TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Dr. rer. nat. Maximilian Wechsung
M. Sc. Hendrik Dohme

Author: Kartik Kamboj

Group number: 1

Group members: Bidhan Chandra Roy, Hemani Chand, Jatin Rattan, Kartik Kamboj

May 6, 2022

Contents

1 Introduction

The International Data Base is an online data bank that contains demographic information for countries and regions throughout the world with populations of 5000 or more people. The precise information offered by the IDB is useful for research, program planning and policy decision-making all around the world. As a result, the level of detail given in the IDB provides solid foundations for tracking the demographic effects of global events that affect people all across the world, such as natural catastrophes and the HIV/AIDS pandemic.

The objective of this project is to use proper statistical measurements and graphical approaches to analyze and visualize a small sample of data from the IDB, which includes life expectancy at birth for males and females, as well as total fertility rates. From 2002 through 2022, the retrieved data covers 227 countries grouped into 5 geographical areas and 21 subregions. The methods that need to be followed are explorative and descriptive.

First, we look at the frequency distribution of male and female life expectancy at birth, as well as total fertility rates, taking into account gender disparities. Second, we must examine the correlations between the values of the variables and determine variabilities within individual subregions or across subregions. Finally, we must look at how the variables have changed over the last 20 years (from 2002 to 2022).

In the second section, the structure of the demographic data contained in the dataset is discussed in greater detail, as well as an overview of the presented dataset. The third section shows how to generate and understand histograms, scatter plots and box plots among other statistical and graphical tools. The application of statistical methods to the dataset is given in the fourth section and the findings are interpreted. Finally, the fifth section finishes with a summary of the findings as well as suggestions for additional data analysis.

2 Problem statement

2.1 Data set and data quality

The following data set is an extract from the United States Census Bureau's International Data Base (The U.S. Census Bureau, 2022), which includes data from the years 2002 and 2022. For 5 regions and 21 subregions, the data covers total fertility rate, life expectancy at birth for both sexes, life expectancy at birth of men and life expectancy at birth of females. We have four categorical variables (Country, Region, Subregion and Year) and four continuous variables (Total Fertility Rate, Life Expectancy at Birth for Both Sexes, Life Expectancy at Birth for Males and Life Expectancy at Birth for Females). Total Fertility Rate is defined as the average number of children born per woman if all women lived to the end of their childbearing years and had children according to a set of age-specific fertility rates. Life expectancy is defined as the average number of years that a group of people born at the same time may expect to live if death rates at each age remain constant.

The entire data set is used for statistical analysis. Country, Region, Subregion, and Year are all categorical variables. The countries and subregions are based on the Asia, America, Africa, Europe, and Oceania regions. The year variable is divided into two categories: 2002 and 2022. The total fertility rate is estimated to four decimal places, whereas the life expectancy at birth rate is estimated to two decimal places.

The data set has 6 missing records in the year 2002 and are not included in the analysis. These missing values are in the continuous variables of the data i.e., Total fertility Rate and all Life Expectancy variables. The data quality here is generally good.

2.2 Project Objectives

Initially, to have a glance at the whole data set a graphical representation using frequency distribution considering inequalities between men and women is depicted by the histograms. Secondly, we need to find the dependency of one variable on another. Moreover, we need to determine whether the correlation between variables is monotonic or not. If it is monotonic, is it linear or nonlinear? To determine such relationships between variables, a scatter plot is used which describes the trend more precisely. In addition, the third task requires assessing the variability of all subregions of an individual region and comparing the values between subregions to determine uniformity and non-uniformity. Comparing boxplots from different subregions in terms of factors such as total fertility rate and life expectancy of both sexes is one way to do this. Finally, we must determine the change in the variables over a 20-year period from 2002 to 2022, by comparing the values of each variable for the years 2002 and 2022.

3 Statistical methods

In this section, the statistical methods are given which are later used for analyzing the data set.

3.1 Mean, Median and Quartiles

3.1.1 Mean

Mean is the sum of all observations divided by the total number of observations in the data collection. The arithmetic mean ($\bar{x}$) of n observations x_1, x_2, x_3, ..., x_n is (Black, 2006, p. 49-50)

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

3.1.2 Median

A median is a statistical measure that determines the middle value of a dataset when values are evaluated relative to each other but not defined absolutely. When there are a significant number of outliers, it is one of the most commonly used measures of central tendency for quantitative data since it is unaffected by outliers. (Black, 2006, p. 48-49)

3.1.3 Quartiles

A quartile is another central tendency measure that separates data into four equal parts in order to summarize a large list into a few figures that may be used to gain a good view of the data. It provides information on the data's center, range and spread. (Black, 2006, p. 52-53)

3.2 Histograms

Histograms are graphs that show discrete or continuous data in a graphical format. The frequency distribution is represented graphically in this diagram. The data is divided into class intervals using a rectangle. On the X-axis, a rectangle is drawn. On the Y-axis, data's frequency is plotted. Each rectangle represents the number of frequencies included in each class interval. On the horizontal axis, response variables are defined, and you must determine their frequency.

The histogram can be used to investigate data distributions such as peaks, spreads, and distribution shapes. Determine the highest bar on the chart, which is the spike. The most common values in your data are determined by peaks. The range of the data to explain variation is referred to as the spread. The distribution can take the form of a symmetric, asymmetric, uniform, random, or normal distribution. When data is skewed, it is almost always on one side. When the majority of the sample data is grouped on the left side of the histogram, the data is right-skewed. When the majority of the sample data is grouped on the right side of the histogram, the data is left-skewed. (Hay-Jahans, 2019, p. 131-137)

3.3 Box plot

The box and whisker plot, simply called the box plot, is a graphical representation of statistical data. It doesn't show the distribution in as much detail as histogram does, but it indicates whether the distribution is skewed and are there any outliers in the data set.

The left and right sides of the box are the first and third quartile(hinge). The first quartile (Q1) is a point where 25% of the data values are lower than it and 75% of the data values are higher than it. The third quartile (Q3) is a point where 75% of the data values are lower than it and 25% of the data values are higher than it. The line that splits the box in two is the median or often referred to as second quartile (Q2). It is the middle value of the dataset. The length of the box corresponds to the interquartile range (IQR), where 50% of the data is found [IQR=Q3-Q1]. The smallest value in the dataset is known as minimum and the largest value in the dataset is known as maximum. The lines outside the box that go from the minimum to the lower quartile and then from the upper quartile to the maximum, are known as whiskers. The values which lie outside these whiskers are known as outliers.

The difference between distinct areas of the box determines data spread and skewness. The longer the box means data is more dispersed. The smaller the box means data is less dispersed. Boxplots can be drawn horizontally or vertically. (Hay-Jahans, 2019, p. 137-142)

3.4 Pearson's Correlation Coefficient

The term correlation when applied to two continuous random variables X and Y refers to the existence of a linear relationship between the two variables in common usage. Pearson's correlation coefficient is utilized to reflect the strength and kind of any existing linear relationship. The Pearson's correlation coefficient (r) can be defined by using formula:

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

where, n is the size of the sample, x_1, x_2, x_3, ..., x_n and y_1, y_2, y_3, ..., y_n are the individual samples of variables X and Y respectively and $\bar{x}$ is the mean of samples of X and $\bar{y}$ is the mean of samples of Y. The correlation coefficient value, r, is always between -1 (strong negative relationship) and +1 (strong positive relationship). A value of 0 indicates that the variables are not linearly related. (Hay-Jahans, 2019, p. 321-325)

3.5 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient evaluates the strength of monotonic relationship between two random variables X and Y. In a monotonic relationship, variables change together but not at the same rate. The formula to predict monotonic relationship between bivariate samples (x_i, y_i) of two random variables X and Y is

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where, r_S is Spearman's rank correlation coefficient, d is the pairwise distances of the ranks of the variables x_i and y_i and n is the number of samples. Its interpretation is similar to Pearsons, i.e., the closer r_S is to +1 or -1 the stronger the monotonic relationship and 0 indicates no association. (Hay-Jahans, 2019, p. 329-331)

3.6 Scatter plot

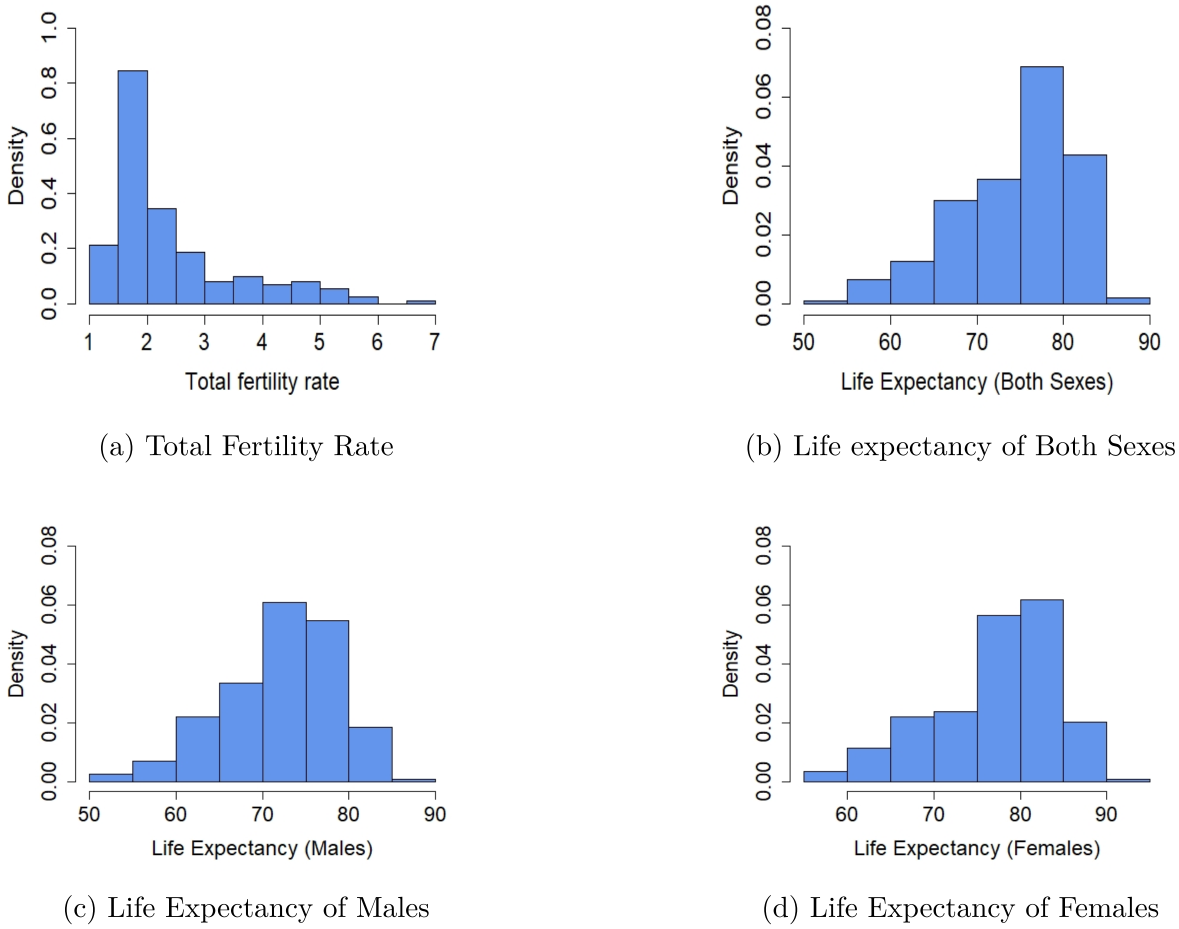
A scatter plot is a type of data visualization that uses points to represent the values of two separate variables. The values for each data point are indicated by the position of each point or dot on the horizontal or vertical axis. They're used to look at how variables relate to one another. In a scatter plot, each point has two coordinates. The first is the X coordinate, which indicates how far left or right you go. The Y coordinate, or the amount you go up or down, is the second. The intersection of the two coordinates is where the observation point is located. If the data shows an upward trend as you move from left to right, this implies that X and Y have a positive association. The Y value rises in lockstep with the X value. If the data shows a downward trend as you move from left to right, this implies that X and Y have a negative connection signifying that the Y value tends to drop as the X value rises. There is no association between X and Y if the data do not reveal any form of pattern. (Hay-Jahans, 2019, p. 159-169)

4 Statistical analysis

In this section, all the discussed statistical methods are used to perform exploratory and descriptive analysis on the data set. All the following graphs were generated using the software R (R Core Team, 2020), version 4.0.5. The results of all the parts are given in the following section. The first three sections contain data only from the year 2022 but for the fourth section we are using the entire data set. The R packages used in the project for the graph plots are tidyverse (Wickham, 2021b), dplyr (Wickham et al., 2021), ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017) and forcats Wickham (2021a).

4.1 Frequency Distribution

The frequency distribution of all quantitative variables from our data set is described in this subsection. The frequency distribution of total fertility rates in 227 countries is depicted in Figure 1(a). For the given data set, it shows that more than 80% of the population density has a fertility rate between 1.5 and 2 in the year 2022. Furthermore, the distribution appears to be right-skewed, i.e., positive skew, following visualization. Whereas, as seen in Figure 1(b), the distribution of life expectancy is skewed to the left. Life expectancy for both sexes gradually rises until it reaches its peak in the range of 75-80 years. It began to decline after the age range of 80-85 years and finally reached its lowest point in the age range of 85-90 years. Figures 1(c) and 1(d) show that females live longer than males, with life expectancies of 80-85 years and 70-75 years, respectively.



(a) Total Fertility Rate

(b) Life expectancy of Both Sexes

(c) Life Expectancy of Males

(d) Life Expectancy of Females

Figure 1: Histogram Plots

Figure 2 combines both male and female life expectancy in a single panel to better understand the disparities between the sexes. From the graph, we can see that life expectancy at birth of males is less than that of females. So, we can conclude that life expectancy of females is better than life expectancy of males which indicates females tend to live longer than males.

4.2 Dependency and monotonicity of the variables

In the following subsection, Pearson's correlation coefficient is used to determine the correlation between variables as well as whether they are dependent on each other. Furthermore, to get an idea about the monotonicity between variables, we use Spearman's rank correlation coefficient, which also hints towards nonlinear relationships.

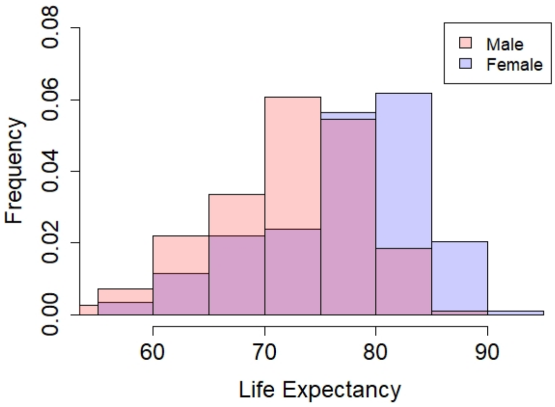


Figure 2: Comparison of Life Expectancy of both the sexes

According to Figure 3, the scatter plots show that all three life expectancy variables have a strong positive correlation or are highly dependent because their Pearson's correlation coefficient values are very close to 1, but the total fertility rate has a negative correlation and also less dependent with the other life expectancy variables because its Pearson's correlation coefficient values are not very close to -1. This means that as the total fertility rate rises, the population's life expectancy falls.

Consider Figure 8 on page 16 in the Appendix, the corresponding Spearman's correlation coefficients values between three life expectancy variables, shows that they have a strong perfect monotonic relationship with each other, which is very close to 1. The total fertility rate does not have a perfect monotonic relationship with the other life expectancy factors because the correlation coefficient values for total fertility rate is in the range of -0.7 to -0.8 with all life expectancy variables, which is not very close to -1. Since, the correlation coefficient values of total fertility rate and other life expectancy variables from both methods is of comparable magnitude this indicates that there is no trace of nonlinear relationship between variables.

4.3 Variables' variabilities within and between subregions

In this subsection, we estimate the variability of total fertility rate and life expectancy at birth for both sexes as well as individual sex using box plots. We are plotting a box plot for each sub-region and sorting it by region with the same color code. The variability of the variable is interpreted with the help of the length of the Interquartile range (IQR)

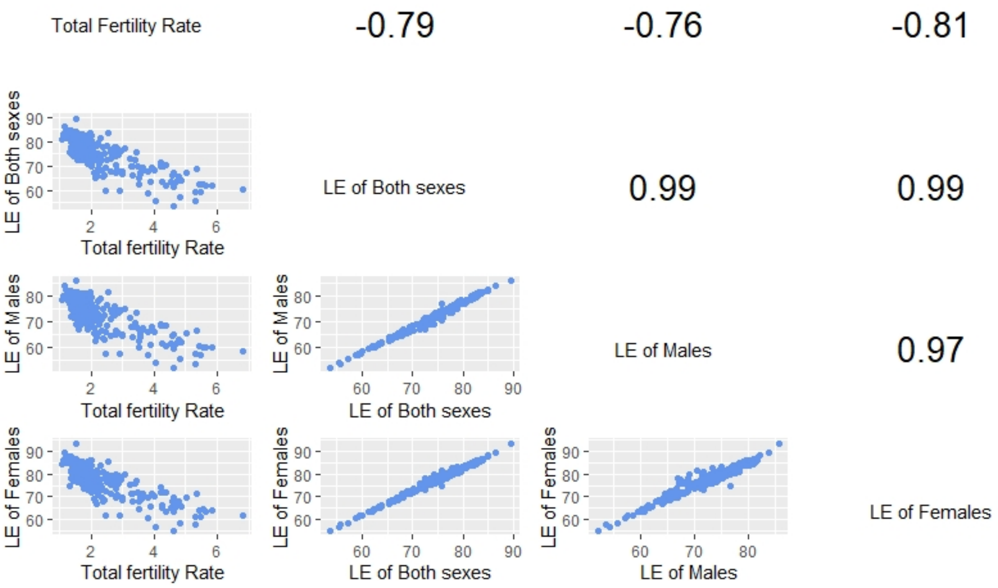


Figure 3: Scatter plot matrix of Pearson Correlation

of the box plot and comparison of the variability of the variables within individual subregions and between various subregions leads to homogeneous and heterogeneous between and within subregions.

Figure 4 shows that in the African region, total fertility rate variability is lower in Southern Africa than in Middle Africa. The variation in box size or IQR between the two subregions is not significant and the whisker length is also varied. Southern Africa's IQR is around 0.7, whereas Middle Africa's is nearly 1.5, indicating a considerable difference in value, implying that Southern Africa and Middle Africa are heterogeneous. Similarly, there are some differences in IQR of Western Asia and Eastern Asia, demonstrating the substantial variability in total fertility rates between the two subregions.

In comparison to subregions in other areas, the variability within African subregions is substantial, indicating that they are the most varied within subregions. The variability in the European region is less as they have thin whiskers with no extreme values like in Eastern Europe and Western Europe. As a result, the European region seems to be homogeneous within subregions. Similarly, subregions within the Oceania, Asia, and America regions are homogeneous. The variability in total fertility rate within Australia/New Zealand subregions is the least among all and highest in Northern Africa subregions.

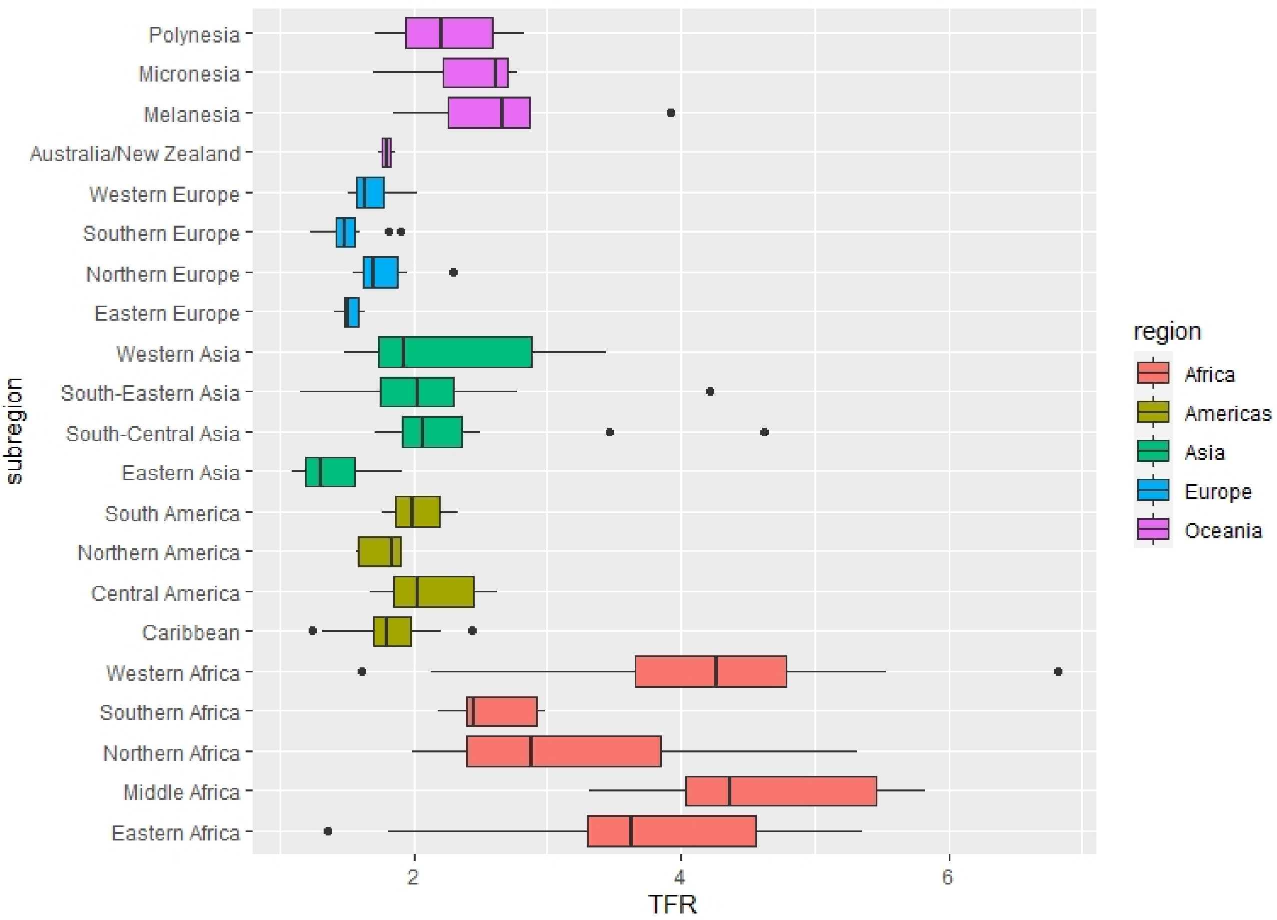


Figure 4: Box plot for Total Fertility Rate

As shown in Figure 5, we generated box plots for both sexes' life expectancy at birth for each subregion. The graph shows that variation in life expectancy is inversely proportional to variation in total fertility rate within subregions. Middle Africa has a lower IQR than Southern Africa, while Western Asia also has a lower dispersion than Eastern Asia. However, the result remains the same: there is a lot of diversity between two subregions, making them heterogeneous. Australia/New Zealand, Western Europe, Northern Europe, Northern America and Central America are homogeneous within their respective subregions. Western Africa, Southern Africa, Northern Africa, Eastern Asia, South-Eastern Asia are heterogeneous within their respective subregions.

The variability in life expectancy at birth for males and females follows a similar trend to that of life expectancy at birth for both sexes, with the least variability in the Oceania subregions of Australia/New Zealand and the most variability in the Asia subregions of Eastern Asia.

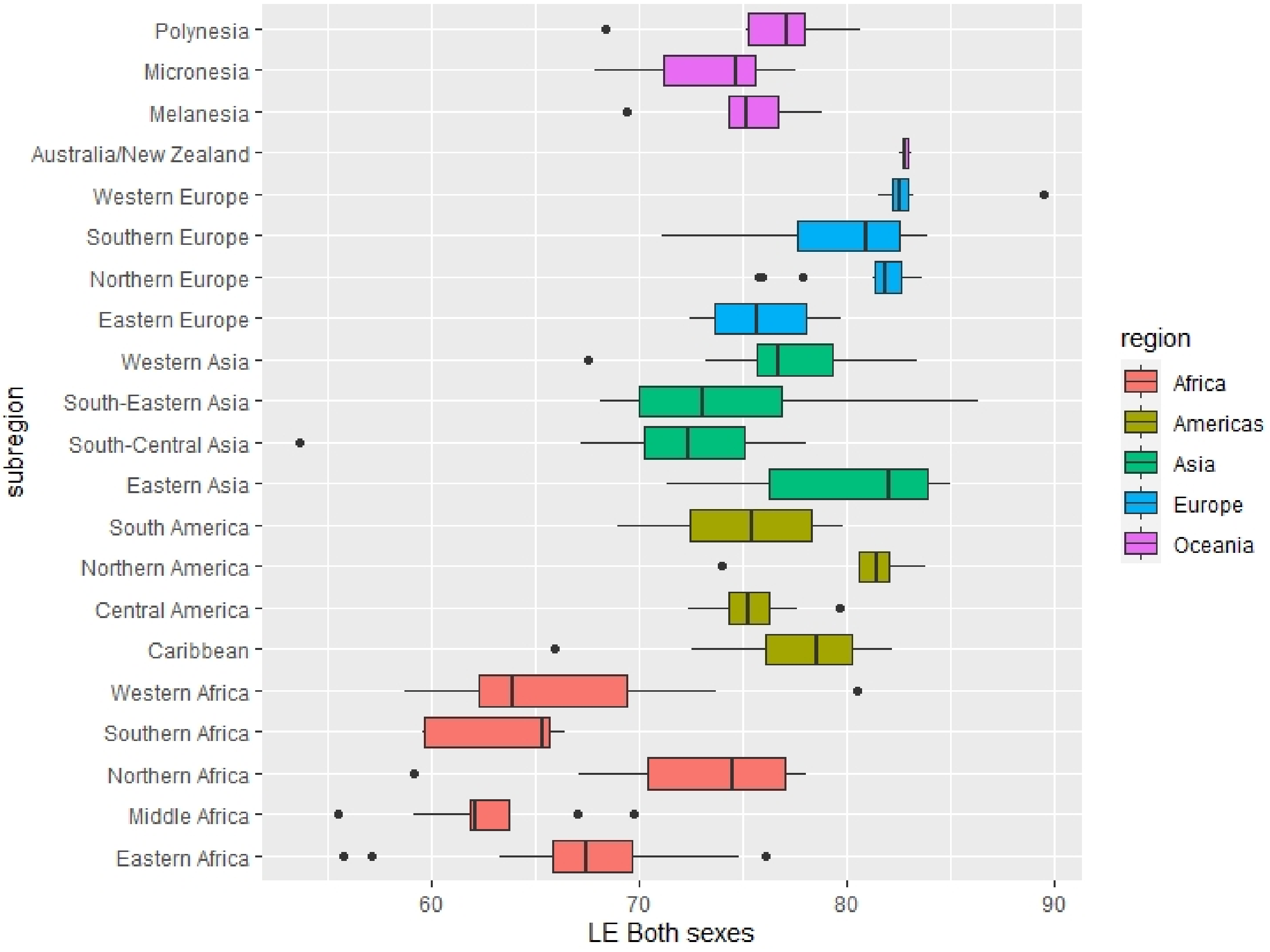


Figure 5: Box plot for Life Expectancy of Both Sexes

Overall, we can claim that all variables exhibit a lot of variation within African subregions and are hence heterogeneous. With a few exceptions, Asia, Africa, America, Europe, and Oceania are all homogeneous within their respective subregions.. The boxplots for the life expectancy of males and females are presented in Figures 9 and 10, respectively, in the Appendix on page 17.

4.4 Comparison of variable values from 2002 to 2022

In the final subsection, we utilize scatter plots to compare how variable values have changed in all regions, as shown by different colors, during a 20-year period (2002-2022).

Even though there is a linear association between these two years, the total fertility rate has declined overall from 3.0 to 2.4 in the last 20 years. Figure 6 shows that over the last

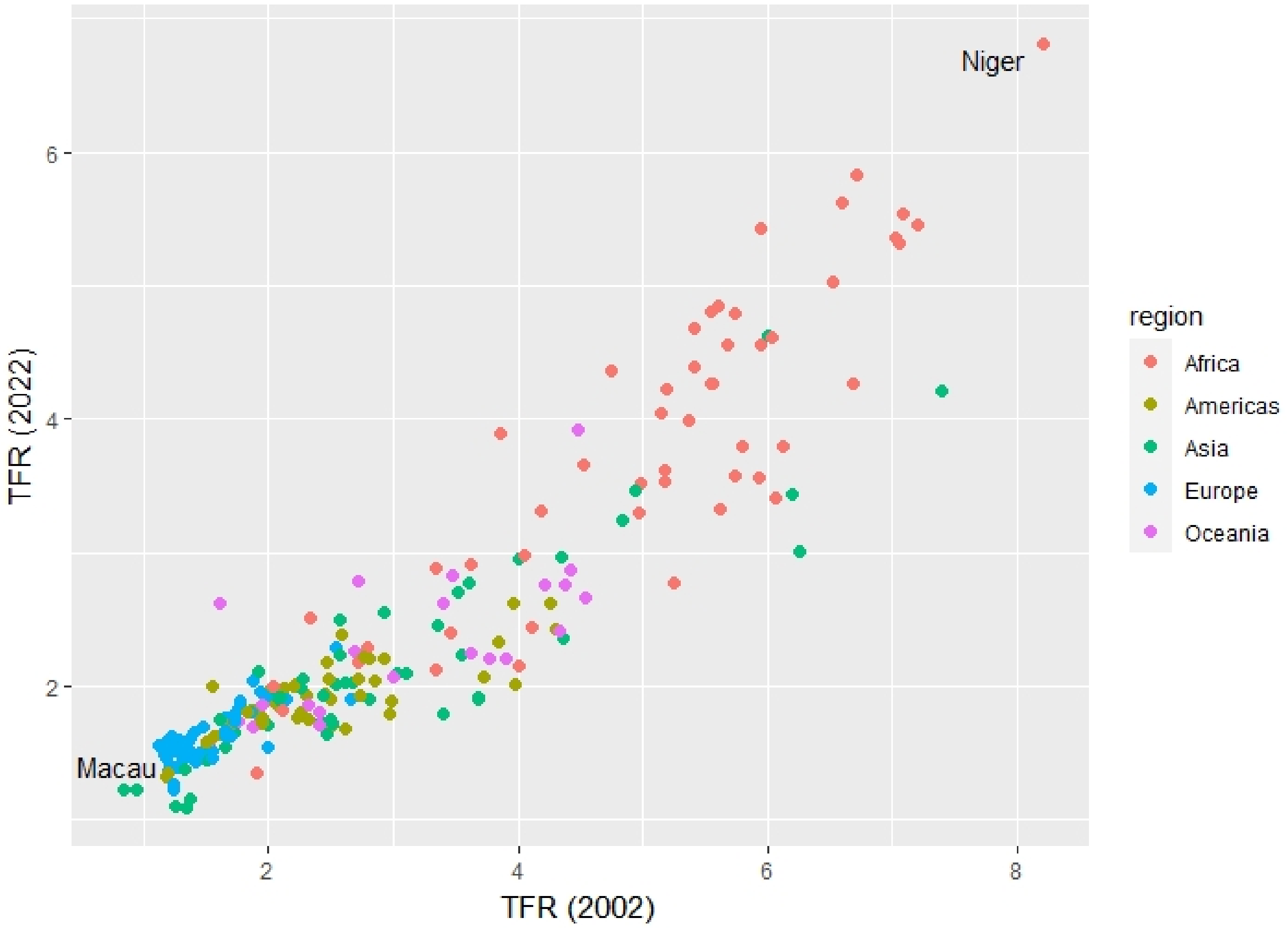


Figure 6: Scatter Plot of TFR for 20 years

20 years, Niger has had the highest fertility rate and Macau has had the lowest fertility rate, which are in the African region and Asian region, respectively. Only the African region has better fertility rates than other regions, even though there is no improvement in it over 20 years. We can see that the European regions have very low fertility rates followed by the American regions. No improvement in fertility rates can also be seen in Asia and Oceania, and some of its subregions show a high reduction in fertility rates.

It can be seen in Figure 7, the trend in life expectancy for both sexes differs from the overall fertility rate. Here, Africa has a lower life expectancy than other regions. Monaco, which belongs to the European zone, has the highest life expectancy, while Afghanistan, which belongs to the Asian region, has the lowest. In Africa, there are some subregions where life expectancy has improved, such as Zimbabwe, Nigeria and Botswana.

Referring to Figures 11 and 12 in the Appendix on page 18, it is evident that all of the life expectancy variables follow the same pattern. In 2002 and 2022, male and female expectancy values in Africa are low, whereas in Europe and America it is high. This pattern is diametrically opposed to the overall fertility rate. As a result, we may conclude that the total fertility rate has declined over the last 20 years while life expectancy has improved.

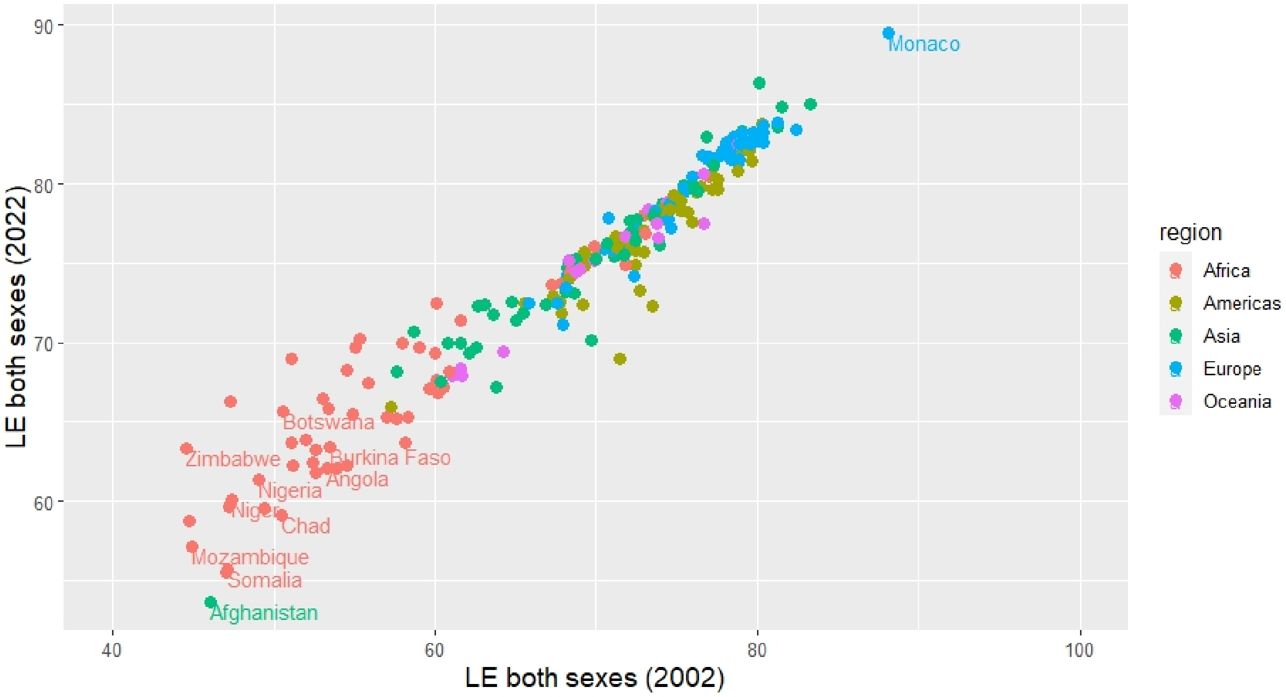


Figure 7: Scatter Plot of Life Expectancy of Both Sexes for 20 years

5 Summary

In this project, we have performed data analysis on a small extract of data from the IDB data set, which includes life expectancy and total fertility rates for 227 countries for the years 2002 and 2022. The analysis of the data is done by descriptive statistical methods such as histograms, box plots, correlation plots and scatter plots. The result is interpreted and visualized.

In the first three tasks, only the records for the year 2022 are used. Initially, we used histograms to describe the full data set in order to comprehend the frequency distribution of the quantitative variables. Based on visualization, we concluded that females have a better life expectancy after birth in 2022 than males. Furthermore, through correlation plots, we found that all the involved variables are highly correlated. However, all life expectancy variables show a negative correlation with the total fertility rate, whereas all life expectancy variables are positively correlated with each other. We further investigated the relationship and discovered that all life expectancy variables show increasing monotonic trend with each other while total fertility rate shows decreasing monotonic trend with all life expectancy variables.

Further, from our analysis, we interpreted that the total fertility rate and life expectancy at birth of both sexes, as well as the life expectancy at birth of males and females for Africa's subregions, are heterogeneous within their respective subregions with respect to the variability. Variability within subregions is very less in Europe followed by Asia, America and Oceania. The values of variabilities between subregions is comapratively homogeneous with few exceptions. At last, we found that life expectancy at birth for females, males and both sexes had improved from 2002 to 2022. In addition, the total fertility rate has not improved much over the last 20 years.

To further improve the analysis and obtain reliable and accurate results, we can extract more records and also the data set can be expanded to include additional parameters such as Sex ratio at birth, Birth rate, Death rate, Age-Specific Fertility Rate, Pandemic, and so on.

Bibliography

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

Ken Black. *Business statistics for contemporary decision making.* John Wiley & Sons, Inc., 2006. ISBN 978-0470-40901-5.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics.* CRC Press, Taylor & Francis Group, Boca Raton, 01 2019. ISBN 9780429448294.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

The U.S. Census Bureau. *International data base*, 2022. URL https://www.census.gov/programs-surveys/international-programs/about/idb.html. (visited on 30th April 2022).

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

Hadley Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*, 2021a. URL https://CRAN.R-project.org/package=forcats. R package version 0.5.1.

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2021b. URL https://cran.r-project.org/web/packages/tidyverse/index.html. R package version 1.3.1.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.4.

Appendix

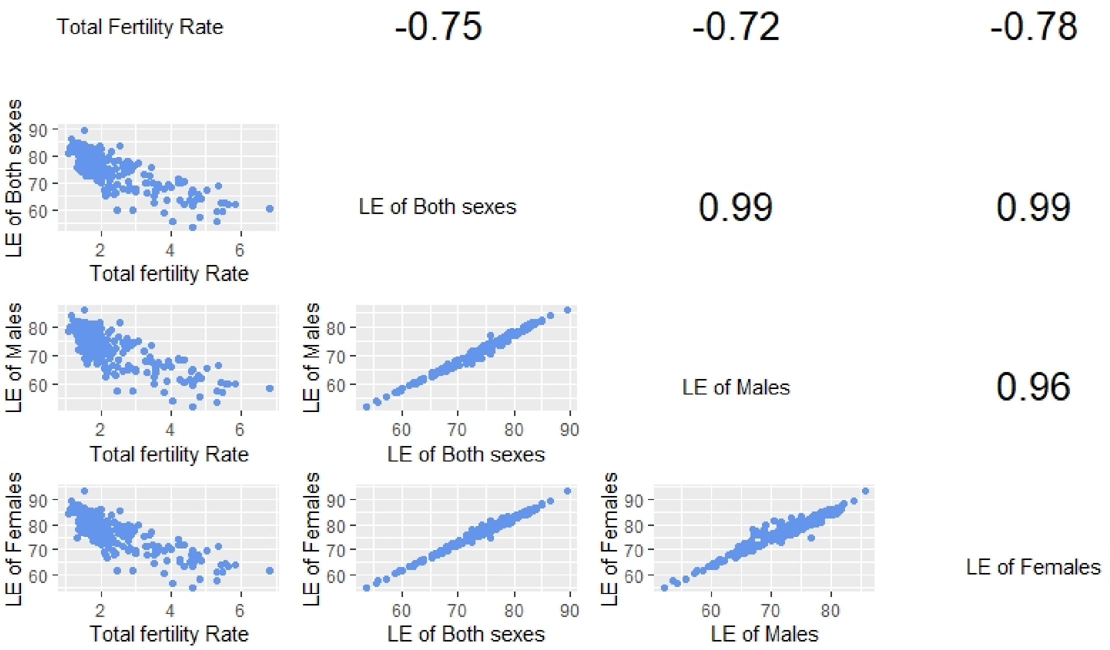


Figure 8: Scatter plot matrix of Spearman Correlation

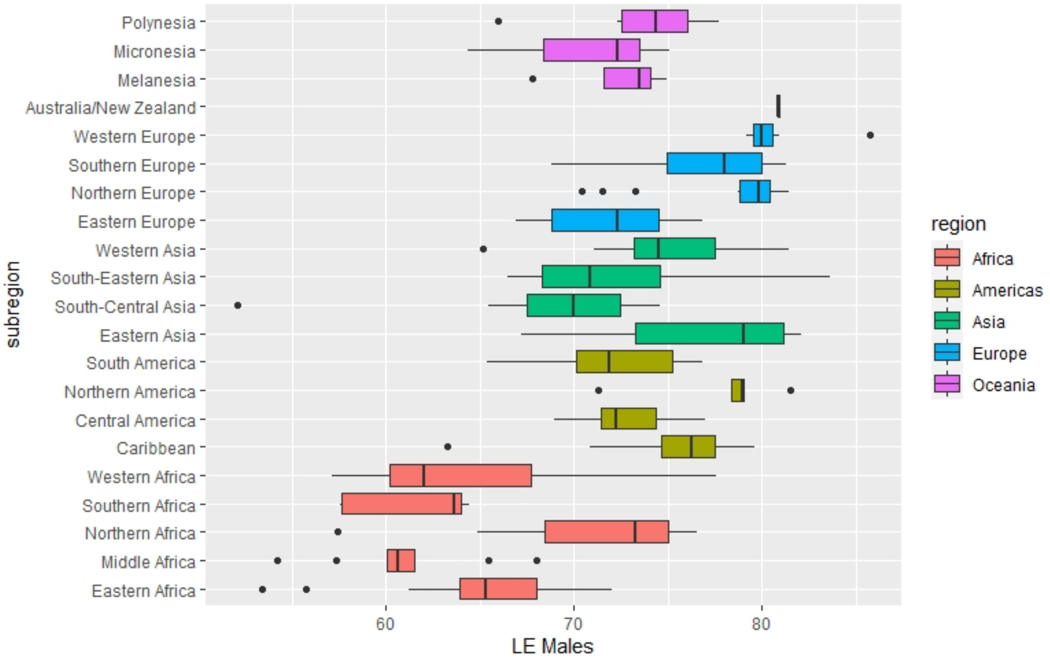


Figure 9: Box plot for Life Expectancy of Males

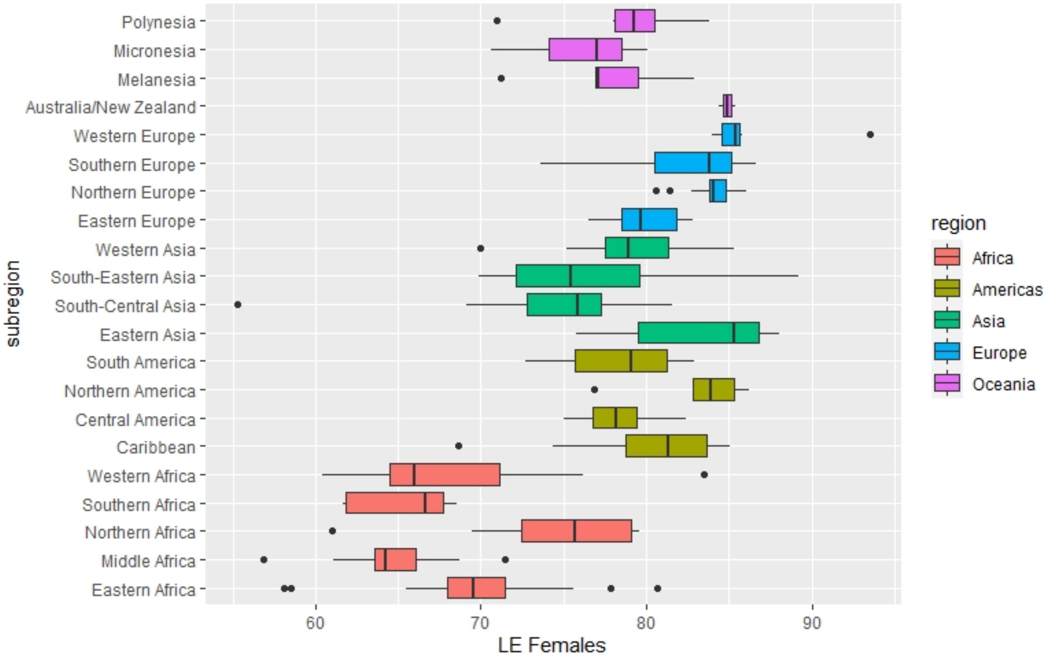


Figure 10: Box plot for Life Expectancy of Females

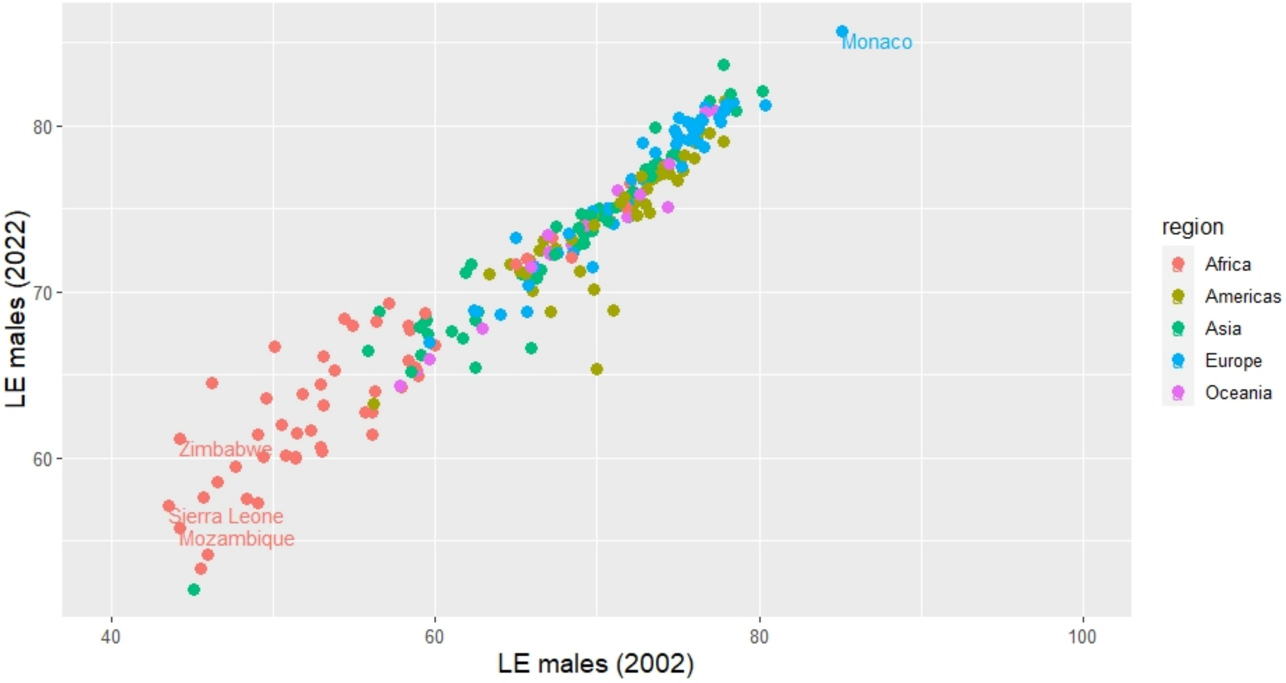


Figure 11: Scatter plot for Life Expectancy of Males for 20 years

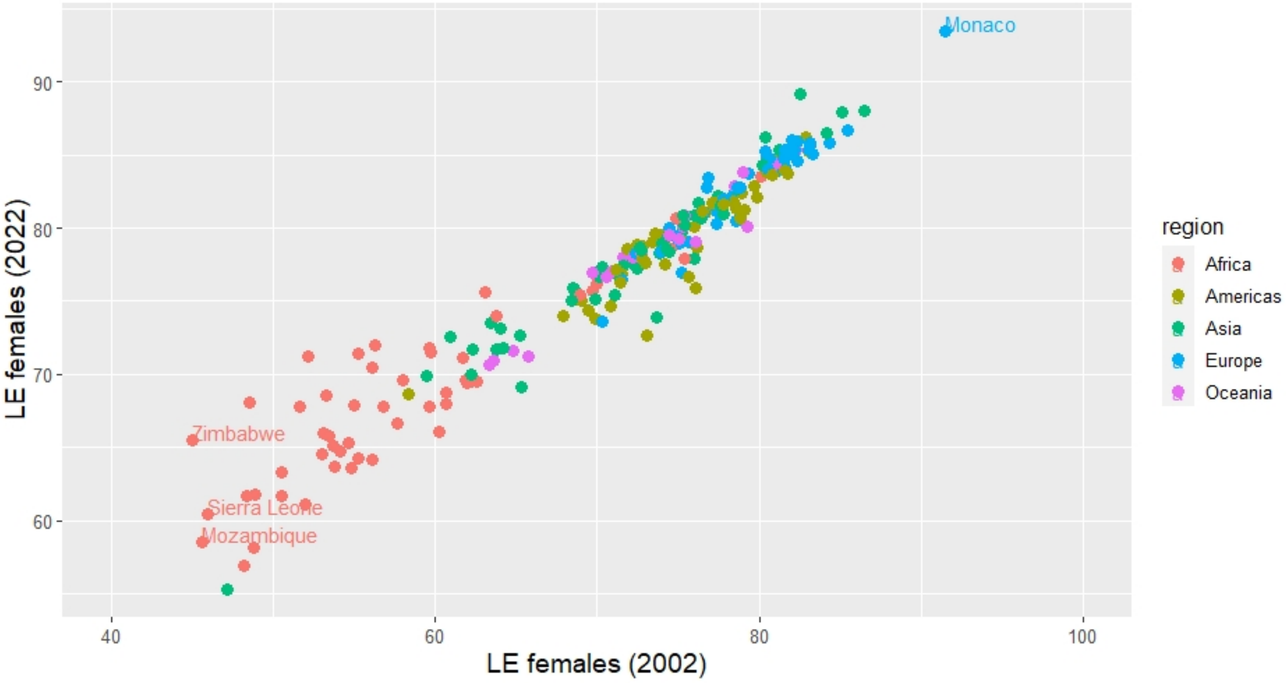


Figure 12: Scatter plot for Life Expectancy of Females for 20 years