TU Dortmund

Introductory Case Studies

Project I: Descriptive Analysis Of Demographic Data

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Franziska Kappenberg

M. Sc. Marieke Stolte

Author: Bushra Tariq Kiyani

Group number: 5

Group members: Prerana Rajeev Chandratre, Sathish Ravindranth, Shivam Shukla, Janani Veeraraghavan

November 11, 2022

Contents

1 Introduction

Demographic data analysis is the study of different population groups based on gender, age, interests, level of education, birth rate, death rate, race etc. Demographic data refers to the collection of such information about specific groups of people. Governments, public and nonpublic institutions use demographic information to learn about the characteristics of a population for policy development. It also helps businesses to understand customer needs and future demands. Market research, customized products, recommendation systems and transport systems all are based on demographic data analysis.

This project aims to provide a descriptive analysis of demographic data collected from 227 countries in 5 regions around the globe. This data includes the infant mortality rate, life expectancy for both sexes, and individual life expectancy for women and men for the years 2001 and 2021. To do this, we first look at the frequency distributions of all variables using histograms. We observe the mean and median of all distributions to compare different distributions. Then we check dependencies between variables by pair plots. Box plots are used to study variability within and between sub-regions. Finally, to observe the changes that occurred in 20 years, we compare the 2001 variables to the 2021 variables.

The second section describes the structure and quality of the data set in more detail. Additionally, the exact goals of the project are stated in the second section. The third section explains the measures of the central tendencies (mean, median), measures of dispersion for example variance and the correlation coefficient. It describes the charts (histograms, boxplot, scatterplot, heatmap, pairplot) used in the statistical analysis. The fourth section focuses on the application of these methods and the interpretation of the graphs. Finally the fifth section summarizes the most important findings and discusses possible further analyses of this data set.

2 Problem statement

2.1 Data set and data quality

This report deals with the analysis of a small sample of the data set for the years 2001 and 2021 taken from the online International Database (IDB) of the U.S. Census

Bureau(Bureau, 2022). The international database contains data from more than 200 countries and areas of the world since the 1960s. The Census Bureau provides population estimates and projections using IDB. The data is gathered through censuses, surveys, administrative records, and vital statistics. IDB is also used for research in education, journalism and business.

The data set consists of 454 observations and eight variables. Description of variables according to the Census Bureau (Glossary, 2021) is given below:

Variable	Type	Description
Country name	Nominal	Name of the country. The data set contains data from 227 different countries.
Subregion	Nominal	The sub-regions of a country and has total 21 unique values (sub-regions).
Region	Nominal	Region name to which the country belongs. The data set contains five regions.
Year	Integer	The year in which the data was collected. In the data set it is either 2001 or 2021.
Life expectancy of both sexes	Float	Average number of years a cohort can be expected to live if mortality at each age remains constant in the future.
Life expectancy of males	Float	Average number of years a group of males can be expected to live if mortality at each age remains constant in the future.
Life expectancy of females	Float	Average number of years a group of females can be expected to live if mortality at each age remains constant in the future.
Infant mortality rate of both sexes	Float	Number of deaths of infants under 1 year of age from a cohort of 1,000 live births. Denoted by 1q0 or IMR, it is the probability of dying between birth and exact age 1.

Table 1: Data Description

For six countries Libya, Puerto Rico, South Sudan, Sudan, Syria and the United States there is no information in the data set on life expectancy and mortality rate for the year 2001. We might omit these missing data observations. Because the US Census Bureau is an official American agency that regularly updates the IDB to provide the latest information, high data quality is expected.

2.2 Project objectives

In this project, descriptive analysis of the given data set is performed. First, the frequency distribution of variables for the year 2021 is analyzed by visualizing frequency distributions using histograms. The result is interpreted by looking at the mean and spread of the distributions. Then bivariate correlations between variables in the year 2021 are observed using a pairplot and the linearity of their relationship is measured using Pearson correlation coefficients.

Homogeneity of all variables of 2021 within the subregions and between different subregions is observed through box plots. For this, we analyze the variability of the values within the subregions and then the central values of the individual variables between different subregions are compared. The change in mortality rate and life expectancy for both sexes from the year 2001 to 2021 is observed through scatter plots.

3 Statistical methods

In this section, statistical methods used for descriptive data analysis are discussed. Statistical measures, coefficients and graphs are presented. For calculation of all statistical measures and graphical representations R software (R Core Team, 2022) Version, 4.2.1 is used with additional packages **GGally** (Schloerke et al., 2021), **dplyr** (Wickham et al., 2022), **ggplot2** (Wickham, 2016), **gridExtra** (Auguie, 2017).

3.1 Statistical Measures

Statistical measures summarize the characteristics of a data set. There are different measures for different types of data. Measures of location/central tendency tell the centre of values for example mean, median and mode. Measures of the dispersion tell dispersion/spread of the values around the centre such as range variance and interquantile range. In measures of correlation, correlation coefficients are used to analyze the bivariate relationships as Pearson correlation coefficient. These measures are used to analyze continuous data. In the following subsections description of these methods is represented.

3.1.1 Measures of Location/Central Tendency

For continuous numerical data mean, median and mode are measures of location ("Center" of the attribute values $x_1, x_2, \ldots, x_n$), which characterize the attribute values of the sample data by a single value. The choice of a measure of location can have a decisive effect on the analysis decision.

Mean: The mean is one of the most common ways of measuring the location of data and is used to quantify the centre of the sampled data. The mean is calculated as the sum of the data values divided by the number of observations, n. Formally, for observations $x_1, x_2, \ldots, x_n$ mean is denoted as $\bar{x}$ and is defined as follows: (Crowder et al., 2020, p. 61)

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

The mean can be calculated for both discrete and continuous values, but not for categorical data because it is not possible to sum categorical data. Although the sample mean is the most commonly used measure, it is sensitive to outliers, which means it can be heavily skewed toward outliers when they are present.

Median: The median ("middle number") is another common type of measure of central tendency. It is defined as $50th$ percentile of the data, which means 50% observations fall above and 50% observations fall below the median. For data with extreme values, the median can be a more meaningful measure of central tendency as it is more robust than the mean in the presence of outliers. (Crowder et al., 2020, p. 61)

The median is a value that divides the ordered data into two equal parts. First we sort the data and then choose the middle value if n is odd or we take the average of two middle values if n is even. So it's not necessarily a value in the data set.(Kaptein and van den Heuvel, 2022, p. 15)

$$Q_{0.5} = \text{median} := \begin{cases} X_{\frac{(n+1)}{2}}, & n \text{ odd} \\ \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, & n \text{ even} \end{cases}$$

The idea of the median can be generalized to the quantiles. A quantile Q_q, $q \epsilon [0, 1]$ is a value that splits the data into two parts and $q * 100\%$ of the attribute values of the observations lie above and $(1 - q) * 100\%$ lie below this value. $q = 0.25, q = 0.5$ and $q =$

0.75 are called the first, second and third quantiles, respectively. The median is the 0.5 quantile ($Q_{0.5}$).(Kaptein and van den Heuvel, 2022, p. 16-17)

3.1.2 Measures of Dispersion/Spread

Distributions that have the same location parameters may differ from each other. Therefore we also look at the dispersion of the values $x_1, x_2, \ldots, x_n$. The measure of dispersion is actually a measure of uncertainty. It quantifies the range of data.

Variance and Standard Deviation: Variance is a popular measure of spread and it quantifies the spread between values of the data set. It is defined as the average of the squared deviations from the mean. Let s^2 denote the variance and $\bar{x}$ the mean of the data set then the variance is calculated as:(Crowder et al., 2020, p. 62)

$$s_x^2 = \text{var}_x := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The sample standard deviation is the square root of the sample variance, denoted by s_x. $s_x := \sqrt{\text{var}_x}$. The standard deviation is the most commonly used measure because it is on the same scale as observed data instead of a squared scale.(Crowder et al., 2020, p. 62)

Skewness: The skewness is a measure of the imbalance/asymmetry of a distribution. A distribution is asymmetric if the spread on one side of the centre of the data is greater than the spread on the other side. Skewness is calculated as:

$$g_1 = \frac{1}{n} \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{s_x})^3$$

If g_1 is positive the distribution is right-skewed means the spread of data is more on the right side. For a negative value of g_1 distribution is left-skewed. For g_1 equal to zero the distribution is assumed to be symmetric about the center.(Kaptein and van den Heuvel, 2022, p. 18)

Range: The range is the simplest measure of dispersion. It is the difference between the maximum (x_n) and minimum (x_1) value in the sample data. The range (R_x) is calculated as:

$$R_x := \max(x) - \min(x) = x_n - x_1$$

The range is the most sensitive to outliers. Secondly, it only tells the difference between the maximum and minimum values in the data set and does not provide any information about the spread of the all other values in the data set. (Kaptein and van den Heuvel, 2022, p. 17)

Interquartile range (IQR): As the range is sensitive to the outliers, the Interquartile range (IQR) is another robust measure of spread which is defined as the difference between the third quantile $Q_{0.75}$ and the first quantile $Q_{0.25}$. It measures the range in which 50% of the middle data falls. (Kaptein and van den Heuvel, 2022, p. 17)

$$\text{IQR} := Q_{0.75} - Q_{0.25}$$

The interquartile range is also visualized in a box plot, which we will discuss later in this section.

3.1.3 Measures of Correlation

The correlation quantifies the strength of the relationship between two variables. The correlation coefficient r_{xy}, a number ranging from -1 to 1, measures the linear dependency between two continuous variables.

Pearson Correlation Coefficient: The Pearson correlation coefficient r_{xy} is the most commonly used correlation coefficient which measures the strength of the linear relationship between two continuous variables. Lets take x_i and y_i as individual observations, $\bar{x}$ and $\bar{y}$ means of variable x and y then the Pearson correlation coefficient r_{xy} is calculated as follows:

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

$r_{xy} = 1$ if y and x lie on a straight line with a positive slope which shows a strong positive correlation. and $r_{xy} = -1$ if the slope is negative, showing a strong negative correlation. While $r_{xy} = 0$ implies no correlation. Correlation is undefined if x or y is not varying. (i.e. s_x or $s_y = 0$). (Crowder et al., 2020, p. 62)

3.2 Graphical Representations

Along with the statistical measures mentioned above, data visualization also can be useful to analyze data distribution, outliers in the data set, trends, correlations and

unusual variations. We choose graphical tools according to the type of data.(Crowder et al., 2020, p. 63) In the following section different graphical methods used in data analysis are represented.

3.2.1 Histogram

A histogram is a graphical chart, commonly used to show frequency distribution through vertical bars. To create a histogram we divide the variable range into equally spaced intervals which are placed on the x-axis and the y-axis shows the total number of data points in each interval. The height of the bars shows the frequency of data points in each interval.

A histogram provides a graphical representation for location measurements and data dispersion and is also helpful in identifying outliers. It also provides significant information about the distribution of the variable. (Crowder et al., 2020, p. 64)

3.2.2 Scatterplot

A scatterplot is a chart which plots individual data points as dots. It is used to visualize the linear relationship between two variables. It gives a visual representation of the correlation between the values of two variables. We can see if the relationship is linear or curved by looking at the linearity of the data points on the chart. The closer the data points are, the more strongly they are correlated, while data points that are further apart show a weak correlation.

In a scatterplot data points that start from the bottom up show a positive correlation, while the spread from the top down shows a negative correlation. Correlation is said to be positive when the values of two variables increase together, and negative when the values of one variable increase when the other decreases.(Crowder et al., 2020, p. 65)

3.2.3 Boxplot

A boxplot is another graphical representation which represents a five-number summary (Minimum, 25%−Quantile, Median,75%−Quantile, Maximum). This provides a visual representation of the location, spread and variation of the variables. In a boxplot, middle line is the median and the box itself with a lower border at 25%−Quantile and an upper border at 75%−Quantile indicates the middle 50% of the data.

Whiskers in "standard boxplot" show the smallest (x_1) and largest values (x_n). While in "modified boxplot" whiskers are drawn from smallest to largest values in the interval $[Q_{0.25} - 1.5\text{IQR}, Q_{0.75} + 1.5\text{IQR}]$. Observations outside this range are said to be potential outliers. Boxplots are useful when we want to compare a continuous variable between two subgroups.(Kaptein and van den Heuvel, 2022, p. 23)

3.2.4 Pairplot

A pairplot plots the pairwise relationships between the variables through a scatter plot in a matrix form. This creates a compact and nice visualization that is helpful for understanding the data at a glance in one plot.

4 Statistical analysis

In this section, detailed descriptive analysis of the data set is discussed using statistical measures and plots mentioned in the previous section. We also omitted the six countries with missing values.

4.1 Frequency Distributions of Variables

The frequency distribution of all the variables in the year 2021 is analyzed by using histograms. Figure 1 shows histograms of all four variables in the data set. On the x-axis class intervals of the respective variable are shown while the y-axis shows the frequency of each interval. Light blue shows the higher frequency while dark blue color shows the lower frequency of the interval.

Figure 1(a) shows the frequency distribution of life expectancy of females. The distribution is asymmetric, which means the frequency is not equally distributed on both sides of the mean value. The highest frequency is in the interval of $80 - 85$, which means in most regions life expectancy of females is between $80 - 85$ years. An outlier can also be observed in this distribution which shows the life expectancy of females higher than 90. It is Monaco, which is a country in Western Europe, life expectancy of females is 93.4 here. The lowest female life expectancy is observed in Afghanistan which is 54.85.

In Figure 1(b) frequency distribution of life expectancy of males is shown. It can be seen that distribution is asymmetric, the highest frequency for males lies between $70 -$

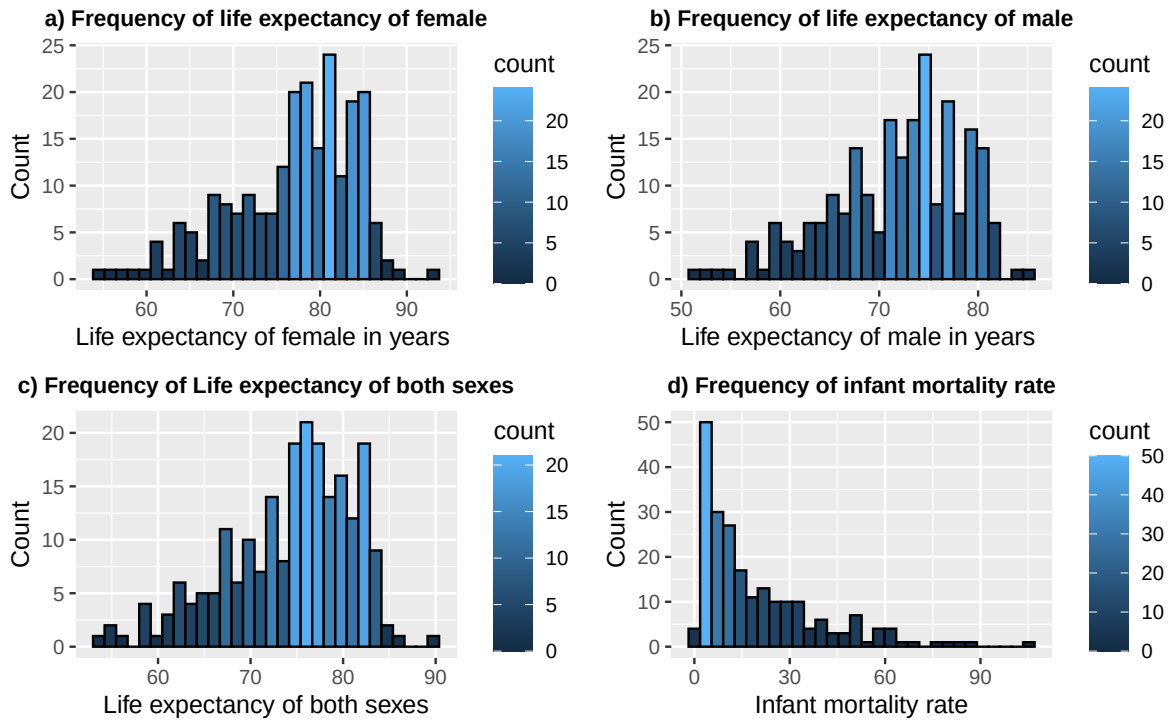

Figure 1: Histograms for (a) Life expectancy at birth of females, (b) Life expectancy at birth of males, (c) Life expectancy at birth of both sexes, (d) Infant mortality rate in the year 2021

75, which means the life expectancy of males is between $70 - 75$ years in most of the regions. Again Monaco is showing as a potential outlier with a male life expectancy 85.55, the highest among all others. The lowest male life expectancy range is $50 - 55$ with Afghanistan having the lowest male life expectancy 51.73.

Figure 1(c) represents the frequency distribution of life expectancy of males and females both. This distribution is asymmetric with most values in the upper middle range and very few in the upper and lower range. It shows life expectancy of the whole population in most countries lies between $75 - 80$. The lowest expectancy range is $50 - 55$ only Afghanistan lies in this range. The highest range is $85 - 90$ years, two countries (Singapore and Monaco) lie in this range.

The frequency distribution of infant mortality rate is shown in Figure 1(d). Distribution is right skewed which shows that the mean will be higher (overestimate). We see 50 countries have infant mortality rates between $1 - 5$. Six countries have mortality rate between $65 - 90$ from which two countries Somalia and the Central African Republic

have the mortality rates between $84-90$. An extreme deviation can be seen in the graph it is Afghanistan, which has the largest infant mortality rate 106.75.

Next, we analyze the difference between the life expectancy of females and males. Figure 2 represents the distribution of the difference. Almost 45 countries lie between $5-6$ and 101 countries lie between $4-6$. This means that women have $4-6$ years more life expectancy than men. In Belarus, Russia and Lithuania this difference is highest, almost 11 years. We observe one deviation on yhe left side of the plot, where this difference is almost -2. This is because we calculated the difference as:

$$\text{Life expectancy of females} - \text{Life expectancy of males}$$

In all the regions females' life expectancy is greater than males. But in Montserrat males' life expectancy is almost 2 years more than females, this deviating behavior can be seen in the plot.

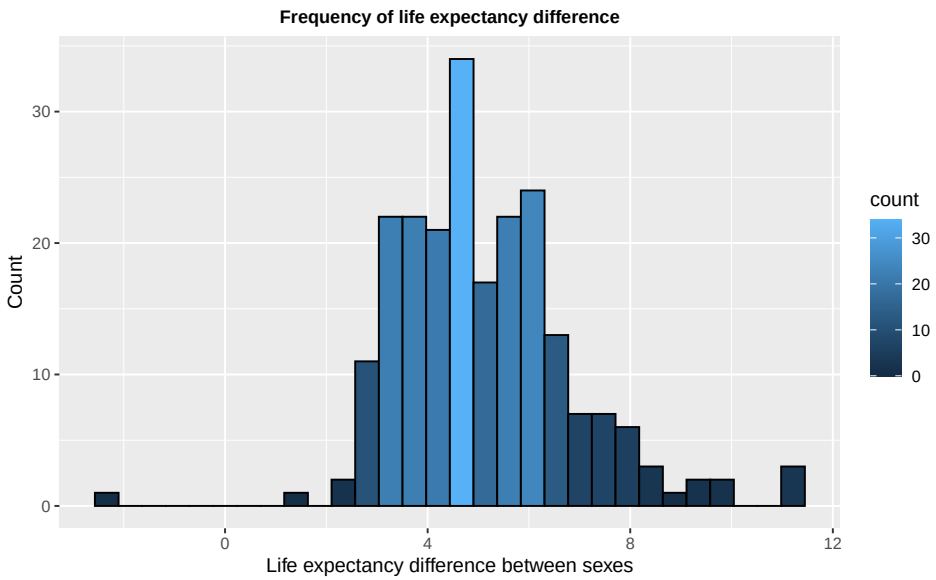

Figure 2: Histogram for the difference between male and female life expectancy

4.2 Bivariate Correlations Analysis

In this section, we'll discuss dependencies between all variables in the year 2021. For that, we plotted a pairplot shown in Figure 3. In the lower half of the chart scatter plots between all variables are displayed while in the upper half values of the Pearson

correlation coefficient are shown. Because relationships between variables are linear we are using the Pearson correlation coefficient. Region-wise correlation coefficient values are also displayed in plots which show region-wise strength of dependencies between variables.

From the data points spreading top to bottom in the scatter plots and the negative correlation coefficient values it can be seen that the mortality rate is negatively correlated between all the other three variables. Large overall correlation coefficient values show a strong negative correlation. This means when other variables increase mortality rate decreases and vice versa.

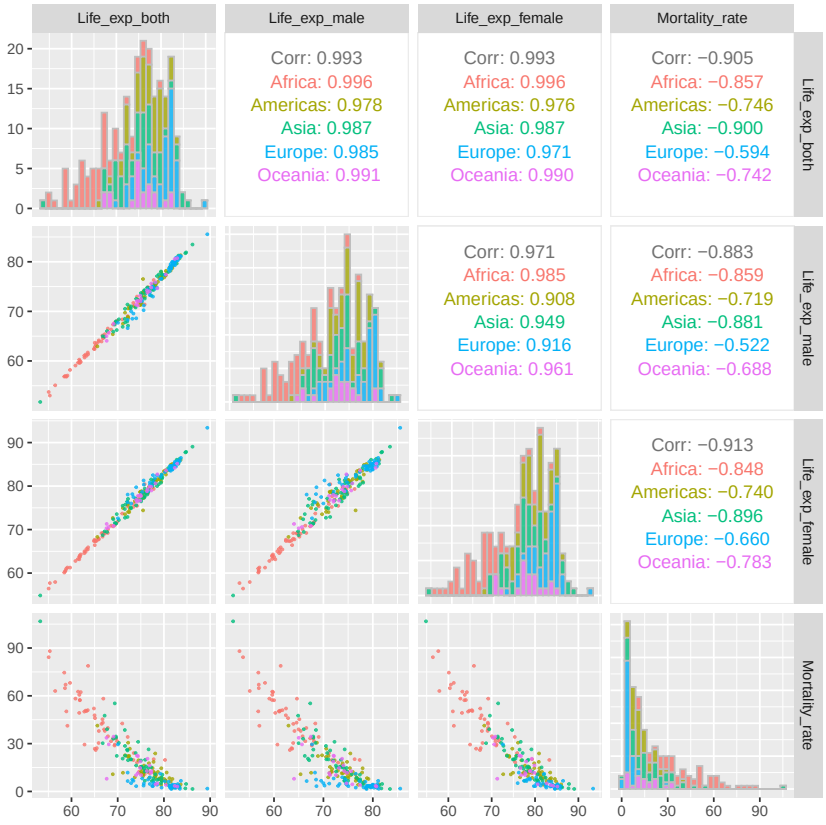

Figure 3: Pairplot of all the variables in 2021

We see small correlation coefficient values between the mortality rate and the other three variables for Europe which could mean that these variables do not play a key role in determining the mortality rate in Europe. For Asia, these coefficient values are highest than all other regions. The correlation between mortality rate and female life expectancy is strongest than other variables.

Scatter plots and coefficient values show that the life expectancy of both sexes has a strong positive linear correlation with female and male life expectancy. Here we see for all the regions also correlation coefficient values are high. Female life expectancy and male life expectancy also have a strong positive linear correlation.

4.3 Analysis of Variability of the Variables

In this section homogeneity and heterogeneity of variables within and between sub-regions is discussed. Boxplots are used to visualize the variability of variables.

Figure 4 shows that life expectancy is highest in Europe. We observed some outliers, especially in Western Europe as we have seen earlier Monaco has the highest life expectancy.

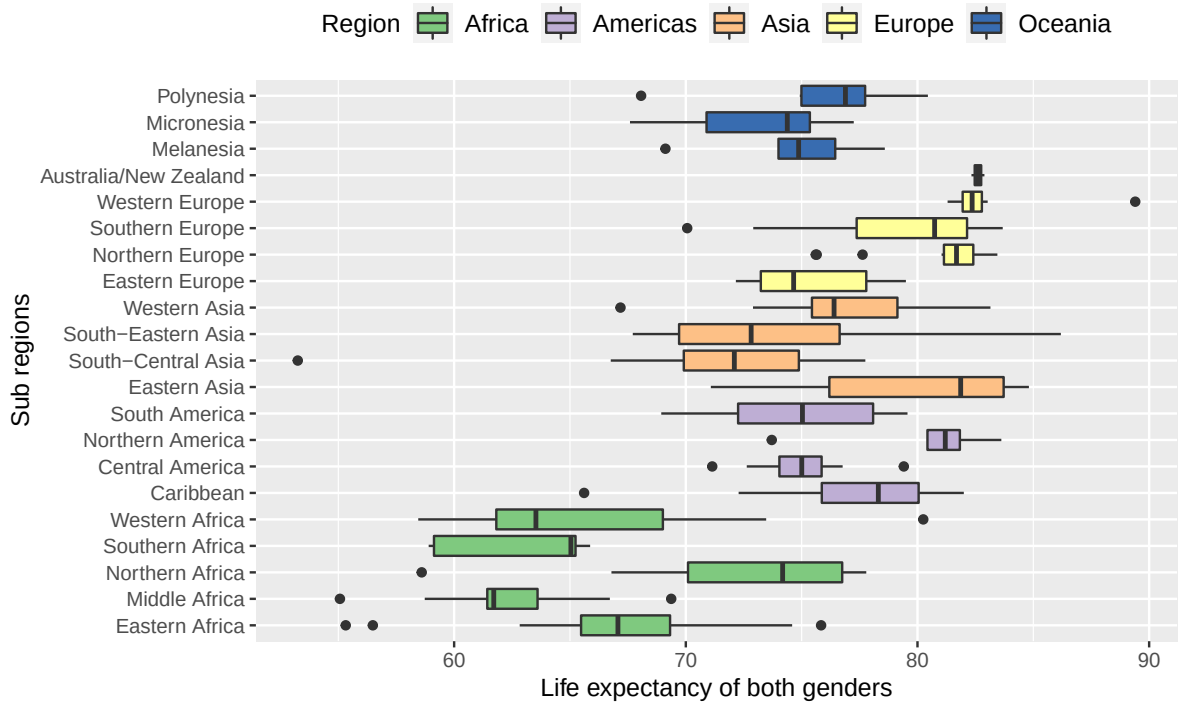

Figure 4: Boxplots of Life expectancy of both genders in 2021

Nevertheless, we see in Western Europe and Northern Europe values are homogeneous. After Europe, America is showing a larger overall life expectancy. In Northern and Central America homogeneity is observed. In Oceania the life expectancy rate is also high especially in Australia and New Zealand it's higher than in Europe, with a constant variance. The highest spread and lowest life expectancy are visible in Africa.

Asia is showing high spread, especially in South-Eastern and South-central Asia. Life expectancy in Asia is higher than in Africa. An outlier, Afghanistan can be seen in South-central Asia which has the lowest life expectancy rate. Overall values are not homogeneous in Asia.

In Figure 5 boxplots of region-wise infant mortality rates are shown. The mortality rate in Europe is the lowest and we observe homogeneity in all the regions of Europe.

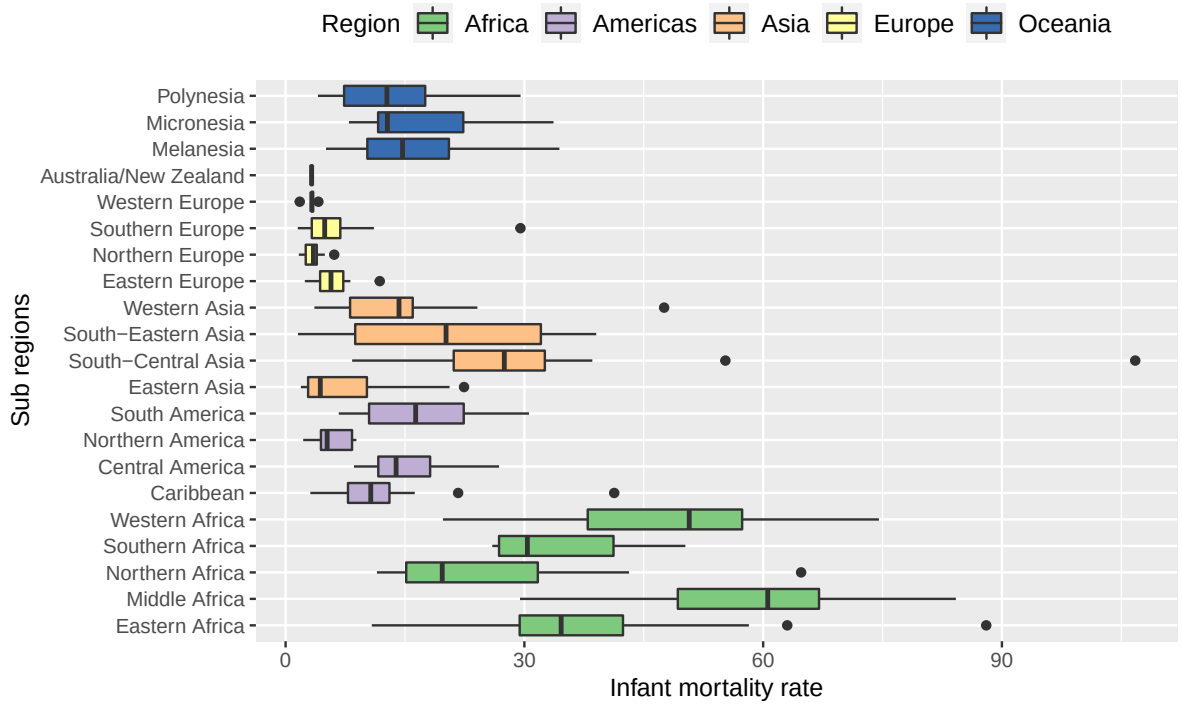

Figure 5: Boxplots of infant mortality rate in 2021

We observe one outlier in Southern Europe which is Kosovo where the mortality rate is almost 30. America shows the second lowest overall mortality rate, especially in Northern America it's homogeneous. In Oceania, Australia and New Zealand show the lowest mortality rate compared to all other subregions. The mortality rate is highest in Africa and values are heterogeneous in all sub-regions of Africa. In Asia Eastern Asia has the lowest mortality rate and less spread compared to the other subregions in this region. We observe values in South Eastern Asia are heterogeneous. In South Central Asia, Afghanistan is shown as an outlier, which has the highest mortality rate of any country.

4.4 Comparison of Variables from 2001 with 2021

Figure 6 and Figure 7 illustrate the comparison of life expectancy of both genders and mortality rate in all sub-regions in the year 2001 with 2021. On the x-axis values of 2001 are shown while the y-axis shows the values from 2021. The diagonal line is an indicator of the change. Points lie below the line representing the values of the variables which decreased in the year 2022, and points lie above the diagonal line indicate an increment in the values of the variables in the year 2021. Points lie on the line that show no change has occurred in the values of variables in 2021 compared to 2001. Different colors indicate different regions while individual data points indicate countries.

In Figure 6 we see almost all the points are above the diagonal line which shows that life expectancy is increased for almost all the countries in 2021 compared to 2001. There are only three countries Mexico, Peru and Venezuela in America for which it is decreased 2021.

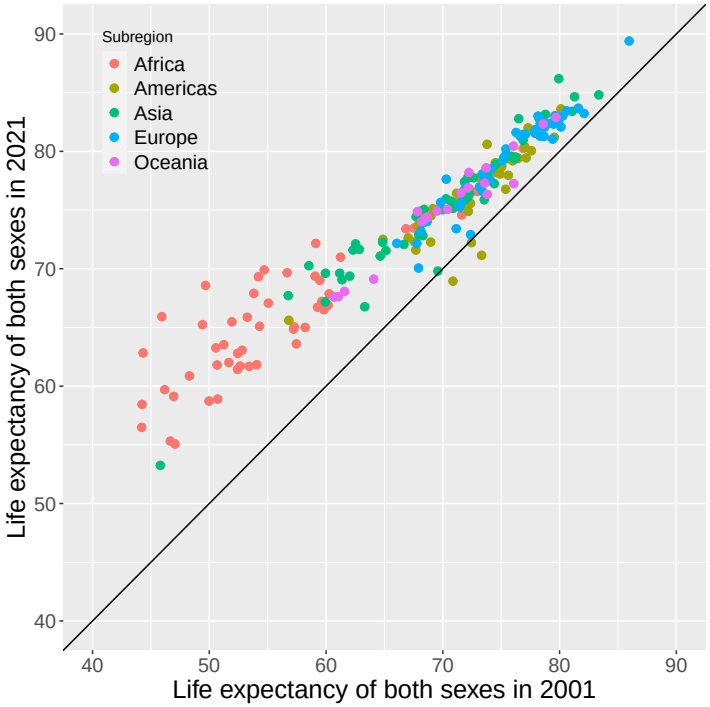

Figure 6: Change in life expectancy from 2001 to 2021

Figure 7 shows a decreasing trend in mortality rate. It indicates that the overall mortality rate decreased in 2021 as compared to 2001. In Europe, America and New Zealand/Australia it is almost constant. In six countries Panama, Venezuela, Malaysia, Croatia, Malta and Guam mortality rate increased in 2021 as compared to 2001.

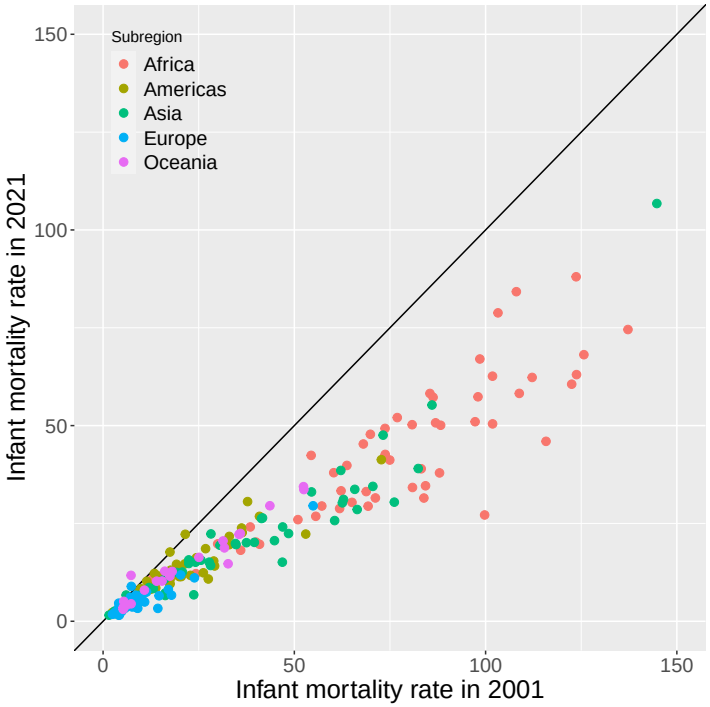

Figure 7: Change in mortality rate from 2001 to 2021

5 Summary

In this project, a descriptive analysis of the data of 227 countries obtained from the US Census Bureau(Bureau, 2022) was performed. Frequency distributions of all variables, bivariate correlations, variability of variables and changes in mortality rate and life expectancy from 2001 to 2021 were analyzed. The study showed that women are expected to live longer than men in almost all countries. This may be due to behavioral or biological differences between females and males. The analysis of dependencies showed a negative correlation between mortality rate and life expectancy. The highest mortality rates were observed in Africa, while it had the lowest life expectancy of all other regions. On the other hand, the subregions of Europe had lower mortality and higher life expectancy. This may be due to the difference in socioeconomic conditions in the two regions. Afghanistan (a country in Central Asia) was observed as an extreme outlier, which had the highest death rate and lowest life expectancy of any country. For further analysis, more information about men and women would provide insight into individual differences in life expectancy between men and women. Additionally, we can add more variables in the study to reveal the factors that change mortality rate and life expectancy.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3.

U.S. Census Bureau. International database: World population estimates and projections, 2022. URL https://www.census.gov/programs-surveys/international-programs/about/idb.html. Accessed: 04-11-2022.

Stephen Crowder, Collin Delker, Eric Forrest, and Nevin Martin. *Introduction to Statistics and Probability*, pages 59–80. Springer International Publishing, Cham, 2020. ISBN 978-3-030-53329-8. URL https://doi.org/10.1007/978-3-030-53329-8$_4$.

U.S. Census Bureau Glossary, 2021. URL https://www.census.gov/programssurveys/international-programs/about/glossary.html. Accessed: 04-11-2022.

Maurits Kaptein and Edwin van den Heuvel. *A First Look at Data*, pages 1–37. Springer International Publishing, Cham, 2022. ISBN 978-3-030-10531-0. URL https://doi.org/10.1007/978-3-030-10531-0$_1$.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.