TU Dortmund

Introductory Case Studies

Project 1: Descriptive Analysis of Demographic Data

Lecturers:
Dr. rer. nat. Maximilian Wechsung
M. Sc. Hendrik Dohme

Author: Akshay Choudhary

Group number: 16

Group members: Alina Imtiaz, Vaibhav Grover, Vipul Chauhan, Nidhi Kiritbhai Patel

May 6, 2022

Contents

1 Introduction **1**

2 Problem statement **1**

2.1 Dataset Description . . . 1

2.2 Project Objectives . . . 2

3 Statistical methods **2**

3.1 Measures of Central Tendency . . . 2

3.1.1 Arithmetic Mean . . . 2

3.1.2 Median . . . 3

3.1.3 Mode . . . 3

3.2 Measures of Variability . . . 3

3.2.1 Range . . . 4

3.2.2 Inter Quartile range . . . 4

3.2.3 Skewness . . . 4

3.2.4 Standard Deviation . . . 4

3.2.5 Variance . . . 5

3.3 Visualization Methods . . . 5

3.3.1 Histogram . . . 5

3.3.2 Box Plot . . . 6

3.3.3 Scatter Plot . . . 7

3.4 Correlation . . . 8

4 Statistical analysis **8**

4.1 Frequency Distribution of Variables . . . 8

4.2 Bivariate Correlation between Variables . . . 10

4.3 Homogeneity and Heterogeneity in Subregions . . . 11

4.4 Data Comparison between 2002 and 2022 . . . 13

5 Summary **15**

Bibliography **17**

Appendix **18**

1 Introduction

The average number of years a newborn is expected to live is life expectancy. It is fascinating because it explains some of the characteristics of the country, such as the living conditions of the country, the health of its citizens, and so on. It is fascinating to see how life expectancy has evolved over time for each country in the world because it explains a lot of information about the countries. Different countries have different lifestyles depending on their demographics, which affects life expectancy and the impact of gender on life expectancy from one region to another.

In this study, sample data from 228 countries between 2002 and 2022 are used to better understand the distribution of life expectancy and average fertility rates across world regions. The data from these 228 countries are organized into five regions and 21 subregions and include observations and variables such as life expectancy for both sexes, men and women, and the average fertility rate. This report addresses descriptive and visualization methods in this context.

The data set is described in detail in Section 2. The objectives of the report are also briefly mentioned. Section 3 explains the descriptive approaches, including mean, median, and interquartile range, as well as visualization techniques. Section 3 also describes scatter plots, box plots, and histograms. Section 4 explains the descriptive analysis of the data set using the methods presented in Section 3. Finally, the last section of the report discusses the overall summary and possible future discussions.

2 Problem statement

2.1 Dataset Description

The data set used in this study was compiled by the United States Census Bureau using census (IDB), survey, and administrative techniques each year from 1950 to 2060. The dataset contains demographic statistics for the years 2002 and 2022. The dataset contains 454 observations, with 6 observations missing from 2002. The missing observations are not included in the analysis in this report. Libya, Puerto Rico, South Sudan, Sudan, Syria, and the United States are among the countries for which 2002 data are not available. The year has a binary scale, while the region, subregion, and country have a nominal scale. The fertility rate and life expectancy for both sexes, males and females, are scaled numerically-rationally for each country.

The region is the group of countries that belong to the same demographic area. The subregion is the part of a larger region. Country is the name of the country to which the record belongs. Total fertility rate is defined as the average number of children a woman has. Life expectancy for both sexes is defined as the average number of years a person expects to live. Similarly, life expectancy for men and women is the average number of years a man expects to live and the average number of years a woman expects to live.

2.2 Project Objectives

In this study, the observations include region, subregion, and country, while the variables include total fertility and life expectancy for both sexes, males, and females. The primary objective is to understand the frequency distributions of each variable. The second step is to find out how these random variables are related. The third step is to look for and assess variables variability within and across subregions. The next step is to examine how the data for all variables have changed between 2002 and 2022. For the first three objectives, only data from 2022 are used, while for the last objective, data from both years are considered.

3 Statistical methods

3.1 Measures of Central Tendency

The central tendency estimates a single value that represents the entire set of data. Common measures of central tendency include mean, median, and mode.

3.1.1 Arithmetic Mean

This method gives the center of the dataset by the sum of values present in the dataset divided by the count of values in the dataset.

The formula to calculate the mean of n numbers:

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

n is the total number of observations in the dataset, x_i is the individual observation $x_1, x_2, ...x_n$ and $n \in N$. The mean is affected by the extreme observations as it considers all the observations in the dataset(Hay-Jahans, 2018).

3.1.2 Median

Median represents the exact midpoint of a data set, separating it into two parts. One is the upper half and another one is the lower half. For this, numbers must be in sorted order, either ascending or descending. The value found by the median formula is the exact middle value of the dataset. Thus, there are two ways to find the median of the dataset depending on the number of observations.

Let us assume the induvial observation $x_1, x_2, ...x_n$ where $x_{(1)} \leq x_{(2)} \leq ...x_{(n)}$ as the dataset in sorted order and $n \in N$ then the formula to find median is:

$$x_{med} = x_{(\frac{n+1}{2})}, \text{ if } n \text{ is odd}$$

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}), \text{ if } n \text{ is even.}$$

The median is not distorted by extreme observations because it is considered only the middle observations of the dataset(Hay-Jahans, 2018).

3.1.3 Mode

The mode of a dataset represents the value that occurs most frequently. Some values in a dataset can occur more than once in a count, which means a dataset can have multiple mode values. When all the values in a dataset occur exactly once, there is no mode in the dataset(Hay-Jahans, 2018).

3.2 Measures of Variability

Variability of the dataset estimates the dispersion of the dataset. To understand the variability of a dataset, Range, Interquartile Range, Standard Deviation, Variance, and Skewness can be used. The similarity and dissimilarity of the variables can be explained by the variability of the dataset.

3.2.1 Range

The range is measured by subtracting the maximum value from the minimum value in the dataset. It gives the idea of how much variation is in the dataset. The range can only be calculated for numerical data points(Hay-Jahans, 2018).

3.2.2 Inter Quartile range

Inter Quartile Range (IQR) represents the middle half of a dataset. A sorted dataset can be represented in quarters Q1, Q2, Q3, Q4 where each quarter represents the 25% of the dataset and calls it a quartile. As mentioned above, IQR takes precisely half of the dataset that means 50% of the dataset, which is calculated by Q3-Q1. These quartiles are arranged in increasing orders of the dataset, which means values in the Q1 quartile are smaller or equal to the Q2 quartile, as this method applies only to the sorted dataset. IQR is not sensitive to extreme values, which is unusual observations in the dataset as it considers only the middle 50% of the dataset(Hay-Jahans, 2018).

3.2.3 Skewness

Skewness measures the symmetry of the distribution. A dataset may have symmetric distribution. In this case, the skewness is zero when mean=median=mode of a dataset. If the dataset has mean>median>mode, then the distribution is skewed to the right. If the dataset has mean<median<mode, then the distribution is skewed to the left. The more the differences between the mean, median and mode of the dataset the more the skewness in the distribution(Hay-Jahans, 2018).

3.2.4 Standard Deviation

This method measures the dispersion of each data point in the dataset from the arithmetic mean of the dataset. The standard deviation of the dataset is low if data points are closed together and high if data points are spread out more. The unit of this method is the same as the mean.

Let us assume the individual observation $x_1, x_2, ...x_n$ where $n \in N$ then the formula of standard deviation is:

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

In this $\bar{x}$ is the mean of the dataset and x_i is the individual observation(Hay-Jahans, 2018; Lane, 2003)

3.2.5 Variance

The variance is similar to the standard deviation, which measures the variability of the data based on the mean of the data set. However, the difference is that it is the square of the standard deviation value. So, there is no need to calculate it separately. Calculating any one of them is sufficient to find the other one. It is the squared unit of the mean. It is represented σ^2(Hay-Jahans, 2018).

3.3 Visualization Methods

3.3.1 Histogram

This visualization technique represents the variable values along the x-axis as bars, where the bars have the range of these values. Y-axis represent the frequency of these values. There are two ways to represent the frequency. One is the range count, and the other is the probability of the range within the data set; this is called the relative frequency.

In this project count of the value is used to represent the frequency. As shown in Figure 1 the x-axis shows the range and, on that range, the y-axis shows the frequency with respect to that range. There is no perfect rule to define the size of the bar and the number of bars used in the histogram, and it depends on the variable values or data points used to generate the histogram. There are different forms of the distribution of data represented by a histogram. Unimodal, when there is only one peak in the distribution. Bimodal, two peaks. Uniform model, all bar size with same height and range. Additionally, skewness is also determined by the histogram of the dataset(Lane, 2003).

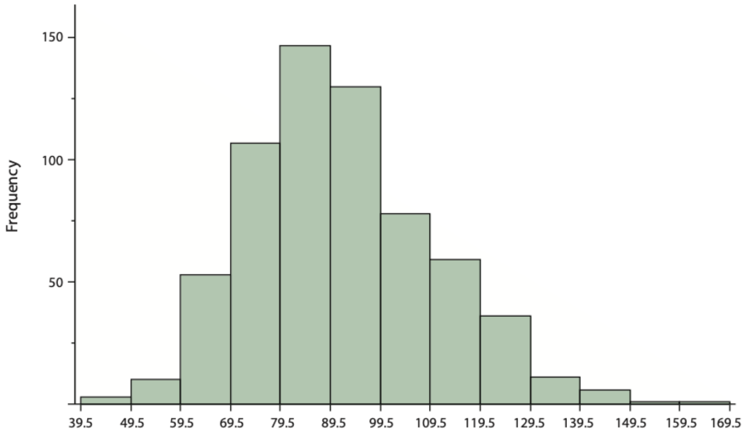

Figure 1: Sample example of Histogram(Lane, 2003)

3.3.2 Box Plot

Apart from the central tendency method, the variability of a dataset also contains a lot of information. To visualize this dispersion in a dataset, a box plot is used usually. Box plots use the IQR method on the dataset and plot the distribution of data into boxes.

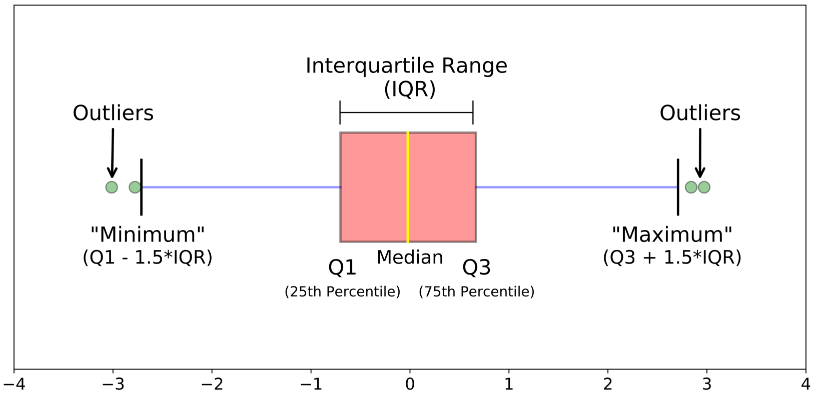

Figure 2: Sample example of Box Plot(Galarnyk, 2018)

As shown in Figure 2 first one is the minimum score which is calculated as Q1-1.5*IQR. The second one is Q1/25th Percentile. The third is the median which is 50% of the dataset. The fourth one is Q3/75th percentile. The last one is a maximum score which is calculated as Q3+1.5*IQR. Additionally, some other points in the figure are beyond the minimum and maximum scores; they are called extreme values. An extreme value is an unusual observation in the dataset, which can be very high or very low from the rest of the values in the dataset, which affects the standard deviation and variance of the dataset(Lane, 2003).

3.3.3 Scatter Plot

To display the bivariate relationship of two random variables, a random variable is mapping of a random experiment to a random number, for which scatter plot is used. It gives information about the relationship between two numeric, random variables. This plot can be modified by adding the third categorical or numeric, random variable. There are three possibilities that can be reflected by the random variables, one is positive, the second is negative, and the last one is no linear relation at all. In a positive relationship if one random variable value is rising then other random variable values also rise respectively. If the values of one random variable is increasing respectively the values of other random variable also increasing and never decreasing, this behaviour is called monotonic. Similar for the decreasing values of the random variables. If one random variable value is rising in a negative relationship, then other random variable values decline. In no linear relation, then values of both random variables scattered in random order in the plot. As shown in Figure 3 the diagonal line passes through the origin which is known as the best fit line, expresses the relationship between the variables (Lane, 2003).

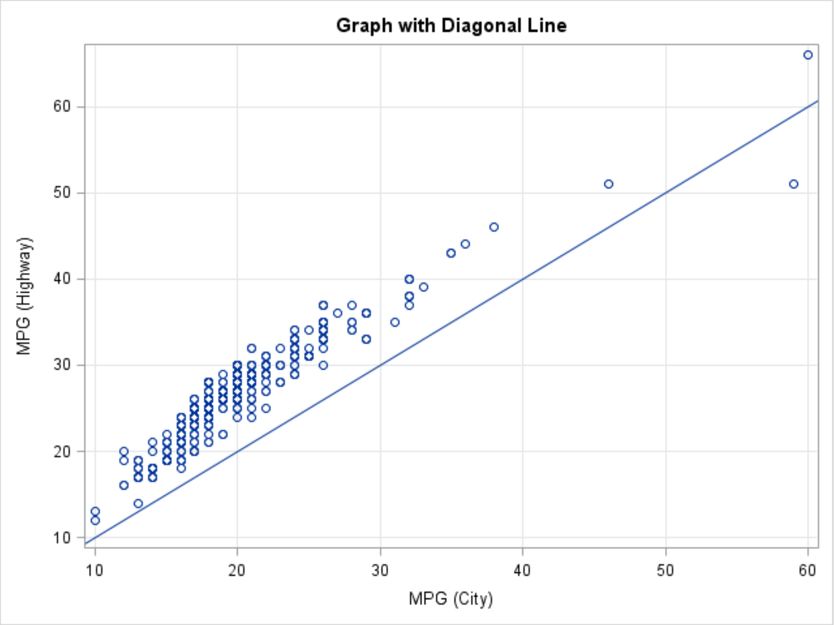

Figure 3: Sample example of Scatter Plot(Wicklin, 2011)

3.4 Correlation

Correlation measures the linear relationship between two random variables. This method is used to find how one variable behaves when another one is changing. Thus, the Pearson correlation coefficient is calculated, which describes the strength and direction of these two random variables. Let X, Y are two random variables. X_i is the individual observation of random variable X and Y_i the individual observation of random variable Y. $\bar{X}$ or $\bar{Y}$ is the mean of the random variable X, Y respectively. σ_X, σ_Y is the standard deviation of X, Y respectively. The formula of the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sigma_X \sigma_Y}.$$

In this project, all variables are quantitative and continuous, and this method works only on continuous (interval or ratio) data points. The range of this coefficient is from +1 to –1, where coefficient value 0 shows no linear relationship between these two random variables. Value 1 shows the perfect negative correlation between the random variables, and +1 shows a perfect positive correlation between the random variables. These two variables can have different scales, and still, the correlation coefficient is calculated(Lane, 2003).

4 Statistical analysis

In this section, The statistical methods explained above are applied to the provided data, and the results of the analysis are interpreted using software (R Development Core Team, 2020).

4.1 Frequency Distribution of Variables

The frequency distribution of any variable explains the count of distinct values of that variable. In this section, to visualize the frequency distribution, histograms of variables are generated.

Figure 4(A) shows that the distribution of life expectancy for males is high between 73-75 years. The mean, median, and mode of the distribution are 72.09, 73.26 and 76.8 years, respectively. Figure 4(B) shows that most of the data is distributed between 75-85

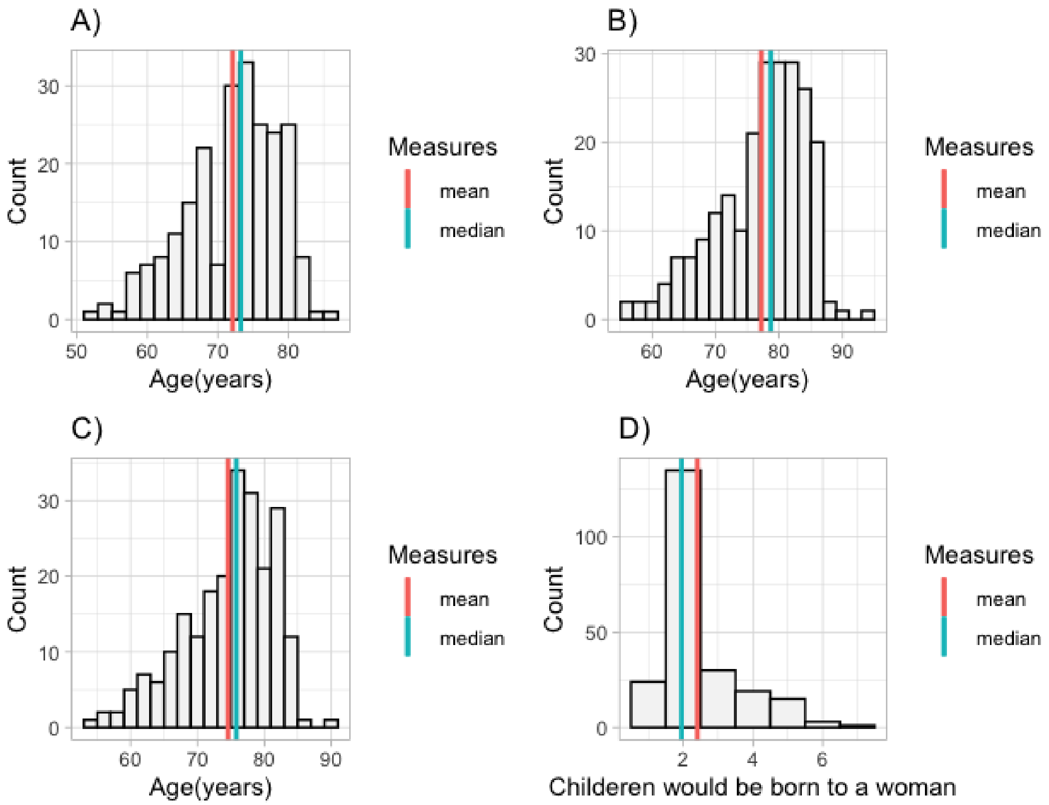

Figure 4: Histograms of frequency distribution of (A)-Average life expectancy for males, (B)-females, (C)-both-sexes and (D)-Average total fertility rate

years. The mean, median, and mode of the distribution are 77.18, 78.69 and 84.79 years. It explains that the life expectancy for females is higher than males. The average year's difference in life expectancy between males and females is 5.09 years and the data is left skewed. The median for the males is 73.26, where for females, it is 78.69, which is 5.43 years higher for females. Most females have the life expectancy 84.79 years, whereas males have 76.8 years, which is lower than females. Figure 4(C) shows the distribution of data for the life expectancy for both sexes, like the distribution of life expectancy for males and females. The mean value is 74.57 years, and the median value is 75.82 years. The range of both sexes is somewhere between the range of males and females mean and median, which is understandable because this variable considers both sexes. Figure 4(D) shows that the distribution is right skewed because the difference between mean and median is 0.45. The mean value 2.40 is very high from all the modes and the median, which means most of the data lies before the mean of the dataset.

4.2 Bivariate Correlation between Variables

In this section, the relationship between two random variables considered. For this, data of two random variables plotted on a scatter plot, and by calculating the Pearson correlation coefficient, the relationship is defined which explains the linear relationship between these two random variables.

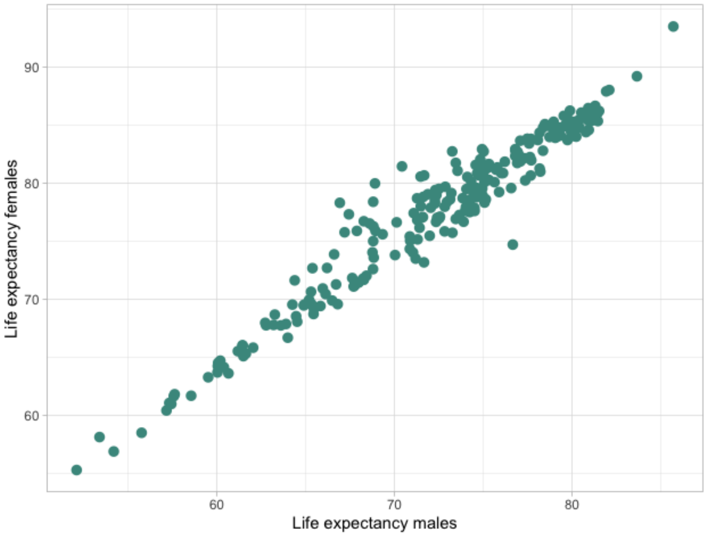

Figure 5: Scatter plot of average life expectancy for males and females

Figure 5 shows that there is a positive correlation, which is strong between the variables life expectancy males and females as the Pearson correlation coefficient value is 0.97. This scatter plot explains that with the years of life expectancy for males increase, the life expectancy for females also increases. Also, it behaves monotone increasing as all the values for both random variable increases.

Figure 6 shows the negative correlation between the life expectancy for both sexes and the total fertility rate which suggests that as the life expectancy increase for both the sexes total fertility rate decrease. The Pearson correlation coefficient value is -0.79. Also, it behaves monotone decreasing as all the values for both random variable decreases. Figure 11 in the Appendix shows that life expectancy for both sexes has high positive correlation with males and females because the Pearson correlation coefficient is 0.99 and 0.99 respectively which is understandable as it includes both the sexes. On the other side Pearson correlation coefficient of males and females with total fertility rate is negative which explains that there is a negative correlation exist.

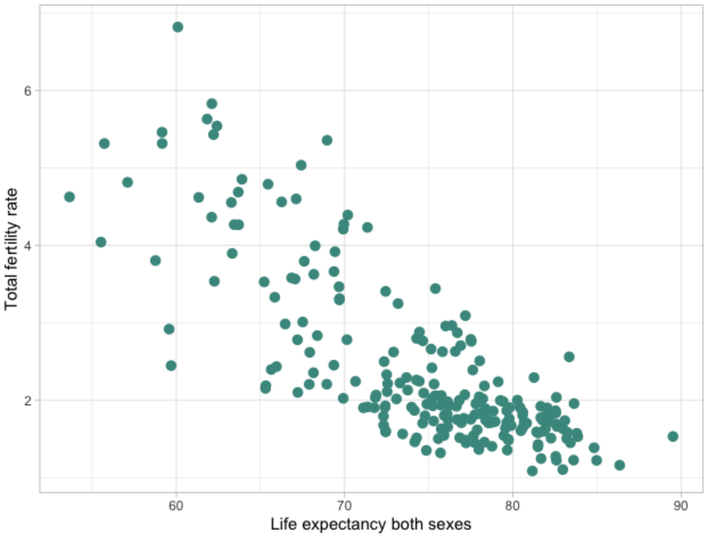

Figure 6: Scatter plot of average life expectancy for both sexes and total fertility rate

4.3 Homogeneity and Heterogeneity in Subregions

Homogeneity within the subregion and heterogeneity between subregions are discussed in this section. If countries from the same subregion have similar data distributions, this is referred to as homogeneity; if countries from different subregions have no similar data distributions, this is referred to as heterogeneity. This box plot is considered to understand.

Table 1 on page 19 in the Appendix present the measure of dispersion in total fertility rate for every subregion. The Africa region has a high standard deviation, indicating that it lacks homogeneity because data from all subregions is distributed across broad areas for the total fertility rate. As Figure 7 shows, Eastern Europe, Australia/New Zealand, and Southern Europe data are distributed in the small interval as their IQR is 0.09, 0.06, and 0.13 respectively. Although the data of subregion Australia/New Zealand is few (only 2 observations), so to get the conclusion out of it from figure is not correct. Western Africa has a median of 4.26, whereas Southern Africa has a median of 2.44, which is different, it shows that these subregions are heterogeneous. Similarly, Eastern Asia data is differently distributed than South-Central Asia, which shows heterogeneity between these subregions, but South-Eastern Asia and South-Central Asia have almost identical median with a difference of 0.04, which shows that these subregions have the same data distribution.

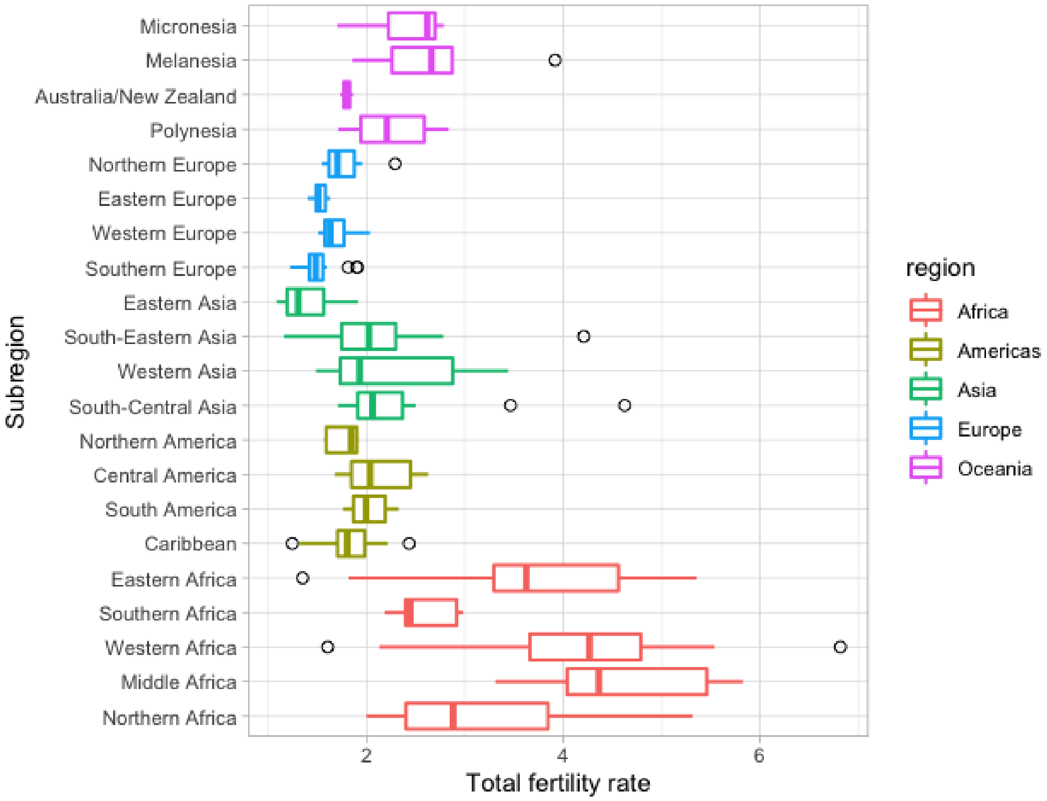

Figure 7: Box Plots of average total fertility rate

Table 2 on page 20 in the Appendix present the measure of dispersion in life expectancy for both sexes for every subregion. Australia/New Zealand shows the homogeneity among other subregions with the lowest standard deviation of 0.27 but as mentioned earlier with few data conclusion for this subregion might not be correct, and Western Europe has IQR 0.82 which shows the homogeneity within the data as well. Figure 8 shows that Northern Africa is not homogeneous because of the 47.18 variances in the data distribution. Eastern Asia and South America have IQR 7.62 and 5.84 respectively, which explains that data is distributed in a large area, so there is no homogeneity in these subregions. Africa region is shows heterogeneity because all the subregions have different median. South-Eastern Asia and South-Central Asia have almost the same mean and median, which shows that these subregions have the same data distribution. As section 4.2 explains, the correlation between the life expectancy for both sexes with males and females is very high because of that interpretation of these two variables will be the same as life expectancy for both sexes.

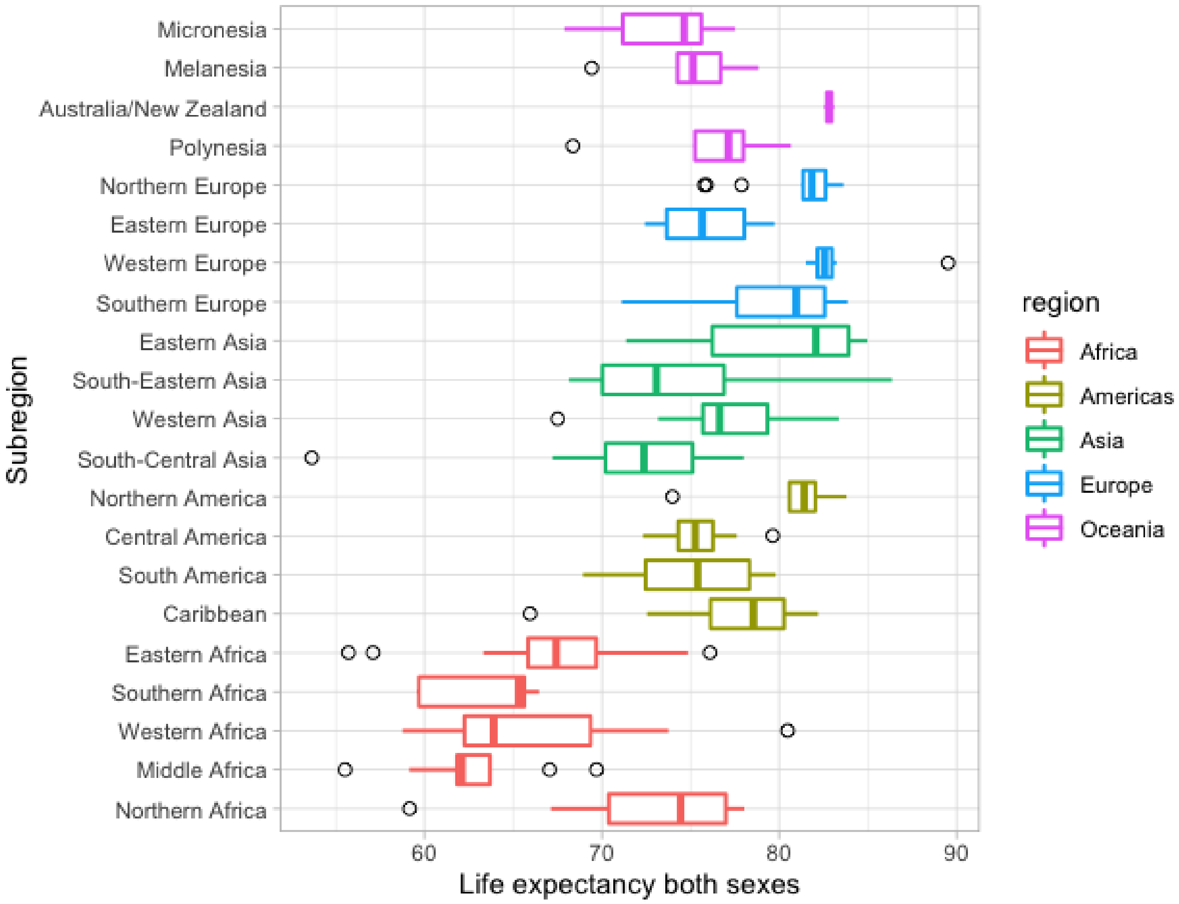

Figure 8: Box plots of average life expectancy for both sexes

4.4 Data Comparison between 2002 and 2022

The datasets from the years 2002 and 2022 is compared in this section. The analysis is carried by using a scatter plot. The variation in the variable data throughout time is examined.

Figure 9 shows that the Africa Region total fertility rate is higher in both the years other then any other region, and country Niger has the highest total fertility rate in the year 2002 as well as in year 2022. The total fertility rate of the Asia region changed in the past 20 years and reduced. Timor-Leste has a total fertility rate of approx.7.5 in the year 2002, but in the year 2022 it changed and become less than 5. The total fertility rate in Europe has not changed much over the years, but America and Oceania have reduced slightly. As the diagonal line passes through the origin and most of the data

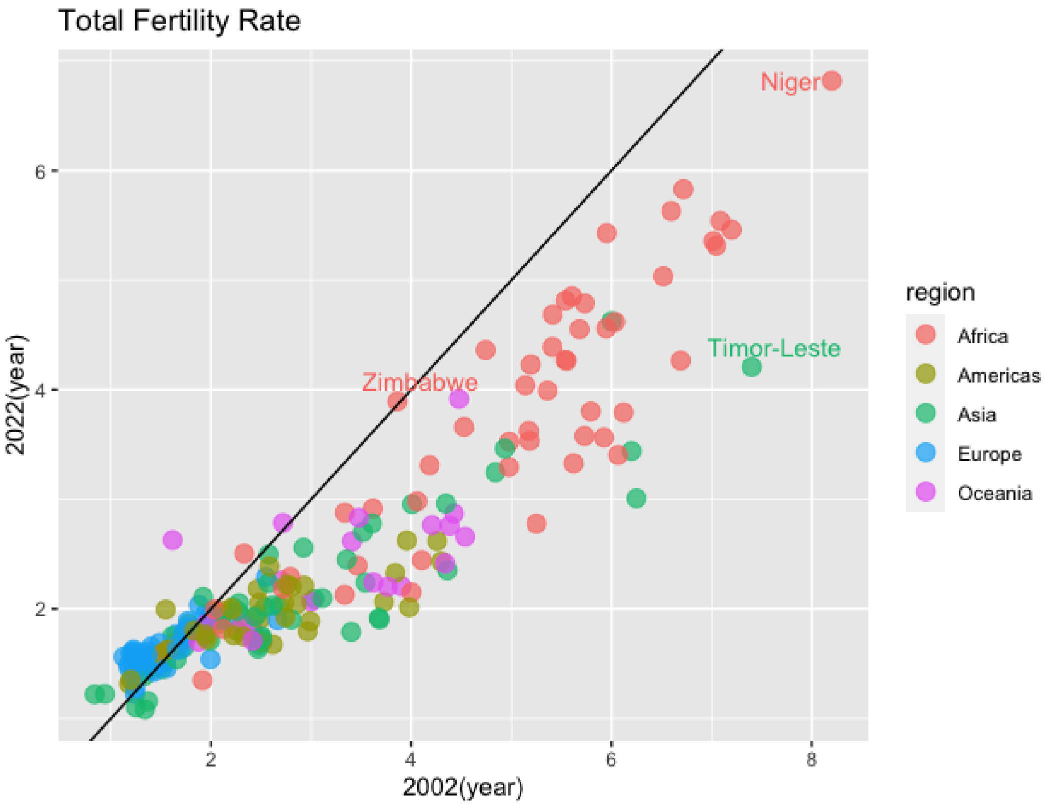

Figure 9: Scatter plot of average total fertility rate

points are below the line, it explains that total fertility rate reduced over the past two decades.

Figure 10 describes that life expectancy for both sexes in the year 2002, the overall life expectancy for both sexes is lower than the year 2022. Africa's region has lower life expectancy of both sexes rather then any other regions but it is increased in past 20 years. As graph shows most of the other regions have more then 60 years of life expectancy for both sexes and it is increase slightly with period of time. Monaco has the highest life expectancy of both sexes in the year 2002 and year 2022 also the change is minimal. Europe and America region does not change much in the past two decades. As the diagonal line passes through the origin and most of the data points are above the line, it explains that Life expectancy of both sexes increased over the past two decades. As the correlation between the life expectancy of both sexes with male and female is very high, the variation in these 2 variables values in past two decades will be similar to the life expectancy of both sexes.

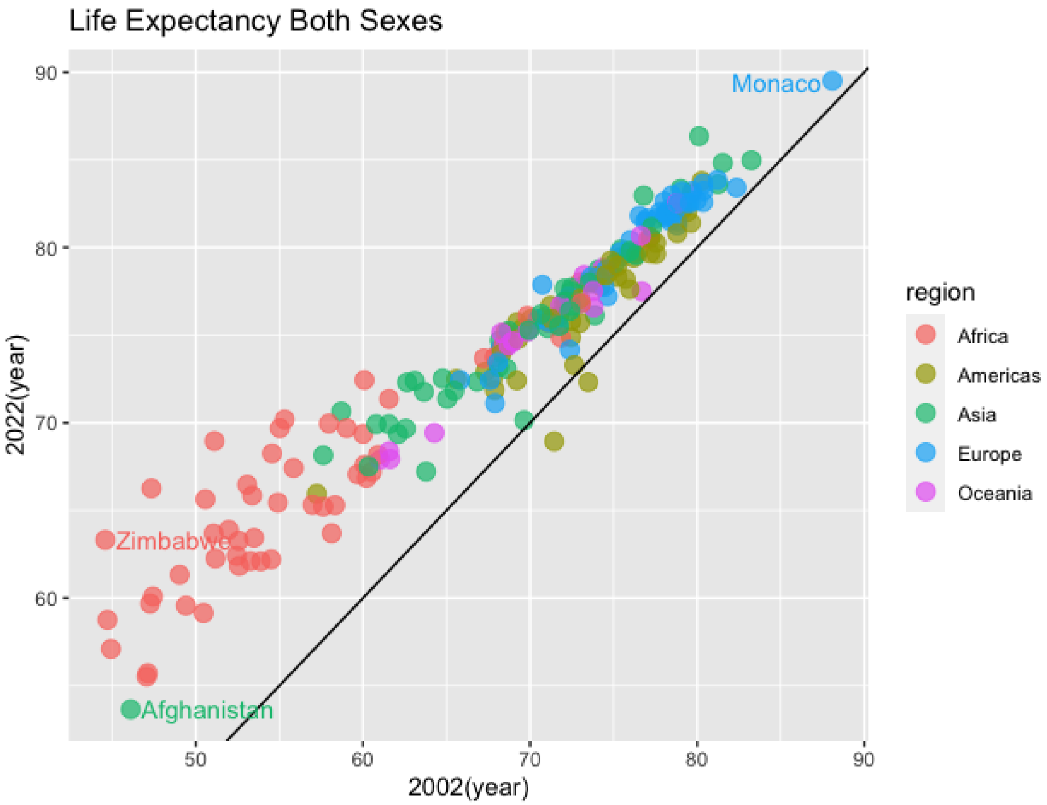

Figure 10: Scatter plot of average life expectancy for both sexes

5 Summary

In this report, the datasets for 2002 and 2022 years are taken from the U.S. Census Bureau's yearly database. In this dataset, total fertility rate, life expectancy for both sexes, males and females, considered variables, region, subregion, and country were considered observations. The dataset contains some missing values that were not taken into account during the study. The frequency distribution of each variable, the bivariate correlation between the variables, variables variability within and between the subregions, and change of the variable data in 20 years are the objectives of this report. Only the dataset for the year 2022 was evaluated for the first three objectives.

Life expectancy for females has a higher mean and median than Life expectancy for males, representing that female live more than males. Total fertility rate data is right-skewed, which represents that the average number of children born to a woman is not more than 2. There is a positive correlation between the life expectancy for males and females as the correlation coefficient is 0.97, and it behaves as monotone increasing.

Whereas correlation between the life expectancy for both sexes and the total fertility rate is negative, which is explained by the coefficient value of -0.79, and it behaves as monotone decreasing. As the correlation between the life expectancy for both sexes, males, and females, is positive, these two variables also negatively correlate with the total fertility rate. Western Europe subregion shows homogeneity in all variables, whereas the Africa region does not show any homogeneity in all variables. There is heterogeneity in subregions of Africa because they all have different medians. The total fertility rate has been reduced in the last 20 years because in the year 2002 average total fertility rate was 3.00, whereas, in the year 2022, it is reduced to 2.38.

On the other hand, life expectancy for both sexes, males and females, increased over the past two decades. Interestingly, Monaco has the highest life expectancy in these two years, but because we do not have the year's data, it is not sure that this country has the highest life expectancy in all the years. In conclusion, with this analysis, it is considered that life expectancy for males and females is increased, and the total fertility rate decreased.

Suppose this dataset is used for further investigation. It is interesting to find out why the life expectancy for females is higher than for males and why there are so many distribution differences in the African region rather than in other regions.

Bibliography

URL https://www.census.gov/ programs-surveys/international-programs/about/glossary.ht

Michael Galarnyk. Understanding boxplots. *towardsDataScience*, 2018. URL https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics.* [CRC Press], 2018.

David Lane. *Introduction to Statistics.* [Rice University], 2003.

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

Rick Wicklin. Scatter plots. *Sas Blogs*, 2011. URL https://blogs.sas.com/content/iml/2011/07/27/add-a-diagonal-line-to-a-scatter-plot-

Appendix

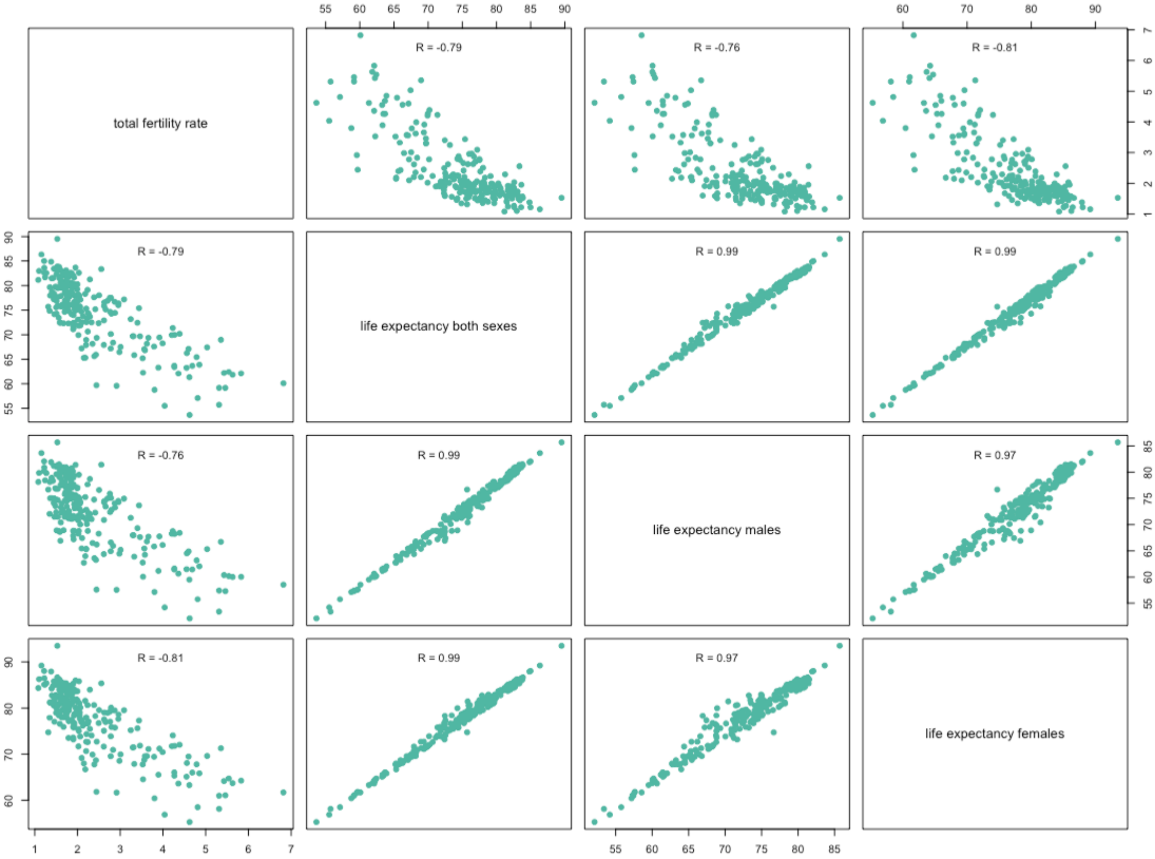

Figure 11: Scatter plot of correlation matrix of all variables

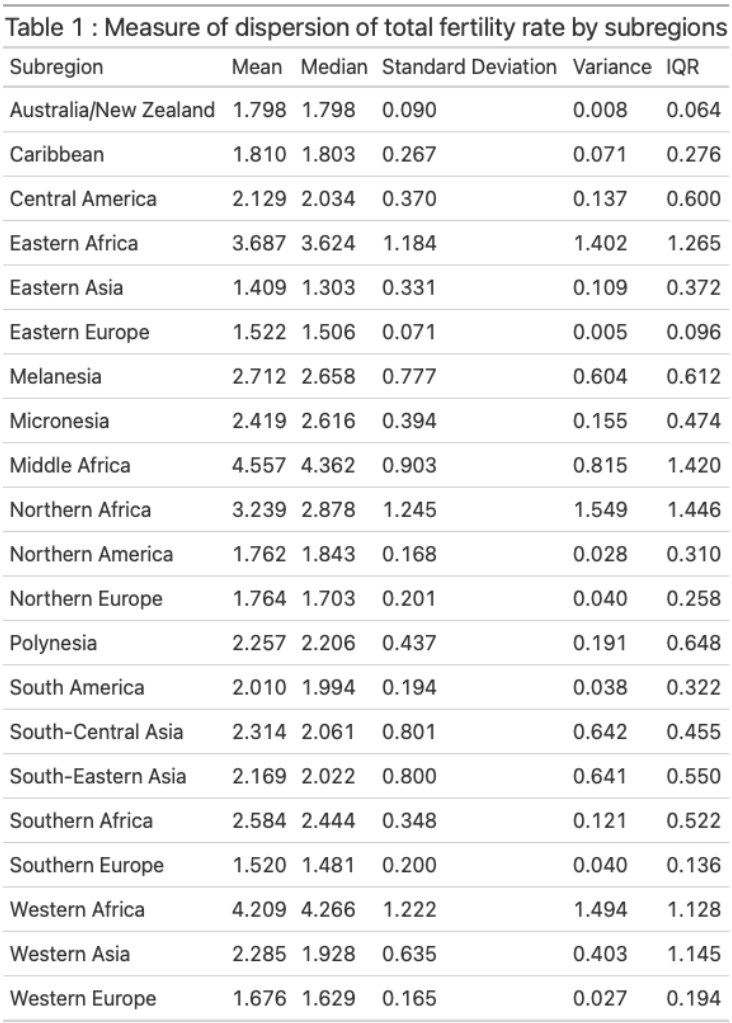

Table 1 : Measure of dispersion of total fertility rate by subregions

Subregion	Mean	Median	Standard Deviation	Variance	IQR
Australia/New Zealand	1.798	1.798	0.090	0.008	0.064
Caribbean	1.810	1.803	0.267	0.071	0.276
Central America	2.129	2.034	0.370	0.137	0.600
Eastern Africa	3.687	3.624	1.184	1.402	1.265
Eastern Asia	1.409	1.303	0.331	0.109	0.372
Eastern Europe	1.522	1.506	0.071	0.005	0.096
Melanesia	2.712	2.658	0.777	0.604	0.612
Micronesia	2.419	2.616	0.394	0.155	0.474
Middle Africa	4.557	4.362	0.903	0.815	1.420
Northern Africa	3.239	2.878	1.245	1.549	1.446
Northern America	1.762	1.843	0.168	0.028	0.310
Northern Europe	1.764	1.703	0.201	0.040	0.258
Polynesia	2.257	2.206	0.437	0.191	0.648
South America	2.010	1.994	0.194	0.038	0.322
South-Central Asia	2.314	2.061	0.801	0.642	0.455
South-Eastern Asia	2.169	2.022	0.800	0.641	0.550
Southern Africa	2.584	2.444	0.348	0.121	0.522
Southern Europe	1.520	1.481	0.200	0.040	0.136
Western Africa	4.209	4.266	1.222	1.494	1.128
Western Asia	2.285	1.928	0.635	0.403	1.145
Western Europe	1.676	1.629	0.165	0.027	0.194

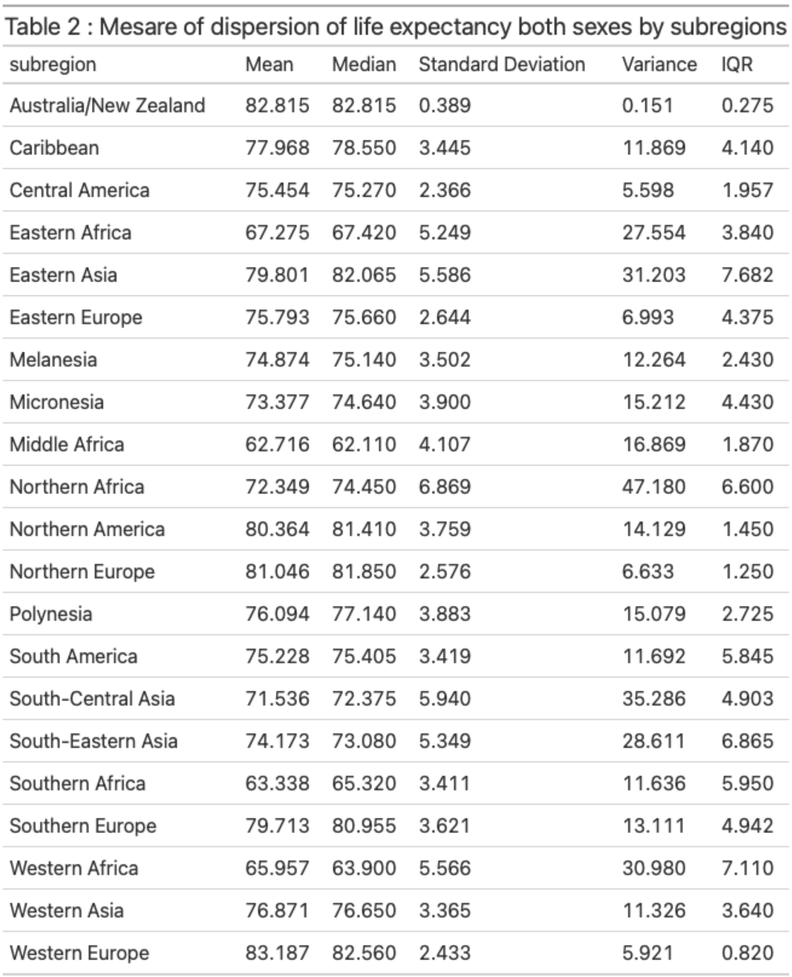

Table 2 : Mesare of dispersion of life expectancy both sexes by subregions

subregion	Mean	Median	Standard Deviation	Variance	IQR
Australia/New Zealand	82.815	82.815	0.389	0.151	0.275
Caribbean	77.968	78.550	3.445	11.869	4.140
Central America	75.454	75.270	2.366	5.598	1.957
Eastern Africa	67.275	67.420	5.249	27.554	3.840
Eastern Asia	79.801	82.065	5.586	31.203	7.682
Eastern Europe	75.793	75.660	2.644	6.993	4.375
Melanesia	74.874	75.140	3.502	12.264	2.430
Micronesia	73.377	74.640	3.900	15.212	4.430
Middle Africa	62.716	62.110	4.107	16.869	1.870
Northern Africa	72.349	74.450	6.869	47.180	6.600
Northern America	80.364	81.410	3.759	14.129	1.450
Northern Europe	81.046	81.850	2.576	6.633	1.250
Polynesia	76.094	77.140	3.883	15.079	2.725
South America	75.228	75.405	3.419	11.692	5.845
South-Central Asia	71.536	72.375	5.940	35.286	4.903
South-Eastern Asia	74.173	73.080	5.349	28.611	6.865
Southern Africa	63.338	65.320	3.411	11.636	5.950
Southern Europe	79.713	80.955	3.621	13.111	4.942
Western Africa	65.957	63.900	5.566	30.980	7.110
Western Asia	76.871	76.650	3.365	11.326	3.640
Western Europe	83.187	82.560	2.433	5.921	0.820