DIIG Data Challenge 2025

Liane Ma

2025-01-25

Factors that Impact Life Expectancy: Prioritizing WHO Efforts

Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(dplyr)
library(tibble)

life_exp <- read_csv("data/Life Expectancy Data.csv")
```

```
## Rows: 2938 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Country, Status
## dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, pe...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

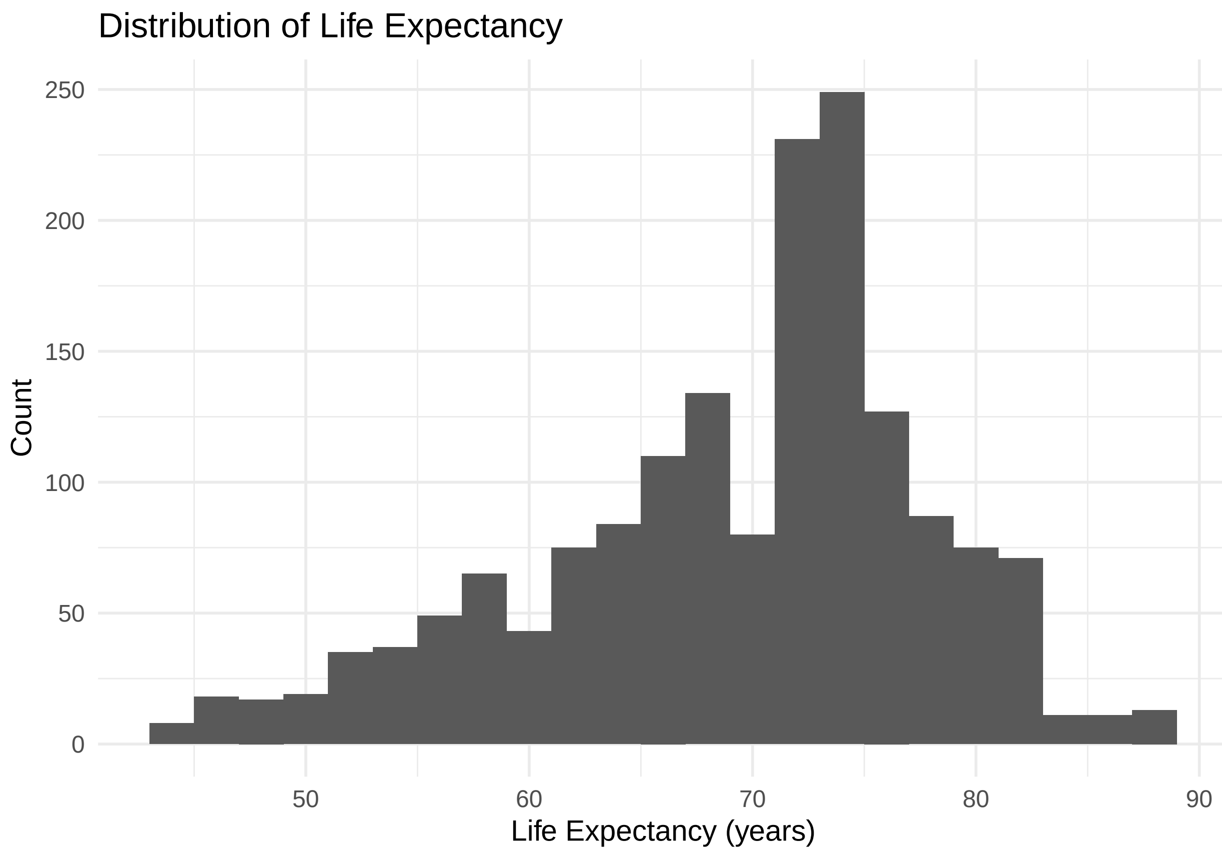
```
#cleaning
colnames(life_exp) <- gsub(" ", "_", colnames(life_exp))
colnames(life_exp) <- gsub("-", "_", colnames(life_exp))
colnames(life_exp) <- gsub("__", "_", colnames(life_exp))

life_exp_clean <- na.omit(life_exp)
```

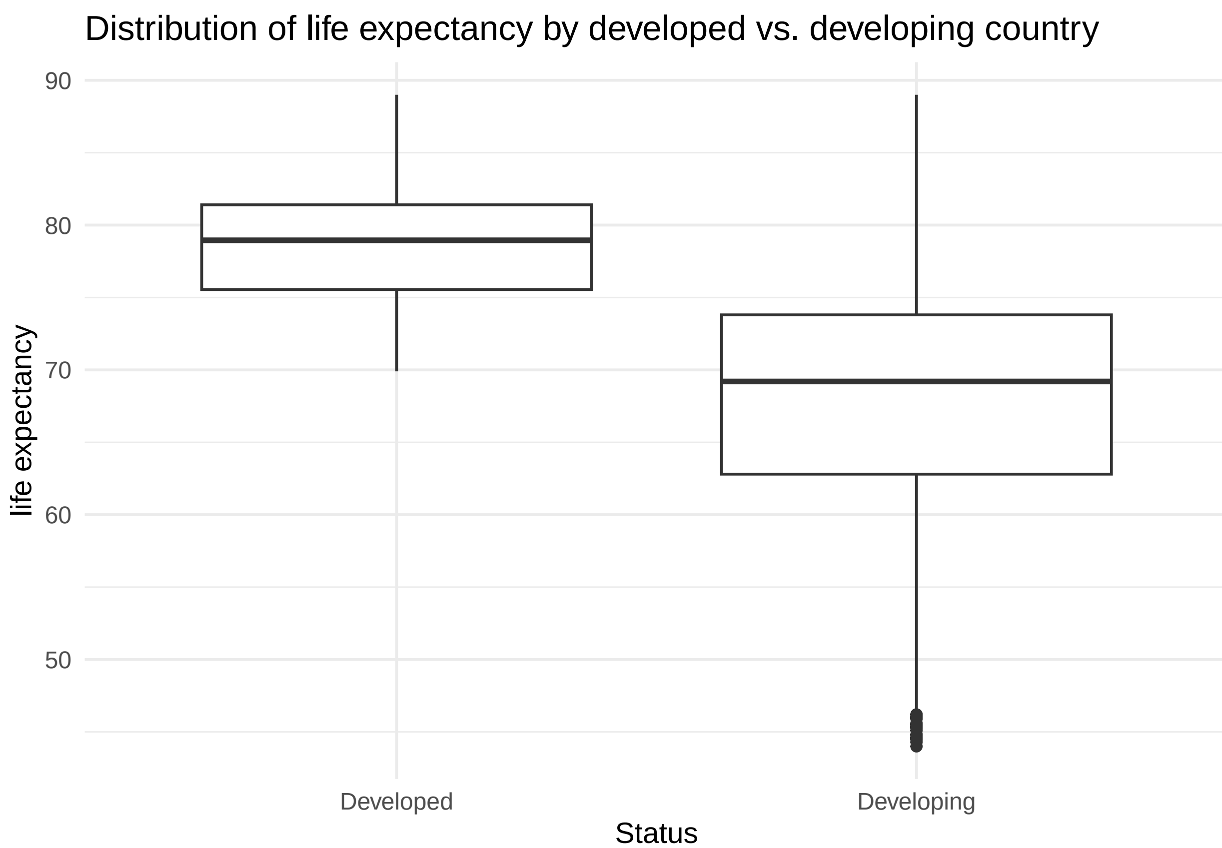
Exploratory Data Analysis

```
life_exp_clean |>
  ggplot(aes(x = Life_expectancy))+
  geom_histogram(binwidth = 2) +
```

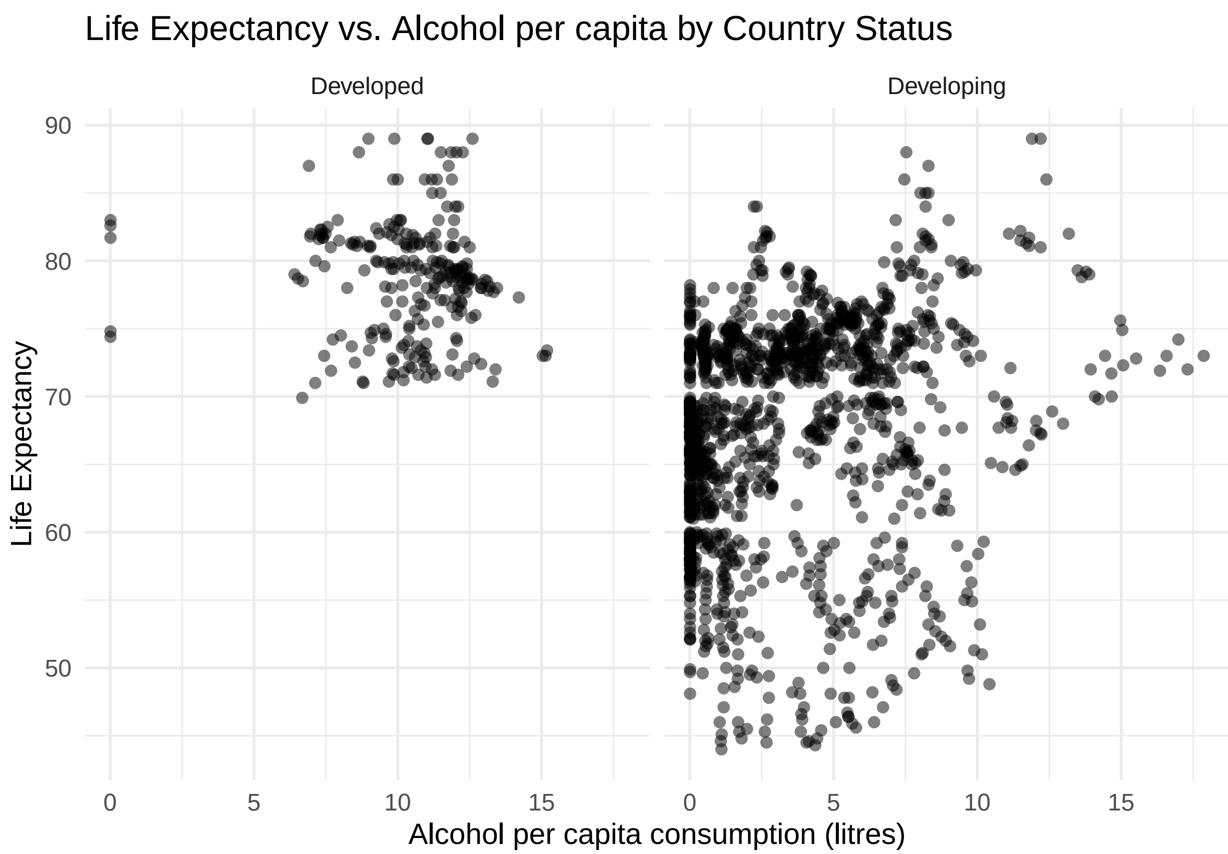
```
  labs(
    title = "Distribution of Life Expectancy",
    y = "Count",
    x = "Life Expectancy (years)"
  ) +
  theme_minimal()
```




```
life_exp_clean |>
  ggplot(aes(x = Status, y = Life_expectancy)) +
  geom_boxplot() +
  labs(
    title = "Distribution of life expectancy by developed vs. developing country",
    y = "life expectancy"
  ) +
  theme_minimal()
```

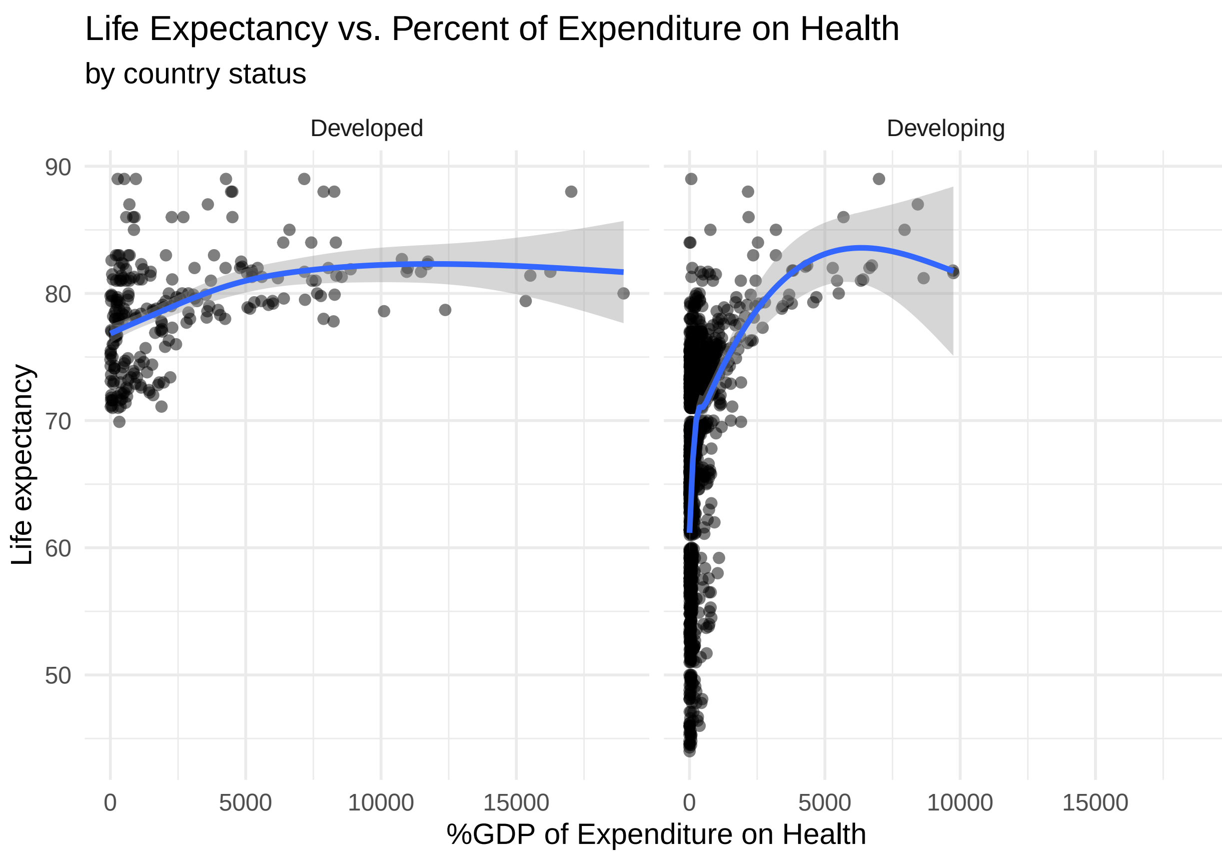



```
life_exp_clean |>
  ggplot(aes(x = Alcohol, y = Life_expectancy)) +
  geom_point(alpha = 0.5) +
  facet_wrap(~Status) +
  labs(
    title = "Life Expectancy vs. Alcohol per capita by Country Status",
    y = "Life Expectancy",
    x = "Alcohol per capita consumption (litres)"
  ) +
  theme_minimal()
```

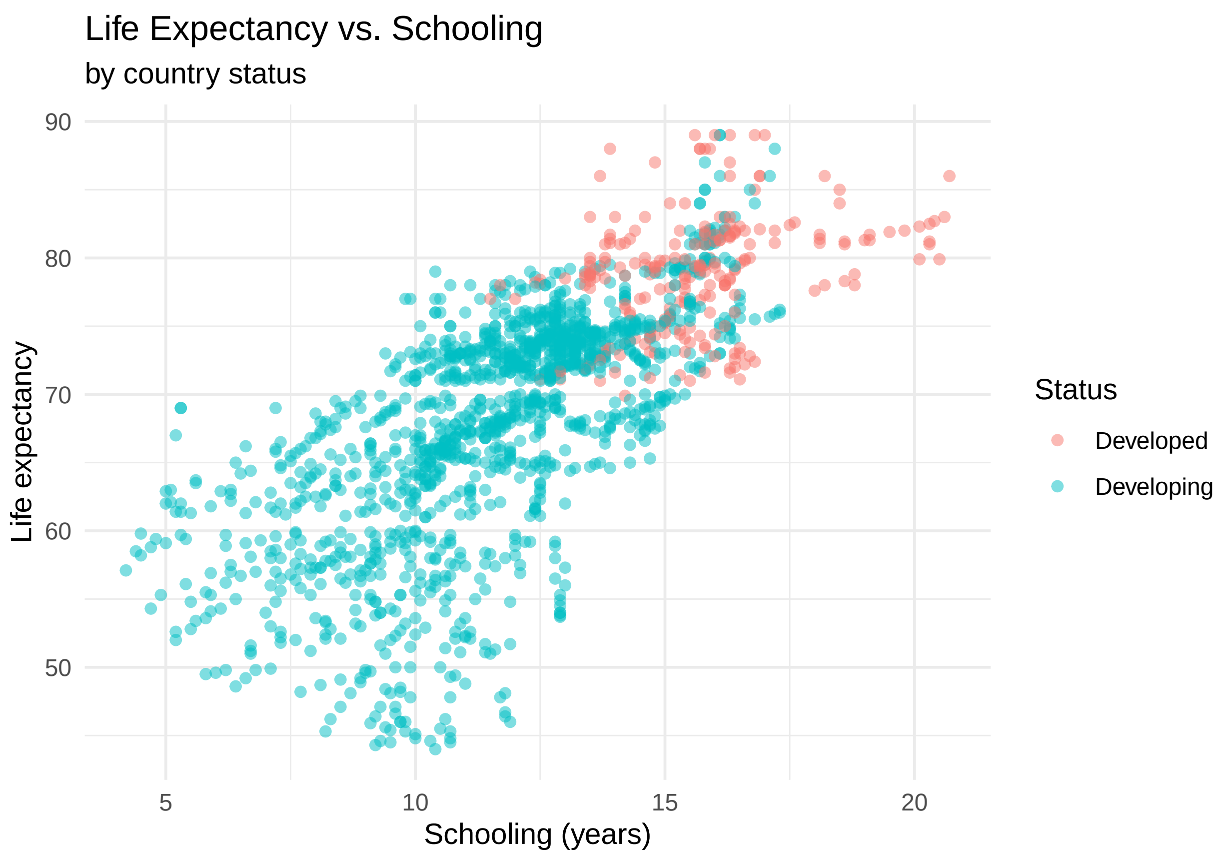



```
life_exp_clean |>
  ggplot(aes(x = percentage_expenditure, y = Life_expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth() +
  facet_wrap(~Status) +
  labs(
    title = "Life Expectancy vs. Percent of Expenditure on Health",
    subtitle = "by country status",
    y = "Life expectancy",
    x = "%GDP of Expenditure on Health"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
life_exp_clean |>
  ggplot(aes(x = Schooling, y = Life_expectancy, color = Status)) +
  geom_point(alpha = 0.5) +
  labs(
    title = "Life Expectancy vs. Schooling",
    subtitle = "by country status",
    y = "Life expectancy",
    x = "Schooling (years)"
  ) +
  theme_minimal()
```



Methodology

```
correlations <- cor(life_exp_clean[, c("Adult_Mortality", "infant_deaths", "Alcohol",
                                       "percentage_expenditure", "Hepatitis_B",
                                       "Measles", "BMI", "under_five_deaths",
                                       "Polio", "Total_expenditure",
                                       "Diphtheria", "HIV/AIDS", "GDP",
                                       "Population","thinness_1_19_years",
                                       "thinness_5_9_years",
                                       "Income_composition_of_resources",
                                       "Schooling")],
                    life_exp_clean$Life_expectancy)

correlations_named <- setNames(correlations,
                               c("Adult_Mortality", "infant_deaths", "Alcohol",
                                       "percentage_expenditure", "Hepatitis_B",
                                       "Measles", "BMI", "under_five_deaths",
                                       "Polio", "Total_expenditure",
                                       "Diphtheria", "HIV/AIDS", "GDP",
                                       "Population","thinness_1_19_years",
                                       "thinness_5_9_years",
                                       "Income_composition_of_resources",
                                       "Schooling"))

correlations_tbl <- enframe(correlations_named,
```

```r
                            name = "Variable",
                            value = "Correlation") |>
  arrange(desc(Correlation))

positive_corr <- correlations_tbl |> filter(Correlation > 0)
```

Correlation Analysis (general)

```
## Warning: Using one column matrices in `filter()` was deprecated in dplyr 1.1.0.
## i Please use one dimensional logical vectors instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
negative_corr <- correlations_tbl |> filter(Correlation < 0)
weak_corr <- correlations_tbl |> filter(abs(Correlation) < 0.1)

cat("Strongest Positive Correlations with Life Expectancy:\n")
```

```
## Strongest Positive Correlations with Life Expectancy:
```

```r
positive_corr
```

```
## # A tibble: 10 x 2
##    Variable                        Correlation[,1]
##    <chr>                                     <dbl>
##  1 Schooling                                 0.728
##  2 Income_composition_of_resources           0.721
##  3 BMI                                       0.542
##  4 GDP                                       0.441
##  5 percentage_expenditure                    0.410
##  6 Alcohol                                   0.403
##  7 Diphtheria                                0.341
##  8 Polio                                     0.327
##  9 Hepatitis_B                               0.200
## 10 Total_expenditure                         0.175
```

```r
cat("\nStrongest Negative Correlations with Life Expectancy:\n")
```

```
##
## Strongest Negative Correlations with Life Expectancy:
```

```r
negative_corr
```

```
## # A tibble: 8 x 2
##   Variable            Correlation[,1]
##   <chr>                         <dbl>
## 1 Population                  -0.0223
## 2 Measles                     -0.0689
## 3 infant_deaths               -0.169
## 4 under_five_deaths           -0.192
## 5 thinness_5_9_years          -0.458
## 6 thinness_1_19_years         -0.458
## 7 HIV/AIDS                    -0.592
## 8 Adult_Mortality             -0.703
```

```r
cat("\nWeak Correlations (Close to 0):\n")
```

```
##
## Weak Correlations (Close to 0):
```

```r
weak_corr
```

```
## # A tibble: 2 x 2
##   Variable   Correlation[,1]
##   <chr>                <dbl>
## 1 Population         -0.0223
## 2 Measles            -0.0689
```

```r
developed <- life_exp_clean |> filter(Status == "Developed")
developing <- life_exp_clean |> filter(Status == "Developing")

correlations_dev <- cor(developed[, c("Adult_Mortality", "infant_deaths",
                                      "Alcohol", "percentage_expenditure",
                                      "Hepatitis_B", "Measles", "BMI",
                                      "under_five_deaths", "Polio",
                                      "Total_expenditure","Diphtheria",
                                      "HIV/AIDS", "GDP", "Population",
                                      "thinness_1_19_years",
                                      "thinness_5_9_years",
                                      "Income_composition_of_resources",
                                      "Schooling")],
                       developed$Life_expectancy)
```

Correlation Analysis by country status

```
## Warning in cor(developed[, c("Adult_Mortality", "infant_deaths", "Alcohol", :
## the standard deviation is zero
```

```r
correlations_devp <- cor(developing[, c("Adult_Mortality", "infant_deaths",
                                      "Alcohol", "percentage_expenditure",
                                      "Hepatitis_B", "Measles", "BMI",
                                      "under_five_deaths", "Polio",
                                      "Total_expenditure","Diphtheria",
                                      "HIV/AIDS", "GDP", "Population",
                                      "thinness_1_19_years",
                                      "thinness_5_9_years",
                                      "Income_composition_of_resources",
                                      "Schooling")],
                         developing$Life_expectancy)

correlations_named_dev <- setNames(correlations_dev,
                                   c("Adult_Mortality", "infant_deaths",
                                      "Alcohol", "percentage_expenditure",
                                      "Hepatitis_B", "Measles", "BMI",
                                      "under_five_deaths", "Polio",
                                      "Total_expenditure","Diphtheria",
                                      "HIV/AIDS", "GDP", "Population",
                                      "thinness_1_19_years",
                                      "thinness_5_9_years",
                                      "Income_composition_of_resources",
                                      "Schooling"))
```

```r
correlations_named_devp <- setNames(correlations_devp,
                                    c("Adult_Mortality", "infant_deaths",
                                      "Alcohol", "percentage_expenditure",
                                      "Hepatitis_B", "Measles", "BMI",
                                      "under_five_deaths", "Polio",
                                      "Total_expenditure","Diphtheria",
                                      "HIV/AIDS", "GDP", "Population",
                                      "thinness_1_19_years",
                                      "thinness_5_9_years",
                                      "Income_composition_of_resources",
                                      "Schooling"))

correlations_tbl_dev <- enframe(correlations_named_dev,
                                name = "Variable",
                                value = "Correlation_Developed") |>
  arrange(desc(Correlation_Developed))

correlations_tbl_devp <- enframe(correlations_named_devp,
                                 name = "Variable",
                                 value = "Correlation_Developing") |>
  arrange(desc(Correlation_Developing))
correlations_tbl_dev
```

```
## # A tibble: 18 x 2
##    Variable                        Correlation_Developed[,1]
##    <chr>                                               <dbl>
##  1 Income_composition_of_resources                    0.721
##  2 percentage_expenditure                             0.392
##  3 GDP                                                0.387
##  4 Schooling                                          0.357
##  5 Total_expenditure                                  0.179
##  6 Population                                         0.123
##  7 Polio                                              0.0598
##  8 BMI                                                0.0108
##  9 Diphtheria                                        -0.0153
## 10 under_five_deaths                                 -0.0316
## 11 Measles                                           -0.0513
## 12 Alcohol                                           -0.0728
## 13 Hepatitis_B                                       -0.0776
## 14 infant_deaths                                     -0.0794
## 15 Adult_Mortality                                   -0.456
## 16 thinness_5_9_years                                -0.717
## 17 thinness_1_19_years                               -0.735
## 18 HIV/AIDS                                          NA
```

```r
correlations_tbl_devp
```

```
## # A tibble: 18 x 2
##    Variable                        Correlation_Developing[,1]
##    <chr>                                                <dbl>
##  1 Schooling                                            0.670
##  2 Income_composition_of_resources                      0.650
##  3 BMI                                                  0.524
##  4 GDP                                                  0.416
```

```
##   5 percentage_expenditure                             0.375
##   6 Diphtheria                                         0.296
##   7 Polio                                              0.278
##   8 Alcohol                                            0.204
##   9 Hepatitis_B                                        0.171
##  10 Total_expenditure                                  0.0972
##  11 Population                                        -0.0104
##  12 Measles                                           -0.0416
##  13 infant_deaths                                     -0.139
##  14 under_five_deaths                                 -0.165
##  15 thinness_5_9_years                                -0.374
##  16 thinness_1_19_years                               -0.374
##  17 HIV/AIDS                                          -0.615
##  18 Adult_Mortality                                   -0.681
```

```r
correlation_comparison <- left_join(correlations_tbl_dev, correlations_tbl_devp,
                                    by = "Variable")



cat("Comparison of Correlations with Life Expectancy:
    Developed vs Developing Countries")
```

```
## Comparison of Correlations with Life Expectancy:
##     Developed vs Developing Countries
```

```r
correlation_comparison
```

```
## # A tibble: 18 x 3
##    Variable                       Correlation_Develope~1 Correlation_Developi~2
##    <chr>                                           <dbl>                  <dbl>
##  1 Income_composition_of_resources                0.721                  0.650
##  2 percentage_expenditure                         0.392                  0.375
##  3 GDP                                            0.387                  0.416
##  4 Schooling                                      0.357                  0.670
##  5 Total_expenditure                              0.179                  0.0972
##  6 Population                                     0.123                 -0.0104
##  7 Polio                                          0.0598                 0.278
##  8 BMI                                            0.0108                 0.524
##  9 Diphtheria                                    -0.0153                 0.296
## 10 under_five_deaths                             -0.0316                -0.165
## 11 Measles                                       -0.0513                -0.0416
## 12 Alcohol                                       -0.0728                 0.204
## 13 Hepatitis_B                                   -0.0776                 0.171
## 14 infant_deaths                                 -0.0794                -0.139
## 15 Adult_Mortality                               -0.456                 -0.681
## 16 thinness_5_9_years                            -0.717                 -0.374
## 17 thinness_1_19_years                           -0.735                 -0.374
## 18 HIV/AIDS                                      NA                     -0.615
## # i abbreviated names: 1: Correlation_Developed[,1],
## #   2: Correlation_Developing[,1]
```

For developed countries, income, percentage expenditure, and GDP showed positive, strong correlation with life expectancy. Polio, BMI, diphtheria and under-five-deaths all had a weak correlation with life expectancy with Pearson coefficients ~0.

in contrast, developing countries had a strong, positive correlation between life expectancy and schooling, income composition index, BMI, GDP, and percentage expenditure, showing that for developing countries, BMI is still an important factor that plays a role in life expectancy, but that this correlation strength decreases as a country becomes a developed country. Percentage of government expenditure on health had a weak correlation with life expectancy in developing countries, likely due to the fact that other factors (poverty, malnutrition, sanitation, etc.) may have a stronger impact on life expectancy, healthcare expenditure may not be evenly distributed across the country, and current government focus is on addressing infectious diseases/emergency healthcare.

Interestingly, total-expenditure had a week correlation in developing countries, and various diseases and alcohol had a mildly strong correlation of around 0.2. Increased years in schooling were strongly associated with higher life expectancy in developing countries, but only moderately so in developed countries.

```r
#developed
model_developed <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                        Adult_Mortality + infant_deaths + Polio +
                        Total_expenditure + Diphtheria +
                        Income_composition_of_resources, data = developed)

#developing
model_developing <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                        Adult_Mortality + infant_deaths + Polio +
                        Total_expenditure + Diphtheria +
                        Income_composition_of_resources,
                      data = developing)

summary(model_developed)
```

Linear Regression Models by country status

```
##
## Call:
## lm(formula = Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
##     Adult_Mortality + infant_deaths + Polio + Total_expenditure +
##     Diphtheria + Income_composition_of_resources, data = developed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0533 -1.5420 -0.6574  1.1440 10.4802
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.475e+01  4.858e+00   7.153 1.12e-11 ***
## GDP                             -3.296e-06  1.009e-05  -0.327 0.744196
## Schooling                       -5.216e-01  1.507e-01  -3.461 0.000641 ***
## BMI                             -1.393e-03  1.042e-02  -0.134 0.893774
## Alcohol                         -1.391e-01  8.282e-02  -1.679 0.094442 .
## Adult_Mortality                 -1.161e-02  3.782e-03  -3.071 0.002390 **
## infant_deaths                    1.554e-01  1.807e-01   0.860 0.390648
## Polio                            1.821e-02  2.858e-02   0.637 0.524741
## Total_expenditure                1.263e-01  7.479e-02   1.688 0.092740 .
## Diphtheria                      -2.852e-02  2.826e-02  -1.009 0.313881
## Income_composition_of_resources  6.529e+01  6.056e+00  10.781  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.824 on 231 degrees of freedom
## Multiple R-squared:  0.5814, Adjusted R-squared:  0.5633
## F-statistic: 32.08 on 10 and 231 DF,  p-value: < 2.2e-16
```

```r
summary(model_developing)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
##     Adult_Mortality + infant_deaths + Polio + Total_expenditure +
##     Diphtheria + Income_composition_of_resources, data = developing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6381  -2.1660   0.4305   2.8267  12.2660
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.283e+01  8.646e-01  61.104  < 2e-16 ***
## GDP                             1.270e-04  2.317e-05   5.480 5.04e-08 ***
## Schooling                       9.387e-01  8.022e-02  11.701  < 2e-16 ***
## BMI                             5.028e-02  7.660e-03   6.564 7.35e-11 ***
## Alcohol                        -2.441e-01  4.225e-02  -5.778 9.30e-09 ***
## Adult_Mortality                -2.882e-02  1.032e-03 -27.939  < 2e-16 ***
## infant_deaths                  -1.249e-03  9.477e-04  -1.318 0.187708
## Polio                           6.193e-03  6.277e-03   0.987 0.324001
## Total_expenditure              -3.465e-02  5.535e-02  -0.626 0.531433
## Diphtheria                      2.167e-02  6.570e-03   3.298 0.000998 ***
## Income_composition_of_resources 9.911e+00  1.030e+00   9.624  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.392 on 1396 degrees of freedom
## Multiple R-squared:  0.7256, Adjusted R-squared:  0.7236
## F-statistic: 369.1 on 10 and 1396 DF,  p-value: < 2.2e-16
```

```r
model_dev_int <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                      Adult_Mortality + infant_deaths + Polio +
                      Total_expenditure + Diphtheria +
                      Income_composition_of_resources + GDP*Schooling,
                      data = developed)
model_dvl_int <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                      Adult_Mortality + infant_deaths + Polio +
                      Total_expenditure + Diphtheria +
                      Income_composition_of_resources + GDP*Schooling,
                      data = developing)
anova(model_developed, model_dev_int)
```

Testing Interaction Effects of Interest with Drop-in Deviance Tests

```
## Analysis of Variance Table
##
```

```
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + GDP * Schooling
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    231 1842.5
## 2    230 1842.5  1 0.0069196 9e-04 0.9766
```

```r
anova(model_developing, model_dvl_int)
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + GDP * Schooling
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   1396 26929
## 2   1395 26858  1    70.807 3.6777 0.05535 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction effect of GDP and Schooling does not significantly improve model fit in either model; pval > 0.05.

```r
model_dev_int2 <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                       Adult_Mortality + infant_deaths + Polio +
                       Total_expenditure + Diphtheria +
                       Income_composition_of_resources +
                       Income_composition_of_resources*Schooling,
                     data = developed)
model_dvl_int2 <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                       Adult_Mortality + infant_deaths + Polio +
                       Total_expenditure + Diphtheria +
                       Income_composition_of_resources +
                       Income_composition_of_resources*Schooling,
                     data = developing)
anova(model_developed, model_dev_int2)
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + Income_composition_of_resources *
##     Schooling
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    231 1842.5
## 2    230 1837.7  1    4.8052 0.6014 0.4388
```

```r
anova(model_developing, model_dvl_int2)
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + Income_composition_of_resources *
##     Schooling
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   1396 26929
## 2   1395 26878  1     50.84 2.6387 0.1045
```

Interaction effect of Income Composition Index and Schooling does not significantly improve model fit in either model; pval > 0.05.

Interested in lifestyle:

```r
model_dev_int3 <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                       Adult_Mortality + infant_deaths + Polio +
                       Total_expenditure + Diphtheria +
                       Income_composition_of_resources + BMI*Alcohol,
                     data = developed)
model_dvl_int3 <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                       Adult_Mortality + infant_deaths + Polio +
                       Total_expenditure + Diphtheria +
                       Income_composition_of_resources + BMI*Alcohol,
                     data = developing)
anova(model_developed, model_dev_int3)
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + BMI * Alcohol
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    231 1842.5
## 2    230 1841.8  1   0.71231 0.089 0.7658
```

```r
anova(model_developing, model_dvl_int3)
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + BMI * Alcohol
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   1396 26929
```

```
## 2   1395 26826  1    102.46 5.3279 0.02113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model_developing)$r.squared
```

```
## [1] 0.7255623
```

```r
summary(model_dvl_int3)$r.squared
```

```
## [1] 0.7266065
```

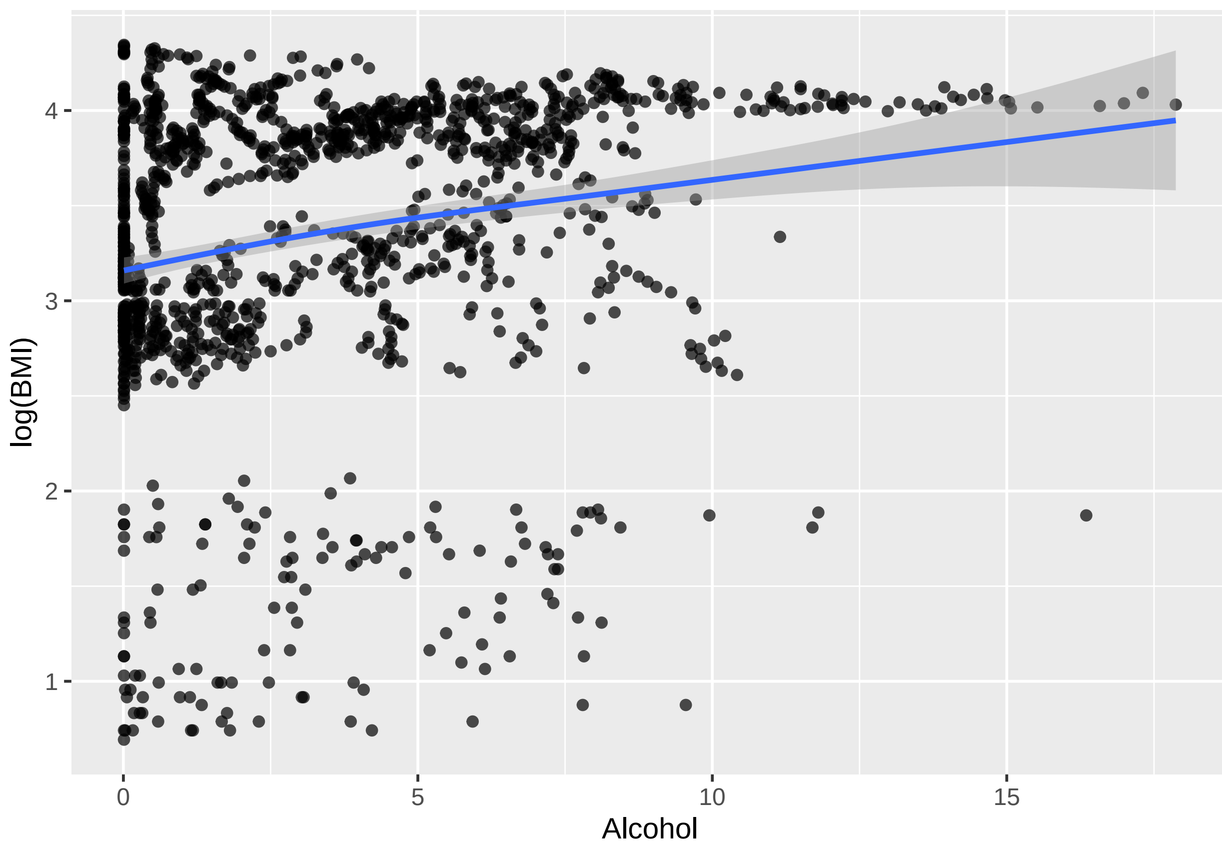
Interaction effect of BMI and Alcohol does not significantly improve model fit in developed countries, but it is significant in developing countries with a p-value of 0.02, less than the threshold of 0.05.

- This suggests that the relationship between **alcohol consumption** and **BMI** in relation to **life expectancy** might differ between developed and developing countries.
- R^2 increased minutely, indicating better model fit
- Countries with lower avg BMI experience less negative impact of alcohol consumption on life expectancy, whereas those with higher BMI may experience a greater reduction in life expectancy at similar levels of alcohol consumption.

Further exploration of interaction effect:

```r
developing |>
  ggplot(aes(x = Alcohol, y = log(BMI))) +
  geom_point(alpha = 0.7) +
  geom_smooth()
```

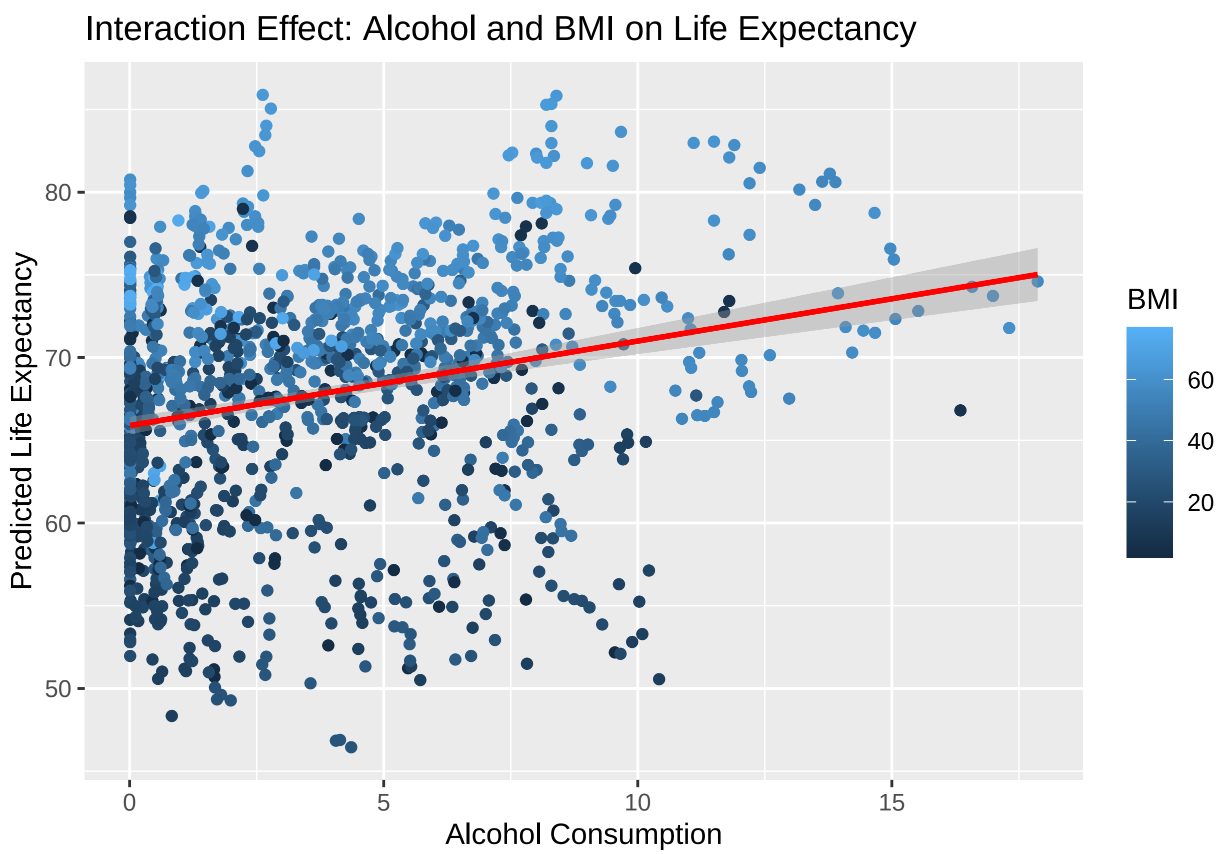
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
developing$predicted_life_expectancy <- predict(model_dvl_int3,
                                                newdata = developing)

ggplot(developing, aes(x = Alcohol, y = predicted_life_expectancy, color = BMI)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Interaction Effect: Alcohol and BMI on Life Expectancy",
       x = "Alcohol Consumption", y = "Predicted Life Expectancy")
```

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




```
model_dev_int4 <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                       Adult_Mortality + infant_deaths + Polio +
                       Total_expenditure + Diphtheria +
                       Income_composition_of_resources + Total_expenditure*Polio,
                     data = developed)
model_dvl_int4 <- lm(Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
                       Adult_Mortality + infant_deaths + Polio +
                       Total_expenditure + Diphtheria +
                       Income_composition_of_resources + Total_expenditure*Polio,
                     data = developing)
anova(model_developed, model_dev_int4)
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + Total_expenditure * Polio
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    231 1842.5
## 2    230 1838.9  1    3.5893 0.4489 0.5035
```

```
anova(model_developing, model_dvl_int4)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources
## Model 2: Life_expectancy ~ GDP + Schooling + BMI + Alcohol + Adult_Mortality +
##     infant_deaths + Polio + Total_expenditure + Diphtheria +
##     Income_composition_of_resources + Total_expenditure * Polio
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   1396 26929
## 2   1395 26926  1    2.7572 0.1428 0.7055
```

Interaction effect of total expenditure and Polio does not significantly improve model fit in either model; pval > 0.05.

```r
# install.packages("car")
library(car)
```

Assessing Final Model (including interaction effect BMI*Alcohol)

```
## 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
```

```r
vif(model_dvl_int3, type = "predictor")
```

```
## GVIFs computed for predictors

##                                     GVIF Df GVIF^(1/(2*Df)) Interacts With
## GDP                             1.419009  1        1.191222             --
## Schooling                       2.928210  1        1.711201             --
## BMI                             2.157746  3        1.136755        Alcohol
## Alcohol                         2.157746  3        1.136755            BMI
## Adult_Mortality                 1.276873  1        1.129988             --
## infant_deaths                   1.107361  1        1.052312             --
## Polio                           1.602373  1        1.265849             --
## Total_expenditure               1.066298  1        1.032617             --
## Diphtheria                      1.620819  1        1.273114             --
## Income_composition_of_resources 2.348231  1        1.532394             --
##
## GDP                             Schooling, BMI, Alcohol, Adult_Mortality, infant_deaths, Polio, Tota
## Schooling                             GDP, BMI, Alcohol, Adult_Mortality, infant_deaths, Polio, Tota
## BMI                                      GDP, Schooling, Adult_Mortality, infant_deaths, Polio, Tota
## Alcohol                                  GDP, Schooling, Adult_Mortality, infant_deaths, Polio, Tota
## Adult_Mortality                             GDP, Schooling, BMI, Alcohol, infant_deaths, Polio, Tota
## infant_deaths                             GDP, Schooling, BMI, Alcohol, Adult_Mortality, Polio, Tota
## Polio                             GDP, Schooling, BMI, Alcohol, Adult_Mortality, infant_deaths, Tota
## Total_expenditure                               GDP, Schooling, BMI, Alcohol, Adult_Mortality, infant_
## Diphtheria                              GDP, Schooling, BMI, Alcohol, Adult_Mortality, infant_deaths,
```

```
## Income_composition_of_resources                          GDP, Schooling, BMI, Alcohol, Adult_Morta
```

```
model_dvl_int3
```

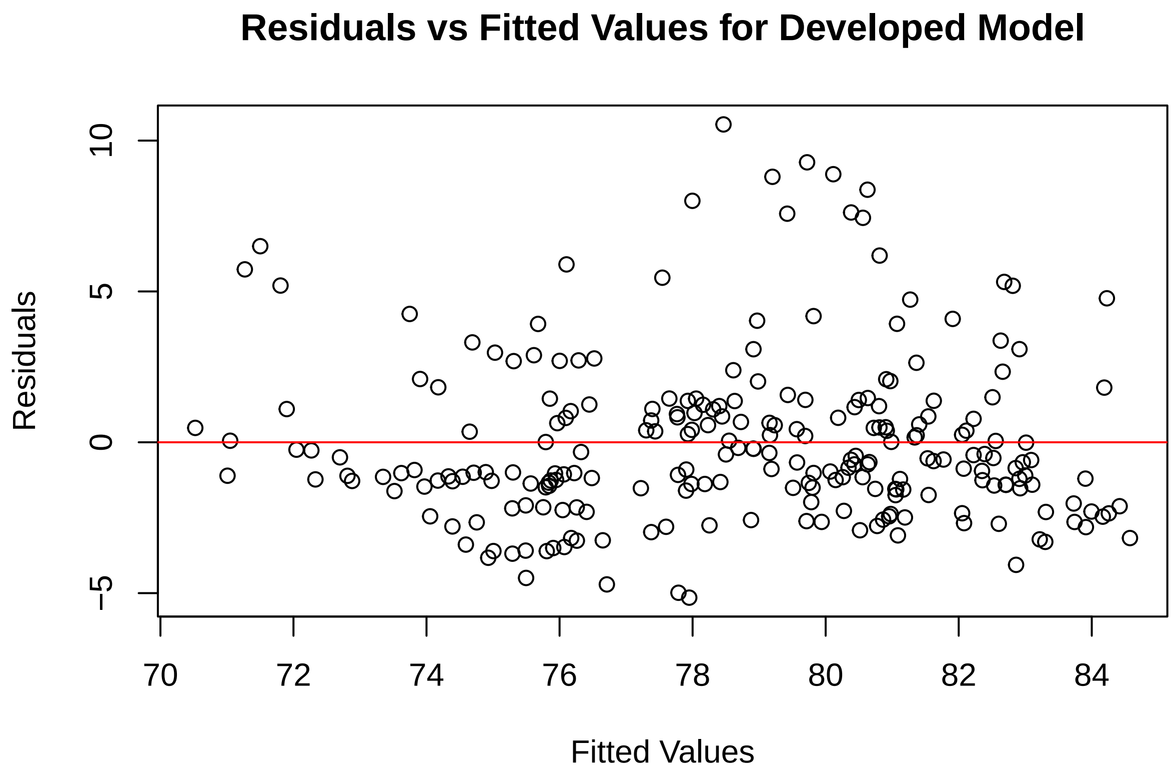
```
##
## Call:
## lm(formula = Life_expectancy ~ GDP + Schooling + BMI + Alcohol +
##     Adult_Mortality + infant_deaths + Polio + Total_expenditure +
##     Diphtheria + Income_composition_of_resources + BMI * Alcohol,
##     data = developing)
##
## Coefficients:
##                     (Intercept)                              GDP
##                      53.2265020                        0.0001172
##                       Schooling                              BMI
##                       0.9494894                        0.0357676
##                         Alcohol                  Adult_Mortality
##                      -0.4191192                       -0.0288127
##                   infant_deaths                            Polio
##                      -0.0012378                        0.0063658
##               Total_expenditure                       Diphtheria
##                      -0.0297684                        0.0212777
## Income_composition_of_resources                      BMI:Alcohol
##                       9.9368478                        0.0043319
```

None have GVIF > 10, so multicollinearity is not an issue with this model

Checking Model Assumptions Constant Variance assumption satisfied.

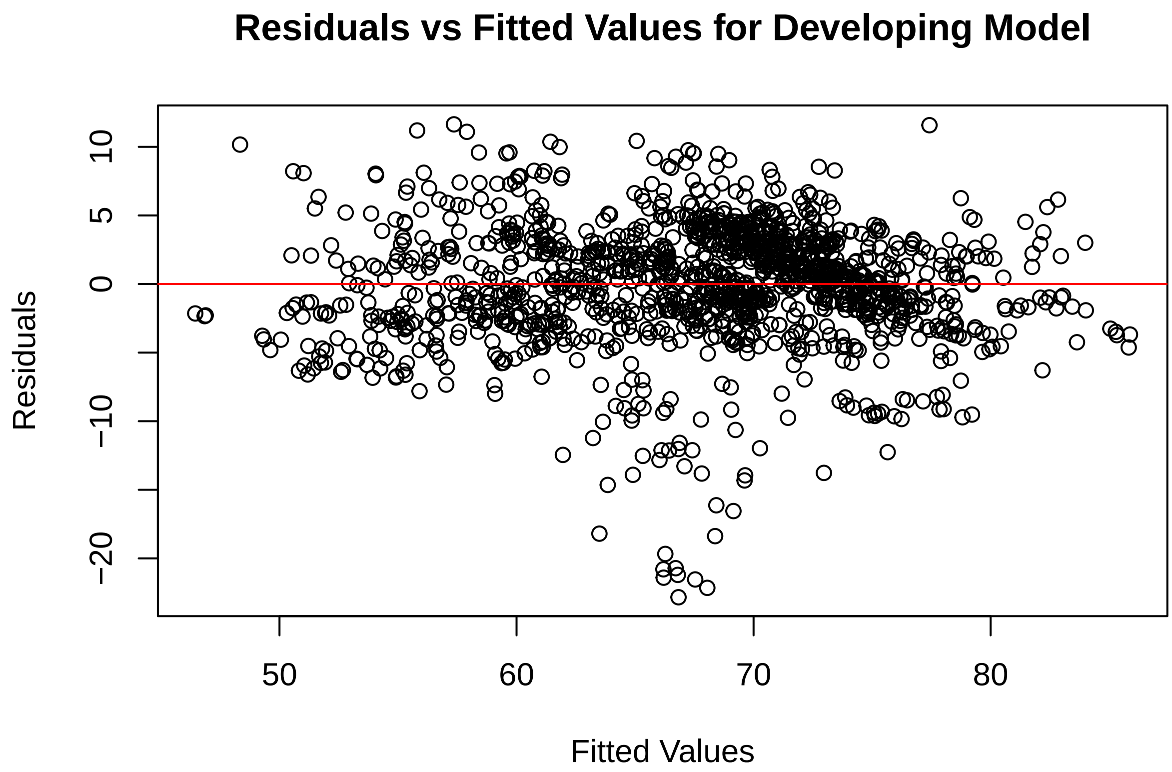
```
residuals_dev <- residuals(model_dev_int3)
fitted_values_dev <- fitted(model_dev_int3)


plot(fitted_values_dev, residuals_dev,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values for Developed Model")
abline(h = 0, col = "red")
```



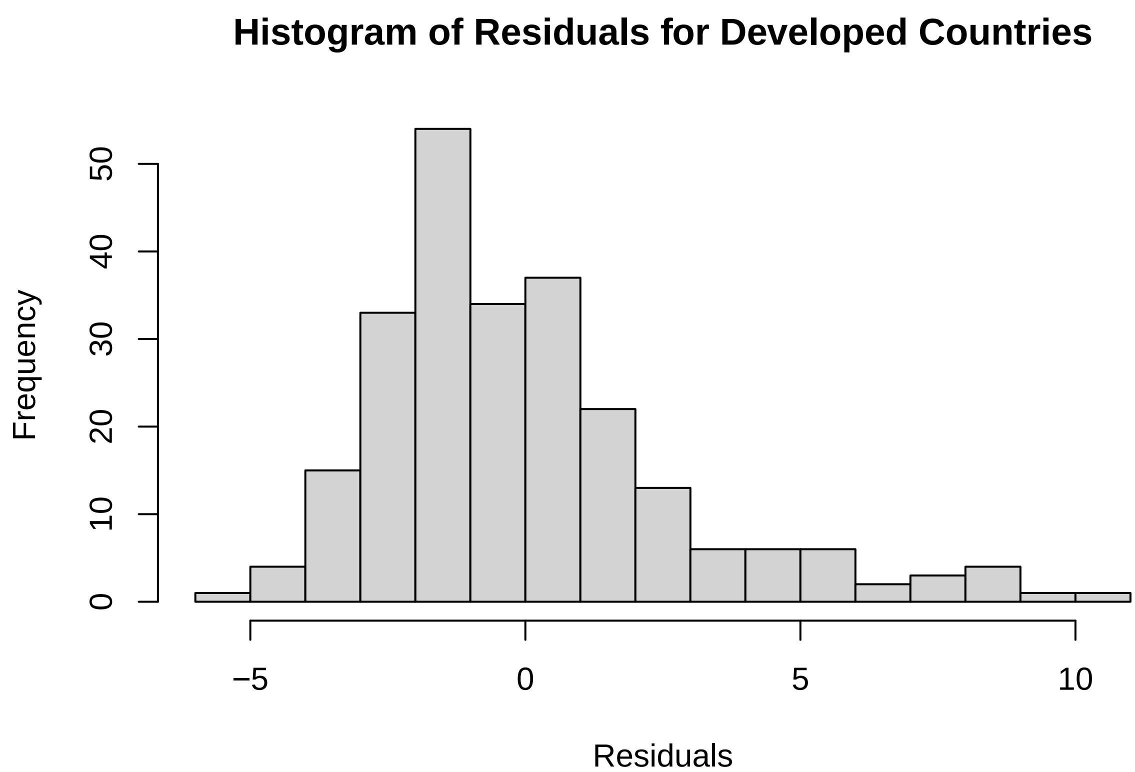

```r
residuals_dvl <- residuals(model_dvl_int3)
fitted_values_dvl <- fitted(model_dvl_int3)


plot(fitted_values_dvl, residuals_dvl,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values for Developing Model")
abline(h = 0, col = "red")
```

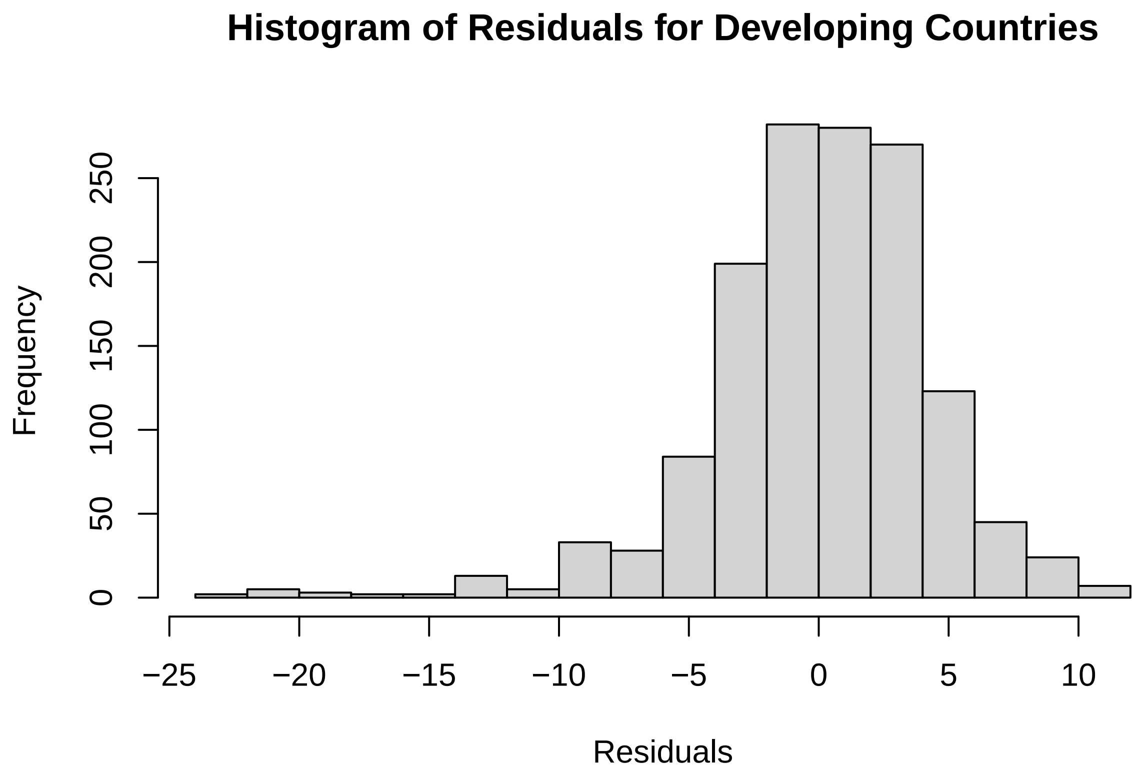


Normality Assumption satisfied; residuals are normally distributed.

```
hist(residuals_dev,
     main = "Histogram of Residuals for Developed Countries",
     xlab = "Residuals",
     breaks = 20)
```




```r
hist(residuals_dvl,
     main = "Histogram of Residuals for Developing Countries",
     xlab = "Residuals",
     breaks = 20)
```



Conclusion

Emphasis should be placed on:

1. **Education Programs:** Better education results in improved public health awareness, lifestyle choices, and access to resources.

- Investments in education systems with a rural/underserved focus
- Teacher training programs
- Collaboration w/ national education ministries
- Health expenditure monitoring program to fairly distribute resources

2. **Economic Development:** Investments aimed at economic growth and resource allocation would significantly positively impact life expectancy.

- Develop Global Income Composition Index and GDP Goals that countries have incentive to meet
- Promote health-sensitive economic policies, health financing models
- Support mobilization of resources through taxation policy, international health funding

3. **Health System Infrastructure:** Strengthening weak correlations between total health expenditure and life expectancy

- Make sure health spending directly benefits the population through efficient healthcare delivery
- Address inequalities in health system and target vulnerable populations
- Supporting disease prevention programs

- Increasing vaccination support and infectious disease research for diphtheria, polio, etc.

Limitations

Independence of observations is not met, since there could be interdependence between countries due to geographic proximity, trade relations, shared economic conditions, or regional policies. However, I proceeded with the analysis because:

- Model was built with control for observable characteristics, or relevant covariates that account for country-specific differences (GDP, population, etc)., reducing risk of bias due to country-specific interdependencies
- Time period of interest is relatively short, spanning from 2000 to 2015, so temporal autocorrelation (correlation across years within a country) is limited

Additional limitations include omission of NAs from dataset, which may lead to differences in results.

Future Work

- **Inclusion of fixed effects** (year-level) in linear regression to control for time-invariant characteristics that could affect the relationship between the variables
- Comparison between results with NAs and without NAs
- **Time-Series Analysis** w/ ARIMA to determine if there are any trends over time