Analyze the factors that influence happiness in different regions

1 Introduction

Promoting the development of society is human's immutable goals and motives. In recent years, with the development of material life, people begin to pursue the spiritual satisfaction. The focus of happiness research has shifted from the measurement of life quality with an emphasis on material well-being to subjective well-being.(Lin K,2016) And the definition of social development is constantly updating. At present, people's pursuit of life has gradually shifted from the improving at material level to gradually improving quality of life. In this process, the quality of life has become one of the most important concerns of both the government and massive residents.

Happiness is an important standard to evaluate people's living standard. The National Happiness Survey is a landmark survey that can be used by governments to make important international and national decisions. It is also an effective measure of national progress.Through the analysis of the relationship between happiness and several important factors that affect it, the most significant influencing factors can be obtained, and the results can be applied to the field of social psychology, which is conducive to social construction.(Helliwell J F, Huang H, Wang S,2016) According to the data from the World Happiness Report 2021(Rowan A N,2021), we believe that there are some important factors that affect the happiness index:GDP, Support, Freedom, Expectancy, Generosity, Corruption. We will examine the relationship between happiness and these factors. We are interested in performing the analysis for predicting Happiness score from these factors and which out of them are the most significant. These factors are vital to a country's standard of living. In addition, happiness index scores are derived from people's questionnaires.

This article mainly consists of five parts. Section 1 is introduction of proposal. Section 2 is Main definitions and method of the factors used in the analysis. Section 3 is the exploratory data analysis of happiness index and these factors in countries around the world, and studies the relationship between them. Section 4 is the concrete analysis of data by fitting linear regression model. Section 5 is the summary of analysis.

2 Main Definitions And Method

2.1 Population and Data

2.1.1 Population

147 countries in World Happiness Report 2021

2.1.2 Data Collection

We chose the data from World Happiness Report 2021. There is 149 regions in total but here we summarized the Taiwan(China) ,Hongkong(China) and China in one region named China. So the data has 147 observations The variables are shown in the followed table.

Table 1: variable definition

category	variable	definition	type
Dependent variable	happiness	state of well-being characterized by emotions ranging from contentment to intense joy	continuous variable
Independent variable	Region	one of the areas that a country is divided into, that has its own customs and/or its own government	classified variable
Independent variable	GDP	the total value of all the goods and services produced by a country in one year	continuous variable
Independent variable	Support	encouragement and help that government and society give to people	continuous variable
Independent variable	Freedom	he right to do or say what you want without anyone stopping you	continuous variable
Independent variable	Expectancy	he number of years that a person is likely to live	continuous variable
Independent variable	Generosity	the fact of being generous	continuous variable
Independent variable	Corruption	dishonest or illegal behaviour, especially of people in authority	continuous variable

Table 2: Region

Region	Abbreviation_for_region
South Asia	SA
Central and Eastern Europe	C&EE
Middle East and North Africa	ME&NA
Latin America and Caribbean	LA&C
Commonwealth of Independent States	CIS
North America and ANZ	NA&A
Western Europe	WE
Sub-Saharan Africa	SSA
Southeast Asia	SEA
East Asia	EA

2.2 Objective

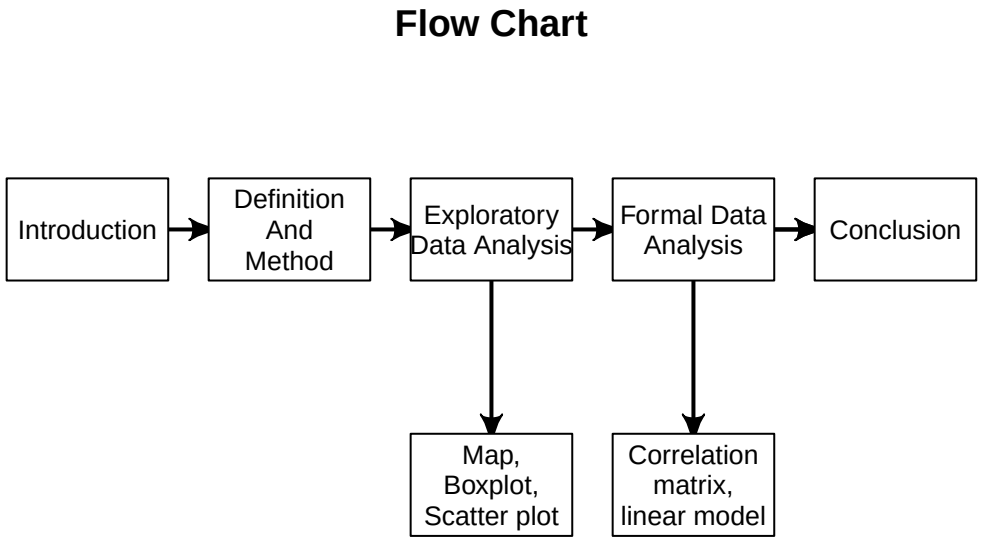
2.2.1 Primary Objective

Analyze the influencing factors of happiness

2.2.2 Secondary Objectives

Analyze the difference of happiness of residents around the world

2.3 Work-flow



3 Exploratory Data Analysis

3.1 Analysis of the development situation of different regions

The table shows aggregate statistics on happiness and other variables for 147 countries.The standard deviation of happiness(1.1)is 11 times greater than the standard deviation of support(0.1), freedom.Support, freedom(0.1) is relatively stable. Expectancy had the highest standard deviation (6.7), while freedom and support had the lowest standard deviation (0.1).

Table 3: Summary statistics

name	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
Happiness	147	5.5	1.1	2.523	4.8430	5.534	6.2390	7.842
GDP	147	9.4	1.2	6.635	8.5395	9.557	10.3755	11.647
Support	147	0.8	0.1	0.463	0.7480	0.832	0.9055	0.983
Expectancy	147	64.9	6.7	48.478	59.5535	66.601	69.5440	76.953
Freedom	147	0.8	0.1	0.382	0.7185	0.806	0.8780	0.970
Generosity	147	0.0	0.2	-0.288	-0.1285	-0.036	0.0800	0.542
Corruption	147	0.7	0.2	0.082	0.6700	0.787	0.8460	0.939

3.1.1 Analysis of the happiness of different regions

Figure 1 shows the distribution of happiness score across the countries around the world along with the boxplots which summarizes the geographical distribution numerically. The color gradient in the geographical plot gradually changes from blue to red. Bluer the region, it depicts a less happy country. On the other hand, the darker the gradient of red, the country is considered relatively happier.

Table 4 gives us the numerical values of the mean and median of happiness across the 10 regions. The table shows that North America and ANZ(NA&A) region has the highest median and mean happiness score compared to other regions whereas Sub-Saharan Africa(SSA) has the lowest median happiness score.

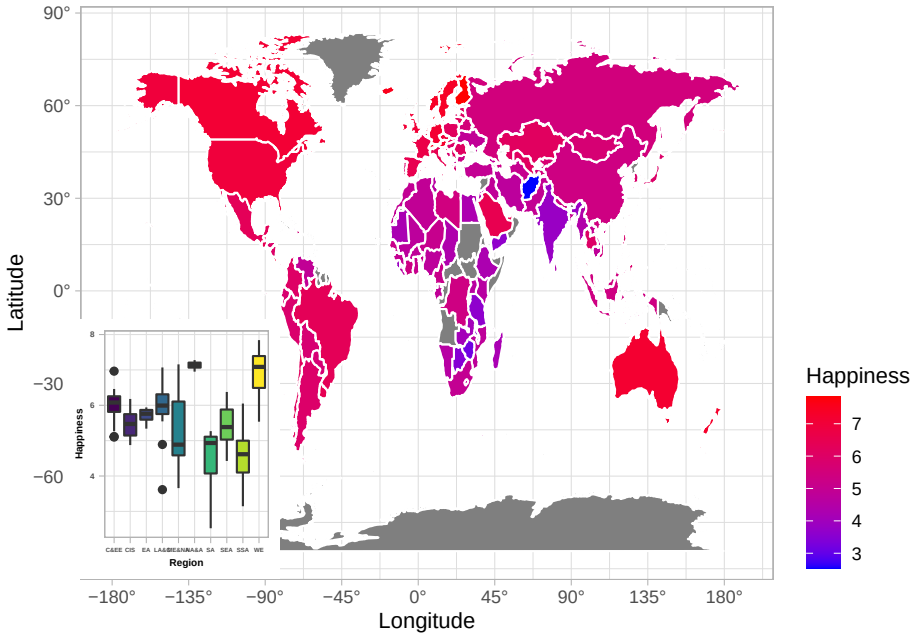


Figure 1: Happiness plot

Table 4: Happiness Table

Region	Mean of happiness	Median of happiness
C&EE	5.984765	6.0780
CIS	5.467000	5.4715
EA	5.700250	5.7610
LA&C	5.908050	5.9920
ME&NA	5.219765	4.8870
NA&A	7.128500	7.1430
SA	4.441857	4.9340
SEA	5.407556	5.3840
SSA	4.494472	4.6160
WE	6.914905	7.0850

3.1.2 Analysis of the GDP of different regions

The geographical variations in the GDP has been depicted in the Figure 2. The color gradient shifts from yellow to orange stating the increase in GDP. The boxplot along with the map distribution gives us a pictorial

idea of the GDP score of different regions.

Table 5 is the table of mean and median GDP score for the different geographical regions. It can be seen from the table and the boxplot that the region Western Europe(WE) has highest mean and median GDP while Sub-Saharan Africa(SSA) has the lowest mean and median GDP Scores.

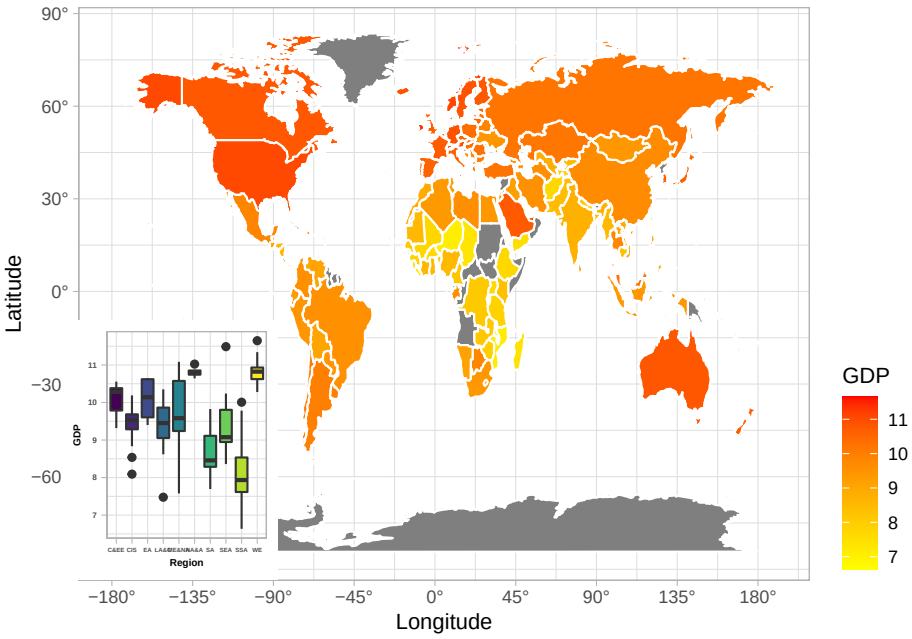


Figure 2: GDP plot

Table 5: GDP Table

Region	Mean of GDP	Median of GDP
C&EE	10.109059	10.2840
CIS	9.401833	9.5280
EA	10.083750	10.1420
LA&C	9.370000	9.4530
ME&NA	9.666118	9.5840
NA&A	10.809500	10.7860
SA	8.682571	8.4580
SEA	9.421444	9.0760
SSA	8.075194	7.9345
WE	10.822714	10.8230

3.1.3 Analysis of the Support of different regions

Support score is the numerical value which determines the extend to which the people receive support from the government. Figure 3 is the support plot which shows the distribution of support score across different geographies. The colour ascends from yellow to orange (yellow depicting the regions with lower support and orange depicting region with greater support). The boxplot gives us a pictorial summary of the Support score across 10 different regions and also shows us potential outliers from that region.

Table 6 gives us the exact mean and median value of the Support score across the different regions. From the table, it can be concluded that Western Europe(WE) and North America and ANZ(NA&A) have almost the similar mean and median support score. This is similar to the distribution of happiness.North America and Australia (NA&A) scored the highest. The region with lowest mean Support score is Sub-Saharan Africa(SSA) and lowest median Support score is South Asia(SA).

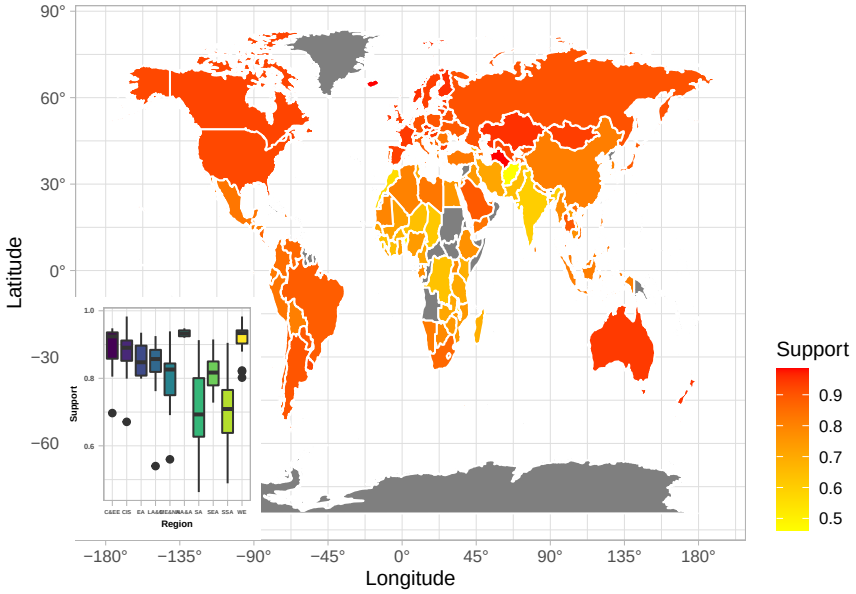


Figure 3: Support plot

Table 6: Support Table

Region	Mean of Support	Median of Support
C&EE	0.8874118	0.9240
CIS	0.8725000	0.8905
EA	0.8572500	0.8475
LA&C	0.8395000	0.8570
ME&NA	0.7976471	0.8260
NA&A	0.9335000	0.9330
SA	0.7034286	0.6930
SEA	0.8203333	0.8170
SSA	0.6967500	0.7090
WE	0.9144762	0.9340

3.1.4 Analysis of the Expectancy of different regions

Figure 4 is the Expectancy plot which gives the distribution of Expectancy across the regions with a boxplot giving representing numerical summaries. It can be seen from the boxplot that SSA region has the lowest median Expectancy score whereas North America and ANZ(NA&A) has the highest median Expectancy score.

Table 12 summarises for us, the exact numerical values of mean and median expectancy score across all 10 regions. As anticipated from the boxplots, the table verifies our conclusion that NA&A has highest mean and median Expectancy score and SSA has the lowest score.

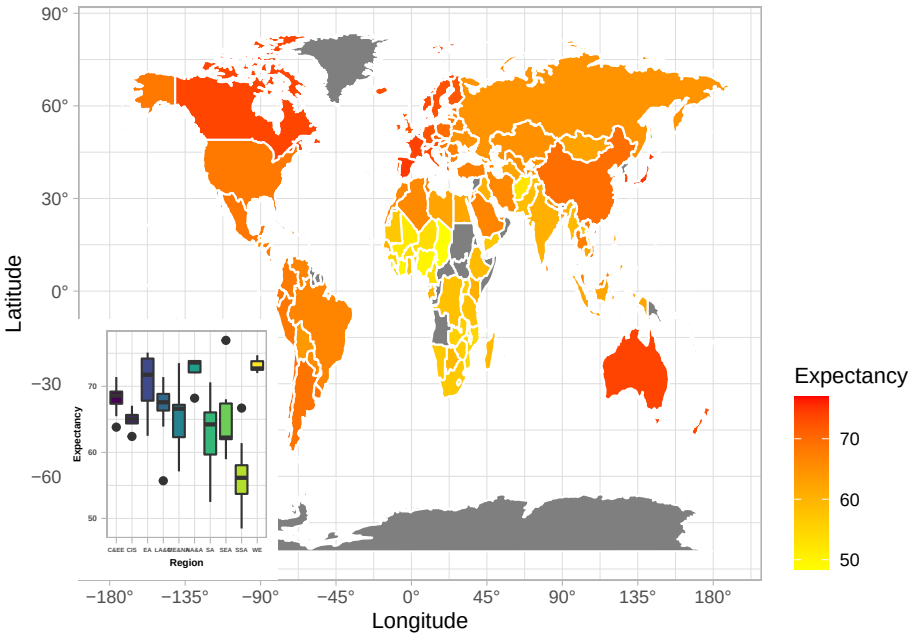


Figure 4: Expectancy plot

Table 7: Expectancy Table

Region	Mean of Expectancy	Median of Expectancy
C&EE	68.33841	68.6000
CIS	65.00950	65.0510
EA	70.27325	71.7465
LA&C	67.07605	67.5785
ME&NA	65.60912	66.6030
NA&A	72.32500	73.6000
SA	62.68100	64.2330
SEA	64.88844	62.2360
SSA	55.88647	56.1510
WE	73.03310	72.7000

3.1.5 Analysis of the Freedom of different regions

Figure 5 is a Freedom plot which shows us the distribution of freedom score across the regions. The boxplot is helpful to five number summaries of the freedom score for different regions. It can be seen from the boxplots that WE, NA&A and SEA have almost similar median freedom score whereas SSA has the lowest. The IQR for the Middle East and North Africa (ME&NA) is the highest.

Table 8 is the table of mean and median Freedom score. It can be concluded from the table that NA&A has the highest median score whereas SEA has the highest mean freedom score. SSA has the least mean and median Freedom score.

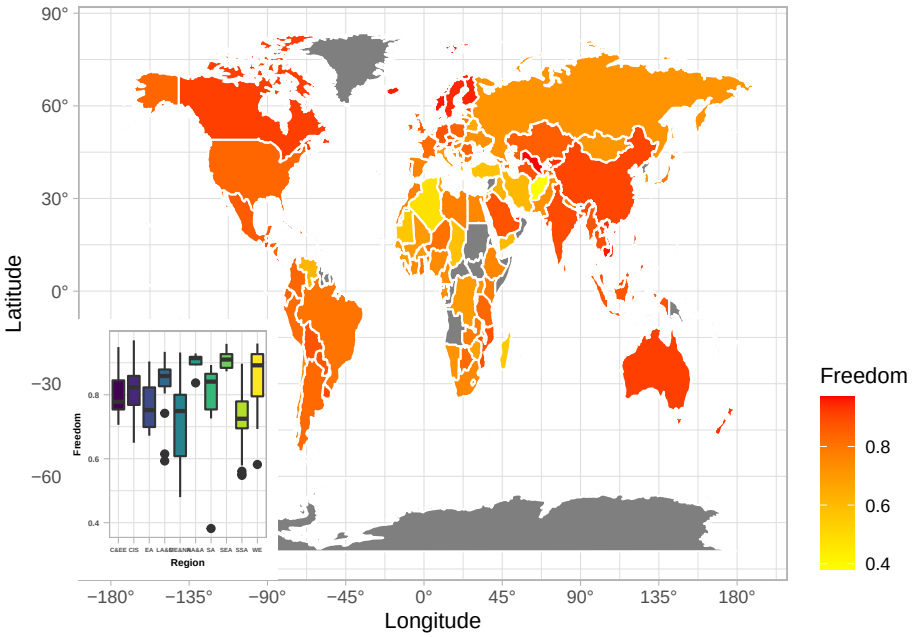


Figure 5: Freedom plot

Table 8: Freedom Table

Region	Mean of Freedom	Median of Freedom
C&EE	0.7970588	0.7780
CIS	0.8169167	0.8235
EA	0.7700000	0.7520
LA&C	0.8317500	0.8585
ME&NA	0.7164706	0.7490
NA&A	0.8987500	0.9145
SA	0.7650000	0.8410
SEA	0.9090000	0.9100
SSA	0.7231944	0.7250
WE	0.8587143	0.8920

3.1.6 Analysis of the Generosity of different regions

Generosity Scores distribution can be seen in Figure 6. The boxplot at left bottom left can help us analyse the generosity score amongst different regions. We can see that NA&A region and SEA region has almost similar median generosity score.

To verify the judgements that can be made from the boxplot, Table 9 can be helpful which gives numerical values of mean and median Generosity score for different regions. It can be concluded from the table that SEA has the highest median generosity score and LA&C region has the lowest median generosity score.

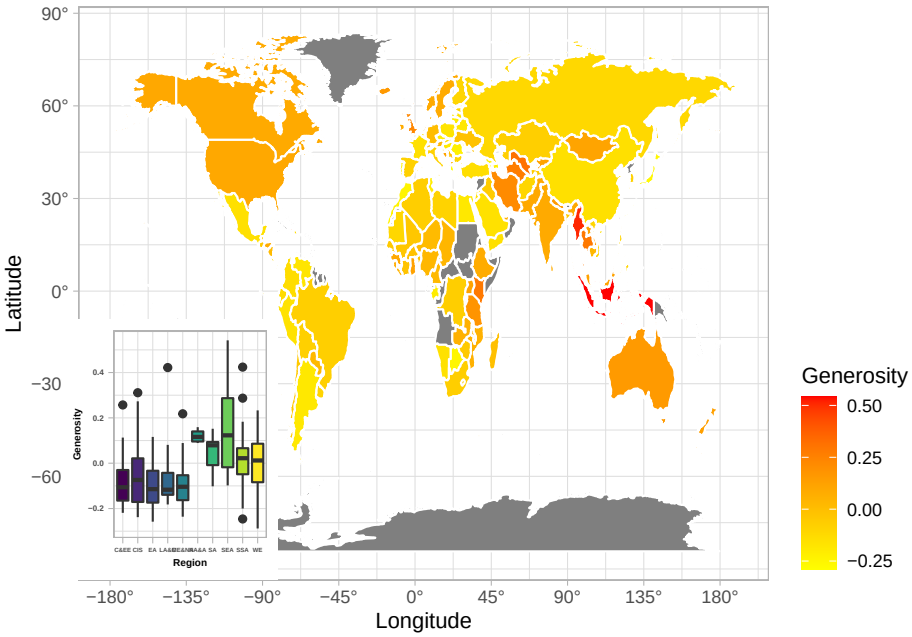


Figure 6: Generosity plot

Table 9: Generosity Table

Region	Mean of Generosity	Median of Generosity
C&EE	-0.0789412	-0.1060
CIS	-0.0360000	-0.0740
EA	-0.0927500	-0.1145
LA&C	-0.0677000	-0.1170
ME&NA	-0.0797647	-0.1040
NA&A	0.1200000	0.1160
SA	0.0427143	0.0790
SEA	0.1563333	0.1230
SSA	0.0134444	0.0220
WE	-0.0031905	0.0120

3.1.7 Analysis of the Corruption of different regions

Variations in corruption score across different geographies can be seen from Figure 7. The colour gradient ascends from yellow to orange. Yellow depicting the lowest corruption score and orange depicting the highest. The boxplot along with the geographical map summarises graphically the variations in freedom score for different region. It can be seen from the boxplot that WE region has the highest IQR. C&EE region has the highest Corruption score and NA&A has the lowest.

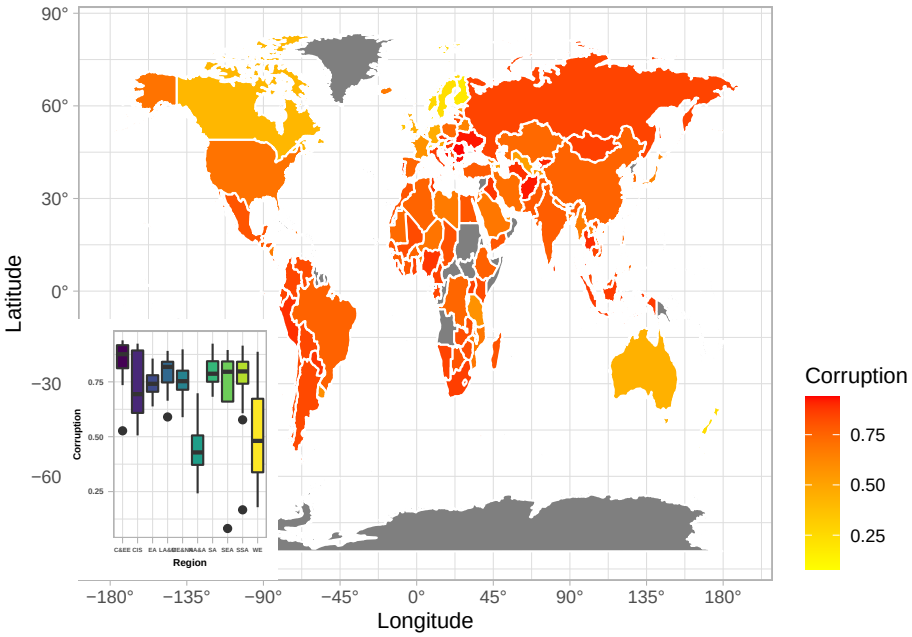


Figure 7: Corruption plot

Table 10: Corruption Table

Region	Mean of Corruption	Median of Corruption
C&EE	0.8505294	0.8760
CIS	0.7250833	0.6940
EA	0.7440000	0.7410
LA&C	0.7926000	0.8180
ME&NA	0.7622353	0.7530
NA&A	0.4492500	0.4285
SA	0.7974286	0.7870
SEA	0.7091111	0.7960
SSA	0.7659444	0.7975
WE	0.5230952	0.4810

3.2 Analysis of the influencing factors of happiness

3.2.1 Analysis of correlation between Happiness and GDP

Figure 8 is a scatterplot of happiness scores and GDP level in ten regions of the map. In most of the region, the GDP level and happiness scores show a linear positive relationship.Therefore, it is necessary to use a linear regression model to evaluate the relationship between GDP and happiness.

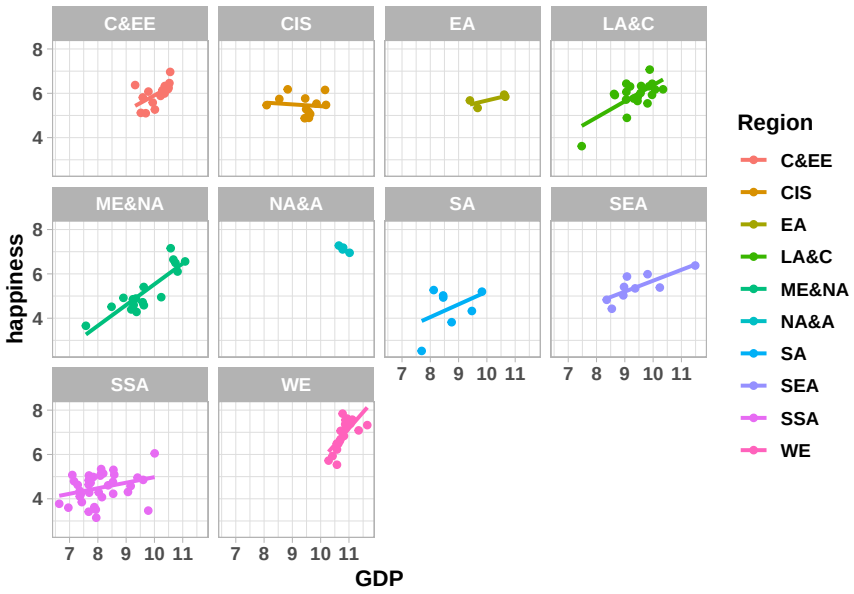


Figure 8: The relationship between happiness and GDP

3.2.2 Analysis of correlation between Happiness and Support

Figure 9 is a scatterplot of happiness scores and support scores in ten regions of the map. ME&NA,SEA,WE,NA&A show strong positive correlation,and most of the rest are Moderate Correlation (Because some regions only have few countries and the scale of the data is not very large to check if it is positive correlation). Therefore,it is necessary to use a linear regression model to evaluate the relationship between support and happiness.

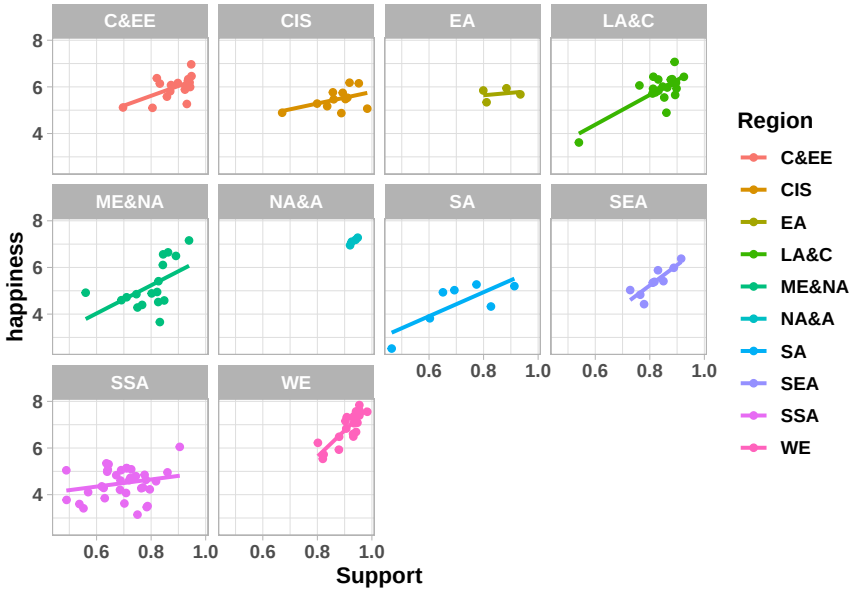


Figure 9: The relationship between happiness and support

3.2.3 Analysis of correlation between Happiness and Expectancy

Figure 10 is a scatterplot of happiness scores and Expectancy in ten regions of the map. In most of the region, the Expectancy and happiness scores show a linear positive relationship. Therefore, it is necessary to use a linear regression model to evaluate the relationship between Expectancy and happiness.

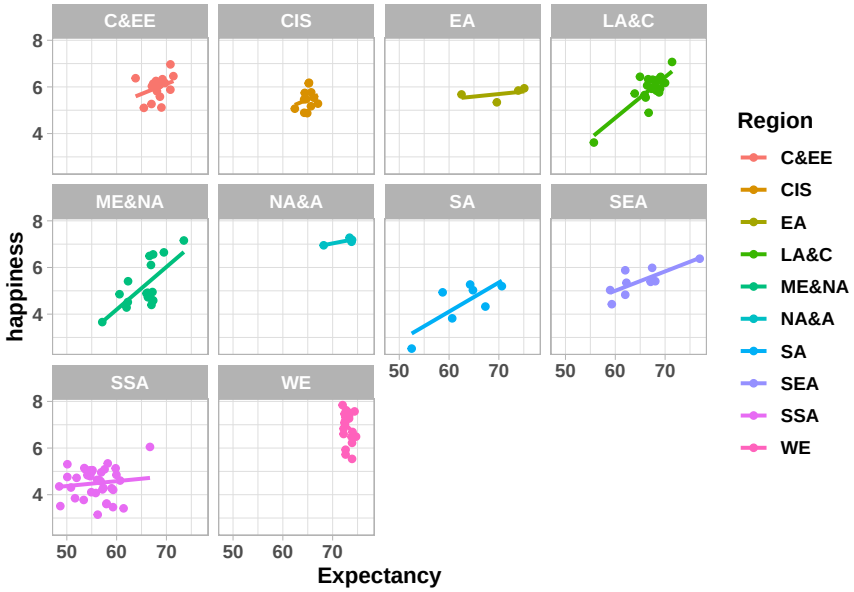


Figure 10: The relationship between happiness and Expectancy

3.2.4 Analysis of correlation between Happiness and Freedom

Figure 11 is a scatterplot of happiness scores and Freedom in ten regions of the map. In most of the region, the Freedom and happiness scores show a linear positive relationship. Therefore, it is necessary to use a linear regression model to evaluate the relationship between Freedom and happiness.

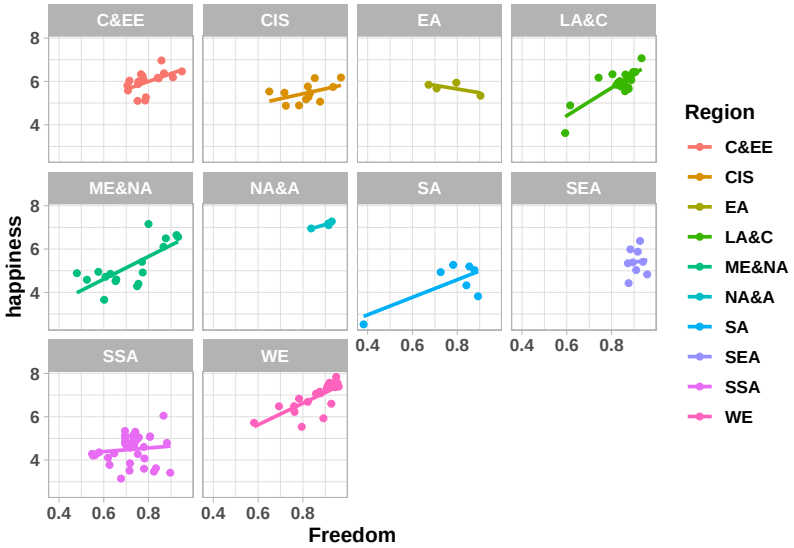


Figure 11: The relationship between happiness and Freedom

3.2.5 Analysis of correlation between Happiness and Generosity

Figure 12 is a scatterplot of happiness scores and Generosity in ten regions of the map. In most of the region, the Generosity and happiness scores show stable lines. Therefore, this variable may not related to the happiness and we should do more analysis to check it.

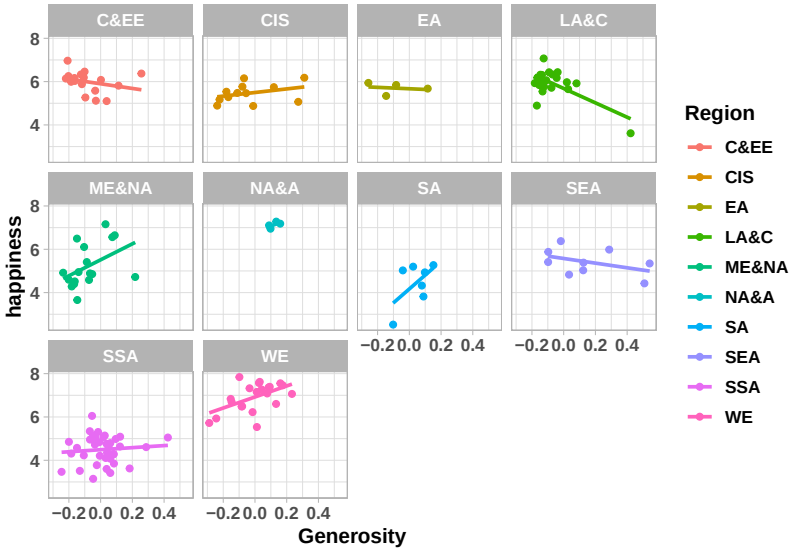


Figure 12: The relationship between happiness and Generosity

3.2.6 Analysis of correlation between Happiness and Corruption

Figure 13 is a scatterplot of happiness scores and Corruption in ten regions of the map. We can see the regions ME&NA,SA and WE show obvious negative relationship. Except region SSA an LA&C, other regions also have a slightly negative correlation. Therefore, we can try to use a linear regression model to evaluate it.

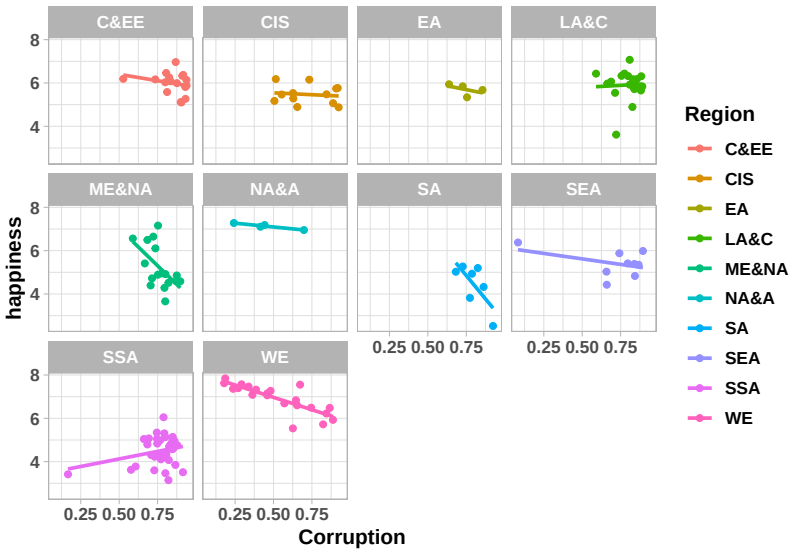


Figure 13: The relationship between happiness and Corruption

3.3 Analysis of correlation between Happiness and its impacted factors

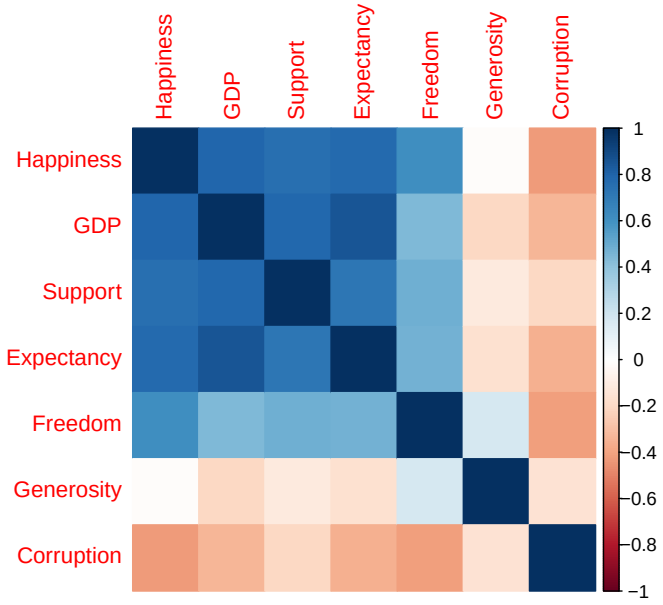


Figure 14: The relationship of factors related to happiness

Apparently, "Corruption" is negatively correlated with all the other numerical variables.One possible explaination here this could be, the lower the Corruption, the higher the happiness value.Generosity seems to have weak negative relationship with happiness. Other variables show positive relation with happiness score. The concrete evidence of this can be shown in Formal Data Analysis.

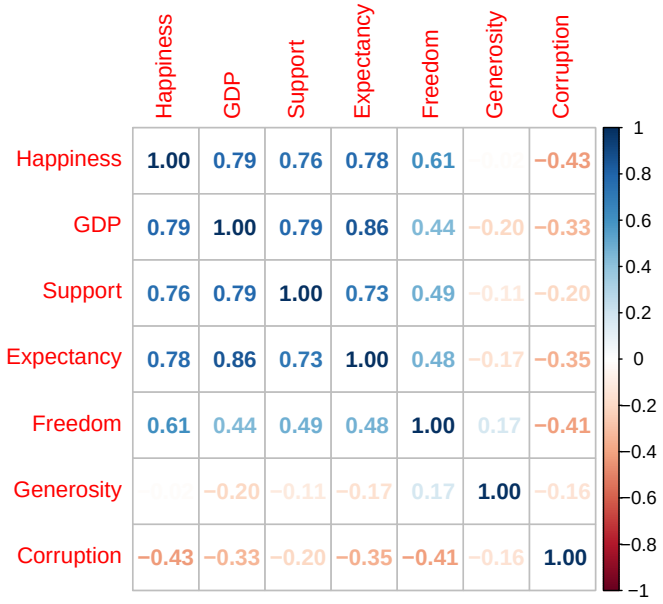


Figure 15: The relationship of correlation coefficient of happiness

4 Formal Data Analysis

4.1 Analysis of correlation Test

Firstly, we can see the correlation between Continuous independent variables GDP, Support, Expectancy, Freedom, Generosity, Corruption and dependent variable Happiness.

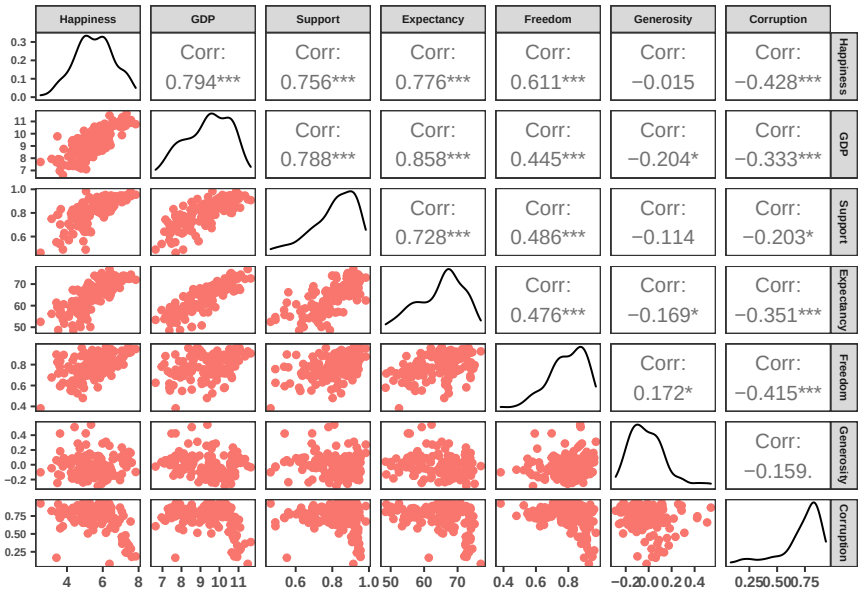


Figure 16: Correlation matrix of data

According to the correlation matrix, we can see there is a positive relationship between GDP, Support, Expectancy, Freedom and Happiness. And the correlation coefficient of Generosity and Happiness equal to - 0.015 which are not very related to Happiness and among the other 5 continuous variables, some have a high correlation coefficient with others which may cause a multicollinearity problem, so it is necessary to do a Multicollinearity test to choose which variables can be used.

4.2 Analysis of multiple collinear test

So we do the test of vif and Here is the result.When VIF>10, there is a serious problem of collinearity between the variables. According to the above table, there is no VIF value greater than 10. Therefore, there is no collinearity among these variables.

Table 11: VIF value Table

	GVIF	Df	GVIF^(1/(2*Df))
Region	9.984789	9	1.136368
GDP	5.741301	1	2.396101
Support	3.384953	1	1.839824
Freedom	1.922859	1	1.386672
Corruption	1.858462	1	1.363254
Expectancy	6.530983	1	2.555579

4.3 Analysis of linear-regression

To begin our analyze of happiness score formally, we fit the following linear model to the data.

$$\widehat{\text{Happiness}}_{\text{ij}} = \widehat{\alpha}+\widehat{\beta}_1\cdot\mathbb{I}_{\text{Region}}(i)+\widehat{\beta}_2\cdot\text{GDP}_{\text{j}}+\widehat{\beta}_3\cdot\text{Support}_{\text{j}}+\widehat{\beta}_4\cdot\text{Expectancy}_{\text{j}}+\widehat{\beta}_5\cdot\text{Freedom}_{\text{j}}+\widehat{\beta}_6\cdot\text{Corruption}_{\text{j}}\epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2)$$

where

the intercept $\widehat{\text{Happiness}}_{\text{ij}}$ is the expected value of the happiness in the ith region and of the jth independent variables in the sample;

the intercept $\widehat{\alpha}$ is the mean Happiness for the baseline category of C&EE which represents Central and Eastern Europe;

$\widehat{\beta}_1$ is the difference in the mean Happiness of 9 other regions;

$\widehat{\beta}_{\text{2-6}}$ is the value of Happiness that changes as the corresponding independent variable changes by one unit;

$\text{variables}_{\text{j}}$ is the jth observations of the continuous variables;

and

$\mathbb{I}_{\text{Region}}(i)$ is an indicator function such that

$$\mathbb{I}_{\text{Region}}(i) = \begin{cases} 1 & \text{if region of } j\text{th observations belongs to the Region}(i), \\ 0 & \text{if region of } j\text{th observations belongs to the region other than Region}(i). \end{cases}$$

But now we have to test if all the four continuous variables in can make the model best using the p-value to check their significance and AIC test and other criterion to solve over-fitting problems.

Table 12: Table of models of the minimum of AIC(Top 10)

Model	AIC	Cp	R2adj
Region GDP Support Freedom Corruption	227.0778	-2.068012	0.7870758
Region GDP Support Freedom	227.6042	-1.763669	0.7850013
Region GDP Support Freedom Corruption Expectancy	228.0436	-1.000000	0.7869669
Region GDP Support Freedom Expectancy	228.1084	-1.132749	0.7855777
Region Support Freedom Corruption Expectancy	233.9977	4.339305	0.7768129
Region Support Freedom Expectancy	234.2100	4.452170	0.7751194
Region GDP Freedom	234.6410	4.810811	0.7730632
Region GDP Freedom Expectancy	234.7018	4.926178	0.7743658
Region GDP Freedom Corruption	235.5653	5.762320	0.7730365
Region GDP Freedom Corruption Expectancy	235.9414	6.193889	0.7738423

According to the table which contains best 10 models, the first two models have lower AIC and higher R2adj. Next we have to check the significance of these four continuous variables.

$$\widehat{\text{Happiness}}_{\text{ij}} = \widehat{\alpha}+\widehat{\beta}_1\cdot\mathbb{I}_{\text{Region}}(i)+\widehat{\beta}_2\cdot\text{GDP}_{\text{j}}+\widehat{\beta}_3\cdot\text{Support}_{\text{j}}+\widehat{\beta}_4\cdot\text{Freedom}_{\text{j}}+\widehat{\beta}_5\cdot\text{Corruption}_{\text{j}}\epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2)$$

We choose the first model on the data first, the following estimates of α (intercept) and $\beta_{\text{1-5}}$ are returned:

Table 13: Estimates of the parameters from the fitted linear regression model 1st.

term	estimate	p_value
intercept	-0.029	0.969
RegionCIS	-0.405	0.044
RegionEA	-0.202	0.471
RegionLA&C	0.109	0.531
RegionME&NA	-0.313	0.087
RegionNA&A	0.430	0.155
RegionSA	-0.734	0.004
RegionSEA	-0.589	0.009
RegionSSA	-0.420	0.036
RegionWE	0.382	0.047
GDP	0.267	0.001
Support	2.048	0.002
Freedom	2.381	0.000
Corruption	-0.473	0.131

We noticed that the p-value of "Corruption" is not significant(0.131), so now we can try the 2nd model:

$$\widehat{\text{Happiness}}_{\text{ij}} = \widehat{\alpha} + \widehat{\beta}_1 \cdot \mathbb{I}_{\text{Region}}(i) + \widehat{\beta}_2 \cdot \text{GDP}_{\text{j}} + \widehat{\beta}_3 \cdot \text{Support}_{\text{j}} + \widehat{\beta}_4 \cdot \text{Freedom}_{\text{j}} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

Table 14: Estimates of the parameters from the fitted linear regression model 2nd.

term	estimate	p_value
intercept	-0.630	0.344
RegionCIS	-0.345	0.080
RegionEA	-0.149	0.594
RegionLA&C	0.128	0.462
RegionME&NA	-0.258	0.152
RegionNA&A	0.591	0.039
RegionSA	-0.712	0.005
RegionSEA	-0.556	0.013
RegionSSA	-0.366	0.064
RegionWE	0.516	0.003
GDP	0.279	0.000
Support	1.887	0.004
Freedom	2.654	0.000

we can see that the variables: GDP, Support, Freedom are all significant, so we prefer to select this model as our final model.

4.4 Plots for checking model assumptions

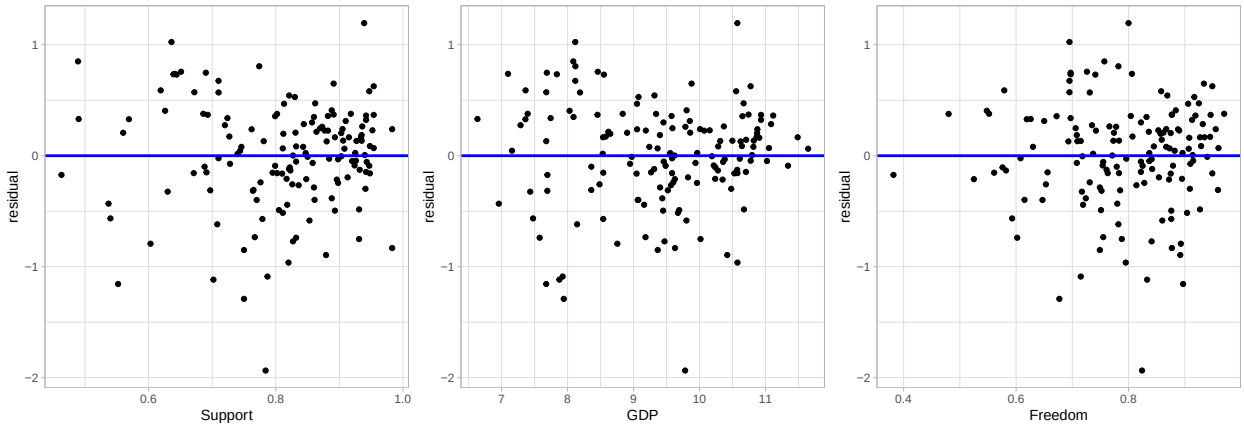


Figure 17: plots of the Residual.

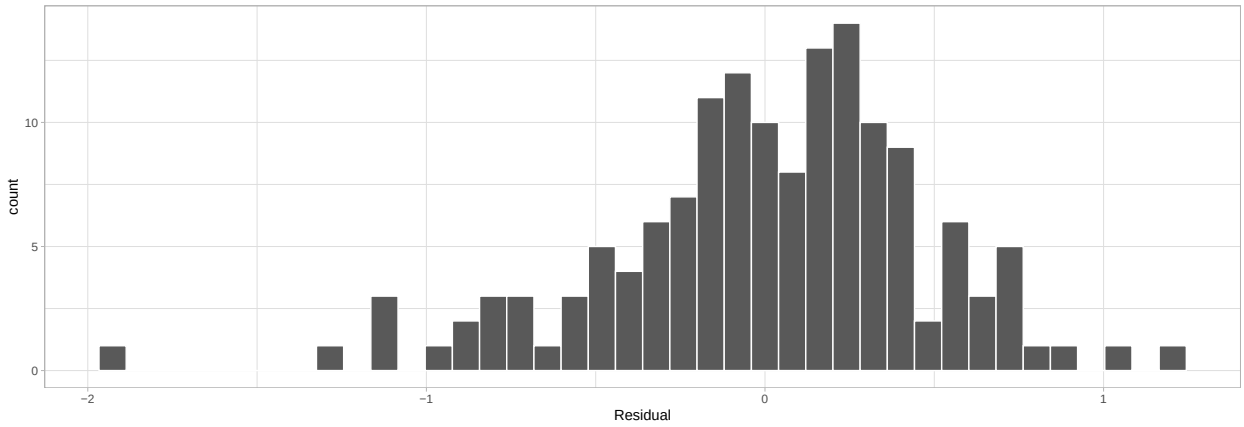


Figure 18: plots of the Residual.

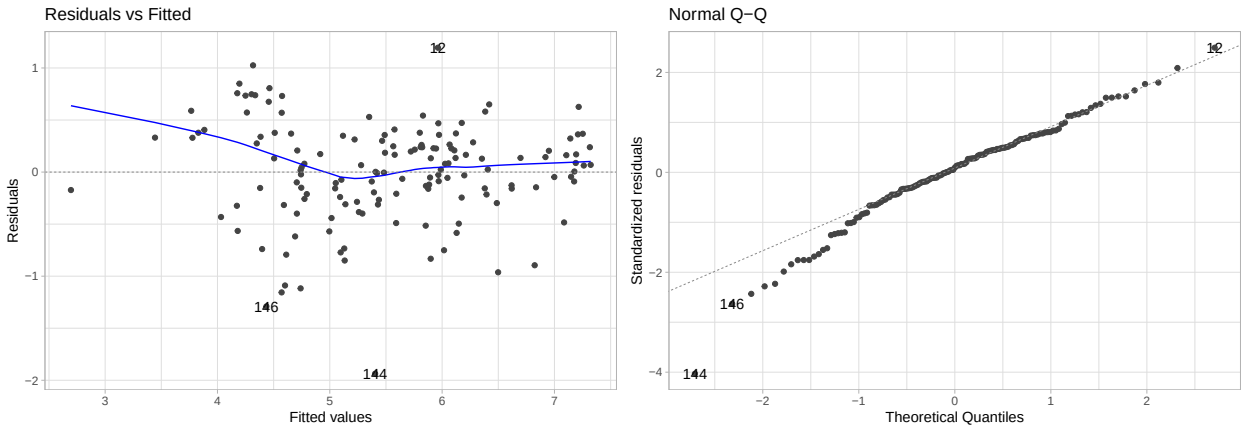


Figure 19: plots of the Residual.

The assumptions of the residuals having mean zero and constant variability across all values of the explanatory variable appear to be valid.

The residuals appear to be normally distributed and centred at zero.

The residuals vs fitted values plot that the residuals are randomly scatter around the zero line.

From the Normal Q-Q plot that the residuals are approximately consistent with a normal distribution.

5 Conclusions

Broadly, happiness is a measure of how happy people are in their lives. It contains many kinds of aspects, and we can summarize these into two main parts: economic level and spiritual level.

With a high level of GDP people can achieve materialistic satisfaction. For example, satisfying food and clothing, may help people pursue higher quality of life. And in linear model, European countries and North American countries have a higher happiness level. According to the "Global economic ranking", many European countries and North American countries come out top.

Apart from economical aspect like GDP, high spiritual level is also helpful for the increasing of happiness. Support and Freedom are two elements of spiritual well-being. A high support can help people for many kinds of things in their life. For example, government can help their children to have a good education; Government can help local people to find a job. Government can help people for their medical treatment and can help provide people with fundamental infrastructure and so on.

In a highly liberal country, people are more open while choosing what they want to do and therefore can achieve spiritual satisfaction with this Freedom. And in some "Relatively Closed countries" people are not free not for making their own choices. People residing in such countries absolutely cannot have a high happiness level. Hence, Freedom plays a vital role in elevating happiness.

In all, the region with high GDP, good support and high freedom score is bound to have high happiness score. We chose the region to be one categorical variable. But it's worth noting that some regions are significantly happier than others, such as North America and Western Europe; Some regions have significantly lower levels of happiness than others: South and Southeast Asia. This may be due to cultural differences, such as religious factors, that cause people in different regions to have different views of happiness. In addition, some countries score more than others because they might have formed complete social systems over hundreds of years, such as more humane employment system, retirement welfare and more habitable environment, which are more conducive to improving resident's happiness score , even though GDP, government support and freedom are the same.

6 Reference

1.Helliwell J F, Huang H, Wang S. The distribution of world happiness[J]. World Happiness, 2016, 8.

2.Lin K. Social quality and happiness—An analysis of the survey data from three Chinese cities[J]. Applied Research in Quality of Life, 2016, 11(1): 23-40.

3.Rowan A N. World Happiness 2021[J]. WellBeing News, 2021, 3(3): 3.