World Happiness Report: Statistical Analysis

Statistical Learning course project

Chiarello Federico Mat. 2058163
Pivato Davide Mat. 2056101
Nanni Sara Mat. 2087621

01 Dataset

Description and **Preprocessing**

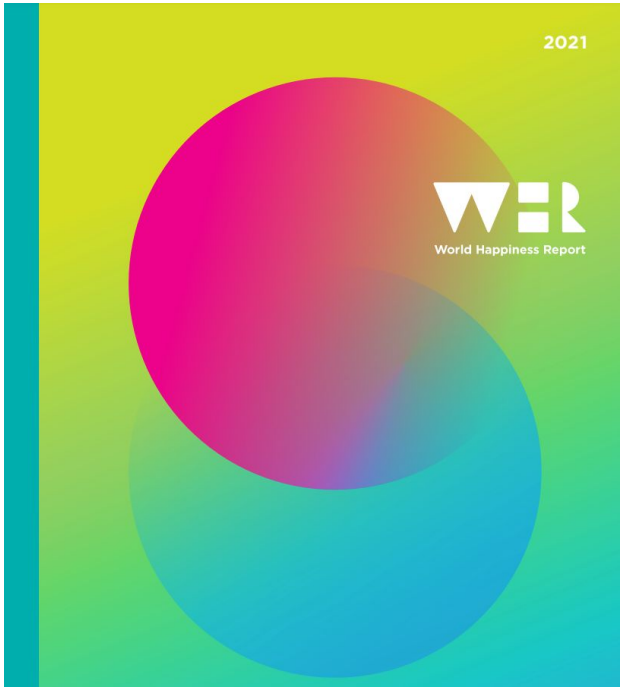
Datasets

World Happiness Report 2021:

social and economical features
for each country

World Data 2021:

demographic informations for each state



Happiness Score

- Score that estimates the happiness in a country
- Range: 0 - 10
- Yearly publication based on data collected from surveys conducted by Gallup World Poll

Happiness Report 2021

GDP per capita

indicator of the economic well-being

Social Support

opportunity of having someone to count on

Healthy Life Expectancy

good physical and mental health

Generosity

engagement in a positive community

Freedom to Make Life Choices

freedom perceived by individuals

Perception of Corruption

corruption perceived within a country

World Data 2021

Fertility	Sex Ratio
Median Age	Life Expectancy
Population Growth	Suicide Rate
Urbanization Rate	

Pre-processing

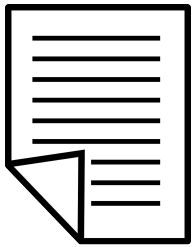
NAs removal

NA value removal is an important step in statistical analysis to address several keys concern and ensure the integrity and validity of the data and subsequent analysis.

Preliminary omissions

Domain specific actions that ensures not to use redundant or useless features that will compromise the effectiveness of the analysis

Final Dataset



Source	Kaggle
Merging procedure	Carried out on ISO codes
Number of rows	138
Number of features	18

```
'data.frame':   138 obs. of  18 variables:
 $ Country                    : chr  "Afghanistan" "Algeria" "Argentina" "Armenia" ...
 $ ISO.code                   : chr  "AFG" "DZA" "ARG" "ARM" ...
 $ Fertility                  : num  4.5 3 2.3 1.8 1.7 1.5 1.7 2 2 1.4 ...
 $ Life.expectancy            : num  64.5 76.7 76.5 74.9 83.3 81.4 72.9 77.2 72.3 74.6 ...
 $ Median.age                 : num  27.4 28.1 31.7 35.1 38.7 44 32.3 32.3 26.7 40 ...
 $ Population.growth          : num  2.41 1.89 0.88 0.17 1.6 0.46 1.35 1.92 1.19 -0.1 ...
 $ Sex.ratio                  : num  1.03 1.03 0.98 0.95 0.99 0.96 0.98 1.53 0.97 0.87 ...
 $ Suicide.rate               : num  6.4 3.3 9.1 5.7 11.7 11.4 2.6 5.7 6.1 21.4 ...
 $ Urbanization.rate          : num  26 73.7 92.1 63.3 86.2 58.7 56.4 89.5 38.2 79.5 ...
 $ Region                     : chr  "South Asia" "Middle East and North Africa" "Latin America and Caribbean"
 $ Continent                  : chr  "Asia" "Africa" "Latin America" "Asia" ...
 $ Score                      : num  2.52 4.89 5.93 5.28 7.18 ...
 $ Logged.GDP.per.capita      : num  7.7 9.34 9.96 9.49 10.8 ...
 $ Social.support             : num  0.463 0.802 0.898 0.799 0.94 0.934 0.836 0.862 0.693 0.91 ...
 $ Healthy.life.expectancy    : num  52.5 66 69 67.1 73.9 ...
 $ Freedom.to.make.life.choices: num  0.382 0.48 0.828 0.825 0.914 0.908 0.814 0.925 0.877 0.65 ...
 $ Generosity                 : num  -0.102 -0.067 -0.182 -0.168 0.159 0.042 -0.223 0.089 -0.041 -0.18 ...
 $ Perceptions.of.corruption  : num  0.924 0.752 0.834 0.629 0.442 0.481 0.506 0.722 0.682 0.627 ...
```

02 Goals

Goals and **Hypothesis**

Project Goals

Our aim is to investigate if there exist and which are the most influential factors that contribute in determining the happiness score of a specific nation

Does money buy happiness?

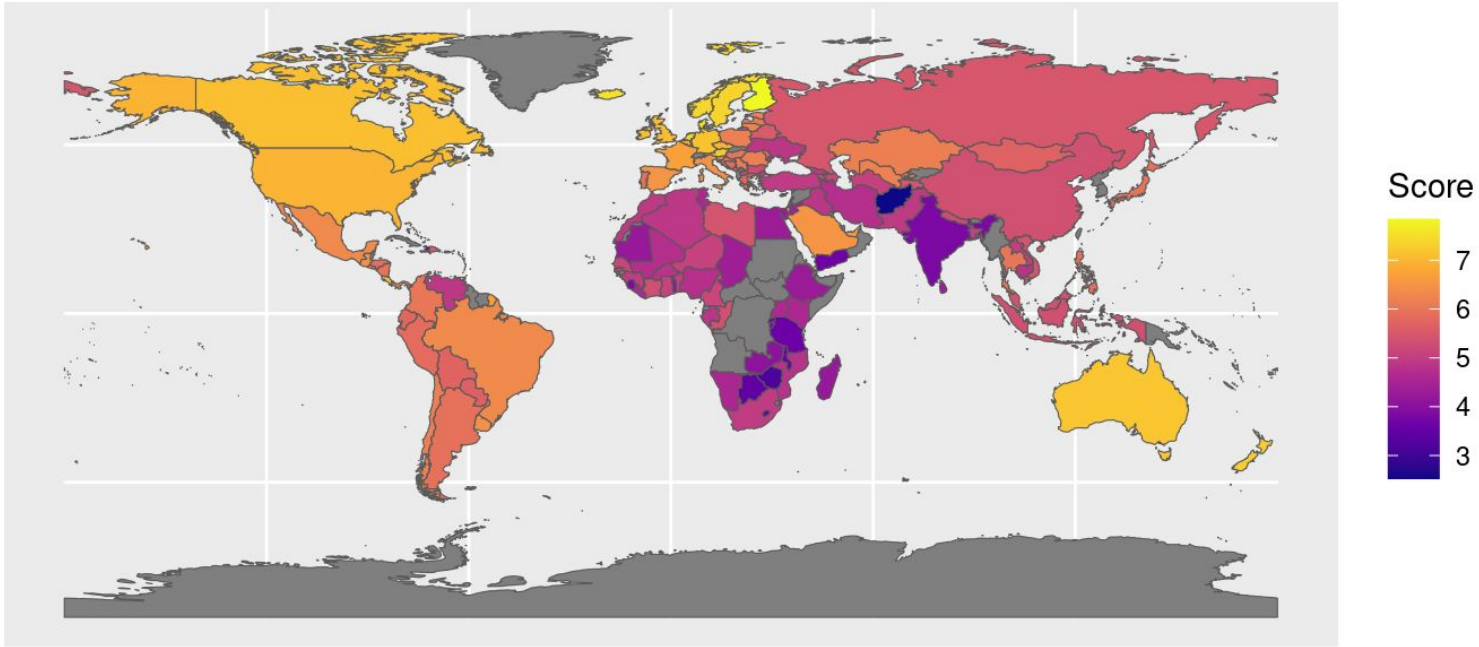
Can other people affect the quality of our lives?

Does live longer bring to a happier life?

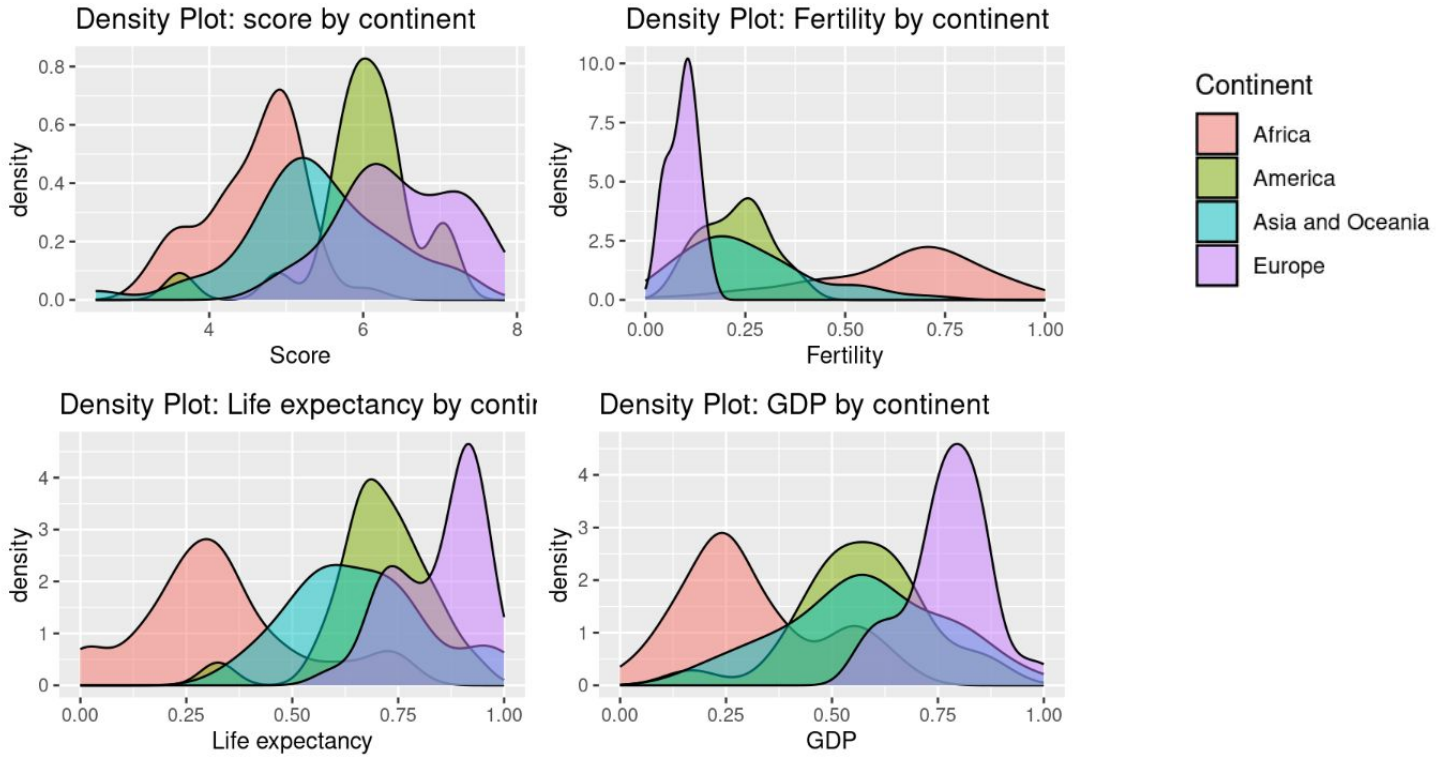
03 Data Exploration

Correlations and **Insights**

Infographic Map on Score



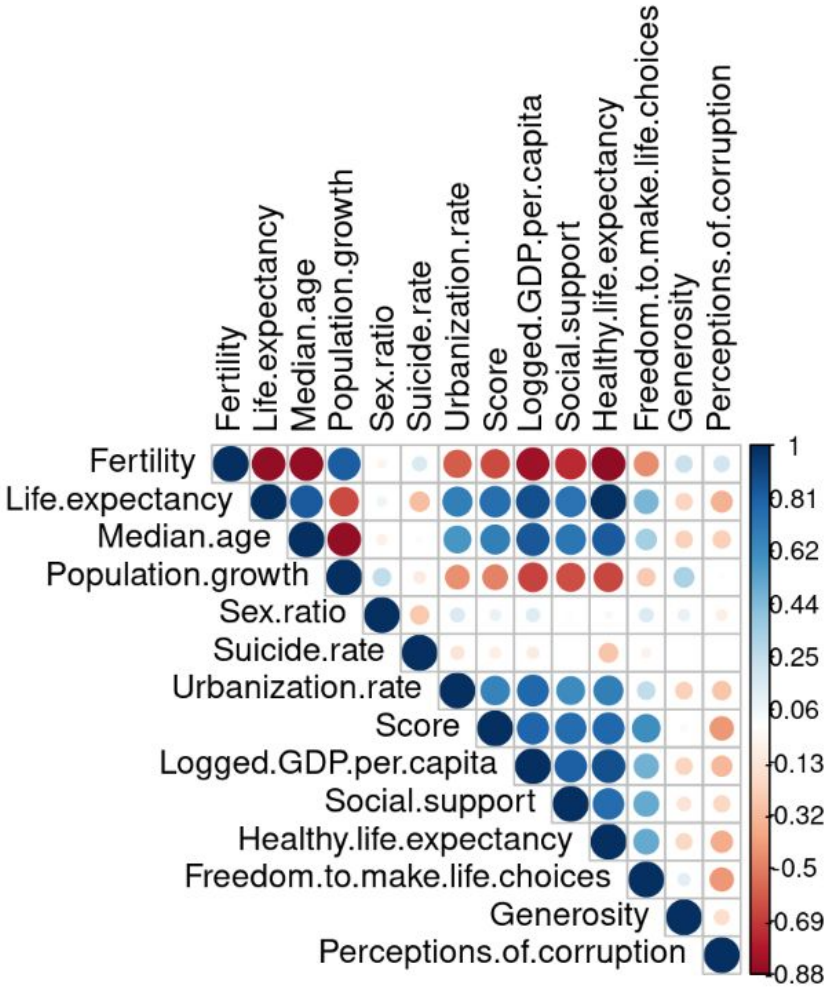
Density Plots



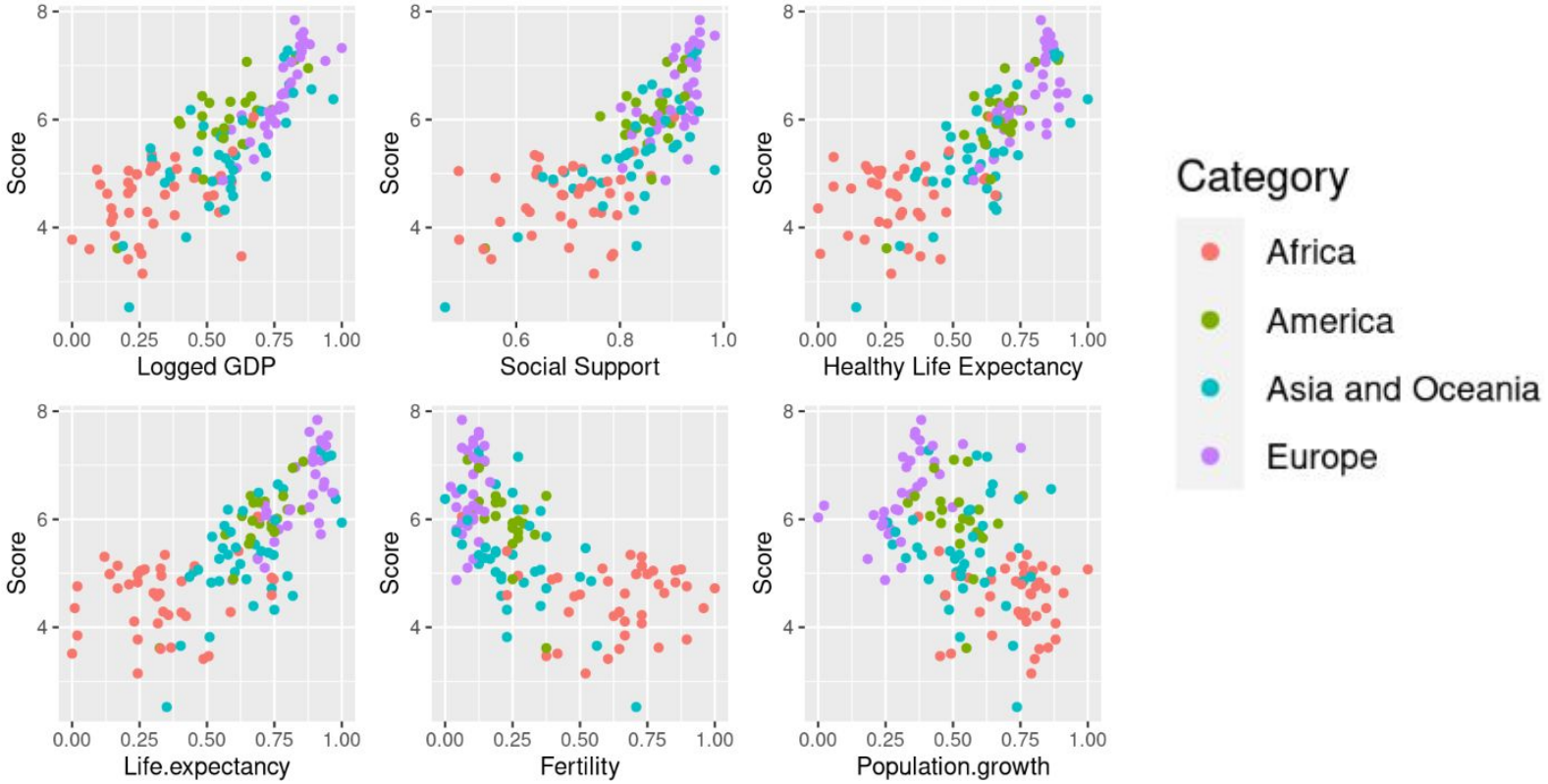
Correlation Analysis

Very **strong correlations**, both positive or negative, existing between some variables

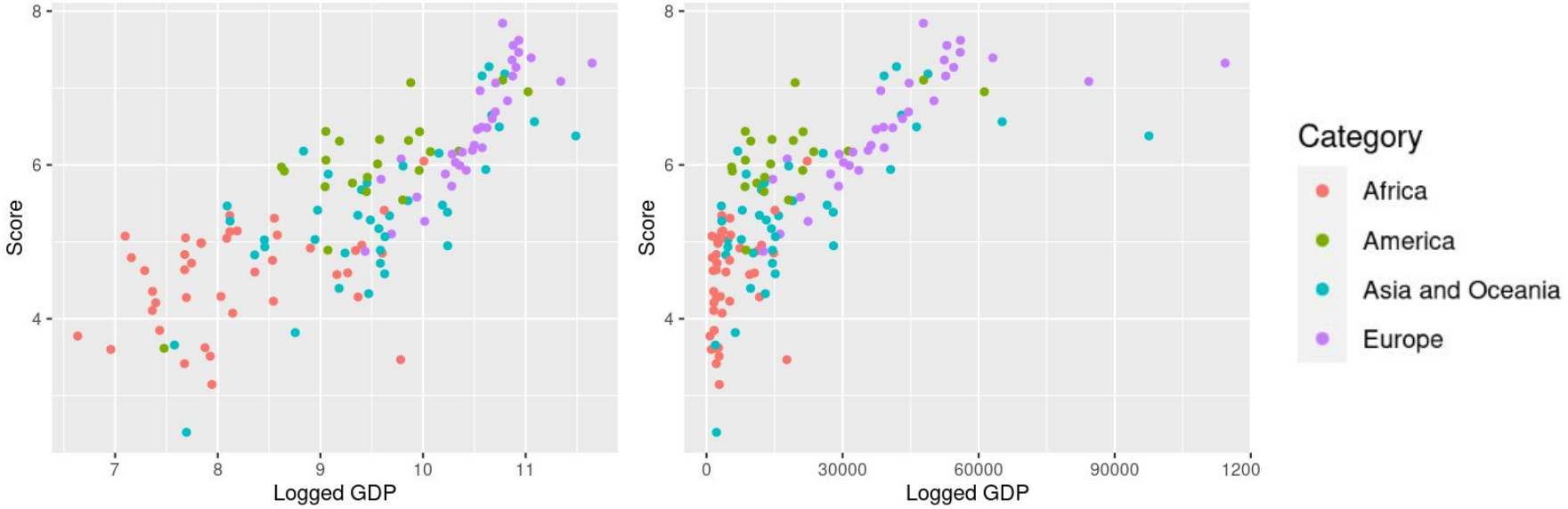
Risk of **redundant information** among data



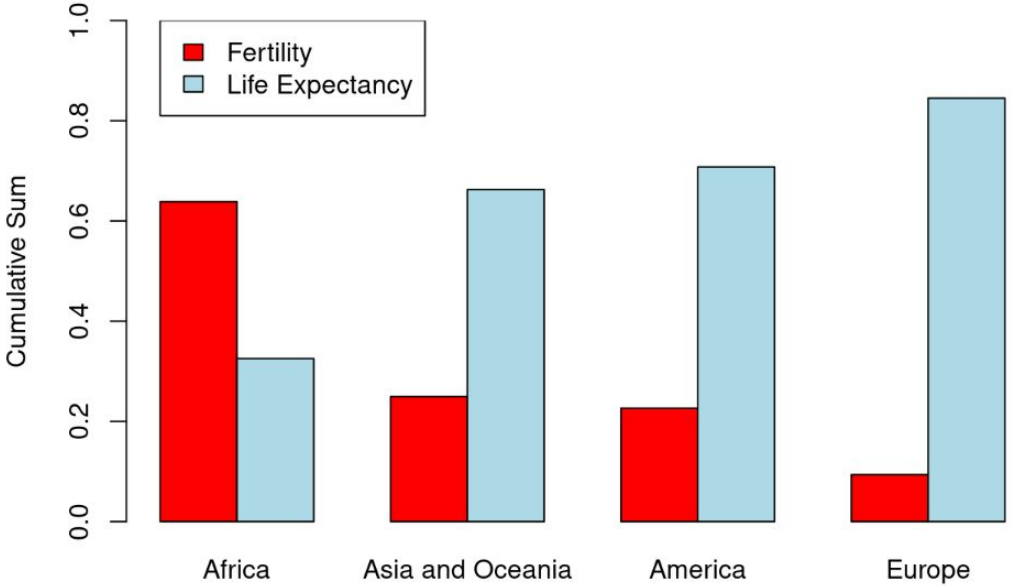
Scatterplots



GDP vs Logged GDP



Fertility vs Life Expectancy



Negative Linear Correlation

Important **Sociological** Insight

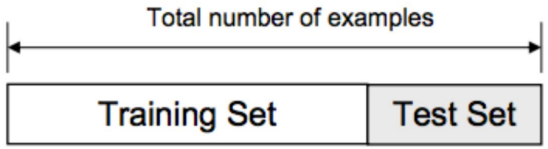
04 Prediction

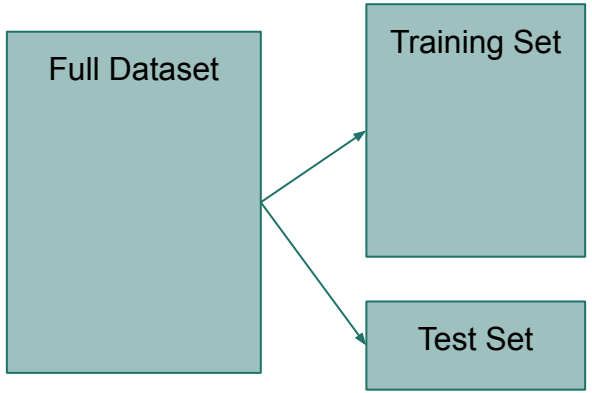
Regression models

Train - Test Split

We splitted the dataset into two subsets:

- **Train (85%)**: used for training
- **Test (15%)**: used for evaluating the models





VIF Analysis

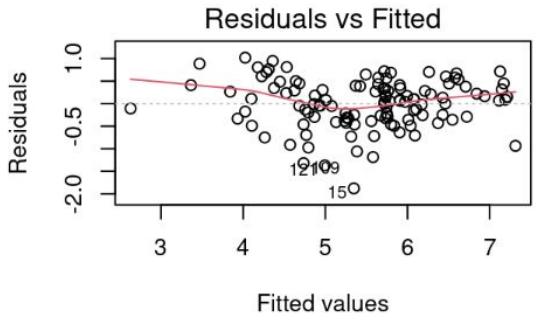
(Variance Inflation Factor)

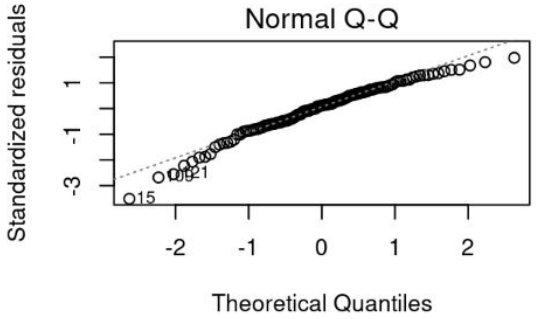
Metric used in **multicollinearity** analysis:

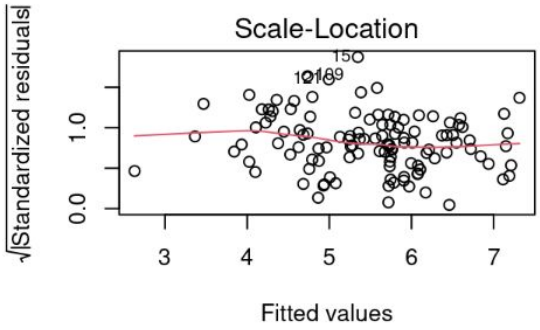
helps assessing the level of multicollinearity among the independent variables in a **regression model**

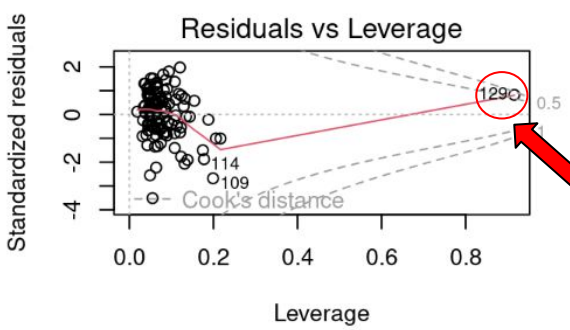
```
##              Life.expectancy                Population.growth
##                     5.176552                         3.069076
##                    Sex.ratio                     Suicide.rate
##                     1.251358                         1.808207
##            Urbanization.rate     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

Residual Analysis





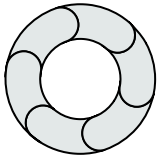




High Leverage Point

How to treat it?

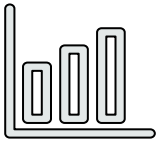
Variable selection



Stepwise procedures

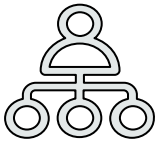
Implemented Forward and Backward stepwise selection

Comparison on different scores



Ridge regression

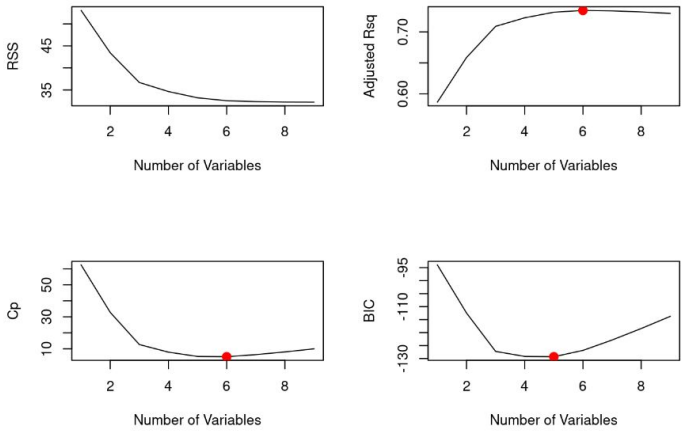
Retaining all the variables available, applying L2 norm penalty



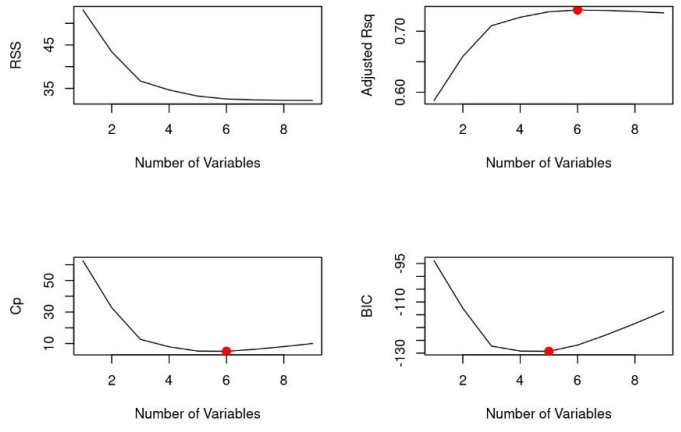
Lasso Regression

Variable selection procedure forcing some of the variables coefficients to 0

Stepwise Variable Selection



Forward Selection

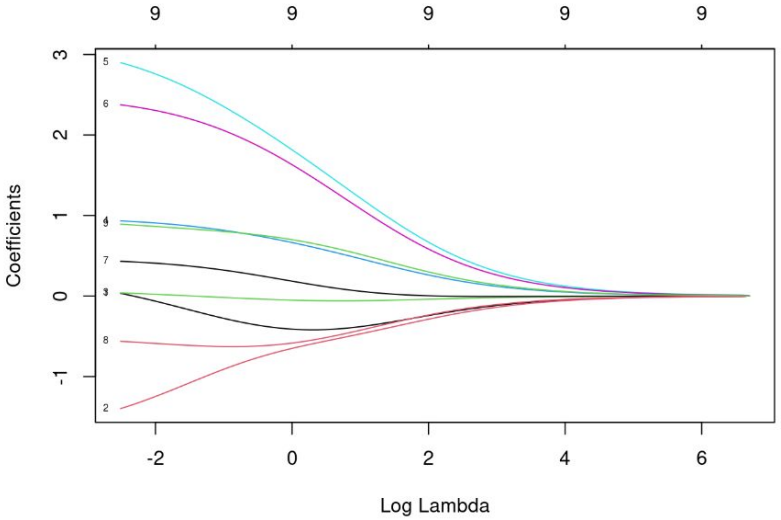


Backward Selection

Majority voting on the number of variables to retain based on different metrics

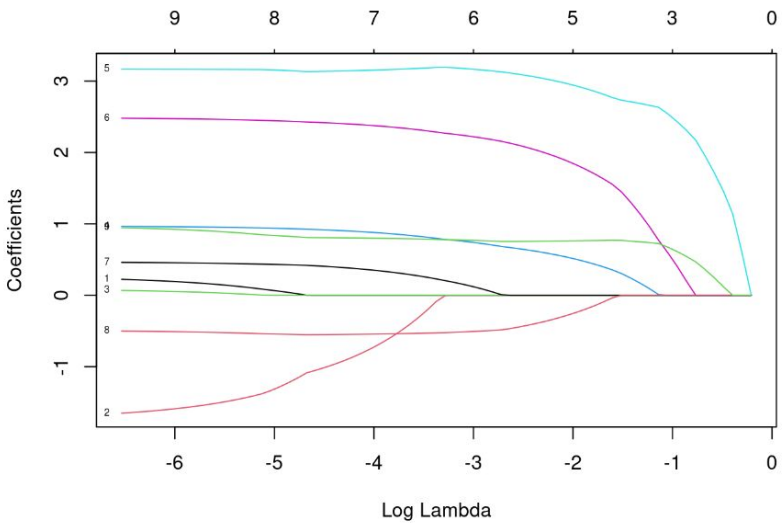
Regularization Techniques

Ridge Regression



Best λ = 0.1177554

Lasso Regression



Best λ = 0.02596661

Results & Model Selection

To perform the **model selection** step, we decided to compute the **MSE** metric for the prediction obtained on the **Test set** for each model we built

Model	MSE
Stepwise models	0.2669179
Ridge Regression	0.2674599
Lasso Regression	0.2764902

Lowest score = Stepwise selection model

Final Regression Model

```
##
## Call:
## lm(formula = 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
```

05 Conclusions

Does money bring happiness?

Key features:

- ***Freedom to Make Life Choices***
- ***Social support***

Even if most significant variables are positively correlated with the GDP per capita, it is interesting to notice that our main predictors are both putting a focus on ***human and social relationships*** and ***personal freedom***

Thanks for your attention!

Chiarello Federico - Pivato Davide - Nanni Sara