Project 1

Aaron Chien and Dorris Wu

2022-10-01

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## Loading required package: lubridate
##
##
## Attaching package: 'lubridate'
##
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
##
##
## Loading required package: PerformanceAnalytics
##
## Loading required package: xts
##
## Loading required package: zoo
##
##
## Attaching package: 'zoo'
##
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
##
## Attaching package: 'xts'
##
##
## The following objects are masked from 'package:dplyr':
##
```

```
##     first, last
##
##
##
## Attaching package: 'PerformanceAnalytics'
##
##
## The following object is masked from 'package:graphics':
##
##     legend
##
##
## Loading required package: quantmod
##
## Loading required package: TTR
##
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
##
##
## Please cite as:
##
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
##
##
## Loading required package: lattice
##
## Loading required package: survival
##
## Loading required package: Formula
##
##
## Attaching package: 'Hmisc'
##
##
## The following object is masked from 'package:quantmod':
##
##     Lag
##
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
##
##
```

```
## Loading required package: usethis
##
## Loading required package: car
##
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
##
##
## Loading required package: sandwich
```

Introduction

Hello, in the project by Aaron Chien and Dorris Wu, we will be exploring the dataset of world happiness from Kaggle exploring different variables and their effects on happiness. Each variable has a value from 0 to 10 where 0 is the worst and 10 is the best base on public polls. The dataset can be found here: world happiness data.

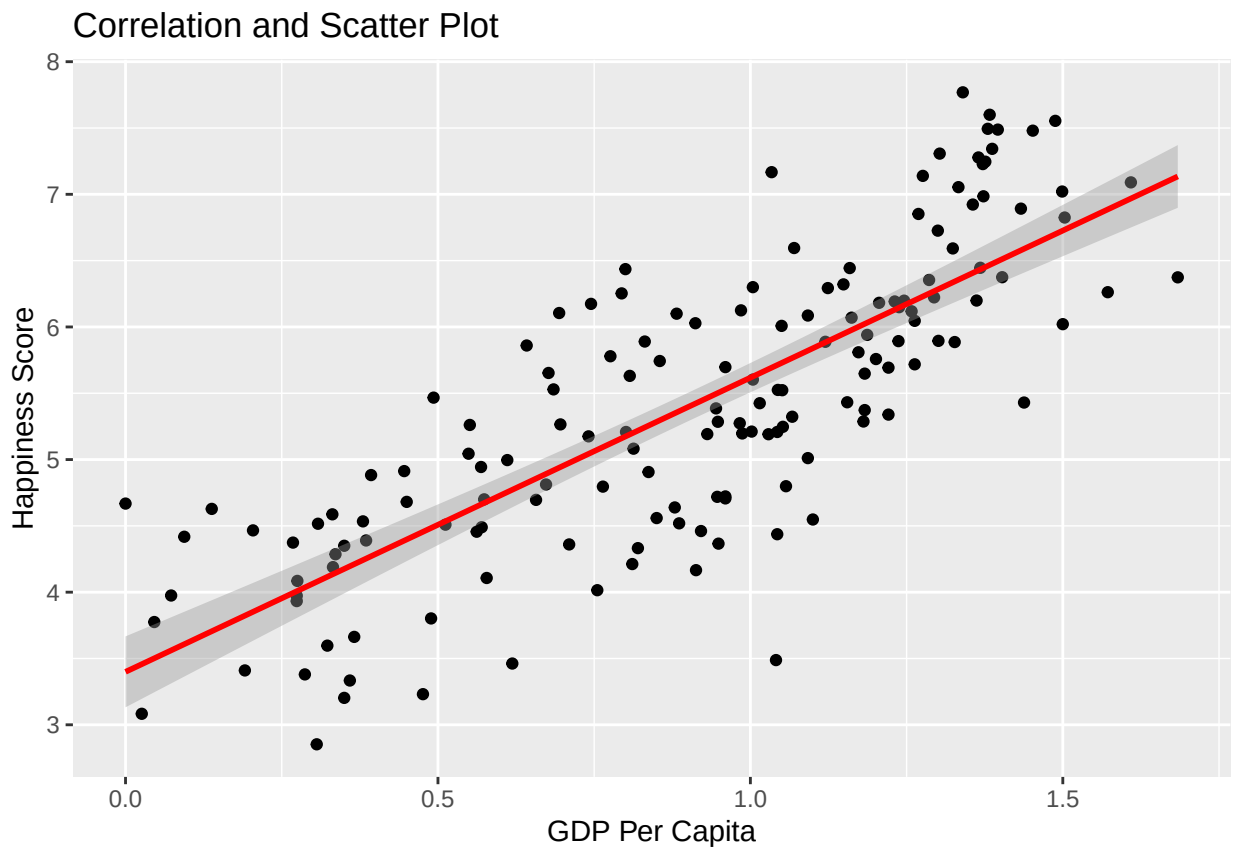
Section 1

```
world_happiness <- read.csv("world_happiness.csv")
```

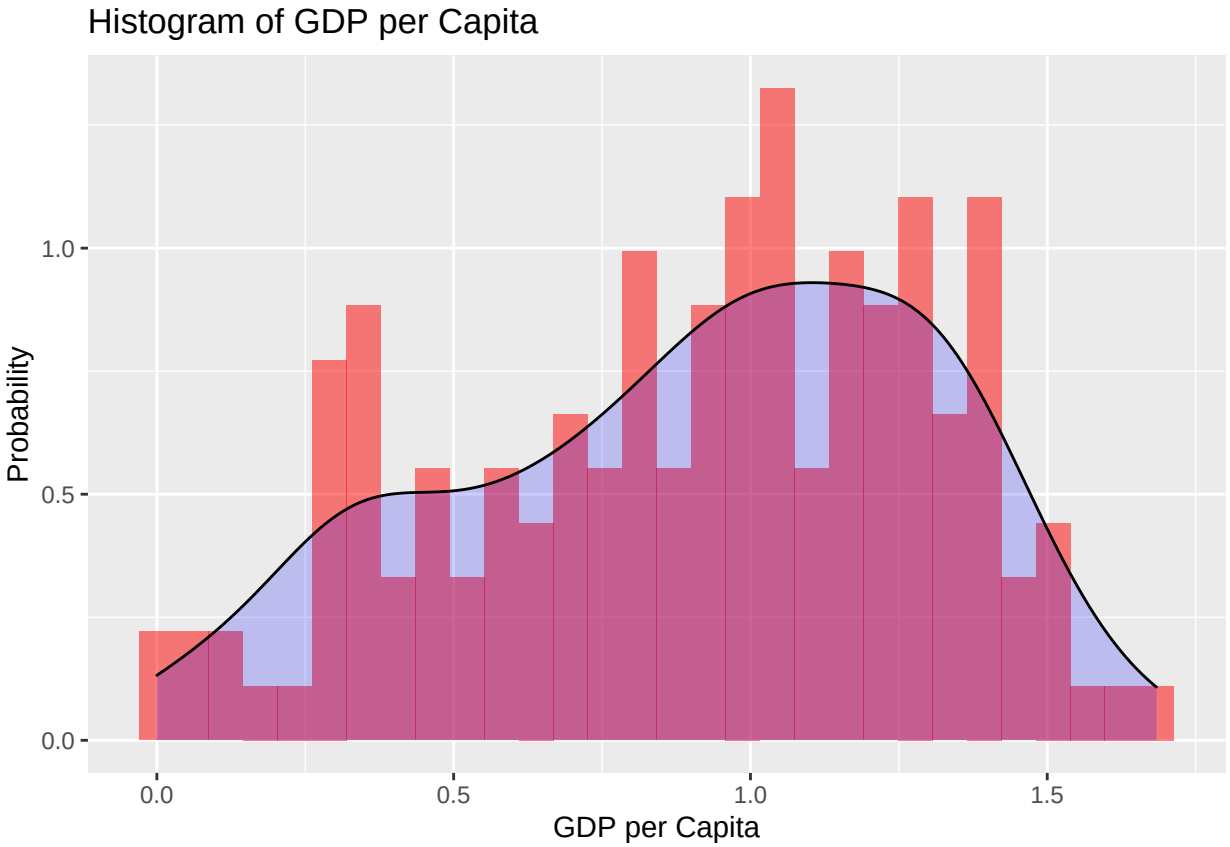
```
colnames(world_happiness) <- c("Rank", "Country", "Score", "GDP", "Support", "Health", "Freedom", "Gene
```

To understand the data better when coding, I will be renaming some of these column names: overall rank - Rank, Country or region - Country, Score - Score, GDP per capita - GDP, Social support - Support, Healthy life expectancy - Health, Freedom to make life choices - Freedom, Generosity - Generosity, Perceptions of corruption - Corruption.

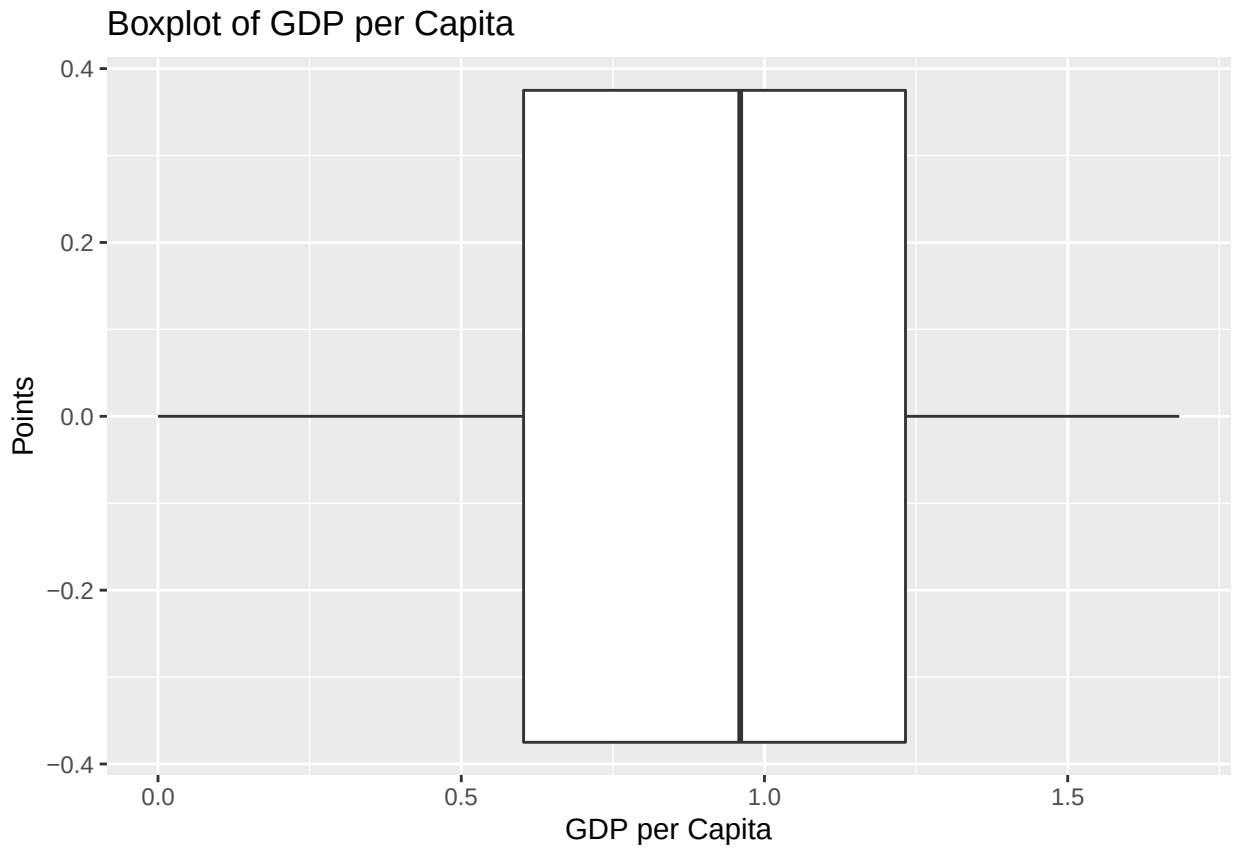
```
world_happiness %>%
  ggplot(mapping = aes(x = GDP, y = Score)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = lm, se = T, col = "red") +
  labs(title = "Correlation and Scatter Plot",
       x = "GDP Per Capita",
       y = "Happiness Score")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = GDP)) +
  geom_histogram(aes(y = ..density..), fill = "red", alpha = 0.5,
                 bins = 30) +
  geom_density(fill = "blue", alpha = 0.2) +
  labs(title = "Histogram of GDP per Capita",
       x = "GDP per Capita",
       y = "Probability")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = GDP)) +
  geom_boxplot() +
  labs(title = "Boxplot of GDP per Capita",
       x = "GDP per Capita",
       y = "Points")
```

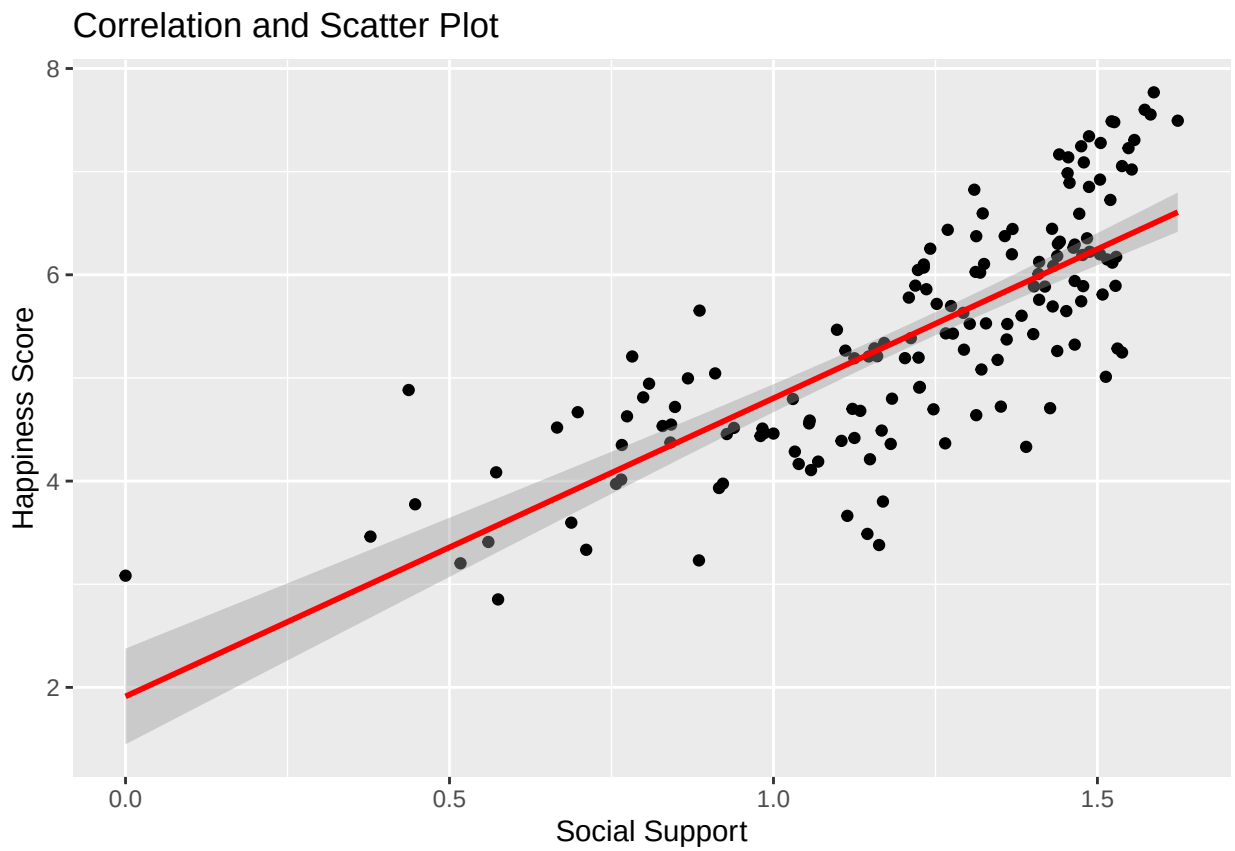



```
summary(world_happiness$GDP)
```

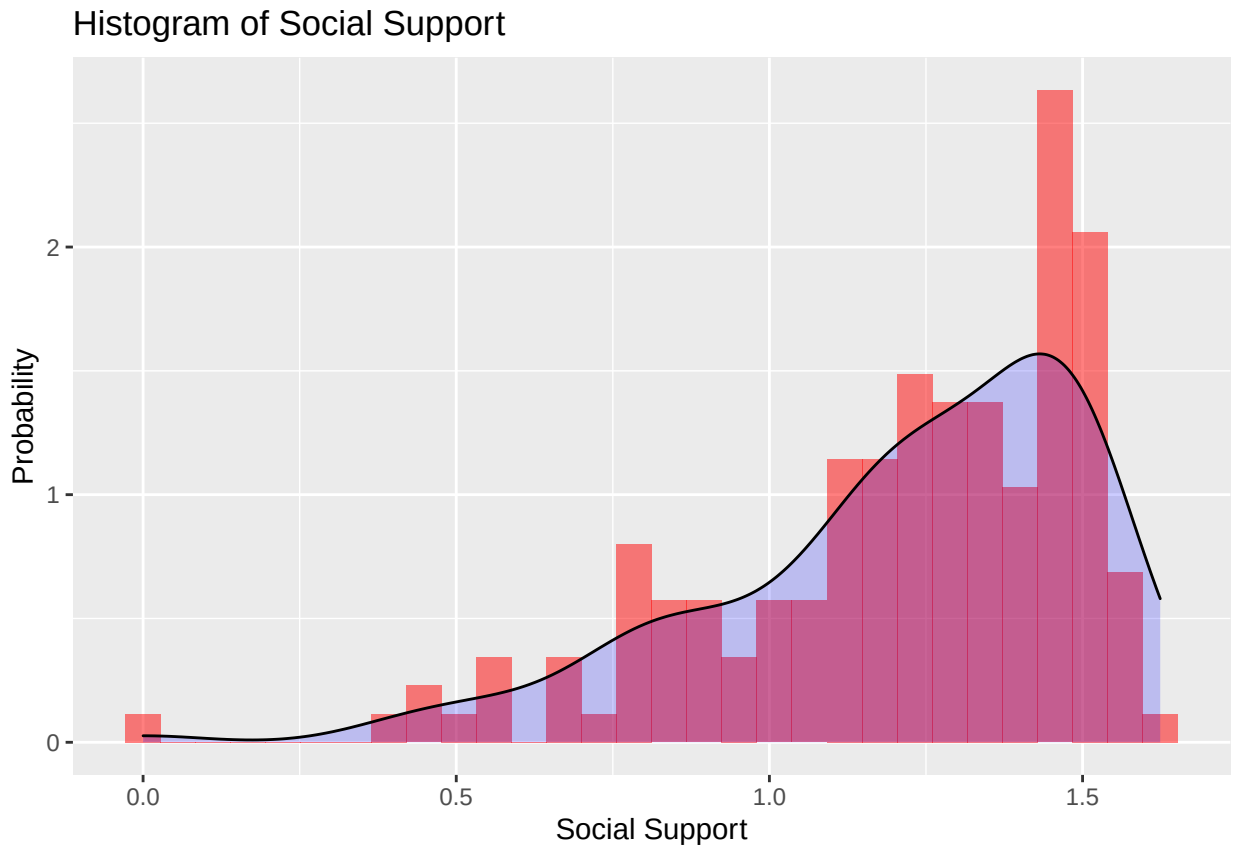
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6028  0.9600  0.9051  1.2325  1.6840
```

Examining GDP per capita first, we can see that the data is a bit left-skewed and that there are no outliers in the boxplot. The skewness is speaking out to me as we can see that most of the data does not lie with the mean. However, I do think the skewness is not that big and we can continue on. The regression line also seems to fit in well with the actual score, indicating that world happiness is correlated with GDP per capita.

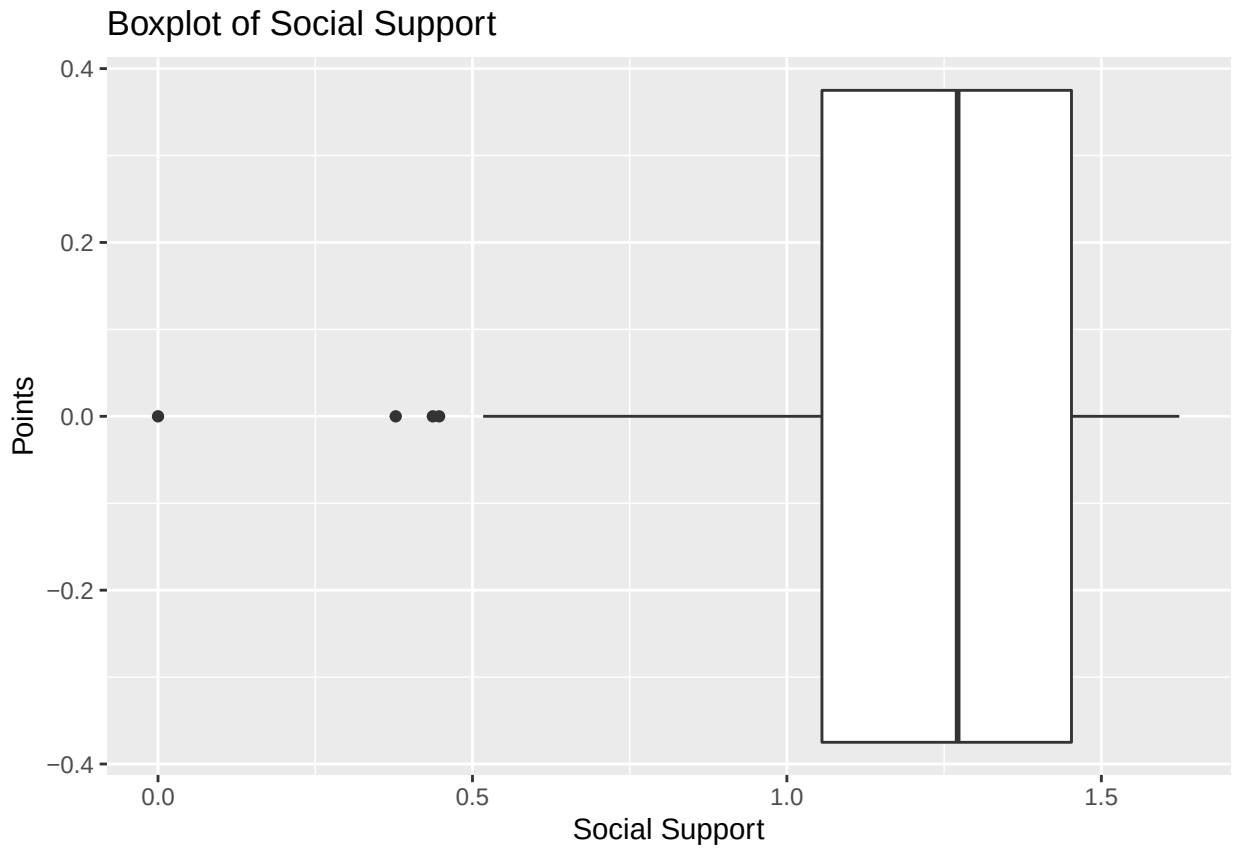
```
world_happiness %>%
  ggplot(mapping = aes(x = Support, y = Score)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = lm, se = T, col = "red") +
  labs(title = "Correlation and Scatter Plot",
       x = "Social Support",
       y = "Happiness Score")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = Support)) +
  geom_histogram(aes(y = ..density..), fill = "red", alpha = 0.5,
                 bins = 30) +
  geom_density(fill = "blue", alpha = 0.2) +
  labs(title = "Histogram of Social Support",
       x = "Social Support",
       y = "Probability")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = Support)) +
  geom_boxplot() +
  labs(title = "Boxplot of Social Support",
       x = "Social Support",
       y = "Points")
```



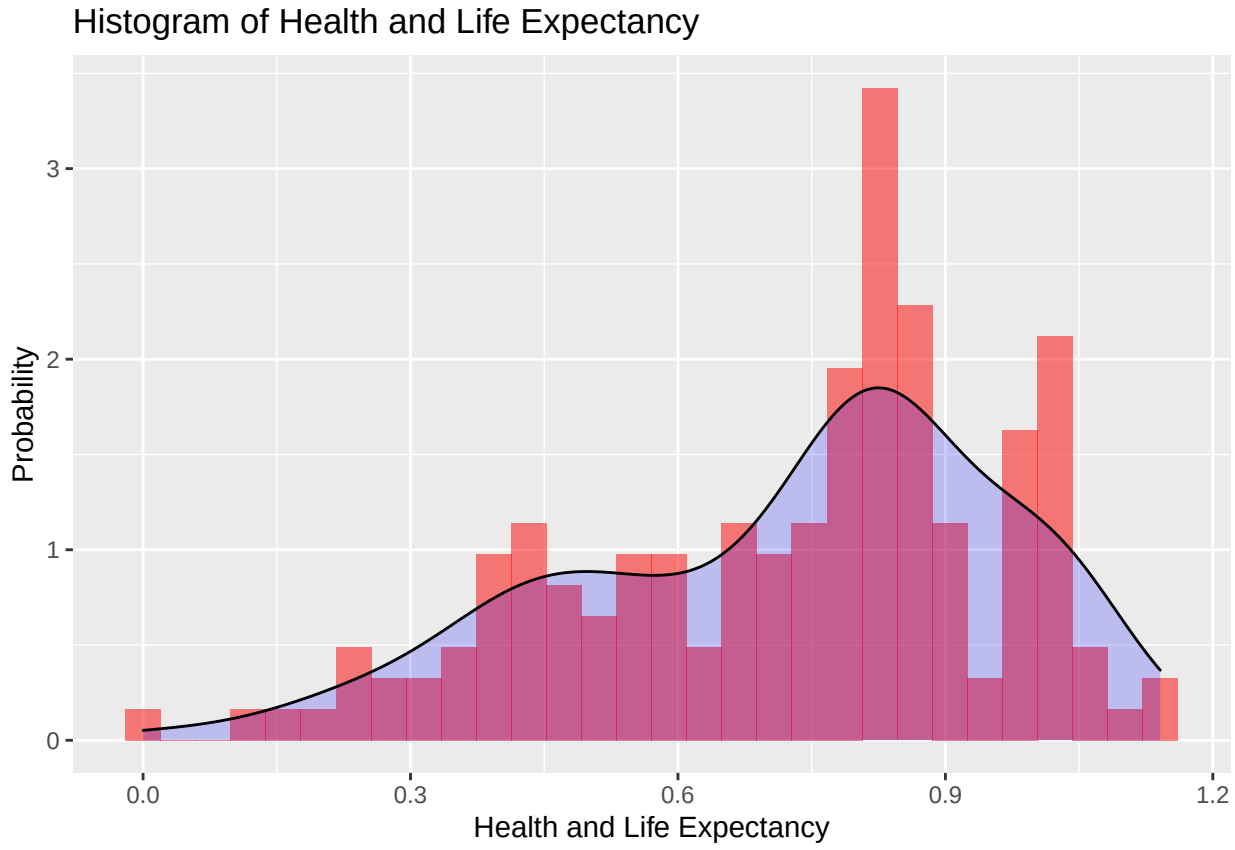

```
summary(world_happiness$Support)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.056   1.272   1.209   1.452   1.624
```

Here, we see that social support is heavily left-skewed and has many outliers, these outliers can pose a problem which is why we will be getting rid of them later on. But looking at the regression line against the score, we can see that they are a bit correlated, so social support may also be an important predictor in world happiness.

```
world_happiness %>%
  ggplot(mapping = aes(x = Health, y = Score)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = lm, se = T, col = "red") +
  labs(title = "Correlation and Scatter Plot",
       x = "Health and Life Expectancy",
       y = "Happiness Score")
```


```
world_happiness %>%
  ggplot(mapping = aes(x = Health)) +
  geom_histogram(aes(y = ..density..), fill = "red", alpha = 0.5,
                 bins = 30) +
  geom_density(fill = "blue", alpha = 0.2) +
  labs(title = "Histogram of Health and Life Expectancy",
       x = "Health and Life Expectancy",
       y = "Probability")
```



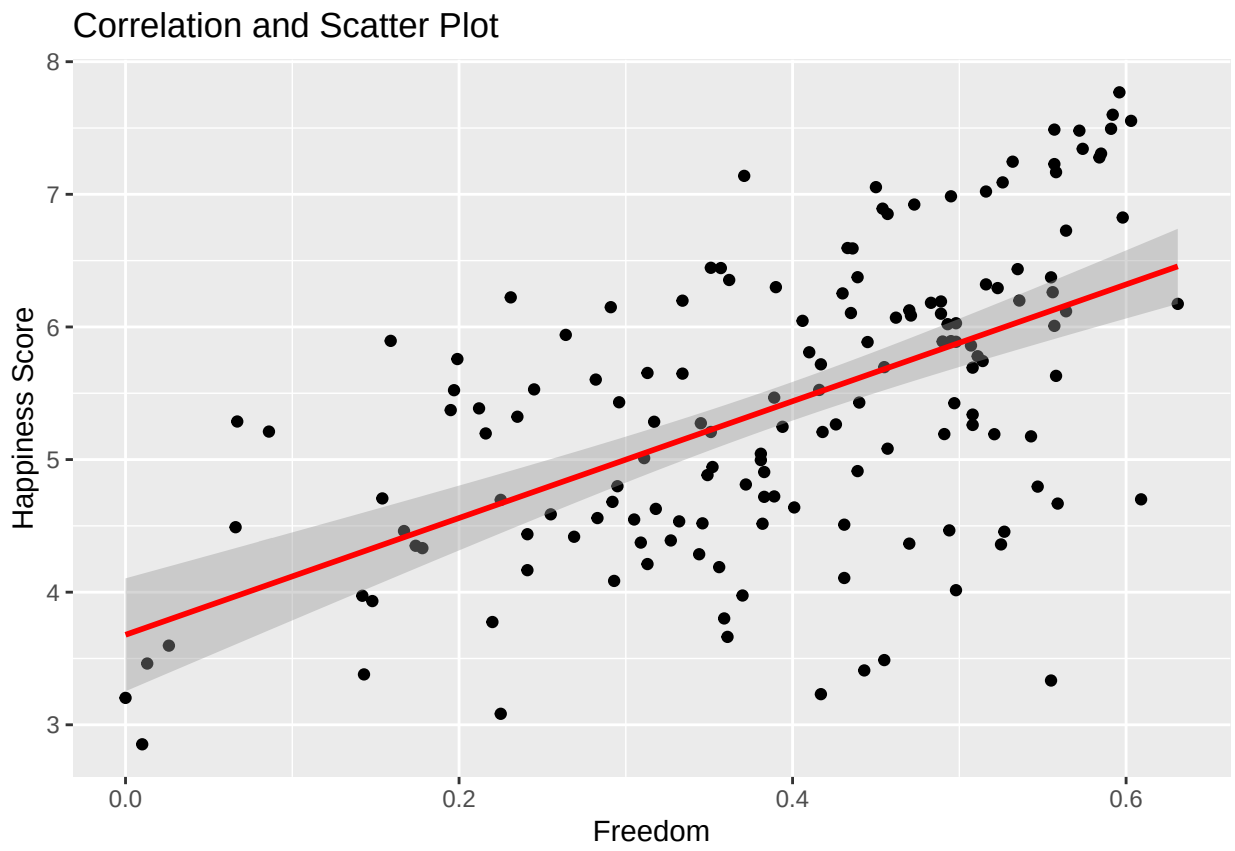

```
world_happiness %>%
  ggplot(mapping = aes(x = Health)) +
  geom_boxplot() +
  labs(title = "Boxplot of Health and Life Expectancy",
       x = "Health and Life Expectancy ",
       y = "Points")
```


```
summary(world_happiness$Health)
```

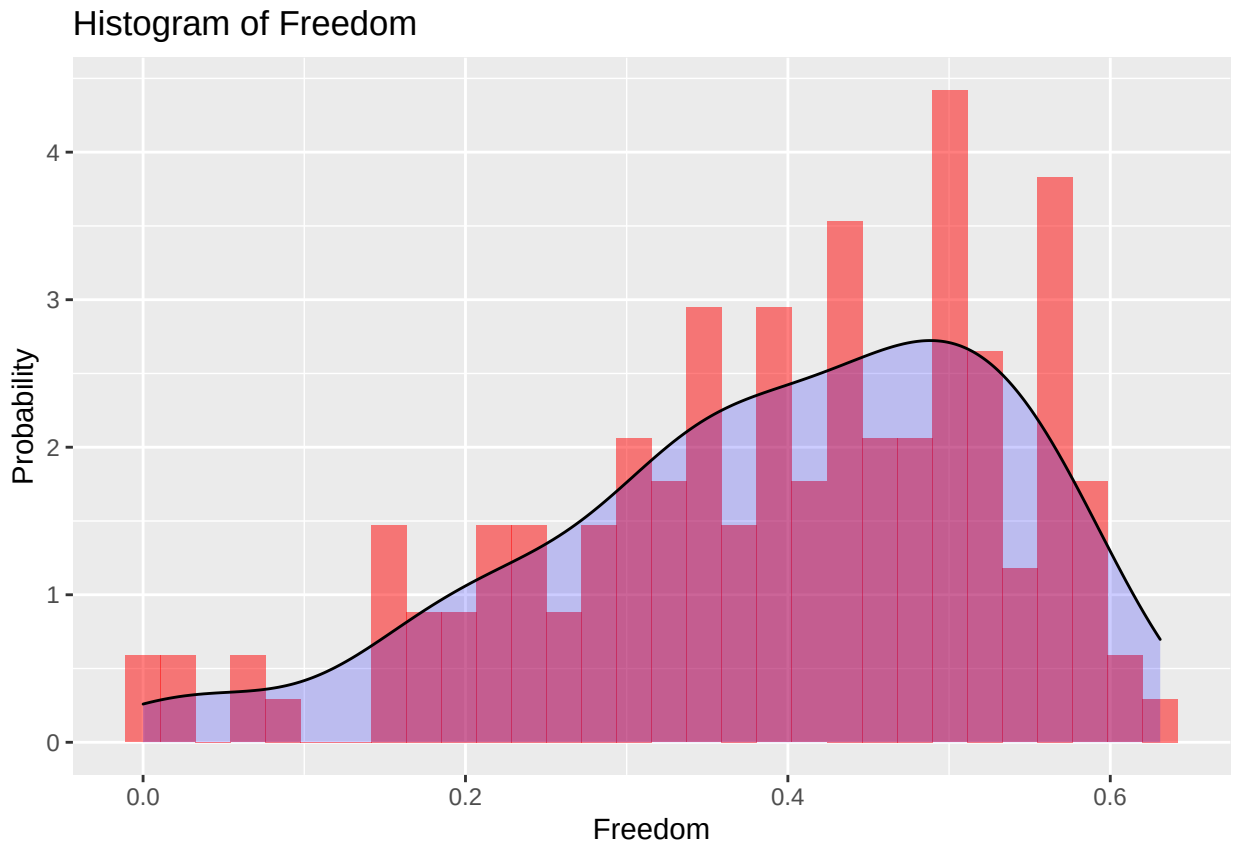
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.5477  0.7890  0.7252  0.8818  1.1410
```

Here, life expectancy has two peaks and has one outlier. This may be problematic as it may skewed the data, so we may have to use life expectancy as the cause of hetereoskedascity in the model if hetereoskedascity is present. However, the score of happiness and life expectancy do seem to be correlated so this may be another predictor that needs to be included in the model.

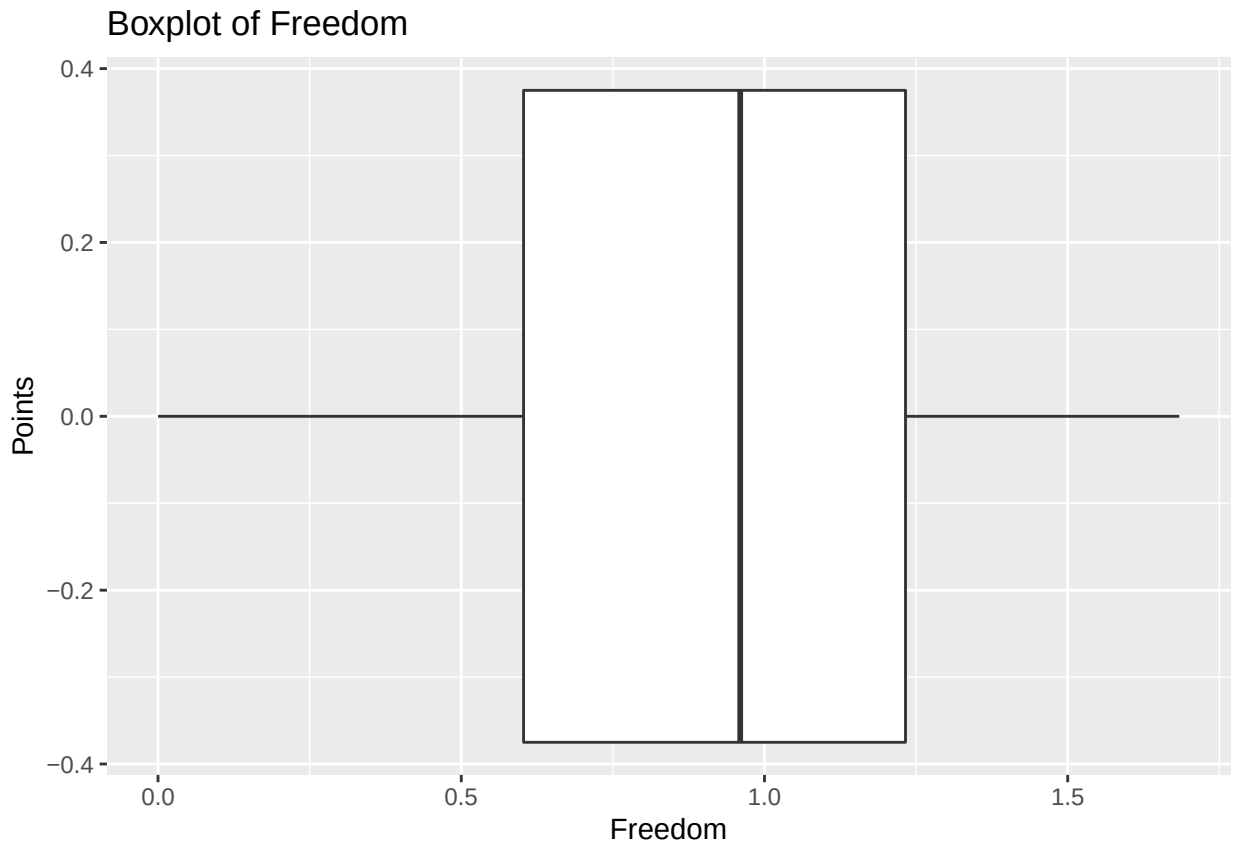
```
world_happiness %>%
  ggplot(mapping = aes(x = Freedom, y = Score)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = lm, se = T, col = "red") +
  labs(title = "Correlation and Scatter Plot",
       x = "Freedom",
       y = "Happiness Score")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = Freedom)) +
  geom_histogram(aes(y = ..density..), fill = "red", alpha = 0.5,
                 bins = 30) +
  geom_density(fill = "blue", alpha = 0.2) +
  labs(title = "Histogram of Freedom",
       x = "Freedom",
       y = "Probability")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = GDP)) +
  geom_boxplot() +
  labs(title = "Boxplot of Freedom",
       x = "Freedom",
       y = "Points")
```

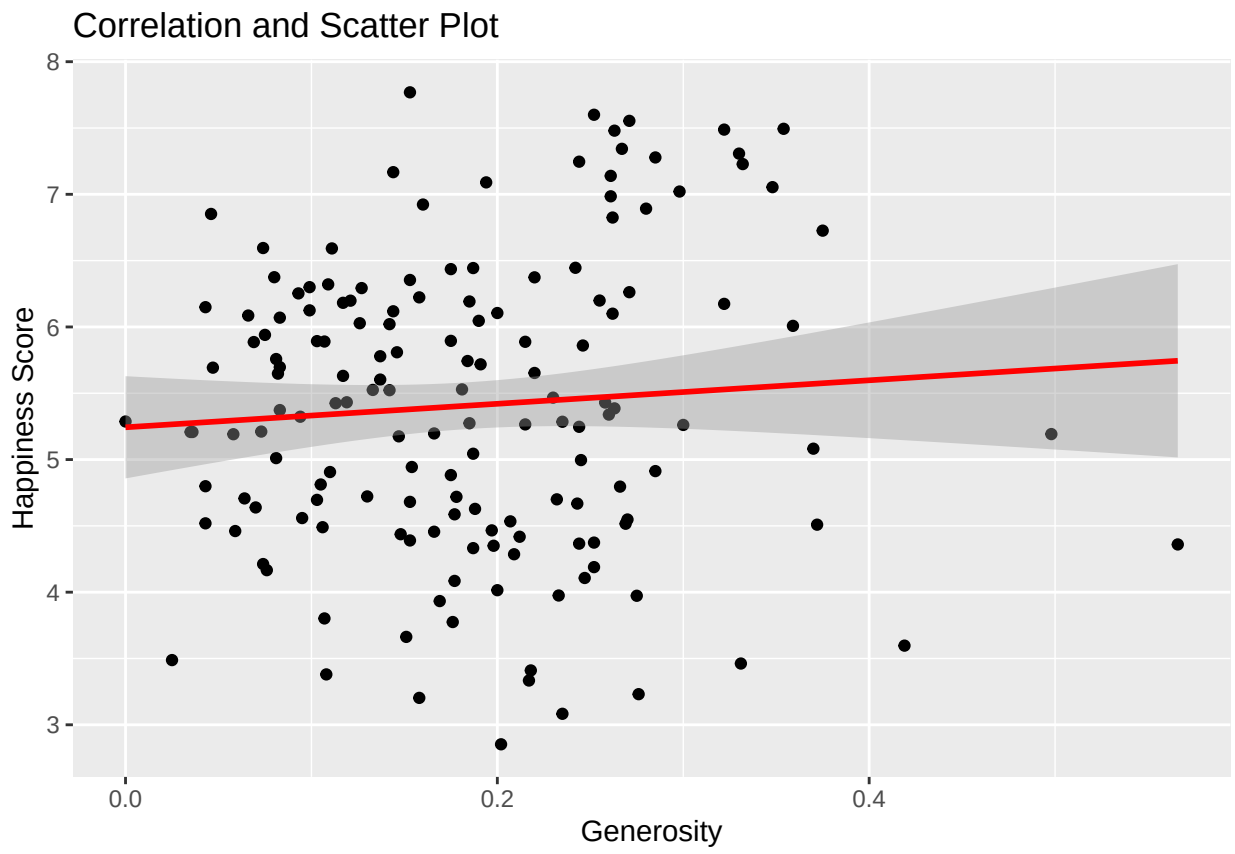



```
summary(world_happiness$Freedom)
```

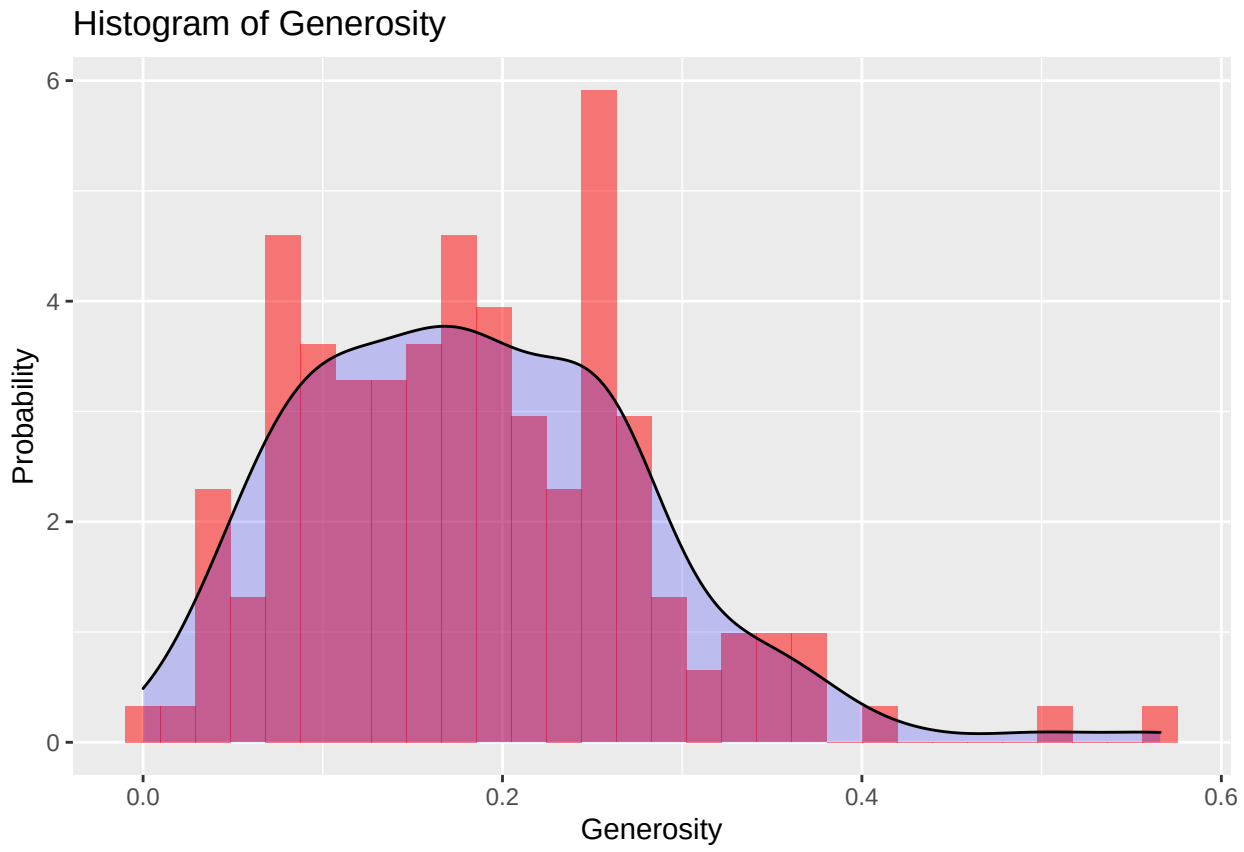
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3080  0.4170  0.3926  0.5072  0.6310
```

In the freedom variable, we can see that there is a strong correlation between happiness and the freedom to make choices. Examining further into the histogram, we see that it is skewed the to left, meaning that most of the data lies in the median rather than the mean.

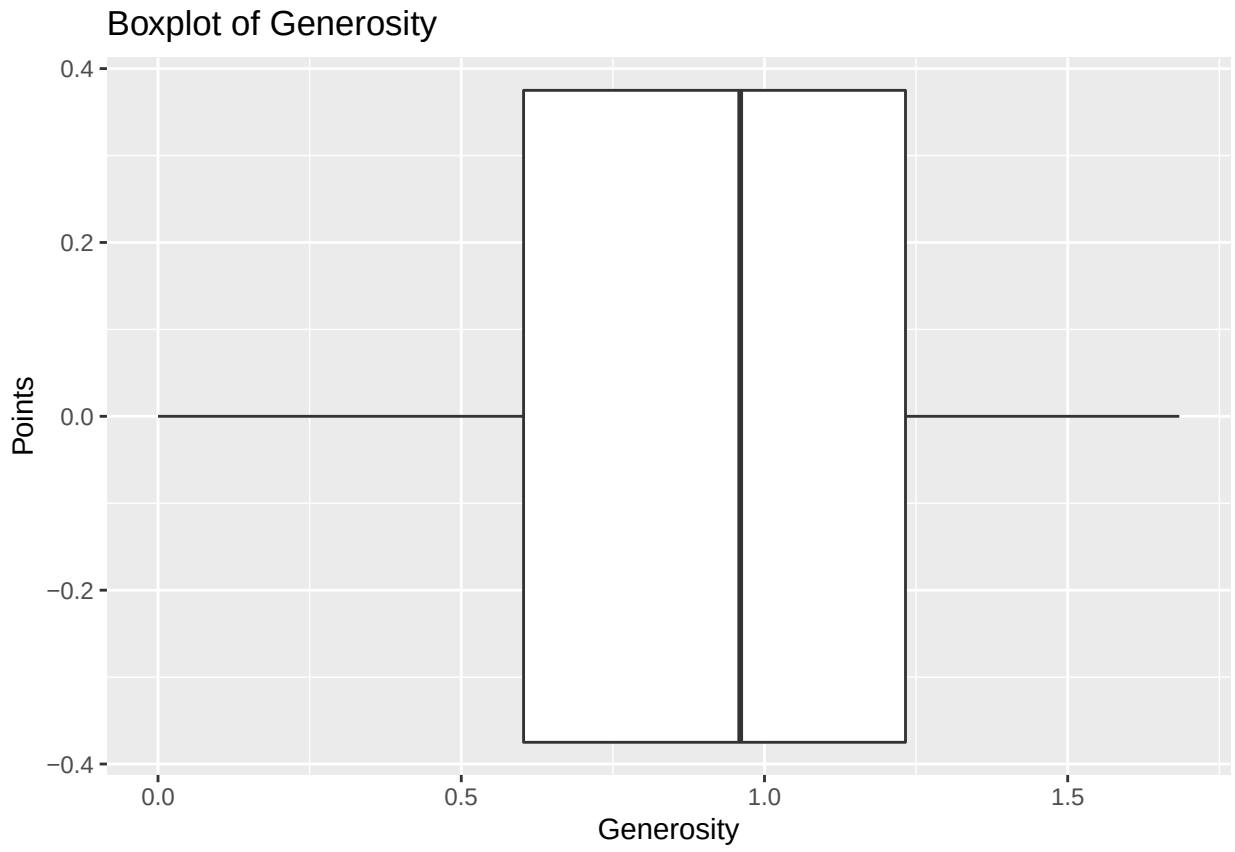
```
world_happiness %>%
  ggplot(mapping = aes(x = Generosity, y = Score)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = lm, se = T, col = "red") +
  labs(title = "Correlation and Scatter Plot",
       x = "Generosity",
       y = "Happiness Score")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = Generosity)) +
  geom_histogram(aes(y = ..density..), fill = "red", alpha = 0.5,
                 bins = 30) +
  geom_density(fill = "blue", alpha = 0.2) +
  labs(title = "Histogram of Generosity",
       x = "Generosity",
       y = "Probability")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = GDP)) +
  geom_boxplot() +
  labs(title = "Boxplot of Generosity",
       x = "Generosity",
       y = "Points")
```

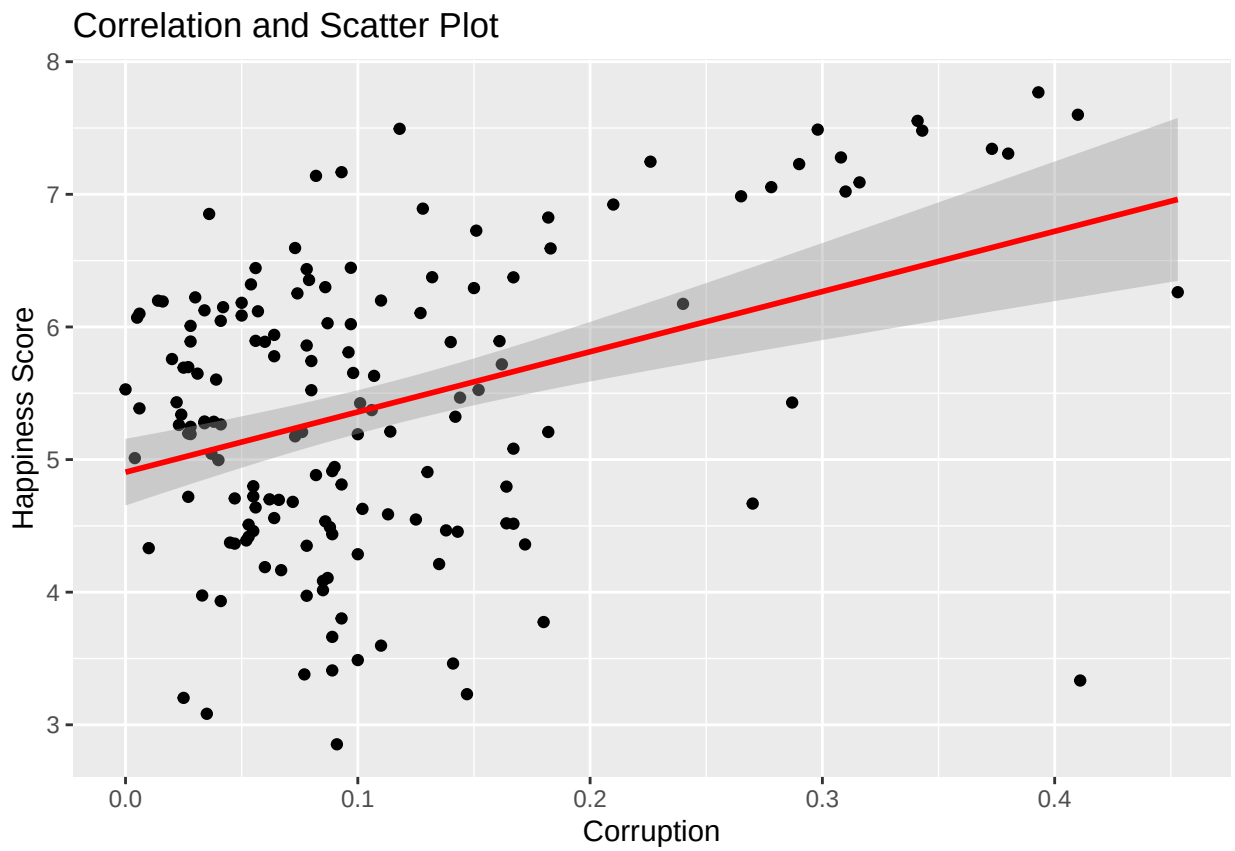



```
summary(world_happiness$Generosity)
```

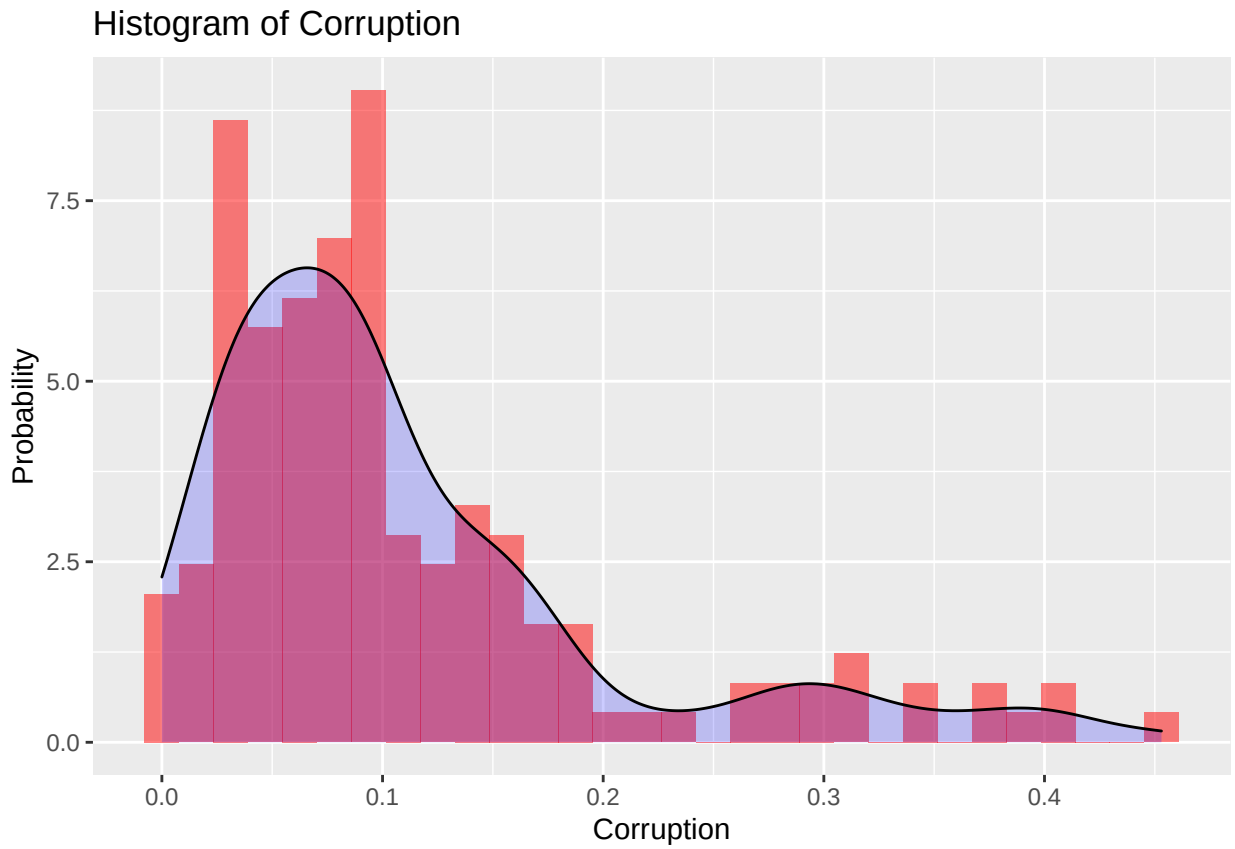
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1087  0.1775  0.1848  0.2482  0.5660
```

Examining generosity, we can see that it is heavily skewed to the right and we also have a few outliers that we may need to remove in order to gain a better model and there also seems to be more variance of the score against the regression line. The regression line is very flat, so generosity may not be the best predictor variable to use here.

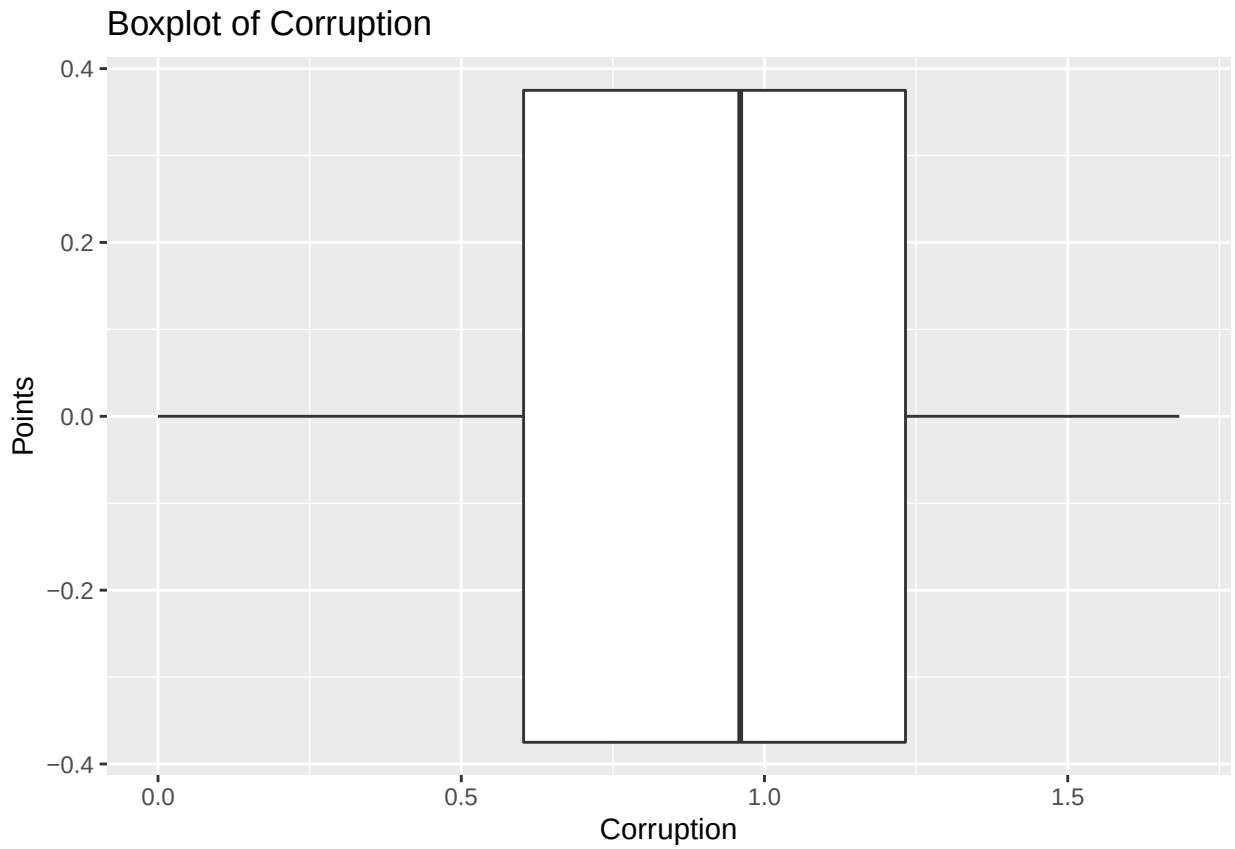
```
world_happiness %>%
  ggplot(mapping = aes(x = Corruption, y = Score)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = lm, se = T, col = "red") +
  labs(title = "Correlation and Scatter Plot",
       x = "Corruption",
       y = "Happiness Score")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = Corruption)) +
  geom_histogram(aes(y = ..density..), fill = "red", alpha = 0.5,
                 bins = 30) +
  geom_density(fill = "blue", alpha = 0.2) +
  labs(title = "Histogram of Corruption",
       x = "Corruption",
       y = "Probability")
```




```
world_happiness %>%
  ggplot(mapping = aes(x = GDP)) +
  geom_boxplot() +
  labs(title = "Boxplot of Corruption",
       x = "Corruption",
       y = "Points")
```




```
summary(world_happiness$Corruption)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0470  0.0855  0.1106  0.1412  0.4530
```

Perceptions of corruption is heavily skewed and we also see in the regression line that there is some correlation, but a lot of the data is scattered towards low levels of corruption. This variable would need to be explored more because the regression line is showing correlation, but there's a high cluster of the scores towards the levels of 0 to 0.2 of corruption.

Section 2

To start off, we create a base model to work from where Score is the dependent variable base on social support, health and life expectancy, freedom to make choices, generosity, and perceptions of world corruption.

```
wh_reg <- lm(Score ~ GDP + Support + Health + Freedom +
               Generosity + Corruption, data = world_happiness)

stargazer(wh_reg, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
```

```
##                       ----------------------------
##                                 Score
## ------------------------------------------------
## GDP                            0.775***
##                                (0.218)
##
## Support                        1.124***
##                                (0.237)
##
## Health                         1.078***
##                                (0.335)
##
## Freedom                        1.455***
##                                (0.375)
##
## Generosity                      0.490
##                                (0.498)
##
## Corruption                      0.972*
##                                (0.542)
##
## Constant                       1.795***
##                                (0.211)
##
## ------------------------------------------------
## Observations                     156
## R2                              0.779
## Adjusted R2                     0.770
## Residual Std. Error      0.534 (df = 149)
## F Statistic           87.618*** (df = 6; 149)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Our intercept, GDP, social support, and freedom estimates are the most significance as they reject the null hypothesis that they are 0 at even the 1% level of significance. The generosity estimate and the level of perceptions of corruption are the least significant. Generosity fails to reject the value that it is 0 at all levels and corruption fails to reject the value that it is 0 at the 10% level of significance.

The full equation is:

$$Score = 1.795 + 0.775 \times GDP + 1.124 \times Support + 1.078 \times Health + 1.455 \times Freedom + 0.490 \times Generosity + 0.972 \times Corruption$$

Section 3 Since the generosity and perceptions of corruption are the least significant, we can probably remove them but first we should remove the outliers to see if they are causing the insignificance. We decided to remove the outliers because the data is skewed, so removing the outliers may help reduce a bit of the skewness.

```
support <- world_happiness$Support
health <- world_happiness$Health
freedom <- world_happiness$Freedom
gen <- world_happiness$Generosity
corrupt <- world_happiness$Corruption

outliers <- c(which(support %in% boxplot.stats(support)$out),
              which(health %in% boxplot.stats(health)$out),
```

```
              which(freedom %in% boxplot.stats(freedom)$out),
              which(gen %in% boxplot.stats(gen)$out),
              which(corrupt %in% boxplot.stats(corrupt)$out))

world_happiness_cl <- world_happiness[-outliers, ]
```

Removing outliers and creating a new regression line.

```
wh_reg2 <- lm(Score ~ GDP + Support + Health + Freedom +
                              Generosity + Corruption,
                              data = world_happiness_cl)

stargazer(wh_reg2, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Score
## ------------------------------------------------
## GDP                           0.709***
##                                (0.225)
##
## Support                       1.196***
##                                (0.261)
##
## Health                        1.183***
##                                (0.357)
##
## Freedom                       1.391***
##                                (0.372)
##
## Generosity                     0.928*
##                                (0.556)
##
## Corruption                      1.135
##                                (0.832)
##
## Constant                      1.618***
##                                (0.259)
##
## ------------------------------------------------
## Observations                     134
## R2                              0.740
## Adjusted R2                     0.728
## Residual Std. Error       0.503 (df = 127)
## F Statistic           60.324*** (df = 6; 127)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

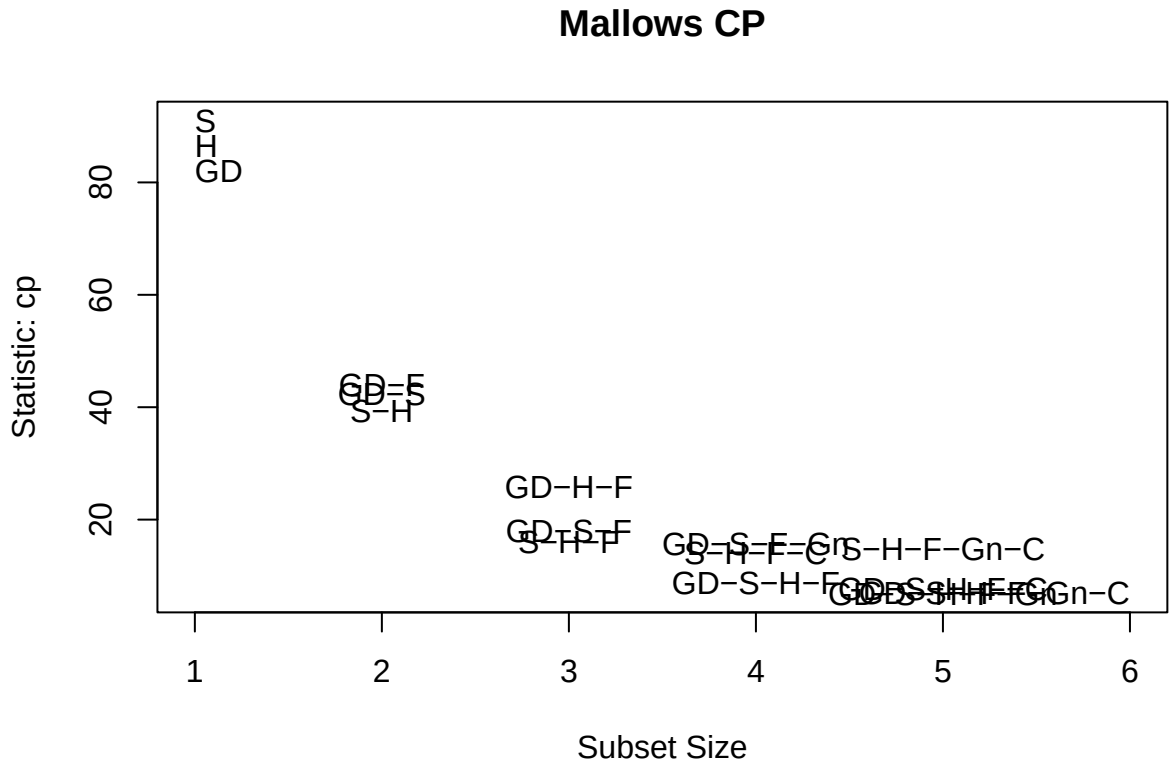
The new equation after removing the outliers is:

$Score = 1.618 + 0.709 \times GDP + 1.196 \times Support + 1.183 \times Health + 1.391 \times Freedom + 0.928 \times Generosity + 1.135 \times Corruption$

4

```
ss <- regsubsets(Score ~ GDP + Support + Health + Freedom +
                         Generosity + Corruption,
                         method = c("exhaustive"), nbest = 3,                    data = wo
```

```
subsets(ss, statistic = "cp", legend = F, main = "Mallows CP")
```




```
##            Abbreviation
## GDP                  GD
## Support               S
## Health                H
## Freedom               F
## Generosity           Gn
## Corruption            C
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
possibilites <- ols_step_all_possible(wh_reg2)
possibilites[c(42, 57, 58, 63), ]
```

```
##    Index N                                        Predictors  R-Square
## 42    42 4                       GDP Support Health Freedom 0.7284859
## 57    57 5            GDP Support Health Freedom Generosity 0.7364510
## 58    58 5            GDP Support Health Freedom Corruption 0.7345501
## 63    63 6 GDP Support Health Freedom Generosity Corruption 0.7402568
##    Adj. R-Square Mallow's Cp
## 42     0.7200668    8.755351
## 57     0.7261561    6.860846
## 58     0.7241810    7.790261
## 63     0.7279855    7.000000
```

Examining the Mallows CP graph, although difficult to point out, we can see that the lowest score is GD - S - H - F - Gn, meaning that we should be adding these variables in our model.

```
## After 10 iterations, +0.57 secs:
```

```
##  confirmed 6 attributes: Corruption, Freedom, GDP, Health, Rank and 1 more;
```

```
##  still have 2 attributes left.
```

```
##              meanImp medianImp     minImp    maxImp  normHits  decision
## Rank       32.230745 32.252553 29.5890284 34.218225 1.0000000 Confirmed
## Country     1.578995  1.591001 -1.0153214  4.026954 0.3838384 Tentative
## GDP        13.199915 13.203453 12.2070644 14.327733 1.0000000 Confirmed
## Support    15.011239 15.055189 13.6524856 16.522652 1.0000000 Confirmed
## Health     13.559989 13.526863 12.1969034 15.713888 1.0000000 Confirmed
## Freedom     6.308410  6.249723  3.6891708  8.184498 0.9898990 Confirmed
## Generosity  1.560502  1.526049 -0.6314611  3.763764 0.4141414 Tentative
## Corruption  3.469006  3.369967  1.6229292  5.787924 0.9393939 Confirmed
```

```
##      [,1]         [,2]
## [1,] "Rank"       "2"
## [2,] "Country"    "1"
## [3,] "GDP"        "2"
## [4,] "Support"    "2"
## [5,] "Health"     "2"
## [6,] "Freedom"    "2"
## [7,] "Generosity" "1"
## [8,] "Corruption" "2"
```

GDP per capita, Social support, Healthy life expectancy, Freedom to make choices, and Perception of corruption are the most important predictors with the equation from the Boruta algorithmn. Comparing the Boruta algorithmn and Mallows CP, GDP + Support + Health + Freedom + Generosity + Corruption would be best as it has a mallows CP of 7, only 0.2 off from GDP + Support + Health + Generosity. The Boruta algorithmn also says that Generosity is tenantive, so it may or may not be significant, but it may be in an interaction variable. So for now, we are able to keep our orginal regression, wh_reg2.

Section 5

After the information we have gained in section 4, we can create a new model where GDP, Social Support, Health and Life Expectancy, Freedom to make choices, and Perceptions of corruption are the predictors.

```
vif(wh_reg2)
```

```
##        GDP    Support     Health    Freedom Generosity Corruption
##   3.679138   2.300507   3.237709   1.301888   1.192665   1.177559
```

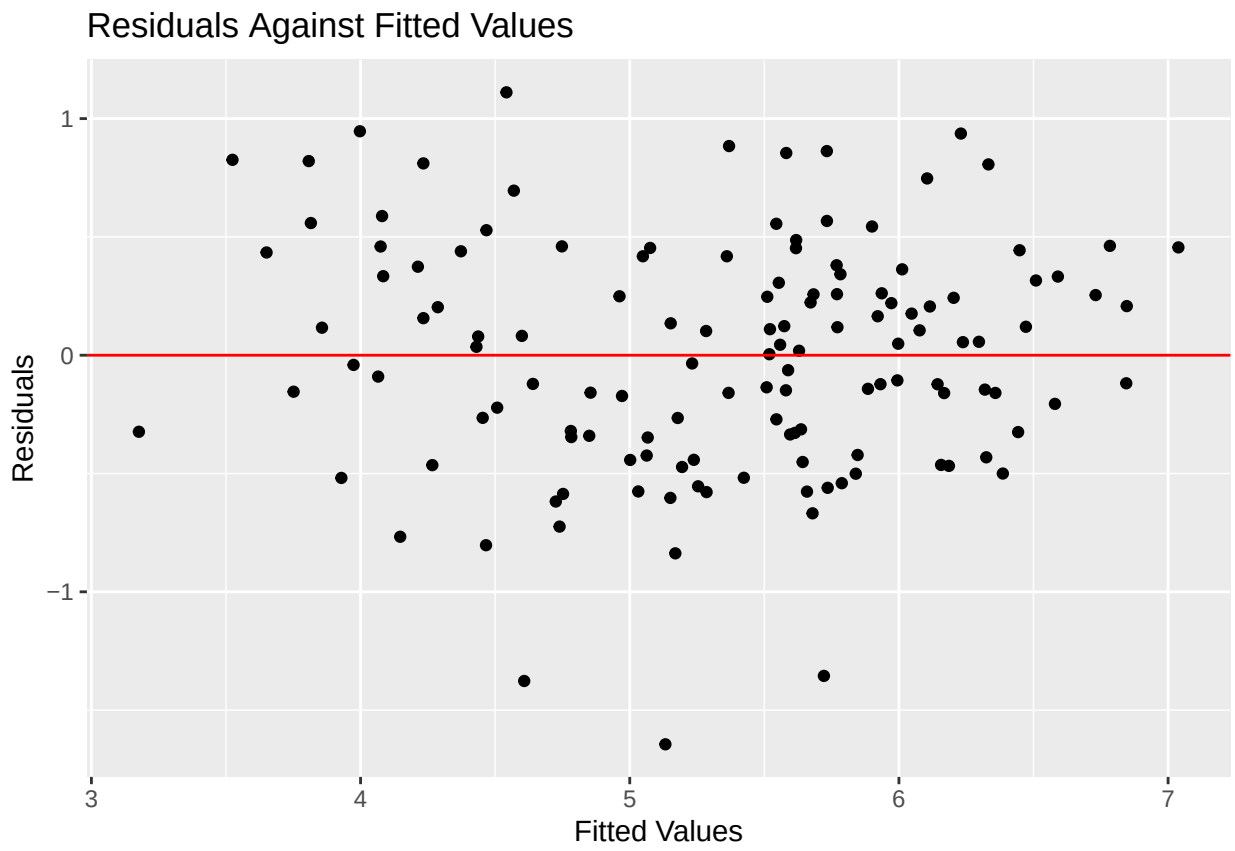
The VIF score is relatively small for all of my variables, so I am deciding to keep them. However, GDP and Health are the largest, but since they are not near a score of 5, I do not believe I need to change my equation.

Section 6

```
residuals_plot <- data_frame(Residuals = wh_reg2$residuals,
                             Fitted_values = wh_reg2$fitted.values)
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
```

```
residuals_plot %>%
  ggplot(mapping = aes(x = Fitted_values, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Fitted Values",
      y = "Residuals",
      title = "Residuals Against Fitted Values")
```



The residuals do not appear to follow a function, so I do not believe there are any forms of hetereoskedascity present, but testing shall double check this.

Section 7

```
resettest(formula = wh_reg2, power = 2, type = "fitted")
```

```
##
##  RESET test
##
## data:  wh_reg2
## RESET = 10.887, df1 = 1, df2 = 126, p-value = 0.001259
```

$Score = 1.618 + 0.709 \times GDP + 1.196 \times Support + 1.183 \times Health + 1.391 \times Freedom + 0.928 \times Generosity + 1.135 \times Corruption + \beta \times$

We reject the null that the β for Score^2 is 0, so there is at least an interaction between two variables.

```
AIC(wh_reg2)
```

```
## [1] 204.7837
```

```
BIC(wh_reg2)
```

```
## [1] 227.9664
```

When creating different models, we can compare the different models to the AIC and BIC of wh_reg2 to have a sense of which model is the best.

```
wh_reset_model <- lm(Score ~ GDP + Support  + Health + Freedom + I(Support * Health), data = world_happ

summary(wh_reset_model)
```

```
##
## Call:
## lm(formula = Score ~ GDP + Support + Health + Freedom + I(Support *
##     Health), data = world_happiness_cl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61711 -0.22222 -0.00931  0.27430  1.19022
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.8328     0.6369   7.588 5.90e-12 ***
## GDP                   0.6672     0.2051   3.254 0.001458 **
## Support              -1.5758     0.5924  -2.660 0.008811 **
## Health               -3.5287     1.0047  -3.512 0.000615 ***
## Freedom               1.6700     0.3214   5.197 7.78e-07 ***
## I(Support * Health)   4.0916     0.8209   4.984 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4685 on 128 degrees of freedom
## Multiple R-squared:  0.7726, Adjusted R-squared:  0.7637
## F-statistic: 86.98 on 5 and 128 DF,  p-value: < 2.2e-16
```

After testing many models, I concluded that this model would be the best as it had the lowest AIC and BIC score, but we still need to examine for heteroskedascity. I also removed generosity and corruption as they were proven to be insignificant in my summary of the regression model and did not improve the model at all in interactions.

Section 8

```
gqtest(wh_reset_model, fraction = 0.20, alternative = "greater",
       order.by = ~ GDP + Support + Health + Freedom +
       I(Support * Health), data = world_happiness_cl)
```

```
##
##  Goldfeld-Quandt test
##
```

```
## data:  wh_reset_model
## GQ = 0.57986, df1 = 48, df2 = 47, p-value = 0.9685
## alternative hypothesis: variance increases from segment 1 to 2
```

According to the GQ test, our p value is EXTREMELY high, meaning that we fail to reject the null so our data is relatively homoskedastic.

```
wh_reset_summ <- summary(wh_reset_model)

N <- nrow(world_happiness_cl)
resid_sq <- (wh_reset_summ$residuals)^2
world_happiness_cl$resid_sq <- resid_sq

wh_resid <- lm(resid_sq ~ GDP + Support + Health + Freedom +
                          I(Support * Health),
                          data = world_happiness_cl)
```

```
qchisq(0.95, 6 - 1)
```

```
## [1] 11.0705
```

```
N * summary(wh_resid)$r.squared
```

```
## [1] 3.934106
```

Using the BP test on the full mode, we see that the BP value is NOT greater than the chi square test, so we fail to reject the null that it is homoskedastic.

```
GDP_BP <- lm(resid_sq ~ world_happiness_cl$GDP)

Support_BP <- lm(resid_sq ~ world_happiness_cl$Support)

Health_BP <- lm(resid_sq ~ world_happiness_cl$Health)

Freedom_BP <- lm(resid_sq ~ world_happiness_cl$Freedom)

Support_Health_BP <- lm(resid_sq ~ I(world_happiness_cl$Support *
                                     world_happiness_cl$Health))
```

I decided to go with individual BP tests on each variable so I can better examine which variable is causing heteroskedascity if any exists.

```
qchisq(0.95, 2 - 1)
```

```
## [1] 3.841459
```

```
N * summary(GDP_BP)$r.squared
```

```
## [1] 0.8104497
```

```
N * summary(Support_BP)$r.squared
```

```
## [1] 2.115717
```

```
N * summary(Health_BP)$r.squared
```

```
## [1] 1.319235
```

```
N * summary(Freedom_BP)$r.squared
```

```
## [1] 0.5171469
```

```
N * summary(Support_Health_BP)$r.squared
```

```
## [1] 1.822997
```

We see that none of our variables are greater than the chisq, so our data is relatively homoskedastic.

Section 9

```
AIC(wh_reg)
```

```
## [1] 255.5295
```

```
BIC(wh_reg)
```

```
## [1] 279.9284
```

The AIC and BIC are well above 200, which is really high. But let's examine our other model.

```
AIC(wh_reg2)
```

```
## [1] 204.7837
```

```
BIC(wh_reg2)
```

```
## [1] 227.9664
```

Removing the outliers helped improve our model, but the AIC and BIC score is still quite high.

```
AIC(wh_reset_model)
```

```
## [1] 184.9556
```

```
BIC(wh_reset_model)
```

```
## [1] 205.2404
```

This model has an AIC and BIC score that is less than our original model, indicating a better fit and predictor.

Section 9

```
stargazer(wh_reset_model, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 Score
## ------------------------------------------------
## GDP                           0.667***
##                                (0.205)
##
## Support                       -1.576***
##                                (0.592)
##
## Health                        -3.529***
##                                (1.005)
##
## Freedom                       1.670***
##                                (0.321)
##
## I(Support * Health)           4.092***
##                                (0.821)
##
## Constant                      4.833***
##                                (0.637)
##
## ------------------------------------------------
## Observations                     134
## R2                              0.773
## Adjusted R2                     0.764
## Residual Std. Error       0.469 (df = 128)
## F Statistic           86.984*** (df = 5; 128)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Our final model is

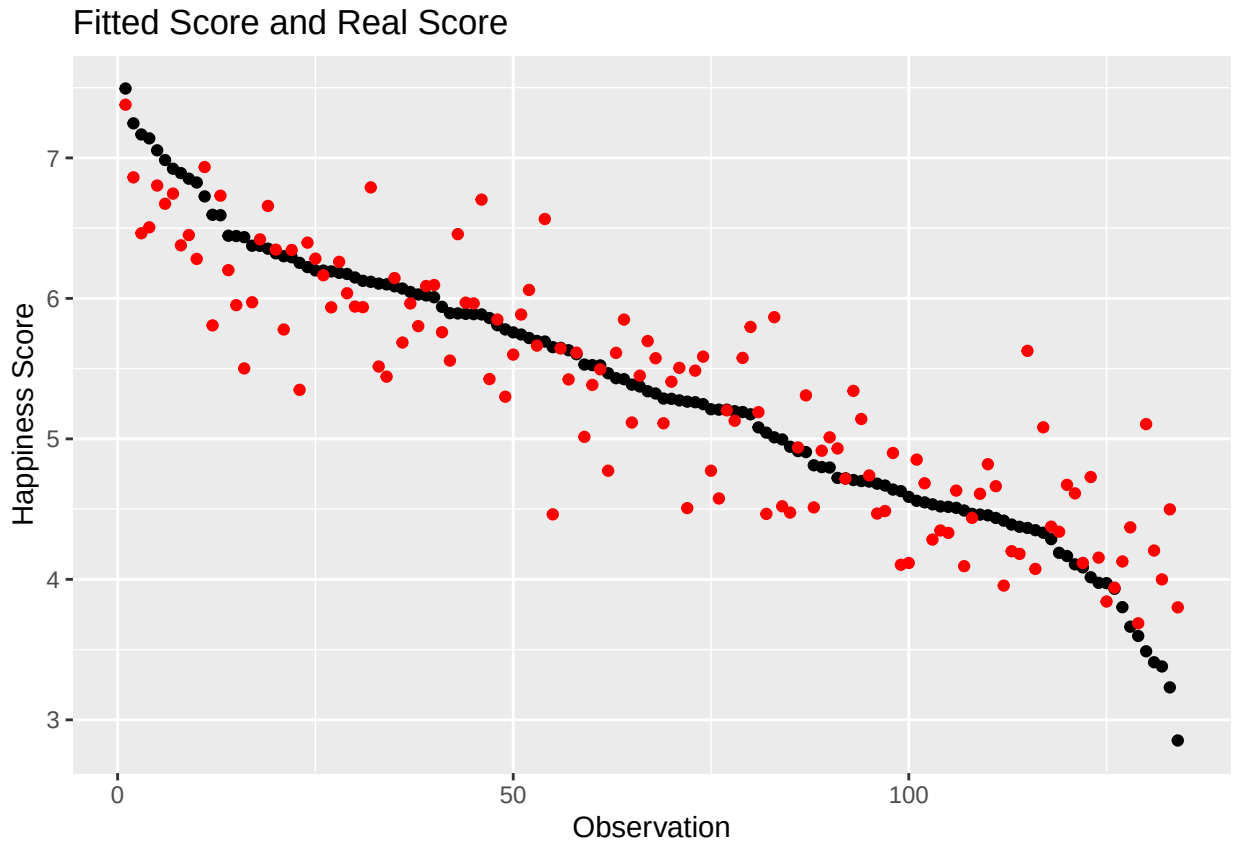
$$Score = 4.833 + 0.667 \times GDP + -1.576 \times Support + -3.529 \times Health + 1.670 \times Freedom + 4.092 \times Support \times Health$$

The F stats has a p value that is extremely small, so all the coefficients are significant enough and not 0. The R^2 is also 0.773 so 77.30% of the score is explained by the model itself, which is a majority and I

am satisfied with that. The standard errors for each coefficient is quite small, meaning that the coefficient for each variable would not differ too much. For GDP, an increase in the score increases the happiness by 0.667 and increase in the score for freedom increases happiness by 1.670. Since Support and Health have an interaction, an increase in the Support score increases happiness by $-1.576 + 4.092 \times Health$ and an increase in the Health score increases happiness by $-3.529 + 4.092 \times Support$.

```
results <- data.frame(Row = 1:N,
                      Score = (world_happiness_cl$Score),
                      Predicted = (wh_reset_model$fitted.values))

results %>%
  ggplot(aes(Row)) +
  geom_point(aes(y = Score)) +
  geom_point(aes(y = Predicted), col = "red") +
  labs(title = "Fitted Score and Real Score",
       x = "Observation",
       y = "Happiness Score") +
  scale_color_manual(breaks=c('Observation', 'Fitted'),
                     values=c('Observation'='black',
                              'Fitted'='red'))
```



As you can see, they both follow along the same trend and path, but of course there are some residuals as in the world of data science, it is very difficult to find the true parameters and true equations, but we can do our best to estimate it as close as possible.