1 Introduction

The younger people under age 5 mortality rate is an important parameter to know the total mortality and health status in every country in the world. There are several reasons this can happen. Health condition is one of the factors. Moreover, the social and economic disparities within the country also cause under age 5 mortality. This project focuses on the analysis of the under age 5 mortality rate for both sexes and life expectancy at birth for both sexes in countries around the world. The dataset of this project is a collection of values of life expectancy at birth and under age 5 morality rates in 2002 and 2022 from various regions in 227 countries. The purpose of this assessment is to understand the trends in life expectancy at birth and under age 5 mortality rates. Furthermore, the difference in life expectancy between males and females is also analysed. The difference in under age 5 mortality rate and life expectancy at birth within and between subregions are compared to see the variability of the values. Another goal is to analyse the changes of life expectancy and under age 5 mortality rates over time. The second chapter describes the dataset and the goals of this project.

In the third section, the calculation methods used for analysing the dataset, such as mean, variance, median, and correlation coefficient are explained. The visualization methods such as histogram, boxplots, scatterplots are illustrated. In the fourth section, graphs are drawn based on the methods introduced in the previous section for a more detailed analysis. The frequency distributions between life expectancy for both sexes and under age 5 mortality rate for both sexes are described using histogram. Additionally, the differences in life expectancy between the sexes is visualized with histograms. Boxplots are used to see the variability within and between subregions. Pearson correlation is used to see the relationship between variables life expectancy at birth and under age 5 mortality rate. Scatterplots are used to describe the changing between years 2002 and 2022. The last chapter is about summaries, the conclusion, and the possibility for further analysis that can be done using this dataset.

2 Problem statement

2.1 Description of the dataset

This project uses a dataset from International Data Base (IDB) of the US Census Bureau. The dataset provided for this report covers the under age 5 mortality rate and life expectancy in 227 countries, divided into 5 regions and 21 subregions for the years 2002 and 2022. The data was collected by surveys and census. The under age 5 mortality rate is the number of young children that die before they reach their first year of age expressed as a rate in every 1,000 live births (Bureau, 2021). The variables country, region, and subregion are categorical. Year is a binary variable because there are only two years, 2002 and 2022. Life expectancy at birth and under 5 age mortality are continuous variables that can have all consecutive real values (Forsyth, 2018). There are 4 missing values in countries which are CuraÃ§ao and CÃ´te d'Ivoire in year 2002 and 2022 in column region and subregion. There are another 6 missing values in countries which are Libya, Puerto Rico, South Sudan, Sudan, Syria, United States in year 2002. Imputing missing values are necessary to analysed the data (Hastie and Tibshirani, 2008).

2.2 Project goals

The goal of this project is to analyse the life expectancy and under 5 age mortality rate using several statistical methods and visualize the result in a graph. In addition, the life expectancy between males and females can be illustrated as a graph and the variability of the value between subregions can be compared. The main purpose is to investigate the correlation between the under age 5 mortality rate for both sexes and life expectancy at birth for both sexes. By visualizing the variables between years, the changes between 2002 and 2022 can be showed.

3 Statistical methods

In this chapter, mathematical methods, graphical methods and correlation coefficients are explained. R is the programming language that is used for this project and R studio is the software editor that is also used for applying mathematical methods and creating a visualization of the result (R Development Core Team, 2020). R provides packages that include a generous amount of libraries that can be downloaded for the analysis and data plotting. Used packages are dplyr, xtable, ggplot, lessR, readr (Grolemund and Wickham, 2017).

3.1 Calculation methods

Mean

In mathematical expressions, the arithmetic mean is also known as average and is measured to calculate the central tendency (Forsyth, 2018), where the variable size of data is n and consists of values $x_i, .., x_n$. The formula to calculate the mean is :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

It is easy to calculate, but it is highly affected by extreme values and very small value.

Variance and Standard deviation

Variance is one way to measure dispersion that takes the spread of all data points in a dataset (Forsyth, 2018). Normally, the measuring of dispersion is used alongside standard deviation, which is the square root of variance. Standard deviation is used to measure how far the number of observations differs from the mean value (Jahans, 2019). The variance formula is:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Median

The median is the value that divides the data into equal parts and represents the middle value of the dataset (Forsyth, 2018). By sorting the data and finding the middle value in a set of data the median, can be obtained. The data in the middle of a dataset may contain more than a single piece of data. This happens when the dataset is divided into two groups with the exact same length, then there are still two data points left in the middle of the dataset. When this occurs, the median can be obtained by finding the average between those two data points.

Quartile

The quartile is quite identical to the median. The median splits the data into two equal parts while the quartile partitions the data according to the percentage of proportions. The quartile divides the data into four equal values after the data is sorted, starting from the smallest value to the largest one (Forsyth, 2018). The first quartile, denoted as Q1, is the value located at the cumulative 25% rate which is the $\frac{1}{4}(n+1)$ th value. The second quartile, denoted as Q2, is the $\frac{2}{4}(n+1)$ th value, and the third quartile or upper quartile is 75% which is $\frac{3}{4}(n+1)$ th value. It denoted as Q3.

3.2 Visualization methods

Histogram

Histograms are used to visualize the frequency distribution of a continuos variable (Jahans, 2019). Every bar in histogram describes the range of value which is called bins. The height of the bar shows the frequency of the corresponding range. The difference between the smallest value x_{min} and the largest value x_{max} of a variable is divided into n intervals: $(x_{max} - x_{min})/n$. A set of bars is built, one per interval. The height of each bar is the number of data items in the corresponding interval.

Boxplot

Boxplot is also called box whisker plot and is used to analyse the spread of the data. The boxplot is a graph that illustrates the summary of five values from the dataset (Jahans, 2019). Figure 1 shows an example of boxplot.

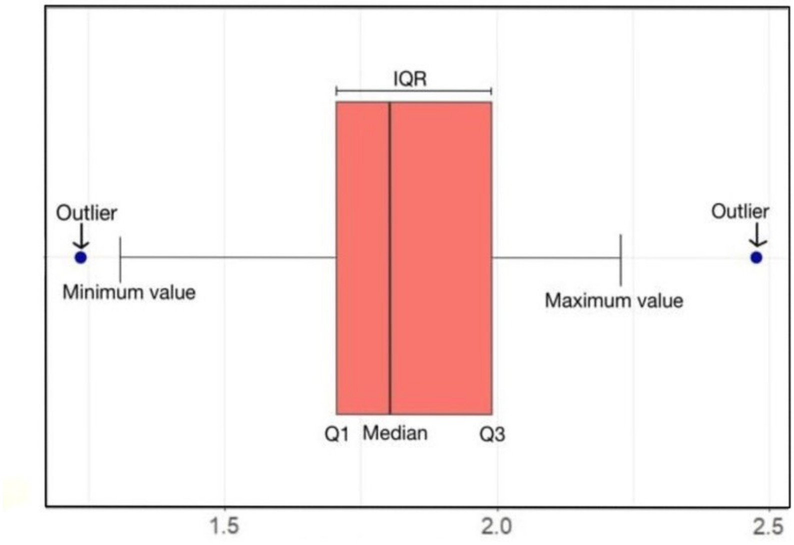


Figure 1: Example of boxplot

The summary of five values are the minimum value, the first quartile (Q1), the second quartile (Q2) which is equal to the median, the third quartile (Q3) and the maximum value when the values are arranged in order of size.The interquartile range (IQR) is a way to measure the spread. The interquartile range (IQR) is the difference between the first quartile (Q1) and the third quartile (Q3) or Q3-Q1. In addition, boxplots also indicate the range of extreme values from the dataset. When the value is larger than Q3 + 1.5(IQR) or the value is less than Q1 - 1.5(IQR) then it can be referred to as an extreme value or outlier. The whiskers are two lines outside the box extend to minimum or maximum values within Q3+1.5(IQR) or Q1-1.5(IQR).

Scatterplot

Dots in scatterplot represent the value of each data point. The scatterplot is used to observe the relationship between two variables, in which the relationship may vary (Jahans, 2019). It can be a positive, negative, linear or non linear relationship. The relationship between two variables can be observed by plotting the values of a variable y against the respective values of a variable x. Positive relationship is the condition when the x values are increases and values in y also increases. On the other hand, negative correlation is the condition when y values tends to decrease and x values increase.

3.3 Correlation coefficients

A correlation coefficient determines the quality of correlation between two variables. There are ways to measure a correlation coefficient and Pearson's correlation is very well known among others, although compatibility is the main reason to use Pearson's correlations. The range of correlation coefficient values is between -1.0 and 1.0. Positive correlation refers to relationship in which the value of one variable increase as the value of other variable also increases. On contrary, if the value is negative, it refers to a relationship where the value of one variable increase if the value of the other variable decreases. If the correlation coefficient is 0 means there is no linear relationship between two variables (Kronthaler and Zöllner, 2018). The Pearson's correlation formula :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

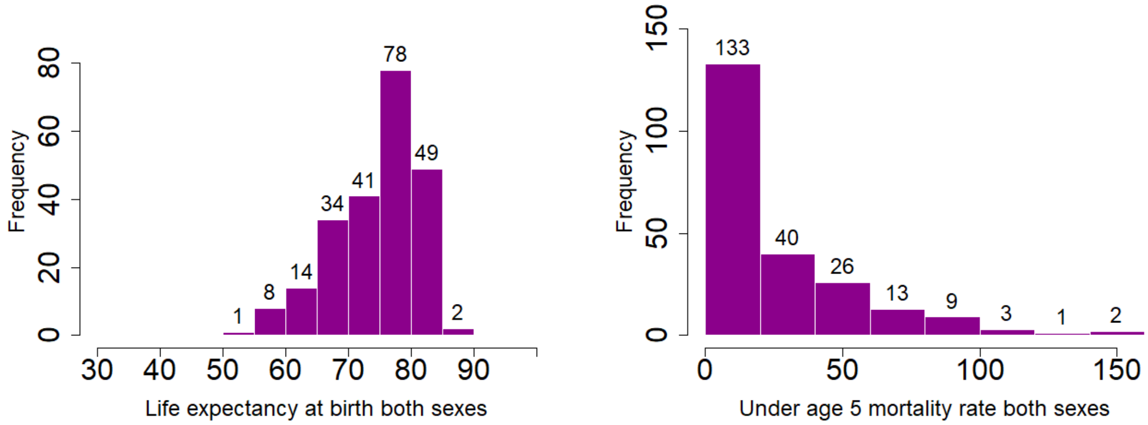
where $\bar{x}$ and $\bar{y}$ are corresponding means of each variables.

4 Statistical analysis

This chapter covers the end result graphs from a calculation using mathematical methods in the previous chapter. It also covers the Pearson correlation, which describes the relationship between under age 5 mortality rate and life expectancy. The table and several graphs are shown in the appendix A and B.

4.1 Frequency distribution

There are 10 missing values that are already mentioned in the previous chapter. Before doing the calculation methods, are necessary to replace the missing value. There are 2 Country missing the region and subregion in 2002 and 2022 that need to be fulfilled manually. Moreover, the other missing value is replaced using median. Since it suits the variables with skewed and extreme values. Due to a large number of different values, in comparison to other types of charts, a histogram seems the more appropriate choice to visualize the frequency distribution. Figure 2 shows frequency distributions, respectively Figure 2(a) shows the frequency distribution of life expectancy at birth for both sexes and Figure 2(b) describes the frequency distribution of under age 5 mortality rate for both sexes.



(a) Life expectancy at birth for both sexes

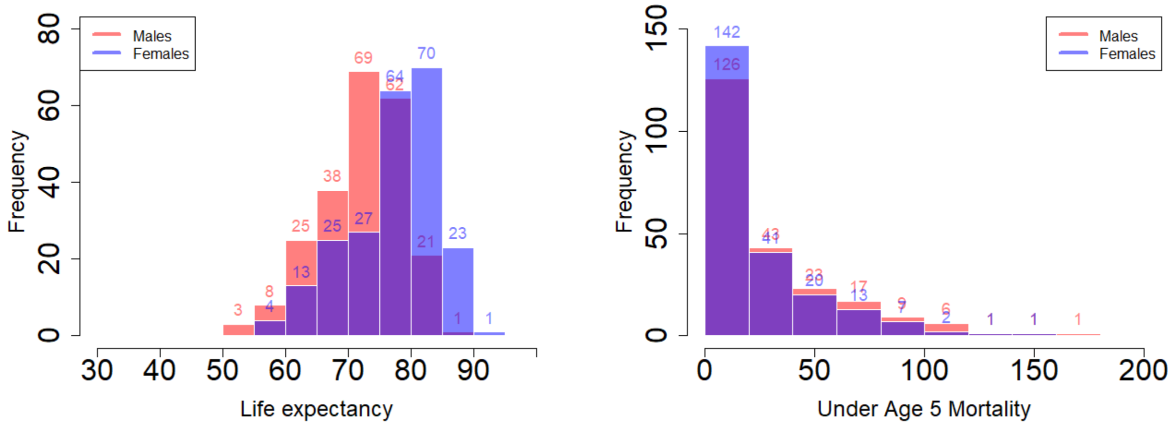
(b) Under age 5 mortality rate for both sexes

Figure 2: Frequency distributions

There are 227 observations that are used on both histograms, with X-axis for life expectancy for both sexes in Figure 2(a) and under age 5 mortality rate for both sexes in Figure 2(b). The Y-axis for Figures 2(a) and 2(b) are absolute frequency. Figure 2(a) shows the left-skewed distribution or negative distribution while the mean value is 74.58. The maximum value of the life expectancy at birth for both sexes is occupied by Western Europe, with 89.52, whilst South-Central Asia held the lowest value of life expectancy at birth for both sexes with 53.65. On the contrary, Figure 2(b) shows the right-skewed distribution or positive distribution with the mean value of under age 5 mortality rate for both sexes 26.67. The maximum value of life expectancy at birth for both sexes is 154.13 held by South Central Asia and the minimum held by South Eastern Asia with 1.94. All the values represented in those two figures can be seen in the Appendix B tables 1 and tables 2. The histograms of frequency distribution for life expectancy for both males and females are also shown in Appendix A Figure 8. Besides that, histograms for frequency distribution fo under age 5 mortality rate for both males and females are also shown in Appendix A Figure 9. The mean, maximum, and minimum values for the life expectancy of males and females are also included in Appendix B table 1.

Difference between sexes

Figure 3 below describes the difference in life expectancy and under age 5 mortality rate between the sexes. Figure 3(a) is life expectancy difference between sexes and Figure 3(b) is under 5 mortality rate between sexes.



(a) Life expectancy for both sexes

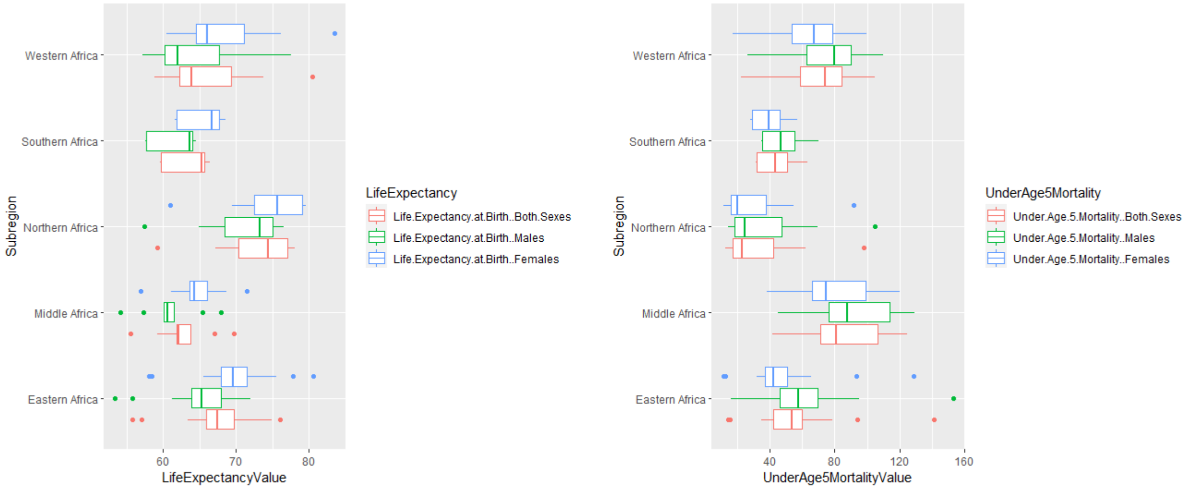
(b) Under age 5 mortality rate for both sexes

Figure 3: Difference in life expectancy and under age 5 mortality rate between sexes

From Figure 3 most of values are positive which means the life expectancy of females are higher than life expectancy of males. Moreover, the value of under age 5 mortality rate tends to be higher for females than for males.

4.2 Homogeinity within subregions

Boxplots of life expectancy for both sexes and under age 5 mortality rate for both sexes per subregion and only for region Africa are shown in Figure 4(a) and Figure 4(b).



(a) Life expectancy at birth

(b) Under age 5 mortality rate

Figure 4: Boxplots by subregion

The IQRs are visualize in Figure 4 and it can indeed be determine to distinguish heterogeneous and homogeneous behaviour. The length of the quartile gives a hint which one is homogeneous and which one is heterogeneous. If the box is long, then it is classified as heterogeneous. In contrast, when the length is short then it is homogeneous. Figure 4(a) indicates that Western Africa in all three life expectancy for both sexes, males, and females is more heterogeneous in comparison to other subregions with 7.11, 7,5 and 6.5 respectively, while Middle Africa is homogeneous with 1.87 for life expectancy at birth for both sexes, 1.4 for males, and 2.4 for females. In Figure 4(b), the value of under age 5 mortality rate for both sexes is shown. Based on the boxplots, it can be seen that Southern Africa are homogeneous. However, Middle Africa is heterogeneous with IQR value of under age 5 mortality rate for both sexes 35.31, 37.63 for males and 32.91 for females. The values that mentioned above can be seen in Appendix B table 4 and table 5.

4.3 Correlation coefficients

There are several ways to calculate the correlation coefficient. According to the data that are given, Pearson seems suitable to calculate the coefficient correlation between the variables. Figure 5 shows the correlation between under age 5 mortality rate and life expectancy at birth. 'LE Both' in scatterplots implies the life expectancy at birth for both sexes. Moreover, 'LE Males' and 'LE Females' both indicate the life expectancy at birth for each of their own. Therefore, 'MR Both' implies to under age 5 mortality rate for both sexes. 'MR Males' and 'MR Females' are implies to under age 5 mortality rate for males and females.

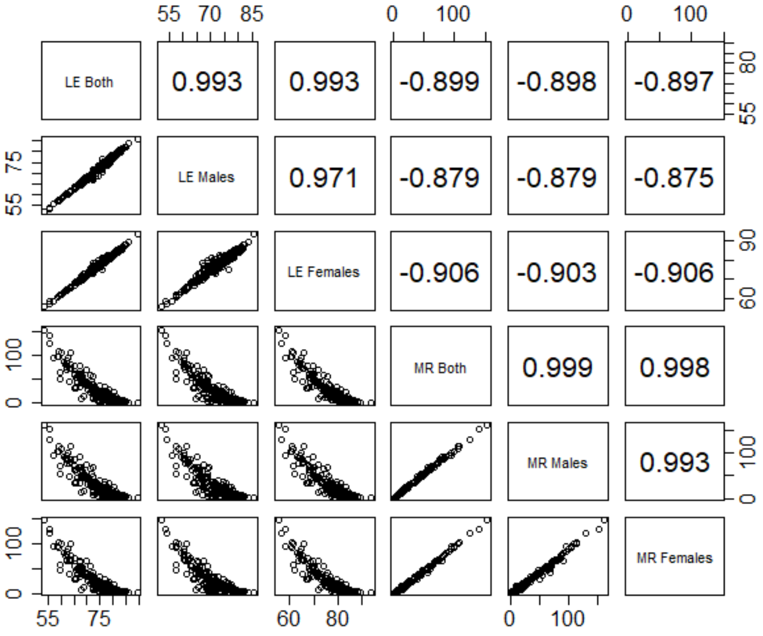
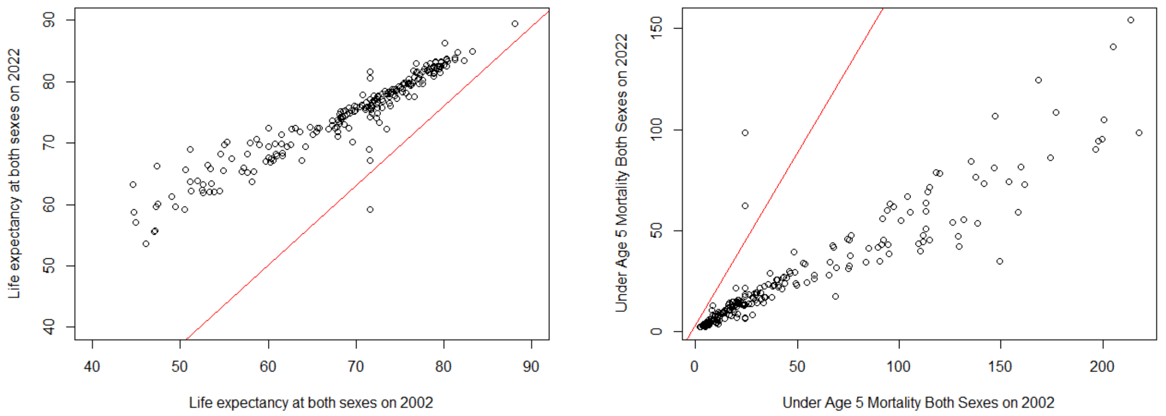


Figure 5: Correlation between the variables

The scatterplots are used to ease the visualization of the two variables. The scatterplots matrix shows the plots and correlation coefficient number between life expectancy at birth and under age 5 mortality rate. The correlation coefficient between life expectancy at birth for both sexes, for males and for females and under age 5 mortality rate for both sexes, for males and for females are negative correlation. On the other hand, correlation between life expectancy at birth for both sexes and for males and females are positive correlation and also between under age 5 mortality rate for both sexes and for males and for females. The correlation value between life expectancy at birth for both sexes and life expectancy at birth for males and females are 0.993 and 0.993 respectively. The correlation value between under age 5 mortality rate for both sexes and under age 5 mortality rate for males and females are 0.999 and 0.993 respectively. Moreover, the correlation between life expectancy at birth for both sexes and under age 5 mortality rate for both sexes is negative correlation which is -0.899.

4.4 Comparison between years

The scatterplots are used to compare the values between years. Figure 6(a) shows the difference between years of life expectancy for both sexes and Figure 6(b) shows the difference between years of under age 5 mortality rate for both sexes.

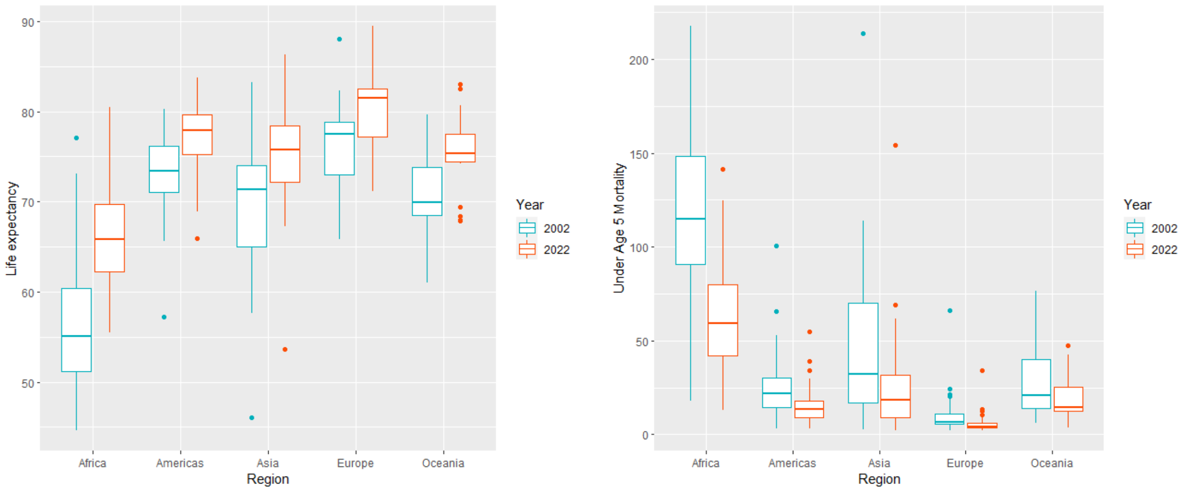


(a) Life expectancy at birth for both sexes

(b) Under age 5 mortality rate for both sexes

Figure 6: Difference between years of 2002 and 2022

Looking at life expectancy for both sexes in Figure 6(a) the data points of most countries are above the line which means the values are increasing in most countries. According to under age 5 mortality rate for both sexes in Figure 6(b) the values are decreasing in most countries.



(a) Life expectancy at birth for both sexes

(b) Under age 5 mortality rate for both sexes

Figure 7: Boxplot difference between years of 2002 and 2022 per region

Figure 7 shows the boxplot of differences between year 2002 and 2022 for life expectancy at birth for both sexes and under age 5 mortality rate for both sexes. In Figure 7(a) can be seen the life expectancy at birth for both sexes in each region is obviously increasing. But in Oceania, the life expectancy there is slightly increasing. On the other hand, in Figure 7(b) under age 5 mortality rate for both sexes are decreasing. Especially in Africa, it can be seen the under age 5 mortality rate from 2002 and 2022 is significantly decrease.

5 Summary

In this project, the dataset originated from The International Data Base of the U.S Census Bureau was analysed. The dataset has observations of 227 countries, 5 regions and 21 subregions. Each country has variables life expectancy at birth for both sexes, males and females also variables under age 5 mortality rate for both sexes, males and females respectively in year 2002 and 2022. The main purpose of this project was to analyse the correlation between the under age 5 mortality rate for both sexes rate and life expectancy at birth for both sexes. Comparing the under age 5 mortality rate for both sexes and life expectancy at birth for both sexes to see the variability were also illustrated. Another goal was to see the changes over the time. The data was illustrated using histograms, boxplots and scatterplots. Based on the mean value, most countries were centered at 74.58 for life expectancy at birth for both sexes and 26.68 under age 5 mortality rate for both sexes. Based on the difference between sexes, it represented that life expectancy at birth for females is higher than males. The correlation between life expectancy for both sexes and under age 5 mortality rate for both sexes was a negative relationship which is -0.899. The correlation between those variables turned out to be quite influential. To see the variability within and between subregions for region Africa, boxplots are used. Notable result was that life expectancy at birth for both sexes, males and females in Middle Africa is homogeneous. On contrary the result for under age 5 mortality rate for both sexes, males and females in Middle Africa was heterogeneous. From the analysis of the difference in life expectancy at birth between 2002 and 2022, it was found that life expectancy at birth for both sexes are increasing in each subregion, especially in Africa. On the other hand, the under age 5 mortality rate was decreasing from 2002 to 2022 especially in Africa was significantly decreasing. In future work, other influences on economic, alongside health conditions can be researched. To summarize, extensive work has been done on exploring, analysing and applying mathematical methods.

Bibliography

Bureau, U. C. (2021), ‘International data base (idb)’. visited on 2012-10-18.

Forsyth, D. (2018), *Probability and Statistics for Computer Science*, Springer, Urbana.

Grolemund, G. and Wickham, H. (2017), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.*, O’Reilly Media.

Hastie, T. and Tibshirani, R. (2008), *The Elements of Statistical Learning*, Springer, California.

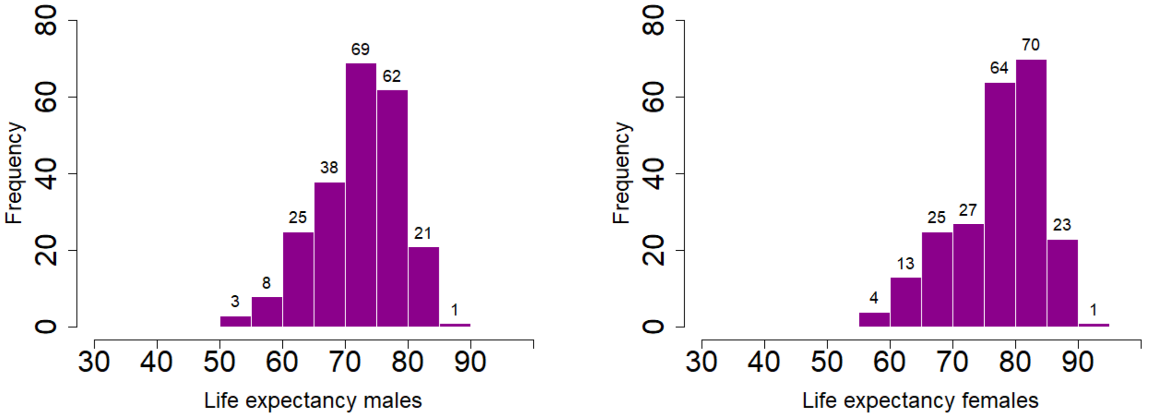
Jahans, C. H. (2019), *R Companion to Elemntary Applied Statistics*, Taylor and Francis Group, New York.

Kronthaler, F. and Zöllner, S. (2018), *Data Analysis with RStudio*, Springer, Chur.

R Development Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Appendix

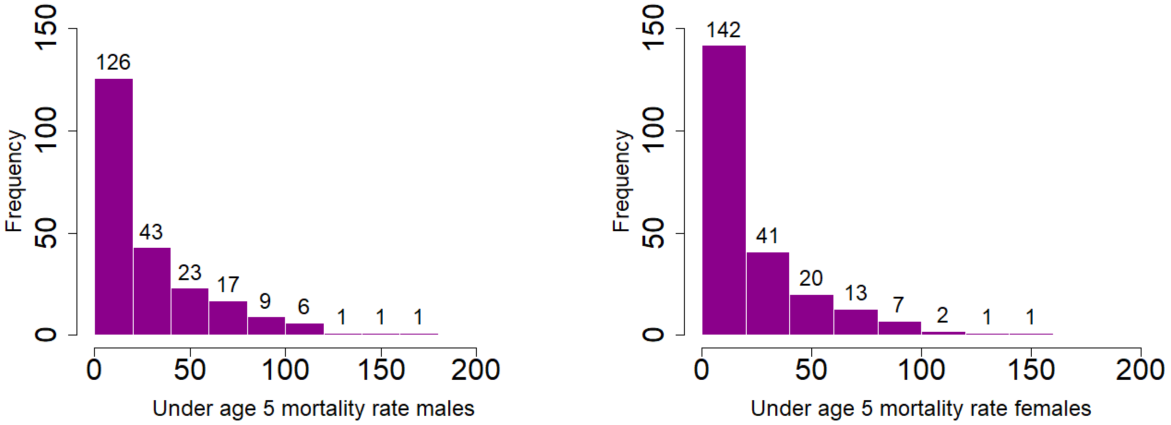
A Additional figures



(a) Life expectancy at birth for males

(b) Life expectancy at birth for females

Figure 8: Frequently distributions of life expectancy



(a) Under age 5 mortality rate for males

(b) Under age 5 mortality rate for females

Figure 9: Frequently distributions of under age 5 mortality rate

B Additional tables

	Life Expectancy both sexes	LE Males	LE Females
Minimum	53.65	52.10	55.28
Q1	70.05	67.93	72.63
Median	75.82	73.26	78.69
Mean	74.58	72.10	77.18
Q3	79.66	77.19	82.56
Maximum	89.52	85.70	93.49
Variance	46.7769	44.5409	50.8327

Table 1: Summary for frequency distributions of life expectancy

	UA5 mortality rate both sexes	UA5MR Males	UA5MR Females
Minimum	1.940	2.03	1.640
Q1	7.415	8.32	6.345
Median	15.080	17.55	13.620
Mean	26.677	29.23	24.012
Q3	37.775	41.02	34.470
Maximum	154.130	161.78	146.090
Variance	789.2912	905.1974	683.3034

Table 2: Summary for frequency distributions of under age 5 mortality rate

	LE both	LE males	LE females	UA5MR both	UA5MR males	UA5MR females
LE both	1					
LE males	0.9926	1				
LE females	0.9929	0.971	1			
UA5MR both	-0.8989	-0.8789	-0.9059	1		
UA5MR males	-0.8976	-0.8794	-0.9029	0.9985	1	
UA5MR females	-0.897	-0.8749	-0.906	0.9979	0.9929	1

Table 3: Correlation coefficient

Subregions	LE both	LE males	LE females
	IQR	IQR	IQR
Western Africa	7.11	7.51	6.57
Southern Africa	5.95	6.37	5.93
Northern Africa	6.60	6.60	6.60
Middle Africa	1.87	1.46	2.42
Eastern Africa	3.84	4.09	3.47

Table 4: Life expectancy for both sexes in Africa by subregions

Subregions	UA5MR both	UA5MR males	UA5MR females
	IQR	IQR	IQR
Western Africa	26.38	27.09	24.81
Southern Africa	18.45	20.03	16.83
Northern Africa	25.33	28.93	22.06
Middle Africa	35.31	37.63	32.91
Eastern Africa	17.74	23.20	13.76

Table 5: Under age 5 mortality rate for both sexes in Africa by subregions