Analysis on Life Expectancy Report Paper

Krissa Joy Tabornal & Jan Ira Villacarlos

6/1/2021

Introduction

Life expectancy is the estimated hypothetical average number of years that a person of a certain age is expected to live (Bezy, 2020). The measure is influenced by numerous factors and differs by sex, age, race, and geographic location. It reflects many local conditions, such as cultural lifestyles, urbanization, healthcare access, and economical factors. In this capstone project, life expectancy data from 2000 to 2015 will be analyzed. The main goal of this project is to determine the predicting factors that significantly affect life expectancy. The analysis will focus on immunization factors, mortality factors, economic factors, social factors and health factors. This can help in gaining insight about specific factors that should be given more importance to effectively improve the life expectancy of the population.

The project aims to answer the following questions: 1. What are the predicting factors that actually affect life expectancy? Has life expectancy improved over the years? Is there a difference between the life expectancy of developed countries and developing countries? 2. How does healthcare expenditure influence average lifespan? 3. How does schooling or education affect life expectancy? 4. How does immunization affect life expectancy? What kind of immunizations have a greater influence? 5. How do economic factors affect life expectancy?

Data

The 'Life Expectancy' dataset was posted by KumarRajarshi in Kaggle and can be accessed at https://www.kaggle.com/kumarajarshi/life-expectancy-who. This data was collected from the Global Health Observatory (GHO), a data repository, owned by the World Health Organization (WHO) and United Nations (UN), which records data about health status and other factors for all countries. The data includes a set of predictor variables that might have an impact on life expectancy in different countries. The variables included in the data set are the following: country, year, status, life expectancy, 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, and schooling. These variables were classified into immunization related, mortality, health, economic, and social factors. Overall, the data set contains 22 variables and 2938 observations containing data from 2000-2015 for 193 countries. ### Loading Packages

```
#packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package ’ggplot2’ was built under R version 4.0.5

## Warning: package ’tidyr’ was built under R version 4.0.5

## Warning: package ’readr’ was built under R version 4.0.5

## Warning: package ’forcats’ was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(ggplot2)
library(ggpubr)
```

```
## Warning: package ’ggpubr’ was built under R version 4.0.5
```

```
library(ggthemes)
```

```
## Warning: package ’ggthemes’ was built under R version 4.0.5
```

```
library(GGally)
```

```
## Warning: package ’GGally’ was built under R version 4.0.5

## Registered S3 method overwritten by ’GGally’:
##   method from
##   +.gg   ggplot2
```

```
library(reshape2)
```

```
## Warning: package ’reshape2’ was built under R version 4.0.5

##
## Attaching package: ’reshape2’

## The following object is masked from ’package:tidyr’:
##
##     smiths
```

Checking data set structure

```r
# importing data set
life_expectancy <- read_csv("Life Expectancy Data.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   Country = col_character(),
##   Status = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```r
# checking data set
head(life_expectancy)
```

```
## # A tibble: 6 x 22
##   Country  Year Status `Life expectanc~ `Adult Mortalit~ `infant deaths` Alcohol
##   <chr>   <dbl> <chr>             <dbl>            <dbl>           <dbl>   <dbl>
## 1 Afghan~  2015 Devel~             65                263              62    0.01
## 2 Afghan~  2014 Devel~             59.9              271              64    0.01
## 3 Afghan~  2013 Devel~             59.9              268              66    0.01
## 4 Afghan~  2012 Devel~             59.5              272              69    0.01
## 5 Afghan~  2011 Devel~             59.2              275              71    0.01
## 6 Afghan~  2010 Devel~             58.8              279              74    0.01
## # ... with 15 more variables: percentage expenditure <dbl>, Hepatitis B <dbl>,
## #   Measles <dbl>, BMI <dbl>, under-five deaths <dbl>, Polio <dbl>,
## #   Total expenditure <dbl>, Diphtheria <dbl>, HIV/AIDS <dbl>, GDP <dbl>,
## #   Population <dbl>, thinness  1-19 years <dbl>, thinness 5-9 years <dbl>,
## #   Income composition of resources <dbl>, Schooling <dbl>
```

```r
names(life_expectancy)
```

```
##  [1] "Country"                         "Year"
##  [3] "Status"                          "Life expectancy"
##  [5] "Adult Mortality"                 "infant deaths"
##  [7] "Alcohol"                         "percentage expenditure"
##  [9] "Hepatitis B"                     "Measles"
## [11] "BMI"                             "under-five deaths"
## [13] "Polio"                           "Total expenditure"
## [15] "Diphtheria"                      "HIV/AIDS"
## [17] "GDP"                             "Population"
## [19] "thinness  1-19 years"            "thinness 5-9 years"
## [21] "Income composition of resources" "Schooling"
```

```r
str(life_expectancy)
```

```
## spec_tbl_df [2,938 x 22] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Country                        : chr [1:2938] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanis
##  $ Year                           : num [1:2938] 2015 2014 2013 2012 2011 ...
##  $ Status                         : chr [1:2938] "Developing" "Developing" "Developing" "Developing"
##  $ Life expectancy                : num [1:2938] 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult Mortality                : num [1:2938] 263 271 268 272 275 279 281 287 295 295 ...
##  $ infant deaths                  : num [1:2938] 62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num [1:2938] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 .
##  $ percentage expenditure         : num [1:2938] 71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis B                    : num [1:2938] 65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : num [1:2938] 1154 492 430 2787 3013 ...
##  $ BMI                            : num [1:2938] 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 .
##  $ under-five deaths              : num [1:2938] 83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : num [1:2938] 6 58 62 67 68 66 63 64 63 58 ...
##  $ Total expenditure              : num [1:2938] 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ..
##  $ Diphtheria                     : num [1:2938] 65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV/AIDS                       : num [1:2938] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num [1:2938] 584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num [1:2938] 33736494 327582 31731688 3696958 2978599 ...
##  $ thinness  1-19 years           : num [1:2938] 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness 5-9 years             : num [1:2938] 17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income composition of resources: num [1:2938] 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.41
##  $ Schooling                      : num [1:2938] 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Country = col_character(),
##   ..   Year = col_double(),
##   ..   Status = col_character(),
##   ..   ‘Life expectancy‘ = col_double(),
##   ..   ‘Adult Mortality‘ = col_double(),
##   ..   ‘infant deaths‘ = col_double(),
##   ..   Alcohol = col_double(),
##   ..   ‘percentage expenditure‘ = col_double(),
##   ..   ‘Hepatitis B‘ = col_double(),
##   ..   Measles = col_double(),
##   ..   BMI = col_double(),
##   ..   ‘under-five deaths‘ = col_double(),
##   ..   Polio = col_double(),
##   ..   ‘Total expenditure‘ = col_double(),
##   ..   Diphtheria = col_double(),
##   ..   ‘HIV/AIDS‘ = col_double(),
##   ..   GDP = col_double(),
##   ..   Population = col_double(),
##   ..   ‘thinness  1-19 years‘ = col_double(),
##   ..   ‘thinness 5-9 years‘ = col_double(),
##   ..   ‘Income composition of resources‘ = col_double(),
##   ..   Schooling = col_double()
##   .. )
```

```
glimpse(life_expectancy)
```

```
## Rows: 2,938
## Columns: 22
## $ Country                           <chr> "Afghanistan", "Afghanistan", "Afgha~
## $ Year                              <dbl> 2015, 2014, 2013, 2012, 2011, 2010, ~
## $ Status                            <chr> "Developing", "Developing", "Develop~
## $ ‘Life expectancy‘                 <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, ~
## $ ‘Adult Mortality‘                 <dbl> 263, 271, 268, 272, 275, 279, 281, 2~
## $ ‘infant deaths‘                 <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, ~
## $ Alcohol                         <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, ~
## $ ‘percentage expenditure‘        <dbl> 71.279624, 73.523582, 73.219243, 78.~
## $ ‘Hepatitis B‘                   <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, ~
## $ Measles                         <dbl> 1154, 492, 430, 2787, 3013, 1989, 28~
## $ BMI                             <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, ~
## $ ‘under-five deaths‘             <dbl> 83, 86, 89, 93, 97, 102, 106, 110, 1~
## $ Polio                           <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 5~
## $ ‘Total expenditure‘             <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, ~
## $ Diphtheria                      <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, ~
## $ ‘HIV/AIDS‘                      <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0~
## $ GDP                             <dbl> 584.25921, 612.69651, 631.74498, 669~
## $ Population                      <dbl> 33736494, 327582, 31731688, 3696958,~
## $ ‘thinness  1-19 years‘          <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, ~
## $ ‘thinness 5-9 years‘            <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, ~
## $ ‘Income composition of resources‘ <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0~
## $ Schooling                       <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9,~
```

```
dim(life_expectancy)
```

```
## [1] 2938   22
```

Data Analysis

Life expectancy was the main outcome of this analysis. The aforementioned predictor variables were used to answer the research questions. The tidyverse was the main package used in the analysis. First, the data was first examined using head(), names(), str(), dim(), and summary(). The summary results showed that there were missing values in some variables. The developing and developed countries were also counted from 2000-2015. Preliminary visualization using geom_count() was performed to get a glimpse of the range of life expectancy for each year and differentiate data values from developed and developing countries. Correlation test was also conducted using cor() to know which variables are good candidates in predicting life expectancy.

```
#check range of values and to see number of NAs
summary(life_expectancy)
```

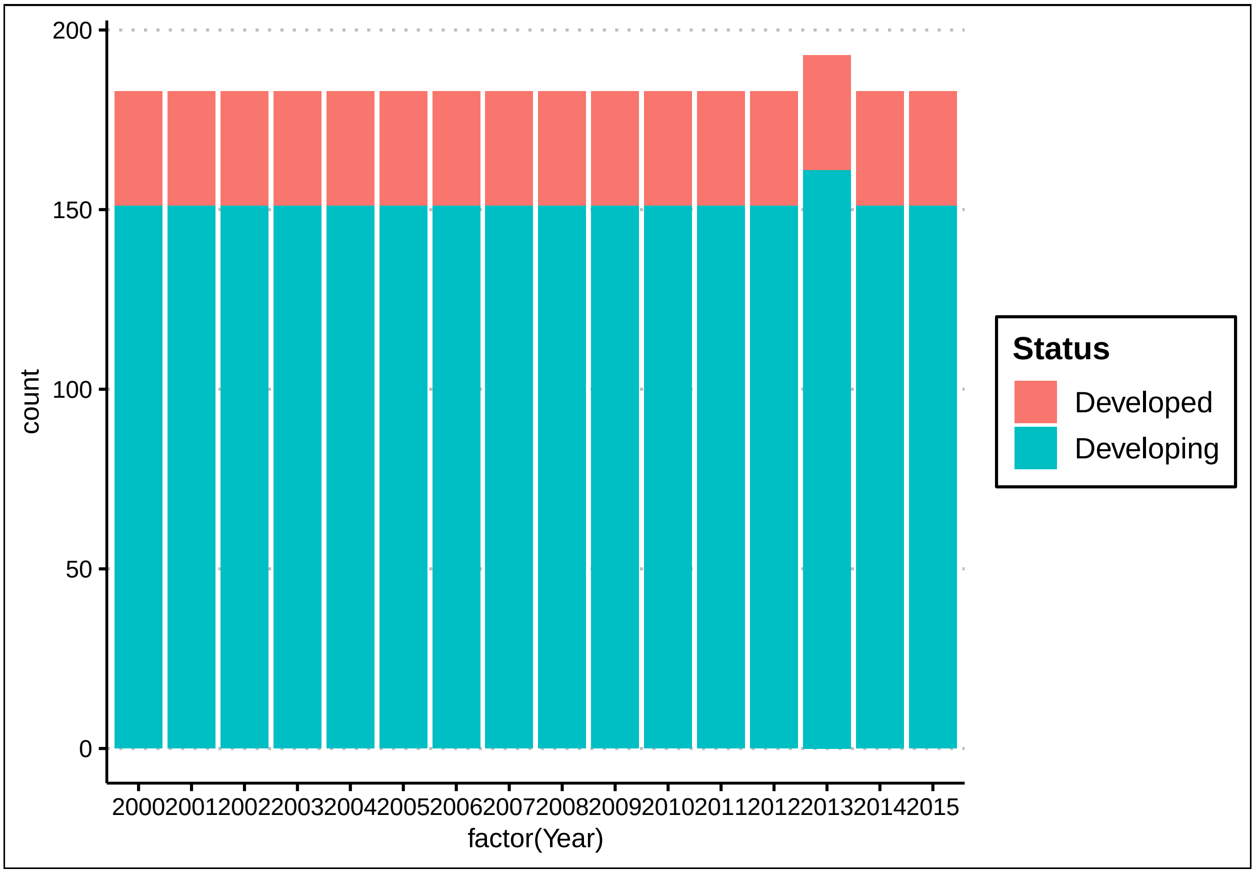
Checking general visualizations of the data

```
##    Country               Year         Status          Life expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult Mortality infant deaths       Alcohol        percentage expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis B       Measles              BMI        under-five deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total expenditure   Diphtheria       HIV/AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness  1-19 years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness 5-9 years Income composition of resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

```
#count of developed and developing countries
status <- life_expectancy %>%
  select(Year, Status) %>%
  group_by(Status, Year) %>%
  summarise(count = n())
```

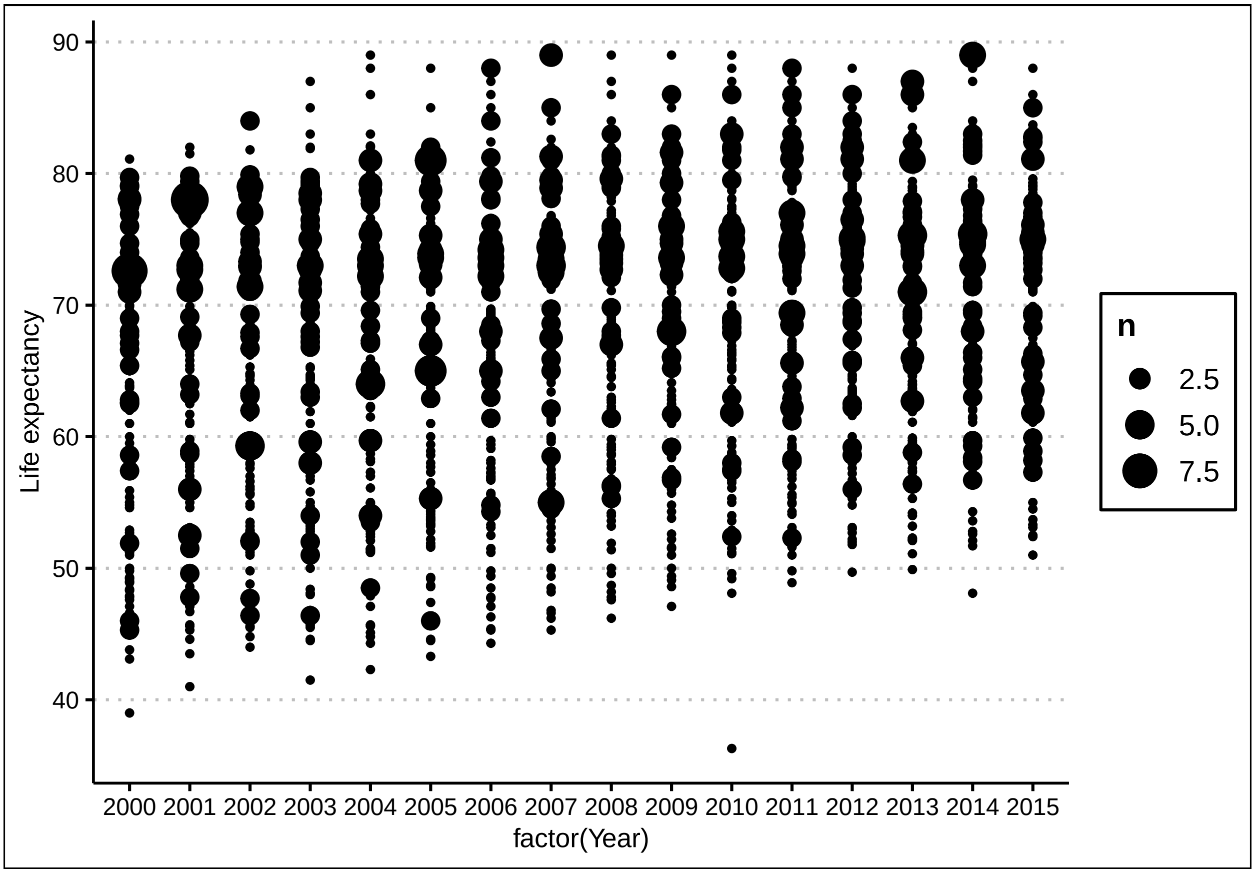
```
## `summarise()` has grouped output by 'Status'. You can override using the `.groups` argument.
```

```
#the status of countries per year did not change as much over the years

status <- life_expectancy %>%
  select(Year, Status) %>%
  group_by(Status, Year)

ggplot(data = status, aes(x = factor(Year), fill = Status)) + geom_bar() + theme_clean()
```



```
#life expectancy per year
ggplot(data = life_expectancy, aes(y = `Life expectancy`, x = factor(Year))) + geom_count() + theme_clea
```

```
## Warning: Removed 10 rows containing non-finite values (stat_sum).
```



```
#we can see that over the years countries are moving towards longer life expectancy, very few points le

#check correlations between variables
shapiro.test(life_expectancy$`Life expectancy`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  life_expectancy$`Life expectancy`
## W = 0.95605, p-value < 2.2e-16
```

```
#nonnormal so must perform nonparametric correlation tests

#check possible correlations
life_cor <- life_expectancy %>%
              select(-Status, -Country, -Year)
life_cor <- cor(life_cor, use = "na.or.complete", method = "spearman")
life_cor
```

```
##                                  Life expectancy Adult Mortality infant deaths
## Life expectancy                       1.00000000      -0.6556208    -0.5214551
## Adult Mortality                      -0.65562081       1.0000000     0.3424661
## infant deaths                        -0.52145513       0.3424661     1.0000000
## Alcohol                               0.45156710      -0.2217074    -0.3370801
## percentage expenditure                0.57668638      -0.3230491    -0.4242543
## Hepatitis B                       0.29839200      -0.1653819    -0.2940386
## Measles                          -0.24957375       0.1377107     0.5574775
## BMI                               0.58363707      -0.3882226    -0.4539651
## under-five deaths                -0.53843735       0.3579002     0.9928293
## Polio                             0.43534266      -0.2551687    -0.3552821
## Total expenditure                 0.27107970      -0.1759035    -0.2115073
## Diphtheria                        0.44819623      -0.2651449    -0.3434294
## HIV/AIDS                         -0.72026470       0.5236847     0.4267495
## GDP                               0.57079637      -0.3287902    -0.4107712
## Population                       -0.07978953       0.1004794     0.4858230
## thinness  1-19 years             -0.61960047       0.3957256     0.4357573
## thinness 5-9 years               -0.62981655       0.4135373     0.4516389
## Income composition of resources   0.84902333      -0.5237509    -0.5248356
## Schooling                         0.77476912      -0.4643623    -0.5467229
##                                     Alcohol percentage expenditure Hepatitis B
## Life expectancy                 0.45156710             0.57668638  0.29839200
## Adult Mortality                -0.22170740            -0.32304912 -0.16538192
## infant deaths                  -0.33708006            -0.42425435 -0.29403863
## Alcohol                         1.00000000             0.44814169  0.17742513
## percentage expenditure          0.44814169             1.00000000  0.18995676
## Hepatitis B                     0.17742513             0.18995676  1.00000000
## Measles                        -0.12127338            -0.19947285 -0.21228217
## BMI                             0.37501012             0.43188405  0.18899847
## under-five deaths              -0.33294586            -0.42741034 -0.29166860
## Polio                           0.32437796             0.27470259  0.75551471
## Total expenditure               0.25749475             0.26276857  0.11445678
## Diphtheria                      0.33825603             0.28713663  0.78496127
## HIV/AIDS                       -0.19955154            -0.36572118 -0.28829180
## GDP                             0.45409082             0.92721062  0.21748459
## Population                      0.03337387            -0.02947787 -0.11605589
## thinness  1-19 years           -0.43240480            -0.43709323 -0.07736988
## thinness 5-9 years             -0.41530516            -0.44430201 -0.09219078
## Income composition of resources  0.61620145             0.62405465  0.33708626
## Schooling                       0.60470712             0.61309962  0.35066907
##                                    Measles         BMI under-five deaths
## Life expectancy                -0.2495738  0.58363707        -0.5384373
## Adult Mortality                 0.1377107 -0.38822258         0.3579002
## infant deaths                   0.5574775 -0.45396505         0.9928293
## Alcohol                        -0.1212734  0.37501012        -0.3329459
## percentage expenditure         -0.1994729  0.43188405        -0.4274103
## Hepatitis B                    -0.2122822  0.18899847        -0.2916686
## Measles                         1.0000000 -0.24535508         0.5595391
## BMI                            -0.2453551  1.00000000        -0.4671115
## under-five deaths               0.5595391 -0.46711149         1.0000000
## Polio                          -0.1990552  0.24807828        -0.3531983
## Total expenditure              -0.1719058  0.24671464        -0.2142392
## Diphtheria                     -0.2012445  0.24920857        -0.3411521
## HIV/AIDS                        0.1293674 -0.47688931         0.4541495
## GDP                            -0.1599364  0.44602786        -0.4162367
## Population                      0.3105939 -0.06504985         0.4822878
## thinness  1-19 years            0.3115006 -0.59347779         0.4446156
## thinness 5-9 years              0.3357185 -0.60882909         0.4589028
## Income composition of resources -0.2030251  0.62343384        -0.5358550
## Schooling                      -0.2262436  0.62133279        -0.5569545
##                                       Polio Total expenditure Diphtheria
## Life expectancy                  0.43534266        0.27107970  0.4481962
## Adult Mortality                 -0.25516875       -0.17590353 -0.2651449
## infant deaths                   -0.35528210       -0.21150731 -0.3434294
## Alcohol                          0.32437796        0.25749475  0.3382560
## percentage expenditure           0.27470259        0.26276857  0.2871366
## Hepatitis B                      0.75551471        0.11445678  0.7849613
## Measles                         -0.19905520       -0.17190577 -0.2012445
## BMI                              0.24807828        0.24671464  0.2492086
## under-five deaths               -0.35319833       -0.21423918 -0.3411521
## Polio                            1.00000000        0.14075507  0.9311072
## Total expenditure                0.14075507        1.00000000  0.1525230
## Diphtheria                       0.93110717        0.15252303  1.0000000
## HIV/AIDS                        -0.37401715       -0.09179743 -0.3665770
## GDP                              0.30373162        0.18150863  0.3129743
## Population                      -0.08978785       -0.08619511 -0.0762494
## thinness  1-19 years            -0.19504705       -0.27888963 -0.2014556
## thinness 5-9 years              -0.20850611       -0.30636942 -0.2098635
## Income composition of resources  0.47894406        0.23594238  0.4913116
## Schooling                        0.46771150        0.27224112  0.4800501
##                                    HIV/AIDS        GDP  Population
## Life expectancy                 -0.72026470  0.5707964 -0.07978953
## Adult Mortality                  0.52368474 -0.3287902  0.10047938
## infant deaths                    0.42674946 -0.4107712  0.48582304
## Alcohol                         -0.19955154  0.4540908  0.03337387
## percentage expenditure          -0.36572118  0.9272106 -0.02947787
## Hepatitis B                     -0.28829180  0.2174846 -0.11605589
## Measles                          0.12936740 -0.1599364  0.31059388
## BMI                             -0.47688931  0.4460279 -0.06504985
## under-five deaths                0.45414950 -0.4162367  0.48228777
## Polio                           -0.37401715  0.3037316 -0.08978785
## Total expenditure               -0.09179743  0.1815086 -0.08619511
## Diphtheria                      -0.36657703  0.3129743 -0.07624940
## HIV/AIDS                         1.00000000 -0.4030439  0.09371056
## GDP                             -0.40304391  1.0000000 -0.02740440
## Population                       0.09371056 -0.0274044  1.00000000
## thinness  1-19 years             0.47895572 -0.4031451  0.07953644
## thinness 5-9 years               0.45827955 -0.4072094  0.09015301
## Income composition of resources -0.63010400  0.6529581 -0.03782649
## Schooling                       -0.57967397  0.6319253 -0.04715817
##                                 thinness  1-19 years thinness 5-9 years
## Life expectancy                          -0.61960047        -0.62981655
## Adult Mortality                           0.39572558         0.41353731
## infant deaths                             0.43575734         0.45163890
## Alcohol                                  -0.43240480        -0.41530516
## percentage expenditure                   -0.43709323        -0.44430201
## Hepatitis B                              -0.07736988        -0.09219078
## Measles                                   0.31150064         0.33571848
## BMI                                      -0.59347779        -0.60882909
## under-five deaths                         0.44461563         0.45890283
## Polio                                    -0.19504705        -0.20850611
## Total expenditure                        -0.27888963        -0.30636942
## Diphtheria                               -0.20145563        -0.20986350
## HIV/AIDS                                  0.47895572         0.45827955
## GDP                                      -0.40314513        -0.40720937
## Population                                0.07953644         0.09015301
## thinness  1-19 years                      1.00000000         0.92959088
## thinness 5-9 years                        0.92959088         1.00000000
## Income composition of resources          -0.59981895        -0.59082453
## Schooling                                -0.58072178        -0.57416719
##                                  Income composition of resources   Schooling
## Life expectancy                                      0.84902333  0.77476912
## Adult Mortality                                     -0.52375093 -0.46436232
## infant deaths                                       -0.52483560 -0.54672287
## Alcohol                                              0.61620145  0.60470712
## percentage expenditure                               0.62405465  0.61309962
## Hepatitis B                                          0.33708626  0.35066907
## Measles                                             -0.20302506 -0.22624362
## BMI                                                  0.62343384  0.62133279
## under-five deaths                                   -0.53585501 -0.55695449
## Polio                                                0.47894406  0.46771150
## Total expenditure                                    0.23594238  0.27224112
## Diphtheria                                           0.49131155  0.48005007
## HIV/AIDS                                            -0.63010400 -0.57967397
## GDP                                                  0.65295806  0.63192528
## Population                                          -0.03782649 -0.04715817
## thinness  1-19 years                                -0.59981895 -0.58072178
## thinness 5-9 years                                  -0.59082453 -0.57416719
## Income composition of resources                      1.00000000  0.90945734
## Schooling                                            0.90945734  1.00000000
```

```
#almost all variables have high correlation coefficients, except for Hepatitis B and Population
```

Second, tidying the data was performed. Separate data frames were created based on the focus of analysis per section. The variables were grouped into categories: immunization factors, mortality factors, economic factors, social factors, and health factors. This was done to easily organize the data and to minimize the impact of removing the missing values. Data values for GDP and percentage expenditure were rounded off to three decimal places.

```
#round off numbers
life_expectancy$`percentage expenditure` <- round(life_expectancy$`percentage expenditure`, digits = 3)
life_expectancy$GDP <- round(life_expectancy$GDP, digits = 3)

#inspect the years with data
unique(life_expectancy$Year)
```

```
##  [1] 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001
## [16] 2000
```

```
#count how many countries have data for each year
life_expectancy %>%
  group_by(Year) %>%
  summarise(count = n())
```

```
## # A tibble: 16 x 2
##     Year count
##    <dbl> <int>
##  1  2000   183
##  2  2001   183
##  3  2002   183
##  4  2003   183
##  5  2004   183
##  6  2005   183
##  7  2006   183
##  8  2007   183
##  9  2008   183
## 10  2009   183
## 11  2010   183
## 12  2011   183
## 13  2012   183
## 14  2013   193
## 15  2014   183
## 16  2015   183
```

```
# creating a tidy dataframe for each categories

#immunization related factors
immunization <- life_expectancy %>%
  select(Country, Year, Status,`Life expectancy`,`Hepatitis B`,Measles,Polio,Diphtheria) %>%
  remove_missing()
```

```
## Warning: Removed 563 rows containing missing values.
```

```
#economical related factors
economical <- life_expectancy %>%
  select(Country, Year, Status,`Life expectancy`,`percentage expenditure`,`Total expenditure`,GDP,`Inco
  remove_missing()
```

```
## Warning: Removed 611 rows containing missing values.
```

```
#social related factors
social <- life_expectancy %>%
  select(Country, Year, Status,`Life expectancy`,Schooling) %>%
  remove_missing()
```

```
## Warning: Removed 170 rows containing missing values.
```

```
#health related factors
body <- life_expectancy %>%
  select(Country, Year, Status,`Life expectancy`,`thinness  1-19 years`,`thinness 5-9 years`,BMI) %>%
  remove_missing()
```

```
## Warning: Removed 42 rows containing missing values.
```

```
#lifestyle related factor
alcohol <- life_expectancy %>%
  select(Country, Year, Status,`Life expectancy`,Alcohol) %>%
  remove_missing()
```

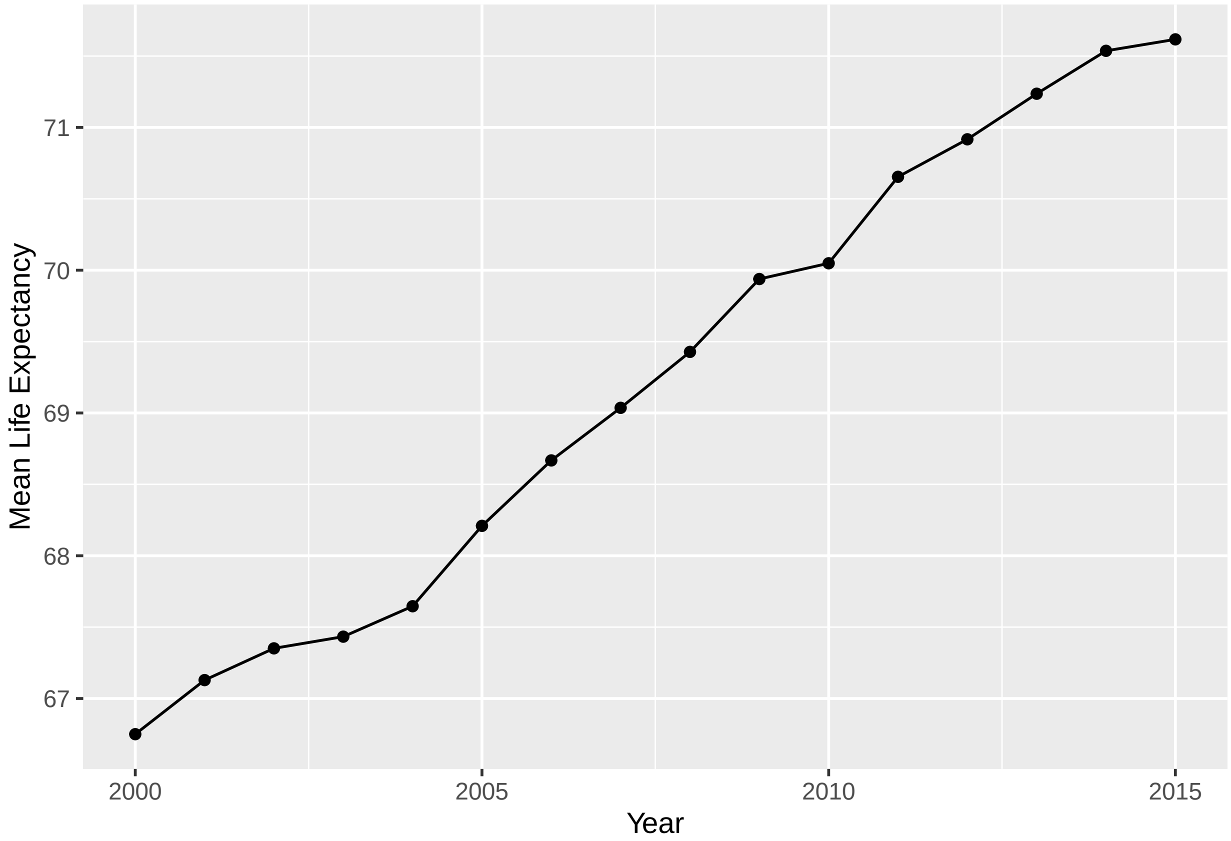
```
## Warning: Removed 203 rows containing missing values.
```

```
#mortality and health related factors
mortality_health <- life_expectancy %>%
  select(Country, Year, Status,`Life expectancy`,`Adult Mortality`,`infant deaths`,`under-five deaths`,
  remove_missing()
```

```
## Warning: Removed 236 rows containing missing values.
```

The mean life expectancy from 2000-2015 were visualized using geom_point() and geom_line() to examine if life expectancy has changed over the years. The time series graph showed that the average life span has increased from 66.75 years in 2000 to 71.62 years in 2015.

```
mean_life <- life_expectancy %>%
                select(Year, `Life expectancy`) %>%
                filter(!is.na(`Life expectancy`)) %>%
                group_by(Year) %>%
                arrange(Year) %>%
                summarise(mean = mean(`Life expectancy`))

ggplot(data = mean_life, aes(x = Year, y = `mean`)) + geom_point() + geom_line() + ylab("Mean Life Expe
```



```
#shows increase in life expectancy from 2000-2015
```

Life Expectancy and Predicting Factors

We hypothesized that life expectancy is greatly influenced by health related factors, such as immunization, health care expenditure, and child and adult mortality. To know which predictors significantly affect life expectancy, multiple linear regression was used. Backward regression was implemented to end up with the variables that have the greatest effect on life expectancy. To make sure that the model is valid, the p-value for each variable and the adjusted R2 value and residual error for each model were examined. The AIC method was used in choosing the best model; the least value corresponds to the model that can explain most variability in life expectancy.

The regression model that showed the lowest AIC and the highest adjusted R2 value contained the following variables: adult mortality, infant deaths, alcohol, percentage expenditure, BMI, under-five deaths, total expenditure, Diphtheria, HIV/AIDS, thinness 5-9 years, income composition of resources, and schooling. All predictor values included in this model have a p-value lower than 0.05, except for alcohol and total expenditure. The histogram of the residuals showed a normal distribution.

This model can explain 83.31% of variability in life expectancy and only has 3.594 residual standard error. The final equation of the model is as follows. Life expectancy = 53.30 - 0.02 (Adult Mortality) + 0.09 (Infant deaths) - 0.05 (Alcohol) + 0.0005 (Percentage expenditure) + 0.033(BMI) - 0.07 (under-five deaths) + 0.08 (Total Expenditure) + 0.01 (Diphtheria) - 0.44 (HIV/AIDS) - 0.06 (thinness 5-9 years) + 9.88 (income composition of resources) + 0.89 (Schooling) - 0.0028 error

This agrees with our hypothesis, given that most of these significant variables are health related factors. Income composition resources and schooling have the highest slope values, emphasizing the critical role of economic factors and education in improving people's lives. The slope value for the infant deaths was the only one that did not meet with our expectations, since we expected a negative relationship with life expectancy. However, according to Murray, 1988, infant mortality rate is not a good indicator of overall mortality. It could be argued that this slight positive relationship can be due to differences between the developing and developed countries. Overall, the model has a p-value of 2.2x10-16, confirming that it is statistically significant.

```
#remove all rows with NA
life_complete <- life_expectancy %>%
                  drop_na() %>%
                  select(-Country, -Year, -Status)
dim(life_complete)
```

```
## [1] 1649   19
```

```
#removed 1289 values

#know which predictor variables are most important
life_lm <- lm(`Life expectancy` ~., data = life_complete)
summary(life_lm) #check variables with lowest p values
```

```
##
## Call:
## lm(formula = `Life expectancy` ~ ., data = life_complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0176  -2.0454  -0.0185   2.2260  11.9157
##
## Coefficients:
```

```
##                                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                        5.328e+01  7.358e-01  72.412  < 2e-16 ***
## ‘Adult Mortality‘                 -1.689e-02  9.473e-04 -17.828  < 2e-16 ***
## ‘infant deaths‘                    9.369e-02  1.068e-02   8.776  < 2e-16 ***
## Alcohol                           -5.435e-02  3.061e-02  -1.776   0.0760 .  
## ‘percentage expenditure‘           3.777e-04  1.805e-04   2.093   0.0365 *  
## ‘Hepatitis B‘                     -5.582e-03  4.446e-03  -1.256   0.2095    
## Measles                           -8.617e-06  1.081e-05  -0.797   0.4253    
## BMI                                3.350e-02  6.011e-03   5.573 2.92e-08 ***
## ‘under-five deaths‘               -7.047e-02  7.728e-03  -9.119  < 2e-16 ***
## Polio                              7.836e-03  5.163e-03   1.518   0.1293    
## ‘Total expenditure‘                7.975e-02  4.074e-02   1.958   0.0505 .  
## Diphtheria                         1.439e-02  5.938e-03   2.423   0.0155 *  
## ‘HIV/AIDS‘                        -4.383e-01  1.788e-02 -24.519  < 2e-16 ***
## GDP                                1.383e-05  2.838e-05   0.487   0.6260    
## Population                        -6.917e-10  1.753e-09  -0.395   0.6931    
## ‘thinness  1-19 years‘            -8.670e-03  5.310e-02  -0.163   0.8703    
## ‘thinness 5-9 years‘              -5.123e-02  5.242e-02  -0.977   0.3286    
## ‘Income composition of resources‘  9.824e+00  8.340e-01  11.780  < 2e-16 ***
## Schooling                          8.783e-01  5.939e-02  14.789  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 3.596 on 1630 degrees of freedom
## Multiple R-squared:  0.8347, Adjusted R-squared:  0.8329 
## F-statistic: 457.4 on 18 and 1630 DF,  p-value: < 2.2e-16
```

```
#perform backward regression
backward_life <- step(life_lm, direction = "backward", scope = formula(life_lm))
```

```
## Start:  AIC=4239.58
## ‘Life expectancy‘ ~ ‘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
## 
##                                     Df Sum of Sq   RSS    AIC
## - ‘thinness  1-19 years‘             1       0.3 21076 4237.6
## - Population                         1       2.0 21078 4237.7
## - GDP                                1       3.1 21079 4237.8
## - Measles                            1       8.2 21084 4238.2
## - ‘thinness 5-9 years‘               1      12.3 21088 4238.5
## - ‘Hepatitis B‘                      1      20.4 21096 4239.2
## <none>                                           21076 4239.6
## - Polio                              1      29.8 21106 4239.9
## - Alcohol                            1      40.8 21117 4240.8
## - ‘Total expenditure‘                1      49.5 21125 4241.5
## - ‘percentage expenditure‘           1      56.6 21132 4242.0
## - Diphtheria                         1      75.9 21152 4243.5
## - BMI                                1     401.6 21477 4268.7
## - ‘infant deaths‘                    1     995.8 22072 4313.7
## - ‘under-five deaths‘                1    1075.1 22151 4319.6
## - ‘Income composition of resources‘  1    1794.3 22870 4372.3
## - Schooling                          1    2828.0 23904 4445.2
## - ‘Adult Mortality‘                  1    4109.7 25186 4531.3
## - ‘HIV/AIDS‘                         1    7773.2 28849 4755.3
##
## Step:  AIC=4237.61
## ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ + Alcohol +
##     ‘percentage expenditure‘ + ‘Hepatitis B‘ + Measles + BMI +
##     ‘under-five deaths‘ + Polio + ‘Total expenditure‘ + Diphtheria +
##     ‘HIV/AIDS‘ + GDP + Population + ‘thinness 5-9 years‘ + ‘Income composition of resources‘ +
##     Schooling
##
##                                      Df Sum of Sq   RSS    AIC
## - Population                         1       2.1 21078 4235.8
## - GDP                                1       3.1 21079 4235.8
## - Measles                            1       8.2 21084 4236.2
## - ‘Hepatitis B‘                      1      20.5 21097 4237.2
## <none>                                           21076 4237.6
## - Polio                              1      29.5 21106 4237.9
## - Alcohol                            1      40.4 21117 4238.8
## - ‘Total expenditure‘                1      49.5 21126 4239.5
## - ‘percentage expenditure‘           1      56.6 21133 4240.0
## - ‘thinness 5-9 years‘               1      61.9 21138 4240.4
## - Diphtheria                         1      76.3 21153 4241.6
## - BMI                                1     404.0 21480 4266.9
## - ‘infant deaths‘                    1     997.1 22073 4311.8
## - ‘under-five deaths‘                1    1077.5 22154 4317.8
## - ‘Income composition of resources‘  1    1798.6 22875 4370.6
## - Schooling                          1    2840.6 23917 4444.1
## - ‘Adult Mortality‘                  1    4111.9 25188 4529.5
## - ‘HIV/AIDS‘                         1    7775.7 28852 4753.4
##
## Step:  AIC=4235.77
## ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ + Alcohol +
##     ‘percentage expenditure‘ + ‘Hepatitis B‘ + Measles + BMI +
##     ‘under-five deaths‘ + Polio + ‘Total expenditure‘ + Diphtheria +
##     ‘HIV/AIDS‘ + GDP + ‘thinness 5-9 years‘ + ‘Income composition of resources‘ +
##     Schooling
##
##                                      Df Sum of Sq   RSS    AIC
## - GDP                                1       3.1 21081 4234.0
## - Measles                            1       7.5 21086 4234.4
## - ‘Hepatitis B‘                      1      20.2 21098 4235.3
## <none>                                           21078 4235.8
## - Polio                              1      29.4 21108 4236.1
## - Alcohol                            1      40.4 21119 4236.9
## - ‘Total expenditure‘                1      49.6 21128 4237.6
## - ‘percentage expenditure‘           1      56.3 21135 4238.2
## - ‘thinness 5-9 years‘               1      61.9 21140 4238.6
## - Diphtheria                         1      75.7 21154 4239.7
## - BMI                                1     402.5 21481 4265.0
## - ‘infant deaths‘                    1    1036.2 22114 4312.9
## - ‘under-five deaths‘                1    1095.1 22173 4317.3
## - ‘Income composition of resources‘  1    1800.7 22879 4368.9
## - Schooling                          1    2843.2 23921 4442.4
## - ‘Adult Mortality‘                  1    4125.2 25203 4528.5
## - ‘HIV/AIDS‘                         1    7773.9 28852 4751.5
##
## Step:  AIC=4234.01
## ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ + Alcohol +
##     ‘percentage expenditure‘ + ‘Hepatitis B‘ + Measles + BMI +
##     ‘under-five deaths‘ + Polio + ‘Total expenditure‘ + Diphtheria +
##     ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ + ‘Income composition of resources‘ +
##     Schooling
##
##                                      Df Sum of Sq   RSS    AIC
## - Measles                            1       7.5 21089 4232.6
## - ‘Hepatitis B‘                      1      19.8 21101 4233.6
## <none>                                           21081 4234.0
## - Polio                              1      29.9 21111 4234.4
## - Alcohol                            1      39.7 21121 4235.1
## - ‘Total expenditure‘                1      48.7 21130 4235.8
## - ‘thinness 5-9 years‘               1      62.5 21144 4236.9
## - Diphtheria                         1      75.2 21157 4237.9
## - BMI                                1     401.1 21482 4263.1
## - ‘percentage expenditure‘           1     818.5 21900 4294.8
## - ‘infant deaths‘                    1    1036.2 22118 4311.1
## - ‘under-five deaths‘                1    1094.9 22176 4315.5
## - ‘Income composition of resources‘  1    1821.0 22902 4368.6
## - Schooling                          1    2882.9 23964 4443.4
## - ‘Adult Mortality‘                  1    4124.4 25206 4526.7
## - ‘HIV/AIDS‘                         1    7775.2 28857 4749.7
##
## Step:  AIC=4232.6
## ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ + Alcohol +
##     ‘percentage expenditure‘ + ‘Hepatitis B‘ + BMI + ‘under-five deaths‘ +
##     Polio + ‘Total expenditure‘ + Diphtheria + ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ +
##     ‘Income composition of resources‘ + Schooling
##
##                                      Df Sum of Sq   RSS    AIC
## - ‘Hepatitis B‘                      1      19.4 21108 4232.1
## <none>                                           21089 4232.6
## - Polio                              1      29.7 21119 4232.9
## - Alcohol                            1      41.0 21130 4233.8
## - ‘Total expenditure‘                1      50.6 21139 4234.5
## - ‘thinness 5-9 years‘               1      57.6 21147 4235.1
## - Diphtheria                         1      75.2 21164 4236.5
## - BMI                                1     415.3 21504 4262.8
## - ‘percentage expenditure‘           1     822.7 21912 4293.7
## - ‘infant deaths‘                    1    1062.7 22152 4311.7
## - ‘under-five deaths‘                1    1107.0 22196 4315.0
## - ‘Income composition of resources‘  1    1821.5 22910 4367.2
## - Schooling                          1    2895.0 23984 4442.7
## - ‘Adult Mortality‘                  1    4121.9 25211 4525.0
## - ‘HIV/AIDS‘                         1    7790.7 28880 4749.0
##
## Step:  AIC=4232.12
## ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ + Alcohol +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + Polio +
##     ‘Total expenditure‘ + Diphtheria + ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ +
##     ‘Income composition of resources‘ + Schooling
##
##                                     Df Sum of Sq   RSS    AIC
## - Polio                              1      23.1 21131 4231.9
## <none>                                           21108 4232.1
## - Alcohol                            1      39.3 21148 4233.2
## - ‘Total expenditure‘                1      48.0 21156 4233.9
## - Diphtheria                         1      56.3 21165 4234.5
## - ‘thinness 5-9 years‘               1      61.0 21169 4234.9
## - BMI                                1     408.0 21516 4261.7
## - ‘percentage expenditure‘           1     847.5 21956 4295.0
## - ‘infant deaths‘                    1    1073.1 22181 4311.9
## - ‘under-five deaths‘                1    1112.9 22221 4314.8
## - ‘Income composition of resources‘  1    1830.3 22939 4367.2
## - Schooling                          1    2888.1 23996 4441.6
## - ‘Adult Mortality‘                  1    4140.8 25249 4525.5
## - ‘HIV/AIDS‘                         1    7771.3 28880 4747.0
##
## Step:  AIC=4231.92
## ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ + Alcohol +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + ‘Total expenditure‘ +
##     Diphtheria + ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ + ‘Income composition of resources‘ +
##     Schooling
##
##                                     Df Sum of Sq   RSS    AIC
## <none>                                           21131 4231.9
## - Alcohol                            1      37.0 21168 4232.8
## - ‘Total expenditure‘                1      48.9 21180 4233.7
## - ‘thinness 5-9 years‘               1      59.1 21191 4234.5
## - Diphtheria                         1     141.0 21272 4240.9
## - BMI                                1     407.4 21539 4261.4
## - ‘percentage expenditure‘           1     841.0 21972 4294.3
## - ‘infant deaths‘                    1    1092.4 22224 4313.0
## - ‘under-five deaths‘                1    1134.8 22266 4316.2
## - ‘Income composition of resources‘  1    1829.1 22961 4366.8
## - Schooling                          1    2954.7 24086 4445.7
## - ‘Adult Mortality‘                  1    4171.8 25303 4527.0
## - ‘HIV/AIDS‘                         1    7768.2 28900 4746.2
```

```
backward_life
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     Alcohol + ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ +
##     ‘Total expenditure‘ + Diphtheria + ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ +
##     ‘Income composition of resources‘ + Schooling, data = life_complete)
##
## Coefficients:
##                       (Intercept)                  ‘Adult Mortality‘
##                        53.2986964                         -0.0169728
##                   ‘infant deaths‘                            Alcohol
##                         0.0917273                         -0.0514941
##          ‘percentage expenditure‘                                BMI
##                         0.0004651                          0.0334791
##               ‘under-five deaths‘                ‘Total expenditure‘
##                        -0.0694512                          0.0789781
##                        Diphtheria                         ‘HIV/AIDS‘
##                         0.0149712                         -0.4374122
##              ‘thinness 5-9 years‘  ‘Income composition of resources‘
##                        -0.0565893                          9.8831238
##                         Schooling
##                         0.8869012
```

```
#result of backward regression
backward_lifelm <- lm(`Life expectancy` ~ `Adult Mortality` + `infant deaths` +
    Alcohol + `percentage expenditure` + BMI + `under-five deaths` +
    `Total expenditure` + Diphtheria + `HIV/AIDS` + `thinness 5-9 years` +
    `Income composition of resources` + Schooling, data = life_complete)

backward_lifelm #use to find equation
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     Alcohol + ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ +
##     ‘Total expenditure‘ + Diphtheria + ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ +
##     ‘Income composition of resources‘ + Schooling, data = life_complete)
##
## Coefficients:
##                       (Intercept)                  ‘Adult Mortality‘
##                        53.2986964                         -0.0169728
##                   ‘infant deaths‘                            Alcohol
##                         0.0917273                         -0.0514941
##          ‘percentage expenditure‘                                BMI
##                         0.0004651                          0.0334791
##               ‘under-five deaths‘                ‘Total expenditure‘
##                        -0.0694512                          0.0789781
##                        Diphtheria                         ‘HIV/AIDS‘
##                         0.0149712                         -0.4374122
##              ‘thinness 5-9 years‘  ‘Income composition of resources‘
##                        -0.0565893                          9.8831238
##                         Schooling
##                         0.8869012
```

```
summary(backward_lifelm) #used to find adjusted R squared, residual standard error, and median error
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     Alcohol + ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ +
##     ‘Total expenditure‘ + Diphtheria + ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ +
##     ‘Income composition of resources‘ + Schooling, data = life_complete)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -17.0473  -2.0562  -0.0251   2.2284  11.8901
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.330e+01  7.071e-01  75.382  < 2e-16 ***
## ‘Adult Mortality‘                -1.697e-02  9.444e-04 -17.972  < 2e-16 ***
## ‘infant deaths‘                   9.173e-02  9.974e-03   9.196  < 2e-16 ***
## Alcohol                          -5.149e-02  3.044e-02  -1.692 0.090924 .
## ‘percentage expenditure‘          4.651e-04  5.764e-05   8.069 1.36e-15 ***
## BMI                               3.348e-02  5.961e-03   5.616 2.29e-08 ***
## ‘under-five deaths‘              -6.945e-02  7.410e-03  -9.373  < 2e-16 ***
## ‘Total expenditure‘               7.898e-02  4.061e-02   1.945 0.051968 .
## Diphtheria                        1.497e-02  4.532e-03   3.304 0.000975 ***
## ‘HIV/AIDS‘                       -4.374e-01  1.784e-02 -24.524  < 2e-16 ***
## ‘thinness 5-9 years‘             -5.659e-02  2.645e-02  -2.140 0.032512 *
## ‘Income composition of resources‘  9.883e+00  8.305e-01  11.900  < 2e-16 ***
## Schooling                         8.869e-01  5.864e-02  15.125  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.594 on 1636 degrees of freedom
## Multiple R-squared:  0.8343, Adjusted R-squared:  0.8331
## F-statistic: 686.4 on 12 and 1636 DF,  p-value: < 2.2e-16
```

```
#trying to increase adjusted R2 by removing those with low p values
backward_lifelm1 <- lm(`Life expectancy` ~ `Adult Mortality` + `infant deaths` +
    `percentage expenditure` + BMI + `under-five deaths` + Diphtheria + `HIV/AIDS` + `thinness 5-9 years
    `Income composition of resources` + Schooling, data = life_complete)

backward_lifelm1
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + Diphtheria +
##     ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ + ‘Income composition of resources‘ +
##     Schooling, data = life_complete)
##
## Coefficients:
##                     (Intercept)                ‘Adult Mortality‘
##                      53.9623486                      -0.0171698
##                 ‘infant deaths‘         ‘percentage expenditure‘
##                       0.0939272                       0.0004554
##                             BMI              ‘under-five deaths‘
##                       0.0341701                      -0.0712594
##                      Diphtheria                       ‘HIV/AIDS‘
##                       0.0149217                      -0.4366019
##            ‘thinness 5-9 years‘  ‘Income composition of resources‘
##                      -0.0534609                       9.6168023
##                       Schooling
##                       0.8663266
```

```
summary(backward_lifelm1)
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + Diphtheria +
##     ‘HIV/AIDS‘ + ‘thinness 5-9 years‘ + ‘Income composition of resources‘ +
##     Schooling, data = life_complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5586  -2.0531  -0.0028   2.2237  12.0221
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       53.9623486  0.6563592  82.215  < 2e-16 ***
## ‘Adult Mortality‘                 -0.0171698  0.0009414 -18.239  < 2e-16 ***
## ‘infant deaths‘                    0.0939272  0.0098250   9.560  < 2e-16 ***
## ‘percentage expenditure‘           0.0004554  0.0000564   8.074 1.31e-15 ***
## BMI                                0.0341702  0.0059626   5.731 1.19e-08 ***
## ‘under-five deaths‘               -0.0712594  0.0072860  -9.780  < 2e-16 ***
## Diphtheria                         0.0149217  0.0045181   3.303 0.000978 ***
## ‘HIV/AIDS‘                        -0.4366019  0.0176828 -24.691  < 2e-16 ***
## ‘thinness 5-9 years‘              -0.0534609  0.0261065  -2.048 0.040738 *
## ‘Income composition of resources‘  9.6168023  0.8210847  11.712  < 2e-16 ***
## Schooling                          0.8663266  0.0556497  15.567  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.599 on 1638 degrees of freedom
## Multiple R-squared:  0.8337, Adjusted R-squared:  0.8326
## F-statistic: 820.9 on 10 and 1638 DF,  p-value: < 2.2e-16
```

```
backward_lifelm2 <- lm(`Life expectancy` ~ `Adult Mortality` + `infant deaths` +
    `percentage expenditure` + BMI + `under-five deaths` + Diphtheria + `HIV/AIDS`  +
    `Income composition of resources` + Schooling, data = life_complete)

backward_lifelm2
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + Diphtheria +
##     ‘HIV/AIDS‘ + ‘Income composition of resources‘ + Schooling,
##     data = life_complete)
##
## Coefficients:
##                       (Intercept)                  ‘Adult Mortality‘
##                         53.394564                          -0.017266
##                   ‘infant deaths‘           ‘percentage expenditure‘
##                          0.091575                           0.000462
##                               BMI                ‘under-five deaths‘
##                           0.038435                          -0.070036
##                         Diphtheria                         ‘HIV/AIDS‘
##                           0.014797                          -0.438326
## ‘Income composition of resources‘                          Schooling
##                           9.778268                           0.873643
```

```r
summary(backward_lifelm2)
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + Diphtheria +
##     ‘HIV/AIDS‘ + ‘Income composition of resources‘ + Schooling,
##     data = life_complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5647  -2.0290   0.0135   2.2187  12.2937
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         5.339e+01  5.955e-01  89.663  < 2e-16 ***
## ‘Adult Mortality‘                  -1.727e-02  9.411e-04 -18.345  < 2e-16 ***
## ‘infant deaths‘                     9.157e-02  9.767e-03   9.376  < 2e-16 ***
## ‘percentage expenditure‘            4.620e-04  5.636e-05   8.198 4.89e-16 ***
## BMI                                 3.844e-02  5.592e-03   6.873 8.92e-12 ***
## ‘under-five deaths‘                -7.004e-02  7.269e-03  -9.636  < 2e-16 ***
## Diphtheria                          1.480e-02  4.522e-03   3.272  0.00109 **
## ‘HIV/AIDS‘                         -4.383e-01  1.768e-02 -24.792  < 2e-16 ***
## ‘Income composition of resources‘  9.778e+00  8.181e-01  11.953  < 2e-16 ***
## Schooling                           8.736e-01  5.559e-02  15.716  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.602 on 1639 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8323
## F-statistic: 909.9 on 9 and 1639 DF,  p-value: < 2.2e-16
```

```r
backward_lifelm3 <- lm(`Life expectancy` ~ `Adult Mortality` + `infant deaths` +
    `percentage expenditure` + BMI + `under-five deaths` + `HIV/AIDS`  +
    `Income composition of resources` + Schooling, data = life_complete)

backward_lifelm3
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + ‘HIV/AIDS‘ +
##     ‘Income composition of resources‘ + Schooling, data = life_complete)
##
## Coefficients:
##                        (Intercept)                  ‘Adult Mortality‘
##                          54.222912                          -0.017278
##                    ‘infant deaths‘           ‘percentage expenditure‘
##                           0.096574                           0.000458
##                                BMI                ‘under-five deaths‘
##                           0.037267                          -0.073977
##                         ‘HIV/AIDS‘  ‘Income composition of resources‘
##                          -0.439380                          10.089825
##                          Schooling
##                           0.897017
```

```
summary(backward_lifelm3)
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Adult Mortality‘ + ‘infant deaths‘ +
##     ‘percentage expenditure‘ + BMI + ‘under-five deaths‘ + ‘HIV/AIDS‘ +
##     ‘Income composition of resources‘ + Schooling, data = life_complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7109  -2.0552   0.0037   2.2774  12.2723
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        5.422e+01  5.406e-01 100.300  < 2e-16 ***
## ‘Adult Mortality‘                 -1.728e-02  9.439e-04 -18.305  < 2e-16 ***
## ‘infant deaths‘                    9.657e-02  9.675e-03   9.981  < 2e-16 ***
## ‘percentage expenditure‘           4.580e-04  5.652e-05   8.103 1.04e-15 ***
## BMI                                3.727e-02  5.598e-03   6.658 3.78e-11 ***
## ‘under-five deaths‘               -7.398e-02  7.189e-03 -10.290  < 2e-16 ***
## ‘HIV/AIDS‘                        -4.394e-01  1.773e-02 -24.783  < 2e-16 ***
## ‘Income composition of resources‘  1.009e+01  8.149e-01  12.381  < 2e-16 ***
## Schooling                          8.970e-01  5.529e-02  16.224  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.613 on 1640 degrees of freedom
## Multiple R-squared:  0.8321, Adjusted R-squared:  0.8313
## F-statistic:  1016 on 8 and 1640 DF,  p-value: < 2.2e-16
```

```
#compare AIC values
AIC(backward_lifelm)
```

```
## [1] 8913.582
```

```
AIC(backward_lifelm1)
```

```
## [1] 8916.004
```

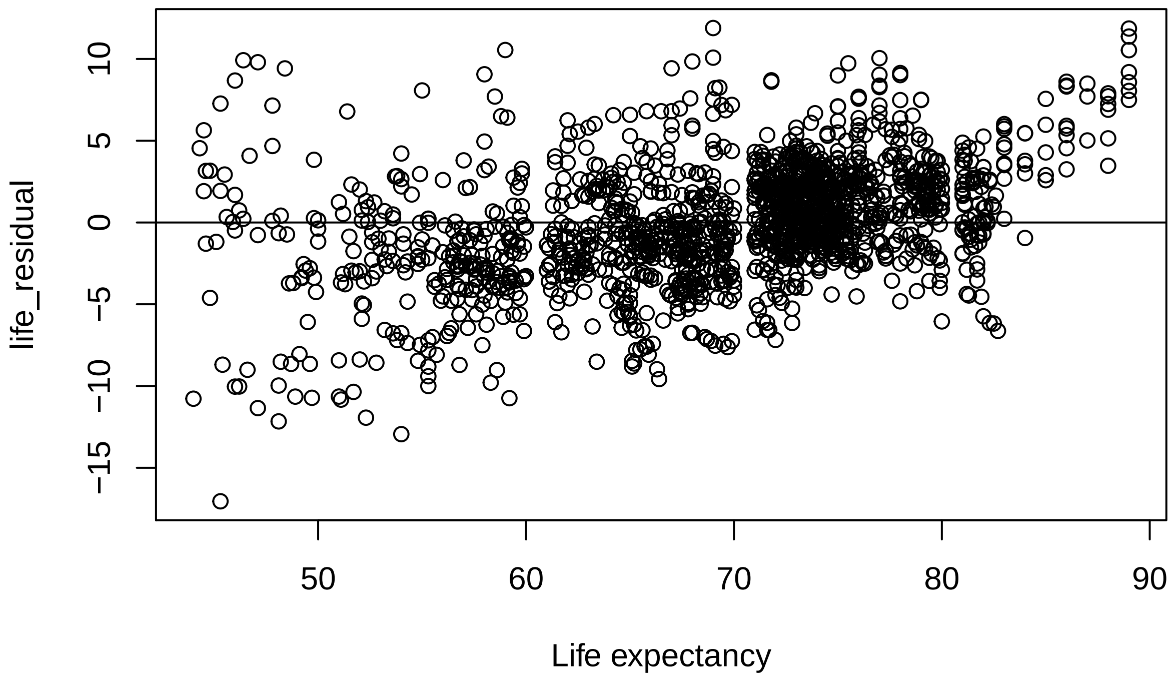
```
AIC(backward_lifelm2)
```

```
## [1] 8918.221
```

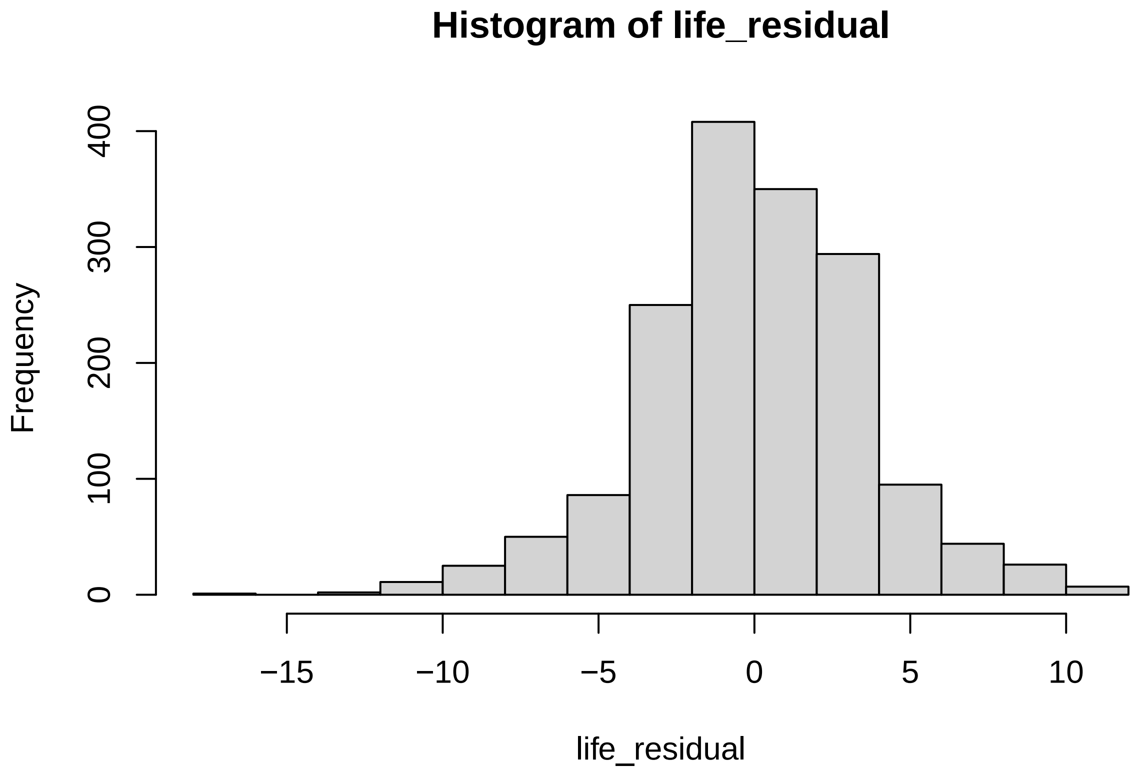
```
AIC(backward_lifelm3)
```

```
## [1] 8926.958
```

```
#best model is the model found using the backward regression, has the lowest AIC

#check the residuals of the best model
life_residual <-residuals(backward_lifelm)
plot(life_residual  ~ `Life expectancy`, data = life_complete)
abline(h = 0)
```



```
hist(life_residual) #showing normality
```



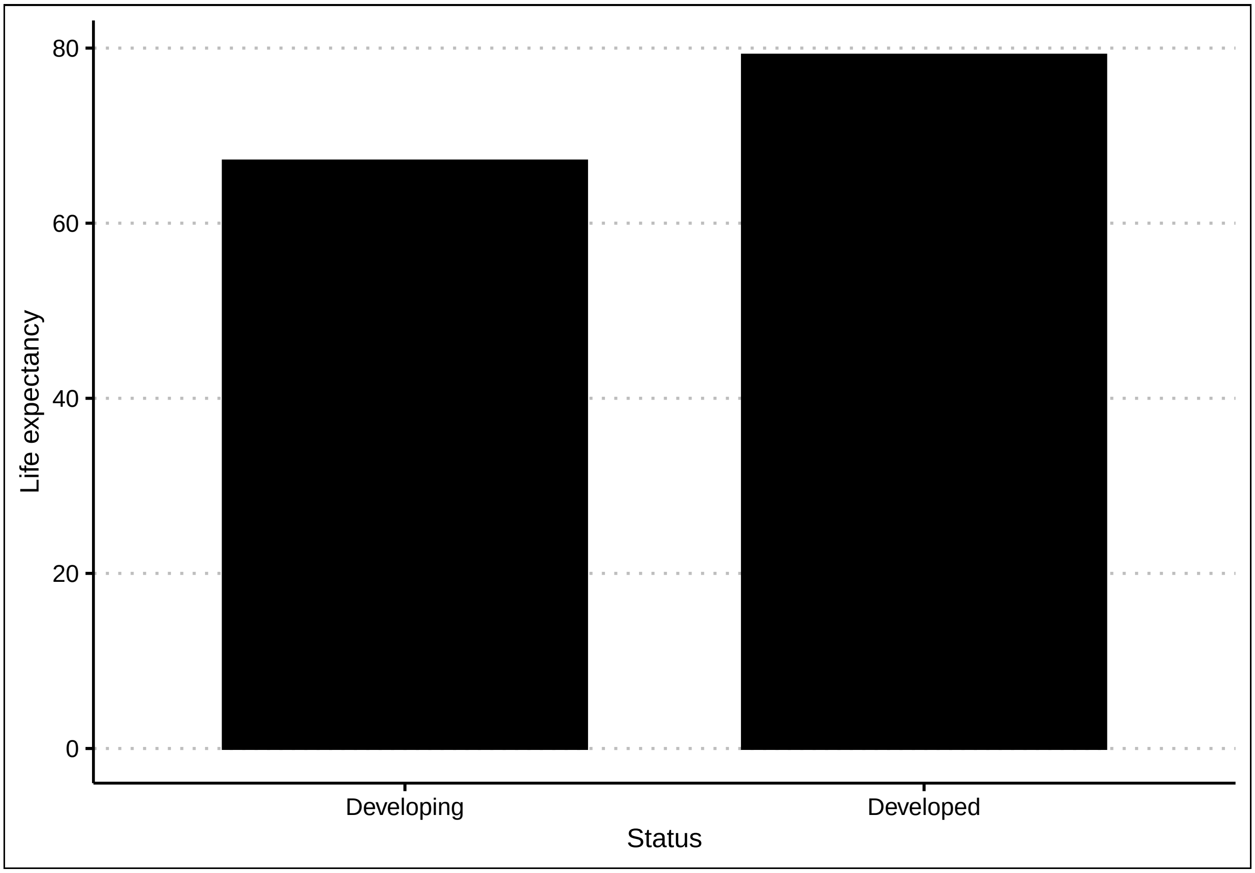
Life Expectancy Between Developed and Developing Countries

The mean life expectancy between developed and developing countries was visualized in a bar plot. The normality of the data was tested using the Shapiro-Wilk test, showing a non-normal distribution of life expectancy between developed and developing countries. Thus, the difference was tested with a Mann-Whitney test, wherein results showed that there is a significant difference between the life expectancy of developed and developing countries (p-value < 2.2 x 10-16). The life expectancy is higher in developed countries with an average and a median of 79 years than in developing countries with an average life expectancy of 67 years and a median of 69 years.

```
# df for life expectancy and status
expectancy_status <- life_expectancy %>%
  select(`Life expectancy`,Status, Country) %>%
  remove_missing()
```

```
## Warning: Removed 10 rows containing missing values.
```

```
expectancy_status <- rename(expectancy_status,life_expectancy=`Life expectancy`)

# bar plot computing life expectancy based on status
status <- ggbarplot(data = expectancy_status,
                    y = "life_expectancy", x = "Status",
                    add = "mean", fill = "black", ylab = "Life expectancy",
               ggtheme = theme_clean())
status
```



```
# normality of the data
shapiro.test(expectancy_status$life_expectancy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  expectancy_status$life_expectancy
## W = 0.95605, p-value < 2.2e-16
```

```
# non-normal distribution

#perform Mann Whitney Test
wilcox.test(`life_expectancy` ~ Status, data = expectancy_status, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  life_expectancy by Status
## W = 1131521, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
#significant difference

#summarize mean and median results
```

```
expectancy_status %>%
group_by(Status) %>%
summarize(Mean = mean(life_expectancy, na.rm=TRUE),
          Median = median(life_expectancy), Count = n())
```

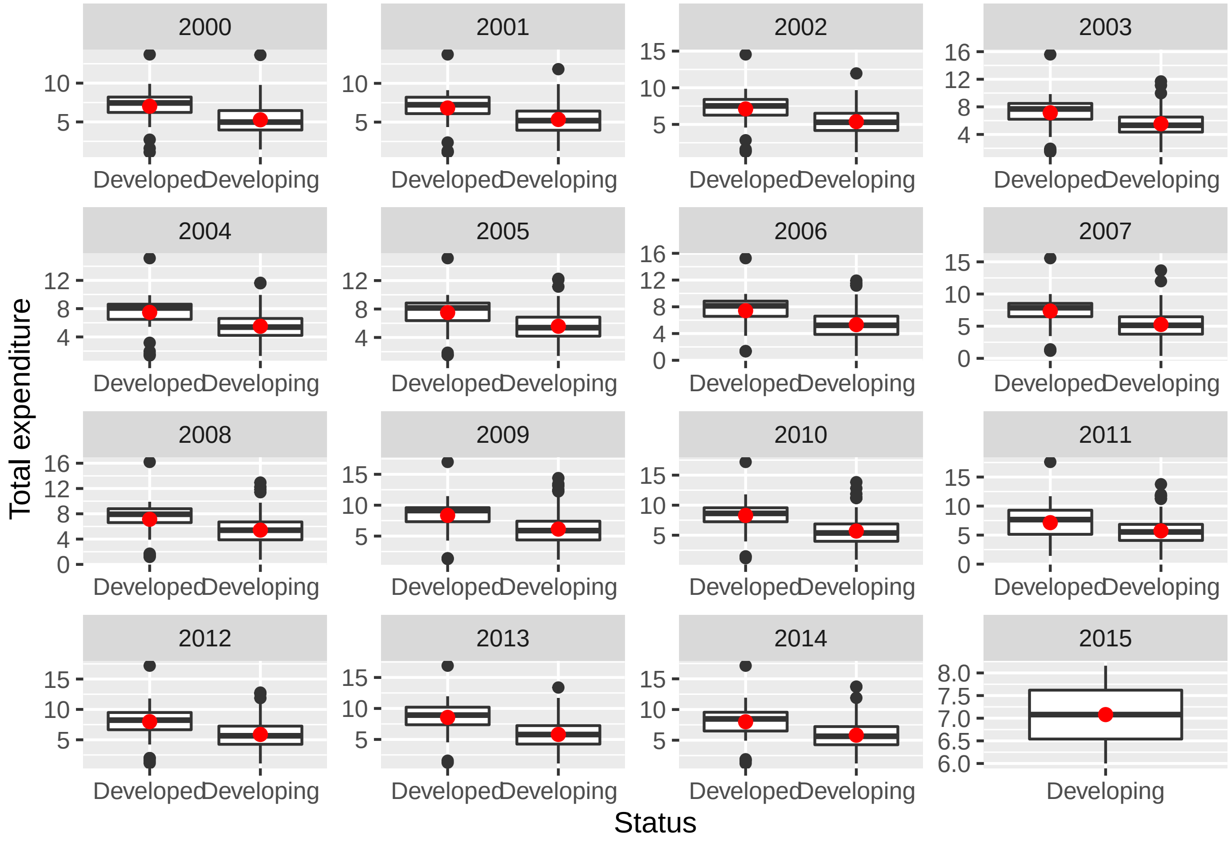
```
## # A tibble: 2 x 4
##   Status      Mean Median Count
##   <chr>      <dbl>  <dbl> <int>
## 1 Developed   79.2   79.2   512
## 2 Developing  67.1   69    2416
```

Life Expectancy and Health Expenditure

We hypothesized that health expenditure is positively associated with life expectancy. Larger budget for public health access will contribute positively to people's quality of life. We chose to analyze the total expenditure rather than the percentage expenditure as it gives more information about the government's allocation of its spendings for healthcare purposes. The total expenditure for both developed and developing countries were visualized. The graph shows expectedly that most of the developed countries have higher spendings for health related resources relative to developing countries. Since the graph showed no data from developed countries in 2015, all datas in 2015 was removed to avoid bias in analyzing total expenditure based on the status of the countries.

The total expenditure was not normally distributed, so a Spearman correlation test was used. Total expenditure was found to have a weak correlation with life expectancy and mortality rates, with correlation coefficients that ranges from -0.22 to 0.29. The low correlation coefficient predicts that total expenditure does not have a significant, high causal factor for life expectancy. This was confirmed using simple linear regression, showing that total expenditure can only explain 4.77% variability in life expectancy. This low correlation can be due to how the health expenses were converted to actual health care services that people have accessed. This could be influenced by the population, general education about health care, and information access on services in each country.

```
#show distribution of total expenditure in developed and developing countries over the years
ggplot(data = mortality_health, aes(x = Status, y = `Total expenditure`)) + geom_boxplot()  + stat_summa
```



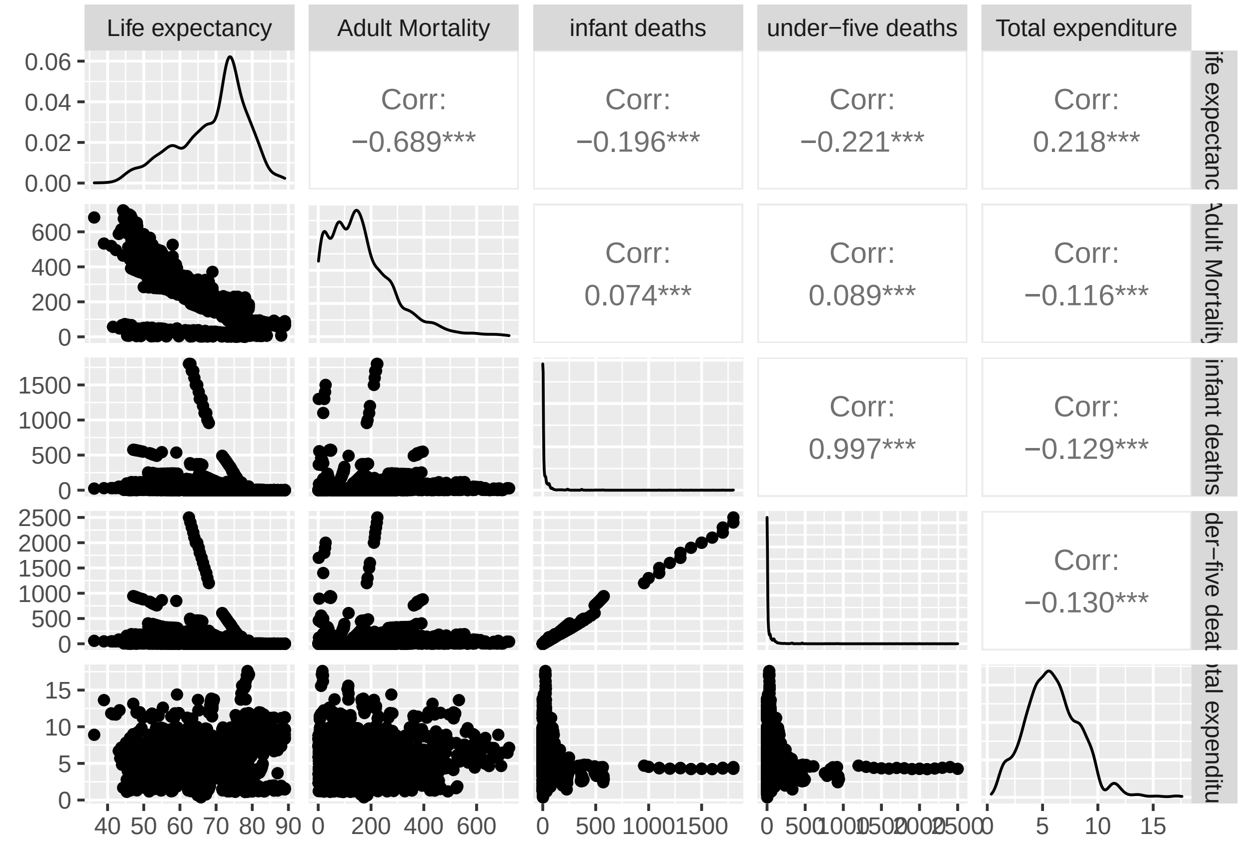
```
#shows no data available for developed countries in 2015

#remove data values in 2015
mortality_health <- mortality_health %>%
                    filter(Year != 2015)

#make df for easy correlation , so remove nonnumerical variables
health_spend <- mortality_health %>%
                  select(-Year, -Status, -Country)

#determine distribution
shapiro.test(mortality_health$`Total expenditure`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mortality_health$‘Total expenditure‘
## W = 0.9788, p-value < 2.2e-16
```

```
#non normal so nonparametric

cor(health_spend, method = "spearman")
```

```
##                      Life expectancy Adult Mortality infant deaths
## Life expectancy            1.0000000      -0.6399910    -0.5940201
## Adult Mortality        -0.6399910       1.0000000     0.3790075
## infant deaths          -0.5940201       0.3790075     1.0000000
## under-five deaths      -0.6117321       0.3919333     0.9930902
## Total expenditure       0.2938650      -0.1755103    -0.2177268
##                    under-five deaths Total expenditure
## Life expectancy           -0.6117321         0.2938650
## Adult Mortality            0.3919333        -0.1755103
## infant deaths              0.9930902        -0.2177268
## under-five deaths          1.0000000        -0.2231802
## Total expenditure         -0.2231802         1.0000000
```

```
#only small correlation between variables so unlikely causation

ggpairs(health_spend)
```



```
#visualize relationships on expenditures and mortality to see if there is improvement in health sector

#perform simple linear regression
health_lm1 <- lm(`Life expectancy` ~ `Total expenditure`, data = mortality_health)
summary(health_lm1) #only 4.77% explained
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘Total expenditure‘, data = mortality_health)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.675  -5.390   2.453   6.374  23.529
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          64.2566     0.4635  138.64   <2e-16 ***
## ‘Total expenditure‘   0.8377     0.0721   11.62   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.303 on 2698 degrees of freedom
## Multiple R-squared:  0.04765,    Adjusted R-squared:  0.0473
## F-statistic:   135 on 1 and 2698 DF,  p-value: < 2.2e-16
```

Life Expectancy and Education

We hypothesized that education will positively affect life expectancy. The box plot showed that developed countries have longer schooling years compared to the developing countries. The Spearman correlation test showed a high correlation coefficient of 0.814, indicating that schooling and life expectancy have a strong positive relationship with each other. Moreover, the simple linear regression demonstrated that schooling years can explain 56.55% of the variation in life expectancy. The equation of the model is as follows: Life expectancy = 44.109 + 2.103(Schooling) + 0.6186 error. Long schooling years is a good indicator of quality education that can help individuals in leading healthy and productive lives in their community.

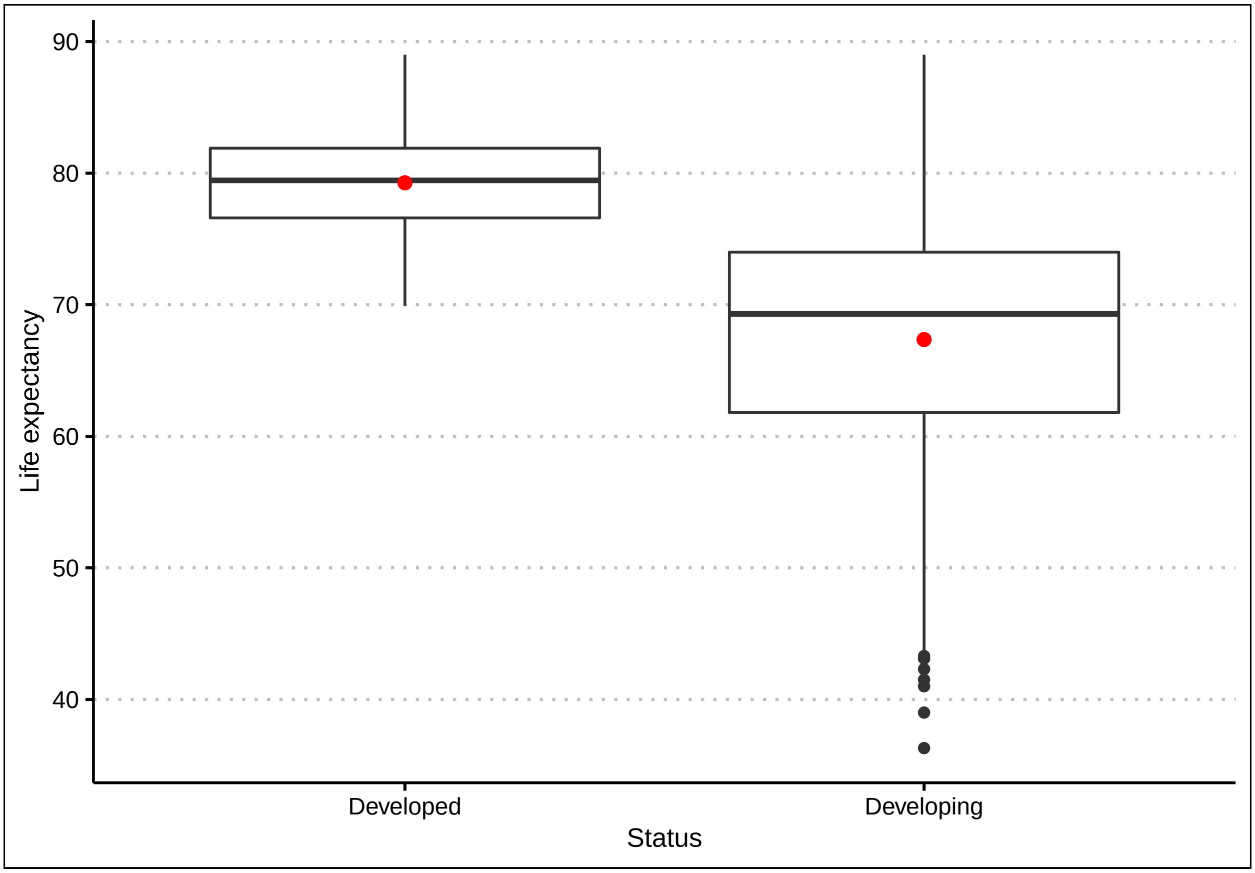
```
#show summary of schooling years and life expectancy in developing and developed countries
developedS <- social %>%
  filter(Status == "Developed") %>%
  select(-Year, -Status)
summary(developedS)
```

```
##    Country          Life expectancy   Schooling
##  Length:464         Min.   :69.90   Min.   :11.50
##  Class :character   1st Qu.:76.60   1st Qu.:14.70
##  Mode  :character   Median :79.45   Median :15.80
##                     Mean   :79.27   Mean   :15.85
##                     3rd Qu.:81.90   3rd Qu.:16.80
##                     Max.   :89.00   Max.   :20.70
```

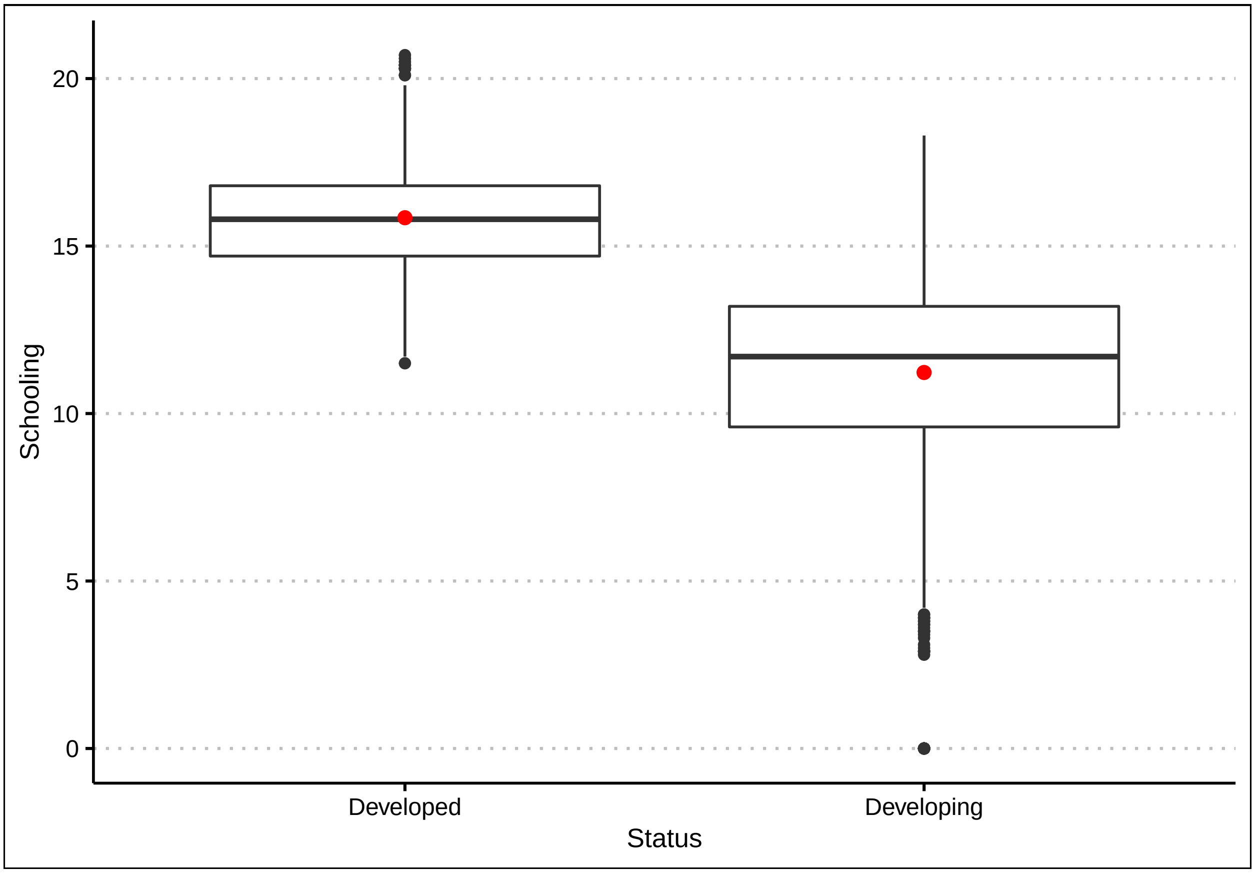
```
developingS <- social %>%
  filter(Status == "Developing") %>%
  select(-Year, -Status)
summary(developingS)
```

```
##    Country          Life expectancy   Schooling
##  Length:2304        Min.   :36.30   Min.   : 0.00
##  Class :character   1st Qu.:61.80   1st Qu.: 9.60
##  Mode  :character   Median :69.30   Median :11.70
##                     Mean   :67.35   Mean   :11.23
##                     3rd Qu.:74.00   3rd Qu.:13.20
##                     Max.   :89.00   Max.   :18.30
```

```
#visualize the summaries
ggplot(data = social, aes(x = Status, y = `Life expectancy`)) + geom_boxplot()  + stat_summary(fun = mea
```



```
ggplot(data = social, aes(x = Status, y = `Schooling`)) + geom_boxplot() + stat_summary(fun = mean, geom
```



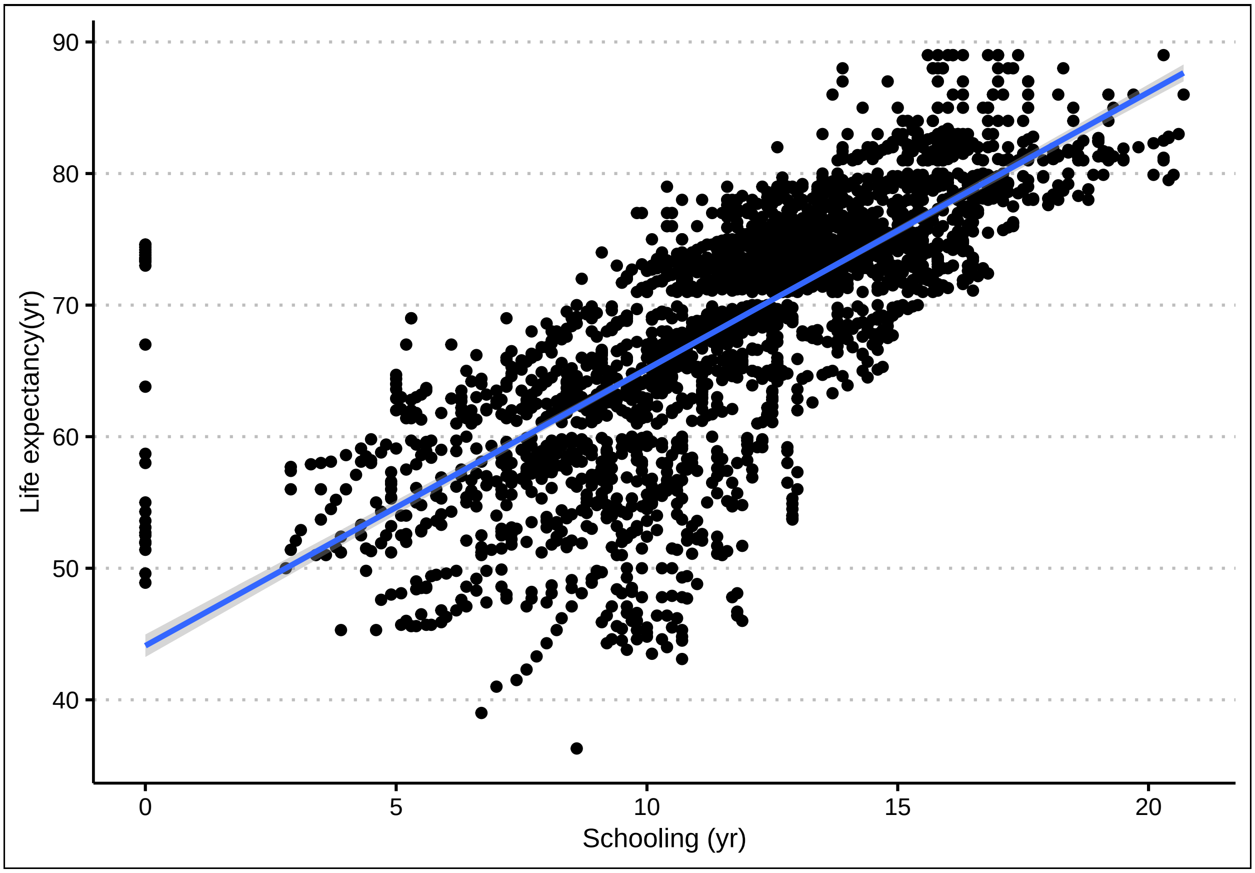
```
#shows that there is a difference in schooling years and life expectancy for developing and developed c

#determine if there is a correlation and causation between schooling and life expectancy
cor.test(social$`Life expectancy`, social$Schooling, method = "spearman")
```

```
## Warning in cor.test.default(social$`Life expectancy`, social$Schooling, : Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  social$`Life expectancy` and social$Schooling
## S = 659068097, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.813541
```

```
#indicates high correlation so can be modeled using linear regression

#visualize the relationship using lm
ggplot(data = social, aes(x = Schooling, y = `Life expectancy`)) + geom_point() + geom_smooth(method =
```

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



```r
#test if there is causation
social_lm <- lm(`Life expectancy` ~ Schooling, data = social)
summary(social_lm)
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ Schooling, data = social)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8986  -2.8210   0.6186   3.8186  30.4911
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.10889    0.43676  100.99   <2e-16 ***
## Schooling    2.10345    0.03506   59.99   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.172 on 2766 degrees of freedom
## Multiple R-squared:  0.5655, Adjusted R-squared:  0.5653
## F-statistic:  3599 on 1 and 2766 DF,  p-value: < 2.2e-16
```

```r
#only 56.55% of the variability in life expectancy can be explained by years of schooling
```

Life Expectancy and Immunization

Immunization for Hepatitis B, Diphtheria, and Polio were reported as immunization coverage among one-year-olds in percentage. The relationship between life expectancy by immunization factor was tested using the Spearman correlation test. All immunization factors showed a positive correlation with life expectancy. An increasing percentage of immunization coverage among one-year-olds, corresponds to increase in life expectancy. Immunization for Polio and Diphtheria showed a moderately high positive correlation with life expectancy while hepatitis B only showed low positive correlation. The relationship between immunization and life expectancy were further analyzed using a multiple linear regression model. The results showed that immunization for diphtheria and polio explains 16% of the variability in life expectancy. The model has an AIC value of 9700.55 and the residuals are normally distributed. Overall, the model has a p-value of 2.2x10-16, affirming that it is statistically significant. This shows that immunization is an important health factor that improves life expectancy by protecting the individual from succumbing to preventable deadly diseases, such as Polio and Diphtheria. The equation of the model is as follows. Life expectancy = 54.838 + 0.0852(Polio) + 0.0929 (Diphtheria) + 1.333 error

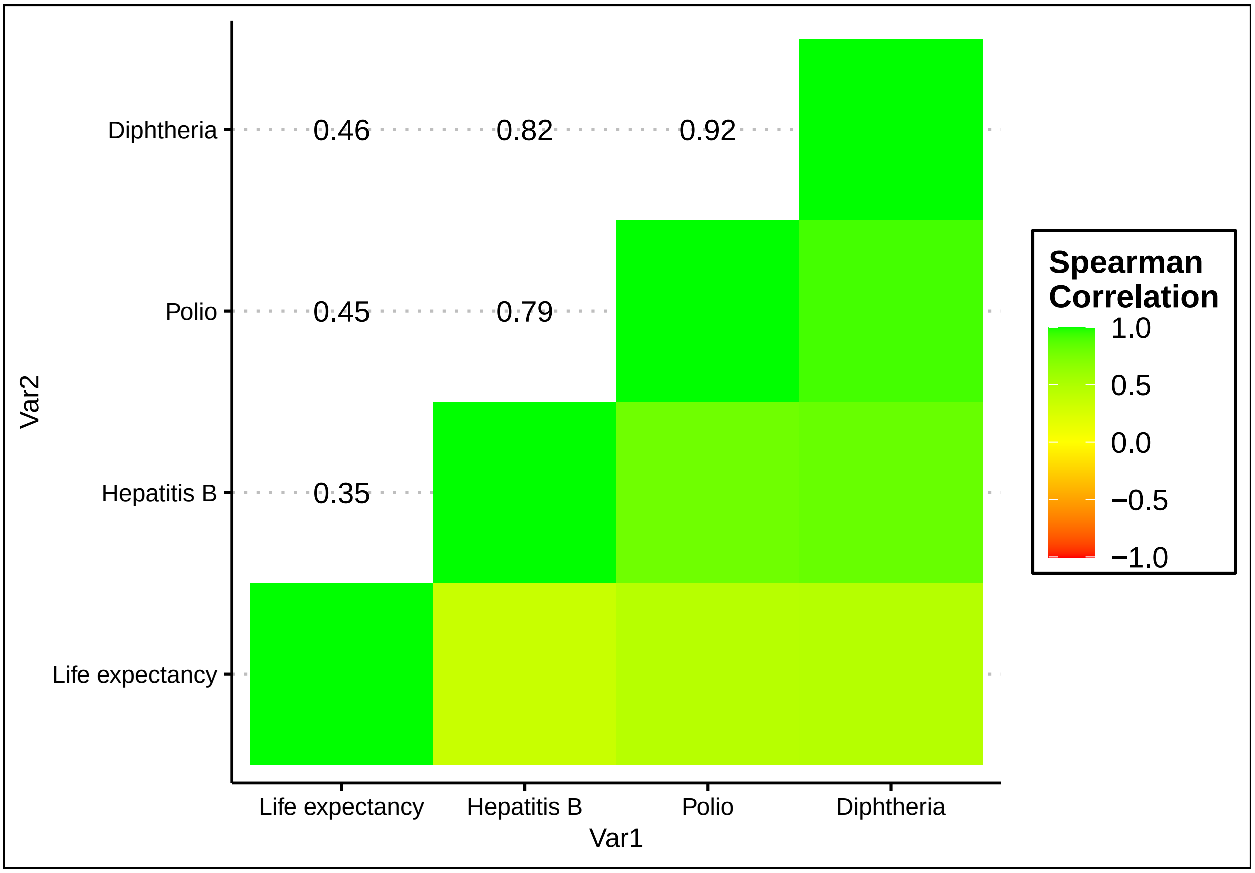
```
# df for life expectancy and different kinds of immunization
expectancy_immunization <- immunization %>%
  select(`Life expectancy`,`Hepatitis B`,Polio,Diphtheria)

# compute correlation
immunization_cormat <- round(cor(expectancy_immunization, method = "spearman"),2)
immunization_cormat[upper.tri(immunization_cormat)] <- NA
head(immunization_cormat)
```

```
##                  Life expectancy Hepatitis B Polio Diphtheria
## Life expectancy            1.00          NA    NA         NA
## Hepatitis B                0.35        1.00    NA         NA
## Polio                      0.45        0.79  1.00         NA
## Diphtheria                 0.46        0.82  0.92          1
```

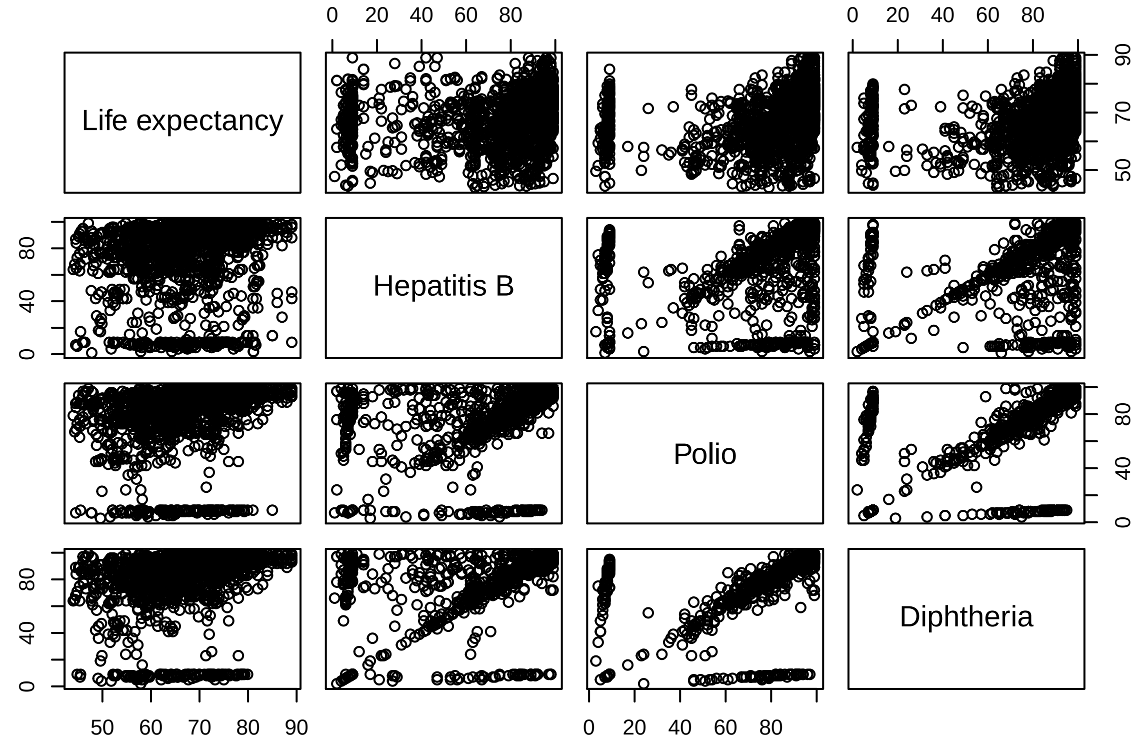
```
# Hepatitis B, Polio, and Diphtheria shows a slight correlation

# reshape correlation
immunization_melted_cormat <- melt(immunization_cormat, na.rm = TRUE)
head(immunization_melted_cormat)
```

```
##              Var1            Var2 value
## 1 Life expectancy Life expectancy  1.00
## 2     Hepatitis B Life expectancy  0.35
## 3           Polio Life expectancy  0.45
## 4      Diphtheria Life expectancy  0.46
## 6     Hepatitis B     Hepatitis B  1.00
## 7           Polio     Hepatitis B  0.79
```

```
# plotting melted cormat
ggplot(immunization_melted_cormat, aes(x = Var1, y = Var2, fill = value)) +
  scale_fill_gradient2(low = 'red', mid = 'yellow', high = 'green', limit = c(-1,1), midpoint = 0, name
  geom_text(aes(Var2,Var1,label = value), color = "black",size = 4) +
  geom_tile() + theme_clean()
```



```
# multiple linear regression model for immunization
plot(expectancy_immunization)
```



```
immunization_lm <- lm(`Life expectancy` ~., data = expectancy_immunization)
summary(immunization_lm)
```

```
##
## Call:
## lm(formula = `Life expectancy` ~ ., data = expectancy_immunization)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.111  -4.391   1.325   4.707  23.497
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.794213   0.723601  75.724   <2e-16 ***
## `Hepatitis B`  0.003428   0.008140   0.421    0.674
## Polio          0.084413   0.009132   9.244   <2e-16 ***
## Diphtheria     0.090905   0.010241   8.877   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.704 on 2371 degrees of freedom
## Multiple R-squared:  0.1649, Adjusted R-squared:  0.1639
## F-statistic: 156.1 on 3 and 2371 DF,  p-value: < 2.2e-16
```

```
backward_immunization <- step(immunization_lm, direction = "backward", scope = formula(immunization_lm)
```

```
## Start:  AIC=9702.37
## ‘Life expectancy‘ ~ ‘Hepatitis B‘ + Polio + Diphtheria
##
##                 Df Sum of Sq    RSS    AIC
## - ‘Hepatitis B‘  1      10.5 140740 9700.6
## <none>                       140729 9702.4
## - Diphtheria     1    4676.9 145406 9778.0
## - Polio          1    5072.0 145801 9784.5
##
## Step:  AIC=9700.55
## ‘Life expectancy‘ ~ Polio + Diphtheria
##
##              Df Sum of Sq    RSS    AIC
## <none>                    140740 9700.6
## - Polio       1    5376.8 146117 9787.6
## - Diphtheria  1    6203.6 146944 9801.0
```

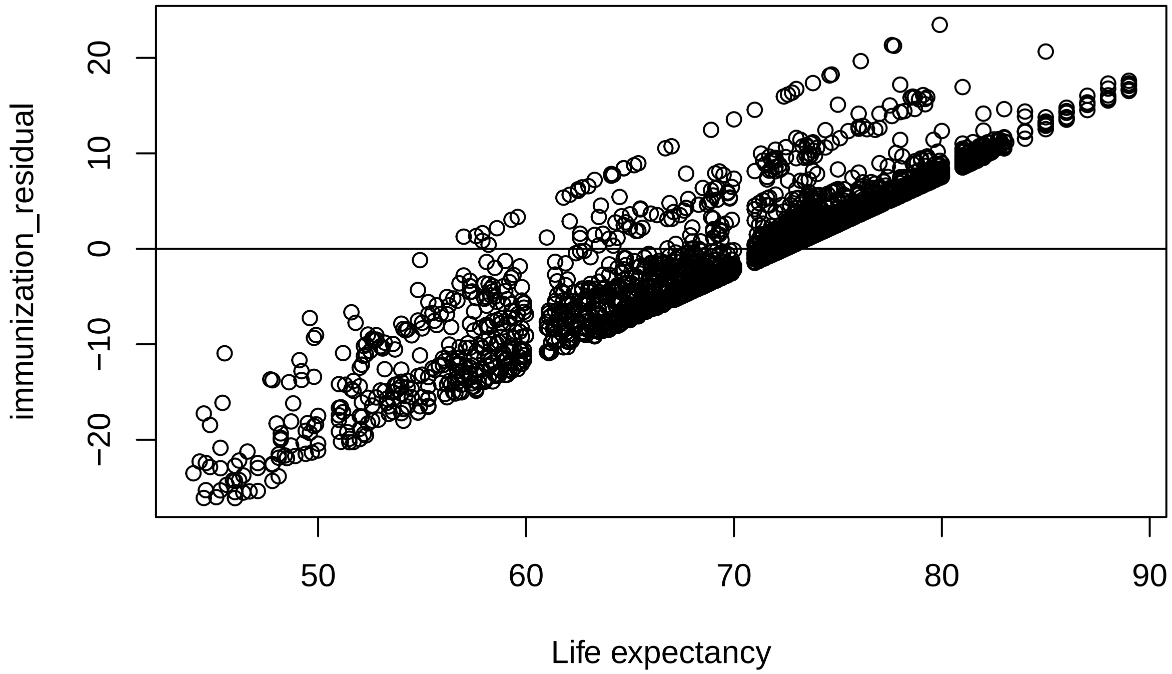
```
immune_lm <-lm(`Life expectancy` ~ Polio + Diphtheria, data = expectancy_immunization)
immune_lm
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ Polio + Diphtheria, data = expectancy_immunization)
##
## Coefficients:
## (Intercept)        Polio   Diphtheria
##    54.83801      0.08518      0.09289
```

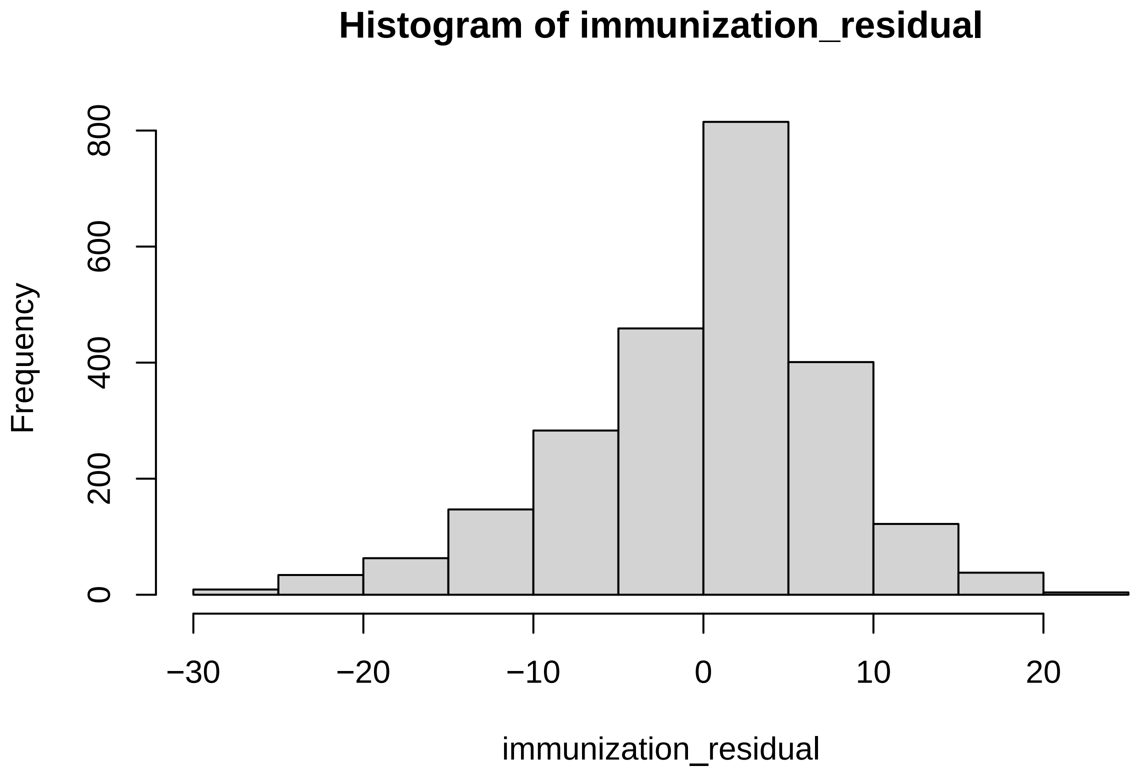
```
summary(immune_lm)
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ Polio + Diphtheria, data = expectancy_immunization)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.111  -4.405   1.333   4.717  23.459
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.838007   0.715964  76.593   <2e-16 ***
## Polio        0.085178   0.008948   9.519   <2e-16 ***
## Diphtheria   0.092894   0.009085  10.225   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.703 on 2372 degrees of freedom
## Multiple R-squared:  0.1649, Adjusted R-squared:  0.1642
## F-statistic: 234.1 on 2 and 2372 DF,  p-value: < 2.2e-16
```

```
# Only the immunization for Diphtheria and Polio shows a significant relationship with Life expectancy

# check the residuals of the model
immunization_residual <- residuals(backward_immunization)
plot(immunization_residual  ~ `Life expectancy`, data = expectancy_immunization)
abline(h = 0)
```



```
hist(immunization_residual) #showing normality
```



Life Expectancy and Economic Factors

There are four predictor variables classified as part of the economic factors, namely, percentage expenditure reported as expenditure on health as percentage of GDP per capita, total expenditure reported as general government expenditure on health as a percentage of total government expenditure, GDP reported in USD, and income composition of resources reported as Human Development Index ranging from 0 to 1. All of the economic factors exhibit a positive correlation with life expectancy using the Spearman correlation. Income composition of resources shows a very strong correlation to life expectancy with a correlation coefficient of 0.91. Percentage expenditure (r = 0.65) and GDP (r = 0.64) has a strong correlation to life expectancy. On the other hand, total expenditure (r = 0.26) is only weakly correlated with life expectancy. Additionally, the multiple linear regression results showed that all economic factors except for total expenditure has a significant relationship with life expectancy. The model explains 79% of the variability in life expectancy and is significant (p = 2.2x 10-16). Moreover, the model has an AIC value of 6587.29 and the residuals show a normal distribution. This demonstrates the critical role of the country's economy in ensuring good, quality life to its population by providing enough goods to meet the basic needs and providing high quality healthcare and public services. The equation of the model is as follows. Life expectancy = 35.11 + 0.00042 (percentage expenditure) - 0.000074(GDP) + 52.43(Income composition of resources) + 0.3022 error

```
# df for life expectancy and economical factors
expectancy_economy <- economical %>%
  select(`Life expectancy`,`percentage expenditure`,`Total expenditure`,GDP,
         `Income composition of resources`) %>%
  filter(`Income composition of resources` > 0)
```

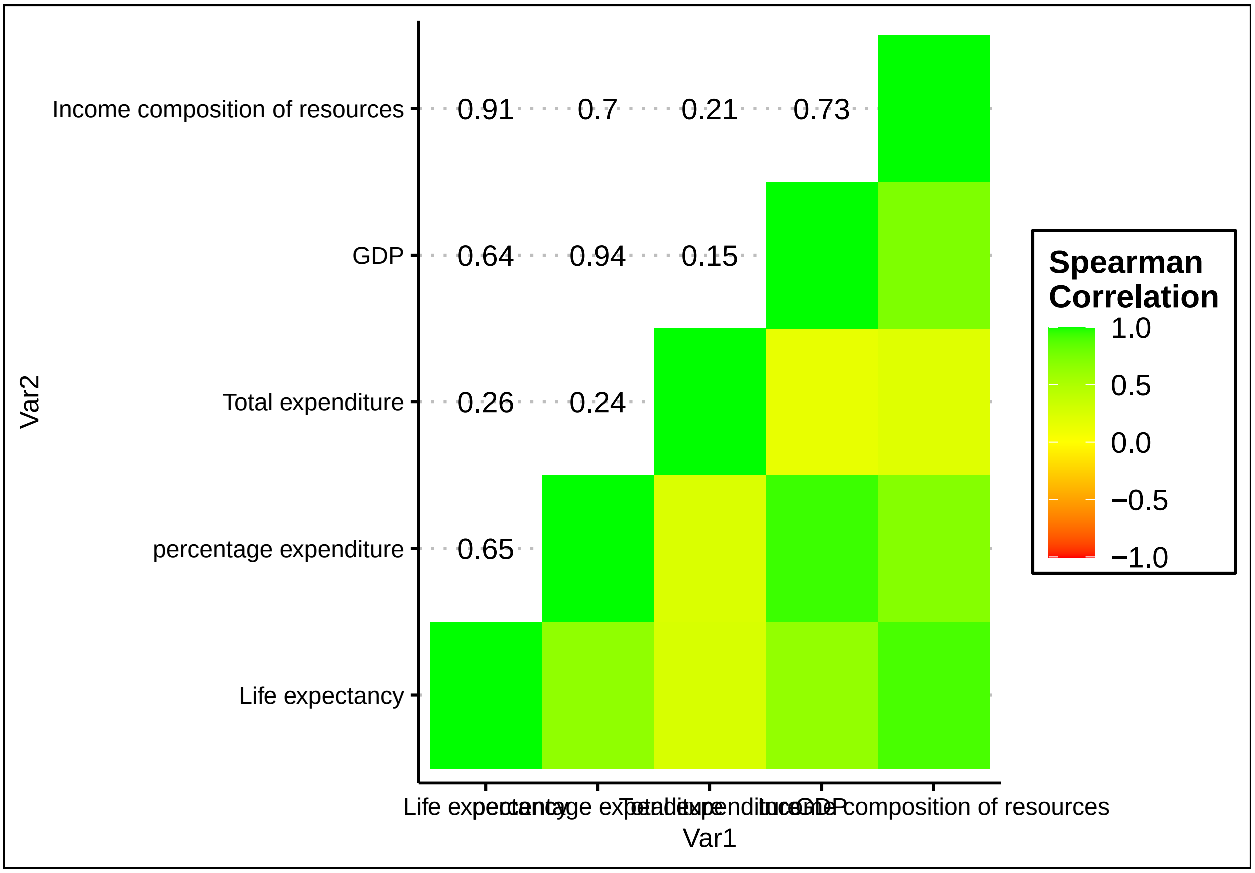
```
# compute correlation
economy_cormat <- round(cor(expectancy_economy, method = "spearman"),2)
economy_cormat[upper.tri(economy_cormat)] <- NA
economy_cormat
```

```
##                                 Life expectancy percentage expenditure
## Life expectancy                            1.00                     NA
## percentage expenditure                     0.65                   1.00
## Total expenditure                          0.26                   0.24
## GDP                                        0.64                   0.94
## Income composition of resources            0.91                   0.70
##                                 Total expenditure  GDP
## Life expectancy                                NA   NA
## percentage expenditure                         NA   NA
## Total expenditure                            1.00   NA
## GDP                                          0.15 1.00
## Income composition of resources              0.21 0.73
##                                 Income composition of resources
## Life expectancy                                              NA
## percentage expenditure                                       NA
## Total expenditure                                            NA
## GDP                                                          NA
## Income composition of resources                               1
```

```
# Percentage expenditure, GDP, and income composition of resources shows a high correlation with life e

# reshape correlation
economy_melted_cormat <- melt(economy_cormat, na.rm = TRUE)
head(economy_melted_cormat)
```

```
##                               Var1                   Var2 value
## 1                  Life expectancy        Life expectancy  1.00
## 2           percentage expenditure        Life expectancy  0.65
## 3                Total expenditure        Life expectancy  0.26
## 4                              GDP        Life expectancy  0.64
## 5  Income composition of resources        Life expectancy  0.91
## 7           percentage expenditure percentage expenditure  1.00
```

```
# plotting melted cormat
ggplot(economy_melted_cormat, aes(x = Var1, y = Var2, fill = value)) +
  scale_fill_gradient2(low = 'red', mid = 'yellow', high = 'green', limit = c(-1,1), midpoint = 0, name
  geom_text(aes(Var2,Var1,label = value), color = "black",size = 4)+
  geom_tile() + theme_clean()
```



```
# multiple linear regression model for immunization
economy_lm <- lm(`Life expectancy` ~., data = expectancy_economy)
summary(economy_lm)
```

```
##
## Call:
## lm(formula = `Life expectancy` ~ ., data = expectancy_economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.5664  -2.0487   0.3542   2.6315  15.9991
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.492e+01  4.563e-01  76.527  < 2e-16 ***
## `percentage expenditure`          3.936e-04  1.115e-04   3.531 0.000423 ***
## `Total expenditure`               4.333e-02  4.060e-02   1.067 0.285962
## GDP                              -7.069e-05  1.744e-05  -4.054 5.21e-05 ***
## `Income composition of resources` 5.233e+01  6.705e-01  78.046  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.411 on 2213 degrees of freedom
## Multiple R-squared:  0.7939, Adjusted R-squared:  0.7935
## F-statistic:  2131 on 4 and 2213 DF,  p-value: < 2.2e-16
```

```
economy_lm1 <- lm(`Life expectancy` ~ `percentage expenditure` + GDP + `Income composition of resources
summary(economy_lm1)
```

```
##
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘percentage expenditure‘ + GDP +
##     ‘Income composition of resources‘, data = expectancy_economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4171  -2.0305   0.3022   2.6469  16.0302
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        3.511e+01  4.193e-01  83.738  < 2e-16 ***
## ‘percentage expenditure‘           4.199e-04  1.087e-04   3.863 0.000115 ***
## GDP                               -7.399e-05  1.716e-05  -4.311 1.69e-05 ***
## ‘Income composition of resources‘  5.243e+01  6.644e-01  78.911  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.411 on 2214 degrees of freedom
## Multiple R-squared:  0.7938, Adjusted R-squared:  0.7935
## F-statistic:  2840 on 3 and 2214 DF,  p-value: < 2.2e-16
```

```
# [economy_lm1], percentage expenditure, GDP, and income composition explains 79% of the variability in
backward_economy <- step(economy_lm, direction = "backward", scope = formula(economy_lm))
```

```
## Start:  AIC=6588.15
## ‘Life expectancy‘ ~ ‘percentage expenditure‘ + ‘Total expenditure‘ +
##     GDP + ‘Income composition of resources‘
##
##                                     Df Sum of Sq    RSS    AIC
## - ‘Total expenditure‘                1        22  43074 6587.3
## <none>                                            43052 6588.2
## - ‘percentage expenditure‘           1       243  43295 6598.6
## - GDP                                1       320  43372 6602.6
## - ‘Income composition of resources‘  1    118499 161551 9519.3
##
## Step:  AIC=6587.29
## ‘Life expectancy‘ ~ ‘percentage expenditure‘ + GDP + ‘Income composition of resources‘
##
##                                     Df Sum of Sq    RSS    AIC
## <none>                                            43074 6587.3
## - ‘percentage expenditure‘           1       290  43365 6600.2
## - GDP                                1       362  43436 6603.8
## - ‘Income composition of resources‘  1    121146 164221 9553.6
```

```
economy_lm <- lm(`Life expectancy` ~ `percentage expenditure` + GDP + `Income composition of resources`
economy_lm
```

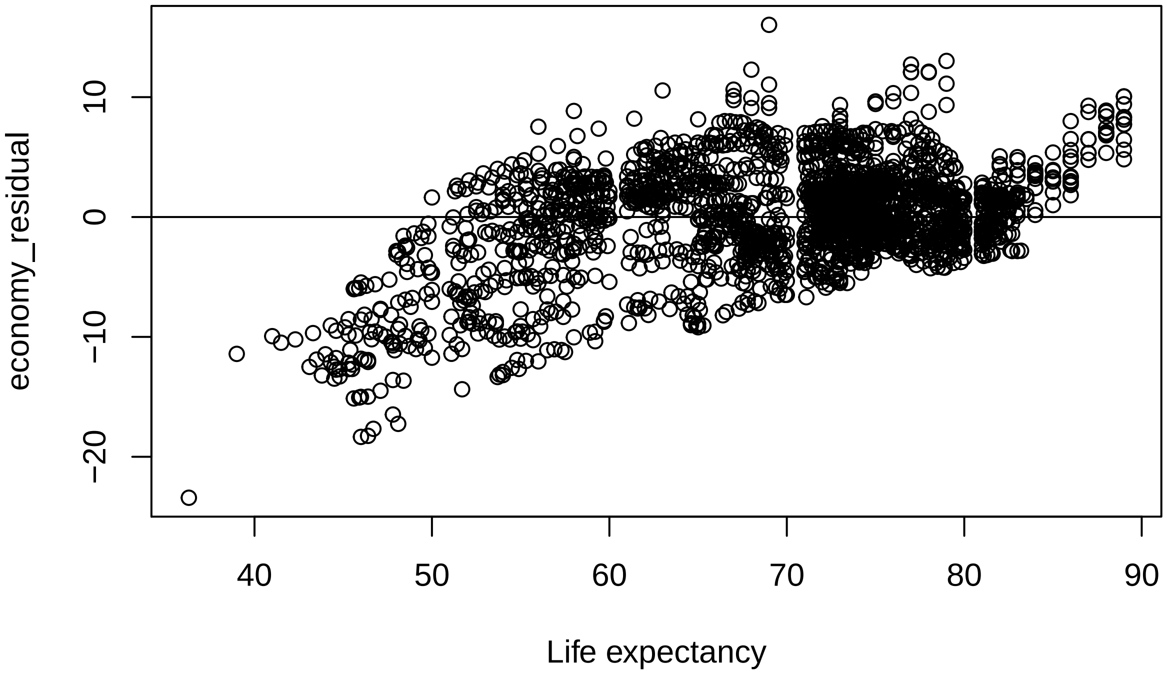
```
##
```

```
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘percentage expenditure‘ + GDP + 
##     ‘Income composition of resources‘, data = expectancy_economy)
## 
## Coefficients:
##                       (Intercept)           ‘percentage expenditure‘  
##                         3.511e+01                          4.199e-04  
##                               GDP  ‘Income composition of resources‘  
##                        -7.399e-05                          5.243e+01  
```

```
summary(economy_lm)
```

```
## 
## Call:
## lm(formula = ‘Life expectancy‘ ~ ‘percentage expenditure‘ + GDP + 
##     ‘Income composition of resources‘, data = expectancy_economy)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -23.4171  -2.0305   0.3022   2.6469  16.0302 
## 
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                        3.511e+01  4.193e-01  83.738  < 2e-16 ***
## ‘percentage expenditure‘           4.199e-04  1.087e-04   3.863 0.000115 ***
## GDP                               -7.399e-05  1.716e-05  -4.311 1.69e-05 ***
## ‘Income composition of resources‘  5.243e+01  6.644e-01  78.911  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 4.411 on 2214 degrees of freedom
## Multiple R-squared:  0.7938, Adjusted R-squared:  0.7935 
## F-statistic:  2840 on 3 and 2214 DF,  p-value: < 2.2e-16
```

```
#check the residuals of the model
economy_residual <-residuals(backward_economy)
plot(economy_residual  ~ `Life expectancy`, data = expectancy_economy)
abline(h = 0)
```



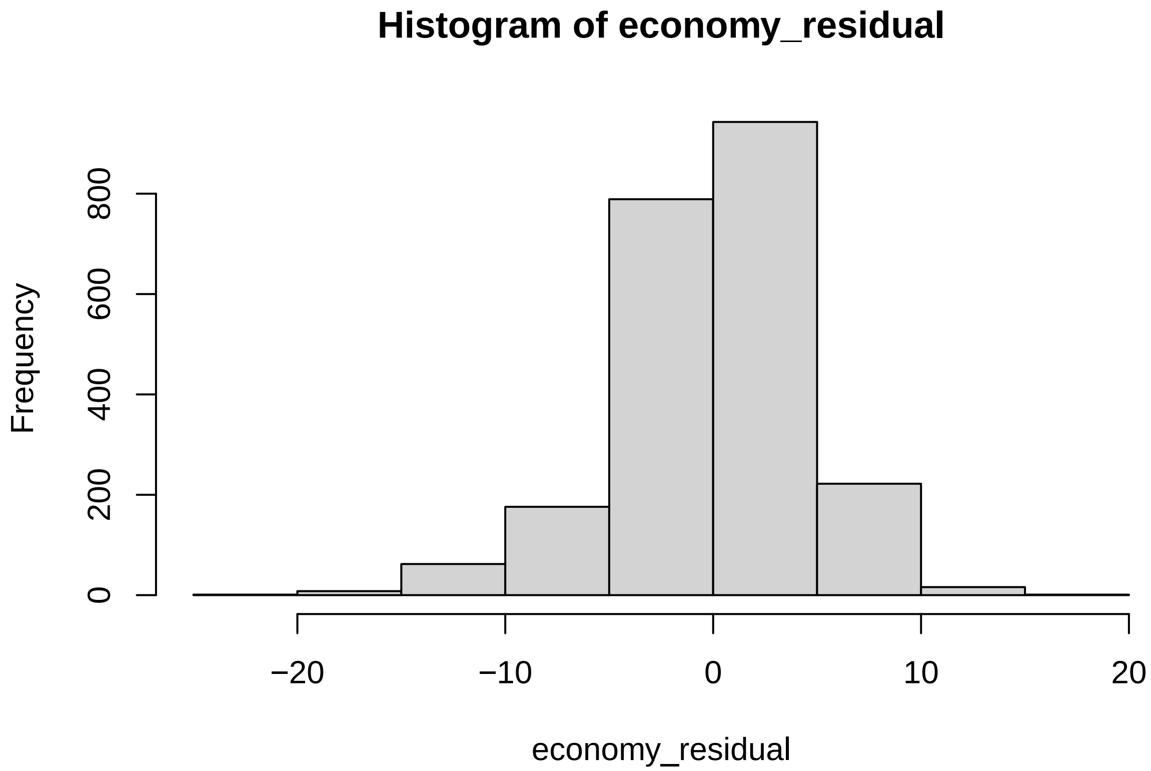
```
hist(economy_residual,bins = 5) #skewed to the right
```

```
## Warning in plot.window(xlim, ylim, "", ...): "bins" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "bins"
## is not a graphical parameter
```

```
## Warning in axis(1, ...): "bins" is not a graphical parameter
```

```
## Warning in axis(2, ...): "bins" is not a graphical parameter
```



References

Bezy, J. Marie. 2020. Life expectancy. Encyclopedia Britannica. https://www.britannica.com/science/life-expectancy

Murray, C.J., 1988. The infant mortality rate, life expectancy at birth, and a linear index of mortality as measures of general health status. International journal of epidemiology, 17(1), pp.122-128.