VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

VIT CHENNAI

Vandalur- Kelambakkam Road

Chennai - 600127

ANALYSIS OF LIFE EXPECTANCY

by

20BCE1235 - Khushi Kashyap
20BCE1722 - Priyanka Gyanchandani

J Component Report submitted to

Prof. Joshan Athanesious J
School of Computer Science and Engineering

In partial fulfilment of the requirements
for the course of
CSE3020 - DATA VISUALIZATION

in

B.TECH COMPUTER SCIENCE AND ENGINEERING
April 2023

TABLE OF CONTENTS

ABSTRACT

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-sets are made available to public for the purpose of health data analysis. The data-set is related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. A lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that the effect of immunization and human development index was not taken into account in the past. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

The goal is to find a set of features that affect Life Expectancy. The main idea of the project is that this project depends on the model's correctness. We want to use a variety of machine learning methods, including Multiple Linear Regression, Decision Tree, to predict life expectancy. We will assess each algorithm's accuracy before choosing the best one. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years.

OBJECTIVE

The objective of the study described is to analyze the factors that affect life expectancy in different countries, including immunization factors, mortality factors, economic factors, social factors, and other health-related factors. The study aims to consider important immunizations such as Hepatitis B, Polio, and Diphtheria, which have not been taken into account in previous studies. The goal is to determine the predicting factor(s) that contribute

to a lower life expectancy in different countries, by analyzing data collected from the WHO and United Nations data repositories.

By analyzing a life expectancy dataset is to gain insights into the factors that affect life expectancy and to identify any patterns or trends in the data. By analyzing the dataset, we could determine which factors are most strongly correlated with life expectancy, such as healthcare access, income level, education level, and social determinants of health. This analysis could be used to inform public health policies and interventions aimed at improving life expectancy and reducing health disparities across populations. Other objectives could include identifying sub-groups with particularly low life expectancy and targeting interventions to these groups, or examining the impact of specific interventions or policies on life expectancy over time.

METHODS USED

As Life expectancy is the response variable and it is continuous, we need to perform a linear regression to figure out what are the main predictors.

- We need to try different models based on different types of analysis so that we can determine which are the main predictors. and the models are:
  - Clean and filter data - (Full Model)
  - perform Exploratory Data Analysis - (EDA Model)
  - Apply Feature selection methods (Feature selection model)
- After that the following models are applied:
  - Multiple linear regression
  - Random Forest

OUTCOME OF THE PROJECT

1. Identification of factors that are strongly associated with life expectancy, such as income level, education level, access to healthcare, and social determinants of health.
2. Identification of patterns or trends in the data that could be used to inform public health policies and interventions aimed at improving life expectancy and reducing health disparities.
3. Identification of sub-groups with particularly low life expectancy, such as certain racial or ethnic groups or geographic regions, which could be targeted with interventions aimed at improving health outcomes.

4. Assessment of the impact of specific interventions or policies on life expectancy over time, such as the impact of healthcare reform or social welfare programs on population health.
5. Development of predictive models or risk scores to identify individuals or populations at high risk of premature mortality, which could be used to target interventions and improve health outcomes.

SCOPE OF PROJECT

The project on life expectancy using data from the World Health Organization (WHO) could have various scopes depending on the objectives of the project. Here are some potential scopes:

1. Analyzing life expectancy trends: The project could involve analyzing the historical trends in life expectancy across countries or regions using WHO data. This could involve identifying the factors that have influenced changes in life expectancy over time.
2. Identifying key determinants of life expectancy: The project could involve using the WHO data to identify the key determinants of life expectancy. This could involve analyzing the relationship between life expectancy and various factors such as access to healthcare, economic development, education, and lifestyle factors.
3. Predicting future life expectancy: The project could involve developing models to predict future life expectancy in different countries or regions based on current trends and other factors.
4. Comparing life expectancy across countries: The project could involve comparing life expectancy across countries or regions to identify areas where improvements are needed. This could involve analyzing differences in life expectancy based on factors such as healthcare systems, socioeconomic status, and lifestyle factors.
5. Examining health inequalities: The project could involve examining health inequalities within countries or regions by analyzing differences in life expectancy based on factors such as income, education, and geographic location.

Overall, the scope of the project will depend on the specific research questions or objectives that the project aims to address.

1. Introduction

Life expectancy is a statistical measure that reflects the average number of years a person is expected to live based on a specific set of demographic factors, such as age, gender, and socioeconomic status. The analysis of life expectancy involves examining various factors that influence the length of a person's life, such as lifestyle choices, environmental factors, and medical advances.

Based on a particular set of demographic parameters, such as age, gender, and socioeconomic level, life expectancy is a statistical measure that shows the typical number of years a person is expected to live. Examining numerous elements that affect a person's life expectancy, such as lifestyle decisions, environmental influences, and medical advancements, is part of the examination of life expectancy.

Depending on a person's country of origin, race, and other demographic factors, life expectancy might vary significantly. Access to high-quality healthcare, educational attainment, income, nutrition, lifestyle choices like diet and exercise, and exposure to environmental toxins are all factors that affect life expectancy.

Examining life expectancy can give important information about a population's health and happiness. Researchers can pinpoint areas of healthcare achievement, like advancements in medical treatments, and areas in need of improvement, such minimising healthcare inequities among various socioeconomic groups, by looking at changes in life expectancy through time.

Furthermore, a population's health and well-being can be improved with the help of public health policies and treatments that are based on life expectancy analysis. To address health inequities, encourage healthy lifestyle choices, and enhance access to healthcare services, for instance, politicians may utilize this information to allocate resources.

Much research on the factors influencing life expectancy have been conducted in the past, taking into account demographic demographics, income distribution, and death rates. It was discovered that the impact of vaccinations and the human development index had not previously been taken into consideration. Hepatitis B, polio, and diphtheria vaccinations are also important to consider. In summary, this study will concentrate on aspects associated with immunisation, mortality, the economy, society, and other health-related factors. A country will find it easier to identify the predicting factor causing a lower value of life expectancy as the observations in this dataset are based on multiple countries. This will assist in recommending to a nation which region should be prioritized in order to effectively raise the population's life expectancy.

Overall, a population's health and well-being may be understood through the analysis of life expectancy, which also yields useful data that can be utilized to guide public health policies and actions.

The goal is to discover and examine the real influences on life expectancy of the predictive factors, such as mortality, vaccination, economic, social, and other health-related factors.
The project depends on the model's correctness. We want to use a variety of machine learning methods, including Multiple Linear Regression, Decision Tree and Random Forest Classifierto predict life expectancy. We will assess each algorithm's accuracy before choosing the best one.

2. LITERATURE REVIEW

Analysis of Life Expectancy
Life expectancy refers to the average period that a person is expected to live based on factors such as health, social, economic, and environmental conditions. It is an important measure of population health and is used to assess the well-being of a society. The analysis of life expectancy involves studying the factors that influence life expectancy, as well as the trends and patterns of life expectancy over time and across populations.
Numerous factors influence life expectancy, including genetics, lifestyle factors, socioeconomic status, and access to healthcare. In their study, Crimmins and Beltrán-Sánchez (2011) found that improvements in medical technology and public health interventions have contributed significantly to the increase in life expectancy over time. Moreover, they noted that the benefits of these improvements in life expectancy are not equally distributed across populations, with disparities in life expectancy observed among different socioeconomic groups.
One of the major factors affecting life expectancy is access to healthcare. Research has found that access to healthcare, including preventive care and timely medical treatment, is essential for improving life expectancy and reducing health disparities. Socioeconomic factors such as income, education, and occupation also play a crucial role in determining life expectancy . A study in the United States found that income inequality is a significant predictor of life expectancy, with individuals living in more unequal areas having shorter life expectancies compared to those living in more equal areas . Lifestyle factors such as tobacco use, physical activity, and diet are also important determinants of life expectancy .

Studies have shown that life expectancy has been increasing globally over the past few decades. For instance, Roser and Ortiz-Ospina (2019) reported that the global life expectancy at birth increased from 53 years in 1960 to 72 years in 2016. However, the increase in life expectancy is not uniform across regions and countries. For example, a study by Mackenbach et al. (2019) found that the life expectancy gap between the rich and poor in Europe is widening, with disadvantaged groups experiencing slower improvements in life expectancy compared to their affluent counterparts.

Some studies have found that life expectancy patterns vary by gender, with women generally living longer than men. Life expectancy patterns also vary by race and ethnicity, with disparities in life expectancy observed among different racial and ethnic groups . For example, in the United States, African Americans have a shorter life expectancy compared to White Americans, and this disparity is attributed to a variety of factors, including socioeconomic status, access to healthcare, and exposure to environmental and social stressors .

The relationship between life expectancy and economic development is complex, with studies showing mixed results. Some studies have found a positive association between income and life expectancy (e.g., Deaton, 2003), while others have found no significant relationship (e.g., Cutler et al., 2006). Furthermore, studies have shown that the relationship between economic development and life expectancy is mediated by factors such as access to healthcare, education, and sanitation (e.g., Drewnowski and Popkin, 1997).Some studies have found a positive association between economic development and life expectancy, while others have found no significant relationship . A study in China found that economic growth was positively associated with life expectancy, but the relationship was moderated by factors such as access to healthcare and social support . Another study in India found that economic growth had a positive impact on life expectancy only in states with higher levels of human development .

The analysis of life expectancy is an important area of research, with numerous factors influencing life expectancy, including genetics, lifestyle factors, socioeconomic status, and access to healthcare. While life expectancy has been increasing globally, there are disparities in life expectancy across populations, with disadvantaged groups experiencing slower improvements in life expectancy compared to their affluent counterparts. The relationship between life expectancy and economic development is complex, with mixed findings reported in the literature. Overall, understanding the factors that influence life expectancy and the trends and patterns of life expectancy over time and across populations is crucial for improving population health and well-being.

1. In their literature review, Kulkarni et al. (2018) examine the determinants of life expectancy in India, focusing on factors such as income, education, and healthcare access.
2. Kitagawa (2014) provides a review of trends in US life expectancy, including disparities in life expectancy based on race, ethnicity, and socioeconomic status.
3. Drefahl and Koupil (2013) review observational studies examining the socioeconomic and lifestyle factors that influence survival and healthy aging, such as physical activity and social support.

4. Gruenewald and Seeman (2010) review the literature on the role of social support in cardiovascular disease, including the protective effects of social networks and relationships.
5. Crimmins and Beltrán-Sánchez (2011) provide a review of mortality and morbidity trends in aging populations, including the concept of compression of morbidity and the potential for delaying the onset of chronic conditions.
6. Goldman and Smith (2011) review studies that explore the relationship between education and health, finding that higher levels of education are associated with better health outcomes.
7. Mackenbach et al. (2008) review studies examining socioeconomic inequalities in health across 22 European countries, highlighting the role of social determinants such as income, education, and occupation.
8. Ezzati and Riboli (2013) provide a review of the behavioral and dietary risk factors for noncommunicable diseases, such as smoking and poor diet, and the impact these risk factors have on life expectancy.
9. Johnson et al. (2000) review the relationship between marital status and mortality, finding that married individuals have lower mortality rates than unmarried individuals.
10. Cutler and Lleras-Muney (2010) provide a comprehensive review of the theories and evidence linking education and health, finding that education is a significant predictor of health outcomes and that policies aimed at improving educational attainment may have positive health effects.

3. Materials and Methods

3.1 Models Used (Algorithms/Pseudocode):

3.1.1 Multiple regression

Multiple regression goes one step further, and instead of one, there will be two or more independent variables. If we have an additional variable (let's say "experience") in the equation above then it becomes a multiple regression:

Income= b0+ b1*x1+ b2*x2+ e

Where,

Y= Output/Response variable

b0, b1, b2...= Coefficients of the model.

x1, x2,...= Various Independent/feature variable

Assumptions for Multiple Linear Regression:

- A linear relationship should exist between the Target and predictor variables.
- The regression residuals must be normally distributed.
- MLR assumes little or no multicollinearity (correlation between the independent variable) in data.

Algorithm for Multiple Linear Regression:

1. Data Pre-processing Steps
2. Fitting the MLR model to the training set
3. Predicting the result of the test set

3.1.2 Decision Tree

Decision tree algorithm is a machine learning algorithm that is commonly used for classification and regression tasks. Here is the basic algorithm for building a decision tree:

1. Begin with a dataset containing labeled examples (i.e., a set of input/output pairs).
2. Choose an attribute (or feature) that best splits the data into separate classes based on some criterion, such as information gain or Gini impurity.
3. Create a decision node for the chosen attribute, and split the data into subsets based on the attribute values.
4. Recursively apply the above steps to each subset until all data points in a subset belong to the same class or a stopping criterion is met.
5. Create a leaf node for each subset that contains data points of the same class, and label it with the class.
6. Use the decision tree to classify new examples by traversing the tree from the root to a leaf node based on the attribute values of the new example, and outputting the label of the leaf node.
7. Optional: prune the decision tree to reduce overfitting by removing nodes that do not contribute significantly to the accuracy of the tree.

The decision tree algorithm is a popular and effective way to build models for classification and regression problems. It is easy to understand and interpret, and can handle both categorical and continuous input features. However, it can be prone to overfitting and may not perform well on complex datasets with many features.

3.1.3 Random Forest Classifier

Random forest classifier is a popular machine learning algorithm used for classification problems. It is an extension of the decision tree algorithm that uses an ensemble of decision

trees to improve the accuracy and reduce overfitting. Here is the basic algorithm for building a random forest classifier:

1. Begin with a dataset containing labeled examples (i.e., a set of input/output pairs).
2. Choose a random subset of the input features for each decision tree.
3. Randomly select a subset of the training examples with replacement (i.e., bootstrap sampling) to create a new dataset for each decision tree.
4. Use the decision tree algorithm to build a tree using the random subset of input features and the new bootstrap dataset.
5. Repeat steps 2-4 to create a forest of decision trees.
6. To classify a new example, pass it through each decision tree in the forest and record the class predictions.
7. Use the majority vote of the decision trees to classify the new example.

Random forest classifier is a popular and effective way to build models for classification problems. It is robust to noise and can handle both categorical and continuous input features. It is also able to handle missing data and maintain accuracy on imbalanced datasets. However, it can be computationally expensive and may not perform well on high-dimensional datasets.

3.2 Dataset

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status for all countries. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website. Among all categories of health-related factors only those critical factors were chosen which are more representative. In the past 15 years , there has been a huge development in the health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from the years 2005-2020 for 193 countries for further analysis.

Data description

The publicly available dataset provides data for 193 countries spanning from year 2000 to year 2015 and is structured in 2938 rows (data points) which are characterized into a total of 22 columns (features). The features can be categorized into two groups:

- Health factors which are originally provided by the Global Health Observatory (GHO) data repository under the World Health Organization (WHO)
- Economic factors which have been collected by the United Nation (UN) website.

Variable description

1. country (Nominal) - the country in which the indicators are from (i.e. United States of America or Congo)
2. year (Ordinal) - the calendar year the indicators are from (ranging from 2000 to 2015)
3. status (Nominal) - whether a country is considered to be 'Developing' or 'Developed' by WHO standards
4. life_expectancy (Ratio) - the life expectancy of people in years for a particular country and year
5. adult_mortality (Ratio) - the adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population); if the rate is 263 then that means 263 people will die out of 1000 between the ages of 15 and 60; another way to think of this is that the chance an individual will die between 15 and 60 is 26.3%
6. infant_deaths (Ratio) - number of infant deaths per 1000 population; similar to above, but for infants
7. alcohol (Ratio) - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita.
8. percentage_expenditure (Ratio) - expenditure on health as a percentage of Gross Domestic Product (gdp)
9. hepatitis_b (Ratio) - number of 1 year olds with Hepatitis B immunization over all 1 year olds in population
10. measles (Ratio) - number of reported Measles cases per 1000 population
11. bmi (Interval/Ordinal) - average Body Mass Index (BMI) of a country's total population
12. under-five_deaths (Ratio) - number of people under the age of five deaths per 1000 population
13. polio (Ratio) - number of 1 year olds with Polio immunization over the number of all 1 year olds in population
14. total_expenditure (Ratio) - government expenditure on health as a percentage of total government expenditure
15. diphtheria (Ratio) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds
16. 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
17. gdp (Ratio) - Gross Domestic Product per capita
18. population (Ratio) - population of a country
19. thinness_1-19_years (Ratio) - rate of thinness a

Screenshots of the dataset:

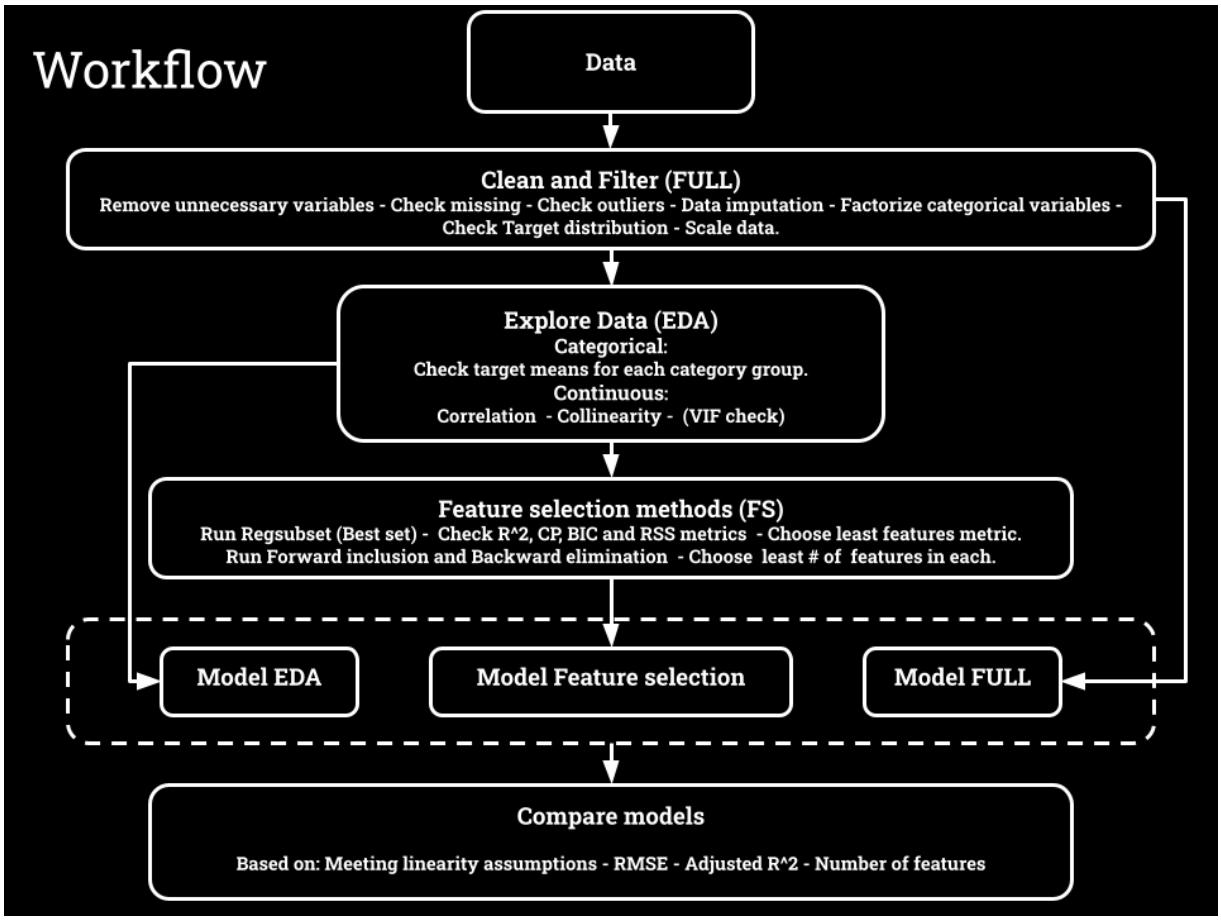
Life Expectancy Data

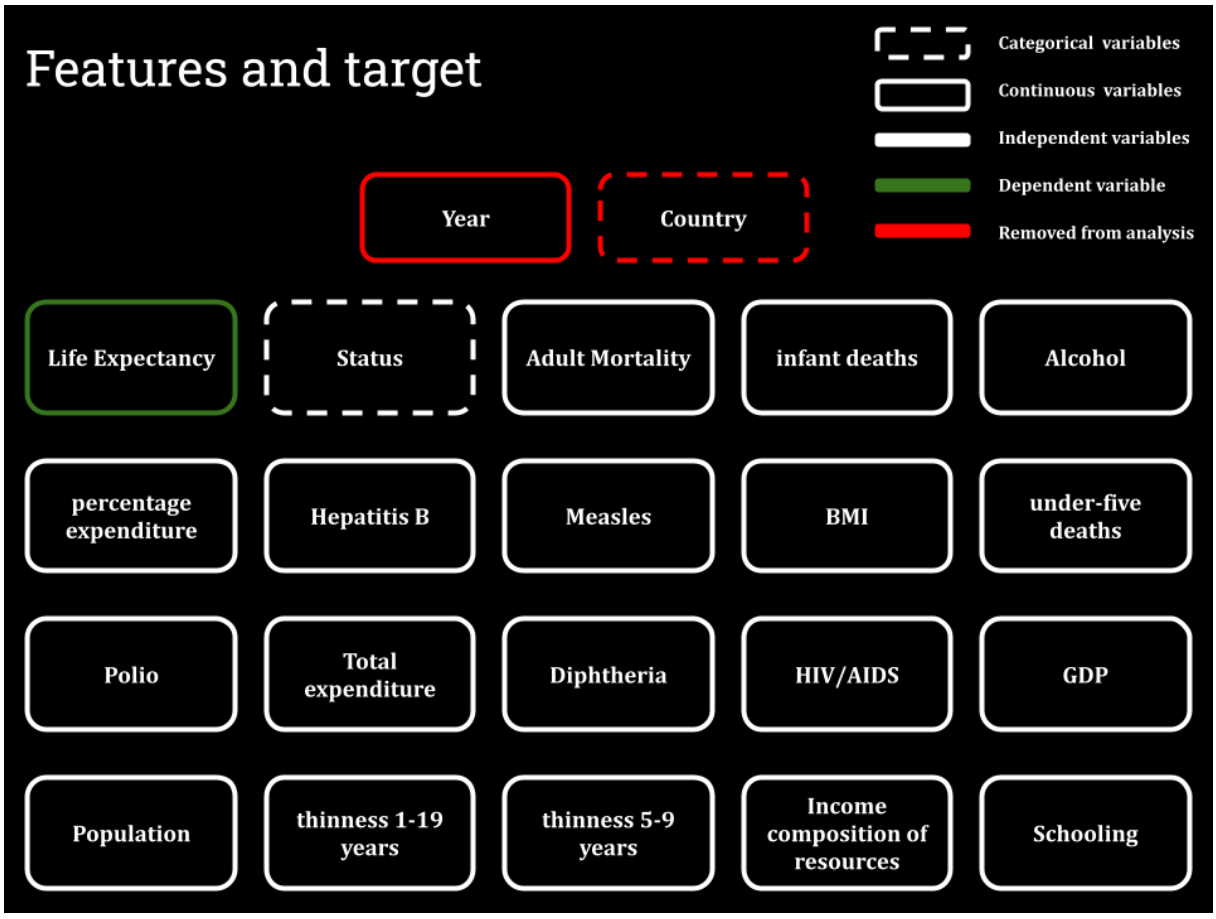
Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
Afghanistan	2015	Developing	65	263	62	0.01	71.27962362	65	1154	19.1	83	6	8.16	65	0.1	584.25921	33736494	17.2	17.3	0.479	10.1
Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358168	62	492	18.6	86	58	8.18	62	0.1	612.696514	327582	17.5	17.5	0.476	10
Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924272	64	430	18.1	89	62	8.13	64	0.1	631.744976	31731688	17.7	17.7	0.47	9.9
Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9	18	0.463	9.8
Afghanistan	2011	Developing	59.2	275	71	0.01	7.097108703	68	3013	17.2	97	68	7.87	68	0.1	63.537231	2978599	18.2	18.2	0.454	9.5
Afghanistan	2010	Developing	58.8	279	74	0.01	79.67936736	66	1989	16.7	102	66	9.2	66	0.1	553.32894	2883167	18.4	18.4	0.448	9.2
Afghanistan	2009	Developing	58.6	281	77	0.01	56.76221682	63	2861	16.2	106	63	9.42	63	0.1	445.8932979	284331	18.6	18.7	0.434	8.9
Afghanistan	2008	Developing	58.1	287	80	0.03	25.87392536	64	1599	15.7	110	64	8.33	64	0.1	373.3611163	2729431	18.8	18.9	0.433	8.7
Afghanistan	2007	Developing	57.5	295	82	0.02	10.91015598	63	1141	15.2	113	63	6.73	63	0.1	369.835796	26616792	19	19.1	0.415	8.4
Afghanistan	2006	Developing	57.3	295	84	0.03	17.17151751	64	1990	14.7	116	58	7.43	58	0.1	272.56377	2589345	19.2	19.3	0.405	8.1
Afghanistan	2005	Developing	57.3	291	85	0.02	1.388647732	66	1296	14.2	118	58	8.7	58	0.1	25.2941299	257798	19.3	19.5	0.396	7.9
Afghanistan	2004	Developing	57	293	87	0.02	15.29606643	67	466	13.8	120	5	8.79	5	0.1	219.1413528	24118979	19.5	19.7	0.381	6.8
Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905273	65	798	13.4	122	41	8.82	41	0.1	198.7285436	2364851	19.7	19.9	0.373	6.5
Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735091	64	2486	13	122	36	7.76	36	0.1	187.84595	21979923	19.9	2.2	0.341	6.2
Afghanistan	2001	Developing	55.3	316	88	0.01	10.5747282	63	8762	12.6	122	35	7.8	33	0.1	117.49698	2966463	2.1	2.4	0.34	5.9
Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496	62	6532	12.2	122	24	8.2	24	0.1	114.56	293756	2.3	2.5	0.338	5.5
Albania	2015	Developing	77.8	74	0	4.6	364.9752287	99	0	58	0	99	6	99	0.1	3954.22783	28873	1.2	1.3	0.762	14.2
Albania	2014	Developing	77.5	8	0	4.51	428.7490668	98	0	57.2	1	98	5.88	98	0.1	4575.763787	288914	1.2	1.3	0.761	14.2
Albania	2013	Developing	77.2	84	0	4.76	430.8769785	99	0	56.5	1	99	5.66	99	0.1	4414.72314	289592	1.3	1.4	0.759	14.2
Albania	2012	Developing	76.9	86	0	5.14	412.4433563	99	9	55.8	1	99	5.59	99	0.1	4247.61438	2941	1.3	1.4	0.752	14.2
Albania	2011	Developing	76.6	88	0	5.37	437.0621	99	28	55.1	1	99	5.71	99	0.1	4437.17868	295195	1.4	1.5	0.738	13.3
Albania	2010	Developing	76.2	91	1	5.28	41.82275719	99	10	54.3	1	99	5.34	99	0.1	494.358832	291321	1.4	1.5	0.725	12.5
Albania	2009	Developing	76.1	91	1	5.79	348.0559517	98	0	53.5	1	98	5.79	98	0.1	4114.136545	2927519	1.5	1.6	0.721	12.2
Albania	2008	Developing	75.3	1	1	5.61	36.62206845	99	0	52.6	1	99	5.87	99	0.1	437.539647	2947314	1.6	1.6	0.713	12
Albania	2007	Developing	75.9	9	1	5.58	32.24655228	98	22	51.7	1	99	6.1	98	0.1	363.13685	29717	1.6	1.7	0.703	11.6
Albania	2006	Developing	74.2	99	1	5.31	3.3021542	98	68	5.8	1	97	5.86	97	0.1	35.1293	2992547	1.7	1.8	0.696	11.4
Albania	2005	Developing	73.5	15	1	5.16	26.99312143	98	6	49.9	1	97	6.12	98	0.1	279.142931	311487	1.8	1.8	0.685	10.8
Albania	2004	Developing	73	17	1	4.54	221.8428	99	7	48.9	1	98	6.38	97	0.1	2416.588235	326939	1.8	1.9	0.681	10.9
Albania	2003	Developing	72.8	18	1	4.29	14.71928882	97	8	47.9	1	97	6.27	97	0.1	189.681557	339616	1.9	2	0.674	10.7
Albania	2002	Developing	73.3	15	1	3.73	104.5169157	96	16	46.9	1	98	6.3	98	0.1	1453.642777	3511	2	2.1	0.67	10.7
Albania	2001	Developing	73.6	14	1	4.25	96.20557078	96	18	46	1	97	6	97	0.1	1326.97339	36173	2.1	2.1	0.662	10.6
Albania	2000	Developing	72.6	11	1	3.66	91.71154052	96	662	45	1	97	6.26	97	0.1	1175.788981	38927	2.1	2.2	0.656	10.7
Algeria	2015	Developing	75.6	19	21		0	95	63	59.5	24	95		95	0.1	4132.76292	39871528	6	5.8	0.743	14.4
Algeria	2014	Developing	75.4	11	21	0.01	54.2373183	95	0	58.4	24	95	7.21	95	0.1	547.8517	39113313	6	5.8	0.741	14.4
Algeria	2013	Developing	75.3	112	21	0.53	544.4507432	95	25	57.2	24	95	7.12	95	0.1	5471.866766	38338562	5.9	5.8	0.737	14.4
Algeria	2012	Developing	75.1	113	21	0.66	555.9260834	95	18	56.1	24	95	6.14	95	0.1	5564.82566	37565847	5.9	5.8	0.732	14.4
Algeria	2011	Developing	74.9	116	21	0.56	509.0020405	95	112	55	24	95	5.29	95	0.1	5432.2523	36819558	5.9	5.8	0.724	14
Algeria	2010	Developing	74.7	119	21	0.45	430.7175861	95	103	53.9	24	95	5.12	95	0.1	4463.394675	36117637	5.9	5.8	0.714	13.6
Algeria	2009	Developing	74.4	123	20	0.5	352.0636419	94	107	52.8	23	94	5.36	95	0.1	3868.83123	3546576	6	5.9	0.705	13.1
Algeria	2008	Developing	74.1	126	20	0.46	43.08717334	91	217	51.8	23	92	4.2	93	0.1	495.254866	3486715	6	5.9	0.697	12.6
Algeria	2007	Developing	73.8	129	20	0.44	320.3239241	9	0	5.8	23	95	3.82	95	0.1	3935.183343	34376	6	5.9	0.69	12.3
Algeria	2006	Developing	73.4	132	20	0.36	270.2401962	8	944	49.8	23	95	3.36	95	0.1	3464.6179	33777915	6.1	6	0.686	12.3
Algeria	2005	Developing	72.9	136	19	0.5	2.548922758	83	2302	48.9	22	88	3.24	88	0.1	31.122378	33288437	6.1	6	0.68	12
Algeria	2004	Developing	72.3	14	19	0.45	220.393699	81	3289	47.9	23	86	3.54	86	0.1	2598.9823	3283196	6.2	6.1	0.673	11.7
Algeria	2003	Developing	71.7	146	20	0.34	25.0185226		15374	47	23	87	3.6	87	0.1	294.33556	3243514	6.3	6.1	0.663	11.5
Algeria	2002	Developing	71.6	145	20	0.36	148.5119843		5862	46.1	23	86	3.73	86	0.1	1774.33673	3199546	6.3	6.2	0.653	11.1

3.3 Architecture and Explanation

As Life expectancy is the response variable and it is continuous, we need to perform a linear regression to figure out what are the main predictors.

- We need to try different models based on different types of analysis so that we can determine which are the main predictors. and the models are:
    - Clean and filter data - (Full Model)
    - perform Exploratory Data Analysis - (EDA Model)
    - Apply Feature selection methods (Feature selection model)
    - Apply various models like Multiple Linear Regression, Decision Tree and Random Forest classifier and choose the best one based on the accuracy percentage.
- The following figure represents the analysis workflow:


Workflow
Data
Clean and Filter (FULL)
Remove unnecessary variables - Check missing - Check outliers - Data imputation - Factorize categorical variables - Check Target distribution - Scale data.
Explore Data (EDA)
Categorical:
Check target means for each category group.
Continuous:
Correlation - Collinearity - (VIF check)
Feature selection methods (FS)
Run Regsubset (Best set) - Check R^2, CP, BIC and RSS metrics - Choose least features metric.
Run Forward inclusion and Backward elimination - Choose least # of features in each.
Model EDA
Model Feature selection
Model FULL
Compare models
Based on: Meeting linearity assumptions - RMSE - Adjusted R^2 - Number of features


Features and target
Categorical variables
Continuous variables
Independent variables
Dependent variable
Removed from analysis
Year
Country
Life Expectancy
Status
Adult Mortality
infant deaths
Alcohol
percentage expenditure
Hepatitis B
Measles
BMI
under-five deaths
Polio
Total expenditure
Diphtheria
HIV/AIDS
GDP
Population
thinness 1-19 years
thinness 5-9 years
Income composition of resources
Schooling

4.Proposed work

4.1 Novelty

To find the root cause and analyze various other factors affecting the life expectancy which will help the government of a particular country to decide where to invest in a particular sector to raise life expectancy.
To find whether the various other factors such as eating habits, lifestyle, exercise, smoking, drinking alcohol has positive or negative correlation with Life Expectancy.
Using the provided datasets, to determine whether nations with a high population density have shorter life expectancies or not?

4.2 Project Contributions

Team Members(s) Contributions:

Worklet Tasks	*Contributor's Names*
Data Cleaning	Khushi Kashyap
Data Preprocessing	Khushi Kashyap
Data Imputation	Priyanka Gyanchandani
Data Visualization	Priyanka Gyanchandani
Algorithm-1 (Multiple Linear Regression)	Khushi Kashyap
Algorithm-2 (Decision Tree)	Priyanka Gyanchandani
Algorithm-3 (K-nearest neighbor)	Priyanka Gyanchandani
Report and Presentation building	Khushi Kashyap, Priyanka Gyanchandani

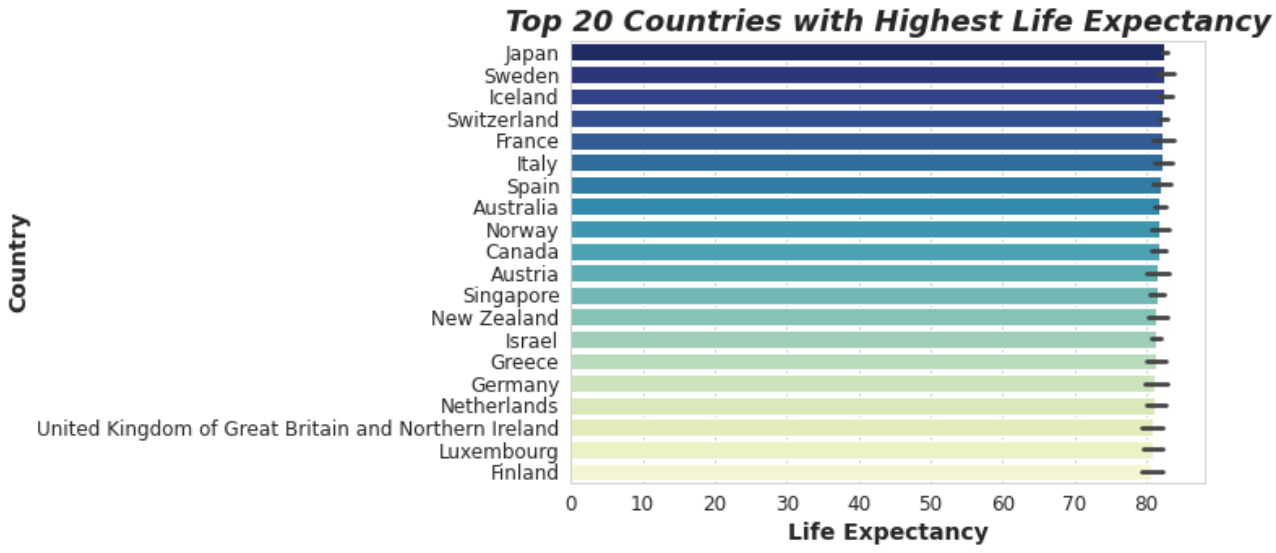
5. Results and Discussion

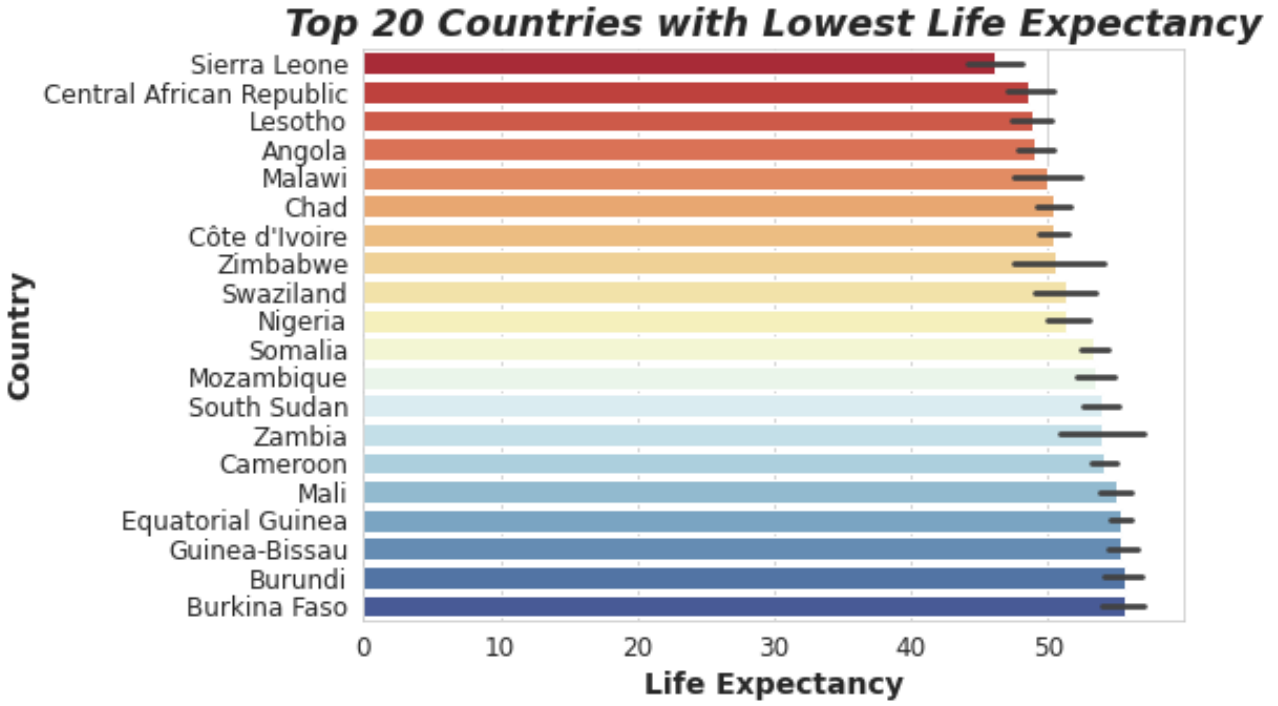
5.1 Results

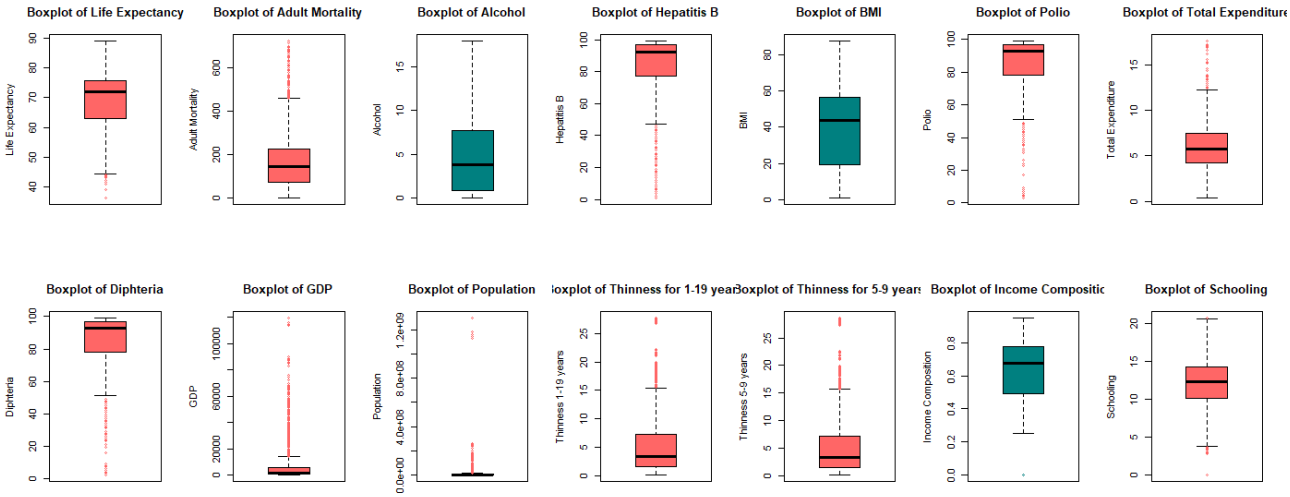
The mean absolute error and root mean square error for test data is 0.061 and 0.076
Based on performance measures among all models regression Tree shows least Mean absolute error and Root mean square error.So Regression tree is considered as best model to predict Life expectancy.

Developed countries seems to have successfully eradicated diphtheria disease because of vaccines whereas in developing countries there was low expectancy value initially but now it is gradually increasing maybe because of proper doses being given.

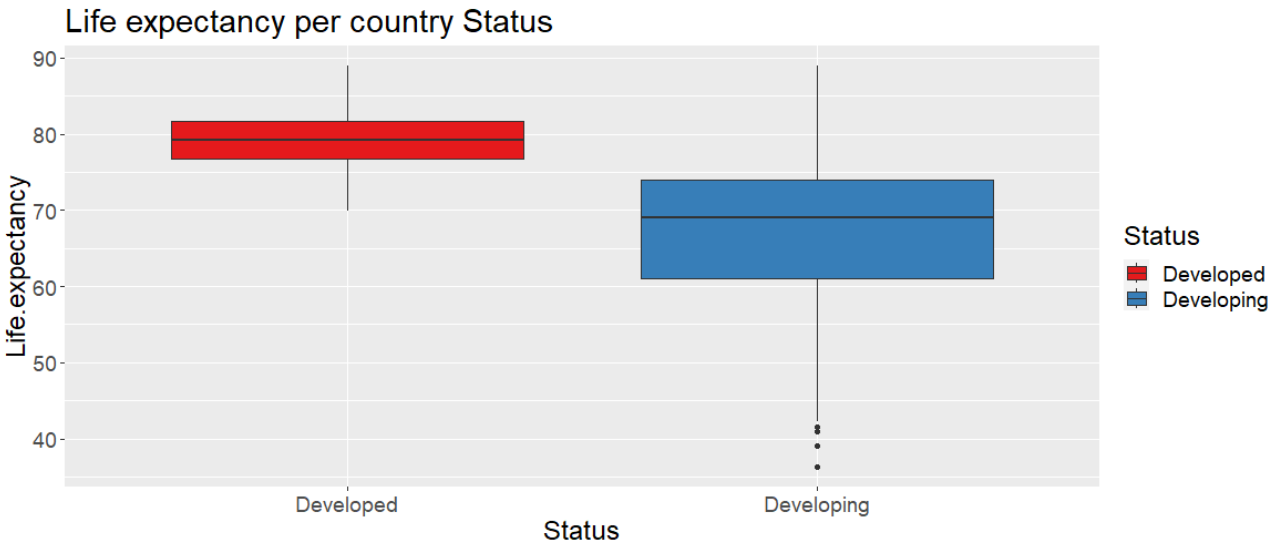
5.2 Figures, Comparison Tables



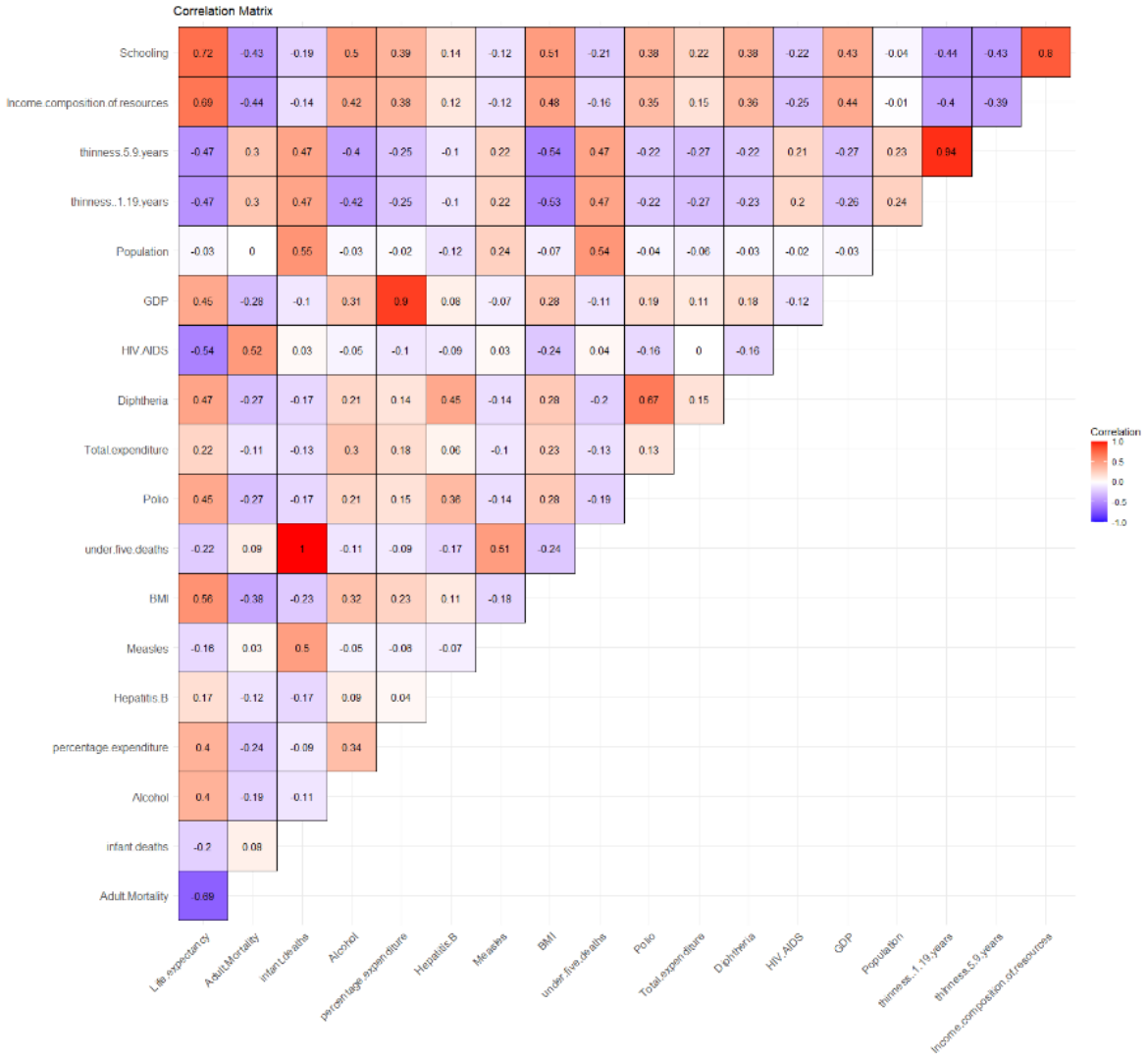




The boxplots that show none or few outliers are: Alcohol, BMI, Income.composition.of.resources.



Boxplots of life expectancy over the two categories developed and developing countries. We can see that the life expectancy in Developed countries is higher which means the categorical variable might be a good predictor for the model.



As

expected, can notice very strong correlation[>= 0.9] between:

- infant.deaths and under.five.deaths.
- GDP and percentage.expenditure.
- thinness..1.19.years and thinness.5.9.years.

Yet, correlation doesn't imply collinearity between independent variables.

5.3 Explanation

The dataset, although collected by WHO, contained a lot of missing values and we saw that most of the missing values were from countries with very less population and where data collection is a very tedious task. A lot of outliers were detected. It has been observed that in the past 15 years , there has been a huge development in the health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Japan is the country with the highest Life expectancy value followed by Sweden and Sierra Leone has the lowest Life expectancy value. Life Expectancy is inversely proportional to Adult mortality and does not have much relation with infant death.It is high with high BMI, which shows costly eating habitsIt is inversely related to thinness, which shows lower eating habits results in lower Life Expectancy. It is high in developed countries, lifestyle is generally higher in developed countries.

We largely saw how developing countries have very less life expectancy when we see diseases like HIV/AIDS, polio etc and how Schooling plays a big role in increasing the life expectancy of developing countries as people become much more educated and help improve the welfare and healthcare of the country along with the economy. There is a slight decrease in the life expectancy value in case of developed countries whereas in case of Developed countries the life expectancy value is gradually rising which means that developing countries are taking measures for setting up vaccine of hepatitis B. In case of Measles, according to the graph the developed countries seems to have vaccines available to tackle measles whereas developing countries life expectancy values is decreasing day by day maybe because of lack of resources to handle measles. Alcoholism is a big issue in developed countries where people have a good amount of money to spend and this shows how careless people are in terms of their health when it comes to alcoholism. We are guessing that this is due to the fact that only wealthier countries can afford alcohol or the consumption of alcohol is more prevalent among wealthier populations.That is why developing countries and alcohol have positive relation and developed countries and alcohol have negative relation. Developed countries seems to have successfully eradicated polio

diseases because of vaccines whereas in developing countries there was low expectancy initially but now it is gradually increasing maybe because of proper doses being given.

6. Conclusion

The Developed countries should help developing countries in eradicating the diseases which are affecting the life of the people by providing vaccinations. The government should focus more on the schooling of the kids which will become the face of the country in future and provide them with good food and educate them properly. The governments of developing countries should launch various schemes to motivate people to send their kids to schools. Government should organize free health care camps to provide free vaccinations for the needy and poor people so that they don't have to spend their precious money and they also stay healthy to treat their families well. The government should increase the subsidy on liquor and increase healthcare and welfare camps to generate awareness among people, how band overdrinking is and how it affects your body. WHO should with the help of developed nations should help the government of developing countries in providing free food and education and organize healthcare camps.

Github link:

https://github.com/KHUSHIKASHYAP-WEB/DV_PROJECT.git

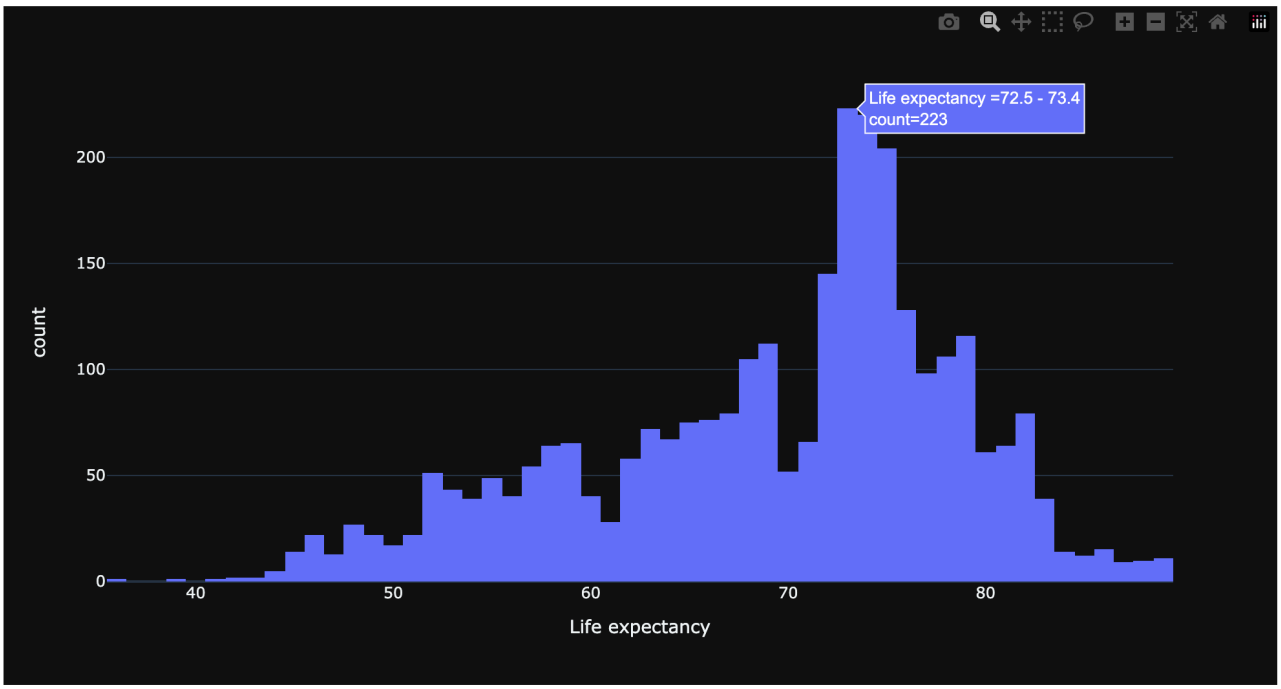
7. References:

- https://www.bmj.com/content/bmj/362/bmj.k2562.full.pdf
- https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31694-5/fulltext
- https://www.researchgate.net/publication/23545170_The_Determinants_of_Life_Expectancy_An_Analysis_of_the_OECD_Health_Data
- Dataset link: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

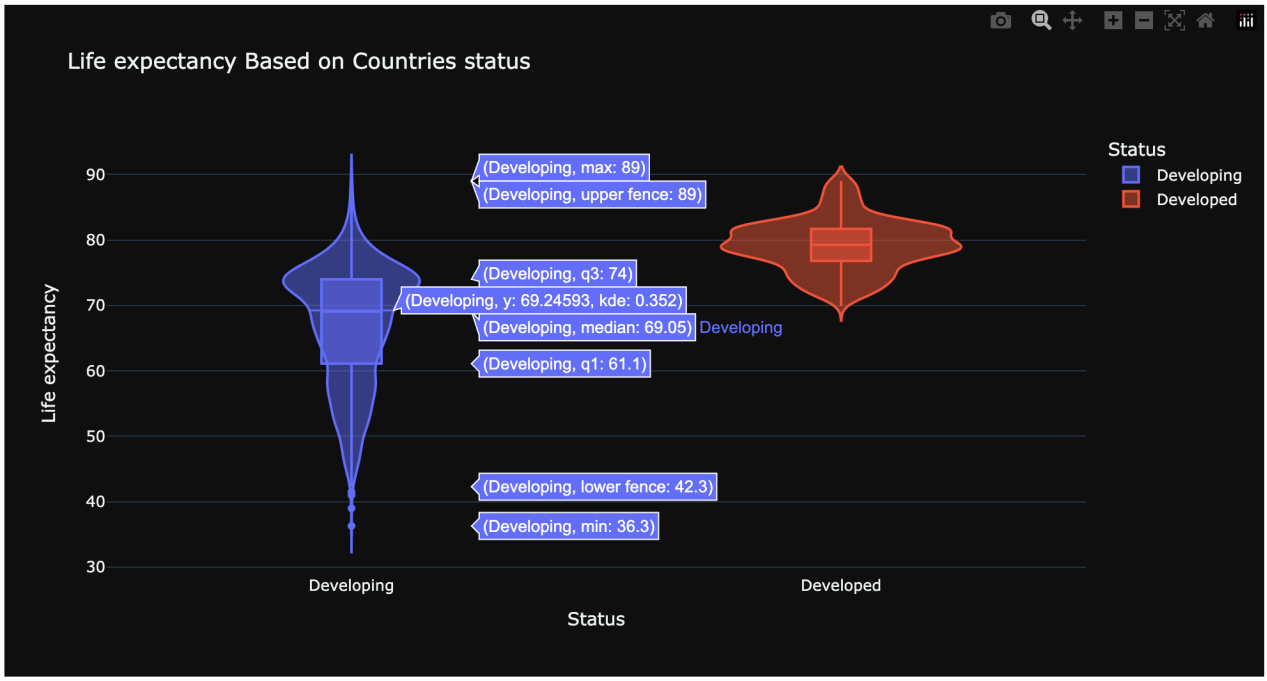
8. Appendix

DATA VISUALIZATION using Python Code:

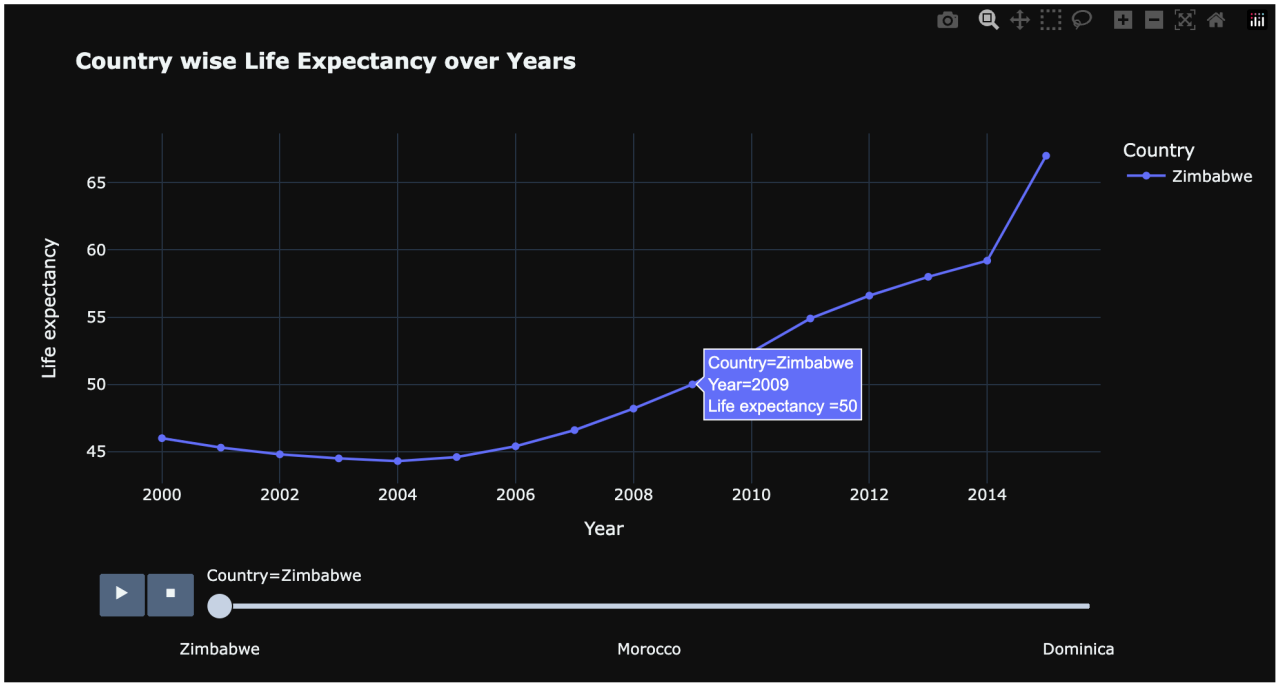
```
fig=px.histogram(df,x='Life expectancy ',template='plotly_dark')
fig.show()
```



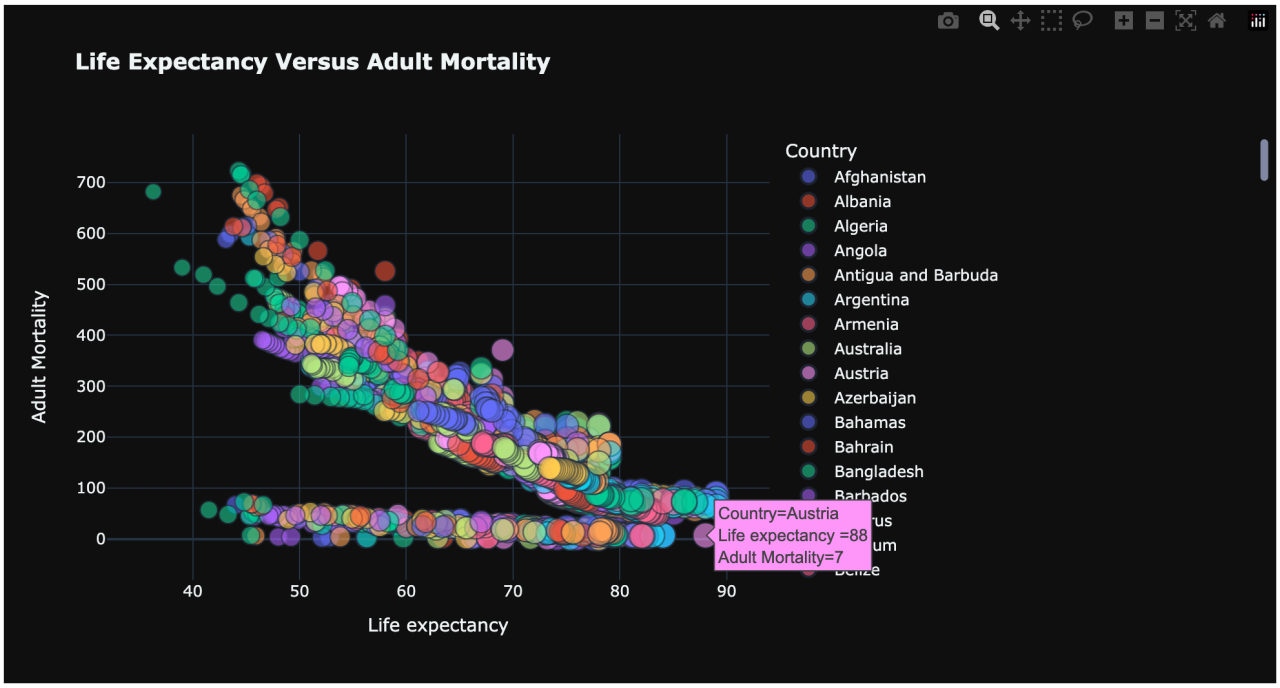

```
fig=px.violin(df,x='Status',y='Life expectancy ',color='Status',template='plotly_dark',box=True,title='Life expectancy Based on Countries status')
fig.show()
```



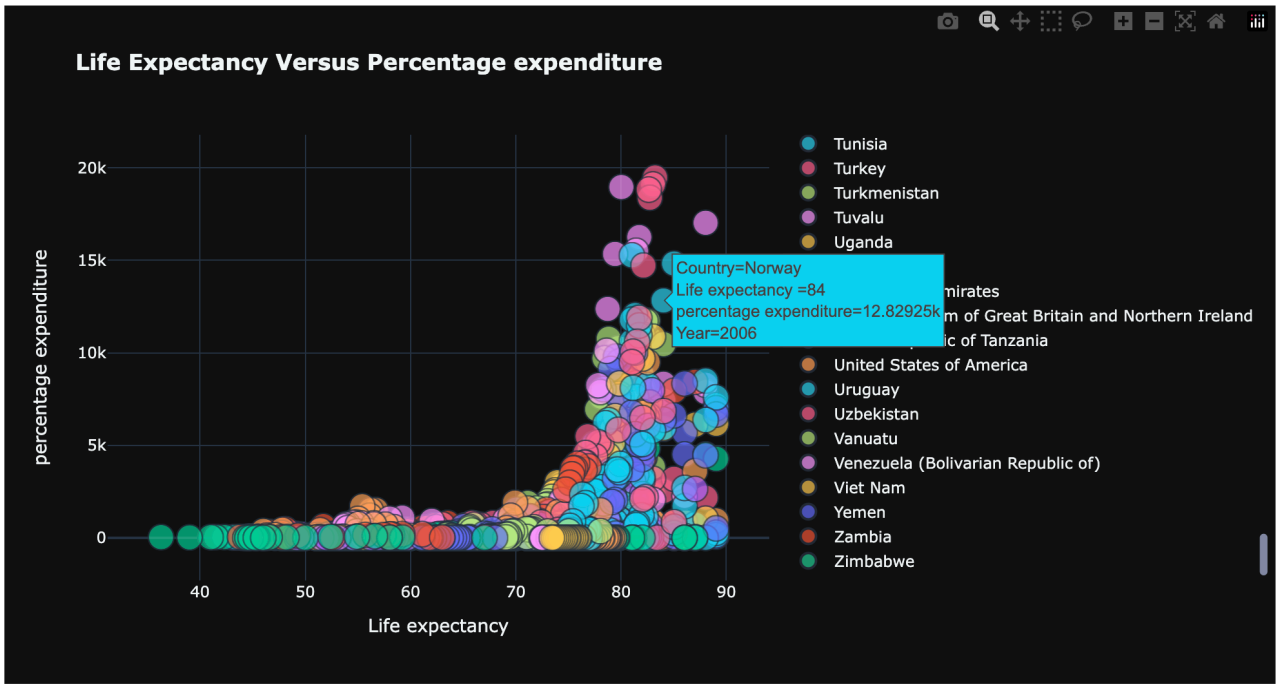
```
fig=px.line(df.sort_values(by='Year'),x='Year',y='Life expectancy
',animation_frame='Country',animation_group='Year',color='Country',markers=True,templat
e='plotly_dark',title='<b> Country wise Life Expectancy over Years')
fig.show()
```



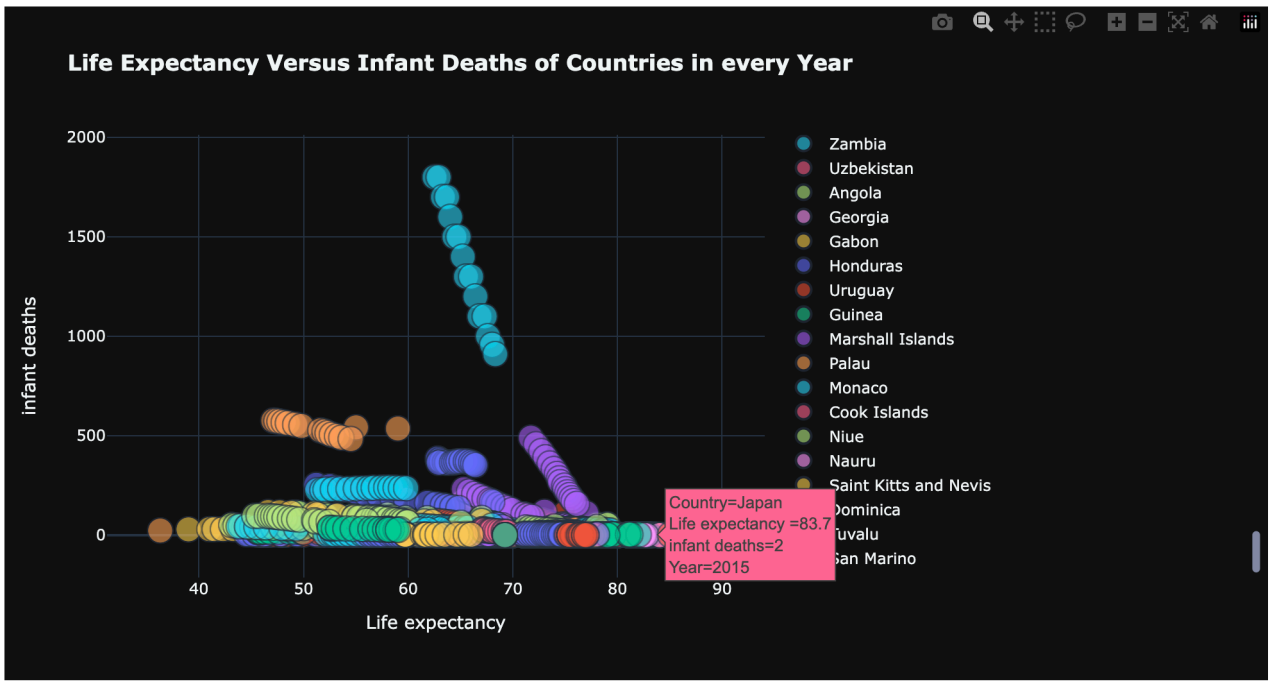

```
px.scatter(df,y='Adult Mortality',x='Life expectancy ',color='Country',size='Life expectancy
',template='plotly_dark',opacity=0.6,title='<b> Life Expectancy Versus Adult Mortality')
```



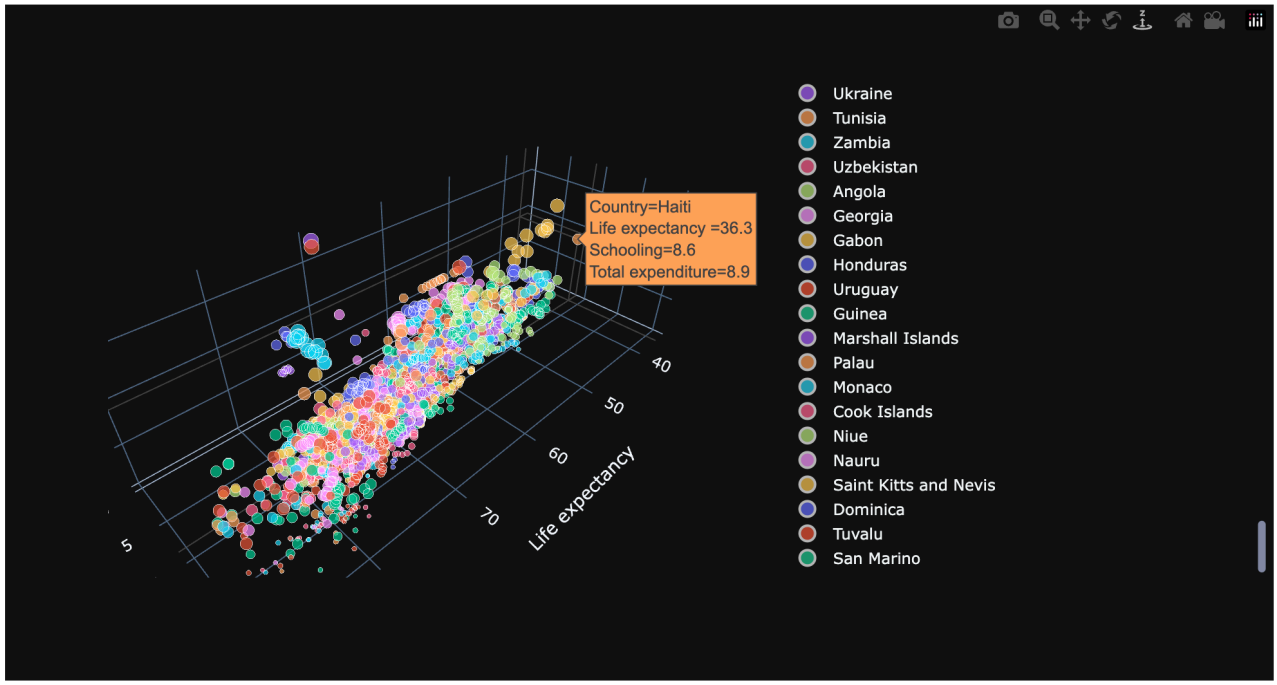
px.scatter(df,x='Life expectancy ',y='percentage expenditure',color='Country',size='Year',template='plotly_dark',title='<b> Life Expectancy Versus Percentage expenditure')



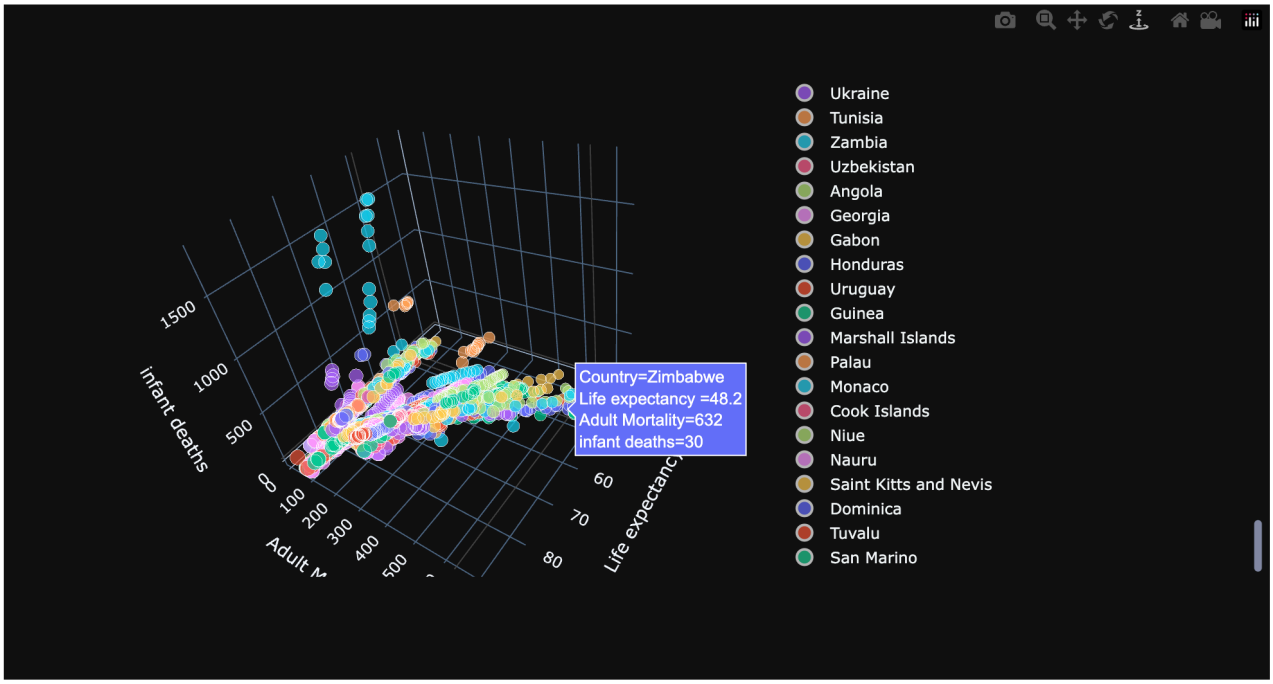
px.scatter(df.sort_values(by='Year'),y='infant deaths',x='Life expectancy ',template='plotly_dark',size='Year',color='Country',opacity=0.6,title='<b>Life Expectancy Versus Infant Deaths of Countries in every Year')



px.scatter_3d(df.sort_values(by='Year'),y='Schooling',x='Life expectancy ',z='Total expenditure',template='plotly_dark',color='Country',size='Total expenditure')



px.scatter_3d(df.sort_values(by='Year'),y='Adult Mortality',x='Life expectancy ',z='infant deaths',size='Life expectancy ',template='plotly_dark',color='Country')



Code in R:

```
library(dplyr)
library(ggplot2)
library(ggcorrplot)
library(corrplot)
library(ggpubr)
library(moments)
library(caret)
library(caTools)
library(Hmisc)
library(lattice)
library(Formula)
library(survival)
library(forecast)
library(corrplot)
library(car)
library(ROCR)
library(Metrics)
library(VIM)
library(rpart)
library(rpart.plot)
library(rattle)
library(FNN)
#install.packages("car")
library(car)



##############################################
#1. Loading Data

#"C:\Users\HP\Downloads\Life Expectancy Data.csv"
data <- read.csv("Life Expectancy Data.csv")
head(data)
sprintf("Dataset size: [%s]", toString(dim(data)))

#2. Clean and filter data.
#2.1 Remove unnecessary variables
```

```
#data <- subset(data, select = -c())

#2.2 Missing data
missing.rows = dim(data)[1] -  dim(na.omit(data))[1]
sprintf("Dataset size: [%s]", toString(dim(data)))
sprintf("Missing rows: %s (%s%%)", missing.rows, round((missing.rows*100)/dim(data)
[1], 2))

missings_df <- data.frame(type=c("missing", "non-missing") ,count = c(missing.rows,
dim(na.omit(data))[1]))



ggplot(missings_df, aes(fill=type, y="", x=count)) +
  geom_bar(position="stack", stat="identity")+
  ggtitle("Missing vs Non-missing row counts") +
  xlab("Missing count") + ylab("") +
  theme(text = element_text(size = 18))+
  scale_fill_brewer(palette="Set1")



missing_counts <- data.frame(feature = factor(names(data)),
                             counts=sapply(data, function(x) sum(is.na(x))))



ggplot(missing_counts,
       aes(x=reorder(feature, -counts), y=counts, fill=counts)) +
  geom_bar(stat="identity") +
  ggtitle("Missing counts in each feature") +
  xlab("Feature") + ylab("Missing count") +
  theme(axis.text.x=element_text(angle=20, hjust=1))+
  theme(text = element_text(size = 18))+
  scale_fill_continuous(trans = 'reverse')
#As we have more than 40% of missing data, we can apply data imputation as following:
#Checking outliers in each variable that contains missings using boxplots
#The variables with high outliers will apply imputation with median
#The variables with low outliers will apply imputation with mean.
```

```
par(mfrow=c(2,7))
boxplot(data$Life.expectancy,
        ylab = "Life Expectancy",
        main = "Boxplot of Life Expectancy",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$Adult.Mortality,
        ylab = "Adult Mortality",
        main = "Boxplot of Adult Mortality",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$Alcohol,
        ylab = "Alcohol",
        main = "Boxplot of Alcohol",
        col= "#008080",
        outcol="#008080")
boxplot(data$Hepatitis.B,
        ylab = "Hepatitis B",
        main = "Boxplot of Hepatitis B",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$BMI,
        ylab = "BMI",
        main = "Boxplot of BMI",
        col= "#008080",
        outcol="#008080")
boxplot(data$Polio,
        ylab = "Polio",
        main = "Boxplot of Polio",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$Total.expenditure,
        ylab = "Total Expenditure",
        main = "Boxplot of Total Expenditure",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$Diphtheria,
        ylab = "Diphteria",
```

```
        main = "Boxplot of Diphteria",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$GDP,
        ylab = "GDP",
        main = "Boxplot of GDP",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$Population,
        ylab = "Population",
        main = "Boxplot of Population",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$thinness..1.19.years,
        ylab = "Thinness 1-19 years",
        main = "Boxplot of Thinness for 1-19 years old",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$thinness.5.9.years,
        ylab = "Thinness 5-9 years",
        main = "Boxplot of Thinness for 5-9 years old",
        col= "#FF6666",
        outcol="#FF6666")
boxplot(data$Income.composition.of.resources,
        ylab = "Income Composition",
        main = "Boxplot of Income Composition",
        col= "#008080",
        outcol="#008080")
boxplot(data$Schooling,
        ylab = "Schooling",
        main = "Boxplot of Schooling",
        col= "#FF6666",
        outcol="#FF6666")


mydata=kNN(data,variable=c("Alcohol","Hepatitis.B","Polio","Total.expenditure","Diphtheria","Life.expectancy","Adult.Mortality","thinness..1.19.years","Income.composition.of.resources","Schooling","thinness.5.9.years","BMI","Population","GDP"),k=9)
```

```
#Removing the unwanted variables created during imputation
mydata1=subset(mydata,select=Country:Schooling)
#Creation of dummy variable
mydata1$Developed=ifelse(mydata$Status=="Developed",1,0)
attach(mydata)


#check for outliers
boxplot(Adult.Mortality~Status)
boxplot(mydata1$infant.deaths~Status)
boxplot(mydata1$Alcohol~Status)
boxplot(mydata1$percentage.expenditure~Status)
boxplot(mydata1$Hepatitis.B~Status)
boxplot(mydata1$Measles~Status)
boxplot(mydata1$BMI~Status)#No outliers
boxplot(mydata1$under.five.deaths~Status)
boxplot(mydata1$Polio~Status)
boxplot(mydata1$Total.expenditure~Status)
boxplot(mydata1$Diphtheria~Status)
boxplot(mydata1$HIV.AIDS~Status)
boxplot(mydata1$GDP~Status)
boxplot(mydata1$Population~Status)
boxplot(mydata1$thinness..1.19.years~Status)
boxplot(mydata1$thinness.5.9.years~Status)
boxplot(mydata$Income.composition.of.resources~Status)
boxplot(mydata1$Schooling~Status)

#Correlation check
corrplot(cor(mydata1[,-c(1:3)]),type="upper",method="circle",title="Correlation plot
between variables",mar=c(0.1,0.1,0.1,0.1))

#Adult.Mortality, thinness 1-19 years, thinness 5-9 years , HIV.AIDS are negatively
correlated with Life expectancy.And there is positive correlation of
Income.composition.of.resources,Schooling, percentage expenditure, Polio,
Diphtheria ,developed and BMI with Life expectancy.
#The variable infant death and under five death are highly correlated
#There is high correlation between thinness..1.19.years and thinness.5.9.years
#GDP and percent expenditure are highly correlated.
```

#There is multicollinearity present because there is correlation between independent variables
#Insights from EDA:-

#Life expectancy increased across years.
#In developed countries adult mortality rate,prevelance of thinness of child between 1 to 19 years,infants deaths and deaths from HIV.AIDS are less comparing to developing countries which ultimately leads higher life expectancy in developed countries.
#On the other hand the other variables such as Income composition of resources ,number of years of schooling,percentage expenditure and Total expenditure on health are more in developed countries than developing countries which again increase the life expectancy in developed countries.
#Also by the increased numbers of immunization coverage against Hepatitis B, Polio and Diphtheria in developed countries results to high life expectancy in developed countries in contrast to developing countries.
#Life expectancy is high in developed countries comparing to developing countries.

#Stepwise removal of variables

```
vif(lm(Life.expectancy~.-Country-Year-Status-Life.expectancy,data=mydata1))
vif(lm(Life.expectancy~.-Country-Year-Status-Life.expectancy- infant.deaths - thinness.5.9.years-GDP,data=mydata1))
```

#The variables such as infant.deaths,GDP and thinness..5.9.years with high VIF(>5) are removed from the final model.
#final data after removing the unwanted variable

```
finaldata=mydata1[,-c(3,6,17,20)]
#Data Partition
train=subset(finaldata, Year <= 2013)
test=subset(finaldata, Year >= 2014)
dim(train)
dim(test)
#Scaling the data
scale = preProcess(train[,-c(1,2)], method = "range")
scaled.train = predict(scale, train[,-c(1,2)])
```

```
scale = preProcess(test[,-c(1,2)], method = "range")
scaled.test = predict(scale, test[,-c(1,2)])
#Data is split into train and test data based on the variable year train data -2000 to 2013 test data - 2014 to 2015
#Data scaling is done for all the numeric variables


###Linear regression model
train.LR=scaled.train
test.LR=scaled.test
#Base model with all variables
model1=lm(Life.expectancy~.,data=train.LR)
summary(model1)
#final model
model2=lm(Life.expectancy~.-Population-Hepatitis.B-Alcohol-Total.expenditure-under.five.deaths,data=train.LR)
summary(model2)
#Alcohol,Total.expenditure,Hepatitis.B,Population,thinness..1.19.years are not significant,so removed from the final model.Top 5 variable effecting Life expectancy are Adult.Mortality,BMI,HIV.AIDS,Schooling and Income.composition.of.resources.
#Prediction using train data
train.pred=train.LR
train.pred$predTrain=predict(model2, newdata = train.pred,type = "response")
#Mean Absolute ERROR
mae(train.pred$Life.expectancy,train.pred$predTrain)

#ROOT MEAN SQUARE ERROR
rmse(train.pred$Life.expectancy,train.pred$predTrain)


#Prediction using test data
test.pred=test.LR
test.pred$predTest=predict(model2, newdata = test.pred,type = "response")

#Mean Absolute ERROR
mae(test.pred$Life.expectancy,test.pred$predTest)
```

```
#ROOT MEAN SQUARE ERROR
rmse(test.pred$Life.expectancy,test.pred$predTest)

#Linear regression

#The mean absolute error and root mean square error for train and test data is 0.058 and
0.078.
#The mean absolute error and root mean square error for test data is 0.064 and 0.087
###Regression tree
train.cart=scaled.train
test.cart=scaled.test
#First tree
m1 <- rpart(
  formula = Life.expectancy ~ .,
  data    = train.cart,
  method  = "anova",
  control = list(cp = 0, xval = 10,minsplit=60, minbucket = 20)
)
plotcp(m1)
abline(v = 15, lty = "dashed")
#Final tree
m2 <- rpart(formula = Life.expectancy ~ .,data=train.cart,method=
"anova",cp=0.003,minsplit=60, minbucket = 20)
m2
rpart.rules(m2,style = "tallw", cover = FALSE, nn = FALSE,
           roundint = TRUE, clip.facs = FALSE,
           varorder = NULL)
#plotting the decision tree
fancyRpartPlot(m2)
##Identify the importance of the variables
m2$variable.importance
#Prediction using train data
train.pred.cart=train.cart
train.pred.cart$predTrain.cart=predict(m2, newdata = train.pred.cart)
#Mean Absolute ERROR
mae(train.pred.cart$Life.expectancy,train.pred.cart$predTrain.cart)
#ROOT MEAN SQUARE ERROR
rmse(train.pred.cart$Life.expectancy,train.pred.cart$predTrain.cart)
```

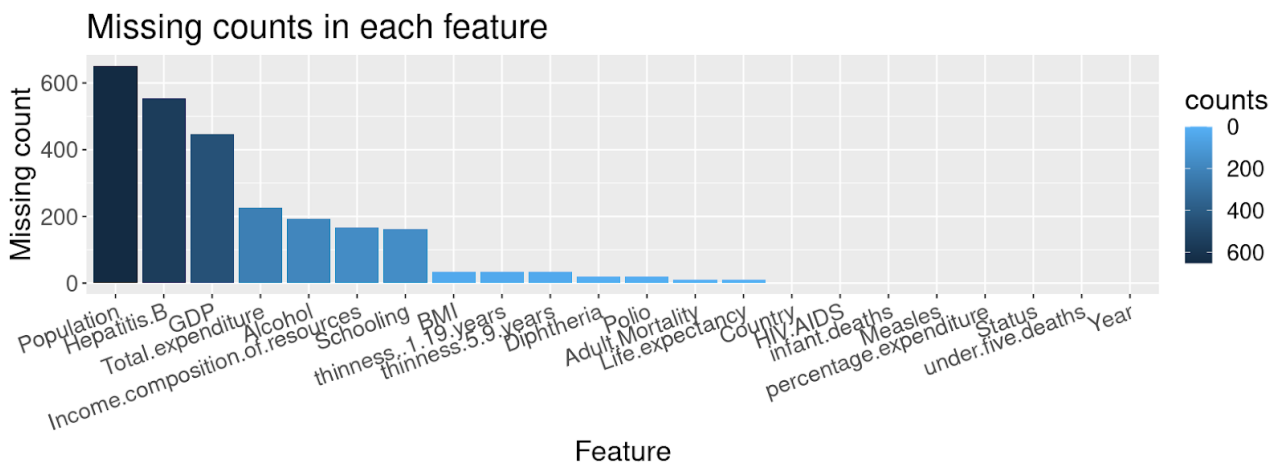
```
#Prediction using test data
test.pred.cart=test.cart
test.pred.cart$predTest.cart=predict(m2, newdata = test.pred.cart)

#Mean Absolute ERROR
mae(test.pred.cart$Life.expectancy,test.pred.cart$predTest.cart)

#ROOT MEAN SQUARE ERROR
rmse(test.pred.cart$Life.expectancy,test.pred.cart$predTest.cart)

#Regression tree

#The mean absolute error and root mean square error for train and test data is 0.043 and
0.057.
#The mean absolute error and root mean square error for test data is 0.055 and 0.076

##KNN Regression
train.knn=scaled.train
test.knn=scaled.test

trControl=trainControl(method='repeatedcv',number=10,repeats=3)
set.seed(100)
fit=train(Life.expectancy~.,data=train.knn,tuneGrid=expand.grid(k=1:20),method='knn',trC
ontrol=trControl)
varImp(fit)
#prediction using KNN
pred=predict(fit,newdata=test.knn)
#Mean Absolute ERROR
mae(test.knn$Life.expectancy,pred)
#ROOT MEAN SQUARE ERROR
rmse(test.knn$Life.expectancy,pred)
#KNN regression

#The mean absolute error and root mean square error for test data is 0.061 and 0.076
#Based on performance measures among all models regression Tree shows least Mean
absolute error and Root mean square error.So Regression tree is considered as best model to
predict Life expectancy.
```

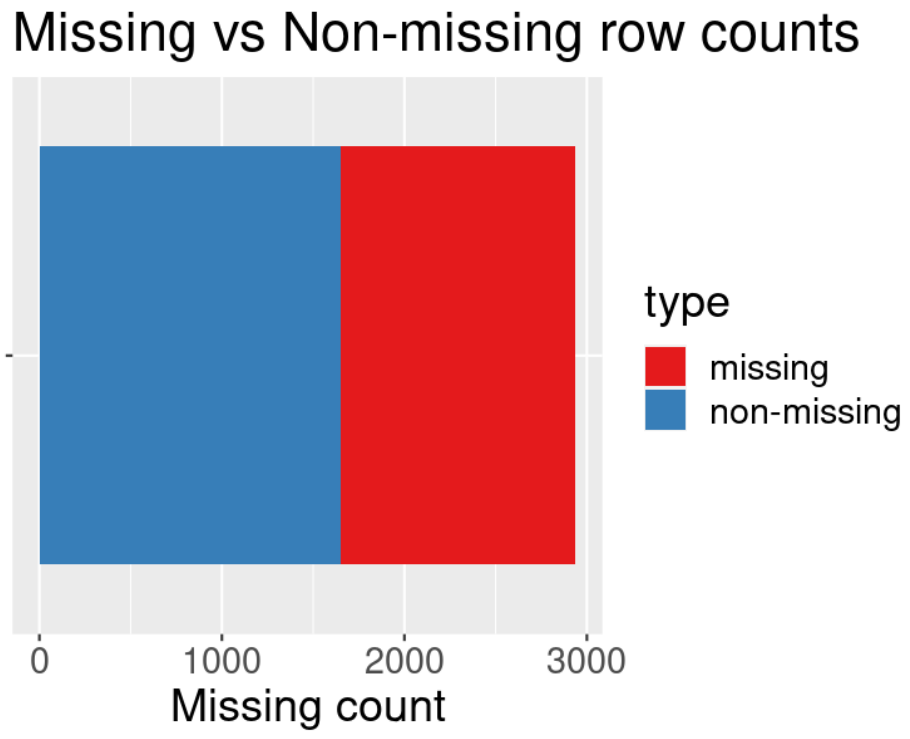
OUTPUT:

```
> library(dplyr)
> library(ggplot2)
> library(ggcorrplot)
> library(corrplot)
> library(ggpubr)
> library(moments)
> library(caret)
> library(caTools)
> library(Hmisc)
> library(lattice)
> library(Formula)
> library(forecast)
> library(corrplot)
> library(car)
> library(ROCR)
> library(Metrics)
> library(VIM)
> library(rpart)
> library(rpart.plot)
> library(rattle)
> library(FNN)
> #install.packages("car")
> library(car)
> #"C:\Users\HP\Downloads\Life Expectancy Data.csv"
> data <- read.csv("Life Expectancy Data.csv")
> head(data)
```

```
> head(data)
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol
1 Afghanistan 2015 Developing            65.0             263            62    0.01
2 Afghanistan 2014 Developing            59.9             271            64    0.01
3 Afghanistan 2013 Developing            59.9             268            66    0.01
4 Afghanistan 2012 Developing            59.5             272            69    0.01
5 Afghanistan 2011 Developing            59.2             275            71    0.01
6 Afghanistan 2010 Developing            58.8             279            74    0.01
  percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio Total.expenditure Diphtheria
1              71.279624          65    1154 19.1                83     6              8.16         65
2              73.523582          62     492 18.6                86    58              8.18         62
3              73.219243          64     430 18.1                89    62              8.13         64
4              78.184215          67    2787 17.6                93    67              8.52         67
5               7.097109          68    3013 17.2                97    68              7.87         68
6              79.679367          66    1989 16.7               102    66              9.20         66
  HIV.AIDS       GDP Population thinness..1.19.years thinness.5.9.years Income.composition.of.resources
1      0.1 584.25921   33736494                17.2              17.3                            0.479
2      0.1 612.69651     327582                17.5              17.5                            0.476
3      0.1 631.74498   31731688                17.7              17.7                            0.470
4      0.1 669.95900    3696958                17.9              18.0                            0.463
5      0.1  63.53723    2978599                18.2              18.2                            0.454
6      0.1 553.32894    2883167                18.4              18.4                            0.448
  Schooling
1      10.1
2      10.0
3       9.9
```

```
[1] "Dataset size: [2938, 22]"
> #2.2 Missing data
> missing.rows = dim(data)[1] -  dim(na.omit(data))[1]
> sprintf("Dataset size: [%s]", toString(dim(data)))
[1] "Dataset size: [2938, 22]"
> sprintf("Missing rows: %s (%s%%)", missing.rows, round((missing.rows*100)/dim(data)[1], 2))
[1] "Missing rows: 1289 (43.87%)"
> missings_df <- data.frame(type=c("missing", "non-missing") ,count = c(missing.rows,  dim(na.omit
(data))[1]))
> ggplot(missings_df, aes(fill=type, y="", x=count)) +
+   geom_bar(position="stack", stat="identity")+
+   ggtitle("Missing vs Non-missing row counts") +
+   xlab("Missing count") + ylab("") +
+   theme(text = element_text(size = 18))+
+   scale_fill_brewer(palette="Set1")
> missing_counts <- data.frame(feature = factor(names(data)),
+                              counts=sapply(data, function(x) sum(is.na(x))))
> ggplot(missing_counts,
+        aes(x=reorder(feature, -counts), y=counts, fill=counts)) +
+   geom_bar(stat="identity") +
+   ggtitle("Missing counts in each feature") +
+   xlab("Feature") + ylab("Missing count") +
+   theme(axis.text.x=element_text(angle=20, hjust=1))+
+   theme(text = element_text(size = 18))+
+   scale_fill_continuous(trans = 'reverse')
```

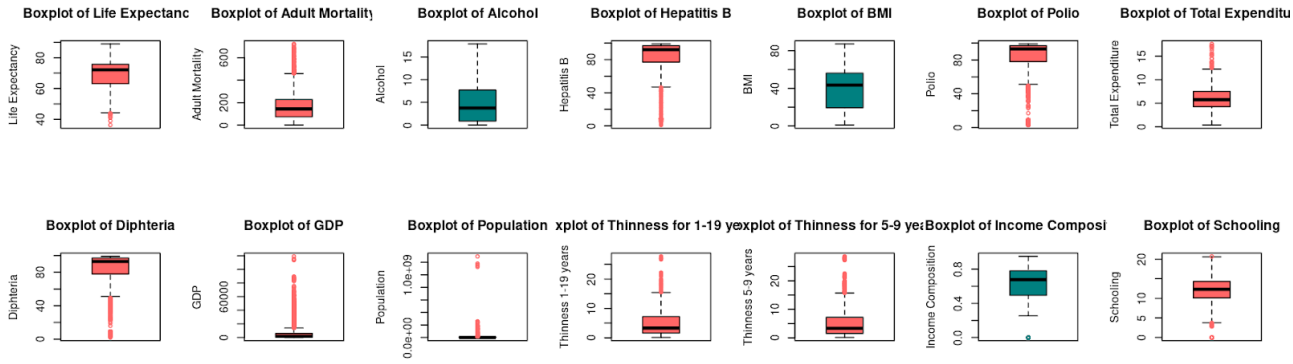




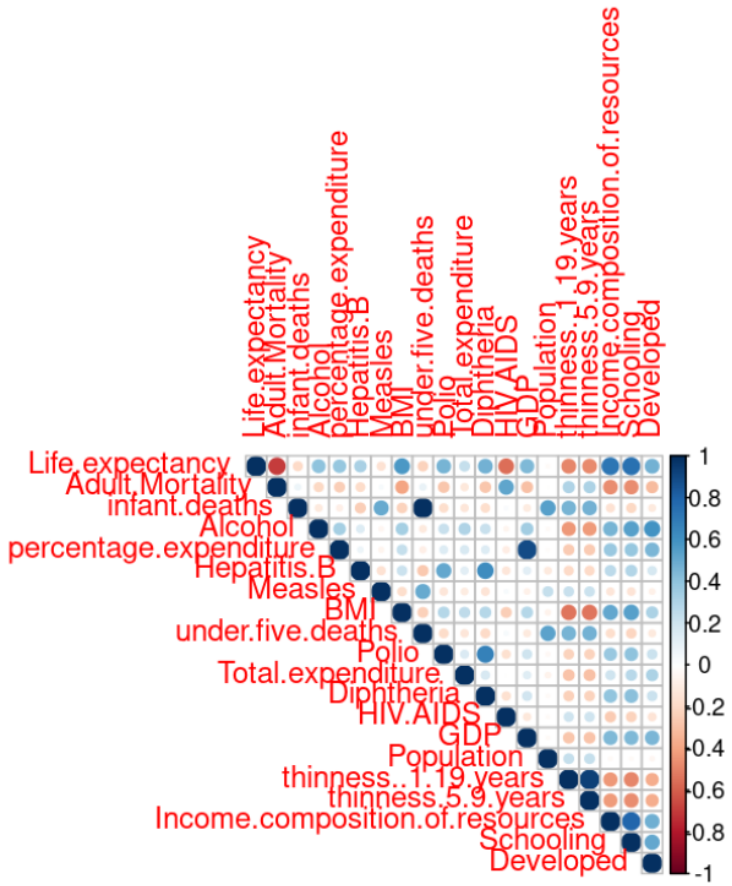
```
> par(mfrow=c(2,7))
> boxplot(data$Life.expectancy,
+         ylab = "Life Expectancy",
+         main = "Boxplot of Life Expectancy",
+         col= "#FF6666",
+         outcol="#FF6666")
> boxplot(data$Adult.Mortality,
+         ylab = "Adult Mortality",
+         main = "Boxplot of Adult Mortality",
+         col= "#FF6666",
+         outcol="#FF6666")
> boxplot(data$Alcohol,
+         ylab = "Alcohol",
+         main = "Boxplot of Alcohol",
+         col= "#008080",
+         outcol="#008080")
> boxplot(data$Hepatitis.B,
+         ylab = "Hepatitis B",
+         main = "Boxplot of Hepatitis B",
+         col= "#FF6666",
+         outcol="#FF6666")
> boxplot(data$BMI,
+         ylab = "BMI",
+         main = "Boxplot of BMI",
+         col= "#008080",
+         outcol="#008080")
```

```
> boxplot(data$Polio,
+          ylab = "Polio",
+          main = "Boxplot of Polio",
+          col= "#FF6666",
+          outcol="#FF6666")
> boxplot(data$Total.expenditure,
+          ylab = "Total Expenditure",
+          main = "Boxplot of Total Expenditure",
+          col= "#FF6666",
+          outcol="#FF6666")
> boxplot(data$Diphtheria,
+          ylab = "Diphteria",
+          main = "Boxplot of Diphteria",
+          col= "#FF6666",
+          outcol="#FF6666")
> boxplot(data$GDP,
+          ylab = "GDP",
+          main = "Boxplot of GDP",
+          col= "#FF6666",
+          outcol="#FF6666")
> boxplot(data$Population,
+          ylab = "Population",
+          main = "Boxplot of Population",
+          col= "#FF6666",
+          outcol="#FF6666")
```

```
> boxplot(data$thinness..1.19.years,
+         ylab = "Thinness 1-19 years",
+         main = "Boxplot of Thinness for 1-19 years old",
+         col= "#FF6666",
+         outcol="#FF6666")
> boxplot(data$thinness.5.9.years,
+         ylab = "Thinness 5-9 years",
+         main = "Boxplot of Thinness for 5-9 years old",
+         col= "#FF6666",
+         outcol="#FF6666")
> boxplot(data$Income.composition.of.resources,
+         ylab = "Income Composition",
+         main = "Boxplot of Income Composition",
+         col= "#008080",
+         outcol="#008080")
> boxplot(data$Schooling,
+         ylab = "Schooling",
+         main = "Boxplot of Schooling",
+         col= "#FF6666",
+         outcol="#FF6666")
```



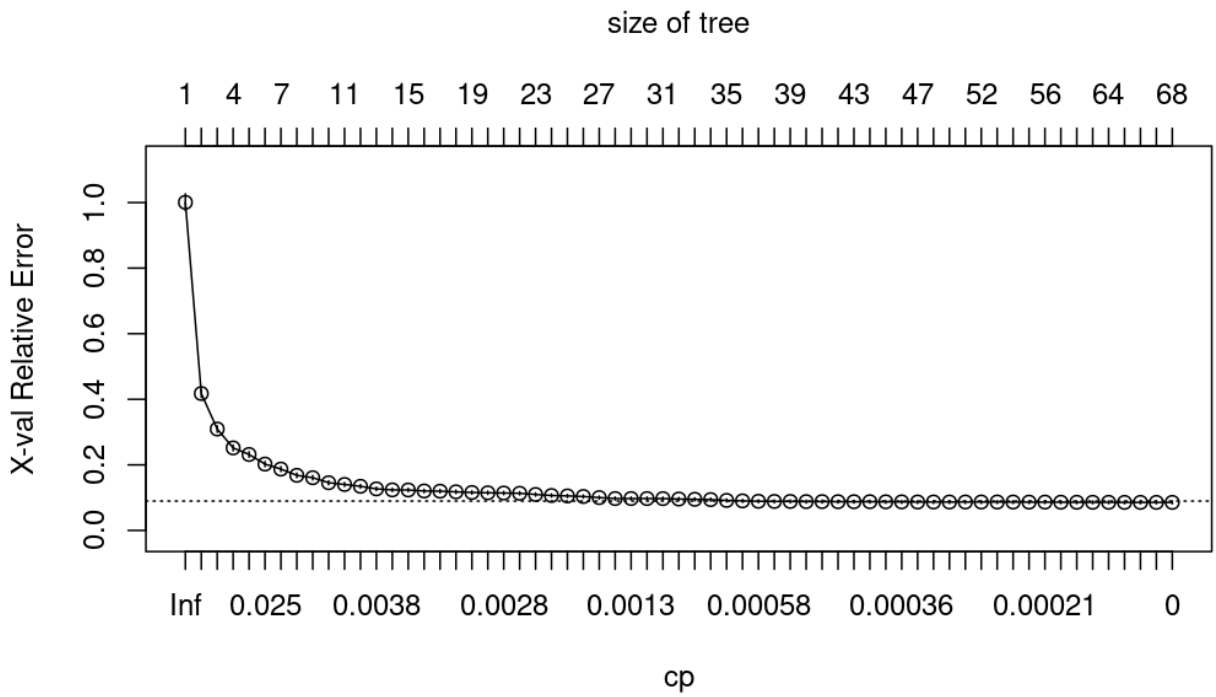

```
> #Correlation check
> corrplot(cor(mydata1[,-c(1:3)]),type="upper",method="circle",title="Correlation plot between variable
s",mar=c(0.1,0.1,0.1,0.1))
> vif(lm(Life.expectancy~.-Country-Year-Status-Life.expectancy,data=mydata1))
             Adult.Mortality                 infant.deaths                       Alcohol
                    1.771714                    177.497629                      1.970904
      percentage.expenditure                 Hepatitis.B                       Measles
                    5.422255                      1.763138                      1.377223
                         BMI               under.five.deaths                         Polio
                    1.783529                    176.681353                      1.969899
       Total.expenditure                    Diphtheria                       HIV.AIDS
                    1.220382                      2.406220                      1.421889
                         GDP                    Population       thinness..1.19.years
                    5.743396                      1.473127                      8.496430
      thinness.5.9.years Income.composition.of.resources                     Schooling
                    8.563711                      3.058925                      3.515479
                   Developed
                    1.940332
```



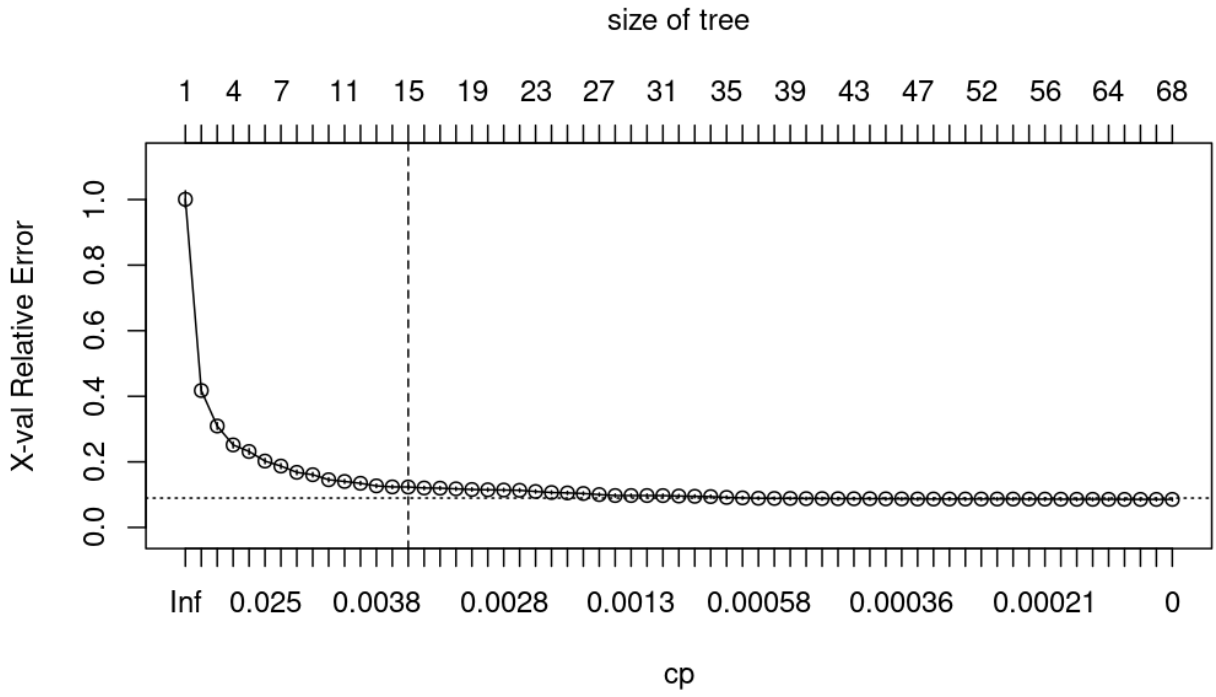
```
> vif(lm(Life.expectancy~.-Country-Year-Status-Life.expectancy- infant.deaths - thinness.5.9.years-GDP,
data=mydata1))
                Adult.Mortality                         Alcohol          percentage.expenditure
                       1.764326                        1.930660                        1.328458
                    Hepatitis.B                         Measles                             BMI
                       1.756141                        1.373377                        1.751227
              under.five.deaths                           Polio             Total.expenditure
                       2.173681                        1.961449                        1.197853
                     Diphtheria                         HIV.AIDS                      Population
                       2.377023                        1.415420                        1.426266
          thinness..1.19.years Income.composition.of.resources                      Schooling
                       1.990257                        2.982343                        3.497921
                      Developed
                       1.936175
```

```
> #The variables such as infant.deaths,GDP and thinness..5.9.years with high VIF(>5) are removed from t
he final model.
> #final data after removing the unwanted variable
> finaldata=mydata1[,-c(3,6,17,20)]
> #Data Partition
> train=subset(finaldata, Year <= 2013)
> test=subset(finaldata, Year >= 2014)
> dim(train)
[1] 2572   19
> dim(test)
[1] 366  19
> #Scaling the data
> scale = preProcess(train[,-c(1,2)], method = "range")
> scaled.train = predict(scale, train[,-c(1,2)])
> scale = preProcess(test[,-c(1,2)], method = "range")
> scaled.test = predict(scale, test[,-c(1,2)])
> ###Linear regression model
> train.LR=scaled.train
> test.LR=scaled.test
> #Base model with all variables
> model1=lm(Life.expectancy~.,data=train.LR)
> summary(model1)

Call:
lm(formula = Life.expectancy ~ ., data = train.LR)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41545 -0.04391 -0.00269  0.04492  0.34911

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.312147   0.011004  28.367  < 2e-16 ***
Adult.Mortality                 -0.248819   0.011645 -21.367  < 2e-16 ***
Alcohol                         -0.007364   0.009647  -0.763  0.44535
percentage.expenditure           0.110314   0.017283   6.383 2.06e-10 ***
Hepatitis.B                     -0.020807   0.007991  -2.604  0.00927 **
Measles                         -0.087849   0.032213  -2.727  0.00643 **
BMI                              0.075730   0.008929   8.481  < 2e-16 ***
under.five.deaths               -0.083996   0.034518  -2.433  0.01503 *
Polio                            0.055692   0.008821   6.314 3.20e-10 ***
Total.expenditure                0.013591   0.011727   1.159  0.24660
Diphtheria                       0.079806   0.009385   8.503  < 2e-16 ***
HIV.AIDS                        -0.468290   0.017309 -27.054  < 2e-16 ***
Population                       0.073580   0.042640   1.726  0.08454 .
thinness..1.19.years            -0.025233   0.013469  -1.873  0.06112 .
Income.composition.of.resources  0.130919   0.011460  11.424  < 2e-16 ***
Schooling                        0.283343   0.017608  16.092  < 2e-16 ***
Developed                        0.023059   0.005733   4.022 5.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.07828 on 2555 degrees of freedom
Multiple R-squared:  0.8181,    Adjusted R-squared:  0.817
F-statistic: 718.4 on 16 and 2555 DF,  p-value: < 2.2e-16

> #final model
> model2=lm(Life.expectancy~.-Population-Hepatitis.B-Alcohol-Total.expenditure-under.five.deaths,data=t
rain.LR)
> summary(model2)

Call:
lm(formula = Life.expectancy ~ . - Population - Hepatitis.B -
    Alcohol - Total.expenditure - under.five.deaths, data = train.LR)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41079 -0.04351 -0.00307  0.04408  0.34359

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.309049   0.010284  30.051  < 2e-16 ***
Adult.Mortality                -0.250429   0.011600 -21.589  < 2e-16 ***
percentage.expenditure          0.113854   0.017270   6.593 5.23e-11 ***
Measles                        -0.115571   0.028686  -4.029 5.77e-05 ***
BMI                             0.076363   0.008896   8.584  < 2e-16 ***
Polio                           0.052859   0.008710   6.069 1.48e-09 ***
Diphtheria                      0.071957   0.008656   8.313  < 2e-16 ***
Polio                           0.052859   0.008710   6.069 1.48e-09 ***
Diphtheria                      0.071957   0.008656   8.313  < 2e-16 ***
HIV.AIDS                       -0.465149   0.017188 -27.063  < 2e-16 ***
thinness..1.19.years           -0.032970   0.012017  -2.744  0.00612 **
Income.composition.of.resources  0.128526   0.011419  11.256  < 2e-16 ***
Schooling                       0.283885   0.017080  16.621  < 2e-16 ***
Developed                       0.020686   0.005154   4.013 6.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07842 on 2560 degrees of freedom
Multiple R-squared:  0.8172,    Adjusted R-squared:  0.8164
F-statistic:  1040 on 11 and 2560 DF,  p-value: < 2.2e-16

> #Alcohol,Total.expenditure,Hepatitis.B,Population,thinness..1.19.years are not significant,so removed
from the final model.Top 5 variable effecting Life expectancy are Adult.Mortality,BMI,HIV.AIDS,Schoolin
g and Income.composition.of.resources.
> #Prediction using train data
> train.pred=train.LR
> train.pred$predTrain=predict(model2, newdata = train.pred,type = "response")
> #Mean Absolute ERROR
> mae(train.pred$Life.expectancy,train.pred$predTrain)
[1] 0.05838886
> #ROOT MEAN SQUARE ERROR
> rmse(train.pred$Life.expectancy,train.pred$predTrain)
[1] 0.07823295
> #Prediction using test data
> test.pred=test.LR
> test.pred$predTest=predict(model2, newdata = test.pred,type = "response")
> #Mean Absolute ERROR
> mae(test.pred$Life.expectancy,test.pred$predTest)
[1] 0.0639056
> #ROOT MEAN SQUARE ERROR
> rmse(test.pred$Life.expectancy,test.pred$predTest)
[1] 0.08722445
> #The mean absolute error and root mean square error for train and test data is 0.058 and 0.078.
> #The mean absolute error and root mean square error for test data is 0.064 and 0.087
> ###Regression tree
> train.cart=scaled.train
> test.cart=scaled.test
> #First tree
> m1 <- rpart(
+   formula = Life.expectancy ~ .,
+   data    = train.cart,
+   method  = "anova",
+   control = list(cp = 0, xval = 10,minsplit=60, minbucket = 20)
+ )
> plotcp(m1)
```




```
> ###Regression tree
> train.cart=scaled.train
> test.cart=scaled.test
> #First tree
> m1 <- rpart(
+   formula = Life.expectancy ~ .,
+   data    = train.cart,
+   method  = "anova",
+   control = list(cp = 0, xval = 10,minsplit=60, minbucket = 20)
+ )
> plotcp(m1)
> abline(v = 15, lty = "dashed")
```



```
> ###Regression tree
> train.cart=scaled.train
> test.cart=scaled.test
> #First tree
> m1 <- rpart(
+   formula = Life.expectancy ~ .,
+   data    = train.cart,
+   method  = "anova",
+   control = list(cp = 0, xval = 10,minsplit=60, minbucket = 20)
+ )
> plotcp(m1)
> abline(v = 15, lty = "dashed")
> #Final tree
> m2 <- rpart(formula = Life.expectancy ~ .,data=train.cart,method= "anova",cp=0.003,m
insplit=60, minbucket = 20)
> m2
n= 2572

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2572 86.0915300 0.6189396
   2) HIV.AIDS>=0.01287129 677 11.9504300 0.3844257
```

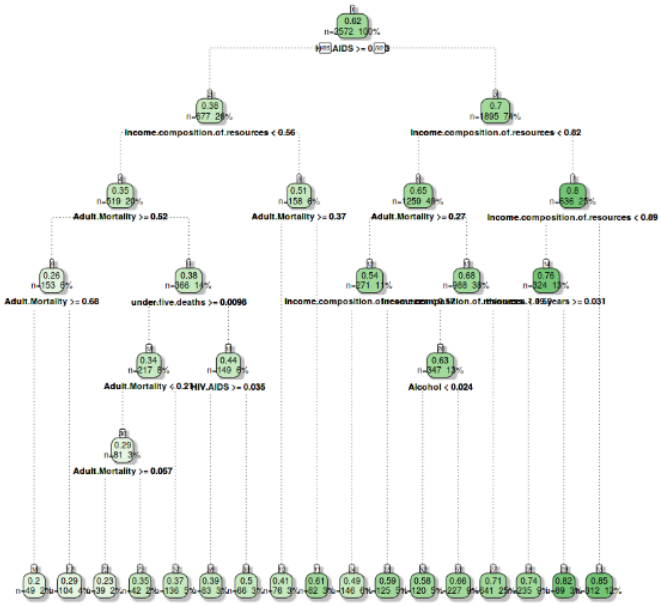
```
0, minbucket = 20)
> m2
n= 2572

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2572 86.0915300 0.6189396
   2) HIV.AIDS>=0.01287129 677 11.9504300 0.3844257
     4) Income.composition.of.resources< 0.5610403 519  5.9419450 0.3464515
       8) Adult.Mortality>=0.515928 153  0.8715913 0.2624301
        16) Adult.Mortality>=0.6814404 49  0.2153100 0.1965690 *
        17) Adult.Mortality< 0.6814404 104  0.3435923 0.2934608 *
       9) Adult.Mortality< 0.515928 366  3.5387090 0.3815753
        18) under.five.deaths>=0.0098 217  1.5957830 0.3426403
          36) Adult.Mortality< 0.2132964 81  0.6671178 0.2940708
            72) Adult.Mortality>=0.05747922 39  0.1796499 0.2314991 *
            73) Adult.Mortality< 0.05747922 42  0.1929871 0.3521731 *
          37) Adult.Mortality>=0.2132964 136  0.6237822 0.3715677 *
        19) under.five.deaths< 0.0098 149  1.1348830 0.4382792
          38) HIV.AIDS>=0.03465347 83  0.4655172 0.3867767 *
          39) HIV.AIDS< 0.03465347 66  0.1723416 0.5030476 *
     5) Income.composition.of.resources>=0.5610403 158  2.8016700 0.5091634
      10) Adult.Mortality>=0.3746537 76  0.5891223 0.4052981 *
      11) Adult.Mortality< 0.3746537 82  0.6327632 0.6054288 *
```

```
      11) Adult.Mortality< 0.3746537 82  0.6327632 0.6054288 *
   3) HIV.AIDS< 0.01287129 1895 23.6066100 0.7027211
     6) Income.composition.of.resources< 0.7446921 808  6.6337320 0.6189246
      12) Adult.Mortality>=0.2915512 214  0.8267738 0.5201812
        24) Adult.Mortality>=0.3691136 46  0.1312249 0.4470753 *
        25) Adult.Mortality< 0.3691136 168  0.3823882 0.5401983 *
      13) Adult.Mortality< 0.2915512 594  2.9686830 0.6544988
        26) Alcohol< 0.04115342 186  1.0267300 0.6108425
          52) Polio< 0.9427083 120  0.5176672 0.5819734 *
          53) Polio>=0.9427083 66  0.2272147 0.6633316 *
        27) Alcohol>=0.04115342 408  1.4258530 0.6744010 *
     7) Income.composition.of.resources>=0.7446921 1087  7.0818180 0.7650096
      14) Income.composition.of.resources< 0.876327 711  2.4234990 0.7244947
        28) Adult.Mortality>=0.248615 97  0.3323306 0.6433420 *
        29) Adult.Mortality< 0.248615 614  1.3514290 0.7373153
          58) thinness..1.19.years>=0.07427536 361  0.4692526 0.7177564 *
          59) thinness..1.19.years< 0.07427536 253  0.5470231 0.7652234 *
      15) Income.composition.of.resources>=0.876327 376  1.2843510 0.8416216
        30) Adult.Mortality>=0.1059557 139  0.3349673 0.8063014 *
        31) Adult.Mortality< 0.1059557 237  0.6742785 0.8623368 *
> rpart.rules(m2,style = "tallw", cover = FALSE, nn = FALSE,
+             roundint = TRUE, clip.facs = FALSE,
+             varorder = NULL)
Life.expectancy is 0.20 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources < 0.56
```

```
Life.expectancy is 0.20 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources < 0.56
                       Adult.Mortality >= 0.681

Life.expectancy is 0.23 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources < 0.56
                       Adult.Mortality is 0.057 to 0.213
                       under.five.deaths >= 0.0098

Life.expectancy is 0.29 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources < 0.56
                       Adult.Mortality is 0.516 to 0.681

Life.expectancy is 0.35 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources < 0.56
                       Adult.Mortality < 0.057
                       under.five.deaths >= 0.0098
```

```
Life.expectancy is 0.37 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources < 0.56
                       Adult.Mortality is 0.213 to 0.516
                       under.five.deaths >= 0.0098

Life.expectancy is 0.39 when
                       HIV.AIDS >= 0.035
                       Income.composition.of.resources < 0.56
                       Adult.Mortality < 0.516
                       under.five.deaths < 0.0098

Life.expectancy is 0.41 when
                       HIV.AIDS >= 0.013
                       Income.composition.of.resources >= 0.56
                       Adult.Mortality >= 0.375

Life.expectancy is 0.45 when
                       HIV.AIDS < 0.013
                       Income.composition.of.resources < 0.74
                       Adult.Mortality >= 0.369
```

```
Life.expectancy is 0.50 when
                        HIV.AIDS is 0.013 to 0.035
                        Income.composition.of.resources < 0.56
                        Adult.Mortality < 0.516
                        under.five.deaths < 0.0098

Life.expectancy is 0.54 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources < 0.74
                        Adult.Mortality is 0.292 to 0.369

Life.expectancy is 0.58 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources < 0.74
                        Adult.Mortality < 0.292
                        Alcohol < 0.041
                        Polio < 0.94

Life.expectancy is 0.61 when
                        HIV.AIDS >= 0.013
                        Income.composition.of.resources >= 0.56
                        Adult.Mortality < 0.375

Life.expectancy is 0.64 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources is 0.74 to 0.88
                        Adult.Mortality >= 0.249

Life.expectancy is 0.66 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources < 0.74
                        Adult.Mortality < 0.292
                        Alcohol < 0.041
                        Polio >= 0.94

Life.expectancy is 0.67 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources < 0.74
                        Adult.Mortality < 0.292
                        Alcohol >= 0.041

Life.expectancy is 0.72 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources is 0.74 to 0.88
                        Adult.Mortality < 0.249
                        thinness..1.19.years >= 0.074

Life.expectancy is 0.77 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources is 0.74 to 0.88
                        Adult.Mortality < 0.249
                        thinness..1.19.years < 0.074

Life.expectancy is 0.81 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources >= 0.88
                        Adult.Mortality >= 0.106

Life.expectancy is 0.86 when
                        HIV.AIDS < 0.013
                        Income.composition.of.resources >= 0.88
                        Adult.Mortality < 0.106
> #plotting the decision tree
> fancyRpartPlot(m2)
```




```
> ##Identify the importance of the variables
> m2$variable.importance
                      HIV.AIDS                 Adult.Mortality Income.composition.of.resources
                     53.2367424                     41.1552305                     36.6336402
                     Schooling           thinness..1.19.years                         Polio
                     27.6509983                     11.7358957                     10.8262353
             under.five.deaths                        Alcohol                            BMI
                      6.7497501                      6.1116737                      6.0375914
        percentage.expenditure                        Measles                    Diphtheria
                      3.1328133                      1.2423417                      1.0929487
                     Developed              Total.expenditure                     Population
                      1.0100585                      0.3708643                      0.3378516
                   Hepatitis.B
                      0.2671022
> #Prediction using train data
> train.pred.cart=train.cart
> train.pred.cart$predTrain.cart=predict(m2, newdata = train.pred.cart)
> #Mean Absolute ERROR
> mae(train.pred.cart$Life.expectancy,train.pred.cart$predTrain.cart)
[1] 0.04303205
> #ROOT MEAN SQUARE ERROR
> rmse(train.pred.cart$Life.expectancy,train.pred.cart$predTrain.cart)
[1] 0.05734288
> #Prediction using test data
> test.pred.cart=test.cart
> test.pred.cart$predTest.cart=predict(m2, newdata = test.pred.cart)


> #Prediction using test data
> test.pred.cart=test.cart
> test.pred.cart$predTest.cart=predict(m2, newdata = test.pred.cart)
> #Mean Absolute ERROR
> mae(test.pred.cart$Life.expectancy,test.pred.cart$predTest.cart)
[1] 0.05301913
> #ROOT MEAN SQUARE ERROR
> rmse(test.pred.cart$Life.expectancy,test.pred.cart$predTest.cart)
[1] 0.07233939
```

```
> ##KNN Regression
> train.knn=scaled.train
> test.knn=scaled.test
> trControl=trainControl(method='repeatedcv',number=10,repeats=3)
> set.seed(100)
> fit=train(Life.expectancy~.,data=train.knn,tuneGrid=expand.grid(k=1:20),method='knn',trControl=trControl)
> varImp(fit)
loess r-squared variable importance

                               Overall
Income.composition.of.resources 100.000
Schooling                        70.903
Adult.Mortality                  62.654
BMI                              58.686
HIV.AIDS                         41.406
percentage.expenditure           38.550
Diphtheria                       29.735
thinness..1.19.years             29.692
Developed                        29.113
Polio                            27.309
Alcohol                          22.510
Hepatitis.B                      11.413
Total.expenditure                11.360
Measles                          10.879
under.five.deaths                 4.666
Population                        0.000
```


```
Developed                        29.113
Polio                            27.309
Alcohol                          22.510
Hepatitis.B                      11.413
Total.expenditure                11.360
Measles                          10.879
under.five.deaths                 4.666
Population                        0.000
```


```
> #prediction using KNN
> pred=predict(fit,newdata=test.knn)
> #Mean Absolute ERROR
> mae(test.knn$Life.expectancy,pred)
[1] 0.06380952
> #ROOT MEAN SQUARE ERROR
> rmse(test.knn$Life.expectancy,pred)
[1] 0.08030729
```

***