STA3000 Final Report

Jessie Wong

Description

The World Happiness Report (WHR) measures the state of happiness of 156 countries by using the Gallup World Poll. Respondents were asked to rate their own lives on a scale from 1 to 10, with 10 being the best life and 1 being the worst life. There are 6 variables in the Happiness score: Social support, Generosity, Perception of corruption, GDP per capita, Freedom to Make Life Choices, and Healthy Life Expectancy, which are used to estimate the extent to which they contribute to happiness. For my analysis, I used the Happiness Score as the y-variable (Score) and the 6 variables as the x-variables, to examine how they individually affect happiness.

Source

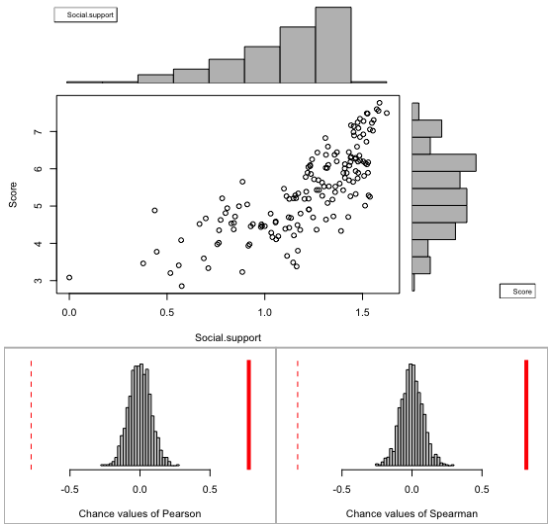
Kaggle: World Happiness Score (2019)

https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv

Association Analysis

Social support vs. Score

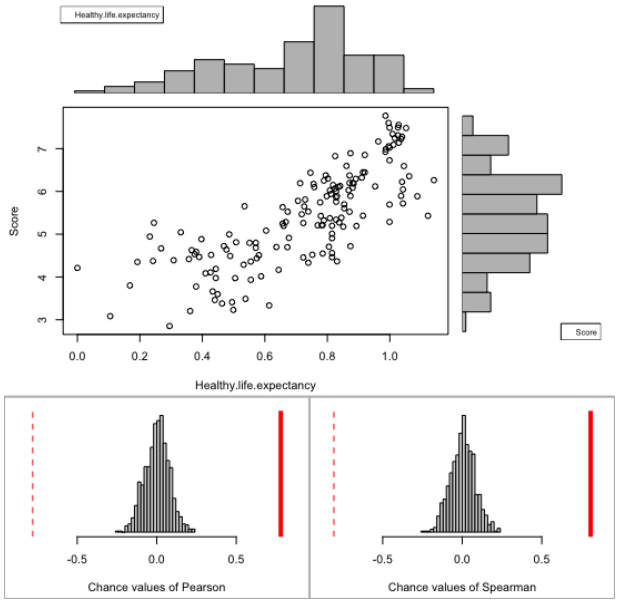
The graph shows outliers and heteroscedasticity, so I used Spearman's p-value, which is 0. The result is conclusive because 0.05 isn't between 0 and 0.004, and it's statistically significant because the p-value of <2.2e-16 is smaller than the alpha value of 0.05. Therefore, there's a statistically significant positive association between Social Support and Score.




```
Permutation procedure:
                            Value Estimated p-value
Pearson's r                 0.7770578               0
Spearman's rank correlation 0.8161807               0
With 1000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.004
 the p-value of Spearman's rank correlation is between 0 and 0.004
```

Healthy Life Expectancy vs. Score

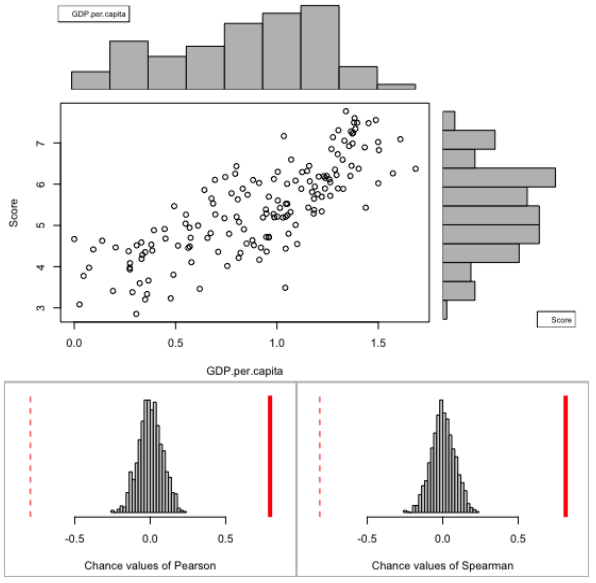
The association shows heteroscedasticity and has several outliers, so I used Spearman's p-value, which is <2.2e-16, which is less than 0.05. Therefore, the relationship is statistically significant, and there's a positive correlation between Healthy Life Expectancy and happiness. The test is conclusive because 0.05 isn't between 0 and 0.004.




```
Permutation procedure:
                                    Value Estimated p-value
Pearson's r                     0.7798831                 0
Spearman's rank correlation 0.8072746                 0
With 1000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.004
 the p-value of Spearman's rank correlation is between 0 and 0.004
```

GDP per capita vs. Score

The graph contains outliers and generally has a positive correlation. I used Spearman's p-value, which is <2e-16. Since the p-value is smaller than 0.05, the relationship between GDP per capita and happiness is statistically significant. The test is also conclusive because 0.05 isn't between 0 and 0.004.

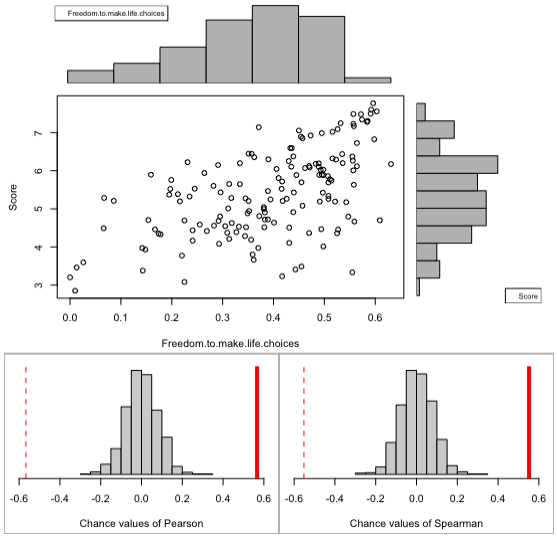



```
Permutation procedure:
                                    Value Estimated p-value
Pearson's r                     0.7938829                 0
Spearman's rank correlation 0.8144834                 0
With 1000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.004
 the p-value of Spearman's rank correlation is between 0 and 0.004
```

Freedom to Make Life Choices vs. Score

There are many outliers and strong evidence of heteroscedasticity, so I used Spearman's p-value to evaluate the association. The p-value is <2e-16 and the points are going in a positive direction.

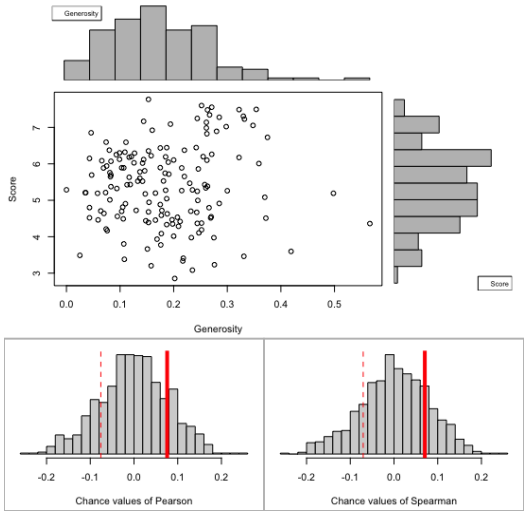
Therefore, the positive association is statistically significant because the p-value of <2e-16 is less than 0.05. The test is also conclusive because 0.05 isn't between 0 and 0.004.




```
Permutation procedure:
                                  Value Estimated p-value
Pearson's r                   0.5667418                   0
Spearman's rank correlation 0.5519742                   0
With 1000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.004
 the p-value of Spearman's rank correlation is between 0 and 0.004
```

Generosity vs. Score

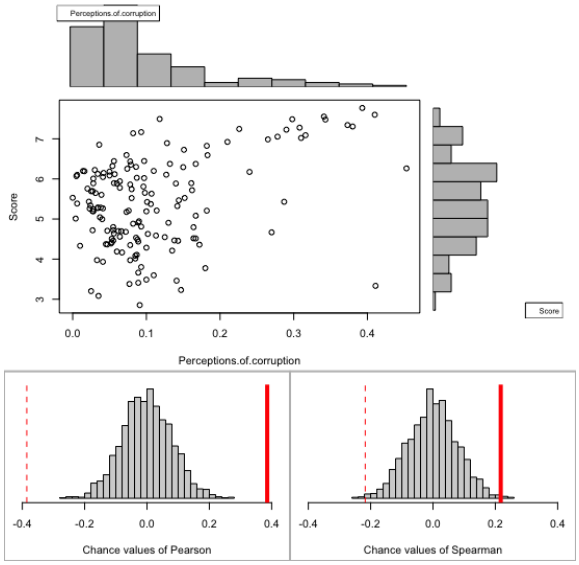
The graph shows many outliers, and it's not statistically significant because Spearman's p-value of 0.38 is greater than 0.05. Therefore, there is no relationship between the Score and Generosity. The test is conclusive because 0.05 isn't between 0.355 and 0.398.




```
Permutation procedure:
                                  Value Estimated p-value
Pearson's r                  0.07582369              0.3495
Spearman's rank correlation 0.07048260              0.3765
With 2000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0.329 and 0.371
 the p-value of Spearman's rank correlation is between 0.355 and 0.398
```

Perceptions of Corruption vs. Score

For this relationship, I used Spearman's p-value because there are extreme outliers. The association is statistically significant because Spearman's p-value is 0.0064, which is smaller than 0.05, and the association is conclusive because 0.05 isn't between 0.004 and 0.009.

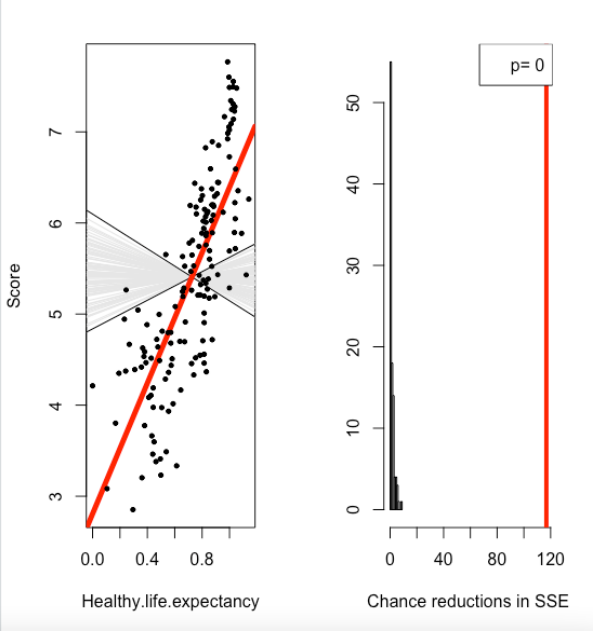



```
Permutation procedure:
                                    Value Estimated p-value
Pearson's r                     0.3856131            0.0000
Spearman's rank correlation 0.2173484            0.0064
With 5000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.001
 the p-value of Spearman's rank correlation is between 0.004 and 0.009
```

Simple Linear Regression

Healthy Life Expectancy vs. Score

The p-value for this relationship is <2e-16, so this relationship is statistically significant because it's less than 0.05, and it's unlikely that I'll observe the relationship due to chance. The residuals are fairly symmetrical, so the model fits quite well with the data, meaning that the points fall close to the actual points. For the coefficient estimate's slope, for every increase in Healthy Life Expectancy, happiness increases by 3.6. For the coefficient estimate's intercept, the average value of happiness is 2.8 in Healthy Life Expectancy. For the coefficient standard error, the average amount that the coefficient estimate varies from the average value of the Score is by 0.23, which is a small error. The standard error (RMSE) in coefficients is 0.7 and because it's a low number, the error made by the regression model is small. The t-value is about 15 standard deviations from the coefficient estimate away from 0, so it's safe to reject the null hypothesis. For Multiple R-squared, it's 0.6 in this case, and because it's closer to 1 than 0, it explains variance well: 60% of variance found in the Score can be explained by Healthy Life Expectancy. For the F-statistic, the value is 239, so there's a strong relationship between the x and y-variable. The regression line is steep and the SSE reduction is large, and because neither one is due to chance, the regression is statistically significant.



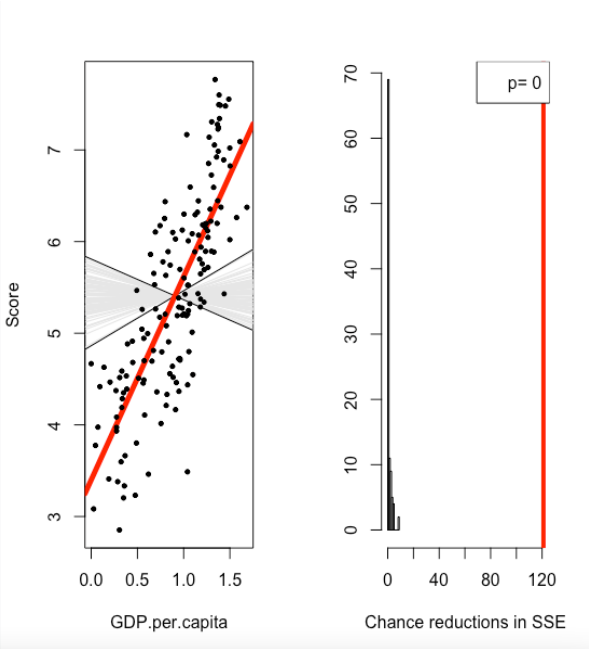

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.6743 -0.4621  0.0863  0.4864  1.5797

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.8068     0.1772   15.84   <2e-16 ***
Healthy.life.expectancy     3.5854     0.2319   15.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.699 on 154 degrees of freedom
Multiple R-squared:  0.6082,    Adjusted R-squared:  0.6057
F-statistic: 239.1 on 1 and 154 DF,  p-value: < 2.2e-16
```

GDP per capita vs. Score

The p-value for this relationship is <2.2e-16, so this relationship is statistically significant because it’s less than 0.05, and it's unlikely that I'll observe a relationship due to chance. The residuals across the 5 summary points have a pretty symmetrical distribution on the mean value of zero so the points fall close to the actual points. For the coefficient estimate slope, for every increase in GPD per capita, happiness increases by 2.2 points. For the coefficient estimate intercept, the value is 3.4, so that’s the average Score in GDP per capita across the countries. For standard error, which is 0.14, it's a small number so the coefficient estimates don't vary much from the actual average value of the Score. For the t-value, the t-value of 16 isn't near zero, so this indicates that a relationship between GDP per capita and Score exists. For Residual standard error, the error value is 0.68, which is a small error made by the regression model when predicting Score. For Multiple R-squared, the value is 0.63, so 63% of variance in Score can be explained by GDP per capita. Therefore, the regression explains the variance in Score well. For the F-statistic, the value is 262.5, and because it's far from 1, there is a strong relationship between GDP per capita and Score. For the charts, the regression line is steep and the SSE reduction is large, so neither of them is due to chance, which makes the relationship statistically significant.



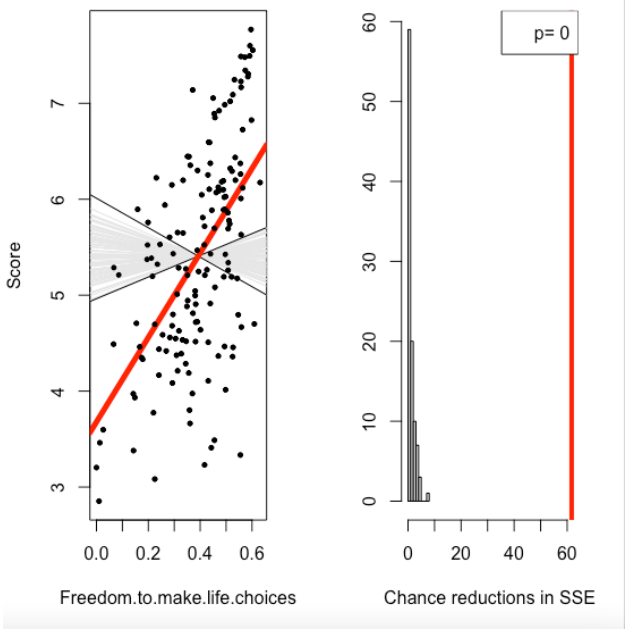

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.22044 -0.48361  0.00828  0.48433  1.47409

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.3993     0.1353   25.12   <2e-16 ***
GDP.per.capita 2.2181     0.1369   16.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.679 on 154 degrees of freedom
Multiple R-squared:  0.6303,    Adjusted R-squared:  0.6278
F-statistic: 262.5 on 1 and 154 DF,  p-value: < 2.2e-16
```

Freedom to Make Life Choices vs. Score

The p-value is 1.238e-14 which is less than 0.05, so the relationship is statistically significant because it’s less than 0.05, and it’s unlikely that I’ll observe a relationship by chance. The residuals have somewhat of a symmetrical distribution on the mean value of zero, so the model predict points that don't fall far from the actual points. For the coefficient estimate’s slope, for every increase in Freedom to Make Life Choices, the Score goes up by 4.4 points. The coefficient estimate’s intercept is 3.7, which is the average score of happiness for Freedom to Make Choices. For the coefficient standard error, the value is 0.5, which means that the Score varies by 0.5 from the actual average value of Score. For the t-value, it’s around 8.5, which tells me that a relationship exists because it's not very close to zero. For Residual standard error, the actual value deviates from the true regression line by about 0.92, and 0.92 is the typical size of the error made by the regression model, which is a small error. For Multiple R-squared, the value is 0.32, which means that 32% of the variance found in the Score can be explained by Freedom to Make Life Choices. Therefore, the model doesn’t quite fit the actual data well. For the F-statistic, it's 72.9, and because it's far from 1, it’s a good measure of the relationship between the Score and Freedom to Make Life Choices. For the regression model, the regression line is steep and the SSE reduction is large, and neither of them is due to chance. Therefore, the regression is statistically significant.



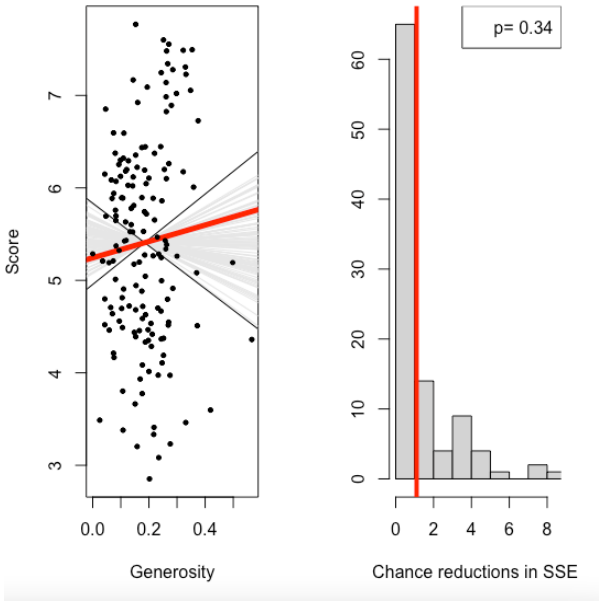

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7882 -0.5838  0.0149  0.7029  1.8269

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.6788     0.2155  17.075  < 2e-16 ***
Freedom.to.make.life.choices 4.4026     0.5158   8.536 1.24e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9201 on 154 degrees of freedom
Multiple R-squared:  0.3212,    Adjusted R-squared:  0.3168
F-statistic: 72.87 on 1 and 154 DF,  p-value: 1.238e-14
```

Generosity vs. Score

The p-value is 0.35, so this relationship isn't statistically significant because 0.35 is greater than 0.05. For the residuals, the 5 summary points appear to be nearly symmetrical, meaning that the model calculates points that are quite close to the actual points. For the coefficient estimate intercept, 5.2 is the average value of happiness in Generosity. The slope, which is 0.89, means that for every increase in Generosity, the Score goes up by 0.89 points. For coefficient standard error, the average amount that the coefficient estimates vary is about 0.94 from the actual average of the Score, which is a small error. For the t-value, the value of 0.9 means there's a possibility that we can't reject the null hypothesis and a relationship doesn't exist, because 0.9 is really close to zero. For the Residual standard error, the typical size of the error made by the regression model is 1.11, so it’s a larger error than the other regression models. For Multiple R-squared, the value is 0.005749 so the model represents a regression that doesn't explain the variance in Score well, because only 0.5% of variance found in Score can be explained by Generosity. For the f-statistic, the value is 0.8905, which is close to zero, so there is no relationship between Generosity and Score. For the charts, the regression line has a very weak positive relationship and the SSE reduction is small, so the regression isn’t statistically significant.



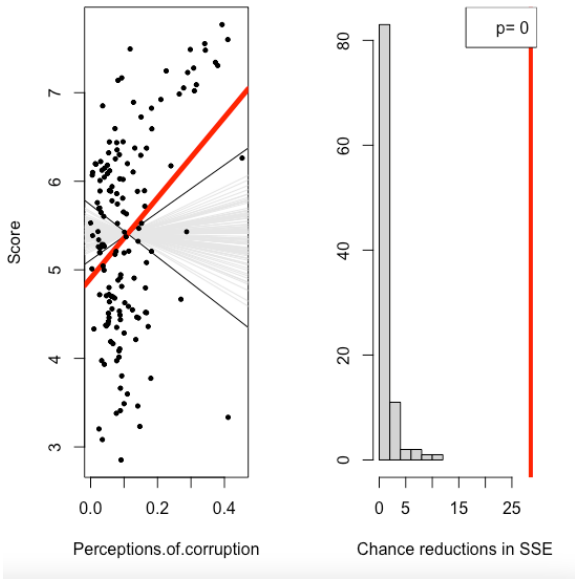

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.56930 -0.81851  0.00815  0.78707  2.39012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.2433     0.1951  26.872   <2e-16 ***
Generosity    0.8861     0.9390   0.944    0.347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.114 on 154 degrees of freedom
Multiple R-squared:  0.005749,  Adjusted R-squared:  -0.0007069
F-statistic: 0.8905 on 1 and 154 DF,  p-value: 0.3468
```

Perceptions of Corruption vs. Score

The p-value is 6.65e-07, so the relationship is statistically significant because it’s less than 0.05. The residuals are somewhat symmetrical across the 5 summary points, so the model predicts points that fall closer to the real observed points. For the coefficient estimate’s slope, for every increase in Perceptions of Corruption, happiness increases by 4.5. For the coefficient estimate’s intercept of 4.9, it means that the average Score in Perceptions of Corruption is 4.9. For the coefficient standard error, the average amount that the coefficient estimates vary from the actual average value of the Score is about 0.88. The t- values 5 and 38 aren’t very close to zero, so we can reject the null hypothesis. For the Residual standard error, the average amount that the Score deviates from the true regression line is about 1.03. In other words, the regression model makes an error of about 1.03, which isn’t a large error. For the Multiple R-squared, the value is 0.15. Therefore, about 15% of the variance found in the Score can be explained by Perceptions of Corruption, which means that the regression model doesn’t fit the actual data well. For the f-statistic, the value is 30, and because it's far from 1, there's evidence of a relationship between Perceptions of Corruption and the Score. As for the regression model, the regression line is steep and the SSE reduction is large, so neither of them is due to chance. Therefore, the regression is statistically significant.



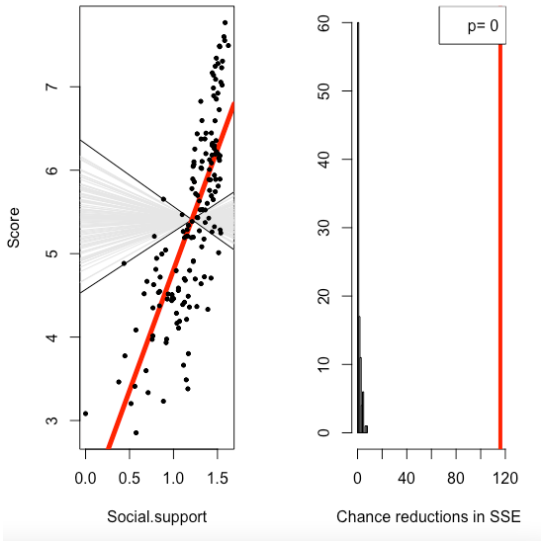

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4370 -0.7428  0.1767  0.8566  2.0533

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.9049     0.1272  38.560  < 2e-16 ***
Perceptions.of.corruption    4.5403     0.8754   5.186 6.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.03 on 154 degrees of freedom
Multiple R-squared:  0.1487,    Adjusted R-squared:  0.1432
F-statistic:  26.9 on 1 and 154 DF,  p-value: 6.654e-07
```

Social Support vs. Score

The p-value is <2.2e-16 so the relationship is statistically significant because it’s less than 0.05. The residuals are quite symmetrical across the 5 summary points, meaning that the regression model predicts points that fall close to the actual observed points. For the coefficient estimate slope, for every increase in Social Support, the happiness score increases by about 2.9. For the coefficient estimate intercept, the average happiness score in Social Support is 1.9. For standard error, the average amount that the coefficient estimates vary from the average value of the Score is 0.19, so the error is small. For the t-value, the value is 15.32, and because it's not close to 0, I can reject the null hypothesis. For the Residual standard error, it's 0.7, meaning that the regression model makes an error of 0.7 when predicting the Score, which is a small error. For Multiple R- squared, the value is 0.6. Therefore, about 60% of the variance found in the Score can be explained by Social Support, so the model fits the actual data well. For the F-statistic, the value is 234.7, and because it's very far from 1, there's a strong relationship between Social support and Score. For the charts, the regression line is really steep and the SSE reduction is large, so neither one of them is due to chance. Therefore, the regression is statistically significant.




```
Residuals:
     Min       1Q   Median       3Q      Max
-1.89465 -0.45762 -0.01993  0.54720  1.70721

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.9124     0.2349    8.14 1.25e-13 ***
Social.support   2.8910     0.1887   15.32  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7029 on 154 degrees of freedom
Multiple R-squared:  0.6038,    Adjusted R-squared:  0.6012
F-statistic: 234.7 on 1 and 154 DF,  p-value: < 2.2e-16
```