TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Prof. Dr. Katja Ickstadt
M. Sc. Zeyu Ding
M. Sc. Yassine Talleb

Author: Saptarsi Bhattacharya

Group number: 15

Group members: Burak Dede, Syed Hassan Saqlain Tayyab, Saptarsi Bhattacharya, Hemalatha Sekar

May 12, 2023

Contents

1 Introduction

Census statistics are collected every ten years to examine the impact of population expansion on various characteristics. Understanding the differences in life expectancy between continents can also help us better understand how lifestyle, education, and access to high-quality healthcare services have affected it throughout the years. It's been really interesting to realize the difference, where we are 20 years from now, especially with the global fight against the dreaded COVID-19.

This research examines 227 nations to identify regional variations or convergences in life expectancy at birth and under-five mortality rates between 2002 and 2022. To this end, we first performed a uni-variate analysis of all numeric variables in 2022. Bi-variate associations between these numerical variables are also visible. We also investigate the relationships between and within the various sub-regions of these variables.

The project report that is being presented here has 4 additional sections in addition to the introductory section. The description of the used data set, including the definitions of the variables, and the calibre of the data provided, is covered in Section 2. The definition of various statistical methods needed for data analysis, such as mean, median, variance, etc., is covered in Section 3. In addition to Correlation, here is an explanation of the concept of correlation. Interpreting the results of the final analysis using the statistical methods discussed in Section 3 is covered in Section 4, which follows. The final section: Section 5 serves as the summation. It includes an overview of all the results and explanations. It also offers a perspective for a more in-depth investigation.

2 Problem statement

2.1 Description of Data Set and Quality

A tiny sample of demographic information from the U.S. Census Bureau's (currently from 1950 to 2100) IDB (International Data Base) was utilized in this research. Between 2002 and 2022, it provides data on life expectancy and under-five mortality rates for more than 227 nations. Eight columns, also known as features, and 457 rows, sometimes known as records, make up the data set utilized in this report. Three of the eight features—Country, Subregion, and Region—are categorical type variables. Country-specific information is provided, including the name of the nation where the observations

were made. Five regions are created by grouping together the 227 nations and 21 sub-regions. The element the year in which the data was captured (2002 or 2022) is identified by Year, which is an ordinal variable with a specified order. Life expectancy for both sexes, which is the number of years a person is expected to live on average., and under-five mortality for both sexes are two of the four attributes in the eight numeric type variables. Life expectancy at birth for males and females further divides the varying life expectancy by his, her, their, etc. gender.(International Data Base, 2022)

In order to prevent an unintended influence on the measures that may be derived from the data, the only two missing records for four characteristics are disregarded for the purposes of this report. Overall, the data quality was suitable for statistical analysis.

2.2 Project Objective

For each numerical variable in this project, a univariate analysis was performed first. It can be seen that each variable has a frequency distribution around the world. The data set includes information on life expectancy for all persons as well as for males and females individually, thus a scatter plot is also used to show the differences between the sexes. Finding correlation coefficients can also reveal bivariate relationships between variables. Life expectancy and under-5 mortality for these sub-regions are shown, and the results are explained because the dataset contains data for 21 sub-regions. Only the datum for the year 2022 is used for these activities. Finally, the change in the region-wide variable is recorded over a 20-year period.

3 Statistical methods

Getting a general overview of the variables in the data set would be the first stage in the descriptive analysis. For this project, the provided dataset was evaluated using the statistical techniques listed below. All of the analysis was performed using Jupyter notebook environment (Python Core Team, 2021)

See a numerical example of the n observations listed below: x1, x2, x3, ..., xn.

3.1 Uni-variate

3.1.1 Measure of Central Tendency

The central tendency metrics used in this report are listed below.

Arithmetic Mean The arithmetic mean is calculated by adding a set of numbers and dividing by the number of numbers used in the set. The sample mean can be calculated using the following formula:

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

The population mean (µ) of a known population may be calculated using the method above.(R Development Core Team, 2020)

Median The middle element in an ascending data set is called the median. It can be compared to the geometric middle, as opposed to the mathematical middle represented by the mean. Its computation presumes that the samples x1, x2,.., xn above are arranged in ascending order. The median (x) is then determined as:

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

(R Development Core Team, 2020)

3.1.2 Measure of Spread

The term "spread" refers to the data's distribution. The typical spread metrics the list of sources utilized in this paper.

Standard Deviation and Variance It may be characterized as the measurement of the datum's deviation from the mean, or the spread of the provided data set in relation to the mean. The standard deviation (s)and sample variance (s^2) can be calculated as follows:

You may determine sample variance (s^2) and standard deviation (s) as follows:

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

$$s = \sqrt{s^2}$$

(R Development Core Team, 2020)

3.1.3 Measure of position

Interquartile Range Before understanding the interquartile range (IQR), it is important to understand interquartiles. The five factors that make up the quartiles are the smallest number in a data set, Q1 or quartile 1, which separates the lowest 25 percent of the data set, Q2 or quartile 2, also known as the median, which is the number in the middle of the sorted data set, Q3 or quartile 3, which separates the lowest 75 percent of the data set, and the last factor is the largest number in the data set. The difference between the third and first quartiles then yields the interquartile range.(R Development Core Team, 2020)

3.2 Bi-variate

3.2.1 Correlation Coefficient

According to the correlation coefficient (r), there is a linear link between two random variables. The correlation coefficient is calculated as follows:

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

(R Development Core Team, 2020)

Interpreting values of r Interpreting values of r The correlation coefficient ranges from +1 to -1; the closer the coefficient is to either extreme, the stronger the association. A strong positive relation is shown by r values that are close to 1, a strong negative relationship is indicated by r values that are close to -1, and a linear relationship is not indicated at all by r values that are close to 0.

3.3 Histogram

A histogram shows a variable's frequency distribution. The X-axis represents the variable whose frequency is to be measured, and the Y-axis represents the frequency of the

variables. It separates measurements into bins, or value ranges, whether they are continuous or discontinuous. The different heights of the bins show how frequently the datum occurs. In this report, a density histogram is used instead of a frequency histogram. The density histogram's vertical scale denotes fractions of the total area of all the bars that sum to one. The histogram is used to determine the distribution's shape, outliers, and central tendency of the data.(R Development Core Team, 2020)

3.4 Scatterplot

Two-dimensional graphics that may be used to illustrate pairings of data are called scatterplots. The locations of each point show the values of the associated variables as well as any relationships between them. There is a linear relationship between these two variables when these patterns resemble a line. The scatterplot may be understood as follows, assuming that the horizontal and vertical axes are referred to as X and Y axes: If there is an upward trend, it means that the X-axis and Y-axis are in a positive connection, and if there is a downward trend, it means that the axes are in a negative connection. No pattern or trend means there is no connection between the X and Y axes.(R Development Core Team, 2020)

Pair plots are matrices of scatterplots. It creates a pairwise incidence matrix for each variable in the data set, allowing for fast data analysis.

3.5 Boxplot

The first quartile (Q1), median (Q2), third quartile (Q3), and maximum (Q3) of the data set are used to create a boxplot, which is a visual depiction of the data distribution based on these five numerical values. This five-number summary makes it easier to compare and analyze data. The interquartile range is the space inside the box between the whiskers, which are the horizontal lines that extend out from it. Outliers are data points that are outside the whiskers of the box plot. In order to determine the skewness, dispersion, symmetry, and outliers within a data set, boxplots are utilized.(R Development Core Team, 2020)

3.6 Barplot

Bar charts can be used to display frequencies by using equal-width bar heights corresponding to various categorical variables or component levels (interval ranges). In a bar graph, the frequency is shown by the height of the bar. The width of the bar can be used as interval in some cases, such as interval range. It makes it simple to compare frequencies across multiple categories or time intervals.(R Development Core Team, 2020)

4 Statistical analysis

The full dataset was subjected to the statistical methods described in Section 3 to allow for a better understanding of the data.

Description of the Data Set Since Table 1 describes each numerical variable's central tendency, dispersion, and shape, it may be utilized to comprehend the description of the data set.

	Year	Life.Expectancy.at.Birth..Both.Sexes	Life.Expectancy.at.Birth..Males	Life.Expectancy.at.Birth..Females	Under.Age.5.Mortality..Both.Sexes	Under.Age.5.Mortality..Males	Under.Age.5.Mortality..Females
count	454.000000	448.000000	448.000000	448.000000	448.000000	448.000000	448.000000
mean	2012.000000	71.758080	69.362076	74.278080	37.703862	40.664174	34.619219
std	10.011031	8.728865	8.429339	9.162633	43.278855	45.477918	41.125455
min	2002.000000	44.580000	43.540000	44.990000	1.940000	2.030000	1.640000
25%	2002.000000	67.877500	65.315000	69.910000	8.665000	9.552500	7.320000
50%	2012.000000	73.765000	71.275000	76.565000	19.490000	21.690000	17.375000
75%	2022.000000	78.032500	75.267500	80.982500	47.312500	52.735000	40.882500
max	2022.000000	89.520000	85.700000	93.490000	217.810000	226.880000	208.470000

Table 1: Description of the Data Set

4.1 Uni-variate analysis

Each numerical variable is subjected to a single-variate analysis. Figure 1 allows for the following analysis to be done: The histogram plot of the six numeric variables found in the data set is displayed in Figure 1.

As we can see from the histogram in Figure 1, the distribution of children under the age of 5 between the three mortality rates rate has right skewness, while the plots of the other three life expectancy plots have little skewness to the left. For all three mortality rates, the histogram peaks for under-five mortality are close to zero, with more women than

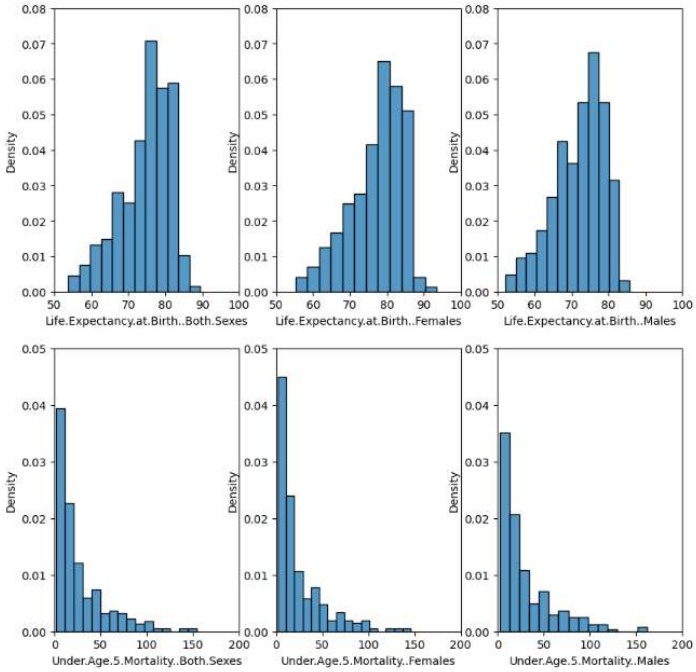

Figure 1: Density distribution of numerical variable

men dying. Correspondingly, the frequency distribution of life expectancy is maximum between 75 to 78 years for both sexes. Correspondingly, the frequency distribution of life expectancy is maximum between 75 to 78 years for both sexes. Considering gender-wise life expectancy, it can be observed that the life expectancy of males is 75 years and females are 80 years. It can be said that, in the year 2022, the female life expectancy is slightly greater than males. Joint scatterplots are used to better understand differences in life expectancy between the sexes.

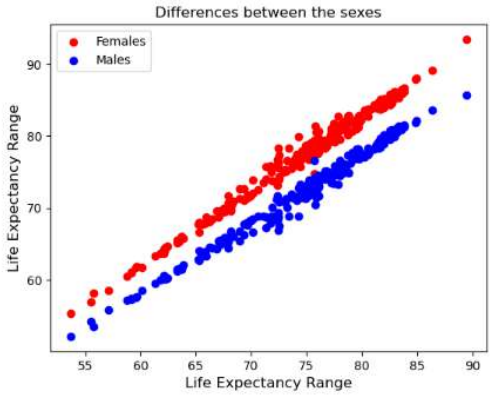

Figure 2: Difference in sexes in terms of life expectancy

As can be seen from Figure 2, depending on gender, women are likely to have a longer life expectancy at birth than men in 2022. While there are numerous ongoing studies and speculations to back this up, we have yet to acquire a definitive basis for this result.

4.2 Variability of the values in the individual and different sub-regions

Under 5 Mortality rate Figure 3 shows the under-five mortality rates for both sexes in the African region. Out of Africa's five subregions, Southern and Eastern Africa has the least amount of datum fluctuation when compared to the other subregions. Northern Africa has the lowest death rate compared to other African subregions, ranging from 15 to 60. The biggest change was in Middle Africa, where the death rate ranged from 40 to 130.

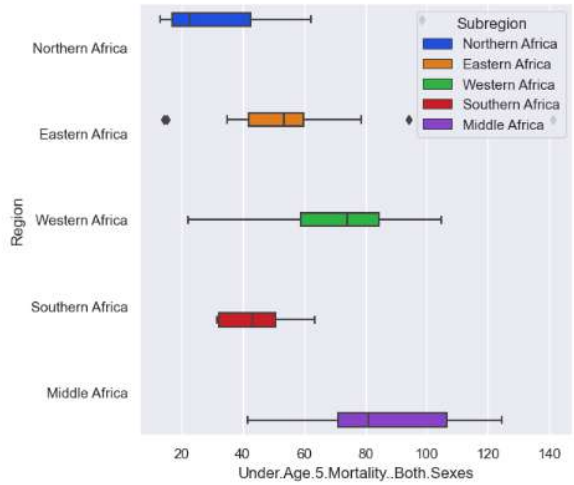

Figure 3: Under 5 Mortality Rate of African Both Sex

Life Expectancy Life expectancy for both sexes in the African region is shown in Figure 4. When compared to the other sub-regions of Africa's five sub-regions, Middle Africa experiences the least datum volatility. Northern Africa has the highest life expectancy compared to other subregions in Africa, ranging from 71 to 77 years. The biggest improvements were seen in West Africa, where life expectancy ranges from 58 to 74 years.

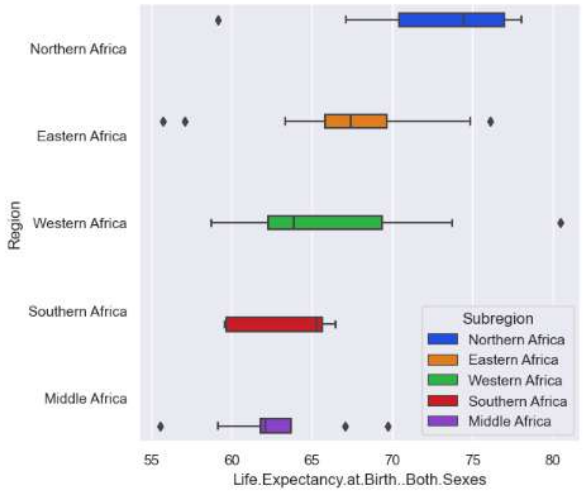

Figure 4: Life Expectancy at Birth for Both Sexes

4.3 Bi-variate analysis

In this part, we'll use the correlation coefficient to see if the variables in the data set have any underlying relationships. Both Table 2 and Table 4 (Appendix) use heat maps to show the values of the correlation coefficients between the numerical variables in the dataset. Since the diagonal value of the table is the coefficient value of the variable and itself, which is always equal to 1, it can be ignored. Table 2 shows that these three variables—life expectancy at birth for men and women—are highly positively correlated. This response is to be expected given that the three variables are linked to and derived from one another.

	Life.Expectancy.at.Birth..Both.Sexes	Life.Expectancy.at.Birth..Males	Life.Expectancy.at.Birth..Females	Under.Age.5.Mortality..Both.Sexes	Under.Age.5.Mortality..Males	Under.Age.5.Mortality..Females
Life.Expectancy.at.Birth..Both.Sexes	1.000	0.993	0.993	-0.899	-0.898	-0.897
Life.Expectancy.at.Birth..Males	0.993	1.000	0.971	-0.879	-0.879	-0.875
Life.Expectancy.at.Birth..Females	0.993	0.971	1.000	-0.906	-0.903	-0.906
Under.Age.5.Mortality..Both.Sexes	-0.899	-0.879	-0.906	1.000	0.999	0.998
Under.Age.5.Mortality..Males	-0.898	-0.879	-0.903	0.999	1.000	0.993
Under.Age.5.Mortality..Females	-0.897	-0.875	-0.906	0.998	0.993	1.000

Table 2: Calculation of correlation coefficients

Using the pair of plots in Figure 5, which depicts a strong positive linear relation between male and female life expectancy, one might draw similar conclusions. On the other hand, a negative correlation coefficient indicates that the link between the under-five death rate and life expectancy is inverse or negative monotonic. This response shows an increase in life expectancy and a decrease in mortality among children under five, respectively. Figure 5 shows a line with a negative slope between under-five mortality

and life expectancy, with no sign of a curve, resulting in a negative linear relationship between the two.

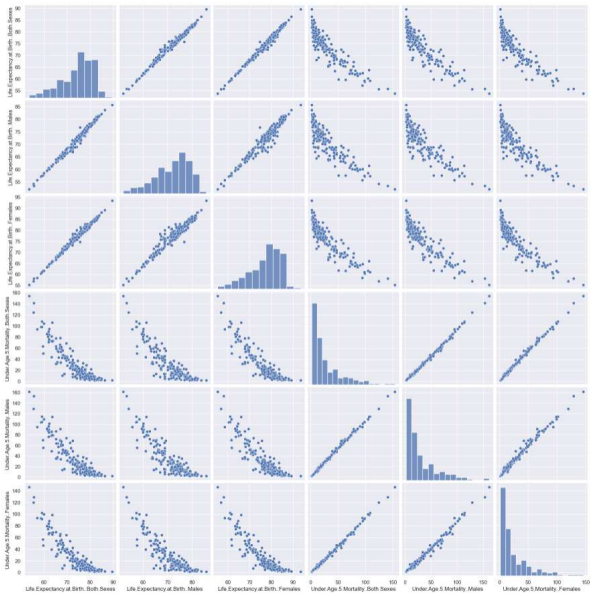

Figure 5: Bi-variate analysis of numerical data using pair-plot

4.4 Comparison of variables from 2002 to 2022

Table 3 shows the average change in variables over the past 20 years. It can be interpreted that the under 5 mortality rate for both sexes has declined from 49 to 26.6, under age mortality rate for females has declined from 24 to 45.5 and the under 5 mortality rate for males has declined from 29.2 to 52.4, but life expectancy has improved by 6 years. It should also be noted that the preceding phrase refers to the global population, not the population by region, subregion or country. We will now look at the same data by region. For this, we will use box plots to compare the results.

Year	Life.Expectancy.at.Birth..Both.Sexes	Life.Expectancy.at.Birth..Females	Life.Expectancy.at.Birth..Males	Under.Age.5.Mortality..Both.Sexes	Under.Age.5.Mortality..Females	Under.Age.5.Mortality..Males
2002	68.862	71.295	66.553	49.030	45.514	52.405
2022	74.578	77.182	72.097	26.677	24.012	29.234

Table 3: Variable change over 20 Years

As can be seen from Figure 6, the African region has seen a large increase in life expectancy, while other regions have seen much smaller increases. On the other hand, we have witnessed a significant decline in the under-five mortality rate in all regions except

Europe, where it has been relatively stable over the past 20 years. The following statement can also be made based on Figure 6 that the European average under 5 mortality rate is lower than the world's average under 5 mortality rate, while Africa's under 5 mortality rate is higher than the world's average under 5 mortality rate.

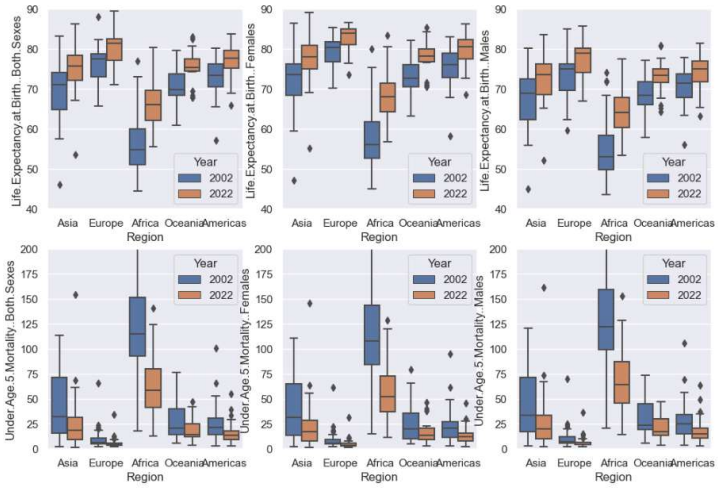

Figure 6: Comparison of change in variables in every region over year

5 Summary

A brief excerpt from the U.S. Census Bureau's IDB (International Data Base) served as the data set for this study. It consists of data on 227 nations, which are divided into 21 subregions and 5 regions according to their demographics. It has 10 features and 454 records. The provided data set is the full datum for 2002 and 2022.

This report's main objective is to give a thorough descriptive analysis of the data. The first step was to define and depict the frequency distribution for each variable using histograms, which led to the worldwide under-5 mortality rate for both sexes and the life expectancy at birth for both sexes. A scatterplot is used to explain the difference in life expectancy between the sexes, which shows that women have a higher average life expectancy at birth than men. Second, bivariate links between variables were found using correlation coefficients, revealing an inverse relationship between under-5 mortality and life expectancy. We also note regional and subregional differences in life expectancy in under-five mortality. Contrasted with other regions and subregions, Africa has considerable variability, while the Americas and Europe have less volatility. This response may be related to regional differences in lifestyle preferences and healthcare system standards. Finally, comparing the statistics for 2002 and 2022, it is found that the increase in life expectancy in 2022 is accompanied by a decrease in under-five mortality. However, this information is lacking since the population movement over time is not taken into consideration.

It would be more interesting and valuable for future studies if the population could also be taken into account for the yearly change in variables. It's more useful to know the elements that generate it, because it produces actual results.

Bibliography

International Data Base. *Glossary of census data*. United States Census Bureau, 2022. URL `https://www.census.gov/glossary/`. (visited on 2nd May 2022).

Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, Version: 3.9.7, 2021. URL `https://www.python.org`. (visited on 10th April 2022).

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

Appendix

A Additional figures

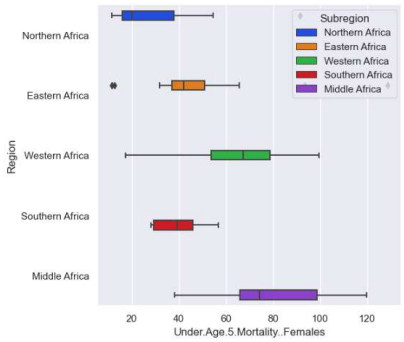

Figure 7: Under 5 Mortality Rate of African Females

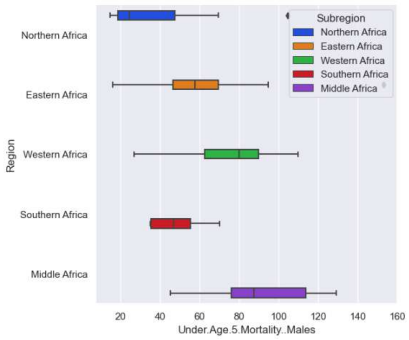

Figure 8: Under 5 Mortality Rate of African Males

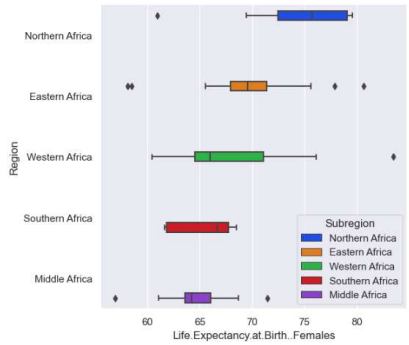

Figure 9: Life expectancy at birth for females

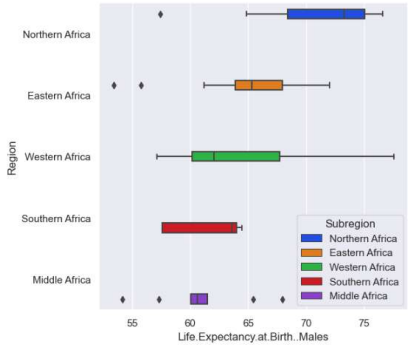

Figure 10: Life expectancy at birth for males

B Additional tables

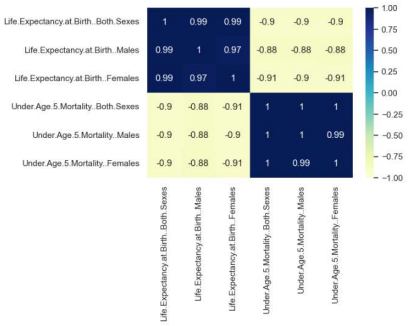

Table 4: Correlation Coefficient using Heatmap