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, Aijaz Ahmed Afzal, Samanvita Chormalle , Yat Chun Fung

November 10, 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 fertility 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).

As the IDB's website is in agreement with Information quality standards, which consists of guidelines for statistical quality standards, information quality, and procedures to correct inaccurate information. IDB permits census.gov to access the records since 1996.

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:** (Binary data type) - As the considered data, in this case, is just for the years 2001 and 2021, the data type for the variable 'year' is binary.
- **Life.Expectancy..Both.Sexes:** (Numeric data type) - is the average number of children born per woman, considering that all women have survived through their whole child-bearing phase and have given birth at the specified set of age-specific fertility rates.
- **Infant.Mortality.Rate..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
- 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 fertility 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

citations or just type them 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 3.7.3)

language is scripted with Jupyter Notebook (Version 6.4.5) launched using the software - Anaconda Navigator. The packages utilized are pandas, numpy,matplotlib.pyplot, matplotlib.lines and seaborn.

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) ?. The formula for arithmetic mean of a variable x is -

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

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 ?. 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. 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.

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

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.? 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 $R = x_{max} - x_{min}$. this shoud comeChristopher Hay-Jahans, 2019, p. 77
- **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 25^{th} quartile, meaning that 25% of the data falls below this quartile.

$$Q1 = (\frac{(n+1)}{4})^{th}\text{value}$$

- Second Quartile (or the median), is the 50^{th} quartile, meaning that 50% of the data falls below this quartile.

$$Q2 = (\frac{(n+1)}{2})^{th}\text{value}$$

- Third Quartile, also called as lower quartile, is the 75^{th} quartile, meaning that 75% of the data falls below this quartile.

$$Q3 = (\frac{3(n+1)}{4})^{th}\text{value.}$$

- 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 the third quartile (Q3) minus the first quartile (Q1). This gives us the range of the middle half of a data set. i.e. $IQR = Q_3 - Q_1$?

3.2 Statistical plots

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

3.2.1 Grouped Bar Chart

Grouped Bar Chart is also called a clustered bar chart, a multi-series bar chart. In this numeric values are plotted against the levels of two categorical variables instead of one. This chart is usually used to compare different categories of two or more groups. The bars have equal width and can be positioned horizontally or vertically. If the bar chart is vertical, then the horizontal axis shows different categories and the other axis shows the frequencies as a measure of the height of the bars.

3.2.2 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.

3.2.3 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.

3.2.4 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 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.

3.2.5 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 TFR, LEBS, LEM, LEF from the year 2021 is analyzed in this section. Also, life expectancy at birth across all the regions and sub-regions is visualized for males against females.

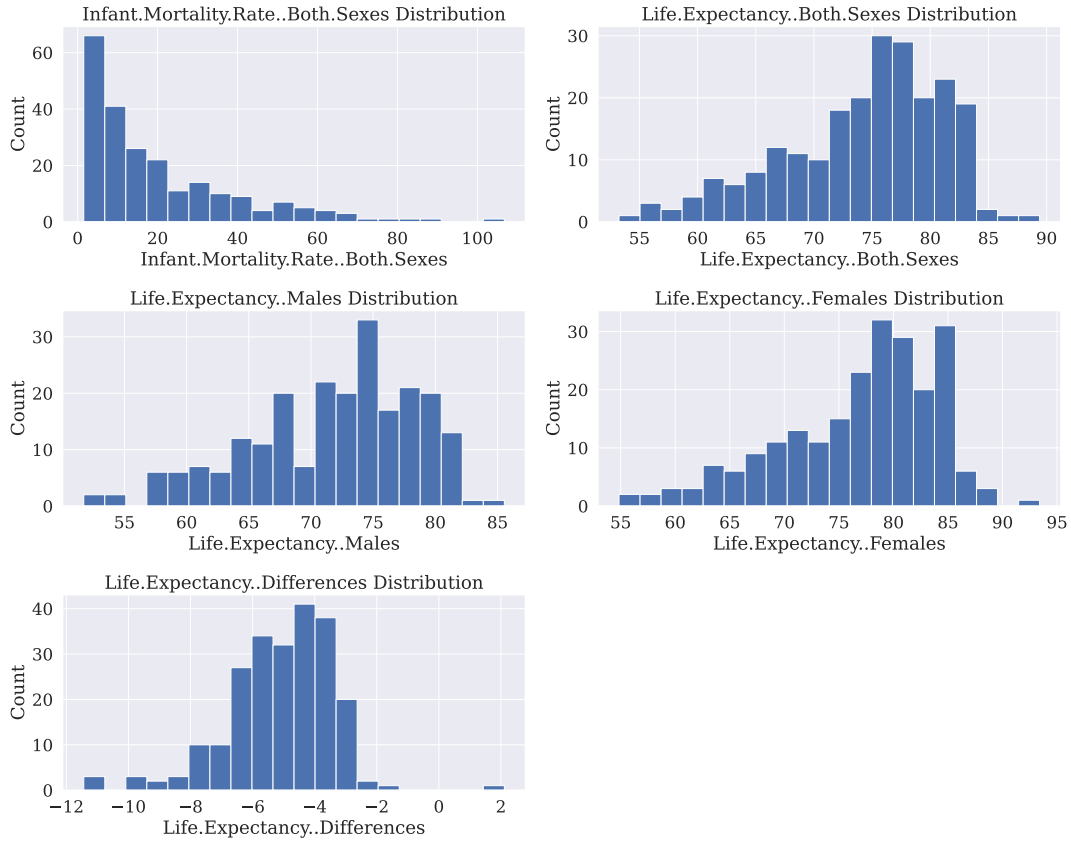

Figure 1: Frequency distribution histograms for different variables

The given Figure 1, shows the histograms for the frequency distribution of the variables: Total.Fertility.Rate, Life.Expectancy.Both.Sexes, Life.Expectancy.Males, Life.Expectancy.Females and further, 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.Expection.Differences.

The frequency distribution for the first chart with the title Infant.Mortality.Rate..Both.Sexes, is skewed to the left and for the rest of the charts, it is skewed to its right. In the first graph, the total fertility rate is centered between 0 and 5. The maximum is at 144.77 for Afghanistan country. One can also observe that in the second graph, life expectancy for both the sexes is centered in between 75.0 and 78.0 and Monaco has 89.4, the maximal value among the given data.??????????????????????check the values ??????????????????

The total life expectancy for males is centered between 74.0 and 75.1, which is less than that of both the sexes. the maximum years lived by males is 85.5 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 figure is the histogram the plotted values are the difference between female life expectancies and male life expectancies. As we 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.

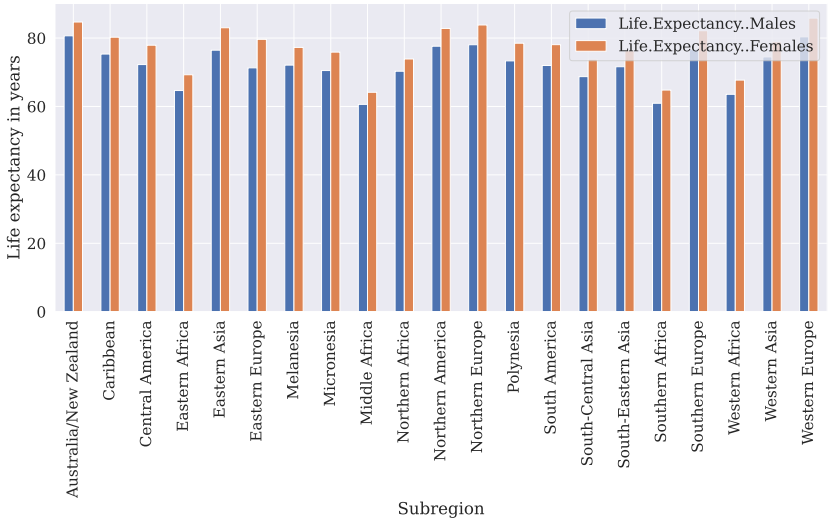

Figure 2: Bar chart of life expectancy at birth for males and females

To analyze more about life expectancy difference between male and female, a grouped bar chart is prepared. Figure 9 shows, the life expectancies of males and females are grouped based on region and Figure 2 shows the life expectancies of males and females are grouped based on sub-region. These 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.

Analysing 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, the correlation between the variables. As we know that the bivariate analysis determines the relationship between the variables. This is done using the tables and pairplots.

	Infant.Mortality.Rate..Both.Sexes	Life.Expectancy..Both.Sexes
Infant.Mortality.Rate..Both.Sexes	1.000	-0.905
Life.Expectancy..Both.Sexes	-0.905	1.000
Life.Expectancy..Males	-0.883	0.993
Life.Expectancy..Females	-0.913	0.993

The abbreviations used in the table are as follows: TFR- Total.Fertility.Rate, 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 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 total fertility and three other variables, life.expectancy.both.sexes, life.expectancy.females, life.expectancy.males, we can see the correlation value is -0.80,

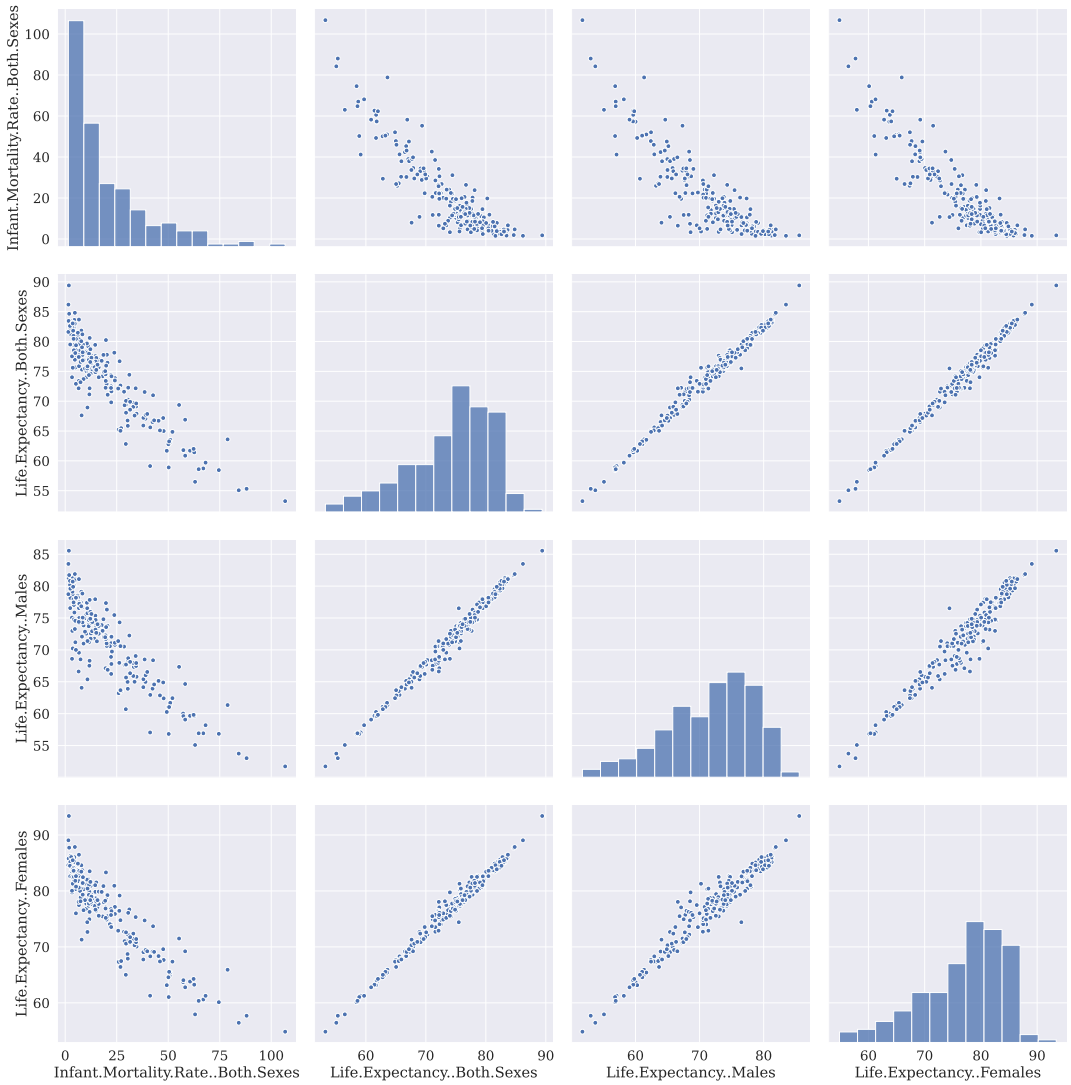

Figure 3: Correlation matrix for the variables

-0.77, -0.81 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. **?**

4.3 Variability analysis

This subsection focuses on the variability or the spread of the variables with observations from the year 2020. The median and the variability for each of the variables are plotted using the box plots. The figures 4, 5 and 7 have lines extending from the boxes are

whiskers, which indicates variability outside the upper and lower quartiles. The Data points (small diamond-shaped points) plotted outside the whiskers are known as outliers.

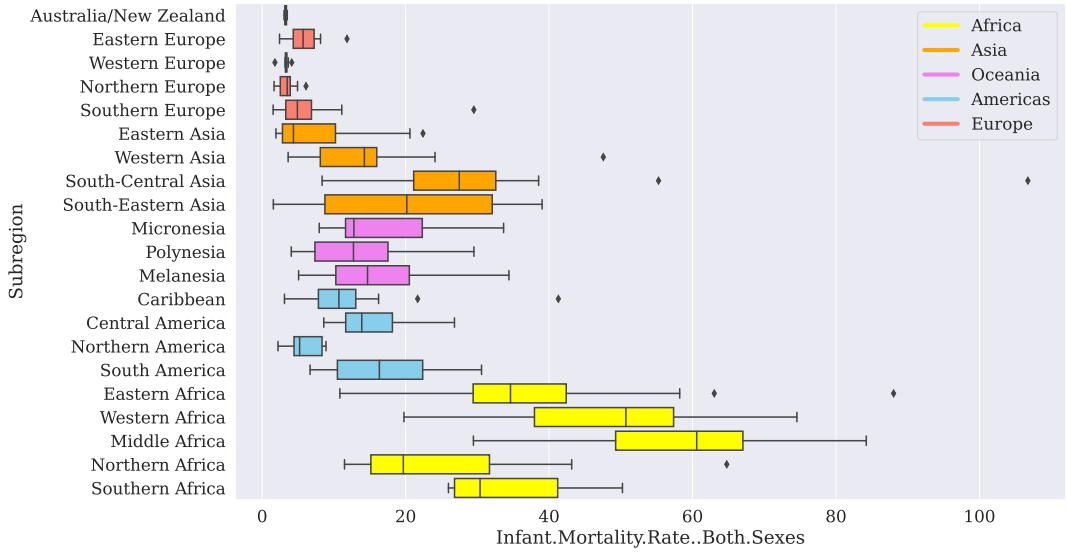

Figure 4: : Box plot showing variability in total fertility rate across the sub-regions

In Figure 4 total fertility rate is plotted against all the subregions. As one can see the African subregions display a higher variability in their fertility rates. The medians of these sub-regions are spread out within a region, this shows that the total fertility 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 this subregion are very close by we can say that they are homogeneous within their sub-regions. The lowest fertility rate in the region Africa is 56 belonging to Middle Africa and The highest fertility rate in the region of Europe is 89, South-Eastern Asia.

The Figures 5, 7 and 6 displays variability in life expectancy at birth for males,females and for both sexes is plotted against the sub-regions respectively. As we've 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 synonymous with that of the individual sexes.

Sub-regions like Northern Africa, Eastern Asia show the highest variability for both males and females and are highly heterogeneous with their sub-regions. This is the same in both sexes' plots as well. On the other hand, Sub-regions like Australia/New Zealand,

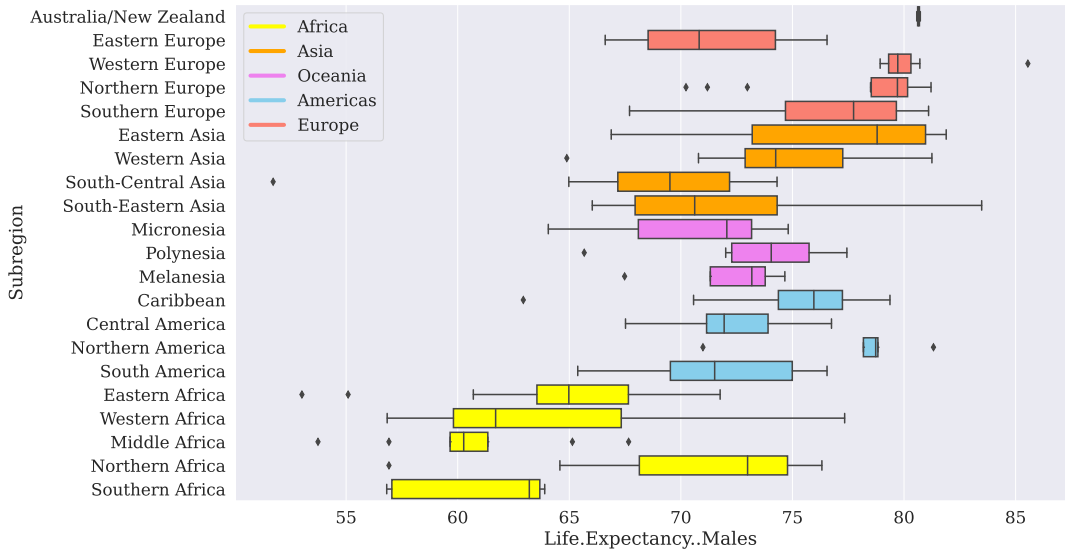

Figure 5: : Bar chart of life expectancy at birth for males

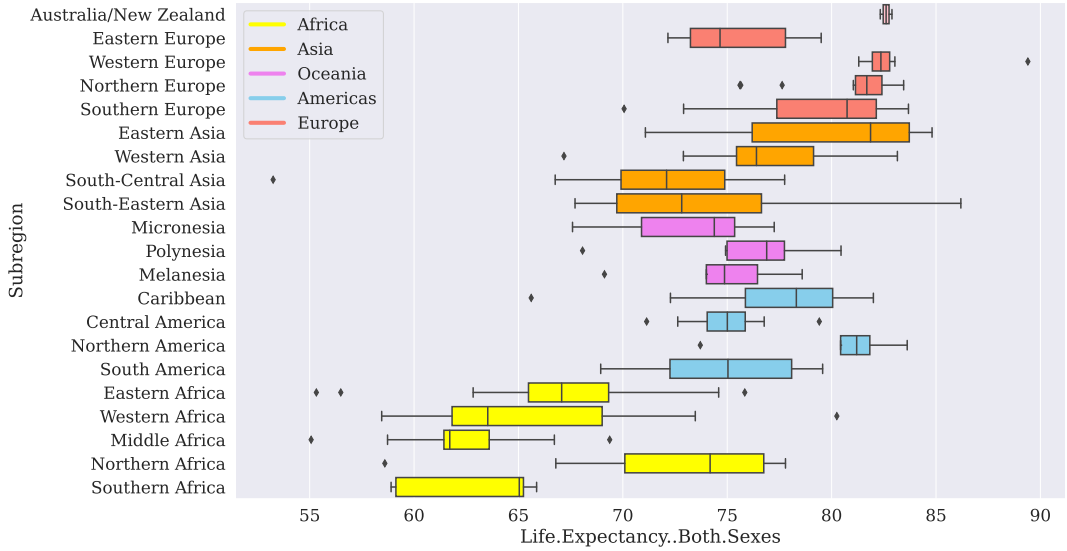

Figure 6: : Box plot showing variability in life expectancy at birth for both sexes the sub-regions

Western and Northern Europe show very little variability for both males and females and are highly homogeneous with their sub-regions. This is the same in both sexes' plots as well.

?

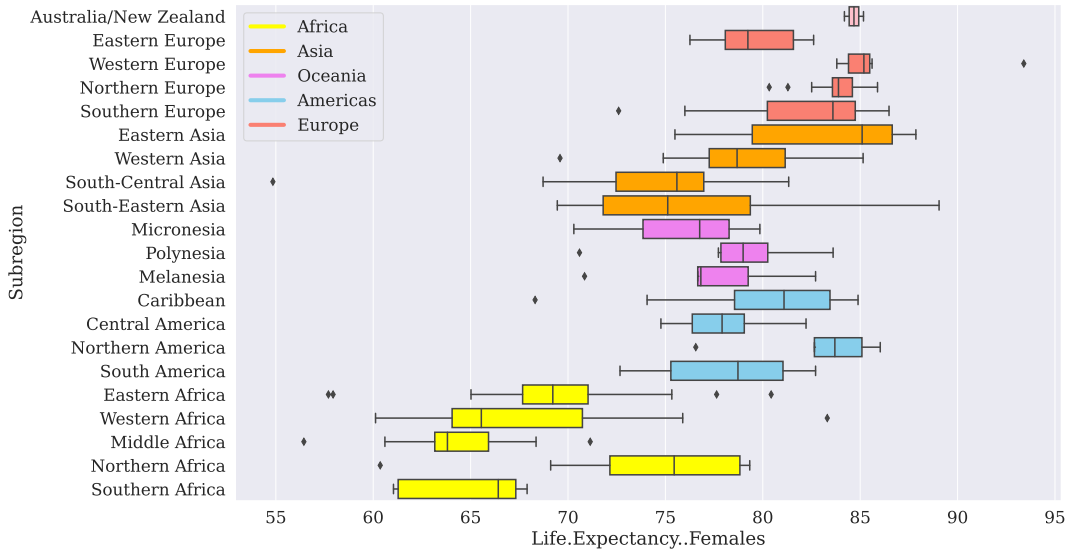

Figure 7: : Box plot showing variability in life expectancy at birth for females across the sub-regions

4.4 Trend analysis

The main focus of this section is to collect the information for 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 are not included in the analysis to avoid inconsistency.

The above-given Figure 10, 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.

For lower fertility rates, the values in 2021 are quite close to the values in 2001. For larger fertility rates in the year 2001, there seems to be a lower fertility rate for the year 2021.The extreme value of fertility rate in the year 2001 is 8.1 but decreases to 6.9 in 2021, belongs to the country Niger. Likewise, for Timor-Leste, the fertility rate in 2001 is 7.5 whereas, in 2021, it is around 4.3. On the whole, the total fertility rate displays a decreasing trend from the year 2001 to 2021.

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

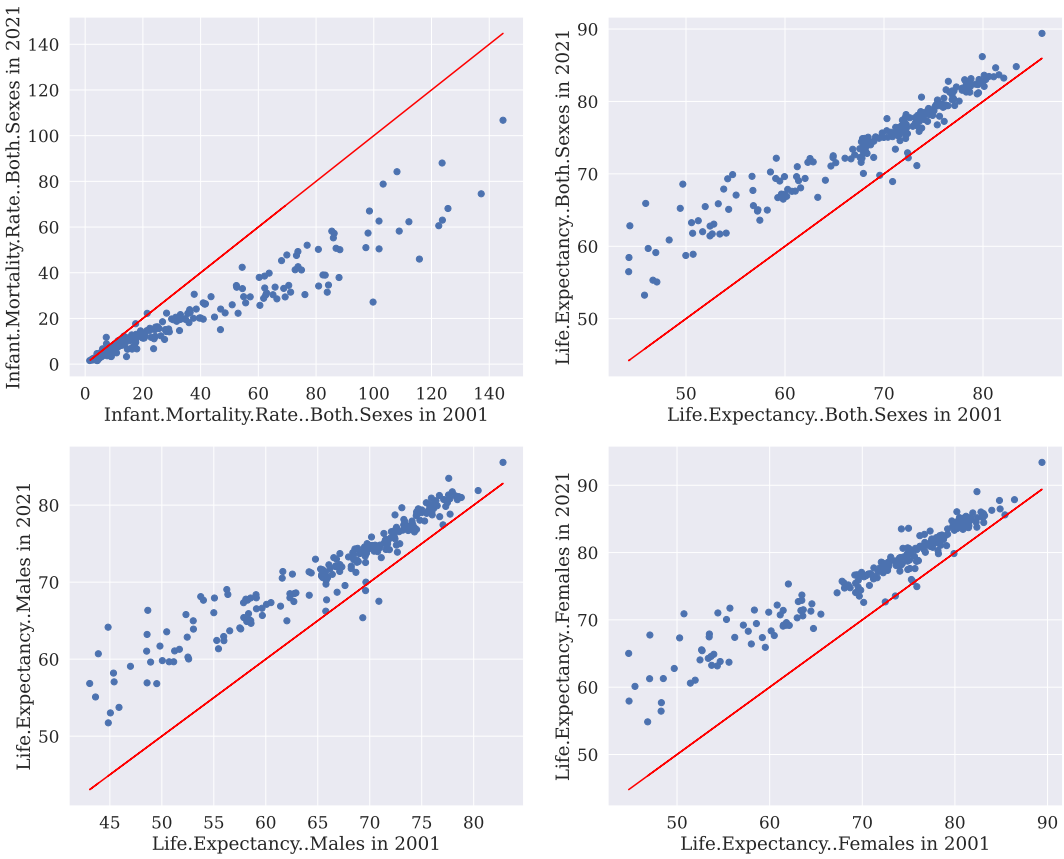

Figure 8: :Scatter plots showing trend of the variables from the year 2001 to 2021

variables in 2021 are higher than those in 2001. Generally, the values appear to be similar for higher life expectancies, approaching the ideal line.

In all cases Venezuela, Andorra have some of the extreme values. For Venezuela, the year 2001 shows 72.45 and 2021 shows 72.22 for life expectancy for both sexes. The value for life expectancy at birth of males in the year 2001 is 69.6 and in 2021, it is 68.9. Similarly, for females, the life expectancy at birth in 2001 is 75.46 and in 2021, 75.7 and and also for the country Andorra with 85.4 in 2001 and 85.6 in 2021. Thus, there is an slight increase in the overall life expectancy for most of the sub-regions and sexes with time. **?**

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 using exploratory and descriptive methods. 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 globe.

This project mainly focuses on the Fertility rate and life expectancy between males and females of all the 228 countries, which are grouped into 21 subregions and 5 regions. Examined data is between 2001 and 2021. Graphical and statistical measures were used to analyze the data. It is very important to note that the Fertility rate and life expectancy of both the sexes are closely related. Since the year 2001, the fertility 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. WHO contributors (2020)

Bibliography

evenios publishing Armin Stroß-Radschinski sole trader . Python. ://www.python.org/. [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

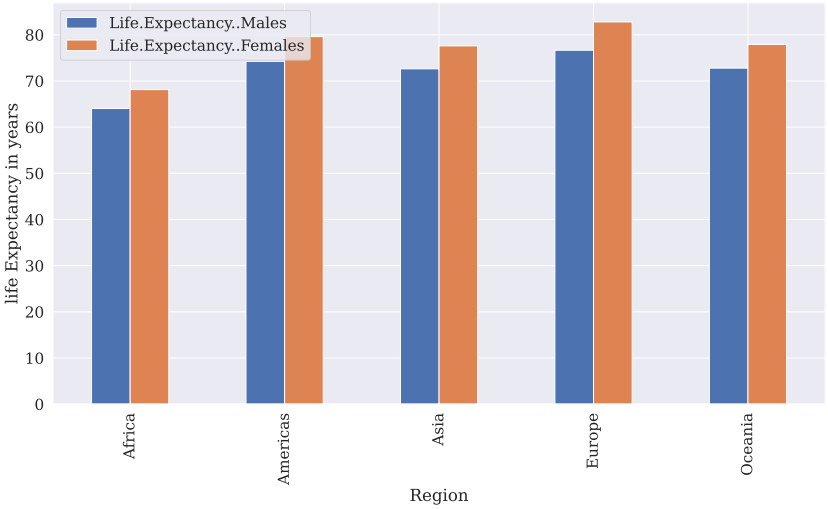

Figure 9: Bar chart of life expectancy at birth for males and females

Figure 10: Heatmap explaining bivariate correlation

B Additional tables

Table 1: Inter-quartile regions of each variable based on the sub-regions.

subregion	TFR	LEBS	LEM	LEF
Australia/New Zealand	0.066	0.280	0.080	0.490
Caribbean	0.286	4.180	2.870	4.900
Central America	0.315	2.755	3.243	3.313
Eastern Africa	1.266	3.840	4.100	3.360
Eastern Asia	0.497	8.575	8.345	8.745
Eastern Europe	0.090	4.402	5.705	3.277
Melanesia	0.506	2.450	2.460	2.590
Micronesia	0.490	4.460	5.075	4.420
Middle Africa	1.548	4.920	5.300	4.620
Northern Africa	1.374	10.305	10.642	9.862
Northern America	0.321	1.400	0.650	2.440
Northern Europe	0.265	1.305	1.627	1.117
Polynesia	0.679	2.750	3.465	2.405
South America	0.325	4.853	4.240	4.993
South-Central Asia	0.498	4.482	4.685	4.460
South-Eastern Asia	0.712	7.020	6.955	7.960
Southern Africa	0.526	6.110	6.630	6.040
Southern Europe	0.111	4.950	4.730	5.005
Western Africa	1.242	4.330	3.990	4.520
Western Asia	1.192	3.690	4.375	3.900
Western Europe	0.249	0.830	1.000	1.090