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: Mohammad Sakhawat Hossain

Group number: 11

Group members: Muhammad Mahir Hasan Chowdhury, Aritra Paul, Sadia Mahjabin

November 10, 2022

Contents

1 Introduction

The new oil is the data. In almost every field today, from politics to decision-making, data are an essential component. A huge amount of accurate data can boost any sector's ability to identify the exact problem or focus on appropriate issues. Demographic data (Coale and Trussell, 1996) also plays a vital role in industry, politics, or any business sector (Winkelmann and Zimmermann, 1994).

The data set given here is a small part of data collected by *International Data Base* (IDB, 2022) of the *U.S Census Bureau*. The objective of this report is to analyze the provided variables and present a summary of the result. The report's main conclusions are as follows: females have a longer average lifespan than males; there is a strong correlation between the sexes in terms of life expectancy; heterogeneity is present among the subregions; and the mortality rate versus life expectancy of both sexes changes in an opposite direction for a while. With the aid of various statistical techniques, such as mean, median, and standard deviation, the analysis begins by explaining the frequency distribution. Then, using the correlation coefficient method (Pearson), the correlation between the variables is explicitly evaluated. Later, with the help of the median, interquartile range (IQR), whiskers, and variance, a comparison between the variables within the sub-region of a particular region and all the sub-regions is presented. Last but not least, increases and decreases of the variables have been compared for the last 20 years by using the mean and a visual presentation (scatter plot).

In the second section, an overview of the given data set as well as it's structure is precisely presented. Here, explanation of all the variables and it's characteristics are given in details. Third section is dedicated to the explorative and descriptive statistical methods(arithmetic mean, median, correlation, standard deviation,variance) which was used in our analysis. Based on this methodology, a thorough and logical analysis is conducted in Section 4 using graphical representations (histogram, box plot, and scatter plot). The final section contains a summary of the findings,a discussion, and an outlook on further possible analysis.

2 Problem Statement

The data set is compiled and given by the instructors of the course "Introductory Case Studies" at TU Dortmund in the Winter 2022 session. This data set is a small part of the original set owned by *International data base* (IDB, 2022) with the help of different state organizations, surveys, and records, as well as estimates and projections by the U.S. Census Bureau itself.

The data set contains 8 variables and 454 observations. It consists of 5 *Regions* (character) and 21 *sub regions* (character) with 227 *Countries* (character) from the *Year*(integer) 2001-2021. It has 24 missing values. Each variables *Life.Expectancy..Both.Sexes*, *Infant.Mortality.Rate..Both.Sexes*, *Life.Expectancy..Males*, *Life.Expectancy..Females* has 6 missing values. These four variables are all numerical in nature. Here, infant mortality rate for both sexes can be referred as the number of babies that die within one year of their birth from a group of 1000 live births. (Glossary, 2021a). The life expectancy can be referred as the average lifespan of a group of people who were all born in the same year, assuming that mortality rates at each age don't change in the future. (Glossary, 2021b).

Although the data set has 24 missing values, the quality of the data seems good as it is supervised and preserved by a renowned organization like *International Data Base(IDB)*. The aim of the data analysis is to figure out the potential relationship between the variables in the region or subregion and head toward a conclusive decision. Therefore, in the first step, a descriptive analysis is carried out using methods of central tendency (mean, median), dispersion (variance, and standard deviation), and a graphical method called a histogram. The second step involved using the Pearson correlation method and a scatter plot to identify any potential linear correlation. In the third part, a box plot is used to assess the characteristics of variables within region and sub-regions. A scatter plot is then used to show how the variables have changed over the 20 years.

3 Statistical Methods

Several statistical methods are presented in this section. The software R (version 4.0.5)(R Core Team, 2021) with library patchwork(Thomas Lin, 2021), tidyverse (Wickham and et al., 2019) and ggplot2 (Hadley, 2016) are used for the visualisation as well as calculation. To merge multiple grid based plot gridExtra (Auguie, 2017) has been used. Initially, we will explain the measures of central tendency, e.g., arithmetic mean, and median, then we will move on with the dispersion, e.g., standard deviation and variance. Next, the Pearson method will be explained and finally different graphical representations are interpreted.

3.1 Arithmetic mean:

The arithmetic mean (Hay-Jahans, 2019a, P.74) can be applied when the underlying random variable for a sample is continuous. If $x_1, x_2, ..., x_n$ are numeric sample then arithmetic mean is denoted by $\bar{x}$ and defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where n is the total number of observations.

The mean differs for each sample (or population) and isn't always present in the dataset used to compute it. This measure is best suited for fairly homogeneous data because it is sensitive to extreme values in one of the data's tails.

3.2 Median :

The median (Hay-Jahans, 2019b, P.75) is calculated if the sample $y_1, y_2, ..., y_n$ of a particular variable Y is sorted in ascending order. It is defined as

$$\tilde{y} = \begin{cases} y_{\frac{n+1}{2}} & \text{if n is odd.} \\ \frac{1}{2}\Big(y_{\frac{n}{2}} + y_{(\frac{n}{2}+1)}\Big) & \text{if n is even.} \end{cases}$$

When the data contain extreme values in one of the tails, this measure of central tendency is frequently chosen since it is indifferent to extreme values in a sample.

3.3 Standard deviation(SD) and variance:

To determine how data are spread out from the mean value, we normally use empirical variance or standard deviation (Lee et al., 2015). If $x_1, x_2, ..., x_n$ are sample data then the variance defined as

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

In the formula squared difference helps to get non-negative values where $\bar{x}$ represent mean value of the sample data. The standard deviation is simply $\sqrt{\hat{\sigma}^2}$.

While a high standard deviation suggests that the values are dispersed throughout a larger range, a low standard deviation suggests that the values tend to be near to the mean of the set.

3.4 Pearson correlation coefficient:

A correlation coefficient or Pearson correlation (Benesty et al., 2009) is a statistical measure which is used to determine the linear relationship between two quantitative variables. The linear relationship can be defined when two variables are directly related, meaning that if the value of x changes, y must likewise change in the same manner. We can use scatter plot to ascertain the correlation between the variables. Strength of the relationship can be presented by the correlation coefficient(r). The range of coefficient lies -1 to 1.

If there are two variables X and Y then the samples for these variables be $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$. The Pearson correlation can be denoted by r and defined as

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

If the value of r is between 0 and 1 then it's positively correlated and if less than 1 then negatively correlated.
1 : Exact positive linear correlation i.e when one variable increases another variable also increases in the same direction.

0 : No linear correlation exists i.e there is no linear relationship between the variables
-1 : Negative linear correlation i.e when one variable increases another variable decrease.

3.5 Boxplot:

A Boxplot (Hay-Jahans, 2019c, P.137-142) is a graphical view with five number summary i.e maximum (the highest data point, without outliers, in the data set), minimum (the lowest data point, without outliers, in the data set), 1st quartile(Q1), median and third quartile(Q3). This summary can be used as an early indicator of symmetry violation or the potential existence of extreme values (outliers) in the data. In a box plot, a longer right whisker denotes right-skewed data, while a longer left whisker denotes left-skewed data. The sample is deemed symmetrical if the whiskers are roughly equal in length and the median is roughly in the center of the box. Since median is more robust (not influenced) to extreme values than median, therefore often box plot is used in the statistical measure. The whisker (T shaped line) represent how data are skewed. The box always represent the range of the middle 50 percent of data values which is called interquartile range (IQR).

$IQR = Q_3 - Q_1$

Here, First quartile (Q1), also referred to as the lower quartile (0.25), is the median value between the minimum value and the median (of entire box plot). Similarly, Third quartile (Q3), also referred to as the upper quartile (0.75), is the median value between the maximum value and the median (of entire box plot).

When observations fall 1.5 IQR or more units below the first quartile or rise above the third quartile, they are classified as extreme values (or outliers) in the sample.

3.6 Scatterplot:

When one continuous variable (X) depends on another variable (Y) or when the two continuous variables are independent, a scatter plot (Hay-Jahans, 2019e, P.159-168) can be used. In the scatter plot, points are on the 2-dimensional coordinate axes for ordered pairs (x_i, y_i), where $i = 1, 2, ..., n$ used to define different values along X and Y axis.
The relationship between the variables can be determined by drawing a best fit line. A positive correlation exists between the variables if the pattern of the dots slopes from

lower left to upper right. If the dot pattern slopes downward from upper left to lower right, there is a negative correlation.

3.7 Histogram:

Frequency distribution (a list or table that represents the frequency of the variables) which are grouped can be presented graphically using a histogram (Hay-Jahans, 2019d, P.131-136). The horizontal axis shows the ranges of values (grouped), and the density of the distribution is represented by vertically scaled bars. We can count how many values fall into each interval after dividing the entire range of values into a series of intervals, which are typically referred to as bins. The area of the constructed rectangle is proportional to the number of cases in the bin even if the bins are not of equal width and in the density histogram the area of the entire histogram equals to 1.

4 Statistical analysis

The data set given here has 454 observations and 8 variables. It also includes 24 missing values: 6 missing values for each variable *Infant.Mortality.Rate..Both Sexes*, *Life.Expectancy..Both.Sexes*, *Life.Expectancy..Females* and *Life.Expectancy..Males*. At first, descriptive analysis is carried out over the 4 variables mentioned above for the year 2021, where females have the highest mean life expectancy. Next, bivariate correlation will be identified for the aforementioned 4 variables. We found a positive correlation between female and male expectancies in 2021 and a negative correlation with respect to the mortality rate. In the third step, we will use all the variables except *Country* for the year 2021 and observe the relationship of the variables within *subregions* and outside the *subregions*. Our analysis says there is no homogeneity among the subregions. The last step is to figure out how the variables have changed over the past 20 years. In general, the death rate has gone down and life expectancy has gone up.

4.1 Descriptive analysis:

In this section we will analyse our data with histogram and descriptive method. First, we observe the frequency distribution of *Infant.Mortality.Rate..Both.Sexes* with respect to the density for the year 2021. The figure (1) shows that when the mortality is increased,

the density is decreased and the right-skewed distribution shows that more countries have a lower mortality rate. Here, the mean value of infant mortality of both sexes is 20.25 (rounded to 2 decimal places), where the maximum value is 106.75 and the standard deviation is 19.19. The standard deviation is comparatively large indicating there is a high variation in the mortality rate. Maximum values (106.75) and ranges (105.22) are also almost the same, which also indicates high variation in the data set. The median is almost half of the mean, referring the data distribution is right-skewed and can be seen from the figure as well.

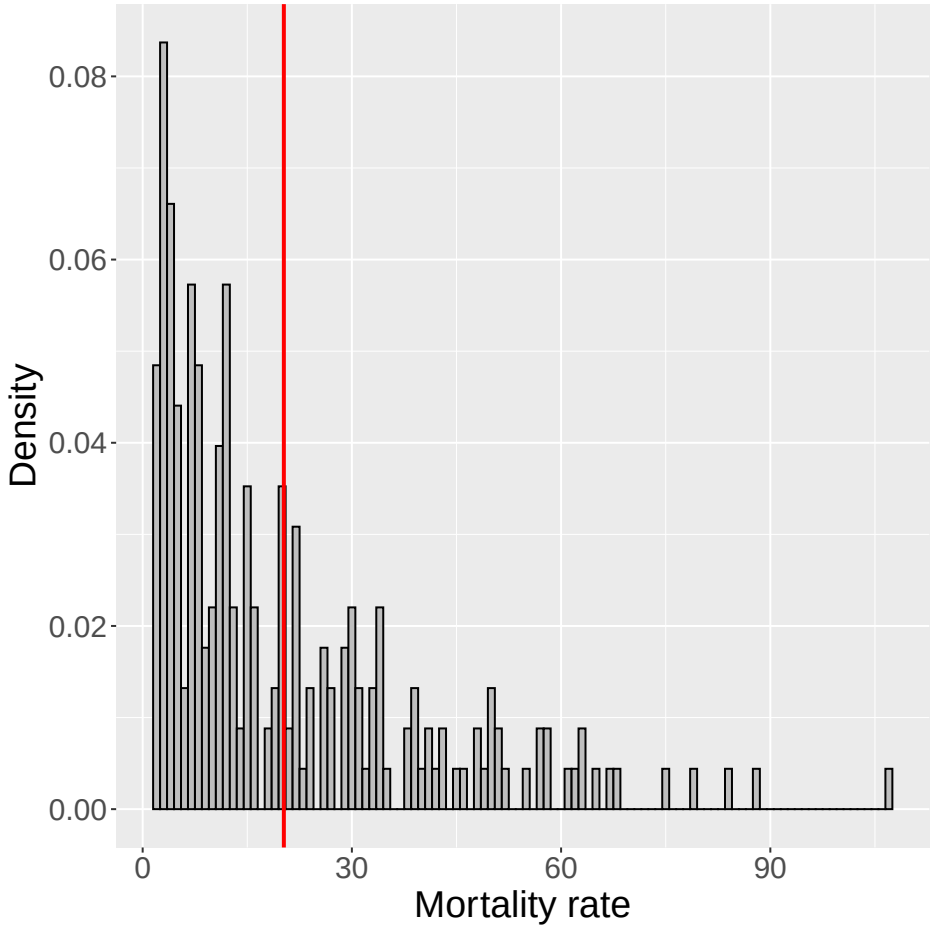

Figure 1: Histogram with respect to mortality rate. Red line indicate mean value.

Therefore, we can draw the conclusion that as more nations exist, the mortality rate is declining. All the values can be found in the table placed in the appendix section (Table 1) on page 18.

According to the data and figure (9, 10) attached on page 19 in the appendix, we observe the average life expectancy for females (76.89) is nearly 5 years higher than the average life expectancy for males (71.78). The median value of female life expectancy is very close to the mean value (1.47 difference) and the standard deviation is much less than the mean value, which indicates the distribution is left-skewed and more countries have a higher life expectancy. Therefore, we can conclude that, female has more life expectancy than male.

In addition, we see that for both sexes (Figure 2), the mean (74.36), median (75.80), and standard deviation (6.92) are usually higher than females but lower than males and overall data are left skewed.

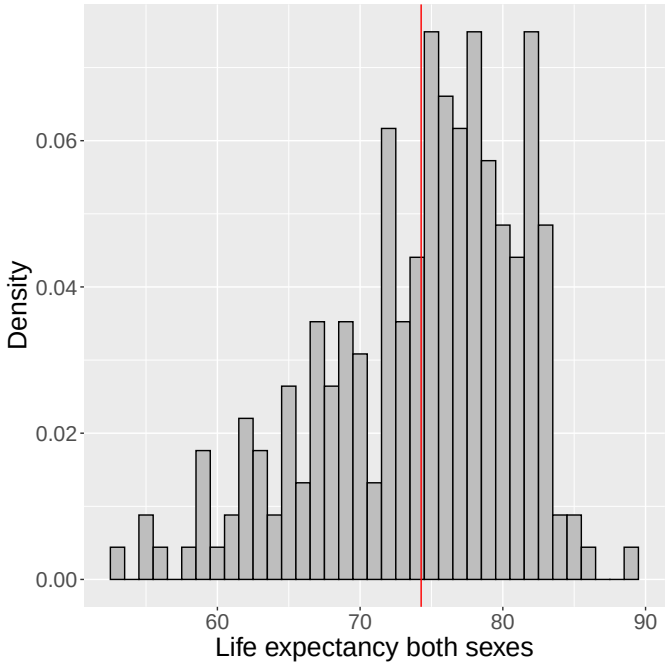

Figure 2: Histogram with respect to *both sexes life expectancy*

4.2 Relationship between the variables:

In this part, we'll try to figure out if any two variables are related to each other or not. Only 4 variables are used for this analysis such as *Infant.Mortality.Rate..Both.Sexes*, *Life.Expectancy..Males* *Life.Expectancy..Both.Sexes* and *Life.Expectancy..Females* in 2021 as they are the continuous variables in the given data set. We find that as the mortality rate increases along the X axis, the life expectancy of all sexes decreases along the Y axis. Most preciously, mortality rate in both sexes, males and females are nearly

identical and negatively correlated (Figure 3) where the values are -0.90,-.88 and -.91 respectively. This indicates there is a strong negative linear correlation between all the sexes and the mortality rate, where females have the largest negative linear correlation.

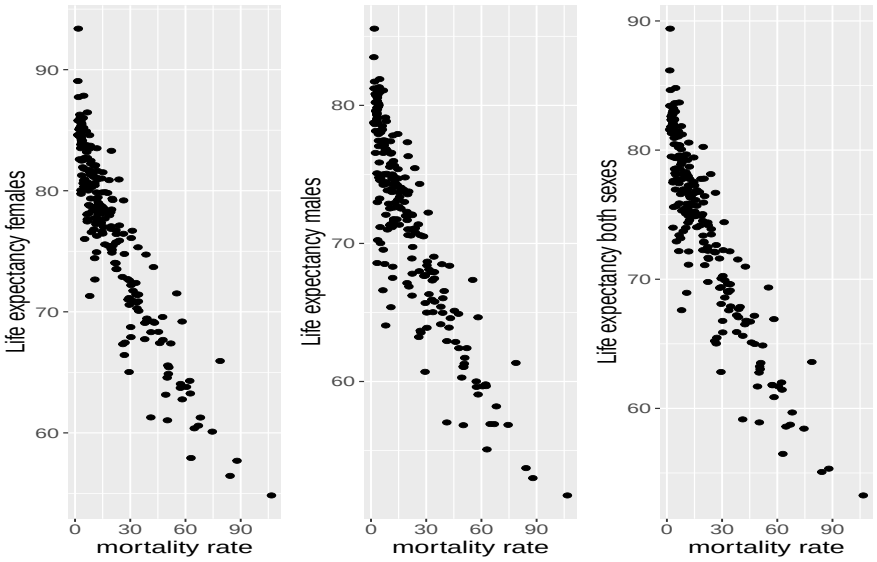

Figure 3: Scatter plot of all sexes with respect to mortality rate

Contrarily, since the line (Figure 4) is nearly linear, there is a positive linear correlation between life expectancy for both sexes when compared to female life expectancy (.99) and male life expectancy (.99). Furthermore, there is a strong correlation between male and female life expectancy (.97). As a result, we can draw the conclusion that there are positive correlations between the sexes, meaning that as female life expectancy rises, so does male life expectancy.

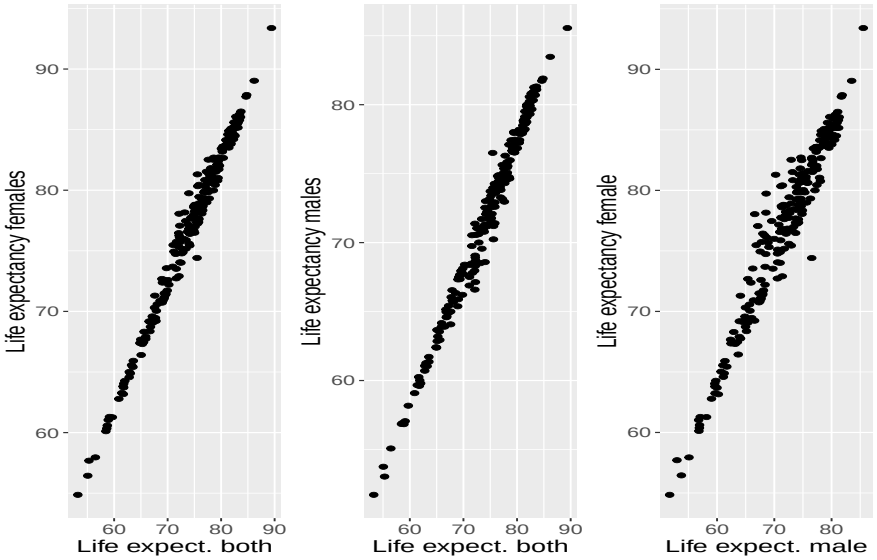

Figure 4: Scatter plot for different combination of sexes.

4.3 Homogeneity or heterogeneity checking:

The analysis of data by region and subregion will be the main focus of this section. Since a box plot (Figure 5) allows us to compare data from various subregions, we will use it to analyze the data. Our data set consists of 5 regions and 21 subregions. All the regions are illustrated with individual colors, e.g., the red color indicates the African region. When the data are examined carefully, it is clear that Western Africa and Middle Africa have roughly symmetric distributions, with median values of 50.71 and 60.58, respectively. Middle Africa's median value is the highest of any of the subregions, indicating that the mortality rate of 50 % countries children is higher than around 70. The table (2) on page 18 shows that the average variance is highest in the African regions. Since the whisker is longer on the right side, central America is right skewed compared to the south American sub region, where the mean and median values are nearly equal at 16.79 and 16.34. Similarly, in the Asian region, western Asia and, especially, south-east Asia, have almost similar means and medians, indicating the data are symmetric (both sides have the same whiskers) and the median is almost in the middle position. However, among all the subregions, south-central Asia has the highest variance (571.48) because of some exceptionally high values (outliers). In contrast, all the European subregions have comparatively lower variance and a consistently lower mortality rate. So, we can say that there is no homogeneity among the subregions when it comes to the death rates.

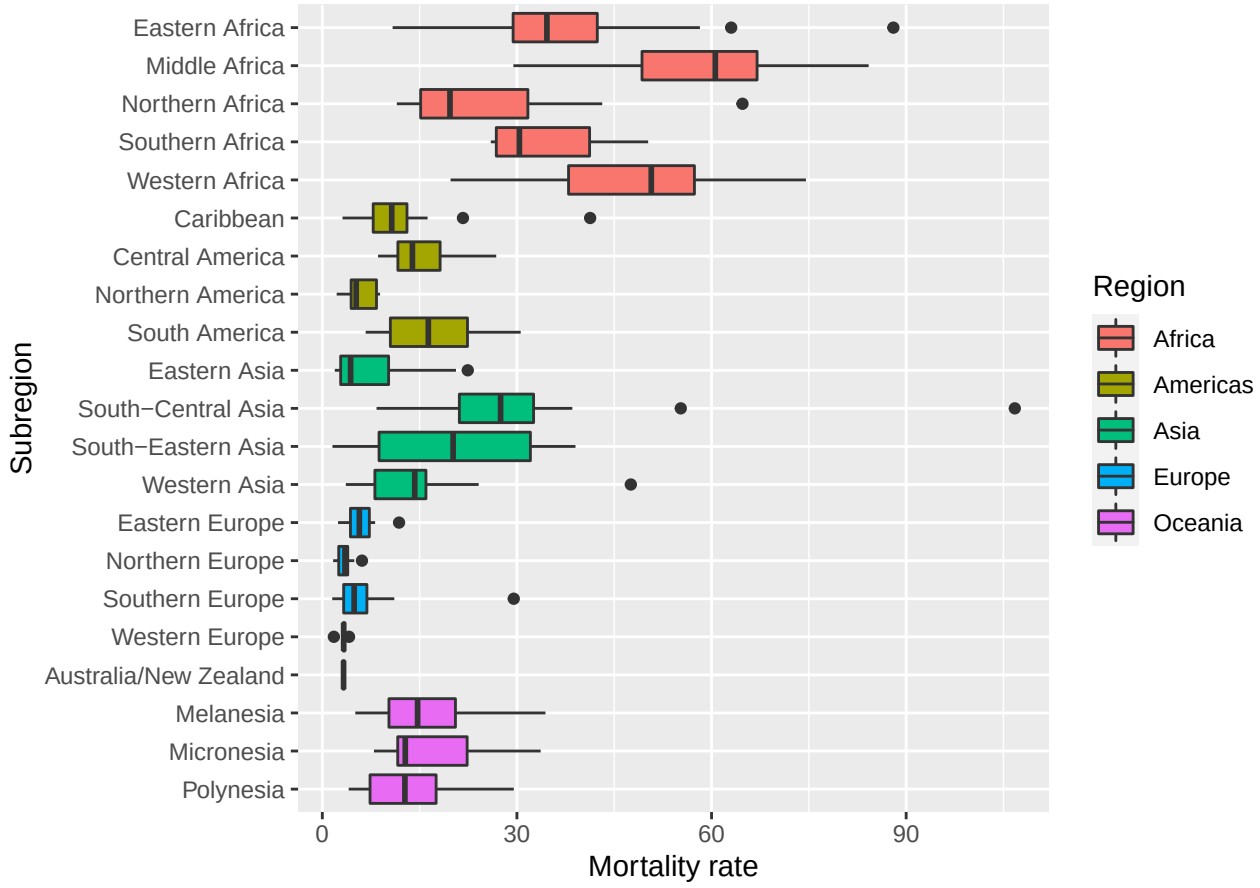

Figure 5: Box plot of all subregions with respect to mortality rate

In the next figure (Figure 6) life expectancy of both sexes among the subregions has been analyzed. Here we see that the life expectancy for the African subregion is very low, where the median value for Middle Africa is the lowest (61.71). Among the Asian subregions, East Asia has the highest median, which indicates that 50 % countries people have a life expectancy greater than 82 years. From the table 2 we see that the IQR is also higher for East Asia (7.53). Europe, on the other hand, has a higher life expectancy than every other subregion except Eastern Europe. Among all the subregions, the median value of western Europe is the highest at 82.36. Thus, we can conclude that there is actually no homogeneity among the subregions in terms of life expectancy for both sexes.

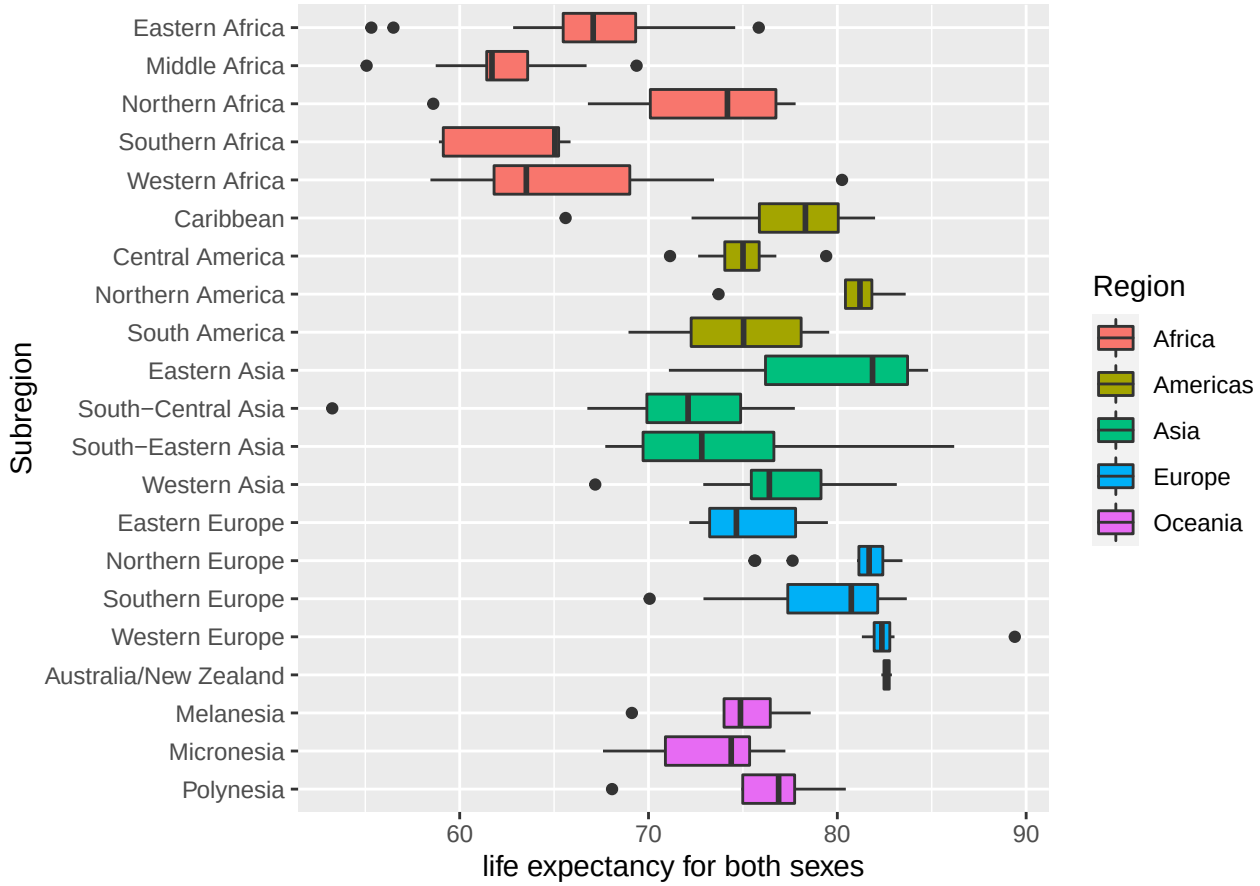

Figure 6: Box plot of all subregions with respect to both sexes life expectancy

4.4 Changes over the time period:

For the past 20 years (2001-2021), changes in the final four numerical variables in our data set have been observed.

A scatter plot (Figure 7) is drawn to demonstrate the changes for the 20-year period of time. All the regions are marked with different colors. The horizontal axis shows data for 2001 and the vertical axis for 2021. A midline in the scatter plot helps to distinguish values from different years. Points falling above this line are considered to be decreased compared to the previous years, and we see two American countries falling into this category. One Asian, American, and European country lie on the line, which indicates no changes in the last 20 years in terms of life expectancy for those countries. However, we see that *Life.Expectancy..Both.Sexes* increased in 2021 compared to 2001 for almost all the regions. Compared to other regions, Europe, America, and some Asian countries

will have a higher life expectancy in 2021. In contrast, the African region tends to have the lowest life expectancy among the regions, and one Asian country will have an exceptionally low life expectancy in 2021.

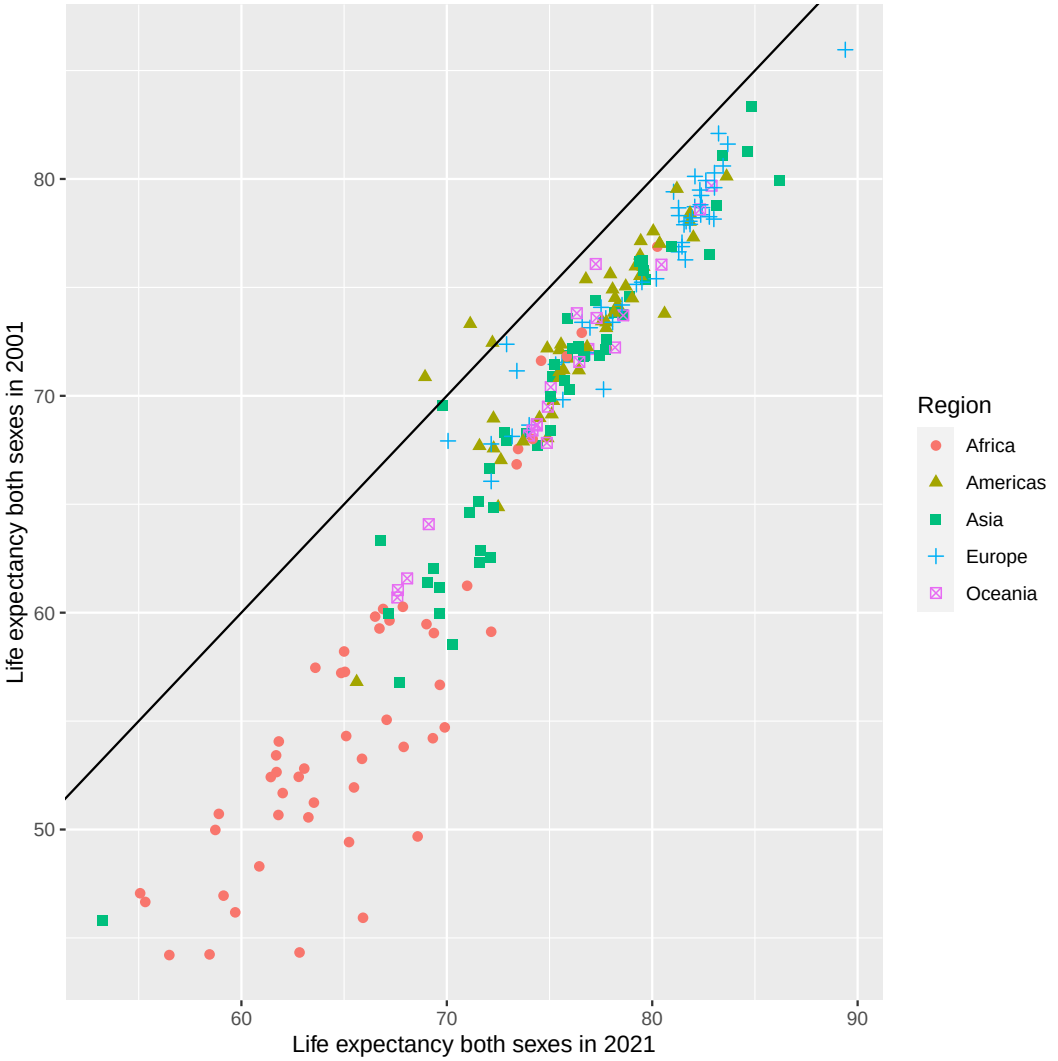

Figure 7: Scatter plot to show the life expectancy of both sexes in 2001 and 2021

We also observe that (Figure 8) overall death rate of both sexes has sharply decreased in 2021 compare to 2001 since most of the countries are lying above the line. However, entire European region tends to have consistence mortality rate throughout the 20 years of period. Some Asian and American region also falls in the same category. In contrast, the African region, including one exceptional Asian country, has higher mortality rate in 2021 than other regions.

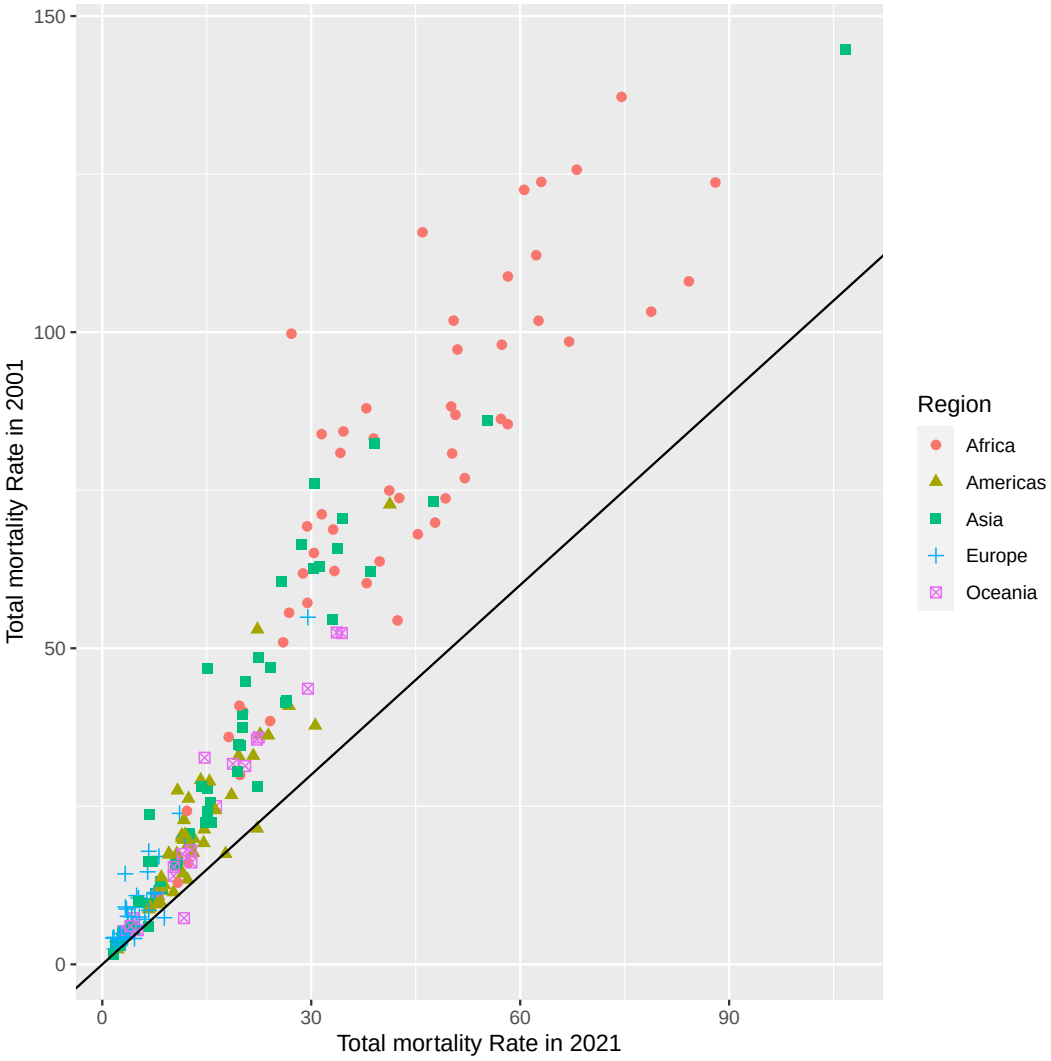

Figure 8: Scatter plot to show the difference of mortality rate in 20 years

Therefore, we can conclude that mortality rate and life expectancy have changed reversely in 20 years, i.e., the former has decreased and the latter has increased over the time period.

5 Summary

The analysis used a small portion of the International Data Base of the U.S. Census Bureau, which was compiled by the instructors of the course Introductory Case Studies. An extensive analysis of data consisting of 454 observations, 8 variables, 24 missing values, 227 countries, 5 regions, 21 sub-regions, and combinations of numerical, integer, and character data showed that women had a longer life expectancy than men. Less prone to accidents might be the substantial reason for longer female life expectancy. Presenting with a histogram, we saw that mortality rate and life expectancy are changing in reverse. This suggests that life expectancy is higher in the countries with lower infant mortality rates. Reducing infant mortality rates may be aided by better sanitation, access to clean drinking water, immunization against infectious diseases, and other public health initiatives. We also observed that positive correlations exist for different combinations (male, female, both) of sexes. In contrast, a negative correlation exists between the mortality rate and life expectancy for all sexes. Heterogeneity and homogeneity were checked, where some European countries have comparatively high levels of homogeneity and overall heterogeneity found among the subregions. Additionally, in 20 years, life expectancy increased for almost all countries, whereas mortality rates decreased. The main reasons for the longer life expectancy may be better lifestyles, better education, and improvements in health care and medicine.

We analyzed only 4 numerical variables, and other external factors like weather, lifestyle, and economy might have a significant effect in any particular region, so the data sets might not be completely perfect. In the future, finding a regression model between the fertility rate and life expectancy related to each subregion might be interesting. In addition, it might be of interest to include other variables such as "high-risk populations (low life expectancy)" and "developed countries" to figure out whether those new variables affect our main measurements or not.

Bibliography

International data base, 2022. URL `https://www.census.gov/programs-surveys/international-programs/about/idb.html`. [Visited on 05-11-2022].

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.3.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

Ansley Coale and James Trussell. The development and use of demographic models. *Population studies*, 50(3):469–484, 1996.

Glossary. Glossary, 2021a. URL `https://www.census.gov/glossary/`. [Visited on 05-11-2022].

Glossary. Glossary, 2021b. URL `https://www.census.gov/glossary/`. [Visited on 05-11-2022].

Wickham Hadley. ggplot2: Elegant graphics for data analysis, 2016. URL `https://ggplot2.tidyverse.org`.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019a.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019b.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019c.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019d.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019e.

Dong Lee, Junyong In, and Sangseok Lee. Standard deviation and standard error of the mean. *Korean journal of anesthesiology*, 68:220, 06 2015. doi: 10.4097/kjae.2015.68.3.220.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Pedersen Thomas Lin. Patchwork, 2021. URL https://patchwork.data-imaginist.com.

Hadley Wickham and et al. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686. URL https://doi.org/10.21105/joss.01686.

Rainer Winkelmann and Klaus F. Zimmermann. Count data models for demographic data. *Mathematical Population Studies*, 4(3):205–221, 1994. doi: 10.1080/08898489409525374. URL https://doi.org/10.1080/08898489409525374. PMID: 12287090.

Appendix

A Additional tables

Variable names	mean($\overline{X}$)	median (X~)	SD (S_x)	1st quartile	3rd quartile
Total mortality rate	20.25	12.58	19.19	6.27	29.48
life expectancy both sexes	74.28	75.56	6.91	69.73	79.42
life expectancy female	76.89	78.36	7.21	72.29	82.34
life expectancy male	71.78	73.99	6.74	67.58	76.94

Table 1: Measure of central tendency and dispersion.

Subregions	Mean	Median	Variance	IQR
Western Africa	48.17	50.71	223	19.4
Southern Africa	34.92	30.38	110.04	14.38
Northern Africa	27.09	19.68	387.70	16.54
Middle Africa	58.60	60.58	292.72	17.74
Eastern Africa	38.62	34.62	339.62	2.98
South America	16.79	16.34	57.30	11.89
Northern America	5.82	5.22	7.80	3.91
Central America	15.45	13.89	33.97	6.51
Caribbean	11.69	10.7	54.13	5.21
Western Asia	14.62	14.25	93.05	7.87
South-Eastern Asia	20.08	20.16	167.98	23.33
South-Central Asia	33.17	27.48	571.48	11.45
Eastern Asia	8.24	4.36	69.52	7.39
Western Europe	3.25	3.29	0.39	0.21
Southern Europe	6.59	4.91	43.89	3.59
Northern Europe	3.48	3.50	1.38	1.39
Eastern Europe	6.01	5.70	7.50	2.91
Polynesia	13.73	12.73	78.97	10.19
Micronesia	17.47	12.79	81.66	10.7
Melanesia	16.99	14.69	126.97	10.25
Australia/New Zealand	3.28	3.28	.10	.23

Table 2: mean, median, variance and IQR of total mortality rate in all subregions

B Additional figures

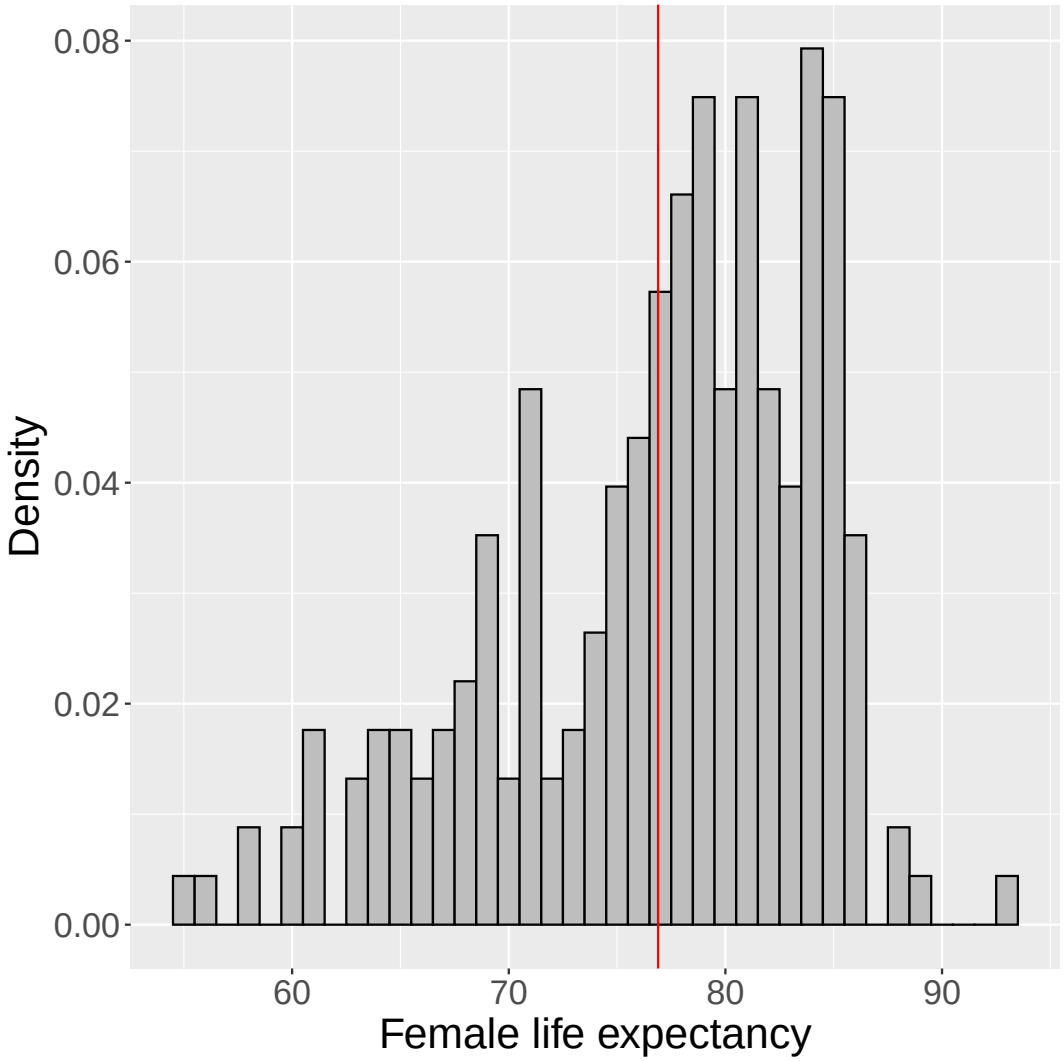

Figure 9: Histogram with respect to *female life expectancy*

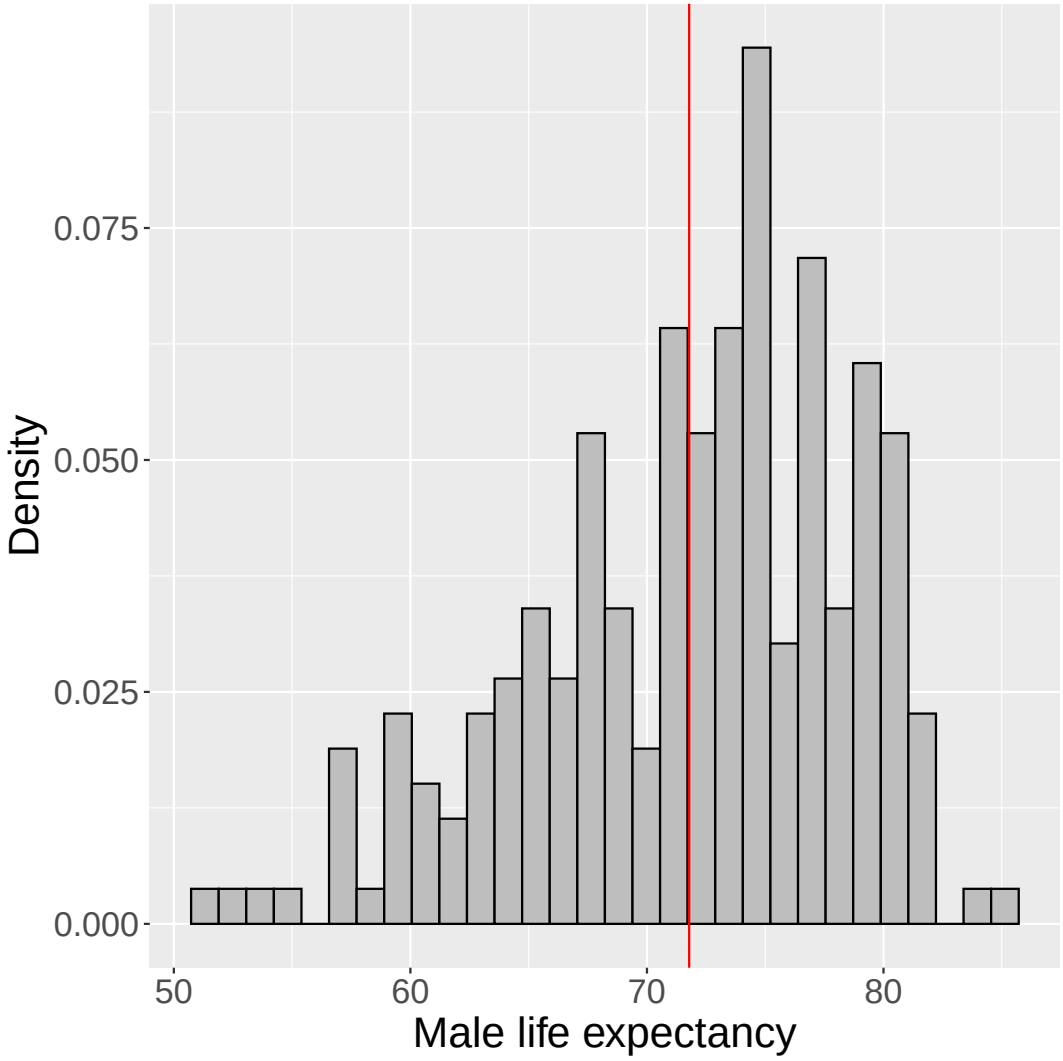

Figure 10: Histogram with respect to male life expectancy