MONASH
BUSINESS
SCHOOL

# Exploratory analysing on World Happiness Report

**Weihao Li**
28723740

**Jinhao Luo**
29012449

**Xitong He**
29026342

Report for
ETC5513 Assignment 4

**8 June 2020**

**Department of Econometrics & Business Statistics**

☎ (03) 9905 2478
✉ BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

# 1 Introduction

Helliwell, Layard, and Sachs (2019) has indicated that there is an inner connection between government and happiness, it means the jobs of government would influence the happiness of citizens, while the people feeling of happiness could guide them to decide which kind of government to support. In addition, Helliwell, Layard, and Sachs (2019) has also explained that the living quality of people could be measured by the happiness of each country. Therefore, this report has utilized the happiness score of each country to analyse the changing of well-being according to countries, regions, and the world. Meanwhile, the happiness scores have been allocated into the world map to explain the distribution. In addition, the report has also analysed the importance of each modelling variables, as well as the relationships between happiness scores and other factors. During the study, this report has found the trends and distribution of happiness scores. Meanwhile, this report has explored the ———-what do you find(Karen)——. ——what do you find(weihao's part)——. However, the number of countries observed was not consistent across the datasets, which might generate errors when calculating the average level and analyse the trends over the years. ——- other limitations———.

# 2 Q1.The evolution of world happiness

Helliwell, Layard, and Sachs (2019) has indicated that the world has become a rapidly changing place, and such the fastest changing might influence many aspects of different countries on their people. Therefore, in order to explore how satisfied people are with their countries over years, this report has used the happiness score to consider it.

## 2.1 The trends in happiness 2015-2019

In this part, the top ten countries in 2015 have been extracted to consider how their happiness changing during the five years. According to table 1, it has shown that Switzerland

has occupied first place in 2015, which was around 7.59 points. While Iceland was around 0.02 standing in second place. In addition, Australia was the tenth country, which was around 7.28. Based on figure 1, which has shown that Finland was the country with a significant increase over

five years. Finland has increased from just over 7.4 to around 7.77 and has become the happiest country in 2019. Meanwhile, Denmark, Norway, Netherlands, and New Zealand have also experienced an increase in both scores and places compared with 2015. Although the happiness score of Sweden has decreased in the same period, however, the rank of Sweden has improved in 2019. The rest of the four countries have experienced a decrease in both ranks and scores over five years. The reasons for the changing in score and ranks might because of the social welfare, country economic environment, and government policies. Such aspects have all related to each nation a within country so that impacts the satisfaction level.

**Table 1:** *Top10 countries in 2015*

| Country | 2015 |
|---|---|
| Switzerland | 7.587 |
| Iceland | 7.561 |
| Denmark | 7.527 |
| Norway | 7.522 |
| Canada | 7.427 |
| Finland | 7.406 |
| Netherlands | 7.378 |
| Sweden | 7.364 |
| New Zealand | 7.286 |
| Australia | 7.284 |

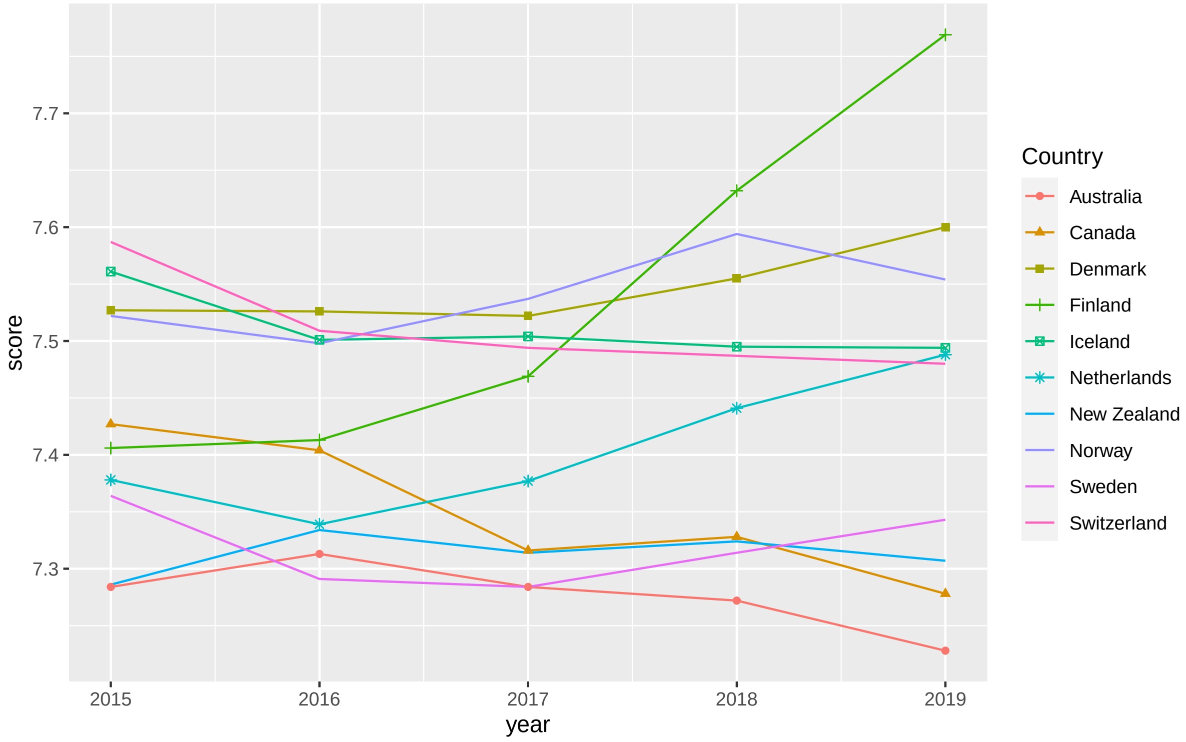


**Figure 1:** *The trends in happiness scores from 2015-2019 (based on the top10 countries in 2015)*

## 2.2 The dynamics of the world and regional happiness 2015-2019

Besides, this report has also considered the changing of regional happiness. Firstly, this report has explored the percentage of the number of countries which happiness scores above or below the world average level in 2015 and 2019 to consider the overall condition. According to figure 2, the percentage of the above-average level has accounted for

46.84% in 2015, which was 74 countries. While there were still 84 countries that did not reach the average level. In 2019, the percentage of the above-average level has increased to 49.36%, which was 77 countries. And the percentage of the below-average level has decreased to 50.64%. In general, more and more people have satisfied their countries. To study the deeper level, figure 3 has explained the number of

countries that above-average level of each region in 2015. According to that figure, Western Europe, and the region of Latin America and the Caribbean have the highest number, which was 19 countries respectively. While the region of Sub-Saharan Africa has only one country that above the average level in 2015. Furthermore, based on figure 4, western Europe has still stood

at the dominant place, which was 19 countries. While Latin America and the Caribbean have lost a country, which totals 18 countries above average level in 2019. Central and Northern Europe

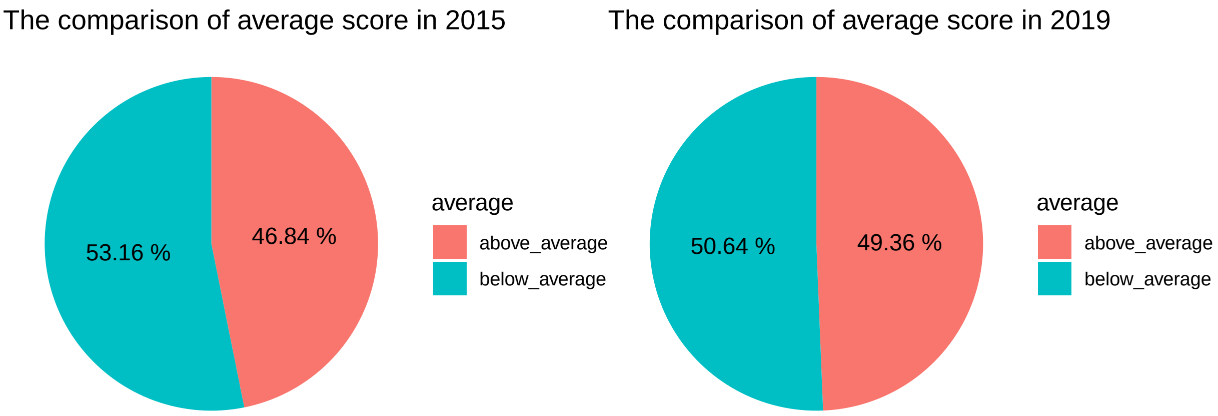


**Figure 2:** *The percentage of the number of countries above or below average happiness scores in 2015 and 2019*

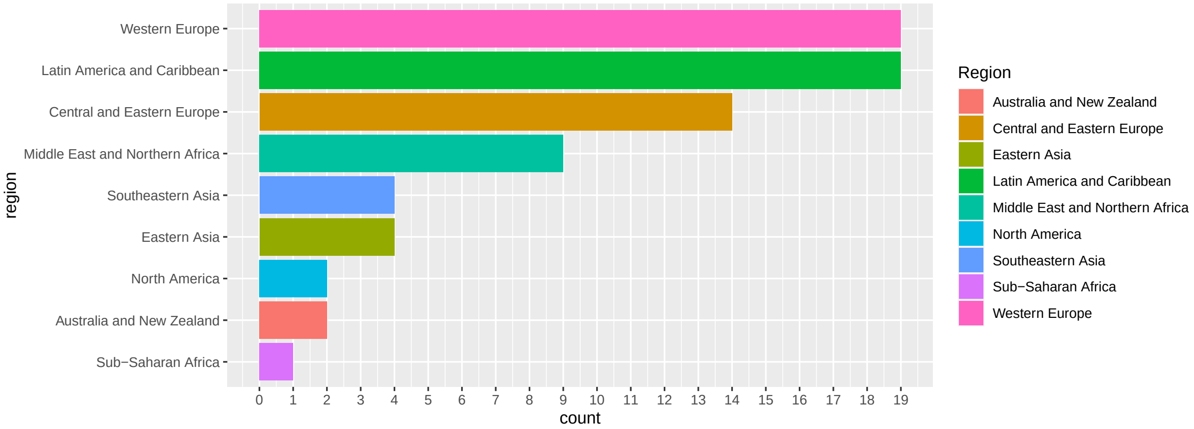


**Figure 3:** *The number of above average happiness scores across regions in 2015*

have the same number as Latin America and the Caribbean in 2019. Meanwhile, Southern Asia has appeared in the figure in 2019, which was one country above the average level. In addition, North America, the region of Australia and New Zealand, and Sub-Saharan Africa have remained the same level compared with 2015. In general, according to the figure of 2015 and 2019, the countries that happiness scores above average level have concentrated on the region of Europe, and Latin America and Caribbean. While the other regions have less number of countries compared with them.

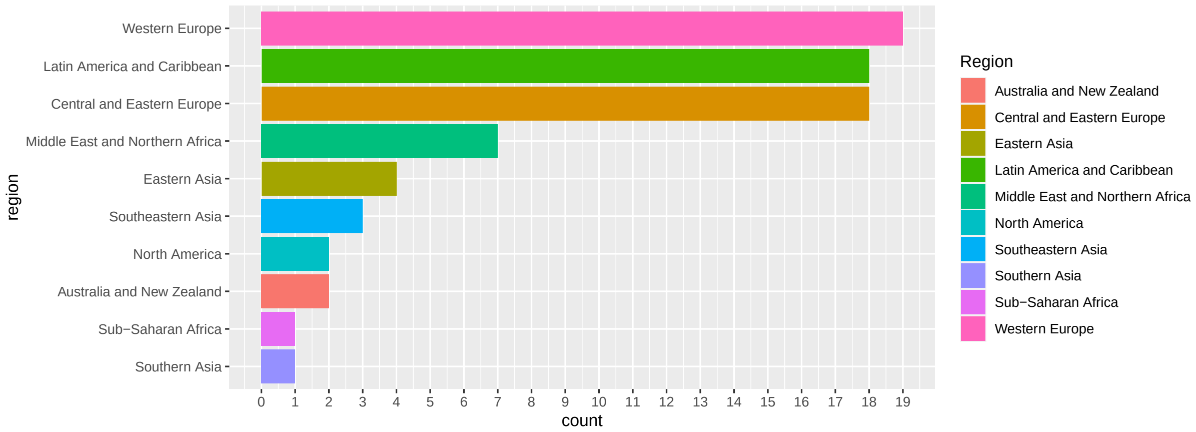


**Figure 4:** *The number of above average happiness scores across regions in 2019*

# 3 Question 3 : Which variables are useful for modelling happiness scores and how the relationship between happiness scores and other factors?

To explore the factors that could be contributing to the happiness score differences between each year,a **linear regression model** is established by testing the relationship between predictor and response variables.From the **LM** result, we can conclude the estimated closed form of happiness score:

$$\begin{aligned}\text{Happiness score} = \text{GDP per Capital } + \text{ Family } + \text{ Life expectancy } + \text{ Freedom} + \\ \text{Government Corruption } + \text{ Generosity } + \text{Dystopia Residual}\end{aligned}$$

The **Table** 2 is shown that there is strong evidence that the happiness score is and only is a simple average of these factors. In addition, all the R squared are close to 1 from 2015 to 2017 in the **Table** 3 ,indicating that these factors are standardized and no longer to use as building up a model. Besides, for 2018 and 2019 data, the organization provides another set of variables such as social support has been changed from family factor, which are also highly correlated with the happiness score, but we have no method to understand what those values represent since they have been transformed in an unknown manner.

Instead of using the factors in the given dataset, we join other datasets to analysis the happiness score, including GDP, CPI, population and area by different countries. There are another three datasets such as GDP, CPI and population data which are downloaded from The World Bank. And area of each country is computed using the simple feature polygons with the package `rnaturalearth` (South, 2017). In addition, we pick up thee variables such as GDP,CPI and population between 2015-2019 to put them into new happiness score dataset. After matching the area of each country and drop out the missing value, the new dataset with happiness score for analyzing the relative variable importance is established in the **Table** 4.

**Table 2:** *Linear regression model for happiness scores in 2015*

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.0000640 | 0.0001246 | 5.142216e-01 | 0.6078538 |
| Economy..GDP.per.Capita. | 1.0001014 | 0.0001129 | 8.855750e+03 | 0.0000000 |
| Family | 0.9999703 | 0.0001153 | 8.675863e+03 | 0.0000000 |
| Health..Life.Expectancy. | 0.9998826 | 0.0001619 | 6.175103e+03 | 0.0000000 |
| Freedom | 0.9996953 | 0.0001976 | 5.059468e+03 | 0.0000000 |
| Trust..Government.Corruption. | 0.9999191 | 0.0002237 | 4.470866e+03 | 0.0000000 |
| Generosity | 1.0000613 | 0.0002018 | 4.956272e+03 | 0.0000000 |
| Dystopia.Residual | 1.0000304 | 0.0000417 | 2.400390e+04 | 0.0000000 |

**Table 3:** *R Squared in 2015-2017*

| 2015 | 2016 | 2017 |
|---|---|---|
| 0.9999999 | 0.9999999 | 0.9991408 |

**Table 4:** *A clean dataset with Happiness Score*

| area | name | year | cpi | gdp | population | Happiness Score |
|---|---|---|---|---|---|---|
| 6.522701e+11 | Afghanistan | 2015 | -0.6617092 | 19907111419 | 34413603 | 3.575 |
| 6.522701e+11 | Afghanistan | 2016 | 4.3838920 | 19362642267 | 35383128 | 3.360 |
| 6.522701e+11 | Afghanistan | 2017 | 4.9759515 | 20191764940 | 36296400 | 3.794 |
| 1.245464e+12 | Angola | 2015 | -21.5316935 | 116193649124 | 27884381 | 4.033 |
| 1.245464e+12 | Angola | 2016 | 32.3777341 | 101123851090 | 28842484 | 3.866 |
| 1.245464e+12 | Angola | 2017 | 31.6916861 | 122123822334 | 29816748 | 3.795 |

### 3.1 Variable importance

For comparing the relative importance with each predictors, the **lmg** approach with package `relaipmo` (Grömping, 2006) is used for evaluating the variable importance in modelling happiness score. By using **lmg** approach to calculate the relative contribution of each predictor to the R squared with the consideration of the sequence of predictors appearing in the model.

Firstly, we built up a **LM** model for the new happiness score dataset. By eliminating the heteroskedasticity in the residual and interpreting the percentage marginal effect of each variable, using logarithmic method for setting up a linear regression model is an appropriate way. Therefore, a **LM** model is shown :

$$\log(\text{Happiness Score}) = \text{CPI} + \log(\text{GDP}) + \log(\text{Area}) + \log(\text{Population})$$

From **Figure 5**,we can see that the total proportion of variance explained by the model with all four predictors is around 72.88%. For the four predictors, GDP with the highest value in four predictors contributed to around 65% for the happiness scores,indicating that GDP factor is the most important for happiness score and it is more likely to affect the change of happiness scores when the GDP variable change (Grömping, 2015). On the other hand,by using **variance inflation factor** (VIF) measure to evaluate the impact of the correlation among the explanatory variables. In the **Table 5**, the VIF value in each predictor is not too high but for the population variable is the highest one around at 3, indicating that there is a slight correlation of population with other variables in the model (Daoud, 2017).

### 3.2 Scatter matrix

To analyze the relationship between different predictors, we use scatter matrix which can be visualized easily how much one variable is affected by another or the relationship between them. The **Figure 6** compares all the predictors in the model and represents the relationship between these variables. If there is a strong positive linear correlation between two factors, we can say that if one factor is important in evaluating a country's overall happiness, it is likely that the other factor is important as well. Based on the plots, it seems that the importance of GDP and area are strongly correlated, as well as population and area. Therefore,GDP is the most important factors for evaluating the happiness scores as it has the highest positive correlation with happiness score,but also for the developed countries in most of western area with highest GDP, the happiness scores will be higher than the development countries with lower GDP.

In conclusion, GDP is the most influential factors for assessing the happiness scores but the area variable represents a higher correlation with GDP. Thus, Even in very poor countries, GDP comparisons seem to influence the happiness score mostly; however, it may look more likely to happen in rich-country phenomenon (Clark & Senik, 2011).

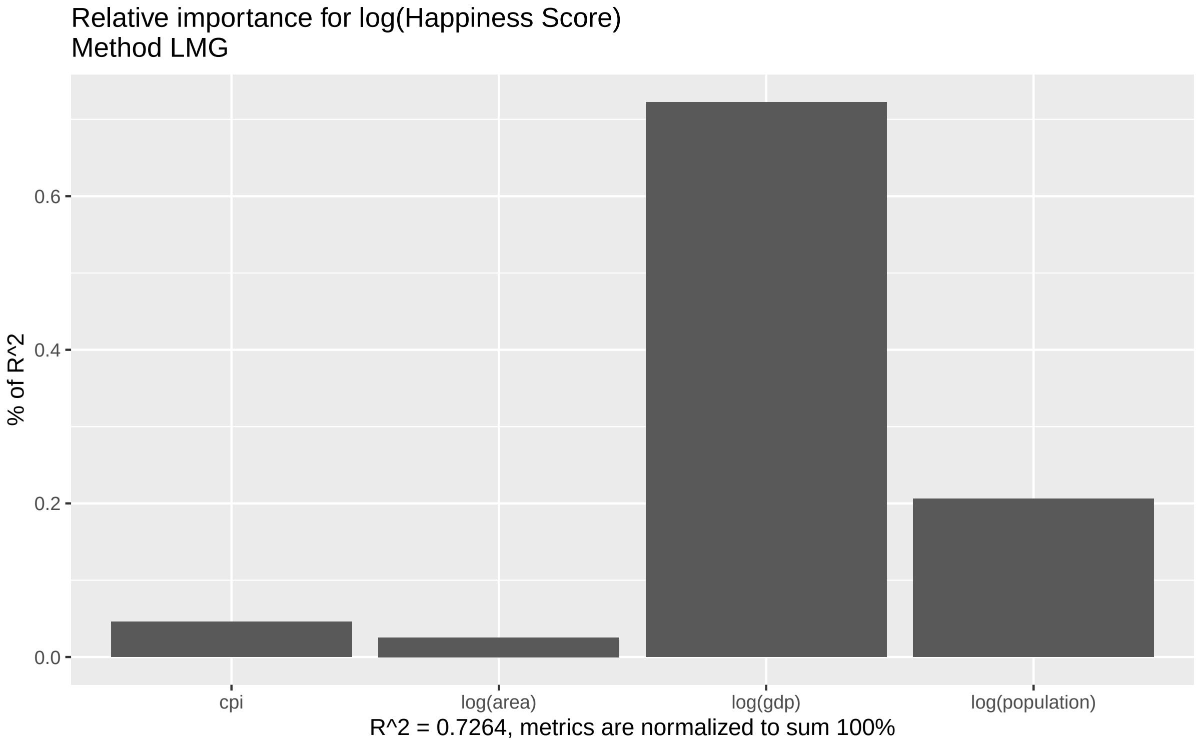


**Figure 5:** *Variable importance in happiness score*

**Table 5:** *VIF procedures for checking multicollinearity with each predictors*

| Var1 | Freq |
|---|---:|
| log(area) | 1.985872 |
| cpi | 1.137452 |
| log(gdp) | 1.903231 |
| log(population) | 2.742265 |

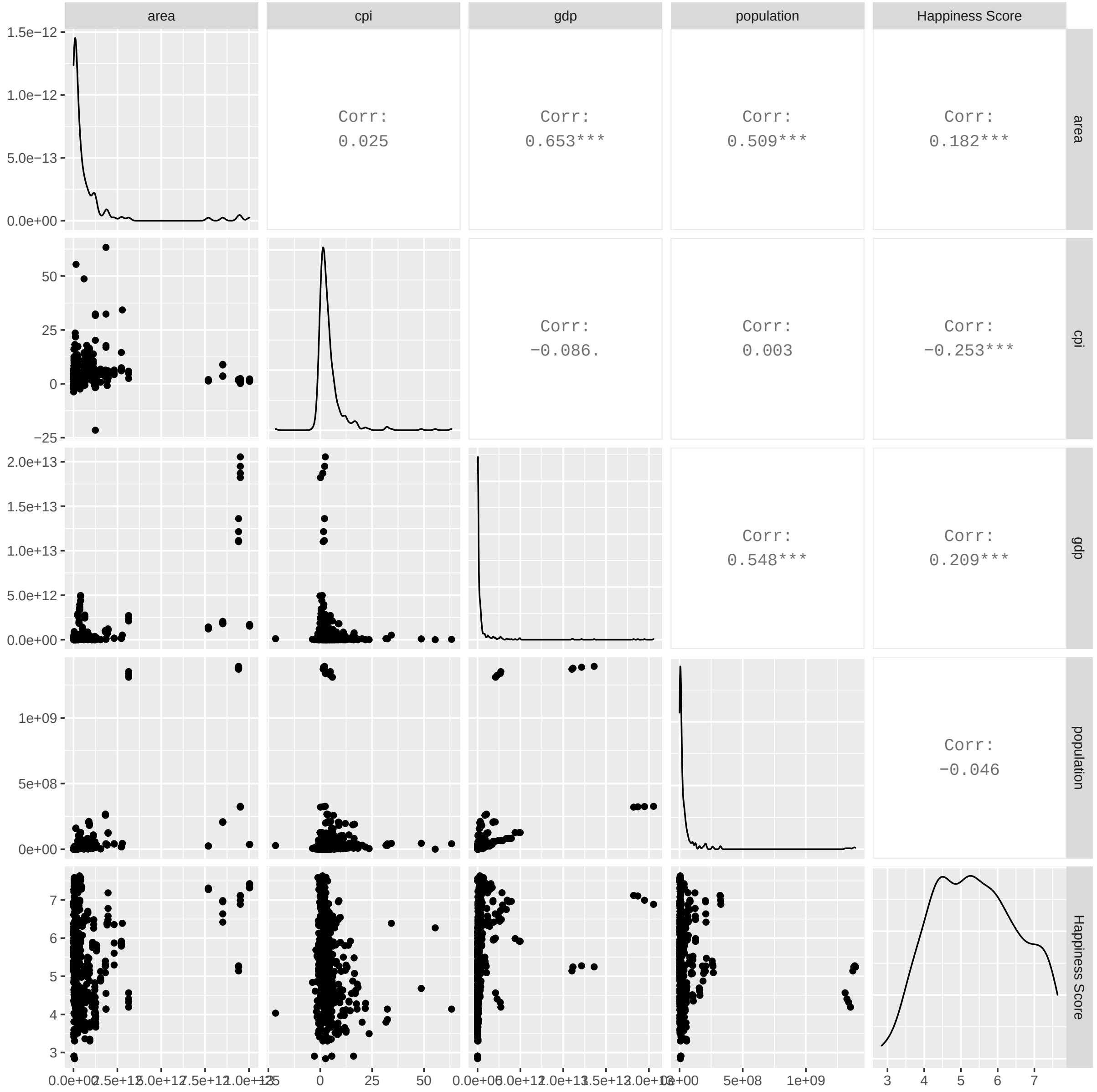


**Figure 6:** *Happiness Score Scatter Matrix*

## 4 Q4. What is the relationship between predictors and happiness score in linear regression analysis?

We use variables mentioned in Q3 as predictors, the regression result is shown in **Table** 6. According to the result, we can write down the formula for this model:

$$\begin{aligned}\log(\text{Happiness Score}) = 0.3208 + 0.0005\ \text{CPI} + 0.1285\ \log(\text{GDP})\\ - 0.1218\ \log(\text{Population}) + 0.0056\ \log(\text{Area})\end{aligned}$$

Notice that only the intercept, log of GDP and the log of the population are significant with $\alpha = 5\%$. Since we take logarithm on both response variable and regressors except CPI given it is a percentage index, we can interpret them as the elasticity of happiness score of GDP and the elasticity of happiness score of the population. This can be proved by:

$$\frac{\partial \log(\text{Happiness Score})}{\partial \log(\text{GDP})} = 0.1285 \tag{1}$$

From (1) we can say, when other variables remain constant

$$\Delta\ \log(\text{Happiness Score}) = 0.1285\Delta\ \log(\text{GDP}) \tag{2}$$

By using the linear approximation of log function, we can know

**Table 6:** *Linear regression model for happiness scores. The $R^2$ of this model is 0.7264.*

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.2927 | 0.0914 | 3.2020 | 0.0015 |
| cpi | 0.0005 | 0.0009 | 0.5327 | 0.5945 |
| log_gdp | 0.1285 | 0.0038 | 33.9384 | 0.0000 |
| log_population | -0.1218 | 0.0057 | -21.2290 | 0.0000 |
| log_area | 0.0056 | 0.0044 | 1.2847 | 0.1995 |

$$\begin{aligned} log(x_0 + \Delta x) - log(x_0) &\approx log(x_0) + log'(x_0)\Delta x - log(x_0) \\ &= \frac{\Delta x}{x_0} = \text{percentage change in x} \end{aligned} \tag{3}$$

Therefore

$$\begin{aligned} \frac{\Delta \text{ log(Happiness Score)}}{\Delta \text{ log(GDP)}} &\approx \frac{\Delta \text{ Happiness Score}}{\Delta \text{ GDP}} \frac{\text{GDP}}{\text{Happiness Score}} \\ &= \frac{\%\Delta \text{ Happiness Score}}{\%\Delta \text{ GDP}} \\ &= 0.1285 \end{aligned} \tag{4}$$

Now we can interpret the result from **Table** 6. Only variables that are statistically significant will be reported. Roughly, a 1% increase in GDP leads to a 0.1285% rise in the happiness score, and a 1% increase in population leads to a 0.1218% fall in the happiness score.

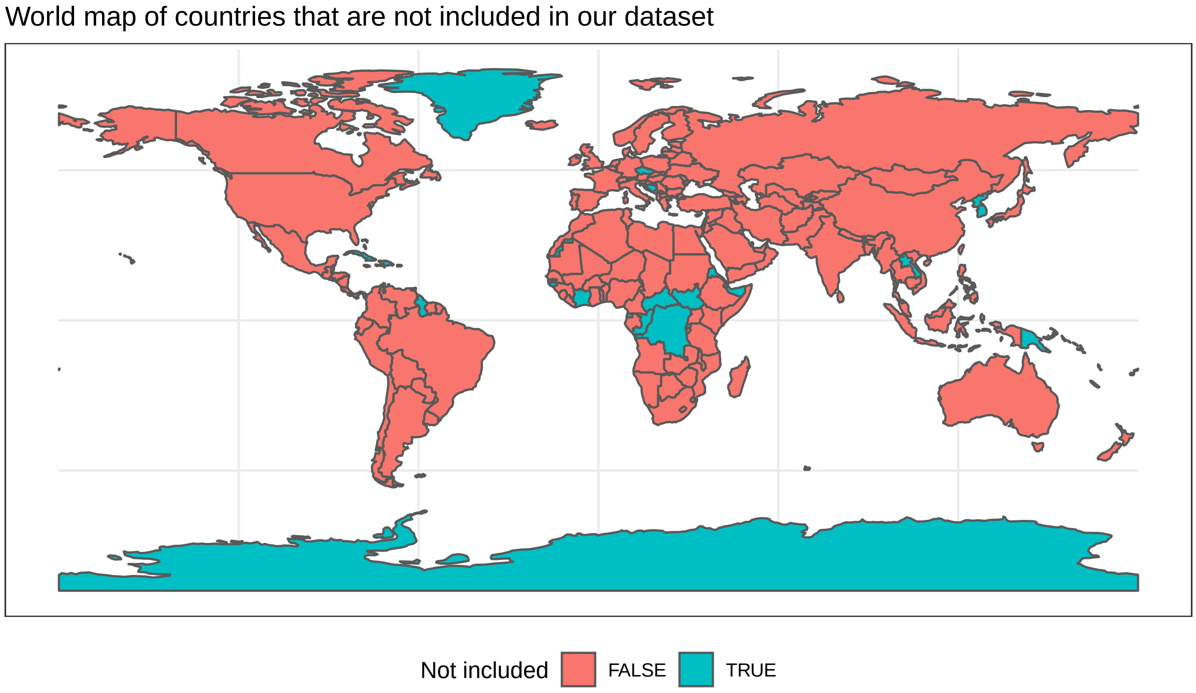


**Figure 7:** *The world map of missing countries in our dataset*

Apparently, the happiness score can be affected by other factors that correlated with regressors and doesn't include in our model, such as the education index and the social welfare system. It means an endogeneity issue occurs in our model, which leads our estimator to be bias and inconsistent. We can either include other variables that we think will be useful for happiness score prediction, or find instrumental variable and fit a Two-Stage least squares regression. Besides, another issue with our model is the sample selection problem. There are only 126 countries with completed cases in our final dataset, which means 69 countries are not considered. **Figure** 7 shows the map of missing countries. According to the map, most of the countries being omitted are developing countries. It indicates that our samples are not randomly selected from the population, and we are facing an observability issue. It can be expected that the happiness scores in those countries are lower than the average, which suggests we potentially underestimate our coefficients in OLS.

The diagnostics of our regression is shown in **Figure** 8. From the Residuals vs Fitted plot, we can see the non-constant variance across the fitted value. The model can be adjusted using Heteroskedasticity and Autocorrelation Consistent (HAC) covariance matrix estimation (Newey & West, 1986) to let our inference to be credible. The QQ plot suggests the density of residuals is asymmetric, but it is very close to a normal distribution. In the Cooks distances vs Fitted plot, we use 4/(n-k-1) as the threshold suggested in the book written by Fox (2019). The most influential points are from Suriname, Angola, Botswana, Sudan, Gabon, Sierra Leone, Togo and Burundi. **Figure** 9 shows the pattern of these countries. Most of them are in Africa, which contains extreme values in either happiness scores, GDP, CPI or area. It tells us that Africa is different from other places in the world, and we can consider putting a dummy variable in our model to indicate whether the country is in Africa.

In conclusion, this model does a solid job in explaining the relationship between happiness score and our regressors.

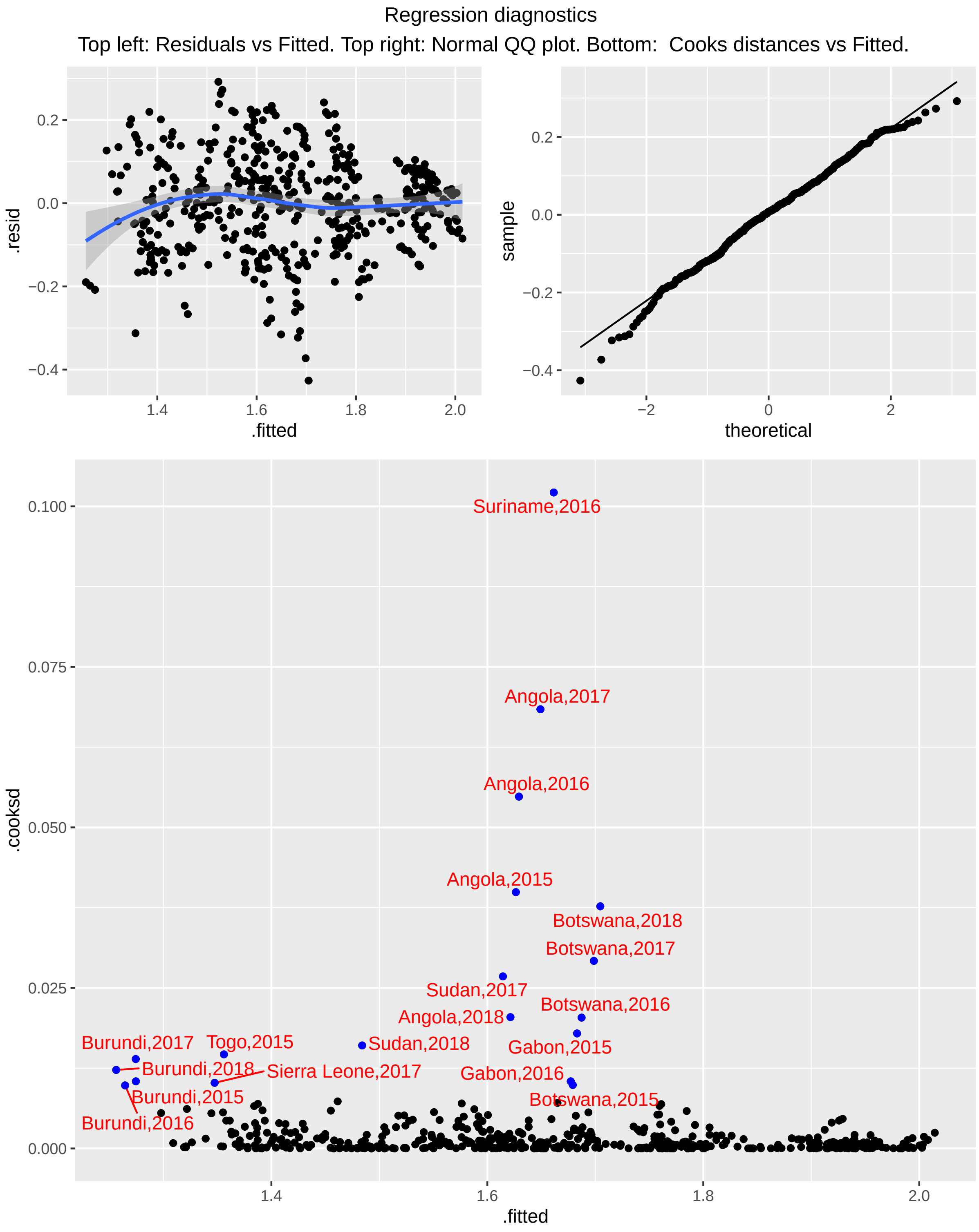
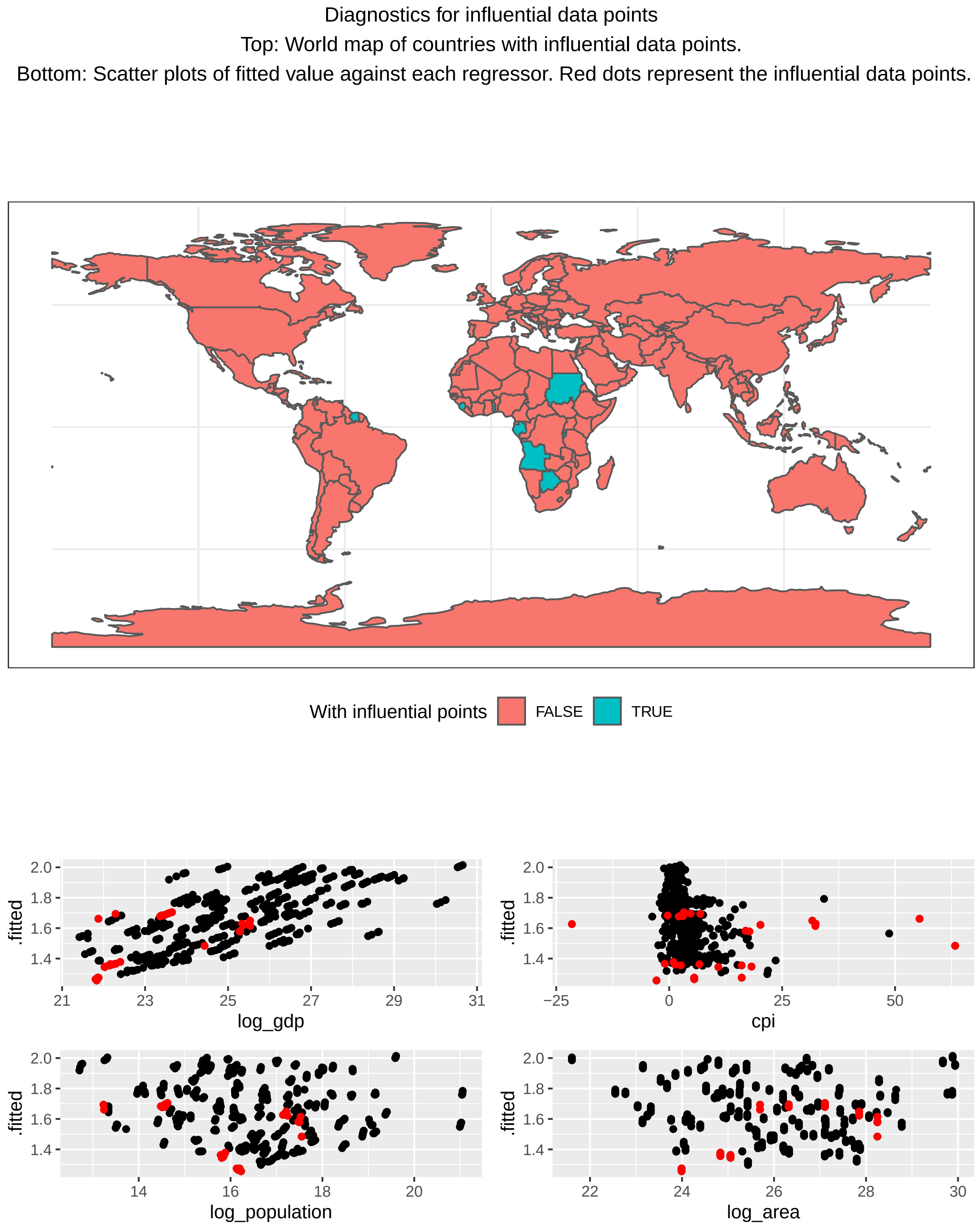


**Figure 8:** *Regression diagnostics for checking heteroskedasticity, non-normality and influential points*



Diagnostics for influential data points

Top: World map of countries with influential data points.

Bottom: Scatter plots of fitted value against each regressor. Red dots represent the influential data points.

**Figure 9:** *Diagnostics for influential data points*

## 5 Acknowledgement

- package ggplot2 (Wickham, 2016),
- package tidyverse (Wickham, 2017),
- package dplyr (Wickham et al., 2020),
- package relaimpo (Grömping, 2006),
- package GGally (Schloerke et al., 2020),
- package kableExtra (Zhu, 2019),
- package rnaturalearth (South, 2017),
- package sf (Pebesma, 2018),
- package rgeos (Bivand & Rundel, 2018),
- package lwgeom (Pebesma, 2020),
- package car (Fox & Weisberg, 2011),
- package here (Müller, 2017),
- package lubridate (Grolemund & Wickham, 2011),
- package cowplot (Wilke, 2019)

## References

Bivand, R, & Rundel, C. (2018). *Rgeos: Interface to geometry engine - open source ('geos')* [R package version 0.4-2]. https://CRAN.R-project.org/package=rgeos

Clark, A, & Senik, C. (2011). Will gdp growth increase happiness in developing countries?

Daoud, JI. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, **949**(1), 012009.

Fox, J. (2019). *Regression diagnostics: An introduction* (Vol. 79). SAGE Publications, Incorporated.

Fox, J, & Weisberg, S. (2011). *An R companion to applied regression* (Second). Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Grolemund, G, & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, **40**(3), 1–25.

Grömping, U. (2006). Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software*, **17**(1), 1–27.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**(2), 137–152.

Helliwell, JF, Layard, R, & Sachs, JD. (2019). World happiness report 2019. new york: Sustainable development solutions network.

Müller, K. (2017). *Here: A simpler way to find your files* [R package version 0.1]. https://CRAN.R-project.org/package=here

Newey, WK, & West, KD. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, **10**(1), 439–446.

Pebesma, E. (2020). *Lwgeom: Bindings to selected 'liblwgeom' functions for simple features* [R package version 0.2-1]. https://CRAN.R-project.org/package=lwgeom

Schloerke, B, Cook, D, Larmarange, J, Briatte, F, Marbach, M, Thoen, E, Elberg, A, & Crowley, J. (2020). *Ggally: Extension to 'ggplot2'* [R package version 2.0.0]. https://CRAN.R-project.org/package=GGally

South, A. (2017). *Rnaturalearth: World map data from natural earth* [R package version 0.1.0]. https://CRAN.R-project.org/package=rnaturalearth

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'* [R package version 1.2.1]. https://CRAN.R-project.org/package=tidyverse

Wickham, H, Fran?ois, R, Henry, L, & Müller, K. (2020). *Dplyr: A grammar of data manipulation* [R package version 1.0.0]. https://CRAN.R-project.org/package=dplyr

Wilke, CO. (2019). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'* [R package version 1.0.0]. https://CRAN.R-project.org/package=cowplot

Zhu, H. (2019). *Kableextra: Construct complex table with 'kable' and pipe syntax* [R package version 1.1.0]. https://CRAN.R-project.org/package=kableExtra