Predicting Infant Mortality:
A Global Analysis of 2023 Data

Aurora Degree & P.G College | M. Sc Applied Statistics

Ruchitha | Shravanthika | Ramya | Madirai | Deepika

Abstract

This study explores various machine learning models to predict infant mortality rates across different countries.

We evaluate multiple algorithms, including linear regression, random forests, support vector machines and more.

Content

INTRODUCTION

The infant mortality rate (IMR) is often regarded as a barometer for overall welfare of a community or country.

It refers to the number of deaths per 1000 live births within the year

OBJECTIVES

Objective 1

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

Objective 2

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

LITERATURE REVIEW

1. In a study by Zakir Hossain, Enamul Kabir, and Rumana Rois (November 2021), machine learning algorithms were used to identify predictors of infant mortality in Bangladesh using data from the 2017–18 Demographic and Health Survey. Key features like age at first marriage, birth interval, and education were found significant. Among various ML models, random forest performed best with an accuracy of 89.3% and an AUC of 0.6613. The findings can guide policy-makers and public health interventions aimed at reducing infant mortality

2. Leonardo Matsuno da Frota et al. (March 2024) used machine learning to predict infant mortality in Brazil from 2.9 million Brazil's Unique Health System(SUS) data points. Survival Support Vector Machines and Extreme Gradient Boosting achieved high accuracy (c-index: 0.84 and 0.83). The Cox model also performed well (c-index: 0.83), highlighting machine learning's role in enhancing mortality predictions and informing health policy.

About the Dataset

- ## Data Source:

The Dataset for this project is a secondary data taken from Kaggle repository.

https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023/data

- ## Description:

The Dataset contains 195 countries as rows and 27 features.

some of the features include Country, Population, Land Area, Agriculture Land, Co2- Emissions , CPI, Fertility Rate, Forested Area, Gasoline Price, GDP, Life expectancy, maternal mortality ratio, minimum wage, Density, Longitude, Latitude.

Data overview

	Country	Density\n(P/Km2)	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Co2-Emissions	CPI	CPI Change (%)	Fertility Rate	Forested Area (%)	Gasoline Price	GDP	Gross primary education enrollment (%)	Gross tertiary education enrollment (%)	Infant mortality	Life expectancy	Maternal mortality ratio	Minimum wage	Out of pocket health expenditure
0	Afghanistan	60	58.10%	652,230	323,000	32.49	8,672	149.9	2.30%	4.47	2.10%	$0.70	$19,101,353,833	104.00%	9.70%	47.9	64.5	638.0	$0.43	78.40%
1	Albania	105	43.10%	28,748	9,000	11.78	4,536	119.05	1.40%	1.62	28.10%	$1.36	$15,278,077,447	107.00%	55.00%	7.8	78.5	15.0	$1.12	56.90%
2	Algeria	18	17.40%	2,381,741	317,000	24.28	150,006	151.36	2.00%	3.02	0.80%	$0.28	$169,988,236,398	109.90%	51.40%	20.1	76.7	112.0	$0.95	28.10%
3	Andorra	164	40.00%	468	NaN	7.20	469	NaN	NaN	1.27	34.00%	$1.51	$3,154,057,987	106.40%	NaN	2.7	NaN	NaN	$6.63	36.40%
4	Angola	26	47.50%	1,246,700	117,000	40.73	34,693	261.73	17.10%	5.52	46.30%	$0.97	$94,635,415,870	113.50%	9.30%	51.6	60.8	241.0	$0.71	33.40%
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
190	Venezuela	32	24.50%	912,050	343,000	17.88	164,175	2,740.27	254.90%	2.27	52.70%	$0.00	$482,359,318,768	97.20%	79.30%	21.4	72.1	125.0	$0.01	45.80%
191	Vietnam	314	39.30%	331,210	522,000	16.75	192,668	163.52	2.80%	2.05	48.10%	$0.80	$261,921,244,843	110.60%	28.50%	16.5	75.3	43.0	$0.73	43.50%
192	Yemen	56	44.60%	527,968	40,000	30.45	10,609	157.58	8.10%	3.79	1.00%	$0.92	$26,914,402,224	93.60%	10.20%	42.9	66.1	164.0	NaN	81.00%
193	Zambia	25	32.10%	752,618	16,000	36.19	5,141	212.31	9.20%	4.63	65.20%	$1.40	$23,064,722,446	98.70%	4.10%	40.4	63.5	213.0	$0.24	27.50%
194	Zimbabwe	38	41.90%	390,757	51,000	30.68	10,983	105.51	0.90%	3.62	35.50%	$1.34	$21,440,758,800	109.90%	10.00%	33.9	61.2	458.0	NaN	25.80%

195 rows × 27 columns

Data Cleaning

Handling Missing Data

feature wise missing percentage

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
Total tax rate	2.660
Gross tertiary education enrollment (%)	2.660
Out of pocket health expenditure	2.128
Physicians per thousand	1.064
Infant mortality	0.532
Life expectancy	0.532
Gross primary education enrollment (%)	0.532
Co2-Emissions	0.532
Agricultural Land(%)	0.532
Forested Area (%)	0.532
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
Urban_population	0.000

dropped country with more than 90% missing values

Country	null_percentage
Eswatini	74.074
Vatican City	85.185
Monaco	55.556
Nauru	81.481
North Macedonia	77.778
Palestinian National Authority	92.593
Tuvalu	51.852

** Please refer to Appendix page 31 for a data descriptions*

Exploratory Data Analysis

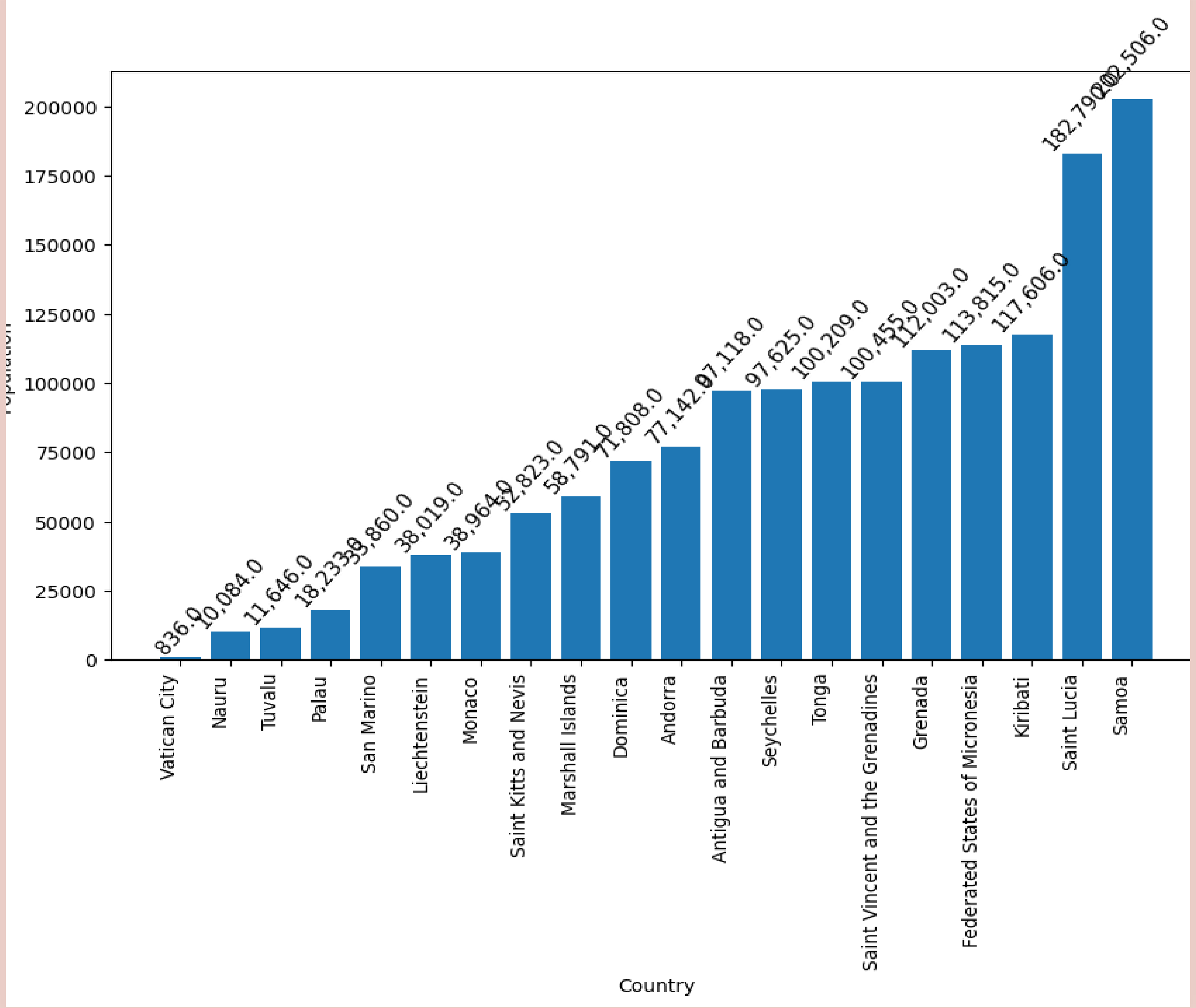
Correlation Matrix

	Density (P/Km2)	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Co2-Emissions	CPI	CPI Change (%)	Fertility Rate	Forested Area (%)	Gasoline Price	GDP	Gross primary education enrollment (%)	Gross tertiary education enrollment (%)	Infant mortality	Life expectancy	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	GDP per capita
Density (P/Km2)	1	-0.035	-0.053	0.00017	-0.15	-0.0046	-0.014	-0.017	-0.049	-0.033	0.22	-0.018	0.0032	0.036	-0.11	0.055	-0.03	0.26	-0.092	0.21	-0.018	0.021	-0.012	-0.027	-0.033	-0.022	0.06
Agricultural Land(%)	-0.035	1	-0.031	0.029	0.2	0.061	-0.0014	-0.007	0.18	-0.43	0.044	0.052	-0.046	-0.11	0.2	-0.24	0.2	-0.023	0.13	-0.036	0.12	-0.11	-0.052	0.16	0.06	0.1	-9.9e-05
Land Area(Km2)	-0.053	-0.031	1	0.55	-0.066	0.59	0.021	0.034	-0.065	-0.014	-0.19	0.55	0.0061	0.22	-0.066	0.055	-0.054	0.12	-0.014	0.072	0.45	-0.007	-0.17	0.09	0.043	0.55	-0.026
Armed Forces size	0.00017	0.029	0.55	1	-0.14	0.74	0.039	0.059	-0.14	-0.018	-0.15	0.6	0.057	0.11	-0.076	0.074	-0.1	-0.019	0.14	0.013	0.87	-0.08	-0.17	0.11	-0.024	0.85	-1.4e-05
Birth Rate	-0.15	0.2	-0.066	-0.14	1	-0.15	0.14	0.1	0.98	-0.073	-0.2	-0.18	-0.082	-0.71	0.86	-0.87	0.75	-0.45	0.25	-0.73	-0.053	0.17	-0.35	0.19	-0.042	-0.11	-0.00032
Co2-Emissions	-0.0046	0.061	0.59	0.74	-0.15	1	-0.017	-0.0042	-0.14	-0.029	-0.067	0.92	0.0024	0.16	-0.12	0.12	-0.11	0.08	-0.033	0.06	0.81	-0.014	-0.14	0.061	0.0065	0.93	5e-05
CPI	-0.014	-0.0014	0.021	0.039	0.14	-0.017	1	0.89	0.14	0.0022	-0.25	-0.027	-0.17	0.012	0.18	-0.18	0.3	-0.07	0.16	-0.037	-0.0049	0.021	-0.054	0.042	0.12	-0.0093	-4.2e-05
CPI Change (%)	-0.017	-0.007	0.034	0.059	0.1	-0.0042	0.89	1	0.094	0.012	-0.25	-0.011	-0.14	0.052	0.15	-0.15	0.22	-0.082	0.13	-0.029	0.0056	-0.016	-0.075	0.11	0.11	0.0073	-3.9e-05
Fertility Rate	-0.049	0.18	-0.065	-0.14	0.98	-0.14	0.14	0.094	1	-0.063	-0.12	-0.16	-0.14	-0.67	0.85	-0.85	0.77	-0.37	0.2	-0.66	-0.055	0.15	-0.35	0.21	-0.071	-0.1	-0.00026
Forested Area (%)	-0.033	-0.43	-0.014	-0.018	-0.073	-0.029	0.0022	0.012	-0.063	1	0.13	-0.00056	0.14	-0.01	-0.055	0.0058	-0.06	-0.028	-0.24	-0.081	-0.056	0.14	0.11	0.033	-0.077	-0.035	-2.7e-06
Gasoline Price	0.22	0.044	-0.19	-0.15	-0.2	-0.067	-0.25	-0.25	-0.12	0.13	1	-0.02	-0.069	0.15	-0.16	0.23	-0.066	0.35	-0.27	0.24	-0.071	-0.01	0.33	0.051	-0.039	-0.074	0.00024
GDP	-0.018	0.052	0.55	0.6	-0.18	0.92	-0.027	-0.011	-0.16	-0.00056	-0.02	1	-0.0041	0.21	-0.15	0.18	-0.12	0.21	-0.1	0.092	0.63	-0.007	-0.12	0.051	0.031	0.78	0.0001
Gross primary education enrollment (%)	0.0032	-0.046	0.0061	0.057	-0.082	0.0024	-0.17	-0.14	-0.14	0.14	-0.069	-0.0041	1	-0.0089	-0.14	0.094	-0.17	-0.05	-0.16	-0.026	0.04	0.22	0.16	-0.062	-0.034	0.029	1.9e-06
Gross tertiary education enrollment (%)	0.036	-0.11	0.22	0.11	-0.71	0.16	0.012	0.052	-0.67	-0.01	0.15	0.21	-0.0089	1	-0.7	0.71	-0.56	0.46	-0.17	0.7	0.025	-0.15	0.27	-0.032	0.037	0.097	0.00025
Infant mortality	-0.11	0.2	-0.066	-0.076	0.86	-0.12	0.18	0.15	0.85	-0.055	-0.16	-0.15	-0.14	-0.7	1	-0.92	0.87	-0.43	0.34	-0.68	0.0047	0.12	-0.36	0.23	0.0017	-0.057	-0.00027
Life expectancy	0.055	-0.24	0.055	0.074	-0.87	0.12	-0.18	-0.15	-0.85	0.0058	0.23	0.18	0.094	0.71	-0.92	1	-0.82	0.5	-0.31	0.68	0.0088	-0.15	0.34	-0.19	-0.039	0.07	0.00033
Maternal mortality ratio	-0.03	0.2	-0.054	-0.1	0.75	-0.11	0.3	0.22	0.77	-0.06	-0.066	-0.12	-0.17	-0.56	0.87	-0.82	1	-0.3	0.27	-0.53	-0.024	0.19	-0.32	0.2	-0.037	-0.067	-0.00018
Minimum wage	0.26	-0.023	0.12	-0.019	-0.45	0.08	-0.07	-0.082	-0.37	-0.028	0.35	0.21	-0.05	0.46	-0.43	0.5	-0.3	1	-0.37	0.41	-0.047	-0.046	0.24	-0.063	-0.044	0.0085	0.00039
Out of pocket health expenditure	-0.092	0.13	-0.014	0.14	0.25	-0.033	0.16	0.13	0.2	-0.24	-0.27	-0.1	-0.16	-0.17	0.34	-0.31	0.27	-0.37	1	-0.18	0.13	-0.19	-0.32	0.24	0.027	0.064	-0.0002
Physicians per thousand	0.21	-0.036	0.072	0.013	-0.73	0.06	-0.037	-0.029	-0.66	-0.081	0.24	0.092	-0.026	0.7	-0.68	0.68	-0.53	0.41	-0.18	1	-0.049	-0.14	0.26	-0.1	0.0022	0.00046	0.00031
Population	-0.018	0.12	0.45	0.87	-0.053	0.81	-0.0049	0.0056	-0.055	-0.056	-0.071	0.63	0.04	0.025	0.0047	0.0088	-0.024	-0.047	0.13	-0.049	1	-0.057	-0.19	0.082	-0.042	0.95	-0.02
Population: Labor force participation (%)	0.021	-0.11	-0.007	-0.08	0.17	-0.014	0.021	-0.016	0.15	0.14	-0.01	-0.007	0.22	-0.15	0.12	-0.15	0.19	-0.046	-0.19	-0.14	-0.057	1	-0.11	-0.17	-0.43	-0.049	2.1e-05
Tax revenue (%)	-0.012	-0.052	-0.17	-0.17	-0.35	-0.14	-0.054	-0.075	-0.35	0.11	0.33	-0.12	0.16	0.27	-0.36	0.34	-0.32	0.24	-0.32	0.26	-0.19	-0.11	1	-0.081	0.21	-0.18	9.9e-05
Total tax rate	-0.027	0.16	0.09	0.11	0.19	0.061	0.042	0.11	0.21	0.033	0.051	0.051	-0.062	-0.032	0.23	-0.19	0.2	-0.063	0.24	-0.1	0.082	-0.17	-0.081	1	-0.034	0.099	-9.9e-05
Unemployment rate	-0.033	0.06	0.043	-0.024	-0.042	0.0065	0.12	0.11	-0.071	-0.077	-0.039	0.031	-0.034	0.037	0.0017	-0.039	-0.037	-0.044	0.027	0.0022	-0.042	-0.43	0.21	-0.034	1	-0.015	-5.8e-05
Urban_population	-0.022	0.1	0.55	0.85	-0.11	0.93	-0.0093	0.0073	-0.1	-0.035	-0.074	0.78	0.029	0.097	-0.057	0.07	-0.067	0.0085	0.064	0.00046	0.95	-0.049	-0.18	0.099	-0.015	1	1.2e-06
GDP per capita	0.06	-9.9e-05	-0.026	-1.4e-05	-0.00032	5e-05	-4.2e-05	-3.9e-05	-0.00026	-2.7e-06	0.00024	0.0001	1.9e-06	0.00025	-0.00027	0.00033	-0.00018	0.00039	-0.0002	0.00031	-0.02	2.1e-05	9.9e-05	-9.9e-05	-5.8e-05	1.2e-06	1

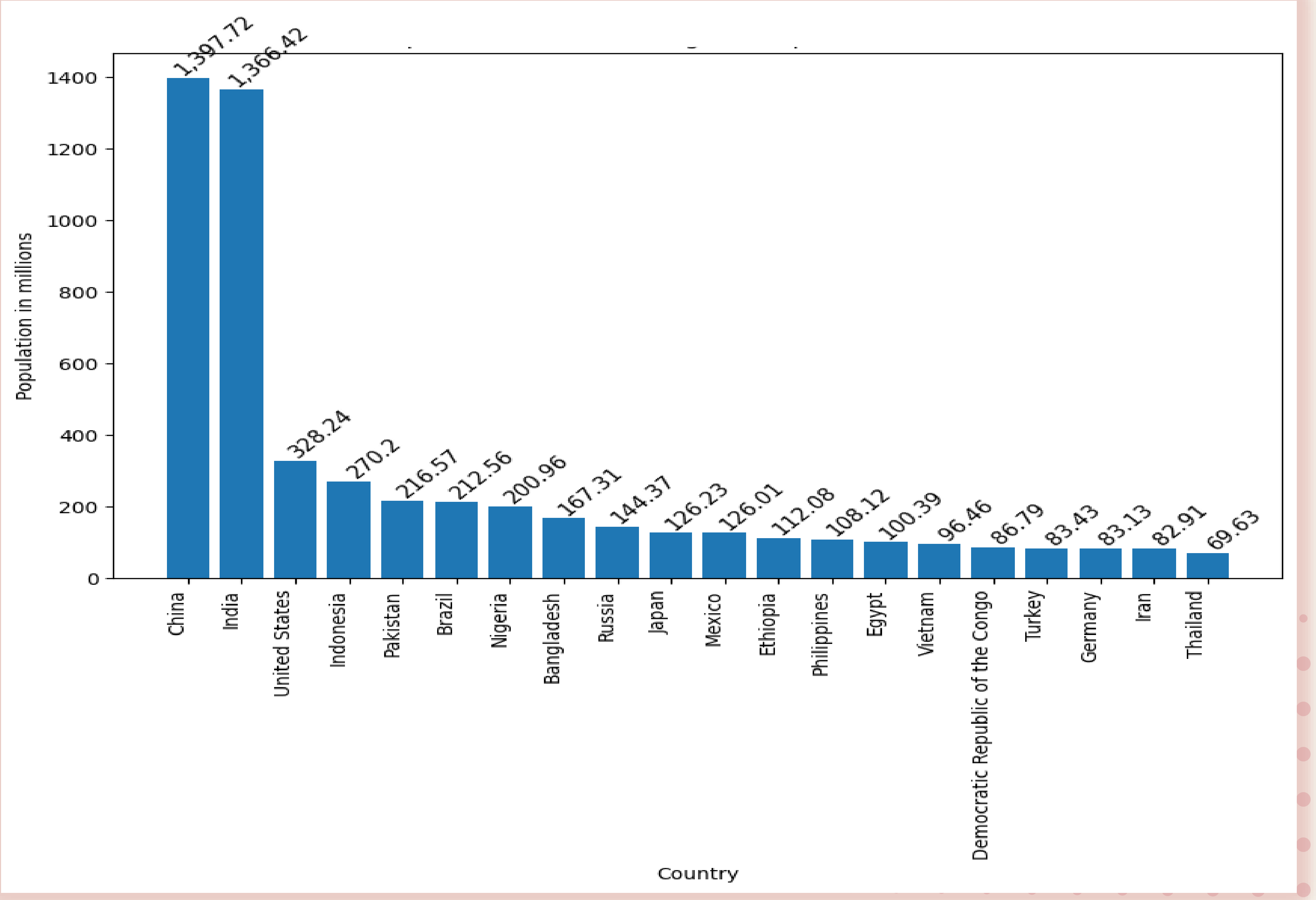
Bar plot

Countries vs Population

Bottom 20 Countries with lower Population

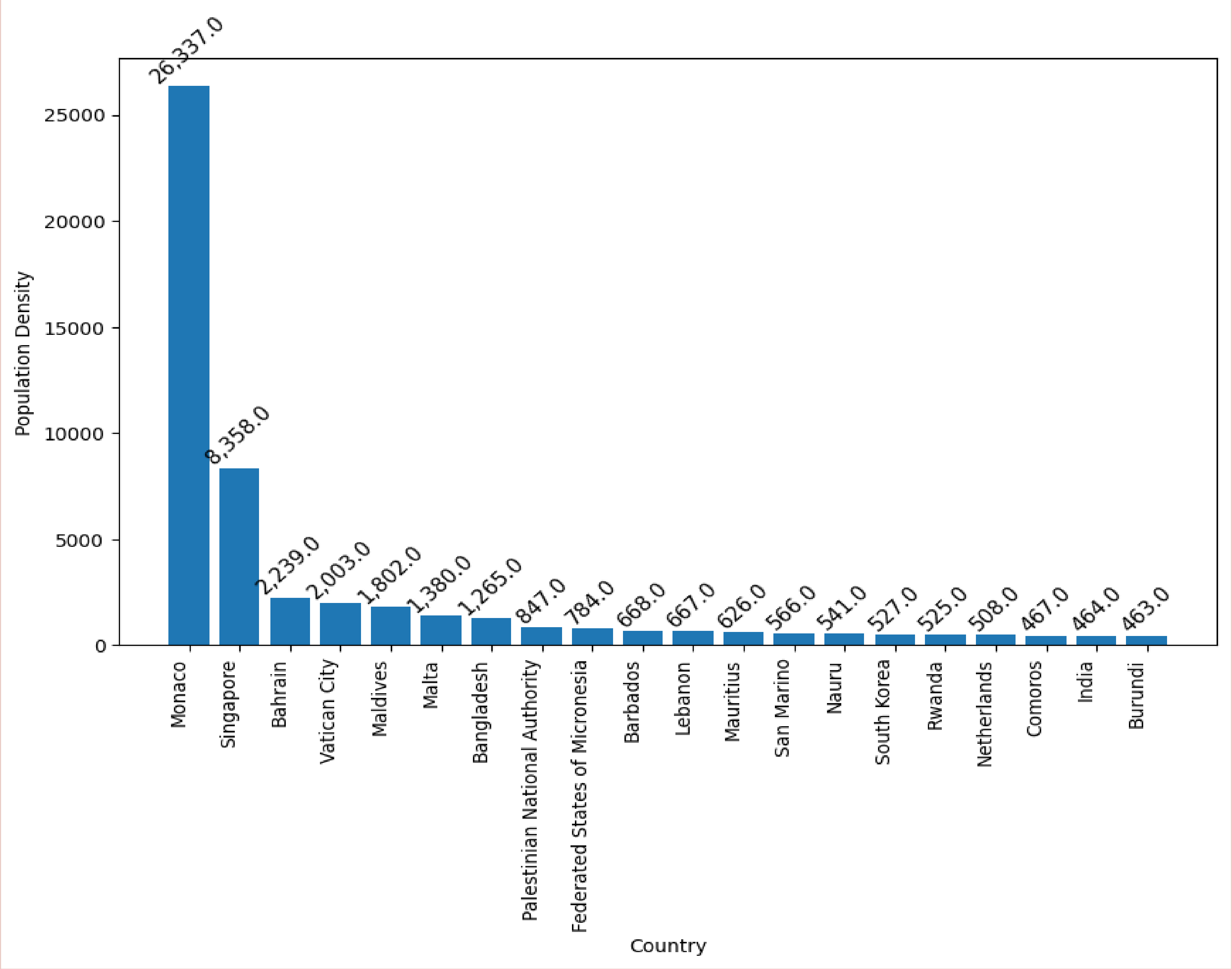


Top 20 Countries with High Population (in M)

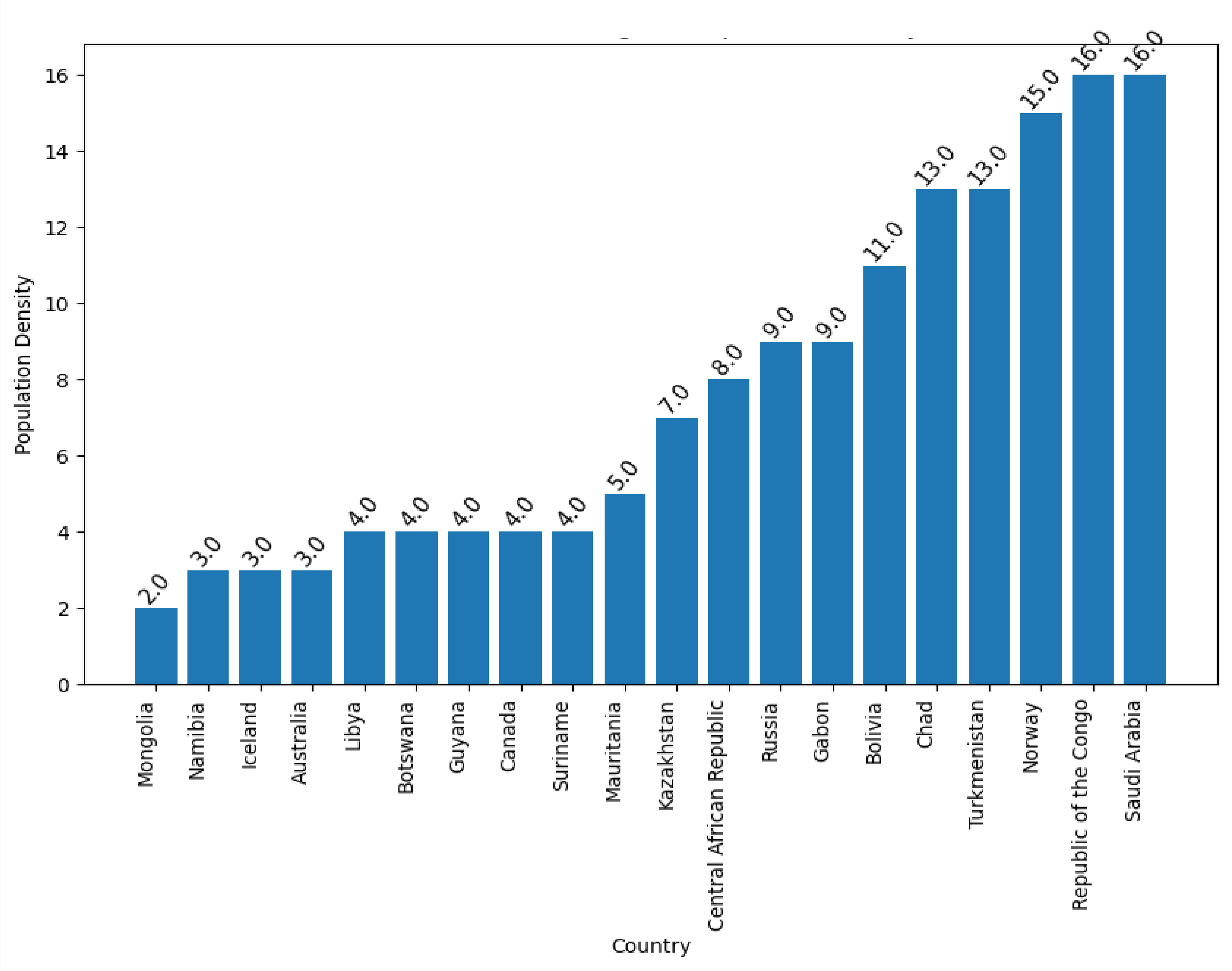


Countries with Population Density

Top 20 Countries with Highest Population Density

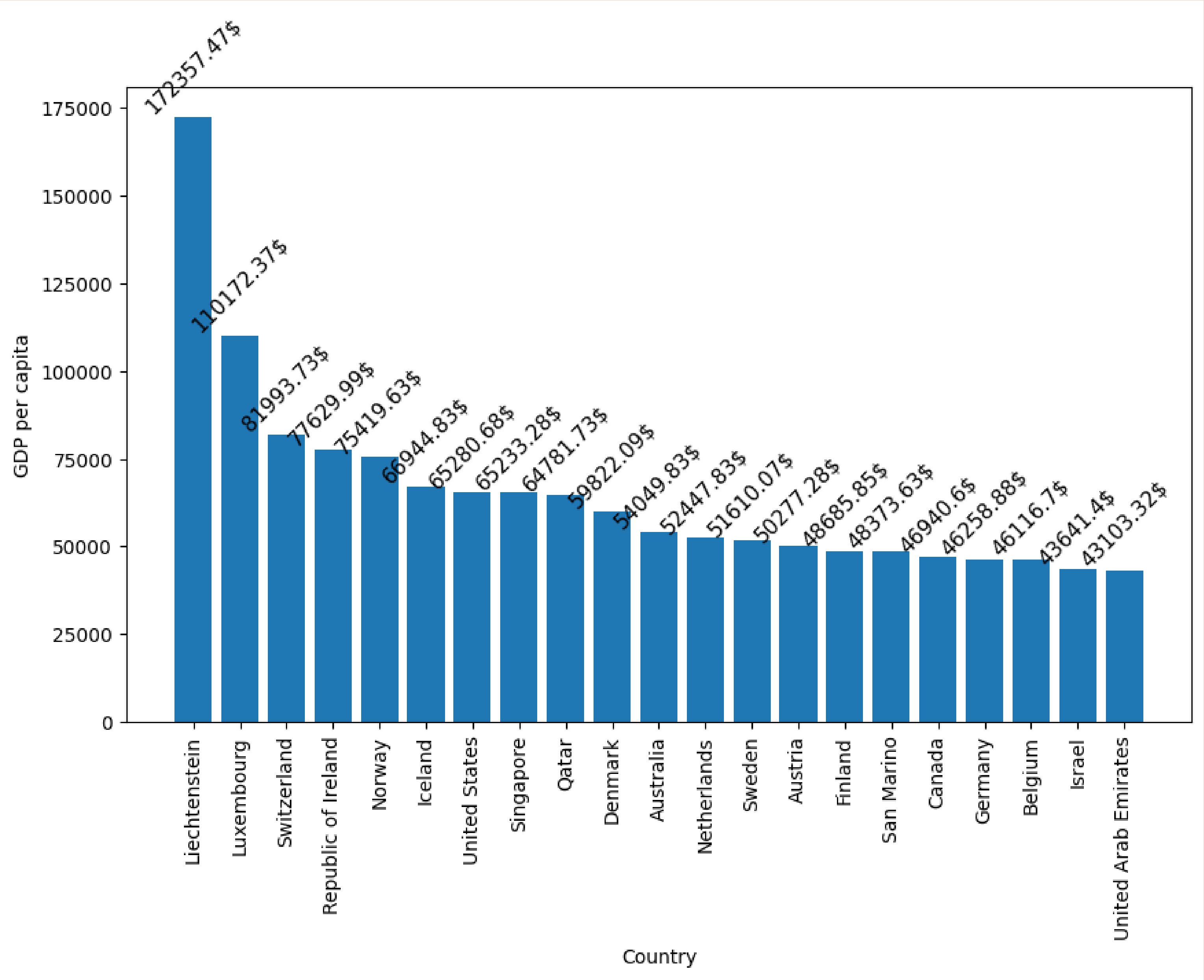


Bottom 20 Countries with Lowest population Density

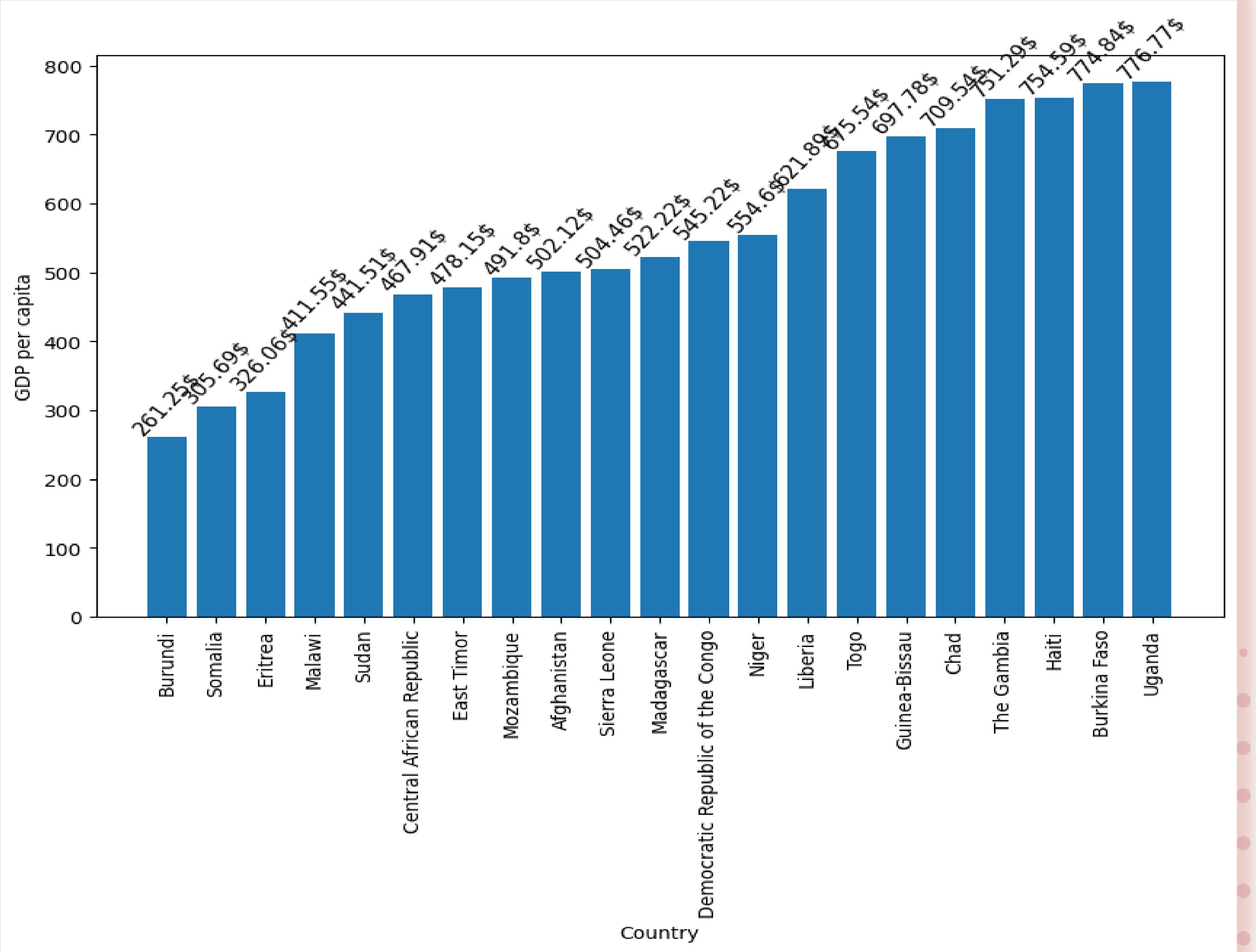


Countries with GDP Per Capita

Top 20 Countries with GDP per capita

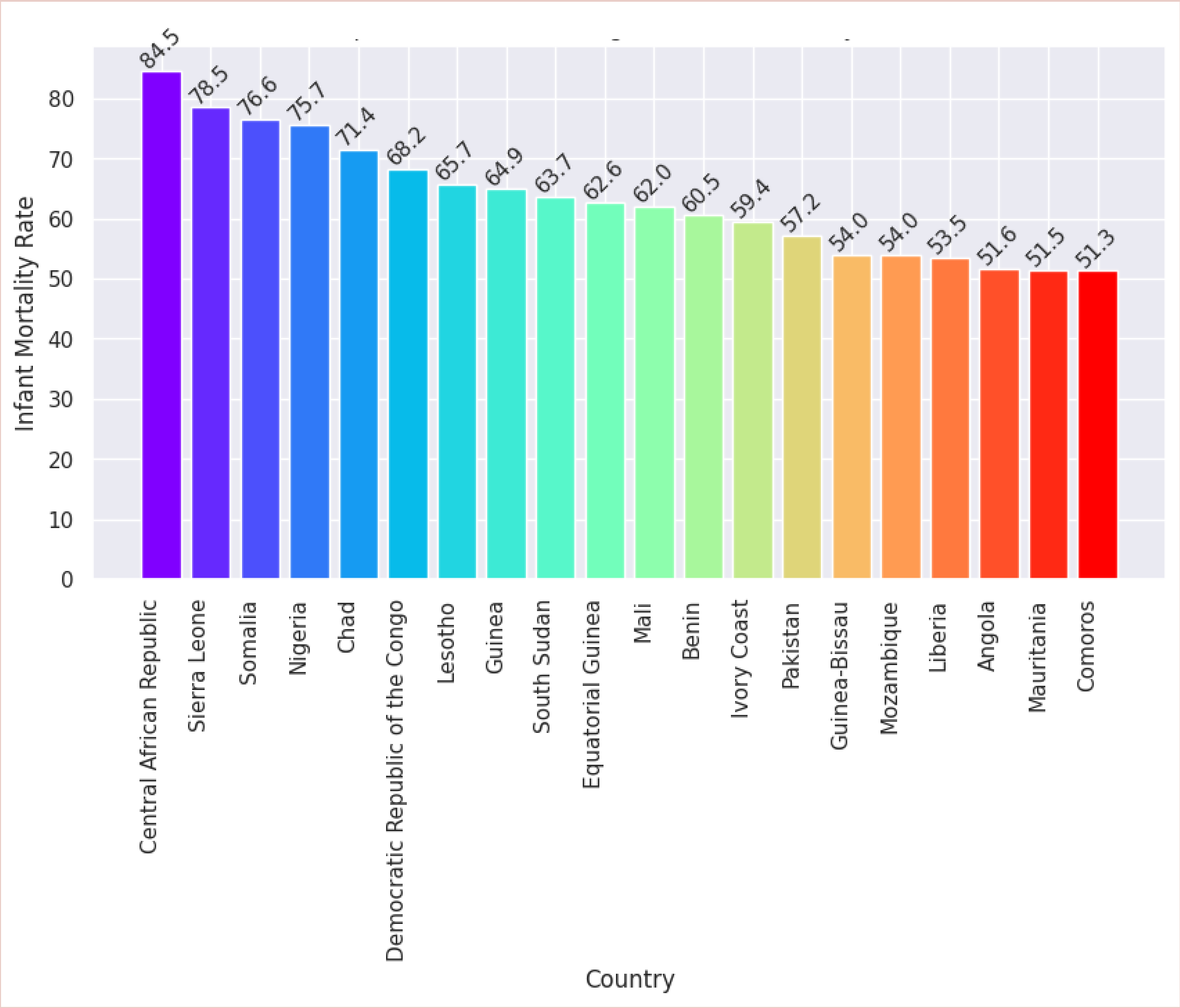


Bottom 20 Countries with GDP per capita

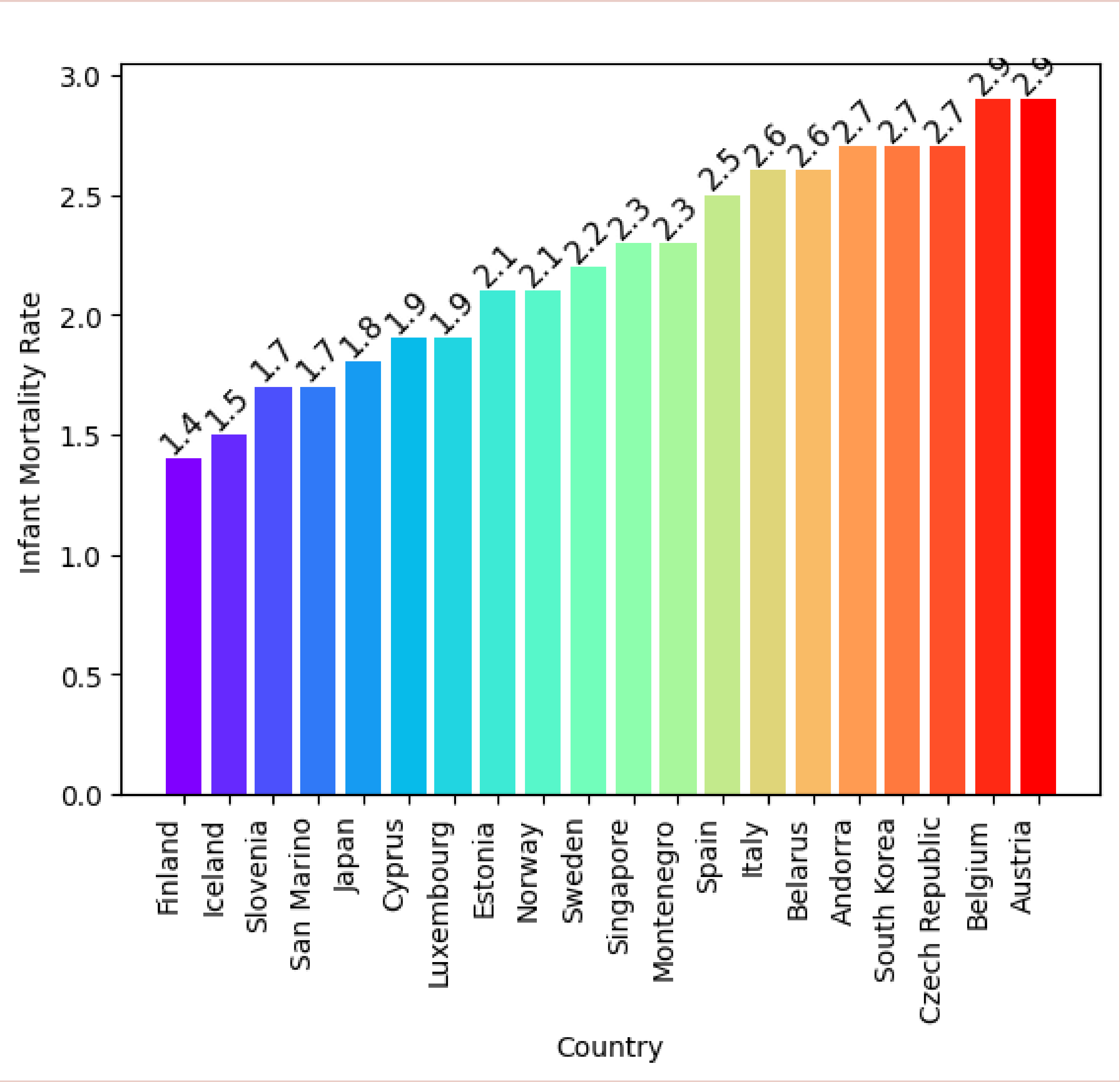


Country with Infant Mortality

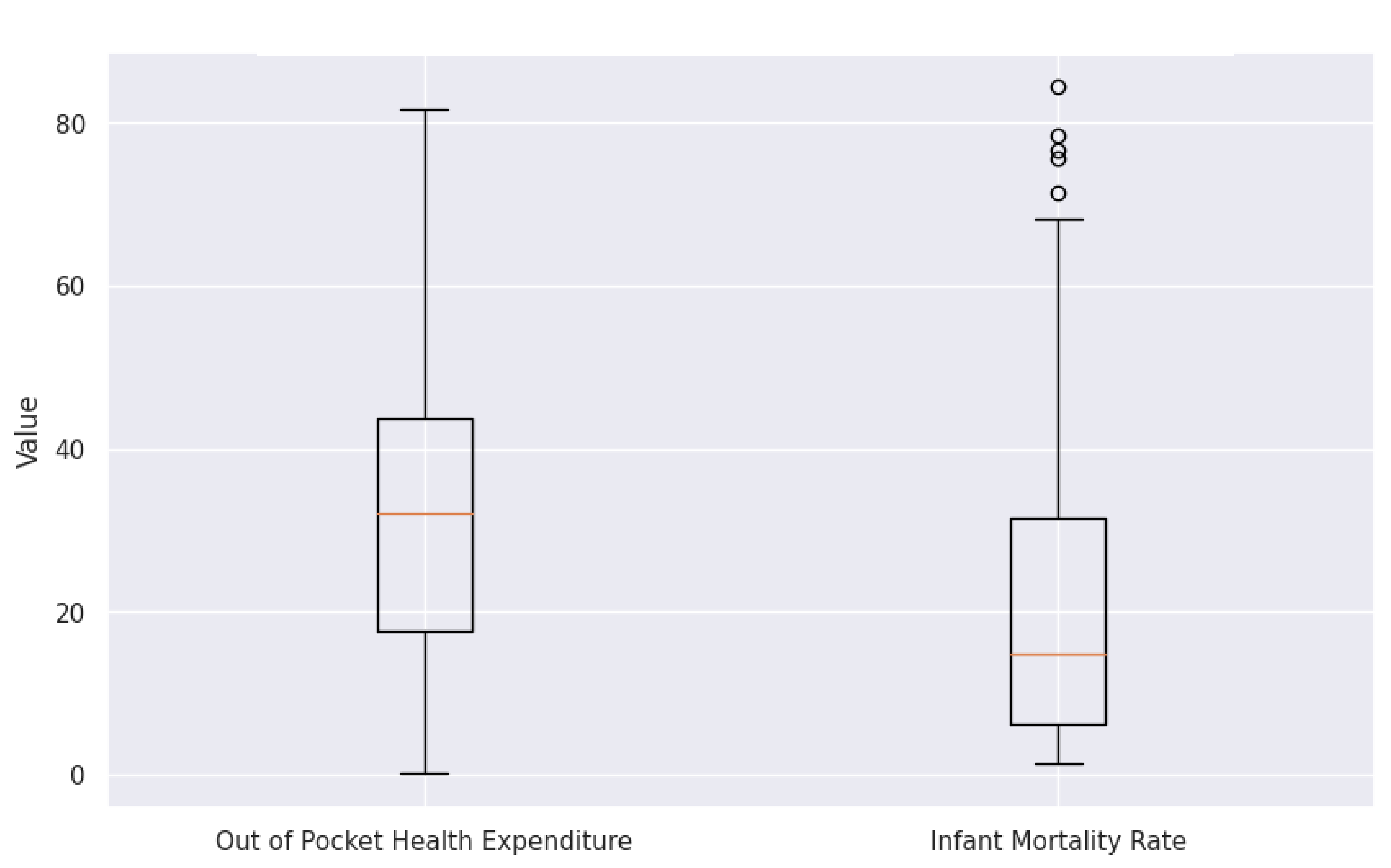
Top 20 Countries with Highest IMR



Bottom 20 Countries with lowest IMR



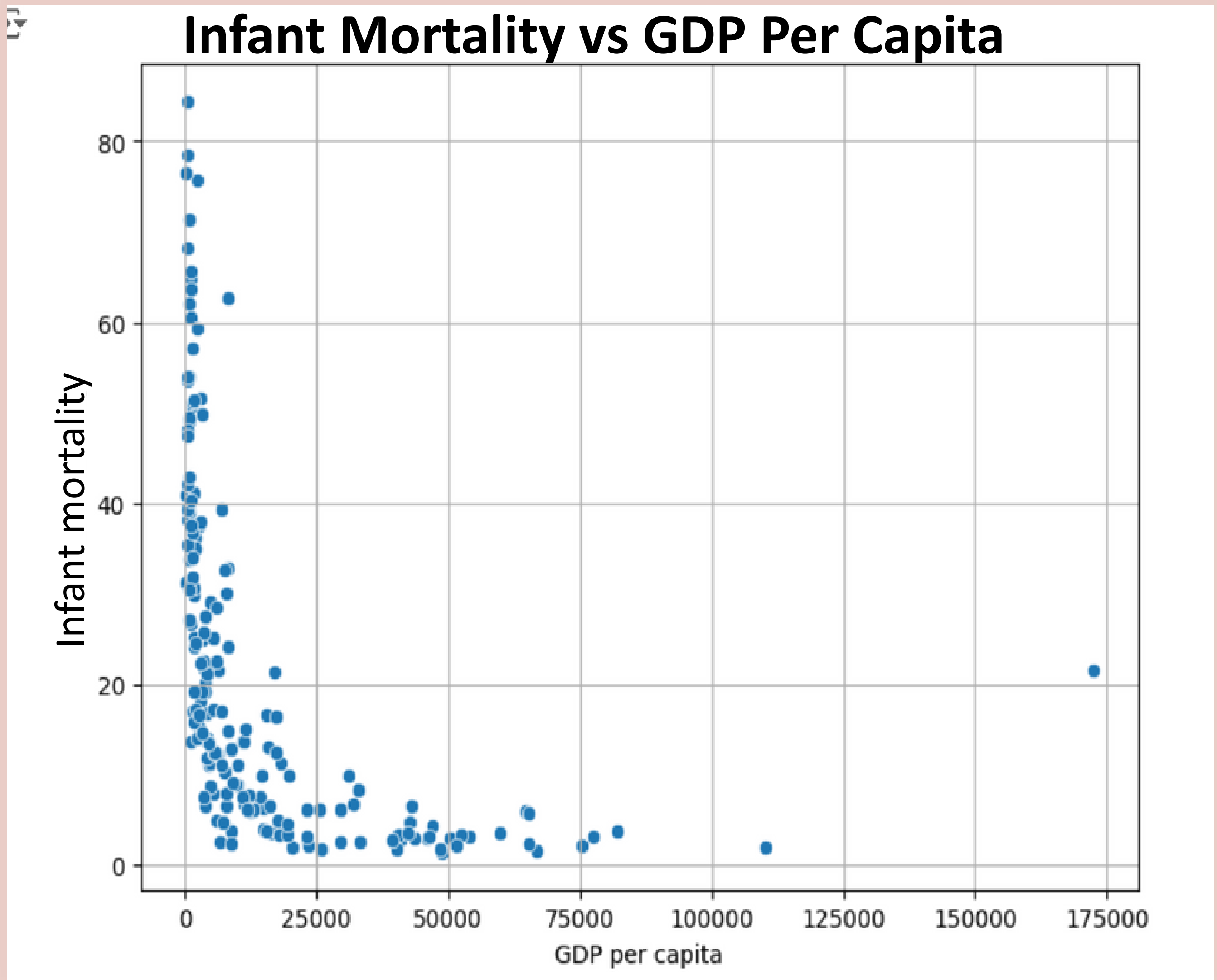
Box Plot of Health Expenditure and Infant Mortality Rate



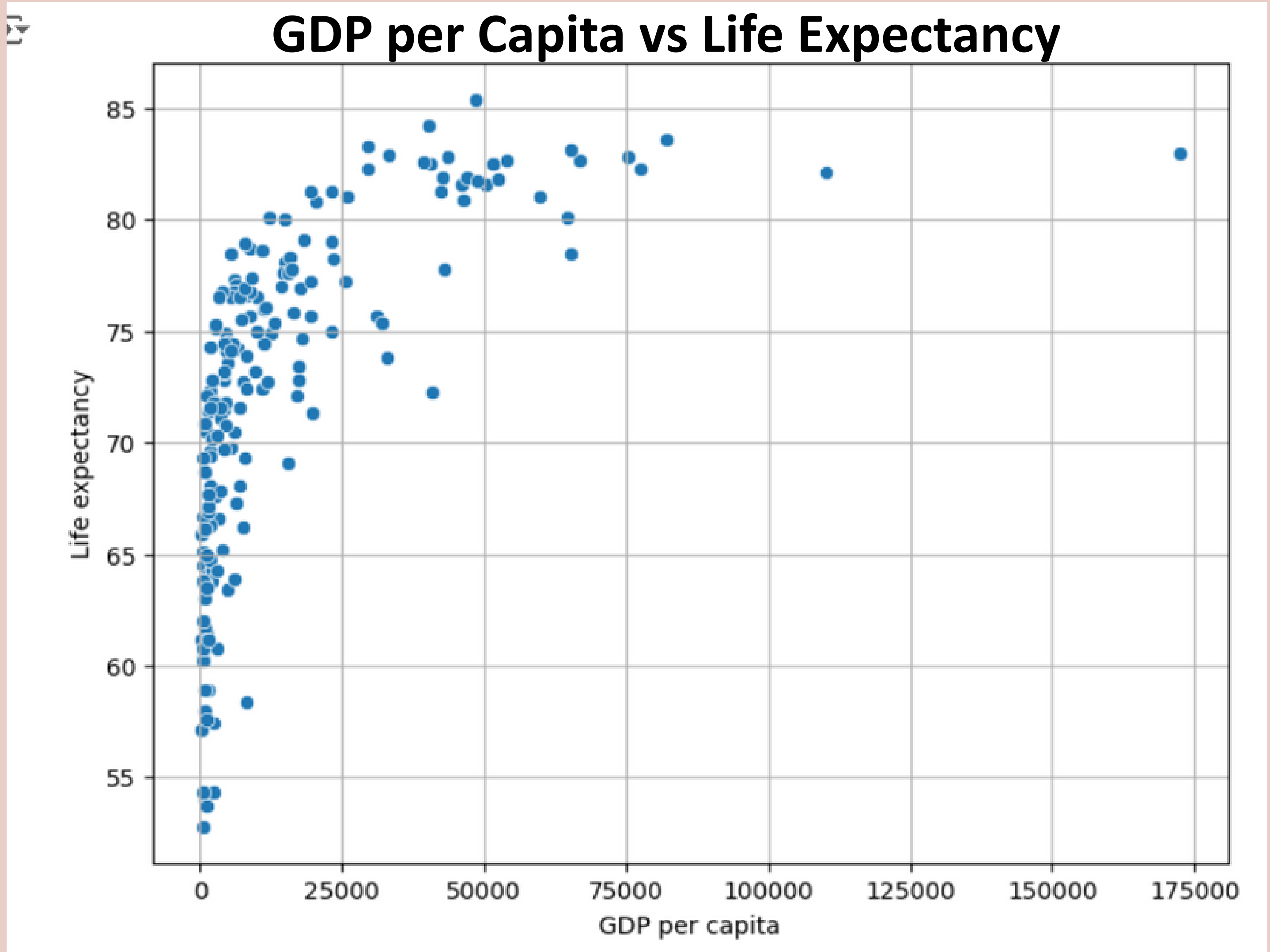
This means that in half of the countries, people spend 32% of their healthcare costs out of their own pockets.

This indicates that in half of the countries, the infant mortality rate is 14 or lower.

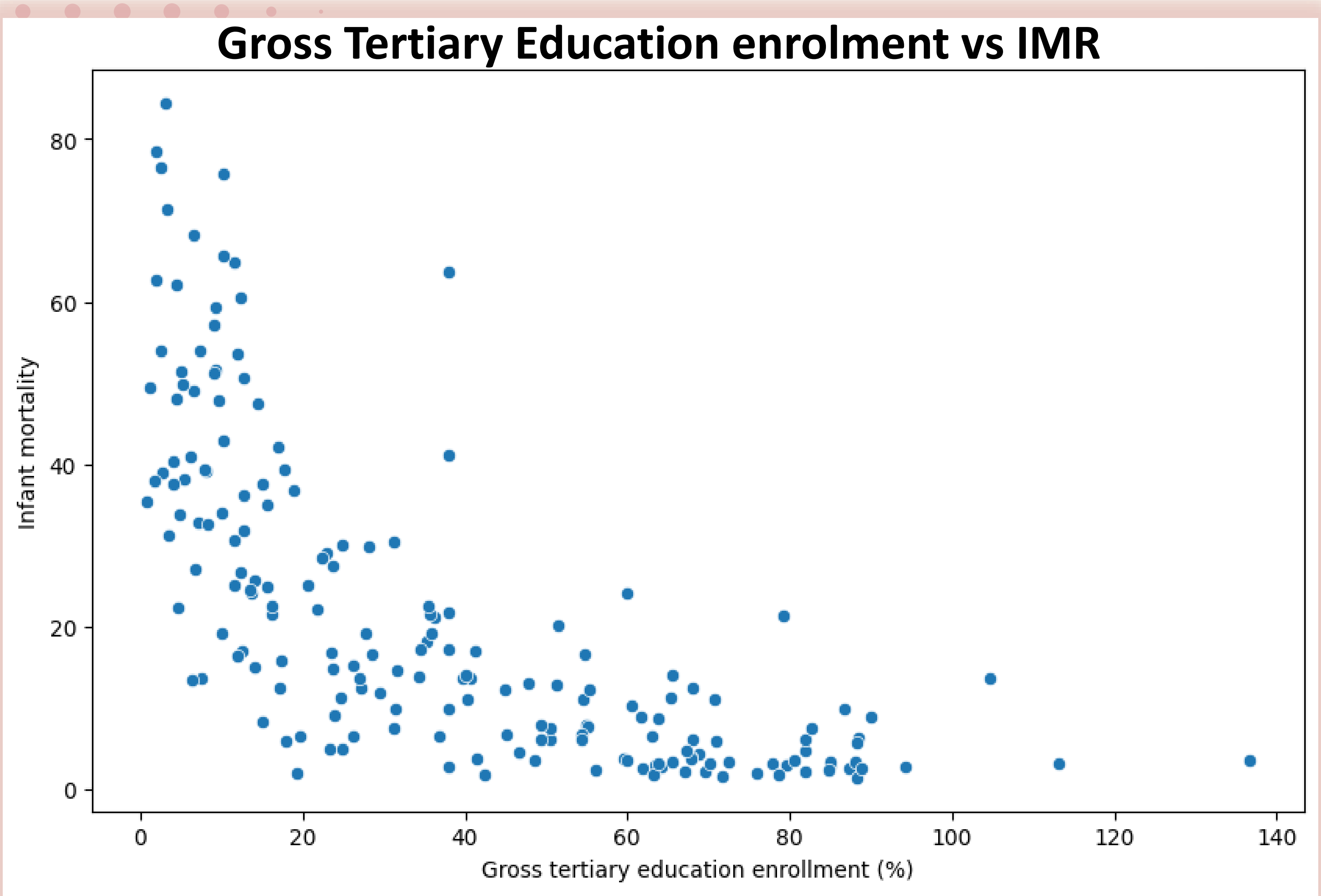
Scatter Plot

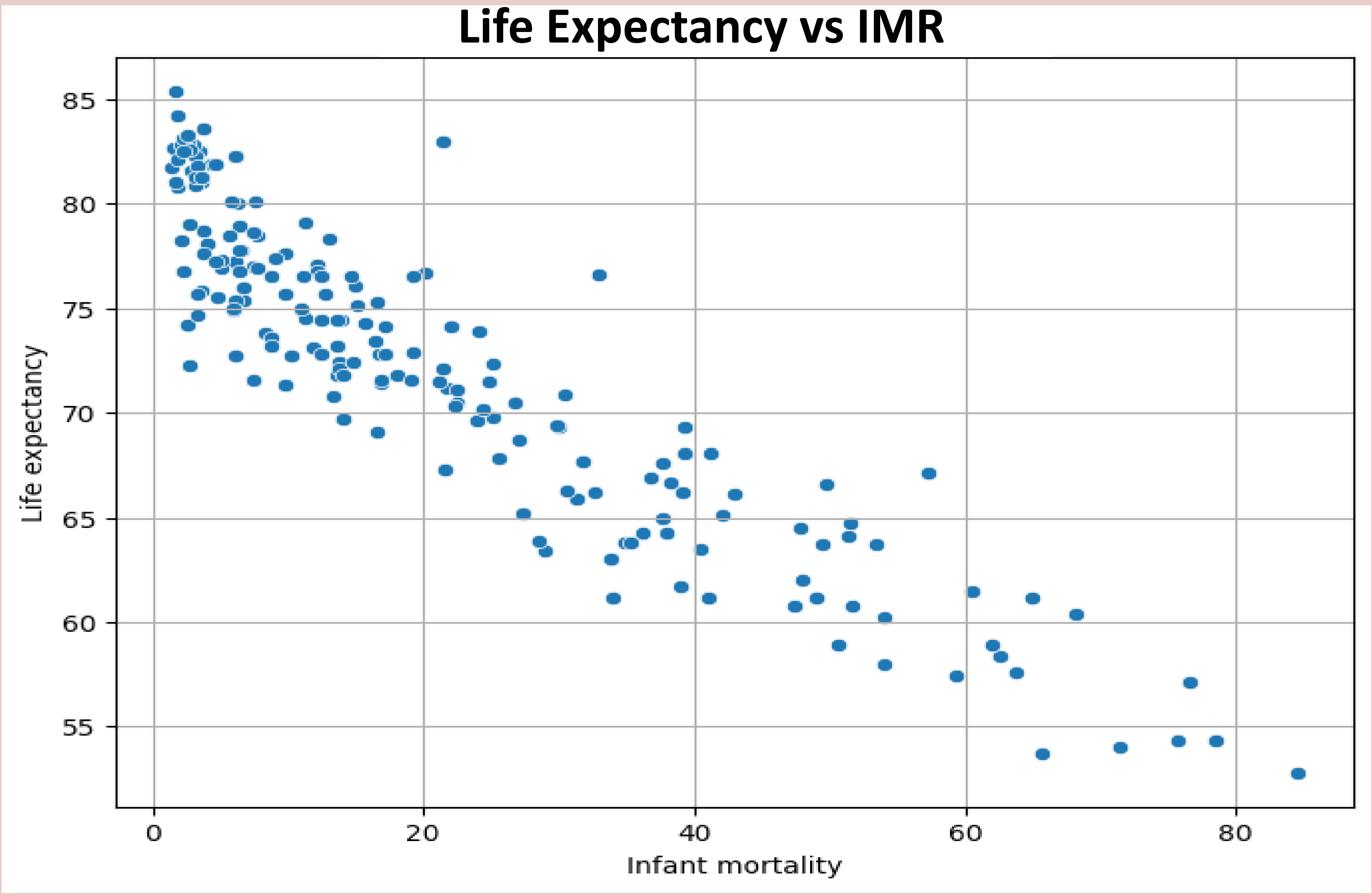


Most of the countries have less than $50,000 GDP per capita
having Infant mortality is upto 80 years.



Most of the countries are concentrated with
Life expectancy <80 and GDP per capita $25,000.





The countries with higher education enrolment have lower infant mortality
And higher infant mortality have lower education enrolment.

The countries with higher life expectancy have lower infant mortality
And higher infant mortality have lower life expectancy.

** Please refer to Appendix page 36 for more plots*

Multicollinearity check

Table 1

variables	VIF
Birth Rate	54.6
Urban_population	52.6
Fertility Rate	44.9
Population	28.7
Co2-Emissions	25.9
GDP	10.1
CPI	6
Armed Forces size	5.7
CPI Change (%)	5.6
Maternal mortality ratio	3.7
Physicians per thousand	3.2
Gross tertiary education enrollment (%)	3.1
GDP per capita	2.3
Land Area(Km2)	2.2
Minimum wage	2.1
Gasoline Price	1.8
Out of pocket health expenditure	1.7
Tax revenue (%)	1.6
Agricultural Land(%)	1.5
Life expectancy	1.5
Forested Area (%)	1.5
Population: Labor force participation (%)	1.5
Total tax rate	1.4
Unemployment rate	1.4
Gross primary education enrollment (%)	1.3
Density\n(P/Km2)	1.2

Table 2

variables	VIF
Urban_population	52.4
Population	28.6
Co2-Emissions	25.9
GDP	10.1
CPI	5.9
Armed Forces size	5.6
CPI Change (%)	5.5
Maternal mortality ratio	3.7
Fertility Rate	3.5
Gross tertiary education enrollment (%)	3
Physicians per thousand	2.8
GDP per capita	2.3
Minimum wage	2.1
Land Area(Km2)	2.1
Gasoline Price	1.7
Out of pocket health expenditure	1.6
Tax revenue (%)	1.6
Agricultural Land(%)	1.5
Life expectancy	1.5
Forested Area (%)	1.5
Population: Labor force participation (%)	1.5
Total tax rate	1.4
Unemployment rate	1.4
Gross primary education enrollment (%)	1.2
Density\n(P/Km2)	1.2

Table 3

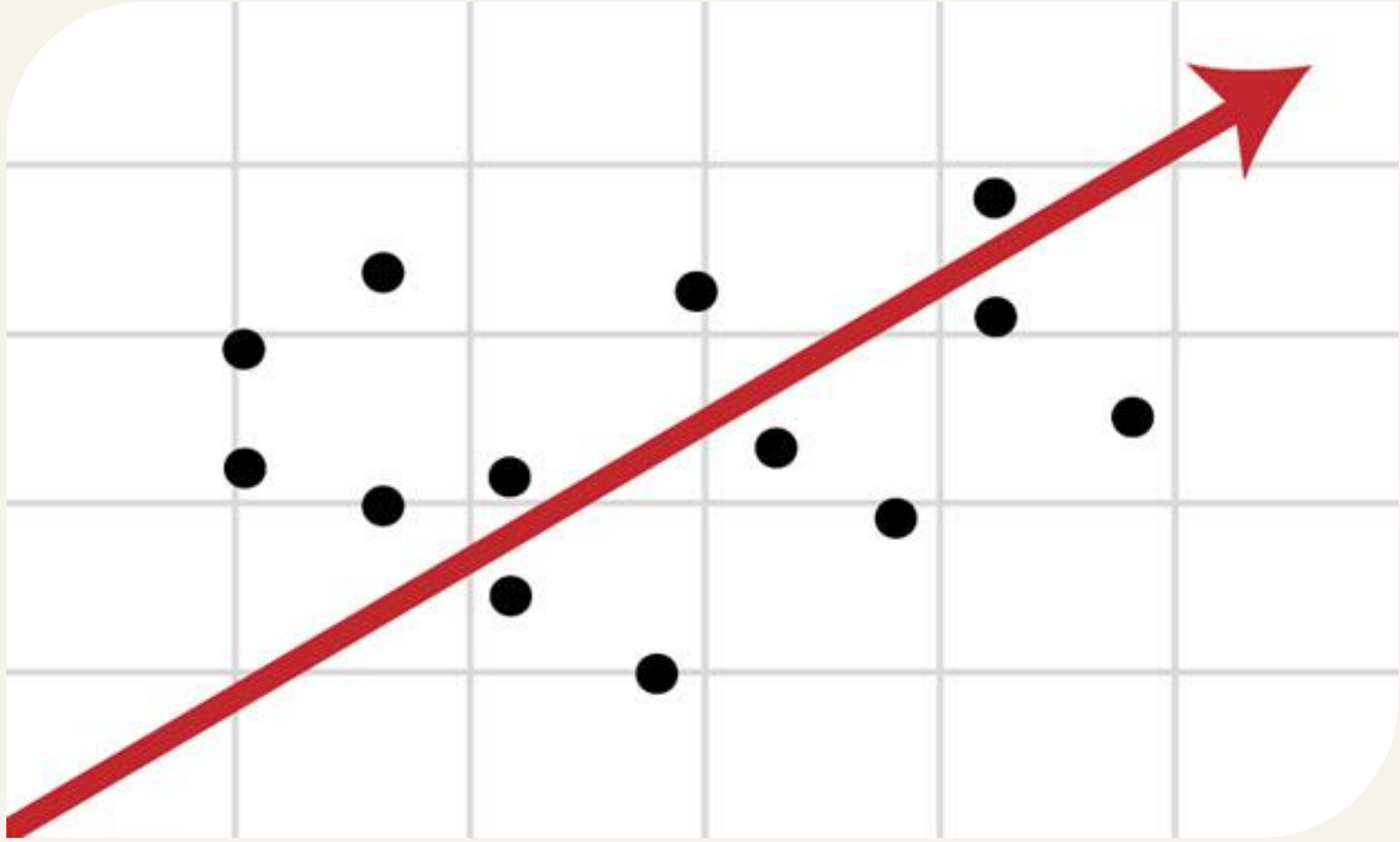
variables	VIF
Co2-Emissions	16.9
GDP	10.1
Population	8.3
CPI	5.9
Armed Forces size	5.6
CPI Change (%)	5.5
Maternal mortality ratio	3.7
Fertility Rate	3.5
Gross tertiary education enrollment (%)	3
Physicians per thousand	2.8
GDP per capita	2.3
Minimum wage	2.1
Land Area(Km2)	2
Gasoline Price	1.7
Out of pocket health expenditure	1.6
Tax revenue (%)	1.6
Agricultural Land(%)	1.5
Life expectancy	1.5
Forested Area (%)	1.5
Population: Labor force participation (%)	1.5
Unemployment rate	1.4
Total tax rate	1.3
Gross primary education enrollment (%)	1.2
Density\n(P/Km2)	1.2

Table 4

variables	VIF
CPI	5.9
CPI Change (%)	5.5
Armed Forces size	5.4
Population	5.2
Maternal mortality ratio	3.6
Fertility Rate	3.5
Gross tertiary education enrollment (%)	3
Physicians per thousand	2.8
GDP	2.4
GDP per capita	2.2
Minimum wage	2
Land Area(Km2)	1.9
Gasoline Price	1.7
Tax revenue (%)	1.6
Agricultural Land(%)	1.5
Life expectancy	1.5
Out of pocket health expenditure	1.5
Forested Area (%)	1.5
Population: Labor force participation (%)	1.5
Unemployment rate	1.4
Total tax rate	1.3
Gross primary education enrollment (%)	1.2
Density\n(P/Km2)	1.2

The features like **Birth Rate, Urban_ population, Co2-Emission** were multicollinear in our dataset.

DATA MODELLING



Significance Test for Multiple Linear Regression

Ordinary Least squares

OLS Regression Results

Dep. Variable:	Infant mortality	R-squared:	0.865
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	294.2
Date:	Sun, 18 Aug 2024	Prob (F-statistic):	1.59e-78
Time:	12:25:20	Log-Likelihood:	-636.66
No. Observations:	188	AIC:	1283.
Df Residuals:	183	BIC:	1300.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9679	2.490	0.389	0.698	-3.945	5.881
Maternal mortality ratio	0.0411	0.004	11.164	0.000	0.034	0.048
Fertility Rate	5.1354	0.740	6.942	0.000	3.676	6.595
Physicians per thousand	-2.1225	0.438	-4.846	0.000	-2.987	-1.258
Out of pocket health expenditure	0.1174	0.029	4.017	0.000	0.060	0.175

Omnibus:	34.512	Durbin-Watson:	2.209
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59.001
Skew:	0.941	Prob(JB):	1.54e-13
Kurtosis:	4.997	Cond. No.	1.37e+03

Multiple Linear Regression

★ Y(**Infant Mortality**)=0.9679

+0.0411(**Maternal mortality ratio**)

+5.1354(**Fertility Rate**)

-2.1225(**Physicians per thousand**)

+0.1174(**Out of pocket health expenditure**)

	Train-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

Here the Multiple Regression is the good fit with 88% R^2

Decision Tree

Train-test	max_features	min_samples_leaf	max_depth	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

Here the Decision Tree is consistent for different train-test ratio's as R^2 is about 70%

** Please refer to Appendix page 43 for a detailed Decision Tree diagram*

Random Forest

Train-test	max_features	min_samples_leaf	max_depth	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

Here the Random Forest Tree is consistent for different train-test ratio's as R^2 is in around 83% to 86%.

** Please refer to Appendix page 44 for a details code*

KNN

Train-test	n	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

Here the KNN is **consistent** for different train-test ratio's as R^2 is about 80%.

XGBoosting

Train-test	max_depth	n_estimators	sub_sample	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

Here the XGBoosting is **not consistent** for different train-test ratio's as R^2 is ranging between 50% - 70%.

Bagging

train-test	n_estimator	max_samples	max_features	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

Here the Bagging is consistent for different train-test ratio's as R^2 is ranging is around 86%.

Results

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

Here the Multiple linear is consistent for different train-test ratio's as R^2 is 88%.

Conclusion

- On comparing different machine learning algorithms the Multiple Linear Regression is the best model with best fit of 88%.

- **From Multiple Linear Regression:**

  Having more doctors and longer life expectancy for countries helps reduce infant deaths, while having more children per family and higher health costs tend to increase it.

Linear model

Y(**Infant Mortality**)=0.9679

+0.0411(**Maternal mortality ratio**)

+5.1354(**Fertility Rate**)

-2.1225(**Physicians per thousand**)

+0.1174(**Out of pocket health expenditure**)

Recommendation

- Number of doctors(Physicians) of the country can reduce the infant mortality.
- In order to reduce the infant mortality in the country, government’s must focus on improving the numbers of doctors and improving the Health care infrastructure.
- Promote Family Planning and Education, a need for policies that promote family planning, reproductive health education, and women’s empowerment, aiming to reduce high fertility rates

Thank you

Ruchitha | Shravanthika | Ramya | Madirai | Deepika

Contributions

<u>Ruchitha</u>

Worked on Exploratory Data Analysis:

<u>Shravanthika</u>

Worked on Data Cleaning, and imputing methods

<u>Ramya</u>

Worked on Multicolinearity

<u>Madirai</u>

Worked on data modeling

<u>Deepika</u>

Worked on data modeling

Appendix

Data Description:

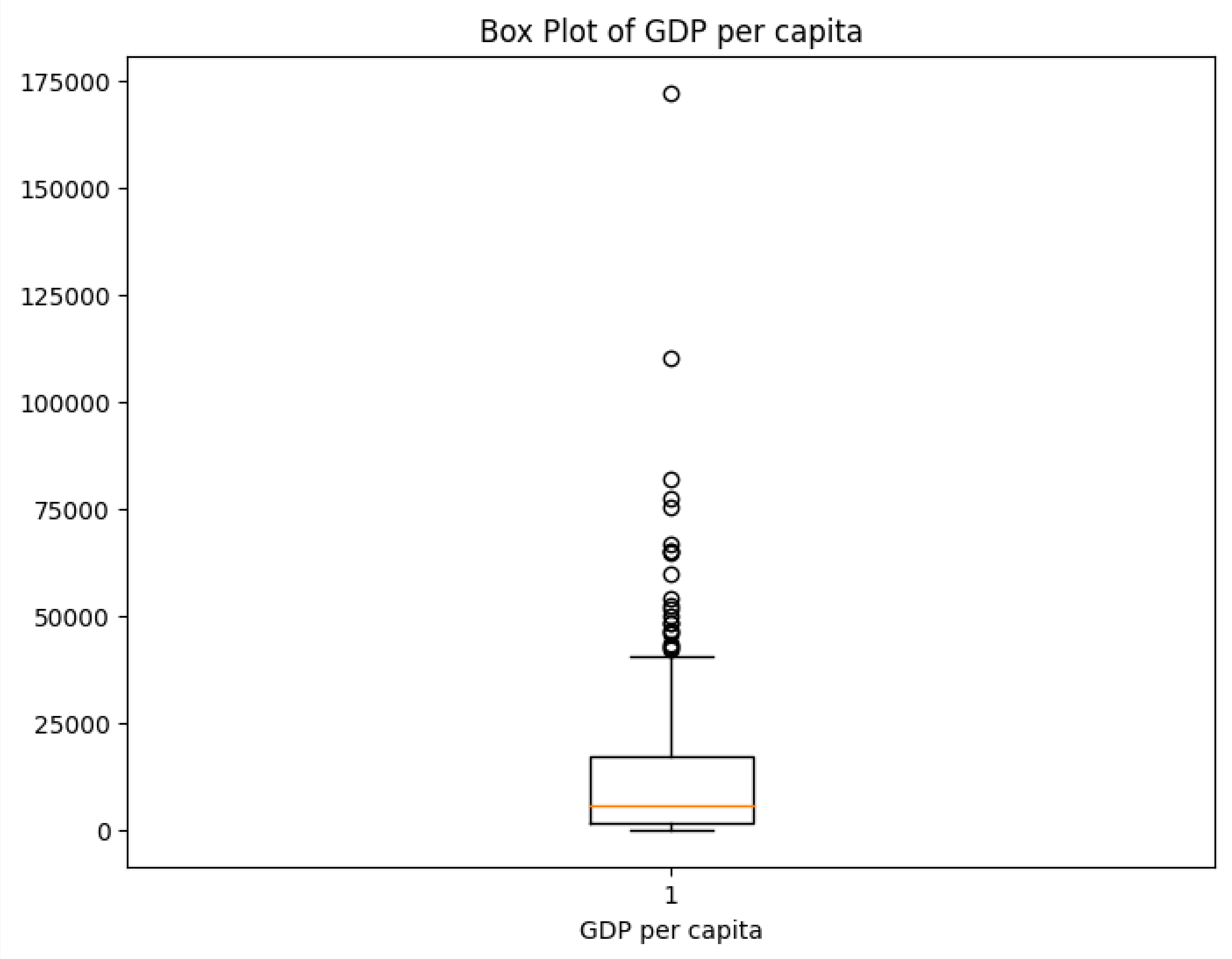
Variable Name	Description
Country	Name of the country.
Density (P/Km²)	Population density measured in persons per square kilometer.
Agricultural Land (%)	Percentage of total land area that is used for agriculture.
Land Area (Km²)	Total land area of the country in square kilometers.
Armed Forces Size	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.

```python
# Get pairs of variables with correlation greater than 0.5
corr_matrix = data1.iloc[:,1:].corr()
high_corr_pairs = []
for i in range(len(corr_matrix.columns)):
  for j in range(i + 1, len(corr_matrix.columns)):
    if abs(corr_matrix.iloc[i, j]) > 0.5:
      high_corr_pairs.append((corr_matrix.columns[i], corr_matrix.columns[j]))

print("Pairs of variables with correlation greater than 0.5:")
for pair in high_corr_pairs:
  print(pair)

```

```
Pairs of variables with correlation greater than 0.5:
('Land Area(Km2)', 'Armed Forces size')
('Land Area(Km2)', 'Co2-Emissions')
('Land Area(Km2)', 'GDP')
('Land Area(Km2)', 'Urban_population')
('Armed Forces size', 'Co2-Emissions')
('Armed Forces size', 'GDP')
('Armed Forces size', 'Population')
('Armed Forces size', 'Urban_population')
('Birth Rate', 'Fertility Rate')
('Birth Rate', 'Gross tertiary education enrollment (%)')
('Birth Rate', 'Infant mortality')
('Birth Rate', 'Life expectancy')
('Birth Rate', 'Maternal mortality ratio')
('Birth Rate', 'Physicians per thousand')
('Birth Rate', 'GDP per capita')
('Co2-Emissions', 'GDP')
('Co2-Emissions', 'Population')
('Co2-Emissions', 'Urban_population')
```

Pair of Features which are highly Correlated with each other

Box Plot of GDP per Capita

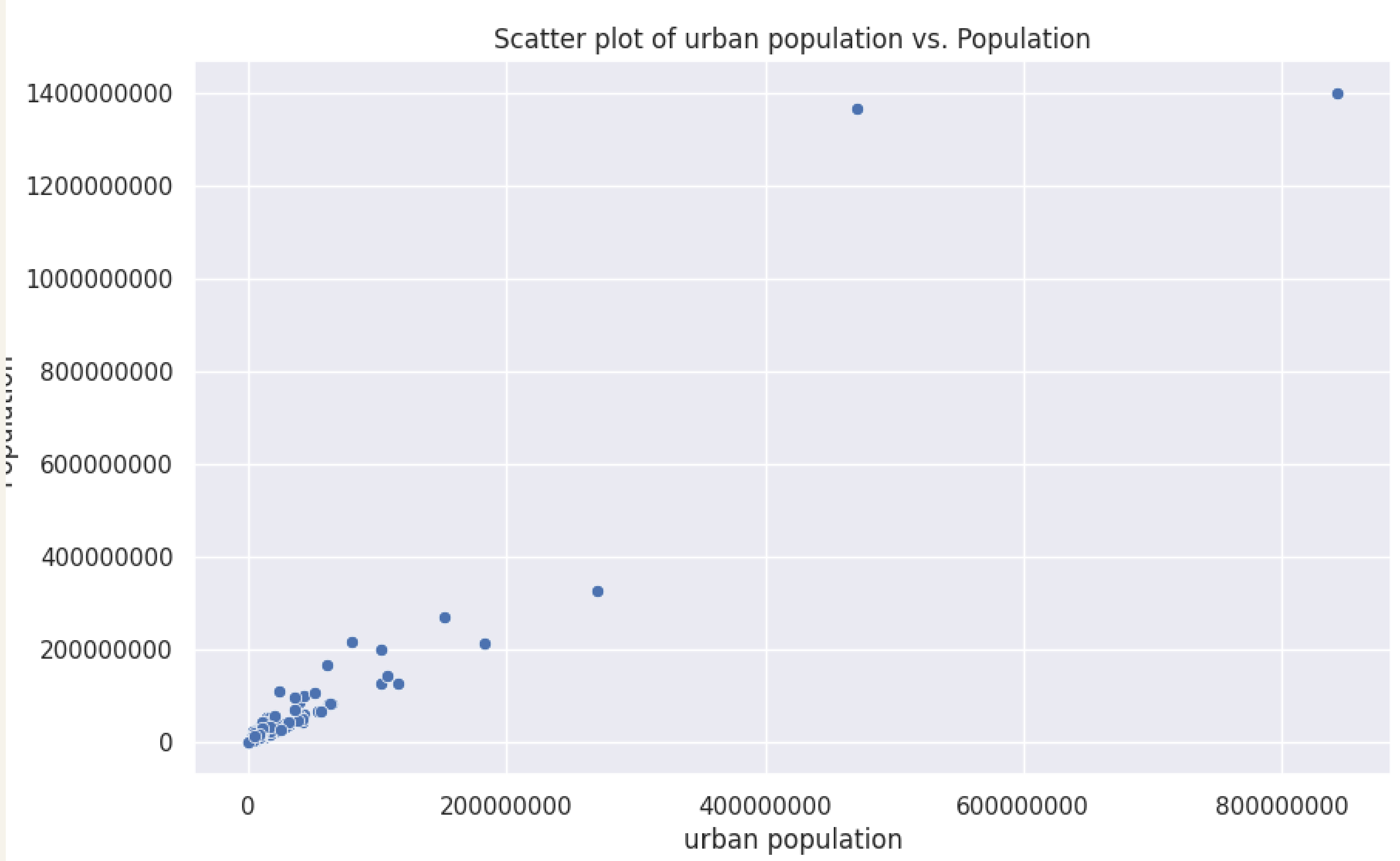



```
count 188.00
mean 14987.56
std 22377.72
min 261.25
25% 1887.26
50% 5978.25
75% 17498.87
max 172357.47
```

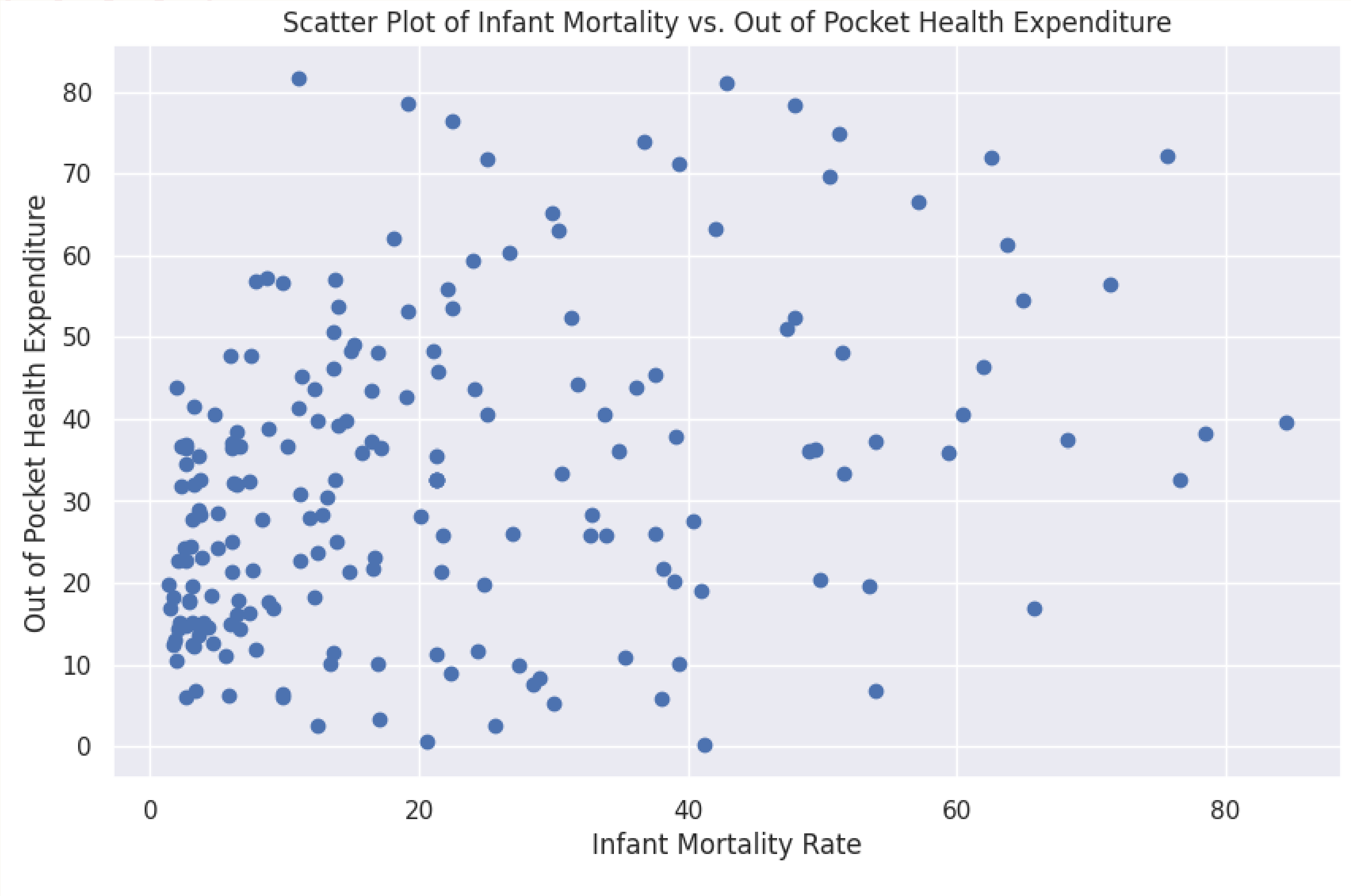
Statistics of GDP per capita

Scatter Plot between Urban Population and Population

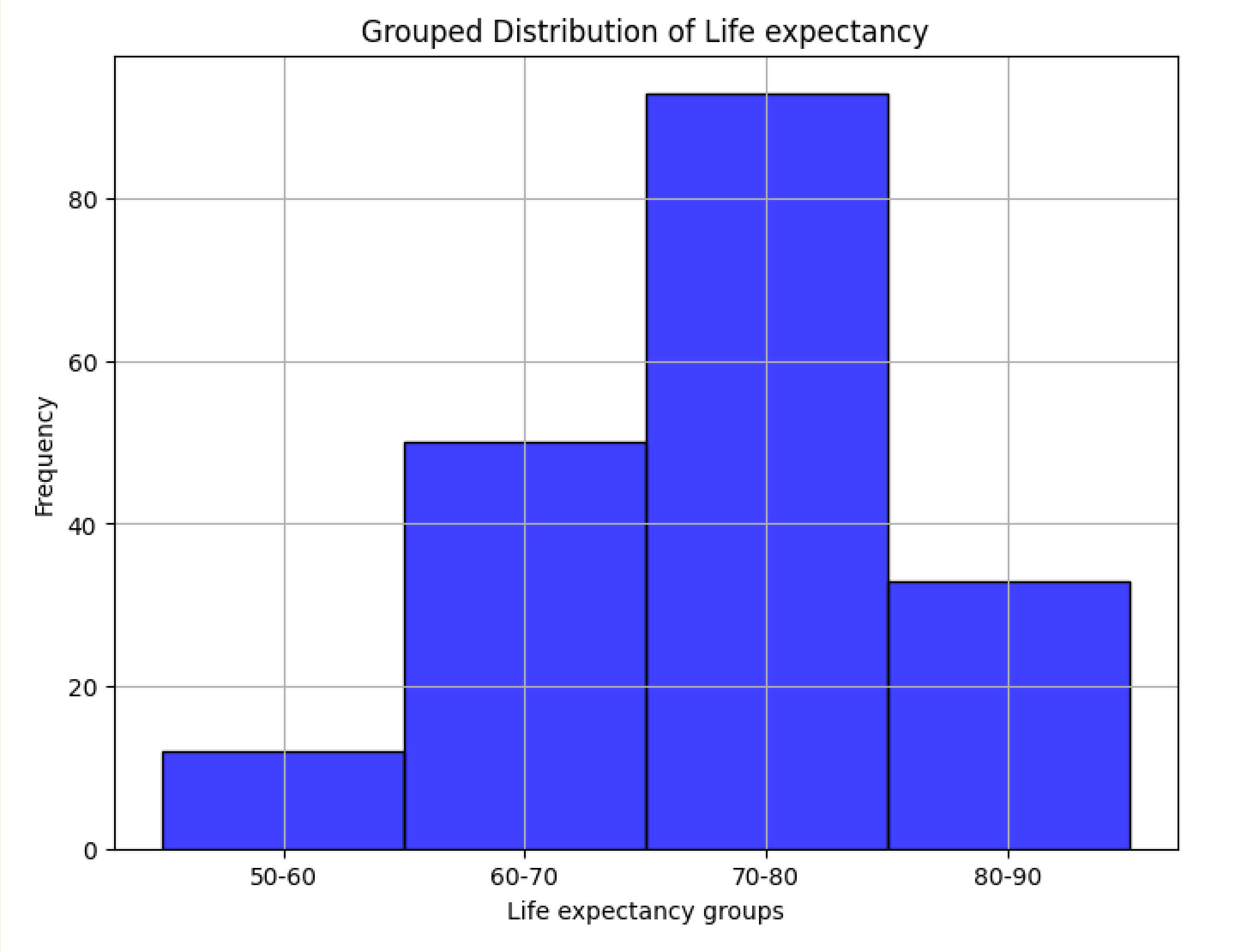
```python
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Urban_population', y='Population', data=data1)
plt.title('Scatter plot of urban population vs. Population')
plt.ticklabel_format(style='plain', axis='y')
plt.ticklabel_format(style='plain', axis='x')
plt.xlabel('urban population')
plt.ylabel('Population')
plt.show()
```



Scatter Plot of IMR and Pocket health Expenditure



Grouped Life Expectancy into Group



Code for Variance Inflation Factor (VIF)

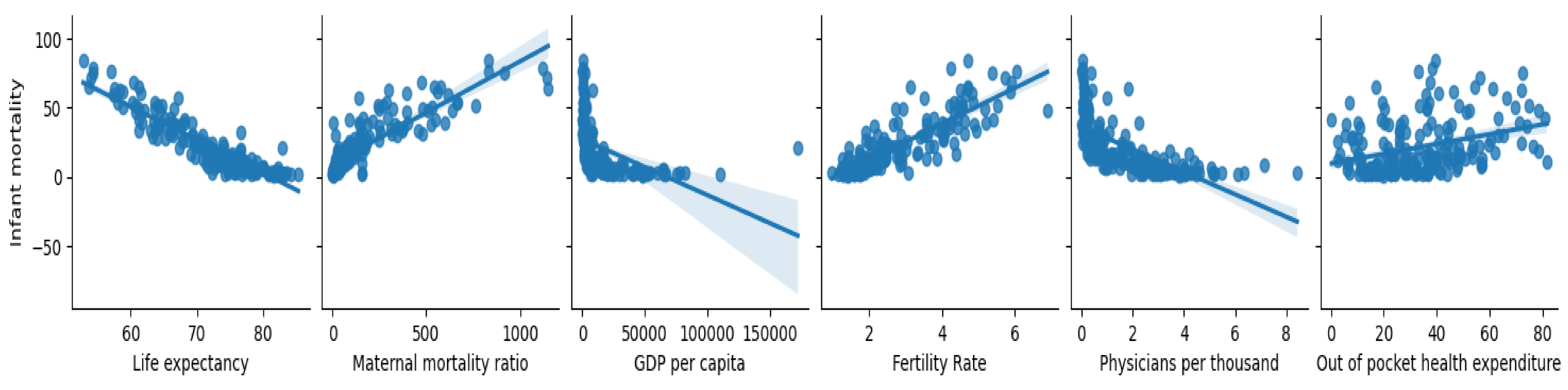
```
# Import library for VIF
from statsmodels.stats.outliers_influence import variance_inflation_factor

def calc_vif(X):

    # Calculating VIF
    vif = pd.DataFrame()
    vif["variables"] = X.columns
    vif["VIF"] = [variance_inflation_factor(X.values, i).round(1) for i in range(X.shape[1])]

    return(vif)

vif_df = calc_vif(X)
vif_df.sort_values(by='VIF',ascending=False)
```

	variables	VIF
4	Birth Rate	54.6
24	Urban_population	52.6
8	Fertility Rate	44.9
19	Population	28.7
5	Co2-Emissions	25.9
11	GDP	10.1
6	CPI	6.0
3	Armed Forces size	5.7
7	CPI Change (%)	5.6
15	Maternal mortality ratio	3.7
18	Physicians per thousand	3.2
13	Gross tertiary education enrollment (%)	3.1
25	GDP per capita	2.3
2	Land Area(Km2)	2.2
16	Minimum wage	2.1
10	Gasoline Price	1.8
17	Out of pocket health expenditure	1.7
21	Tax revenue (%)	1.6
1	Agricultural Land(%)	1.5
14	Life expectancy	1.5

Code for Ordinary Least Square

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

import statsmodels.api as sm

x_train_constant = sm.add_constant(x_nomulti_colinearity)

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

# Print the summary to get p-values
print(model.summary())
print(model.pvalues.sort_values(ascending=False).round(4))
```

Pair plot for Significant Features



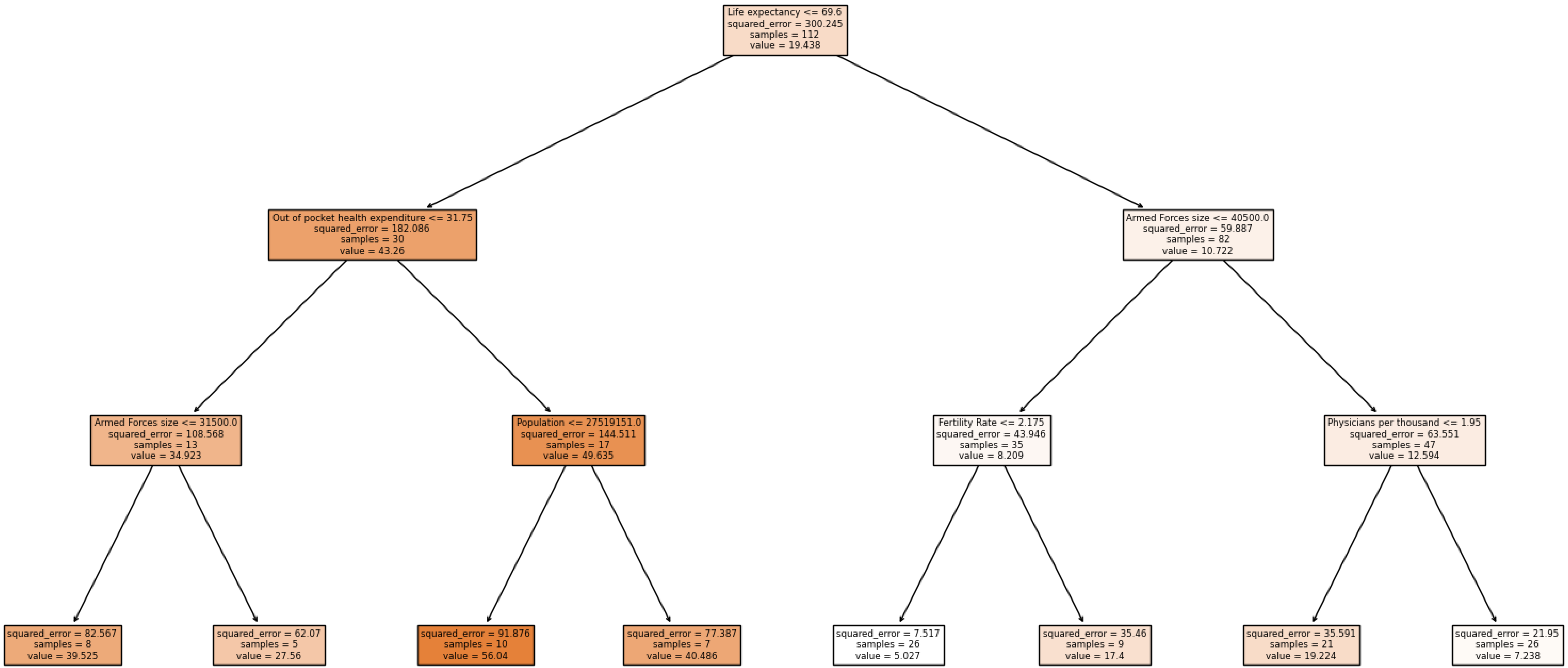
Multiple Linear Regression

```
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: MULTIPLEREGRESSION")
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)

  # 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)))
```

Decision Tree

```
from sklearn.tree import DecisionTreeRegressor, plot_tree
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import GridSearchCV

for i in [x_nomulti_colinearity]:
  X_train, X_test, y_train, y_test = train_test_split(i, y, test_size=0.2, random_state=0)
  regressor = DecisionTreeRegressor(max_features=4,min_samples_leaf=3,max_depth=3)
  regressor.fit(X_train, y_train)
  y_pred = regressor.predict(X_test)
  print("\nTrain Mean Squared Error: ", mean_squared_error(y_train, regressor.predict(X_train)).round(3))
  print("Test Mean Squared Error: ", mean_squared_error(y_test, y_pred).round(3))
  print("\nTrain R^2 Score: ", r2_score(y_train, regressor.predict(X_train)).round(3))
  print("Test R^2 Score: ", r2_score(y_test, y_pred).round(3))
```

Decision Tree Diagram



Random Forest


```python
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)

  # Creating the Random Forest Regressor model
  model = RandomForestRegressor(n_estimators=100,max_depth=3,min_samples_leaf=2,min_samples_split=2,max_features=4)

  # Training the model on the training data
  model.fit(X_train, y_train)

  # Predicting on the test data
  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)))
```

KNN

```
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import accuracy_score, r2_score
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
y_array = y.to_numpy()
y_reshaped = y_array.reshape(-1, 1)


y_scaled = pd.DataFrame(scaler.fit_transform(y_reshaped), columns=['Infant mortality'])

for i in [x_nomulti_colinearity]:
  # Fit the scaler to the data and transform it
  df_standardized = pd.DataFrame(scaler.fit_transform(i), columns=i.columns)
  X_train, X_test, y_train, y_test = train_test_split(df_standardized, y_scaled, test_size=0.2, random_state=0)
  for j in range(1,20):
    model=KNeighborsRegressor(n_neighbors=j)
    model.fit(X_train,y_train)
    y_pred=model.predict(X_test)
    print(f'-----N={j}----')
    print("Train Mean Squared Error: ", mean_squared_error(y_train, model.predict(X_train)).round(3))
    print("Test Mean Squared Error: ", mean_squared_error(y_test, y_pred).round(3))
    print()
    print("Train R^2 Score: ", r2_score(y_train, model.predict(X_train)).round(3))
    print("Test R^2 Score: ", r2_score(y_test, y_pred).round(3))

```

Support Vector Machine


```
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)))

```

XGBossting

```
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)

  # Creating the XGBoost Regressor model
  model = XGBRegressor(max_depth=4,
                       n_estimators=100,
                       subsample=0.0310,

                       random_state=0)

  # Training the model on the training data
  model.fit(X_train, y_train)

  # Predicting on the test data


  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))
```

Bagging


```
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)

  # Creating the Bagging Regressor model with Decision Tree as base estimator
  model = BaggingRegressor(
      base_estimator=DecisionTreeRegressor(),
      n_estimators=99,  # Number of trees in the ensemble
      random_state=42,
      max_samples=0.8, # Fraction of samples to use for each tree
      max_features=4
  )

  # Training the model on the training data
  model.fit(X_train, y_train)

  # Predicting on the test data
  y_pred = model.predict(X_test)

  # Calculating the R² score
  r2 = r2_score(y_test, y_pred)

  # Output the R² score
  print(f'\nR² for test ration {i*100}% score: {r2:.4f}')
  print("Mean squared error:",(metrics.mean_squared_error(y_test, y_pred)))
```

Thank you

Ruchitha | Shravanthika | Ramya | Madirai | Deepika

80-20 split

Models/splits	R^2
Multiple Regression	0.927
Random Forest	0.853
Dession Tree	0.731
KNN	0.803
SVR	0.735
Bagging	0.870
XGBoosting	0.712

Here for 80-20 train-test split Multiple Regression is the best model as compared to the other models.

75-25 split

Models/splits	R^2
Multiple Regression	0.933
Random Forest	0.873
Dession Tree	0.744
KNN	0.820
SVR	0.749
Bagging	0.862
XGBoosting	0.647

Here for 75-25 train-test split Multiple Regression is the best model as compared to the other models.

70-30 split

Models/splits	R^2
Multiple Regression	0.926
Random Forest	0.864
Dession Tree	0.693
KNN	0.807
SVR	0.758
Bagging	0.866
XGBoosting	0.539

Here for 70-30 train-test split Multiple Regression is the best model as compared to the other models.

60-40 split

Models/splits	R^2
Multiple Regression	0.913
Random Forest	0.878
Dession Tree	0.729
KNN	0.802
SVR	0.736
Bagging	0.869
XGBoosting	0.558

Here for 60-40 train-test split Multiple Regression is the best model as compared to the other models.