Report on New Insights on Poverty

Yanhan Si

12/2/2019

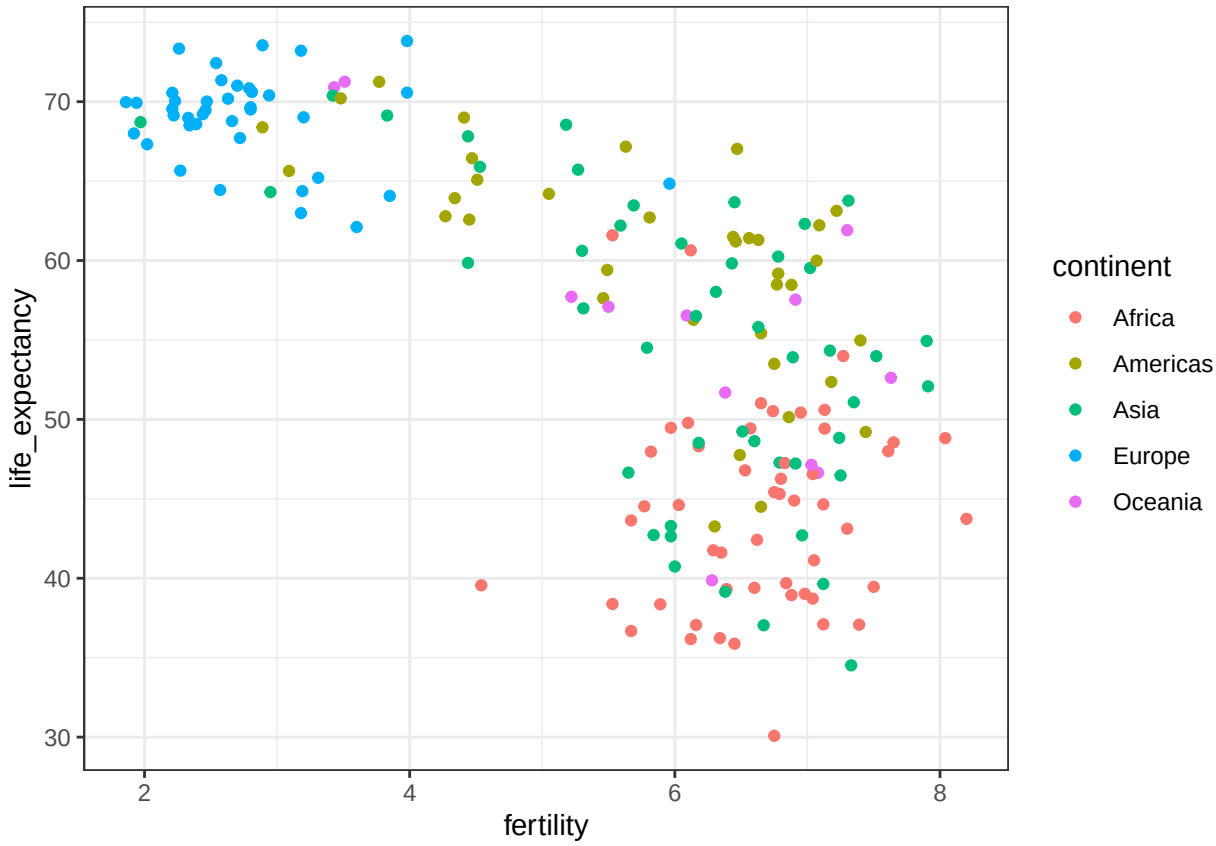
Introduction

This report shows trends in world health and economics using data visualization, exploratory data analysis and summarization. The original data are from Gapminder https://www.gapminder.org. The results will show how actual trends in health and economics contradict the narratives that emanate from sensationalist media coverage of catastrophes, tragedies, and other unfortunate events.

Part 1: the West versus developing world view

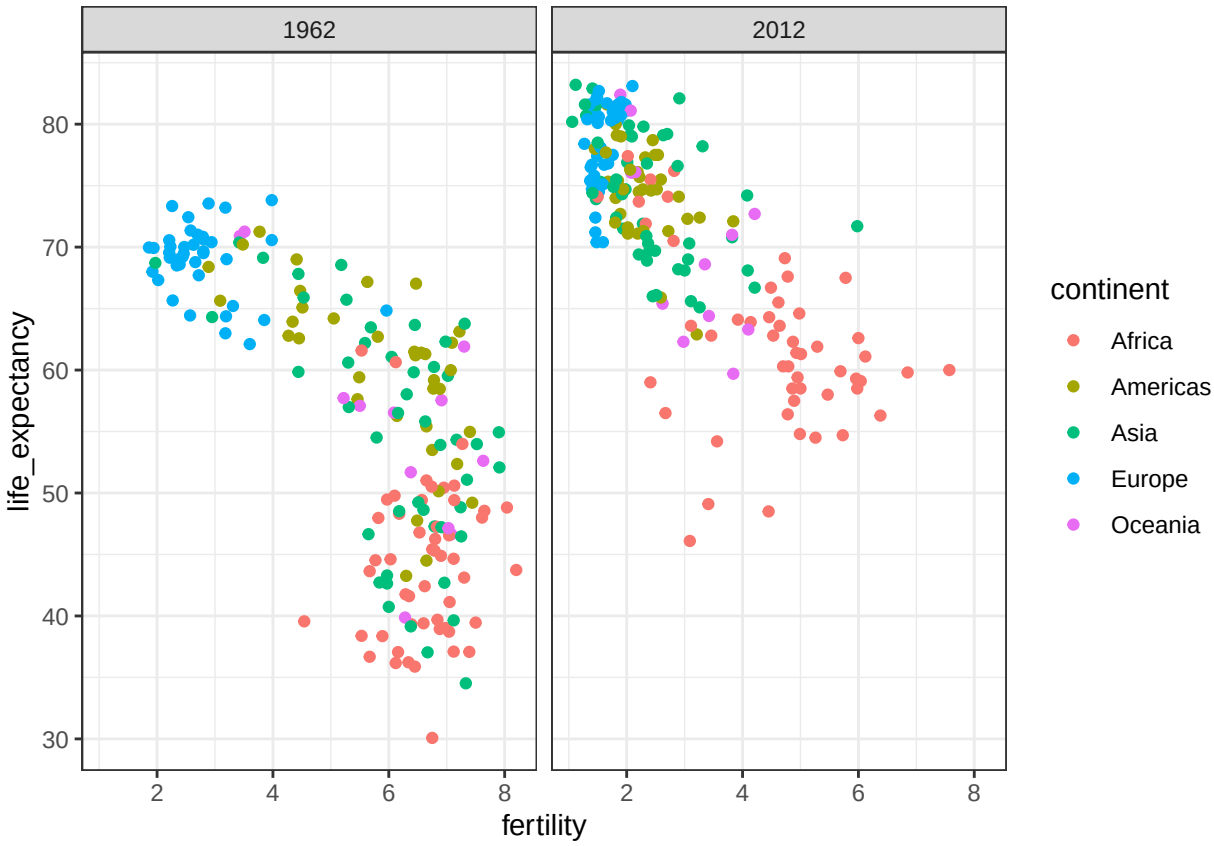
This scatter plot shows life expectancy versus fertility among countries in 1962. From this point, most points falls into two distinct categories. 1. Life expectancy around 70 years and 3 or fewer child per family 2. Life expectancy lower than 65 years and more than 5 child per family We use color to distinguish the different continent to see if this explains the clusters. And in 1962, "the West versus developing world" view was grounded in some reality. However, this is not case 50 years later.

```
ds_theme_set()
filter(gapminder, year == 1962) %>%
  ggplot(aes(fertility, life_expectancy, color = continent)) +
  geom_point()
```



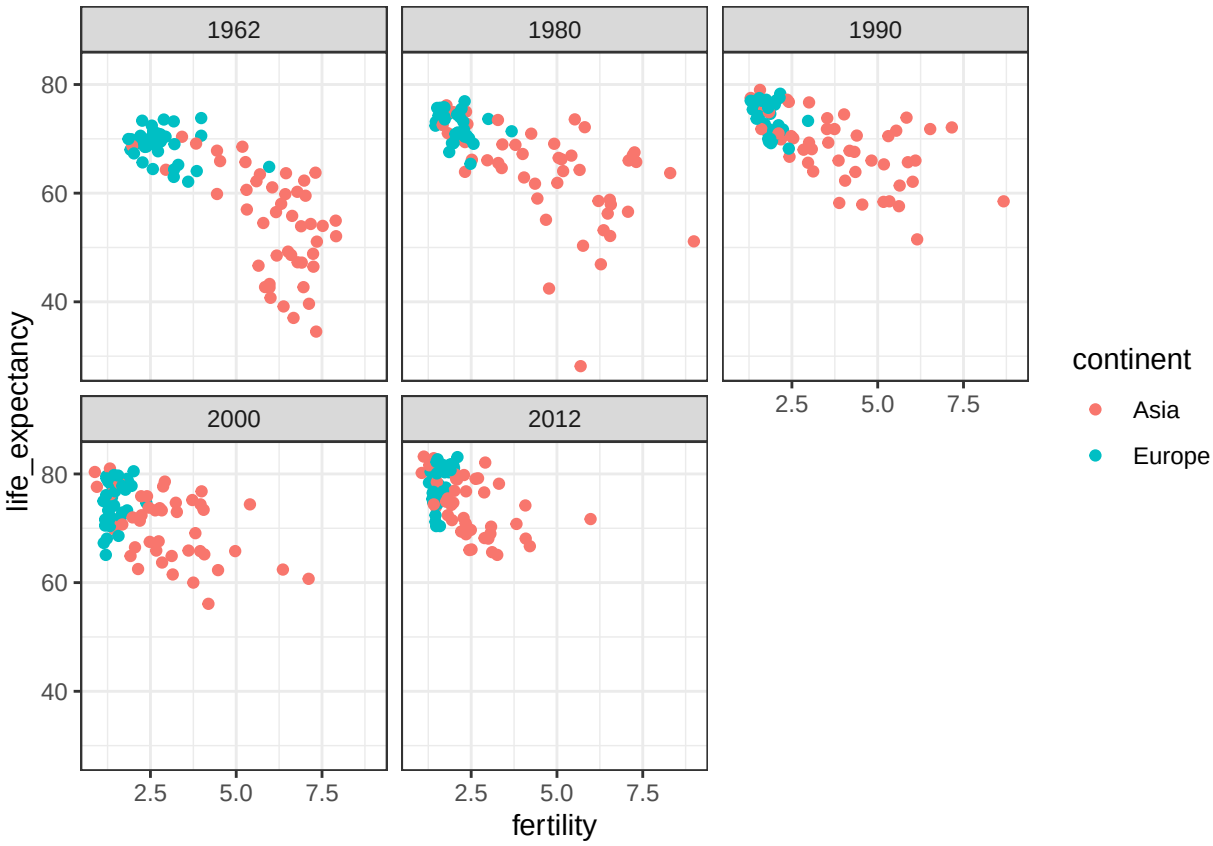
To make a comparison of how the life expectancy versus fertility changes over the past decades, we make side by side plots of 2012 data and 1962 data. This plot clearly shows that the majority of countries have moved from the `developing world` cluster to the Western world one. They went from having large families and short lifespans to having smaller families and longer lifespans. In 2012, the Western versus developing world view no longer makes sense.

```
filter(gapminder, year %in% c(1962, 2012)) %>%
  ggplot(aes(fertility, life_expectancy, col = continent)) +
  geom_point() +
  facet_grid(. ~ year)
```



To explore how this transformation happened through the years, we show the plots for sereral years. This plot clearly shows how most Asian countries have improved at a much faster rate than European ones.

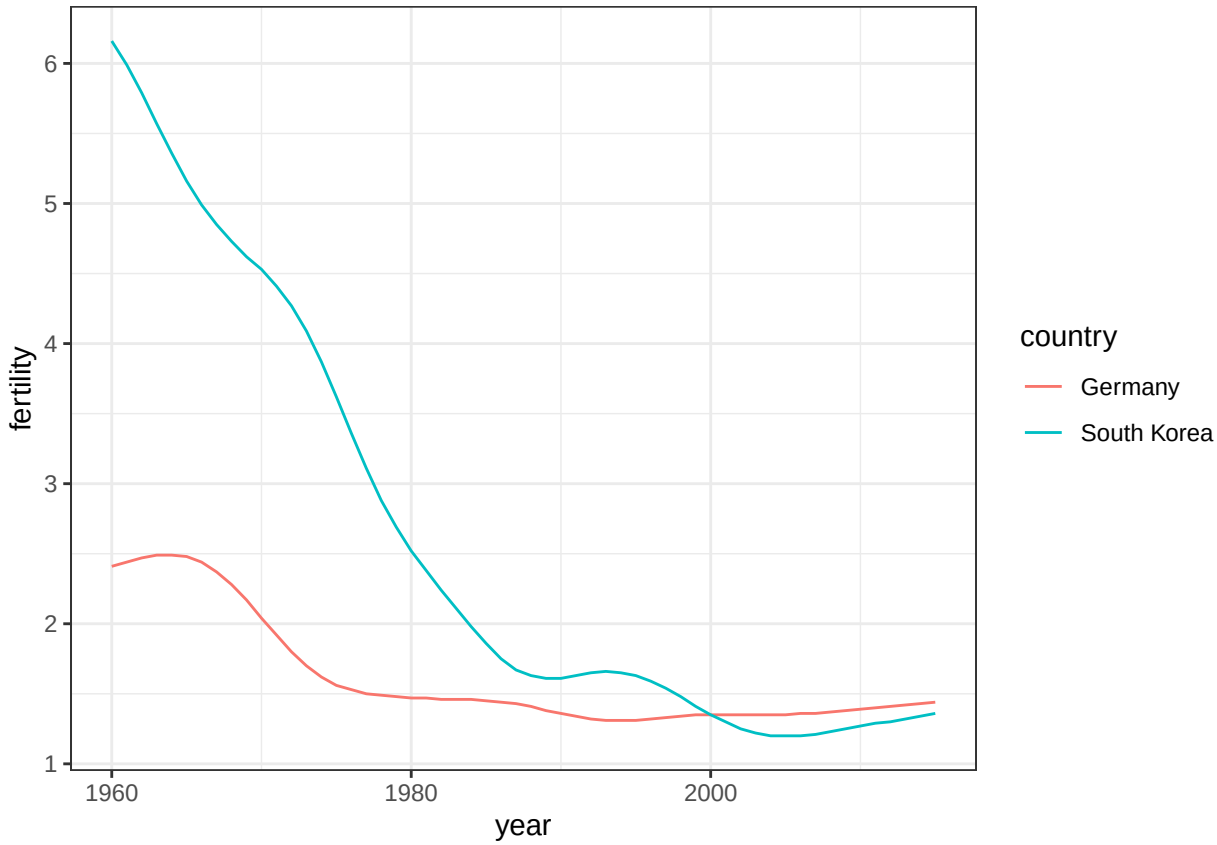
```
years <- c(1962, 1980, 1990, 2000, 2012)
continents <- c("Europe", "Asia")
gapminder %>%
  filter(year %in% years & continent %in% continents) %>%
  ggplot(aes(fertility, life_expectancy, col = continent)) +
  geom_point() +
  facet_wrap(~year)
```



Part 2

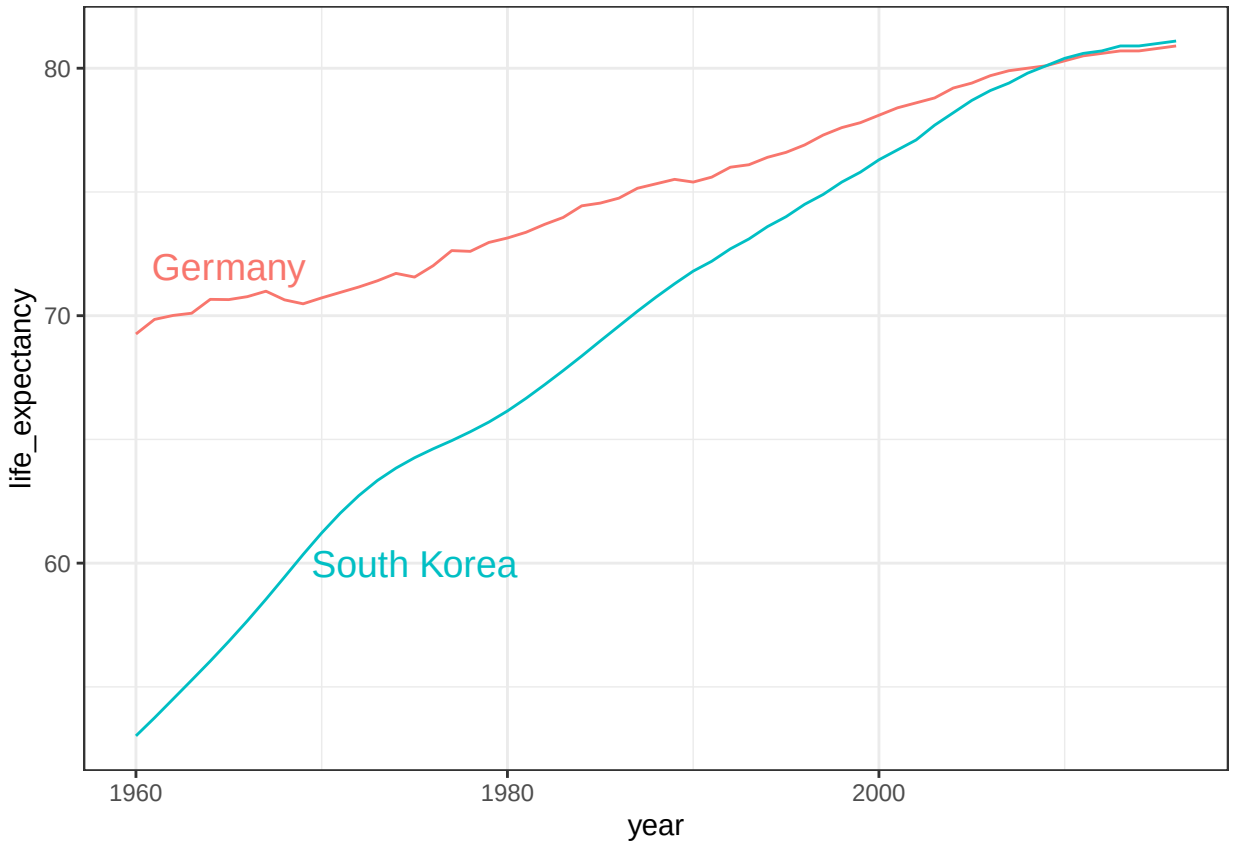
The visualizations of Part 1 illustrate that data no longer supports the Western versus developing worldview. We will see how this change happened over time. This plot clearly shows how South Korea's fertility rate dropped drastically during the 1960s and 1970s, and by 1990 has a similar rate to that of Germany.

```
countries <- c("South Korea","Germany")
gapminder %>% filter(country %in% countries & !is.na(fertility)) %>% ggplot(aes(year,fertility, col = c
geom_line()
```



This is the life expectancy plot. This plot clearly shows how an improvement in life expectancy followed the drops in fertility rates. In 1960, Germans lived 15 years longer than South Koreans, although by 2010 the gap is completely closed. It exemplifies the improvement that many non-western countries have achieved in the last 40 years.

```
countries <- c("South Korea", "Germany")
labels <- data.frame(country = countries, x = c(1975, 1965), y = c(60, 72))
gapminder %>% filter(country %in% countries) %>%
  ggplot(aes(year, life_expectancy, col = country)) +
  geom_line() +
  geom_text(data = labels, aes(x, y, label = country), size = 5) +
  theme(legend.position = "none")
```

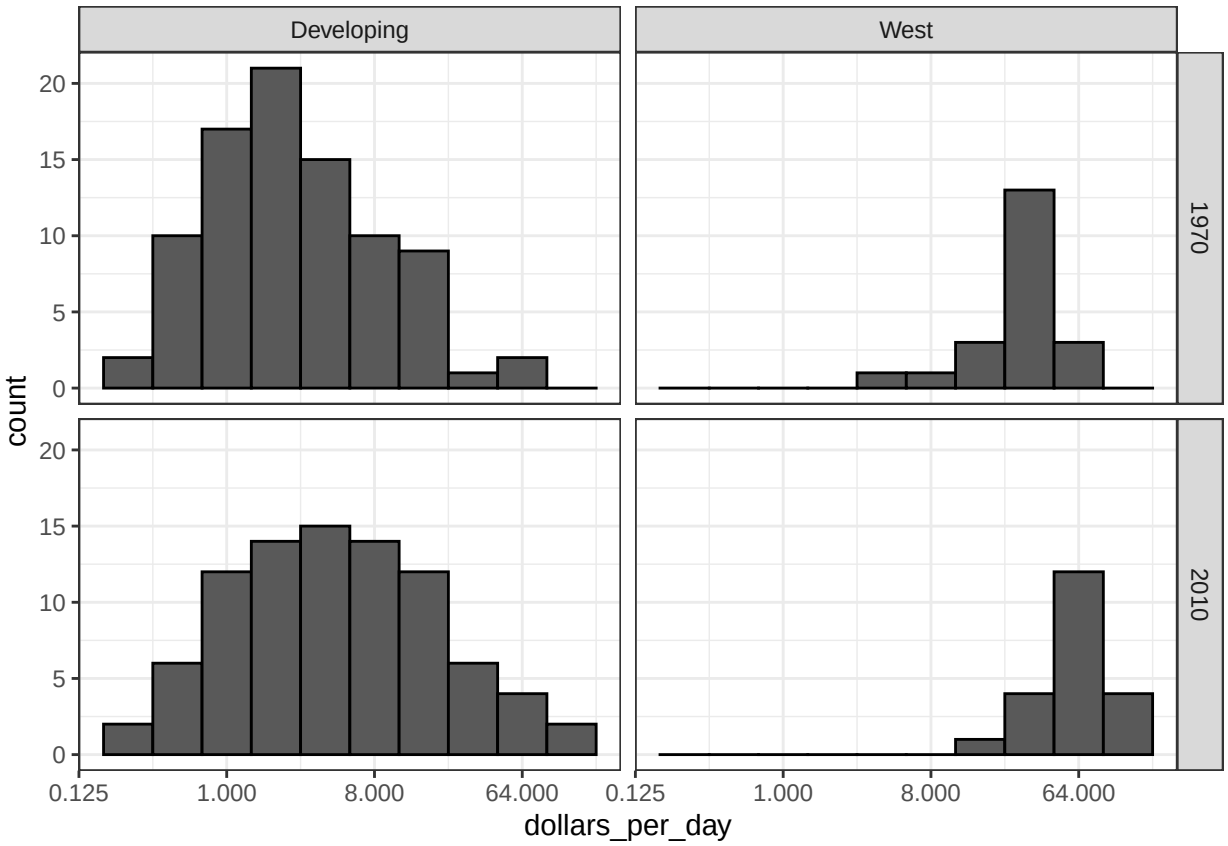


Part3

In this part, we will explore the income distribution on 1970 versus 2010.

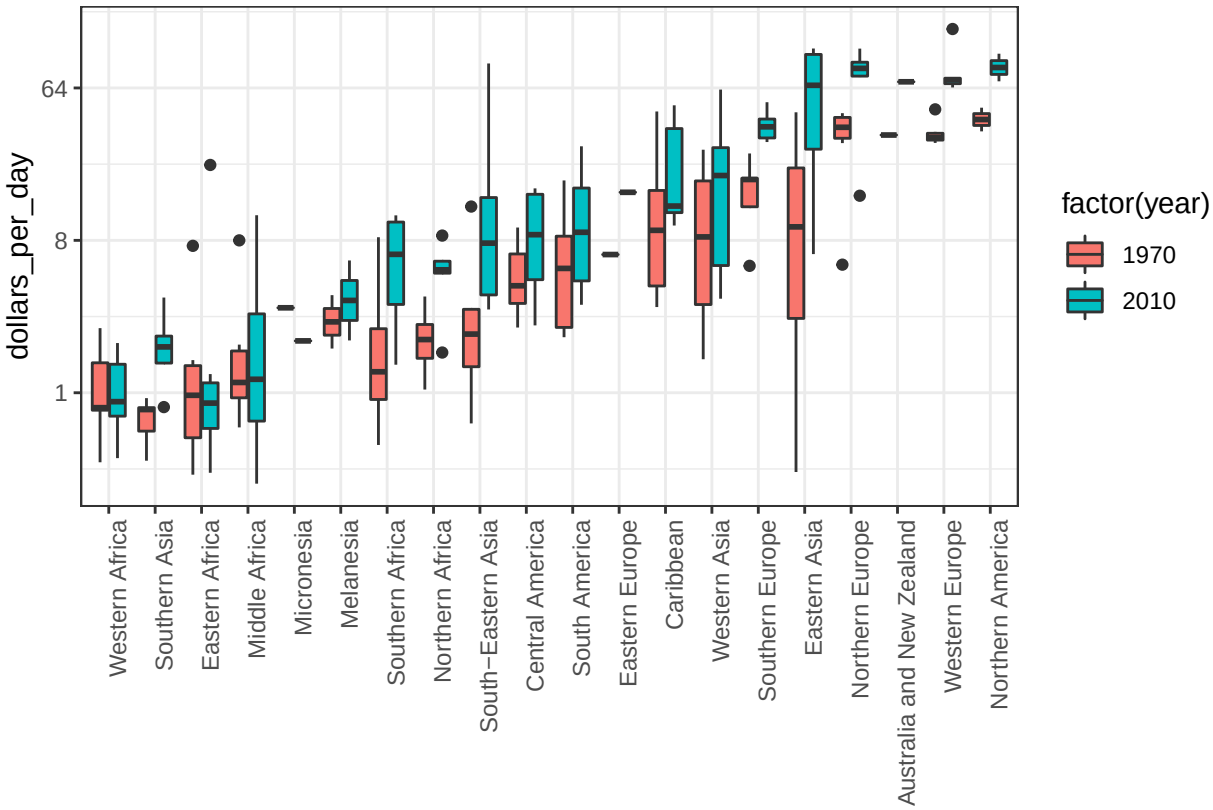
```
# add dollars per day variable and define past year
gapminder <- gapminder %>%
  mutate(dollars_per_day = gdp/population/365)
past_year <- 1970
# define Western countries
west <- c("Western Europe", "Northern Europe", "Southern Europe", "Northern America", "Australia and Ne
present_year <- 2010
# define countries that have data available in both years
country_list_1 <- gapminder %>%
  filter(year == past_year & !is.na(dollars_per_day)) %>% .$country
country_list_2 <- gapminder %>%
  filter(year == present_year & !is.na(dollars_per_day)) %>% .$country
country_list <- intersect(country_list_1, country_list_2)

# make histogram including only countries with data available in both years
gapminder %>%
  filter(year %in% c(past_year, present_year) & country %in% country_list) %>%    # keep only selected
  mutate(group = ifelse(region %in% west, "West", "Developing")) %>%
  ggplot(aes(dollars_per_day)) +
  geom_histogram(binwidth = 1, color = "black") +
  scale_x_continuous(trans = "log2") +
  facet_grid(year ~ group)
```



From this histogram for the four different groups, we can see that the rich countries have become a bit richer percentage wise, the poorer countries appear to have improved more.The histogram has shifted more to the right than for the rich countries. In particular, we see that the proportion of developing countries earning more than $16 a day increases substantially. The income distribution of the developing countries has gotten closer to those from the west. In order to see which specific regions improve the most, we made boxplots of 2010 data and 1970 data.

```
# arrange matching boxplots next to each other, colored by year
gapminder %>%
  filter(year %in% c(past_year, present_year) & country %in% country_list) %>%
  mutate(region = reorder(region, dollars_per_day, FUN = median)) %>%
  ggplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  xlab("") + scale_y_continuous(trans = "log2") +
  geom_boxplot(aes(region, dollars_per_day, fill = factor(year)))
```



From this boxplot, we can see that Eastern Asia went from way down around 8 all the way up almost to 64.

Part 4

We used a series of histograms and box plots to discover the income gap between rich and poor countries has closed considerably during the last forty years. Here we will use the smooth density plots to examine the income distribution again.

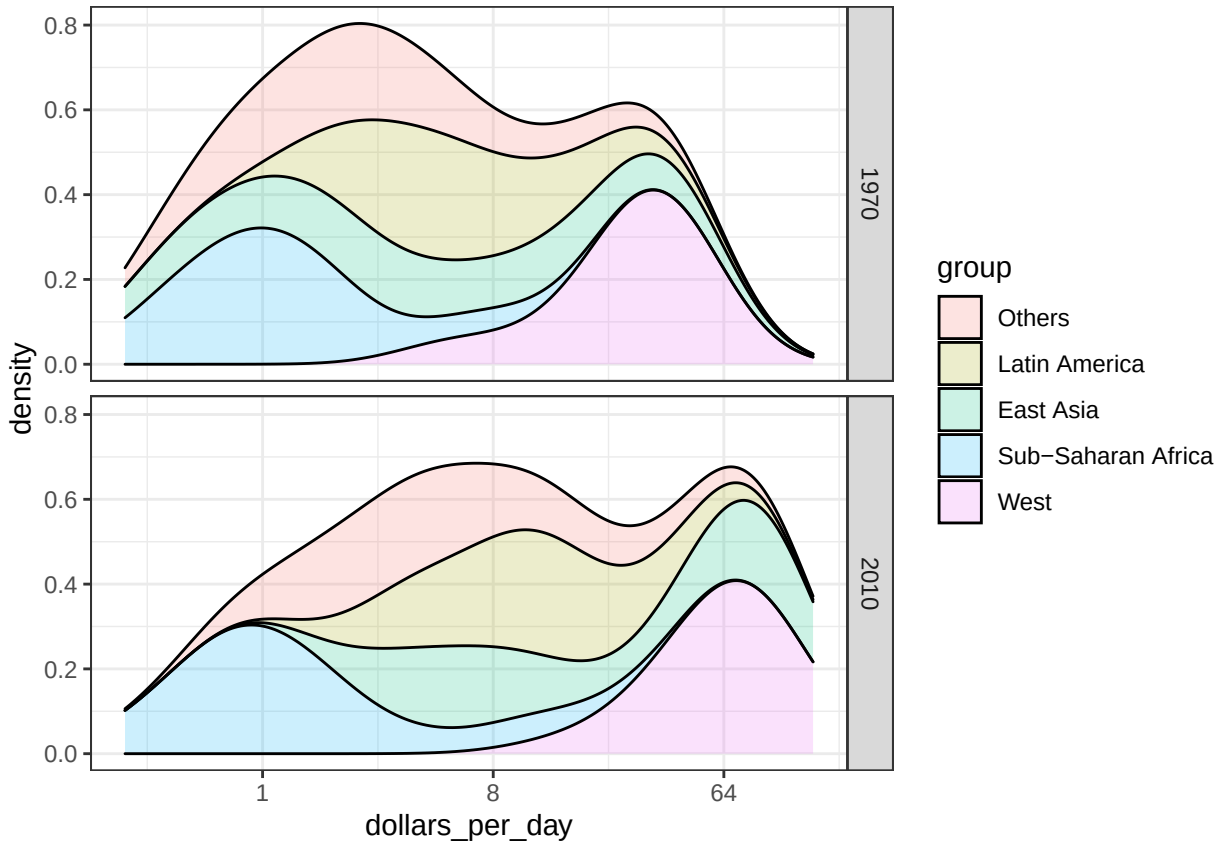
```
##Code: Add new region groups with case_when

# add group as a factor, grouping regions
gapminder <- gapminder %>%
  mutate(group = case_when(
    .$region %in% west ~ "West",
    .$region %in% c("Eastern Asia", "South-Eastern Asia") ~ "East Asia",
    .$region %in% c("Caribbean", "Central America", "South America") ~ "Latin America",
    .$continent == "Africa" & .$region != "Northern Africa" ~ "Sub-Saharan Africa",
    TRUE ~ "Others"))
# reorder factor levels
gapminder <- gapminder %>%
  mutate(group = factor(group, levels = c("Others", "Latin America", "East Asia", "Sub-Saharan Africa",


##Code: Stacked density plot

# note you must redefine p with the new gapminder object first
p <- gapminder %>%
  filter(year %in% c(past_year, present_year) & country %in% country_list) %>%
  ggplot(aes(dollars_per_day, fill = group)) +
```

```
  scale_x_continuous(trans = "log2")
# stacked density plot
p + geom_density(alpha = 0.2, bw = 0.75, position = "stack") +
  facet_grid(year ~ .)
```



Here we can clearly see how the distributions for East Asia, Latin America, and others shift markedly to the right. While Sub-Saharan Africa remains stagnant.

Part 5

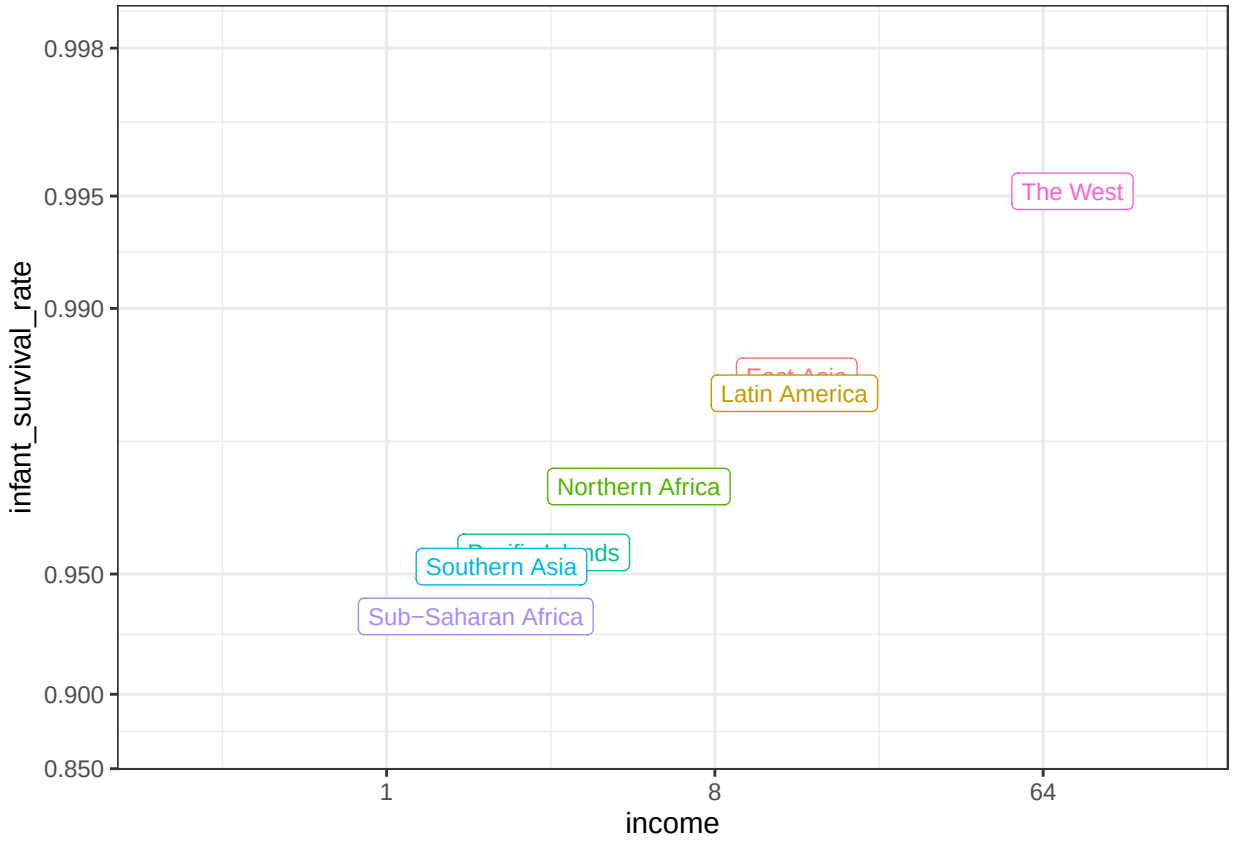
In this report, we have been comparing regions of the world. We have seen that, on average, some regions do better than others. In this part, we focus on describing the importance of variability within the groups when examining the relationship between a country's infant mortality rates and average income.

```
# add additional cases
gapminder <- gapminder %>%
  mutate(group = case_when(
    .$region %in% west ~ "The West",
    .$region %in% "Northern Africa" ~ "Northern Africa",
    .$region %in% c("Eastern Asia", "South-Eastern Asia") ~ "East Asia",
    .$region == "Southern Asia" ~ "Southern Asia",
    .$region %in% c("Central America", "South America", "Caribbean") ~ "Latin America",
    .$continent == "Africa" & .$region != "Northern Africa" ~ "Sub-Saharan Africa",
    .$region %in% c("Melanesia", "Micronesia", "Polynesia") ~ "Pacific Islands"))
# define a data frame with group average income and average infant survival rate
surv_income <- gapminder %>%
  filter(year %in% present_year & !is.na(gdp) & !is.na(infant_mortality) & !is.na(group)) %>%
```

```
  group_by(group) %>%
  summarize(income = sum(gdp)/sum(population)/365,
            infant_survival_rate = 1 - sum(infant_mortality/1000*population)/sum(population))
surv_income %>% arrange(income)
```

```
## # A tibble: 7 x 3
##   group              income infant_survival_rate
##   <chr>               <dbl>                <dbl>
## 1 Sub-Saharan Africa   1.76                0.936
## 2 Southern Asia        2.07                0.952
## 3 Pacific Islands      2.70                0.956
## 4 Northern Africa      4.94                0.970
## 5 Latin America       13.2                 0.983
## 6 East Asia           13.4                 0.985
## 7 The West            77.1                 0.995
```

```
# plot infant survival versus income, with transformed axes
surv_income %>% ggplot(aes(income, infant_survival_rate, label = group, color = group)) +
  scale_x_continuous(trans = "log2", limit = c(0.25, 150)) +
  scale_y_continuous(trans = "logit", limit = c(0.875, .9981),
                     breaks = c(.85, .90, .95, .99, .995, .998)) +
  geom_label(size = 3, show.legend = FALSE)
```



The relationship between these two variables is almost perfectly linear and the graph shows a dramatic difference. While in the West less than 0.5% of infants die, in Sub-Saharan Africa the rate is higher than 6%!