TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Franziska Kappenberg

Author: Aishwarya Dinni

Group number: 16

Group members: Aishwarya Dinni, Samanvita Chormalle , Aijaz
Afzaal Ahmed and Yat Chun Fung

November 11, 2022

Contents

1 Introduction

Demographic data is a type of data that is statistically socio-economic in nature. Demographic analysis is the study of a population based on factors such as age, race, and sex. Governments and other organizations use demographics to learn more about the population's characteristics for many purposes. Demographic studies help us to understand existing trends and evolution, present distribution, and future implications of the social, economic, cultural, and political behavior within the population of an area. This is the key for public policy analysis, disaster management, and formulating welfare and developmental plans for a country. This case study analysis the demographic data provided by the International Data Base (IDB) of the U.S. Census Bureau using exploratory and descriptive methods. This database contains data such as total population, population with respect to age, sex, and also demographic characteristics such as fertility, mortality, and migration. The data is collected from many sources such as information from state institutions, censuses, surveys or administrative records, and also the estimates and projections by the U.S. Census Bureau. The main focus of this analysis is to use explorative and descriptive methods to analyze the demographic data, mainly the ~~mortality and fertility rate~~ among men and women, spread throughout various regions across the world between 2001 and 2021.

In this project, the data comparison is done using graphical methods, such as bar graph, boxplot, histogram, heat map, scatter plot, and also statistical measures like mean, median, variance, correlation. The main factors explored are:

- The comparison of life expectancy of females against that of males,

- Relationship between the life expectancy and the infant mortality rate and

- Changes in above mentioned features over the years and across different given regions.

Section 2, the problem statement consists of an illustration of the data set and the concerns examined. Whereas, section 3 includes briefs on the statistical methods such as mean, median, etc., and graphical methods including histogram, bar graph, and many others involved in this study. At last Section 4 includes the observations, summary, and conclusions on analyzing this data.

2 Problem statement

2.1 Data set description

The motive of this report is to analyze the given data set [taken from the International Data Base (IDB) of the U.S. Census Bureau(2021)]. The database contains accurate demographic measures and characteristics including population by sex and age, total population and fertility, mortality, migration for many countries respectively.

As earlier mentioned in the introduction, data collected from various sources is documented with cohort-component methodology. In this method, the components of population change (fertility, mortality, and net migration) are projected separately for each birth cohort (persons born in a given year).

The variables involved in this analysis are as follows:

- **Country.Name:** (Nominal data type) - This variable describes the 228 countries that are found in this data set.

- **Subregion:** (Nominal data type) - The above mentioned 228 countries are geographically categorized into 21 subregions.

- **Region:** (Nominal data type) - The contients/regions- Africa, Americas, Asia, Europe, Oceania are given by the variable "RegionName".

- **Year:** (Numeric data type) - As the considered data, in this case, is just for the years 2001 and 2021.

- **Life.Expectancy..Both.Sexes:** (Numeric data type) - displays the average number of years people are expected to live if they were born in the same year and assuming that mortality remains unchanged at each age in the future..

- **Life.Expectancy..Males:** (Numeric data type) - Life expectancy of males is given by this variable.

- **Life.Expectancy..Females:** (Numeric data type) - gives Life expectancy of females

- **Infant.Mortality.Rate..Both.Sexes** (Numeric data type) - The number of infants dying before reaching one year of age, per 1,000 live births in a given year.

- Note: Years are used to measure life expectancy.

Note: In the given data ~~seven~~ countries ~~including~~ Libya, Puerto Rico, South Sudan, Sudan, Syria, United States have missing values for the year 2001. These records are not considered in this analysis.

2.2 Objectives

The aim of this project is to execute a descriptive analysis of the census data. The data is analyzed using various statistical methods such as central tendency, spread, and some graphical representation methods.

The initial analysis is done using the data for the year 2021. Histograms represent the frequency distribution of the variable. Univariate analysis is done on both the sexes for infant mortality rate and life expectancy, then the outcome is compared. The correlation matrix is used to assess the bivariate correlation between the variables and then it is visualized using pairwise plots and heat maps. Across the subregions, variability within continuous numerical data is calculated and then is visualized and is analyzed using box plots.

Homogeneity within the subregions and heterogeneity between different subregions are determined for all the individual variable values. Ultimately, the changes in the values of the variables from 2001 to 2021 are examined and the trend is expressed using scatter plots.

3 Statistical methods

This statistical methods section includes many statistical measures and graphical methods. Later these are used for analyzing data set according to the issues examined. These ~~models~~ and methods are examined based on their attributes and assumptions. Initial examination of the given data in the csv file is done using Microsoft Excel(Version 2010) and this is done in order to understand the variables and their properties. For all calculations and visualizations, Python (Version 3.9.0) language is scripted with Jupyter Notebook (Version 6.4.5) launched using the software - Anaconda Navigator (Python Development Core Team (2020)). The packages utilized are pandas (development team (2020)), numpy (Harris and Millman (2020)), matplotlib.pyplot, matplotlib.lines (Hunter (2007)).

3.1 Statistical measures

3.1.1 Measures of central tendency

Measure of central tendency is a typical value for a sample. It may also be called a center or location of the distribution. Few of the common measures of central tendency are the arithmetic mean, the median, and the mode among many other. These measures can be defined as follows:

Arithmetic mean: It is also called as "mean" or "average" and is usually denoted as $\bar{x}$. This can be defined as sum of all the observations $(x_1 + x_2 + \cdots + x_n)$ divided by number of observations (n) (Christian Heumann, Michael Schomaker, Shalab (2017)). The formula for arithmetic mean of a variable x is -

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

Median: In simple terms median is the middle value in a ordered set of data. It separates the higher half from the lower half of the data set (Christian Heumann, Michael Schomaker, Shalab (2017)). For a sample $x_1, ..., x_n$ of size n arranged in ascending order, the formula for calculating the median is

$$\tilde{x} = \begin{cases} \frac{x_{(n+1)}}{2} & \text{- If n is an odd number} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2})+1}}{2} & \text{- If n is an even number} \end{cases}$$

3.1.2 Measures of variability

- **Variance:** The term variance denoted by s^2 or σ^2 refers to a statistical of dispersion in a data set, to be specific variance determines how far each number in the set is from the average (mean), and thus from every other number in the set (Christopher Hay-Jahans (2019)). Considering a sample of size n with observations $x_1, ..., x_n$, and $\bar{x}$ being the mean value of all observations. the variance is calculated as:

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

- **Standard Deviation:** is simply the square root of variance $\sqrt{\sigma^2}$. A standard deviation with a low value indicates that the values tend to be close to the mean

of the set, while a high standard deviation indicates that the values are spread out over a wider range.

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

(Christopher Hay-Jahans (2019)).

3.1.3 Pearson correlation coefficient

Pearson correlation coefficient is also called as bivariate correlation and is denoted by r, it can be defined as a measure of linear correlation between two continuous variables. Correlation defines the direction and strength of a linear relationship that exists between pair of variables. (Christopher Hay-Jahans (2019)) In the Pearson correlation coefficient, the results values are always between -1 and 1.

- -1 and values closer to -1 show that there is a negative correlation between the variables,

- 0 denoting that there is no correlation. There might still be a relation, but not in terms of a positive or negative correlation,

- +1 and values closer to +1 depicts that there is a strong positive correlation,

- Considering two continuous random variables x and y, which have samples $x_1, ..., x_n$ and $y_1, ..., y_n$, mean $\bar{x}$ and $\bar{y}$ respectively with sample size n, the correlation coefficient r is defined as
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

3.1.4 Range, quartile and inter-quartile range

- **Range:** The range of a set of data is the difference between the maximum and minimum values. Range $R = x_{max} - x_{min}$. (Christopher Hay-Jahans (2019))

- **Quartile:** Based on values of the data, the observations are divided into four intervals and how they compare to the total set of observations. Assuming that the set of observations are arranged in their ascending order, the quartiles are represented as -

- First Quartile: also called as lowest quartile, is the 25th quartile, meaning that 25% of the data falls below this quartile. $Q1 = (\frac{(n+1)}{4})^{th}$ value.

- Second Quartile (or the median), is the 50th quartile, meaning that 50% of the data falls below this quartile. $Q2 = (\frac{(n+1)}{2})^{th}$ value.

- Third Quartile, also called as lower quartile, is the 75th quartile, meaning that 75% of the data falls below this quartile. $Q3 = (\frac{3(n+1)}{4})^{th}$ value.

- Note: If $(\frac{(n+1)}{2})$ is not an integer then the result is rounded

- **Interquartile Range:** Other names for Interquartile Range are midspread, middle 50%, or H-spread. The interquartile range is defined as the difference between the 75th and 25th quartiles i.e. $IQR = Q_3 - Q_1$. This gives us 50% of the observations that lie in the range of middle half of a data. (Christian Heumann, Michael Schomaker, Shalab (2017))

3.2 Statistical plots

Statistical plots represent the statistical analysis of the data set. Bar graphs, histograms, box plots, scatter plots are a few of the statistical plots among many others.

3.2.1 Histograms

A histogram is a graphical plot that represents the distribution of the numerical data. It illustrates the count or frequency of defined values of a univariate feature. The observed variable values are plotted on the x-axis and their counts are plotted on the y-axis. The frequency of each value is represented by the height of each respective bin (Christian Heumann, Michael Schomaker, Shalab (2017)).

3.2.2 Box plots

A boxplot is a graphical method in which the quartiles depict their groups of numerical data. Sometimes it is also termed as box-and-whisker plot and box-and-whisker diagram as they may also have lines extending from the boxes are called whiskers, which indicates variability outside the upper and lower quartiles. The data points plotted outside the whiskers are known as outliers.

A boxplot is considered to be a standardized way of representing the dataset based on a five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles. The interquartile range is determined by the length of the box, and the median is determined by the line in the middle of the box (Christopher Hay-Jahans (2019)).

3.2.3 Scatter Plot

It is also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram. A scatter plot is a type of mathematical diagram or plot that uses cartesian coordinates to represent values for typically two quantitative variables for a set of data. The points can also be coded (color/shape/size) to differentiate multiple variables. In simple terms, this graph uses dots to represent values for two different numeric variables (Howard J. Seltman (2012)).

3.2.4 Pairwise Plot

Pairwise plots, also termed matrix scatter plots, are used to represent the correlation (relationship) between pair of variables. This plot is used to generate a plot matrix. The set of variables considered are the same in both the horizontal and vertical axes. For each subplot in the matrix, the values of the variables in the x-axis are plotted against the values of the variable in the y-axis respectively.

4 Statistical analysis

This section includes the detailed analysis of the given data set using the statistical methods mentioned earlier in the project.

4.1 Univariate analysis

Univariate analysis is the simplest form of analyzing data. It takes data, summarizes that data, and finds patterns in the data. The univariate frequency distribution for each of the continuous variables such as LEBS, LEM, LEF and IMRBS from the year 2021 is analyzed in this section. Life expectancy across all the regions and sub-regions is visualized for males against females.

Figure 1: Frequency distribution histograms for <u>different variables</u>

~~The given~~ Figure 1 shows the histograms for the frequency distribution of the variables: Infant.Mortality.Rate..Both.Sexes, Life.Expectancy..Both.Sexes, Life.Expectancy..Males, Life.Expectancy..Females and further, <u>the female life expectancies is subtracted from the male life expectancies and the frequency distribution of this difference is also plotted as a histogram with the title Life.Expectancy..Differences.</u>

The frequency distribution for the first chart with the title Infant.Mortality.Rate..B<u>oth.Sexes,</u> is <u>skewed to the right</u> and for the rest of the ~~charts,~~ it is ~~skewed to its left~~. In the first graph, the infant mortality rate is centered between 1 and 5. The maximum is at 144.77 for Afghanistan. One can also observe that in the second graph, life expectancy for both the sexes is <u>centered</u> in between 75.0 and 78.0 and Monaco has 89.4, the maximal value among the given data.

The total life expectancy for males is centered between 74.0 and 75.5, which is less than that of both the sexes. ~~t~~he maximum years lived by males is 85.55 years, which is less than that of both the sexes. On the other hand for females, it can be observed that it

is centered between 78.0 and 80.0, which is more than that of the males and is similar to that of both the sexes. The value of total life expectancy of females peaks at around 93.4. This again, more than the maximum value for males and is also more than total life expectancy for both the sexes.

The last histogram graph depicts the plotted values for the difference between female life expectancies and male life expectancies. As it can be noticed that most of the values are negative, which means the life expectancy of females is comparatively higher. Most differences here are found between -4.0 and -4.5. This implies that in most countries, the life expectancy of females is about four to four and half years higher than that of males.
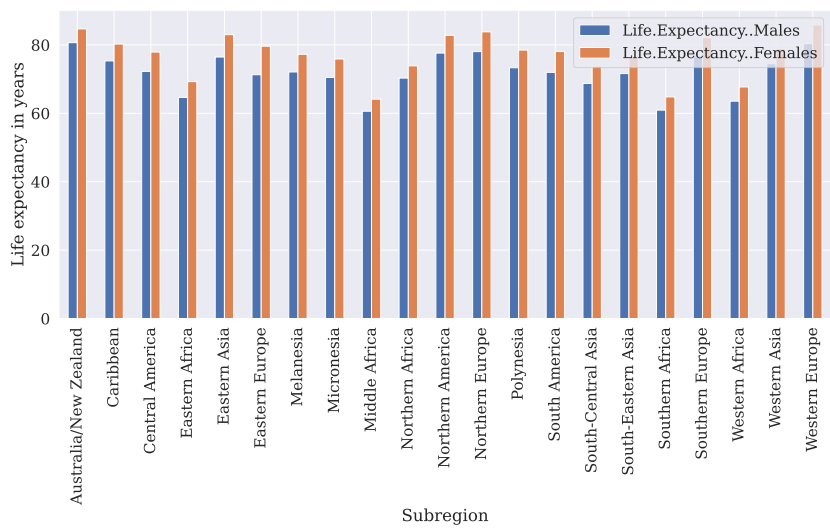

Figure 2: Bar chart of life expectancy at birth for males and females

To analyze more about life expectancy difference between male and female, a grouped bar chart is prepared. Figure 10 displays, the life expectancies of males and female grouped together based on region and Figure 2 shows the life expectancies of males and females are grouped based on subregion. Both the graphs have two bars, the blue-colored bar represents the life expectancy at birth in years for females and the other for males.

Analyzing both the grouped bar graphs, it is sure that the life expectancy of females is higher than that of males within all the subregions and also within all the regions.

4.2 Bivariate analysis

This section consists of the correlation between the variables. The bivariate analysis determines the relationship between the variables.

Table 1: Correlation matrix for the variables.

	IMRBS	LEBS	LEM	LEF
IMRBS	1.000	-0.905	-0.883	-0.913
LEBS	-0.905	1.000	0.993	0.993
LEM	-0.883	0.993	1.000	0.971
LEF	-0.913	0.993	0.971	1.000

The abbreviations used in the table are as follows:
IMRBS - Infant.Mortality.Rate..Both.Sexes, LEBS- Life.Expectancy..Both.Sexes, LEM - Life.expectancy..Males, LEF - Life.Expectancy..Females. The table 1 shows the correlation coefficients for every pair of continuous numeric variables and this can be plotted in scatter plots as given in the figure 3.

Analyzing the table 1 and the figure 3 we can determine that there's a positive correlation of 0.97 that exists between the male and female life expectancy. As we know that positive r values indicate a positive correlation, where the values of both variables tend to increase together. So, we can conclude that as the life expectancy of females increases the life expectancy of males also increase and vice versa.

There exists a linear relationship between the life expectancy of females at birth and the life expectancy of both the sexes, which is positive correlation 0.993 and it remains the same between the life expectancy of males at birth and the life expectancy of both the sexes. Meaning that the higher the life expectancy for both the sexes, the higher is the life expectancy for males and also for females.

Now comparing the relation between infant mortality rate and three other variables, Life.Expectancy..Both.Sexes, Life.Expectancy..Females, Life.Expectancy..Males, we can see the correlation value is -0.90, -0.88, -0.91 respectively. As we know that negative correlation values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.
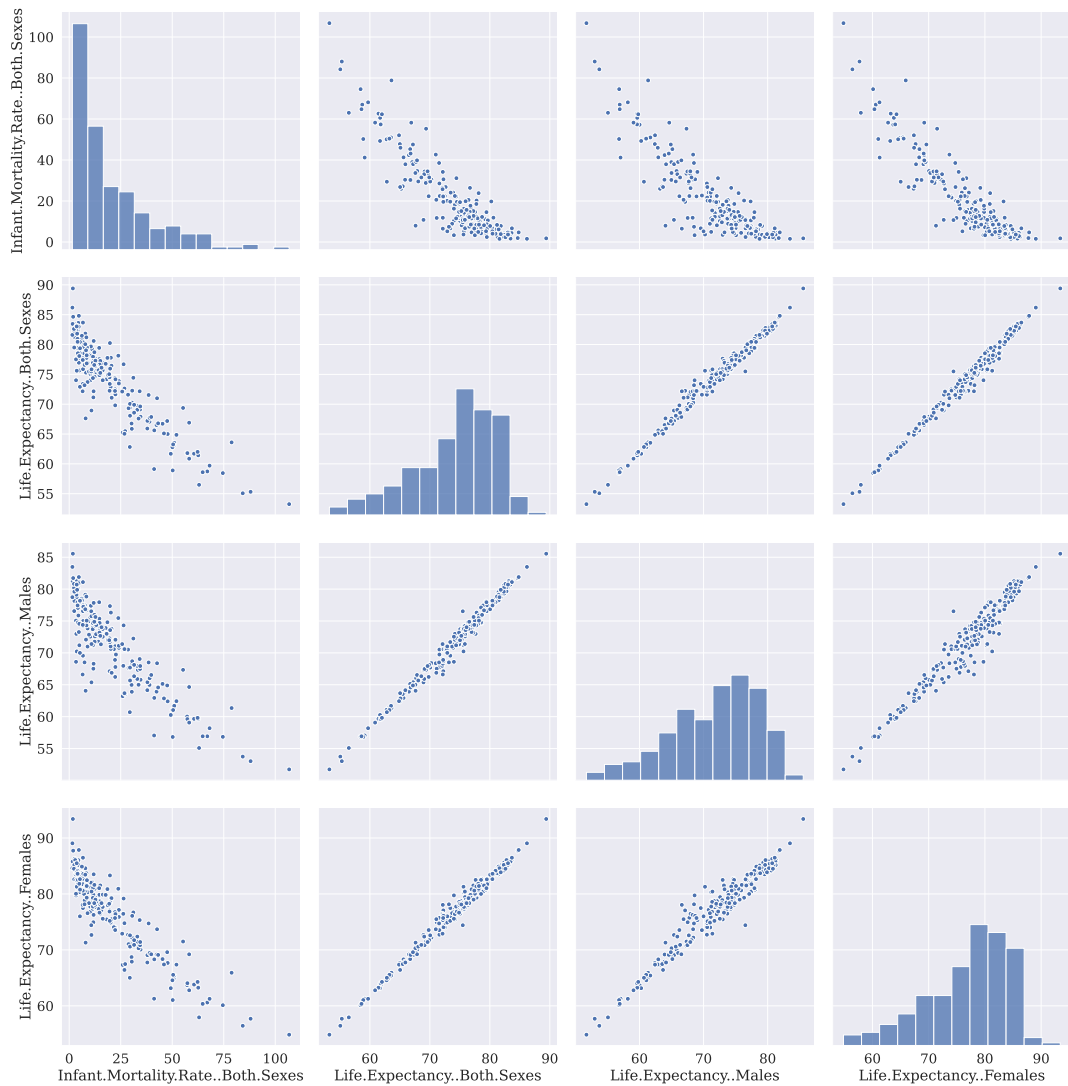
Figure 3: Correlation matrix for the variables

4.3 Variability analysis

This subsection focuses on the variability or the spread of the variables with observations from the year 2021. The median and the variability for each of the variables are obtained using the box plots. The figures 4, 6, 7 and 8 have lines extending from the boxes, these are called as whiskers. These indicate variability outside the upper and lower quartiles. The data points (small diamond-shaped points) plotted outside the whiskers are known as outliers.

In Figure 4 infant mortality rate is plotted against all the subregions. As one can see the African subregions display a higher variability in their infant mortality rates.
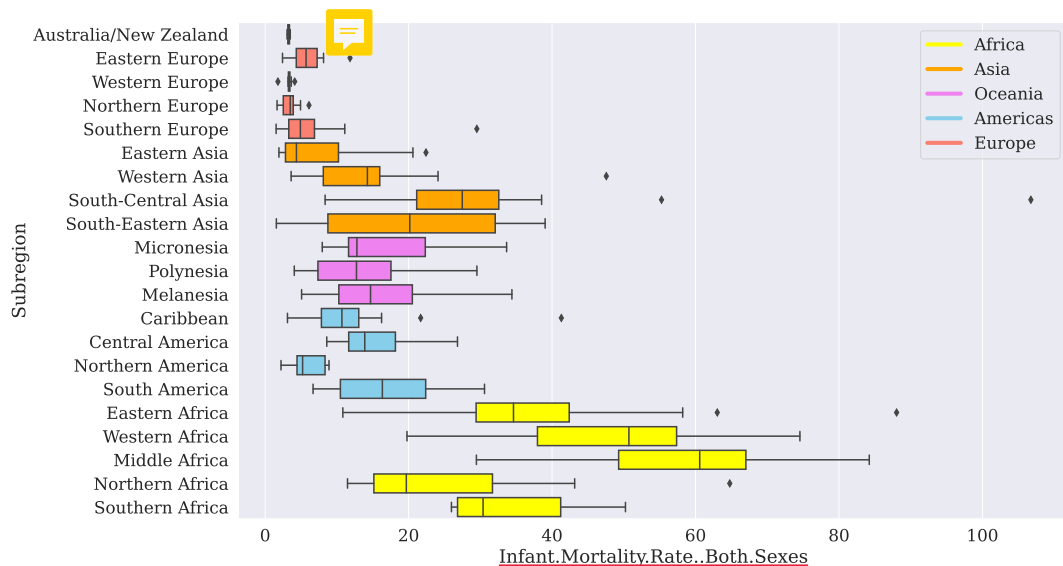
Figure 4: Box plot showing variability in infant mortality rate across the sub-regions

The medians of these sub-regions are spread out within a region, this shows that the infant mortality rate is heterogeneous between subregions. The subregions of Asia also show higher variability when compared to the other three subregions and comparatively subregions of Oceania and America show minimum variability. The subregions of Europe show minimal variability compared to subregions of other regions. The medians of
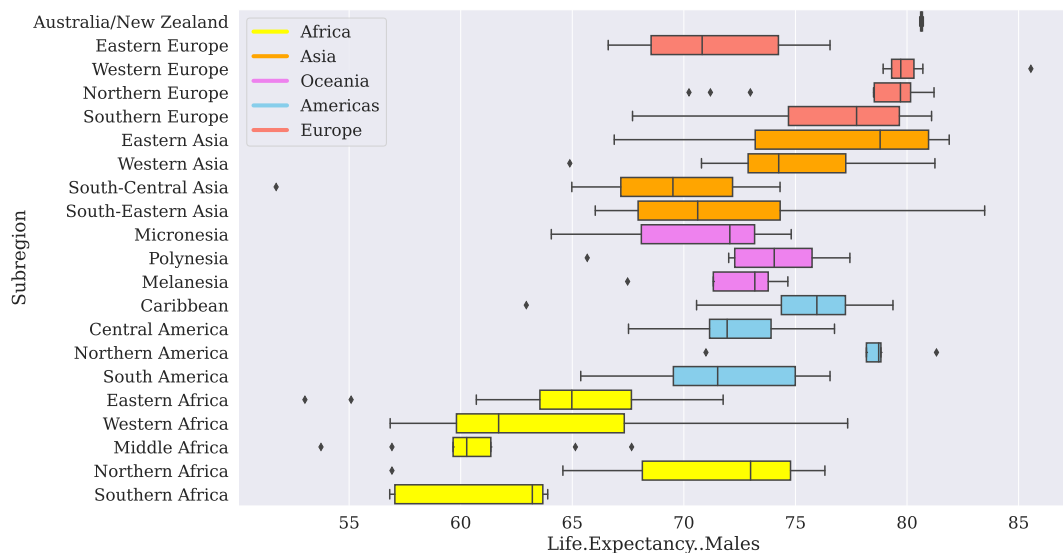

Figure 5: Box plot showing variability in Life.Expectancy..Male across the sub-regions

European subregion are very close by we can say that they are homogeneous within their sub-regions. The lowest infant mortality rate in the region Europe is 1.53 belonging to

Southern Europe and ~~T~~he highest infant mortality rate in the region of Asia is 144.77, South-Central Asia.
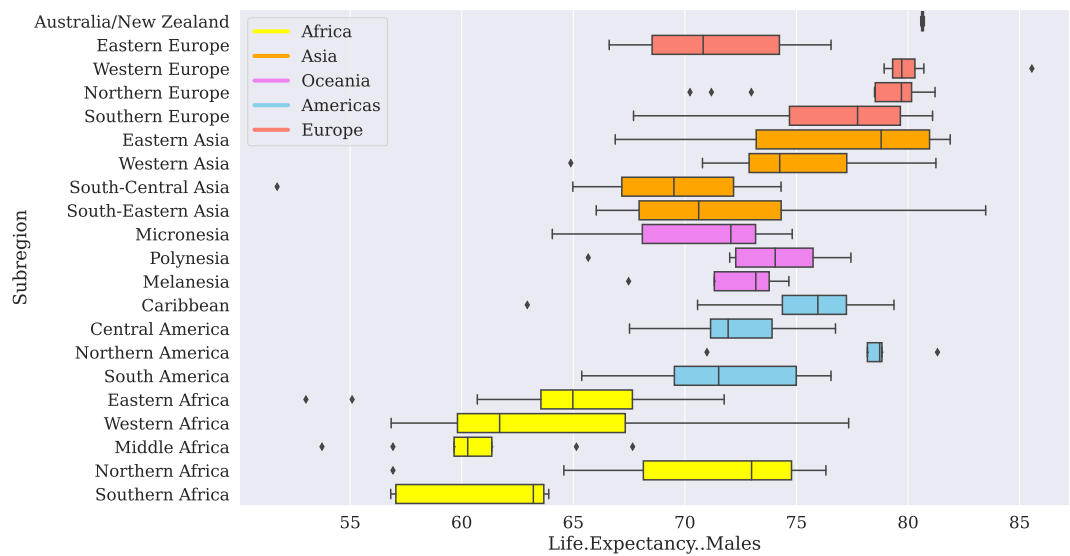

Figure 6: Box plot showing variability in Life.Expectancy..Male across the sub-regions
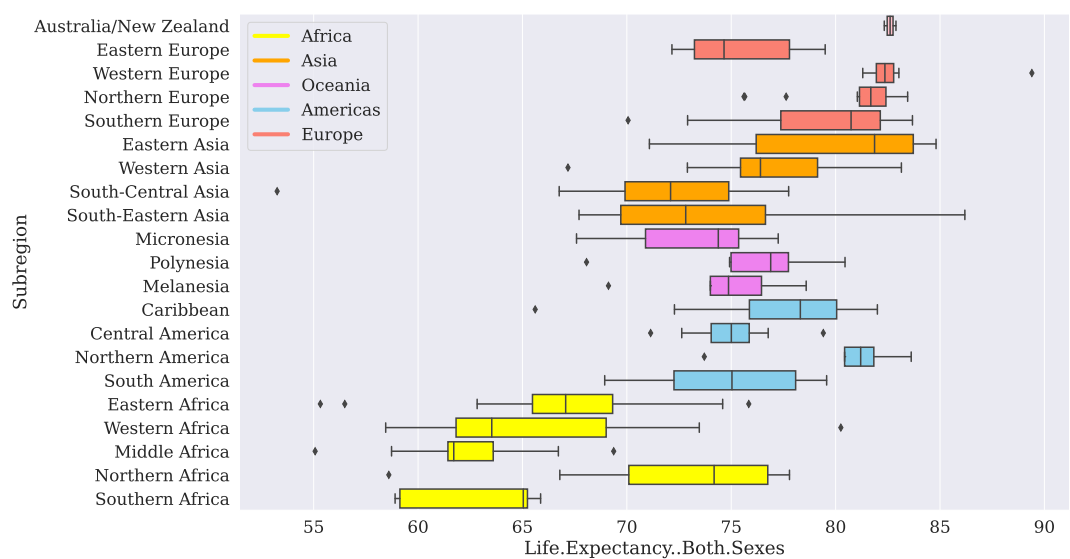

Figure 7: Box plot showing variability in Life.Expectancy..Both.Sexes across the sub-regions

The Figures 6, 7 and 8 displays variability in life expectancy of males,females and both sexes respectively plotted ~~against the~~ subregions. As we have already seen that there is a linear relationship between the life expectancy for both the sexes and the life expectancy

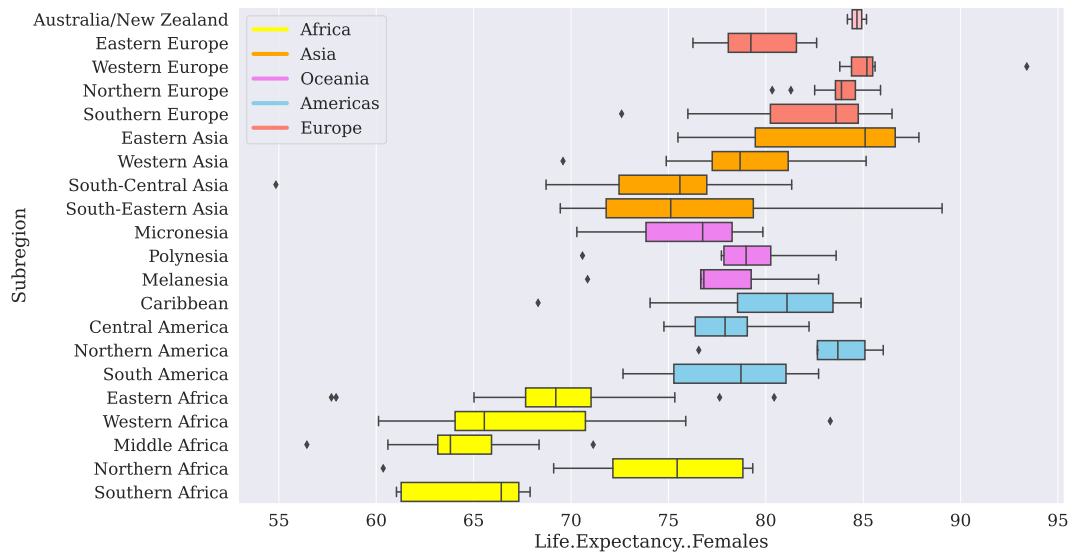of males and females. Thus, the box plot for both the sexes is ~~same as~~ that of the individual sexes.

Figure 8: Box plot showing variability in life.Expectancy..Females across the sub-regions

Subregions like Northern Africa, Eastern Asia show the highest variability for both males and females and are highly heterogeneous within their sub-regions. This is the same in both sexes' plots as well. On the other hand, subregions like Australia/New Zealand, Western and Northern Europe show very little variability for both males and females and are highly homogeneous within their subregions. This is the same in both sexes' plots as well.

4.4 Trend analysis

The main focus of this section is to collect the information over a longer period and attempt to find a pattern. So, the data from the past 20 years is collected and is analyzed. As mentioned at the start of the project that values for a few countries are missing for the year 2001 and are not included in the analysis to avoid inconsistency.

The figure 11, is the scatter plot that displays the values of each variable for the year 2001 which is plotted against the values for the year 2021. As one can see the values for the year 2001 are plotted on the x-axis and the values for the year 2021 are plotted on the y-axis. For each of the plots the identity $f(x) = y$ line is drawn to visualize the trend.
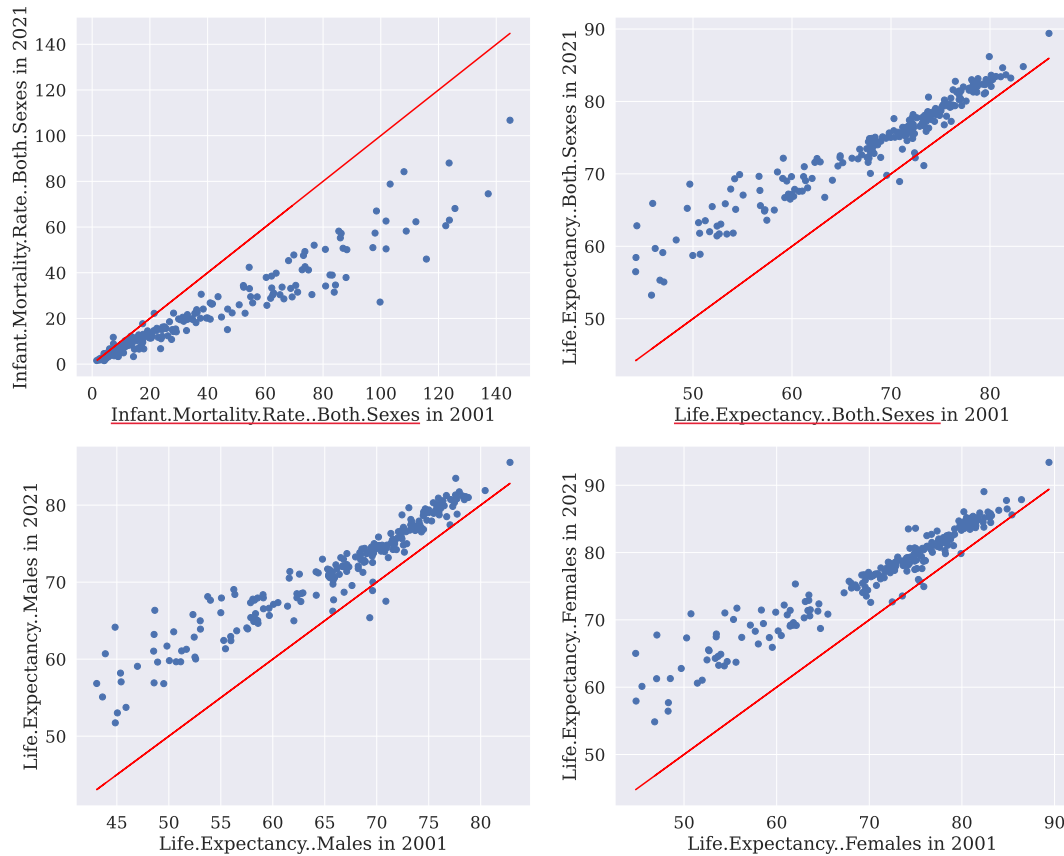
14

Figure 9: Scatter plots showing <u>trend</u> of the <u>variables</u> ~~from~~ the year 2001 ~~to~~ 2021

For lower infant mortality rates, the values in 2021 are quite close to the values in 2001. For larger infant mortality rates in the year 2001, there ~~seems to be~~ a lower mortality rate for the year 2021. <u>The extreme value of infant mortality rate in the year 2001 is 144.77 but decreases to 106.75 in 2021, belongs to the country Afghanistan</u>. Likewise, for Mozambique, the mortality rate in 2001 is 123.77 whereas, in 2021, it is around 63.03. To sum up, the infant mortality rate displays a decreasing trend from the year 2001 to <u>2021</u>.

After analyzing the given data and their correlations, we can say that the trend for all the three life expectancy variables seems alike. The values of the life expectancy variables in 2021 are higher than those in 2001. Generally, there is an increase in the overall life expectancy for all the subregions and sexes within time, also the values appear to be closer for higher life expectancies.

5 Summary

~~As mentioned earlier in the introduction,~~ the data considered in this project is a subset of the demographic data which is provided by the International Data Base (IDB) of the U.S. Census Bureau <u>using exploratory and descriptive methods</u>. The database consists of factors like population characterized by their age, sex, etc., and also the demographic characteristics such as mortality, migration, etc. The IDB is timely, precise, and is always kept updated for research, program planning, and other decision-making throughout the <u>globe</u>.

This project mainly focuses on the infant mortality rate and life expectancy between males and females of all the ~~228~~ countries, which are grouped into 21 subregions and 5 regions. <u>Examined data is between 2001 and 2021</u>. <u>Graphical and statistical measures were used to analyze the data</u>. It is very important to note that the infant mortality rate and life expectancy of both the sexes are closely related. Since the year 2001, the infant mortality rate has drastically reduced across most of the regions and on the other hand, the life expectancy is increased for both sexes.

As per the World Health Organization the life expectancy around the globe has increased by more than 6 years within the span of 2000 and 2019(from 66.8 years to 73.4 years). Along with the life expectancy, the While life expectancy abbreviated as "HALE" has also increased by 8%~~.~~ (from 58.3 years to 63.7 years) within the same span. According to the WHO the reason for this increase is due to a decline in mortality rate~~.~~ <u>WHO contributors (2020)</u>

Bibliography

Christian Heumann, Michael Schomaker, <u>Shalab.</u> Introduction to statistics and data analysis. pages 37–66, 2017. ISSN 978-3-319-46160-1. ~~URL~~ ~~https://link.springer.~~ ~~com/chapter/10.1007/978-3-319-46162-5_3~~. ~~[Online; accessed 10-November-~~ ~~2022]~~.

Christopher Hay-Jahans. R companion to elementary applied statistics. 2019. ISSN 9780429827273. ~~URL~~ ~~https://books.google.de/books?id=8UOCDwAAQBAJ~~. ~~[Online;~~ ~~accessed 10-November-2022]~~.

Pandas development team. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`. [Online; accessed 10-November-2022].

Charles R. Harris and K. Jarrod Millman. Array programming with NumPy. *Nature*, 585 (7825):357–362, 2020. ~~URL~~ ~~https://doi.org/10.1038/s41586-020-2649-2~~. ~~[On-~~ ~~line; accessed 10-November-2022]~~.

Howard J. Seltman. Experimental design and analysis. 153, 2012. ISSN 0000000120970344. ~~URL~~ ~~https://books.google.de/books?id=8UOCDwAAQBAJ~~. ~~[On-~~ ~~line; accessed 10-November-2022]~~.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. ~~[Online; accessed 10-November-~~ ~~2022]~~.

Python Development Core Team. *Python: A Language and Environment for Statistical Computing*, 2020. [Online; accessed 10-November-2022].

WHO contributors. Variance — WHO, life expectancy and healthy life expectancy. `https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy`, 2020. [Online; accessed 10-November-2022].

Appendix

A Additional figures

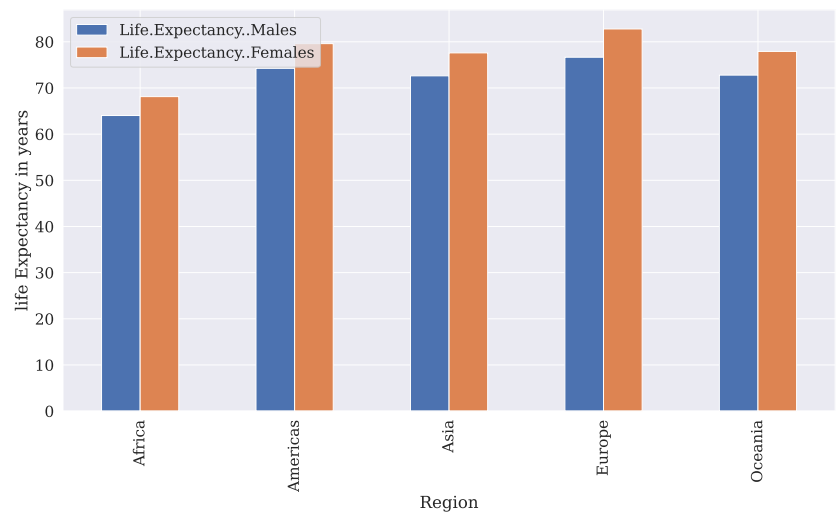

Figure 10: Bar chart of life expectancy at birth for males and females

Figure 11: Heatmap explaining bivariate correlation

B Additional tables

Table 2: Inter-quartile regions of each variable based on the sub-regions.

subregion	IMRBS	LEBS	LEM	LEF
Australia/New Zealand	0.225	0.280	0.080	0.490
Caribbean	5.210	4.180	2.870	4.900
Central America	6.512	1.822	2.755	2.670
Eastern Africa	12.980	3.840	4.100	3.360
Eastern Asia	7.390	7.525	7.767	7.185
Eastern Europe	2.910	4.552	5.705	3.497
Melanesia	10.250	2.450	2.460	2.590
Micronesia	10.700	4.460	5.075	4.420
Middle Africa	17.740	2.170	1.690	2.760
Northern Africa	16.535	6.655	6.640	6.675
Northern America	3.910	1.400	0.650	2.440
Northern Europe	1.390	1.267	1.627	1.025
Polynesia	10.185	2.750	3.465	2.405
South America	11.893	5.828	5.463	5.758
South-Central Asia	11.445	4.967	5.007	4.505
South-Eastern Asia	23.330	6.930	6.365	7.555
Southern Africa	14.380	6.110	6.630	6.040
Southern Europe	3.585	4.765	4.970	4.513
Western Africa	19.400	7.190	7.520	6.690
Western Asia	7.870	3.690	4.375	3.900
Western Europe	0.210	0.830	1.000	1.090