D20125700

Luis Martinez

TU256

Option A

ONE ANALYSIS OF HAPPINESS WORLDWIDE

INTRODUCTION

The ultimate goal of the project is to find out the key factors that make people happy nationwide.

The data considered in this work have been collected from Kaggle website[1], corresponding to the World Happiness Report, consisting of historical surveys of the corresponding to the years 2015, 2016 and 2017 in more than 100 countries a world level. The top six factors that cost in each of the datasets and that serve to value happiness in each country are: production economic, life expectancy, social support, freedom, absence of corruption and generosity.

The Happiness Report of the World ranks more than 100 countries according to their level of happiness, it was launched in the United Nations in an event celebrating the International Day of Happiness on March 20. So, this work is done with three of those datasets and corresponding to the years 2015, 2016 and 2017; Data from 2018 and 2019 were not used in the project, since there are few country records and less variables, which cannot be compared with the number of existing records in the years 2015, 2016 and 2017 regarding the happiness score which will be analysed.

MOTIVATION

[1]Kaggle website: https://www.kaggle.com/unsdsn/world-happiness?select=2018.csv

My personal motivation for carrying out this work is because of what I have learned so far in the master's degree I like the study of predictions more in which it works mainly with contents of the subjects of Statistics and R language.

I selected the topic of happiness because I would like to determine the importance of the key factors that play a role in this feeling of self-realization and compromise of our wishes and aspirations. personally, I consider that happiness is quite an abstract concept and perhaps not so easy to measure.

Happiness is a state and at the same time, a dynamic process, which is generated by the interaction of a large number of conditions or variables that act on the individual eliciting terminal responses of a positive nature. These variables can be grouped in different ways: Biological (gender, health, malformations), psychological (personality traits, self-esteem, values, beliefs, affections) and sociocultural (marriage, income, family, marginalization, etc.).

In any case, I believe that my experience living in other countries and doing a constant fieldwork experience observing the character, personality and type of social relations between the specific population of those places in which I have lived, has contributed to my interest in this topic and carry out research using this database, which is at the national level (not individual) and has a more general character for the population, being quite interesting likewise.

The World Happiness Report is a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be.[2]

The research will analyse the importance of each of the factors that intervene in happiness, as well as determine the relationship between these, identify, among the variables, which or which are the best predictors of the happiness.

The sample consisted of 156 countries around the world. There are fundamentally 6 factors that influence happiness. Then, each of them is detailed:

[2] https://worldhappiness.report/

a) **Gross domestic product per capita**: It is the final sum of quantities of goods and services that are produced in a country, at the monetary value of a country reference.

b) **Social support**: It is the help from family or friends in case of having problems.

c) **Life expectancy**: It is an index that determines how much expects a person to live in a particular social context.

d) **The freedom to make decisions**: It is the faculty that the person has to decision making, that is; who is directly responsible for their acts.

e) **Generosity**: It is the virtue of giving and sharing over one's own interest or utility.

f) **The perception of corruption**: It is the concept that citizens have with respect to the government and/or companies.

Other studies that have investigating the happiness at global level are World Happiness REPORT Edited by John Helliwell, Richard Layard and Jeffrey Sachs [3,] with same name but with a different perspective, The report emphasizes that money is an important factor in the life of a human being, but only with this happiness is not achieved. That the land must be protected to achieve a good quality of life, which will be adopted better lifestyles and technologies that help improve happiness. The search for happiness is closely linked to the search for sustainable development. Important external factors are: income, work, community and government, values and religion. Personal factors include: physical and mental health, family experience, education, gender and age. Many of these factors have a two-way interaction with happiness.

There are many other variables that have a greater effect on happiness, which are: confidence social, the quality of work and freedom of choice and political participation.

Other interesting work that deal with this topic is DATA1001 Project #1 Chandler Elissa, Nikolovski Mihael, Villar Miguel, Ando Koki [4], in this report made by experts from

[3] World Happiness REPORT: https://www.researchgate.net/publication/233401584_World_Happiness_Report
[4] DATA1001 Project #1: https://rpubs.com/koki25ando/DATA1001TeamBver1

different fields of psychology, health, economics, etc. Happiness is analysed from the same approach as the dataset using on this investigation, considering the the six key conditions in each country:

1. economic prosperity, including decent work for all who want it;

2. the physical and mental health of citizens;

3. the freedom of individuals to make key decisions in life;

4. strong and dynamic social support networks (social capital);

5. shared public values of generosity;

6. social trust, including trust in the honesty of business and government.

The sum of the conditions determines the global score for each country.

DATA

As mention before, for this work have been selected three datasets corresponding to the years 2015, 2016 and 2017.

2015: Composed of 158 rows and 12 columns. The detail of the columns is:

Number	Name	description
1	Country	Name of the country
2	Region	Region to which the country belongs
3	Happiness Rank	Country rating based on happiness score
4	Happiness Score	Metric measured in 2015 by formulating the ask the people included in the sample: "How would you rate your happiness on a scale from 0 to 10, where 10 is the happiest? "
5	Standard Error	The standard error of the happiness score
6	Economy (GDP per Capita)	The extent to which GDP contributes to the calculation of the happiness score
7	Family	

		The extent to which the family contributes to the calculation of happiness score
8	Health (Life Expectancy)	The extent to which life expectancy contributed to happiness score calculation
9	Freedom	The extent to which freedom contributed to the calculation of happiness score
10	Trust (Government Corruption)	The extent to which the perception of corruption contributes to the happiness score
11	Generosity	The extent to which generosity contributed to the calculation of the happiness score
12	Dystopia Residual	The extent to which Dystopia Residual contributed to happiness score calculation

2016: Composed of 157 rows and 13 columns. The detail of the columns is:

Number	Name	description
1	Country	Name of the country
2	Region	Region to which the country belongs
3	Happiness Rank	Country rating based on happiness score
4	Happiness Score	Metric measured in 2016 by formulating the ask the people included in the sample: "How would you rate your happiness on a scale from 0 to 10, where 10 is the happiest? "
5	Lower Confidence Interval	Lower confidence interval of the score happiness
6	Standard Error	The standard error of the happiness score
7	Economy (GDP per Capita)	The extent to which GDP contributes to the calculation of the happiness score
8	Family	The extent to which the family contributes to the calculation of happiness score
9	Health (Life Expectancy)	The extent to which life expectancy contributed to happiness score calculation
10	Freedom	The extent to which freedom contributed to the calculation of

		happiness score
11	Trust (Government Corruption)	The extent to which the perception of corruption contributes to the happiness score
12	Generosity	The extent to which generosity contributed to the calculation of the happiness score
13	Dystopia Residual	The extent to which Dystopia Residual contributed to happiness score calculation

2017: Composed of 155 rows and 12 columns. The detail of the columns is:

Number	Name	description
1	Country	Name of the country
2	Happiness Rank	Country rating based on happiness score
3	Happiness Score	Metric measured in 2017 by formulating the ask the people included in the sample: "How would you rate your happiness on a scale from 0 to 10, where 10 is the happiest? "
4	Whisker.high	High margin
5	Whisker.low	Low margin
6	Economy (GDP per Capita)	The extent to which GDP contributes to the calculation of the happiness score
7	Family	The extent to which the family contributes to the calculation of happiness score
8	Health (Life Expectancy)	The extent to which life expectancy contributed to happiness score calculation
9	Freedom	The extent to which freedom contributed to the calculation of happiness score
10	Trust (Government Corruption)	The extent to which the perception of corruption contributes to the happiness score
11	Generosity	The extent to which generosity contributed to the calculation

		of the happiness score
12	Dystopia Residual	The extent to which Dystopia Residual contributed to happiness score calculation

ANALYSIS

The first thing I did was to investigate whether there are differences of happiness between the 3 years.

To answer this question, the method that I did was to use is One-way ANOVA in order to determine whether there are significant differences in the level of happiness among the three years, I follow these steps:

Firstly, the data preparation, in order to be able to compare the data for each country over the three years, I must group the information, for this consider:

The first step is to recode the name of the columns, the variables of the datasets of the three years, then I generate a data frame that contains the data of all years, consisting of three columns(Country, Hscore and Group) with the Happiness score of all the years include in it (2015, 2016 and 2017 respectively), this was done merging the variables from the all data frames.

```
In [28]: head(Countries2)
```

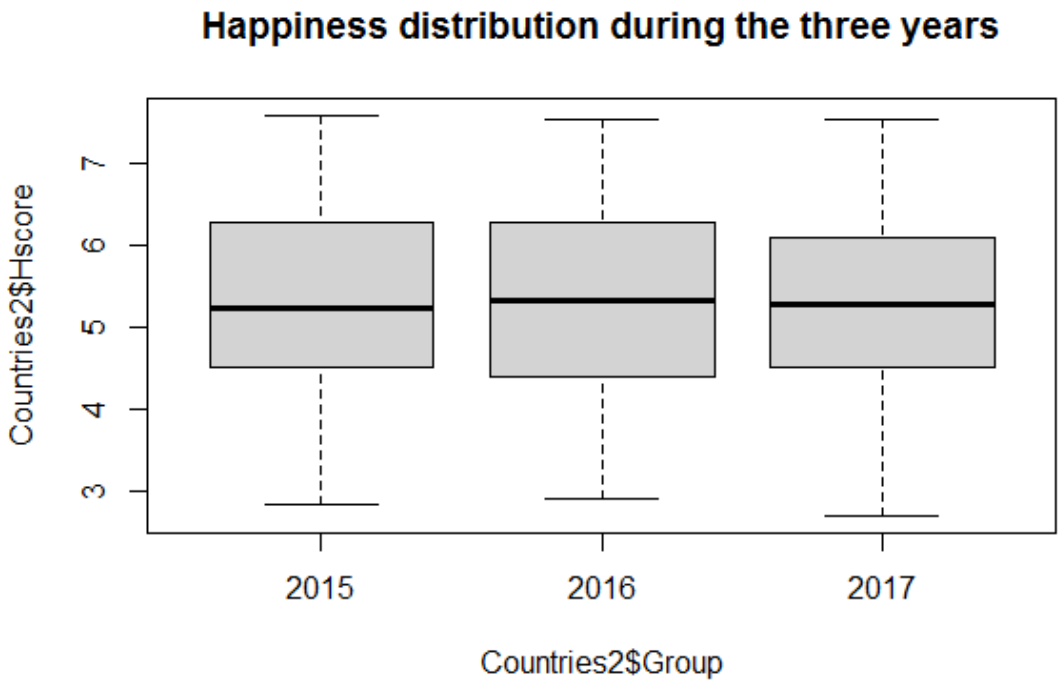
A data.frame: 6 × 3

	Country	Hscore	Group
	<chr>	<dbl>	<chr>
1	Switzerland	7.587	2015
2	Iceland	7.561	2015
3	Denmark	7.527	2015
4	Norway	7.522	2015
5	Canada	7.427	2015
6	Finland	7.406	2015

Then visualize the type of variables of the datasets, let's see for 2015:

```
|var         |class     |
|:----------|:---------|
|Country     |character |
|Region      |character |
|Hrank       |integer   |
|Hscore      |numeric   |
|Stnderror   |numeric   |
|GDP         |numeric   |
|family      |numeric   |
|lifexp      |numeric   |
|Freedom     |numeric   |
|Trust       |numeric   |
|Generosity  |numeric   |
|Resid       |numeric   |
```

The next step is to visualize the happiness scores for the three years with box plots, one of them corresponding to each year:



Graph 1. Happiness distribution during the three years.

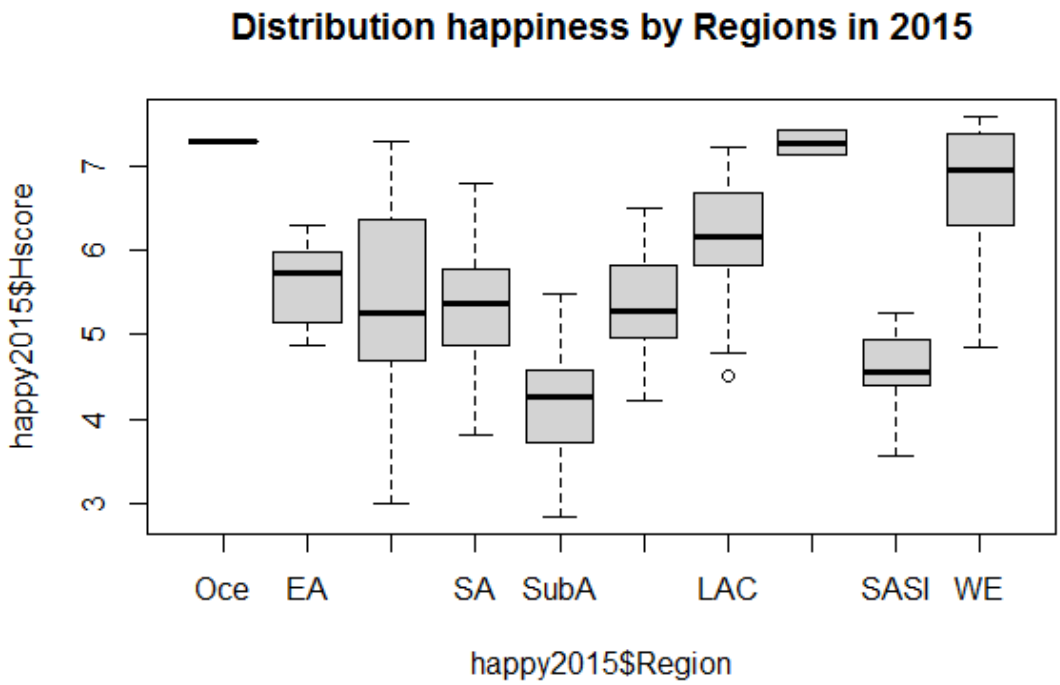
Then I investigate if there are differences in the level of happiness between the three years, throw a calculation of ANOVA.

```
Hscoreaov <- aov(Countries2$Hscore~Countries2$Group, data=Countries2)
Hscoreaov

Call:
   aov(formula = Countries2$Hscore ~ Countries2$Group, data = Countries2)

Terms:
                Countries2$Group Residuals
Sum of Squares            0.0678  606.2388
Deg. of Freedom                2       467

Residual standard error: 1.139366
Estimated effects may be unbalanced

summary (aov(Countries2$Hscore~Countries2$Group))

                  Df Sum Sq Mean Sq F value Pr(>F)
Countries2$Group   2    0.1  0.0339   0.026  0.974
Residuals        467  606.2  1.2982
```

```
numSummary(Countries2$Hscore, groups=Countries2$Group, statistics=c('mean','sd'))

          mean       sd data:n
2015 5.375734 1.145010    158
2016 5.382185 1.141674    157
2017 5.354019 1.131230    155
```

According to the data obtained, As the p-value is greater than 0.05 then it is not ruled out a hypothesis that happiness levels are similar in the last three years with 95% confidence. Although we can see that it looks to be little differences in the levels of happiness in the last three years even if the values of the means are almost similar. We could say that in 2016 there are higher level of happiness than in the other two years.

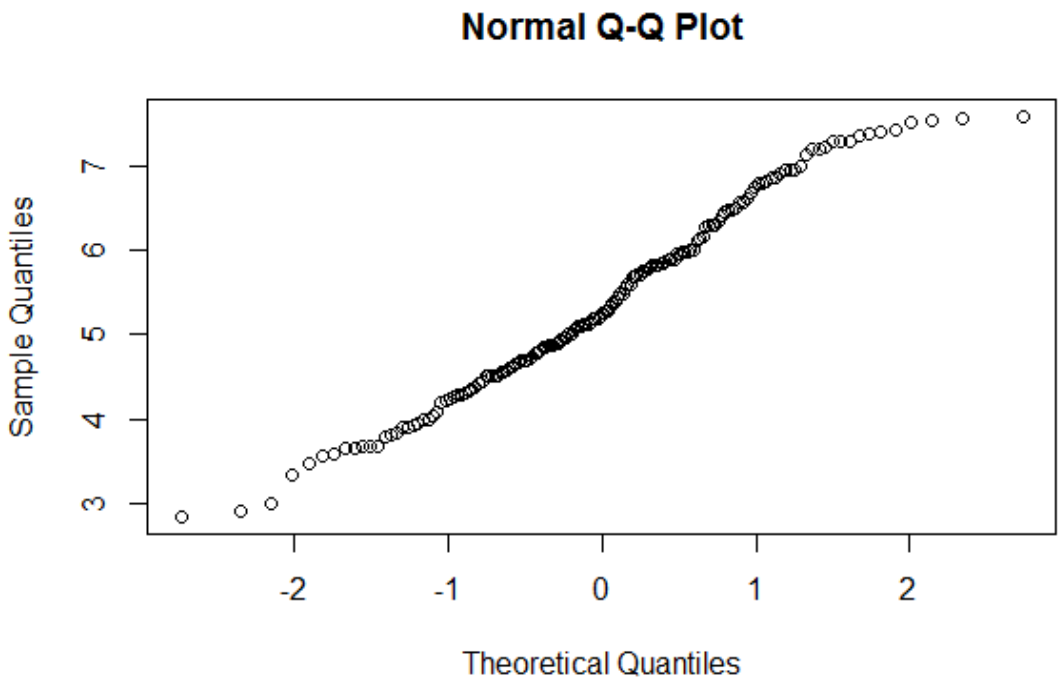
Other question in the investigation is whether are there differences in the level of happiness between the different regions. I used an ANOVA for independent samples to calculate it in the same year (2015). Firstly, I updated the names of the region in the column 'Region' to be able to work better. Let's visualize a box pot to see the distribution of the happiness around the countries for 2015:



Graph 2 Distribution happiness by Regions in 2015.

The ANOVA test requires that the sample data meet two basic assumptions: normality and equality of variances (homoscedasticity, which occurs if the error made by the model always has the same variance). It then checks whether meet these conditions. In fact, the assumptions have to be verified before apply ANOVA.

Checking the normality, it can be represented graphically with the function qqnorm using as a parameter the Hscore value (happiness score). This verification is visualized in the next graph:



Graph 3. Checking normality

Preparation of the data frame to apply ANOVA. In this case, the data frame must contain the variable "Region" as a factor, this is the independent variable, and the variable "Hscore" which is of the numeric type, is the dependent variable. Each row represents a region and a value specific Hscore for the specific region.

```
Hscore2015 <- aov(happy2015$Hscore~happy2015$Region)
Hscore2015

Call:
   aov(formula = happy2015$Hscore ~ happy2015$Region)

Terms:
                happy2015$Region Residuals
Sum of Squares         123.68339  82.15118
Deg. of Freedom                9       148

Residual standard error: 0.7450339
Estimated effects may be unbalanced

summary(Hscore2015)

                 Df Sum Sq Mean Sq F value Pr(>F)
happy2015$Region  9 123.68  13.743   24.76 <2e-16 ***
Residuals       148  82.15   0.555
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
numSummary(happy2015$Hscore, groups=happy2015$Region, statistics = c('mean','sd'))

         mean          sd data:n
Oce   7.285000 0.001414214      2
EA    5.626167 0.554052855      6
MeNa  5.406900 1.101381902     20
SA    5.317444 0.950020146      9
SubA  4.202800 0.609557099     40
CEE   5.332931 0.570445811     29
LAC   6.144682 0.728560053     22
NAM   7.273000 0.217788889      2
SASI  4.580857 0.570526490      7
WE    6.689619 0.824581802     21
```

```
model.tables(Hscore2015)

Tables of effects

 happy2015$Region
      Oce      EA     MeNa       SA   SubA     CEE     LAC   NAM    SASI     WE
    1.909 0.2504  0.03117 -0.05829 -1.173 -0.0428  0.7689 1.897 -0.7949  1.314
rep 2.000 6.0000 20.00000  9.00000 40.000 29.0000 22.0000 2.000  7.0000 21.000
```

According to the data of the analysis and the boxplots graph, the result of P-Value with ANOVA (2e- 16) and the data obtained by modelling ANOVA tables, it is evidenced that there are no relationships between each of the regions. It can be seen that between Australia and New Zealand (Oce) and North America (NAM) there is a slight relationship in their variances, but the P value is very small so we can conclude the non-relation of Hscore in the analysed regions (Rejection of null hypothesis).

Next, the verification is carried out on whether the assumption of homogeneity of variances. For this, the "bartlett.test" test is applied to check it:

```
bartlett.test(happy2015$Hscore,happy2015$Region)

        Bartlett test of homogeneity of variances

data:  happy2015$Hscore and happy2015$Region
Bartlett's K-squared = 27.188, df = 9, p-value = 0.001302
```

According to the data, verifying the homogeneity of variances, it is identified that at least two of them are different, this data is verified in the graph of the set of samples and using the Bartlett test that specifies the value of p-value less than 0.05, in this way the null hypothesis is rejected.

REGRESSION

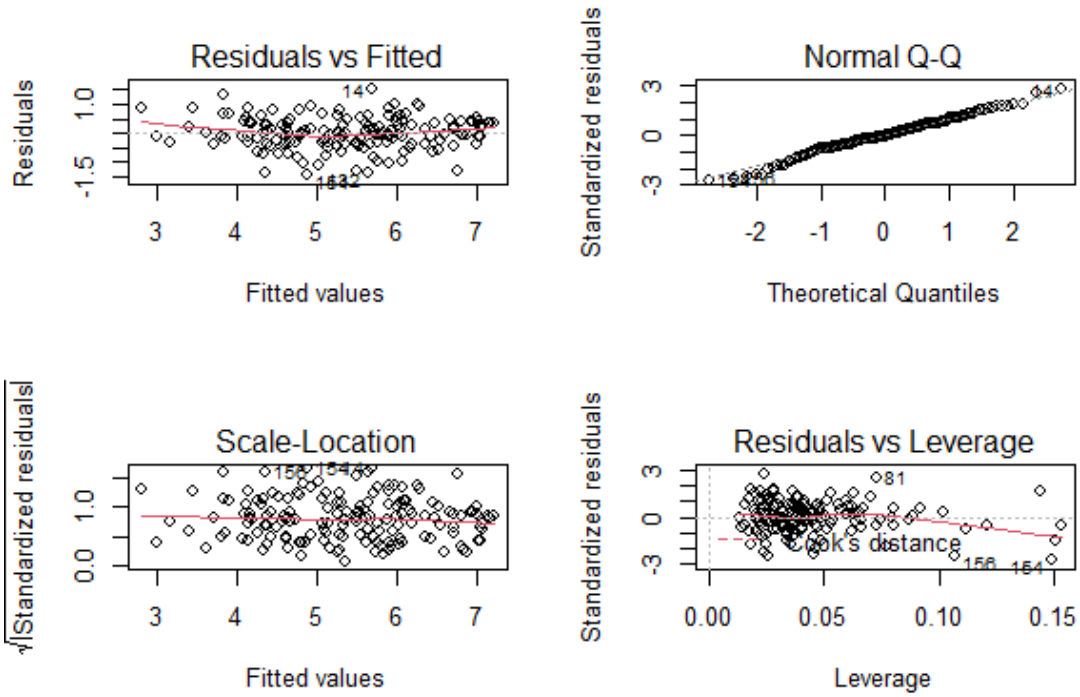
A Multiple linear regression analysis has been carried out with two data from 2015, for which the model has been done considering the dependent variable happiness

(Hscore)and as explanatory variables: GDP, family, life expectancy, freedom, trust and generosity.

```
ModeloRM <- lm(Hscore~GDP+family+lifexp+Freedom+Trust+Generosity, data = happy2015)
summary(ModeloRM)

Call:
lm(formula = Hscore ~ GDP + family + lifexp + Freedom + Trust +
    Generosity, data = happy2015)

Residuals:
     Min       1Q   Median       3Q      Max
-1.40484 -0.31734 -0.02814  0.37189  1.50130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8602     0.1905   9.766  < 2e-16 ***
GDP           0.8607     0.2203   3.907 0.000141 ***
family        1.4089     0.2227   6.327 2.69e-09 ***
lifexp        0.9753     0.3163   3.084 0.002433 **
Freedom       1.3334     0.3850   3.463 0.000694 ***
Trust         0.7845     0.4365   1.797 0.074302 .
Generosity    0.3889     0.3910   0.995 0.321471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.551 on 151 degrees of freedom
Multiple R-squared:  0.7772,    Adjusted R-squared:  0.7684
F-statistic: 87.81 on 6 and 151 DF,  p-value: < 2.2e-16
```

According the p-value reflects that the model is statistically significant and a happiness coefficient of 0.7772; while, in the analysis of the variables that interfere with the level of happiness in countries, GDP per capita has a greater impact, Family and Freedom, Life Expectancy has an effect of 95% with respect to the rest of predictors.

Then an Analysis of the model residuals and interpretation of the result has been done as well:



Graph 4. Residuals and results.

In the residual analysis, it is evident that there are no patterns different, although the lower part of the graph shows the presence of some high values. In the analysis of the normal, in the Q-Q graph it is verified that the variable dependent is normally distributed, whereas in the Scale review Location, the data is equally distributed throughout the ranges of the predictors. In the ratio of residuals vs leverage, it is the typical aspect when there are no cases influential. In conclusion, the independent variables influence the result of Happiness score.

Adding the nominal variable to the model:

```
ModeloRM2 <- lm(Hscore~GDP+family+lifexp+Freedom+Trust+Generosity+Region, data = happy2015)
summary(ModeloRM2)

Call:
lm(formula = Hscore ~ GDP + family + lifexp + Freedom + Trust +
    Generosity + Region, data = happy2015)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43699 -0.26343 -0.01104  0.29632  1.24909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.87864    0.56746   5.073 1.21e-06 ***
GDP          0.95771    0.22274   4.300 3.16e-05 ***
family       1.23995    0.22722   5.457 2.10e-07 ***
lifexp       0.43928    0.42000   1.046   0.2974
Freedom      0.90377    0.39168   2.307   0.0225 *
Trust        0.96687    0.43847   2.205   0.0291 *
Generosity   0.37807    0.43048   0.878   0.3813
RegionEA    -0.73105    0.43951  -1.663   0.0985 .
RegionMeNa  -0.51951    0.41164  -1.262   0.2090
RegionSA    -0.58883    0.42460  -1.387   0.1677
RegionSubA  -0.70158    0.45843  -1.530   0.1281
RegionCEE   -0.53492    0.41352  -1.294   0.1979
RegionLAC    0.09822    0.41045   0.239   0.8112
RegionNAM    0.17880    0.51794   0.345   0.7304
RegionSASI  -0.43794    0.45661  -0.959   0.3391
```

Not very big different found after adding the nominal variable Region to the model. With p-value: < 2.2e-16 it keep been statisticly significant, and get to explain 81.79% of the variance in happiness according to the Multiple R-squared and 79.87% considering the Adjusted R-squared

DISCUSION

In general, we can conclude that all the variables selected for the analysis have a certain degree of influence on the happiness levels for countries.

According to all the analysis it can be said that, just visualizing the data before analysis, there are important differences in the level of happiness between countries.

The family score tends to have the greatest impact on the Happiness score, Economy (GDP per Capita) have the second biggest impact. Trust has the lowest score of all conditions observed along with generosity.

Regarding the differences in happiness among the three years, according to the data obtained, yes, there are differences regarding the levels of happiness in the last three years. Although the values are almost similar regarding the calculation of the mean, it can be said that in 2016 there is a higher level of happiness than in the rest of years. In any case, it is not very representative, its minimal.

Regarding the relationships between the different regions according to the level of happiness, there are no according to the investigation. As seen in the boxplots, there are no relationships between each of the regions. It can be seen that between Australia and New Zealand (AUSNZ) and North America (NAM) there is a slight relationship in their variances, but the P value is very small, so it can be concluded that there is no HS relationship (level of happiness) in the analysed regions.

Also, according to the analysis it can be affirmed that The "happiest" countries are located in Europe, meanwhile, the "least happy" countries are located in Africa.

REFERENCES

Kaggle, W. H. (2018). Kaggle. Obtained from https://www.kaggle.com/unsdsn/world-happiness?select=2018.csv

Ludwigs, K., Lucas, R., Veenhoven, R. et al. Applied Research Quality Life (2019). https://doi.org/10.1007/s11482-019-09723-2

DATA1001 Project #1 Chandler Elissa, Nikolovski Mihael, Villar Miguel, Ando Koki https://rpubs.com/koki25ando/DATA1001TeamBver1

World Happiness REPORT Edited by John Helliwell, Richard Layard and Jeffrey Sachs https://www.researchgate.net/publication/233401584_World_Happiness_Report