TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

July 27, 2023

Contents

1 Introduction

The International Data Base (IDB) includes demographic measures across more than 200 countries. The detailed database is updated on an ongoing basis with the latest findings. These findings are then used for decision-making and strategic planning. In addition, the IDB also facilitates the assessment of the demographic impacts of worldwide events on individuals. With the rapid growth of the global population, the threat of overpopulation poses a severe threat. Therefore, numerous nations have also proposed population control strategies to resolve this problem.

The goal of the report is to analyze the data set to obtain important demographic insights. The dataset consists of life expectancy and mortality rates under age 5 for 227 countries with two distinct years. The report examines descriptive studies on the frequency distribution, measures of central tendency and dispersion of essential variables such as life expectancy and mortality rates for children under age 5 for both sexes and separately. The report highlights the differences between genders and regions. Furthermore, the report explores the heterogeneity and homogeneity of subregions between and within regions respectively. The report also discusses relationship between two different variables by analyzing the bivariate correlation coefficient values. In the end, it concludes by examining the changes in life expectancy and child mortality rates for both sexes, males and females from 2002 to 2022 (U.S. Census Bureau, 2022).

Section 2 analyzes the dataset and its corresponding characteristics. In Section 3, the dataset is subjected to various statistical methods such as central tendency and spread measures, Pearson's correlation, histograms, scatterplots, and boxplots. Section 4 elaborates on the analysis with relevant data visualization. Finally, Section 5 evaluates the primary findings of the study and identifies potential topics for future research.

2 Problem Statement

2.1 Dataset Description and Data Quality

The dataset analyzed in this study is a small sample of the International Data Base (IDB) provided by the US Census Bureau. It contains information on life expectancy for males and females at birth and children mortality rate under age 5 across 227 countries between 2002 and 2022. A closer inspection reveals that these countries are classified geographically into five regions and twenty-one subregions where each country has two observations from 2002 to 2022. The dataset containing a total of 454 observations initially consists of 11 variables including X, country, subregion, region, year and life expectancy at birth for both sexes, males, and females individually and under age 5 child mortality rates for both sexes, males and females separately. The definition of life expectancy is the average lifespan of a group of people born in the same year are expected to live. In the case of child mortality, it is defined as the number of children under age 5 who pass away within the initial year of their existence among a group of 1000 babies. Here, the country variable refers to a name of a country. Moreover, the region is a collection of subregions consisting of different countries. We have data for 2002 and 2022 throughout the whole dataset for each country (U.S. Census Bureau, 2022).

From the eleven variables, X is used as an index. Countries, years, regions, and subregions are categorical variables. The variables measuring child mortality rates and life expectancy values for males and females under age 5 differ in their degree of precision but are all continuous. These continuous values of the quantitative variables from the dataset contain positive integers, one-decimal and two-decimal point values. The dataset needs to include data for some subregions and regions. The entire dataset has 44 missing values, whereas each region and subregion has four missing values individually. The remaining 36 missing data exist for life expectancy and under age five mortality variables. In addition, the study has addressed this by adding observations with missing values in subregion columns and performing a mean imputation on several numerical variables based on subregions of each region. These techniques ensure a robust approach to handling missing data and preserving data integrity for statistical analysis.

To conclude, the quality of the data appears to be satisfactory given that the dataset originates from IDB. This study ensures the accuracy of its results by employing effective statistical methods (U.S. Census Bureau, 2022).

2.2 Project Objective

This study examines four fundamental analytical tasks related to the research questions. We provide first a comprehensive summary of the variables included in our dataset. Next, we use a histogram to compare male and female life expectancy in 2022. We then use boxplots to show differences between all regions. The histograms and boxplots showcase the frequency distribution of values and summarise all the numerical variables. To evaluate the bivariate correlation values in scatterplots, we employ the widely-used Pearson correlation coefficient. Furthermore, we again use boxplots to evaluate the homogeneity and heterogeneity of subregions. Finally, since all continuous variables have changed between 2002 and 2022, we employ scatterplots to visualize the patterns and relationships between them.

3 Statistical Methods

This section encompasses a range of statistical methodologies and visual depictions that are employed for analytical purposes. The statistical software package utilised in this study is R, version 4.0.5 (R Core Team, 2020). The present study employs R packages along with the ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017), cowplot (Wilke, 2020), reshape2 (Wickham, 2007) and tidyverse (Wickham and et al., 2019). This section presents statistical methods for analysing the research question.

3.1 Mesures of Central Tendancy

3.1.1 Mean

Mean is the arithmetic average in which the total number of observations $x_1, x_2, \cdots, x_n$ of a variable is divided by the sum of all values. It is expressed as $\sum_{i=1}^{n} x_i$, where i represents each value in the sample. The arithmetic mean is defined by,

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

where n is the total number of observations, the mean is sensitive to their presence when there are extreme values. The values significantly differ from the rest of the data. Extreme values can substantially affect the mean since it considers all values in the calculation (Hay-Jahans, 2019, pp. 73-74).

3.1.2 Median

The median is an important measure of central tendency that helps us understand the central location of data particularly when we deal with extreme values. To calculate the median, we sort the sample data in either ascending or descending order and then we identify the median value in the sorted data. If the sample size n is odd, then the middle of the data set is the median value. On the other hand, if the sample size is even, we define the median as the average of the two middle values. We can define the median by,

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & , \text{ if odd } n \\ \frac{[x_{(n/2)} + x_{((n/2)+1)}]}{2} & , \text{ if even } n \end{cases}$$

where the set of observations in the sorted set $(x_1, x_2, \cdots, x_n)$ is defined by n. In summary, the median divides the dataset into two halves, with half of the observations above it and the other half below it (Hay-Jahans, 2019, pp. 75-76).

3.2 Measures of Spread

3.2.1 Variance and Standard Deviation

Variance measures dispersion to determine whether data values are densely packed or widely dispersed. Consider an example of numeric data, $x_1, x_2, ..., x_n$, where n is the number of observations. Here, the variance of the sample, denoted by s^2, is the sum of the squared difference from the mean divided by $n - 1$. The following formula for the variance can be defined by using sample data to estimate the variance of the population,

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

It is the mathematical sum of the squares of the deviation from the mean. Conceptually, the variance indicates the population's diversity.

The sample standard deviation s is determined as the square root of the variance using the formula $s = \sqrt{s^2}$. Extreme values have an effect on the variance as well as the standard deviation (Hay-Jahans, 2019, pp. 76-77).

3.3 Correlation

3.3.1 Pearson's Correlation

The Pearson correlation coefficient is a linear relationship between two continuous random variables X and Y. It also measures the strength and nature of linear relationships.

It is defined by ρ and shows how a linear relationship is set up. The correlation coefficient ρ is defined as,

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \ ,$$

where the covariance of X and Y is represented by σ_{XY} and the variances of X and Y are represented by σ_X and σ_Y, respectively. The following methods assist in producing a fair estimate of this parameter. Here, the observations of the random variables are $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ for n objects respectively. The range for Pearson's correlation coefficient lies between -1 and +1. If X and Y follow a linear relationship, their value is +1. Moreover, it is -1 if they do not follow a linear relationship. It is defined by,

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \ ,$$

where the symbol s_{xy} stands for the sample covariance of the variables X and Y. The sample variances of X and Y are denoted by the symbols s_X^2 and s_Y^2 respectively (Hay-Jahans, 2019, pp. 321-322).

3.4 Graphical Methods

3.4.1 Histogram

Histograms help determine how numerical data are roughly distributed. Interval of different lengths decomposes the values for a continuous variable They are displayed in a frequency table along with the total number of occurrences of each value. The use of vertical bars allows for the segment in each bin to be identified. The frequency tables as well as histograms with *y-axis* scaled in a way that it is comparable to a density are presented as a histogram. There is a diverse selection of bin widths although in most cases, they are ordered sequentially and do not overlap. Histograms can also provide information on the density of the distribution which is another valuable piece of data (Bruce and Bruce, 2017, pp. 21-23)

3.4.2 Scatterplots

A scatter plot is created by ordering the numeric variables x and y in some way and then plotting the result. The axes of this plot only exist in two-dimensional space. In order to generate a scatterplot depicting the relationship between two distinct variables, we use one numeric variable on the *x-axis* and the other on the *y-axis*. The variables may have a significant relationship if the plot points are close to a straight line. The correlation between the variables is diminished when the points are dispersed arbitrarily (Hay-Jahans, 2019, pp. 159-160).

3.4.3 Boxplots

A boxplot is data visualization based on the five-number summary which provides a concise summary of the set of observations. The five-numbers simplify dataset comparison. The box depicts the upper and lower quartiles of the distribution and the area between them accounts for fifty per cent of the entire set of observations. The parameters for the boxplots include the minimum value, the first quartile (Q_1), the median, the third quartile (Q_3) and the highest value. The quantiles divides observation sets into small subgroups of equal size. The quartile value then divides the observation sets into four parts which are all equal. Thus, the values that fall below twenty-five per cent are defined as the first quartile whereas values that fall below seventy-five per cent are defined as the third quartile. The right whisker extends to the data point representing the most significant observation. Similarly, the left whisker extends to the observation with the lowest observation. Thus, right-skewed whiskers have longer right whiskers. In addition, the longer left whisker indicates that the set of observations is left-skewed. In this case, the median we find is the centre of the box since the two whiskers are equal. Calculating the interquartile range (IQR) involves subtracting the first quartile from the third quartile and then drawing whiskers to the minimum and maximum values of the data set or to a multiple of the IQR. The line that runs horizontally through the middle of the box illustrates the median value of the set of observations. Boxplots that have considerable variation are more vertical in appearance than boxplots that have low variation (Hay-Jahans, 2019, pp. 137-141).

4 Statistical Analysis

In this section of the report, we apply the statistical methods described earlier to evaluate the given tasks. We define under age mortality rate as U5 mortality rates and life expectancy as LE at birth throughout the whole analysis and plots.

4.1 Frequency Distributions Analysis

Our initial objective in this subsection is to analyze the frequency distributions of different variables. The variables contain life expectancy at birth and under age 5 mortality rates for both sexes, males and females for the year 2022. We begin by describing the frequency distribution of the variables by employing histograms. Then we explain the ways in which the sexes under regions are distinct from one another by introducing boxplots.

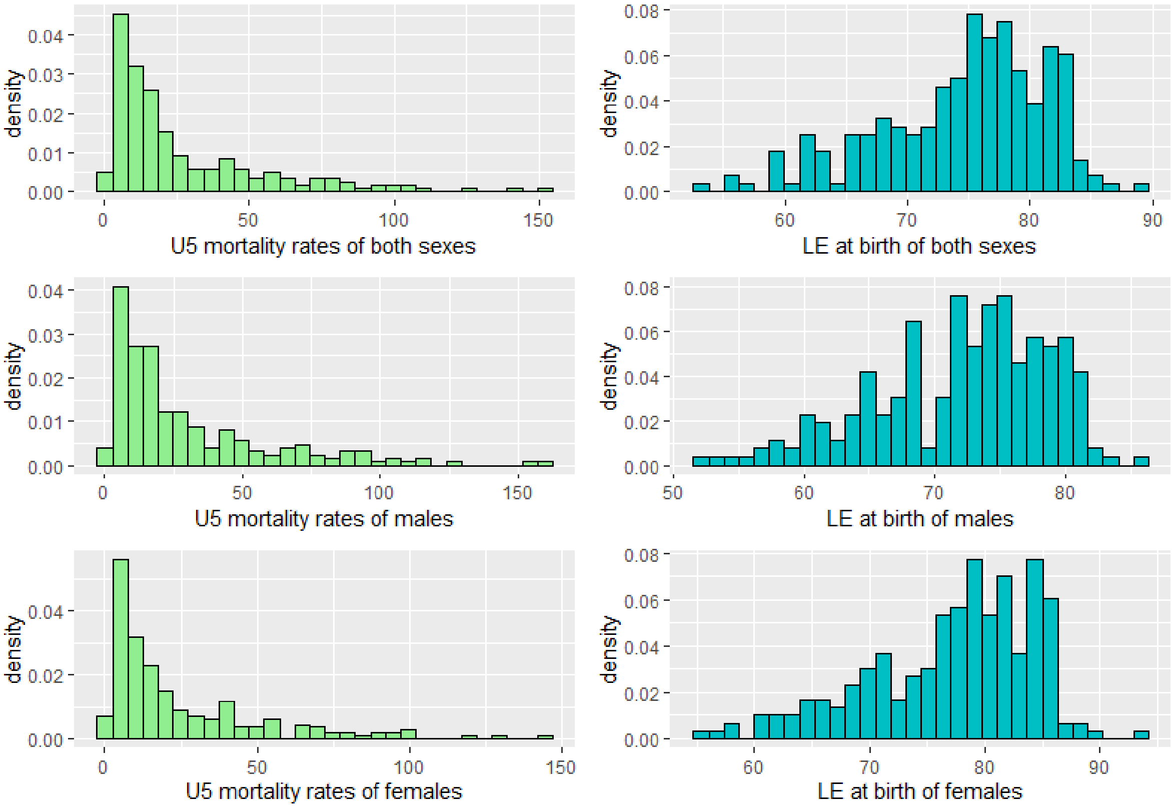

Figure 1: Distribution of frequencies over all of the variables

Figure 1 shows the frequency distribution of these variables within the dataset. If we closely observe the Figure 1 and Table 1 of the Appendix on page 18, the mean life expectancy at birth for females is 77.18 years with a median of 78.69 years. In contrast, the median life expectancy for males at birth is 73.26 years and the mean is 72.10 years.

Negative skewness indicates that most values are positioned on the right side of the histogram which we discover on the right side of Figure 1 for life expectancies of both sexes, males and females. Moreover, we also discover positive skewness on the left side of the histogram where plots of under age 5 mortality variables are present. As most observations fall on the left side, the corresponding histograms are positively skewed. From Table 1 of the Appendix on page 18, the mean mortality rate for children under 5 for both sexes is 26.68 and the median is 15.08. The mean mortality rates of males and females are 29.23 and 24.01 with medians of 17.55 and 13.62 respectively.

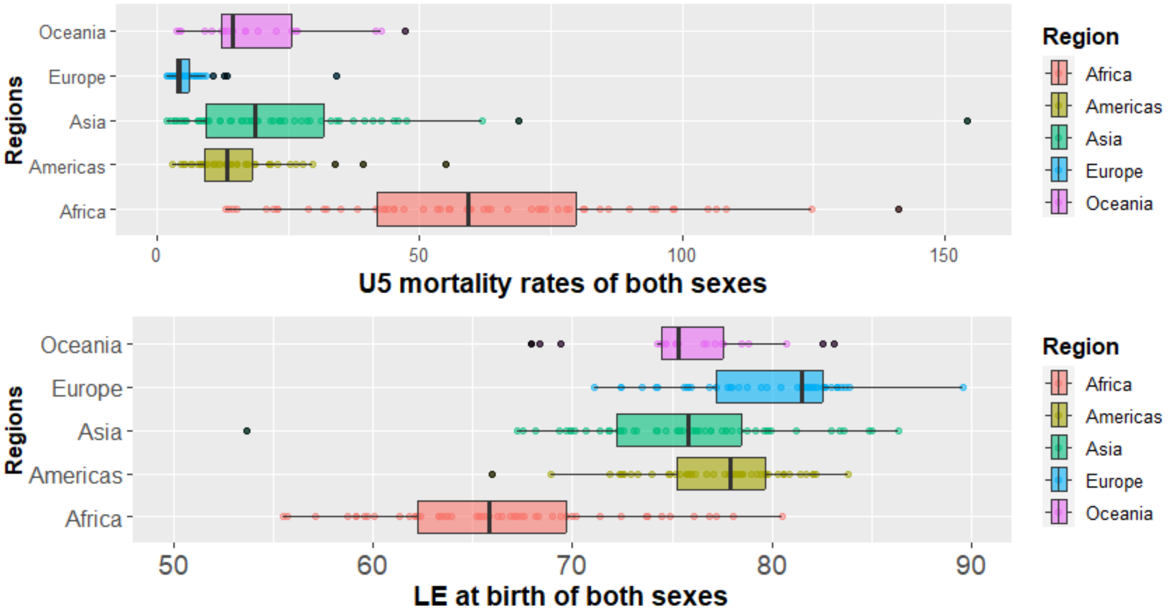

Figure 2: Boxplot for region differences of mortality rates and life expectancy of both sexes

Here, we discuss the second part of the question. From Figure 2, we show the difference between the regions. Now, we consider six variables which contain under age 5 mortality rates and the life expectancy of both sexes, males and females separately. The above figure shows the difference between each region for both sexes. After analyzing the boxplots, we observe that the highest mean (slightly greater than 80) and median (slightly less than 80) of life expectancy are found in Europe. The Americas and Oceania also followed similarly. On the other hand, Africa has the lowest mean (slightly greater than 65) and median (slightly greater than 65 and the lowest mean) regarding life expectancy. The data variability in life expectancy is the least in Oceania ranging from 72 to 81 whereas if we closely look at Africa, it has the most significant variation ranging from 51 to 81 approximately. Most of the countries in Europe have the highest life expectancy while most of the subregions of Africa is the lowest.

Now we observe the mortality of both sexes under age 5. Here we see that Africa has the highest mean (approximately 60) and median (close to 60) while Europe has the lowest. In Europe, the variation in mortality is also the lowest. However, Africa still has the most significant variation ranging from 12 to 125 whereas Europe has the lowest ranging from 1 to 11 approximately.

Similarly, the life expectancy of females in the Appendix in Figure 8 has a higher mean and median than males in all the regions. Also, the variability of the male plot is longer than the box in the female plot. Moreover, males under the age of 5 children in African regions have higher mean and median mortality rates than females in all regions. If we closely investigate the difference between the highest and lowest variance from each group (males or females), we observe that the mortality rate under age 5 for males has higher variability compared to under age 5 for female mortality.

4.2 Homogeneity and Heterogeneity Variability Analysis

In this section, we measure differences in life expectancy for males and females and under age 5 child mortality rate for both sexes in 2022 between and within subregions. There are important observations on page 19 of the Appendix which we discuss here too.

Figure 3 shows two boxplots for under age 5 mortality rates and life expectancy of both sexes. We observe that there is a significant amount of variation between each subregion from Table 2 of the Appendix on page 19. Hence, we compare the minimum and maximum values from the same table for the variable under age 5 mortality rates for both sexes across subregions in Africa. We observe a minimum and maximum value of 12.95 and 141.20 respectively for Northern and Eastern Africa. In addition, the interquartile range (IQR) for each subregion is quite distinct indicating that the data are heterogeneous. Comparing the central tendency measures of various subregions reveals that there are significant differences. As an illustration, the mean and the lowest mortality rate under age 5 for both sexes in Northern Africa is 36.21 while in Middle Africa it is 84.76 which is also the highest. This suggests that there is heterogeneity between subregions.

Analyzing the variable for life expectancy for both sexes between African subregions reveals a range of values for each subregion with distinct minimum and maximum values. The interquartile range (IQR) for each subregion also indicates significant heterogeneity in the data points. Comparing central tendency measures between subregions reveals

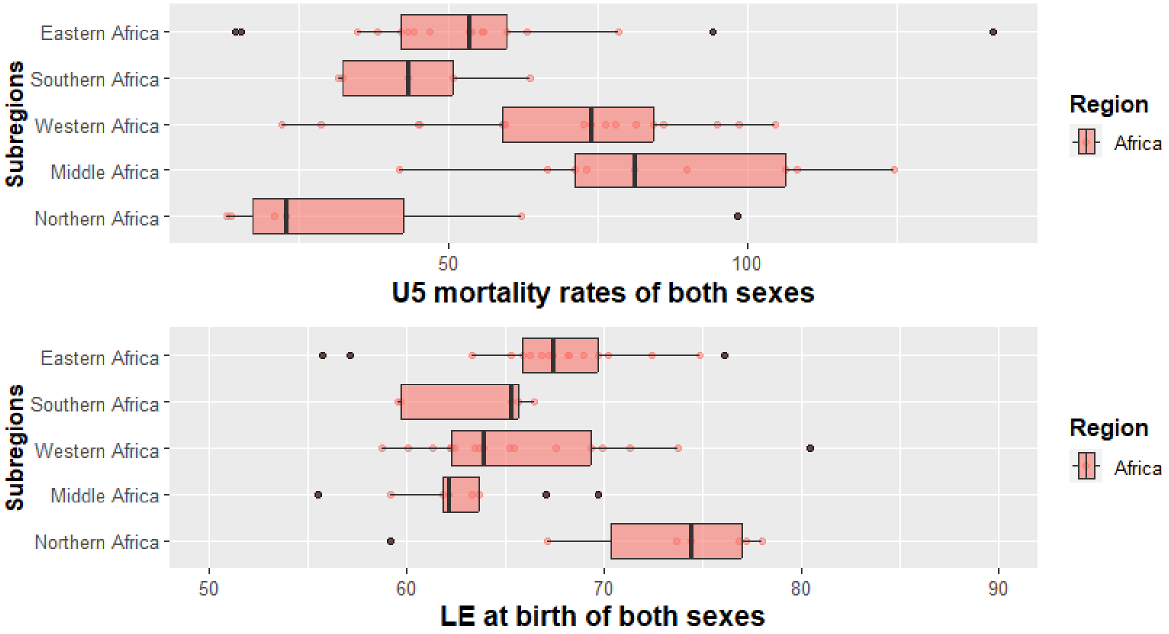

Figure 3: Boxplots for both sexes displaying homogeneity and heterogeneity between subregions

notable differences such as the disparity between the mean values of Northern Africa (72.35) and Middle Africa (62.72) from Table 2 indicating heterogeneity. Besides, the (IQR) of Northern Africa within the life expectancy for both sexes, males and females separately are 6.60, 6.60, and 6.61 respectively which is an indication that these variables are homogenous within the individual subregions.

In Figure 9 and Table 2 of the Appendix page 19, we analyze the mean and median values for the variable mortality rate under age 5 males vary across the subregions of Africa with Middle Africa having the highest mean and median (90.11 and 87.58 respectively) and Northern Africa having the lowest (39.30 and 24.40 respectively). Northern Africa has the most significant variance (1182.78) whereas Southern Africa has the lowest variance (218.78). The interquartile range (IQR) varies from 20.03 in Southern Africa to 37.63 in Middle Africa. However, in Table 2, the variation between subregions is also substantial with a range of approximately 50.81 for the mean, 63.18 for the median and 964 for the variance from the lowest to the highest subregion. Overall, the boxplot shows the considerable variation of mortality rates under the age of 5 males between subregions in Africa demonstrating the heterogeneity of the continent.

We also observe the data for mortality rates under age 5 females in Figure 9 and Table 2 of the Appendix page 19. The mean values are relatively consistent across the subregions of

Africa ranging from 32.95 in Northern Africa to 79.25 in Middle Africa. Also, the median values are consistent which range from 20.23 in Northern Africa to 74.39 in Middle Africa. The variance values range from 144.17 in Southern Africa to 880.19 in Northern Africa indicating that the data for this variable between each subregion are more dispersed indicating heterogeneity. The variance of mortality rates under age five of both sexes, males and females for Western Africa are 590.99, 618.94 and 565.58 respectively which also follow homogeneity. The (IQR) ranges also vary ranging from 13.76 to 32.91 in Eastern and Middle Africa respectively. This signifies that the information for mortality rates under age 5 females is more dispersed between each subregion.

4.3 Bivariate Analysis

In this subsection, we investigate if there are any two variables which have a relationship with each other or not. We observe the difference between the lifespan of average men and women from the scatter plot in Figure 4. We calculate Pearson's correlation coefficient to assess the strength of the relationship between the variables.

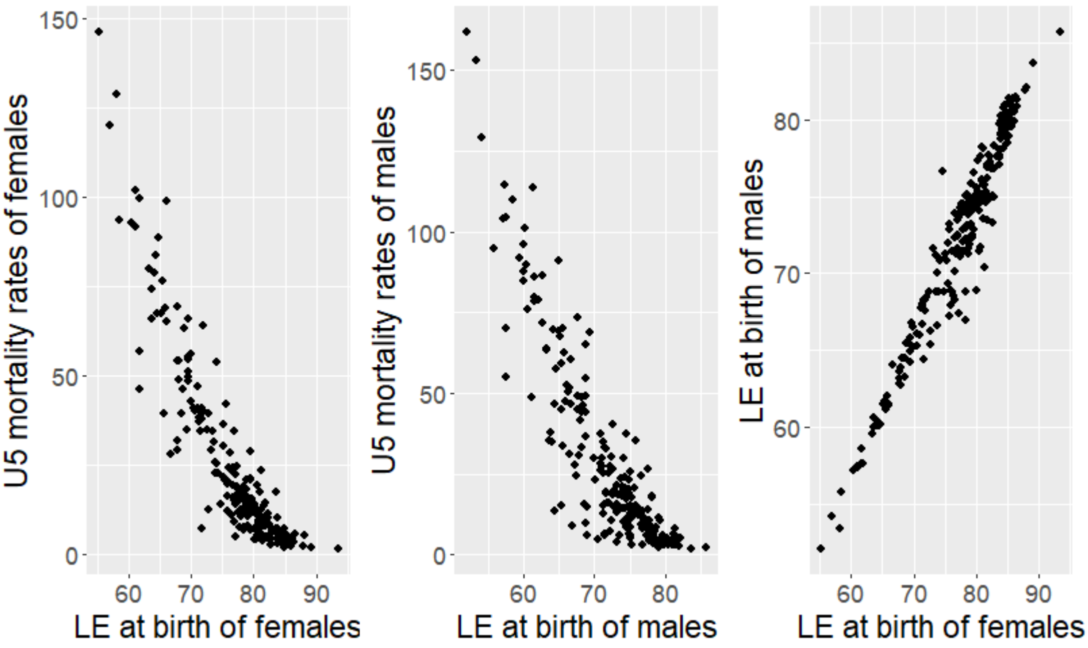

Figure 4: Scatter plot for displaying relationship between variables

The scatter plots presented in Figure 4 depict the correlation between various variables in 2022. The scatter plots of child mortality rates under 5 for both sexes, males and females show a downward trend indicating a negative correlation presented in the first two parts of Figure 4. On the contrary, the third part of Figure 4 which depicts life expectancy for males and females demonstrate a linearly increasing pattern. This increasing pattern indicates a positive correlation between these variables. We can interpret from the scatter plot in Figure 4 that an increase in the mortality rate under 5 is associated with a decrease in life expectancy for both sexes, males and females. Therefore, the child mortality rate under age 5 for males and females decreases when life expectancy increase.

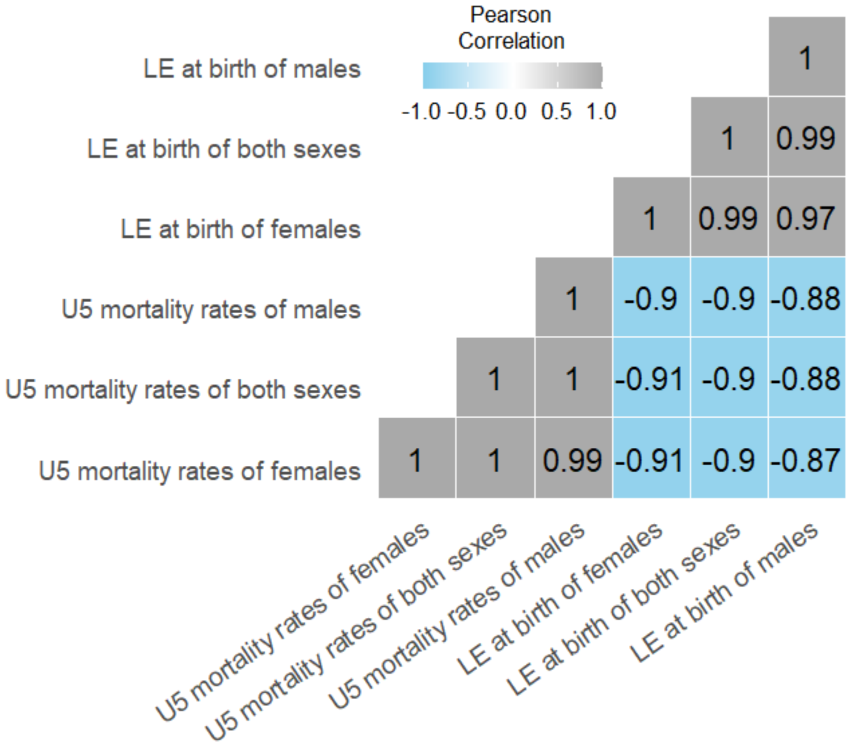

Figure 5: Pearson correlation across all variables

We also examine the correlation in Figure 5 between life expectancy and under age 5 mortality across countries. As there are various indicators of these variables, we investigate the correlation coefficients. We acknowledge a strong positive linear correlation between life expectancy for both sexes when compared to male (0.99) and female (0.99) respectively in Figure 5 according to the correlation coefficient matrix presented. Furthermore, we also investigate the correlation between the under age 5 mortality rate for both sexes and the life expectancy of both sexes. The correlations for child mortality under age 5 for both sexes against the life expectancy of both sexes are negative with correlation coefficients of -0.9. And the correlation coefficient for under age 5 mortality for males and females separately against life expectancy for males and females are -0.88 and -0.91 respectively.

4.4 Variable Change Between 2002 and 2022

The scatterplots in Figure 6 between 2002 and 2022 depict the changes in mortality rates under age 5 and life expectancy for both sexes across all regions. On *x-axis*, we plot the mortality of both sexes under age 5 and the life expectancy of both sexes for 2022 and *y-axis*, we plan with the same variables but for 2002. Comparing life expectancy and mortality rates under age 5 of both sexes, we analyze changes in these variables from 2002 to 2022. We compare the averages of the two variables for the years 2002 and 2022 to perform this. We can obtain these values from the R code (R Core Team, 2020)

When examining the mortality rates under age 5 on the scatterplot, most points are located in the lower left quadrant above the line indicating a decline in mortality rates across most regions. This indicates that there has been a substantial decrease in the mortality rates across all regions between 2002 and 2022 as previously stated.

Meanwhile, the scatterplot of life expectancy indicates that there has been an increase in life expectancy across most regions in 2022. The majority of the data points fall on the upper right quadrant below the line. The observations from 2002 indicate a noteworthy enhancement in life expectancy across all geographical areas during the period spanning from 2002 to 2022. According to the scatterplot, life expectancy (57.17) from 2002 in some regions, such as Africa, raised significantly (66.41) in 2022. In contrast to other regions, such as Europe, where life expectancy (76.06) was relatively higher in 2002, we observe the rise in life expectancy (79.93) is relatively moderate in 2022. The mean life expectancy of both sexes for all regions has increased. In 2002, the mean for Africa, Americas, Asia, Europe and Oceania are 57.17, 73.43, 69.94, 76.06 and 70.62 respectively whereas in 2022 the mean is 66.41, 77.15, 75.31, 79.93 and 75.54. The second largest growth we observe for Asia at 5.37 years after Africa at 9.24 years. Moreover, America and Europe have grown by 3.72 and 3.87 years respectively. Africa has the most significant rise at 9.24 years while America has the lowest at 3.72 years.

Similarly, we observe a decreasing trend in both males and females for mortality rates under age 5 for most of the countries for the year 2022. Here, Africa has higher variability compared to other regions for mortality. In contrast, we acknowledge an increasing trend in the case of life expectancy for males and females for the year 2022. Most of the

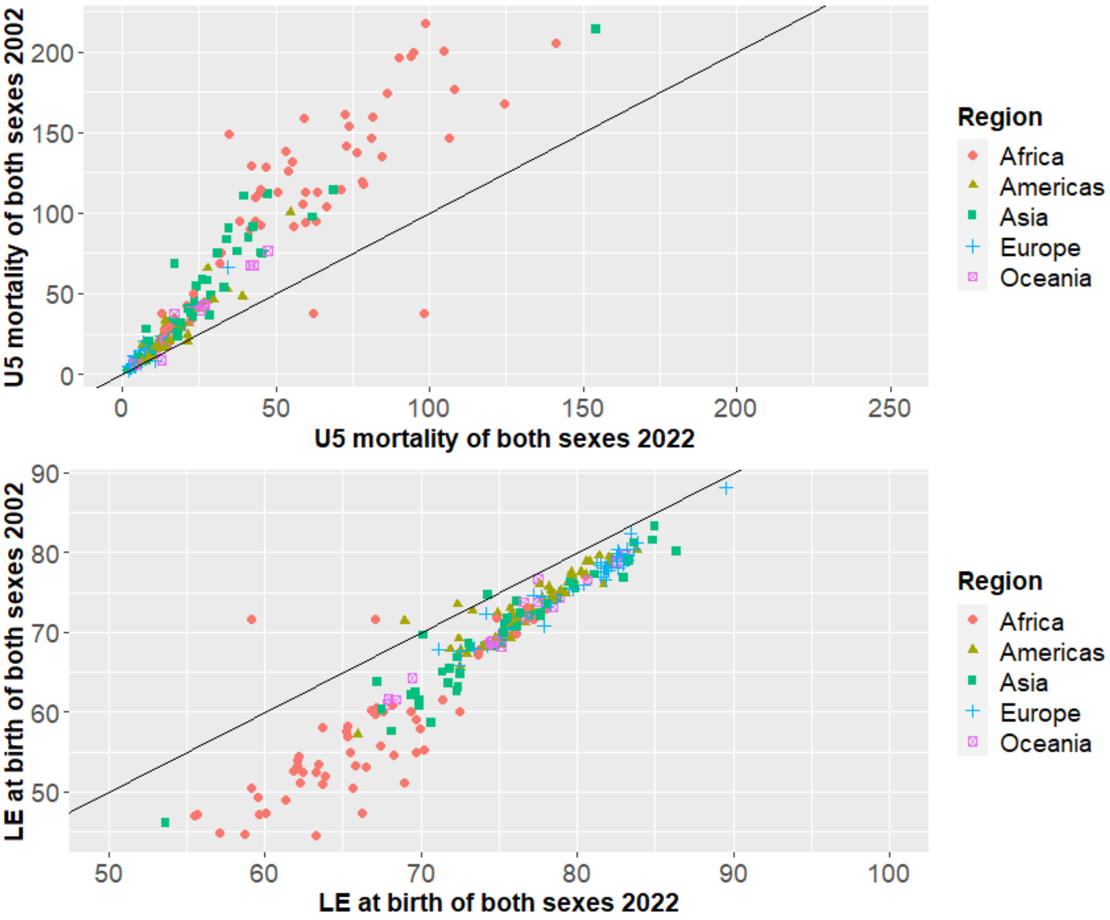

Figure 6: Mortality rates and life expectancy comparison of both sexes

countries follow the increasing trend for the year 2022 compared to 2002. We discover that Africa has a life expectancy (64.39) and mortality rate (65.95) for males for the year 2022 whereas we found much lower life expectancy (55.65) and higher mortality rates (121.72) for the males of 2002. In addition, the rate of growth of life expectancy of females here is also higher compared to the males of all African regions. On the other hand, the mortality rate of females under age 5 decreases more than the mortality of males in Africa.

5 Summary

In this study, we analyzed the demographic data of IDB which is maintained by the U.S. Census Bureau. The dataset analyzed for the entire study contained 454 observations with a total of 227 countries. There were 5 regions and 21 subregions. The dataset had

11 variables including categorical variables such as country, region and subregion names. There were also 6 continuous variables for under 5 mortality rates and life expectancy for both sexes, males and females individually. We also dealt with missing values by implying mean imputation based on subregions. As per the given problem, we analyzed the observations from the year 2022 for the first three tasks. We also investigated the difference between 2002 and 2022 in the last part of our research question. The first task reveals significant differences in life expectancy and under age 5 mortality rates particularly by gender. We observed Europe to be the highest life expectancy while Africa has the highest mortality rates. In addition, females have higher life expectancy than males whereas males have higher mortality rates than females.

Furthermore, we observed differences in life expectancy and mortality rates across the regions. By analyzing the boxplots, we observed Europe to be the highest life expectancy while Africa has the highest mortality rates. And for the second task, the values of the individual variables are relatively homogeneous within the individual subregions, but more heterogeneous between different subregions due to the larger variability in the data. In the third task, the analysis of the bivariate correlations among the variables indicates an inverse correlation between life expectancy and under age 5 mortality rates. In the final part of the report, life expectancy increased significantly in Africa and Asia worldwide for the years 2002 and 2022, but there was a slight increase in the Americas, Europe and Oceania. In contrast, under age 5 mortality rates have decreased in all regions.

In conclusion, further investigation is possible to understand the underlying causes of demographic trends including factors such as healthcare infrastructure and environmental factors. The implications of these trends and patterns for global health and economic advancement could also be explored through further research.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. 01 2019. ISBN 9780429448294. doi: 10.1201/9780429448294.

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

U.S. Census Bureau. International database. https://www.census.gov/programs-surveys/international-programs/about/idb.html, 2022. Accessed: 2023-04-20.

Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL http://www.jstatsoft.org/v21/i12/.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham and et al. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686. URL https://doi.org/10.21105/joss.01686.

Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2020. URL https://CRAN.R-project.org/package=cowplot. R package version 1.1.1.

Appendix

A Additional Figures

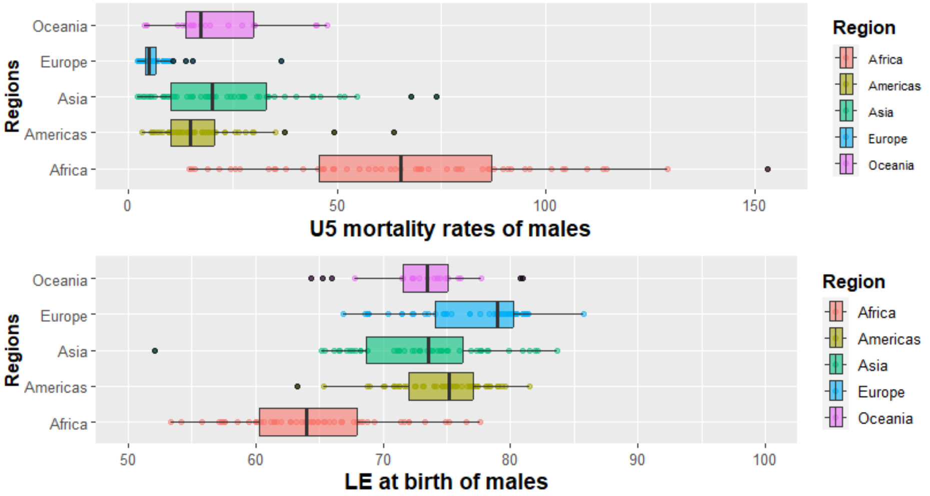

Figure 7: Boxplots for males displaying homogeneity and heterogeneity between subregions

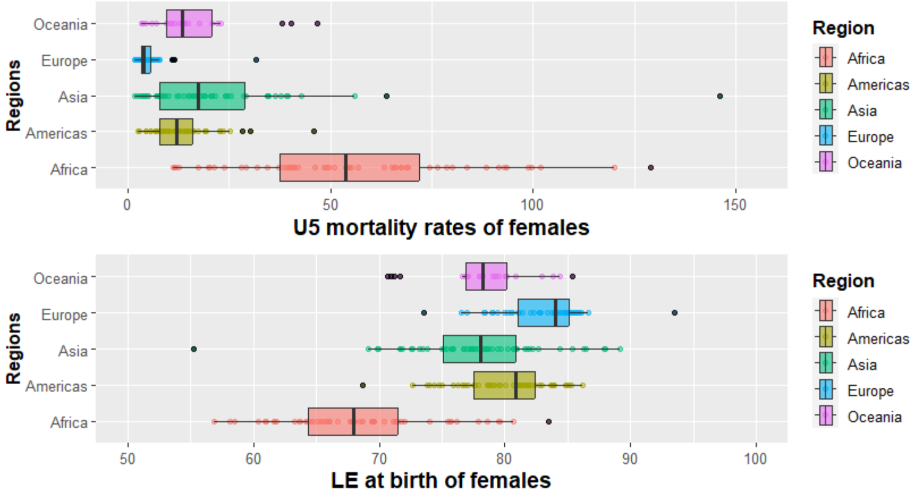

Figure 8: Boxplot for females illustrating homogeneity and heterogeneity between subregions

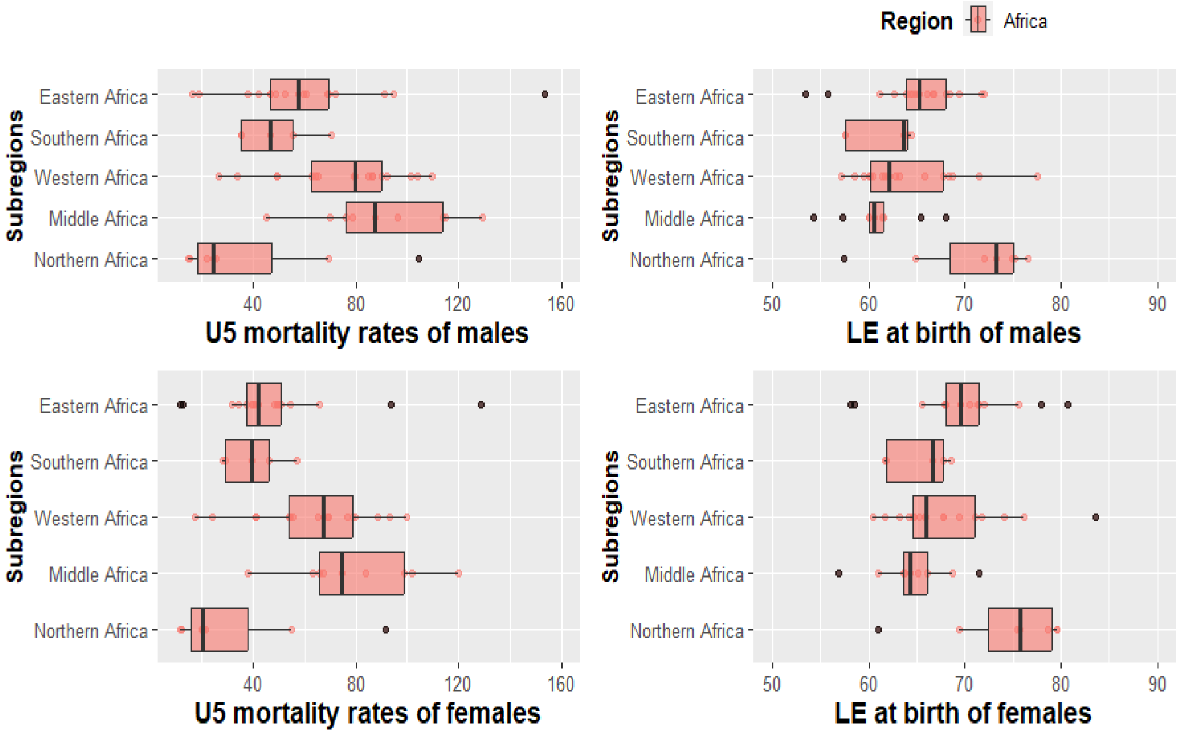

Figure 9: Boxplots showing variability across subregions between four variables

B Additional Tables

Table 1: Mean, Median and Variance for the variables in 2022

Variables	Mean	Median	Variance
U5 mortality of both sexes	26.68	15.08	789.29
U5 mortality of males	29.23	17.55	905.20
U5 mortality of females	24.01	13.62	683.30
LE at birth of both sexes	74.58	75.82	46.78
LE at birth of males	72.10	73.26	44.54
LE at birth females	77.18	78.69	50.83

Table 2: Summary statistics for subregions in Africa for 2022

Variables	**Northern**	**Middle**	**Western**	**Southern**	**Eastern**
U5 mortality of both sexes					
Mean	36.21	84.76	68.79	44.25	54.95
Median	22.85	81.09	73.93	43.11	53.35
Variance	1028.07	644.27	557.68	179.98	876.53
IQR	25.33	35.31	25.38	18.45	17.74
Min	12.95	41.73	22.14	31.61	14.30
Max	98.26	124.58	104.72	63.57	141.20
U5 mortality of males					
Mean	39.30	90.11	74.31	48.48	60.92
Median	24.40	87.58	79.88	46.80	57.61
Variance	1182.78	682.80	587.54	218.78	1009.72
IQR	28.93	37.63	27.09	20.03	23.20
Min	14.43	45.29	26.61	35.07	15.97
Max	104.67	129.08	109.77	70.16	153.23
U5 mortality of females					
Mean	32.95	79.25	63.08	39.91	48.80
Median	20.23	74.39	67.33	39.32	42.02
Variance	880.19	608.09	531.43	144.17	770.16
IQR	22.06	32.91	24.81	16.83	13.76
Min	11.38	38.07	17.43	28.09	11.51
Max	91.52	119.95	99.52	56.79	128.81
LE at birth both sexes					
Mean	72.35	62.72	65.96	63.34	67.28
Median	74.45	62.11	63.90	65.32	67.42
Variance	47.18	16.87	30.98	11.64	27.55
IQR	6.60	1.87	7.11	5.95	3.84
Min	59.16	55.52	58.76	59.57	55.72
Max	78.03	69.70	80.48	66.47	76.10
LE at birth males					
Mean	70.60	60.95	63.91	61.45	65.00
Median	73.26	60.65	62.04	63.60	65.32
Variance	48.35	16.34	28.41	12.46	23.93
IQR	6.60	1.46	7.51	6.37	4.09
Min	57.43	54.19	57.16	57.57	53.39
Max	76.57	67.98	77.58	64.46	72.04
LE at birth females					
Mean	74.19	64.53	68.08	65.28	69.62
Median	75.72	64.24	65.99	66.68	69.57
Variance	46.33	17.58	33.93	10.97	32.46
IQR	6.61	2.42	6.57	5.93	3.47
Min	60.97	56.88	60.41	61.64	58.12
Max	79.57	71.48	83.51	68.53	80.66