CLUSTERING AND FITTING

Introduction:

This research study is based on clustering and fitting of dataset. In this task a dataset of forest area has been selected for clustering and model fitting that includes different countries and its forest area based on the year from 1990 to 2021. In this part the target variable "Forest area" has been selected from the given dataset on which the "Linear regression", "K means clustering" and "Lasso Regression" have been performed to obtain the accuracy score of the data.

Methodology:

The entire research is based on clustering methods performed using machine learning algorithms to obtain the accuracy score of the model. The dataset of forest area is to be imported using required codes so that the clustering can be performed. In this study, data visualisation are also performed so that the relationship among the variables can be understood. After the dataset has been imported the required libraries have also been imported for certain purposes. The dataset has been preprocessed before implementing the required algorithms for clustering. In this case, data preprocessing included null value checking, and all the bull values have been filled so that the accuracy of the model becomes increased. Then the k-means clustering method has been implemented. Then the lasso regression and linear regression have been implemented along with two data visualisation techniques such as scatter plot, and bar plot.

Results and Analysis

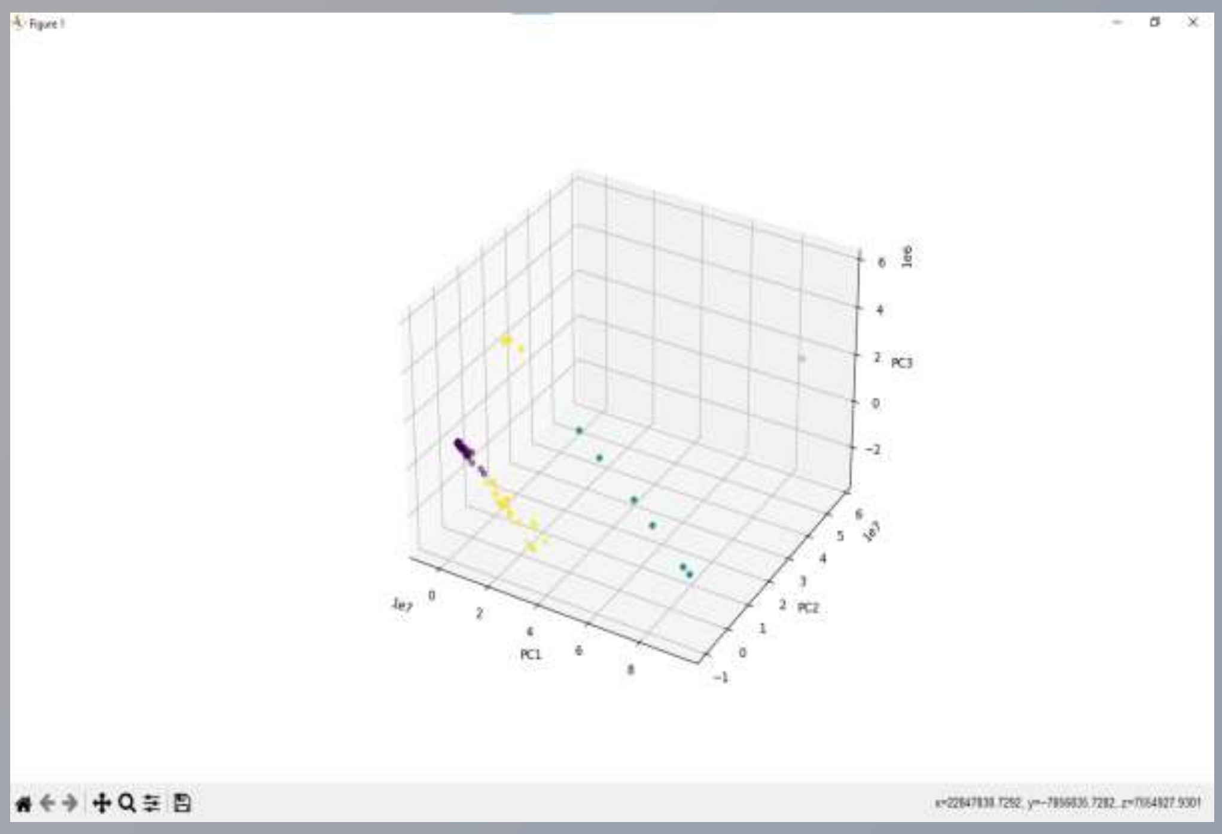

Figure 1: 3D scatter plot of K-means clustering

(Source: Generated on jupyter notebook)

This above figure has shown the 3D scatter plot of pc1, pc2, and pc3 are obtained after implementing data visualisation technique.

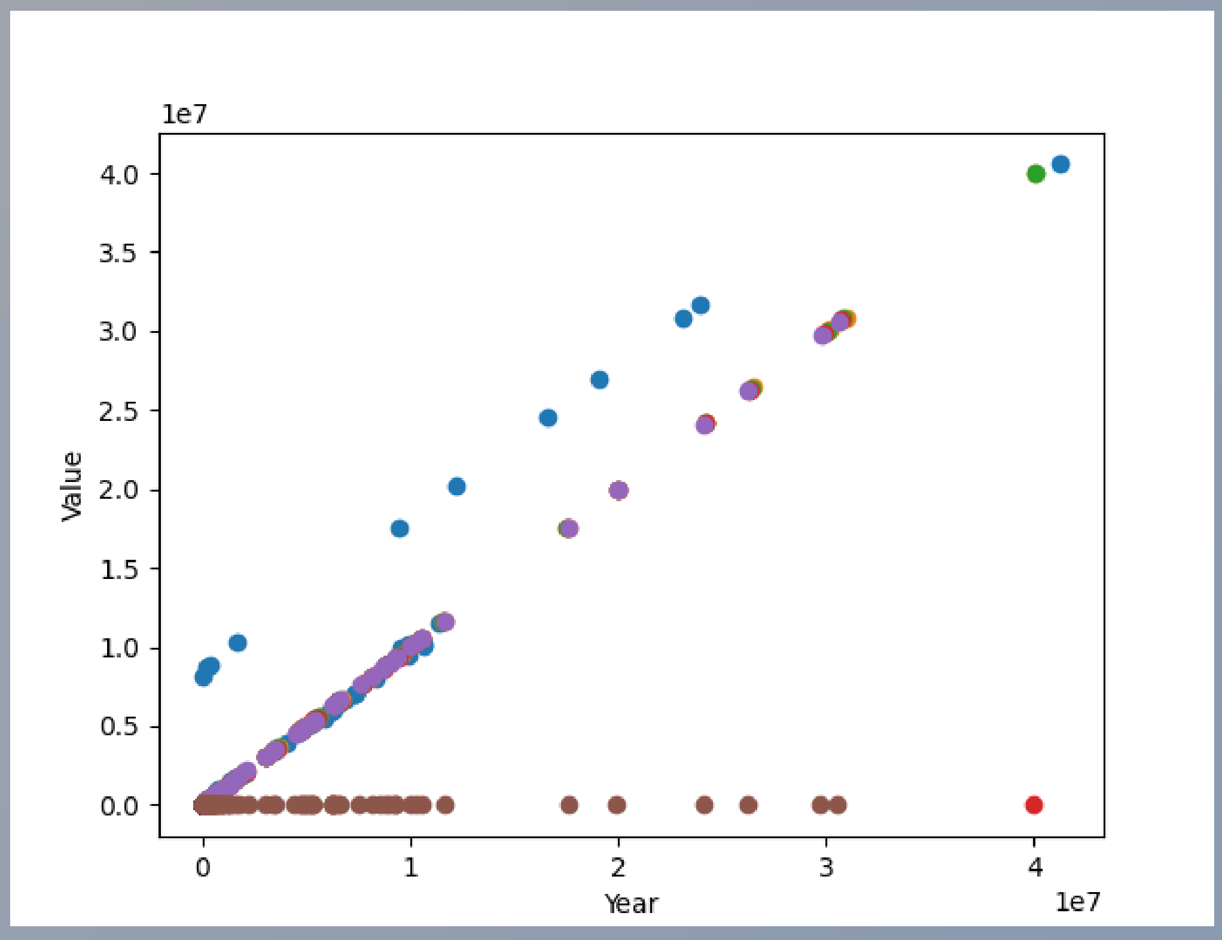

Figure 2: Scatter plot of value based on year

(Source: Generated on jupyter notebook)

This figure shows that the scatter plot has been obtained within the variable "value" of forest area based on the year from 2012 to 2021.

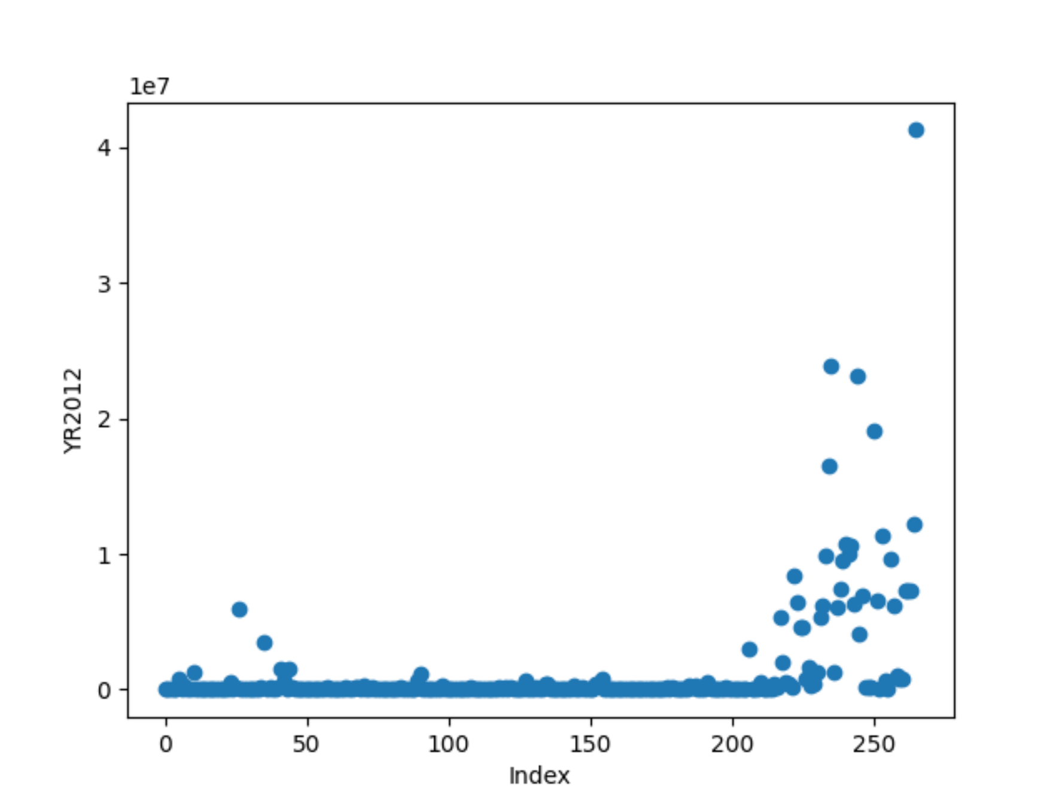

Figure 3: Scatter plot of year 2012 vs Index of forest area

(Source: Generated on jupyter notebook)

This above figure has shown the scatter plot obtained for the particular year 2012 based on the variable index.

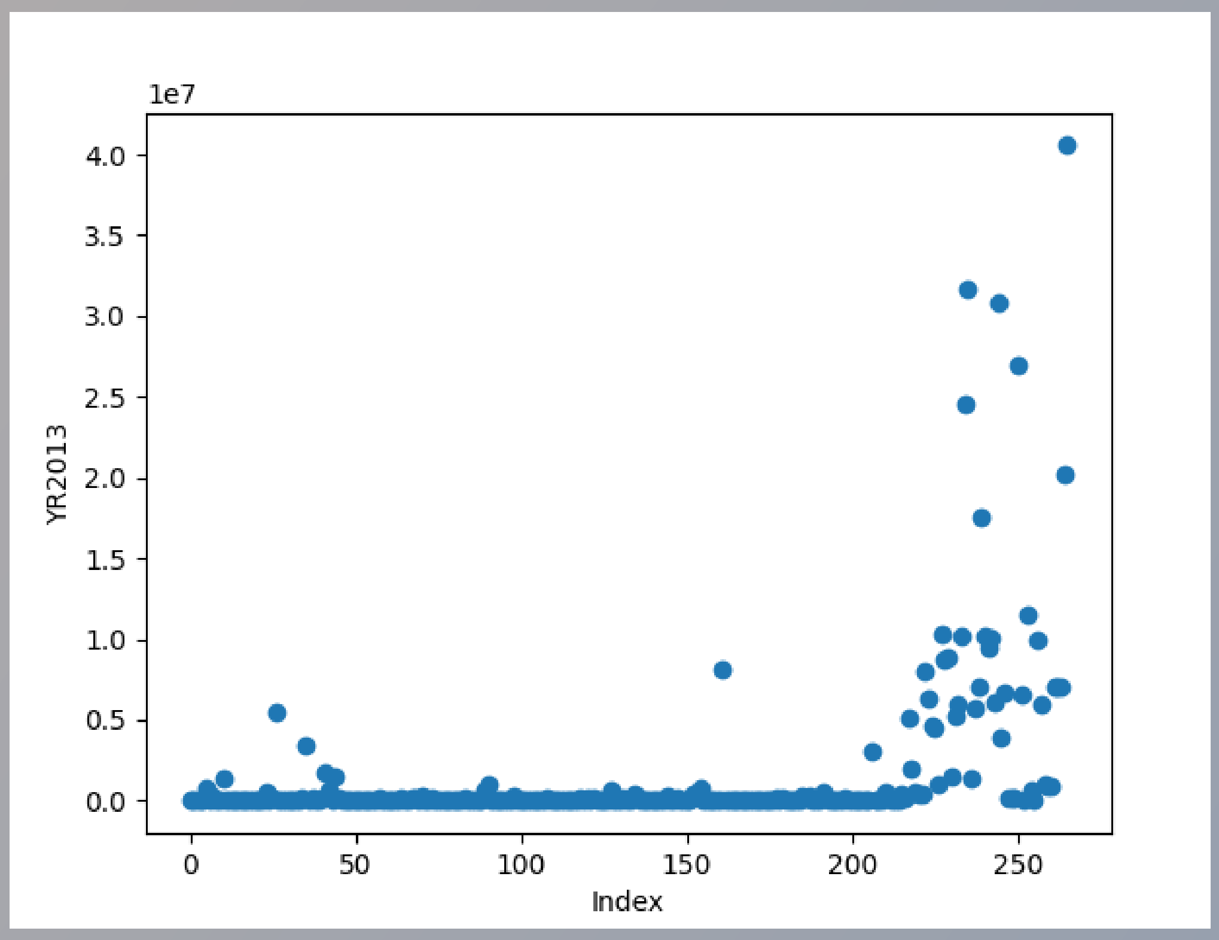

Figure 4: Scatter plot of the year 2013 vs Index of forest area

(Source: Generated on jupyter notebook)

This figure also shows the index value of forest area has been obtained using data visualisation technique for the year 2013.

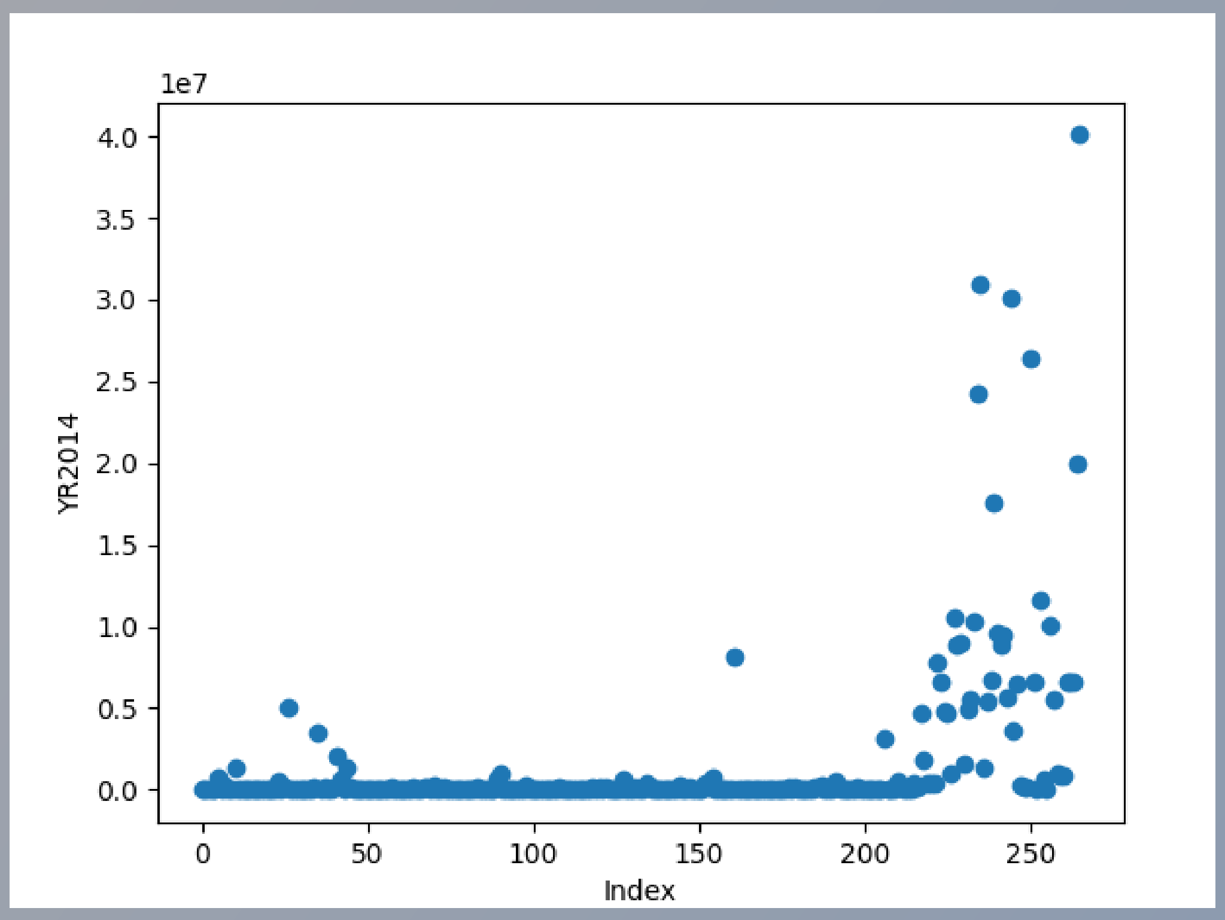

Figure 5: Scatter plot of year 2014 vs Index of forest area

(Source: Generated on jupyter notebook)

The above figure has also been implemented for getting the index value of the forest area dataset depending on the year 2014.

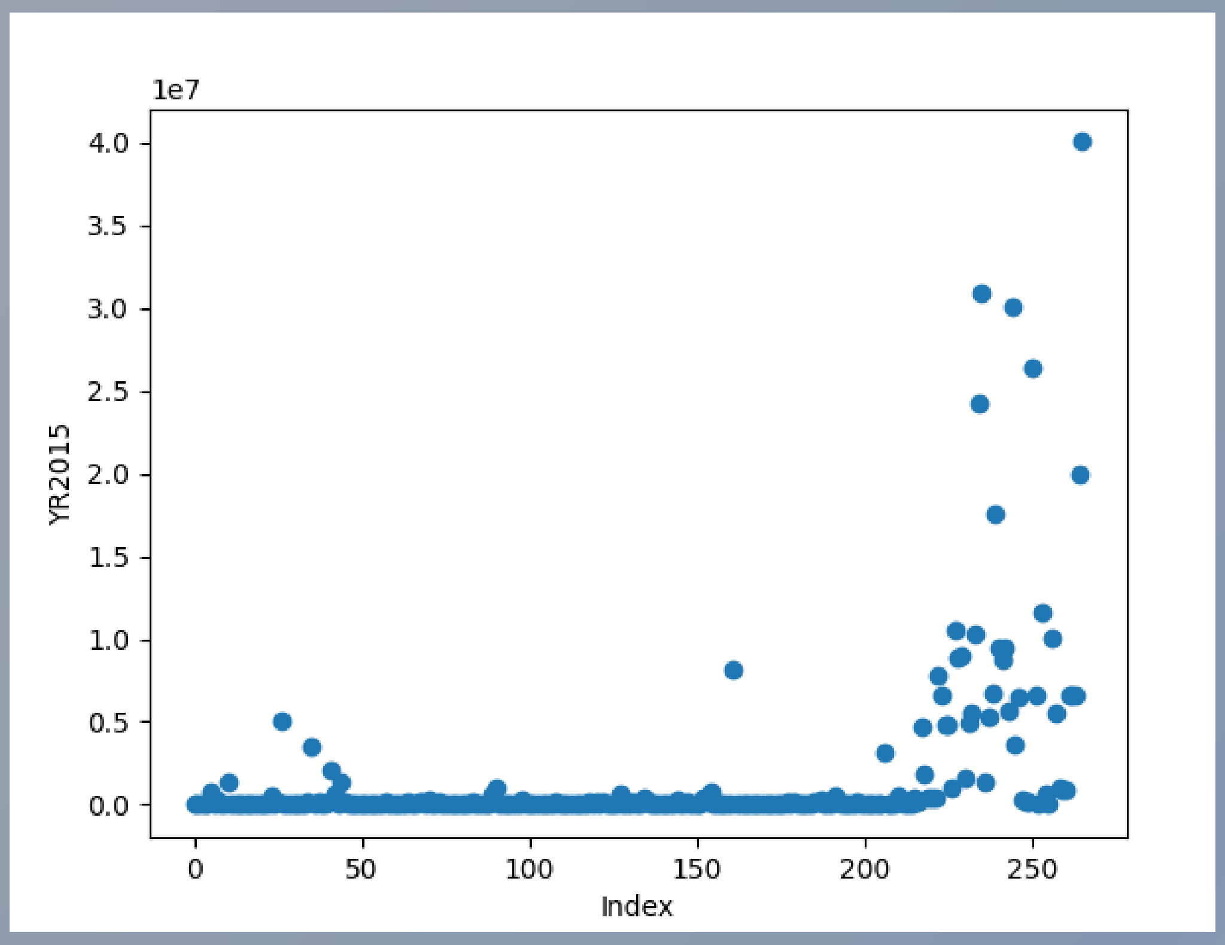

Figure 6: Scatter plot of year 2015 vs Index of forest area

(Source: Generated on jupyter notebook)

The above figure also obtained after implementing data visualisation technique of scatter plot to get the "index" value of the dataset for the year 2015.

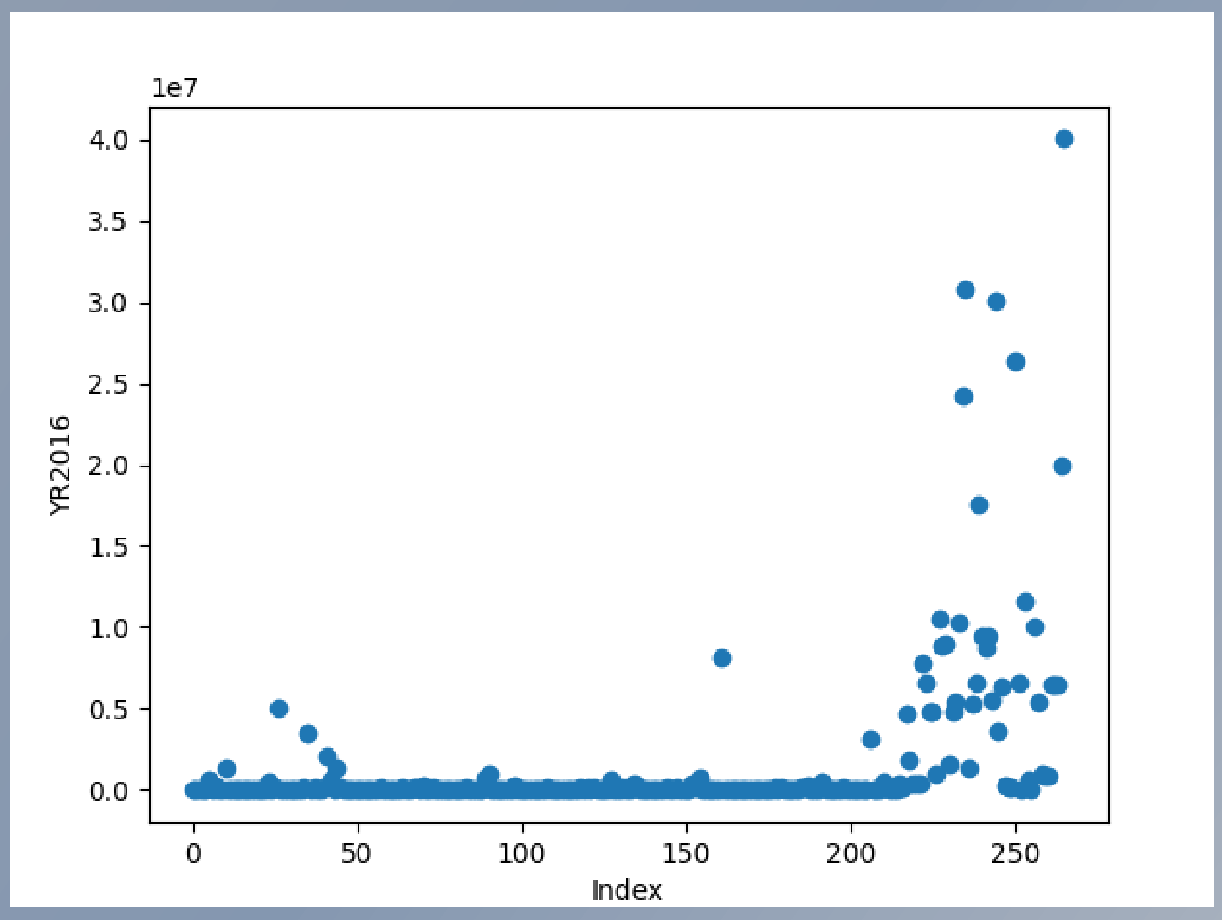

Figure 7: Scatter plot of year 2016 vs Index of forest area

(Source: Generated on jupyter notebook)

From this figure the index value of 2016 of the forest area dataset has been obtained using scatter plot visualisation technique.

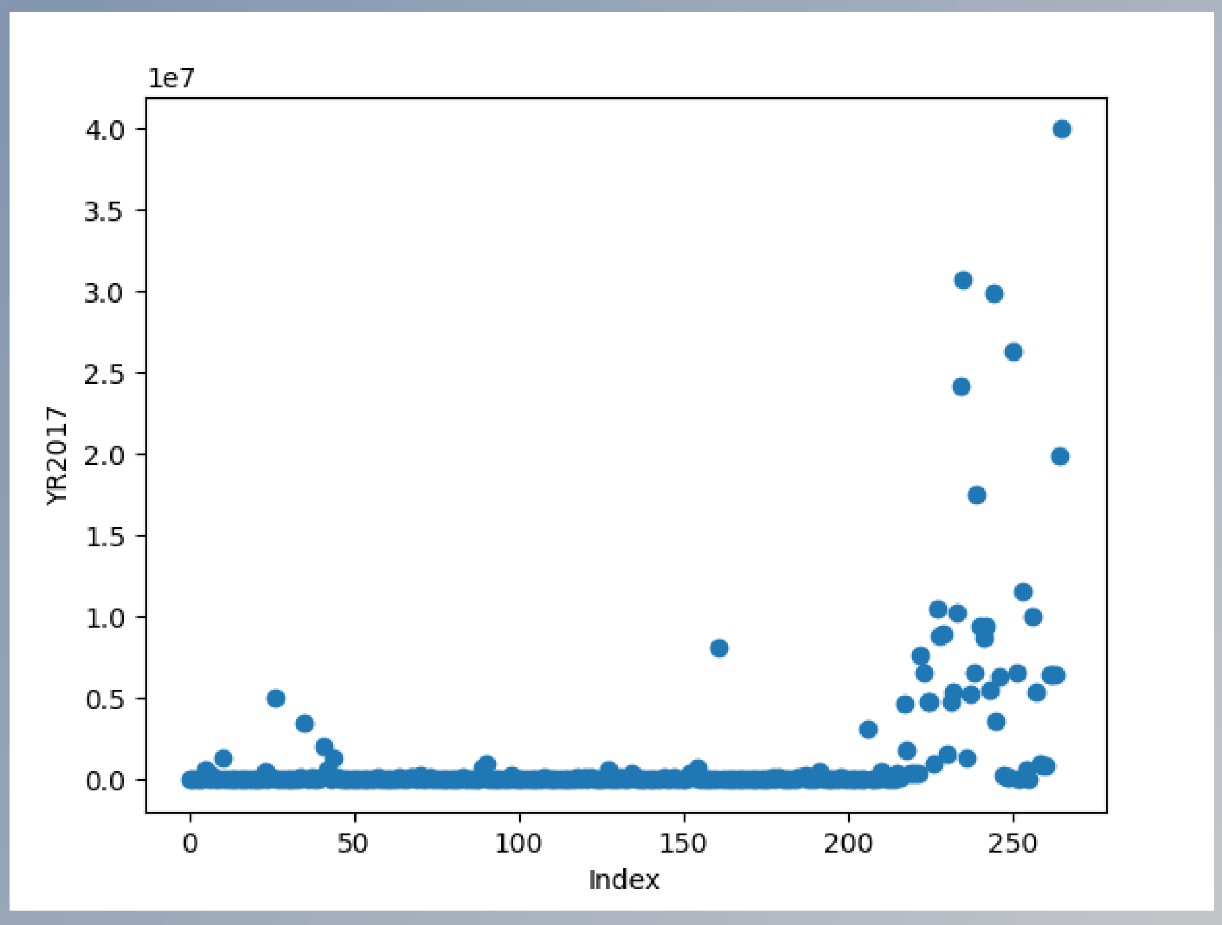

Figure 8: Scatter plot of year 2017 vs Index of forest area

(Source: Generated on jupyter notebook)

It is the plot of index of the year 2017 of the given dataset that has been plotted to visualise.

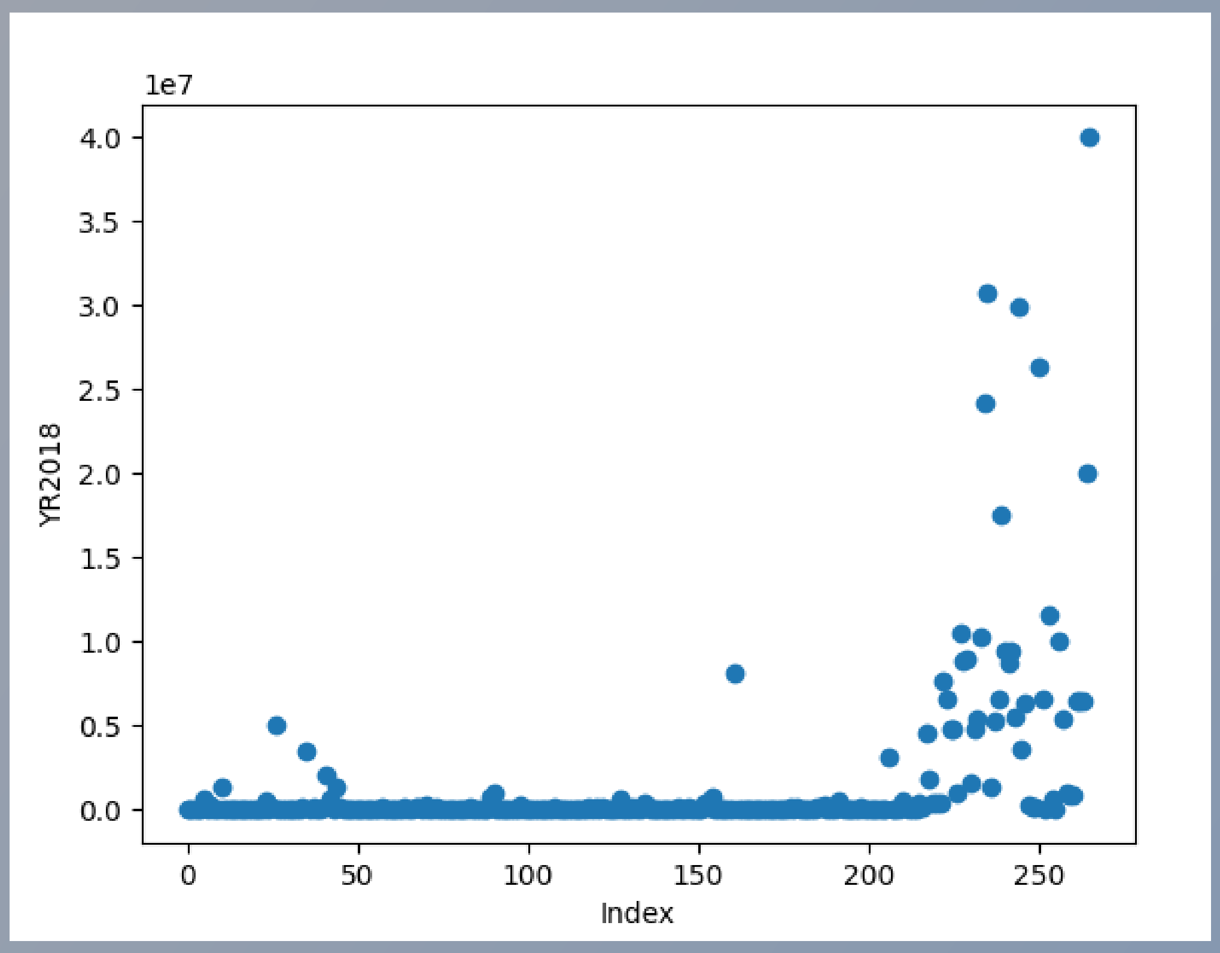

Figure 9: Scatter plot of year 2018 vs Index of forest area

(Source: Generated on jupyter notebook)

The above figure has shown the scatter plot of visualisation of forest area index of the year 2018.

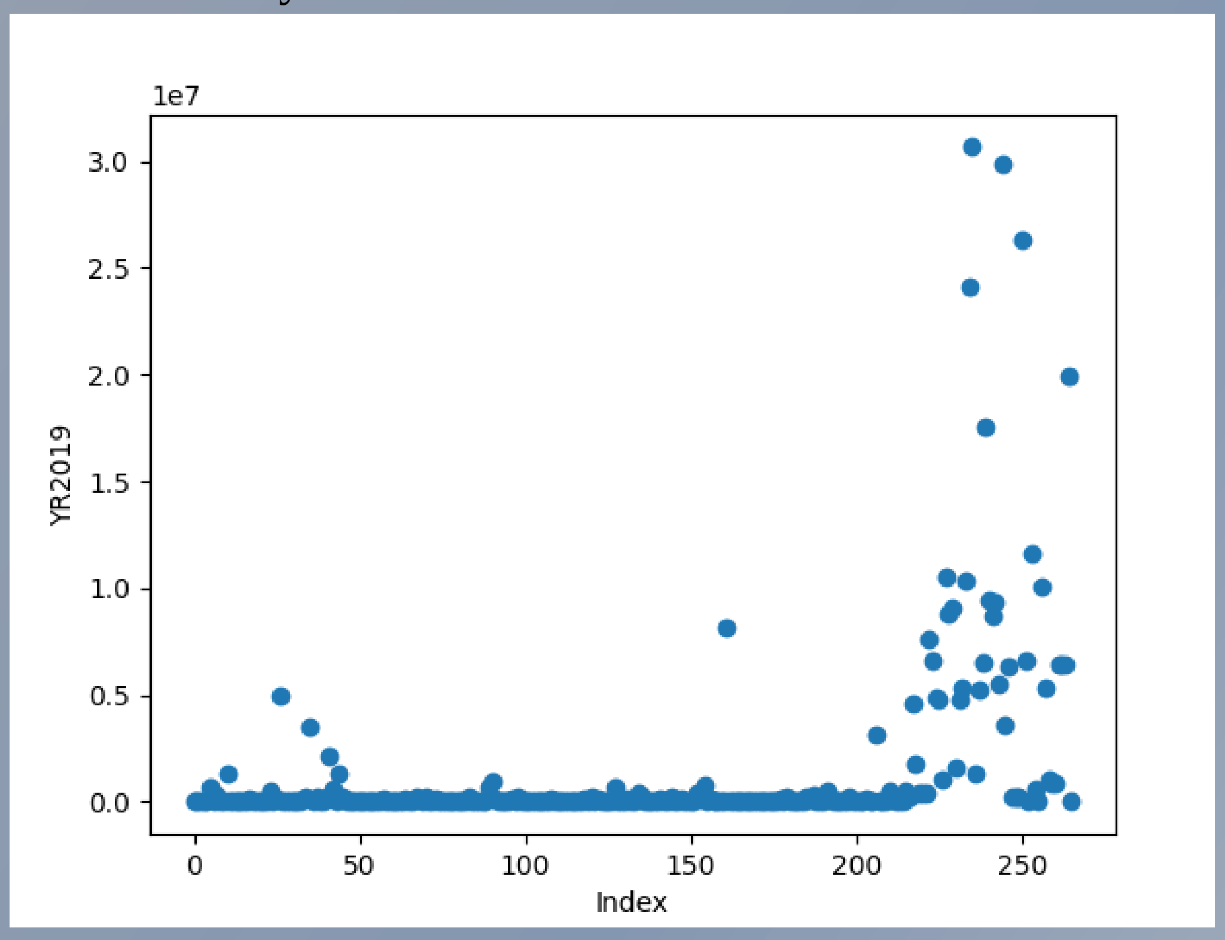

Figure 10: Scatter plot of year 2019 vs Index of forest area

(Source: Generated on jupyter notebook)

This also shows the similar plot of the similar variables but for the year of 2019.

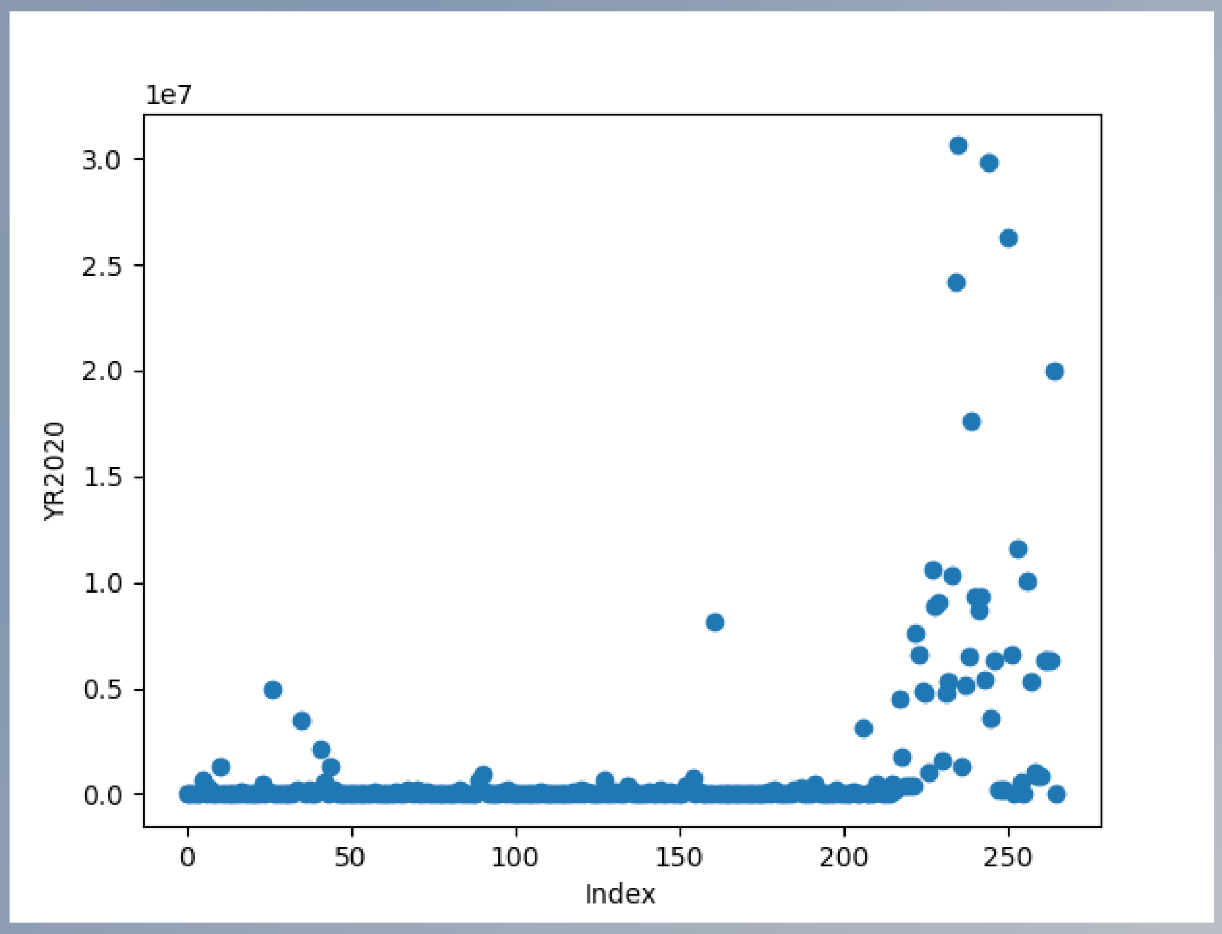

Figure 11: Scatter plot of year 2020 vs Index of forest area

(Source: Generated on jupyter notebook)

This figure has shown the scatter plot of "index" vs "year" of 2020 to show the relationship among these two variables.

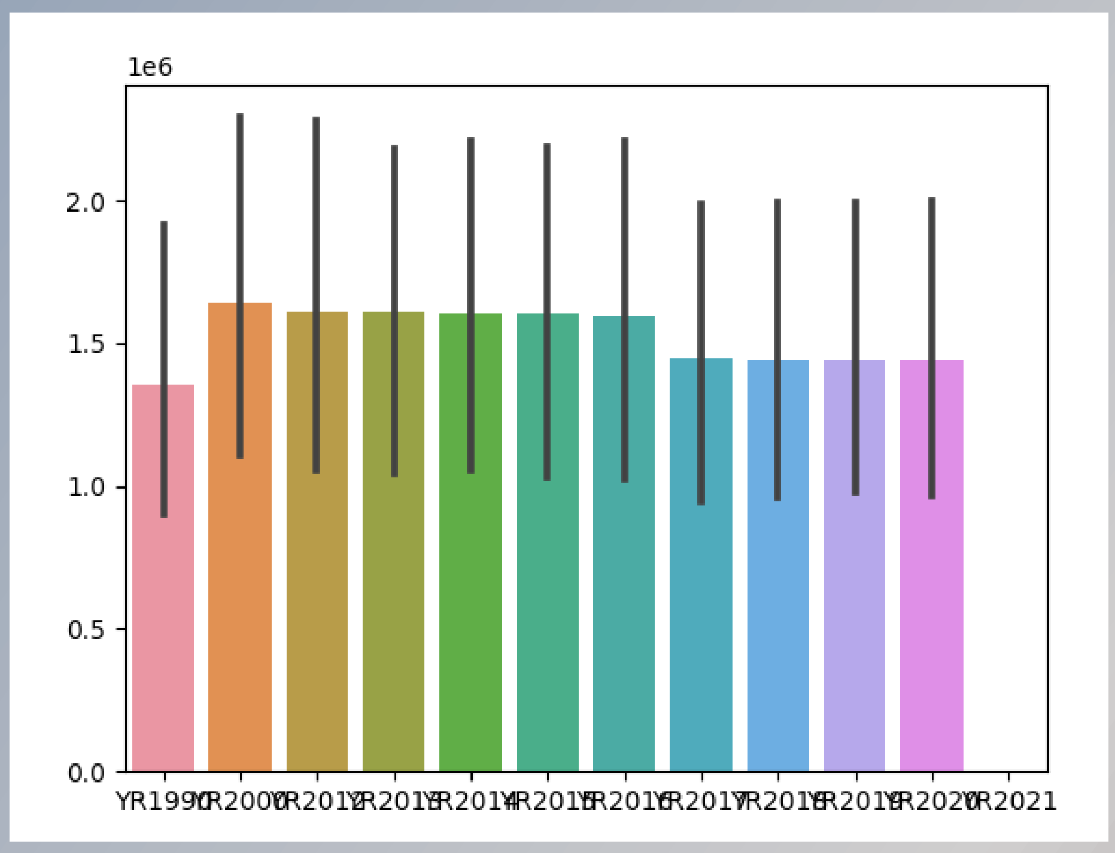

Figure 14: Bar plot of year vs Index of forest area

(Source: Generated on jupyter notebook)

This figure has shown the bar plot implemented on the dataset for visualisation. In this case the plot has been obtained depending on forest area of the year 1990, and from 2012 to 2022.

Conclusion:

This research study is entirely performed using machine learning algorithms of clustering such as linear regression, k-means clustering, and Lasso regression. The dataset of forest area has been chosen to visualise the plot of data variables for the year 2012 to 2022. The scatter plot and bar plot have been performed to see the relationship within two variables of the dataset. Based on this study it can be concluded that the main thing that should be done in the clustering and model fitting using machine learning is preprocessing of the dataset. Since the dataset consists of null values, it could lead to undesired visualisation or the inaccuracy in model fitting.

GitHub Link:

https://github.com/Salehabailey/Clustering-fitting