Project 1 by the group MNT

Exploring the GapMinder dataset using different graphs

Manuel Roosevelt Lamptey, Rabeya Illyas Noon, and Tarana Ferdous

2022/06/03

Table of Contents

[Introduction to the GapMinder World Health Data 2](#_Toc106717570)

[Installing library 2](#_Toc106717571)

[Creating a data subset 2](#_Toc106717572)

[Exploring the new data set 2](#_Toc106717573)

[Finding the most recent year with complete data 2](#_Toc106717574)

[The Bubble Chart 2](#_Toc106717575)

[The Bubble Chart–Single Year 3](#_Toc106717576)

[The Bubble Chart over Time 5](#_Toc106717577)

[The Growth Chart 7](#_Toc106717578)

[GDP per Capita 7](#_Toc106717579)

[Infant Mortality 9](#_Toc106717580)

[Rank Charts 12](#_Toc106717581)

[Population Rank by GDP/Capita 12](#_Toc106717582)

[Infant Mortality by Population 14](#_Toc106717583)

[Conclusion 15](#_Toc106717584)

[Project group members and Acknowledgment: 15](#_Toc106717585)

# Introduction to the GapMinder World Health Data

The data set GapMinder is a health data set that reflects the global population-level data with most of the countries from the early 20th century. In this project 1, we will be illustrating three interactive charts, bubble chart, growth chart and rank chart by using this data set and provide details of each graph to explain what they depict.

# Installing library

To proceed further, we need to first install required packages and library. For this project, we loaded the library for dslabs, plotly, ggplot2 and tidyverse.

# Creating a data subset

A subset of the this data is available through the R package dslabs, which we have already loaded. From here, we have now created the following data set named gapminder\_df. Here, we added one new variable, gdpPerCap to represent the GDP per capita, which measures the economic output of a nation per person, in other words, income per person of a country. We will be using the modified data set, gapminder\_df, for the rest of the project 1.

# Exploring the new data set

By using the following function str and summary, we can see the new data set has now ten (10) variables and 10545 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

From the function summary, we can identify the latest year of the data set from the variable year. We can see the earliest year was 1960 and latest year was 2016 in the data set. However, as we are looking for the most complete data, we used the function table1 to see the missing percentage for each year.

As per the result, year 2009 has the lowest missing value, which is 2.2% as the most recent year. Thus, we will be selecting 2009 as our most recent year with complete data.

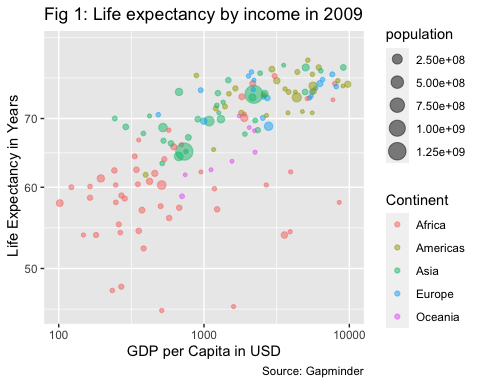
# The Bubble Chart

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 project 1, we will first show 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 in gapminder website, we are going to prepare an interactive bubble chart where we will show ‘Life Expectancy by Income’ interaction for the most recent year with complete data, 2009. In our data, we do not have income variable. However, we will use gdpPerCapvariable, which will represent the income per person, as described earlier.

**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.



**Graph orientation:** In our bubble chart, figure 1, at first, we selected our desired year, 2009, by using filter function in ggplot at the data layer.

Next, at the aesthetic layer, first we added the independent variable, GDP per Capita in the x-axis, which reflects the annual income per person in USD for a particular country. Next, we added the dependent variable, life expectancy in the y-axis, which reflects the mean number of years for people in a particular country might expect to live according to their current age-specific mortality rates. We want to show the linear relation between income and life expectancy, as both variables are continuous. Furthermore, we added color as continent to highlight the five different continents in our graph to show the distribution of geographical location. Next, we added size as population of each country, to depict the bubble size by highlighting the total population of respective country. Finally, we positioned the scale for our continuous variables, x and y by transforming them into log10 and set the limit between 100 and 10000.

At our last layer, geometric layer, we added geom\_point at alpha value 0.05. Here, alpha is used to represent the opacity of the bubbles. The range of alpha can vary between 0 to 1, and here we kept the opacity in the middle by choosing 0.5 to notice if any country with smaller population is merged inside the bubble of a larger country population.

**Overall comment:** 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 were mostly between 50 to 65 years, in contrast with high-income countries, where the value shows above 70 years.

This graphical distribution aliens with our previously stated hypothesis. Therefore, we can say, high income countries had higher life expectancy in 2009.

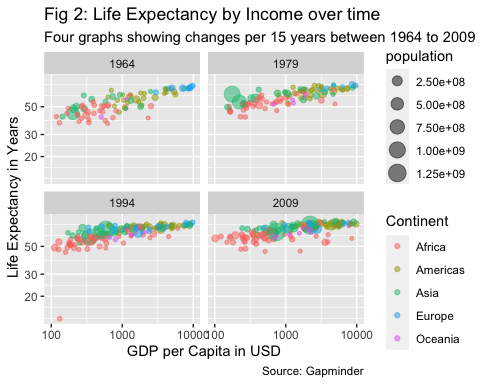
**Comment on outliers:** We can see two larger green bubbles, representing two highly populated countries located in Asia continent, one with high life expectancy, whereas another with lower life expectancy and income. Furthermore, we can see two African countries have life expectancy around 45 years,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

**Public health implication:** Our bubble chart data shows in 2009, 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 and low income countries.

## The Bubble Chart over Time

In this bubble chart, we will be showing 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.



**Graph orientation:** In our bubble chart, figure 2, at first, we selected four (4) specific years with a 15 years gap, starting from 1964 until 2009, by using filter function in ggplot at the data layer. For the aesthetic layer and geometric layer, we have applied the same codes as we described above. At the end of the code, we added a function facet\_wrap to show all four bubble chart in one figure.

**Overall comment:** Figure 2 shows that for each year in 15 years apart starting from 1964 until 2009, 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 over the time trend, at the beginning of modern era, in 1964, the bubble plots were wider in the y-axis and bubbles were relatively distance from each other than compared to that of 2009. The bubbles started clustering closely in the next 15 years trends and it was mostly clustered in 2009 compared to that of 1964. In a nut shell, 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 1964, 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 2009.

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.

**Comment on outliers:** In 1964, we can see one larger green bubble, representing a highly populated country located in Asia continent with lower life expectancy and income. In 1979, we can see there are two large green bubbles, depicting drastic population increase among two countries in Asia continent, which can be explained by the two revolutionary events, the civil rights movement and the end of the Vietnam War, occurred between 1960 to 1980. In 1994, 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 introduction of HIV/AIDS in the early 80s and 90’s in African countries, with limited treatment and healthcare facility. : <https://www.sciencedirect.com/science/article/pii/S0304387805000775>. Interestingly, in 2009, 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 introduction 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.

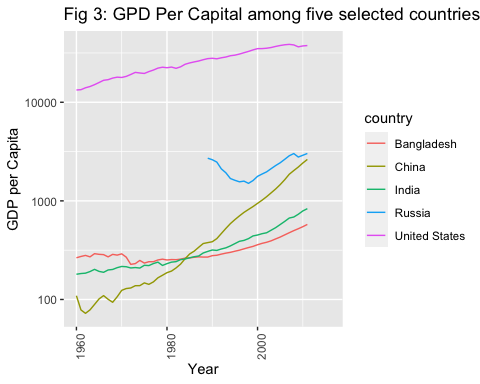
# The Growth Chart

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



**Graph orientation:** In our growth chart, figure 3, at first, we selected our desired countries, China, Russia, the US, India, and Bangladesh, by using filter function on country variable and created a new subset of data called growth\_gdp\_per\_cap. Next, we used ggplot for the data set growth\_gdp\_per\_capat the data layer.

Next, at the aesthetic layer, first we added the independent variable, year, in the x-axis from 1960 to 2000 range where each grid shows 10 years, and added the dependent variable, GDP per Capita in the y-axis, which reflects the annual income per person in USD for a particular country ranging between 100 to 10000, where each two grid shows a 10-fold of increase. We want to show the linear relation between income and year, as both variables are continuous. Furthermore, we added color as country to highlight the five different countries in our graph to show the differences between five curves. Next, we positioned the scale for y variable by transforming it into log10. Next, by using the labs function, we added the title and axis for the growth curve. Finally, we used the theme function to change the x-axis text angle into 90 degree.

At our last layer, geometric layer, we added geom\_line to create the growth curve.

**Overall comment:** Figure 3 shows that the curve for USA is on the top 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 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. The curve was not steady for next 10 years, but it started increasing afterwards. 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.

The relation between India and Bangladesh is very interesting over the years. In 1060 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, whereas Bangladesh, as a new independent nation, started from scratch. Around 1985, the GDP per capita for both nations were similar, and in early 1990, Bangladesh overtook India in its GDP per Capita indicator. Over the next 20-25 years, GDP per Capita for Bangladesh became near 1000 USD whereas, for India, it is around 600 USD. As per World Bank data, India and Bangladesh became a lower middle income country from low income in in 2007 and 2014, respectively. A recent article, : <https://foreignpolicy.com/2020/10/20/india-falls-behind-bangladesh-gdp/> argued that Bangladesh has utilized the development of a low-skill, low-wage manufacturing export sector, whereas in India, the highest populated states do not have any export-friendly structure.

However, the above discussion follows the trend given in the world bank ranking in the reference 3. : <https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html>

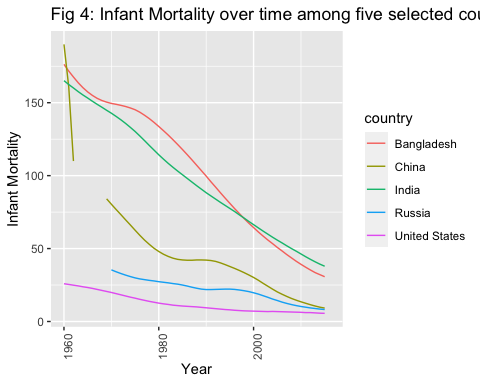
**Comment on outliers:** Data were missing for Russia until 1990, hence, we were unable to conclude the trend for Russia over the period as we did for other four countries. The USA was the high income country from the beginning, and with time, the income per person in the USA has increased steadily.

**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, 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.



**Graph orientation:** This graph is similar to figure 3, except in the y-axis we used infant mortality instead of GDP per capita variable and in scale we consider both x and y as continuous. We made change in the respective layers of the graph as required.

**Overall comment:** 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, probably because of under reporting or data missing at that time as the population was booming at that time in China. The figure shows, infant mortality drastically decreased between 1970 to 1980, from more than 80 to less than 50 and China has introduced ‘one child policy’ after that in 1980 as low infant mortality rate has increased its population drastically. 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 to 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.

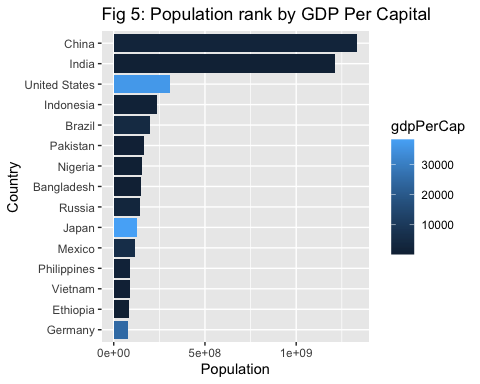
**Public health implication:** Public health interventions were implemented effectively for all five countries over the years, which helped to reduce the burden of infant mortality among all five countries. Such interventions were more effective in low and middle income countries like Bangladesh and 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 will rank the countries by their population sizes (for the most recent year with data) and then we will fill the bars by the GDP per capita of the countries. We will show the top 15 rows of countries. Next, we will rank the countries by their infant mortality rates (for the most recent year with data), and then we will fill 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. 

**Graph orientation:** In our rank chart, figure 5, at first, we selected our desired year, 2009, by using filter function on year variable of gapminder\_df data set and arranged the population in descending order by using arrange function. We also used the head function to show the first 15 rows in the rank chart. Here, we created a new subset of data called rank\_chart and used ggplot for this data set at the data layer.

Next, at the aesthetic layer, first we added the independent variable, population, in the x-axis. In the y-axis, we added the countries after reordering them according to the highest to lowest population. We filled the bars 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. Next, we added the title and labels for x and y axis.

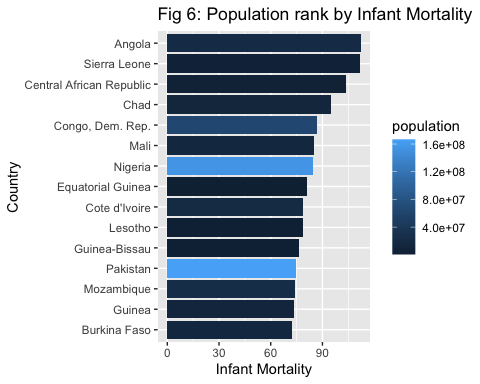
At our last layer, geometric layer, we added geom\_bar to create the rank chart. Here we used stat="identity" function to keep the actual values of y variable.

**Overall comment:** 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 are among the high income countries and they are sparsely populated, therefore, bigger in size compared to other countries.

**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. 

**Graph orientation:** In our rank chart, figure 6, at first, we selected our desired year, 2009, by using filter function on year variable of gapminder\_df data set and arranged the infant mortality in descending order by using arrange function. We also used the head function to show the first 15 rows in the rank chart. Here, we created a new subset of data called infant\_rank\_chart and used ggplot for this data set at the data layer.

Next, at the aesthetic layer, first we added the independent variable, infant mortality, in the x-axis. In the y-axis, we added the countries after reordering them according to the lowest to highest infant mortality. We filled the bars with the value of population, where the darker blue shade indicates the lower population and lighter blue shade indicates the higher population. Next, we added the title and labels for x and y axis.

At our last layer, geometric layer, we added geom\_bar to create the rank chart. Here we used stat="identity" function to keep the actual values of y variable.

**Overall comment:** 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 50 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.