TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

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

Author: Ankita Sarkar

Group number: 22

Group members: name1, name2, name3, name4

April 30, 2022

Contents

1 Introduction

The analysis of demographic data helps us to keep track the impact of major events such as disasters, pandemic etc on populations around the world. The purpose of this project is to analyse the variables namely total fertility rate and life expectancy in the year 2022 through frequency distributions. Also, the frequency distributions of the differences of life expectancy between males and females are analysed. The relationships between the variables are checked to reveal any dependency structures along with the examination of the nature of functional relationships. The variability within and between the subregions are assessed. Finally, the change in values of the variables from 2002 to 2022 is compared graphically.

In section 2, the description of data, methods of data collections and data quality are described. Section 3 briefly explains the statistical methods which includes histograms,boxplots and correlation measures as well as the software tools that are being used. Section 4 gives the detailed presentations of the results and their interpretations. The last section concludes the project and also discusses the possibilities of further analyses.

2 Problem statement

2.1 Description of the Dataset

The dataset used for this project is a small sample from the International Data Base (IDB) of the U.S. Census Bureau. The Census Bureau collects data from a variety of sources namely censuses, surveys, vital registries and also administrative records and makes them available through the IDB. National statistical offices of other countries are other primary sources for the data collection. The IDB is updated regularly so that it stays relevant to research, program-planning and policy making plan in US and around the globe.Henderson and Velleman (1981)

The IDB gives the estimates and projections for U.S government includes demographic measures of over 200 countries and areas all over the globe with populations 5000 or more.

2.2 Subsection 2

In this section, various statistical methods and graphs are described which are used to analyse the dataset. Python programming language (?) version 3.9.7 is used.

3 Statistical methods

- description of the statistical methods, models, etc. that are used, including their properties and the assumptions on which they are based (mathematical formulas are also required here)
- details of the tools that are used (software including version number, statistical tables, etc.)

3.1 Subsection 1

This is a formula:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

You can also use equations in-line with the text: The arithmetic mean $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ is a measure of central tendency.

3.2 Subsection 2

This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text.

3.3 Subsection 3

This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text. This is your text.

The statistical software R (**?**), version 4.0.3 was used for analysis.

4 Statistical analysis

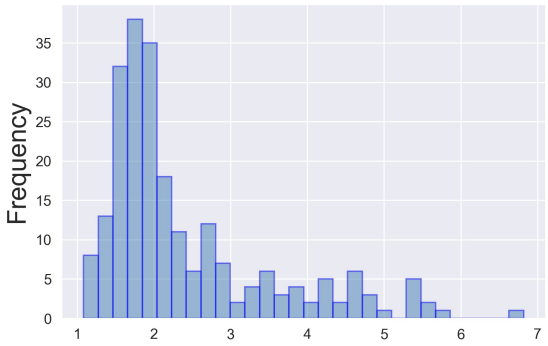
The statistical methods explained in the section 3 are applied on the dataset to explore the data. First, the frequency distributions of the variables Total Fertility Rate, Life Expectancy of both sexes, Life Expectancy of males and Life Expectancy of females in the year 2022 are presented to get a better understanding of the data. Next, the differences of life expectancy data between male and female are considered to directly compare the data. To explore the monotonic nature and linearity of the relationships of the variables, scatter plots are used. The variability of the variables within and between the subregions are observed with boxplots. Finally, how the data for the variables the Total fertility rate and the life expectancy of both sexes between the years 2002 and 2022 are displayed with the help of boxplots again.

4.1 Frequency distributions

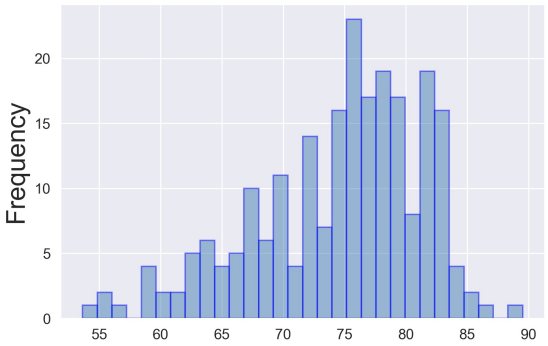
The frequency distributions of the four continuous variables for the year 2022 are considered. There are no missing data in the dataset being used. Figure 1 and Figure 7 show histogram with x-axis as the respective variable and the y-axis as the frequency of the corresponding variable.

Figure 1(a) shows the distribution of the Total fertility rate which is positively skewed. There are maximum 6.82 children per woman and minimum 1.08 child per woman. Niger, a country in Western Africa has the highest total fertility rate and Taiwan in Eastern Asia has the lowest total fertility rate. If we look at the interquartile range, we see that the 50% of the data are spread between 1.68 and 2.77 which indicates that the women in the half of the countries of the data have 2 to 3 children. On average each woman has 2.40 children.

Figure 1(b) is slightly negatively skewed. The median of the data is 75.82 and mean is 74.58. The range of the total life expectancy of both sexes is 35.87. Afghanistan in South-Central Asia has the minimum life expectancy at age 53.65 whereas Monaco in Western Europe has the maximum life expectancy at age 89.52. The interquartile range is 9.6. So 50 % of the countries have life expectancy of both sexes are between 79.65 and 70.05.



(a) Total fertility rate



(b) Life expectancy of both sexes

Figure 1: Histogram for a) Total fertility rate and b) Life expectancy of both sexes in 2022.

Table 1 and Table 3 give the summaries of the data for Total fertility rate and Life expectancy of both sexes respectively.

Table 1: Summary table for total fertility rate

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.08	1.68	1.95	2.40	2.77	6.82

Table 2: Summary table for life expectancy of both sexes

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
53.65	70.05	75.82	74.58	79.65	89.52

Figure 7 in Appendix shows that the histograms for the Life expectancy of males and females separately in the year of 2022. The distributions are similar with the previous distribution of the life expectancy of both sexes. However we see that males live maximum of 85.70 while females live maximum of 93.49 years. On average males live 72.09 years and females live 77.18 years. The minimum years for males are 52.10 and for females are 55.28. We can better explore the differences between the two sexes with the next distribution.

4.1.1 Differences between the sexes

Figure 2 shows two overlapping histograms for males and females consecutively to highlight the differences between the life expectancies in both the genders. Here we clearly see that females outlive males on average by extra 5 years (Appendix : Table **??**).

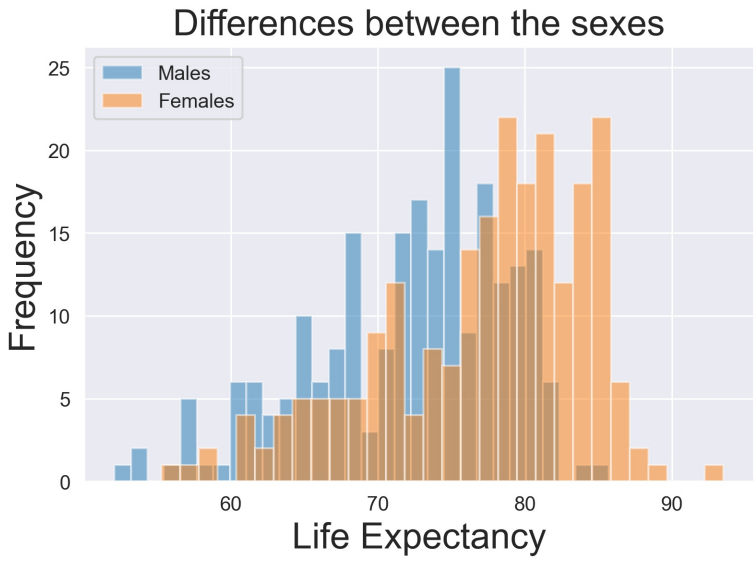
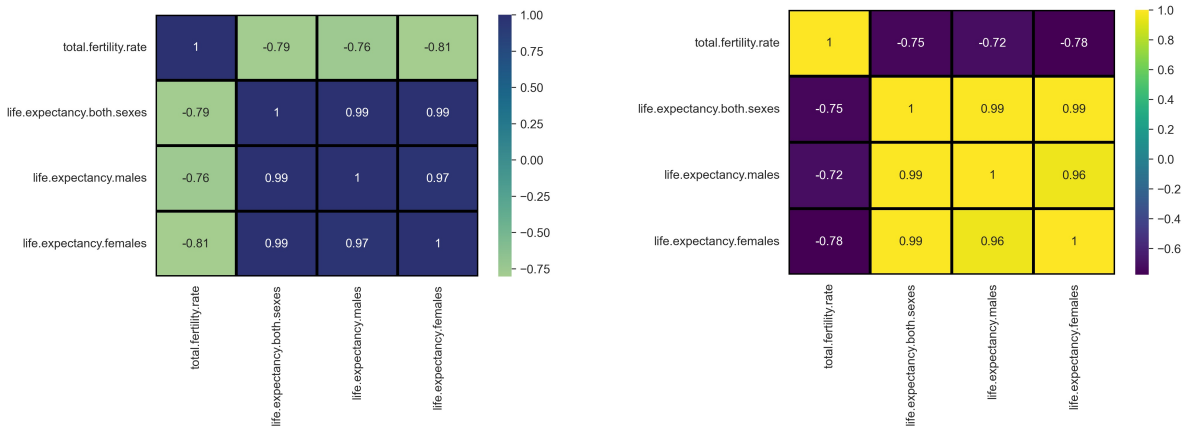


Figure 2: Differences between the sexes

4.2 Analysis of Correlation between the variables

To compare the linear relationship between the variables, Pearson correlation coefficient is used whereas Spearman correlation coefficient is used to check the monotonic nature of the relationships. Figure 3(a) shows heatmap for the Pearson correlation coefficients and Figure 3(b) shows the same for the Spearson correlation coefficients between the variables.

From the heatmaps below, we see that Total fertility rate and Life expectancy of females are strongly negatively correlated. Pearson correlation coeeficient of -0.81 between the two variables indicates a strong negative relationship. However the relationship is not exactly linear rather monotonically decreases in non-linear fashion. Spearman correlation coefficient of -0.78 between the two same variables confirms that. So the Life expectancy of females decreases with high Total fertility rate.



(a) Pearson correlation coeeficient

(b) Spearson correlation coeeficient

Figure 3: Correlation coefficient between variables in 2022.

The correlations between the variables are best viewed through scatterplots which can be found in Appendix (Figure 8). The Life expectancy of both sexes and Life expectancy of females or males are strongly positively correlated. The relationship is linear and monotonically increasing. The Life expectancies are observed as more positively correlated in case of females than for males.

4.3 Comparison of variability within and between subregions

The variability of the Total fertility rate within and between the subregions are analysed in this subsection. Multiple boxplots are used to explain the variabilities among the regions. The regions are alphabetically ordered. Different colours for each particular region are helpful in understanding the spread of data within the subregions.

4.3.1 Total fertility rate

In Figure 4, we see that the subregions in Africa show the highest variability followed by Asia and then Oceania. Subregions in America and Europe show the least variability compared to other regions. The median values and the interquartile range in African subregions vary the most. For majority of European subregions and Australia/New Zealand in Oceania, the interquartile range is more compact indicating the Total fertility rate in 50% of the countries in these regions are closer to each other. It is interesting to see that a country named Togo in Western Africa and Uzbekistan in South-Central Asia

have the highest extreme values namely 6.82 and 4.62 respectively. But these extreme values do not add much to the analysis.

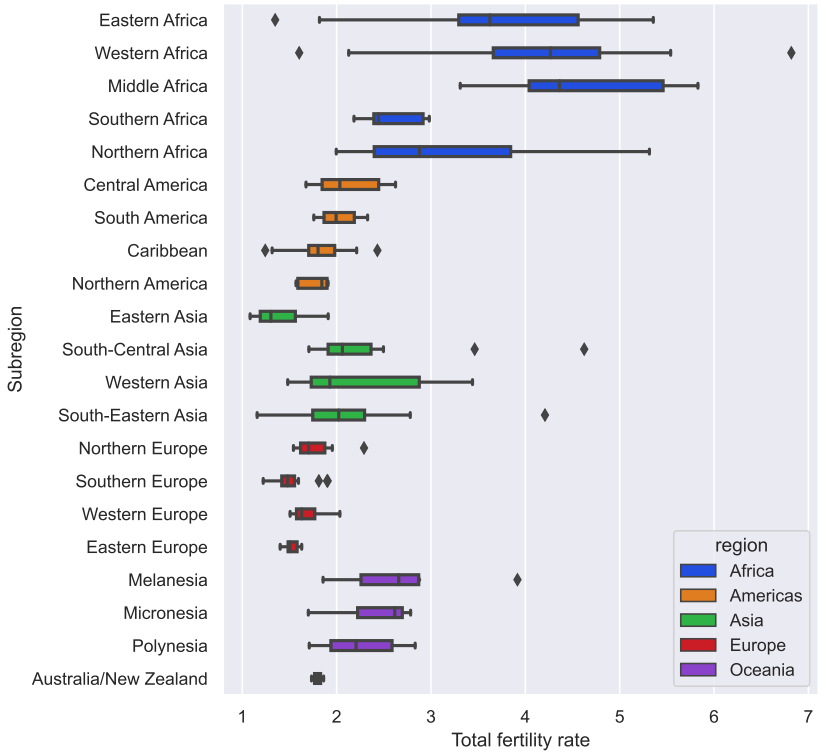


Figure 4: Variability comparison within and between subregions for total fertility rate.

4.3.2 Life expectancy of both sexes

Figure 5 represents the variability of the variable Life expectancy of both sexes in different subregions. The change in the shape of the boxplots from the previous plot discussed just above are specifically noted as a result of the negative correlation between the Total fertility rate and Life expectancy of both sexes. The subregions which have higher Total fertility rate have the lower Life expectancy. The interquartile range varies quite a lot in African and Asian subregions indicating high variance in life expectancy. For America, Europe and Oceania, the variation in the interquartile range are relatively small. Monaco in Western Europe has the highest life expectancy of 89.52 years whereas Afghanistan in South-Central Asia has the lowest life expectancy of 53.65 years. Again, these are the extreme values and these do not influence the analysis much.

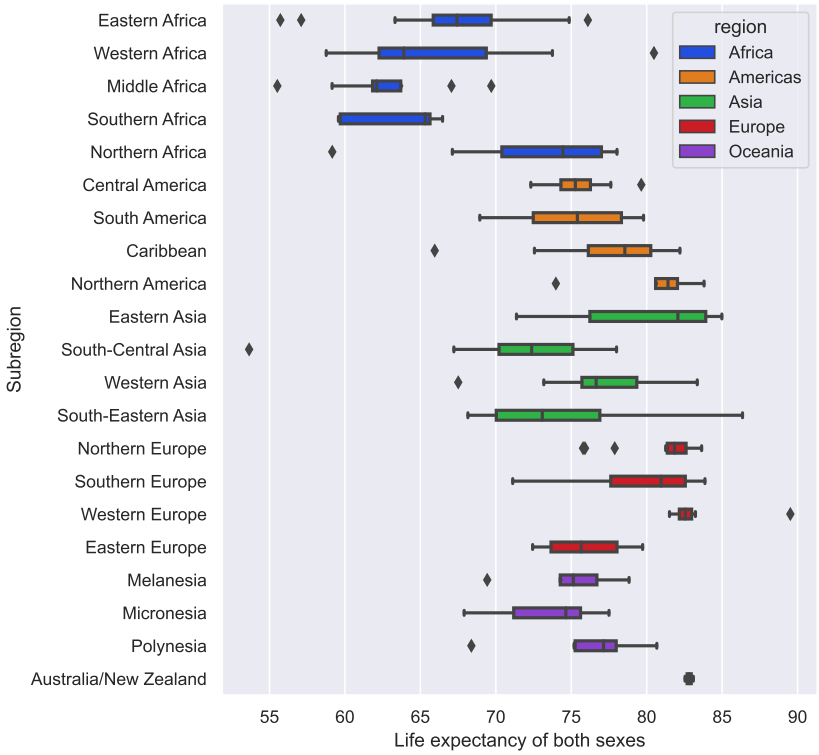


Figure 5: Variability comparison within and between subregions for life expectancy of both sexes.

Figure 9 and Figure 10 in the Appendix display the variabilities for the life expectancy of males and females separately. We observe high variability in the African and Asian subregions. The interquartile ranges in Middle Africa, Caribbean subregion in Americas region, Western Europe and Australia/New Zealand in Oceania display very less variance for both males and females.

4.4 Change of variables between 2002 and 2022

In this subsection, we see how the values for the variables have been changed over the years. To illustrate the changes, boxplots are used in the graphs for the two years. The six regions are plotted on the x-axis and the respective variables are plotted on the y-axis. Data for the two years are displayed in two different colours.

In Table 4 in Appendix the Total fertility rate has decreased from 3.00 in 2002 to 2.41 in 2022. This can be observed in the graph as well. The orange boxes are placed lower

than the blue boxes for all the regions except for Europe. The mean Total fertility rate which is 1.62 has increased in 2022 in European subregions than 1.51 from 2002.

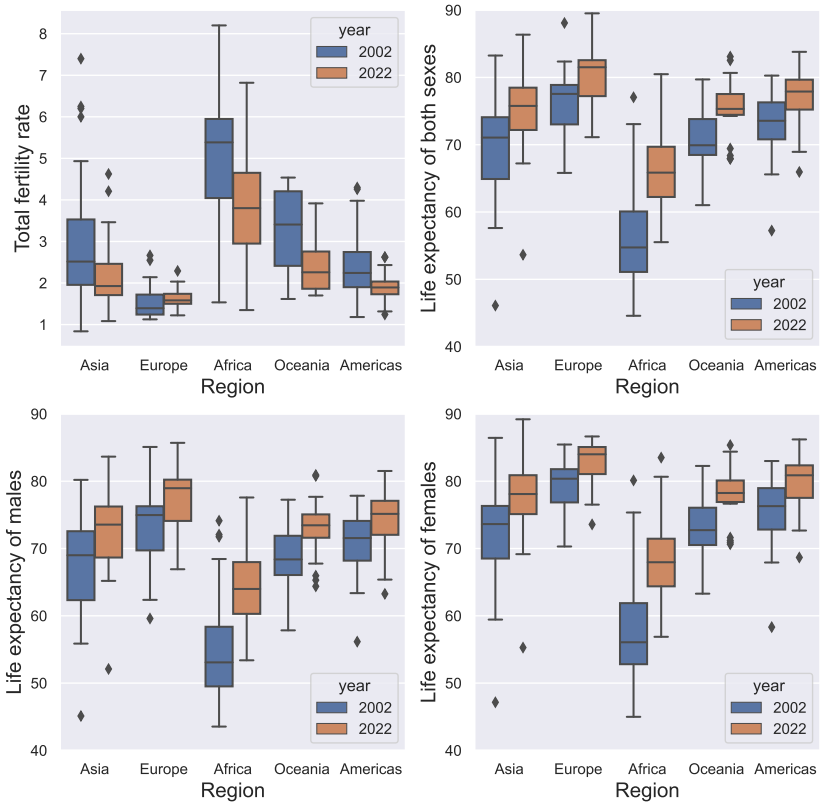


Figure 6: Region wise change in values of variables between the year 2002 & 2022.

On average, life expectancy have increased in 2022 from that in 2002. This is confirmed from Table 4 in Appendix with mean life expectancy values of both sexes which are 68.86 in 2002 and 74.58 in 2022.

5 Summary

For the purpose of this project, a small subset of the International Data Base (IDB) maintained by U.S. Census Bureau which contains demographic data for over 200 countries with 500 populations or more (from 1950 till now) has been used. Four variables namely Total fertility rate, Life expectancy of both sexes, Life expectancy of males and Life expectancy of females are considered. The dataset also includes the information of country, region, subregion and for the year 2002 and 2022. To analyse the data, the

frequency distribution of the continuous variables are shown. It has been observed that the Total fertility rate of 50% of the countries lie between 1.68 to 2.77. One extreme value such as 6.82 per female is found in Niger in Western Africa. The life expectancy of both sexes are found to be in between 79.65 to 70.05 for 50% of the countries. One case of highest life expectancy of both sexes is 89.52 which is found to be in Monaco in Europe and one lowest case of 53.65 in South-Central Asia country of Afghanistan. Although the extreme values do not contribute much to the analysis of the variables in this project. It is found that women on average live 5 years longer than men. The variables Total fertility rate and Life expectancy are negatively correlated. The relationship is monotonic in nature though not linear. There are maximum variance within the African subregions followed by Asian subregions for the studied variables. European subregions, Americas and Australia/New Zealand in Oceania region show less variabilities within themselves. The change in the values of the variables between the year of 2002 and 2022 are observed. It is found that the Total fertility rate has decreased (expect for Europe) and the Life expectancy has increased in 2022 across all the countries.

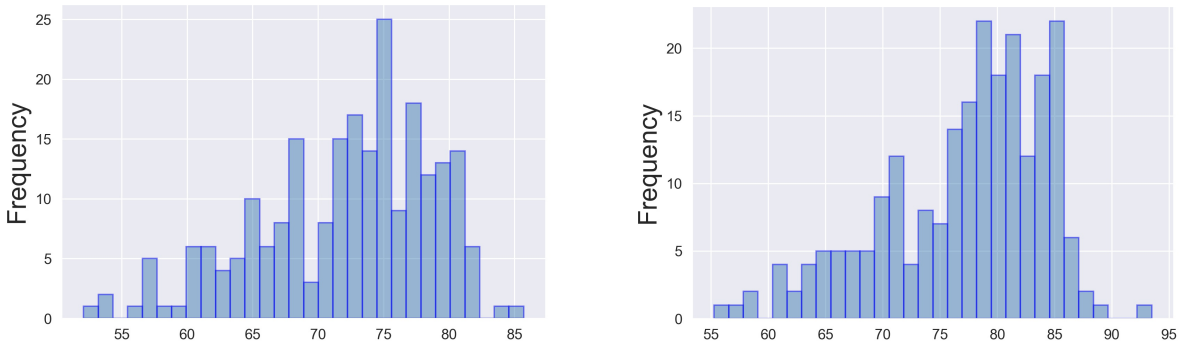
For further investigations, it can be useful to include variables such as child mortality rate, countries' GDP, economic standards, health care conditions etc in the project to better understand the demographic data to make strong conclusions about the increasing or decreasing trend of the variables and their inter relationships.

Bibliography

Harold V. Henderson and Paul F. Velleman. Building multiple regression models interactively. *Biometrics*, 37(2):391–411, 1981. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2530428.

Appendix

A Additional figures



(a) Life expectancy of males

(b) Life expectancy of females

Figure 7: Histogram for a) Life expectancy of males and b) Life expectancy of females in 2022.

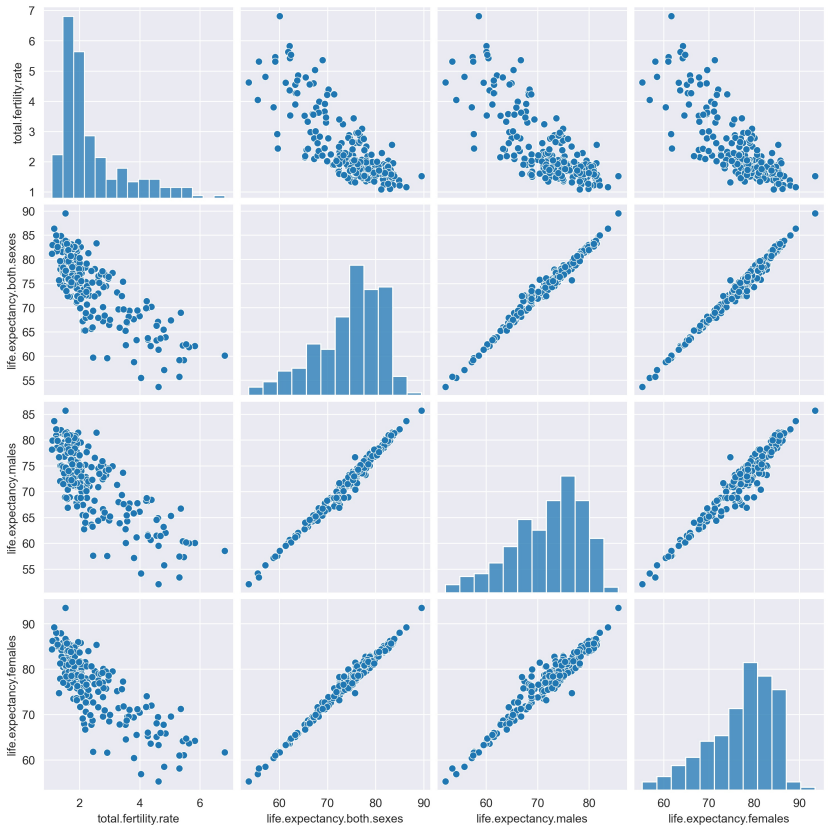


Figure 8: Pairplots between the variables in 2022.

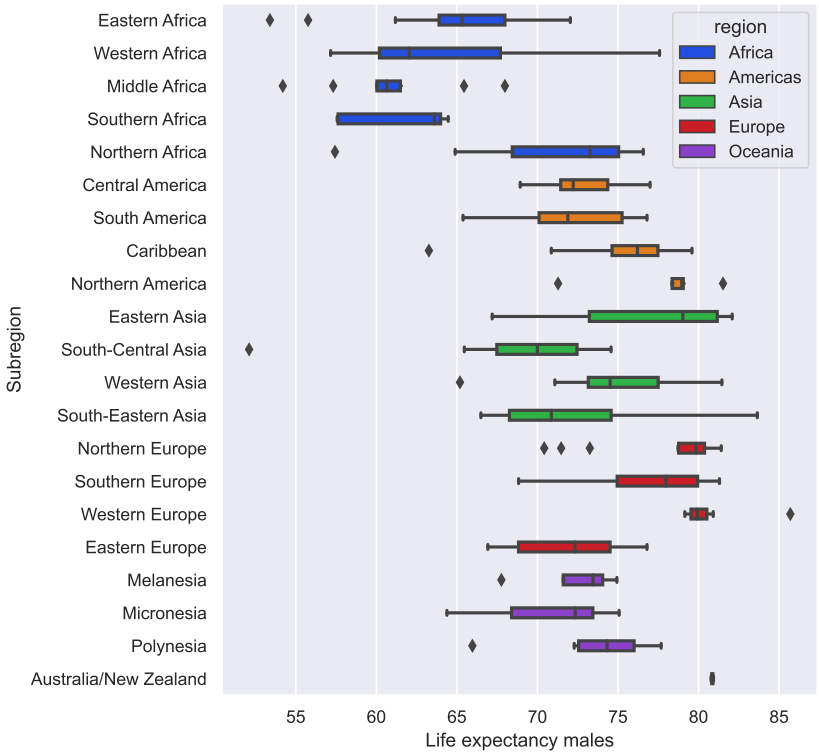


Figure 9: Variability comparison within and between subregions for life expectancy of males.

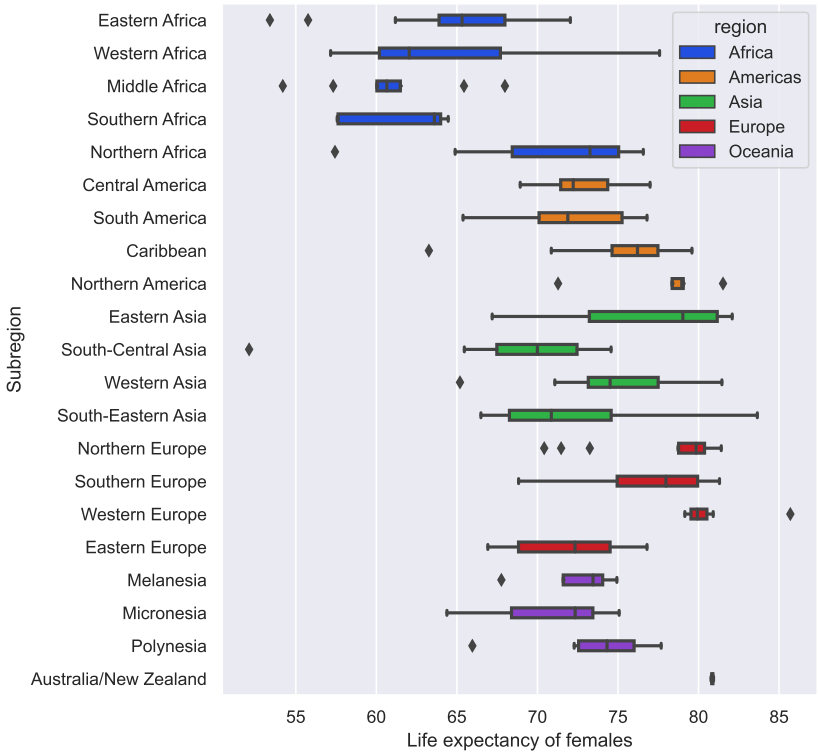


Figure 10: Variability comparison within and between subregions for life expectancy of females.

B Additional tables

Table 3: Summary table for life expectancy of males & females

	Mean	Min.	25%	50%	75%	Max.
Males	72.10	52.10	67.93	73.26	77.19	85.70
Females	77.18	55.28	72.63	78.69	82.55	93.49

Table 4: Pivot table for the mean value of the variables in 2002 & 2022

Year	Total fertility rate	Life expectancy of of both sexes	Life expectancy of males	Life expectancy of females
2002	3.00	68.86	66.55	71.29
2022	2.41	74.58	72.10	77.18