□ DATA AND DATA ANALYSIS

Income Inequality findings from Pakistan Data:

Income inequality has been a persistent issue in Pakistan, despite its economic growth in recent years. This brief report highlights the main trends and factors contributing to income inequality in the country using data up to 2021.

Key Findings:

A measure of income inequality known as the Gini coefficient has revealed that Pakistan's income distribution is unbalanced. The Gini coefficient, which measures income inequality, was 0.41 in 2021. A Gini number of 0 denotes perfect equality, while a Gini coefficient of 1 indicates the greatest degree of inequality.

Income Distribution:

The richest 20% of people make up more than 50% of the nation's total income, while thepoorest 20% make up less than 5%. Over time, the rich have gotten richer and the poor have gotten poorer, widening the income gap.

Urban-Rural Divide:

Pakistan has a discernible income disparity between urban and rural areas. Because there are more job opportunities, resources, and social services available in urban areas, the average income there is significantly higher than it is in rural areas.

Education and Income:

In Pakistan, a person's salary is mostly determined by their level of education. Higher education is typically associated with better paying jobs and more stable income. For those whoare economically disadvantaged, illiteracy and a lack of quality education worsen income disparity.

Gender Wage Gap:

In Pakistan, women make significantly less money than males, which contributes to wealth disparity. Cultural traditions, women's lower labor market participation rates, and occupational segregation all contribute to the gender wage gap.

Data:

The dataset used for this research is taken from the World Income Inequality Database (WIID). The World Income Inequality Database (WIID) presents information on income inequality for developed, developing, and transition countries. It provides the most comprehensive set of income inequality statistics available.

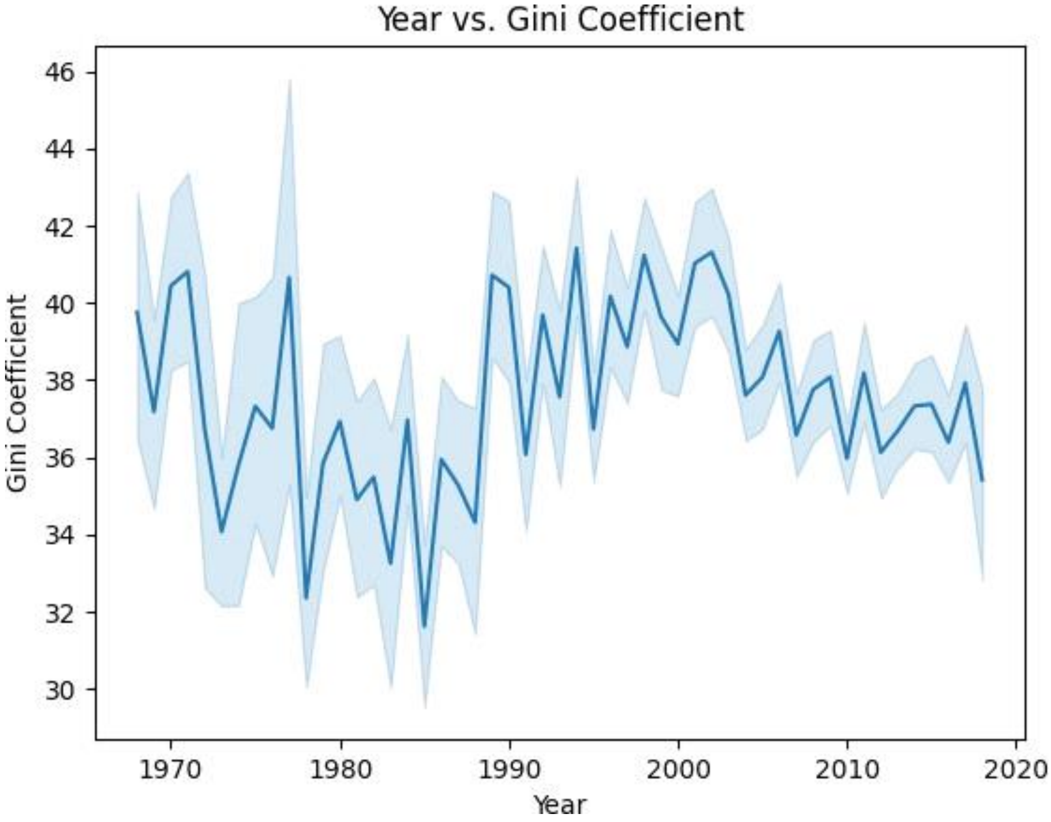
Data Preprocessing:

The dataset we were analyzing contained numerous columns that were redundant or irrelevant to our research, as well as many missing values that could have impacted the accuracyof our analysis. In order to overcome these challenges, we meticulously selected only the pertinent columns for our study, which included country, year, Gini reported, Palma, and population. Both Palma and Gini reported are widely used measures of income inequality, with the former focusing on the income share of the top 10% versus the bottom 40%, and the latter ranging from 0 to 1 with 0 representing perfect equality and 1 indicating perfect inequality. By doing this, we ensured that our dataset was streamlined and devoid of unnecessary variables.

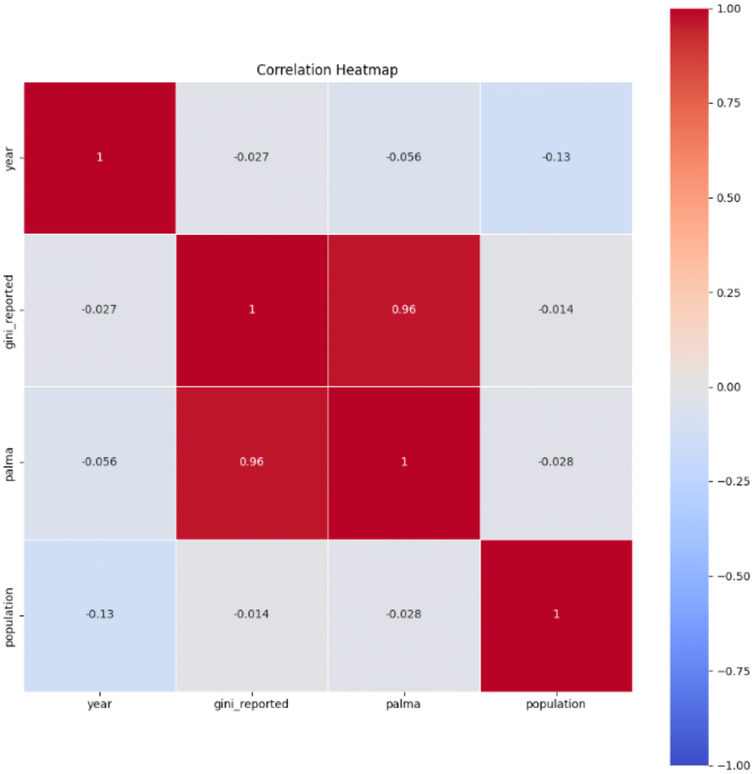
To ensure the accuracy of our analysis, we removed all rows containing missing values from our dataset. We then employed the interquartile range (IQR) method to detect and eliminateany potential outliers. Any data points that were found to be outside of 1.5 times the IQR were deemed outliers and excluded from the dataset. This enabled us to produce reliable and precise results, unimpeded by any anomalies that could have skewed our findings.

Data Visualization

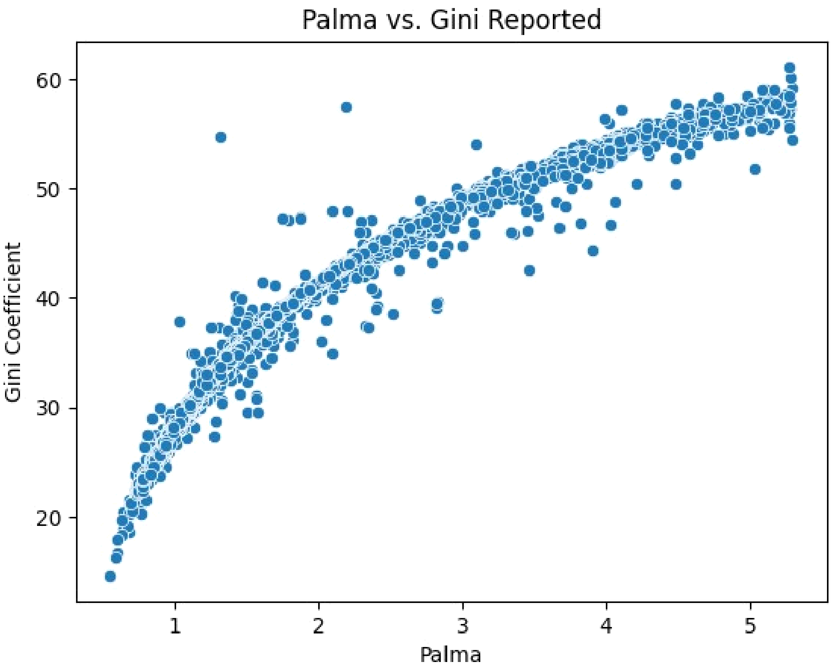
- The graph of year vs Gini Coefficient is shown. It explores how the trend of Ginicoefficient changes over time.



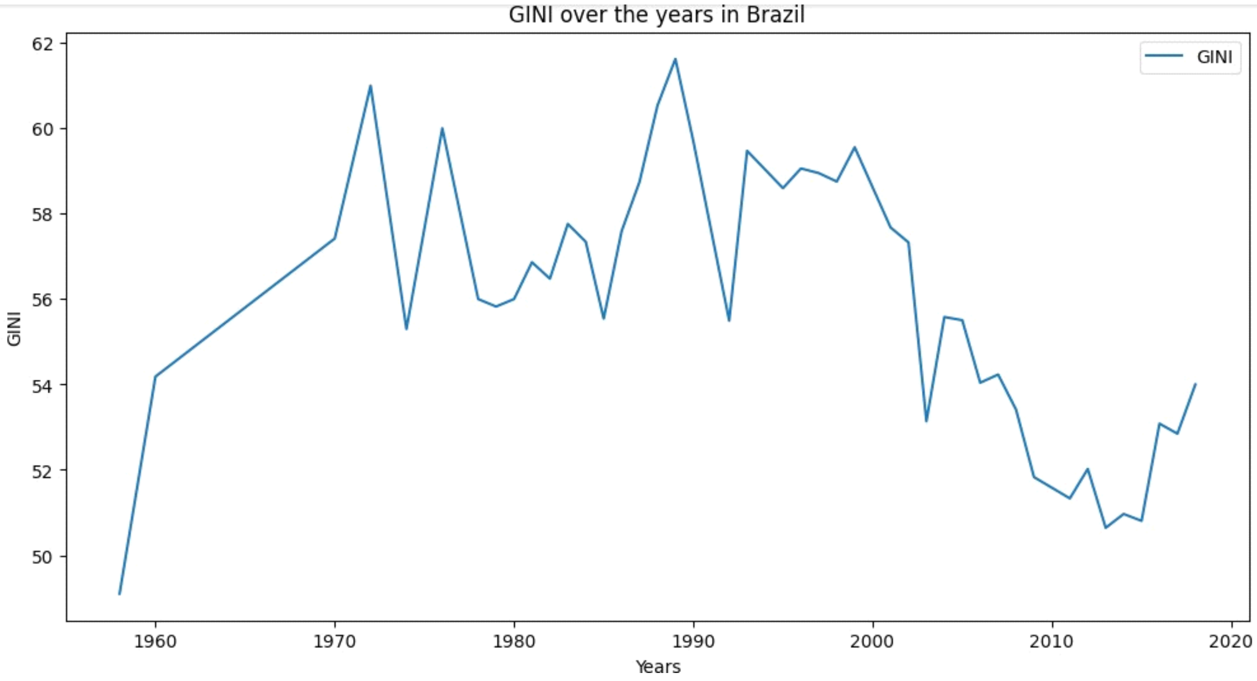
- The correlation heatmap of the columns in the database is shown. Notice that there is a close correlation only between Palma and Gini reported. Therefore, we will use these twocolumns to train our machine learning models.



- The graph of Palma Vs Gini Reported is shown. From the graph it can be seen that the graph follows a close parabolic relation. From the graph it can be seen that as the Palmarises, the value of Gini coefficient also rises. It rises more steeply at the start but eventually the slope of the graph becomes lesser and lesser.

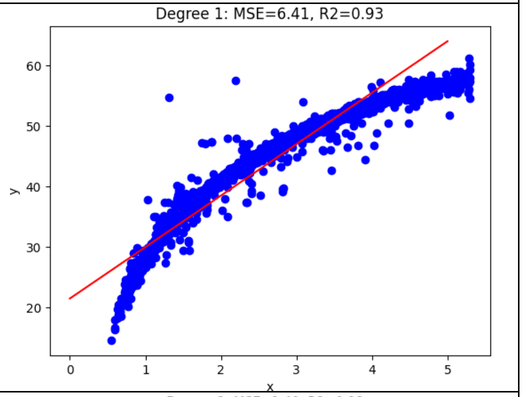
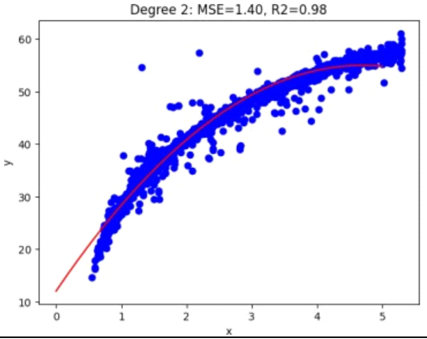
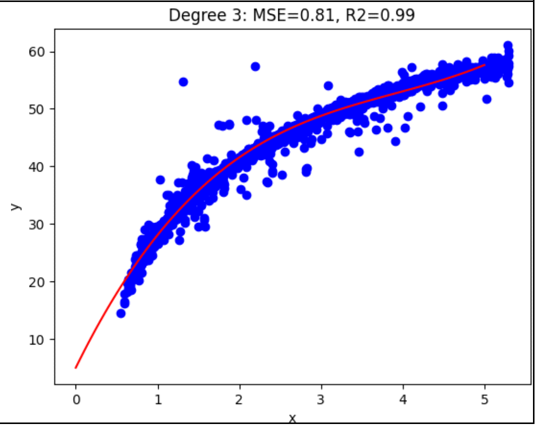


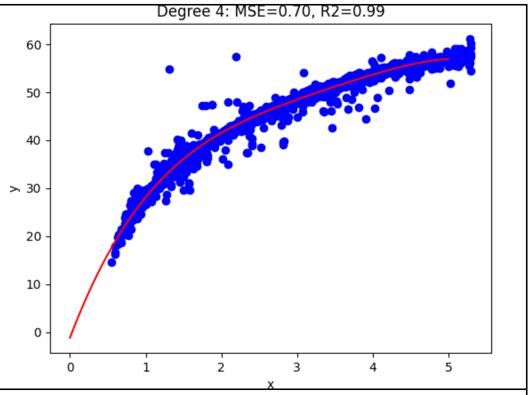
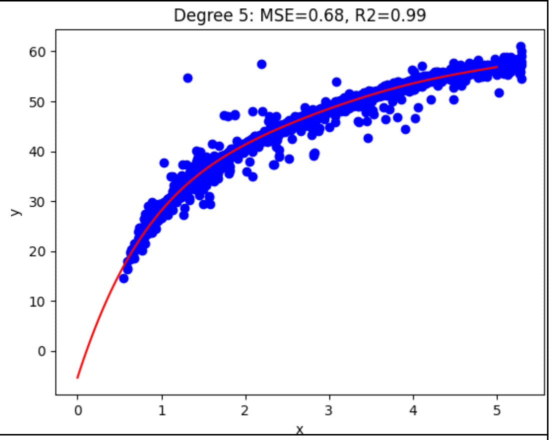
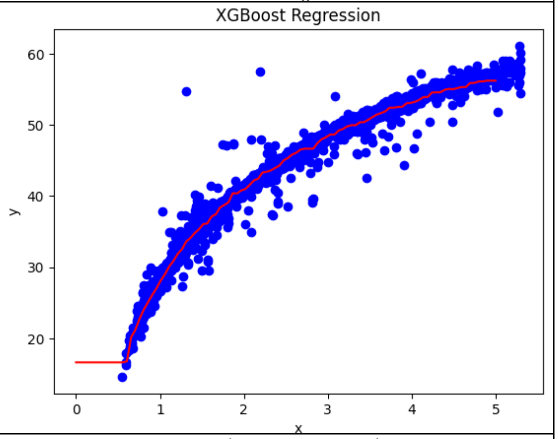
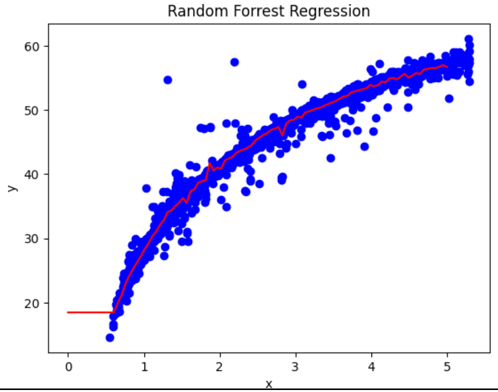
- An example graph of Brazil's Gini index over the years is shown:

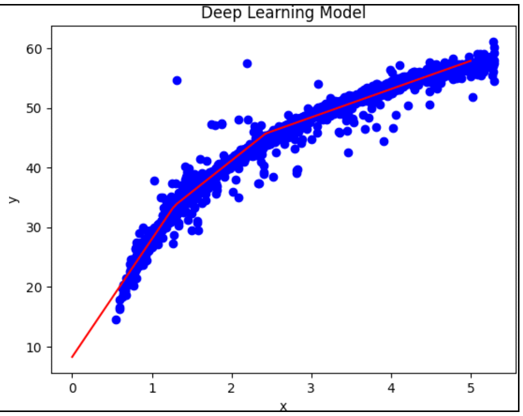


- **Machine Learning:**

From the correlation matrix it can be seen that the Palma and Gini Index follow a close correlation, therefore, we used these two variables for our machine learning models. We used several models. Since we want to predict numerical values of Gini coefficient, we should minimize the mean squared error to assess the performance of each model. The performance ofeach model is as follows:

Model	**Mean-Squared Error**	**Performance Graphs**
Linear Regression Order 1	6.405676091820359	
Linear Regression Order 2	1.395597037632316	
Linear Regression Order 3	0.8092766593903405	
Linear Regression Order 4	0.6965922304001532	

		
Linear Regression Order 5	0.6787099743056815	
XGBRegressor	0.7634068933586372	
Random Forest Regressor	0.747558603002395	

Deep Learning Model	0.8522720963282366	

The goal is to minimize the mean squared error. Linear Regression model of order 5 had the least mean-squared error of 0.6787099743056815, therefore, it is the best model to use in thiscase.