BIOS 512 - Final Project

Raphael Kim

I am working with data from https://www.kaggle.com/augustus0498/life-expectancy-who.

This dataset, from WHO, has data on the country level to better understand factors that influence life expectancy. There are separate observations by country and by year, along with corresponding statistics related to relevant risk factors.

For this project, I will primarily be looking at the relationship between Lifeexpectancy (in years) and AdultMortality (probability of dying between 15 and 60 years per 1000 population), GDP (Gross Domestic Product per capita (in USD)), Schooling (in years), while facetting by Status (Developed or Not Developed).

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

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

```
library(dplyr)
library(stats) # for complete.cases
```

```
path="/Users/Raphael/Desktop/Class/BIOS512/final_project"
setwd(path)

data<-read.csv("led.csv")
data %>% head
```

```
##       Country Year     Status Lifeexpectancy AdultMortality infantdeaths
## 1 Afghanistan 2015 Developing           65.0            263           62
## 2 Afghanistan 2014 Developing           59.9            271           64
## 3 Afghanistan 2013 Developing           59.9            268           66
## 4 Afghanistan 2012 Developing           59.5            272           69
## 5 Afghanistan 2011 Developing           59.2            275           71
## 6 Afghanistan 2010 Developing           58.8            279           74
##   Alcohol percentageexpenditure HepatitisB Measles  BMI under.fivedeaths
## 1    0.01             71.279624         65    1154 19.1               83
## 2    0.01             73.523582         62     492 18.6               86
## 3    0.01             73.219243         64     430 18.1               89
## 4    0.01             78.184215         67    2787 17.6               93
## 5    0.01              7.097109         68    3013 17.2               97
## 6    0.01             79.679367         66    1989 16.7              102
##   Polio Totalexpenditure Diphtheria HIV.AIDS       GDP Population
## 1     6             8.16         65      0.1 584.25921   33736494
## 2    58             8.18         62      0.1 612.69651     327582
## 3    62             8.13         64      0.1 631.74498   31731688
## 4    67             8.52         67      0.1 669.95900    3696958
## 5    68             7.87         68      0.1  63.53723    2978599
## 6    66              9.20         66      0.1 553.32894    2883167
##   thinness1.19years thinness5.9years Incomecompositionofresources
## 1              17.2             17.3                        0.479
## 2              17.5             17.5                        0.476
## 3              17.7             17.7                        0.470
## 4              17.9             18.0                        0.463
## 5              18.2             18.2                        0.454
## 6              18.4             18.4                        0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

```
# Make sure no NAs exist:
sum(is.na(data))
```

```
## [1] 2563
```

```
data=data[complete.cases(data),]
dim(data)
```

```
## [1] 1649   22
```

Now we work off of the remaining 1649 samples.

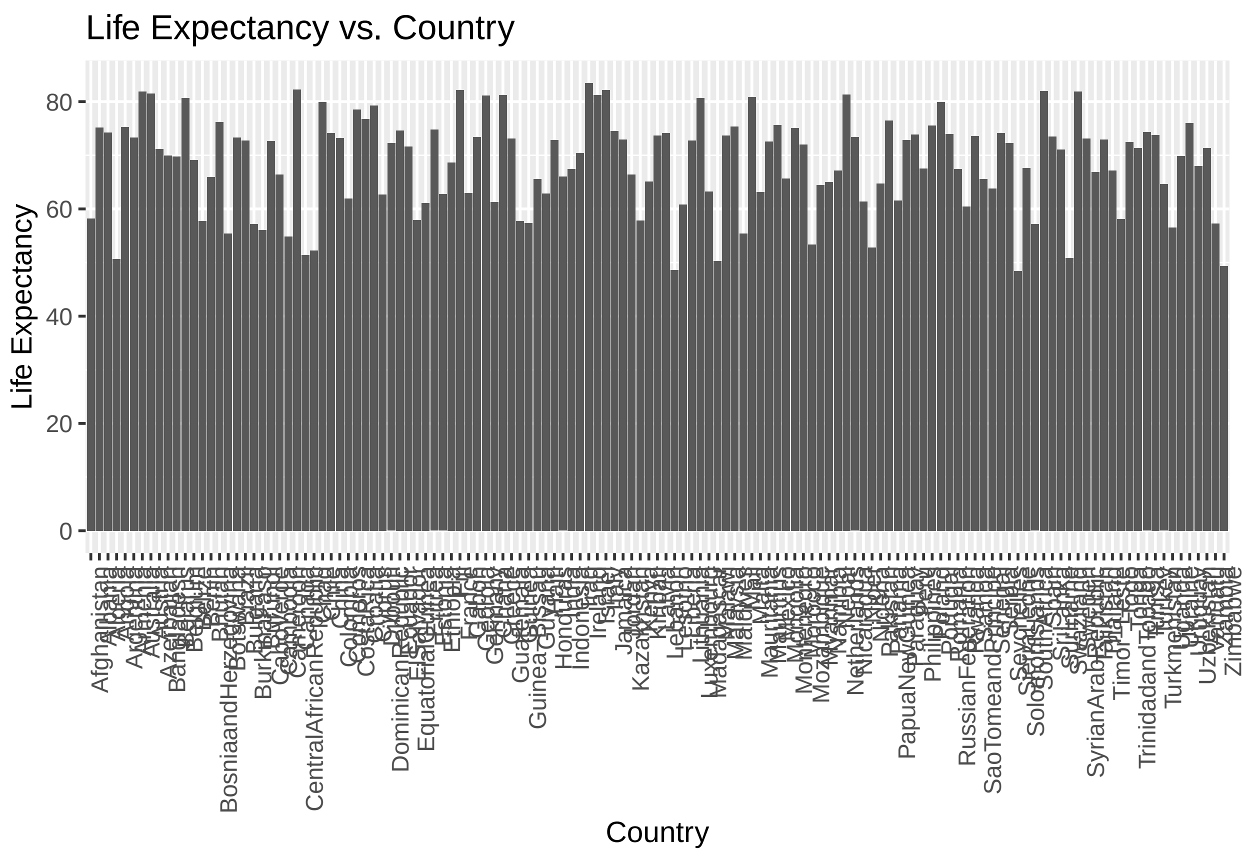
We first begin by grouping these samples on country, and aggregating our measures of interest (Lifeexpectancy, GDP, Schooling, Adult Mortality) for each country by Expectation or mean.

```
aggregatedData = data %>%
  group_by(Country) %>%
  summarize(mean_Lifeexpectancy = mean(Lifeexpectancy), mean_GDP = mean(GDP), mean_Schooling = mean(Scho
```

Life Expectancy vs. Country

Then we look into the countries and their respective (average) measures.

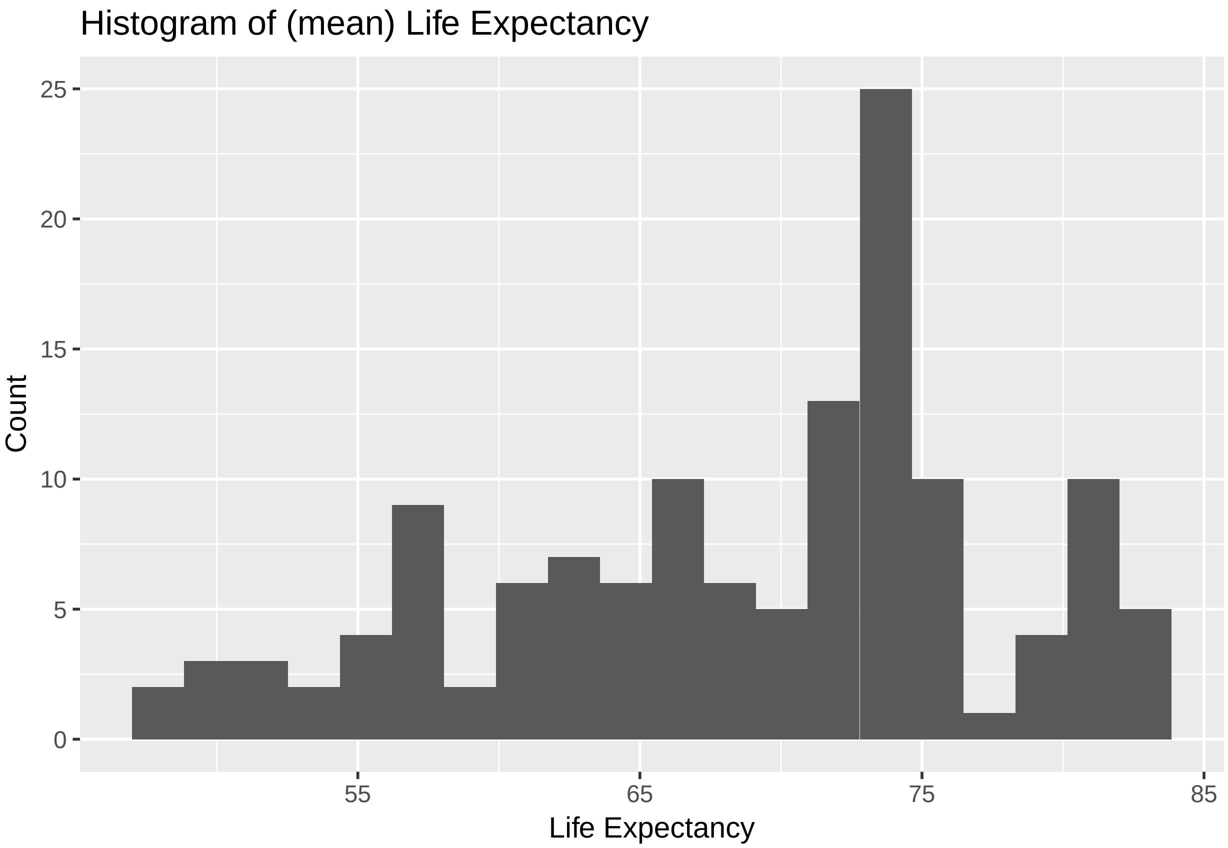
```
p = ggplot(aggregatedData, aes(x=Country))
p = p + geom_bar(aes(y=mean_Lifeexpectancy), stat='identity')
p = p + theme(axis.text.x = element_text(angle = 90, hjust = 1))
p = p + labs(x = 'Country', y = 'Life Expectancy')
p = p + ggtitle("Life Expectancy vs. Country")
p
```



This is messy so let's filter to countries on the extremes, or those that are above and below a certain threshold.

To find a threshold, let's briefly look at the distribution.

```
p = ggplot(aggregatedData, aes(x = mean_Lifeexpectancy))
p = p + geom_histogram(bins = 20)
p = p + ggtitle("Histogram of (mean) Life Expectancy")
p = p + labs(x = 'Life Expectancy', y = 'Count')
p
```



```
IQR(data$Lifeexpectancy) # inter quartile Range
```

```
## [1] 10.6
```

```
quantile(data$Lifeexpectancy) # quantiles
```

```
##   0%  25%  50%  75% 100%
## 44.0 64.4 71.7 75.0 89.0
```

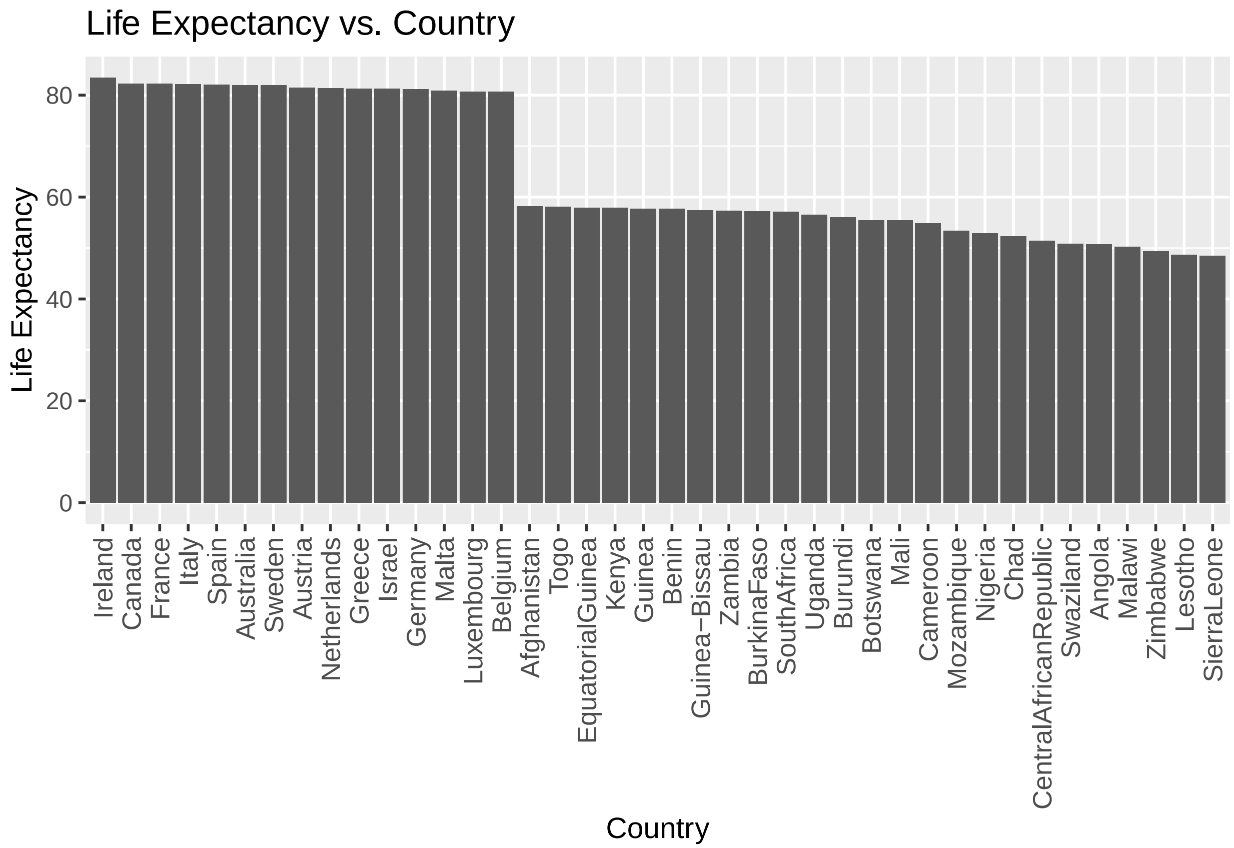
IQR=10.6

0%: 44.0 25%: 64.4 50%: 71.7 75%: 75.0 100%: 89.0

So we will take out the range from (first quartile - 1/2 * IQR) to (3rd quartile + 1/2 *IQR) , just to get the extremes, and replot (sorted by life expectancy).

```
data.extremes <- aggregatedData %>%
  filter(mean_Lifeexpectancy < (64.4-10.6/2) | mean_Lifeexpectancy > (75.0+10.6/2))

data.extremes %>%
  mutate(Country = fct_reorder(Country, desc(mean_Lifeexpectancy))) %>%
  ggplot(aes(x=Country, weight=mean_Lifeexpectancy)) +
    geom_bar() +
    ggtitle("Life Expectancy vs. Country") +
    xlab("Country") +
    ylab("Life Expectancy") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 10, vjust=0.5  )
  )
```



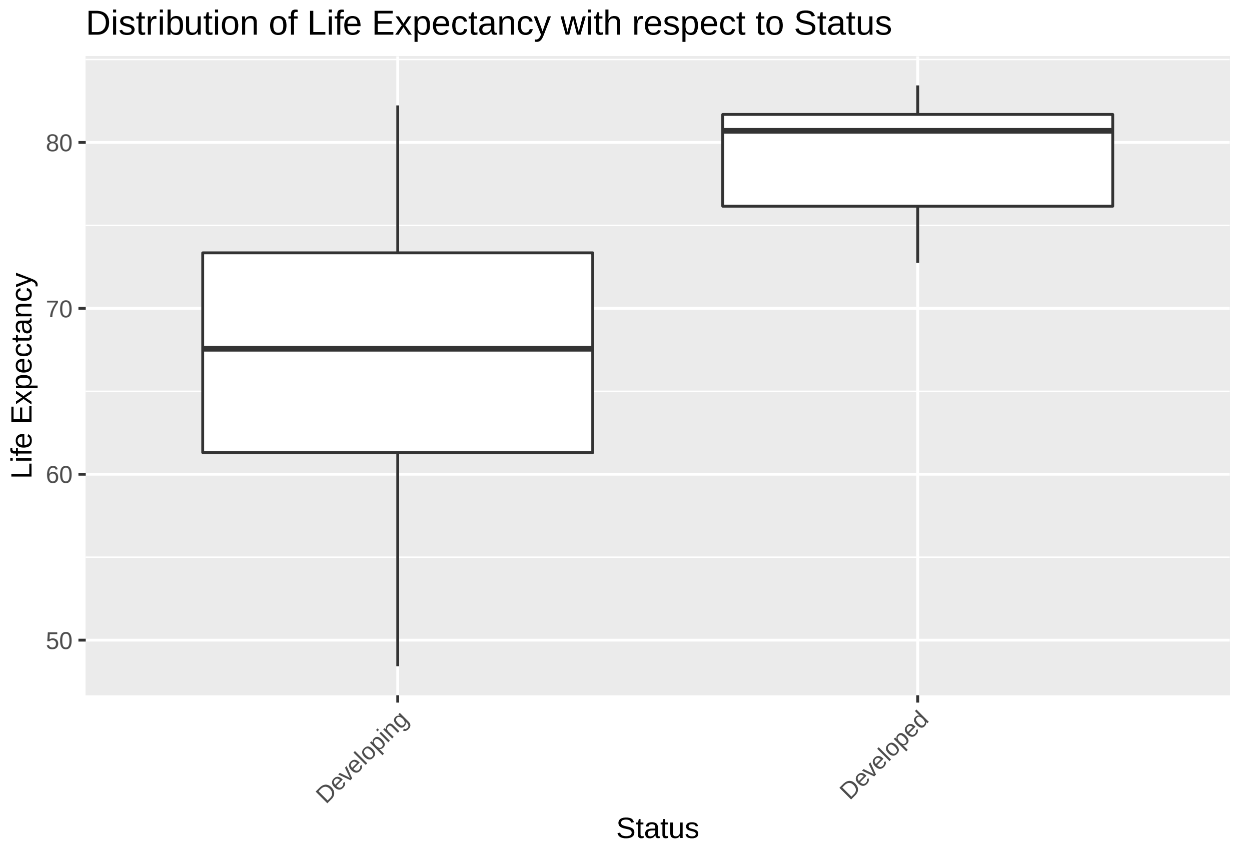
Indeed, we see that highly developed countries like Spain, Germany,... have higher life expectancy than more developing countries like Angola or Lesotho.

Life Expectancy by distribution by Status

Now we plot the distribution of Life Expectancy for each country separating by Status (developed or not developed).

```
aggregatedData_StatusSeparation = aggregatedData %>%
    mutate(mode_Status = fct_reorder(mode_Status, mean_Lifeexpectancy)) # reorder on status to have it

p = ggplot(aggregatedData_StatusSeparation, aes(x = mode_Status, y = mean_Lifeexpectancy))
p = p + geom_boxplot()
p = p + ggtitle("Distribution of Life Expectancy with respect to Status")
p = p + labs(x = 'Status', y = 'Life Expectancy')
p = p + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p
```



As expected, there is higher life expectancy in developed countries.

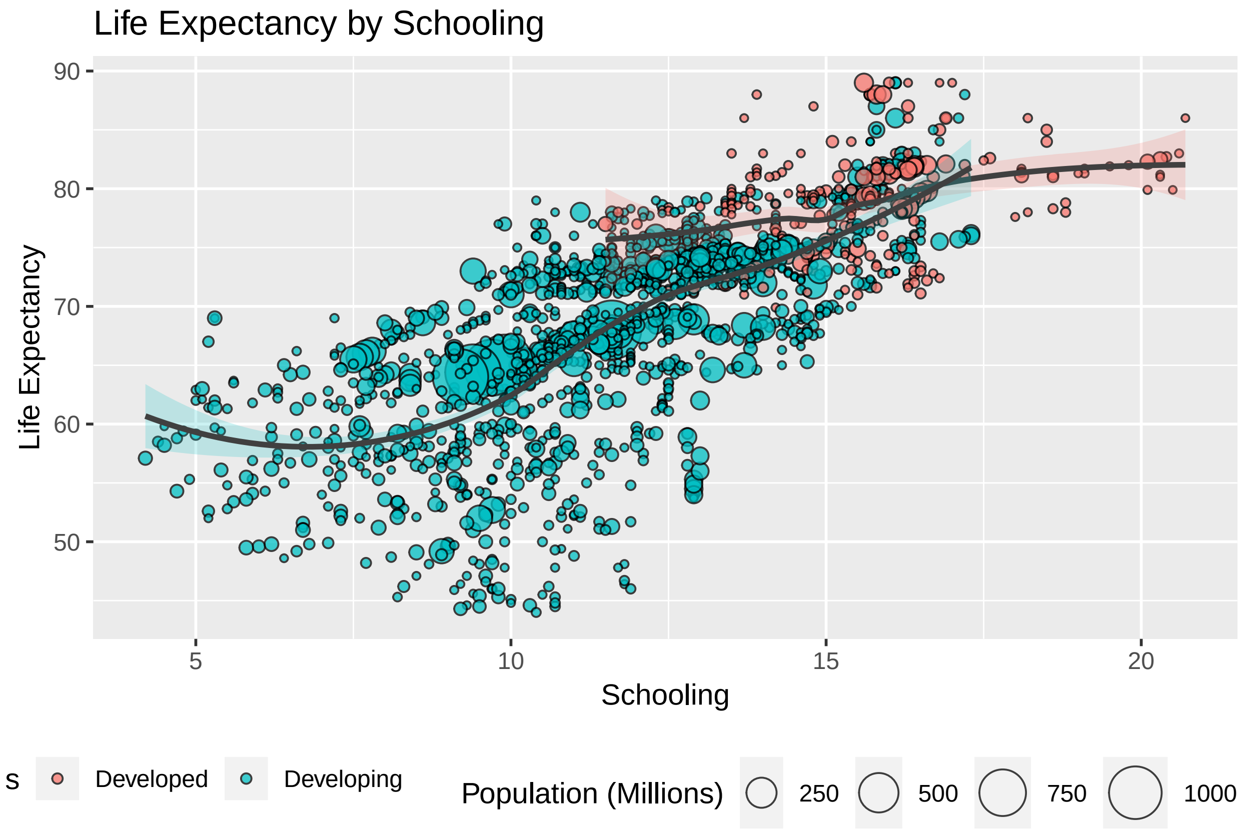
Now we will look more specifically at how Life Expectancy is related to each covariate of interest (Schooling, Adult Mortality and GDP), labeling for Developing or Developed, as that is an important factor as confirmed above.

With some simple exploration that confirms our intuitions, we move on to more interesting plots.

Life Expectancy vs. Schooling

Now we look at Life Expectancy vs. Schooling, separated by Status, for each country.

```
p = ggplot(data, aes(x=Schooling, y=Lifeexpectancy, size=Population/1000000, fill=Status))+
  geom_point(alpha=0.75, shape = 21)+
  geom_smooth(method="loess", color="gray25", alpha=.2, show.legend=FALSE)  +
  scale_size(range = c(1, 10), name="Population (Millions)")+
  labs(x="Schooling", y="Life Expectancy", title="Life Expectancy by Schooling") +
  theme(legend.position="bottom")
p
```

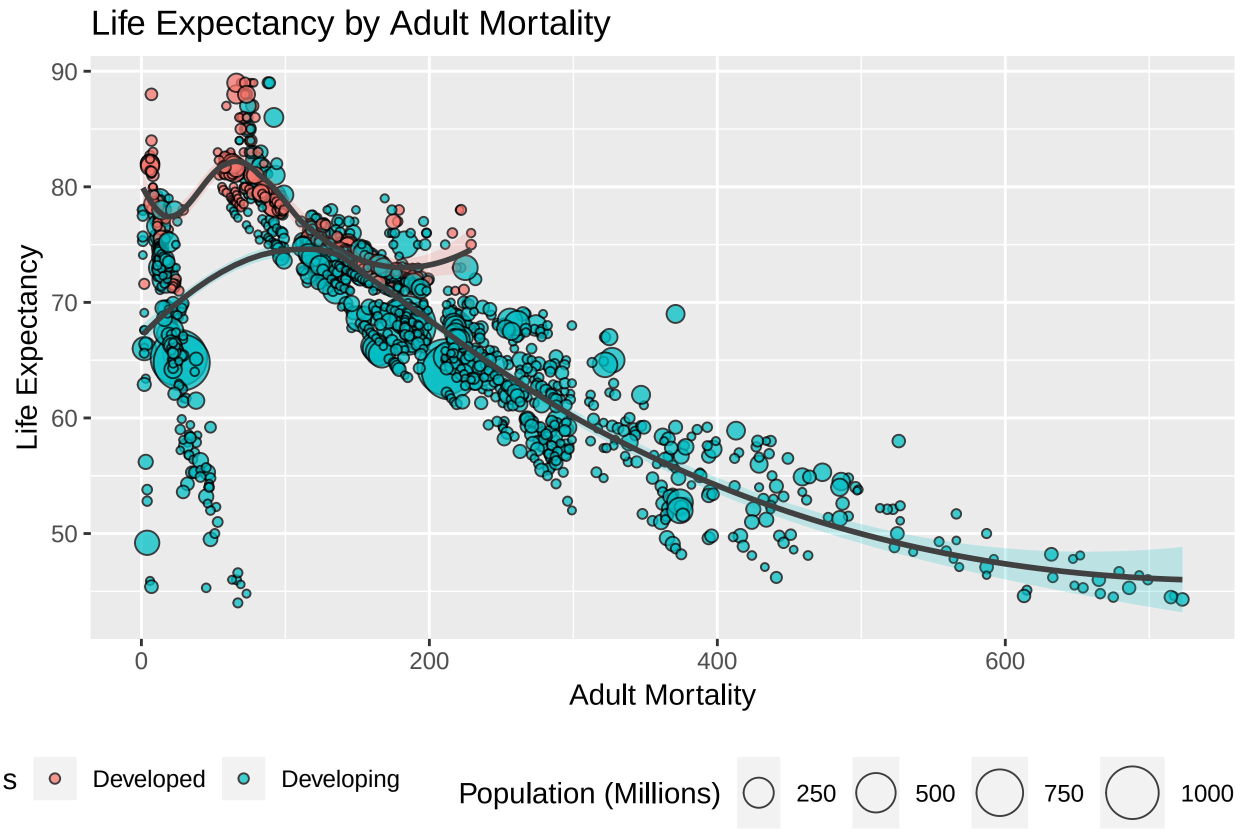


We note a general upward trend in that more schooling, or more education, is related to higher life expectancy. This would make sense in that if education can be focused on and done for this long, then a country most likely has the other basic needs taken care of for the most part. In fact, a quick google search (https://scholar.google.com/scholar?q=life+expectancy+and+education&hl=en&as_sdt=0&as_vis=1&oi=scholart) reveals several studies investigating this relationship, so there may be many more, say socioeconomic, factors to consider in this relationship as well.

Life Expectancy vs. Adult Mortality

Now we look at Life Expectancy vs. AdultMortality, separated by Status, for each country.

```
p = ggplot(data, aes(x=AdultMortality, y=Lifeexpectancy, size=Population/1000000, fill=Status))+
  geom_point(alpha=0.75, shape = 21)+
  geom_smooth(method="loess", color="gray25", alpha=.2, show.legend=FALSE)  +
  scale_size(range = c(1, 10), name="Population (Millions)")+
  labs(x="Adult Mortality", y="Life Expectancy", title="Life Expectancy by Adult Mortality") +
  theme(legend.position="bottom")
p
```

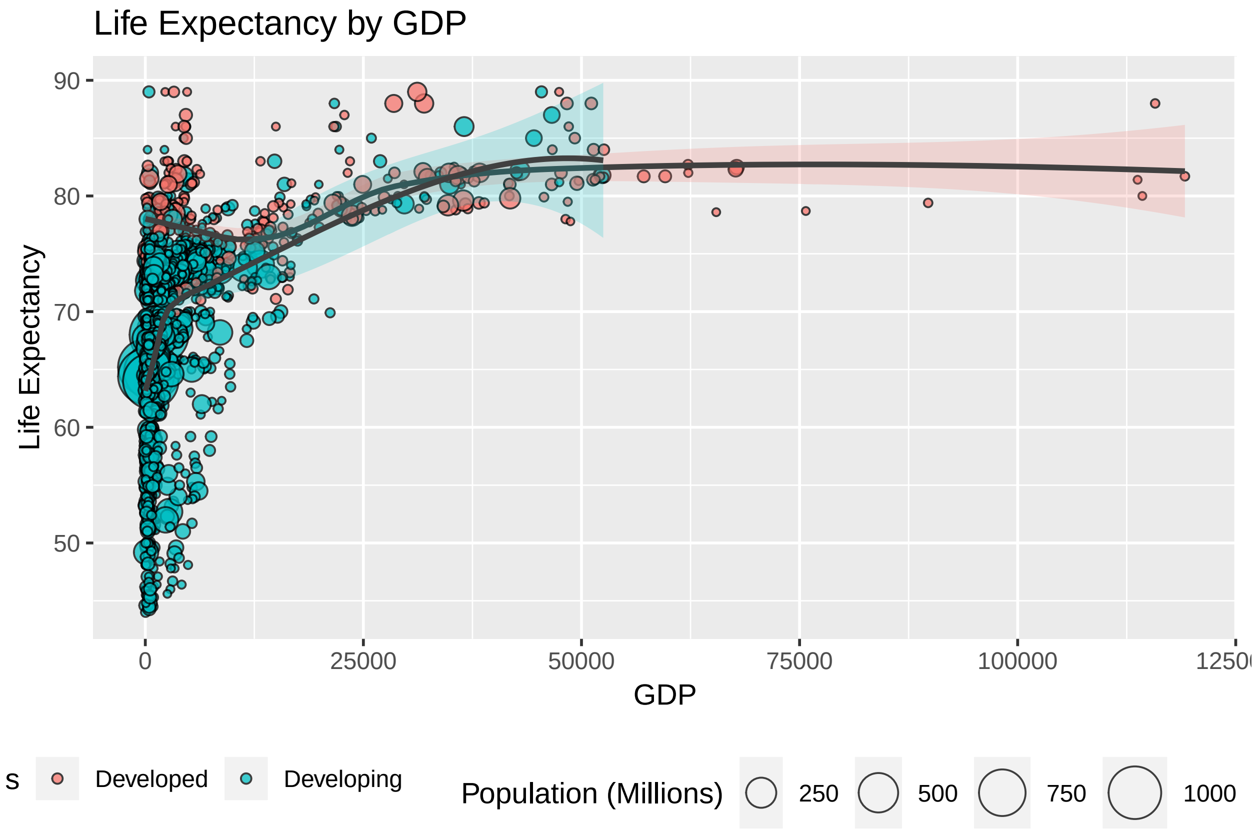


For the most part, we see higher adult mortality is related to lower life expectancies. In addition, it makes sense that the higher mortalities are related to Developing countries much more than Developed countries. However, we do see a few developed countries there with relatively higher adult mortalities and some developing countries with low adult mortality and high and low life expectancy despite lower adult mortality. This may be a result of the data quality, but further investigation into what countries these points are, and learning more about the country could be interesting.

Life Expectancy vs. GDP

Finally, we look at Life Expectancy vs. GDP, separated by Status, for each country.

```
p = ggplot(data, aes(x=GDP, y=Lifeexpectancy, size=Population/1000000, fill=Status))+
  geom_point(alpha=0.75, shape = 21)+
  geom_smooth(method="loess", color="gray25", alpha=.2, show.legend=FALSE)  +
  scale_size(range = c(1, 10), name="Population (Millions)")+
  labs(x="GDP", y="Life Expectancy", title="Life Expectancy by GDP") +
  theme(legend.position="bottom")
p
```



Based on this plot, higher GDP, or higher monetary value of finished goods and products in the country, tends to be related to higher life expectancy, and lower GDP is more related to lower life expectancy. High GDP could mean higher resources and work for citizens, allowing for greater sustenance of a steady economy, and thus a more developed country, leading to higher life expectancies with more basic needs met.

There are several cases where developing and developed countries have low GDP but high life expectancy. Further investigation into what countries these points are could be interesting, but it is worth noting that GDP itself can be a misleading metric (https://www.investopedia.com/articles/economics/08/genuine-progress-indicator-gpi.asp).

How does Life Expectancy of [selected] countries change over time?

Finally we look at how Life Expectancy of given countries change over time. For a start, we use 6 countries, namely the extremes in mean life expectancy.

```
data.filtered = filter(data, data$Country %in% c("Ireland", "Canada", "France", "Zimbabwe", "Lesotho",

p = ggplot(data.filtered, aes(x=Year, y=Lifeexpectancy, size=Population/1000000, fill=Country))+
  geom_point(alpha=0.75, shape = 21)+
  geom_smooth(method="loess", color="gray25", alpha=.2, show.legend=FALSE)  +
  scale_size(range = c(1, 10), name="Population (Millions)")+
  labs(x="Year", y="Life Expectancy", title="Life Expectancy by Year") +
  theme(legend.position="bottom")
p
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 2010
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 4.0804
```

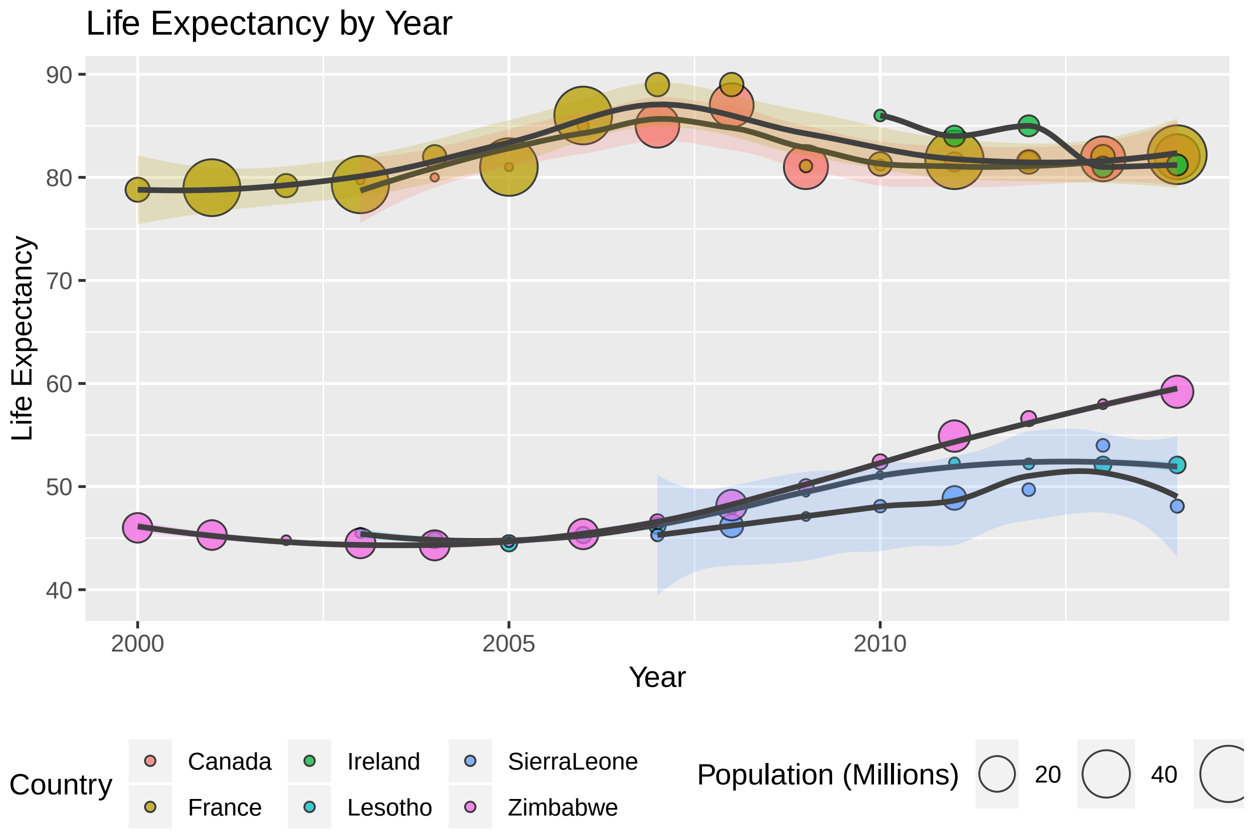
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small.
## fewer data values than degrees of freedom.
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
## at 2010
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 2.02
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
## condition number 0
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other
## near singularities as well. 4.0804
```



We can see that for the most part, as time goes on, developing countries are continuing to develop in the 'right' direction, as indicated by their increasing life expectancies. Although Sierra Leone and Zimbabwe have started to taper off. Sierra Leone, most likely due to the recently tamed Ebola outbreaks, and Lesotho most likely due to its political strife.

On the other hand, more developed countries have maintained high life expectancy, experiencing peaks around 2007-2008. These expectancies later tapered off, for a variety of factors. Interestingly enough, CNBC states 3 primary reasons for this in very well developed countries could be because of a rise in the following: * Drug Overdoses * Liver Disease * Suicide

(https://www.cnbc.com/2019/07/09/us-life-expectancy-has-been-declining-heres-why.html)

Next Steps

In the future, it would be interesting to see trajectories of the other covariates with respect to other countries over time (especially with the animation to really see how each changes year by year). Overall, there are many more relationships to analyze with respect to covariates and countries, and this offers a start plotting wise to continue into much more in depth analysis.

Of course, higher life expectancy does not exactly mean life is 'better' in that country, as with advancing technology and medicine, we are getting better at keeping people alive, and simply that - perhaps not really helping people to truly 'live' longer. So a more in depth analysis with a variety of other factors, perhaps reported happiness and autonomy in old age, could help narrow down on which countries really offer higher quality of life in old age as well. Regardless, past this more nuanced view of aging and quality of life, this data can provide several insights into overall prosperity of a country and general quality of life.