Gapminder Data Set Visualization

In this case study, I am going to visualize the gapminder dataset and answer a few questions. Gapminder is a dataset by the Gapminder Foundation about the world health and economics. It contains various details about each country from 1960 to 2016 like infant mortality, fertility, life expectancy, GDP and many more. It was used by Hans Rosling which was the cofounder of Gapminder foundation to educate the public so that they dispel common myths about the so called developing world. It has nine variables which are country, year, infant\_mortality, life\_expectancy, fertility, population, GDP, continent and region.

In this small case study, I am going to answer the following two questions by using the gapminder dataset. I will be using the ggplot2 library in R for our data visualization.

1. Is it fair to say that the world is divided into the western rich nations and the developing world in Africa, Asia and Latin America?
2. Has income inequality across countries worsened during the last forty years?

I am going to use the tidyverse and dslabs library which contain the gapminder data set. You can access the data using the following R code.

Library (tidyverse)

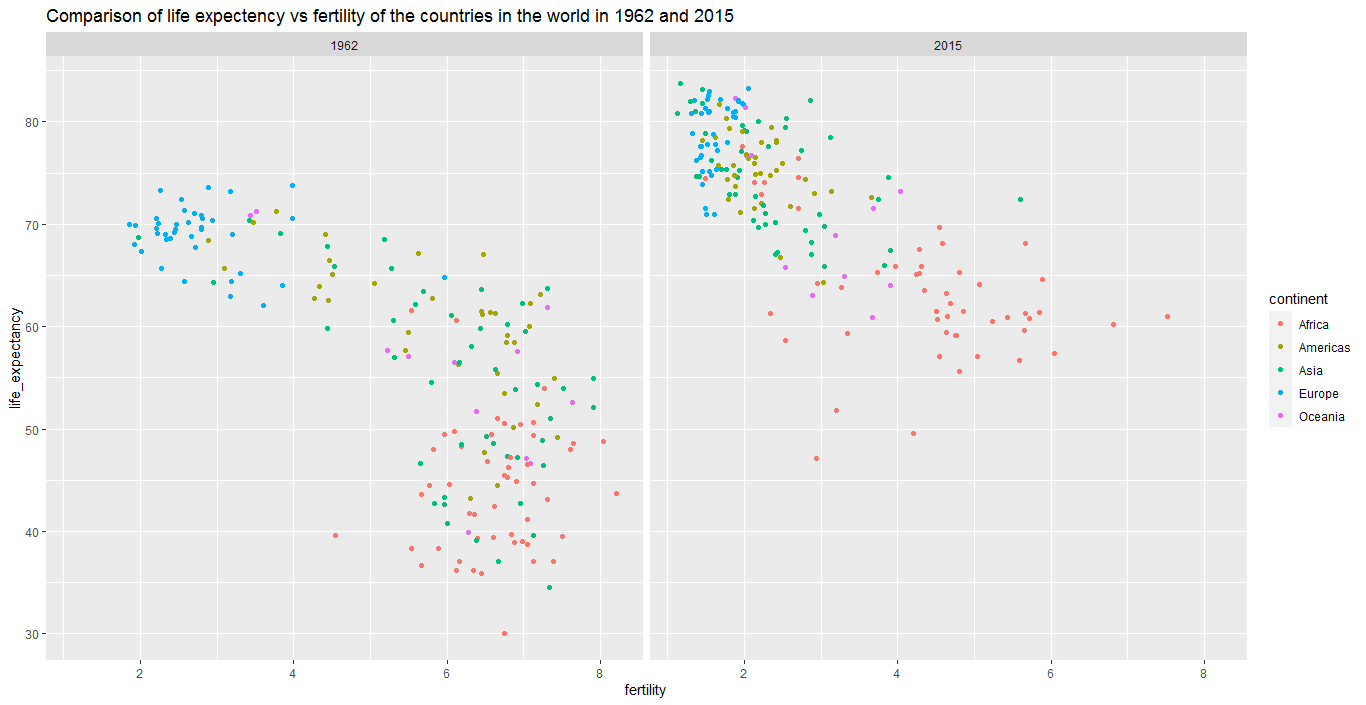
Library (dslabs)

Data (gapminder)

Gapminder %> % as\_tibble ()

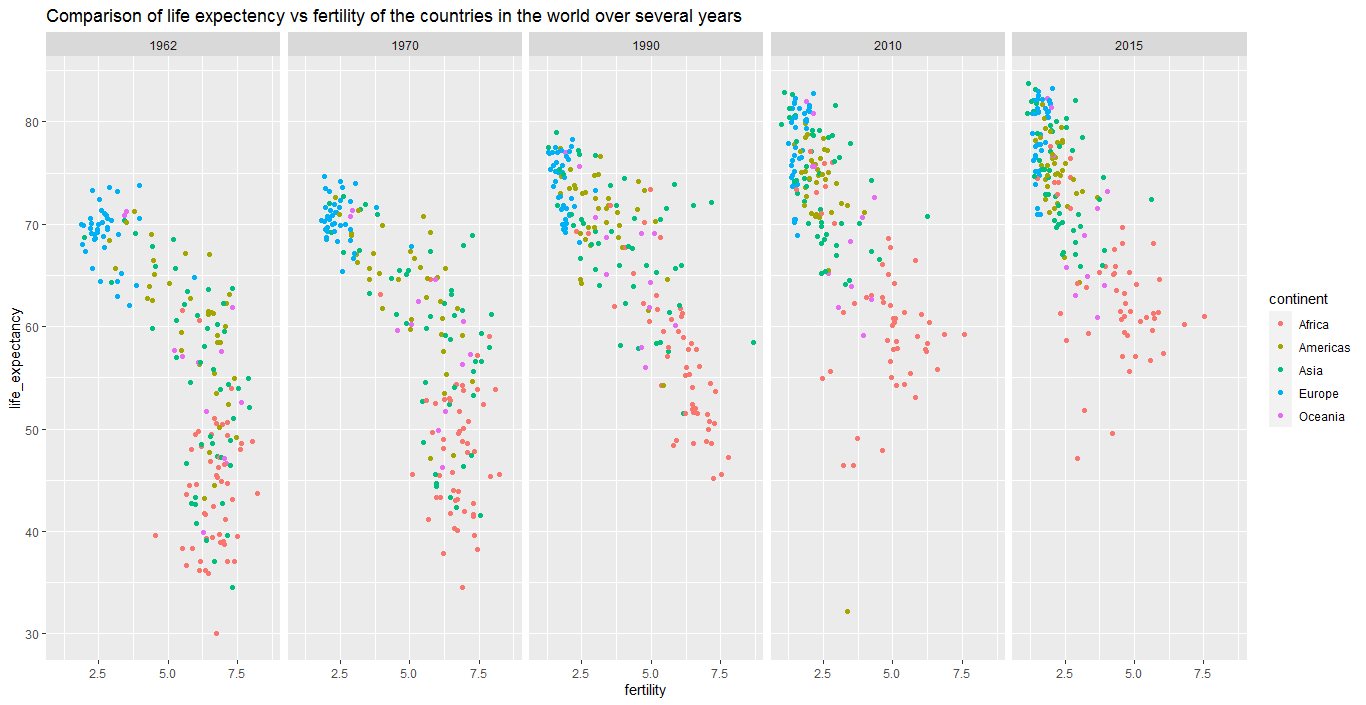
To answer the first question, I am going to plot fertility versus life expectancy of each country in 1962 as well as 2015. Following is the code and plot.

gapminder %>% filter(year %in% c(1962, 2015)) %>% ggplot(aes(fertility, life\_expectancy, col = continent)) + geom\_point() + facet\_grid(.~ year) + ggtitle("Comparison of life expectency vs fertility of the countries in the world in 1962 and 2015")



In the plot we can clearly see that most of the countries have moved from developing world to western world. In 2015, the western vs the developing world view no longer make sense. Now I am going to compare the fertility vs life expectancy of each country over several years and see how this transformation happened.

gapminder %>% filter(year %in% c(1962, 1970, 1990, 2010, 2015)) %>% ggplot(aes(fertility, life\_expectancy, col = continent)) + geom\_point() + facet\_grid(.~ year) + ggtitle("Comparison of life expectency vs fertility of the countries in the world over several years")



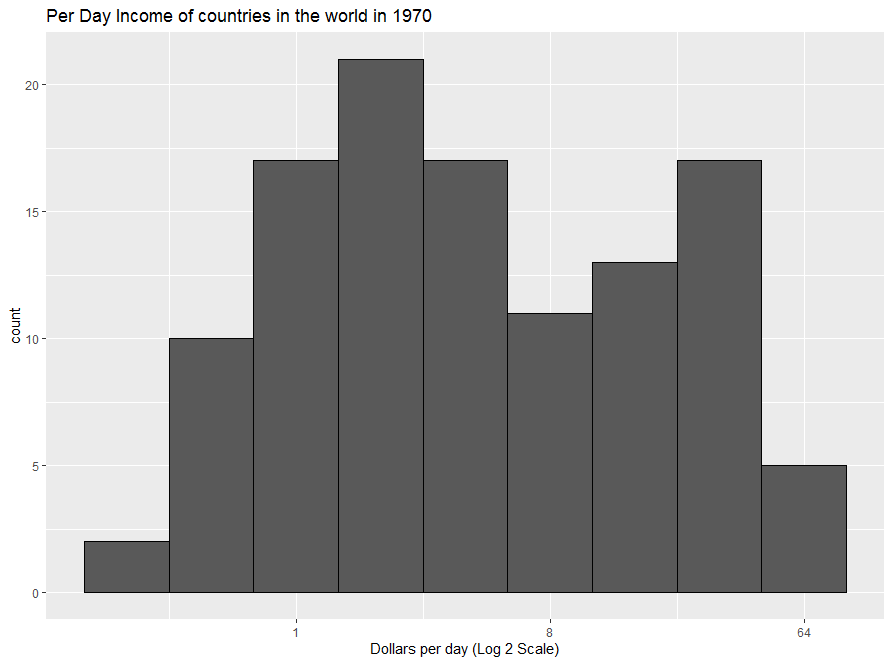
Here we can see that most of the Asian countries have improved more in the last 55 years.

To answer the second question, we are going to compare the income distribution across countries and see whether it has improved or worsened in the last 40 years. Before going to plot a graph, I am going to add another variable to our dataset which is dollars per day. We can obtain this variable by dividing the GDP of each country by its population and then by number of days in a year. The GDP per person is often used as a rough summary of a country’s wealth. So, we are using the GDP per person of each country in our analysis. The code for adding the variable is given below.

gapminder <- gapminder %> % mutate (dollars\_per\_day = gdp/population/365)

Now, let’s draw a histogram and see the income distribution in the year 1970. Below, is the code and the plot.

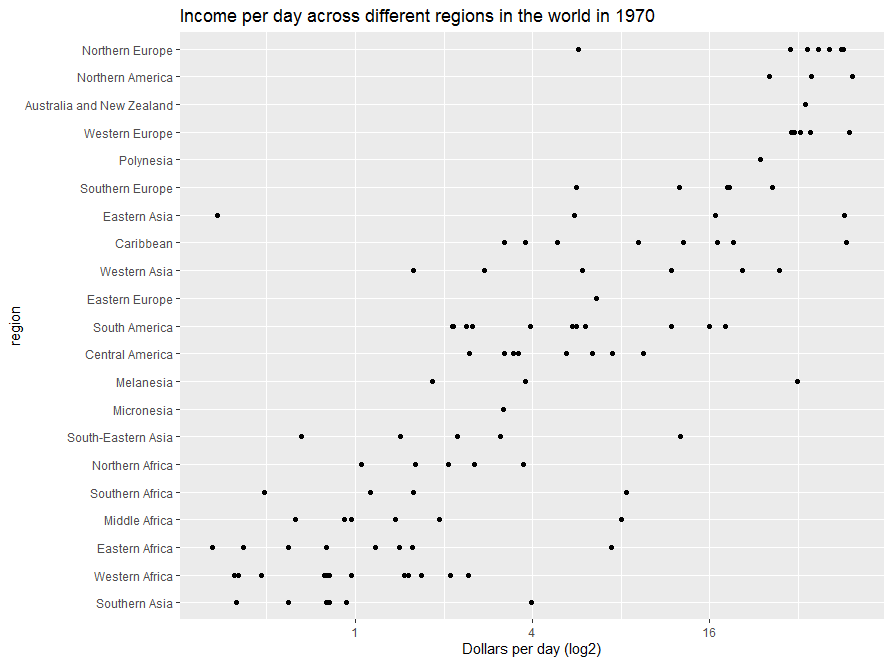
gapminder %>% filter(year == 1970 & !is.na(gdp)) %>% ggplot(aes(dollars\_per\_day)) + geom\_histogram(binwidth = 1, color = "black") + scale\_x\_continuous(trans = "log2") + ggtitle("Per Day Income of countries in the world in 1970") + xlab("Dollars per day (Log 2 Scale)")



The histogram above suggests that the income distribution across countries in 1970 has two modes: one at about 2 dollars per day and another at about 32 dollars per day. This bimodality is consistent with a dichotomous world made up of countries with average income less than $8 a day and countries above that. So, the above histogram clearly shows that there was a “west versus the rest” dichotomy in 1970.

Now let’s see how the distribution of income has changed after forty years. Here, we will focus on just two groups: the west and the rest. To see which countries come in the west and which in the rest, we will use a scatter plot which is given below.

gapminder %>% filter(year == 1970 & !is.na(gdp)) %>% ggplot(aes(dollars\_per\_day)) + geom\_histogram(binwidth = 1, color = "black") + scale\_x\_continuous(trans = "log2") + ggtitle("Per Day Income of countries in the world in 1970") + xlab("Dollars per day (Log 2 Scale)")

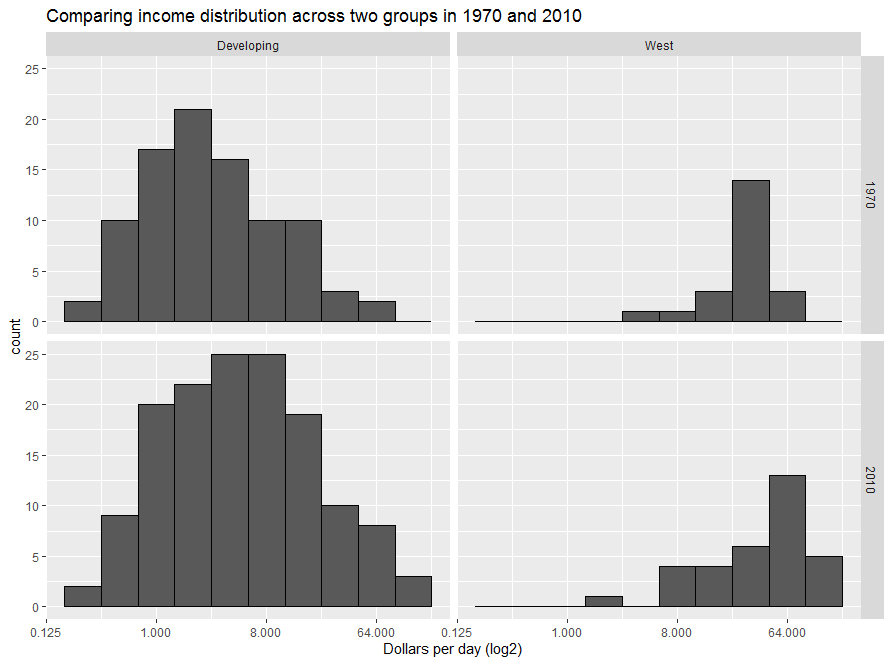


Now, on the basis of above graph we can clearly divide the countries into west versus the rest or the developing countries. Let’s use the following code to do this.

gapminder <- gapminder %>% mutate(group = case\_when( region %in% c("Western Europe", "Northern Europe", "Northern America", "Australia and New Zealand", "Southern Europe") ~ "West", TRUE ~ "Developing"))

Now let’s make four histograms to compare income distribution between the west and the developing regions in 1970 and 2010.

gapminder %>% filter(year %in% c(1970, 2010) & !is.na(gdp)) %>% ggplot(aes(dollars\_per\_day)) + geom\_histogram(binwidth = 1, color = "black") + scale\_x\_continuous(trans = "log2") + facet\_grid(year~group) + ggtitle("Comparing income distribution across two groups in 1970 and 2010") + xlab("Dollars per day (log2)")



Before we interpret the findings of this plot, we notice that there are more countries represented in 2010 Histogram than in 1970. One reason for this is several countries were founded after 1970. So we are going to remake the plots using only countries with data available for both years. The code for finding the list of countries for which data is available for both years is given below.

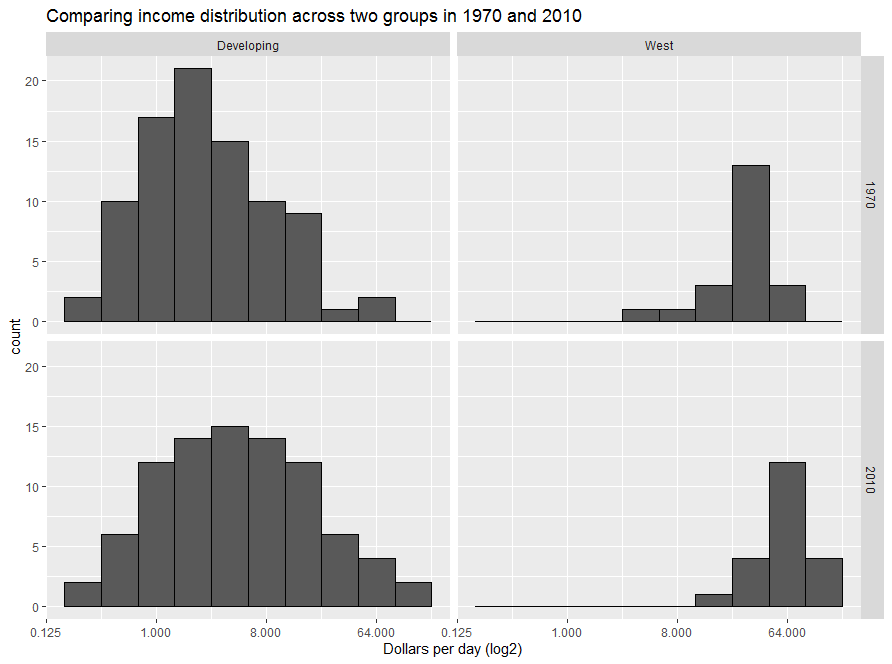
country\_list1 <- gapminder % > % filter (year == 1970 & !is.na(dollars\_per\_day)) % > % pull(country)

country\_list2 <- gapminder % > % filter (year == 2010 & !is.na(dollars\_per\_day)) % > % pull (country)

country\_list <- intersect (country\_list1, country\_list2)

Now let’s remake the plots by using the below code.

gapminder %>% filter(year %in% c(1970, 2010) & !is.na(gdp) & country %in% country\_list) %>% ggplot(aes(dollars\_per\_day)) + geom\_histogram(binwidth = 1, color = "black") + scale\_x\_continuous(trans = "log2") + facet\_grid(year~group) + ggtitle("Comparing income distribution across two groups in 1970 and 2010") + xlab("Dollars per day (log2)")



We now can see that the rich countries have become a little richer, but percentage-wise, the poor countries appear to have improved more. In particular, we see that the proportion of developing countries earning more than $16 a day increased substantially. So, the answer of the second question of our analysis is that the income distribution has actually improved in the last 40 years.

Now to see which specific region has improved the most in the last 40 years, we are going to make a boxplot for the analysis. Since, we have a lot of regions in our dataset, we are going to group the similar regions according to the scatterplot that we have shown previously. Below is the code for this specific grouping as well as the boxplot.

gapminder <- gapminder % > %

+ mutate (group1 = case\_when (

+ region %in% c ("Western Europe", "Northern Europe”, “Southern Europe",

+ "Northern America",

+ "Australia and New Zealand") ~ "West",

+ region %in% c ("Eastern Asia", "South-Eastern Asia") ~ "East Asia",

+ region %in% c ("Caribbean", "Central America",

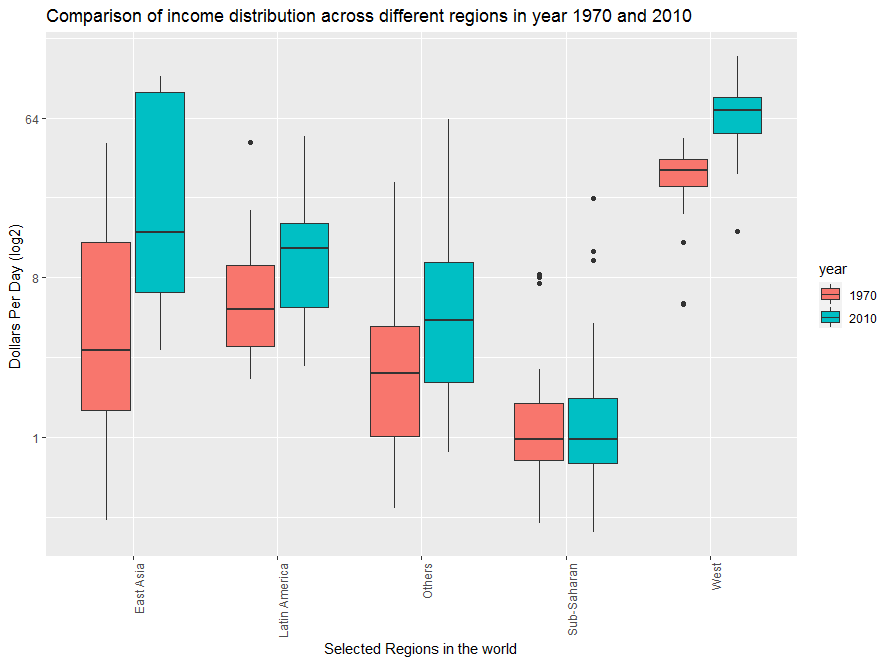
+ "South America") ~ "Latin America",

+ continent == "Africa" &

+ region != "Northern Africa" ~ "Sub-Saharan",

+ TRUE ~ "Others"))

gapminder %>% filter(year %in% c(1970, 2010) & country %in% country\_list) %>% mutate(year = factor(year)) %>% ggplot(aes(group1, dollars\_per\_day, fill = year)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + scale\_y\_continuous(trans = "log2") + ggtitle("Comparison of income distribution across different regions in year 1970 and 2010") + ylab("Dollars Per Day (log2) ")



Here, we can see that East Asia has improved the most in terms of its income per day in the last forty years.

Now, we come to the conclusion of our analysis. We have used ggplot2 library of the R programming language for making the plots. We have used the gapminder dataset and made the world view about the west and developing world wrong by making plots. We have also answered the two questions that we have raised in the start of our analysis.