Life Expectancy Prediction

Mohd. Rizwan Shaikh
IMT2019513

October 3, 2021

Abstract

The aim of this machine learning assignment is to predict the life expectancy of a person based on certain features. For this, we implement univariate and multivariate linear regression. We implement these models using closed form solution, gradient descent approach and newton's method of optimization.

Contents

1 Introduction

The purpose of this assignment is to predict the life expectancy of a person using various features. We implement both univariate and multivariate linear regression using closed form solution, gradient descent approach and newton's method of optimization.

To train our model, we use 'Life expectancy' dataset which consists of 21 features that represent various factors of over 193 countries over a span of 2000-2015.

2 Initial Data Processing

The dataset consists of about 21 features. We will first look at NaN values and try to get rid of the same. The following table gives the count of NaN values in each feature.

Table 1: Featurewise count of NaN values

Feature	No. of NaN values
Country	0
Year	0
Status	0
Life Expectancy	10
Adult Mortality	10
Infant Deaths	0
Alcohol	194
Percentage Expenditure	0
Hepatitis B	553
Measles	0
BMI	34
Under five deaths	0
Polio	19
Total Expenditure	226
Diphtheria	19
HIV/AIDS	0
GDP	448
Population	652
Thinness 1-19 years	34
Thinness 5-9 years	34
Income Composition of Resources	167
Schooling	163

We drop rows with NaN values in Life expectancy, Adult Mortality, Alcohol, BMI, Polio, Total Expenditure, Diphtheria, Thinness 1-19 years, Thinness 5-9 years, Income

Composition of Resources and Schooling because of comparably less number of NaN values in them.

We drop Country and Status column because these are categorical data.

We drop Hepatitis B, GDP and Population because of large number of NaN values.

There are no duplicate rows and '?' in the data.

3 Exploratory Data Analysis

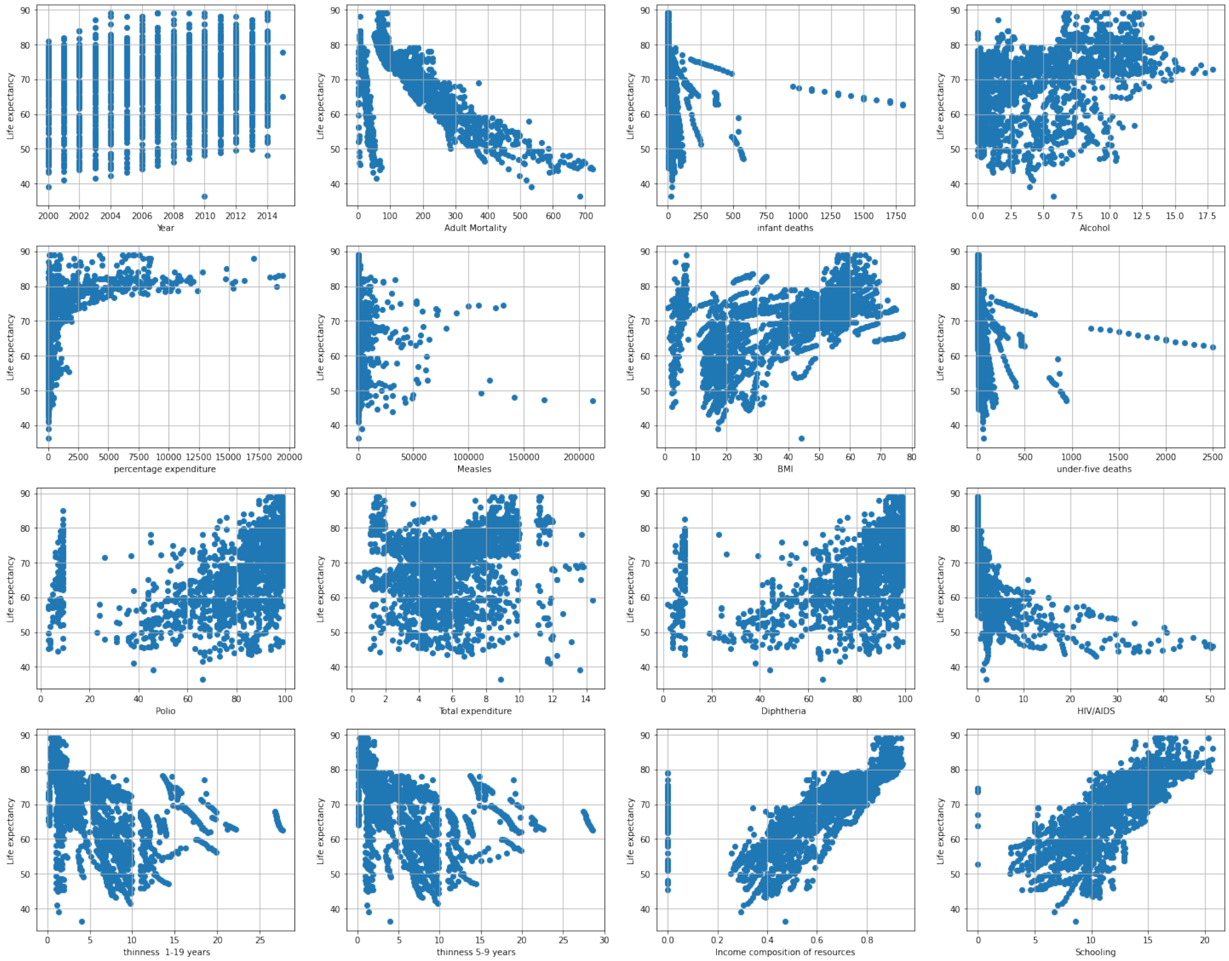

Figure 1: Life expectancy vs feature after initial preprocessing

At first glimpse, we see that there are illogical values in some feature. There are discussed pointwise below:

- Infant deaths is measured per 1000 people. It cannot be more than 1000. Here, there are 13 values above 1000. So we drop rows with infant deaths greater than 1000.

- Percentage expenditure on health as a measure of GDP cannot be more than 100. There are 1292 such values. So we drop this feature.

- Number of Measles cases per 1000 people cannot be more than 1000. Here there are 427 such values. We drop this feature.

- Under five deaths more than 1000 cannot be more than 1000. Here there are 2 such values. So we drop rows with under five deaths more than 2.

- In school, there are some zeros as well. Most probably these are illogical values. There are 10 of them. So we drop rows with school equal to 0.

- In income composition of resources, there are 104 zeros. We will drop rows with 0 income composition of resources.

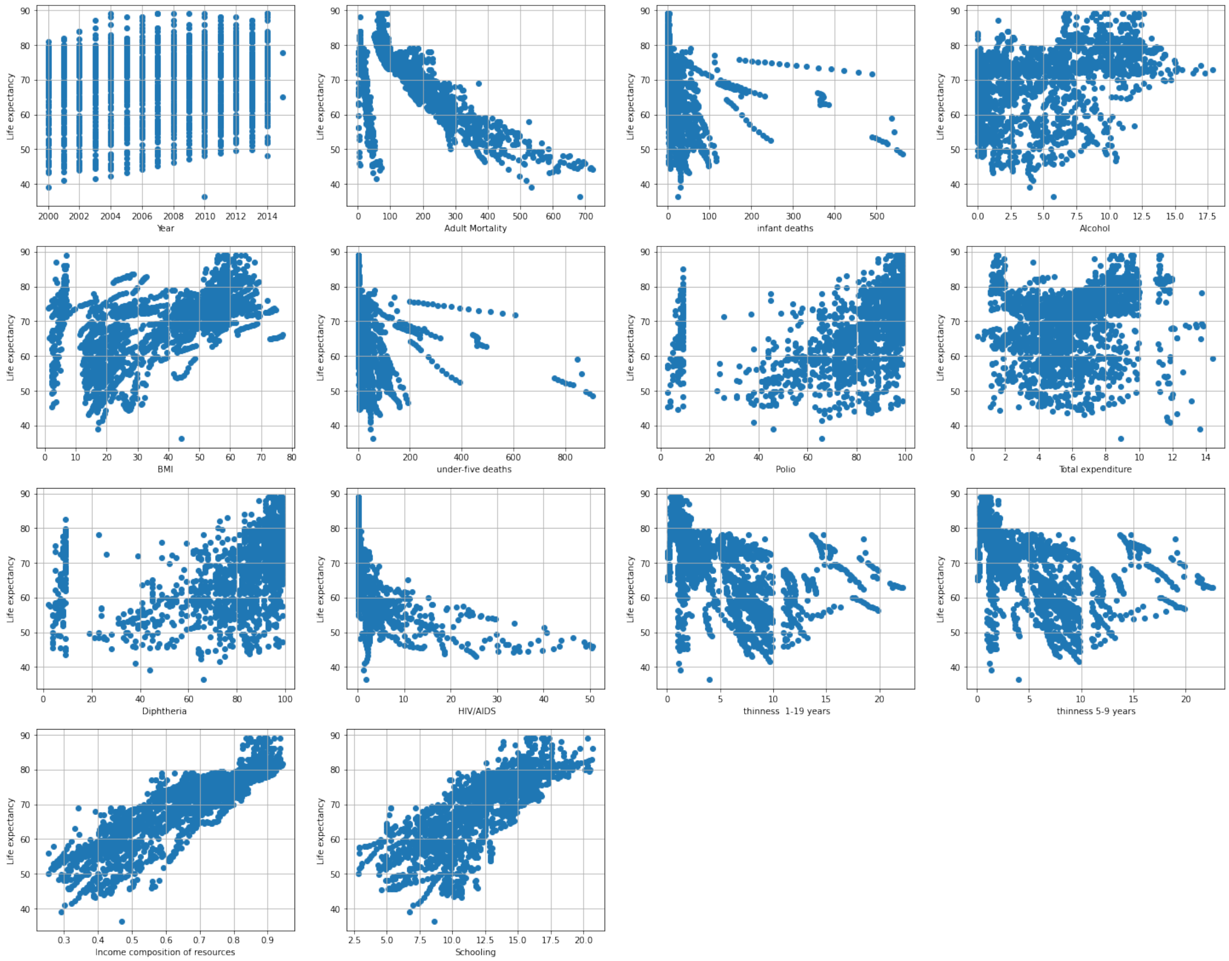

Figure 2: Life expectancy vs feature after removing illogical values

Now, we will see the boxplot to check for outlier values.

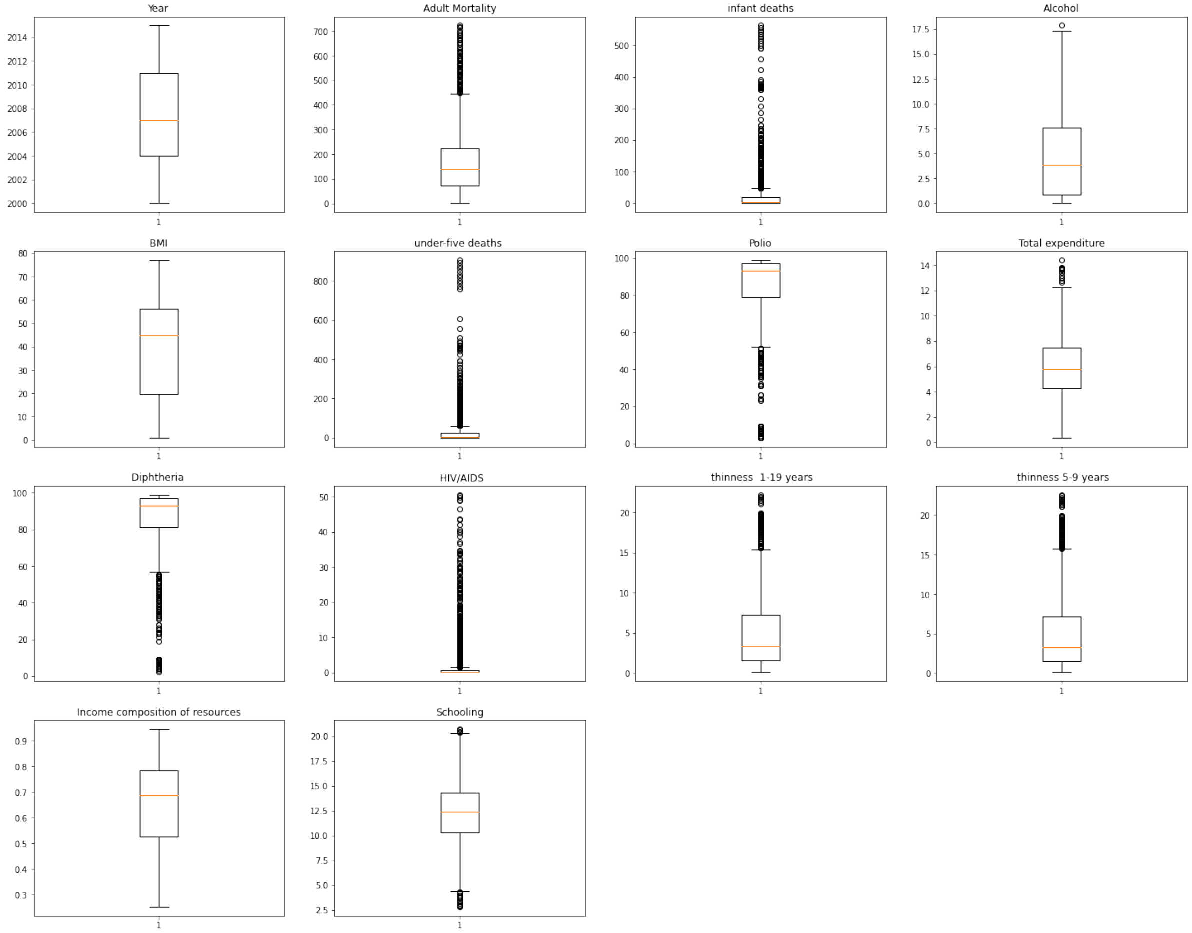

Figure 3: Boxplot of outlier values in the features before removing them

There are decent amount of outlier values in most of the features. The table given below shows the count of outlier values for each feature:

Table 2: Featurewise count of outlier values

Feature	No. outliers	Feature	No. of outliers
Country	0	Year	0
Adult Mortality	80	Infant Deaths	274
Alcohol	1	BMI	0
Under five deaths	321	Polio	218
Total Expenditure	13	Diphtheria	250
HIV/AIDS	477	Thinness 1-19 years	59
Thinness 5-9 years	65	Income Composition of Resources	0
Schooling	23		

The outliers which are above upperlimit are changed to the upperlimit of the data. Similarly, the outliers which are below the lowerlimit are changed to lowerlimit of data.

The boxplot after shifting the outlier values is shown below:

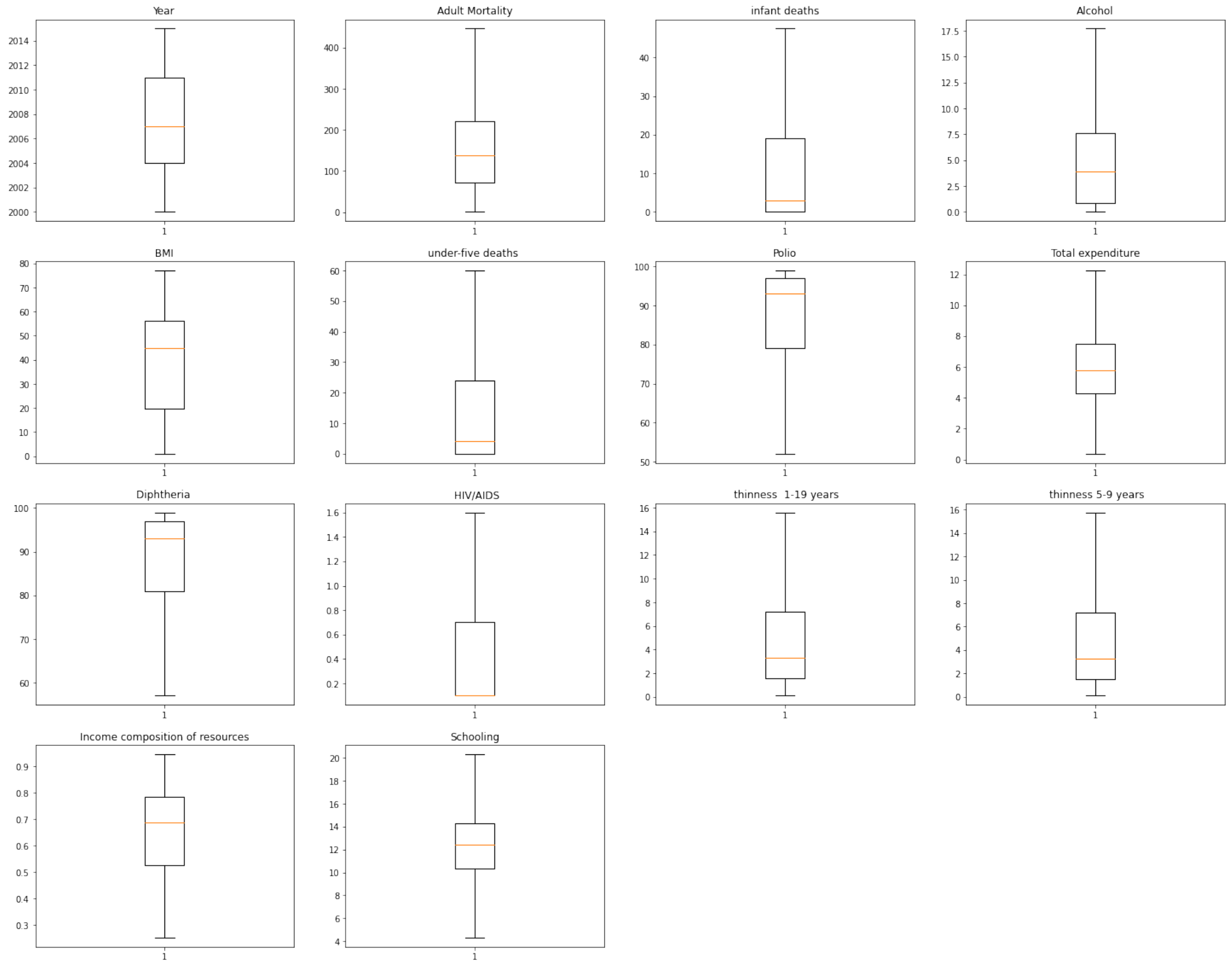

Figure 4: Boxplot of outlier values in the features after shifting the outliers

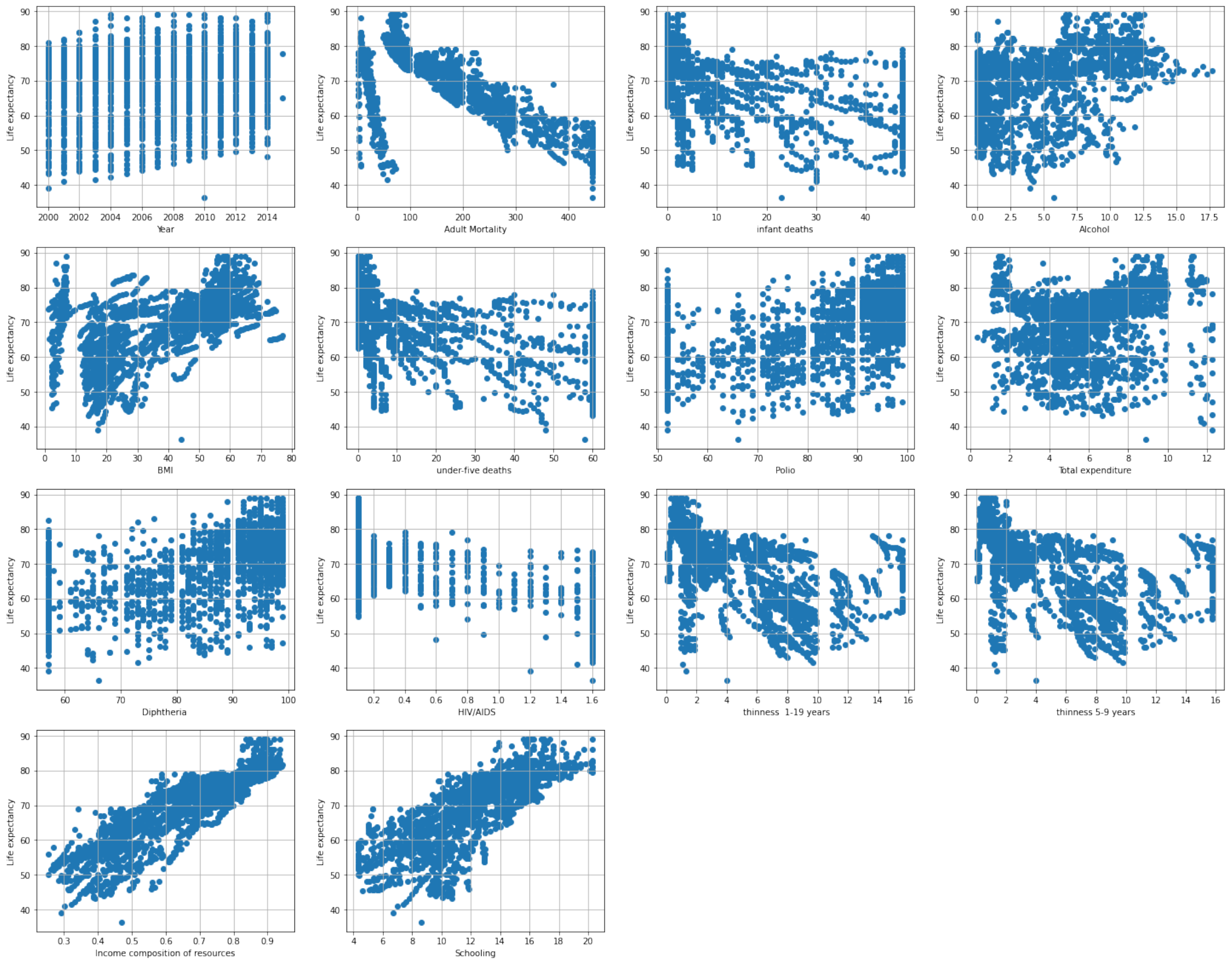

Figure 5: Life expectancy vs feature after shifting the outlier values

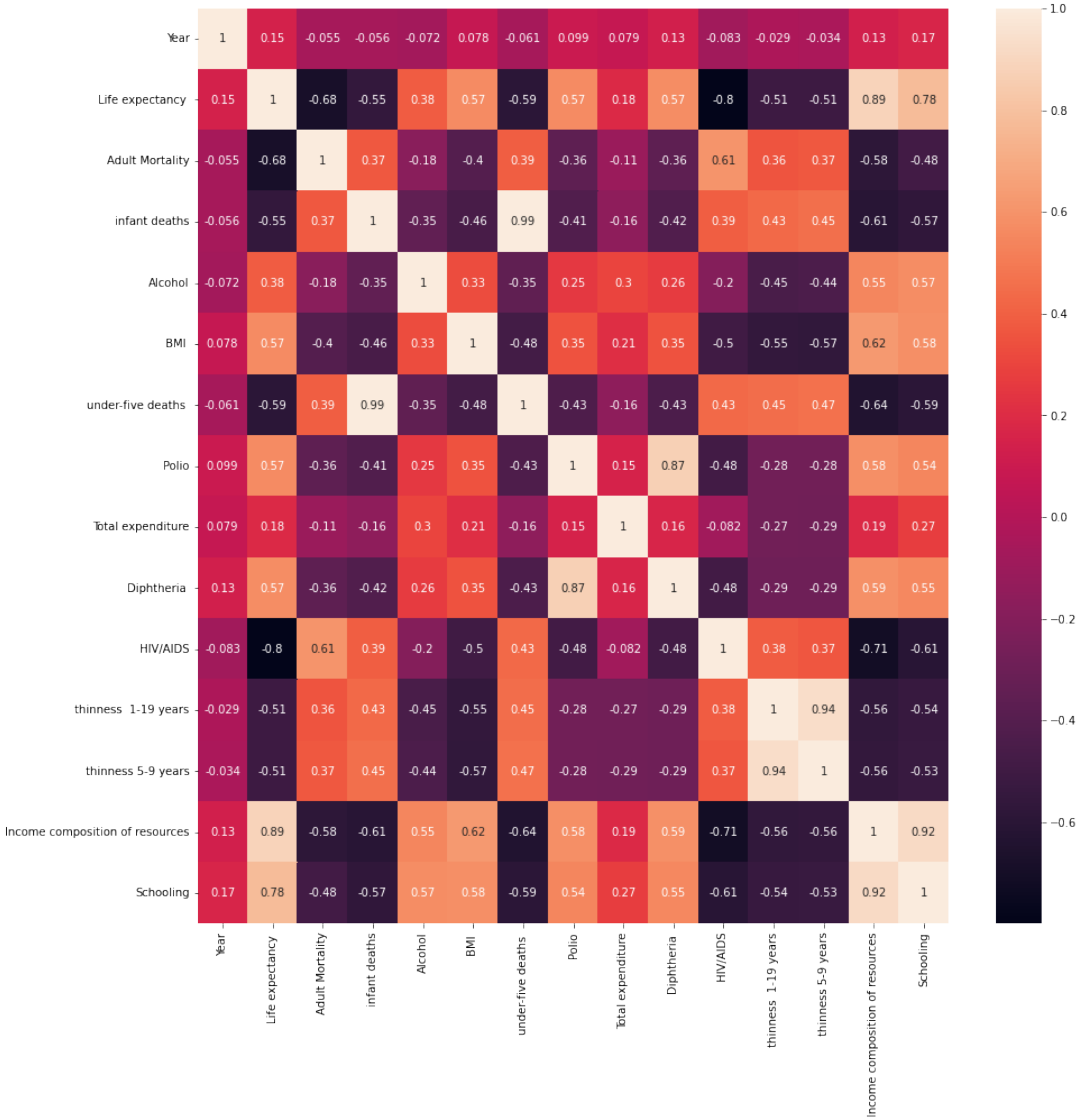

Figure 6: Heatmap of correlation matrix

Here, we see that there are many correlated features. We can safely drop the following features because of their high correlation with other feature.

- Infant Deaths (Correlation of 0.99 with under five deaths)
- Thinness 5-9 years (Correlation of 0.94 with thinness 1-19 years)
- Diphtheria (Correlation of 0.87 with Polio)

- Schooling (Correlation of 0.92 with income composition of resources)

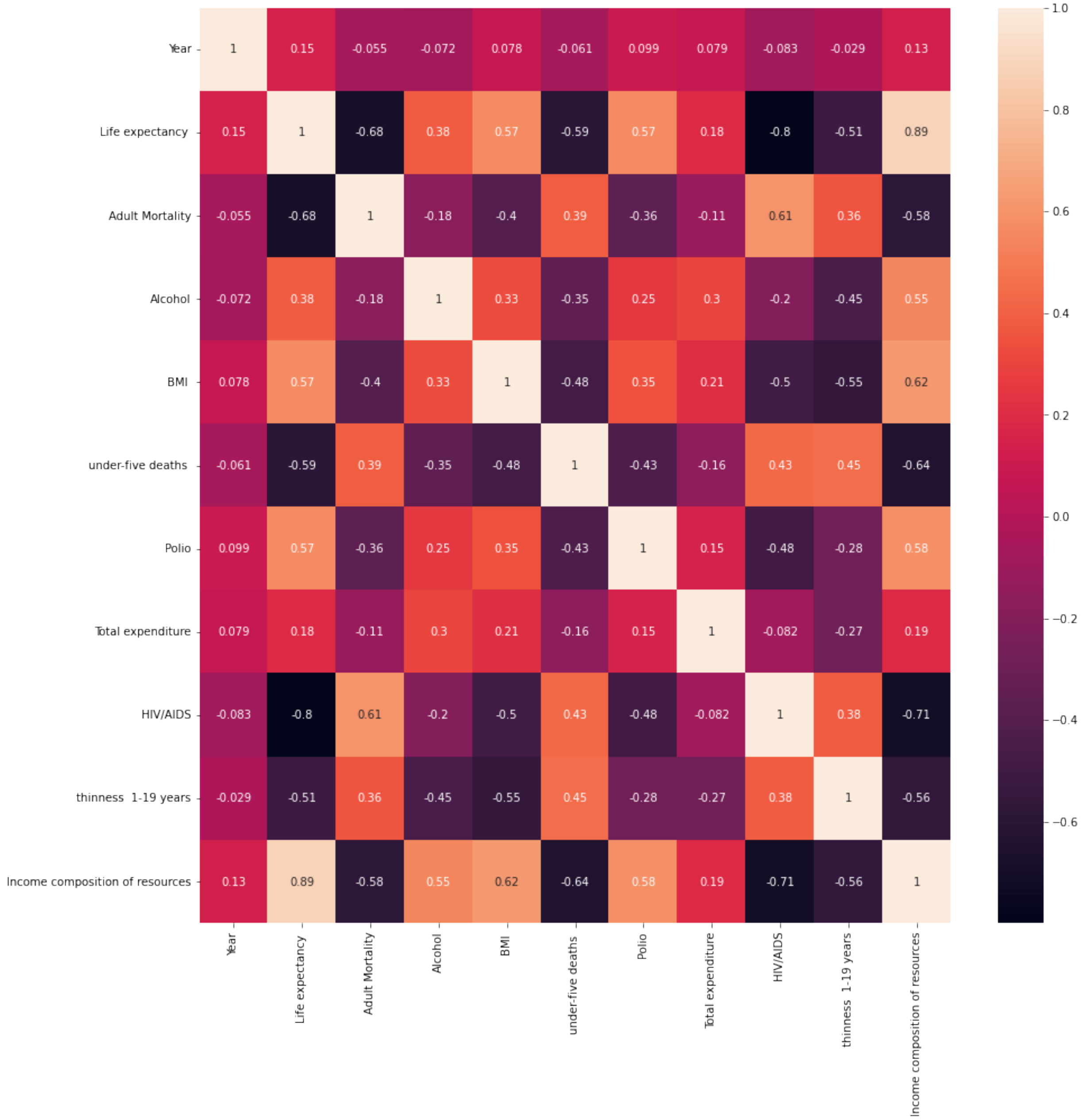

Figure 7: Heatmap of correlation matrix after removing correlated features

We consider 75% of this dataset for training the linear regression models. Before training, we normalize the data.

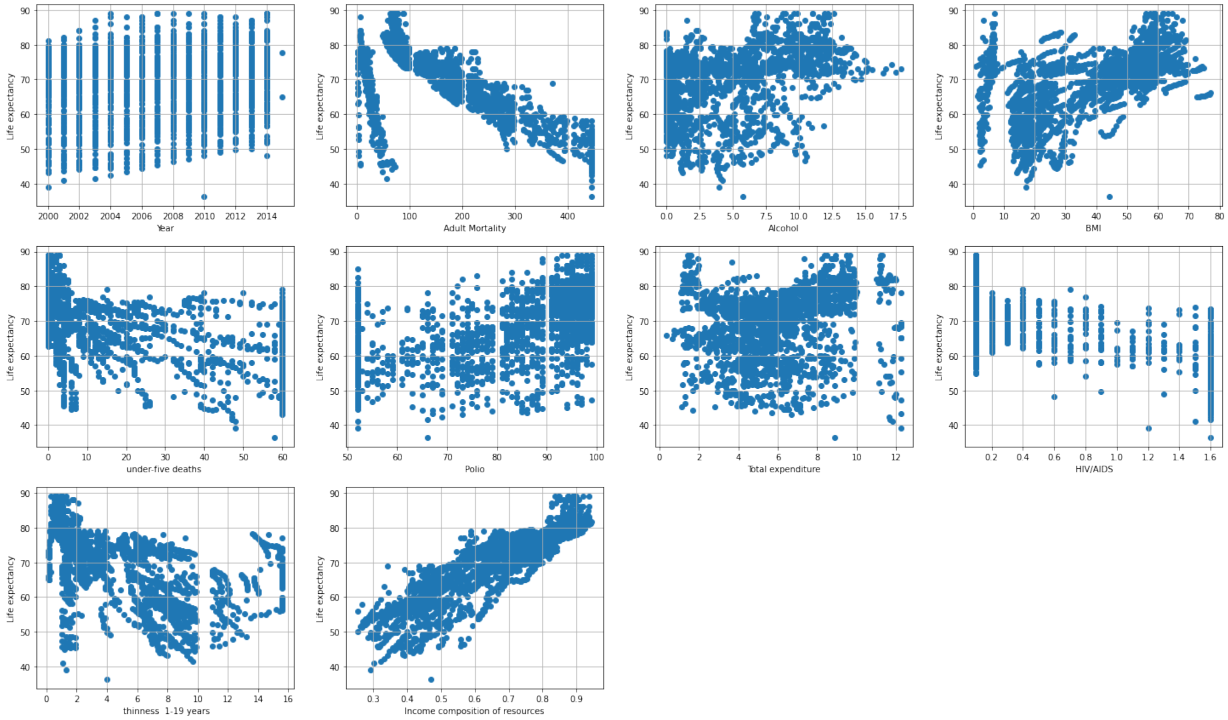

Figure 8: Life expectancy vs feature after cleansing of the data and removal of relevant features

4 Regression Model

The output of the model for closed form solution, gradient descent method and newton's method of optimization is shown below for all the features as well as multivariate model.

4.1 Year

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 7.87        Mean Squared Error: 90.45

Gradient Descent solution
Mean Absolute Error: 7.87        Mean Squared Error: 90.45

Newton Method solution
Mean Absolute Error: 7.87        Mean Squared Error: 90.45
```

ON TRAINING DATA:

```
Closed Form solution
Mean Absolute Error: 7.53        Mean Squared Error: 85.69
```

```
Gradient Descent solution
Mean Absolute Error: 7.53        Mean Squared Error: 85.69

Newton Method solution
Mean Absolute Error: 7.53        Mean Squared Error: 85.69
```

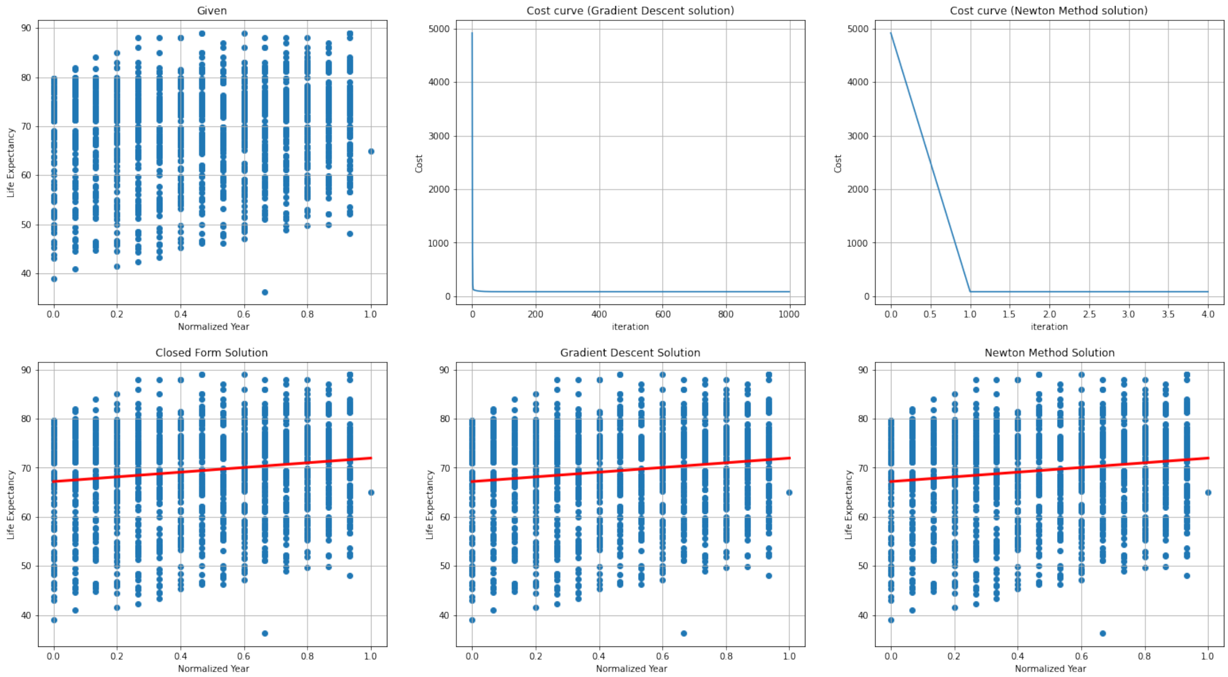

Figure 9:

4.2 Adult Mortality

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 5.01        Mean Squared Error: 49.61

Gradient Descent solution
Mean Absolute Error: 5.01        Mean Squared Error: 49.61

Newton Method solution
Mean Absolute Error: 5.01        Mean Squared Error: 49.61
```

ON TRAINING DATA:

```
Closed Form solution
```

```
Mean Absolute Error: 4.8        Mean Squared Error: 46.51

Gradient Descent solution
Mean Absolute Error: 4.8        Mean Squared Error: 46.51

Newton Method solution
Mean Absolute Error: 4.8        Mean Squared Error: 46.51
```

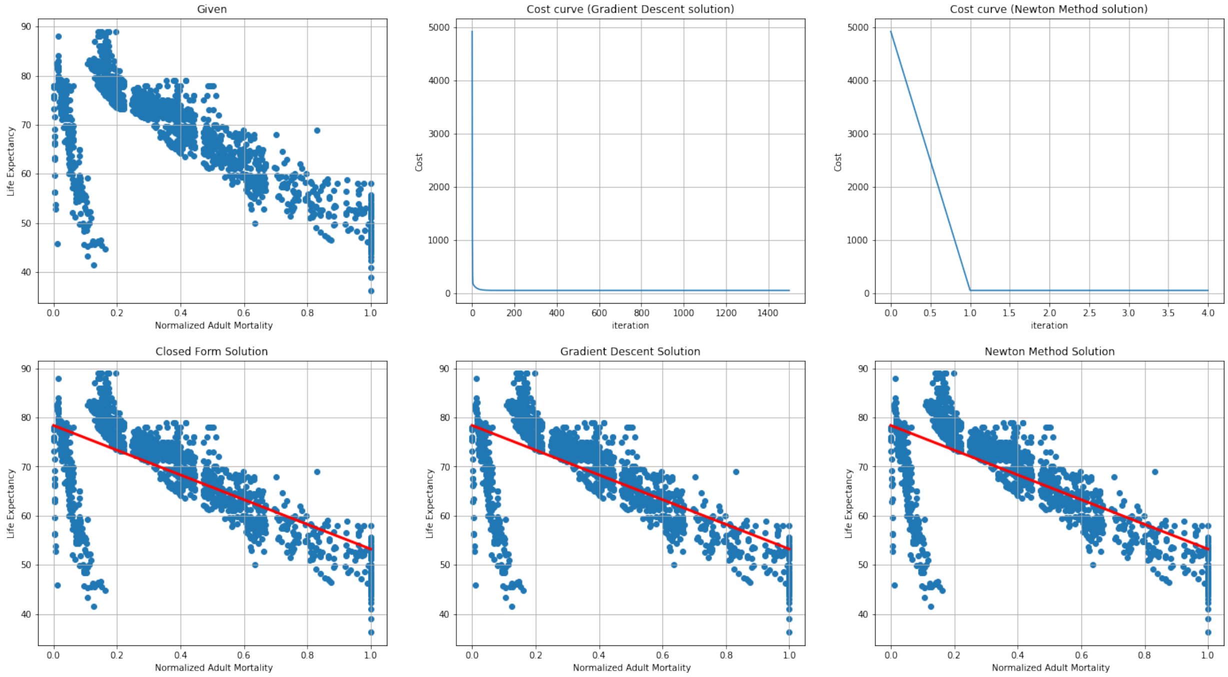

Figure 10:

4.3 Alcohol

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 7.02        Mean Squared Error: 81.8

Gradient Descent solution
Mean Absolute Error: 7.02        Mean Squared Error: 81.8

Newton Method solution
Mean Absolute Error: 7.02        Mean Squared Error: 81.8
```

ON TRAINING DATA:

```
Closed Form solution
Mean Absolute Error: 6.74       Mean Squared Error: 74.19

Gradient Descent solution
Mean Absolute Error: 6.74       Mean Squared Error: 74.19

Newton Method solution
Mean Absolute Error: 6.74       Mean Squared Error: 74.19
```

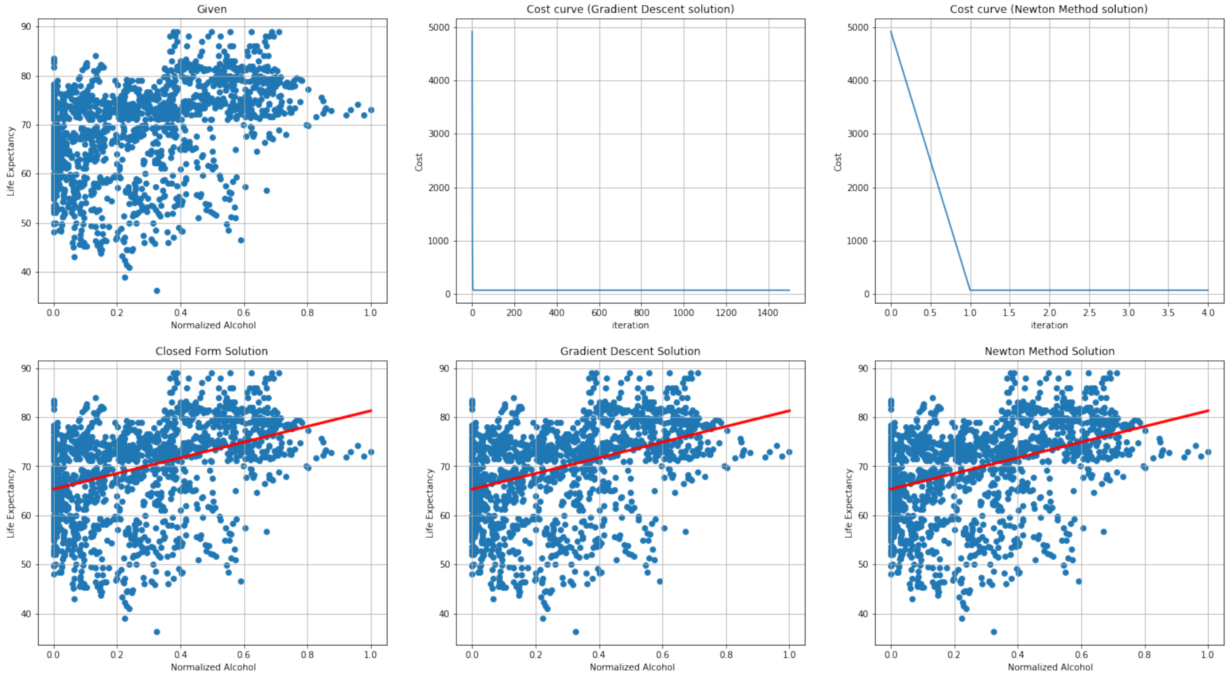

Figure 11:

4.4 BMI

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 5.99       Mean Squared Error: 63.04

Gradient Descent solution
Mean Absolute Error: 5.99       Mean Squared Error: 63.04

Newton Method solution
Mean Absolute Error: 5.99       Mean Squared Error: 63.04
```

```
ON TRAINING DATA:
 Closed Form solution
Mean Absolute Error: 5.78        Mean Squared Error: 59.39

 Gradient Descent solution
Mean Absolute Error: 5.78        Mean Squared Error: 59.39

Newton Method solution
Mean Absolute Error: 5.78        Mean Squared Error: 59.39
```

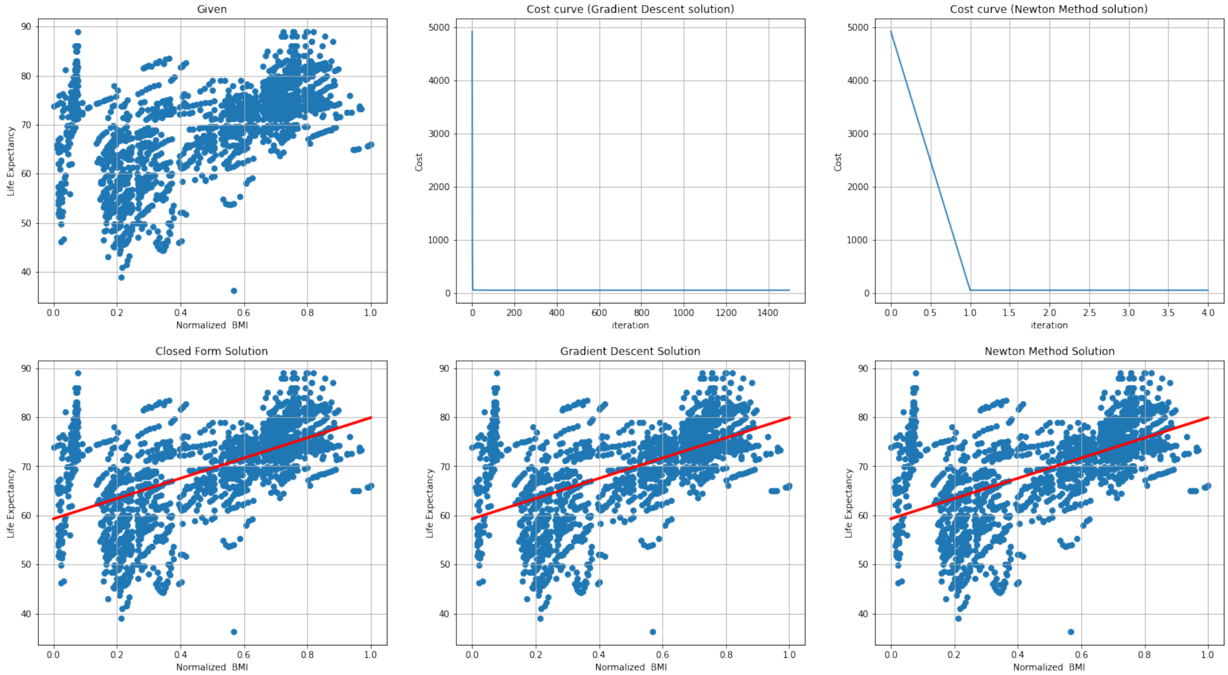

Figure 12:

4.5 Under five deaths

```
ON TEST DATA:
 Closed Form solution
Mean Absolute Error: 6.1        Mean Squared Error: 63.74

 Gradient Descent solution
Mean Absolute Error: 6.1        Mean Squared Error: 63.74

Newton Method solution
Mean Absolute Error: 6.1        Mean Squared Error: 63.74
```

```
ON TRAINING DATA:
Closed Form solution
Mean Absolute Error: 5.68        Mean Squared Error: 55.84

Gradient Descent solution
Mean Absolute Error: 5.68        Mean Squared Error: 55.84

Newton Method solution
Mean Absolute Error: 5.68        Mean Squared Error: 55.84
```

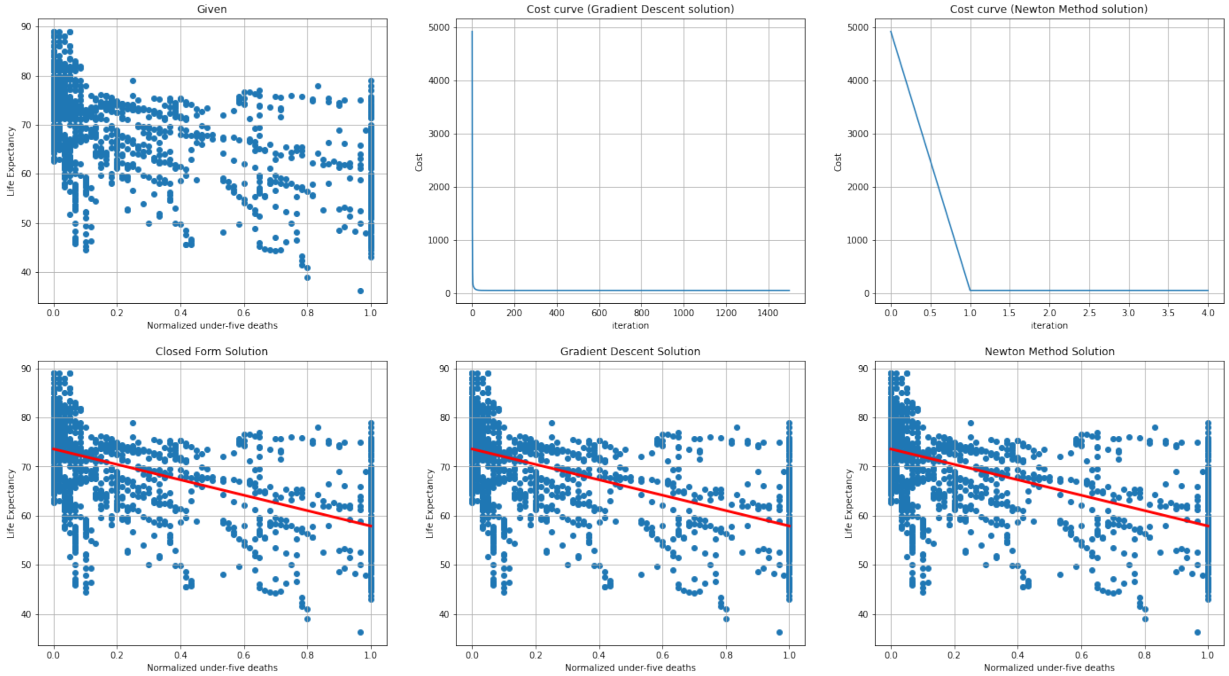

Figure 13:

4.6 Polio

```
ON TEST DATA:
Closed Form solution
Mean Absolute Error: 5.87        Mean Squared Error: 61.74

Gradient Descent solution
Mean Absolute Error: 5.87        Mean Squared Error: 61.74

Newton Method solution
```

```
Mean Absolute Error: 5.87        Mean Squared Error: 61.74
```

ON TRAINING DATA:

```
 Closed Form solution
Mean Absolute Error: 5.78        Mean Squared Error: 59.87

 Gradient Descent solution
Mean Absolute Error: 5.78        Mean Squared Error: 59.87

 Newton Method solution
Mean Absolute Error: 5.78        Mean Squared Error: 59.87
```

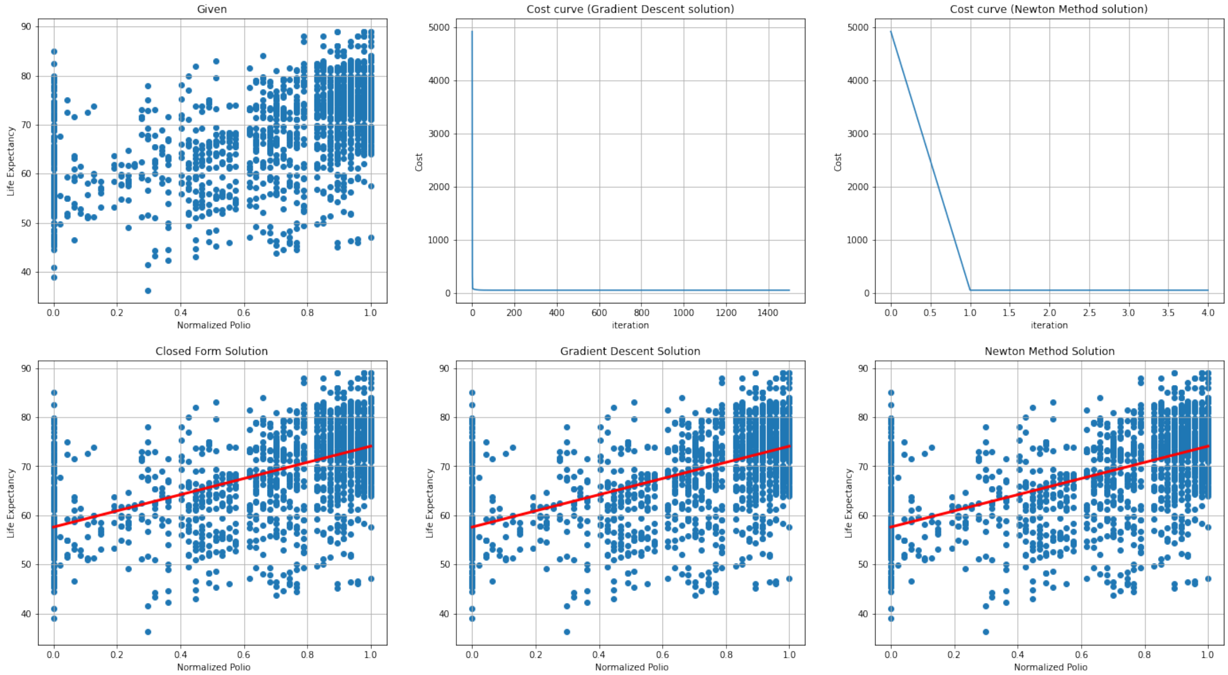

Figure 14:

4.7 Total Expenditure

ON TEST DATA:

```
 Closed Form solution
Mean Absolute Error: 7.68        Mean Squared Error: 89.0

 Gradient Descent solution
Mean Absolute Error: 7.68        Mean Squared Error: 89.0
```

```
Newton Method solution
Mean Absolute Error: 7.68        Mean Squared Error: 89.0
```

ON TRAINING DATA:

```
 Closed Form solution
Mean Absolute Error: 7.29        Mean Squared Error: 84.76

 Gradient Descent solution
Mean Absolute Error: 7.29        Mean Squared Error: 84.76

Newton Method solution
Mean Absolute Error: 7.29        Mean Squared Error: 84.76
```

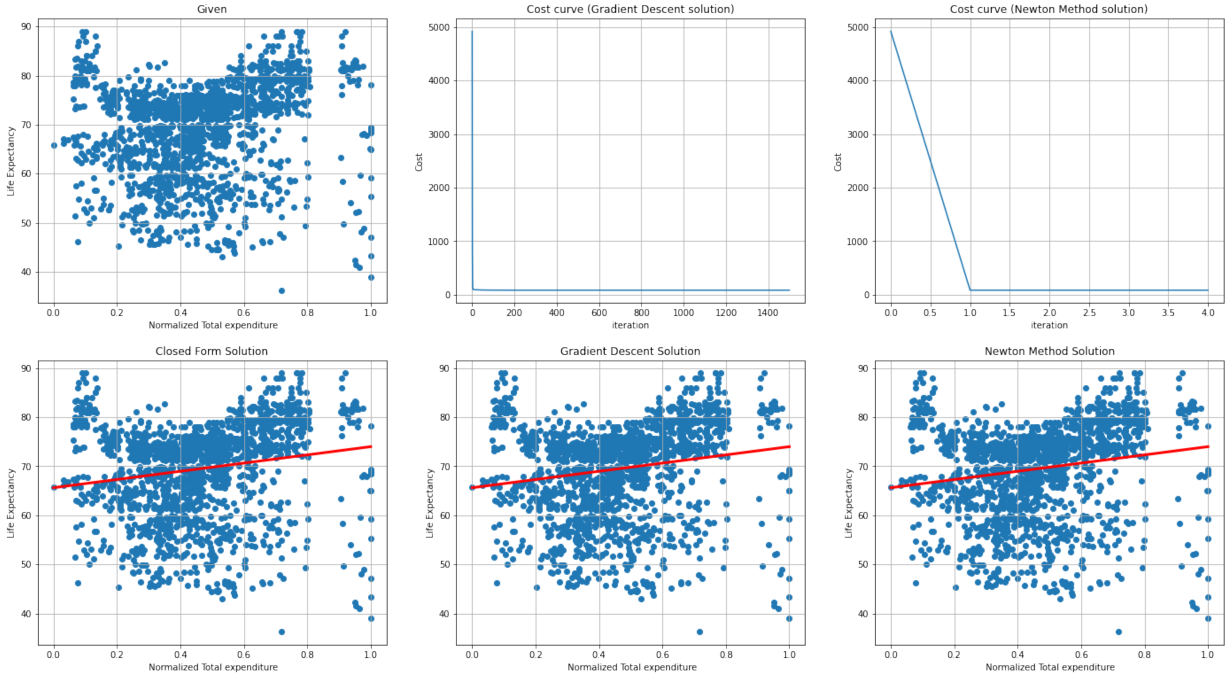

Figure 15:

4.8 HIV/AIDS

ON TEST DATA:

```
 Closed Form solution
Mean Absolute Error: 4.12        Mean Squared Error: 30.51

 Gradient Descent solution
Mean Absolute Error: 4.12        Mean Squared Error: 30.51
```

```
Newton Method solution
Mean Absolute Error: 4.12        Mean Squared Error: 30.51
```

ON TRAINING DATA:

```
Closed Form solution
Mean Absolute Error: 4.37        Mean Squared Error: 32.51

Gradient Descent solution
Mean Absolute Error: 4.37        Mean Squared Error: 32.51

Newton Method solution
Mean Absolute Error: 4.37        Mean Squared Error: 32.51
```

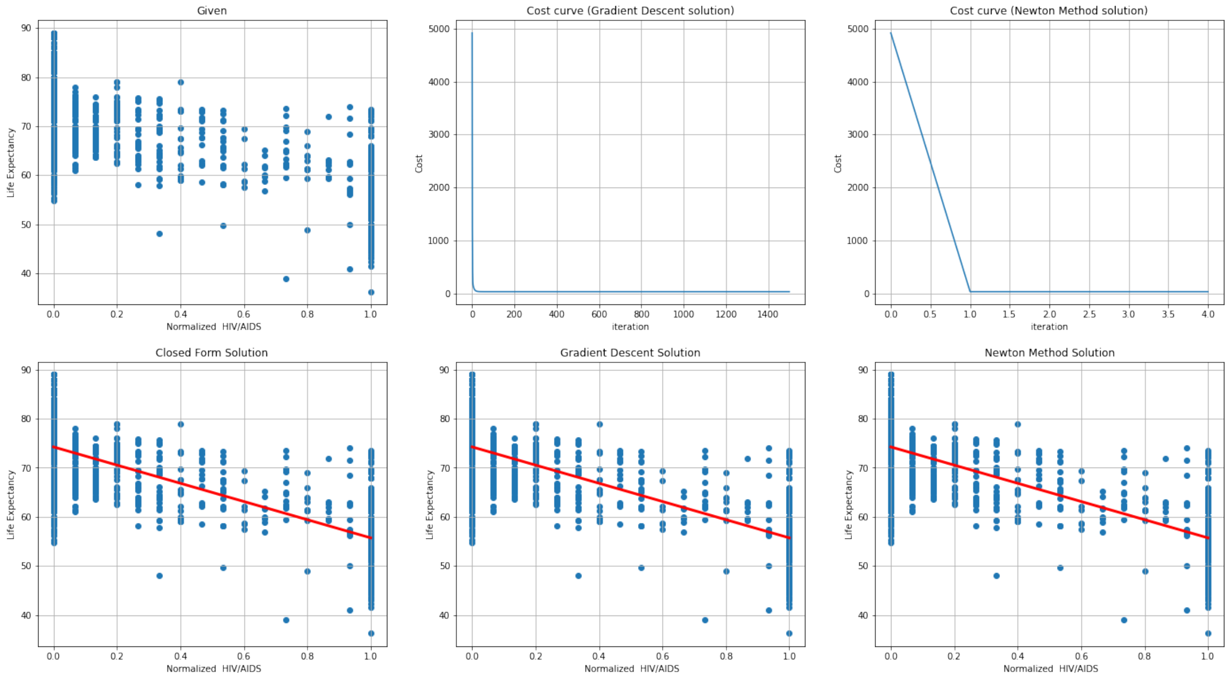

Figure 16:

4.9 Thinness 1-19 years

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 6.4        Mean Squared Error: 66.04

Gradient Descent solution
```

```
Mean Absolute Error: 6.4       Mean Squared Error: 66.04

Newton Method solution
Mean Absolute Error: 6.4       Mean Squared Error: 66.04
```

ON TRAINING DATA:

```
Closed Form solution
Mean Absolute Error: 6.31       Mean Squared Error: 65.75

Gradient Descent solution
Mean Absolute Error: 6.31       Mean Squared Error: 65.75

Newton Method solution
Mean Absolute Error: 6.31       Mean Squared Error: 65.75
```

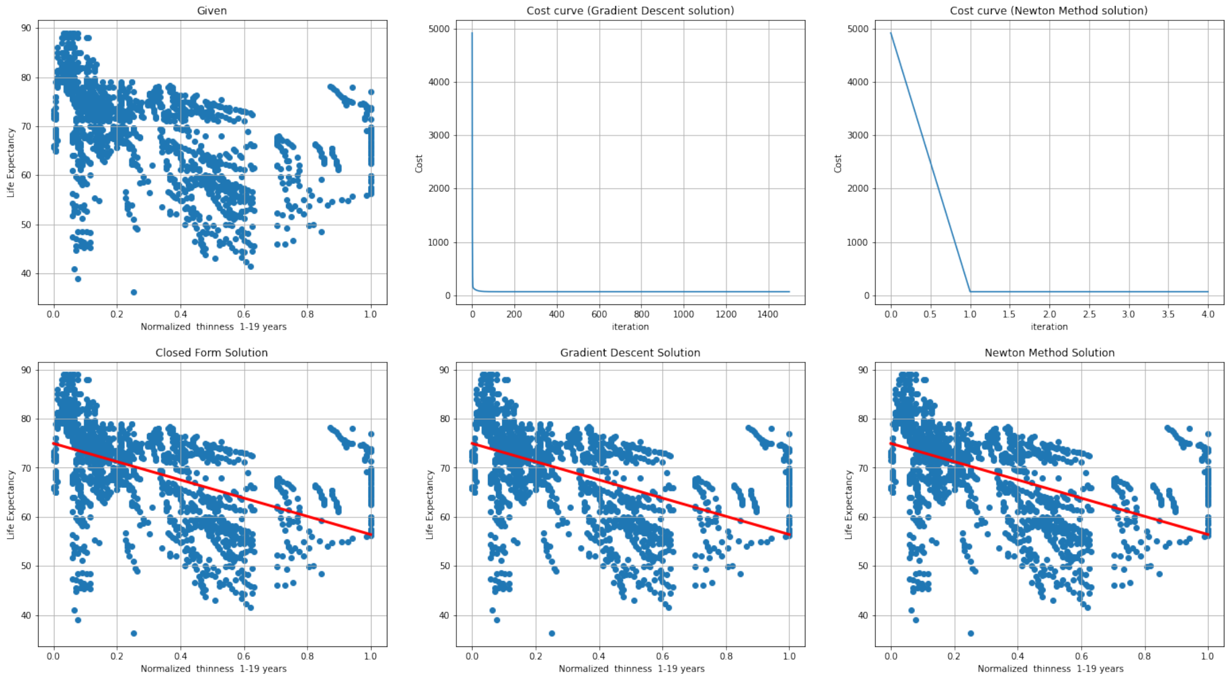

Figure 17:

4.10 Income Composition of Resources

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 3.37       Mean Squared Error: 20.69
```

```
Gradient Descent solution
Mean Absolute Error: 3.37        Mean Squared Error: 20.69

Newton Method solution
Mean Absolute Error: 3.37        Mean Squared Error: 20.69
```

ON TRAINING DATA:

```
Closed Form solution
Mean Absolute Error: 3.17        Mean Squared Error: 18.36

Gradient Descent solution
Mean Absolute Error: 3.17        Mean Squared Error: 18.36

Newton Method solution
Mean Absolute Error: 3.17        Mean Squared Error: 18.36
```

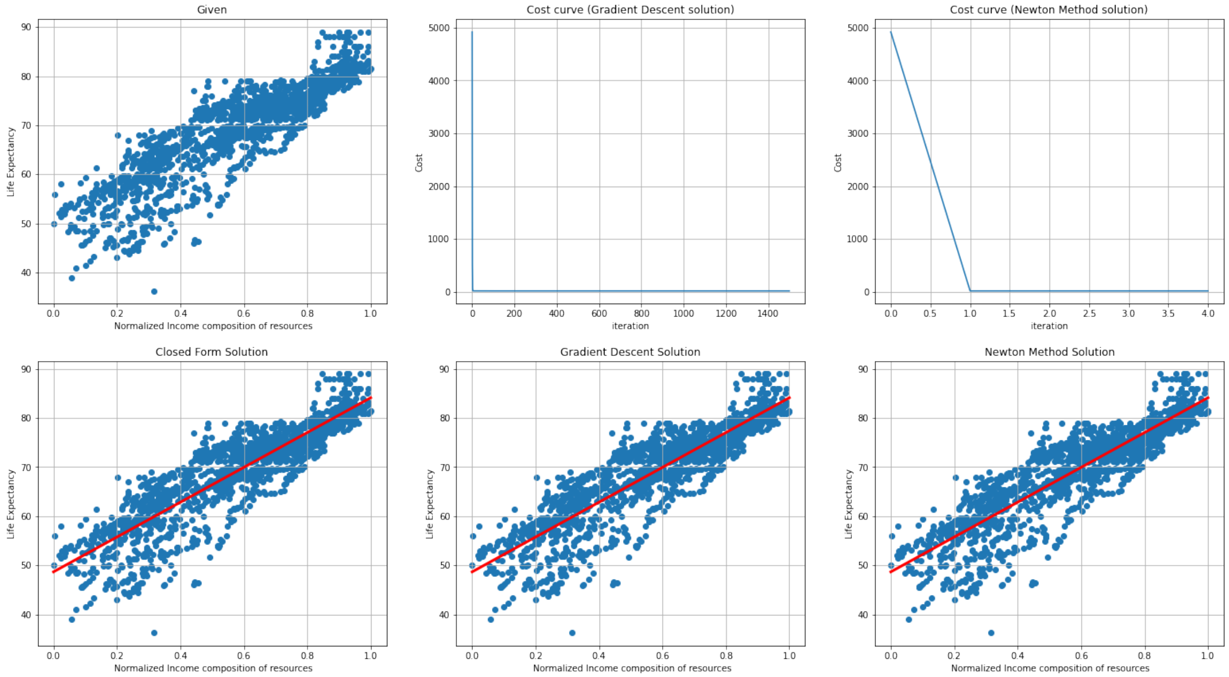

Figure 18:

4.11 Multivariate

ON TEST DATA:

```
Closed Form solution
Mean Absolute Error: 2.58        Mean Squared Error: 12.49
```

```
Gradient Descent solution
Mean Absolute Error: 2.58        Mean Squared Error: 12.49

Newton Method solution
Mean Absolute Error: 2.58        Mean Squared Error: 12.49
```

ON TRAINING DATA:

```
Closed Form solution
Mean Absolute Error: 2.53        Mean Squared Error: 11.59

Gradient Descent solution
Mean Absolute Error: 2.53        Mean Squared Error: 11.59

Newton Method solution
Mean Absolute Error: 2.53        Mean Squared Error: 11.59
```

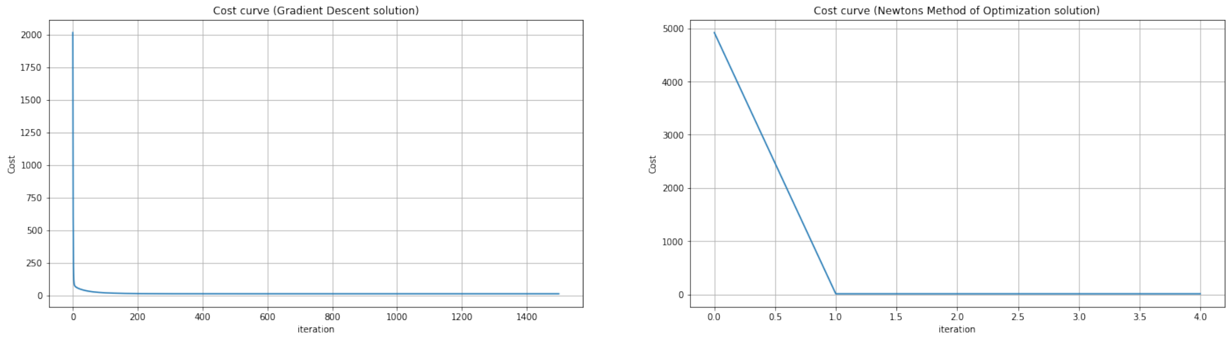

Figure 19: