World Happiness

Linda Jones
6/10/2019

```
setwd("~/R/Dataset")
WorldHappiness<- read.csv("~/R/Dataset/WorldHappiness Capstone.csv")
View(WorldHappiness)
WorldHappiness<- read.csv("~/R/Dataset/WorldHappiness Capstone.csv")
View(WorldHappiness)
attach(WorldHappiness)
```

#INTRODUCTION

This report is being submitted to satisfy the "Choose-your-own-project" Capstone Project requirements of the 'HarvardX: PH125.9x Data Science course. The World Happiness 2017 dataset available on Kaggle (https://www.kaggle.com/unsdsn/world-happiness) was used for this project. The goal was to use R and RStudio and employ skills learned during the course to create a machine learning algorithm that could help solve a problem.

#OVERVIEW

The World Happiness Report is genersted by the United Nations Sustainable Development Solution Network. There are a total of six factors the study uses to calculate happiness in 2017. The six factors are economy, family, health, freedom, trust, and generosity. The six factors were added together to generate the happiness score. The total happiness score is used to determine a countries happiness rank. The highest World Happiness score receives a ranking of one (1). The happiness scores are ranked in descending order from largest to smallest. The World Happiness report suggests that happiness in the United States is falling as a result of social issues. Therefore, the goal for this study is to determine the correlation between happiness.Score, freedom and family.

The freedom and family variables were chosen because these two variables are aligned with, or affected by social issues. A multiple linear regression model is used to predict the likelihood of a countries ranking for overall happiness improving when freedom and family receive high scores. The data is also assessed against the Pearson correlation coefficient.

This report has five parts: (1) Dataset Description, (2) Data Exploration, (3) Methods & Analysis, (4) Results, and (5) Conclusion. The data's dimensions and attributes are assessed in the dataset exploration section to ensure it is suitable for the intended purpose. In the data exploration section, data attributes, such as the country, happiness.Score, and freedom are explored. The methods employed to predict the ranking and the corresponding analysis are presented in the methods and analysis section. The final sections of the report present the results and conclusion.

#DATASET DESCRIPTION

A Worldhappiness 2017 dataset was used for this project. The excel file was converted into cvs format and set as the working directory. Then the file was imported as a text file so it

could be attached to the file making it available for others to view the data and run the script in R. The following libraries were also loaded:

#load libraries library(glm2) library(caret) library(rpart) library(caret) library(dplyr) library(MASS) library(ggplot2) library(reshape2)

#DATA EXPLORATION

Exploring the data involved looking at variables and examining them for accuracy, completeness, emerging patters and to ensure the data can support the statistical analysis. Results of the examination appear below.

#View the headers

```
head(WorldHappiness)
```

```
##    ï..Country Happiness.Rank Happiness.Score Whisker.high Whisker.low
## 1      Norway              1           7.537     7.594445    7.479556
## 2     Denmark              2           7.522     7.581728    7.462272
## 3     Iceland              3           7.504     7.622030    7.385970
## 4 Switzerland              4           7.494     7.561772    7.426227
## 5     Finland              5           7.469     7.527542    7.410458
## 6 Netherlands              6           7.377     7.427426    7.326574
##   Economy..GDP.per.Capita.   Family Health..Life.Expectancy.   Freedom
## 1                1.616463 1.533524                0.7966665 0.6354226
## 2                1.482383 1.551122                0.7925655 0.6260067
## 3                1.480633 1.610574                0.8335521 0.6271626
## 4                1.564980 1.516912                0.8581313 0.6200706
## 5                1.443572 1.540247                0.8091577 0.6179509
## 6                1.503945 1.428939                0.8106961 0.5853845
##   Generosity Trust..Government.Corruption. Dystopia.Residual
## 1  0.3620122                     0.3159638          2.277027
## 2  0.3552805                     0.4007701          2.313707
## 3  0.4755402                     0.1535266          2.322715
## 4  0.2905493                     0.3670073          2.276716
## 5  0.2454828                     0.3826115          2.430182
## 6  0.4704898                     0.2826618          2.294804
```

#View the dimension of the dataset

```
dim(WorldHappiness)
```

```
## [1] 155  12
```

#View the dataset attributes - class

```
sapply(WorldHappiness,class)
```

```
##                    ï..Country                Happiness.Rank
##                      "factor"                     "integer"
##               Happiness.Score                  Whisker.high
##                     "numeric"                     "numeric"
```

```
##                   Whisker.low      Economy..GDP.per.Capita.
##                     "numeric"                     "numeric"
##                        Family      Health..Life.Expectancy.
##                     "numeric"                     "numeric"
##                       Freedom                    Generosity
##                     "numeric"                     "numeric"
## Trust..Government.Corruption.             Dystopia.Residual
##                     "numeric"                     "numeric"
```

#Summary of WorldHappiness 2017 dataset

```
summary(WorldHappiness)
```

```
##        ï..Country  Happiness.Rank  Happiness.Score  Whisker.high
##  Afghanistan:  1   Min.   :  1.0   Min.   :2.693   Min.   :2.865
##  Albania    :  1   1st Qu.: 39.5   1st Qu.:4.505   1st Qu.:4.608
##  Algeria    :  1   Median : 78.0   Median :5.279   Median :5.370
##  Angola     :  1   Mean   : 78.0   Mean   :5.354   Mean   :5.452
##  Argentina  :  1   3rd Qu.:116.5   3rd Qu.:6.101   3rd Qu.:6.195
##  Armenia    :  1   Max.   :155.0   Max.   :7.537   Max.   :7.622
##  (Other)    :149
##   Whisker.low    Economy..GDP.per.Capita.     Family
##  Min.   :2.521   Min.   :0.0000           Min.   :0.000
##  1st Qu.:4.375   1st Qu.:0.6634           1st Qu.:1.043
##  Median :5.193   Median :1.0646           Median :1.254
##  Mean   :5.256   Mean   :0.9847           Mean   :1.189
##  3rd Qu.:6.007   3rd Qu.:1.3180           3rd Qu.:1.414
##  Max.   :7.480   Max.   :1.8708           Max.   :1.611
##
##  Health..Life.Expectancy.    Freedom         Generosity
##  Min.   :0.0000           Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3699           1st Qu.:0.3037   1st Qu.:0.1541
##  Median :0.6060           Median :0.4375   Median :0.2315
##  Mean   :0.5513           Mean   :0.4088   Mean   :0.2469
##  3rd Qu.:0.7230           3rd Qu.:0.5166   3rd Qu.:0.3238
##  Max.   :0.9495           Max.   :0.6582   Max.   :0.8381
##
##  Trust..Government.Corruption. Dystopia.Residual
##  Min.   :0.00000               Min.   :0.3779
##  1st Qu.:0.05727               1st Qu.:1.5913
##  Median :0.08985               Median :1.8329
##  Mean   :0.12312               Mean   :1.8502
##  3rd Qu.:0.15330               3rd Qu.:2.1447
##  Max.   :0.46431               Max.   :3.1175
##
```

#Review the structure of the WorldHappines 2017 dataset

```
str(WorldHappiness)
```

```
## 'data.frame':    155 obs. of  12 variables:
##  $ ï..Country                   : Factor w/ 155 levels "Afghanistan",..: 1
05 38 58 133 45 99 26 100 132 7 ...
##  $ Happiness.Rank               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Happiness.Score              : num  7.54 7.52 7.5 7.49 7.47 ...
##  $ Whisker.high                 : num  7.59 7.58 7.62 7.56 7.53 ...
##  $ Whisker.low                  : num  7.48 7.46 7.39 7.43 7.41 ...
##  $ Economy..GDP.per.Capita.     : num  1.62 1.48 1.48 1.56 1.44 ...
##  $ Family                       : num  1.53 1.55 1.61 1.52 1.54 ...
##  $ Health..Life.Expectancy.     : num  0.797 0.793 0.834 0.858 0.809 ...
##  $ Freedom                      : num  0.635 0.626 0.627 0.62 0.618 ...
##  $ Generosity                   : num  0.362 0.355 0.476 0.291 0.245 ...
##  $ Trust..Government.Corruption.: num  0.316 0.401 0.154 0.367 0.383 ...
##  $ Dystopia.Residual            : num  2.28 2.31 2.32 2.28 2.43 ...
```

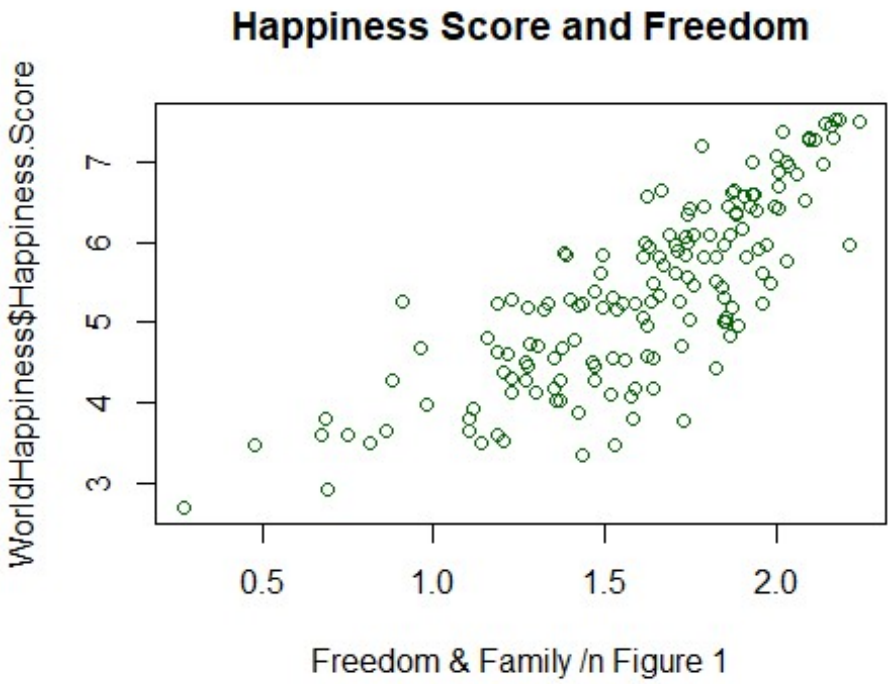
The data is deemded sufficient and ready for analysis after exploring the data, visually inspecting it, and verifying nothing is missing.

#METHODS & ANALYSIS

Multiple linear regression is appropriate to investigate more than one variable. Multiple linear regression goes beyond simple linear regression because it uses two or more variables and can be linear or nonlinear. For this project, the two independent variables are family and freedom. Freedom and family are the x variable and Happiness.Score is the y variable. The goal is to determine if increased levels of freedom and family result in an improved happiness score.

#Multiple Linear Regression #Create a scatterplot with freedom and family as the x value and happiness.score as the y value to visualize the data

```
plot(WorldHappiness$Freedom+WorldHappiness$Family,WorldHappiness$Happiness.Sc
ore,main="Happiness Score and Freedom",col="dark green", xlab="Freedom & Fami
ly /n Figure 1")
```



#Correlation the relationsip between Freedom and Family and happiness.score

```
cor(WorldHappiness$Freedom+WorldHappiness$Family,WorldHappiness$Happiness.Sco
re)

## [1] 0.8017802
```

The correlation is .0817802, which indicates that there is variation around the line of best fit. When assessed together there appers to be a negative correlation between freedom, family and the happiness.Score. This suggests that as one one variable decreasess in value the other will increase. However, does not imply that there is a cause and effect relationship between these two specific variables. There are four other variables that contribute to the happiness score.

#Fit the linear regression model. Using this to scale the relationship between the freedom,family and generosity (x) and happiness.Score (y)

```
lr_model<-lm(WorldHappiness$Happiness.Score~WorldHappiness$Freedom+WorldHappi
ness$Family)
lr_model

##
## Call:
## lm(formula = WorldHappiness$Happiness.Score ~ WorldHappiness$Freedom +
##     WorldHappiness$Family)
##
## Coefficients:
```

```
##            (Intercept)  WorldHappiness$Freedom   WorldHappiness$Family
##                  1.496                   2.303                   2.453
```

A correlation test was performed to assess the association between freedom and family. When using the cor.test function, the Pearson's Correlation is the default. Results appear below:

```
cor.test(WorldHappiness$Freedom,WorldHappiness$Family)

##
##  Pearson's product-moment correlation
##
## data:  WorldHappiness$Freedom and WorldHappiness$Family
## t = 5.807, df = 153, p-value = 3.558e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2865125 0.5460325
## sample estimates:
##       cor
## 0.4249658
```

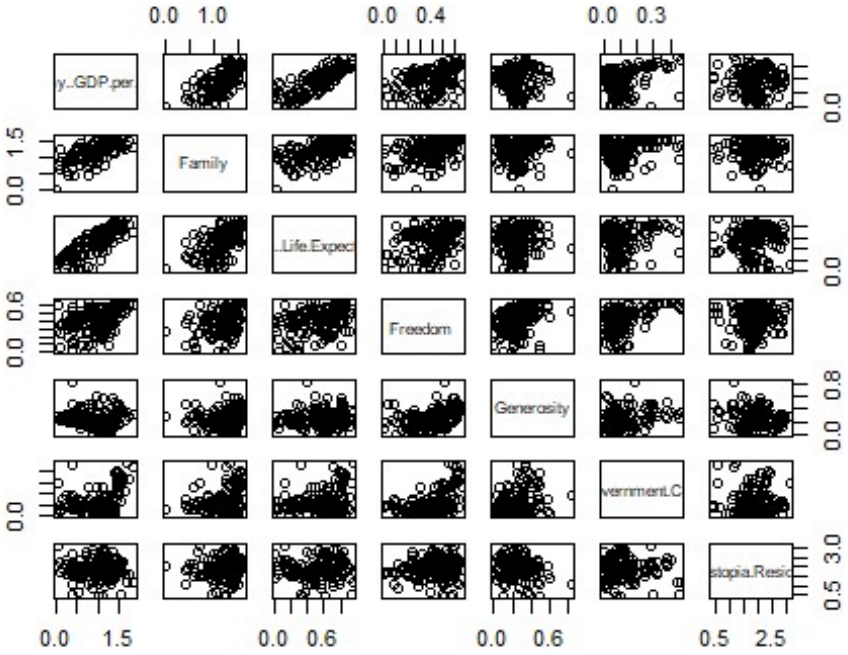
#Plot Pearson's product-moment correlation

```
cor(WorldHappiness[,6:12])

##                                Economy..GDP.per.Capita.     Family
## Economy..GDP.per.Capita.                     1.00000000 0.68829631
## Family                                       0.68829631 1.00000000
## Health..Life.Expectancy.                     0.84307664 0.61208006
## Freedom                                      0.36987339 0.42496576
## Generosity                                  -0.01901125 0.05169263
## Trust..Government.Corruption.                0.35094410 0.23184139
## Dystopia.Residual                            0.02422642 0.07050576
##                                Health..Life.Expectancy.    Freedom
## Economy..GDP.per.Capita.                     0.84307664 0.36987339
## Family                                       0.61208006 0.42496576
## Health..Life.Expectancy.                     1.00000000 0.34982679
## Freedom                                      0.34982679 1.00000000
## Generosity                                   0.06319149 0.31608271
## Trust..Government.Corruption.                0.27975198 0.49918279
## Dystopia.Residual                            0.05496328 0.08192597
##                                 Generosity Trust..Government.Corruption.
## Economy..GDP.per.Capita.      -0.01901125                     0.35094410
## Family                         0.05169263                     0.23184139
## Health..Life.Expectancy.       0.06319149                     0.27975198
## Freedom                        0.31608271                     0.49918279
## Generosity                     1.00000000                     0.29415945
## Trust..Government.Corruption.  0.29415945                     1.00000000
## Dystopia.Residual             -0.11662674                    -0.02275506
##                                Dystopia.Residual
## Economy..GDP.per.Capita.              0.02422642
```

```
## Family                                0.07050576
## Health..Life.Expectancy.              0.05496328
## Freedom                               0.08192597
## Generosity                           -0.11662674
## Trust..Government.Corruption.        -0.02275506
## Dystopia.Residual                     1.00000000
```

```
plot((WorldHappiness[,6:12]))
```



We assess the scatterplots to identify values of zero becuase this indicates the variables are not associated. Values above zero are indicative of positive associations between the variables while negative values are indicative of an inverse relationship between variables, which means as one increases the other decreases. Unlike multiple linear regression, the Pearson's Correlation does not distinguish between the dependent and independent variables. In this plot all six variables appear making it easier to visually inspect the data to determine which, if any, of the variables have a postivie correlation.

#summary of the model results

```
summary(lr_model)
```

```
##
## Call:
## lm(formula = WorldHappiness$Happiness.Score ~ WorldHappiness$Freedom +
##     WorldHappiness$Family)
##
## Residuals:
```

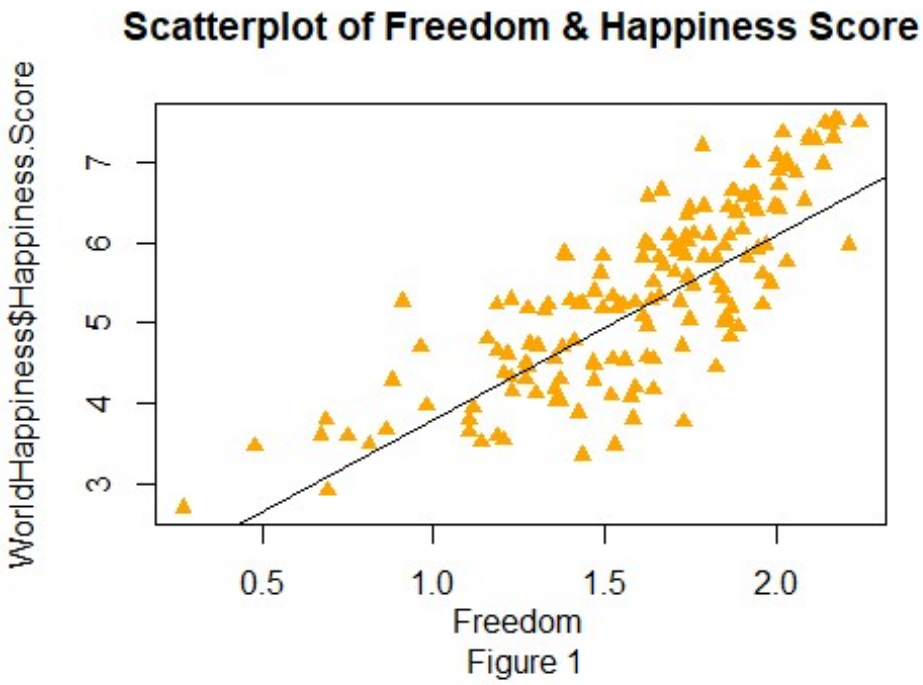
```
##      Min       1Q   Median       3Q      Max
## -1.89019 -0.49980  0.09403  0.51089  1.58109
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.4959     0.2406   6.218 4.65e-09 ***
## WorldHappiness$Freedom     2.3033     0.4037   5.705 5.91e-08 ***
## WorldHappiness$Family      2.4532     0.2108  11.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6803 on 152 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.6383
## F-statistic: 136.9 on 2 and 152 DF,  p-value: < 2.2e-16
```

The standard error in this model is .2522. The slope for freedom is 1.4532. The Root Means Squared Error or Residual standard error is 0.6818 on 151 degrees of freedom.

#Scatterplot with Freedom

```
plot(WorldHappiness$Freedom+WorldHappiness$Family,WorldHappiness$Happiness.Score,main="Scatterplot of Freedom & Happiness Score",pch=17, col="orange",xlab="Freedom \n Figure 1")
abline(lr_model)
```

```
## Warning in abline(lr_model): only using the first two of 3 regression
## coefficients
```



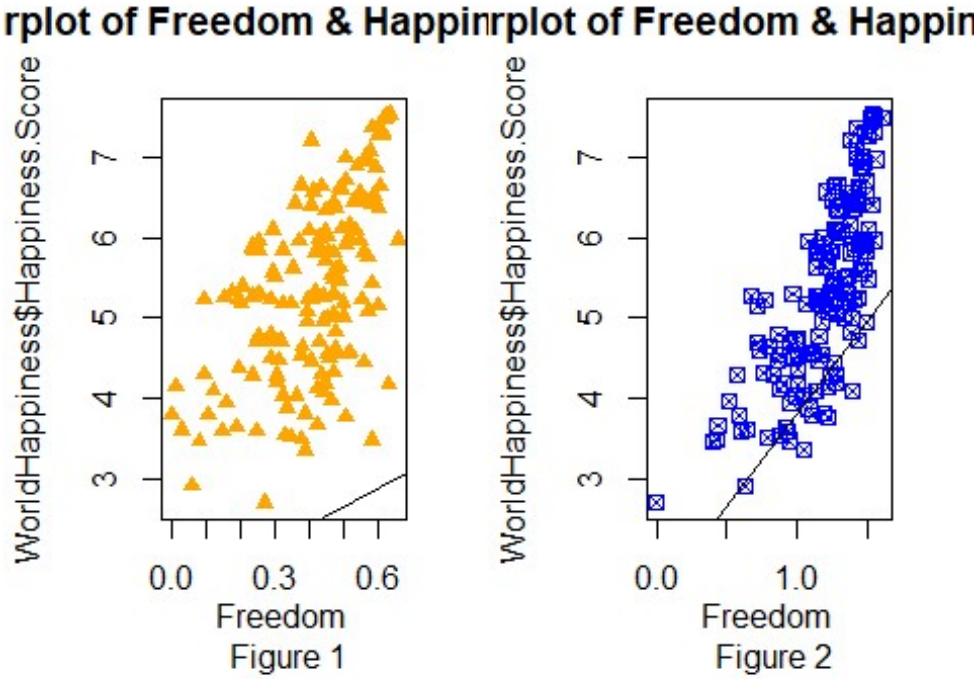
There appears to be a linear relationship between freedom, family and the happiness score.

```
par(mfrow=c(1,2))
plot(WorldHappiness$Freedom,WorldHappiness$Happiness.Score,main="Scatterplot
of Freedom & Happiness Score",pch=17, col="orange",xlab="Freedom \n Figure 1"
)
abline(lr_model)

## Warning in abline(lr_model): only using the first two of 3 regression
## coefficients

plot(WorldHappiness$Family,WorldHappiness$Happiness.Score,main="Scatterplot o
f Freedom & Happiness Score",pch=7, col="blue",xlab="Freedom \n Figure 2")
abline(lr_model)

## Warning in abline(lr_model): only using the first two of 3 regression
## coefficients
```



#Identify the coefficient intercept

```
confint(lr_model)

##                              2.5 %   97.5 %
## (Intercept)            1.020580 1.971281
## WorldHappiness$Freedom 1.505586 3.100937
## WorldHappiness$Family  2.036639 2.869668
```

This value helps to determine or measure the usefulness of the regression prediction. Based on the results there is a 97.5% likelihood that freedom and family will be within the confidence interval of the regression line.

#Identify the model coefficient

```
coef(lr_model)
```

```
##            (Intercept) WorldHappiness$Freedom  WorldHappiness$Family
##               1.495930               2.303262               2.453153
```

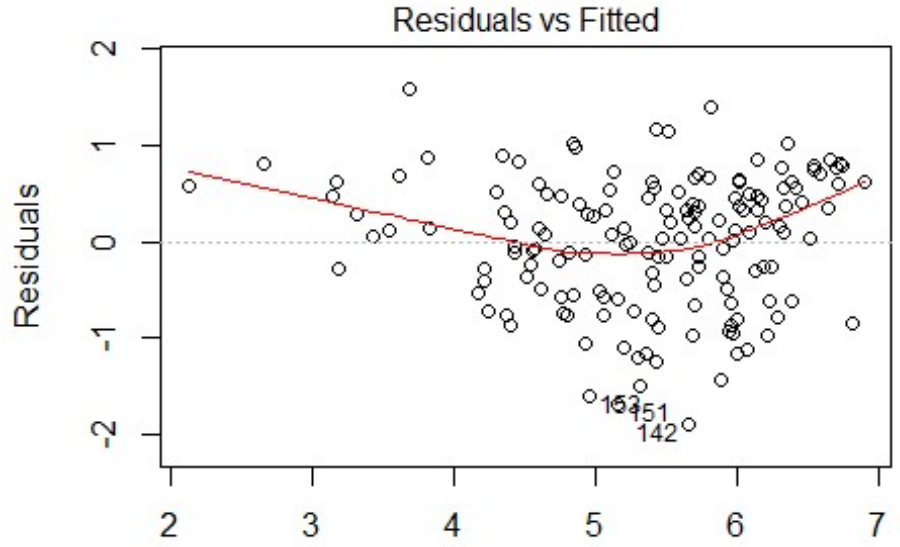
The coefficient value is used to measure the strength of the linear relationship between freedom, family, and the happiness score. Based on the results the intercept is 3.596327.

```
anova(lr_model)
```

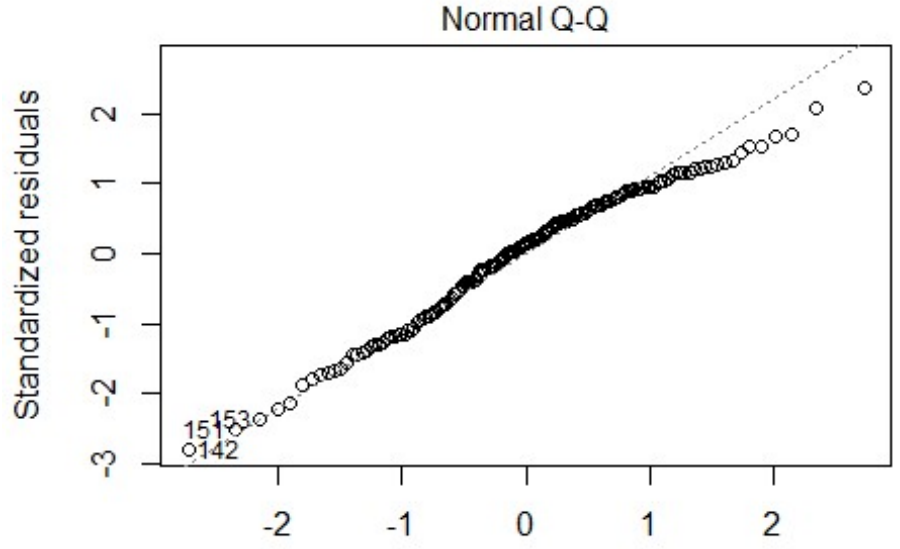
```
## Analysis of Variance Table
##
## Response: WorldHappiness$Happiness.Score
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## WorldHappiness$Freedom     1 64.059  64.059  138.41 < 2.2e-16 ***
## WorldHappiness$Family      1 62.665  62.665  135.40 < 2.2e-16 ***
## Residuals                152 70.346   0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Plot the lr_model

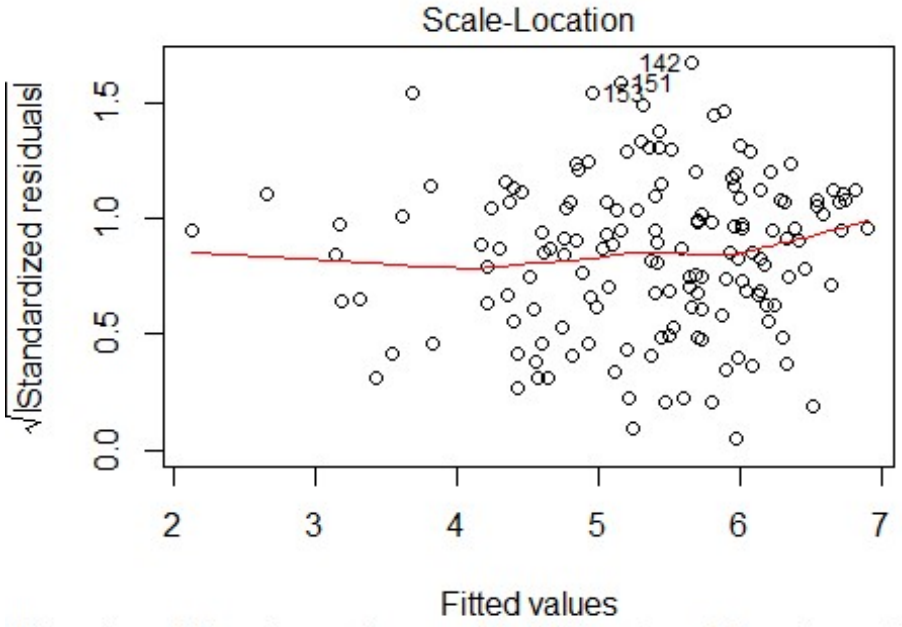
```
plot(lr_model)
```



dHappiness$Happiness.Score ~ WorldHappiness$Freedom + World



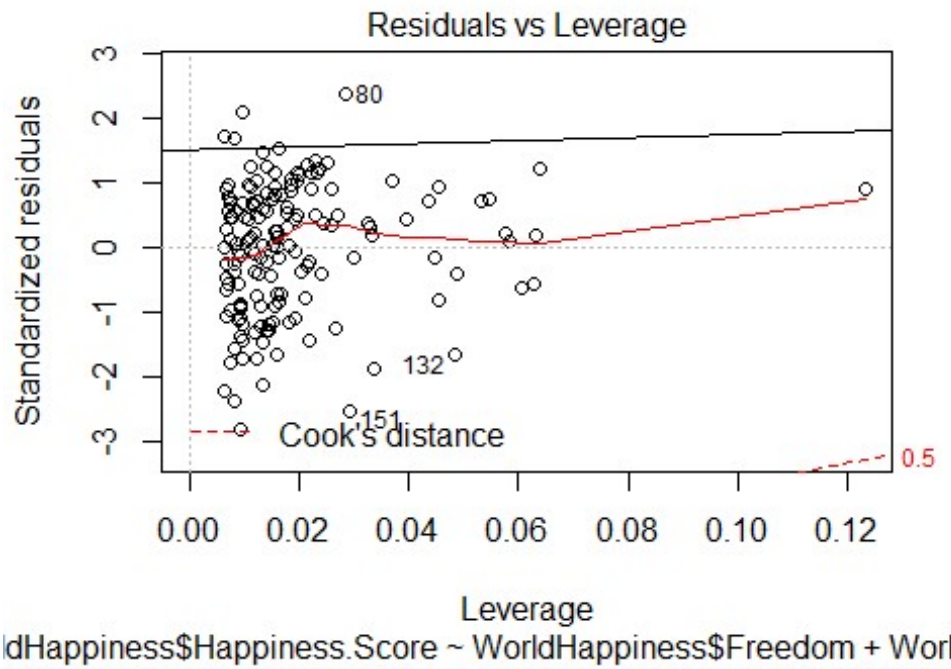
dHappiness$Happiness.Score ~ WorldHappiness$Freedom + World



dHappiness$Happiness.Score ~ WorldHappiness$Freedom + World

```
abline(lr_model)

## Warning in abline(lr_model): only using the first two of 3 regression
## coefficients
```



dHappiness$Happiness.Score ~ WorldHappiness$Freedom + World

```
par(mfrow=c(2,2))
```

The residual vs fitted graph shows that happiness scores between 3.5 and 5.0 are more linear than happiness scores between 5.0 and 6.5. By looking at the QQ plot, we see that the errors are normally distributed because most of the variables align with the regression line.

#Make a prediction

```
predict(lr_model)
```

```
##        1        2        3        4        5        6        7        8
## 6.721443 6.742926 6.891434 6.645332 6.697694 6.349630 6.537431 6.708236
##        9       10       11       12       13       14       15       16
## 6.533813 6.585953 5.807278 6.306780 6.385113 6.144065 6.638533 6.403859
##       17       18       19       20       21       22       23       24
## 6.326440 6.445796 6.337470 5.513928 6.003807 6.014713 6.144248 6.168584
##       25       26       27       28       29       30       31       32
## 5.416986 6.083471 6.503316 6.294833 5.799820 6.131437 5.983404 6.329971
##       33       34       35       36       37       38       39       40
## 5.725033 6.195914 6.013447 6.035478 5.688240 6.076384 5.729359 5.869416
##       41       42       43       44       45       46       47       48
## 5.977991 5.573885 5.684496 5.699540 5.391915 6.240967 6.811907 5.631221
##       49       50       51       52       53       54       55       56
## 5.961225 5.409871 6.184075 5.650188 4.845210 5.692101 4.857821 5.121312
##       57       58       59       60       61       62       63       64
## 5.369879 5.794376 6.122243 5.899571 5.491419 6.377973 5.522235 5.594640
##       65       66       67       68       69       70       71       72
```

```
## 5.089252 6.221995 5.720986 5.898949 5.470328 6.284143 5.724579 5.919884
##       73       74       75       76       77       78       79       80
## 5.063189 5.499338 5.195901 5.944832 4.459651 4.886770 5.432596 3.687906
##       81       82       83       84       85       86       87       88
## 5.643428 5.361916 4.982968 4.336887 5.240001 6.209032 4.756898 4.931599
##       89       90       91       92       93       94       95       96
## 5.997592 4.589969 5.103921 5.209723 4.651881 5.945102 5.388080 5.701166
##       97       98       99      100      101      102      103      104
## 5.938638 5.964064 5.403032 6.066332 5.996859 4.293340 4.917576 4.592654
##      105      106      107      108      109      110      111      112
## 5.680859 4.644211 4.807956 3.815729 4.346184 4.401101 5.389042 5.158610
##      113      114      115      116      117      118      119      120
## 4.739281 5.436410 5.260349 5.020096 4.561237 5.050218 4.558768 5.879466
##      121      122      123      124      125      126      127      128
## 4.424186 4.429025 4.837755 4.539041 3.608388 5.054394 5.345792 4.746679
##      129      130      131      132      133      134      135      136
## 5.423217 4.508618 4.613504 5.200187 5.288344 4.802360 4.769574 3.828766
##      137      138      139      140      141      142      143      144
## 4.209801 4.927711 5.315211 4.205222 3.167104 5.656191 3.544891 4.171479
##      145      146      147      148      149      150      151      152
## 3.137001 4.365146 3.309791 4.402081 4.239863 3.431625 5.156033 2.655437
##      153      154      155
## 4.950404 3.178877 2.119750
```

#RESULTS

Based on the results, this model has a good confidence level and can be used to make predictions.

#CONCLUSION

The algorithm can to predict happiness based on the level of freedom and family with a confidence level of 97.5%, making this a good model. There are six factors that contribute to the happiness score. Further analysis investigating the effects of multiple variables on the happiness score may prove to provide useful information for the researchers associated with the study. They may find that some variables influence happiness more than others.