Econometrics Data Assignment 1
Report

Hardik Singh-2021390 *Bhawna Kapoor-2019356* *Nikhil Kumar-2021174*
Shlok Vinodkumar Mehroliya -2021421

In this research project, we aim to investigate the correlation between economic growth, income inequality, and environmental quality in India, particularly regarding groundwater level. Our analysis will explore the Environmental Kuznets Curve theory, which proposes that as per capita income increases, environmental quality declines before eventually improving.
To accomplish this objective, we will utilize district-level data on economic output, income inequality, and groundwater levels to construct regression models and examine the relationship between these variables. The ultimate goal of our study is to provide valuable insights into how economic growth and income inequality impact the environment in India, with a specific focus on the groundwater level.
By conducting this research, we hope to contribute to a better understanding of the complex interplay between economic growth, income inequality, and environmental quality and to offer policymakers recommendations on promoting sustainable development in India.

Reference:

- Data source: Ministry of Jal Shakti
- Data Aggregation Platform: National Data Analytics Portal (NDAP)
- Net State Domestic Product (SDP) provided by the Reserve Bank of India(RBI), was accessed on Database for the Indian Economy (DBIE) portal
- District-Level Gini index from the paper “Estimates of Poverty and Inequality in the Districts of India,2011–2012” by Mohanty et al. (2016)

PART 1 + 2

In the first question of our project, we were required to select an environmental quality measure from a given list of options. We decided to work with Ground Water Level as our quality measure and, thus, loaded the corresponding NDAP data into a data frame as ‘dat.’
We then transformed the data into a district-year level dataset, including a unique “district-year ID” as “ID” for each row in the sample. Also, the data contained monthly fetched data for which we have taken the mean groundwater level of the months of a specific year.

PART 3

In order to analyze the relationship between economic output and groundwater level in India, we needed to merge district-level groundwater level data with state-year-wise economic output data. We obtained the net state domestic product (SDP) data at constant prices and loaded it into a data frame called 'df3'. Using the 'gather' function, we transformed the data frame into a Year-State-SDP format. Finally, we merged the 'df2' and 'df4' data frames based on the 'YEAR' and 'State' columns and created a new data frame called 'df5'. This new dataset allowed us to examine the relationship between economic output and groundwater level in India and can provide valuable insights for policy decisions promoting sustainable development.

PART 4

Now we needed to incorporate the District wise Gini index from Mohanty et al. (2016).We cleaned the table and merged the GINI index data by district with the already merged data by NDAP and SDP and stored in ‘final’, on which statistical work is done and the filtered data is stored in ‘specific’ for regression analysis.

PART 5

Summary Statistics

Descriptive statistics of variables

Statistic	N	Mean	St. Dev.	Min	Max
Ground.water.level	5,350	9.0	10.9	0.02	160.1
SDP	5,242	520,957.2	375,217.1	10,229	1,782,903
GI	5,262	0.3	0.1	0.2	0.5

The above table shows the descriptive statistics of three variables:

- **Ground.water.level**
- **GI**
- **SDP**

The "N" column indicates the number of observations for each variable. For Ground.water.level, there are 5,350 observations; for SDP, there are 5,242 observations, and for GI, there are 5,262 observations. The different number of observations is caused due to NA values.

The "Mean" column shows the average value of each variable across all observations. For example, the mean value of Ground.water.level is 9.0.

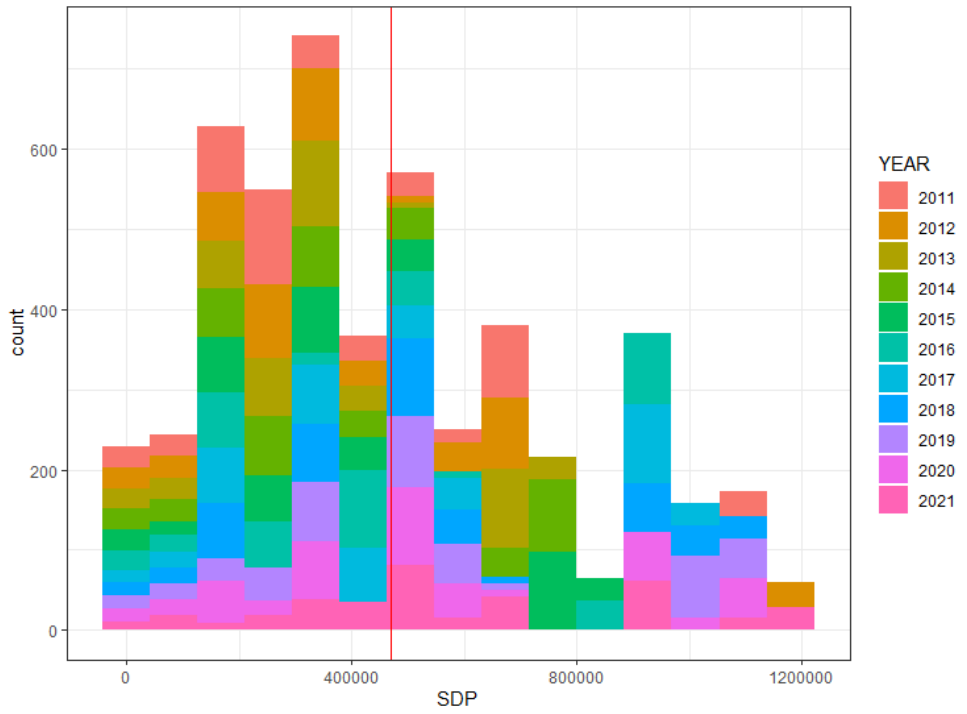
The "St. Dev." column shows the standard deviation of each variable, which is a measure of how much the values of the variable vary from the mean.

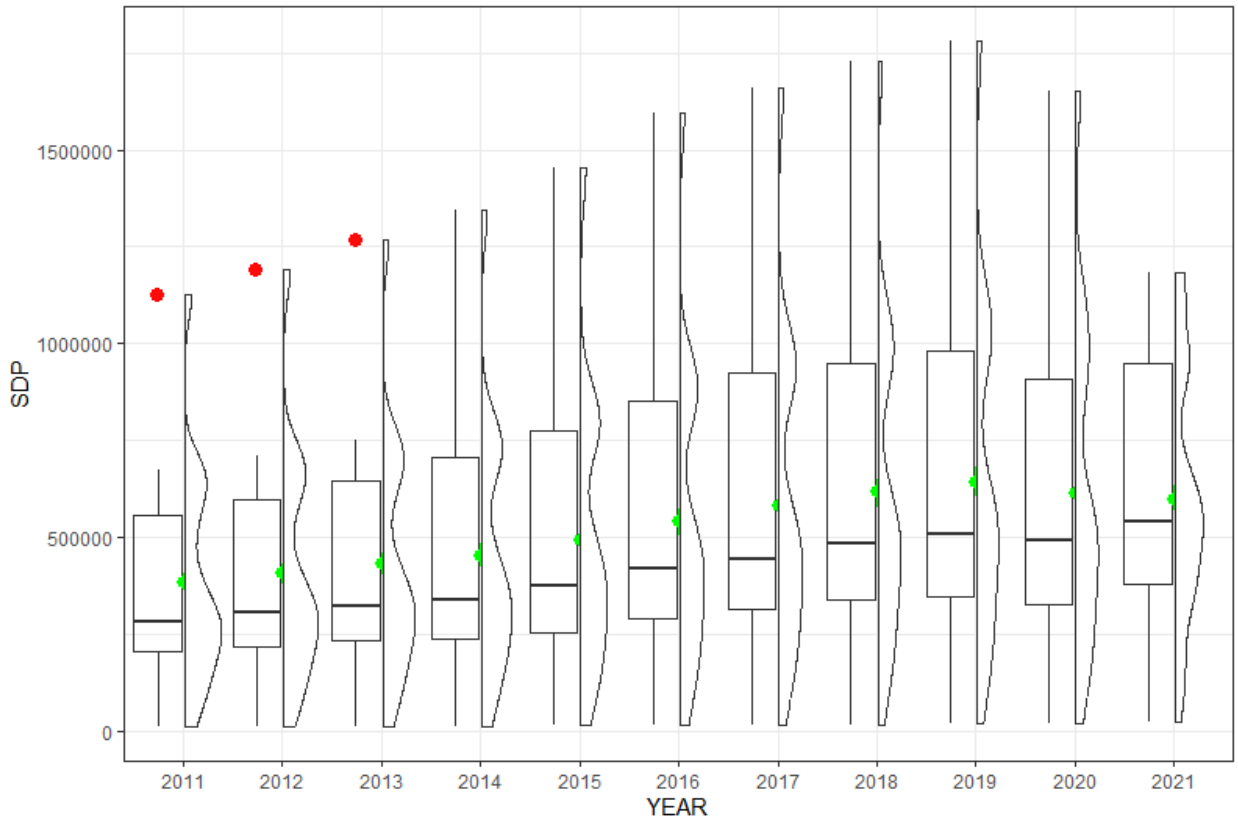
The "Min" and "Max" columns show the minimum and maximum values of each variable, respectively. For example, the minimum value of Ground.water.level is 0.02, and the maximum value is 160.1.

Detailed Statistics

Variables

1.**SDP**





Skewness: 1.119683

A skewness value of 1.119683 indicates that the distribution is highly skewed to the right and has a high degree of positive skewness.

More specifically, the skewness value of 1.119683 indicates that the majority of the data points are clustered on the left side of the distribution and that the tail of the distribution is significantly longer on the right side than it is on the left.

Interpretation:

Histogram

The histogram shows that the distribution is not normally distributed, which can be verified by the skewness. The red line represents the mean of the SDP, which also does not intersect with the mode. Also, the graph is right skewed as the distribution's right tail is longer.

The histogram has also been color coded for the different years, so the distribution over the years can also be seen.

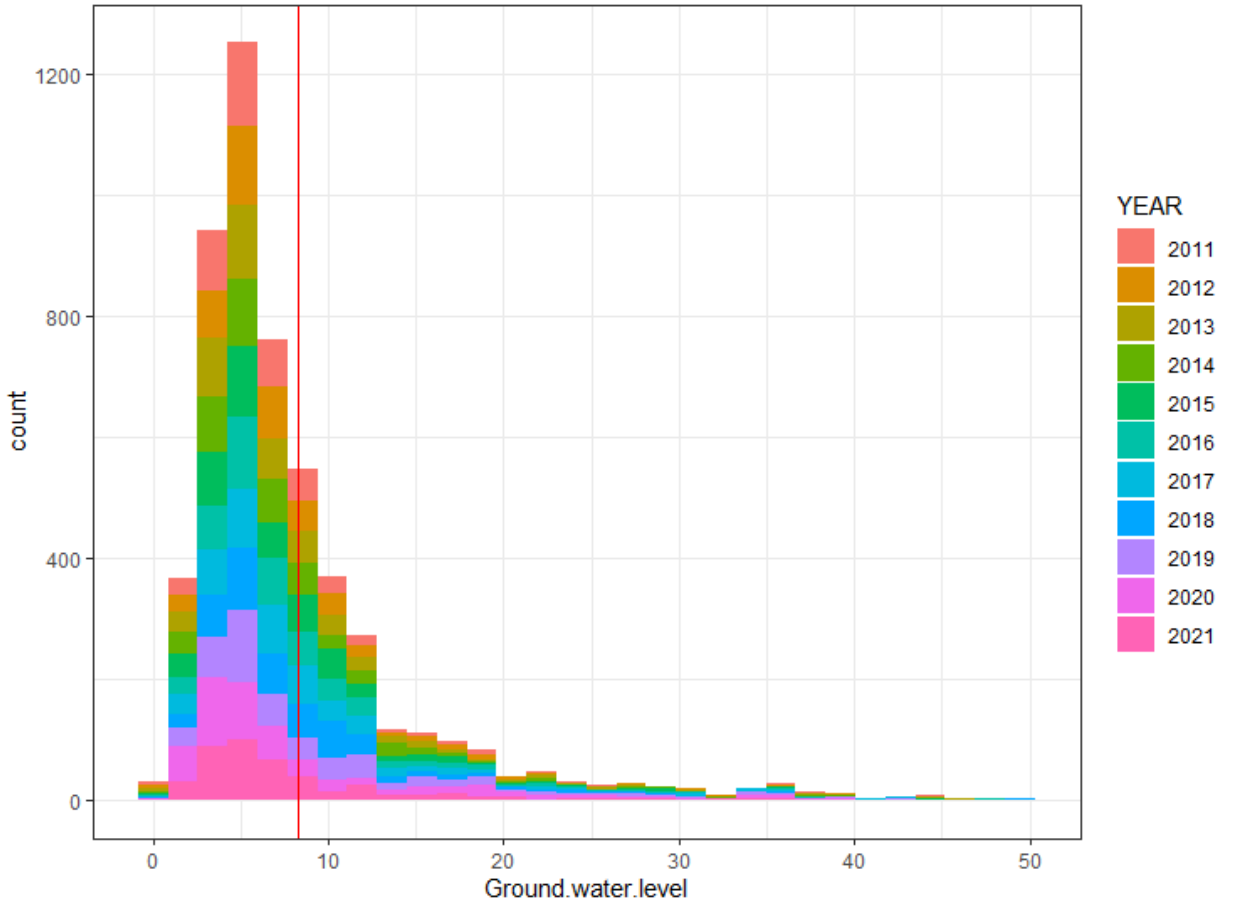
BoxPlot

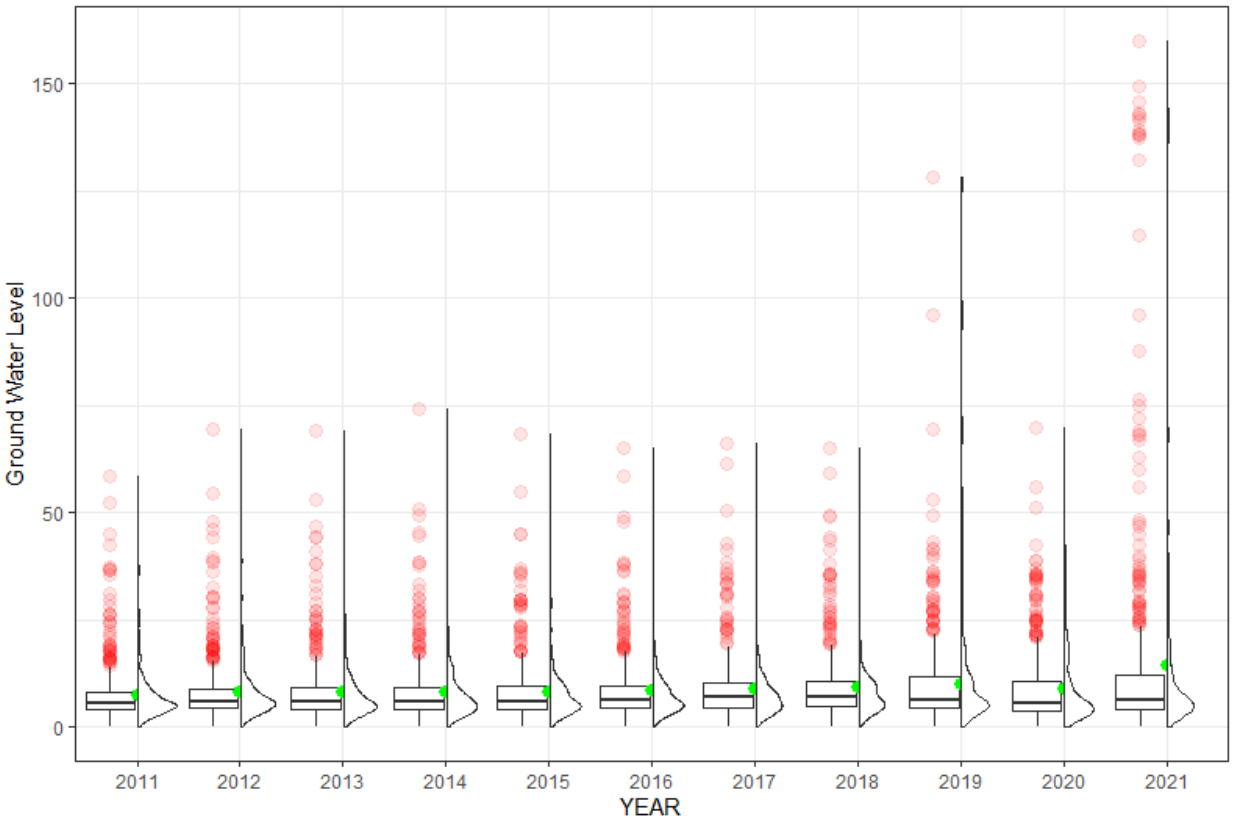
In the boxplot, the mean is represented with a green dot. The outliers have been marked in red. The boxplot also verifies the above claim of skewed distribution as none

of the medians lies on the mean. As all the medians are lower than the mean, the graph is right-skewed, which verifies the above claim from the histogram.
The distribution is plotted using the curve on the right of the boxplot.
We can also generalize that the SDP has increased positively over the years(particularly till 2019).

Summary:
We can conclude that the distribution is right-skewed with a high skewness value. There are outliers in the distribution as well.

2. Ground Water Level





Skewness: 6.548

A skewness value of 6.548 indicates that the distribution is significantly skewed to the right and has a high degree of positive skewness.

More specifically, the skewness value of 6.548 indicates that the majority of the data points are clustered on the left side of the distribution and that the tail of the distribution is significantly longer on the right side than it is on the left.

Interpretation:

Histogram

The histogram shows that the distribution is not normally distributed, which can be verified by the skewness. The red line represents the mean of the SDP, which also does not intersect with the mode. Also, the graph is right skewed as the distribution's right tail is longer.

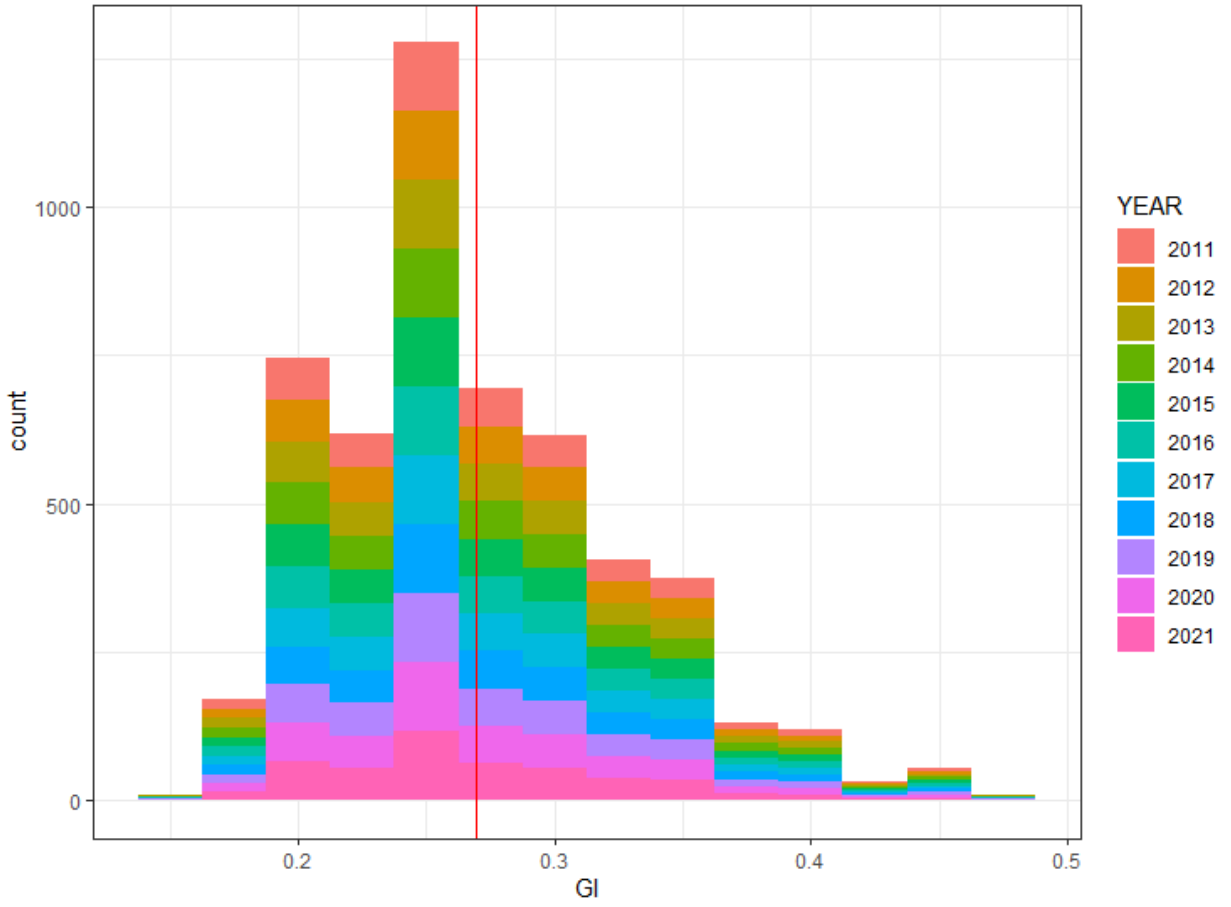
The histogram has also been color coded for the different years, so the distribution over the years can also be seen.

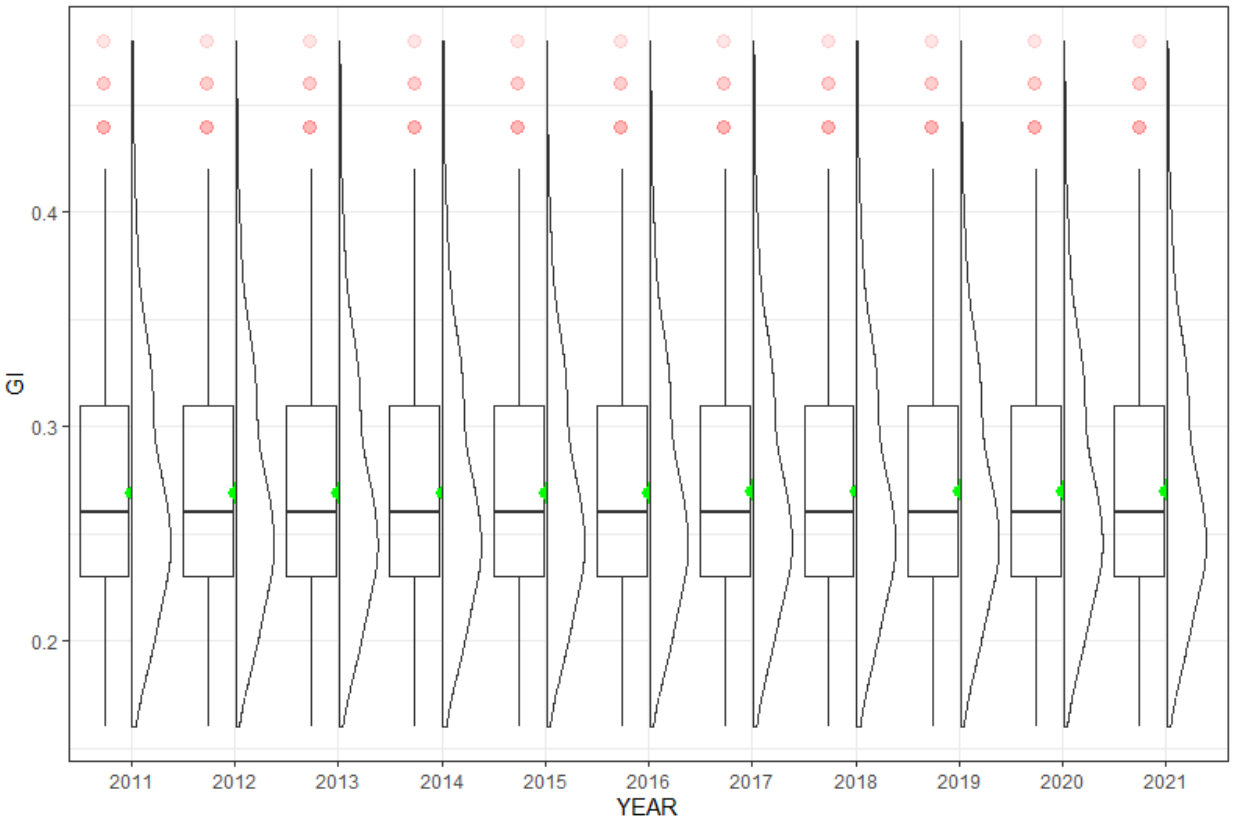
BoxPlot

In the boxplot, the mean is represented with a green dot. The outliers have been marked in red. The boxplot also verifies the above claim of skewed distribution as none

of the medians lies on the mean. As all the medians are lower than the mean, we can say that the graph is right-skewed, which verifies the above claim from the histogram. The distribution is plotted using the curve on the right of the boxplot.
We can also see a slight upward trend in the values of Ground Water Level over the years.
Summary:
We can conclude that the distribution is right-skewed with a high skewness value. There are outliers in the distribution as well.

3. Gini Index(GI)





Skewness: 0.7327

A skewness value of 0.7327 indicates that the distribution is moderately skewed to the right.

Interpretation:

Histogram

The histogram shows that the distribution is not normally distributed, which can be verified by the skewness. However, the skewness is small as the mean is very close to the mode, and there are only a few outliers.

The red line represents the mean of the SDP. Also, the graph is right skewed as the distribution's right tail is longer.

The histogram has also been color coded for the different years, so the distribution over the years can also be seen.

BoxPlot

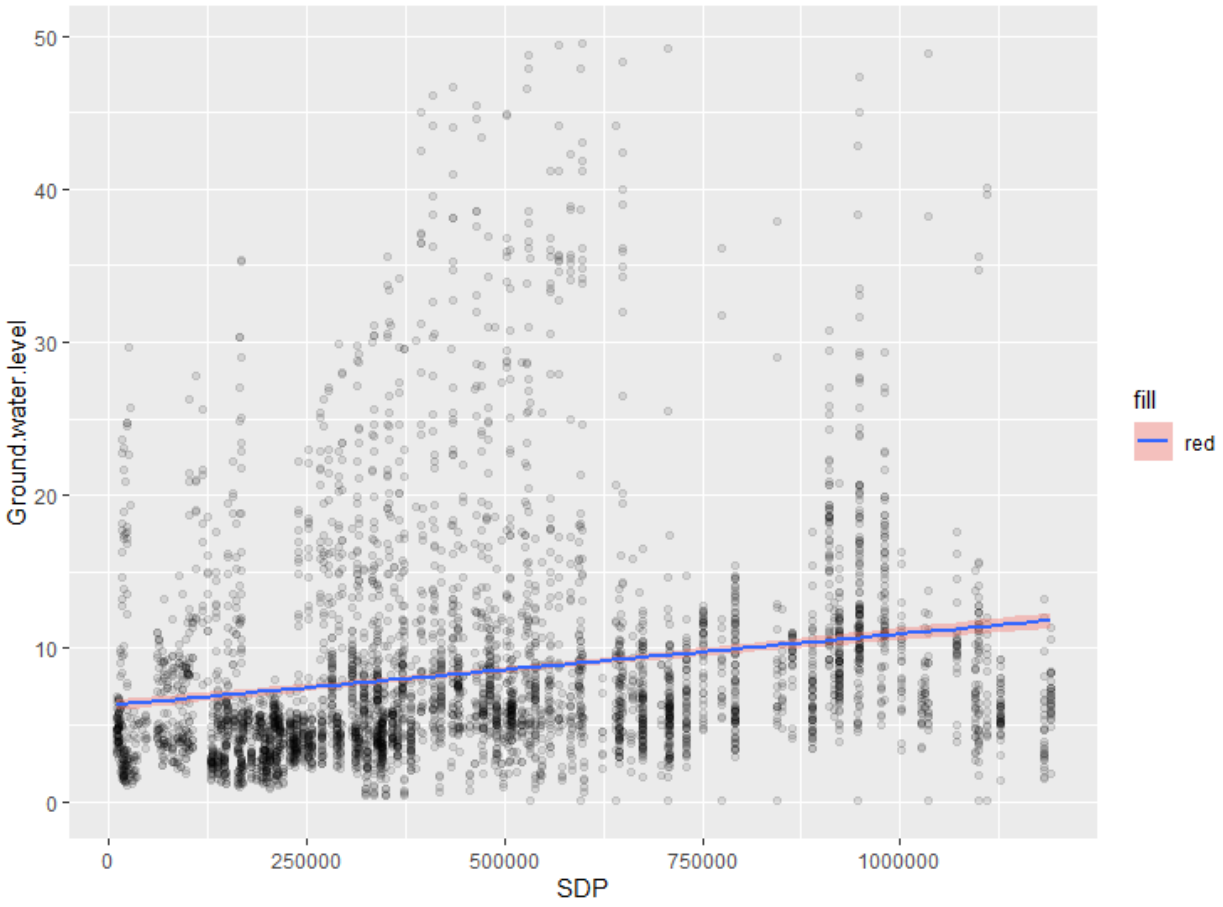
In the boxplot, the mean is represented with a green dot. The outliers have been marked in red. The boxplot also verifies the above claim of skewed distribution, as none of the medians lies on the mean. However, the difference between the mean and median is not very large and is uniform over all the years, which verifies the moderate skewness value.

The distribution is plotted using the curve on the right of the boxplot.
GI has also been pretty uniform over the years.

Summary:
We can conclude that the distribution is right-skewed with a moderate skewness value.
There are a few outliers in the distribution as well.

PART 6

	Dependent variable:
	Ground.water.level
SDP	0.00000***
	(0.00000)
Constant	6.255***
	(0.186)
Observations	4,867
R2	0.038
Adjusted R2	0.038
Residual Std. Error	7.026 (df = 4865)
F Statistic	192.755*** (df = 1; 4865)
Note:	*p<0.1; **p<0.05; ***p<0.01



The table shows the results of the regression analysis with Ground.water.level as the dependent variable and one independent variable(SDP).

The coefficient for the independent variable (SDP) is 0.00000, which indicates that a one unit increase in the SDP of the independent variable is associated with no increase in the dependent variable (Ground.water.level). The notation "***" indicates that this coefficient is statistically significant at the 0.01 level, which means that there is a less than 1% chance that this observed relationship is due to chance.

However the graph shows us a visible positive relation. The red highlight provides us with the standard error.

This discrepancy is due to the fact that the range in which the value of SDP lies is very wide so the increase in dependent variable (Ground.water.level) on unit increase of SDP is negligible, however when we increase SDP by a significant amount(in range of 10000) we see a visible increase in Ground.water.level.

The constant term in the regression is 6.255, which is the expected value of the dependent variable when the independent variable is zero. The notation "***" indicates that this constant term is statistically significant at the 0.01 level.

Observations

This is the number of observations (occurrences) in our dataset. In our case, the total observations are 4,867.

R2 (r^2) is also known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the predictor variable/independent variable.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A variable of 1 indicates that the response variable can be perfectly explained without error by the predictor variable/ independent variable.

In the above case, the R2 is 0.038, which indicates that 3.8% of the variance in the dependent variable (ground water level) can be explained by the independent variable (GI).

Adjusted R2

This is a modified version of R2 that has been adjusted for the number of predictors in the model. It is always lower than the R2. The adjusted R2 can be useful for comparing the fit of different regression models to one another.

In the above case, the Adjusted-R2 is 0.038, which is the same as the R2. This means that no adjustments were made to the model to account for the number of independent variables.

Residual Std. error

The residual standard error of 7.026 indicates the average amount of error or variability in the model that is not explained by the independent variable.

F-statistic

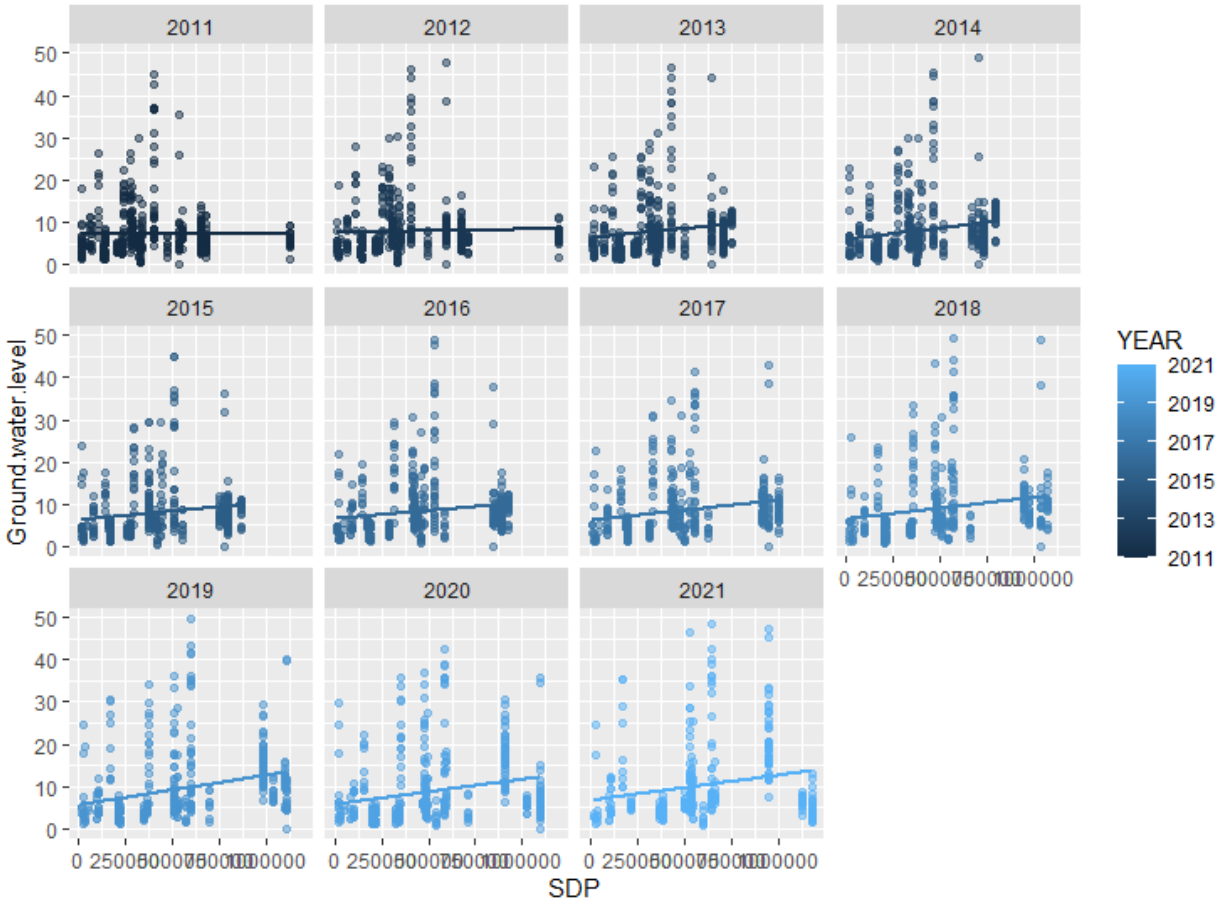
This statistic indicates whether the regression model provides a better fit to the data than a model that contains no independent variables.

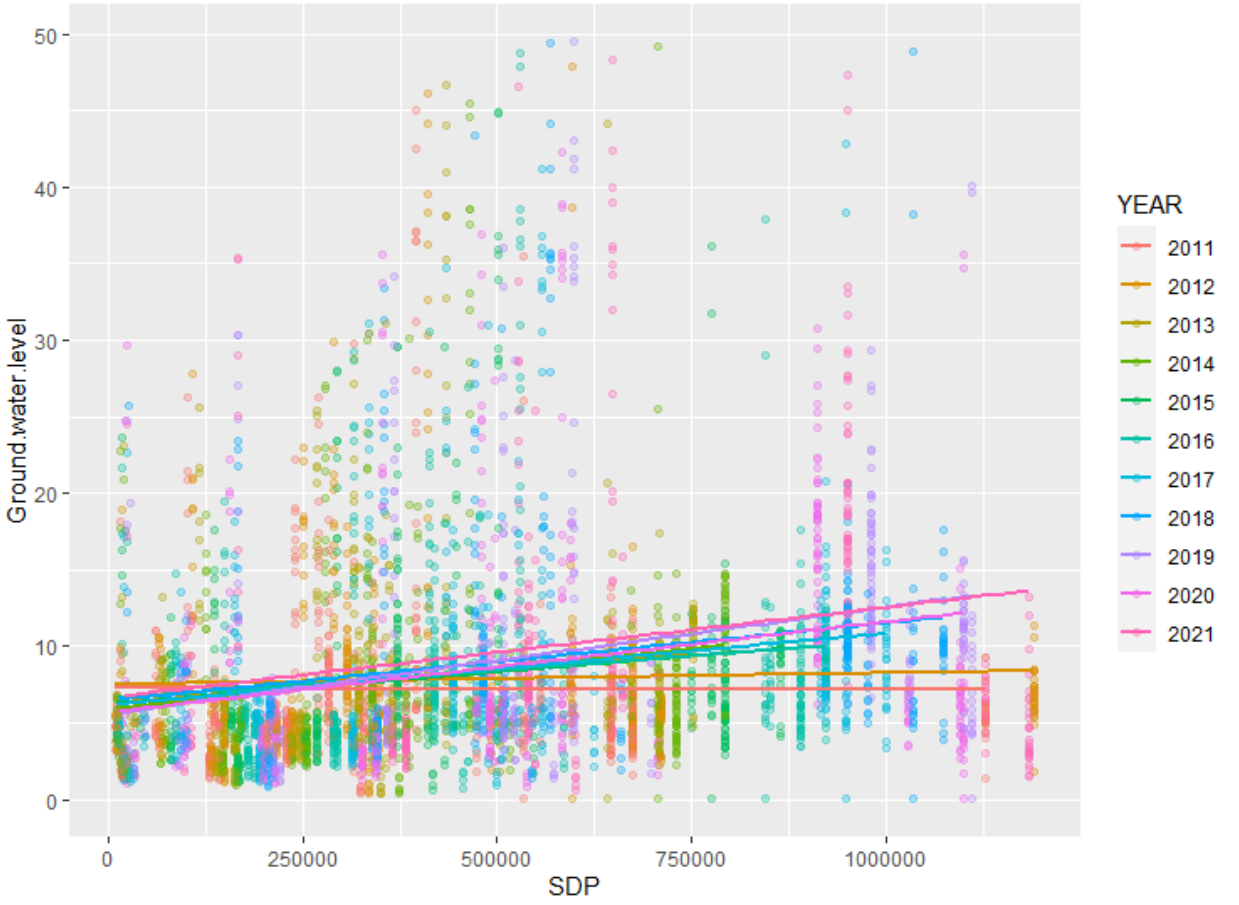
Basically, it measures whether there is a significant linear relationship between the independent variables and the dependent variable.

In our case, the F-statistic is 192.755.

The notation "*p<0.01" indicates the level of statistical significance associated with the F-statistic. In this case, the* "" indicates that the F-statistic is statistically significant at the 0.01 level, which means that there is a less than 1% chance that the observed relationship between the variables is due to chance.

Overall, if this F-statistic is greater than the critical value then the independent variable(SDP) is a significant determinant of the dependent variable(Ground.water.level).





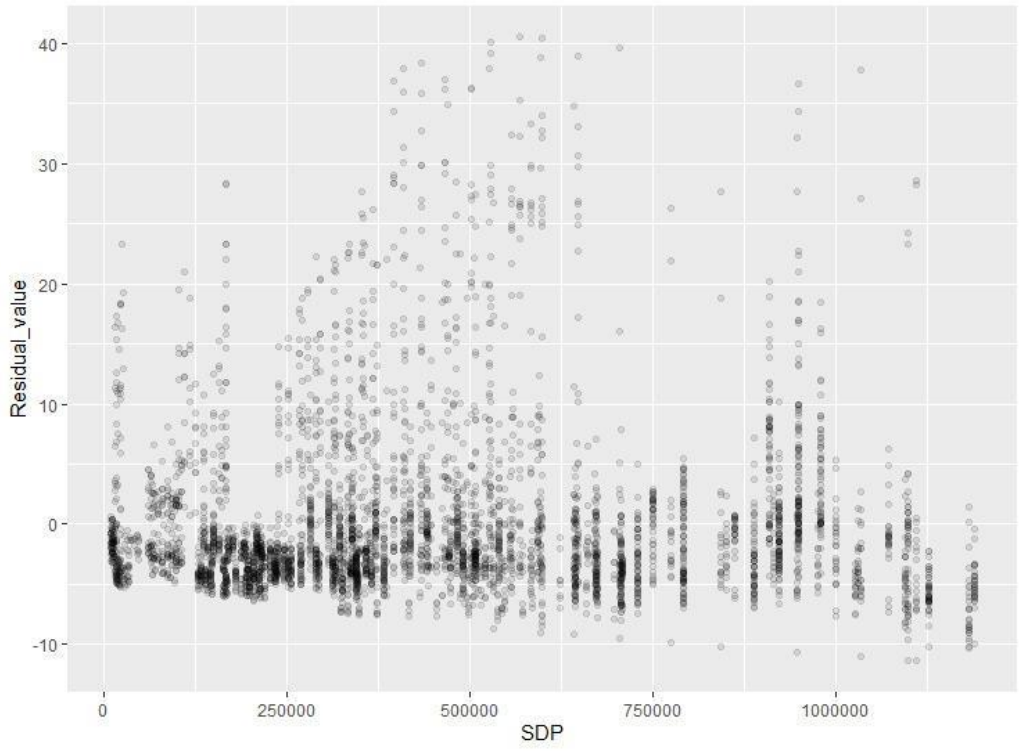
Interpretation:

The above two graphs provide a year-wise dependence of Ground.water.level on SDP. Upon inspection, we see that generally, as the year increases, the dependency of Ground.water.level on SDP increases. The graph shows that the initial relationship was horizontal and became more and more positive as the year passed.

Summary:

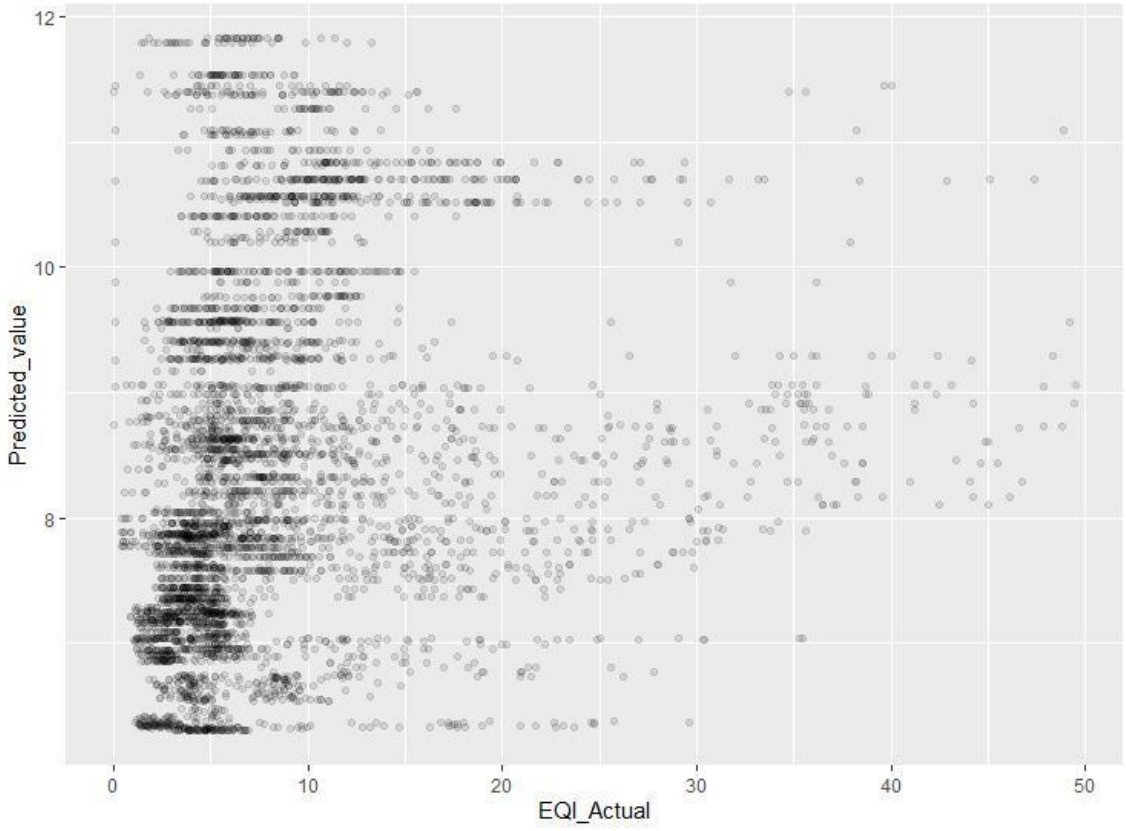
This result aligns with the expectation from Environmental Kuznets Curve theory which expects the relationship between environmental quality and the economy to improve positively as time passes. This has been shown by the relationship between EQI(Environment Indicator, i.e., Ground Water Level) and Economy(State Domestic Product).

PART 7



Interpretation:

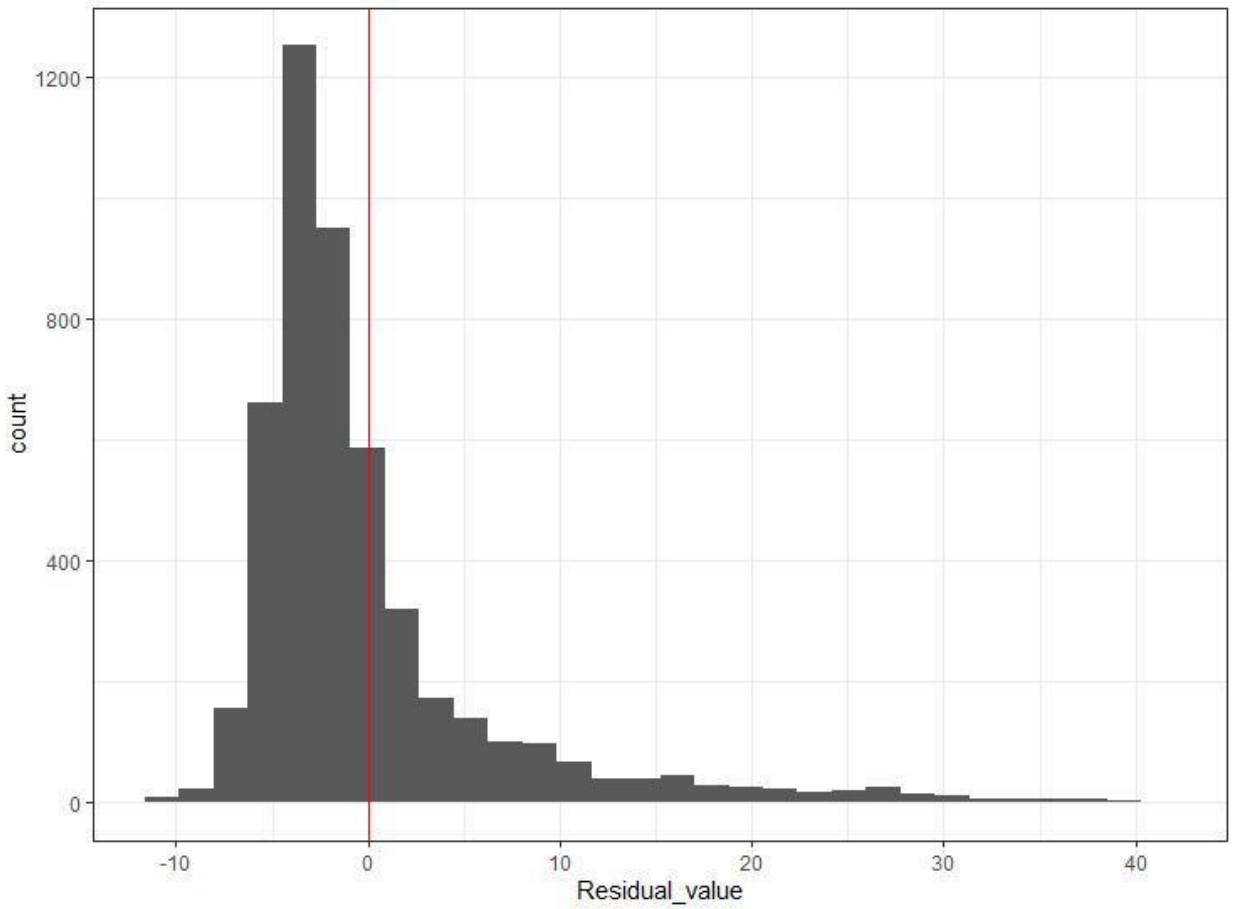
The above graph represents the difference between the observed value and the predicted value(i.e., the residuals) of the dependent variable being plotted against the values of SDP. Upon inspection, it can show the validity of the assumption of a linear regression model and identify a potential problem in our model(if any).



Interpretation:

The above plot is a graphical representation of how well the linear regression model fits the data. We have plotted the predicted values(from our regression model) against our actual values(of Ground Water Level).

PART 8



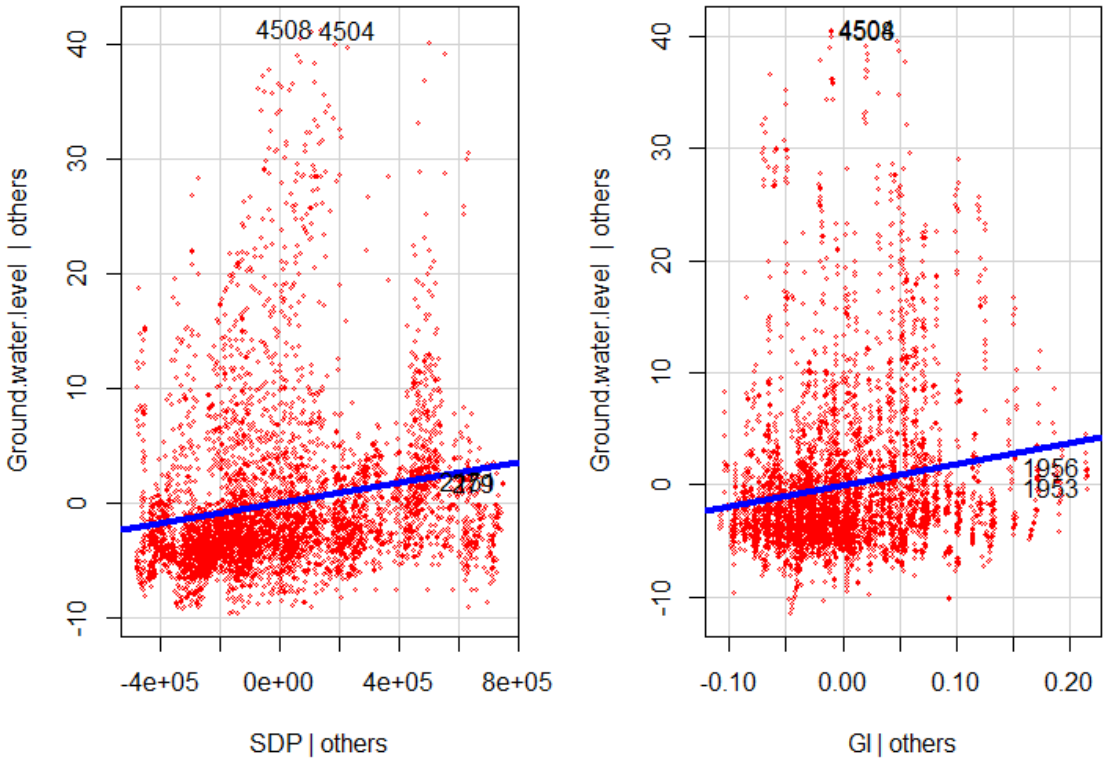
Interpretation:

The above graph represents the distribution of residuals in a linear regression model for the given set of observations. The above histogram shows the frequency distribution of the residuals, with the mean represented by the red line. The mean(red line) also verifies the assumption that the residuals' mean is zero.

PART 9

	Dependent variable:
	Ground.water.level
SDP	0.00000***
	(0.00000)
GI	18.720***
	(1.729)
Constant	1.308***
	(0.493)
Observations	4,867
R2	0.061
Adjusted R2	0.060
Residual Std. Error	6.943 (df = 4864)
F Statistic	157.307*** (df = 2; 4864)
Note:	*p<0.1; **p<0.05; ***p<0.01

Dependence on each of the Independent Variables while holding others fixed



The table shows the results of a linear regression model with Ground.water.level as the dependent variable and SDPand GI as the independent variables.

The coefficients for each independent variable are shown in the table.

The coefficient for SDP is 0.00000, which indicates that a one unit increase in the SDP of the independent variable is associated with no increase in the dependent variable (Ground.water.level). The notation "***" indicates that this coefficient is statistically significant at the 0.01 level, which means that there is a less than 1% chance that this observed relationship is due to chance.

Similarly, the coefficient for GI is 18.720, which means that for every unit increase in GI, Ground.water.level is expected to increase by 18.720 units.The notation "***" indicates that the GI term is statistically significant at the 0.01 level.

The constant term in the regression is 1.308, which is the expected value of the dependent variable when the independent variable is zero.

The table also shows the standard error for each coefficient, which is a measure of the uncertainty in the estimate of the coefficient. The notation "***" indicates that this constant term is statistically significant at the 0.01 level.

- The standard error for the coefficient of SDP is 0.00000.
- The standard error for the coefficient of GI is 1.729 .
- The constant term in the regression model is shown in the table with a coefficient of 0.493.

Observations

This is the number of observations (occurrences) in our dataset. In our case, the total observations are 4,867.

R2 (r^2) is also known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the predictor variable/independent variable.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A variable of 1 indicates that the response variable can be perfectly explained without error by the predictor variable/ independent variable.

In the above case, the R2 is 0.061, which indicates that 6.1% of the variance in the dependent variable (Ground.water.level) can be explained by the independent variables (SDP , GI) in the regression model.

Adjusted R2

This is a modified version of R2 that has been adjusted for the number of predictors in the model. It is always lower than the R2. The adjusted R2 can be useful for comparing the fit of different regression models to one another.

In the above case, the Adjusted-R2 is 0.060, which is the same as the R2. This means that no adjustments were made to the model to account for the number of independent variables.

Residual Std. error

The residual standard error of 6.943 indicates the average amount of error or variability in the model that is not explained by the independent variables.

F-statistic

This statistic indicates whether the regression model provides a better fit to the data than a model that contains no independent variables.

Basically, it measures whether there is a significant linear relationship between the independent variables and the dependent variable.

A larger F statistic indicates a better fit of the model to the data.

The "df" values indicate the degrees of freedom for the model, which are used to calculate the p-value associated with the F statistic. In our case, the F statistic is 157.307 with 2 and 4864 degrees of freedom, and the p-value is less than 0.01, which indicates that the regression model is statistically significant.

Interpretation from graph:

The graph shows the effect of each independent variable(SDP and GI) on the dependent variable independently.
The graph shows that the Dependent variable(Ground Water Level) is positively related to both SDP and the Gini Index(GI); this contradicts the above statement that SDP is unrelated to Ground Water Level. This discrepancy is due to the fact that the range of values for SDP is vast, so the change in the Dependent variable

for a unit change is negligible. However, when we increase SDP by a significant amount(in the range of 10000), we see a visible increase in Groundwater level which is evident from the graph.