Happiness Data Analysis Report

Mason Hu

2022-12-15

INTRODUCTION

Happiness is crucial to our livelihood. In fact, it IS our livelihood. In this research, I am going to answer:

- How does a country's GDP, corruption, and Gini coefficient, location, life expectancy, and its median age correlate to happiness?
- And how can we predict the world's happiness when a pandemic (COVID-19) hits?

so that people will be better equipped with the ability to choose happiness, expect happiness, and change happiness, especially in globally influential events.

There are three papers that motivated my research: *Trust and Deaths under COVID-19* accounts for the governmental factors in my research. *Happiness in Czechia during the COVID-19 Pandemic* supports my conjecture that the location of a country is deterministic for their happiness. *Why Countries Differ Greatly in the Effects of COVID-19* analyzed the causal relation of governmental indices on COVID impact, but it's not quite my research question. My research question is the reverse, where COVID deaths is a predictor instead of a response. The three papers coupled with my motivation, imply that the significance and necessity of this research are undeniable.

METHODS

1. **Dataset obtainment and splitting**

   The dataset I use is obtained from kaggle by merging together "world-happiness-report-2021.csv" and "WHRData2021.csv". There are 146 complete observations indexed by the name of the country.

   I split my dataset into training: 76 and testing: 70 so that there are enough data for the most important training part of our model. This is done at the beginning of all the process. But validation is done at the end. This and all further model training is done in Rstudio.

2. **Preliminary EDA**

   Preliminary assessment of linear model assumptions are done in the training dataset by plotting out univariate boxplots/barplot for the response:

   - Ladder.score (measure of average happiness)

   and for the predictors:

   - Logged.GDP.per.capita

- Healthy.life.expectancy
- Perceptions.of.corruption
- Median.age
- Gini.coefficient.of.income
- COVID.19.deaths.per.100.000.population.in.2020
- Europe.or.north.america (1 if the country is in Europe or North America and 0 if not)

and pairwise scatterplots with the responses are plotted to discover the linear relationship.

3. **Initial model fitting and formal assumption checking**

   I fitted an initial/full linear model to the seven predictors described above and plotted out the residuals against the fitted values and the seven predictors. I also plot out the qqplot of errors to assess normality. This way I can easily observe any violations to the four assumptions of linear regression:

   (a) uncorrelated errors

   (b) normality of errors

   (c) linearity

   (d) constant variance

4. **Additional condition checking model diagnostics**

   After (If) I observe assumption violations in the residual plots, I check the 2 additional conditions to see if the residual plot tells us exactly what and how to remedy the assumptions.

   Condition 1: I plotted the true versus fitted values and check if they are resemble the identity function relationship.

   Condition 2: I plotted the pairwise scatterplots for the 7 predictors and observe if there are patterns other than linearity between them.

   If these conditions hold, as remedy, I either apply variance stabilizing transformations to violation of constant variance. Or I apply box-cox transformations to violation of linearity or normality of errors. I will denote the transformed model as the new model.

   However, if these conditions fail to hold, there is no guarantee of the validity of our transformation and we will have to note that as a limitation.

5. **Partial F tests and VIF**

   I do a partial F test on the variables that does not indicate a significant linear relationship as specifies by the T-test in the linear model summary and remove the variables to get a reduced model if the anova partial F test DOES NOT reject the null hypothesis. If it does, I retrieve some of the variables and see if the new reduced model rejects the null hypothesis in the new partial F test.

   I also do VIF tests to see if there's variance inflating predictors that are multicollinear, and can try further reducing the model.

   After reducing the model I will ONCE AGAIN check the model assumptions by plotting out the new relevant plots.

6. **Checking for outliers, influential points, and leverages**

After reducing the model, I check for leverages, influential points, and outliers. I check the leverages using the Hat matrix, outliers using the standardized residuals, and influential points using the following three methods: Cooks distance (influence on the regression line as a whole), DFFITS (influence on its own fitted value), and DFBETAS (influence on the coefficients from β_0 to β_n).

7. **Goodness of model**

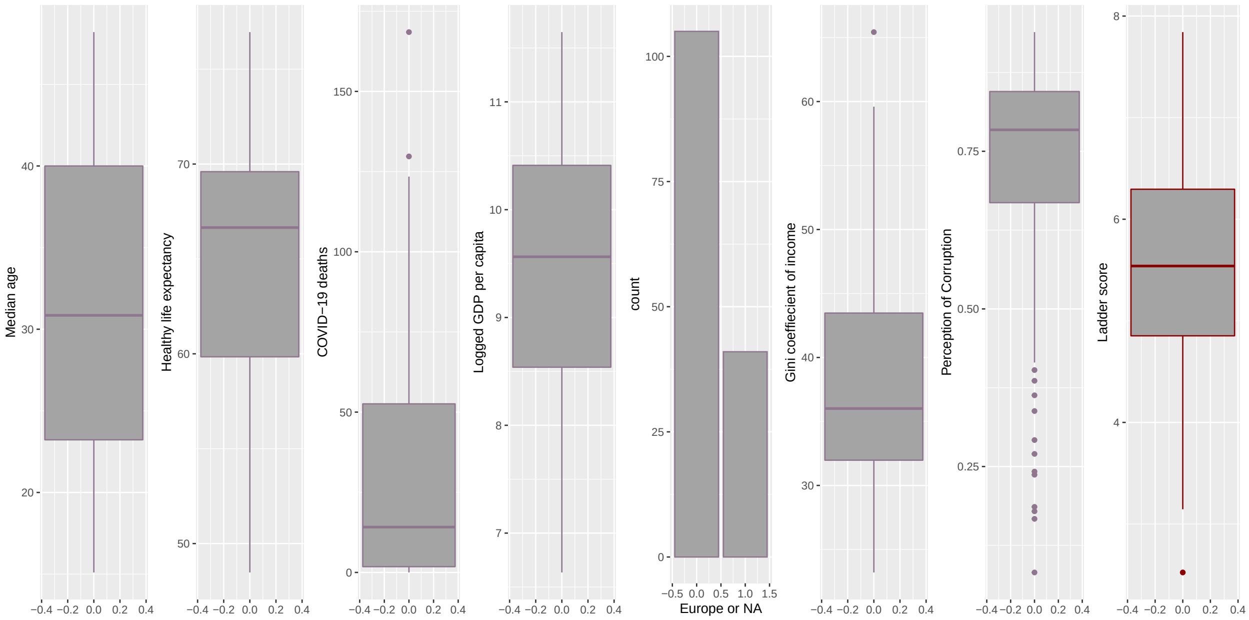
   As the ultimate model selection process, I use adjusted coefficient of determination R^2, AIC, AIC_C, and BIC on six of the previous models and summarize their results. The model with the highest in R^2, lowest in AIC, AIC_C, and BIC will be the best model. And I will if the four different tests favor different models, I will favor my variable of interest–Log.death and Europe.or.north.america–as a tie breaker.

8. **Final model validation and assumption check**

   Finally I choose the best one/two model(s) in the previous part and perform model validation by fitting the same model to the testing dataset and compare their characteristics side-by-side. I will compare model coefficients and various characteristics like R^2 and VIF and influential points. After this step, I compare the validity based on the differences in training and testing. I will then check the model assumptions once again and I can finally declare my final model.

RESULTS

Variable visualizations



Three things to highlight:

- COVID-19 deaths is extremely right-skewed. This is my potential subject of future transformation.
- Perception of corruption is relatively right skewed. Not necessarily a subject of transformation but can possibly violate assumptions.

- Europe or North America is also not evenly distributed. This is inevitable since there are less countries in the northern hemisphere than the southern hemisphere.

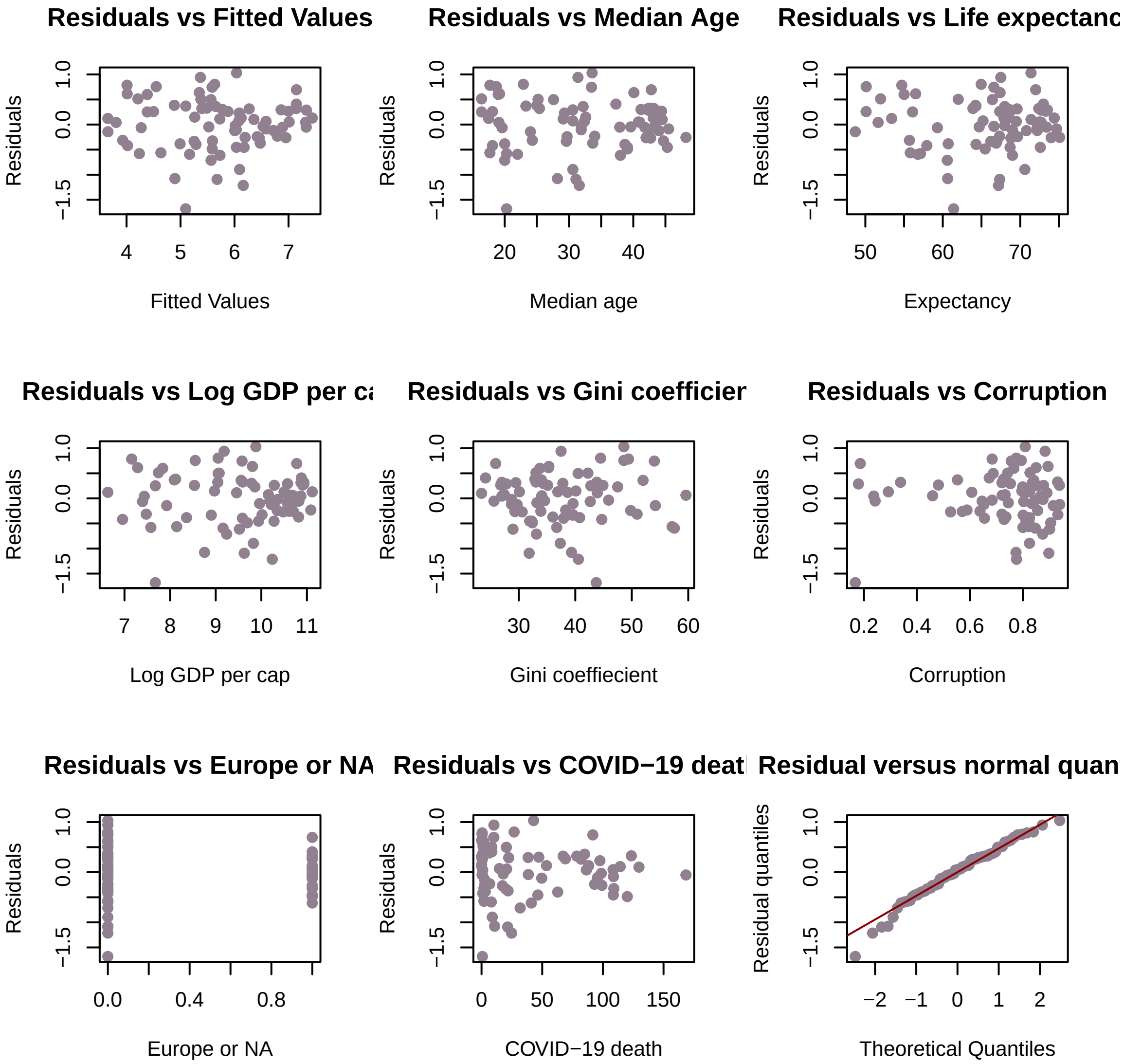
(Preliminary randomness check of training and testing datasets is provided in appendix.)

I fitted the initial model

$\hat{y}_i = -2.815335 - 0.050050M_i + 0.061812H_i + 0.693651L_i - 0.001176G_i - 0.844528P_i + 0.365679E_i + 0.000217C_i$

where M,H,L,G,P,E,C are the intials of the seven predictors.

I checked assumptions

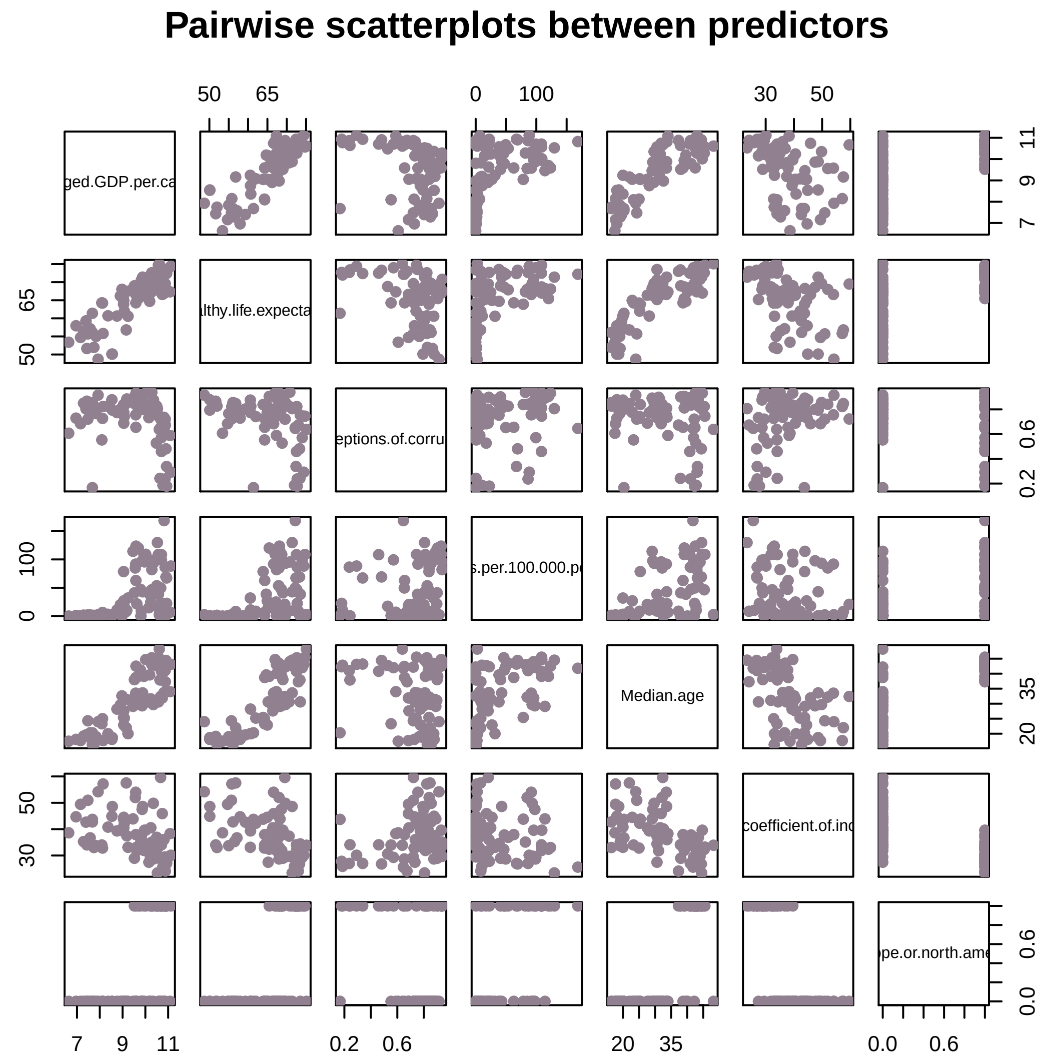


There are no obvious patterns in the first six residual plots.

- For the residuals vs Europe or North America plot, there also seems to be violation of constant variance. This inevitability was discussed in my EDA, but I will still address it in the limitations.

- For the residuals vs COVID-19 death plot, there seems to be a mild fanning pattern. This is indication of violation of constant variance. I will proceed to perform remedies. However, I first need to check two additional conditions.

Additional conditions checking



The left pairwise scatterplot checks our condition 2. There is no obvious patterns besides linear relationship. The appendix plot (Fitted values versus true response) verifies condition 1: fitted values and true values are roughly is related via identity function.

So our residual plots are helpful in determining the violated assumptions. According to the above discussion, we should perform a variance stabilizing transformation.

I transformed COVID-19 deaths

By observing the patterns and by result from a power transformation test:

Likelihood ratio test that transformation parameter is equal to 0 (log transformation)

I conclude I should take the natural logarithm of COVID-19 death and store it in Log.death. The new transformed linear model satisfy the assumptions. See appendix(after log transform).

I reduced the model

I tried removing four relatively insignificant predictors(M, G, C, and my variable of interest: Logged COVID deaths)

Pr(>F)=0.001869

which rejects the null hypothesis that all removed coefficients should be zero, So at least one of the variables should not have been removed.

Then I tried removing only two insignificant predictors(G, C) and performed partial F test.

Pr(>F)=0.263

failing to reject the null hypothesis that all removed coefficients are zero, meaning I can safely remove both Gini coefficient and Corruption from my model.

I chose the best model

Model	Adjusted R^2	AIC	BIC
Removed(M, G, P)	0.7309242	-79.0782599	-61.0938599
Removed(L, G, P)	0.6369921	-56.3214722	-38.3370722
Removed(G, P)	0.7722444	-90.8270123	-70.5118789
Transformed(Unreduced)	0.7745771	-89.8124891	-64.835889
First/Full	0.7706106	-88.4868303	-63.5102302

According to our adjusted R^2, AIC and BIC, all models have their lengths are shortcomings. In light of my research question, prediction for happiness in future pandemic is more important than explaining the correlation. Therefore, I choose to include and Median age and Log.death and choose the model with the highest R^2, but also higher AIC , BIC.

I validated the final model

Characteristics	Removed(G, P) train	Removed(G, P) test
no. Cooks Distance	0	0
no. DFFITS	2	4
Largest VIF	7.159883	6.392747
Violation	None*	None*
Intercept	-2.8153351 ± 0.9(*)	-0.5079115± 1.184
Healthy.life.expectancy	-0.0500499 ± 0.02(*)	0.0713507± 0.027(*)
Logged.GDP.per.capita	0.0618125 ± 0.117(*)	0.1584616± 0.167
Europe.or.north.america	0.6936512 ± 0.209(*)	1.0205957± 0.243(*)
Median.age	-0.0011759 ± 0.017(*)	-0.0128812± 0.02
Log.death	-0.8445282 ± 0.036(.)	-0.0101405± 0.034

*model assumptions are verified in appendix (process same as the first model)

This validation table exhibits similar traits characteristics when it comes to problematic points or variance inflation factor. However, the model coefficients have drastically different results. This is mainly because this is a small dataset and the training and testing data could have potentially altered the results drastically.

DISCUSSIONS

The final model (assumption checked) is as follows (C for COVID death(logged))

$$\hat{y}_i = -0.50791152 + 0.07135067H_i + 0.15846156L_i + 1.02059571E_i - 0.01288123M_i - 0.01014049C_i$$

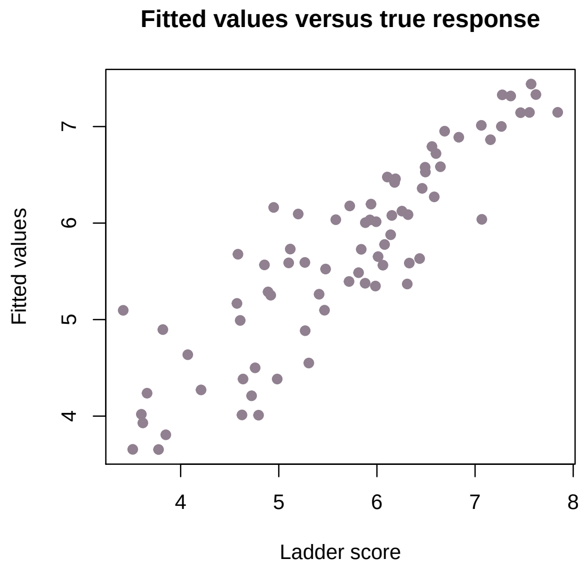
This shows that a country's average happiness is positively correlated with Life expectancy, GDP, and negatively correlated with Age and death in pandemics. Also you will be happier if you live in Europe.

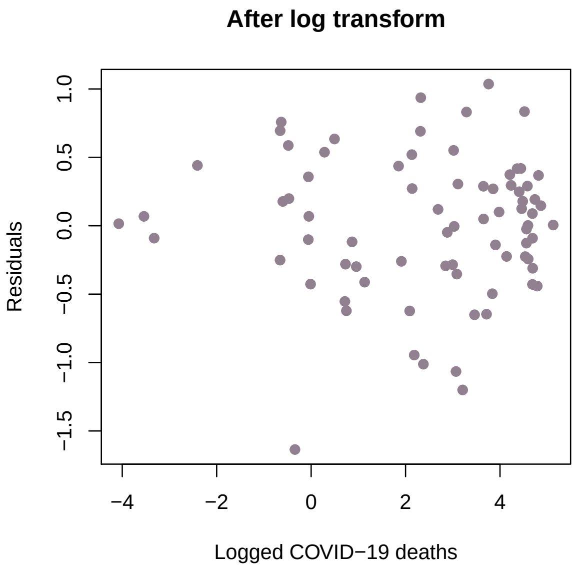
This model also answers my second question of predicting happiness in future pandemics: for every unit increase in the logarithm of the country's death count, everybody in that country will be 0.01 less happy on average.

However, there are some limitations:

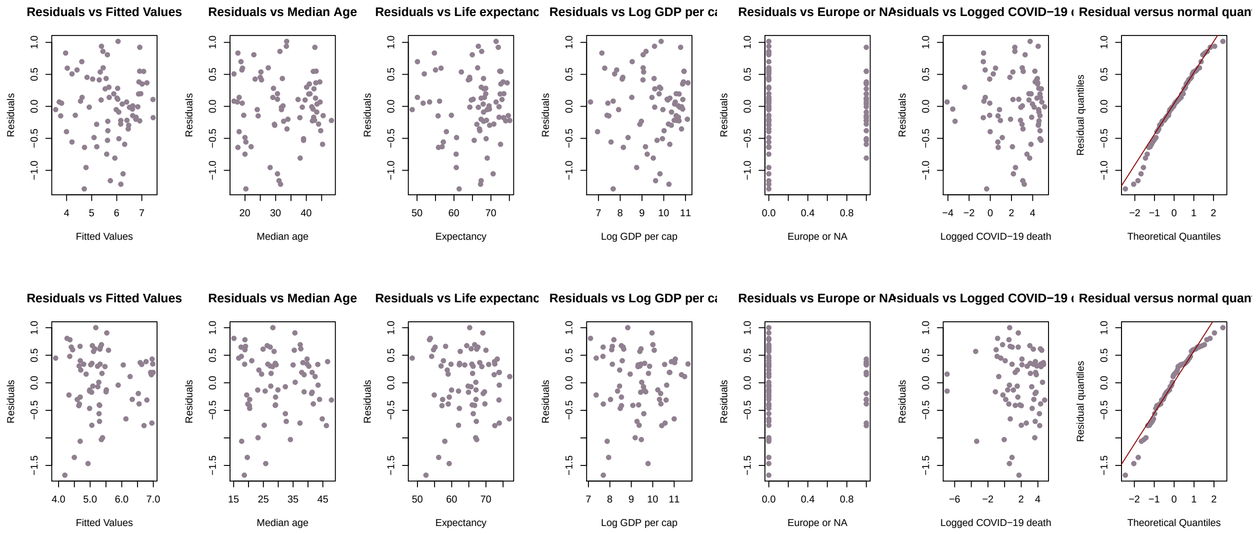
1. The main limitation of this analysis is that this dataset is too small, but it is unchangeable since there are only almost 200 countries on earth. This makes our validation process hard and unreliable.
2. the disproportionate distribution of countries in Europe or North America versus not is also inevitable. But it might not influence our model too much since the violations is not terrible.
3. there is a lot of (10) leverage point (countries like Cambodia that are uniques and extreme in the collected data) meaning these 10 countries could be dominating the whole model in the future as they become more extreme.

APPENDIX





Final model assumption check (training upper, testing lower)



Mean and standard error variables of randomized training and testing datasets

Variable	mean (s.d.) in training	mean (s.d.) in test
Ladder.score	5.386 (5.691)	1.012 (1.125)
Logged.GDP.per.capita	9.342 (9.503)	1.114 (1.214)
Healthy.life.expectancy	64.5 (65.55)	6.557 (6.805)
Perceptions.of.corruption	0.728 (0.727)	0.17 (0.191)
COVID-19 deaths	24.478 (40.832)	32.027 (43.88)
Median.age	30.304 (32.233)	9.077 (9.533)
Gini.coefficient.of.income	37.228 (37.739)	8.079 (37.739)

Variable	mean (s.d.) in training	mean (s.d.) in test
Europe.or.north.america	0.214 (0.342)	0.413 (0.478)

REFERENCES

by John F. Helliwell from UBC. World Happiness, Trust and Deaths under COVID-19 https://www.researchgate.net/profile/Shun-Wang-31/publication/350511691_World_Happiness_Trust_and_Deaths_under_COVID-19/links/6063d19b299bf173677dc90c/World-Happiness-Trust-and-Deaths-under-COVID-19.pdf

by František Petrovič: Happiness in Czechia during the COVID-19 Pandemic, https://www.mdpi.com/2071-1050/13/19/10826/htm

Why Countries Differ Greatly in the Effects of COVID-19 https://www.researchgate.net/profile/Victor-Dementiev/publication/355224679_Why_Countries_Differ_Greatly_in_the_Effects_of_COVID-19/links/6176cd86a767a03c14b0ee7d/Why-Countries-Differ-Greatly-in-the-Effects-of-COVID-19.pdf?_sg%5B0%5D=started_experiment_milestone&origin=journalDetail

Datasets: Kaggle “world happiness report 2021” and Kaggle “WHRdata2021”