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: Sadat Mahmud

Group number: 2

Group members: Sadat Mahmud, Ahmet Emircan Coskun, Montasir Hasan Chowdhury, Manouchehr Norouzi

November 17, 2023

Contents

1 Introduction

Demographic data generally carry information of the population for a specific geographical area. Here geographic area could be limited to a town or several regions of the planet Earth. In demographic data, the population can be separated by different factors such as gender, age, religious views, and so on. It generally consists of different types of social and economic factors such as the percentage of education, mortality rate, birth rate, income, etc. The general purpose of analyzing the demographic data is to make future predictions and obtain historic developments about the targeted group of the population.

This data set contains information about the years 2003 and 2023. The project aims to use proper graphical visualization, use proper statistical measures, and statistical analysis of the given demographic data.

Different types of visualization tools are used in this case study. For instance, box charts, heat maps, and scatter plots are used appropriately based on the problems. The main purpose behind choosing the visualization approach is to get a clear visual presentation of the changes in trends and changes in the value of variables between the years 2003 and 2023. Also, how it changes based on the different regions, subregions, and countries.

In the problem statement, data quality, and data type are discussed. Also, various properties of the dataset are discussed. In short, a complete overview of the dataset is discussed in section 2.

Several statistical methods such as mean, median, and standard deviation are used in this project to solve the given tasks. The properties of statistical methods and relevant preconditions to use these methods in the specific problem are discussed.

The interpretation of the results based on the given data set with proper statistical methods and visualization are discussed in section 4.

The last section contains the summary of the whole work. Also, short future predictions based on the whole project's work.

2 Problem statement

2.1 Description of Data set

The dataset is collected from The International Database (IDB)(U.S Census Bureau (2023)) of the United States of America Census Bureau. The dataset census2003_2023. csv is a brief excerpt from the IDB which contains the data for total fertility rates, Infant mortality rates, and Median ages of the population of 227 countries for the years 2003 and 2023. It has a total of 453 observations distributed in 11 variables. The variables are Country.Name, Region, Subregion, Year, Median.age..both.sexes, Median.age..females, Median.age..males, Total.Fertility.Rate, Infant.Mortality.Rate..Both.Sexes, Infant. Mortality.Rate.. Males, Infant. Mortality.Rate..Females. It has 7 missing values in Median.age..both.sexes, 7 missing values for Median.age..females, Median.age..males have 7 missing values, Infant. Mortality. Rate.. Both.Sexes have 6 missing values, Total.Fertility.Rate have 6 missing values, Infant.Mortality.Rate..Males have 6 missing values, Infant.Mortality.Rate.. Females have 6 missing values. The total missing value for the given dataset is 45.In the entire project for each task, the missing value is dropped as it does not have any significant differences. This dataset contains both categorical and numerical data. This dataset contains data from 227 countries, 21 subregions, and 5 regions. According to the Census Bureau, the definition of Fertility rate is "A fertility rate is typically expressed as the number of births per 1,000 women" (U.S. Census Bureau,2023b). The United States Census Bureau also defines as the "The infant mortality rate (IMR), also denoted as $1q_0$, represents the number of deaths of infants under 1 year of age per 1,000 live births. It quantifies the probability of dying between birth and reaching the exact age of 1." (U.S. Census Bureau,2023a).
The IDB is worldwide very popular for its high-quality data. They collect data through surveys, censuses, administrative records, and vital statistics (U.S Census Bureau (2023)). Hence, this dataset is considered to be of decent-quality data. Consequently, all variables provided in the dataset are utilized for analysis.

2.2 Project objective

Descriptive analysis is used to analyze this dataset. Descriptive analysis is a method to recapitulate a dataset, providing patterns to understand the problem or generate clear

insights about the variables by using its subsets. Four specific tasks are completed using several statistical methodologies. Frequency distribution is investigated, considering variations based on gender, regions, and subregions. A comparison is made between two regions, Europe and Africa, to demonstrate the differences in individual variables within the subregion. Additionally, the homogeneity and heterogeneity of values are analyzed between these two continents. The potential of bivariate correlations between the variables are explored. Lastly, the change of values in the variables over the last 20 years is compared using statistical methods and visualized with the graphical representations. Mean, Median, and standard deviation are used to achieve this goal. The Pearson correlation method is applied to compare correlation values and scatter plots and heatmaps are used to visualize the correlation. Several plots such as Barplots, scatterplots, and heatmaps are used for visualization. .

3 Statistical methods

In this part, statistical measures are discussed and those are used in this project. Different statistical methods are applied for solving the given problems. For statistical analysis and visualization, python version 3.11.3 is used(Python Software Foundation (2023)). Several packages such as pandas (pandas development team (2023)), Numpy (NumPy Community (2023)), matplotlib (Matplotlib Development Team (2023)), and seaborn (Seaborn Development Team (2023)) are used.

3.1 Statistical Measures

Mean

According to Jim Frost "In arithmetic and statistics, a mean is a single number that represents the middle point value of a dataset. It is also known as the average." (Jim Frost (2021)).
If the numeric observations are $x_1, x_2, x_3, ...x_n$ then the arithmetic mean can be formulated as:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

Where $\bar{x}$ represents the mean of total observation. n represents the total number of observations. Authors Hay-Jahans and Christopher state in their book "R Companion to Elementary Applied Statistics" that "The mean is distinctive for every sample (or population), and it must not necessarily hold true for the dataset for which it is computed. This measure is influenced by extreme values on one end of the data distribution and is best suited for homogeneous data or, at the very least, exhibit symmetry"(Hay-Jahans,p. 73-75).

Median

The median is the central value within a set of observed values, whether arranged in ascending or descending order. In terms of the odd number of observations, the middle value is considered as the median, and in the case of an even number of observations, the middle pair is determined by identifying the middle pair and taking their averages. The formula of the median is according to authors Hay-Jahans and Christopher (Hay-Jahans,p. 76)

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

Here $n+1$ = odd number, n is number of observations,x_n represents the n-th observation in the dataset, and $x_1, x_2, ..x_n$ represents the value of the variables X with n observations.

Variance and Standard Deviation

Measures of variability are known as Variance. It provides the idea about how data is from the mean value. High variance means the greater spread of data and low variance means the low spread of data. The formula of the variance is given below (Hay-Jahans, p. 76-77)

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

Here $\bar{x}$ denote mean value, total observation is n , x_i represents the i^th observation of

the variables X where i =1, 2, ...n.
Standard deviation is the square root of the variance. The formula is (Hay-Jahans, p. 76-77)

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

Where s generally represents the simple standard deviation and s^2 represents the variance.

Pearson Correlation Coefficient

The linear dependencies between two continuous variables are measured by the Pearson correlation coefficient. The strength and direction of the linear relationship between two variables are calculated through the Pearson correlation coefficient. The value of the Pearson correlation coefficient generally varies from +1 to -1. If the value is near 1 or 1 that means the variables are positively correlated and if the values are close to -1 or -1 then it indicates that the variables are negatively correlated. The formula of the Pearson correlation coefficient is (Hay-Jahans, p.321-322).

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

Here r is the correlation coefficient, the total observation number is denoted by n, x_i, and y_i is the i^{th} observations of the variables X and Y.

3.2 Plots for Data Visualization

Bar Charts

According to Hay-Jahans and Christopher in their book "R Companion to Elementary Applied Statistics," the authors state that " A bar chart typically visualizes frequencies corresponding to factor levels, which are represented by lengths of bars having equal

width". The bar charts can be oriented both horizontally and vertically. Although, a vertical orientation is more common. In vertically oriented bar charts, the horizontal axis indicates factor levels, and the vertical axis measures the bar lengths in terms of frequencies (Hay-Jahans, p. 111-112).

Scatter plot

A scatter plot is a graph where two axes represent values from two variables. To visualize the relationship between the two variables scatter plot uses dots. It generally shows the linear or curved scatter plot based on the relationship between two variables. In the scatter plot, no pattern means there is no relationship between the two variables. A scatter plot is ideal to visualize a correlation or predictive relationship. (Hay-Jahans, p.159-169)

Heat Map

Dr. Argenis Leon and Luis Aguirre state that " A heat map is a type of graphical presentation in Cartesian space that visualizes information about two variables. It measures the magnitude of a phenomenon in two dimensions with a color variation"(Leon and Aguirre (Nov,2021)). Here, Cartesian space states a two - dimensional coordinate system where data points are represented by pair of values along the x and y axes.

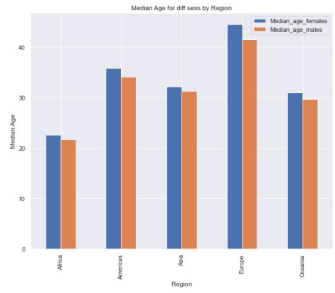
4 Statistical Analysis

4.1 Descriptive Analysis

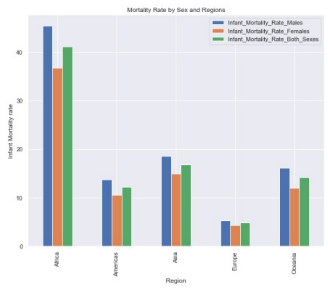
In this segment, all seven numerical variables are analyzed using data from the year 2023. The frequency distribution among the variables is examined. A detailed comparison between the regions of Europe and Africa is also investigated. Descriptive analysis is conducted to check the homogeneity and heterogeneity between different variables among regions and subregions. The bivariate correlation between the variables "Median Age" and Infant "Mortality Rate" is explored. Finally, A comparison between the years 2003 and 2023 is explored.

4.2 Frequency distribution of the variables

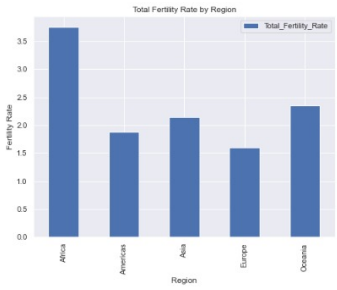
In this segment, The data will be analyzed using bar plots. To begin, the frequency distribution of the median age for the different sexes (Male and female) is analyzed using bar plots. Referring to Figure 1 and Table A on page 13 in the appendix section, the highest mean value for the median age of females, based on region, is 44.47 in Europe. The lowest number is found in Africa, measuring 22.62. For males, the highest mean value of the median age is 41.49, also in the European region. The lowest mean value for the males is in the African region, with a value of 21.70. Comparing these two value it can be said that female has a greater mean value for the median age compared to male. Now, the Infant mortality rate for males, females, and both sexes are analyzed and visualized using bar plots based on regions. In all the categories Africa has the largest mean value in terms of infant mortality rate, measuring 45.32 for males, 36.69 for females, and 41.07 for both sexes. The lowest value is in Europe in all the categories, measuring 5.27 for males, 4.35 for females, and 4.81 for both sexes. Now, the total fertility rate is visualized using a bar plot. Here, Africa has the highest mean value in terms of fertility rate, measuring 3.76 and the lowest mean value is in Europe, measuring 1.60.



(a) Median Age for different sexes by Region



(b) Infant Mortality rate by sexes and region



(c) Total Fertility rate by region

Figure 1: Frequency distribution for the Median Age, Infant Mortality, and Total Fertility.

4.3 Variability between two regions

The primary objective of this section is to conduct a thorough analysis of data about regions and subregions to explore the homogeneity and heterogeneity among the individual subregions. Table 1 presents the mean and standard deviation for the variables such as "Median Age of Both Sexes", "Total Fertility Rate", and "Infant Mortality Rate

of Both Sexes" for the regions and subregions. Specifically, Africa and Europe have been selected as the focus regions regions. In the region of Africa, It includes 5 subregions and in the region of Europe, It includes 4 subregions. When examining the variable "Median age of both sexes" it becomes apparent that European subregions exhibit a higher degree of homogeneity compared to African subregions. Notably, The most homogeneous variable for both regions is the "Total Fertility Rate". Conversely, The most heterogeneous variable for both regions is the "Infant mortality rate for both sexes". Finally, A visible pattern emerges, indicating that European subregions demonstrate greater homogeneity and African Subregions exhibit a higher level of heterogeneity.

Table 1: Details Differences between the regions Africa and Europe

Region	Subregion	Median Age		Total Fertility Rate		Infant Mortality Rate	
		Mean	Std	Mean	Std	Mean	Std
Africa	Eastern Africa	22.24	6.60	3.60	1.16	37.29	1.16
	Middle Africa	19.23	2.25	4.43	0.92	54.43	18.81
	Northern Africa	25.87	5.59	3.25	1.19	25.77	18.79
	Southern Africa	25.50	3.01	2.58	0.30	32.52	9.84
	Western Africa	21.14	6.70	4.10	1.20	46.59	14.57
Europe	Eastern Europe	43.10	1.92	1.48	0.17	6.31	3.45
	Northern Europe	41.75	2.95	1.73	0.20	3.34	1.18
	Southern Europe	42.97	4.75	1.53	0.19	6.09	5.57
	Western Europe	44.68	4.75	1.65	0.12	3.14	0.60

4.4 Bivariate Correlation with Pearson Correlation

In this segment, the relationship between the two specific variables, "Median Age" and "Infant Mortality Rate" will be explored. The bivariate Correlation using Pearson Correlation is investigated. The correlation is visualized using a scatter plot and heat map. Here the "Median age" and "Infant Mortality Rate" are negatively correlated with each other as indicated by the correlation value of value -0.80 in Figure 2.
Upon Comparing the scatter plots in Figure 3, it can be inferred the variable "Infant mortality rate Male" is the most negatively correlated with "Median age".

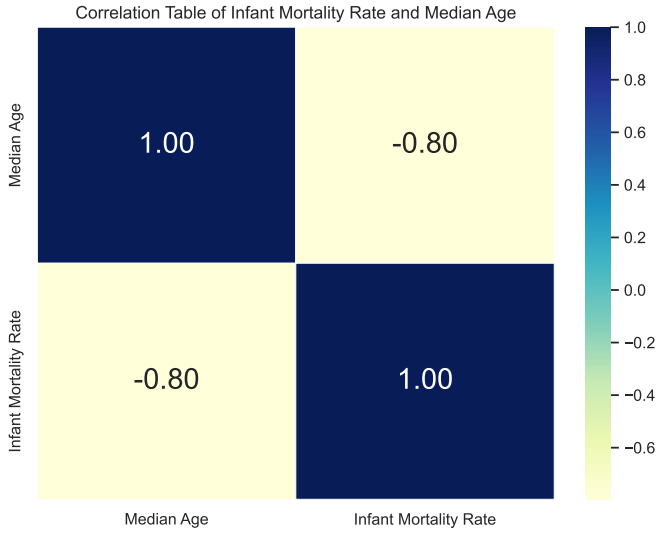
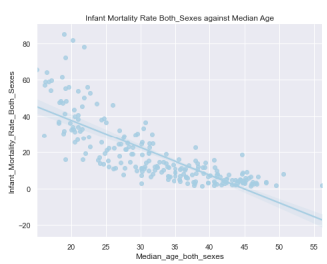
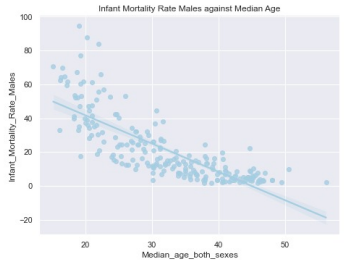


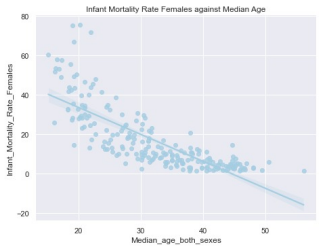
Figure 2: Correlation table of Infant Mortality Rate and Median Age.



(a) Infant Mortality both Sexes Vs Median Age



(b) Infant Mortality Rate of Male Vs Median Age



(c) Infant Mortality Rate of Female Vs Median Age

Figure 3: Scatter plot with different correlation between Median Age and Infant Mortality Rate

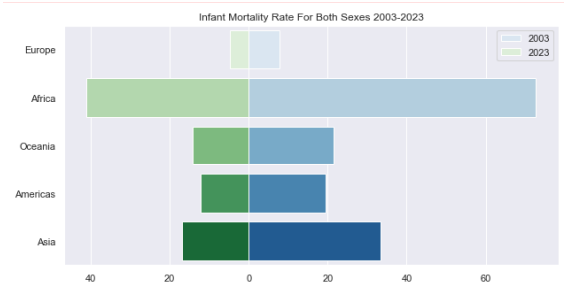
4.5 Change of Variables Over The Last 20 Years

The primary goal of this section is to analyze the variations in overall variables between the year 2003 and 2023. The correlation is visualized through two bar plots, each labeled with two distinct colors to signify the respective time frames. To effectively illustrate two differences across these time periods, horizontally oriented bar charts are employed. Furthermore, Table 2 provides the mean values of the variables for the years 2003 and 2023..

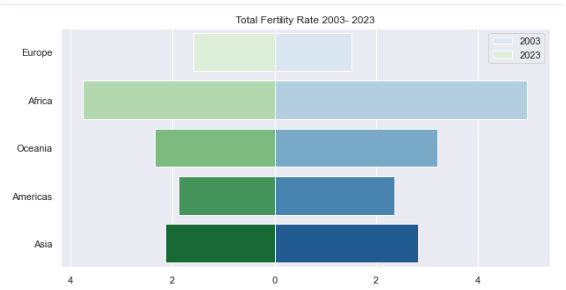
Table 2: For tables, the caption belongs above the table.

Region	Year	Median Age Both Sexes	Total Fertility Rate	Infant Mortality Both Sexes
Africa	2003	19.08	4.97	78.82
Africa	2023	22.17	3.76	45.33
America	2003	27.78	2.36	21.44
America	2023	34.95	1.88	13.70
Asia	2003	25.35	2.83	36.00
Asia	2023	31.83	2.15	18.53
Europe	2003	37.46	1.53	8.66
Europe	2023	42.97	1.60	5.26
Oceania	2003	24.24	3.21	24.01
Oceania	2023	30.31	2.35	16.08

In 2003, Africa exhibited an average "Median Age of Both sexes" was 19.08, Which has increased to 22.17 in 2023. During the same period, the average total fertility rate decreased from 4.97 to 3.76. The average infant mortality rate for both sexes in 2003 was 78.82, and this figure declined to 45.33 in 2023.
In America, both The median age of both sexes and the Infant mortality rate demonstrated significant improvement from 2003 to 2023.
Asia witnessed notable progress in the realm of infant mortality rates for both sexes between 2003 and 2023.



(a) Infant Mortality Both Sexes 2003-2023



(b) Total Fertility rate 2003-2023

Figure 4: Significant Changes of values over past 2003-2023

This analysis reflects key demographic changes over the specified timeframe and provides valuable insights for a comprehensive report.

5 Summary

This report is conducted based on a descriptive analysis utilizing a dataset obtained from the International Data Base (IDB) of the U.S. Census Bureau (U.S Census Bureau (2023)). The dataset comprises data from 227 countries, 21 subregions, and 5 regions, encompassing a total of 11 variables and 353 observations. The investigation focuses on the frequency distributions of median age for different sexes. The European region exhibits the highest mean value, while the African region has the lowest. Notably, a discernible difference is observed between the median age for females and males.
The analysis extends to exploring homogeneity and heterogeneity across regions and subregions, revealing that the European region displays greater homogeneity compared to Africa. Bivariate correlation is examined for the variables "Median Age" and "Infant Mortality Rate" using Pearson correlation. Additionally, changes in variables over the last two decades are scrutinized. America and Asia achieved notable success in controlling infant mortality rates between 2003 and 2023.

While this dataset includes seven variables, it's acknowledged that other factors, such as air and weather quality, also influence demography. Nevertheless, this descriptive analysis of demographic data provides a comprehensive graphical overview.

Bibliography

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Chapman Hall/CRC. ISBN 9781138329164.

Jim Frost. Statistics by jim, 2021. URL https://statisticsbyjim.com/basics/mean$_a verage$/.

Dr. Argenis Leon and Luis Aguirre. *Data Processing with Optimus*. Packt Publishing, September Nov,2021. ISBN 9781801077750.

Matplotlib Development Team. *Matplotlib Documentation*. Matplotlib Development Team, 2023. URL https://matplotlib.org/stable/contents.html.

NumPy Community. *NumPy Documentation*. NumPy Community, 2023. URL https://numpy.org/doc/stable/.

pandas development team. *pandas Documentation*. pandas development team, 2023. URL https://pandas.pydata.org/pandas-docs/stable/.

Python Software Foundation. *Python Language Reference*. Python Software Foundation, 3.11.3 edition, Apr 2023. URL https://docs.python.org/3/reference/. Version 3.11.3 (tags/v3.11.3:f3909b8, Apr 4 2023, 23:49:59) [MSC v.1934 64 bit (AMD64)].

Seaborn Development Team. *Seaborn Documentation*. Seaborn Development Team, 2023. URL https://seaborn.pydata.org/.

U.S. Census Bureau, 2023a. URL https://www.census.gov/glossary/term$_I nfantmortalityrate?term = Infant + mortality + rate$.

U.S. Census Bureau, 2023b. URL https://www.census.gov/topics/health/fertility/about.html.

U.S Census Bureau. International data base, 2023. URL https://www.census.gov/programs surveys/international-programs/about/idb .html.

U.S. Census Bureau, 2023. URL https://www.census.gov/topics/health/fertility/about.html.

Appendix

A Additional Table

Table 3: Mean of the variables according to region

	Region	Medion age females	Medion age males	Infant Mortality Rate Males	Infant Mortality Rate Females	Infant Mortality Rate Both Sexes	Total Fertility Rate
0	Africa	2.62	21.70	45.32	36.68	41.07	3.76
1	Americas	35.82	34.07	13.53	10.40	12.00	1.88
2	Asia	32.18	31.24	18.53	14.95	16.78	2.14
3	Europe	44.47	41.49	5.25	4.35	4.81	1.60
4	Oceania	30.98	29.71	16.09	12.01	14.10	2.34