Factors Influening Life Expectancy

Ezra Abah

01/06/2021

1.0 Problem Definition

The World Health Organization (WHO) requires analysis to be carried out on data gathered of different countries from 2000 to 2015. The analysis needed to help them understand factors that affect life expectancy. In understanding these factor, recommendations can then be made to governments to help them improve life expectancy.

2.0 Dataset and dataset preview

The data set contains economic, health care, immunization data as well as other data. It contains 19 rows and 2,938 columns. This section imports the data and gives a general overview of it.

```
# Import libraries required for analysis

library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
# install.packages("zoo")
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
# Import data and save it in a dataframe "data"
data <- read_csv ('life_expectancy.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## 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(),
##   schooling = col_double()
## )
```

```r
# Show top six rows
head(data)
```

```
## # A tibble: 6 x 19
##   country    year status   life_expectancy adult_mortality infant_deaths alcohol
##   <chr>     <dbl> <chr>              <dbl>           <dbl>         <dbl>   <dbl>
## 1 Afghanis~  2015 Develop~            65               263            62    0.01
## 2 Afghanis~  2014 Develop~            59.9             271            64    0.01
## 3 Afghanis~  2013 Develop~            59.9             268            66    0.01
## 4 Afghanis~  2012 Develop~            59.5             272            69    0.01
## 5 Afghanis~  2011 Develop~            59.2             275            71    0.01
## 6 Afghanis~  2010 Develop~            58.8             279            74    0.01
## # ... with 12 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>, schooling <dbl>
```

```r
# Show bottom six rows
tail(data)
```

```
## # A tibble: 6 x 19
##   country    year status    life_expectancy adult_mortality infant_deaths alcohol
##   <chr>     <dbl> <chr>               <dbl>           <dbl>         <dbl>   <dbl>
## 1 Zimbabwe  2005 Developi~            44.6             717            28    4.14
## 2 Zimbabwe  2004 Developi~            44.3             723            27    4.36
```

```
## 3 Zimbabwe  2003 Developi~            44.5             715            26    4.06
## 4 Zimbabwe  2002 Developi~            44.8              73            25    4.43
## 5 Zimbabwe  2001 Developi~            45.3             686            25    1.72
## 6 Zimbabwe  2000 Developi~            46               665            24    1.68
## # ... with 12 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>, schooling <dbl>
```

```
str(data)
```

```
## spec_tbl_df [2,938 x 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ country               : chr [1:2938] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ 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 ...
##  $ 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(),
##   ..   schooling = col_double()
##   .. )
```

This shows that all 19 columns and some of their characteristics. It should be noted that it shows that all columns except two (country and status) are numerical. Measures will be taken to convert them to numerical values in preprocessing session.

3.0 Data Cleaning and preprocessing

3.1 Check and Removal of duplicate rows

```
# Check and removal of duplicate rows
## Check for duplicate rows(s)
sum(duplicated(data))
```

```
## [1] 0
```

There are no duplicate rows

3.2 Check and handling of missing data

To handle null values, Rows of null values from target variable "life_expectancy" ar all removed. Rows of null values will also be removed for features with less than 5% null values as the removal will not severely reduce number of rows For features with null values greater than 5% but less than 30%, values will be estimated by interpolation of already available data. This is done using zoo library. With features with more than 30% null values, remove feature

```
# Check and handling of missing data
colSums(is.na(data))
```

```
##                country                   year                 status
##                      0                      0                      0
##        life_expectancy        adult_mortality          infant_deaths
##                     10                     10                      0
##                alcohol percentage_expenditure            hepatitis_b
##                    194                      0                    553
##                measles                    bmi      under_five_deaths
##                      0                     34                      0
##                  polio      total_expenditure             diphtheria
##                     19                    226                     19
##               hiv_aids                    gdp             population
##                      0                    448                    652
##              schooling
##                    163
```

```
sum(is.na(data))
```

```
## [1] 2328
```

There are 2328 null entries.

Target variable, life_expectancy

```
# find missing cells in life_expectancy column
data[which(is.na(data$life_expectancy)),]
```

```
## # A tibble: 10 x 19
##    country     year status life_expectancy adult_mortality infant_deaths alcohol
##    <chr>      <dbl> <chr>            <dbl>           <dbl>         <dbl>   <dbl>
##  1 Cook Isla~  2013 Devel~              NA              NA             0    0.01
##  2 Dominica    2013 Devel~              NA              NA             0    0.01
##  3 Marshall ~  2013 Devel~              NA              NA             0    0.01
##  4 Monaco      2013 Devel~              NA              NA             0    0.01
##  5 Nauru       2013 Devel~              NA              NA             0    0.01
##  6 Niue        2013 Devel~              NA              NA             0    0.01
##  7 Palau       2013 Devel~              NA              NA             0   NA
##  8 Saint Kit~  2013 Devel~              NA              NA             0    8.54
##  9 San Marino  2013 Devel~              NA              NA             0    0.01
## 10 Tuvalu      2013 Devel~              NA              NA             0    0.01
## # ... with 12 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>, schooling <dbl>
```

All 10 missing values of life expectancy and adult mortality are from 10 countries which data points are un-available for only one year. So we can remove these data points

```
# find missing cells in life_expectancy column
data <- filter(data, !is.na(life_expectancy))
sum(is.na(data$life_expectancy))
```

```
## [1] 0
```

```
# Check percentage of missing data per cell
round((colSums(is.na(data))/nrow(data)*100),2)
```

```
##                country                     year                   status
##                   0.00                     0.00                     0.00
##        life_expectancy           adult_mortality            infant_deaths
##                   0.00                     0.00                     0.00
##                alcohol percentage_expenditure              hepatitis_b
##                   6.59                     0.00                    18.89
##                measles                      bmi       under_five_deaths
##                   0.00                     1.09                     0.00
##                  polio      total_expenditure              diphtheria
##                   0.65                     7.72                     0.65
##               hiv_aids                      gdp               population
##                   0.00                    15.13                    21.99
##              schooling
##                   5.46
```

Remove rows with missing values for bmi, polio, diptheria as they have less than 5% null values

```
# Remove rows with empty life_expectancy cells
data <- data %>%
  filter(!is.na(bmi), !is.na(polio), !is.na(diphtheria))
dim(data)
```

```
## [1] 2888   19
```

```
sum(is.na(data))
```

```
## [1] 2151
```

```
colSums(is.na(data))
```

```
##                country                   year                 status
##                      0                      0                      0
##        life_expectancy        adult_mortality          infant_deaths
##                      0                      0                      0
##                alcohol percentage_expenditure            hepatitis_b
##                    175                      0                    525
##                measles                    bmi      under_five_deaths
##                      0                      0                      0
##                  polio      total_expenditure             diphtheria
##                      0                    212                      0
##               hiv_aids                    gdp             population
##                      0                    435                    644
##             schooling
##                    160
```

```
data$alcohol=na.approx(data$alcohol)
data$hepatitis_b=na.approx(data$hepatitis_b)
data$total_expenditure=na.approx(data$total_expenditure)
data$gdp=na.approx(data$gdp)
data$population=na.approx(data$population)
data$schooling=na.approx(data$schooling)
```

```
colSums(is.na(data))
```

```
##                country                   year                 status
##                      0                      0                      0
##        life_expectancy        adult_mortality          infant_deaths
##                      0                      0                      0
##                alcohol percentage_expenditure            hepatitis_b
##                      0                      0                      0
##                measles                    bmi      under_five_deaths
##                      0                      0                      0
##                  polio      total_expenditure             diphtheria
##                      0                      0                      0
##               hiv_aids                    gdp             population
##                      0                      0                      0
##             schooling
##                      0
```

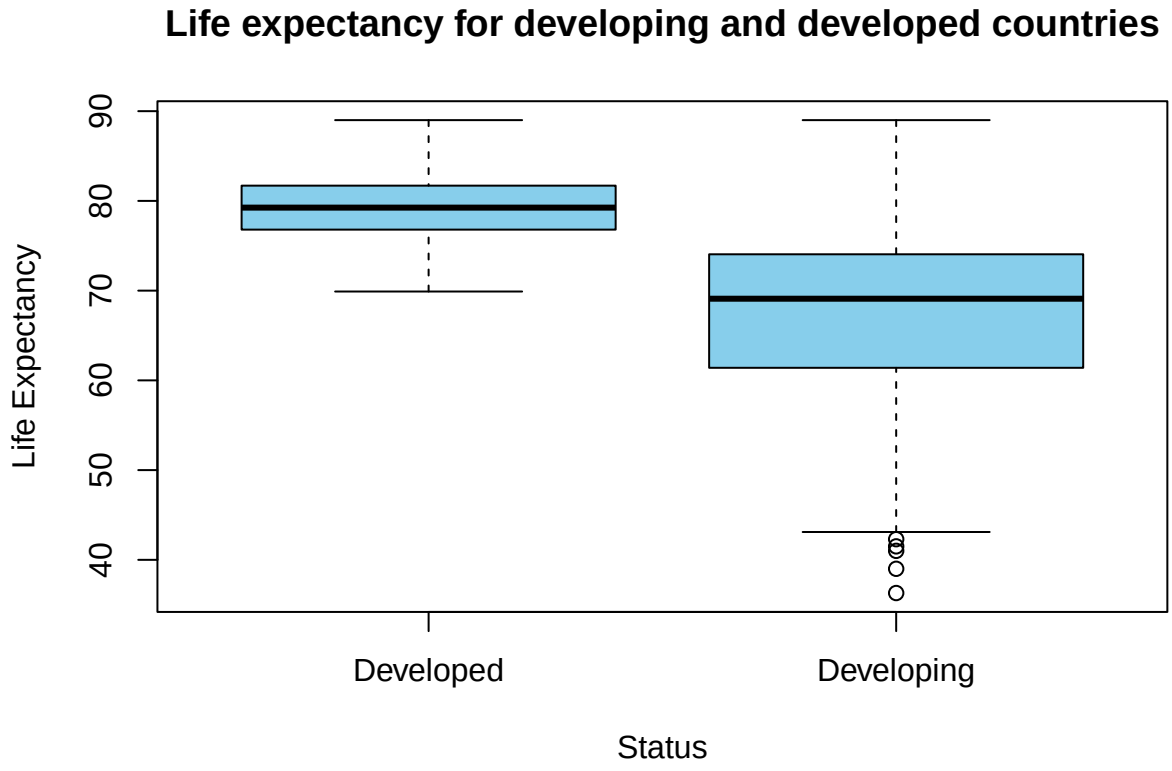
No features have more than 30% null values

3.3 Convert strings to numerical variables

```
data$country <- as.factor(data$country)
data$status <- as.factor(data$status)
```

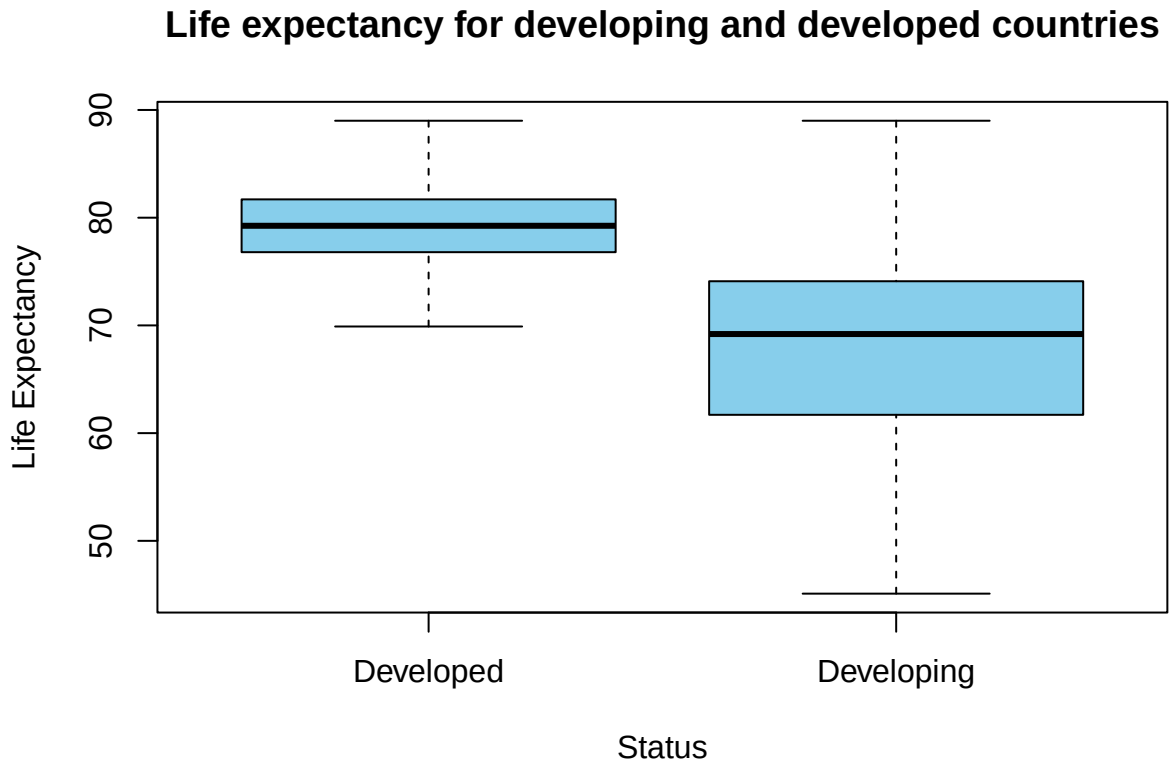
3.4 Outlier handling

```
#Outlier handling

boxplot(life_expectancy~status,
data=data,
main="Life expectancy for developing and developed countries",
xlab="Status",
ylab="Life Expectancy",
col="skyblue",
border="black"
)
```




```
#Outlier handling
data <- data[data$life_expectancy > quantile(data$life_expectancy, .25) - 1.5*IQR(data$life_expectancy)
             data$life_expectancy < quantile(data$life_expectancy, .75) + 1.5*IQR(data$life_expectanc
```

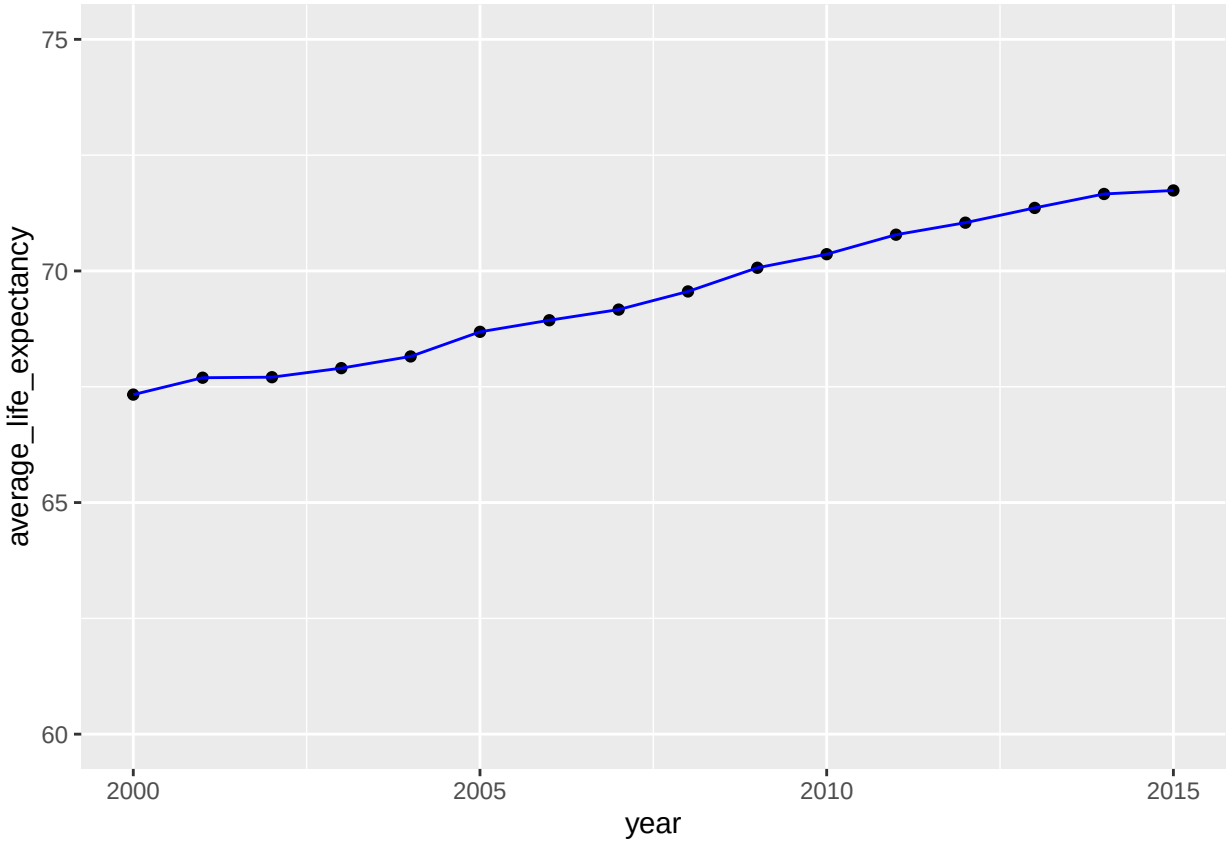
```
#Outlier handling

boxplot(life_expectancy~status,
data=data,
main="Life expectancy for developing and developed countries",
xlab="Status",
ylab="Life Expectancy",
col="skyblue",
border="black"
)
```



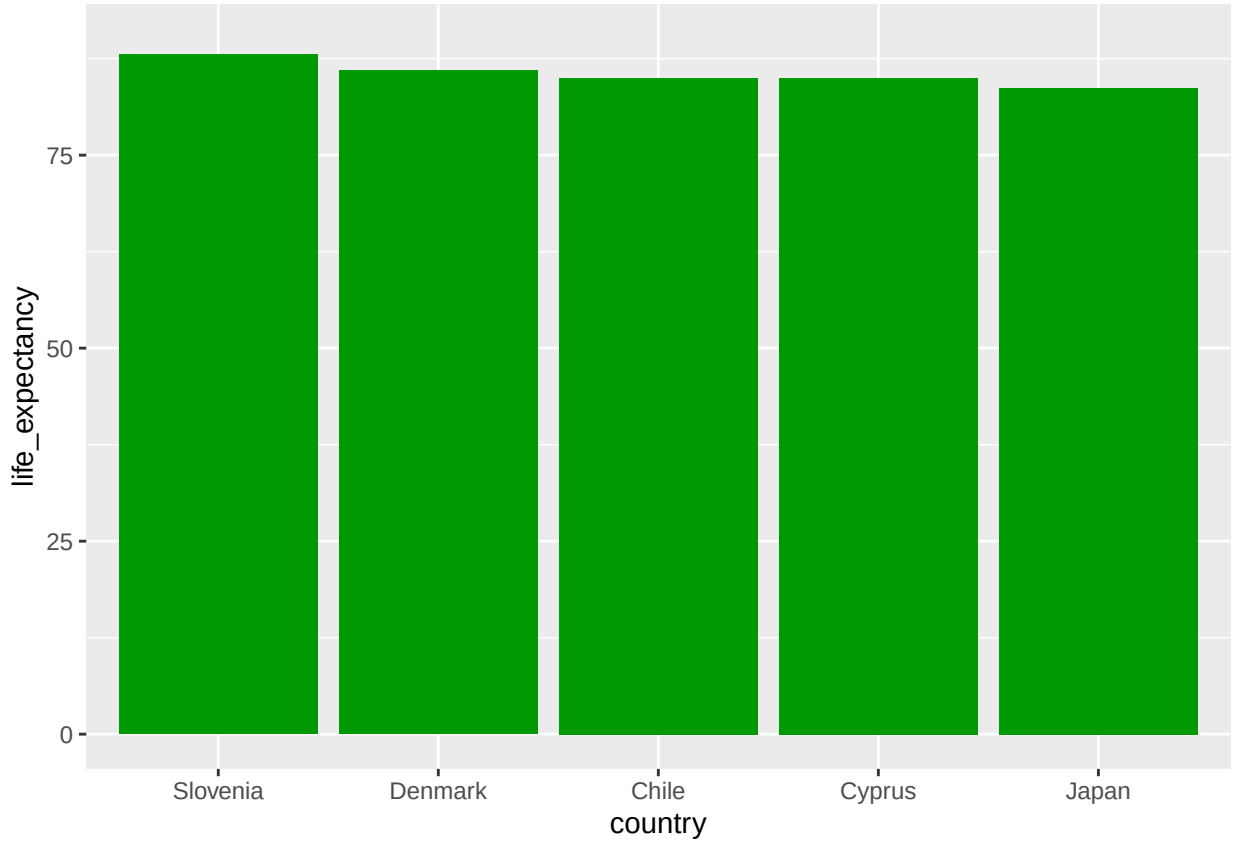

```
# View dimension of data
dim(data)
```

```
## [1] 2869   19
```

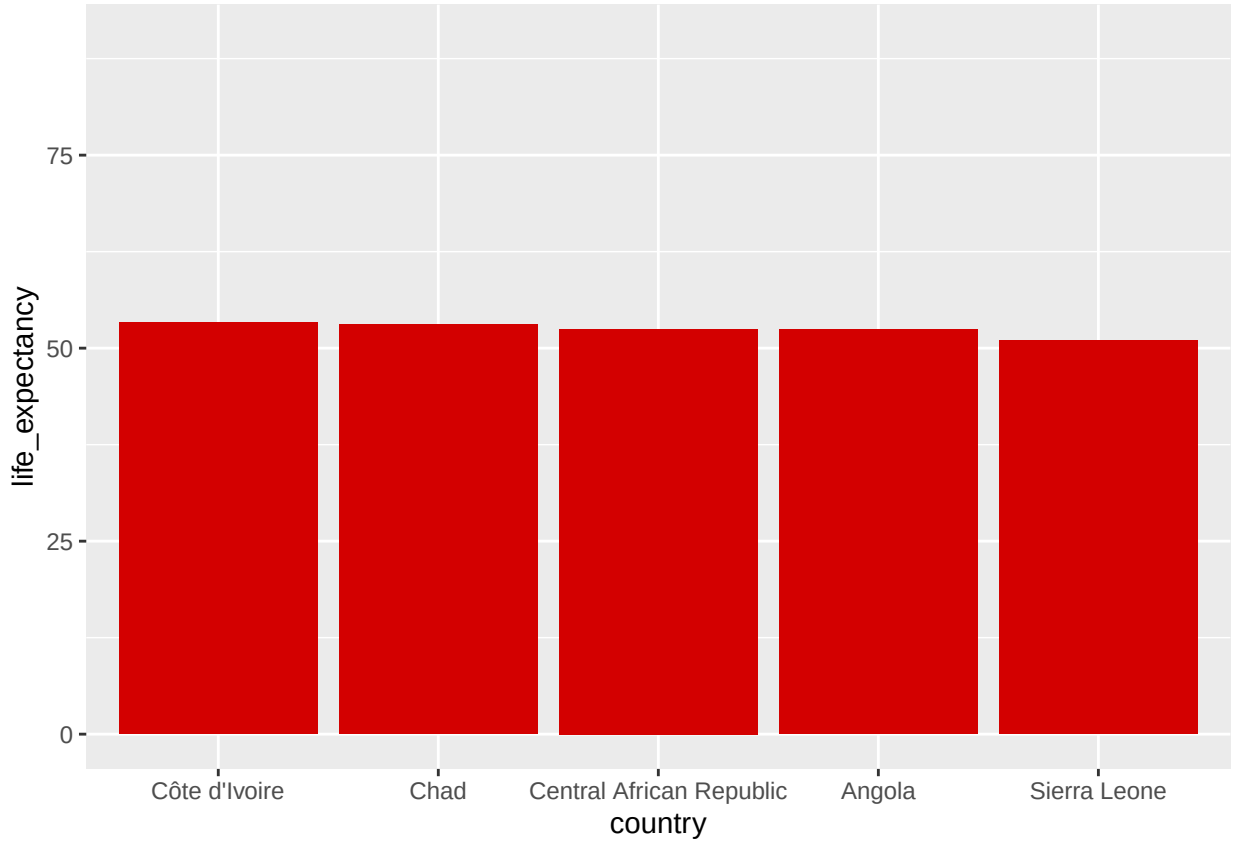
```
# Average life expectancy per year
data %>%
  group_by(year) %>%
  summarise (average_life_expectancy = mean(life_expectancy, na.rm = TRUE))%>%
  arrange(year) %>%
  ggplot(aes(x=year, y= average_life_expectancy)) +
  geom_point()+
  geom_line(color='blue')+
  ylim(60,75)
```




```
# 2015 Life expectancies, top 5
data %>%
  filter(year == 2015) %>%
  top_n(5, life_expectancy) %>%
  arrange(desc(life_expectancy)) %>%
  mutate(country= factor(country, levels = unique(country))) %>%
  ggplot(aes(country,life_expectancy)) +
  geom_col(fill='#009904')+
  ylim(0,90)
```




```
# 2015 Life expectancies, bottom 5
data %>%
  filter(year == 2015) %>%
  top_n(-5, life_expectancy) %>%
  arrange(desc(life_expectancy)) %>%
  mutate(country= factor(country, levels = unique(country))) %>%
  ggplot(aes(country,life_expectancy)) +
  geom_col(fill='#d30000')+
  ylim(0,90)
```



4.0 Exporatory Analysis

4.1 Model Selection

Backward selection is used for selecting the most appropriate model. To start with, a final model is created using all potential variables after which elimination is done to remove less influential variables. The process is repeated till variable remaining after elimination all meet standard for neccessity ($P < 0.05$)

Step 1: Full model creation

```
data <- subset(data, select = -c(country, year))
```

```
full_model <- lm(life_expectancy ~ ., data = data)
summary(full_model)
```

```
##
## Call:
## lm(formula = life_expectancy ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -18.3870  -2.3646   0.0198   2.4996  15.6842
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.621e+01  6.176e-01  91.016  < 2e-16 ***
## statusDeveloping        -1.616e+00  2.694e-01  -5.998 2.25e-09 ***
## adult_mortality         -2.037e-02  8.135e-04 -25.037  < 2e-16 ***
## infant_deaths            1.021e-01  8.321e-03  12.269  < 2e-16 ***
## alcohol                  9.188e-02  2.519e-02   3.648 0.000270 ***
## percentage_expenditure   1.049e-04  7.878e-05   1.331 0.183261
## hepatitis_b             -7.831e-03  3.673e-03  -2.132 0.033099 *
## measles                 -3.206e-05  7.635e-06  -4.199 2.77e-05 ***
## bmi                      5.223e-02  4.679e-03  11.161  < 2e-16 ***
## under_five_deaths       -7.709e-02  6.121e-03 -12.595  < 2e-16 ***
## polio                    2.658e-02  4.505e-03   5.900 4.07e-09 ***
## total_expenditure        6.095e-02  3.388e-02   1.799 0.072146 .
## diphtheria               4.212e-02  4.738e-03   8.890  < 2e-16 ***
## hiv_aids                -4.898e-01  1.841e-02 -26.608  < 2e-16 ***
## gdp                      4.240e-05  1.189e-05   3.567 0.000367 ***
## population               2.139e-09  1.643e-09   1.302 0.193066
## schooling                8.733e-01  3.473e-02  25.144  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.033 on 2852 degrees of freedom
## Multiple R-squared:  0.8122, Adjusted R-squared:  0.8111
## F-statistic: 770.9 on 16 and 2852 DF,  p-value: < 2.2e-16
```

Step 2: Elimination

population is first to be eliminated as it is the least statistically significant and has a highest p-value of 0.193 which is greater than 0.05

```
model_1 <- lm(life_expectancy ~ .-population, data = data)
summary(model_1)
```

```
##
## Call:
## lm(formula = life_expectancy ~ . - population, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4077  -2.3633   0.0264   2.5000  15.7442
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.619e+01  6.175e-01  90.997  < 2e-16 ***
## statusDeveloping        -1.605e+00  2.693e-01  -5.960 2.84e-09 ***
## adult_mortality         -2.040e-02  8.132e-04 -25.086  < 2e-16 ***
## infant_deaths            1.040e-01  8.196e-03  12.684  < 2e-16 ***
## alcohol                  9.280e-02  2.518e-02   3.685 0.000233 ***
## percentage_expenditure   1.038e-04  7.879e-05   1.317 0.187916
## hepatitis_b             -7.891e-03  3.673e-03  -2.148 0.031770 *
```

```
## measles               -3.245e-05  7.631e-06  -4.252 2.19e-05 ***
## bmi                    5.235e-02  4.679e-03  11.188  < 2e-16 ***
## under_five_deaths     -7.806e-02  6.076e-03 -12.846  < 2e-16 ***
## polio                  2.657e-02  4.506e-03   5.898 4.11e-09 ***
## total_expenditure      5.975e-02  3.387e-02   1.764 0.077845 .
## diphtheria             4.232e-02  4.736e-03   8.935  < 2e-16 ***
## hiv_aids              -4.897e-01  1.841e-02 -26.597  < 2e-16 ***
## gdp                    4.271e-05  1.189e-05   3.593 0.000332 ***
## schooling              8.739e-01  3.473e-02  25.161  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.034 on 2853 degrees of freedom
## Multiple R-squared:  0.8121, Adjusted R-squared:  0.8111
## F-statistic:   822 on 15 and 2853 DF,  p-value: < 2.2e-16
```

Percentage expenditure which is the least significant in this new model has a p-value of 0.188 which is greater than 0.05 so is eliminated

```
model_2 <- lm(life_expectancy ~ .-population-percentage_expenditure, data = data)
summary(model_2)
```

```
##
## Call:
## lm(formula = life_expectancy ~ . - population - percentage_expenditure,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3805  -2.3807   0.0402   2.4924  15.7391
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.624e+01  6.164e-01  91.244  < 2e-16 ***
## statusDeveloping  -1.642e+00  2.679e-01  -6.128 1.01e-09 ***
## adult_mortality   -2.040e-02  8.133e-04 -25.088  < 2e-16 ***
## infant_deaths      1.040e-01  8.197e-03  12.690  < 2e-16 ***
## alcohol            9.370e-02  2.518e-02   3.722 0.000202 ***
## hepatitis_b       -8.528e-03  3.642e-03  -2.342 0.019262 *
## measles           -3.251e-05  7.631e-06  -4.260 2.11e-05 ***
## bmi                5.224e-02  4.679e-03  11.165  < 2e-16 ***
## under_five_deaths -7.810e-02  6.077e-03 -12.852  < 2e-16 ***
## polio              2.649e-02  4.506e-03   5.878 4.63e-09 ***
## total_expenditure  6.415e-02  3.371e-02   1.903 0.057157 .
## diphtheria         4.254e-02  4.733e-03   8.988  < 2e-16 ***
## hiv_aids          -4.892e-01  1.841e-02 -26.571  < 2e-16 ***
## gdp                5.575e-05  6.571e-06   8.486  < 2e-16 ***
## schooling          8.716e-01  3.469e-02  25.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.034 on 2854 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.8111
## F-statistic: 880.3 on 14 and 2854 DF,  p-value: < 2.2e-16
```

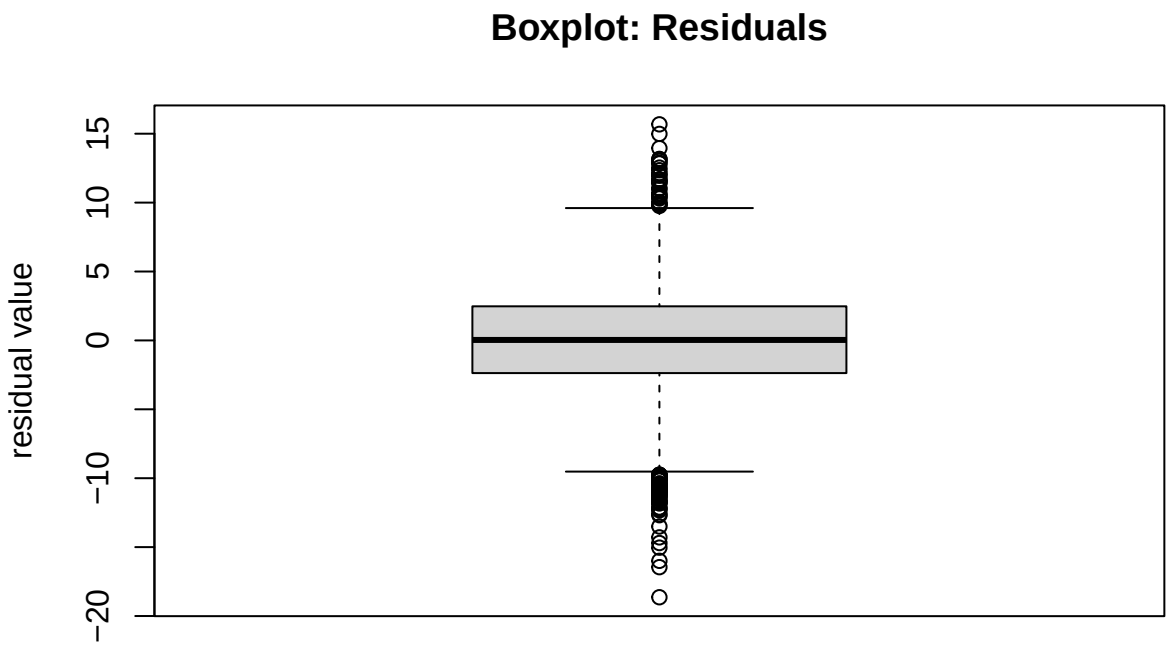
total expenditure which is the least significant in this new model has a p-value of 0.057 which is greater than 0.05 so is eliminated

```
model_3 <- lm(life_expectancy ~ .-population-percentage_expenditure-total_expenditure, data = data)
summary(model_3)
```

```
##
## Call:
## lm(formula = life_expectancy ~ . - population - percentage_expenditure -
##     total_expenditure, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6366  -2.3706   0.0335   2.4745  15.6780
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.658e+01  5.907e-01  95.779  < 2e-16 ***
## statusDeveloping   -1.714e+00  2.653e-01  -6.461 1.21e-10 ***
## adult_mortality    -2.043e-02  8.135e-04 -25.112  < 2e-16 ***
## infant_deaths       1.040e-01  8.201e-03  12.677  < 2e-16 ***
## alcohol             9.930e-02  2.502e-02   3.969 7.38e-05 ***
## hepatitis_b        -8.465e-03  3.643e-03  -2.324   0.0202 *
## measles            -3.305e-05  7.630e-06  -4.332 1.53e-05 ***
## bmi                 5.308e-02  4.660e-03  11.390  < 2e-16 ***
## under_five_deaths  -7.809e-02  6.080e-03 -12.844  < 2e-16 ***
## polio               2.650e-02  4.508e-03   5.878 4.64e-09 ***
## diphtheria          4.275e-02  4.734e-03   9.029  < 2e-16 ***
## hiv_aids           -4.867e-01  1.837e-02 -26.491  < 2e-16 ***
## gdp                 5.513e-05  6.565e-06   8.397  < 2e-16 ***
## schooling           8.740e-01  3.469e-02  25.198  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.036 on 2855 degrees of freedom
## Multiple R-squared:  0.8117, Adjusted R-squared:  0.8109
## F-statistic: 946.9 on 13 and 2855 DF,  p-value: < 2.2e-16
```

The largest p-value which is "hepatitis_b" is less than 0.05, so we do not need to eliminate any predictors. The current model is the best-fitting model. Dropped variables include population, percentage_expenditure and total_expenditure

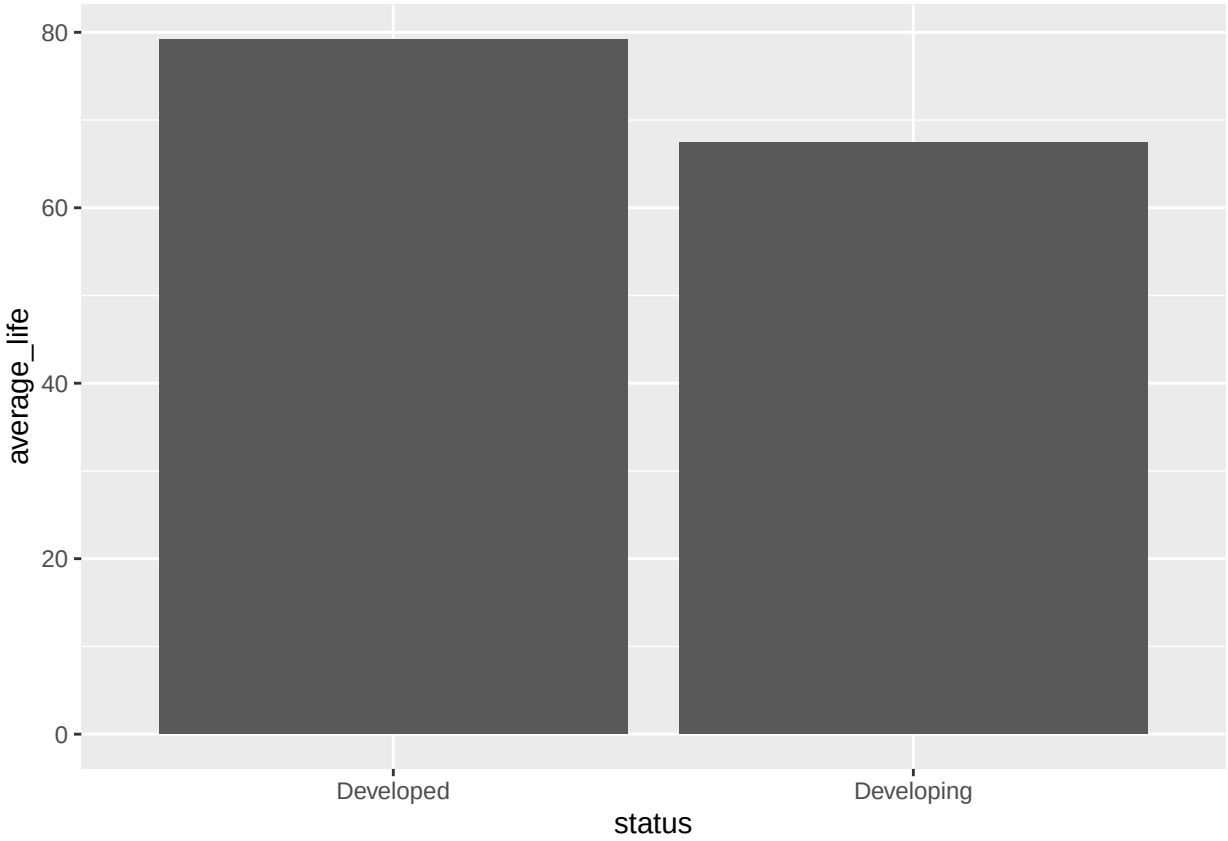
```
# Plot of Residuals of final  model
boxplot(model_3[['residuals']],main='Boxplot: Residuals',ylab='residual value')
```



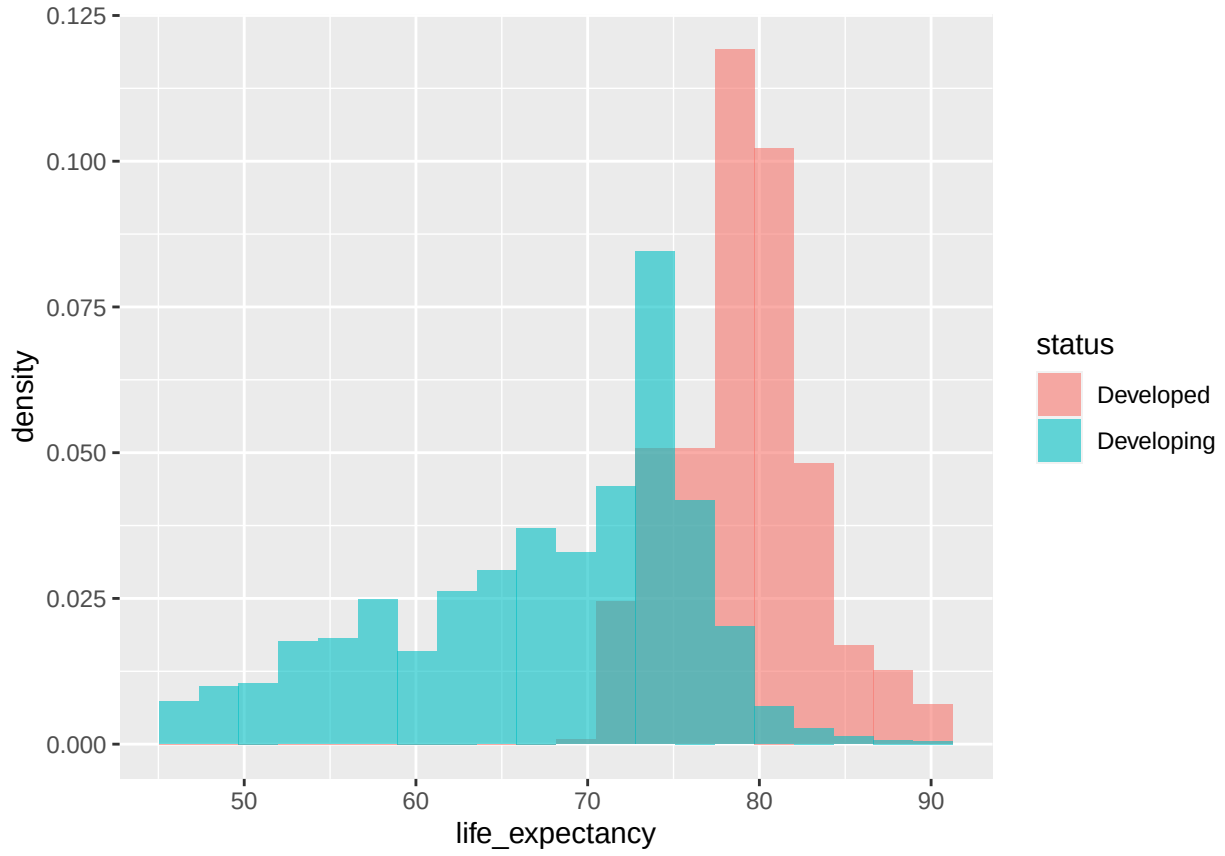
4.2 Variable Analysis

4.2.1 Levels of development

```
# Status
dev_data <- data %>%
  group_by(status)%>%
  summarise(
    average_life = mean(life_expectancy, na.rm = TRUE)
  )



dev_data%>%
  ggplot(aes(x=status, y= average_life)) +
  geom_bar(stat='identity')
```



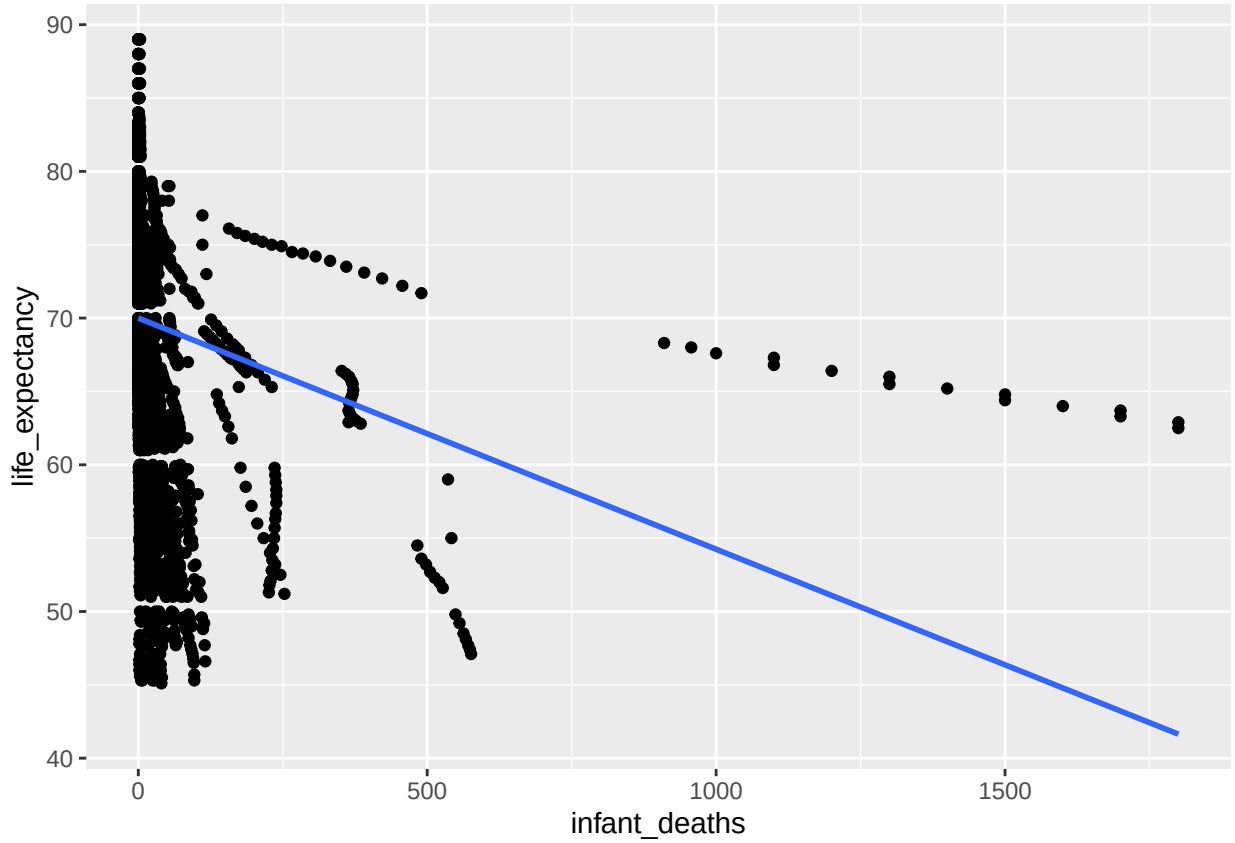

```
# View distribution of life expectanc for developed and developing countries

ggplot(data) +
  aes(x = life_expectancy, fill=status) +
  geom_histogram(bins = 20, alpha=0.6, position='identity', aes(y = ..density..))
```



4.2.2 Mortality

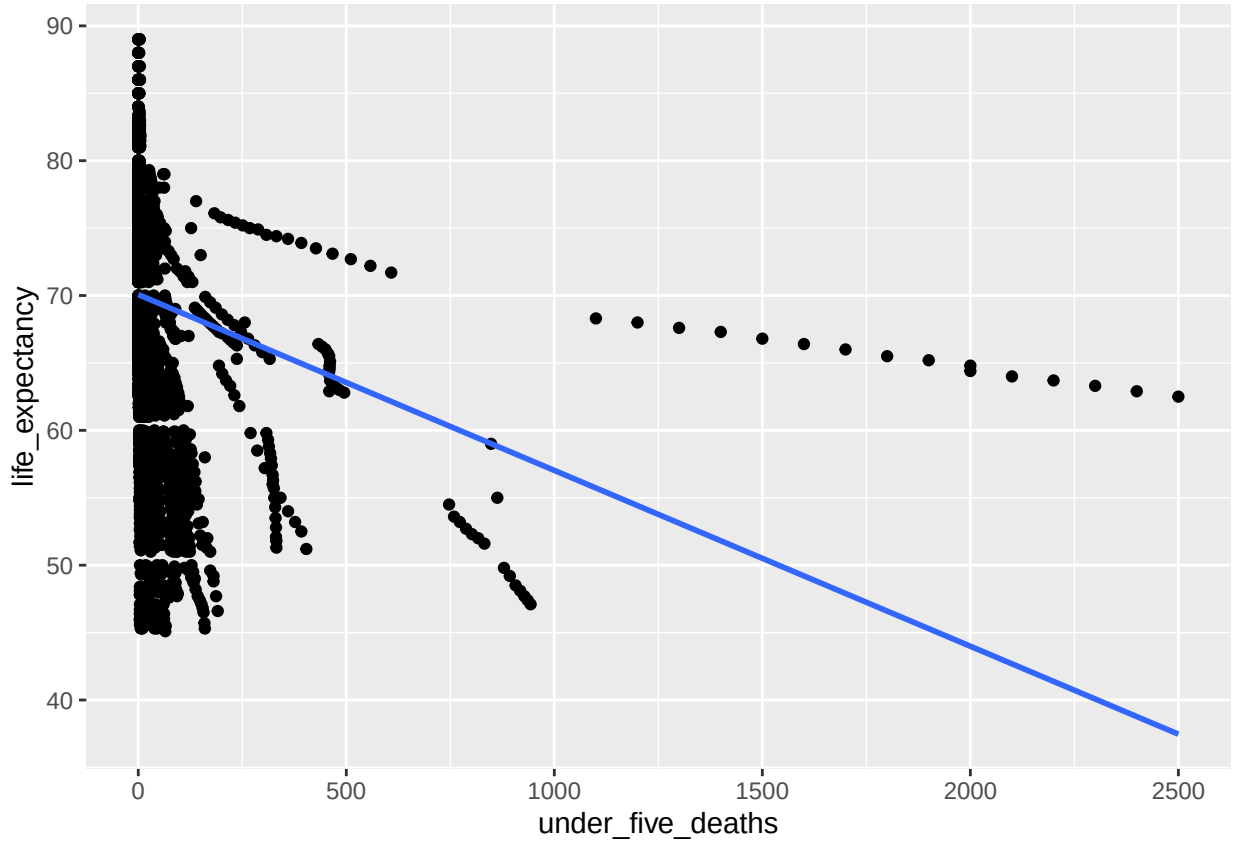
4.2.2.1 Infant Mortality

```
# infant_mortality
data %>%
  ggplot() +
  aes(x = infant_deaths, y = life_expectancy) +
  geom_point(stat='identity')+
  geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



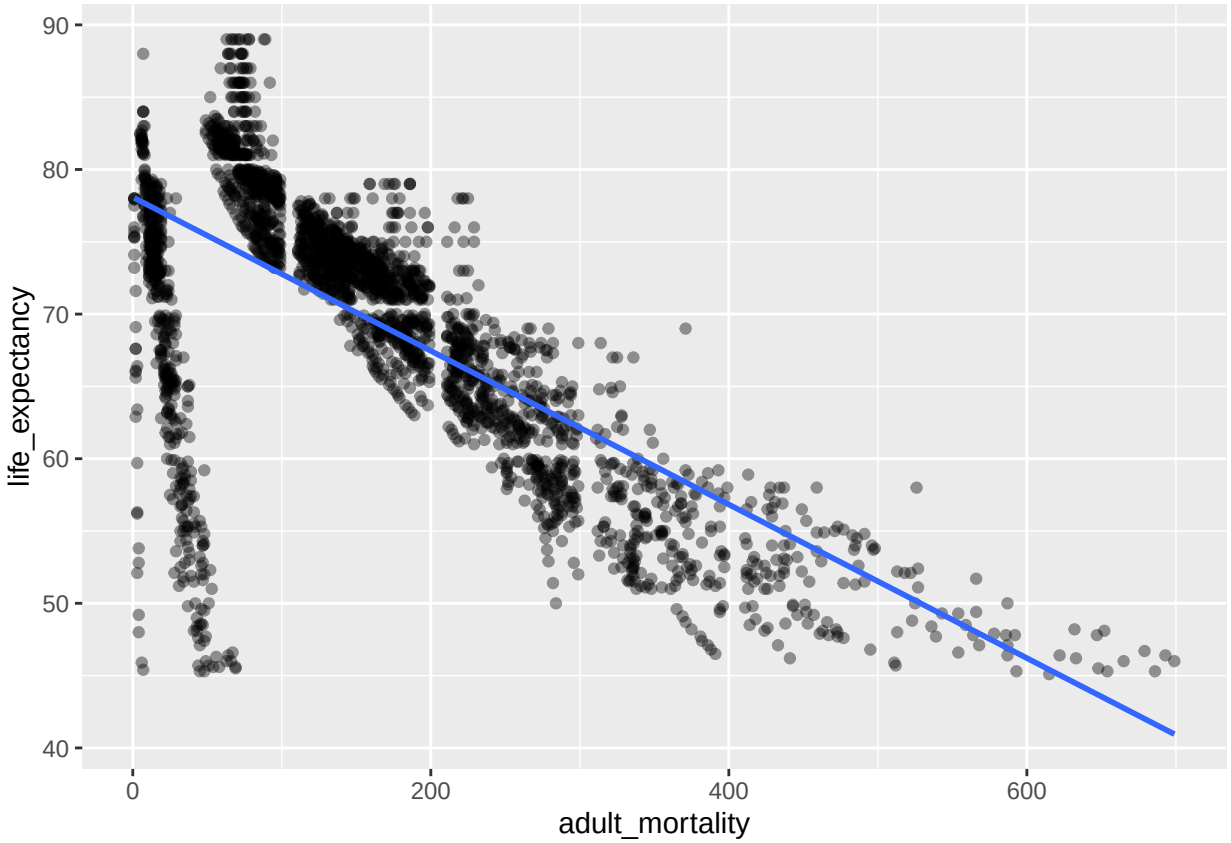
4.2.2.2 Under-five death

```
# infant_mortality
data %>%
  ggplot() +
  aes(x = under_five_deaths, y = life_expectancy) +
  geom_point(stat='identity')+
  geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



4.2.2.3 Adult mortality

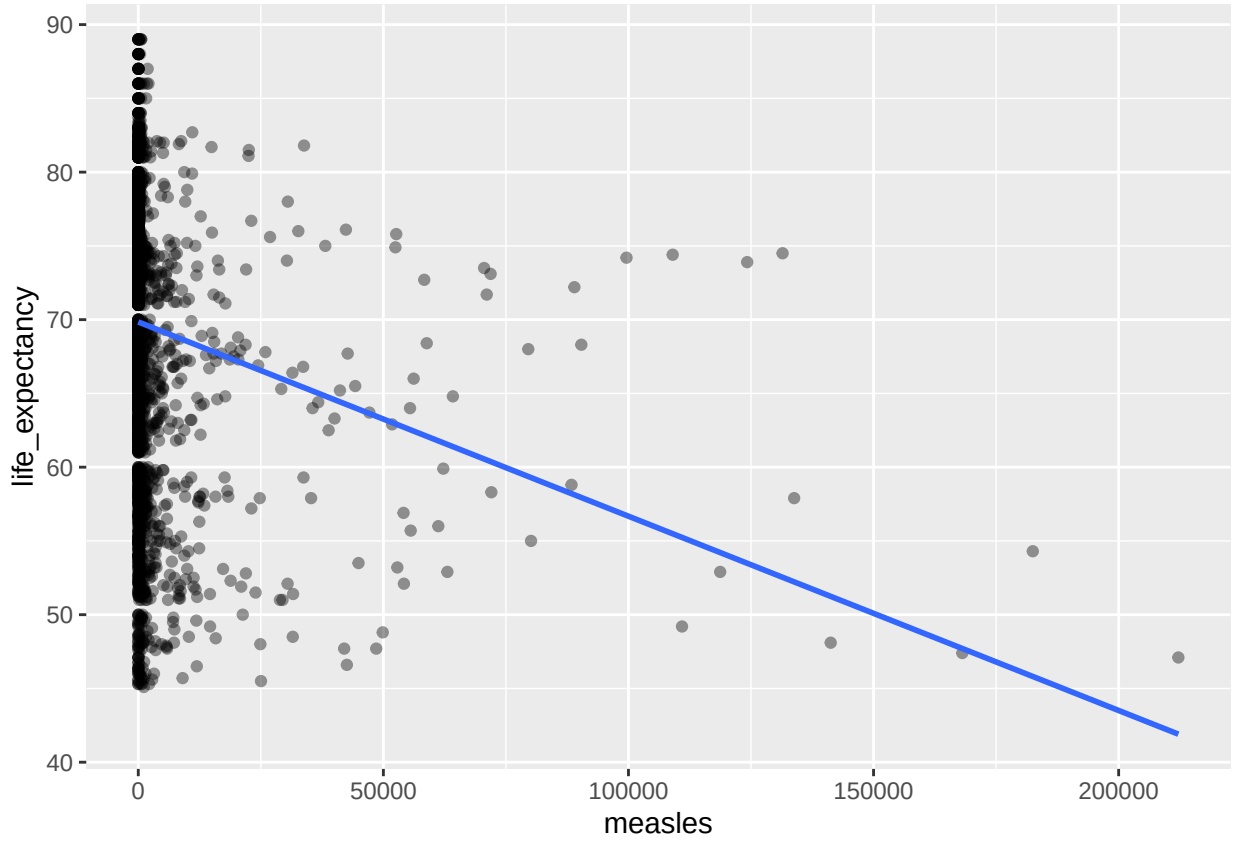
```
# Adult_mortality
ggplot(data) +
aes(x = adult_mortality, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



4.2.3 Death rate due to health condition

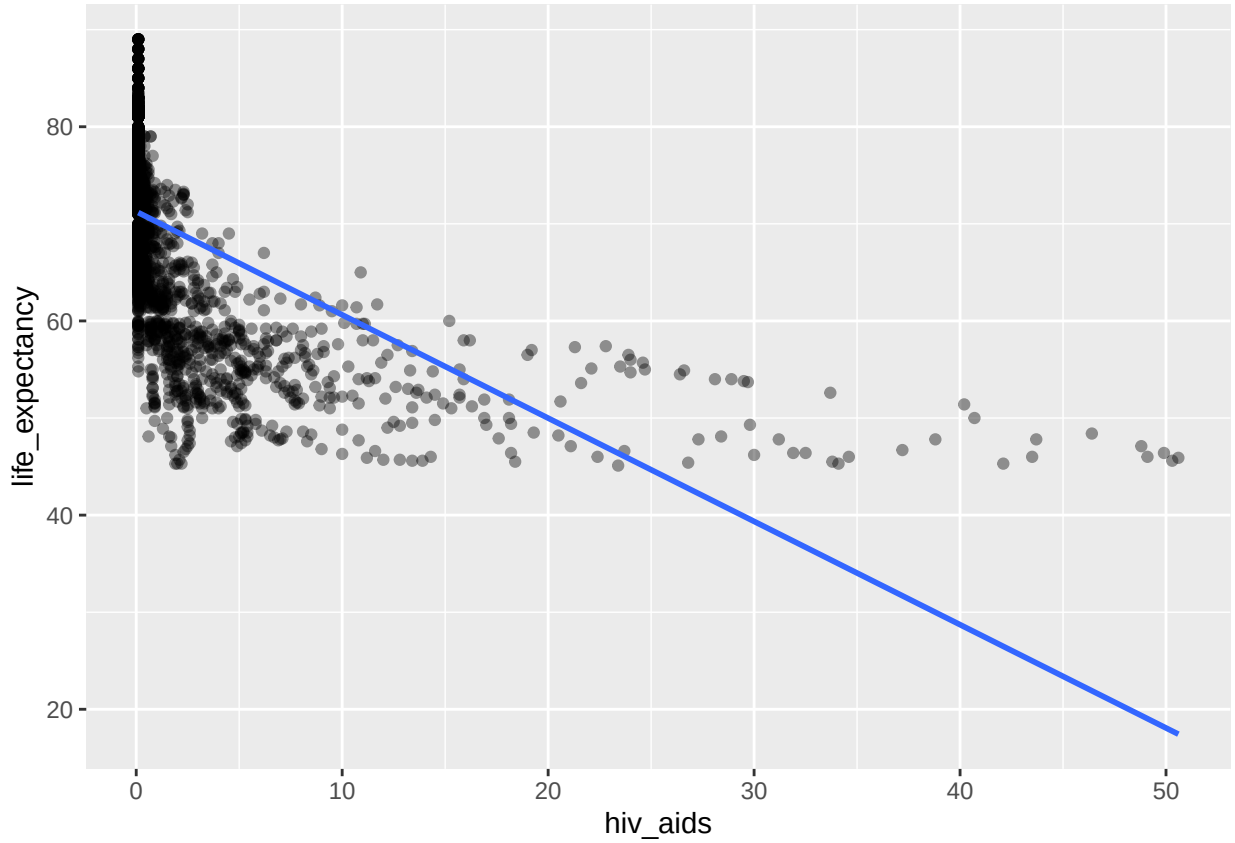
4.2.3.1 Measles

```
# Adult_mortality
ggplot(data) +
aes(x = measles, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



4.2.3.2 HIV/AIDS

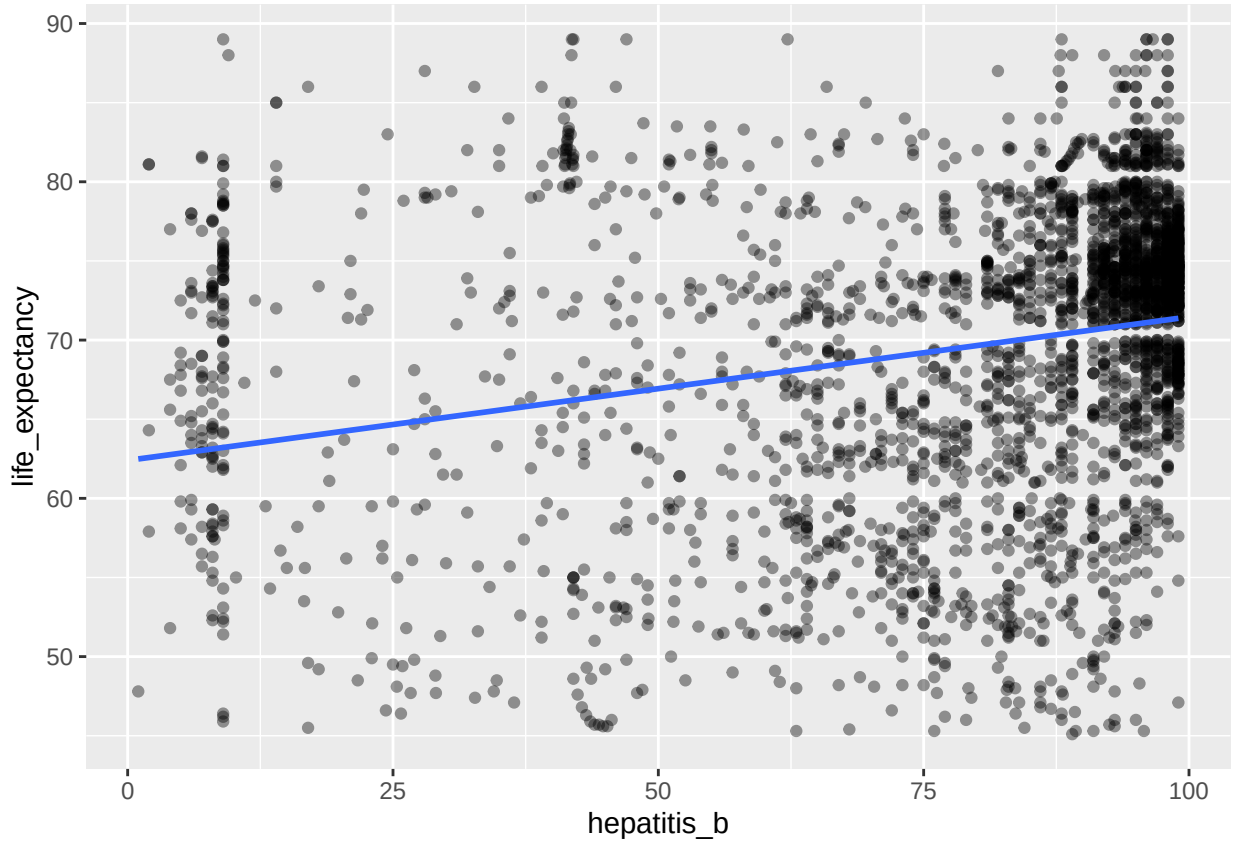
```
# Adult_mortality
ggplot(data) +
aes(x = hiv_aids, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



4.2.4 Immunisation

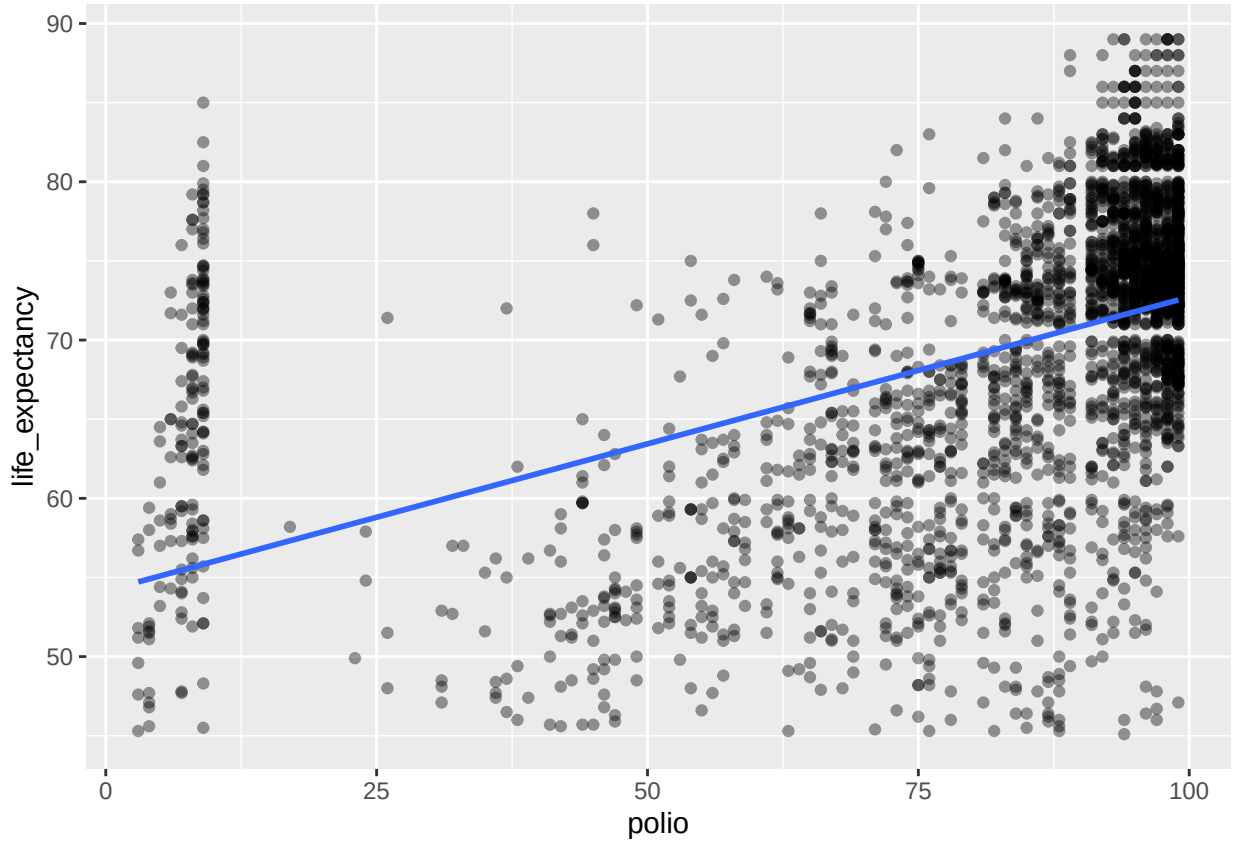
4.2.4.1 Hepatitis B

```
# Hepatitis B
ggplot(data) +
aes(x = hepatitis_b, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



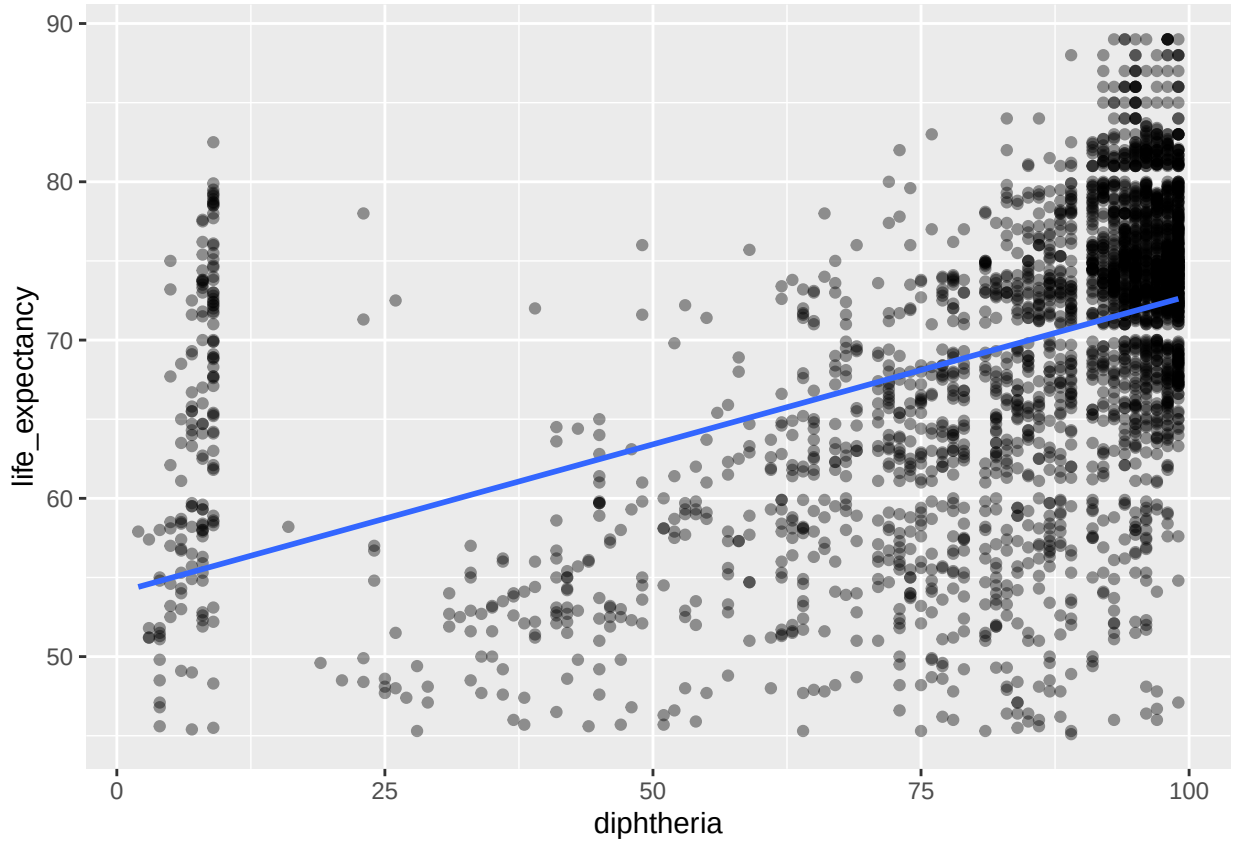
4.2.4.2 Polio

```
# Polio
ggplot(data) +
aes(x = polio, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



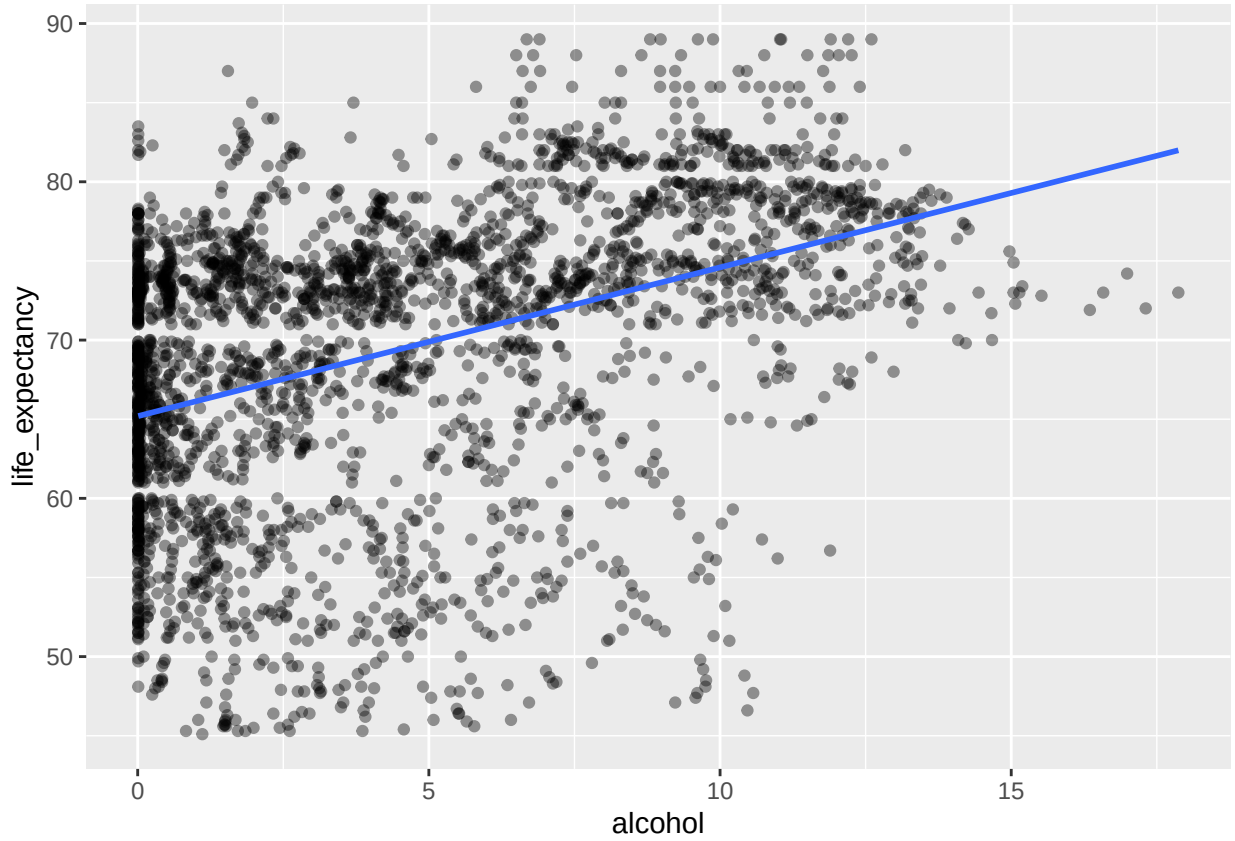
4.2.4.3 Diphtheria Tetanus Toxoid and Pertussis

```
# Diphtheria
ggplot(data) +
aes(x = diphtheria, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



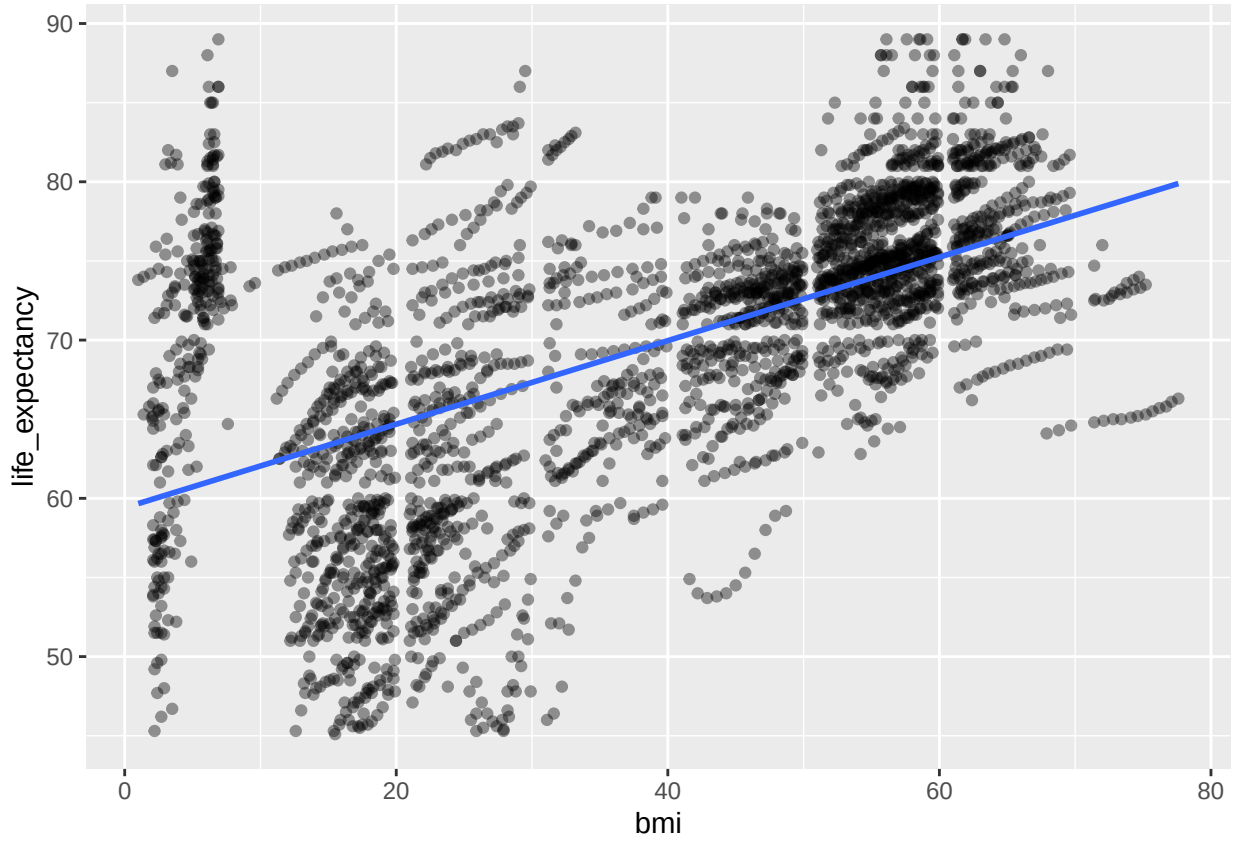
4.2.5 Alcohol

```
# Adult_mortality
ggplot(data) +
aes(x = alcohol, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



4.2.6 Body mass index

```
# Diphtheria
ggplot(data) +
aes(x = bmi, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```

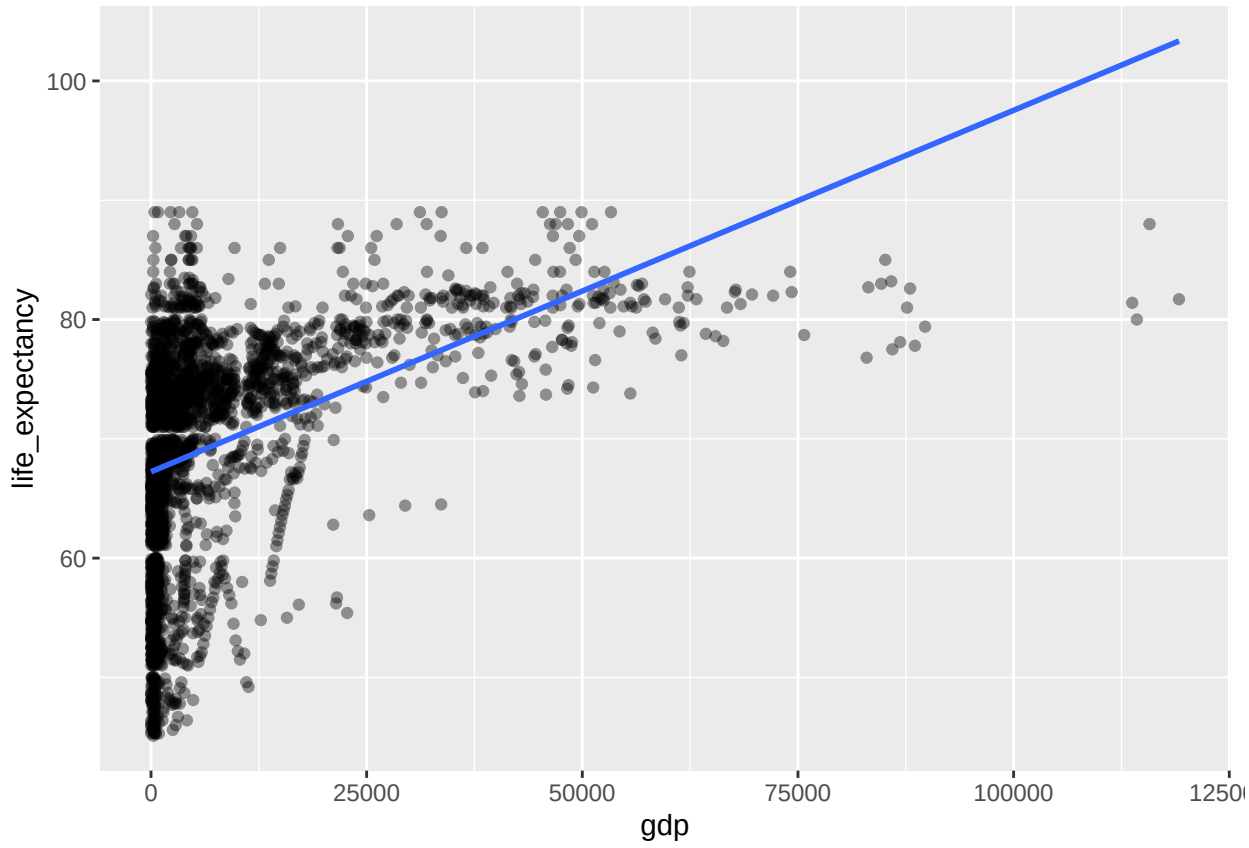



```
df <-
data %>%
  mutate(
    bmi_group = cut(bmi, 3, c('low', 'Moderate', 'high'))
  )

df %>%
  filter(!is.na(bmi_group)) %>%
  group_by(bmi_group) %>%
  summarise(
    average_life_expectancy = mean(life_expectancy, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 2
##   bmi_group average_life_expectancy
##   <fct>                       <dbl>
## 1 low                          62.8
## 2 Moderate                     69.3
## 3 high                         76.2
```

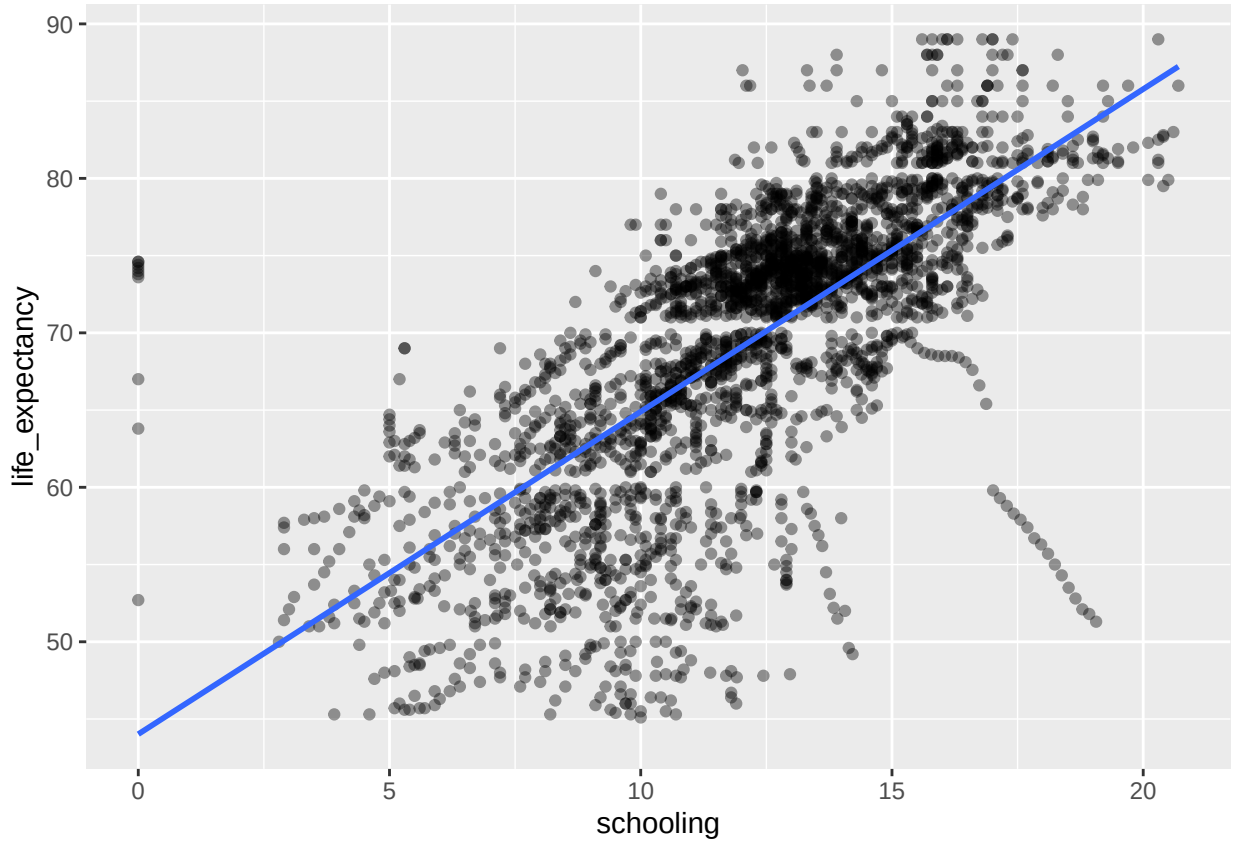
4.2.7 Gross Domestic product (GDP)

```
# GDP
ggplot(data) +
aes(x = gdp, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```



4.2.8 Schooling

```
# Schooling
ggplot(data) +
aes(x = schooling, y = life_expectancy) +
geom_point(alpha = 0.4)+
geom_smooth(method = "lm",formula = y ~ x, se = FALSE)
```




```
df <-
data %>%
  mutate(
    school_group = cut(schooling, 3, c('low', 'Moderate', 'high'))
  )

df %>%
  filter(!is.na(school_group)) %>%
  group_by(school_group) %>%
  summarise(
    average_life_expectancy = mean(life_expectancy, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 2
##   school_group average_life_expectancy
##   <fct>                          <dbl>
## 1 low                             56.4
## 2 Moderate                        67.2
## 3 high                            76.6
```