TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Franziska Kappenberg

M. Sc. Marieke Stolte

Author: Raj Anilbhai Pawar

Group number: 17

Group members: Amritha Sukhdev Singh Agarwal, Sagar Basnet, Muhammad Fahad, Siddhartha Karki

November 11, 2022

Contents

1 Introduction

The pandemic has shown more than ever the importance of collecting demographic information to help us better draw conclusions about population. Estimates and projections of world population over the last few decades show significant progress toward some of the Sustainable Development Goals (Census, 2020). Demographic information allows us to better understand certain background characteristics of an audience. Governments and organizations could create strategies for legislation and social services based on the intended audience using these data. The United States Census Bureau has published varied demographic data from the year 1950 to the year 2021, as well as forecasts till the year 2060 and 227 countries are represented in the collected data.

The primary objective of this report is to apply descriptive statistics techniques to a sample of data from International Database (IDB, 2021). It includes birth and overall infant mortality rates for 227 nations between 2001 and 2021 and also to see how the values of the variables have changed over the last 20 years. By using statistical approaches, an overview of each variable's characteristics and relationships between them has been shown. To examine the relationship between the variables, the distribution of the variables and their correlation were performed. A suitable graphical form has been used to compare the mortality rates of infants and the life expectancy rates of males and females.

In the remaining section of the reports apart from the introduction part there is second section which consists of the problem statement, it describes the quality of the data and its source from where the data has been taken also the objective has been discussed. In the third section of the report, descriptive statistical methods like mean, variance, standard deviation, five-number summary, correlation, and pearson correlation have been described. Visualization techniques have been used to study the distribution of the variables using histograms, scatter plots, and box plots. In the fourth section of this report, with graphs the interpretations has been discusses after using statistical techniques on the provided dataset. The results that have been achieved with the major discoveries from the demographic data are summarized in the report's concluding part.

2 Problem statement

2.1 Source and Quality of Data

Every state and territory in the globe that is recognized by the US Department of State and has a population of 5000 or more is included in the International Base (IDB, 2021) of the US Census Bureau currently from 1950 to 2100. Data from state agencies, such as censuses, surveys, or administrative records, as well as projections and estimates from the U.S. Census Bureau, are among the sources used in the provided dataset.

This report utilises a dataset from file "census2001_2021.cs" an extract from the International Database (IDB, 2021), which contains data on infant mortality rates and gender specific life expectancy rates. There are 227 distinct countries, 21 subregions, and a total of 5 regions in this dataset. The column names from the data file have been modified for simplicity and to improve readability.

- Infant_Mortality_Rate_Both_Sexes: It is the number of very young children dying before becoming one year old, per 1,000 live births in a particular year (U.S. Census Bureau, 2021). This variable does not contains the values for the countries Libya, Puerto Rico, South Sudan, Sudan, Syria and United States in year 2001.
- Life_Expectancy_Both_Sexes: This variable represents the typical age of males and females. This variable is a statistical measure which is calculated based on birth year, present age and demographic factors parameters like sex (U.S. Census Bureau, 2021).
- Life_Expectancy_Males: This variable represents the typical age of males who were born in the same year (U.S. Census Bureau, 2021). Its unit is represented in years.
- Life_Expectancy_Females: This variable represents the typical age of females who were born in the same year (U.S. Census Bureau, 2021).
- Region: The region variable shows each record's continent. There are in total 5 regions given in the data file named Asia, Europe, Africa, Oceania, and Americas. This categorical variable is filled in for each data record since the regions are only labels without any arbitrary order.
- Subregion: Every region is then further divided into subregions there are in total of 21 unique subregions. So the subregion variable is also a categorical variable as it is part of the region.

- Country.Name: Every subregion is then further divided into countries. There are in total of 227 unique countries. So this variable will also be considered as a categorical variable.
- Year: The variable year consists of only either of 2 values 2001 or 2021. This variable form can be considered as binary variable. It shows that which year does the data is belonging to.

When comparing the data for the years 2001 and 2021, the six countries with null values in the field Infant_Mortality_Rate_Both_Sexes for the year 2001 will be excluded. There are no null values for the mortality rates or life expectancy rates in the data for year 2021 therefore the data for analysis of year 2021 is of good quality.

2.2 Objectives

In this report we have discussed four objectives. The first objective is to show the frequency counts to interpret the information more easily. Each numerical variable's frequency distribution is examined in order to accomplish this. Also, the comparison has been made for the life expectancy at birth between males and females. The second objective is to check the bivariate correlations between the numerical variables. It is achieved by determining the relationship between the numerical variables weather they are dependent on each other or not. Also the monotonous relationships between numerical variables is checked and described weather they are linear or not. The third objective is to see whether the values of the individual variables are comparably heterogeneous between different subregions and homogeneous same within each individual subregion. The first, second and third objective in this report has exclusively used the data for year 2021. The fourth objective of the report is to compare the year 2021 and 2001 data. The comparison has been done to check how the infant mortality rate and life expectancies have been changed for the year 2021 and 2001.

3 Statistical methods

Statistical analytic techniques with their mathematical justifications will be covered in this part along with ways for data visualization.

3.1 Descriptive statistics

3.1.1 Mean

A finite collection of numbers, mean serves as a measure of central tendency. The arithmetic mean, which is calculated by adding up all of the numbers in a group and dividing by the total number of numbers, is the average of that group. Greek letter μ is used to denote it, and formula for its calculation is given by,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_N = \frac{1}{N}\left(x_1 + \cdots + x_N\right)$$

Here, the capital Greek letter $\sum$ denotes summation of all the numbers in data. $x_1, ..., x_N$ denotes the numbers present in the data and N denotes the count of such numbers. (Black, 2009)

3.1.2 Variance and Standard deviation

The squared variations about the arithmetic mean for a group of numbers are averaged to form the variance. It serves to demonstrate how widely the data is dispersed. The variance increases as the data deviates further from the mean. The variance is indicated by σ and its mathematical formula is,

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

Here, $x_1, ..., x_N$ are the samples of variable X, μ represents the mean of dataset, N represents the count of values in the dataset (Black, 2009)

The standard deviation is indicated by the symbol σ, which is the square root of variance. It is employed to determine how far the observations deviate from the mean and to summarize the degree of data dispersion. The standard deviation is expressed in the same units as the raw data, unlike the variance, which is written in those units squared. This is one element that sets the standard deviation apart from the variance. The data's unit of measurement is the same for the standard deviation. Standard Deviation mathematical formula is $\sigma = \sqrt{\sigma^2}$

3.1.3 Five-Number summary

The minimum value, first quartile, maximum value, median, third quartile, and maximum value are the five numbers that make up the five-number summary. If there is a divergence from symmetry or the presence of outliers, it can be shown using a five-number summary. The value known as a quartile is used to partition data into four groups. A value that falls between the minimum value and median and in which at least 25% of values are less than or equal to it is referred to as the first quartile. The third quartile is a value that falls between the median and maximum value and at least 75% of the values are less than or equal to it. The first quartile is subtracted from the third quartile to obtain the interquartile range (IQR), which contains 50% of the observations. In order for at least half of the values to be less than or equal to the median and for the other half to be greater than or equal to the median, the values of the observations must be divided into two equal portions. According to the number of samples, the median is determined. If the number is odd, the median is the middle 4 value, and if it is even, the number is the mean of the two middle values. The observations that are 1.5∗IQR or more units below or above the first quartile are considered outliers. (Hay-Jahans, 2017)

3.1.4 Correlation

The correlation coefficient measures how closely connected two variables are. If the values of the two variables fluctuate in opposite directions, that is if one variable increases or decreases while the other lowers or increases. This is referred to as a negative correlation between the two variables. When one variable's value increases and the other variable's value decreases, or when the two variables fluctuate in the same direction. The two variables are said to be positively linked. The coefficient of correlation, a number that can range from -1 to +1 is used to quantify the relationship between two variables. (Black, 2009)

3.1.5 Pearson correlation

The degree and direction of the linear link between continuous variables are assessed using the Pearson correlation coefficient. The i^{th} sample for the two variables X and Y, respectively, will be denoted by x_i and y_i. Let $\bar{x}$ and $\bar{y}$ represent the means of X and Y. The variances of the variables X and Y are, respectively, σ_X^2 and σ_Y^2, while σ_{XY} is their covariance. The mathematical formula is,

$$r = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

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

The correlation coefficient has a value range of [-1, 1], regardless of the units used to measure the two variables. Additionally, if the absolute value of $|r|$ is close to 1, a strong linear relationship between these two variables can be deduced. Otherwise, if this value is close to or equal to 0 means that there is no linear relationship between variables, but it does not indicate that these two variables are independent as their relationship could be non-linear. (Hay-Jahans, 2017)

3.2 Visualization techniques

3.2.1 Histogram

A distribution of numerical data is represented by a histogram. The idea behind it is to divide the data into distinct intervals, count how many values are in each interval, and then represent each group as a segmented column. The area of each bar in a histogram, or the product of its width and height, roughly represents the relative frequency of it's category. Each bar's width might or might not be the same size. The heights of bars correspond to the number of data points in each group when all bars are the same width. There are numerous methods for selecting the width size, and many of them advise selecting k groups after which the width size can be calculated as $(max - min)$ / k. In this report, the Freedman-Diaconis formula has been used to obtain the value of k.

$$k = \frac{[max(x) - min(x)]\sqrt[3]{n}}{2IQR}$$

where sample size is n, and IQR stands for interquartile range. (Hay-Jahans, 2017)

3.2.2 Scatter plot

Dot's are used in a scatter plot to show the values of two different numerical variables. Each dot's location on the horizontal and vertical axes represents that data point's values. By watching how the value of one variable varies with the value of the other variable, we may compare the relationship between the two variables. We can claim that there is a linear relationship between the two variables when the data points on the scatter plot are in a straight line. We say two variables are positively linked if an increase in one variable results in a rise in the other. We say two variables are negatively linked if an increase in one variable results in a decrease in the other. Additionally, we may use a matrix scatter plot to simultaneously display scatter plots of a multi-variate dataset if we have more than two continuous variables. (Hay-Jahans, 2017)

3.2.3 Boxplot

A boxplot is a visual representation of the location, dispersion, and skewness groups of quartiles of numerical data. A box and a whisker line serve as symbols for the boxplot. The box shows the median, which is shown by the middle line, as well as the interquartile range (IQR), which is the difference between the first and third quartiles, which are represented by the two hinges. In order to show the variability of data outside the box, whiskers are two lines that extend from the lower quartile to the minimum value and the upper quartile to the maximum value. Right-skewed data are indicated by a longer right-whisker or by the presence of a median to the left of the box's center. Left-skewed data are shown by a longer left-whisker or a median that trends to the right of the box's center. It is termed symmetric data if both whiskers are roughly equal and the median is roughly in the middle of the box. (Hay-Jahans, 2017)

3.3 Software

The Python programming language (Python Software Foundation, 2020), version 3.9.0 was used for analysis and compilation. Visual Studio Code Text Editor (Microsoft Corporation, 2022) was used for editing the python code files. All the used packages are described in the code file.

4 Statistical analysis

Here we have used L.E. and IMR as an abbreviation for Life Expectancy and Infant Mortality Rate respectively.

4.1 Descriptive Analysis

Here the distribution feature of the four variables will be analysed. The variables are named as Infant Mortality Rate, Life Expectancy at birth of Males, Life Expectancy at birth of Females and Life Expectancy at birth of Both Sexes. Here the data has been given for 227 unique countries. This analysis is done for the year 2021.

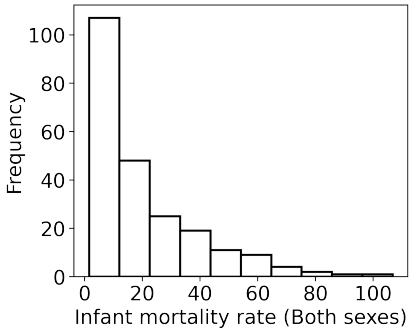


Figure 1: IMR Both Sexes

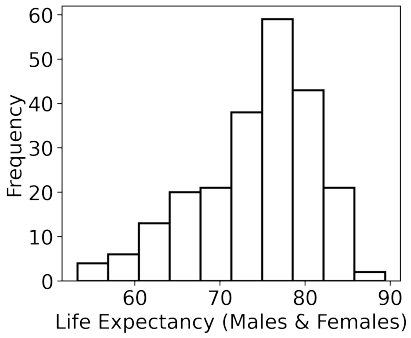


Figure 2: L.E. Both Sexes

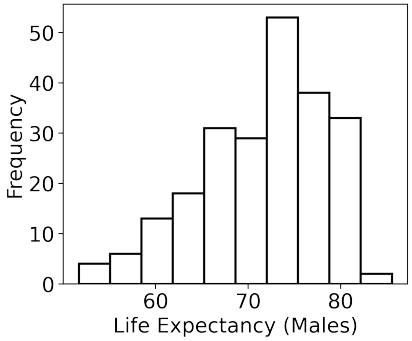


Figure 3: L.E. Females

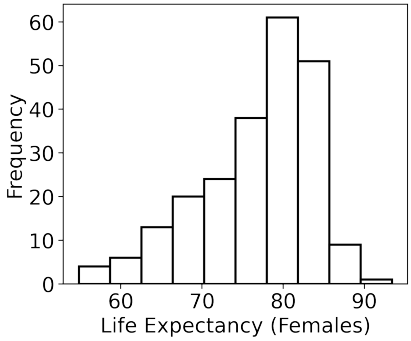


Figure 4: L.E. Males

We may deduce that the distribution is right skewed from the infant mortality rate histogram in Figure 1. The bin with a width of 20 is represented on the x-axis, while the frequency is shown on the y-axis. Most countries under observation have infant mortality

rates between 2 and 12. The infant mortality rate mean is 20.25. These findings suggest that most countries have infant mortality rates that are lower than the global average. Only a small number of nations have death rates higher than 20.25.

Additionally, the data in each of the other three histograms for life expectancies at birth of both sexes, males and females in Figures 2, 3, and 4 respectively is left-skewed. Here, the y-axis denotes the frequency, and the x-axis indicates a bin with a width of 10 years. The average life expectancy at birth for both sexes is between 65 and 82 years with a global average of 74 years. There are only a few countries where the life expectancy is less than 60 years. From figure 3 we can see males have life expectancy at birth between 73 and 75 years. Males have a 71.78 year average life expectancy. Nevertheless from figure 4 we see life expectancy of women is on average of 78 to 82 years so we deduce the majority of nations have life expectancies for females that are higher than the global average. At birth, women may expect to live an average of 76.89 years. The average difference in life expectancy between men and women is 5 years, according to the data.

We have now taken the difference in life expectancy at birth between the sexes and plotted it in Figure 9 of the appendix in order to compare the frequency distribution between the sexes. The difference is distributed in a symmetrical and positive way. Additionally, women have higher mean life expectancies at birth than men do. Therefore, we can draw the conclusion that women generally live longer than men.

4.2 Analysis of Correlation

We examine any possible dependence structures between the relevant continuous variables for the year 2021 in this subsection. We examine whether or not there are bivariate correlations between the variables. The scatter plot and the calculation of the Pearson correlation coefficients are used to achieve this. Figure 8 in the appendix shows scatter plot matrix associations between the four variables. Additionally, Tables 1 exhibit the Pearson correlation coefficients.

Table 1: Pearson Correlation coefficient between four continuous variable

	L.E. Both Sexes	L.E. Males	L.E. Females	IMR Both Sexes
L.E. Both Sexes	1	0.99	0.99	-0.90
L.E. Males	0.99	1	0.97	-0.88
L.E. Females	0.99	0.97	1	-0.91
IMR Both Sexes	-0.90	-0.88	-0.91	1

The bivariate correlations between the variables, which show how strongly or weakly the variables are associated, are similarly described by these coefficients. The scatter plot matrix make it easy to see the linear relationships.

In appendix the Figure 8 scatter plot matrix demonstrates how the infant mortality rate increased when the life expectancy at birth of both sexes declined. Mortality rate has a stronger linear and monotonic relationship with the life expectancy at birth of female than with the life expectancy at birth of male which can be seen from the Pearson coefficient whose magnitude is close to 1.

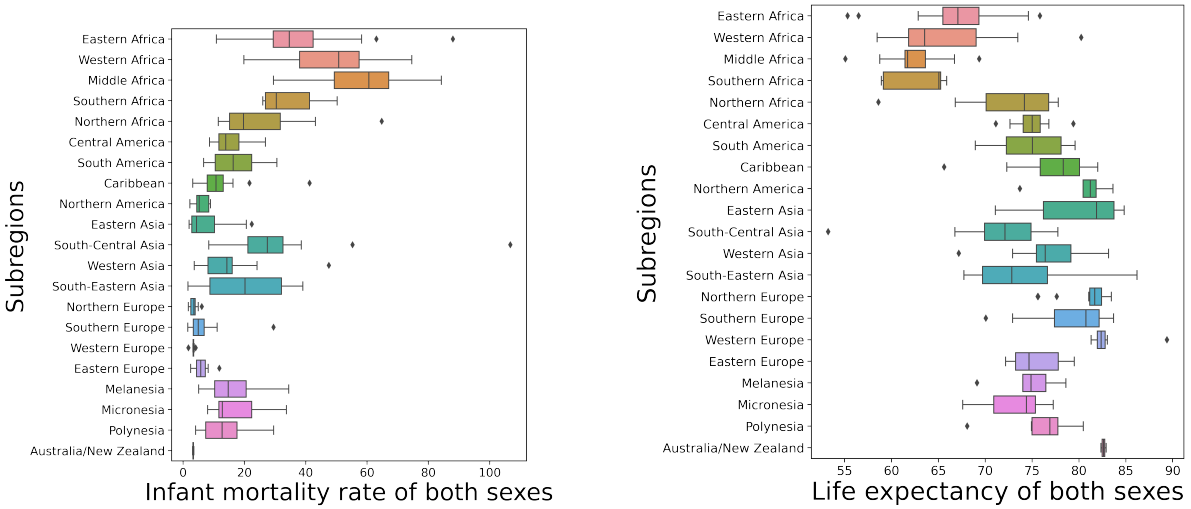
With a very significant correlation of 0.99, the life expectancy at birth of both sexes with the life expectancy at birth of female and male exhibited the strongest linear and monotonic connections among the four variables. These findings indicate that there is no non-linear relationship between the variables and there is present a strong linear and monotonic relationship.

4.3 Variable's value comparison

Here, a boxplot illustrates the variation in mortality rate and life expectancy for each subregion for the year 2021. The interquartile range of the box plot can be used to verify the variability. The associated subregion can be assumed to be homogenous if the length between the quartile 1 and quartile 3 is less from the remaining subregions. First, we examine the variability of each subregion, and then we examine the variability between various subregions, to determine whether the variable is homogenous both within and across the subregion.

Figure 5(a) demonstrates that the infant mortality rate is comparatively equal throughout all Europe and Australia/New Zealand subregions. This indicates that the mortality rates of the countries in a certain subregion are nearly the same. The mortality rates in South-Eastern Asia and all of Africa's subregions vary greatly between their respective nations. As a result, mortality rates around the world vary considerably between different subregions.

Figure 5(b) displays the life expectancy at birth for both sexes. Western Europe, Northern Europe, and Australia/New Zealand are the subregions that have remarkably similar values. These subregions also have the greatest average life expectancy, which ranges from 80 to 85 years. Overall, there are significant differences in life expectancy amongst subregions, ranging from 5 to 20 years. While essentially every region has a life ex-



(a) Boxplot (Mortality rates between subregions)

(b) Boxplot (Life expectancy at birth of both sexes between subregion)

Figure 5: Box Plot for Infant Mortality rate both sexes and Life Expectancy both sexes

pectancy at birth of at least 70 years, only a few exceptional African countries have a life expectancy above 70 years.

Figure 10 in the appendix section also shows the box plot for life expectancy at birth for male and life expectancy at birth for female for the year 2021.

4.4 Comparison over last 20 years

In this part, we'll examine how the variables' values have changed between the years 2001 and 2021. A scatter plot has been utilized to show the difference. The values for the countries in the years 2001 and 2021 are shown by the x-axis and the y-axis, respectively. Here in scatter plots each and every data point represents a data point and the graphs also contains an identity line.

Figure 6 displays the majority of the nations close to the identity line, showing changes in infant mortality rates for most of the nations. Afghanistan is an anomaly among the Asia region since it has the highest infant mortality rate. In comparison to other countries in the Africa region, Seychelles has the lowest infant mortality rate and has not changed considerably.

Comparing all regions, the Europe region has the lowest infant mortality rate. Kosovo is a country in the Europe region that exhibits unusual trends in comparison to other

European region countries and has a high infant mortality rate. The average infant mortality rate declined from 34.98 in year 2001 to 20.13 in year 2021, showing a decreasing trend over time.

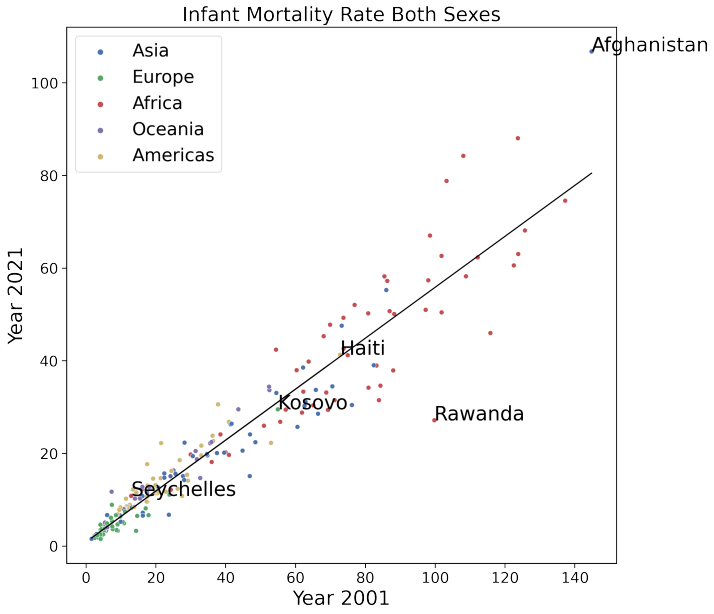


Figure 6: Boxplot (Life expectancy at birth of both sexes between subregion)

In figure 7 the life expectancy of both sexes is shown and in figure 11 the life expectancy of males and females is shown respectively. For the life expectancy of both sexes Monaco is the country which has the highest life expectancy rate. Unlike other European countries it shows some irregular trends which makes it as an outlier. Afghanistan is the only country in the Asia region that shows the lowest life expectancy rate, hence it is also considered as an outlier among the Asia region.

The countries in Europe region shows the highest life expectancy rates among all other regions. The mean value for life expectancy for both sexes has increased from 68.53 in year 2001 to 74.31 in year 2021. For life expectancy of females, the mean has increased from 70.97 in year 2001 to 76.93 in year 2021. Male life expectancy on the other hand increased from 66.23 in year 2001 to 71.81 in year 2021. For females the life expectancy is approximately 5 years greater than males.

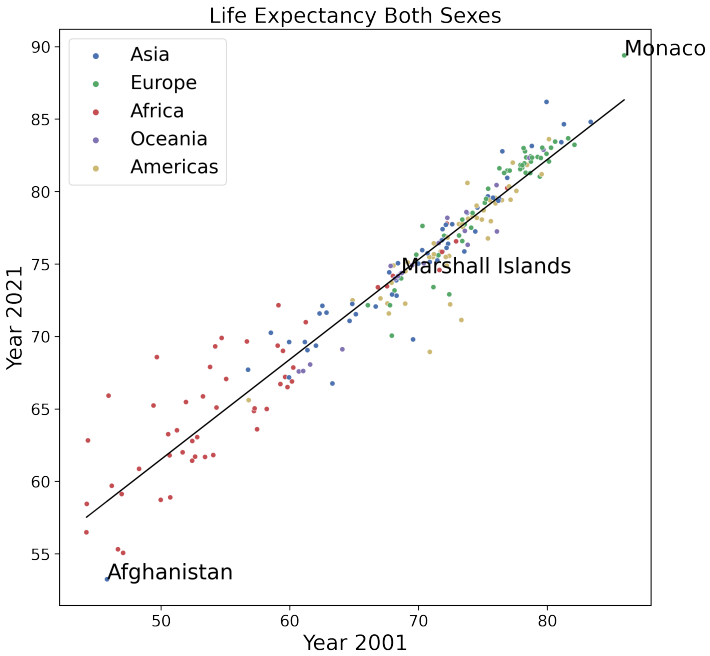


Figure 7: Boxplot (Life expectancy at birth of both sexes between subregion)

5 Summary

The International Data Base (IDB, 2021) served as the source of the dataset examined in this report. The data file includes observations for variables including mortality rate, life expectancy for both sexes, life expectancy for males and life expectancy for females from 227 countries distributed throughout 5 regions and 21 subregions. In order to understand how life expectancy and mortality rates have changed among regions between the years 2001 and 2021, one objective of the report was to study data variation. We used descriptive statistical techniques including the histogram, correlation, boxplot, and scatter plot for this goal. The necessary data was displayed and examined.

In order to compare the life expectancies of males and females, we approximated the frequency distribution of the continuous variables during our analysis. By calculating the mean result, we have discovered that females typically live for five more years than males do. Further study discovered that the mortality rate has a greater linear and monotonic relationship with female birth life expectancy than it does with male birth life expectancy. In terms of overall infant mortality rates, we found that the African region has the lowest rates while the European region has the highest, and vice versa. Subregions of America, Europe, and Oceania had homogeneous distributions of life expectancy variables, whereas subregions of Africa and Asia had heterogeneous distributions.

We compared the values of each variable in 2021 to their respective values in 2001 in order to determine how the trend changed between those two years. All of the variables had a positive slope and a linear relationship. With a few exceptions, life expectancies for both men and women have risen in the past 20 years.

To better understand the mortality rates and life expectancies of the various regions, it could be interesting to include other variables in future research, such as per capita GDP, homeownership, or employment rates of the countries.

Bibliography

Ken Black. *Business Statistics: For Contemporary Decision Making, Sixth Edition*. John Wiley & Sons, Inc., 2009. ISBN 9780470409015.

Census. Sustainable development goals and the 2020 round of censuses. https://www.census.gov/content/dam/Census/library/working-papers/2018/demo/sdg-2020.pdf, 2020. Accessed: 2022-11-06.

Hay-Jahans. *An R Companion to Linear Statistical Models*. CRC Press LLC, 2017, 2017. ISBN 9781138116030.

IDB. International database. https://www.census.gov/data-tools/demo/idb/#/country?COUNTRY_YEAR=2021&COUNTRY_YR_ANIM=2021, 2021. Accessed: 2022-11-06.

Microsoft Corporation. Visual studio code, 2022. URL https://code.visualstudio.com/.

Python Software Foundation. Python, 2020. URL https://www.python.org/.

U.S. Census Bureau. U.s. census bureau. glossary. https://www.census.gov/programs-surveys/international-programs/about/idb.html, 2021. Accessed: 2022-11-06.

Appendix

A Additional figures

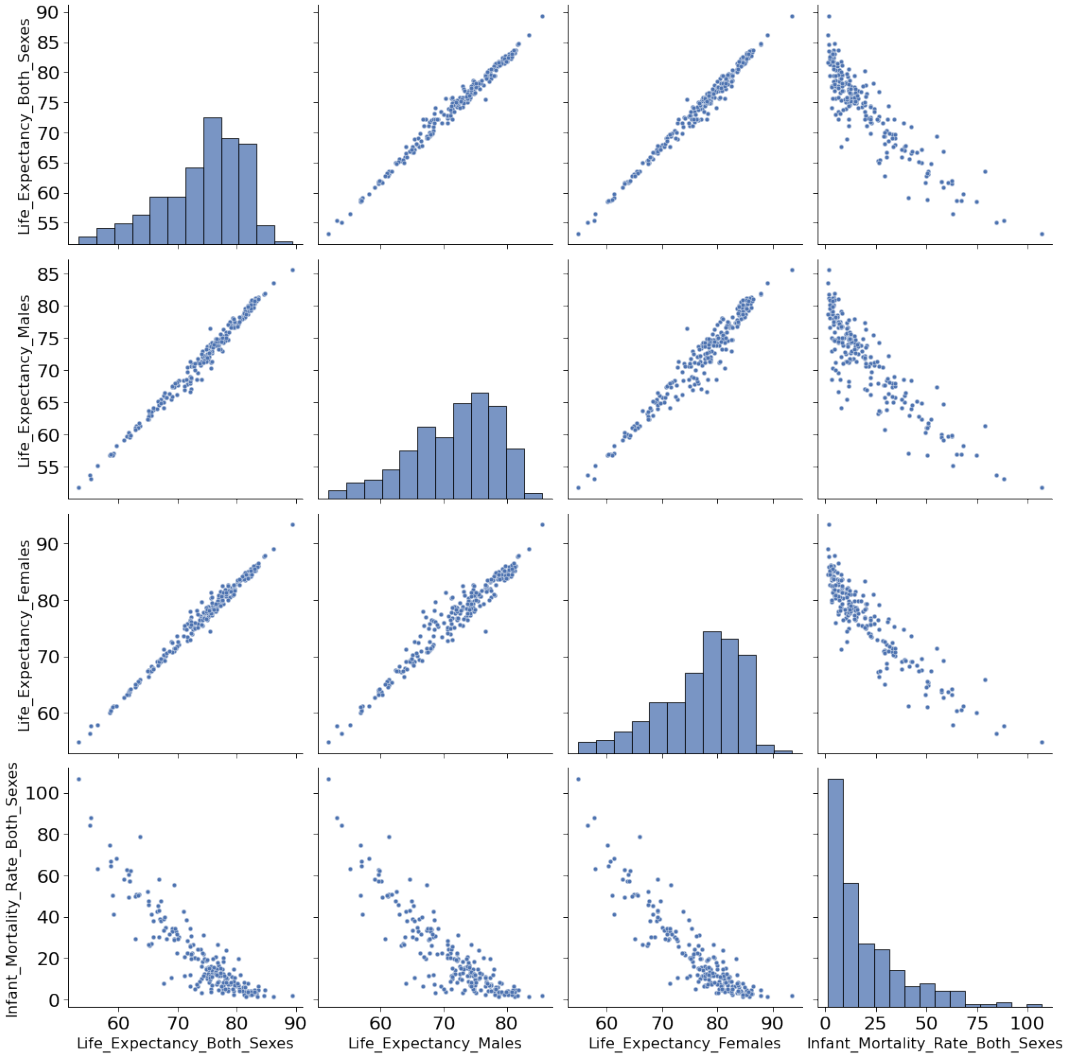


Figure 8: Scatter plot matrix for four continous variables

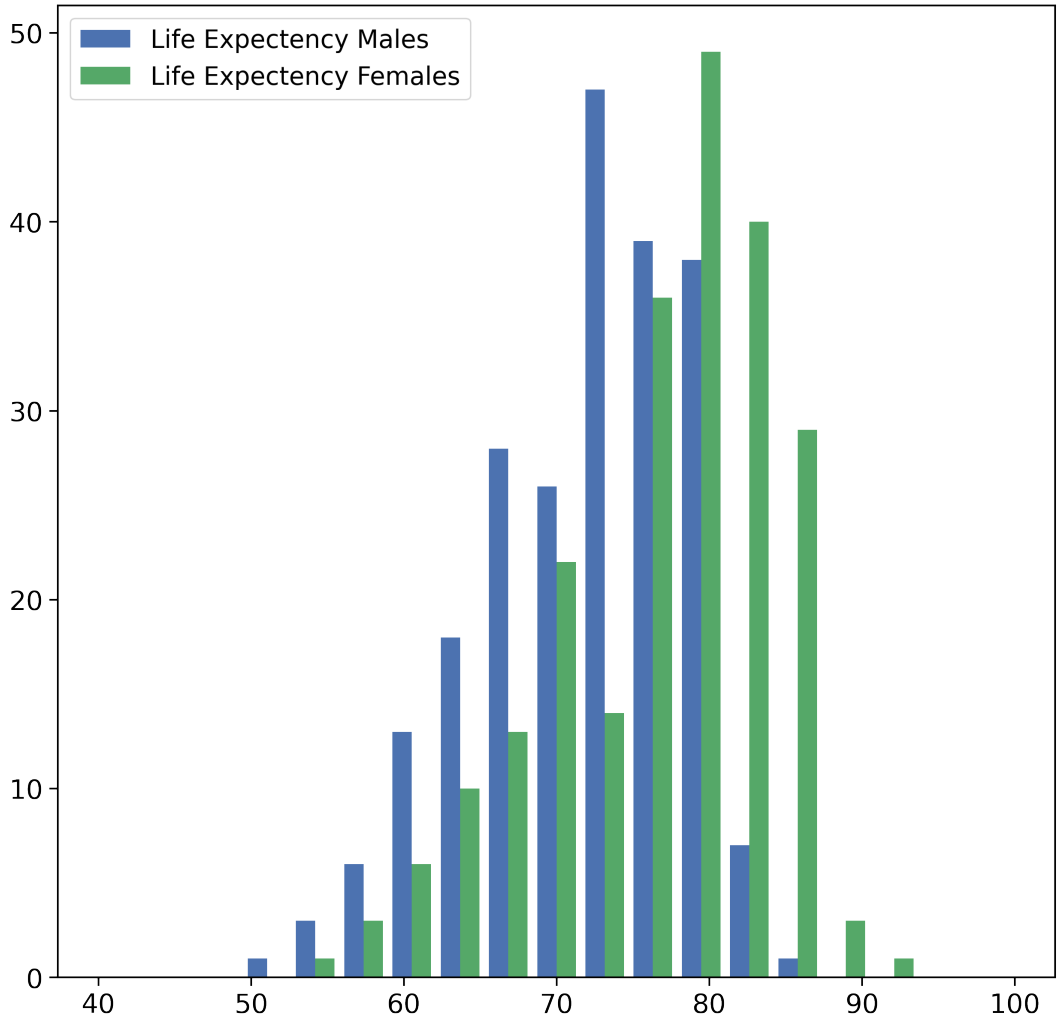
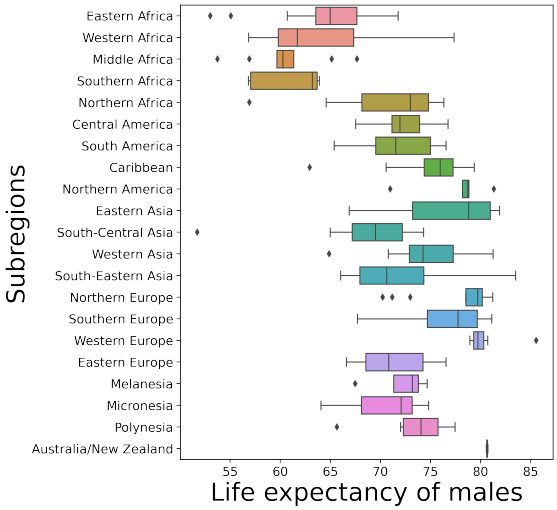


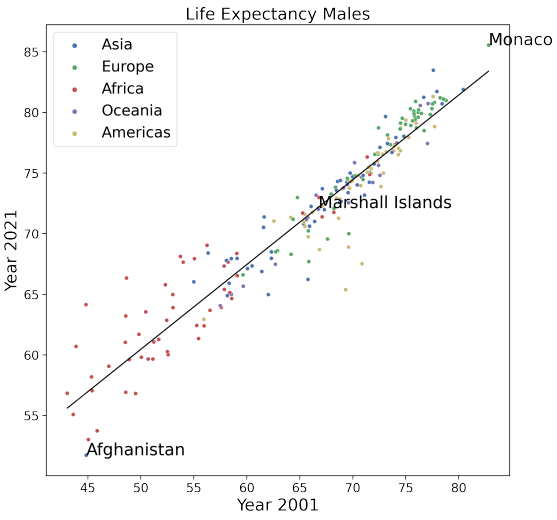
Figure 9: Histogram for comparison of life expectancy of males and females in year 2021



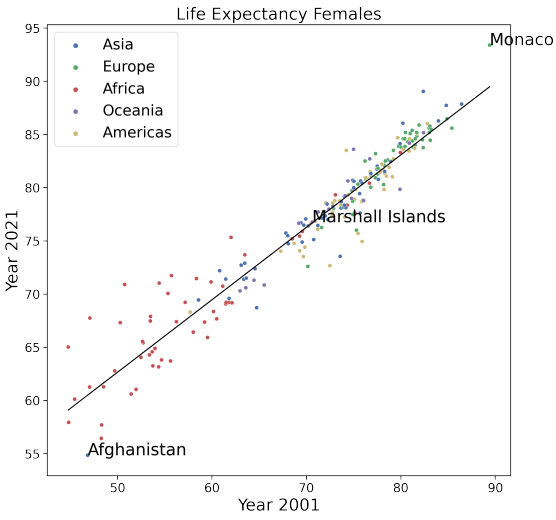
(a) Life expectancy at birth for males

(b) Life expectancy at birth for females

Figure 10: Box plot for life expectancy at birth for males and females



(a) Life expectancy at birth for males



(b) Life expectancy at birth for females

Figure 11: Scatter plot for life expectancy at birth of males and females