M130 Lab Project

Jack Fogliasso

September 25, 2018

```
knitr::opts_chunk$set(echo = TRUE, warning=FALSE, message=FALSE)
library(dplyr)
library(tidyr)
library(knitr)
library(ggplot2)
```

Introduction

I am looking at a couple of datasets from the World Bank. The variables I will be looking at are fertility rates and infant mortality rates in various countries, in various years. Total fertility rate is given as births per woman, per year and infant mortality rate is the number of infant deaths (before the age of 1) per 1,000 live births, per year.

First I need to import, clean and reshape the data. Since I am using multiple datasets, I utilised `list.files()` and `read.csv()` to retrieve the names of the files in my data folder and to read them in. Then I combined the two datasets. Next, I selected only the most recent data, from the years 2000-2016. Then I reshaped the data. Next, I added the "metadata" – I was specifically interested in the variables `Region` (the geographic region of each country) and `IncomeGroup` (a category based on a country's income) and their relationships with the two variables of interest. Finally, I gave new names to the variables and changed the levels for `income` to be in order of increasing income.

```
myList <- list.files(path="C:/Users/Jack Fogliasso/Documents/F18/MATH130/lab project/data/",
                     pattern = "*.csv", full.names = TRUE)

a <- read.csv(myList[1], skip=4, header=TRUE)
b <- read.csv(myList[2], skip=4, header=TRUE)
metadata <- read.csv(myList[3], header=TRUE)
a$var <- "mortality"
b$var <- "fertility"

raw <- rbind(a,b)

rawLastDecade <- raw %>%
                 dplyr:: select(Country.Code, var, X2000: X2016) %>%
                 mutate(var = factor(var))
rawLong <- gather(rawLastDecade, year, rate, X2000:X2016, factor_key=TRUE)
clean <- spread(rawLong, var, rate)
names(clean) <- c("countryCode", "year", "fertility", "mortality")
clean$year2 <- substr(clean$year, 2, 5)
clean$year <- clean$year2
clean$year2 <- NULL

metadata$countryCode <- metadata$ï..Country.Code
clean2 <- clean %>% left_join(metadata)
clean2 <- select(clean2, -(SpecialNotes:X), -ï..Country.Code) %>%
          filter(Region != "", IncomeGroup !="") %>%
          droplevels()
names(clean2) <- c("countryCode", "year", "fertility", "mortality", "region", "income")
clean2$income <- factor(clean2$income, levels = c("Low income", "Lower middle income",
                                                   "Upper middle income", "High income"))

clean <- clean2
clean2 <- NULL
```

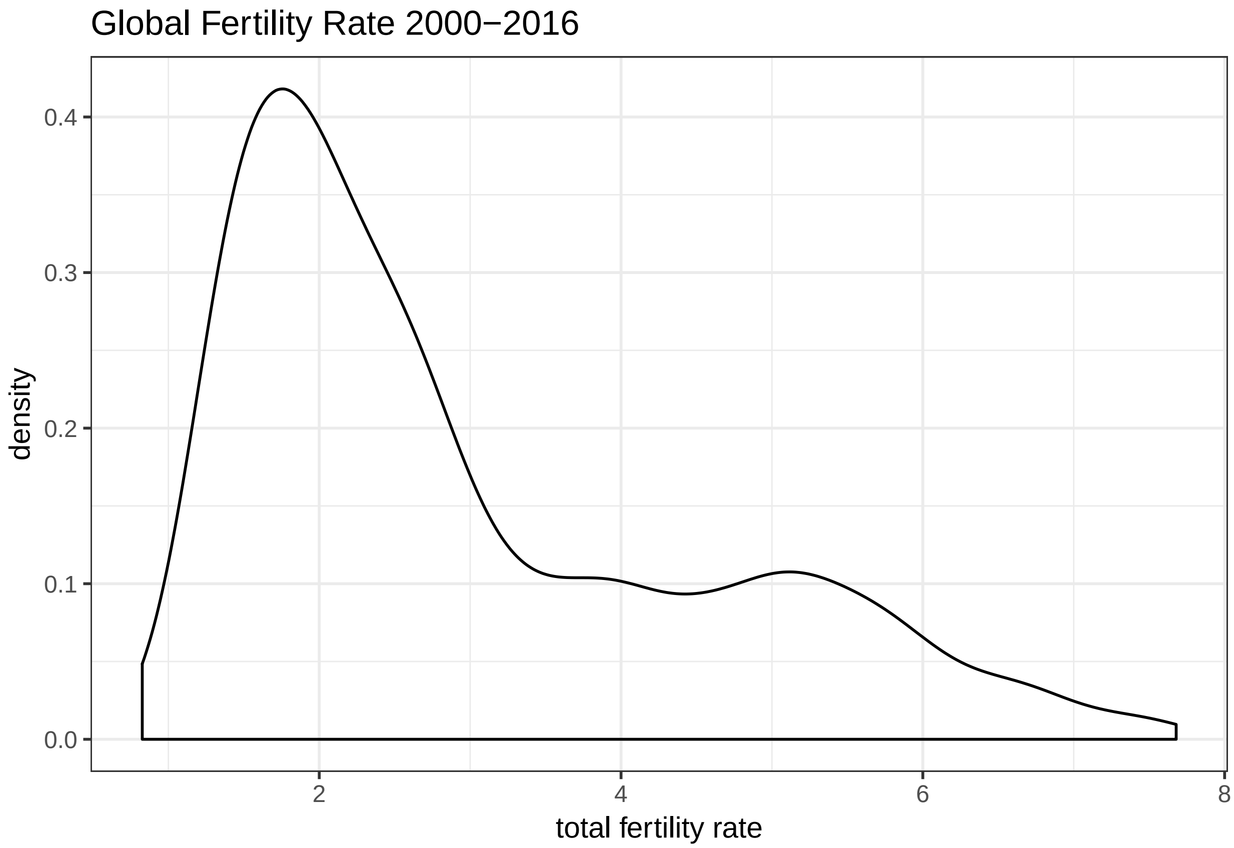
Univariate

Fertility

```
kable(t(round(summary(clean$fertility),3)))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.827	1.76	2.417	2.959	3.972	7.679	272

```
ggplot(clean, aes(x=fertility)) + geom_density() + theme_bw() + xlab("total fertility rate") +
  ggtitle("Global Fertility Rate 2000-2016")
```



During 2000-2016, the mean fertility rate was about 3 births per woman, with a median of 2.4

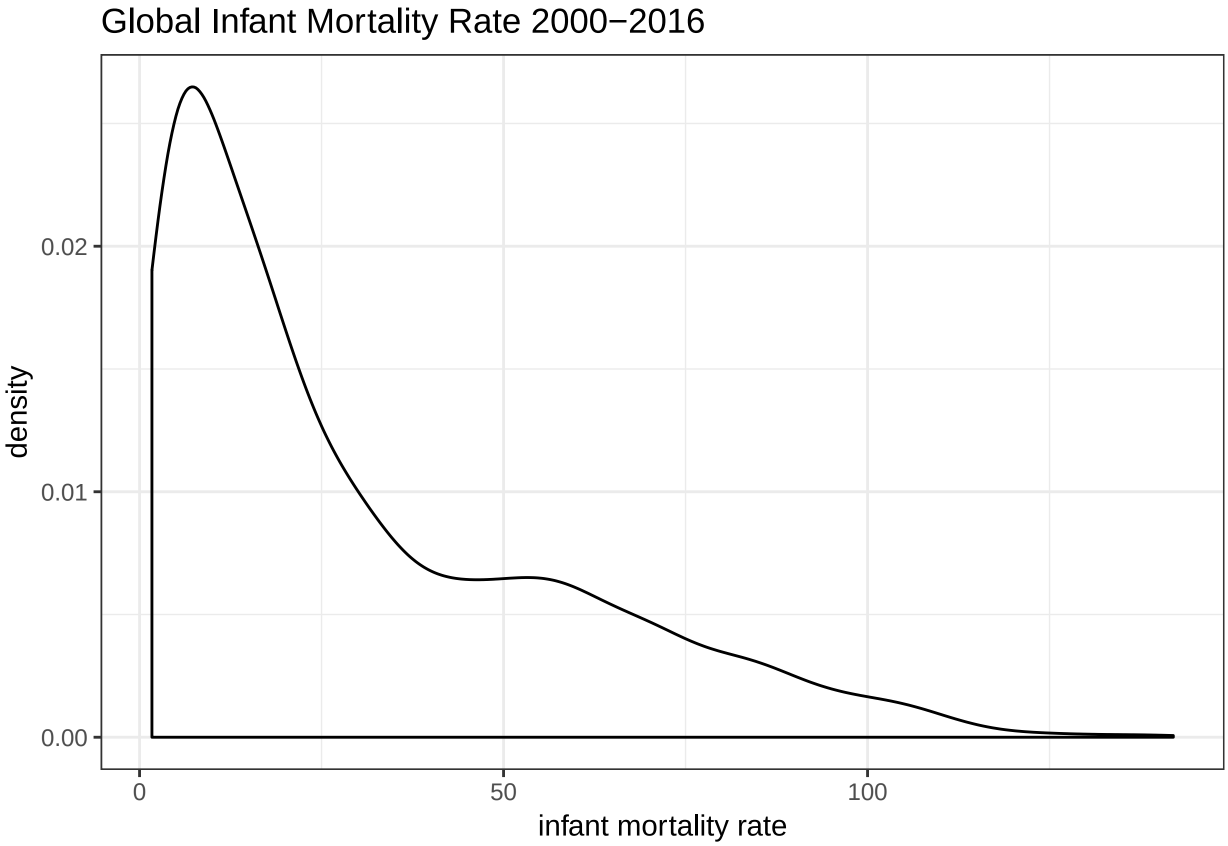
and range of 6.9. The variable is somewhat right-skewed.

Infant Mortality

```
kable(t(round(summary(clean$mortality),3)))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1.7	8	19.6	30.04	47.1	142	408

```
ggplot(clean, aes(x=mortality)) + geom_density() + theme_bw() + xlab("infant mortality rate") +
  ggtitle("Global Infant Mortality Rate 2000-2016")
```



Infant mortality rate is somewhat right skewed, with a mean of 30 that is to the right of the median, 19.6.

Region and Income

```
kable(table(clean$region), col.names=c("Region","Count"))
```

Region	Count
East Asia & Pacific	629
Europe & Central Asia	986

Region	Count
Latin America & Caribbean	714
Middle East & North Africa	357
North America	51
South Asia	136
Sub-Saharan Africa	816

```r
kable(round(prop.table(table(clean$income)),3), col.names=c("Income", "Frequency"))
```

Income	Frequency
Low income	0.157
Lower middle income	0.217
Upper middle income	0.258
High income	0.369

The region with the most countries in this dataset is Europe/Central Asia, while the region with the fewest countries is North America. As for income, there seems to be a trend toward higher income.

Bivariate

I used `dplyr` to create some grouped summary statistics before creating graphs.

Fertility by year

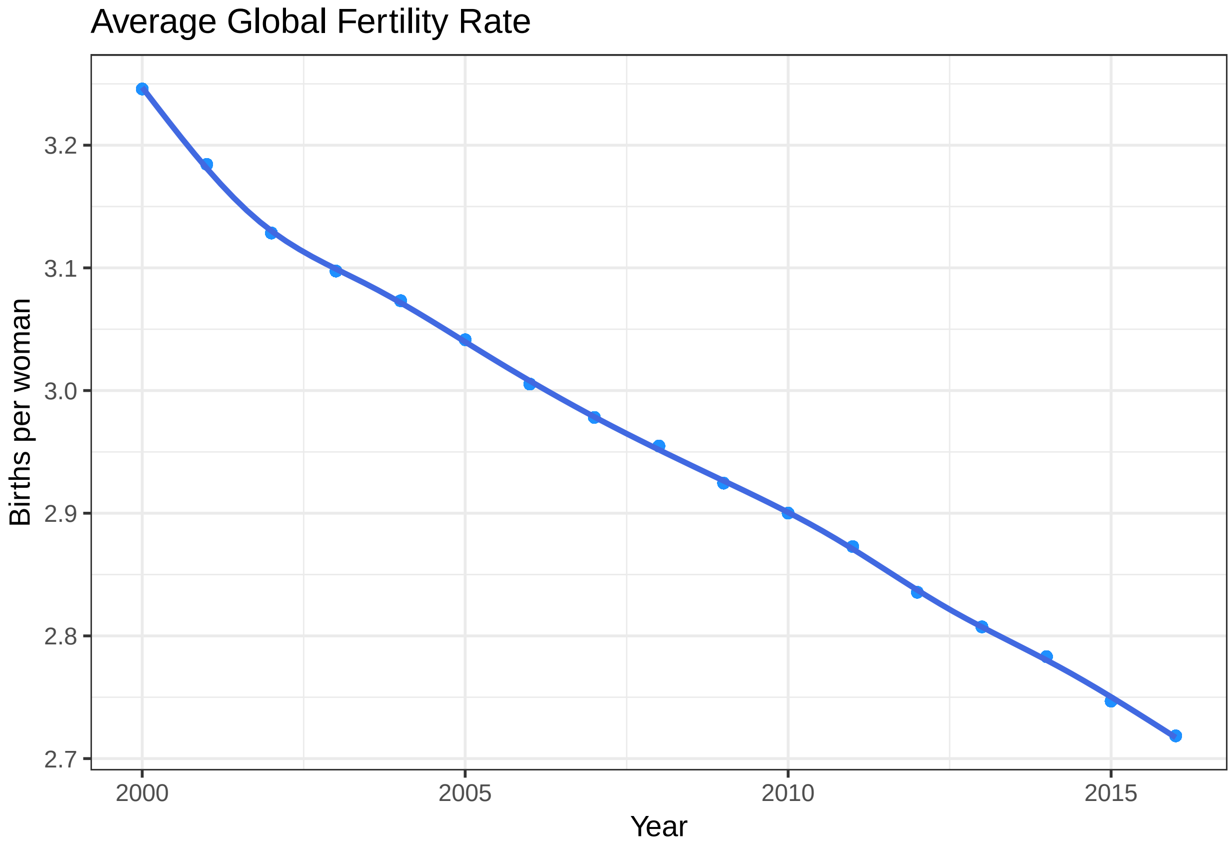
```r
byYear <- group_by(clean,year)

kable(summarise(byYear, avgFertRate = mean(fertility, na.rm=TRUE),
                median = median(fertility, na.rm=TRUE),
                range = max(fertility, na.rm=TRUE)- min(fertility, na.rm=TRUE)),
                digits = 1, col.names = c("year", "average", "median", "range"),
                caption="Global Fertility Rate Stats", align=c('c','c','c','c'))
```

Table 5: Global Fertility Rate Stats

year	average	median	range
2000	3.2	2.8	6.7
2001	3.2	2.7	6.8
2002	3.1	2.6	6.8
2003	3.1	2.5	6.8
2004	3.1	2.5	6.8
2005	3.0	2.5	6.8
2006	3.0	2.5	6.7
2007	3.0	2.4	6.7
2008	3.0	2.4	6.6
2009	2.9	2.4	6.5
2010	2.9	2.3	6.4
2011	2.9	2.3	6.3

year	average	median	range
2012	2.8	2.3	6.3
2013	2.8	2.3	6.3
2014	2.8	2.2	6.1
2015	2.7	2.2	6.1
2016	2.7	2.2	6.1

```r
group_by(clean,year) %>%
  mutate(avgFert = mean(fertility, na.rm=TRUE)) %>%
  ggplot(aes(x=as.numeric(year), y=avgFert)) +
  geom_point(color="dodgerblue") + geom_smooth(color="royalblue") +
  labs(title="Average Global Fertility Rate",
                 x="Year", y="Births per woman") +
  theme_bw()
```



Average global fertility rates are declining, going from 3.2 to 2.7 births per woman.

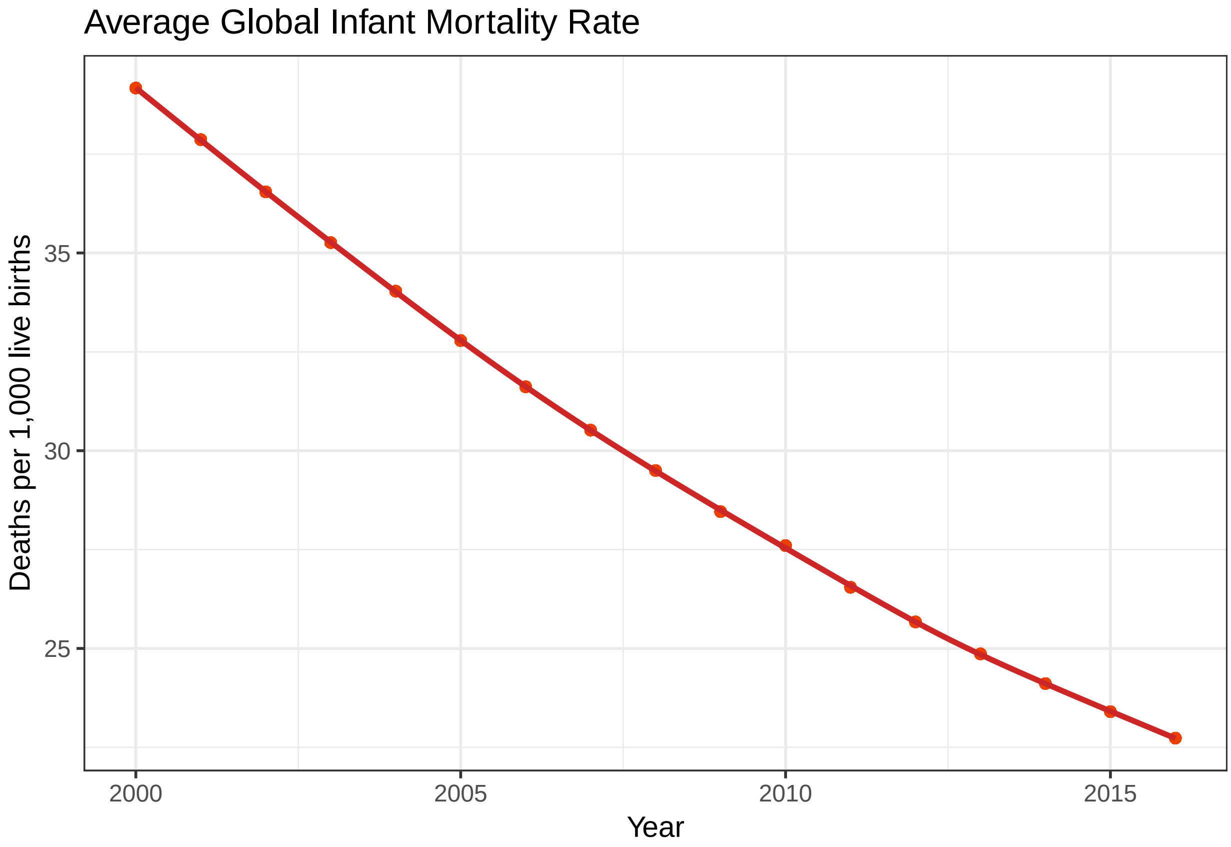
Infant mortality by year

```r
kable(summarise(byYear, avgMortRate = mean(mortality, na.rm=TRUE),
                median = median(mortality, na.rm=TRUE),
                range = max(mortality, na.rm=TRUE)- min(mortality, na.rm=TRUE)),
                digits = 1,
                col.names = c("year", "average", "median", "range"),
                caption="Global Infant Mortality Rate Stats",
                align=c('c','c','c','c'))
```

Table 6: Global Infant Mortality Rate Stats

year	average	median	range
2000	39.2	26.5	139.0
2001	37.9	25.2	136.7
2002	36.5	24.5	134.1
2003	35.3	23.4	131.4
2004	34.0	22.2	128.6
2005	32.8	20.8	125.5
2006	31.6	20.0	122.1
2007	30.5	19.5	118.5
2008	29.5	18.6	114.7
2009	28.5	17.6	110.6
2010	27.6	17.2	106.4
2011	26.5	16.4	102.1
2012	25.7	16.1	97.9
2013	24.9	15.7	94.5
2014	24.1	15.8	92.1
2015	23.4	15.5	89.9
2016	22.7	15.1	87.5

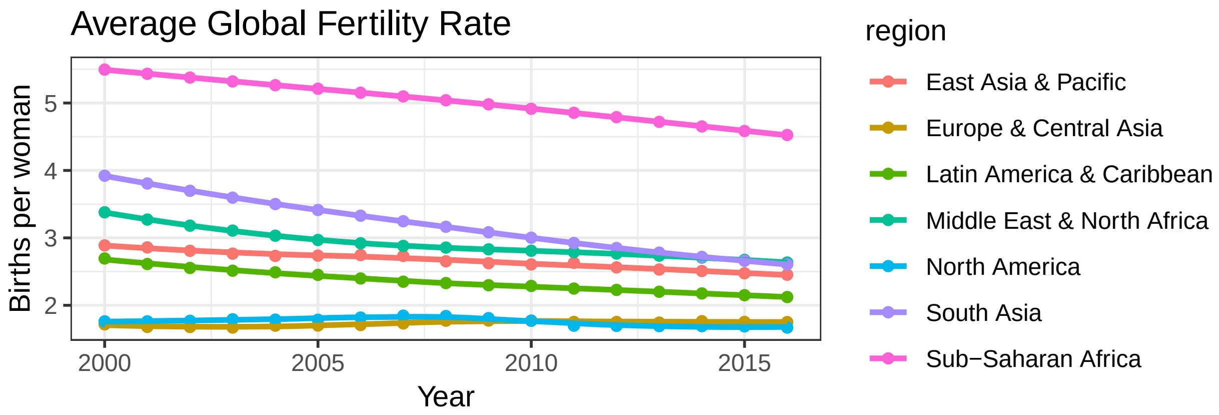
```
group_by(clean,year) %>%
  mutate(avgMort = mean(mortality, na.rm=TRUE)) %>%
  ggplot(aes(x=as.numeric(year), y=avgMort)) +
  geom_point(color="orangered2") + geom_smooth(color="firebrick3") +
  labs(title="Average Global Infant Mortality Rate",
                 x="Year", y="Deaths per 1,000 live births") +
  theme_bw()
```

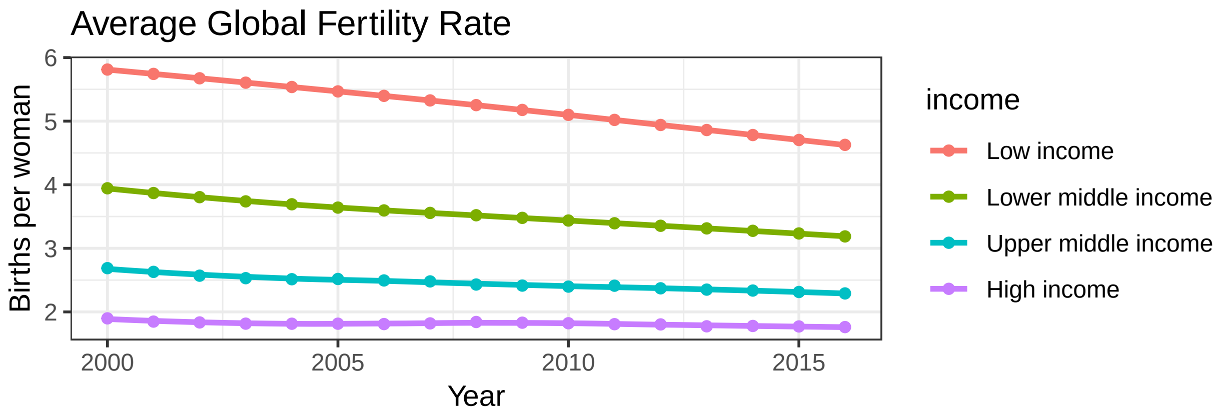


Average infant mortality decreased from 39.2 deaths per 1,000 live births to 22.7, with the range decreasing as well.

Fertility rates, income and region

```
plot1 <- group_by(clean, year, region) %>%
  summarise(avgFert = mean(fertility, na.rm=TRUE)) %>%
  ggplot(aes(x=as.numeric(year), y=avgFert, colour = region)) +
  geom_point() + geom_smooth(se=FALSE) +
  labs(title="Average Global Fertility Rate",
                 x="Year", y="Births per woman") +
  theme_bw()

plot2 <- group_by(clean, year, income) %>%
  summarise(avgFert = mean(fertility, na.rm=TRUE)) %>%
  ggplot(aes(x=as.numeric(year), y=avgFert, colour = income)) +
  geom_point() + geom_smooth(se=FALSE) +
  labs(title="Average Global Fertility Rate",
                 x="Year", y="Births per woman") +
  theme_bw()

gridExtra::grid.arrange(plot1, plot2)
```



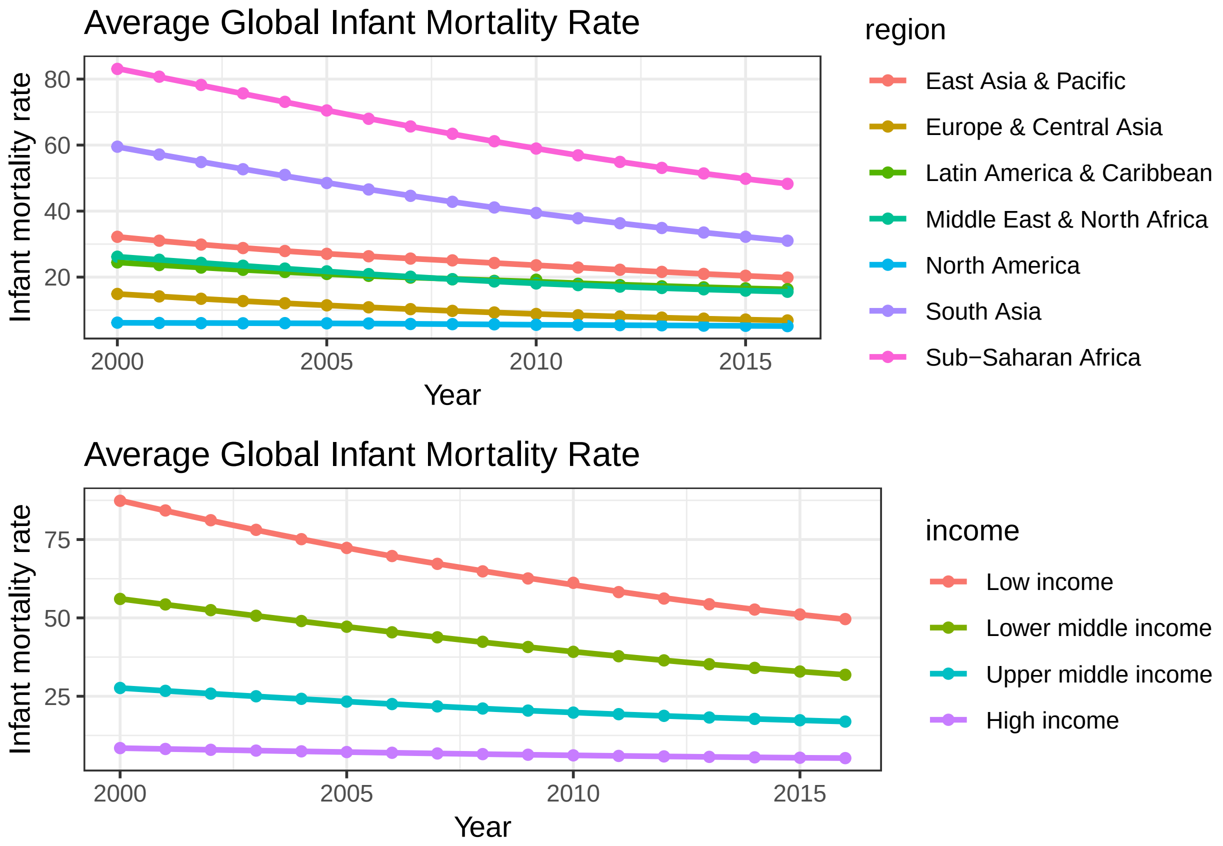


These graphs seem to reveal relationships between fertility rates, income and geographic region. We see that the rates are either decreasing or stagnant. Core/developed regions (e.g. North America, Europe and Central Asia) have the lowest fertility rates, and lowest rates of change. On the other hand, periphery/developing regions (e.g. Subsaharan Africa, South Asia) have seen a drop in these rates. Although the gaps have begun to close, there are still large gaps between Subsaharan Africa and the rest of the world, and between low income countries and the rest of the world.

Infant mortality rates, income and region

```
plota <- group_by(clean, year, region) %>%
  summarise(avgMort = mean(mortality, na.rm=TRUE)) %>%
  ggplot(aes(x=as.numeric(year), y=avgMort, colour = region)) +
  geom_point() + geom_smooth(se=FALSE) +
  labs(title="Average Global Infant Mortality Rate",
                   x="Year", y="Infant mortality rate") +
  theme_bw()

plotb <- group_by(clean, year, income) %>%
  summarise(avgMort = mean(mortality, na.rm=TRUE)) %>%
  ggplot(aes(x=as.numeric(year), y=avgMort, colour = income)) +
  geom_point() + geom_smooth(se=FALSE) +
  labs(title="Average Global Infant Mortality Rate",
                   x="Year", y = "Infant mortality rate") +
  theme_bw()
```

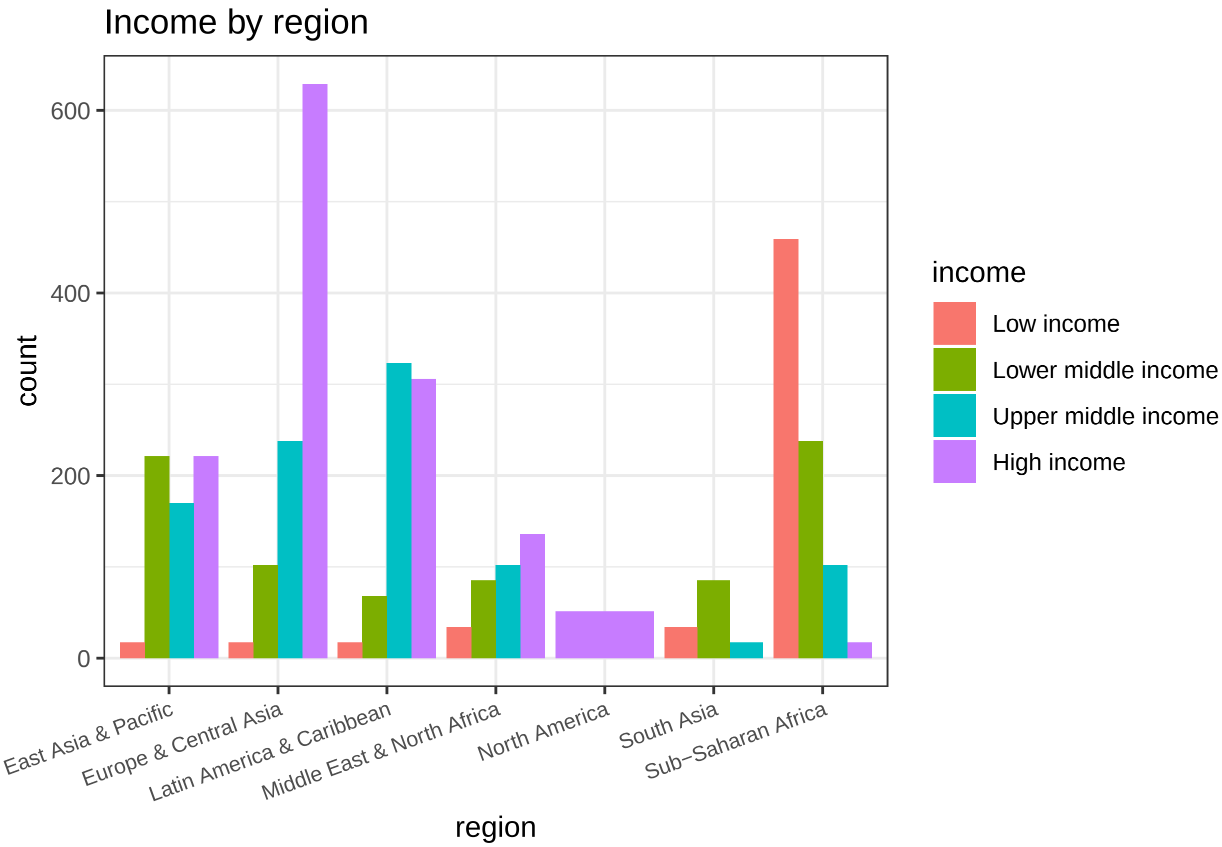
```
gridExtra::grid.arrange(plota, plotb)
```



We see simliar trends with infant mortality, however the gaps between developed and developing regions and between rich and poor countries seem to be closing faster. The downward trends for Subsaharan Africa and South Asia are steeper than the fertility rates.

Income by region

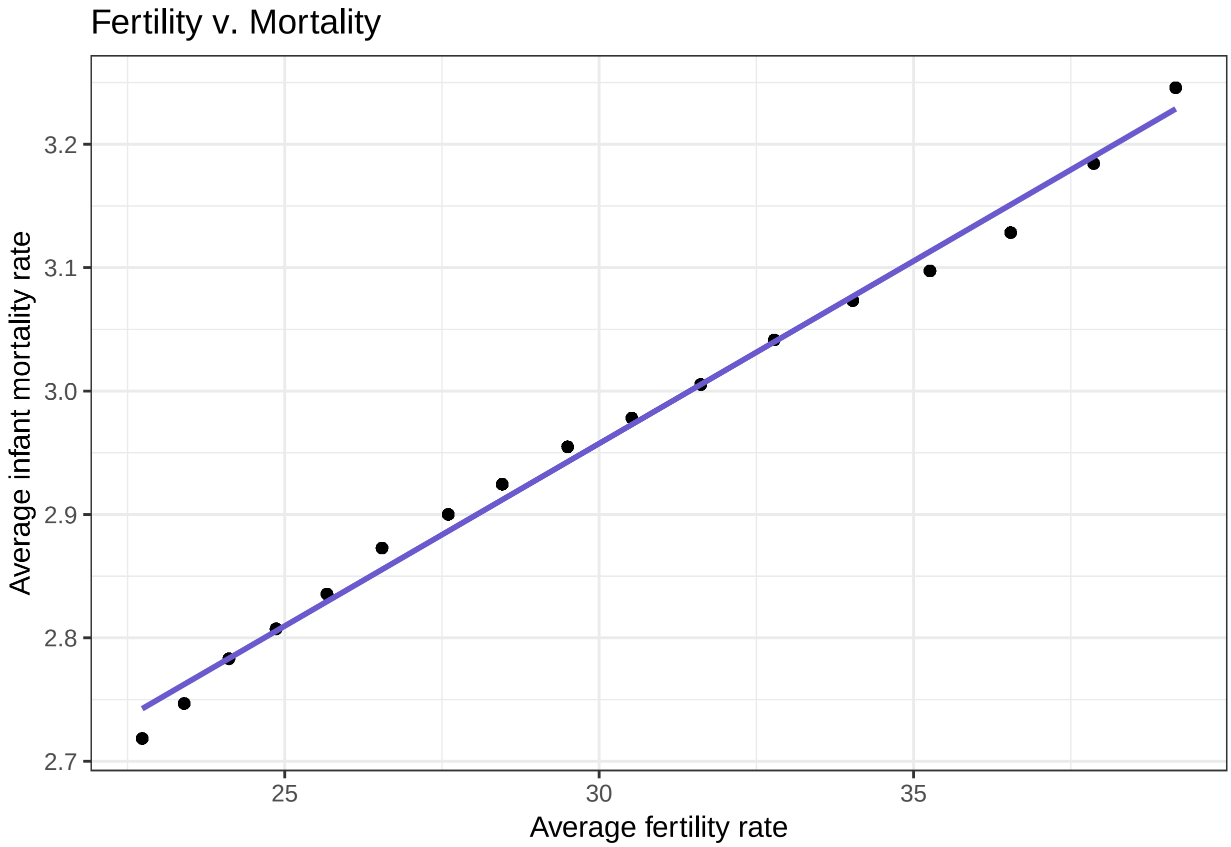
```
ggplot(clean, aes(x=region, fill=income)) + geom_bar(position="dodge") + theme_bw() +
  theme(axis.text.x=element_text(angle=20, hjust=1, size=8)) + ggtitle("Income by region")
```



This graph helps show the relationship between income and region of the globe. Regions like Subsaharan Africa and South Asia contain more low income and lower middle income countries, while regions like Europe and Central Asia and North America are made up of higher income countries.

Fertility and mortality

```
group_by(clean,year) %>%
  mutate(avgMort = mean(fertility, na.rm=TRUE)) %>%
  mutate(avgFert = mean(mortality, na.rm=TRUE)) %>%
  ggplot(aes(x=avgFert, y=avgMort)) +
  geom_point() + geom_smooth(color="slateblue", method = "lm") +
  labs(title="Fertility v. Mortality", x="Average fertility rate",
       y="Average infant mortality rate")+
  theme_bw()
```



```
round(cor(clean$fertility, clean$mortality, use="complete.obs"),3)
```

```
## [1] 0.861
```

There seems to be a strong, positive, linear association between fertility rates and infant mortality rates, with a correlation coefficient of 0.861. One explanation for this could be that people in developing countries have more children because their children are less likely to survive.