TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:

Dr. rer. nat. Maximilian Wechsung

M. Sc. Hendrik Dohme

Author: Ankita Sarkar

Group number: 22

Group members: Avisha Anilkumar Bhiryani, Janani Veeraraghavan , Kaushal Tajane, Shivam Shukla, Shubham Khochare

May 6, 2022

Contents

1 Introduction

The analysis of demographic data helps us to keep track the impact of major events such as disasters, pandemic etc on populations around the world. This kind of estimates and projections from the demographic data are great educational resources and are used in research, businesses and government decision making system. In this project, the statistical distributions of certain demographic variables such as Total Fertility Rate and Life Expectancy for the year 2022 have been considered. Secondly, the relationships between the variables are observed to find out the dependencies with each other. In the third step, the variability of the data across the subregions and regions are investigated. Finally, the change in values of the variables between 2002 and 2022 is compared.

In section 2, the description of dataset, methods of data collections and data quality are described. Section 3 briefly explains the statistical methods involving the mean, median, quartiles, histograms, boxplots and correlation coefficients as well as the software tools that are being used. Section 4 gives the detailed presentations of the results and their interpretations. The last section concludes the project and also discusses the possibilities of further analyses.

2 Problem statement

2.1 Description of the Dataset

The dataset used for this project is a small sample from the International Data Base (IDB) of the U.S. Census Bureau. The Census Bureau collects data from a variety of sources namely censuses, surveys, vital registries and also administrative records and makes them available through the IDB. National statistical offices of other countries are other primary sources for the data collection. The dataset contains different demographic measures for over 200 countries with populations 5000 or more. The US Census Bureau has produced the data since 1950. The IDB is updated regularly so that it stays relevant to research, program-planning and policy making plan in the US and around the globe (United States Census Bureau, 2020).

The sample used for this project contains 448 total observations from 227 countries from all over the world for the year 2002 and 2022. The 5 regions of Africa, Americas, Asia, Europe and Oceania are divided into different subregions (21 in total). For instance, 55 countries across 5 subregions in Africa, 52 countries in Asia, 50 in Americas, 49 in Europe and 21 in Oceania where each region is subdivided into 4 subregions. There are four categorical variables - country, region, subregion and year and four numeric variables - Total Fertility Rate, Life Expectancy of both Sexes, Life Expectancy of Males and Life Expectancy of Females. The Total Fertility Rate is defined as the average number of children that a woman can have if the woman being considered lives till the end of the child bearing age and bears children according to the age-specific fertility rates. Life Expectancy is defined to be the average number of years a group of people born in the same year are expected to live if mortality at each age remains constant in the future (United States Census Bureau, 2021). There are 6 missing observations found in the year 2002. The countries are Libya, South Sudan and Sudan from Africa, Puerto Rico and United States from Americas and Syria from Asia. The missing data are ignored in this project as we try to analyse the dataset based on region or subregion rather than individual country.

2.2 Project objectives

The purpose of this project is to carry out the descriptive statistical analysis of the four numeric demographic variables. First, the analysis of the frequency distributions of the variables of the year 2022 are displayed through histograms. The differences of life expectancy between males and females are shown with two overlapping histograms. Next, the Pearson and the Spearman correlation coefficients are calculated to check the linearity and monotonicity of the relationships between the variables. Third, the spread of data across the regions and the subregions are assessed with the help of boxplots. Finally, boxplots are used again to see how the numeric values between the year 2002 and 2022 have been changed.

3 Statistical methods

In this section, various statistical methods and plots are described which are used to analyse the dataset. Python programming language (Vanderplas, 2016) version 3.9.7 is used. The packages used are pandas (McKinney et al., 2011), matplotlib (Hunter, 2007) and seaborn (Waskom, 2021).

Arithmetic mean

The arithmetic mean is the measure of the central tendency of any given dataset. It is calculated as the sum of all the observations divided by the total number of observations (Heumann et al., 2016, p. 38).

The arithmatic mean of $\bar{x}$ of n observations $x_1, x_2, ..., x_n$ is given by

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

Median

To obtain the median in a dataset, the data needs to be ordered from the lowest to the highest value. Now if the data is divided into two equal parts, then at least 50% of the data is greater than or equal to the value in the middle or at least 50% of the data is less than the middle value. The so called value in the middle is the median. However the median depends on whether the total count of observations is odd or even. The median is denoted by $\tilde{x}_{0.5}$. Let us consider n number of observations $x_1, x_2, ..., x_n$ which are ordered as $x_1 \leq x_2 \leq, ..., \leq x_n$. If n is odd, then $\tilde{x}_{0.5}$ is the middle ordered value and if n is even, then $\tilde{x}_{0.5}$ is the arithmetic mean of the two middle ordered values:

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases}$$

If the data is distributed symmetrically around the centre, the mean and median become the same. In other cases, the mean and the median might differ. If the data has outliers, then it is useful to use median as the mean is sensitive to outliers (Heumann et al., 2016, p. 40-42).

Quartiles

The spread of the data can nicely be measured by quartiles. The first quartile, denoted as Q_1 is the 25th percentile which is the lowest 25% of the whole dataset. Q_3 known as the third quartile which is the lowest 75% of the whole dataset. The difference between the Q_3 and Q_1 is known as the interquartile range. The second quartile, Q_2, also known as the median.(Han et al., 2011, p. 49).

$$IQR = Q_3 - Q_1$$

Another good way to asses the data distribution is to look at the range of the data which is the difference between the largest value (max) and the smallest value (min) of a dataset.

Standard deviation

The standard deviation is useful in measuring the dispersion of the data. It tells us how spread out the data is around it's center. It is usually denoted by the Greek letter sigma (σ). For a set of n observations, $x_i, i = 1, .., n$,

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

where $\bar{x}$ is the arithmetic mean of the dataset. The square of the Standard Deviation is called the Variance. A low standard deviation means that the data is concentrated around it's centre and a high standard deviation indicates that the data is spread away from the centre (Han et al., 2011, p. 50-51).

Correlation

Suppose that two variables X and Y are continuous and are linearly related with each other i.e. $Y = a + bX$ where a and b are constants. Then r(X,Y) is called the Bravais–Pearson correlation coefficient which measures how X and Y are linearly related with each other.

$$r(X,Y) = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where

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

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and } S_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

The limit of r is between -1 to +1. X and Y is perfectly linear and decreasing if $r = -1$ and perfectly linear and increasing if $r = +1$. If the value of r is close to zero, it indicates that X and Y are independent and not linear.

Another measure of correlation between the ordinal or continuous variables is the Spearman's rank correlation coefficient which is useful in understanding the monotonicity of the relationships. (Han et al., 2011, p. 84-86)

Histogram

The distribution of continuous variables can well be understood with histograms. Suppose that X is a numeric variable. Then if the range of X is divided into equal consecutive parts, each part is called a bin. The range of each bin is the width. A bar is drawn in each bin. The height of the bar represents the total number of observations in a particular bin. The area of a bar (height $*$ width) is proportional to the relative frequency. The relative frequency is calculated as the number of counts in a particular bin divided by the total number of counts. The choice of selecting the number of bins is context specific. (Han et al., 2011, p. 54)

Skewness

The skewness of a distribution of data shows the relationships between mean, median and mode (maximum occurrence of any data point). A distribution of data has no skewness when the data is evenly spread out on either side of the center for unimodal case. If either of the tail is longer than the other, then the distribution is said to be either skewed left (negatively skewed) or skewed right (positively skewed). In a bell-shaped curve, the mean, median and mode tend to be at the centre whereas in a skewed curve, the mean tends to be located towards to tail of the distribution (Black, 2019, p. 77).

Boxplot

Boxplots are useful to compare several compatible datasets. The two ends of the box are the quartiles. So the box length is the interquartile range. The median is identified with a line within the box. Two dashed lines which are called whiskers are outstretched from either ends of the box are extended till the highest and the lowest value in the dataset.

The extreme values can be plotted individually if these are less than the $1.5 * IQR$ distance from both the quartiles (Han et al., 2011, p. 49-50).

Heatmap

A heatmap is usually used to visualize two dimensional data with colours. In this project, we have used heatmap to display the correlation between bivariate variables. The stronger the relationship is between the variables, the brighter the colour becomes on the heatmap. For better understanding, the graph is also annotated with correlation coefficient values (Waskom, 2021).

4 Statistical analysis

The statistical methods which have been explained in the section 3 are applied on the dataset to explore the data. First, the frequency distributions of the variables Total Fertility Rate, Life Expectancy of both Sexes, Life Expectancy of Males and Life Expectancy of Females in the year 2022 are presented to get a better understanding of the data. Next, the differences of life expectancy data between males and females are considered to compare the data directly. Then heatmaps are used to find out the monotonic and linear relationships between the variables. After that the variability of the variables within and between the subregions and regions are explored with boxplots. Finally, the change of values in the Total Fertility Rate and the Life Expectancy between the years 2002 and 2022 are displayed again with boxplots.

4.1 Frequency distributions

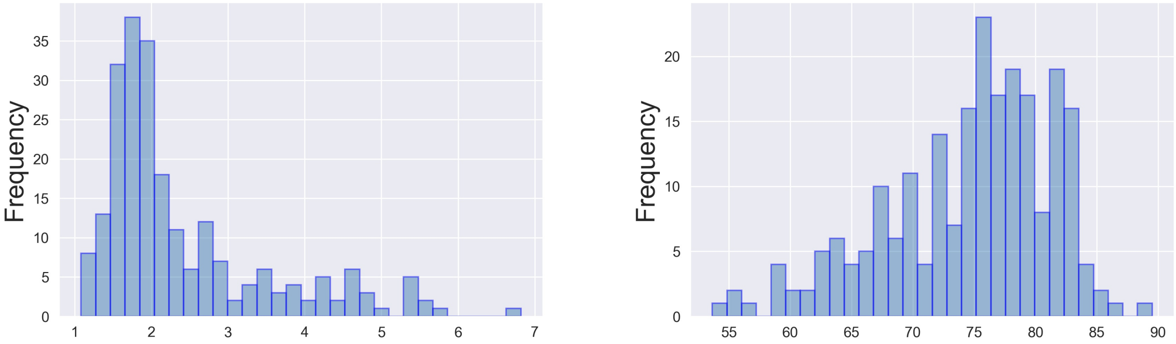
The frequency distributions of the four continuous variables for the year 2022 are considered. There are currently no missing data. Figure 1 and Figure 6 show histograms with x-axis as the respective variable and the y-axis as the frequency of the corresponding variable.

Figure 1(a) shows the distribution of the Total Fertility Rate which is positively skewed. There are maximum 6.82 children and minimum 1.08 child per woman. Niger, a country in Western Africa has the highest Total Fertility Rate and Taiwan in Eastern Asia has the lowest one. If we look at the interquartile range, we see that the 50% of the data are spread between 1.68 and 2.77 which indicates that the women in the half of the countries of the data have 2 to 3 children. On average each woman has 2.40 children. Table 1 gives the summaries for Total Fertility Rate.

Figure 1(b) is slightly negatively skewed. The median of the data is 75.82 and mean is 74.58. The range of the Life Expectancy of both Sexes is 35.87. Afghanistan in South-Central Asia has the minimum life expectancy at age 53.65 whereas Monaco in Western Europe has the maximum life expectancy at age 89.52. The interquartile range is 9.6. So 50 % of the countries have Life Expectancy of both Sexes between 79.65 and 70.05. Table 2 summaries the data for the Life Expectancy of both Sexes.

Table 1: Summary table for Total Fertility Rate

Min.	Q_1	Median	Mean	Q_3	Max.
1.08	1.68	1.95	2.40	2.77	6.82



(a) Total Fertility Rate

(b) Life Expectancy of both Sexes

Figure 1: Histogram for a) Total Fertility Rate and b) Life Expectancy of both Sexes in 2022.

Table 2: Summary table for Life Expectancy of both Sexes

Min.	Q_1	Median	Mean	Q_3	Max.
53.65	70.05	75.82	74.58	79.65	89.52

Figure 6 (Appendix, p. 16) shows that the histograms for the Life Expectancy of Males and Females separately of year 2022. The distributions are similar with that of the Life Expectancy of both Sexes. However we see that males live maximum of 85.70 while

females live maximum of 93.49 years. On average males live 72.10 years and females live 77.18 years. The minimum years of life expectancy for males is 52.10 and for females 55.28. We can better explore the differences of the same between the two sexes with the next distribution. The summaries of the data can be found in Table 3 (Appendix, p. 17).

Differences between the sexes

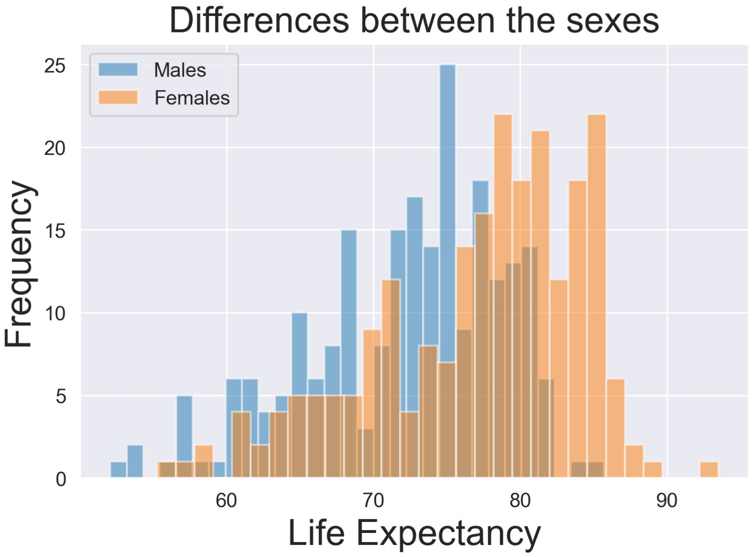
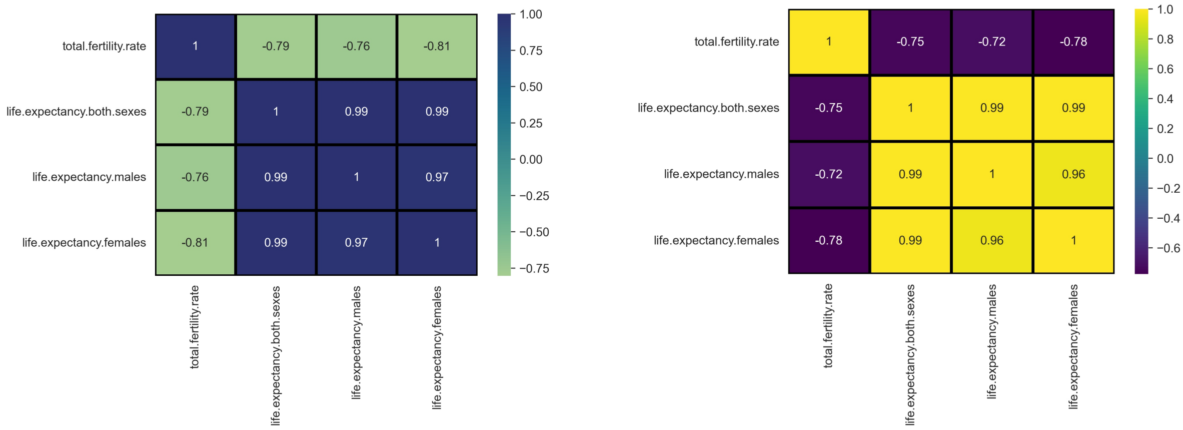


Figure 2: Differences between the sexes

Figure 2 shows two overlapping histograms for males and females consecutively to highlight the differences between the life expectancies of the two genders. Here we clearly see that females outlive males by 5 years on average. The column Differences between sexes in Table 4 (Appendix, p.18) shows that the mean is 5.10 years and median 4.85 years. The interquartile range is 2.21 meaning that the difference of the life expectancy between male and female is 2.21 years in 50% of the countries.

4.2 Analysis of Correlation between the variables

To compare the linear relationship between the variables, Pearson correlation coefficient is used whereas to check the monotonic nature of the relationships Spearman correlation coefficient is used. Figure 3(a) shows the heatmap for the Pearson correlation coefficients and Figure 3(b) shows the same for the Spearson correlation coefficients between the variables.



	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
total.fertility.rate	1	-0.79	-0.76	-0.81
life.expectancy.both.sexes	-0.79	1	0.99	0.99
life.expectancy.males	-0.76	0.99	1	0.97
life.expectancy.females	-0.81	0.99	0.97	1

(a) Pearson correlation coeeficient

	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
total.fertility.rate	1	-0.75	-0.72	-0.78
life.expectancy.both.sexes	-0.75	1	0.99	0.99
life.expectancy.males	-0.72	0.99	1	0.96
life.expectancy.females	-0.78	0.99	0.96	1

(b) Spearson correlation coeeficient

Figure 3: Correlation coefficient between variables in 2022.

We see that Total Fertility Rate and Life Expectancy of Females are strongly negatively correlated in Figure 3(a). The value of -0.81 between the two variables indicates a strong negative relationship. Spearman correlation coefficient of -0.78 between the two same variables in Figure 3(b) indicates monotonically decreasing nature. So we can conclude that the females with higher Total Fertility Rate tend to have lower Life Expectancy.

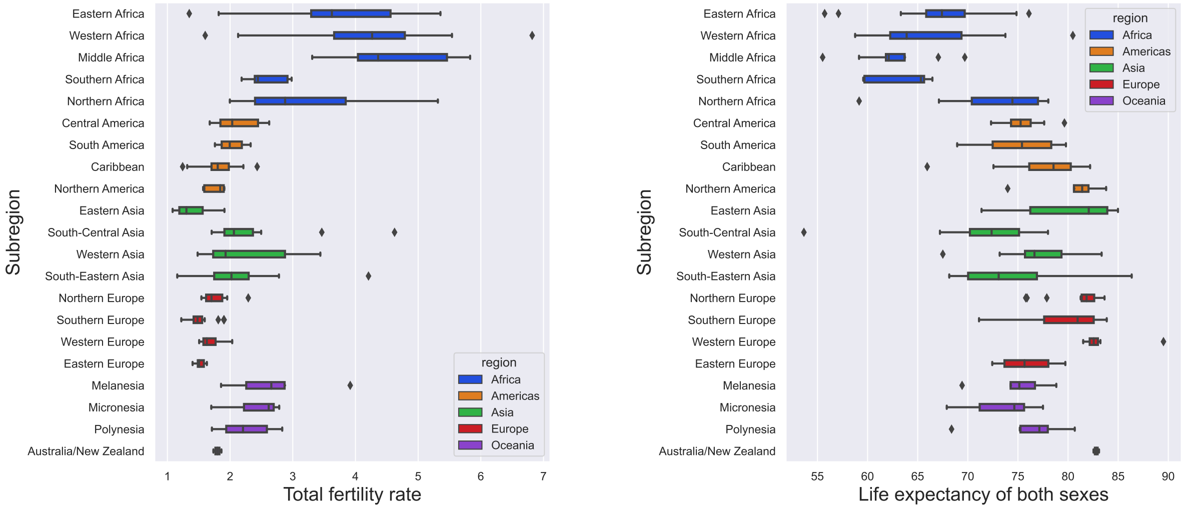
The correlations between the variables are best viewed through scatterplots which we have showed in Figure 7 (Appendix, p.16). The relationship between the Total Fertility Rate and Life Expectancy of Females is not exactly linear. However the relationships between the Life Expectancy variables is linear and monotonically increasing. The Life expectancies are observed as more positively correlated in case of females than for males.

4.3 Comparison of variability within and between subregions

The variability of the Total Fertility Rate within and between the subregions are analysed in this subsection. Multiple boxplots are used to explain the variabilities among the regions. The regions are alphabetically ordered. Different colours for each particular region are helpful in understanding the spread of data within the subregions in it.

Total Fertility Rate

In Figure 4(a), we see that the subregions in Africa show the highest variability followed by the subregions in Asia and then Oceania. Subregions in Americas and Europe show the least variability compared to other regions. The median values and the interquartile range in African subregions vary the most. For majority of European subregions and Australia/New Zealand in Oceania, the interquartile range is more compact indicating that the Total Fertility Rate in 50% of the countries in these regions do not vary much. It is interesting to see that Niger in Western Africa and Afghanistan in South-Central Asia have the highest and second highest extreme values namely 6.82 and 4.62 respectively. But these individual extreme values do not add much insight to the analysis.



(a) Total Fertility Rate

(b) Life Expectancy of both Sexes

Figure 4: Variability comparison within and between subregions for a) Total Fertility Rate and b) Life Expectancy of both Sexes in 2022.

Life Expectancy of both Sexes

Figure 4(b) represents the variability of the variable Life Expectancy of both Sexes in different subregions. The change in the shape of the boxplots are specifically noted as a result of the negative correlation between the Total Fertility Rate and Life Expectancy of both Sexes. The subregions which have higher Total Fertility Rate are now seen to have lower Life Expectancy. So we can see that African subregions tend to have the lowest Life Expectancy followed by the subregions in Asia and Oceania. The interquartile range varies quite a lot in African and Asian subregions indicating high variance in the variable. For America, Europe and Oceania, the variation in the interquartile range are relatively small. Monaco in Western Europe has the highest life expectancy of 89.52 years whereas Afghanistan in South-Central Asia has the lowest life expectancy of 53.65 years. Again, these are the extreme values and these do not influence the analysis much.

Figure 8(a) and 8(b) (Appendix, p.17) display the variabilities for the Life Expectancy of Males and Females separately. We again observe high variability in the African and Asian subregions. The interquartile ranges of Northern America, Western Europe and Australia/New Zealand display very less variance for both males and females.

4.4 Change of variables between 2002 and 2022

In this subsection, we see how the values for the variables have been changed over the years. To illustrate the changes, boxplots are used in the graphs for 2002 and 2022. The six regions are plotted on the x-axis and the respective variables are plotted on the y-axis. Two different colours in the graph represent two separate years.

In Table 5 (Appendix, p.18) we see that the mean Total Fertility Rate has decreased from 3.00 in 2002 to 2.41 in 2022. This can be observed in Figure 5 as well. The orange boxes are placed lower than the blue boxes for all the regions except for Europe. In Europe the median value seems to have increased in 2022 than that in 2002.

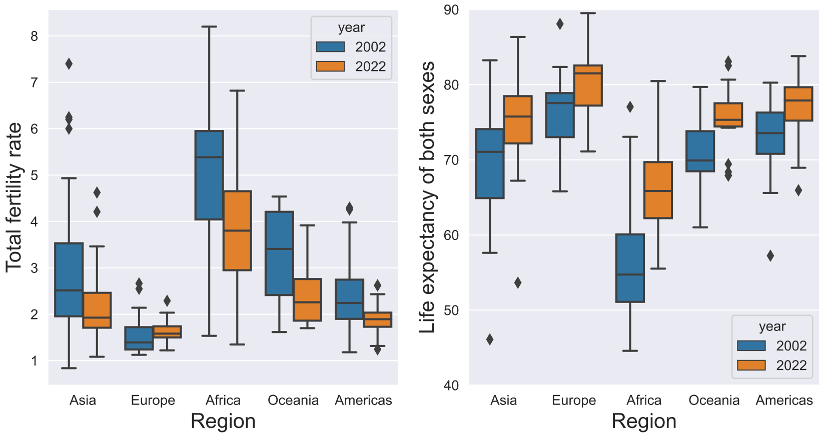


Figure 5: Region wise change in values of Total Fertility Rate & Life Expectancy of both Sexes between the year 2002 & 2022.

On average, life expectancy have increased in 2022 from that in 2002. This can be confirmed from Table 5 (Appendix, p.18) with mean life expectancy values of both sexes as 68.86 in 2002 and 74.58 in 2022. Figure 5 reflects these results. The change in values for the Life Expectancy of Males and Life Expectancy of Females can be found in Figure 9 (Appendix, p.17).

5 Summary

For the purpose of this project, a small subset of the International Data Base (IDB) maintained by U.S. Census Bureau which contained demographic data for over 200 countries with 500 populations or more (from 1950 till now) were used. Four variables namely Total Fertility Rate, Life Expectancy of both Sexes, Life Expectancy of Males and Life Expectancy of Females were considered. The dataset also included the information of country, region, subregion and year. Data from 5 regions only for the year 2002 and 2022 had been taken into account. To better understand the data, the frequency distribution of the numeric variables were shown. It had been observed that the Total Fertility Rate of 50% of the countries lie between 1.68 to 2.77. Niger in Western Africa was observed to have the highest as 6.82 children per female and Taiwan in Eastern Asia the lowest as 1.08 child per female. The Life Expectancy of both Sexes were found to be in between 79.65 to 70.05 for 50% of the countries. One case of highest Life Expectancy of both Sexes was 89.52 found in Monaco in Europe and one lowest case of 53.65 was in South-Central Asia country of Afghanistan. It was found that women on average live 5 years longer than men. The variables Total Fertility Rate and Life Expectancy of Females are negatively correlated with each other. The relationship is monotonic though not exactly linear. There were maximum variance within the African subregions followed by Asia and Oceania for the studied variables. European subregions, Americas and Australia/New Zealand in Oceania region displayed less variabilities within themselves. The negative correlation between the Total Fertility Rate and the Life Expectancy is also reflected in the variability analysis. The change in the values of the variables between the year of 2002 and 2022 were noted as well. It was found that the Total Fertility Rate had decreased (expect for Europe) and the Life Expectancy had increased in 2022 across all the countries than those in 2002..

For further investigations, it can be useful to include variables such as child mortality rate, countries' GDP, economic standards, health care conditions etc in the project to make strong conclusions about the increasing or decreasing trend of the variables and their inter relationships.

Bibliography

Black, K. (2019). *Business statistics : for contemporary decision making, 6th Edition.* Wiley.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques, 3rd Edition.* Elsevier.

Heumann, C., Schomaker, M., and Shalabh (2016). *Introduction to statistics and data analysis.* Springer.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. URL: `https://matplotlib.org/`.

McKinney, W. et al. (2011). Pandas: Python library for statistics. URL: `https://pandas.pydata.org/`.

United States Census Bureau, I. D. B. (2020). International data base: Population estimates and projections methodology. URL: `https://www2.census.gov/programs-surveys/international-programs/technical-documentation/methodology/idb-methodology.pdf`. (Visited on 1st May 2022).

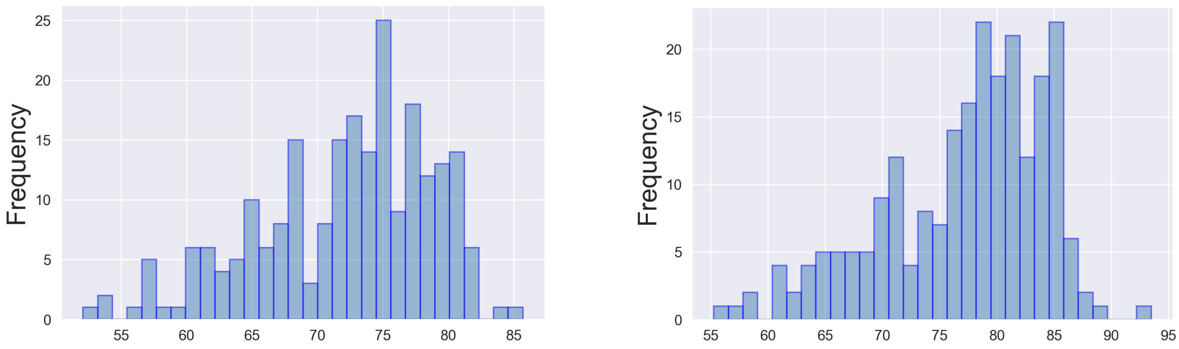
United States Census Bureau, I. D. B. (2021). Glossary census bureau. URL: `https://www.census.gov/glossary/`. (Visited on 1st May 2022).

Vanderplas, J. T. (2016). *Python Data Science Handbook: Tools and Techniques for Developers.* O'Reilly.

Waskom, M. L. (2021). Seaborn: Data visualization library in python. URL: `https://seaborn.pydata.org/`.

Appendix

A Additional figures



(a) Life Expectancy of Males

(b) Life Expectancy of Females

Figure 6: Histogram for a) Life Expectancy of Males and b) Life Expectancy of Females in 2022.

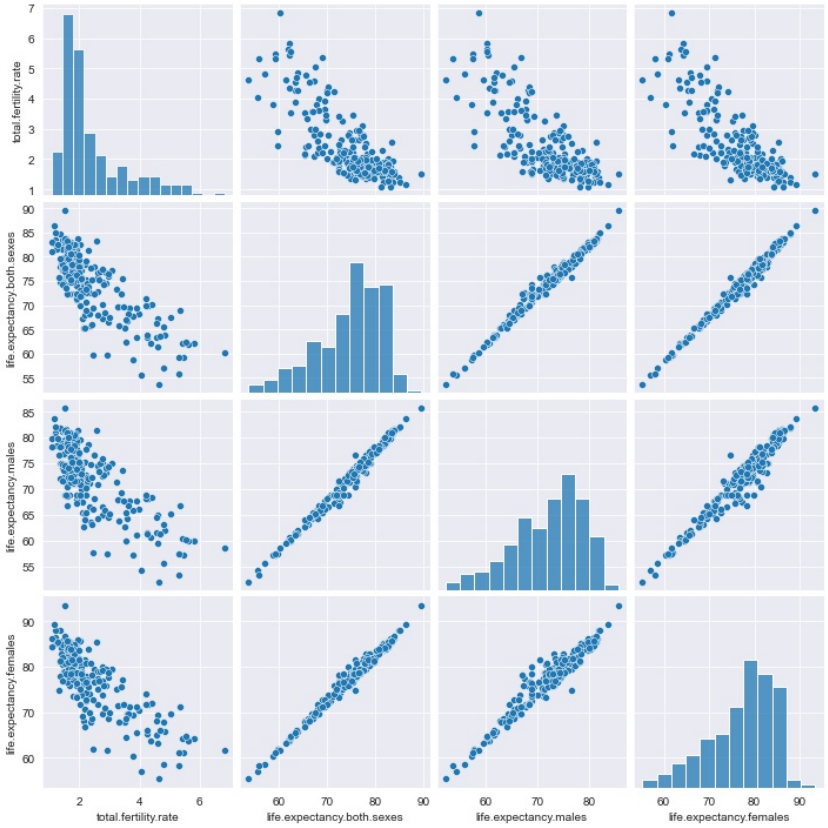
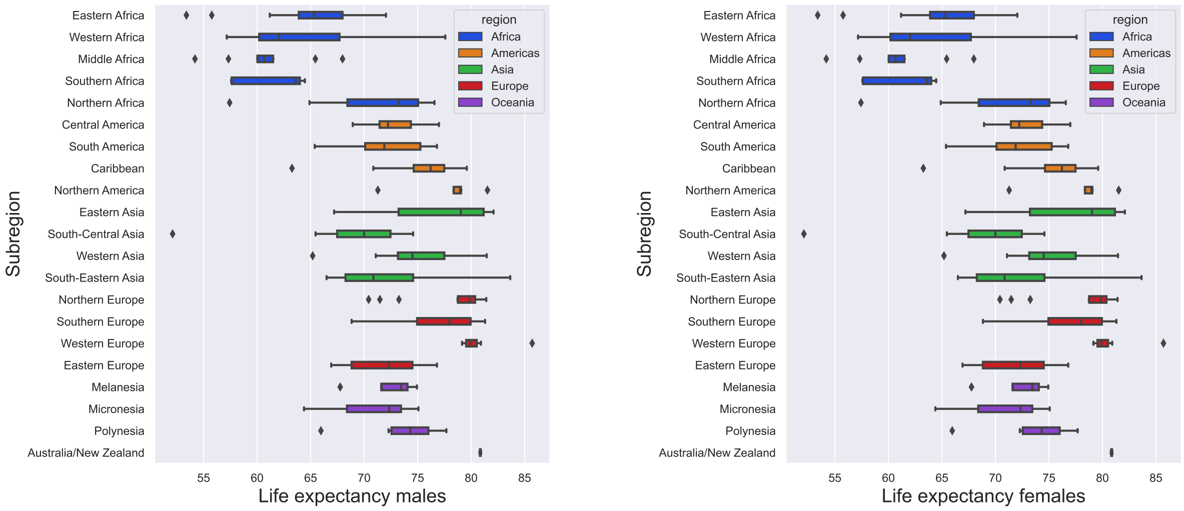


Figure 7: Pairplots between the variables in 2022.



(a) Life Expectancy of Males

(b) Life Expectancy of Females

Figure 8: Histogram for a) Life Expectancy of Males and b) Life Expectancy of Females in 2022.

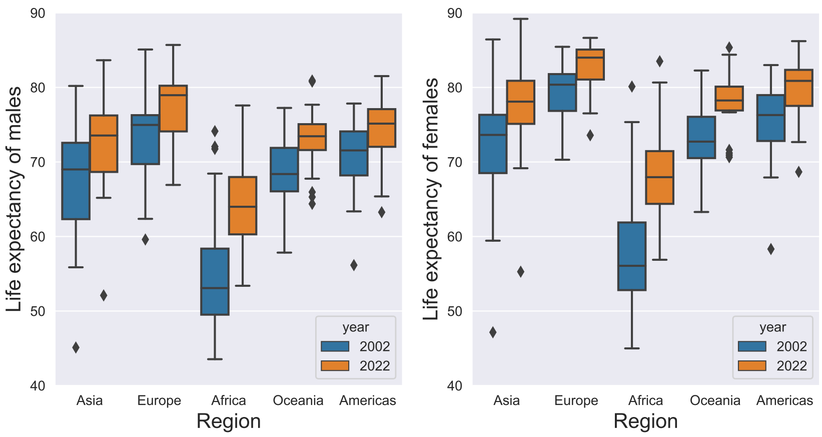


Figure 9: Region wise change in values of Life Expectancy of Males & Life Expectancy of Females between the year 2002 & 2022.

B Additional tables

Table 3: Summary table for Life Expectancy of Males & Females

	Min.	Q_1	Mean	Median	Q_3	Max.
Males	52.10	67.93	72.10	73.26	77.19	85.70
Females	55.28	72.63	77.18	78.69	82.55	93.49

Table 4: Description of data in 2022

	Total fertility rate	Life expectancy of of both sexes	Life expectancy of males	Life expectancy of females	Differences between sexes
count	227	227	227	227	227
mean	2.41	74.58	72.10	77.18	5.10
std	1.11	6.84	6.67	7.13	1.66
min	1.08	53.65	52.10	55.28	1.51
25%	1.68	70.05	67.93	72.63	3.82
50%	1.95	75.82	73.26	78.69	4.85
75%	2.78	79.65	77.19	82.55	6.03
max	6.82	89.52	85.70	93.49	11.38

Table 5: Pivot table for the mean value of the variables in 2002 & 2022

Year	Total fertility rate	Life expectancy of of both sexes	Life expectancy of males	Life expectancy of females
2002	3.00	68.86	66.55	71.29
2022	2.41	74.58	72.10	77.18