TU Dortmund

Introductory Case Studies

Project I: Descriptive analysis of demographic data

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Birte Hellwig
Dr. Paul Wiemann
M. Sc. Hendrik Dohme

Author: Supritha Palguna

Group number: 20

Group members: Kunal Kochar, Dhanunjaya Elluri, Harshini Eggoni, Naveen Kumar Bhageradhi, Ashish Saini

May 6, 2022

Contents

1 Introduction

Demographics data encompass a wide range of socioeconomic data, such as a population's breakdown by gender, age, income and so on. Not all data is equal in the world of demographics. Before using data to make important decisions, one needs to understand where information originates from and how it is derived. The U.S Census, which is updated every ten years, provides the most thorough population sample from which demographic statistics may be derived. These statistics aid in determining who receives Federal assistance, how your vote is counted, how long you need to drive to the store, how much tax you pay and how much funding your child's school might receive from local, state, and federal sources (French, 2014).

This project aims to perform demographic analysis of demographic data of 228 countries from 2002 and 2022. The data is comprised of life expectancy at birth and total fertility rates for these countries. At first, statistical analysis is performed for the year 2022 with frequency distributions and measures of central tendency. Next, we aim to find monotonic relationship between the variables. Further, the variability of variables within and between subregions is discussed. Finally, data from 2022 is compared to data from 2002 to see how trends have altered over the last two decades.

Section 2 describes the dataset, which involves data collection technique, type and size of the sample data and description of variables. It also gives an overview of the objectives of the project. Section 3 provides a description of several statistical methods employed in this study. This comprises of graphs such as histograms, scatter plots, and box plots, as well as measurements like mean, median, variance, correlation, and so on. In section 4, results are presented with tables and graphs and also interpreted in relation to the problem statement. Last section includes a succinct summary of the important findings and an outlook on further analysis.

2 Problem statement

2.1 Dataset

The United States Census Bureau's International Data Base (IDB) contains demographic data for over 200 nations with a population of 5,000 or more people from 1950 to the present (with projections until 2060). The data is based on official demographic infor-

mation from the individual countries by the U.S. Census Bureau. To offer estimates and projections, the Census Bureau examines and assesses these records. The dataset used in this project is a small extract from IDB (International data base, 2022).

The dataset contains 454 observations which includes fertility rate and life expectancy for 228 nations during the years 2002 and 2022. Geographically, the countries are divided into five regions and 21 subregions. The dataset incorporates eight variables. *country* specifies the name of the country. *region* corresponds to the country's continent which includes Africa, Asia, Americas, Europe and Oceania. *subregion* indicates subregion within the *region*. There are 21 subregions in the dataset. *year* determines which year the observation refers to, 2002 or 2022. *total.fertility.rate* is a continuous numeric variable which indicates the average number of children born per woman assuming that all women survive to the end of their childbearing years. *life.expectancy.both.sexes* is a continuous numeric variable measured in years which implies the expected number of years that a group of people both male and female born in the same year will live, given that the mortality rate at any age remains constant in the future. *life.expectancy.males* is a continuous numeric variable measured in years which implies the expected number of years that a group of males born in the same year will live, given that the mortality rate at any age remains constant in the future. *life.expectancy.females* is a continuous numeric variable measured in years which implies the expected number of years that a group of females born in the same year will live, given that the mortality rate at any age remains constant in the future.

For the year 2002, fertility rate and life expectancy at birth values are missing for six countries. These countries are Libya, Puerto Rico, South Sudan, Sudan, Syria and United States

2.2 Project Objective

The goal of this project is to conduct a descriptive analysis of the demographic data provided. Initially, several statistical measures and mathematical graphs are used to perform a univariate analysis of continuous variables from the year 2022. The central tendency measures and histograms are used to summarize the dataset for the year 2022. By plotting a box plot against each other, fertility rate and life expectancy at birth for males and females in different countries is compared. The bi-variate analysis of continuous data is performed using the statistical measure correlation. The measure of

dispersion variance is then used to investigate and compare the variability of data within and between subregions. Finally, the fertility rate and life expectancy are compared for 2002 and 2022.

3 Statistical methods

Various statistical measures and graphs used for analysis of the dataset are explained in this section. The Python programming language (van Rossum, 2022), version 3.9.7, is used for analysis and visualization. The packages used here are pandas (Community, 2022), matplotlib (Hunter, 2021) and seaborn (sea, 2022).

3.1 Statistical Measures

Arithmetic mean

The summation of all the observations divided by the total number of observations is the arithmetic mean of a variable.The arithmetic mean $\bar{x}$ is given by:

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

where n is the number of observations (Heumann et al., 2016). When there are no extreme values in the data set, the mean is beneficial for predicting future results. The impact of outliers on the mean, on the other hand, may be significant and should therefore be acknowledged.

Median

The median splits the observations into two equal sections, with at least 50% of the values being more than or equal to the median and 50% of the values being less than or equal to the median. Let $x_1, x_2, ..., x_n$ be n observations in the order $x_1 \leq x_2 \leq ... \leq x_n$. Then the median $\bar{x}_{0.5}$ is given by (Heumann et al., 2016):

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\Big(x_{(n/2)} + x_{(n/2+1)}\Big) & \text{if } n \text{ is even} \end{cases}$$

For distributions that are skewed or have outliers, the median is frequently utilized as a measure of central tendency.

Standard deviation

The standard deviation expresses how far the observations differ from one another or how scattered they are around the arithmetic mean. When the standard deviation is low, it means that the observations are tightly clustered around the mean. A large standard deviation shows that the observations are less concentrated around the mean. Standard deviation is given by:

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

where $\bar{x}$ is the mean and $x_1, x_2, ..., x_n$ are observations of the variable (Heumann et al., 2016).

Variance

The average value of the squared deviation from the mean $\bar{x}$ for observations $x_1, x_2, ..., x_n$ is called variance (s^2). It is given by (Heumann et al., 2016):

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

Correlation

The degree of relatedness between variables is measured by correlation. The correlation coefficient is a tool for determining the correlation between two variables. The coefficient r for variables X and Y having observations $x_1, x_2, ...x_n$ and $y_1, y_2, ...y_n$ respectively and with means $\bar{x}$ and $\bar{y}$ respectively is given by:

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

The covariance S_{xy} with variances S_{xx} and S_{yy} of the variables X and Y respectively is given by:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

The strength of the association between the variables is represented by r that varies from 1 to +1. A perfect positive association between two sets of values is seen in a r value of +1. A perfect negative correlation has a r value of -1, indicating an inverse relationship between two variables. As one variable increases, the other decreases. An r value of 0 indicates that the two variables have no linear relationship (Ken, 2014) (Heumann et al., 2016).

Range, quartile and inter-quartile range

The difference between the greatest and smallest data values is the range. It is given by, $R = x_{max} - x_{min}$.

Quartiles are central tendency measurements that divide a set of data into four segments. $Q1$, $Q2$, and $Q3$ refer to the three quartiles. The first quartile, Q_1, is equal to the 25th percentile and separates the first one-fourth of the data from the upper three-fourths. The second quartile, Q_2, distinguishes the second and third quarters of the data. Q_2 is the data's median and is positioned at the 50th percentile. The third quartile, Q_3, equals the value of the 75th percentile and splits the first three-quarters of the data from the previous quarter (Ken, 2014).

The range of data between the first and third quartiles is known as the interquartile range. It is the range of the middle 50% of the data, which is calculated by $Q_3 - Q_1$. When the users are more concerned about the values in the center and less interested in extremes, the interquartile range is very beneficial (Ken, 2014).

3.2 Statistical Plots and Graphs

Bar plot

The bar plot is an extensively used qualitative data charting tool. It has two or more categories on one axis and a series of bars on the other axis, one for each category. The scale of the measure (sum, frequency, cash, proportion, etc.) for each category is usually

represented by the length of the bar. Because the categories are not numerical, the bar plot is qualitative, and it might be horizontal or vertical (Ken, 2014).

Histogram

The number of occurrences of a particular type of observation or during a given class period is defined as frequency. In terms of frequencies and class intervals, frequency distribution is one of the ways to display, group, or arrange data. The frequency distribution can be represented graphically or in tabular format.

The frequency of data in specific class intervals is represented by a histogram, which is a collection of continuous rectangles. The heights of the rectangles show the frequency of values in a specific class interval given that the class intervals used along the horizontal axis are identical. If the class intervals are unequal, the areas of the rectangles can be used to compare class frequencies relative to each other (Ken, 2014).

Scatter plot

A scatter plot is a two-dimensional graph which is used to plot observations from two numerical variables of pairs of points. It acts as a tool to examine possible association between two variables.

The identity line is the diagonal of a scatter plot with x-axis values identical to y-axis values. When a point is above the identity line, the value of the y-axis variable is larger than the value of the x-axis variable. The opposite is true for points located below the diagonal (Ken, 2014).

Box plot

A box is an approach to represent a data distribution. The upper and lower quartiles, as well as the median and two most extreme values, are used to graphically display a distribution in a box plot. The plot is made by enclosing the median using a box. This box is expanded outwards along a continuum from the median to the lower and upper quartiles, enclosing not only the median but also the middle half of the data. Whisker lines are extended out from the box toward the outermost data values from the lower and higher quartiles. The five-number summary is determined by five unique numbers,

- Average (Q_2)
- Lower quartile (Q_1)
- Upper quartile (Q_3)
- Minimum value in the distribution
- Maximum value in the distribution (Ken, 2014)

4 Statistical analysis

The statistical analysis of the continuous variables fertility rate, life expectancy at birth for both sexes, and life expectancy at birth for males and females from the year 2022 is performed and displayed using various graphs in this section. Finally, the data from 2022 is compared to the data from 2002.

4.1 Univariate Analysis

Table 1: Summary of the dataset.

	Total fertility rate	Life expectancy both sexes	Life expectancy males	Life expectancy females
count	227	227	227	227
mean	2.4	74.59	72.1	77.18
std	1.11	6.84	6.67	7.13
min	1.08	53.65	52.1	55.28
25%	1.69	70.05	67.93	72.63
50%	1.95	75.82	73.26	78.69
75%	2.77	79.66	77.19	82.55
max	6.81	89.52	85.7	93.49

Table 1 summarizes the description of dataset. On an average, a women bears around 2 children in the year 2022 across 228 countries. The minimum fertility rate is 1.08 of the country Taiwan in Asia. An African country Nigeria has the highest fertility rate of 6.8. People from over 200 countries are expected to live around 74.59 years on average (both male and female). Life expectancy for both sexes is highest in Monaco of Europe i.e., 89.52 years. The minimum life expectancy of 53.65 years is seen in the country Afghanistan. Life expectancy for males is 72.1 years on an average. Highest

life expectancy of 85.7 years for males is seen in the country Andorra which belongs to Europe and the lowest life expectancy is 52 years of the African country, Uganda. The average life expectancy of a female is 77.18 years. Females from the European country Monaco have the highest life expectancy of 93.49 years, whereas females from Afghanistan have the minimum life expectancy of 55.28 years.

Figure 1 shows the histogram of density distribution for the variable total fertility rate. Here x-axis represents the class interval, while y-axis represents the distribution of the variable. From the graph, it can be seen that majority of the countries have fertility rate around 1.5. The plot is here right-skewed which means that most of the population has fertility less than an average. From the figure 6 in Appendix, page number 15, its seen that majority of the countries have a life expectancy of around 75 years. Figures 7 and 8 in Appendix, page number 15 and 16, shows that most number of countries have life expectancy of 72-80 years for males and 77-85 for females respectively.

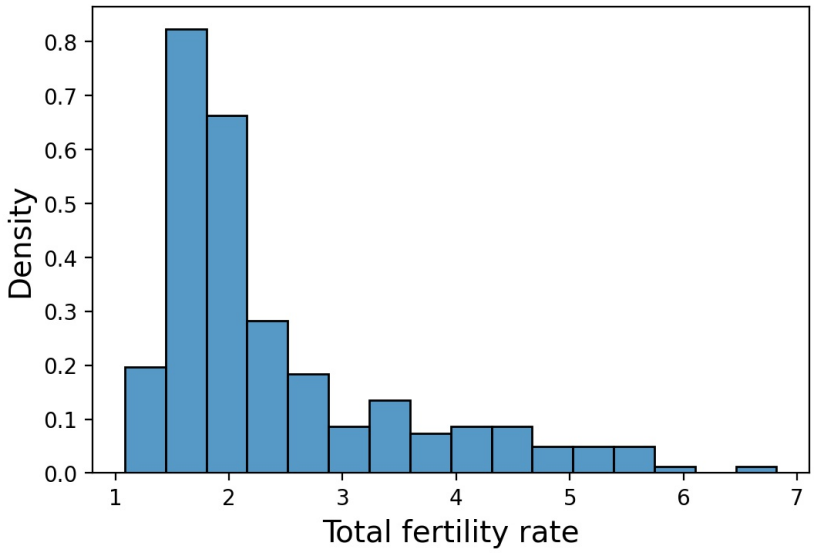

Figure 1: Histogram for total fertility rate.

4.2 Comparison of life expectancies between males and females

Figure 2 shows a scatter plot with male life expectancy represented by blue '+' sign and female life expectancy represented by orange dots. It can be seen that females survive more number of years than males in the majority of the countries. This could be due to the social, genetic, and behavioral disparities between men and women.

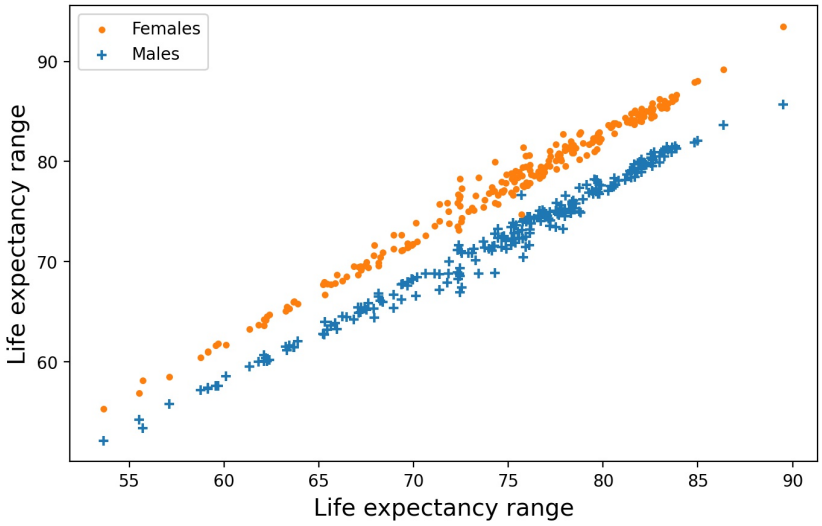

Figure 2: Scatter plot of life expectancy between females and males

4.3 Analysis of relationship between variables

Table 2: Bivariate correlation between the variables.

	Total fertility rate	Life expectancy both sexes	Life expectancy males	Life expectancy females
Total fertility rate	1.0	-0.78	-0.76	-0.8
Life expectancy both sexes	-0.78	1.0	0.99	0.99
Life expectancy males	-0.76	0.99	1.0	0.97
Life expectancy females	-0.8	0.99	0.97	1.0

From table 2 and figure 3, it is seen that there is a negative correlation between total fertility rate and life expectancy variables. As the total fertility rate increases, life expectancy decreases and vice-versa. It can be inferred from the graph that there is a weak linear relationship between the variables which is monotonically decreasing. Life expectancy of females is more negatively correlated to total fertility rate than the other two variables.

Figure 9 in appendix, page number 16, shows that the male and female life expectancy values are substantially positively associated with the variable life expectancy at birth. Furthermore, the male and female life expectancies have a very high positive correlation value, indicating a strong linear association between each other. Hence, there are no significant hints to say that the relationships which are monotone are not linear.

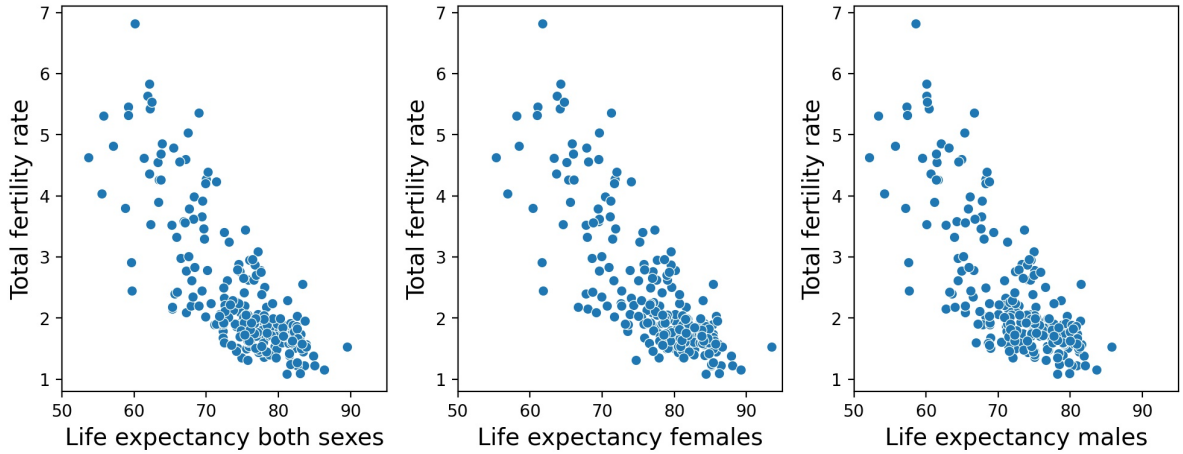

Figure 3: Correlation between fertility rate and life expectancy.

4.4 Analysis of variability of the variables

The variation of the variable fertility rate is plotted for each of the subregions in Figure 4. Eastern Europe and Australia/New Zealand have relatively little variation, indicating that fertility rates are uniform and nearly equal within both Australia/New Zealand and Eastern European countries. African region, on the other hand, has the most variation. Western Africa has the most variation, with the lowest and highest rates of around 2 and 7 children per woman, respectively. This vast disparity could be attributed in large part to the subregion's economic inequality (Hallum and W. Obeng, 2019). It does seem that fertility rates are highly varied in most African subregions, with values that are scattered out and far apart. All subregions within the regions Europe, America, and Oceania have relatively low variance, indicating that fertility rates within these subregions are reasonably uniform.

Figure 10 in the appendix, page number 17, shows a visualization of the variation of the variable life expectancy among subregions for both sexes. Following Australia/New Zealand, Northern Europe has the lowest variance, indicating excellent homogeneity in life expectancy statistics within the subregion. South-Central Asia has the largest variance, indicating that its countries are highly heterogeneous. Sri Lanka has the highest life expectancy in this subregion (78 years), while Afghanistan has the lowest (53.65 years).

The plot of variation of the variables male and female life expectancy is shown in Figures 11 and 12 in the appendix, page number 17 and 18 respectively. It does seem that life expectancy values for both females and males are fairly uniform in the subregion

Australia/New Zealand and highly varied throughout Northern African countries. The majority of African subregions, followed by Asian subregions, are comparatively highly diverse for all variables across their countries. For all factors, most European and American subregions are comparatively homogeneous across their respective countries. For all variables, the variability in the subregion Australia/New Zealand is nearly nil.

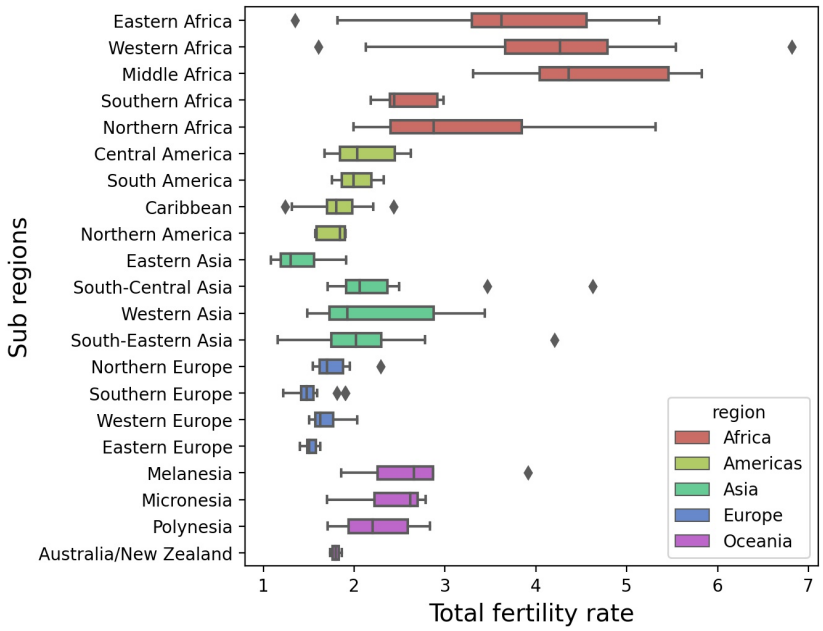

Figure 4: Box plot for total fertility rate.

4.5 Comparison of variables between the years 2002 and 2022

In this section, it is analyzed how the value of total fertility rate and life expectancy variables have changed over two decades. On an average, the fertility rate has decreased from around five children per women in the year 2002 to around 2.4 children per women in the year 2022. According to numerous research, women empowerment, lower child death rates, and other variables have all contributed to lower fertility rates (Roser, 2014). From the figure 5 it can be inferred that the total fertility rate for almost all the regions has decreased in the past 20 years. In comparison to 2002, the fertility rate has very slightly increased in 2022 in the Europe region.

The orange bars in the graph indicate that the life expectancy has increased in the year 2022 when compared to 2002 across all the regions. It can also be seen that the

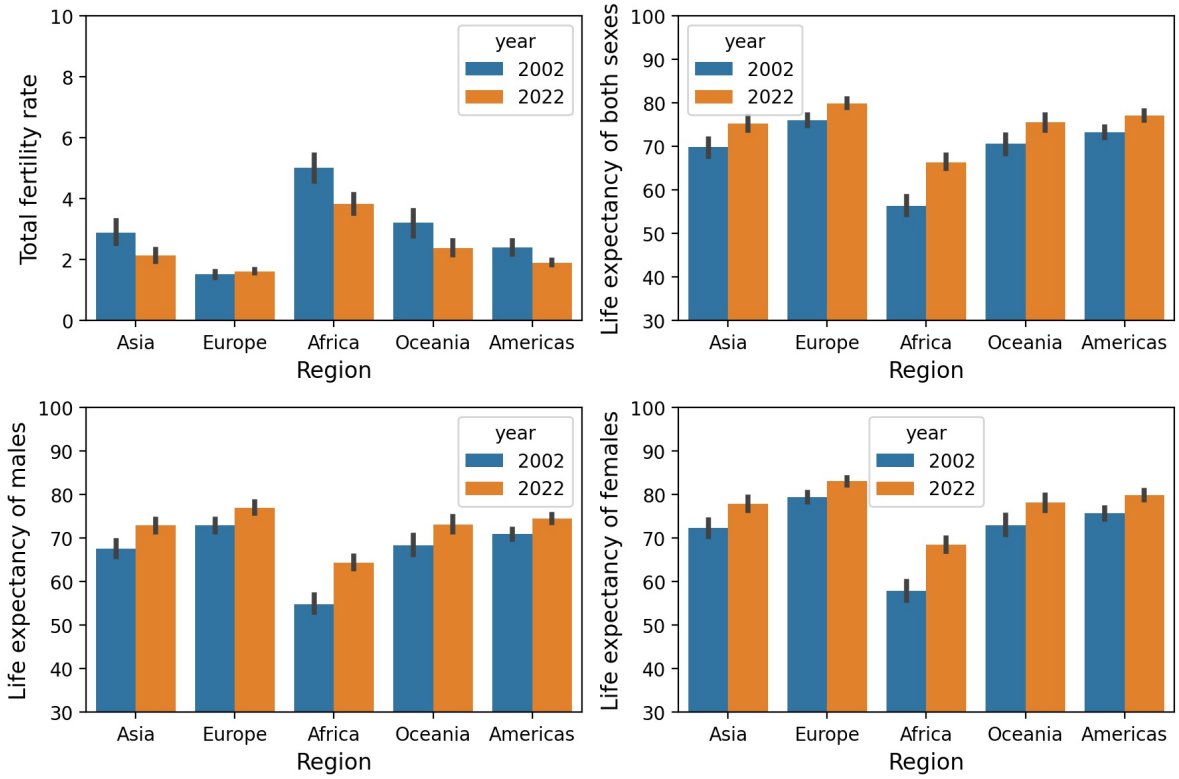

Figure 5: Bar plot of the variables from year 2002 and 2022.

life expectancy has increased more in Africa when compared to other regions over two decades.

5 Summary

The demographic data from the U.S. Census Bureau was used to conduct the project's analysis (International data base, 2022). Life expectancy and fertility rates in 228 countries from all continents were studied. The fertility rates for most countries are 2 children per woman, with Nigeria being the outlier with a fertility rate of around 7 children per woman assuming that the childbirth and childbearing years are constant. The majority of countries have a life expectancy of 75 years. Afghanistan, an Asian country, has the lowest male and female life expectancy, while Monaco, a European country, has the highest. Females have a longer life expectancy than males in general. When the association between the variables was examined, it was discovered that the fertility rate has a negative linear relationship with all life expectancy variables. It can be seen that most

African subregions have a lot of variability, which suggests there are a major differences in fertility rates and life expectancies across its countries. Several European subregions, on the other hand, are comparably identical, indicating less variation in fertility rates and life expectancy across its countries. The significant variability in African countries is attributable to social inequality and poor health infrastructure. In comparison to 2002 data, the average fertility rate has decreased and the average life expectancy has increased in all the countries. Even if most countries' socioeconomic situations have improved over the past 20 years, there is still room for improvement in terms of fertility rates and life expectancy in a few Asian and African countries.

For further analysis, other variables like environment conditions, natural resources, literacy rate of the countries can be considered to analyze fertility rate and life expectancies in a better way.

Bibliography

Seaborn library, 2022. URL `http://seaborn.pydata.org/`. (Visited on 27th April 2022).

Community. Pandas, 2022. URL `https://pandas.pydata.org/`. (Visited on 25th April 2022).

Dr. Charlie French. Community planning new hampshire. 2014.

Christian Hallum and Kwesi W. Obeng. The west africa inequality crisis. page 1–48, 2019. doi: 10.21201/2019.4511.

Christian Heumann, Michael Schomaker, and Shalabh. *Introduction to Statistics and Data Analysis.* 2016. doi: 10.1007/978-3-319-46162-5.

John D. Hunter. Matplotlib, 2021. URL `https://matplotlib.org/`. (Visited on 25th April 2022).

US Census International data base, 2022. URL `https://www.census.gov/programs-surveys/international-programs/about/idb.html`. (Visited on 2nd May 2022).

Black Ken. *Business statistics : for contemporary decision making.* Hoboken, New Jersey : Wiley, 2014.

Max Roser. Fertility rate, 2014. URL `https://ourworldindata.org/fertility-rate`. (Visited on 3rd May 2022).

Guido van Rossum. *Python: A dynamic, open source programming language*, 2022. URL `https://www.python.org/`.

Appendix

A Additional figures

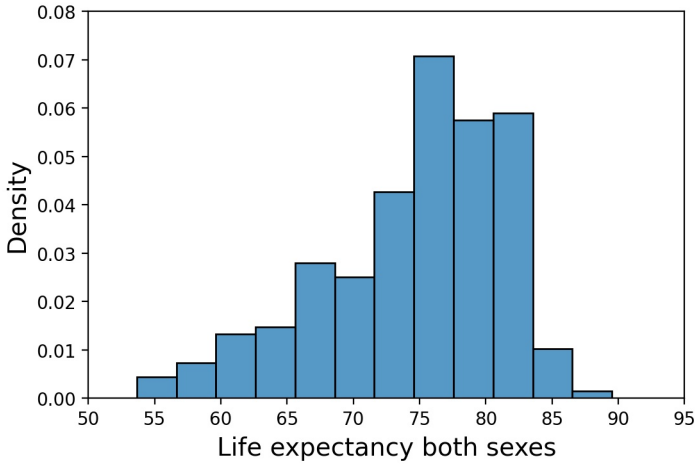

Figure 6: Histogram for life expectancy of both sexes.

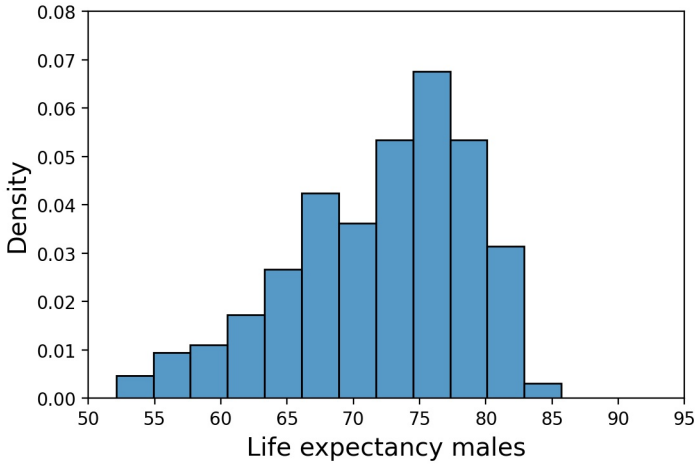

Figure 7: Histogram for life expectancy of males.

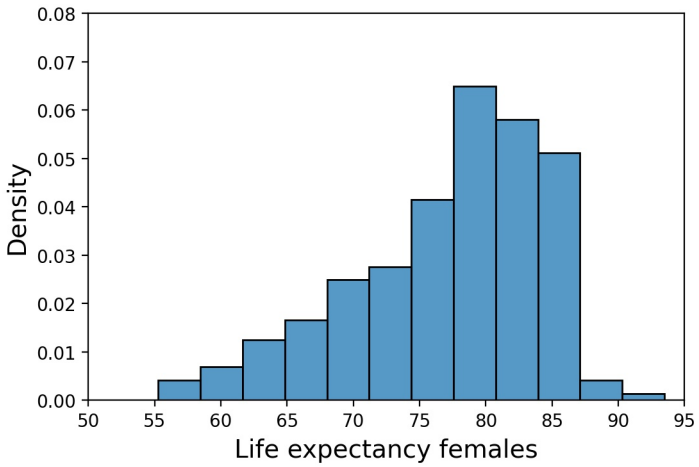

Figure 8: Histogram for life expectancy of females.

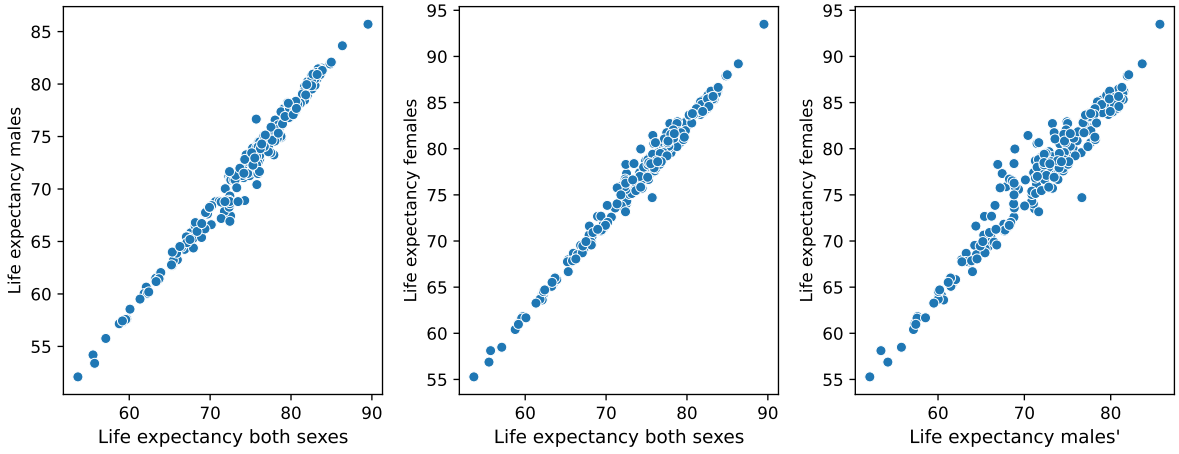

Figure 9: Correlation between life expectancy of box sexes, males and females.

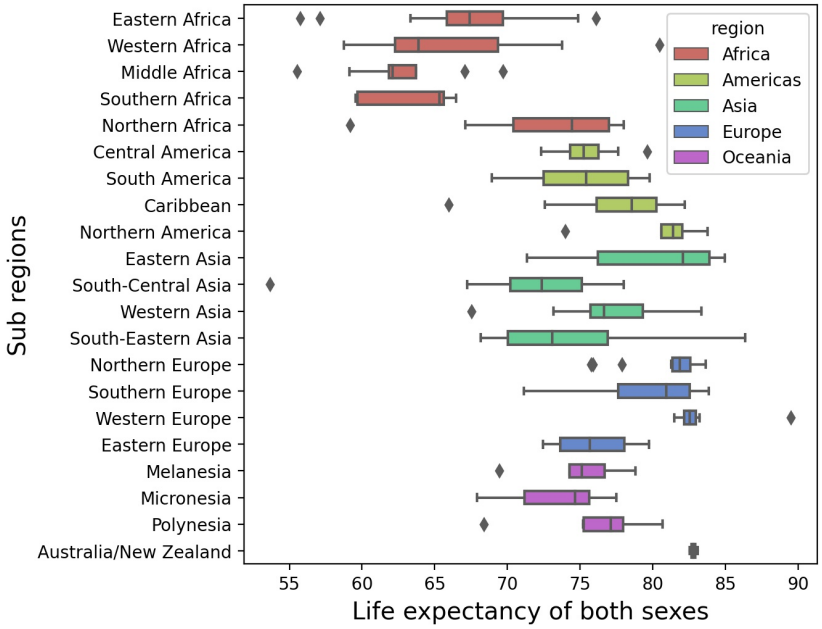

Figure 10: Box plot for life expectancy of both sexes.

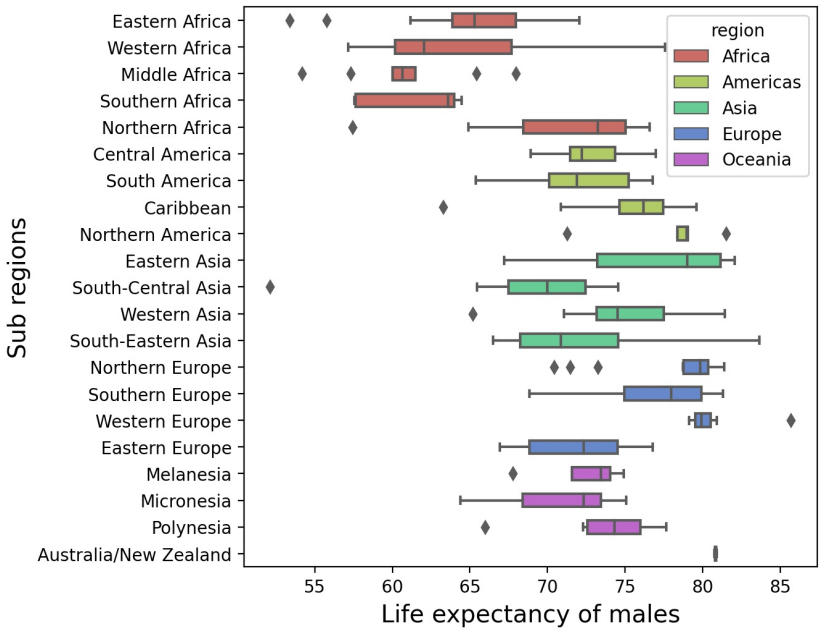

Figure 11: Box plot for life expectancy of males.

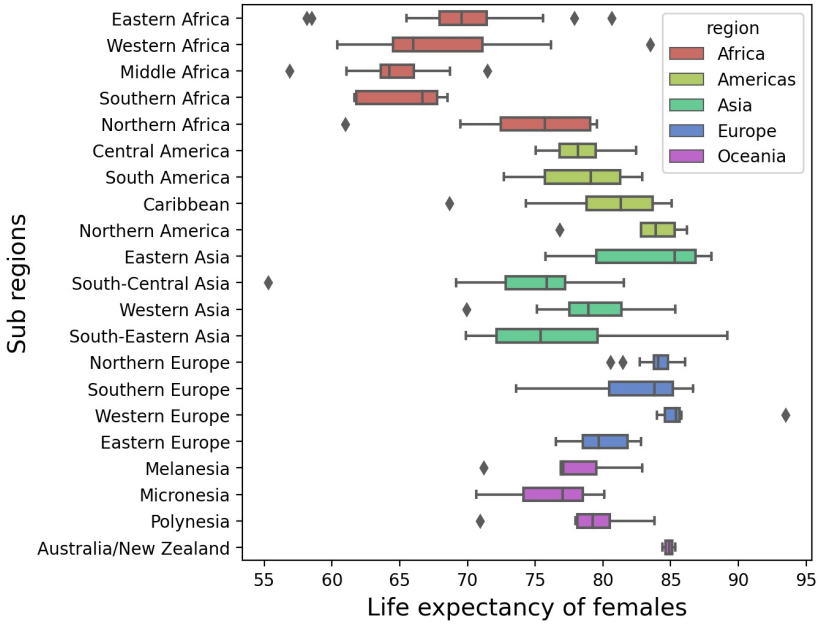

Figure 12: Box plot for life expectancy of females.