Group Name: The Dundies

Members: Harshita V | Krish A | Joan J | Reiyonna S

Register No: 21112041 | 21112016 | 21112037 | 21112020

**Statistical Analysis on Factors Influencing Life Expectancy**

Dataset Used :

Life Expectancy Prediction Data

Brief Summary of the Dataset:

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 dataset 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 United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative.

Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The dataset involves the following column names:

1. Country - the country in which the indicators are from
2. Year - the calendar year the indicators are from (ranging from 2000 to 2015)
3. Status - whether a country is considered to be 'developing' or 'developed' by who standards
4. Life Expectancy - the life expectancy of people in years for a particular country and year
5. Adult Mortality - the adult mortality rate per 1000 population
6. Infant Deaths - number of infant deaths per 1000 population
7. Alcohol - a country's alcohol consumption rate
8. Percentage Expenditure - expenditure on health as a percentage
9. Hepatitis B - number of 1-year olds with Hepatitis B immunization over all 1-year-old in population
10. Polio - number of 1-year olds with polio immunization over the number of all 1-year-old in population.
11. HIV/AIDS - Deaths per 1000 live births caused by HIV/AIDS
12. GDP - Gross Domestic Product per capita
13. Population - Population of a country
14. Income Composition of Resources - Human development index in terms of income composition of resources (index ranging from 0 to 1)
15. Schooling (ratio) - Average number of years of schooling of a population.

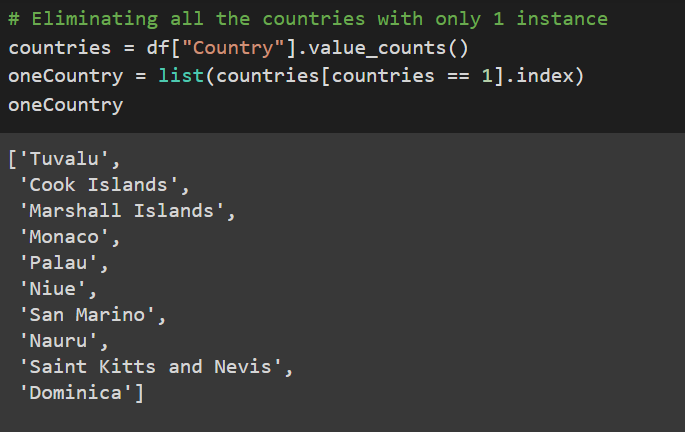
Idea/Approach:

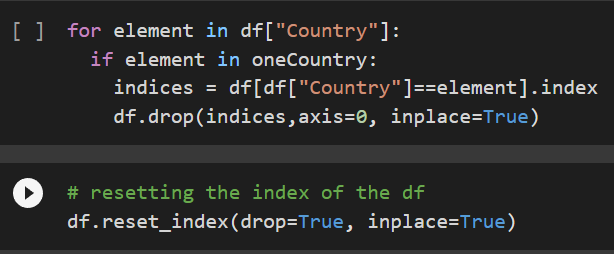
**Step 1** : Data Cleaning and Pre-processing:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Firstly, we had to deal with all the missing/ NA values present in each column of the dataset and find ways to manage or fill those with relevant values.

Part 1: Dropping irrelevant instances from Database:

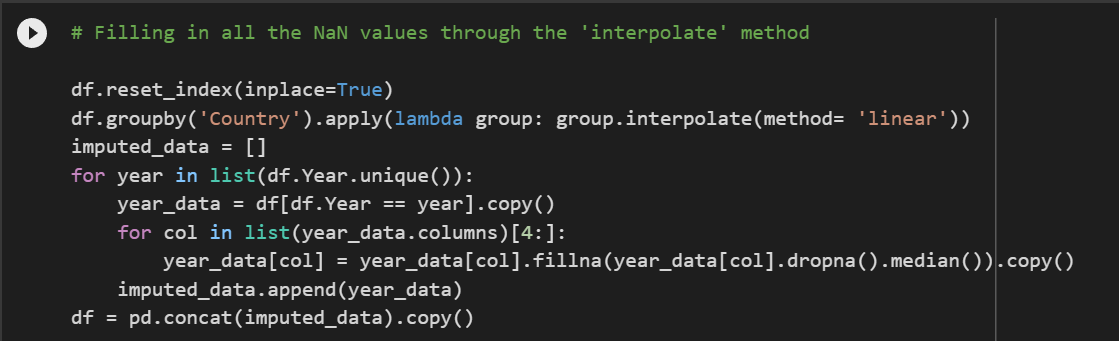
In the provided dataset, each country had multiple entries based on the timeline from 2000-2015. However, there were certain countries which only had single instances and observed to have multiple NA values. Since these cannot be filled with any predefined logic, they were dropped from the dataset.





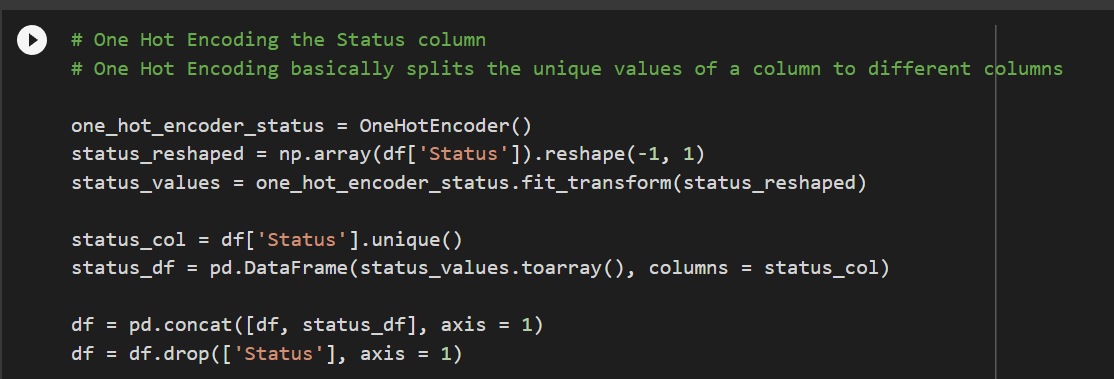
Part 2: Filling NA values with relevant Values using interpolate() function :

Another practical way of filling NA values is using the interpolate() function in Pandas library. Linear Interpolation helps fill gaps within a range of values in a given dataset. In this case, NA values were filled with the most relevant data possible rather than mean or median imputation.



**Step2**: Feature Engineering:

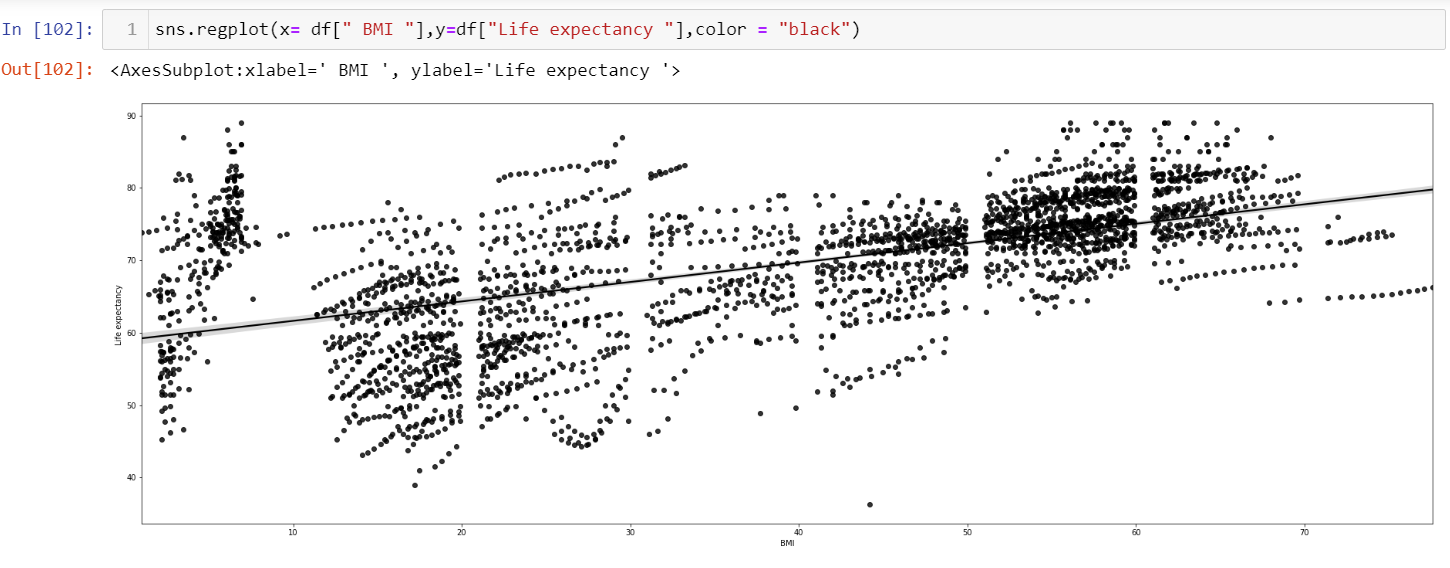
Feature engineering refers to manipulation — addition, deletion, combination, mutation — of your data set to improve machine learning model training, leading to better performance and greater accuracy.

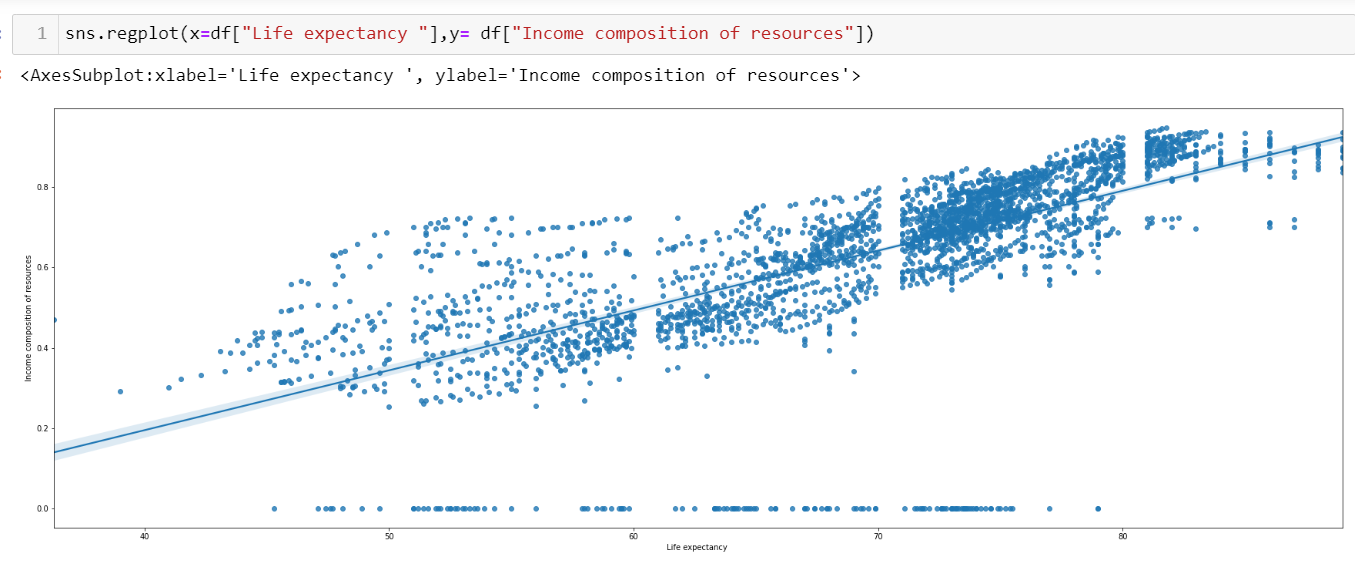


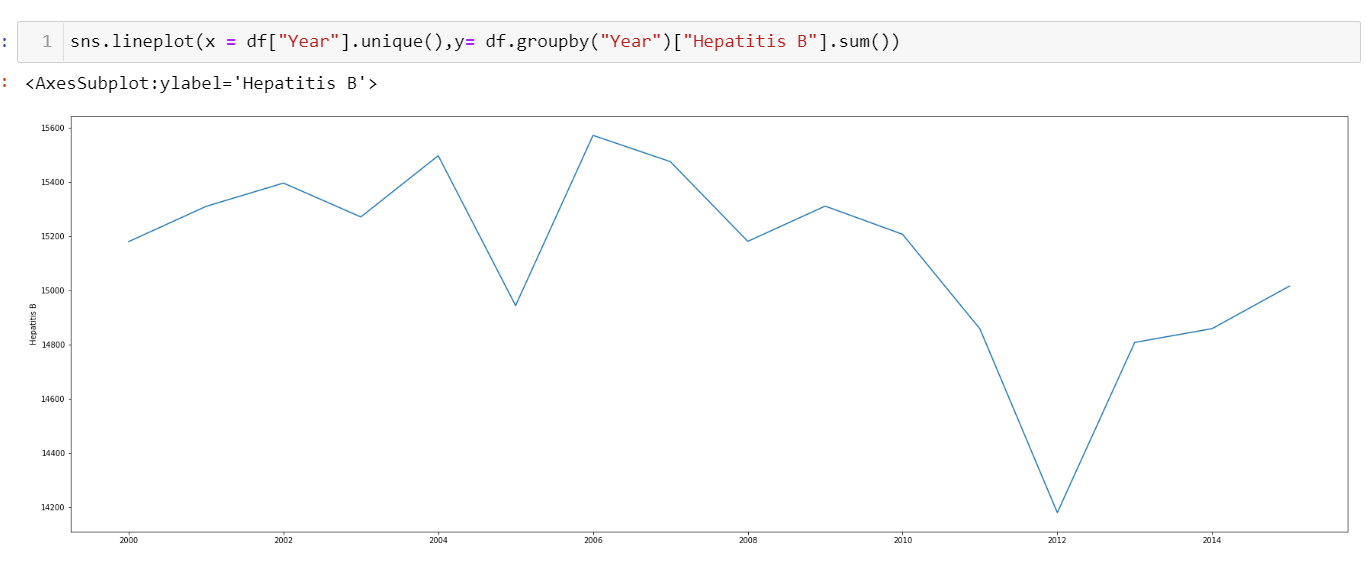
The given dataset has two categorical variables that is, Country and Status. Country column was engineered using Label Encoding. Label Encoding basically changes the unique categorical value to numerical categories in the same column itself. The Status column was encoded using One Hot Encoding which splits the unique values of a column to different columns.

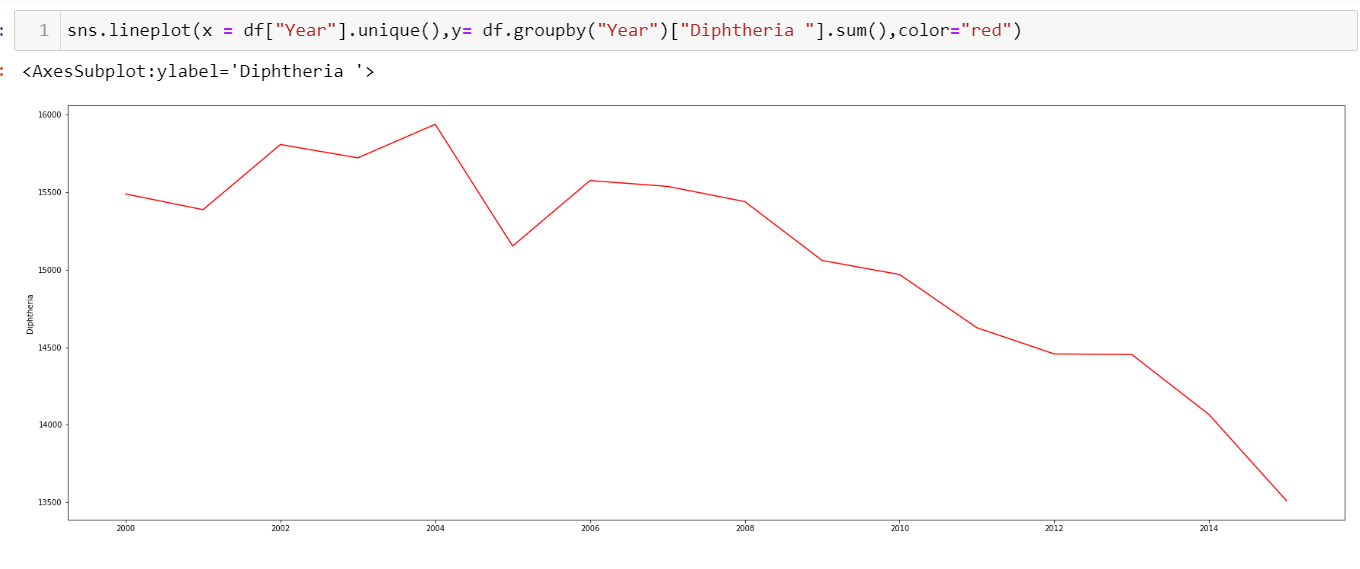
**Step 3**: Exploratory Data Analysis:

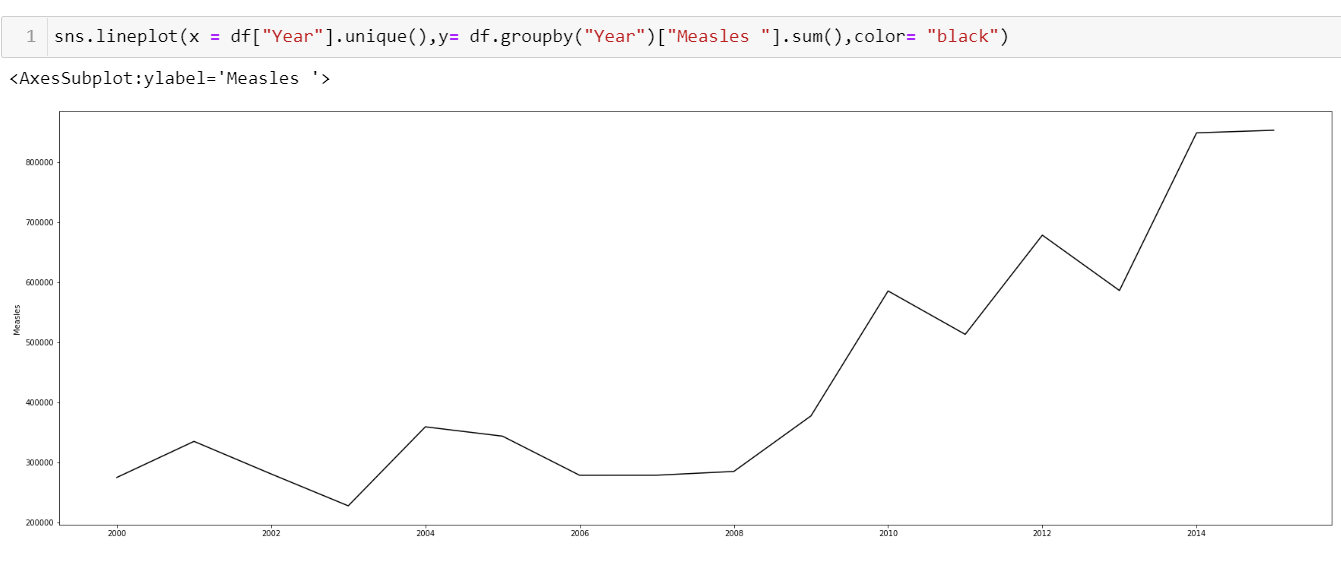
Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. The following insights were taken from the Life Expectancy dataset :



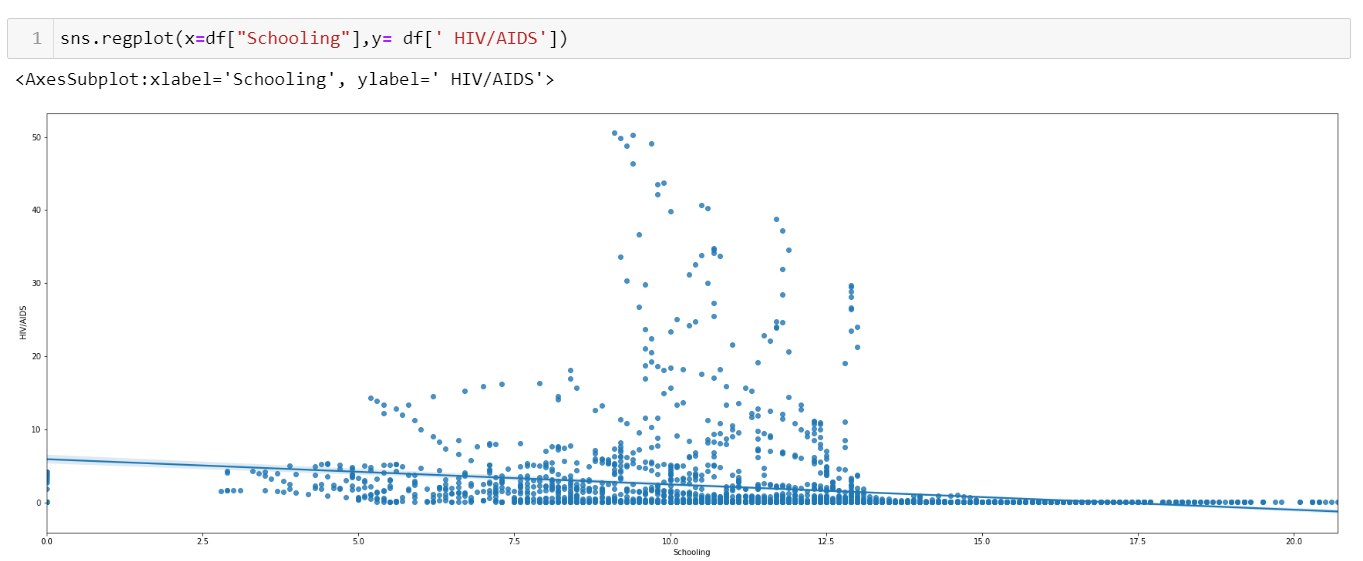
Observation: An Upward trend in life-expectancy vs Income composition







Examples of solo parameter value checks.



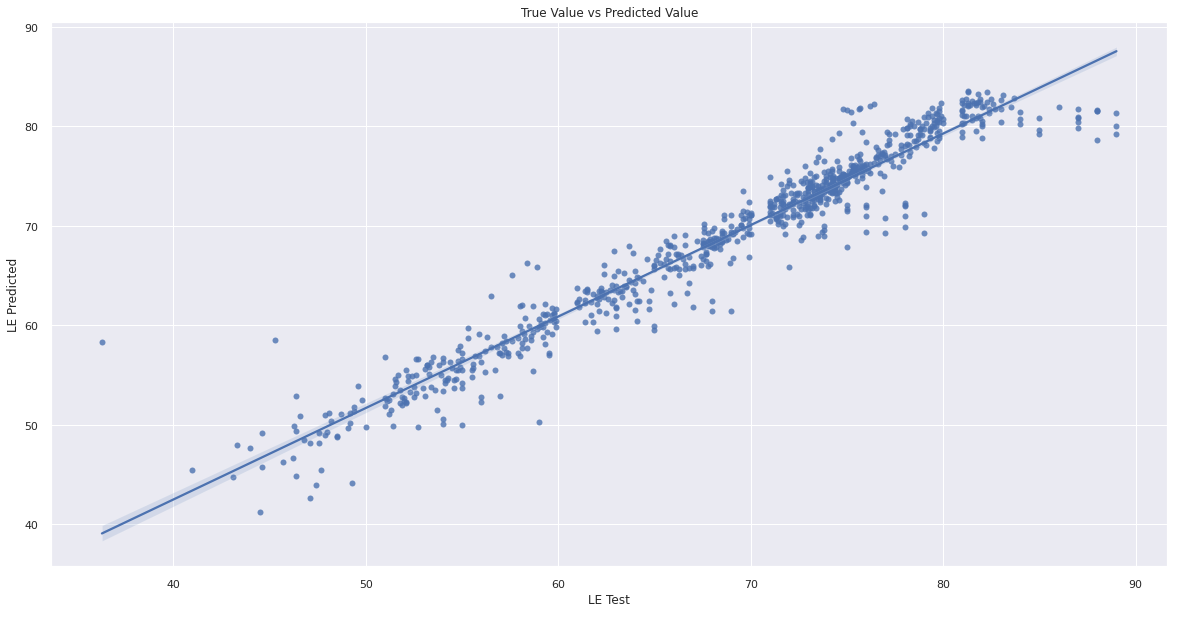
**Step 4**: Applying ML models:

First step before applying any ML algorithm is splitting the dataset in training and testing sets after which algorithms are chosen based on nature of dataset.

In this case, we have used the following algorithms:

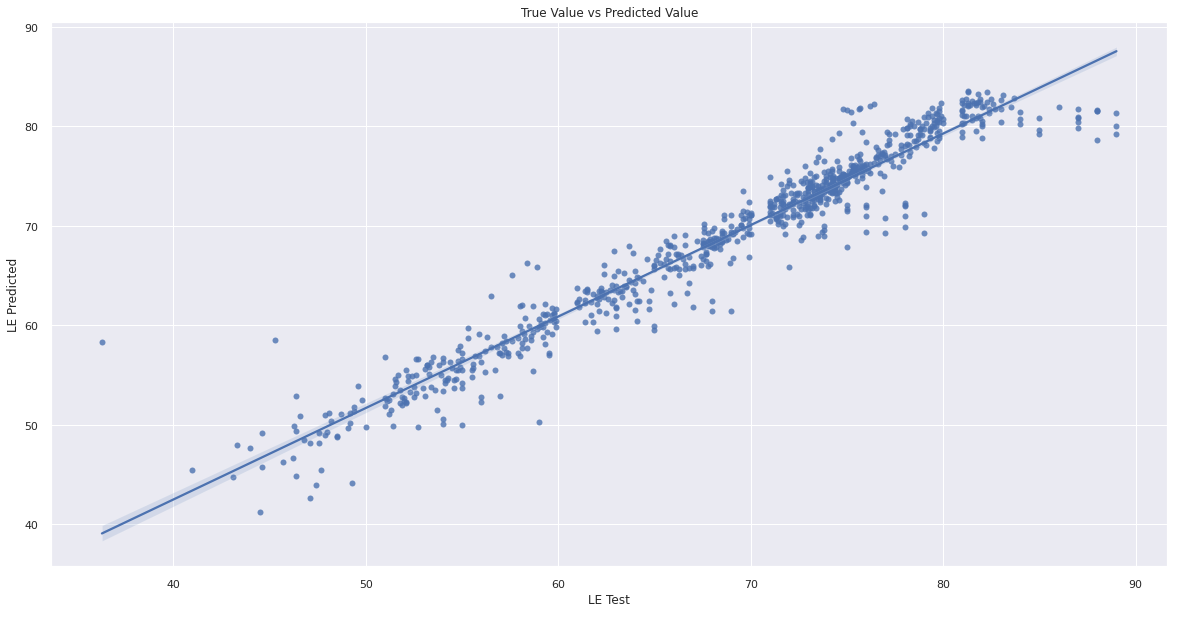
1. Linear Regression:

Linear regression analysis is used to predict the value of a variable based on the value of another variable.



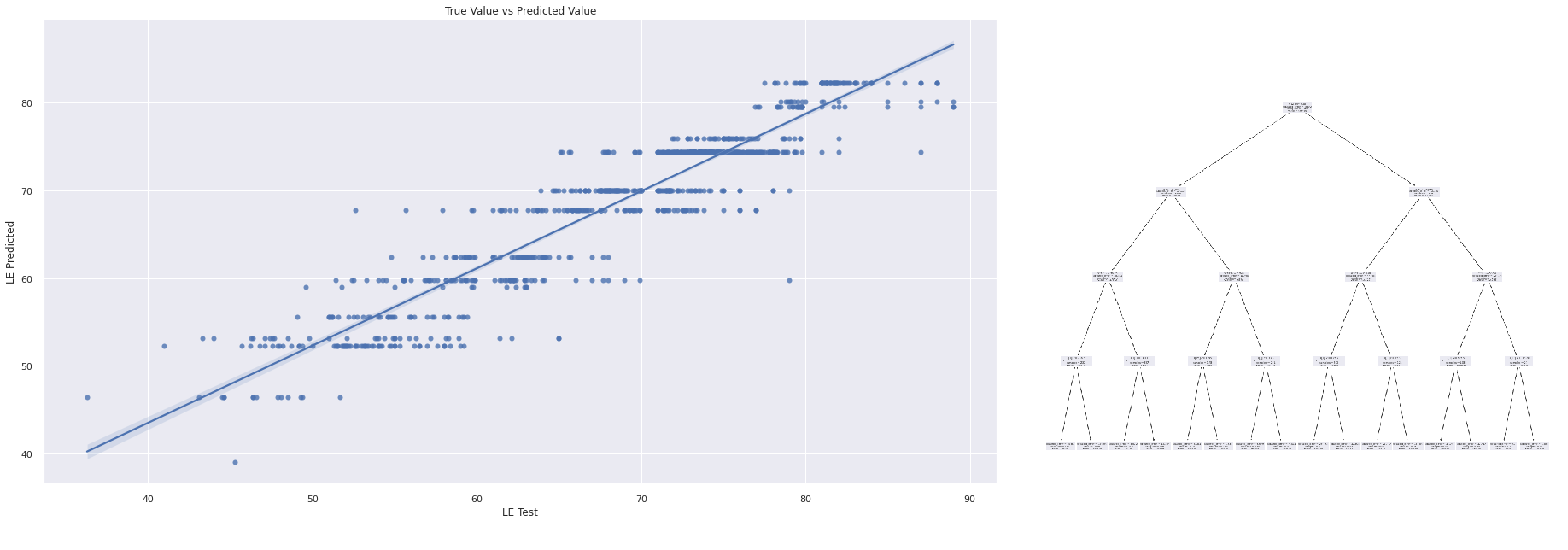
1. Ridge Regression:

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated.



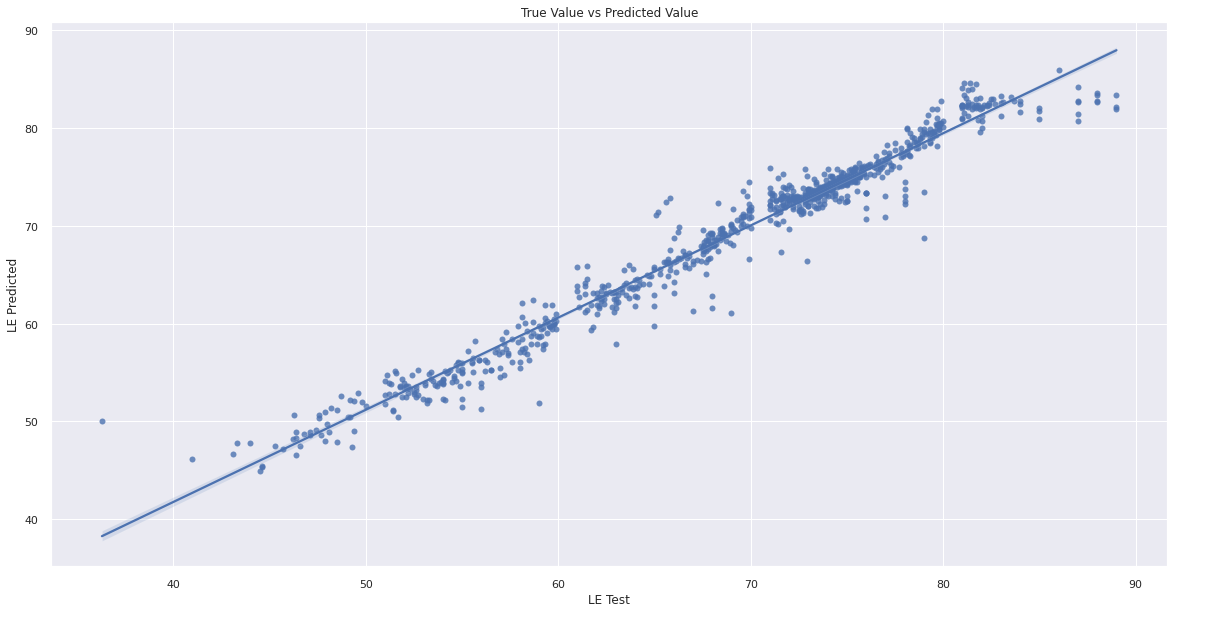
1. Decision Tree:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.



1. Random Forest:

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.



In order to check the accuracy of all the models, the following test metrics were used:

1. Mean Absolute Error:

Mean Absolute Error refers to the magnitude of difference between the prediction of an observation and the true value of that observation.

1. Mean Squared Error:

Mean Squared Error measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

1. R2 score:

The R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.