Statistical Analysis of factors affecting the life expectancy

T.P.N. Silva

Department of Computer Science and Engineering

University of Moratuwa, Sri Lanka

Abstract

World Health Organization has introduced various indicators to determine the health level in countries. Among the introduced indicators, life expectancy is given the highest priority. Many countries focus towards identifying the factors affecting the life expectancy of the citizens and taking necessary steps to improve and adhere to the standards implemented by the World Health Organization.

Introduction

The project aims at identifying the factors affecting the life expectancy of citizens of a country based on the data collected over a duration of fifteen years since 2000 to 2015.

Data Set

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data set comprises of data on 193 countries all over the world during the period 2000 to 2015.

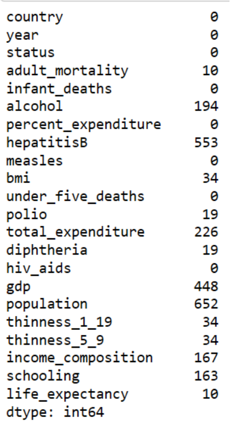
Data Column	Details	Data Type
Country	Country	String
Year	Year	Date
Status	Developed or Developing status	String
Life expectancy	Life Expectancy in age	Decimal
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)	Integer
Infant Deaths	Number of Infant Deaths per 1000 population	Integer
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)	Decimal
Percentage Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)	Decimal
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)	Integer
Measles	Measles - number of reported cases per 1000 population	Integer
BMI	Average Body Mass Index of entire population	Decimal

Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)	Integer
Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)	Decimal
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	Integer
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)	Decimal
GDP	Gross Domestic Product per capita (in USD)	Decimal
Population	Population of the country	Integer
thinness 1-19 years	Prevalence of thinness among children for Age 5 to 9(%)	Decimal
thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9(%)	Decimal
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Decimal
Schooling	Number of years of Schooling(years)	Decimal

Methodology and Results

1) Data Preprocessing

The data used for an analysis may contain duplicate values, missing values as they are user entered values. Before performing a data analysis the data set should be preprocessed by removing missing values or replacing the missing values with mean or by a given value so that the model is not affected and removing duplicate values etc.
The following features contained null values and such records were eliminated from the data set.



```
country                0
year                   0
status                 0
adult_mortality       10
infant_deaths          0
alcohol              194
percent_expenditure    0
hepatitisB           553
measles                0
bmi                   34
under_five_deaths      0
polio                 19
total_expenditure    226
diphtheria            19
hiv_aids               0
gdp                  448
population           652
thinness_1_19         34
thinness_5_9          34
income_composition   167
schooling            163
life_expectancy       10
dtype: int64
```

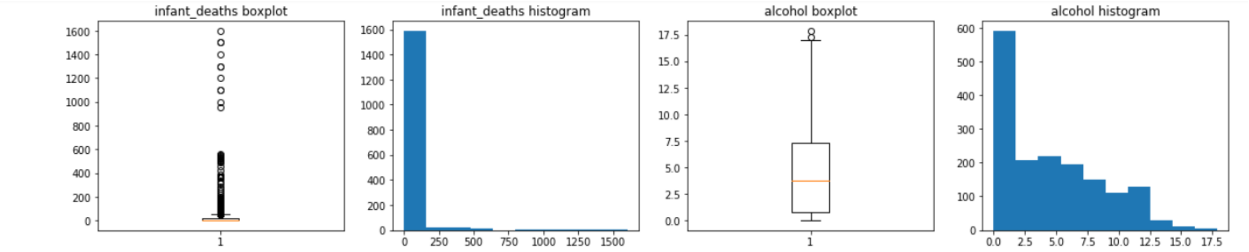
There are 4 phases in data analytics. They are;

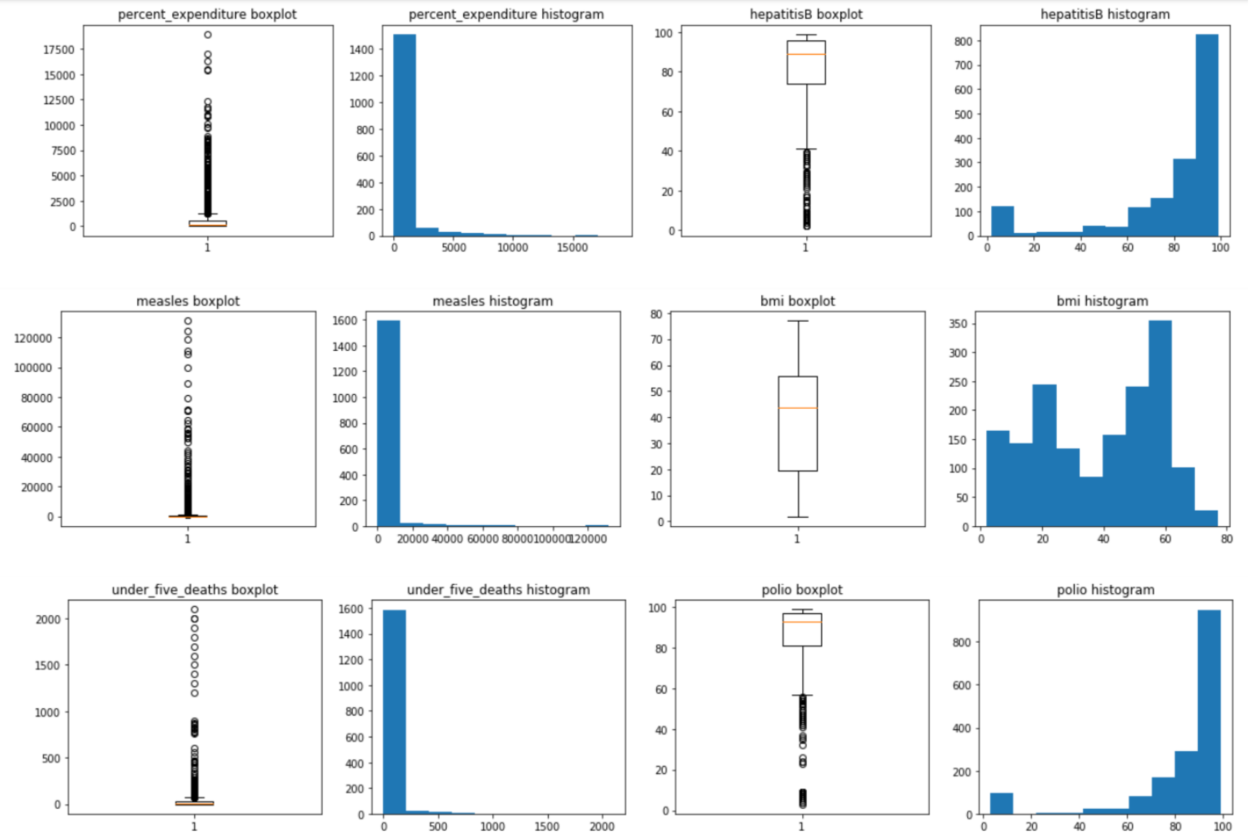
i) Descriptive Analysis
ii) Diagnostic Analysis
iii) Predictive Analysis
iv) Prescriptive Analysis

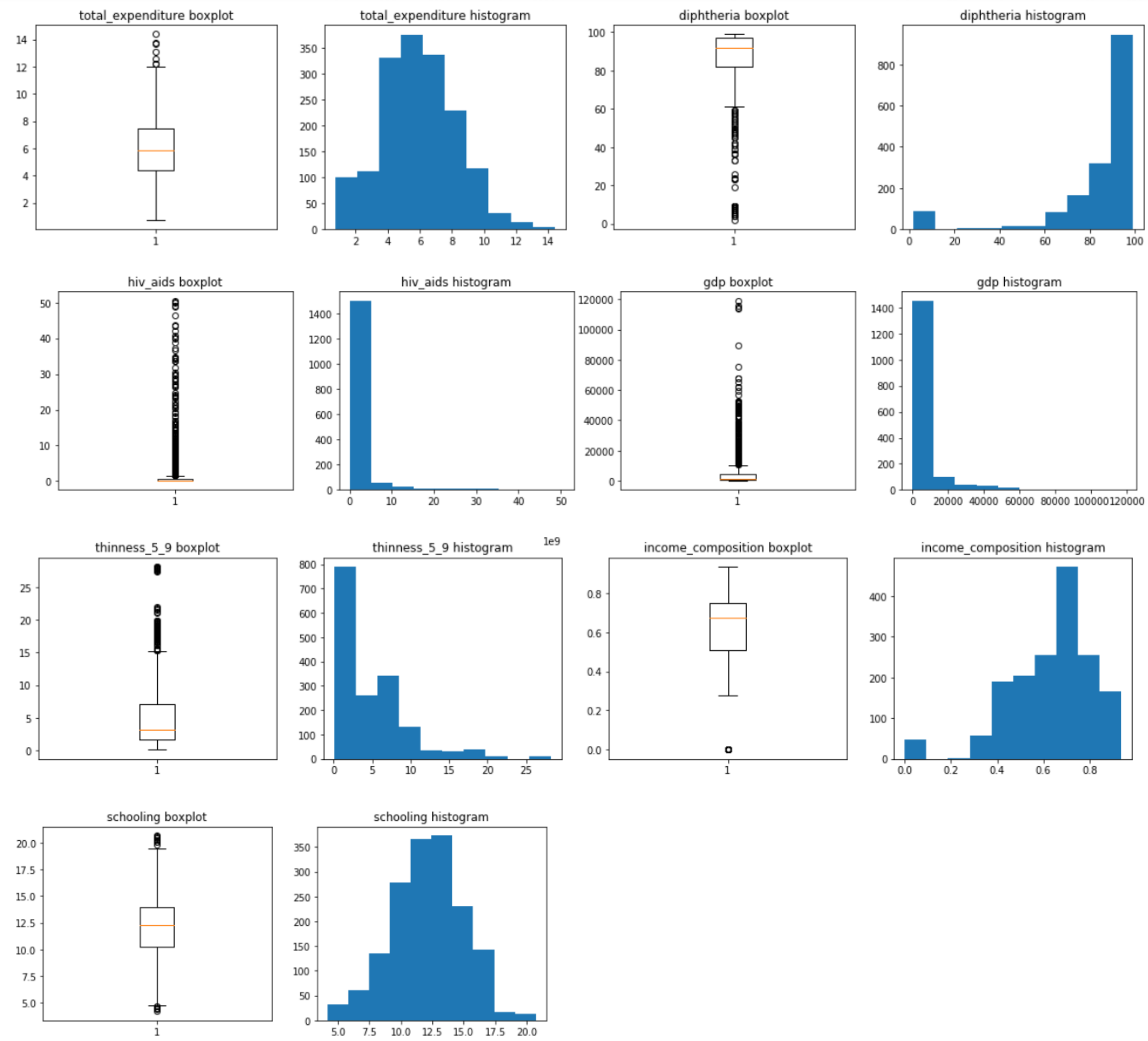
2) Descriptive Analysis

The initial step is to perform a descriptive analysis. Descriptive Analysis is the process of gathering and interpreting data to describe what has occurred.

The box plot and histogram techniques were used to identify the most significant outliers. From the results obtained the features such as ‘alcohol’, ‘bmi’, ‘total_expenditure’, ‘income_composition’ and ‘schooling’ can be selected.

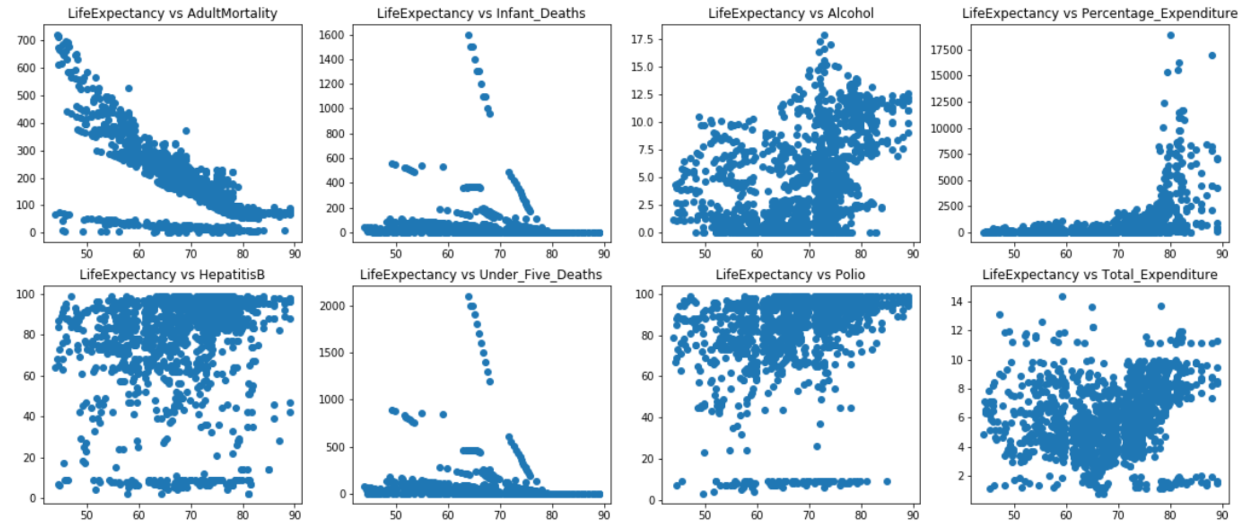


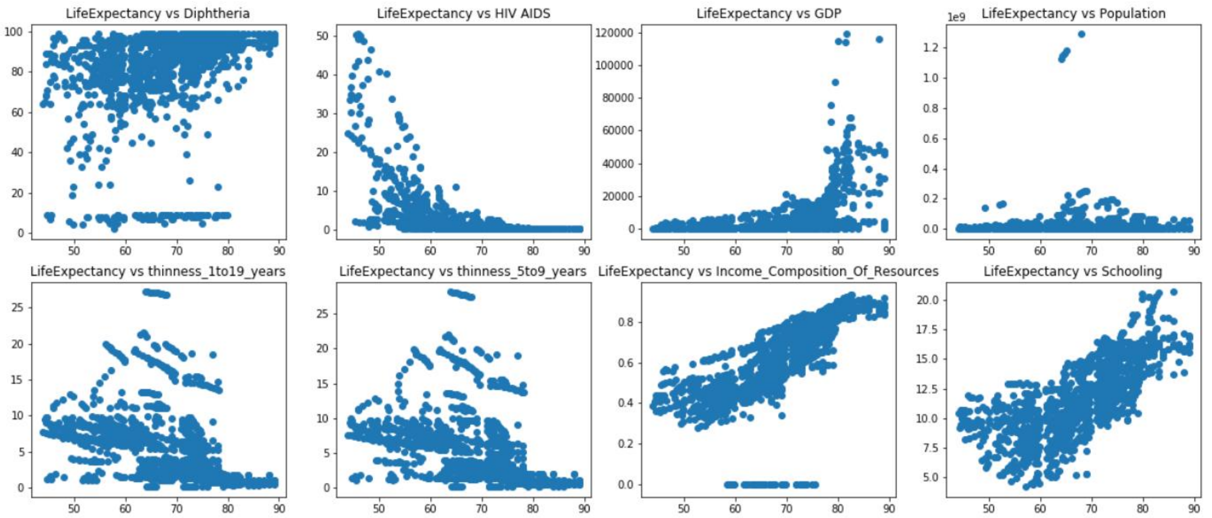




3) Diagnostic Analysis

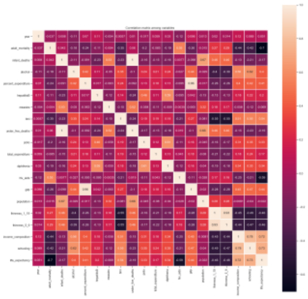
The next step is to perform a diagnostic analysis. In the process of diagnostic analysis, different anomalies and relationships among different features are identified. Here the technique of scatter plot was used to determine the correlation between the features selected and the target feature ‘life_expectancy’.



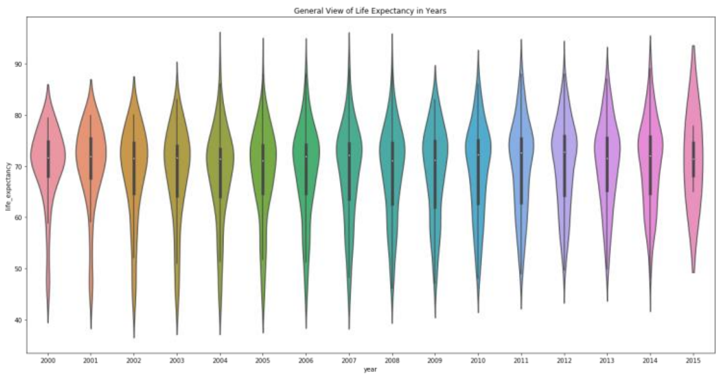


A highly positive correlation could be observed in the features ‘percentage_expenditure’, ‘gdp’, ‘schooling’.

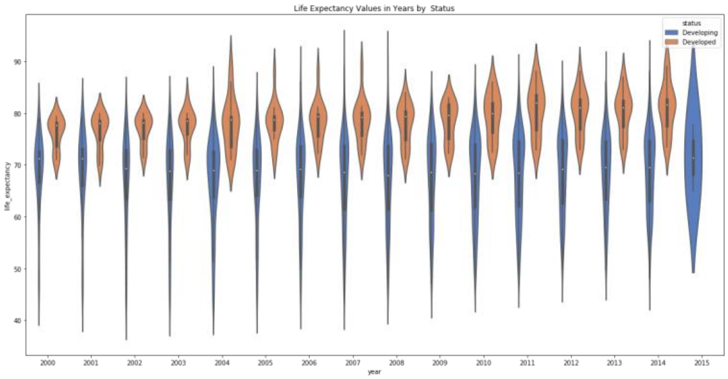
Furthermore a heatmap was generated to identify the correlation values for the features. According to the heatmap the top three highly positive correlated features with the target feature ‘life_expectancy’ were 'schooling' (+0.73), 'income_composition' (+0.72) and 'bmi'(+0.54).

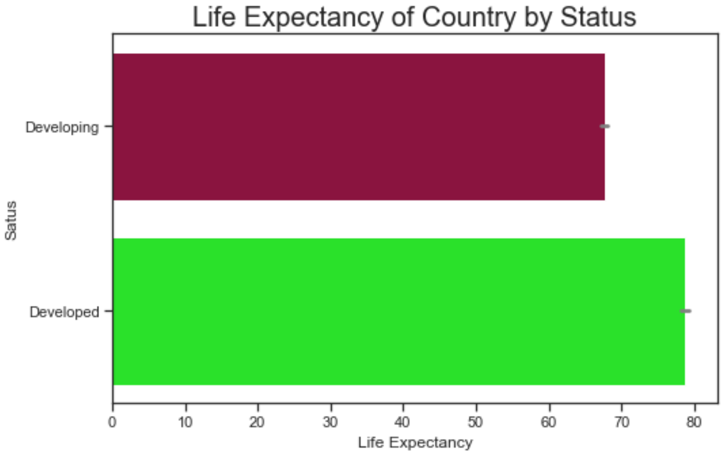


According to the violin plot created based on the life expectancy of all countries over the duration 2000-2015, it can be observed that the density of people with life expectancy greater than 70 is lower when compared to other years.

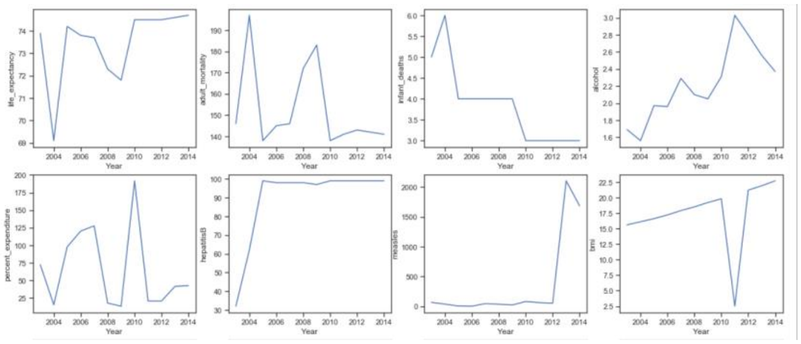


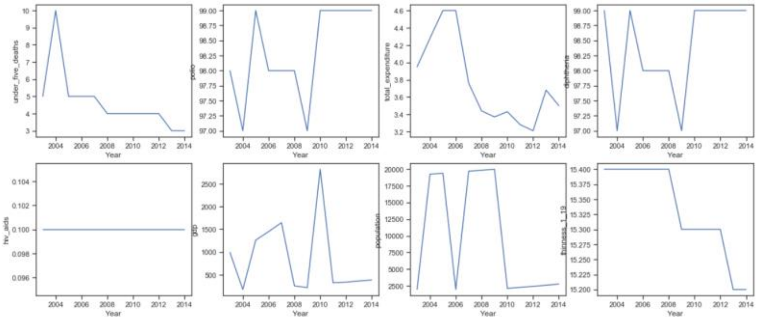
When the feature 'status' (country in 'developing' or 'developed' state) was selected to the violin plot, life expectancy of developed countries is higher than developing countries. The bar chart depicts the same.

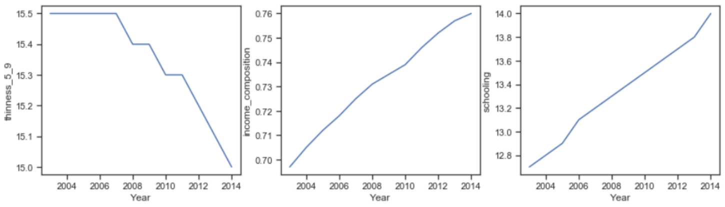




The dataset consisting of the same features was extracted separately for the country Sri Lanka during the period of 2004 – 2014 and then each feature was plotted against the target feature life expectancy over the period.







Accordingly, in 2014 the features adult mortality, infant deaths, under five year deaths, thinness 1-19, thinness 5-9 are lowest. The income composition and the schooling has increased in a directly proportional manner. Due to these reasons, the life expectancy is highest in 2014.