TU Dortmund

Introductory Case Studies

Project I: Descriptive analysis of Demographic data

Author: Swapnil Srivastava

May 14, 2021

Contents

1 Introduction

Census data are recorded every 10 years, to analyze the population growth on various features. In addition to understand the population health across the continents, change in life expectancy in years can be because of number of factors, including change in living standards, lifestyle and education, as well as access to quality health services. Especially in the current scenario where the world is fighting with the deadly COVID-19, it is of great interest to actually understand this difference, where we stand after 10 years.

In this project, 228 countries of the world are examined to understand the regional differences or similarities in the Life Expectancy and Total fertility rates in 2020 from 2000. For this purpose, first we perform the uni-variate analysis of all the numerical variables for the year 2020. Furthermore, we see the bi-variate correlation between these numerical variables. We also study how these variables are linked within/among the various sub-regions.

Apart from the Introductory section, the project report presented here contains additional 4 sections. The Section 2 deals with the description of the used data set in terms of the definitions of the variables and the quality of the data provided. The Section 3 deals with the explanation of various statistical terms used for analyzing the data like, mean, variance etc. In addition to it the concept of correlation is also introduced here. The next section i.e. Section 4 deals with the interpretation of the final analysis by using the Statistical methods (Section 3) on the provided data set. The last section: Section 5 is the concluding section. It contains the summary of all the interpretations and results. It also provides an insight for the further analysis.

2 Problem statement

2.1 Description of Data Set and Quality

The dataset used in this project contains an extract from the International Data Base (IDB) of the U.S. Census Bureau. It includes data about Life expectancy and Fertility rates for 228 countries from 2000 and 2020. The countries taken into account here are the ones which are recognized by the US Department of State and have population more than or equal to 5000. (International Data Base)

The data set comprises of data from 228 countries which are then clubbed in 21 Sub-regions which are in turn clubbed in 5 regions, demographically. There are 10 columns which are the features for the data set. Each column has different corresponding values to each row which are the observations for the data set being used. The *Country.Area.Name* corresponds to the name of the Country for which the observations are recorded. Federal Information Processing Series (*(FIPS)* Codes are the standardized system which are used to improve the use of data and avoid unnecessary duplication and incompatibility in the collection, processing, and dissemination of data. Geopolitical Entities Names and Code *(GENC)* are also similar demographically assigned codes. The *Subregions* corresponds to the either of 21 Sub region in which that particular country belongs to. *Region* corresponds to the one of the 5 regions in which that Sub region (or Country) belongs to. *Year* states the particular year (2010 or 2020) of which the data is recorded. *Total fertility rate* is the average number of children that would be born per woman provided, all women live to the end of their childbearing years. *Life Expectancy* is 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. This data is recorded in general and then specifically for Males and Females as well differently. (International Data Base)

The data set which is used for this project has 10 feature and 456 observations. Out of the 10 features, 6 are categorical variables and rest 4 are numerical variables. Table 1 below, shows the description of the variables used in data set.

Coulmn Name	Data Type	Type
Country.Area.Name	object	Categorical
FIPS	object	Categorical
GENC	object	Categorical
Subregion	object	Categorical
Region	object	Categorical
Year	int64	Categorical
Total.Fertility.Rate	float64	Numerical
Life.Expectancy.at.Birth..Both.Sexes	float64	Numerical
Life.Expectancy.at.Birth..Males	float64	Numerical
Life.Expectancy.at.Birth..Females	float64	Numerical

Table 1: The description of the data set used.

The Data Quality of the data set is good and there are only 7 missing data for 4 features out of 456 rows. Also, the 2 missing values in the GENC feature is misleading as 'NA' is the encoding present in the feature

2.2 Project Objective

In this project, firstly the Uni -Variate analysis is performed for every Numerical Variable. The frequency distribution is observed for every variable across the world. Since the data set involves Life Expectancy data of people in general, and for female and male separately as well, the difference among the sexes is also observed using scatter plot. Furthermore, the bi-variate correlations between the variables is also observed by finding the correlation coefficient. Since data set contains data of 21 sub regions, the Life Expectancy and Total Fertility Rate for these sub regions are visualized and the results are interpreted. For these tasks, the data for only year 2020 is taken into account. Lastly, the variables are observed over all the Region, how it changed over the 10 years.

3 Statistical methods

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

Arithmetic Mean Arithmetic mean may be defined as the average of all the n observations. 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. Since arithmetic mean depends on each and every sample data observations, hence it is very sensitive to extreme values. We don't prefer finding the central tendency of the data set if our data set is not symmetric or homogeneous. (Christopher Hay-Jahans, 2020).

Median Median may be defined as computing the central value of the data set. For its computation, the above sample $x_1, x_2, .., x_n$ is assumed to be sorted in ascending order. $Median(\bar{x})$, is then calculated as:

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

Contrary to mean, *Median* does not depend on the values of whole data set but only the central value instead. Hence it is not sensitive in case of non symmetrical or heterogeneous data, making it a preferable choice for computing the central tendency in such cases. (Christopher Hay-Jahans, 2020).

Frequency It gives the number of occurrences of a particular data in whole data set. Like for ex, if our data set is : $x_1, x_1, x_2, x_2, x_2, x_2$, frequency of x_1 is 2 while that of x_2 is 4.

3.1.2 Measure of Spread

Standard Deviation and Variance It can be defined as the measure of how far the data is from the mean i.e. measuring the spread of the given data set with respect to the mean. *Sample* variance (s^2) and Standard Deviation (s) can be calculated as:

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

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

For a *known* population, *population* Variance (σ^2) and *population* Standard Deviation (σ) can also be computed using the above formula. (Christopher Hay-Jahans, 2020).

3.1.3 Measure of Position

Percentile and Quantiles In general, the p^{th} percentile is the number q_p that corresponds to a percentile rank 100 * p% and is obtained by :

$$q_p = \begin{cases} x_k, & \text{where } k = \lceil np \rceil \; and \; np \, is \, not \, an \, integer \\ \frac{(x_k + x_{k+1})}{2}, & \text{where } k = \; np \, and \, np \, is \, an \, integer \end{cases}$$

Following are the three major Quartiles which are of high significance :

- First Quartile (Q_1) = 25^{th} Percentile
- Second Quartile (Q_2 *or Median*) = 50^{th} Percentile
- Third Quartile (Q_3) = 75^{th} Percentile. (Ludwig Fahrmeir, 2020)

3.2 Bi-variate

3.2.1 Correlation

It can be defined as the statistical relationship between two variables. We use correlation coefficients (r) in order to calculate the strength of relationship between them. It can be calculated as:

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

Here the positive value of r shows that the 2 variables are positively correlated where as a negative value shows that they are negatively correlated. The coefficient has possible values in the range from -1 to +1, where *higher* the absolute value of coefficient , *higher* is the strength of correlation. (Ludwig Fahrmeir, 2020)

4 Statistical analysis

All the following measures and plots were created using the software Python in version 3.8.5 (Jupyter Notebooks)

Description of Data Set 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:

	Year	Total.Fertility.Rate	Life.Expectancy.at.Birth..Both.Sexes	Life.Expectancy.at.Birth..Males	Life.Expectancy.at.Birth..Females
count	456.000000	449.000000	449.000000	449.000000	449.000000
mean	2010.000000	2.781093	71.135635	68.723296	73.673474
std	10.010983	1.478454	8.983394	8.658055	9.437928
min	2000.000000	0.892400	39.840000	37.470000	42.290000
25%	2000.000000	1.710000	66.740000	64.560000	68.880000
50%	2010.000000	2.191000	73.430000	70.860000	76.130000
75%	2020.000000	3.603500	77.500000	74.660000	80.550000
max	2020.000000	8.090000	89.270000	85.400000	93.300000

Table 2: Description of Data Set

4.1 Uni-variate Analysis

As mentioned in Section 3.1, for the Uni-variate analysis the numerical features for the year 2020 are used.

Frequency Distribution of the Uni-Variate Variables It can be graphically represented as below:

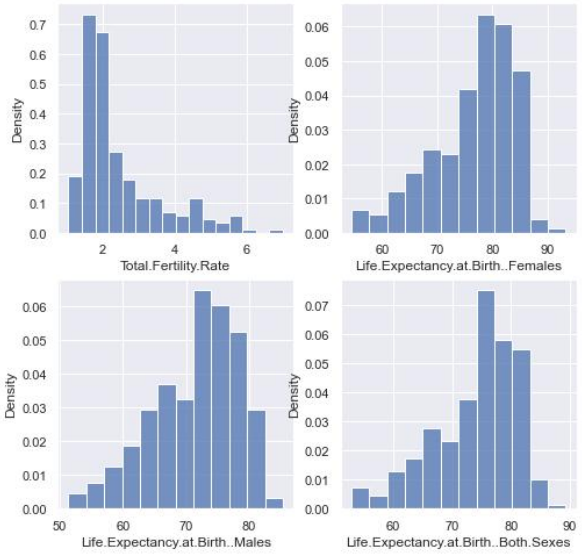

Figure 1: Frequency Distribution of Numerical Variables

We can see from the above table and graph that the *frequency* of Total Fertility Rate is maximum around 2. This means that maximum number of women around the continents for the year 2020 had 2 children. While it is also observed that there are few who had lesser number of children than 2 but there are very few who had children more than 2 and almost null have more than 6. The frequency distribution of Life Expectancy (*for both sexes*) is maximum around 75. This means that maximum number of people, in

general born in year 2020 are expected to live around 75 years. It can also be concluded that few people live below 75 years but there are very few people who are expected to live more than 90 years. The **Differences in sexes** in terms of Life Expectancy, can be observed clearly in the below graph:

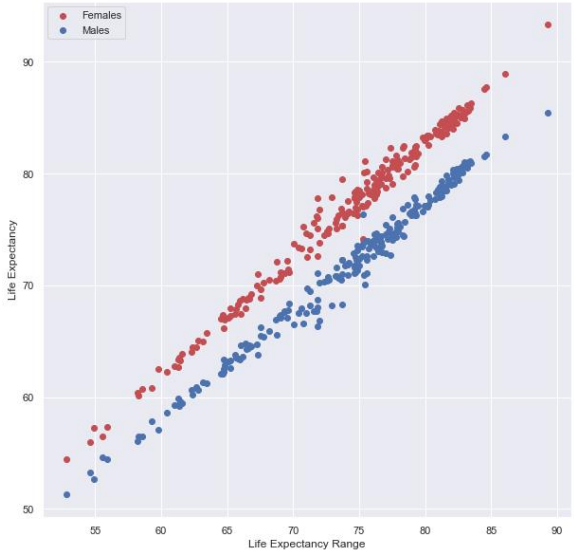

Figure 2: Difference in Sexes in terms of Life Expectancy

We can observe in Figure 2 that the red line (*female*) is slightly above than the Blue Line (*male*) in Figure 2, and in Figure 1 the frequency of Women for the same is the highest at around age 85 while that of men is the highest around 72. Hence, it can be concluded that women have *higher* life expectancy than men in year 2020. While there are number of on going researches and theories to support this, we are yet to receive a conclusive reason of the above conclusion.

4.2 Bi-variate Analysis

The Correlation coefficients were calculated for all the numerical features in the data set. It can be seen in the below Figure 3. The numerical values in the figure represent the value of the correlation coefficient and the color represent the strength of the correlation. The positive value represent Positive correlation while Negative the negative correlation. Also, darker the color, higher is the strength of correlation between the variables. The diagonal of the below figure contains only 1 as it represents the correlation of the variable with itself, hence the strongest.

Figure 3: Bi-variate Analysis of the Numerical Feature

It is important to note that in general the Life Expectancy is negatively correlated to Total Fertility Rate. This implies that in general, people who tend to have lesser number of children are expected to live longer and vice-versa. There is an interesting theory by (Fabrizio Ardiles) that this may be because people who have lesser number of children give greater importance to their own quality of life, since having a children represents a significant cost and hence tend to live longer.

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.

Total Fertility Rate (TFR): The Total Fertility Rate among various Sub regions can be visualized in Figure 4. We can observe that within *Africa*, except for Northern Africa, rest every Sub-region has its median value inclined to one of the Quartiles, hence the Total Fertility rate is *not* homogeneous within Africa and except for Southern Africa, it can be observed that there is a large variability of data. Where there are 2 extremes for Western Africa having Total Fertility Rate as 7 (*Niger*) and 1.5 (*Saint Helena, Ascension, and Tristan da Cunha*), the rest of the Africa ranges between 1.8 to 6. On

the other hand, we see a *less* variability of TFR among all the sub regions of *Americas*. Most of the Americas have number of children around 2 and none have more than 3. If we discard *North America*, we can say that rest of the America have *almost* homogeneous TFR. Similarly, *Europe* and *Oceania* also have less variability of TFR ranging from 1.5 to 2 and 1.8 to 3 *respectively*. While it is interesting to note that almost every sub region of Europe has *homogeneous* data, Oceania does not. *Australia/ New Zealand* stands out in Oceania as *almost* every person in here has TFR has 1.8 and hence *homogeneous*. The most *homogeneous* TFR is observed within Asia with few extremes. Except for few countries like *Timor-Leste* and *Afghanistan* which have TFR more than 4, rest of Asia has it below 4. Hence it is safe to conclude that *Africa* has the most variability of data while *Europe* has the least. With TFR as 7, *Niger* of Western Africa has the maximum, while *Taiwan* of Eastern Asia has the lowest TFR of 1.06 across the continents.

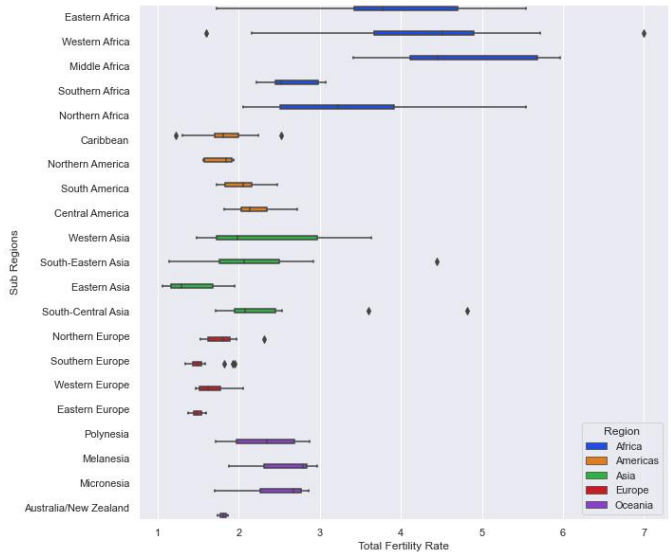

Figure 4: Total Fertility Rate across the world

Life Expectancy at Birth: The Life Expectancy at Birth for both the sexes can be visualized in Figure 5. It can be observed that Unlike Total fertility Rate, Life Expectancy is *almost* homogeneously distributed within the sub regions. While the most variability is observed in Africa, the least is observed in Europe. In *Africa* we see a large range Life Expectancy from 55 to 80, while in *Europe*, it is mostly between 72 to around 82 with an exception of *Monaco* in Europe of 89 Life Expectancy. In general, the *least* Life Expectancy is observed in Africa, the most is observed in Europe. This means, that the

people of Africa live for lesser number of years than the rest of the continent. 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 owing to the excellent health care system there. An Exception of *Afghanistan* can be observed in South central Asia which has the least Life Expectancy of 52 years. As per news, Children here suffer from malnutrition which is the most important reason of this low rate. The Life Expectancy Rate of year 2020 was also affected due to the on going pandemic of COVID-19. While most of *Oceania* ranges from 68 to 80, *America* ranges from 70 to 85 with few exceptions. In *Asia*, the large variability of Life Expectancy can be observed in *South-Central Asia* which ranges from approximately 65 to 85. Hence to conclude, where almost every sub region is homogeneous within itself, *Oceania* is also more homogeneous among every sub region, while *Africa* the least.

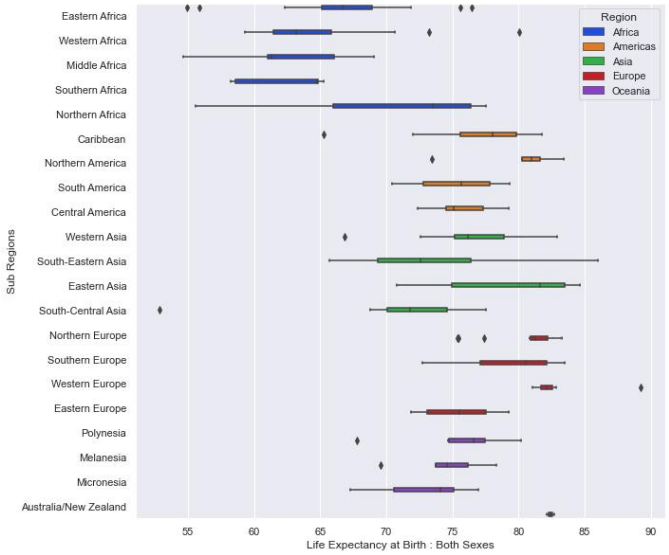

Figure 5: Life Expectancy at Birth for Both the sexes

4.4 Comparison of variables from 2000 to 2020

Arithmetic Mean: As mentioned in the table below, it can be observed that in general, the Total Fertility Rate is decreased from 2000 to 2020 while the Life Expectancy increased from 2000 to 2020. But it is also important to note that this data might be misleading as the population change has not been accounted into.

Year	Life.Expectancy.at.Birth..Both.Sexes	Life.Expectancy.at.Birth..Females	Life.Expectancy.at.Birth..Males	Total.Fertility.Rate
2000	68.142443	70.585882	65.828462	3.111771
2020	74.036930	76.666272	71.529254	2.460568

Table 3: Calculation of Mean of various numerical variables for 2020 and 2000

Numerical Variables: As observed in Figure 6, where the Life Expectancy *increased* from 2000 to 2020, the Total Fertility Rate *decreased* from 2000 to 2020. We can also observe that for Life Expectancy, there is a *significant* increase for Africa, all the other regions show similar lesser 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 10 years.

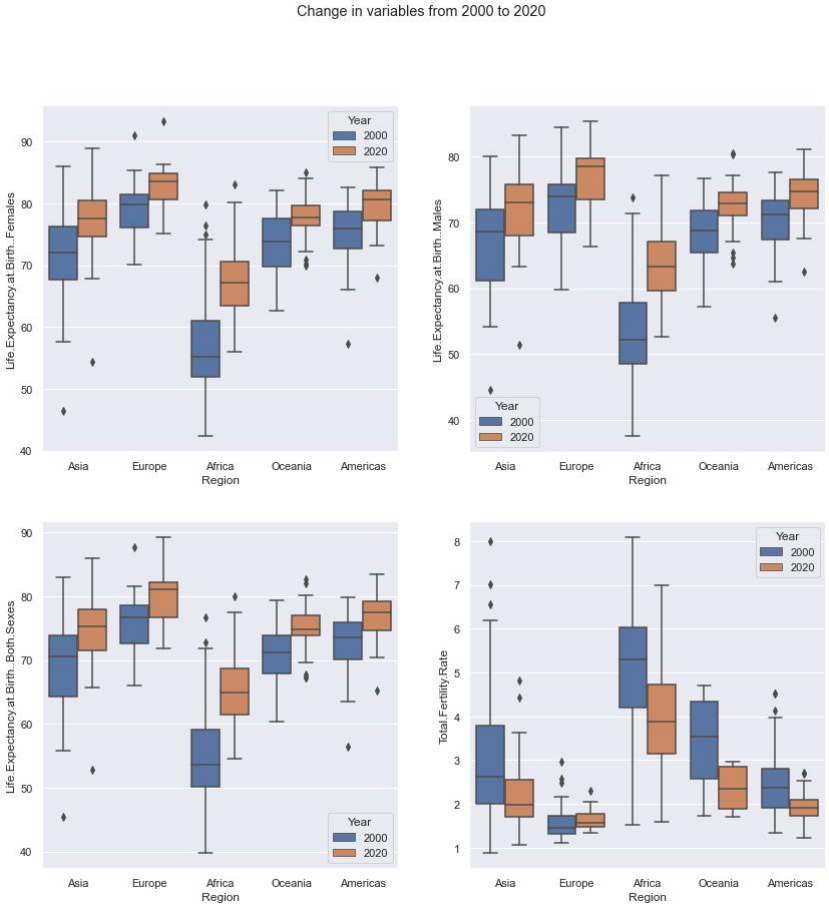

Figure 6: Comparison of change in variables in every Region over Year

5 Summary

This project report basically deals with the Descriptive analysis of the Data Set given. The Data Set was compiled by the instructors of the course Introductory Case Studies at TU Dortmund University (Summer Semester 21). The Data set is a small extract from the *International Data Base* of the *U.S Census Bureau* which contains information from census,surveys etc. It contains information about 228 countries divided demographically into 21 sub regions and 7 regions. It has 456 observations over 10 features including data from 2 years i.e. 2000 and 2020.

The report provides *Uni-variate* and *Bi-variate* analysis of all the numerical variables involved, basically *Total Fertility Rate* and *Life Expectancy at Birth* which is in-turn stratified by sex. It is here that we observe that women tend to live *longer* than men owing to many possible theories and researches. One of the possible theory includes that since the Gender Discrimination still prevails in our society, we see number of jobs where sex ratio is imbalanced. Men, tend to work more in stressful jobs and have the burden to be the only bread owner of the family. Hence due to more stress and more open to world, they tend to live shorter than women, in general. (Robert H. Shmerling, MD) Furthermore in the report, we also observe the Life Expectancy and Total Fertility Rate in various regions and sub regions. We can see that Africa has the *least* Life Expectancy and Europe has the *most*. This is significantly due to the difference in the quality of health care system in the two regions. Lastly, the data is also compared in the 2 years and we find out that while Total Fertility Rate has *decreased* in 2020, 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, 2020.

Fabrizio Ardiles. Correlation between life expectancy and fertility. URL *http* : $//vox.lacea.org/?q = blog/life_expectancy_fertility$.

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

Thomas Kneib Ludwig Fahrmeir. *Regression Models, Methods and Applications*. Springer, London, NewYork, 2020.

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`

Appendix

A Additional tables

	Total.Fertility.Rate	Life.Expectancy.at.Birth..Both.Sexes	Life.Expectancy.at.Birth..Males	Life.Expectancy.at.Birth..Females
Total.Fertility.Rate	1.000000	-0.802146	-0.774607	-0.818479
Life.Expectancy.at.Birth..Both.Sexes	-0.802146	1.000000	0.992988	0.993381
Life.Expectancy.at.Birth..Males	-0.774607	0.992988	1.000000	0.972876
Life.Expectancy.at.Birth..Females	-0.818479	0.993381	0.972876	1.000000

Table 4: Calculation of Correlation Coefficients

B Additional figures

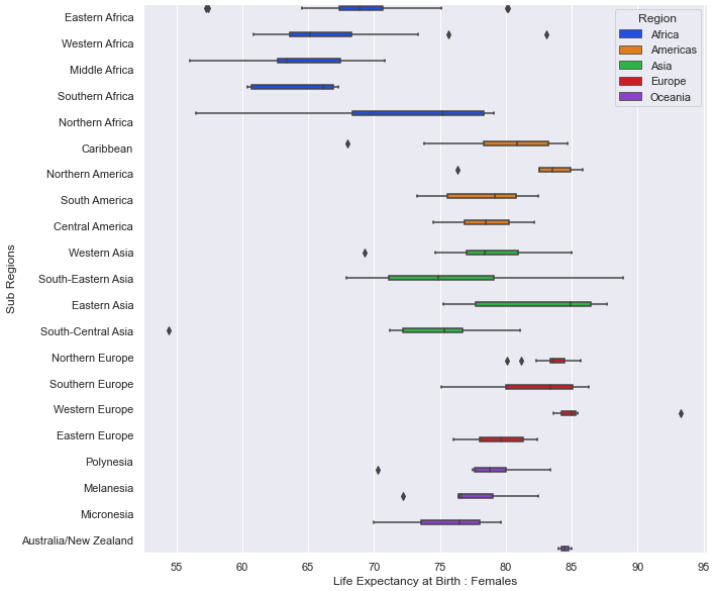

Figure 7: Life Expectancy of Females for 2020 across the world

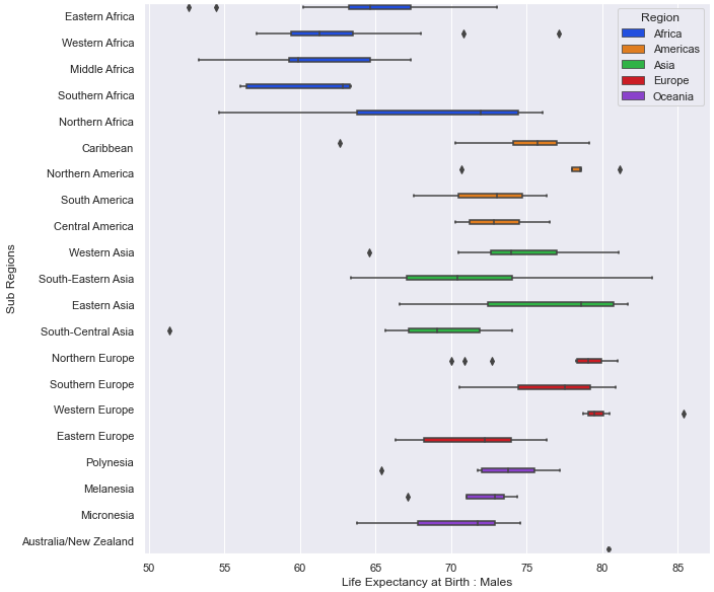

Figure 8: Life Expectancy of Males for 2020 across the world

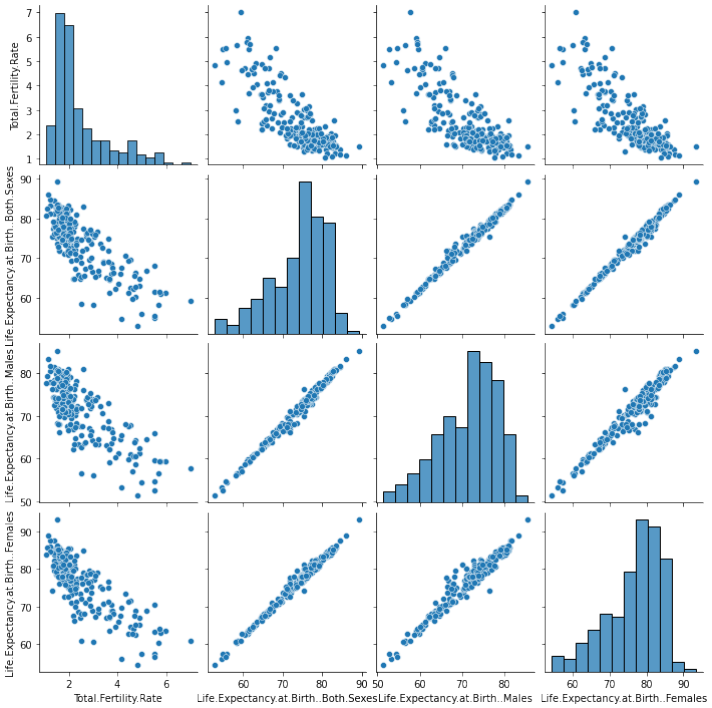

Figure 9: Bi-Variate Analysis of Numerical Data