Analysisi of the World happiness

Group 21

2021/6/24

Introduction

The World Happiness Report is a landmark survey of the state of global happiness. However, because of the covid-19 prevalence, people's lives get reshaped largely. More seriously, As the pandemic struck, there was a large and immediate decline in mental health in many countries worldwide. Therefore, it seems necessary to give more importance to happiness and well-being thereby better achieving and measuring social and economic development. Our study mainly serves two purposes. One is about finding and evaluating the influential factors of happiness. The other is about tracing the size and distribution of happiness impacts. We adopts the multiple regression analysis, using the data from the Gallup World Poll. There are six variables to measure happiness scores. After processing the data, we finally get three appropriate influential factors.

At first, we choose seven variables with one outcome variable, Happy score and six explanatory variables, they are Logged GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity and Perceptions_of_corruption. The following is interpretations of each variable.

Happy score: respondents were asked to rate their possible lives from one to ten, with 1 being the worst and 10 being the best. Logged GDP per capita: it means logged GDP index from 149 countries.

Social support: it means if the participants could find someone's help when they are in trouble.

Healthy life expectancy: it means the expected life expectancy.

Freedom to make life choices: it means if the participants are satisfied with their freedom to choose what they do in life.

Generosity: it means if the participants are generous after asking if they have donated money to charity organizations.

Perceptions of corruption: it means if the participants are conscious of the corruption within the government and business.

Exploratory Data Analysis

We visualize the values as boxplots:

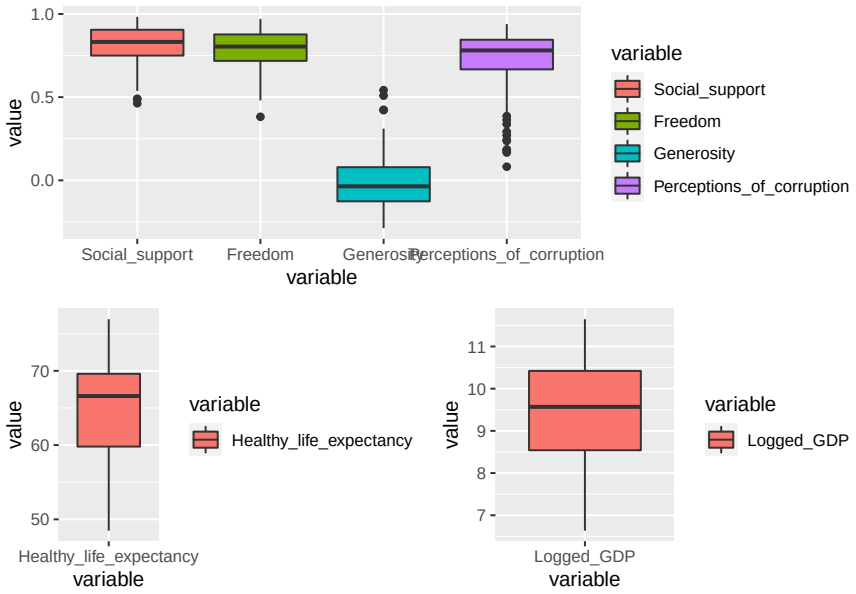

Figure 1: Boxplots.

Table 1: Summary statistics for observations with chosen variables.

	score	LoggedGDP	Social	expectancy	Freedom	Generosity	corruption
	Min. :2.523	Min. : 6.635	Min. :0.4630	Min. :48.48	Min. :0.3820	Min. :-0.28800	Min. :0.082
	1st Qu.:4.852	1st Qu.: 8.541	1st Qu.:0.7500	1st Qu.:59.80	1st Qu.:0.7180	1st Qu.:-0.12600	1st Qu.:0.66
	Median :5.534	Median : 9.569	Median :0.8320	Median :66.60	Median :0.8040	Median :-0.03600	Median :0.78
	Mean :5.533	Mean : 9.432	Mean :0.8147	Mean :64.99	Mean :0.7916	Mean :-0.01513	Mean :0.727
	3rd Qu.:6.255	3rd Qu.:10.421	3rd Qu.:0.9050	3rd Qu.:69.60	3rd Qu.:0.8770	3rd Qu.: 0.07900	3rd Qu.:0.84
	Max. :7.842	Max. :11.647	Max. :0.9830	Max. :76.95	Max. :0.9700	Max. : 0.54200	Max. :0.939

Visualization of the data

We can visualize our data by producing histogram and radarplot, where seeing as we have several countries, we shall plot the bars or lines using different colours for each country:

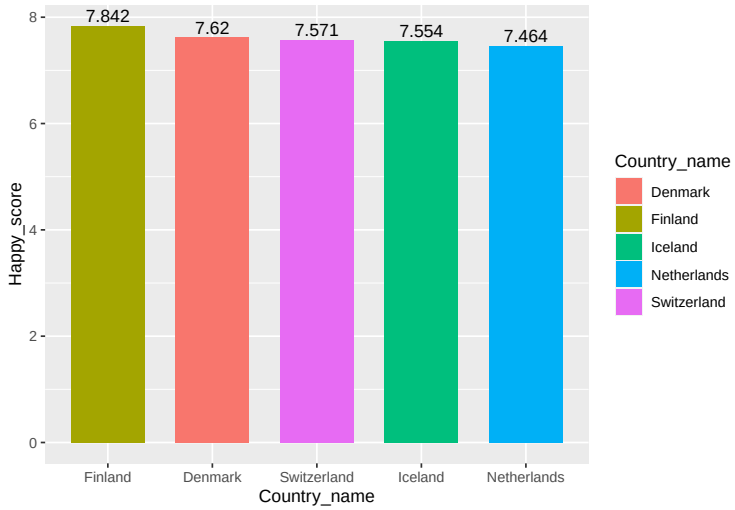

Figure 2: The top 5 happiest contries.

Figure 2 shows the 5 happiest countries. Finland gets the highest happy score with 7.842, followed by Denmark at 7.62. in addition, the differences between those five countries are small, less than 0.4, and all scores are greater than 7.4.

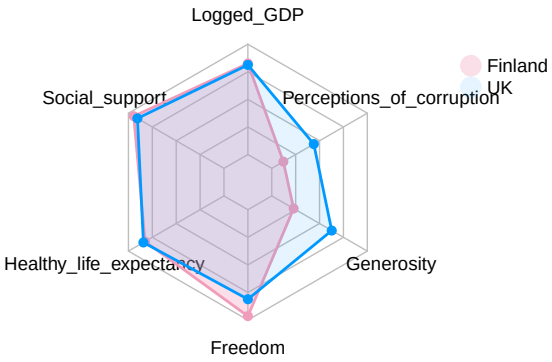

Figure 3: Comparisons between Finland and the UK from six index.

Figure 3 compares those six dimensions that Finland obtains with those that UK obtains. There are obvious differences between the perceptions of corruption and generosity, with the UK getting relatively much higher values. Conversely, when referring to Freedom indicators, Finland has a slightly higher value. Besides, it owns similar grades among the other three aspects.

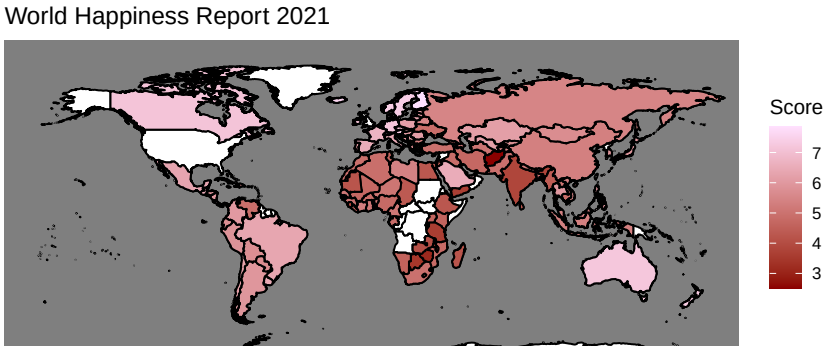

Figure 4: World happiness score map.

Figure 4 has displayed a happiness score map globally based on the World Happiness Report 2021. We can see that most countries in the western area have high scores whereas the eastern countries have low scores, which demonstrates that people in western countries lead happier and more fulfilling lives.

Formal Data Analysis

To begin to analysis the world happiness dataset, we need to check the correlation between the six explanatory variables to avoid the problem of multicollinearity.

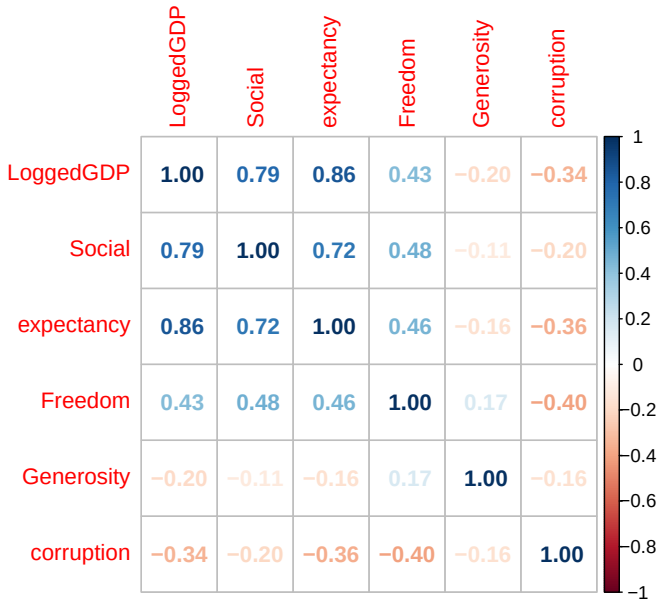

Figure 5: Correlation coefficient table.

From our correlation table we can see that the correlation between our Logged GDP and Healthy life expectancy is 0.859, which is a strong positive linear relationship.And the Logged GDP and Social support also have the high degree of collinearity, the correlation between this two variables is 0.785.So we remove Healthy life expectancy and Logged GDP.Then, using the remaining 4 explanatory variables to perform stepwise regression, and observe whether the remaining variables need to be eliminated.

```
Start:  AIC=-141.88
score ~ Social + Freedom + Generosity + corruption

             Df Sum of Sq    RSS      AIC
- Generosity  1     0.091 53.856 -143.628
<none>                    53.765 -141.879
- corruption  1     6.078 59.843 -127.920
- Freedom     1     6.364 60.129 -127.210
- Social      1    43.953 97.718  -54.857

Step:  AIC=-143.63
score ~ Social + Freedom + corruption

             Df Sum of Sq     RSS      AIC
<none>                     53.856 -143.628
- corruption  1     5.988  59.844 -129.919
- Freedom     1     6.325  60.181 -129.082
- Social      1    47.398 101.254  -51.561

Call:
lm(formula = score ~ Social + Freedom + corruption, data = happiness)

Coefficients:
(Intercept)       Social      Freedom   corruption
     0.0779       5.6256       2.2271      -1.2254
```

According to the results of stepwise regression, we choose the model with the smallest AIC as the final model.Then,we fit the following linear model to the data.

$$\widehat{\text{score}}_{\text{i}} = \widehat{\alpha} + \widehat{\beta} * \text{Social}_{\text{i}} + \widehat{\gamma} * \text{Freedom}_{\text{i}} + \widehat{\delta} * \text{corruption}_{\text{i}}$$

where

- the $\widehat{\text{score}}_{\text{i}}$: the happiness score of the ith country.
- the $\widehat{\alpha}$: the intercept of the regression line.
- the $\widehat{\beta}$: the coefficient for the first explanatory variable Social.
- the $\widehat{\gamma}$: the coefficient for the second explanatory variable Freedom.
- the $\widehat{\delta}$: the coefficient for the second explanatory variable corruption.

When this model is fitted to the data, the following estimates of α (intercept) and β,γ and δ are returned:

Table 2: Estimates of the parameters from the fitted linear regression model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.078	0.559	0.139	0.889	-1.028	1.184
Social	5.626	0.498	11.297	0.000	4.641	6.610
Freedom	2.227	0.540	4.127	0.000	1.160	3.294
corruption	-1.225	0.305	-4.015	0.000	-1.829	-0.622

According to this table, the coefficient for social support tells us that, taking all other variables in the model into account and holding them constant, there is an associated increase, on average,every increase of 1 unit in the social support score increases the happiness index score by approximately 5.63 units.In the same way, when the freedom score of life choice increases by 1 unit, the happiness index score also increases by approximately 2.23 units.On the contrary, for every increase of 1 unit in the score for corruption, the total score of happiness index decreases by 1.23 units.

Before we can continue to use the fitted model, we must check the model's assumptions. It is best to consider these according to the residual plot in Figure 6.

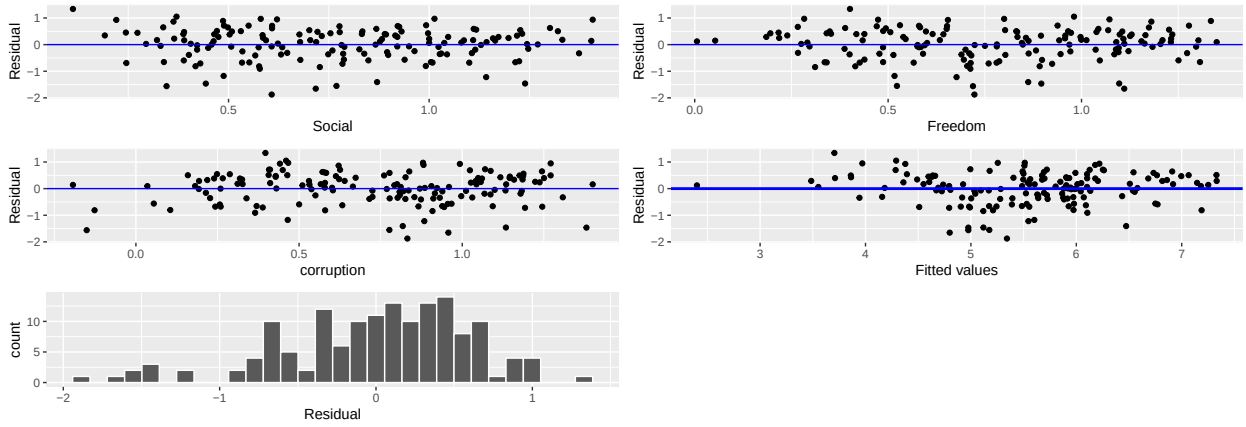

Figure 6: Scatterplots of the residuals by Social,Freedom,corruption,fitted value and the histogram of residuals.

The assumptions of the residuals having mean zero and constant variability across all values of the explanatory variable appear to be valid in this case.According to the three different explanatory variables scatter plots, it can be concluded that the residuals are uniformly distributed above and below the zero line, so the mean is 0. The residuals are randomly distributed around the zero line, and the distribution of the residuals is constant across all fitted values, so there is no obvious pattern or change in the variant.And also the histogram supports the assumption of normal distribution error. After hypothesis testing, we will analyze the confidence interval and set the confidence level to 95%.

```
                2.5 %     97.5 %
(Intercept) -1.027760  1.1835605
Social       4.641318  6.6098138
Freedom      1.160482  3.2937585
corruption  -1.828636 -0.6222416
```

The data in the table shows that the coefficient range of Social support affecting happiness score is 4.64 to 6.61, and the coefficient range of Freedom is 1.16 to 3.29. The most explanatory variable corruption affects the corresponding variable in the range of -1.83 to -0.62. Does not contain 0, indicating that the selected independent variables and their coefficients are valid, and as analyzed above, their given coefficients are the mean value of the interval.

Conclusions