A STUDY ON THE LIFE EXPECTANCIES OF 183 COUNTRIES FROM THE YEAR 2000 TO 2015

Project report submitted in the partial fulfilment of the requirement for the Bachelor's Degree in STATISTICS under

MAHATMA GANDHI UNIVERSITY

Kottayam

SUBMITTED BY

Anet Mary George

Anaswara P Biju

Ann Mariya Moncy

Rahul G Nair

Arun P S

Nidheesh Sahadevan

MAR ATHANASIUS COLLEGE (AUTONOMOUS)

KOTHAMANGALAM

2020 - 2023

MAR ATHANASIUS COLLEGE (AUTONOMOUS)

KOTHAMANGALAM

Department of Statistics

CERTIFICATE

This is to certify that the project report entitled **“A STUDY ON THE LIFE EXPECTANCIES OF 183 COUNTRIES FROM 2000 TO 2015 ”** is an original and authentic record of study carried out by Anet Mary George, Anaswara P Biju, Ann Mariya Moncy, Rahul G Nair, Arun P S and Nidheesh Sahadevan under my guidance and supervision, in the partial fulfilment of the requirement for the award of the degree of Bachelor of Science, in Statistics during the academic year 2022 – 2023 and that it has not been previously submitted to the award of any degree, diploma, fellowship or any other similar title or recognition.

By

Sudha V

Head of the Department of Statistics

Mar Athanasius College (Autonomous)

Kothamangalam

DECLARATION

We, Anet Mary George, Anaswara P Biju, Ann Mariya Moncy, Rahul G Nair, Arun P S and Nidheesh Sahadevan, hereby declare that the project work entitled **"A STUDY ON THE LIFE EXPECTANCIES OF 183 COUNTRIES FROM 2000 TO 2015"** is a sincere work done by us in 2023 for the partial fulfilment of the requirement for a Bachelor's Degree in Statistics at Mar Athanasius College (Autonomous), Kothamangalam under the guidance of Ms. Sudha V, Head of the Department of Statistics, Mar Athanasius College (Autonomous), Kothamangalam.

By

Name: Anet Mary George
Anaswara P Biju
Ann Mariya Moncy
Rahul G Nair
Arun P S
Nidheesh Sahadevan

ACKNOWLEDGEMENT

At the outset, we thank God Almighty for making the endeavour a success. We express our deep sense of gratitude and thank SUDHA.V, the Head of the Department of Statistics for her valuable guidance, suggestions and support throughout the completion of this project.

We are thankful to Dr T.M Jacob , the Head of the Statistics Department, at Nirmala College, Muvattupuzha and his MSc. Statistics students (2021-2023), for their valuable guidance and mentorship throughout our project. We are also grateful to Mr. Basil Mathew Biju for his support in the completion of this project.

Last but not the least, we also express our gratitude to all other faculty members of the department and our friends who fostered us in various occasions during the project work.

Kothamangalam: By

Date: Name: Anet Mary George

Anaswara P Biju

Ann Mariya Moncy

Rahul G Nair

Arun P S

Nidheesh Sahadevan

CONTENTS

INTRODUCTION

WHAT IS STATISTICS?

Statistics constitutes an integral part of every scientific and economic industry. Social and economic studies without Statistics are inconceivable. Statistics is indispensable in almost all the spheres of human activity and knowledge. It has an important and essential place in the development of science, both pure and social. Statistics plays a multifarious role and Tippett has rightly remarked “affects everybody and touches life at many places”.

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data or as a branch of Mathematics. Some consider Statistics to be a distinct mathematical science rather than a branch of Mathematics. While many scientific investigations make use of data, statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty.

In applying Statistics to a problem, it is common practice to start with a population or process to be studied. Populations can be diverse topics such as “all people living in a country” or “every atom composing a crystal”. Ideally, statisticians compile data about the entire population (an operation called census). This may be organized by governmental statistical institutes. Descriptive Statistics can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data (like income), while frequency and percentage are more useful in terms of describing categorical data (like education).

When a census is not feasible, a chosen subset of the population called a sample is studied. Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting. Again, Descriptive Statistics can be used to summarize the sample data. However, drawing the sample contains an element of randomness; hence, the numerical descriptors from the sample are also prone to uncertainty. To draw meaningful conclusions about the entire population, Inferential Statistics is needed. It uses patterns in the sample data to draw inferences about the population represented while accounting for randomness. These inferences may take the form of answering yes/no questions about the data (hypothesis testing), estimating numerical characteristics of the data (estimation), describing associations within the data (correlation), and modelling relationships within the data (for example, using regression analysis). Inference can extend to forecasting, prediction, and estimation of unobserved values either in or associated with the population being studied. It can include extrapolation and interpolation of time series or spatial data, and data mining.

Mathematical Statistics is the application of Mathematics to Statistics. Mathematical techniques used for this include mathematical analysis, linear algebra, stochastic analysis, differential equations, and measure-theoretic probability theory

DEFINITIONS OF STATISTICS

Statistics is the practice or science of collecting and analysing numerical data in large quantities, especially for the propose of inferring proportion in a whole from those in a representative sample.

Different authors had given different definition of statistics. Some of the definition of statistic describing it as quantitatively are:

- " Statistics is the science that deals with the analysis of masses of quantitative data. it includes the collection, classification, presentation and interpretation of such data."

- R.H. WESSEL

- "Statistical methods are methods especially adapted to the elucidation of quantitative data affected by multiplicity of causes "

- G.U. YULE

- "Statistics is a system of analysis and synthesis of numerical data for the purpose of obtaining and diffusing knowledge "

-HAWARD.L. BALSLEY

- "Statistics are measurements, enumerations or estimates of natural or social phenomena, systematically arranged to exhibit their inner relations."

- CONNER

- "The theory and methods of collecting, tabulating and analysing numerical data comprise the study of Statistics as the subject "

- TARO YAMANE

- "Statistics may be defined as the principles and methods used on the collection, presentation, analysis and interpretation of numerical data ".

- LAWRENZ.G. KAPLAN

Statistics is a numerical statement of facts in any department of enquiry placed in relation to each other “.

- BOWLEY

- “By Statistics we mean the aggregate of facts affected to a marked extent by multiplying of causes numerically expressed, enumerated or estimated according to reasonable standard of accuracy collected in a systematic manner for a pre - determined purpose and placed in relation to each other "

- SECRIST

- “Statistics is a body of theories and methods which have been developed for handling the collection, analysis and description of sample data for drawing useful conclusions "

- LIN COLN L CHEO

ORIGIN OF STATISTICS

The subject of Statistics, as it seems, is not a new discipline but it is as old as human society itself. Its origin can be traced to the old days when it was regarded as the “Science of Statecraft” and was the by-product of the administrative activity of the State. The word Statistics was first used by a German Scholar Gottfried Achenwall in the middle of the 18th century as the science of statecraft concerning the collection and use of data by the state. The word Statistics comes from the Latin word “Status” or Italian word “Statista” or German word “Statistik” or the French word “Statistique”, meaning a political state, and originally meant information useful to the state, such as information about sizes of the population (human, animal, products, etc.) and armed forces.

According to Pioneer Statistician Yule, the word Statistics occurred at the earliest in the book “The Element of Universal Erudition” by Baron (1770). In 1787, a wider definition was used by E.A.W. Zimmermann, “A Political survey of the present state of Europe”. It appeared in the Encyclopedia of Britannica in 1797 and was used by Sir John Sinclair in Britain in a series of volumes published between 1791 and 1799 giving a statistical account of Scotland. The theoretical development of the so-called modern Statistics came during the mid-seventeenth century with the introduction of Theory of Probability and Theory of Games and Chance, the chief contributors being mathematicians and gamblers of France, Germany and England. In the 19th century, the word Statistics acquired a wider meaning covering numerical data of almost any subject whatever and also interpretation of data through appropriate analysis. That’s all about the short history of Statistics. Now let us see how statistics is being used in different meanings nowadays.

•Statistics refers to “numerical facts that are arranged systematically in the form of tables or charts etc. In this sense, it is always used a plural i.e., a set of numerical information. For instance, Statistics of prices, road accidents, crimes, births, educational institutions, etc.

•The word Statistics is defined as a discipline that includes procedures and techniques used to collect, process, and analyse the numerical data to make inferences and to reach appropriate decisions in situations of uncertainty (uncertainty refers to incompleteness, it does not imply ignorance). In this sense word Statistic is used in the singular sense. It denotes the science of basing decisions on numerical data.

•The word Statistics are numerical quantities calculated from sample observations; a single quantity calculated from sample observations is called Statistics such as the mean. Here word Statistics is plural.

THE GROWTH OF STATISTICS

The growth of Statistics was tremendous in the twentieth century. During this period, lot of new theories, applications in various disciplines were introduced. With the contribution of renowned Statisticians several theories and methods were introduced, naming a few are Probability Theory, Sampling Theory, Statistical Inference, Design of Experiments, Correlation and Regression Methods, Time Series and Forecasting Techniques.

In early 1900s, Statistics and Statisticians were not given much importance but over the years due to advancement of technology it had its wider scope and gained attention in all fields of science and management. We also tend to think Statistician as a small profession but a steady growth in the last century is impressive. It is pertinent to note that the continued growth of Statistics is closely associated with information technology. As a result, several new interdisciplines have emerged. They are Data Mining, Data Warehousing, Geographic Information System, Artificial Intelligence etc. Nowadays, Statistics can be applied in hardcore technological spheres such as Bioinformatics, Signal processing, Telecommunications, Engineering, Medicine, Crimes, Ecology, etc.

Today’s business managers need to learn how analytics can help them make better decisions that can generate better business outcomes. They need to have an understanding of the statistical concepts that can help analyse and simplify the flood of data around them. They should be able to leverage analytical techniques like decision trees, regression analysis, clustering and association to improve business processes.

FUNCTIONS OF STATISTICS

- The science of Statistics presents complex and complicated data into an understandable way. Thus, it simplifies complicated data.
- Statistics helps in planning the economic policy of the state.
- Statistics does not simplify the complex data, but compares their relation with other simplified data also.

- Statistics are necessary for efficient and sound administration.
- Statistical methods form the key note of a quantitative inquiry.

DIVISIONS OF STATISTICS

The field of Statistics is divided into two major divisions: Descriptive Statistics and Inferential Statistics. Each of these segments is important, offering different techniques that accomplish different objectives.

❖ Descriptive Statistics

Descriptive Statistics is the type of Statistics that probably springs to most people's minds when they hear the word "Statistics." Descriptive Statistics describe what is going on in a population or data set. In this branch of Statistics, the goal is to describe. Numerical measures are used to talk about features of a set of data. Measures of central tendency, dispersion, skewness and kurtosis constitute the essence of Descriptive Statistics.

❖ Inferential Statistics

Inferential Statistics are produced through complex mathematical calculations that allow scientists to infer trends about a larger population based on a study of a sample taken from it. Scientists use inferential statistics to examine the relationships between variables within a sample and then make generalizations or predictions about how those variables will relate to a larger population.

SCOPE OF STATISTICS

Statistics plays a vital role in every field of human activity. Statistics helps in determining the existing position of per capita income, unemployment, population growth rates, housing, schooling medical facilities, etc., in a country.

Now Statistics holds a central position in almost every field, including industry, commerce, trade, Physics, Chemistry, Economics, Mathematics, Biology, Botany, Psychology, Astronomy, etc., so the application of Statistics is very wide. Following are the importance and scope of Statistics.

• Business

Statistics plays an important role in business. A successful businessman must be very quick and accurate in decision making. Statistics helps businessmen to plan production according to the taste of the customers, and the quality of the products can also be checked more efficiently by using statistical methods. Thus, it can be seen that all business activities are based on statistical information. Businessmen can make correct decisions about the location of business, marketing of the products, financial resources, etc.

• Economics

Economics largely depends upon Statistics. National income accounts are multipurpose accounts. In Economics research, statistical methods are used to collect and analyse the data and test hypotheses. The relationship between supply and demand is studied by statistical methods; imports and exports, inflation rates, and per capita income are problems which require a good knowledge of statistics.

• Mathematics

Statistics helps in describing the measurements more precisely. Statistics is a branch of applied Mathematics. A large number of statistical methods like probability averages, dispersions, estimation, etc., is used in Mathematics, and different techniques of pure Mathematics like integration, differentiation and algebra are used in statistics.

• Banking

Statistics plays an important role in banking. Banks make use of Statistics for a number of purposes. They work on the principle that everyone who deposits their money with the banks does not withdraw it at the same time. The bank earns profits out of these deposits by lending it to others on interest. Bankers use statistical approaches based on probability to estimate the number of deposits and their claims for a certain day.

• State Management (Administration)

Statistics is essential to a country. Different governmental policies are based on Statistics. Statistical data are now widely used in making all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the cost of living, and statistical methods will be used to determine the rise in the cost of living. The preparation of federal and provincial government budgets mainly depends upon Statistics because it helps in estimating the expected expenditures and revenue from different sources. So Statistics are the eyes of the administration of the state.

• Accounting and Auditing

Accounting is impossible without exactness. But for decision making purposes, so much precision is not essential; the decision may be made on the basis of approximation, know as Statistics. The correction of the values of current assets is made on the basis of the purchasing power of money or its current value. In auditing, sampling techniques are commonly used. An auditor determines the sample size to be audited on the basis of error.

• Natural and Social Sciences

Statistics plays a vital role in almost all the natural and social sciences. Statistical methods are commonly used for analysing experiments results, and testing their significance in Biology, Physics, Chemistry, Mathematics, Meteorology, research, chambers of commerce, Sociology, Business, Public administration, Communications and Information Technology, etc.

• Astronomy

Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, and sizes, masses and densities of heavenly bodies by means of observations. During these measurements errors are unavoidable, so the most probable measurements are found by using statistical methods. Example: This distance of the moon from the earth is measured. Since history, astronomers have been using statistical methods like method of least squares to find the movements of stars.

INDIAN STATISTICAL ORGANISATIONS

- Central Statistics Office
- Ministry of Statistics and Programme Implementation
- Indian Statistical Institute, Kolkata
- National Statistical Commission
- Economic and Statistical Organization Punjab (ESO)
- Department of Economic and Statistical Analysis, Haryana
- Ministry of Finance

APPLICATION OF STATISTICS

Statistics and its principles are applied to a wide variety of problems in government, business, and industry as well as problems in biological, physical, and social sciences. They are applied to operating as well as research problems. Following are a few applications of Statistics.

• Government Agencies

The government uses Statistics to make decisions about populations, health, education, etc. It may conduct research on education to check the progress of high school students using a specific curriculum or collect characteristic information about the population using a census.

• Science and Medicine

Medical field would be far less effective without research to see which medicines or interventions work best and how the human bodies react to treatment. Medical professionals also perform studies by race, age, or nationality to see the effect of these characteristics on health.

• Psychology

Although this is attached to both the science and medical field, success in Psychology would be impossible without the systematic study of human behaviour, often analysing results statistically.

• Education

Teachers are encouraged to be researchers in their classrooms, to see what teaching methods work on which students and understand why. They also should evaluate test items to determine if students are performing in a statistically expected way. At all levels of education and testing, there are statistical reports about student performance, from kindergarten to an SAT or GRE

• Large Companies

Every large company employs its own statistical research divisions or firms to research issues related to products, employees, customer service, etc. Business success relies on knowing what is working and what isn't.

• Sector

In agricultural experiment stations and in some scientific and industrial laboratories, experiment designs based upon statistical principles are being used to test differences between alternative methods, process and products. Statistics and its principles are basic requirements for the efficient design of laboratory and field experiments.

LIMITATIONS OF STATISTICS

• **Study of Numerical Facts only:**

Statistics studies only such facts as can be expressed in numerical terms. It does not study qualitative phenomena, like honesty, friendship, wisdom, health, patriotism, justice, etc.

• **Study of Aggregates only:**

Statistics studies only the aggregates of quantitative facts. It does not study any particular unit

• **Not the only Method:**

Statistical method is not the only method to study. Many a time this method does not suggest the best solution of each problem. The conclusions drawn on the basis of Statistics should be verified with the help of the conclusion drawn with the help of qualitative methods.

• **Homogeneity of Data:**

Quantitative data must be uniform and homogeneous. To compare the data, it is essential that whatever Statistics are collected, the same must be uniform in quality. Data of different qualities and kinds cannot be compared.

• **Results are true only on an Average:**

Laws of Statistics are true only on an average. They express tendencies. Unlike the laws of Physical Science or Chemistry, they are not absolutely true. They are not valid always and under all conditions. For instance, if it is said that per capita income in India is 6000 per annum, it does not mean that the income of each and every Indian is 6000 per annum. Some may have more and some may have less than it. It is true only on an average.

• **Without Reference Results May Prove Wrong:**

In order to understand the conclusions very well, it is necessary that the circumstances and conditions under which these conclusions have been drawn are also studied, otherwise they may prove wrong.

• **Can be used only by Experts:**

Statistics can be used only by those persons who have special knowledge of Statistical methods. Those who are ignorant about these methods cannot make use of it. It can, therefore, be said that data in the hands of an unqualified person is like a medicine in the hands of a quack who may abuse it out of ignorance leading to dangerous results. In the words of Yule and Kendall, “Statistical Methods are most dangerous tools in the hands of an inexpert”.

• **Misuse of Statistics is Possible:**

Misuse of Statistics is possible. It may prove true what actually is not true. It is usually said, “Statistics are like clay of which you can make a god or devil, as you please”. Misuse of Statistics is, therefore, its greatest misuse.

• **Only Means and not a Solution:**

Some scholars are of the opinion that Statistics are only a means in the solution of any problem. It is not a solution to the problem. To check the misuse of Statistics, conclusions should be drawn impartially and without any selfish interest. Otherwise, Statistics may not become a proper means for the solution of any problem In short, while making use of Statistics, its limitations as discussed above, must always be kept in mind.

• **Misuses of Statistics**

Statistics can be easily misused. Because of this distrust of everything Statistics has developed. Statements like ‘there are black lies, white lies and Statistics ‘, and ' Statistics can prove anything ' have gained currency because of this. But it is doubtful whether the fault is in the science or with people who handle it. Politicians and advertisers often deceive people by wrongly using statistical information. The figures they quote may be incomplete manipulated or quoted without giving the circumstances under which they were collected. All these may lead to wrong conclusions. Statistics is only a method of approach and analysing the same data in different ways can arrive at different conclusions. So even when there is no intention of misleading others, if the data is not analysed by an expert, the conclusions may be misleading. Only one who has the theoretical as well as practical knowledge of statistical methods can efficiently analyse a data. Statistics is a good servant but a bad master.

MINISTRY OF STATISTICS AND PROGRAMME IMPLEMENTATION (MoSPI)

The Ministry of Statistics and Programme Implementation (MoSPI) is a ministry of Government of India concerned with coverage and quality aspects of statistics released. The surveys conducted by the Ministry are based on scientific sampling methods.

The Ministry of Statistics and Programme Implementation (MOSPI) came into existence as an Independent Ministry on 15 October 1999 after the merger of the Department of Statistics and the Department of Programme Implementation. The Ministry has two wings, one relating to Statistics and the other Programme Implementation.

The Statistics Wing called the National Statistical Office (NSO) consists of the Central Statistical Office (CSO), the computer centre and the National Sample Survey Office (NSSO). The Programme Implementation Wing has three Divisions, namely, (i) Twenty Point Programme (ii) Infrastructure Monitoring and Project Monitoring and (iii) Member of Parliament Local Area Development Scheme. Besides these two wings, there is National Statistical Commission created through a Resolution of Government of India (MOSPI) and one autonomous Institute, viz., Indian Statistical Institute declared as an institute of National importance by an Act of Parliament.

The Ministry of Statistics and Programme Implementation attaches considerable importance to coverage and quality aspects of statistics released in the country. The Statistics released are based on administrative sources, surveys and censuses conducted by the centre and State Governments and nonofficial sources and studies. The surveys conducted by the Ministry are based on scientific sampling methods. Field data are collected through dedicated field staff. In line with the emphasis on the quality of Statistics released by the Ministry, the methodological issues concerning the compilation of national accounts are overseen Committees like Advisory Committee on National Accounts, Standing Committee on Industrial Statistics, Technical Advisory Committee on Price Indices. The Ministry compiles data sets based on current data, after applying standard statistical techniques and extensive scrutiny and supervision.

❖RESPONSIBILITIES

- National Statistical Office (NSO) is mandated with the following responsibilities:

- Acts as the nodal agency for planned development of the statistical system in the country, lays down and maintains norms and standards in the field of statistics, involving concepts and definitions, methodology of data collection, processing of data and dissemination of results;

- Coordinates the statistical work in respect of the Ministries/Departments of the Government of India and State Statistical Bureaus (SSBs), advises the Ministries/Departments of the Government of India on statistical methodology and on statistical analysis of data.

- Prepares national accounts as well as publishes annual estimates of national product, government and private consumption expenditure, capital formation, savings, estimates of capital stock and consumption of fixed capital, as also the state level gross capital formation of supra-regional sectors and prepares comparable estimates of State Domestic Product (SDP) at current prices;

- .Maintains liaison with international statistical organizations, such as, the United Nations Statistical Division (UNSD), the Economic and Social Commission for Asia and the Pacific (ESCAP), the Statistical Institute for Asia and the Pacific (SIAP), the International Monetary Fund (IMF), the Asian Development Bank (ADB), the Food and Agriculture Organizations (FAO), the International Labor Organizations (ILO), etc.

- Compiles and releases the Index of Industrial Production (IIP) every month in the form of 'quick estimates'; conducts the Annual Survey of Industries (ASI); and provides statistical information to assess and evaluate the changes in the growth, composition and structure of the organized manufacturing sector;

- Organizes and conducts periodic all-India Economic Censuses and follow-up enterprise surveys, provides an in-house facility to process the data collected through various socioeconomic surveys and follow-up enterprise surveys of Economic Censuses;

- Conducts large scale all-India sample surveys for creating the database needed for studying the impact of specific problems for the benefit of different population groups in diverse socioeconomic areas, such as employment, consumer expenditure, housing conditions and environment, literacy levels, health, nutrition, family welfare, etc.;

- Examines the survey reports from the technical angle and evaluates the sampling design including survey feasibility studies in respect of surveys conducted by the National Sample Survey Organizations and other Central Ministries and Departments;

- Dissemination of statistical information on various aspects through a number of publications distributed to Government, semi-Government, or private data users/ agencies; and disseminates data, on request, to the United Nations agencies like the UNSD, the ESCAP, the ILO and other international agencies;

- Releases grants-in-aid to registered Non-Governmental Organizations and research institutions of repute for undertaking special studies or surveys, printing of statistical reports, and financing seminars, workshops and conferences relating to different subject areas of official Statistics.

❖Programme Implementation Wing Responsibilities

- Monitoring of the Twenty Point Programme (TPP)

- Monitoring the performance of the country's 11 key infrastructure sectors, viz., Power, Coal, Steel, Railways, Telecommunications, Ports, Fertilizers, Cement, Petroleum & Natural Gas, Roads and Civil Aviation (IPMD);

- Monitoring of all Central Sector Projects costing Rs. 150 crore and above (IPMD)

- Monitoring the implementation of Members of Parliament Local Area Development Scheme (MPLADS).

STATISTICS IN EVERYDAY LIFE

- Medical Studies: Scientists ought to show a statistically legitimate price of effectiveness earlier than any drug can be prescribed. Statistics are at the back of each clinical study

- Predicting Diseases: A lot of times on information reviews, facts about sickness are said. When facts turn out to be involved, we will have a higher idea of how that sickness may additionally affect you.

- Weather Forecast: The computer models are built using facts that compare prior climate stipulations with the modern climate to predict future climate.

- Emergency Preparedness: Emergency administration organizations cross into high alert to be prepared to rescue humans. Emergency teams remember on Statistics to inform them when risk may show up.

- Genetics: Statistics are necessary in determining the possibilities of a new baby being affected by the disease.

- Insurance: The price that on insurance plan organization fees is based totally upon facts from all drivers or home owners in our area.

ABOUT THE DATA

Life expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, current age, and other demographic factors like income level , education etc... Shortly, it is the average number of years a person is expected to live. Our project is focused on studying the life expectancies of 183 countries between the years 2000 and 2015 with variables as Life expectancy, Adult mortality, Infant death, GDP, Schooling, Alcohol consumption, Income composition of resources etc…Since the observations of this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to the higher as well as lower value of life expectancy. Analysing the key determinants of Life Expectancy can provide valuable insights for a country to prioritize and focus on specific areas for improvement, ultimately leading to a healthier and more prosperous population.

The Global Health Observatory (GHO) data repository under World Health Organisation (WHO) keeps track of the health status as well as many other related factors for all countries, the data-sets are made available to public for the purpose of health data analysis. The data set related to life expectancy has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. It has been observed that there was a huge development in health sector , in comparison with the past 30 years. Therefore, in this project , we have considered the data from the year 2000 to 2015. On initial visual inspection , the data showed some missing values. These missing values were handled in R Software. Kaggle, a popular online platform for data science competitions and hosting datasets is the source of this particular dataset.

VARIABLES UNDER STUDY

Life expectancy: This variable represents the average number of years a person can expect to live in a given country. It is typically calculated using data on deaths and population size.

Adult mortality: This variable represents the probability of dying between the ages of 15 and 60 in a given country. It is often used as a measure of the overall health of a population.

Infant deaths: This variable represents the number of deaths of children under the age of one in a given country. It is often used as an indicator of the quality of healthcare and other services available to mothers and infants.

Alcohol consumption: This variable represents the amount of alcohol consumed per capita in a given country. It is often used as an indicator of risky health behaviours and may have an impact on life expectancy.

Schooling: This variable represents the average number of years of schooling completed by individuals in a given country. It is often used as an indicator of educational attainment and may be related to health outcomes.

Income composition of resources: This variable typically refers to the distribution or composition of income sources within a country. It provides insights into the economic structure and the extent to which different sectors contribute to national income. For example, it may include factors such as the share of income derived from agriculture, industry, or services.

GDP (Gross Domestic Product): The total value of goods and services produced within a country's borders in a given year.

Total expenditure: The total amount of money spent by a country's government on goods and services, including investments and transfers.

Percentage expenditure: The proportion of a country's GDP that is spent on government expenditure.

HIV/AIDS: The variable is the measure of number of deaths per 1000 live births due to HIV/AIDS in children between the ages of 0 to 4 years.

Under five deaths : The number of deaths that occur in children under the age of five**.**

Population : The total number of individuals living in a particular geographic area or country at a given point in time.

SOFTWARE USED

R is a programming language and open-source software environment used for statistical computing, data analysis, and graphics. R was first developed in 1993 by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and more.

R is widely used in data science, machine learning, and artificial intelligence research and applications. It has a large and active user community and offers a vast library of packages, which provide additional functionality for specific purposes. R is available for free under the GNU General Public License and can be downloaded and installed on various platforms, including Windows, macOS, and Linux.

OBJECTIVES OF THE STUDY

The objective of our project appears to analyse the factors that affect life expectancy in different countries and to understand how these factors have changed over time. The variables such as Adult mortality, Infant deaths, Alcohol consumption, Schooling, GDP and income consumption of resources, are all important factors that can influence life expectancy.

By analysing these variables across different countries and over different years, we may be able to identify patterns and trends in life expectancy and the factors that affect it. For example, we may be able to identify which variables have the strongest correlations with life expectancy, which countries have seen the biggest improvements in life expectancy over time, and which countries still face challenges in improving life expectancy.

Ultimately, the goal of our project is likely to increase our understanding of the complex factors that influence life expectancy and to help inform public policy decisions related to healthcare, education, economic development, and other areas that can impact population health.

OBJECTIVES

- Plot a boxplot to analyse the variation in Life Expectancy through the years.
- Construct correlation matrix to study whether there is any relationship between the variables.
- To check whether the variables (Income, Schooling, & Adult Mortality) associated with Life Expectancy are significant or not. (ANOVA)
- To fit a regression model using the variables Life Expectancy, Income, Schooling, & Adult Mortality. (Multiple Regression)
- To cluster the countries based on the variables Life expectancy , Adult mortality, schooling , income composition resources of the year 2015 only.(K - Means Clustering).

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory analysis is a statistical approach to analysing data that focuses on discovering patterns, trends and relationships within a dataset. The primary goal of the exploratory analysis is to gain insights into the data and identify potential areas of interest or further investigation. Exploratory analysis typically involves visualising and summarising the data using various statistical and graphical techniques, such as histograms, scatterplots, boxplots and summary techniques.

It is particularly useful when working with large and complex datasets, as it can help you quickly identify potential patterns or anomalies that might not be immediately apparent from the raw data. Exploratory analysis can also help you generate hypothesis and guide further analysis or modelling.

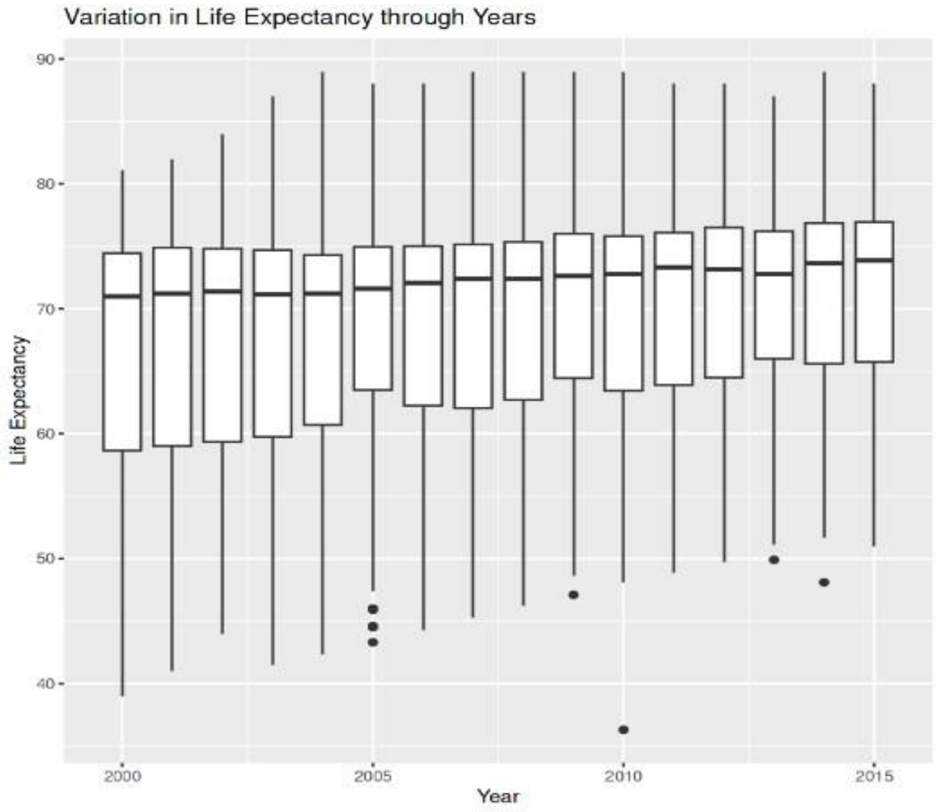
Overall, exploratory analysis is an essential first step in any statistical analysis or modelling project, as it provides a foundation for understanding the data and generating insights that can inform subsequent analysis and decision making.

OBJECTIVE

Plot a boxplot to analyse the variation in Life Expectancy through the years, 2000 to 2015.

BOX PLOT

Box plot is a graphical representation of a data distribution through five statistical summary measures: minimum, first quartile, median, third quartile and maximum. The box in the plot represents the middle 50% of the data, with the bottom and top edges of the box representing the first and third quartiles. The line within the box represents the median. The whiskers extending from the box represent the range of the data, excluding any outliers. Outliers, if present, are plotted as individual points beyond the whiskers. Box plots are useful for identifying any skewness, symmetry or outliers in the dataset and for comparing the distributions of multiple datasets side by side.



INFERENCE

The median life expectancy appears to have increased over time , as the median line within each boxplot appears to be moving upwards, as we move towards more recent years .
The plot clearly suggests that , the data is negatively skewed.
The minimum and maximum values also appear to be increasing over time, suggesting an overall increase in the trend of the life expectancy values.
The Box plot demonstrates a decreasing dispersion and Inter Quartile Range (IQR).

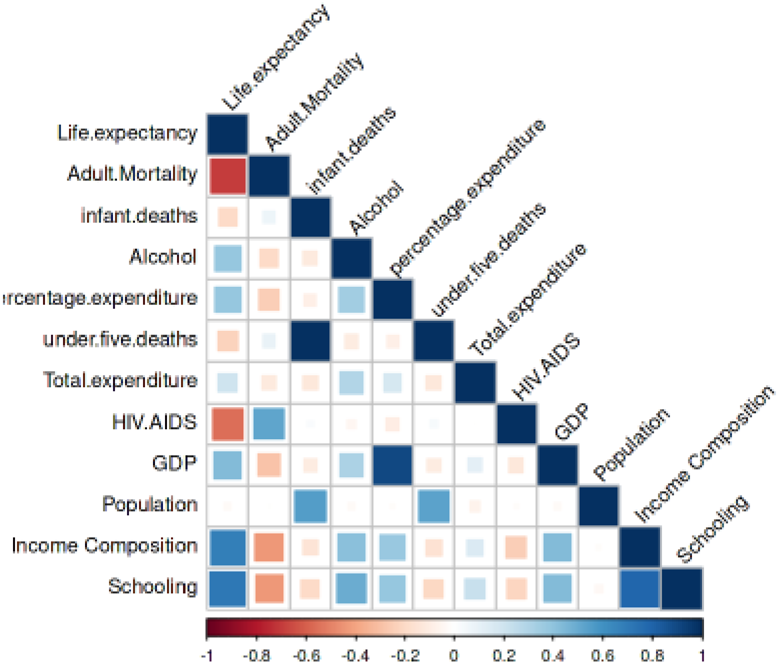
OBJECTIVE

Plotting a correlation matrix between the variables of the Life Expectancies of 183 countries from 2000 to 2015.

CORRELATION MATRIX

A correlation matrix is used to show the degree of the linear relationship between variables in a dataset. It indicates the correlation using the correlation coefficient.

Correlation coefficients measure the strength and direction of the linear relationship between two variables. The coefficients range from -1 to +1, with -1 indicating a perfect negative correlation, +1 indicating a perfect positive correlation, and 0 indicating no correlation. A correlation matrix shows the correlation coefficients between each pair of variables in a dataset, with each variable appearing both as a row and a column.



INFERENCE

The variable considered here are Life expectancy, Adult mortality, Infant deaths, Alcohol consumption, percentage expenditure, under five deaths, total expenditure, HIV/AIDS,GDP, population, income composition of resources and schooling.

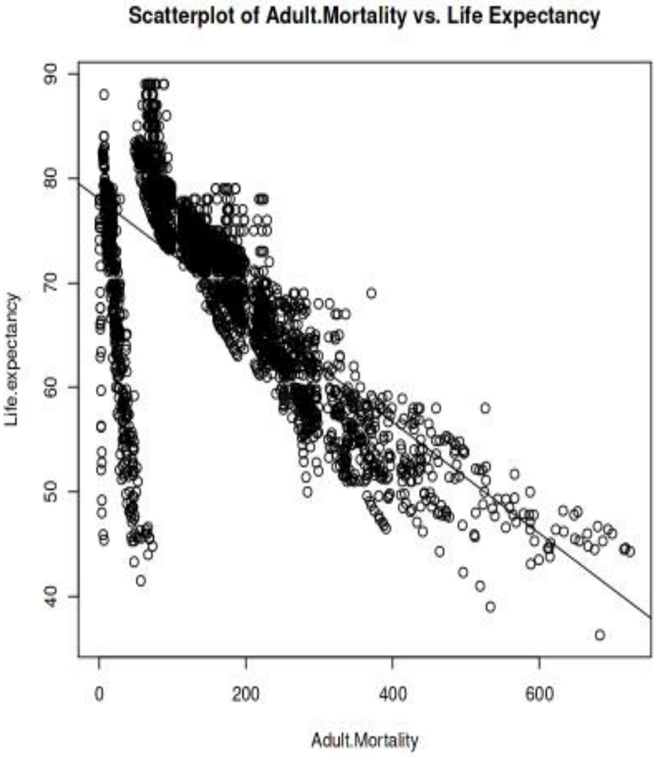
Life Expectancy has a high degree of negative correlation (-0.7) with Adult Mortality , suggesting that as Adult Mortality increases , Life Expectancy decreases.

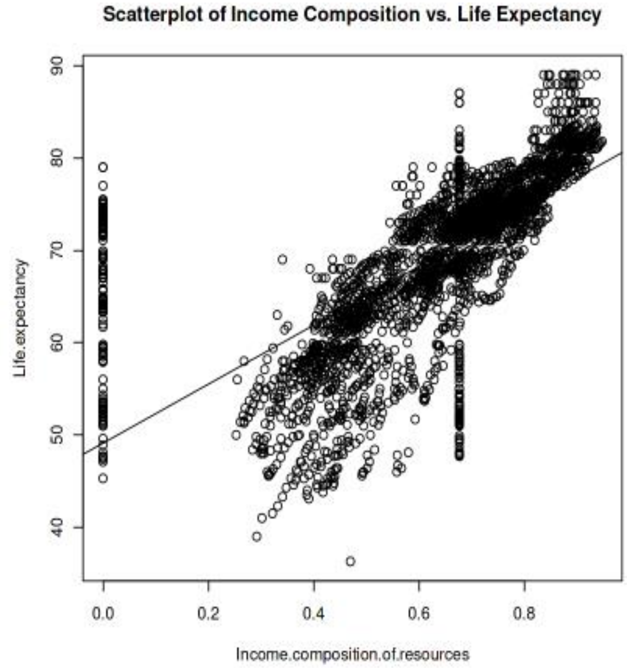
The analysis also reveals a strong positive correlation (0.71) between Life Expectancy and Schooling indicating that as the levels of schooling increases there is a corresponding increase in the Life Expectancy.

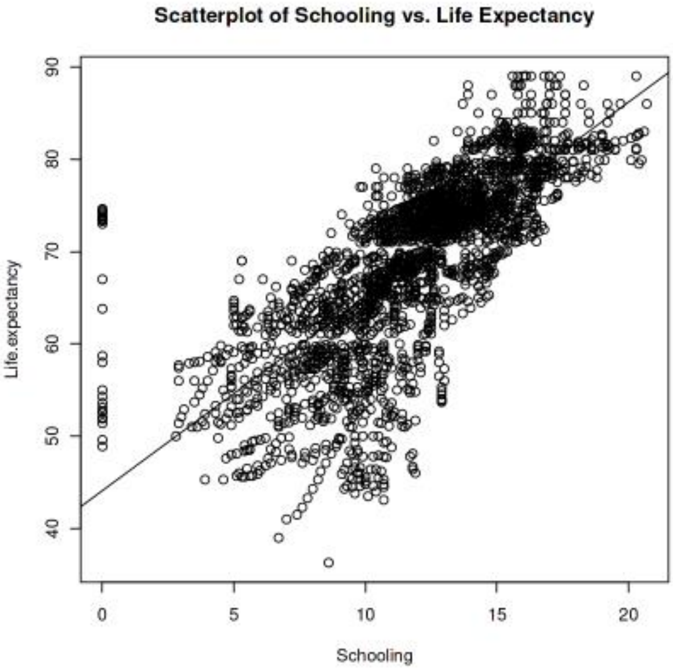
There is also a high degree of positive correlation (0.69) between Income Composition of Resources and Life Expectancy. So as Income Composition of Resources increases, Life Expectancy also increases.

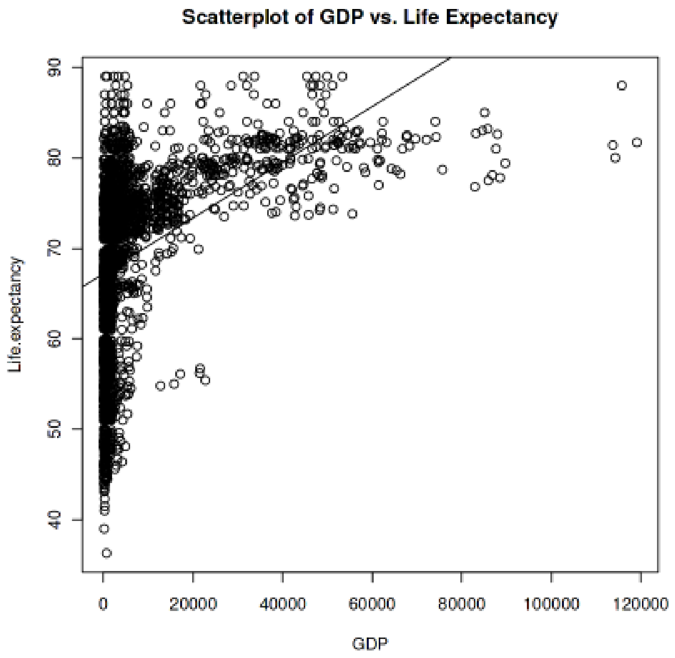
There is a weak negative correlation observed between infant deaths and under-five deaths with life expectancy. Conversely, alcohol consumption, percentage expenditure, and GDP exhibit a moderate positive correlation with life expectancy. Moreover, a moderate negative correlation is found between HIV/AIDS and life expectancy. It is noteworthy that there is a weak positive correlation between total expenditure and life expectancy.

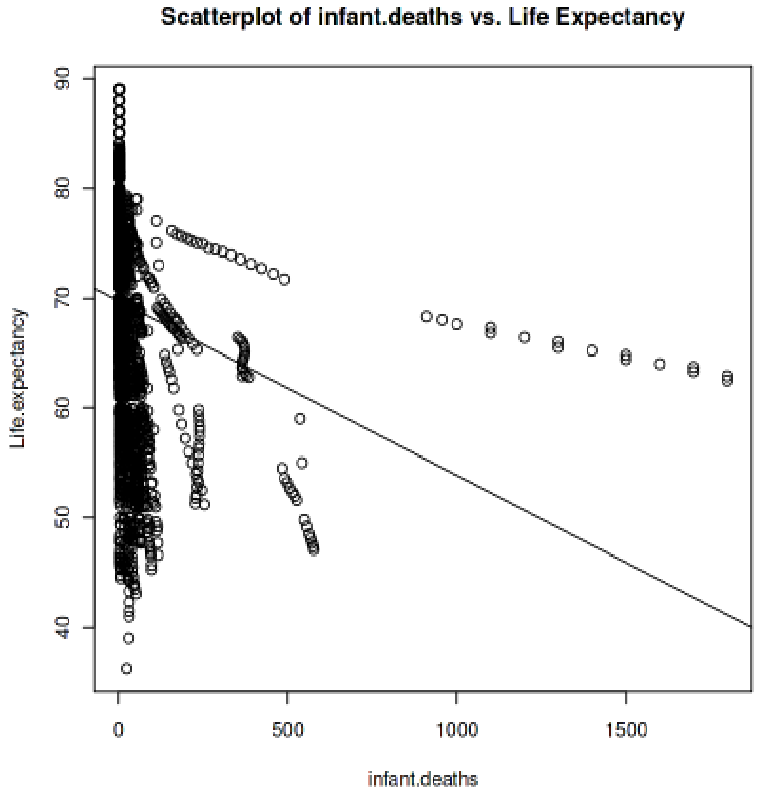
SCATTER PLOTS

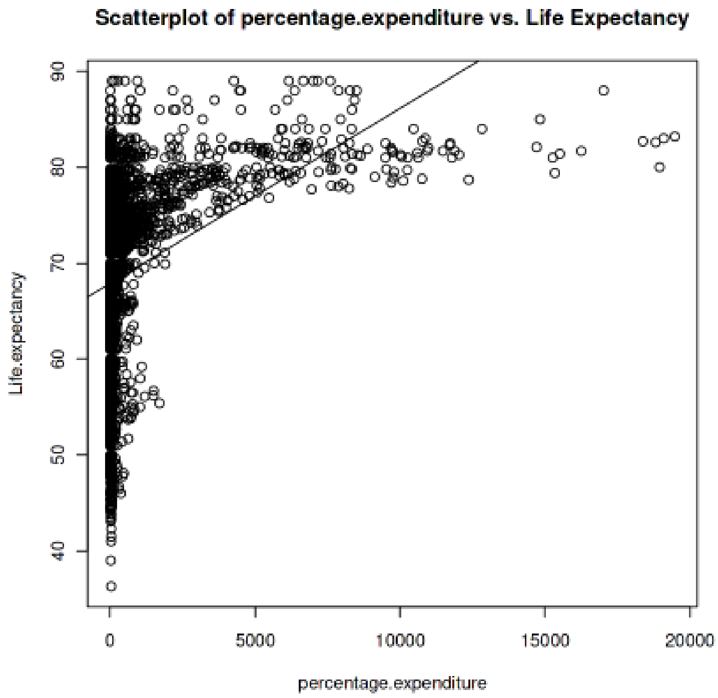


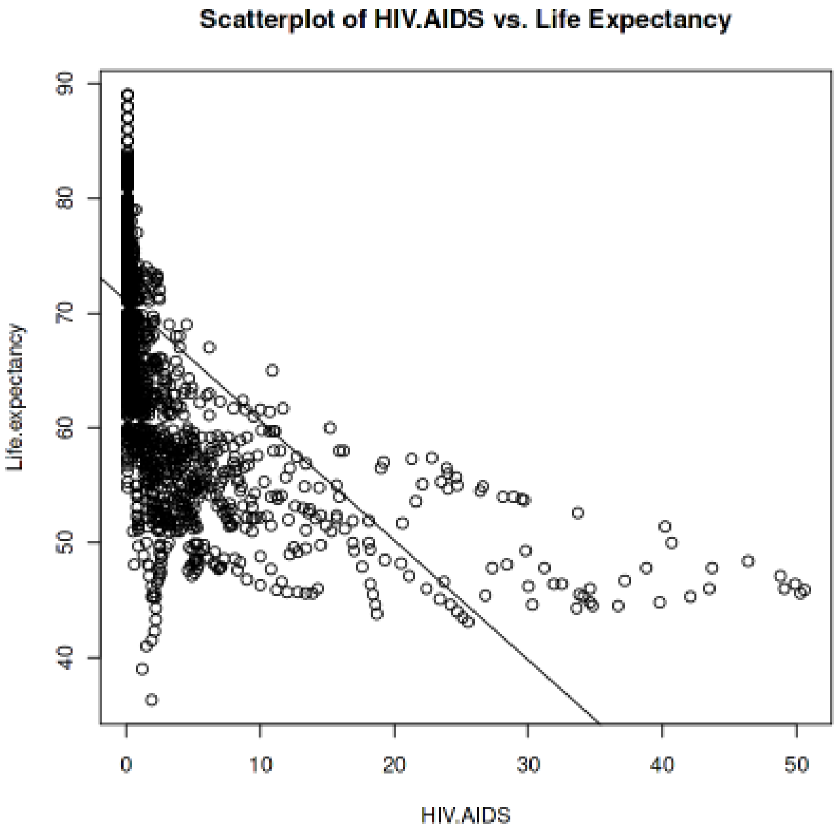

Scatterplot of Income Composition vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0.0
0.2
0.4
0.6
0.8
Income.composition.of.resources

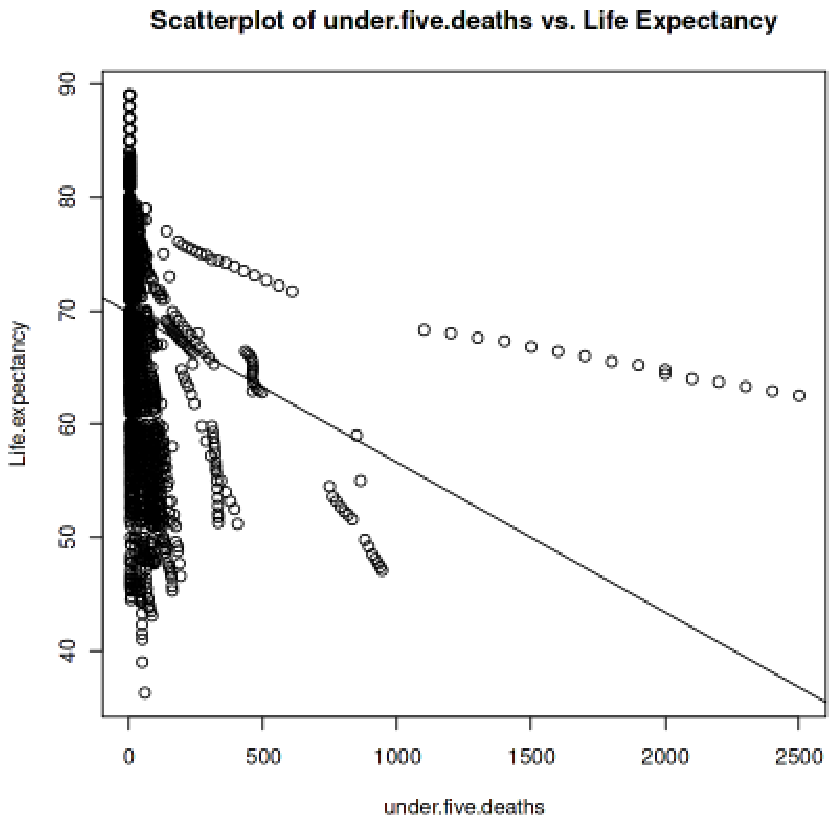

Scatterplot of Schooling vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
5
10
15
20
Schooling

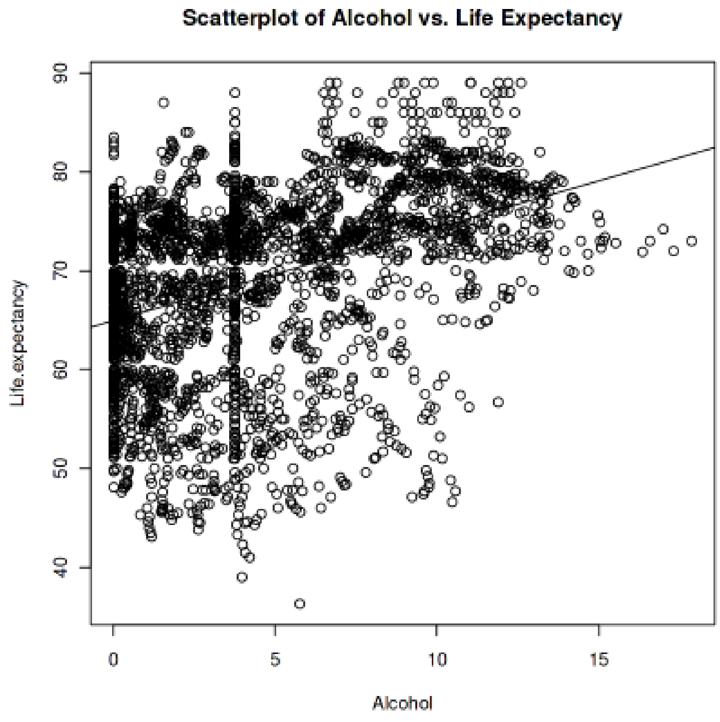

Scatterplot of GDP vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
20000
40000
60000
80000
100000
120000
GDP

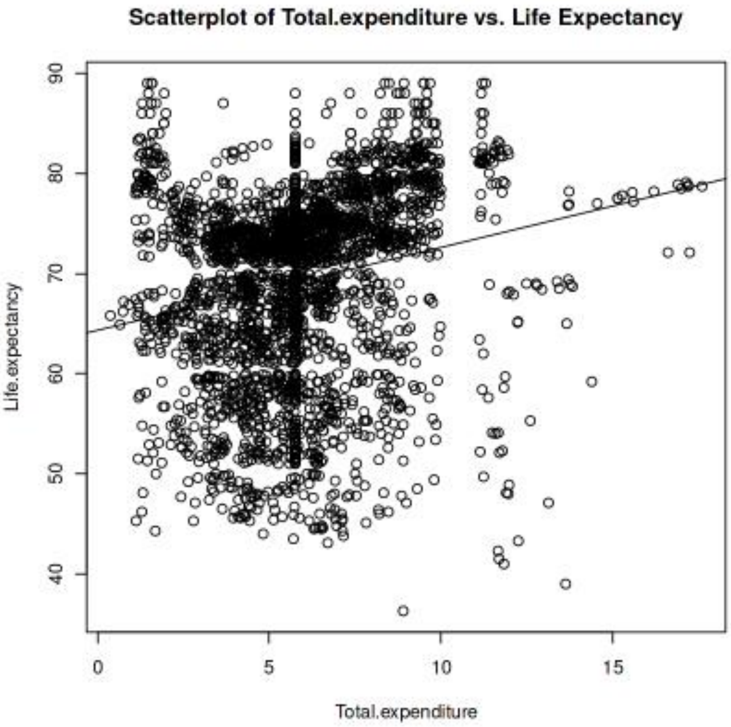

Scatterplot of infant.deaths vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
500
1000
1500
infant.deaths


Scatterplot of percentage.expenditure vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
5000
10000
15000
20000
percentage.expenditure


Scatterplot of HIV.AIDS vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
10
20
30
40
50
HIV.AIDS


Scatterplot of under.five.deaths vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
500
1000
1500
2000
2500
under.five.deaths


Scatterplot of Alcohol vs. Life Expectancy
Life.expectancy
40
50
60
70
80
90
0
5
10
15
Alcohol


Scatterplot of Total.expenditure vs. Life Expectancy
Life.expectancy
Total.expenditure
90
80
70
60
50
40
0
5
10
15

INFERENCE

SCATTER PLOT OF ADULT MORTALITY VS LIFE EXPECTANCY.

The scatter plot forms a tight downward sloping cluster , it suggests a negative association between life expectancy and adult mortality. It indicates that adult mortality rates serve as an informative indicator of life expectancy.

SCATTER PLOT OF INCOME COMPOSITION OF RESOURCES VS LIFE EXPECTANCY.

The scatter plot exhibits an upward trend, it suggests a positive association between life expectancy and income composition of resources. It indicates that income composition of resources serves as an informative indicator of life expectancy.

SCATTER PLOT OF SCHOOLING VS LIFE EXPECTANCY.

The scatter plot exhibits an upward trend, it suggests a positive association between life expectancy and schooling. It indicates that schooling serves as an informative indicator of life expectancy.

ANOVA (Analysis Of Variances)

NORMALITY TEST

A Normality test is a statistical test used to determine if a given dataset follows a normal distribution .The Shapiro-Wilk test is a statistical test used to determine whether a given dataset follows a normal distribution. It is one of the most widely used tests for assessing normality. The test is based on the null hypothesis that the data is drawn from a normally distributed population .The Shapiro-Wilk test works by calculating a test statistic (W) that measures the correlation between the observed data and the expected values under the assumption of normality. The test statistic ranges from 0 to 1, where a value closer to 1 indicates a closer fit to a normal distribution.

We started with a population dataset, and from that, we took a sample of 30 countries for each of the variables: Schooling, Adult Mortality, and Income Composition of Resources. To assess the normality of these variables, we performed the Shapiro-Wilk test, yielding the following W statistics:

- Schooling, the W statistic is 0.97103.
- Adult Mortality, the W statistic is 0.91892.
- Income Composition of Resources, the W statistic is 0.89787.

These W statistics provide insights into how closely the distributions of the variables match a normal distribution.

OBJECTIVE

To check whether the variables such as Income, Schooling, & Adult Mortality, associated with Life Expectancy are significant or not.

ANOVA

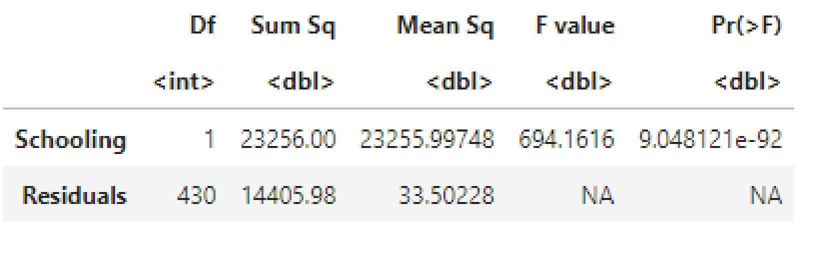
ANOVA stands for Analysis of Variance, which is a statistical method used to test for significant differences between means of three or more groups. ANOVA is used when comparing the means of multiple groups to determine if there is a significant difference between them. It measures the variability within groups and compares it to the variability between groups.

There are several types of ANOVA, including one-way ANOVA, two-way ANOVA, and repeated-measures ANOVA. One-way ANOVA is used when there is one independent variable, while two-way ANOVA is used when there are two independent variables. Repeated-measures ANOVA is used when the same group of participants is tested multiple times under different conditions.

ANOVA produces an F-statistic, which is used to determine whether the means of the groups are significantly different. If the F-value is greater than the critical value, then the null hypothesis is rejected, indicating that there is a significant difference between the means of the groups. ANOVA is commonly used in social science, medical research, and other fields where multiple groups need to be compared ,

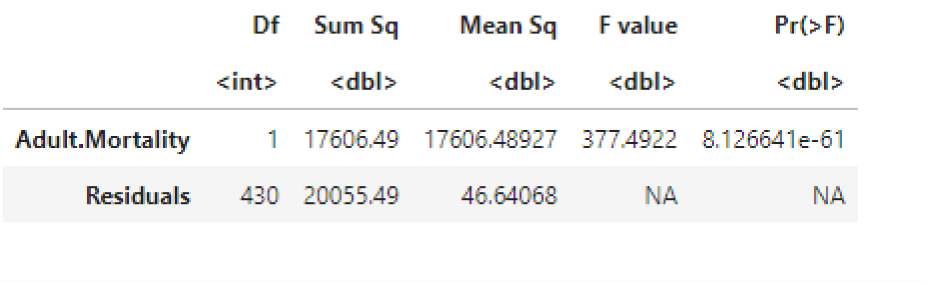
PERFORMING ANOVA

- SCHOOLING



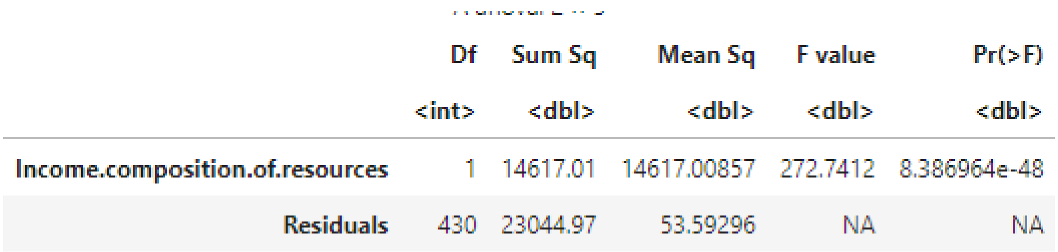
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Schooling	1	23256.00	23255.99748	694.1616	9.048121e-92
Residuals	430	14405.98	33.50228	NA	NA

- ADULT MORTALITY



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Adult.Mortality	1	17606.49	17606.48927	377.4922	8.126641e-61
Residuals	430	20055.49	46.64068	NA	NA

- INCOME COMPOSITION OF RESOURCES



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Income.composition.of.resources	1	14617.01	14617.00857	272.7412	8.386964e-48
Residuals	430	23044.97	53.59296	NA	NA

INFERENCE

The p-values obtained are less than 0.05,indicating that the results are statistically significant and we reject the null hypothesis (The variables associated are not significant). Hence accepting the alternative hypothesis (The variables associated are significant).

The variables such as Income, Schooling, & Adult Mortality, associated with Life Expectancy are significant.

REGRESSION ANALYSIS

OBJECTIVE

To fit a regression model using the variables Life Expectancy, Income, Schooling, & Adult Mortality.

MULTIPLE REGRESSION

Multiple regression is a statistical analysis technique used to examine the relationship between a dependent variable and two or more independent variables. It aims to model and predict the value of the dependent variable based on the values of the independent variables.

The goal of multiple regression is to determine the extent to which the independent variables collectively contribute to explaining or predicting the variation in the dependent variable.

The multiple regression model is based on some assumptions such as;

- There is a linear relationship between the dependent variable and the independent variables.
- The independent variables are not too highly correlated with each other.
- The observations used in the regression analysis should be independent of each other.
- The residuals should be normally distributed.

The general formula for multiple regression is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + \varepsilon$$

Where Y is the dependent variable, X_1, X_2, X_3, …, X_k are independent variables, β0 is the intercept or constant term, β_1, β_2, β_3, …, β_k are the coefficients, and ε is the error term.

The objective of multiple regression is to estimate the coefficients (β) that best fit the data so that we can make predictions about the dependent variable (Y) based on the values of the independent variables (X_1, X_2, X_3, …, X_k). The coefficients (β) represent the change in Y for a one-unit change in X_1, X_2, X_3, …, X_k, while holding all other independent variables constant.

Multiple regression is a more complex form of linear regression, which only has one independent variable. The addition of multiple independent variables allows us to explore more complex relationships between the variables and the dependent variable. However, it also makes the analysis more difficult, as there are more variables to consider and the potential for multicollinearity (where the independent variables are highly correlated with each other).

To perform multiple regression, we typically use a software package such as R or Python. The analysis involves several steps, including data cleaning, exploratory data analysis, model fitting, and model evaluation. We need to ensure that the assumptions of multiple regression are met, such as the normality of residuals and linearity of the relationship between the independent variables and the dependent variable.

The multiple regression analysis provides estimates of the regression coefficients, which represent the direction and magnitude of the relationship between the dependent variable and each independent variable. It also provides information on the overall fit of the model, such as the coefficient of determination (R-squared), which indicates the proportion of variance in the dependent variable explained by the independent variables.

However, it requires careful analysis and interpretation, and should only be used when the assumptions of the technique are met . VIF stands for Variance Inflation Factor. It is a statistical measure used to assess multicollinearity in regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other. This high correlation can pose problems in the regression analysis, such as unstable or unreliable estimates of the regression coefficients. The VIF is calculated for each predictor variable and provides a numerical value that indicates how much the variance of the estimated regression coefficient for that variable is inflated due to multicollinearity. It measures the impact of multicollinearity on the precision of the coefficient estimates.

The equation for calculating VIF is ,

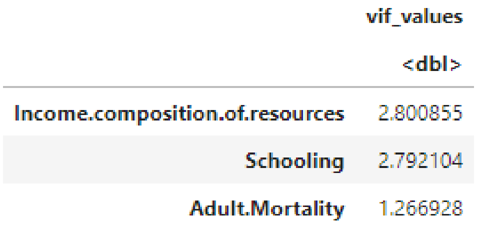
VIF=1/(1-R^2)

- VIF less than 1: A VIF value less than 1 suggests that there is no multicollinearity between the independent variables. This situation is highly unlikely and can indicate an error or an issue with the data.
- VIF around 1: A VIF value around 1 indicates that there is minimal or no multicollinearity between the independent variables. This is generally considered desirable, as it means the variables are not highly correlated.
- VIF between 1 and 5: VIF values between 1 and 5 are generally considered acceptable and indicate a moderate level of multicollinearity. While some correlation between variables exists, it is not severe enough to cause significant issues with the regression analysis.

- VIF greater than 5: VIF values above 5 suggest a high degree of multicollinearity between the independent variables. This indicates a strong correlation, which can lead to unstable coefficient estimates and challenges in interpreting the effects of individual variables.

CALCULATION OF VIF

The Variance Inflation Factor (VIF) was calculated for the independent variables, namely income composition of resources, schooling, and adult mortality. The VIF values for these variables are as follows:



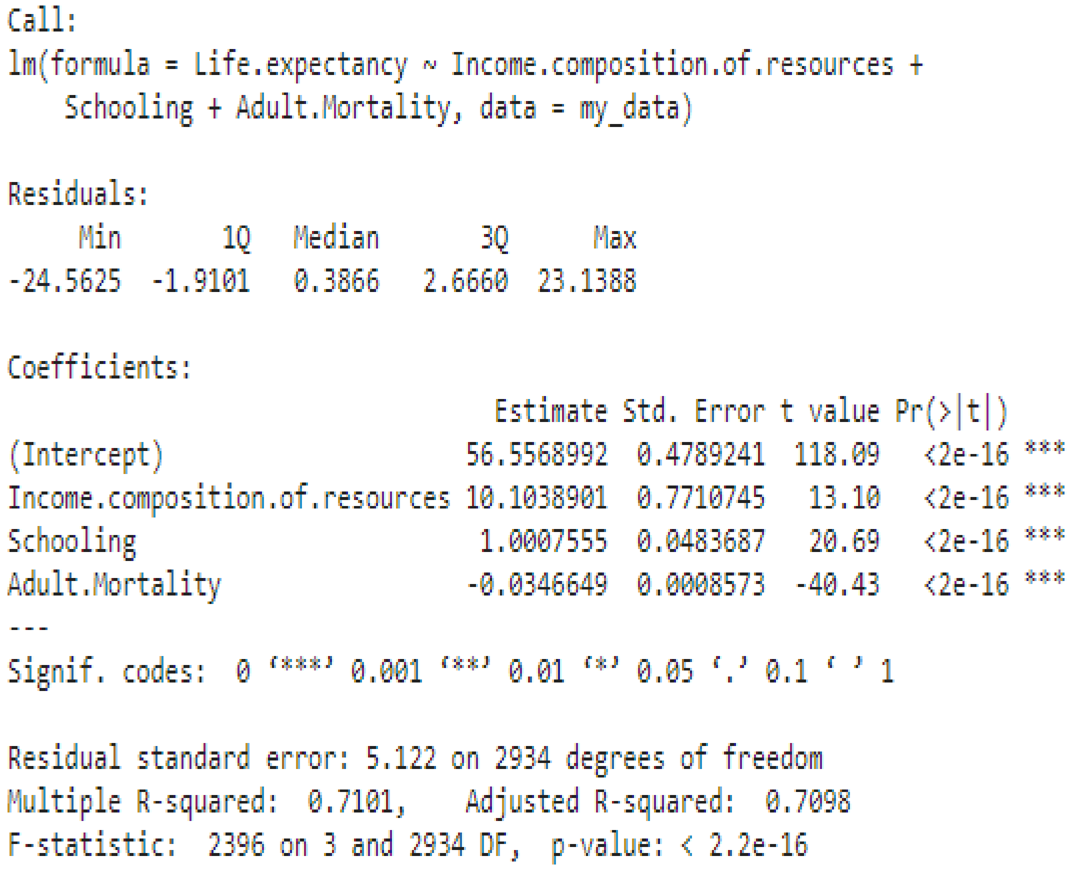
	vif_values <dbl>
Income.composition.of.resources	2.800855
Schooling	2.792104
Adult.Mortality	1.266928

- Income composition of resources: VIF = 2.800855
- Schooling: VIF = 2.792104
- Adult Mortality: VIF = 1.266928

These VIF values indicate the extent of multicollinearity between each independent variable and the other independent variables in the multilinear regression model.

In this case, all three variables have relatively low VIF values. Income composition of resources and Schooling have VIF values around 2.8, indicating some degree of multicollinearity but not severe. Adult Mortality has a lower VIF value of 1.266928, suggesting even less correlation with the other independent variables.

FITTING THE MODEL



```
Call:
lm(formula = Life.expectancy ~ Income.composition.of.resources +
    Schooling + Adult.Mortality, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-24.5625  -1.9101   0.3866   2.6660  23.1388

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     56.5568992  0.4789241  118.09   <2e-16 ***
Income.composition.of.resources 10.1038901  0.7710745   13.10   <2e-16 ***
Schooling                        1.0007555  0.0483687   20.69   <2e-16 ***
Adult.Mortality                 -0.0346649  0.0008573  -40.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.122 on 2934 degrees of freedom
Multiple R-squared:  0.7101,	Adjusted R-squared:  0.7098 
F-statistic:  2396 on 3 and 2934 DF,  p-value: < 2.2e-16
```

INFERENCE

The p-value of 2.2e-16 indicates that the regression model is statistically significant, meaning that there is a relationship between the independent variables (income consumption of resources, schooling, and adult mortality) and the dependent variable (life expectancy).

The R-squared value of 0.7101 indicates that approximately 71% of the variation in life expectancy can be explained by the independent variables included in the model. This is a relatively high R-squared value, suggesting that the model is a good fit for the data.

Therefore, based on the p-value and R-squared value, we can infer that the independent variables (Income Composition of Resources, Schooling and Adult Mortality) are important predictors of life expectancy and that the regression model has a strong ability to explain the variation in life expectancy based on these variables.

K MEANS CLUSTERING

OBJECTIVE

To cluster the countries based on the variables Life expectancy , Adult Mortality, schooling and income composition of resources of 2015 only.

K MEANS CLUSTERING

K-means clustering is a popular algorithm used to group or cluster a dataset based on similarities in their features. It aims to find K clusters, where K is a pre-defined number set by the user. The algorithm works by iteratively reassigning data points to the nearest cluster centroid and recalculating the centroid until convergence.

Rather than defining groups before looking at the data, clustering allows you to find and analyse the groups that have formed organically. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithms start with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two

1. Data assignment step:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance.

2. Centroid update step:

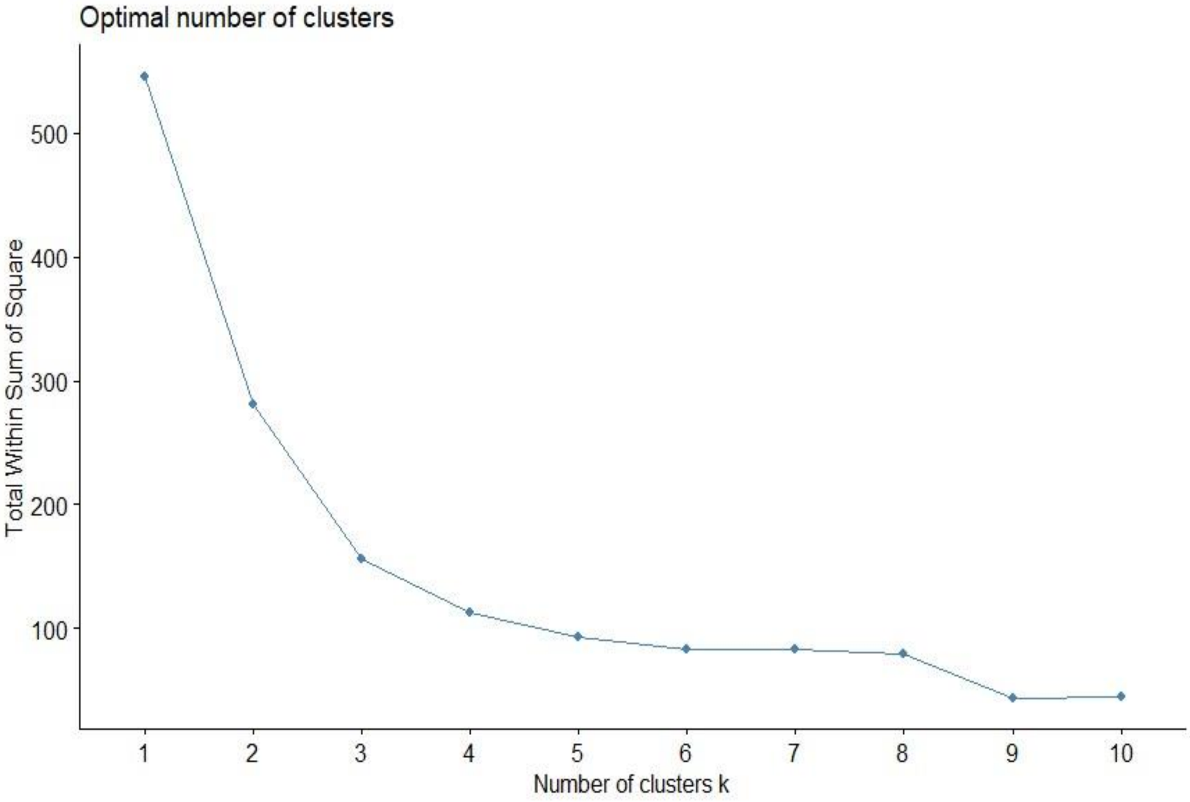
In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between steps one and two until a stopping criterion is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

The algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e., not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome

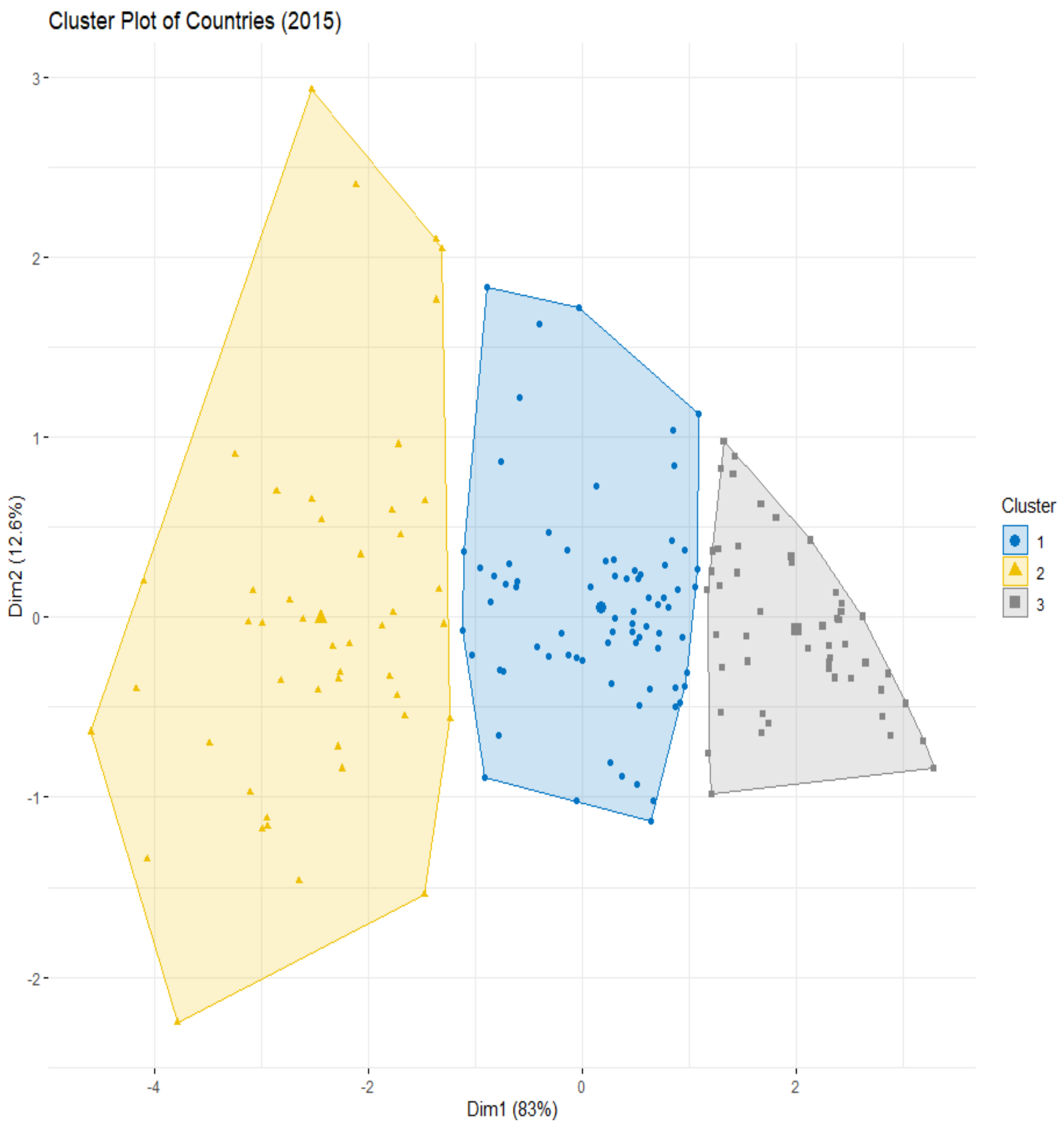
To find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of K values and compare the results. One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K.

SCREEPLOT



The optimal number of clusters that can be formed is 3.

CLUSTER PLOT



INFERENCE

The resulting clusters are then visualized using a scatter plot with data points coloured by cluster membership

The three clusters identified by k-means algorithm appear to be relatively well separated, with limited overlap between them.

CLUSTER 1 (BLUE DOTS)

The countries belonging to cluster 1 are developing and underdeveloped countries , which are mostly Asian countries (Bangladesh, Bhutan, Thailand, Brazil, Mexico, Jordan, Oman, Tongo etc…). They are characterised by moderate values of life expectancy, schooling, income composition of resources and adult mortality.

CLUSTER 2 (YELLOW DOTS)

The countries belonging to cluster 2 are predominantly underdeveloped which are mostly Asian and African countries (Afghanistan ,Chad, Congo, Togo, Yemen, Sudan , Uganda, Ghana etc…).They are characterised by lower values of life expectancy, schooling, income composition of resources and higher values of adult mortality.

CLUSTER 3 (ASH DOTS)

The countries belonging to cluster 3 are developed countries , which are mostly European countries(Ireland, Italy, Netherlands, Norway, Spain, Germany, Austria, Belgium, Singapore, Canada, France etc…). They are characterised by higher values of life expectancy, schooling, income composition of resources and lower values of adult mortality.

CONCLUSION

Based on our study of the Life Expectancies of 183 countries between 2000 and 2015, the following are the conclusions:

- The Exploratory Analysis on the data was performed and it was observed that the trend of Life Expectancy had an overall positive increase from 2000 to 2015, and the data was negatively skewed. The median life expectancy appeared to have increased over time. The IQR and dispersion seemed to be decreasing.
- The correlation matrix revealed that life expectancy had a high degree of negative correlation (-0.7)with Adult Mortality and a high degree of positive correlation with Schooling (0.71) and Income Composition of Resources (0.69). There is a weak negative correlation observed between infant deaths and under-five deaths with life expectancy. Conversely, alcohol consumption, percentage expenditure, and GDP exhibit a moderate positive correlation with life expectancy. Moreover, a moderate negative correlation is found between HIV/AIDS and life expectancy. It is noteworthy that there is a weak positive correlation between total expenditure and life expectancy.
- Our analysis of the scatter plots revealed that life expectancy is negatively associated with adult mortality, while being positively associated with schooling and income composition of resources, suggesting that these factors may be important predictors of life expectancy and could be used as informative indicators in public health policy and future research.
- The ANOVA results suggested that Income Composition of Resources, Schooling, and Adult Mortality were significant variables associated with Life Expectancy.
- Furthermore, Multiple Regression analysis demonstrated that the independent variables (Income Composition of Resources, Schooling, and Adult Mortality) were important predictors of Life Expectancy, and the Regression Model had a strong ability to explain the variation in Life Expectancy. The R-squared value of 0.7101 indicates that approximately 71% of the variation in life expectancy can be explained by the independent variables included in the model.
- By K-means Clustering we came to understand that out of the three clusters formed, Cluster 3 is mainly comprised of Developed European countries ,showing higher values of Life Expectancy, Income Composition of Resources, and Schooling, and lower values of Adult Mortality. In contrast, Cluster 2 mainly comprised of Underdeveloped Asian and African countries showing lower values of Life Expectancy, Income Composition of Resources, and Schooling, and higher values of Adult Mortality. India belongs to Cluster 2. Cluster 1, contains Developing and Underdeveloped countries with moderate values of life expectancy, income composition of resources and schooling.
- In conclusion, this analysis provides valuable insights into the factors that contribute to differences in Life Expectancy among countries, and can guide policy decisions aimed at improving life expectancy globally.

APPENDIX

- **IMPORTING THE DATASET**

```
mydata <- read.csv("/kaggle/input/life-expectancy-who/Life Expectancy Data.csv")
```

- **BOX PLOT OF LIFE EXPECTANCY THROUGH YEARS**

```
#ploting a box plot

ggplot(my_data, aes(x = Year, y = Life. expectancy, group = Year)) +

  geom_boxplot() +

  labs(x = "Year", y = "Life Expectancy") +

  ggtitle("Variation in Life Expectancy through Years")
```

- **CORRELATION MATRIX**

```
data <- my_data[, !(names(my_data) %in% c("Country","Status", "Year"))]

cor_matrix <- cor(data, use = "pairwise.complete.obs")

corrplot(cor_matrix, method = "square", type = "lower", tl.col="black", tl.srt = 45,
title="CORRELATION MATRIX")
```

- **SCATTER PLOTS**

```
# Creating a scatterplot

plot(mydata$Income.composition.of.resources, mydata$Life.expectancy,

xlab = "Income.composition.of.resources", ylab = "Life.expectancy",

main = "Scatterplot of Income Composition vs. Life Expectancy")

# Adding a regression line to the scatterplot

abline(lm(mydata$ Life.expectancy ~ mydata$Income.composition.of.resources))
```

```
# Creating a scatterplot
plot(mydata$Adult.Mortality, mydata$Life.expectancy,
xlab = "Adult.Mortality", ylab = "Life.expectancy",
main = "Scatterplot of Adult.Mortality vs. Life Expectancy")
# Adding a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$Adult.Mortality))

# Creating a scatterplot
plot(mydata$Schooling, mydata$Life.expectancy,
xlab = "Schooling", ylab = "Life.expectancy",
main = "Scatterplot of Schooling vs. Life Expectancy")
# Adding a regression line to the scatterplot
abline(lm(mydata$Life.expectancy
~ mydata$Schooling))

# Create a scatterplot
plot(mydata$infant.deaths, mydata$Life.expectancy,
 xlab = "infant.deaths", ylab = "Life.expectancy",
main = "Scatterplot of infant.deaths vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$infant.deaths))
```

```
# Create a scatterplot
plot(mydata$Alcohol, mydata$Life.expectancy,
 xlab = "Alcohol", ylab = "Life.expectancy",
 main = "Scatterplot of Alcohol vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$Alcohol))

# Create a scatterplot
plot(mydata$percentage.expenditure, mydata$Life.expectancy,
xlab = "percentage.expenditure", ylab = "Life.expectancy",
 main = "Scatterplot of percentage.expenditure vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$percentage.expenditure))

# Create a scatterplot
plot(mydata$under.five.deaths, mydata$Life.expectancy,
xlab = "under.five.deaths", ylab = "Life.expectancy",
 main = "Scatterplot of under.five.deaths vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$under.five.deaths))
```

```
# Create a scatterplot
plot(mydata$Total.expenditure, mydata$Life.expectancy,
 xlab = "Total.expenditure", ylab = "Life.expectancy",
main = "Scatterplot of Total.expenditure vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$Total.expenditure))


# Create a scatterplot
plot(mydata$HIV.AIDS, mydata$Life.expectancy,
xlab = "HIV.AIDS", ylab = "Life.expectancy",
main = "Scatterplot of HIV.AIDS vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$HIV.AIDS))


# Create a scatterplot
plot(mydata$GDP, mydata$Life.expectancy,
xlab = "GDP", ylab = "Life.expectancy",
main = "Scatterplot of GDP vs. Life Expectancy")
# Add a regression line to the scatterplot
abline(lm(mydata$Life.expectancy ~ mydata$GDP))
```

GENERATING THE SAMPLE

```
# Load the population dataset

population_data <- my_data

# Determine the desired sample size

sample_size <- 30

# Randomly select countries

random_countries <- sample(population_data$Country, size = sample_size)

# Extract life expectancy data for the selected countries

sample_data <- population_data %>% filter(Country %in% random_countries)

# View the sample dataset

View(sample_data)
```

NORMALITY TEST

```
# Perform Shapiro-Wilk test for normality

shapiro_test <- shapiro.test(sample_data$Adult.Mortality)

# Print the test results

print(shapiro_test)


# Perform Shapiro-Wilk test for normality

shapiro_test <- shapiro.test(sample_data$Schooling)

# Print the test results

print(shapiro_test)

# Perform Shapiro-Wilk test for normality
```

```
shapiro_test <- shapiro.test(sample_data$Income.composition.of.resources)

# Print the test results

print(shapiro_test)
```

- **ANOVA**

```
#Checking the significance between Life.expectancy,Income.composition

model <- lm(Life.expectancy ~ Income.composition.of.resources ,data = sample_data)

anova(model)

#Checking the significance between Life.expectancy,Schooling

model <- lm(Life.expectancy ~ Schooling, data = sample_data)

anova(model)


#Checking the significance between Life.expectancy,Adult.Mortality

model <- lm(Life.expectancy ~Adult.Mortality, data = sample_data)

anova(model)
```

- **CALCULATING VIF VALUES**

```
# Fit your multilinear regression model

model <- lm(Life.expectancy ~ Income.composition.of.resources + Schooling +
Adult.Mortality, data = my_data)

# Calculate VIF values for the independent variables

vif_values <- vif(model)

# Print the VIF values

print(vif_values)

data.frame(vif_values)
```

MULTIPLE REGRESSION

```
# Fitting the multiple regression model

model <- lm(Life.expectancy ~ Income.composition.of.resources + Schooling + Adult.Mortality, data = my_data)

summary(model)
```

K MEANS CLUSTERING

```
# Load the necessary libraries

library(dplyr) # For data manipulation

library(ggplot2) # For data visualization

install.packages("factoextra")

library(factoextra)

# Read in the dataset

life_exp <- read.csv("C:/Users/HP/Downloads/Life Expectancy Data.csv")

names(life_exp)

life_exp

# Filter the data for the variables of interest (life expectancy, infant deaths, income composition of resources) for the year 2015

life_exp_2015 <- life_exp %>%

 filter(Year == 2015) %>%

  select(Country, Life.expectancy, Schooling, Income.composition.of.resources, Adult.Mortality)

# Remove rows with missing or infinite values

life_exp_2015 <- na.omit(life_exp_2015)

# Standardize the variables
```

```
life_exp_2015_std <- scale(life_exp_2015[,2:5])

library(factoextra)

# Compute the cluster using the k-means algorithm

set.seed(123)

k <- 3 # number of clusters

life_exp_2015_cluster <- kmeans(life_exp_2015_std, centers = k, nstart = 25)

# Visualize the cluster plot

fviz_cluster(life_exp_2015_cluster, data = life_exp_2015_std,

  geom = "point",

  palette = "jco",

   main = "Cluster Plot of Countries (2015)",

   ggtheme = theme_minimal(),

   stand = FALSE,

    legend.title = "Cluster")

# Add cluster assignments to the original data frame

life_exp_2015_clustered <- life_exp_2015 %>%

 mutate(cluster = life_exp_2015_cluster$cluster)

# Get a list of the countries in each cluster

countries_by_cluster <- life_exp_2015_clustered %>%

 group_by(cluster) %>%

  summarise(countries = paste(Country, collapse = ", "))

countries_by_cluster

# Add the cluster assignment as a new column to the life_exp_2015 data frame

life_exp_2015$cluster <- life_exp_2015_cluster$cluster

# Split the data frame by cluster
```

```
life_exp_2015_by_cluster <- split(life_exp_2015, life_exp_2015$cluster)
# Get the names of the countries in each cluster
country_names_by_cluster <- lapply(life_exp_2015_by_cluster, function(df) df$Country)
# Get a summary of the clusters
summary(life_exp_2015_clustered)
library(dplyr)
# Group the data by cluster and calculate the summary statistics
life_exp_2015_clustered %>%
  group_by(cluster) %>%
  summarise(mean_life_exp = mean(Life.expectancy),
   median_life_exp = median(Life.expectancy),
   max_life_exp = max(Life.expectancy),
   min_life_exp = min(Life.expectancy))
# Group the data by cluster and calculate the summary statistics
life_exp_2015_clustered %>%
 group_by(cluster) %>%
  summarise(mean_sch = mean(Schooling),
   median_sch = median(Schooling),
   max_sch = max(Schooling),
   min_sch = min(Schooling))
# Group the data by cluster and calculate the summary statistics
life_exp_2015_clustered %>%
  group_by(cluster) %>%
  summarise(mean_inc = mean(Income.composition.of.resources),
   median_inc = median(Income.composition.of.resources),
   max_inc = max(Income.composition.of.resources),

   min_inc = min(Income.composition.of.resources))

# Group the data by cluster and calculate the summary statistics

life_exp_2015_clustered %>%

  group_by(cluster) %>%

  summarise(mean_adu = mean(Adult.Mortality),

   median_adu = median(Adult.Mortality),

   max_adu = max(Adult.Mortality),

   min_adu = min(Adult.Mortality))
```

BIBLIOGRAPHY

BOOKS REFFERED

- Chang, W. (2018). R Graphics Cookbook: Practical Recipes for Visualizing Data. O'Reilly Media, Inc.

- Gupta S.C and Kapoor V.K, (2021). Fundamentals of Mathematical Statistics, (11th Edition). Sultan Chand & Sons, New Delhi.

- Matloff, N. (2011). The Art of R Programming: A Tour of Statistical Software Design. No Starch Press.

- Miller, A. (2017). Review of R for Data Science: Import, Tidy, Transform, Visualize, and Model Data by Hadley Wickham and Garrett Grolemund.

- Nolan, D., & Lang, D. T. (2015) Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. CRC Press.

- Pandya Kiran, Joshi Prashant, Bulsari Smruti, (2018) Statistical Analysis in Simple Steps Using R. SAGE Publications.

ONLINE SOURCE

- Kaggle (Life e\Expectancy dataset from WHO) .
  https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

- Multiple Regression, How to perform and Interpretation of results.
  https://www.scribbr.com/statistics/multiple-linear-regression/

- K Means Clustering , Understanding it and How does it work?
  https://neptune.ai/blog/k-means-clustering