Studying Income Inequality

by Suhas Gumma

Submission date: 22-Apr-2020 05:29PM (UTC+0530)
Submission ID: 1304404528
File name: STUDYING_INCOME_INEQUALITY_2.docx (1.33M)
Word count: 1974
Character count: 9785

STUDYING ECONOMIC INEQUALITY

A Project Report

Submitted by

SUHAS GUMMA - 1700278C203.
M NAVANEETH NANDA-1700244C203.
P SRINIVAS PRANAY-1700262C203
KOKA SIVA PRAKASH - 1700223C203

Batch - 2017 CSE 2

Course
MACHINE LEARNING and DATA MINING - CSE2702

SCHOOL OF ENGINEERING & TECHNOLOGY
BML MUNJAL UNIVERSITY

ACKNOWLEDGEMENT

This project is all about the income inequality between the countries and this has knowledge how the economy is depended on every single individual in a country. This project has given us a great opportunity to gain good practical experience.

We would like to sincerely thank **Bernie Sanders**, the US senator for showing the US and the world that Economic Inequality is a real problem and We CAN solve it by CHANGING Little things Mindfully with the help of **DATA**.

WE SINCERELY THANK OUR COURSE FACULTY DR. ATUL MISHRA TO GIVE THIS OPPORTUNITY FOR DOING THIS PROJECT AND ALSO FOR HELPING OUT WHENEVER HIS ASSISTANCE IS NEEDED EVEN WHEN HE IS BUSY WITH HIS DUTIES.

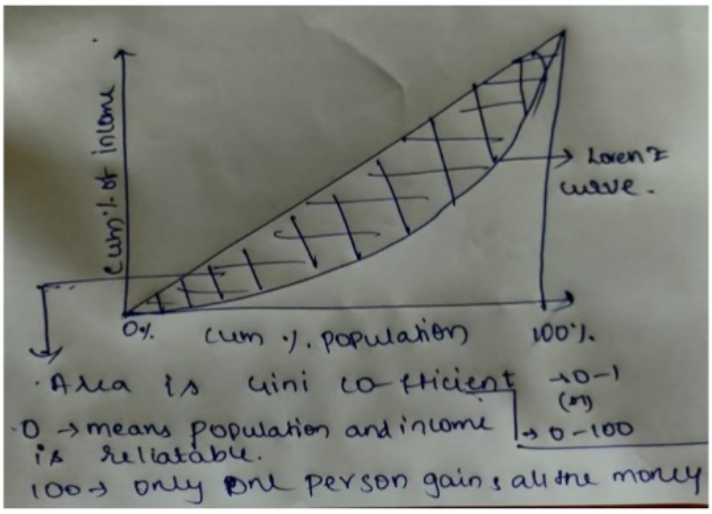
Abstract *– Economic Inequality is a very complex issue as it depends on many factors. Many economists and politicians try to change many of those factors which are not practically feasible. In this project, we concentrated on some important factors that can be easily changed and can have drastic effects on Income Inequality. The project focuses on developing a simple regression model. The independent factors are Investments percentage in GDP made by the government, Tax Revenue percentage in GDP, Taxes on different classes based on income and per capita Income of a country.*

Keywords –

- *Artificial Neural Networks*
- *Economic Inequality*
- *GDP*
- *Gini Index*
- *Investments percentage in GDP made by the government.*
- *Left Wing Political Party and Right Wing Political Party.*
- *Multiple Linear Regression.*
- *Progressive Taxation.*
- *Per Capita Income.*
- *Regressive Taxation.*
- *Tax Revenue percentage in GDP.*
- *The 1 percent.*

I. INTRODUCTION

In recent years, Income inequality is growing at a faster rate even in the developed economies. Many argue that it is due to the policies introduced by the government favouring the top 1 percent. It might be a reason but not the only reason. The chief economists and politicians are ignoring the growing income inequality in a zeal to push the economy forward. The project focuses on some simple factors and tries to show how drastically they can affect income inequality. For starters, income inequality is measured by the Gini index. It ranges from 0 to 100. Zero value of the Gini index indicates that everyone is earning the same and hundred indicates the opposite scenario where a single person earns everything. For many decades, there have been many arguments between the left-wing political parties and the right-wing political parties that their respective economic policies are better than the other. Showing proof with the data is a nice way to answer the problem. The other debate going on is about the progressive taxation. As the income of the individual grows, she should pay more percentage of her income. This way of taxation is called progressive taxation. Everyone agrees that the progressive taxation is the best way to collect taxes,but the concern is 'Is it progressive enough?'. If we charge more on the higher class of income, does income inequality decrease? It may seem plain, but it is more complex than it looks.The economy of a country is like a spaghetti. It is entwined with various factors. So, we will cross check if the data will backup the hypothesis that more taxing on the class earning the most will reduce income inequality. One of the factors is tax revenue percentage in GDP. It tells how much the tax revenue is contributing to the GDP. And the final factor is the percentage in GDP made by the government. It may seem out of the context for studying economic inequality,but it also affects the inequality significantly. It is more beneficial to the top 10 percent than the rest.



II. LITERATURE REVIEW

To this project we did a lot of background research on how does GDP impacts the countries and how the GINI index works.And what are the algorithms are needed to use in order to work on this model.We have checked this paper as background to work on it.

1. http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8904917&fileOId=8906382

In this paper, they used the time series data of thirty to find the causes of income inequality in a society. They have done that by using dynamic regression models. They found that they found that affects the income inequality are trade to GDP ratio, wing of the government(left or right) and the employments share in industries

$$\log y_{it} = \varphi \log (y_{it-1}) + \beta\, 0 x_{it} + u_{it}$$
$$i = 1, \ldots, 30 \text{ and } t = 1985, \ldots, 201.$$

- x1 → log(government expenditure /GDP),
- x2 → log(sum of exports and imports / GDP),
- x3 → log(domestic credit to the private sector as share of GDP).
- x4 → industry's share of total employment.
- x5 a dichotomous variable.
  - 1 if left-wing party is in power
  - 0 if right-wing party is in power

2. **The Gini Index and Measures of Inequality. The American Mathematical Monthly.** Gini index is a statistic how uniformly a particular resource is dispersed in the society. One of the examples is income. The lower rates of Gini indicates that the particular resource is distributed almost equally and the higher values indicates that the resource is held by a few people

III. METHODOLOGY

- The project is all about predicting the Gini Index(The measure of Inequality) by using five simple factors and getting to know how they affect the Gini index i.e economic inequality.
- So, we are going to build models and compare them.

1. Collecting the data

- There is no ready made data for the factors choosed. We have to manually search the google and record values in excel for 2-3 factors.

- The other factors were found in kaggle individually.

2. Preprocessing the data

The main challenge with the collected data is that it has a considerable number of missing values. One way of dealing with it is to ignore the entire row. It would be an option if the data was huge.But, it's not the case in this situation. The best way to deal with the missing values is to fill them by the mean of the whole column. The process is called impute.

There are various options to impute the missing values. In this project we used SimpleImputer class from sci-kit learn. We've chosen it because it is simple and reliable and it can impute all the data at once.

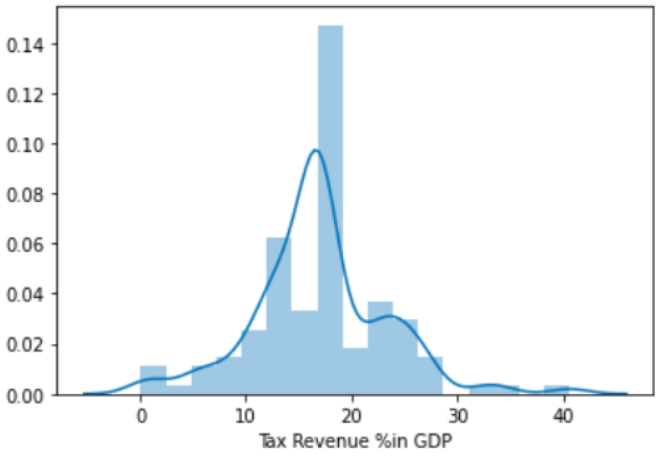
```
from sklearn.impute import SimpleImputer

#Create an object of SimpleImputer class
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')

imputed_data = data
imputer.fit(imputed_data)

#Returns a numpy array
imputed_data = imputer.transform(imputed_data)

imputed_data = pd.DataFrame(imputed_data)
```

- It would be better if you normalized the data. The entire data falls between 0 and 1. So it would be easy to scale, especially when you are doing regression. It's fine even if the normalization is not done to the data.
    - **data = data-min(data)/(max(data) - min(data))**

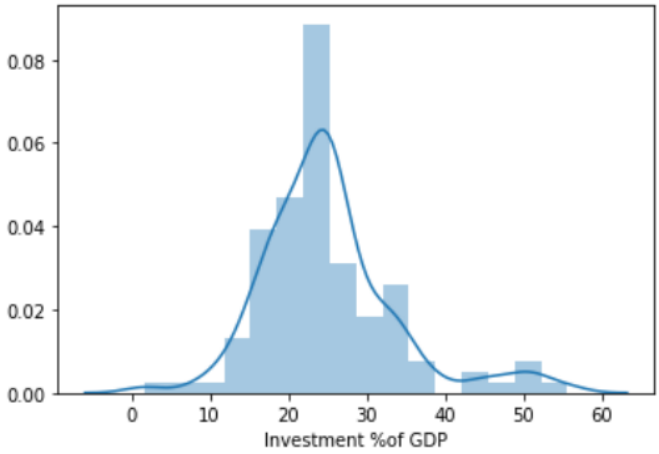
- **Split the data into Train data and test data**
    - If there is one thing you should not forget about the regression,it is overfitting. You might think that you built the perfect model, but instead of learning from the data, the model literally remembers the data. This is called overfitting. To be aware of this, we keep some data data aside for testing.

3. Statistical Analysis of the data(Mainly Descriptive Statistics)

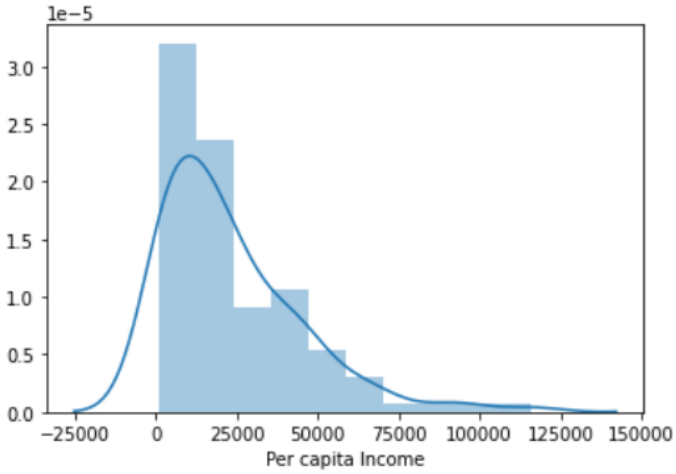
- **Percentage of Tax Revenue in GDP**
    - **Mean(μ) = 16.90**
    - **Standard Deviation(σ) = 6.31**



- **Investment % in GDP**
    - **Mean(μ) = 25.14**
    - **Standard Deviation(σ) = 8.74**

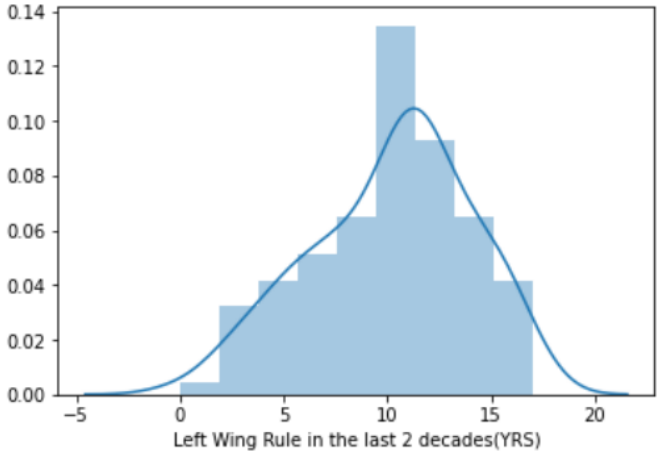


- **PerCapita Income**
    - **Mean(μ) = 23396.52**
    - **Standard Deviation(σ) = 21705.98**

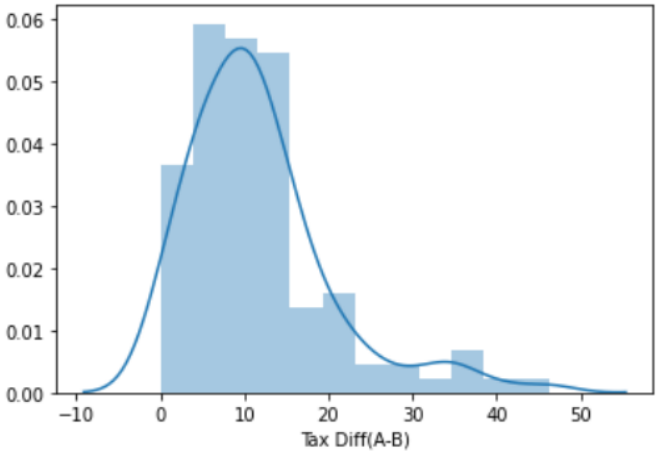


- **Number of years the left wing party in power over the last 2 decades**
    - **Mean(μ) = 10.18**
    - **Standard Deviation(σ) = 3.82**

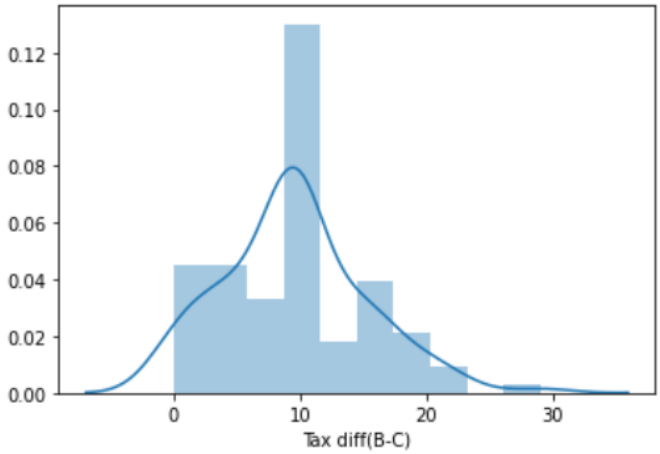
Studying Income Inequality



- **Difference between Taxes of class A(High) and Class B(Middle)**
  - **Mean(μ) = 11.6**
  - **Standard Deviation(σ) = 8.6**

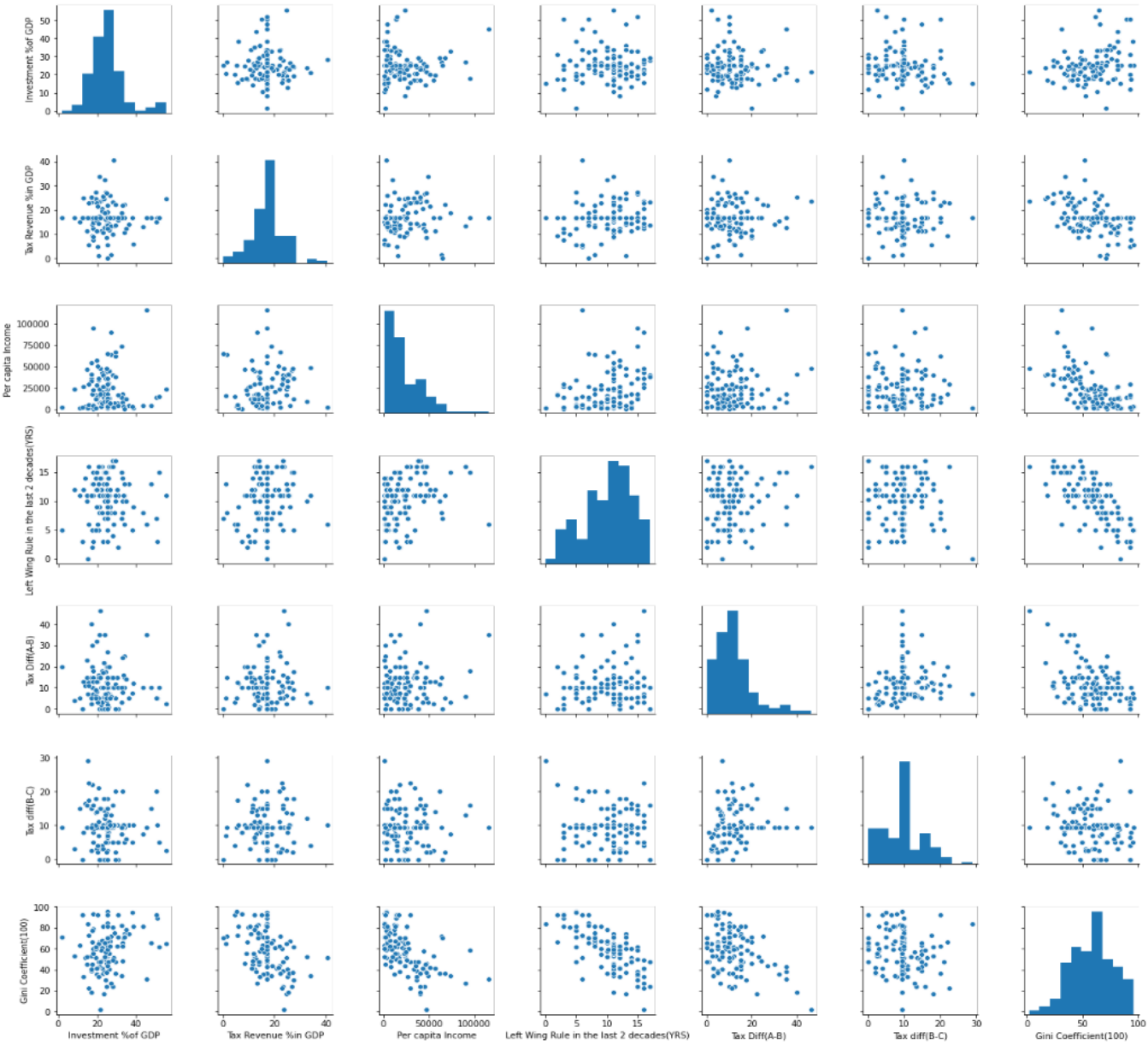


- **Difference between Taxes of class B(Middle) and Class C(Low)**
  - **Mean(μ) = 9.4**
  - **Standard Deviation(σ) = 5.6**



- **Scatter Plots of every possible Combination**

Studying Income Inequality

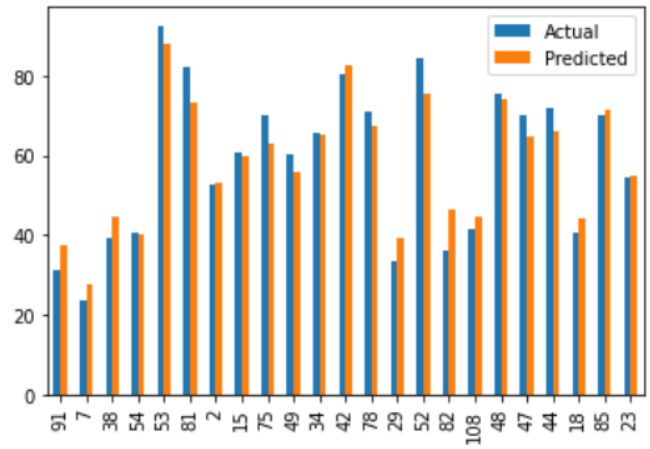


MODEL 1:

Multiple Linear Regression

target_variable =(coefficients * independent_varibales) + bias

- **For the first model every variable is taken into account**
- **Adjusted R2 = 0.862**
- **Mean of error = 1.23**



OLS Regression Results

Dep. Variable:	Gini Coefficient(100)	R-squared (uncentered):	0.869
Model:	OLS	Adj. R-squared (uncentered):	0.862
Method:	Least Squares	F-statistic:	119.9
Date:	Wed, 22 Apr 2020	Prob (F-statistic):	2.09e-45
Time:	05:26:46	Log-Likelihood:	-515.23
No. Observations:	114	AIC:	1042.
Df Residuals:	108	BIC:	1059.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Investment %of GDP	1.9266	0.195	9.866	0.000	1.540	2.314
Tax Revenue %in GDP	0.5029	0.309	1.626	0.107	-0.110	1.116
Per capita Income	-0.0003	0.000	-2.696	0.008	-0.000	-7.32e-05
Left Wing Rule in the last 2 decades(YRS)	-0.2563	0.539	-0.476	0.635	-1.324	0.812
Tax Diff(A-B)	-0.2326	0.257	-0.904	0.368	-0.743	0.278
Tax diff(B-C)	0.9545	0.375	2.547	0.012	0.212	1.697

Omnibus:	0.611	Durbin-Watson:	1.694
Prob(Omnibus):	0.737	Jarque-Bera (JB):	0.442
Skew:	0.152	Prob(JB):	0.802
Kurtosis:	3.020	Cond. No.	8.39e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.39e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Model 2:

- **Going through hierarchical building of multiple linear regression, we found this model to be effective**

- **giniIndex = b0 + b1*(investment) + b2*(Tax Revenue) + b3*(perCapita) +b4*(left_wing_rule) + b5* (tax_diffA_B * tax_diffB_C)**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Gini Coefficient(100)   R-squared (uncentered):                   0.863
Model:                               OLS   Adj. R-squared (uncentered):              0.856
Method:                    Least Squares   F-statistic:                              136.8
Date:                   Wed, 22 Apr 2020   Prob (F-statistic):                    2.68e-45
Time:                           06:19:46   Log-Likelihood:                         -518.16
No. Observations:                    114   AIC:                                      1046.
Df Residuals:                        109   BIC:                                      1060.
Df Model:                              5
Covariance Type:               nonrobust
==============================================================================================
                                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Investment %of GDP                           2.0090      0.196     10.268      0.000       1.621       2.397
Tax Revenue %in GDP                          0.6062      0.309      1.962      0.052      -0.006       1.219
Per capita Income                           -0.0003      0.000     -2.702      0.008      -0.000   -7.51e-05
Left Wing Rule in the last 2 decades(YRS)   -0.1721      0.538     -0.320      0.750      -1.238       0.894
TaxA*TaxB                                    0.0186      0.021      0.872      0.385      -0.024       0.061
==============================================================================
Omnibus:                        0.403   Durbin-Watson:                   1.712
Prob(Omnibus):                  0.817   Jarque-Bera (JB):                0.206
Skew:                           0.100   Prob(JB):                        0.902
Kurtosis:                       3.059   Cond. No.                     8.25e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

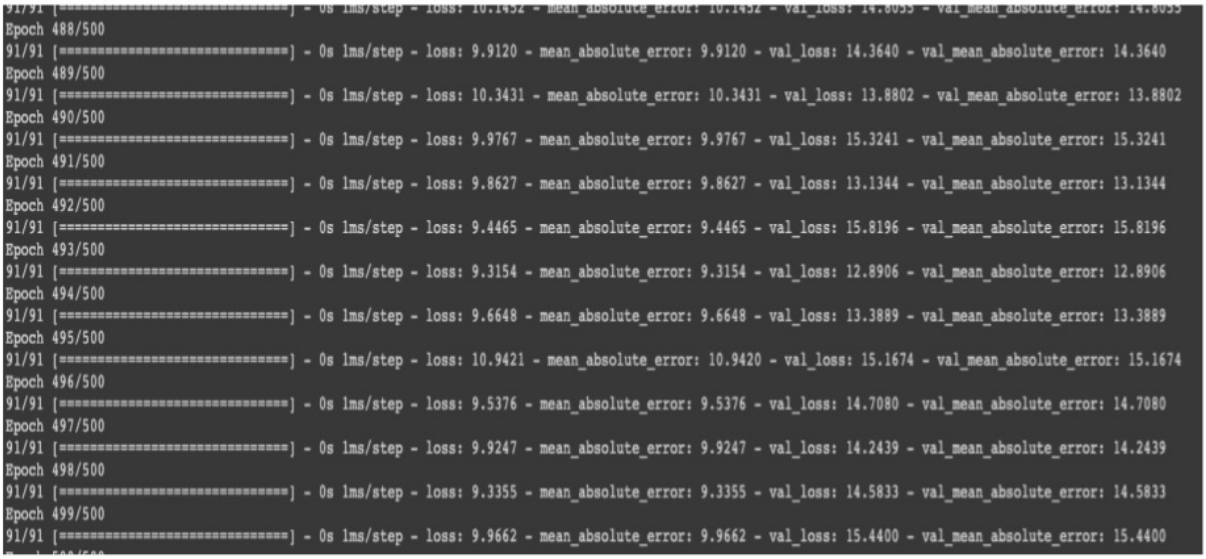
Model3:

Artificial Neural Networks Using keras

- **Building a neural network**

```
#Building Neural Network Using Keeras for regression
from keras.models import Sequential
from keras.layers import Dense, Activation, Flatten

NN_model = Sequential()

# The Input Layer :
NN_model.add(Dense(64, kernel_initializer='normal',input_dim = train_X.shape[1], activation='relu'))

# The Hidden Layers :
NN_model.add(Dense(128, kernel_initializer='normal',activation='relu'))
NN_model.add(Dense(128, kernel_initializer='normal',activation='relu'))
# NN_model.add(Dense(128, kernel_initializer='normal',activation='relu'))

# The Output Layer :
NN_model.add(Dense(1, kernel_initializer='normal',activation='linear'))

# Compile the network :
NN_model.compile(loss='mean_absolute_error', optimizer='adam', metrics=['mean_absolute_error'])
NN_model.summary()
```

- **Training the neural network**



IV. PERFORMANCE

Model-1:

R-Square = 0.869
Adjusted R-square of the model = 0.862
MSE = 5.33

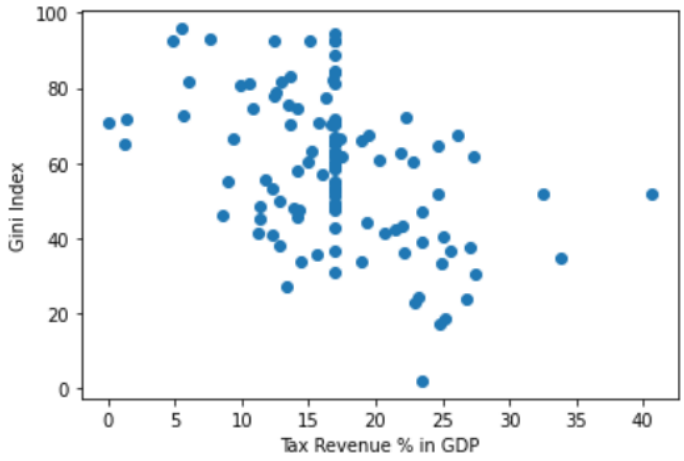
Model-2:
R-Square = 0.863
Adjusted R-square of the model = 0.856
MSE = 4.84

VIII. RESULT

What are the factors that affect the gini index are calculated and relation between the normal curve and lorenzcurve

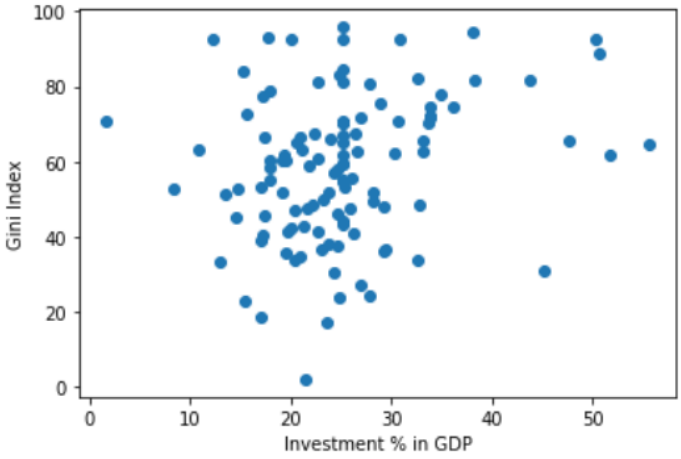
- There is a correlation of **-0.45** between Gini Index and tax Revenue percentage

Studying Income Inequality

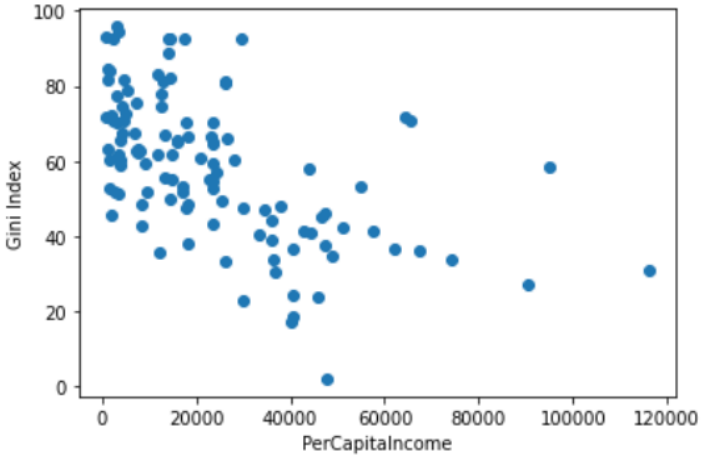


* It Indicates that the Gini Index decreases as tax revenue increases. Decrease in Gini Index indicates decrease in inequality. So, if the governments try to increase the tax revenue percentage in GDP, inequality may decrease.

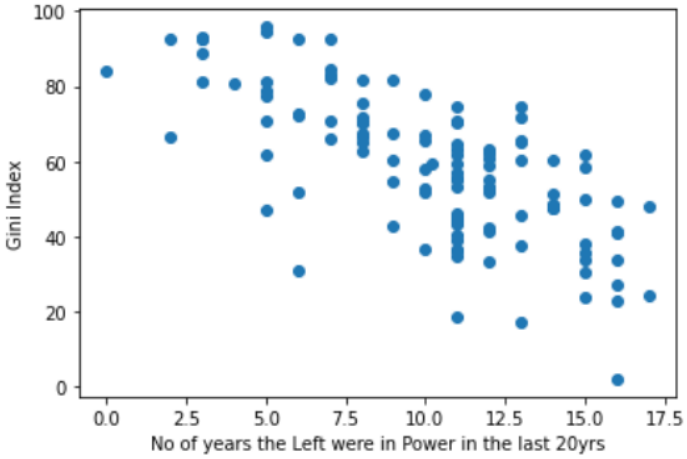
- There is a correlation of **0.22** between Gini Index and Investments % in GDP.
    - It indicates that if investments increase, inequality also increases. It is understandable because the investments benefit the rich in the short term.



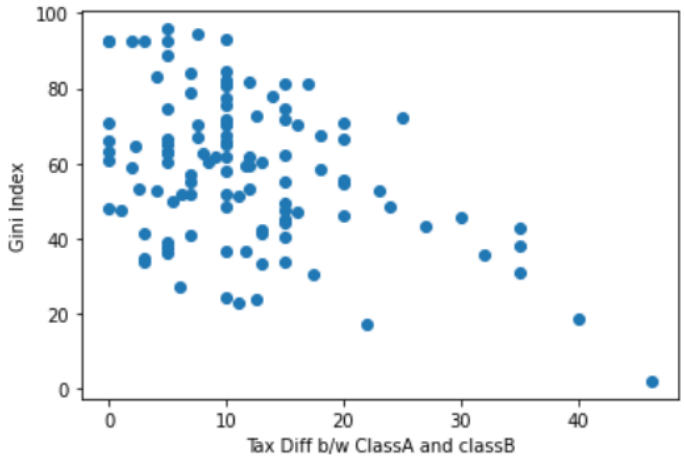
- There is a correlation of**. -0.54** between Gini Index and PerCapita income.
    - This Indicates that as the countries grow richer, the inequality will also vanish if other factors are held constant.



- There is a correlation of **.-0.70** between Gini Index and Left Wing rule.
    - So, this alone should prove that the policies of left wing parties are more effective in reducing economic inequality in the society



- Finally, the taxes part. There is only a little relationship(positive) between taxes paid by the upper class and economic inequality. So, increasing heavy taxes on the rich is not a good idea as it hits the corporate morale hard.



- Model 1 is the most accurate model we designed. The model parameters are :

Investment %of GDP	1.9266
Tax Revenue %in GDP	0.5029
Per capita Income	-0.0003
Left Wing Rule in the last 2 decades(YRS)	-0.2563
Tax Diff(A-B)	-0.2326
Tax diff(B-C)	0.9545

IX. CONCLUSION AND FUTURE SCOPE

No matter what way you look at it, solving an economic problem is always going to be a complex task.The only way to solve this complex task is to make it as simple as possible. That's what we did in this project. Everyone thinks that **raising taxes on the rich drastically will eradicate** economic inequality **but the data doesn't say so**. Raising tax revenue will decrease economic inequality. **More investments by the government will increase inequality** as most of the benefits go to the rich. But, it's not that significant. As the **countries get richer i.e per capita income increases, the economic inequality fades away** sooner or later. And the most important conclusion of this project is that the **left wing party policies were the most successful ones**. The data suggests that they suit the present better.

Coming to the Machine Learning part, for regression, the ordinary least squares method is far better than training the neural networks. The OLS method completes in no time and training a neural network takes forever. The final error obtained by OLS method is far less than the neural networks.

The future scope of this project is to add in more factors and bring it closer and closer to real life application. We would like to develop ANN and make it better than the OLS method. The present model has hetero-schedacity at untolerable levels. we would like to decrease it. The p-values for some factors are far more than they should be. In the future, we will minimize them.

REFERENCES:

1 http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8904917&fileOId=8906382
Adrian Mehic

2 http://arc.hhs.se/download.aspx?MediumId=680

3https://towardsdatascience.com/gini-coefficient-and-lorenz-curve-f19bb8f46d66

4 Farris, Frank. (2010). The Gini Index and Measures of Inequality. American Mathematical Monthly. 117. 851-864. 10.4169/000298910X523344.

Studying Income Inequality

ORIGINALITY REPORT

3%	2%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sunway College Student Paper	1%
2	etd.ohiolink.edu Internet Source	1%
3	etd.lib.nsysu.edu.tw Internet Source	1%
4	writemyclassessay.com Internet Source	1%

Exclude quotes	On	Exclude matches	< 11 words
Exclude bibliography	On		