Understanding the Happiness Score in 2020

Prepared by Kayde Varona

I. Introduction

2020 has been a challenging year, mainly because of the Covid-19. It affects everyone and posted economic instability, social distress, and emotional challenges worldwide. This project wants to understand what affects the world's happiness score during the pandemic, specifically during 2020. This study is inspired by the World Happiness Report 2021, where they compared happiness with government trust and Covid-19 related deaths. Click this link (https://worldhappiness.report/ed/2021/) for more information regarding this report.

This project aims to understand the factors that affect the happiness score in 2020. This project focused mainly on answering three questions:

- First, what are the factors that affect the happiness score in 2020?
- Second, is there a strong relationship between COVID factors and the happiness score?
- Third, do countries with higher GDP per Capita tend to have a higher happiness score for 2020?

The results of the analysis shows that:

- GDP per Capita, social support, freedom to make life choices, perception to corruption and COVID-19 related deaths are affects the happiness score.
- There is a relationship between COVID-19 to the happiness score but it is not strong.
- Countries with higher GDP have higher happiness scores compared to low GDP regions.

II. Data Sources and Variable Definitions

The data was gathered from the appendix of the World Happiness Report 2021.

a. The significant variables used in this study are

- **ladder score** – this is utilized as the dependent variable for this study. This is also known as the happiness score measured by subjective well-being, from the Gallup World Poll (GWP). The top ladder being 10 represents the "best possible" life and 0 as the worst.
- **log gdp per capita** – is the Gross Domestic Product per capita is countries' economic performance. It is to compare average standard living between countries
- **freedom** – or freedom to make life choices - this is the national average response to the GWP question to "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
- **social support** – this is a 0 or 1 response to the GWP survey question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?". This is computed as the national average.
- **healthy life** – healthy life expectancies at birth from the World Health Organization
- **covid deaths** – Covid-19 deaths per 100,000 population in 2020
- **corruption** – also known as corruption perception –is from the GWP survey averaging the national response from business and government perception.

b. Two variables considered factors or characters for this project are:

- **country** – this refers to 149 countries included in this study. Countries that had nulls are removed from the dataset
- **region** – the countries are grouped by areas, namely Western Europe, Central and Eastern Europe, Commonwealth of Independent States, Southeast Asia, South Asia, East Asia, Latin America and Caribbean, North America and ANZ, Middle East and North Africa, and Sub-Saharan Africa

c. Other variables are used in the model but not considered significant after the regression model are the following:

- **generosity** – this is the national average response for the GWP survey question about donating money to the charity.
- **gini index** – represents the gini coefficient of all countries – used to compare income inequalities of the countries
- **exposure index** – is one country's exposure to infections in other countries. It is the sum of all other countries' infections weighted by the inverse of the bilateral distance of their capital cities.

III. Data Preparation

There are going to be two parts of data preparation to answer the three main objectives of this project. The first part is for the hypothesis testing. The second part for the regression analysis. Before importing the CSV file, the mortality and happiness data were combined into a single table. In addition to that, the following columns from the happiness table were removed: standard error of ladder score, upper whisker, lower whisker, and ladder score in dystopia from the happiness table. Similarly, the columns population 2020, median age, island, log of the average distance to SARS countries, WHO Western Pacific Region and female head of government were removed from the mortality table.

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.5     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.0.2     v forcats 0.5.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

Rows: 148 Columns: 13

-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): country, region
dbl (11): ladder_score, log_gdp, social_support, healthy_life, freedom, gene...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

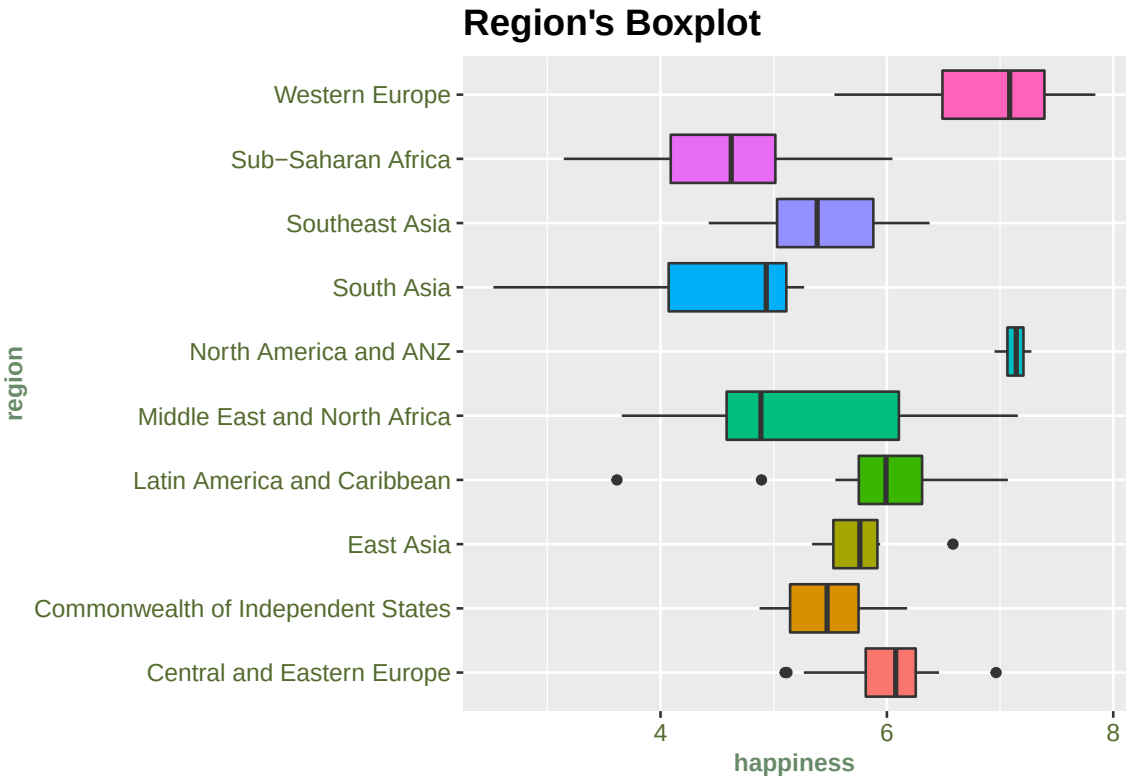
```
Rows: 148
Columns: 13
$ country                   <chr> "Finland", "Denmark", "Switzerland", "Icelan~
$ region                    <chr> "Western Europe", "Western Europe", "Western~
$ ladder_score              <dbl> 7.842, 7.620, 7.571, 7.554, 7.464, 7.392, 7.~
$ log_gdp                   <dbl> 10.775, 10.933, 11.117, 10.878, 10.932, 11.0~
$ social_support            <dbl> 0.954, 0.954, 0.942, 0.983, 0.942, 0.954, 0.~
$ healthy_life              <dbl> 72.000, 72.700, 74.400, 73.000, 72.400, 73.3~
$ freedom                   <dbl> 0.949, 0.946, 0.919, 0.955, 0.913, 0.960, 0.~
$ generosity                <dbl> -0.098, 0.030, 0.025, 0.160, 0.175, 0.093, 0~
$ corruption                <dbl> 0.186, 0.179, 0.292, 0.673, 0.338, 0.270, 0.~
$ covid_deaths              <dbl> 10.1250, 22.4094, 88.3343, 8.4982, 67.2605, ~
$ exposure_index            <dbl> 2.2250834, 3.9530561, 5.5835218, 1.6493348, ~
$ gini_index                <dbl> 25.90000, 27.80000, 30.10000, 24.10000, 27.0~
$ institutional_trust_index <dbl> 2.2250834, 3.9530561, 5.5835218, 1.6493348, ~
```

1. Data Preparation for Hypothesis Testing: All the character columns, namely region, and countries are mutated into factors. The countries are divided into ten regions, Central and Eastern Europe, Western Europe, Commonwealth of Independent States, East Asia, South Asia, Southeast Asian, Latin America and Caribbean, Middle East and North Africa, North America and ANZ, and Sub-Saharan Africa. For the t-test, the regions are grouped into two. The first group comprises the high GDP regions: Western Europe, North America and ANZ, East Asia, Central and Eastern Europe, and lastly, the Middle East and North Africa. The remaining regions are grouped into low GDP countries. The median GDP identifies the subgroup of the areas.

The Shapiro-Wilk test is used to perform a normality test between the two groups before t-testing. Both groups of regions is normally distributed with a p-value greater than 0.05.

A One-Sided T-test is performed to see if log GDP countries have less than happy than high GDP countries.

Box Plot of the Happiness Score of Different Regions The box plot shows that Western Europe has the highest median happiness score of all the regions. Another interesting note is a significant inequality in the happiness score in South Asia, where only 20% of the countries are considered relatively happy. In addition to that, even though Latin America and the Caribbean have an overall average happiness score, the two countries are outliers and far below the region's average.



Scatter plot of the Relationship between GDP per Capita The scatter plot shows a positive relationship between GDP per capita and the happiness score, meaning happiness also increases as the GDP per capita of a country increases. Most Western European countries have high GDP and high happiness score.


```
# A tibble: 10 x 3
   region                           log_gdp_median ladder_median
   <fct>                                     <dbl>         <dbl>
 1 Western Europe                            10.8           7.08
 2 North America and ANZ                     10.8           7.14
 3 East Asia                                 10.6           5.76
 4 Central and Eastern Europe                10.3           6.08
 5 Middle East and North Africa               9.58          4.89
 6 Commonwealth of Independent States         9.53          5.47
 7 Latin America and Caribbean                9.45          5.99
 8 Southeast Asia                             9.08          5.38
 9 South Asia                                 8.46          4.93
10 Sub-Saharan Africa                         7.93          4.62
```

```
# A tibble: 10 x 3
   region                           log_gdp_median ladder_median
   <fct>                                     <dbl>         <dbl>
 1 Western Europe                            10.8           7.08
 2 North America and ANZ                     10.8           7.14
 3 East Asia                                 10.6           5.76
 4 Central and Eastern Europe                10.3           6.08
 5 Middle East and North Africa               9.58          4.89
 6 Commonwealth of Independent States         9.53          5.47
 7 Latin America and Caribbean                9.45          5.99
 8 Southeast Asia                             9.08          5.38
 9 South Asia                                 8.46          4.93
10 Sub-Saharan Africa                         7.93          4.62

Rows: 10
```

```
Columns: 2
$ group        <chr> "high_gdp", "high_gdp", "high_gdp", "high_gdp", "high_gdp~
$ ladder_score <dbl> 7.08, 7.14, 5.76, 6.08, 4.89, 5.47, 5.99, 5.38, 4.93, 4.62
```

Normality Test Performing a normality test to see if the two regions are normally distributed.

The Shipiro-Wilk test shows that both High GDP and Low GDP group are normally distributed with p-values greater than 0.05

```
**** Regions with  High GDP ****

    Shapiro-Wilk normality test

data:  ladder_score[group == "high_gdp"]
W = 0.91878, p-value = 0.5221

**** Regions with  Low GDP ****

    Shapiro-Wilk normality test

data:  ladder_score[group == "low_gdp"]
W = 0.9776, p-value = 0.9214
```

T- Testing To answer the question if Low GDP countries have lower happiness score than High GDP countries, one-sided t-test is needed to check if this is true.

The hypotheses are:

- Null Hypothesis: *low_gdp is less than or equal to high_gdp*
- Alternative Hypothesis: *low_gdp is greater than or high_gdp*

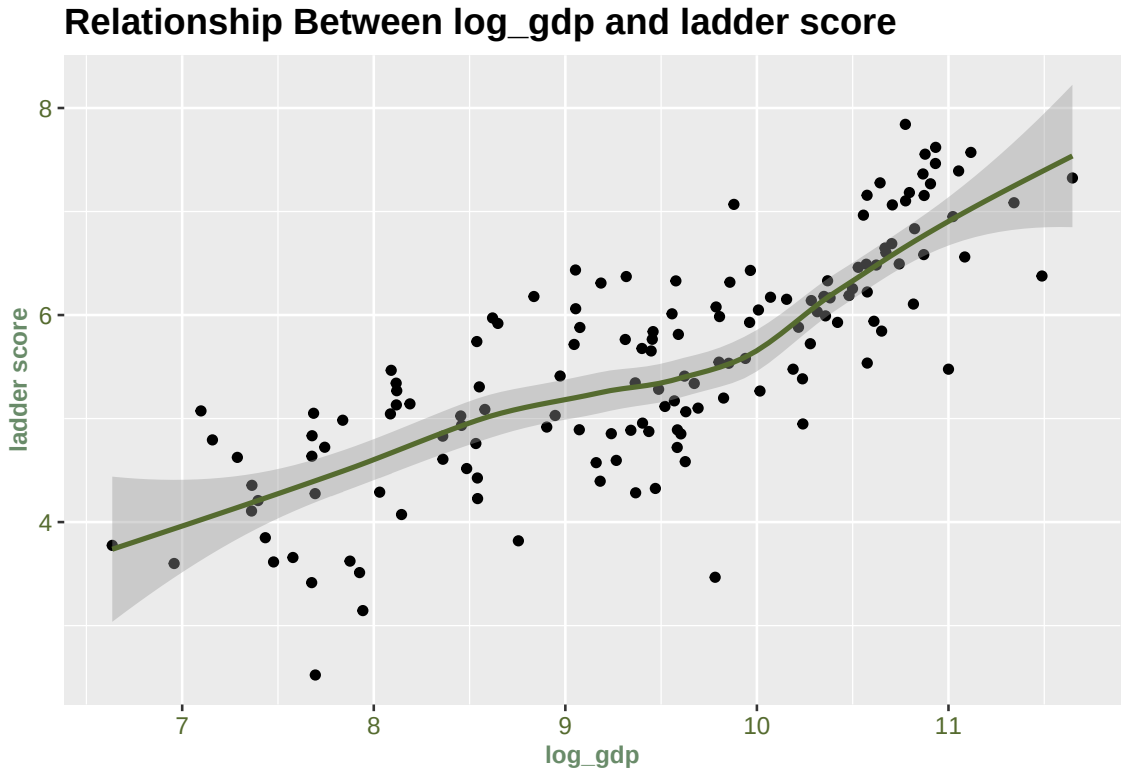
```
**** Low GDP  < High GDP ****

    Welch Two Sample t-test

data:  df_low_gdp$ladder_score and df_high_gdp$ladder_score
t = -1.8833, df = 6.2592, p-value = 0.05331
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.02204472
sample estimates:
mean of x mean of y
    5.278     6.190
```

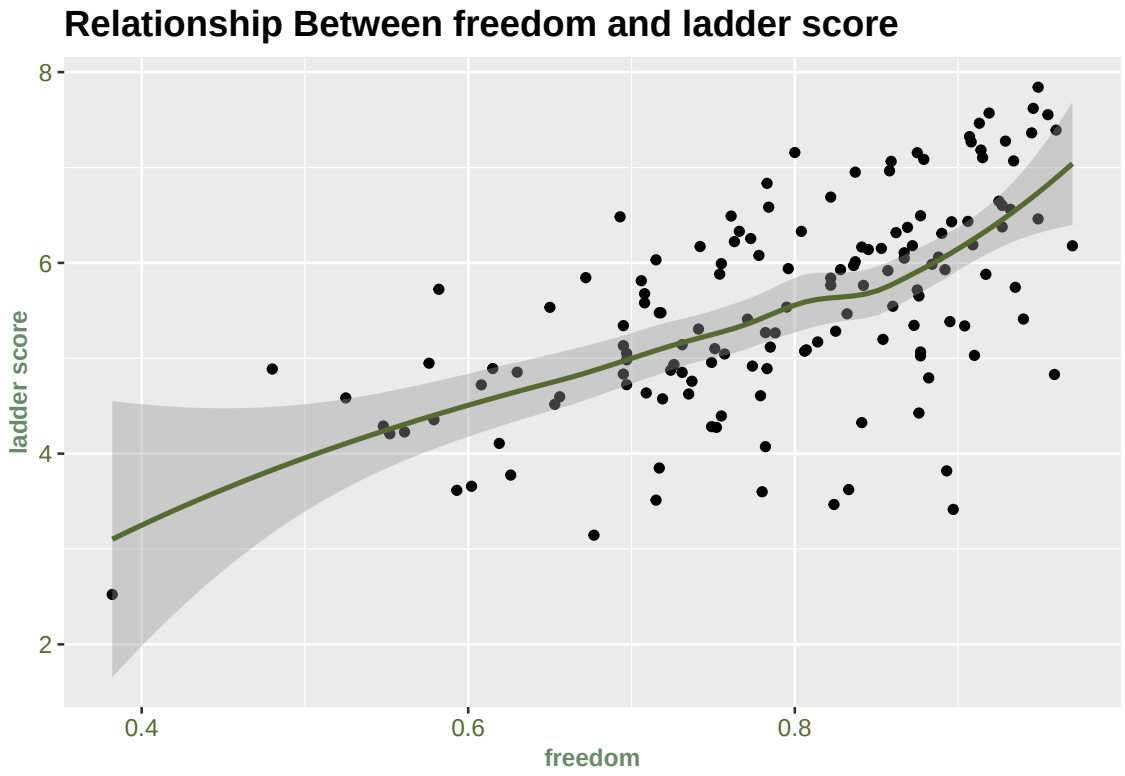
Results and Interpretation The result of the One Sided T-Test shows that countries that have lower GDP tends to have a lower happiness score than countries with high GDP.

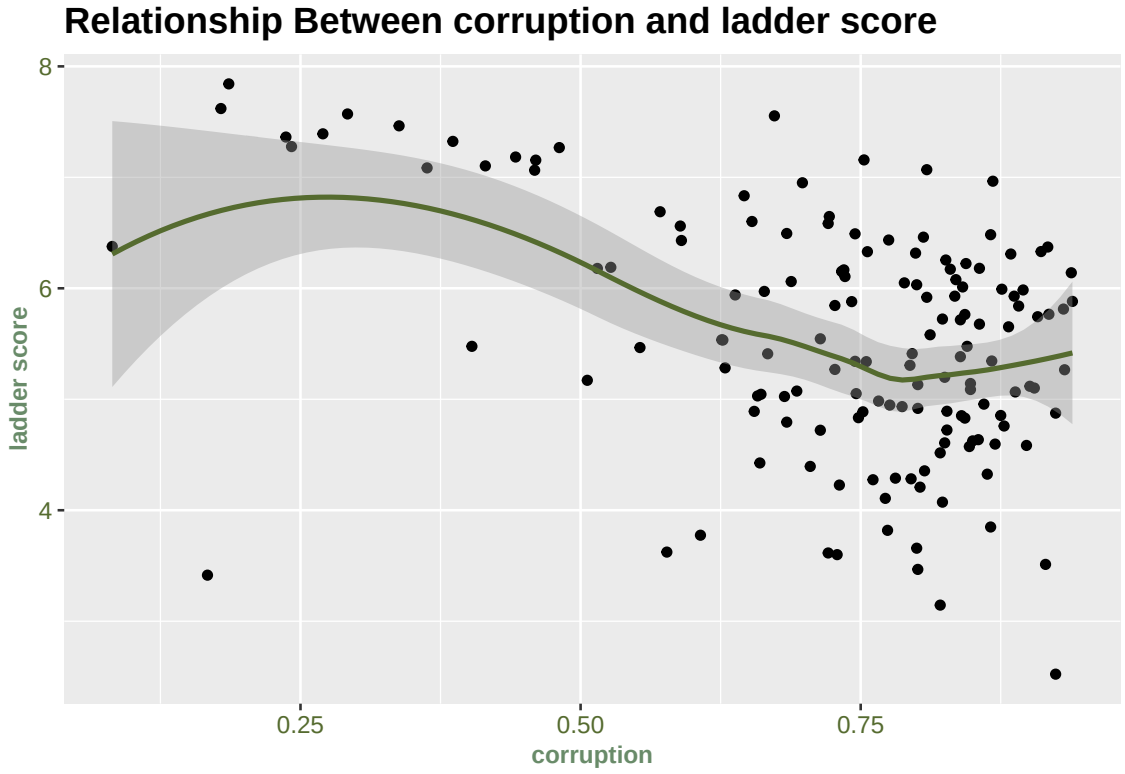
- *p-value* - the p-value = 0.05331 $>$ 0.05 it fails to reject the null hypothesis that low GDP $<$ high GDP
- *confidence interval* - the lower end of the confidence interval is negative infinity and the upper end is 0.022. The confidence interval contains 0, failing to reject the null hypothesis.

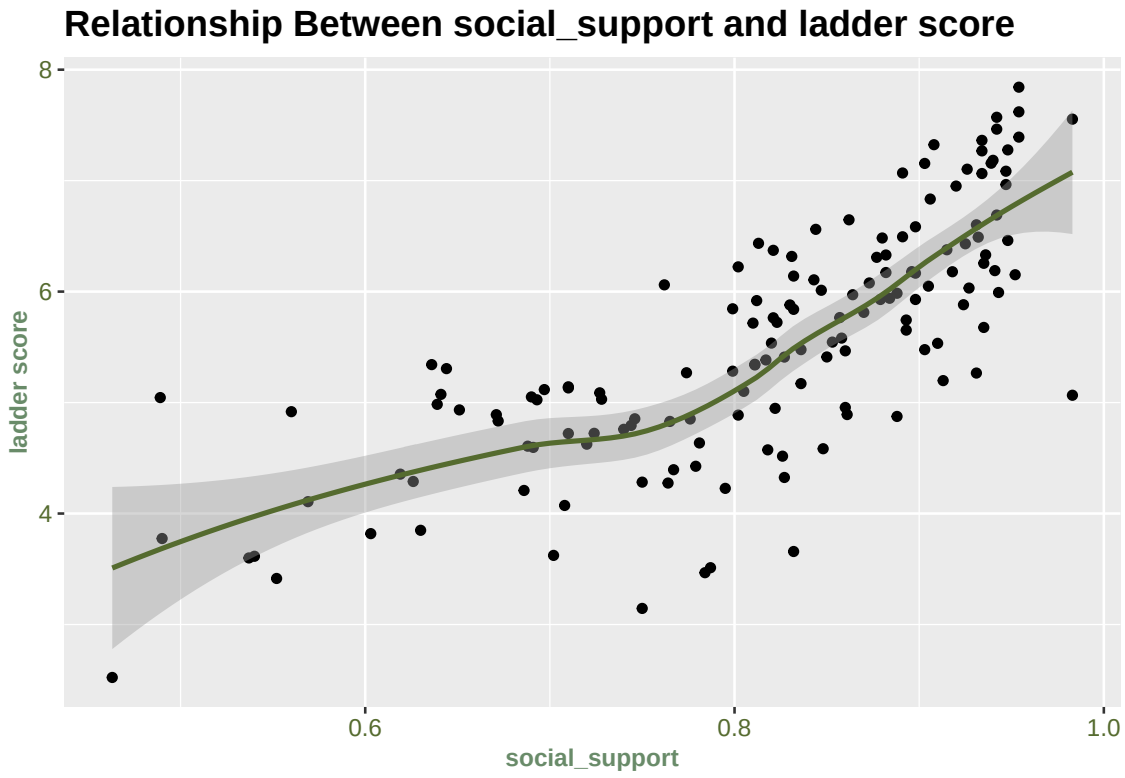
2. Data Preparation for Regression: Splitting the data into a training set and a testing set is done to wrangle the data for regression. The data split is 70/30. The significant independent variables are also plotted to have an overview of their relationship with the dependent variable. The variables in the linear model are accepted when the p-value is greater than 0.05. The better model is chosen by comparing the error value.

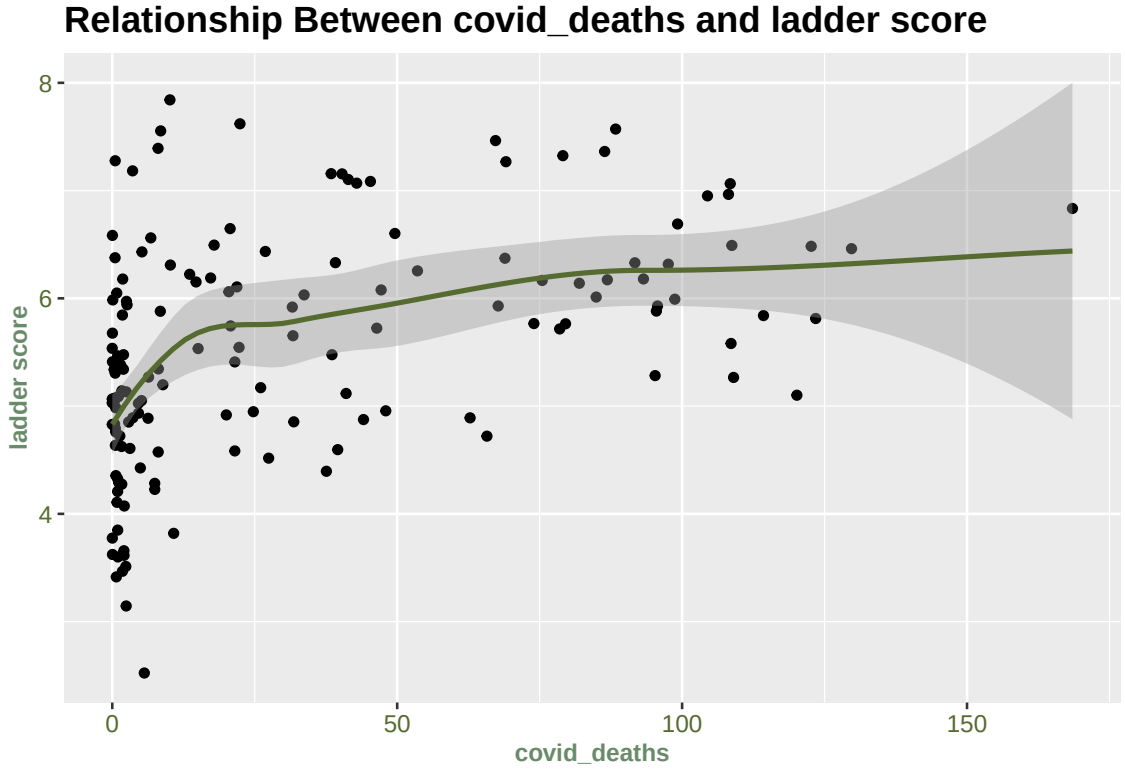
Visualizing Ladder Score in Relation to the Independent Variables The following graphs plots the ladder score to various independent variables to see their relationship. GDP per Capita, freedom, corruption, social support, Covid deaths shows a direct relationship between ladder score, while corruption shows an indirect relationship between the label/dependent variable.




Relationship Between freedom and ladder score
ladder score
8
6
4
2
0.4
0.6
0.8
freedom


Relationship Between corruption and ladder score
ladder score
8
6
4
0.25
0.50
0.75
corruption





Splitting the dataset into testing and training The dataset is split between 70/30. 70% of the data is used for testing, and 30% is for testing, and the seed is set at 1222.

IV. Modeling

a. Training the Models This first model has variables that are not significant which means that it is an overfitting model. To fix this the insignificant variables should be dropped

```
Call:
lm(formula = ladder_score ~ log_gdp + social_support + healthy_life +
    freedom + generosity + corruption + exposure_index + covid_deaths +
    gini_index, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66603 -0.28215  0.09007  0.30169  1.02844

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.715602   0.973772  -1.762  0.08139 .
log_gdp         0.213768   0.100755   2.122  0.03653 *
social_support  2.485906   0.774985   3.208  0.00184 **
healthy_life    0.026622   0.016564   1.607  0.11140
freedom         1.835716   0.570410   3.218  0.00178 **
generosity      0.717836   0.401976   1.786  0.07740 .
corruption     -0.664274   0.397153  -1.673  0.09777 .
exposure_index  0.064885   0.066467   0.976  0.33150
covid_deaths    0.002914   0.002019   1.443  0.15246
gini_index      0.007141   0.007099   1.006  0.31701
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5241 on 93 degrees of freedom
Multiple R-squared:  0.7469,    Adjusted R-squared:  0.7224
F-statistic: 30.49 on 9 and 93 DF,  p-value: < 2.2e-16
```

Testing Model 1

$ladder\_score_1 = 0.25 * log\_gdp + 2.83 * social\_support + 1.93 * freedom - 1.20 * corruption + covid\_deaths - 1$

The first model shows the ladder score with a Covid related deaths.

This model shows that **ladder_score** is significantly dependent on *log_gdp, social_support, freedom, corruption and covid_deaths.*

These independent variables are accepted when p-value < 0.05.

The model also shows the 99% of the variances was reduces as shown by the adjusted R- squared.

The F-static is high at 2284 which represents how model is explaining the variables.

```
Call:
lm(formula = ladder_score ~ log_gdp + social_support + freedom +
    corruption + covid_deaths - 1, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84485 -0.24209  0.07645  0.33206  1.03986
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
log_gdp         0.250036   0.064453   3.879 0.000190 ***
social_support  2.837029   0.751047   3.777 0.000272 ***
freedom         1.936595   0.468363   4.135 7.50e-05 ***
corruption     -1.201081   0.249970  -4.805 5.58e-06 ***
covid_deaths    0.004520   0.001453   3.111 0.002446 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5272 on 98 degrees of freedom
Multiple R-squared:  0.9915,    Adjusted R-squared:  0.9911
F-statistic:  2284 on 5 and 98 DF,  p-value: < 2.2e-16


                    2.5 %       97.5 %
log_gdp         0.12213092  0.377941499
social_support  1.34659952  4.327457529
freedom         1.00714485  2.866045949
corruption     -1.69713760 -0.705023831
covid_deaths    0.00163649  0.007404122
```

Training Model 2

$$ladder\_score_2 = 0.26 * log\_gdp + 2.37 * social\_support + 0.02 * healthy\_life + 2.34 * freedom + 1$$

The second model shows the regression model without any Covid related variables, ladder score is significantly dependent on log gdp and

This model shows that **ladder_score** is significantly dependent on *log_gdp, social_support, healthy_life* and *freedom* at p value < 0.05

```
Call:
lm(formula = ladder_score ~ log_gdp + social_support + healthy_life +
    freedom, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91823 -0.30290  0.09006  0.30008  1.10627

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.79614    0.56033  -4.990 2.62e-06 ***
log_gdp         0.26775    0.09829   2.724  0.00764 **
social_support  2.37076    0.76063   3.117  0.00240 **
healthy_life    0.02983    0.01555   1.918  0.05808 .
freedom         2.34926    0.51958   4.521 1.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5423 on 98 degrees of freedom
Multiple R-squared:  0.7144,    Adjusted R-squared:  0.7027
F-statistic: 61.28 on 4 and 98 DF,  p-value: < 2.2e-16
```

```
                     2.5 %      97.5 %
(Intercept)    -3.908105942 -1.68417756
log_gdp         0.072684723  0.46280620
social_support  0.861310110  3.88021741
healthy_life   -0.001040861  0.06069482
freedom         1.318173015  3.38035000
```

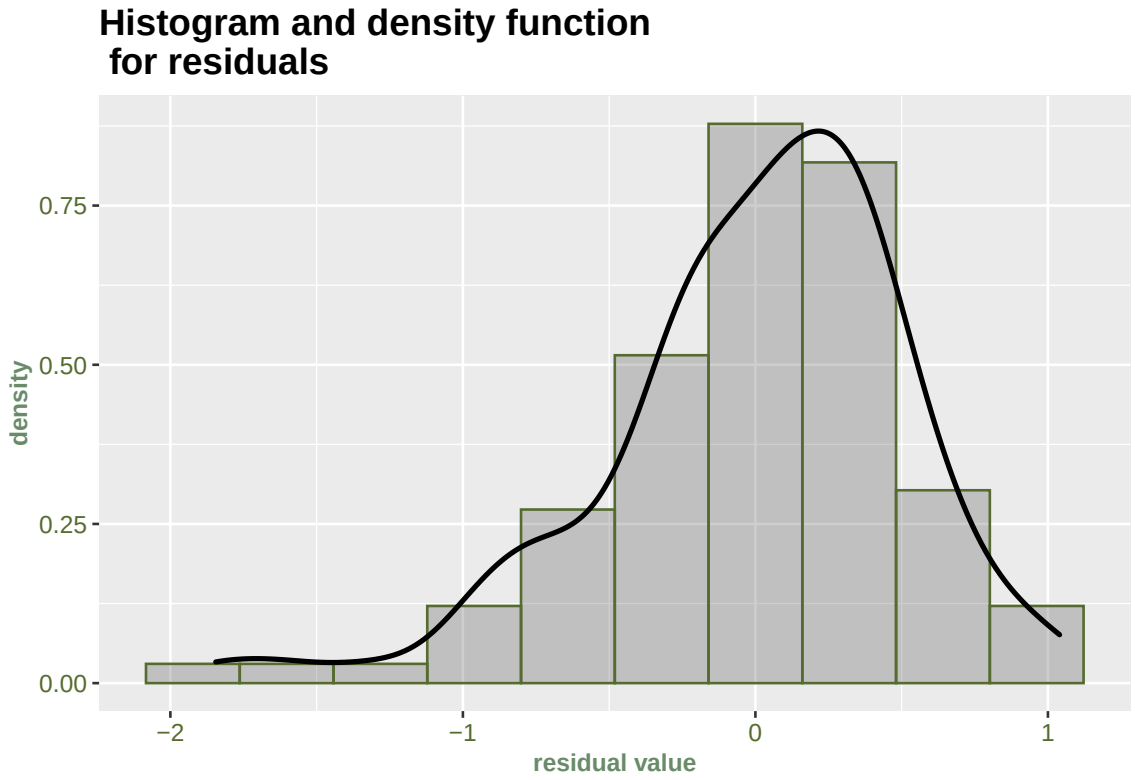
b. Testing the Models

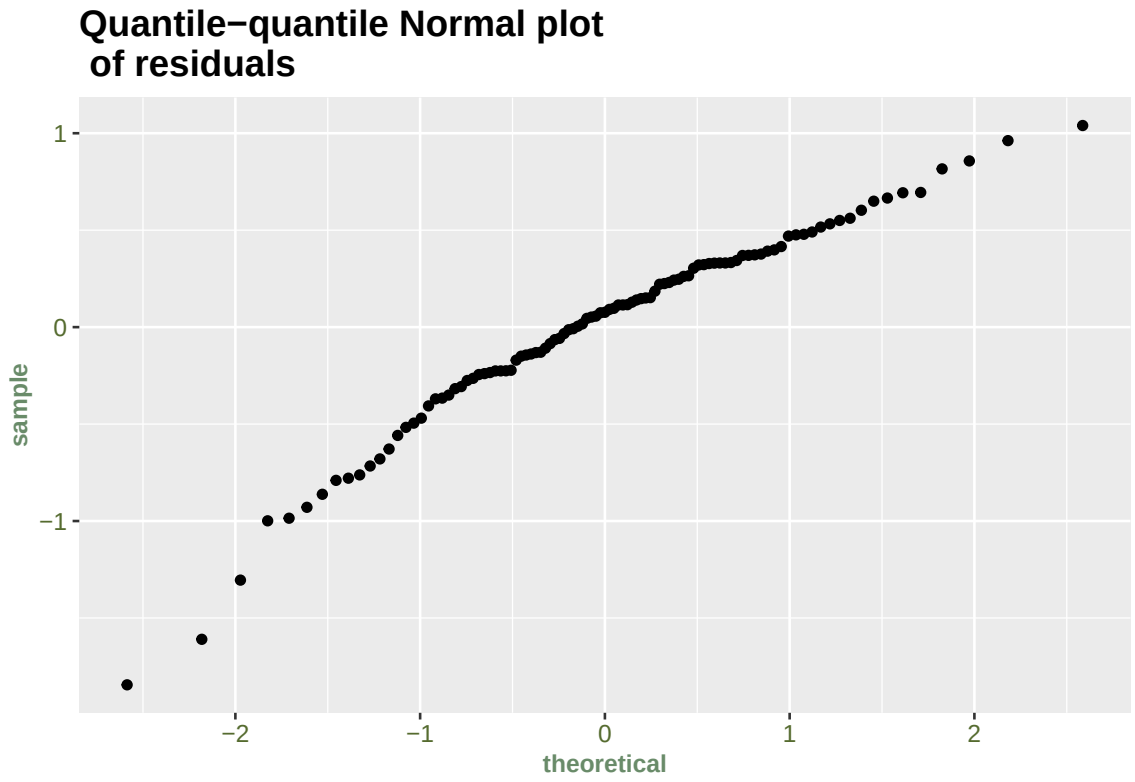
Model 1 Prediction Score and Error A prediction column is included to test the model. The error in model one is equal to 0.57

```
# A tibble: 45 x 2
   ladder_score prediction
          <dbl>      <dbl>
 1         7.84       7.06
 2         7.62       7.16
 3         7.57       7.28
 4         7.55       6.59
 5         7.39       7.04
 6         7.36       7.30
 7         7.16       6.13
 8         7.06       6.93
 9         6.96       6.43
10         6.83       6.78
# ... with 35 more rows
```

```
[1] "Model 1 Error = 0.565099224135639"
```

Visualizing the Residuals for Model 1



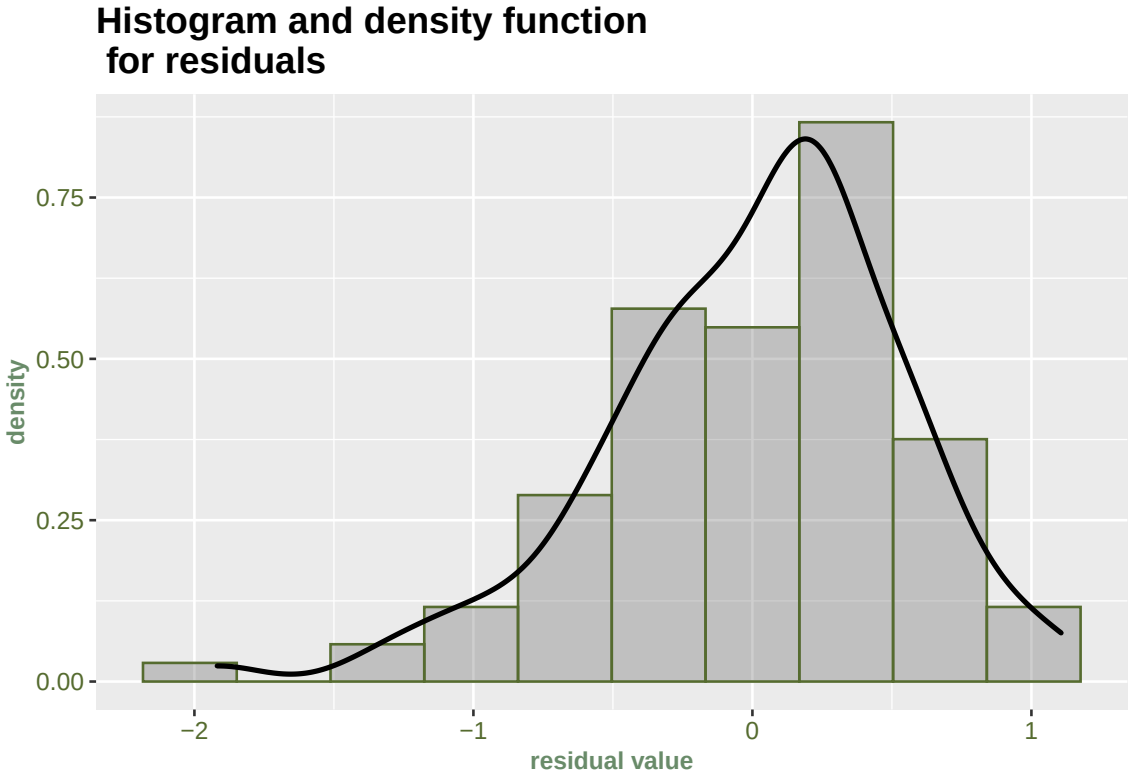


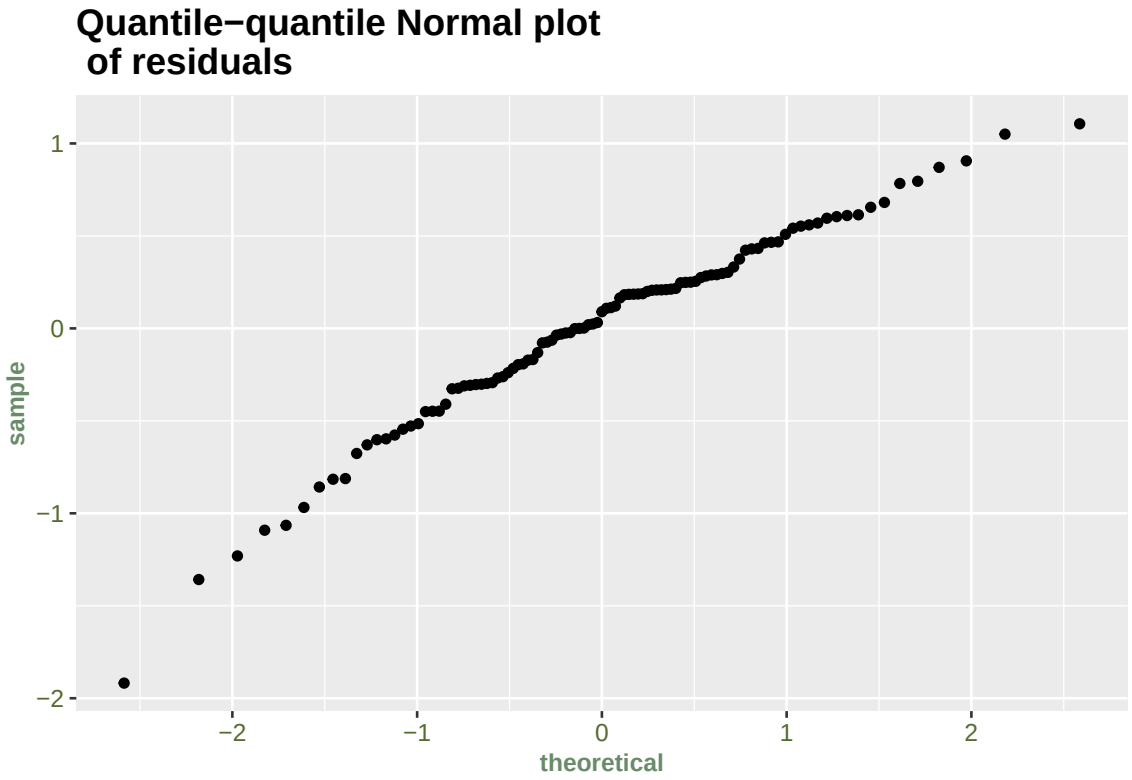
Model 2 Prediction Score and Error The prediction column in model two is computed the same as the first model. The result of the error for model two is equal to 0.59

```
# A tibble: 45 x 2
   ladder_score prediction
          <dbl>      <dbl>
 1         7.84       6.73
 2         7.62       6.78
 3         7.57       6.79
 4         7.55       6.87
 5         7.39       6.87
 6         7.36       6.72
 7         7.16       6.33
 8         7.06       6.47
 9         6.96       6.40
10         6.83       6.24
# ... with 35 more rows
```

```
[1] "Model 2 Error = 0.59393596868558"
```

Visualizing the Residuals for Model 2





c. Comparing and Evaluating the model

- Model 1 $ladderscore = 0.25 * log\_gdp + 2.83 * social\_support + 1.93 * freedom - 1.20 * corruption + covid\_deaths - 1$
- Model 2 $ladder_score = 0.26 * log\_gdp + 2.37 * social\_support + 0.02 * healthy\_life + 2.34 * freedom + 1$

Both models have almost the same variables except a variable gauging Covid 19 related deaths are added to the first model. Although Covid deaths were added to model one, it didn't make a lot of impact on the ladder score. This is explained by the almost zero value in the estimate. In addition to that, the variable healthy life is present on the second model, which, when compared to the first model it is not considered significant. It is safe to assume that without any Covid-related variables in the model, healthy life affects the happiness score. In addition, the intercept in model one is insignificant compared to model two. When comparing the errors, model one has a minor error at 0.57 compared to model two's 0.59.

IV. Results and Conclusions

Understanding the happiness scores of 148 countries for 2020 has provided this study with fascinating insights. It is evident that the pandemic has changed day-to-day life and challenged everyone in the world. The Covid-19 had an effect in 2020, but that effect is very low compared to the other factors that affect people's happiness. What countries value the most is the social support they can get from their friends and families during these difficult times. In addition to that, countries value their freedom to make their own decision to be happy. This is very interesting because this was the time period when there were many restrictions and mandatory lockdowns took place. The perception of corruption is also one of the factors that affect happiness. This perception covers not only the government but for businesses. Relying on the fact that the government is making the right decisions during unpresedented times is very important. Lastly, the GDP per capita is also evident in increasing a countries' happiness. How significant the impact is on the happiness score is represented by the estimates in the regression.

Overall, regions with high GDP, namely, North America, East Asia, Middle East, North Africa, and Europe, have a higher happiness score than regions with lower GDP. Out of all the regions, countries in Western Europe have the happiest people.