Statistical Learning Project: World Happines Report

Federico Chiarello, Sara Nanni, Davide Pivato

2023-06-27

Contents

Introduction

The **World Happines Report** is a yearly pubblication that provides useful insights about the wellness of the people from all over the world. It is generated from a survey of the state of global happiness computed as a numeric score that takes into account both social and economical features. We choosed to analyze the 2021 World Happines Report and we are interested in finding meaningful relationships among the impact of those features on the Happines Score registered in each country.

Dataset

Happiness Report 2021

The World Happiness Report computes various features based on data collected from surveys conducted by Gallup World Poll. The original dataset that we used has been collected from Kaggle website, it contains 149 observations on 8 different variables. The key features included in the World Happiness Report are:

- **Happiness Score**: It is a composite score that estimates the happiness in a country. It is typically derived from survey questions asking individuals to rate different aspects of their life. It ranges between 0 and 10.
- **Region**: Region the country belongs to.
- **Logged GDP per capita** - How much each country produces (indicated by the GDP - Gross Domestic Product), divided by the number of people in the country. GDP per capita gives information about the size of the economy and how well it is performing.
- **Social support**: It captures the perception of individual on the opportunity of having someone to count on in times of trouble.
- **Healthy Life Expectancy**: It estimates the number of years an individual can expect to live in good physical and mental health.
- **Freedom to Make Life Choices**: It measures the extent to which individuals feel they have the freedom to make decisions about their lives, including personal and career choices.
- **Generosity**: It captures the inclination of individuals to engage in charitable activities and their perception of being part of a positive community.

- **Perceptions of Corruption**: It captures the perception of the level of corruption perceived within a country, both in business and in governments.

We think it is important to clarify a couple of details regarding the nature of the data presented in the World Happiness Report and how they were collected or computed.

The World Happiness Report 2021 use data from the Gallup World Poll surveys from the previous two years. They are based on answers to the some life evaluation question. The interviewed are asked to rate their own current lives on a 0 to 10 scale. The number of people and countries surveyed varies year to year, but largely more than 100,000 people in 150 countries participate in the Gallup World Poll each year. The values for the GDP per capita and for healthy life expectancy are computed on the 2021 year.

World Data

We then decided to merge our dataset with the World Data 2021 dataset present on Kaggle, which contains other useful informations about each country. The merge was carried on the ISO code of each state.

- **Fertility**: The average number of children that would be born to a woman over her lifetime.
- **Sex ratio**: Ratio of males for each female in a population.
- **Median Age**: Median age of the population of that country.
- **Life Expectancy**: It estimates the number of years an individual can expect to live.
- **Population growth**: Population growth rate over the last year and its projections over the world's countries.
- **Suicide rate**: Suicide rate as per data published by the World Health Organization (WHO).
- **Urbanization rate**: The population shift from rural to urban areas, the decrease in the proportion of people living in rural areas, and the ways in which societies adapt to this change.

Library imports

```
library(dplyr)
```

```
## Warning: il pacchetto 'dplyr' è stato creato con R versione 4.2.2
```

```
library(scales)
```

```
## Warning: il pacchetto 'scales' è stato creato con R versione 4.2.2
```

```
library(corrplot)
```

```
## Warning: il pacchetto 'corrplot' è stato creato con R versione 4.2.2
```

```
library(ggplot2)
```

```
## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.2.3
```

```
library(gridExtra)
```

```
## Warning: il pacchetto 'gridExtra' è stato creato con R versione 4.2.3
```

```
library(sf)
```

```
## Warning: il pacchetto 'sf' è stato creato con R versione 4.2.3
```

```
library(rnaturalearth)
```

```
## Warning: il pacchetto 'rnaturalearth' è stato creato con R versione 4.2.3
```

```
library(rnaturalearthdata)
```

```
## Warning: il pacchetto 'rnaturalearthdata' è stato creato con R versione 4.2.3
```

```
library(car)
```

```
## Warning: il pacchetto 'car' è stato creato con R versione 4.2.3
```

```
## Warning: il pacchetto 'carData' è stato creato con R versione 4.2.3
```

```
library(leaps)
```

```
## Warning: il pacchetto 'leaps' è stato creato con R versione 4.2.3
```

```
library(glmnet)
```

```
## Warning: il pacchetto 'glmnet' è stato creato con R versione 4.2.3
```

```
## Warning: il pacchetto 'Matrix' è stato creato con R versione 4.2.3
```

Data Pre-Processing

After having merged our two sources of data, we carried on the cleaning process detecting and deleting the NAs values from our final dataset. The original dataset presents an unsuitable distinction of countries under a statistical point of view, so in order to perform a richer analysis we decided to shift from a continental subdivision to a macro-region one, mapping some countries to their closest continent. In particular we chose to aggregate Australia and New Zeland to the Asia region and merge all of the American countries under the same partition.

```
happiness_2021 <- happiness_2021 %>%
  mutate(Continent = ifelse(Continent %in% c("Oceania", "Asia"), "Asia and Oceania", Continent))

happiness_2021 <- happiness_2021 %>%
  mutate(Continent = ifelse(Continent %in% c("North America", "Latin America"), "America", Continent))
```

GDP vs Logged GDP

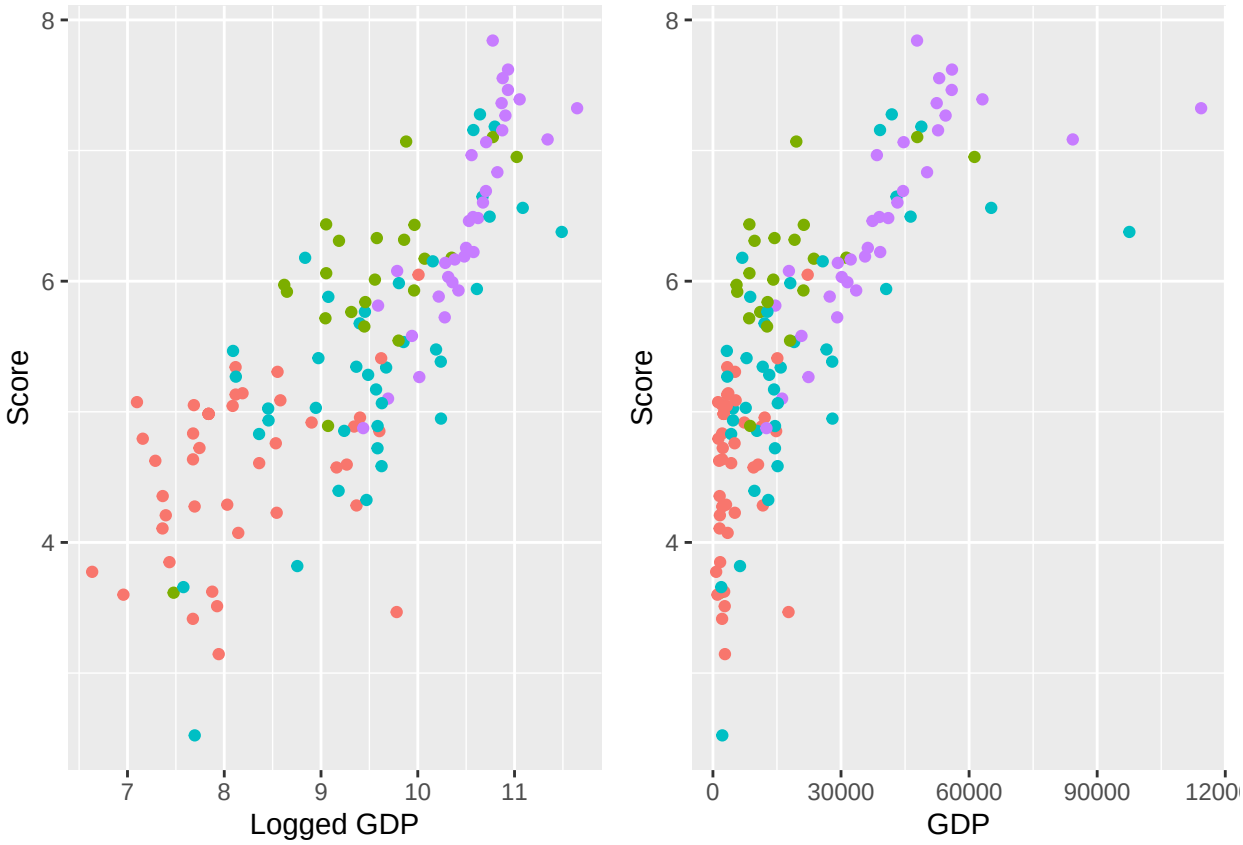
GDP represents the total economic value of all goods and services produced within a specific geographic region over a given period, in our case a year. It includes the value of consumer goods, investments, government spending, and net exports (exports minus imports).

GDP per capita is obtained by dividing the GDP by the population, it provides an average measure of economic output per person. It serves as an indicator of the economic well-being and standard of living within a region. A higher GDP per capita generally suggests higher average incomes and a potentially higher standard of living, although it doesn't necessarily reflect the distribution of wealth within the population.

Our dataset directly presented the **Logged GDP per capita**, that is the natural logarithm of the GDP per capita. By using the logarithm, extreme values are compressed, making the data distribution more symmetrical. This kind of preprocessing has several advantages, it helps to transform the data to a more linear scale, which can be useful for our statistical analysis. It also helps to stabilize the variance of the data, as GDP per capita values can vary widely across countries.

Here we compared the *Logged* GDP per capita and the GDP per capita over the happiness score values using two scatterplots. We can note that the non-logged version presents a clear non linear relationship between the GDP and the happiness score, instead the logged version present a clear linear relationship. It is also evident that the values of the GDP, in the non-logged version, are more widespread, taking really high values for a bunch of countries.

Those results explain the choice of the organization behind the World Happiness Report to stick with the logged version of the GDP per capita.



Scale variables

In order to obtain more interpretable values we decide to scale all our features to the same range between 0 and 1. This will help in making more compact and robust graphical representation and especially later in the report in the comparison between scores in the model creation process.

```
happiness_2021$Life.expectancy <- rescale(happiness_2021$Life.expectancy)
happiness_2021$Fertility <- rescale(happiness_2021$Fertility)
happiness_2021$Urbanization.rate <- rescale(happiness_2021$Urbanization.rate)
happiness_2021$Median.age <- rescale(happiness_2021$Median.age)
happiness_2021$Fertility <- rescale(happiness_2021$Fertility)
happiness_2021$Population.growth <- rescale(happiness_2021$Population.growth)
happiness_2021$Sex.ratio <- rescale(happiness_2021$Sex.ratio)
happiness_2021$Suicide.rate <- rescale(happiness_2021$Suicide.rate)
happiness_2021$Logged.GDP.per.capita <- rescale(happiness_2021$Logged.GDP.per.capita)
happiness_2021$Healthy.life.expectancy <- rescale(happiness_2021$Healthy.life.expectancy)
happiness_2021$Generosity <- rescale(happiness_2021$Generosity)
```

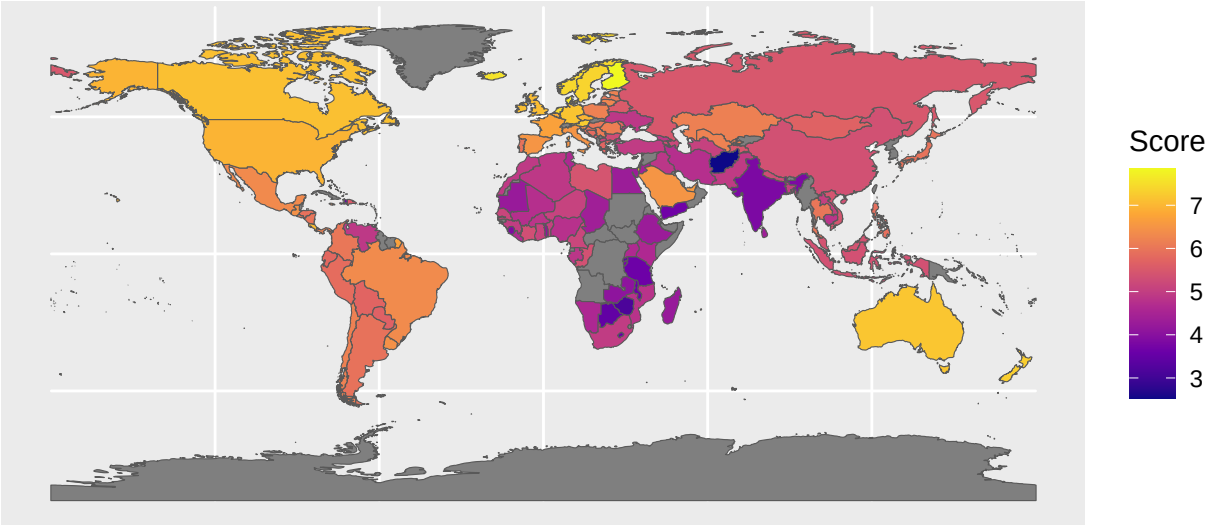
Exploratory Data Analysis

Maps

To obtain a first overview of the worldwide distribution of the happiness score we decided to plot an infographic map. This plot indicates the amount of happiness score for each specific nation of which we have available

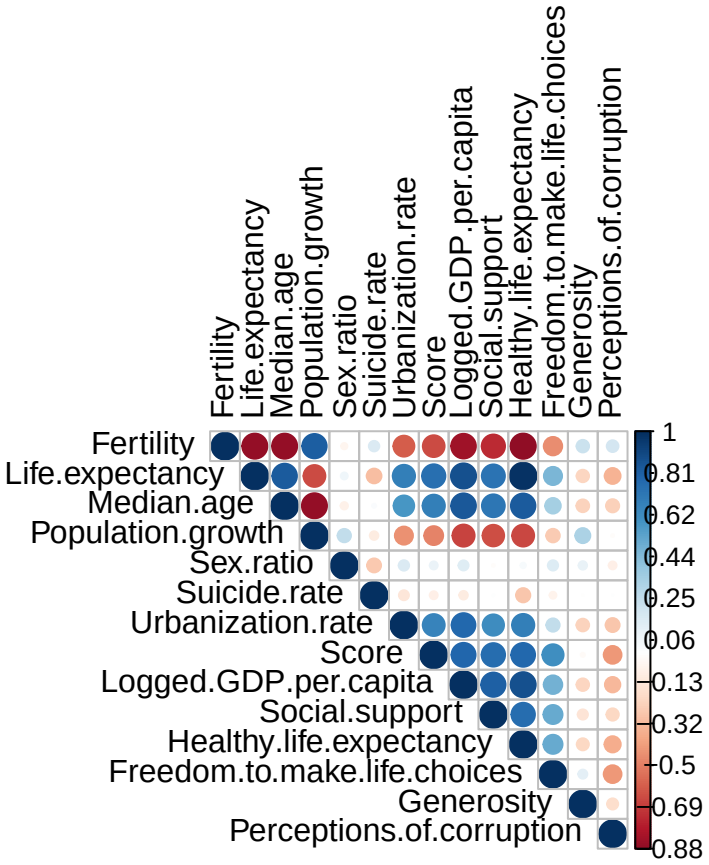
data.

```
world <- ne_countries(scale = "medium", returnclass = "sf")

merged_df <- merge(happiness_2021, world, by.x = "ISO.code", by.y="iso_a3", all.y=TRUE)
df1_selected <- merged_df[c("ISO.code", "Country", "Score")]
merged_2_df <- merge(world, df1_selected, by.x = "iso_a3", by.y="ISO.code", all.x=TRUE)

ggplot(data = merged_2_df) + geom_sf(aes(fill=Score)) + scale_fill_viridis_c(option = "plasma")
```



Correlations

In order to discover the first significant relationships between the features of our dataset we started computing the matrix of the correlations among them. In this plot the strength of the correlation is represented by the size of the circle in the corresponding pairing cell, while the gradient of the color of the circle refers to the nature of the correlation, blue for the positive correlations and red for the negative ones.



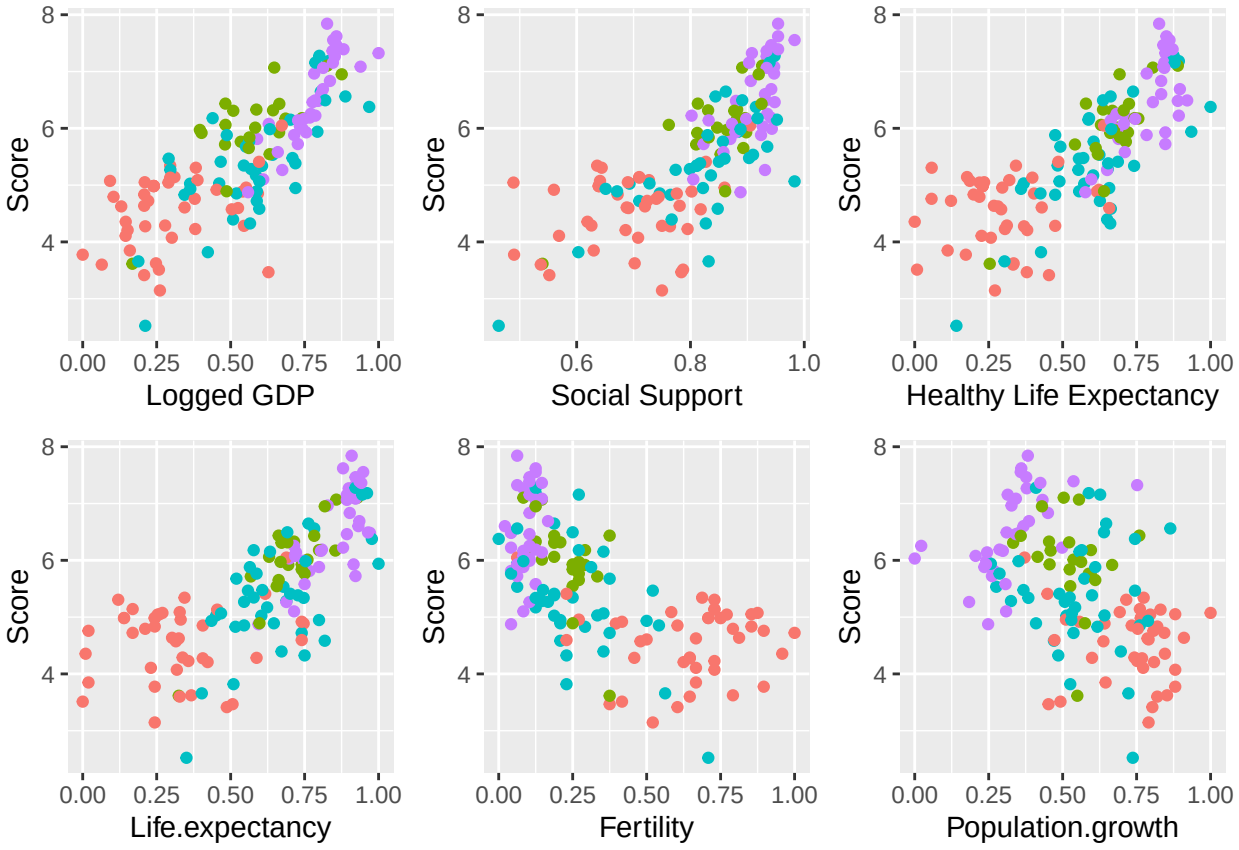
The most relevant insights we can retrieve by looking at the plot are that between several variables exist a very strong correlation (either positive or negative), meaning that there is the risk that they could contain redundant piece of information, this issue will be addressed later in the report in the variable's selection section. By now we will limit to analyze the feature that domain-wise could be more significant to a good description of the happiness score.

Scatterplots

Scatterplots on score Now we want to focus on a better understanding of the most influent variables with the target response, and to do this we show the partial correlations between the latter and the relevant formers.

Variable	Correlation with Happiness Score
Logged.GDP.per.capita	0.79
Healthy.life.expectancy	0.78
Social.Support	0.76
Life.expectancy	0.75
...	...
Fertility	-0.65

To better visualize this strongest correlations with the Happiness Score we proceed now to show the relative scatter plots in which each nation is colored by continent. As a confirm of this influence we can notice some strong linear nature in the relationship showed in the following plots.



Life Duration and Population Grow Variables Analysis

In this section we briefly address the problems brought up by the presence of multiple variables related to the duration and quality of life and by population grow variables.

In particular we will address the following variables:

- Healthy Life Expectancy
- Life Expectancy
- Median Age
- Fertility
- Population Grow

As we have already seen from the correlation matrix we have strong positive linear correlations between life expectancy, healthy life expectancy and median age. This make sense because they all refers to duration of lives but with a slightly different focus. We will address this issue in the following paragraph.

A strong positive linear correlation can be found between the population grow and fertility. This also matches with our intuition.

Then all the variables present in the first group are negatively correlated with all the variables of the second group. In particular in a following paragraph we will focus on the relationship between Fertility and Healthy Life Expectancy, but a similar analysis can be carried out on the other variables as well, obtaining similar results.

Life Expactency vs Healthy Life Expectancy As we have seen in the correlation matrix Life Expectancy and Healthy Life Expactancy are strongly correlate to each other, with a correlation value above 0.98 . This

was expected since they carry a similar information, in fact they both refers to the amount of living years, but with a different focus:

- Life Expectancy represents the number of years an individual can expect to live;
- Healthy Life Expectancy represents the number of years an individual can expect to live in good physical and mental health.

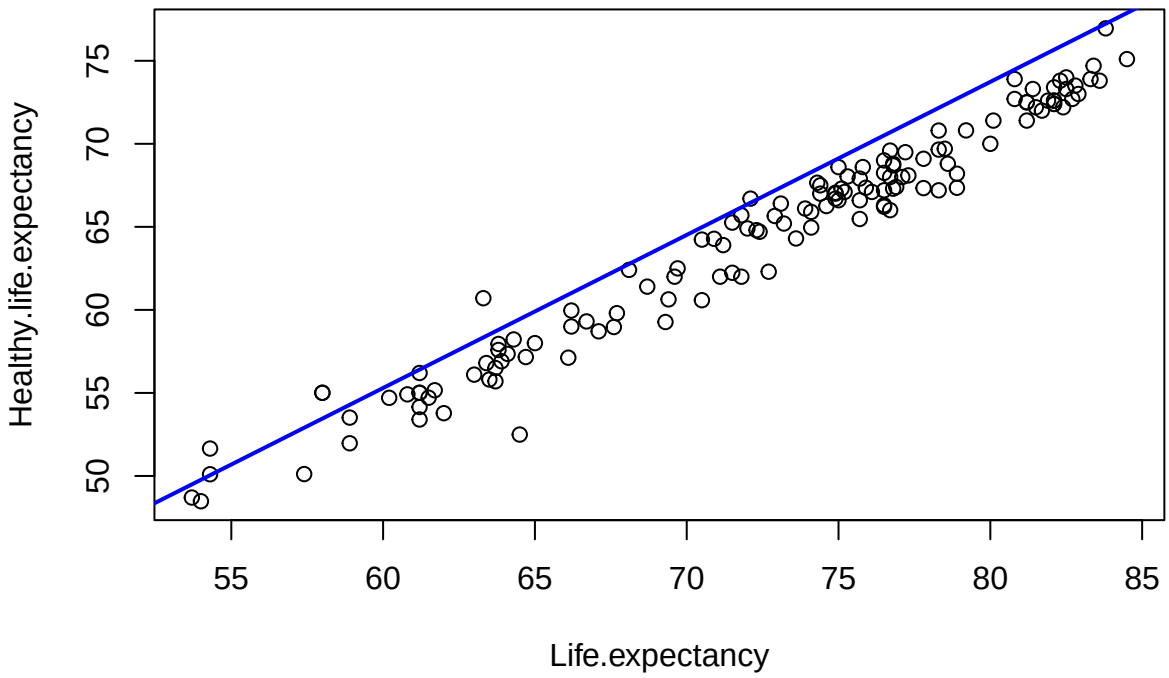
We expect that the possibility of reaching an higher age, in particular in good physical and mental conditions, will increase the happiness score. In the variable selection paragraph we will proceed to remove one of them using both VIF score and domain knowledge.

Simple Linear Regression We provide here a simple linear regression, using the life expectancy as an explanatory variable and the healthy life expectancy as a target. This model provides us a simple interpretation of the linear relations that is present between the two variables: an increase of one year in the life expectancy corresponds to an increase of slightly less than a year (0.92 years, corresponding to 11 months) in the healthy life expectancy. In the scatterplot is also present a line correspondent to the values predicted by the model. As we can see the line is slightly shifted from the position in which we would have expected to find it, this is due to the presence of some outliers. We won't address this issue here as this does not represent the main focus of our project.

```
life.exp.reg <- lm(Healthy.life.expectancy ~ Life.expectancy, data = happiness_2021)
summary(life.exp.reg)
```

```
##
## Call:
## lm(formula = Healthy.life.expectancy ~ Life.expectancy, data = happiness_2021)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.183709 -0.025168  0.001746  0.023764  0.140557
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.001544   0.009950   0.155    0.877
## Life.expectancy 0.921620   0.015072  61.149   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04484 on 136 degrees of freedom
## Multiple R-squared:  0.9649, Adjusted R-squared:  0.9646
## F-statistic:  3739 on 1 and 136 DF,  p-value: < 2.2e-16
```

```
plot(Healthy.life.expectancy~Life.expectancy)
abline(life.exp.reg, col="blue", lwd=2)
```



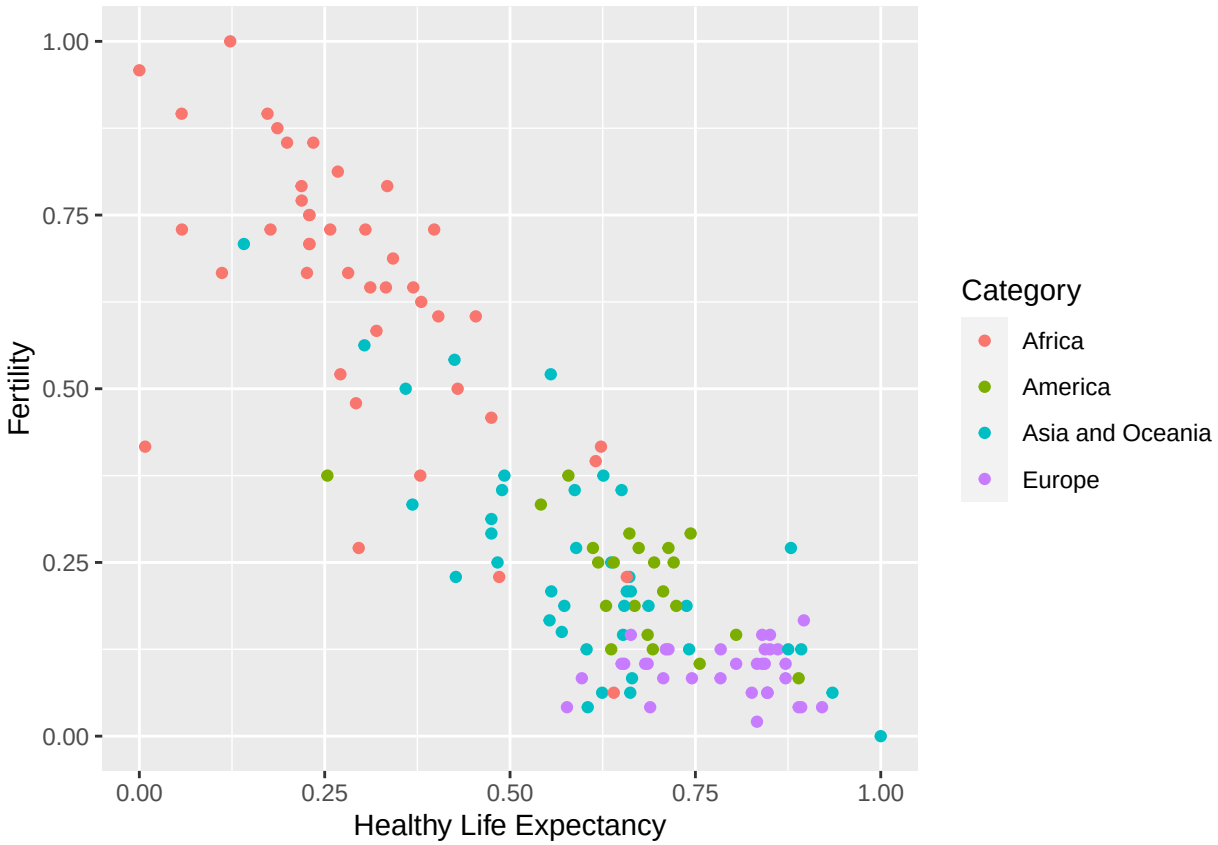
Fertility vs Healthy Life Expectancy Here we highlights the negative linear correlation between Fertility and Healthy Life Expectancy. We defined fertility as the average number of children that would be born to a woman over her lifetime.

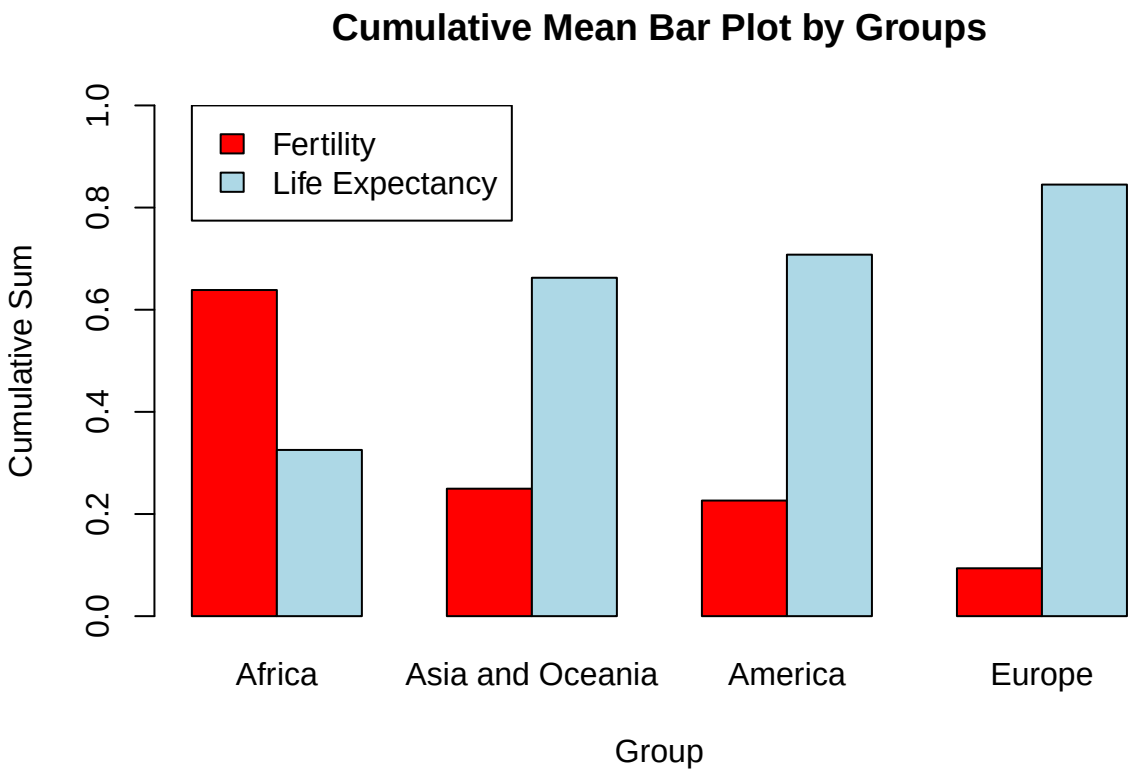
This result matches with general statement that life expectancy of a country and the number of children born are often inversely proportional due to various social, economic, and demographic factors.

Here are a few reasons for this relationship:

1. **Economic development**: As countries develop economically, they typically experience improved healthcare systems and have better access to medical facilities. This leads to a decrease in infant mortality rates and an increase in life expectancy. Consequently, when people have more confidence in the survival of their children, they tend to have fewer children overall.
2. **Education and empowerment of women**: With increased education and empowerment, women tend to have more opportunities for personal and professional growth. When women have access to education, employment, and reproductive healthcare, they often choose to delay marriage and childbearing. This results in smaller family sizes and a decline in the total fertility rate.
3. **Social security and elderly care**: In countries with robust social security systems and comprehensive elderly care, individuals may feel more secure about their future and retirement. They may rely less on having more children to support them in old age, knowing that there are alternative means of support available. This can contribute to a decrease in the number of children born.

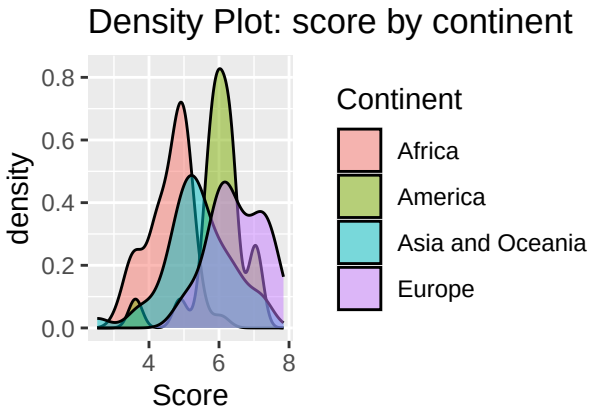
It's important to note that while a general inverse relationship exists between life expectancy and the fertility, there can be exceptions across countries and cultures. Societal norms, religious beliefs, government policies, and other factors can influence the specific dynamics of population growth and life expectancy in different regions.

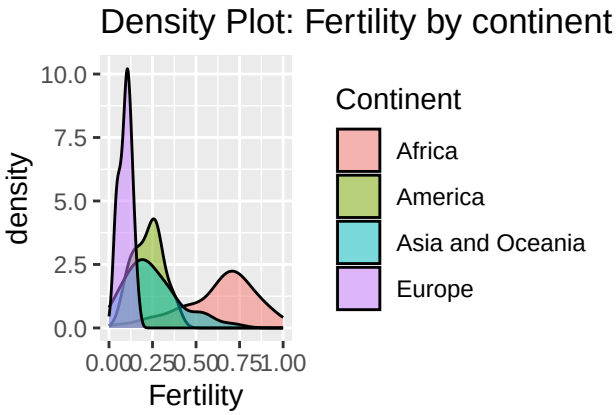

1.00
0.75
0.50
0.25
0.00
Fertility
0.00
0.25
0.50
0.75
1.00
Healthy Life Expectancy
Category
Africa
America
Asia and Oceania
Europe

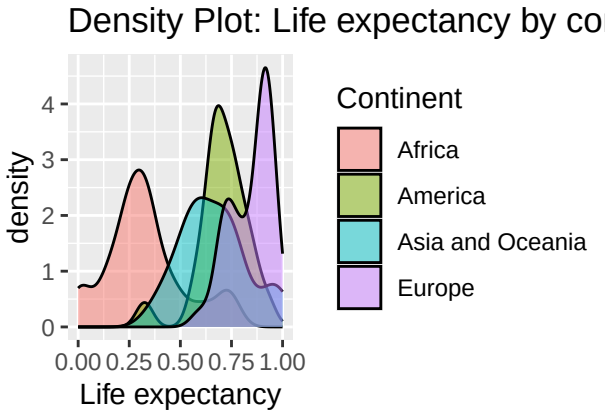


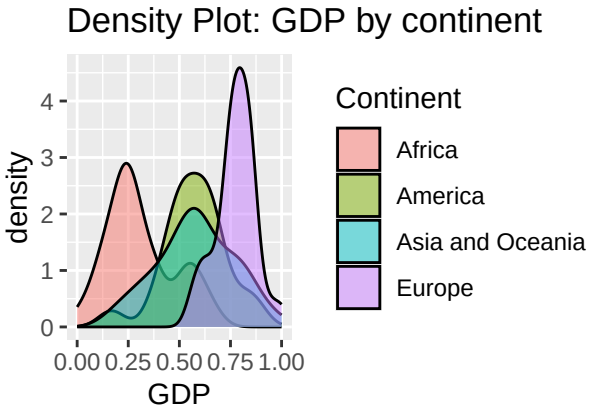
Density Plots

Consequently we plotted the density plots of the features to show the behavior of each region and to discover the adhesion to a particular trend by all of them or if there exists an anomaly way of acting by a particular part of the world. What we can notice is that the African continent is distinguished by very low values, exception made for the Fertility variable, in opposition to the Europe region. For what concerns the Asian and American region they show a very similar behavior in every feature that we observed.



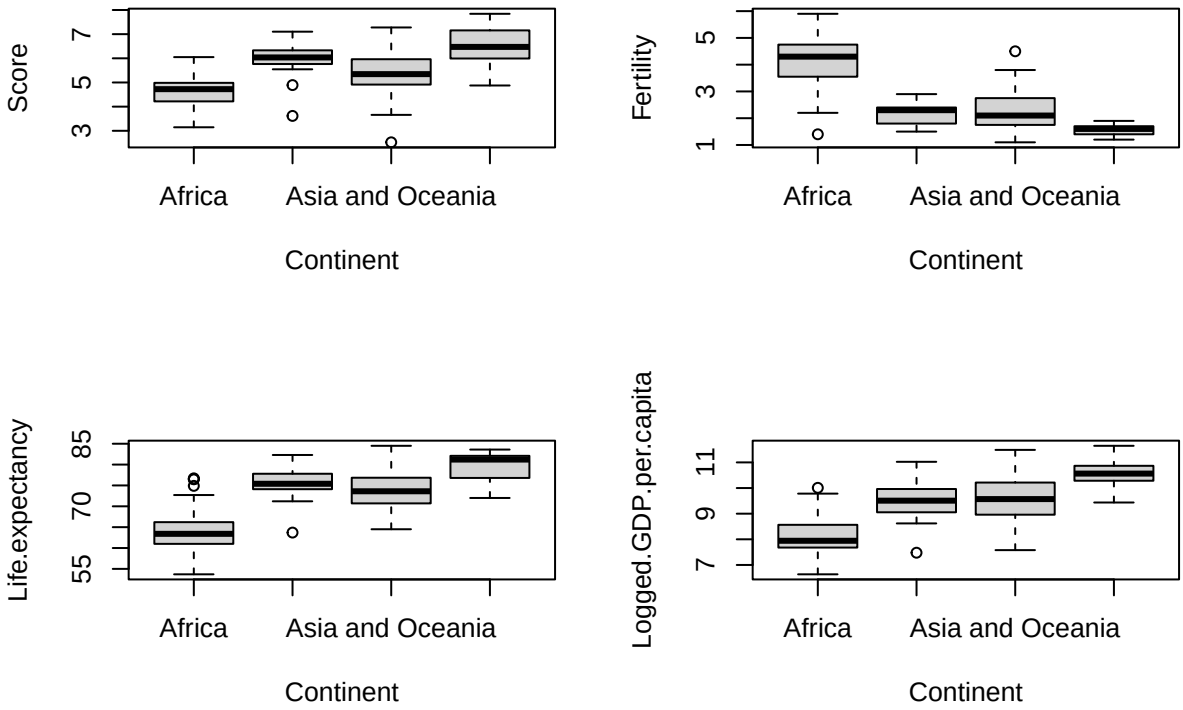






Boxplots

What we observed with the density plots is confirmed by the boxplots that we plotted below. A relevant thing to notice is the presence of outliers in the American and Asian regions, due to the subdivision that we performed earlier. Beside this few observation we can see that all the informations are in line with the previous analysis.



Regression

In this section we performed regression on the Happiness Score target variable. We tried to predict the value of the target variable happiness score based on the available explanatory variables.

Train/Test Split

In order to properly evaluate the goodness of our procedure, we split our dataset in two subsets, one for the training process and the other as assessment of the performance of our models.

```
set.seed(1)


data_size <- floor(0.85 * nrow(happiness_2021))

train_sample <- sample(seq_len(nrow(happiness_2021)), size = data_size, replace = FALSE)

train <- happiness_2021[train_sample, ]
test <- happiness_2021[-train_sample, ]
```

Variance Inflation Factor

We used VIF (Variance Inflation Factor) to assess the presence of multicollinearity and perform a first step in variable selection. Multicollinearity occurs when there is a high correlation between predictor variables in a regression model, which can cause issues in interpreting the model's coefficients and make the model less reliable.

```
#COLLINEARITY CHECK

lin_all <- lm(Score ~ .-Score-ISO.code-Continent-Region-Country, data = train)
vif(lin_all)
```

```
##                     Fertility               Life.expectancy
##                     10.645318                     39.218032
##                    Median.age             Population.growth
##                     13.230830                     11.070451
##                     Sex.ratio                  Suicide.rate
##                      1.703124                      1.992916
##             Urbanization.rate        Logged.GDP.per.capita
##                      3.228483                     12.168019
##                Social.support      Healthy.life.expectancy
##                      4.069715                     42.796362
## Freedom.to.make.life.choices                    Generosity
##                      2.279954                      1.306783
##    Perceptions.of.corruption
##                      1.805690
```

As we can see there are numerous variables with a really high VIF score, to perform a logical procedure we decide to remove one at the time the variables with the most multicollinearity issues, still taking into account some domain specific pieces of information in the removal process.

```
lin_to_try <- lm (Score ~ .-Score-ISO.code-Continent-Region-Country-Logged.GDP.per.capita
                  -Healthy.life.expectancy-Median.age-Fertility, data=train)
lin_m <- lm(Score ~ .-Score-ISO.code-Continent-Region-Country
            -Healthy.life.expectancy-Median.age-Fertility-Logged.GDP.per.capita
            -Life.expectancy, data=train )

vif(lin_to_try)
```

```
##               Life.expectancy             Population.growth
##                      5.176552                      3.069076
##                     Sex.ratio                  Suicide.rate
##                      1.251358                      1.808207
##             Urbanization.rate                Social.support
##                      2.149803                      3.298859
## Freedom.to.make.life.choices                    Generosity
##                      1.706787                      1.215899
##    Perceptions.of.corruption
##                      1.687128
```

```
vif(lin_m)
```

```
##             Population.growth                     Sex.ratio
##                      2.159489                      1.248035
##                  Suicide.rate             Urbanization.rate
##                      1.150243                      2.024450
##                Social.support Freedom.to.make.life.choices
##                      2.865775                      1.691033
##                    Generosity    Perceptions.of.corruption
##                      1.209460                      1.432150
```

```
columns_to_retain <- c("Population.growth", "Sex.ratio",
                       "Suicide.rate", "Urbanization.rate", "Social.support",
```

```
                        "Freedom.to.make.life.choices", "Generosity", "Perceptions.of.corruption",
                        "Life.expectancy", "Score")

train <- train[, columns_to_retain]
test <- test[, columns_to_retain]
```

So we performed two variables selection procedures in this step, in the first we procedeed to remove mechanically one by one the variables solely based on their value of the VIF score, resulting in the removal of the Life expectancy variable that in our opinion could have carried some useful information for the prediction task.

So what we tried to do was to remove the variables that most could contain redundant informations starting with the Logged.GDP, this resulted in keeping at the end the Life expectancy feature that finished in an acceptable range of VIF score.

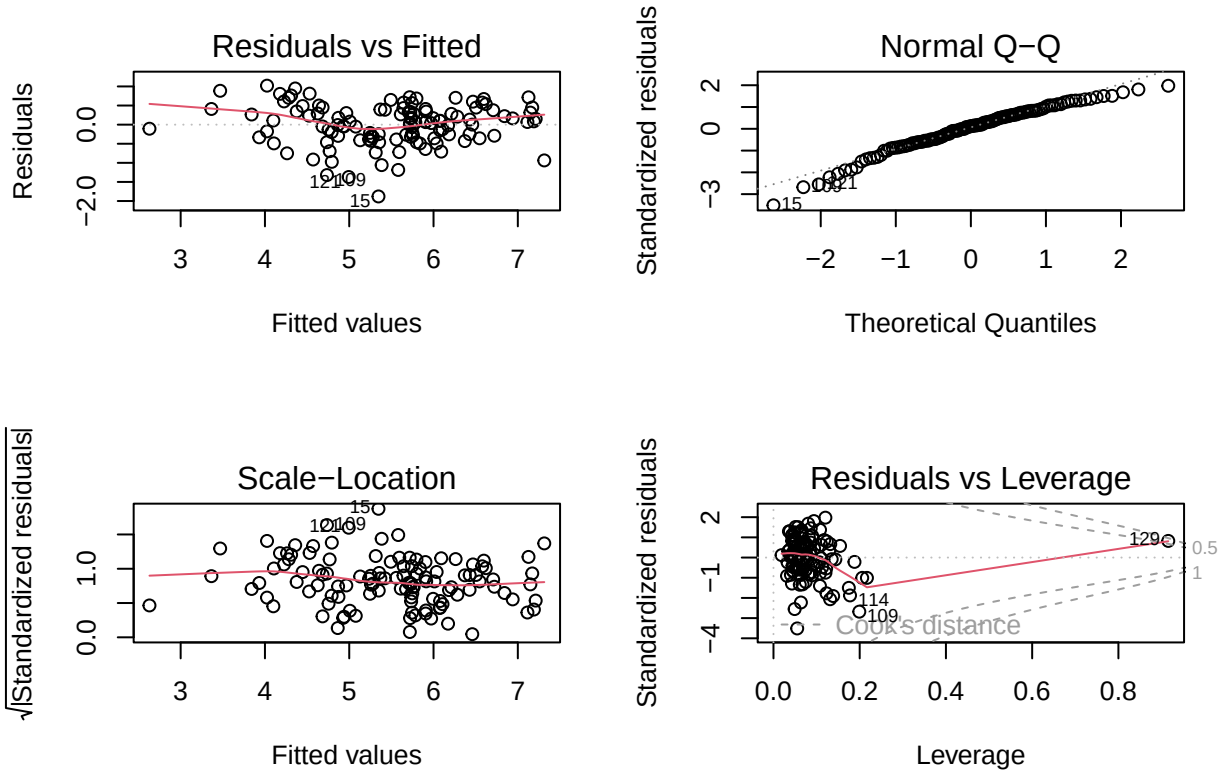
At the end we modified the training and test sets retrieving only the column that survived the VIF elimination procedure.

Here we present the summary of the a simple regression model on the new training set and the corresponding residuals plots, showing that all the linear assumption are satisfied and suggesting that a linear regression model is a suitable choice.

```
summary(lin_to_try)
```

```
##
## Call:
## lm(formula = Score ~ . - Score - ISO.code - Continent - Region -
##     Country - Logged.GDP.per.capita - Healthy.life.expectancy -
##     Median.age - Fertility, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88004 -0.32103  0.06553  0.37737  1.02010
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.08291    0.86019  -0.096  0.92339
## Life.expectancy               1.00251    0.46753   2.144  0.03427 *
## Population.growth             0.19365    0.43643   0.444  0.65814
## Sex.ratio                    -0.13404    0.62539  -0.214  0.83070
## Suicide.rate                  0.18514    0.34371   0.539  0.59124
## Urbanization.rate             0.92937    0.30672   3.030  0.00307 **
## Social.support                3.19341    0.76653   4.166 6.30e-05 ***
## Freedom.to.make.life.choices  2.45572    0.58996   4.162 6.39e-05 ***
## Generosity                    0.46896    0.32055   1.463  0.14641
## Perceptions.of.corruption    -0.48954    0.36631  -1.336  0.18425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5506 on 107 degrees of freedom
## Multiple R-squared:  0.7516, Adjusted R-squared:  0.7307
## F-statistic: 35.98 on 9 and 107 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lin_to_try)
```

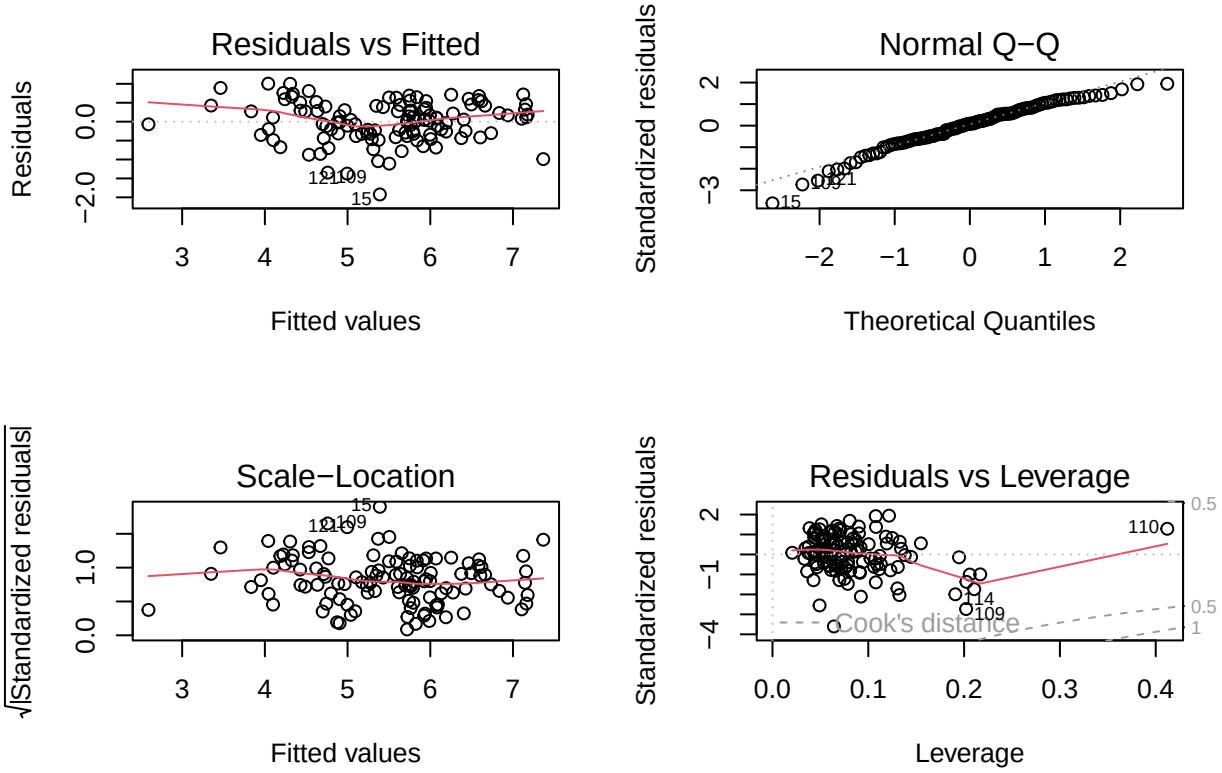


Leverage Point Analysis The residuals vs leverage plot highlights the presence of a leverage point (with index 129), that corresponds to the United Arab Emirates state. The presence of a leverage point could influence the performances of a linear regression model, by changing significantly some of its coefficients.

After careful considerations we decided to keep this country in our training set, since it could be representative of a specific class of states and shouldn't in this way deviate strongly our model.

This has been confirmed by a naive implementation of a linear regression model trained on the same dataset but without the presence of that particular point. The performance slightly decreased, showing that its presence provides valuable insights.

```
# Removing United Arab Emirates (index 129)
train <- train[-2,]
lin_m <- lm(Score ~ .-Score, data=train)

par(mfrow=c(2,2))
plot(lin_m)
```




```
summary(lin_m)
```

```
##
## Call:
## lm(formula = Score ~ . - Score, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92326 -0.31914  0.04752  0.38821  1.00418
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.002126   0.867659   0.002  0.99805
## Population.growth               0.267723   0.446255   0.600  0.54983
## Sex.ratio                      -1.739499   2.048491  -0.849  0.39771
## Suicide.rate                    0.088861   0.363562   0.244  0.80738
## Urbanization.rate               0.971777   0.311468   3.120  0.00233 **
## Social.support                  3.168216   0.768296   4.124 7.43e-05 ***
## Freedom.to.make.life.choices    2.490553   0.592368   4.204 5.48e-05 ***
## Generosity                      0.468348   0.321035   1.459  0.14756
## Perceptions.of.corruption      -0.489388   0.366867  -1.334  0.18507
## Life.expectancy                 0.978132   0.469168   2.085  0.03949 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5515 on 106 degrees of freedom
```

```
## Multiple R-squared:  0.751,  Adjusted R-squared:  0.7299
## F-statistic: 35.52 on 9 and 106 DF,  p-value: < 2.2e-16
```

Stepwise Variable selection

A deeper study of the choice of the most relevant features for the prediction task has been conducted by the mean of the implementation of a stepwise variable selection procedure. We both implemented the Forward and the Backward selection in order to obtain a solid comparison, and doing so, assess the most robust and consistent methodology for the creation of our final model.

```
varmax <-10
mod.bw <- regsubsets(Score ~ ., nvmax=varmax, method= "backward", data=train)
summary_bw <- summary(mod.bw)
summary_bw$rsq
```

```
## [1] 0.5902389 0.6645931 0.7165735 0.7324398 0.7434193 0.7485131 0.7501519
## [8] 0.7508664 0.7510067
```

```
mod.fw <- regsubsets(Score ~ ., nvmax=varmax, method= "forward", data=train)
summary_fw <- summary(mod.fw)

summary_bw$rsq
```

```
## [1] 0.5902389 0.6645931 0.7165735 0.7324398 0.7434193 0.7485131 0.7501519
## [8] 0.7508664 0.7510067
```

We plotted the loss curve of four different scores:

- RSS
- Adjusted R^2
- Mallow's Cp
- BIC

We than identified at which number of the selected variables the different models, obtained with this techniques, performed the best, applying a sort of majority voting to select the final number of parameters to adopt. We then compared forward and backward variables selection, and both methods converged to the same subset choices of variables.

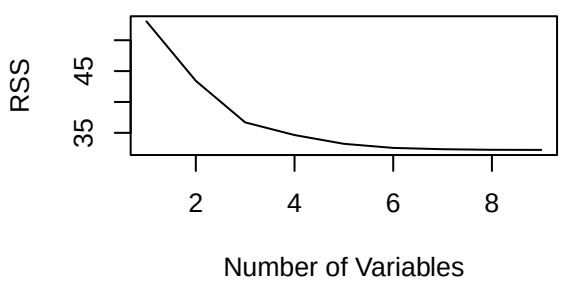
```
par(mfrow=c(2,2))

#FORWARD SELECTION PLOT ANALYSIS

# residual sum of squares
plot(summary_fw$rss,xlab="Number of Variables",ylab="RSS",type="l")

# adjusted-R^2 with its largest value
plot(summary_fw$adjr2,xlab="Number of Variables",ylab="Adjusted Rsq",type="l")
max_r2 <- which.max(summary_fw$adjr2)
points(max_r2,summary_fw$adjr2[max_r2], col="red",cex=2,pch=20)

# Selected variables:
summary_fw$which[max_r2,]
```

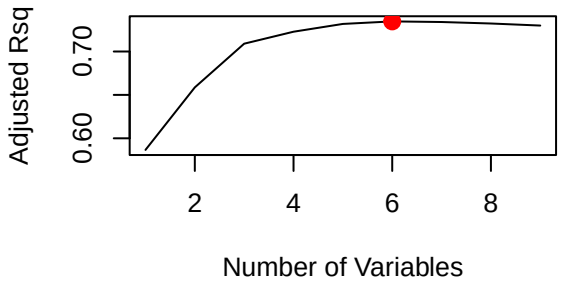
Forward Stepwise

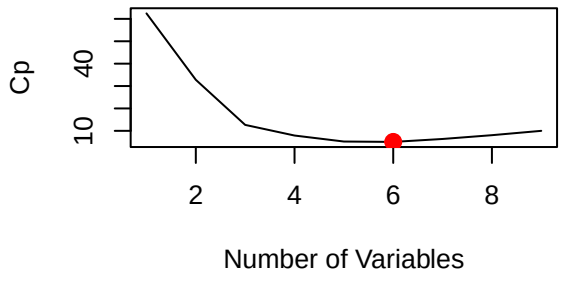
```
##                  (Intercept)            Population.growth
##                         TRUE                        FALSE
```

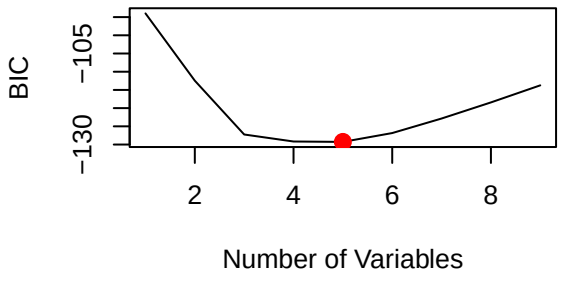
```
##                     Sex.ratio                Suicide.rate
##                         FALSE                       FALSE
##           Urbanization.rate              Social.support
##                          TRUE                        TRUE
## Freedom.to.make.life.choices                  Generosity
##                          TRUE                        TRUE
##    Perceptions.of.corruption             Life.expectancy
##                          TRUE                        TRUE
```

```
# Mallow's Cp with its smallest value
plot(summary_fw$cp,xlab="Number of Variables",ylab="Cp",type='l')
min_cp <-which.min(summary_fw$cp)
points(min_cp,summary_fw$cp[min_cp],col="red",cex=2,pch=20)

# BIC with its smallest value
plot(summary_fw$bic,xlab="Number of Variables",ylab="BIC",type='l')
min_bc <- which.min(summary_fw$bic)
points(min_bc,summary_fw$bic[min_bc],col="red",cex=2,pch=20)
```









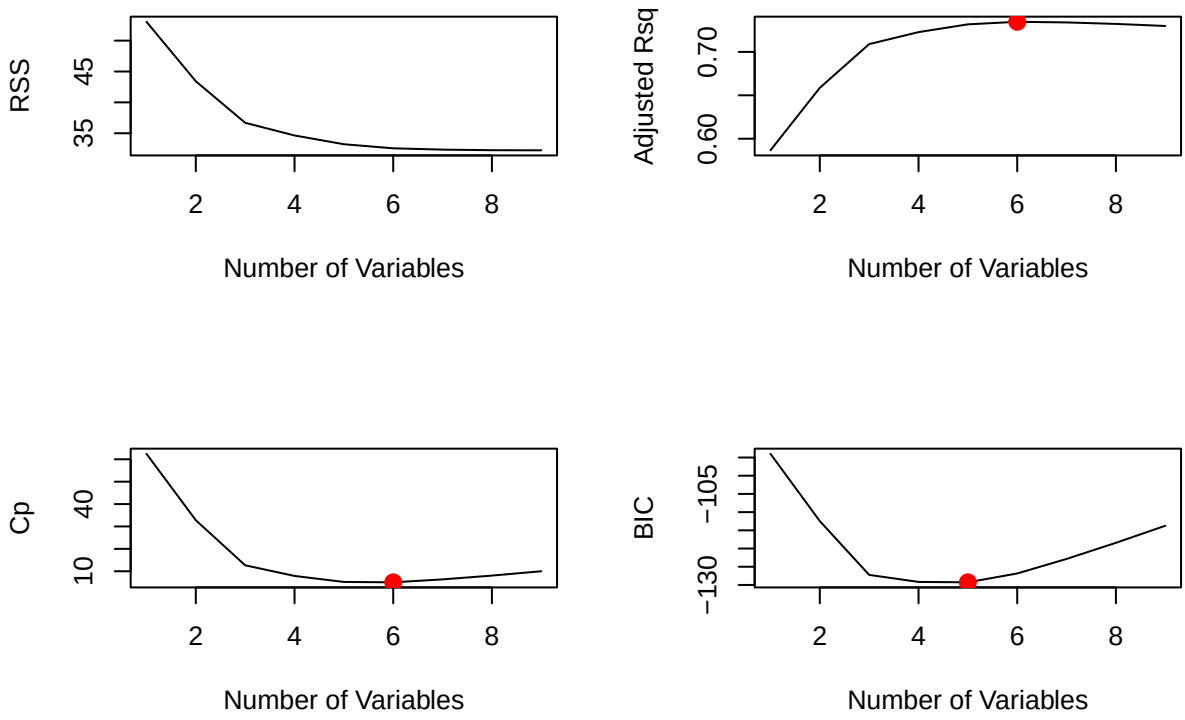

```
par(mfrow=c(2,2))

#BACKWARD SELECTION PLOT ANALYSIS

# residual sum of squares
plot(summary_bw$rss,xlab="Number of Variables",ylab="RSS",type="l")
```

```
# adjusted-R^2 with its largest value
plot(summary_bw$adjr2,xlab="Number of Variables",ylab="Adjusted Rsq",type="l")
max_r2 <- which.max(summary_bw$adjr2)
points(max_r2,summary_bw$adjr2[max_r2], col="red",cex=2,pch=20)

# Selected variables:
summary_bw$which[max_r2,]
```

Backward Stepwise

```
##                   (Intercept)             Population.growth
##                          TRUE                         FALSE
##                     Sex.ratio                  Suicide.rate
##                         FALSE                         FALSE
##            Urbanization.rate                Social.support
##                          TRUE                          TRUE
## Freedom.to.make.life.choices                    Generosity
##                          TRUE                          TRUE
##    Perceptions.of.corruption              Life.expectancy
##                          TRUE                          TRUE
```

```
# Mallow's Cp with its smallest value
plot(summary_bw$cp,xlab="Number of Variables",ylab="Cp",type='l')
min_cp <- which.min(summary_bw$cp)
points(min_cp,summary_bw$cp[min_cp],col="red",cex=2,pch=20)

# BIC with its smallest value
plot(summary_bw$bic,xlab="Number of Variables",ylab="BIC",type='l')
min_bc <- which.min(summary_bw$bic)
points(min_bc,summary_bw$bic[min_bc],col="red",cex=2,pch=20)
```



Regularization Techniques

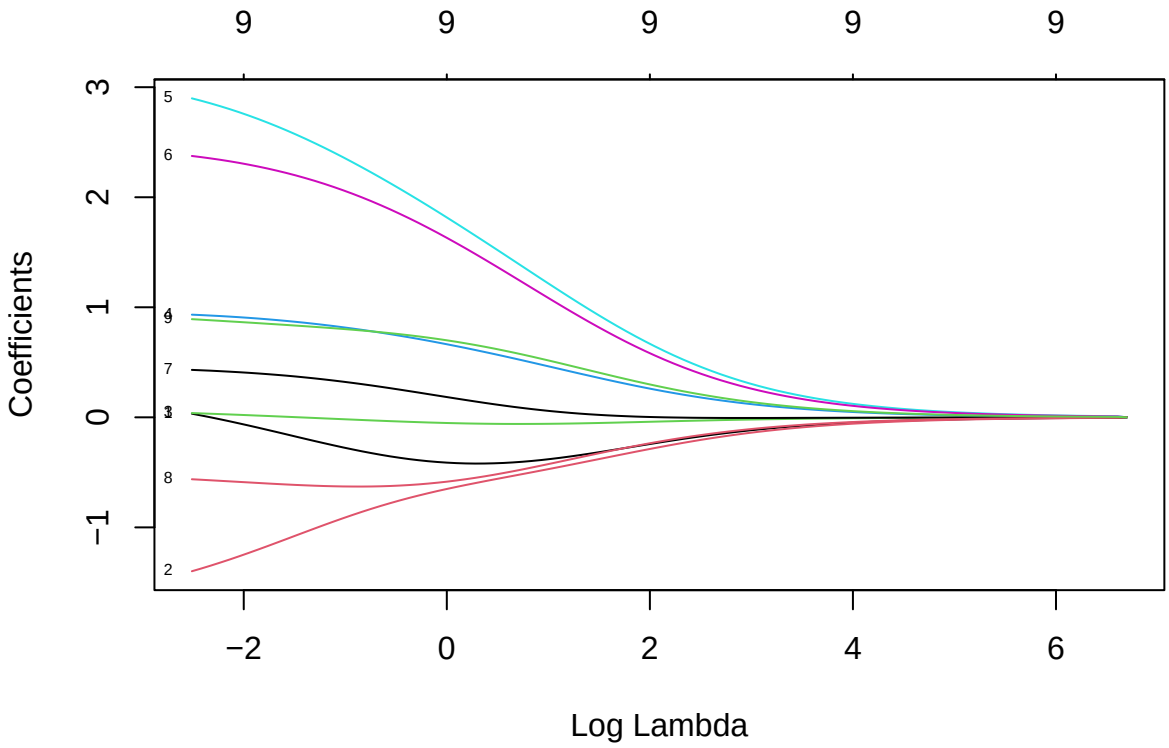
Another way to approach the variable selection step can be done by using Ridge and Lasso regression: both procedures consist in fitting a model that contains all the predictors, using at the same time the introduction of a penalty that constrains or regularizes the coefficient estimates.

Ridge regression assumes that all predictors are relevant to the model and it applies non-zero shrinkage to all coefficients.

Lasso regression instead has the ability to perform variable selection by driving the coefficients of irrelevant, or less important, predictors to exactly zero.

Ridge Regression

```
ridge.mod <- glmnet(X, train$Score, alpha=0, thresh = 1e-12, data=train)

# add labels to identify the variables
plot(ridge.mod, xvar="lambda", label=TRUE)
```

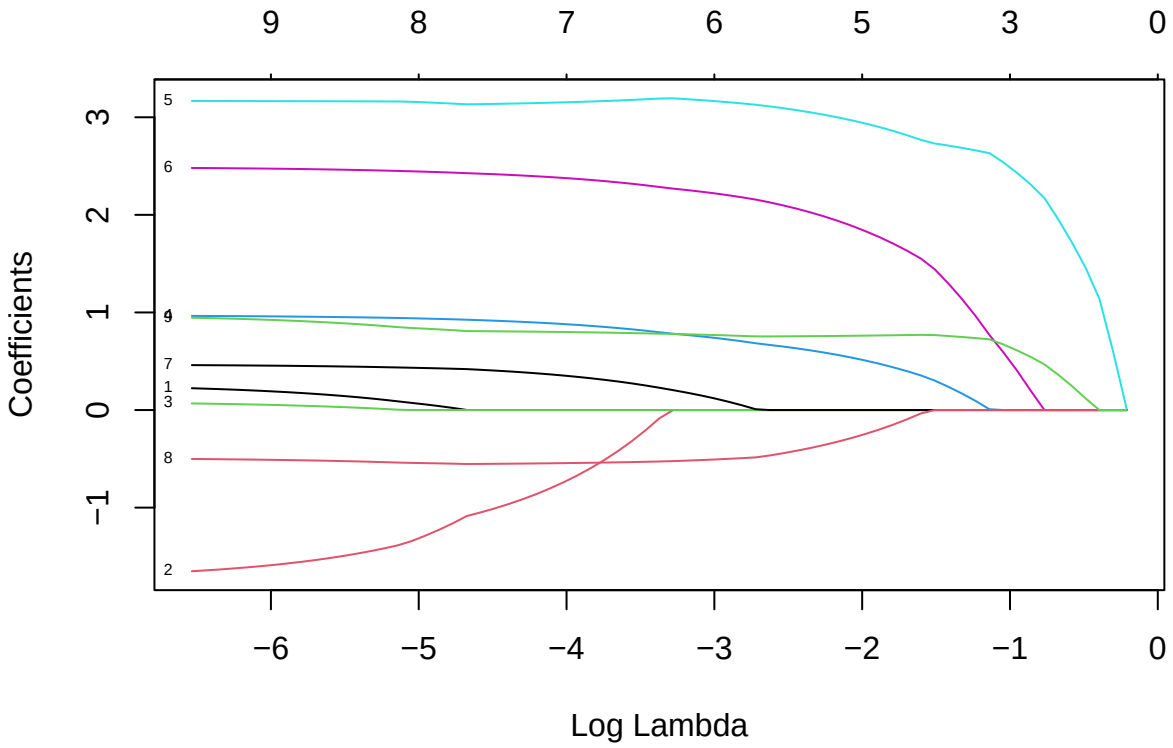


Each curve corresponds to the regression coefficient estimates for one of the variables, plotted as a function of log lambda.

Lasso Regression

In the case of the Lasso regression the penalty forces some of the coefficient estimates to be exactly zero with a proper large value for log lambda.

```
# apply lasso to the training set
lasso.mod <- glmnet(X, train$Score, alpha=1, thresh = 1e-12, data=train)

# add labels to identify the variables
plot(lasso.mod, xvar="lambda", label=TRUE)
```



Models Selection Here we performed a 10 fold cross-validation on the training set for both methods in order to find the best lambda values and then we compared the two obtained models.

```
# apply 10 fold cross-validation to the training set

set.seed(10)
cv.out.lasso <- cv.glmnet(as.matrix(X), train$Score, alpha=1)

# estimated test MSE

bestlambda <- cv.out.lasso$lambda.min
lasso.pred <- predict(lasso.mod, s=bestlambda, newx=as.matrix(X_test))
mse.lasso <- mean((lasso.pred-test$Score)^2)

mse.lasso
```

```
## [1] 0.2764902
```

```
bestlambda
```

```
## [1] 0.02596661
```

```
best_lasso <- glmnet(X, train$Score, alpha=1, lambda = bestlambda, thresh = 1e-12, data=train)

coefficients <- coef(best_lasso)
coefficients
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                                    s0
## (Intercept)                 0.4798075
## Population.growth           .
## Sex.ratio                  -0.4193130
## Suicide.rate                .
## Urbanization.rate           0.8399457
## Social.support              3.1718549
## Freedom.to.make.life.choices  2.3329579
## Generosity                  0.2922046
## Perceptions.of.corruption  -0.5357756
## Life.expectancy             0.7912851
```

```
# apply 10 fold cross-validation to the training set

set.seed(10)
cv.out.ridge <- cv.glmnet(as.matrix(X), train$Score, alpha=0)

# estimated test MSE

bestlambda <- cv.out.ridge$lambda.min
ridge.pred <- predict(ridge.mod, s=bestlambda, newx=as.matrix(X_test))
mse.ridge <- mean((ridge.pred-test$Score)^2)

mse.ridge
```

```
## [1] 0.2674599
```

```
bestlambda
```

```
## [1] 0.1177554
```

```
best_ridge <- glmnet(X, train$Score, alpha=0, lambda = bestlambda, thresh = 1e-12, data=train)


coefficients <- coef(best_ridge)
coefficients
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                                     s0
## (Intercept)                  0.75985642
## Population.growth           -0.03611376
## Sex.ratio                   -1.29196173
## Suicide.rate                 0.02610171
## Urbanization.rate            0.91528063
## Social.support               2.80083294
## Freedom.to.make.life.choices  2.32674566
## Generosity                   0.41499781
## Perceptions.of.corruption   -0.58170436
## Life.expectancy              0.87176872
```

```
model_try <- lm(Score ~ .-Score-Population.growth-Sex.ratio-Suicide.rate, data = train)
summary(model_try)
```

```
##
## Call:
## lm(formula = Score ~ . - Score - Population.growth - Sex.ratio -
##     Suicide.rate, data = train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89359 -0.33279  0.06792  0.36556  1.05883
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.2156     0.6085   0.354  0.72382
## Urbanization.rate               0.9125     0.3000   3.041  0.00295 **
## Social.support                  3.2813     0.6824   4.808 4.91e-06 ***
## Freedom.to.make.life.choices    2.4185     0.5823   4.154 6.52e-05 ***
## Generosity                      0.4650     0.3130   1.486  0.14020
## Perceptions.of.corruption      -0.5744     0.3363  -1.708  0.09051 .
## Life.expectancy                 0.8122     0.3370   2.410  0.01761 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5465 on 109 degrees of freedom
## Multiple R-squared:  0.7485, Adjusted R-squared:  0.7347
## F-statistic: 54.07 on 6 and 109 DF,  p-value: < 2.2e-16
```

```
linear_pred <- predict(model_try, X_test)
mse.linear <- mean((linear_pred-test$Score)^2)
mse.linear
```

```
## [1] 0.2669179
```

```
SSE <- sum((lasso.pred - (test$Score))^2)
SST <- sum((test$Score - mean(test$Score))^2)
R_square <- 1 - SSE / SST
n=117
k=7
adj_R_square <- 1 - ((1-R_square)* (n-1)/ (n-k-1))
```

After having explored all these alternatives for the subset selection of our variables we build a model for each technique and compare the results obtained with the MSE metric on our separated test set.

Model	MSE
Stepwise models	0.2669179
Ridge Regression	0.2674599
Lasso Regression	0.2764902

What we can notice is that the best and lowest values has been obtained by the stepwise based regression models, so what we can state looking at the coefficients of each variable retrieved by this model is that the variables that with a one-unit increase of its value brings the strongest contribution to the predicted happiness score are Social support and Freedom to make life choices.

This is in line with our initial expectations even though those weren't the ones that we would have guessed the most.

Discussion and Conclusion

The analysis we carried out so far confirmed the hypothesis made in the beginning regarding the existence of interesting and useful linear relationships between some explanatory variables and the target happiness score.

The two most significant explanatory variables in the final linear regression model were **Freedom to make Life Choices** and **Social Support**. Both variables obtained also the highest coefficients.

Even though both variables are inherently positively correlated with the GDP per capita, and so we cannot completely ignore the inference of economical wellness on the happiness of peoples, it is reassuring and interesting to notice that our main predictors are both putting a focus on human and social relationships and personal freedom.