TU Dortmund

Introductory Case Studies

Project 1 : Descriptive Analysis of Demographic Data

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Birte Hellwig
Dr. Paul Wiemann
M. Sc. Hendrik Dohme

Author: Kunal Kochar

Group number: 22

Group members: Kunal Kochar, Harshini Eggoni, Supritha Palguna, Natasha Sahare

November 12, 2021

Contents

1 Introduction 3

2 Problem statement 3

2.1 Overview of the Data Set and Data Quality 3

2.2 Report Objective and Task Details 4

3 Statistical methods 4

3.1 Uni-variate 5

3.1.1 Measure of Central Tendency 5

3.1.2 Measure of Spread 6

3.1.3 Measure of Position 6

3.2 Bi-variate 7

3.2.1 Correlation 7

3.3 Histogram 8

3.4 Scatterplot 8

3.5 Boxplot 9

4 Statistical analysis 10

4.1 Uni-variate Analysis 10

4.2 Bi-variate Analysis 11

4.3 Variability of the values in the individual and different Subregions 13

4.4 Comparison of variables from 2001 to 2021 15

5 Summary 16

Bibliography 18

Appendix 19

A Additional tables 19

B Additional figures 19

1 Introduction

Exact and up-to-date evaluation of demographic data and their measurements are vital for understanding social, financial and open well being issues that influence population around the world. An increase or decrease in population has both good and bad impact on development of specific country. It hence makes sense to analyze from a wide variety of parameters which is best to explain the affect. Two such crucial parameters are life expectancy and fertility rate. In this report we will have a detailed overview of these two parameters that are key indicators for global health systems based on different regions, sub-regions and countries. The general thought or hypotheses is countries with high life expectancy rate prefer having low number of children compared to other countries. Therefore, individuals live longer in country with low fertility rate than individuals living in a country with high fertility rate.

In the second part of this report, an outline of data set is given which includes all the necessary variables or features to analyze the data. A brief description of the report objective is also mentioned here. On the third segment, statistical methods are defined which have been used in exploratory data analysis including the software packages and libraries with their version are mentioned. All the statistical methods that are being used in this project are explained in this section along-with their formula. In the fourth section, analysis is being performed with the help of the statistical methods described in third segment. All the crucial variables are being examined and the results with the visuals are presented here. Finally, the fifth section provides summary of all the interpretations and results. It also provides an insight for the further analysis.

2 Problem statement

2.1 Overview of the Data Set and Data Quality

In this study, the demographic data was extracted from IDB (International Data Base) of the U.S. Census Bureau. It contains data about life expectancy and fertility rates of 228 countries for the year 2001 and 2021. The entire data set is used for descriptive analysis and it contains 9 features and 456 records. Each feature has different values to each row which are the observations for the data set being used. Out of the nine features, four are categorical variables i.e. Country, GENC, Subregion, Region. The *country* corresponds to the name of the Country for which the observations are recorded. There are 228 countries, 21 sub-regions which are clubbed to provide 5 regions or continents. The categorical variables which is of least use for our study is *GENC* which are developed to allow government to have a standard/abbreviated way to refer the countries. The variable *Year* is an ordinal variable which have a clear order and states the particular year (2001 or 2021) of which the data is recorded. The data set consists of four numerical variables and these are *total fertility rate* which measures the average number of children a woman would have within her life, *life expectancy both sexes* which refers to average number of years an individual is expected to live. The variable life expectancy is further separated gender-wise: *life expectancy at birth for males and females.*

There are only 6 missing records for 4 features which can be ignored for this report so that it does not have a unintended effect on the measures that can be drawn from the data. There are 2 missing values in the variable GENC which is misleading because the 'NA' is encoding/abbreviation of the country present in the data set and not NA values. So, this is a problem of data preprocessing.

2.2 Report Objective and Task Details

In this report, detailed exploratory data analysis is performed on our data set. In order to perform analysis, we have few packages and libraries and their functions which provides better visuals and describes our data best. In the first task, uni-variate analysis of numerical variables are performed with the help of histogram plot. After the uni-variate analysis, relation amongst the variable is explored with the help of bi-variate correlation and heat-map. In the third task, the variables are compared within sub-regions and between different subregions to check for homogeneity and heterogeneity respectively with the help of box plots. Lastly, variables are compared between year 2001 and 2021 to check for change in the variables over the last 20 years.

3 Statistical methods

In this section, we will look into several statistical methods which have been used for the data analysis. All the measures and analysis were created in Jupyter notebook using *Python Programming Language.*

Table 1 gives an overview of packages and libraries that are used in this report and their respective versions and short description.

Table 1: Packages and Libraries used in this project

Library	Version	Short Description
Pandas	1.1.3	Used for Data Analysis and Manipulation
Matplotlib	3.3.2	Used for Plotting Graphs
Seaborn	0.11.0	Used for Data Visualization

For the descriptive analysis, the first step would be to get the overview of the variables in the data set. The below mentioned statistical methods are used to evaluate the given data set, for the current project.

Lets consider a numerical sample of the following *n observations*: $x_1, x_2, x_3, ..., x_n$.

3.1 Uni-variate

The following Statistical methods are applied to the numerical variables.

3.1.1 Measure of Central Tendency

Central tendency describes the middle of the data. Below are typical measure of central tendency that are used in this report.

Arithmetic Mean The arithmetic mean is the most widely used measure of averages. It is calculated by taking sum of all the numbers in the series and dividing it by the count of the numbers in the series. The *sample mean* can be computed using the below formula:

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

For a *known* population, *population* mean (μ) can also be computed using the above formula.

Note: This method is sensitive to extreme values and hence it is best to be used when the data distribution is continuous and symmetrical. (Christopher Hay-Jahans, 2018).

Median The median is the middle element of the data set when arranged into an array. It can be thought as geometric middle while mean can be thought as arithmetic middle. For its computation, the above sample $x_1, x_2, .., x_n$ is assumed to be sorted in ascending order. $Median(\hat{x})$, is then calculated as:

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

Note: This method is insensitive to extreme values and is therefore preferred more than arithmetic mean if there are extreme values present in the sample data. (Christopher Hay-Jahans, 2018).

3.1.2 Measure of Spread

Spread refers to the dispersion of the data and it is important because it determines the reliability of central tendency measurement. Below are typical measures of deviation that are used in this report.

Variance and Standard Deviation The variance is the average of the squared deviations of the data from their mean. The standard deviation is given by the square root of the variance. *Sample* variance (s^2) and Standard Deviation (s) can be calculated as follows:

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

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

For a *known* population, *population* Variance (σ^2) and *population* Standard Deviation (σ) can also be computed using the above formula. (Christopher Hay-Jahans, 2018).

3.1.3 Measure of Position

Range and Interquartile Range The range is the difference between the smallest and the largest data values. Before describing Interquartile range (IQR), we will understand about quartiles. The quartiles separates the data into four equal sized groups. There are five factors that make up the quartiles which are mentioned below in Table 2:

Table 2: Overview of Quartiles and their respective Percentiles

Quartile	Percentile	Short Description
-	0th	minimum or the smallest number within the data-set
Q1	25th	number that separates the lowest 25 percent of the group
Q2 or Median	50th	number in the middle of the sorted data set
Q3	75th	number that separates the lowest 75 percent of the group
-	100th	maximum or the largest number within the data-set

Quartiles are used to summarize a group of number. Rather than looking into a big list of numbers, it is easier to look into a small list of numbers which gives the intuition what exactly is happening in the big list. Quartiles are widely used in Box-plots which will be defined and used in the later section. The interquartile range refers to the middle 50 percent of an ordered data set and it is equal to the third quartile minus the first quartile. (Walter Antoniotti, 2001).

3.2 Bi-variate

3.2.1 Correlation

The Coefficient of Correlation (r) measures the strength of relationship between two variables. After understanding about variance and standard deviation it is easy to calculate the correlation coefficient. The correlation coefficient is determined by dividing the covariance by the product of the two variables standard deviation. It takes values between +1 and -1 inclusive and the closer the coefficient to either of extreme, the stronger the relationship is. The correlation coefficient value and the corresponding interpretation is mentioned in Table 3.

The correlation coefficient is given by:

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

(Walter Antoniotti, 2001)

Table 3: Interpretation of Correlation Coefficient

Correlation Coefficient value	Interpretation
+1	Indicates a strong positive relationship between variables
-1	Indicates a strong negative relationship between variables
0	Indicates no relationship at all

3.3 Histogram

A histogram is the most commonly used graph to show frequency distribution. Histogram is used to summarize discrete or continuous measurements. The X-axis contains the variable whose frequency is to be counted and the Y-axis contains the actual frequency of the variables. It takes the continuous or discrete measurement and places them into ranges of values known as bins. The different heights of the bins shows frequency of occurrence of the data. The histogram plot reveals the properties of the sample data using the summary statistic in a graphical way. We use summary statistic to describe an entire data set and it uses a single value to represent these measure like mean and median. Histogram uses these summary statistic and plots the measure and data set which helps in understanding and interpreting the data very easily and clearly. In this report, we are using density histogram and not frequency histogram. The density histogram is a smooth version of frequency histogram which shows the percentage of each unit to make the total area of all the bins equal to 1.

Histogram is used to identify,

1. Central Tendency of the data (mean, median).
2. Spread/dispersion of the data (variance, standard deviation).
3. Shape of the distribution (symmetric, asymmetric, uniform).
4. Outliers (unusual values in the data set).

3.4 Scatterplot

Scatterplot helps in exploring the relationship between two numerical variables. Scatterplot uses a collection of points or dots to represent the numerical variables. The position of each point represents the value for the corresponding variables and also report the patterns among the variables. Suppose that the horizontal and vertical axes are referred to as X and Y axes, the following can be interpreted from the scatterplot:

Table 4: Interpretation based on the scatterplot of a data set

Trend of the data (from left to right)	Interpretation
Upward trend	Indicates positive relationship between X-axis and Y-axis
Downward trend	Indicates negative relationship between X-axis and Y-axis
No pattern found	No relationship exists between X-axis and Y-axis

3.5 Boxplot

Boxplot helps to describe the distribution of data based on five number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum). The horizontal line that extends out from the box are called whiskers and the box itself in between is called the interquartile range. Below is the short description on each term and an example of boxplot,

1. Minimum: It is the smallest value in the data set (excluding outliers).

2. First Quartile (Q1) : Number which separates the lower 25 percentage of the data with the above 75 percentage of the data.

3. Median : The middle value of the data set or it can be also be referred as second quantile (Q2) which is a point where 50 percentage of data are above and below it.

4. Third Quartile (Q3) : Number which separates the lowest 75 percentage of the data with the above 25 percentage of the data

5. Maximum : It is the largest value in the data set (excluding outliers).

Boxplot are used to identify the average score of the data set, skewness of the data set, variability in the data set, outliers within the data set.

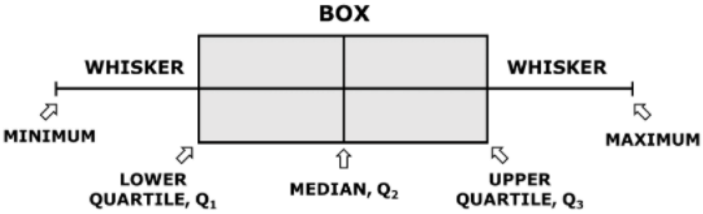
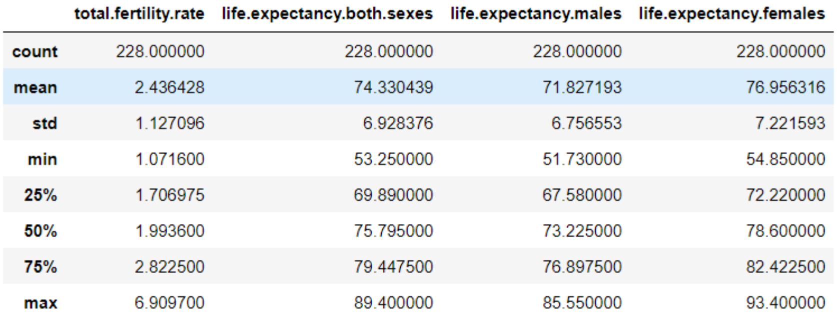


Figure 1: Box Plot

4 Statistical analysis

The Statistical methods described in Section 3 are implemented on the whole data set to get the better understanding of data. It can be summarized as below:

Description of the Data Set Table 5 can be referred for this section as it provides descriptive statistical analysis of all numerical variables and summarizes the central tendency, dispersion and shape of each variables. It is to be noted that we will be considering only year 2021 for the first three tasks as per the report requirement.



	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
count	228.000000	228.000000	228.000000	228.000000
mean	2.436428	74.330439	71.827193	76.956316
std	1.127096	6.928376	6.756553	7.221593
min	1.071600	53.250000	51.730000	54.850000
25%	1.706975	69.890000	67.580000	72.220000
50%	1.993600	75.795000	73.225000	78.600000
75%	2.822500	79.447500	76.897500	82.422500
max	6.909700	89.400000	85.550000	93.400000

Table 5: Description of the Data Set

4.1 Uni-variate Analysis

Univariate Analysis is the simplest form of analyzing the data where we examine each numerical variable individually. We will use frequency distribution plot to look at the distribution of each variables.

The following analysis can be made from Figure 2,

We can see from the histogram plot that the distribution of total fertility rate shows right skewness whereas life expectancy is left skewed. The frequency of the fertility rate is maximum around 2 which means that maximum number of women had 2 children for the year 2021. The frequency distribution of life expectancy (for both sexes) is maximum around 75. This means that the maximum number of people, in general, born in the year 2021 are expected to live around 75 years. The frequency distribution of life expectancy of females is maximum around 77 whereas for males it is maximum around 72. Thus it can be stated that the frequency distribution of life expectancy of females is slightly

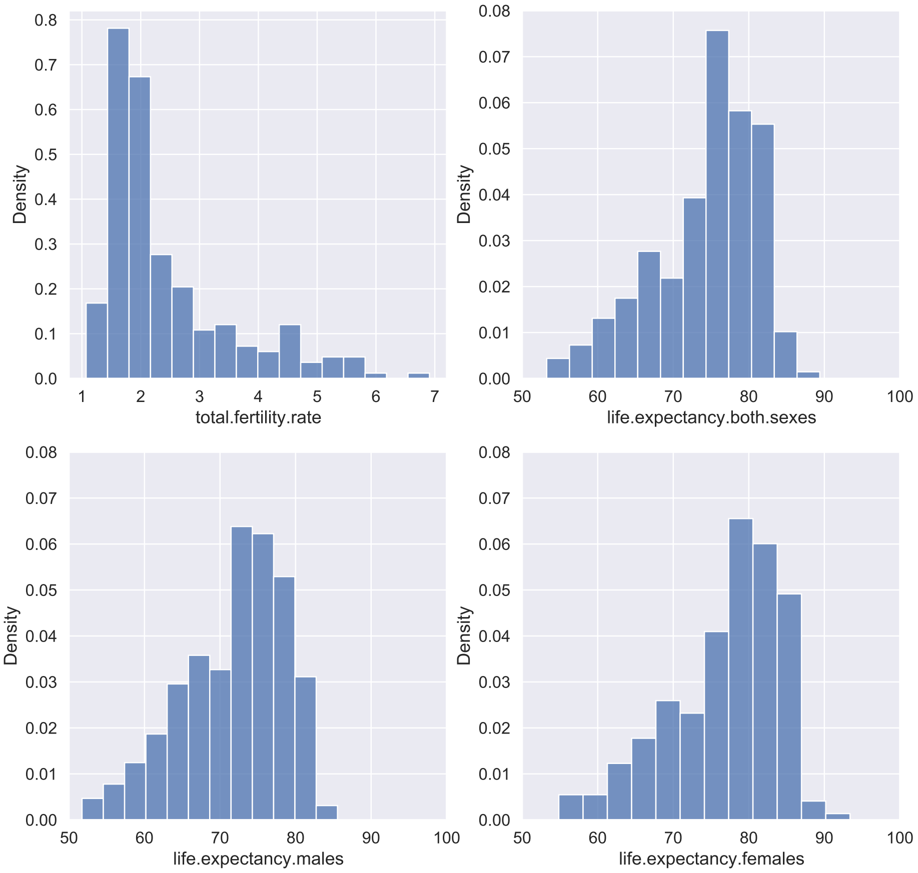


Figure 2: Density Distribution of Numerical Variables

higher than males in the year 2021, this can be inferred from the mean of Table 5 and also from the scatter plot of Figure 3.

4.2 Bi-variate Analysis

After looking at each variable individually, we will now look into the relationship between two variables. As mentioned in the above section, we will use correlation coefficient to check the relation amongst variables. Table 6 provides the coefficient values between the variables and also the same has been visualized using heat map which can also be referred from the appendix (Table 8).

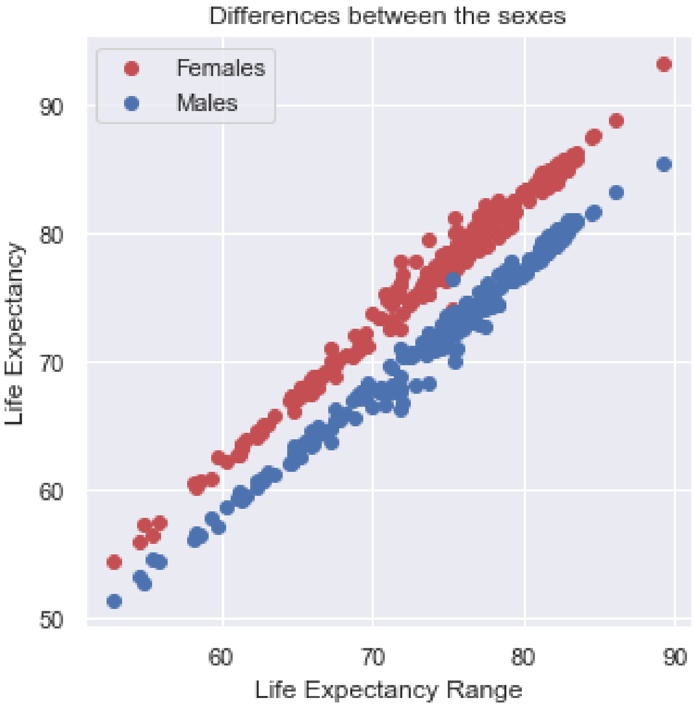
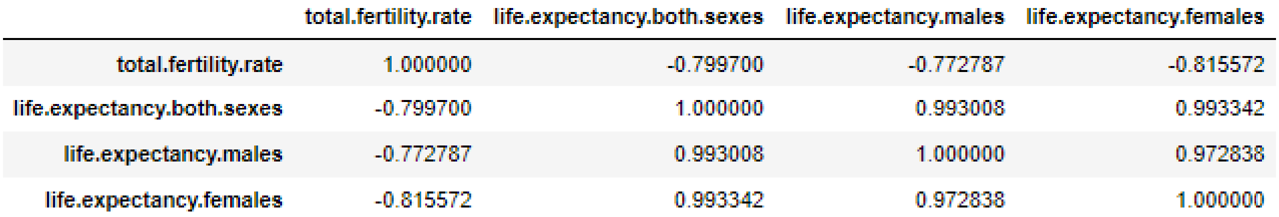


Figure 3: Difference in Sexes in terms of Life Expectancy



	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
total.fertility.rate	1.000000	-0.799700	-0.772787	-0.815572
life.expectancy.both.sexes	-0.799700	1.000000	0.993008	0.993342
life.expectancy.males	-0.772787	0.993008	1.000000	0.972838
life.expectancy.females	-0.815572	0.993342	0.972838	1.000000

Table 6: Calculation of Correlation Coefficients

It can be inferred from Table 6 that the total fertility rate and life expectancy show strong negative correlation trend. This implies that, people who tend to have lesser number of children are expected to live longer and vice-versa. In a more abstract way it can be said that increased life expectancy comes at a cost of decreased fertility rate. Thus the hypothesis that was made in the introduction can be said to be true. In the section below, we will analyze the same data country-wise to get a better understanding.

4.3 Variability of the values in the individual and different Subregions

As mentioned in the Section 2, there are 21 sub regions and 7 regions. It is therefore of great interest to observe the data within and across the sub-region.

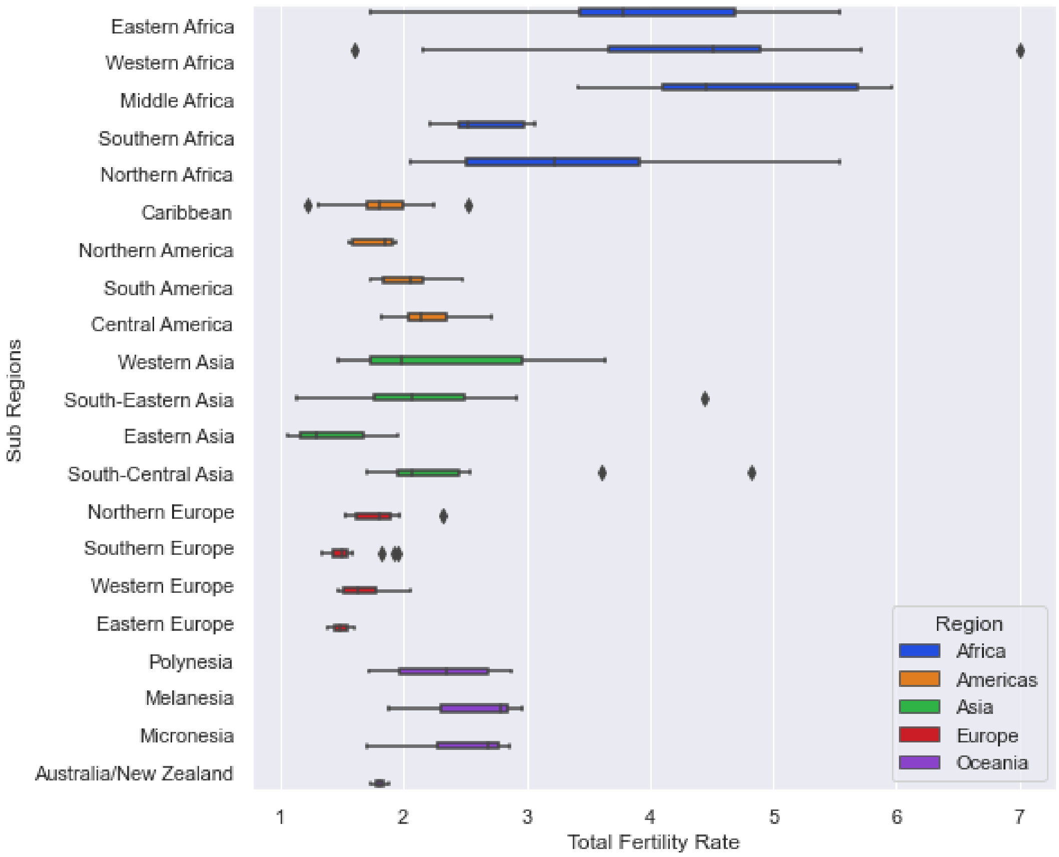


Figure 4: Total Fertility Rate across the world

Below is the inference made from Figure 4,

For the variable total fertility rate, we can observe that Africa region has higher spread of the data or the interquartile range for Africa and its sub-regions is relatively higher as compared to other regions and sub-regions, hence the total fertility rate is not homogeneous within Africa whereas Europe and America region has less variability within their sub-regions among all regions and it is interesting to note that sub-regions of Europe and America has homogeneous data. Oceania and Asia has variability ranging from 1.8 to 3 and 1 to 3.5 respectively. For these two regions, not every sub-regions has homogeneous data, Australia stands out in Oceania as the total fertility rate of Australia is 1.8 with

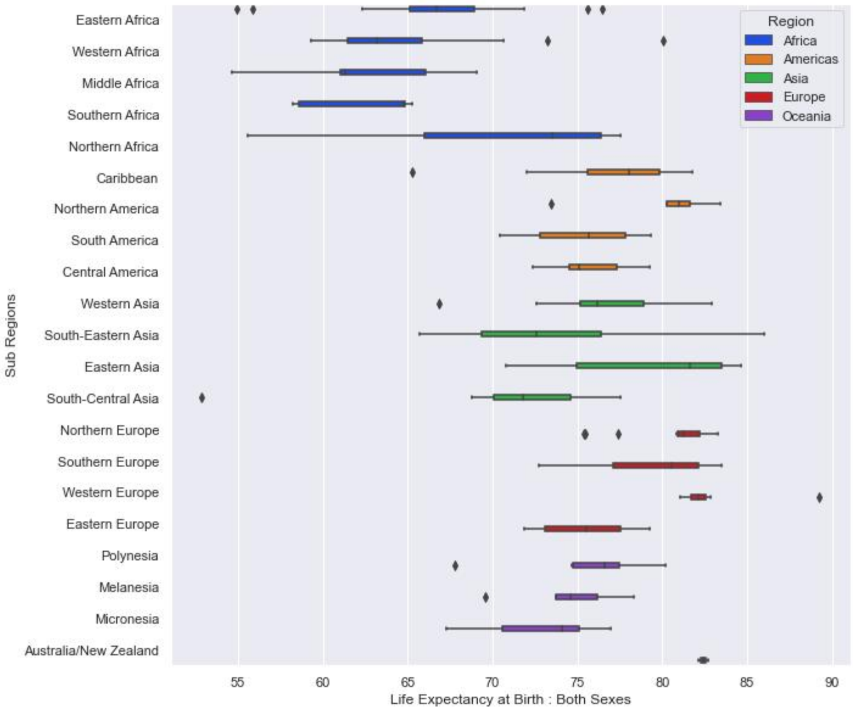


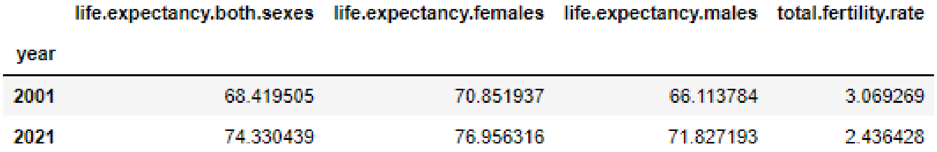
Figure 5: Life Expectancy at birth for both sexes

almost negligible variability. It can also be observed that the minimum total fertility rate for Middle Africa is higher than maximum of total fertility rate of Oceania, America and Europe regions.

Below is the inference made from Figure 5,

For the variable life expectancy at birth, we can observe that here again sub-region of Africa has the highest spread of data as compared to other regions and thus it is heterogeneous. It can also be observed from the Figure 5, Europe has the highest life expectancy at birth whereas Africa has the lowest. This means, that the Africans live for lesser number of years than the rest of the regions. This low life expectancy is due to the major health crisis that is prevailing in Africa. On the other hand, people of Europe live for most number of years due to the excellent health care system. The life expectancy for the regions Asia and America are ranging from 65 to 86 and 70 to 84 respectively, the subregion Northern America is standing out from all the other sub-region of America as it has the most homogeneous life expectancy ranging from 80 to 84.

4.4 Comparison of variables from 2001 to 2021



	life.expectancy.both.sexes	life.expectancy.females	life.expectancy.males	total.fertility.rate
year				
2001	68.419505	70.851937	66.113784	3.069269
2021	74.330439	76.956316	71.827193	2.436428

Table 7: Variable Change over 20 Years

In this section, we will check how have the variables changed over the past 20 years. Table 7 shows the average change in variables. It can be seen from Table 7 that the fertility rate decreased from 3 to 2.4 over the past 20 years and life expectancy increased by 6 years. This is also to be noted that the above statement is for world population and not region/subregion or country wise. It cannot be said that any women picked at

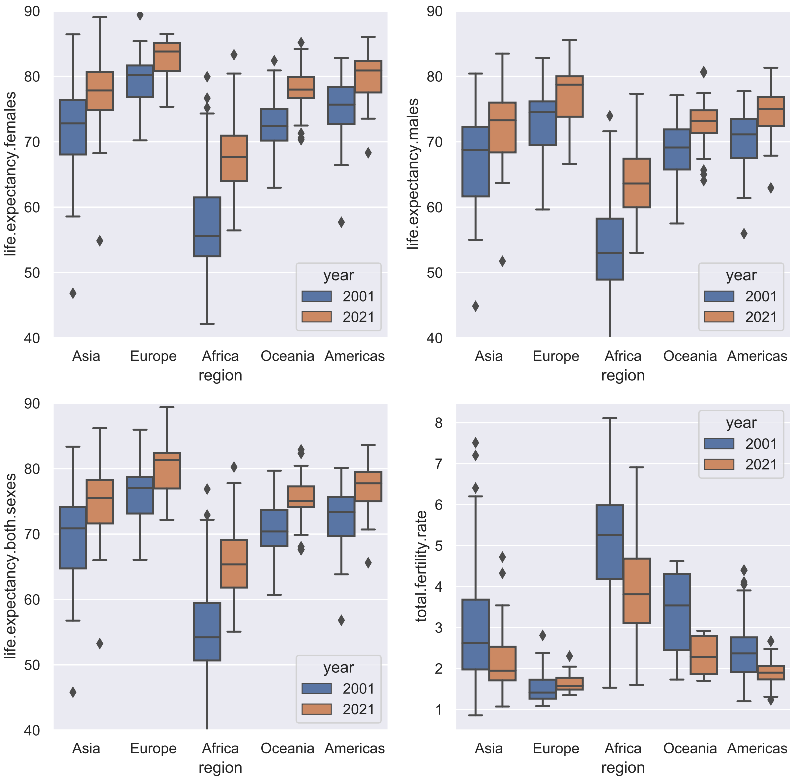


Figure 6: Comparison of change in variables in every Region over Year

random has fertility rate of 2 to 3 because population is different for different countries and hence the fertility rate and life expectancy also varies. We will now analyze the same data 'region-wise'. To do so, we will use box-plot to compare as shown in Figure 6

As observed from Figure 6, the life expectancy increased and the total fertility rate decreased from year 2001 to 2021. We can also observe that for life expectancy, there is a significant increase for Africa region and all the other regions show relatively less increment. On the other hand, for total fertility rate we see a significant decrease for every region except Europe which remains more or less the same even after 20 years. The following statement can also be made based on the observation from Figure 6 that the average fertility rate for European region is lower than the world's average fertility rate and Africa has higher fertility rate as compared to world's average fertility rate.

5 Summary

In this report the data-set was extracted from (International Data Base) of the U.S. Census Bureau. The data set contains 9 features and 456 records. Out of 9 features, 4 were numerical and 5 were categorical variables. It contains information about 228 Countries divided demographically into 21 sub regions and 5 regions. The entire dataset provided is for the year 2001 and 2021. The main objective of this report is to provide a detailed descriptive analysis of the data. To do so, we have used Jupyter Notebook software and Python Programming Language. In Python Programming Language, we have used few libraries to perform our analysis such as Pandas, Matplotlib and Seaborn. Few statistical methods were introduced in this report such as Arithmetic Mean, Median, Variance, Standard Deviation, Quantiles and Correlation Coefficient.

The analysis which was made in this report is summarized as follows, global average total fertility rate is around 2.4 and global life expectancy for both sexes is around 74 years. Life Expectancy at Birth is in-turn stratified by sex and it has been observed that the average life expectancy at birth for females is more than the average life expectancy of males and this can be due to many reasons. Few possible reasons are men have more dangerous jobs including firefighting, military combat. Men commit suicide more often than women. Men tend to avoid medical care and regular health check-up as compared to women. (Robert H. Shmerling, MD). Furthermore in the report, we also observed that the life expectancy and total fertility rate in various regions and sub regions. America and Europe tend to show homogeneity and low fertility rate whereas Africa depicts heterogeneity and has high fertility rate. This can be due to lifestyle choices and quality of health care system between different countries. In a more general way, it can be said that the countries with high life expectancy rate prefer having low number of children compared to other countries or vice-versa. Therefore, individual live longer in country with low fertility rate than individual living in a country with high fertility rate. Lastly, the data is also compared for year 2001 and 2021 and we observed that while total fertility rate has decreased in 2021, life expectancy has increased. But this data is not fully appropriate as the population change in the years is not considered.

For further analysis, it would be of more interest and value if the population can also be taken into account for the yearly change in variables. This would lead to actual results and hence it would be of more use to understand the factors which are causing so.

Bibliography

Christopher Hay-Jahans. *An R Companion to Elementary Applied Statistics*. Taylor and Francis Group, London, NewYork, 2018.

(International Data Base). Glossary for census data. URL `https://www.census.gov/glossary/`.

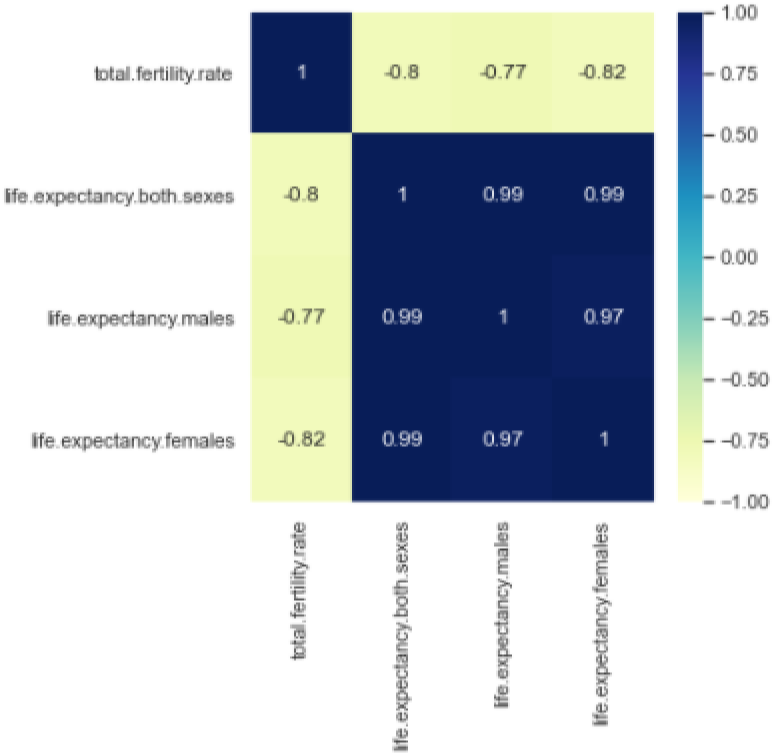
Python Programming Language. Python software foundation. URL *http* : *//www.python.org*.

(Robert H. Shmerling, MD). Why men often die earlier than women. URL `https://www.health.harvard.edu/blog/why-men-often-die-earlier-than-women-2016021991`

(Walter Antoniotti, 2001). *Statistics by Walter Antoniotti*. 21stCenturyLearningProducts, 227 Baboosic Lake Road Merrimack, NH03054.

Appendix

A Additional tables



	total.fertility.rate	life.expectancy.both.sexes	life.expectancy.males	life.expectancy.females
total.fertility.rate	1	-0.8	-0.77	-0.82
life.expectancy.both.sexes	-0.8	1	0.99	0.99
life.expectancy.males	-0.77	0.99	1	0.97
life.expectancy.females	-0.82	0.99	0.97	1

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Table 8: Correlation Coefficient between the variables

B Additional figures

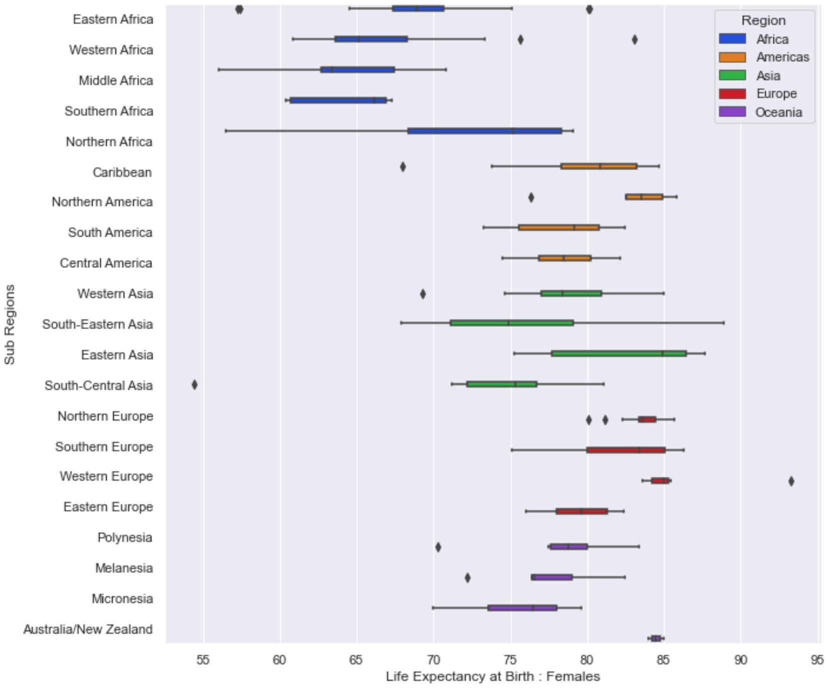


Figure 7: Life Expectancy of Females for 2021 across the world

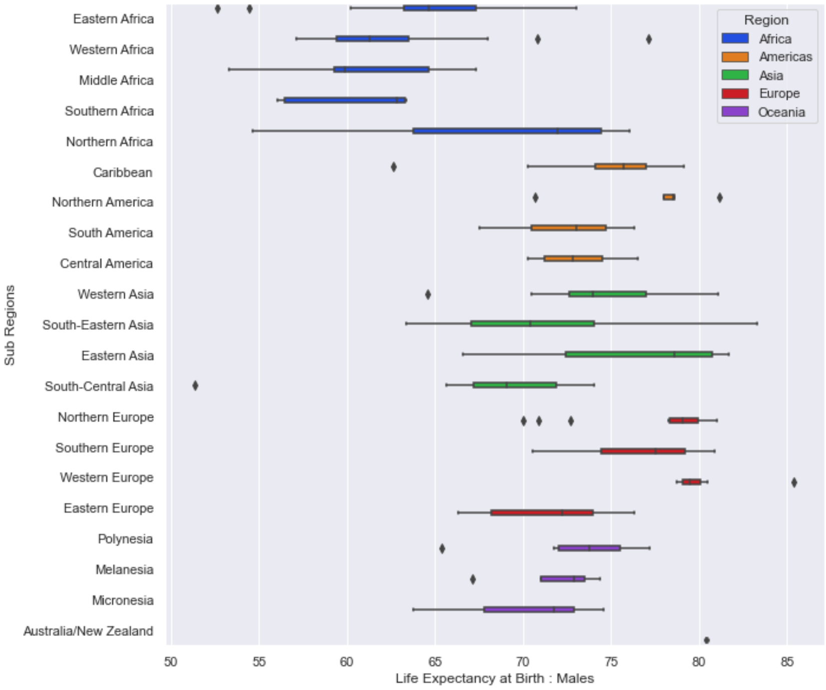


Figure 8: Life Expectancy of Males for 2021 across the world

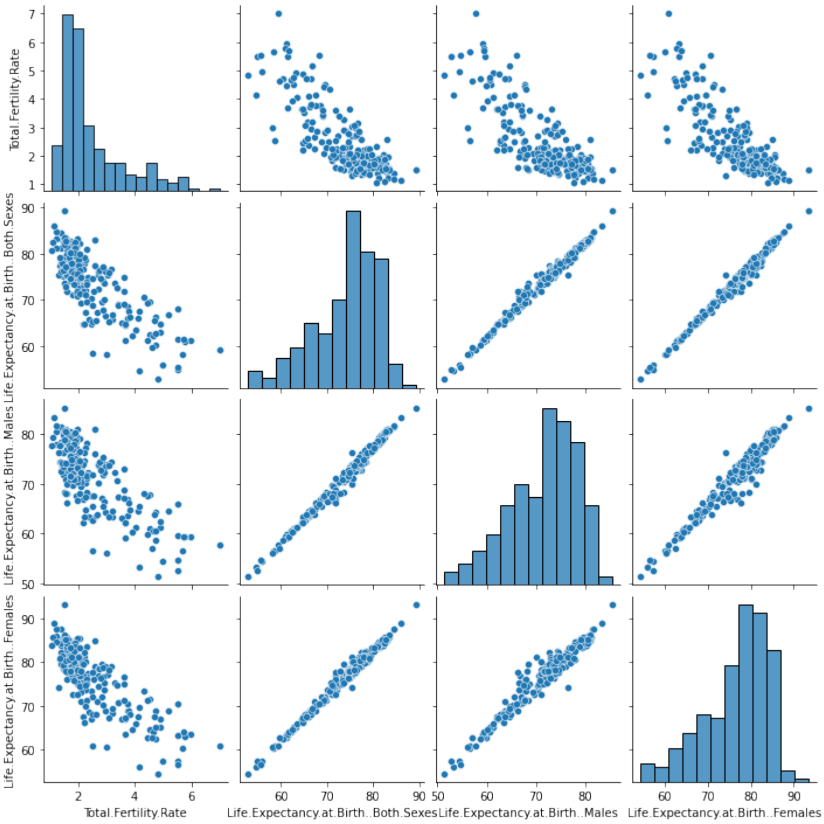


Figure 9: Bi-Variate Analysis of Numerical Data