TU Dortmund

Introductory Case Studies

Project 1: Descriptive data analysis

Lecturers:

Dr. Crystal Wiedner

Dr. Marlies Hafer

Dr. Rouven Michels

Author: Mohaiminul Islam

Group number: 5

Group members: Md Zihad Hossain, , Mosfiqun Nahid Hassan, Ilham Pambudi

November 3, 2024

Contents

1 Introduction

Life expectancy and infant mortality rates are critical indicators of public health and quality of life. These metrics reflect the overall health of populations and reveal disparities driven by demographic, socio-economic, and geographical factors. In this report, we examine variations in life expectancy and under-5 children children mortality rates influenced by factors such as gender, year and regional difference, aiming to identify and analyze these differences using both statistical measures and visualization techniques. To address these objectives, we employ descriptive statistics such as the mean, median, inter-quartile range (IQR) to conduct a detailed numerical analysis. Additionally, visual tools such as box plots and histograms are utilized to illustrate the distribution and variability of life expectancy and under-5 children mortality rates in terms of gender and different regions. This combined approach of statistical analysis and visual representation offers a comprehensive understanding to assess the variability associated with these influential factors .

In **Section 2**, we provide an overview of the dataset used in this study and evaluate its quality. The specific objectives of the project are also outlined. **Section 3** details the statistical methods employed for the analysis, including their rationale and application. In **Section 4**, these statistical techniques are applied to the dataset, and the results are interpreted in the context of regional and gender-based disparities. Finally, **Section 5** presents a summary of the findings, a discussion of the implications, and suggestions for potential future research.

2 Problem statement

2.1 Data description and data quality

The dataset analyzed in this report was provided by the instructors of the "Case Studies I" course at TU Dortmund University during the winter term 2023/24. It represents a subset from the U.S. Census Bureau's International Data Base (IDB), which compiles demographic data for internationally recognized states and regions with populations exceeding 5,000, spanning the years 1950 to 2100.

This dataset includes a sample size of 455 observations (n = 455) across five major regions—Asia, Europe, Africa, Oceania, and the Americas—each further divided

into sub-regions .The primary variables in the dataset include *Name* (object), *Year* (int64), *Life.Expectancy.at.Birth..Both.Sexes* (float64), *Life.Expectancy.at.Birth..Males* (float64), *Life.Expectancy.at.Birth..Females* (float64), *Under.Age.5.Mortality..Both.Sexes* (float64), *Under.Age.5.Mortality..Males* (float64), *Under.Age.5.Mortality..Females* (float64), *Subregion* (object), and *Region* (object). For clarity, the *Life.Expectancy.at.Birth..* variables were renamed to *Life Expectancy (Both)*, *Life Expectancy (Male)*, and *Life Expectancy (Female)*, while *Under.Age.5.Mortality..* variables were renamed to *Mortality Under 5 (Both)*, *Mortality Under 5 (Male)*, and *Mortality Under 5 (Female)*. This renaming is consistently applied throughout the report.In this context, *Life Expectancy* refers to the average expected lifespan at birth, assuming constant mortality rates over time, while *Mortality Under 5* denotes the number of deaths of children under age five per 1,000 live births. These indicators are critical for understanding disparities in public health by region and gender.

The dataset initially contained several missing values across both numerical and categorical variables. Specifically, there were five missing entries for each of the *Life Expectancy* and *Mortality Under 5* variables (recorded for both sex, male and female), and four missing entries for the *Region* and *Subregion* variables.Altogether, these missing values were distributed across nine rows, each containing at least one missing entry. For consistency in analysis, these rows with incomplete data were excluded from the dataset, reducing the original count from 455 to 446 complete records .

2.2 project goals

The primary objective of this report is to identify and analyze variations in life expectancy and under-5 children mortality, with a specific emphasis on gender disparities and socio-economic regions. This report examines the linear dependency between variables, as well as compares changes in life expectancy and mortality rates across two distinct time frames: 2004 and 2024.

3 Statistical methods

3.1 Mean

The mean represents the average of the data and is commonly used as an indicator of a "Central" value within the dataset. It is calculated by summing all measurements in the dataset and then dividing by the total number of measurements. As it summarizes all the observations, with the presence of extreme values, the mean can be distorted, making it an unreliable measure of central value. For a set of n quantitative observations $x_1, x_2, \ldots, x_n$, the formula for the mean $\bar{x}$ is:

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

((3)McClave et al.,2018, p. 81-82)

3.2 Median

The median represents the middle value of a dataset when the quantitative observations are arranged in ascending or descending order. Unlike the mean, the median provides a better measure of central tendency in the presence of outliers (extremely large or small values). For n quantitative observations:

- If n is odd, the median is the middle observation in the ordered list.

- If n is even, the median is the average of the two middle observations. ((3)McClave et al., 2018, p. 84)

3.3 Standard deviation

Standard deviation is a measure of the amount of variation or dispersion in a set of values. It quantifies how much individual data points differ from the mean of the dataset. A low standard deviation indicates that the values tend to be close to the mean, while a high standard deviation indicates a wider spread of values.((3) McClave et al.,2018, p.93-94)

The formula for calculating the standard deviation s for a sample is:

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

3.4 Inter quartile Range (IQR)

The inter-quartile range (IQR) is a measure of statistical dispersion that indicates the range within which the central 50% of the data points lie. It is calculated by subtracting the first quartile Q_1 (the 25th percentile) from the third quartile Q_3 (the 75th percentile). The formula for the IQR is given by:

$$\text{IQR} = Q_3 - Q_1$$

To determine the quartiles, the dataset must be sorted in ascending order, and the median of the lower and upper halves is used to find Q_1 and Q_3, respectively. The IQR is particularly useful for identifying outliers and understanding the spread of the dataset while minimizing the influence of extreme values.((2)Illowsky B., & Dean S. ,2018 , p.87)

3.5 Correlation

The correlation coefficient quantifies the strength and direction of the relationship between two variables. Among various methods to measure correlation, Pearson's correlation coefficient is widely used, primarily due to its compatibility with linear relationships. The values of the correlation coefficient range from -1.0 to 1.0. A positive correlation indicates that as one variable increases, the other variable also tends to increase. Conversely, a negative correlation suggests that an increase in one variable is associated with a decrease in the other. A correlation coefficient of 0 signifies no linear relationship between the two variables.((2)Illowsky B., & Dean S. ,2018, p.631-633). The formula for Pearson's correlation coefficient r is:

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

where $\bar{x}$ and $\bar{y}$ represent the means of the respective variables.

3.6 Histogram

A histogram is composed of adjacent bars that collectively represent the distribution of data. The horizontal axis x represents the variable being measured or the range of data

values. It is divided into intervals, or bins, which indicate the groups or classes of data. Each bin corresponds to a specific range of values, and the width of the bin reflects the size of that range. The vertical axis y is labeled with either frequency, relative frequency, percent frequency, or probability. Regardless of the label used, the overall shape of the histogram remains the same. Histograms provide insights into the data distribution, including its shape, central tendency, and variability.((2)Illowsky B., & Dean S. 2018, p. 75). Figure 1 provides an illustration of the Histogram.

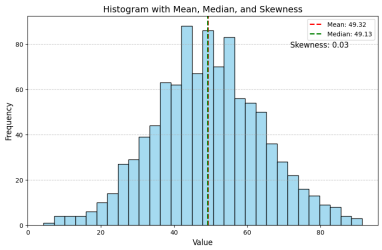


Figure 1: Histogram

3.7 Box Plot

Box plots, also known as box-and-whisker plots, effectively illustrate the concentration and distribution of data, highlighting the relationship of extreme values to the rest of the dataset. A box plot is constructed using five key values: the minimum, first quartile Q_1, median Q_2, third quartile Q_3, and maximum. The box plot is typically created with a horizontal or vertical number line along with a rectangular box. The endpoints of the axis represent the smallest and largest data values, while the box is defined by the first and third quartiles, encompassing approximately 50% of the data. Whiskers extend from the box to the minimum and maximum values. The median is positioned within the box, providing a clear visual representation of the data distribution.((2) Illowsky B., & Dean S. 2018, p. 93-94). Figure 2 provides an example of a box plot that displays all the relevant metrics.

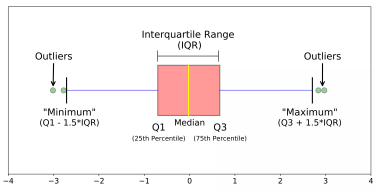


Figure 2: Box Plot

3.8 Scatter plot Matrix

A scatter plot matrix is a grid of scatter plots that visually assesses the relationships between multiple pairs of variables in a dataset. To construct a scatter plot matrix, first, select the quantitative variables to include, as these will effectively reveal relationships. For each pair of selected variables, create a scatter plot where one variable is represented on the x-axis and the other on the y-axis, with each point corresponding to an observation. Organize these scatter plots into a matrix format, with each cell displaying the scatter plot for a unique pair of variables; diagonal cells can be utilized to present histograms or density plots of individual variables for added insight. Finally, analyze the scatter plot matrix for visible patterns, correlations, or anomalies across the different variable pairs. This visualization is particularly beneficial for exploratory data analysis, providing a comprehensive overview of the relationships among multiple variables within a single display.((1) Fox J. ,2016, p.120-122).Figure 3 shows an example of a scatter plot matrix, featuring scatter plots in the lower triangle, histograms along the diagonal, and correlation coefficients in the upper triangle.

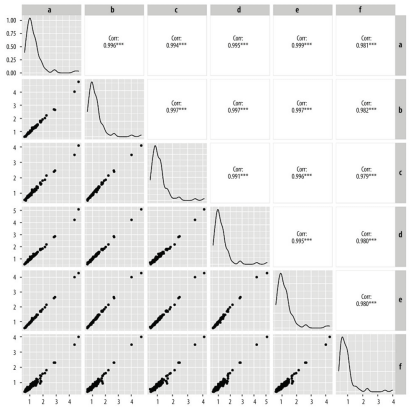


Figure 3: Scatterplot metrix

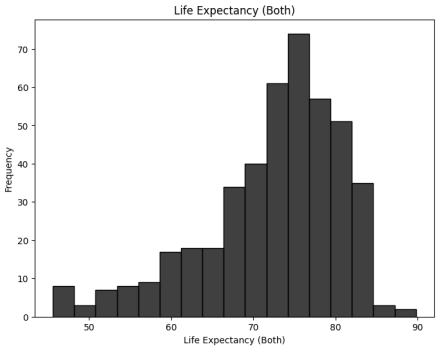
4 Statistical analysis

The statistical analysis was conducted using **Python** == 3.12.6 with VScode==1.94.2 as an IDE. The following Python packages Numpy==2.1.2, pandas==2.2.3, matplotlib==3.9.2, seaborn==0.13.2 are used for data manipulation, computation, and visualization((4)Python Software Foundation). The analysis is based on a dataset containing 446 records, after excluding 9 rows with missing values from the original 455 records.

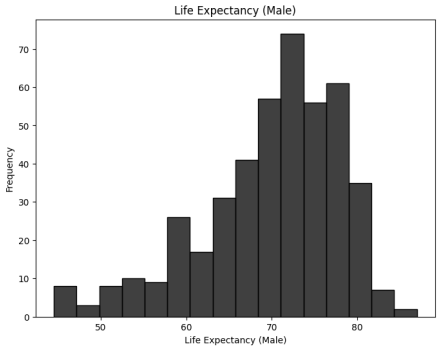
4.1 Frequency Distributions

The histogram in Figure 4 illustrates the life expectancy distribution .Each sub-figure represents a category of life expectancy distribution: 4(a) for both sexes, 4(b) for males, and 4(c) for females. Across all three histograms, the distributions are left-skewed, meaning that most values cluster around the higher end of the life expectancy range. For both sexes , the distribution centers around between 65 and 85 years, while for males it clusters around 65 to 80 years, and for females from 70 to 85 years. This range suggests that, on average, females have a higher life expectancy than males. According to Appendix Table 1(p.14), the mean life expectancy is 74.88 years for females, compared to 70 years for males and 72.38 years for both sexes combined.

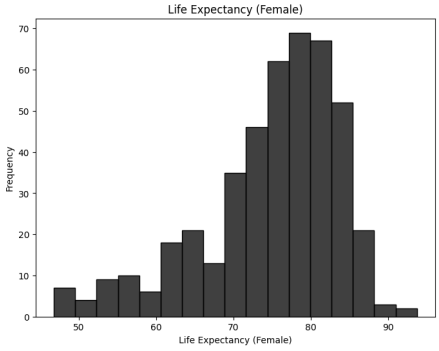
Examining variability, all three distributions exhibit similar spreads, but life expectancy for females shows slightly higher dispersion. The standard deviation (Sd) values for males, both sexes, and females, as presented in Appendix Table 1(p.14), are 8.19, 8.46, and 8.86, respectively. The marginally higher Sd for females suggest a greater degree of variability in female life expectancy.



(a) Life expectancy for both sex



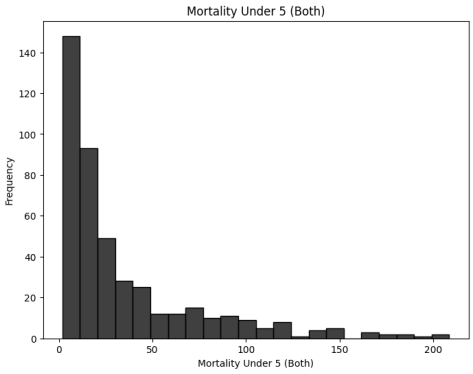
(b) Life expectancy for male



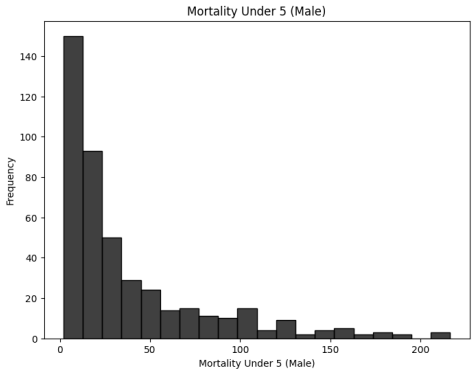
(c) Life expectancy for female

Figure 4: Histograms for comparing the distribution of life expectancy variables.

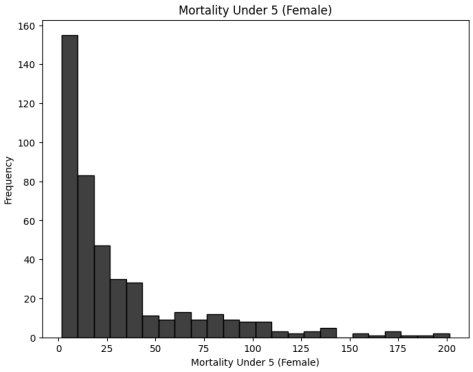
Figure 5 illustrates the distribution of mortality rates for individuals under five years of age. Each sub-figure represents a specific category of mortality: 5(a) for both sexes, 5(b) for males, and 5(c) for females. All histograms are right-skewed, indicating that the majority of countries in the dataset have lower mortality rates. From Figures 5(b) and 5(c), it is observed that the highly concentrated region for males is between 0 and 50, while for females, it is between 0 and 30, suggesting a higher average mortality rate for males compared to females. According to Appendix Table 1(p.14), the average mortality rate is 37.80 for males and 32.13 for females, which supports the observations made from the histograms. Regarding variability, the male distribution (sub-figure 5(b)) indicates a steady increase in mortality rates compared to that of females, indicating a higher variability in males compared to females. From Appendix Table 1(p.14), The standard deviation (Sd) for males is 42.34, while for females, it is 38.13 This higher Sd for males reinforces the notion of higher variability compared to female.



(a) Mortality under-5 for both sex



(b) Mortality under-5 for male

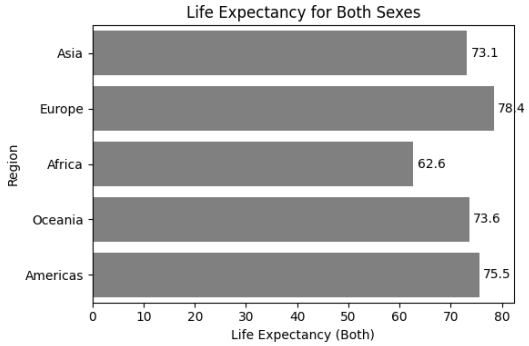


(c) Mortality under-5 for female

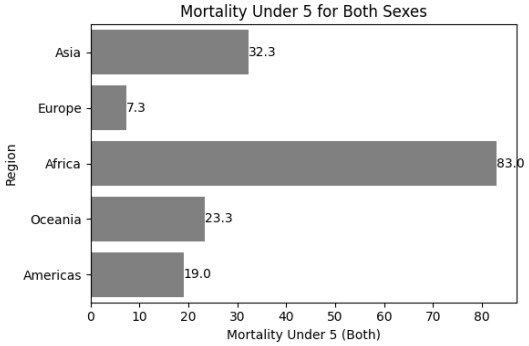
Figure 5: Histograms for comparing distributions across under 5 age mortality Variables

4.2 Comparison between regions

Figure 6 Life expectancy and mortality rates for children under age five vary significantly across regions. Figures 6(a)and 6(b) compare these metrics with respect to the mean values for both sexes across different regions. The average life expectancy is highest in Europe at 78.4 years, followed by the Americas at 75.5 years and Asia at 73.1 years, while Africa records the lowest average life expectancy at 62.6 years. In contrast, the average mortality rate under age five is highest in Africa at 83.0, whereas Europe has the lowest at 7.3. This contrast highlights disparities in health outcomes across regions.



(a) Average life expectancy for both sex



(b) Average under-5 children mortality rate for both sex

Figure 6: histogram of mean value for life expectancy and under-5 children mortality across regions

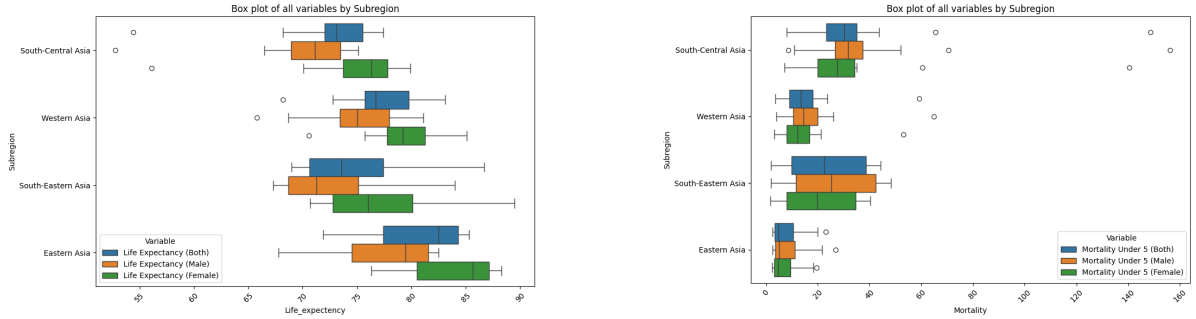
4.3 Homogeneity within sub regions

Figure 7 presents box plots that summarize the distribution of life expectancy and mortality rates of under-5 children across four different sub-regions in Asia, with **7(a)** illustrating life expectancy and **7(b)** depicting child mortality rates.

Regarding life expectancy, South Central Asia demonstrates a narrower inter-quartile range (IQR) of 3.475, 4.55, and 4.075 for both sexes, males, and females, respectively (Appendix Table 2,p.14). This indicates a relative homogeneity in life expectancy within these regions compared to others. Following South Central Asia, Western Asia exhibits the second lowest inter-quartile range, reflecting a similar trend of uniformity in life expectancy among its population. However, the median life expectancy values for South Central Asia with 73.5 for both sexes, 71.15 for males, and 76.30 for females(Appendix Table 3,p.14) are lower than those of Western Asia with 76.70 for both sexes, 75 for

males, and 79.20 for females . In contrast, Eastern Asia and South Eastern Asia display comparatively larger inter-quartile ranges, indicating increased variability in life expectancy values within these regions. Notably, Eastern Asia boasts the highest median life expectancies across all categories: 82.50 for both sexes, 79.45 for males, and 85.65 for females while South Eastern Asia presents much lower medians across all categories. These differences in IQRs and medians indicates the heterogeneity in life expectancy among these sub-regions(Appendix Table 3,p.14).

In terms of mortality rates, Eastern Asia shows the lowest IQR of 7.075,7.675,6.40 for both sexes, males and females respectively, followed closely by Western Asia(Appendix Table 2,p.14). This trend suggests a greater homogeneity in mortality rates within these regions. Conversely, South Eastern Asia displays the highest degree of variability in mortality rates, as observed by the length of box (IQR) . The median values across these sub-regions vary distinctly, with Eastern Asia reporting the lowest medians: 4.85, 5.05,4.5 (Appendix Table 3 ,p.14)for both sexes, males and females, respectively, and South Central Asia exhibiting the highest median mortality rates, with figures of 30.30 for both sexes, 31.65 for males, and 27.50 for females(Appendix Table 3,p.14). This contrast in median mortality rates along with the variation in inter quartile ranges, serves as an evidence of the heterogeneity in under 5 mortality exists among these sub-regions.



(a) Life expectancy box plot

(b) Mortality under -5 boxplot

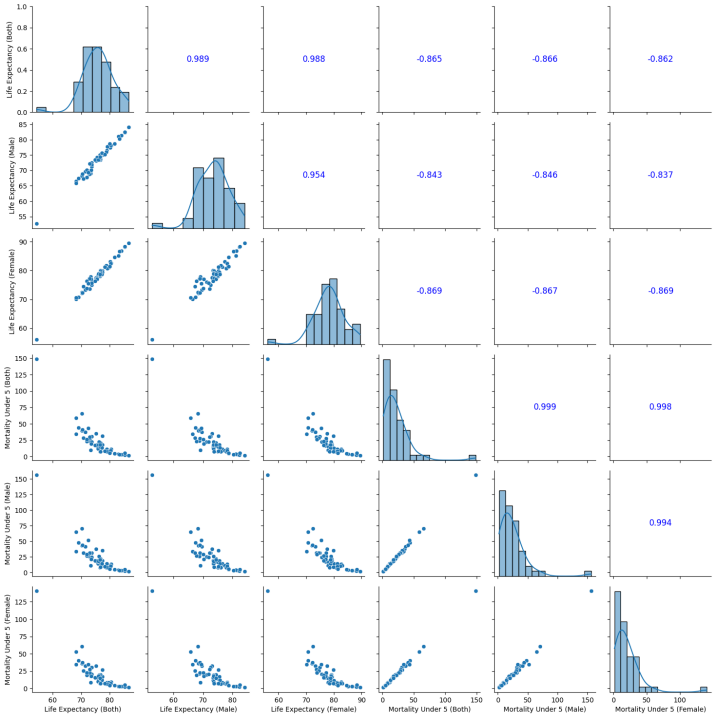
Figure 7: Box-plot for subregions of Asia

4.4 Correlation between variables

Figure 8 provides a scatter plot matrix illustrating the relationships between variable of interests with correlation coefficients along with scatter plot to quantify the associations between each pair. From Figure 8(a), it is evident that life expectancy for both

sexes shows a highly positive correlation with life expectancy for males and females, with correlation coefficients of 0.989 and 0.988, respectively. Additionally, the correlation between male and female life expectancy is 0.954. This strong positive correlation indicates that changes in any of these life expectancy variables will result in nearly uniform changes in the others. The same trend of strong positive correlation is observed in mortality rates, where the mortality rates for both males and females exhibit high correlation coefficients, suggesting that as the mortality rate for one sex increases, the mortality rate for the other sex tends to increase in a similar manner.

In contrast, the correlation between life expectancy (for both sexes) and under-5 children mortality rates (for both sexes) is highly negative, with a correlation coefficient of -0.865. This indicates that as life expectancy increases, under-5 children mortality rates decrease sharply. This negative relationship is consistent across pairs of life expectancy (males) and under-5 children mortality (males), as well as life expectancy (females) and under-5 children mortality (females).



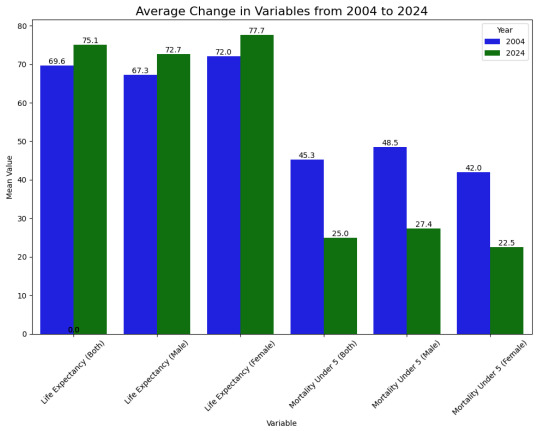
(a) correlation between variables

Figure 8: Scatter plot Matrix with Correlation Coefficients

4.5 Comparison between two time frame

Figure 9 compares the average changes in life expectancy and under-5 children mortality rates between 2004 and 2024. In 2024, the life expectancy for both sexes, as well as for males and females, has increased, with values of 75.1, 72.7, and 77.7, respectively. In contrast, the life expectancy figures for 2004 were 69.6 for both sexes, 67.3 for males, and 72.0 for females.

Conversely, the under-5 children mortality rates have shown a marked decrease, with rates in 2024 being 25.0 for both sexes, 27.4 for males, and 22.5 for females from the rates in 2004 with 45.3 for both sexes, 48.5 for males, and 42.0 for females.



(a) Difference between years of 2004 and 2024

Figure 9: Histograms for mean value in 2004 and 2024

5 Summary

In this project, data originated from the U.S. Census Bureau's International Data Base (IDB) on life expectancy and under-5 children mortality rates -encompassing regions, sub-regions, and countries, was analyzed to explore averages and variations influenced by gender, year and socio-economic areas. The analysis was conducted on 446 completed records after eliminating the 9 rows with missing values from the original dataset. Statistical methods and visualization techniques were employed to illustrate research goals, identify trends, and understand underlying patterns. The analysis revealed notable gender disparities with male having the lower average life expectancy(70) compared to female (74.88) along with having higher average under-5 children mortality rates for male(37.80) compared to female(32.13).Regional differences were also evident, with Europe having the highest average life expectancy at 78.4 years, and Africa reporting the highest under-5 children mortality rate at 83 deaths per 1,000 live births. Further analysis focused on subregions within Asia, where box plots showed varied inter quartile ranges for life expectancy and mortality rates across different subregions. South Central Asia exhibits the most homogenous data distribution with respect to IQR for life expectancy while for the under-5 children mortality rates Eastern Asia displays the lowest IQR. It was also observed that, each of these subregions having distinct median and IQR , are heterogenous among themselves. A key finding was the strong negative correlation of -0.866 between life expectancy(both sex) and under-5 children mortality rates(both sex), suggesting that the increase of one variable influences the rapid decrease of other variable. Overall, the data indicated improvements in health outcomes over time, with the increased average life expectancy and the decreased average under-5 children mortality in 2024 compared to 2004. Future research could expand on this analysis by investigating additional socio-economic and environmental factors, such as income levels, healthcare access, and educational attainment, to better understand their impact on life expectancy and under-5 children mortality rates

Bibliography

[1] Fox J. (2016): *Applied Regression Analysis and Generalized Linear Models*, 3rd edition, Sage Publications.

[2] Illowsky B., & Dean S. (2018): *Introductory Statistics*, 2nd edition, OpenStax.

[3] McClave J. T., Benson P. G., & Sincich T. (2018): *Statistics for Business and Economics*, 14th edition, Pearson.

[4] Python Software Foundation. (2024). *Python Language* (Version 3.12.6). https://www.python.org/

Appendix

A Additional tables

Table 1: Mean and Standard Deviation of Variables

Variable	**Mean ($\bar{x}$)**	**Std Dev (s)**
Life Expectancy (Both)	72.38	8.46
Life Expectancy (Male)	70.00	8.19
Life Expectancy (Female)	74.88	8.86
Mortality Under 5 (Both)	35.03	40.22
Mortality Under 5 (Male)	37.80	42.34
Mortality Under 5 (Female)	32.13	38.13

Table 2: Interquartile Range (IQR) Values by Subregion

Subregion	**Life Exp. (Both)**	**Life Exp. (Male)**	**Life Exp. (Female)**	**Mortality Under 5 (Both)**	**Mortality Under 5 (Male)**	**Mortality Under 5 (Female)**
Eastern Asia	6.900	7.00	6.625	7.075	7.675	6.400
South-Central Asia	3.475	4.55	4.075	11.750	10.725	14.175
South-Eastern Asia	6.750	6.40	7.300	28.600	30.750	26.500
Western Asia	4.050	4.50	3.500	9.100	9.400	8.900

Table 3: Median Values by Subregion

Subregion	**Life Exp. (Both)**	**Life Exp. (Male)**	**Life Exp. (Female)**	**Mortality Under 5 (Both)**	**Mortality Under 5 (Male)**	**Mortality Under 5 (Female)**
Eastern Asia	82.50	79.45	85.65	4.85	5.05	4.55
South-Central Asia	73.15	71.15	76.30	30.30	31.65	27.50
South-Eastern Asia	73.60	71.30	76.00	22.50	25.20	19.70
Western Asia	76.70	75.00	79.20	13.40	14.60	12.10