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: Prerana Rajeev Chandratre

Group number: 5

Group members: Janani Veeraghavan, Bushra Kiyani, Sathish Ravindranath Kabatkar, Shivam Shukla

November 10, 2022

Contents

1 Introduction

In the development of a country, aspects such as demographic and economic growth play a vital role. The collection of demographic data aids in the analysis of health issues, medical conditions, and diseases. Understanding and analyzing the previous demographic data will help to improve human life in the future. Such analysis could be helpful for the growth and welfare of the country. The U.S. Census Bureau collects demographic statistics around the world in 200 countries with populations of 5,000 or more.

For this project, the data set holds the details of 227 administrative regions around the world, which include states, counties, regions, and subregions of different countries. The countries are divided into 5 regions and 21 sub-regions. The variables of the given data include the life expectancy of males, the life expectancy of females, the life expectancy of both sexes, and infant mortality rate. The given data has information from the years 2001 and 2021 (U.S Census, 2022).

The goal of this project is to analyze the provided dataset to find the frequency distribution of the variables. The second task of the project is to check the relationship between infant mortality and life expectancy for both sexes. Another task is to find the variability of life expectancy of both sexes and infant mortality rate within and across all the subregions. Finally, the comparison of data from the year 2001 and data from the year 2021 has to be done. Descriptive statistical measures and methods such as central tendency measures, variability measures, and graphical representations have been used as part of this project to achieve the mentioned tasks.

Including the introduction section, this report is divided into five sections. The objectives of this report, along with a brief description of the dataset and its variables, are discussed in Section 2. Section 3 introduces and discusses the statistical methods like mean, median, and quartile, and the graphs such as histogram, box plot, and scatter plot used in this report.Then, the methods discussed in the previous section are applied to the dataset in Section 4, and the analysis and results are discussed. Finally, a brief summary of the findings and a discussion of the project's objectives are presented in Section 5.

2 Problem Statement

2.1 Description of the Dataset

This project is based on an analysis of the International Database (IDB) collected by the U.S. Census Bureau. The Census Bureau has collected U.S. population data since 1960, and this dataset is available on their website. The given dataset "census 2001–2021.csv" contains demographic data of 227 countries and cities around the world, collected in the year 2001 and 2021. Data includes information such as life expectancy of men and women and infant mortality rate. This project uses a small extract of the demographic data from the IDB (U.S Census, 2022).

2.2 Variables and Data Quality

In this section, all the variables of the data set are given in Table 1. There are 454 observations in total. From the year 2001 data, 6 observations are missing. The missing observations are not considered for the analysis. The variables are explained with detailed descriptions (U.S Census, 2022), and the type of each variable is mentioned below:

Variable	Data Type
Country	Nominal
Region	Nominal
Sub region	Nominal
Year	Numeric(Discrete)
Life Expectancy both sexes	Numeric (Continuous)
Life Expectancy Males	Numeric (Continuous)
Life Expectancy females	Numeric (Continuous)
Infant mortality rate	Numeric (Continuous)

Table 1: Variables and their data type

The variable *Country* which is a character type lists the 227 country names. The variable *Region* has five values: Asia, Europe, Africa, America, and Oceania. These regions are divided into smaller territories called *Sub region*. There are 21 sub-regions in total. The numeric variable *Year* has two values, 2001 and 2021, indicating the year in which the information is collected. The biological factors such as life expectancies, mortality rate,

population measures are included in the different variables. The variable *Life expectancy of both sexes* gives the average of the life expectancies of both the males and females. The variables *Life expectancy of females* and *Life expectancy of males* shows how long a newborn female and a male child is expected to live on an average, correspondingly, assuming the current mortality rates remain constant. Also, the variable *Infant mortality rate* shows the number of infant dying under a year of age, per 1,000 births. It tells the probability between death and birth.

2.3 Project Objective

The main objective of this project is to perform descriptive analysis on the data set. The analysis tasks are carried out in four stages. For the first three tasks, the information (data values) of the year 2021 are considered. First, the frequency distributions of the variables, life expectancy of both sexes and an infant mortality rate are observed with the help of histogram. Next, the bivariate correlation is observed between life expectancy and infant mortality rate with the help of the correlation methods. Afterwards, the homogeneity and heterogeneity of the variables across the regions and sub-regions have been observed. Finally, the differences in the variables from the year 2001 and 2021 are examined, and the results are summarized.

3 Statistical Methods

In this section, descriptive statistical measures such as mean, median, variance, interquartile range, and graphical representations, i.e. histograms, box plots, and scatter plots are explained in detail with formulas. In this analysis tasks, software Jupyter Notebook (Fernando Perez, 2015) Version 6.3.0 running on R (R Core Team, 2020) Version 4.1.0 is used. The graphs are created using R packages, namely ggplot(Wickham, 2016), dplyr (Wickham, 2022), GGally (Schloerke, 2020) and gridExtra (Auguie, 2017).

3.1 Measures of Central Tendency

3.1.1 Mean

The sample mean, also known as the arithmetic mean or mean, is one of the commonly used descriptive measures of central tendency is found in Sheldon (2010, p. 73-76). Its

value is often affected by the extreme data points in the vector. Consider a sample data vector $x_1, x_2, \ldots, x_n$ with n data points, then the sample mean $\bar{x}$ is given by the formula:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

3.1.2 Median

In the arithmetic average, data points with extreme values often affect the measurement of the center. In this case, the sample median is a better measure. The data points are sorted in an assending order and the middle value is referred as the median is found in Sheldon (2010, p. 83). If the data points are odd in number, the middle value of those data points is considered as a median and for an even number of data points, average of the two middle data points is taken as median. For a sample $x_1, x_2, \ldots, x_n$ of size n arranged in ascending order, the formula for calculating the median is:

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

3.2 Measures of Variability

3.2.1 Range

The range is the difference between the maximum and minimum value in the data vector can be found in Hay-Jahans (2019, p. 77). Consider the data vector $x_1, x_2, \ldots, x_n$, let x_1 be the minimum value $x_{\min}$ and x_n be the maximum value $x_{\max}$, then range R, is given by the formula:

$$R = x_{\max} - x_{\min}.$$

3.2.2 Interquartile Range

The inverse of a cumulative distribution function gives the p^{th} quantile, which is the probability of getting a value in the data less or equal to it is found in Dalgaard (2008, p. 63). Since it is a probability value, it takes values between 0 and 1. The values of p, such as 0.25, 0.5 and 0.75, which is often referred as 1^{st} quartile, median or 2^{nd} quartile and 3^{rd} quartile respectively is found in Dalgaard (2008, p. 68). These three values

divides the data into four parts. The interquartile range is the difference between the 3^{rd} quartile and 1^{st} quartile.

3.2.3 Variance

The sample variance is the most commonly used measure of spread. The distance between two points is also known as the deviation. The deviations from each data point to its mean is calculated, then the sum of squared deviations gives the sample variance is found in Sheldon (2010, p. 99). For a given data vector $x_1, x_2, ..., x_n$ with mean value $\bar{x}$, the sample variance s^2 is given by below formula:

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

3.2.4 Standard Deviation

The units of the measurement changes as the deviations are squared. To get the units same as the data points and mean value, standard deviation is used. The sample standard deviation is equal to the square root of the sample variance found in Sheldon (2010, p. 103). The formula for sample standard deviation is:

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

3.3 Pearson Correlation Coefficient

The correlation coefficient is a measure that quantifies the linear relationship between two variables. It takes values between -1 and +1. If an increase in the value of one variable, increases the other variable, and similarly, if a decrease in one variable decreases the other variable, is said to be positively correlated. Or if an increase in the value of one variable decreases the other, then it is said to be negatively correlated. If, a correlation value is 0.0 then it interpret that, there is no linear relation among those two variables is found in Sheldon (2010, p. 123). For two numeric variables $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ with mean value of $\bar{x}$ and $\bar{y}$ respectively, the correlation coefficient r

is defined by the formula:

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

If $r > 0$, the variables are positively correlated else if, $r < 0$, the variables are negatively correlated.

3.4 Graphical Methods

In this sub section, different types of graphs are explained which are used in the project for graphical representation.

3.4.1 Histogram

The histogram is a simple plot to visualize the distribution of a given variable. The data points in the variable are sorted in ascending order and split into equal intervals. The data points in the particular intervals are counted and a bar is drawn above the intervals. The height of the bars represents the frequency of the data points is found in Sheldon (2010, p. 32). The histogram or frequency distribution is said to be symmetric around its middle bin or mean value if the height of the bins decreases on both sides equally in a bell-shape. If the height of bins on the right side is higher than on the left side, the histogram will be called as left-skewed histogram. Correspondingly, if the bins' height is more on the left side than the right side, they are right-skewed histograms.

3.4.2 Scatter Plot

A scatter plot is a simple way to visualize the relationship between two variables. The data points are plotted as coordinates (x, y) in a two-dimensional plot, where the X-axis represent one variable and the Y-axis represents another variable, is found in Sheldon (2010, p. 51). There is another popular way to draw the scatter plots in pairwise combinations of variables; it is called a matrix scatter plot or pair plot. It helps to analyze and plot the pairwise combinations of multiple variables that are in a multivariate data set is found in Hay-Jahans (2019, p. 167)

3.4.3 Box Plot

The box plot helps to visualize how the data is distributed. It gives the summary of five measures namely minimum of the variable, first quartile (Q1), median, third quartile (Q3), and maximum of the variable is found in Hay-Jahans (2019, p. 137). In box plots, the height of the box represents the interquartile range and the line in the middle of the box is the median. Lines drawn from the box to the minimum "$(1.5 * 1^{st} quartile)$" and maximum "$(1.5 * 3^{rd} quartile)$" are called whiskers. Outliers are data points that are outside the whiskers.

4 Statistical Analysis

A brief explanation of the results of this project is given in this section. Results of the tasks mentioned in the project objective are obtained by using the discussed statistical methods.

4.1 Frequency Distribution of Variables

This section uses histograms to infer frequency distributions of variables life expectancy and infant mortality rate. The distributions of life expectancies of females and males are shown in Figures 1(a) and (b), respectively. Both the distributions are left skewed, as shown in the plot, the height of the bins on the right is greater than on the left. The mean values of the life expectancies of women and men are 76.89 and 71.78 respectively. On an average, life expectancy of women is higher as compared to men.

In addition, In Figure 1(c), the distribution of the differences of life expectancies of both sexes is shown, with a mean value of 5.11 and a median of 4.87. These values conveys that, in the year 2021, the average rate of life expectancies of both the sexes has increased. The distribution of infant mortality rate is shown in the Figure 1(d). The height of the right bin is smaller compared to the left bin, so the distribution is right skewed, with a mean of 20.25 and a median of 12.58. The summary of central measure tendency for the year 2021 is given in the Table 2.

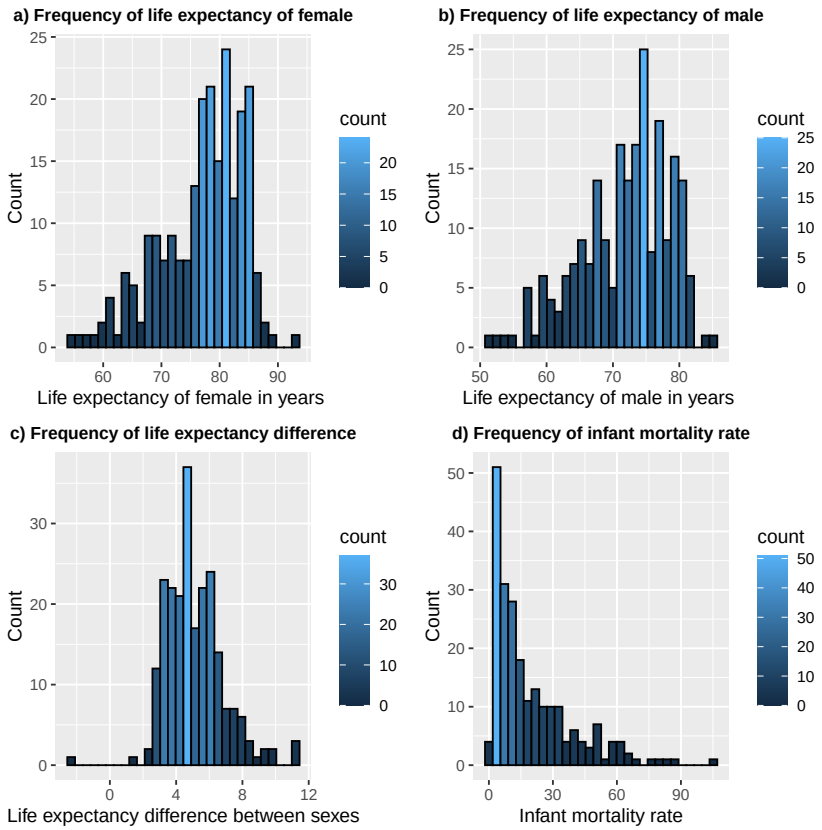


Figure 1: Histograms of frequency distribution of difference in life expectancy of both sexes and infant mortality rate.

	LE_male	LE_female	LE_both	LE_diff_btw_sexes	Mortality_rate
Min	51.73	54.85	53.25	-2.11	1.53
Q_1	67.59	72.29	69.73	3.84	6.27
Median	72.99	78.36	75.56	4.87	12.58
Mean	71.78	76.89	74.28	5.11	20.25
Q_3	76.94	82.34	79.42	6.07	29.48
Max.	85.55	93.40	89.40	11.44	106.75
Std	6.74	7.21	6.91	1.74	19.19

Table 2: Measures of central Tendency and spread in the year 2021 (Std - standard deviation, LE - life expectancy at birth)

4.2 Correlation between the Variables

Correlations between variables, infant mortality rate and life expectancy for men and women are shown in Figure 2 with the help of a scatter plot and a histogram. The scatter plot shows that all points are close to each other linearly and interprets the linear relationship between the variables. Pearson's correlation coefficient method is used for this analysis.

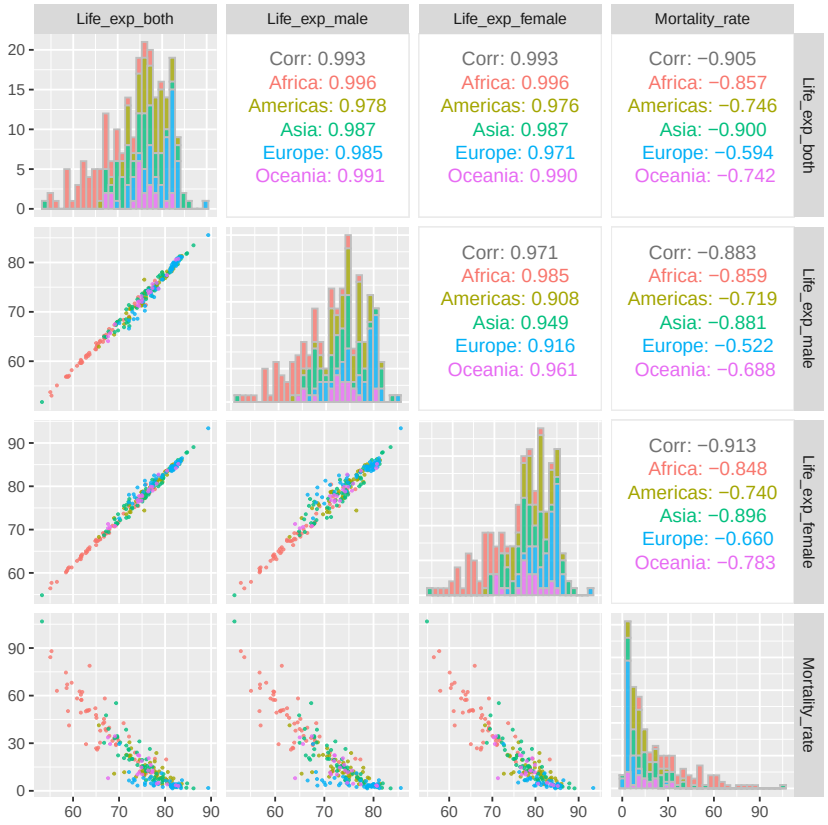


Figure 2: Pair plot of correlation between the life expectancy of males, females with their mortality rate

The correlation coefficient value for life expectancies of both sexes with their mortality rate is -0.905 which indicates that there is strong negative linear relationship between life expectancy and infant mortality rate. The life expectancy of men and women increases as the mortality rate decreases. Also, from Figure 2, it is observed that, the mortality rate of the African and Asian region is relatively high with the correlation value -0.859 and -0.881 respectively as compared to the European region which has the low correlation

value i.e -0.522. The life expectancy of both the sexes shows the positive correlation among life expectancy of individual males and females in all the regions.

4.3 Variability Analysis of the Data

The variability of the life expectancies of both sexes across all the sub regions is analyzed with the box plots shown in Figure 3. The median value and the interquartile range in each of the sub-regions is different from each other. The variability of life expectancy is not consistent even within the regions that exhibit non homogeneity, i.e., the life expectancies of men and women are different in different regions. Almost all sub regions of Oceania, Africa are Europe heterogeneous within the sub regions. Meanwhile, some of the sub regions of Asia, America are homogeneous within the sub region but heterogeneous within the regions. The sub regions like South-Eastern Asia, South Central Asia, South America, Central America and Caribbean are homogeneous within the sub region.

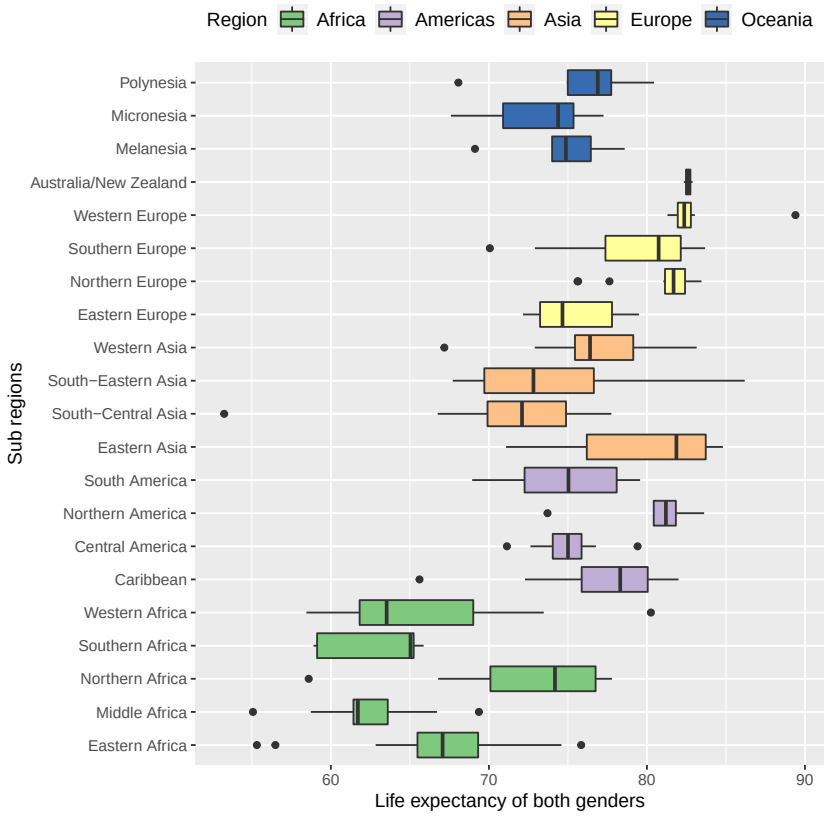


Figure 3: Box plots for life expectancy of both sexes in all sub-regions.

By the measures of central tendency, statistical summary of all the 5 regions is presented in the following Table 3: Africa region has the lowest mean and median values of 55.07 and 65.48 respectively. Contradictorily, the European region has the high mean and median values 70.06 and 81.30 respectively, for the life expectancies of male and females.

	Region	min	Q_1	median	Q_3	max
1	Africa	55.07	61.81	65.48	69.34	80.25
2	Americas	65.61	74.95	77.66	79.43	83.62
3	Asia	53.25	71.64	75.50	78.24	86.19
4	Europe	70.06	76.97	81.30	82.35	89.40
5	Oceania	67.59	74.17	75.06	77.29	82.89

Table 3: Box plot summary for the variable life expectancy of both sexes in all the 5 regions

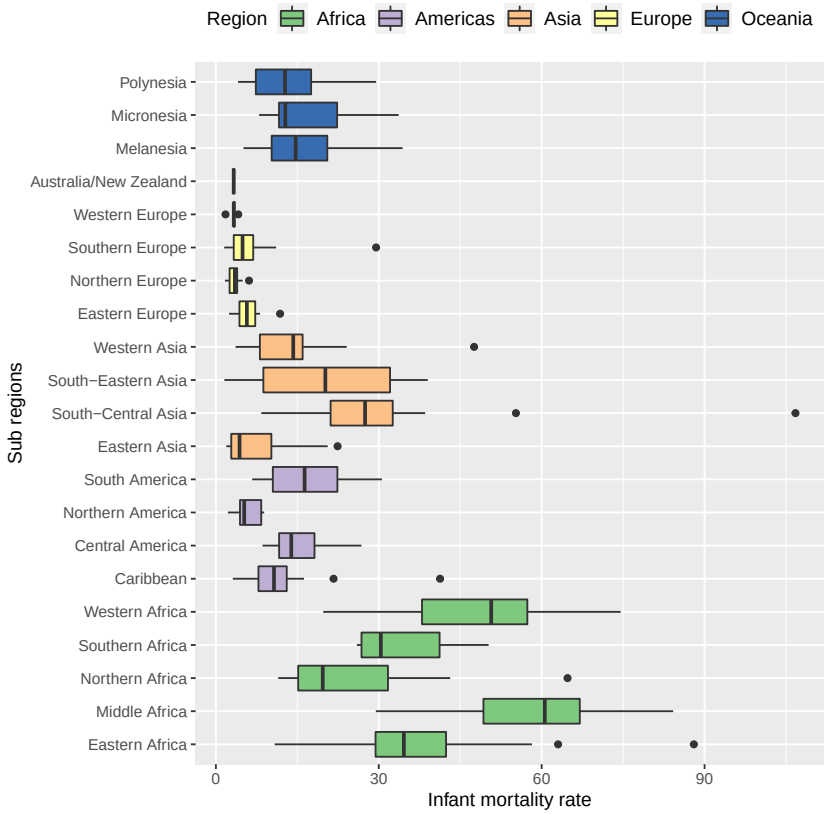


Figure 4: Box plots for infant mortality rate in all sub-regions.

Figure 4 shows the variability of the infant mortality rate across all 21 sub-regions. The interquartile range for sub-regions within a region is roughly consistent. However, the

sub-regions in Africa are not consistent. The variability of the infant mortality rate for both the sexes is not homogeneous within the regions but within some sub regions like Melanesia, South Eastern Asia, South America, and Caribbean it is homogeneous. It also shows that the mortality rate in the Africa is heterogeneous within regions as well as sub regions. The statistical five number summary for all 5 regions is added in the Table 4. The mortality rate of the Africa region is the highest with the mean value 10.83 and median value 42.39. Meanwhile, the European region has the lowest death rate with a mean value of 1.53.

The statistical measures of central tendency are also calculated for life expectancy of both sexes and mortality rate for each sub region. It is given in the Table 6 and Table 7 respectively on Appendix pages 17 and 18.

	Region	min	Q_1	median	Q_3	max
1	Africa	10.83	29.43	42.39	57.30	88.03
2	Americas	2.21	8.21	11.49	15.21	41.29
3	Asia	1.56	7.50	15.64	26.30	106.75
4	Europe	1.53	3.24	3.64	5.32	29.51
5	Oceania	3.05	7.96	12.73	20.52	34.40

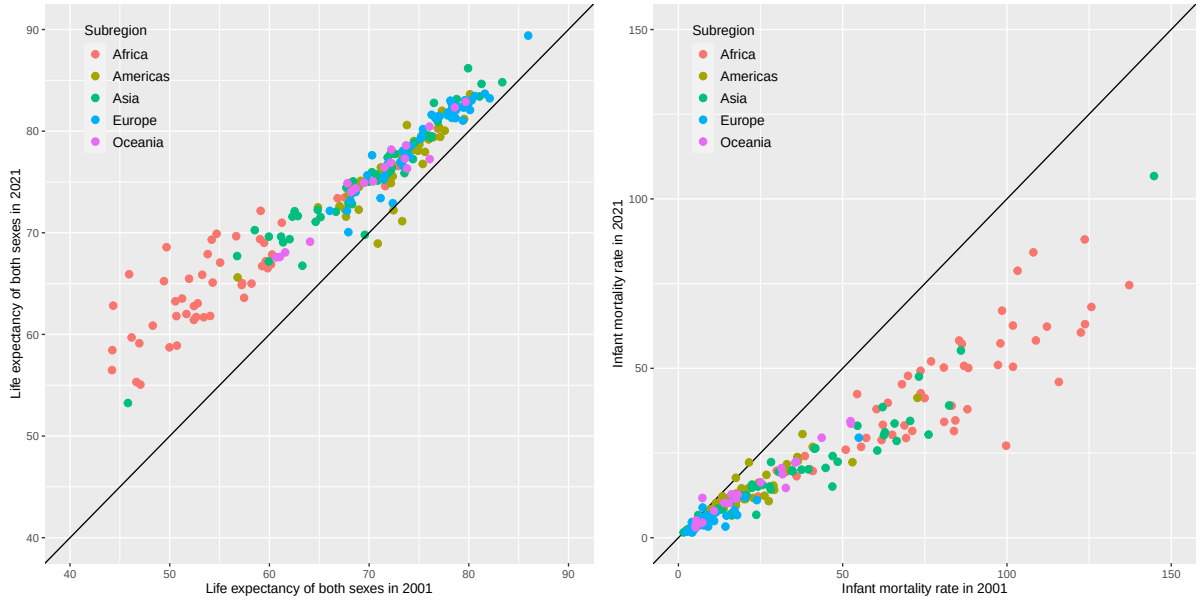
Table 4: Box plot summary for the variable infant mortality rate in all the 5 regions

4.4 Comparison of Data from years 2001 and 2021

As shown in Figure 5(a), the average life expectancy for both males and females has increased in the year 2021 when compared to 2001. The scatter plot shows that over the last 20 years there has been a significant change in the life expectancies of both the sexes. The European region has the highest life expectancy, which is around 85.00 and the African region has low life expectancy comparatively. The diagonal line in the Figure 5(a) represents that there is a high positive linear correlation between the variables and constant improvements in the life expectancies of both sexes in the last 2 decades.

The Figure 5(b) imitates the comparison of infant mortality rate in the year 2001 and 2021. It is observed that there is a remarkable decrease in the mortality rate over the last 20 years. It is also evident that the life expectancy has increased as the infant mortality rate decreased, showing a strong negative correlation. The mortality rate for Europe and Oceania region has almost reached to 0.00 followed by other regions such as America,

Asia and Africa, respectively. All the points that are below, under the diagonal lines in the scatter plot, indicates the inverse relationship of life expectancy of both sexes with infant mortality rate over 2 decades. By the measures of central tendency, summary of the year 2021 and 2001 is presented in the Table 2 and Table 5, on page 16 of the Appendix respectively.



(a) Life expectancies of both the sexes

(b) Infant mortality rate

Figure 5: Scatter plot for the comparison of variables life expectancy of both sexes and infant mortality rate in 2001 and 2021

5 Summary

In this report, the demographic data collected in 227 countries in the years 2001 and 2021 by the U.S. Census was analyzed. A small extract of the collected data was provided by the lecturers for the analysis. A detailed overview of the data set and the objectives of the analysis were given in the introduction section. The statistical methods used to analyze the data set were explained in detail with statistical formulas and graphical representations in section 3.

At first, the frequency distributions of the variables, life expectancies of males and females and infant mortality rate for both sexes were discussed in section 4. The distributions were skewed as shown in the histograms. From the plots, it was observed

that, females have higher life expectancy (on an average 76 years) than males who has 71 years as life expectancy. Then, the relationship between the life expectancy of men and women and the infant mortality rate was found to be a monotonic, negative linear relationship. A decrease in the mortality rate tends to improve both men's and women's life expectancies. In the third task, the variability of the variables life expectancy and infant mortality rate between all the sub-regions was examined. The overall variability of the infant mortality rate was approximately not the same between the regions but same within their sub-regions, except in the African region. However, the life expectancies' were different between the regions, including the sub-regions, as it was inferred from the box plot.

Finally, the change in the variables, life expectancy and infant mortality rate was found to be increased and decreased, respectively, over the span of 20 years. From the plots, it was concluded that over the last 2 decades, there was a significant change in the life expectancies of both the sexes as compared to the mortality rate. The rate of life span was notably increased and improved in this last 20 years.

The variables given in the data set were not sufficient to conclude that the reason of increase in the life expectancies in both sexes. Further study including more information like food habits and lifestyles of people would help in better understanding of social and economic status of a country which can be helpful in the country's growth around the globe.

Bibliography

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

Dalgaard, P. (2008), *Introductory Statistics with R*, Springer. ISBN 978-0-387-79053-4.

Fernando Perez, B. G. (2015), *Jupyter: Interactive data science and scientific computing in all programming languages URL: https://jupyter.org*. version 6.3.0.

Hay-Jahans, C. (2019), *R Companion to Elementary Applied Statistics*, CRC Press. ISBN 978-1-138-32916-4.

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

Schloerke, B. (2020), *GGally: Extension to ggplot2*. R package version 2.1.2.

Sheldon, M. (2010), *Introductory Statistics*, Elsevier. ISBN 978-0-12-374388-6.

U.S Census, B. (2022), *International Database URL: https://www.census.gov/programs-surveys/international-program/about/idb.html (visited on 10th November 2022)*.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag. R package version 3.4.0.

Wickham, H. (2022), *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.

Appendix

A Additional Tables

	LE_male	LE_female	LE_both	LE_diff_btw_sexes	Mortality_rate
Min.	43.06	44.78	44.21	-4.47	1.57
1st Qu.	59.64	63.63	62.03	3.21	9.04
Median	68.78	74.06	71.54	4.53	20.74
Mean	66.23	70.97	68.53	4.74	34.98
3rd Qu.	73.27	78.29	75.61	6.00	54.91
Max.	82.82	89.40	85.96	13.23	144.77
Std	9.17	10.14	9.57	2.34	33.20

Table 5: Measures of central Tendency and spread in the year 2001 (Std - standard deviation, IMR - infant mortality Rate, LE - life expectancy at birth)

	Subregion	min	Q_1	median	Q_3	max
1	Eastern Africa	55.32	65.48	67.07	69.32	75.84
2	Middle Africa	55.07	61.43	61.71	63.60	69.37
3	Northern Africa	58.60	70.09	74.18	76.75	77.79
4	Southern Africa	58.90	59.13	65.04	65.24	65.87
5	Western Africa	58.45	61.82	63.53	69.01	80.25
6	Caribbean	65.61	75.87	78.31	80.05	82.00
7	Central America	71.14	74.04	75.00	75.86	79.41
8	Northern America	73.71	80.43	81.20	81.83	83.62
9	South America	68.94	72.26	75.03	78.08	79.57
10	Eastern Asia	71.08	76.19	81.87	83.72	84.81
11	South-Central Asia	53.25	69.91	72.09	74.88	77.75
12	South-Eastern Asia	67.71	69.71	72.82	76.64	86.19
13	Western Asia	67.18	75.44	76.40	79.14	83.15
14	Eastern Europe	72.16	73.24	74.66	77.79	79.50
15	Northern Europe	75.61	81.14	81.69	82.41	83.45
16	Southern Europe	70.06	77.38	80.74	82.14	83.68
17	Western Europe	81.30	81.95	82.36	82.78	89.40
18	Australia/New Zealand	82.33	82.47	82.61	82.75	82.89
19	Melanesia	69.12	74.00	74.87	76.45	78.59
20	Micronesia	67.59	70.90	74.38	75.35	77.25
21	Polynesia	68.07	74.99	76.89	77.74	80.45

Table 6: Box plot summary for the variable life expectancy of both sexes in all the 21 sub-regions

	Subregion	min	Q_1	median	Q_3	max
1	Eastern Africa	10.83	29.41	34.62	42.39	88.03
2	Middle Africa	29.45	49.28	60.58	67.02	84.22
3	Northern Africa	11.48	15.15	19.68	31.69	64.77
4	Southern Africa	25.97	26.82	30.38	41.20	50.23
5	Western Africa	19.77	37.96	50.71	57.36	74.55
6	Caribbean	3.11	7.84	10.70	13.05	41.29
7	Central America	8.59	11.65	13.89	18.16	26.81
8	Northern America	2.21	4.44	5.22	8.35	8.90
9	South America	6.68	10.49	16.34	22.38	30.58
10	Eastern Asia	1.92	2.83	4.36	10.22	22.42
11	South-Central Asia	8.36	21.12	27.48	32.57	106.75
12	South-Eastern Asia	1.56	8.74	20.16	32.08	39.03
13	Western Asia	3.62	8.11	14.25	15.97	47.56
14	Eastern Europe	2.42	4.34	5.71	7.25	11.83
15	Northern Europe	1.66	2.53	3.50	3.92	6.11
16	Southern Europe	1.53	3.29	4.91	6.88	29.51
17	Western Europe	1.78	3.24	3.29	3.45	4.12
18	Australia/New Zealand	3.05	3.16	3.27	3.39	3.50
19	Melanesia	5.08	10.27	14.69	20.52	34.40
20	Micronesia	7.96	11.62	12.79	22.32	33.66
21	Polynesia	4.06	7.36	12.73	17.54	29.52

Table 7: Box plot summary for the variable infant mortality rate in all the 21 sub-regions