TU Dortmund

Introductory Case Studies

Project 1:Descriptive Analysis of Demographic Data

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

Author: Opeyemi Ayanwale

Group number: 7

Group members: Minjae Ok, Opeyemi Ayanwale, Mustapha Azeezat Mosunmade, Fyalisia Amanda Putri

November 8, 2021

Contents

1 Introduction

Demography is the study of populations; it entails the long-term tracking of key demographic indicators such as mortality, migration and fertility over time (Max Planck Institute, 2021). Demography is gaining popularity among the general public, as demographic transition has become a topic of political debate in many developed countries (Max Planck Institute, 2021). Prior to now, demographic patterns were relatively stable, populations growth is slow, birth rates and death rates changed slowly. This long period of stability however has changed as most countries have birth rates that are lower than the replacement level of 2.1 children per woman, while life expectancy has increased significantly and continue to rise (Max Planck Institute, 2021; Max Roser, 2014).

This report's primary goal is to investigate demographic data from 228 countries over the last 20 years(2001-2021). This project's specific goals are to describe the distribution of variables of interest, which are total fertility rate and life expectancy, to see if there is a relationship between the variables, to look at variability within subregions and to look at changes over the last 20 years.

For each variable of interest, descriptive statistics and a frequency table were first generated. The frequency distribution of the variables is described using a histogram. To determine whether or not there is a relationship between two variables, a correlation analysis was performed. Finally, data from 2001 and 2021 were compared to see how the variables' values changed over the last 20 years. A boxplot is used to show how variables changes over time.

The second section provides a more detailed overview of the data set, description of all variables, and information on data quality. The statistical analysis methods are presented and explained in the third section. And in the fourth section, presentation, analysis and interpretation of the results are presented. Finally, in the fifth section, the main findings and future research directions are summarized.

2 Problem statement

2.1 Data set and data quality

This report uses the data set of the International Data Base (IDB) of U.S. Census Bureau, which contains various demographic data (currently from 1950 to 2060) on all

states and regions of the world recognized by the US Department of State and with a population of 5000 or more. The database's sources include information from state institutions such as censuses, surveys, and administrative records, as well as estimates and projections from the United States Census Bureau. The data set analysed in this report is a small extract from the IDB for the year 2001-2021. The dataset includes country names, GENC a two-letter code which represent the geopolitical entities, names and codes of a particular country, subregion, region, year, total fertility rate and life expectancy stratified by sex (U.S. Census Bureau , 2021).

Between 2001 and 2021, there are 228 countries, which are divided into 5 regions and 21 subregions. Country, genc, region and subregion are all strings variables, whereas year, total fertility rate, female life expectancy, male life expectancy and life expectancy for both sexes are all numerical variable (U.S. Census Bureau , 2021). The total fertility rate is defined as the average number of children that would be born to a woman over her lifetime if the woman were to experience the current age-specific fertility rates throughout her lifetime. The total fertility rate is generally computed by summing up the age-specific fertility rates defined over a five-year interval. Life expectancy is the average number of years that a group of people born in the same year can be expected to live if mortality at each age remains constant in the future. The most commonly used measure is life expectancy at birth. All variable definitions were obtained from U.S. Census Bureau (U.S. Census Bureau, 2021)

There are 26 missing data 2 of which are from GENC "NA" code for "Namibia" and remaining 24 are from total fertility rate, life expectancy at birth for male, life expectancy at birth for female and life expectancy for both sexes. The 6 rows containing the 24 missing data are removed, while the missing data (NA) from GENC which is an abbreviation for the country "Namibia" is replaced with "NM".

This dataset contains nine variables in total, but only six to seven will be used for this analysis.

2.2 Project objectives

The year 2021 is the focus of this report, with other variables of interest including total fertility rate, male life expectancy, female life expectancy, life expectancy for both sexes, regions and subregions. First, descriptive statistics are generated. A histogram is used to depict the frequency distribution of the variables.

A correlation analysis was carried out to determine whether or not there is a relationship between two variables, namely total fertility rate and life expectancy. Data from 2001 and 2021 were compared to see how the values of the variables changed over the last 20 years. A boxplot and a scatter plot are used to show how data is distributed and how variables have changed over time.

3 Statistical methods

3.1 Histogram

The histogram is a well-known graphing tool. It is used to summarize discrete or continuous data on an interval scale. It is widely used to depict the major elements of data distribution in an easy-to-understand style (Dharmaraja Selvamuthu and Dipayan Das, 2018). Each data point in the histogram is represented by a single bar. The data points are aggregated and presented based on the bin value. The complete range of data values is separated into a series of non-overlapping intervals. (Dharmaraja Selvamuthu and Dipayan Das, 2018).

Mathematical definitions:

$$n = \sum_{i=1}^{k} mi$$

where mi is the number of observations that falls into each of the disjoint categories known as bins, n be the total number of observations and k be the total number of bins.

Frequency, relative frequency, and frequency density can all be used in histograms. The numbers along the vertical axis will change, but the overall shape of the histogram will not. This report makes use of frequency density.

Mathematical definitions:

$$FDi = \frac{F_i}{w_i}$$

where FDi is the frequency density of bin i, Fi (absolute) frequency of bin i and wi bin size of bin i.

3.2 Correlation

Correlation is a statistical term that describes how closely two variables move in sync with one another. When two variables move in the same direction, a positive correlation exists. When they move in opposite directions, a negative correlation exists (Dharmaraja Selvamuthu and Dipayan Das, 2018). The are several correlation coefficient (r) or (p) which ranges from -1.0 to 1.0. expresses the strength of a relationship between two variables. The Pearson correlation coefficient, which is only sensitive to a linear relationship between two variables, is used in this report (Dharmaraja Selvamuthu and Dipayan Das, 2018). The Pearson correlation coefficient mathematically can be defined as:

$$rxy = \frac{\sum(X_i - X)(Y_i - Y)}{\sqrt{\sum(X_i - X)^2}\sqrt{\sum(Y_i - Y)^2}}$$

Where correlation coefficient is rxy between two random variables Xi and Yi are the individual sample points indexed with i. $X\ and\ Y$ are the sample mean.

Positive r values suggest a positive correlation, which occurs when the values of both variables tend to rise together. Negative r values show a negative correlation, which occurs when the values of one variable tend to increase while the values of the other variable tend to drop (Dharmaraja Selvamuthu and Dipayan Das, 2018).

3.3 Scatter plot

A scatter plot is a mathematical graphic used to show the relationship between two variables in a set of data. A scatter plot might imply several types of relationships between variables. Positive (increasing), negative (falling), or null correlations are all possible (uncorrelated). If the pattern of the dots from lower left to higher right shows a positive correlation between the variables under consideration. A negative association is indicated if the dot pattern slopes from higher left to lower right. To investigate the relationship between the variables, a line of best fit (also known as a 'trendline') can be created (Dharmaraja Selvamuthu and Dipayan Das, 2018).

3.4 Box plot

A boxplot is a way for graphically representing groups of numerical data based on their quartiles. A boxplot is made up of two parts: a box and a set of whiskers. The lowest point represents the data set's minimum (Q0 / 0th percentile) which is the lowest data point excluding outliers, while the highest point represents the data set's maximum (Q4 / 100th percentile) the largest data point excluding outliers. The box is drawn from first quartile (Q1 / 25th percentile) the median of the lower half of the dataset to the third quartile (Q3 / 75th percentile) which is the median of the upper half of the dataset, with a horizontal line placed in the middle to represent the median (Q2) which is the middle value of the dataset. The interquartile range (IQR) is the distance between the upper and lower quarter (Q3-Q1) (Dharmaraja Selvamuthu and Dipayan Das, 2018).

4 Statistical analysis

Several statistical methods are presented in this section, which are later used to analyze the data set based on the questions investigated. The statistical software R (R Development Core Team , 2020), version 4.0.3 was used for all analysis and visualizations.

4.1 Frequency distribution of the variables

To visualize the frequency distribution of the variables, a histogram of numeric variables is plotted on the y-axis using frequency density. The histogram of total fertility rate is shown in Figure 1. Total fertility rates are skewed to the right. The histogram shows that the majority of the sample values are clustered on the right side. The data peaks at about 1.5 to 2, indicating that many countries have low total fertility rates, and it declines towards the right side of the graph (see Figure 1). Figure 2 shows a figure with three subfigures. Figure 2(a) depicts a histogram of life expectancy at birth for both sexes, 2(b) depicts a histogram of male life expectancy at birth, and 2(c) depicts female life expectancy at birth. The three histograms are skewed to the left, indicating that most countries have a high life expectancy. Many countries' life expectancy ranges between 70 and 85 years (see Figure 2).

When comparing male and female life expectancy, the chart shows that the tallest bin for males is 75-80years, while the tallest bin for females is 75-85years. The highest life

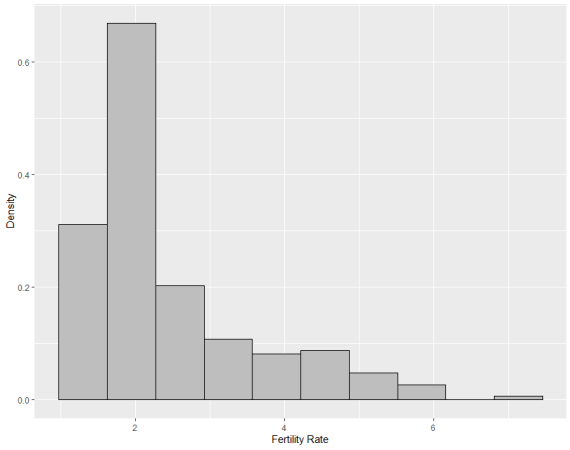
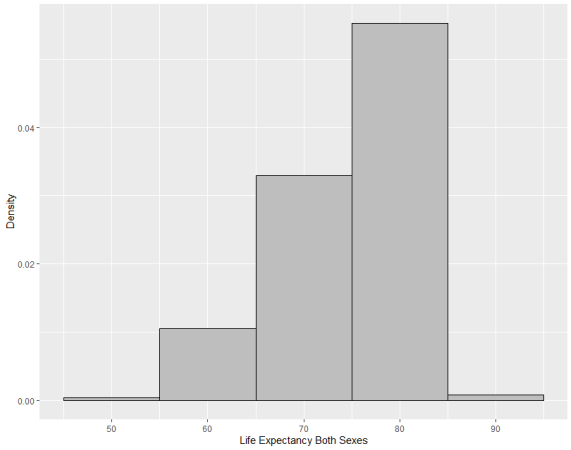
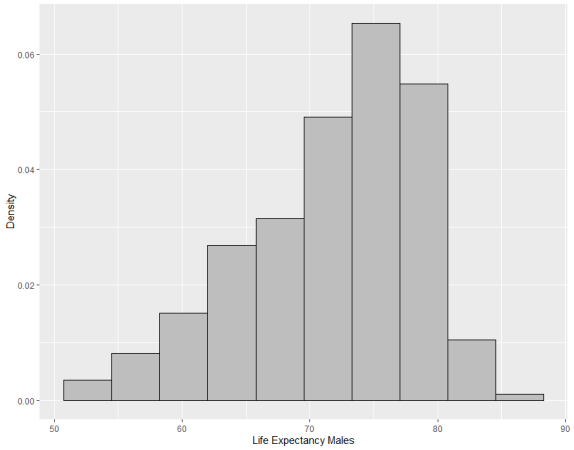


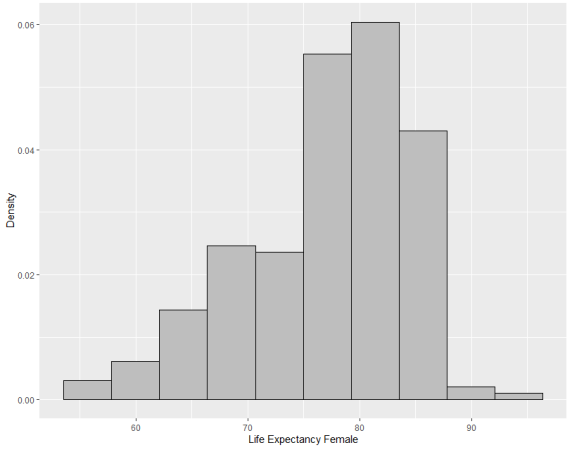
Figure 1: Histogram of total fertility rate for year 2021



(a) Life expectancy at both sexes



(b) Life expectancy male



(c) life expectancy female

Figure 2: Histogram of life expectancy at birth by sex.

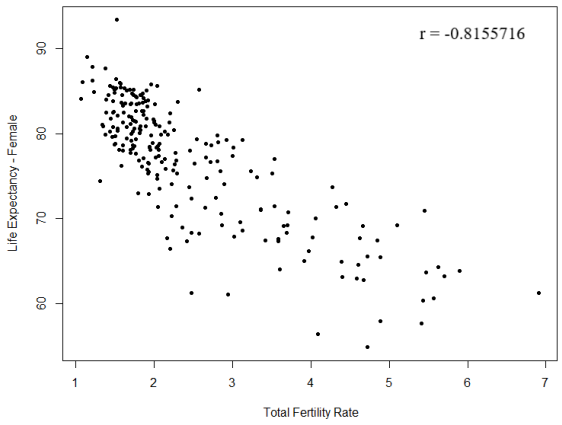
expectancy represented is between 85 and 95 years, with females outnumbering males in this category. Similarly, from the descriptive summary of the data the largest data point for female is 93, for male is 85 and for both sexes it's 89.This means that females have a longer life expectancy than males (see Figure 2).

4.2 Correlation between variables

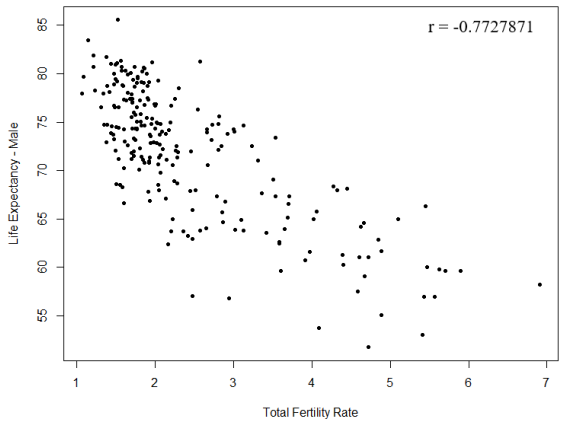
In this section, we use Pearson correlation coefficients and scatter plots to explore the relationships that exist between the variables. Table 1 summarizes the correlation coefficient between total fertility rate and life expectancy at birth based on sex. A scatter plot was created to further investigate and demonstrate the relationship between total fertility rate and life expectancy. Figures 3a, 3b, and 4 depict a scatter plot of total fertility rate versus life expectancy for females, males and both sexes.

Table 1: Correlation coefficient between total fertility rate and life expectancy at birth based on sex.

Life expectancy	Total fertility rate
Life expectancy for female	-0.8155716
Life expectancy for male	-0.7727871
Life expectancy for both sexes	-0.7997004



(a) Life expectancy of female versus Total fertility rate



(b) Life expectancy male versus Total fertility rate

Figure 3: Life expectancy at birth by sex versus Total fertility rate.

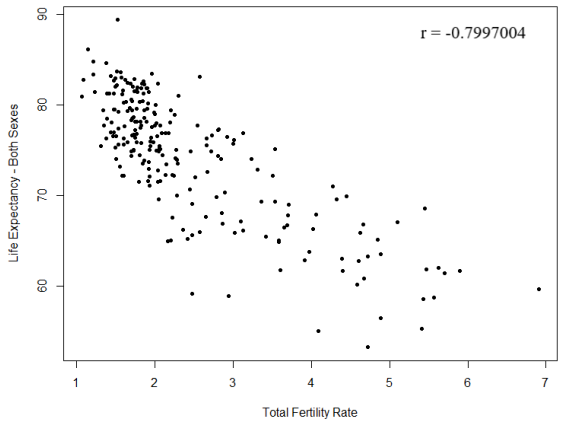


Figure 4: Life expectancy both sexes versus Total fertility rate

The analysis results show that the correlation coefficient (r) values are all negative, with values close to -1 indicating a strong negative association between the variables. The scatter plot shows a negative relationship between total fertility rate and life expectancy by birth for females, men, and both sexes. Life expectancy tends to decrease as the total fertility rate decreases (see Figures 3 and 4).

4.3 Comparing the variability of the values between subregions

Box plots are used to depict variability in several subregions. Figure 5 show boxplots of life expectancy by subregion. Northern Africa and Eastern Asia have large differences between their third quartile (Q3) and the first quartile (Q1) values, the data is skewed to the right, and the medians (Q2) are skewed to the right, indicating large differences between the data for the subregion, whereas Western Europe, Northern Europe, Northern America and Australia have small differences between their Q3 and Q1 values, the data is not skewed, indicating very low variability. Life expectancy at birth in subregions is often similar to that of other subregions within the same region, but not to that of subregions within different regions (see Figure 5). Figure 6 show boxplots of total fertility by subregion. The vast majority of the countries with the highest fertility rates (with the highest data point Q4) are in Africa, from the result of this analysis, Western Africa, Middle Africa, Northern Africa and Easterner Africa have the highest fertility rate while Austrialia/New Zealand, Eastern Europe and Southern Europe have the lowest fertility rate (see Figure 6). Multiple factors, including planned family size,

low levels of usage of modern contraception, and high levels of adolescent pregnancy, may contribute to Africa's high fertility rate (World Population Review , 2021).

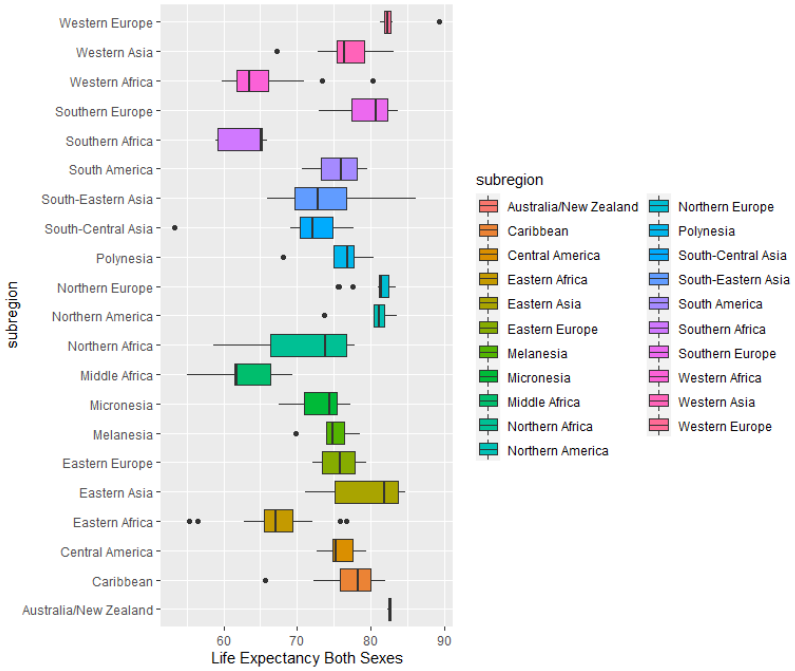


Figure 5: Boxplots Life Expectancy by Subregion

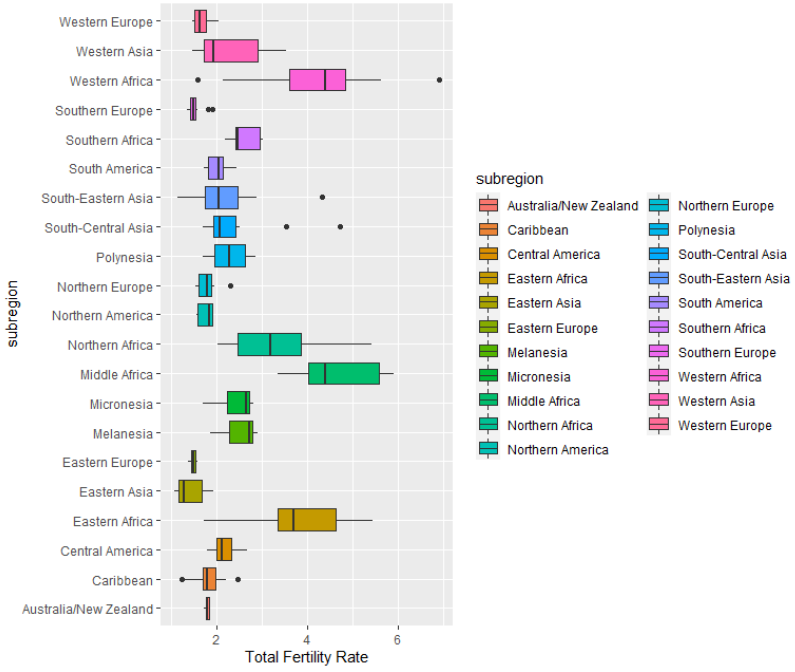
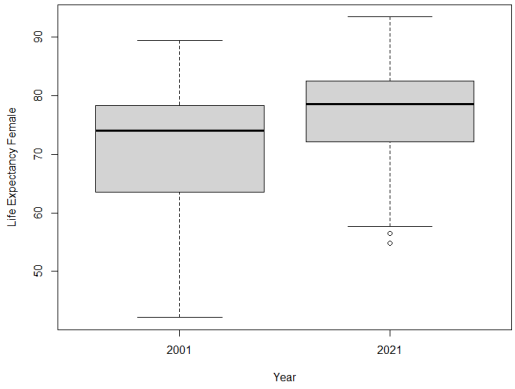


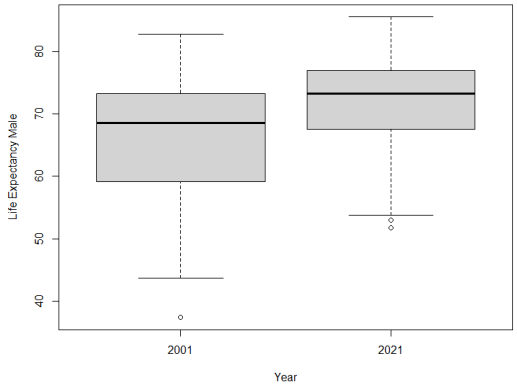
Figure 6: Boxplots Fertility Rate by Subregion

4.4 Comparing the variability of the values between 2001 and 2021

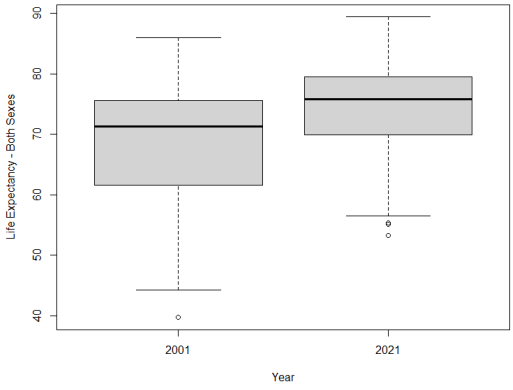
A boxplot is used to depict life expectancy by year and fertility rate by year to see how the values of the variables have evolved over the last 20 years. Figure 7 show the boxplots of life expectancy by sex between the year 2001 and 2021. The boxplots look similar regardless of the sex, and life expectancy has increased for both sexes over the last 20 years, with less fluctuation in the data. The median life expectancy (Q2) for male in 2001 is slightly below 70 and in 2021 the median increases sightly above 70 (see Figure 7b). And the median life expectancy for female(Q2) in 2001 is above 70 but below 80 and in 2021 the median life expectancy is around 80 (see Figure 7a). In 2021, life expectancy for women is still higher than the life expectancy for men (see Figure 7). The world is progressing toward a higher life expectancy at birth.



(a) Life expectancy of female by year



(b) Life expectancy of male by year



(c) Life expectancy of both sexes by year

Figure 7: Life expectancy at birth by sex for year 2001 and 2021.

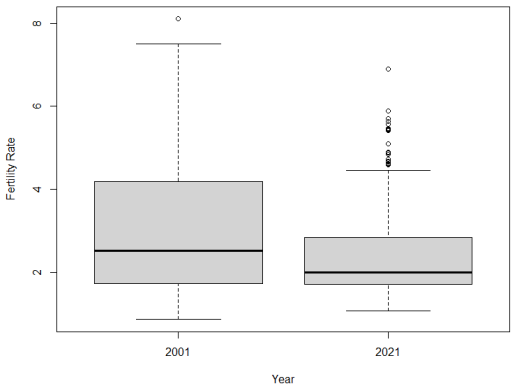


Figure 8: Boxplot of total fertility rate for year 2001 and 2021

Figure 8 depicts boxplots of total fertility rates between 2001 and 2021. The boxplot for 2021 shows that the general fertility rate has fallen and that data variability is lower than it was in 2001. Nonetheless, there are several outliers in the data, suggesting that while many countries' total fertility rates have declined, some still have extremely high total fertility rates (see Figure 8).

5 Summary

This project uses the data set of the United States Census Bureau's International Data Base (IDB), which provides diverse demographic data (currently from 1950 to 2060) on all states and regions of the world recognized by the United States Department of State and having a population of 5000 or more. The database's sources include data from state organizations including censuses, surveys, and administrative records, as well as estimates and projections from the US Census Bureau. This report's data set is a short excerpt from the IDB for the years 2001-2021. It comprises 228 nations' life expectancy and fertility rates from 2001 and 2021. Geographically, the countries are organized into 5 regions and 21 subregions. The purpose of this report was to conduct explanatory data analysis on each variable in order to understand the distribution of the variables, the relationships that exist between the variables, how similar the subregions are and how the data had changed from 2001 to 2021.

The frequency distribution of the variables is depicted using a histogram. A boxplot and a scatter plot are used to demonstrate how data is distributed and how variables change

over time. Using a histogram, many countries have low overall fertility rates, which decrease towards the right side of the graph. The findings also suggest that the majority of countries have a high life expectancy. Life expectancy in many countries is between 70 and 85 years, and when male and female life expectancy is compared, females have a higher life expectancy than males. To further analyze and highlight the relationship between total fertility rate and life expectancy, a scatter plot was developed. The scatter plot demonstrates a negative relationship between total fertility rate and life expectancy by birth for women, men, and both sexes. As the total fertility rate falls, so does life expectancy.

The variability by subregion was visualized using a box plot. Subregional life expectancy at birth is frequently comparable to that of other subregions within the same region, but not to that of subregions within different regions. Africa has the vast majority of the countries with the highest fertility rates. Over the last 20 years, both sexes' life expectancy has improved of which Women's life expectancy remains higher than men's life expectancy in 2021. The overall fertility rate has decreased and data variability is lower now than it was in 2001. As a result, today's women have half the number of children they did 20 years ago. This transition has happened as a result of increased contraception access and reliability, a major reduction in infant and child mortality rates and increased educational and occupational training for women (Statista , 2021).

The results of this analysis cannot be generalized because we only used a subset of the population data. A larger sample size is necessary to improve the generalizability of the findings. Further research into the significance of the Correlation Coefficient could add to our understanding of the outcomes of this analysis.

Bibliography

Dharmaraja Selvamuthu and Dipayan Das (2018). *Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control*. Springer.

Max Planck Institute (2021). What is demography. https://www.demogr.mpg.de/en/about$_u s_6$113/$what_i s_d emography_6$674/($visitedon22ndOctober$2021).

Max Roser (2014). Fertility rate. *Our World in Data*. https://ourworldindata.org/fertility-rate.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Statista (2021). World and continental fertility rate. https://www.statista.com/statistics/1034075/ fertility-rate-world-continents-1950-2020/ (visited on 27nd October 2021).

U.S. Census Bureau (2021). International data base (idb) demograph data. https://www.census.gov/programs-surveys/international-programs/about/idb.html (visited on 22nd October 2021).

U.S. Census Bureau (2021). Glossary. https://www.census.gov/programs-surveys/international-programs/about/glossary.html (visited on 22nd October 2021).

World Population Review (2021). Total fertility. https://worldpopulationreview.com/country-rankings/total-fertility-rate (visited on 27nd October 2021).

Appendix

A Additional figures

B Additional tables