World Happiness Analysis

Introduction:

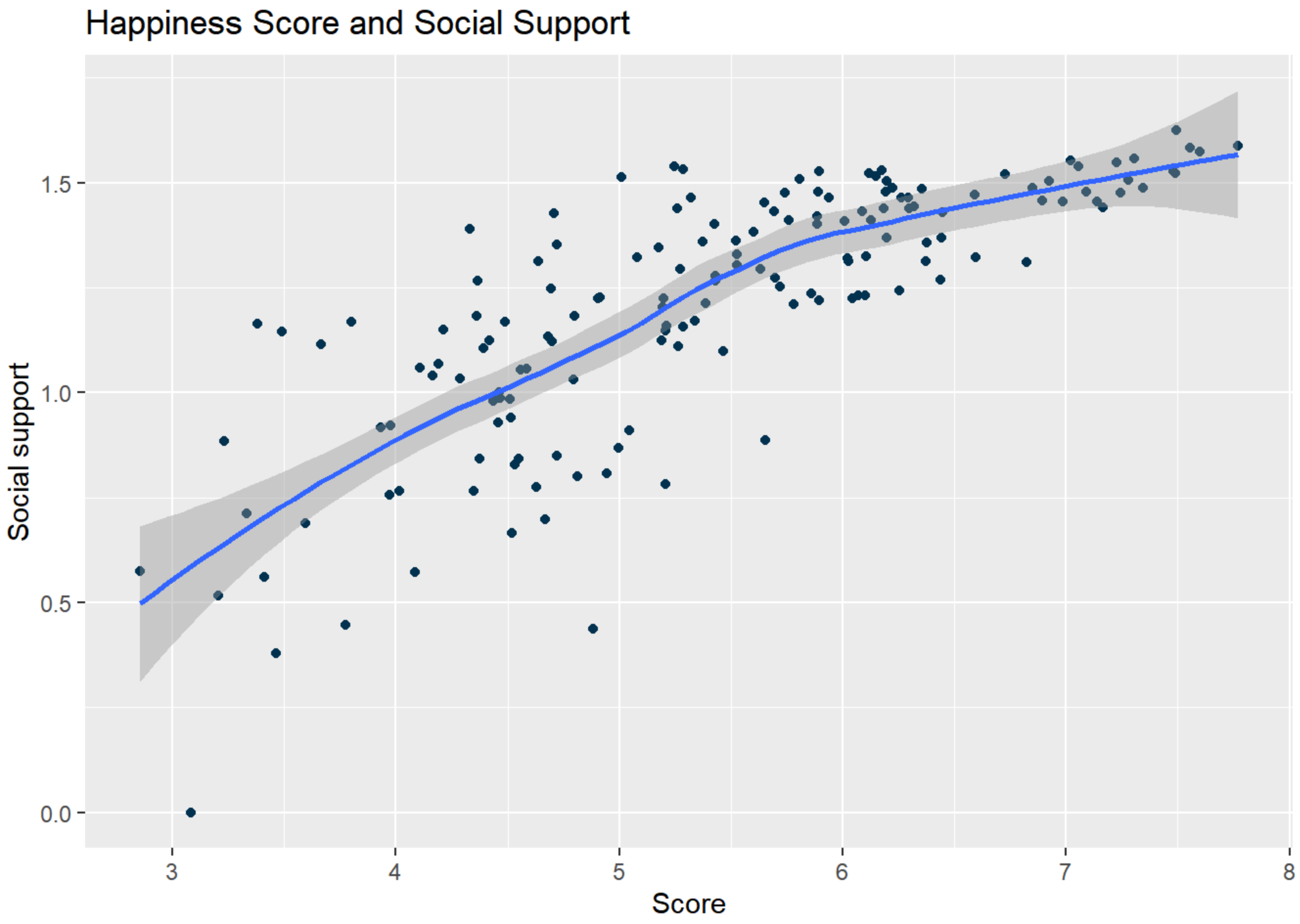
This is an analysis of the world happiness data from 2019. The data was sourced from Kaggle, and it includes the happiness data of 156 countries/regions. I will focus on predicting a country’s overall happiness score using GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption as predictors.

```
summary(happiness)
```

```
##   Overall rank    Country or region      Score       GDP per capita
##  Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000
##  1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028
##  Median : 78.50   Mode  :character   Median :5.380   Median :0.9600
##  Mean   : 78.50                      Mean   :5.407   Mean   :0.9051
##  3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325
##  Max.   :156.00                      Max.   :7.769   Max.   :1.6840
##  Social support  Healthy life expectancy Freedom to make life choices
##  Min.   :0.000   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.056   1st Qu.:0.5477          1st Qu.:0.3080
##  Median :1.272   Median :0.7890          Median :0.4170
##  Mean   :1.209   Mean   :0.7252          Mean   :0.3926
##  3rd Qu.:1.452   3rd Qu.:0.8818          3rd Qu.:0.5072
##  Max.   :1.624   Max.   :1.1410          Max.   :0.6310
##    Generosity     Perceptions of corruption
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1087   1st Qu.:0.0470
##  Median :0.1775   Median :0.0855
##  Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.5660   Max.   :0.4530
```

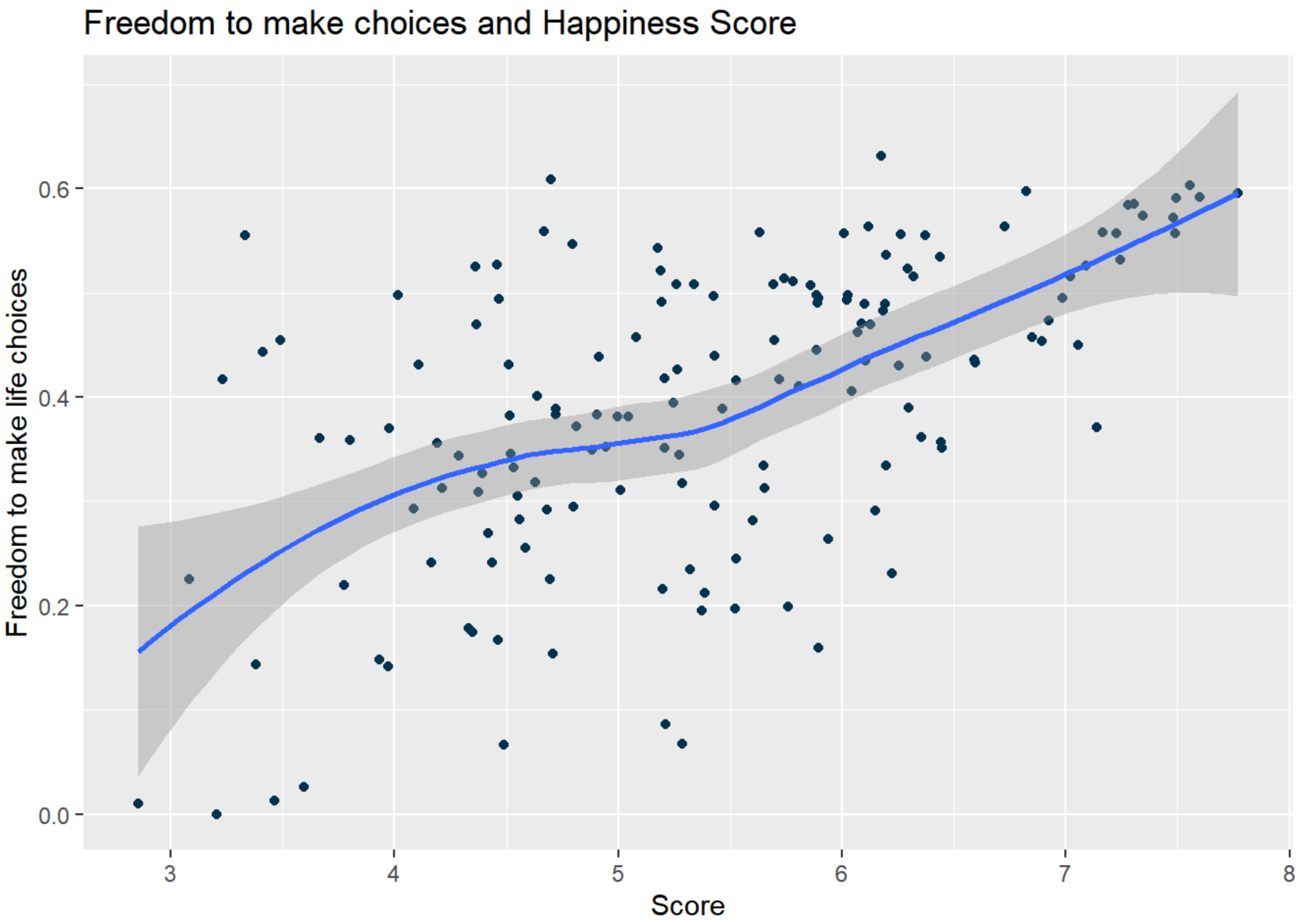
```
ggplot(happiness, aes(x = Score, y = `Social support`)) + geom_point(color = "#003554") + labs(t
itle = "Happiness Score and Social Support") + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



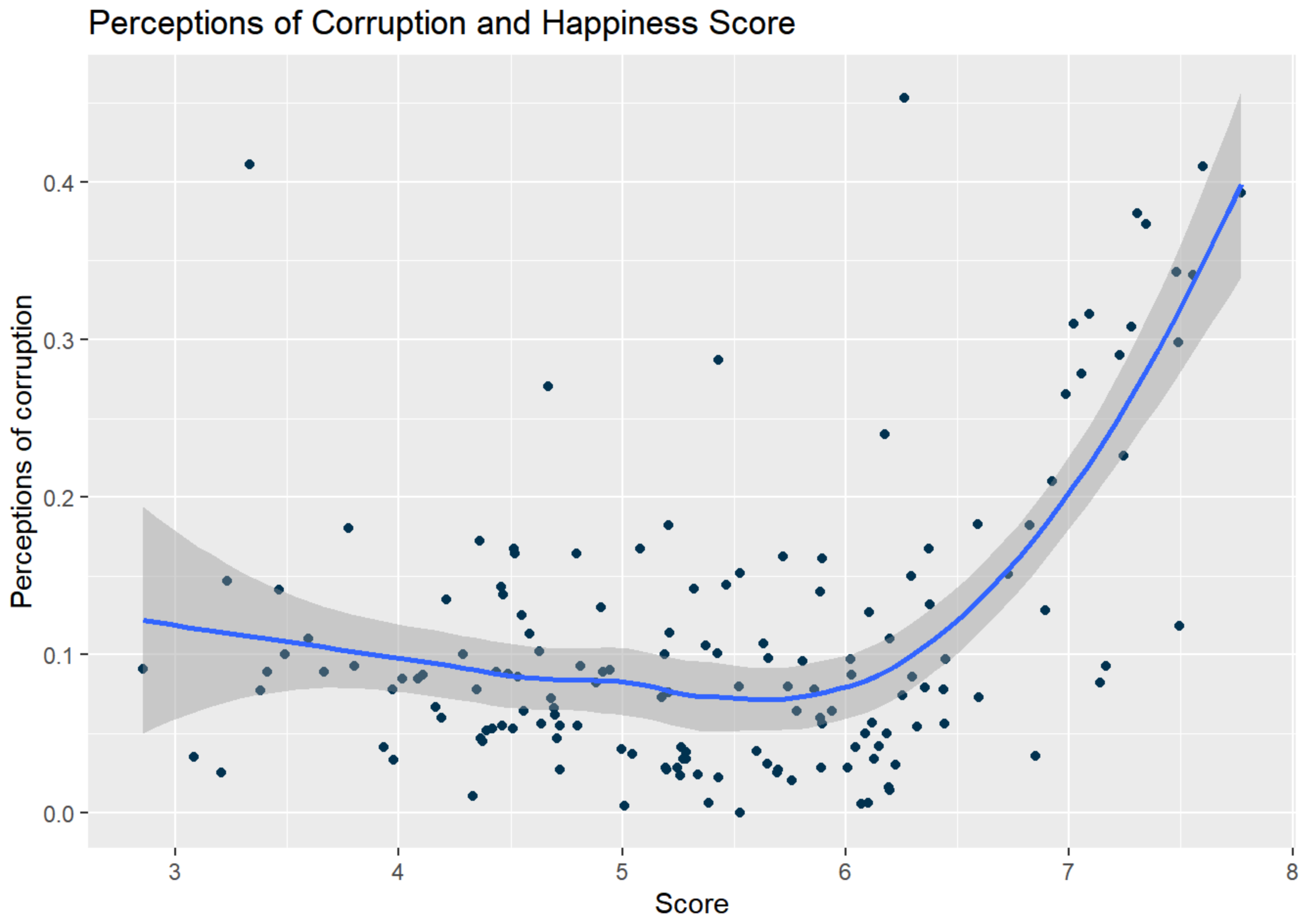
```
ggplot(happiness, aes(x = Score, y = `Freedom to make life choices`)) + geom_point(color = "#003
554") + geom_smooth() + labs(title = "Freedom to make choices and Happiness Score")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



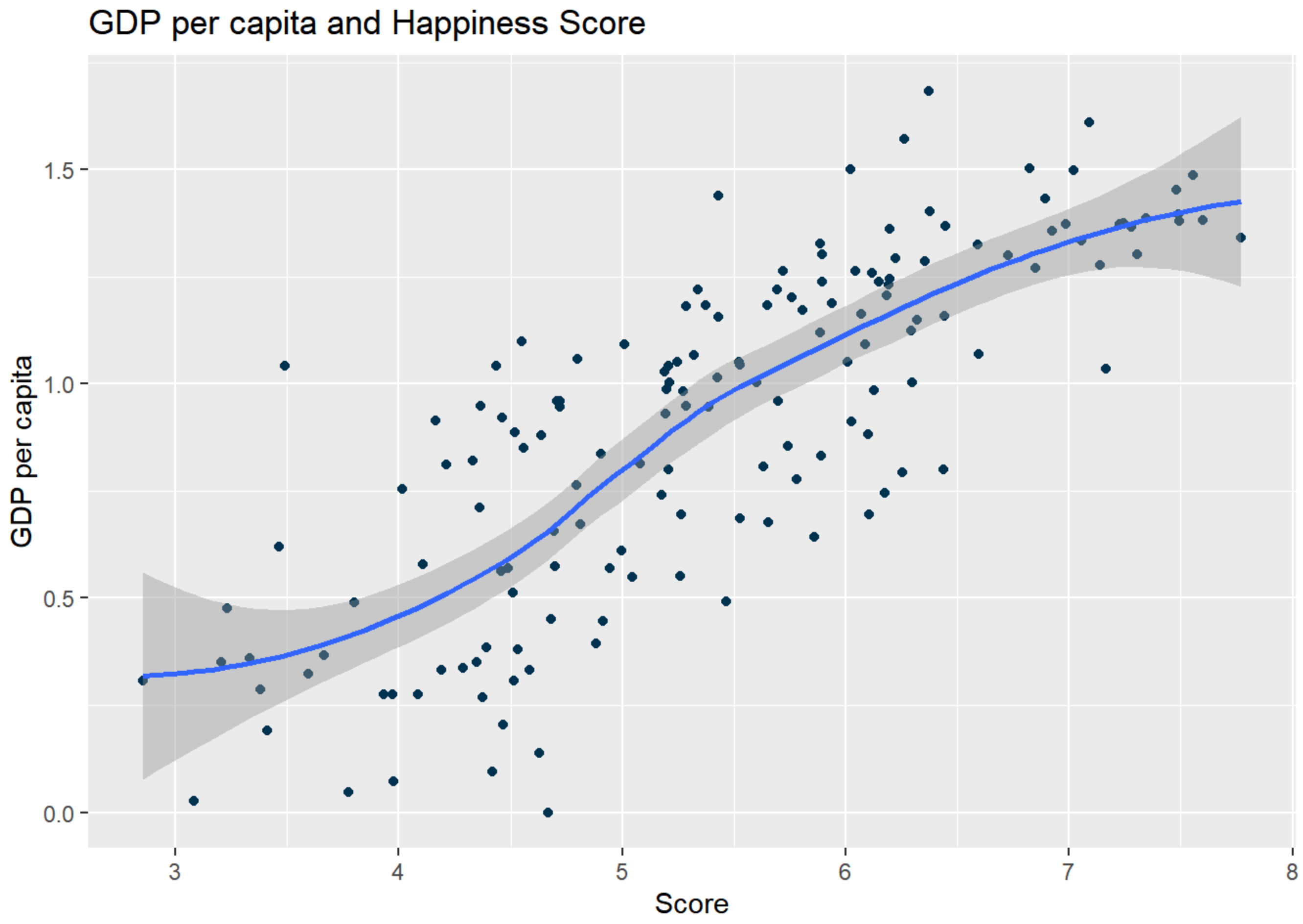
```
ggplot(happiness, aes(x = Score, y = `Perceptions of corruption`)) + geom_point(color = "#00355
4") + geom_smooth() + labs(title = "Perceptions of Corruption and Happiness Score")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



```
ggplot(happiness, aes(x = Score, y = `GDP per capita`)) + geom_point(color = "#003554") + geom_smooth() + labs(title = "GDP per capita and Happiness Score")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Numerical Summary

The lowest happiness score is 2.853 in South Sudan while the highest score is 7.769 in Finland. The only country with a Social support score of 0 is the Central African Republic. Greece has a generosity of 0, Swaziland has a health life expectancy of 0, Somalia has a GDP per capita of 0, Afghanistan has a 0 for freedom to make life choices (closely followed by South Sudan with 0.010), and Moldova has a 0 for Perceptions of corruption. These might be data entry errors, but I do not have enough information to exclude these points from the data.

There doesn't appear to be any concerning skews in the data. There is a positive, exponential relationship between Perception of corruption and the happiness score. This could be a result of better education in countries that are happier and therefore the citizens are more aware of how their government functions. However, there is not education measurement in this data set, so we will not be able to explore this possible trend.

Building Linear Models to Predict Happiness Scores

Full Model

```
#full model predicting happiness score
lm_mod <- lm(Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + `Freedom
to make life choices` + Generosity + `Perceptions of corruption`, happiness)
summary(lm_mod)
```

```
## 
## Call:
## lm(formula = Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + 
##     `Freedom to make life choices` + Generosity + `Perceptions of corruption`, 
##     data = happiness)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.75304 -0.35306  0.05703  0.36695  1.19059 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      1.7952     0.2111   8.505 1.77e-14 ***
## `GDP per capita`                 0.7754     0.2182   3.553 0.000510 ***
## `Social support`                 1.1242     0.2369   4.745 4.83e-06 ***
## `Healthy life expectancy`        1.0781     0.3345   3.223 0.001560 ** 
## `Freedom to make life choices`   1.4548     0.3753   3.876 0.000159 ***
## Generosity                       0.4898     0.4977   0.984 0.326709    
## `Perceptions of corruption`      0.9723     0.5424   1.793 0.075053 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5335 on 149 degrees of freedom
## Multiple R-squared:  0.7792, Adjusted R-squared:  0.7703 
## F-statistic: 87.62 on 6 and 149 DF,  p-value: < 2.2e-16
```

GDP per capita, Social support, Healthy life expectancy, and freedom to make life choices are all significant predictors of a country’s overall happiness score. Generosity and perceptions of corruption are not significant predictors in this model, but this does not mean that they are never significant predictors of the happiness score. The EDA shows that there is collinearity between many of the predictors, so it is possible that generosity and perceptions of corruption simply don’t provide anymore information than the predictors already in the model.

Correlations between predictors

```
cor(happiness$Generosity, happiness$`Perceptions of corruption`)
```

```
## [1] 0.3265375
```

```
round(cor(happiness[,-2]),2)
```

```
##                                Overall rank Score GDP per capita Social support
## Overall rank                           1.00 -0.99          -0.80          -0.77
## Score                                 -0.99  1.00           0.79           0.78
## GDP per capita                        -0.80  0.79           1.00           0.75
## Social support                        -0.77  0.78           0.75           1.00
## Healthy life expectancy               -0.79  0.78           0.84           0.72
## Freedom to make life choices          -0.55  0.57           0.38           0.45
## Generosity                            -0.05  0.08          -0.08          -0.05
## Perceptions of corruption             -0.35  0.39           0.30           0.18
##                                Healthy life expectancy
## Overall rank                                     -0.79
## Score                                             0.78
## GDP per capita                                    0.84
## Social support                                    0.72
## Healthy life expectancy                           1.00
## Freedom to make life choices                      0.39
## Generosity                                       -0.03
## Perceptions of corruption                         0.30
##                                Freedom to make life choices Generosity
## Overall rank                                          -0.55      -0.05
## Score                                                  0.57       0.08
## GDP per capita                                         0.38      -0.08
## Social support                                         0.45      -0.05
## Healthy life expectancy                                0.39      -0.03
## Freedom to make life choices                           1.00       0.27
## Generosity                                             0.27       1.00
## Perceptions of corruption                              0.44       0.33
##                                Perceptions of corruption
## Overall rank                                       -0.35
## Score                                               0.39
## GDP per capita                                      0.30
## Social support                                      0.18
## Healthy life expectancy                             0.30
## Freedom to make life choices                        0.44
## Generosity                                          0.33
## Perceptions of corruption                           1.00
```

Generosity doesn’t have particularly strong correlation with any of the other predictors, but the correlation between perceptions of corruption and freedom to make life choices is .44 which might be why perceptions of corruption is not significant in the model.

Is Perceptions of corruptions significant when freedom to make life choices is removed from the model?

```
model <- lm(Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + Generosity
+ `Perceptions of corruption`, happiness)
summary(model)
```

```
##
## Call:
## lm(formula = Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     Generosity + `Perceptions of corruption`, data = happiness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5590 -0.3089 -0.0166  0.3919  1.3653
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.8706     0.2198   8.511 1.65e-14 ***
## `GDP per capita`              0.7545     0.2281   3.307  0.00118 **
## `Social support`              1.3940     0.2368   5.887 2.48e-08 ***
## `Healthy life expectancy`     1.1318     0.3495   3.238  0.00148 **
## Generosity                    0.9034     0.5084   1.777  0.07761 .
## `Perceptions of corruption`   1.6343     0.5383   3.036  0.00283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5579 on 150 degrees of freedom
## Multiple R-squared:  0.7569, Adjusted R-squared:  0.7488
## F-statistic:  93.4 on 5 and 150 DF,  p-value: < 2.2e-16
```

When freedom to make life choices is removed from the model, perception of corruption is significant when $\alpha = .05$, but Generosity is not significant.

Confidence Intervals for the full model

```
confint(lm_mod)
```

```
##                                      2.5 %   97.5 %
## (Intercept)                     1.37813642 2.212304
## `GDP per capita`                0.34415545 1.206588
## `Social support`                0.65607391 1.592309
## `Healthy life expectancy`       0.41709028 1.739195
## `Freedom to make life choices`  0.71315985 2.196505
## Generosity                     -0.49376827 1.473335
## `Perceptions of corruption`    -0.09943173 2.043992
```

When we check the confidence intervals for the predictors, we see that 0 is the interval for Generosity and Perceptions of corruption. This means that 95% the coefficents for these variables can equal 0, and these predictors are likely not significant for the model.

Model without Generosity as a predictor and confidence interval

```
#model excluding generosity as a predictor
lm_mod2 <- lm(Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + `Freedom
to make life choices` + `Perceptions of corruption`, happiness)
summary(lm_mod2)
```

```
##
## Call:
## lm(formula = Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + `Perceptions of corruption`,
##     data = happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82997 -0.35344  0.05803  0.35977  1.17522
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.8689     0.1973   9.471  < 2e-16 ***
## `GDP per capita`                 0.7455     0.2161   3.450 0.000728 ***
## `Social support`                 1.1180     0.2368   4.722 5.33e-06 ***
## `Healthy life expectancy`        1.0840     0.3344   3.241 0.001467 **
## `Freedom to make life choices`   1.5340     0.3666   4.185 4.84e-05 ***
## `Perceptions of corruption`      1.1176     0.5218   2.142 0.033839 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5335 on 150 degrees of freedom
## Multiple R-squared:  0.7777, Adjusted R-squared:  0.7703
## F-statistic:   105 on 5 and 150 DF,  p-value: < 2.2e-16
```

```
confint(lm_mod2)
```

```
##                                     2.5 %   97.5 %
## (Intercept)                    1.47895629 2.258789
## `GDP per capita`               0.31851105 1.172394
## `Social support`               0.65015235 1.585911
## `Healthy life expectancy`      0.42317492 1.744857
## `Freedom to make life choices` 0.80969634 2.258322
## `Perceptions of corruption`    0.08647588 2.148630
```

All the predictors are significant when Generosity is removed from the model. The confidence intervals for the predictors also show that 0 is not in the 95% confidence interval for any of the predictors. However, we should run a conclusive test to see if the models with and without Generosity are significantly different from each other. The null hypothesis will be that Generosity = 0 and should therefore not be included in the model and the alternative hypothesis is that Generosity $\neq$ 0 and should be included in the model.

ANOVA for full model vs model without generosity

```
#ANOVA test to see if the model with generosity is significantly different from the model withou
t generosity

anova(lm_mod2, lm_mod)
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + `Perceptions of corruption`
## Model 2: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + Generosity + `Perceptions of corruption`
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    150 42.687
## 2    149 42.412  1   0.27561 0.9683 0.3267
```

The p-value from the ANOVA test is .32. Since this is greater than .05, we fail to reject the null meaning that we do not have sufficient evidence to support that Generosity $\neq 0$, and therefore we should exclude it from the model because the simpler model is always preferred.

The Perceptions of corruption had 0 in the confidence interval for the full model, so we may want to test if it can also be removed from the model. We will run 2 tests, the first test will compare a model without Perceptions of corruption and Generosity to the model without Generosity. The second tests will compare a model without Perceptions of corruption and Generosity to the model with all the predictors.

More ANOVAs

```
lm_mod3 <- lm(Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + `Freedom
to make life choices`, happiness)
summary(lm_mod3)
```

```
##
## Call:
## lm(formula = Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices`, data = happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86584 -0.34594  0.03403  0.43676  1.13076
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.8921     0.1994   9.491  < 2e-16 ***
## `GDP per capita`                 0.8105     0.2165   3.745 0.000256 ***
## `Social support`                 1.0166     0.2347   4.331 2.70e-05 ***
## `Healthy life expectancy`        1.1414     0.3373   3.384 0.000910 ***
## `Freedom to make life choices`   1.8458     0.3404   5.423 2.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5398 on 151 degrees of freedom
## Multiple R-squared:  0.7709, Adjusted R-squared:  0.7649
## F-statistic:   127 on 4 and 151 DF,  p-value: < 2.2e-16
```

```
#model without Perceptions and generosity vs model without generosity
anova(lm_mod3, lm_mod2)
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices`
## Model 2: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + `Perceptions of corruption`
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    151 43.993
## 2    150 42.687  1    1.3053 4.5866 0.03384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value from the ANOVA test is .03. Since this is less than .05, we reject the null meaning that we have sufficient evidence to support that Perceptions of corruptions $\neq$ 0, and therefore we should not exclude it from the model.

```
#model without Perceptions and generosity vs full model
anova(lm_mod3, lm_mod)
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices`
## Model 2: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + Generosity + `Perceptions of corruption`
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    151 43.993
## 2    149 42.412  2    1.5809 2.7769 0.06545 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value from the ANOVA test is .06. Since this is not less than .05, we fail to reject the null meaning that we do not have sufficient evidence to support that Perceptions of corruptions $\neq$ 0 and Generosity $\neq$ 0, and therefore we should exclude these predictors from the model.

These tests tell us that it is better to use a model without Perceptions of corruption and Generosity than to use the full model; however, it is best to only exclude Generosity and keep Perceptions of corruption as a predictor.

Is a model without generosity always better?

This is not a situation where any model without Generosity is the best choice. For example, let’s exclude Social support and Generosity from the model and compare it to the full model.

```
#model with Social Support and Generosity compared to the full model
mod4 <- lm_mod3 <- lm(Score ~ `GDP per capita` + `Healthy life expectancy` + `Freedom to make li
fe choices` + `Perceptions of corruption`, happiness)
summary(mod4)
```

```
##
## Call:
## lm(formula = Score ~ `GDP per capita` + `Healthy life expectancy` +
##     `Freedom to make life choices` + `Perceptions of corruption`,
##     data = happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92776 -0.35241  0.08972  0.38163  0.95621
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.4366     0.1671  14.578  < 2e-16 ***
## `GDP per capita`                 1.1622     0.2107   5.517 1.47e-07 ***
## `Healthy life expectancy`        1.4433     0.3479   4.149 5.56e-05 ***
## `Freedom to make life choices`   2.0447     0.3741   5.465 1.87e-07 ***
## `Perceptions of corruption`      0.6248     0.5461   1.144    0.254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5698 on 151 degrees of freedom
## Multiple R-squared:  0.7447, Adjusted R-squared:  0.7379
## F-statistic: 110.1 on 4 and 151 DF,  p-value: < 2.2e-16
```

```
anova(mod4, lm_mod)
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ `GDP per capita` + `Healthy life expectancy` + `Freedom to make life choices
` +
##     `Perceptions of corruption`
## Model 2: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + Generosity + `Perceptions of corruption`
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    151 49.032
## 2    149 42.412  2    6.6199 11.628 2.029e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the p-value for the anova test is close to 0 which says we should reject the null (H_0 = Social support = Generosity = 0) and choose the alternative hypothesis which is the full model in this scenario.