Contents

1 Introduction

Understanding social, financial, and open well-being issues that affect populations around the world requires exact and up-to-date demographic data and their measurements. For evidence-based decision-making for a country's growth and development, descriptive analysis of demographic data is important. A raise or reduction in the population has both positive and negative effects on a country's development. It hence makes sense to look for characteristics to check which one explains the effect best.

We will have a detailed overview of life expectancy and fertility rate in this report, which are the two crucial indicators for global health systems, based on different regions, sub-regions, and countries. A general thought is countries with high life expectancy rate prefer to have a limited number of children compared to other countries. In this report, we explore the regional differences and similarities in life expectancy and total fertility rate for 228 countries between the year 2022 and 2002. To begin, we do a uni-variate study of all numeric variables for the year 2022. The bi-variate connection between these numerical variables is also investigated. We also look at how these variables are related within and between the sub-regions. Lastly, a year comparison from 2002 to 2022 is performed to check how these factors have changed over the last 20 years.

The project report contains four additional sections apart from the introductory section. The upcoming section of this report provides a brief summary on the outline of the data set, which includes all of the required variables or features for data analysis. The third section deals with an explanation of various statistical approaches and methods that have been utilized in exploratory data analysis. The statistical procedures mentioned in the third section is then used to conduct analysis in the fourth section. All the important variables are investigated and the results along with visualizations are presented here. Finally, the fifth section presents a recap of all the previous sections and their interpretations and results. It also provides information for further research.

2 Problem statement

2.1 Overview of the data set and data quality

The demographic data used in this report is a small extract from the IDB (International Data Base) of the U.S Census Bureau. It contains information about life expectancy and fertility rates of over 200 countries between the year 2002 and 2022. The data set used in this report contains eight columns that are also known as features and 454 rows that are also called as records. Three of the Eight features are categorical type variables namely Country, Subregion, and Region. The name of the country for which the observations are recorded is indicated by *country.* There are 228 countries and 21 sub-regions, which are grouped together to provide five regions. The variable *Year* is an ordinal variable with a defined order that identifies the year in which the data is recorded (2002 or 2022). Four out of Eight features are numerical type variable and these are, *total fertility rate* which is the natural measure of fertility because it is calculated as the average number of children per woman, *life expectancy for both sexes* which indicates number of years an person is expected to live on average. The variable life expectancy is further stratified by gender: *life expectancy at birth for males and females.*

There are only six missing records for four attributes, which are ignored for the purpose of this report to avoid having an unwanted impact on the measures which can be derived from the data. Overall, the quality of data is appropriate to perform statistical analysis.

2.2 Report objective and task details

In this report, we do a comprehensive exploratory data analysis of our data collection. In the first task, a histogram plot is used to perform frequency distribution of the numerical variables in the data set. Succeeding the uni-variate analysis, bi-variate correlation is used to investigate the relationship between the variables, and the dependent structure of the variables involved. With the use of box plots, the variables are compared within sub-regions and between different subregions in the third task. Finally, the fourth task deals with reviewing the variables from the year 2002 to 2022 to see if they have changed over the last 20 years.

3 Statistical methods

In this section, we will look into several statistical approaches that were utilized to analyze the data. All of the analysis was performed using *Python Programming Language* in a Jupyter notebook.

All the upcoming statistical methods in this report would be explained considering the following n observations as a numerical sample : $x_1, x_2, x_3, ..., x_n$

3.1 Uni-variate

3.1.1 Measure of central tendency

Below mentioned are the measures of central tendency used in this report.

Arithmetic Mean The arithmetic mean is computed by adding a group of numbers and dividing that by the count of the numbers used in the group. The *sample mean* can be computed using the below formula:

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

The above formula can also be used to compute the population mean (μ) of a known population. (Christopher Hay-Jahans, 2018, p. 73)

Median When a data collection is organized in ascending order, the median is the middle element. It can be thought of as the geometric middle, whereas the mean is the arithmetic middle. For its computation, the above sample $x_1, x_2, .., x_n$ is assumed to be sorted in ascending order. *Median*($\hat{x}$) is then calculated as:

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

(Christopher Hay-Jahans, 2018, p. 75)

3.1.2 Measure of spread

The dispersion of the data is referred to as the spread. The common measures of spread used in this report are listed below.

Variance and Standard Deviation The standard deviation is a measure that describes how far each data point deviates from the actual mean. Squaring the standard deviation yields variance.

Sample variance (s^2) and Standard Deviation (s) can be calculated as follows:

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

$$s = \sqrt{s^2}$$

(Christopher Hay-Jahans, 2018, p. 76)

3.1.3 Measure of position

Interquartile Range We will understand quartiles before discussing the interquartile range (IQR). The quartiles divide the data into four groups of equal size and are made up of five factors, that are: the smallest number in a data-set, Q1 or quartile 1 which is the number that separates the lowest 25 percent of the data-set, Q2 or quartile 2 or also known as median, which is the number in the middle of sorted data-set, Q3 or quartile 3 which is the number that separates the lowest 75 percent of the data-set and the last factor is the largest number in the data-set. The interquartile range is then given by the difference between the third quartile and the first quartile. (Christopher Hay-Jahans, 2018, p. 77)

3.2 Bi-variate

3.2.1 Correlation Coefficient

Correlation coefficient *(r)* indicates the existence of a linear relationship between two random variables.

The correlation coefficient is given by:

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

(Christopher Hay-Jahans, 2018, p. 321)

Interpreting values of r Coefficient of correlation accepts values ranging from +1 to -1, with closer the coefficient is to either extreme, the stronger the relationship is. Value of r, that are close to 1 indicate a strong positive relationship, value of r close to -1 indicate a strong negative relationship and lastly the value of r close to 0 indicate the absence of a linear relationship at all. (Christopher Hay-Jahans, 2018, p. 322)

3.3 Histogram

A histogram displays a frequency distribution of a variable. The variable whose frequency is to be measured is on the X-axis, while the actual frequency of the variables is on the Y-axis. It takes continuous or discrete measurements and divides them into bins, or value ranges. The varied heights of the bins represent the data's frequency of occurrence. We are using a density histogram rather than a frequency histogram in this report. The vertical scale of density histogram indicates units that add up to one in the total area of all the bars.

The histogram is used in identifying the data central tendency, spread, the shape of the distribution, and outliers. (Christopher Hay-Jahans, 2018, p. 131)

3.4 Scatterplot

A scatterplot is a two-dimensional graphical representation that can be used to display pairs of data. Each point's position represents the value of the related variables, as well as the patterns between them. When these patterns resembles a line then an linear relationship exists between these two variables. Assuming that the horizontal and vertical axes are referred to as X and Y axes, the scatterplot can be interpreted as follows: If there is an upward trend then it indicates a positive relationship between X-axis and Y-axis and if there is a downward trend then it indicates a negative relationship between the axes. If there is no pattern or trend then it indicates no relationship between the X and Y axis. (Christopher Hay-Jahans, 2018, p. 159)

A matrix of scatterplot is known as pair plot. It generates a matrix of pairwise relationship between each variable in the data set which helps to quickly examine the data.

3.5 Boxplot

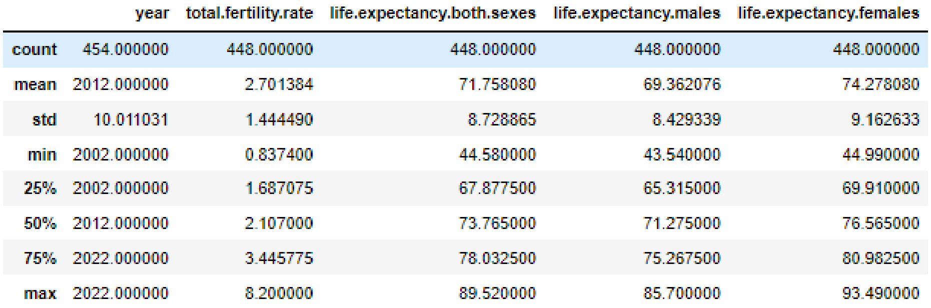
A boxplot is a visual representation of data distribution based on five numerical values, which are: minimum (smallest value in the data set), first quartile (Q1), median (Q2), third quartile (Q3) and maximum (highest value in the data set). This five number summary helps in comparing and examine data more clear. Whiskers are the horizontal lines that stretch out from the box, and the interquartile range is the box itself in between. Data points outside the box plot's whiskers are known as outliers.

Boxplot are used to estimate the skewness, spread, symmetry and outliers within a data set. (Christopher Hay-Jahans, 2018, p. 137)

4 Statistical analysis

To gain a better understanding of data, the statistical approaches mentioned in Section 3 are applied to the entire data set.

Description of the Data Set Table 1 can be used to understand description of the data set since it summarizes the central tendency, dispersion, and form of each numerical variable.



	year	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
count	454.000000	448.000000	448.000000	448.000000	448.000000
mean	2012.000000	2.701384	71.758080	69.362076	74.278080
std	10.011031	1.444490	8.728865	8.429339	9.162633
min	2002.000000	0.837400	44.580000	43.540000	44.990000
25%	2002.000000	1.687075	67.877500	65.315000	69.910000
50%	2012.000000	2.107000	73.765000	71.275000	76.565000
75%	2022.000000	3.445775	78.032500	75.267500	80.982500
max	2022.000000	8.200000	89.520000	85.700000	93.490000

Table 1: Description of the Data Set

4.1 Uni-variate analysis

Uni-variate analysis is performed on each numerical variable individually. The following analysis can be made from Figure 1:

Figure 1 shows the histogram plot of the four numeric variables present in the data set. As seen from the histogram plot of Figure 1, the distribution of total fertility rate has right skewness, whereas the plot of the other three life expectancy plots shows left skewness. The histogram of total fertility rate is at peak at around two indicating

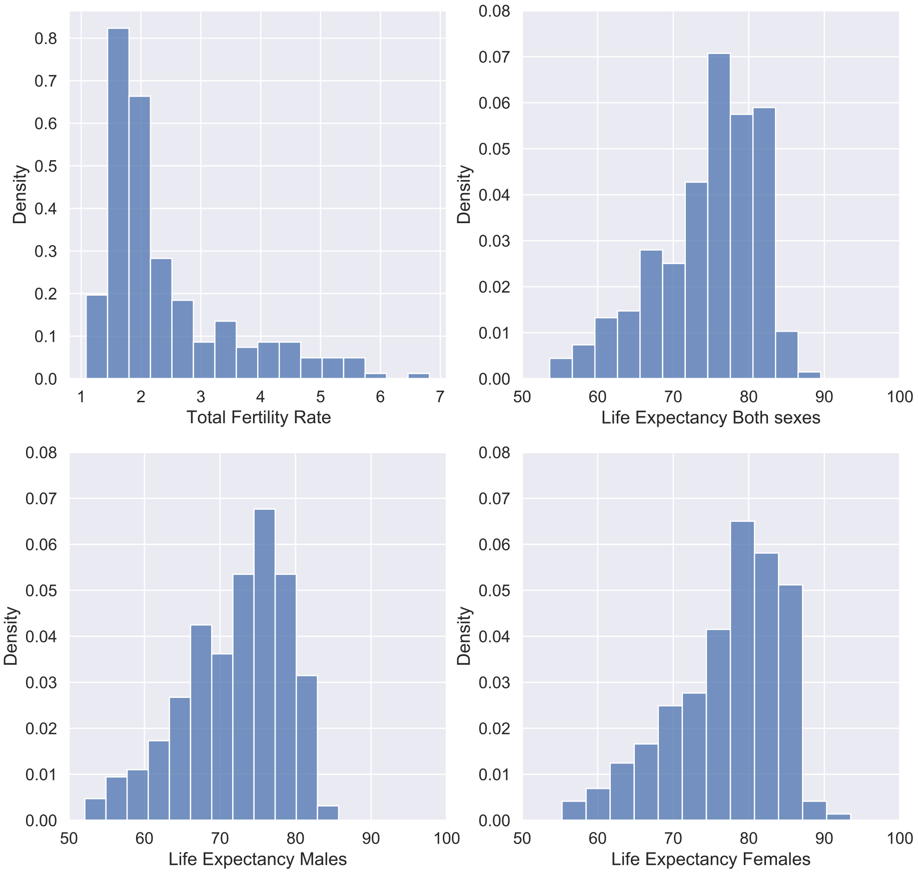


Figure 1: Density distribution of numerical variables

that the maximum number of women had two children on an average in the year 2022. Similarly, the frequency distribution of life expectancy is maximum between 75 to 78 years for both sexes. Considering gender-wise life expectancy, it can be observed that the average life expectancy of males and females are 75 and 80 years respectively. It can be said that, in the year 2022, the female life expectancy is slightly greater than males. A joint scatterplot is used to gain better understanding on how life expectancy varies between the sexes.

It can be illustrated from Figure 2 that, depending on sexes, females might be deemed to have a longer life expectancy at birth than males in the year 2022. While there are

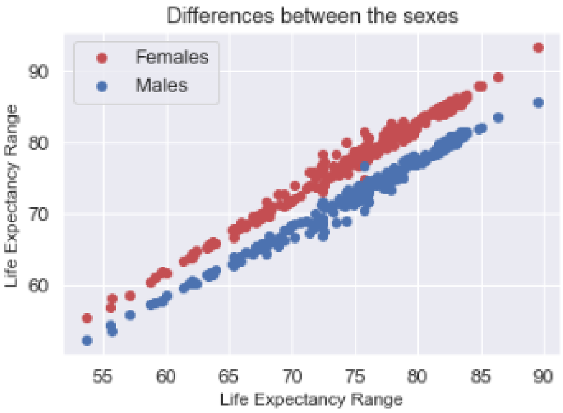
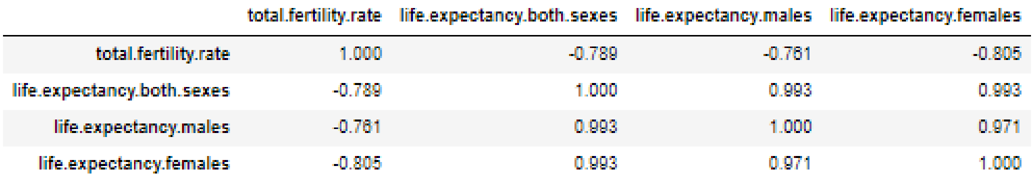


Figure 2: Difference in sexes in terms of life expectancy

numerous ongoing studies and speculations to back this up, we have yet to acquire a definitive basis for the aforementioned result.

4.2 Bi-variate analysis

In this section, we will use the correlation coefficient to check for any underlying relationship between the variables in the data set. The correlation coefficient values between numerical variables in the data set are shown in Table 2 and also through a heat map in Table 4 (Appendix). The table's diagonal value can be ignored because it is the coefficient value of variables with itself, which is always equal to one. It can be observed from Table 2 that there exists a strong positive correlation between the three variables i.e. life expectancy at birth for both sexes, males and females which is expected given that these factors are related and derived from one another.



	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
total.fertility.rate	1.000	-0.789	-0.761	-0.805
life.expectancy.both.sexes	-0.789	1.000	0.993	0.993
life.expectancy.males	-0.761	0.993	1.000	0.971
life.expectancy.females	-0.805	0.993	0.971	1.000

Table 2: Calculation of correlation coefficients

A similar result could be seen using a pair plot from Figure 3, that shows a strong positive linear (scatterplot resembles straight line) relationship between life expectancy for both sexes, males and females. On the other hand, negative value of correlation coefficient implies an inverse or negative monotonic relationship between total fertility rate and life expectancy. This implies that decrease in total fertility rate shows increase in the life expectancy value and vice versa. Figure 3 illustrates a negative slope of line between total fertility and life expectancy without any trace of curve and this results in negative linear relationship between total fertility rate and life expectancy at birth.

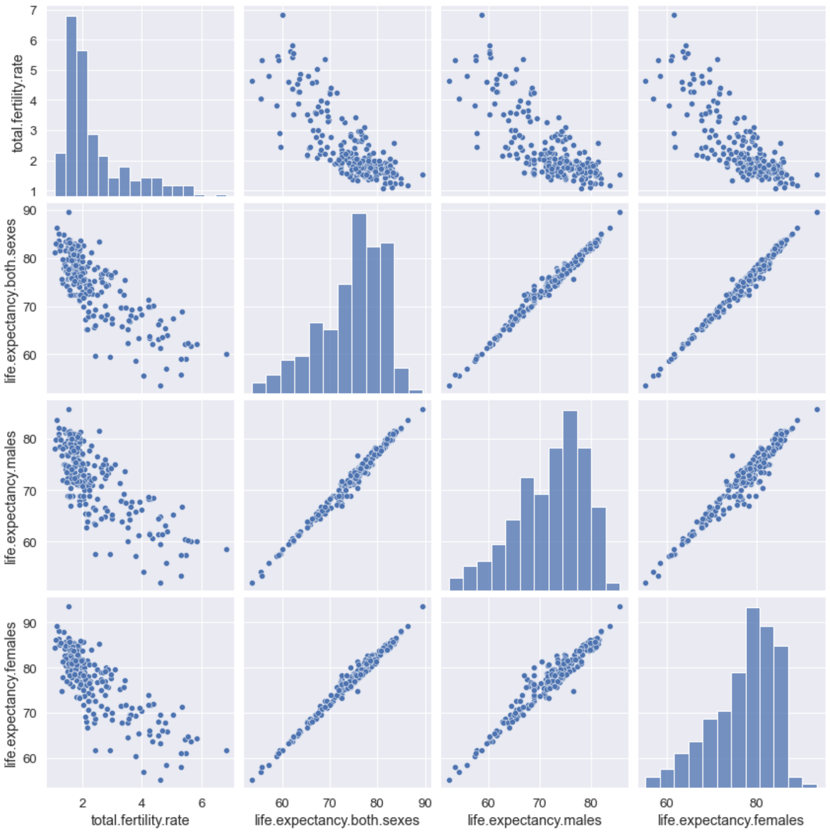


Figure 3: Bi-variate analysis of numerical data using pair-plot

4.3 Variability of the values in the individual and different sub-regions

There are 21 sub-regions and 5 regions, as described in Section 2. Observing data within and across the sub-regions is therefore quite interesting.

Total Fertility Rate Figure 4 depicts the total fertility rate across different sub-regions. We can observe that the data spread or variability of the total fertility rate for Africa and its sub-regions is higher compared to other regions and sub-regions. Southern Africa

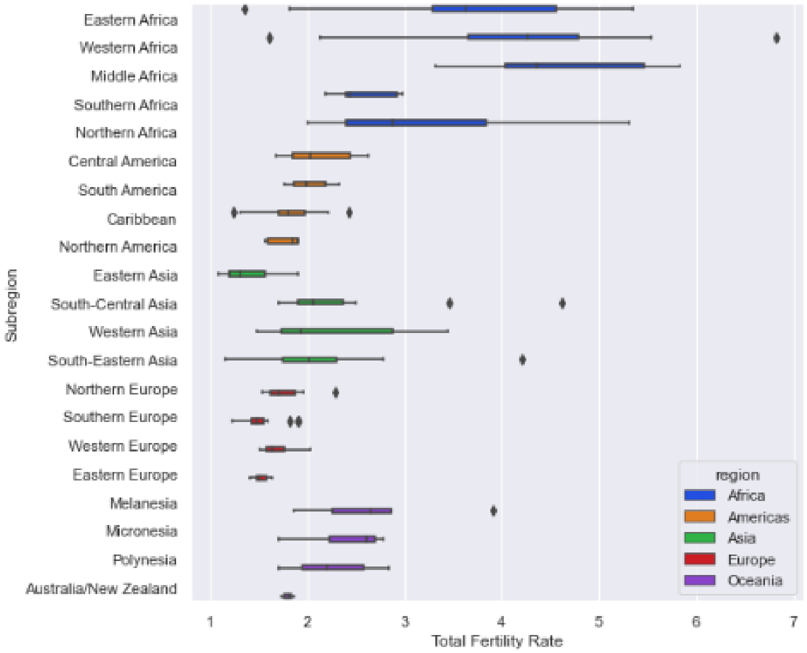


Figure 4: Total fertility rate across the world

shows a very less variability in its data as compared to other African sub-regions. On the other hand, the variability of the total fertility rate for America and Europe is less when compared to other regions and sub-regions. All the sub-regions of America and Europe have a total fertility rate ranging between 1.5 and 2.5 and the range for European regions is much smaller compared to America. Most of the European have children between 1 and 2 and none more than 2. Oceania and Asia region have variability ranging from 1.8 to 3 and 1 to 3.5 respectively. Australia and Western Asia stand out when compared within their sub-regions as the total fertility rate of Australia is 1.8 with negligible variability and Western Asia is between 1.5 and 3.5 with maximum variability within its sub-regions.

Life Expectancy at Birth Figure 5 depicts the life expectancy at birth for both sexes across different sub-regions. We can observe a large variability in the data for Africa where the life expectancy ranges between 58 to 75 years. On the other hand, European sub-regions show less variability when compared to other sub-regions. The severe health crisis that is raging in Africa is to blame for the poor life expectancy and people of Europe live for the most number of years due to the excellent healthcare system. The life expectancy for Asia and America ranges between 67 to 87 years and 68 to 83 respectively. The sub-region, Northern America stands out among the other sub-regions of America

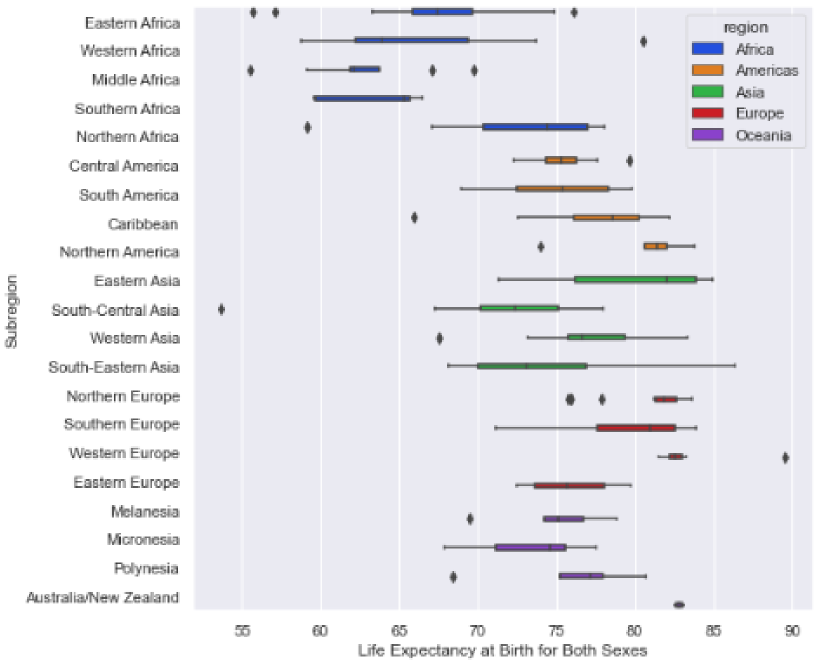
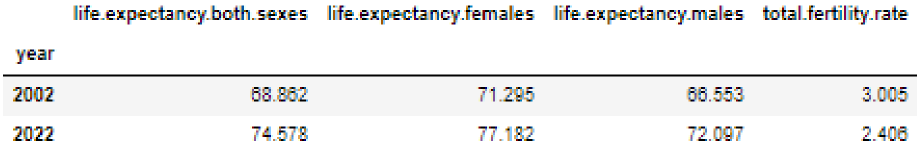


Figure 5: Life expectancy at birth for both sexes

because it has very less variability in life expectancy ranging from 81 to 83 years. Figure 7 and Figure 8 from the Appendix can be referred to observe life expectancy at birth for males and females across different sub-regions..

4.4 Comparison of variables from 2002 to 2022

Table 3 shows the average change in variables during the last 20 years. It can be interpreted that the fertility rate has declined from 3 to 2.4, but life expectancy has improved by 6 years. It should also be noted that the preceding phrase refers to the global population rather than population by region, subregion, or country. We will now look at the same data region by region. To accomplish so, we will use a box plot to compare the results.



	life.expectancy.both.sexes	life.expectancy.females	life.expectancy.males	total.fertility.rate
year				
2002	68.862	71.295	66.553	3.005
2022	74.578	77.182	72.097	2.406

Table 3: Variable change over 20 Years

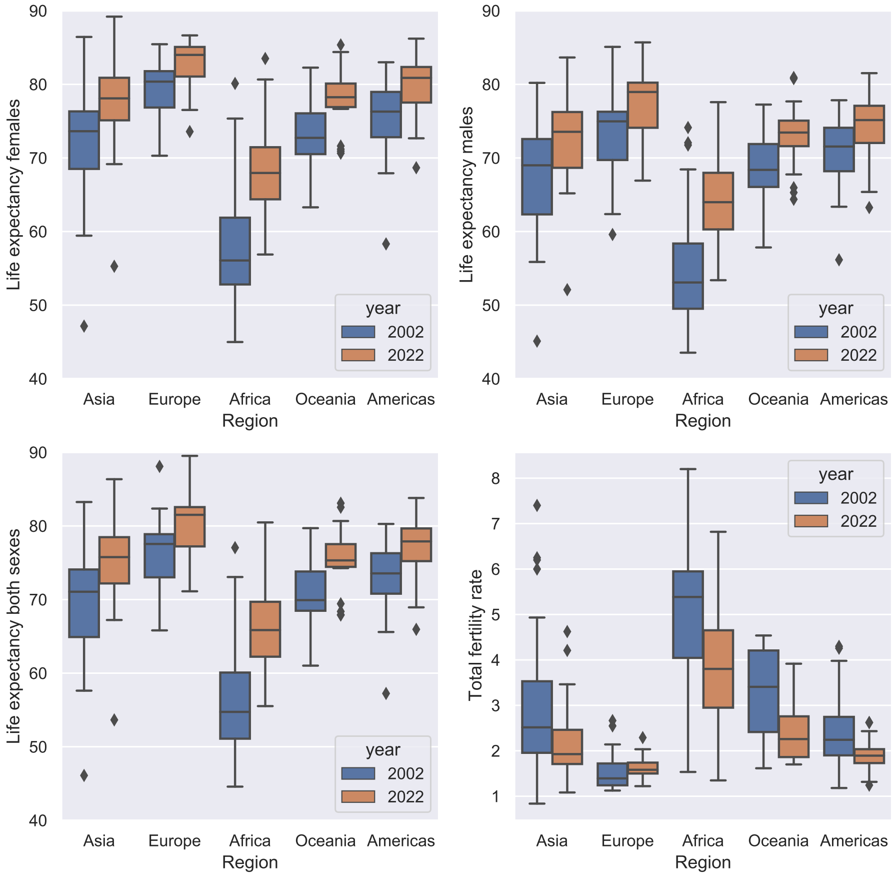


Figure 6: Comparison of change in variables in every region over year

It can be depicted from Figure 6 that, in terms of life expectancy, African region has seen a big increase, whereas the rest of the regions have seen a much smaller gain. On the other hand, we witness a significant decline in overall fertility rates in every region except Europe, which has been relatively constant over the last 20 years. The following statement can also be made based on Figure 6 that the European average fertility rate is lower than the world's average fertility rate, while Africa's fertility rate is higher than the world's average fertility rate.

5 Summary

The data set used in this report is a small extract from IDB (International Data Base) of the U.S. Census Bureau. It contains eight features and 454 records and composes data on 228 countries, which are grouped into 21 subregions and five regions based on their demographics. The entire dataset provided is for the year 2002 and 2022.

The primary goal of this report is to present a comprehensive descriptive analysis of the data. Firstly, frequency distribution for each individual variable was described and displayed using histograms that resulted in global average fertility rate and life expectancy at birth for both sexes at around 2 and 75 respectively. The differences between the life expectancy of individual sexes were explained using a scatterplot that revealed that the average life expectancy at birth for females is more than that of males. Secondly, the bivariate association between variables was discovered using a correlation coefficient which resulted in a negative connection between fertility rate and life expectancy. We also observed variability in total fertility rate and life expectancy for various regions and sub-regions. America and Europe tend to show less variability in the data whereas Africa depicted high variability compared to other regions and subregions. This could be related to differences in lifestyle choices and the quality of healthcare systems between countries. Finally, the data were compared for the years 2002 and 2022, and it was discovered that, while total fertility rates declined in 2022, life expectancy improved. However, because the population shift over time is not taken into account, this data is incomplete.

It would be more interesting for further research if the country's population, birth rate and death rate could also be taken into account for the yearly change in variables as this will lead to a more accurate study by allowing us to pinpoint the particular reasons that cause such changes to occur.

Bibliography

Christopher Hay-Jahans. *An R Companion to Elementary Applied Statistics*. Taylor and Francis Group, London, NewYork, 2018.

(International Data Base). Glossary for census data. URL `https://www.census.gov/glossary/`.

Python Programming Language. Python software foundation. URL $http://www.python.org$.

Appendix

A Additional tables

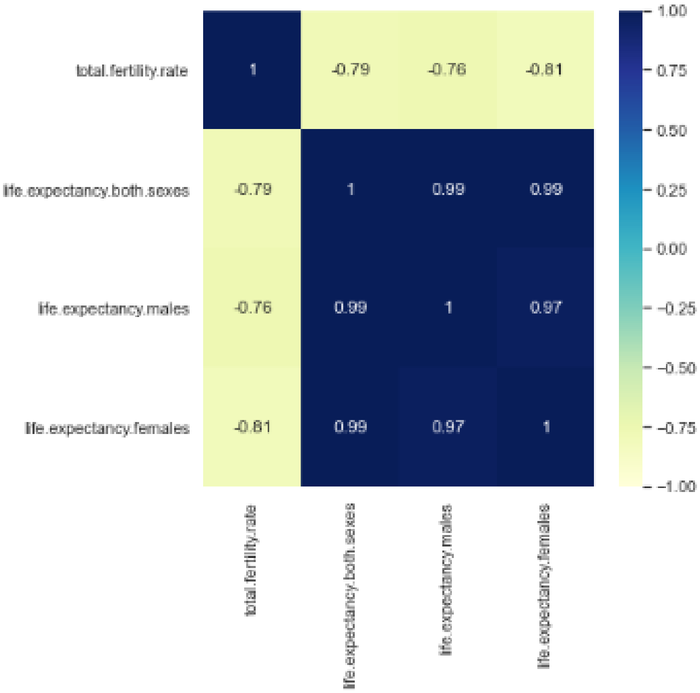


Table 4: Correlation Coefficient using Heatmap

B Additional figures

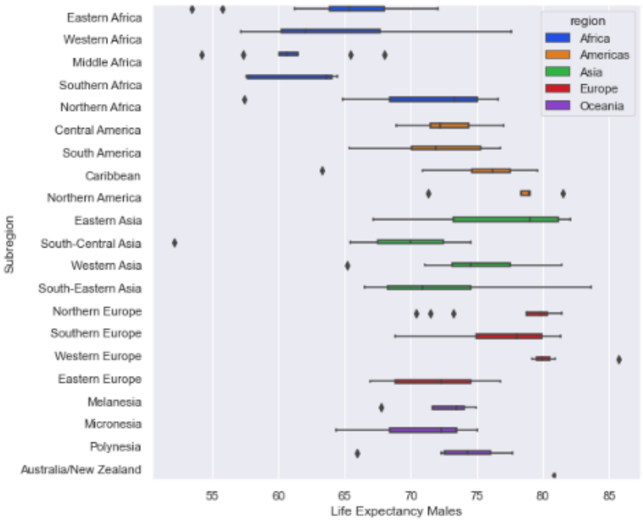


Figure 7: Life expectancy of males for 2022 across the world

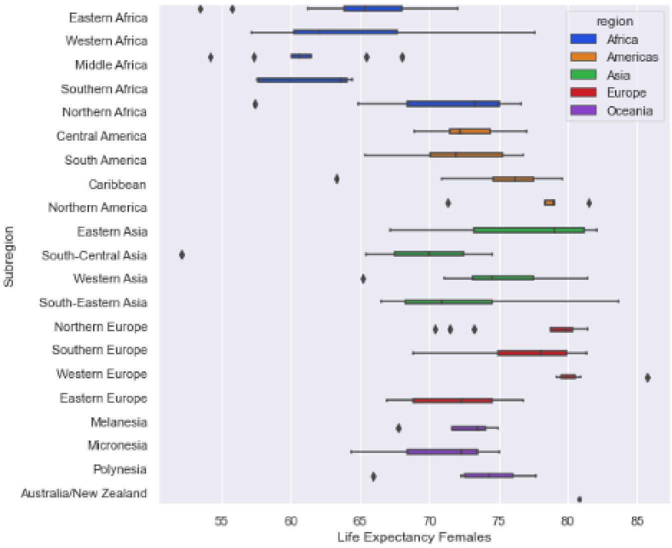


Figure 8: Life expectancy of females for 2022 across the world