Report on Global Health and Income Level: Gapminder Dataset (1799-2016)

Wrangling data sets with dplyr and exploring data

Manuel Roosevelt Lamptey, Rabeya Illyas Noon, and Tarana Ferdous

2022/07/08

Table of Contents

# Introduction

This report is aimed to provide an insight into the global population-level health data with the countries available from 1799 to 2016. In this report, we showed the process of curating the GapMinder World Health Data and recreated three interactive charts from a previous report (project 01), such as, bubble chart, growth chart, and rank chart by using the newly created data set, gapminder\_df, and provided details of each graph to explain what they depict.

# Curating the GapMinder World Health Data

To create this report, we have used six raw data sets from the website <https://www.gapminder.org/data/> to explore the global health and income level data. To execute the curating process, we followed steps including importing raw data, checking Patterns, joining data sets into one, calculating a new variable, performing ‘sanity checks’, and saving a new data file as a .csv file. The details of the procedure is provided below.

## Importing raw data

After loading the required libraries and importing the data sets, we reformed them from wide to long format.

## Checking Patterns

We reformed the data to a tidy format by checking if any data had a mix of numeric and character values. We checked the patterns, and if found anything, we replaced the character values with numeric values. For example, we found there are character values such as “k”, “B” “” in population data and “M”, “TR”, NA, and “B” in GDP data. Here, k, B, TR, and NA stand for thousand, billion, trillion, and missing value, respectively. We have replaced these characters as k = “e3”, M = “e6”, B = “e9”.

## Joining these six data sets into one

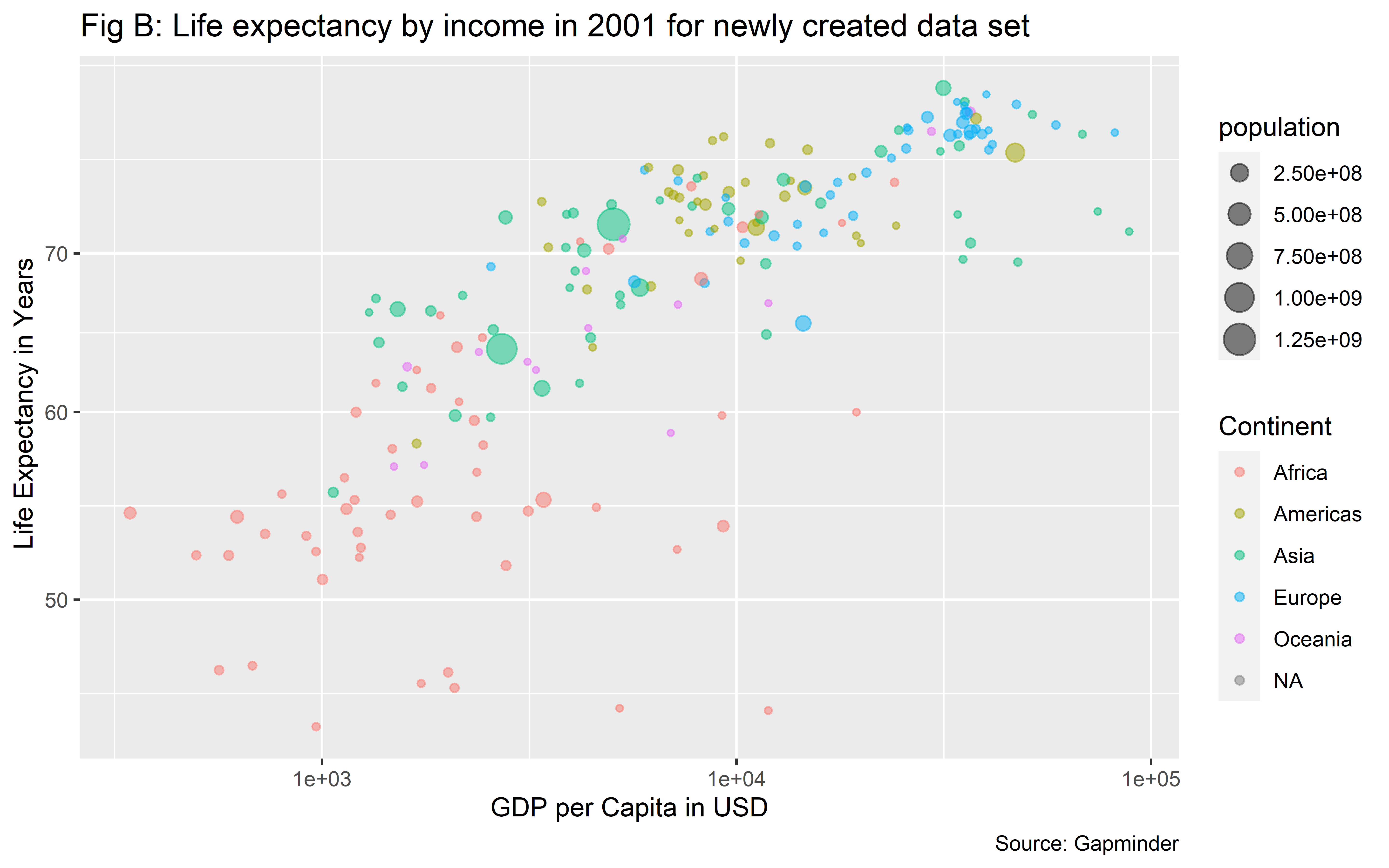
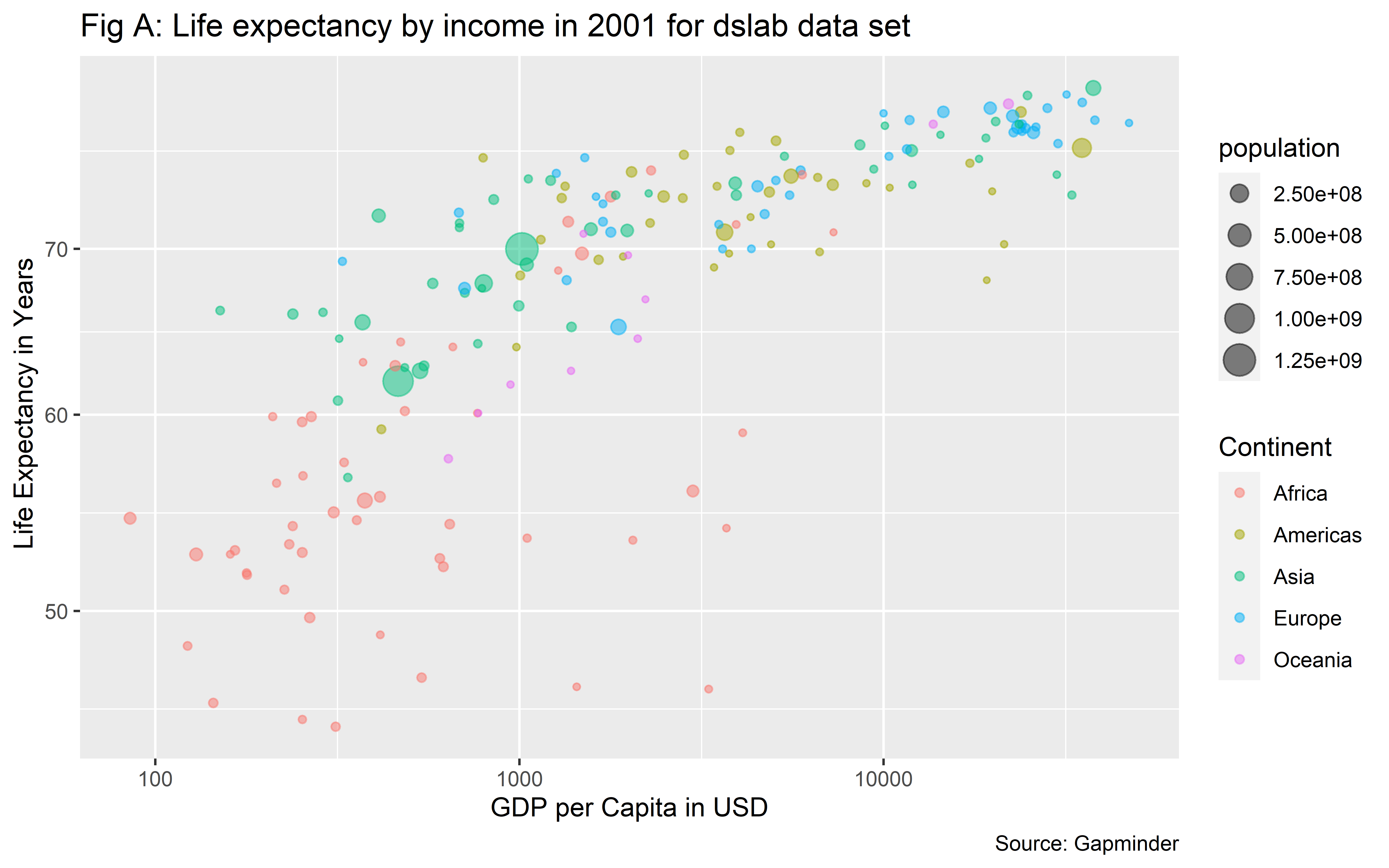
Next, we joined all six data sets to get one single data set, which we denoted as gapminder\_joined.

## Calculating the GDP/capita

We calculated a new variable, gdpPerCap, which is the value of GDP per Capita, by dividing the GDP by the population of a country for each year. We considered this as our final data set, and denoted it as gapminder\_df.

## Performing “sanity checks”

Before we tried to recreate the graphs from the previous report, we wanted to check if our newly created data set is aligned with the previous data set, gapminderold\_df. To check this, we performed a sanity check, where we created the bubble charts from both data sets (newly created vs old), and found both data sets showed almost similar bubble charts. Hence, we considered that our newly created data set aligns with the previous data set.



## Saving the wrangled data set as a .csv file.

We saved our data as a csv file and named it as gapminderwrangled\_df.csv.

# Rebuilding Project 1 figures by using newly created gapminder\_df data set.

Our gapminder\_df data set has 13 variables and 69731 observation. This data set has important demographic health indicators such as life expectancy, GDP, infant mortality and others, which will help to understand the global progress of health in terms of mortality, life expectancy and income among the population of different geographical locations.

## Finding the most recent year with complete data

To identify the most complete data, we filtered data from 2000 to 2016, and found that 2010 has the lowest missing value, which is 18.4% as the most recent year. Thus, we will be selecting 2010 as our most recent year with complete data.

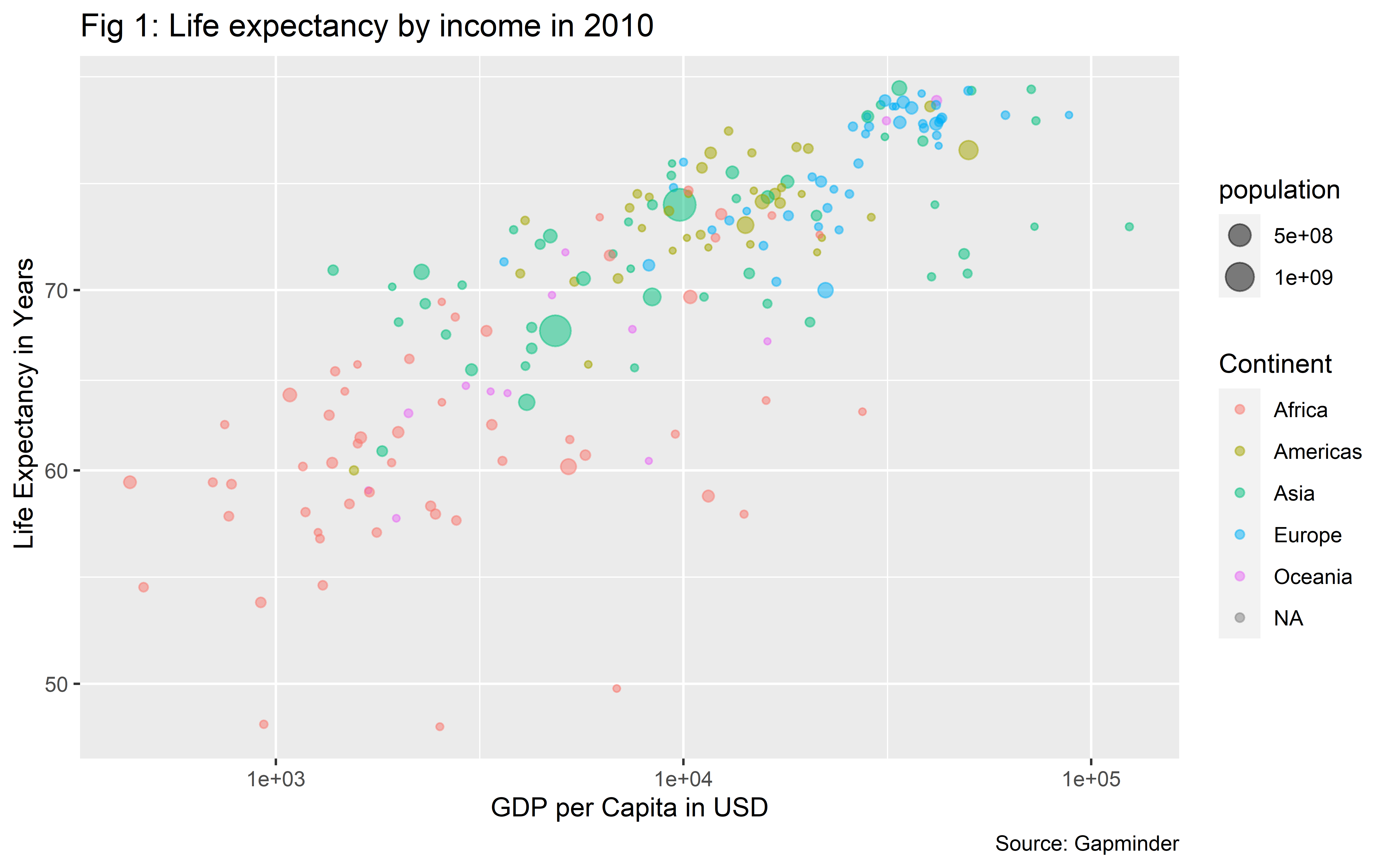
## The Bubble Chart

A bubble chart is commonly used to visualize the graphical relation or patter between three or more numerical variables. It can be used to show the relation between two variables for a fixed third variable or the trends over time between the other variables. In this report, we first showed a bubble chart for a single year and then, a trend for over a time period.

### The Bubble Chart-Single Year

As per the most famous tool on gapminder website, we prepared an interactive bubble chart where we showed ‘Life Expectancy by Income’ interaction for the most recent year with complete data, 2010. In our data, we do not have an income variable. However, we used gdpPerCap variable, which will represent the income per person.

**Hypothesis:** We hypothesize that countries with high GDP per Capita will show higher life expectancy in our bubble chart, which means, bubbles for such countries will be on the upper right corner of the chart.



**Overall comment:** In figure 1, the horizontal axis is GDP per Capita in USD and the vertical axis is life expectancy in year. In addition, the the size of the bubbles represents population of different countries in different continents.

Overall, figure 1 shows that countries with higher income (high GDP per Capita) have higher life expectancy, as depicted by the bubbles on the upper right corner of the graph. If we look at the color of the bubbles, we can see the red bubbles, representing African countries are clustering on the lower left corner of the graph, indicating this continent has the lower income per person and life expectancy. Next, the size of the bubble that represents the population size, shows countries with high income and life expectancy have smaller bubbles (Europeans and Americans) compared to the countries with low life expectancy and income (Africans).

According to the World Bank rankings for 2011, *Ref1:* <https://blogs.worldbank.org/opendata/changes-country-classifications> we can see the high-income countries had average incomes of $12,276 or more. In the x-axis of the graph, we can see most of the African countries (red bubbles) fell on the left side of GDP per Capita value = $10,000, which supports the world bank ranking that low income countries had average incomes less than $12,276 . On the other hand, in the y-axis, these countries fell on the lower part of the chart, representing life expectancy was mostly between 50 to 65 years, in contrast with high-income countries, where the value shows above 70 years.

This graphical distribution aligns with our previously stated hypothesis. Therefore, we can say, high-income countries had higher life expectancy in 2010.

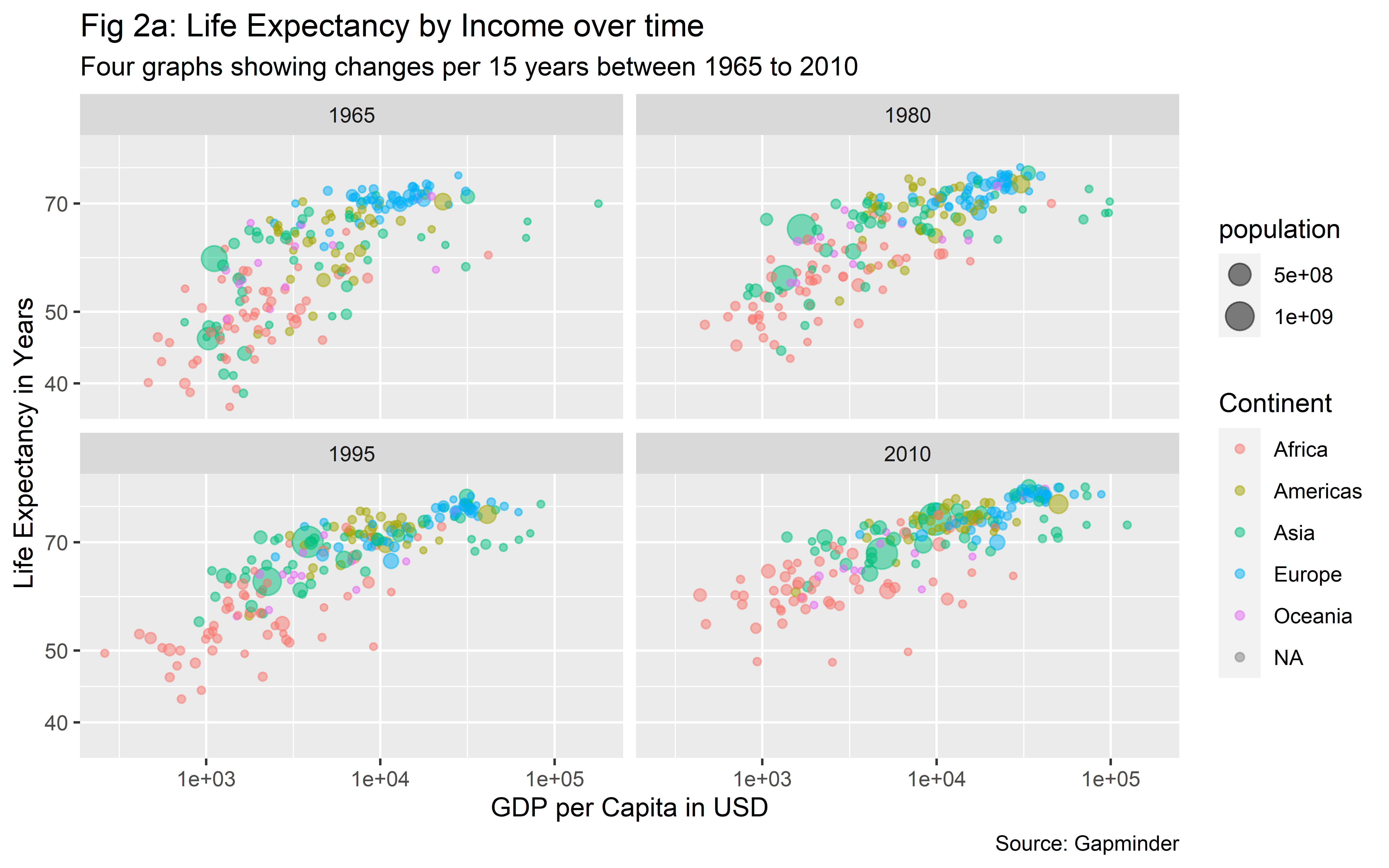
**Comment on outliers:** We can see two larger green bubbles, representing two highly populated countries located in the Asia continent, one with high life expectancy, and another with lower life expectancy and income. Furthermore, we can see two African countries have a life expectancy of around 45 years, representing countries with inadequate health infrastructure, public health regulation and health budget.

**Public health implication:** Our bubble chart data shows in 2010, most of the low-income countries had life expectancy between 50 to 60 years, whereas, the high-income countries had between 70 to 80 years. Therefore, adequate public health strategies should have been taken to reduce this gap of life expectancy between high-income and low-income countries.

### The Bubble Chart over Time

In this bubble chart, we showed the same ‘Life Expectancy by Income’ interaction over the time period instead of one single year to understand the trend of change.

**Hypothesis:** We hypothesize that over time we will see bubbles (countries) will be shifting on the upper right corner of the graph, as we expect over the years income and life expectancy have meaningfully improved.



**Overall comment:** In figure 2a, the horizontal axis is GDP per Capita in USD and the vertical axis is life expectancy in years. In addition, the size of the bubbles represents the population of different countries in different continents. The bubble charts are shown for four specific years with a 15 years gap, starting from 1965 until 2010.

Figure 2a shows that over time countries were shifted on the upper right corner of the graph, as over the years life expectancy have improved meaningfully for most of the countries, specifically for African countries and income has increased slightly. For each year in 15 years apart starting from 1965 until 2010, countries with higher income (high GDP per Capita) have higher life expectancy, as depicted by the bubbles on the upper right corner of each graph. If we look at the color of the bubbles, we can see the red bubbles, representing African countries are clustering on the lower left corner of the graph, indicating this continent has the lower income per person and life expectancy. Next, the size of the bubble that represents the population size, shows countries with high income and life expectancy have smaller bubbles (Europeans and Americans) compared to the countries with low life expectancy and income (Africans).

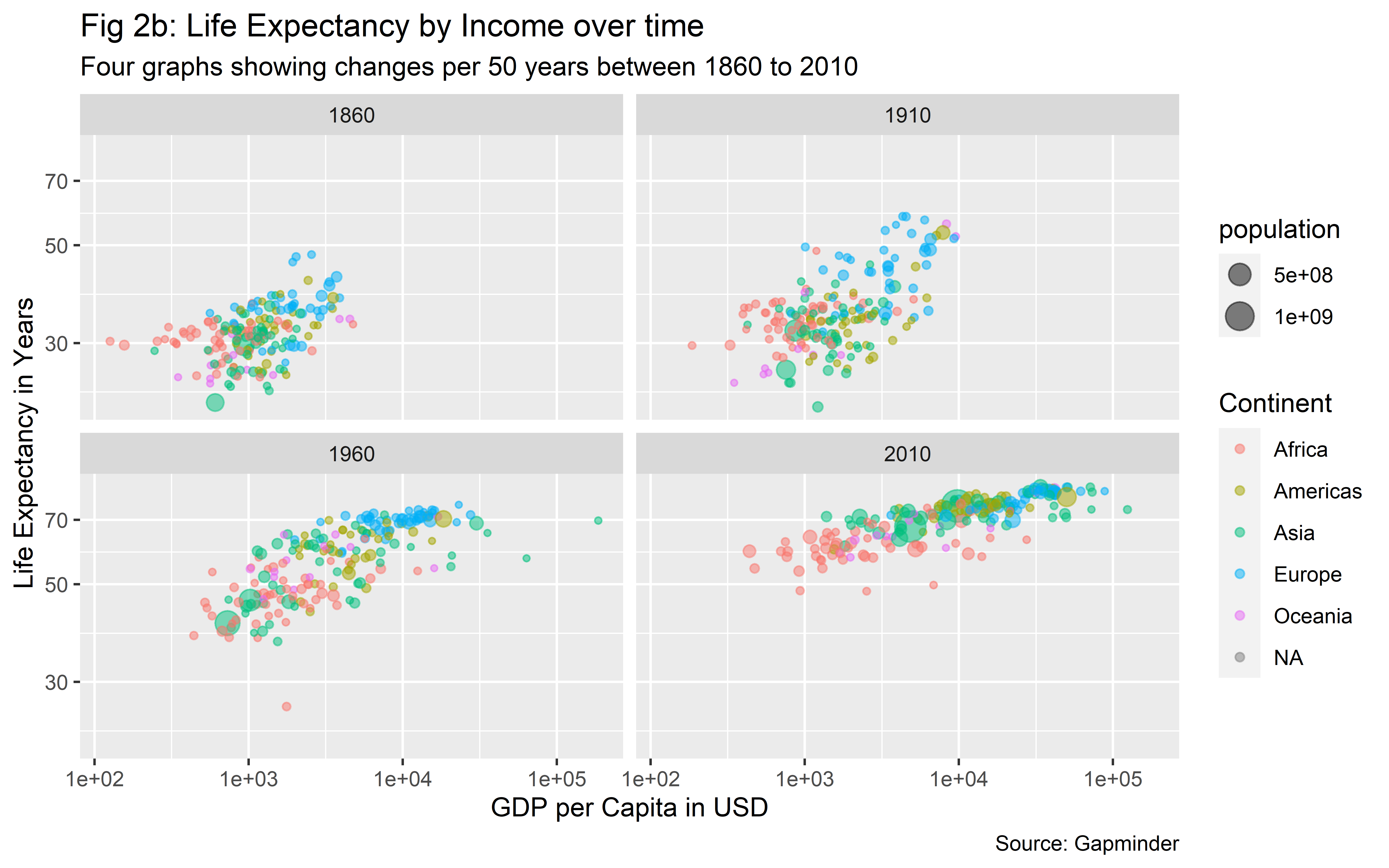
If we look at the time trend, at the beginning of modern era, in 1965, the bubble plots were wider in the y-axis and bubbles were relatively distance from each other than compared to that of 2010. The bubbles started clustering closely in the next 15 years trends and it was mostly clustered in 2010 compared to that of 1965. In a nutshell, this figure depicts that over time in the modern era, life expectancy increased, specially for the African continent, however, the change in GDP per capita for this continent was not much.

In the x-axis of the graph, we can see most of the African countries (red bubbles) fell on the left side of GDP per Capita value = $10,000 in 1965, which supports the world bank ranking that low income countries had average incomes less than $12,276 (). On the other hand, in the y-axis, these countries fell on the lower part of the chart, representing life expectancy were mostly between 50 to 65 years, in contrast with high-income countries, where the value shows above 70 years. If we look at the other three graphs, we can see, for each 15 years trends, the red bubbles were going up and slightly on the right side, indicating that life expectancy is increasing and income in also slowly increased for some African countries.

This graphical distribution aliens with our previously stated hypothesis. Therefore, we can say, high income countries had higher life expectancy in 2010.

**Comment on outliers:** In 1965, we can see one larger green bubble, representing a highly populated country located in Asia continent (China) with lower life expectancy and income. In 1980, we can see there are two large green bubbles, depicting drastic population increase among two countries in Asia continent (China and India), which can be explained by the revolutionary event, the end of the Vietnam War, occurred between 1955 to 1975. In 1995, both the green bubbles have shifted toward right side of x-axis, showing their increase in income over next 15 years after the war ended. However, the African countries were still on the left side of the graph, showing low income and short life expectancy. One reason could be the epidemic of HIV/AIDS in the early 80s and 90’s in African countries, with limited treatment and healthcare facilities. : <https://www.sciencedirect.com/science/article/pii/S0304387805000775>. Interestingly, in 2010, both the green bubbles shifted towards the high income side, however, African countries were still under the short life expectancy category. Another interesting fact is, for Asian countries, we can see a trend of increased GDP per capita over the years.

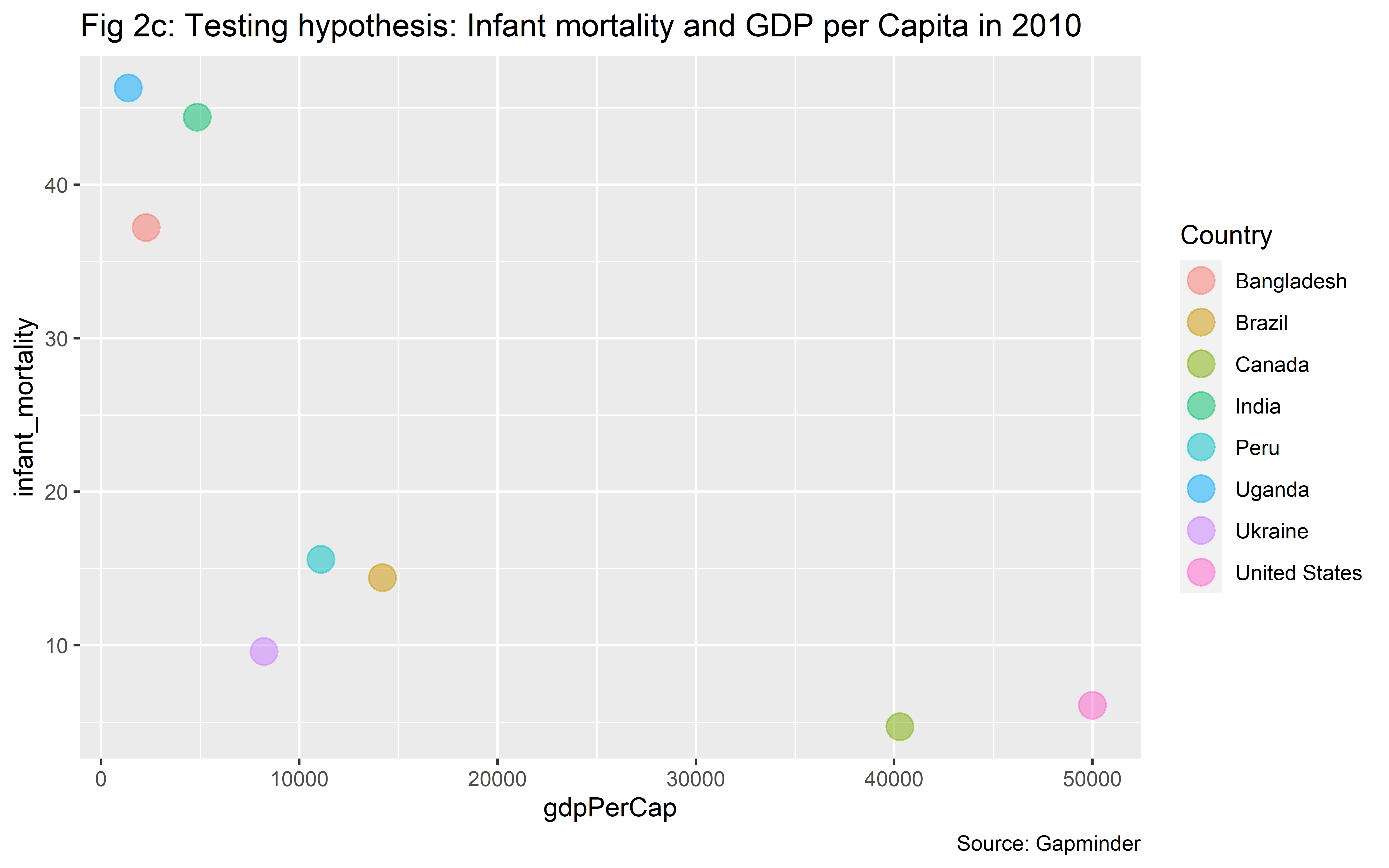
**Public health implication:** Our bubble chart data over time shows that despite the high prevalence of deadly HIV/AIDS, African countries with limited healthcare facilities kept improving their life expectancy rate. This indicates that public health measures were implemented in this continent to fight against the AIDS and other communicable diseases. However, health care is a complex system and has interaction with multiple components. Therefore, the life-expectancy change was not drastic among this continent.



**Overall comment:** In figure 2b, the horizontal axis is GDP per Capita in USD and the vertical axis is life expectancy over the years. In addition, the size of the bubbles represents population of different countries in different continents. The bubble charts are shown for four specific years with a 50 years gap, starting from 1860 until 2010.

Figure 2b shows that over time countries were more scattered in the earlier years compared to the recent years. In addition, over time, countries have shifted on the upper right corner of the graph, as over the years life expectancy has improved meaningfully for most of the countries, specifically for African countries and income has increased slightly. For each year in 50 years apart starting from 1860 until 2010, countries with higher income (high GDP per Capita) have higher life expectancy, as depicted by the bubbles on the upper right corner of each graph. If we look at the color of the bubbles, we can see the red bubbles, representing African countries are clustering on the lower left corner of the graph, indicating this continent has the lower income per person and life expectancy. Next, the size of the bubble that represents the population size, shows countries with high income and life expectancy have smaller bubbles (Europeans and Americans) compared to the countries with low life expectancy and income (Africans).

The overall comments on outliers and public health implication are reflecting the same as discussed above in figure 2a.



**Overall comment:** In figure 2c, the horizontal axis is GDP per Capita in USD and the vertical axis is infant mortality for 2010. In addition, the color of the bubbles represents country.

We selected our countries according to the world bank ranking, <https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html> for the year of 2010. We selected a pair of countries for high-income (United States and Canada), upper-middle income (Brazil and Peru), lower-middle-income (India and Ukraine) and low-income (Bangladesh and Uganda).

Figure 2c shows a negative relation between infant mortality and GDP per capita, as the infant mortality was lower for countries those who had higher GDP per capita and vice versa. High-income and higher-middle-income countries have more access to healthcare and infrastructural facilities, therefore, the child mortality is much lower than lower-middle and low-income countries.

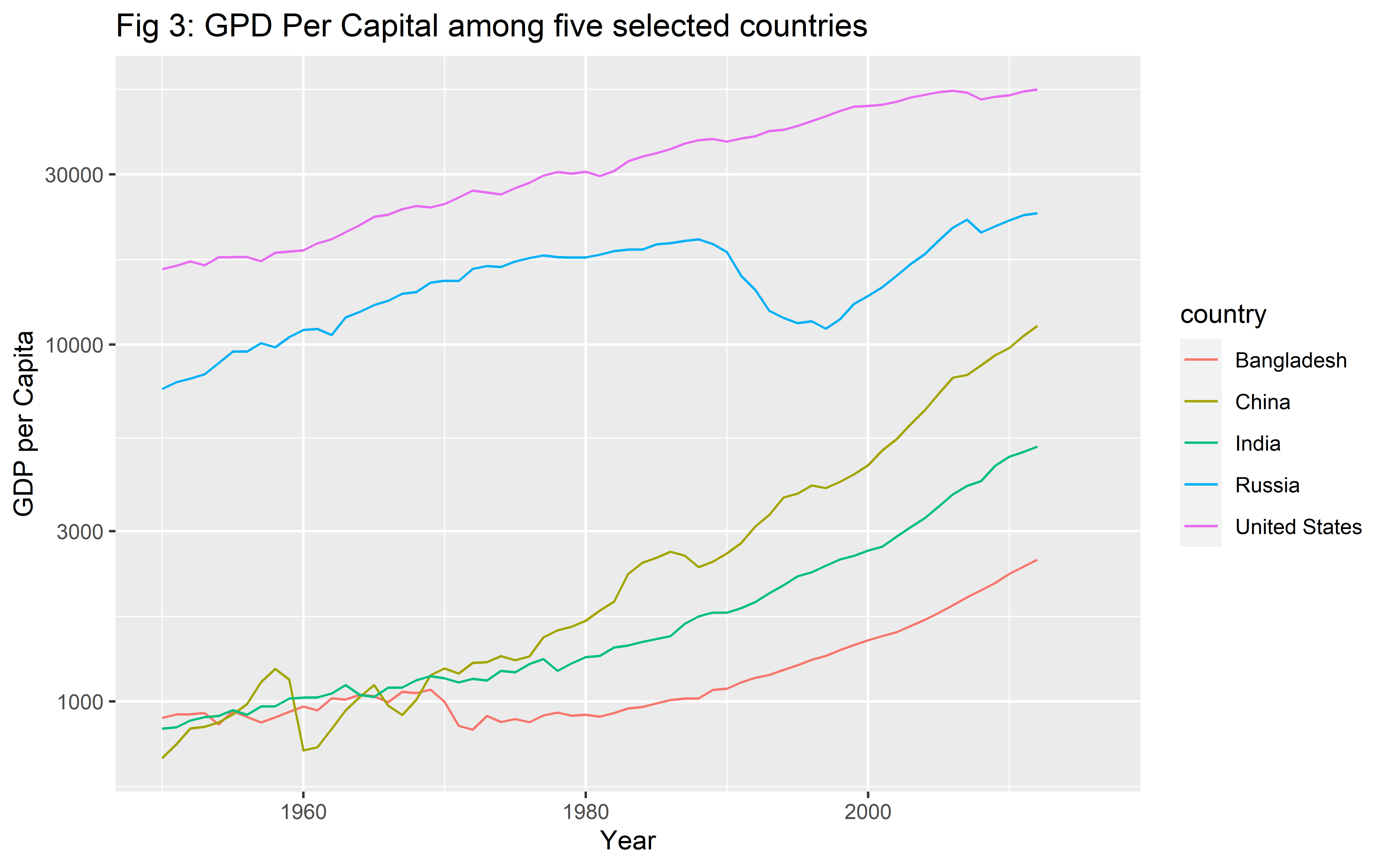
## The Growth Chart

Here, we described the data by using growth chart to illustrate the growth of our expected variables over time. It showed us the direction of our expected variables both in past and into the future time period. At first, we showed growth chart for GDP per Capita (income) for five selected countries over time, and then infant mortality for the same countries over time.

### GDP per Capita

GDP per Capita of a nation reflects the annual income per person in USD for a particular country. It is also called the income per person for a given year.

**Hypothesis:** We hypothesize that over time, curve of our selected five countries will be shifting on the upper right corner of the graph and rise, as we expect over the years income per person has improved for all five countries.



**Overall comment:** In figure 3, the horizontal axis is Year and the vertical axis is GDP per Capita. In addition, the color of the lines represents country for five selected countries.

Figure 3 shows that the curve for the USA is on the top of than rest of all countries, which shows it always had GDP more than 10000, and it increased over time. In 2016, its GDP per capita increased to more than 50,000 USD. It remained as a high-income country over the whole time period.

For Russia, the data were missing until around 1990, then the curve went downwards, and then near 2000, income of Russia started to increase until 2016. The data before 1990 were missing for Russia because it was a part of the Soviet Union before that, and in the late 1991, with the collapse of the Soviet Union, Russia was considered as an independent country afterwards. As a new nation it struggled initially, so the GDP per capita tend to decrease and Russia was considered as lower middle income country. However, from 2000 it started to increase the income, and from 2004 it was considered as Upper middle income country.

For China, the change was very significant over the years if we compare it with USA. In 1960, it was at the bottom of graph, depicting a low-income country with GDP per capita around 100 USD. This is because China faced the largest or second-largest famine in human history, the Great Chinese Famine (Great Leap) between 1959 and 1961, where millions of people died in China (estimates ranging from 15 to 55 million). : <https://www.everycrsreport.com/reports/RL33534.html> The curve was not steady for next 10 years, but it started increasing afterward, as the Great Leap Forward’, a five year economic recovery plan was introduced by the Chinese Communist Party (CCP). Just in 40 years, the GDP per capita for China increased to 10 folds (100 to 1000 USD), and it became a lower-middle income country from a low-income country in 1999. The increase was steady and steep for the next 10 to 15 years for China, and in 2010 it became an upper middle income country by reaching the income level of Russia. It has crossed the GDP per capita over 4000 USD at the end of 2016. add what happened in china in 1960

The relation between India and Bangladesh is interesting over the years. In 1960 the GDP per Capita for India was almost double than Bangladesh, around 400 UDS and 250 USD, respectively. In 1970, income started to coming down for India, and it came down close to Bangladesh. This is because of the effect of war between Pakistan and Bangladesh in 1971 as the geographical location of India lies in the middle of these two war zones. Around 1980, GDP for Bangladesh came down as it was recovering after its independence in 1971 from Pakistan and started from scratch as a new independent nation. Meanwhile, India doubled their per capita GDP. For the next two decades, both countries had increased income, however, the gap between the countries remained the same until 2010.

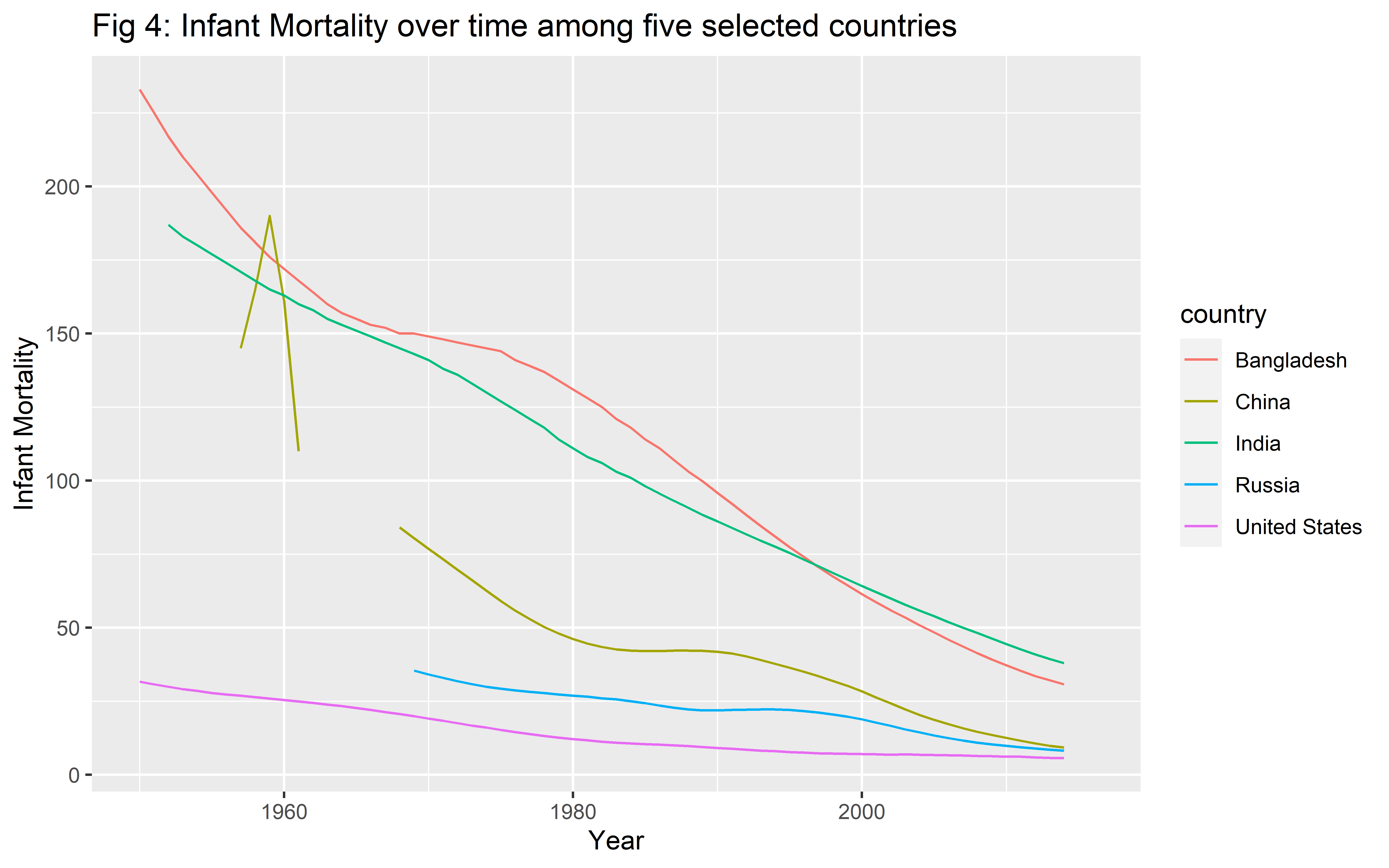
**Comment on outliers:** Data were missing for Russia until 1990, hence, we were unable to conclude the trend for Russia over the period as we did for other four countries. The USA was the high income country from the beginning, and with time, the income per person in the USA has increased steadily.

**Public health implication:** Each country’s national health expenditure is dependent on the GDP per capita. Therefore, for all there five countries, we can say, as the income per person increased, they were able to spend more on health expenditure, which eventually improved other health indicators. Thus, a healthy nation contributed more on earning higher GDP.

### Infant Mortality

Infant mortality is defined as the death of an infant before his or her first birthday. It is an important health indicator for a nation that indicates the number of infant deaths for every 1,000 live births per year.

**Hypothesis:** We hypothesize that over time, the curve of our selected five countries will be shifting on the lower right corner of the graph and show a downfall, as we expect over the years infant mortality will be decreased in all five countries.



**Overall comment:** In figure 4, the horizontal axis is Year and the vertical axis is Infant Mortality. In addition, the color of the lines represents country for five selected countries.

Figure 4 shows that in 1960, the US had the lowest infant mortality (around 25 children were dying for every 1,000 live births) whereas China had the highest infant mortality (around 190 children were dying for every 1,000 live births). Bangladesh and India were near China, 175 and 165, respectively.

For China, there was a sharp fall of infant mortality after 1960, however, the curve shows it is missing a part between 1963 to 1970. This is because China faced the largest or second-largest famine in human history, the Great Chinese Famine (Great Leap) between 1959 and 1961, where millions of people died in China (estimates ranging from 15 to 55 million). : <https://www.everycrsreport.com/reports/RL33534.html> The figure shows, infant mortality drastically decreased between 1970 to 1980, from more than 80 to less than 50 as the Great Leap Forward’, a five year economic recovery plan was introduced by the Chinese Communist Party (CCP). Over the next 35 years China has decreased its infant mortality to near the rate of Russia and the US, which is around 12.

Data for Russia in infant mortality is available after 1970, thus, we can see it has also decreased its infant mortality from 32 to 10 over the 45 years. This trend of reduction in the infant mortality was similar to the US, which reached around 5 over the 55 years from 25.

For Bangladesh and India, the trend of curves were highly steep, showing that in 55 years both countries have reduced the infant mortality from around 200 to around 40 or less. One main factor was both countries have taken initiatives in improving maternal and newborn child health care by reducing the share of births that were taking place in homes instead of health-care facilities. Increase of education level, breaking the stigma of seeing a male healthcare person, improvement of diet of pregnant women were other key factors that played important role in this.

**Comment on outliers:** Data were missing for Russia and China for few years. This is because Russia was a part of Soviet Union from 1964 to 1982 and China was facing the the largest or second-largest famine in human history (Great Leap) between 1959 and 1961.

**Public health implication:** Public health interventions were implemented effectively for all five countries over the years, which helped to reduce the burden of infant mortality among all five countries. Such interventions were more effective in low and middle income countries like Bangladesh and India compared to high income countries like the USA.

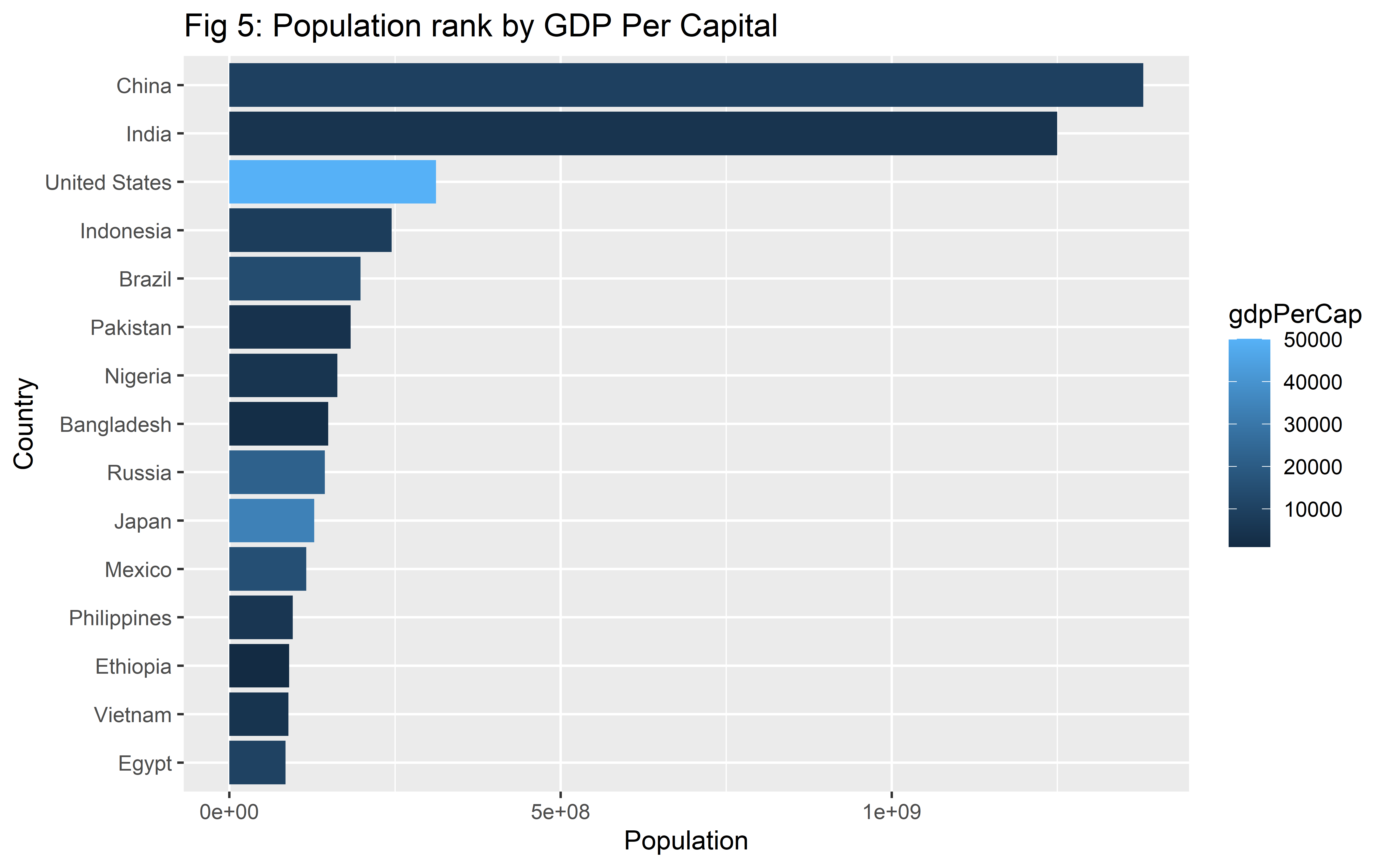
## Rank Charts

The rank chart shows the ranking or arrangement of the dimension of desired variables based on certain measures over a period of time.

Here, first, we ranked the countries by their population sizes (for the most recent year with data) and then we filled the bars by the GDP per capita of the countries. We showed the top 15 rows of countries. Next, we ranked the countries by their infant mortality rates (for the most recent year with data), and then we filled the bars by the population of the countries.

### Population Rank by GDP/Capita

**Hypothesis:** We hypothesize that countries with high population will have low GDP per capita compared to countries with lower population, because the earned GDP of high populated countries are divided by the higher number of people, hence the GDP per capita becomes smaller. On the other hand, low populated countries will have higher GDP per capita. Thus, in the rank chart we expect to see countries with high population and low GDP per Capita in the top and countries with low population and high GDP per Capita in the bottom.



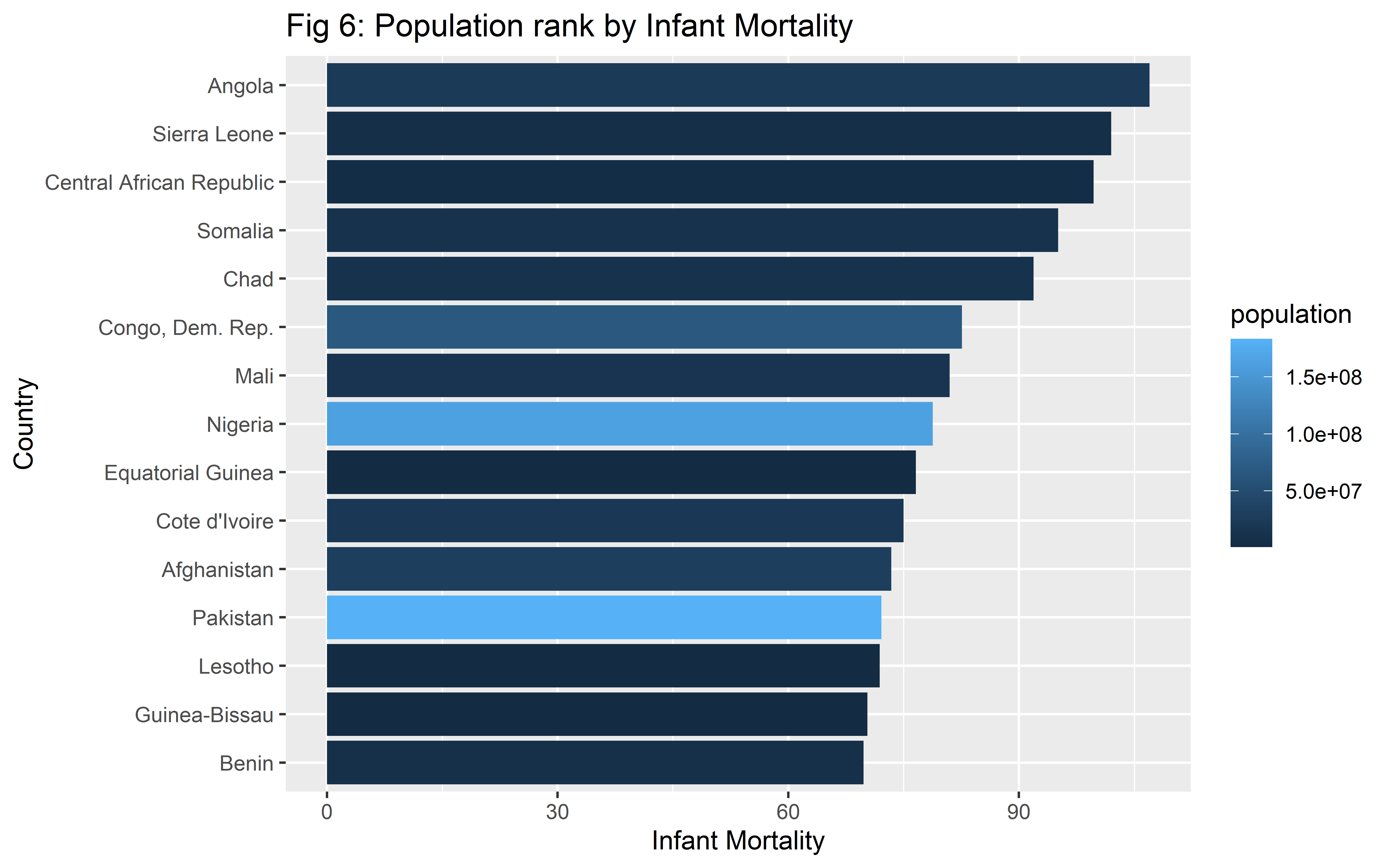
**Overall comment:** In figure 5, the horizontal axis is population and the vertical axis is country. The countries are reordered according to the highest to lowest population. In addition, the bars are filled with the value of GDP per Capita, where the darker blue shade indicates the lower GDP per Capita and lighter blue shade indicates the higher GDP per Capita.

Figure 5 shows that the most highly populated country is China, with dark shade indicating it has lower GDP per Capita. On the other hand, Germany is least populated country among the first 15 highly populated country, and it shows a lighter blue shade indicating the higher GDP per Capita. China and India have the highest population, more than 1 billion, and all other countries in this list have comparatively less population.

**Comment on outliers:** For the US and Japan, despite being in the 15 highly populated county list, they show a light shade in their bars, indicating high GDP per capita. This is because both countries have abundant natural resources, a well-developed infrastructure, and high productivity compared to most countries in this list.

**Public health implication:** This rank chart helps to understand how to implement public health strategies for highly populated countries based on their national income. It is helpful for the nations to understand how to prioritize, budget and spend the health and other expenditure. Overall, it loweres the costs of hospital and patient, and increases the equity in life.

### Infant Mortality by Population

**Hypothesis:** We hypothesize that countries with low population will have more infant mortality compared to countries with higher population, because we think countries with low population have less healthcare facilities and fewer healthcare service provider, hence, high infant mortality will occur. Thus, in the rank chart we expect to see countries with low population and high infant mortality rate in the top and countries with high population and low infant mortality rate in the bottom. 

**Overall comment:** In figure 6, the horizontal axis is infant mortality and the vertical axis is country. The countries are reordered according to the highest to lowest infant mortality. In addition, the bars are filled with the value of population, where the darker blue shade indicates the lower population and lighter blue shade indicates the higher population.

Figure 6 shows that highest infant mortality in 2009 was in Angola, with dark blue shade bar indicating it has lower population. Sierra Leone was the second highest country with almost similar high infant mortality and lower population of Angola. The 15th ranked country is Burkina Faso, which had the lowest infant mortality among the 15 countries. This country also has darker shade, indicating lower population.

**Comment on outliers:** Since these 15 countries are ranked as the most highest rate of infant mortality in 2009, we expected all bars will be filled with darker shades, indicating low income countries. However, we can see for Nigeria and Pakistan, although they have high infant mortality rate, their light blue shade indicates that they belong to highly populated countries.

**Public health implication:** This rank chart helps to understand what to prioritize in implementing public health strategies in the maternal and neonatal health care sector for countries with high infant mortality rate and low population. For example, during health budgeting, they should give importance in healthcare for neonatal care. Moreover, public health interventions to reduce infant mortality should be applied in countries like Nigeria and Pakistan despite having larger population.

# Conclusion

Overall, different graphs from the data depict the difference between countries and continents in terms of population, income, infant mortality over the period of 200 years time. Through these graphs it is easy to visualize the differences, which helps to take timely initiatives in terms of public health concerns. By using the bubble chart, growth curve and rank chart, it is easier to understand the trend of change over a certain years or for a fixed year between our interest of variables. The outliers help to figure out what makes a difference in a given trend of change. Based on these important data, many international health organizations and fund provider can make wise and timely decision on how to spend the proper use of money to facilitate the health care system and public health globally.

# Project group members and Acknowledgment:

In this project 1, we worked in a group of three members. We are, Manuel Roosevelt Lamptey, Rabeya Illyas Noon, and Tarana Ferdous. We are from Statistics, Biostatistics and Epidemiology departments, respectively. We would like to acknowledge our TA, Catalina and Anny for their explanation on our queries and Dr. Odom for this opportunity to learn through this project.