Life Expectancy in Kenya

Hello Reader. This Rmd-file is my first basic analysis of a dataset in my journey to learn R. I really enjoy learning, and learning with R. It is however difficult. As it is with all types of learning though, slow and steady progress brings you further towards your goal. This has taken many, many hours - far more than I thought it would take. I think many who have learnt R can relate to a similar experience in the first phase of their learning. However, this is my first example of how progress is slow, but progress nonetheless. This notebook gives some insight, perhaps more into me and my coding rather than the dataset itself. Thus, this serves as a display of my initial skills in R.

Please note that this dataset has some data inconsistencies and is not guaranteed to be accurate.

Ulrik

I start with:

```
library(tidyverse) #
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.0
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
life_expect <- read_csv("Life Expectancy Data.csv") # Reading in data
```

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

```
# Initial exploratory analysis. Just want to see what this data is made up
of.
view(life_expect)
str(life_expect)
```

```
## spc_tbl_ [2,938 × 22] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Country                        : chr [1:2938] "Afghanistan"
"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 ...
##  $ 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 0.415 0.405 ...
##  $ 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

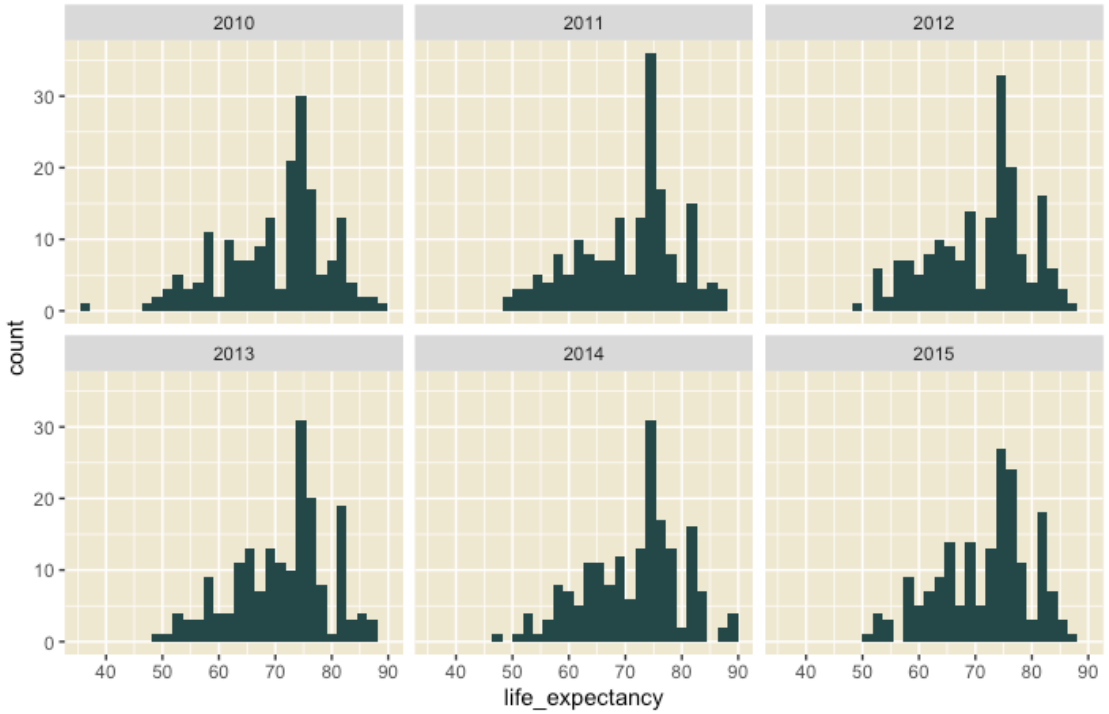
After loading in the data, I created a table with the columns I wanted to look at. These names were unpractical. Some had single quotes around them and they didn't have underscores between words.

```
# Creating a first basic selection of collumns I want to look at. Renaming
for ease.
basics_df <- life_expect %>%
  rename(
    country = Country,
    year = Year,
    life_expectancy = `Life expectancy`,
    status = Status,
    adult_mortality = `Adult Mortality`,
    infant_mortality = `infant deaths`,
    population = Population
  )
```

Let's first make a general histogram of what the general life expectancy is. To not overpopulate the graph too much, we're only looking at the period 2010-2015.

```
life_year <- basics_df %>%
  select(life_expectancy,
         year) %>%
```

```
  filter(year >= 2010)

ggplot(life_year) +
  geom_histogram(aes(x = life_expectancy), fill = "#244747") +
  theme(
    panel.background = element_rect(fill = "#efe8d1")) +
  facet_wrap(~ year)

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

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

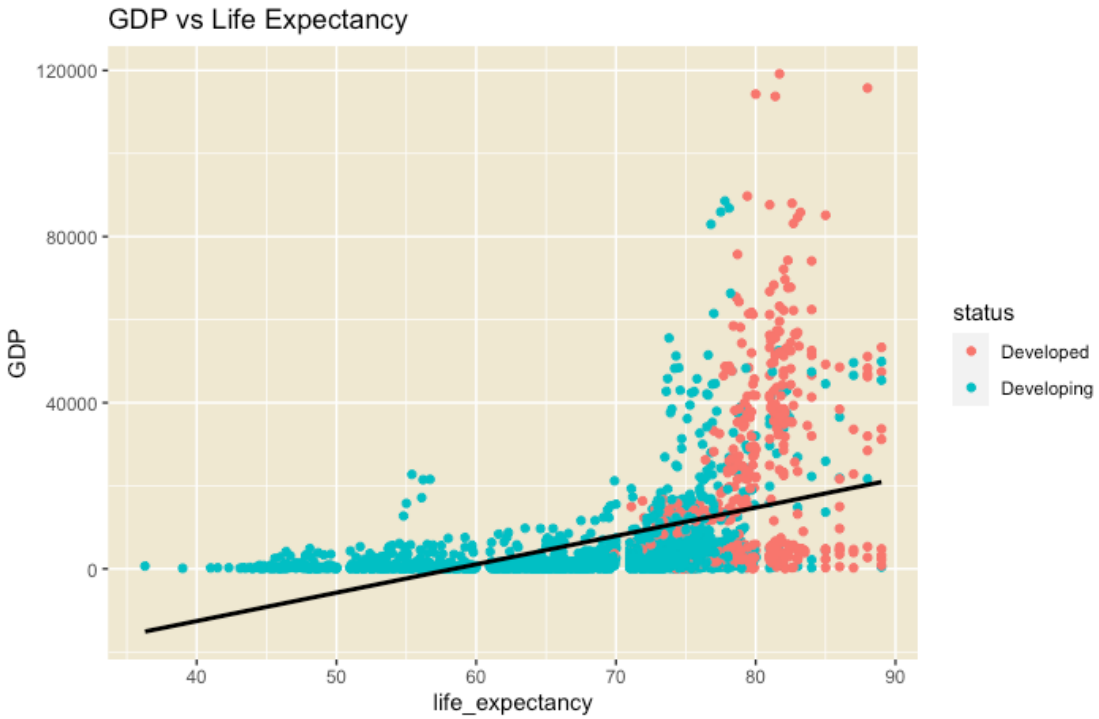


This histogram shows the number of countries and their corresponding life expectancy. It's an interesting chart, perhaps most interestingly is the large spike in the ≈ 75 range. However, much is unsaid from this graph. Let's move on.

I'm fascinated by what impacts life expectancy. The most obvious assessment is that GDP will correlate with the life expectancy. Let's see if that's the case with this dataset.

```
gdp_life <- basics_df %>% # Creating my dataset.
  select(life_expectancy,
         GDP,
         status,
         country)

# I'm mapping this out with a scatter plot.
ggplot(gdp_life) +
  geom_point(mapping = aes(x = life_expectancy, y = GDP, color = status)) +
  theme(
```

```
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "GDP vs Life Expectancy") +
  geom_smooth(method = lm, formula = y ~ x, aes(life_expectancy, GDP), color
= "black", se = FALSE) # Simple regression added to this model.

## Warning: Removed 453 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 453 rows containing missing values (`geom_point()`).
```



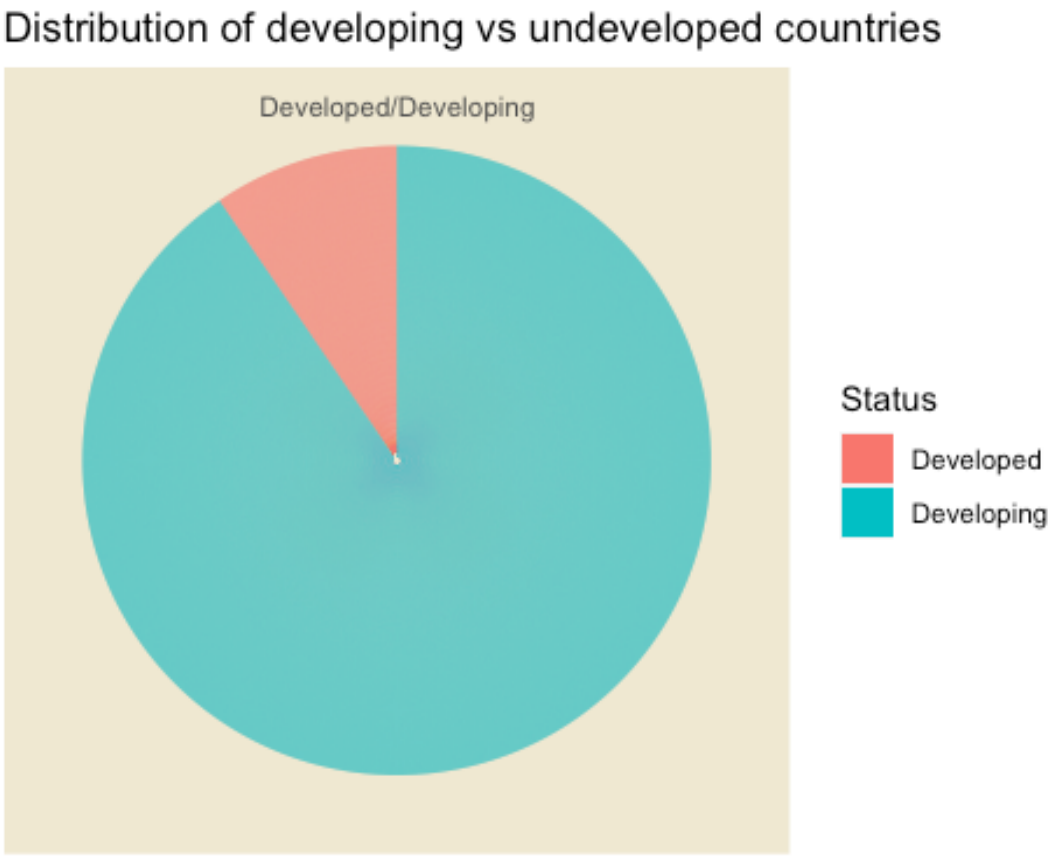
So this graph is quite interesting to me. What we can clearly see is that there are many countries that have a low life expectancy, and they hava correspondingly low GDP, and vice versa. There are however exceptions as can be seen. The simple linear regression show's the simple trend between higher life expectancy and GDP. In other words - our earlier assumption was correct.

This graph also discerns between "developed" and "developing" countries and color them accordingly. Here we see that developed countries exclusively live longer, whilst often having higher GDP as well.

I want to dive into this more since this data set provides it. Note that this is without a context for what a "developing" or "developed" country is. So, what is the distribution of developed vs developing countries?

```
ggplot(basics_df) +
  geom_bar(aes(x = "", y = status, fill = status), stat = "identity") +
  coord_polar("y", start = 0) +
  theme(
```

```
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank(),
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "Distribution of developing vs undeveloped countries") +
  scale_fill_discrete(name = "Status")
```

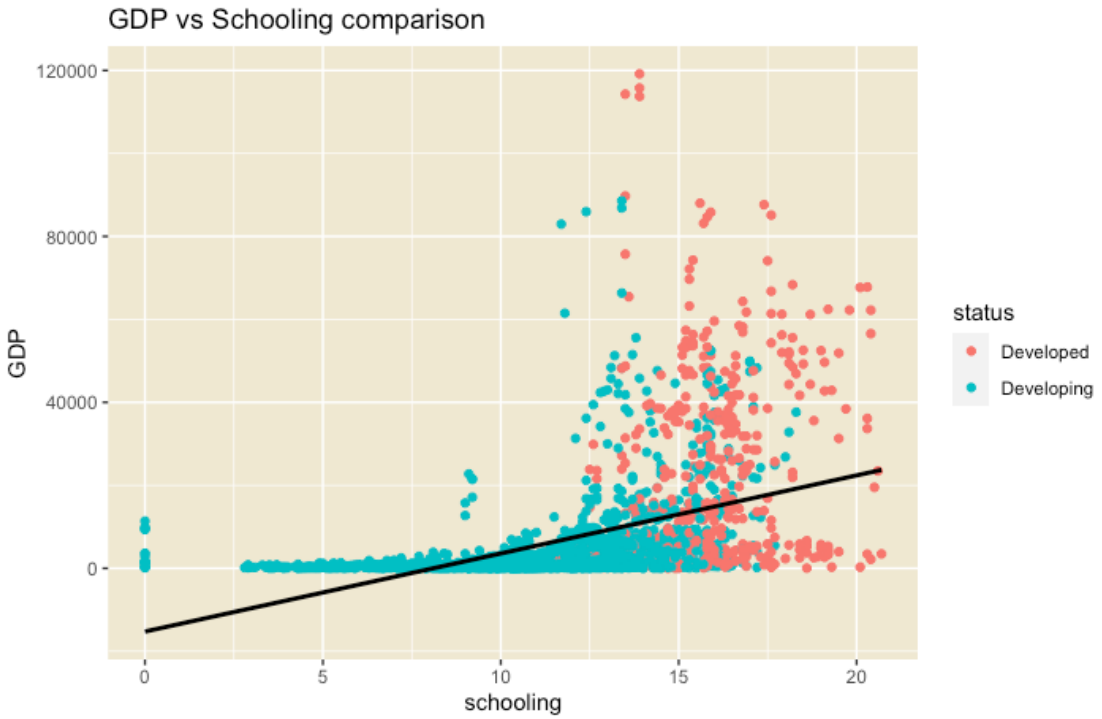


This pie chart shows that there's far more developing countries than there are developed countries. In my eyes, it's far more interesting to look at the developing countries, so that's what we're going to be doing now.

It's natural to think that higher GDP will have an impact on whether a country is developed or developing. But what can influence the GDP? Well, let's see. First: GDP vs schooling

```
gdp_school <- basics_df %>%
  select(
    Schooling,
    GDP,
    status)

gdp_school_ren <- gdp_school %>% # Again, renaming for some consistency.
```

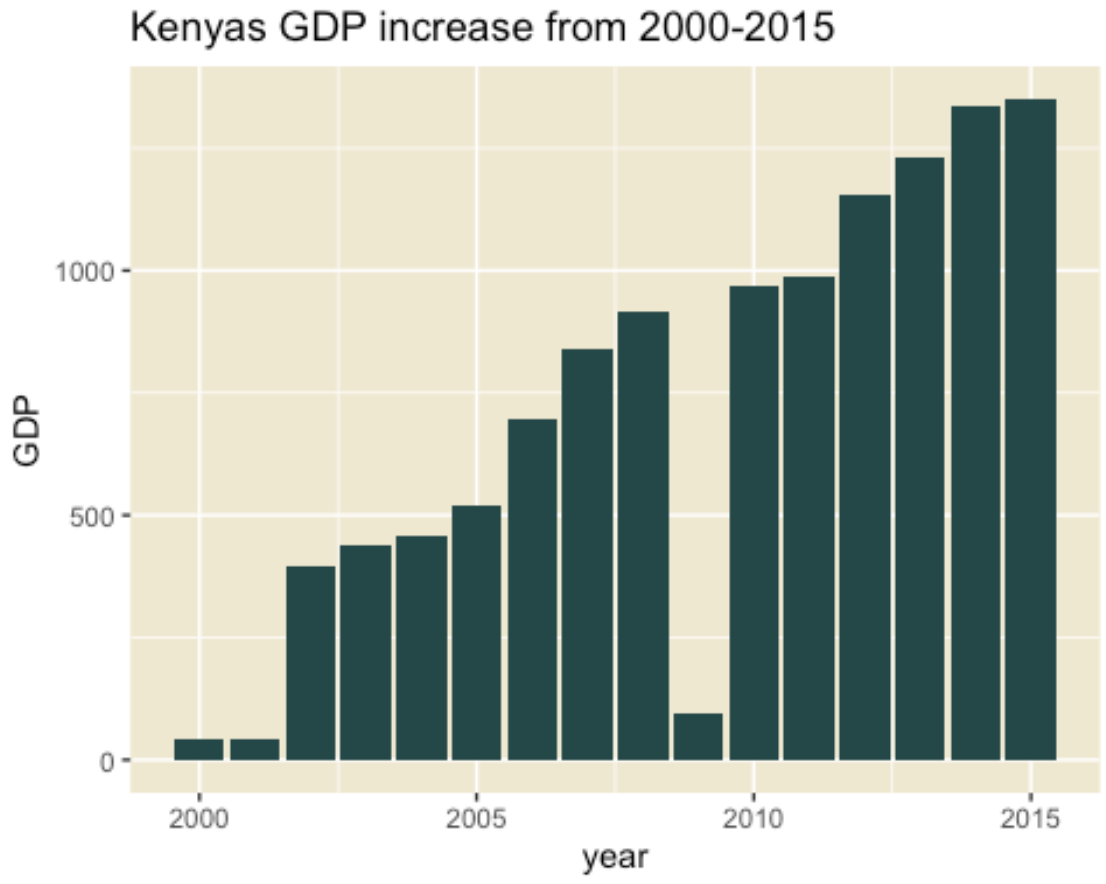
```
  rename(
    schooling = Schooling)

ggplot(gdp_school_ren) +
  geom_point(mapping = aes(x = schooling, y = GDP, color = status)) +
  theme(
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "GDP vs Schooling comparison") +
  geom_smooth(method = lm, formula = y ~ x, aes(schooling, GDP), color =
"black", se = FALSE)

## Warning: Removed 451 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 451 rows containing missing values (`geom_point()`).
```

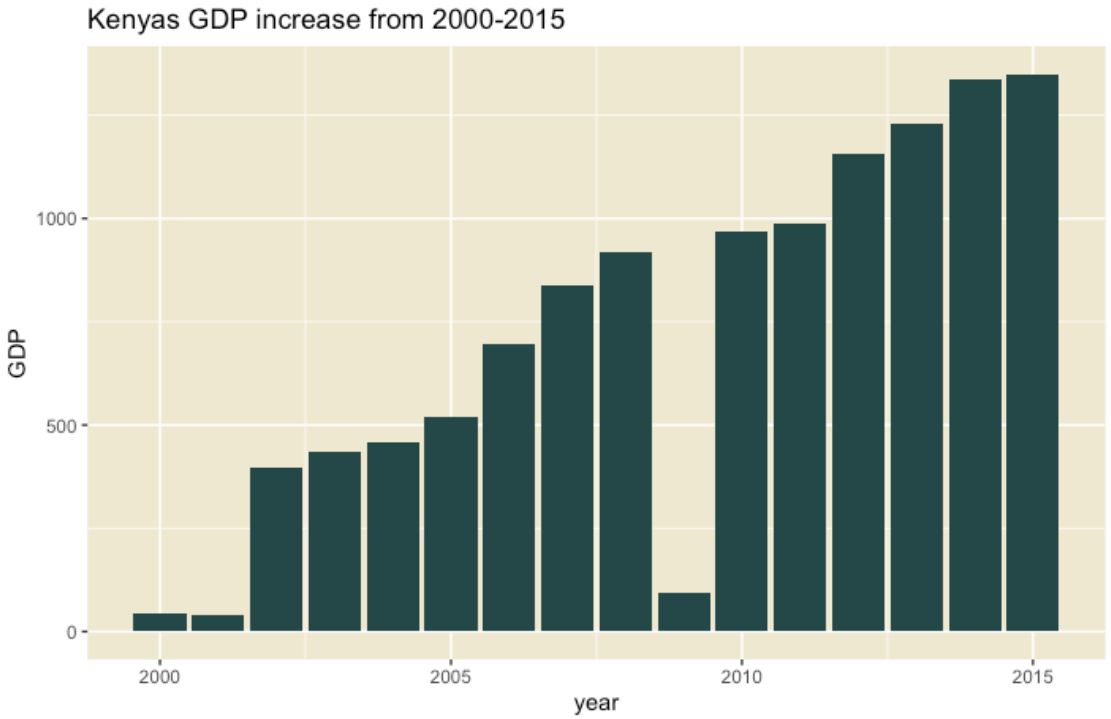


Similar trend to the previous scatter plot. Higher levels of schooling means higher GDP. Looking at these kinds of population-based metrics is quite interesting, and is something I want to dive a little deeper into, specifically with Kenya as a case.

```
kenya_df <- basics_df %>%
  select(
    country,
    population,
    year,
    Schooling,
    Diphtheria,
    Polio,
```

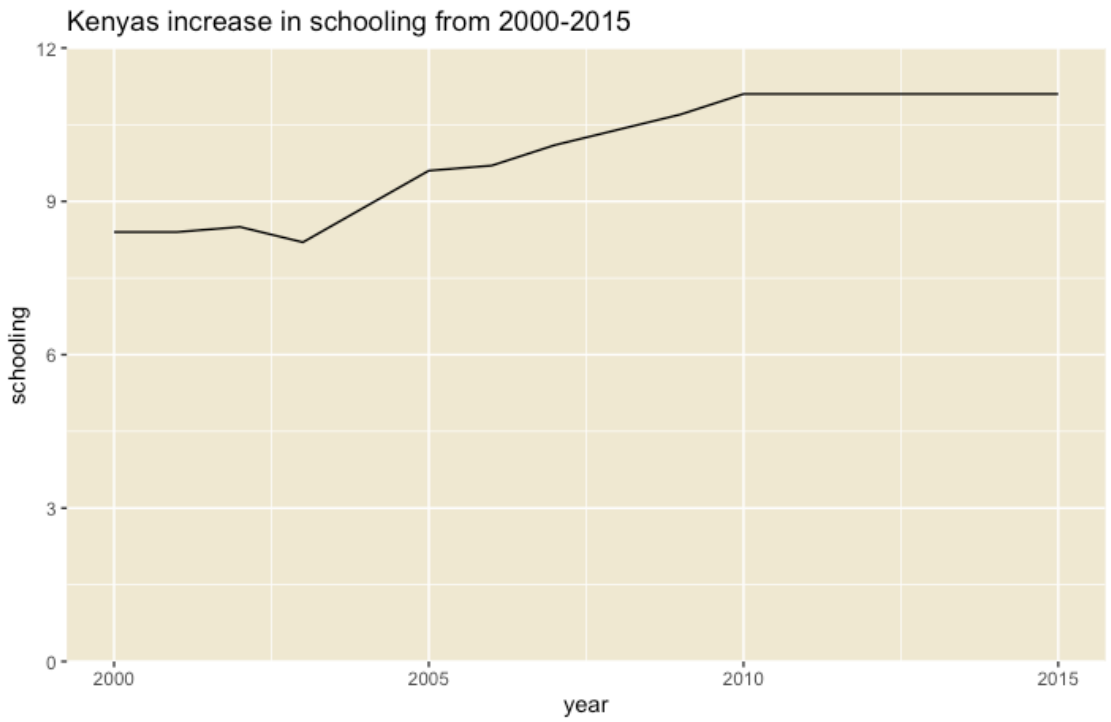
```
    `Hepatitis B`,
    Measles,
    `HIV/AIDS`,
    infant_mortality,
    adult_mortality,
    GDP) %>%
  rename(diphteria = Diphtheria, # again, rename for ease.
    schooling = Schooling,
    polio = Polio,
    hep_b = `Hepatitis B`,
    measles = Measles,
    HIV_AIDS = `HIV/AIDS`) %>%
  filter(country == "Kenya")

ggplot(kenya_df) +
  geom_col(aes(x = year, y = GDP), fill = "#244747") +
  theme(
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "Kenyas GDP increase from 2000-2015")
```





As we can see, this dataset has inaccuracies which I haven't been able to resolve. Despite that however, it's clear that Kenya has had a stark increase in GDP over the period from 2000-2015. May this be impacted by schooling as previously charted?

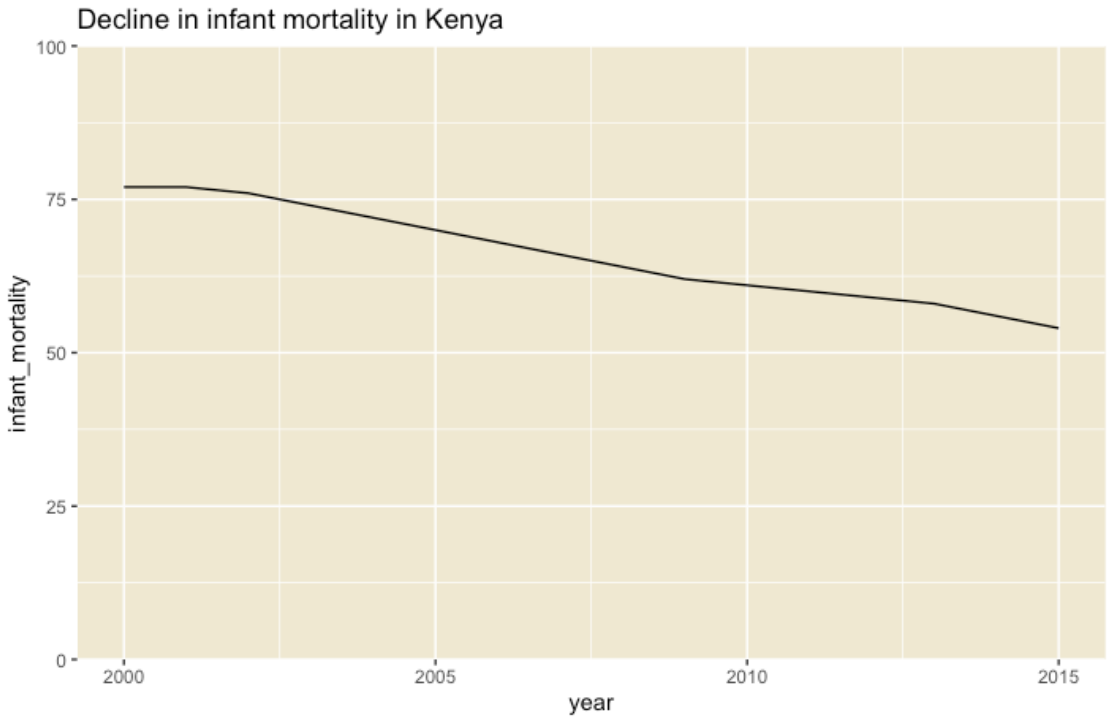
```
kenya_schools <- kenya_df %>%
  select(
    schooling,
    year,
    GDP)

ggplot(kenya_schools) +
  geom_line(mapping = aes(x = year, y = schooling)) +
  theme(
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "Kenyas increase in schooling from 2000-2015") +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 12))
```



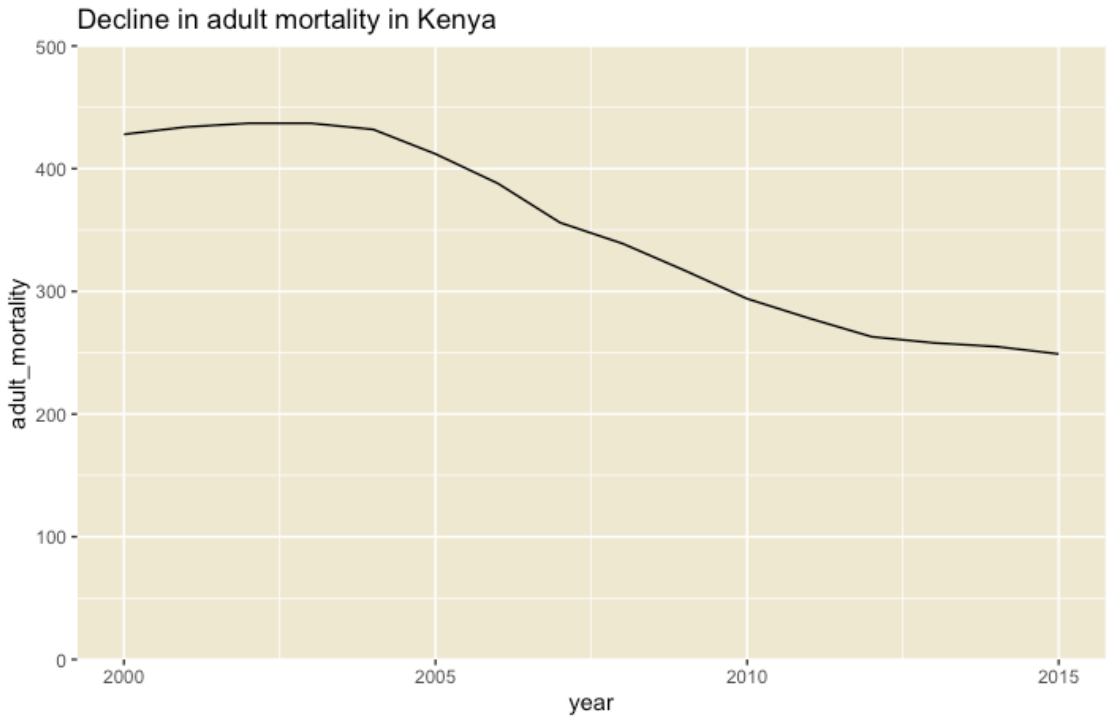
At the very least we see a stark increase in the level of schooling over this period. However, as with many of these analyses, it's hard to discern the complete picture. Regardless, we have up to this point shown how things move in the positive direction for Kenya, such as the GDP and level of schooling.

On the opposite end I would like to finish with some charts on the mortality rates for adults and children. These have declined in this period.

```
mortality_yy <- basics_df %>%
  select(
    country,
    infant_mortality,
    adult_mortality,
    year) %>%
  filter(
    country == "Kenya"
  )

ggplot(mortality_yy) +
  geom_line(aes(x = year, y = infant_mortality), color = "black") +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 100)) +
  theme(
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "Decline in infant mortality in Kenya")
```




```
ggplot(mortality_yy) +
  geom_line(aes(x = year, y = adult_mortality), color = "black") +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 500)) +
  theme(
    panel.background = element_rect(fill = "#efe8d1")) +
  labs(title = "Decline in adult mortality in Kenya")
```



This analysis has been quite small and is very limited, but I hope it can serve as some introduction to me and my skills with R.

Thanks,

Ulrik