Final Project Stat 5310

Otavek Aliev | Oscar Pacheco

2022-12-08

Introduction

Several psychologists agree that the main goal of rational human beings is to find happiness irrespective of their geographical locations, the time they lived or other demographic variables. Happiness is defined differently in each community, hence, there is no common ground to describe happiness (McMahon, 2008). However, there are entities that go deep in trying to answer and assess the happiness of the world, by taking worldwide surveys in all the countries, and utilizing distinct factors, to measure happiness.

For our project, we will use the World Happiness Reports from the year 2021, provided by the Gallup World Survey year 2021, which is a poll that continually surveys citizens in 160 countries, representing more than 98% of the world's adult population. The poll has over 100 global questions and region-specific items. Gallup also works with organizations, cities, governments, and countries to create custom items and indexes to gather information to provide a score to the "happiness" of each country.

The main purpose of our regression analysis is to predict the world Ladder.score - Happiness score (Dependent variable) based on different variables like GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity (independent variables). We will also determine which variables have more influence on the response variable. Lastly, we hope to provide a distinct perspective to our readers, as often, we define happiness with material things, and we do not measure it with other variables that are important in our lives, like peace, freedom of choices, generosity, etc.

Approach for Regression analysis

We will start with data exploration by viewing the basic statistics of the variables , different data types, and review if there is need for data preprocessing. We then will proceed to visualize the data, plotting the different variables and see the correlation with a scatterplot matrix.

We will develop our regression models based on the most relevant features for our dependent variable. Run different models to pick the most accurate one and improve the model. Finally, we will show our prediction based on the selected best fitted model. # Library, packages and data

```r
library(ggplot2)
library(tidyverse)
library(GGally)
library(corrplot)
library(compareDF)

hr_2021 <- read.csv("2021.csv", stringsAsFactors = TRUE) #World Happiness Report 2021
```

Data Preparation

Data Source

Dataset "World Happiness Report 2021", was obtained on Kaggle, the link is attached below:

https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2020.csv

The data set "The World Happiness Report", is a publication of the Network, powered by the Gallup World Poll data.

Dataset

The World Happiness Report 2021 contains 149 observations and 20. It contains 18 numerical variables like "Ladder.score"(happiness score), "Logged.GDP.per.capita", "Generosity" among other scores. Also contains 2 nominal variables: "Country" and "Regional.indicator"(Most useful for our analysis).

Top 3 countries: Finland (1) | Denmark (2) | Switzerland (2)

Bottom 3 countries: Rwanda (147) | Zimbabwe (148) | Afghanistan (149)

Rank of the Unites States: United States (19)

```
#Top 3 countries
head(hr_2021, n = 3)
```

```
##   Country.name Regional.indicator Ladder.score Standard.error.of.ladder.score
## 1      Finland     Western Europe        7.842                          0.032
## 2      Denmark     Western Europe        7.620                          0.035
## 3  Switzerland     Western Europe        7.571                          0.036
##   upperwhisker lowerwhisker Logged.GDP.per.capita Social.support
## 1        7.904        7.780                10.775          0.954
## 2        7.687        7.552                10.933          0.954
## 3        7.643        7.500                11.117          0.942
##   Healthy.life.expectancy Freedom.to.make.life.choices Generosity
## 1                    72.0                        0.949     -0.098
## 2                    72.7                        0.946      0.030
## 3                    74.4                        0.919      0.025
##   Perceptions.of.corruption Ladder.score.in.Dystopia
## 1                     0.186                     2.43
## 2                     0.179                     2.43
## 3                     0.292                     2.43
##   Explained.by..Log.GDP.per.capita Explained.by..Social.support
## 1                            1.446                        1.106
## 2                            1.502                        1.108
## 3                            1.566                        1.079
##   Explained.by..Healthy.life.expectancy
## 1                                 0.741
## 2                                 0.763
## 3                                 0.816
##   Explained.by..Freedom.to.make.life.choices Explained.by..Generosity
## 1                                      0.691                    0.124
## 2                                      0.686                    0.208
## 3                                      0.653                    0.204
##   Explained.by..Perceptions.of.corruption Dystopia...residual
```

```
## 1                                   0.481              3.253
## 2                                   0.485              2.868
## 3                                   0.413              2.839
```

```r
#Bottom 3 countries
tail(hr_2021, n = 3)
```

```
##     Country.name Regional.indicator Ladder.score Standard.error.of.ladder.score
## 147       Rwanda Sub-Saharan Africa        3.415                          0.068
## 148     Zimbabwe Sub-Saharan Africa        3.145                          0.058
## 149  Afghanistan         South Asia        2.523                          0.038
##     upperwhisker lowerwhisker Logged.GDP.per.capita Social.support
## 147        3.548        3.282                 7.676          0.552
## 148        3.259        3.030                 7.943          0.750
## 149        2.596        2.449                 7.695          0.463
##     Healthy.life.expectancy Freedom.to.make.life.choices Generosity
## 147                  61.400                        0.897      0.061
## 148                  56.201                        0.677     -0.047
## 149                  52.493                        0.382     -0.102
##     Perceptions.of.corruption Ladder.score.in.Dystopia
## 147                     0.167                     2.43
## 148                     0.821                     2.43
## 149                     0.924                     2.43
##     Explained.by..Log.GDP.per.capita Explained.by..Social.support
## 147                            0.364                        0.202
## 148                            0.457                        0.649
## 149                            0.370                        0.000
##     Explained.by..Healthy.life.expectancy
## 147                                 0.407
## 148                                 0.243
## 149                                 0.126
##     Explained.by..Freedom.to.make.life.choices Explained.by..Generosity
## 147                                      0.627                    0.227
## 148                                      0.359                    0.157
## 149                                      0.000                    0.122
##     Explained.by..Perceptions.of.corruption Dystopia...residual
## 147                                   0.493               1.095
## 148                                   0.075               1.205
## 149                                   0.010               1.895
```

```r
#United States Rank
hr_2021[which(hr_2021$Country.name == "United States"), ]
```

```
##     Country.name    Regional.indicator Ladder.score
## 19 United States North America and ANZ        6.951
##    Standard.error.of.ladder.score upperwhisker lowerwhisker
## 19                          0.049        7.047        6.856
##    Logged.GDP.per.capita Social.support Healthy.life.expectancy
## 19                11.023           0.92                    68.2
##    Freedom.to.make.life.choices Generosity Perceptions.of.corruption
## 19                        0.837      0.098                     0.698
##    Ladder.score.in.Dystopia Explained.by..Log.GDP.per.capita
## 19                     2.43                            1.533
```

```
##    Explained.by..Social.support Explained.by..Healthy.life.expectancy
## 19                         1.03                                 0.621
##    Explained.by..Freedom.to.make.life.choices Explained.by..Generosity
## 19                                      0.554                    0.252
##    Explained.by..Perceptions.of.corruption Dystopia...residual
## 19                                   0.154               2.807
```

Data Variables

Nomival variables

We can see that "Regional.indicator" is not distributed in a proportional way, since there are regions in the world with more countries than others.

```
table(hr_2021$Regional.indicator)
```

```
##
##         Central and Eastern Europe Commonwealth of Independent States
##                                 17                                 12
##                          East Asia        Latin America and Caribbean
##                                  6                                 20
##       Middle East and North Africa              North America and ANZ
##                                 17                                  4
##                         South Asia                     Southeast Asia
##                                  7                                  9
##                 Sub-Saharan Africa                     Western Europe
##                                 36                                 21
```

Numerical variables

Our dependent variable (Ladder.score) shows that the scores range from 2 - 7 approx., in a scale from 0 - 10, with a median of 5 and a mean also of 5, showing how most of the country's "happiness" is. The summary of the data set also allows us to see that some variables are whole numbers of scores and some are percentages such as "perceptions.of.corruptions" and "Generosity".

```
summary(hr_2021)
```

```
##       Country.name                          Regional.indicator  Ladder.score
##  Afghanistan:  1   Sub-Saharan Africa                :36       Min.   :2.523
##  Albania    :  1   Western Europe                    :21       1st Qu.:4.852
##  Algeria    :  1   Latin America and Caribbean       :20       Median :5.534
##  Argentina  :  1   Central and Eastern Europe        :17       Mean   :5.533
##  Armenia    :  1   Middle East and North Africa      :17       3rd Qu.:6.255
##  Australia  :  1   Commonwealth of Independent States:12       Max.   :7.842
##  (Other)    :143   (Other)                           :26
##  Standard.error.of.ladder.score  upperwhisker    lowerwhisker
##  Min.   :0.02600                Min.   :2.596   Min.   :2.449
##  1st Qu.:0.04300                1st Qu.:4.991   1st Qu.:4.706
##  Median :0.05400                Median :5.625   Median :5.413
##  Mean   :0.05875                Mean   :5.648   Mean   :5.418
```

```
##  3rd Qu.:0.07000                3rd Qu.:6.344   3rd Qu.:6.128
##  Max.   :0.17300                Max.   :7.904   Max.   :7.780
##
##  Logged.GDP.per.capita Social.support   Healthy.life.expectancy
##  Min.   : 6.635        Min.   :0.4630   Min.   :48.48
##  1st Qu.: 8.541        1st Qu.:0.7500   1st Qu.:59.80
##  Median : 9.569        Median :0.8320   Median :66.60
##  Mean   : 9.432        Mean   :0.8147   Mean   :64.99
##  3rd Qu.:10.421        3rd Qu.:0.9050   3rd Qu.:69.60
##  Max.   :11.647        Max.   :0.9830   Max.   :76.95
##
##  Freedom.to.make.life.choices   Generosity       Perceptions.of.corruption
##  Min.   :0.3820               Min.   :-0.28800   Min.   :0.0820
##  1st Qu.:0.7180               1st Qu.:-0.12600   1st Qu.:0.6670
##  Median :0.8040               Median :-0.03600   Median :0.7810
##  Mean   :0.7916               Mean   :-0.01513   Mean   :0.7274
##  3rd Qu.:0.8770               3rd Qu.: 0.07900   3rd Qu.:0.8450
##  Max.   :0.9700               Max.   : 0.54200   Max.   :0.9390
##
##  Ladder.score.in.Dystopia Explained.by..Log.GDP.per.capita
##  Min.   :2.43             Min.   :0.0000
##  1st Qu.:2.43             1st Qu.:0.6660
##  Median :2.43             Median :1.0250
##  Mean   :2.43             Mean   :0.9772
##  3rd Qu.:2.43             3rd Qu.:1.3230
##  Max.   :2.43             Max.   :1.7510
##
##  Explained.by..Social.support Explained.by..Healthy.life.expectancy
##  Min.   :0.0000               Min.   :0.0000
##  1st Qu.:0.6470               1st Qu.:0.3570
##  Median :0.8320               Median :0.5710
##  Mean   :0.7933               Mean   :0.5202
##  3rd Qu.:0.9960               3rd Qu.:0.6650
##  Max.   :1.1720               Max.   :0.8970
##
##  Explained.by..Freedom.to.make.life.choices Explained.by..Generosity
##  Min.   :0.0000                             Min.   :0.000
##  1st Qu.:0.4090                             1st Qu.:0.105
##  Median :0.5140                             Median :0.164
##  Mean   :0.4987                             Mean   :0.178
##  3rd Qu.:0.6030                             3rd Qu.:0.239
##  Max.   :0.7160                             Max.   :0.541
##
##  Explained.by..Perceptions.of.corruption Dystopia...residual
##  Min.   :0.0000                          Min.   :0.648
##  1st Qu.:0.0600                          1st Qu.:2.138
##  Median :0.1010                          Median :2.509
##  Mean   :0.1351                          Mean   :2.430
##  3rd Qu.:0.1740                          3rd Qu.:2.794
##  Max.   :0.5470                          Max.   :3.482
##
```

Data Cleaning

We have no missing values in our data set, and with "str" function, we can check if there are any inconsistent factors, which we do not have in our data set. We also were able to see some variables have data that's repetitive so we will make them "NULL", data ready for exploratory analysis.

```
sum(is.na(hr_2021))
```

```
## [1] 0
```

```
str(hr_2021)
```

```
## 'data.frame':    149 obs. of  20 variables:
##  $ Country.name                              : Factor w/ 149 levels "Afghanistan",..: 41 34 129 55 97
##  $ Regional.indicator                        : Factor w/ 10 levels "Central and Eastern Europe",..: 
##  $ Ladder.score                              : num  7.84 7.62 7.57 7.55 7.46 ...
##  $ Standard.error.of.ladder.score            : num  0.032 0.035 0.036 0.059 0.027 0.035 0.036 0.037 0
##  $ upperwhisker                              : num  7.9 7.69 7.64 7.67 7.52 ...
##  $ lowerwhisker                              : num  7.78 7.55 7.5 7.44 7.41 ...
##  $ Logged.GDP.per.capita                     : num  10.8 10.9 11.1 10.9 10.9 ...
##  $ Social.support                            : num  0.954 0.954 0.942 0.983 0.942 0.954 0.934 0.908 0
##  $ Healthy.life.expectancy                   : num  72 72.7 74.4 73 72.4 73.3 72.7 72.6 73.4 73.3 ..
##  $ Freedom.to.make.life.choices              : num  0.949 0.946 0.919 0.955 0.913 0.96 0.945 0.907 0
##  $ Generosity                                : num  -0.098 0.03 0.025 0.16 0.175 0.093 0.086 -0.034 0
##  $ Perceptions.of.corruption                 : num  0.186 0.179 0.292 0.673 0.338 0.27 0.237 0.386 0
##  $ Ladder.score.in.Dystopia                  : num  2.43 2.43 2.43 2.43 2.43 2.43 2.43 2.43 2.43 2.43
##  $ Explained.by..Log.GDP.per.capita          : num  1.45 1.5 1.57 1.48 1.5 ...
##  $ Explained.by..Social.support              : num  1.11 1.11 1.08 1.17 1.08 ...
##  $ Explained.by..Healthy.life.expectancy     : num  0.741 0.763 0.816 0.772 0.753 0.782 0.763 0.76 0
##  $ Explained.by..Freedom.to.make.life.choices: num  0.691 0.686 0.653 0.698 0.647 0.703 0.685 0.639 0
##  $ Explained.by..Generosity                  : num  0.124 0.208 0.204 0.293 0.302 0.249 0.244 0.166 0
##  $ Explained.by..Perceptions.of.corruption   : num  0.481 0.485 0.413 0.17 0.384 0.427 0.448 0.353 0
##  $ Dystopia...residual                       : num  3.25 2.87 2.84 2.97 2.8 ...
```

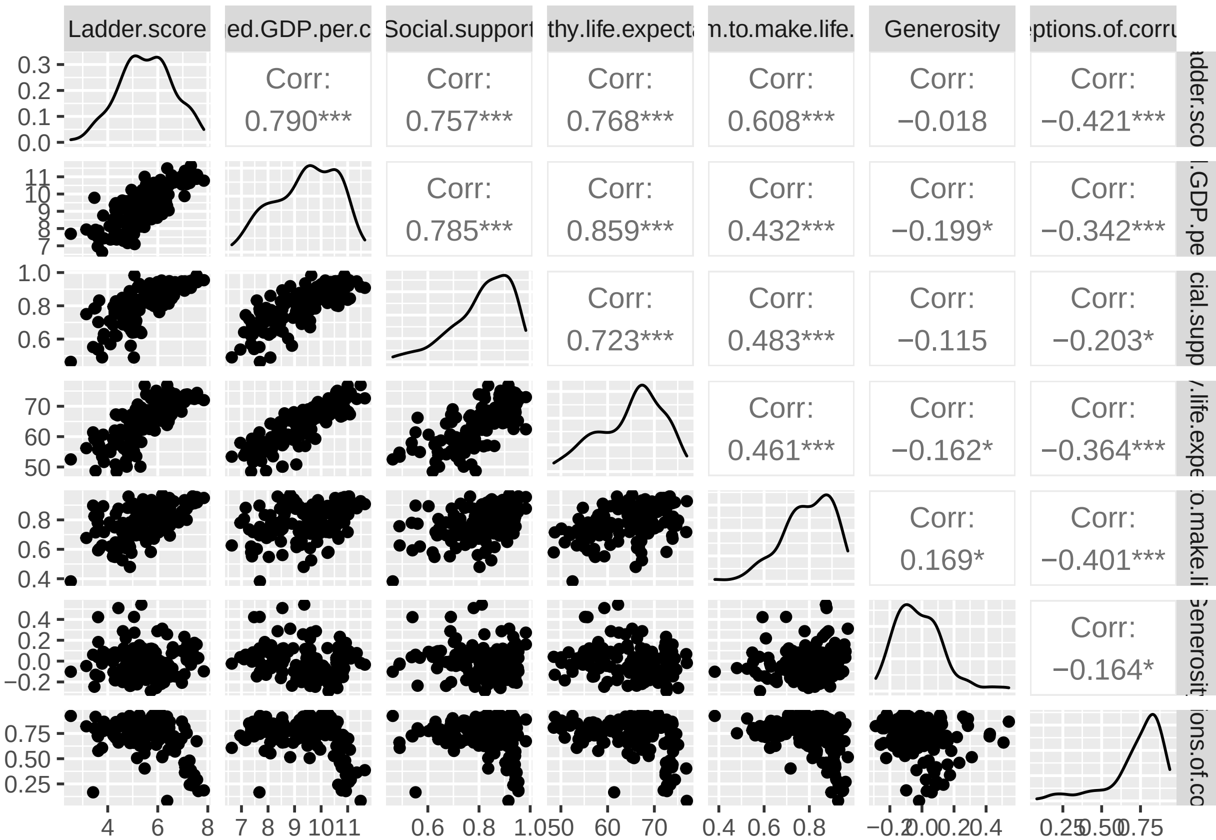
```
# Make NULL variables with repetitive data
hr_2021[13:20] <- NULL
hr_2021[4:6] <-NULL
```

Exploratory Data analysis

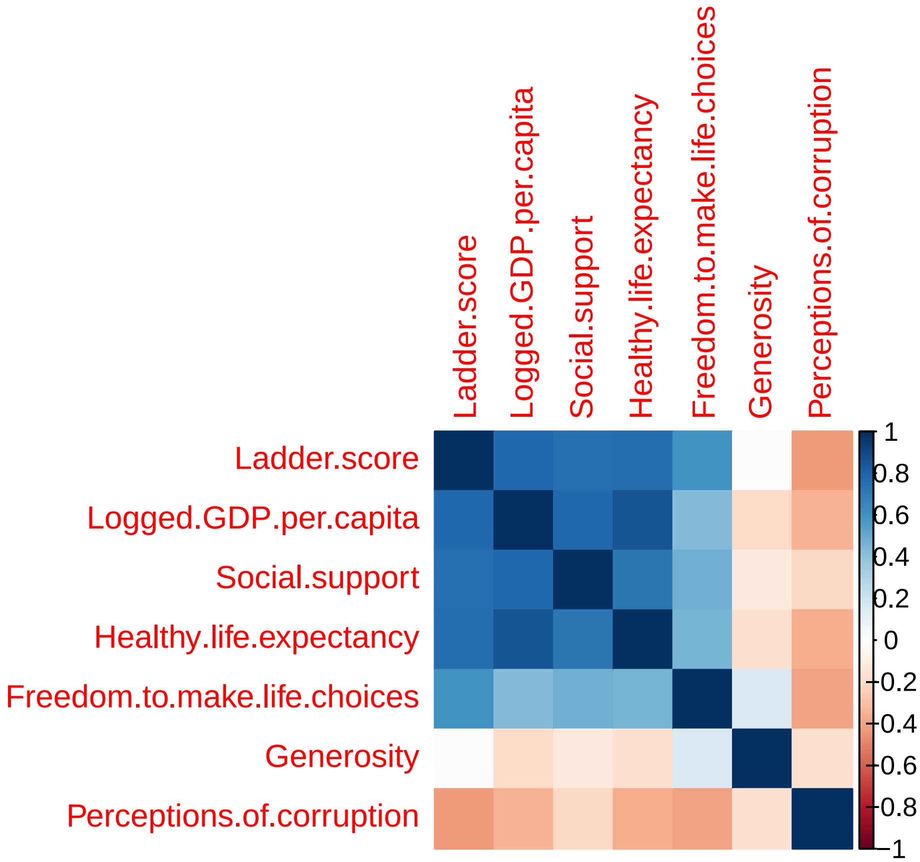
Correlation

Scatter-plot matrix

```
#Scatter-plot matrix
ggpairs(hr_2021[3:9])
```




```
#heat map
#Positive correlations are displayed in blue and negative correlations in red.
#Color intensity and the size of the circle are proportional to the correlation coefficients.
corrplot(cor(hr_2021[3:9]), method = "color")
```



With "ggpairs" from the libraries "ggplot2" and "GGally", we visualize
the different correlations between variables in a scatter plot matrix, better than the usual matrix that we obtain with "plot", as it provides one score, and proper
visualization of the distribution of data in each of the variables. We can see that the variables about GDP, social support, and healthy life expectancy, have the highest correlation with our dependent variables "Ladder.score". Meaning that has these variables increase, so does the ladder score. We can still see correlation between the variables, freedom to make life choices, generosity, and perception of corruption, but they are not as strong as the other three variables mentioned before. Our heat map with "corrplot" function, from the package "corrplot", supports our previous observations, as positive correlations are displayed in blue and negative correlations in red. Th color intensity and the size of the are proportional to the correlation coefficients.

Regression Models

Our goal with our regression analysis, will be to predict Ladder.score . We will perform different models, using singular, multiple and all the variables.

Our exploratory data analysis showed us the variables that are more correlated with Ladder.score.

Model 1 will be a simple linear regression will be a simple linear regression L adder.score ~ Logged.GDP.per.capita.

Model 2 will contain all the variables that showed more correlation with Ladder.score, which are: Logged.GDP.per.capita, Social.support, Healthy.life.expectancy and Freedom.to.make.life.choices.

Model 3 will have all the variables in the dataset. Including interaction between the nominal variable of region.

We will use the summary() function to obtain an analysis of each model, which will be of use when selecting the best one.

```r
#Model 1
model_1 <- lm(Ladder.score  ~ Logged.GDP.per.capita , data = hr_2021)
##Model 1 summary
summary(model_1)
```

```
##
## Call:
## lm(formula = Ladder.score ~ Logged.GDP.per.capita, data = hr_2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32190 -0.46198  0.08206  0.50740  1.32618
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.3719     0.4456  -3.079  0.00248 **
## Logged.GDP.per.capita   0.7320     0.0469  15.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.661 on 147 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6212
## F-statistic: 243.7 on 1 and 147 DF,  p-value: < 2.2e-16
```

```r
#Model 2
model_2 <- lm(Ladder.score ~ Logged.GDP.per.capita + Social.support +
                Healthy.life.expectancy + Freedom.to.make.life.choices, data = hr_2021)
##Model 2 summary
summary(model_2)
```

```
##
## Call:
## lm(formula = Ladder.score ~ Logged.GDP.per.capita + Social.support +
##     Healthy.life.expectancy + Freedom.to.make.life.choices, data = hr_2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99295 -0.32898  0.07857  0.38442  1.04540
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3.11157    0.45703  -6.808 2.48e-10 ***
## Logged.GDP.per.capita         0.29182    0.08632   3.381 0.000931 ***
## Social.support                2.16586    0.66232   3.270 0.001345 **
## Healthy.life.expectancy       0.03308    0.01343   2.463 0.014954 *
## Freedom.to.make.life.choices  2.49817    0.46452   5.378 2.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5506 on 144 degrees of freedom
```

```
## Multiple R-squared:  0.7442, Adjusted R-squared:  0.7371
## F-statistic: 104.7 on 4 and 144 DF,  p-value: < 2.2e-16
```

```
#Model 3
#Data used for model 3, since we dont want name of countries on the linear regression
model_3_data <- hr_2021[3:9]
model_3 <- lm(Ladder.score ~ ., data = model_3_data)
##Model 3 summary
summary(model_3)
```

```
##
## Call:
## lm(formula = Ladder.score ~ ., data = model_3_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85049 -0.30026  0.05735  0.33368  1.04878
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.23722    0.63049  -3.548 0.000526 ***
## Logged.GDP.per.capita          0.27953    0.08684   3.219 0.001595 **
## Social.support                 2.47621    0.66822   3.706 0.000301 ***
## Healthy.life.expectancy        0.03031    0.01333   2.274 0.024494 *
## Freedom.to.make.life.choices  2.01046    0.49480   4.063 7.98e-05 ***
## Generosity                     0.36438    0.32121   1.134 0.258541
## Perceptions.of.corruption     -0.60509    0.29051  -2.083 0.039058 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5417 on 142 degrees of freedom
## Multiple R-squared:  0.7558, Adjusted R-squared:  0.7455
## F-statistic: 73.27 on 6 and 142 DF,  p-value: < 2.2e-16
```

We will select model 2, since it explains variance of 74.42% compared to model 1 which explains 62.37%, and compared to model 3, which explains variance of 75.58%, that is higher than our model 2, howevere model 2 has only 4 variables while model 3 has 7.

```
summary(model_2)
```

```
##
## Call:
## lm(formula = Ladder.score ~ Logged.GDP.per.capita + Social.support +
##     Healthy.life.expectancy + Freedom.to.make.life.choices, data = hr_2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99295 -0.32898  0.07857  0.38442  1.04540
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3.11157    0.45703  -6.808 2.48e-10 ***
## Logged.GDP.per.capita          0.29182    0.08632   3.381 0.000931 ***
```

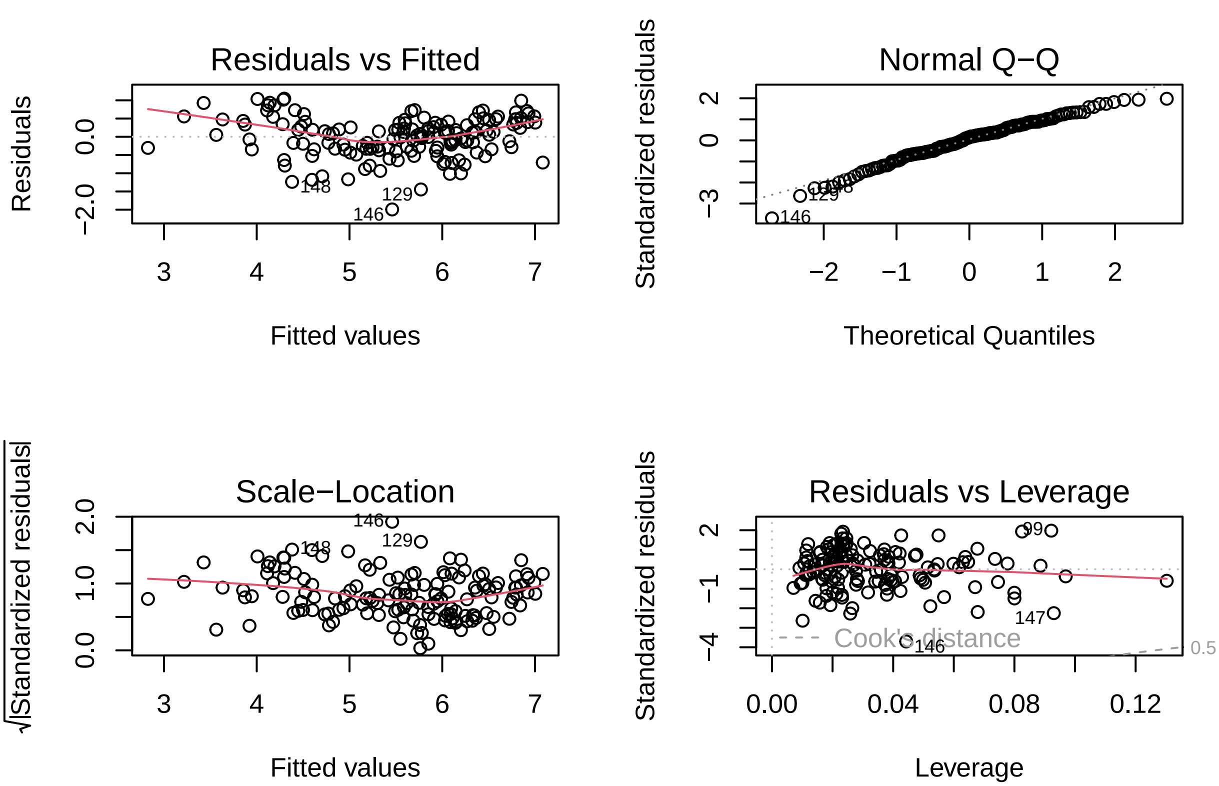
```
## Social.support                 2.16586    0.66232   3.270 0.001345 **
## Healthy.life.expectancy        0.03308    0.01343   2.463 0.014954 *
## Freedom.to.make.life.choices   2.49817    0.46452   5.378 2.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5506 on 144 degrees of freedom
## Multiple R-squared:  0.7442, Adjusted R-squared:  0.7371
## F-statistic: 104.7 on 4 and 144 DF,  p-value: < 2.2e-16
```

Summary of model 2 provide us with insights about the performance of the model, such intercept and the coefficients of the independent variables and the dependent variable. A unit increase in Logged.GDP.per.capita will increase the score by 0.292 in Ladder.score. Summary of model 2 also shows us that all the 4 variables are statistically significant, with values lower that 0.05. In addition R square of 74.42% indicates that the variation is explained by the model is good. For now, we will move to try to improve the results of the summary of model 2.

Diagnostic and transformations

Regression diagnostic plots

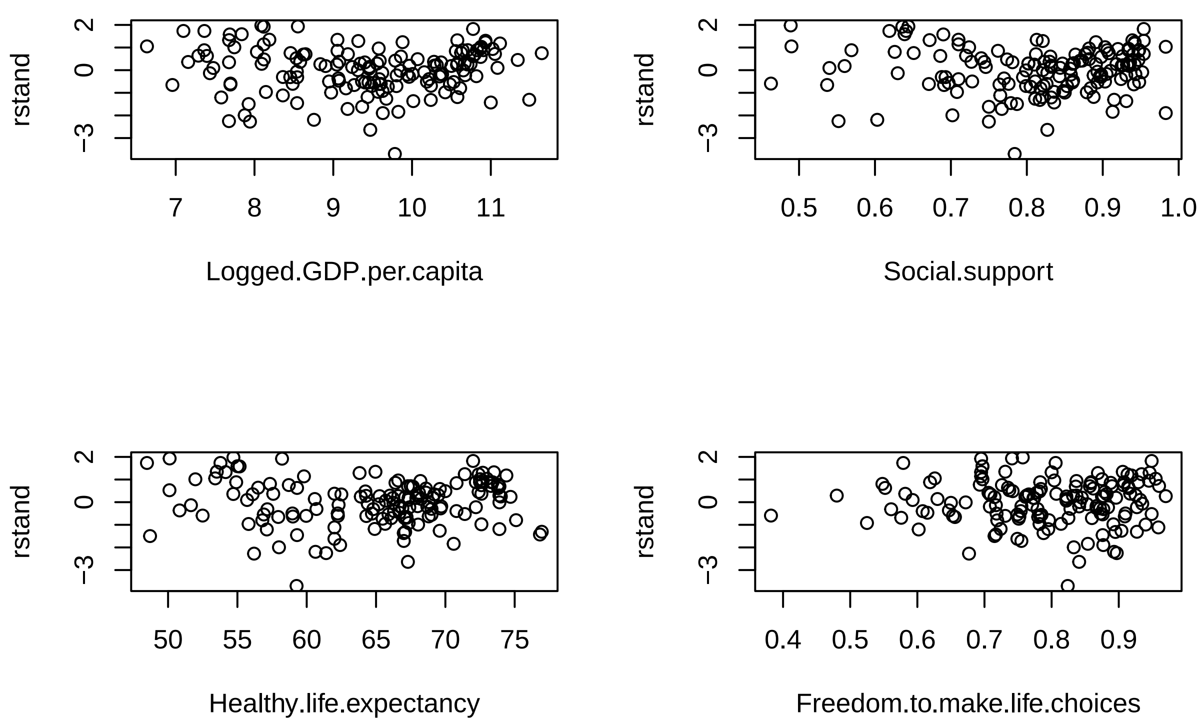
```
#Regression diagnostics on model 2
par(mfrow = c(2,2))
plot(model_2)
```



- Residual vs Fitted: Doesn't show a great horizontal line, looks more like a U shape. At least there is not a distinct pattern which is an indicator for relationship.
- Normal Q-Q: Good residual distribution as the points follow the straight dashed line. just a few outliers on the lower left side, and the upper right side.
- Scale-Location (or Spread-Location): Shows some variance in our residuals, as we don't have a horizontal line, but some inclination.
- Residuals vs Leverage: Shows there are few extreme values that would affect our regression, However, nothing too severe.

Standarized residuals against each predictor.

```
par(mfrow = c(2,2))
rstand <- rstandard(model_2)
plot(rstand ~ Logged.GDP.per.capita, data = hr_2021)
plot(rstand ~ Social.support, data = hr_2021)
plot(rstand ~ Healthy.life.expectancy, data = hr_2021)
plot(rstand ~ Freedom.to.make.life.choices, data = hr_2021)
```



The standardized residuals plot shows Homoscedasticity, among the variables, since there are no cluster of data points. This plot also strength, our initial visualization of the correlation of the data, where we could see that most of the data showed normal distribution, good histogram. Implying that even transformations such as applying

log() on our linear regression mode, will not have a significant change in the results
previous results of our model.

Transformation

Developing model based on Log transformations

```r
#log transformation of model 1
log_model_1 <- lm(log(Ladder.score) ~ Logged.GDP.per.capita + Social.support +
    Healthy.life.expectancy + Freedom.to.make.life.choices, data = hr_2021)

summary(log_model_1)$r.sq
```

```
## [1] 0.7289735
```

```r
#log transformation on model 2
log_model_2 <- lm(log(Ladder.score) ~ Logged.GDP.per.capita + Social.support *
                    Freedom.to.make.life.choices + Healthy.life.expectancy, data = hr_2021)

summary(log_model_2)$r.sq
```

```
## [1] 0.7317535
```

As expected, since the data on the World Happiness Report is on a good part, very good distributed, utilizing log() on our dependent variable, did not improve the results of our previous model 2, even after adding interaction between two of the variables that have a significant impact on our dependent variable (GDP and Social Support).

Conclusions

It seems out model 2 is the best fit, the summary also supports this, since the model has p-value less than 0.05, all the variables are statistically significant, low residual to standard error ratio, and well distributed data, as well, both previous log() on lm() didn't surpass the 74.42% obtained previously.

We can see which variables are related the most to the "ladder.score (happiness score)" and it is remarkably interesting to see that happiness cannot be measured just by money (GDP per capita), or only for freedom. However, our model shows us that freedom of choice and life expectancy, has a significant impact on the happiness score. Of course, a combination on high scores on the other variables is important, but these two are the ones with the biggest impact on the score.

Predictions based on selected model

Our fitted regression model is the next: Ladder.Score = -3.11157 + 0.29182(Logged.GDP.per.capita) − 2.16586(Social.support) + 0.03308(Healthy.life.expectancy) + 2.49817(Freedom.to.make.life.choices)

Prediction case 1 - Balanced GDP - Freedom and life expectancy Country with scores of: GDP = 9.750, Social support = 0.854, healthy life expectancy = 70.000, freedom to make life choices = 0.9000

```r
#define new country
new <- data.frame(Logged.GDP.per.capita=c(9.775), Social.support=c(0.854),
                  Healthy.life.expectancy=c(70.000), Freedom.to.make.life.choices = 0.900)

#use the fitted model to predict the rating for the new player
predict(model_2, newdata=new)
```

```
##        1
## 6.154323
```

Prediction case 2 - High GDP and low life expectancy and freedom Country with scores of: GDP = 6.750, Social support = 0.8000, healthy life expectancy = 65, freedom to make life choices = 0.8000

```r
#define new country
new <- data.frame(Logged.GDP.per.capita=c(11.500), Social.support=c(0.8000),
                  Healthy.life.expectancy=c(50), Freedom.to.make.life.choices = 0.4000)

#use the fitted model to predict the rating for the new player
predict(model_2, newdata=new)
```

```
##        1
## 4.630125
```

Prediction case 3 - Low High GDP and high life expectancy and freedom Country with scores of: GDP = 6.750, Social support = 0.8000, healthy life expectancy = 65, freedom to make life choices = 0.8000

```r
#define new country
new <- data.frame(Logged.GDP.per.capita=c(6.635), Social.support=c(0.983),
                  Healthy.life.expectancy=c(76.953), Freedom.to.make.life.choices = 0.970)

#use the fitted model to predict the rating for the new player
predict(model_2, newdata=new)
```

```
##        1
## 5.922274
```

Limitation and future improvements

There are improvements that could be made to the analysis, the data was well distributed overall, however, for a future project, we could gather data from previous year, and that will improve our prediction power, as all models feed of the amount of data available. We could also do comparison on year that there were some catastrophes or big event for the world and see if the independent variables correlation with the happiness score changed.

Thank you