****

An Analysis Report on

**A Data-Driven Analysis of Factors Affecting Life Expectancy**

**Submitted to: Submitted By:**

Abhinav Khandelwal

24122002

**Introduction:**

Life expectancy is a measure that indicates a person's average lifespan. Several factors influence a person's life expectancy, such as whether the person lives in a developed or developing country, their economic condition, the country's population, GDP, and other social and geographical factors.

This study aims to build a regression model to predict the average life expectancy based on economic, social, geographical, and mortality factors.

**Dependent And Independent Columns:**

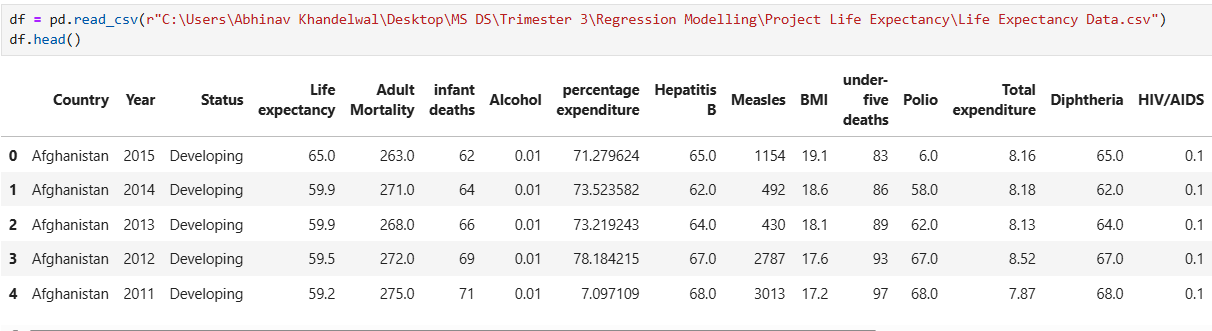
In this study the Dependent Column will be Life Expectancy which will be predicted and mainly the independent variable are status of the country like the country is developed or developing, Alcohol, Infant Deaths, Adult Mortality and etc.

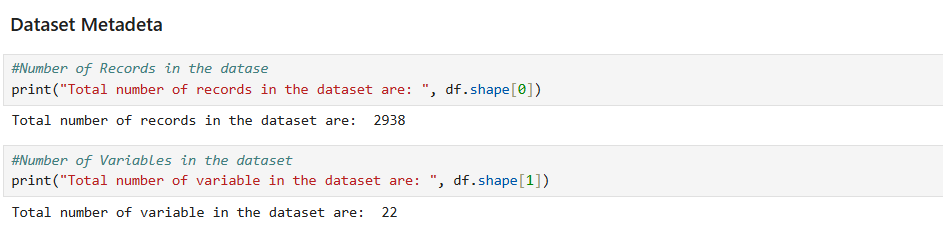
**Collect and Understand the Dataset:**

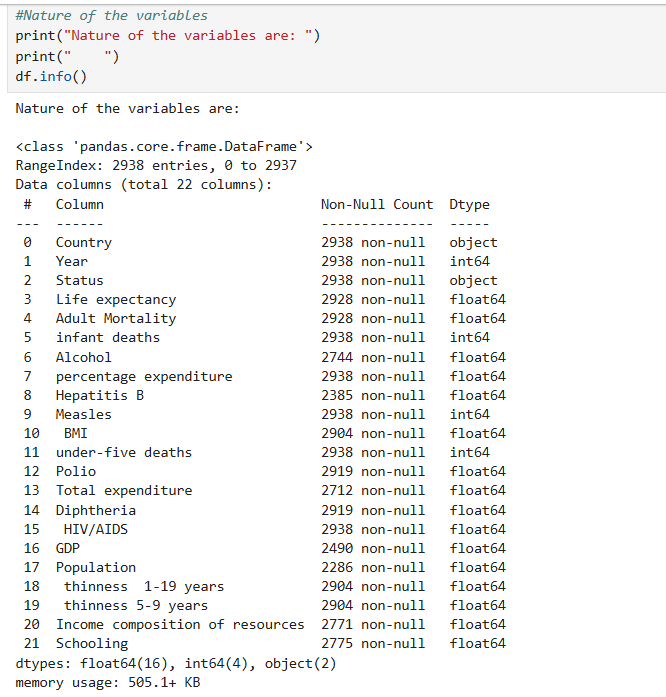
The Life Expectancy Dataset, which was initially based on information gathered by the World Health Organization (WHO), was obtained from Kaggle and used in this investigation. It includes comprehensive data on mortality rates over a number of years, socioeconomic variables, and health indicators for different nations.

* Features like life expectancy, adult mortality, infant mortality, alcohol use, GDP, education, BMI, vaccination rates, and population are all included in the dataset.
* The inclusion of both developed and developing nations enables comparisons across various developmental stages.

The information is meant to aid in the analysis of the ways in which social, economic, and health-related factors affect the population's average lifespan.

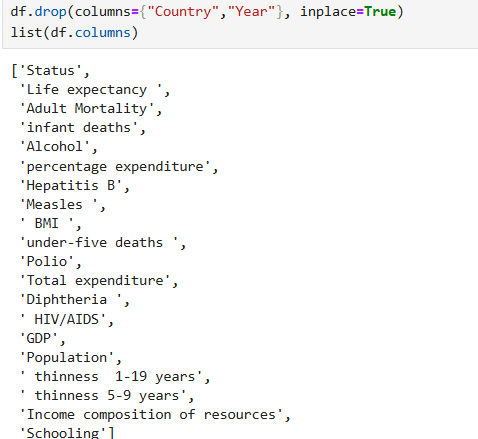
****

****

****

**Exploratory Data Analysis and Data Preprocessing:**

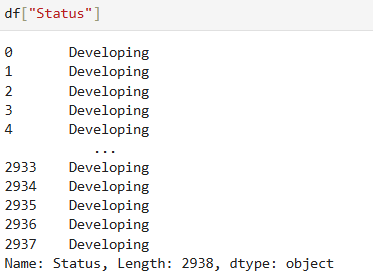
After reviewing the dataset, we can remove the year and the country column and not contributing to the analysis. Instead of a country we will focus on Status of the country whether it is developed or developing.

****

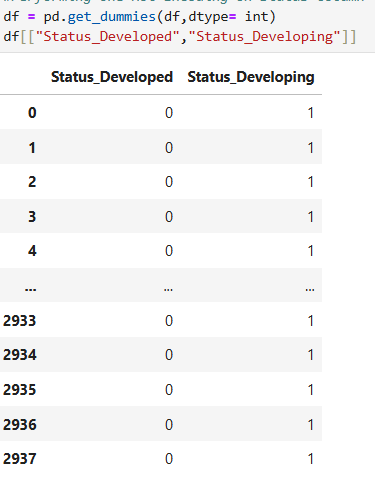
**Encoding Categorical Column:**

There is a Categorical feature in the dataset which is status of the country which tells about whether the country is developed or developing. **One hot encoding** technique will be used to encode the categorical column to numerical column for the machine learning to train.

Before Encoding:



After Encoding:



**Correlation Matrix:**

Correlation Matrix is used to plot to check the correlation between the independent variables to check if there is Multicollinearity and then heat map is used to visualize the correlation matrix for clear understanding.

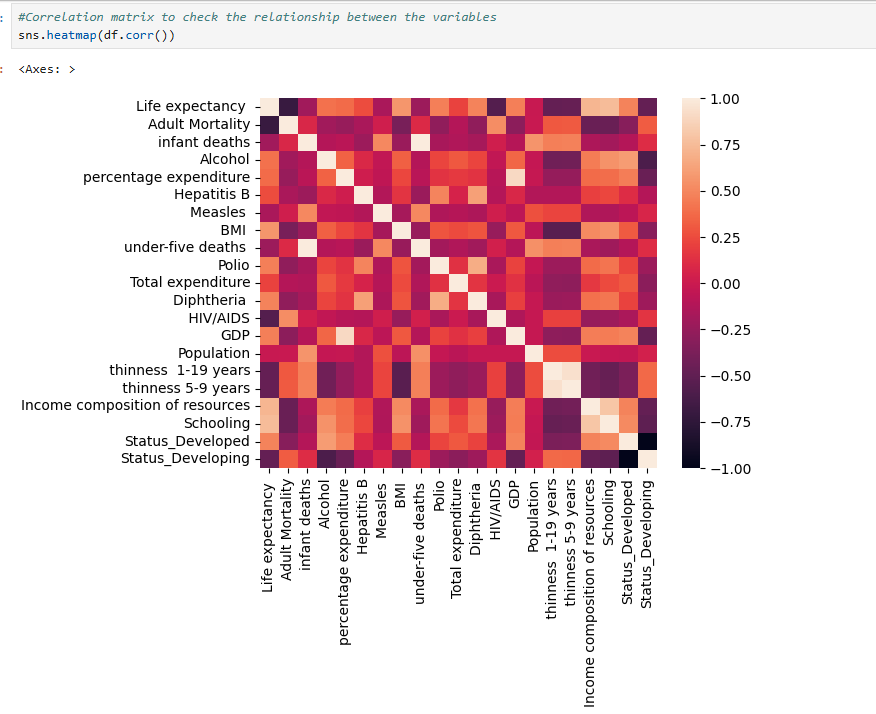


Figure 1: Correlation Matrix

From the above visualization we can conclude that there are some variables (independent variables) which are highly correlated having the correlation greater than 0.5 in both positive and negative direction which shows close to strong correlation this can affect out model.

**Distribution of the Dependent Variable:**

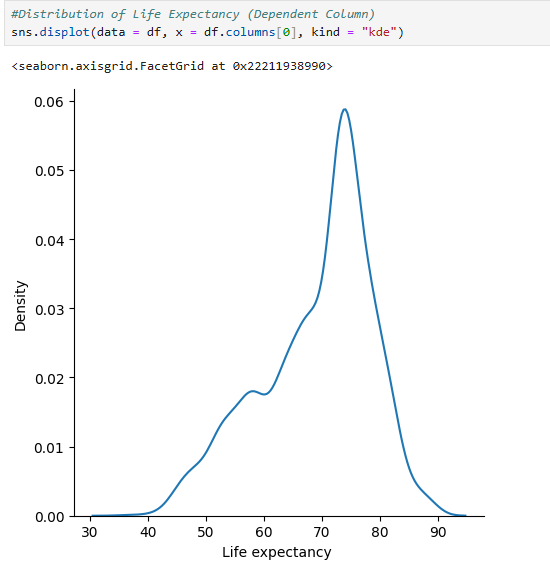


Figure 2 Life Expectancy Distribution

From the distribution we can clearly that the distribution is negatively skewed which shows that the majority of the people has average age in between 62 and 85 but let’s see how this distribution is when we check Developing vs Developed Countries.

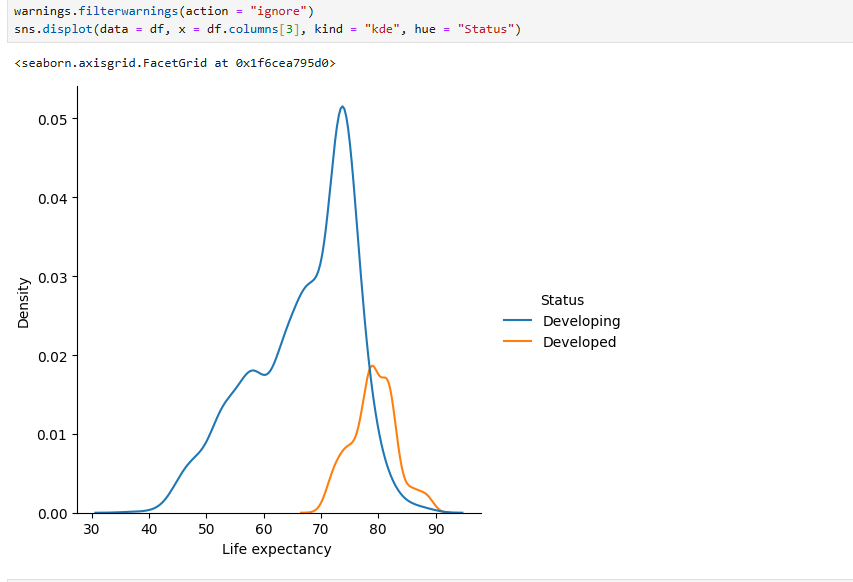


Figure 3 Life Expectancy based on Developed and Developing Countries

From the above distribution we can see that the distribution of Developed Countries are close to normal and they have the average age is between 65 – 90 and majority are in between 78 – 85. While in the case of Developing countries they have the range from 30 – 90 and majority are lying in between 60 – 80. We can say that the average age in the Developed countries are higher and also not highly spread.

**Identifying Outlier:**

Outlier are the datapoints which are away from the natural behaviour of the data. They affect the model performance and linear regression is highly sensitive to the outliers because of the loss function it is used to train the model. It is important to identify the outliers and handled it.

Box plot is used to identify the outliers.



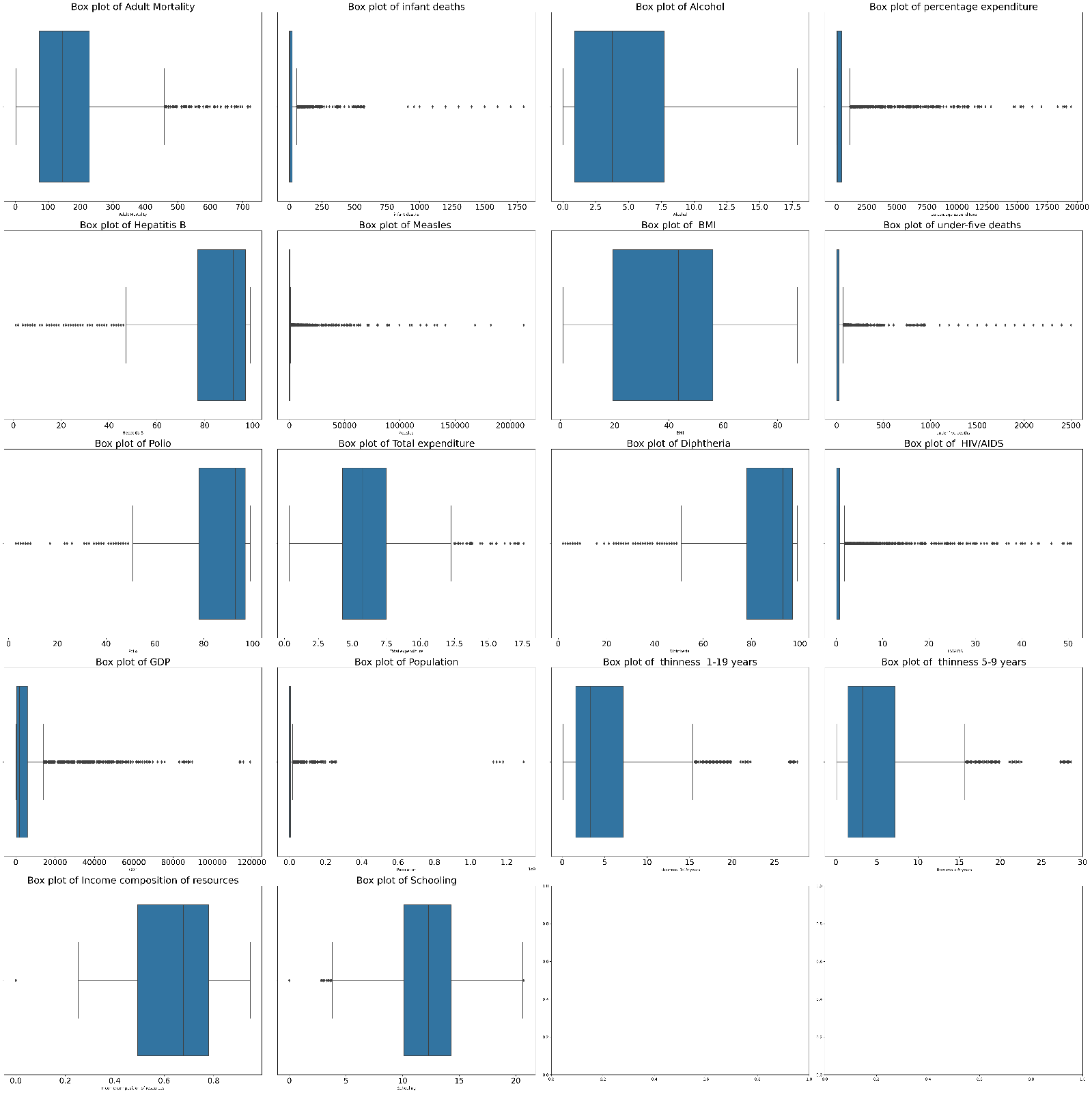
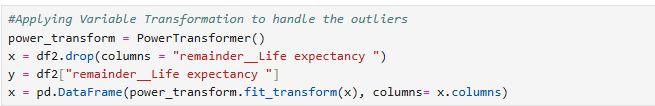


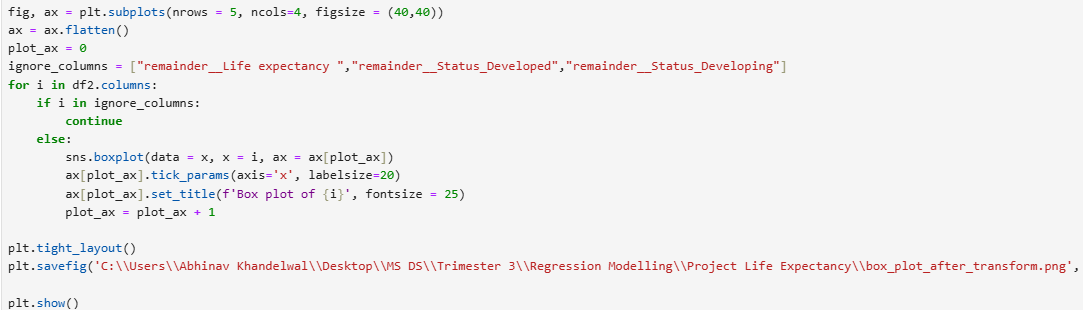
Figure 4 Box Plot of Independent Variables

From the above plots it is very clear that all the independent variables have some outlier’s columns like GDP, thinness 5-9 years, HIV/AIDS. Measles, percentage expenditure these variables have a lot of outliers which is not good for the analysis. So these outliers need to be handled there are many techniques but, in this Transformation, will be used because if we remove the outliers then we will end up losing the dataset.

**Yeo-Johnson Power Transformation:**

Before applying Yeo-Johnson, log transformation, sqrt transformation and other transformation was tried but the performance of the model didn’t improve. Yeo-Johnson Transformation gave the good result and it worked with negative values very well.





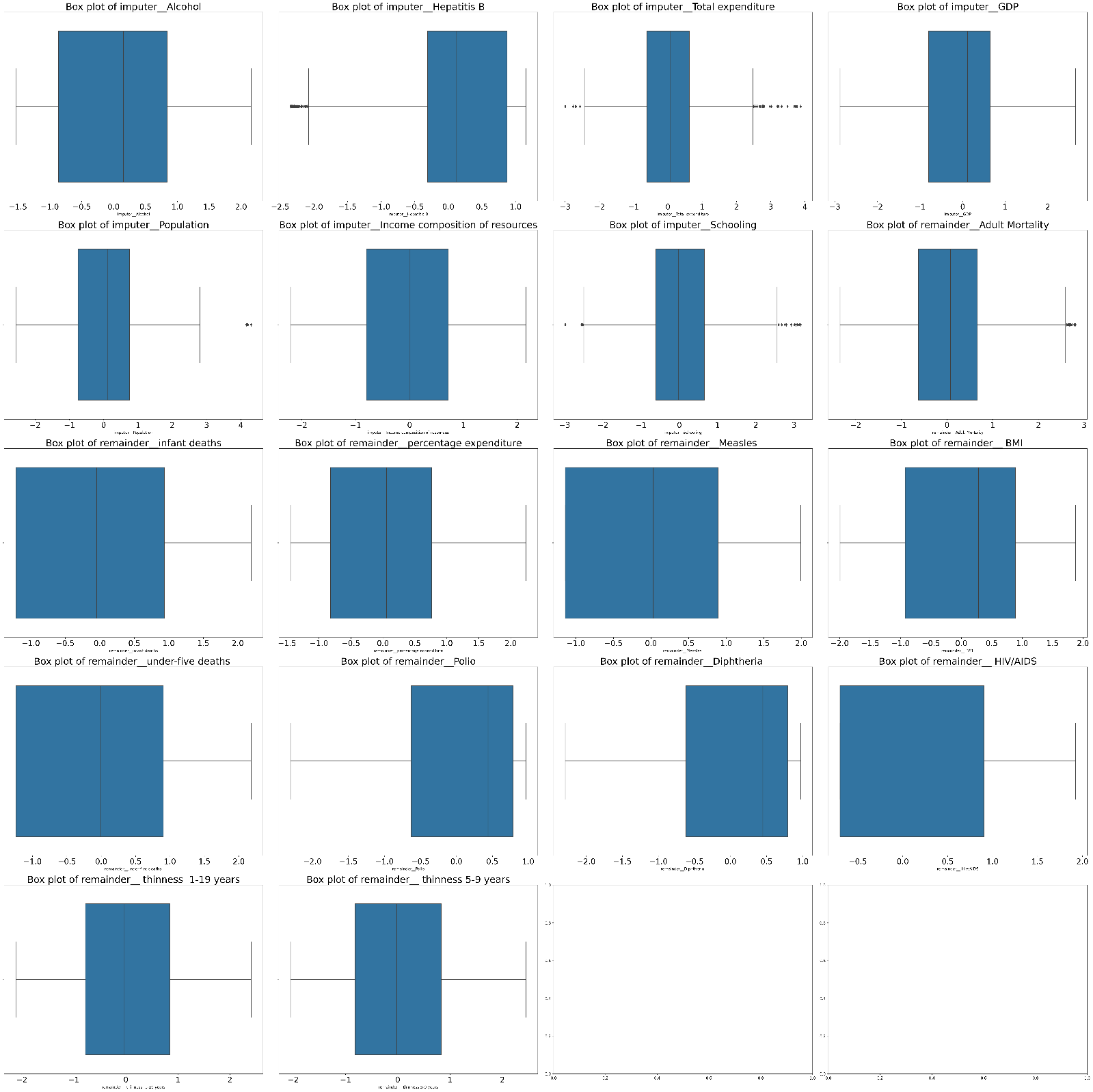
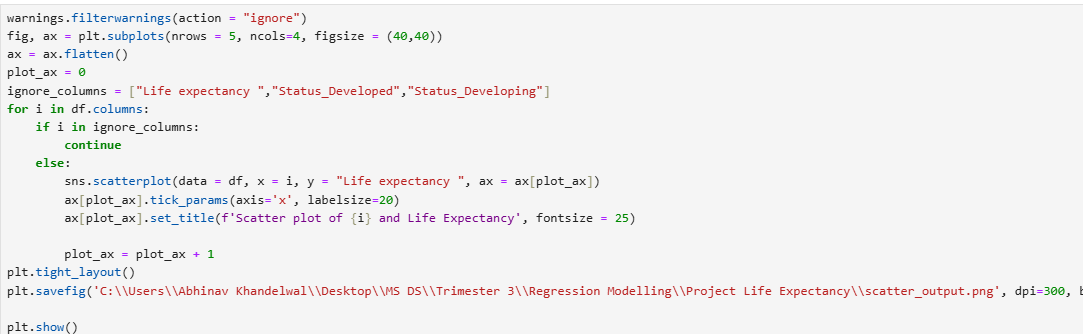


Figure 5 Box Plot after Yeo-Johnson Transformation

After applying the Yeo-Johnson Transformation we can see in the same columns like GDP, thinness 5-9 years, HIV/AIDS. Measles, percentage expenditure the outliers in these columns are handled very well.

**Relationship between Independent Variables and Dependent Variable:**



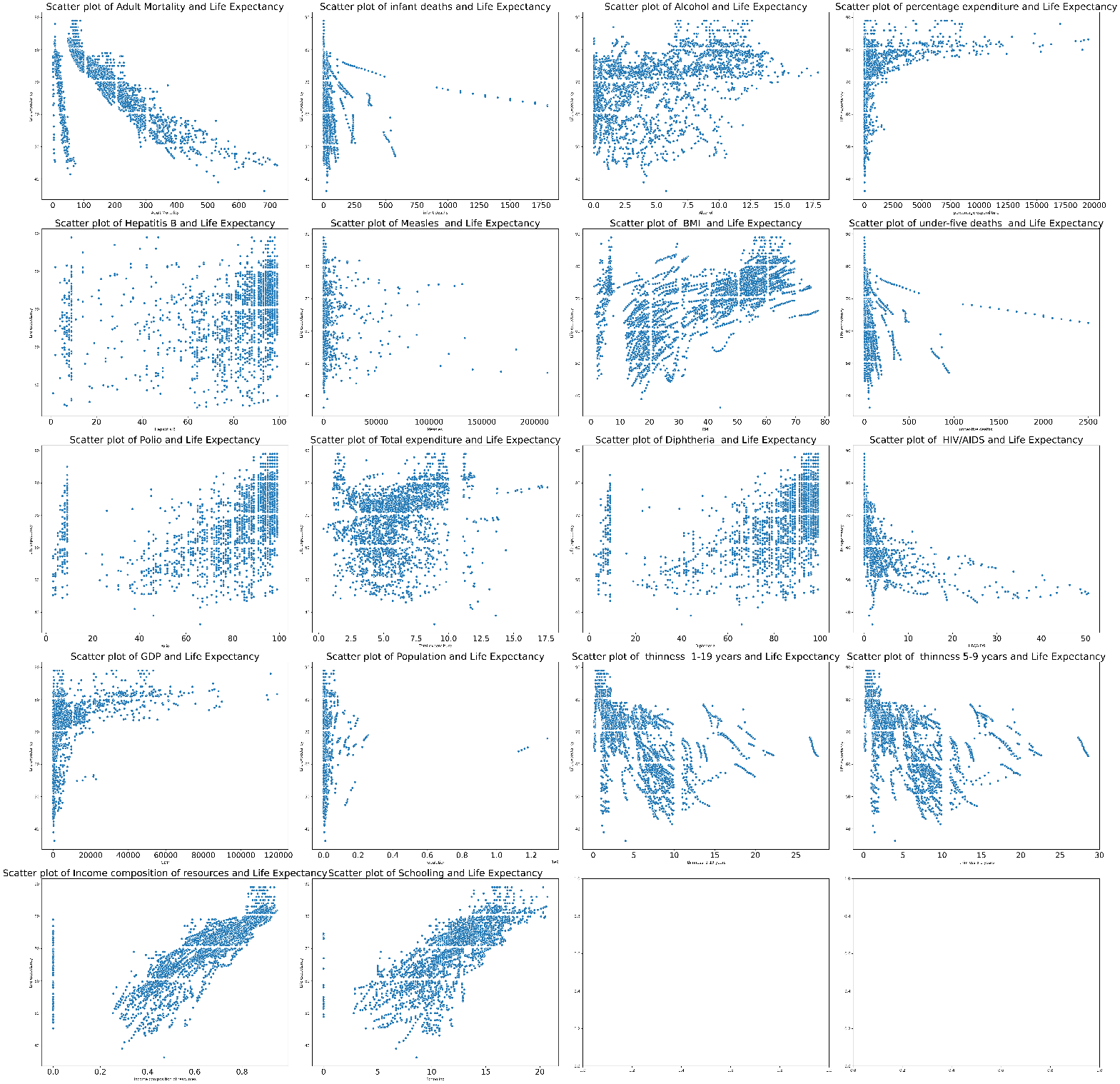


Figure 6 Scatter Plot of the Independent Variables with Dependent Variable

From the scatter plot between independent variable and dependent variable we can observe that not all are having linear relationship with the dependent variable. Some variables like Income composition of resources, Alcohol, Schooling line some sort of linear relationship.

**Identifying and Handling Null Values:**

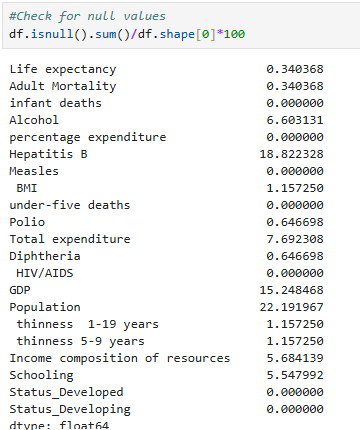
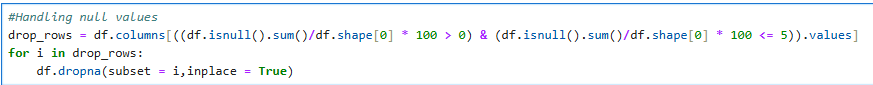
****

Figure 7 Column with Null Values Percentage

Columns with null values less than 5% the rows will be dropped and other columns null values will be imputed through Simple Imputer.

**Dropping the columns having null values less than 5%:**



**Imputing columns having null values greater than 5%:**

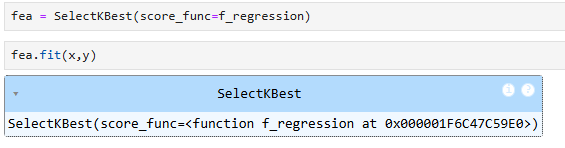


All the null values have been handled as the percentage of the null values are 0.

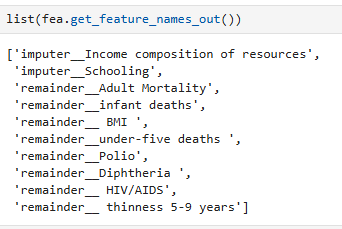
**Model Building:**

**Feature Selection:**

Instead of working on all the feature, feature selection technique will be applied to identify 10 important features in the dataset. In feature selection technique SelectKBest is used with f\_regression score. F\_regression can handle the negative values efficiently as compared to chisq test.



As per the test below are the important features for our analysis.



**Splitting the dataset into training and testing:**

****

**Creating the model:**

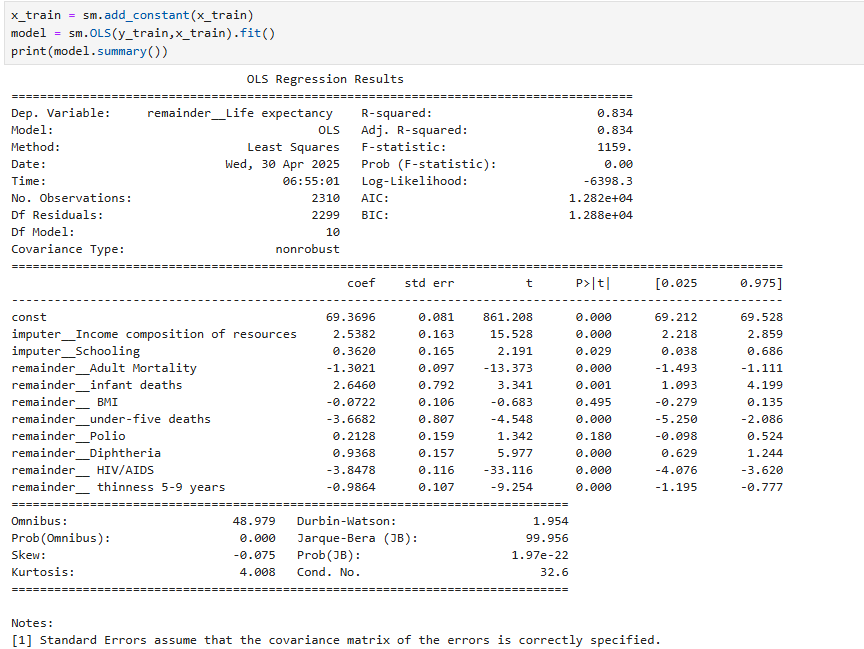
****

Figure 8 Trained Model Parameters

From the OLS Table we can observe that R2 value is 0.834 which means that the independent variables in the model is able to capture 83% of the variance of the dependent variable and Adjusted R2 is also 0.834 which indicated that R2 is not inflated because of increase of the variables.

**Interpretation of p-values:**

p-value less than 0.05 means we have enough evidence to reject the null hypothesis and accept the alternate hypothesis. Below are the columns which has p-value <0.05 and are significant for the analysis.

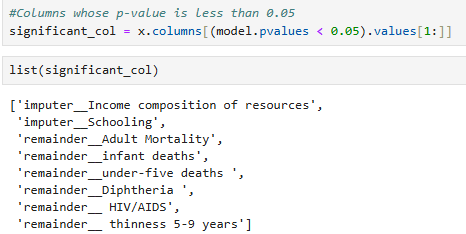
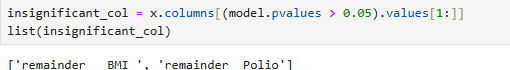


Figure 9 Significant Columns as per p-value

Below are the columns which are not significant has p-value greater than 0.05:



**Interpretation of Coefficients:**

Variables like **Income composition of resources, Adult Mortality, infant deaths, under-five deaths, HIV/AIDS** these variables have high impact on the dependent variables as their coefficient values are higher between the range of 1 and 3 as we have transformed the data so we can’t interpret the exact impact. Other variables which are significant has less impact on the Life Expectancy.

**Conclusion and Recommendations:**

Predicting life expectancy is a sensitive and complex task influenced by multiple socio-economic and health-related factors. This project explored various features, including adult mortality, infant deaths, alcohol consumption, GDP, education, BMI, vaccination rates, and population metrics.

To ensure data quality, outliers in the independent variables were treated using the Yeo-Johnson transformation, and missing values were addressed using a simple imputation technique. After preprocessing, feature selection techniques were employed to identify the top 10 most influential features for predicting life expectancy. The final model was trained on these selected features.

The analysis revealed that the most significant contributors to life expectancy include:

* Income composition of resources
* Schooling
* Adult mortality
* Infant deaths
* Under-five deaths
* Diphtheria immunization rate
* HIV/AIDS prevalence
* Thinness among children aged 5–9

These findings can help policymakers and health organizations focus on the most impactful areas to improve life expectancy in various populations.

The model performance can potentially be improved with access to a larger dataset, as the current model was trained on approximately 2,800 data points, which limited its predictive power and resulted in a modest R² score. In future data collection efforts, focusing on the most significant features identified in this analysis—rather than gathering a broad range of variables—could enhance model accuracy and efficiency.