MATH7343 Applied Statistics Final Project

Team E

April 25, 2022

Contents

1 Introduction **3**

1.1 Background . . . 3

1.2 Data Source . . . 3

2 Data Summary and Visualization **3**

2.1 Happiness Scores by Country . . . 3

2.2 Happiness Scores by Year . . . 4

2.3 Other Indicators . . . 4

2.3.1 Economy / GDP . . . 4

2.3.2 Health / Life Expectancy . . . 5

3 Regression **6**

3.1 Preprocessing . . . 6

3.1.1 Dealing with Missing Values and Outliers . . . 6

3.1.2 Correlation/Collinearity/Distribution Analysis . . . 6

3.2 Normalization . . . 6

3.3 Split the Dataset into Training Data and Testing Data . . . 7

3.4 Assumptions . . . 7

3.5 Result of the Model . . . 9

3.6 Evaluation of the Model . . . 9

4 ANOVA **12**

5 Covid Hypothesis Test **13**

6 Conclusion **14**

7 Team Information **15**

8 References **15**

1 Introduction

1.1 Background

The World Happiness Report is a landmark survey of the state of global happiness. The reports use data drawn from the Gallup World Pull (GWP), as well as some other resources. The Primary Report was issued in 2012. The most recent 2022 Report will be released on March 18, 2022. Information is available at https://worldhappiness.report.

1.2 Data Source

The data comes from World Happiness Report by Helliwell, John et al., available on Kaggle(Aché)[1].
The data files contain reports from 2015 to 2021. Each file contains the Happiness Score along with the factors used to explain the score. The Happiness Score is a national average of the responses to the main life evaluation question asked in the GWP, which uses the Cantril Ladder. Each data contains: Happiness Score, GDP per capita, Healthy Life Expectancy, Social Support, Freedom to make life choices, Generosity, Corruption Perception, Residual error. Though, there may exist small variations of indicators.

2 Data Summary and Visualization

2.1 Happiness Scores by Country

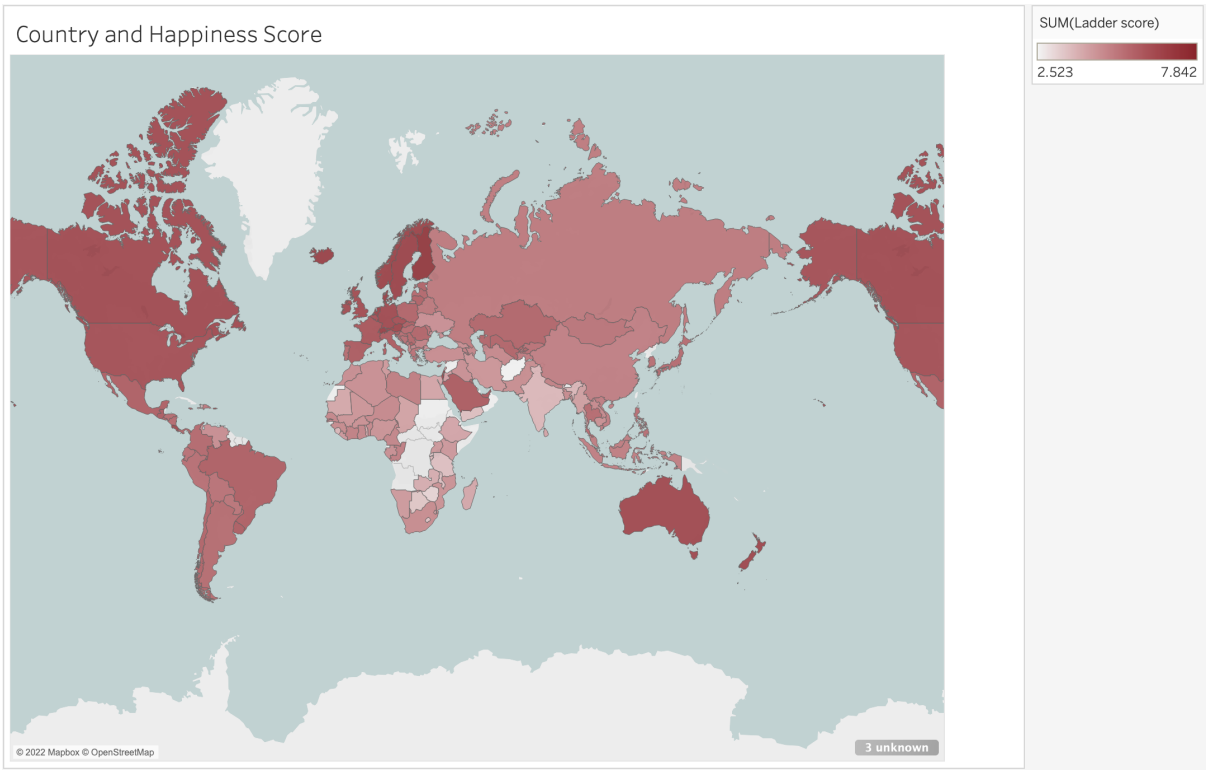
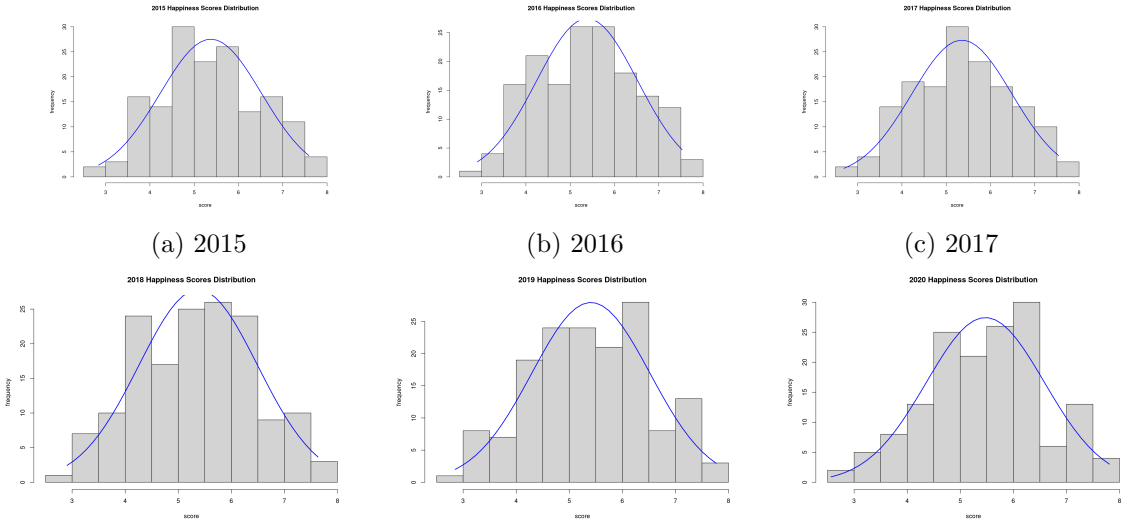


Figure 1: Happiness Score map in 2021

We can see from the map plot (Figure 1) that Afghanistan has the lowest happiness score. Asia and Africa have relatively low happiness scores, while North America, Europe and Australia have relatively high happiness scores.

2.2 Happiness Scores by Year

The happiness scores are approximately normally distributed (Figures 2 and 3). The average score at 2021 ticked up by roughly 3% from 5.376 to 5.533 since 2015 (The percentage change is calculated by $(avg_{2021} - avg_{2015})/avg_{2015}$). Although the average dropped during 2016 and 2017, it began to rise after 2017. Meanwhile, the standard deviation has been lowered each year. The minimum was 2.839 in 2015 but dropped to 2.523 in 2021, whereas the maximum of 7.587 improved to 7.842. If we examine closely, we notice that the minimum happiness score descends while the maximum happiness score has improved in the past two years. In fact, the range (maximum minus minimum) of the data is growing over the years. It may be a sign of discrepancy among countries; that is, citizens in a few countries feel happier while those in other countries feel worse. However, reduced standard deviations may indicate that the majority of people are in the middle and that level of happiness are becoming similar. We are unable to perform root cause analysis at this time based on available data. Some reasonable guesses may be geopolitical conflicts, certain diseases including but not limited to covid, etc.



(a) 2015 (b) 2016 (c) 2017

(d) 2018 (e) 2019 (f) 2020

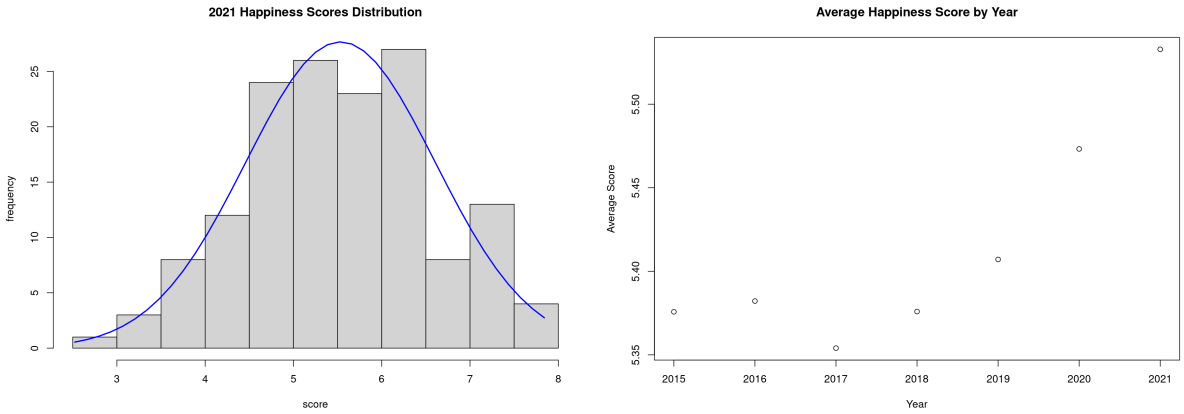
Figure 2: Happiness Score Distribution from 2015 to 2020

2.3 Other Indicators

We look at selected indicators that will be used in regression or anova later.

2.3.1 Economy / GDP

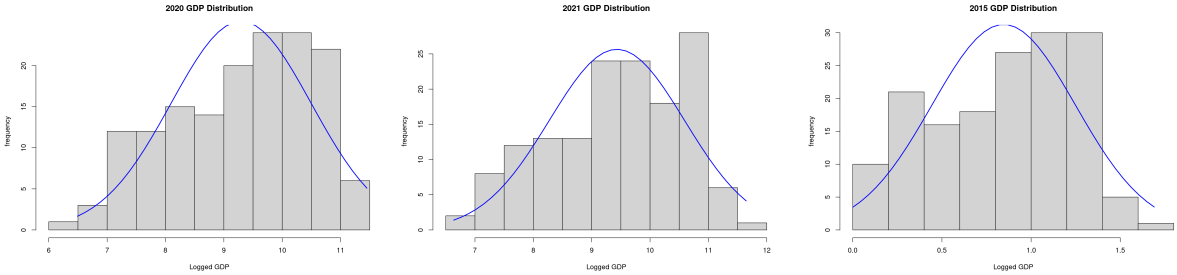
The Economy (GDP) data has two formats. Data prior to 2020 are original GDP per capita., which ranges between 0 and 2.1, with medians around 1. Data after 2021 are displayed in logged



(a) 2021 (b)

Figure 3

form. A GDP of zero likely means that some countries do not report or researchers do not have access to the data. The advantage of logged data is reduced skewed data to approximately confirm normality. From the frequency distribution (Figure 4) , we see that 2015 GDP is skewed to the left. Logged 2020 and logged 2021 GDP are also skewed to the left after the transformation. The difference between the country with the lowest GDP and the country with the highest GDP is getting larger, although the range has been reduced from 2018 to 2019. However, we are unable to tell specifically in which direction GDP changes from 2019 to 2020 due to the difference in the original data. Also, we cannot conclude the overall trends. However, based on 2020 and 2021 data, we can see an increase in GDP across two years. Also, data prior to 2020 also display an overall upward trend except for the year of 2019.



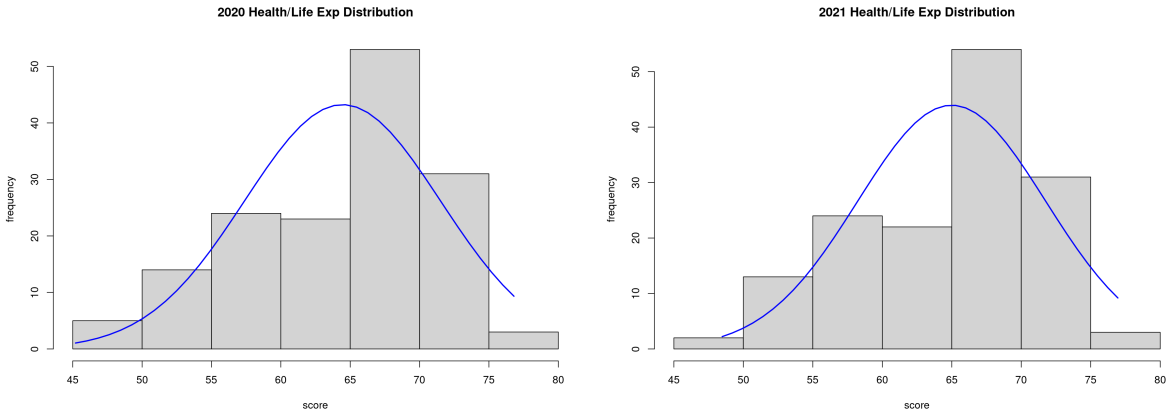
(a) 2020 (b) 2021 (c) 2015

Figure 4

2.3.2 Health / Life Expectancy

The data prior to 2020 ranged between 0 and 1.2, while the 2020 and 2021 data range between 64 and 77. There may be calculation differences when scholars record the data. Nonetheless, the data appear to be skewed to the left (Figure 5) and mode occurs between 65 and 70. Comparing 2020 and 2021 data, we see that the life expectancy is improving - the average has increased by 0.55, the median is up by 0.3, the standard deviation has decreased by 4% from 7.058 to 6.762, the range has reduced by 3.13, the minimum life expectancy is 7% stronger from 45.2 to 48.478, and the maximum is risen from 76.8 to 76.95. Such changes may reveal that living standard are improving worldwide. The countries with the lowest life expectancy are performing better

and that people in those countries live longer than before.



(a) 2020 (b) 2021

Figure 5

3 Regression

We would like to predict the happiness score in 2015 by using Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual.

3.1 Preprocessing

3.1.1 Dealing with Missing Values and Outliers

There are no missing values.
We use boxplot to drop the points below Q1-1.5*IQR or above Q3+1.5*IQR for every variable. The number of rows drops from 158 to 136 after deleting the outliers.

3.1.2 Correlation/Collinearity/Distribution Analysis

We can see the scatter plot (Figure 6), bar plot, and correlation index for all the variables in the plot. All the independent variables are related to the happiness score.

3.2 Normalization

We use the scale() function to do the normalization for all the variables except the predicted variable - happiness score. The normalizing of a dataset using the mean value and standard deviation is known as scaling.

Below is the summary of the dataset after normalization.

We choose the linear regression model.

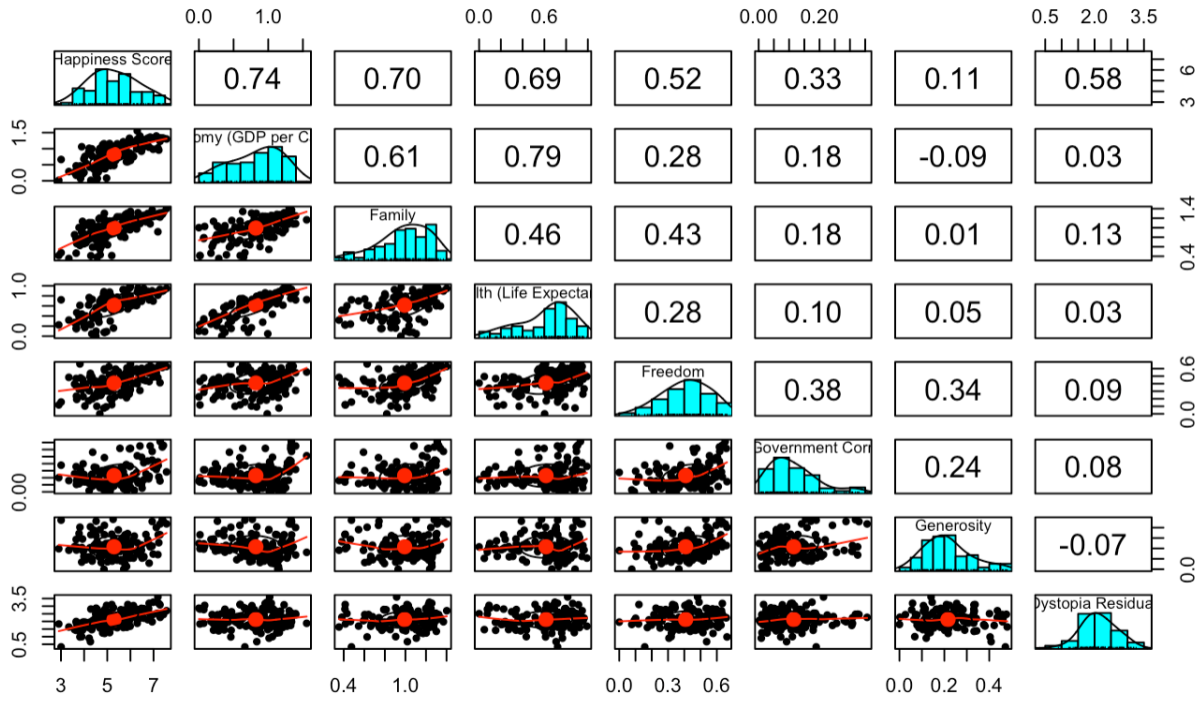


Figure 6: correlation between all the variables

```
Economy (GDP per Capita)     Family       Health (Life Expectancy)    Freedom       Trust (Government Corruption)   Generosity
Min.   :-2.1986          Min.   :-2.6319   Min.   :-2.5513          Min.   :-2.7932   Min.   :-1.4012               Min.   :-2.0283
1st Qu.:-0.7369          1st Qu.:-0.5477   1st Qu.:-0.7481          1st Qu.:-0.6338   1st Qu.:-0.7024               1st Qu.:-0.7382
Median : 0.1902          Median : 0.1038   Median : 0.2775          Median : 0.0506   Median :-0.2662               Median :-0.1176
Mean   : 0.0000          Mean   : 0.0000   Mean   : 0.0000          Mean   : 0.0000   Mean   : 0.0000               Mean   : 0.0000
3rd Qu.: 0.8066          3rd Qu.: 0.8554   3rd Qu.: 0.7130          3rd Qu.: 0.7650   3rd Qu.: 0.4927               3rd Qu.: 0.5345
Max.   : 1.9601          Max.   : 1.6919   Max.   : 1.5573          Max.   : 1.7374   Max.   : 2.9942               Max.   : 2.4781
Dystopia Residual   Happyiness
Min.   :-3.2584   Min.   :2.905
1st Qu.:-0.6240   1st Qu.:4.542
Median :-0.0293   Median :5.166
Mean   : 0.0000   Mean   :5.292
3rd Qu.: 0.6260   3rd Qu.:5.977
Max.   : 2.6852   Max.   :7.561
```

Figure 7: summary of the dataset after normalization

3.3 Split the Dataset into Training Data and Testing Data

Choose 70% of the data randomly to be the training data, the remaining data to be the testing data.

3.4 Assumptions

There are four assumptions of linear regression.

1.Linear Relationship
It means there exists a linear relationship between the independent variable, x, and the dependent variable, y. We can check for linearity by using scatter plots.

From the plots, the linear relationship is not good for variable Generosity, for other variables it is good.

2. Independence
It means the residuals are independent.

It is not a longitudinal data set, so we do not need to worry about independence assumptions. It is “assumed” to be met.

3. Homoscedasticity
It means the residuals have constant variance at every level of x.

4.Normality
It means the residuals of the model are normally distributed.
We will check for Homoscedasticity and Normality later.

3.5 Result of the Model

```
Call:
lm(formula = score ~ ., data = train)

Residuals:
       Min          1Q      Median          3Q         Max
-5.486e-04 -2.176e-04 -2.703e-05  2.417e-04  5.585e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.292e+00  3.034e-05  174425   <2e-16 ***
GDP         3.738e-01  5.764e-05    6485   <2e-16 ***
Family      2.425e-01  3.962e-05    6120   <2e-16 ***
Health      2.411e-01  5.154e-05    4679   <2e-16 ***
Freedom     1.462e-01  3.704e-05    3947   <2e-16 ***
Trust       8.053e-02  3.772e-05    2135   <2e-16 ***
Generosity  1.065e-01  3.257e-05    3271   <2e-16 ***
Dystopia    5.508e-01  2.946e-05   18695   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0002934 on 87 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.881e+08 on 7 and 87 DF,  p-value: < 2.2e-16
```

Figure 8: summary of the linear regression model

3.6 Evaluation of the Model

With R-squared, and Adjusted R-squared of 1, the model is good, but we are concerned about overfitting. Conducting the regression model on testing data, we find that the R squared is 0.9999, and the mse is 8.452302e-08.

From the plot (Figure 9 to 12):

1. Residuals vs Fitted

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable, and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

From the residual plot of our model below, we can see that the red line is basically horizontal

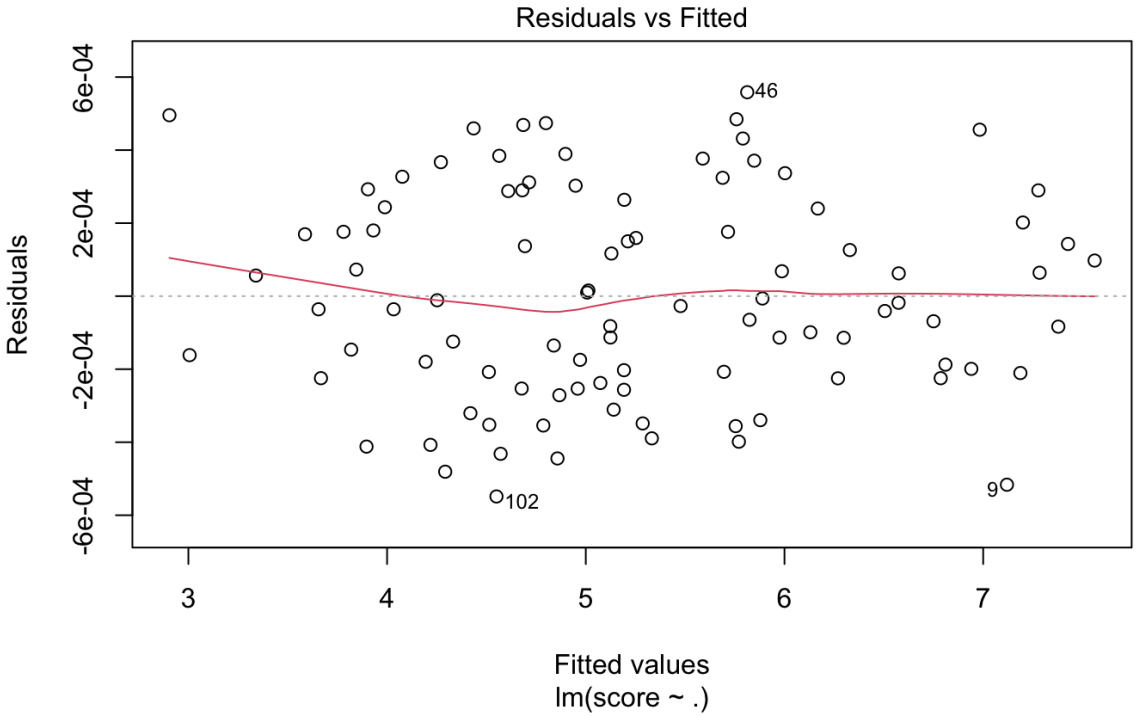


Figure 9: Residuals vs Fitted

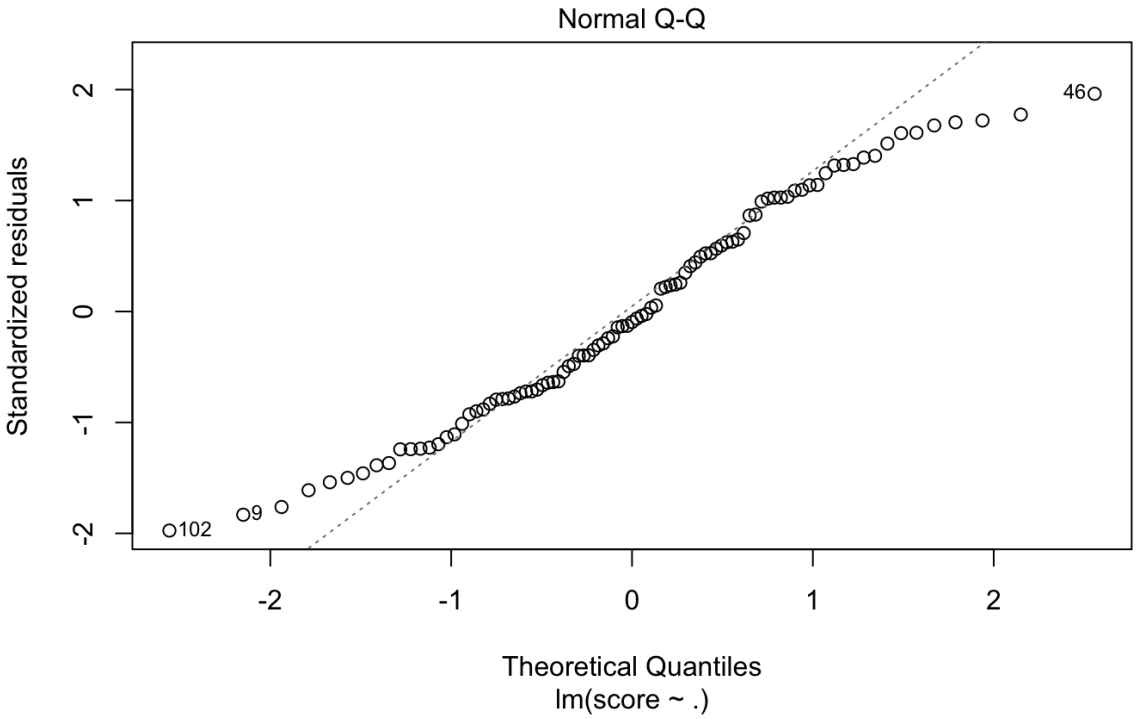


Figure 10: Normal Q-Q

and centered around zero. This means most of the data meets the regression assumption well.

2. Normal Q-Q

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or

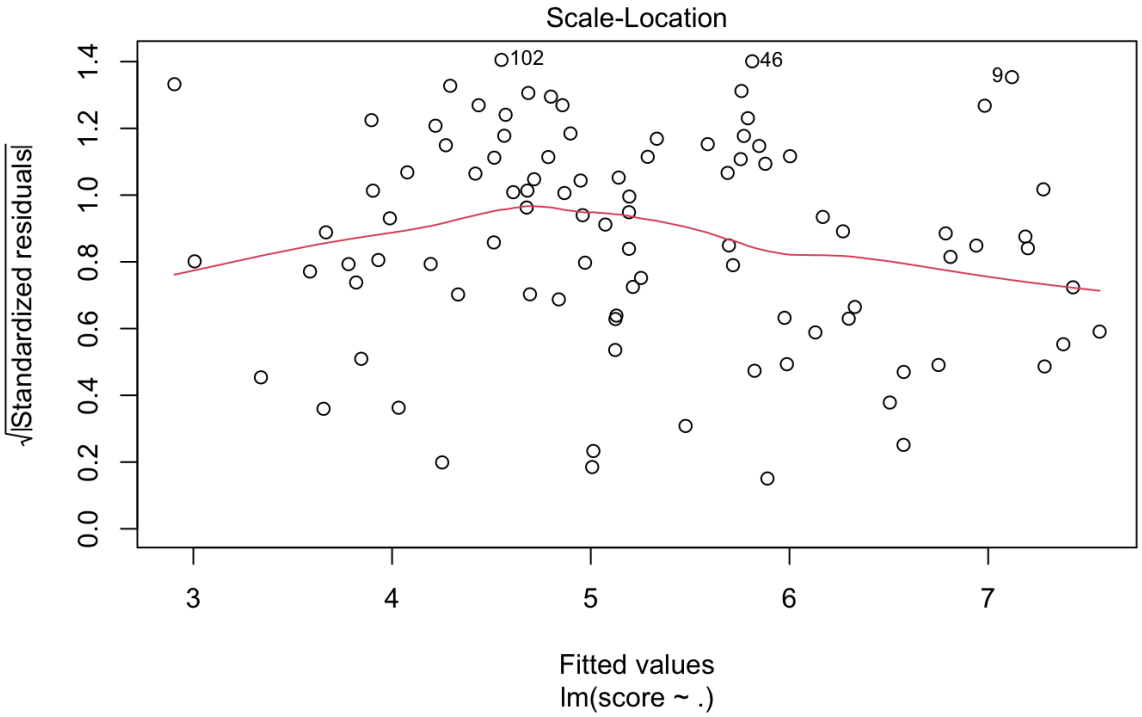


Figure 11: Scale-Location

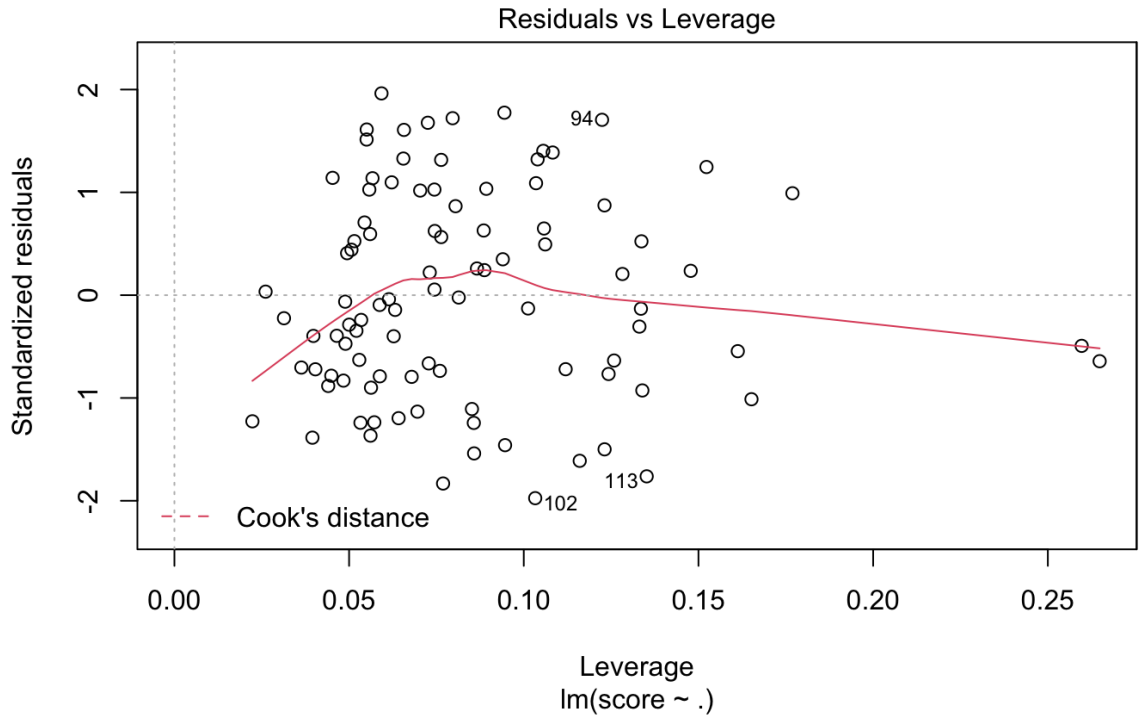


Figure 12: Cook's Distance

do they deviate severely? It's good if residuals are lined well on the straight dashed line. Our Q-Q plot is basically good, but the 3 observations look a little off.

3. Scale-Location

It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.

In the Scale-Location plot of our model, the residuals appear randomly spread. The red smooth line is basically horizontal, although not perfect.

4. Residuals vs Leverage

We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

In our Residuals vs Leverage, 3 points are beyond the Cook's distance lines. But basically it's good.

From the above(Bommae)[2], it can be concluded that the assumptions are met basically.

In conclusion, it is a good model to predict the happinees score in year 2015 given the data of Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual.

4 ANOVA

In order to perform the analysis between 2017 ,2018 ,2019 happiness scores, we used an analysis of variance (ANOVA) table. We chose 2017,2018,2019 for ANOVA since the datasets in these three years have the same amount of countries.The null hypothesis is H_0:$\mu_{2017} = \mu_{2018} = \mu_{2019}$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	0.3	0.1724	0.134	0.875
Residuals	438	563.6	1.2867		

Figure 13: results of ANOVA test

In Figure 13, the ANOVA table lists the test statistic F and its associated p-value. A p-value displayed as 0.875 means that p-value > 0.05.

In Figure 14 and 15, the table has lwr, upr, and p adjusted columns between each groups. Lwr indicates lower bound of the difference in mean between groups; upr indicates upper bound of the difference in mean between groups; p adj indicated p-value between each group. The p-values are respectively 0.8641,0.9756,0.9504 (Finnstats)[3]. Overall, we can see that the conclusion is consistent that there is no significant difference between the means of 2017,2018,2019 happiness score.We will not reject the null hypothesis.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Score ~ group, data = df)

$group
                 diff        lwr       upr     p adj
three-one  0.06812245 -0.2430339 0.3792788 0.8641404
two-one    0.02796599 -0.2831903 0.3391223 0.9756736
two-three -0.04015646 -0.3513128 0.2709998 0.9504985
```

Figure 14: results of ANOVA test

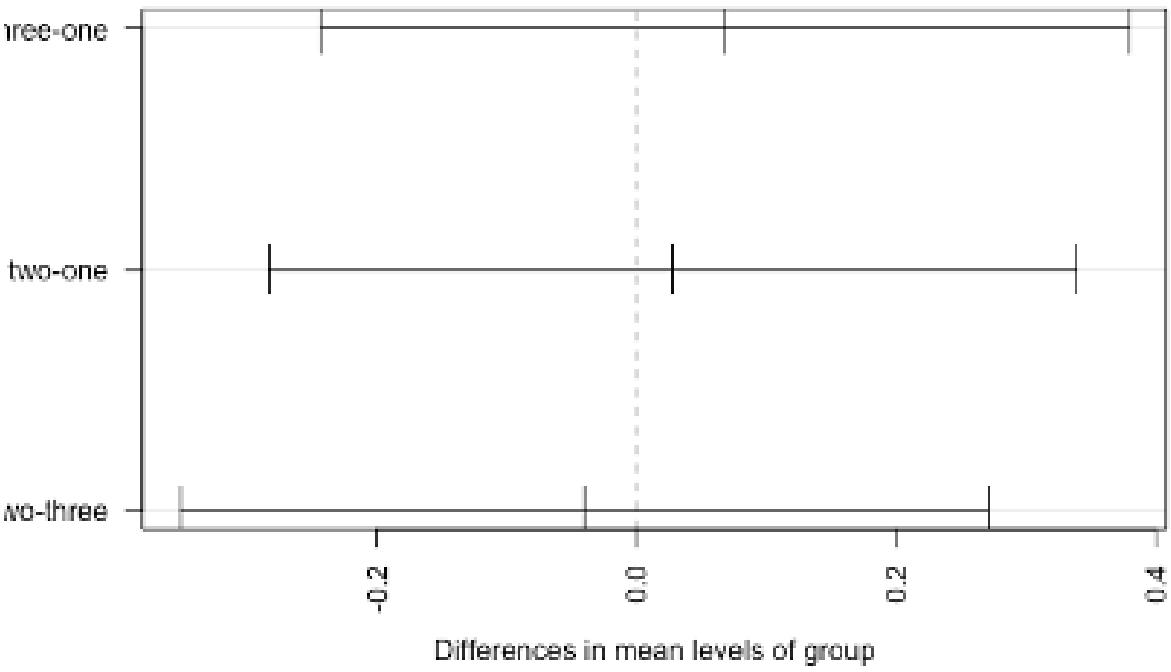


Figure 15: results of ANOVA test

5 Covid Hypothesis Test

Since Covid-19 started in 2020, we suspect that maybe covid-19 may have some negative effect on happiness scores. So, we decided to do a hypothesis test on the happiness score between 2019 and 2020. For the reason why we chose 2019 is because we want to minimize the effect on other variables.

For data processing, we didn't remove the outliers, we wanted to use paired data, so we removed the data from countries that not in both years. Then we use the Wilcoxon-signed rank test (STHDA)[4] to do the hypothesis test, where H_0:$\mu_{2019} = \mu_{2020}$

```
        Wilcoxon signed rank test with continuity correction

data:  Y2019$Score and Y2020$`Ladder score`
V = 1440, p-value = 1.057e-14
alternative hypothesis: true location shift is not equal to 0
```

Figure 16: results of wilcoxon signed rank test

The above figure 16 shows the result of the Wilcoxon-signed rank test. Since p-value is less than 0.05, we reject H_0 so $\mu_{2019} \neq \mu_{2020}$. Then we could conclude that covid-19 does have some effects on happiness. But when we calculate the mean value for happiness score for each year, we found out that the score of 2020 is even higher than the score of 2019, which means the negative effect caused by the covid-19 may not be as big as we though. So we think this study needs further investigation. And now we have some assumptions for that:
1.Due to Covid-19, people need to work at home in some countries, so they may have more spare time; thus the happiness score got higher.
2. Covid-19 may have some huge effect on some countries, but most of the countries may not be greatly affected.

6 Conclusion

Countries in Middle-East, Asia, and Africa have relatively lower happiness scores, whereas those in North America, Europe, and Australia have relatively higher happiness scores. The average happiness score has been improving over the years. The range is growing, but the standard deviation is getting smaller. Such change may indicate that, as time goes, the majority of people are feeling happier, except for people in countries that are in crisis or trouble.

The regression model we constructed does a good prediction about the happiness score in year 2015 with high R-squared and basically met assumptions. From the result of linear regression, all the seven independent variables are important at 0.05 level of significance.

The Economy (GDP) and Health (Life Expectancy) indicators are probably unreported or inaccessible, which could be the reason why there are zeroes in these data. Also, data before and after 2020 may have been reported in separate ways. We are unable to see the overall trends from 2015 to 2021. However, both indicators have grown from 2020 to 2021, which means that the economy and living standard have improved worldwide.

The ANOVA result indicates there is no difference between 2017,2018,2019 mean of happiness score since p-value between groups are both grater than 0.05.

Covid-19 do have some effect on the happiness score. But from the results of the calculation and the test, there was no evidence that a large difference between the data of those two years exists. So this study needs more research to conclude the final results.

7 Team Information

Donghao Liu	Covid Hypothesis Test, ANOVA
Jie Ji	Data Cleaning, Data Visulization and Regression
Hui Du	Regression
Danxi Wu	Data Summary and Visualization
Kachun Lee	Covid Hypothesis Test, ANOVA
Jianjian Liu	Regression

8 References

1. Aché, Mathurin. "World Happiness Report 2015-2021." Kaggle, 19 Mar. 2021
https://www.kaggle.com/datasets/mathurinache/
world-happiness-report-20152021?resource=download.
2. Bommae, Written by. "University of Virginia Library Research Data Services
+ Sciences." Research Data Services + Sciences,
https://data.library.virginia.edu/diagnostic-plots/.
3. Finnstats. "How to Perform Tukey HSD Test in R: R-Bloggers." R, 28 Aug. 2021,
http://www.r-bloggers.com/2021/08/how-to-perform-tukey-hsd-test-in-r/.
4. "Paired Samples Wilcoxon Test in R." STHDA,
http://www.sthda.com/english/wiki/paired-samples-wilcoxon-test-in-r.