A Statistical Approach to Improving Happiness

Introduction to Statistics for Data Science

Group Project Report

Year of Study: 2023/24

Part 1: Introduction

The reported happiness of a country's population has continued to be a critical area of focus for countries, as it is often used as a proxy for a country's success. While improving happiness can be a common goal that governments would like to achieve, understanding the mechanics of how to do this can be overwhelming as multiple key variables need to be considered. To gain a better technical understanding of how to improve happiness, we can look to the World Happiness Report. This dataset contains 137 countries and countries can see their reported happiness in the ladder_score column. This report focuses on six variables: Natural Logarithmic Gross Domestic Product (LGPD), Social Support, Healthy Life Expectancy (HLE), Freedom, Corruption, and Continent. To gain a better understanding of how these factors contribute to a higher ladder score, we can use descriptive statistics and a multivariate regression model to analyze happiness ladder scores. Doing so will allow us to provide recommendations for governments on how to improve their happiness over the next decade by focusing on which key factors could provide the best yield when invested.
From an initial look at our data, Finland and Denmark stick out as leaders in the levels of happiness that their residents experience and are leaders in most of the six factors. Since there is no clear pattern of which factors contribute to a higher ladder score, does this mean that countries should approach this problem by trying to incrementally improve all factors? Or, are there some key factors that are more responsible for these countries' happiness? After building our model using all the data from the WHR, here are the key recommendations::

We propose the government puts a primary focus on Social Support due to its strong relative correlation with happiness. Here is what additional focus could look like:

- The government could strengthen the foundations of social welfare systems to provide safety for their most vulnerable populations. This includes extending unemployment benefits, healthcare access and social security.

While Social Support has the largest impact on happiness, other key variables have a sizable impact on happiness. This means that further investments into these variables, namely Freedom, LGDPHLE, and Corruption, can yield a great ladder score as they have a significant correlation with happiness. Here are further recommendations for these variables:

- Continue to grow a robust economy (LGPDHLE) whilst focusing on social support. A stable economy can help provide funding for social programs through investments, higher tax revenue, and privatized industries that offer support.
- Promote freedom within the country; governments should ensure citizens are free to express themselves and have control over their lives.
- Address and lower Corruption; tackling corruption is crucial as it can damage the trust in the government and its effectiveness in social programs. Our results show that corruption has the most negative impact on happiness, so additional focus on preventing it can help.

Part 2: Findings

Continent	LGDP Rank	Support Rank	HLE Rank	Freedom Rank	Corruption Rank	Aggregated Score	Happiness Rank
Oceania	1	1	1	1	1	5	1
Europe	2	2	2	3	2	11	2
North America	3	4	3	2	4	16	3
Asia	4	5	5	5	3	22	5
South America	5	3	4	4	6	22	4
Africa	6	6	6	6	5	29	6

Figure 1: Rankings by mean for each variable

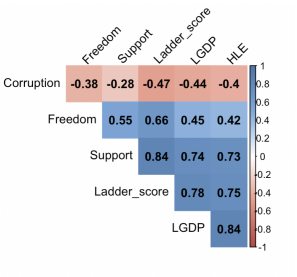


Figure 2: Correlation matrix

Correlation

We find that various factors such as Social Support, LGDP, HLE, Freedom, and Corruption correlate with the Ladder Score (happiness, see Figure 2). Social Support is the highest correlated to the ladder Score. However, these correlations do not necessarily mean causation. For instance, higher LGDP might boost economic success, increasing government revenues, which can then be invested into Social Support programs. This suggests an interplay between LGDP and Social Support. A broader more holistic approach might be needed to increase happiness considering this interplay between variables.

LGDP followed by HLE shows the second strongest positive correlation with happiness. HLE and LGDP are extremely statistically significant between each other. Potentially because of the high funding needed for healthcare. In our analysis, these factors influenced each other too much which can skew our result in the analysis of our multivariate model, therefore, we treat them as interdependent in our regression model. We have a new variable LGDPHLE, to represent the combined effect of these factors. In contrast, increased corruption has a negative influence on happiness and has the weakest correlation. It also has a significant but weaker correlation.

Mean, Variance and Outliers

Corruption has the highest number of Outliers, with notable outliers like Finland, Denmark, and the Netherlands, which have extremely low corruption levels and rank among the happiest. These cases, such as Finland's low corruption score of 0.187 suggest corruption has a high influence on happiness, however, the high average corruption score of 0.7267 implies corruption alone may not directly increase happiness.

The variables have a broad range, indicating strong differences between countries. In particular, HLE and Happiness have high variance, which means their values spread out further away from the mean. In contrast, Social Support, Corruption and Freedom have lower relative variation. Therefore, the effect of Social Support, Corruption and Freedom on happiness might be different, but it is more consistent globally.

The top 10 are spread across Europe and Oceania, with Israel being an exception from Asia. This is consistent with the broader analysis from Figure 2, which identifies Europe, Oceania and North America as continents with the happiest countries on average. The higher score is consistent with higher rankings in LGDP, Social Support, HLE, Freedom and Corruption, the happiest countries like Finland, Denmark and Iceland abide by this rule. In contrast, continents with lower happiness rankings, and corresponding scores rank lower in these factors. The data corresponds that improvement in these variables is associated with increased happiness regardless of continent.

Interaction between variables
We aim to quantify the influence of various factors on happiness scores. Using normalized coefficients, we directly compare their impact, except for continents which are categorical. The analysis has shown that the greatest effect is given by Social Support (with a coefficient of 0.402421), followed by the combined LGDP + HLE (0.28098), Freedom (0.2241) and Corruption (-0.15325). Consequently, support was found to be approximately 1.8 times more influential than Freedom, and Freedom was about 1.46 times more influential than Corruption.

Furthermore, when we predict a 10% increase in each variable mean has the following result (we increase the non-normalised variable by 10%) :

- 10% increase in Support for continent Europe: Ladder Score goes from 5.68 to 5.97 (**5.5% increase**)
- 10% increase in Freedom for continent Europe: Ladder Score goes from 5.68 to 5.855666 (**3.3% increase**)
- 10% decrease in Corruption for continent Europe: Ladder Score goes from 5.68 to 5.727862 (**1.1% increase**)
- 10% increase in LGDPHLE for continent Europe: Ladder Score goes from 5.68 to 5.998779 (**5.8% increase**)

Additionally, if we increase Social Support by 10% and decrease the other variables by 10%. We find that the ladder score decreases, despite Social Support increasing.

The R-squared score (which is used to measure the accuracy of a model, the higher the better) is higher when we add more variables. With only Support, we have a score of 0.702, 0.812 with LGDP, HLE, Support, and Freedom, and 0.847 with all four variables included.

Consistent with our findings, the result from our multivariate model shows that Europe, North America, and Oceania have higher average happiness scores than Africa. Conversely, Asia and South America score lower than Africa. Given that Africa is the baseline. Consistent with our analysis of Continent means.

Country.name	LGDP	Support	HLE	Freedom	Corruption	Continent	Ladder_score
Finland	10.792	0.969	71.150	0.961	0.182	Europe	7.804
Denmark	10.962	0.954	71.250	0.934	0.196	Europe	7.586
Iceland	10.896	0.983	72.050	0.936	0.668	Europe	7.530
Israel	10.639	0.943	72.697	0.809	0.708	Asia	7.473
Netherlands	10.942	0.930	71.550	0.887	0.379	Europe	7.403

Figure 3: Highest-ranked countries

Analysis
Social Support has the highest normalized unit increase and strongest correlation with happiness. It is around 1.43 times the next most influential variable LGDPHLE, indicating that Social Support plays a crucial influence in determining happiness levels.
When we model a real-life scenario by examining the impact of a 10% increase in the mean of LGDP+HLE and Social Support we find that LGDP+HLE performs slightly better despite it having a lower standardized coefficient. This could be because of the higher variability Social Support has with a range of 80.93% as a percentage of its mean, compared to LGDPHLE's relatively small 39.77%. This results in LGDPHLE having a more consistent impact on happiness score. The difference between variance might signify an underlying issue, increasing

LGDP might be harder across countries than increasing Social Support. There is more room for improvement in Social Support globally.

Despite their weaker correlations with happiness compared to Social Support and LGDPHLE, Corruption and Freedoms statistically significant correlation should not be overlooked. As evidenced by the notable highest-ranked countries which have high Freedom and low Corruption Scores.

In conclusion, the complex nature of happiness suggests a holistic approach is needed. While certain factors such as LGDP+HLE and Social Support have more of an impact on happiness score, the interdependence and combined influence of all the variables need to be considered. This is reinforced by the higher R-squared score when we include more variables in the model, indicating that a comprehensive approach considering all these aspects is likely to be more effective in enhancing happiness.

Recommendations

The analysis highlights the huge potential for global improvement in Support, which our model suggests has the greatest influence on happiness. Governments could prioritize this area, perhaps through fiscal policies aimed at broadening social welfare programs, to assist people in disadvantaged positions e.g. enhancing unemployment benefits.

However, any fiscal policy to improve Social Support should be done responsibly to prevent a potential negative impact on the economy (LGDP). If policies enhance Social Support but negatively affect economic stability, the initial benefit could be offset by the economic decline. Over time, there might be reduced funding for social support programs due to the declining economy. Therefore, a holistic approach is needed that ensures all the variable scores are considered in the policy. We should not increase a variable to the detriment of another variable. Instead, we should aim to maintain and improve the variable scores collectively.

With that being said, the key areas to focus on for a country to improve happiness would be to focus on social support while growing a robust economy (LGPDHLE). Because a strong economy can help provide funding for social programs through investments, higher tax revenue, and privatized industries that offer support, the combined focus on these two areas has a strong potential to increase happiness. Furthermore, countries should focus on increasing the freedom of their citizens & keeping corruption as low as possible; while our model shows that these variables aren't of the highest importance, they do still have a sizable impact on happiness and could hurt other variables thus further focus on these areas is recommended.

Limitation

The world happiness report has become influential in analysis of well-being, however, its methodology is primarily based on survey data. This has several flaws that might impact our analysis. For example, there is a lot of reliance on self-reported data that opens up the survey to respondent bias - the current mood of the respondent can affect his answers.

We base our analysis a lot on the result of our multivariate regression model, despite our efforts to ensure the reliability of our model, by removing influential points and NA values we might have removed invaluable insight about edge cases in our data. We also oversimplify our model by assuming LGDP and HLE are dependent. We also ignore other factors that might affect both variables. Finally, we do not fully consider the impact of our approach on all Continents. We give a general approach to improving happiness, but perhaps a Continent-to-Continent approach might be needed (We could test different baselines).

Part 3: Statistical Approaches Used

In this part, the report will detail how we describe the dataset and the pre-processing steps of building our multivariate regression model in the next part

3.1 Dataset

Step 1: Data types

The approaches we take at building our multivariate regression model depend on the data types of our Dataset e.g. We have categorical data to see how we deal with it in the pre-processing stage

```
> str(datasets)
'data.frame':   137 obs. of  8 variables:
 $ Country.name: chr  "Finland" "Denmark" "Iceland" "Israel" ...
 $ LGDP        : num  10.8 11 10.9 10.6 10.9 ...
 $ Support     : num  0.969 0.954 0.983 0.943 0.93 0.939 0.943 0.92 0.879 0.952 ...
 $ HLE         : num  71.2 71.2 72 72.7 71.5 ...
 $ Freedom     : num  0.961 0.934 0.936 0.809 0.887 0.948 0.947 0.891 0.915 0.887 ...
 $ Corruption  : num  0.182 0.196 0.668 0.708 0.379 0.202 0.283 0.266 0.345 0.271 ...
 $ Continent   : chr  "Europe" "Europe" "Europe" "Asia" ...
 $ Ladder_score: num  7.8 7.59 7.53 7.47 7.4 ...
```

Step 2: Descriptive Statistics (including Figure 1, Figure 2 and Figure 3)

The analysis of our multivariate regression model and its result is meaningless without the context of the dataset. We consider key concepts, such as mean, outliers, IQR and more… to get insight about the data and our result.

```
Country.name          LGDP           Support           HLE             Freedom          Corruption          Continent
Length:128         Min.   : 7.007   Min.   :0.3410   Min.   :51.53   Min.   :0.3820   Min.   :0.1460   Africa       :32
Class :character   1st Qu.: 8.595   1st Qu.:0.7272   1st Qu.:60.76   1st Qu.:0.7300   1st Qu.:0.6777   Asia         :33
Mode  :character   Median : 9.583   Median :0.8355   Median :65.88   Median :0.8030   Median :0.7745   Europe       :40
                   Mean   : 9.491   Mean   :0.8021   Mean   :65.03   Mean   :0.7936   Mean   :0.7267   North America:12
                   3rd Qu.:10.540   3rd Qu.:0.8982   3rd Qu.:69.41   3rd Qu.:0.8772   3rd Qu.:0.8460   Oceania      : 2
                   Max.   :11.660   Max.   :0.9830   Max.   :74.35   Max.   :0.9610   Max.   :0.9290   South America: 9
 Ladder_score      Combined
Min.   :1.859   Min.   :58.65
1st Qu.:4.896   1st Qu.:69.73
Median :5.750   Median :75.84
Mean   :5.618   Mean   :74.52
```

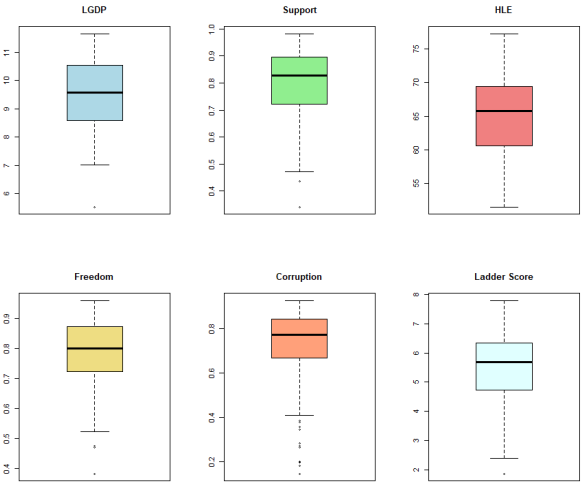


Figure 4: Box plot of variables

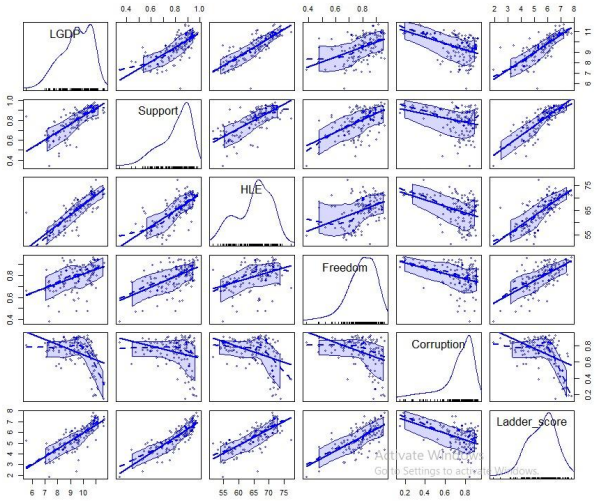


Figure 5 Scatter plots

Step 3: Null values

```
> rows_with_missing_values <- datasets[!complete.cases(datasets),]
>
> rows_with_missing_values
        Country.name  LGDP Support HLE Freedom Corruption Continent Ladder_score
99 State of Palestine 8.716   0.859  NA   0.694      0.836      Asia        4.908
```

3.2 Data Pre-Processing

In this stage, we will outline how we have prepared and cleaned the data for multivariate regression.

Outliers

Figure 4 of the boxplots reveals that there is a variety of outliers in different columns of the dataset, more prominently, in Corruption where we can see several points extending outside the lower quartile. Consequently, there may be a non-uniform distribution with some values deviating from the majority. This can skew the predictions of our model, potentially leading to less accurate results. Regardless, with further analysis, we discover that these outliers represent true observations and provide valuable insight into the behavior of the data particularly at the lower tails. For example, Finland's extremely low corruption score can have a huge impact on statistical analysis, but when we explore further we find that the top happiest countries all have extremely low corruption scores. This is reinforced by the fact that removing outliers causes a decrease in R-squared from 0.83 to 0.785, indicating that keeping outliers can help explain some of the variability in happiness score. Therefore, we chose to keep the outliers in our model.

Null Values

In the original dataset, the State of Palestine had an NA value for their HLE value. Since many statistical techniques assume the data is complete we considered removing this from our dataset. Before doing so, we wanted to make sure that this removal wouldn't cause any noticeable impacts on our model. In other words, we wanted to ensure that the State of Palestine didn't have any significant points of influence that may alter our results. By comparing the summary of our dataset with and without the State of Palestine, the mean of all values were completely unchanged when rounding to .01, except Corruption which saw a .01 decrease when the State of Palestine was removed from our data. This negligible impact on our data set gave us the confidence to remove the State of Palestine from our dataset.

Dealing with Categorical Data

To prepare our data for multivariate regression we use R's "as.factor" method to convert the Continents column into a factor, this is R's way of telling the computer we are dealing with categorical rather than continuous data. This has one main implication for our analysis, we need to choose a baseline Continent which will act as a reference point, and the effect of each variable will be quantified against this Continent. In our case, we are not comparing Continents directly, instead we are comparing the impact of our variables on the Continents as their effect is consistent for all Continents. Therefore, we let R choose the baseline for us. By default the variables are chosen based on alphabetical order, hence our baseline is Africa.

Normalization

We standardize all these variables to a mean of 0 and a standard deviation of 1 using z-score normalization. This brings the variables onto a common scale, allowing us to compare their variance directly irrespective of their original scale. This is particularly useful in the analysis of our multivariate regression model, where the coefficients can now be compared directly.

We transform each point separately according to their column using the following equation:

$$z = \frac{x - \mu}{\sigma}$$

Where z is the new value in the column σ is the variance of the column x is the current value in the column μ is the mean of the column. It should be noted that as Continents are categorical

and as a result, we have not normalized them, we should not be comparing our coefficients from our multivariate regression model with continents.

3.2 Building the Model

In this section, we will explore how we build the model and perform a preliminary analysis of our model. We will refine our model such that it meets all the required assumptions for the model to be considered accurate.

Building the Model

We will model the independent factors Freedom, Support, LGDP, HLE and Continent on the dependent factor Ladder Score. We do this to determine how changes in the independent variables are associated with the changes in the dependent variable. We do this by fitting our data to a multivariate regression model.

The general form for multivariate regression with p predictors is given by the formula below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Where Y is the dependent variable, β_0 is the intercept, $\beta_0, \beta_1 \ldots . \beta_p$ are the dependent variables of Y and ϵ is the error term representing variation in Y not explained by the X variable.

We estimate the β values using the least squares method which minimizes the sum of the squared differences between observed and predicted values. The equation can be seen below.

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}))^2$$

Variable combination

We attempted a brute-force attempt at trying to determine which combination of variables has the highest model accuracy. We ran our multivariate regression model under every possible combination of the variables. When we model Support as the only dependent factor of happiness we get an R-squared of 0.7023386, when we add LGDP we get a score of 0.7608284, when we add Corruption we get a score of 0.8116132, when we add Freedom we get score of 0.825144 and finally, when we add Continent we get a score of 0.8468101. The effect of adding HLE is nominal.

We aim to have a model that is the least complex (has the least variables) but has the highest R-squared score. In our case, we find that all these variables add significant performance to our model, except HLE perhaps because it is dependent on LGDP and adds no new information to the model

Influential Points

Influential points have a disproportionate effect on multivariate regression lines. In general, the model should avoid points with such huge influence, as they may skew the result of the model leading to less reliable predictions. This occurs when the anomalies are not representative of the general trend which we are trying to generalize. This can lead to issues such as heteroscedasticity due to less accurate predictions at certain points, causing uneven residuals.

In our model, Figure 6 suggests that the model has no visible influential points. However, further analysis of the crook distance reveals that countries like {Israel, Hong Kong, Botswana and Lebanon} are above the threshold of 4/(n-p-1) (where n is the number of observations and p is the number of model parameters).This threshold is associated with the points having a disproportionate influence on the model in a variety of literature.

When we remove these influential points our model accuracy (from an R-squared score of 0.8297 to 0.8506). This suggests that these points are masking the true trend of the data that is consistent with our analysis e.g. Lebanon has an extremely low happiness score despite its high LGDP, leading to skewed regression lines and less accurate predictions likely causing issues such as heteroscedasticity. Interestingly, if we measure the performance of our model using K-fold we get a R-squared of 0.836 instead of 0.8506 without K-fold. This suggests although the model improved by the removal of outliers, this may not generalize across all subsets. Instead removing the influential points might cause overfitting. See next section to see why we decided to remove the influential points

Homoscedasticity

To rule out heteroscedasticity, we can hope to see an even distribution of residuals around the zero y plane in our graph comparing the fitted values of our model (see Figure 7) to the square root of the residuals.

From this graph, we do see a pretty even distribution of residuals around the y plane of zero, indicating no heteroscedasticity. We do notice that the median line of these residuals peaks in the middle and tails downward at each end, so performing a Breusch-Pagan test can help us confirm homoscedasticity. After doing so, we received a p-value of 0.08635. Since the p-value is greater than 0.05, we fail to reject the null hypothesis and can not say that our regression model does not suffer from heteroscedasticity.

Furthermore, in our analysis we find that removing HLE, the p-value dropped to 0.02493, suggesting the presence of heteroscedasticity. This result, combined with the need to avoid multicollinearity, led us to combine HLE with LGDP in our final model, which showed stronger evidence of homoscedasticity. It should be noted that before we removed the influential points the p-value was 0.04149, indicating statistical significance of heteroscedasticity. Therefore, to prevent heteroscedasticity we also need to remove the influential points.

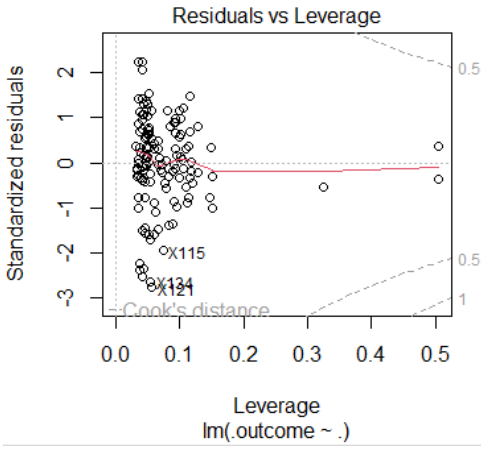


Figure 6: Residual vs Leverage plot

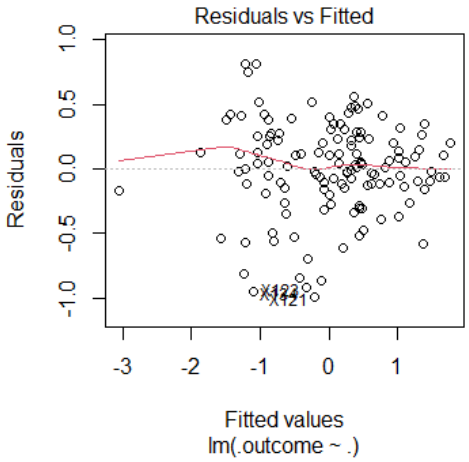


Figure 7: Residual vs Fitted plot

Multicollinearity

The high correlation of HLE and LGDP raises concerns about multicollinearity. By performing a VIF test, we confirm that multicollinearity is indeed an issue. In the first iteration of the model,

the VIF values for LGDP and HLE were 4.53 and 5.302042 respectively. A value of 5 and above suggests high multicollinearity, HLE with a value of 5.3 falls in that criteria.

There are a couple of ways we can address this problem. The two ways that worked best were to remove HLE or make a composite variable LGDPHLE which is column HLE + LGDP. The R-squared score for removing HLE is 0.8656 (there is no change) combining HLE and LGDP has a lower score of 0.8506 (a decrease in score) assuming that we have removed the influential points.

Indeed, if we remove the HLE column we get a VIF score of 3.35 for LGDP and if we create a composite variable we get a VIF score of 4.21. These are both below 5, therefore multicollinearity is not an issue.

Given that the R-squared score is higher and the VIF score is lower when we consider removing HLE, however, see the section on Homoscedastic for why we ended up combining HLE and LGDP instead.

Linearity

The statistically significant correlation between our independent and dependent variables suggests that our dependent variables have a strong linear relationship. In the inspection of our qqnorm plot vs residuals (see Figure 8) and predicted we observe that the points in the middle section of the plot followed the (x = y) dashed line quite closely, which suggests the central part of the data is normally distributed. However, at the tails, particularly the lower tails, some points deviate significantly from the line; this could be because of the presence of outliers like columns 123 and 134.

Furthermore, in our residual vs fitted plot (Figure 7), we see that our values are evenly distributed around the y-axis. There are also no clear non-linear patterns present when comparing our residuals with fitted values. The insights from both of these plots give us the confidence to continue with our assumption that our model has a linear relationship between the predicted values and their residuals.

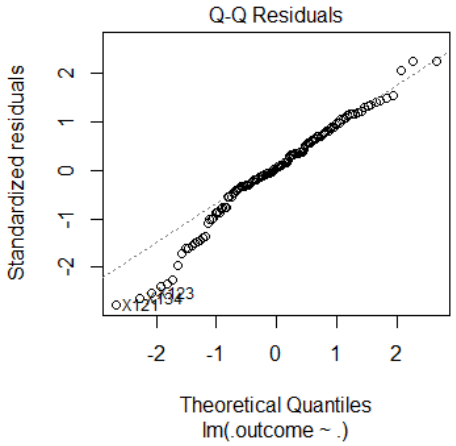


Figure 8: QQ-plot of residuals

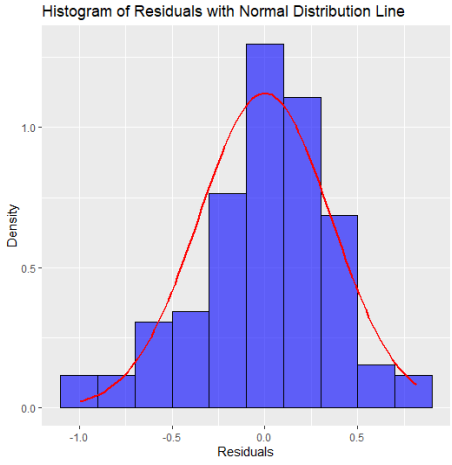


Figure 9: Histogram Plot of Residuals

Normal Distribution

The sample size is greater than 50, therefore slight deviations in normal distributions can be less of a concern due to the Central Limit Theorem which suggests the distribution of means will

be near normal if the sample size is large. Therefore, we can assume our dataset is normally distributed.

Analysis of Performance of Model
K-fold Validation
Overfitting occurs when our model fails to generalize patterns and similarities between input data, instead, the model becomes too complex and becomes too closely fitted to our dataset. Overfitted models are not useful in real-life scenarios as their predictions are less accurate when dealing with new data as they fail to generalize the underlying structure of our data.

We use a technique called K-fold cross-validation which splits our dataset into K equally sized subsets and trains the model on each of the subsets. For each subset, it calculates the R-squared metric and finds the average performance across all of the test sets.

This serves as a more robust method for measuring the performance of our model than calculating the R-squared score on a single training set as the model has to perform well on each K subset of the data. Thereby, by using cross-validation we can build a model that does not just memorize the training data but learns to generalize its patterns.

K-fold has only one hyperparameter for our model, in general, the larger K the less biased towards overestimating the true expected error but higher variance and higher running time. In most cases k is usually 5 or 10, we will stick to this and use k = 10 considering our dataset is not too large.
Result

y = 0.12241 + 0.40421 × Support + 0.22410 × Freedom - 0.32680 × ContinentAsia - 0.05557 × ContinentAfrica + 0.11738 × ‘ContinentNorth America’ - 0.12573 × ‘ContinentSouth America’ + 0.01137 × ContinentOceania - 0.15325 × Corruption + 0.28098 × LGDPHLE

By assuming linearity, homoscedasticity, and the absence of multicollinearity we ensure that the estimated coefficients are accurate and meaningful. A model with these assumptions is more likely to produce reliable predictions.

Since our model incorporates multiple predictor variables, we’ll also be looking at the adjusted R-squared value instead of the R-squared value of the model to measure the performance of our model. Looking at the result of K-fold Validation, we obtain an R-squared of 0.836, which means that 83.6% of the variance in ladder score is explained by our model. An RMSE of 0.375 when compared to the range of the ladder score, is relatively low indicating that the model’s predictions are generally close to the actual values. The MAE is even smaller than RMSE, indicating that the average error in predictions is quite low. Furthermore, the qqnorm plot (see Figure 8) looks to be normally distributed, this is reinforced by the histogram which appears to closely resemble the red line which resembles a normal distribution (see Figure 9). Hence, our data fit well with the model and generalizes well.

Figure 10: Result of K-fold Cross validation

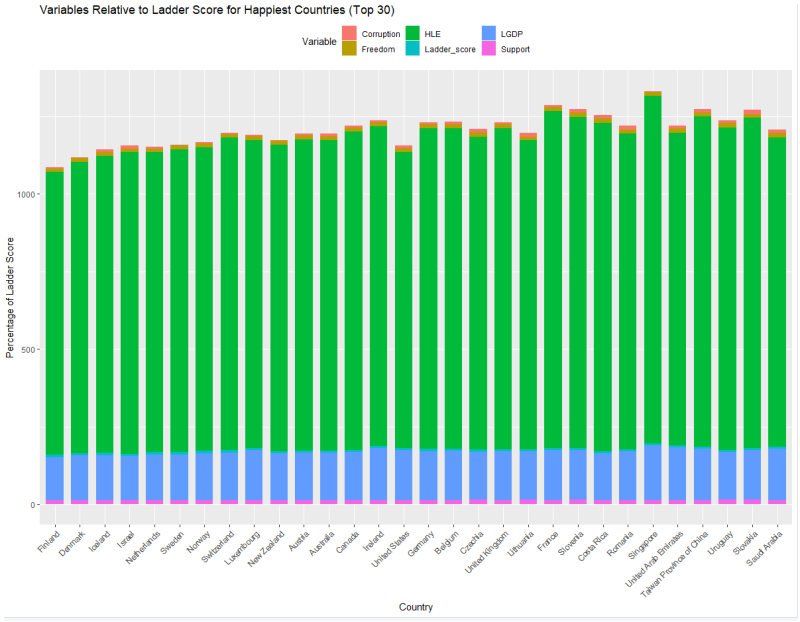
```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 118, 117, 117, 119, 118, 119, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3633554  0.8363516  0.2850699
```

Appendix A

```
################################################Libraries#########################################
library(caret)
library(dplyr)
library(ggplot2)
library(car)
library(corrplot)
library(RColorBrewer)
library(maps)
library(sf)
library(tidyverse)
library(stringr)
library(dplyr)  # For data manipulation
library(broom)  # For tidying model results
library(lmtest)

################################################Libraries#########################################



setwd("C:/Users/Matty/Desktop") # Set directory

set.seed(123) #Define a seed for replicable results

datasets <- read.csv("happy.csv") #open happy.csv

summary(datasets) #summary of dataset

str(datasets)  #summary of datatypes

datasets <- na.omit(datasets)

# Convert 'Continent' to a factor
datasets$Continent <- as.factor(datasets$Continent)

# Creating new variable that combines LGDP and HLE
datasets$LGDPHLE <- datasets$HLE + datasets$LGDP



################################################ Boxplots of variables ##############################

boxplot(datasets$LGDP, main = "LGDP", col = "lightblue")
boxplot(datasets$Support, main = "Support", col = "lightgreen")
boxplot(datasets$HLE, main = "HLE", col = "lightcoral")
boxplot(datasets$Freedom, main = "Freedom", col = "lightgoldenrod")
boxplot(datasets$Corruption, main = "Corruption", col = "lightsalmon")
boxplot(datasets$Ladder_score, main = "Ladder Score", col = "lightcyan")
boxplot(datasets$LGDPHLE, main = "LGDP+HLE", col = "purple")
################################################ Boxplots of variables ##############################

################################################ Normality test for variables ######################

qqnorm(datasets$LGDP)
qqline(datasets$LGDP, col = 2)  # Add a reference line
mtext("QQ Plot - LGDP", side = 3, line = -3)  # Adjust the line parameter

qqnorm(datasets$Support)
qqline(datasets$Support, col = 2)  # Add a reference line
mtext("QQ Plot - Support", side = 3, line = -3)  # Adjust the line parameter

qqnorm(datasets$HLE)
qqline(datasets$HLE, col = 2)  # Add a reference line
mtext("QQ Plot - HLE", side = 3, line = -3)  # Adjust the line parameter

qqnorm(datasets$Freedom)
qqline(datasets$Freedom, col = 2)  # Add a reference line
mtext("QQ Plot - Freedom", side = 3, line = -3)  # Adjust the line parameter

qqnorm(datasets$Corruption)
qqline(datasets$Corruption, col = 2)  # Add a reference line
mtext("QQ Plot - Corruption", side = 3, line = -3)  # Adjust the line parameter

qqnorm(datasets$Ladder_score)
qqline(datasets$Ladder_score, col = 2)  # Add a reference line
mtext("QQ Plot - Ladder Score", side = 3)

################################################ Normality test for variables ######################


################################################ Plot Correlation Matrix ############################

corr <- cor(datasets[, c("LGDP", "Support", "HLE", "Freedom", "Corruption", "Ladder_score")])

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))(200)

corrplot(
  corr,
  method = "color",
  col = col,
  type = "upper",
  order = "hclust",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45,
  diag = FALSE
)

################################################ Plot Correlation Matrix ############################
```

```
################################################## Find outliers##############################################

get_outliers <- function(data, variable) {
  q1 <- quantile(data[[variable]], 0.25)
  q3 <- quantile(data[[variable]], 0.75)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
  outliers <- data[data[[variable]] < lower_bound | data[[variable]] > upper_bound, ]
  if (nrow(outliers) > 0) {
    outliers$Outlier_Variable <- variable
  }
  return(outliers)
}

# Find Outliers in All Numeric Variables
variables_to_check <- names(datasets)[sapply(datasets, is.numeric)]
all_outliers <- data.frame()

for (variable in variables_to_check) {
  outliers_for_variable <- get_outliers(datasets, variable)
  all_outliers <- rbind(all_outliers, outliers_for_variable)
}

print(all_outliers)
################################################## Find outliers ############################################

##################################################Normalize Dataset ############################################

# Normalize the numerical variables
numerical_cols <- c("Support", "Freedom", "Corruption", "LGDPHLE", "Ladder_score")  # specify your numerical columns

datasets[numerical_cols] <- scale(datasets[numerical_cols])

##################################################Normalize Dataset#############################################

##################################################Initial Model#################################################

# Linear Model
model <- lm(Ladder_score ~ Support + Freedom + Continent + Corruption + LGDP + HLE, datasets)

summary(model)

print(vif(model))

plot(model)
##################################################Initial Model#################################################


################################################## Find and Remove Influential Points############################################

# Cook's Distance
cooks_dist <- cooks.distance(model)

# Threshold for identifying influential points
n <- nrow(datasets)
k <- length(coefficients(model))
threshold <- 4 / (n - k - 1)

# Identify Influential Points
influential_points <- which(cooks_dist > threshold)
influential_countries <- datasets$Country[influential_points]

print(influential_countries)
print(influential_points)

influential_points

# Removing Influential Points
cleaned_data <- datasets[-influential_points, ]

################################################## Find and Remove Influential Points############################################


################################################## Cross-Validation ############################################

# Cross-Validation
train_control <- trainControl(method = "cv", number = 10)
cv_model <- train(Ladder_score ~ Support + Freedom + Continent + Corruption + LGDPHLE + (Support * Corruption) + (Support & Freedom),
                  data = cleaned_data,
                  method = "lm",
                  trControl = train_control)

################################################## Cross-Validation ############################################
```

```
################################################ Assumption tests ###########################################

# Checking Normality of Residuals
qqnorm(residuals(cv_model$finalModel))
qqline(residuals(cv_model$finalModel))

# Checking Homoscedasticity
plot(cv_model$finalModel)

# Check for Multicollinearity
print(vif(cv_model$finalModel))

residuals_cv <- residuals(cv_model$finalModel)

# Create a data frame for the residuals
residuals_df <- data.frame(Residuals = residuals_cv)


dw_test <- durbinWatsonTest(cv_model$finalModel)

# Print the result
print(dw_test)

bp_test <- bptest(cv_model$finalModel)

bp_test

shapiro_test <- shapiro.test(residuals(cv_model$finalModel))

residuals_df <- data.frame(Residuals = residuals(cv_model$finalModel))

# Create a histogram of the residuals
ggplot(residuals_df, aes(x = Residuals)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, fill = "blue", color = "black", alpha = 0.6) +
  stat_function(fun = dnorm, args = list(mean = mean(residuals_df$Residuals), sd = sd(residuals_df$Residuals)), color = "red", size = 1) +
  labs(title = "Histogram of Residuals with Normal Distribution Line", x = "Residuals", y = "Density")

################################################ Assumption tests ###########################################

print(shapiro_test)
print(summary(cv_model))
```



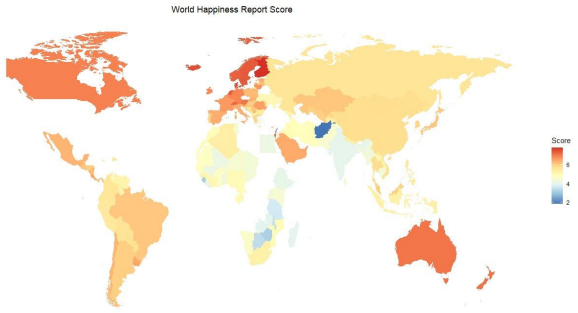


Figure 11: Variables Relative to Ladder Score for Happiest Countries

Figure 12: Happiness Score World Map