

TU DORTMUND

INTRODUCTORY CASE STUDIES

Project 1: Descriptive analysis of demographic data

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Philipp Adämmer

Dr. Andrea Bommert

M. Sc. Hendrik Dohme

Author: Aishwarya Dinni

Group number: 13

Group members: Aishwarya Dinni, Vinay Kumarmath, Bhavana
Thirunagari, Charishma Chinthakayala

November 12, 2021

Contents

1	Introduction	1
2	Problem statement	2
2.1	Data set description	2
2.2	Objectives	3
3	Statistical methods	3
3.1	Statistical measures	4
3.1.1	Measures of central tendency	4
3.1.2	Measures of variability	5
3.1.3	Pearson correlation coefficient	5
3.1.4	Range, quartile and inter-quartile range	6
3.2	Statistical plots	7
3.2.1	Grouped Bar Chart	7
3.2.2	Histograms	7
3.2.3	Box plots	7
3.2.4	Scatter Plot	8
3.2.5	Pairwise Plot	8
4	Statistical analysis	8
4.1	Univariate analysis	8
4.2	Bivariate analysis	10
4.3	Variability analysis	11
4.4	Trend analysis	15
5	Summary	16
	Bibliography	18
	Appendix	19
A	Additional figures	19
B	Additional tables	20



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. **International Data Base (IDB):** 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 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. **In Section 4 the observations, summary, and conclusions on analyzing this data.**

Section 2, the problem statement consists of an illustration of the data set, objectives, and the concerns examined.

2 Problem statement


2.1 Data set description

Description of the dataset: The motive of this report is to analyze an **extricate** from the International Data Base (IDB) of the U.S. Census Bureau(2021). The mentioned database contains accurate demographic measures and characteristics including population by sex and age, total population and fertility, mortality, migration for many countries respectively. 

As **above** 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:

- **CountryName:** (Nominal data type) - This variable describes the 228 countries that are found in this data set.
- **SubRegionName:** (Nominal data type) - The above mentioned 228 countries are geographically categorized into 21 subregions.
- **RegionName:** (Nominal data type) - The continents/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. 
- **Total.Fertility.Rate:** (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.
- **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 and on the other hand,
- **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, Sri Lanka, Sudan, Syria, United States have missing values for the year 2001. These records are not considered in this analysis.

2.2 Objectives

The motive 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 the 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 and this is done in order to understand the variables and their properties

3.1 Statistical measures

3.1.1 Measures of central tendency

Measure of central tendency is a typical value for a probability distribution. 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 divided by number of observations. The formula for arithmetic mean of a variable x is -

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

where,

- \bar{x} is the mean.
- n is the total number of observations that belong to \mathbb{R} , $x_1, \dots, x_n \in \mathbb{R}$.
- x_1, \dots, x_n are the observations of a data set.

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, \dots, x_n of size n arranged in ascending order, the formula for calculating the median is

- When n is an odd number:

$$\text{median}(x) = \frac{x_{(n+1)}}{2}.$$

- When n is an even number:

$$\text{median}(x) = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2})+1}}{2}.$$

?

3.1.2 Measures of variability

- **Variance:** Variance is a measure of dispersion, meaning it determines how far it is from their center of a distribution (average value). Considering a sample of size n with observations x_1, \dots, x_n , the variance is calculated as:

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

Where,

- S^2 is the sample variance.
 - x_i is the value of the one observation.
 - \bar{x} is the mean value of all observations.
 - n is the number of observations.
- **Standard Deviation:** is simply the square root of variance $\sqrt{S^2}$. 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.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}.$$

- σ is the population standard deviation
- N is the size of the population
- x_i = each value from the population
- μ = the population mean

3.1.3 Pearson correlation coefficient

is also called as bi-variate correlation, Pearson product-moment correlation coefficient (PPMCC). It can be defined as a measure of linear correlation between two continuous variables. On the other hand, correlation defines the kind 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 showing that there is negative correlation between the variables,

- 0 denoting that there is no relationship and
- 1 depicts that there is a strong positive correlation.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

- r = correlation coefficient
- x_i = values of the x-variable in a sample,
- \bar{x} = mean of the values of the x-variable,
- y_i = values of the y-variable in a sample and
- \bar{y} = mean of the values of the y-variable.

?

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.

$$\text{Range}(X) = \text{Max}(X) - \text{Min}(X).$$

Where,

- $\text{Max}(X)$ is the maximum value
- $\text{Min}(X)$ is the minimum value

?

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

- First Quartile, $Q1 = (\frac{n+1}{4})^{th}$ value,
- Second Quartile, $Q2 = (\frac{n+1}{2})^{th}$ value and
- Third Quartile, $Q3 = (\frac{3(n+1)}{4})^{th}$ value.



The first and third quartile are also called as upper and lower quartile respectively.

- **Inner-Quartile Range:** Other names for Inner-Quartile Range are midspread, middle 50%, or H-spread. It is defined as the difference between the 75th and 25th percentiles of the data. It is the measure of the variability of values about the median. It is the distance between the quartiles Q1 and Q3 i.e. $IQR = Q_3 - Q_1$



3.2 Statistical plots

Statistical plots present the results of some statistical analysis of the data. 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. Wikipedia contributors (2021)

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.

The above 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.Female.

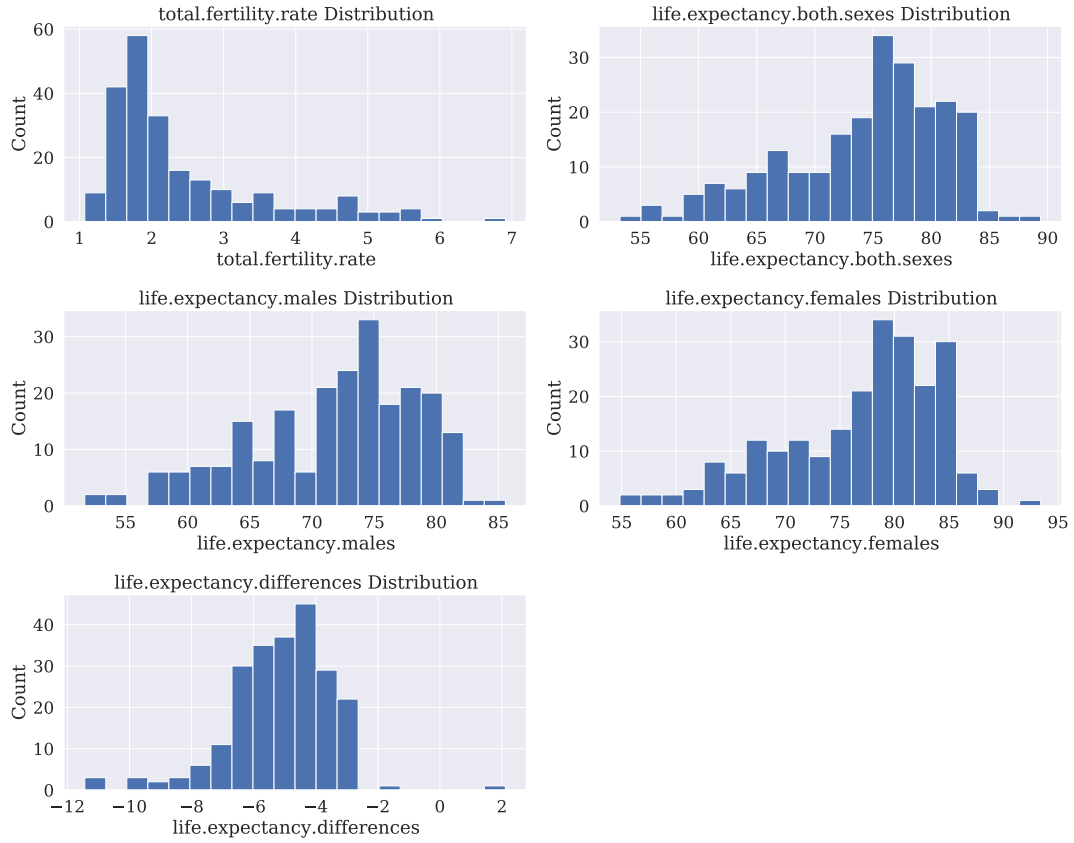


Figure 1: Frequency distribution histograms for different variables

and further, the female life expectancies is **excluded** 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 total.fertility.rate, is skewed to the left and for the rest of the charts, **it's** skewed to its right. In the first graph, the total fertility rate is centered between 1.7 and 1.9. The maximum is at 8.11 for Niger country. One can also observe that in the second graph, life expectancy for both the sexes is centered in between 75.0 and 77.0 and Monaco has 89.4, the maximal value among the given data.

The total life expectancy for males is centered between 74.0 and 75.1, 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.

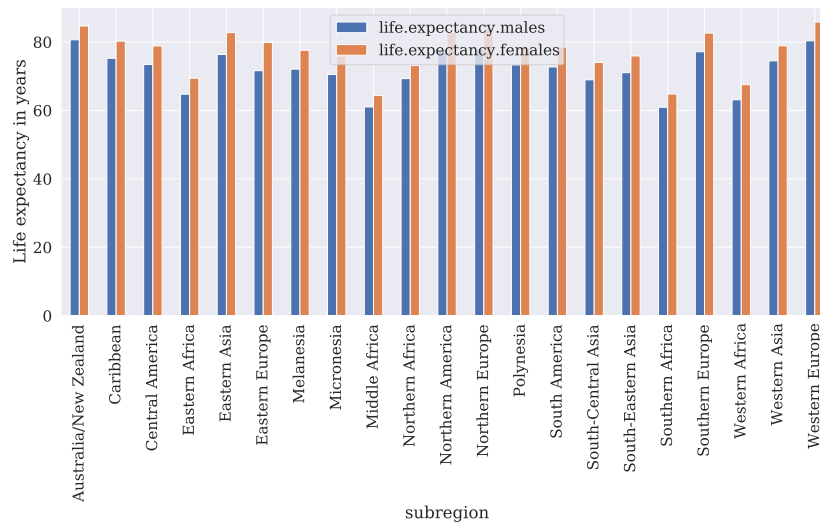


Figure 2: Bar chart of life expectancy at birth for males and females

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.

Table 1: Correlation matrix for the variables.

	TFR	LEBS	LEM	LEF
TFR	1.00	-0.800	-0.773	-0.816
LEBS	-0.800	1.000	0.993	0.993
LEM	-0.773	0.993	1.000	0.973
LEF	-0.816	0.993	0.973	1.000

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 2 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 2 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, -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.

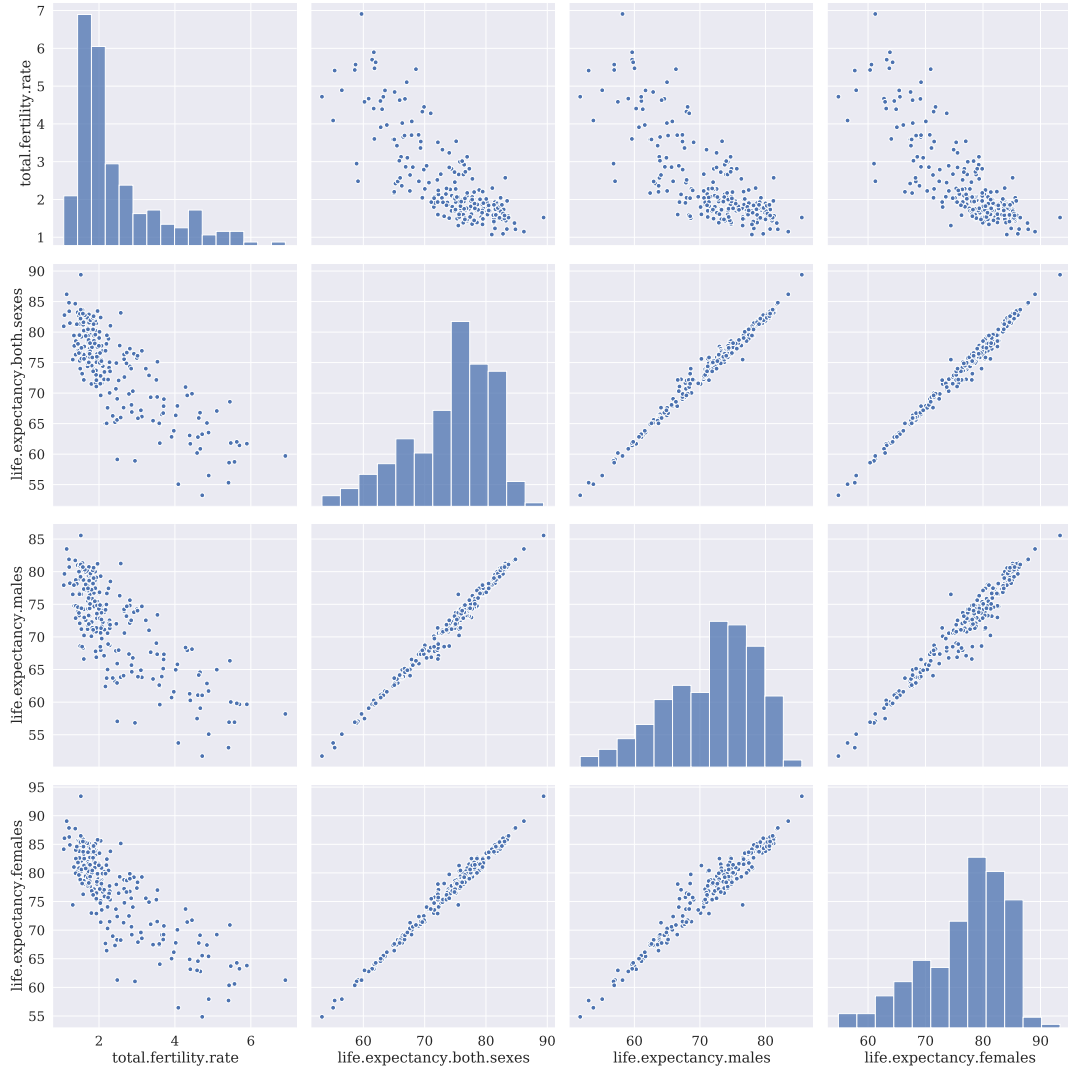


Figure 3: Correlation matrix for the variables

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.

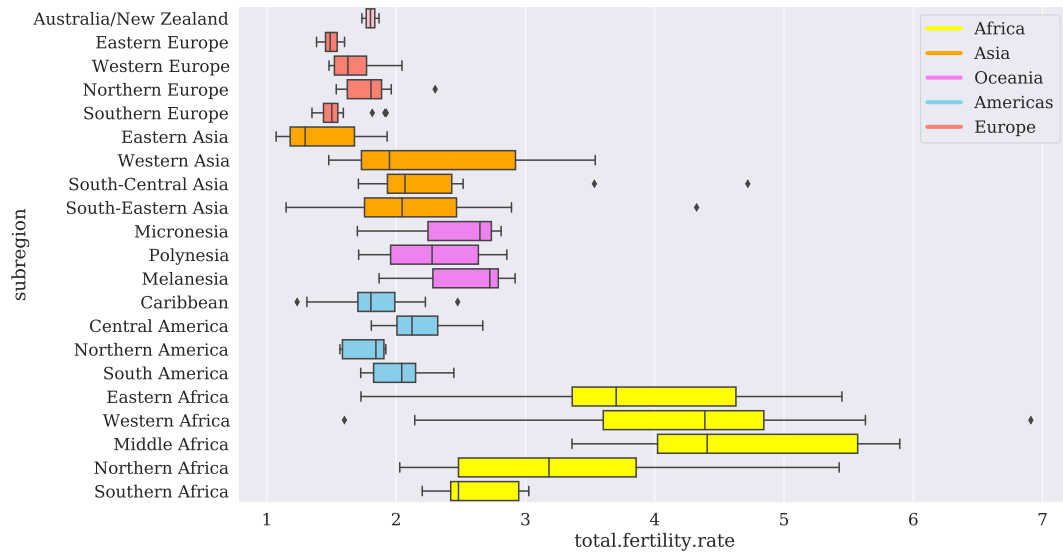


Figure 4: : Box plot showing variability in total fertility rate across the sub-regions

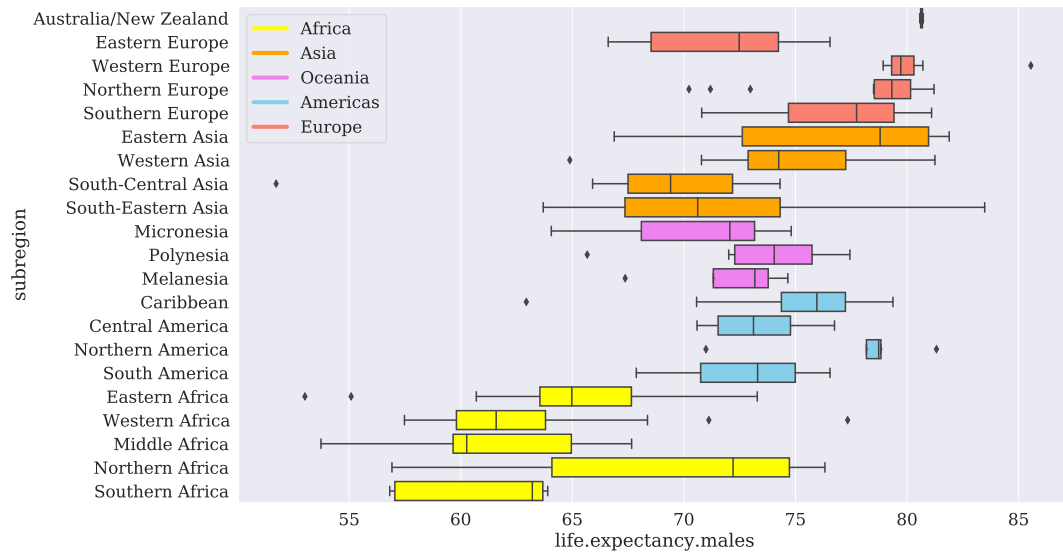


Figure 5: : Bar chart of life expectancy at birth for males

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

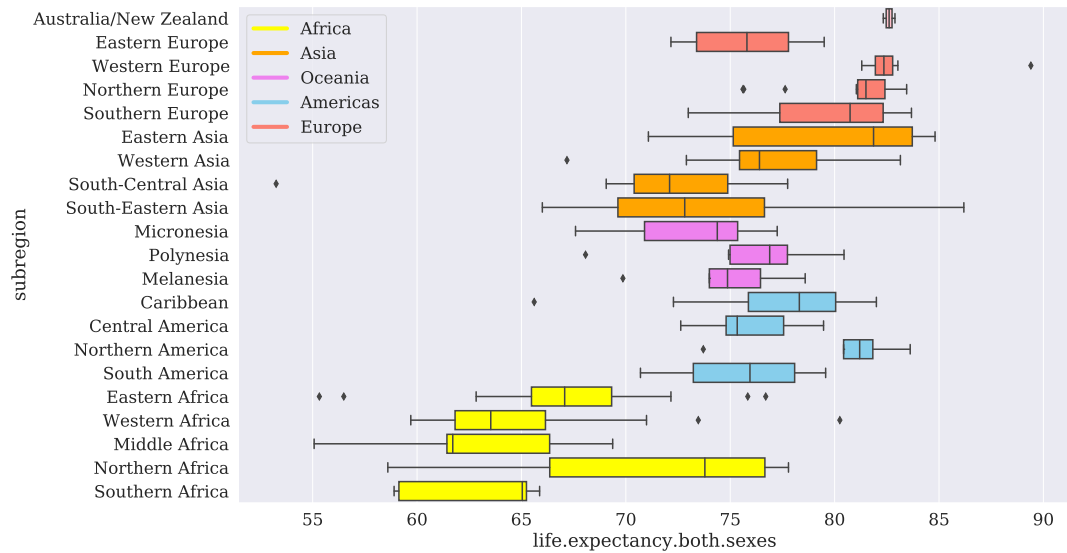


Figure 6: : Box plot showing variability in life expectancy at birth for both sexes the sub-regions

in both sexes' plots as well. On the other hand, Sub-regions like Australia/New Zealand, 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.

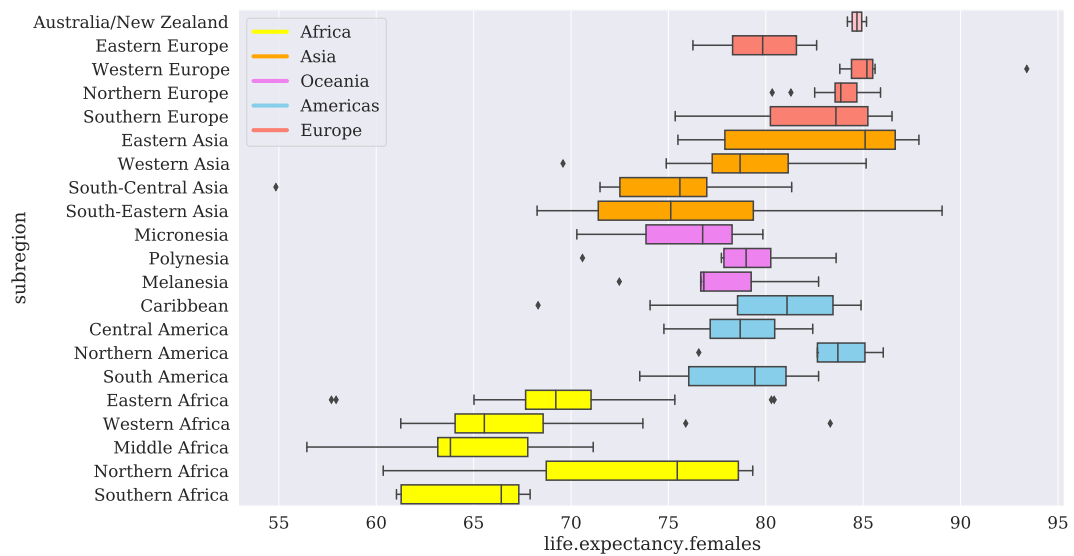


Figure 7: : Box plot showing variability in life expectancy at birth for females across the sub-regions

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.

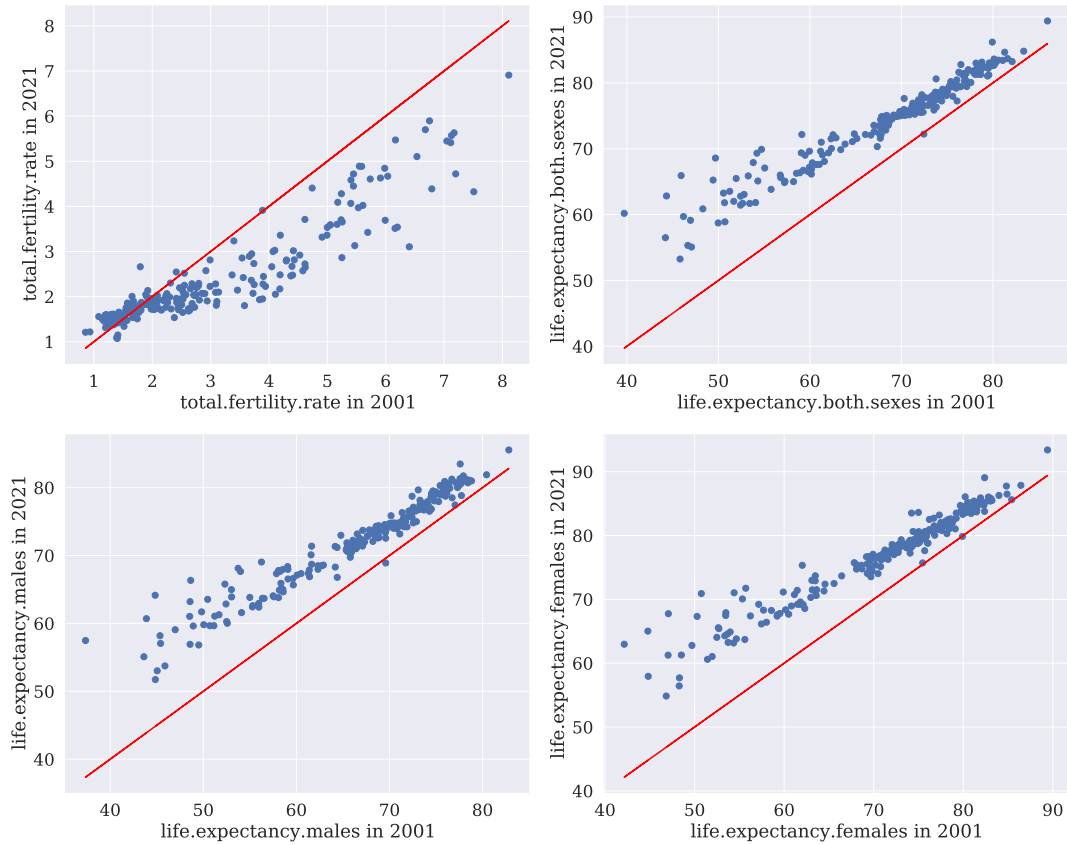


Figure 8: :Scatter plots showing trend of the variables from the year 2001 to 2021

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 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 also for the country Andorra with 85.4 in 2001 and 85.6 in 2021. Thus, there is a 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.



Bibliography

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



Appendix

A Additional figures

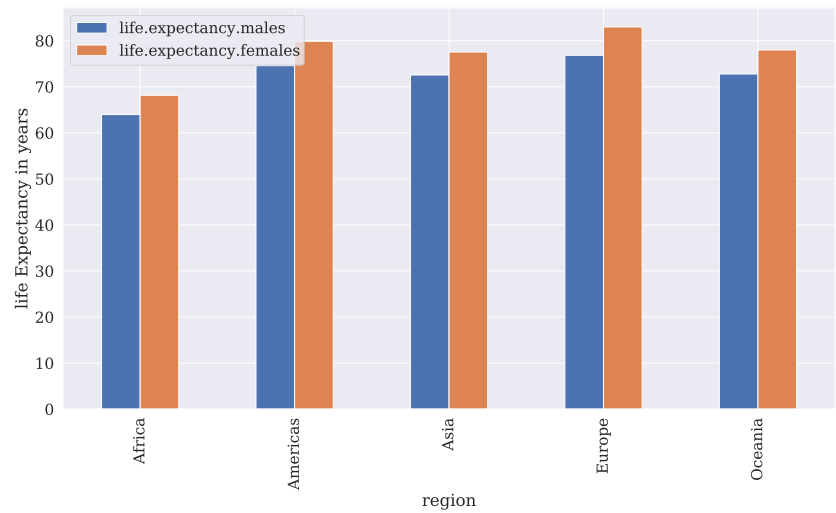


Figure 9: Bar chart of life expectancy at birth for males and females



Figure 10: Heatmap explaining bivariate correlation

B Additional tables

Table 2: 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