TU Dortmund

Introductory Case Studies

Project 1: Descriptive data analysis

Lecturers:

Dr. Crystal Wiedner

Dr. Marlies Hafer

Dr. Rouven Michels

Author: Sabrina Sultana

Matriculation Number : 235062

Group number: 6

Group members: Md Ryad Ahmed Biplob, Nafisa Farhin, Md Mahmudul Hasan Bhuiyan

November 3, 2024

Contents

1 Introduction

Demographic data includes a population's gender, income, social status, etc., in a geographic area. They provide an overview of the various changes in a geographic location based on demographic data at different points in time in the exact location [French, 2014]. Analyzing demographic data offers valuable insights into how various attributes evolve over time and allows for comparisons between different countries, regions, or genders. This information can be beneficial to multiple fields, including economics, social studies, healthcare, and education systems. This project aims to analyze demographic data from 227 countries, focusing on life expectancy at birth and under-5 mortality rates for the years 2004 and 2024. First, descriptive statistics such as frequency distribution and central tendency are used to analyze each variable. Variability measures are employed to assess variance between regions and subregions. Correlation coefficients are utilized to determine the dependence between data pairs. Finally, a comparison is conducted between the datasets from 2004 and 2024.

Section 2 provides an overview of the dataset and its characteristics. Section 3 outlines the statistical methodologies employed to analyze the data. Section 4 presents and interprets the results of the analysis. Finally, Section 5 summarizes the main results and presents an outlook on further possible analyses of the given data set.

2 Problem statement

2.1 Description of the dataset

The dataset for this project is sourced from the International Database (IDB) of the U.S. Census Bureau. The IDB provides demographic data from 1950 to 2100 for all countries and regions recognized by the U.S. Department of State, with a population of 5,000 or more. The data is gathered from state institutions, including censuses, surveys, and administrative records, and supplemented by estimates and projections by the U.S. Census Bureau itself [U.S. Census Bureau, 2023]. The dataset used in the project is an extract from the IDB, containing life expectancy at birth and under-5 mortality rates for 227 countries for the years 2004 and 2024.

The dataset consists of the life expectancy at birth and under age 5 mortality rates of both males and females in 2004 and 2024. This includes a total of 227 countries divided

into 5 regions. These regions are further divided into 5 subregions. The dataset includes a total of 454 observations and comprises ten variables. A brief description of each variable available on the IDB website [U.S. Census Bureau, 2021] is provided below:

Country: This dataset contains 227 different countries over the world. The country name describes the name of those countries..
Region: A region consists of subregions and countries. Africa, Asia, the Americas, Oceania, and Europe are the regions included in the dataset.
Subregion: A subregion consists of many countries. For example, the Western Europe subregion includes nine countries, namely Belgium, the Netherlands, Germany, Luxembourg, Austria, Liechtenstein, France, Switzerland, and Monaco.
Year: The dataset contains two years which are 2001 and 2021.
Life expectancy at birth for both sexes: The average number of years a group of people born in the same year can be expected to live if mortality at each age remains constant in the future.
Life expectancy at birth for males: Life expectancy for males is described as the average number of years that a group of males born in the same year is expected to live if mortality at each age in the future remains constant.
Life expectancy at birth for females: Life expectancy for females is defined as the average number of years that a group of women born in the same year is expected to live if mortality at each age in the future remains constant.

Under-5 mortality for both sexes: Number of deaths of children under 5 years of age from a cohort of 1,000 live births. It is the probability of dying between birth and the exact age of 5.
Under-5 mortality for males: Number of deaths of male children under 5 years of age from a cohort of 1,000 live births.
Under-5 mortality for females: Number of deaths of female children under 5 years of age from a cohort of 1,000 live births.

2.2 Project objectives

The main objective of this project is to analyze the given demographic data to find possible distribution, correlation, and relationship between variables. First, the frequency distributions of the variables for the year 2024 are analyzed using a histogram and measures of central tendency—mean and median. Second, the variability between the

regions and subregions is described with the help of box plots. Then, the dependency between the variables in the year 2024 is calculated using the pair plot matrix as well as the linearity and monotonicity of their relationship are measured using Pearson correlation coefficients. Finally, the observations between 2004 and 2024 for the variables of Life expectancy at birth and under 5 age mortality of both sexes are compared using the box plots.

3 Statistical methods

This section provides an overview of the statistical methods, models, and plots utilized to achieve the objectives of this project. The statistical software R [R Development Core Team, 2020], version 4.4.1, was used for data calculation, modeling, and visualization.

3.1 Statistical Measures

3.1.1 Mean

The arithmetic mean, or mean, of a set of measurements is defined to be the sum of the measurements divided by the total number of measurements. The arithmetic mean, commonly referred to as the average, is calculated by dividing the sum of all the values in a dataset by the number of values. It provides a single representative figure for the entire dataset, but extreme values can significantly influence it. Therefore, it is typically preferred when dealing with relatively homogeneous data. If we let $y_1, y_2, \ldots, y_n$ denote the measurements observed in a sample of size n, then the sample mean $\bar{y}$ can be written as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1}$$

where the symbol appearing in the numerator, $\sum_{i=1}^{n} y_i$, is the notation used to designate a sum of n measurements, y_i:

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n \tag{2}$$

The corresponding population mean is denoted by μ[Ott and Longnecker, 2016, p. 86]. The mean serves as a central value that best represents the data; however, it is sensitive to outliers and is most suitable for datasets with relatively uniform data points.

3.1.2 Median

The median is a measure of central tendency that represents the middle value of an ordered dataset, dividing the data into two equal halves. For an ascending ordered dataset $x_1, x_2, \ldots, x_n$, the median $\tilde{x}$ is defined as follows:

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

where n is the size of the sample. The median is preferred over the mean in situations where the dataset contains extreme values, as it is not affected by outliers [Hay-Jahans, 2019, pp. 75–76].

3.1.3 Correlation Coefficient

The correlation coefficient is a numerical measure of the strength and direction of a linear relationship between two variables. The correlation coefficient, denoted by r, ranges from -1 to 1. A value of $r = 1$ or $r = -1$ indicates a perfect positive or negative linear relationship, respectively, while a value of $r = 0$ suggests no linear relationship. The sign of the correlation coefficient represents the direction of the relationship; positive values mean that as one variable increases, the other also tends to increase. In contrast, negative values indicate an inverse relationship. Using the Pearson correlation coefficient to quantify these relationships is beneficial when analyzing the linearity between two continuous variables [Snedecor and Cochran, 1989].

The Pearson Coefficient: The Pearson correlation coefficient, denoted by r, is used to quantify the strength and direction of a linear relationship between two variables. The formula for calculating the Pearson correlation coefficient is:

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

where: x_i and y_i are the individual sample points, $\bar{x}$ and $\bar{y}$ are the means of the variables x and y, respectively, and n is the number of observations. This measure is beneficial for identifying the linearity of relationships between variables but is sensitive to extreme values.[Moore et al., 2018].

3.2 Statistical Plots

3.2.1 Scatter plot

A scatter plot is a graphical representation that helps visualize the relationship between two quantitative variables. It is created by plotting one point for each pair of variables representing an observation in the dataset [Newbold et al., 2013, p. 47]. Each point on a scatter plot represents an observation from the dataset, with its position determined by the values of two variables on the horizontal and vertical axes. Scatter plots are particularly useful for identifying patterns, such as linear relationships (either positive or negative), non-linear trends, and clusters. They also detect outliers—data points that deviate significantly from the overall pattern, providing insights into potential anomalies in the data.

3.2.2 Box plot

A box plot is a graphical representation that summarizes the distribution of numerical values. It consists of a box formed by the first and third quartiles, with the difference between them called the interquartile range (IQR). The median, or second quartile, is displayed in the middle of the box. Two whiskers extend below and above the box, indicating the minimum and maximum values. To identify outliers, fences are used. The upper fence is drawn at the third quartile plus 1.5 times the IQR, while the lower fence is at the first quartile minus 1.5 times the IQR. Observations beyond these fences are considered outliers [Kohler and Kreuter, 2005, p. 161–162].

3.2.3 Histogram

A histogram is a graphical representation that helps display grouped frequency distributions of data. It divides the data into classes, using bars to represent the frequency of each class interval. The height of each bar corresponds to the frequency, density, or

relative frequency of the values in each class [Hay-Jahans, 2019, p. 131]. When the class intervals have the same width, the heights of the bars in the histogram are proportionate to the absolute frequencies, making understanding more accessible. A histogram is symmetric if the dataset's mean and median are equal. If the mean is more than the median, the histogram has a right skewness; if the median is greater than the mean, the histogram displays a left skewness.

4 Statistical analysis

4.1 Frequency Distribution Analysis

In this subsection, histograms are used for visualizing absolute frequencies using the bin range for the continuous variables where bin ranges are on the x-axis and frequencies are on the y-axis.

In Figure 1, most countries' life expectancy for females typically ranges from 78 to 85 years, with a mean of 77.6 years and a median of 78.8 years. This indicates a slight positive skew in the data, meaning that more countries have life expectancies above the mean compared to below. For males, life expectancy is typically between 70 and 75 years, with a mean of 72.6 years and a median of 73.5 years, indicating a slight positive skew. The higher median than the mean suggests that some countries have lower life expectancy values, pulling down the average. As reflected in these distributions, females have a higher life expectancy than males.

For the under age 5 mortality rate in 2024 is shown by Figure 2. It can be observed that the mortality rate is primarily concentrated below 40 for both males and females. The highest frequency for females is within the range of 0 to 20, with over 80 observations in this category. Similarly, the male mortality rate shows the highest frequency within the 0 to 20 range, though with fewer observations compared to females. The mean mortality rate for females is 22.5, while the median is 12.6, indicating a right-skewed distribution. Similarly, for males, the mean is 27.4, and the median is 16.4, also suggesting a right-skewed distribution. This skewness implies that there are some countries with very high under-5 mortality rates, which increase the mean compared to the median. Overall, the distributions for both sexes reflect that most countries have lower mortality rates, but a few countries with high rates skew the average upwards.

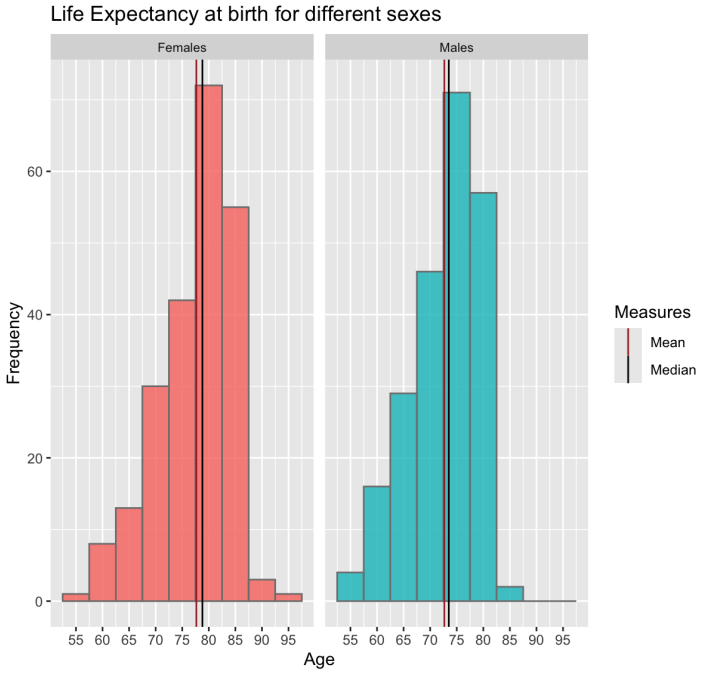


Figure 1: Histogram for life expectancy at birth of different sexes in 2024

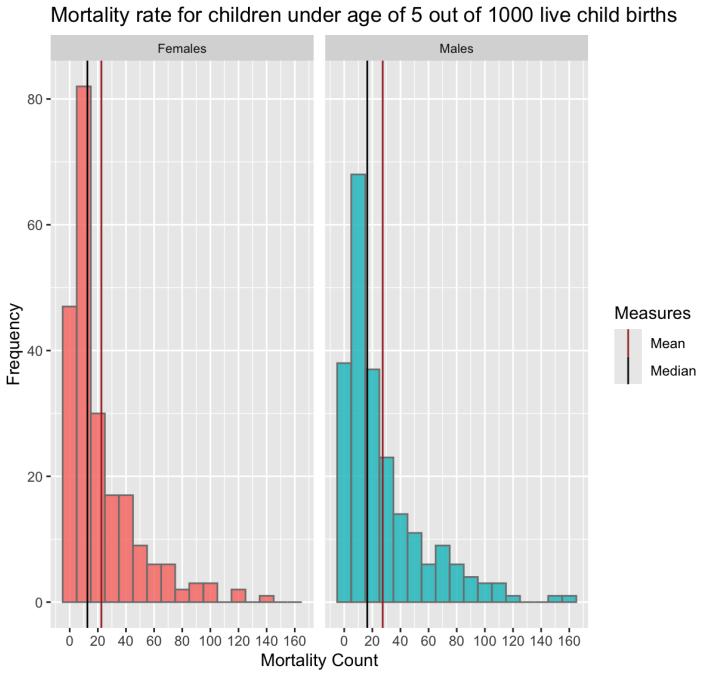


Figure 2: Histogram for life expectancy at birth of different sexes in 2024

4.2 Data Variability Analysis

The boxplots in figure 3 depict life expectancy at birth and under-5 mortality rates for different subregions of Europe, including Eastern, Northern, Southern, and Western Europe. Here, In the left figure, Northern Europe and Western Europe has the least variability, with life expectancy values clustering tightly between 82-85 years, indicated by a small interquartile range (IQR) and few outliers, which suggests a more homogeneous distribution. In contrast, Southern Europe displays a relatively wider spread in the data points compared to other subregions, with life expectancy ranging from approximately 75 to 85 years. The median is around 82 years, with a broad interquartile range (IQR) indicating higher variability, suggesting a more heterogeneous distribution. Similarly, Eastern Europe shows a median life expectancy of around 76 years, the lowest among the European subregions. The interquartile range (IQR) is relatively narrow, indicating limited variability within most of the data, with life expectancy values mostly clustering close to the median, suggesting a less heterogeneous distribution.

The plot on the right illustrates all the median values for under age 5 mortality (Both Sexes) across the European subregions are below 10, indicating generally low mortality rates across these regions. Among them, the under age 5 mortality rates are highest and most variable in Eastern and Southern Europe, while Northern and Western Europe exhibit lower and more consistent rates. Northern and Western Europe both display homogeneous under-age-5 mortality rates, with medians well below 5 and a very narrow IQR, indicating uniformity. Each region has a lower outlier, suggesting exceptional outcomes in certain areas. Eastern Europe has the highest median, a wide IQR, and several high outliers, indicating considerable variability and some areas with notably higher mortality rates whereas Southern Europe also exhibits a wide IQR and long whiskers, reflecting substantial variability across the subregion.

4.3 Relationship Analysis

The plots in figure 4 represent the relationship between each pair of variables in 2024. Figure 4(a) displays a pair plot, specifically a scatter plot matrix, showing the relationship between two variables: life expectancy at birth (both sexes) and Under age 5 mortality (both sexes). As life expectancy at birth for both sexes increases, under age 5 mortality decreases. The scatter plots show an inverse relationship: increased life expectancy correlates with lower under age 5 mortality rates. The color-coded points

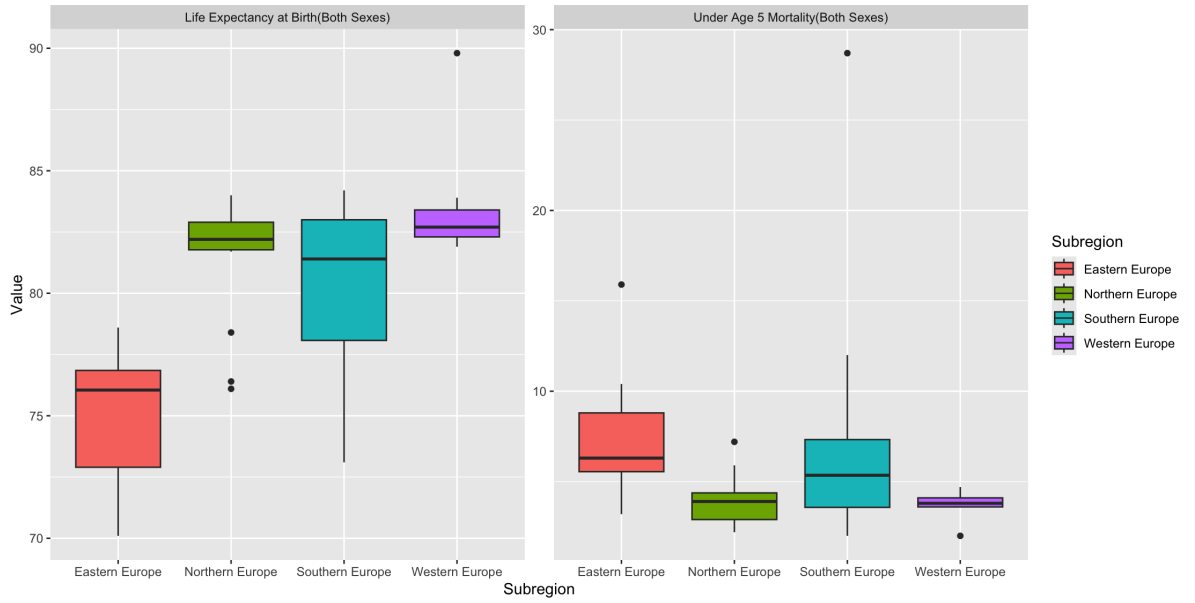


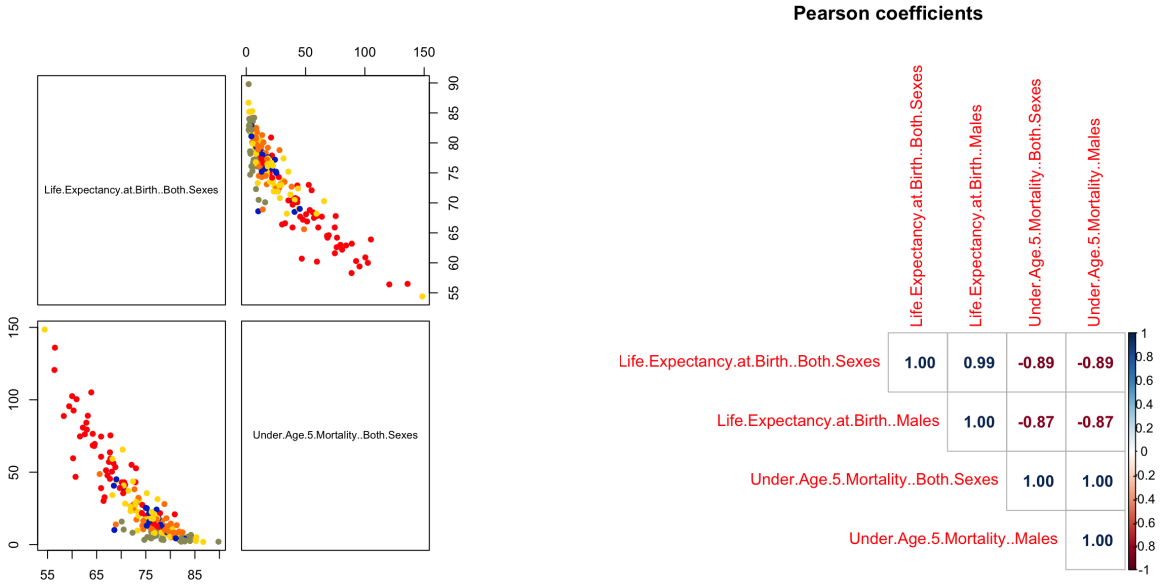
Figure 3: Box plots of variables for Europe Region in 2024

most likely represent distinct regions and consistently follow the overall negative pattern. Figure 4(b) indicates a strong negative relationship between life expectancy (for both sexes) and under-age-5 mortality (for both sexes), with Pearson correlation coefficients of -0.89. This suggests a strong inverse relationship, where an increase in life expectancy corresponds to a significant decrease in under-age-5 mortality. The correlation between life expectancy for both sexes and males is extremely high (0.99), indicating a strong positive linear relationship between them. By analyzing both the pair plot and the coefficient value, it can be concluded that the life expectancy of males has a a strong positive linear relationship with the life expectancy of both sexes and a strong negative linear relationship between the variables for both sexes as well as for males.

4.4 Variability Comparison

The box plots illustrate in figure 5 and figure 6 compare the life expectancy at birth and under 5 age mortality for both sexes across different regions between 2004 and 2024. In each plot, 2004 is represented by the red box and 2024 by the blue box, highlighting life expectancy evolution over the two decades.

In figure 5, the plots show that life expectancy has increased in 2024 across all regions compared to 2004. Africa, while still having the lowest life expectancy, improved significantly from around 55 years in 2004 to about 65 years in 2024.Both the Americas and



(a) (b)

Figure 4: (a) Scatter plot matrix, (b) Pearson coefficients between the pairs of variables in 2024

Europe currently have high life expectancy values. In Europe, the median life expectancy ascended from around 78 years to 83 years, whereas in the Americas, it increased slightly from about 75 years in 2004 to roughly 78 years in 2024. Both regions show consistent improvement, with Europe maintaining the highest median life expectancy. In Asia, life expectancy has moderately increased, accompanied by a narrowing interquartile range from 2004 to 2024, indicating reduced variance in regional health outcomes. Oceania shows improvement, with median life expectancy rising significantly. Although there are a few outliers, the overall upward trend points to progress in health and living standards.

From Figure 6, it can be declared that under 5 age mortality rates have decreased in 2024 in most of the regions. The under-5 mortality rate in Africa has significantly decreased from 2004 to 2024, as shown by a lower median rate (from around 120 in 2004 to approximately 60 in 2024) and reduced variability. Similarly, Asia has experienced a downward trend, with the median mortality rate decreasing from about 50 in 2004 to around 25 in 2024. This suggests a considerable improvement in child health outcomes in both regions. The Americas and Europe exhibit relatively stable and low child mortality rates, with minimal changes between 2004 and 2024. This stability indicates that child mortality was already low in these regions. Oceania also shows a decline in the median mortality rate.

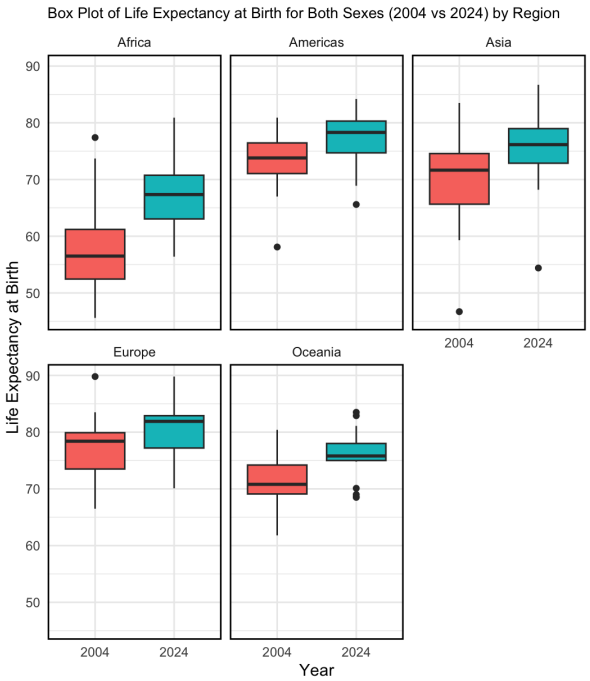


Figure 5: Box plots of life expectancy for both sexes (2004 vs 2024) by region

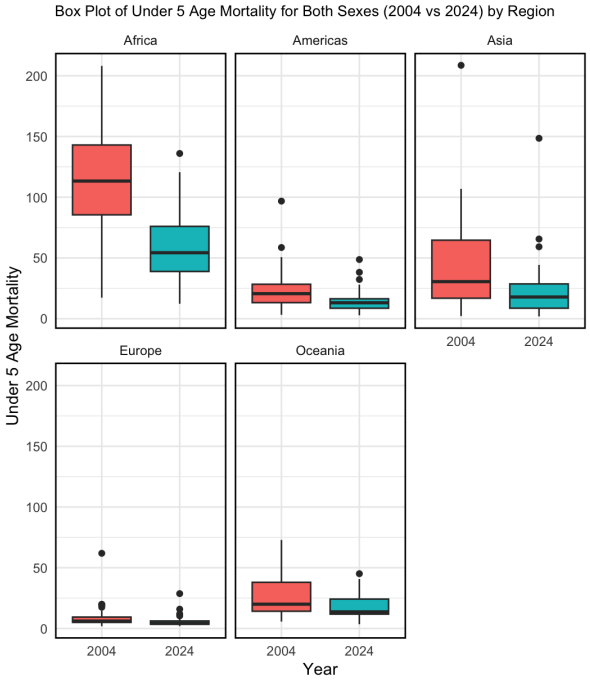


Figure 6: Box plots of under 5 age mortality for both sexes (2004 vs 2024) by region

5 Summary

In this project, demographic data from 227 countries in 2004 and 2024 were used to analyze. The main objectives of this project were to find the possible distribution of observations about fertility rate and life expectancy in different regions and sub-regions. And also compare their frequency distribution for both sexes. Additionally, relationships between variables were examined. Changes in these two parameters from 2004 to 2024 were among the project's objectives. All the given data were interpreted and visualized in figures. The most apparent result of data analysis is that while the fertility rate fell from 2004 to 2024, growth in life expectancy can be observed. Changes in lifestyles, high living costs, and being busier especially having more career women might be mentioned as the substantial reasons for under age 5 mortality rate decline. In addition, improvements in sanitation, medical care, and education might cause a rise in life expectancy. Life expectancy for males and females are very similar to each other and they also have a similar distribution. The only difference is that females in all regions have a higher life expectancy. The life expectancy of different sexes is highly correlated to each other and with a rise in one of them, the other one experiences growth after 20 years.

As mentioned, the mortality rate has reversed, and a negative correlation is observed between under age 5 mortality rate and life expectancy. Another observation is that Eupore has the highest life expectancy among other regions and the highest expectancy in its sub-regions like Northern Europe and Western Europe. On the contrary, African regions' life expectancy is the lowest, and also the under age 5 mortality rate is the highest among all regions. Considering changes over a period of 20 years, the European sub-regions experience the lowest value of mortality rate. For further studies, it would be useful to search for more factors that may affect under 5 mortality rate and life expectancy. For instance, including data about lifestyle, food habits, hygiene, and available facilities in different regions would be interesting. Finally, if observations of all years are available to study changes and predict the future, the interpretation of the result will be more instructive.

Bibliography

Charlie French. Why demographic data matters, 2014. URL https://extension.unh.edu/sites/default/files/migrated_unmanaged_files/Resource004765_Rep6784.pdf.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, Taylor & Francis Group, 2019. ISBN 9781138329256. URL https://books.google.de/books?id=uXy_uAEACAAJ.

Ulrich Kohler and Frauke Kreuter. *Data Analysis Using Stata*. Taylor & Francis, 2005. ISBN 9781597180078. URL https://books.google.de/books?id=4rrsRqUSls8C.

David S. Moore, William I. Notz, and Michael A. Fligner. *The Basic Practice of Statistics*. W.H. Freeman, 8th edition, 2018. ISBN 9781319187601. URL https://books.google.de/books?id=JOMQKI8zj_EC.

Paul Newbold, William L. Carlson, and Betty Thorne. *Statistics for Business and Economics*. Pearson, 8th edition, 2013. ISBN 9780132745659. URL https://books.google.de/books?id=uPOoBwAAQBAJ.

R. Lyman Ott and Michael Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, 2016. ISBN 9781305465513. URL https://books.google.de/books?id=ypGFCwAAQBAJ.

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

George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, 8th edition, 1989. ISBN 9780813815619. URL https://books.google.de/books?id=N09SxgEACAAJ.

U.S. Census Bureau. Glossary, 2021. URL https://www.census.gov/programs-surveys/international-programs/about/glossary.html.

U.S. Census Bureau. International database: World population estimates and projections, 2023. URL https://www.census.gov/programs-surveys/international-programs/about/idb.html.