PROJECT REPORT ON

PREDICTING INFANT MORTALITY: A GLOBAL ANALYSIS OF 2023 DATA

K.Ruchitha	1051-22-508-001
K. Shravantika	1051-22-508-003
M.Ramya	1051-22-508-011
Madirai Nag	1051-22-508-014
V.Deepika	1051-22-508-017

Project submitted for the award of the degree of

MASTER OF SCIENCE

BY

OSMANIA UNIVERSITY, HYDERABAD - 500007

Aurora's Degree & PG College

Reaccredited by NAAC with 'B' grade

Department of Statistics, Chikkadpally - 500020

PREDICTING INFANT MORTALITY: A GLOBAL ANALYSIS OF 2023 DATA

This Project work dissertation is submitted to

Osmania University, as partial fulfillment for the completion of

IV semester of M. Sc. (Applied Statistics)

BY

K.Ruchitha	1051-22-508-001
K. Shravantika	1051-22-508-003
M.Ramya	1051-22-508-011
Madirai Nag	1051-22-508-014
V.Deepika	1051-22-508-017

Under the supervision of

Mrs. Rajya Lakshmi,

Department of Statistics, Aurora's Degree, and PG College

2022 – 2024

Date : 28/08/2024

DECLARATION

I hereby declare that the project work presented in this dissertation, entitled

"**PREDICTING INFANT MORTALITY: A GLOBAL ANALYSIS OF 2023 DATA**" has been carried out by us under the supervision of Mrs. Rajya Lakshmi, in the Department of Statistics, Aurora's Degree and PG College, Chikkadpally,Hyderabad – 500020. The work done is original and has not been submitted so far, in part or full, for any other degree or diploma to any University or institution.

K.Ruchitha	1051-22-508-001
K. Shravantika	1051-22-508-003
M.Ramya	1051-22-508-011
Madirai Nag	1051-22-508-014
V.Deepika	1051-22-508-017

Aurora's Degree & PG College

(Reaccredited by NAAC with 'B' grade)

CERTIFICATE

DATE:28/08/2024

This is to certify that the research work presented in this thesis, entitled "**PREDICTING INFANT MORTALITY: A GLOBAL ANALYSIS OF 2023 DATA** " has been carried out by

K.Ruchitha	1051-22-508-001
K. Shravantika	1051-22-508-003
M.Ramya	1051-22-508-011
Madirai Nag	1051-22-508-014
V.Deepika	1051-22-508-017

registered M. Sc. (Applied Statistics) students of this college. The work done is original and has not been submitted so far, in part or full, for any other degree or diploma to any University or institution.

Signature of head with office seal **Signature of the Supervisor**

Internal Examiner Signature with date:

External Examiner Signature with date:

ACKNOWLEDGMENT

I am thankful to my supervisor Mrs. Rajya Lakshmi, Assistant Professor, Statistics Department Aurora's Degree, and PG College, Chikkadpally, Hyderabad – 500020, for her guidance.

I would also thank United Network of Professioals(UNP) and their entire team, for guidance and support throughtout this project.

I would like to express my sincere thanks to all the Faculty, Department of Statistics, Aurora's Degree, and PG College for their encouragement throughout the period of our coursework.

K.Ruchitha	1051-22-508-001
K. Shravantika	1051-22-508-003
M.Ramya	1051-22-508-011
Madirai Nag	1051-22-508-014
V.Deepika	1051-22-508-017

Predicting Infant Mortality: A Global Analysis of 2023 Data

Abstract:

This study aims to investigate the complex interrelationships between socio-economic, health, and environmental indicators that influence infant mortality across various countries to identify key determinants of national development and well-being. Using a dataset from kaggle named Global macro dataset, that includes variables such as GDP, tax revenue, minimum wage, educational enrollment rates, healthcare expenditure, life expectancy, infant mortality, CO2 emissions, and forested area, this study includes statistical methods including correlation analysis, multiple regression, significance tests and evaluate various machine learning models at different train test split ratios in enhance better predictability..

The primary objective is to understand how economic factors impact health outcomes and to model the influence of education on economic growth on infant mortality using machine learning models.

The results and machine learning models provide valuable insights for policy makers, international organizations, and other stakeholders, offering evidence-based recommendations for enhancing national development strategies. This study helps in tackling the complex socio-economic and healthcare factors driving infant mortality. Which at-large goals like UN's Sustainable Development Goal 3 good health and well-being can be achieved.

INDEX

Chapter - 1

Background

1. BACKGROUND

In recent years, the global community has made significant efforts to reduce infant mortality, particularly through initiatives such as the United Nations Sustainable Development Goals (SDGs), one such goal to improve good health and well-being, which set ambitious targets for improving child health and reducing mortality rates. However, progress has been uneven, with many countries still grappling with high infant mortality rates due to a complex interplay of factors including poverty, inadequate healthcare infrastructure, poor maternal health, and environmental challenges.

As per the United Nations International Children's Emergency Fund (UNICEF) and World Health Organization (WHO) in 2022, approximately 2.3 million children died in their first month of life, which translates to about 6,300 neonatal deaths daily. The average global neonatal mortality rate was 17 deaths per 1,000 live births, a significant reduction from 37 deaths per 1,000 live births in 1990. However, the decline in neonatal mortality has been slower compared to post-neonatal under-five mortality rates, with 47% of all child deaths under the age of five occurring during the neonatal period in 2022.

Causes of Infant Mortality

The leading causes of neonatal deaths include:

- Premature birth
- Birth complications (e.g., birth asphyxia)
- Neonatal infections
- Congenital anomalies

These factors account for nearly 40% of deaths in children under five. Most neonatal deaths occur within the first week of life, with about 1 million newborns dying within the first 24 hours

Despite progress in reducing infant mortality rates globally, challenges persist, particularly in vulnerable regions. Addressing the underlying causes of neonatal deaths and improving healthcare access are critical for further reductions in infant mortality. Continued monitoring and targeted interventions are essential to meet the Sustainable Development Goals (SDGs) for child health by 2030

Chapter - 2

Introduction

2. INTRODUCTION

This study focuses on the study of various indicators of development across countries in the year 2023. Infant mortality, defined as the number of deaths of infants under one year of age per 1,000 live births, is a crucial indicator of a country's overall health and development status. It reflects the effectiveness of healthcare systems, the availability of maternal and child health services, and the socio-economic conditions that influence early life outcomes. In this study, we explore the interrelationships between infant mortality and a range of socio-economic, health, and environmental variables across different countries. By analyzing these relationships, we aim to identify key determinants that contribute to variations in infant mortality rates globally, thereby providing insights into the factors that most significantly impact national well-being and development. This investigation not only highlights the direct influences on infant survival but also underscores the broader socio-economic and environmental context that shapes health outcomes across populations.

Infant mortality has been identified as a key indicator of the general welfare in communities and nations, especially amongst low- and middle income countries (LMIC) who bear an immense share of injustice. The Infant Mortality Rate (IMR), an indicator that has been widely used to indicate health and the status of socio-economic development in a country, A high IMR is often an indicator of problems with health care access, maternal and child health in general (e.g. poor pregnancy interventions), and socioeconomic conditions influencing both the quality of community based medical services as well as secondary referral facilities such emergency obstetric care.

The ultimate goal is to contribute towards these world targets by providing new knowledge about factors associated with infant mortality and ways that can be changed to make a positive impact. In doing so we aim to contribute towards the achievement of SDG 3, improve infant mortality rates and promote overall health outcomes among newborns around the world. Last, this model does not only consider the immediate health issues but also supports both short- and long-term changes that promote greater equity in terms of overall quality-of-life for any community

2.1 OBJECTIVE

- To identify the key influential factors in determining Infant mortality using the multiple linear regression and various machine learning models.

- To identify the significant factors that influence infant mortality globally.

- To develop the most accurate machine learning model for predicting infant mortality on a global scale

2.2 HYPOTHESIS

- **H0:** The coefficient of the independent variable is equal to zero. This means that the independent variable has no effect on the dependent variable i.e Infant mortality.

  i.e $\beta_i = 0$

  H1: The coefficient of the independent variable is not equal to zero. This means that the independent variable has a statistically significant effect on the dependent variable i.e Infant mortality.

  i.e $\beta_i \neq 0$

2.3 LITERATURE REVIEW:

- **"Advanced Machine Learning Models for Predicting Infant Mortality in Brazil: A Comparative Analysis"**

  Authors: Leonardo Matsuno da Frota, Marcos Hasegawa, Paulo Jacinto

  Summary: This study explores infant mortality prediction in Brazil using advanced machine learning techniques. The researchers address the limitations of the Cox proportional hazards model, particularly in the presence of non-proportional hazards, by employing alternative methods such as Survival Support Vector Machines, Random Survival Forest, and Extreme Gradient Boosting. Analyzing a substantial dataset of 2.9 million observations from Brazil's Unique Health System (SUS), the study reports high model accuracy with concordance indices of 0.84, 0.83, and 0.81 for the machine learning models, compared to 0.83 for the Cox model. The SHAP (SHapley Additive exPlanations) framework was utilized to interpret the model results, revealing that factors such as cesarean sections and gestational weeks impact infant mortality in complex, nonlinear ways. The findings suggest that interpretable machine learning models can be valuable tools for policymakers in developing effective strategies to reduce infant mortality in middle-income countries.

- **Application of machine learning methods for predicting infant mortality in Rwanda: analysis of Rwanda demographic health survey 2014–15 dataset**

  Authors: Emmanuel Mfateneza, Pierre Claver Rutayisire, Emmanuel Biracyaza, Sanctus Musafiri & Willy Gasafari Mpabuka

  Published in: *BMC Pregnancy Childbirth* **22**

  Summary: In a cross-sectional study using the 2014–15 Rwanda Demographic and Health Survey, various predictive models were assessed using Python and STATA. Machine Learning methods, including Random Forest (RF), Decision Tree, Support Vector Machine (SVM), and Logistic Regression, were evaluated for their ability to predict Infant Mortality (IM). The RF model emerged as the most effective, achieving an accuracy of 84.3%, with high recall, precision, and F1 score. The Decision Tree model followed closely, while SVM and Logistic Regression showed lower performance. Key predictors of IM included marital status, number of children ever born, birth order, and wealth index. The study highlights that ML methods, particularly Random Forest, can provide valuable predictive insights beyond traditional statistical approaches.

- **“Multilevel log linear model to estimate the risk factors associated with infant mortality in Ethiopia: further analysis of 2016 EDHS”**

  Authors: Solomon Sisay Mulugeta, Mitiku Wale Muluneh, Alebachew Taye Belay, Yikeber Abebaw Moyehodie, Setegn Bayabil Agegn, Bezanesh Melese Masresha, and Selamawit Getachew Wassihun

  Published in: *National Library of Medicine, July 2002*

  Summary: The study provides valuable insights into the multifaceted factors contributing to infant mortality in Ethiopia. By employing a multilevel log-linear model, the researchers were able to delve into the complex interplay between individual, household, and regional factors. The findings highlight the disproportionate impact of infant mortality on certain regions, such as Somali, and underscore the urgent need for targeted interventions. The study's identification of key risk factors, including maternal age, birth weight, socioeconomic status, and regional disparities, offers crucial guidance for policymakers and healthcare providers. Addressing these factors through comprehensive strategies, such as improving access to prenatal care, promoting breastfeeding, and enhancing maternal and child health services, is essential for reducing infant mortality rates and improving the overall health and well-being of children in Ethiopia

- **"Trends, patterns and predictive factors of infant and child mortality in well-performing and underperforming states of India: a secondary analysis using National Family Health Surveys"**

  Authors: Mrigesh Bhatia, Laxmi Kant Dwivedi, Mukesh Ranjan, Priyanka Dixit, Venkata Putcha

  Published in: *BMJ Open, March 2019*

  Summary: The paper analyzes infant and child mortality trends in India from 1992 to 2016 using data from three rounds of cross-sectional household surveys. The study, which covers both high-performing states (Haryana, Kerala, Maharashtra, Punjab, Tamil Nadu) and poor-performing states (Bihar, Chhattisgarh, Madhya Pradesh, Uttar Pradesh), reveals a significant overall decline in mortality rates. However, this progress is unevenly distributed, with states like Uttar Pradesh, Bihar, and Madhya Pradesh showing persistent underperformance. Regression analysis identifies higher mortality risks among female infants and those with shorter birth intervals, particularly in poorer states. Despite the overall positive trend, the findings emphasize the need for targeted strategies to address

regional disparities and focus on reducing neonatal mortality to further improve child health outcomes in high-impact states.

- **"Infant and Child Mortality and its major determinants: a case study of Uttarakhand state, India"**

  Authors: Annu Chahal, Satya Parkash Kaushik

  Published in: R*esearchgate, 18 October 2022*

  Summary: This study investigates spatial variations in infant and child mortality rates in Uttarakhand, analyzing data from the Annual Health Survey (AHS) for 2010-11 to 2012-13 and the National Family Health Survey (NFHS) from 1998-99 to 2019-21. Using coefficients of variation and linear regression for data processing, the study finds a notable decline in Infant Mortality Rate (IMR) from 58 in 1998-99 to 39 in 2019-21, and a reduction in Under-5 Mortality Rate (U5MR) from 79 to 46 over the same period. Significant declines in mortality rates are observed in Rudraprayag and Almora districts. Conversely, Uttarkashi and Pithoragarh districts experienced increases in both IMR and U5MR. The findings highlight overall improvements in child mortality in Uttarakhand, alongside regional disparities that warrant further attention

- **"High infant and child mortality rates in Orissa: An assessment of major reasons"**

  Authors: Jalandhar Pradhan, Perianayagam Arokiasamy

  Published in: R*esearchgate*, 2018

  Summary: Infant and child mortality rates in Orissa are the highest among the Indian states. This is surprising, given other demographic indicators, such as the relatively rapid fertility decline and the quite high levels of antenatal care coverage in this state compared with other comparatively poor states. In this article, the macro- and micro-level determinants of high infant and child mortality are assessed using multivariate logistic regression analyses. The results demonstrate that the major contributors to the high infant and child mortality rates are the extremely low levels of health sector investments and the associated quality of care.

- **"Precision-weighted estimates of neonatal, post-neonatal and child mortality for 640 districts in India, National Family Health Survey 2016"**

  Authors: Rockli Kim, Lathan Liou, Yun Xu, Rakesh Kumar

  Published in: R*esearchgate*, 2020

  Summary: This study highlights the importance of disaggregating infant and under-five mortality rates into neonatal (<1 month), post-neonatal (1-11 months), and child (12-59 months) periods for more targeted interventions. Utilizing data from the 2015-2016 Indian National Family Health Survey with a sample of 259,627 children, the study applied a random effects model to predict district-specific mortality probabilities. Findings indicate that most under-five deaths occur in the neonatal period, with mortality rates varying significantly across districts. The study found moderate correlations between mortality estimates for different age groups and identified strong spatial clustering of high burden districts, often crossing state boundaries. This approach of precision-weighted estimates and detailed age-group analysis is recommended for better resource allocation and more effective child survival strategies.

- **"Predictive modeling of infant mortality"**

  Authors: Antonia Saravanou, Clemens Noelke, Nicholas Huntington, Dolores Acevedo-Garcia & Dimitrios Gunopulos

  Published in: *Springer link, January 2021*

  Summary: This paper investigates predictive models for Infant Mortality Rate (IMR), a critical metric that reflects both infant survival rates and overall societal health. Despite the high prosperity in the United States, the IMR remains relatively high compared to other developed countries and shows significant racial and ethnic disparities. The study employs both traditional machine learning techniques and advanced neural network models to analyze features from birth certificates, including socio-economic, ethnic, and medical data. It explores binary and multi-class classification approaches to predict whether an infant will survive past their first birthday and evaluates model performance across different population subsets, including by race. The paper highlights the importance of feature selection and examines various models and training set distributions, providing insights into improving predictive accuracy and understanding the factors influencing infant mortality.

- **"The Differential Effect of Foreign-Born Status on Low Birth Weight by Race/Ethnicity and Education"**

  Authors: Dolores Acevedo-Garcia, PhD, MPA-URP; Mah-J Soobader, PhD; Lisa F. Berkman, PhD

  Published in: *American Academy of Pediatrics, 2005*

  Summary: This article explores the impact of foreign-born status on low birth weight (LBW) and examines how this effect differs across racial/ethnic groups and educational levels. Using logistic regression analyses of the 1998 Detail Natality Data (n = 2,436,890), the study finds that foreign-born status does not significantly affect LBW risk among white women but increases the risk by 24% among Asian women. Conversely, it reduces LBW risk by approximately 25% among black women and 19% among Hispanic women. The protective effect of being foreign-born is more pronounced among women with lower educational attainment (0–11 years) compared to those with higher education levels. Additionally, the educational gradient in LBW is less distinct among foreign-born women of all racial/ethnic groups compared to their US-born counterparts. The study suggests that future research should investigate the underlying mechanisms of these variations, including health selection, cultural influences, and social support.

- **"Causes of death and infant mortality rates among full-term births in the United States between 2010 and 2012: An observational study"**

  Authors: Neha Bairoliya, Günther Fink

  Published in:*Plos journals, march 2018*

  Summary: This paper investigates the high mortality rates among full-term infants (born at 37–42 weeks of gestation) in the US, comparing them to those in European countries with low infant mortality rates. Analyzing linked birth and death records from 2010–2012, the study uses multivariable logistic and random effects models to examine state-specific variations in full-term infant mortality rates (FTIMR). The overall FTIMR was 2.2 per 1,000 live births, with significant state-level variations ranging from 1.29 in Connecticut to 3.77 in Mississippi. The study found that sudden unexpected death in infancy (SUDI), including sudden infant death syndrome (SIDS) and suffocation, was the leading cause of FTIM, contributing to 43% of cases, while congenital malformations and perinatal conditions contributed 31% and 11.3%, respectively. State-level disparities in FTIMR persisted even after adjusting for maternal education, race, and health. The paper

estimates that aligning all states with the best-performing states could prevent up to 4,003 infant deaths annually. Limitations include the lack of data on state-level termination rates and potential variations in death coding practices.

- **"Determinants of infant mortality in Pakistan: evidence from Pakistan Demographic and Health Survey 2017–18"**

  Authors: Kamalesh Kumar Patel, Rashmi Rai & Ambarish Kumar Rai
  Published in: *Springer Link, January 2020*
  Summary: This study assesses infant mortality in Pakistan using data from the 2018 Pakistan Demographic and Health Survey. Despite a decline from 86 to 62 deaths per 1,000 live births over the past three decades, Pakistan remains among countries with high infant mortality rates. Bivariate analyses and the Cox proportional hazard model were employed to examine risk factors for infant mortality, revealing significant disparities based on socioeconomic and demographic factors. The analysis found that mothers who did not use antenatal care (ANC) had a 1.5 times higher risk of infant mortality compared to those who did. Additionally, infants from wealthy households experienced 30% less mortality than those from poorer households. The study identifies rural-urban disparities in health services and gender inequities as key factors contributing to the persistence of high infant mortality rates in Pakistan..

2.4 ABOUT THE DATASET:

The dataset taken for our analysis is from Kaggle, an online repository of data.
Global Country Information Dataset 2023 (kaggle.com)
(https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023/code)
This comprehensive dataset provides a wealth of information exactlty **195 countries worldwide**, covering 27 features are columns in the dataset. It encompasses demographic statistics, economic indicators, environmental factors, healthcare metrics, education statistics, and much more. With every country represented, this dataset offers a complete global perspective on various aspects of nations.

The following variables present in our study are:

- Country : Name of the country.
- Density (P/Km²) : Population density measured in persons per square kilometer.
- Agricultural Land (%) : Percentage of total land area that is used for agriculture.
- Land Area (Km²) : Total land area of the country in square kilometers.
- Armed Forces Size : Number of personnel in the country's armed forces.
- Birth Rate : Number of live births per 1,000 people in a year.
- CO2 Emissions : Amount of carbon dioxide emissions in metric tons.
- CPI : Consumer Price Index, a measure of the average change in prices over time.
- CPI Change (%) : Percentage change in the Consumer Price Index over a specified period.
- Fertility Rate : Average number of children born to a woman during her lifetime.
- Forested Area (%) : Percentage of the total land area covered by forests.
- Gasoline Price : Price of gasoline per liter.
- GDP : Gross Domestic Product, total monetary value of all goods and services produced in a country.
- Gross Primary Education Enrollment (%) : Percentage of children of official primary school age enrolled in primary school.
- Gross Tertiary Education Enrollment (%) : Percentage of individuals of official tertiary education age enrolled in tertiary education.
- Infant Mortality : Number of deaths of infants under one year old per 1,000 live births.

- Life Expectancy : Average number of years a person is expected to live.
- Maternal Mortality Ratio : Number of maternal deaths per 100,000 live births.
- Minimum Wage : Lowest legal wage that can be paid to workers.
- Out of Pocket Health Expenditure : Percentage of health expenses paid directly by individuals rather than covered by insurance.
- Physicians per Thousand : Number of physicians per 1,000 people in the population.
- Population : Total number of people living in the country.
- Population: Labor Force Participation (%) : Percentage of the working-age population that is part of the labor force.
- Tax Revenue (%) : Government tax revenue as a percentage of GDP.
- Total Tax Rate : Total tax burden on businesses as a percentage of commercial profits.
- Unemployment Rate : Percentage of the labor force that is unemployed and actively seeking employment.
- Urban Population : Percentage of the total population living in urban areas.
- GDP per Capita : GDP divided by the total population, representing average economic output per person.

Chapter - 3

Methodology

3 METHODOLOGY:

1. Data Collection

- **Secondary Data**: Data sources for this project is from Kaggle which was curated from the databases like national health surveys, demographic health surveys (DHS), or databases maintained by organizations like the World Health Organization (WHO) or the United Nations (UN) for 2023.

2. Preprocessing Data

- **Handling Missing:** The necessary checks like presence of outliers, datatypes are checked and corrective mesarures are
- **Transforming the variables:**

3. Statistical Methods

- **Regression Analysis**: Particularly Multiple linear regression is used to predict the likelihood of infant mortality based on various factors.
- **Machine Learning Models**: Advanced models like Random Forests, Support Vector Machines (SVM) are also employed to predict infant mortality by learning patterns in the data.

4. Validation and Testing

- **Model Validation:** It's crucial to validate models using techniques like cross-validation or split-sample validation to ensure the model's accuracy and reliability**.** The models are also tested for overfitting of the data by considering the trained data and compared to the test data predictions.

5. Conclusion and Recommendation

- **Conclusion:** Based on the data insights the conclusions are drawn.
- **Recommendation:** The recommendations help the stake holders, governments and NGO's in reducing the infant mortality rates significantly.

3.1 DATA CLEANING

Data cleaning is a crucial preprocessing step in any data analysis project. It involves identifying and correcting inaccuracies, inconsistencies, and missing values in a dataset to ensure the integrity and quality of the data used for analysis. Common data cleaning tasks include handling missing data through imputation or deletion, correcting errors such as typos or formatting issues, and ensuring that data is consistent across different variables.

Furthermore, data cleaning often involves more complex tasks such as merging datasets, removing duplicates, and transforming variables to better suit the analysis goals. For instance, variables may need to be scaled or normalized if different units are used, or encoded if they are categorical. This process not only improves the quality of the data but also enhances the accuracy and reliability of any models or analyses conducted later. Ultimately, thorough data cleaning ensures that the dataset is accurate, complete, and ready for exploratory data analysis, model building, or other advanced statistical methods.

Check1: Look for the duplicate records values in the the entire dataset:

Observation: We have no duplicate values,

Check2: Removal of unnecessary columns

Observation: the columns namely Abbreviation, Calling code, Major city, Largest city, Official language, Latitude, Longitude are removed as these not necessary for the study.

Check3: Look for any missing values in records:

Observation: the records which contain the more than 50% null values or empty cells of the following countries and their respective percentage values as mentioned below are dropped from our dataset.

Index	Countries	Null_Percentage
56	Eswatini	74.074
73	Vatican City	85.185
113	Monaco	55.556
120	Nauru	81.481
128	North Macedonia	77.778
133	Palestinian National Authority	92.593
181	Tuvalu	51.852

Check3: Check for the missing values in the columns:

Observation: The following percentages of null values were observed in each column. Since the missing values constitute no more than 20% of the complete dataset, we have decided to impute them using the mean to avoid impacting the other values. If the missing values had exceeded this threshold, we would have considered alternative imputation methods.

Percentage wise missing values.

Variables	**Percentage missing values**
Minimum wage	20.745
Tax revenue (%)	10.638
Armed Forces size	9.043
Gasoline Price	7.447
Unemployment rate	6.383
Population: Labor force participation (%)	6.383
CPI	5.319

CPI Change (%)	4.787
Maternal mortality ratio	3.723
Gross tertiary education enrollment (%)	2.660
Total tax rate	2.660
Out of pocket health expenditure	2.128
Physicians per thousand	1.064
Gross primary education enrollment (%)	0.532
Infant mortality	0.532
Life expectancy	0.532
Co2-Emissions	0.532
Agricultural Land(%)	0.532
Forested Area (%)	0.532
Population	0.000
Urban_population	0.000
Country	0.000
Density\n(P/Km2)	0.000
GDP	0.000
Fertility Rate	0.000
Birth Rate	0.000
Land Area(Km2)	0.000

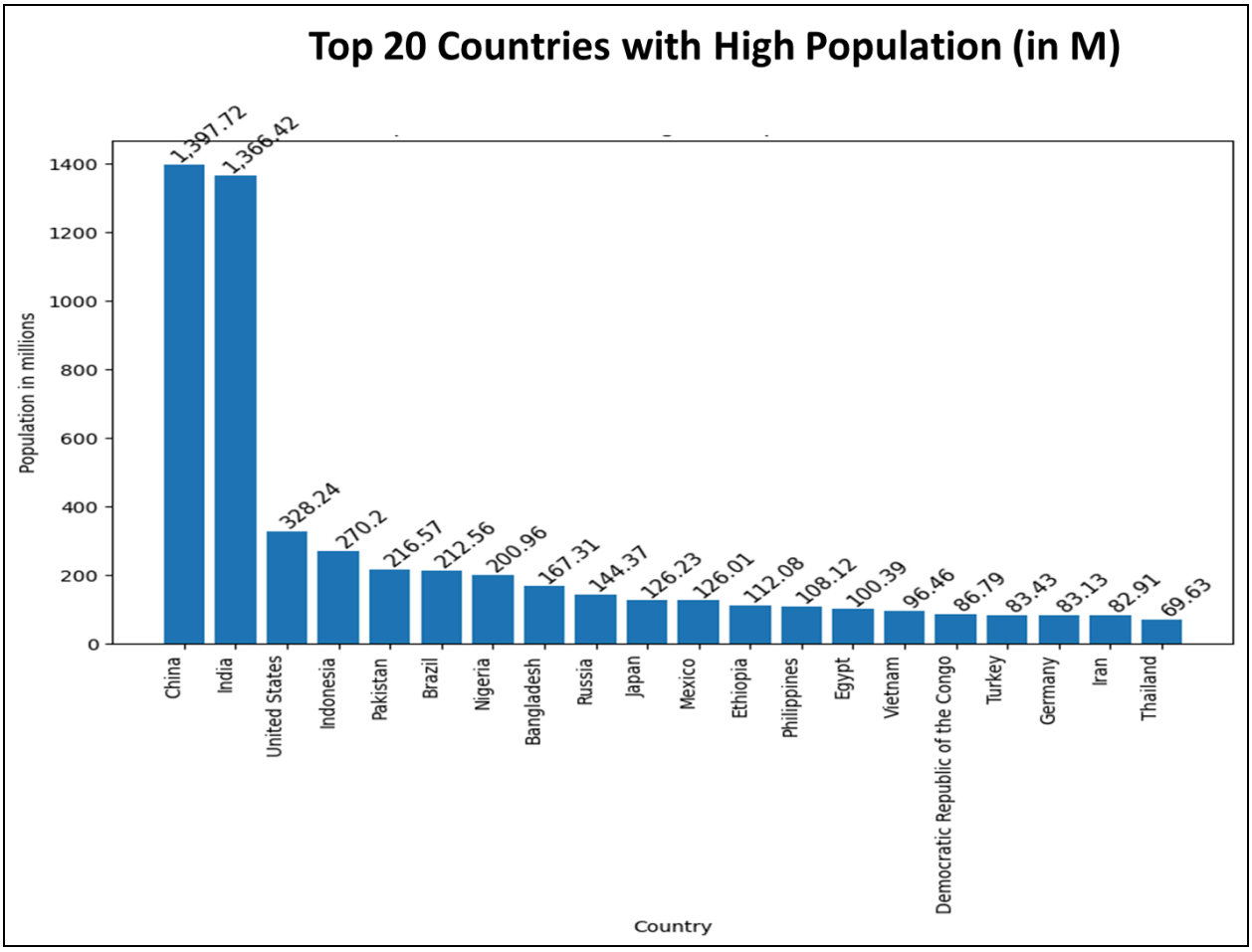
3.2 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a process of describing the data by means of statistical and visualization techniques in order to bring important aspects of that data into focus for further analysis. This involves inspecting the dataset from many angles, describing & summarizing it without making any assumptions about its contents.

Here's a Typical Process

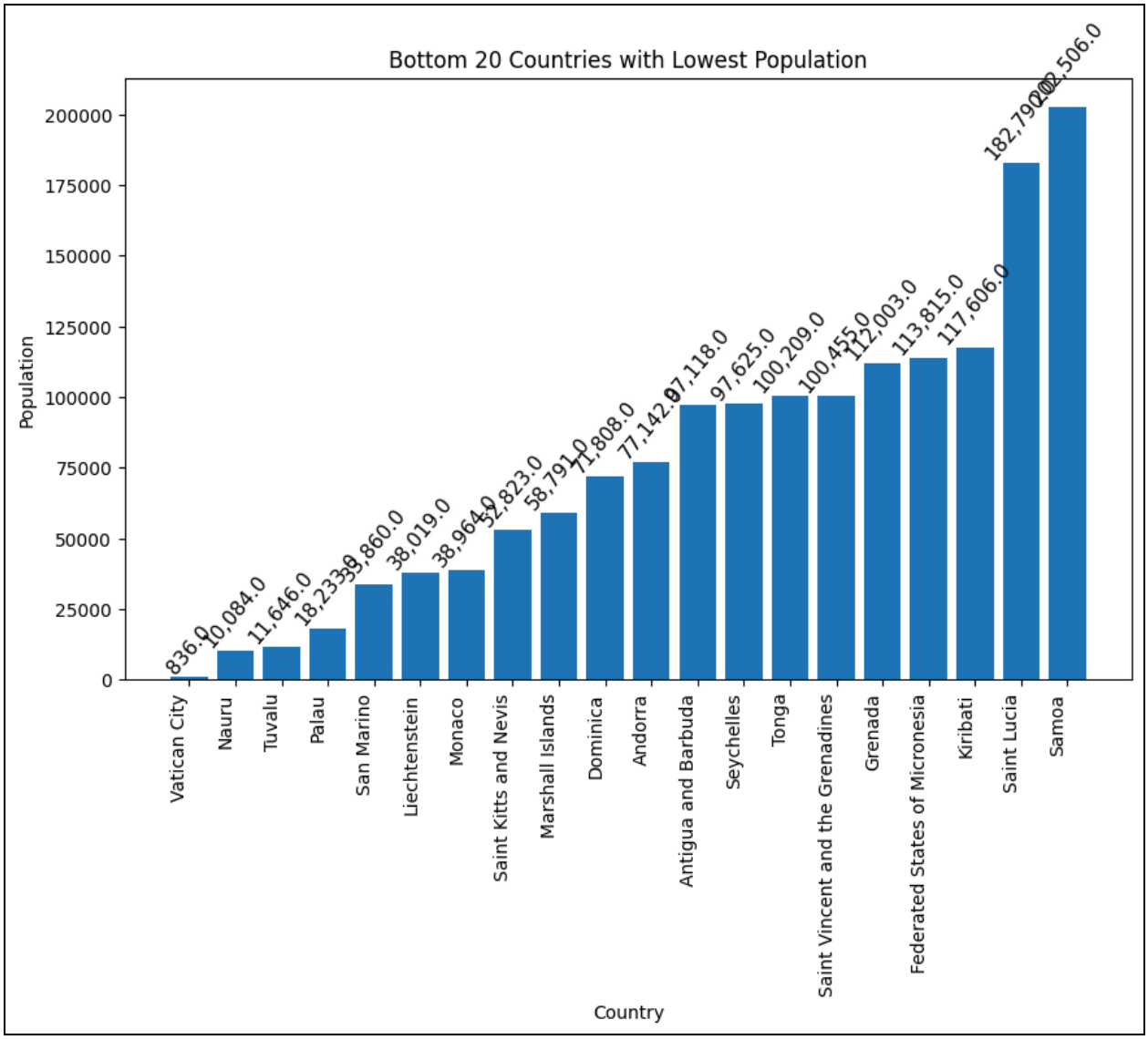
- Look at the Data: Gather information about the data, such as the number of rows and columns, and the type of information each column contains. This includes understanding single variables and their distributions.
- Clean the Data: Fix issues like missing or incorrect values. Preprocessing is essential to ensure the data is ready for analysis and predictive modeling.
- Make Summaries: Summarize the data to get a general idea of its contents, such as average values, common values, or value distributions. Calculating quantiles and checking for skewness can provide insights into the data's distribution.
- Visualize the Data: Use interactive charts and graphs to spot trends, patterns, or anomalies. Bar plots, scatter plots, and other visualizations help in understanding relationships between variables. Python libraries like pandas, NumPy, Matplotlib, Seaborn, and Plotly are commonly used for this purpose.
- Ask Questions: Formulate questions based on your observations, such as why certain data points differ or if there are relationships between different parts of the data.
- Find Answers: Dig deeper into the data to answer these questions, which may involve further analysis or creating models, including regression or linear regression models

Plot - 1)



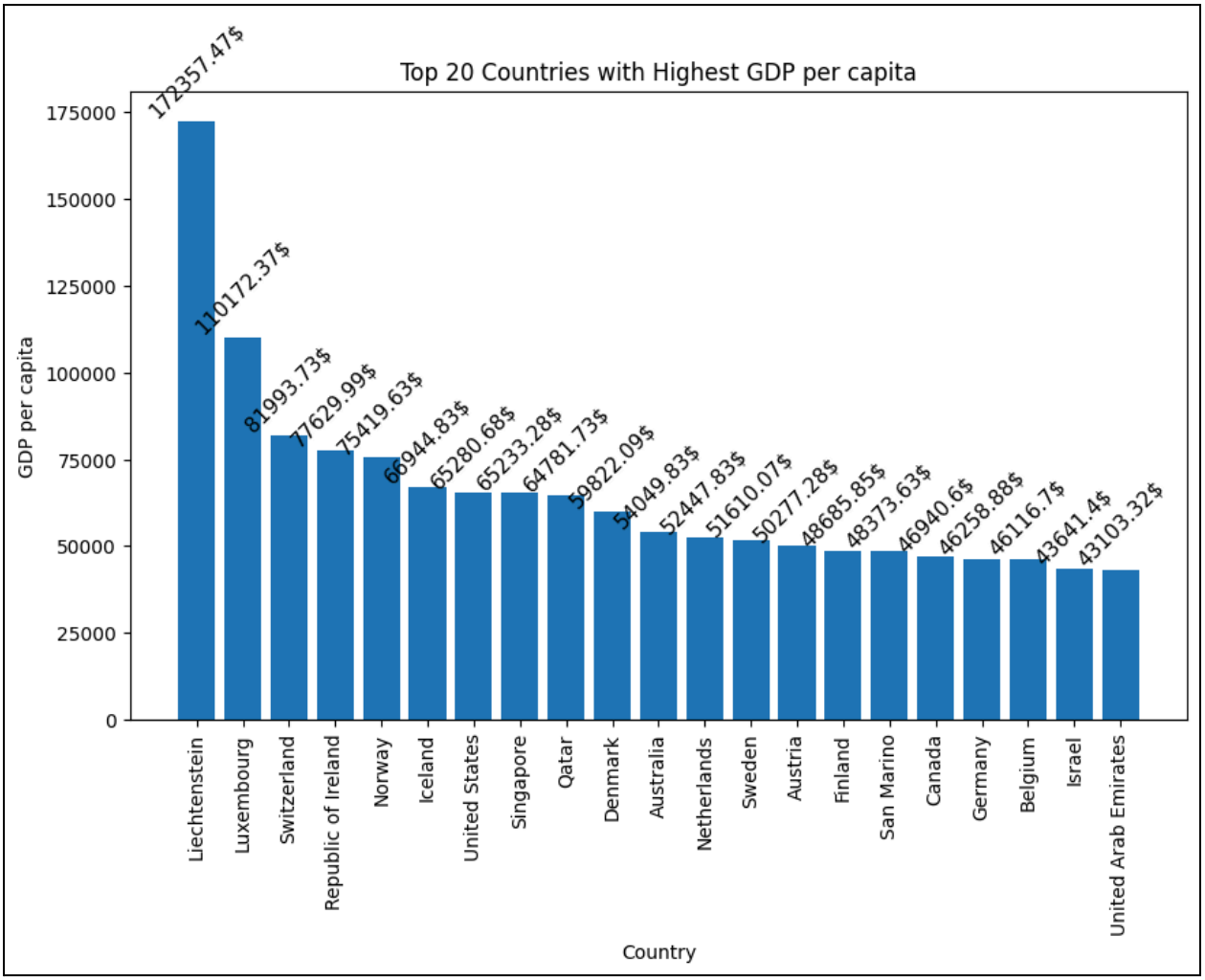
This bar plot depicts the top 20 countries with highest population, with China and India leading the most highly populated countries with 1,397.72 million and 1,366.42 respectively as compared to the rest of the countries like the United states with 328 million followed by Indonesia at 270 million .
China and India approximately hold 4 times the population of the United states and 5 times the Indonesia population.

Plot - 2)



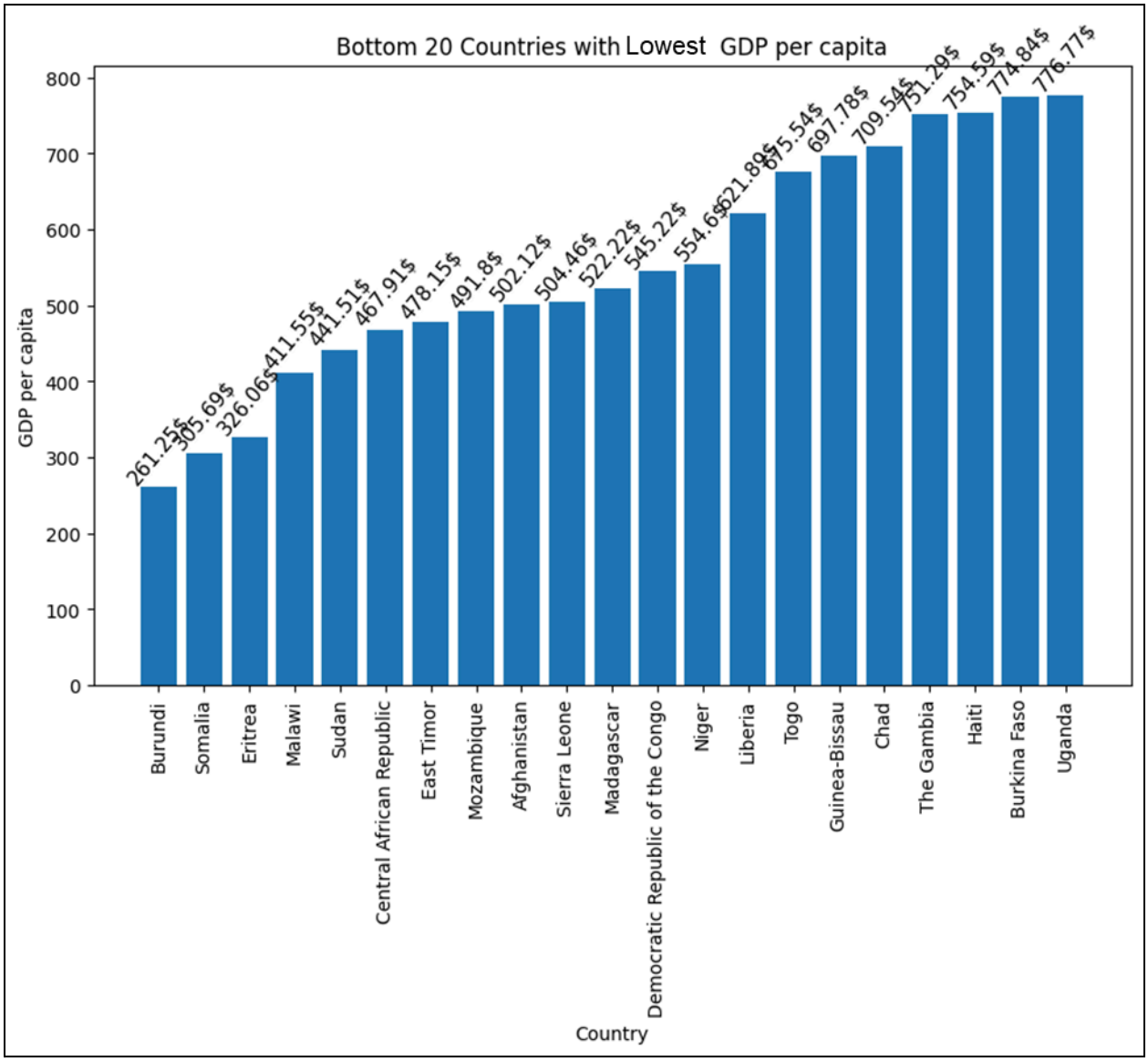
This bar plot depicts the bottom 20 Countries with lowest population during 2023 with Vatican City the least with 836 people which is a part of Europe and followed by Nauru with 10,084 people which is an oceanic country and other constitute like Andorra, Antigua and Barbuda, Seychelles, Tonga , Saint Vincent and the Grenada these countries span across Europe, North America, Africa, and Oceania.

Plot - 3)



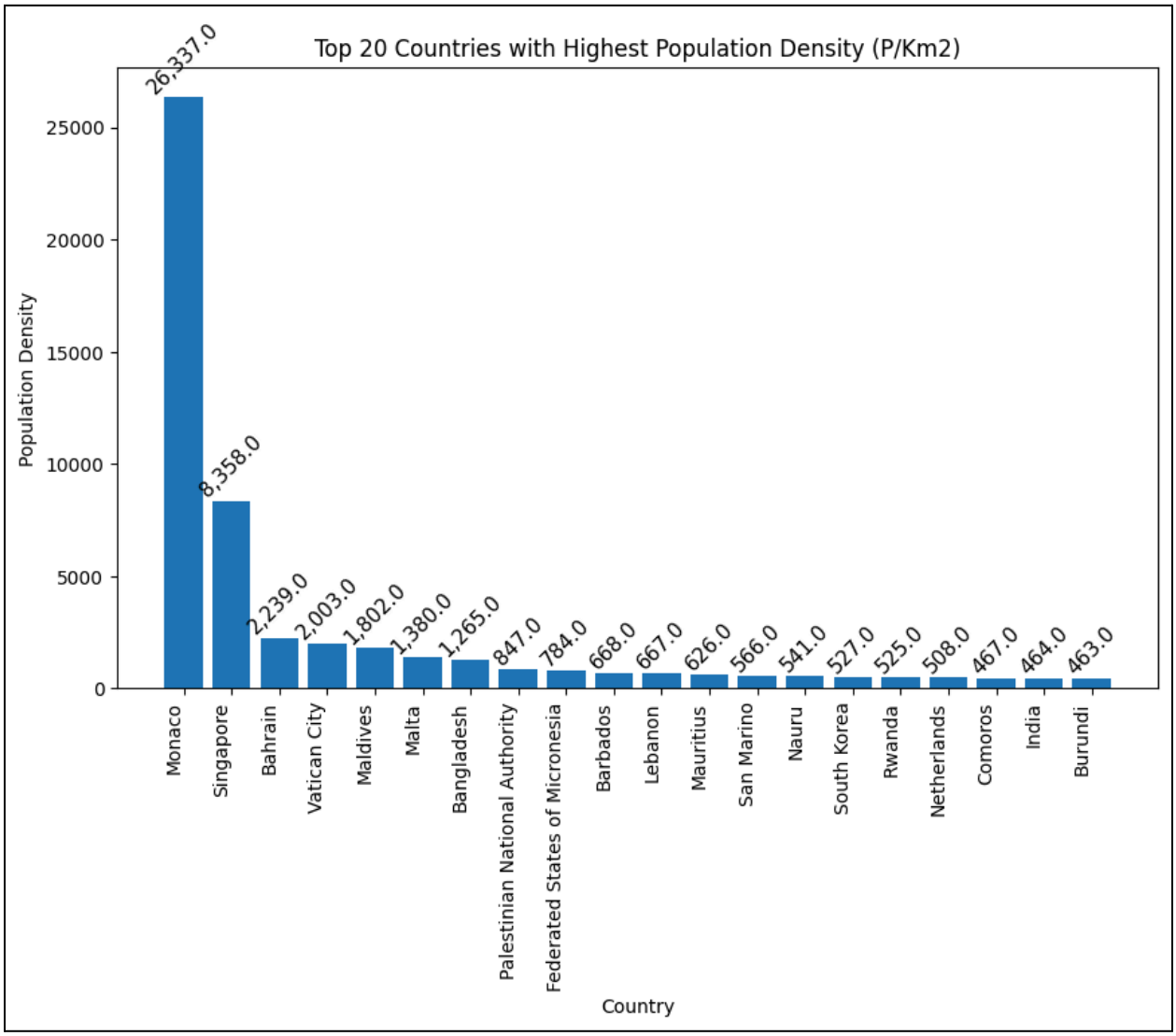
This plot depicts the Top 20 Countries with Highest GDP per capita, Liechtenstein is the leading country with 172357.47 dollars followed by Luxembourg with 110172.37 dollars and next followed countries include Switzerland, Republic of Ireland, Norway, Iceland, United States, Denmark, Netherlands, Sweden, Austria, Finland, San Marino, Canada, Germany, Belgium. These countries span across the European countries and central North America. The western countries have higher GDP per Capita than the other countries.

Plot - 4)



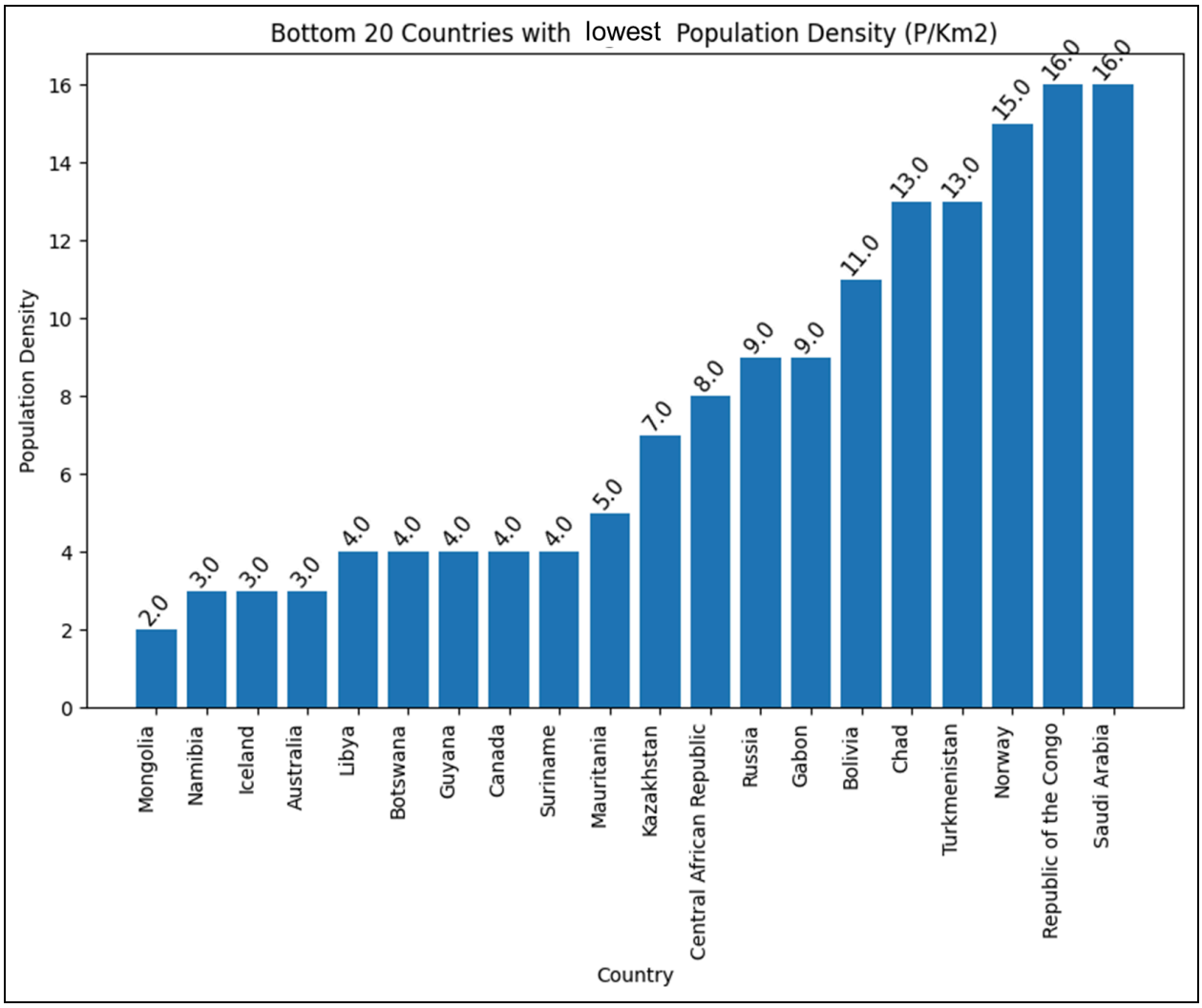
This bar plot shows the bottom 20 countries with lowest GDP per capita, where countries like Burundi, Somalia, Eritrea, Malawi are having lower GDP per capita as compared to the other countries like Uganda, Burkina Faso, Haiti, The Gambia, Chad, Guinea-Bissau, Togo.
It can be observed that some of the African countries mentioned above have lower GDP per capita.

Plot - 5)



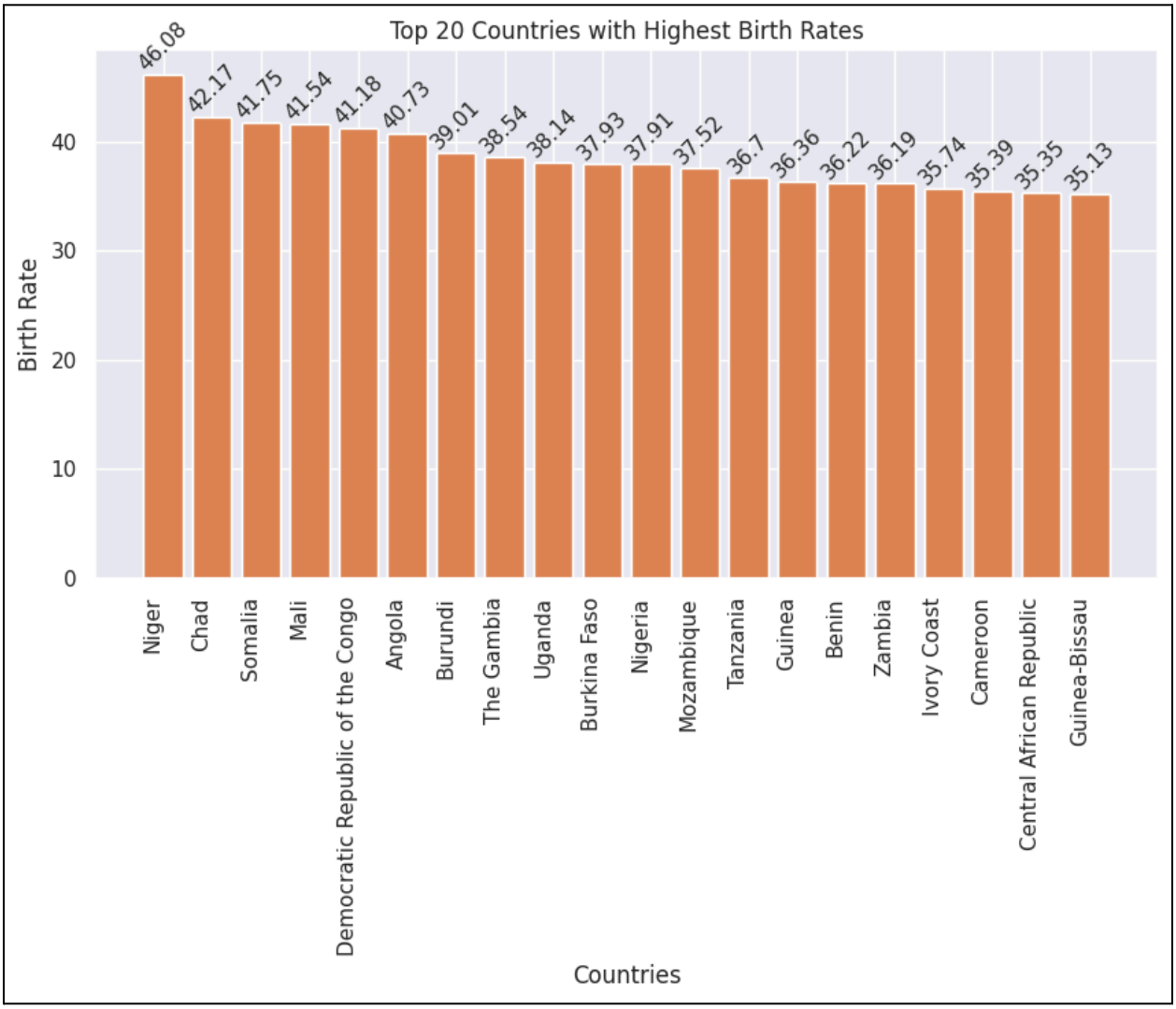
This bar plot depicts the Top 20 Countries with highest Population Density, where Monaco, a European country, has the highest density with 26,337 people living per square kilometer of land, followed by Singapore, Bahrain, Vatican City and maldives. These countries are located in Asia and Europe.

Plot - 6)



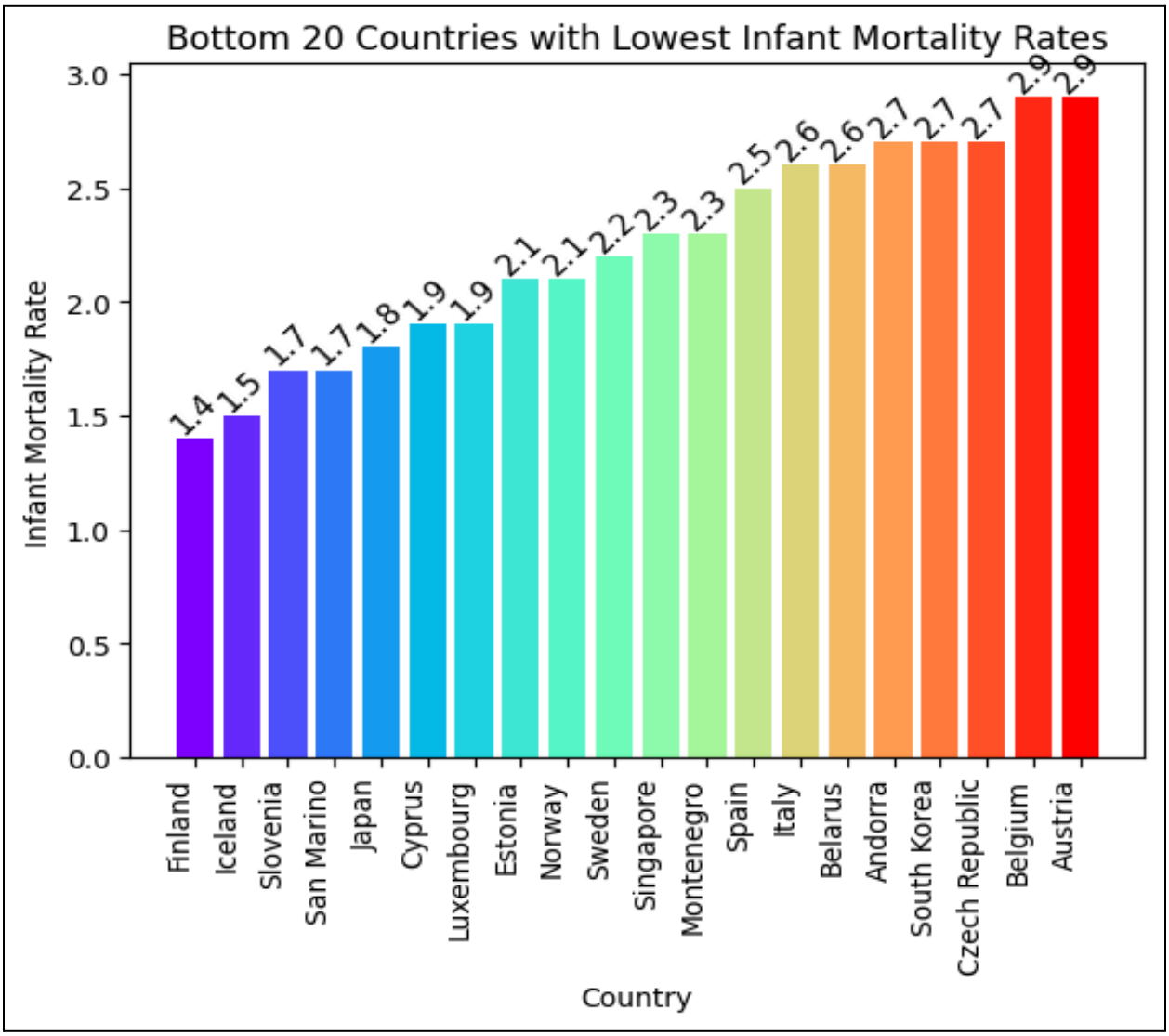
This bar plot depicts the bottom 20 Countries with lowest Population Density, with Mongolia, Namibia, Iceland, Australia, Libya with 2 to 3 people living per square kilometer.
Namibia and Libya are in Africa, Iceland is in Europe, and Australia is in Oceania.

Plot - 7)



This graph depicts the top 20 countries with highest birth rates, here Niger has 46.08 birth rate followed by Chad, Somalia, Mali and Democratic Republic of the Congo with 42.17,41.75,41.54,41.8 respectively, these countries are in Africa. Here the African countries have the highest birth rate compared to the other countries in the world.

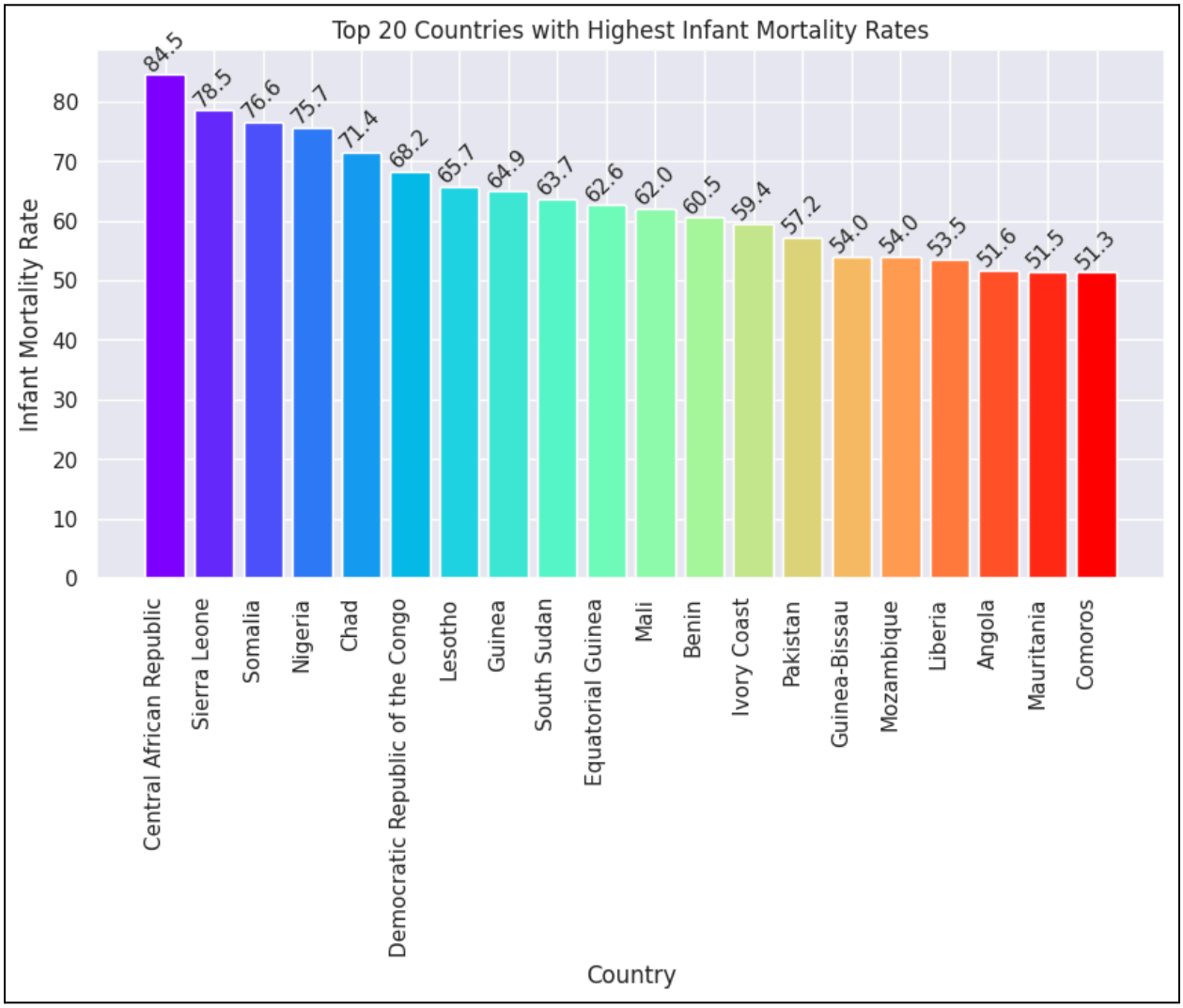
Plot - 8)



Interpretation:

This graph illustrates the bottom 20 countries with the lowest infant mortality rates. Leading the list are Finland (1.4), Iceland (1.5), Slovenia (1.7), San Marino (1.7), Japan (1.8), Cyprus (1.9), and Luxembourg (1.9), all with infant mortality rates below 2 deaths per 1,000 live births. These low rates are partly due to the smaller populations in these countries, contributing to their relatively low infant mortality figures compared to other countries worldwide. The list continues with Estonia (2.1), Norway (2.1), Sweden (2.2), Singapore (2.3), Montenegro (2.3), Spain (2.5), Italy (2.6), Belarus (2.6), Andorra (2.7), South Korea (2.7), Czech Republic (2.7), Belgium (2.9), and Austria (2.9).

Plot - 9)

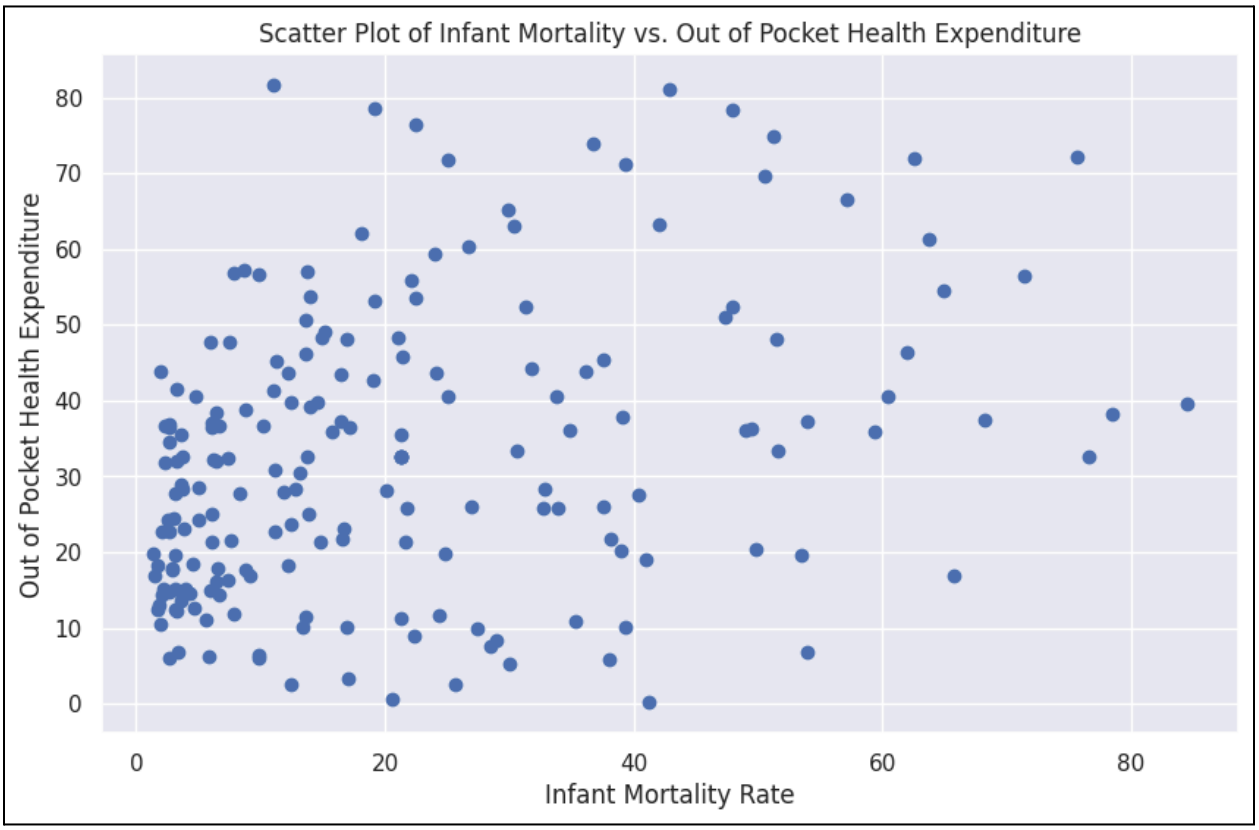


Interpretation:

This plot displays the top 20 countries with the highest infant mortality rates. Leading the list is the Central African Republic, with an infant mortality rate of 84.5 deaths per 1,000 live births. It is followed by Sierra Leone (78.5), Somalia (76.6), Nigeria (75.7), and Chad (71.4). Other countries in the top 20 include the Democratic Republic of the Congo (68.2), Lesotho (65.7), Guinea (64.9), South Sudan (63.7), and Equatorial Guinea (62.6). The list continues with Mali (62.0), Benin (60.5), Ivory Coast (59.4), Pakistan (57.2), Guinea-Bissau (54.0), Mozambique (54.0), Liberia (53.5), Angola (51.6), Mauritania (51.5), and Comoros (51.3).

Health indicators:

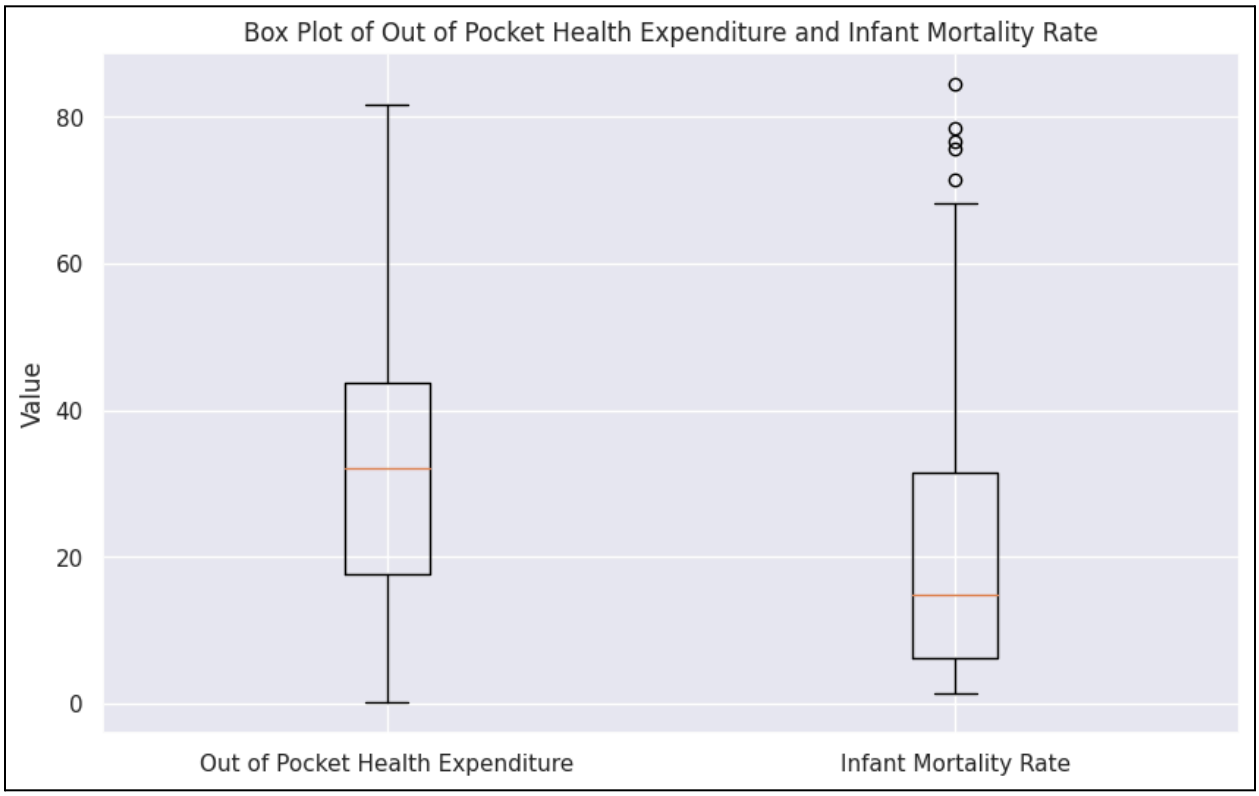
Plot - 10)



Interpretation : This scatter plot illustrates the relationship between Infant Mortality Rate and Out-of-Pocket Health Expenditure. It reveals that countries with an infant mortality rate below 20 and out-of-pocket health expenditures up to 40% tend to show less variability in their data points.

In contrast, countries with an infant mortality rate above 20 exhibit much greater scatter, indicating more diverse patterns of health expenditure and infant mortality.

Plot - 11)



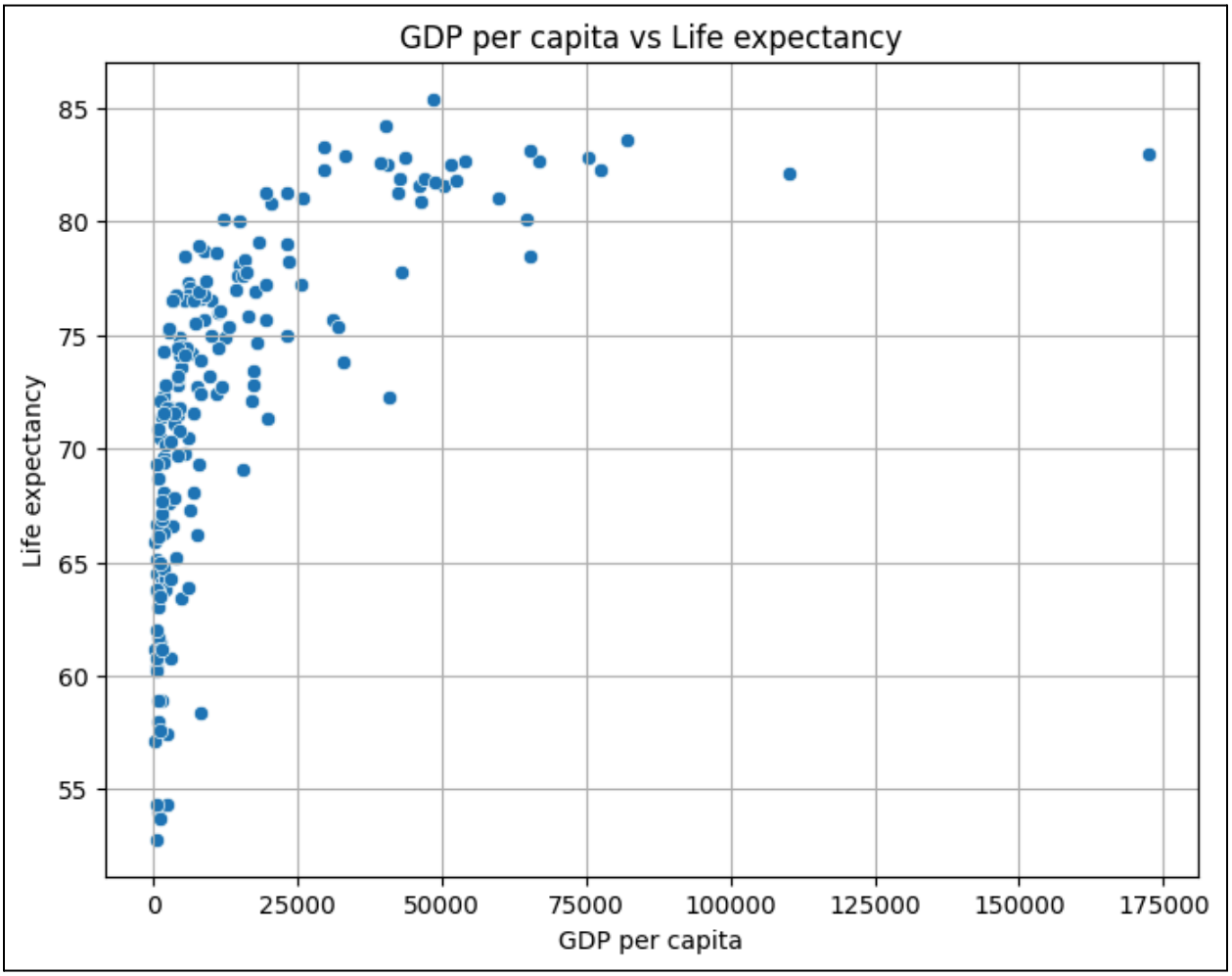
Interpretation:

This box plot compares "Out-of-Pocket Expenditure" with "Infant Mortality Rate," offering insightful observations. The median value indicates that approximately 50% of the population spends around 32.1% of their income on healthcare expenses not covered by insurance. Additionally, the median infant mortality rate shows that about 50% of countries have an infant mortality rate of 14 per 1,000 births. The plot also highlights that there are countries with significantly higher infant mortality rates, suggesting notable disparities in healthcare outcomes across different regions.

Countries that having higher infant mortality rates are Central African Republic, Sierra Leone, Somalia, Nigeria and Chad.

Economic indicators:

Plot -12)



Interpretation:

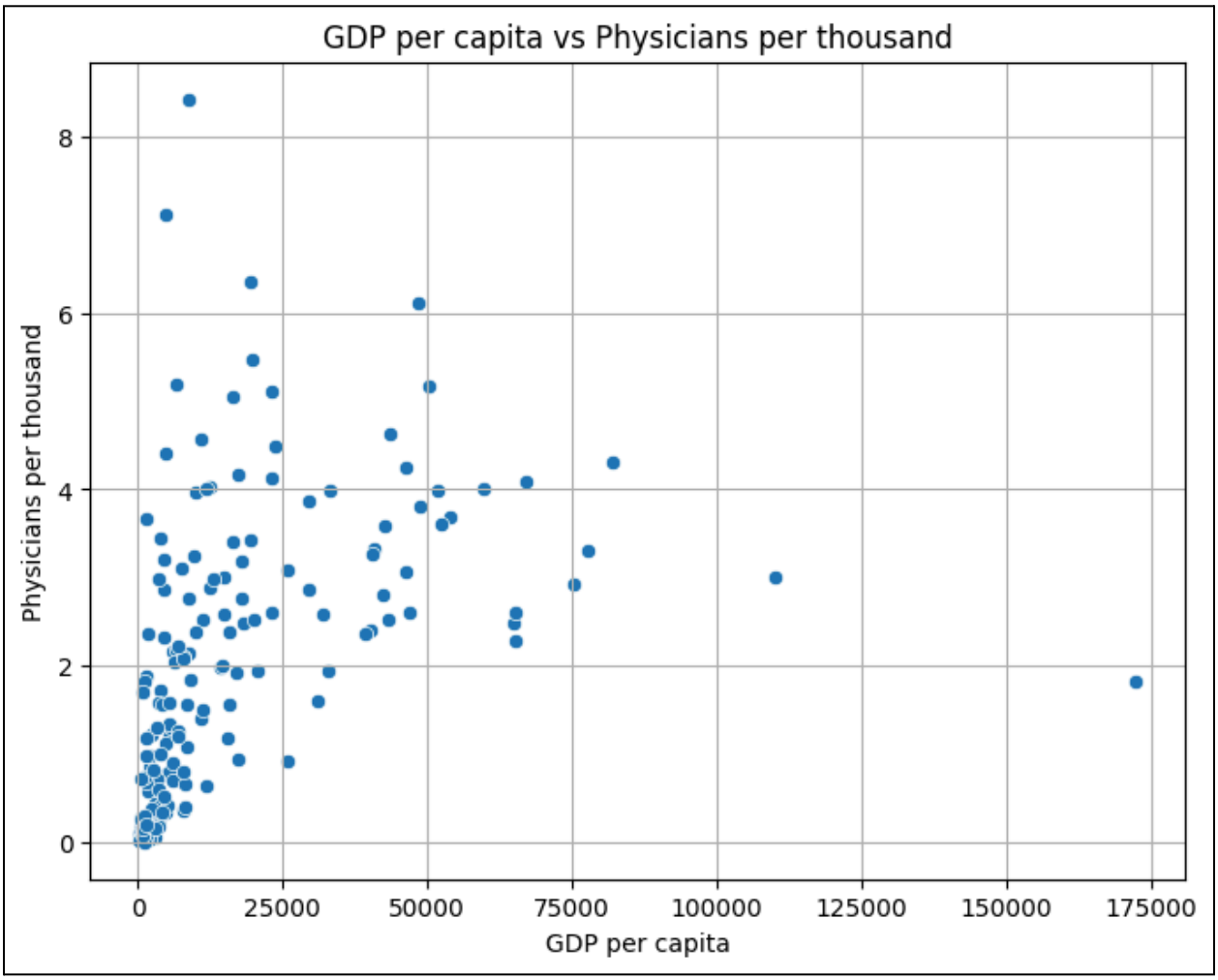
This scatter plot illustrates the relationship between GDP per capita and life expectancy across different countries. It shows that in countries with a GDP per capita below $25,000, life expectancy typically ranges from 50 to 80 years.

This suggests that lower economic resources are associated with a wider range of life expectancies, possibly due to varying levels of healthcare access and quality.

Conversely, countries where life expectancy exceeds 80 years generally have a GDP per capita above $25,000.

This pattern implies that higher economic wealth is often linked to longer life expectancies, likely due to better healthcare systems, higher living standards, and improved overall quality of life. Thus, the data indicates a clear trend where increased economic prosperity correlates with longer life expectancy.

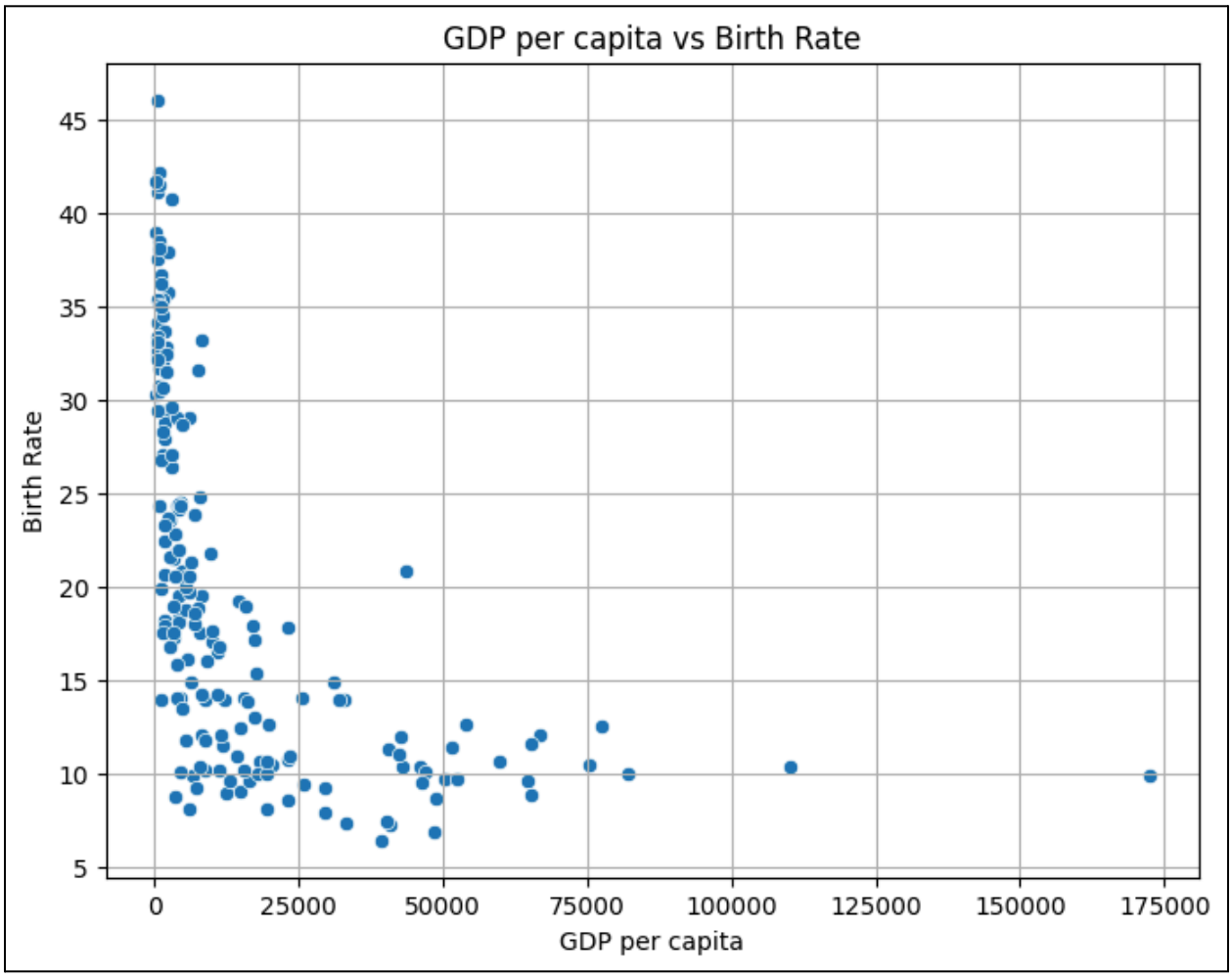
Plot - 13)



Interpretation:

This scatter plot displays the relationship between the number of physicians per thousand people and GDP per capita. It reveals that most countries are concentrated in a region where GDP per capita is below $50,000 and the number of physicians per thousand people is around 4.

Plot - 14)

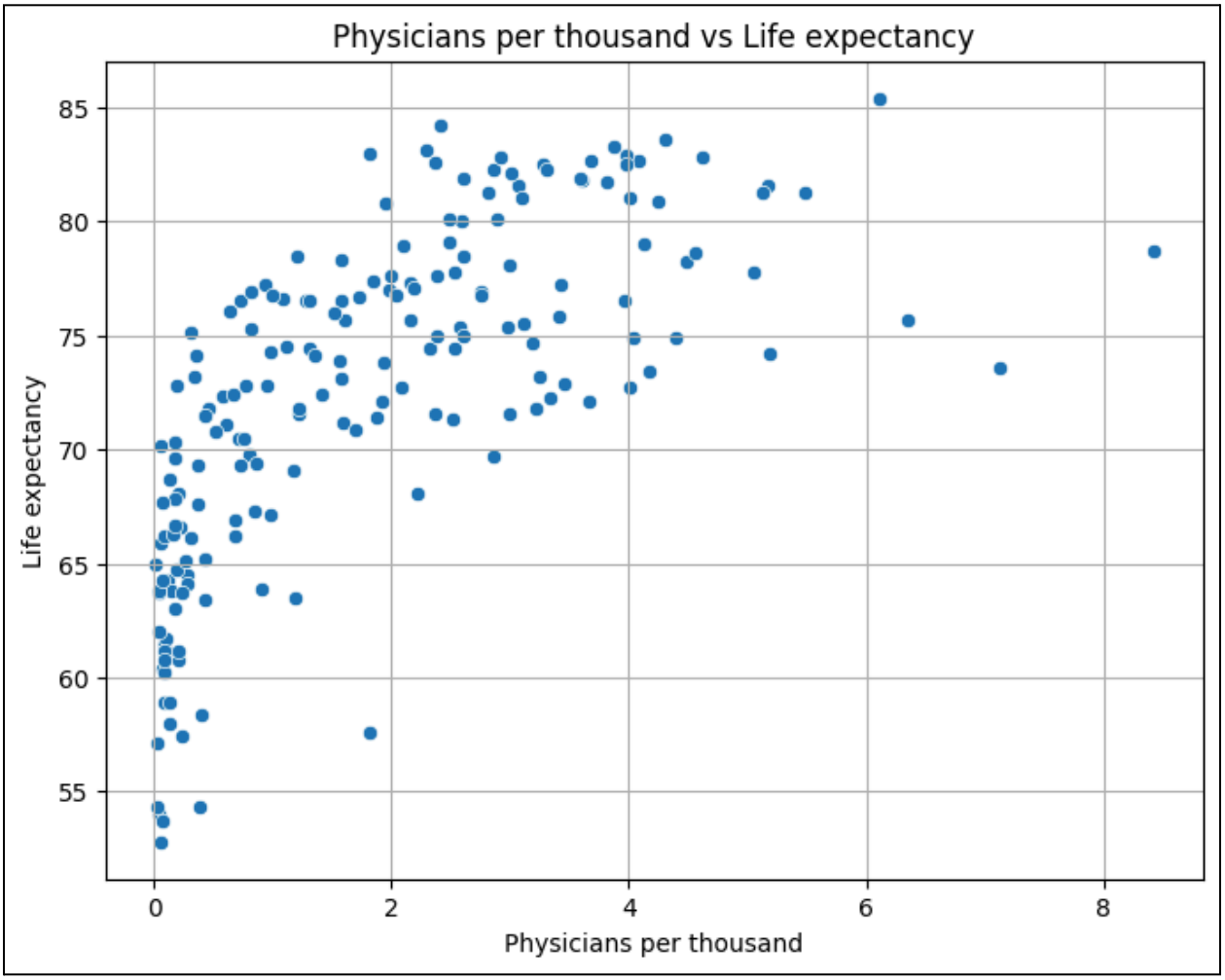


Interpretation:

This scatter plot reveals a clear trend in the relationship between GDP per capita and birth rates across different countries. Specifically, it shows that nations with a GDP per capita below $25,000 tend to have higher birth rates, ranging between 10 and 45 births per 1,000 people. In contrast, countries with a GDP per capita exceeding $25,000 generally experience lower birth rates, often falling below 15 births per 1,000 people.

This pattern suggests that as a country's economic wealth increases, its birth rate tends to decrease, potentially due to factors such as improved access to education, healthcare, and family planning services, which are commonly associated with higher income levels.

Plot - 15)

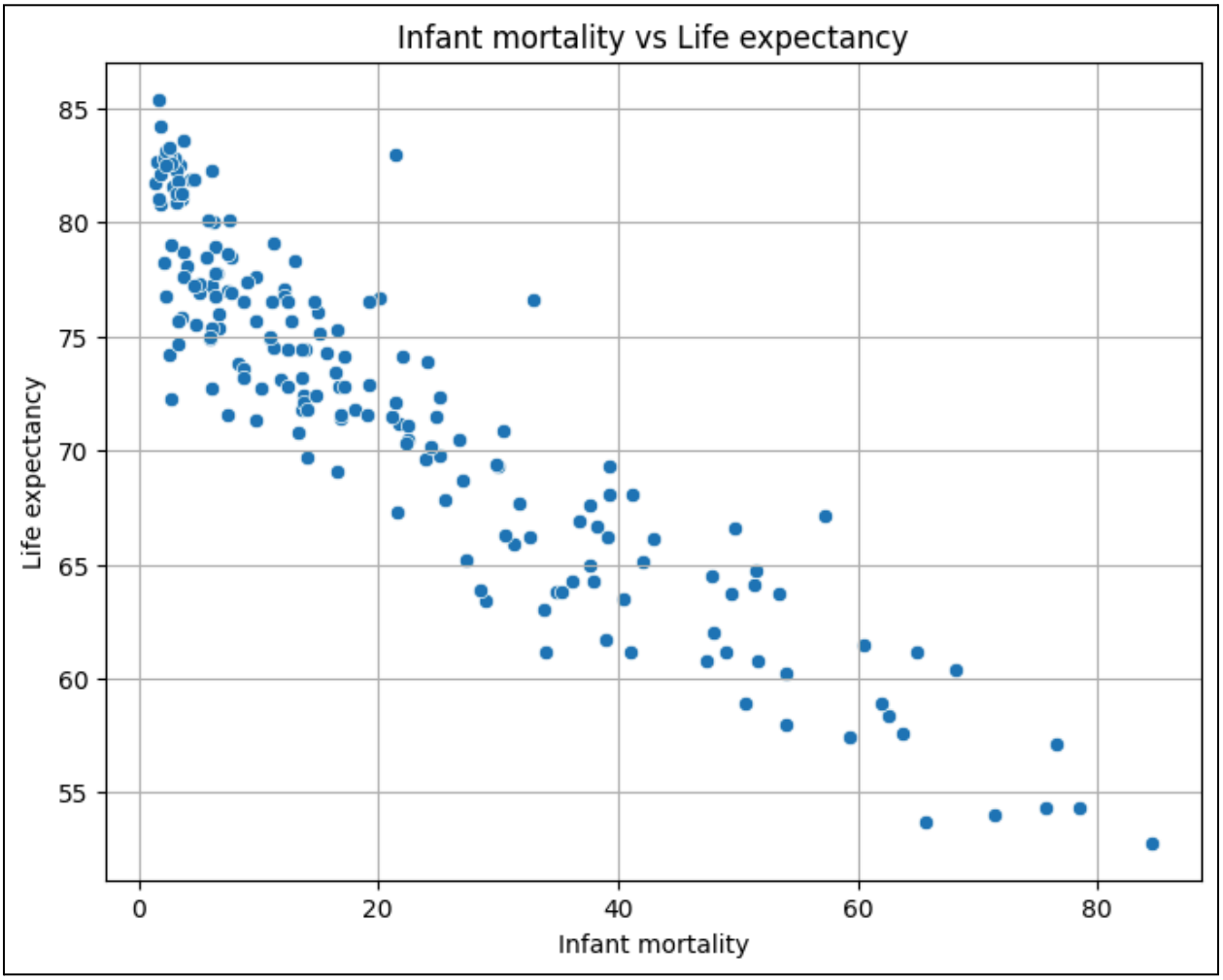


Interpretation:

This scatter plot between Physicians per Thousand and Life Expectancy illustrates a positive relationship between the two variables. Countries with a higher density of physicians, particularly those with more than 2 physicians per thousand people, tend to have life expectancies of 70 years and above. In contrast, countries with fewer than 2 physicians per thousand typically have lower life expectancies, ranging from 50 to 65 years on average.

This trend suggests that access to medical professionals plays a crucial role in extending life expectancy and improving overall public health.

Plot - 16)

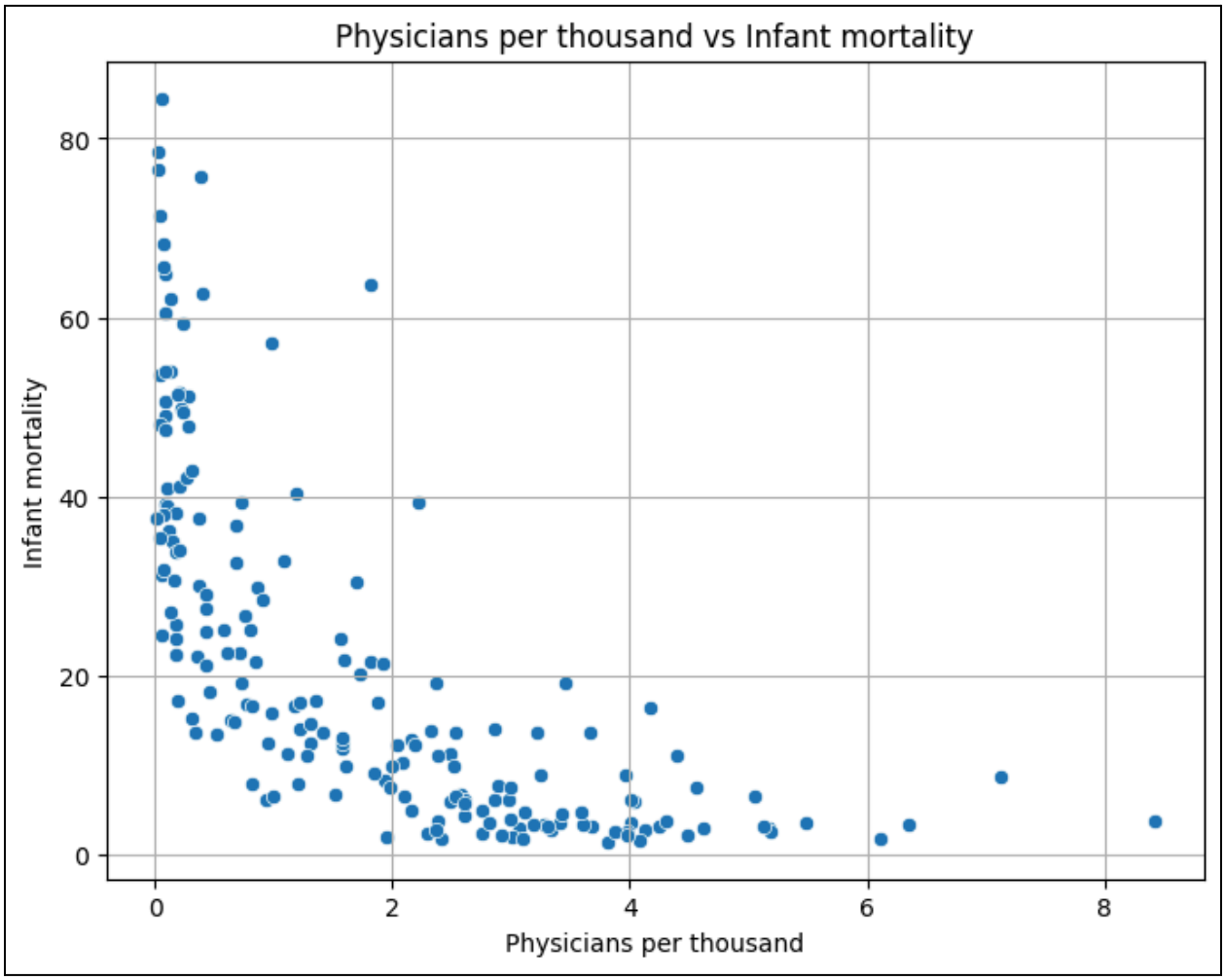


Interpretation:

The scatter plot illustrating the relationship between Infant Mortality and Life Expectancy reveals a strong inverse correlation between the two variables. Countries with higher life expectancy consistently exhibit lower infant mortality rates, indicating better overall health outcomes and more advanced healthcare systems.

Conversely, nations with elevated infant mortality rates tend to have significantly lower life expectancy, reflecting broader challenges in healthcare access, quality, and socioeconomic conditions. This trend highlights the critical impact that infant mortality has on a population's overall health and longevity, making it a key indicator of national well-being.

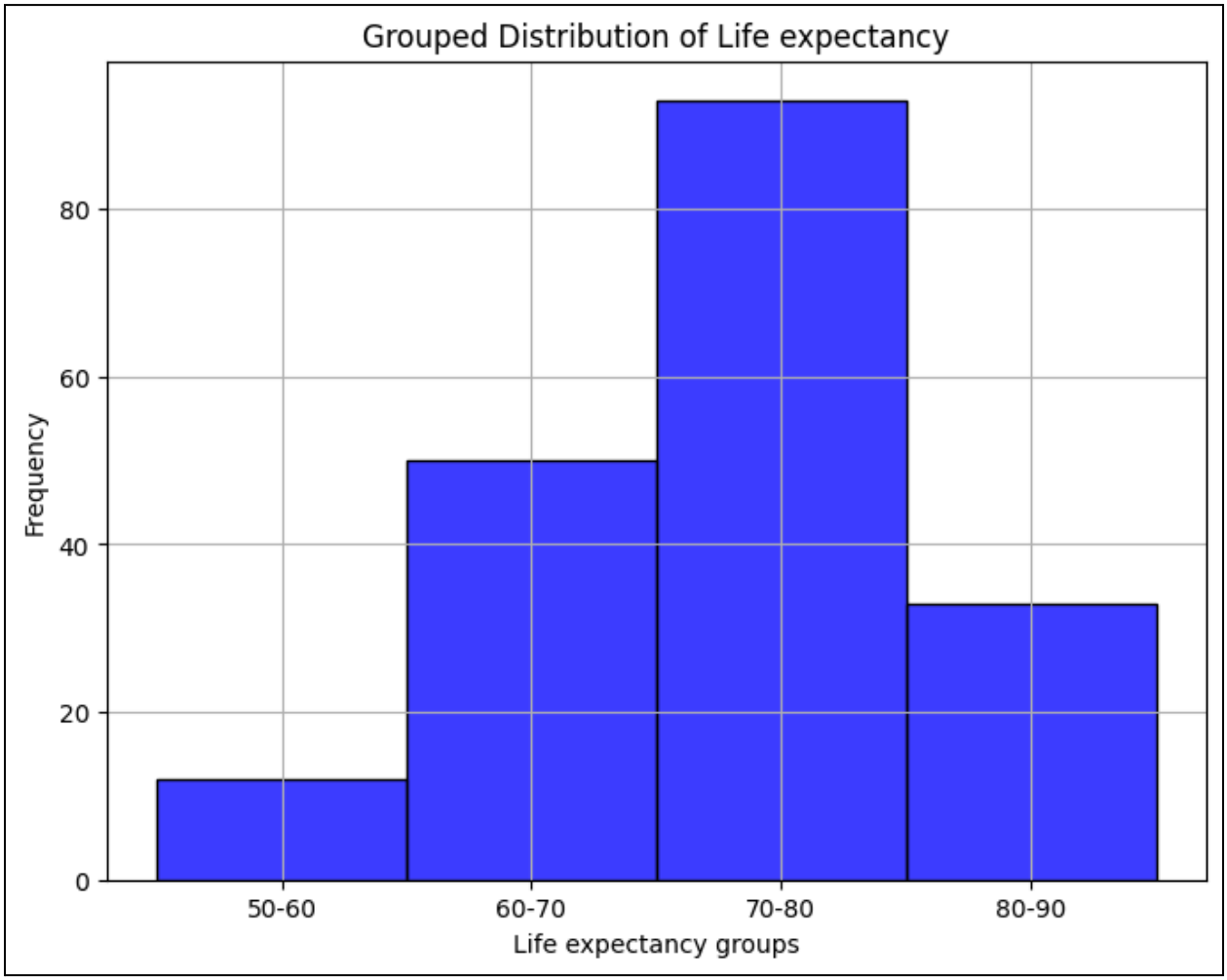
Plot - 17)



Interpretation:

This is the scatter plot of Physicians per thousand vs Infant mortality, from the graph it is clear that the countries with higher Infant mortality have lower physicians per thousand, likewise the higher physicians per thousand have lower lower Infant mortality.

Plot - 18)



This histogram depicts the grouped distribution of life expectancy, it can be observed that in most of the countries across the world the life expectancy for age groups between 70-80 years are comparatively more than the other age groups.

3.3 MULTICOLINEARITY:

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

Effect of Multicollinearity:

Although multicollinearity does not affect the regression estimates, it makes them vague, imprecise, and unreliable. Thus, it can be hard to determine how the independent variables influence the dependent variable individually. This inflates the standard errors of some or all of the regression coefficients.

Detecting Multicollinearity

A statistical technique called the variance inflation factor (VIF) can detect and measure the amount of collinearity in a multiple regression model. VIF measures how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not linearly related. A VIF of 1 will mean that the variables are not correlated; a VIF between 1 and 5 shows that variables are moderately correlated, and a VIF between 5 and 10 will mean that variables are highly correlated

Variance Inflation Factor(VIF):

A variance inflation factor is a tool to help identify the degree of multicollinearity. Multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables—the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

Computed VIF values for the features is as follows in form of table.

Table :1

S.no	variables	VIF
1	Birth Rate	54.6
2	Urban_population	52.6
3	Fertility Rate	44.9
4	Population	28.7
5	Co2-Emissions	25.9
6	GDP	10.1
7	CPI	6
8	Armed Forces size	5.7
9	CPI Change (%)	5.6
10	Maternal mortality ratio	3.7
11	Physicians per thousand	3.2
12	Gross tertiary education enrollment (%)	3.1
13	GDP per capita	2.3
14	Land Area(Km2)	2.2
15	Minimum wage	2.1
16	Gasoline Price	1.8
17	Out of pocket health expenditure	1.7
18	Tax revenue (%)	1.6
19	Agricultural Land(%)	1.5
20	Life expectancy	1.5
21	Forested Area (%)	1.5
22	Population: Labor force participation (%)	1.5
23	Total tax rate	1.4

24	Unemployment rate	1.4
25	Gross primary education enrollment (%)	1.3
26	Density\n(P/Km2)	1.2

From the above table the VIF(Variance Inflation Factor) for Birth Rate, Urban Population, Fertility Rate, Population, CO2 - Emissions and GDP are having much higher values of VIF than compared to other variables. So Birth Rate is highly collinear with other variables, hence removing this variable from the dataset.

The obtained table is as follows

Table :2

S.no	**variables**	**VIF**
1	Urban_population	52.4
2	Population	28.6
3	Co2-Emissions	25.9
4	GDP	10.1
5	CPI	5.9
6	Armed Forces size	5.6
7	CPI Change (%)	5.5
8	Maternal mortality ratio	3.7
9	Fertility Rate	3.5
10	Gross tertiary education enrollment (%)	3
11	Physicians per thousand	2.8
12	GDP per capita	2.3
13	Minimum wage	2.1
14	Land Area(Km2)	2.1
15	Gasoline Price	1.7
16	Out of pocket health expenditure	1.6

17	Tax revenue (%)	1.6
18	Agricultural Land(%)	1.5
19	Life expectancy	1.5
20	Forested Area (%)	1.5
21	Population: Labor force participation (%)	1.5
22	Total tax rate	1.4
23	Unemployment rate	1.4
24	Gross primary education enrollment (%)	1.2
25	Density\n(P/Km2)	1.2

Fertility Rate exhibits a sharp drop in the VIF value from 44.9 to 3.5 when we compare this table 2 with that of table 1. These two variables were collinear with one another. Similarly, the variable Urban Population's next higher VIF value is 52.4; as a result, this variable is eliminated from the dataset.

The resultant table is as follows:

Table: 3

S.no	**variables**	**VIF**
1	Co2-Emissions	16.9
2	GDP	10.1
3	Population	8.3
4	CPI	5.9
5	Armed Forces size	5.6
6	CPI Change (%)	5.5
7	Maternal mortality ratio	3.7
8	Fertility Rate	3.5
9	Gross tertiary education enrollment (%)	3

10	Physicians per thousand	2.8
11	GDP per capita	2.3
12	Minimum wage	2.1
13	Land Area(Km2)	2
14	Gasoline Price	1.7
15	Out of pocket health expenditure	1.6
16	Tax revenue (%)	1.6
17	Agricultural Land(%)	1.5
18	Life expectancy	1.5
19	Forested Area (%)	1.5
20	Population: Labor force participation (%)	1.5
21	Unemployment rate	1.4
22	Total tax rate	1.3
23	Gross primary education enrollment (%)	1.2
24	Density\n(P/Km2)	1.2

Table 2 and Table 3 show that Population and CO2 emissions were substantially correlated with Urban Population; CO2 emissions had the second highest VIF value, 16.9 > 10 (threshold value), so this variable was removed from the dataset.

The resultant Table is as follows:

Table: 4

S. no	**variables**	**VIF**
1	CPI	5.9
2	CPI Change (%)	5.5

3	Armed Forces size	5.4
4	Population	5.2
5	Maternal mortality ratio	3.6
6	Fertility Rate	3.5
7	Gross tertiary education enrollment (%)	3
8	Physicians per thousand	2.8
9	GDP	2.4
10	GDP per capita	2.2
11	Minimum wage	2
12	Land Area(Km2)	1.9
13	Gasoline Price	1.7
14	Tax revenue (%)	1.6
15	Agricultural Land(%)	1.5
16	Life expectancy	1.5
17	Out of pocket health expenditure	1.5
18	Forested Area (%)	1.5
19	Population: Labor force participation (%)	1.5
20	Unemployment rate	1.4
21	Total tax rate	1.3
22	Gross primary education enrollment (%)	1.2
23	Density\n(P/Km2)	1.2

Since the VIF values for all the variables are less than the 10 (Threshold value), it can be said that there is no multicollinearity between the variables in the dataset, further analysis like model building can be done on the data with above variables.

3.4 SIGNIFICANCE TEST FOR MULTIPLE LINEAR REGRESSION

Once we have addressed multicollinearity, we proceed to train the OLS regression model. The OLS model is a widely used technique for linear regression, aiming to find the best-fitting line through the observed data points.

Using the statsmodels library in Python, we construct the formula for our regression model. The formula follows the syntax: "dependent_variable ~ independent_variables"

Next, we fit the model to our preprocessed dataset using the fit method. This process estimates the coefficients of the independent variables and calculates other statistical measures associated with the regression model.

Hypothesis Testing plays a crucial role in evaluating the significance of the coefficients in the regression model. In this analysis, we specifically used p- value to test the null hypothesis that the coefficient for the independent variables are zeros. A statistically significant coefficient would suggest that there is a significant relationship between Infant mortality and the independent variables like Density, Agriculture Land, Land Area, Armed Forces size, CPI, CPI change, Fertility Rate, Forested Area, Gasoline Prices, GDP, Gross Primary education enrollment, Gross tertiary education enrollment, Life expectancy, Maternal mortality ratio, Minimum wage, Out of Pocket health expenditure, Physicians perthousand, Population, Population: Labour force participation, Tax revenue, Unemployment, GDP per Capita.. Indicating differences Infant mortality between the countries

Following is the python code for OLS (Ordinary Least Square)

```
import statsmodels.api as sm

x_nomulti_colinearity=X.drop(['Urban_population','Birth Rate','Co2-Emissions'], axis=1)

x_train_constant = sm.add_constant(x_nomulti_colinearity)

model = sm.OLS(y, x_train_constant).fit()

print(model.summary())
```

OLS Regression Results

Dep. Variable:	Infant mortality	R-squared:	0.912
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	73.82
Date:	Sun, 18 Aug 2024	Prob (F-statistic):	2.47e-74
Time:	10:34:39	Log-Likelihood:	-596.83
No. Observations:	188	AIC:	1242.
Df Residuals:	164	BIC:	1319.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	114.2204	16.072	7.107	0.000	82.486	145.955
Density (P/Km2)	-7.612e-06	0.001	-0.010	0.992	-0.001	0.001
Agricultural Land(%)	-0.0238	0.026	-0.899	0.370	-0.076	0.028
Land Area(Km2)	-3.303e-08	3.22e-07	-0.103	0.918	-6.68e-07	6.02e-07
Armed Forces size	-1.508e-07	2.92e-06	-0.052	0.959	-5.91e-06	5.61e-06
CPI	-0.0036	0.003	-1.249	0.214	-0.009	0.002
CPI Change (%)	0.0559	0.045	1.245	0.215	-0.033	0.145
Fertility Rate	1.9974	0.791	2.526	0.012	0.436	3.559
Forested Area (%)	-0.0216	0.024	-0.886	0.377	-0.070	0.027
Gasoline Price	1.5096	1.726	0.875	0.383	-1.898	4.918
GDP	-1.206e-13	3.13e-13	-0.386	0.700	-7.38e-13	4.97e-13
Gross primary education enrollment (%)	-0.0076	0.039	-0.196	0.845	-0.085	0.069
Gross tertiary education enrollment (%)	-0.0529	0.027	-1.931	0.055	-0.107	0.001
Life expectancy	-1.3623	0.182	-7.474	0.000	-1.722	-1.002
Maternal mortality ratio	0.0240	0.004	5.735	0.000	0.016	0.032
Minimum wage	0.0207	0.251	0.083	0.934	-0.474	0.515
Out of pocket health expenditure	0.0667	0.031	2.134	0.034	0.005	0.129
Physicians per thousand	-0.8517	0.458	-1.861	0.065	-1.755	0.052
Population	3.319e-09	7.03e-09	0.472	0.637	-1.06e-08	1.72e-08
Population: Labor force participation (%)	-0.0550	0.057	-0.965	0.336	-0.168	0.058
Tax revenue (%)	-0.0348	0.088	-0.395	0.693	-0.209	0.139
Total tax rate	0.0345	0.026	1.332	0.185	-0.017	0.086
Unemployment rate	0.0299	0.114	0.263	0.793	-0.195	0.254
GDP per capita	8.95e-05	3.11e-05	2.874	0.005	2.8e-05	0.000

Omnibus:	22.897	Durbin-Watson:	2.136
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.577
Skew:	0.599	Prob(JB):	3.45e-10

The variables Density\n(P/Km2),Armed Forces size,Minimum wage, Land Area(Km2), Gross primary education enrollment (%), Unemployment rate, GDP, Tax revenue (%), Population, Gasoline Price, Forested Area (%), Agricultural Land(%), Population: Labor force participation (%), CPI Change (%) have greater P value to the alpha 0.05, indicating that they are not significant to the model in predicting the infant mortality. Where as Maternal mortality ratio, physicians per thousand, out of pocket expenditure and fertility rate are significant features for modeling infant mortality.

Ordinary Least Squares Regression for the significant and transformed variables:

The need for log transformation of variables arised since the variance of the residuals increased with the magnitude of the dependent variable (i.e infant mortality) and to stabilizes variance by compressing larger values i.e countries with high infant mortality, making the residuals more homoscedastic.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Infant mortality   R-squared:                       0.804
Model:                           OLS   Adj. R-squared:                  0.799
Method:                Least Squares   F-statistic:                     187.1
Date:               Wed, 28 Aug 2024   Prob (F-statistic):           1.62e-63
Time:                       05:27:11   Log-Likelihood:                -125.40
No. Observations:                188   AIC:                             260.8
Df Residuals:                    183   BIC:                             277.0
Df Model:                          4
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                             1.8277      0.164     11.136      0.000       1.504       2.152
Maternal mortality ratio          0.0007      0.000      2.803      0.006       0.000       0.001
Fertility Rate                    0.2982      0.049      6.116      0.000       0.202       0.394
Physicians per thousand          -0.2814      0.029     -9.749      0.000      -0.338      -0.224
Out of pocket health expenditure  0.0107      0.002      5.536      0.000       0.007       0.014
==============================================================================
Omnibus:                        1.755   Durbin-Watson:                   2.058
Prob(Omnibus):                  0.416   Jarque-Bera (JB):                1.366
Skew:                          -0.155   Prob(JB):                        0.505
```

F-statistic: 187.1 and Prob (F-statistic): 1.62e-63: The F-statistic is highly significant, indicating that the overall model is statistically significant, meaning that at least one of the predictors is significantly associated with the log-transformed infant mortality rate.

R-squared: 0.804: This indicates that approximately 80.4% of the variation in the log-transformed infant mortality rate is explained by the independent variables in the model.

3.5 MACHINE LEARNING MODELS:

3.5.1 MULTIPLE LINEAR REGRESSION:

Multiple linear regression is one of the most fundamental statistical models due to its simplicity and interpretability of results. For prediction purposes, linear models can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio, or sparse data (Hastie et al., 2009). In these models, as their name suggests, a predicted (or response) variable is described by a linear combination of predictors. The term "*multiple"* refers to the predictor variables.

the underlying principles of the Ordinary Least-Squares (OLS) regression model. Linear regression is already available in many Python frameworks. Therefore, in practice, one does not need to implement it from scratch to estimate regression coefficients and make predictions. However, our goal here is to gain insight into how these models work and their assumptions to be more effective when tackling future projects.

Linear least-squares
Before diving into equations, I would like to define some notation guidelines.

- Matrices: uppercase italic bold.
- Vectors: lowercase italic bold.
- Scalars: regular italic.

A multiple linear regression model, or an OLS, can be described by the equation below.

$$y_i = \beta_0 + \sum_{j=1}^{M} \beta_j x_{i,j} + \epsilon_i$$

In which y_i is the dependent variable (or response) of observation i, β_0 is the regression intercept, β_j are coefficients associated with decision variables j, x_{ij} is the decision variable j of observation i, and ε is the residual term. In matrix notation, it can be described by

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

In which $\boldsymbol{\beta}$ is a column vector of parameters.

From the OLS, the model is as follows:

log(**Infant Mortality**)=1.8277+0.0007×(**Maternal Mortality Ratio**)+0.2982×(**Fertility Rate**)−0.2814×(**Physicians per Thousand**)+0.0107×(**Out of Pocket Health Expenditure**)

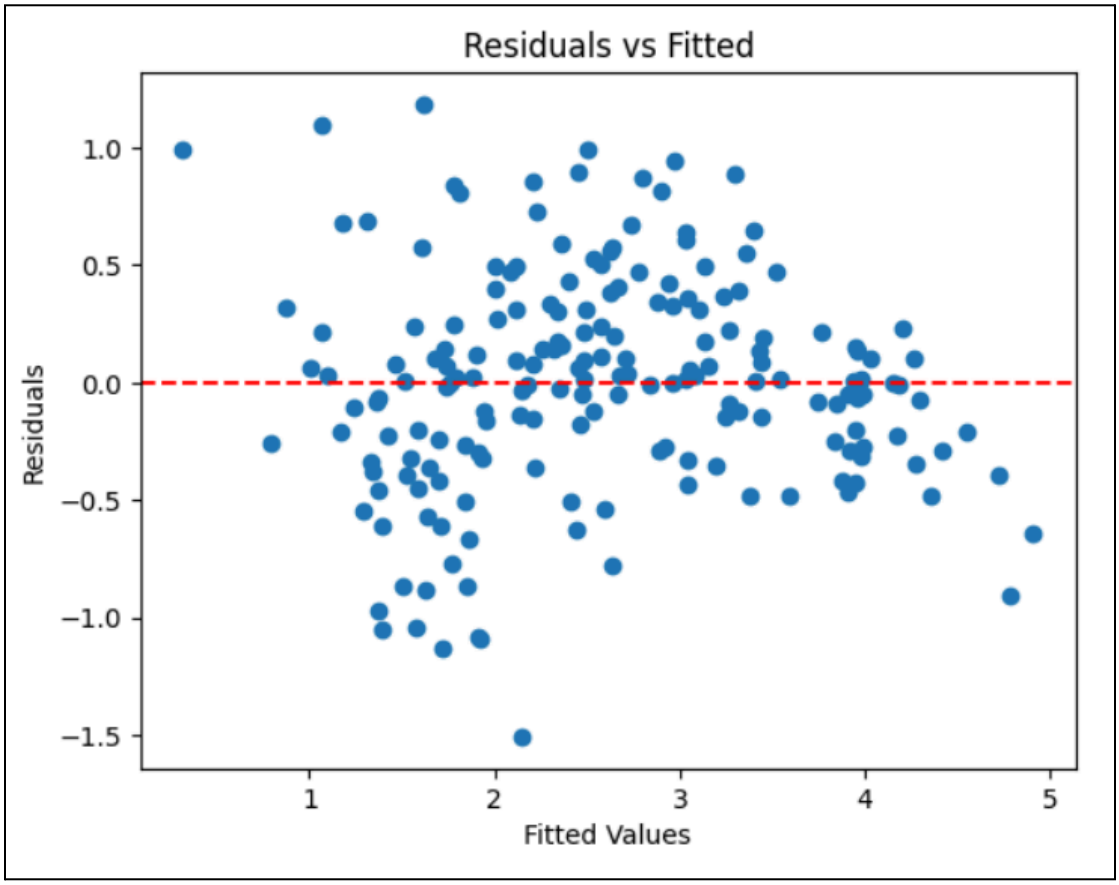
Here, infant mortality has positive coefficients with maternal mortality ratio, fertility rate, and out-of-pocket health expenditure and a negative coefficient with physicians per thousand. Which means that the infant mortality rate can be reduced by reducing the fertility rate and increasing the number of physicians per thousand, i.e., the number of doctors in the country.

To obtain the predicted infant mortality rate (not in log form), multiplying exponentiate on both sides

Predicted Infant Mortality=exp(1.8277+0.0007×(**Maternal Mortality Ratio**)+0.2982×(**Fertility Rate**)−0.2814×(**Physicians per Thousand**)+0.0107×(**Out of Pocket Health Expenditure**))

3.5.1.1 Homoscedasticity

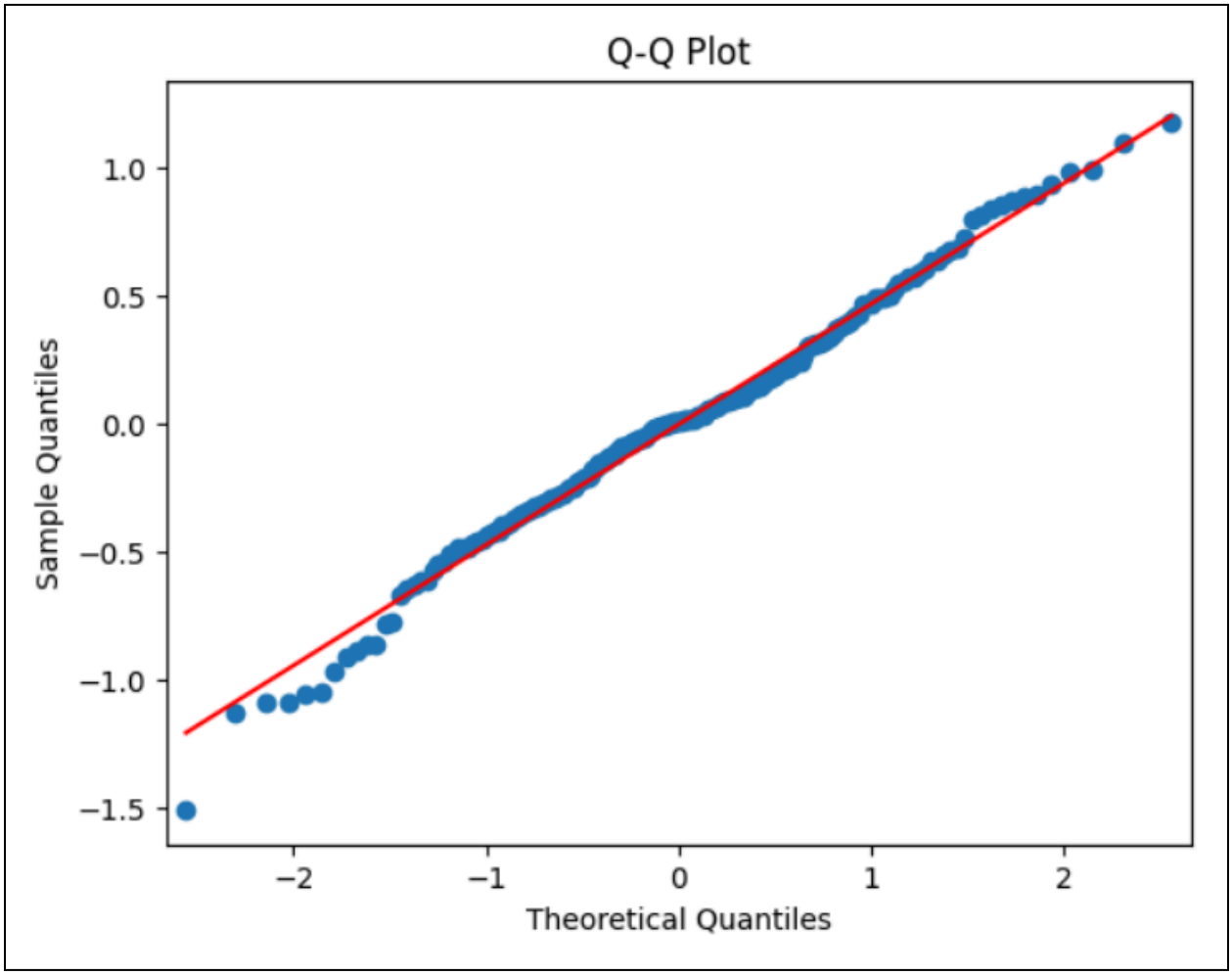
Homoscedasticity is a term used in statistics to describe a situation where the variance of the error term, or the "noise" in the relationship between independent variables and a dependent variable, is constant across all levels of the independent variable.



This plot shows that errors variability are constant across the horizontal line. This indicates there is homoscedasticity condition is met.

3.5.1.2 Normality of errors:

The following Q-Q plot shows the residuals are follow a normal distribution
A Q-Q plot compares the distribution of the residuals to a normal distribution. The residuals follow a normal distribution, the points should lie roughly along the 45-degree line.



For the non - transformed variable the regression for multiple train-test split is:

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
import sklearn as sk
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
print("FOR DIFFERENT TRAIN TEST SPLITS: MULTIPLE REGRESSION")
for i in [0.2,0.25,0.3,0.4]:
  X_train, X_test, y_train, y_test = train_test_split(X.drop(['Life expectancy'], axis = 1), y,
test_size=i, random_state=0)

  # instantiate and fit
  lm2 = LinearRegression()
  lm2.fit(X_train, y_train)
  y_pred = lm2.predict(X_train)

  r2 = r2_score(y_train, y_pred)
  print("---------------\n")
  print(f"R^2 Score for the training set {(1-i)*100}% is : {r2.round(4)}")

  y_pred = lm2.predict(X_test)
  r2 = r2_score(y_test, y_pred)
  print(f"R^2 Score for the testing {(i)*100}%  set is : {r2.round(4)}")
  print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Output:

```
FOR DIFFERENT TRAIN TEST SPLITS: MULTIPLE REGRESSION
---------------

R^2 Score for the training set 80.0% is : 0.8457
R^2 Score for the testing 20.0%  set is : 0.8714
Mean squared error: 82.35338251505438
---------------
```

R^2 Score for the training set 75.0% is : 0.8436
R^2 Score for the testing 25.0% set is : 0.8824
Mean squared error: 68.53965129216951

R^2 Score for the training set 70.0% is : 0.8514
R^2 Score for the testing 30.0% set is : 0.8747
Mean squared error: 64.71131244567077

R^2 Score for the training set 60.0% is : 0.8523
R^2 Score for the testing 40.0% set is : 0.8627
Mean squared error: 66.41585377285963

Train-test	Test R^2	Train R^2
80-20	0.87	0.84
75-25	0.88	0.84
70-30	0.87	0.85
60-40	0.86	0.85

Interpretation: The dataset is divided into four different train-test splits i.e 80% train - 20% test, 75% train - 25% test, 70% train - 30% test, and 60% train - 40% test.

Consistency in Train R^2: The training R^2 scores are very consistent across different splits, ranging from 0.84 to 0.85, indicating that the model is performing similarly on the training data regardless of the split ratio.

Test R^2 Stability: The test R^2 scores also remain stable, with values between 0.86 and 0.88. This suggests that your model is generalizing well to the test data across different train-test splits.
Optimal Split: The 75-25 split has the highest test R^2 (0.88) while maintaining a stable train R^2 (0.84). This could be a good balance between training the model sufficiently and testing its performance.

3.5.2 Decision Tree:

A decision tree is a type of algorithm used in machine learning and data science for both classification and regression tasks. It is a tree-like model that makes decisions by splitting the data into subsets based on certain criteria, usually in a hierarchical manner.

The decision-making process within the tree is intuitive and easy to interpret, making it a popular choice for explaining model predictions. However, decision trees are prone to overfitting, particularly with complex datasets, as they can become too finely tuned to the training data, capturing noise as if it were a pattern. Despite this, decision trees remain a powerful tool, especially when combined with techniques like pruning or when used in ensemble methods such as Random Forests to improve accuracy and reduce overfitting.

1. **Node**: Represents a point where a decision is made.
    - **Root Node**: The topmost node in a tree, representing the entire dataset.
    - **Internal Nodes**: Nodes where the data is split based on a feature.
    - **Leaf Nodes (Terminal Nodes)**: The final nodes that represent the outcome or prediction.
2. **Splitting**: The process of dividing a node into two or more sub-nodes based on a condition (usually a threshold value for continuous data or a category for categorical data).
3. **Branch**: A subsection of the decision tree that connects nodes based on the outcome of the split.
4. **Gini Index/Entropy**: Common criteria used to determine the best split at each node. The goal is to maximize the separation of the classes (for classification) or minimize the variance (for regression).
5. **Pruning**: The process of removing branches that have little importance, which helps prevent overfitting. Pruning can be done by setting a minimum number of samples required to split a node or by setting a maximum depth of the tree.

```
from sklearn.tree import DecisionTreeRegressor, plot_tree

for i in [0.2,0.25,0.3,0.4]:
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=i, random_state=0)
  model = DecisionTreeRegressor(max_features=3,min_samples_leaf=4,max_depth=3)
  model.fit(X_train, y_train)
```

```
y_pred = model.predict(X_train)
r2 = r2_score(y_train, y_pred)
print("---------------\n")
print(f"R^2 Score for the training set {(1-i)*100}% is : {r2.round(4)}")
y_pred = model.predict(X_test)
r2 = r2_score(y_test, y_pred)
print(f"R^2 Score for the testing {(i)*100}%  set is : {r2.round(4)}")
print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Output:

```
R^2 Score for the training set 80.0% is : 0.8572
R^2 Score for the testing 20.0%  set is : 0.7305
Mean squared error: 172.66358681506784
---------------

R^2 Score for the training set 75.0% is : 0.8007
R^2 Score for the testing 25.0%  set is : 0.744
Mean squared error: 149.21363454395362
---------------

R^2 Score for the training set 70.0% is : 0.8617
R^2 Score for the testing 30.0%  set is : 0.6925
Mean squared error: 158.81483104972946
---------------

R^2 Score for the training set 60.0% is : 0.8732
R^2 Score for the testing 40.0%  set is : 0.7293
Mean squared error: 130.90427746828232
```

Train-test	max_features	min_samples_leaf	max_depth	Test R^2	Train R^2
80-20	4	3	4	0.73	0.85
75-25	4	3	4	0.74	0.80
70-30	4	3	4	0.69	0.86
60-40	4	3	4	0.73	0.87

Interpretation:

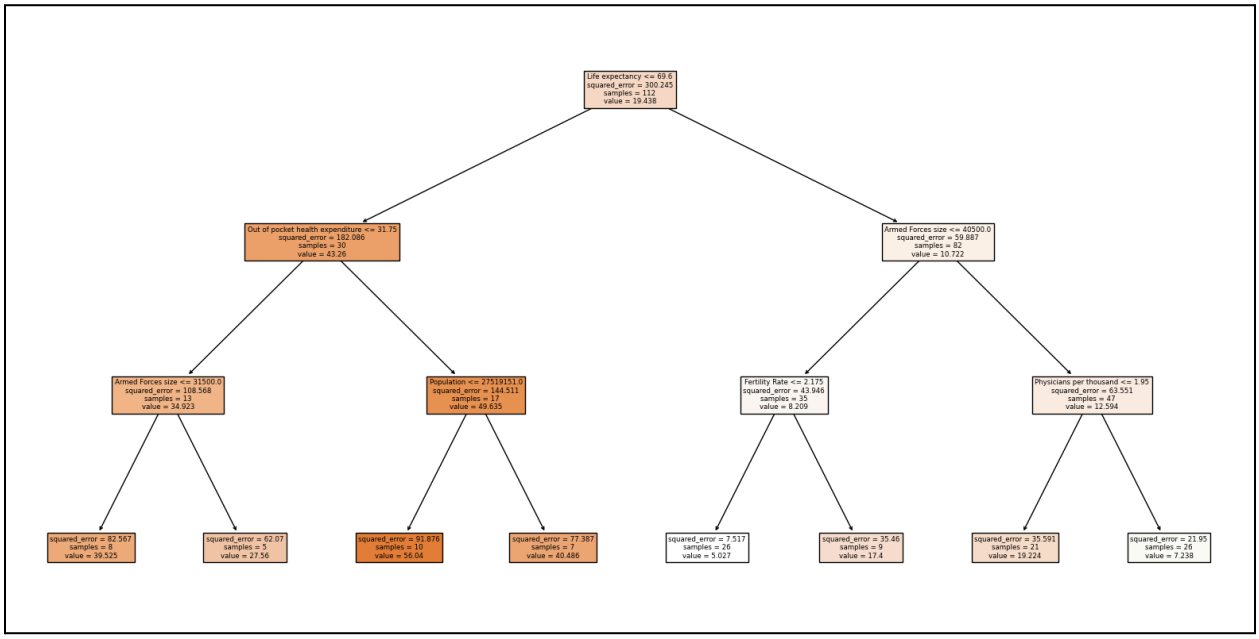
75-25 Split shows the best balance with a decent test R^2 (0.74) and a relatively close train R^2 (0.80), indicating the model is generalizing well.

70-30 Split may indicate some overfitting, as the test R^2 drops more significantly compared to the train R^2.

Overfitting Signs:

The larger discrepancy between train and test R^2 in some splits (e.g., 70-30 split with train R^2 = 0.86 and test R^2 = 0.69) suggests that the model might be overfitting on the training data for those cases.

Decision tree:



3.5.3 Random Forest:

Random Forest is an ensemble machine learning algorithm that builds multiple decision trees and merges them together to produce a more accurate and stable prediction. It is used for both classification and regression tasks.

The core idea behind Random Forest is to combine the predictions of several decision trees, each trained on a random subset of the data, to make the final prediction more robust and less prone to overfitting compared to individual decision trees.

How Random Forest Works:

1. **Bootstrap Sampling**: Random Forest creates multiple decision trees by training each tree on a different subset of the data. These subsets are generated through a process called bootstrap sampling, where each subset is created by randomly selecting data points with replacement from the original dataset.
2. **Random Feature Selection**: When splitting nodes in each decision tree, Random Forest only considers a random subset of features rather than all features. This reduces the correlation between the individual trees and increases diversity among them.
3. **Tree Aggregation**: After all the trees have been constructed, the Random Forest algorithm aggregates their predictions. For classification tasks, it uses majority voting to decide the final class label, and for regression tasks, it averages the predictions from all trees.

```
from sklearn.ensemble import RandomForestRegressor

for i in [0.2,0.25,0.3,0.4]:

 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=i, random_state=0)

 model = RandomForestRegressor(n_estimators=100,

                               max_depth=3,

                               min_samples_leaf=2,

                               min_samples_split=2,

                               max_features=4)
```

```
model.fit(X_train, y_train)

y_pred = model.predict(X_train)

r2 = r2_score(y_train, y_pred)

print("---------------\n")

print(f"R^2 Score for the training set {(1-i)*100}% is : {r2.round(4)}")

y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)

print(f"R^2 Score for the testing {(i)*100}%  set is : {r2.round(4)}")

print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Output:

```
R^2 Score for the training set 80.0% is : 0.9448
R^2 Score for the testing 20.0%  set is : 0.8356
Mean squared error: 105.33856928232758
---------------
R^2 Score for the training set 75.0% is : 0.9454
R^2 Score for the testing 25.0%  set is : 0.8594
Mean squared error: 81.95517532375042
---------------
R^2 Score for the training set 70.0% is : 0.9511
R^2 Score for the testing 30.0%  set is : 0.8603
Mean squared error: 72.14949593202137
---------------
R^2 Score for the training set 60.0% is : 0.955
R^2 Score for the testing 40.0%  set is : 0.8472
Mean squared error: 73.86934966045081
```

Train-test	max_features	min_samples_leaf	max_depth	Test R^2	Train R^2
80-20	4	3	4	0.83	0.94
75-25	4	3	4	0.85	0.94
70-30	4	3	4	0.86	0.95
60-40	4	3	4	0.84	0.95

Interpretation:

Best Split: The 70-30 split has the highest test R^2 (0.86) and a high train R^2 (0.95). However, the performance across the splits is fairly consistent, with only small variations in the test R^2 values.

Generalization and Overfitting: The relatively small difference between train and test R^2 (around 0.09 to 0.11) suggests that while the model is fitting well on the training data, there might be some overfitting. Given the complexity of the Random Forest model, reducing max_depth or increasing min_samples_leaf could help in improving generalization by reducing overfitting.

The Random Forest model shows strong performance across different train-test splits, with fairly consistent test R^2 values

3.5.4 K-Nearest Neighbour(KNN):

K-Nearest Neighbors (KNN) is a simple, yet powerful, machine learning algorithm used for both classification and regression tasks. It is a non-parametric, instance-based learning method, meaning that it makes predictions based on the similarity of the input data to other data points in the training set without making any assumptions about the underlying data distribution.

The choice of 'k' significantly affects the performance of the algorithm. A small 'k' makes the model sensitive to noise, while a large 'k' smoothens the decision boundary but might lead to over-generalization.

The distance metric used to find the nearest neighbors can also influence the algorithm's performance, with Euclidean distance being the most commonly used.

How KNN Works:

1. **Storing Training Data**: KNN stores all the training data points as reference points. This means that no actual learning takes place during the training phase, and all the work is done during the prediction phase.
2. **Distance Calculation**: To make a prediction for a new data point, KNN calculates the distance between this new point and all the points in the training set. Common distance metrics include Euclidean distance, Manhattan distance, or Minkowski distance.
3. **Finding Nearest Neighbors**: Once the distances are calculated, KNN identifies the 'k' nearest neighbors (where 'k' is a user-defined parameter) based on the smallest distances.
4. **Making Predictions**:

- **Classification**: KNN assigns the class label that is most common among the 'k' nearest neighbors.
- **Regression**: KNN predicts the value as the average of the values of the 'k' nearest neighbors.

```
from sklearn.neighbors import KNeighborsRegressor

for i in [0.2,0.25,0.3,0.4]:

  X_train, X_test, y_train, y_test = train_test_split(df_standardized, y_scaled, test_size=i,
random_state=0)
```

```
model = KNeighborsRegressor(n_neighbors=10)

model.fit(X_train, y_train)

y_pred = model.predict(X_train)

r2 = r2_score(y_train, y_pred)

print("---------------\n")

print(f"R^2 Score for the training set {(1-i)*100}% is : {r2.round(4)}")

y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)

print(f"R^2 Score for the testing {(i)*100}%  set is : {r2.round(4)}")

print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Output:

```
R^2 Score for the training set 80.0% is : 0.8531
R^2 Score for the testing 20.0%  set is : 0.8028
Mean squared error: 0.3321298770843237
---------------

R^2 Score for the training set 75.0% is : 0.8488
R^2 Score for the testing 25.0%  set is : 0.8204
Mean squared error: 0.27530792557404
---------------

R^2 Score for the training set 70.0% is : 0.8585
R^2 Score for the testing 30.0%  set is : 0.807
Mean squared error: 0.2621990206455496
---------------

R^2 Score for the training set 60.0% is : 0.8614
R^2 Score for the testing 40.0%  set is : 0.8018
Mean squared error: 0.25211832441689647
```

Train-test	n	Test R^2	Train R^2
80-20	10	0.80	0.85
75-25	10	0.82	0.84
70-30	10	0.80	0.85
60-40	10	0.80	0.86

Interpretation:

The KNN model with n=10 neighbors shows consistent performance across different train-test splits, with stable R^2 values on both the training and testing sets. The model appears to generalize well, with no significant overfitting observed.

The 75-25 split shows the highest test R^2 (0.82), indicating slightly better generalization with this configuration. However, the differences across splits are minor, and all configurations appear to perform similarly.

The gap between the train R^2 and test R^2 is small, which is a good sign. This suggests that the KNN model is generalizing well to the test data and is not significantly overfitting the training data.

3.5.5 Support Vector Machine:

Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm used primarily for classification tasks, though it can also be adapted for regression (as in Support Vector Regression, or SVR). SVM is well-suited for both linear and non-linear classification problems, and it excels in scenarios where the data is not linearly separable.

How SVM Works:

1. **Hyperplane and Decision Boundary**:
    - The core idea of SVM is to find the optimal hyperplane that best separates the data into different classes. In a two-dimensional space, this hyperplane is simply a line, but in higher dimensions, it becomes a flat surface (e.g., a plane in 3D space).
    - The optimal hyperplane is the one that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class. These closest data points are called support vectors.
2. **Support Vectors**:
    - Support vectors are the data points that lie closest to the decision boundary. These points are crucial because they define the position and orientation of the hyperplane. The SVM algorithm tries to position the hyperplane in such a way that the margin is maximized while ensuring these support vectors are correctly classified.

3. **Margin**:

    - The margin is the distance between the hyperplane and the nearest support vectors. A large margin is desirable because it implies that the model is more confident in its classification decisions and is likely to generalize better to unseen data.

4. **Linear vs. Non-Linear SVM**:

    - **Linear SVM**: If the data is linearly separable, SVM can easily find a hyperplane that separates the classes.
    - **Non-Linear SVM**: When the data is not linearly separable, SVM can use the **kernel trick** to project the data into a higher-dimensional space where a linear hyperplane can separate the classes. Common kernels include the radial basis function (RBF), polynomial, and sigmoid kernels.

5. **Kernel Trick**:

   - The kernel trick allows SVM to operate in a higher-dimensional space without explicitly computing the coordinates of the data in that space. Instead, it uses a kernel function to compute the dot product of the data points in the transformed space, allowing for efficient computation even in very high-dimensional spaces.

6. **Regularization Parameter (C)**:

   - The regularization parameter, C, controls the trade-off between maximizing the margin and minimizing the classification error. A high value of C tries to classify all training examples correctly, which may lead to overfitting, while a low value of C allows some misclassifications but aims to achieve a larger margin.

```
from sklearn.svm import SVR

for i in [0.2,0.25,0.3,0.4]:

 X_train, X_test, y_train, y_test = train_test_split(df_standardized, y_scaled, test_size=i, random_state=0)

 model=SVR(C=0.89)

 model.fit(X_train,y_train)

 y_pred = model.predict(X_train)

 r2 = r2_score(y_train, y_pred)

 print("----------\n")

 print(f"R^2 Score for the training set {(1-i)*100}% is : {r2.round(4)}")

 y_pred = model.predict(X_test)

 r2 = r2_score(y_test, y_pred)

 print(f"R^2 Score for the testing {(i)*100}%  set is : {r2.round(4)}")

 print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Output:

R^2 Score for the training set 80.0% is : 0.9565
R^2 Score for the testing 20.0% set is : 0.7348
Mean squared error: 0.4467359195818995

R^2 Score for the training set 75.0% is : 0.9581
R^2 Score for the testing 25.0% set is : 0.7487
Mean squared error: 0.3852336634412157

R^2 Score for the training set 70.0% is : 0.958
R^2 Score for the testing 30.0% set is : 0.7577
Mean squared error: 0.32910725200059543

R^2 Score for the training set 60.0% is : 0.9687
R^2 Score for the testing 40.0% set is : 0.7363
Mean squared error: 0.3353451659713853

train-test	**C**	**Test R^2**	**Train R^2**
80-20	0.89	0.7348	0.9565
75-25	0.89	0.7487	0.9581
70-30	0.89	0.7577	0.958
60-40	0.89	0.7363	0.9687

Interpretation:

The difference between train and test R^2 values (approximately 0.20 to 0.23) suggests that the model might be overfitting. This is especially evident in the 60-40 split, where the train R^2 is extremely high (0.9687) while the test R^2 drops to 0.7363.

The 70-30 split shows the highest test R^2 (0.7577), indicating better generalization compared to the other splits. However, the improvement is marginal, and overfitting remains a concern

3.5.6 XG Boosting:

XGBoost (eXtreme Gradient Boosting) is a highly efficient and powerful machine learning algorithm based on gradient boosting. It is widely used for supervised learning tasks, including classification and regression, and has become one of the most popular algorithms in data science competitions, like those on Kaggle, due to its accuracy, speed, and scalability.

How XGBoost Works:

1. **Gradient Boosting Framework**: XGBoost is built on the gradient boosting framework, which is an ensemble technique that combines the predictions of multiple weak learners (usually decision trees) to create a strong learner. The algorithm works by iteratively adding new models (trees) that correct the errors made by the previous ones.
2. **Additive Model**: XGBoost builds the model in stages. It starts with an initial prediction (usually a constant value), and then fits subsequent models to the residuals (errors) of the previous model to improve accuracy. Each new tree tries to minimize a loss function (e.g., mean squared error) by focusing on the mistakes made by the prior trees.
3. **Regularization**: XGBoost includes a regularization term in its objective function to prevent overfitting. This is one of the key differences between XGBoost and other gradient boosting algorithms. Regularization helps in controlling the complexity of the model, ensuring it generalizes well to unseen data.
4. **Shrinkage (Learning Rate)**: XGBoost uses a learning rate to scale the contribution of each new tree, which helps in making the learning process more robust and reduces the risk of overfitting. A lower learning rate typically means that more trees will be needed, but the model will be more accurate.
5. **Handling Missing Data**: XGBoost is designed to handle missing data efficiently by learning which branch to take in the tree based on the training data.
6. **Parallel Processing**: One of the reasons for XGBoost's popularity is its ability to perform parallel computation, making it faster than other implementations of gradient boosting.

```
from sklearn.model_selection import train_test_split

from xgboost import XGBRegressor

from sklearn.metrics import r2_score

for i in [0.2,0.25,0.3,0.4]:
```

```
 X_train, X_test, y_train, y_test = train_test_split(x_nomulti_colinearity, y, test_size=i,
random_state=0)

 model = XGBRegressor(max_depth=4,

                      n_estimators=100,

                      subsample=0.0310,

                      random_state=0)

 model.fit(X_train, y_train)

 y_pred = model.predict(X_train)

 r2 = r2_score(y_train, y_pred)

 print("---------------\n")

 print(f"R^2 Score for the training set {(1-i)*100}% is : {r2.round(4)}")

 y_pred = model.predict(X_test)

 r2 = r2_score(y_test, y_pred)

 print(f"R^2 Score for the testing {(i)*100}%  set is : {r2.round(4)}")

 print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)).round(3))
```

Output:

```
R^2 Score for the training set 80.0% is : 0.7852
R^2 Score for the testing 20.0%  set is : 0.7122
Mean squared error: 0.485
---------------
R^2 Score for the training set 75.0% is : 0.6542
R^2 Score for the testing 25.0%  set is : 0.6465
Mean squared error: 0.542
---------------
R^2 Score for the training set 70.0% is : 0.7017
R^2 Score for the testing 30.0%  set is : 0.5394
Mean squared error: 0.626
---------------
R^2 Score for the training set 60.0% is : 0.736
R^2 Score for the testing 40.0%  set is : 0.5584
Mean squared error: 0.562
```

Train-test	max_depth	n_estimators	sub_sample	Test R^2	Train R^2
80-20	4	100	0.031	0.71	0.78
75-25	4	100	0.031	0.64	0.65
70-30	4	100	0.031	0.53	0.70
60-40	4	100	0.031	0.55	0.76

Interpretation:

The best performance is observed with the 80-20 split, where the test R^2 is 0.71 and train R^2 is 0.78. This indicates a good balance between training and testing performance.

The 70-30 and 60-40 splits show lower test R^2 values (0.53 and 0.55, respectively), suggesting that the model may struggle more with generalization when more data is reserved for testing.

3.5.7 Bagging:

Bagging, short for Bootstrap Aggregating, is an ensemble learning technique designed to improve the stability and accuracy of machine learning algorithms, particularly those that are prone to high variance, such as decision trees. The main idea behind bagging is to reduce overfitting by combining the predictions of multiple models, each trained on a different random subset of the original dataset.

How Bagging Works:

1. **Bootstrap Sampling**: Bagging begins by creating multiple different subsets of the original training dataset. Each subset is created by randomly selecting samples from the original dataset with replacement (meaning the same data point can appear multiple times in a subset). This process is known as bootstrap sampling.
2. **Model Training**: A separate model is trained on each of these subsets. Typically, bagging is used with decision trees, where each tree is trained independently on a different subset.
3. **Aggregation**:
    - **For Classification**: The final prediction is made by combining the predictions from all the models. In classification tasks, this is usually done by majority voting, where the class label predicted by the most models is chosen as the final output.
    - **For Regression**: In regression tasks, the final prediction is often the average of all the individual model predictions.

```
from sklearn.ensemble import BaggingRegressor

from sklearn.tree import DecisionTreeRegressor

X=x_nomulti_colinearity=data1.drop(['Country','Urban_population','Birth
Rate','Co2-Emissions'],axis=1)

y=data1['Infant mortality']

for i in [0.2,0.25,0.3,0.4]:

 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=i, random_state=0)

 model = BaggingRegressor(

    base_estimator=DecisionTreeRegressor(),
```

```
    n_estimators=99,
    random_state=42,
    max_samples=0.8,
    max_features=4)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
r2 = r2_score(y_test, y_pred)
print(f'\nR² for test ration {i*100}% score: {r2:.4f}')
print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Output:

R² for test ration 20.0% score: 0.8745
Mean squared error: 33.37818726680242

R² for test ration 25.0% score: 0.8621
Mean squared error: 44.922776092507114

R² for test ration 30.0% score: 0.8661
Mean squared error: 48.88321748593158

R² for test ration 40.0% score: 0.8691
Mean squared error: 47.493583361340136

train-test	n_estimator	max_samples	max_features	Test R^2	MSE
80-20	99	0.8	4	0.87	33.37
75-25	99	0.8	4	0.86	44.92
70-30	99	0.8	4	0.86	48.88
60-40	99	0.8	4	0.86	47.49

The R^2 values are consistently high across different splits, ranging from 0.86 to 0.87. This indicates that the Bagging model is performing well and generalizing effectively across different train-test splits.

Summary of the models :

Models/splits	80-20	75-25	70-30	60-40
Multiple Linear Regression	0.87	0.88	0.87	0.86
Random Forest	0.85	0.87	0.86	0.87
Decision Tree	0.73	0.74	0.69	0.72
KNN	0.80	0.82	0.80	0.80
SVR	0.73	0.74	0.75	0.73
Bagging	0.87	0.86	0.86	0.86
XGBoosting	0.71	0.64	0.53	0.55

In summary, Multiple Linear Regression and Bagging consistently performed well across different splits, while XGBoosting showed a decline in performance with smaller training sets. Random Forest and KNN also exhibited robust performance, particularly with the 75-25 and 60-40 splits. Conversely, Decision Tree and SVR displayed more variability in performance depending on the training/testing split.

Chapter - 4

Results and Recommendations

4.1 RESULTS AND FINDINGS:

- With an 88% best fit, Multiple Linear Regression is the strongest model when evaluating various machine learning techniques .

- A significant inverse correlation is found between the two variables when examining the association between infant mortality and life expectancy.

- **From Multiple Linear Regression:**

$$\hat{Y}(\textbf{Predicted Infant Mortality}) = \exp(1.8277 + 0.0007\ (\textbf{Maternal Mortality Ratio}) + 0.2982(\textbf{Fertility Rate}) - 0.2814(\textbf{Physicians per Thousand}) + 0.0107(\textbf{Out of Pocket Health Expenditure}))$$

The reduction of infant mortality can be achieved through having more doctors and longer life expectancy in countries on one hand, while on the other hand, more children per family and higher health costs tend to increase it.

4.2 RECOMMENDATION:

- Reducing infant mortality can be achieved by increasing the number of physicians in the nation.

- In order to lower the infant mortality in the country, government's must focus on raising the numbers of doctors and strengthening the Health care infrastructure.

- Encourage Family Planning and Education: In order to lower high fertility rates, policies that support family planning, reproductive health education, and women's empowerment are needed.

4.3 FUTURE SUGGESTIONS:

- One can conduct a geospatial analysis to identify regions within countries where infant mortality rates are highest and investigate the contributing factors.
- One can evaluate the effectiveness of existing family planning programs in reducing fertility rates and, consequently, infant mortality various countries.
- Could compare the effectiveness of healthcare policies across different countries in reducing infant mortality rates.
- Could conduct an economic evaluation of the cost-effectiveness of various strategies to reduce infant mortality

REFERENCES:

- Factors Affecting Infant Mortality Rate in India: An Analysis of Indian States | SpringerLink https://link.springer.com/chapter/10.1007/978-3-319-47952-1_57

- (PDF) INFANT AND CHILD MORTALITY AND ITS MAJOR DETERMINANTS: A CASE STUDY OF UTTARAKHAND STATE, INDIA (researchgate.net) https://www.researchgate.net/publication/365373615
- High infant and child mortality rates in Orissa: An assessment of major reasons | Request PDF (researchgate.net) https://www.researchgate.net/publication/227530987_High_infant_and_child_mortality_rates_in_Orissa_An_assessment_of_major_reasons
- Predictive modeling of infant mortality | Data Mining and Knowledge Discovery (springer.com) https://link.springer.com/article/10.1007/s10618-020-00728-2
- The Differential Effect of Foreign-Born Status on Low Birth Weight by Race/Ethnicity and Education | Pediatrics | American Academy of Pediatrics (aap.org) https://publications.aap.org/pediatrics/article-abstract/115/1/e20/66937/The-Differential-Effect-of-Foreign-Born-Status-on
- Findings of National Family Health Survey: Regional Analysis on JSTOR https://www.jstor.org/stable/4408531
- Determinants of infant mortality in Pakistan: evidence from Pakistan Demographic and Health Survey 2017–18 | Journal of Public Health (springer.com) https://link.springer.com/article/10.1007/s10389-019-01175-0

APPENDIX

Sample Data Table.

First 10 rows 27columns from the dataset .

Country	Density (P/Km2)	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Calling Code	Co2-Emissions	CPI
Austria	109	32.40%	83,871	21,000	9.7	43	61,448	118.06
Australia	3	48.20%	77,41,220	58,000	12.6	61	3,75,908	119.8
Armenia	104	58.90%	29,743	49,000	13.99	374	5,156	129.18
Argentina	17	54.30%	27,80,400	1,05,000	17.02	54	2,01,348	232.75
Antigua and Barbuda	223	20.50%	443	0	15.33	1	557	113.81
Angola	26	47.50%	12,46,700	1,17,000	40.73	244	34,693	261.73
Andorra	164	40.00%	468		7.2	376	469	
Algeria	18	17.40%	23,81,741	3,17,000	24.28	213	1,50,006	151.36
Albania	105	43.10%	28,748	9,000	11.78	355	4,536	119.05
Afghanistan	60	58.10%	6,52,230	3,23,000	32.49	93	8,672	149.9

CPI Change (%)	Fertility Rate	Forested Area (%)	Gasoline Price	GDP	Gross primary education enrollment (%)	Gross tertiary education enrollment (%)	Infant mortality	Life expectancy
2.30%	4.47	2.10%	$0.70	$19,101,353,833	104.00%	9.70%	47.9	64.5
1.40%	1.62	28.10 %	$1.36	$15,278,077,447	107.00%	55.00%	7.8	78.5

2.00%	3.02	0.80%	$0.28	$169,988,236,398	109.90%	51.40%	20.1	76.7
	1.27	34.00 %	$1.51	$3,154,057,987	106.40%		2.7	
17.10 %	5.52	46.30 %	$0.97	$94,635,415,870	113.50%	9.30%	51.6	60.8
1.20%	1.99	22.30 %	$0.99	$1,727,759,259	105.00%	24.80%	5	76.9
53.50 %	2.26	9.80%	$1.10	$449,663,446,954	109.70%	90.00%	8.8	76.5
1.40%	1.76	11.70 %	$0.77	$13,672,802,158	92.70%	54.60%	11	74.9
1.60%	1.74	16.30 %	$0.93	$1,392,680,589,329	100.30%	113.10%	3.1	82.7
1.50%	1.47	46.90 %	$1.20	$446,314,739,528	103.10%	85.10%	2.9	81.6

Maternal mortality ratio	Minimum wage	Out of pocket health expenditure	Physicians per thousand	Population	Population:Labor force participation (%)	Tax revenue (%)	Total tax rate	Unemployment rate	Urban_population
638	$0.43	78.40%	0.28	3,80,41,754	48.90%	9.30 %	71.40%	11.12%	97,97,273
15	$1.12	56.90%	1.2	28,54,191	55.70%	18.60%	36.60%	12.33%	17,47,593
112	$0.95	28.10%	1.72	4,30,53,054	41.20%	37.20%	66.10%	11.70%	3,15,10,100
	$6.63	36.40%	3.33	77,142					67,873
241	$0.71	33.40%	0.21	3,18,25,295	77.50%	9.20 %	49.10%	6.89%	2,10,61,025

42	$3.04	24.30%	2.76	97,118		16.50%	43.00%		23,800
39	$3.35	17.60%	3.96	4,49,38,712	61.30%	10.10%	106.30%	9.79%	4,13,39,571
26	$0.66	81.60%	4.4	29,57,731	55.60%	20.90%	22.60%	16.99%	18,69,848
6	$13.59	19.60%	3.68	2,57,66,605	65.50%	23.00%	47.40%	5.27%	2,18,44,756
5		17.90%	5.17	88,77,067	60.70%	25.40%	51.40%	4.67%	51,94,416