TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Birte Hellwig
Dr. Paul Wiemann
M. Sc. Hendrik Dohme

Author: Bhavesh Jain

Group number: 14

Group members: Nidhi Patel, Akanksha Tanwar, Sohith.D, Bhavesh Jain

November 11, 2021

Contents

1 Introduction

Collecting demographic data helps us to analyze long range trends of various factors on the population. This various factors can be age, sex ratio, life expectancy, income, etc. For this report, we have used a small extract of the demographic data which is a part of The International Data Base (IDB) of the U.S. Census Bureau (U.S. Census Bureau. Population Division. International Programs Center (U.S.) (2021)). Currently, the main database has information from 1950 to 2100, but in this project only the information during the year 2001 and 2021 is used. The variables mainly used for the analysis are Total Fertility rate, Life Expectancy for Both sexes, Life Expectancy for Male and Life Expectancy for Female.

The aim of the projects includes describing the frequency distributions of the variables and then determine if there exists any correlation between them. As the countries in our data set are divided into 5 regions and 21 sub regions, the next task is to determine the homogeneity within sub regions and heterogeneity between different sub regions for every individual variable respectively. The last objective is comparison of values of the variables from 2001 to 2021.

This project is divided into 5 sections. The second section talks about the dataset and describes the variables involved in further analysis to accomplish the objectives stated above. The third section contains detailed information about various statistical methods such as Mean, Median, Boxplots, Scatterplots, Histogram and Correlation Coefficient. These methods are then implemented on the dataset provided and the detailed analysis are stated in the fourth section. There were 4 tasks to be accomplished in the project. For the first task, where we need to determine the frequency distributions, we use the statistical methods Mean, Median, and Histograms for graphical representation. For the second task, to check if there exist any bi variate correlations between variables, the Pearson's correlation coefficient was used for numerical representation and scatterplots were used to determine the relations graphically. To accomplish the third task, where homogeneity within sub regions and heterogeneity between sub regions were to be determined, the IQR and boxplots were used. For the fourth task, we had to determine how the variables changed between 2001 and 2021, this was completed using scatterplots. The fourth section contains various plots that were analyzed and detailed information is provided on their making too. The analyzed results are summarized in the summary which is the fifth section and then followed by Bibliography and Appendix for further references to the project.

2 Problem statement

2.1 Data set and data quality

In this project, the dataset was provided by our lecturers, which contains a small extract from the International Data Base (IDB) of the U.S. Census Bureau. The sources used to create the database are in the form of information which was received from state institutions such as censuses, surveys and administrative methods. The database contains data from 1950 to 2100 on all state and regions of our world, which even involves the estimates and projections by the U.S. Census Bureau. This dataset contains 456 observations which were taken from 228 countries for the year 2001 and 2021 respectively, where each country's population is a minimum of 5000. The countries are divided into 5 regions and 21 subregions. There are 9 variables defined in the dataset, which are the following- *Country* - this tells the name of the country to which the corresponding data belong, *GENC* - General expenditure for purposes and activities not falling within any standard functional category and unallocated amounts relating to two or more functions, *Sub region*- cluster of countries based on geographic location in the continent, *Region*- Name of the continent to which the country belong, *Total Fertility Rate*- average number of children born per woman, *Life Expectancy Both Sexes*- average of number of years any person can expect to live, *Life Expectancy Males*- average of number of years, any male can expect to live, *Life Expectancy Females*- average of number of years, any female can expect to live, Year - the year to which the data belong. The sub-region, region and country are in nominal scale, whereas the total fertility rate, life expectancy of both sexes, life expectancy of male, life expectancy of females are averages which is numerically scaled. The variables which are considered further in the report are all excluding *GENC*. The values for the life expectancy of both sexes, life expectancy of males and life expectancy of females which belong to countries "Libya, Puerto Rico, South Sudan, Sudan, Syria, United States" for the year 2001 are missing. These missing values are hence ignored for further analysis.

2.2 Project objectives

In this project, the first objective is to determine the distribution of the data for all individual variables followed by analyzing the correlation between them. After that, the next objective is to analyze the homogeneity and heterogeneity within and between

different subregions respectively, for this objective we first compare the variability of the values in each subregion and compare those values between other subregions. For these objectives stated above, we subset the dataset where year is equal to 2021. Lastly, we analyze and conclude the changes drawn from the data over the last 20 years, where data from both the years 2001 and 2021 are used.

3 Statistical methods

In this section, several statistical methods are presented which will be later used for analyzing the provided dataset. For calculations and visualization, the programming language R (R Development Core Team, 2020) (version 4.0.3) and RStudio (version 1.3.1093) (RStudio Team, 2020) software were used.

3.1 Central tendency methods

These methods summarize about where the central value of the data distribution is likely to be. The most common measures are the arithmetic mean, median and mode. For this project, arithmetic mean and median are used.

3.1.1 Arithmetic Mean

Arithmetic mean is a value which is determined on a finite data which can be both discrete and continuous. It provides the central location of the data by the formula given below.

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

Here $\bar{x}$ denotes the mean and is determined by summing up all the values of x and then dividing it by the count of the numbers of x which is n (here $n \in N^+ = (1, 2, 3, .., n)$). For example: $X = (4, 8, 3, 5)$, the mean $\bar{x}$ = (4+8+3+5)/ (4), which is equal to 5. Hence, the mean of X is 5.(Mood and Graybill, 1963)

3.1.2 Median

Median is another measure of central tendency which provides the central value based on the distribution of the data. The data or the values are arranged in the ascending

order, the value that divides the data into two equal proportions is known as median. The median is determined by the following formula:

$$Median(X) = \begin{cases} \frac{x_{n+1}}{2} & \text{if } n \text{ is even} \\ = \frac{[x_{n/1} + x_{n/2+1}]}{2} & \text{if } n \text{ is odd} \end{cases}$$

Where x is a set of n odd number of elements. For a set of even number of elements, the median is determined by taking the average of the two values at the position's n/2 and (n/2)+1. For example: $X = (2, 4, 6, 8, 10)$, as there are 5 values in total, the median is the $((5+1)/2)^{th}$ value, i.e., it's the 3rd value in the list, which is equal to 6.(Mood and Graybill, 1963)

3.2 Graphical Methods

3.2.1 Histogram

Histogram is basically a graph which is used to represent the distribution of the data with the help of vertical bars. To construct a histogram, firstly, divide the range of values into series of intervals which are known as bins and then add the number of observations that fall into that particular bin and increase the bin size vertically upon each observation falling under the same bin. If there are lot of observations, then the histogram can also be normalized which can show the relative frequency via bin's height such that summing up the heights of the bins equals to 1. Here we can see that x axis shows the range of variables and y axis talks about the frequency in each bin (range). This gives us a clear picture for distribution of the variables. It accepts both continuous and discrete values for distribution. With the help of histogram, we can also determine the density estimation f(x) of each interval of the distribution, for this we use the following formula.

$$f(x) = \frac{h_i}{n * (b_i)}$$

Where h_i represents the frequency or the count of the values which falls in the bin respectively, b_i represents the width of each bin, which is determined by the range of the data and the number of total bins. The index i takes the value from 1, 2, 3 and so on up to k, where k represents the total number of bins and n be the total number of observations used to plot the graph. (Chen et al., 2008)

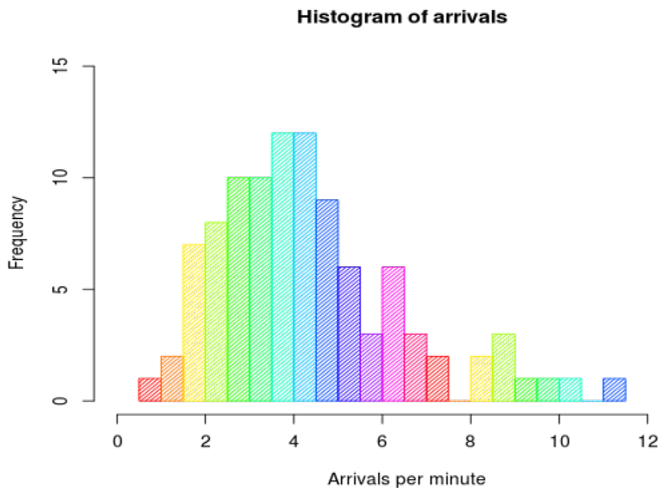


Figure 1: Example of a Histogram
(Wikipedia contributors, 2021a)

3.2.2 Boxplot

Boxplots are the boxed graphical representation of the distribution of the data sets where the box is based on quartiles. Quartile divide the observations into 4 parts, where each part represents 25% of the total of observations. An example of a boxplot is shown below. The box plot shows us the Minimum value, Maximum value, First quartile(Q1),

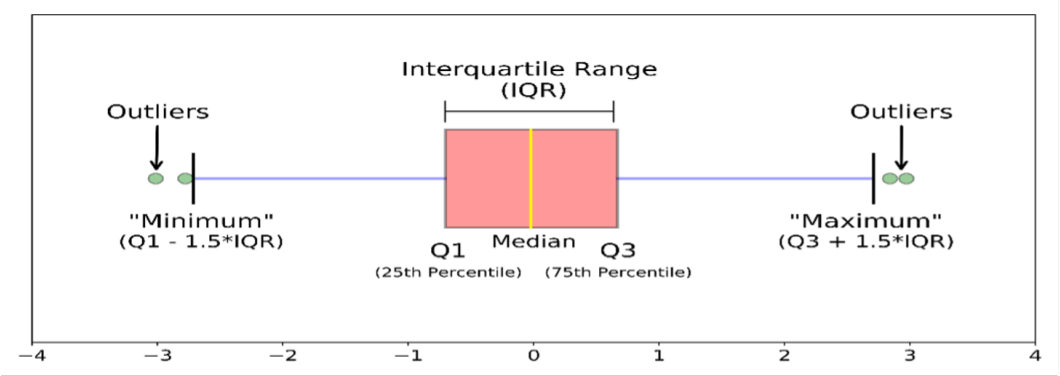


Figure 2: Example of a Boxplot
(Galarnyk, 2018)

Second quartile (Median), Third quartile(Q3), Interquartile range (IQR), whiskers (blue line), outliers. The Interquartile range shows how the middle 50% of the data is spread and is given by the formula IQR = Q3 – Q1. The Q1 quartile is a value such that 25% of the observations in the dataset are less than Q1 and 75 percent of the observations

are greater than Q1. The Q3 quartile is a value such that 75% of the observations are less than Q3 and 25% of the observations are greater than Q3. The minimum and the maximum value are derived by the formula "Minimum = $(Q1-1.5*IQR)$ and Maximum = $(Q3+1.5*IQR)$" respectively and the values preceding the minimum and exceeding the maximum values are considered outliers.(Chen et al., 2008)

3.2.3 Scatterplot

Scatter plots show how two variables relate to each other. The relation between the two variables is seen by how the points are scattered over the plot. There can be a positive, negative and no correlation between the variables, if the variables tend to move in same direction, then we can say that there exists a positive correlation and a negative correlation exists if they tend to move in opposite direction and for random scatter of the data, we say that there is no relation. While making use of scatterplot, we usually

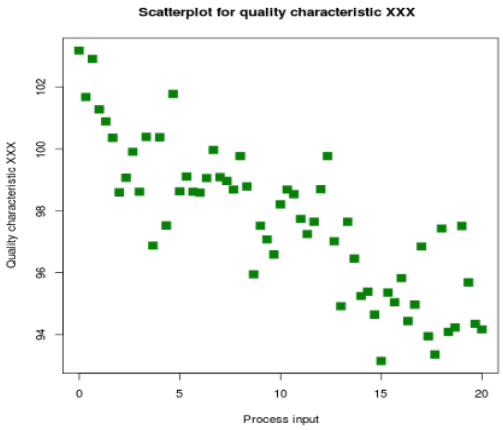


Figure 3: Example of a Scatterplot
(Wikipedia contributors, 2021b)

take one variable on $xaxis$ and another variable on $yaxis$. To get a more clarity on the correlation, we even plot a diagonal line with the equation $y = x$, where x and y are both the variables respectively. Here the diagonal line helps to interpret the linearity among the variables. (Chen et al., 2008)

3.3 Measures of dispersion

3.3.1 Variance

The variance is a measure which defines the average of the squared differences of each value when we consider the differences from its mean. If the observations of the variables are more spread out, we get larger variance. The formula for variance is as follows:

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

Here, we first calculate the mean $\bar{x}$ of the observations and then for every value x_i we determine the difference from the mean. After we get a whole set of differences, we add them up and square it and then divide by n-1 when we consider a sample and divide by n when we consider a population, where n= number of observations. (Rose and Smith, 2002)

3.3.2 Standard deviation

To determine the stand deviation, we simply take the square root of the variance. Standard deviation is basically just a number which tells how the values for a group are spread out from the mean of the values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a greater standard deviation indicates that the values are spread out over a wider range. The formula for standard deviation is given below:

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

Here S^2 is the variance and the x_i, $\bar{x}$, n, denotes the same of which are stated in the formulation of variance. (Mood and Graybill, 1963)

3.3.3 Skewness

Skewness is a measure which tells us about how the curve/bars in a frequency distribution deviated when considering the graph about its mean. The value for skewness can be negative, positive, zero, or can be undefined. Negative skew: In the Figure 4, we can see that the left tail is longer; Tail here means when the curve reached the end (*orangeline*). Most of the elements in the data are on the right of the figure. In this

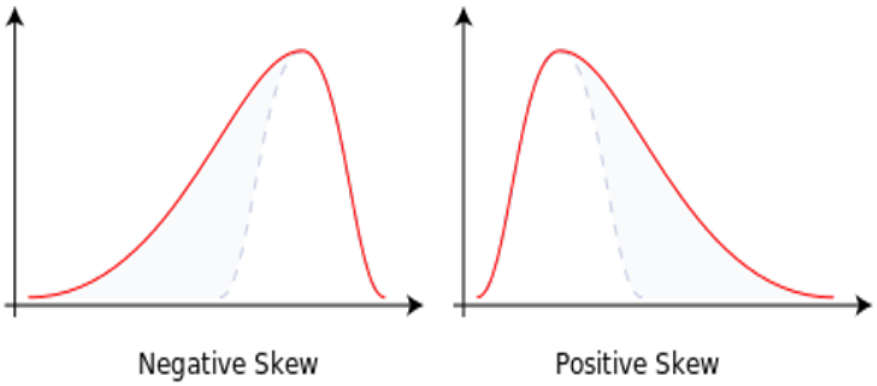


Figure 4: Example of Skewness
(Wikipedia contributors, 2021c)

case, the distribution is said to be left-skewed. It is noticed that left skewed data leans to the right; and left in the left skewed refers to the left tail being drawn out or elongated towards left. A left-skewed distribution has a steep decrease on the right side of the curve. Positive skew: The positive skew follows the same concept as the negative skew, but here the tail is longer on the right side. Here, most of the elements are on the left side of the distribution of the data and the curve descends steep on the left side. (Hastie et al., 2001)

3.4 Measures of Association

3.4.1 Pearson correlation coefficient

Correlation coefficient is a value that defines how two variables are related to each other. The formula for determining the Pearson's correlation coefficient is given by:

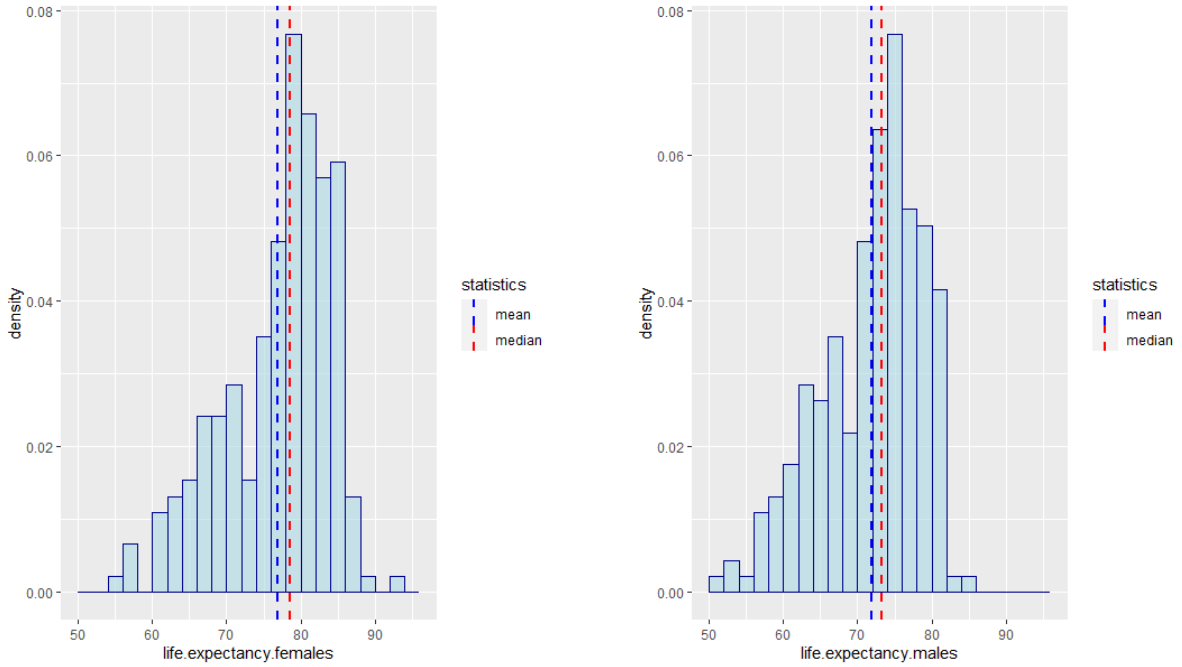
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\sqrt{(y_i - \bar{y})^2}}}$$

In the above formula r_{xy} denotes the correlation coefficient, n = Number of elements where the number n belongs to N^+, x_i = an element in X (set of elements), y_i = an element in Y (set of elements), $\bar{x}$ and $\bar{y}$ denotes the mean of the elements in the respective sets. The correlation coefficient can take value between -1 and +1. A negative correlation is obtained if the coefficient has the value -1 and it's said to be positive when its +1. (Mood and Graybill, 1963)

4 Statistical Analysis

In this section, we make use of the statistical methods which are described in the section 3 on the provided dataset to accomplish the proposed objectives.

4.1 Frequency distribution of the variables



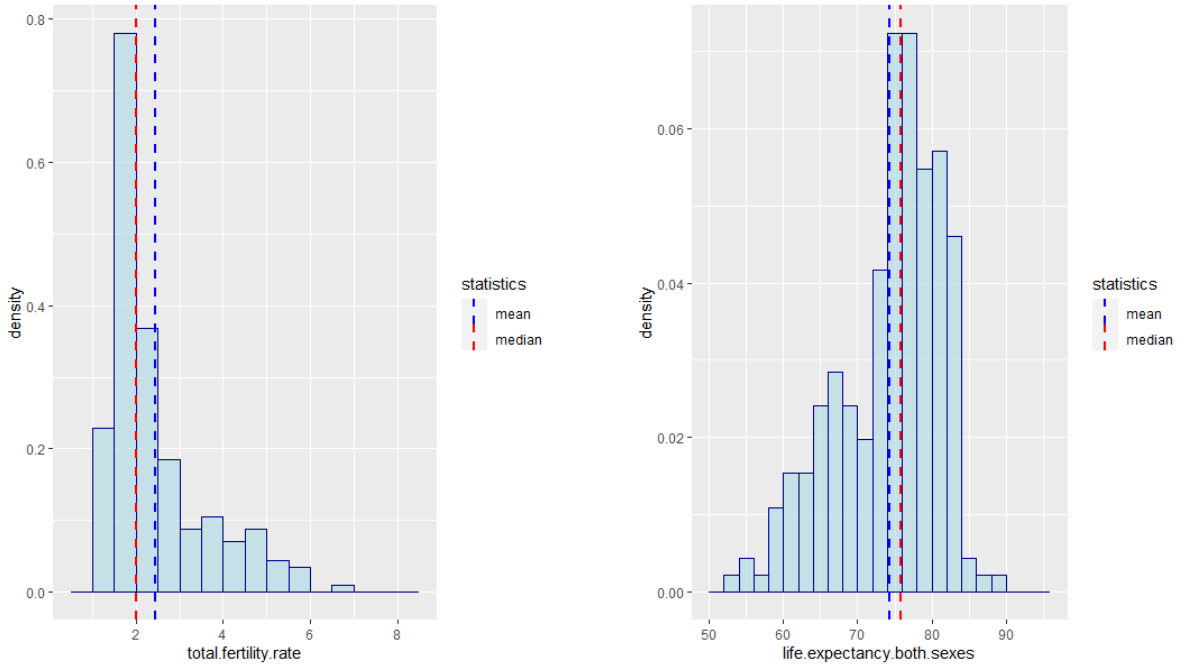
(a) Life Expectancy Females

(b) Life Expectancy Males

Figure 5: Histograms for Frequency distribution.

This subsection describes the distribution of the values for each variable. For this task, we first need to subset the dataset and only consider values for the year 2021. The next step is to determine the arithmetic mean and median of the variables- Life expectancy of males, life expectancy of females and Difference in Life expectancy in both sexes respectively. Then, we plot a histogram where the x axis represents the variables and the y axis represents the density for each range of values in the variable. As all our values for the variables were greater than 50 for females and greater than 50 for males, we considered having the bins range from 30 and 40 for males and females respectively, if a value falls in the range of the bin then the height of the bin is increased with respect to the value on the y axis. Figure 5(a) shows the distribution of the Life expectancy for females, from the figure we can see that distribution of the life expectancy for females is

left skewed with having 78.60 years as median and 76.96 years as the average. It can be seen that maximum number of females have life expectancy between 75 to 85 and very few has the life expectancy above 90. The Figure 5(b) shows the frequency distribution of the life expectancy of males. Here we again use a histogram to plot the frequencies and we notice that life expectancy of males is also negatively skewed. The median of the life expectancy for males is 73.22 years and males have an average life expectancy of 71.83. Males have minimum life expectancy of 50 and can be maximum around 85.55. There are no males whose life expectancy is more than 90, but not the same stands for females. For the life expectancy for males, we can see that it has the density value of 0.07 for the range of 73-75 and for the females, we can see that for the bins ranging from 76-80, we have the density value as 0.07.



(a) Total Fertility Rate (b) Life Expectancy both sexes

Figure 6: Histograms for Frequency distribution.

Similarly, we plot the frequency distributions for Total fertility rate and overall life expectancy. From figure 6(a) we can determine that the median fertility rate is 1.99 and the average is 2.43 and the maximum is 6.91, It is also seen that the distribution is positive skewed and most of the countries have the value between 1 and 3 and there are very few where fertility rate is beyond 5. From the figure 6(b), the median is 75.80 and the average is 74.33. It can be interpreted from the figure below, Now when we

consider both the sexes individually and relate the difference between them for the Life expectancy, there is a noticeable fact that for males, many countries have the life expectancy ranges from 70 to 80 years and just a couple of countries have life expectancy above 80 years for males, whereas there are many countries where life expectancy of the female is between 80 and 85.

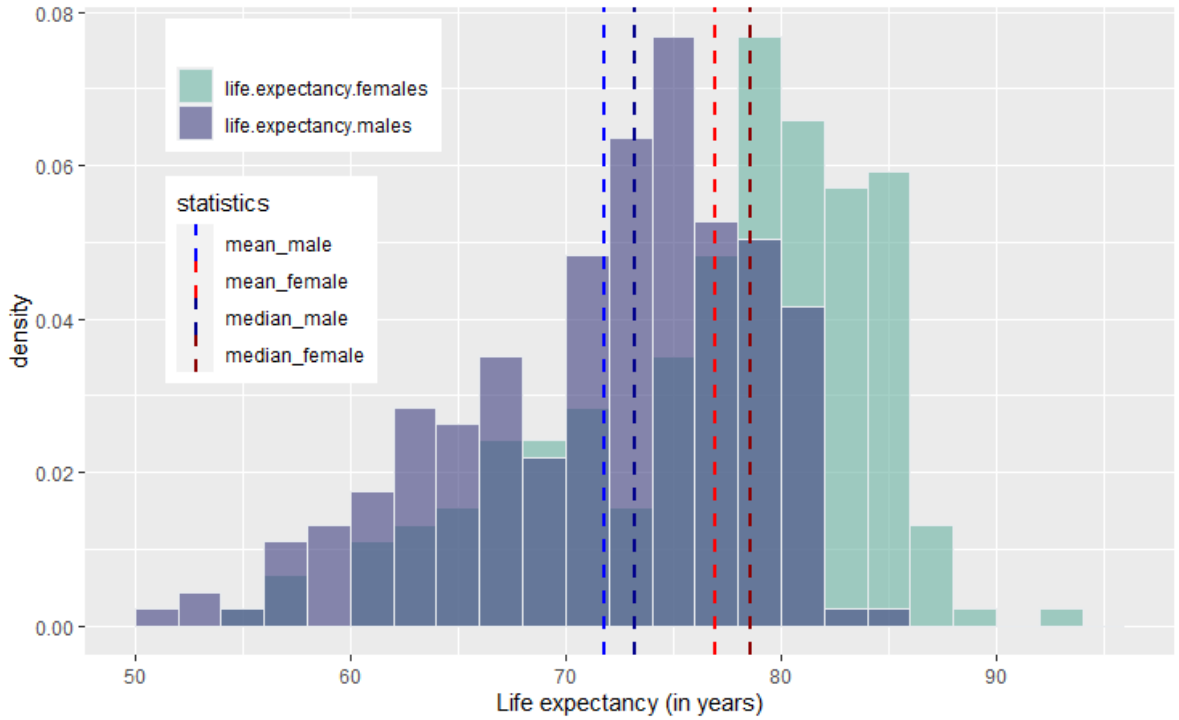


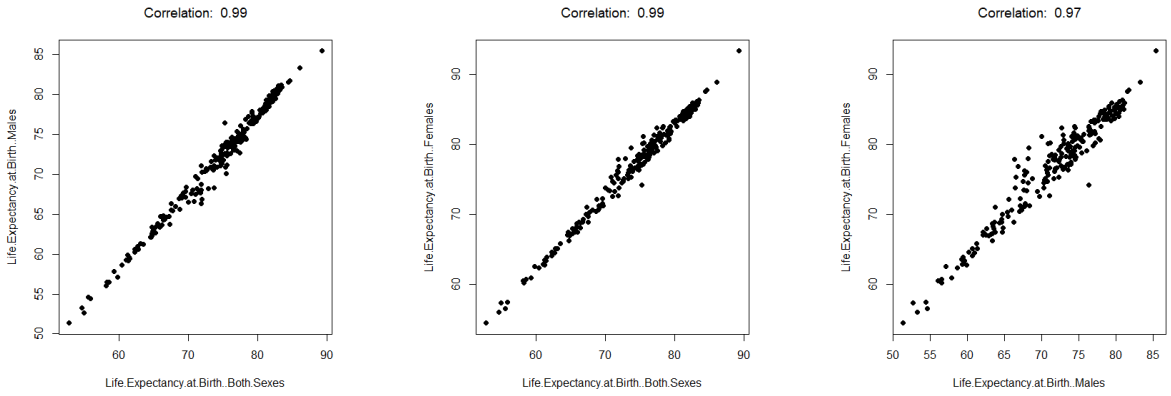
Figure 7: Histogram for Frequency distribution of Life Expectancy Males vs Life Expectancy Females.

4.2 Bivariate correlations between the variables

In this subsection, we derive how one variable is related to each other with the help of scatterplots and Pearson's correlation coefficient. We determine the correlation between the variables *Total Fertility Rate, Life Expectancy Both Sexes, Life.Expectancy Males, Life Expectancy Females*. If we need to derive the values manually by using a simple calculator, It can be done by using the formula stated in the section 3.4.1, But there a function named cor() in the RStudio which automates the process for us and outputs the correlation coefficient values for the input variables. So, after proceeding with the stated process. We get the following results:

	TFR	LEB	LEM	LEF
TFR	1.00	-0.80	-0.77	-0.82
LEB	-0.80	1.00	0.99	0.99
LEM	-0.77	0.99	1.00	0.97
LEF	-0.82	0.99	0.97	1.00

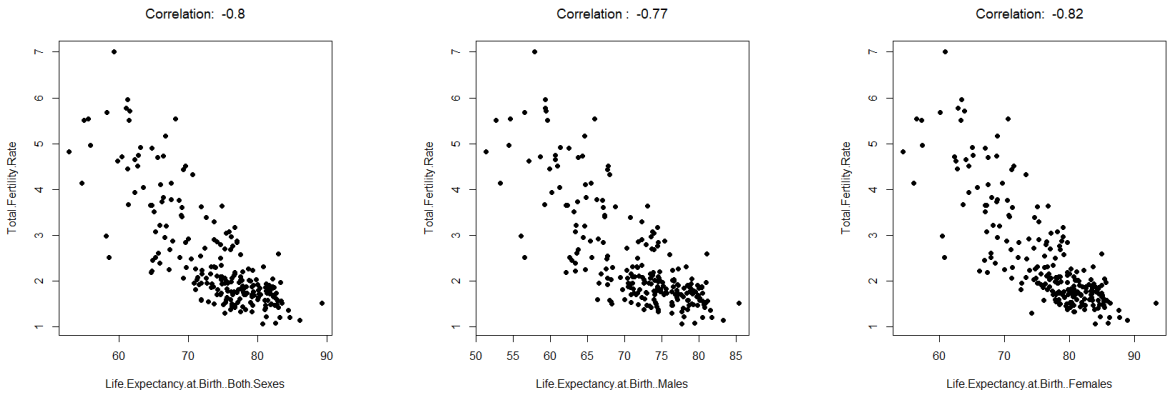
Here, TFR, LEB, LEM, LEF represents the variables 'Total Fertility Rate', 'Life Expectancy both sexes', 'Life Expectancy males', 'Life Expectancy females' respectively. We can see that the values on the diagonals are all 1's as correlation between the variables itself is always 1. As we know that the value of the coefficient can range from -1 to +1 where -1 is negatively correlated and +1 stated that the variables are positively correlated. We observe that there is a strong positive correlation between the following pair of variables 1) LEM and LEB with having the value of .99, 2) LEF and LEB having value of 0.99, 3) LEM and LEF with 0.97. Whereas there is a negative correlation noticed between the variable TFR with the other variables – LEB, LEM and LEF having the coefficient values -0.82, -0.77, -0.82 respectively. To have a better clarity, we now make use of the following scatterplots which determines how a change in one effect the change in another variable. On the basis of the figure above, it is easier to tell that the



(a) (b) (c)

Figure 8: Scatter plot for Bivariate Correlation.

scatter that ascends from the bottom right to the top left are positively correlated and the scattering of points from top left of the plot to the bottom right shows the existence of negative correlation. Therefore, the figure 8 shows the positive correlation and the figure 9 shows the negative correlation.



(a) (b) (c)

Figure 9: Scatter plot for Bivariate Correlation.

4.3 Homogeneity and heterogeneity of values with respect to subregions

As the countries are divided into 21 subregions, here in this subsection, our aim is to verify the homogeneity of the values of the individual variables within subregions and even check for heterogeneity between different subregions. Here, Homogeneity means when the data tends to follow uniformity and display similar properties and whereas heterogeneity means when data shows non-uniformity and has diverse properties. So, to check how the values vary in individual subregion and then compare it between other subregions, we make use of boxplots. On the *xaxis* we take the values of the individual variables and on *yaxis* we take different Subregions into consideration. For the variable *Total Fertility Rate*, the following figure 10 is the result.

If the box plots are widely dispersed for the subregions, it is less probable of being homogenous, because when the plot becomes more dispersed, the homogeneous nature of the variable decreases as variability increases among other observation within a subregion. The countries Australia and New Zealand belong to a single subregion and they show homogeneity for the total fertility rate, as it can be seen that the median value for both the countries is around 1.80. For the subregion of Eastern Africa, Northern Africa and Western Africa, the value ranges from minimum of 1.60 to 6.91, which says that the countries which belong to these subregions are less homogenous among their individual subregions and these subregions tend to show less heterogeneity among each other. Having less IQR value also proves high possibility of homogeneity. From the table 2, we see that the value of the Inter quartile range (IQR = Q3-Q1) of the subregions Northern

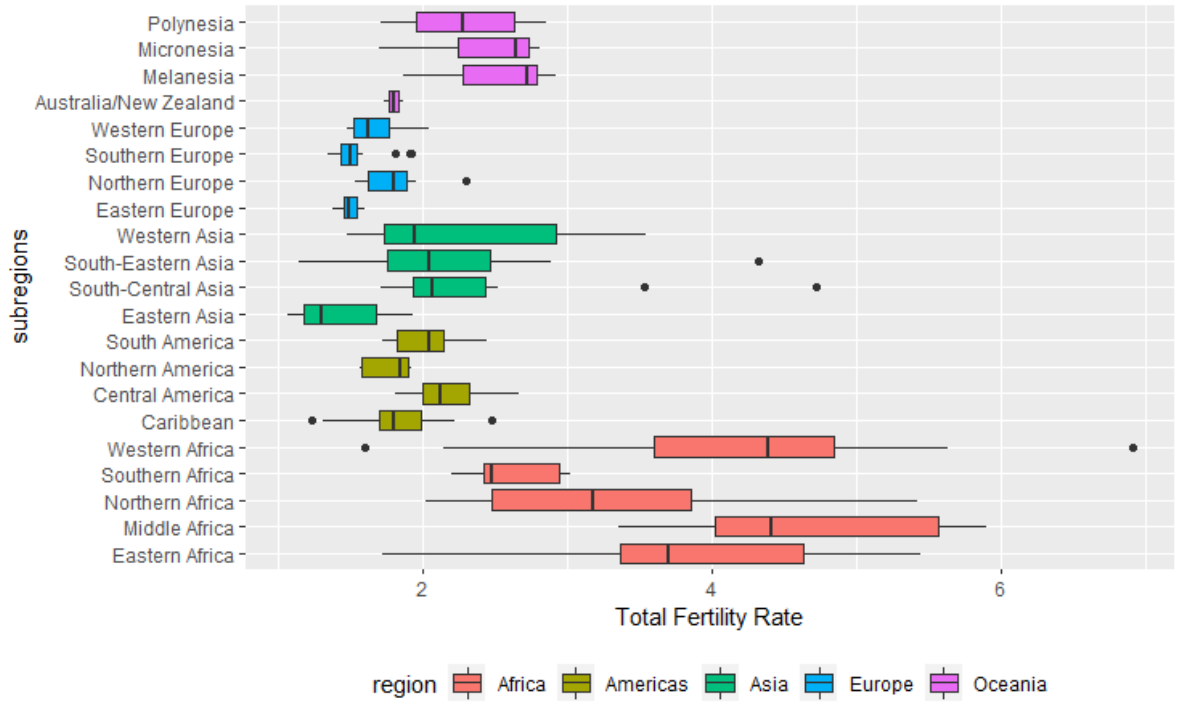


Figure 10: Total fertility rate vs Subregions.

Africa, Eastern Africa and Western Africa are 1.37, 1.27, and 1.24 respectively, which says that they have large IQR values that prove it not to be homogenous within individual subregion and also the IQR values are close to each other proving homogeneity between the subregions. The subregions Southern Europe and Eastern Europe have a similar boxplot which provides evidence of homogeneity within its individual subregion and also between each other, whereas it tends to be heterogenous with the Northern Europe and Western Europe. Similarly, the countries which belong to Northern America are homogenous having the median around 1.84. For South America, some country takes the value of 2 and some take the value 3, so it can be said it is heterogenous. In similar way, we plot the values for the life expectancy for both sexes and obtain this:

Here, from the figure 11 we can observe the following properties. For, Northern Africa the life expectancy shows signs of heterogeneity as the life expectancy varies from 55 to 80 whereas for countries in Southern Africa, life expectancy is from 60 to 65 which shows less variation and proves homogeneity among the countries of Southern Africa. The subregions Middle Africa and Western Africa tend to show little homogeneity between them as the median tends to be constant or having minor changes. For a person born in South

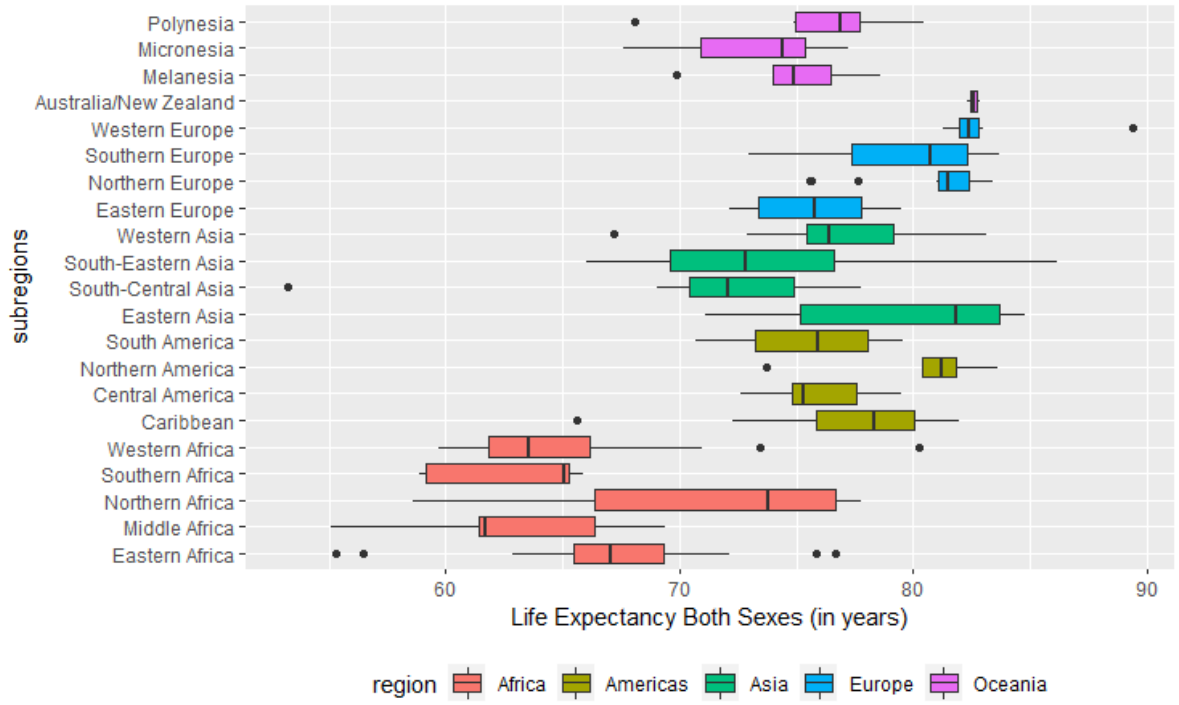
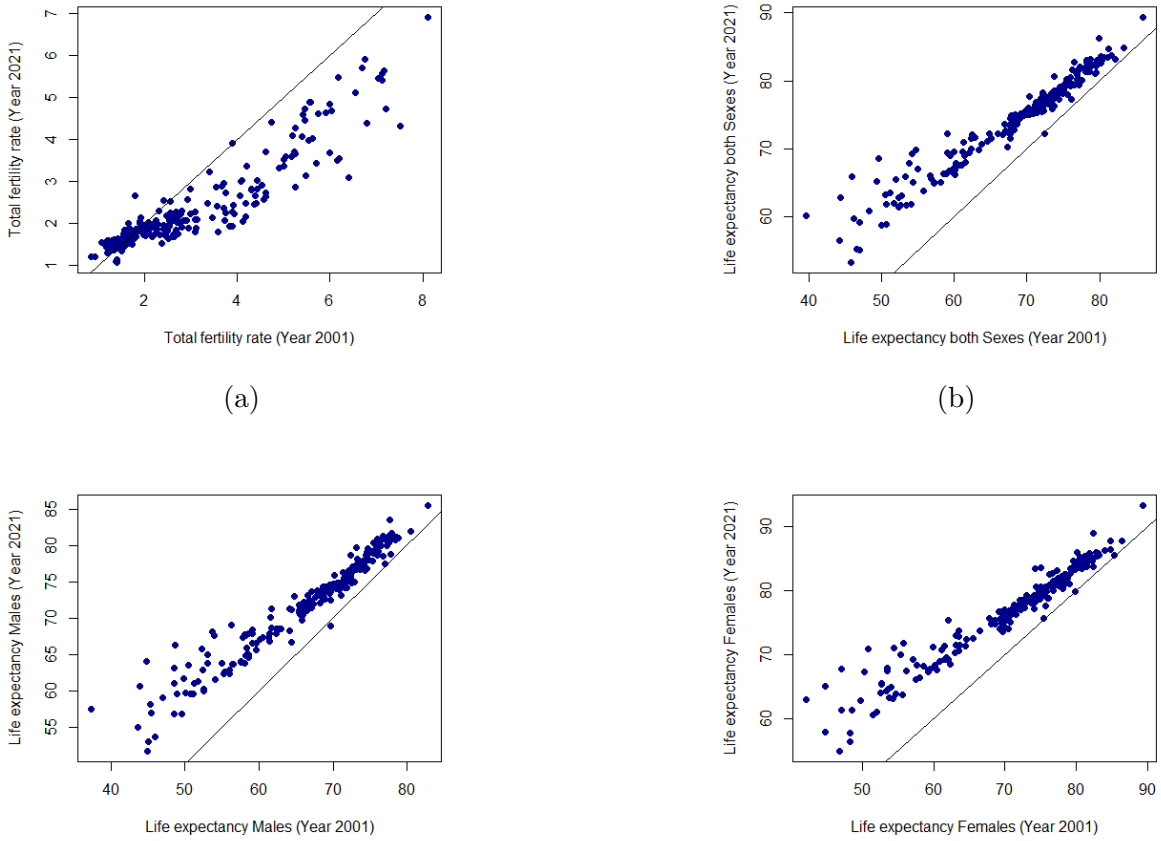


Figure 11: Life Expectancy both sexes vs Subregions

Eastern Asia, the life expectancy can be expected to be anywhere between 65 and 85, which shows large variability among the countries of the South Eastern Asia's subregion. The life expectancy in the subregions of Europe has unequal variation between them. For the Northern Europe, apart from few outliers where the life expectancy is little less, and majority of the countries have life expectancy around 80 years and are homogenous among the subregion. In the subregion of Southern Europe, the life expectancy varies from 72 to 84, but majority of the countries have their life expectancy between 76 and 81, which shows it is little less homogenous. Lastly, Majority of the people born in Western Europe live longer on average when compared to people born in the different subregions of the Europe. For the variables Life expectancy for males and Life expectancy for females, similar plots can be found in the Appendix section of the report, where analysis can be made by using similar methodology.

4.4 Data comparison between 2001 and 2021

For this subsection, we use the dataset without any sub setting. The figure 12 shows the changes in the values of the variable from 2001 to that of 2021. To obtain the above plots, we use scatterplots for every single variable to be analyzed, the values of the variables that belong to the year 2001 are considered on $xaxis$ of the plot and the values for the year 2021 are on the $yaxis$. The upper left plot shows how the Total Fertility rate has changed between 2001 and 2021. We can see that there is cluster formed in the bottom left corner of the plot which says that there are majority of the countries where the fertility rate was less than 3 for the year 2001 and by the year 2021, it has been decrease to 2 and further below. For countries having fertility rate between 4 and 5 in the year 2001 are now having fertility rate ranging between 3 and 4 and only few having above 4. Similarly, we can see that the overall the total fertility rate for each individual country had been decrease over time. The upper right plot shows how the life expectancy of both sexes over time. It can be seen that the minimum value of the life expectancy has increased as during the year 2001, the minimum value was around 45 and during the year 2021 the minimum life expectancy was lifted above 50. It's also says that the majority of the countries where the life expectancy was around 70 to 80 years had an increase in their life expectancy ranging from 75 to 80 years. Therefore, by looking the plot we can say that the there was a linear increase in the life expectancy of both sexes from the year 2001 to the year 2021. The plots that on the bottom left and bottom right of the figure shows the behavior of Life expectancy of males and females respectively. With some exceptional countries where either there was a steep increase in the life expectancy or there was a decrease in the life expectancy, this can be identified by the looking at the observations which are far from the diagonal line which shows a linearity in each of the scatterplots. Here, In all the plots a diagonal line is drawn which helps us to interpret the linearity and also tells us if there exists any points away from the line.



(a) (b)

(c) (d)

Figure 12: Scatter plot: Change in value from 2001 to 2021.

5 Summary

In this project, we have been given a dataset which is an extract from the International Database (IDB), where the dataset contains values for various variables of different countries over the years 2001 and 2021. Now, we subset the data and only consider observations of the year 2021 until the last task. We ignored all the missing to perform each task. With the provided dataset, our main aim of the project is to perform a descriptive analysis of the demographic data with the following four tasks.

In the first task we first determine the frequency distribution of each variables -Total fertility rate, Life expectancy for both sexes, Life expectancy of males and Life expectancy of females, by using histograms and then we determine the mean and median for each variable. We found that the mean is 2.43 for Total fertility rate, and 71.83 years and

76.96 years of life expectancy for males and females respectively. Here, the median stands at 73.22 years for males and 78.60 years for females. We can see that on an average, the life expectancy for females is higher than that of males and both the frequency distributions are negative skewed, where negative skewed means that the majority of the observations belong to the right side of the values in a frequency distribution graph.

For the second task, we checked for correlation between the variables with the means of determining the Pearson's correlation coefficient values for the same variables which were used for to accomplish the Task1. We then determined the Pearson's correlation coefficient values for the "Total fertility rate" against the variables "Life expectancy for both sexes, Life expectancy for males, Life expectancy for females" are -0.80, -0.77, -0.82 respectively. This shows that there is a negative correlation between the variables stated. As we know, the value for correlation coefficient can take values from -1 to +1, where -1 shows high negative correlation and +1 shows high positive correlation. The coefficient value obtained for Life expectancy for males against Life expectancy of both sexes and Life expectancy of females are 0.99 and 0.97, which shows there is a strong correlation between them. Similarly, the values obtained for Life expectancy for females against Life expectancy for both sexes is 0.99, which shows a high positive correlation. To view this correlation graphically, we used scatterplots for each variable against other variables individually. Here, the positive correlations show the scatter moving from bottom left to top right and the negative correlation is seen where the scatter of the points move from top left of the plot to the bottom right of the plot.

After the determination of correlation numerically and graphically, we check for existence of the Homogeneity in a sub region and then verify heterogeneity among other subregions. Here, we consider to make use of boxplots for each subregion, where subregion is a cluster of countries that belong to that particular subregion. After observing the boxplots, we found that the Total fertility rate for the subregions Northern Africa, Eastern Africa, Western Africa and Western Asia are less homogenous in their individual subregion. The values of the subregions of Southern Europe, Eastern Europe and Northern America are more homogenous in their own subregion. Similarly, we found that the Northern, Eastern and Western African subregions are less heterogenous among each other. Using same methodology, we found that Life expectancy for both sexes for the subregions Western Europe, Northern America, Northern Europe, Australia/New Zealand are more homogeneous in their individual subregions and all the subregions of the Africa show high chances of heterogeneity among each other.

Lastly, we compared the variables of the dataset between the year 2001 and 2021 to accomplish the final task of the project. Where we found that the life expectancy for both sexes, males and females have slightly increased from year 2001 to 2021. The total fertility rate for many countries have been decreased over time. As there exists a strong negative correlation between the total fertility rate and the life expectancy for both sexes, we can say that the decrease in fertility rate is one of the factors affecting the increase in life expectancy for both sexes.

For further studies, it would be useful to study what other factors are responsible for the increase in the life expectancy and what changes in the factors have caused the total fertility to decline over time. Finally, we can determine how these changes affect the world population in a whole.

Bibliography

Chun-houh Chen, Wolfgang Hrdle, Antony Unwin, Chun-houh Chen, Wolfgang Hrdle, and Antony Unwin. *Handbook of Data Visualization (Springer Handbooks of Computational Statistics)*. Springer-Verlag TELOS, Santa Clara, CA, USA, 1 edition, 2008. ISBN 3540330364.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

A.M.F. Mood and F.A. Graybill. *Introduction to the Theory of Statistics*. McGraw-Hill series in probability and statistics. McGraw-Hill, 1963. URL `https://books.google.de/books?id=FPRQAAAAMAAJ`.

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

Colin Rose and Murray D. Smith. *mathStatica: Mathematical Statistics with Mathematica*. Physica-Verlag HD, Heidelberg, 2002.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020. URL `http://www.rstudio.com/`.

U.S. Census Bureau. Population Division. International Programs Center (U.S.). *International data base (IDB)*. Washington, D.C. : U.S. Census Bureau, Washington, D.C. : U.S. Census Bureau, 2021. URL `https://www.census.gov/glossary`.

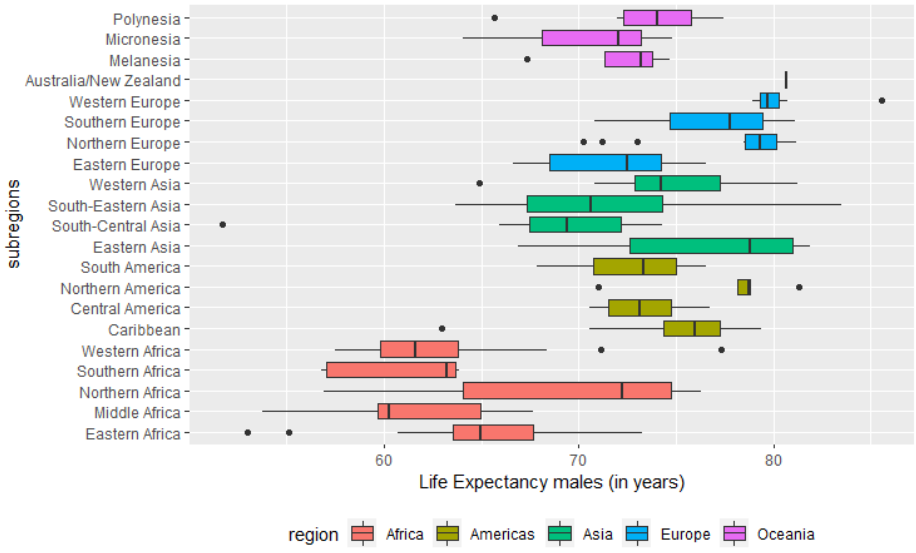
Wikipedia contributors. Histogram — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Histogram&oldid=1051560529`, 2021a. [Online; accessed 11-November-2021].

Wikipedia contributors. Scatter plot — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Scatter_plot&oldid=1054512191`, 2021b. [Online; accessed 11-November-2021].

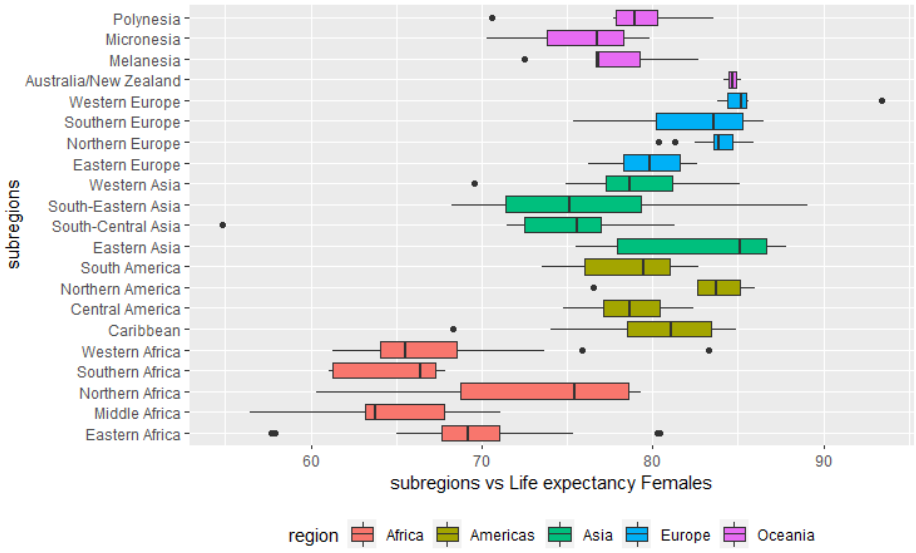
Wikipedia contributors. Skewness — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Skewness&oldid=1049351964, 2021c. [Online; accessed 11-November-2021].

Appendix

A Additional figures



(a) Life Expectancy Males vs Subregions



(b) Life Expectancy Females vs Subregions

Figure 13: Boxplots for life expectancy of males and females vs Subregions.

B Additional tables

Table 1: Quartile Measure of Life Expectancy both sexes values for Subregions

	subregion	min	q1	median	q3	max	iqr
1	Eastern Africa	55.32	65.48	67.07	69.32	76.70	3.84
2	Middle Africa	55.07	61.43	61.71	66.35	69.37	4.92
3	Northern Africa	58.60	66.36	73.78	76.66	77.79	10.30
4	Southern Africa	58.90	59.13	65.04	65.24	65.87	6.11
5	Western Africa	59.70	61.82	63.53	66.15	80.25	4.33
6	Caribbean	65.61	75.87	78.31	80.05	82.00	4.18
7	Central America	72.63	74.80	75.33	77.56	79.47	2.75
8	Northern America	73.71	80.43	81.20	81.83	83.62	1.40
9	South America	70.70	73.23	75.94	78.08	79.57	4.85
10	Eastern Asia	71.08	75.14	81.86	83.72	84.81	8.57
11	South-Central Asia	53.25	70.40	72.09	74.88	77.75	4.48
12	South-Eastern Asia	66.00	69.62	72.82	76.64	86.19	7.02
13	Western Asia	67.18	75.44	76.40	79.14	83.15	3.69
14	Eastern Europe	72.16	73.39	75.80	77.79	79.50	4.40
15	Northern Europe	75.61	81.11	81.50	82.41	83.45	1.30
16	Southern Europe	72.99	77.38	80.74	82.33	83.68	4.95
17	Western Europe	81.30	81.95	82.36	82.78	89.40	0.83
18	Australia/New Zealand	82.33	82.47	82.61	82.75	82.89	0.28
19	Melanesia	69.86	74.00	74.87	76.45	78.59	2.45
20	Micronesia	67.59	70.89	74.38	75.36	77.25	4.46
21	Polynesia	68.07	74.99	76.89	77.74	80.45	2.75

Table 2: Quartile Measure of Total.Fertility.Rate values for Subregions

	subregion	min	q1	median	q3	max	iqr
1	Eastern Africa	1.73	3.36	3.70	4.63	5.45	1.27
2	Middle Africa	3.36	4.02	4.41	5.57	5.90	1.55
3	Northern Africa	2.03	2.48	3.18	3.86	5.43	1.37
4	Southern Africa	2.20	2.42	2.48	2.95	3.03	0.53
5	Western Africa	1.60	3.60	4.39	4.84	6.91	1.24
6	Caribbean	1.23	1.70	1.80	1.99	2.48	0.29
7	Central America	1.81	2.01	2.12	2.32	2.67	0.32
8	Northern America	1.56	1.58	1.84	1.90	1.92	0.32
9	South America	1.73	1.83	2.04	2.15	2.45	0.32
10	Eastern Asia	1.07	1.18	1.30	1.68	1.93	0.50
11	South-Central Asia	1.71	1.93	2.07	2.43	4.72	0.50
12	South-Eastern Asia	1.15	1.75	2.05	2.47	4.32	0.71
13	Western Asia	1.48	1.73	1.95	2.92	3.54	1.19
14	Eastern Europe	1.38	1.45	1.49	1.54	1.60	0.09
15	Northern Europe	1.54	1.62	1.80	1.89	2.30	0.26
16	Southern Europe	1.35	1.44	1.50	1.55	1.92	0.11
17	Western Europe	1.48	1.52	1.63	1.77	2.04	0.25
18	Australia/New Zealand	1.74	1.77	1.80	1.83	1.87	0.07
19	Melanesia	1.87	2.28	2.72	2.79	2.92	0.51
20	Micronesia	1.70	2.25	2.65	2.74	2.81	0.49
21	Polynesia	1.71	1.96	2.28	2.64	2.86	0.68