TU Dortmund

Introductory Case Studies

Project 1: Descriptive Analysis of Demographic data

Lecturers:

Prof. Dr. Katja Ickstadt

M. Sc. Zeyu Ding

M. Sc. Marvin Napps

Author: Sagar Basnet (Matr. No : 236912)

Group number: 6

Group members: Siddhartha Karki, Subarna Subedi, Mukta Ghosh

November 13, 2023

Contents

1 Introduction

Demographical data refers to a compilation of information about a population based on various regions, subregions, and countries. This type of data encompasses a range of statistics, such as life expectancy, mortality rate, birth rate and ages of population. The statistical method employed for analyzing this type of data is known as Demographic Analysis or Demographics. Demographical data is valuable for several applications, including policy development and economic market research, as it provides insights into the main characteristics of a population(Bureau, 2022).

In addition to demographical data, we are performing Descriptive analysis on the given data set. The data set in this case study is an extraction from the original data that is collected by IDB (International Data Base) of the U.S. Census Bureau. The extracted dataset is generated and given to us by Statistical department of TU Dortmund for the course ICS (Introductory Case Study) as a part of our first case study.

Descriptive Analysis means analysis of data that helps to represents the summary of data points in such a way that it describes details of information from the provided data. It helps us to reach to the conclusion by providing the features of detecting typos and outliers, identifying the similarities between the variables which help us to perform statistical analysis.

In this case study, we are performing analysis based on the data extracted from U.S. Census Bureau that contains the data from two different year i.e. year 2003 and year 2023. The data set contains some meaning column to be calculated. The variables in this table are: '*Country name*', '*Region*', '*Sub-region*', '*Year*', '*Medium Age of Both Sexes*', '*Medium Age of Females*', '*Medium Age of Males*', '*Total Fertility Rate*', '*Infant Mortality Rate of Both Sexes*','*Infant Mortality Rate of Females*' and '*Infant Mortality Rate of Males*'. Data Set contains 453 entries including both year 2003 and 2023 and 225 prepared entries for year 2023. The main purpose behind working in these data is to find the frequency distribution of individual variables i.e. how the data are distributed, showing the differences of genders and regions, to calculate the homogeneity of individual variables within the individual subregions and heterogeneity of variables between different subregions. Similarly, we find if there exist any bi-variate correlation relationship between the variables (concerned data are Median Age and Infant Mortality Rate). And, we compute the differences in the change of values of variables in between the year 2003 and 2023.

To perform frequency distribution, bi-variate correlation and homogeneity of variables, we use the data only from the year 2023 which is extracted from the real data provided from Census Bureau. We perform frequency distribution by using 'histogram' and mean frequency distribution method in it. We also show the 'boxplot' to show the differences of genders and regions. For analyzing the homogeneity of individual variables in region Africa and Europe within the subregions and heterogeneity between different subregions, we perform central tendency i.e. Mean, Median and moreover standard deviation. Similarly, we analyze bi-variate correlation relationship to show how the variables are related to each other. Finally, we compute the differences in value of variables in between 2003 and 2023 years.

We are performing this case study, using the references from various sections. Section 2 consist of "Problem Statement". We discuss about the data, how the data is extracted and how the data can be error free in this section. We also discuss about the units and meaning of individual variables. In the section 3, we discuss about "Statistical Methods". In this section, we show the description of statistical method that are used in this case study. For example: we discuss about the method and technique of mean, correlation, mode and homogeneity. Similarly, the section 4 consists of "Statistical Analysis". In this section, we show the methods explained above in section 2 with graphical and tabular representation which proves the existence of "Statistical Method". The assumptions created in section 2 are checked here. And finally, this section consists of illustration and explanation of outcome. Finally, we have section 5 "Summary", where we summarize the whole report clarifying the method used for problems with the outcomes.

2 Problem Statement

Analysis of demographic data from the extracted data has some problems in it because of data set and its quality.

2.1 Data Set and Data Quality

This case study contains the dataset extracted from the demographical data which is managed by IDB (International Data Base) of the U.S. Census Bureau. IDB has performed many surveys among all over 200 countries to build this database. (Bureau, 2022)

The provided data set contains median age, total fertility rate and infant mortality rate from the year 2003 and 2023. The data set here is categorized into 227 countries which is later also sub-categorized by regions and subregions. It contains 5 regions and 21 subregions. Here, "*Median Age*" is divided into three columns: "*Median Age of both sexes*", "*Median Age of males*" and "*Median age of females*". And, "*Infant Mortality rate*" is also divided into three columns: "*Infant Mortality rate of both sexes*", "*Infant Mortality rate of males*" and "*Infant Mortality rate of females*".The characteristics of given dataset and variable is listed in a table below:

Table 1: Characteristics of variables

Variable Name	**Data Type**	**Description**
Country	Categorical	Names of 225 countries
Region	Categorical	It includes 5 different continents
Sub region	Categorical	It includes a part of continents based on location
Median age of both sexes	Numeric	It includes median age of both sexes
Median age of males	Numeric	It includes median age of males of all regions and sub regions
Median age of females	Numeric	It includes median age of females of all regions and sub regions
Infant Mortality rate of both sexes	Numeric	The average number of infants dying before reaching 1 year of age
Infant Mortality rate of males	Numeric	The average number of male infants dying before reaching 1 year of age
Infant Mortality rate of females	Numeric	The average number of female infants dying before reaching 1 year of age
Total Fertility Rate	Numeric	The total number of children that would be born to each woman

While checking the information of variables, we found 7 special characters values in the variables "*Median Age*". The unidentified values seems to be smaller portion as compared to the original data set. So, we cleaned data by deleting the rows containing that value and yet still can have good quality of data.

3 Project Objective

This case study aims to perform a statistical analysis on demographic data by using descriptive analysis techniques. First, the frequency distribution of the variables is examined using mean, quartiles, minimum, and maximum statistical methods. Then, bivariate correlations are established between variables by using the Pearson's Correlation Coefficient technique. Each variable is correlated with the other variables individually. For example, the correlation between "*Median age of both sexes*" and other variables is computed in the same way as for the other variables.In addition, the homogeneity relationship between variables within and among subregions is examined by computing central tendency. Finally, the data for the years 2003 and 2023 are compared to assess the change in the values of variables between those years. Further details about the methods and analysis for achieving these objectives are explained in the subsequent sections of this report.

4 Statistical Methods

4.1 Central Tendency

A measure of central tendency, also known as measures of center or central location, is a summary statistic that aims to represent the middle or center of a distribution by using a single value (Hartmann and Waske, 2018). This value is intended to describe the entire dataset as a whole.

In this case study, the performance of mean and median is done for the finding the frequency distribution of variables.

4.1.1 Mean

Mean is one of the central tendency measurements. The mean (arithmetic mean) calculated for sample data is denoted by $\overline{x}$ (Hartmann and Waske, 2018).

Let $x_1, x_2, \ldots, x_n$ be 'n' number of data then, mean is given by:

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

where $x_1, x_2, \ldots, x_n$ are sample data from 'n' observed data.

4.1.2 Median

Median can be defined by the middle value of ordered data. The data is first arranged in ascending order i.e., from smallest to largest. Median is denoted by M_d and is given by:

$$M_d = \frac{n+1}{2}$$

Here, the way of finding median differs from the size of samples. If the sample size is odd, then the result value of M_d is the location of median value. And if the sample size is even, then the result indicates the mid value between two data. The mean of two data is produced which is median value(State, 2022).

4.2 Variance

In this case study, variance is used for finding the homogeneity between the individual variables within the subregions. Variance is a statistical technique employed to compute the anticipated deviation between values within a dataset(WallstreetMojo, 2021). The primary goal of this technique is to ascertain the overall spread or variability among the variables. It is denoted by σ^2 and is mathematically given by :

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

where 'N' is total number of values in data set, $\overline{x}$ is mean and x_i is the values in data set.

4.3 Inter-Quartile Range (IQR)

Inter-Quartile Range in descriptive analysis tells how the data is spread in middle half of distribution.(Bhandari, 2022)

It is denoted by IQR and is given by:

$$IQR = Q_3 - Q_1$$

where, Q_1 (first quartile) indicates the how the data is distributed in first 25% of distribution and Q_3 (third quartile) indicates how the data is distributed in 75% of distribution.

4.4 Correlation Coefficient

The Correlation Coefficient is a statistical method used to assess the relationship between two variables in a dataset. It is represented by a value that ranges from +1 to -1. A value close to +1 indicates a strong positive relationship between the variables, while a value close to -1 indicates a strong negative relationship(Glen, 2022). Conversely, a value close to 0 suggests no significant relationship between the variables.

Several methods can be employed to calculate the correlation coefficient, and one of these methods is Pearson's Correlation Coefficient, which will be further elaborated upon.

4.4.1 Pearson's Correlation Coefficient

Pearson's Correlation Coefficient is one of the widely used correlation coefficient approach. It helps to show if there is any linear relationship between any two variables(Glen, 2022). It is denoted by "r" and is given by:

$$r_{xy} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum(x^2) - (\sum x)^2}\sqrt{n\sum(y^2) - (\sum y)^2}}$$

where x=$x_1, x_2, \ldots, x_n$ and y=$y_1, y_2, \ldots, y_n$ are sample data and 'n' is the number of observation of sample data.

4.5 Graphical Method

In this section, we discuss about the graphs and plot used in the case study to explain the solution of found problems.

4.5.1 Histogram

A histogram is an excellent visualizing tool that understands the probabilistic distribution of numerical data or image data. It estimates the probabilistic distribution of continuous variable. It is constructed by initializing the bins values (class) of numerical

data. Once bins are set, it counts the number of values that falls in each bins or class and plot it in a graph.The histogram sums the area of bars with each individual class or bins(Aditya, 2019).

4.5.2 Box plot

A box plot is a graphical way of displaying the distribution of data based on minimum value, First Quartile(Q1), Median, Third Quartile(Q3) and maximum value. A box plot can show how the data is distributed i.e., if the data is symmetrical. It also shows how the data are grouped and how the data are skewed(Galarnyk, 2022).

4.5.3 Scatter plot

A scatter plot is a graphical representation of showing the relationship between two variables(Academy, 2017). Variables are represented by x-axis and y-axis and is plotted as (x,y). Scatter plot shows the correlation between two variables.If the value of x increases with increase in value of y, it indicates positive correlation. Similarly, if the value of x increases with the decrease in value of y, it indicates negative correlation. And if, it does not show clear relationship then it indicates zero correlation.

5 Statistical Analysis

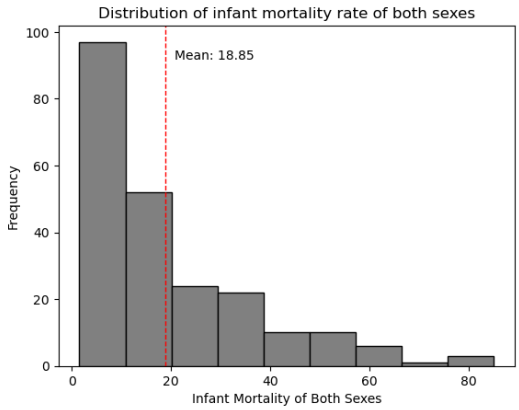
In this section, the statistical methods explained above are applied to the presented data set and the results are interpreted. For all calculations and visualization, python with packages like numpy, pandas, matplotlib and seaborn are used.

5.1 Frequency Distribution of variables

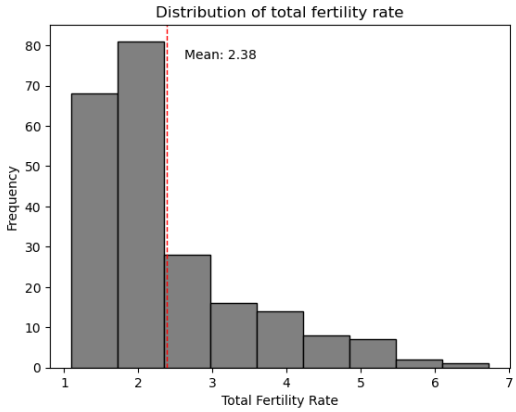
This case study includes histogram to show how the frequency distribution of given variables looks like.

From figure 1 (a), it is concluded that in the year '2023' the mean of infant mortality rate is found to be 18.85 and distribution in between the 0-10 per 1000 live birth is found in 98 countries. Similarly, the mortality rate of 20-40 per 1000 live birth is found in 75 countries and the mortality rate of 60-80 per 1000 live birth is only found in 9

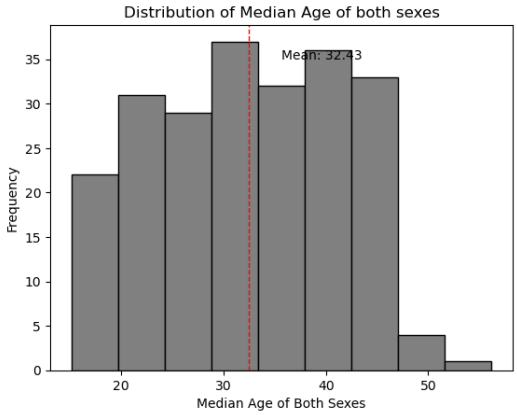
countries. From figure 1 (b), it is found that the mean of total fertility rate is 2.38 i.e.



(a) Infant Mortality rate of both sexes



(b) Total Fertility Rate



(c) Median Age of both sexes

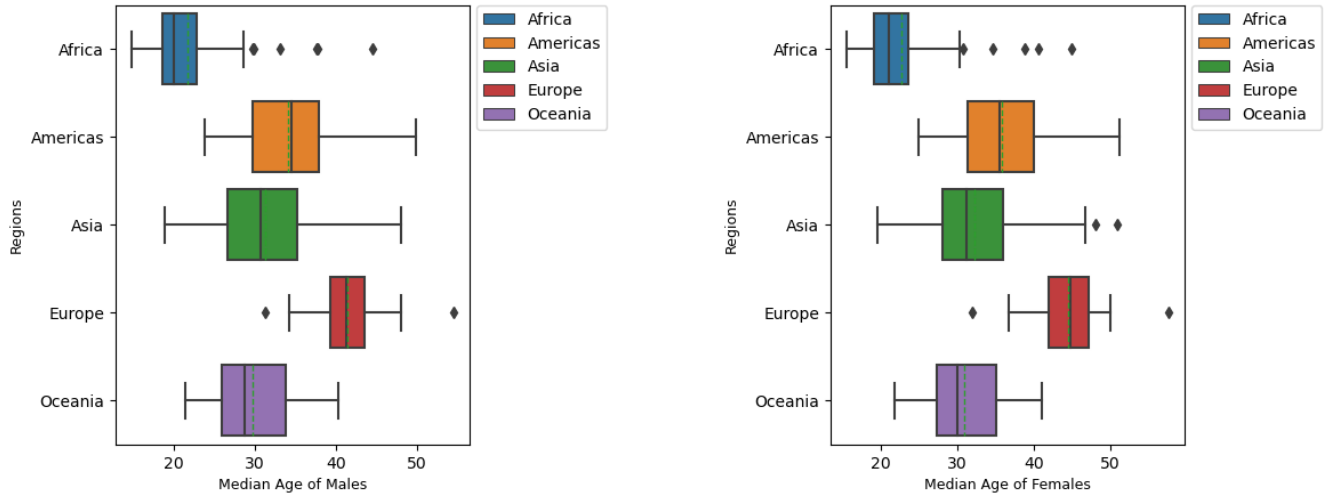
Figure 1: Frequency distribution of variables

the average 2.38 number of children are born to a woman over her lifetime. Moreover, it is also found that 2 children are born to a woman over her lifetime in 80 countries, 1 in 67 countries and 3-6 in 78 countries. And from figure 1 (c), it is seen that the mean of median age of both sexes is found be around 32.43 which means half of the world's population was below the age of 31 and the other half was above it.

5.2 Differences between the sexes and regions

In this subsection, it is shown that how the variables of two different genders are different within and between regions.

Figure 2 (a) and (b) represents the median ages of males and females in different regions. From figure 2 (a) and 2(b), it is found that the mean of median age of males and females is found higher in Europe as compared to others regions i.e. 41.48 and 44.46, which means around 50% of population from males and females are young in Europe and other 50% are considered to be old. Similarly, it is followed by the America with the mean value of 34.07 and 35.81 and Asia with mean value of 31.23 and 32.18. From the analysis, Africa is found to have less mean value of median age of males i.e. 21.70 and 22.61 which means half of the population of male is below the age of 21.70 and female is below 22.61 and half of the population are found to be above 22.61 and 21.70.

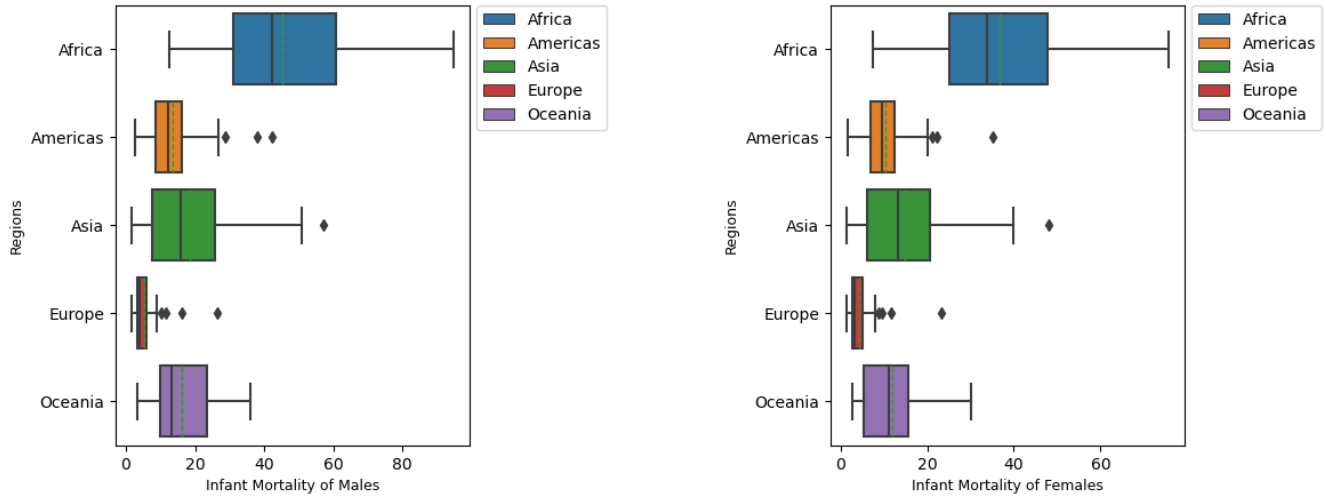


(a) Median age of males in different regions

(b) Median age of females in different regions

Figure 2: Differences between median age of sexes and regions

Figure 3 (a) and (b) represents the Infant mortality rate of males and females in different regions. From analysis as shown in Figure 3, it is found that in average around 46 per 1000 of male infant dies before reaching age of one years and around 37 per 1000 of female infant dies before reaching age of one years in Africa. And, Europe is found to have less infant mortality rate with the mean rate of 5.25 per 1000 for male infant and 4.34 per 1000 for female infant. Similarly, Asia, America and Oceania has male infant mortality rate 18.52, 13.69 and 16.08 per 1000 and has female infant mortality rate 14.94, 10.51 and 12.01 per 1000. It is also found that female are likely to survive as infant more than male.



(a) IMR of males in different regions

(b) IMR of females in different regions

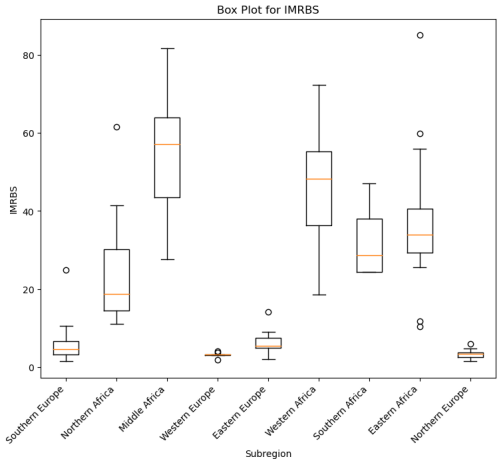
Figure 3: Differences between Infant mortality rate of sexes and regions

5.3 Variability of Variables within sub regions

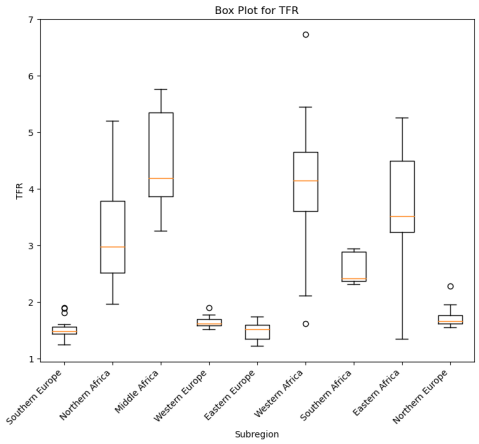
Variability in statistics pertains to how data points within a dataset differ from each other or from the mean. The variability in this case study is analysed by IQR for checking homogeneity and central tendency for checking heterogeneity.

The figure 4 shows the variability of 3 different variables i.e., '*Infant Mortality Rate of Both Sexes*','*Total Fertility Rate*' and '*Median Age of Both Sexes*' with different subregions of Africa and Europe. While analyzing figure 4 (a), it is found that infant mortality rate of both sexes in Western Europe has homogeneity because the IQR value is found to be 0.10. Similarly, while analyzing figure 4 (b), it is found that Southern Europe, Western Europe, Eastern Europe and Northern Europe has the homogeneous variability in variable '*Total Fertility Rate*' with the IQR value 0.13,0.10,0.25,0.14 respectively whereas figure 4 (c) showed that there is not homogeneity in variable '*Median Age of both sexes*'.

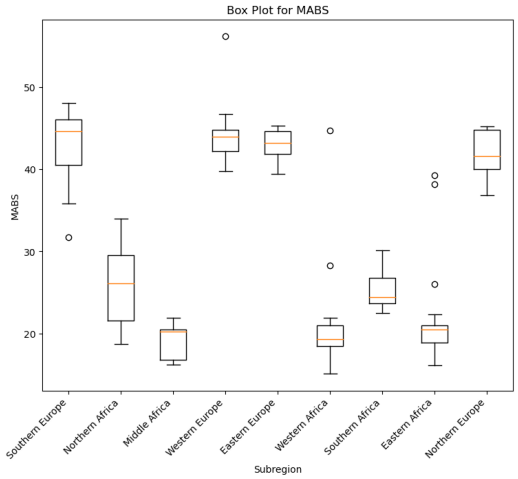
From the figure 4, while analyzing heterogeneity it is found that the sub regions of region Africa has significance differences in individual variables hence, individual variables are found to be heterogeneous between all the subregions of Africa. And, from figure 4 (c), the variable '*Median Age of both sexes*' has heterogeneous property even in the subregions of Europe.



(a) Infant Mortality rate of both sexes



(b) Total Fertility Rate



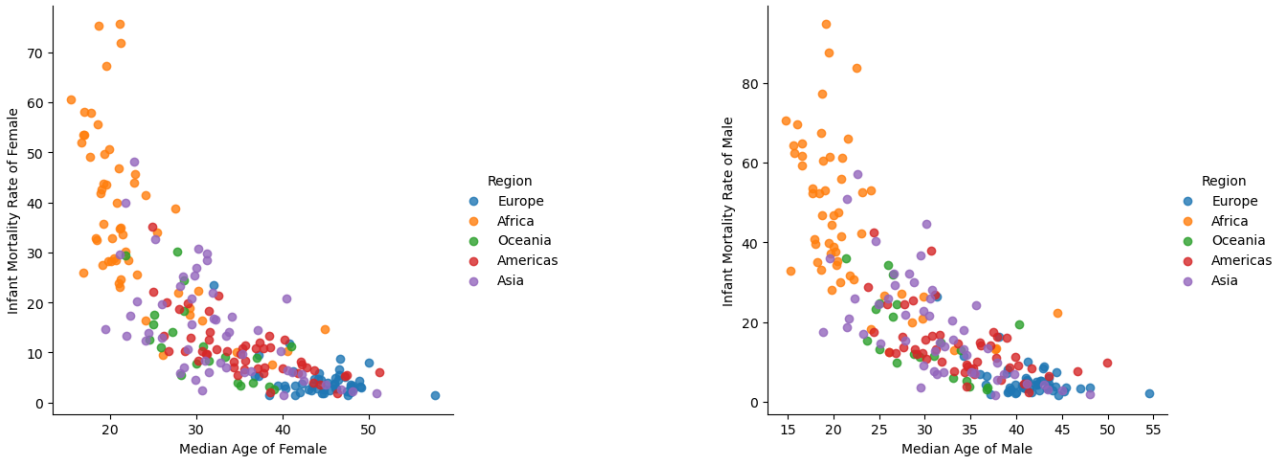
(c) Median Age of both sexes

Figure 4: Variability of Variables

5.4 Bi-variate Correlation

Bi-variate correlation represents the relationship between variables. It shows how the data are correlated with each other. The variables '*Median Age of Both Sexes*','*Median Age of Males*','*Median Age of Females*','*Infant Mortality Rate of Both Sexes*','*Infant Mortality Rate of Males*' and '*Infant Mortality Rate of Females*' are used for performing correlation.

While performing correlation, it is found that the relation between individual median age and infant mortality rate has some how negative correlation with the value around -0.79. From figure 5 (a), it is found that the '*Infant Mortality rate of female*' has strong negative correlation with the variable '*Median Age of Female*' with value -0.78.Similarly,



(a) Median Age Vs Infant Mortality rate of female

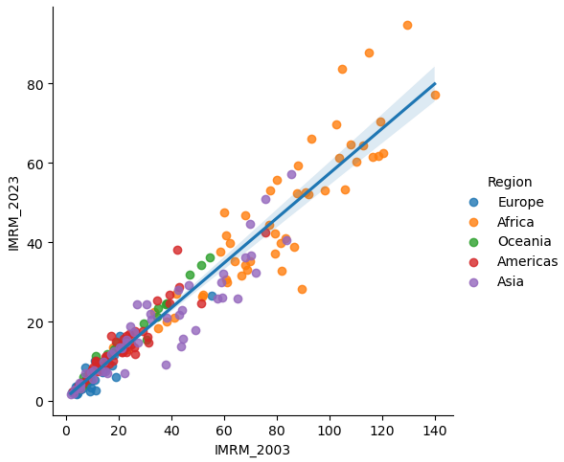
(b) Median Age vs Infant Mortality rate of male

Figure 5: Median Age and Infant Mortality Rate

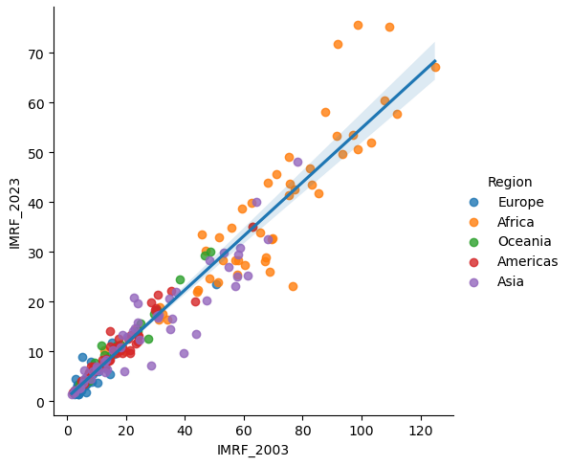
from figure 5 (b), the relation between '*Median Age of male*' and'*Infant Mortality rate of male*' is found to have a strong negative correlation with value of -0.79. The relation between '*Median Age of Both Sexes*' and '*Infant Mortality Rate of Both Sexes*' is found to have negative correlation with the value -0.7964. From analysis it is also found that the variable '*Median Age of Both Sexes*' is indeed depends upon the variables '*Median Age of Males*' and '*Median Age of Females*'. And similarly, the variable '*Infant Mortality Rate of Both Sexes*' is totally depends upon the variable '*Infant Mortality Rate of Males*' and '*Infant Mortality Rate of Females*'. This analysis provided us the fact about how the value of variable '*Median Age*' increases when value of '*Infant Mortality rate*' decreases.

5.5 Changes in variable over 20 years

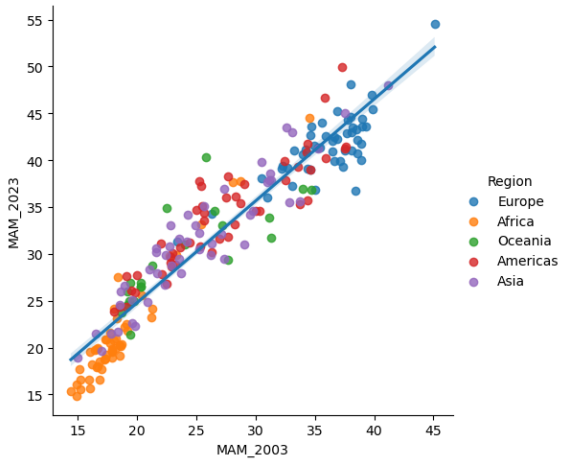
For finding the changes in variables, the given dataset was divided according to the variable '*Year*'. Later, scatter plot was drawn to see the change of variables between the year 2003 and 2023.



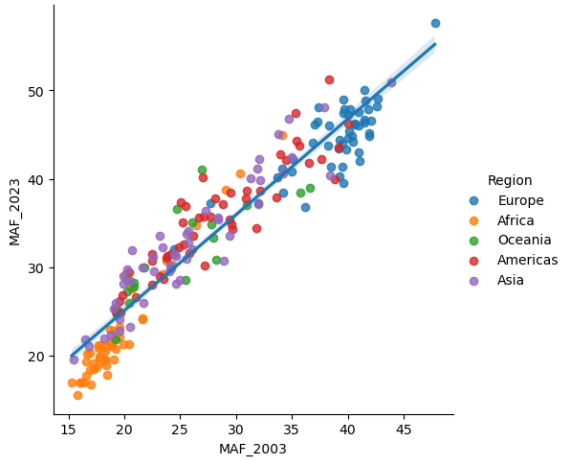
(a) IMR of male 2003 vs 2023



(b) IMR of female 2003 vs 2023



(c) Median Age of male 2003 vs 2023



(d) Median Age of female 2003 vs 2023

Figure 6: Median Age and Infant Mortality Rate

From the figure 5 (a) and (b), the variables '*Infant Mortality Rate of Female*' and the variable '*Infant Mortality rate of Male*' is found to be decreasing over 20 years in all of the regions. Similarly, from figure 5 (c) and (d), the median age of Africa is also found to be increasing in year 2023 as compared to 2003.

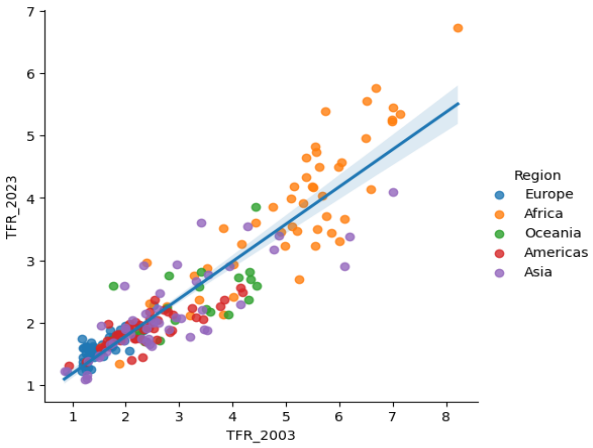


Figure 7: Total Fertility Rate 2003 vs 2023

Analyzing figure 7, the variable '*Total Fertility Rate*' is found to have decreasing in general from year 2003 to 2023. However, the variable '*Total Fertility Rate*' is found increasing in Africa in year 2023 as compared to 2003.

6 Summary

This case study was based on the data extracted from International Database of U.S. Census Bureau which contains categorical variables like '*Country*', '*Region*', '*Subregion*' and '*Years*' and numerical variables like '*Median Age of Both Sexes*', '*Median Age of Males*', '*Median Age of Females*','*Total Fertility Rate*', '*Infant Mortality Rate of Both Sexes*','*Infant Mortality Rate of Males*' and '*Infant Mortality Rate of Females*'. The frequency distribution of variables were discussed using the measure of central tendency and was visualized using histogram. Similarly, we have performed bi-variate correlation to show the relation between the variables using Pearson's Correlation coefficient method. The relationship was then visualized using pair plot. After onward, we have evaluated the variables variability within and between the sub region using Inter-Quartile Range and was visualized using boxplot. To end, we have evaluated the changes of variables in between the year 2003 and 2023 and is presented through line chart.

From frequency distribution, median age of both sexes between 30-40 was found in more than 100 countries, median age of males between 30-40 was also found in 100 countries and average median age of females was found higher than that of average median age of males. From bi-variate correlation, it was found that the relation between infant mortality rate and median age has negative correlation with the value of -0.79. While studying the variability of variables in year 2023 for subregions of Africa and Europe, it was found that the variable '*Infant Mortality Rate of Both Sexes*' had highest degree of variability in Middle Africa with variance 18.81 and lowest in Western Europe with variance 0.60. Similarly, for the variable '*Total Feritlity Rate*', it was found that the variability has highest degree in Western Africa with variance 1.20 and lowest in Southern Europe with variance 0.19. The variability of variable '*Median Age of Both Sexes*' was found highest in Western Africa with variance 6.70 and lowest in Middle Africa with variance 2.25.

While analyzing the changes of variable in over 20 years, it was found that the variable median age has been increased over the period of 20 years whereas the variable infant mortality rate has been decreased over this period. Similarly, the variable '*Total Fertility Rate*' was detected to had slight reduction. This result showed that, the scenario of 2023 is better than that of the scenario of 2003. In future,the next step after descriptive analysis can be Statistical analysis performing some hypothesis test on how accurate the descriptive analysis have been found.

Bibliography

K. Academy. Scatterplot and correlation review. https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-scatterplots/a/scatterplots-and-correlation-review, Aug. 2017.

S. Aditya. Histogram in matplotlib. https://www.datacamp.com/tutorial/histograms-matplotlib, June 2019.

P. Bhandari. Interquartile range :understand, calculate and visualize iqr. https://www.scribbr.com/statistics/interquartile-range/, May 2022.

U. S. C. Bureau. Demographic research. https://www.census.gov/programs-surveys/international-programs/about/dem-soc-analysis.html, Jan. 2022.

M. Galarnyk. Understanding boxplots. https://builtin.com/data-science/boxplot, Aug. 2022.

S. Glen. Correlation coefficient: Simple definition, formula, easy steps. https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/, Jan. 2022.

J. Hartmann, K.and Krois and B. Waske. Measure of central tendency. https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Descriptive-Statistics/Measures-of-Central-Tendency/The-Mean/index.html, Jan. 2018.

P. State. Measure of central tendency. https://online.stat.psu.edu/stat500/lesson/1/1.5/1.5.1, Jan. 2022.

WallstreetMojo. Variance. https://www.wallstreetmojo.com/variance/#h-formula, 2021.