pruned tree splits between Score 3.7 and 5.9 first depending on the Healthy.life.expectancy variable, and on the Freedom to make life choices for the scores bigger than that.

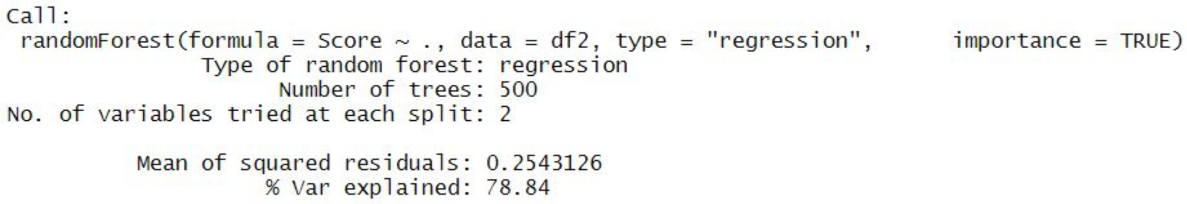


pruned decision tree with optimal complexity parameter

Using this tree against the data from the 2019 report, we obtain a **MSE (Mean squared prediction error) of ~0.42**.

Random Forest

Random forests are a strong modeling technique and more robust than a single decision tree. In this case, we rely on the `randomForest` package in R[6], using the default hyperparameters (ntree = 500 and two features at a time), we get a *Mean of Square Residuals* of ~0.254 and a *Mean Squared Prediction Error* of ~0.231 when tested on the 2019 dataset.

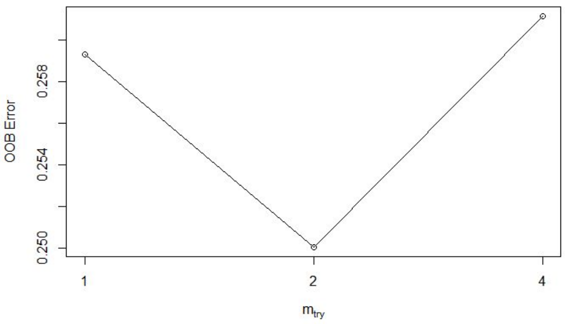


```
Call:
 randomForest(formula = Score ~ ., data = df2, type = "regression",      importance = TRUE) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.2543126
                    % Var explained: 78.84
```

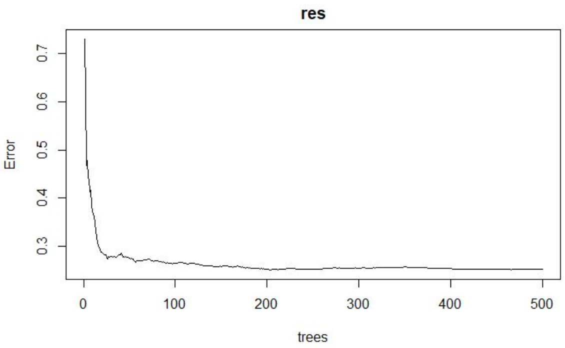
Best parametrization and its generalization error

Using tuneRF

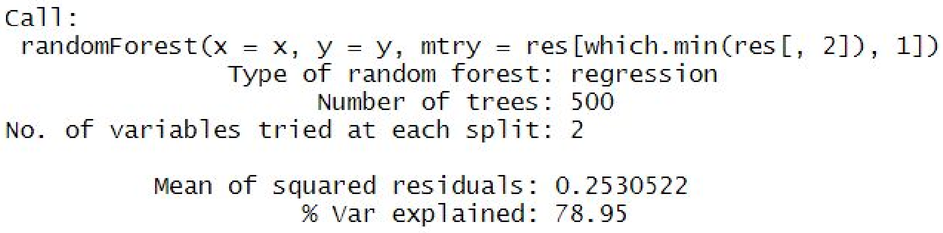
We use the *tuneRF* function from the *randomForest* package to obtain the optimal number of variables to be used in creating the trees in the forest



OOB error estimate with different values of mtry



Error in function of the number of trees



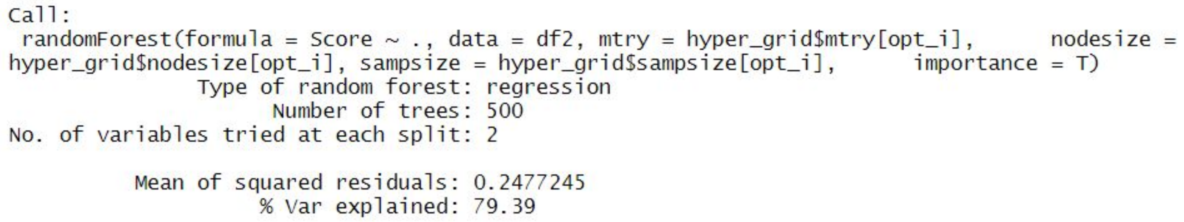
```
Call:
 randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1]) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.2530522
                    % Var explained: 78.95
```

With the proposed tuning we achieve a Mean of Square Residuals of ~0.253 and a Mean Squared Prediction Error of ~0.4835 when tested on the 2019 dataset.

[6] https://www.rdocumentation.org/packages/randomForest/versions/4.6-14

Using full grid search

We created a matrix of values for the parameters *mtry* (6 different values), *sampsize* (3 different values) and *nodesize* (4 different values) of the randomForest function. This matrix contains the 72 different permutations of the values described, and it was used to train 72 different models. We stored the MSE obtained from each training session, and picked the one with the lowest value. The Mean of Square Residuals of ~0.248 and a Mean Squared Prediction Error of ~0.238 with the 2019 dataset.



```
Call:
 randomForest(formula = Score ~ ., data = df2, mtry = hyper_grid$mtry[opt_i],      nodesize =
hyper_grid$nodesize[opt_i], sampsize = hyper_grid$sampsize[opt_i],      importance = T)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.2477245
                    % Var explained: 79.39
```

Summarizing, the best performance of the different random forest was tuneRF.

testing error	**no tuning**	**tuneRF**	**grid search**
MSE Training Model	0.254	0.253	0.248
2019 MSE Prediction	0.231	0.234	0.238

Conclusions

We analyzed the World Happiness Report through multivariate techniques such as PCA, where we observed how the dataset can be reduced to fewer dimensions, with 2 of them accounting for ~75% of the total information contained. This suggests that there is a latent concept which is aligned with the target of the report: researching the happiness of the different countries around the world. This can be confirmed visually by observing the Guttman effect on the plot of individuals.

The most important variables observed during this principal components analysis were GDP.per.capita, Health.life.expectancy, and Social.support.

Next, we created different clusters of countries given the coordinates of the individuals in the new PCA dimensions. We found 5 clusters, being the first made of the least happiest countries, while the fifth cluster was composed of the most happiest countries. Again, one of the most important variables in the clustering is the GDP.per.capita.

Finally, we created a decision tree and random forest model in order to predict new Happiness Scores for 2019.

We can also say that the model is not that good as we would like to be. If well decision trees provide us a clear and easy interpretation, they may fail sometimes in achieving better results. One example of this situation is when all top happiest countries obtain the same Happiness Score with the decision tree model. In other words, the tree model maps the continuous variable Score to a subset of Score values in the leaves of the tree.

It is interesting to notice once again that the most important variables for the optimal tree were GDP.per.capita, Health.life.expectancy, and Social.support, however, this time Generosity was the least important variable, whereas in PCA it was the fourth.

The best decision tree had a MSE of ~0.231, which is the same that the best of the random forest models produced, probably because our dataset is not large enough and we would need more data in order to improve the models.

Bibliography

1. World Happiness Report, https://worldhappiness.report/
2. Datasets from World Happiness Reports, https://www.kaggle.com/unsdsn/world-happiness
3. Camiz, Sergio. "*The Guttman Effect: its Interpretation and a New Redressing Method*" - ISSN 11094192. - 52005, pp. 734.
4. van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1-67. https://www.jstatsoft.org/v45/i03/.