Analysing Life Expectancy With Linear Regression

MATH 1312 Regression Analysis Final Project

Rinku Bajaj (s3672522), Arion Barzoucas-Evans (s3650046), Deepika Joshi (s3672595)

02/06/2019

Contents

1 Introduction

The main aim of this project is to identify the most influential factors affecting life expectancy in developed and developing countries from 2000-2015. The project considers factors traditionally studied for life expectancy predictions such as demographic variables, mortality rate and income composition as well as the effects of new factors such as immunization and human development index. The data is sourced from Kaggle Repository and is a combination of life-expectancy and health data from Global Health Observatory (GHO) published by World Health Organisation (WHO) and economic data published by United Nations. It consists of data spanning over 15 years from 2000-2015 for 193 countries and can be broadly divided into Economic Factors, Social Factors, Immunization Factors and Mortality Factors. The dataset has 2938 observations for 20 predictor variables describing the life expectancy of each nation for that year.

2 Methodology

The countries present in the dataset can be categorised into developed and developing. Different linear regression models will be fitted to each group using the traditionally considered factors, the extended factors, and using stepwise regression. The adequacy of the fitted models will be assessed with a combination of residual diagnostics and statistical tests. Using the models with the best fit to the data, the most influential factors will be identified for both developed and developing countries.

3 Results

Figure 1 displays the distribution of the life expectancy variable in the dataset. This shows that life expectancy ranges from around 40 to 90 years with 50% of the values being over 72 years. This indicates that the variable is slightly skewed to the left. The bar chart in Figure 2 shows that the majority of countries in the dataset are developing countries with only around 30 countries being developed.

Figure 3 shows the correlation between the numeric variables in the dataset. According to this, life expectancy has a high negative correlation with adult mortality, HIV, and thinness. Conversely, there is a high positive correlation between life expectancy and BMI, schooling, polio, diphteria, and GDP. Additionaly, many of the independent variables are correlated to each other and may lead to multi-colinearity issues.

Life expectancy is also strongly correlated to a country's status (Fig 4). It can be seen that developed countries generally have a life expectancy around 80 with little variation. On the other hand, developing countries have an average life expectancy of just under 70 with much greater variance and several outliers with life expectancy of around 40. Life expectancy for both developed and developing countries has an upward trend having a 5 year improvement from 2000 to 2015 (Fig 5). Figure 6 shows a strong negative correlation between life expectancy and adult mortality for both developed and developing countries. However, there are many cases of countries with a low adult mortality rate that still have low life expectancy.This may indicate that adult mortality alone is not a sufficient indicator of life expectancy.

```r
# load libraries
library(ggplot2)
library(data.table)
library(knitr)
library(corrplot)
```

```r
# read data
data = setDT(read.csv("Life Expectancy Data.csv"))

# replace/remove NAs
data = data[!is.na(Life.expectancy)]
cols = names(data)[!grepl("Country|Status", names(data))]
data[, (cols) := lapply(.SD, function(x) {ifelse(is.na(x), -1, x)}), .SDcols = cols]
```

```r
# life expectancy
ggplot(data, aes(x = Life.expectancy)) +
  geom_histogram(fill = "aquamarine4", colour = "white") +
  geom_vline(xintercept = quantile(data$Life.expectancy, 0.5), linetype = "dashed", colour = "red", siz
  annotate("text", x = quantile(data$Life.expectancy, 0.5) - 3, y = 400, label = "Median", colour = "re
  theme_minimal() +
  labs(x = "Life Expectancy", y = "Count", title = "Life expectancy distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
# status
ggplot(data[, length(unique(Country)), Status], aes(x = Status, y = V1)) +
  geom_bar(stat = "identity", fill = "aquamarine4") +
  theme_minimal() +
  labs(y = "Countries", title = "Distribution of countries by status")
```

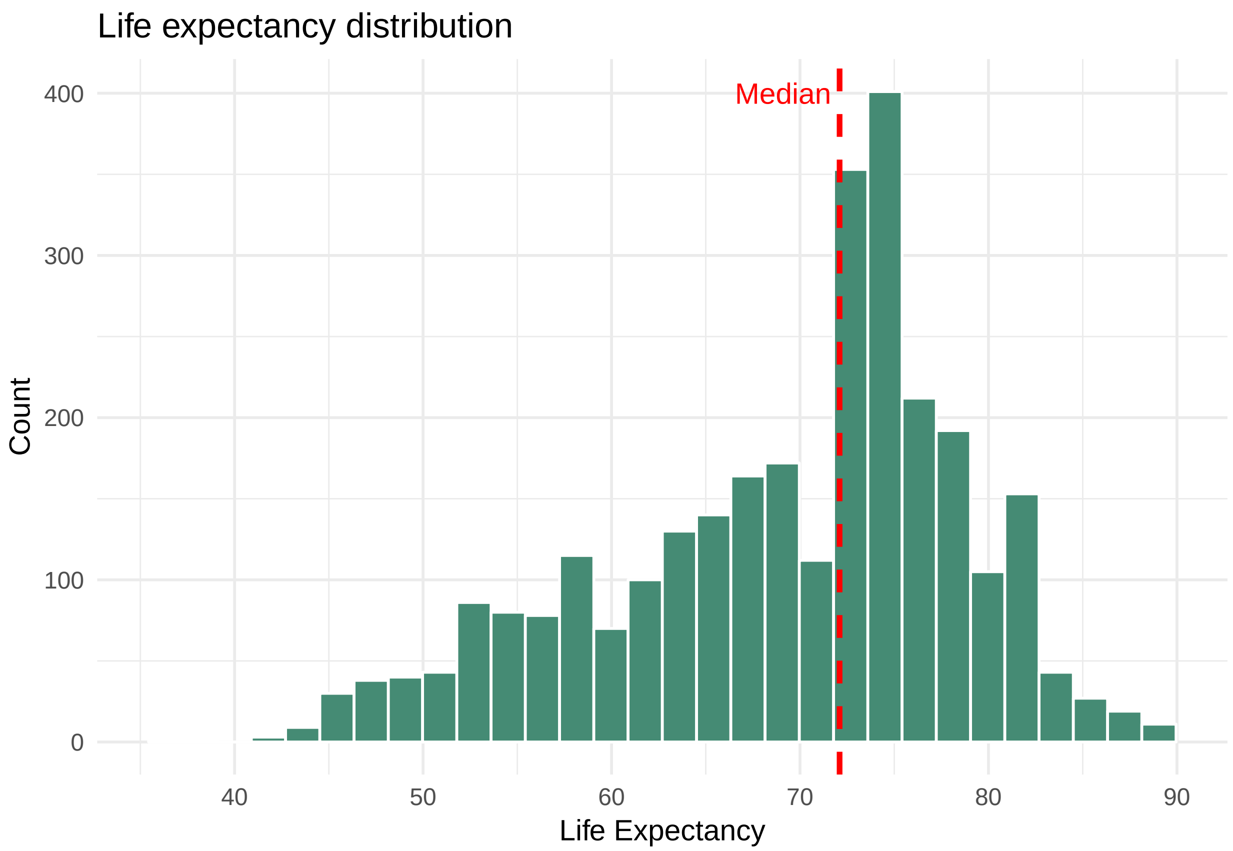

Figure 1: Status distribution in the dataset.

```
# correlation plot
corr = cor(data[, c(2,4:22)], use = "complete.obs")
corrplot(corr, type = "lower")
```

```
# status vs expectancy
ggplot(data, aes(x = Status, y = Life.expectancy)) +
  geom_boxplot( fill = "aquamarine4") +
  theme_minimal() +
  labs(y = "Life Expectancy", title = "Life expectancy by country status")
```

```
# timeseries by status
ggplot(data[, mean(Life.expectancy, na.rm = T), .(Year, Status)], aes(x = Year, y = V1, colour = Status
  geom_point()+
  geom_line() +
  theme_minimal() +
  labs(y = "Average Life Expectancy", title = "Life expectancy timeseries by country status",
       x = "Year")
```

```
# status & adult mortality
ggplot(data, aes(x = Adult.Mortality, y = Life.expectancy, colour = Status)) +
  geom_point()+
  theme_minimal() +
  labs(y = "Life Expectancy", title = "Life Expectancy by Adult Mortality & Country Status",
```

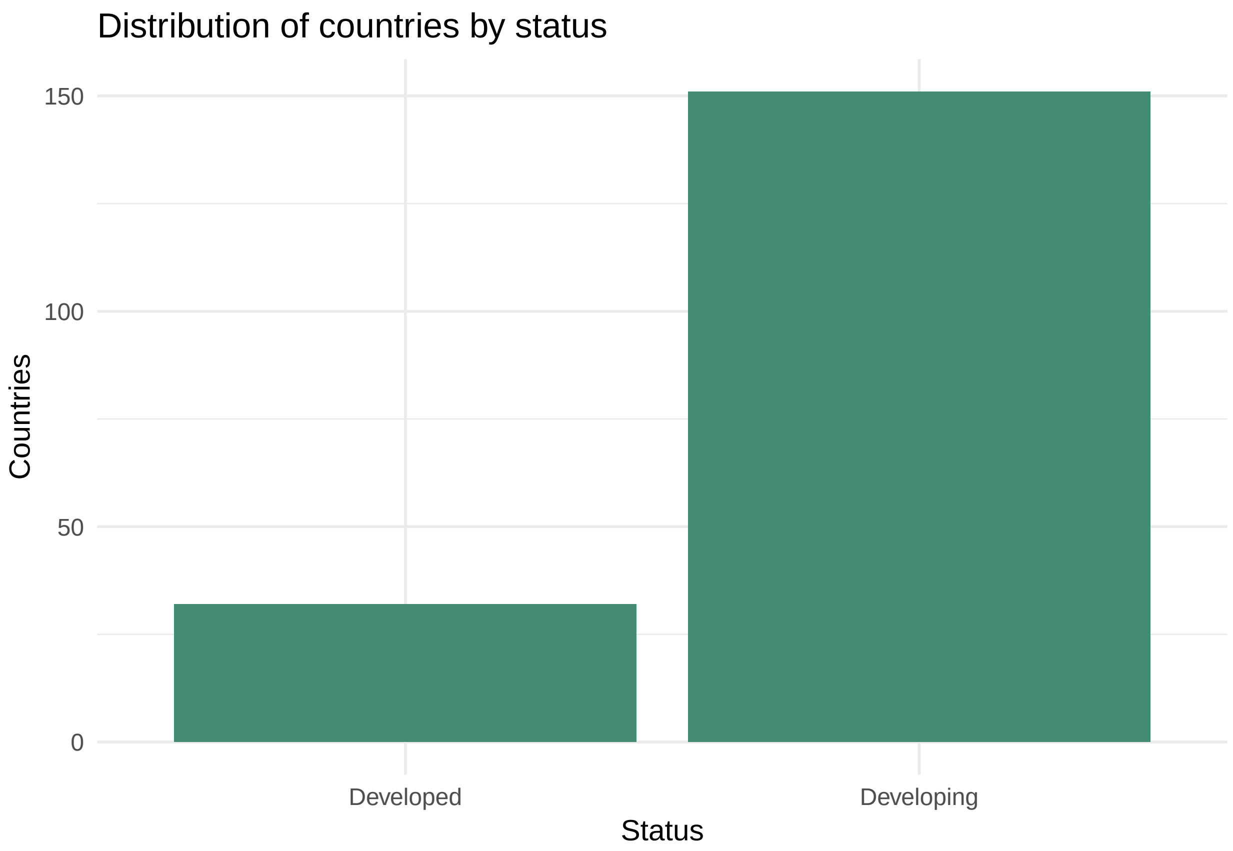

Figure 2: Life expectancy distribution in the dataset.

```
        x = "Adult Mortality")
```

4 Discussion

5 Conclusion

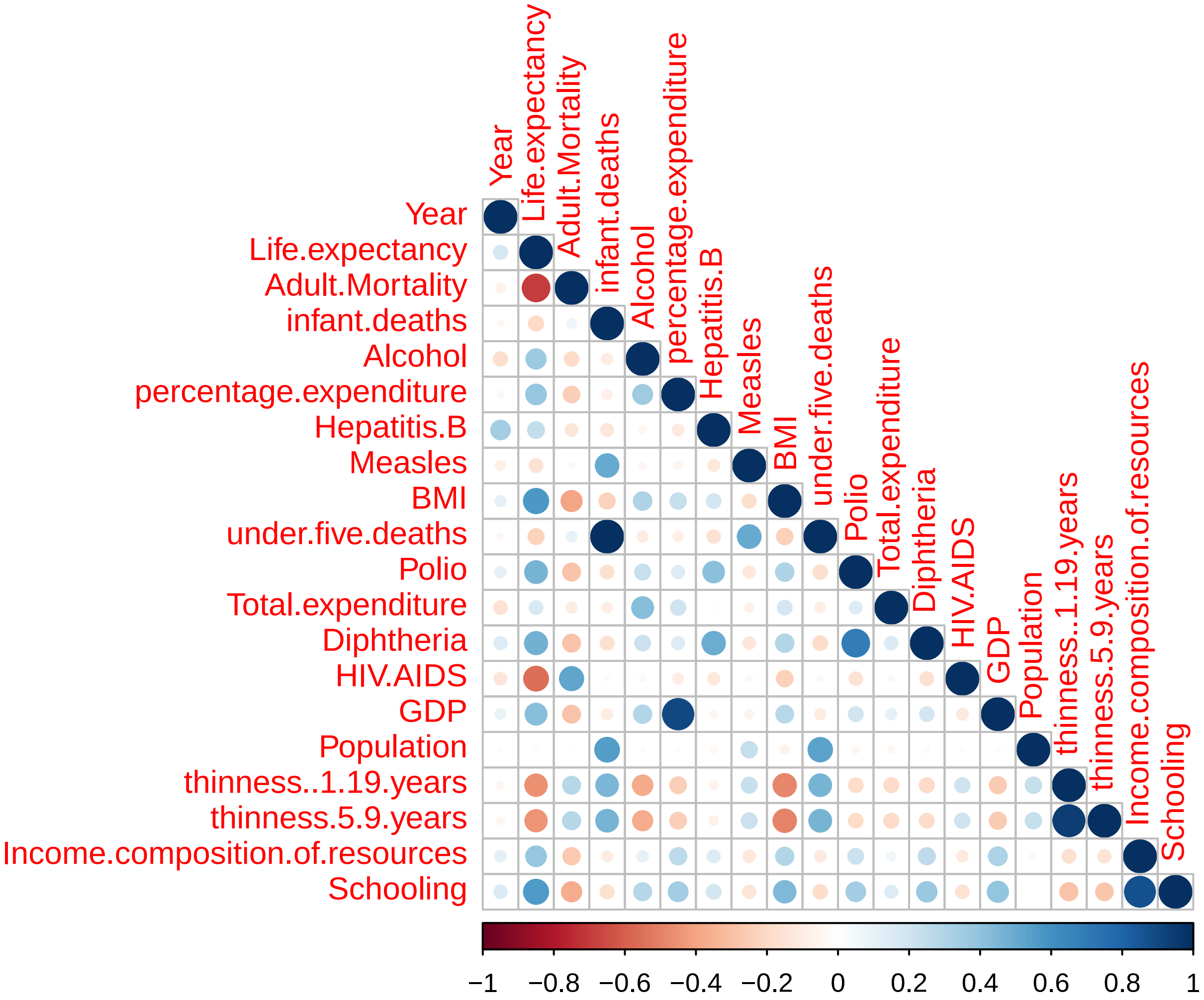

Figure 3: Correlation plot for all numeric variables in the dataset.

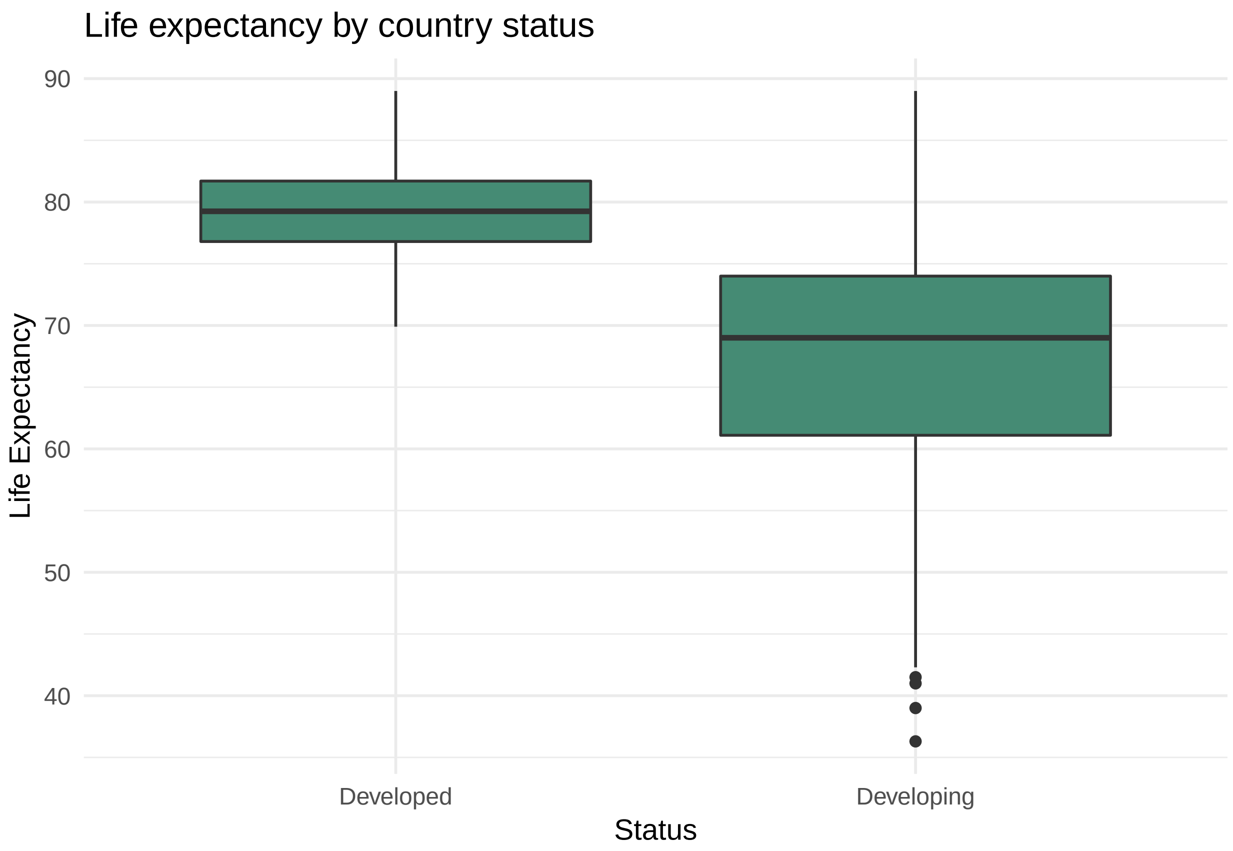

Figure 4: Life expectancy by status.

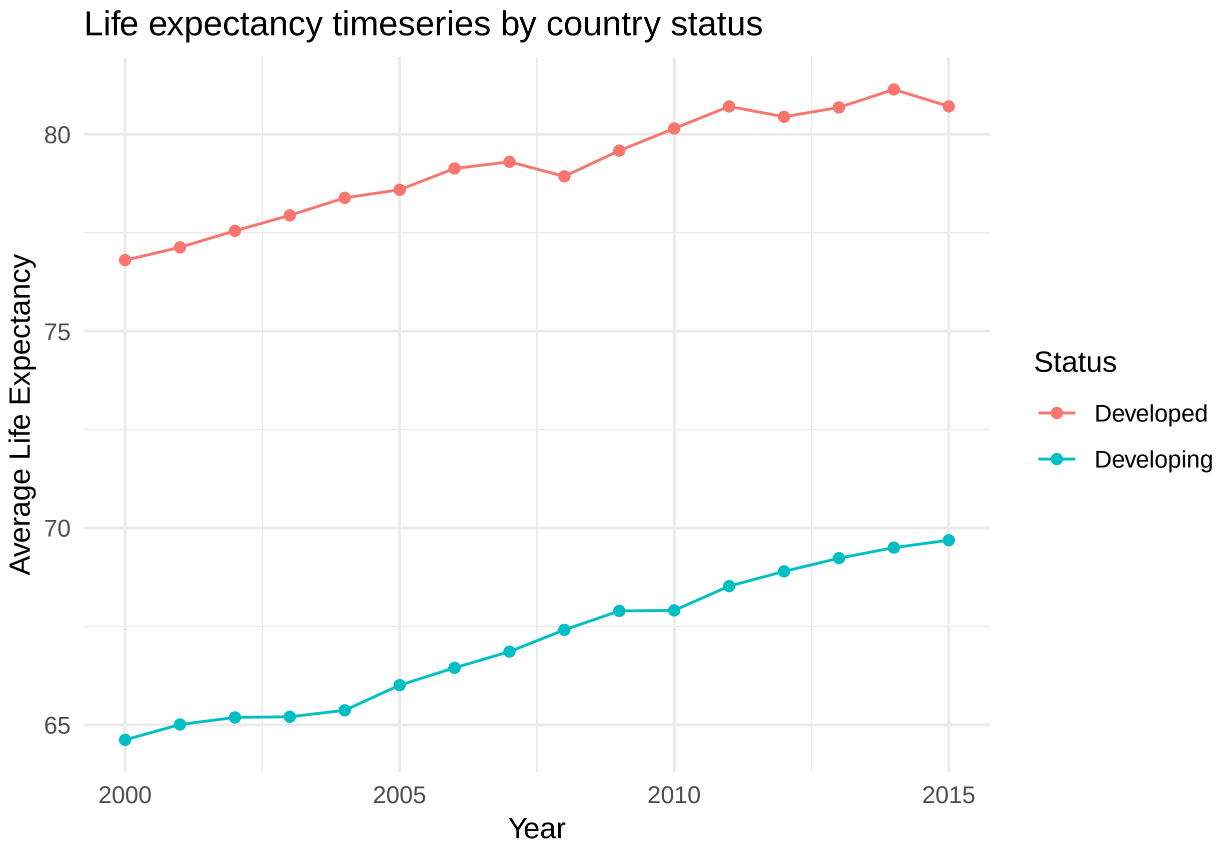

Figure 5: Historical life expectancy by status.

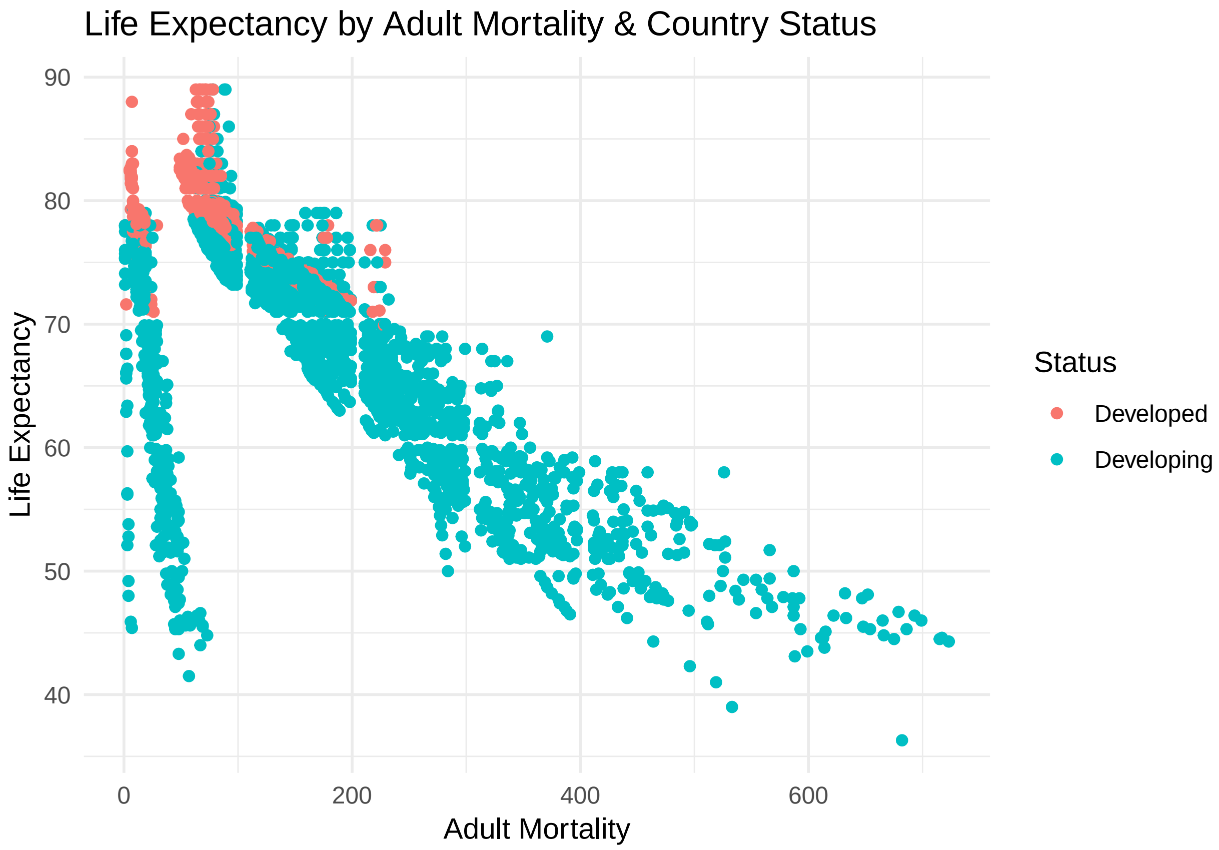

Figure 6: Life expectancy by adult mortality and status.