Analysis of World Happiness Report

Dikshant Gupta

29/03/2022

Contents

1 Background

We live in an age of stark contradictions. The world enjoys technologies of unimaginable sophistication; yet has at least one billion people without enough to eat each day. The world economy is propelled to soaring new heights of productivity through ongoing technological and organizational advance; yet is relentlessly destroying the natural environment in the process. Countries achieve great progress in economic development as conventionally measured; yet along the way succumb to new crises of obesity, smoking, diabetes, depression,and other ills of modern life.[1]

These contradictions would not come as a shock to the greatest sages of humanity, including Aristotle and the Buddha. The sages taught humanity, time and again, that material gain alone will not fulfill our deepest needs. Material life must be harnessed to meet these human needs, most importantly to promote the end of suffering, social justice, and the attainment of happiness.

Happiness has always been the objective of human beings,from the time of Aristotle who mentioned "Happiness is the meaning and purpose of life, the general and final goal of human existence". Later, Maslow (1943) defined happiness as the self fulfillment that individuals achieve after satisfying, partially or totally, certain hierarchically ordered needs[2]

Welfare and well-being have traditionally been gauged by using income and employment statistics, life expectancy, and other objective measures. The Economics of Happiness, which is based on people's reports of how their lives are going, provides a complementary yet radically different approach to studying human well-being. Typically, subjective well-being measures include positive and negative feelings (e.g., momentary experiences of happiness or stress), life evaluations (e.g., life satisfaction), and feelings of having a life purpose. [3]

Similarly, happiness is gaining more and more attention from politicians and decision makers in both developed and developing countries in search of answers. Social scientists often recommend that measures of subjective well-being should augment the usual measures of economic prosperity, such as GDP per capita.[4] Questions such as: What factors influence social well-being to be included in public policy guides? Is GDP the only factor?

Recent authors state that happiness is a much more complete measure than GDP because it focuses only on production and income and is related to the economic aspects of life. [4]

2 Data description

Since 2012, the World Happiness Report has been published [5].This report shows the state of happiness of 156 participating countries.

From Kaggle World Happiness Report of 2021 is being used.This dataset uses the data from the Gallup World Poll.Data is collected from people in over 156 countries. Each variable measured reveals a populated-weighted average score on a scale running from 0 to 10 that is tracked over time and compared against other countries.The rankings of national happiness are based on a Cantril ladder survey. Nationally representative samples of respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. Apart from happiness score which is called Ladder.score in dataset scoring is done based on the following features :

1. GPD per capita
2. Social support
3. Healthy life expectancy
4. Freedom to make life choices
5. Generosity
6. Perception of corruption

Ladder.score.in.Dystopia tell happiness of Dystopia. Dystopia is an imaginary country that has the world's least-happy people. The purpose in establishing Dystopia is to have a benchmark against which all countries can be favorably compared (no country performs more poorly than Dystopia) in terms of each of the six key

variables, thus allowing each sub-bar to be of positive (or zero, in six instances) width. The lowest scores observed for the six key variables, therefore, characterize Dystopia.

Residual which refer to unexplianed components is also another factor , These residuals have an average value of approximately zero over the whole set of countries.

Following are the short data description of all the variables included in the analysis.

Column Names	Description
Country.name	The name of the country
Regional.indicator	Region that specific country belongs to
Ladder.score	Happiness scores based on Contril ladder from 0 to 10
Standard.error.of.ladder.score	Standard Error in Happiness Score
upperwhisker	Upper Confidence Interval of the Happiness Score
lowerwhisker	Lower Confidence Interval of the Happiness Score
Logged.GDP.per.capita	GDP per capita of each country in terms of purchasing power parity (PPP) (in USD).
Social.support	Individual rating that determines whether, when you have problems, your family or friends would help you. Binary responses (0 or 1).
Healthy.life.expectancy	Healthy life expectancy at birth is based on data from the World Health Organization (WHO)
Freedom.to.make.life.choices	Individual rating that determines whether you are satisfied or dissatisfied with your freedom to choose what you do with your life. Binary responses (0 or 1).
Generosity	Generosity is the residual from the regression of the national mean of responses to the question "Have you donated money to a charity in the last month?" on GDP per capita.
Perceptions.of.corruption	Average of binary responses to two GWP questions: corruption in government and corruption in business.
Ladder.score.in.Dystopia	Happiness score in hypothetical country
Explained.by..Log.GDP.per.capita	Ranking of countries based on Logged.GDP.per.capita
Explained.by..Social.support	Ranking of countries based on Social.support
Explained.by..Healthy.life.expectancy	Ranking of countries based on Healthy.life.expectancy
Explained.by..Freedom.to.make.life.choices	Ranking of countries based on Freedom.to.make.life.choices
Explained.by..Generosity	Ranking of countries based on Generosity
Explained.by..Perceptions.of.corruption	Ranking of countries based on Perceptions.of.corruption
Dystopia. . . residual	The residuals, or unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average 2019-2021 life evaluations.

3 Methods

3.1 Data cleaning

Data cleaning is required for checking missing values and making data more structurally correct .

3.1.1 Missing data

Calculating the percentages of missing values in each of the column

```
#check if there is any null data in the dataset
na_percentages <- data.frame(Feature = colnames(full_df),
```

```
Percentage_of_NA = colSums(is.na(full_df))/nrow(full_df)* 100)
rownames(na_percentages) <- NULL
kable(na_percentages, format = "markdown",caption = "Percentage of NA in the dataset")
```

Table 2: Percentage of NA in the dataset

Feature	Percentage_of_NA
Country.name	0
Regional.indicator	0
Ladder.score	0
Standard.error.of.ladder.score	0
upperwhisker	0
lowerwhisker	0
Logged.GDP.per.capita	0
Social.support	0
Healthy.life.expectancy	0
Freedom.to.make.life.choices	0
Generosity	0
Perceptions.of.corruption	0
Ladder.score.in.Dystopia	0
Explained.by..Log.GDP.per.capita	0
Explained.by..Social.support	0
Explained.by..Healthy.life.expectancy	0
Explained.by..Freedom.to.make.life.choices	0
Explained.by..Generosity	0
Explained.by..Perceptions.of.corruption	0
Dystopia. . . residual	0

3.1.2 Feature Engineering and Selection

I will be not using features 'Explained by' as they have no impact on the total score reported for each country of happiness . I will be only using raw features (Country.name, Region.indicator,Logged.GDP.per.capita, Social.support , Healthy.life.expectancy ,Freedom.to.make.life.choices, Generosity, Perceptions.of.corruption) and dependent variable Ladder.score . I will be renaming the features for better manipulation

```
#Selecting columns from main dataset
full_df_final <- subset(full_df, select=c("Country.name", "Regional.indicator",
                                          "Ladder.score", "Logged.GDP.per.capita",
                                          "Social.support", "Healthy.life.expectancy",
                                          "Freedom.to.make.life.choices", "Generosity",
                                          "Perceptions.of.corruption"))
#Renaming the column
full_df_final <- full_df_final %>% rename( Country=Country.name, Region=Regional.indicator,
                                           Happiness_score=Ladder.score,
                                           GDP=Logged.GDP.per.capita,
                                           Social_support=Social.support,
                                           Life_expectancy=Healthy.life.expectancy,
                                           Freedom=Freedom.to.make.life.choices,
                                           Corruption=Perceptions.of.corruption)
```

3.2 Hypothesis Testing

3.2.1 Hypothesis test for South Asia region and Europe region happiness

Filtering Europe and South Asia region data from the dataset

```
#Filtering European region and South Asia region from all the regions
American_region <- subset(full_df_final, Region == "North America and ANZ")
southasianregion <- subset(full_df_final,Region == "South Asia")
american_region_data <-  rbind(American_region,southasianregion)
```

Getting the summary statistics of data -

```
#Summary stats of the data
summary_stat_american <- group_by(american_region_data, Region) %>%
  summarise(count = n(), mean = mean(Happiness_score, na.rm = TRUE),
            sd = sd(Happiness_score, na.rm = TRUE))
```

Checking for outliers in the data set

```
#outliers check
american_region_outliers <- identify_outliers(group_by(american_region_data, Region)
                                              , Happiness_score)
```

Running Shapiro test to check the normality assumption of the Happiness

```
#shapiro test
american_shapiro_test_result <- shapiro_test(group_by(american_region_data, Region)
                                             , Happiness_score)
```

Running Levene's test to check the assumption of homogeneity of variance.

```
#Levene Test
american_levtest <- levene_test(Happiness_score ~ Region, data = american_region_data)
```

Levene's test shows the p-value of 0.2083197 and if p is $>$ .05, variances are homogeneous.

Performing T test -

```
#t-test
t_result <- t_test(american_region_data, Happiness_score ~ Region, var.equal=TRUE)
```

```
#adding significance to t test
add_significance_result <- add_significance(t_result)
```

Running post hoc test of cohen's d

```
#cohen d test
cohen_result <- cohens_d(american_region_data, Happiness_score ~ Region, var.equal=TRUE)
```

Scatter plot between GDP and Happiness:

```
#Scatter plot between GDP and Happiness based on region
gdpplot <- full_df_final %>%
  ggplot(aes(Happiness_score,GDP))+
  geom_point(aes(color=Region))+
  labs(title = "Happiness_Index increases with GDP per Capita",
       subtitle = "Economic importance in the state of a country")+
  xlab("Happy Score / 10")+
  ylab("GDP per Capita")+
  theme_bw()
```

Box plot between American and ANZ region and South Asian Region and Happiness:

```r
#Box plot
bp <- ggboxplot(american_region_data, x = "Region", y = "Happiness_score",
          color = "Happiness_score",
           ylab = "Happiness_score", xlab = "Region")
```

Corelation plot

```r
#Corelation plot
crplt <- ggcorr(full_df_final,
       method = c("everything", "pearson"),
       size = 3, hjust = 0.88,
       low = 'steelblue', mid = 'white', high = "#923b39",
       label = TRUE, label_size = 4,
       layout.exp = 1) + theme(axis.text=element_text(colour="#923b39")) +
  labs(
        subtitle = 'Corelation matrix between Happiness and other indicators') +
  theme(plot.title = element_text(size=10),
        plot.subtitle = element_text(size = 10),
        legend.text = element_text(size = 10))
```

3.3 Linear Model

```r
#Hapiness linear model with corelated features
happiness_linear_model <- lm(data = full_df_final ,
formula = full_df_final$Happiness_score ~ full_df_final$GDP + full_df_final$Freedom)
happiness_linear_modelsummary <- summary(happiness_linear_model)
```

```r
#Hapiness linear model with only gdp
happiness_linear_model_gdp <- lm(data = full_df_final,
formula = full_df_final$Happiness_score ~ full_df_final$GDP)
happiness_linear_model_gdpsummary <- summary(happiness_linear_model_gdp)
```

4 Results

4.1 Running the hypothesis test

The goal is to test that whether the developed countries of American region and Australia-New Zeland Region are more happy than the developing nations of South Asia like India , Pakistan etc.

Hypothesis formulation -

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

Summary Stats of the data set

```
## # A tibble: 2 x 4
##   Region                count  mean    sd
##   <chr>                 <int> <dbl> <dbl>
## 1 North America and ANZ     4  7.13 0.138
## 2 South Asia                7  4.44 0.993
```

Looking for outliers

```
##  [1] Region          Country         Happiness_score GDP
##  [5] Social_support  Life_expectancy Freedom         Generosity
##  [9] Corruption      is.outlier      is.extreme
```

```
## <0 rows> (or 0-length row.names)
```

Running Shapiro Test for Normality

```
## # A tibble: 2 x 4
##   Region                variable        statistic      p
##   <chr>                 <chr>               <dbl>  <dbl>
## 1 North America and ANZ Happiness_score     0.985 0.932
## 2 South Asia            Happiness_score     0.838 0.0952
```

Running Levene's test of homogenity

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1     9      1.84 0.208
```

Running t-test

```
## # A tibble: 1 x 8
##   .y.           group1             group2       n1    n2 statistic    df       p
## * <chr>         <chr>              <chr>     <int> <int>     <dbl> <dbl>   <dbl>
## 1 Happiness_sc~ North America and~ South As~     4     7      5.26     9 5.21e-4
```

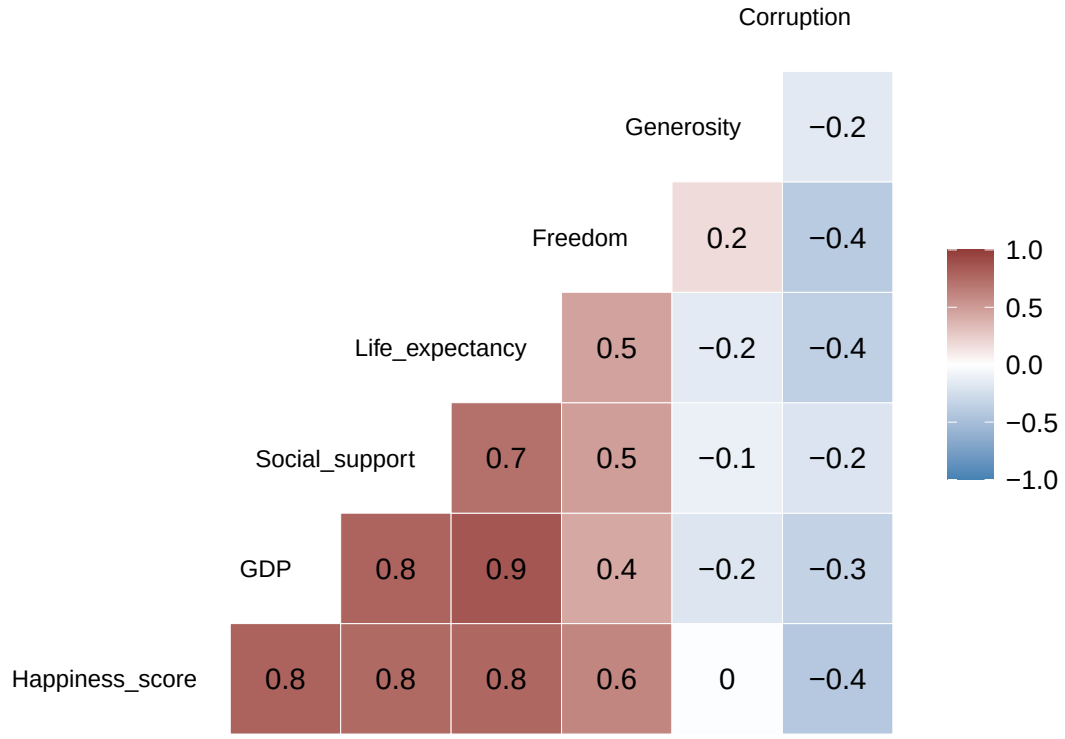
Adding significance to the result

```
## # A tibble: 1 x 9
##   .y.        group1        group2      n1    n2 statistic    df       p p.signif
## * <chr>      <chr>         <chr>    <int> <int>     <dbl> <dbl>   <dbl> <chr>
## 1 Happiness~ North Americ~ South A~     4     7      5.26     9 5.21e-4 ***
```

Running cohen's test

```
## # A tibble: 1 x 7
##   .y.             group1                group2     effsize    n1    n2 magnitude
## * <chr>           <chr>                 <chr>        <dbl> <int> <int> <ord>
## 1 Happiness_score North America and ANZ South Asia    3.30     4     7 large
```

Corelation matrix between Happiness and other indicators



Running Linear Model Multiple :-

```
##
## Call:
## lm(formula = full_df_final$Happiness_score ~ full_df_final$GDP +
##     full_df_final$Freedom, data = full_df_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37656 -0.40059  0.01935  0.45091  1.01389
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.59081    0.43280  -5.986 1.59e-08 ***
## full_df_final$GDP      0.60079    0.04574  13.135  < 2e-16 ***
## full_df_final$Freedom  3.10374    0.46758   6.638 5.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5813 on 146 degrees of freedom
## Multiple R-squared:  0.711,  Adjusted R-squared:  0.707
## F-statistic: 179.6 on 2 and 146 DF,  p-value: < 2.2e-16
```

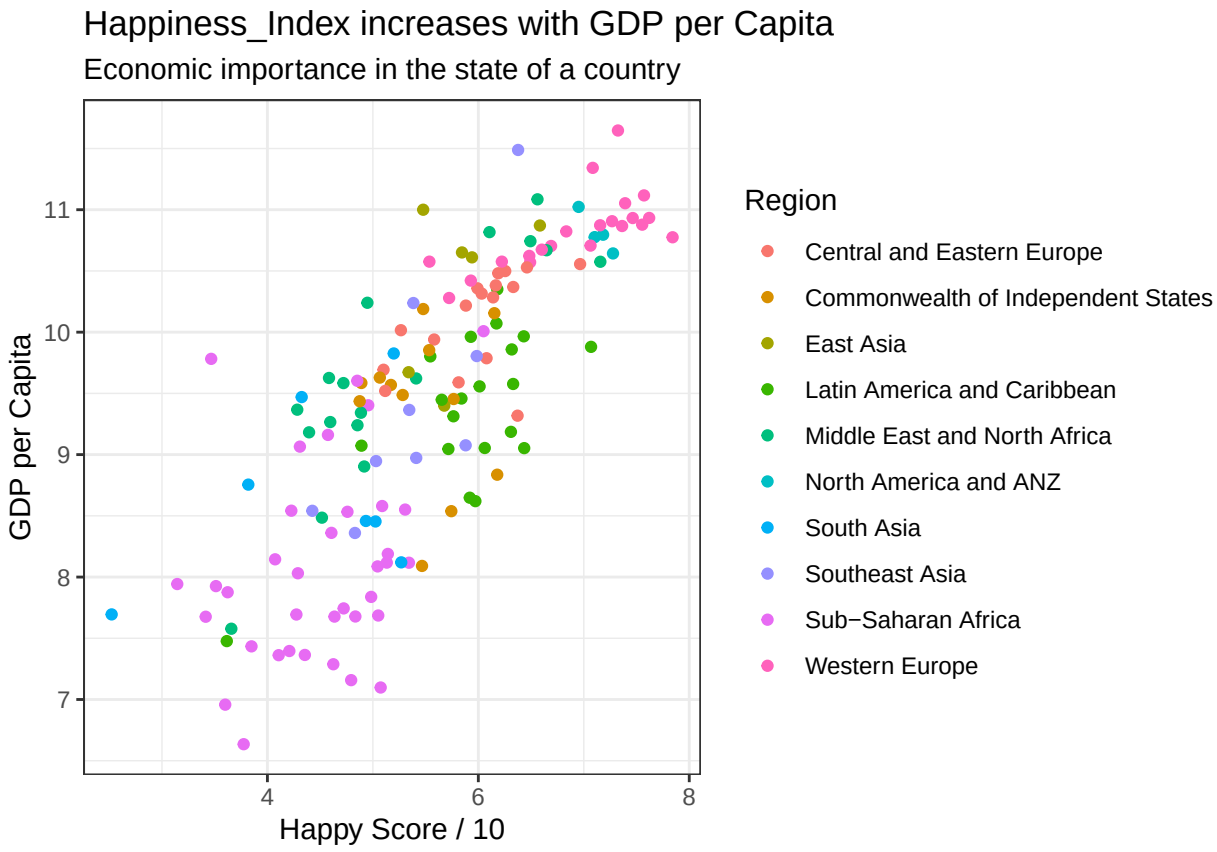
Running Linear Model single variable(GDP) :-

```
##
## Call:
## lm(formula = full_df_final$Happiness_score ~ full_df_final$GDP,
##     data = full_df_final)
```

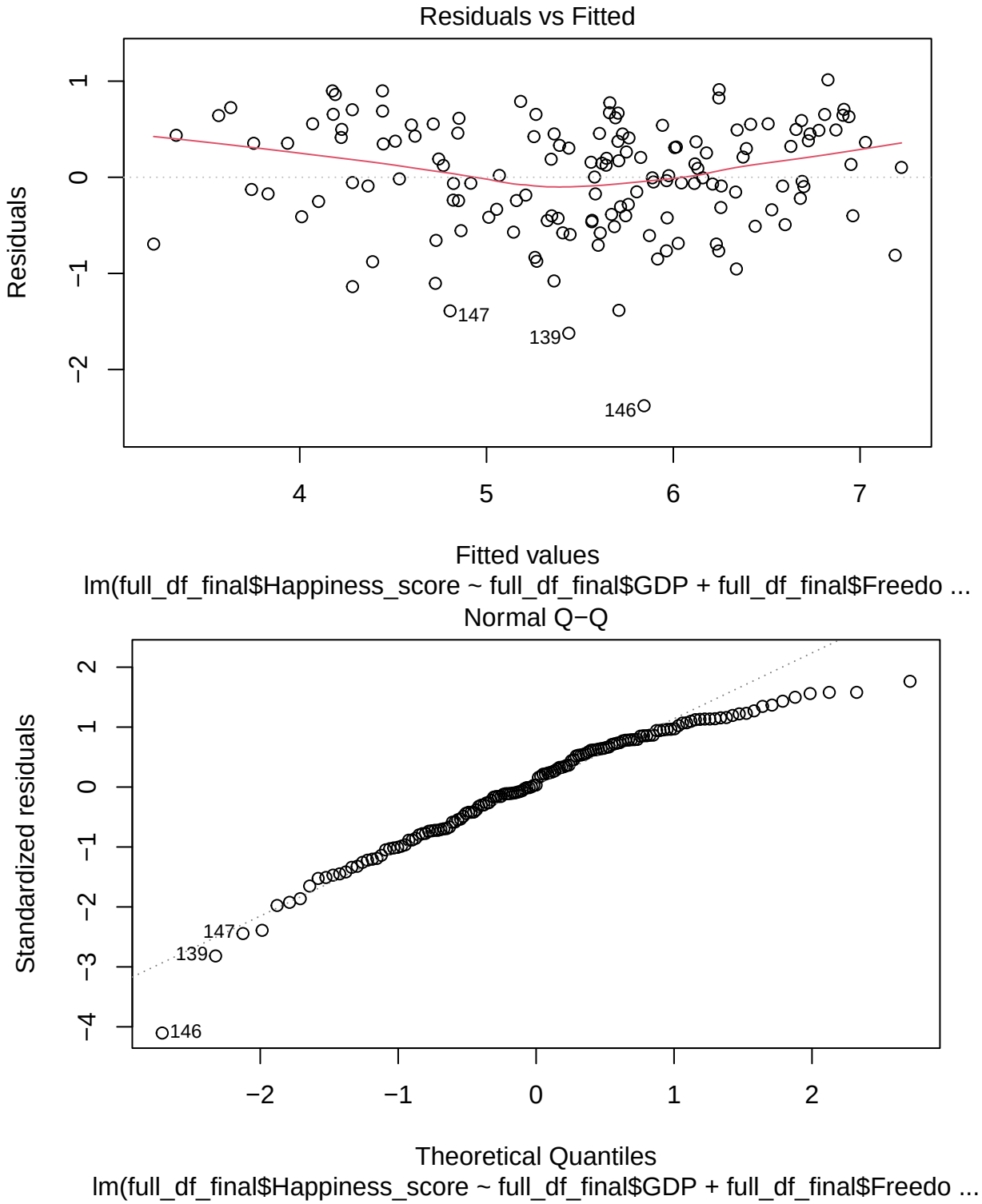
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32190 -0.46198  0.08206  0.50740  1.32618
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.3719     0.4456  -3.079  0.00248 **
## full_df_final$GDP   0.7320     0.0469  15.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.661 on 147 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6212
## F-statistic: 243.7 on 1 and 147 DF,  p-value: < 2.2e-16
```

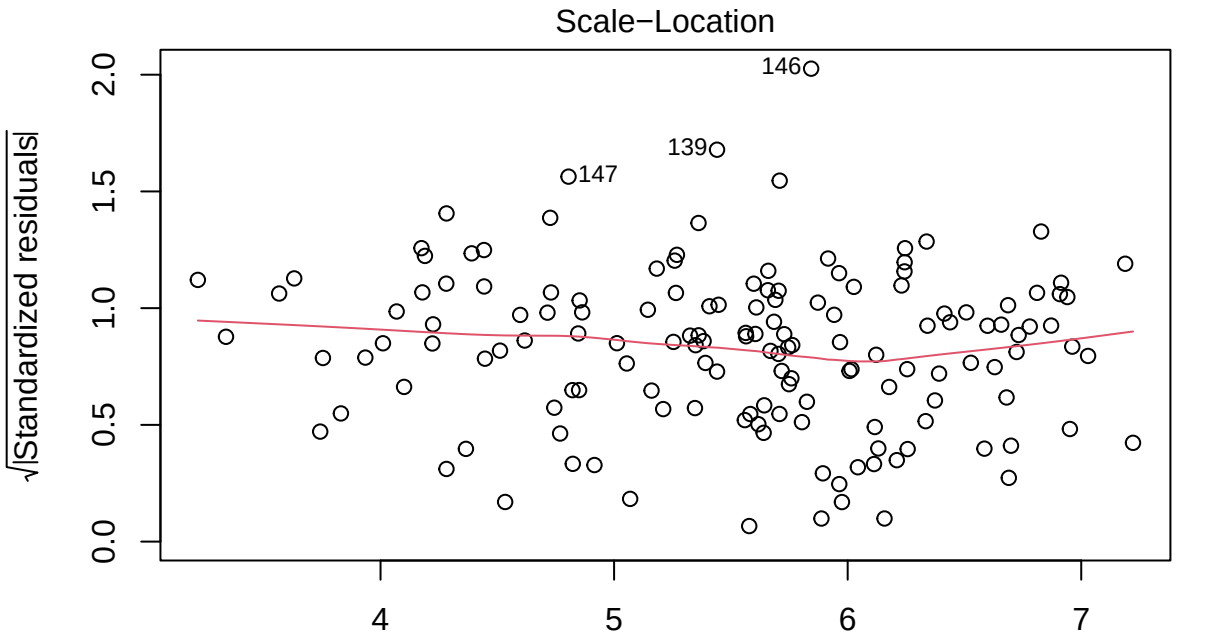
4.2 Visualizations

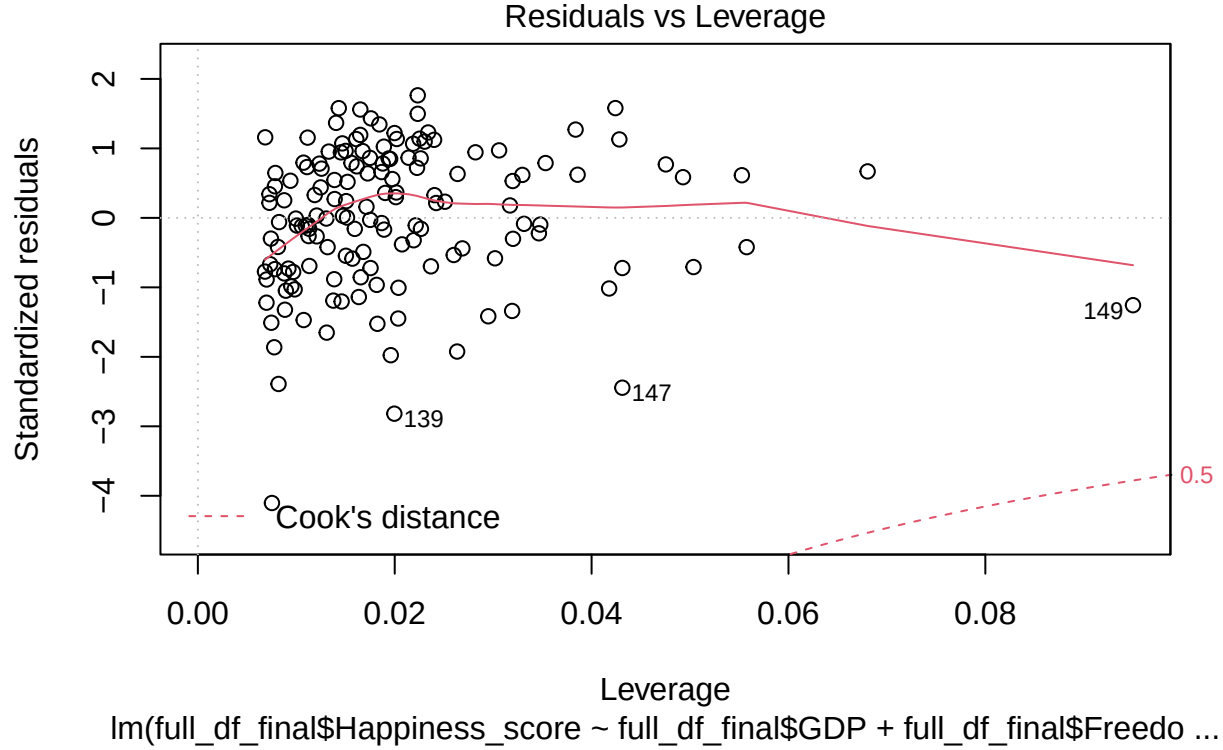
4.2.1 Scatter Plot between GDP and Happiness Score



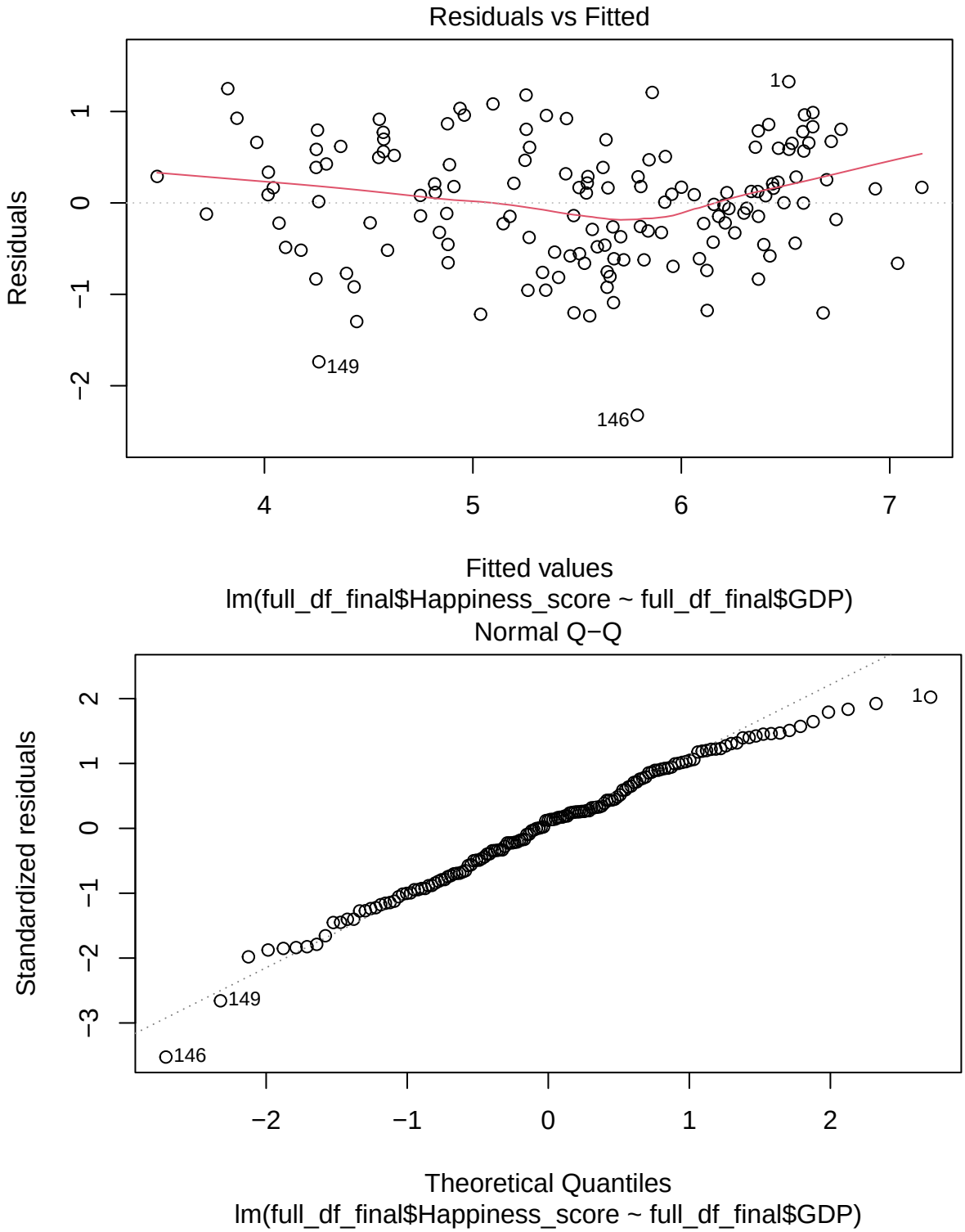
4.2.2 Residual, qq plot for Linear Model 1

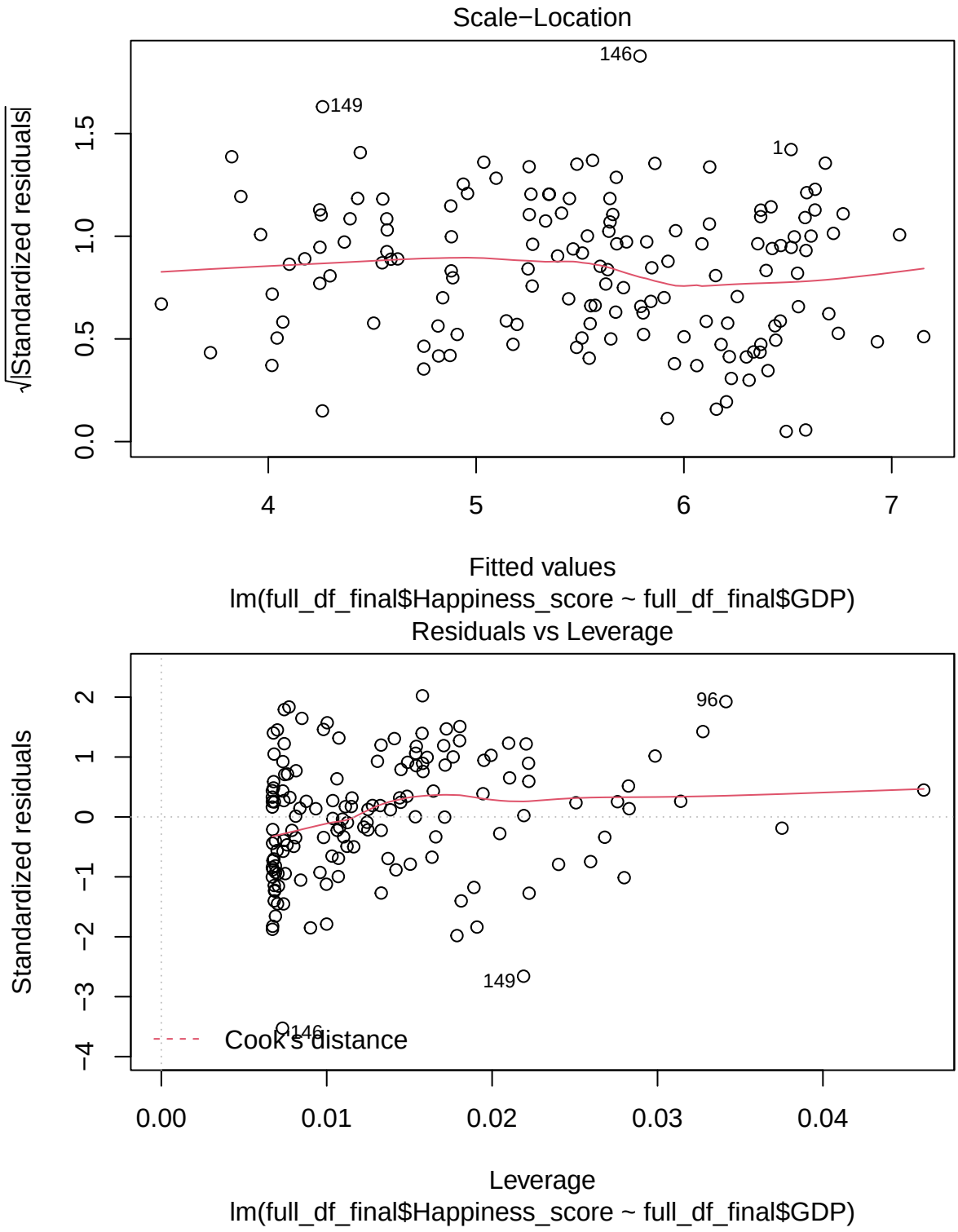



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
146
139
147
4
5
6
7
Fitted values
lm(full_df_final$Happiness_score ~ full_df_final$GDP + full_df_final$Freedo ...

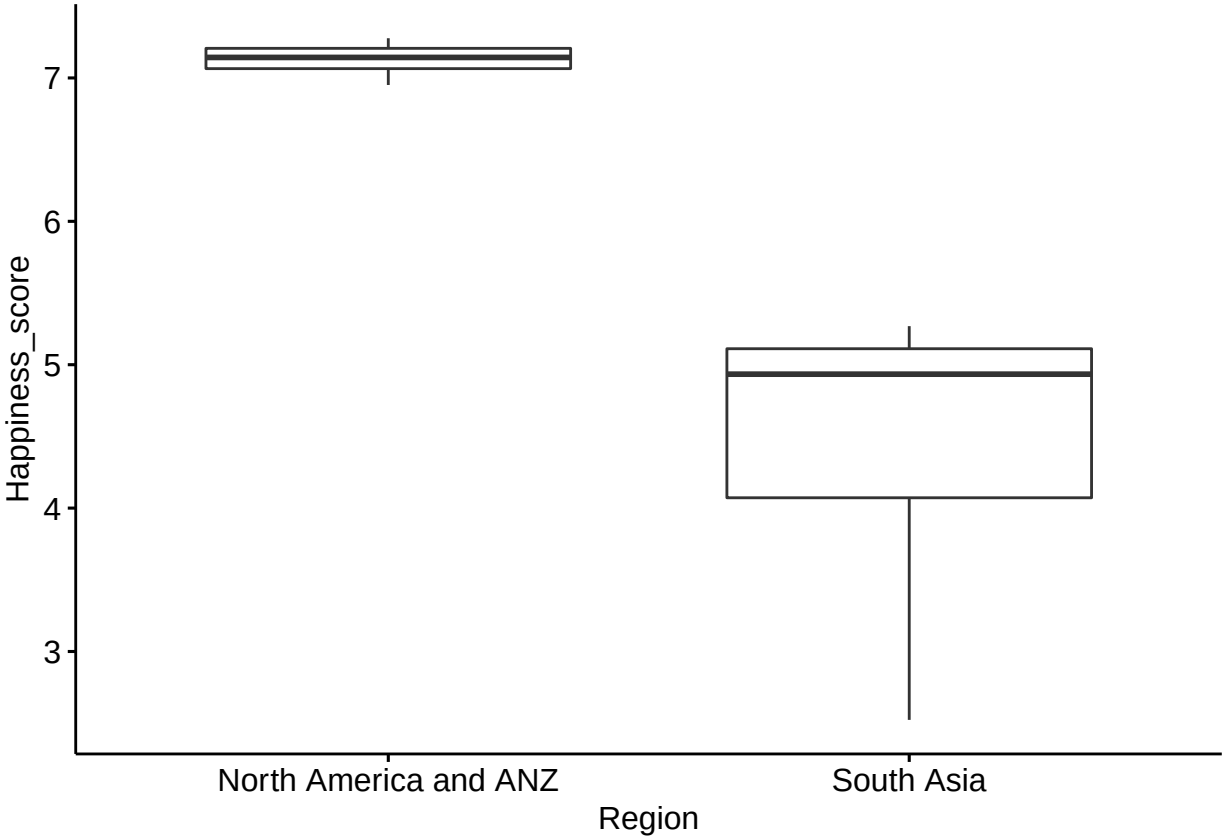

Residuals vs Leverage
Standardized residuals
−4
−3
−2
−1
0
1
2
149
147
139
Cook's distance
0.5
0.00
0.02
0.04
0.06
0.08
Leverage
lm(full_df_final$Happiness_score ~ full_df_final$GDP + full_df_final$Freedo ...

4.2.3 Residual, qq plot for Linear Model 2




Scale−Location
146
149
1
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
5
6
7
Fitted values
lm(full_df_final$Happiness_score ~ full_df_final$GDP)
Residuals vs Leverage
96
149
146
Cook's distance
Standardized residuals
−4
−3
−2
−1
0
1
2
0.00
0.01
0.02
0.03
0.04
Leverage
lm(full_df_final$Happiness_score ~ full_df_final$GDP)

4.2.4 Box Plot Happiness Score and Region



This study determines whether Developed nations of America and ANZ region are more happier than developing nations of South Asia . It also determines whether gdp factor is sufficient for happiness. The sample contained no extreme outliers. A Shapiro-Wilk test demonstrated normality by group, and Levene's test demonstrated homogeneity of variance.

The mean Happiness Score of the America and ANZ region in the sample was 7.128500 (SD = 0.1380568) whereas the mean Happiness Score of the South Asian region in the sample was 4.441857 (SD = 0.9934617). A Welch's independent t-test showed that the mean difference in Happiness Score between America and ANZ region and South Asian region in the sample was statistically significant, $t(1458) = 5.36$, $p < 0.0006$, $d = 3.296$, with America and ANZ region Happiness score be greater than Happiness Score of South Asian Region with large magnitude.

We also observe that GDP,Social Support,Life Expectancy and Freedom are the most highly co related values with Happiness Score . But GDP is also highly co related with Social Support and Life Expectancy . So to avoid multicollinearity we use only gdp and freedom as the independent varaibles for the formula for Happiness . In the 2nd model we only use GDP as the independent varaiable for our model . But by looking at Adjusted R square and F-statistic which is better for model 1(adjR =0.707) than model 2(adjR = 0.62) we can conclude that model 1 which has variable gdp and freedom is better than the model which has only GDP.

The residual vs fitted graph also shows that the Model with multiple factors fit more and has less errors than model with only gdp as factor. The scatter plot shows a linear relationship between gdp and happiness score with western europe having highest happiness and sub sahran African being the lowest. The box plot shows how the Happiness score is Higher in American and ANZ region and lower in South Asian region

5 Conclusion

We saw in background research how some initial papers suggested that GDP is the main factor in determining the happiness but recent studies and social well being reports has shown that the GDP is not an adequate factor in representing well being of a person in a country.In our analysis we found that GDP alone doesn't define happiness better, other factors such as freedom also contribute in the happiness of the country. The initial co relation plot showed us how Happiness Score is related strongly to GDP but deeper dive in linear model showed it is not the only factor although it is a strong factor in determining the happiness.We found many interesting insights such as how different region has different happiness and developed nations has more happiness score than the developing nations.We found out that western europe has highest happiness score whereas sub saharan African countries have low happiness score . Our hypothesis test revelead that American and ANZ region have more happiness score than the South Asia region .

More statistical and analytic tests can be conducted to find factors affecting happiness such as we can see how in low happiness scoring countries what factors more than the high scoring happiness countries . More predictive modelling can be done to find the interesting insights related to social support and life expatancy factors affecting the happiness score . Since we dont have vaccine data and covid deaths in the datset we can't determine how happiness was affected in the pandemic times due to these issues . It would be interesting to see if the vaccinated country is the happier one also ? Does imposing restrictions reduce the happiness of the people. It would be interesting to analyse these hypothesis for further studies.

References :

1. By, Edited & Helliwell, John & Layard, Richard & Sachs, Jeffrey. (2012). World Happiness Report.
2. Maslow, "A Theory of Human Motivation", Psychological Review, 50 (4),pp. 370-396, https://doi.org/10.1037/h0054346
3. E. Diener, M. Tamir and C. Scollon, "Happiness, life satisfaction, and fulfillment: The social psychology of subjective well-being.", 2006
4.J. Ott, "Beyond Economics, happiness as a standard in our personal life and politics" pp71, 2020
5.World Happiness Report https://worldhappiness.report/