Final Project

Xinpeng Hua

2022-12-10

Abstract

This project is to explore the relationship between happiness and some basic indexes, which is based on the World Happiness Report from 2015 to 2021. There are several variables which may influence the happiness score and I choose to clean the data sets and create EDA and the best multilevel model with the group level `Region` for this data. This result indicates the relationship between happiness score and different variables. The report can be divided into 4 parts: Introduction, Method, Result and Discussiion.

Introduction

The happiness is vital for the human. Throughout history, wise rulers have wanted to create a utopia, which means that the happiness score is close to the full score.The social order is unprecedentedly perfect, and there is no evil in this country. But this is impossible. There is always a dark side in society, and the happiness score will not tend to be perfect. It is obvious that every country, every region has different level of happiness. And there are 6 potential factor:`GDP`,`social support`,`life expectancy`, `freedom`,`generosity`,`corruption`. And this report will divide countries in the world into 10 regions: `Australia and New Zealand`,`Central and Eastern Europe`, `Eastern Asia Latin America and Caribbean`, `Middle East and Northern Africa`, `North America`, `Southeastern Asia`, `Southern Asia`, `Sub-Saharan Africa Western Europe`. In our common sense, we always think that more money, more freedom, more help, less corruption and so on can make people happier. However, I am interested in this point, and I want to figure out how they influence and the difference of each region.

Method

Data preprocessing

I select the happiness report from 2015 to 2019. Actually I want to use the latest report to 2022, but I found that the 2022 data set has some problem which is not easy to solve, and the data may not real, so I change it to 2021. There are 7 CSV documents totally for me to combine them to one data set. Through the data cleaning, I found that the the categories of regions of the reports are different from 2015 to 2019. And I decide to use 2015 categories of regions for unifying the standard.

Exploratory Data Analysis

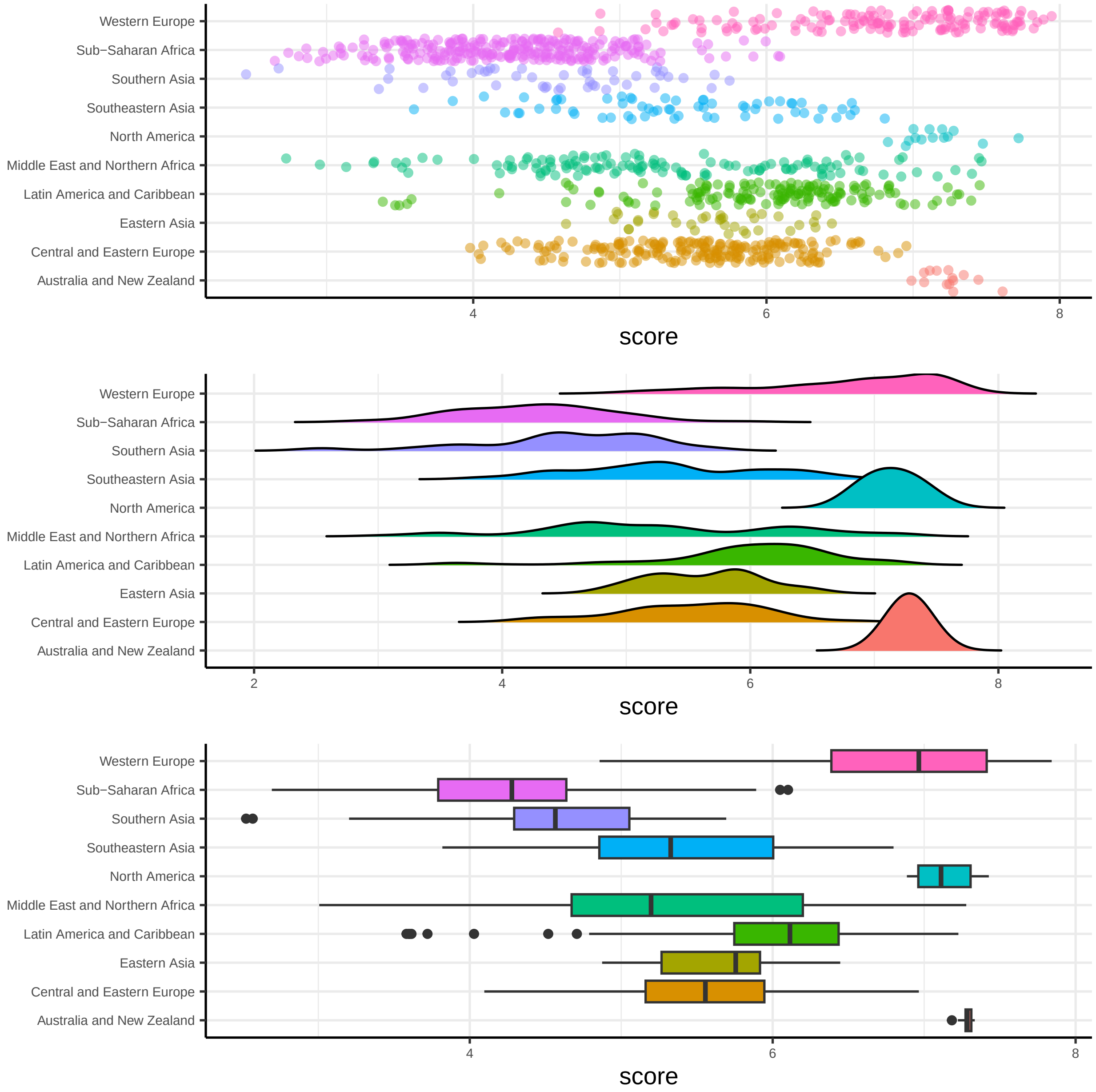


Figure 1: Compare the distributions of happiness score

As during these years, there is little difference between each year. So I create the Figure 1 based on the different regions of whole period. Through the three basic plots, it is clear that Australia and New Zealand region always has high score during this period. And the main point is that these two areas are developed, which means their people get good quality of life. It is true that some developed countries have low happiness score like Japan, so our opinion is based on the common condition. Besides, western Europe and North America also get the high score relatively. However, Sub-Saharan Africa have the lowest score during the years relatively. From my perspective, the wars made these area lose the peace and people lost their homes so that they may get poor. Middle East and Northern Africa and Southeastern Asia have the high level of dispersion. Maybe the reason is that these regions include the countries which have large wealth gap between rich and poor.

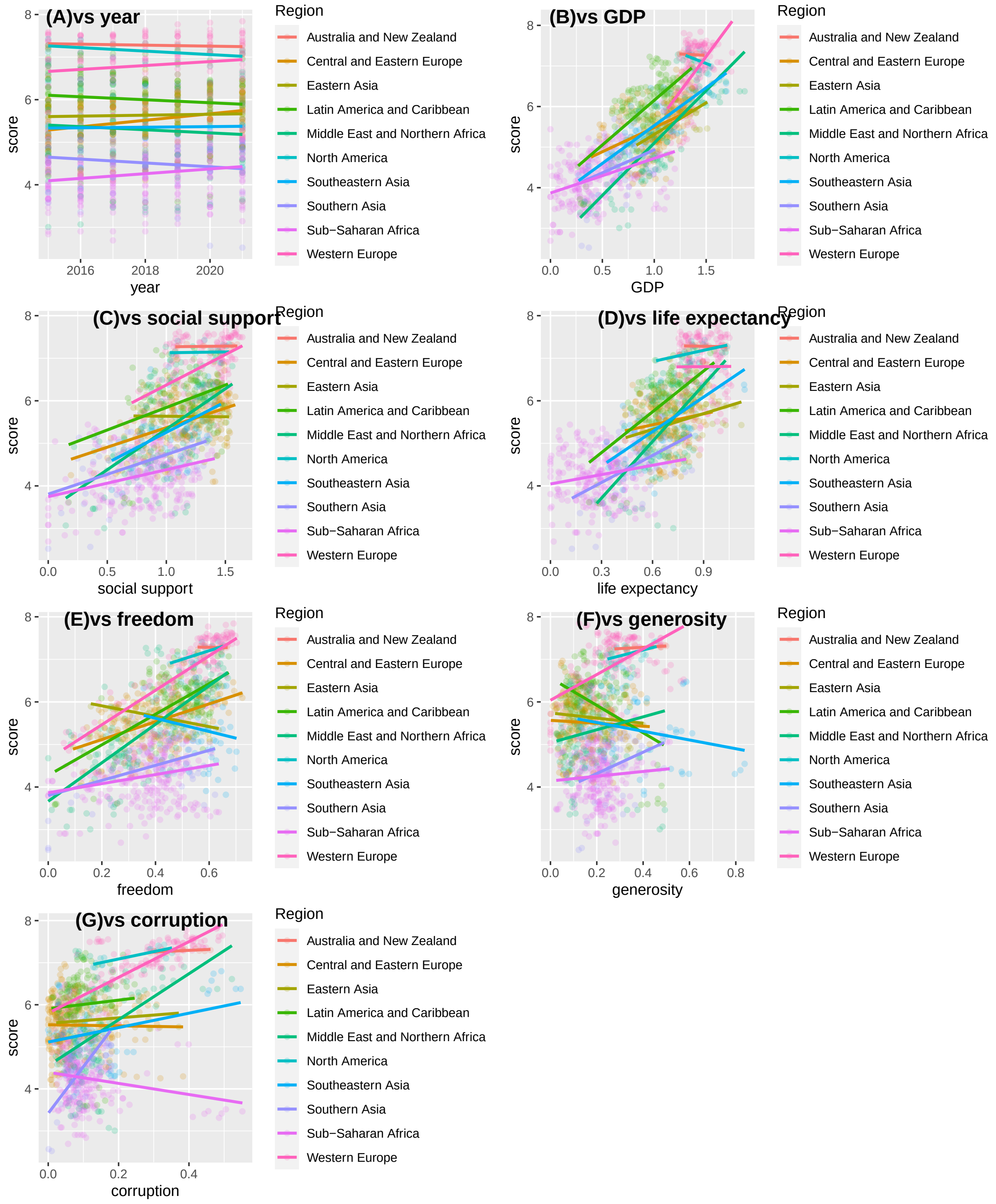


Figure 2: Relationship between happiness scores and different variables

And there are more plots of relationship between happiness scores and different variables. I create the Figure 2 to explore it. Firstly, plot A reveals that the year-to-year changes in scores for different regions are small. This is why we can use the data of different years before to create the EDA basically. Plot B, Plot C and Plot D all illustrate the positive relationship between scores and the variables, GDP, social support, life expectancy on the whole. But Plot E, F, G shows that more regions like north America and eastern Asia have negative relationship between scores and freedom, generosity, corruption.

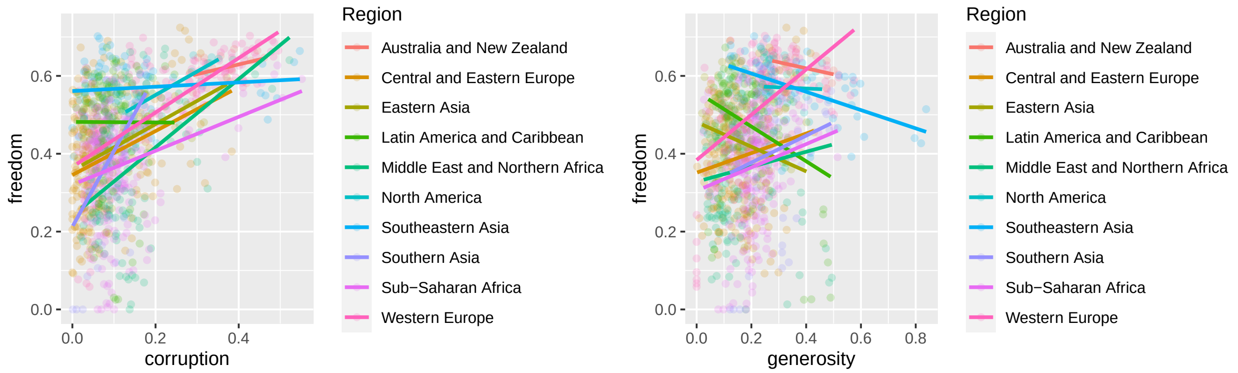


Figure 3: Relationship between freedom and corruption and generosity

Therefore, I create the Figure 3 to try to get the relationship between corruption and generosity with freedom. It turns out that even half regions have negative relationship between freedom and generosity. It is a really special find for me, in my opinion, this is because the more freedom people of those regions have, the more things they can do, which means they can do the good things, but they can do bad things either so that they lose the generosity.

Model Fitting

I try to find the correlations between the variables so that I create the Figure 4. It is obvious that all the variables have good correlation with happiness score. So firstly I choose all the variables to fit the multilevel model. Firstly I assume the region effect all the variables. But the model `fit` I fit was not good as the p-value of generosity or corruption was too large which means the model is not fitted well. Therefore, I choose to delete the `corruption` and `Region` cannot influence `generosity`. And then I found that the model `fit2` is fitted well, all the p-values of the coefficients are less than 0.05.

```
fit <- lmer(score ~ GDP + `social support` + `life expectancy` + freedom + generosity +
corruption + (1 + GDP + `social support` + `life expectancy` + freedom + generosity+
                corruption| Region), data = happiness)

fit2 <- lmer(score ~ GDP + `social support` + `life expectancy` + freedom + generosity+
(1 + GDP + `social support` + `life expectancy` + freedom | Region), data = happiness)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.4992	0.2656	9.408	9.29e-05
GDP	1.2892	0.1398	9.223	2.96e-05
social support	0.3211	0.1207	2.660	0.031129
life expectancy	0.8440	0.2340	3.607	0.007141
freedom	1.5040	0.3580	4.201	0.005853
generosity	0.5256	0.1519	3.461	0.000579

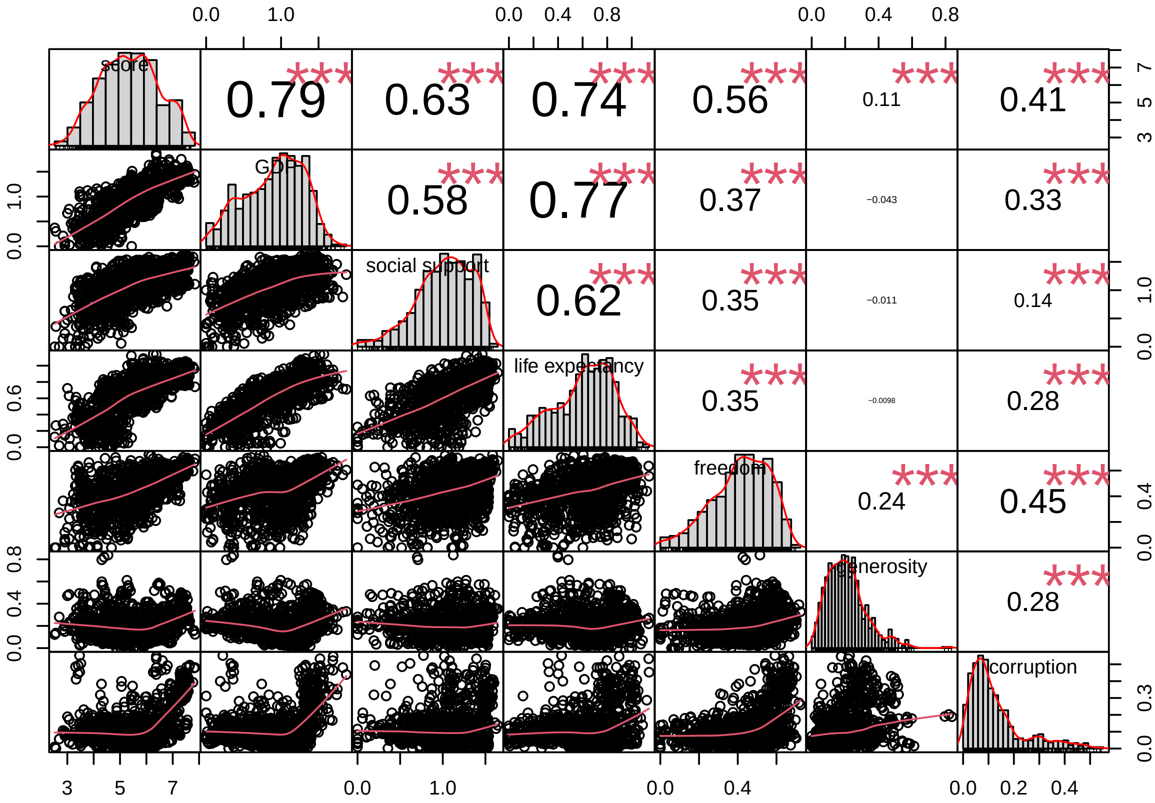


Figure 4: Correlation Matrix

Result

Coefficient

As a result, the formula for happiness scores is

$$happinessscore = 2.4992 + 1.2892 * GDP + 0.3211 * socialsupport$$
$$+0.8440 * lifeexpectancy + 1.5040 * freedom + 0.5256 * generosity$$

And the following table is the summary of random effects.

```
## $Region
##                                  (Intercept)       GDP `social support`
## Australia and New Zealand           2.059833 1.5720800        0.2182191
## Central and Eastern Europe          2.677014 1.2131800        0.4651604
## Eastern Asia                        3.332917 1.0929120        0.2296202
## Latin America and Caribbean         2.512630 1.4706974       -0.2057119
## Middle East and Northern Africa     1.762405 1.5475246        0.5256438
## North America                       2.094855 1.5662867        0.2034429
## Southeastern Asia                   2.574428 1.3761949        0.4093877
## Southern Asia                       3.217085 0.7953694        0.3817467
## Sub-Saharan Africa                  3.229537 0.6781080        0.4026720
## Western Europe                      1.531127 1.5794203        0.5804446
##                                  `life expectancy`    freedom generosity
## Australia and New Zealand                1.2704427 2.27740294   0.525629
## Central and Eastern Europe               0.3501108 1.84949567   0.525629
## Eastern Asia                             0.6750541 0.08613464   0.525629
## Latin America and Caribbean              1.7660207 2.19935678   0.525629
## Middle East and Northern Africa          0.9846416 1.72975074   0.525629
## North America                            1.2855040 2.21651033   0.525629
## Southeastern Asia                        0.8046625 0.71291063   0.525629
## Southern Asia                            0.2860793 0.53987470   0.525629
## Sub-Saharan Africa                       0.1801235 0.65950243   0.525629
## Western Europe                           0.8375368 2.76872700   0.525629
##
## attr(,"class")
## [1] "coef.mer"
```

It is clear that the variables I choose from the data have positive relationship with the happiness score, and we can conclude that GDP and freedom influence the most for the happiness. It was also predictable. Because personal freedom and material needs are the two most important things for human beings. So they can affect people's happiness a lot. From the random effect table, there is only one negative number. It is the coefficient of social support in Latin America and Caribbean which is -0.2. Actually, it is hard to explain why this area have this phenomenon because the number is close to 0 which means the influence is little but exists. Besides, the life expectancy and freedom influence a lot for the Latin America and Caribbean. And for the eastern Asia, every variable influence the least relatively to the happiness score, that is possibly because of the culture of eastern Asia which decide the happiness score of the society.

Checking

Through Figure 5, I use the qqplot to check whether or not the model is fitted well. The qqplot shows that all the points basically form a straight line and are close to the fitting line, which means the model fit well as I predict.

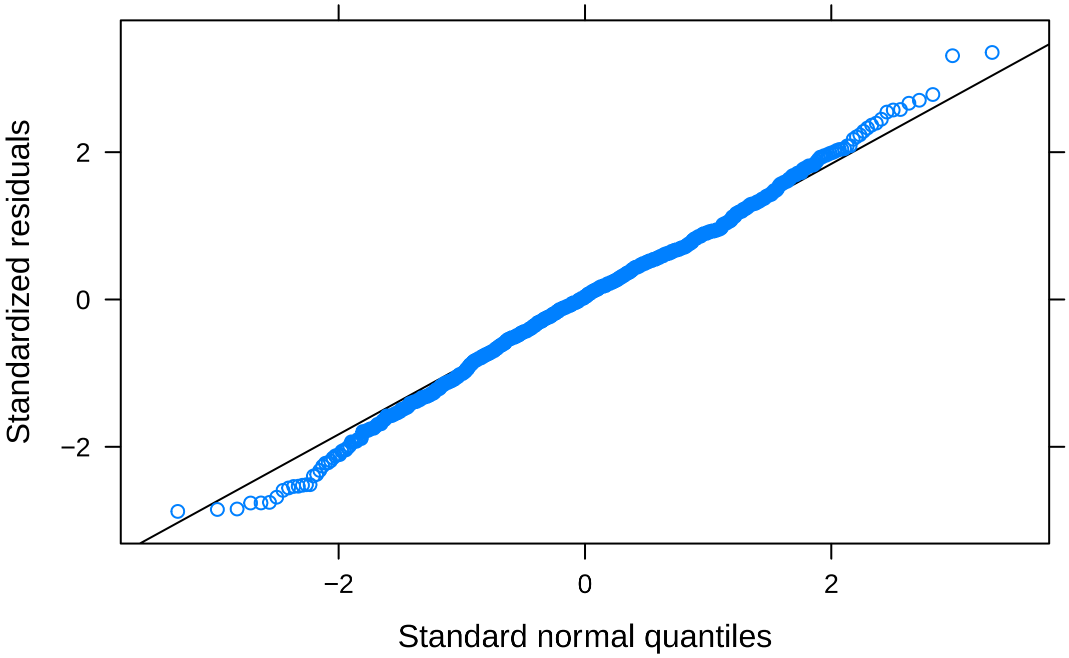


Figure 5: qqplot

Discussion

In this report, I use the multilevel model to figure out the relationship between happiness scores and several factors as I mentioned before. The result is not beyond our prediction, all the variables have positive relationship with happiness score, which means that the effects of some of the factors that make people happier in our common sense are all positive. However, there are some limitations of the model. I deleted the `corruption` to make the model fit better. Actually, `corruption` should have some effects on scores. I am supposed to learn more models to fit it well without deleting a variable. And I ignored the little effect of the year on the score either, which may become the reason of deviation of model coefficient.

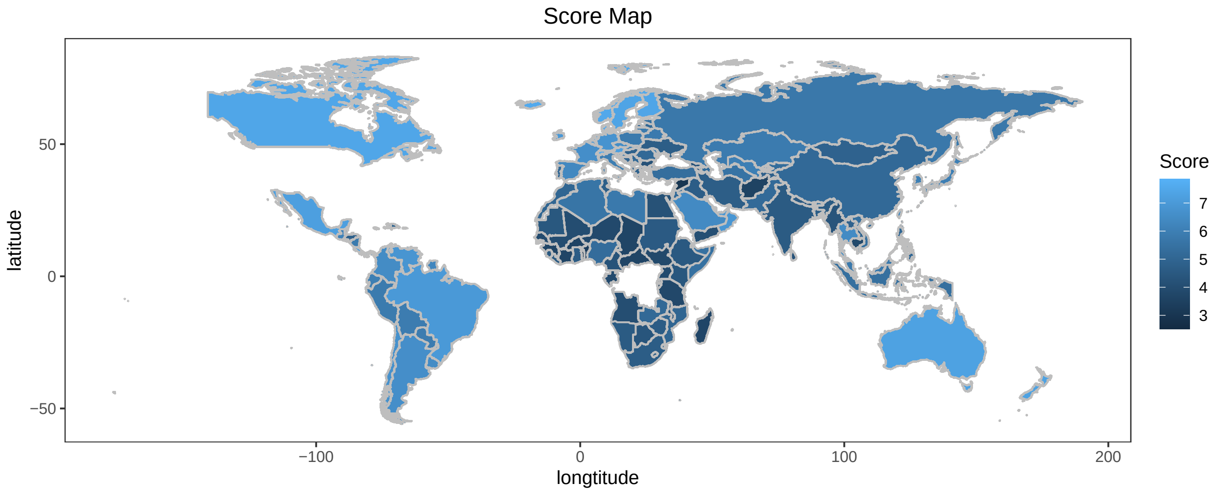
Reference

[1] World Happiness Report. https://www.kaggle.com/datasets/mathurinache/world-happiness-report-20152021

[2] Micheal, P. Chapter 18: Testing the Assumptions of Multilevel Models. https://ademos.people.uic.edu/Chapter18.html

[3] ALEXANDRU CERNAT. Cross-national research using multilevel model in R.https://www.alexcernat.com/cross-national-research-using-multilevel-model-in-r/

Appendix


Score Map
latitude
50
0
−50
−100
0
100
200
longtitude
Score
7
6
5
4
3