Contents

1 Introduction **1**

2 Problem statement **1**

2.1 Description of dataset 1

2.2 Project objective 2

3 Statistical methods **3**

3.1 Measures of central tendency 3

3.2 Measures of spread 4

3.3 Measure of position 4

3.4 Correlation coefficient 5

3.5 Plots for numeric data 5

4 Statistical analysis **6**

4.1 Frequency distribution of variables 6

4.2 Bivariate correlations 8

4.3 Variability analysis 9

4.4 Comparison of variables between 2002 and 2022 10

5 Summary **11**

Bibliography **13**

Appendix **14**

A Additional figures 14

1 Introduction

A country's demographic statistics explains the present state of the people in country in terms of literacy, employment, mortality rate, and various other variables. Furthermore, it is beneficial for comparing these variables within country's regions, as well as between nations. Collecting this data on a regular basis aids in the monitoring of changes in the variables and components that cause these changes. Moreover, it can also be helpful to make decisions related to the country's progress.

The aim of this report is to compare the values of life expectancy and total fertility rate and notice the changes in the values of these variables between the years 2002 and 2022. To achieve this, we perform the following tasks. First, we see the frequency distribution of the variables total fertility rate, and the life expectancy using graphs and notice the patterns of the variables. Second, we use the bivariate correlation to check the connection between the variables. In addition, we also check for monotonic relations between the variables. Third, we interpret how the values of these variables are scattered within and between subregions. Finally, we report the changes in the values of the variables from 2002 to 2022.

In Section 2, a brief description of the dataset is provided. In addition, the variable definitions and the project objectives are presented in more detail. In Section 3, different statistical methods relevant to this project are elaborated on and explained. In Section 4, we apply the statistical methods explained in the previous section to the project dataset and interpret the results. The final section (Section 5) will summarize our findings and ideas for further possible analysis of the given dataset.

2 Problem statement

2.1 Description of dataset

This project makes use of a small part of demographic data from the U.S Census Bureau's International Data Base (IDB) for the years 2002 and 2022. The IDB is a collection of demographic statistics for more than 200 countries with populations greater than or equal to 5,000 from 1950 to this year. It also holds the projection of demographic data until 2060. Variables in the dataset include *total fertility rate* and *life expectancy* at birth for 228 countries. (International Data Base, 2022).

The dataset includes 454 observations and 8 columns in total. The columns can also be termed as independent variables. This data is a survey on demographic statistics from 228 countries. The countries are grouped into 21 subregions and again grouped into 5 regions based on geography. The variable *country* refers to the name of the country from which the observation is collected (among 228 countries). *Subregion* refers to the name of the subregion that contains the country, (among 21 subregions). *Region* refers to the name of the geographic region that contains the country and the subregion, (among 5 regions). *Year* represents the particular year in which the data is noted (2002 or 2022). *Total fertility rate* refers to the average number of children a woman could give birth to, assuming that each woman lived until their childbearing years. *Life expectancy* refers to the average number of years a group of people born in the same year can expect to live if death rates at each age remain constant in the future. The life expectancy is recorded in general and then separately for males and females.

The data set contains missing values in 6 observations out of 454 observations. As the data set is an extract from standard and recognized webpage and missing values doesn't from a major part of the dataset, the overall data quality is decent. The missing values are excluded to simplify the analysis, to maintain the representativeness of the sample, etc.

2.2 Project objective

In this report, first, we examine the dataset and plot graphs for the variables to understand the frequency distribution of the variables. In each task, the variables life expectancy and total fertility rate are examined in great depth. A scatterplot is used to illustrate the disparities in life expectancy between the sexes. Second, bivariate analysis is used to figure out how one variable is linked to others (correlation). We further check for monotonic relationships. Third, the boxplots are used to assess the variability of the variables within and between the subregions. The countries are split into 5 regions and 21 subregions in our dataset. For the above tasks, we only use the data from the year 2022. Finally, we interpret the changes in values of the variables over the course of last 20 years (2002 and 2022).

3 Statistical methods

The statistical methods essential for the analysis of the data set are explained in detail, along with their formulas. All the following graphs and calculations are created using the software Python in version 3.9.7.

To understand formulas better, let us consider a sample $x_1, x_2, x_3, ..., x_n$ with n *observations*. (Python Core Team, 2021).

3.1 Measures of central tendency

Arithmetic Mean

The *mean* can be stated as the addition of all the n values in the sample divided by the total number of values or observations. The formula to calculate the *mean* is as follows

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

The *mean* is different for each sample and depends on the sample observations. This calculation is sensitive to far left values or far right values in the sample. This makes it highly suitable for the fairly homogeneous datasets. If the dataset includes far left or far right values then it is preferred to consider other measures of central tendency like *median* etc. (Christopher Hay-Jahans, 2020, p.73).

Median

Given that the observations are grouped in ascending order, the *median* can be expressed as the mid value of the sample observations. If the sample is grouped in ascending order, the formula to find the *median* is as follows

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

This measure is not affected by the leftmost or rightmost values in the sample, it depends on the sample observations. Hence, it may be the preferred measure if the dataset contains outliers. (Christopher Hay-Jahans, 2020, p.75).

3.2 Measures of spread

Variance and Standard Deviation

The degree of dispersion between observations in the sample and it's *mean* can be expressed as the *variance* of the sample. It measures the deviation of a value from the *mean*. The square root of the *variance* can be expressed as the *standard deviation* of the sample. The formula to find the *variance* and *standard deviation* of a sample is as follows

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

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

(Christopher Hay-Jahans, 2020, p.76).

3.3 Measure of position

Percentiles and Quantiles

Consider the observations in the sample arranged in ascending order, the p^{th} percentile is the number q_p corresponding to the percentile rank 100 * p%, and calculated using the following rule

$$q_p = \begin{cases} x_k, & \text{where } k = \lceil np \rceil \text{ and } np \text{ is not an integer.} \\ (x_k + x_{k+1})/2, & \text{where } k = np \text{ and } np \text{ is an integer.} \end{cases}$$

, here $\lceil . \rceil$ represents ceiling function.

- First quartile (Q_1) = 25^{th} percentile
- Second quartile (Q_2 *or median*) = 50^{th} percentile
- Third quartile (Q_3) = 75^{th} percentile.

(Ludwig Fahrmeir, 2020).

3.4 Correlation coefficient

For any two continuous variables, if there exists a linear relationship between them then it can be expressed as *correlation coefficient*. The formula for the *correlation* is as follows

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

r_{xy} will have the values from -1 to +1. For a positive correlation, the value of r_{xy} range from 0 to 1 and for a strong positive correlation, the value is close to 1. Correspondingly, in the case of negative correlation the value ranges from -1 to 0 and for a strong negative correlation the value is close to -1. Finally, if there is weak or no correlation between the variables then the value of r_{xy} is close to zero.

(Christopher Hay-Jahans, 2020, p.321).

3.5 Plots for numeric data

Histogram

A *histogram* is an efficient way to show the frequency distribution of continuous variables in the data set. It can be stated as graphical representation of a continuous variable, where the variable values are grouped into number ranges. The horizontal axis can contain a set of values and can be termed as buckets. The vertical axis can contain the frequency of the data present in that range as a percentage. The frequency is indicated by the height of the bin. The height of the bin is shorter for data with fewer occurrences, and it is longer for data with more occurrences. (Christopher Hay-Jahans, 2020, p.131).

Boxplot

The *boxplot* can be defined as representing the 5-digit summary (minimum, first quartile, median, third quartile, maximum) of the variable values in the graphical format. *Boxplots* are more useful for understanding variability in data. The interquartile range (IQR) can be defined as difference between third quartile (Q_3) and first quartile (Q_1). Outliers can also be visualized in plots and are also useful when comparing multiple datasets. *Boxplots* are visualized in both horizontal and vertical directions according to the user's interests.

$IQR = Q_3 - Q_1$.

(Christopher Hay-Jahans, 2020, p.137).

Scatterplot

A scatterplot consists of predictor variable (horizontal axis), response variable (vertical axis) and dots of these ordered pairs. It can be used to understand the connection between two variables by analyzing the pattern they follow from the ordered pair graph. The graph can also be helpful to get information on outliers. (Christopher Hay-Jahans, 2020, p.159).

4 Statistical analysis

The statistical methods described in the previous section are tested on the dataset used in this report and the results are interpreted in this section.

In Table 1, we can see the summary of the results.

	total.fertility.rate	**life.expectancy.both.sexes**	**life.expectancy.males**	**life.expectancy.females**
count	448.0	448.0	448.0	448.0
mean	2.70	71.75	69.36	74.27
std	1.44	8.72	8.42	9.16
min	0.83	44.58	43.54	44.99
25%	1.68	67.87	65.31	69.91
50%	2.10	73.76	71.27	76.56
75%	3.44	78.03	75.26	80.98
max	8.20	89.52	85.70	93.49

Table 1: Descriptive statistics of dataset

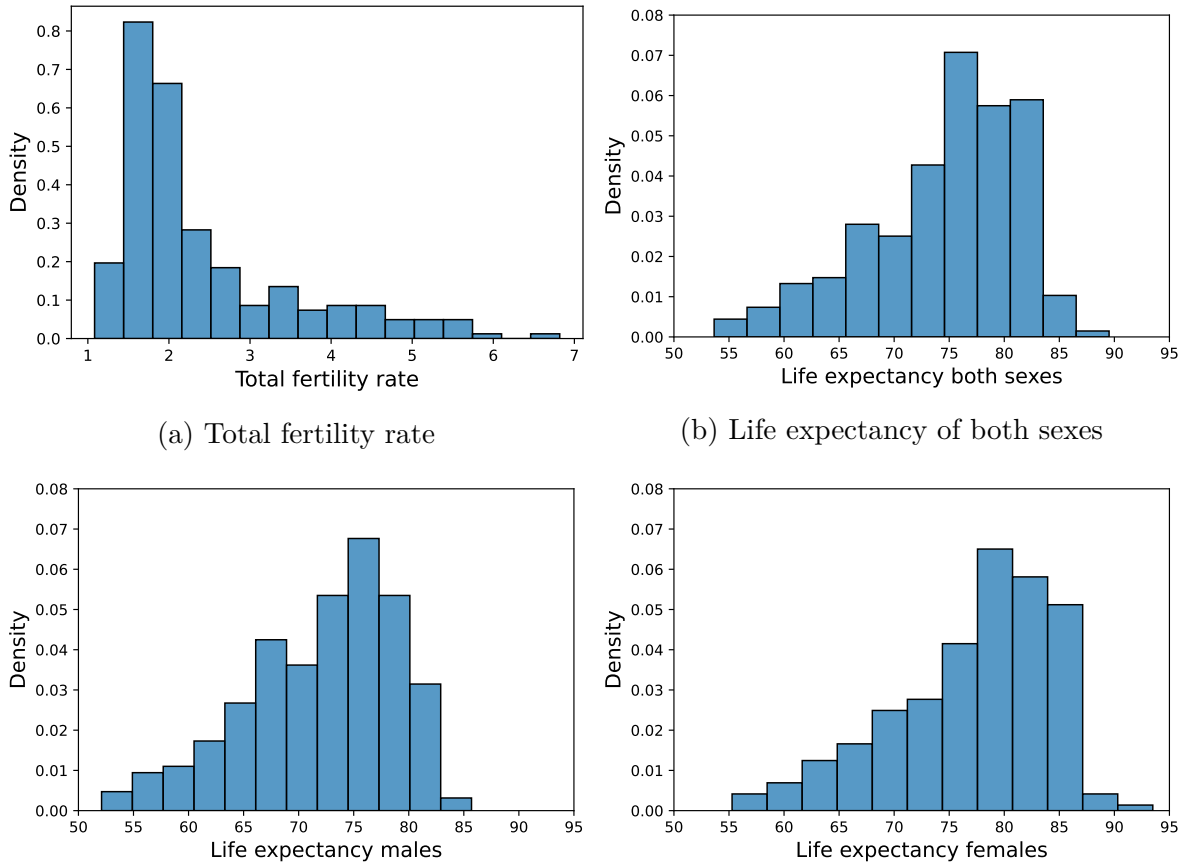
The count refers to the number of non-null observations in the dataset. The table also includes the mean, standard deviation, minimum, maximum and 3 percentiles of each variable. This is a general summary including both years (2002 and 2022).

4.1 Frequency distribution of variables

Figure 1 consists of 4 subfigures, which are the frequency distributions of variables *total fertility rate, life expectancy of both sexes, life expectancy of males, and life expectancy of females*. From the graph 1(a), we can see that it's a positive or right-skewed distribution. We can also see that the graph has a maximum point between 1 and 2 and values range from 1 to 7. From this, it is evident that in the year 2022, the majority of the women

have 1 or 2 children. Moreover, the frequency seems to decline as the fertility rate increases and no women have more than 7 children.

The frequency distribution of *life expectancy for both sexes* in 2022 is shown in graph 1(b). We can see that both sexes have a maximum frequency of *life expectancy* of around 75 years and a range of 50 to 90. From the graph and above table, it is evident that there are no people with a *life expectancy* less than 44 and greater than 90. From table 1, we can say in general that a person born in 2022 can expect a life expectancy between 71 and 72 years (mean value). We can notice a similar pattern in the graph of the *life expectancy of females* (graph 1(d)) in the same year.



(a) Total fertility rate

(b) Life expectancy of both sexes

(c) Life expectancy of males

(d) Life expectancy of females

Figure 1: Frequency density distribution of numerical variables

Figure 2 shows the differences between the sexes with respect to the variable *life expectancy*. We can see that the points are scattered in the form of a line. The orange points (representing females) have greater values than the blue points (representing males). From this, we can say that females have a longer life expectancy than males.

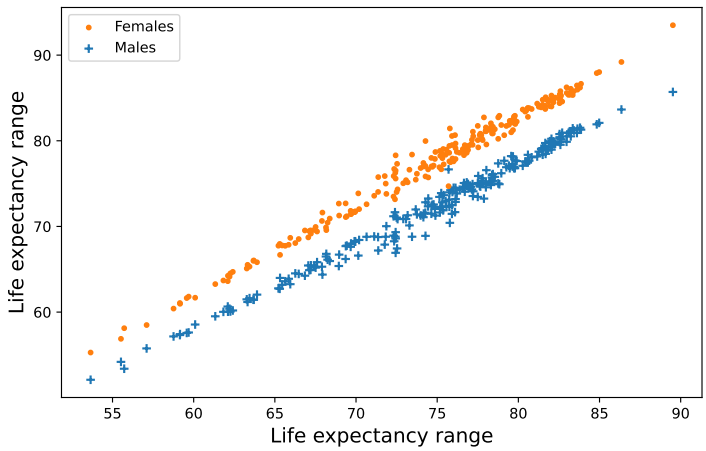


Figure 2: Difference in life expectancy between sexes

In addition, from the table and figures 1(c) and 1(d), we can notice that the frequency of *life expectancy of males* attains a maximum of around 75 whereas the frequency of *life expectancy of females* attains a maximum of around 80. According to the above statements, we can say that women have a longer life expectancy in 2022 than men.

4.2 Bivariate correlations

In this part, we do a bivariate analysis on the variables *life expectancy* and *total fertility rate* to understand the relationship between them. The pair plot is shown in figure 3.

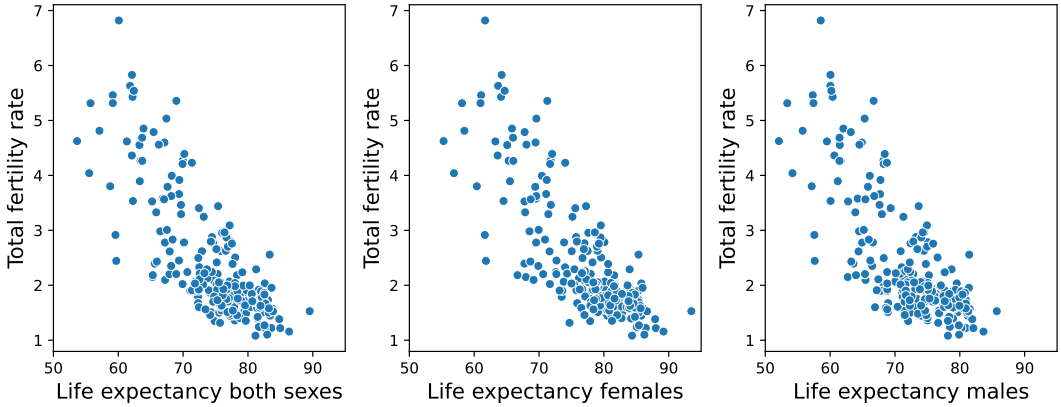


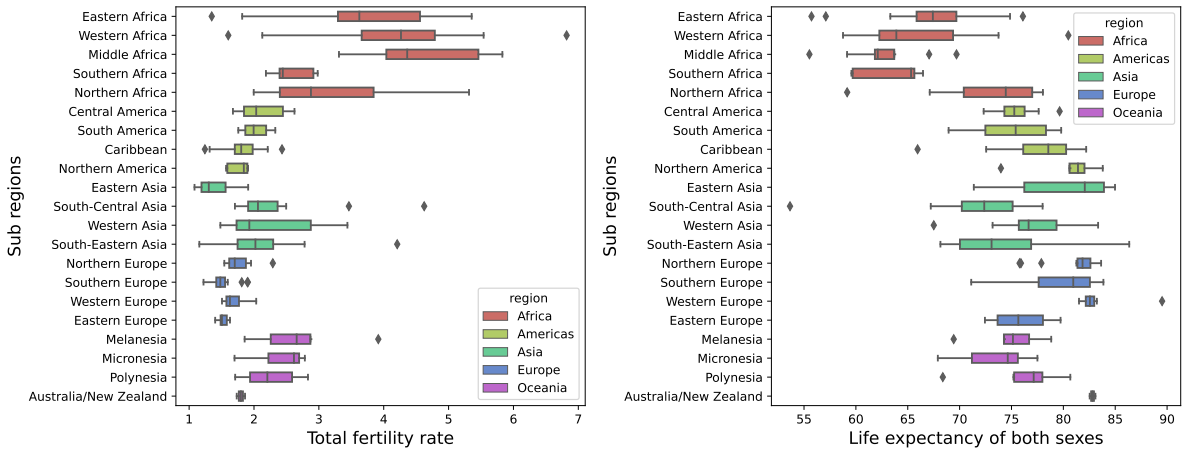
Figure 3: Correlation between Total fertility rate and Life expectancy

From the figure, we can notice that as the *life expectancy* value increases the *total fertility rate* value decreases or vice versa. From this, it is evident that the variables *life expectancy* and the *total fertility rate* are negatively correlated.

From the above relation, we can also say that there exists a hint for the monotonic relationship as it satisfies one of it's properties (i.e. the value of one variable increases, and the other variable value decreases). Therefore, we can comment that there exists a monotonically decreasing linear relationship.

4.3 Variability analysis

In this part, we will perform an analysis to check how the variables *total fertility rate* and *life expectancy* are varied inside the subregions and between the subregions. The observations from the dataset belong to 21 subregions in total. To analyze the variability within the subregion, we check the length of the interquartile range. If the length of the interquartile range is high, then there exists a large variability between the values or else less variability. Correspondingly, to analyze the variability between the subregions, we check the range of the boxplot. If the value of the range is low, then there is less variability in variable values or else high variability.



(a) Total fertility rate (b) Life expectancy of both sexes

Figure 4: Comparison of variability within and different subregions

Total fertility rate

Figure 4(a) shows a boxplot of the variable *total fertility rate* in all 21 subregions. From the plot, we can notice that within each subregion of *Africa* the interquartile range and the range of the boxplot are high except for *southern Africa*. This shows that there is a huge variability of values within and between the subregions of *Africa* (except *southern Africa*). A reverse pattern can be observed in the subregions of *Asia*, where the length

of the interquartile range and the range of the boxplot is low except for *western Asia*. Therefore, the subregions of *Asia* will have low variability of values within and between the subregions (except for *western Asia*). We can also see that the subregions of *Europe* and the *Americas* have less variability as both the interquartile range and the range of the boxplot looks small and lies between 1 and 3. We can also see some outliers in some subregions of *Europe*. The subregions of *Oceania* also have less variability of values within and between subregions except for *Australia /New Zealand*, which has even less variability. Finally, we can comment that the subregions of *Africa* have high variability of values within and between subregions, whereas the subregions of *Europe* have low variability of values.

Life expectancy

Figure 4(b) shows a boxplot of the variable *life expectancy* rate in all 21 subregions. We can notice that in the subregions of *Africa* there is a huge variability of the data within and between subregions as the interquartile range and range of the boxplot is very high (except in *Middle Africa*). We can also see many outliers in the subregions of *Africa*. A similar pattern can be observed in the subregions of *Asia* where the variability of the data is high within and between subregions. In *Europe*, *northern* and *western Europe* has less variability within subregions, whereas *southern* and *eastern Europe* has high variability of the variable values. The *Americas* follow a similar pattern as *Europe*. In subregions of *Oceania* except for *Australia / New Zealand*, other subregions have high variability in both cases. Finally, we can comment that most of the subregions regardless of regions have high variability of values within and between subregions.

4.4 Comparison of variables between 2002 and 2022

In this subsection, we interpret the changes observed between the years 2002 and 2022 with respect to the *total fertility rate* and *life expectancy*. The results are visualized in the below figure.

From figure 5, in the total fertility rate plot, we can notice that the 2022 boxplots are smaller than the 2002 boxplots. On the other hand, the life expectancy graph shows that the *life expectancy* value has increased in 2022. From this, it can be said that from 2002 to 2022, the *total fertility rate* decreased and *life expectancy* increased. The *total fertility rate* in Europe looks similar for both years.

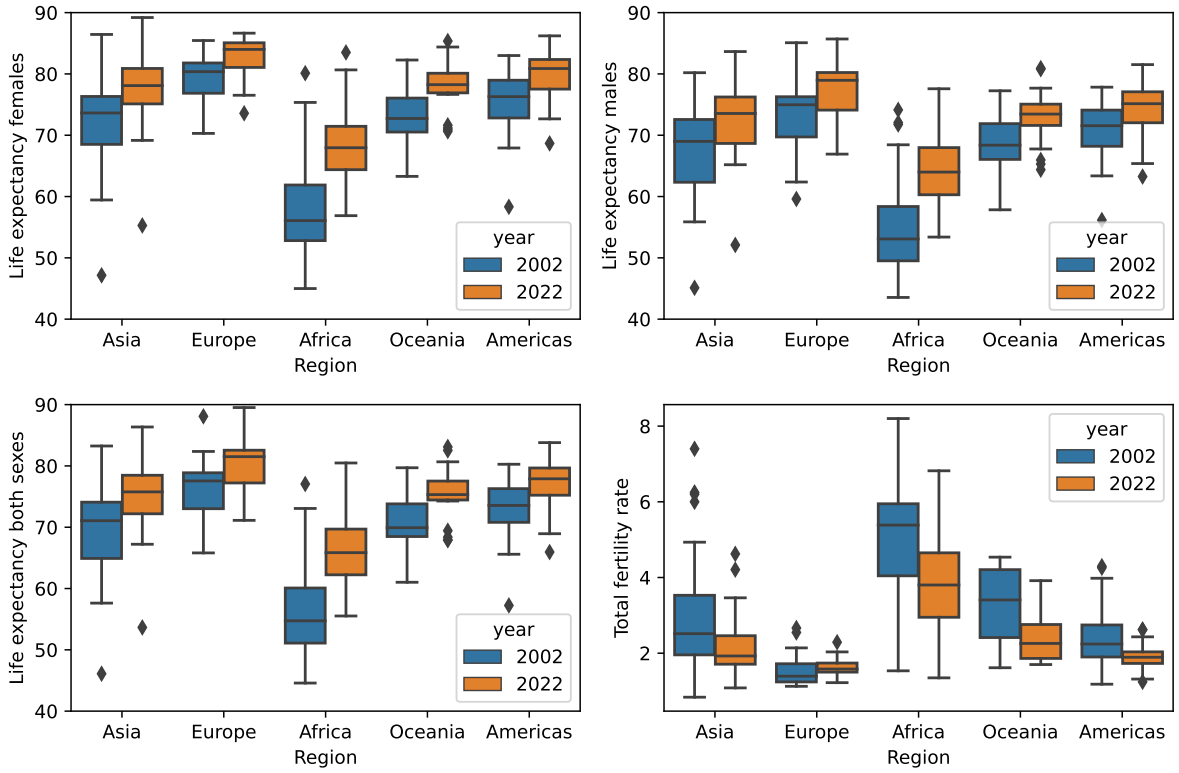


Figure 5: Comparison of variables over years 2002 and 2022

In general life expectancy seems to be increased. This may be because of advancements in healthcare facilities, people shifting towards a healthy lifestyle, etc. We can notice there is an exception in the life expectancy of males in Oceania where the maximum value of life expectancy stays almost same for both years.

5 Summary

In this report, descriptive analysis was performed on the dataset and the results are interpreted. The dataset used in this report was compiled by the instructors of the course Introductory to case studies at TU Dortmund University in the summer semester 2022. The dataset was a tiny extract from the U.S. Census Bureau's International Data Base (IDB), which is an archive for various demographic data from 1950 to till now. The website is a collection of demographic data in the form of the census, and surveys from more than 200 countries, which have a population of 5000 or more. It also has projections until 2060. The observations from the dataset represent data from 228 countries, which are grouped into 21 subregions and 5 regions geographically. The recorded obser-

vations belonged to two years namely, 2002 and 2022, and also each observation contains variables *country, region, subregion, year, total fertility rate, life.expectancy.both.sexes, life.expectancy.males, life.expectancy.females.* (International Data Base, 2022).

The dataset was first analyzed and a brief summary related to the variables of our interest was tabulated. It included the count, mean, standard deviation, and other measures for variables *total fertility rate* and *life expectancy.* In the second step, the frequency distributions of the variables are plotted and the observations from the graph are written. In addition, a scatterplot was used to compare the life expectancy of both genders. From this, we understood that the life expectancy of women is higher than men in 2022. In the third step, bivariate analysis was performed to decrypt the relationship between the variables. Here, we understood that the *life expectancy* and *total fertility rate* are negatively correlated and a monotonically decreasing relationship has been discovered between them. In the fourth step, we understood the variability of *life expectancy* and *total fertility rate* within and between subregions. Here, we noticed that the subregions of *Africa* had high variability for both variables. From boxplots, we understood that regions with low fertility rate have a higher life expectancy. This proves the relation obtained in the third step. Finally, we compared the variables *life expectancy* and *total fertility rate* between the years 2002 and 2022 and we noticed that over the years, the *life expectancy* has been increased and the *total fertility rate* has been decreased.

For further analysis, collecting more observations may give us more information relevant to the study. Also, including more predictor variables that affect the variables we are interested in may help us understand the unknown patterns and state the results more accurately.

Bibliography

Christopher Hay-Jahans. *An R Companion to Elementary Applied Statistics*. Taylor and Francis Group, London, NewYork, 2020.

International Data Base. *Glossary of census data*. United States Census Bureau, 2022. URL `https://www.census.gov/glossary/`. Visited on 3rd May 2022.

Thomas Kneib Ludwig Fahrmeir. *Regression Models, Methods and Applications*. Springer, London, NewYork, 2020.

Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, Version: 3.8.3, 2021. URL `https://www.python.org`. Visited on 3rd May 2022.

Appendix

A Additional figures

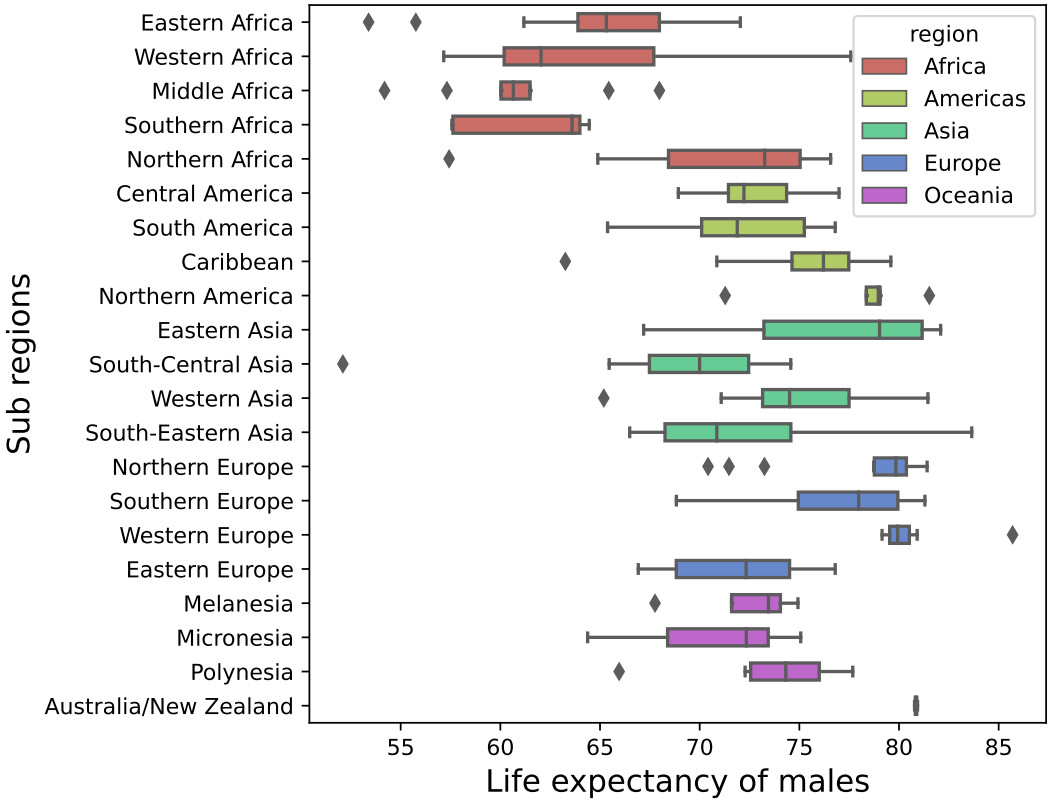


Figure 6: Life Expectancy of males in 2022 across the world

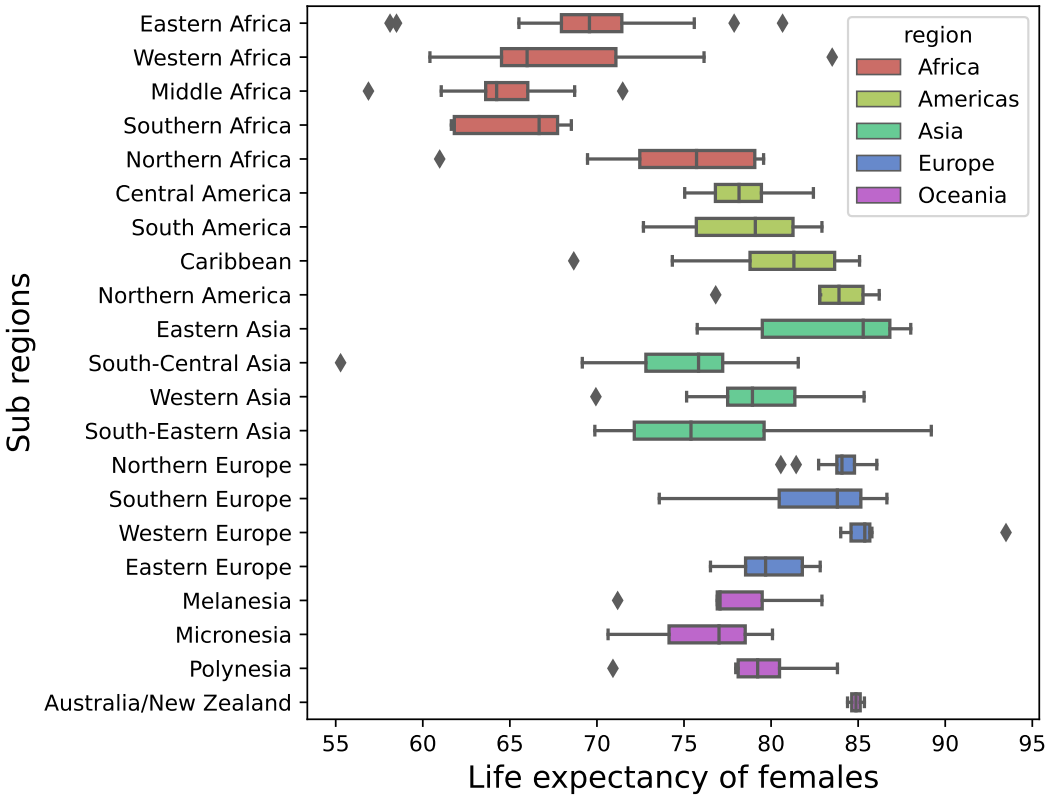


Figure 7: Life Expectancy of females in 2022 across the world