Descriptive analysis of demographic data

Author: Azeezat Mosunmade Mustapha

July 26, 2023

Contents

1 Introduction

Earlier demography research had suggested that no country in the world had life expectancy of more than 40 years, almost all of the world was relatively poor with little to no medical knowledge which had caused people to die younger with high child mortality and low life expectancy (Max Roser and Ritchie, 2013). Child mortality rate is the death of children under the age of 5 per 1,000 live births and also a metric for life expectancy as a large number of children's death means a number of the population not living to old age, thereby reducing the lifespan (Roser et al., 2013).

Due to industrial revolution, new technology, research in medical field, increase in wealth, birth of contraceptive techniques, new abortion laws, education, e.t.c child mortality had reduce all over the world (Roser et al., 2013) with increasing life expectancy. A data set containing life expectancy at birth and under age 5 mortality rates of 227 countries in year 2002 and 2022 is used to examine the relationship between under age 5 mortality rates and life expentancy at birth and how the data change over the years for different regions and subregions. Data from 2022 was examined to understand the relationship between child mortality rate and life expectancy at birth and how the data vary between male, female and different regions. Also, the data from 2022 was compared with data from 2002 to learn how these variables have change over time.

Firstly, Histogram is used to describe the frequency distribution of each of the variables and the differences based on sexes and regions. To check if there are relationships between the variables, correlation and scatter plots are used. Boxplot is used to visualise if variables are homogenous within the subregions and heterogenous between different subregions. Boxplot is also used to check how the values of the variables have changed over the years from 2002 to 2022.

In section two, the data set and the quality of data is examined, additionally the objectives of this report are stated. In section three, the statistical methods used in this report are explained, that is, histogram, boxplot, correlaton and Scatter plots. In the fourth section, the explained Statistical methods in section three are applied to the data set and the results are presented.The concluding section five covers summary of the findings and suggestion on further research.

2 Problem statement

2.1 Data set and data quality

The data set analysed in this report is a small extract from the International Database, (IDB, 2021) U.S. Census Bureau which contains various demographic data (currently from 1950 to 2100) on all states and regions of our world that the US Department of State recognizes and have a population of 5,000 or more. The database sources are information from state institutions, such as censuses, surveys or administrative records, as well as estimates and projections by the U.S. Census Bureau itself which can be assumed to be a reliable data source for geographic and demographic data.

The dataset includes names of countries, subregions, regions, year, under-5 mortality rates of females, under-5 mortality rates of males, under-5 mortality rates of both sexes, life expectancy at birth for males, life expectancy at birth for females and life expectancy at birth for both sexes. Under 5 mortality rate is the chance that a child will die before age 5 for every 1,000 live birth and Life expectancy at birth is defined as the average number of years a cohort of people born in the same year can be anticipated to live if mortality at each age remains constant in the future. (Glossary, 2021).

There are 227 countries from 2002 to 2022. The countries are divided into 5 regions namely: Africa, Americas, Asia, Europe and Oceania and 21 subregions. The countries, regions, subregions are all strings while year, under 5 mortality rates of female,under 5 mortality rates of males, under 5 mortality rates of both sexes, life expectancy at birth for male, life expectancy at birth for female and life expectancy at birth for both sexes are numeric variables. All variable definitions were obtained from U.S. Census Bureau (Glossary, 2021).

There are missing data from subregions and regions of Curaçao and Côte d'Ivoire for years 2002 and 2022 while Libya, Puerto Rico, South Sudan, Sudan, Syria and United States have missing data for the year 2002 from all the numerical variables except for the year making a total of 44 missing data in all. The missing data for regions and subregions are fixed by entering the Region and Subregion while for the numerical values, missing values are replaced with the mean values for each subregion in the year 2002 where the country belongs.

The variables considered in this project are under 5 mortality rates of female,under 5 mortality rates of males, under 5 mortality rates of both sexes, life expectancy at birth

for both sexes, life expectancy at birth for male, life expectancy at birth for females, year, regions, and subregions. Under 5 mortality rates is refered to as Child mortality in this report also, the variable names are shortened to life expectancy at birth for both sexes to LEBS, life expectancy at birth for male to LEM, life expectancy at birth for females to LEF, under 5 mortality rates of female to MRF, under 5 mortality rates of males to MRM and under 5 mortality rates of both sexes to MRBS.

2.2 Project objectives

This project focus on the data from year 2022. Histogram is used to describe the frequency distribution of the variables and differences in sexes and regions.

Data set is grouped into subregions to understand if the variables are homogenous within the subregions and heterogenous between different subregions with Asian region as a focus.Three measures of central measure of central tendency(Median) and Boxplot are used to analyse the variations in the subregions with a focus on one region(Asia).

Scatter plots and correlation are used to anlayse the relationships between life expectancy at birth and under age 5 mortality rates.

Box plot is also used to visualize how the values of the variables have changed from year 2002 to 2022.

3 Statistical methods

3.1 Measure of central tendency

Measure of central tendency summarizes a given data set into a value in order to describe it by identifying the center position in the data. The 3 measures are mean, median and mode (Bluman, 2012). For the purpose of this report, median is explained as it is less affected by outliers in the data(Outliers are values in the data set that are extremely different from other values in the same data set).

Median is the middle value after the values of observations have been arranged in either ascending or descending order. If the number of observations is even, the 2 numbers in the middle are added and divided by 2 to get the median (Hay-Jahans, 2019).

3.2 Histogram

Histogram is a graph of the frequency distribution of a continuous variable with bars of different heights. The data are displayed by dividing the entire range of values into a series of intervals called class width or bins. The height of each bin is determined by the frequency of the class width. The number of bins is at user's discretion (Bluman, 2012). The histograms in this report are plotted using frequency density which is the frequency per unit of the data in each bin, this makes it easy to compare histograms (Hay-Jahans, 2019).

Frequency density is given by:

$$Frequency\ density = \frac{Frequency}{Class\ width}$$

Where: Class width is the size of the bin.

Frequency is the count of values in the bin.

3.3 Scatter plot

Scatter plots are used to display the relationship between two variables, each variable is represented on an axis, and the data is presented as a collection of points (Bluman, 2012). Each point on the chart is determined by the collection of each variable. A line of best fit can be drawn over the points to visualize the pattern on the chart.

3.4 Correlation

Correlation measures the strength and direction of the relationship between two quantitative variables. The value is called correlation coefficient and it is simply represented by r. The value of r ranges from -1 to +1 (Bluman, 2012). This report used the Pearson correlation(r) which measures the linear relationship between two variables. The Pearson correlation coefficient between 2 variables x and y with n number of observations is given below:

$$r = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{[n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2][n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2]}}$$

where x_i = values of x variables and y_i = values of y variables. (Bluman, 2012)

A positive r-value shows that there is a positive linear relationship between the two variables. That is, as one variable increases the other also increases. A negative r-value suggest that as one variable increases the other reduces. 0 r-value indicate that there is no relationship between the variables. The closer r is to 1 or -1 determines how strong the strength of the relationship of the 2 variables being measured.

3.5 Box plot

Box plot is a graphical representation of the five-number summary of a given data. The five numbers include minimum value in the data, maximum value in the data, median, first quantile (Q1), and third quantile(Q3). It can be used to visualize how numerical data is distributed (Bluman, 2012). It can be used to compare the distribution of more than one variable when plotted on the same graph.

Box plot is drawn by drawing a horizontal line from the minimum value to Q1, and a horizontal line from the maximum value to Q3. Then drawing a box to connect the horizontal line of Q1 and Q3 with Q2 falling inside the box.

The longer the box, the higher the interquartile range(IQR). This gives how spread the data is in the middle. It is the difference between Q3 and Q1

To get the quantiles, the data is ordered from lowest to highest. The data that falls into the middle is called the Q2, if there are two numbers in the middle the average of the two numbers is taken as the median.

The median of the group that falls below Q2 is called Q1 and the median of the group that falls above Q2 is called Q3. If the IQR is large it means there is high variability in the data and vice versa (Bluman, 2012).

The statistical software R (R Development Core Team, 2021), version 4.0.3 was used for analysis.

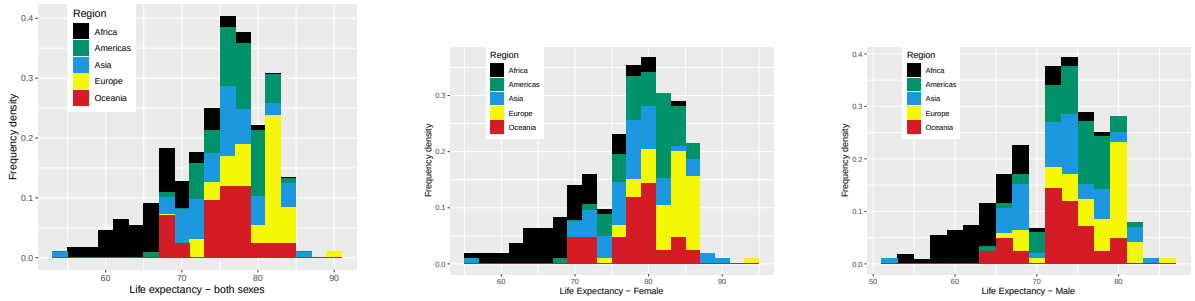
4 Statistical analysis

The following plots and statistical analysis are created using R (R Development Core Team, 2021) in version 4.1.0. The data for the year 2022 was analysed and the vari-

ables considered are subregion, region, year, under 5 mortality rates by sexes and life expectancy at birth by sexes.

4.1 Frequency distributions of the variables

To explore the frequency distribution of the variables, histograms of numeric variables are plotted using frequency density on the y-axis. The same color palette is used for all regions in all the plots. Figure 1(a) - (c) below shows the histogram of life expectancy based on sex. These histograms show that irrespective of the sex there are more countries with high life expectancy at birth. The life expectancy in most countries is between 70 to 85. Despite this, some countries in certain regions have less life expectancy than others in the same regions. The life expectancy of countries in Asia is distributed everywhere while that of Africans is mostly the lowest, countries in Europe have higher life expectancy in general. Countries in Oceania have life expectancy spread in between with Americans having almost similar distribution.



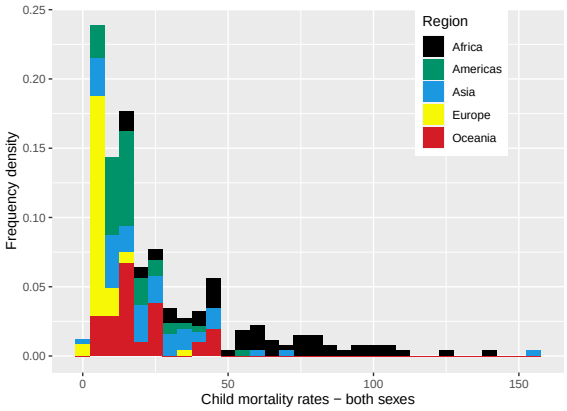
(a) Life expectancy at birth for both sexes

(b) Life expectancy at birth for female

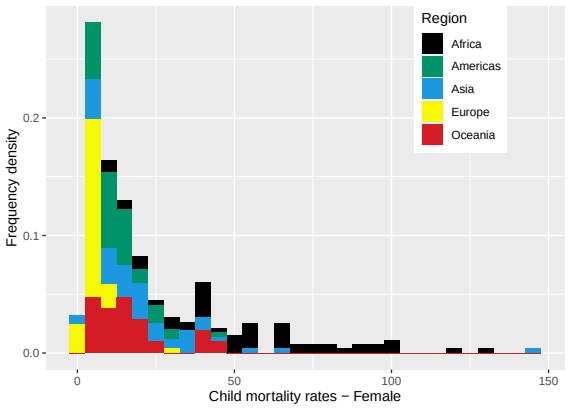
(c) Life expectancy at birth for male

Figure 1: Histograms of life expectancy at birth by gender

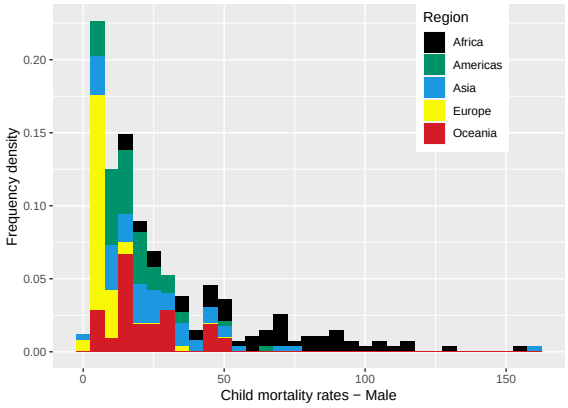
Figure 2(a) - (c) below shows the histogram of child mortality rate by gender. These histograms shows that irrespective of the sex there are more countries with low child mortality. The child mortality rates of most countries is between 0 to 50 in the charts. Despite this low numbers, some countries in certain regions have more child mortality than others in the same regions, say countries in Asia are distributed everywhere with countries here accounting for lowest and highest child mortality rates while that of Africans are mostly the highest, countries in Europe have lower child mortality rates in general. Countries in Oceania have child mortality spread in between with Americans having almost similar distribution.



(a) Child mortality rates for both sexes



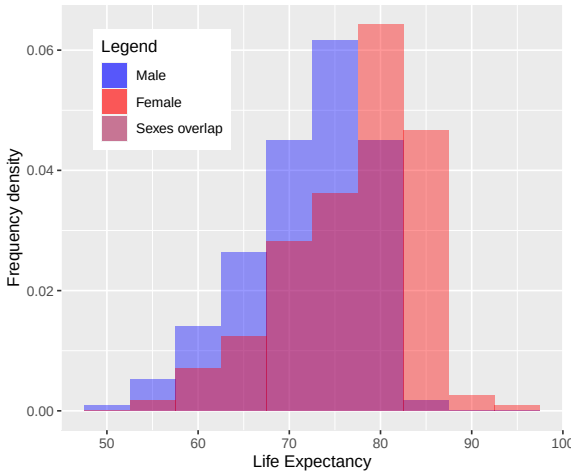
(b) Child mortality rates for female



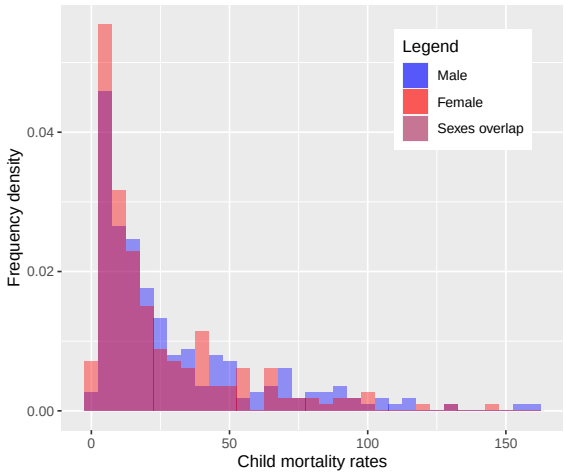
(c) Child mortality rates for male

Figure 2: Histograms of child mortality rates by sexes

The histogram in figure 3(a) shows the distribution of life expectancy of both male and female superimposed on each other. The purpose is to show how the distribution of life expectancy for male and female differs. It shows the data of males in color blue, and females in color red while the third color is where the data overlap and it is a color between blue and red. If the taller bin is male then the smaller bin that overlap represent female and also if the taller bin is female, the smaller bin that overlap is male.



(a) Histogram of life expectancy of male and female superimposed.



(b) Histogram of child mortality of male and female superimposed.

Figure 3: Histograms of life expectancy at birth and Child mortality by gender

The tallest bin for male is 75-78 years while the female is 78-85 years. The least life expectancy represented in the chart is between 45 and 53 years which is only represented by males. The highest life expectancy represented is between 93 and 98 years which is only represented by females. This shows that females have a higher life expectancy at birth than males in general for the year 2022.

However, the histogram in figure 3(b), shows the distribution of child mortality rates of both males and females superimposed on each other. The purpose is to show how the distribution of child mortality for males and females differs. It shows the same color as explained in paragraph above for figure 3(a). Females have the tallest bin indicating that there are more females with low child mortality rates than males. The highest mortality rate presented in the chart is blue which is male.

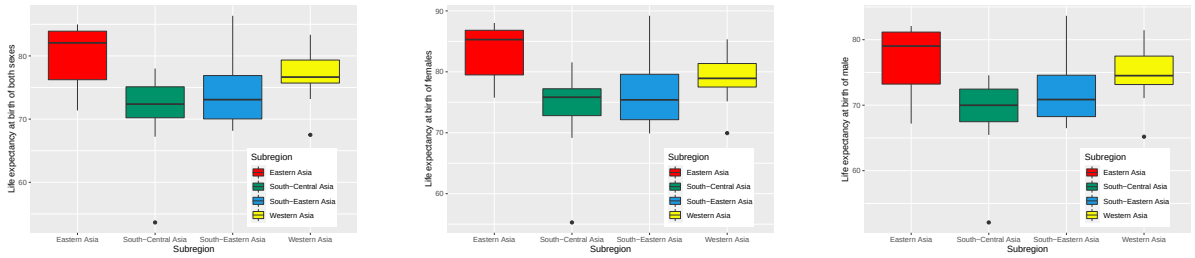
4.2 Exploring variability in Asia and its subregions

Box plots are used to visualize the variability in the 4 subregions in Asia which are Eastern, Western, South-central and South-eastern Asia. Table 1 shows the median values of all 4 subregions for child mortality rates and life expectancy. Figure 4(a) - (c) shows the box plot of life expectancy at birth by subregions in Asia, the charts in this figure are similar irrespective of gender.

Table 1: Median values for Subregions in Asia.

	LEBS	LEM	LEF	MRBS	MRM	MRF
Eastern Asia	82.065	79.025	85.290	5.005	5.240	4.755
South-Central Asia	72.375	69.995	75.835	33.650	34.255	31.815
South-Eastern Asia	73.080	70.860	75.400	23.550	26.270	20.680
Western Asia	76.650	74.510	78.930	16.070	17.220	13.650

Firstly, we look at the length of the box which is the interquantile range(IQR). The box of Western Asia is the shortest indicating that the life expectancy at birth of most countries in this subregion is similar, followed by South-central Asia. The box of Eastern Asia is the longest resulting in a large IQR which indicates that the data are more widespread within this subregion, this is also similar for South-Eastern Asia. Despite countries in Western and South-central Asia being almost homogenous there are countries with extreme life expectancy value that falls outside of the other countries in the same subregion. Comparing the median of the subregions which is the middle line in the box, The median of Eastern Asia is closer to its Q3 which shows that most of the valuse are large and it is also outside of all the other subregions, indicating that the values are different from all the other subregions. South-central and South-Eastern have close medians which shows that the values are similar in these subregions. Western Asia have the median close to Q1, that is, most of the values here are low and few countries in this subregion have higher life expectancy than the others in the same subregion.



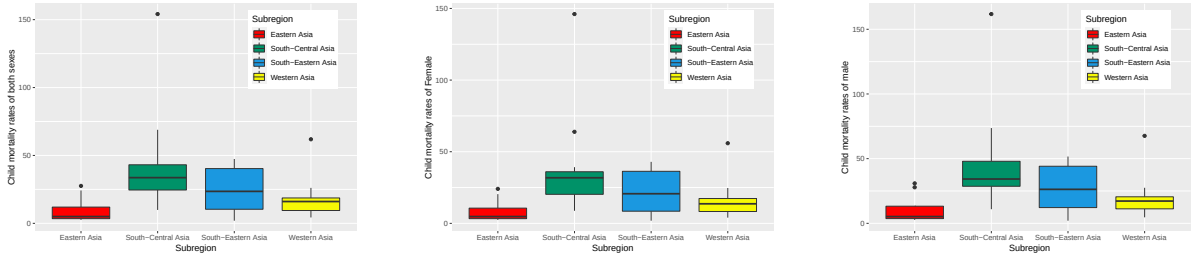
(a) Life expectancy at birth for both sexes

(b) Life expectancy at birth for female

(c) Life expectancy at birth for male

Figure 4: Box plots of life expectancy at birth by gender

In figure 5(a) - (c) below, the box plots for child mortality rates by gender are presented. The child mortality rate for Eastern Asia has the smallest IQR, indicating that the child mortality rates from countries in this subregion are mostly similar, followed by Western Asia.



(a) Child mortality rates for both sexes

(b) Child mortality rates for female

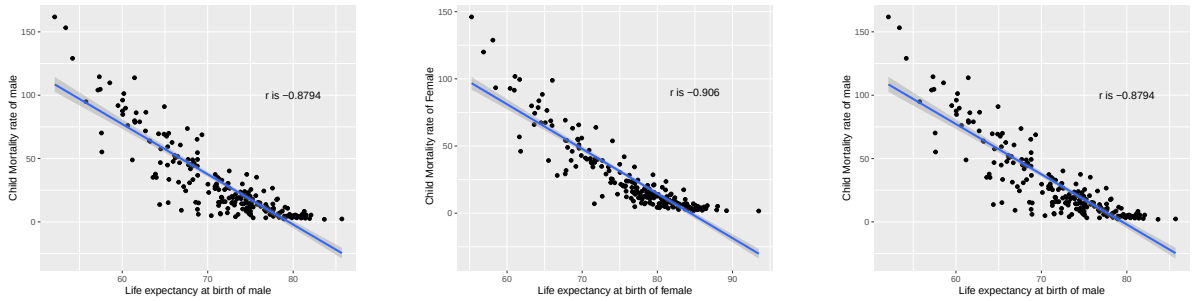
(c) Child mortality rates for male

Figure 5: Box plots of child mortality rates by sexes

The box for South-Eastern Asia is the longest, indicating the child mortality rates from countries in this subregion are widespread and they are not as similar as compared to other subregions in Asia. The median of Eastern Asia is outside the box of all other subregions, showing that the values are more different from every other subregion. Also, the median is very close to its Q1 which means most of the data are small and only a few values are spread. South-Eastern have values that are almost evenly spread, The median of South-central is inside the box of South-Eastern, indicating the values are similar. Also for South-Eastern and Western Asia but Western Asia and South-central values are not similar as the median are far apart from each other. Western and South-central Asia have extreme outliers indicating that there are countries in these subregions with very high child mortality rates. Considering the 5(b) and (c) which shows child mortality rates for females and males respectively, the median value for female is close to Q3 while that of male is closer to Q1, indicating that most of the values of child mortality rates or female in this subregion are more than that of males.

4.3 Correlation and Scatter plots

In this section, we examine the relationship that exists between the variables using the Pearson correlation coefficient and scatter plots. Table 2 summarizes the correlation coefficient between all the variables while Figure 6 shows the scatter plots between child mortality rates and the life expectancy of male, female and both sexes with their correlation coefficients.



(a) Child mortality rates of male versus Life expectancy at birth of male

(b) Child mortality rates of female versus Life expectancy at birth of female

(c) Child mortality rates of both sexes versus Life expectancy at biirth of both sexes

Figure 6: Child mortality rates versus Life expectancy at biirth

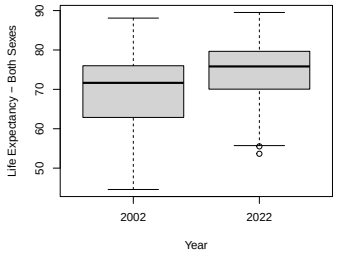
From the plots, there are high negative relationship between life expectancy and Child mortality rates irrespective of gender which indicates that as the values of one increase, the other decreases. From Table 2, there are high positive relationships between the sexes of each variables indicating that as one increases the other also increases.

Table 2: Table showing correlation coefficient between all the variables.

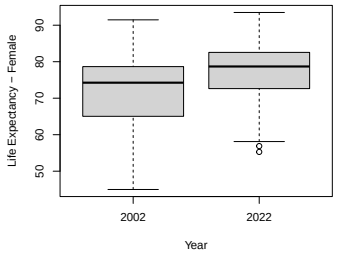
	LEBS	LEM	LEF	MRBS	MRM	MRF
LEBS	1.0000					
LEM	0.9926	1.0000				
LEF	0.9929	0.9712	1.0000			
MRBS	-0.8989	-0.8789	-0.9059	1.0000		
MRM	-0.8976	-0.8794	-0.9029	0.9985	1.0000	
MRF	-0.8970	-0.8749	-0.9060	0.9979	0.9929	1.0000

4.4 Change in the variables from 2002 to 2022

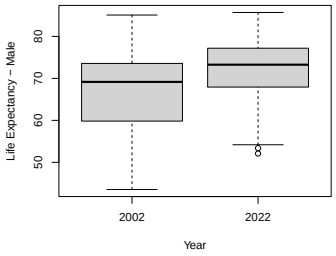
Figure 7 below shows the box plot of life expectancy at birth by gender for the years 2002 and 2022. The box plots look similar irrespective of gender. The box for the year 2022 indicates that life expectancy has increased for all gender and there is less variability in data compared to 2002. The world is moving towards achieving higher life expectancy at birth but despite the increment, there are countries with extremely low life expectancy at birth in the year 2022.



(a) Life expectancy at birth of both sexes



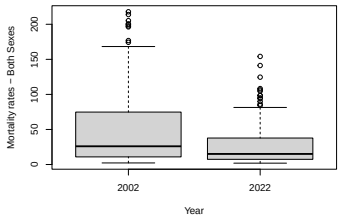
(b) Life expectancy at birth of female



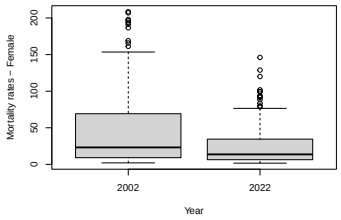
(c) Life expectancy at birth of male

Figure 7: Life expectancy at birth by gender, year 2002 and 2022

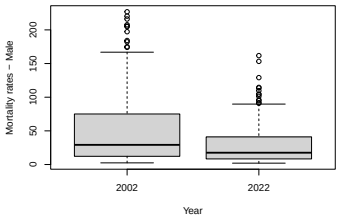
Figure 8 below shows the box plot of child mortality rates by gender for the years 2002 and 2022. The box plots look similar irrespective of gender. The box for the year 2022 indicates that the child mortality rate reduced for all gender and there is less variability in data compared to 2002. The world is moving towards achieving a lower child mortality rate but despite the reduction, there are countries with extreme child mortality rates in the year 2022 similar to the year 2002.



(a) Child mortality rate of both sexes



(b) Child mortality rate of female



(c) Child mortality rate of male

Figure 8: Child mortality rates by gender for year 2002 and 2022

5 Summary

The Data used for the analysis of this report was extracted from International Database (IDB, 2021), The U.S. Census Bureau contains various demographic data (currently from 1950 to 2060) on all states and regions of our world that are recognized by the US Department of State and have a population of 5000 or more. The sources of the

database are information from state institutions, such as censuses, surveys or administrative records, as well as estimates and projections by the U.S. Census Bureau itself. It includes life expectancy at birth and under 5 child mortality rates (child mortality) of 227 countries from the year 2002 and 2022. The countries are divided geographically into 5 regions and 21 subregions. The aim of the report was to perform explanatory data analysis on each of the variables to understand the distribution of the variables, how similar are the subregions, the relationship that exists between the variables and how the data had changed from the year 2002 to 2022.

Histogram, box plot, Scatter plot, and correlation coefficient were used which are presented and interpreted in figures. Examining the frequency distribution of life expectancy using histogram, the life expectancy of a lot of countries is high irrespective of the sexes, the life expectancy of most countries is between 70-85. Some countries in certain regions have less life expectancy than others in the same regions. When comparing the distribution by sexes, females have high life expectancy than males. Histogram was also used to analyse the frequency distribution of child mortality, the child mortality rates of most countries are between 0 and 50 but countries in some regions have more child mortality rates than others in the same regions. The highest and lowest child mortality rates are from countries in Asia, comparing child mortality rates by sex, females have lower child mortality rates than males.

Box plots and median were used to analyse the similarities and differences in the 4 subregions of the Asia region. Box plots of life expectancy at birth show that values from Eastern Asia are widespread and not homogenous within the subregion, this is an almost similar case with South-eastern Asia. Despite values in Western and South-central Asia being almost homogenous within subregions, there are extreme life expectancy values that are outside the subregions. From the median values, Eastern Asia is heterogenous with all other subregions in the regions. South-central and South-Eastern have more similar values. Analysing the Child mortality rates, values from Eastern Asia are more homogenous within the subregion followed by Western Asia. South-Eastern Asia values are more heterogenous within their subregion. From the median values, Eastern Asia is different from all other subregions in the regions while South-central and South-Eastern have more similar values but Western Asia and South-central values are not similar.

Scatter plots and correlations are used to examine the type of relationships that exist between the variables. There is a strong negative relationship between life expectancy

and child mortality rates irrespective of gender which means as life expectancy increases, child mortality reduces and vice versa.

Box plot was also used to visualize how the data had changed from the year 2002 to 2022, which shows that the child mortality rates for a lot of countries have reduced over the years while life expectancy has increased.

For further exploration of the data, the bar chart for each variable can be plotted to visualize the variables by regions and subregions and test for the significance of the Correlation Coefficient since we are using a fraction of the population data.

Bibliography

Bluman, A. G. (2012). *Elementary statistics : a step by step approach*. The McGraw-Hill Companies, Inc., 1221 Avenue of the Americas, New York, NY 10020, eightht edition.

Glossary (2021). United state census bureau glossary. *United state Census Bureau glossary*. URL:https://www.census.gov/programs-surveys/international-programs/about/glossary.html (visited on 10th of May 2023).

Hay-Jahans, C. (2019). *R Companion to Elementary Applied Statistics*. CRC Press Taylor Francis Group, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742.

IDB (2021). The international database. *Demographic Data*. URL: https://www.census.gov/programs-surveys/international-programs/about/idb.html (visited on 3rd of May 2023).

Max Roser, E. O.-O. and Ritchie, H. (2013). Life expectancy. *Our World in Data*. URL: https://ourworldindata.org/life-expectancy (visited on 10th May 2023).

R Development Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roser, M., Ritchie, H., and Dadonaite, B. (2013). Child and infant mortality. *Our World in Data*. https://ourworldindata.org/child-mortality.