Happiness in COVID

By

Group 12
Yuyan Shi, Samir Epili, Bhavna Kaparaju, Mauricio Morales

MIS S381N Data Analytics Programming Project

August 10, 2020

1. Introduction

1.1. Problem & Interest

What started as a single confirmed case early 2020 in China, sporadically spread across the whole world within the first quarter of 2020. COVID-19 has now become a global pandemic and one of the most serious global health crises. While its impact on the world's health systems including fatalities continue to rise, The United Nations published the happiness index report for 2020 in march. The happiness index is calculated by social support, freedom to make life choices, and other key variables. Our team is curious about whether each country's severity of the pandemic is related to the happiness index, including key variables. Figuring out what factors play an important role, the government could have a better understanding of what should be prioritized when the next big health crisis comes up.

1.2. Dataset Description

Our data comes from three sources:

1. 2020 happiness index report:
   In this dataset, we use key variables in the report to predict happiness index. Also, we use the variables in the report to predict the infection rate, which will be defined in the fourth part. (**Link 1.**)
2. COVID 19 confirmed cases:
   The dataset is scraped on the John Hopkins Coronavirus Resource Center. The dataset is the accumulated number of confirmed cases of each province in each country. The dates are from 1/22/2020 to 7/26/2020.
3. World geo json file:
   The world geo json file is downloaded from the website. We will be using the coordinates of points of each country's multipolygon to make heat maps using a folium module. (**Link 2.**)

Exploratory Analysis

Before we dive into the formal model building, we made two heat maps to visually see the relationship between number of confirmed cases and happiness index. We can see some similar patterns between two maps. Basically, the higher the happiness index the fewer confirmed cases, except the United States and countries with missing information. (**Figure 1 - 2.**)

To further visualize the data, we plotted multiple variables of the happiness index data set that we thought were relevant and compared them to each other. We calculated the max infection rate for each country to give all countries an equal playing field and give a more fair comparison given the varying population sizes in different countries. The max infection rate is the highest difference in the number of cases between two dates.

For our happiness index plots, we started by plotting social support versus healthy life expectancy. As expected the plot clearly indicated, the higher the social support, the higher the healthy life expectancy. We then plotted GDP versus life expectancy and got similar results. We also plotted life expectancy versus freedom to make life choices and there was no clear pattern but generally, the higher the freedom to make life choices, the higher the life expectancy. From this, we concluded that GDP and social support were important factors when it came to people living long, healthy, lives. (**Figures 8 - 10.**)

When we plotted these same variables individually against the max infection rate (log) that we calculated there were clear patterns again with GDP and social support. The higher the GDP and social support, the higher the max infection rate was. There were no clear patterns with life expectancy and freedom to make life choices. (**Figures 11 -13.**)

We also found the standard deviations for the variables Ladder score, Logged GDP per capita, Social support , and Healthy life expectancy, and the log of the max infection rate to give us a little glimpse of how out linear regression will be. (**Refer to Table 1**)

Solution and Insights

To further investigate the relationship between happiness and the number of confirmed coronavirus cases, our group created a series of linear regression models.

Our group first created a linear regression model using only the happiness data set to find the most significant predictors of happiness. We found that the p-values of generosity and perceptions of corruption were much higher than 0.05, which indicates that they are unreliable predictors of happiness. Thus, we removed these predictors from our final model in **Figure 3.** The most important, non-zero variables (as indicated by the confidence intervals), were logged GDP per capita, social support, healthy life expectancy, and freedom to make life choices. Performing a 70/30 training/test split, we found an out-of-sample RMSE of 0.600 happiness score units.

Next, our group wanted to observe the effect of using happiness as a predictor of infection rate. Two approaches were taken to correlate the two variables. First, we correlated an estimated overall increase of COVID cases per day with happiness scores. In other words, we calculated an ordinary linear model for each country's data between March and April 2020 to obtain an estimate of how much the country's cases increased per day (slope). For this, we didn't do cross validation, since we are not interested in doing any prediction of this, we wanted to have the best model for the data we have.

Once we obtained the slopes per country, two models were obtained with this data: an OLS and a regression tree. The OLS attempt was done using all the predictors, resulting in a R2: 0.072. This result means that this model describes just 7.2% of the relationship. **(Figure 15)** Going through the p-values of every predictor, we found that the predictor that has more relevance in this relationship is gdp, for which we did another OLS and obtained a R2: 0.044. This R2 is lower than before because we didn't use all of the information of the dataset. However, it's p-value is 0.011 **(Figure 16)**. The model was plotted to graphically see its performance. **(Figure 17)**

Figure 17, shows that there are some values that can be considered as outliers due to its distance to the rest of the data points. Another model was done without these points (U.S.A, Spain, Italy, Germany, Iran, Turkey and Russia) and we got an R2: 0.168. Great improvement from previous model **(Figure 18)**. A scatterplot was done using this new data set **(Figure 19)** and saw that there was a non-linear tendency, so a new model using gdp2 was done, with an R2: 0.178 **(Figure 20)**. Plotted model in **Figure 21.**

For the regression tree we used the whole dataset and obtained quite different results **(Figure 22).** The tree says that the most important factor to predict slope based on happiness data set, is the life expectancy. We went back to OLS and did a new model without the outliers, using life expectancy, life expectancy 2, and life expectancy 3. The R2: is slightly better than the one using gdp with an R2: 0.183 **(Figure 23)**. Plotted model in **Figure 24.**

The second approach was to correlate the maximum spike in confirmed cases within each country between two consecutive days and plotting these against happiness and its various predictors. The reason maximum spike, or maximum infection rate, was considered is because we hypothesized that countries with very low happiness ratings may show varying degrees of infection severity when the pandemic first arrived compared to those with very high happiness ratings, and vice versa.

Initially, the correlation matrix shown in **Table 2** showed that there is very little correlation between happiness and the maximum increase in cases. However, taking the homoscedasticity assumption into consideration we found that the variance is more equally distributed when taking the log of the maximum increase in cases. This may also be due to one of the most significant predictors of happiness, GDP per capita, also being logged in the original data set.

Our regression model using happiness score as a predictor of max rate resulted in a poor R^2 of 0.265 (**Figure 5**) and an out-of-sample RMSE of 1.87, which is within one standard deviation of the log max infection rate from the original data (**Table 1**). Moreover, the p-value of essentially 0, along with non-zero confidence intervals for the happiness score, show that happiness is a significant variable in the model. Overall, we can infer from this model that happiness, or more specifically, some predictors of happiness are also significant predictors of the max infection rate. To investigate this further, our group created a regression model using the most significant predictors of happiness (logged GDP per capita, freedom to make life choices, healthy life expectancy, social support) to predict the max infection rate in a given country (**Figure 6**).

Of the most significant happiness predictors, we found that log GDP per capita is the most significant predictor of max infection rate. This produced a superior R^2 of 0.353 which is higher than that of using happiness rating alone to predict max infection rate. This is likely because of factors that may be effective predictors of happiness, but poor predictors of max infection rate. Additionally, an out-of-sample RMSE of 1.753 was obtained which is improved from the 1.87 received from using happiness score alone.

Another analysis we performed was regression tree analysis to ensure we were looking at the right variables when performing linear regression. The four happiness predictors that were previously determined as important predictors were also used in a regression tree analysis. The results contradicted those found by the linear regression by selecting healthy life expectancy as the most significant predictor of log max infection rate (**Figure 7**). However, the regression tree also produced a significantly worse RMSE of 2.09 which is greater than one standard deviation of the log max infection rate.

From our analyses, we found that a small factor in predicting how severe a spike is in a country where confirmed COVID cases are unknown or inaccurate is happiness - specifically GDP per capita. This can be concluded from our linear regression model which had the lowest RMSE compared to our regression tree model. However, correlation does not necessarily imply causation. It is almost unintuitive to think that wealthier countries have more cases of coronavirus, however, these results can be interpreted as higher GDP likely leads to more reported cases because of increased testing. Only with more statistical testing and the addition of more significant, underlying variables can the model be improved to better predict pandemic response of a country, so their government can better prepare for the future.

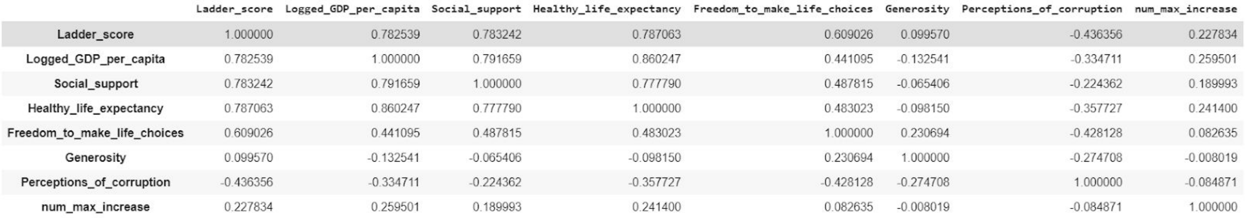
Appendix

Tables and Figures

Table 1. Standard deviations of happiness predictors and maximum infection rate.

Variables	Standard Deviations
Ladder Score	1.16
Logged GDP per capita	1.12
Social support	0.11
Healthy life expectancy	6.66
Freedom to make life choices	0.11
Maximum increase in infections	2.05

Table 2. Correlation matrix of happiness predictors and maximum infection rate.



	Ladder_score	Logged_GDP_per_capita	Social_support	Healthy_life_expectancy	Freedom_to_make_life_choices	Generosity	Perceptions_of_corruption	num_max_increase
Ladder_score	1.000000	0.782539	0.783242	0.787063	0.609026	0.099570	-0.436356	0.227834
Logged_GDP_per_capita	0.782539	1.000000	0.791659	0.860247	0.441095	-0.132541	-0.334711	0.259501
Social_support	0.783242	0.791659	1.000000	0.777790	0.487815	-0.065406	-0.224362	0.189993
Healthy_life_expectancy	0.787063	0.860247	0.777790	1.000000	0.483023	-0.098150	-0.357727	0.241400
Freedom_to_make_life_choices	0.609026	0.441095	0.487815	0.483023	1.000000	0.230694	-0.428128	0.082635
Generosity	0.099570	-0.132541	-0.065406	-0.098150	0.230694	1.000000	-0.274708	-0.008019
Perceptions_of_corruption	-0.436356	-0.334711	-0.224362	-0.357727	-0.428128	-0.274708	1.000000	-0.084871
num_max_increase	0.227834	0.259501	0.189993	0.241400	0.082635	-0.008019	-0.084871	1.000000

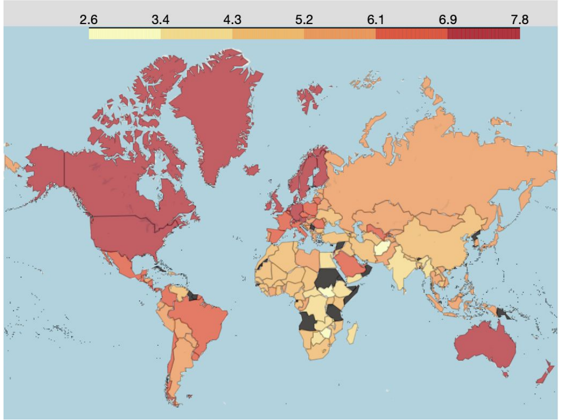


Figure 1. Heat map of happiness index

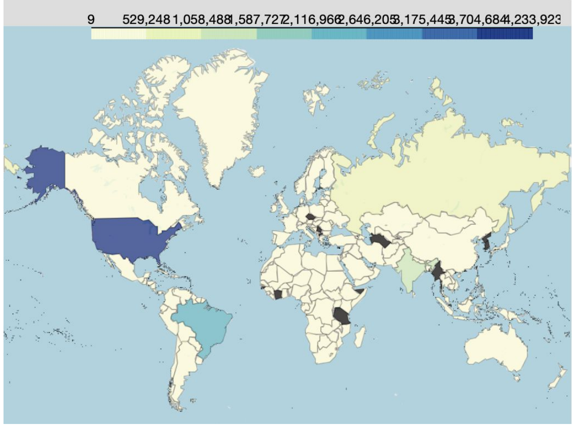
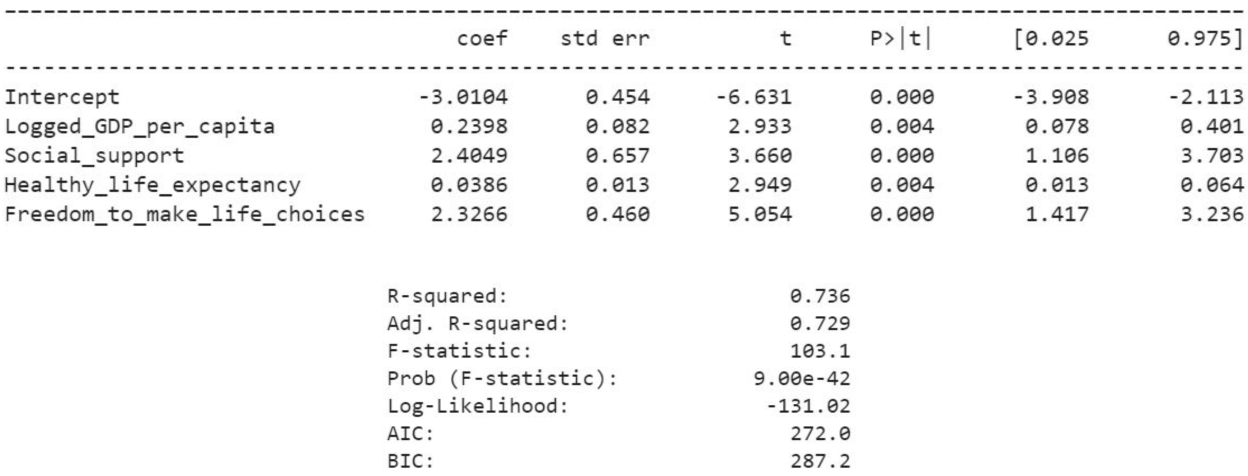


Figure 2. Heat map of confirmed cases (until 7/26/20)



```
==============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  -3.0104      0.454     -6.631      0.000      -3.908      -2.113
Logged_GDP_per_capita       0.2398      0.082      2.933      0.004       0.078       0.401
Social_support              2.4049      0.657      3.660      0.000       1.106       3.703
Healthy_life_expectancy     0.0386      0.013      2.949      0.004       0.013       0.064
Freedom_to_make_life_choices 2.3266     0.460      5.054      0.000       1.417       3.236

                         R-squared:                       0.736
                         Adj. R-squared:                  0.729
                         F-statistic:                     103.1
                         Prob (F-statistic):           9.00e-42
                         Log-Likelihood:                -131.02
                         AIC:                             272.0
                         BIC:                             287.2
```

Figure 3a.,3b. Summary statistics of linear regression of happiness dataset.

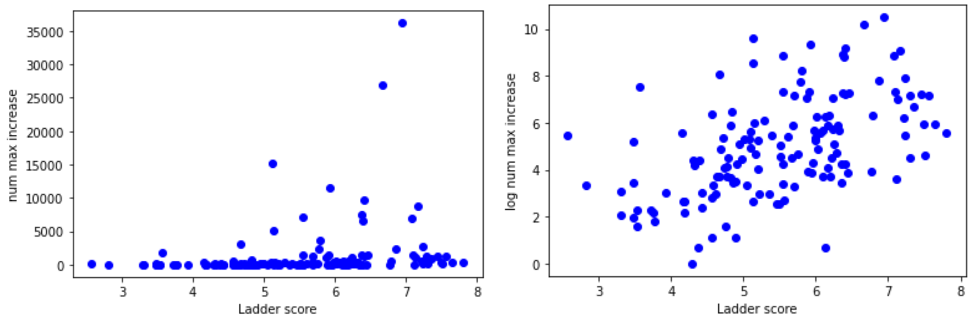
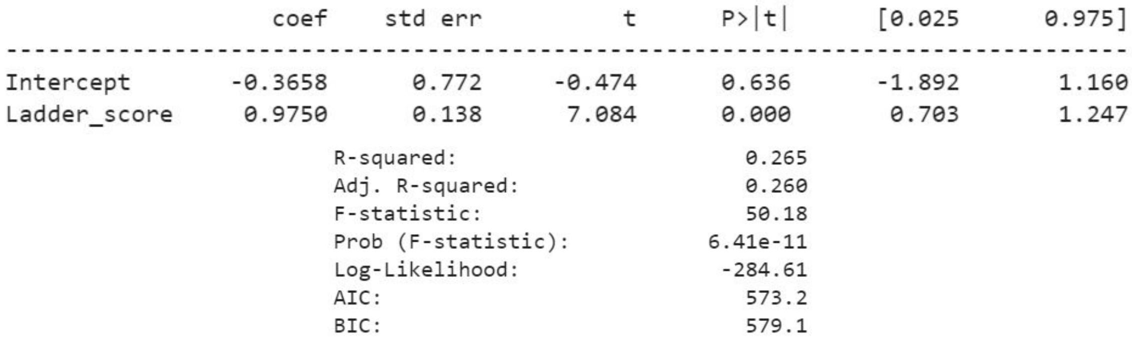


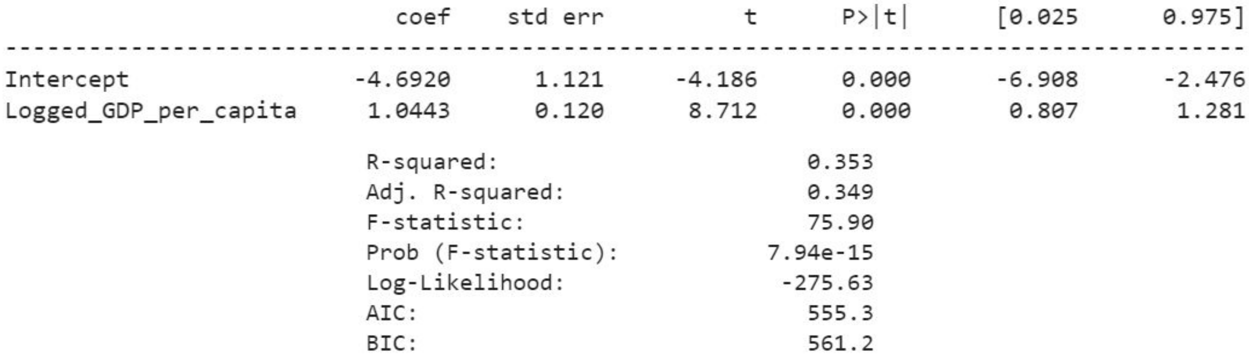
Figure 4a.,4b. 2a(left) shows max infection rate as function of ladder score. 2b(right) shows log(max infection rate) as function of ladder score.



```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.3658      0.772     -0.474      0.636      -1.892       1.160
Ladder_score   0.9750      0.138      7.084      0.000       0.703       1.247
```

```
R-squared:                       0.265
Adj. R-squared:                  0.260
F-statistic:                     50.18
Prob (F-statistic):           6.41e-11
Log-Likelihood:                -284.61
AIC:                             573.2
BIC:                             579.1
```

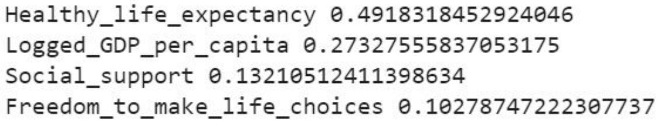
Figure 5a.,5b. Summary statistics of predicting maximum infection rate using happiness rating.



```
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept             -4.6920      1.121     -4.186      0.000      -6.908      -2.476
Logged_GDP_per_capita  1.0443      0.120      8.712      0.000       0.807       1.281
```

```
R-squared:                       0.353
Adj. R-squared:                  0.349
F-statistic:                     75.90
Prob (F-statistic):           7.94e-15
Log-Likelihood:                -275.63
AIC:                             555.3
BIC:                             561.2
```

Figure 6a.,6b. Regression model of Logged GDP per capita vs maximum infection rate.



```
Healthy_life_expectancy 0.4918318452924046
Logged_GDP_per_capita 0.27327555837053175
Social_support 0.13210512411398634
Freedom_to_make_life_choices 0.10278747222307737
```

Figure 7. Regression tree importance output using the four primary predictors of happiness to predict max infection rate.

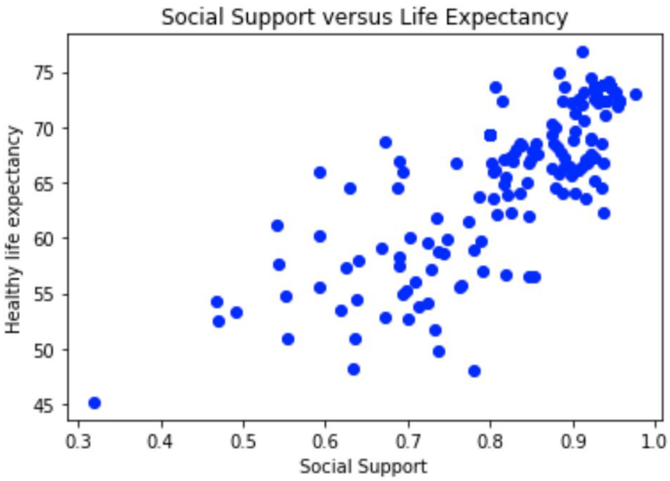


Figure 8. Scatter plot of Social Support versus life Expectancy

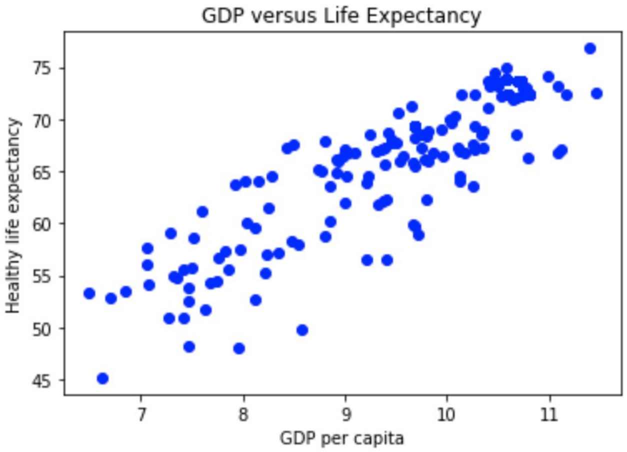


Figure 9. Scatter plot of Logged GDP per capita versus life Expectancy

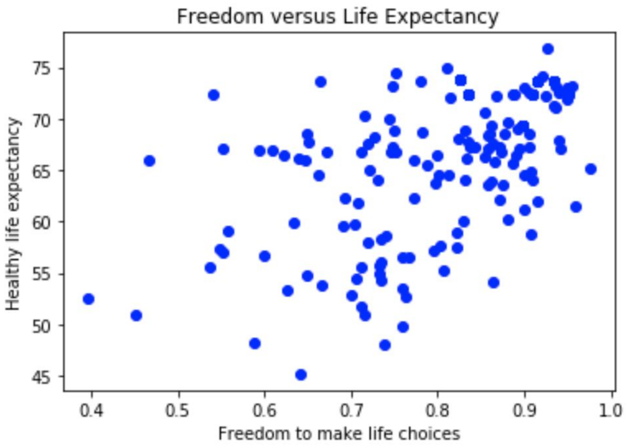


Figure 10. Scatter plot of Freedom versus life Expectancy

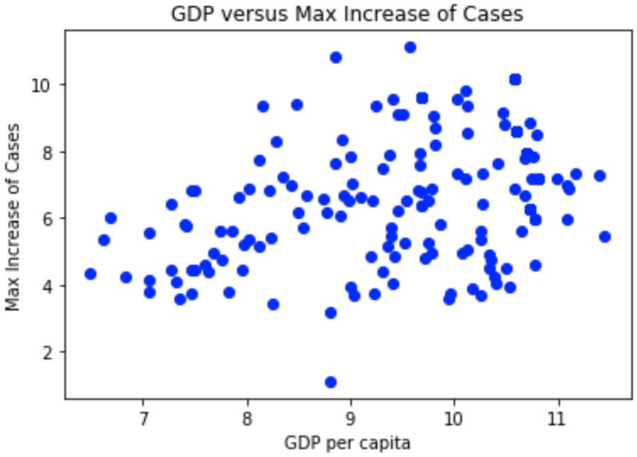


Figure 11. Scatter plot of Logged GDP per capita versus log max increase of cases

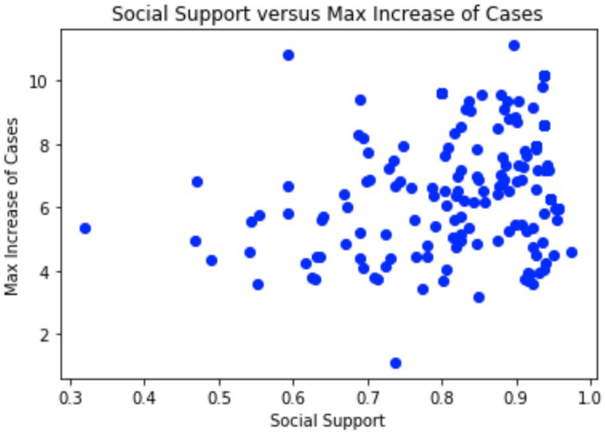


Figure 12. Scatter plot of Social Support versus log max increase of cases

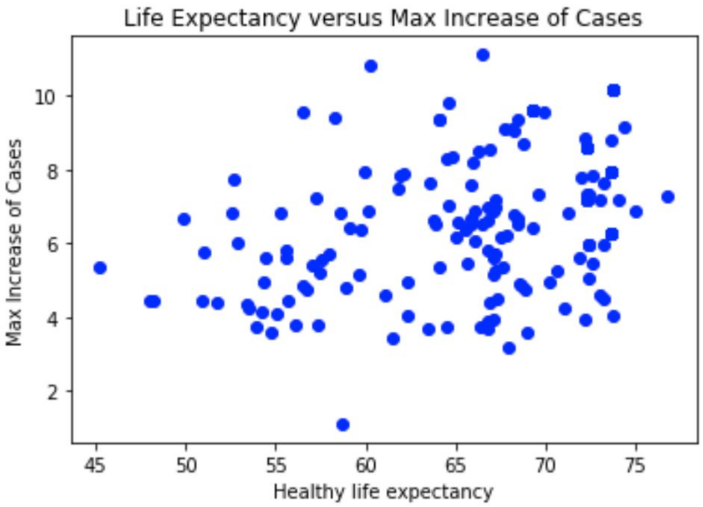


Figure 13. Scatter plot of Life Expectancy versus log max increase of cases

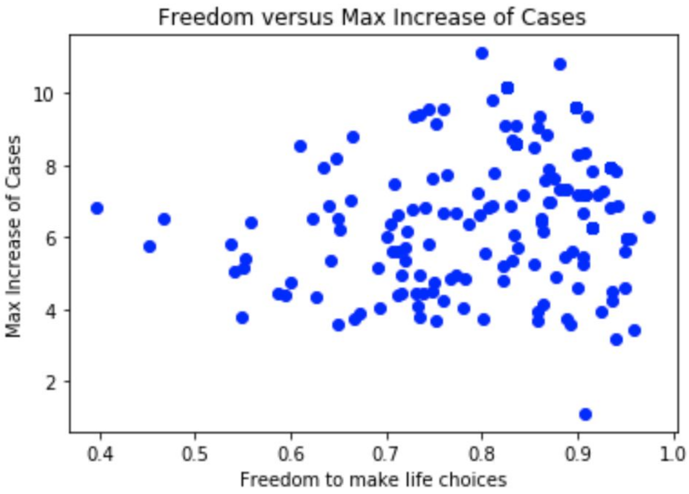
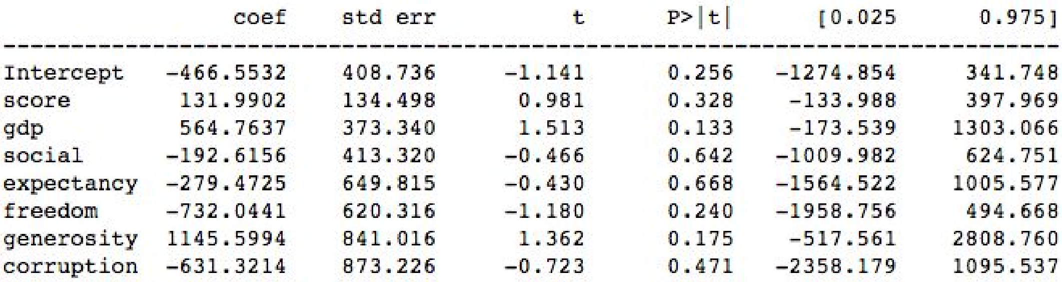
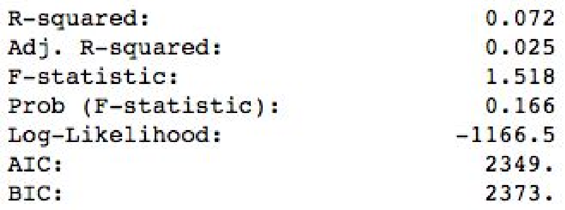


Figure 14. Scatter plot of Freedom versus log max increase of cases

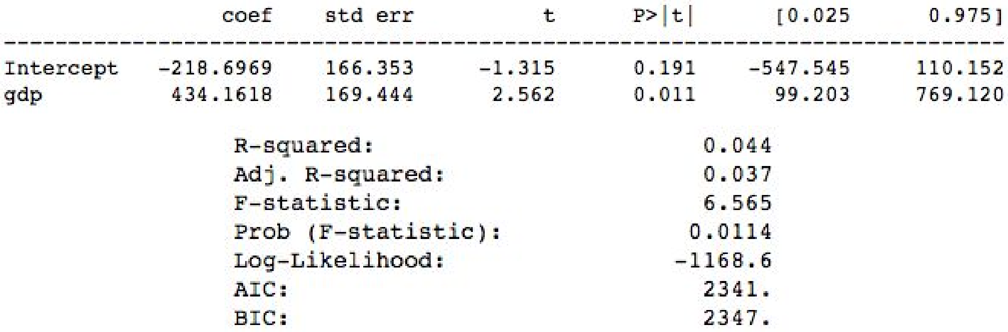


```
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept    -466.5532    408.736     -1.141      0.256   -1274.854     341.748
score         131.9902    134.498      0.981      0.328    -133.988     397.969
gdp           564.7637    373.340      1.513      0.133    -173.539    1303.066
social       -192.6156    413.320     -0.466      0.642   -1009.982     624.751
expectancy   -279.4725    649.815     -0.430      0.668   -1564.522    1005.577
freedom      -732.0441    620.316     -1.180      0.240   -1958.756     494.668
generosity   1145.5994    841.016      1.362      0.175    -517.561    2808.760
corruption   -631.3214    873.226     -0.723      0.471   -2358.179    1095.537
```



```
R-squared:                        0.072
Adj. R-squared:                   0.025
F-statistic:                      1.518
Prob (F-statistic):               0.166
Log-Likelihood:                 -1166.5
AIC:                              2349.
BIC:                              2373.
```

Figure 15. Regression model of happiness predictors against slope per country



```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -218.6969    166.353     -1.315      0.191    -547.545     110.152
gdp          434.1618    169.444      2.562      0.011      99.203     769.120
```

```
R-squared:                        0.044
Adj. R-squared:                   0.037
F-statistic:                      6.565
Prob (F-statistic):              0.0114
Log-Likelihood:                 -1168.6
AIC:                              2341.
BIC:                              2347.
```

Figure 16. Regression model of gdp predictor against slope per country

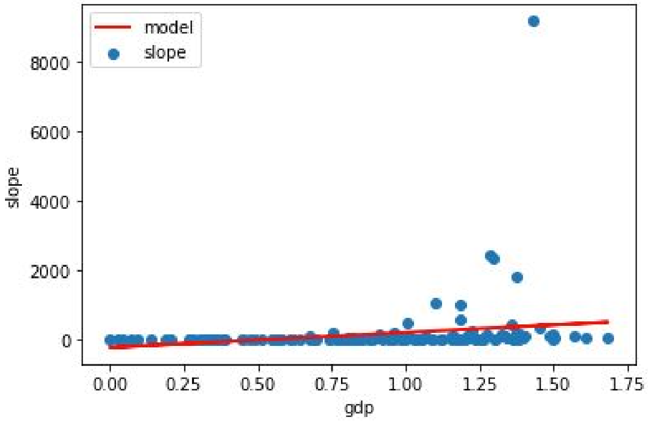
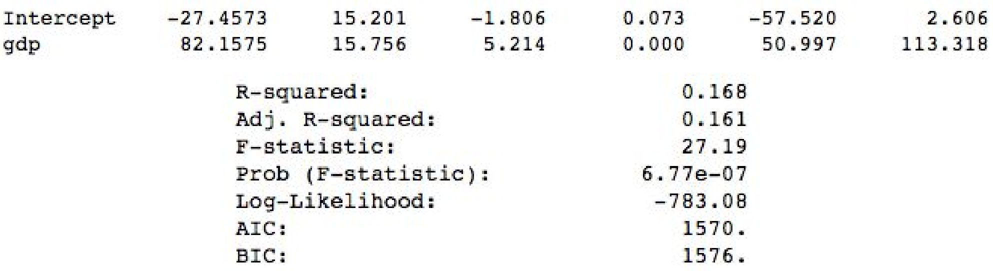


Figure 17. Scatter Plot of gdp vs slope with plotted OLS



```
Intercept     -27.4573     15.201     -1.806      0.073     -57.520       2.606
gdp            82.1575     15.756      5.214      0.000      50.997     113.318
```

```
R-squared:                        0.168
Adj. R-squared:                   0.161
F-statistic:                      27.19
Prob (F-statistic):            6.77e-07
Log-Likelihood:                 -783.08
AIC:                              1570.
BIC:                              1576.
```

Figure 18. Regression model without outliers of gdp predictor against slope per country

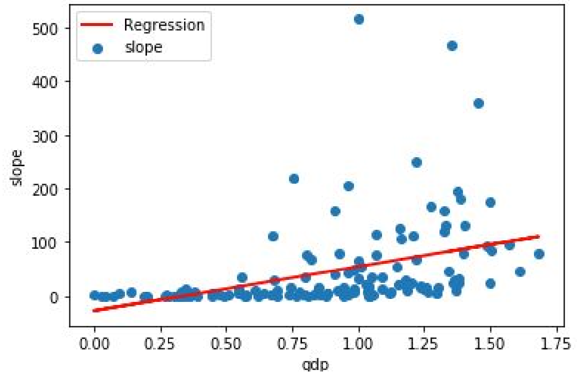
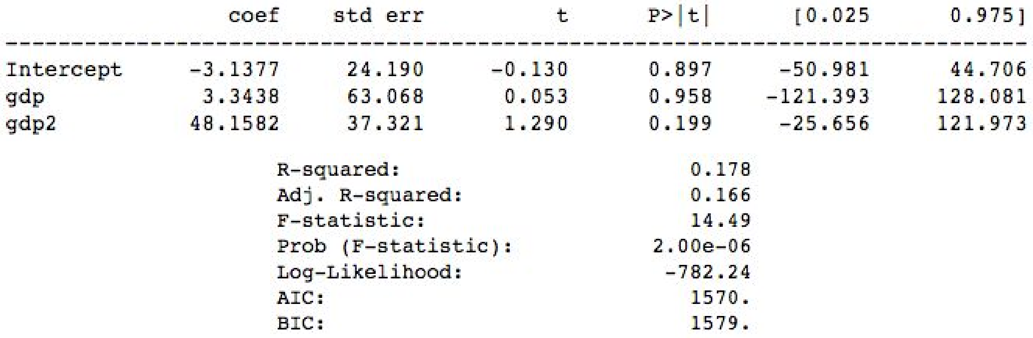


Figure 19. Scatter Plot of gdp vs slope with plotted OLS without outliers



```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.1377     24.190     -0.130      0.897     -50.981      44.706
gdp            3.3438     63.068      0.053      0.958    -121.393     128.081
gdp2          48.1582     37.321      1.290      0.199     -25.656     121.973

                     R-squared:                       0.178
                     Adj. R-squared:                  0.166
                     F-statistic:                     14.49
                     Prob (F-statistic):           2.00e-06
                     Log-Likelihood:                -782.24
                     AIC:                             1570.
                     BIC:                             1579.
```

Figure 20. Regression model without outliers of gdp and gdp2 predictor against slope per country

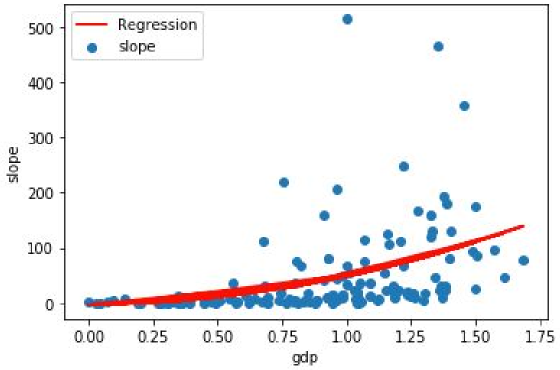


Figure 21. Scatter Plot of gdp vs slope with plotted polynomial OLS

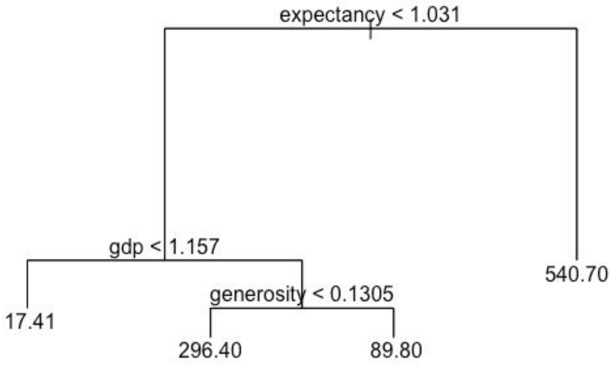
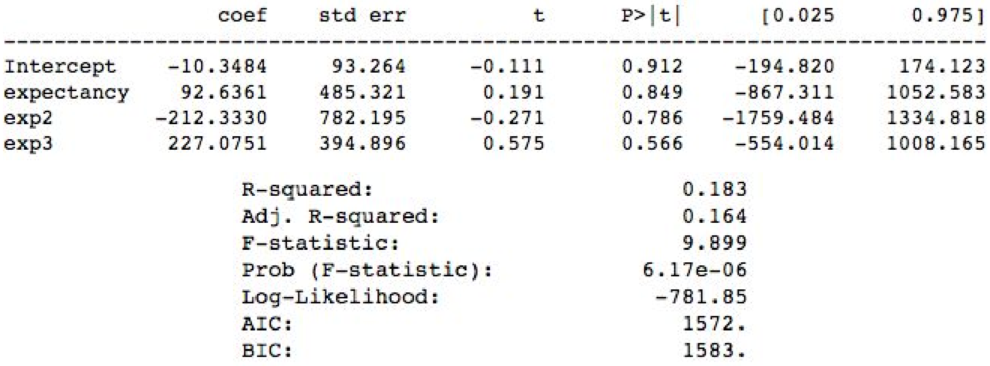


Figure 22. Pruned regression tree with all happiness predictors



```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -10.3484     93.264     -0.111      0.912    -194.820     174.123
expectancy    92.6361    485.321      0.191      0.849    -867.311    1052.583
exp2        -212.3330    782.195     -0.271      0.786   -1759.484    1334.818
exp3         227.0751    394.896      0.575      0.566    -554.014    1008.165

                   R-squared:                       0.183
                   Adj. R-squared:                  0.164
                   F-statistic:                     9.899
                   Prob (F-statistic):           6.17e-06
                   Log-Likelihood:                -781.85
                   AIC:                             1572.
                   BIC:                             1583.
```

Figure 23. Regression model without outliers of polynomial life expectancy predictor against slope per country

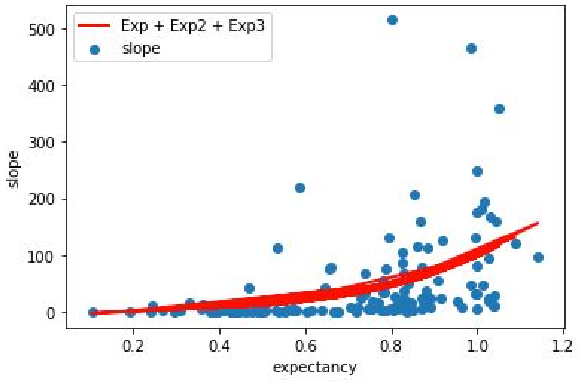


Figure 21. Scatter Plot of life expectancy vs slope with plotted polynomial OLS

Sources

Link1.: Happiness Dataseat
https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv
Link2.: Maps Dataset
https://geojson-maps.ash.ms
Link3.: Confirmed COVID-19 Cases by Country
https://coronavirus.jhu.edu/us-map