JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY,

Sector - 62, Noida

BIG DATA AND DATA ANALYTICS

PROJECT REPORT

Life Expectancy Analysis and Prediction

GROUP MEMBERS (Batch: B3 AND B4)

APEKSHA JAIN	21103081
ANUSHA BHAT	21103084
ARYAN JOLLY	21103110
POOJA SHARMA	21103063
VIVEK RAI	21103059

SUBMITTED TO: Dr Pawan Kumar Upadhyay

ABSTRACT

Life Expectancy is usually used as a term, but hardly understood. As per its definition it's a statistical measure of the typical (see below) time an organism is anticipated to measure, supported the year of its birth, its current age, and other demographic factors including gender; but it's such a lot over that as described within the upcoming sections.

Period lifetime is one among the foremost used summary indicators for the health of a population. Its levels and trends direct health policies, and researchers try and identify the determining risk factors to assess and forecast future developments. the utilization of period lifespan is usually supported the idea that it directly reflects the mortality conditions of a specific year. Accordingly, the reason for changes in lifespan are typically sought in factors that have a right away impact on current mortality conditions.

Although there are lot of studies undertaken within the past on factors affecting anticipation considering demographic variables, income composition and mortality rates. it had been found that effect of immunization and human development index wasn't taken into consideration within the past. Also, a number of the past researches was done considering multiple statistical regression supported data set of 1 year for all the countries. Hence, this offers motivation to resolve both the factors stated previously by formulating a regression model supported mixed effects model and multiple rectilinear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like viral hepatitis, Polio and Diphtheria also will be considered. in an exceedingly nutshell, this study will concentrate on immunization factors, mortality factors, economic factors, social factors and other health related factors also. Since the observations this dataset are supported different countries, it'll be easier for a rustic to work out the predicting factor which is contributing to lower value of lifetime. this can help in suggesting a rustic which area should run importance so as to efficiently improve the life of its population. It's often argued that lifespan across the planet has only increased because child mortality has fallen. If this were true, this is able to mean that we've become far better at preventing young children from dying, but have achieved nothing to enhance the survival of older children, adolescents and adults. Once past childhood, people would be expected to enjoy the identical length of life as they did centuries ago.

This, as we are going to explore, is untrue. expectancy has increased in the least ages. the typical person can expect to measure a extended life than within the past, no matter what age they're.

INTRODUCTION

It is supported the set of observed age-specific death rates, i.e., the quantity of deaths during a certain year and people divided by the typical number of individuals alive during this year and cohort. These death rates are then transformed into probabilities of dying and connected to a survival function from birth to the very best age during which people reside. The mean age at death derived from this survival function is that the PLE. It are often interpreted because the average number of years that newborns of a particular period would live under the hypothetical scenario that the prevailing age-specific death rates remain constant within the future.

The project relies on accuracy of information. the worldwide Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status similarly as many other related factors for all countries the data-sets are made available to public for the aim of health data analysis. The data-set associated with expectancy, health factors for 193 countries has been collected from the identical WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. it's been observed that within the past 15 years, there has been a large development in health sector leading to improvement of human mortality rates especially within the developing nations as compared to the past 30 years. Therefore, during this project we've considered data from year 2000-2015 for 193 countries for further analysis. The individual data files are merged together into one data-set. On initial visual inspection of the information showed some missing values. because the data-sets were from WHO, we found no evident errors. The result indicated that almost all of the missing data was for population, hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it had been decided that we exclude these countries from the ultimate model data-set. the ultimate merged file (final dataset) consists of twenty-two Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors. This project basically focusses on analysis of anticipation using regression. multivariate analysis could be a method that helps to investigate and understand the link between two or more variables of interest. the method that's adapted to perform multivariate analysis helps to grasp which factors are important, which factors are often ignored and the way they're influencing one another. For the multivariate analysis is be a successful method, we understand the subsequent terms:

Dependent Variable: This is often the variable that we try to know or forecast.

Independent Variable: These are factors that influence the analysis or target variable and supply us with information regarding the connection of the variables with the target variable.
Some more terminologies of in multivariate analysis that might be employed in this report, thus it might be great to possess an understanding on these terms:

Outliers

Suppose there's an observation within the dataset that includes a very high or very low value as compared to the opposite observations within the data, i.e. it doesn't belong to the population, such an observation is termed an outlier. In simple words, it's an extreme value. An outlier could be a problem because persistently it hampers the results we get.

Multicollinearity

When the independent variables are highly correlated to every other, then the variables are said to be multicollinear. many varieties of regression techniques assume multicollinearity mustn't be present within the dataset. it's because it causes problems in ranking variables supported its importance, or it makes the task difficult in selecting the foremost important variable.

Heteroscedasticity

When the variation between the target variable and therefore the experimental variable isn't constant, it's called heteroscedasticity. Example-As one’s income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times, eat expensive meals. Those with higher incomes display a greater variability of food consumption.

Underfit and Overfit

When we use unnecessary explanatory variables, it'd result in overfitting. Overfitting implies that our algorithm works well on the training set but is unable to perform better on the test sets. it's also called a controversy of high variance. When our algorithm works so poorly that it's unable to suit even a training set well, then it's said to underfit the info. it's also referred to as a controversy of high bias

LITERATURE SURVEY

Life expectancy is one among the foremost important factors in end-of-life deciding. Good prognostication for instance helps to see the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning improves the standard of the ultimate phase of life by stimulating doctors to explore the preferences for end-of-life care with their patients, and folks near the patients. This research tests the potential of using machine learning and linguistic communication processing techniques for predicting expectancy from electronic medical records. anticipation also received considerable attention worldwide. Because social determinants are treated because the most vital factors of health, lifetime is closely associated with social development. However, in China, different areas have different levels of life expectancy; developed regions, like Beijing and Shanghai, usually have the next level of expectancy than developing regions. as compared, the US citizens don't have a significantly high level of anticipation, despite its high economic development; thus, many non-economic factors also are important in determining life. Therefore, simply estimating the connection between social development and life globally and on the average cannot uncover the underlying association between them and therefore the details of the mechanism by which social development affects life.

The input and development of healthcare are directly related to population health. Before the Chinese economic reform in 1978, an enormous development of basic healthcare greatly contributed to the development of lifespan. After 1978, some researchers argued that attributable to the unequal spatial distribution of healthcare resources and services in China, there have been some disparities within the spatial distribution of lifespan. Ming and Dong reported that always expectancy promotion, it absolutely was not the rise in healthcare provisions but the rise within the usage efficiency of the healthcare services that mattered. However, in an exceedingly transnational study in Europe, Heuvel et al. argued that healthcare investments had little effects on the promotion of lifespan, while investments in social protection could greatly improve lifetime, which also suggested the indirect importance of social protection and harmony

The role of environmental development in lifespan promotion has gradually caught researchers' attention in recent years. supported the comparison of expectancy values between some

developed nations and a few post socialist countries, it's noted that the environmental disadvantages produced many detrimental impacts on anticipation among the residents of post socialist countries. additionally, during a study conducted in 156 nations, it had been also reported that a much better aquatic environment, quite improvements in economy and education development, could most importantly increase lifespan.

There were some disagreements on the effect size of social development on lifetime and therefore the effect of assorted social development dimensions on life in numerous areas and periods. However, the event of spatial analysis techniques in recent years has laid an honest foundation for the understanding of spatial relationships. Nevertheless, within the research area of anticipation, the appliance of spatial analyses is generally limited to univariate analysis, like spatial autocorrelation or hot spot analysis, and people spatial regression models don't seem to be widely utilized in related fields. Thus, only some previous studies have focused on the analysis or control of spatial relationships when examining the influencing factors of specific health outcomes.

Besides, there are 3 major effects we'd like to understand about before we start exploring and predicting with the given data.

Cohort effects

Cohort effects originate from “health conditions that a cohort faces at a given time which have a delayed impact on the cohort’s mortality”. Such conditions include, e.g., exposure to infectious or noninfectious diseases, exposure to radiation during nuclear catastrophes, exposure to malnutrition during and after famines or wars, and health behaviors like diet, physical activity, alcohol consumption, or smoking. Cohort effects and their impact on mortality are extremely heterogeneous. because the given examples illustrate, they'll change gradually over time, interact with one another, affect mortality over a broad range of ages, differ within the length of delay, and that they can affect mortality both within the short in addition as within the future. Last but not least, the short- and long-term impacts may even operate in numerous directions. for example, a more frequent exposure to diseases that immunity will be acquired can generate high childhood mortality on the one hand, but low later life mortality from these diseases on the opposite.

Heterogeneity effects

Heterogeneity describes the circumstance that not all members of a population face the identical risk of dying in a very certain year and age. The characteristics that shape individuals' mortality risks could also be fixed at birth or at a young age (e.g., ethnicity or education) or they'll vary with age (e.g., health status or income). The composition of the full population of low-risk and high-risk individuals would be no problem if it were constant over time and across populations.

In addition to such structural effects of heterogeneity, the various mortality risks of people result in "selection effects". These are closely associated with cohort effects as they also develop their impact on age-specific death rates along the cohorts' life courses

Tempo effects

Tempo Effects emerge in death rates as soon as mortality is changing during the observation period. as an example, when improvements in health and living conditions result in a discount in mortality, a specific number of deaths – which might have occurred under unchanged mortality conditions – are postponed to a later period. Such a postponement consequentially deflates the numerator of death rates by the amount of avoided deaths, while the denominator is inflated by the identical number of saved lives.

The number of deaths decreases by a way larger proportion than the amount of the population in danger within the denominator increases. during this way, TE magnify the effect of the shifted number of deaths within the death rate. The larger the changes in mortality, the larger the magnification effect

NOVELTY

Period life is one amongst the foremost used summary indicators for the health of a population. Its levels and trends direct health policies, and researchers attempt to identify the determining risk factors to assess and forecast future developments. the utilization of period lifespan is usually supported the idea that it directly reflects the mortality conditions of a specific year. Accordingly, the reason for changes in lifetime are typically sought in factors that have an on the spot impact on current mortality conditions. it's frequently overlooked, however, that this indicator also can be suffering from a minimum of three forms of effects, particularly within

the situation of short-term fluctuations: cohort effects, heterogeneity effects, and tempo effects. We aim to boost visualizations and EDA, and further apply linear models to search out the simplest predictor.

Although there are lot of studies undertaken within the past on factors affecting expectancy considering demographic variables, income composition and mortality rates. it absolutely was found that effect of immunization and human development index wasn't taken into consideration within the past. Also, a number of the past research was done considering multiple rectilinear regression supported data set of 1 year for all the countries. Hence, this offers motivation to resolve both the factors stated previously by formulating a regression model supported mixed effects model and multiple rectilinear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like viral hepatitis, Polio and Diphtheria also will be considered. during a nutshell, we aim to specialize in immunization factors, mortality factors, economic factors, social factors and other health related factors similarly. Since the observations this dataset are supported different countries, it'll be easier for a rustic to see the predicting factor which is contributing to lower value of expectancy. this may help in suggesting a rustic which area should incline importance so as to efficiently improve the lifetime of its population.

METHODOLOGY

Description

Exploring life expectancy and looking for data on the following aspects (features):

· Birth Rate

· Cancer Rate

· Dengue Cases

· Environmental Performance Index (EPI)

· Gross Domestic Product (GDP)

· Health Expenditure

· Heart Disease Rate

· Population

· Area

· Population Density

· Stroke Rate

Target is Life Expectancy, measured in number of years.

The assumptions are:

1. These are country level average
2. There is no distinction between male and female

The following data science process is used in the analysis:

- Data collection, data cleaning, Exploratory Data Analysis
- Feature selection, feature engineering
- Model selection, model tuning and hyperparameter tuning
- Model optimization based on selected performance metric

Tools used for this analysis include:

- Python libraries, particularly Numpy and Pandas for manipulating data structures
- Matplotlib and Seaborn for visualization
- Scikit-Learn for regression analysis

Model Selection

The aim is to fit the following models on the train data set:

- **Linear Regression** (a straight line which approximates the relationship between the dependent variables and the independent target variable)
- **Ridge Regression** (this reduces model complexity while keeping all coefficients in the model, known as L2 penalty)
- **LASSO Regression** (Least Absolute Shrinkage and Selection Operator reduces model complexity by penal Qising model coefficients to zero, ie, L1 penalty)
- **Degree 2 Polynomial Regression** (a curve line to approximate the relationship between the dependent variables and the independent target variable)

Dataset Description

In spite of the actual fact that there are a part of studies attempted within the past on components influencing future wondering segment factors, pay piece and death rates. it had been discovered that effect of vaccination and human improvement record wasn't considered

before. Likewise, some of the past examination was finished considering numerous straight relapse addicted to informational index of 1 year for all the nations. Thus, this offers inspiration to see both the components expressed already by detailing a relapse model keen about blended impacts model and diverse direct relapse while brooding about information from a time of 2000 to 2015 for all the nations.

Significant inoculation like viral hepatitis, Polio and Diphtheria will likewise be thought of. More or less, this examination will zero in on inoculation factors, mortality factors, monetary elements, social elements and other wellbeing related factors also. Since the perceptions this dataset rely upon various nations, it'll be simpler for a nation to choose the foreseeing factor which is adding to bring down estimation of future. this may help in proposing a nation which zone should be provided significance to effectively improve the longer term of its populace.

The undertaking depends on exactness of knowledge. the world Health Observatory (GHO) information archive under World Health Organization (WHO) monitors the wellbeing status even as numerous other related elements for all nations The informational indexes are made accessible to public with the top goal of wellbeing information examination. The informational index identified with future, wellbeing factors for 193 nations has been gathered from an analogous WHO information store site and its comparing financial information was gathered from United Nation site. Among all classifications of wellbeing related factors just those basic variables were picked which are more agent. it's been seen that within the previous 15 years , there has been a huge advancement in wellbeing area bringing about progress of decease rates particularly within the agricultural countries in contrast with the previous 30 years. Thusly, during this task we've considered information from year 2000-2015 for 193 nations for extra examination. The individual information records are consolidated into a solitary informational collection. On beginning visual assessment of the data demonstrated some missing qualities. because the informational collections were from WHO, we found no clear mistakes.. the end result showed that an oversized portion of the missing information was for populace, serum hepatitis and GDP. The missing information were from less realized nations like Vanuatu, Tonga, Togo, Cabo Verde and then forth Discovering all information for these nations was troublesome and consequently, it absolutely was concluded that we bar these nations from the last model informational index.

The last combined file(final dataset) comprises of twenty-two Columns and 2938 lines which implied 20 anticipating factors. All anticipating factors was then isolated into some general categories: Immunization related components, Mortality factors, Economical elements and Social elements.

Variables in the dataset

Country- Country

Year- Year

Status- Developed or Developing status

Life Expectancy- Age(years)

Adult Mortality- Adult Mortality Rates of both sexes(probability of dying between 15&60 years per 1000 population)

Infant Deaths- Number of Infant Deaths per 1000 population

Alcohol- Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)

Percent Expenditure- Expenditure on health as a percentage of Gross Domestic Product per capita(%)

Hep B- Hepatitis B (HepB) immunization coverage among 1-year-olds(%)

Measles- number of reported measles cases per 1000 population

BMI- Average Body Mass Index of entire population

U-5 Deaths- Number of under-five deaths per 1000 population

Polio- Polio(Pol3) immunization coverage among 1-year-olds(%)

Total Expenditure- General government expenditure on health as a percentage of total government expenditure(%)

Diphtheria- Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds(%)

HIV/AIDS- Deaths per 1000 live births HIV/AIDS(0-4 years)

GDP- Gross Domestic Product per capita(in USD)

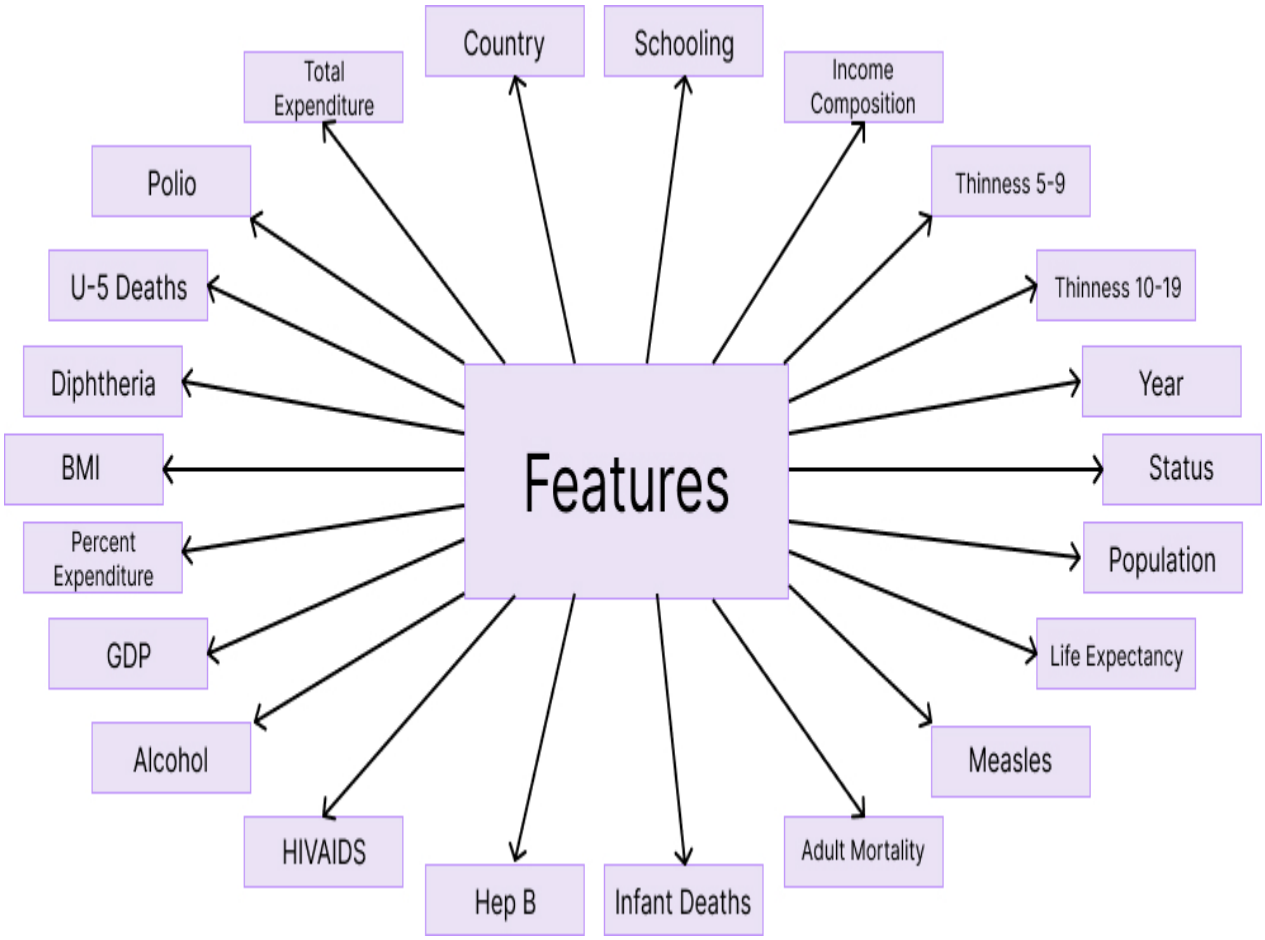
Population- Population

Thinness 10-19- Prevalence of thinness among children and adolescents for Age 10 to 19(%)

Thinness 5-9- Prevalence of thinness among children for Age 5 to 9(%)

Income Composition- Human Development Index in terms of income composition of resources(0-1)

Schooling- Number of years of Schooling

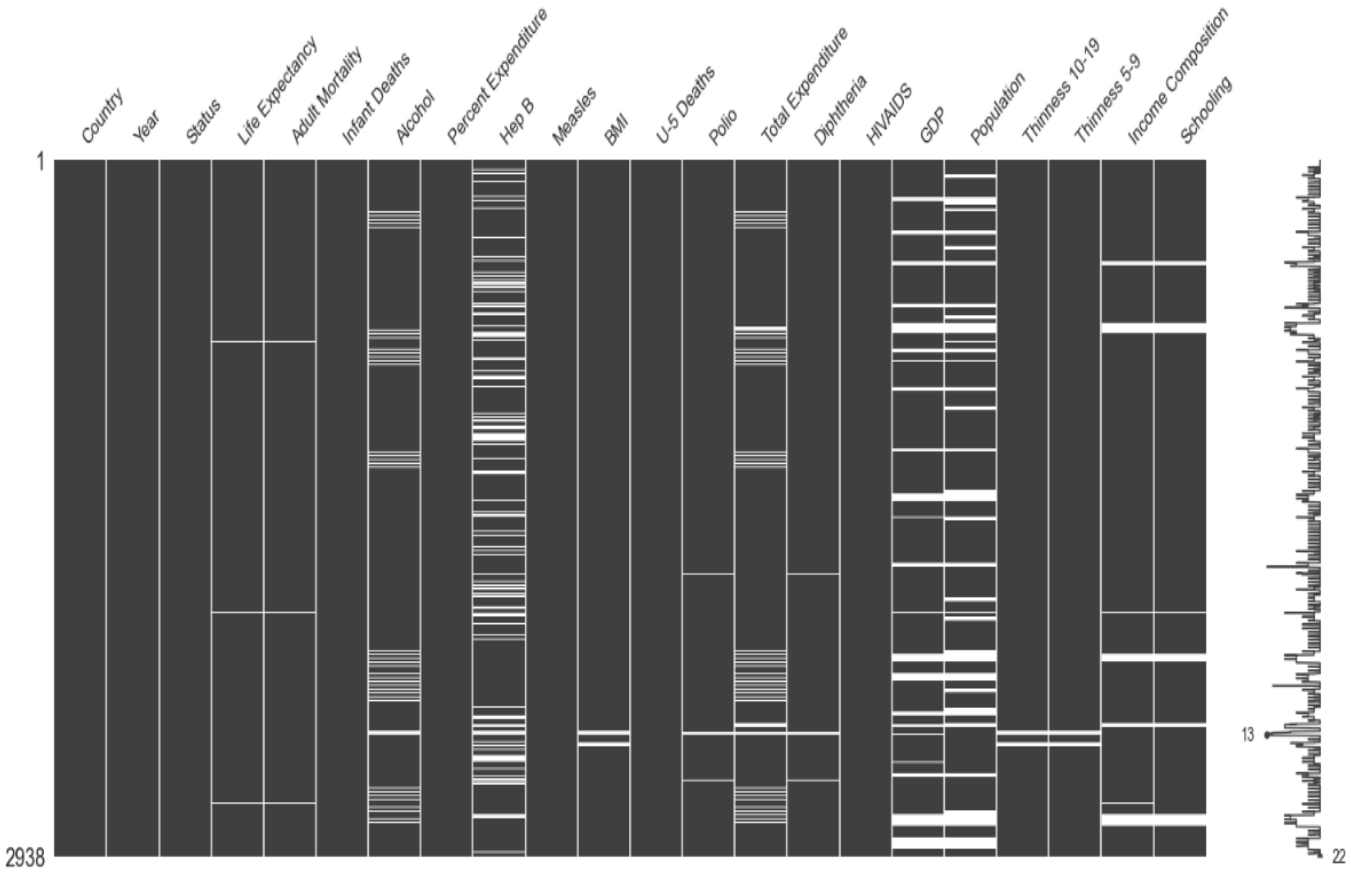

Features
Country
Schooling
Income Composition
Thinness 5-9
Thinness 10-19
Year
Status
Population
Life Expectancy
Measles
Adult Mortality
Infant Deaths
Hep B
HIVAIDS
Alcohol
GDP
Percent Expenditure
BMI
Diphtheria
U-5 Deaths
Polio
Total Expenditure


Start
Importing Libraries
Reading File
Describing and Cleaning Data
Cleaning Missing Data
Data Cleaning
Exploratory Data Analysis EDA
Life Expectancy and QQ plots
Life Expectancy Factor Plot
Regression
Feature Importance

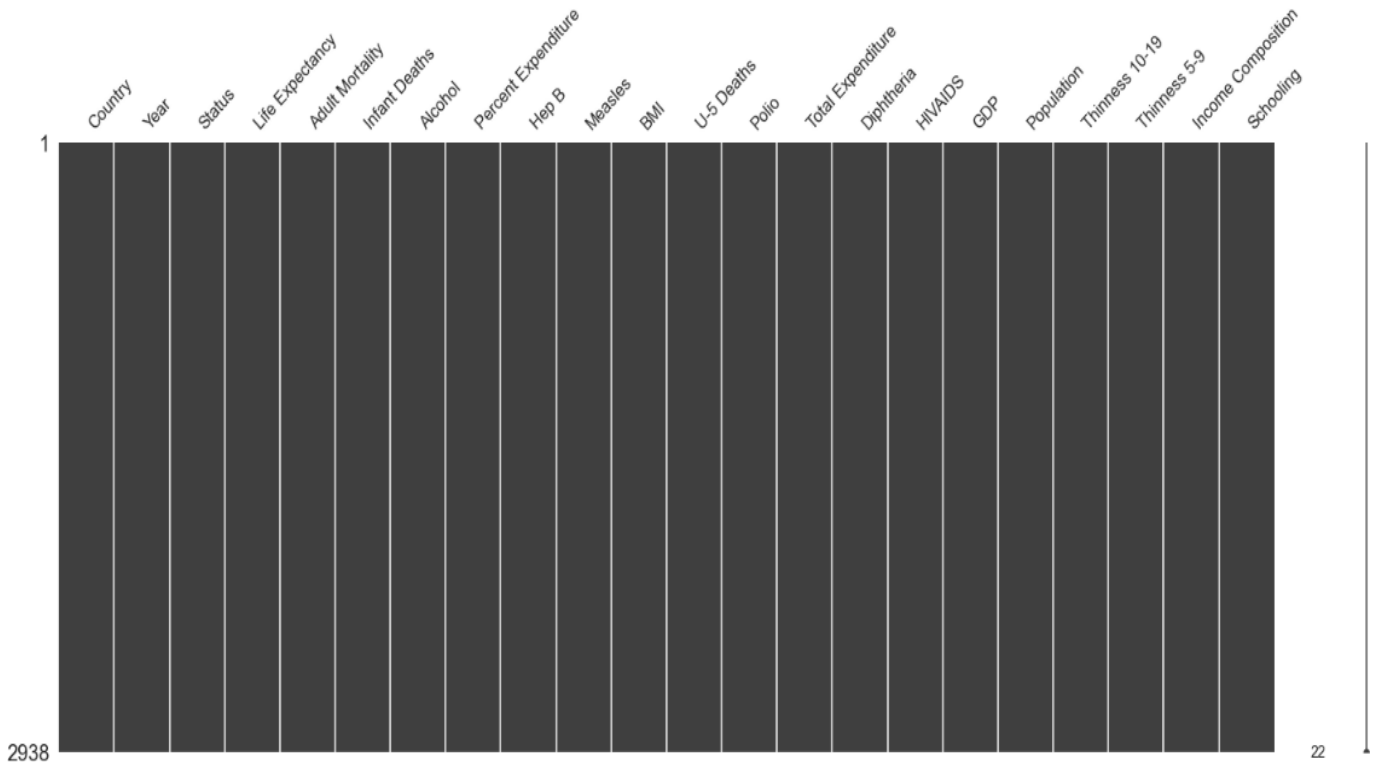
EXPERIMENTAL RESULTS (All of which can be found in the Jupyter notebook attached herewith)

```
In [3]: 1 df.columns

Out[3]: Index(['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'],
      dtype='object')
```

Column names – Depicts the name of various columns in the dataset



Visualizing missing data for better insights – Uses the missing no library in python to visualize the frequency and analyze the pattern of missing data points in each row



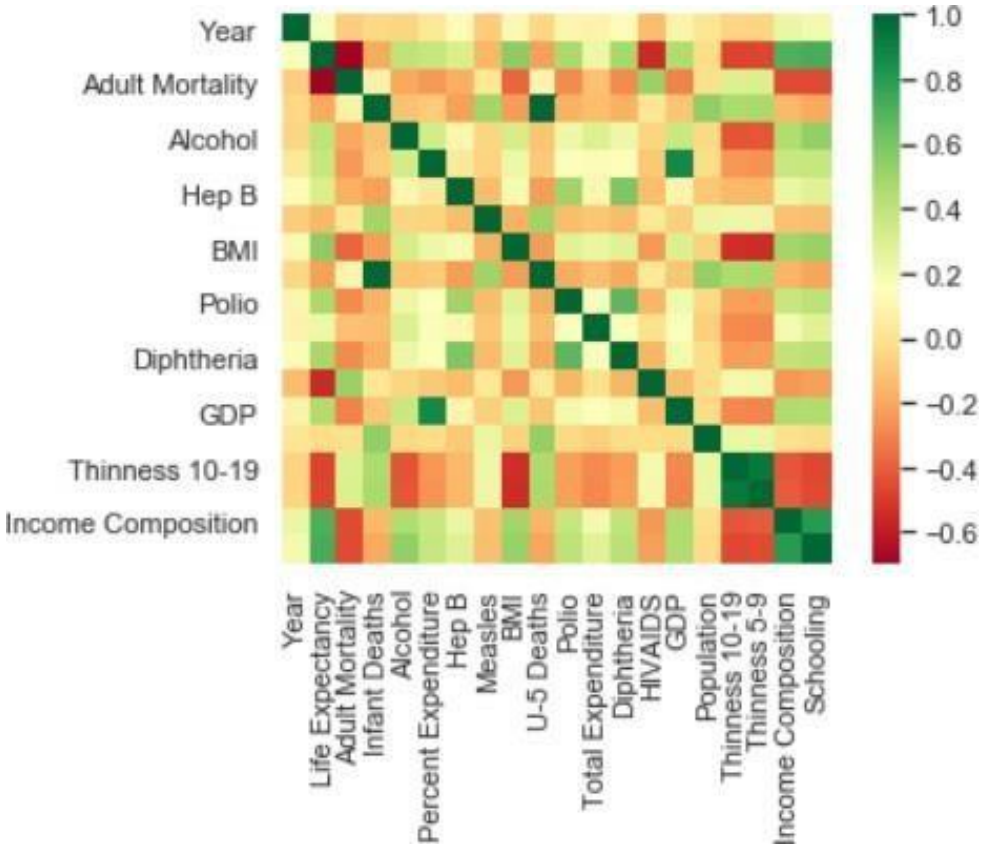
After Replacing Missing Values Associated with Country Feature Mean – Since there are no more white dashes in the middle, it signifies that the dataset now has no missing values

```
In [12]:  1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2938 entries, 0 to 2809
Data columns (total 22 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Country              2938 non-null   object
 1   Year                 2938 non-null   int64
 2   Status               2938 non-null   object
 3   Life Expectancy      2938 non-null   float64
 4   Adult Mortality      2938 non-null   float64
 5   Infant Deaths        2938 non-null   int64
 6   Alcohol              2938 non-null   float64
 7   Percent Expenditure  2938 non-null   float64
 8   Hep B                2938 non-null   float64
 9   Measles              2938 non-null   int64
 10  BMI                  2938 non-null   float64
 11  U-5 Deaths           2938 non-null   int64
 12  Polio                2938 non-null   float64
 13  Total Expenditure    2938 non-null   float64
 14  Diphtheria           2938 non-null   float64
 15  HIVAIDS              2938 non-null   float64
 16  GDP                  2938 non-null   float64
 17  Population           2938 non-null   float64
 18  Thinness 10-19       2938 non-null   float64
 19  Thinness 5-9         2938 non-null   float64
 20  Income Composition   2938 non-null   float64
 21  Schooling            2938 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 527.9+ KB
```

DataFrame Information after handling null values – A final confirmation that the dataset has been cleansed and has no null value

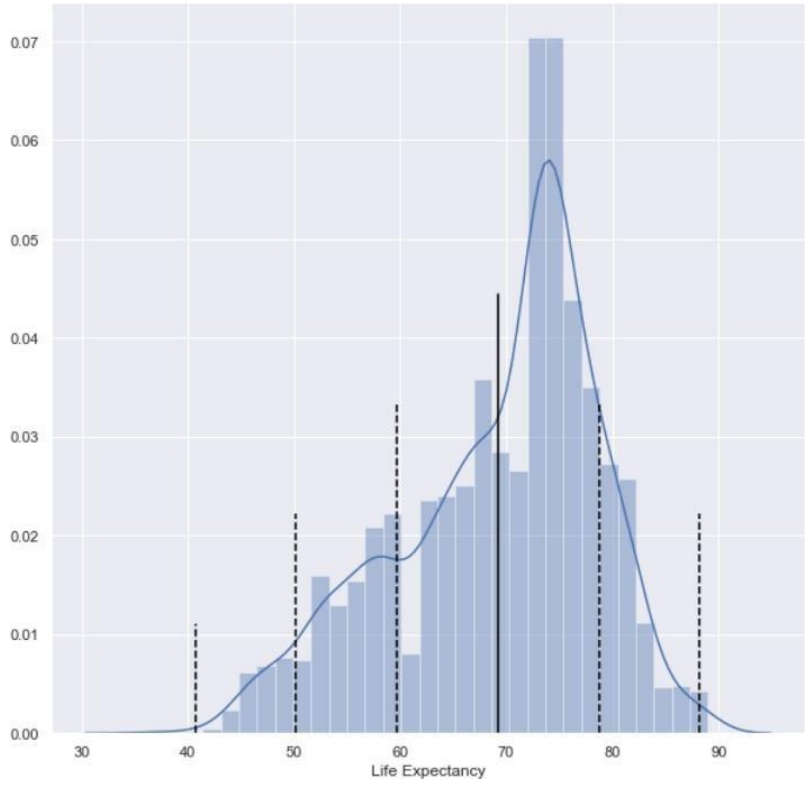
EDA: Some visuals

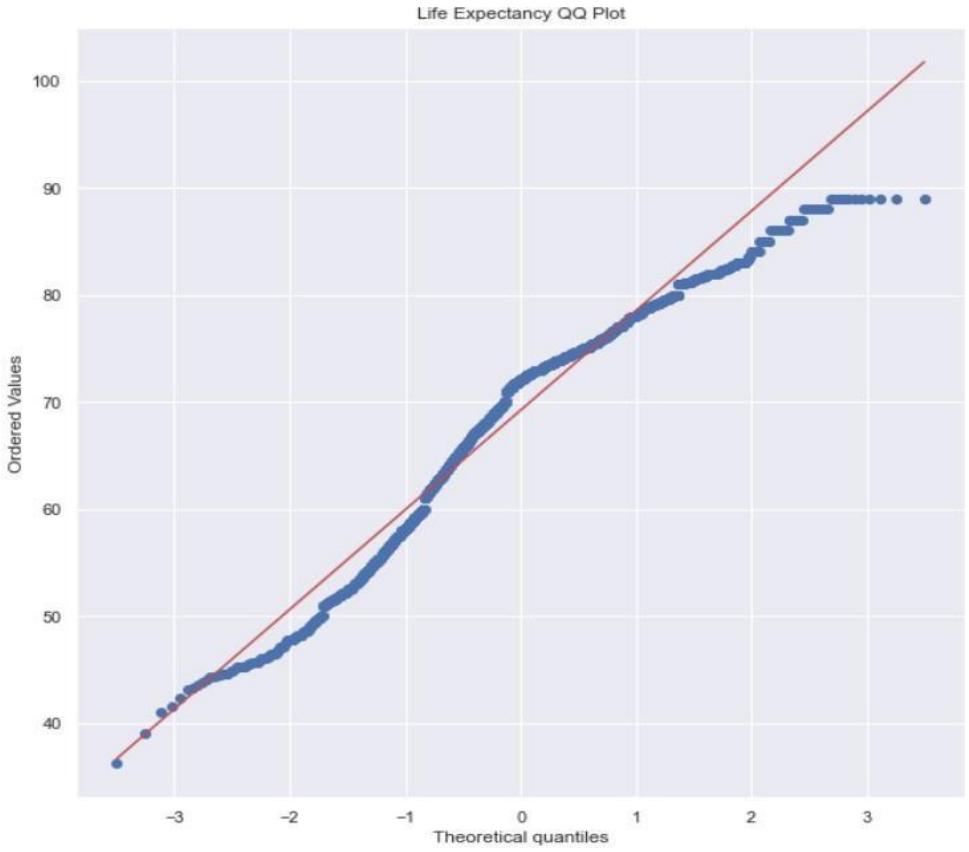


Correlation Heatmap to check for feature relation- A heatmap is a two-dimensional graphical representation of data where the individual values in a matrix, represented as colors.

Q-Q Plots - A Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

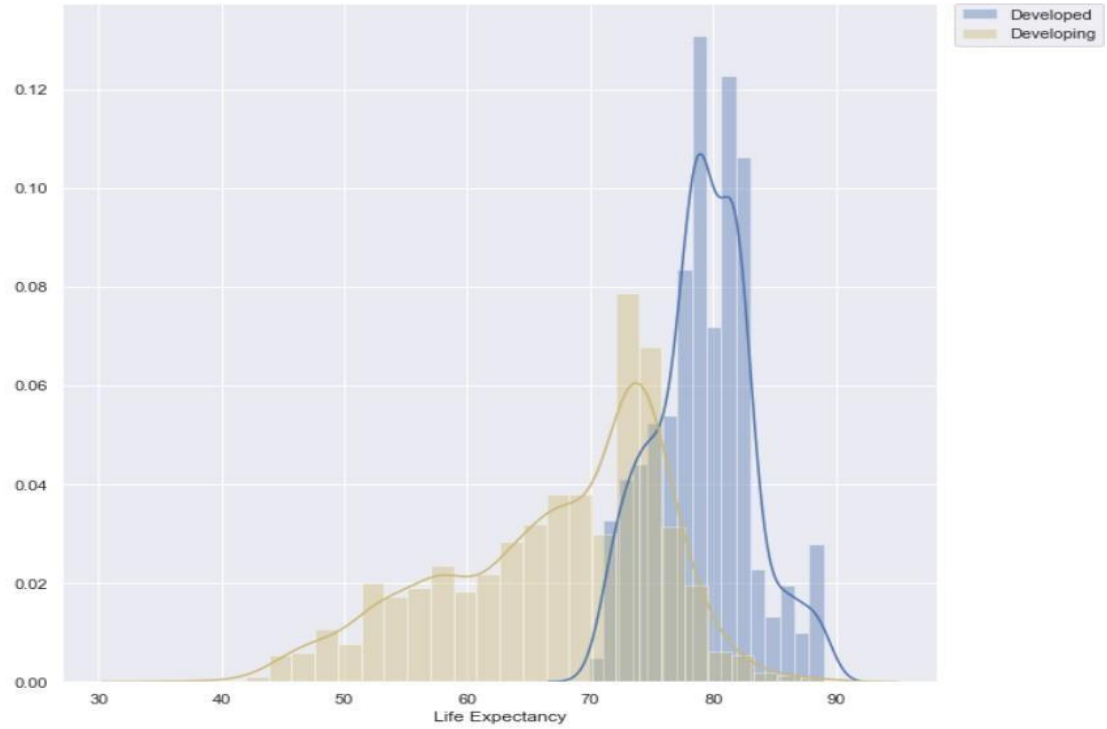
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.





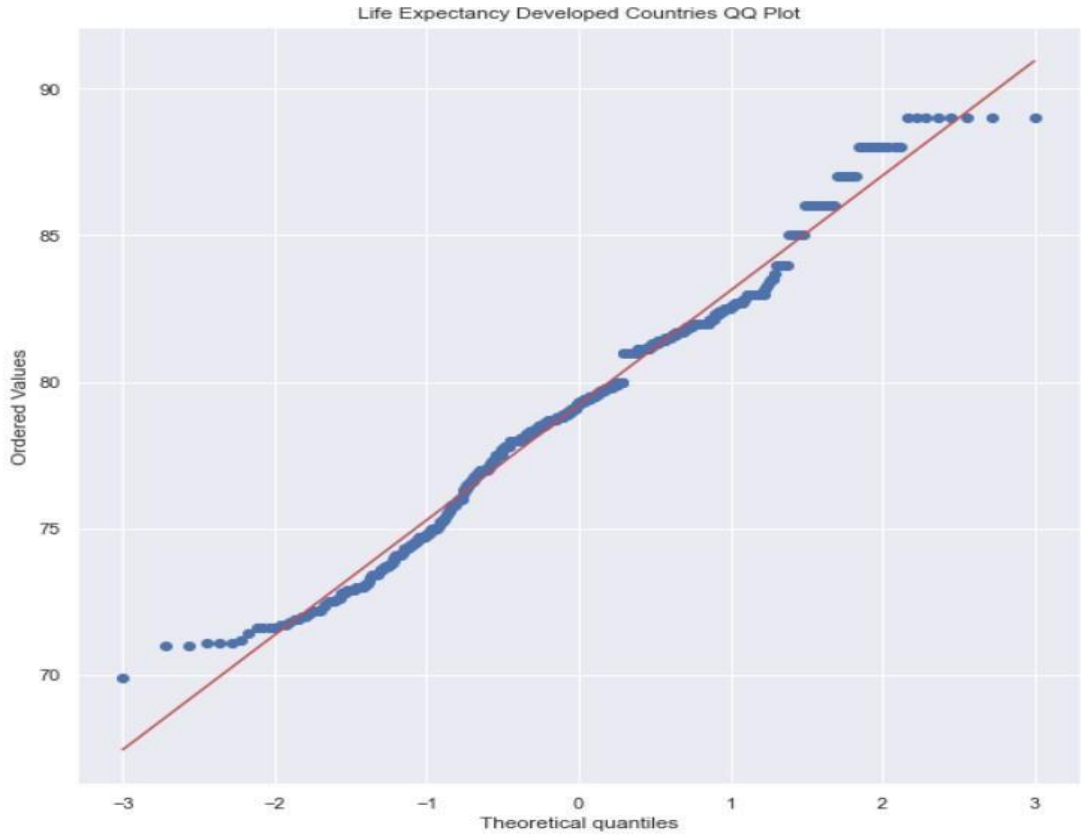

```
ShapiroResult(statistic=0.9566084742546082, pvalue=9.622531605232346e-29)
```

The maximum value 89.0 is about 2.08 standard deviations away from the mean 69.2 while the minimum 36.3 is about 3.46 deviations away. The standard deviation for the whole sample is 9.50 years.

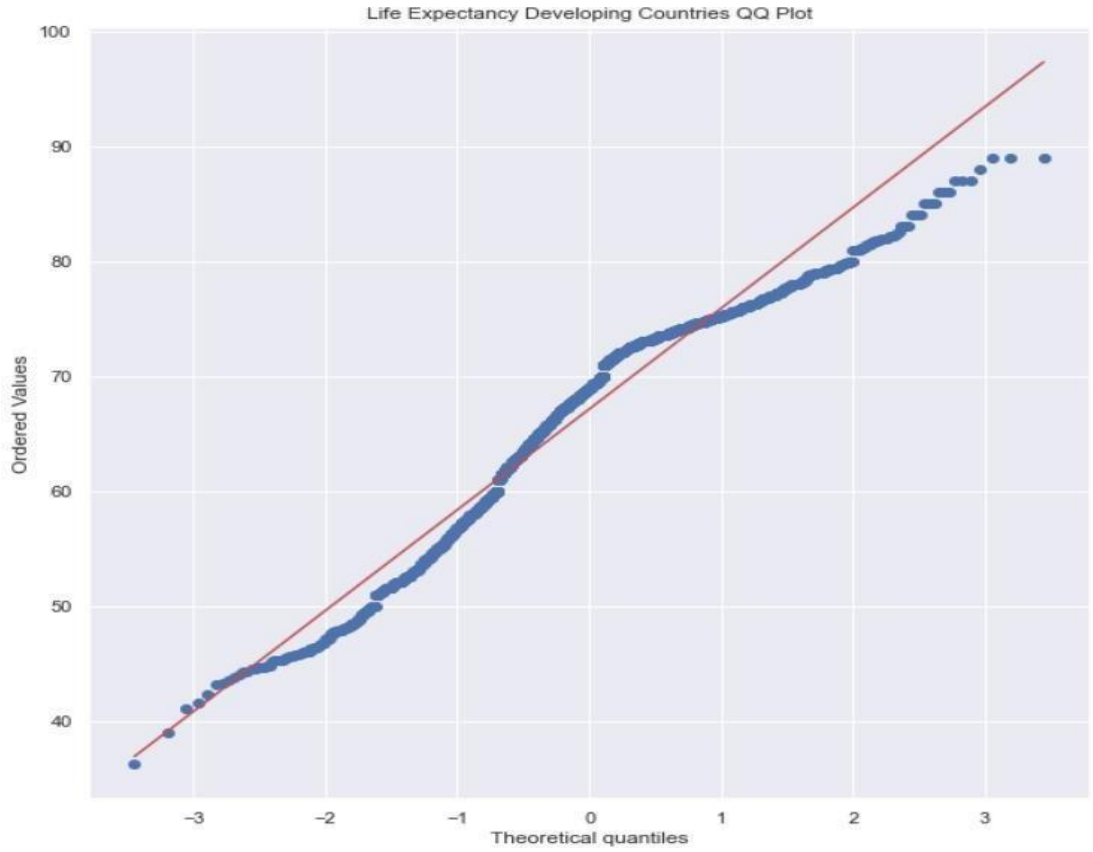



```
ShapiroResult(statistic=0.9566084742546082, pvalue=9.622531605232346e-29)
```

Comparative plot of Developed and Developing countries – Analysing differences in life expectancy curves for developing and developed countries



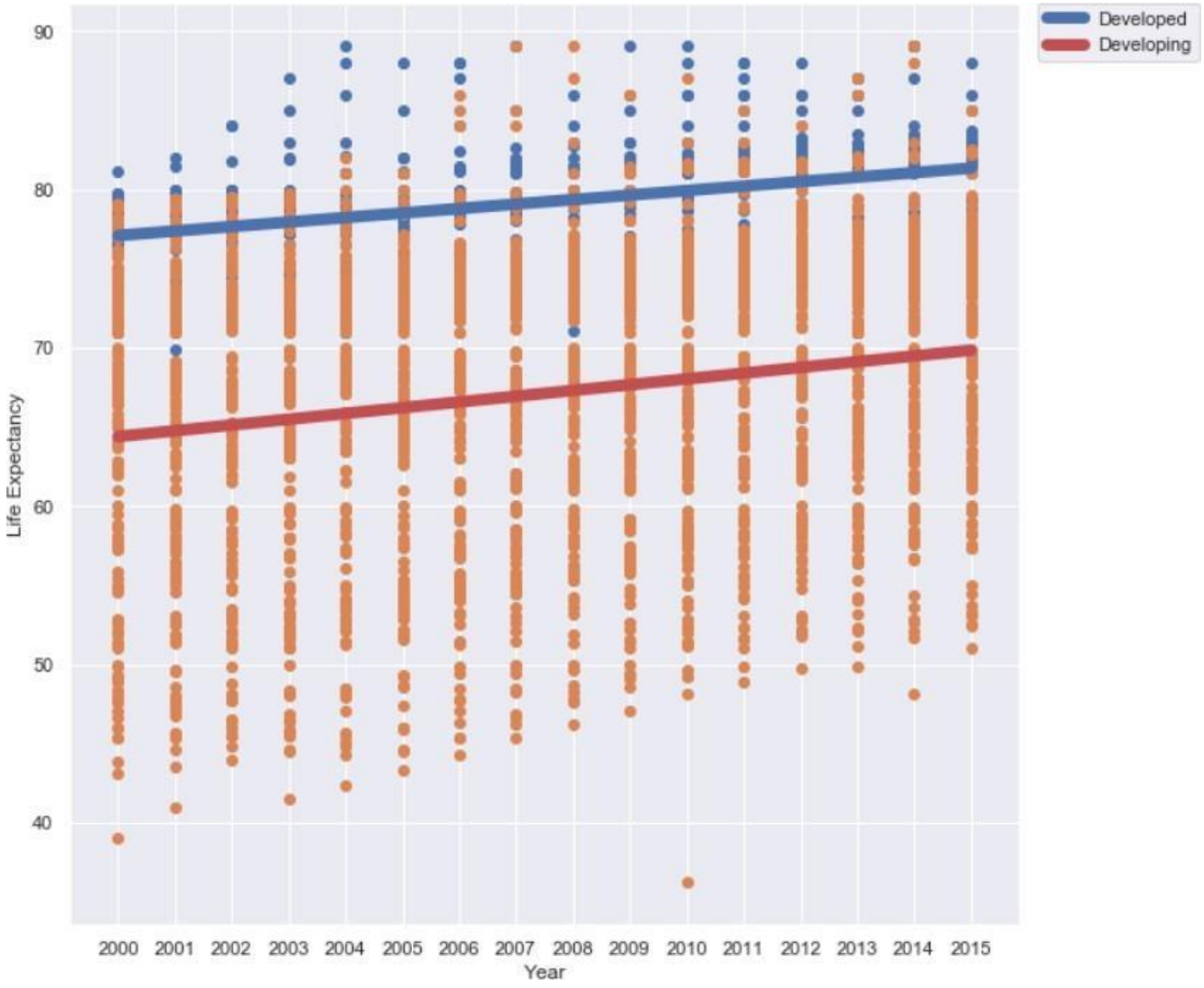

```
ShapiroResult(statistic=0.9566084742546082, pvalue=9.622531605232346e-29)
```



Status	count	mean	std	min	25%	50%	75%	max
Developed	512.0	79.197852	3.930942	69.9	76.8	79.25	81.7	89.0
Developing	2426.0	67.111465	8.987504	36.3	61.1	69.00	74.0	89.0

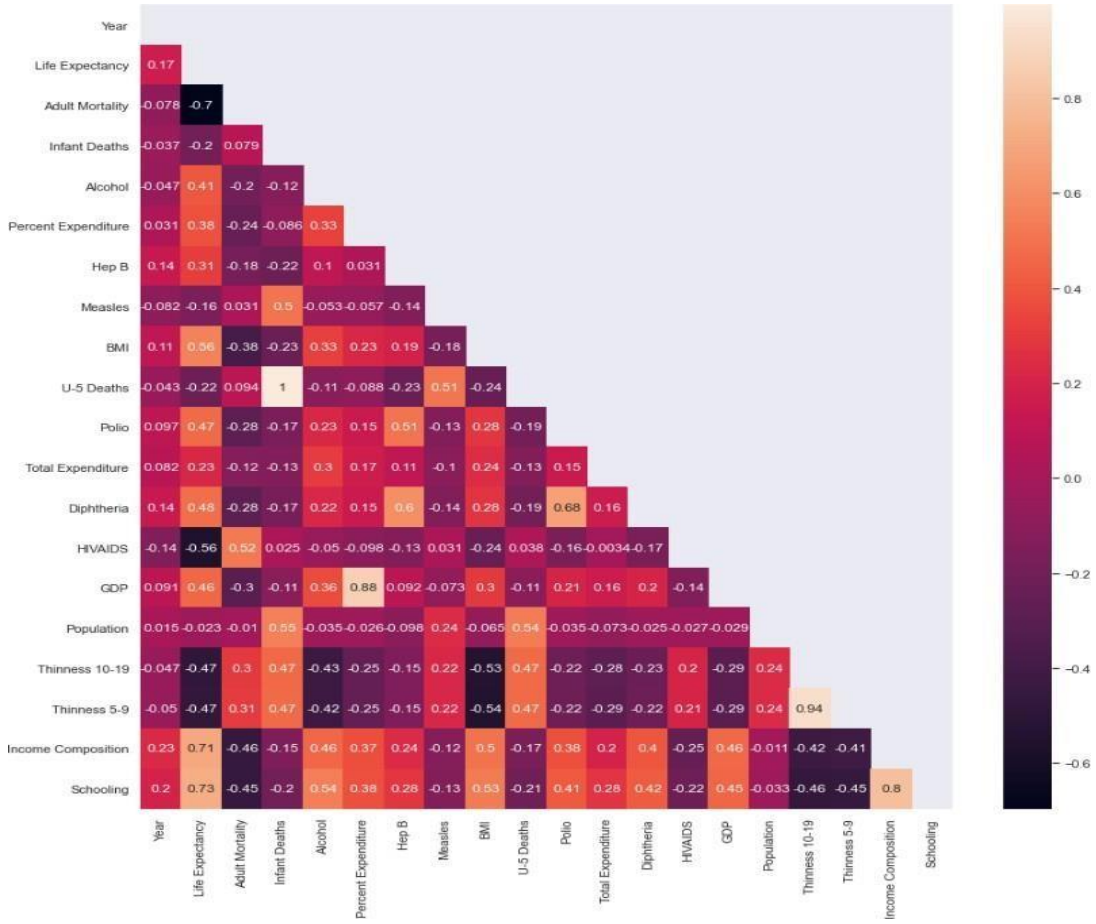
The QQ plots and the table shows the statistics for the developing and developed countries. We could have used them together as a whole but that would lead to a lot of deviation since the average life expectancy for developed and developing countries is very different

(A lor of factors contribute to the life expectancy)



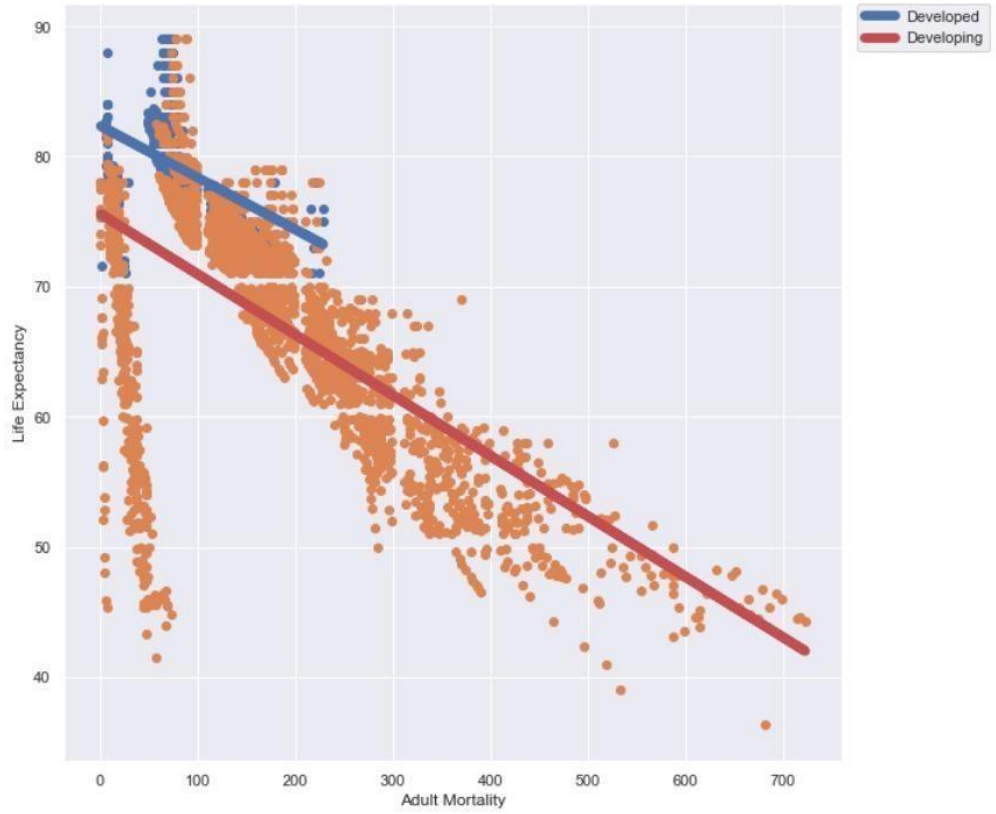
Life Expectancy Factor Plot - We see the difference in life expectancies of developing and developed countries through the LE factor plot we have made

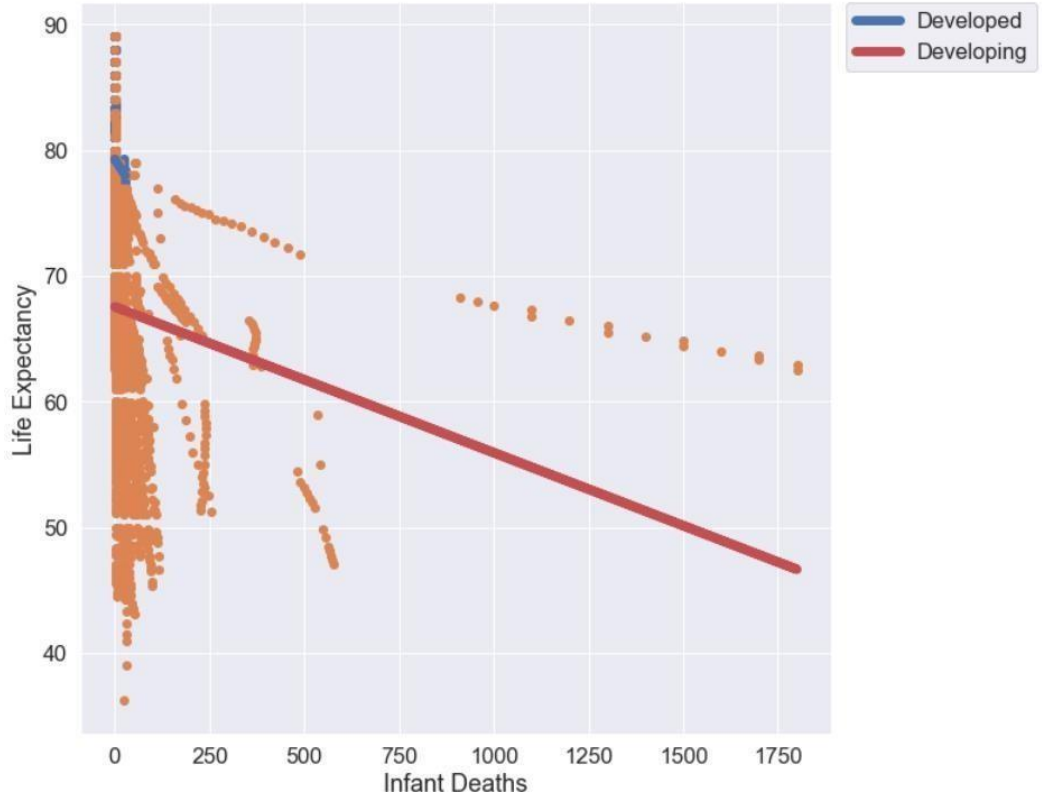
```
Correlation Values for the 193 countries left after Data Cleaning:

Life Expectancy  BMI                    0.560105
                 Diphtheria             0.483574
                 Polio                  0.470396
                 GDP                    0.455359
                 Alcohol                0.407103
                 Percent Expenditure    0.381990
                 Hep B                  0.314744
Year             Income Composition     0.233164
Life Expectancy  Total Expenditure      0.226319
Year             Schooling              0.200663
Life Expectancy  Year                   0.168709
                 Population            -0.022831
                 Measles               -0.157401
                 Infant Deaths         -0.196324
                 U-5 Deaths            -0.222286
                 HIVAIDS               -0.556165
                 Adult Mortality       -0.696386
dtype: float64
```

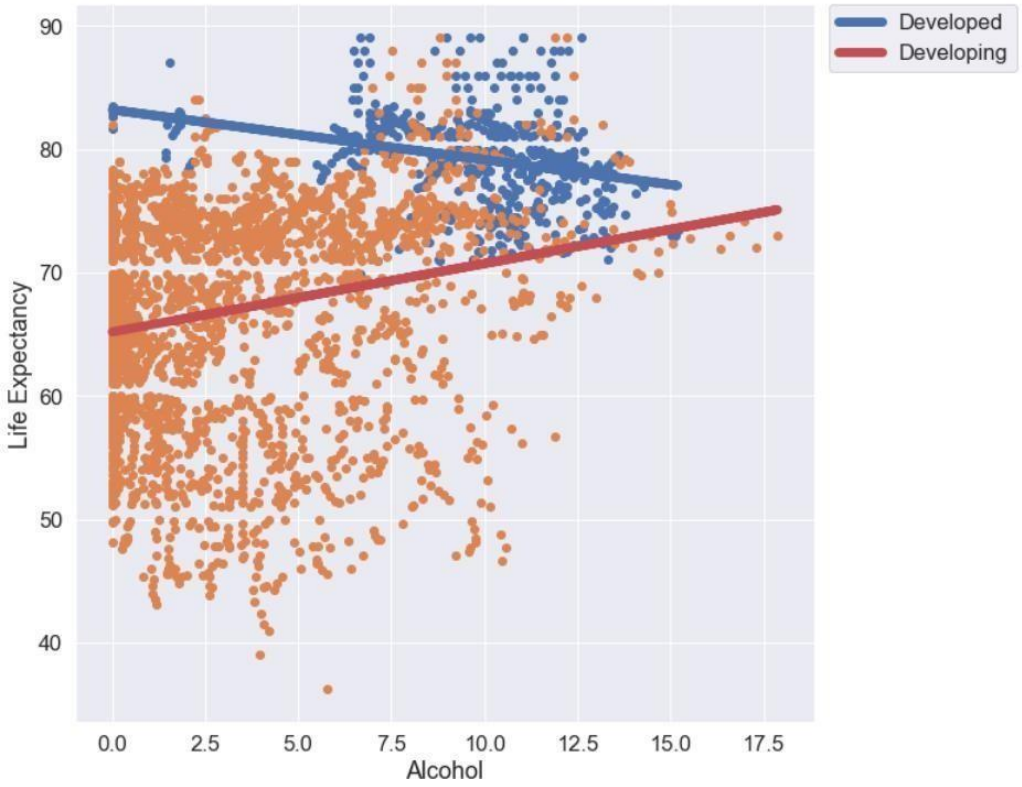


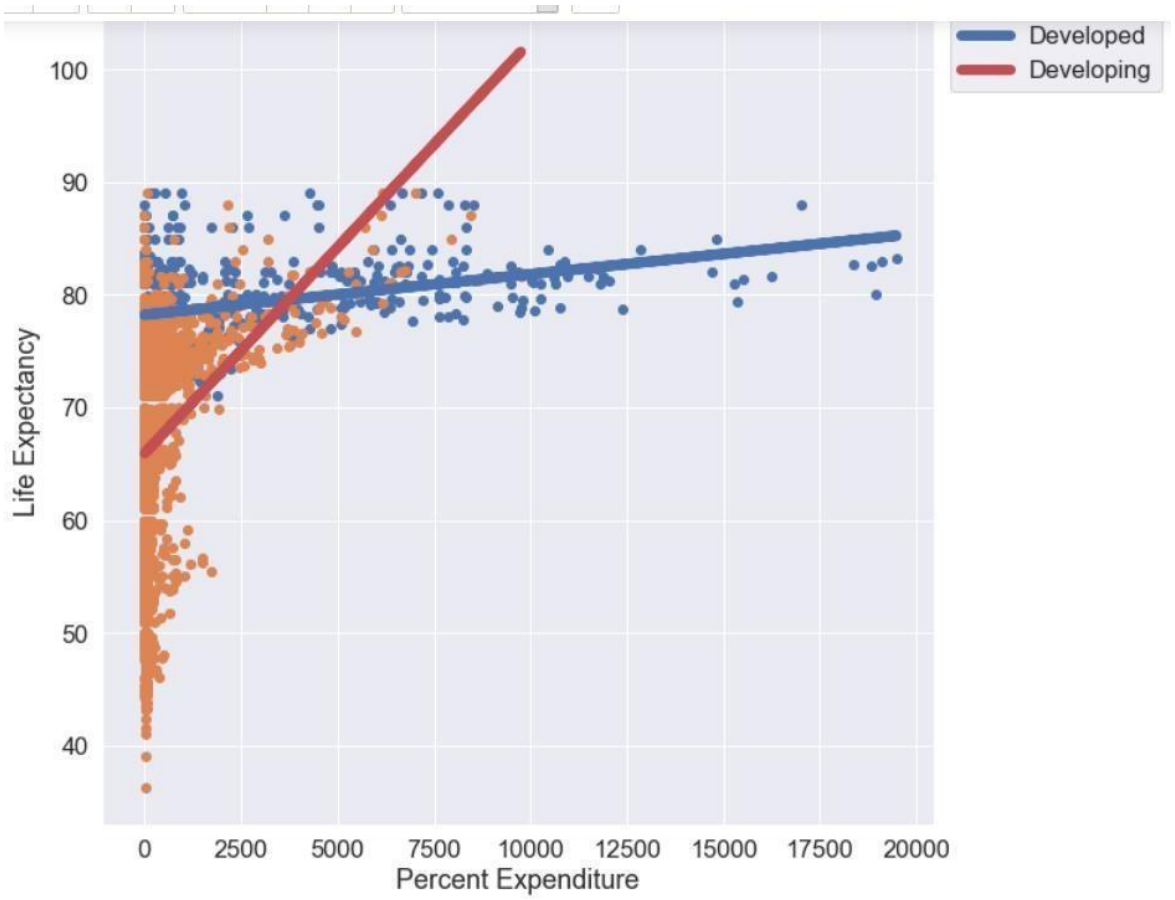
The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. We see that some features are very poorly correlated and some are highly correlated

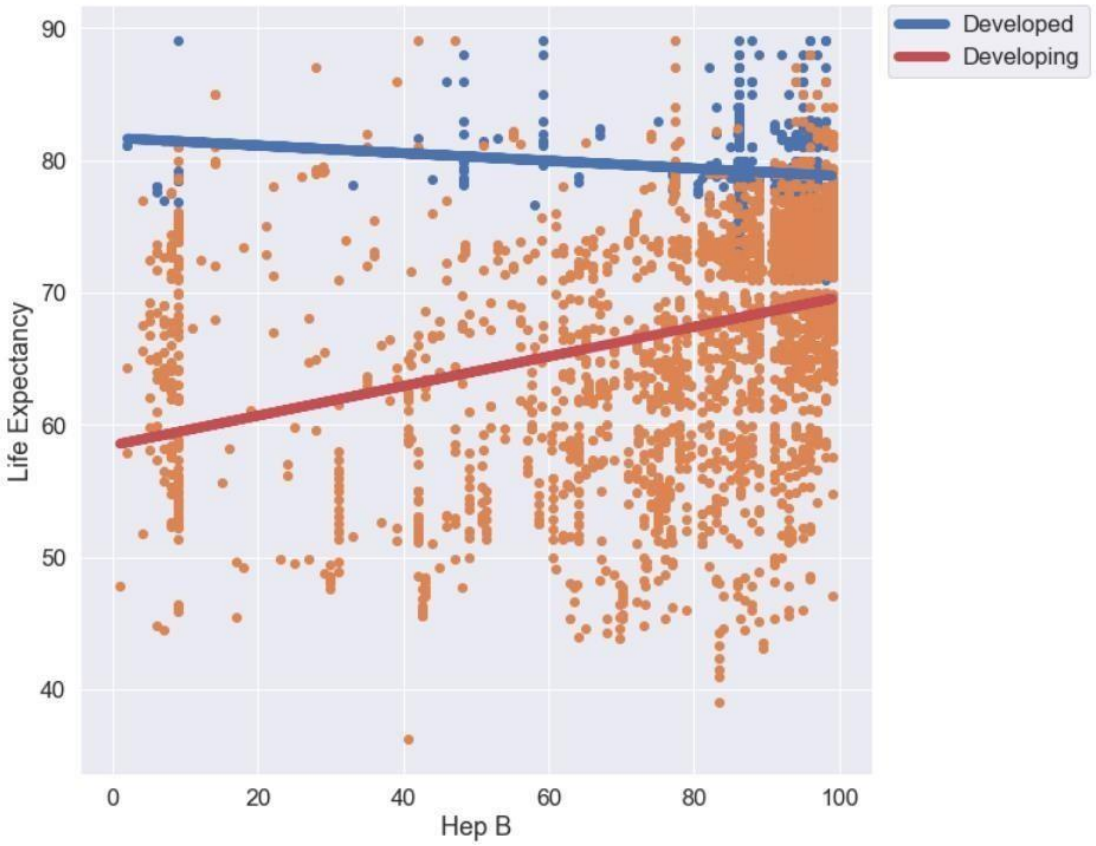
Curves for Life Expectancy v/s features

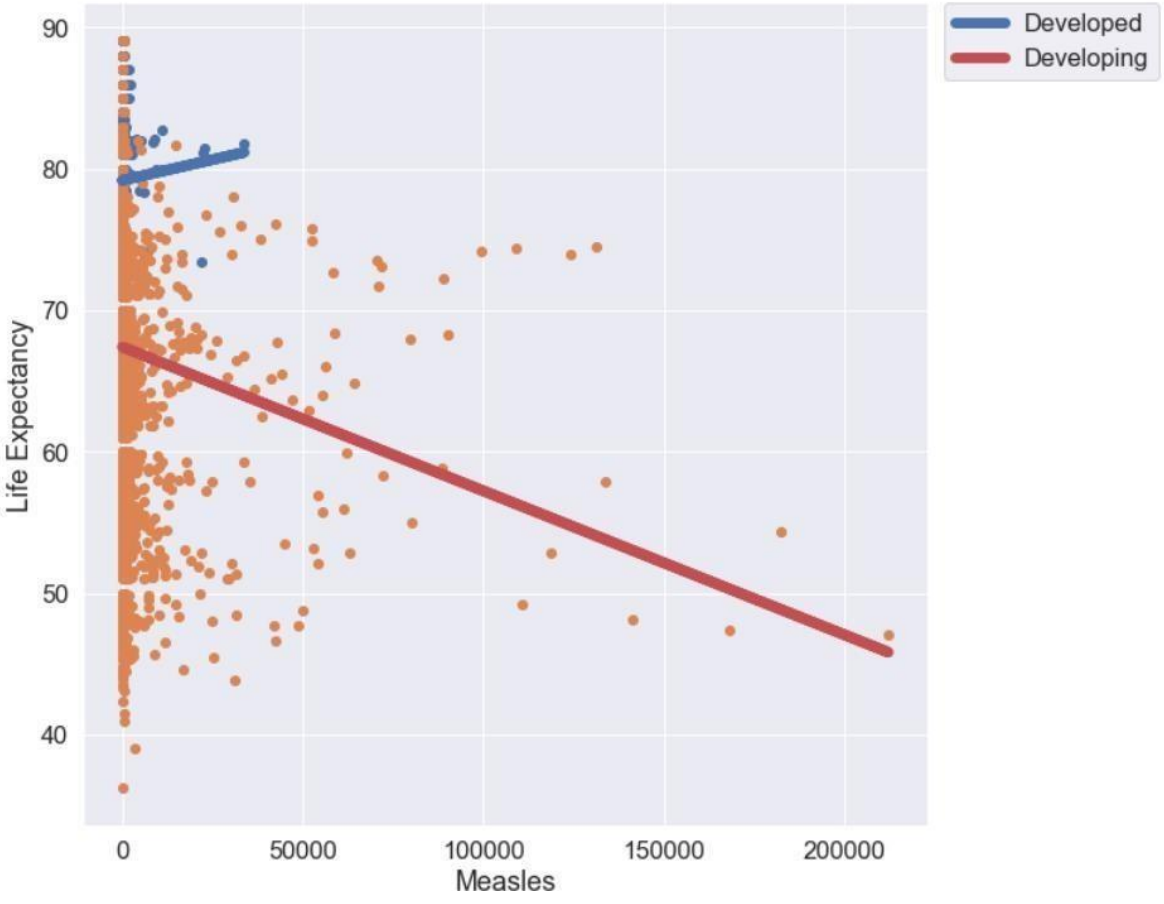


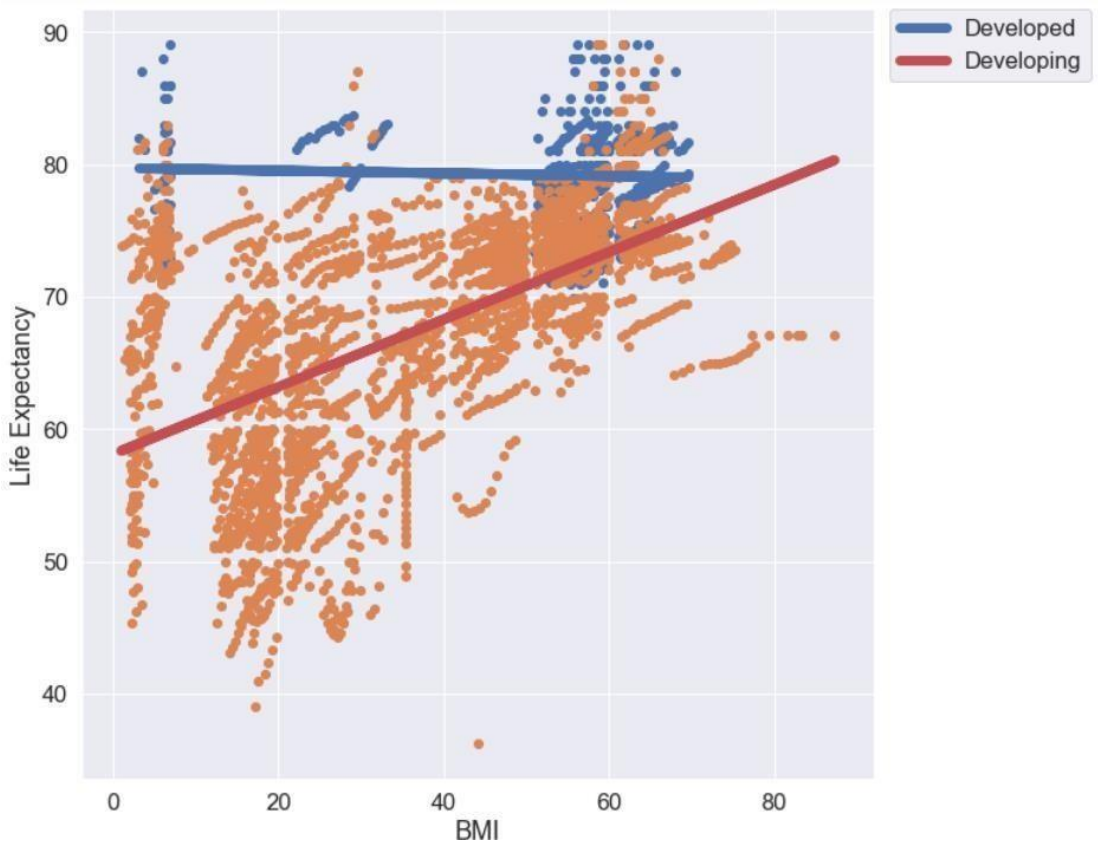


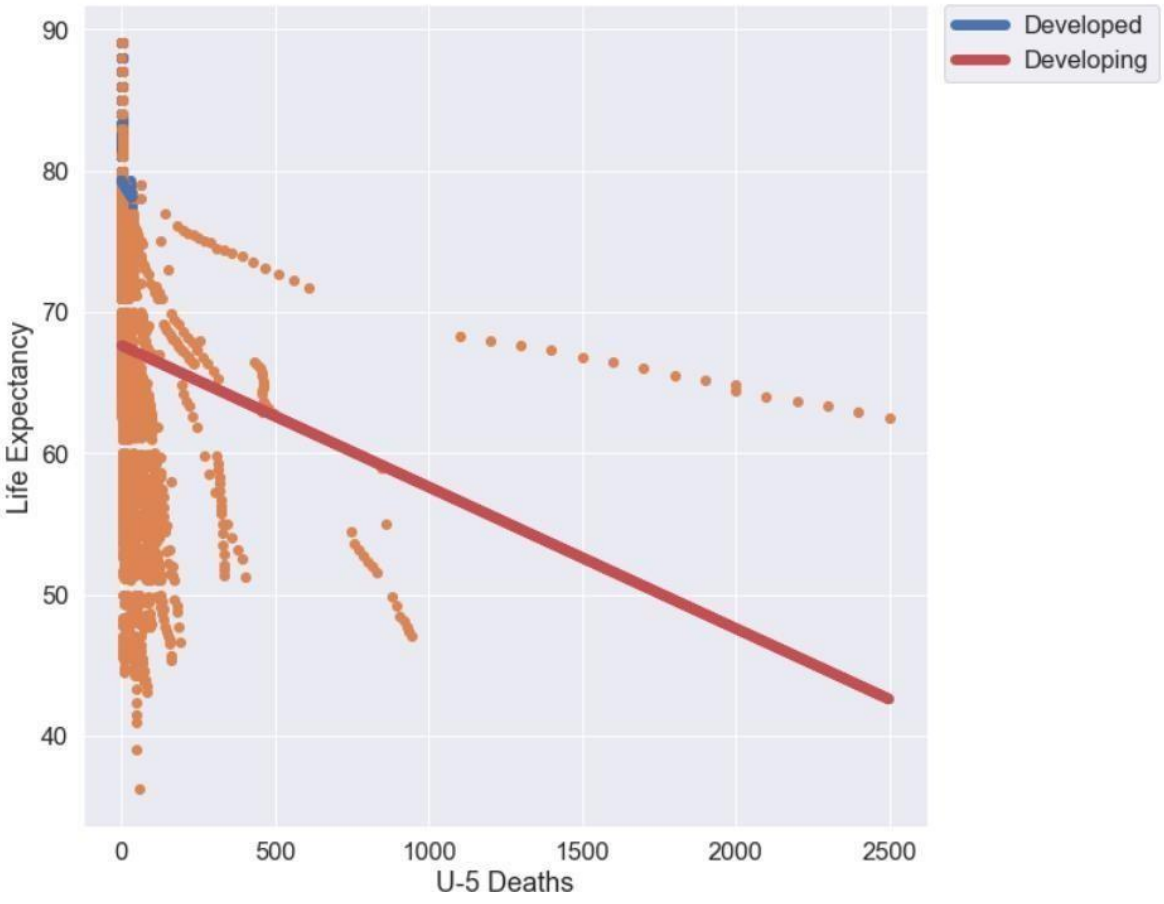

Developed
Developing
Life Expectancy
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Alcohol

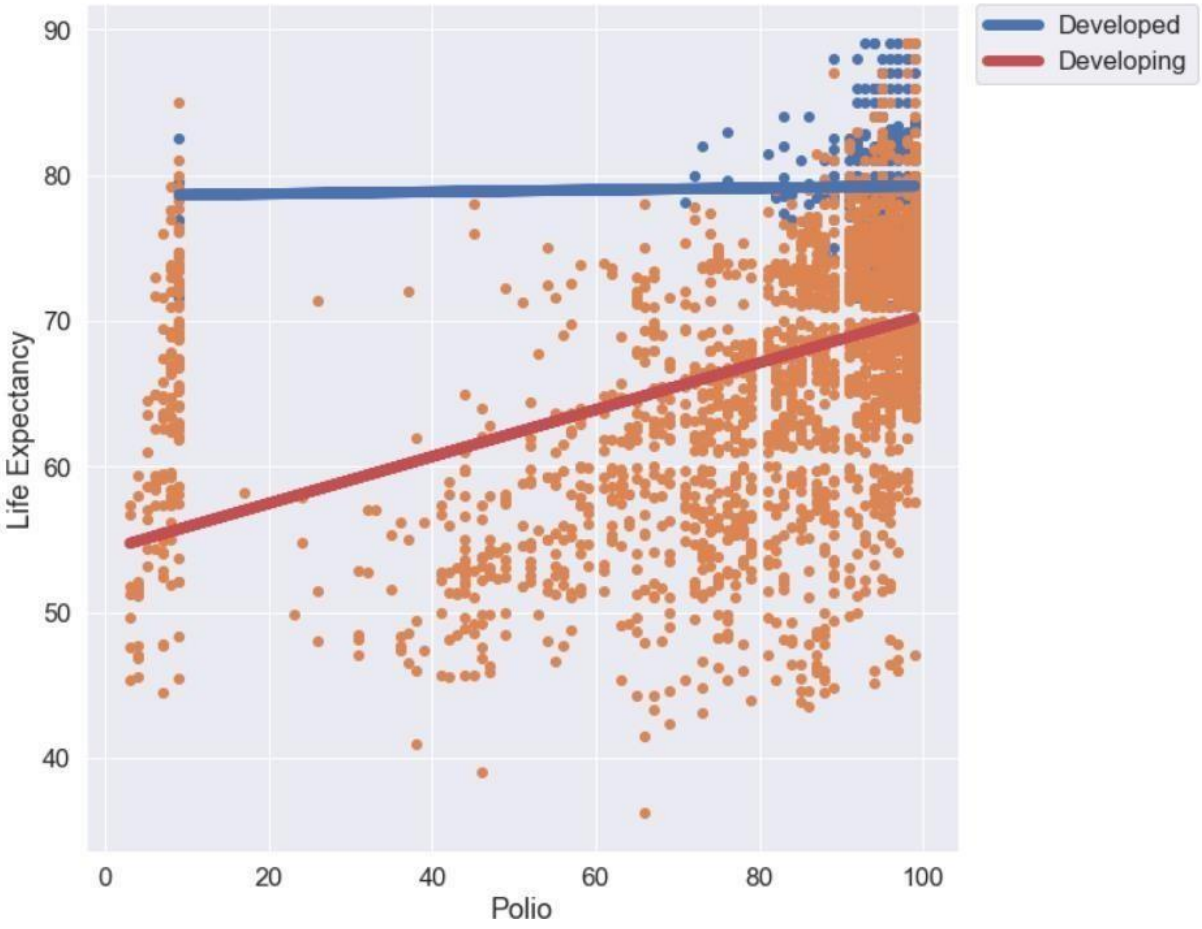

Developed
Developing
Life Expectancy
100
90
80
70
60
50
40
0
2500
5000
7500
10000
12500
15000
17500
20000
Percent Expenditure

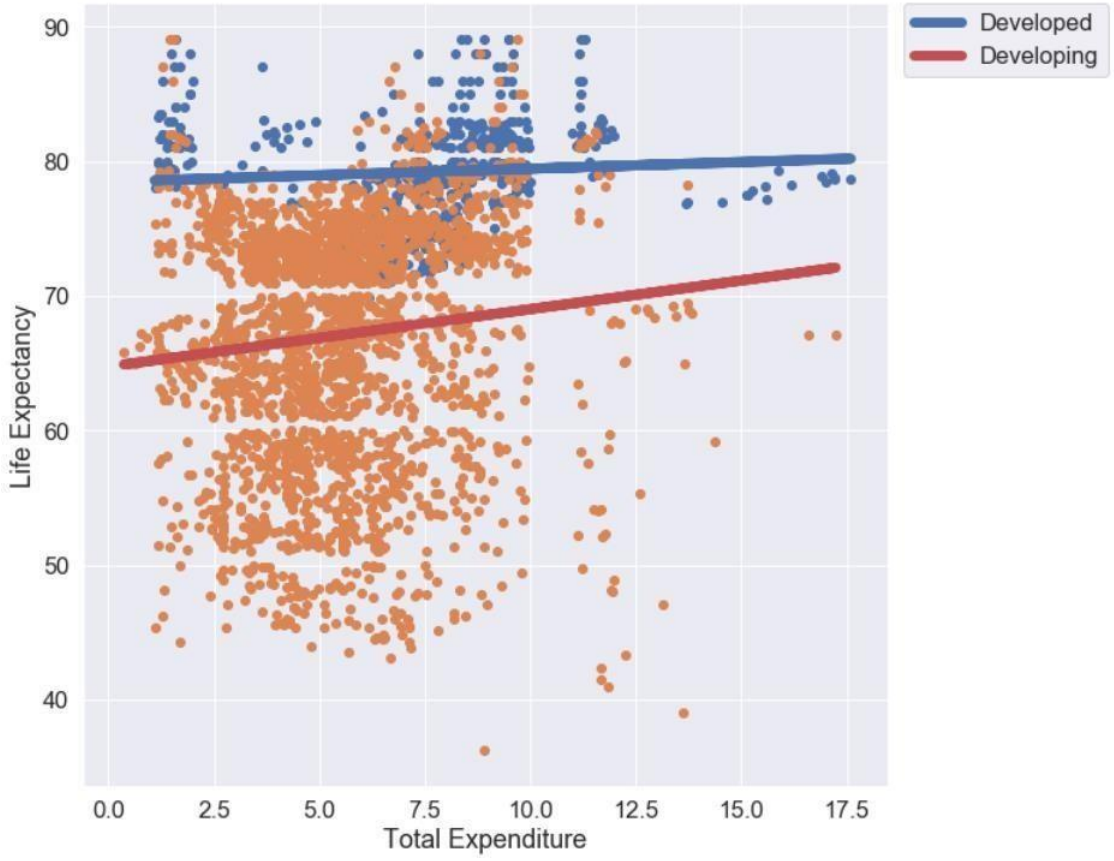

Developed
Developing
Life Expectancy
Hep B
90
80
70
60
50
40
0
20
40
60
80
100

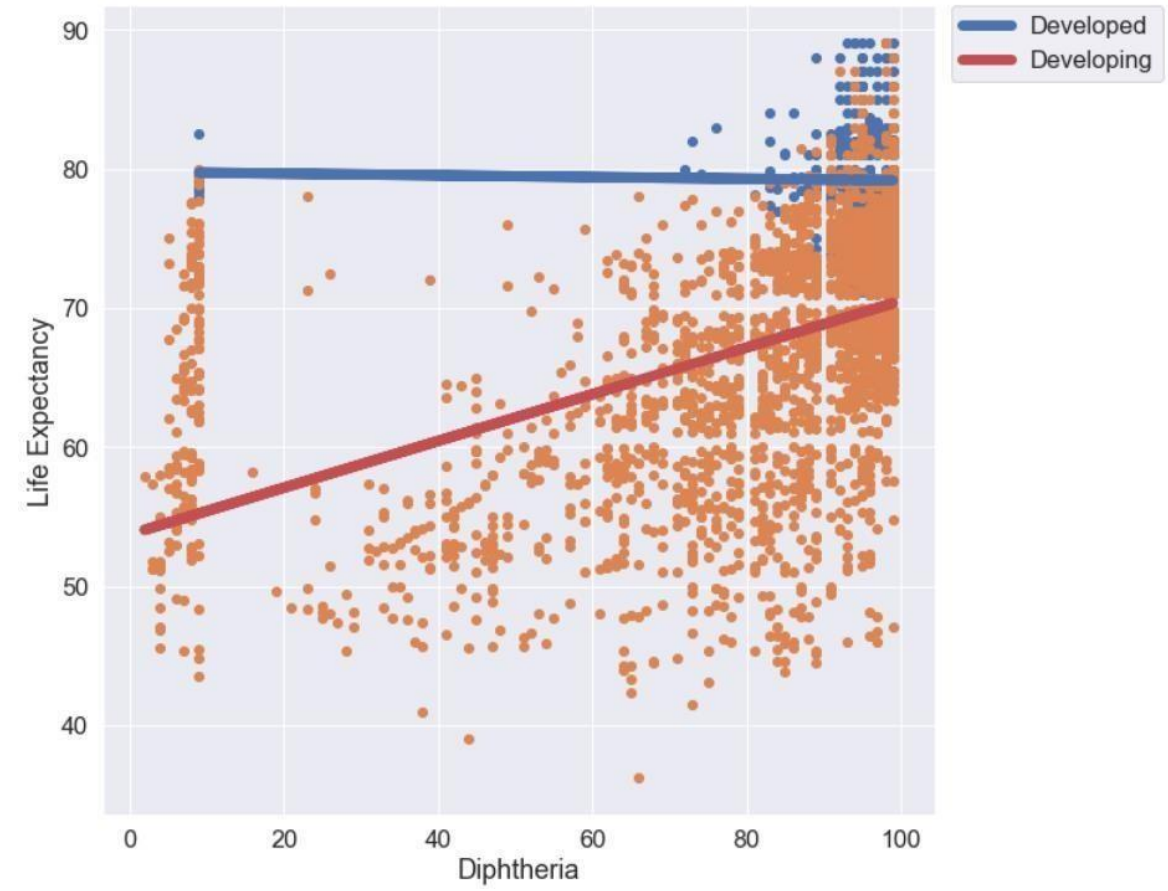

Developed
Developing
Life Expectancy
Measles
90
80
70
60
50
40
0
50000
100000
150000
200000

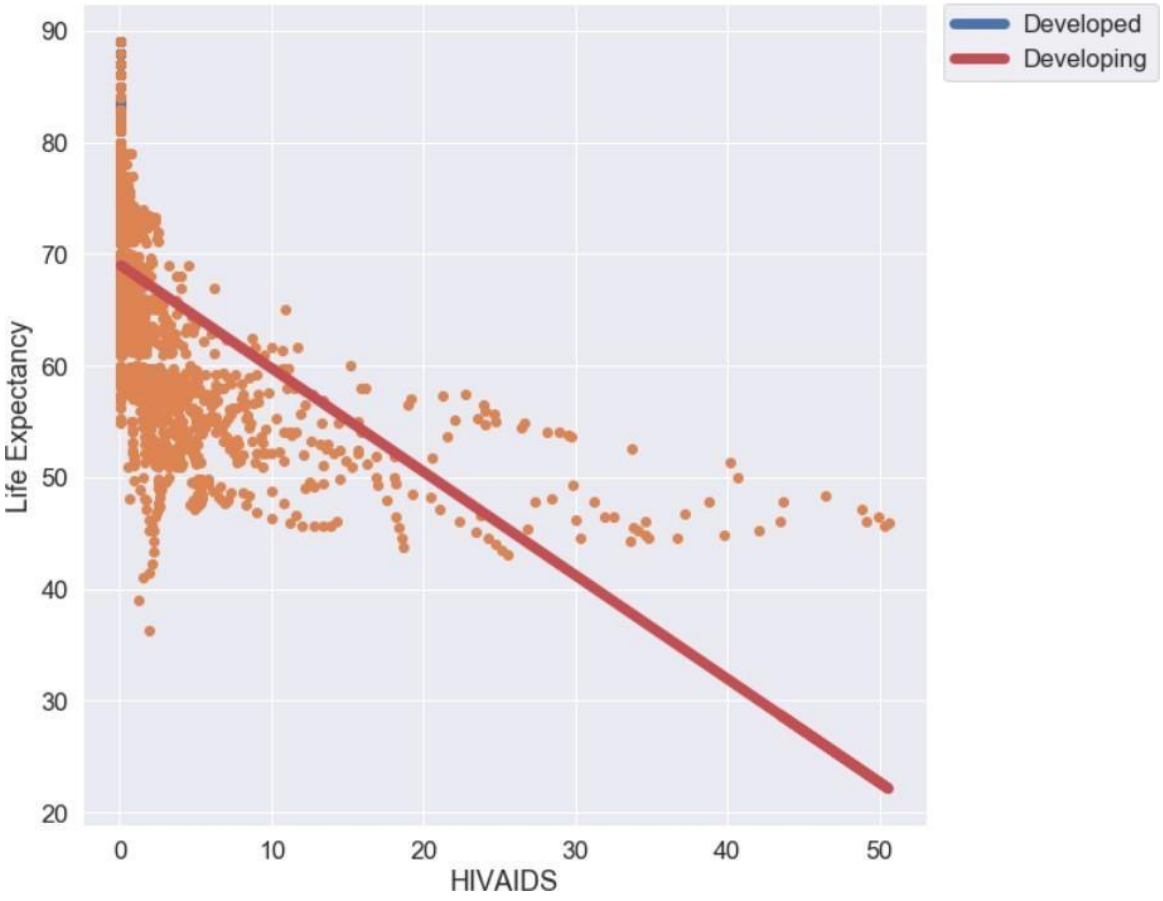

Developed
Developing
Life Expectancy
BMI
90
80
70
60
50
40
0
20
40
60
80

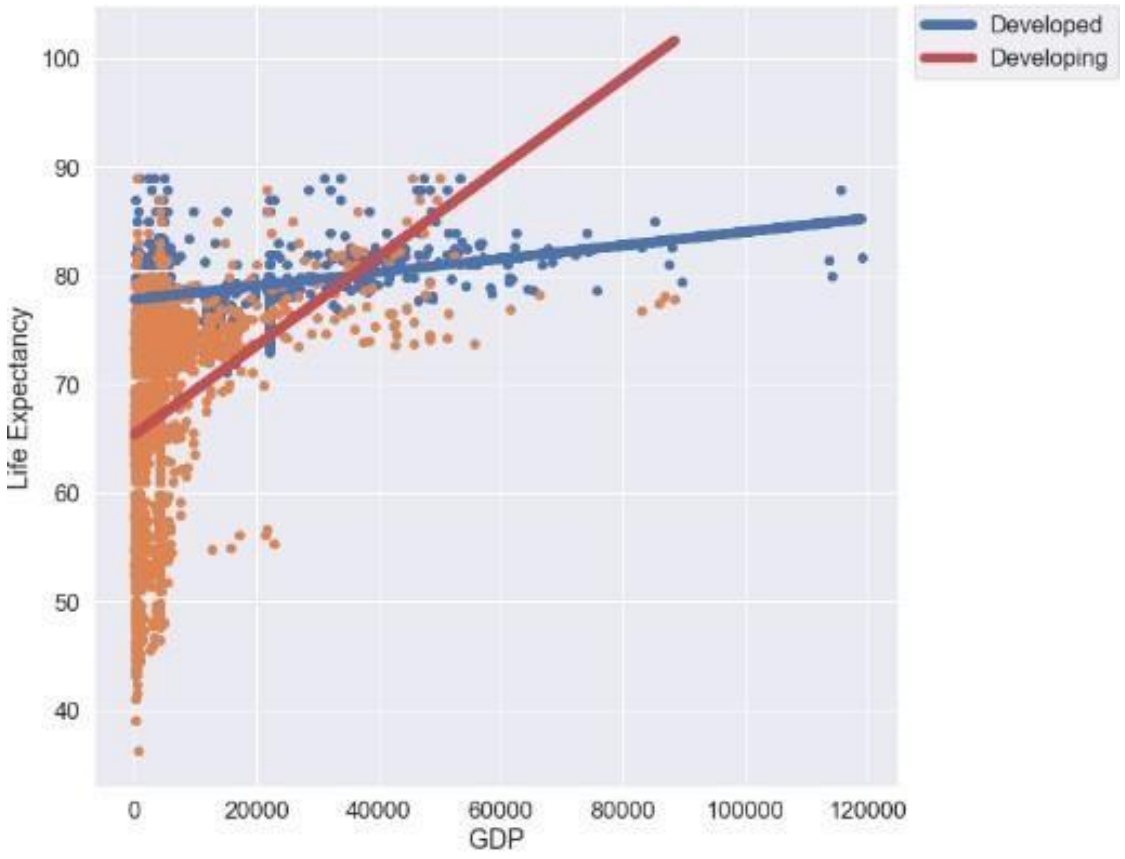

Developed
Developing
Life Expectancy
U-5 Deaths
90
80
70
60
50
40
0
500
1000
1500
2000
2500

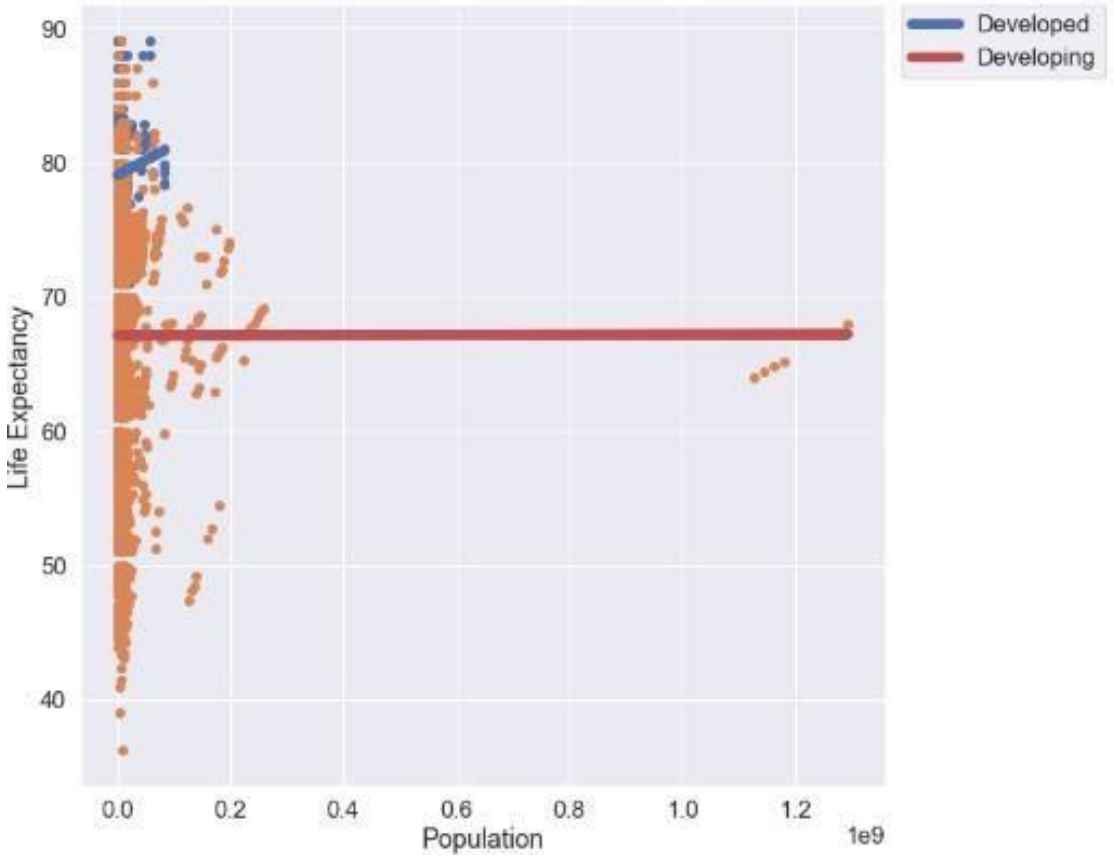

Developed
Developing
Life Expectancy
Polio
90
80
70
60
50
40
0
20
40
60
80
100

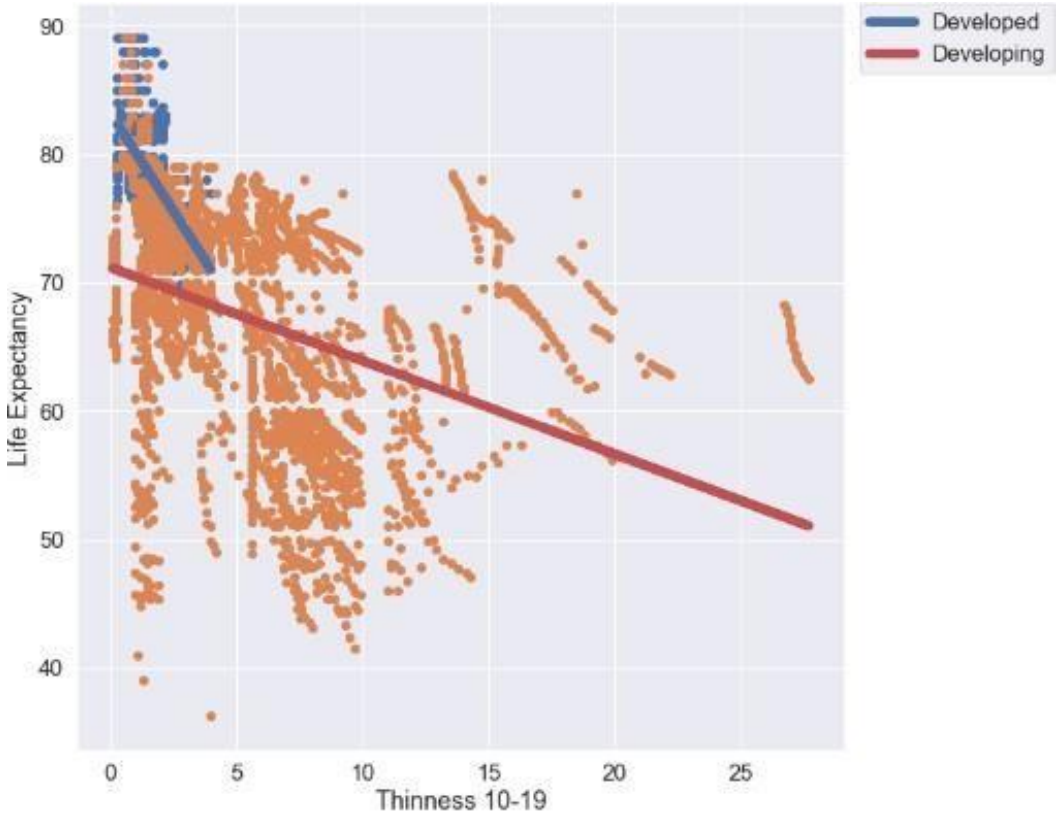

Developed
Developing
Life Expectancy
Total Expenditure
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

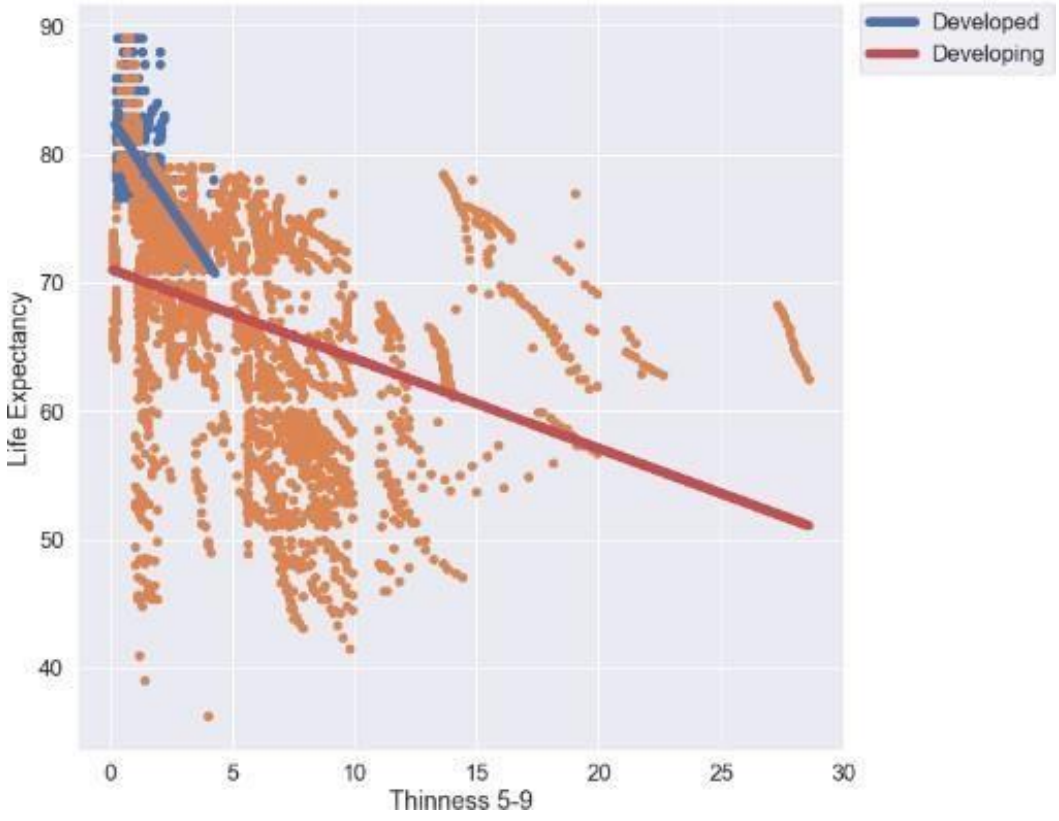

Developed
Developing
Life Expectancy
Diphtheria
90
80
70
60
50
40
0
20
40
60
80
100

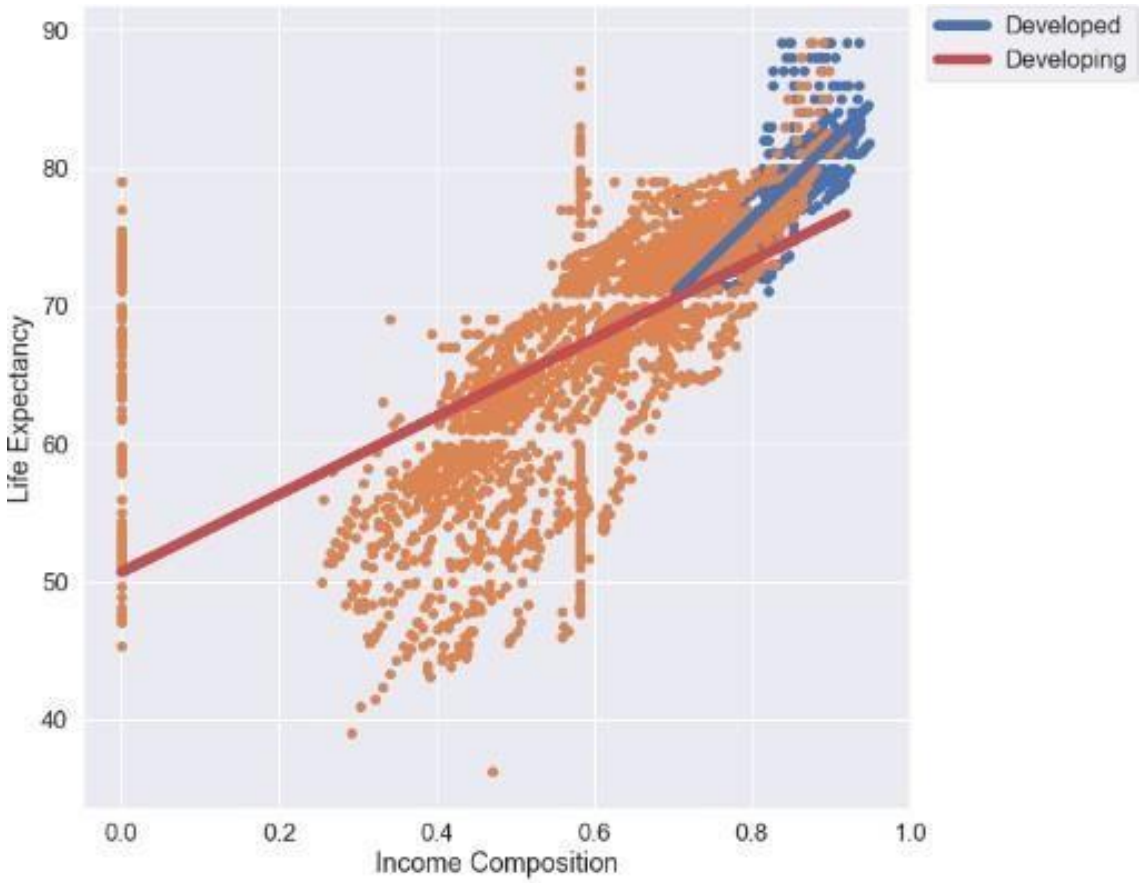

Developed
Developing
Life Expectancy
HIVAIDS
90
80
70
60
50
40
30
20
0
10
20
30
40
50

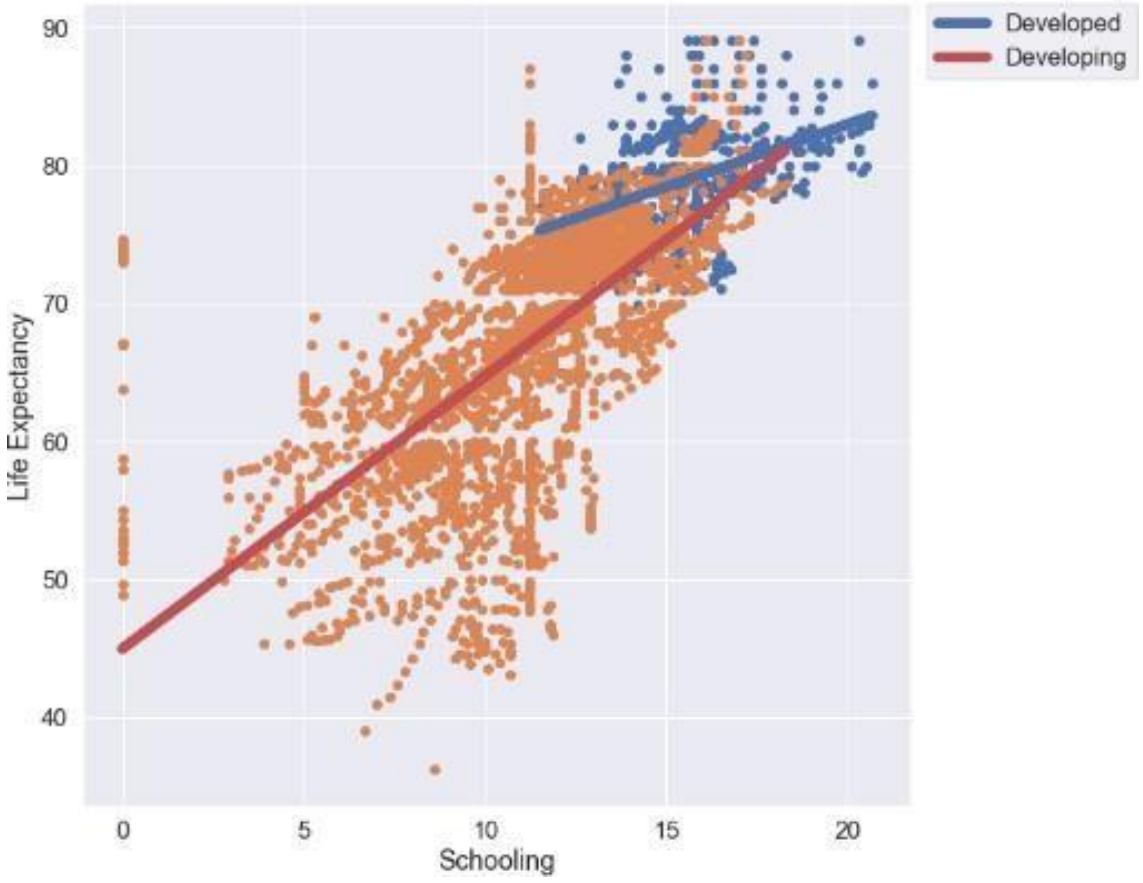

Developed
Developing
Life Expectancy
100
90
80
70
60
50
40
0
20000
40000
60000
80000
100000
120000
GDP


Developed
Developing
Life Expectancy
90
80
70
60
50
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e9
Population


Developed
Developing
Life Expectancy
90
80
70
60
50
40
0
5
10
15
20
25
Thinness 10-19

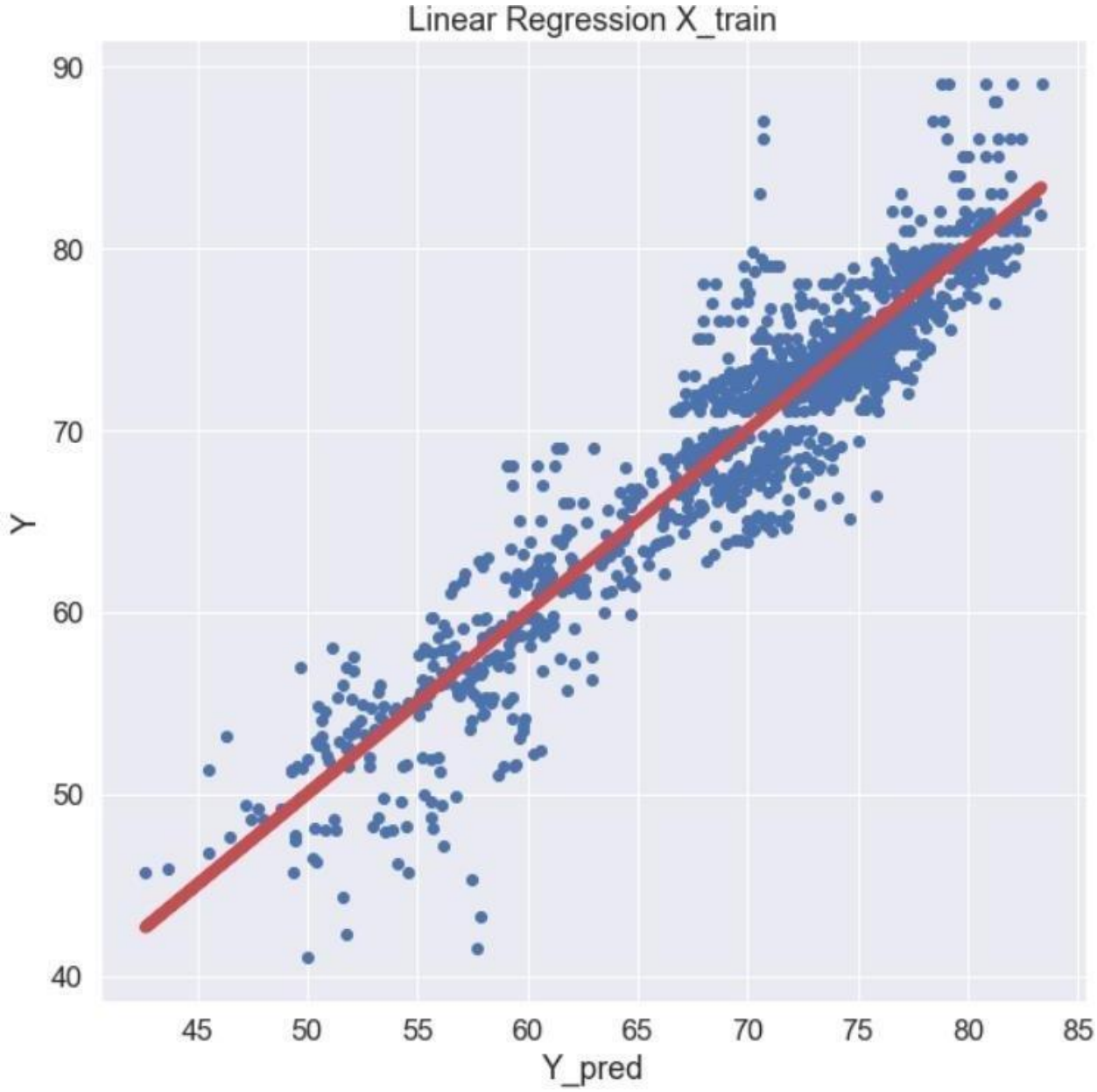

Developed
Developing
Life Expectancy
90
80
70
60
50
40
0
5
10
15
20
25
30
Thinness 5-9


Developed
Developing
Life Expectancy
90
80
70
60
50
40
0.0
0.2
0.4
0.6
0.8
1.0
Income Composition


Developed
Developing
Life Expectancy
90
80
70
60
50
40
0
5
10
15
20
Schooling

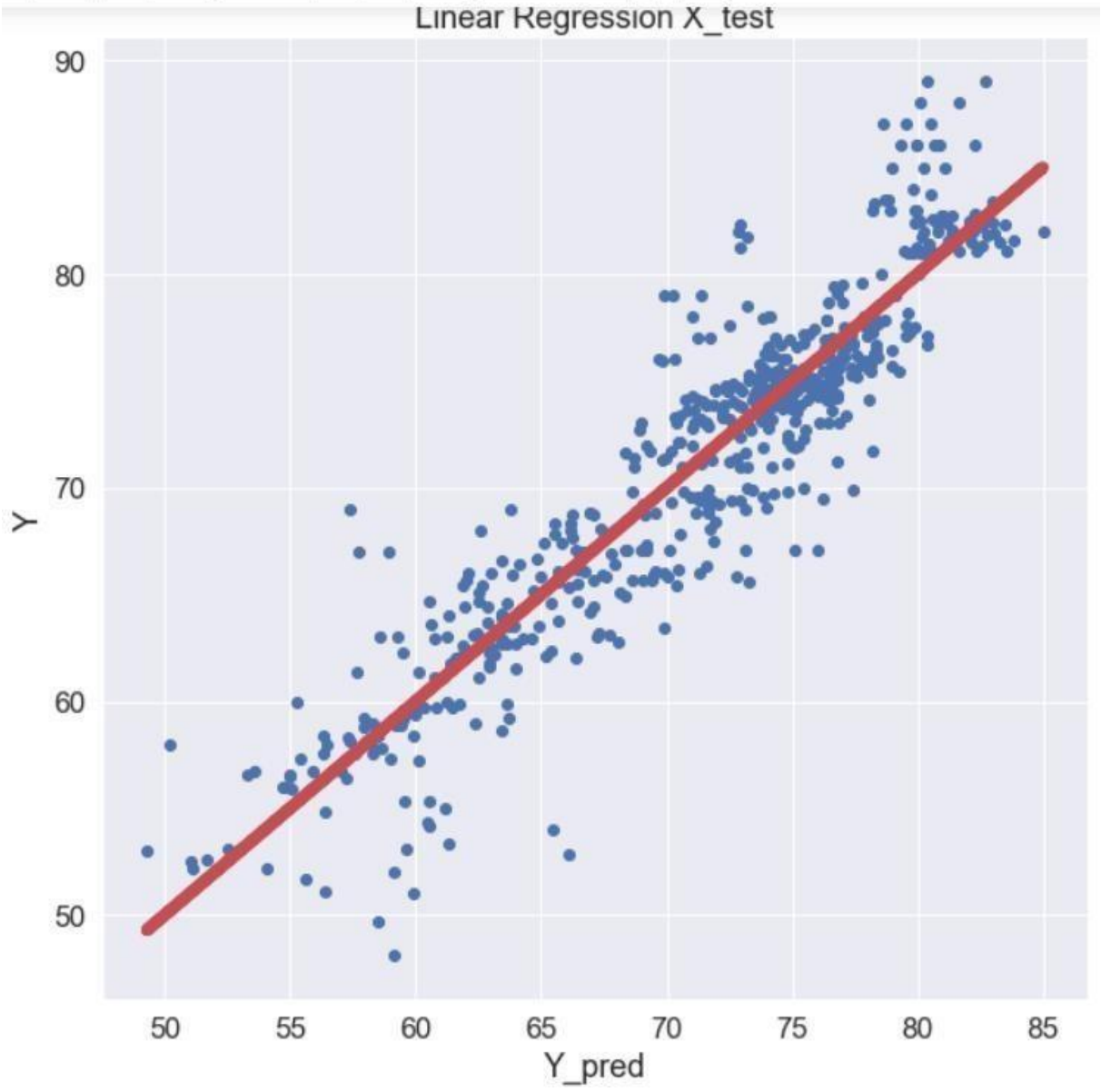
Model Outputs: Linear Regression

```
Linear Regression X_train
X_train (1350, 18)
R^2 Score:0.8725
RMSE: 3.063
Minimum LE: 42.7
Maximum LE: 83.3
Average Predicted LE: 69.9
LE Standard Deviation: 8.012
LE Variance: 64.189
```



On training set – Our model seems to fit well with an accuracy of about 87%

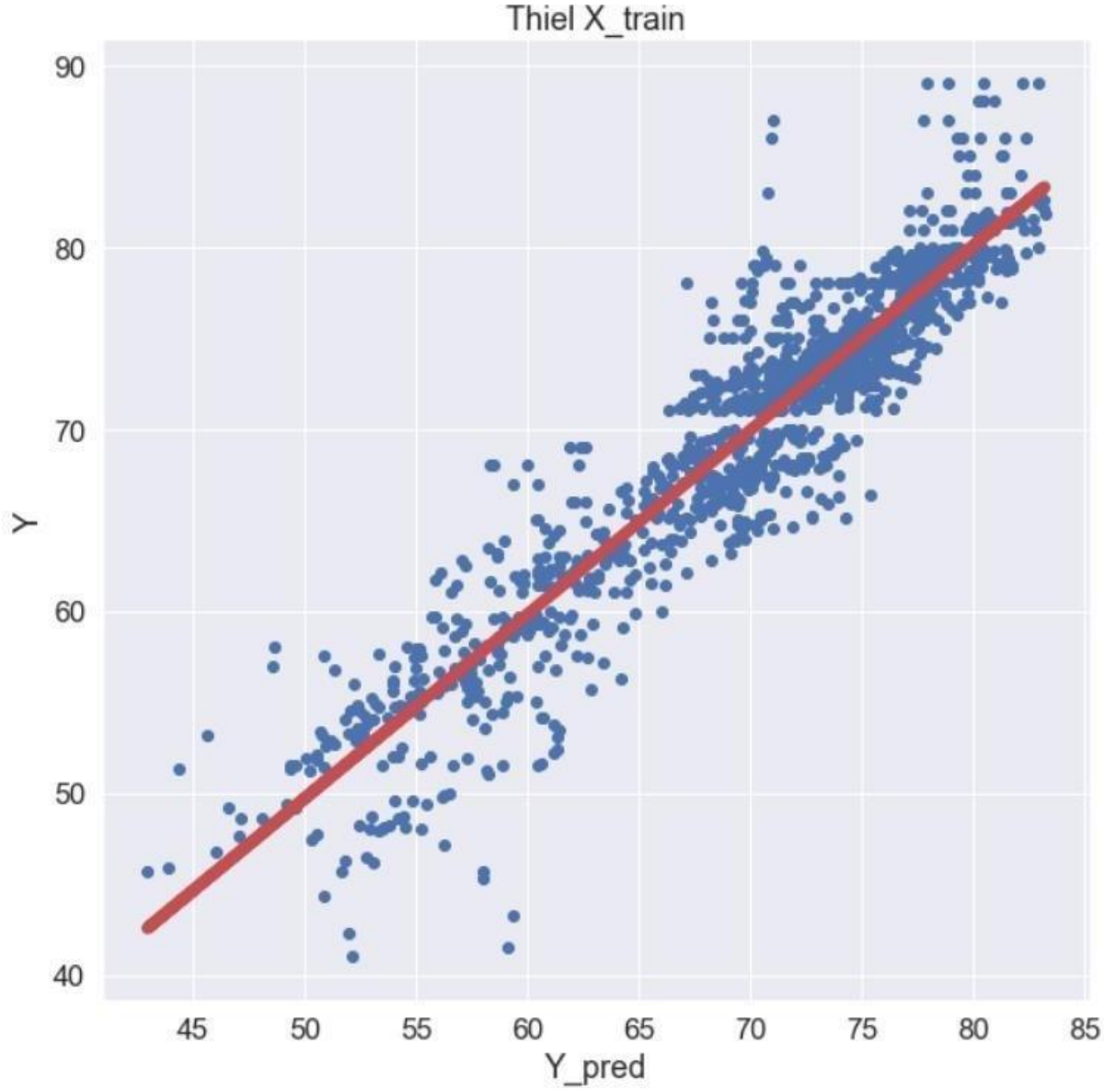
```
Linear Regression X_test
X_test (575, 18)
R^2 Score:0.8500
RMSE: 3.102
Minimum LE: 49.3
Maximum LE: 85.0
Average Predicted LE: 71.3
LE Standard Deviation: 7.384
LE Variance: 54.516
```



On testing set - We see that the Training and Testing accuracies are very similar, hence our model is Fitting properly

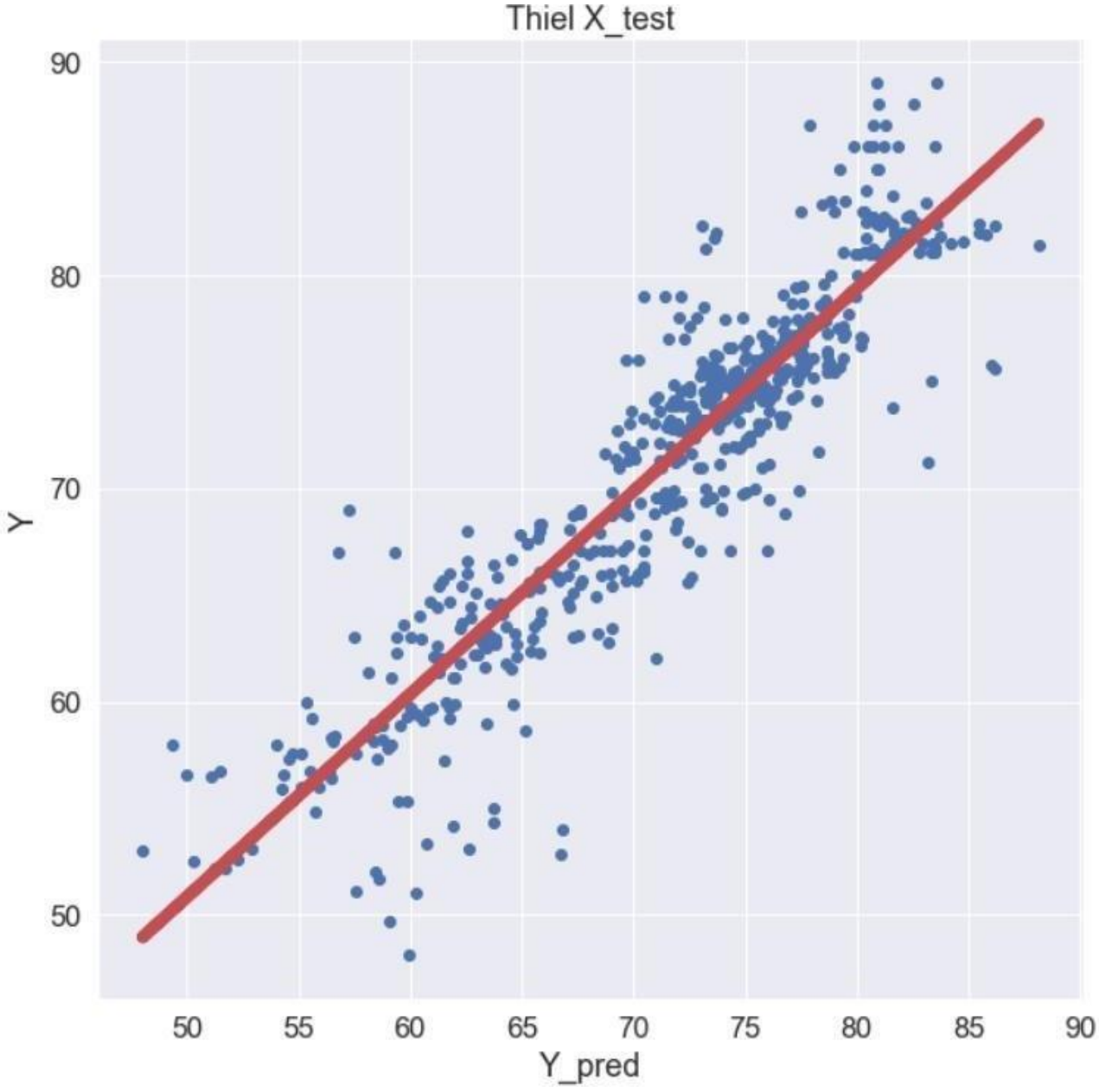
Model Outputs: Theil Sen Regression - Theil–Sen estimator is a method for robustly fitting a line to sample points in the plane by choosing the median of the slopes of all lines through pairs of points.

```
Thiel X_train
X_train (1350, 18)
R^2 Score:0.8652
RMSE: 3.150
Minimum LE: 43.0
Maximum LE: 83.2
Average Predicted LE: 70.0
LE Standard Deviation: 7.880
LE Variance: 62.101
```



On training set

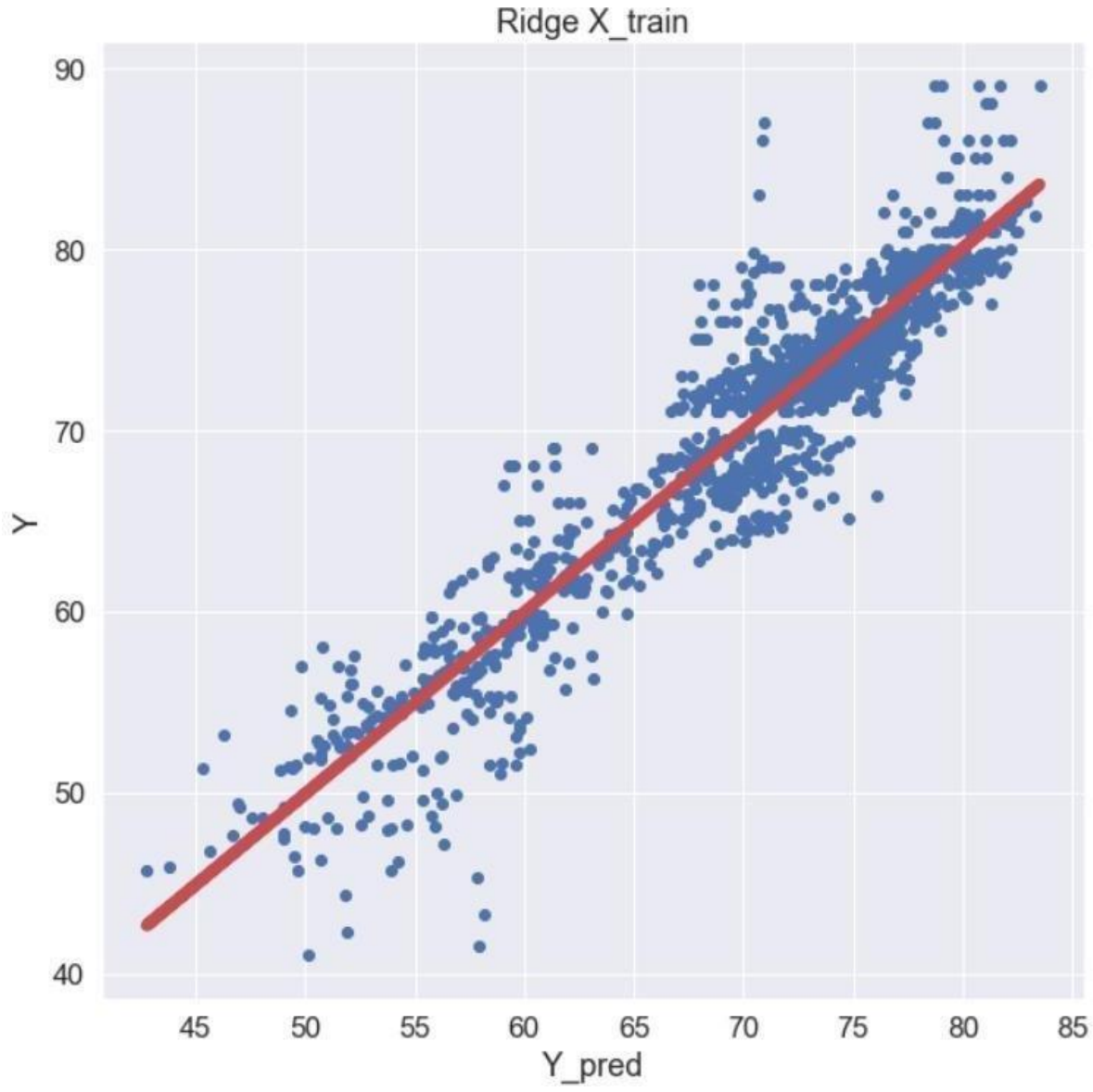
```
Thiel X_test
X_test (575, 18)
R^2 Score:0.8262
RMSE: 3.338
Minimum LE: 48.0
Maximum LE: 88.1
Average Predicted LE: 71.6
LE Standard Deviation: 7.666
LE Variance: 58.767
```



On testing set - Here as well, the training and testing accuracies are very similar, so there is no question of overfitting or underfitting

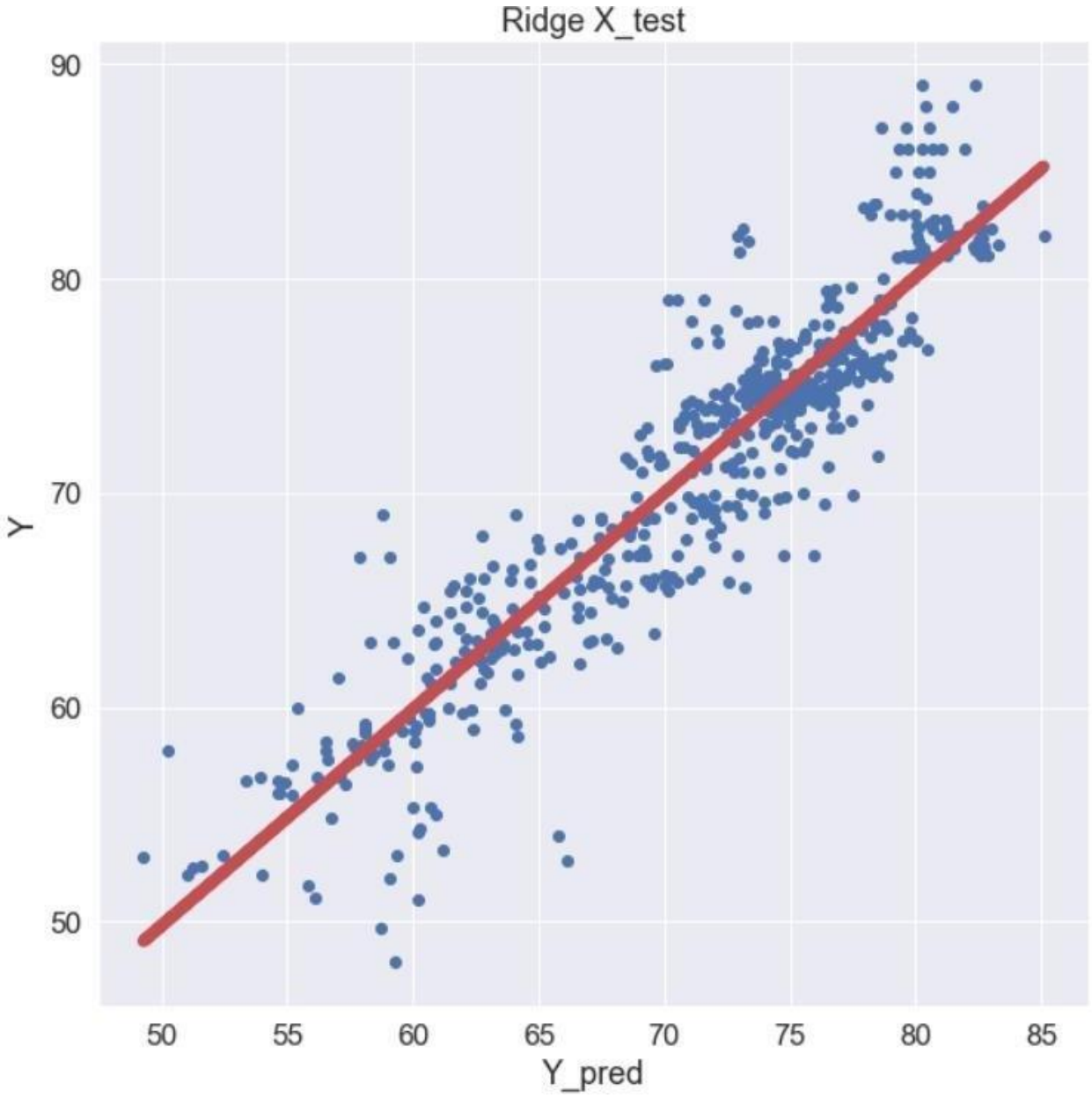
Model Outputs: Ridge Regression: Tikhonov regularization, named for Andrey Tikhonov, is a method of regularization of ill-posed problems. Also known as ridge regression, it is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters.

```
Ridge X_train
X_train (1350, 20)
R^2 Score: 0.8741
RMSE: 3.043
Minimum LE: 42.8
Maximum LE: 83.5
Average Predicted LE: 69.9
LE Standard Deviation: 7.981
LE Variance: 63.693
```



On Training Set - We first use gridsearch to find the best parameters for our Ridge Model

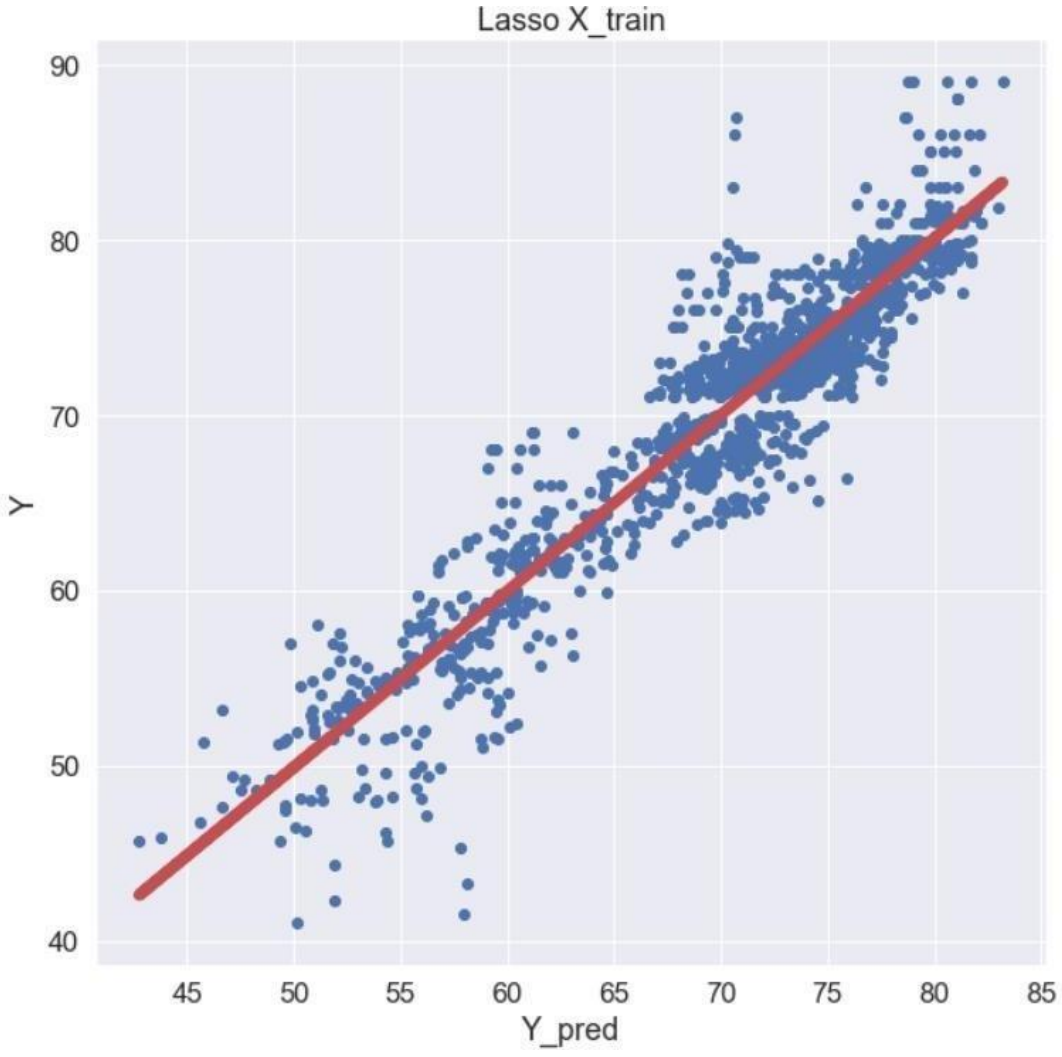
```
Ridge X_test
X_test (575, 20)
R^2 Score: 0.8505
RMSE: 3.096
Minimum LE: 49.3
Maximum LE: 85.1
Average Predicted LE: 71.3
LE Standard Deviation: 7.323
LE Variance: 53.629
```



On Test Set – After the parameters have been found and defined, we run the function of train and test sets

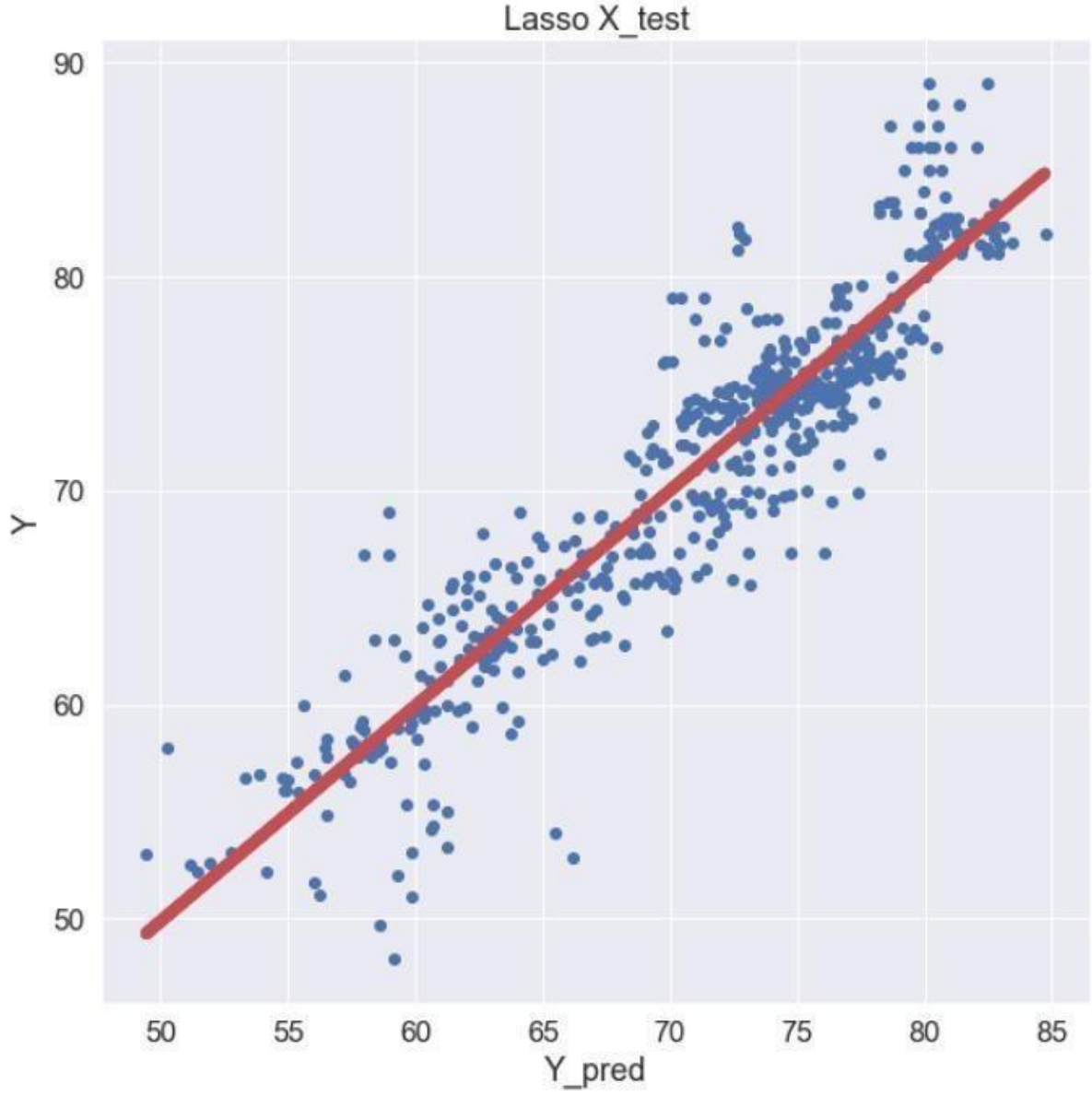
Model Outputs: Lasso Regression - Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when we want to automate certain parts of model selection, like variable selection/parameter elimination. LASSO stands for Least Absolute Shrinkage and Selection Operator

```
Lasso X_train
X_train (1350, 20)
R^2 Score: 0.8726
RMSE: 3.061
Minimum LE: 42.8
Maximum LE: 83.2
Average Predicted LE: 69.9
LE Standard Deviation: 7.967
LE Variance: 63.467
```



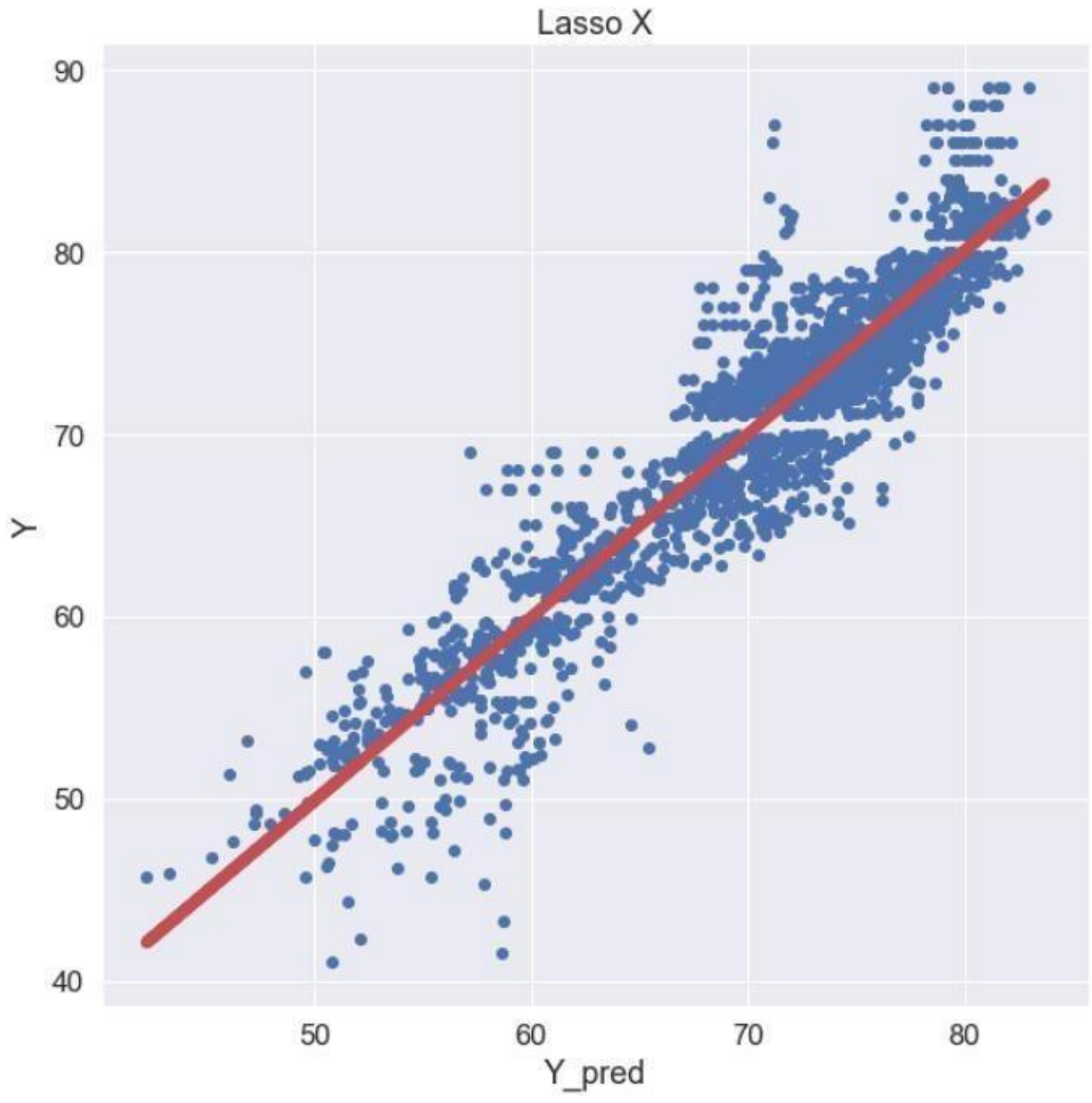
On Training Set

```
Lasso X_test
X_test (575, 20)
R^2 Score: 0.8505
RMSE: 3.097
Minimum LE: 49.4
Maximum LE: 84.7
Average Predicted LE: 71.3
LE Standard Deviation: 7.343
LE Variance: 53.919
```



On test set - We see that all the models we have implemented till now are perfectly fitting as the train and test accuracies are in the +-5% range

```
Lasso X
X (2063, 20)
R^2 Score: 0.8652
RMSE: 3.098
Minimum LE: 42.2
Maximum LE: 83.6
Average Predicted LE: 70.4
LE Standard Deviation: 7.806
LE Variance: 60.932
```



On entire dataset - As we know that ridge regression can't zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficients of collinear variables. Here it helps to select the variable(s) out of given n variables while performing lasso regression.

Model outputs: Gradient Boosting - Gradient boosting involves three elements:

```
A loss function to be optimized.
A weak learner to make predictions.
An additive model to add weak learners to minimize the loss function.
```

1. Loss Function

The loss function used depends on the type of problem being solved.

It must be differentiable, but many standard loss functions are supported and you can define your own.

For example, regression may use a squared error and classification may use logarithmic loss.

A benefit of the gradient boosting framework is that a new boosting algorithm does not have to be derived for each loss function that may want to be used, instead, it is a generic enough framework that any differentiable loss function can be used.

2. Weak Learner

Decision trees are used as the weak learner in gradient boosting.

Specifically regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and "correct" the residuals in the predictions.

Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss.

Initially, such as in the case of AdaBoost, very short decision trees were used that only had a single split, called a decision stump. Larger trees can be used generally with 4-to-8 levels.

It is common to constrain the weak learners in specific ways, such as a maximum number of layers, nodes, splits or leaf nodes.

This is to ensure that the learners remain weak, but can still be constructed in a greedy manner.

3. Additive Model

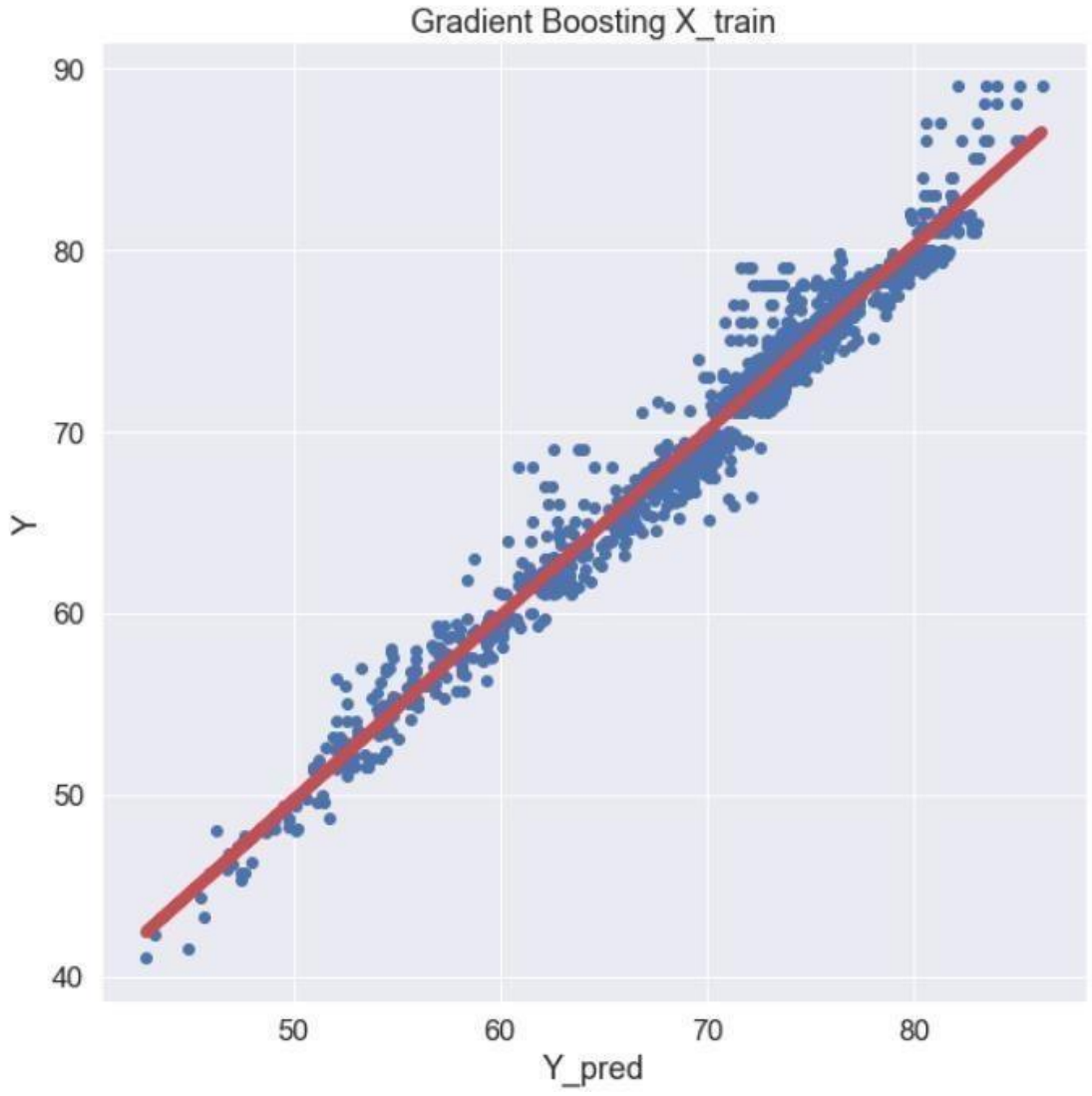
Trees are added one at a time, and existing trees in the model are not changed.

A gradient descent procedure is used to minimize the loss when adding trees.

Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error.

Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure, we must add a tree to the model that reduces the loss (i.e. follow the gradient). We do this by parameterizing the tree, then modify the parameters of the tree and move in the right direction by (reducing the residual loss.

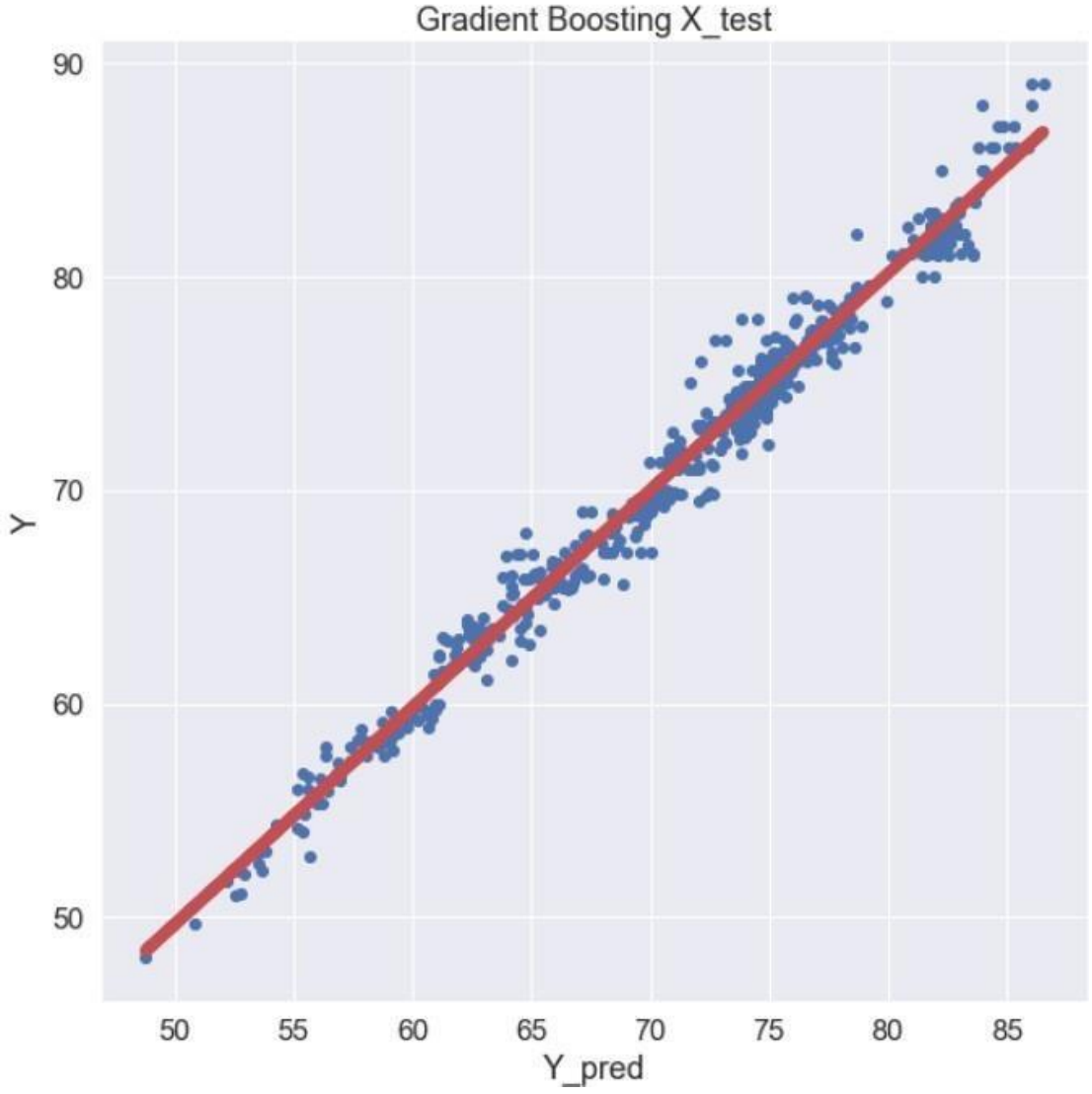
```
Gradient Boosting X_train
X_train (1350, 20)
R^2 Score: 0.9677
RMSE: 1.540
Minimum LE: 42.9
Maximum LE: 86.2
Average Predicted LE: 69.9
LE Standard Deviation: 8.291
LE Variance: 68.733
```




```
Top 5 Features
 IncomeComposition     55.20
TotalExpenditure      20.39
AdultMortality        13.31
Thinness59             2.43
BMI                    2.10
Name: X_train, dtype: float64
```

On training set - A fixed number of trees are added or training stops once loss reaches an acceptable level or no longer improves on an external validation dataset. Neha now tells you to try Gradient boosting, which is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. This model mostly gives the best accuracy

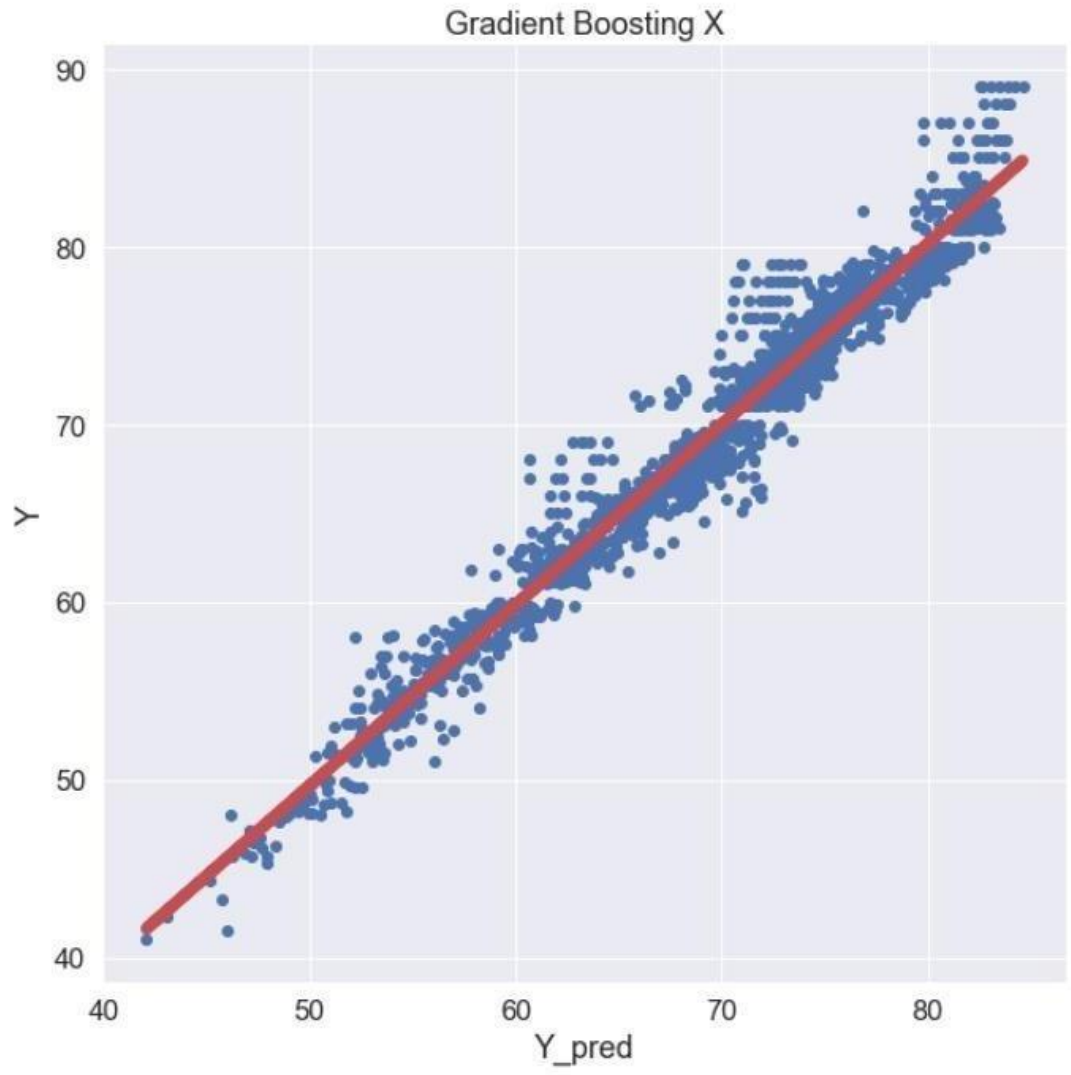
```
Gradient Boosting X_test
X_test (575, 20)
R^2 Score: 0.9813
RMSE: 1.094
Minimum LE: 48.8
Maximum LE: 86.5
Average Predicted LE: 71.3
LE Standard Deviation: 7.809
LE Variance: 60.982
```



```
Top 5 Features
 IncomeComposition     38.32
TotalExpenditure      32.82
AdultMortality         9.84
Schooling              8.00
Thinness59             3.14
Name: X_test, dtype: float64
```

On testing set

```
Gradient Boosting X
X (2063, 20)
R^2 Score: 0.9597
RMSE: 1.693
Minimum LE: 42.1
Maximum LE: 84.6
Average Predicted LE: 70.4
LE Standard Deviation: 8.131
LE Variance: 66.117
```



```
Top 5 Features
 IncomeComposition     57.79
TotalExpenditure      20.02
AdultMortality        11.73
Thinness59             2.50
Measles                1.77
Name: X, dtype: float64
```

On entire dataset - A benefit of using gradient boosting is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function.

We see that The Gradient Boosting Model gives the best accuracy of more than 95%. Hence this is the best linear model for our Dataset

Gradient Boosting is the chosen model for regression due its consistent performance and ability to deal with all types and set of data. It captures the LE range,mean, and standard deviation and allows the user to see which features play a key factor in model performance. Given the lower amount of data for this model, it works well and quickly with more data and factors, it is likely to fall off. This is however very promising given the amount of cleaning present at the beginning.

```
1 feature_importances
```

	X_train	X_test	X
Year	0.05	0.03	0.13
AdultMortality	13.31	9.84	11.73
Infant Deaths	0.43	0.28	0.29
Alcohol	0.66	1.05	0.81
Population	0.28	0.03	0.08
GDP	0.03	0.15	0.01
PercentExpenditure	0.11	0.32	0.10
Hep B	1.19	0.97	0.83
Measles	1.66	0.58	1.77
BMI	2.10	0.51	1.45
U5Deaths	0.22	1.19	0.31
Polio	0.36	0.63	0.76
TotalExpenditure	20.39	32.82	20.02
Diphtheria	0.12	0.24	0.08
HIVAIDS	0.18	0.29	0.07
Thinness1019	0.12	1.12	0.43
Thinness59	2.43	3.14	2.50
IncomeComposition	55.20	38.32	57.79
Schooling	0.34	8.00	0.17
country_code	0.83	0.51	0.65

Feature Importance for Gradient Boosting

CONCLUSION

Gradient Boosting is the chosen model for regression due its consistent performance and ability to deal with all types and set of data. It captures the LE range,mean, and standard deviation and allows the user to see which features play a key factor in model performance. Given the lower amount of data for this model, it works well and quickly with more data and factors, it is likely to fall off. This is however very promising given the amount of cleaning present at the beginning.

Top 5 Average Life Performance Predictors across all the model are Income Composition, Adult Mortality, HIV_AIDS, Schooling, and Thinness 5-9. Life Expectancy ranges are Developed(70-89), Developing (41-86), 1st World(78-88), 2nd World(65-78),3rd World(41-65).

The developed and developing status don't fully cover the different categories of countries. Testing data was generally higher across all groupings due to the year raising LE inherently. In fact, there was a .3 increase every year in Life Expectancy from the models.

Disease and hunger relief are an universal key to improving life expectancy. This can come in the form of just being smarter and preventing outbreaks by vaccinating. This is helpful when traveling to other countries to keep it away from unknowing populations.

There is a strong positive correlation between 'Schooling' and 'Life Expectancy' of 0.73. This may be because education is more established and prevalent in wealthier countries. This means countries with less corruption, infrastructure, healthcare, welfare, and so forth.

Similarly to the point above, there is a moderate positive correlation between 'GDP' and 'Life Expectancy' of 0.44, most likely due to the same reason.

Surprisingly there's a moderate positive correlation between 'Alcohol' and 'Life Expectancy' of 0.40. I'm 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.

Alcohol is negatively correlated with life expectancy for developed countries and 1st world countries while being positively correlated with 2nd, 3rd, and Developing Countries. In this case, it means that population is living longer, but alcohol consumption is not increasing life expectancy.

REFERENCES

Acemoglu, Daron, and Simon Johnson. "Disease and development: the effect of life expectancy on economic growth." *Journal of political Economy* 115.6 (2007): 925-985.

Uyanık, Gülden Kaya, and Neşe Güler. "A study on multiple linear regression analysis." *Procedia-Social and Behavioral Sciences* 106 (2013): 234-240.

Wagner, Harvey M. "Linear programming techniques for regression analysis." *Journal of the American Statistical Association* 54.285 (1959): 206-212.

Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." *R package version 0.4-2* (2015): 1-4.

Escudero, Paola, et al. "Formal modelling of L1 and L2 perceptual learning: Computational linguistics versus machine learning." *Eighth Annual Conference of the International Speech Communication Association*. 2007.

Khan, Alamgir, Salahuddin Khan, and Manzoor Khan. "Factors effecting life expectency in developed and developing countries of the world (An approach to available literature)." *Life* (2010): 3.

Articles:
https://www.statista.com/statistics/274507/life-expectancy-in-industrial-and-developing-countries/

https://medium.com/@nirankari.naveen.13et1136/how-to-predict-life-expectancy-using-machine-learning-5c253ab25125

https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1