Life Expectancy Analysis

2025-06-29

Life Expectancy

```
life_expectancy <- read.csv("C:/Users/autis/OneDrive/Documents/Life Expectancy/Data/life_expectancy.csv
```

Data was sourced from Kaggle: Countries Life Expectancy Dataset.

```
colSums(is.na(life_expectancy))  # Check for missing values
```

```
##              Country                 Year               Status
##                    0                    0                    0
##           Population          Hepatitis.B              Measles
##                  644                  542                    0
##                Polio          Diphtheria             HIV.AIDS
##                   19                   19                    0
##        infant.deaths    under.five.deaths    Total.expenditure
##                    0                    0                  221
##                  GDP                  BMI thinness..1.19.years
##                  442                   32                   32
##              Alcohol            Schooling      Life.expectancy
##                  188                  160                    0
```

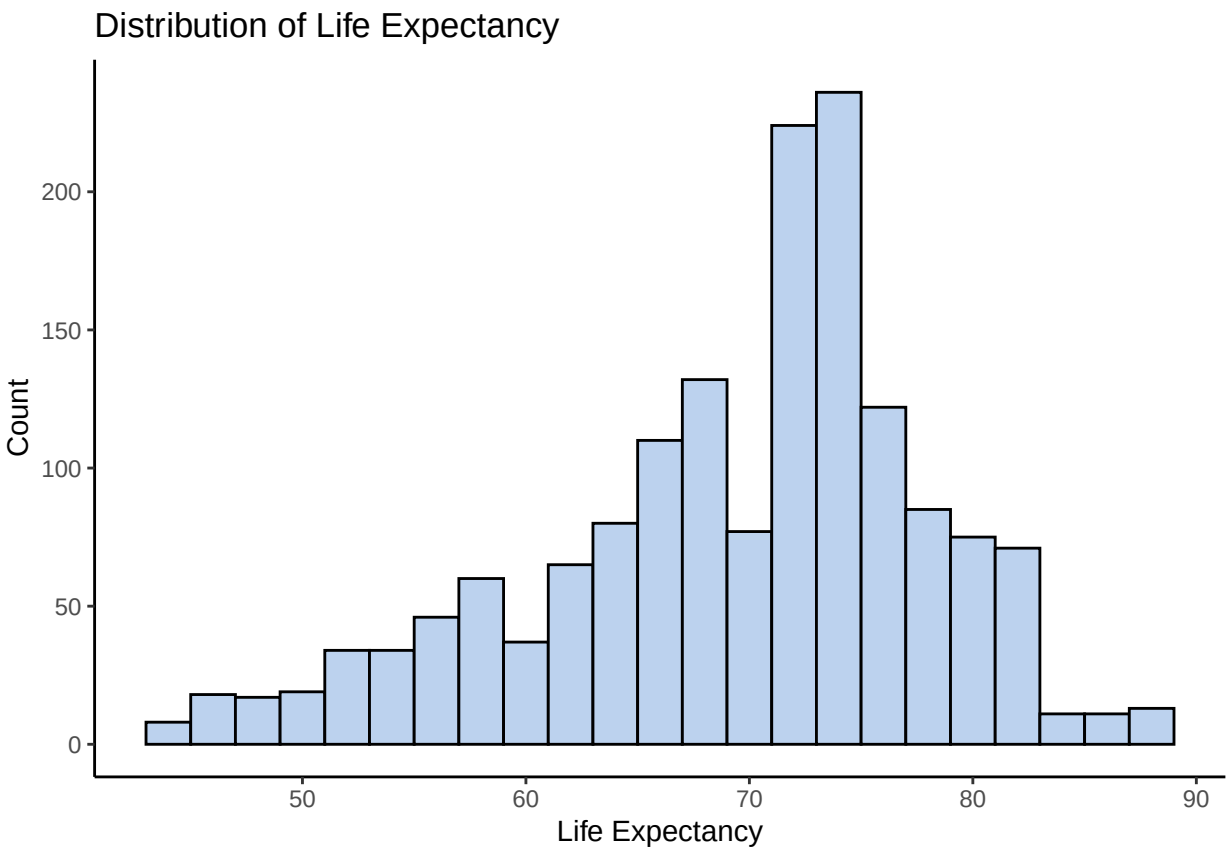
```
life_expectancy <- life_expectancy |>
  drop_na() #Remove missing Values
```

```
#Overall Summary Statistics for every column
summary(life_expectancy)
```

```
##    Country               Year         Status            Population
##  Length:1585        Min.   :2000   Length:1585        Min.   :3.400e+01
##  Class :character   1st Qu.:2005   Class :character   1st Qu.:1.965e+05
##  Mode  :character   Median :2008   Mode  :character   Median :1.432e+06
##                     Mean   :2008                      Mean   :1.487e+07
##                     3rd Qu.:2011                      3rd Qu.:8.121e+06
##                     Max.   :2015                      Max.   :1.294e+09
##   Hepatitis.B       Measles             Polio          Diphtheria
##  Min.   : 2.00   Min.   :     0   Min.   : 3.00   Min.   : 2.00
##  1st Qu.:75.00   1st Qu.:     0   1st Qu.:81.00   1st Qu.:82.00
##  Median :89.00   Median :    13   Median :93.00   Median :93.00
##  Mean   :79.33   Mean   :  1629   Mean   :83.72   Mean   :84.19
##  3rd Qu.:96.00   3rd Qu.:   334   3rd Qu.:97.00   3rd Qu.:97.00
##  Max.   :99.00   Max.   :118712   Max.   :99.00   Max.   :99.00
```

```
##     HIV.AIDS      infant.deaths     under.five.deaths Total.expenditure
##  Min.   : 0.100   Min.   :   0.00   Min.   :   0.00   Min.   : 0.740
##  1st Qu.: 0.100   1st Qu.:   1.00   1st Qu.:   1.00   1st Qu.: 4.380
##  Median : 0.100   Median :   3.00   Median :   4.00   Median : 5.840
##  Mean   : 2.025   Mean   :  29.92   Mean   :  41.01   Mean   : 5.957
##  3rd Qu.: 0.600   3rd Qu.:  21.00   3rd Qu.:  25.00   3rd Qu.: 7.500
##  Max.   :50.600   Max.   :1600.00   Max.   :2100.00   Max.   :14.390
##       GDP                BMI        thinness..1.19.years    Alcohol
##  Min.   :     1.68   Min.   : 2.00   Min.   : 0.100       Min.   : 0.010
##  1st Qu.:   475.11   1st Qu.:19.70   1st Qu.: 1.600       1st Qu.: 0.830
##  Median :  1644.82   Median :44.20   Median : 3.000       Median : 3.790
##  Mean   :  5686.05   Mean   :38.42   Mean   : 4.879       Mean   : 4.558
##  3rd Qu.:  4773.45   3rd Qu.:55.90   3rd Qu.: 7.100       3rd Qu.: 7.380
##  Max.   :119172.74   Max.   :77.10   Max.   :27.200       Max.   :17.870
##    Schooling     Life.expectancy
##  Min.   : 4.20   Min.   :44.00
##  1st Qu.:10.50   1st Qu.:64.70
##  Median :12.30   Median :71.70
##  Mean   :12.18   Mean   :69.41
##  3rd Qu.:14.00   3rd Qu.:75.00
##  Max.   :20.70   Max.   :89.00
```

```
life_expectancy |> #Creating a visual of life expectancy
  ggplot(aes(x = Life.expectancy)) +
  geom_histogram(binwidth = 2, fill = "lightsteelblue2", color = "black") +
  labs(
    title = "Distribution of Life Expectancy",
    x = "Life Expectancy",
    y = "Count"
  ) +
  theme_classic()
```



The overall distribution of life expectancy is left skewed.

What is the average life expectancy globally?

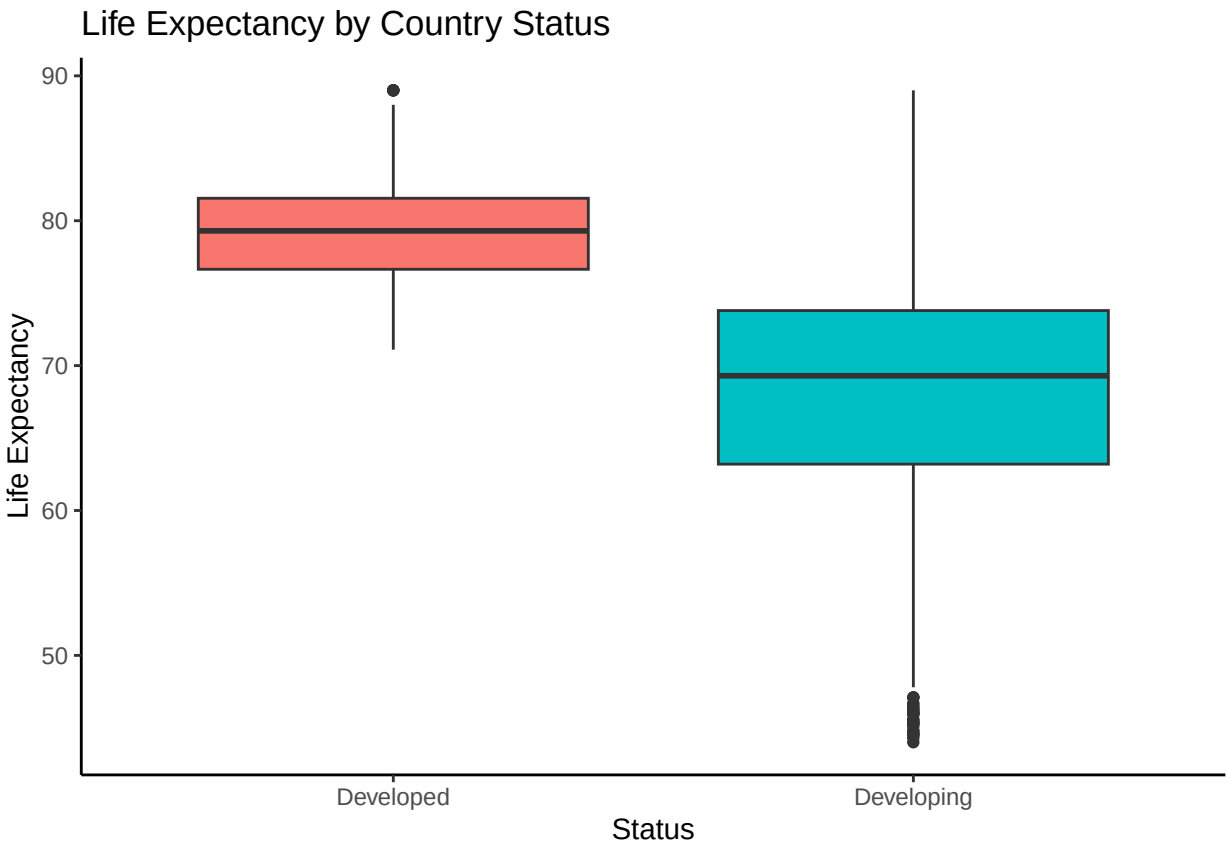
```
summary(life_expectancy$Life.expectancy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.00   64.70   71.70   69.41   75.00   89.00
```

The average global life expectancy is 69.35. The data visualization of life expectancy showed a left skew, making the median the better choice for measure of central tendency.

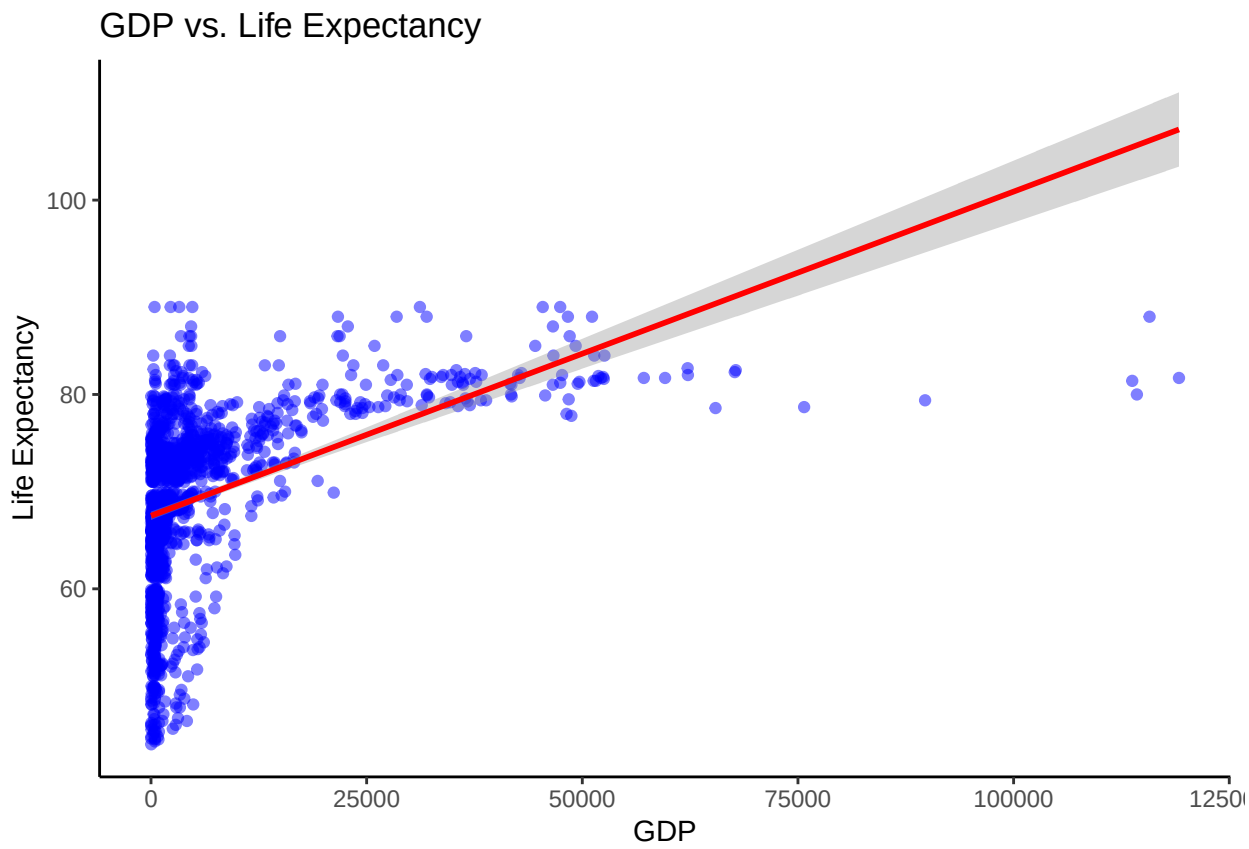
Creating a visual to compare life expectancy with developing countries vs developing countries.

```
life_expectancy |>
  ggplot(aes(x = Status, y = Life.expectancy, fill = Status)) +
  geom_boxplot() +
  labs(
    title = "Life Expectancy by Country Status",
    x = "Status",
    y = "Life Expectancy"
  ) +
  theme_classic() +
  theme(legend.position = "none")
```



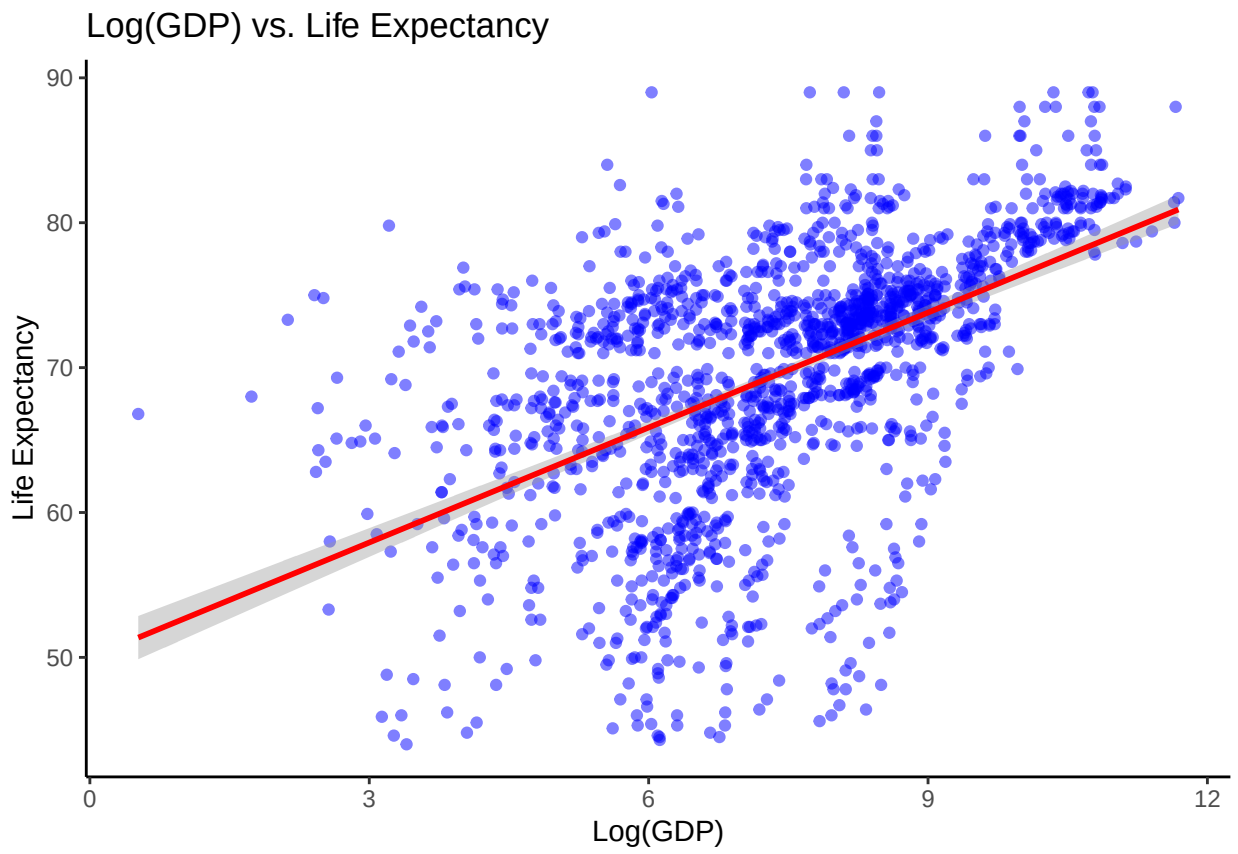
Does a higher GDP predict higher life expectancy?

```
life_expectancy |>
  ggplot(aes(x = GDP, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "GDP vs. Life Expectancy",
    x = "GDP",
    y = "Life Expectancy"
  ) +
  theme_classic()
```



The linearity was not very promising as it had a heavy skew. Attempting a log transformation to possibly linearize.

```
ggplot(life_expectancy, aes(x = log(GDP), y = Life.expectancy)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(title = "Log(GDP) vs. Life Expectancy", x = "Log(GDP)", y = "Life Expectancy") +
  theme_classic()
```



After conducting a log transformation, the data seems to have a more linear appearance despite the scatter, possibly a positive correlation between high GDP and life expectancy.

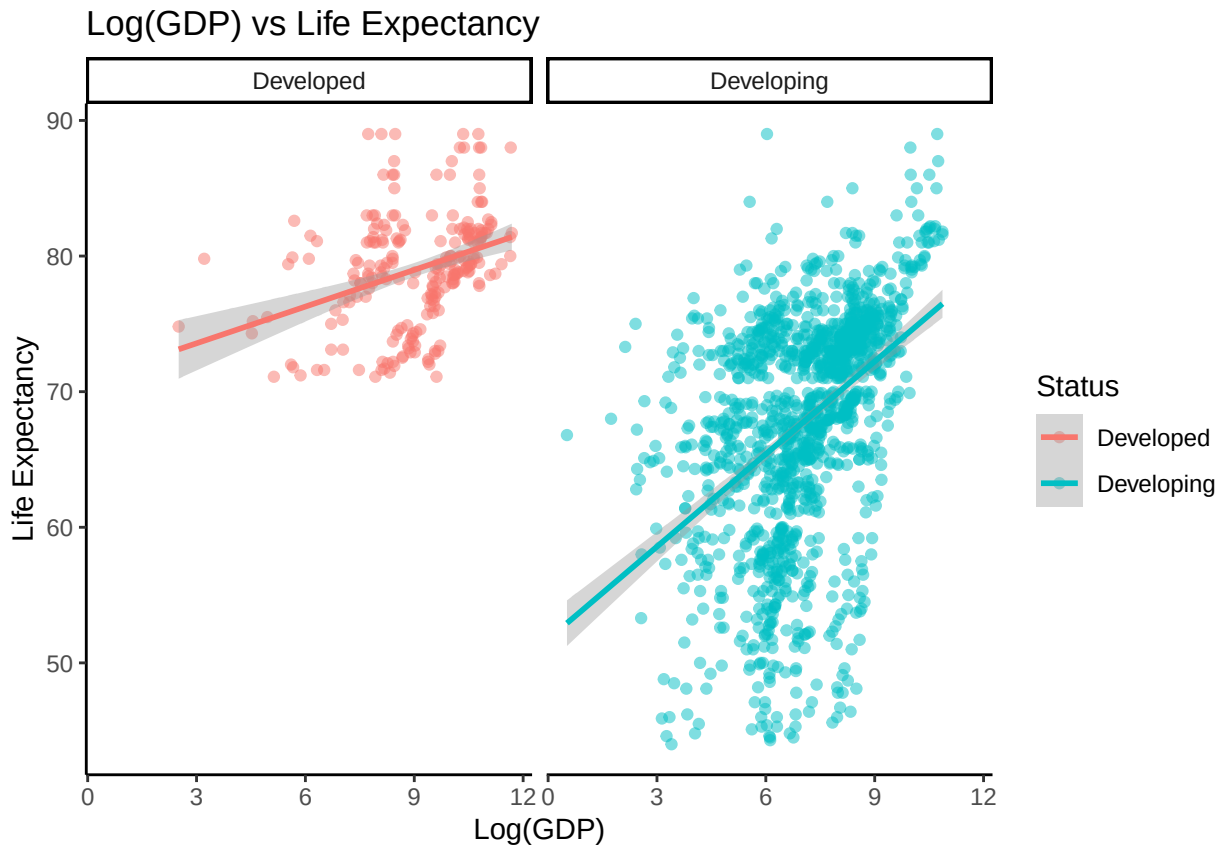
```
life_expectancy$log_gdp <- log(life_expectancy$GDP)
model_log <- lm(Life.expectancy ~ log_gdp, data = life_expectancy)
summary(model_log)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ log_gdp, data = life_expectancy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.637  -3.231   1.169   4.345  23.052
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.9909     0.8154   61.31   <2e-16 ***
## log_gdp       2.6453     0.1080   24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.52 on 1583 degrees of freedom
## Multiple R-squared:  0.2747, Adjusted R-squared:  0.2743
## F-statistic: 599.6 on 1 and 1583 DF,  p-value: < 2.2e-16
```

A small P-value shows a statistically significant relationship. An R-Squared Value of 0.2747 explains 27.47% of the variation in life expectancy. A higher GDP seems to indicate a higher life expectancy.

Check to see if it is also affected by the status of the country.

```
ggplot(life_expectancy, aes(x = log(GDP), y = Life.expectancy, color = Status)) + #Adding status of dev
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
  facet_wrap(~ Status) + #Want the graphs side by side
  theme_classic() +
  labs(
    title = "Log(GDP) vs Life Expectancy",
    x = "Log(GDP)",
    y = "Life Expectancy",
    color = "Status"
  )
```



By doing a comparison between both developing and developed countries the more developed a country is the higher the life expectancy is guaranteed. However there is a much larger disparity with developing counties as the life expectancy ranges from low, to high.

```
GDP_LE_Dev <- lm(Life.expectancy ~ log(GDP), data = filter(life_expectancy, Status == "Developed"))
summary(GDP_LE_Dev)
```

```
##
## Call:
```

```
## lm(formula = Life.expectancy ~ log(GDP), data = filter(life_expectancy,
##     Status == "Developed"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4331 -2.2726 -0.1546  1.9951 11.1653
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.8597     1.4908  47.532  < 2e-16 ***
## log(GDP)      0.9022     0.1622   5.561 7.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.913 on 225 degrees of freedom
## Multiple R-squared:  0.1208, Adjusted R-squared:  0.1169
## F-statistic: 30.93 on 1 and 225 DF,  p-value: 7.567e-08
```

```
GDP_LE_Devg <- lm(Life.expectancy ~ log(GDP), data = filter(life_expectancy, Status == "Developing"))
summary(GDP_LE_Devg)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ log(GDP), data = filter(life_expectancy,
##     Status == "Developing"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.314  -3.688   1.507   4.740  23.526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.7454     0.9225   56.09   <2e-16 ***
## log(GDP)      2.2760     0.1275   17.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.533 on 1356 degrees of freedom
## Multiple R-squared:  0.1903, Adjusted R-squared:  0.1898
## F-statistic: 318.8 on 1 and 1356 DF,  p-value: < 2.2e-16
```

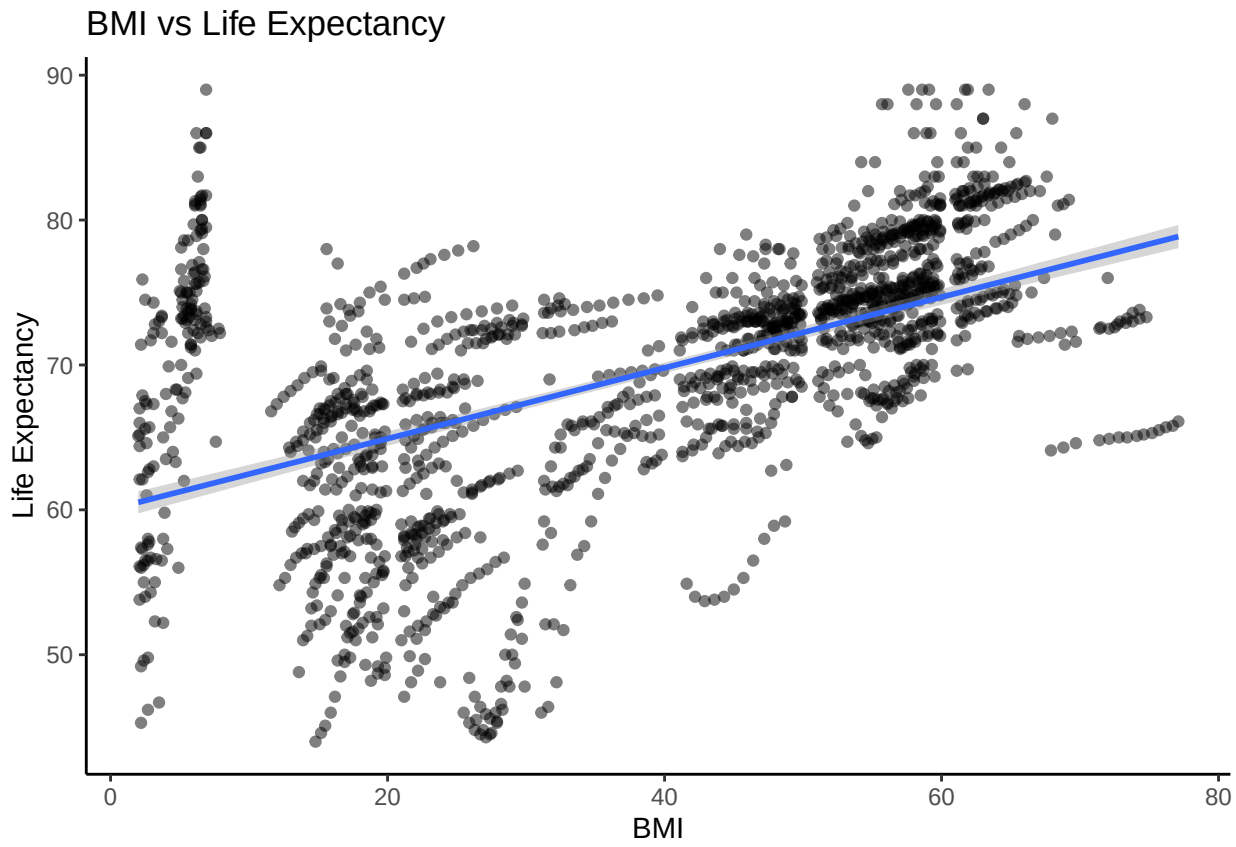
GDP is a significant predictor when it comes to life expectancy and is further confirmed when seperating whether or not the country is developed or developing. Both yield significant statistical results. However their R-squared statistics are closer to 0 than to one which does not explain much of the variation. Therefore life expectancy is determined by a multitude of factors and is not just determined by GDP. Further variables need to be tested.

How much do external factors impact life expectancy *Alcohol, schooling, BMI, HIV*

BMI

```
ggplot(life_expectancy, aes(x = BMI, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
```

```
  theme_classic() +
  labs(
    title = "BMI vs Life Expectancy",
    x = "BMI",
    y = "Life Expectancy",
  )
```




```
BMI_LE <- lm(Life.expectancy ~ BMI, data = life_expectancy)
summary(BMI_LE)
```

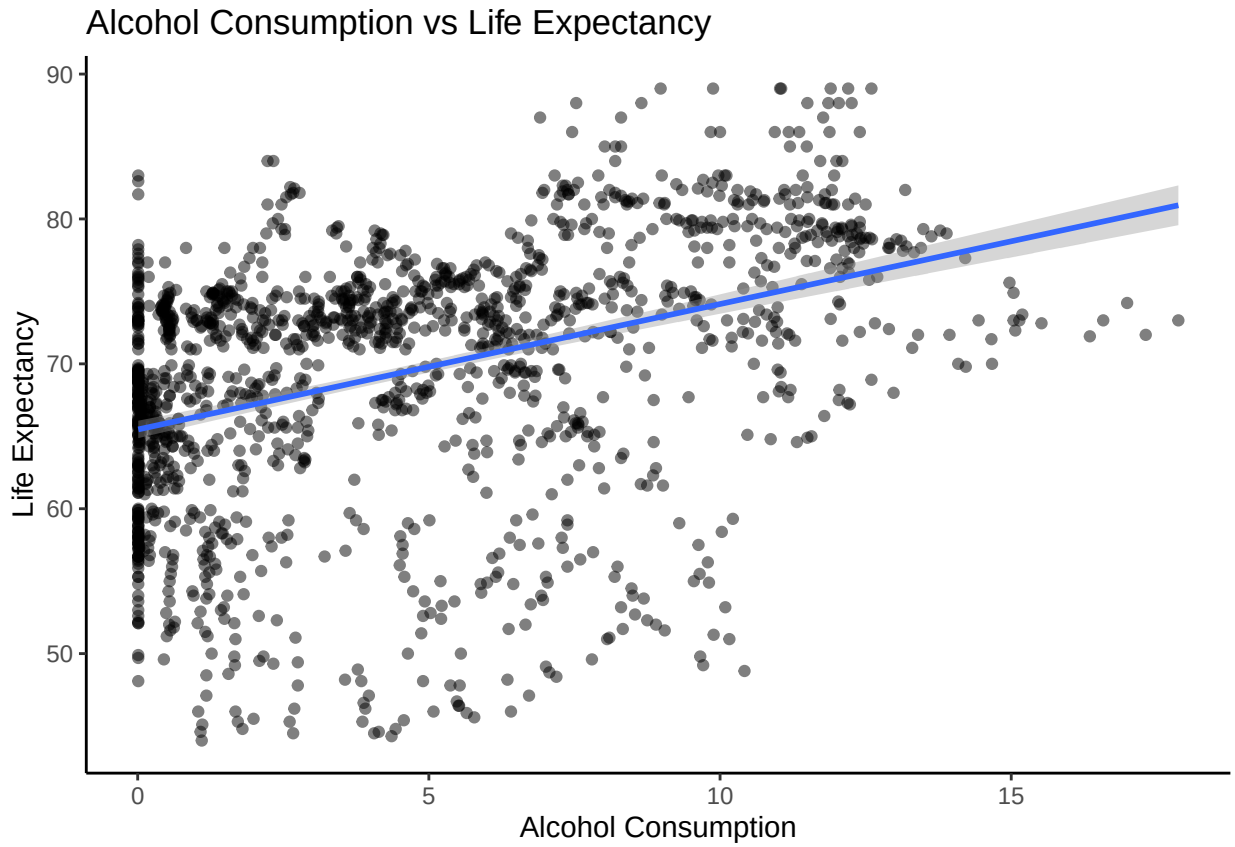
```
##
## Call:
## lm(formula = Life.expectancy ~ BMI, data = life_expectancy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3484  -4.4858   0.6684   4.6394  27.2859
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.028606   0.405755  147.94   <2e-16 ***
## BMI          0.244273   0.009391   26.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.391 on 1583 degrees of freedom
## Multiple R-squared:  0.2994, Adjusted R-squared:  0.299
## F-statistic: 676.6 on 1 and 1583 DF,  p-value: < 2.2e-16
```

p-value: < 2.2e-16 Adjusted R-squared: 0.299

Alcohol

```
ggplot(life_expectancy, aes(x = Alcohol, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
  theme_classic() +
  labs(
    title = "Alcohol Consumption vs Life Expectancy",
    x = "Alcohol Consumption",
    y = "Life Expectancy",
  )
```




```
Alcohol_LE <- lm(Life.expectancy ~ Alcohol, data = life_expectancy)
summary(Alcohol_LE)
```

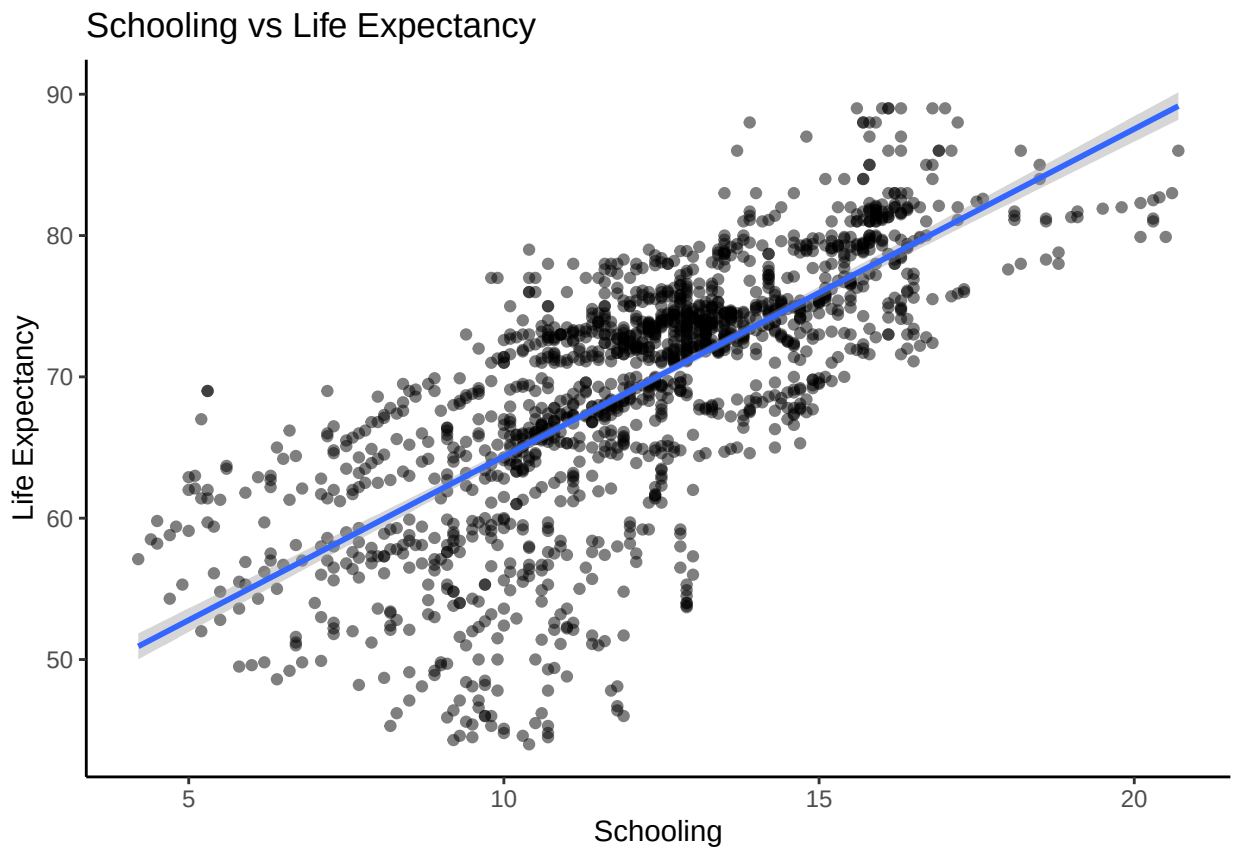
```
##
## Call:
## lm(formula = Life.expectancy ~ Alcohol, data = life_expectancy)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.691  -3.890   1.858   5.668  17.526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.46564    0.30664  213.49   <2e-16 ***
## Alcohol      0.86614    0.05032   17.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.104 on 1583 degrees of freedom
## Multiple R-squared:  0.1577, Adjusted R-squared:  0.1571
## F-statistic: 296.3 on 1 and 1583 DF,  p-value: < 2.2e-16
```

p-value: < 2.2e-16 Adjusted R-squared: 0.1571

Schooling

```
ggplot(life_expectancy, aes(x = Schooling, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
  theme_classic() +
  labs(
    title = "Schooling vs Life Expectancy",
    x = "Schooling",
    y = "Life Expectancy",
  )
```



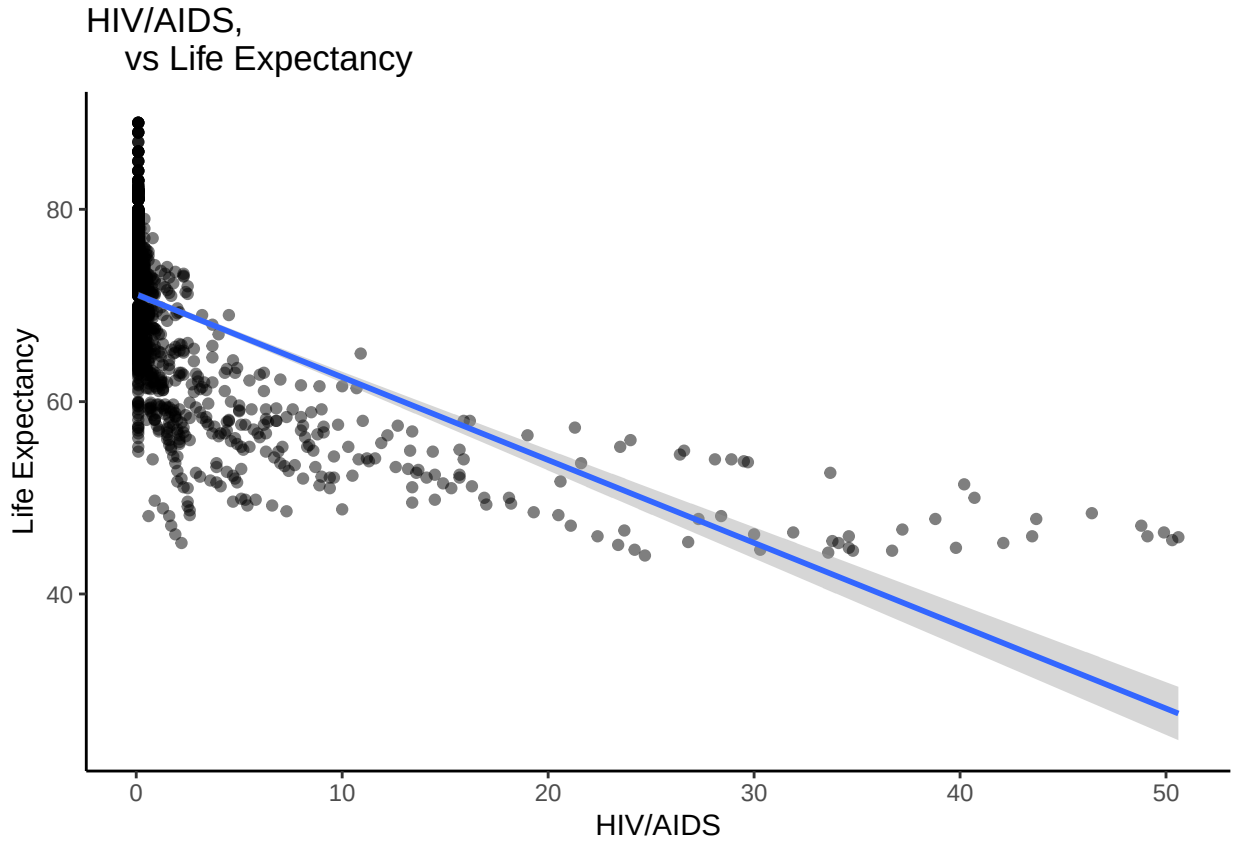

```
Schooling_LE <- lm(Life.expectancy ~ Schooling, data = life_expectancy)
summary(Schooling_LE)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling, data = life_expectancy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7707  -3.1339   0.8244   3.9170  15.5216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.1982     0.6905   59.66   <2e-16 ***
## Schooling     2.3170     0.0553   41.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.08 on 1583 degrees of freedom
## Multiple R-squared:  0.5258, Adjusted R-squared:  0.5255
## F-statistic:  1755 on 1 and 1583 DF,  p-value: < 2.2e-16
```

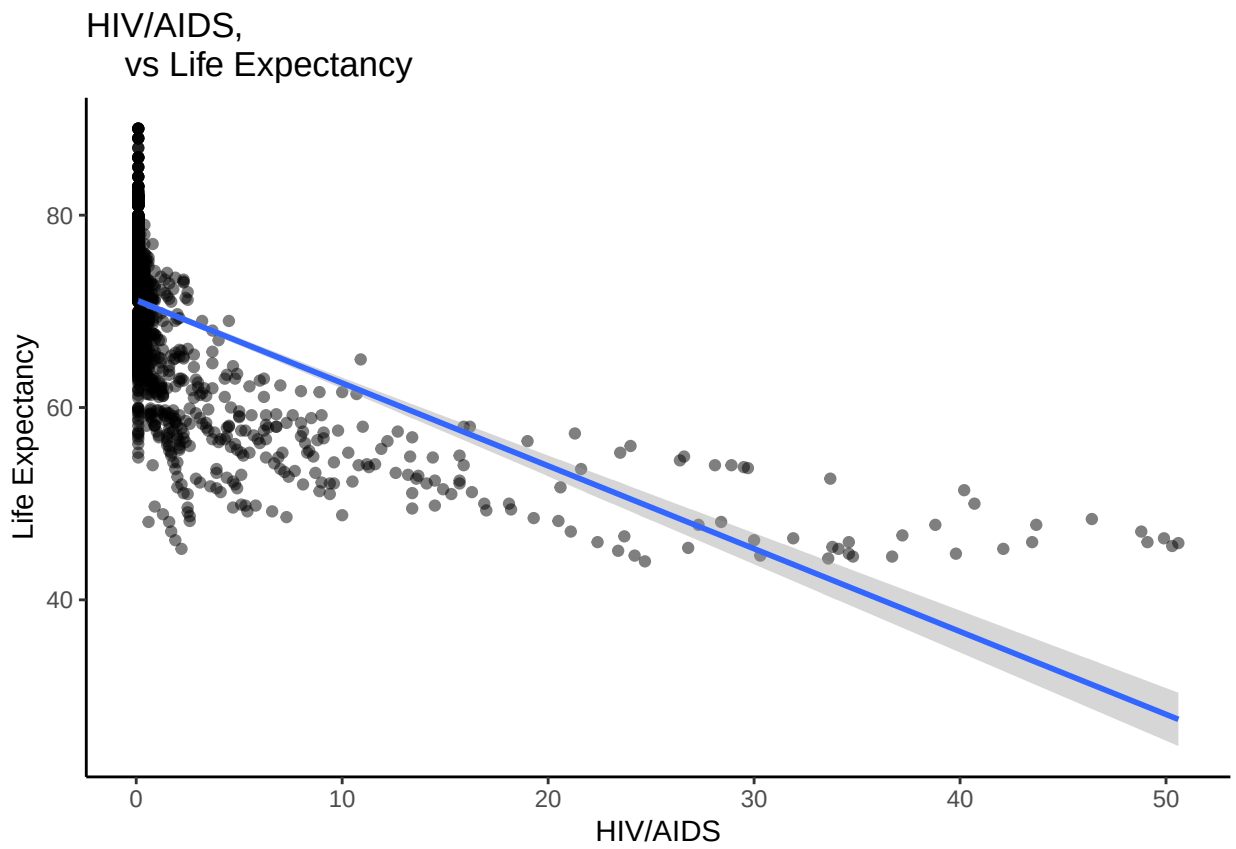
p-value: $<$ 2.2e-16 Adjusted R-squared: 0.5255 With real world data, an Adjusted R-Squared value of 0.5255 indicates that schooling plays a significant factor in life expectancy. A small p-value also indicates statistical significance.

HIV/AIDS

```
ggplot(life_expectancy, aes(x = HIV.AIDS, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
  theme_classic() +
  labs(
    title = "HIV/AIDS,
    vs Life Expectancy",
    x = "HIV/AIDS",
    y = "Life Expectancy",
  )
```




```
ggplot(life_expectancy, aes(x = HIV.AIDS, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
  theme_classic() +
  labs(
    title = "HIV/AIDS,
    vs Life Expectancy",
    x = "HIV/AIDS",
    y = "Life Expectancy",
  )
```



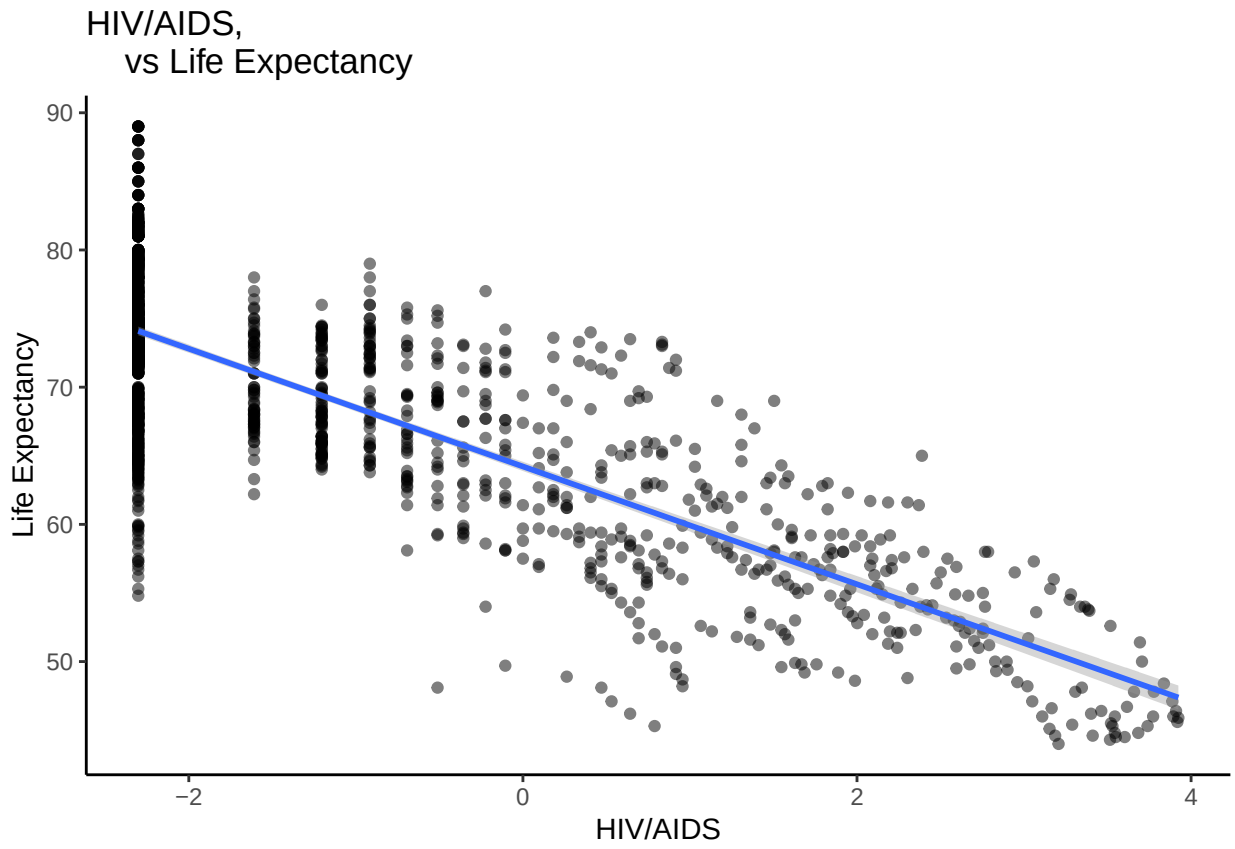

```
HIV_LE <- lm(Life.expectancy ~ HIV.AIDS, data = life_expectancy)
summary(HIV_LE)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ HIV.AIDS, data = life_expectancy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.962  -4.913   1.129   4.329  18.328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.15747    0.18687  380.79   <2e-16 ***
## HIV.AIDS    -0.86137    0.02889  -29.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.066 on 1583 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3592
## F-statistic: 889.1 on 1 and 1583 DF,  p-value: < 2.2e-16
```

p-value: < 2.2e-16 Adjusted R-squared: 0.3592

The HIV/AIDS variable showed right skewness, so I applied a log transformation to better linearize the model.

```r
ggplot(life_expectancy, aes(x = log(HIV.AIDS), y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1) +
  theme_classic() +
  labs(
    title = "HIV/AIDS,
    vs Life Expectancy",
    x = "HIV/AIDS",
    y = "Life Expectancy",
  )
```




```r
HIV_LE <- lm(Life.expectancy ~ log(HIV.AIDS), data = life_expectancy)
summary(HIV_LE)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ log(HIV.AIDS), data = life_expectancy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2991  -3.0248  -0.1471   3.6150  14.9009
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   64.21874    0.16830  381.57   <2e-16 ***
```

```
## log(HIV.AIDS) -4.29097    0.08287  -51.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.38 on 1583 degrees of freedom
## Multiple R-squared:  0.6288, Adjusted R-squared:  0.6285
## F-statistic:  2681 on 1 and 1583 DF,  p-value: < 2.2e-16
```

After a log transformation was applied the Adjusted R-Squared value increased to 0.6285 which indicates that HIV/AIDS is a strong predictor in a large real world data set as it explains 62.85% variation. A p-value of $<$ 2.2e-16 confirms that HIV/AIDS is statistically significant showing that it is a good indicator in predicting life expectancy.

Conclusion

In conclusion, GDP is a good starting point when it comes to life expectancy however there are other factors that seem to be stronger predictors in determining life expectancy. Developed countries seem to have a much higher life expectancy overall. Predictors such as Schooling, HIV/AIDS seem to be the strongest indicators when it comes to life expectancy. For future analysis focusing only on developing countries could yield better results as developed countries seem to overall have higher life expectancy due to access to schooling and medical care.