

Project Report

On

**PREDICTION AND ANALYSIS OF LIFE
EXPECTANCY**

Submitted

in partial fulfilment of the requirements for the degree of
MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

GLENNA MARIA

(Register No. SM21AS008)

(2021-2023)

Under the Supervision of

ROSEWIN MARIYA ROY



DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

MAY 2023

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

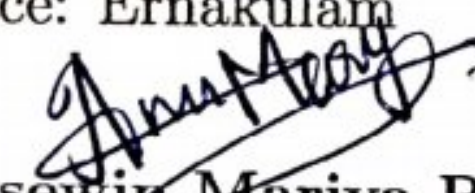


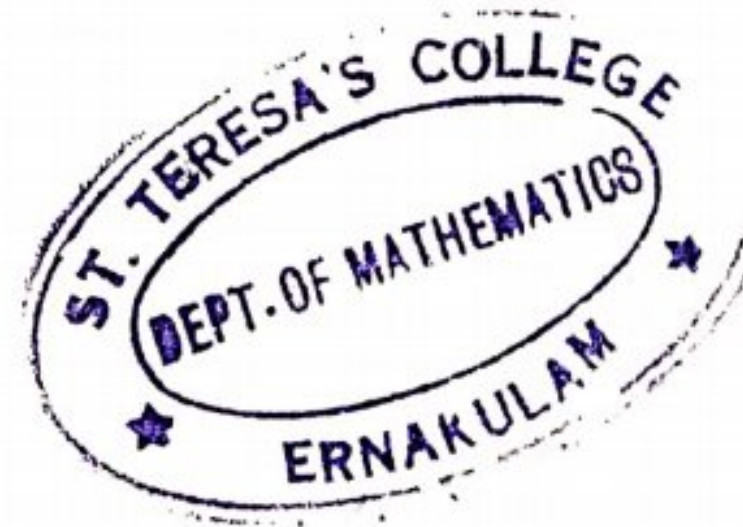
CERTIFICATE


This is to certify that the dissertation entitled, **PREDICTION AND ANALYSIS OF LIFE EXPECTANCY** is a bonafide record of the work done by Ms. **GLENNA MARIA** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 18/05/2023

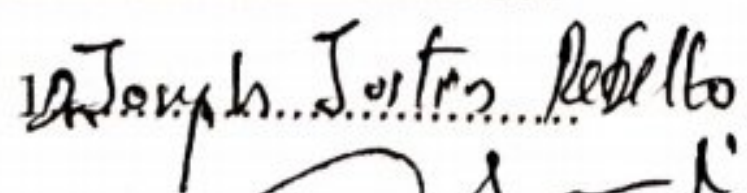
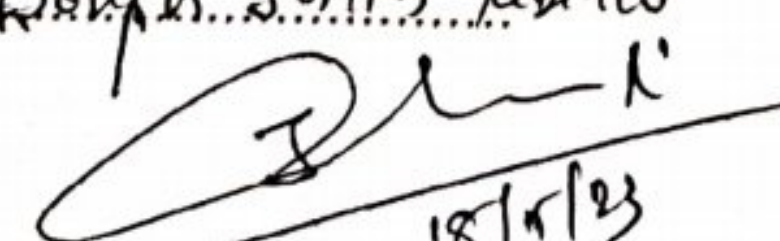
Place: Ernakulam


For 
Rosewin Mariya Roy
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.




Ms Betty Joseph
Associate Professor & HOD,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1: 

18/5/23

2: SARI THOMAS 

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **ROSEWIN MARIYA ROY**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam.

Date: 18/05/2023



GLENNA MARIA

SM21AS008



ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry this work. Their continuous invaluable knowledgeable guidance throughout the course of this study helped me to complete the work up to this stage

I am very grateful to my project guide Rosewin Mariya Roy for the immense help during the period of work

In addition, very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to faculty, teaching and non-teaching staff of the department and Colleagues.

I also very thankful to HoD Ms. Betty for their valuable suggestions, critical examination of work during the progress.

Ernakulam.

Date: 18/05/2023

GLENNA MARIA

SM21AS008



ABSTRACT

Life Expectancy have major effects on the social, health and financial sectors of most of the countries in the World. This paper investigates the Life Expectancy data to identify the factors that contribute highly to Life Expectancy, to compare the mean Life Expectancy of Developing and Developed countries. To fit and compare the Regression, Machine Learning and Time Series models for the data and prediction using the most accurate one. To achieve our goal, we used the data collected from Kaggle in the period of 2000 to 2015.

Factor Analysis was used for the identification of the factors that contribute highly to Life Expectancy. By means of Testing of Hypothesis the mean Life Expectancy of both Developing and Developed countries compared. Models such as Logistic Regression is used to evaluate Life Expectancy data of the World and Multiple Linear Regression, Random Forest Regression and ARIMA models are used to evaluate the Life Expectancy of India.

This study shows that there are mainly 5 factors contribute highly to the Life Expectancy. There is a difference in the mean Life Expectancy between Developing and Developed countries. Logistic Regression models fits good for the whole dataset for evaluating the Status using the other variables of each country. Among the applied models for the data the Random Forest model stands at the top position with 95% accuracy followed by other models.

By this study it is recommended that the research has to be expanded with more data and more Machine Learning Models.



ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM
Certificate of Plagiarism Check for Thesis



Author Name	Glenna Maria
Course of Study	MSc. Applied Statistics & Data Analytics
Name of Guide	Ms. Rosewin Mariya Roy
Department	Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	PREDICTION AND ANALYSIS OF LIFE EXPECTANCY
Similarity	0%
Paper ID	724194
Submission Date	2023-04-18 17:30:05

Signature of Student

For

Signature of Guide

Checked By
College Librarian

* This report has been generated by DrillBit Anti-Plagiarism Software



Contents

<i>CERTIFICATE</i>	ii
<i>DECLARATION</i>	iii
<i>ACKNOWLEDGEMENTS</i>	iv
<i>ABSTRACT</i>	v
<i>CONTENT</i>	vi
1 Introduction	2
1.1 Life Expectancy	2
1.2 Need of Life Expectancy Studies	3
1.3 Specific Objectives of Study	3
2 Literature Review	4
3 Methodology	7
3.1 Data	7
3.2 Exploratory Data Analysis	7
3.3 Factor Analysis	7
3.4 Testing of Hypothesis	8
3.4.1 Normality Tests	8
3.4.2 Z Score	8
3.5 Model building	9
3.5.1 Logistics Regression Model	9
3.5.2 Multiple Linear Regression Model	9
3.5.3 Random Forest Model	10
3.6 Time Series Model	10
3.6.1 Dickey-Fuller test for testing stationarity	10
3.6.2 ARIMA Model	10

4	Dataset and Exploratory Data Analysis	12
4.1	Definitions of attributes	12
4.2	Data Sample	14
4.3	Data Pre-Processing	14
4.4	Descriptive Statistics	14
4.5	Exploratory Data Analysis	15
4.5.1	Univariate Visualisations	15
4.5.2	Multivariate Visualisations	18
5	RESULTS AND DISCUSSION	20
5.1	Factor Analysis	20
5.2	Testing Of Hypothesis	22
5.3	Logistic Regression Model	23
5.4	Fitting of Models and Prediction With Best Model . . .	24
5.4.1	Machine Learning Model (Random Forest Model)	24
5.4.2	Regression Model (Multiple Linear Regression) .	25
5.4.3	Time Series Model (ARIMA Model)	25
5.4.4	Prediction with Best Model	25
6	Conclusion	27
7	References	28

Chapter 1

Introduction

1.1 Life Expectancy

A person's life expectancy, which varies depending on a number of different factors, is measured statistically. Observations of Life Expectancy have been employed in a variety of sectors over the years, including medical and healthcare planning, pricing and underwriting life insurance and insurance products, pension-related services, by concerned government and non-government bodies. LEB has been increasing in most of the countries globally and this is cited as a best indicator of the development of that country. Life Expectancy is a more natural measure of death rather than the mortality rates and it is a useful and most important aggregate measure of death. The increasing population indicates the increasing Life Expectancy with the steady health sectors. In some countries, now also Life Expectancy is very less than the rest of the countries. This inequality in the Life Expectancy values because of the unsettled healthcare facilities in these countries. In this paper to analyse and predict the Life Expectancy I have used Factor Analysis, z-test, Machine Learning model, Regression model and Time series model.

1.2 Need of Life Expectancy Studies

Development of ever-more-reliable predictive models is now possible because to developments in data science technologies and data analysis methodologies. How to determine the retirement age and how to handle the financial concerns associated to public matters are political hot topics in many nations. Expected lifespan predictions solves such problems and also by analyzing Life Expectancy we can find the improving quality of Health sectors. Changes in Life Expectancy can be also used to describes the trends and variations in mortality rates.

1.3 Specific Objectives of Study

- I. To identify the factors that contribute highly to Life Expectancy.
- II. To compare mean Life Expectancy of Developing and Developed countries.
- III. To fit Logistic Regression model for the data and to do prediction using it.
- IV. To fit Regression, Machine Learning and Time Series models for the data of India.
- V. To compare their accuracy and forecast the Life Expectancy using the most accurate model.

Chapter 2

Literature Review

Recently many studies are done on life expectancy. There are several papers, studies, and research articles about life expectancy that have already been completed by a variety of authors. Here are some work reviews are provided below which is related to our work.

- A Alshehri et al. on their paper “Prediction of life expectancy in Saudi Arabia by 2030 using ARIMA models”, they tried to analyse the life expectancy time series from 1960-2012 using Box-Jenkins methodology, the ARIMA (3,2,0) model was constructed, and data from 2013 to 2018 were used for validation. Finally, they have forecasted LEB in Saudi Arabia from 2019 to 2030 and it is expected to reach 78 years by 2030. They have used R programming language for their analysis.
- Maksuda Aktar Rubi et al. on their paper “Life Expectancy Prediction based on GDP and Population size of Bangladesh using Multiple Linear Regression and ANN model”, they tried to predict LEB by two methods and found that ANN models predict better future values. They used data from 1960-2020 and analyse using Python programming language.
- Suresh Kumar Karna et al. on their paper “A research study on the variables affecting Life Expectancy Descriptive and Inferential statistics with Excel and R”, they have studied the variables affecting Life expectancy and found which variable is more significant.

They used Descriptive statistics such as SD, variance Skewness by Excel and inferential statistics such as Multiple Linear Regression by R language for their study. They collected data from World Bank and WHO. They found that variables related to Mortality are much more significant than other variables.

- Akansha Maity et al. in their study “Factors Explaining Average Life Expectancy: An examination across Nations”, they tried to find the significant variables which affect the Life expectancy more. They found that average life expectancy can be influenced by gender, genetics, lifestyle. They put forward a recommendation would to use a dependent variable such as infant mortality to study LE.
- Dr. Vikram Bali et al. in their study “Life Expectancy: Prediction and Analysis using ML” their aim was to analyse the effect of features on the outcome and to check how it varies. Among different models that they have fitted, random forest performs best. This study clearly shows and proves the importance of health, education, and economic features on Life expectancy.
- Yallamati Prakasarao et al. in their study “Life Expectancy Prediction using Machine Learning” found a correlation between attributes like diseases, gender, age, environmental factor and they found that the Random Forest algorithm can forecast the human lifespan with more precision.
- Banhi Guha et al. in their study “Gold Price Forecasting Using ARIMA Model” they predicted the future Gold price in India based on past data from November 2003 to January 2014. ARIMA (1, 1, 1) model which helps them in predicting the future prices of Gold. ARIMA (1, 1, 1) was chosen from six different models as it provides the best model which satisfies all the criteria.
- Surefunmi idowu in her paper “Analysis of Population using Multiple and Logistic Regression Model” used multiple regression analysis to illustrate how well the annual population rate of a country can

be predicted using its total fertility rate, maternal mortality ratio, and life expectancy rank. Furthermore, she used the logistic regression analysis to show how well the life expectancy category/rank of a country can be predicted by its annual population rate and maternal mortality ratio. Her models are precise and more solid for the dataset.

- Shashank Gupta in his study “Life Expectancy using Linear Regression” he tried to analyze the life expectancy data and tried to predict life expectancy using Linear Regression. He started by understanding Life Expectancy and analyzed the factors affecting it. Then he finally implement linear regression for predicting life expectancy. He collected data from World Bank.
- K.C. Arum et al. in their paper “STATISTICAL STUDY OF LIFE EXPECTANCY OF MALE AND FEMALE CHILDREN AT BIRTH IN SOME SELECTED AFRICAN COUNTRIES” tried to determine if there is any significant difference in the distribution of life expectancy of male and female children at birth in some selected African countries and also to determine if there is a common difference in the distributions of life expectancy of male and female in the selected countries. They used 10 African countries for their study and used data from the period 1971 to 2017.

Chapter 3

Methodology

3.1 Data

Data is collected from Kaggle from the period of 2000 – 2015 and the Collected Data is on 21 Variables from 183 Countries out of 195 Countries with 2938 observations.

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA), a technique used in data mining, analyses datasets to highlight their key features, frequently employing data visualisation techniques. EDA is used for knowing what the data can tell before modeling procedures. It is not easy to look at the table and find the characteristics, in this kind of situation EDA can be used as a visual aid. EDA can be used to find the important features and correlations between the variables.

3.3 Factor Analysis

Finding significant underlying factors or latent variables from a set of observed variables is accomplished using the exploratory data analysis technique known as Factor Analysis (FA). By reducing the number of variables, it aids in the understanding of data. It extracts the most common variance possible from each variable and groups them under a

single score. Factors are unobservable(latent) variables that are combined linearly to form observed variables. The goal of FA is to reduce the redundancy among the variables by using a smaller no. of factors. A factor is a latent variable that describes the relationship between several variables that have been observed. A baseline for counting the number of factors is the eigenvalue. A common selection criterion for the feature will be an eigenvalue greater than 1.

3.4 Testing of Hypothesis

3.4.1 Normality Tests

- In statistics, D'Agostino's K2 test, named for Ralph D'Agostino, is a statistical test used to assess whether a given sample of data comes from a normal distribution. It is based on the skewness and kurtosis of the data, which are measures of the asymmetry and peakedness of the distribution, respectively.

The null hypothesis of the test is that the sample data comes from a normal distribution, while the alternative hypothesis is that it comes from a non-normal distribution. This can be used only when the sample size is greater than 20. If the value is less than 0.5, then reject H_0 , otherwise do not reject H_0 .

3.4.2 Z Score

A z-score is a standardized value that measures the distance between an individual data point and the mean of a dataset, expressed in units of standard deviation. It is a useful tool for comparing individual observations across different datasets, and for identifying extreme values or outliers within a dataset.

You must first determine the dataset's mean and standard deviation before you can determine a data point's z-score. Then, you subtract the mean from the data point and divide the result by the standard deviation. The resulting value is the z-score for that data point. A data point's z-score of 0 denotes that it is equal to the dataset's mean.

A point is above the mean if the z-score is positive, whereas the opposite is true if the z-score is negative. The magnitude of the z-score indicates how far the data point is from the mean in units of standard deviation.

Z-scores are often used in statistical analysis, such as hypothesis testing and regression analysis, to identify outliers and to standardize data across different datasets. They are also used in quality control to identify deviations from established norms or standards

3.5 Model building

3.5.1 Logistics Regression Model

Logistic Regression is both a classification and regression technique depending on the scenario used. Logistic regression is a type of analysis method used for predicting the outcome of a categorical dependent variable. The dependent variable should be binary (0,1) and independent variables are continuous in nature.

$$f(z) = \frac{1}{1+e^{-z}}$$

$$\text{where } z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Parameter β_0 is the intercept, Parameters $\beta_1, \beta_2, \dots, \beta_k$ are referred as partial regression coefficients, X_1, X_2, \dots, X_k are independent variables and z is the dependent variable.

3.5.2 Multiple Linear Regression Model

A statistical technique called Multiple Linear Regression (MLR) is used to simulate the relationship between a dependent variable and a number of independent variables. It is used to forecast the value of a dependent variable according to the values of multiple independent variables, in other words. In multiple linear regression, it is important to check for several assumptions, including linearity, normality, constant variance (homoscedasticity), and independence of errors. If these assumptions are not met, it may be necessary to use alternative methods or models, such as nonlinear regression or generalized linear models. It is a powerful tool for identifying significant predictors of a dependent variable and for understanding the relationships between multiple variables.

The equation for MLR is,

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Parameter α is the intercept of the plane, Parameters $\beta_1, \beta_2, \dots, \beta_k$ are referred to as partial regression coefficients, X_1, X_2, \dots, X_k are independent variables and Y is the dependent variable.

3.5.3 Random Forest Model

Random forest is a supervised learning model for classification, regression, and other problems that functions by building a lot of decision trees during training phase. Ensembling is usually done using the concept of bagging with feature sets. The reason for using a large number of trees in the random forests is to train the trees enough such that contribution from each feature comes in a number of models.

3.6 Time Series Model

3.6.1 Dickey-Fuller test for testing stationarity

The first step to work on modelling is to make the time series stationary. Testing for stationarity is a frequently used activity in autoregressive modelling. We have used the Augmented DickeyFuller test for checking stationarity. A statistical technique called the Augmented Dickey-Fuller (ADF) test is used to detect if a time series variable is stationary or not. A key idea in time series analysis is stationarity, which denotes that the series' mean, variance, and autocorrelation are constant across time. We will have a pvalue from which we must draw conclusions about the time series,

3.6.2 ARIMA Model

A generalisation of an Autoregressive Moving Average (ARMA) model is the Auto Regressive Integrated Moving Average (ARIMA) model. To better comprehend the data or to forecast upcoming series points, both of these models are fitted to time series data (forecasting).

The equation of the ARIMA model given as:

There are three terms in the equation:

$$y'_t = c + \phi_1 * y'_{t-1} + \dots + \phi_p * y'_{t-p} + \theta_1 * \epsilon_{t-1} + \dots + \theta_q * \epsilon_{t-q} + \epsilon_t$$

AR (Auto Regression): The time series is regressed with its previous values i.e. y_{t-1} , y_t , etc. The order of the lag is denoted as p .

I(Integration): The time series uses differencing to make it stationary. The order of the difference is denoted as d .

MA (Moving Average): The time series is regressed with residuals of the past observations i.e. error ϵ_{t-1} , error ϵ_t .etc. The order of the error lag is denoted as q . In the above equation, y_t is the differenced series, ϕ_1 is the coefficient of the first AR term, p is the order of the AR term, θ_1 is the coefficient of the first MA term, q is the order of the MA term and ϵ_t is the error.

Chapter 4

Dataset and Exploratory Data Analysis

4.1 Definitions of attributes

Year (Ordinal)	The calendar year the indicators are from (ranging from 2000 to 2015)
Status (Nominal)	Whether a country is considered to be 'Developing' or 'Developed' by WHO standards
Life Expectancy (Ratio)	The life expectancy of people in years for a particular country and year
Adult Mortality (Ratio)	The adult mortality rate per 1000 population
Infant deaths (Ratio)	Number of infant deaths per 1000 population; similar to above, but for infants
Alcohol (Ratio)	A country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
Percentage expenditure (Ratio)	Expenditure on health as a percentage of Gross Domestic Product (GDP)
Hepatitis b (Ratio)	Number of 1 year olds with Hepatitis B immunization over all 1 year olds in population

Measles (Ratio)	Number of reported Measles cases per 1000 population
BMI (Interval/Ordinal)	Average Body Mass Index (BMI) of a country's total population
Under-five deaths (Ratio)	Number of people under the age of five deaths per 1000 population

Polio (Ratio)	Number of 1 year olds with Polio immunization over the number of all 1 year olds in population
Total expenditure (Ratio)	Government expenditure on health as a percentage of total government expenditure
Diphtheria (Ratio)	Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds
HIV/aids (Ratio)	Deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births
GDP (Ratio)	Gross Domestic Product per capita
Population (Ratio)	Population of a country
Thinness_1-19_years (Ratio)	Rate of thinness among people aged 10-19 (Note: variable should be renamed to <i>thinness_10-19_years</i> to

	more accurately represent the variable)
Thinness_5-9_years (Ratio)	Rate of thinness among people aged 5-9
Income composition of resources (Ratio)	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
schooling (Ratio)	Average number of years of schooling of a population

4.2 Data Sample

Country	Year	Status	Life expect	Adult Mort	infant dea	Alcohol	percentag	Hepatitis E	Measles	BMI
Afghanista	2015	Developing	65	263	62	0.01	71.27962	65	1154	19.1
Afghanista	2014	Developing	59.9	271	64	0.01	73.52358	62	492	18.6
Afghanista	2013	Developing	59.9	268	66	0.01	73.21924	64	430	18.1
Afghanista	2012	Developing	59.5	272	69	0.01	78.18422	67	2787	17.6
Afghanista	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	17.2
Afghanista	2010	Developing	58.8	279	74	0.01	79.67937	66	1989	16.7
Afghanista	2009	Developing	58.6	281	77	0.01	56.76222	63	2861	16.2
Afghanista	2008	Developing	58.1	287	80	0.03	25.87393	64	1599	15.7
Afghanista	2007	Developing	57.5	295	82	0.02	10.91016	63	1141	15.2

This figure is a sample of the dataset with 11 columns and 10 rows.

4.3 Data Pre-Processing

To make the original data fit the regression model, the cleaning and transformation process was done manually by selecting relevant fields and values. The process was as follows:

- Cleaning began with removal of null values. 14 variables has missing values (life expectancy, adult mortality, alcohol, hepatitis b, BMI, polio, total expenditure, population, GDP, thinness 1-19 years, thinness 5-9 years, income composition of resources, schooling). After removing the null values dataset is reduced to 1649 observations.
- Then we have identified Outliers using Box Plot. But we cannot remove the outlier because of the status of each Countries (Developed and Developing).

4.4 Descriptive Statistics

Descriptive statistics is a branch of statistics that involves the collection, analysis, and presentation of data in a meaningful way. Descriptive statistics summarizes and describes the main features of a dataset, providing a way to understand the data and make informed decisions.

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS
count	704.00000	704.000000	704.000000	704.000000	662.000000	704.000000	649.000000	704.000000	704.000000	704.000000	704.000000	658.000000	704.000000	704.000000
mean	2007.50000	71.175284	133.276989	64.697443	1.996994	493.776972	83.841294	4461.759943	34.680398	83.250000	85.696023	4.827614	86.313920	0.158523
std	4.61305	5.782441	78.233274	217.157613	2.406438	1169.208395	22.420045	13953.191523	19.503336	287.251858	21.370103	1.982878	19.469553	0.199818
min	2000.00000	54.800000	1.000000	0.000000	0.010000	0.000000	2.000000	0.000000	1.000000	0.000000	5.000000	1.170000	5.000000	0.100000
25%	2003.75000	66.675000	74.000000	1.000000	0.112500	6.060366	78.000000	9.750000	17.275000	1.000000	82.000000	3.400000	81.750000	0.100000
50%	2007.50000	72.700000	132.000000	7.000000	1.375000	63.813282	95.000000	140.000000	33.450000	8.000000	95.000000	4.480000	95.000000	0.100000
75%	2011.25000	74.900000	189.000000	30.000000	2.757500	283.449078	98.000000	1822.750000	52.800000	37.000000	98.000000	5.937500	98.000000	0.100000
max	2015.00000	87.000000	321.000000	1800.000000	13.030000	9498.729062	99.000000	131441.000000	71.400000	2500.000000	99.000000	13.730000	99.000000	2.200000

	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
	612.000000	4.960000e+02	704.000000	704.000000	704.000000	704.000000
	7553.961337	2.829561e+07	7.417330	7.663352	0.646085	11.775994
	13860.093640	1.234376e+08	6.165839	6.284103	0.185160	2.194343
	1.681350	3.400000e+01	0.900000	1.000000	0.000000	0.000000
	538.374968	1.271760e+05	2.800000	3.000000	0.572750	10.500000
	1770.871290	1.874962e+06	5.150000	5.100000	0.674500	12.000000
	5691.199648	1.268933e+07	11.000000	11.200000	0.765250	13.300000
	88564.822980	1.293859e+09	27.700000	28.600000	0.924000	16.100000

4.5 Exploratory Data Analysis

We do some visualizations to obtain some basic insights from the dataset to give ideas regarding the Dataset.

4.5.1 Univariate Visualisations

The Figure 4.1 below shows us that there are more developing countries than developed countries.

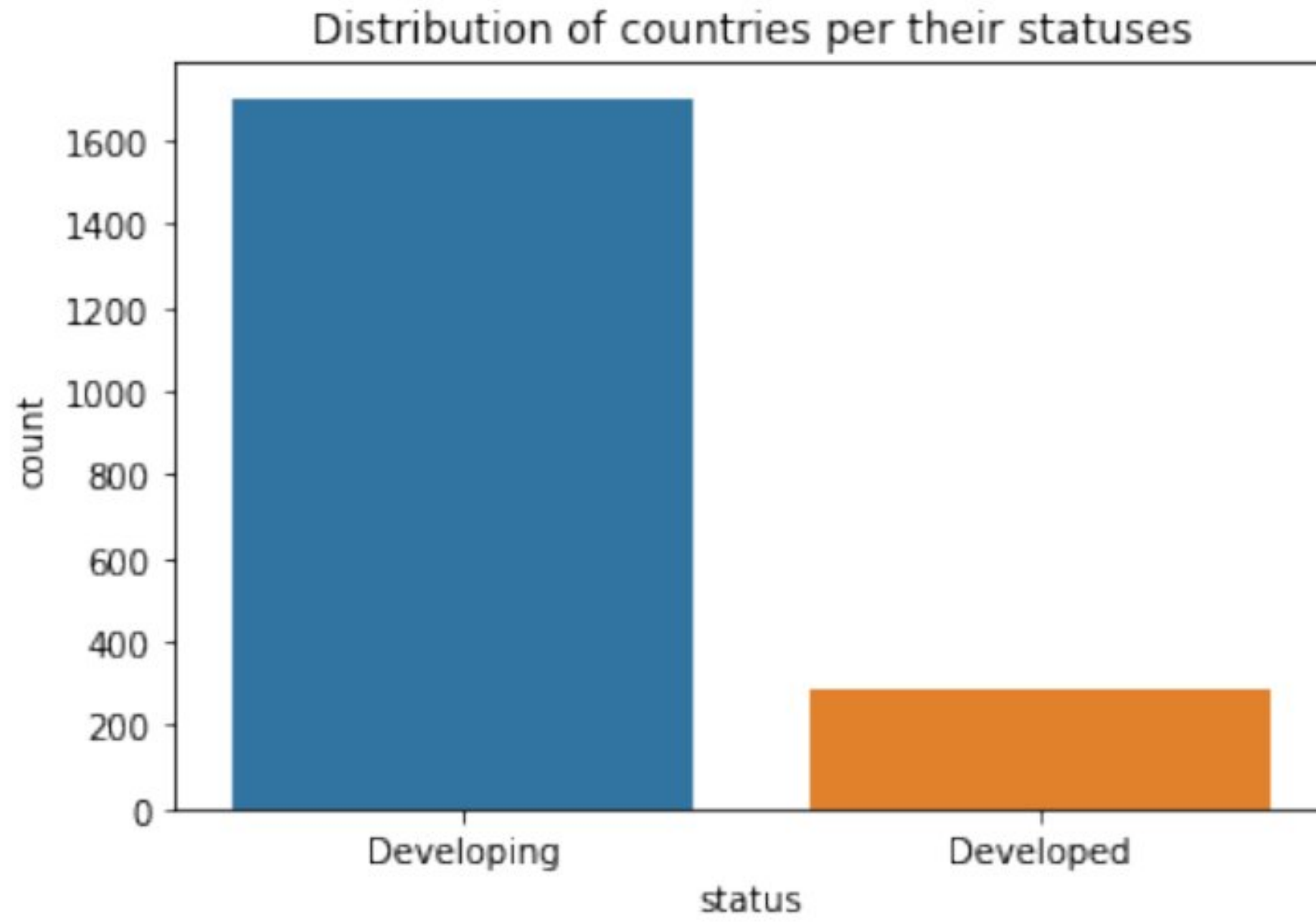


Figure 4.1: Developing vs Developed

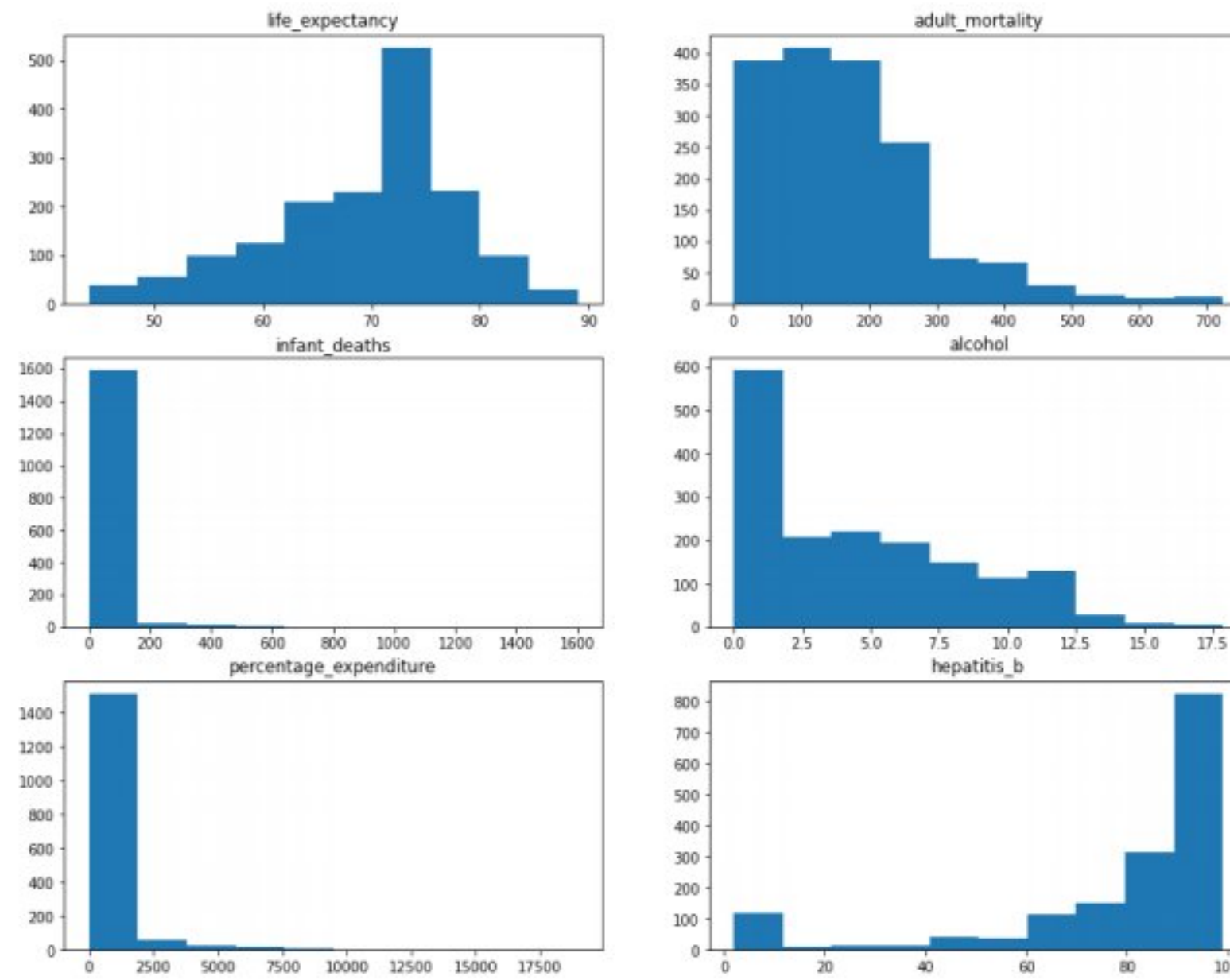


Figure 4.2: Bar graphs

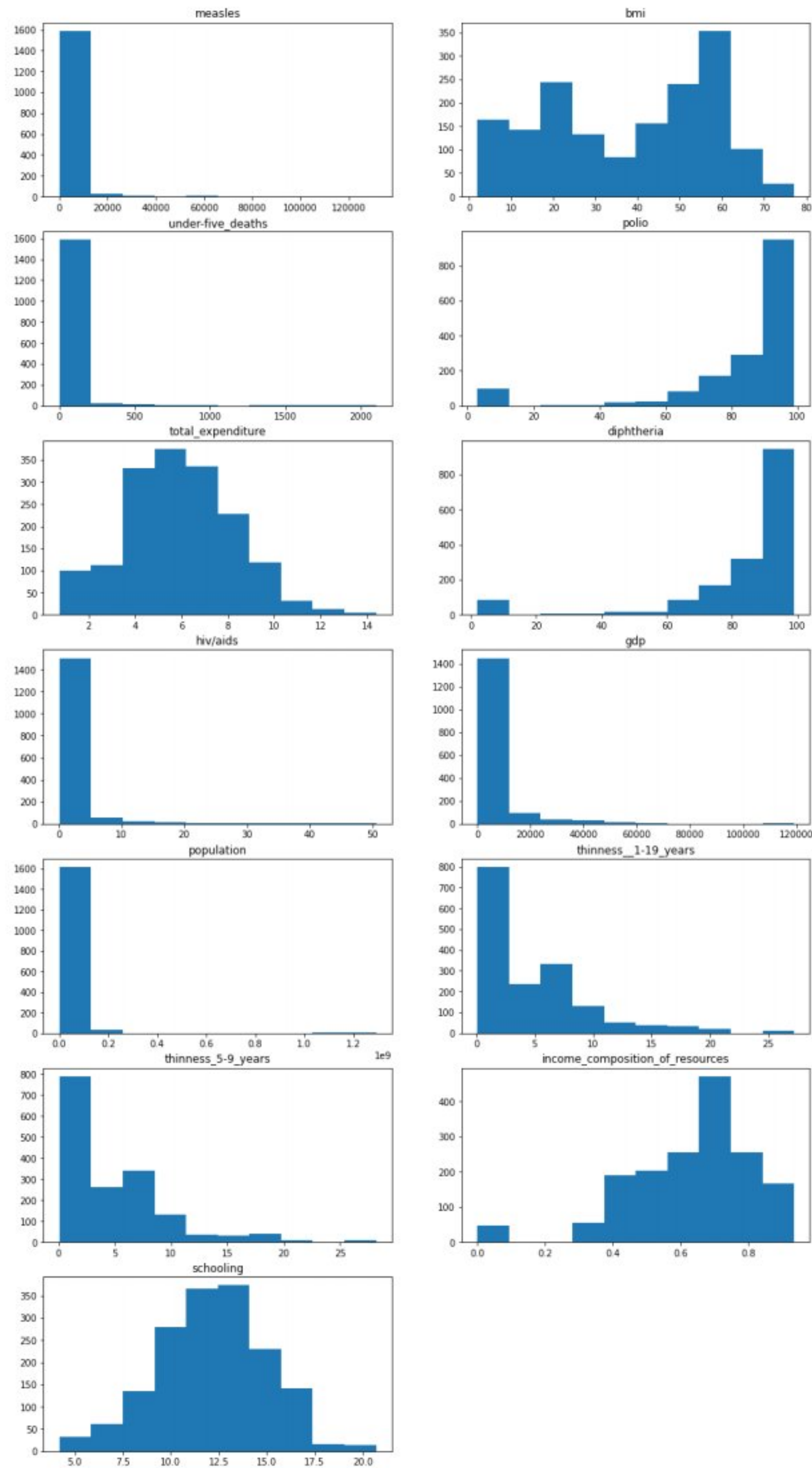


Figure 4.3: Bar graphs

Figure 4.2 and Figure 4.3 show that the columns life expectancy, hepatitis B, BMI, polio, diphtheria, income composition of resources, and schooling are negatively skewed to the left, while the rest are positively skewed to the right.

4.5.2 Multivariate Visualisations

A heat map is a visual representation of data that uses color coding to display the intensity, concentration, or density of values in a dataset. Heat maps are often used to visualize large and complex datasets, making it easier to identify patterns and trends. A Correlation heatmap showing Pearson’s coefficients for the variables was plotted and the findings were as follows:



- Life expectancy has a negative correlation with adult mortality, HIV/aids and thinness of both 10-19 and 5-9 years.
- Life expectancy has positive correlations with BMI, schooling and income composition of resources.
- There is a positive correlation between adult mortality and HIV/aids.
- There is a positive correlation between infant deaths and measles, population, and thinness of both 10-19 years and 5-9 years.
- The positive high correlation between alcohol and income composition of resources and schooling.

- The very high positive correlation between percentage expenditure and the GDP.
- Hepatitis B is positively correlated with polio and diphtheria.
- BMI is positively correlated with schooling and income compositions of the resources and negatively correlated with the thinness of both 10-19 years and 0-5 years.

Chapter 5

RESULTS AND DISCUSSION

5.1 Factor Analysis

Original Eigen values

0	5.541837
1	2.779679
2	1.840680
3	1.496183
4	1.183052
5	0.877362
6	0.793800
7	0.658736
8	0.523713
9	0.512200
10	0.418448
11	0.383610
12	0.347517
13	0.331164
14	0.199853
15	0.070857
16	0.038889
17	0.002420

Here, we can see only for 5 factors eigenvalues are greater than one. It means we only need to choose 5 factors (or unobserved variables). In a standard normal distribution, the mean will be 0, and Standard deviation will be 1, and the variance will be 1. This is the reason for

selecting factors whose eigenvalues are greater than 1.

	Factor1	Factor2	Factor3	Factor4	Factor5
adult_mortality	-0.254474	-0.028083	-0.121766	-0.133926	0.737567
infant_deaths	-0.188340	0.974498	-0.118145	-0.022586	0.003226
alcohol	0.560289	0.046561	0.199740	0.345987	0.044013
percentage_expenditure	0.219896	-0.039144	0.014603	0.917431	-0.081857
hepatitis_b	0.058711	-0.165404	0.654251	-0.032026	-0.049416
measles	-0.063150	0.503061	-0.058231	-0.046115	-0.012019
bmi	0.611919	-0.110399	0.097480	0.096159	-0.234474
under-five_deaths	-0.191196	0.959439	-0.135410	-0.020605	0.022850
polio	0.161092	-0.047436	0.677455	0.074510	-0.081433
total_expenditure	0.242774	-0.094581	0.115179	0.136628	0.055322
diphtheria	0.152717	-0.045454	0.819993	0.065353	-0.067311
hiv/aids	-0.087354	-0.012161	-0.057906	-0.009033	0.704090
gdp	0.247997	-0.036986	0.043503	0.940158	-0.094031
population	-0.070343	0.666409	-0.013525	0.018946	-0.033901
thinness_1-19_years	-0.815483	0.340738	0.009721	-0.020925	0.117108
thinness_5-9_years	-0.797584	0.341491	0.008472	-0.021289	0.135209
income_composition_of_resources	0.615226	0.048289	0.290723	0.304726	-0.268121
schooling	0.659028	-0.014155	0.317113	0.323717	-0.217430

From this figure we can say that,

- Factor 1 has high factor loadings for alcohol, BMI, thinness 10-19, thinness 5-9 and schooling.
- Factor 2 has high factor loadings for infant deaths, under-five deaths and population.
- Factor 3 has high factor loadings for hepatitis-b, polio, diphtheria.
- Factor 4 has high factor loadings for percentage expenditure, GDP.
- Factor 5 has high factor loadings for adult mortality, HIV/aids.

	Factor1	Factor2	Factor3	Factor4	Factor5
SS Loadings	3.177210	2.861406	1.863444	2.104025	1.281857
Proportion Var	0.176512	0.158967	0.103525	0.116890	0.071214
Cumulative Var	0.176512	0.335479	0.439003	0.555894	0.627108

In this figure, it shows us the first row represents the variance explained by each factor. Proportional variance is the variance explained by a factor out of the total variance. Cumulative variance is nothing but the collective sum of proportional variances of each factor. In our case, the 5 factors together are able to explain 62.7% of the total variance.

Findings:

By multivariate analysis, we are able to perform reduction techniques using Factor analysis. Five factors that we obtained can be named as:

- Nutrition (alcohol, BMI, thinness 10-19. thinness 5-9 and schooling)
- Infant deaths per population (infant deaths, under-five deaths and population)
- Immunizable diseases (hepatitis-b, polio, diphtheria)
- Economic factors (percentage Expenditure, GDP)
- Death rates (adult mortality, HIV/aids)

These factors are considered to have a high effect on the life expectancy of the countries.

5.2 Testing Of Hypothesis

We have to investigate that there is any significant difference in the mean life expectancy between developing and developed countries at a significance level of 5%.

Here,

Null Hypothesis, H_0 : The mean life expectancy is equal for developed and developing countries.

v/s

Alternate Hypothesis, H_1 : The mean life expectancy is not equal for developed and developing countries.

By D'Agostino's K2 test, both the population is normally distributed. By using a Z statistic and 5% level of significance, we can test our hy-

pothesis. For this, Critical region for a two-tailed z-test at 0.05 significance level is 1.960. Therefore, Reject the null hypothesis if $Z < -1.960$ or is $Z > 1.960$. For our dataset Z- score is -6.324784780199424.

Findings:

- The z-score is less than -1.960.
- Therefore, we reject the null hypothesis and accept the alternate hypothesis that the mean life expectancy between the developed and developing countries is different.
- Difference in the Mean Life Expectancy of Developed and Developing Countries is 12.191388088032838.

5.3 Logistic Regression Model

The aim is to reveal the class (Developed or Developing) that will occur when a set of x values that are given, to predict a classifier.

For this data, we will examine the status of countries Developed = 0 and Developing = 1.

There are 1407 Developing Countries and 242 Developed Countries.

Here Dependent variable is Status of a country and other variables are independent variables.

The intercept of the model is 5.83327627 and the Coefficient values are 1.08208352, -1.39357844, 3.4092507, 0.470897, -5.35962717, -0.78837698, -1.44070589, 0.37059599, -0.05002695, 0.43758173, 0.22862697, -0.10185549, 0.72963257, 0.51727771, -1.44175632, 0.42676196, 1.95007174, 1.79088634, -1.29839592, -1.92931239

Confusion matrix is given by,

169	73
45	1362

- TP (True Positive)(169): the number of correct predictions that the positive class is positive

- **FP (False Positive)(79):** the number of incorrect predictions that the negative class is positive
- **TN (True Negative)(1362):** the number of correct predictions that the negative class is negative
- **FN (False Negative)(45):** the number of incorrect predictions that the positive class is negative

	precision	recall	f1-score	support
0	0.79	0.70	0.74	242
1	0.95	0.97	0.96	1407
accuracy			0.93	1649
macro avg	0.87	0.83	0.85	1649
weighted avg	0.93	0.93	0.93	1649

Findings:

- Accuracy percentage of the model is 92.8%.
- The first 10 values of test data are predicted correctly.

5.4 Fitting of Models and Prediction With Best Model

The models are fitted for the data of the Country INDIA which is a Developing one.

RMSE value:

RMSE stands for Root Mean Squared Error, which is a commonly used metric to evaluate the accuracy of a predictive model. It is calculated as the average squared difference between both the expected values and the actual ones in a dataset. The RMSE value provides an indication of how well a predictive model performs in terms of predicting the outcome variable.

The model with lowest RMSE is said to be the best model.

5.4.1 Machine Learning Model (Random Forest Model)

The advantage of Random Forest algorithm is very fast and more robust than other regression models. This constructs a multitude of decision

trees and unites their outputs to make a final prediction.

RMSE value of the model is 0.4298348520071278.

5.4.2 Regression Model (Multiple Linear Regression)

Multiple regression models allow you to analyse the relative influences of the independent variables on the dependent variable, these often complex data sets can lead to false conclusions if they aren't analysed properly.

RMSE value of the model is 0.6903477196210235.

5.4.3 Time Series Model (ARIMA Model)

ARIMA (AutoRegressive Integrated Moving Average) model is a popular time series forecasting model used to predict future values of a time series variable based on its past values and it should work better for relatively short series when the number of Observations is not sufficient to apply more flexible methods. Here the most accurate model is ARIMA(1,0,1).

RMSE value of the model 2.0303940504246967.

5.4.4 Prediction with Best Model

If we compare three of the Model using the RMSE value Model with the lowest RMSE value is said to be the best. For our dataset, the Machine Learning model (Random Forest Model) has the lowest RMSE value which is 0.4298348520071278.

The dataset is divided in to two set such as Training and Test dataset. Of the 16 observations it is divided into train data and test data with no. of observations 13 and 3 respectively.

Test data is predicted by Random Forest Model. Then the Predicted values are,

Predicted value	Actual value	Difference
66.863	67.6	0.737
63.376	63.3	-0.076
63.627	63.7	0.073

Chapter 6

Conclusion

To conclude, all the objectives of this project are achieved. One aim of this research was to analyse the most affected features on the outcome. The affected features are Nutrition, Infant deaths per population, Immunizable Diseases, Economic status, Death rates. Also we find that the Mean Life Expectancy of Developed countries are different from Mean Life Expectancy of Developing countries. The aim was achieved that to reveal the class (Developed or Developing) that will occur when a set of x values that are given, to predict a classifier. Accuracy percentage of the model is 92.8%. The best-performing model needed to be found next. Random forest outperforms all other models with an RMSE value 0.4298348520071278.

This study suggests that further research be conducted using more data, machine learning models, and deep learning models. By adding other features like environmental and geographical factors, there is still scope for growth. It is still up for debate whether or not these suggested qualities should be included in studies of life expectancy in this particular field.

Chapter 7

References

1. A Alshehri, F El-Halawany and H Abu-Zinadah(2021). Prediction of life expectancy in Saudi Arabia by 2030 using ARIMA models, *J. Phys.: Conf. Ser.* 1978 012060.
2. Maksuda Aktar Rubi, Md. Hasan Imam Bijoy and Abu Kowshir Bitt(2021). Life Expectancy Prediction Based on GDP and Population Size of Bangladesh using Multiple Linear Regression and ANN Model, DOI: 10.1109/ICCCNT51525.2021.957959.
3. Karna, Suresh and D'Odorico, Elisa. (2020). A research study on the variables affecting Life Expectancy Descriptive and inferential statistics with Excel and R. *Data, Models and Decisions -Professor Pompeo Dalla Posta.* 10.13140/RG.2.2.20690.38080.
4. Akansha Maity, Emelie Rhenman, and Elijah Sanders(2017). Factors Explaining Average Life Expectancy: An Examination Across Nations, *Econometric Analysis Undergraduate Research Papers* [138].
5. Vikram Bali, Deepti Aggarwal and Sumit Singh(2021). Life Expectancy: Prediction and Analysis using ML, DOI-10.1109/ICRITO51393.2021.9596123.
6. Yallamati Prakasarao and Arumalla Nagaraju. Life Expectancy Prediction using Machine Learning. *International Journal for Modern Trends in Science and Technology* 2022, 8(S08), pp. 114-119.

7. Surefunmi Idowu(2019). Analysis of Population Data using Multiple and Logistic Regression Model,
DOI:10.13140/RG.2.2.32336.15364.
8. Banhi Guha, Gautam Bandyopathyay (2016). Gold Price Forecasting Using ARIMA Model,
DOI:10.12720/joams.4.2.117-121
9. Shashank Gupta, Life Expectancy using Linear Regression.
10. Arum, Kingsley,N J Nnanyelu,Ugah T E,Oranye Ebele(2019). Statistical Study of Life Expectancy of Male and Female Children at Birth in Some Selected African Countries, Journal VL 2.
11. Miladinov Genu(2020). Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries, Springer.