TU Dortmund

Introductory Case Studies

Project 1 : Descriptive Analysis of Demographic Data

Lecturers:
Prof. Dr. Katja Ickstadt
M. Sc. Zeyu Ding
M. Sc. Yassine Talleb

Author: Bushra Tariq Kiyani

Group number: 7

Group members: Sarmistha Bhattacharyya , Jatin Rattan , Vikas Kumar

May 11, 2023

Contents

1 Introduction 1

2 Problem statement 1

2.1 Data set and data quality . . . 1

2.2 Project objectives . . . 3

3 Statistical methods 3

3.1 Measures of Location/Central Tendency . . . 3

3.2 Measures of Dispersion/Spread . . . 5

3.3 Measures of Correlation . . . 6

3.4 Graphical Representations . . . 6

3.4.1 Histogram . . . 6

3.4.2 Scatterplot . . . 7

3.4.3 Boxplot . . . 7

3.4.4 Pairplot . . . 7

4 Statistical analysis 7

4.1 Frequency Distributions of Variables . . . 8

4.2 Analysis of Variability of the Variables . . . 11

4.3 Bivariate Correlations Analysis . . . 13

4.4 Comparison of Variables from 2002 with 2022 . . . 14

5 Summary 15

Bibliography 16

Appendix 18

1 Introduction

Demographic data analysis is the study of different population groups based on gender, age, interests, level of education, birth rate, death rate, race etc. Demographic data refers to the collection of such information about specific groups of people. Governments, public and nonpublic institutions use demographic information to learn about the characteristics of a population for policy development. It also helps businesses to understand customer needs and future demands. Market research, customized products, recommendation systems and transport systems all are based on demographic data analysis. (Chi and Zhu, 2008)

The main goal of this project is to offer a descriptive analysis of demographic data collected from 227 countries situated in 21 subregions within 5 global regions. This data includes the under age 5 mortality rates and life expectancy, for both men and women separately, and for both sexes collectively, for the years 2002 and 2022. To do this, we first look at the frequency distributions of all variables using histograms. We observe the mean and median of all distributions to compare different distributions. Box plots are used to study variability within and between sub-regions. Then we check dependencies between variables by pair plots. Finally, to observe the changes that occurred in 20 years, we compare the 2002 variables to the 2022 variables.

The second section describes the structure and quality of the data set in more detail. Additionally, the exact goals of the project are stated in the second section. The third section explains the measures of the central tendencies (mean, median), measures of dispersion for example variance and the correlation coefficient. It describes the charts (histograms, boxplot, scatterplot, heatmap, pairplot) used in the statistical analysis. The fourth section focuses on the application of these methods and the interpretation of the graphs. Finally the fifth section summarizes the most important findings and discusses possible further analyses of this data set.

2 Problem statement

2.1 Data set and data quality

This report deals with the analysis of a small sample of the data set for the years 2002 and 2022 taken from the online International Database (IDB) of the U.S. Census Bureau

(US Census Bureau, 2022). The international database contains data from more than 200 countries and areas of the world since the 1960s. The Census Bureau provides population estimates and projections using IDB. The data is gathered through censuses, surveys, administrative records, and vital statistics. IDB is also used for research in education, journalism and business.

The data set consists of 454 observations and ten variables. Description of variables according to the Census Bureau (International Database Glossary, 2021) is given below:

Variable	Type	Description
Country name	Nominal	Name of the country. The data set contains data from 227 different countries.
Sub-region	Nominal	The sub-regions of a country and has total 21 unique values (sub-regions).
Region	Nominal	Region name to which the country belongs. The data set contains 5 regions.
Year	Numeric	The year in which the data was collected. In the data set it is either 2002 or 2022.
Life expectancy of both sexes	Float	Average number of years a cohort can be expected to live if mortality at each age remains constant in the future.
Life expectancy of males	Float	Average number of years a group of males can be expected to live if mortality at each age remains constant in the future.
Life expectancy of females	Float	Average number of years a group of females can be expected to live if mortality at each age remains constant in the future.
Mortality rate of both sexes	Float	Number of deaths of children under 5 years of age from a cohort of 1,000 live births. Denoted 5q0, it is probability of dying between birth and age 5.
Mortality rate of males	Float	Number of deaths of male children under 5 years of age from a cohort of 1,000 live births. Denoted 5q0, it is probability of dying between birth and age 5.
Mortality rate of females	Float	Number of deaths of female children under 5 years of age from a cohort of 1,000 live births. Denoted 5q0, it is probability of dying between birth and age 5.

Table 1: Data Description

There are four instances where the region and subregion names are missing for the countries Curaçao and Côte d'Ivoire. Additionally, for six countries, namely Libya, Puerto Rico, South Sudan, Sudan, Syria, and the United States, there is no data available in the dataset regarding life expectancy and under 5 mortality rate for the year 2002. The US Census Bureau, being an official American agency that regularly updates the IDB with current information, is expected to maintain a high level of data accuracy.

2.2 Project objectives

This project involves carrying out a descriptive analysis of the provided data set. The first step is to analyze the frequency distribution of variables for the year 2022 through histogram visualizations, followed by interpreting the results based on the mean and distribution of the data. Boxplots are used to examine the homogeneity of all variables within and between different subregions for the year 2022. The analysis includes assessing the variability of values within sub-regions and comparing the central values of individual variables between different sub-regions.

Next, bivariate correlations between variables for the year 2022 are examined using a pairplot and their linearity is assessed using Pearson's correlation coefficients. In addition, changes in the under age 5 mortality rate and life expectancy for both sexes between the years 2002 and 2022 are observed using scatterplots.

3 Statistical methods

In this section, statistical methods used for descriptive data analysis are discussed. Statistical measures, coefficients and graphs are presented. For calculation of all statistical measures and graphical representations R software (R Core Team, 2022) Version, 4.2.1 is used with additional packages **GGally** (Schloerke et al., 2021), **dplyr** (Wickham et al., 2022), **ggplot2** (Wickham, 2016), **gridExtra** (Auguie, 2017), **tidyverse** (Wickham et al., 2019).

3.1 Measures of Location/Central Tendency

For continuous numerical data mean, median and mode are measures of location ("Center" of the attribute values $x_1, x_2, \dots, x_n$), which characterize the attribute values of the sample data by a single value. The choice of a measure of location can have a decisive effect on the analysis decision.

Mean: The mean is one of the most common ways of measuring the location of data and is used to quantify the centre of the sampled data. The mean is calculated as the sum of the data values divided by the number of observations, n. Formally, for observations $x_1, x_2, \ldots, x_n$ mean is denoted as $\bar{x}$ and is defined as follows:

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

The mean can be calculated for both discrete and continuous values, but not for categorical data because it is not possible to sum categorical data. Although the sample mean is the most commonly used measure, it is sensitive to outliers, which means it can be heavily skewed toward outliers when they are present. (Crowder et al., 2020, p. 61)

Median: The median (middle number) is another common type of measure of central tendency. It is defined as $50th$ percentile of the ordered data, which means 50% observations fall above and 50% observations fall below the median. For data with extreme values, the median can be a more meaningful measure of central tendency as it is more robust than the mean in the presence of outliers. (Crowder et al., 2020, p. 61)

The median is a value that divides the ordered data into two equal parts. First we sort the data and then choose the middle value if n is odd or we take the average of two middle values if n is even. So it's not necessarily a value in the data set. (Kaptein and van den Heuvel, 2022, p. 15)

Quantile: The idea of the median can be generalized to the quantiles. A q-quantile is a statistical concept that divides a given ordered dataset into q equal-sized groups, with $q - 1$ dividing points between them. Quartiles, deciles, and percentiles are some of the commonly used types of quantiles, with quartiles being 4-quantiles, deciles being 10-quantiles, and percentiles being 100-quantiles. (Kaptein and van den Heuvel, 2022, p. 16-17)

Quartile: Quartiles are a type of quantile, consisting of three values that divide sorted data into four equal parts, with each part containing an equal number of observations. The first quartile, also referred to as the lower quartile or the 25^{th} percentile, is the value that divides the lowest 25% of the data from the remaining 75%, while the second quartile, also known as the median or the 50^{th} percentile, separates the lowest 50% of the data from the top 50%. The third quartile, also called the upper quartile or the 75^{th} percentile, is the value that separates the lowest 75% of the data from the highest 25%. (Kaptein and van den Heuvel, 2022, p. 16-17)

3.2 Measures of Dispersion/Spread

Distributions that have the same location parameters may differ from each other. Therefore we also look at the dispersion of the values $x_1, x_2, \ldots, x_n$. The measure of dispersion is actually a measure of uncertainty. It quantifies the range of data.

Variance and Standard Deviation: Variance is a popular measure of spread and it quantifies the spread between values of the data set. It is defined as the average of the squared deviations from the mean. Let s^2 denote the variance and $\bar{x}$ the mean of the data set then the variance is calculated as:

$$s_x^2 = \mathrm{var}_x := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The sample standard deviation is the square root of the sample variance, denoted by s_x. $s_x := \sqrt{\mathrm{var}_x}$. The standard deviation is the most commonly used measure because it is on the same scale as observed data instead of a squared scale. (Crowder et al., 2020, p. 62)

Range: The range is the simplest measure of dispersion. It is the difference between the maximum (x_n) and minimum (x_1) value in the ordered sample data. The range (R_x) is calculated as:

$$R_x := \max(x) - \min(x) = x_n - x_1$$

The range is the most sensitive to outliers. Secondly, it only tells the difference between the maximum and minimum values in the data set and does not provide any information about the spread of the all other values in the data set. (Kaptein and van den Heuvel, 2022, p. 17)

Interquartile range (IQR): As the range is sensitive to the outliers, the Interquartile range (IQR) is another robust measure of spread which is defined as the difference between the third quantile $Q_{0.75}$ and the first quantile $Q_{0.25}$. It measures the range in which 50% of the middle data falls. (Kaptein and van den Heuvel, 2022, p. 17)

$$\mathrm{IQR} := Q_{0.75} - Q_{0.25}$$

The interquartile range is also visualized in a box plot, which we will discuss later in this section.

3.3 Measures of Correlation

The correlation quantifies the strength of the relationship between two variables. The Pearson correlation coefficient r_{xy}, a number ranging from -1 to 1, measures the linear dependency between two continuous variables.

Pearson Correlation Coefficient: The Pearson correlation coefficient r_{xy} is the most commonly used correlation coefficient which measures the strength of the linear relationship between two continuous variables. Assume x_i and y_i as individual observations, $\bar{x}$ and $\bar{y}$ means of variable x and y then the Pearson correlation coefficient r_{xy} is calculated as follows:

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

$r_{xy} = 1$ if y and x lie on a straight line with a positive slope which shows a strong positive correlation. and $r_{xy} = -1$ if the slope is negative, showing a strong negative correlation. While $r_{xy} = 0$ implies no correlation. Correlation is undefined if x or y is not varying. (i.e. s_x or $s_y = 0$). (Crowder et al., 2020, p. 62)

3.4 Graphical Representations

Data visualization is useful to analyze data distribution, outliers in the data set, trends, correlations and unusual variations. We choose graphical tools according to the type of data.(Crowder et al., 2020, p. 63) In the following section different graphical methods used in data analysis are represented.

3.4.1 Histogram

A histogram is a graphical chart, commonly used to show frequency distribution through vertical bars. To create a histogram we divide the variable range into equally spaced intervals which are placed on the x-axis and the y-axis shows the total number of data points in each interval. The height of the bars shows the frequency of data points in each interval. A histogram provides a graphical representation for location measurements and data dispersion and is also helpful in identifying outliers. It also provides significant information about the distribution of the variable. (Healy, 2019, p. 80-90)

3.4.2 Scatterplot

A scatterplot is a chart which plots individual data points as dots. It is used to visualize the relationship between two variables. It gives a visual representation of the correlation between the values of two variables. We can see if the relationship is linear or curved by looking at the linearity of the data points on the chart.

In a scatterplot data points that start from the bottom up show a positive correlation, while the spread from the top down shows a negative correlation. Correlation is said to be positive when the values of two variables increase together, and negative when the values of one variable increase when the other decreases. (Healy, 2019, p. 2-5)

3.4.3 Boxplot

A boxplot is another graphical representation which represents a five-number summary (Minimum, 25%−Quantile, Median,75%−Quantile, Maximum). This provides a visual representation of the location, spread and variation of the variables. In a boxplot, middle line is the median and the box itself with a lower border at 25%−Quantile and an upper border at 75%−Quantile indicates the middle 50% of the data.

Whiskers in "standard boxplot" show the smallest (x_1) and largest values (x_n). While in "modified boxplot" whiskers are drawn from smallest to largest values in the interval $[Q_{0.25} - 1.5\text{IQR}, Q_{0.75} + 1.5\text{IQR}]$. Observations outside this range are said to be potential outliers. Boxplots are useful when we want to compare a continuous variable between two subgroups. (Healy, 2019, p. 104-107)

3.4.4 Pairplot

A pairplot plots the pairwise relationships between the variables through a scatter plot in a matrix form. This creates a compact and nice visualization that is helpful for understanding the data at a glance in one plot. (Healy, 2019, p. 170-171)

4 Statistical analysis

This section presents a comprehensive descriptive analysis of the data set using statistical measures and plots described in the previous section. The missing subregions and regions for two countries, namely Curaçao and Côte d'Ivoire, are identified through a brief search and added to the data set. Curaçao is classified under South America, Americas, while Côte d'Ivoire belongs to Western Africa, Africa. The missing values in the six countries are addressed by substituting the median instead of the mean, as the variables in question exhibit skewed distributions. Table 2 on page 19 of the appendix summarizes the descriptive statistical information about the variables in the data set.

4.1 Frequency Distributions of Variables

The frequency distribution of all the variables in the year 2022 is analyzed by using histograms. Figure 1 shows histograms of all four variables in the data set. On the x-axis class intervals of the respective variable are shown while the y-axis shows the frequency of each interval.

Figure 1(a) shows the frequency distribution of female and male life expectancy. These distributions are skewed to the left, indicating that the frequency is not evenly distributed on either side of the mean value. As a result, both distributions are asymmetric. The mean and median values differ for females and male. Specifically, for female, the mean is 77.18 and the median is 78.69, whereas for male, the mean is 72.10 and the median is 73.26. Which shows average life expectancy is higher for females compared to males. The frequency distribution shows that the interval with the highest frequency for female is between $76 - 86$, while for male, it is between $71 - 81$. Additionally, there is an outlier in the distribution where the life expectancy for female exceeds 90. This outlier corresponds to Monaco, where the life expectancy for female is 93.49, Monaco also has the highest male life expectancy of 85.70. The lowest life expectancy, on the other hand, is observed in Afghanistan, where the female and male life expectancy is 55.28 and 52.10, respectively.

The frequency distributions in Figure 1(b) display the under age 5 mortality rates for females and males, both of which exhibit a right-skew. As a consequence of this skewness, the mean values are inflated and tend to overestimate the true central tendency. Again a difference in the mean and median values between females and males is observed. More specifically, the mean value for females is 24.012 and the median is 13.620, whereas for males, the mean is 29.23 and the median is 17.55. Which shows average under age 5 mortality rate is less in females than males. There are 55 countries with under age 5 mortality rates between $1 - 6$ for females, while for males, the corresponding number is

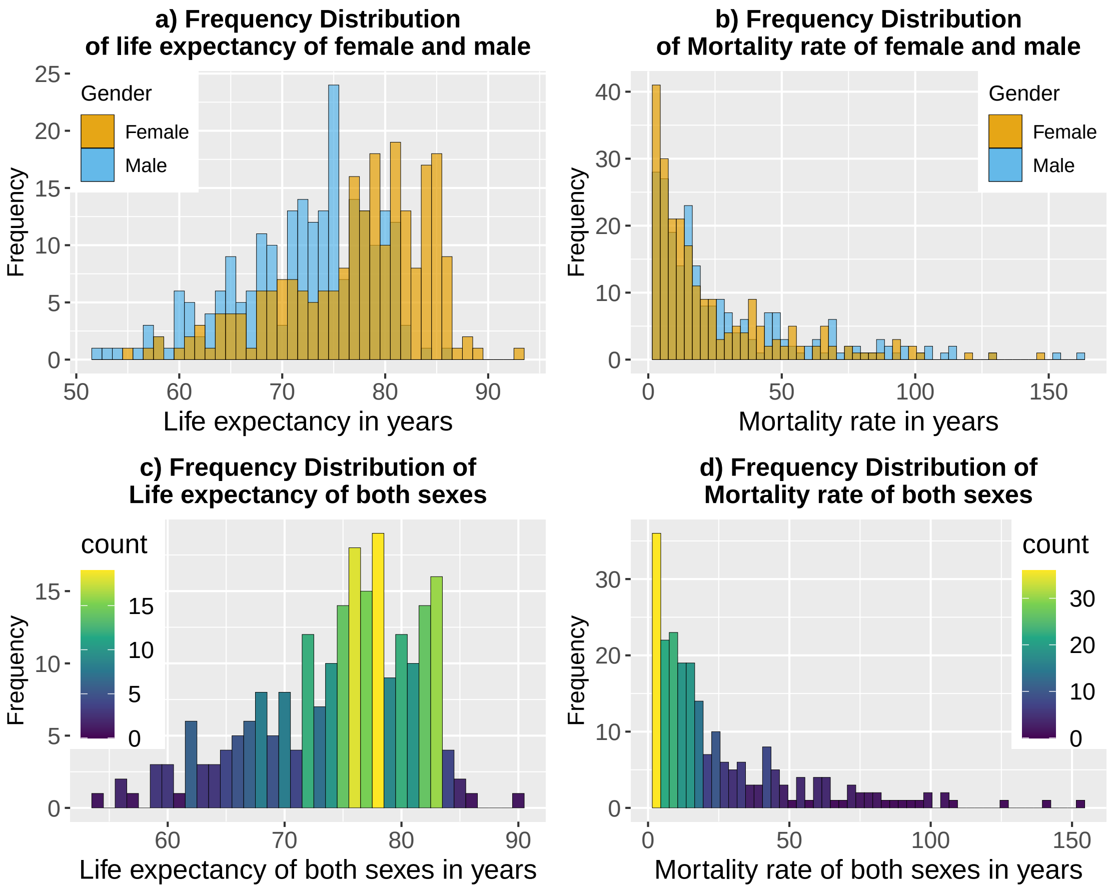

Figure 1: Histograms for (a) Life expectancy at birth of females and males, (b) Under age 5 Mortality rate of females and males, (c) Life expectancy at birth of both sexes, (e) Under age 5 Mortality rate of both sexes

46. The country with the lowest under age 5 mortality rate is Monaco, where the rates are 1.64 and 2.31 for females and males, respectively. Moreover, there are 4 countries with a female mortality rate exceeding 100, while 9 countries have a male mortality rate exceeding 100. Afghanistan has the highest mortality rate of 146.09 and 161.78 for females and males, respectively. For Somalia, the under age 5 mortality rate is the second highest, with rates of 128.81 for females and 153.23 for males.

Figure 1(c) represents the frequency distribution of life expectancy of both sexes. The distribution of life expectancy for the entire population is left-skewed, with the majority of values falling in the upper-middle range and very few in the upper and lower range. Specifically, the life expectancy rate for most countries ranges between $74 - 84$. Only Afghanistan falls within the lowest range of $50 - 55$, with a life expectancy rate of 53.65.

On the other hand, two countries, Singapore (86.35) and Monaco (89.52), have a life expectancy rate in the highest range of $85 - 90$.

The frequency distribution of under age 5 mortality rate for both sexes is shown in Figure 1(d). The distribution is skewed to the right. About 41 countries have under age 5 mortality rates between $1 - 5$. Additionally, there are five countries with mortality rates between $100 - 150$, out of which Somalia and the Central African Republic have the highest mortality rates at 141.20 and 124.58, respectively. An extreme deviation can be seen in the graph it is Afghanistan, which has the highest under age 5 mortality rate 154.13.

There are 55 countries in Africa, 50 in the Americas, 52 in Asia, 49 in Europe, and 21 in Oceania. Figure 2(a) shows the boxplots for life expectancy of both sexes by region. The life expectancy range in Africa is broader, from 53 to 81, than in other regions, indicating greater variation in life expectancy. The median life expectancy (65.85) in Africa is the lowest among all regions, suggesting that people in African countries have, on average, shorter lifespans than those in other regions. The distribution in the Americas is left-skewed, with the second-largest median (77.9) and less variability compared to Africa, Asia, and Europe. An outlier in the Americas is Haiti, which has the lowest life expectancy of 65.95 within the region.

The median life expectancy in Asia is the third highest at 75.765. Afghanistan is the outlier in this region with the lowest life expectancy at 53.65. The distribution of life expectancy in Europe is left-skewed with the highest median (81.51) among all other regions, indicating that people in European countries, on average, have a higher life expectancy than those in other regions. Oceania has fewer countries and less variation in data compared to other regions. The median (75.32) in Oceania is very similar to Asia. Australia (83.09) and New Zealand (82.54) are outliers showing the highest life expectancy rates, while Kiribati (67.90), Nauru (67.93), and Tuvalu (68.38) are outliers showing the lowest life expectancy rates.

Figure 2(b) shows the boxplots for under age 5 mortality rate of both sexes by region. The spread in Africa is broader, than in other regions, indicating greater variation in under age 5 mortality rate. Distribution is right-skewed with the median mortality rate (59.16) in Africa is the highest among all regions, suggesting that people in African countries have, on average, larger under age 5 mortality rate than those in other regions. The distribution in the Americas is right-skewed, with the second-lowest median (13.43) and less variability compared to Africa and Asia. Outliers in the Americas are Haiti

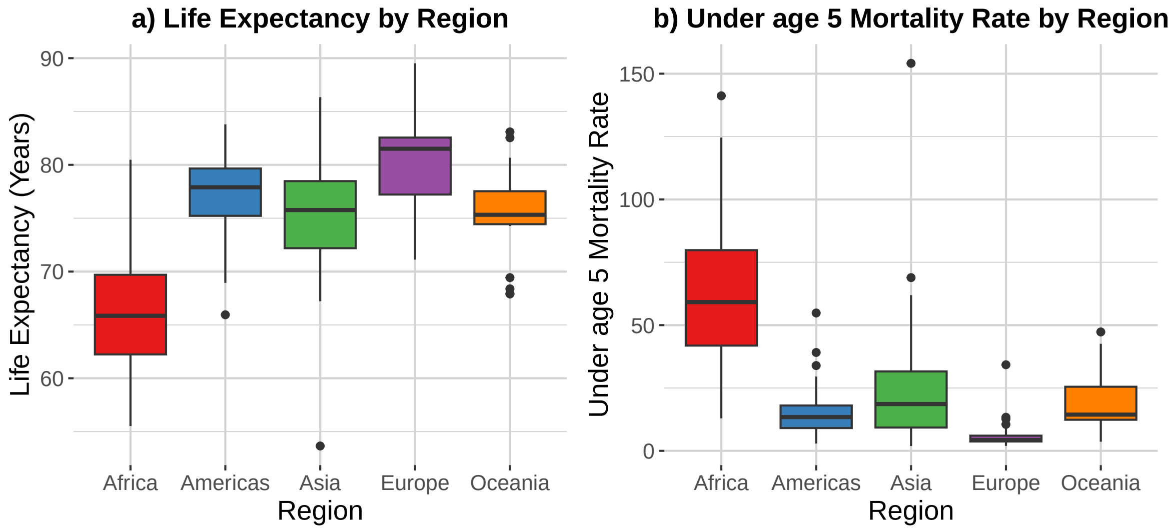

Figure 2: Boxplots for (a) Life expectancy for both sexes by Region, (b) Under age 5 Mortality rate for both sexes by Region

(54.84), Suriname (39.14) and Guatemala (33.92) which have the higher mortality rates within the region. Distribution in Asia is also right-skewed with the median 18.62. Afghanistan (154.13) and Pakistan (68.90) are the outliers in this region with the highest mortality rate. The distribution in Europe is also right-skewed with the lowest median (4.32) among all other regions, indicating that people in European countries, on average, have a lower under age 5 mortality rate than in other regions. In Oceania the distribution is also right-skewed and the median (14.42) is third lowest. Tuvalu (47.29) shows the highest under age 5 mortality rate within the region.

4.2 Analysis of Variability of the Variables

This section discusses the homogeneity and heterogeneity of variables within and between subregions, utilizing boxplots to visualize their variability. As shown in Figure 3, data spread in Western Europe is minimal, resulting in a small interquartile range (IQR) and small box size, indicating that the data is homogeneous within the subregion for life expectancy (male/female/both). Each variable has one outlier, Monaco, which exhibits the highest life expectancy for males (85.70), females (93.49), and both sexes (89.52). The median for life expectancy (both sexes) is situated in the middle, suggesting a symmetric distribution. However, for females, the distribution is left-skewed, while for males, it is slightly right-skewed.

Northern Europe exhibits a high degree of homogeneity, as evidenced by the minimal spread. However, the distribution of life expectancy for males is left-skewed, while for females and both sexes, it is right-skewed. Outliers in female life expectancy are Latvia (80.56) and Lithuania (81.44), in male life expectancy, they are Lithuania (70.42), Latvia (71.47) and Estonia (73.25) while in both sexes, they are Lithuania (75.78) and Latvia (75.91). In terms of median and box sizes (IQR), Western Europe and Northern Europe do not differ a lot. Therefore, we can conclude that the variables are homogeneous between these two subregions.

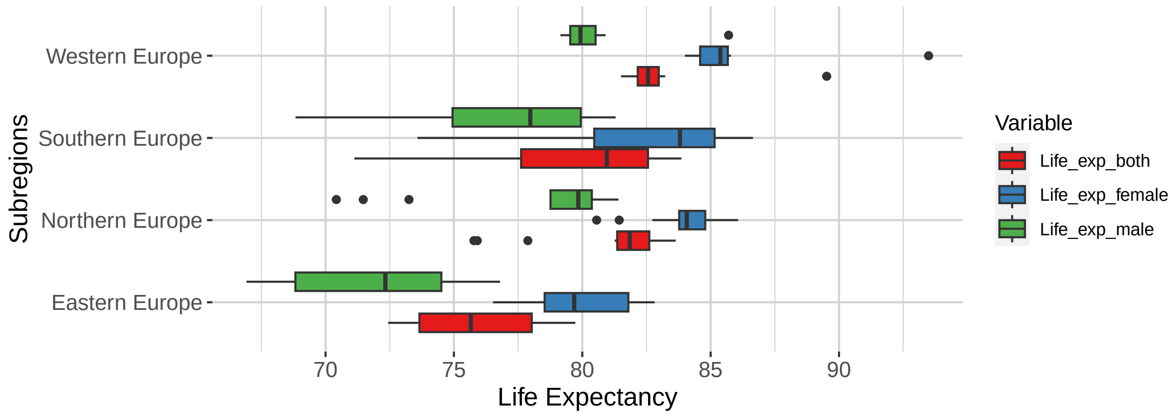

Figure 3: Boxplots illustrating the distribution of life expectancy for Europe in 2022

In Southern Europe, the data spread for all variables is substantial, and each variable exhibits a left-skewed distribution. But median for all variables does not differ noticeably from Northern and Western Europe. In Eastern Europe, the spread is high, particularly for male life expectancy, with a left-skewed distribution, while for females, it is right-skewed. Additionally, the median and IQR exhibit considerable differences between Eastern Europe and all other subregions. Compared to other subregions, we can say that the variables are heterogeneous between Eastern Europe and all other subregions.

The under age 5 mortality rates for Europe's subregions are illustrated in Figure 4. The spread for Western and Northern Europe is minimal, and the median is nearly identical, indicating homogeneity within and between the subregions. Conversely, Southern and Eastern Europe display relatively high spread, although the median of all variables does not differ noticeably in all subregions.

A few outliers are also observed. In Southern Europe, an extreme outlier with the highest mortality rates is Kosovo, where for males, the mortality rate is 36.73, for females 31.58, and for both genders, it is 34.25.

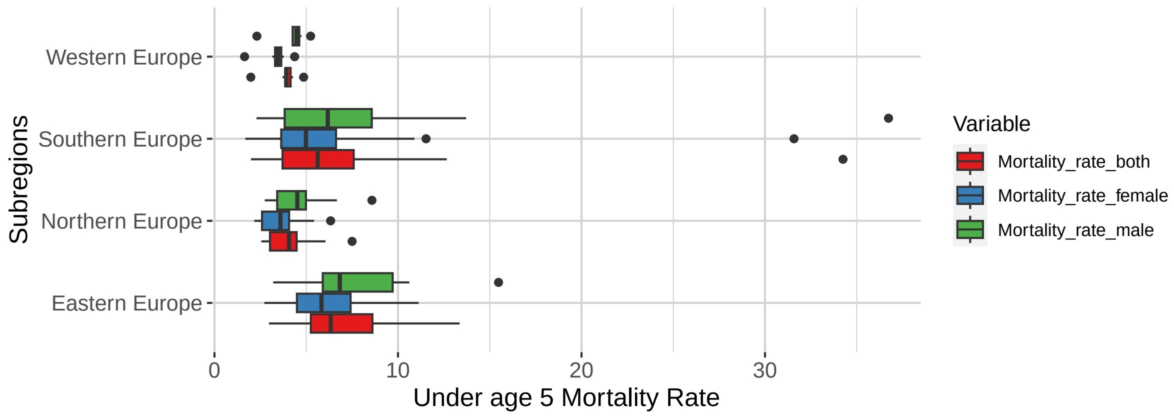

Figure 4: Boxplots illustrating the distribution of Mortality rate for Europe in 2022

4.3 Bivariate Correlations Analysis

This section examines the interdependencies between all variables in 2022. To achieve this, a pairplot is generated as shown in Figure 6 on page 18 of the appendix. The lower half of the chart displays scatter plots demonstrating the relationships between all variables, whereas the upper half shows the values of the Pearson correlation coefficient. Since the relationships between variables are linear, the Pearson correlation coefficient is being used. Additionally, region-wise correlation coefficient values are also displayed in plots to depict the strength of the dependencies between variables across different regions.

By observing the scatter plots and negative correlation coefficient values, it is evident that the under age 5 mortality rate is negatively correlated with life expectancy. The large correlation coefficient values overall indicate a strong negative correlation. This implies that as life expectancy (male/female/both) increases, the under age 5 mortality rate (male/female/both) decreases and vice versa. For Europe, the correlation coefficient values between under age 5 mortality rate (male/female/both) and life expectancy (male/female/both) are moderate or weak, indicating that life expectancy does not have a strong correlation with the under age 5 mortality rate in Europe. However, on average, for Asia, these coefficient values are the highest among all other regions. Additionally, the correlation between under age 5 mortality rate (male/female/both) and female life expectancy is stronger than the correlation with life expectancy (male/both).

The scatter plots and coefficient values indicate a robust positive linear correlation between life expectancy for both sexes and the life expectancy of males and females separately. The high correlation coefficient values observed across all regions suggest a strong positive correlation. Additionally, there is a strong positive linear correlation between female and male life expectancy. Similarly, the under age 5 mortality rate for both sexes shows a positive correlation with the under age 5 mortality rate of males and females, as evidenced by high coefficient values. Moreover, under age 5 mortality rates for males and females are also strongly positively correlated.

4.4 Comparison of Variables from 2002 with 2022

Figure 5(a) and 5(b) illustrate the comparison of life expectancy of both genders and mortality rate in all sub-regions in the year 2002 with 2022. On the x-axis values of 2002 are shown while the y-axis shows the values from 2022. The diagonal line is an indicator of the change. Points lie below the line representing the values of the variables which decreased in the year 2022, and points lie above the diagonal line indicate an increment in the values of the variables in the year 2022. Points lie on the line that show no change has occurred in the values of variables in 2022 compared to 2002. Different colors indicate different regions while individual data points indicate countries.

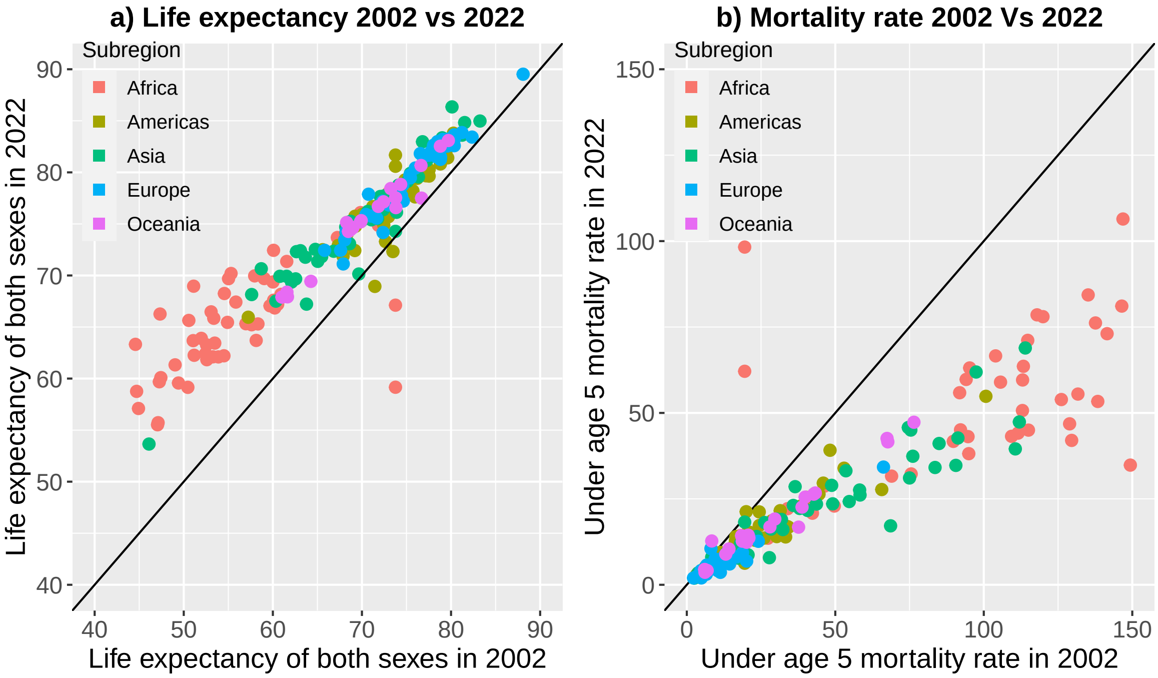

Figure 5: (a) Variation in life expectancy between 2002 and 2022, (b) Variation in mortality rate between 2002 and 2022

In Figure 5(a) it can be observed that a majority of the points are located above the diagonal line, indicating an increase in life expectancy for most countries in 2022 compared to 2002. However, there are a few exceptions where the life expectancy has decreased in 2022 as compared to 2002. The countries South Sudan and Sudan in Africa, as well as Mexico and Peru in the Americas, have experienced a decline in life expectancy.

Figure 5(b) shows a decreasing trend in the under age 5 mortality rate of both sexes, indicating an overall decline in mortality rates in 2022 as compared to 2002. Conversely, six countries - South Sudan, Sudan, Panama, Croatia, and Guam - experienced an increase in mortality rate in 2022 when compared to 2002.

5 Summary

This project involved conducting a descriptive analysis of data from the US Census Bureau (US Census Bureau, 2022), which included information on 227 countries. In this analysis frequency distributions of all variables, bivariate correlations, variability of variables, and changes in mortality rate and life expectancy from 2002 to 2022 are examined.

The study indicated that, frequency distribution of life expectancy is left-skewed while for mortality rate it's right-skewed. Analysis revealed that in nearly all countries, women have a higher life expectancy and lower mortality rate than men. This could potentially be attributed to behavioral or biological differences between the both. The analysis showed a negative correlation between under age 5 mortality rate and life expectancy.

In 2022 a general upward trend in life expectancy compared to 2002 was observed, indicating an overall increase. While, the mortality rate showed a decreasing trend in 2022 compared to 2002, indicating a decrease in mortality. Africa had the higher mortality rates and the lower life expectancy compared to other regions, while Europe had lower mortality rates and higher life expectancy. This may be due to the difference in socioeconomic conditions in the two regions. Afghanistan was observed as an extreme outlier, which had the highest death rate and lowest life expectancy of any country.

For further analysis, more information about men and women would provide insight into individual differences in life expectancy between men and women. Additionally, we can add more variables in the study to reveal the factors that change mortality rate and life expectancy.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3.

Guangqing Chi and Jun Zhu. Spatial regression models for demographic analysis. 2008. URL `https://rdcu.be/daMLahttps://rdcu.be/daMLa`.

Stephen Crowder, Collin Delker, Eric Forrest, and Nevin Martin. *Introduction to Statistics and Probability*, pages 59–80. Springer International Publishing, Cham, 2020. ISBN 978-3-030-53329-8. URL `https://doi.org/10.1007/978-3-030-53329-8`$_4$.

Kieran Healy. *Data Visualization, A PRACTICAL INTRODUCTION.* Princeton University Press, 41 William Street, Princeton, New Jersey 08540 6 Oxford Street, Woodstock, Oxfordshire OX20 1TR, 2019. ISBN 978-0-691-18161-5.

International Database Glossary, 2021. URL `https://www.census.gov/programssurveys/international-programs/about/glossary.html`. Accessed: 27-04-2023.

Maurits Kaptein and Edwin van den Heuvel. *A First Look at Data*, pages 1–37. Springer International Publishing, Cham, 2022. ISBN 978-3-030-10531-0. URL `https://doi.org/10.1007/978-3-030-10531-0`$_1$.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally.

US Census Bureau. International database: World population estimates and projections, 2022. URL `https://www.census.gov/programs-surveys/international-programs/about/idb.html`. Accessed: 27-04-2023.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Appendix

A Additional figures

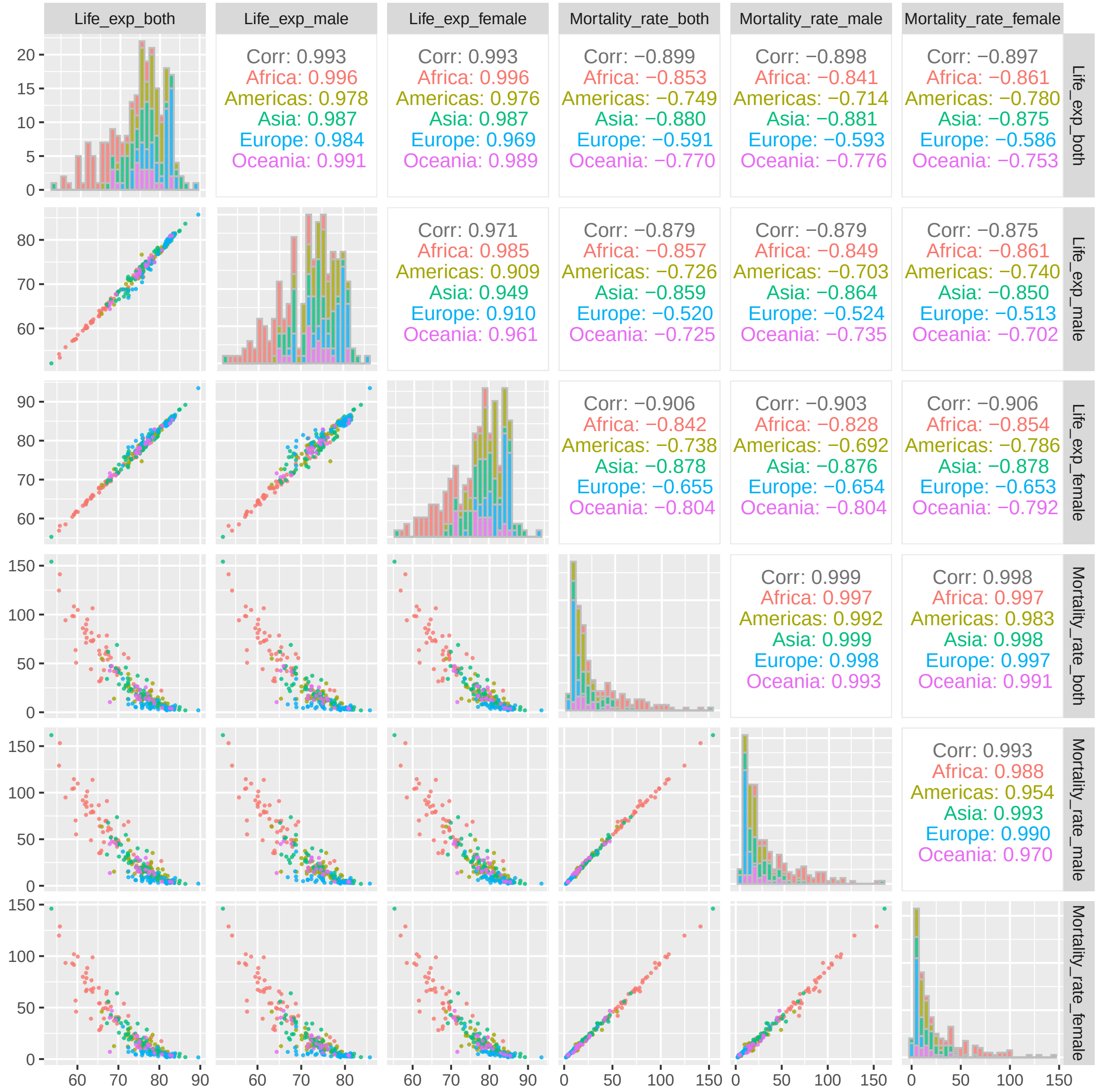

Figure 6: A pairplot displaying the pairwise relationships between all variables in 2022.

B Additional tables

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Life_exp_both	53.65	70.05	75.82	74.58	79.66	89.52
Life_exp_male	52.10	67.93	73.26	72.10	77.19	85.70
Life_exp_female	55.28	72.63	78.69	77.18	82.56	93.49
Mortality_rate_both	1.940	7.415	15.080	26.677	37.775	154.130
Mortality_rate_male	2.03	8.32	17.55	29.23	41.02	161.78
Mortality_rate_female	1.640	6.345	13.620	24.012	34.470	146.090

Table 2: Summary Statistics 2022