Data Assignment 1
ECO221 (Winter 2023)

Introduction

This is the report of Data Assignment1 by Group14. We are here trying to reassess the environmental Kuznets Curve which is a hypothesized relationship between environmental degradation and per capita income. Naturally, for any demand to materialize into policy action - requires political intervention which is often a function of who is demanding what outcomes. Hence, the power and/or income distribution, i.e., inequality also likely contributes to the relationship between income and environmental quality (shown by Torras, M. and Boyce, J. (1998))

Answers

Q1)

We have chosen Ground Water Level as Environmental Quality measure.
In this section, the required libraries are loaded, a CSV file named "NDAP REPORT 7063.csv" is read, and some columns and rows that contain missing values are removed. By joining the "District.lgd.code" and "Yearcode" columns, a new column called "district year id" is produced in the dataset.

```
1
2  rm(list=ls())
3  set.seed(50)
4  library(ggplot2)
5  library(rlang)
6  library(dplyr)
7  library(readxl)
8  library(stargazer)
9
10
11 #Part1
12
13 df <- read.csv("D:/IIITD/Econometrics1/DataAssignment1/Groundwater Level (1)/NDAP_REPORT_7063.csv")
14
15 df <- df[,-c(8,9,10)]
16 df <- df[!(df$Ground.water.level == ""), ]
17 df <- df[!is.na(df$Ground.water.level), ]
18
19 df$district_year_id <- do.call(paste, c(df[c("District.lgd.code", "Yearcode")], sep = "-"))
20
21
```

Q2)

```
#PART2

# Created a new column for district-year-id
s=0
i=0
v=0
df1=data.frame(matrix(ncol=0,nrow=0))
v=0
df1=data.frame(matrix(ncol=0,nrow=0))
#df1[1,1:9]=df[1,1:9]
earlierid=""
currentid=""
flag=1
for(n in 1:nrow(df)){
  currentid=df[n,9]
  if(earlierid==currentid){
    s=s+df[n,8]
    i=i+1
  }
  else{if(flag==0){
    df1[v,1:9]=df[n-1,1:9]
    df1[v,8]=s/i
  }
    else{
      flag=0
    }
    s=df[n,8]
    i=1
    v=v+1
  }
  earlierid=currentid
}
df1[v,1:9]=df[nrow(df),1:9]
df1[v,8]=s/i
df1 <- df1 %>% mutate(State = tolower(State))
df1 <- df1 %>% mutate(District = tolower(District))
```

For each district-year combination, a new data frame called "df1" is constructed to contain the average ground water level. The loop determines the average ground water level for each district-year combination by iterating over the rows of the original dataset. We are taking average of all months present in a year to calculate the Ground water level for that particular year. For uniformity, the resulting data frame is likewise changed to lowercase.

The data includes the ROWID, Country, State.lgd.code, State, District.lgd.code, District, Yearcode, Ground.water.level, and district_year_id columns.

- ROWID: A unique identifier for each row of data.
- Country: The name of the country, which is India.
- State.lgd.code: The code for the state.
- State: The name of the state.
- District.lgd.code: The code for the district.
- District: The name of the district.
- Yearcode: The year for which the groundwater level data is reported.
- Ground.water.level: The groundwater level for the given district and year.
- district_year_id: A combination of the district code and year, used as a unique identifier for each row.

Each row in the dataset refers to a particular district in a certain year, and it also includes the matching ground water level. Using a special code that combines the district code and the year, the district and year are recognised.

Q3)This section reads a CSV file "SDP.csv" into a data frame SDP_df. We then converted the State column to lowercase in SDP_df. Next, we merged df1 and SDP_df data frames by State column into a new data frame dff or we can say that merged the district-year level environmental quality data with the corresponding state-year wise economic output data, i.e., the net state domestic product (SDP) at constant prices.

```
#Part3

SDP_df <- read.csv("D:/IIITD/Econometrics1/DataAssignment1/SDP.csv")

SDP_df <- SDP_df %>% mutate(State = tolower(State))
dff <- merge(df1, SDP_df, by="State")
```

Q4)This component reads a CSV file called "GINI.csv" and renamed the second column to "District". Transformed the "District" column to lowercase and combined "dff" and "gini df" data frames by the "District" column, and assigned it to a new data frame called "merged df". We then displayed the column names of the "merged_df" data frame and removed unwanted columns by using the select function. Finally, renamed the sixth column of "final_df" to "Yearcode" and displayed it in a new window.

```
#Part4

gini_df <- read.csv("D:/IIITD/Econometrics1/DataAssignment1/GINI.csv")
colnames(gini_df)[2] <- "District"
gini_df <- gini_df %>% mutate(District = tolower(District))
merged_df <- merge(dff, gini_df, by="District")


final_df <- select(merged_df, -c(ROWID, Yearcode.y, Sr...No..,))
colnames(final_df)[6] <- "Yearcode"

View(final_df)
```

```
> print(colnames(merged_df))
 [1] "District"           "State"              "ROWID"              "Country"
 [5] "State.lgd.code"     "District.lgd.code"  "Yearcode.x"         "Ground.water.level"
 [9] "district_year_id"   "Yearcode.y"         "SDP"                "Sr...No.."
[13] "Gini"
```

Q5)

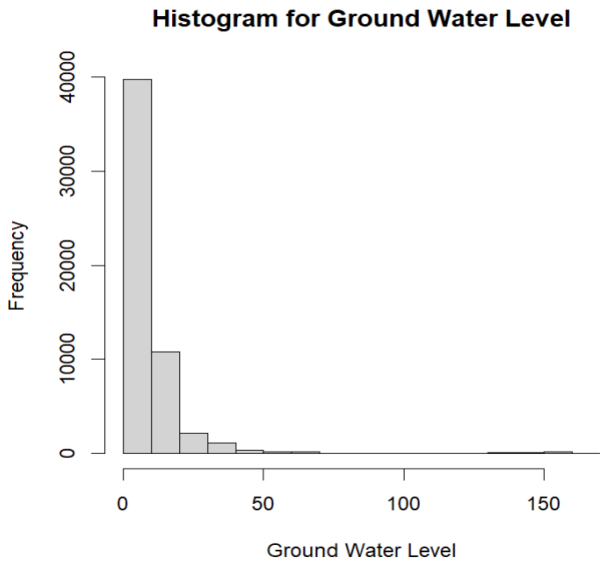
Summary Statistics

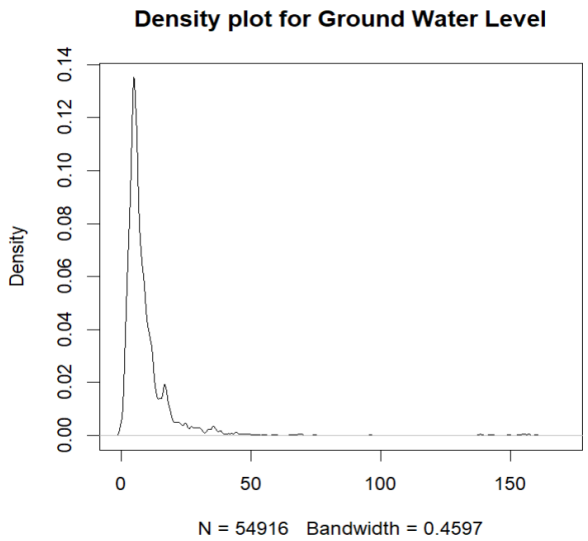
Statistic	N	Mean	St.Dev.	Min	Max	1Q	3Q
Ground.water.level	54,916	9.99	13.88	0.019	169.264	4.53	10.65
SDP	54,916	536,839.1	376,669.5	10,229	1,782,903	255,739	729,686
Gini	54,916	0.27	0.057	0.16	0.48	0.23	0.31

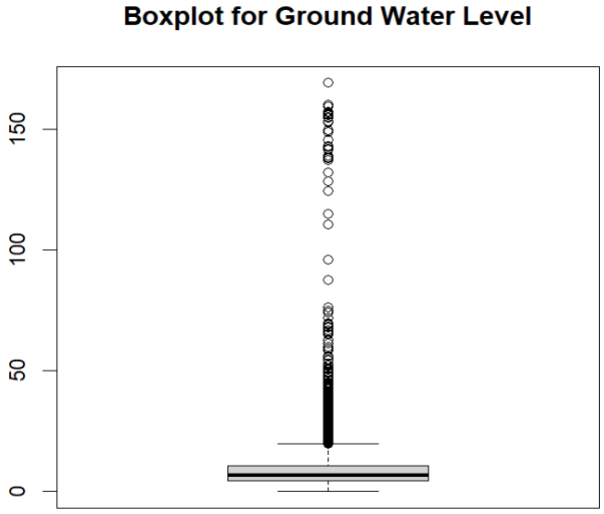
```
> skewness(final_df$Ground.water.level)
[1] 6.897057
> skewness(final_df$SDP)
[1] 1.035898
> skewness(final_df$Gini)
[1] 0.7558638
```

We have plots of all these variables are as as follows →

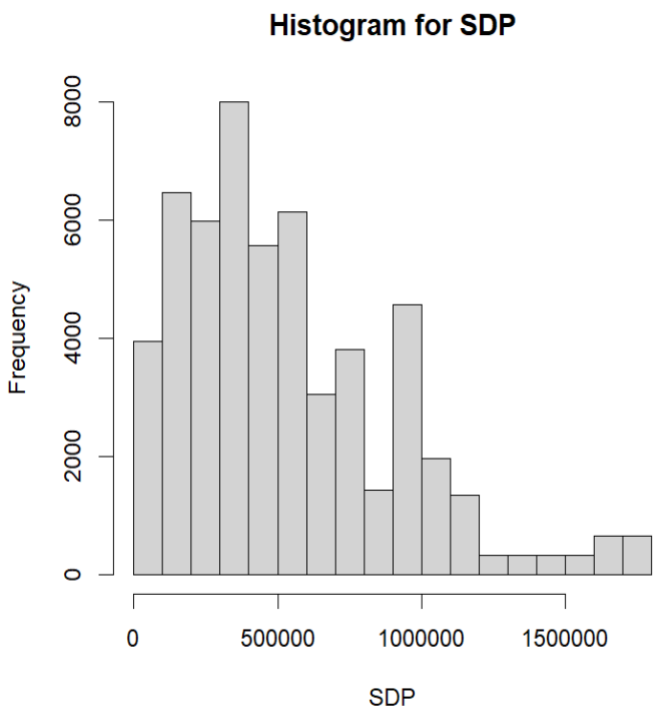
(a) Ground water level

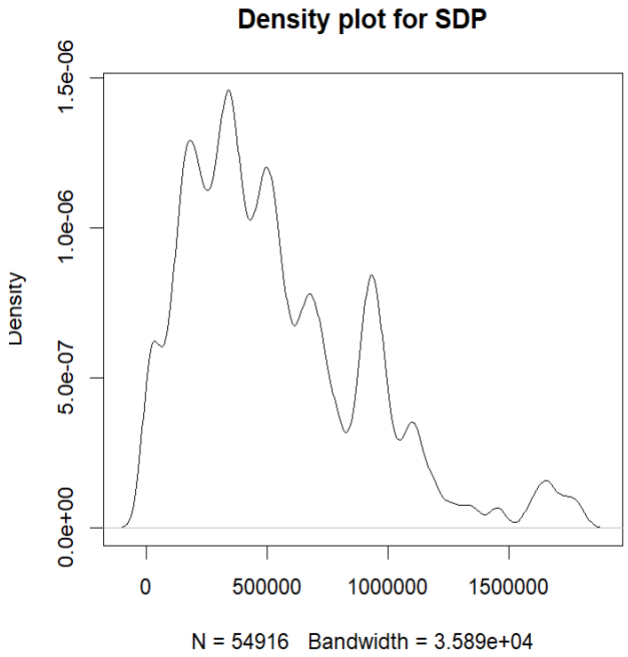


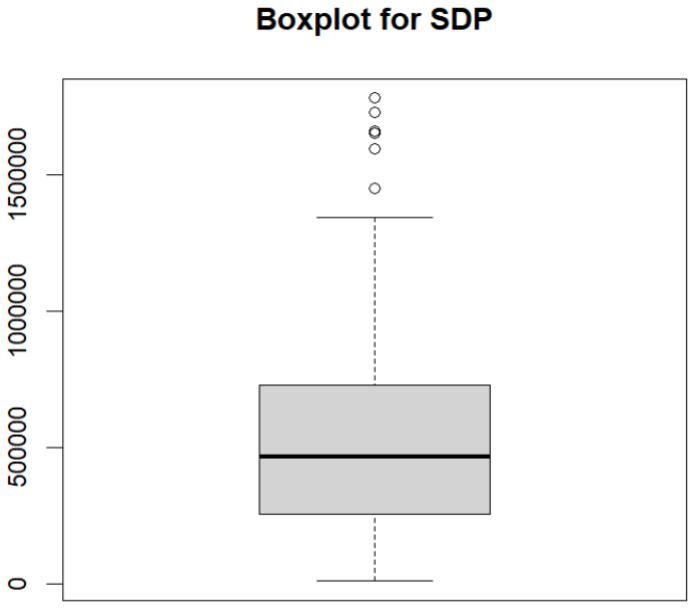




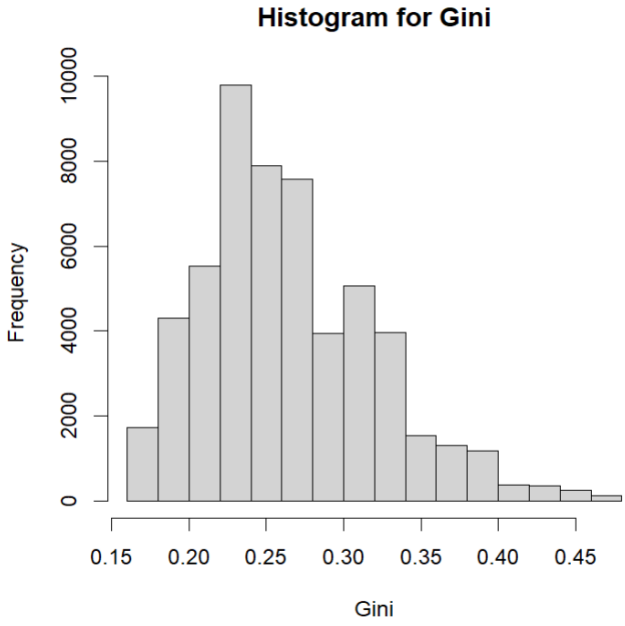
(b) SDP

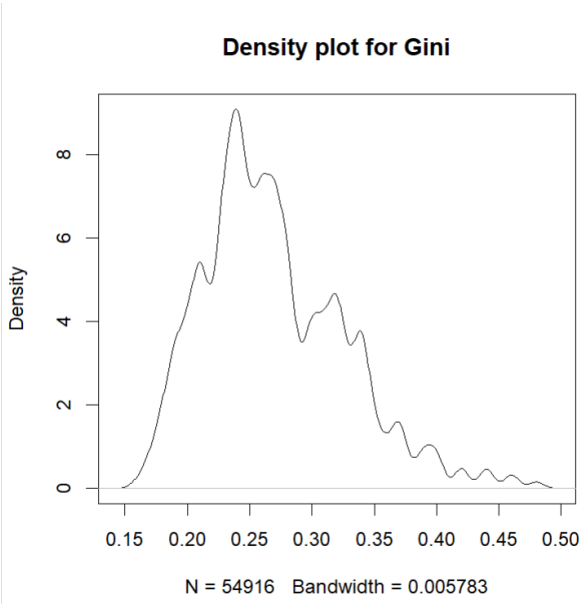

Histogram for SDP
Frequency
0
2000
4000
6000
8000
0
500000
1000000
1500000
SDP

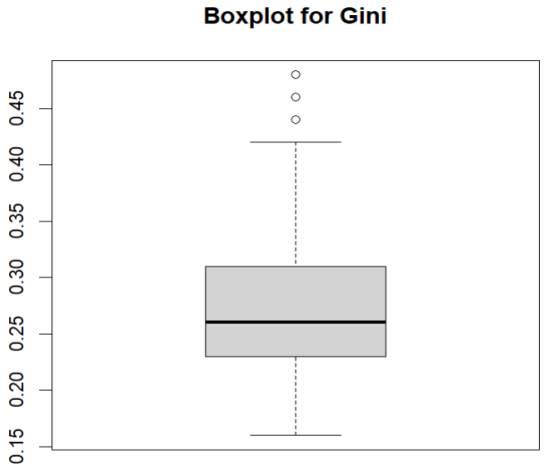

Density plot for SDP
Density
0.0e+00
5.0e-07
1.0e-06
1.5e-06
0
500000
1000000
1500000
N = 54916 Bandwidth = 3.589e+04


Boxplot for SDP
0
500000
1000000
1500000

(c) Gini Index







Yes there are outliers present in the data . The minimum residual of Ground water level is 0.019 and the maximum residual is 169.264 . So since these values are pretty far from the 1Q , median and 3Q values , this suggests that there are outliers in this data . Also in the box plot we can see that there are points that are outside the whiskers of the plot that are the outliers .

```r
#Part5

stargazer(final_df, type = "text", title = "Summary Statistics")


hist(final_df$Ground.water.level, xlab = "Ground Water Level",
     main = "Histogram for Ground Water Level")
hist(final_df$SDP, xlab = "SDP",
     main = "Histogram for SDP")
hist(final_df$Gini, xlab = "Gini",
     main = "Histogram for Gini")


boxplot(final_df$Ground.water.level, main = "Boxplot for Ground Water Level")
boxplot(final_df$SDP, main = "Boxplot for SDP")
boxplot(final_df$Gini, main = "Boxplot for Gini")

plot(density(final_df$Ground.water.level), main = "Density plot for Ground Water Level")
plot(density(final_df$SDP), main = "Density plot for SDP")
plot(density(final_df$Gini), main = "Density plot for Gini")

skewness(final_df$Ground.water.level)
skewness(final_df$SDP)
skewness(final_df$Gini)
```

Q6)

Environmental Quality Indicator $(EQI)_{i,t} = \beta_0 + \beta_1 SDP_{i,t} + u_{i,t}$

In this section we created a linear regression model using the "lm" function, with "Ground.water.level" as the dependent variable and "SDP" as the independent variable.

The coefficients here represent the estimated effect of each SDP variable on the environmental quality indicator, after controlling for the other SDP variables included in the model.

The intercept represents the estimated value of the environmental quality indicator when all the SDP variables are equal to zero.

```r
#Part6

model <- lm(formula = final_df$Ground.water.level ~ final_df$SDP)
summary(model)
```

→ Output :

```
Call:
lm(formula = final_df$Ground.water.level ~ final_df$SDP)

Residuals:
    Min      1Q  Median      3Q     Max
-10.329  -5.417  -3.519   0.566 159.448

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.670e+00  1.031e-01  93.767  < 2e-16 ***
final_df$SDP 6.100e-07  1.573e-07   3.879 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.88 on 54914 degrees of freedom
Multiple R-squared:  0.0002739, Adjusted R-squared:  0.0002557
F-statistic: 15.05 on 1 and 54914 DF,  p-value: 0.000105
```

The output shows a linear regression model fitted to predict the values of Ground water level based on SDP.

The model has an intercept of 9.67 indicating that the expected Ground water level is 9.67m when SDP is 0. Here, slope is 6.1×10^{-7}, meaning that for every unit increase in SDP, the predicted Ground water Level increases by 6.1×10^{-7} , all else held constant. As we can see this change is very small and thus we can say the change in ground water level with the unit change in SDP is negligible.

There is a statistically significant correlation between SDP and Ground water Level as shown by the p-value for the coefficient of SDP which is 0.000105, and less than 0.05.

About 0.027% of the variation in Ground water Level is explained by SDP according to the model's R-squared value of 0.0002739. The corrected R-squared value, which accounts for the number of predictors in the model, is a

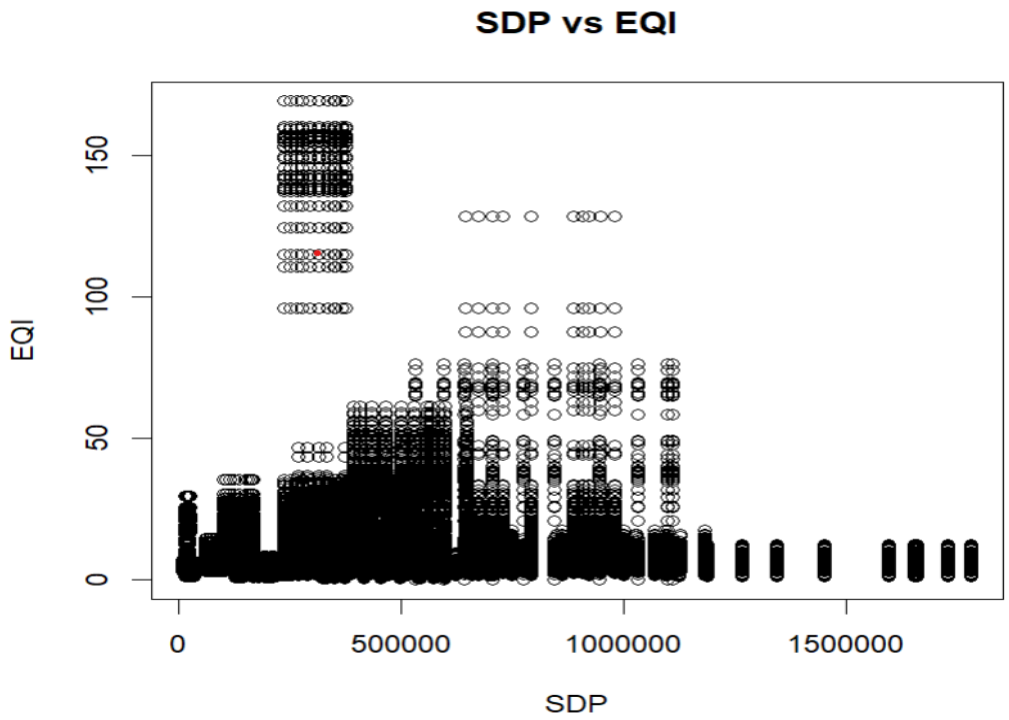
little lower at 0.0002557.The residual standard error (RSE) is 13.88, which is the estimate of the standard deviation of the errors.

In general, the model predicts a weakly positive association between SDP and Ground water Level although the predictor only partially accounts for the response variable's variability.

Q7)

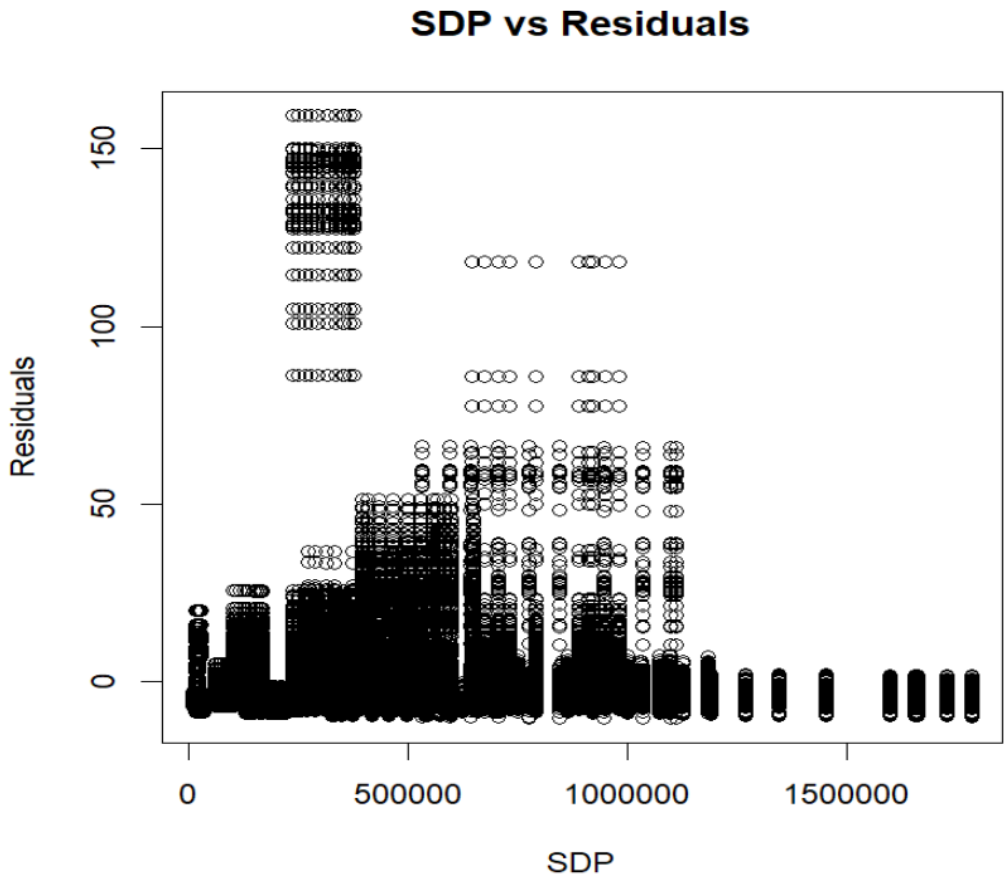
```
#Part7 (Need to confirm this)

residuals = resid(model)
plot(final_df$SDP, final_df$Ground.water.level, xlab = "SDP", ylab = "EQI",
     main = "SDP vs EQI")

plot(final_df$SDP, residuals, xlab = "SDP", ylab = "Residuals",
     main = "SDP vs Residuals")

predicted = predict(model)
plot(final_df$Ground.water.level, predicted, xlab = "True value",
     ylab = "Predicted value", main = "True Value vs Predicted Value of EQI")
```

→ First Plot :



The relationship between the predictor variable (SDP) and the responder variable (Ground.water.level) is depicted in the first plot *(plot(final df$SDP, final df$Ground.water.level, xlab = "SDP", ylab = "EQI", main = "SDP vs EQI")*.
This plot demonstrates how effectively the dependent and independent variables are related in the present regression model. It shows the distribution of the residuals across the plot. The residuals should be distributed symmetrically along the line of best fit in a good fit.

→ Second Plot :

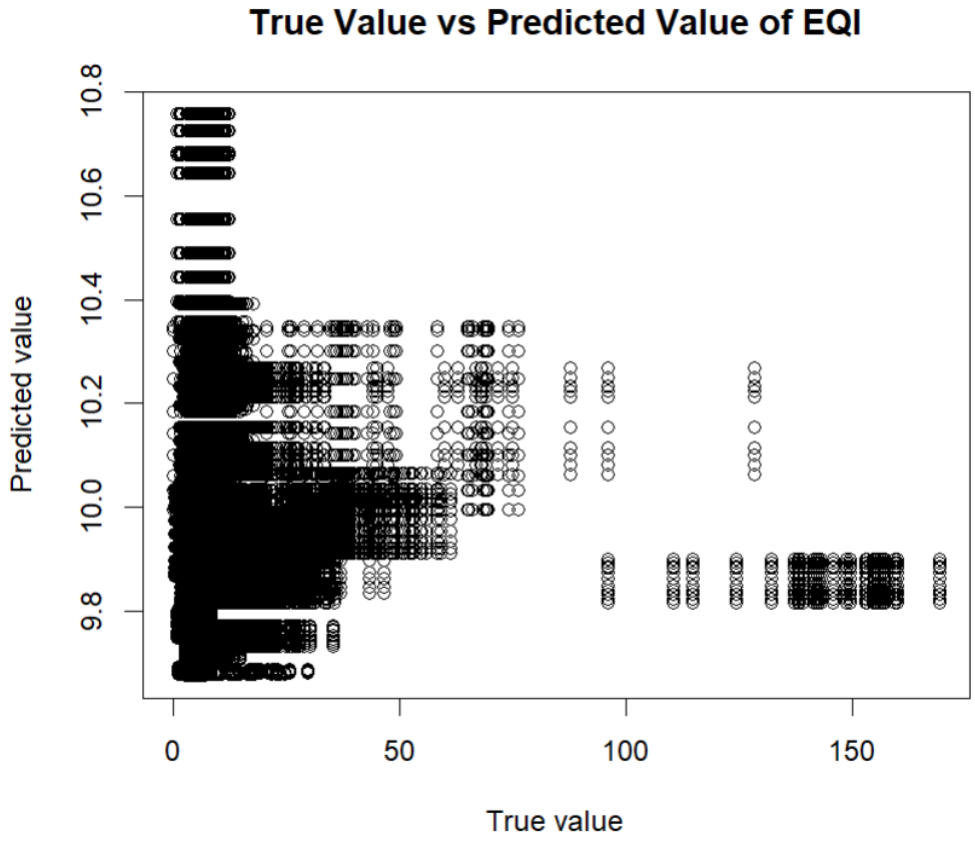


The second plot *(plot(final df$SDP, residuals, xlab = "SDP", ylab = "Residuals", main = "SDP versus Residuals"))* illustrates the link between the predictor variable (SDP) and the residuals of the linear regression model.
SDP is written on the x-axis, and "Residuals" is written on the y-axis. "SDP vs Residuals" is the plot's official title.

The independent variable (SDP) is on the x-axis of the second plot, which is a residual plot with residuals on the y-axis. A regression model makes the assumption that the variance of residuals is constant, i.e., that they are unrelated to one another, and that their anticipated value is zero. This plot aids in determining whether the model adheres to certain presumptions.

A decent fit in this plot would show most of the points close to the y=0 line. A normal residual distribution would be indicated by a symmetric distribution around the y=0 line. Because it depicts the residual distances that were previously plotted—basically, the difference between the true and anticipated values of the EQI—this map is connected to the one before it.

→ Third Plot :



The link between the actual values and anticipated values of the ground water level is depicted in this plot. The y-axis displays predicted values for ground water level, while the x-axis displays true values for ground water level.Predicted values should be near enough to the actual values.

This suggests that a good fit requires a high density of points along the plot's diagonal line. Outliers are points that are far from the diagonal line and are poorly explained by the model. The points on this plot should be somewhat close to the diagonal line if the distribution of residuals is normal, as suggested by the second plot.

Q8)

```
#Part8

sum_residuals = sum(residuals)

hist(residuals, xlab="Residuals", main = "Histogram of model residuals")
```

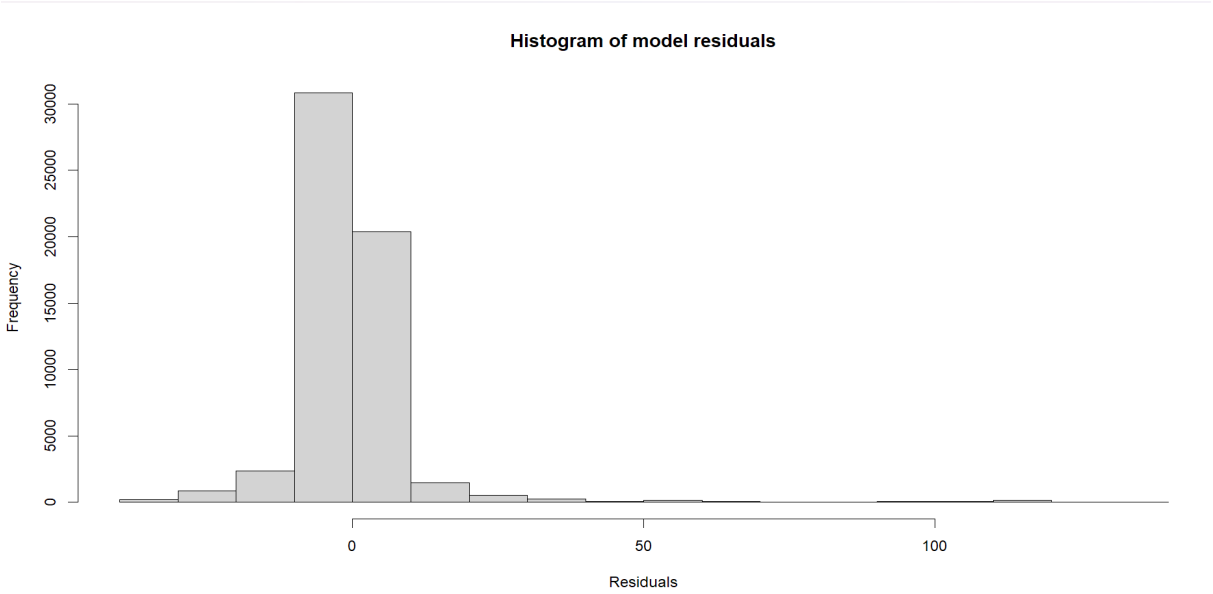
sum_residuals = sum(residuals)

→ The aboveline calculates the sum of the residuals and assigns it to the variable sum_residuals.

hist(residuals, xlab="Residuals", main = "Histogram of model residuals")

→ Using the hist() function, a residuals histogram is produced in the second line. The xlab argument defines the label for the x-axis, which in this case is "Residuals," and the residuals variable represents the input data. The plot's title, "Histogram of model residuals," is given in the main argument.

Histogram :



→ (IInd Part)

To verify that $\sum_{i,t} \hat{u}_{i,t} = 0$:
We can use the following code,

```
if (abs(sum_residuals) < 1e-10) {
  print("The sum of residuals is close to zero.")
} else {
  print("The sum of residuals is not close to zero.")
}
```

→ In order to determine whether the sum of residuals is close to zero or not, the code determines whether the absolute value of sum residuals is less than 1e-10 (i.e., 0.0000000001).

→ The assumption that ûi,t = 0 is verified if the output is "The sum of residuals is near to zero." This signifies that the sum of residuals is indeed close to zero.

→ Our Output :

```
> if (abs(sum_residuals) < 1e-10) {
+   print("The sum of residuals is close to zero.")
+ } else {
+   print("The sum of residuals is not close to zero.")
+ }
[1] "The sum of residuals is close to zero."
```

Hence, $\sum_{i,t} \hat{u}_{i,\,t} \approx 0$ proved.

Q9)

```
#Part9

SDP <- final_df$SDP
SDP2 <- as.numeric(final_df$SDP)^2
SDP3 <- as.numeric(final_df$SDP)^3
gini <- final_df$Gini

model2 <- lm(formula = final_df$Ground.water.level ~ SDP + SDP2 + SDP3 + gini)
summary(model2)
```

→ $EQI_{i,t} = \alpha_0 + \alpha_1 SDP_{i,t} + \alpha_2 SDP_{i.t}^2 + \alpha_3 SDP_{i,t}^3 + \alpha_4 GINI_i + \gamma_{i,t}$

Running This model we get this from R

```
Call:
lm(formula = final_df$Ground.water.level ~ SDP + SDP2 + SDP3 +
    gini)

Residuals:
    Min      1Q  Median      3Q     Max
-13.634  -5.343  -2.822   0.921 158.396

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.388e+00  3.302e-01  -7.233 4.79e-13 ***
SDP          3.013e-05  1.102e-06  27.326  < 2e-16 ***
SDP2        -3.529e-11  1.604e-12 -21.999  < 2e-16 ***
SDP3         1.075e-17  6.384e-19  16.834  < 2e-16 ***
gini         2.476e+01  1.026e+00  24.134  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.66 on 54911 degrees of freedom
Multiple R-squared:  0.03264,   Adjusted R-squared:  0.03257
F-statistic: 463.2 on 4 and 54911 DF,  p-value: < 2.2e-16
```

→ Summarized Results

Dependent Variable >> Ground Water Level [N = 54,916 ; R^2 = 0.033]	
Explanatory Variables	Coefficients
Intercept	-2.388 *** (0.33)
SDP	3.013×10^{-5} *** (1.102×10^{-6})
SDP^2	-3.529×10^{-11} *** (1.604×10^{-12})
SDP^3	1.075×10^{-17} *** (6.384×10^{-19})
GINI	24.76 *** (1.026)

*** : p-value less than 0.001

→ Interpretation in Plain English

α_0 → Represents average Ground water level when SDP=0 and GINI Index =0

$\alpha_1 + \alpha_2 + \alpha_3$ → Represents average decline in Ground water level due to unit increase in SDP and when GINI Index=0

α_4 → Represents average decline in Ground water level due to unit increase in GINI Index and when SDP=0

However, our Summarized Table clearly shows the effect on Ground water Level from SDP^2 and SDP^3 is almost negligible… Thus we have, $\alpha_1 + \alpha_2 + \alpha_3 \approx \alpha_1$ or, in other words, α_1 approximately represents average decline in Ground water level due to unit increase in SDP and when GINI Index=0 (as α_2 and α_3 tend to zero)

--*-* END OF REPORT *-*-*-*

Group-14

Name and Roll no. of Teammates

- Animesh Pareek (2021131)
- Divyam Khorwal (2021388)
- Mohit Bansal (2019257)
- Sejal Khurana (2021418)