TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Prof. Dr. Katja Ickstadt
M. Sc. Zeyu Ding
M. Sc. Yassine Talleb

Author: Abdul Muqsit Farooqi

December 17, 2023

Contents

1 Introduction

In this modern era, the countries are concerned to invest more on living standard, their education, their health care facilities and other essential needs. To provide these or for such progress, requires a structure that brings growth to the society. The population indicators that helps to understand the statistics of population dynamics and characteristics are population size, under age 5 mortality rates and life expectancy at birth. Governments of all countries make balance between both population and economic growth.

The United States Census Bureau (USCB) is an agency of U.S. Federal Statistical System that provides World's Population data. It provides open access of the population data that can also be taken as an educational resource. Th dataset contains small sample of population, life expectancy at birth, and under age 5 mortality rates for 227 countries from 2002 and 2022. The objective of this project is to perform descriptive analysis of sample data and apply appropriate statistical analysis to understand the relationship between life expectancy and mortality. For the analysis, frequency distribution of the variable is determined, and the correlation between variables is calculated using Pearson correlation coefficient.

In Section 2, the dataset is explained briefly the quality of the dataset and the structure of descriptive analysis is discussed. In Section 3, statistical method and its notations is explained. In Section 4, graphical plots such as histograms, scatter plot and box plots are used to interpret the results. Lastly, all the results are summarized in Section 5.

2 Problem statement

2.1 Dataset and Data Quality

The Census Bureau that provides estimations and projections of population is the source of U.S. Government. International Database (IDB) that keeps the record including demographic measures for over 2200 countries and areas of the world having population of 5000 or more. Data in IDB, contains characteristics such as fertility, mortality and migration that provides essential information for tracking the demographic impacts of major events that affects population around the globe.

For this project, the small sample is taken from the existing repository. The dataset containing the 454 observations for eight features. This data contains the observations of 227 countries that are divided into regions and sub-regions. For both genders (male and female), Life expectancy and under age 5 mortality features are collected from the years of 2002 and 2022.

Table 1: Variable types and its description.

Variable Name	Variable Type	Description
country	Categorical	Name of the Country
sub-region	Categorical	Subdivision of a region5
region	Categorical	Continent
year	Categorical	Years of 2002 and 2022
Life expectancy both Sexes	Numeric	The average number of years that a newborn baby can live
Life expectancy Males	Numeric	The average number of years that a male newborn baby can live
Life expectancy Females	Numeric	The average number of years that a female newborn baby can live
Mortality rate both Sexes	Numeric	The average number of deaths of a baby under the age of 5

2.2 Project Objectives

The primary objective of the project is to perform descriptive analysis of the given data and to achieve it there are some tasks defined. These tasks are the following:

1. In first task, frequency distributions of the variables is described while considering the differences between the sexes.

2. In second task, analysing the variability of the values within the individual sub-regions comparing the measures of central tendency of the individual variables between different sub-regions.
3. In third task, carrying out the bivariate correlation analysis between the variables.
4. For the final task, performing the comparison over the last 20 years (i.e., 2002 with 2022).

3 Statistical methods

3.1 Descriptive Statistics

3.1.1 Mean

Mean is one of the important measure of central tendency also referred as arithmetic mean or average. It is the sum of all observation $(x_1, x_2, \ldots, x_n)$ divided by n. There is small difference in the notation of sample mean and population mean, sample mean is denoted by $\overline{x}$ and population mean is denoted by μ.
The standard formula of sample mean is expressed as:

$$x = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Where n represents the no. of observations and the $\sum_{i=i}^{n} x_i$ is the addition of observations xi, for i = 1,...n.

Characteristics of Mean

- Preferred when there are few outliers over median.
- Used with numerical data.
- Helps on comparing the data of various categories.

3.1.2 Median

Median is the second measure of central tendency and is the middle value of the dataset when it is ordered. It is a measure that separates lowest 50% from the highest 50%.

For an even number of observations, the median is calculated as the mean of the two middle values and for the odd number of observations, the median is the middle ordered observation.

$$median(x) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if n is odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} & \text{if n is even} \end{cases} \quad (1)$$

Where n represents the total number of values and x represents the set of numbers.

Characteristics of Median

- Used with numerical data.
- Preferred over mean when the values of data distribution is skewed.

3.1.3 Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) is the common way of measuring a linear correlation. Measures the strength of the relationship between two variables that lies between -1 and 1. When the correlation is 1 then there is a strong linear relationship between two variables means an increase in one variable also increase in the other variable. When correlation is -1 then it is a strong negative linear association between two variables, an increase in one variable results in a decrease in the other variable. When there is no linear association between variables then $r = 0$.
The standard formula of Pearson Correlation Coefficient is expressed as:

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

Where $\overline{x}$ and $\overline{y}$ are the means of x and y.

Characteristics of Pearson Correlation Coefficient

- Used with numerical data.
- It is calculated for the bivariate analysis.
- Suitable to use when the variables show a linear relationship.

3.2 Graphical Methods

3.2.1 Histogram

A histogram is a graphical representation of quantitative data. The representation is like bar graph on a horizontal axis where each bar represents range of numeric values called bins and bins are also referred as intervals or classes. Each interval shares a boundary with another interval representing frequency distribution of numeric values and is the most commonly used.

$$k = \frac{Range}{Number\ of\ Bins}$$

3.2.2 Box Plot

Box plot also known as box-whisker plot gives graphical image where the data is concentrated. The advantage of using box plot to compare distributions between many data sets. A box plot represents five values: minimum value, the lower quartile, median, upper quartile and the maximum value. Median in the line in the box, Lower and Upper quartiles are the starting and the ending line of the box, maximum and minimum are the starting and ending lines of the box plot and the whiskers are the lines between lower quartile to the lowest value and upper quartile to the highest value. It also tells the outliers and their values. An observation is considered as an outlier when it falls beyond 1.5 time the inter-quartile range from both quartiles.

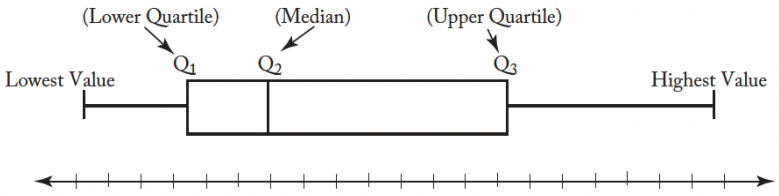
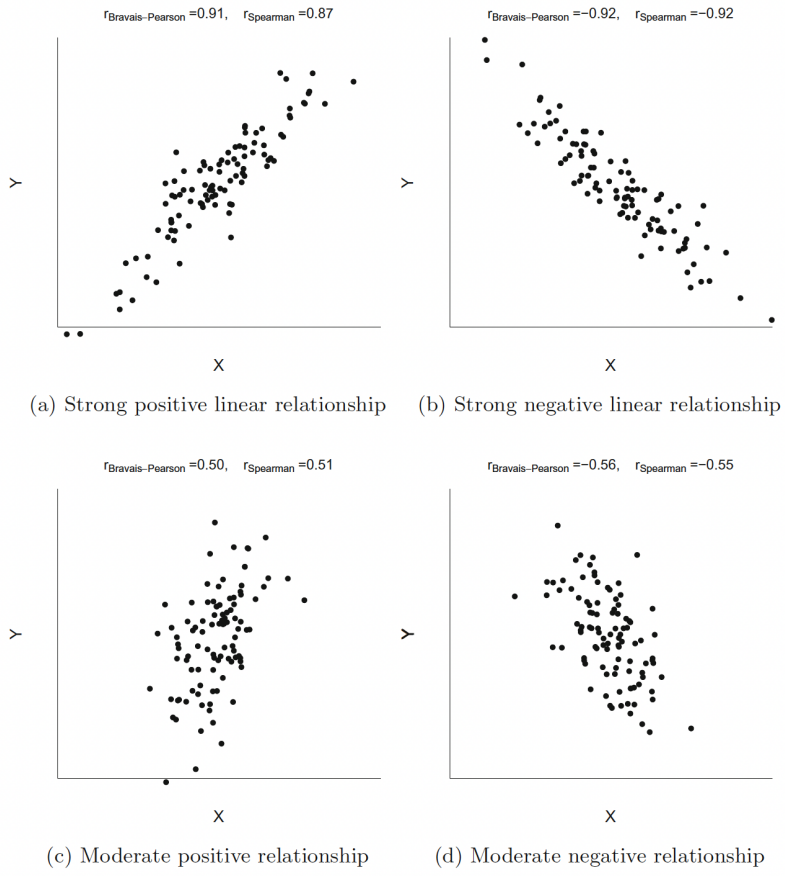


Figure 1: Box plot.(Mendenhall et al., 2009, p. 12)

3.2.3 Scatter Plot

Scatter plot is a graphical representation of showing bivariate relationship between the variables. It is plotted in a 2D graph where one variable is on the x-axis and the other

variable is on the y-axis. Each dot in the graph represents the value of individual variable. The most common application of the scatter plot is to show the correlation between two variables and also show the pattern of the data when taken entirely. Figure 2 shows different kinds of relationships of scatter plot.



(a) Strong positive linear relationship (b) Strong negative linear relationship

(c) Moderate positive relationship (d) Moderate negative relationship

Figure 2: Scatter plot.(Christian Heumann, 2016, p. 12)

4 Statistical analysis

In this section, all statistical methods discussed in the previous section are used to fulfill the project objectives. For the calculation and graphical representation of statistical measures R Software version, 4.2.1 (Core-Team, 2022) is used with the package ggplot2 (Wickham, 2016).

4.1 Frequency distribution of variables

The sample containing demographic data of 454 observations for 8 features. This represent the collection of data for sub-region, region, year, life expectancy at birth for Both Sexes, life expectancy at birth for Males, life expectancy at birth for Females and mortality rate of under age 5. Their variable type and description is discussed in Table 1.

Figure 3 represents histogram of 4 features including life expectancy at birth for Both Sexes, life expectancy at birth for Males, life expectancy at birth for Females and mortality rate of under age 5. Fig 3 represents the univariate distribution of life expectancy at birth for Males and life expectancy at birth for Females respectively for the year 2022. The average life expectancy of male is 72.13 (represented by red line), and its median is 73.26 (represented by blue line). By the values we can see that the distribution illustrates most of the values for males are in between 70-80. For life expectancy of females, the mean is 77.22 and the median is 78.69 which tells females outlive males by 5 years.

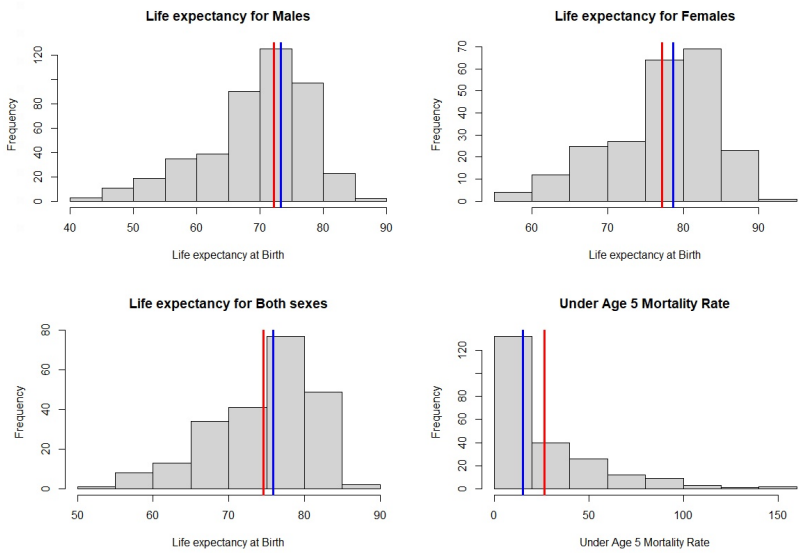


Figure 3: Frequency distribution of life expectancy and under age 5 mortality rate

Similar results can be seen in fig 3 (c) when comparing the life expectancy for both sexes.

In figure 3, the mean of the under age 5 mortality rate for both sexes is 26.53 and the median of 15.08. It shows a positively skewed or right-skewed distribution where the most of the values lies between 0-50. It has a long tail which tells that it is a right-skewed distribution.

4.2 Homogeneity and Heterogeneity with boxplots

This section describes the presence of heterogeneous or homogeneous variability between the numeric variables and sub-regions. In Figure below shows the box plots of sub-regions that are on vertical y-axis and the under age 5 mortality rate represented on the horizontal x-axis. There is a presence of outliers in all sub-regions that are South-Central Asia, South-Eastern Asia, Western Asia and Eastern Asia of the region of Asia. The lowest and the highest under age 5 mortality rate are South-Eastern Asia with a value of 1.94 and South-Central Asia with a value of 213.89. South-Central Asia and South-Eastern Asia shows homogeneous variability.

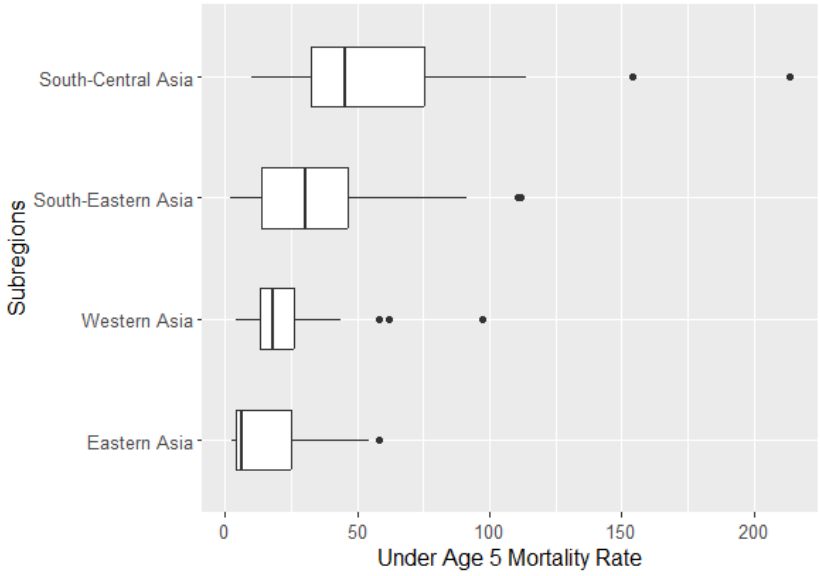
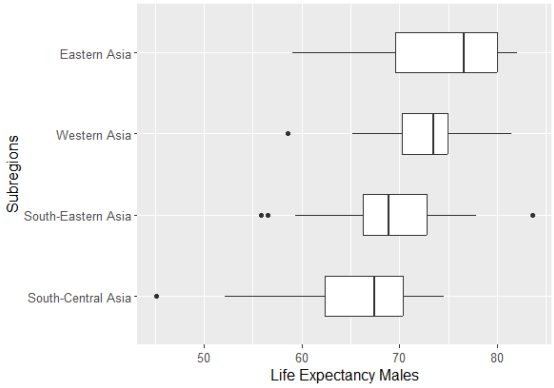
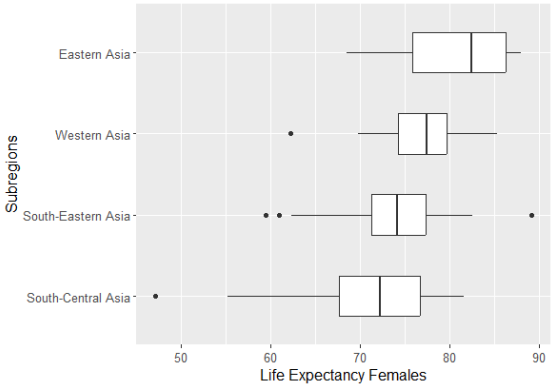


Figure 4: Boxplots for under age 5 mortality rates in sub-regions

Lastly, fig 5 (a) and fig 5 (b) shows boxplot for the life expectancy of male and life expectancy of female where both plots of sub-region Eastern Asia shows high variability. Finally, fig shows the complete picture of the correlation that the lower under age 5 mortality rate leads to a higher life expectancy.



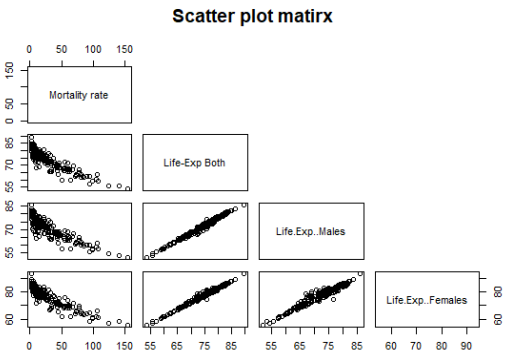
(a) Life expectancy Males



(b) Life expectancy Females

Figure 5: Boxplot of male and female

4.3 Bivariate Correlation



(a) Scatter plot matrix for numeric variables

(b) Heatmap of correlation matrix of numeric variables

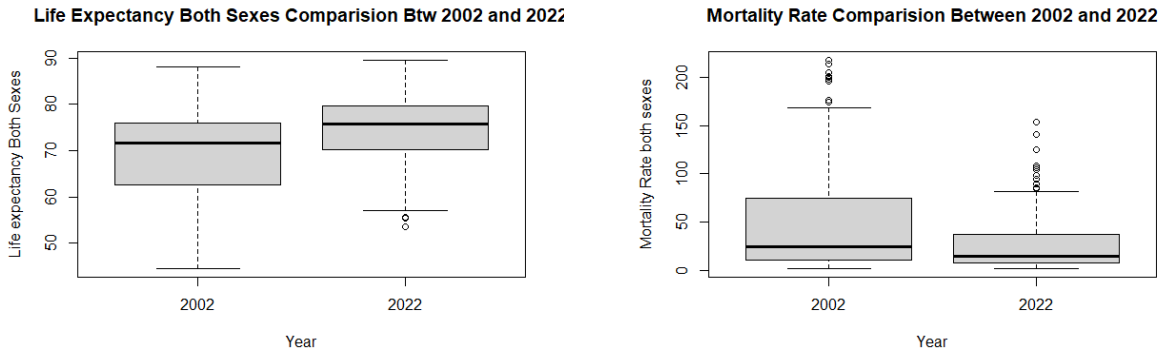
Figure 6: Scatter plot matrix and the Heatmap

To test the strength of linearity between the variables, the Pearson correlation coefficient (r) is used. Fig shows a scatter plot matrix for the numeric variables. There is a strong negative correlation between under age 5 mortality rate and life expectancy for both sexes having a linear correlation of -0.9. This result provides an understanding that decrease in the rate of mortality increases the population of life expectancy for both genders. Similarly, the comparison between mortality rate, life expectancy for males and females have a negative relationship where males have -0.88 and females have -0.904. Additionally, a strong positive linear correlation between life expectancy for both genders, life expectancy for males and life expectancy for females is found with r of 0.993

for each of the variables. This positive relationship shows that both life expectancy of male and female contribute to the common good of life expectancy for both sexes.

4.4 Change of variables over the years

This section illustrates that how mortality and life expectancy rates affected over the last two decades. The fig below shows that there is an increased trend in the life expectancy of both genders for the year 2022 compare to the life expectancy of the year 2002. A median of life expectancy for the year 2022 is 75.82 compare to the year 2002 which is 71.65 that results in 2022 more people under the age of 5 are expected to live through to the age of 75.



(a) Life expectancy both sexes

(b) Under age 5 mortality rate

Figure 7: Boxplots to compare life expectancy and under age 5 mortality rate

As there is an upward trend in the life expectancy for the year 2022 indicates that there is a low trend of mortality rate for the year 2022. Also from the figure above the mortality rate is dropped to the median of 15.08 in comparison to the year 2002 which has a median of 24.42.

The positive impact in the result of mortality rate and life expectancy rate can be because of the development of the society, modern research that brings vaccines to prevent infections and good parental care etc.

5 Summary

In this project, the sample dataset was taken from the International Database (IDB) of the United States Census Bureau (USCB). The primary objective of the project was to perform a descriptive analysis using appropriate statistical methods. The data includes observations collected from 227 countries for the years 2002 and 2022. The data is providing the information of population, life expectancy at birth for the genders and under age 5 mortality rate.

Firstly, the histogram was constructed to illustrate the frequency distribution of variables for the year 2022. For all the histograms, both mean and median were represented on the plot. It was observed that the average life expectancy for both the sexes was around 74.61 and death rate of under age 5 was on the average of 26.54 for the year 2022. Additionally, a bivariate correlation was performed between the variables to determine how mortality rate affected life expectancy. Thereafter, mortality and life expectancy of the sub-regions of Asia were compared to check the variability which resulted in a conclusion that the life expectancy for male and female have homogeneous pattern. Lastly, life expectancy and mortality rate was checked of whole dataset of 2 decades. As a result, life expectancy and under age 5 mortality rate of 2022 was found as a positive change in comparison to the year 2002.

For further studies, there are more factors that can be involved to have a better understanding of the relationship between under age 5 mortality and life expectancy and the factors include medical conditions of a mother, her diet and her age, etc. The methods used in this project can be implemented on the larger dataset for more than two population indices.

Bibliography

Akinkunmi, Mustapha. 2018. Introduction to Statistics using R. Morgan Claypool Publishers.

R Development Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Christian Heumann, Michael Schomaker, Shalabh. 2016. Introduction to Statistics and data analysis. Springer International Publishing AG.

Wiley, Joshua F. 2020. JWileymisc: Miscellaneous Utilities and Functions. R package version 1.2.0.

Wickham, Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Appendix

A Additional figures

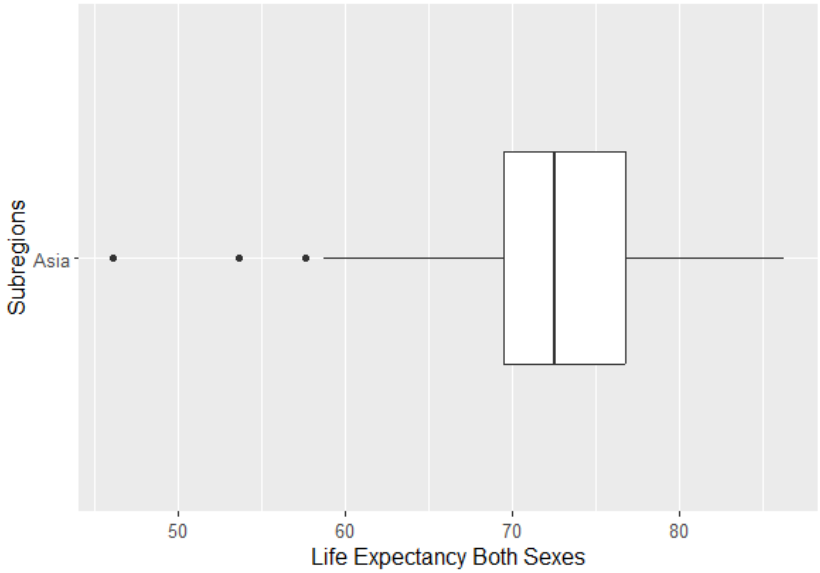


Figure 8: Life expectancy for both sexes

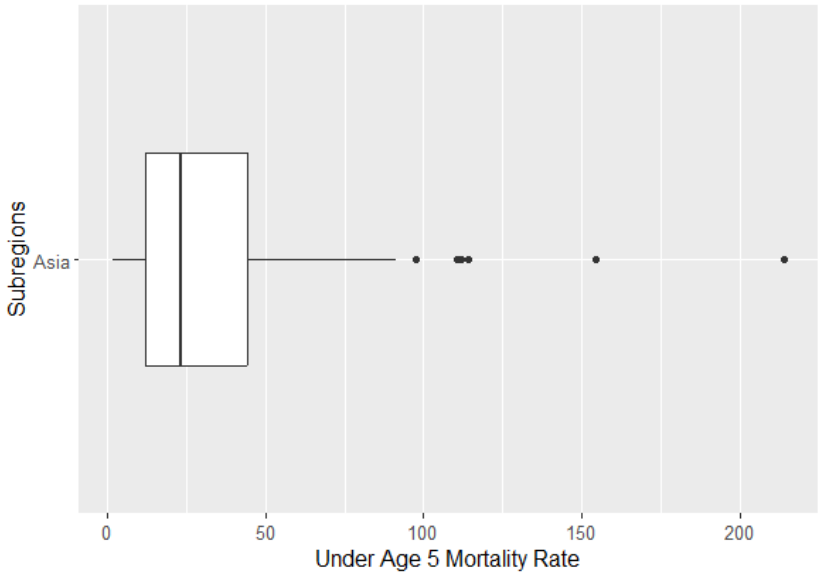


Figure 9: Under Age 5 Mortality Rate for both sexes