Measuring Happiness

Where are we on the Cantril ladder?

Elisa D'Errico 3165871
Arianna Feliziani 3168356

January 9, 2023

Contents

1 Introduction

What we mean by happiness and satisfaction has long been studied by humans as it is, in the end, the final goal of human beings and the driving force of our actions. It is undoubtedly difficult to accurately measure happiness especially due to its multifaceted and complex nature: it is what we aim for and keep chasing over the course of our life, and yet we sometimes do not even realize how lucky we are and how much we should be thankful for what we have.
Our paper is based on surveys conducted in 165 countries about self-reported happiness and life satisfaction, "which have been found to be measuring subjective well-being with reasonable accuracy", [1] and it aims at investigating how different factors can influence it. More specifically, we thought it would have been interesting to focus our research on the data available over 6 years (2014-2019). The scope of our paper, in fact, is not to analyse happiness levels during periods of economic regression and distress. On the contrary, we thought it would have been more useful to select a time period in which, at least for the majority of the countries, welfare was increasing. Thus, after the 2008 crisis and avoiding to include the data collected during the pandemic, the time period chosen seemed the best option. The socio-political aspects of our choice are beyond the scope of this paper, and this is why we do not proceed with further discussion.
To conclude, while happiness is subjective and can be difficult to measure, statistical analysis can provide valuable insights into the factors that contribute to overall well-being in different countries.

2 Data

2.1 Data Source

After some research, we decided to select our datasets from the website "Our World In Data", where datasets about a wide range of different subjects are available. We decided to study if there was some correlation between happiness/life satisfaction and the following socio-economical, political or lifestyle-related factors: GNI, Homicide Rate, Human Rights, Education Levels, Investments in Healthcare, Mental Health, Internet Usage, Income Inequality (see Appendix A for the description of the variables).

2.2 Constructing the table

After the collection of the CSV files, we selected our columns of interest to get the data from the year 2014 unti 2019 (See R Code in Appendix E for more details). As stated in the introduction, we selected this time range as we believe that in these years the variables that we wanted to analyse did not suffer from major changes and could thus explain the degree of happiness of individuals.

2.3 Data Visualization

We started by looking at our object of study, which is happiness levels, for the year 2019. We chose this year as we believe that the pandemic deeply influenced our lifestyles and personalities, thus 2019 was the most recent year in which we still had a "normal" life (although, we might wonder: what does normal mean anymore?).

[1] Esteban Ortiz-Ospina and Max Roser (2013) - "Happiness and Life Satisfaction". Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/happiness-and-life-satisfaction

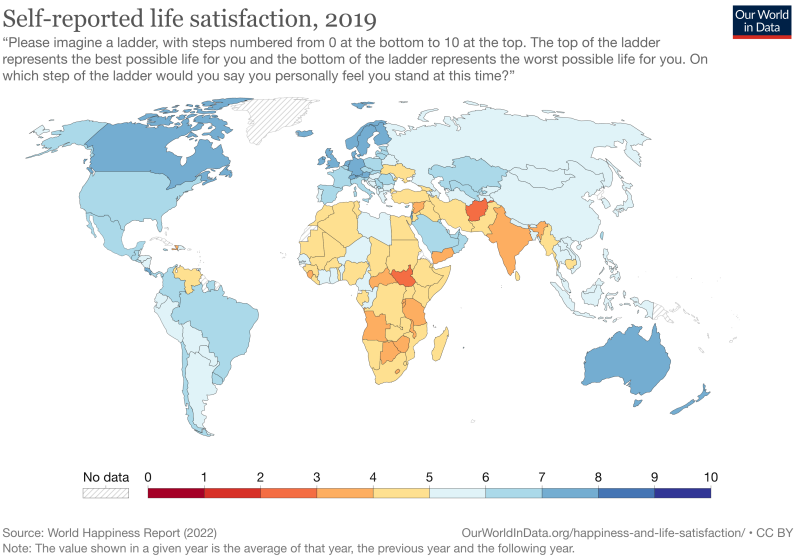

Figure 1: Self-Reported Happiness Levels, 2019

This image, color coded on a scale from 1 to 10, gives us some interesting hints: North America, Northern Europe and Australia have the highest happiness level, followed by central and Southern America, the rest of Europe and a few countries in Asia. Lastly, we have almost every country in South Asia and Africa which report very low levels of happiness.
The problem with this picture is that it is difficult, as we mentioned in the introduction and further discuss in the limitations section, to define a proper way of universally measuring happiness, due to the fact that each of us has a different interpretation of happiness and satisfaction, and this is moreover highly influenced by cultural and social differences.
By using the tools available on R, more specifically the ggplot function of the tidyverse package, we were then able to display some pairwise plots (Figures 2 and 3) that allowed us to understand if there was a graphical relation between the selected variables and life satisfaction/happiness. For some of our chosen factors (Homicides, Mental Health, Income Inequality) we expect a negative correlation, while for the rest of them (Human rights, Investments in Healthcare, GNI, Internet Usage, Education Level) we would expect a positive one. For the sake of consistency, the plots are displayed for our chosen year, 2019, although trends for other years are similar.
Only some of the graphical representations confirm our hypothesis: GNI, healthcare expenditure, human rights, Internet Access. The most surprising result arises from the Expected Years of Education, where we cannot recognise any clear correlation between the variables, but we would have expected a high correlation.
Also, in the cases of income inequality, homicides, mental health diseases we would have expected a clear negative correlation, and instead the plots show none. In particular, for what concerns income inequality, the situation is more complex, as we will discuss in the limitations section.

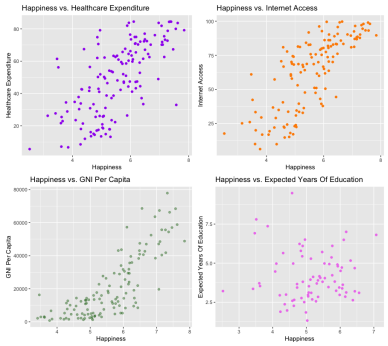

Figure 2: From top to Bottom, Clockwise: Health Expenditure, Internet Access, Expected Years of Education, GNI Per Capita

Figure 3: From top to Bottom, Clockwise: Deaths by Homicide, Mental Disorders, Income Inequality, Human Rights

3 Multivariate Linear Regression

We started with a multivariate linear regression so as to study the relationship between some explanatory variables and one response variable. The table was constructed using data from 165 countries over 6 years, with a total of 303 observations, and thus it has dimensions 303x9.
First and foremost, we performed the linear regression taking into account the eight explanatory variables and the response variable.
Then, we performed the correlation to find Multicollinearity using a graphical method and a qualitative method, respectively we plotted the pairwise scatterplot of the data and we performed the VIF test [2].
Even though from the graphical representation it might seem that there is some correlation among the explanatory variables, by running the VIF test, since no value is greater than 10, we are sure that the correlation is not significant.[3]
We then proceeded to perform the Diagnostic Plot to check if the key assumptions for linear regressions (homoscendasticity and normality of the residuals) are satisfied.

Since the Residuals vs Fitted plot (Figure 4) shows no clear pattern, we can say that there is homoscedasticity, i.e. all independent variables have similar or equal variance. As regards normality of the residuals, we can see graphically that they align very well on the diagonal line (Normal Q-Q plot, Figure 4), thus confirming normality. In addition, we plotted the histogram of the residuals and the boxplot to help us visualise the distribution (Figure 10).
To be sure of our result, we also performed a qualitative method, the Kolmogorov-Smirnov test, which confirms the normality of the distribution since the p-value is greater than 0.05 (see 7 in Appendix B).
We then proceed with analysing the results from the linear regression (see Figure 9 in Appendix B).
First, we see that the p-value of the F-statistic is smaller than 0.05, thus our model is valid. Furthermore, from the Adjusted R-Squared Value, we can see that our chosen factors explain 75% variance of our dependent variable.

[2]Meaures the strength of the correlation between the independent variables. It is calculates through the formula VIF = $\frac{1}{1-R}$, where R is the R-Squared Value.
[3]To view VIF values and pairwise representation refer to Appendix B, figures 6 and 8.

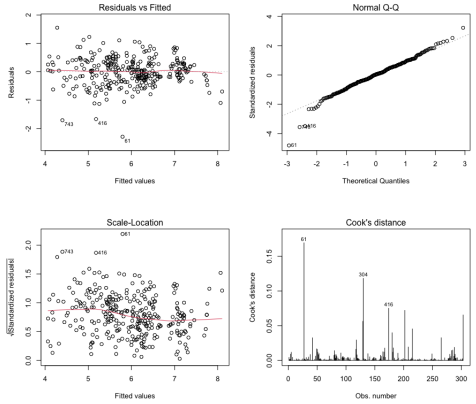

Figure 4: Diagnostic Plot. From top to bottom, clockwise: Residuals vs Fitted, Normal Q-Q, Scale-Location, Cook's Distance

On the other hand, looking at the p-values of the independent variables, we can see that three of those are not statistically significant. In particular, IncomeInequality, MentalHealth and Education all have p-values greater than 0.05.

4 Model Selection

We now look for better models (possibly with fewer variables), that can explain our response variable well enough.
We got the same conclusion using two different approaches, the testing methods (step both, step up and step down [4]) and the information theory approach.
For what concerns the first one, we used two different R packages: MASS and olsrr. Using the first one we performed the step up and step down method, which however yield different results (see Figures 11 and 12 in Appendix C). Thus we also performed the step both method (Figure 13 in Appendix C), which confirms the result of the step down one. For this reason, we consider that result, and thanks to the function ols_step_both_p we obtained the Stepwise Selection Summary Table (See Figure 5).

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	GNI	addition	0.628	0.627	150.0590	553.8843	0.5996
2	Homicides	addition	0.693	0.691	74.3970	498.3496	0.5462
3	HealthExpenditure	addition	0.726	0.723	36.3720	465.7982	0.5168
4	CivilRights	addition	0.743	0.740	16.9390	447.5383	0.5006
5	Internet	addition	0.755	0.750	5.4860	436.0924	0.4905

Figure 5: Stepwise Selection Summary

We see that from the step both method the prediction power is of 0.750 (Adjusted R-Squared) and that the variables selected for the model are five: GNI, Homicides, Internet, Civil Rights and Health Expenditure. Comparing this to the result Table 9, we see that we removed the independent variables that were not relevant (p-value > 0.05).
This result is confirmed by the other approach , which is based on penalisation of bigger models and uses the estimators AIC and SBC, of which we select the lowest value. By

[4]The step both function performs at each step a t-test for each one of the independent variables and checks if there is a lowest p-value. The others two methods have a similar approach

Figure 14, we can see that the minimum AIC value corresponds to the fifth row, i.e. takes into account the five variables GNI, Homicides, Internet, Civil Rights and Health Expenditure just like in the previous case. Likewise, for the SBC criteria, the fifth row is the one we select since it presents the lowest value.
The fact that the two results coincide, is an indicator of the goodness and validity of our model, since they lead to the same conclusion of keeping only five variables to describe the dependent one.

5 Limitations

It can be argued that one major limitation of this case study is the meaning we give to happiness. It is probably one of the few aspects of our lives which is extremely hard to measure in an objective way, but it was exactly this which pushed us to pursue this project question. Some of the limitations of our research relate to the chosen variables, in particular the homicide rate, income inequality, and the mental health disorders. As regards the homicide rate, we see that there is a discrepancy between the graph representation in Figure 3 and the actual result of the analysis, where we see that the homicide rate has the lowest p-value after the GNI, and thus a strong correlation with the study variable. In this case, outliers comprise almost entirely of countries from South America and Islands. One limitation of this variable is for sure the fact that we wanted to look at the level of criminality, and using the number of homicides as a proxy is clearly reductive.
For Income inequality, we would expect a clear negative correlation, but instead the graph is very sparse. We understand that, differently from what we would have expected, Income Inequality does not contribute and affect life satisfaction, but is out the scope of this paper to understand why.
Lastly, the result obtained for mental health is the most unexpected one. We would have thought that as mental health diseases increase, happiness decreases. Instead, the graph shows no correlation and this is confirmed after having carried out the analysis. This is probably due to the fact that increasingly many people with mental health problems are unreported, thus it is hard to take into consideration the correct number. Another variable that we want to analyse more in detail the Education Level. Even though from the graph (Figure 2) one cannot grasp a strong correlation, from the analysis it emerges that the variable is actually relevant, although not as much as we would have believed. This could be due to the dataset that we chose, which was based on the predicted levels of education, and not the actual levels. To conclude, to make the model more accurate, we could have taken into consideration a single continent, for example Europe, so as to increase "fairness" of comparisons: within the same country, people are usually exposed to similar living conditions and have similar access to resources.

6 Conclusions

Looking at the figures 14 and 5, we see that even though the highest Adjusted R-Square value (0.7508) is obtained using six or seven parameters, thanks to the AIC, SBC and testing methods we are able to select a model which has a level of accuracy of 0.7505, almost equal to the one obtained with six or seven variables, but employs less parameters. Thus we prefer this model, of which the reader can see the regression summary on Appendix D, table 15.

A Appendix - Description Of Variables

1. GNI - Gross National Income per capita: sum of the Gross domestic Product and the income from abroad;
2. Homicide Rate: Measure of Interpersonal Violence, excluding civilian and military deaths during wars and genocides;
3. Human Rights: Extent to which physical integrity rights, civil right and political rights are guaranteed to citizens;
4. Education Levels: Number of years a child is expected to study at school;
5. Investments in healthcare: Share of current health expenditures funded from domestic public sources for health;
6. Mental Health: Share of population with mental health disorders;
7. Internet Usage: Share of the population with internet access;
8. Income Inequality: Inequality of incomes measured by the GINI coefficient.

B Appendix - Multivariate Linear Regression Tables

CivilRights	IncomeInequality	HealthExpenditure	MentalHealth	GNI	Homicides
1.835494	1.691975	2.031082	1.270629	3.491453	1.599749
Internet	Education				
3.564430	1.281475				

Figure 6: VIF Test

Test	Statistic	pvalue
Shapiro-Wilk	0.9793	2e-04
Kolmogorov-Smirnov	0.048	0.4873
Cramer-von Mises	38.299	0.0000
Anderson-Darling	0.7667	0.0458

Figure 7: Kolmogorov-Smirnov Test. Since the number of observations is greater than 50 we use this test instead of the Shapiro-Wilk test.

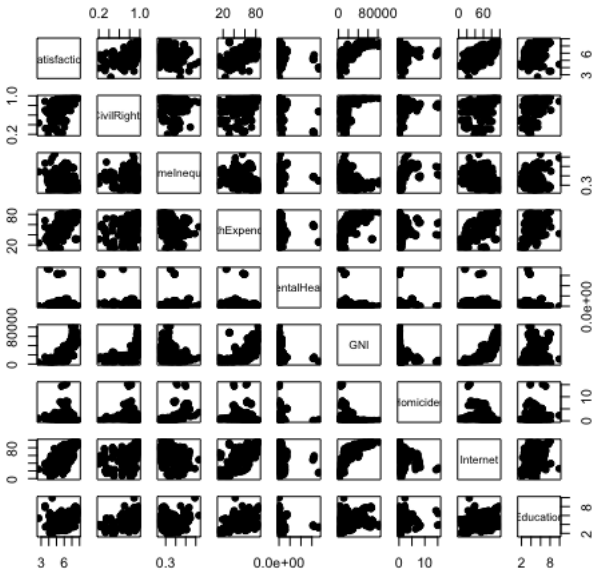

Figure 8: Pairwise representation of variables

```
Call:
lm(formula = Satisfaction ~ ., data = tb.final)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28934 -0.28137  0.02364  0.29451  1.55121

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.173e+00  2.486e-01  12.762  < 2e-16 ***
CivilRights       8.474e-01  2.095e-01   4.044 6.71e-05 ***
IncomeInequality  1.885e-01  4.713e-01   0.400 0.689445
HealthExpenditure 8.232e-03  2.273e-03   3.622 0.000345 ***
MentalHealth      1.133e-09  1.271e-09   0.892 0.373326
GNI               2.918e-05  2.933e-06   9.949  < 2e-16 ***
Homicides         9.976e-02  1.544e-02   6.460 4.32e-10 ***
Internet          8.601e-03  2.336e-03   3.682 0.000275 ***
Education         2.852e-02  2.392e-02   1.192 0.234029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4909 on 294 degrees of freedom
Multiple R-squared:  0.7567,    Adjusted R-squared:  0.7501
F-statistic: 114.3 on 8 and 294 DF,  p-value: < 2.2e-16
```

Figure 9: Linear Regression Summary

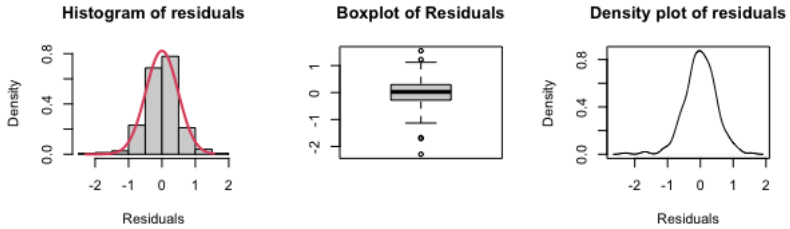

Figure 10: From the different representations of the distribution we can clearly see a high level of normality, with few outliers that are distributed quite evenly on the extremities and equally long whiskers.

C Appendix - Model Selection Tables

```
Call:
lm(formula = Satisfaction ~ CivilRights + IncomeInequality +
    HealthExpenditure + MentalHealth + GNI + Homicides + Internet +
    Education, data = tb.final)

Coefficients:
     (Intercept)       CivilRights   IncomeInequality  HealthExpenditure       MentalHealth               GNI          Homicides           Internet          Education
       3.173e+00          8.474e-01          1.885e-01          8.232e-03          1.133e-09          2.918e-05          9.976e-02          8.601e-03          2.852e-02
```

Figure 11: Step Up Method

```
Call:
lm(formula = Satisfaction ~ CivilRights + HealthExpenditure +
    GNI + Homicides + Internet, data = tb.final)

Coefficients:
     (Intercept)       CivilRights  HealthExpenditure               GNI          Homicides           Internet
       3.374e+00          8.201e-01          8.605e-03          2.959e-05          1.039e-01          8.493e-03
```

Figure 12: Step Down Method

```
Call:
lm(formula = Satisfaction ~ CivilRights + HealthExpenditure +
    GNI + Homicides + Internet, data = tb.final)

Coefficients:
     (Intercept)       CivilRights  HealthExpenditure               GNI          Homicides           Internet
       3.374e+00          8.201e-01          8.605e-03          2.959e-05          1.039e-01          8.493e-03
```

Figure 13: Step Both Method

```
                                    Best Subsets Regression
-----------------------------------------------------------------------------------------------------
Model Index    Predictors
-----------------------------------------------------------------------------------------------------
     1         GNI
     2         GNI Homicides
     3         HealthExpenditure GNI Homicides
     4         CivilRights HealthExpenditure GNI Homicides
     5         CivilRights HealthExpenditure GNI Homicides Internet
     6         CivilRights HealthExpenditure GNI Homicides Internet Education
     7         CivilRights HealthExpenditure MentalHealth GNI Homicides Internet Education
     8         CivilRights IncomeInequality HealthExpenditure MentalHealth GNI Homicides Internet Education
-----------------------------------------------------------------------------------------------------

                                        Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------------------------
                        Adj.        Pred
Model    R-Square    R-Square    R-Square       C(p)        AIC         SBIC         SBC         MSEP       FPE       HSP       APC
-----------------------------------------------------------------------------------------------------------------------------------
  1        0.6283      0.6271       0.623   150.0595   553.8843    -307.5050   565.0255   108.9270   0.3619   0.0012   0.3766
  2        0.6926      0.6906      0.6856    74.3971   498.3496    -362.6999   513.2046    90.3899   0.3013   0.0010   0.3135
  3        0.7257      0.7230      0.7179    36.3717   465.7982    -394.7782   484.3668    80.9193   0.2706    9e-04   0.2816
  4        0.7435      0.7400      0.7346    16.9393   447.5383    -412.5633   469.8207    75.9404   0.2548    8e-04   0.2651
  5        0.7546      0.7505      0.7445     5.4861   436.0924    -423.5217   462.0885    72.8896   0.2453    8e-04   0.2553
  6        0.7558      0.7508      0.7416     6.0730   436.6448    -422.8573   466.3546    72.7881   0.2458    8e-04   0.2558
  7        0.7565      0.7508      0.7396     7.1600   437.7058    -421.6920   471.1294    72.8097   0.2466    8e-04   0.2567
  8        0.7567      0.7501      0.7364     9.0000   439.5409    -419.7867   476.6782    73.0184   0.2481    8e-04   0.2582
-----------------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

Figure 14: Subset Regression Summary

D Appendix - Final Model Summary

```
Call:
lm(formula = Satisfaction ~ CivilRights + Homicides + HealthExpenditure +
    GNI + Internet, data = tb.final)

Residuals:
     Min       1Q   Median       3Q      Max
-2.18032 -0.27610 -0.00039  0.29417  1.50929

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.374e+00  1.570e-01  21.491  < 2e-16 ***
CivilRights       8.201e-01  1.877e-01   4.370 1.72e-05 ***
Homicides         1.039e-01  1.353e-02   7.677 2.37e-13 ***
HealthExpenditure 8.605e-03  2.239e-03   3.844 0.000148 ***
GNI               2.959e-05  2.883e-06  10.262  < 2e-16 ***
Internet          8.493e-03  2.314e-03   3.671 0.000286 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4905 on 297 degrees of freedom
Multiple R-squared:  0.7546,    Adjusted R-squared:  0.7505
F-statistic: 182.7 on 5 and 297 DF,  p-value: < 2.2e-16
```

Figure 15: Final Model Summary

E Appendix - Code R (Graphs)

```
install.packages("dplyr")
install.packages("tidyverse")
install.packages("ggpubr")


library("dplyr")
library("tidyverse")
library("ggpubr")


#Import Data Sets

# Happiness
happy <- read.csv("happiness-cantril-ladder.csv")
dfh <- data.frame(happy)

# Rights
rights <- read.csv("distribution-human-rights-vdem.csv")
dfr <- data.frame(rights)
dfr$region <- NULL

# Income
income <- read.csv("economic-inequality-gini-index.csv")
dfi <- data.frame(income)

# Education
edu <- read.csv("total-government-expenditure-on-education-gdp.csv")
dfe <- data.frame(edu)

# Health Expenditure
health <- read.csv("share-of-public-expenditure-on-healthcare-by-country.csv")
dfhe <- data.frame(health)

# Mental diseases
mental <- read.csv("people-with-mental-health-disorders.csv")
dfm <- data.frame(mental)

# GNI
growth <- read.csv("gross-national-income-per-capita.csv")
dfg <- data.frame(growth)

# Homicides
crime <- read.csv("share-of-deaths-homicides.csv")
dfc <- data.frame(crime)

# Internet
internet <- read.csv("share-of-individuals-using-the-internet.csv")
```

```
dfn <- data.frame(internet)

# Adjust Happiness Data Frame
colnames(dfh)[4] = "Score"
dfh_cys <- dfh %>% select(Code,Year,Score)

#select year 2019
dfh_cs <- dfh_cys[dfh_cys$Year == "2019",]



######### HUMAN RIGHTS ##########
colnames(dfr)[4] = "Score"
dfr_cys <- dfr %>% select(Code,Year,Score)
dfr_cs <- dfr_cys[dfr_cys$Year == "2019",]

#new data frame happiness and rights
df_hr <- merge(x = dfh_cs, y = dfr_cs, by = "Code")
df_hr_toplot <- df_hr %>% select(Code, Score.x , Score.y )

#plot of data happiness rights
hr <- ggplot(data = df_hr_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 0.5, col = "blue")
hrf <- hr + ggtitle("Happiness vs. Human Rights")
+ xlab("Happiness") + ylab("Human Rights")



######### INCOME INEQUALITY ##########
colnames(dfi)[4] = "Score"
dfi_cys <- dfi %>% select(Code,Year,Score)
dfi_cs <- dfi_cys[dfi_cys$Year == "2019",]

#new data frame happiness and income
df_hi <- merge(x = dfh_cs, y = dfi_cs, by = "Code")
df_hi_toplot <- df_hi %>% select(Code, Score.x, Score.y)

#plot of data happiness-income
hi <- ggplot(data = df_hi_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 0.5, col = "red")
hif <- hi + ggtitle("Happiness vs. Income Inequality")
+ xlab("Happiness") + ylab("Income Inequality")


######### EDUCATION ##########
colnames(dfe)[4] = "Score"
dfe_cys <- dfe %>% select(Code,Year,Score)
dfe_cs <- dfe_cys[dfe_cys$Year == "2019",]
```

```
#new data frame happiness and education
df_he <- merge(x = dfh_cs, y = dfe_cs, by = "Code")
df_he_toplot <- df_he %>% select(Code, Score.x, Score.y)

#plot of data happiness education
he <- ggplot(data = df_he_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 1.5, col = "violet")
hef <- he + ggtitle("Happiness vs. Expected Years Of Education")
+ xlab("Happiness") + ylab("Expected Years Of Education")


######### HEALTHCARE EXPENDITURE ##########
colnames(dfhe)[4] = "Score"
dfhe_cys <- dfhe %>% select(Code,Year,Score)
dfhe_cs <- dfhe_cys[dfhe_cys$Year == "2011",]

#new data frame happiness and healthcare expenditure
df_hhe <- merge(x = dfh_cs, y = dfhe_cs, by = "Code")
df_hhe_toplot <- df_hhe %>% select(Code, Score.x, Score.y)

#plot of data happiness healthcare expenditure
hhe <- ggplot(data = df_hhe_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 1.5, col = "purple")
hhef <- hhe + ggtitle("Happiness vs. Healthcare Expenditure")
+ xlab("Happiness") + ylab("Healthcare Expenditure")


######### MENTAL HEALTH AND SUBSTANCE USE ##########
colnames(dfm)[4] = "Score"
dfm_cys <- dfm %>% select(Code,Year,Score)
dfm_cs <- dfm_cys[dfm_cys$Year == "2019",]

#new data frame happiness and mental health
df_hm <- merge(x = dfh_cs, y = dfm_cs, by = "Code")
df_hm_toplot <- df_hm %>% select(Code, Score.x, Score.y)

#plot of data happiness mental health
hm <- ggplot(data = df_hm_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 1.5, col = "orange")
hmf <- hm + ggtitle("Happiness vs. People with Mental Health Disorders")
+ xlab("Happiness") + ylab("People with Mental Health Disorders")


######### GROSS NATIONAL INCOME PER CAPITA ##########
colnames(dfg)[4] = "Score"
dfg_cys <- dfg %>% select(Code,Year,Score)
dfg_cs <- dfg_cys[dfg_cys$Year == "2019",]

#new data frame happiness and GNI
```

```
df_hg <- merge(x = dfh_cs, y = dfg_cs, by = "Code")
df_hg_toplot <- df_hg %>% select(Code, Score.x, Score.y)

#plot of data happiness GNI
hg <- ggplot(data = df_hg_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 0.4, col = "darkgreen")
hgf <- hg + ggtitle("Happiness vs. GNI Per Capita")
+ xlab("Happiness") + ylab("GNI Per Capita")


######### DEATHS BY HOMICIDE ##########
colnames(dfc)[4] = "Score"
dfc_cys <- dfc %>% select(Code,Year,Score)
dfc_cs <- dfc_cys[dfc_cys$Year == "2019",]

#new data frame happiness and deaths by homicide
df_hc <- merge(x = dfh_cs, y = dfc_cs, by = "Code")
df_hc_toplot <- df_hc %>% select(Code, Score.x, Score.y)
country.names <- df_hc_toplot["Code"]
country.names.final <- unique(country.names)

#plot of data happiness deaths by homicide
hc <- ggplot(data = df_hc_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 0.3, col = "darkblue")
hcf <- hc + ggtitle("Happiness vs. Deaths By Homicide")
+ xlab("Happiness") + ylab("Death By Homicide")


######### INTERNET ACCESS ##########
colnames(dfn)[4] = "Score"
dfn_cys <- dfn %>% select(Code,Year,Score)
dfn_cs <- dfn_cys[dfn_cys$Year == "2019",]

#new data frame happiness and internet access
df_hn <- merge(x = dfh_cs, y = dfn_cs, by = "Code")
df_hn_toplot <- df_hn %>% select(Code, Score.x, Score.y)

#plot of data happiness internet access
hn <- ggplot(data = df_hn_toplot,aes(x = Score.x, y = Score.y))
+ geom_point(alpha = 4, col = "darkorange")
hnf <- hn + ggtitle("Happiness vs. Internet Access")
+ xlab("Happiness") + ylab("Internet Access")

# Display 4 in each page
ggarrange(hcf,hmf,hif,hrf)
ggarrange(hhef,hnf,hgf,hef)
```

F Appendix - Code R (Multivariate Linear Regression)

```
###### STATISTICS PROJECT ######
##### Measuring Happiness ######

# Elisa D'Errico
# Arianna Feliziani

# Install Packages
install.packages("dplyr")
install.packages("ggplot2")
install.packages("car")
install.packages("tidyverse")
install.packages("olsrr")
install.packages("MASS")

library("olsrr")
library("tidyverse")
library("dplyr")
library("ggplot2")
library("car")
library("MASS")
select <- dplyr::select

# Import data sets

# Happiness
happy <- read.csv("happiness-cantril-ladder.csv")
dfh <- data.frame(happy)

# Rights
rights <- read.csv("distribution-human-rights-vdem.csv")
dfr <- data.frame(rights)
dfr$region <- NULL

# Income
income <- read.csv("economic-inequality-gini-index.csv")
dfi <- data.frame(income)

# Education
edu <- read.csv("total-government-expenditure-on-education-gdp.csv")
dfe <- data.frame(edu)

# Health Expenditure
health <- read.csv("share-of-public-expenditure-on-healthcare-by-country.csv")
dfhe <- data.frame(health)

# Mental diseases
mental <- read.csv("people-with-mental-health-disorders.csv")
```

```
dfm <- data.frame(mental)

# GNI
growth <- read.csv("gross-national-income-per-capita.csv")
dfg <- data.frame(growth)

# Homicides
crime <- read.csv("share-of-deaths-homicides.csv")
dfc <- data.frame(crime)

# Internet
internet <- read.csv("share-of-individuals-using-the-internet.csv")
dfn <- data.frame(internet)

# Make the table and select the years
df.list <- list(dfh, dfr, dfi, dfhe, dfm, dfg, dfc, dfn, dfe)
tb <- Reduce(function(x, y) merge(x, y, all=FALSE), df.list)
colnames(tb)[4] = "Satisfaction"
colnames(tb)[5] = "CivilRights"
colnames(tb)[6] = "IncomeInequality"
colnames(tb)[7] = "HealthExpenditure"
colnames(tb)[8] = "MentalHealth"
colnames(tb)[9] = "GNI"
colnames(tb)[10] = "Homicides"
colnames(tb)[11] = "Internet"
colnames(tb)[12] = "Education"
#temp <- tb %>% select(-Entity)
tb <- tb[tb$Year > "2013",]
tb.final <- tb %>% select(-Entity, -Year, -Code)

# Run Linear Regression with Multicolinearity Issue

multi.reg <- lm(Satisfaction ~ ., data = tb.final)
summary(multi.reg)

# Perform Correlation to find Multicolinearity

# 1: Pairwise Scatterplot
pairs(tb.final,pch=19)

# 2: Perform VIF's Test to study correlation between the predictors
vif(multi.reg)

# Perform Diagnostic Plot
par(mfrow=c(2,2))
plot(multi.reg,which=1:4)

# Histogram of Residuals
par(mfrow=c(1,3))
```

```
hist.res <- residuals(multi.reg)
n.line <- seq( min(hist.res), max(hist.res), length = 60)
nc <- dnorm( n.line , mean = mean(hist.res), sd = sd(hist.res))
hist( hist.res, main="Histogram of residuals", prob = TRUE,
      ylim = c(0, max(nc)), xlab = "Residuals")
lines(n.line, nc, col = 2, lwd = 2)
boxplot(hist.res, main = "Boxplot of Residuals")
plot(density(hist.res), main="Density plot of residuals",
     ylab="Density",xlab="Residuals")

#kolmogorov-smirnov

ols_test_normality(multi.reg)

######## SUBSET SELECTION SUMMARY ########

# Best Subset Summary
multi.reg.model.subset <- ols_step_best_subset(multi.reg)
multi.reg.model.subset

# Step Both
model <- lm(Satisfaction ~ ., data = tb.final)
ols_step_both_p(model)
both <- stepAIC(multi.reg, direction = "both", trace = FALSE)
both

# Step Forward
forward <- stepAIC(multi.reg, direction = "forward", trace = FALSE)
forward

# Step Backward
backward <- stepAIC(multi.reg, direction = "backward", trace = FALSE)
backward

#Final Model Summary
final.model <- lm(Satisfaction ~ CivilRights + Homicides
                  + HealthExpenditure + GNI + Internet , data = tb.final)
summary(final.model)
```