Econometrics Data Assignment 1
Report

Siddhart Rajput (2021102)

In this project, we are examining the relationship between economic growth, income inequality, and environmental quality in India. We shall be exploring the Environmental Kuznets Curve hypothesis, which suggests that as per capita income increases, environmental quality first deteriorates and then improves. We will use district-level data on environmental quality, economic output, and income inequality to estimate regression models and analyze the relationship between these variables.

The ultimate goal of completing this assignment is to gain a better understanding of how economic growth and income inequality impact the environment in India and provide insights that can inform policy decisions aimed at promoting sustainable development.

1. In the first question we are supposed to choose one environmental quality measure from the sheet provided to us. We have chosen Ground Water Quality as our quality measure and loaded the NDAP data into a dataframe.

2. In this part, we transformed the data into a district-year level dataset that includes a unique district-year ID for each row in the sample. We have converted all the text in the 'State' column of the dataframe to upper case using the 'toupper()' function so that it becomes easier to format the data in the later process.

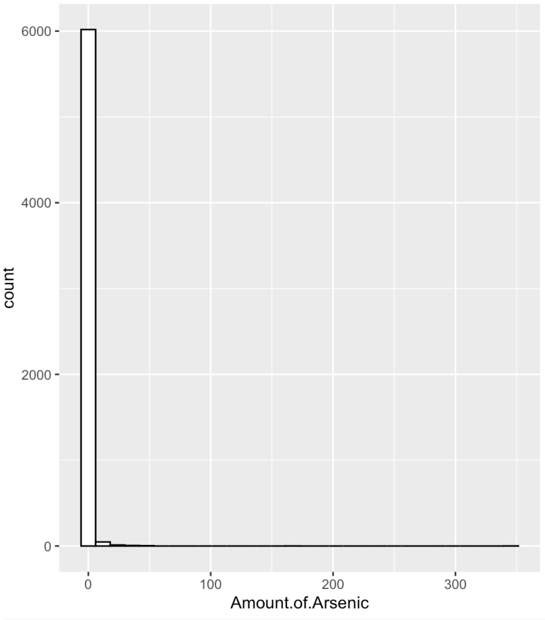
3. Here, we have to merge the district-year level environmental quality data with the corresponding state-year wise economic output data, i.e., the net state domestic product (SDP) at constant prices provided by the Reserve Bank of India accessed on the Database for the Indian Economy (DBIE) portal.

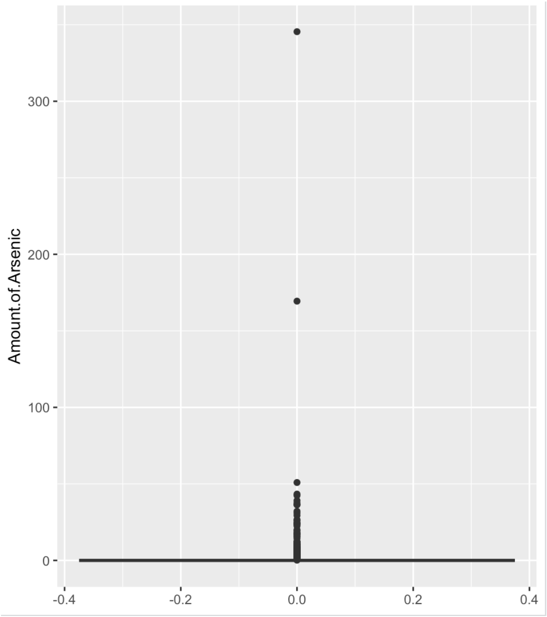
    - So, we first loaded the SDP data in the 'pf' dataframe.
    - We then cleaned the data set by removing the first four rows, changing the format of the 'YEAR' column to include only the first four digits, and including the NA values.
    - Next, we transformed the dataframe to have a Year-State-SDP format using the 'gather' function from the 'tidyr' package.
    - Finally, the 'merged_data' data frame is created by merging the 'pf_f' and 'df' data frames based on the 'YEAR' and 'State' columns.

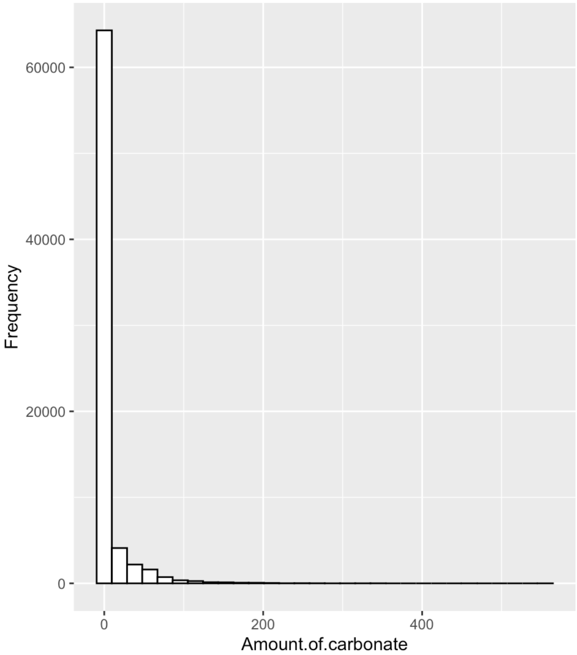
4. We are then required to merge the dataset with the district-level Gini index from the paper Estimates of Poverty and Inequality in the Districts of India, 2011–2012 by Mohanty et al. (2016).

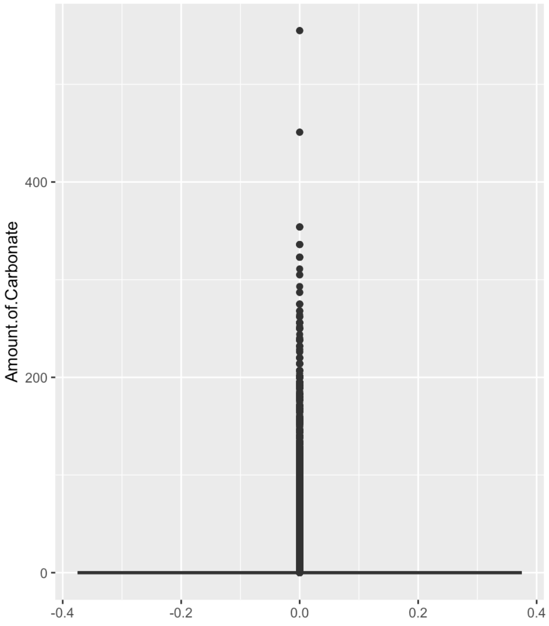
- First, we loaded a csv file containing GINI index data from the pdf file provided to us and renamed the columns.
- Next, we merged the GINI index data by district with the already merged data by NDAP and SDP.
- Finally, we renamed the column in the final merged dataset and saved the data in a csv file naming it ‘merged.csv’.

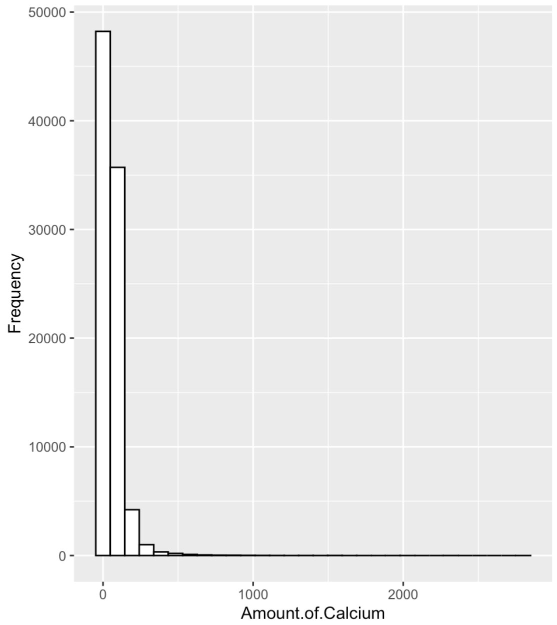
5. Histogram and Box-plot

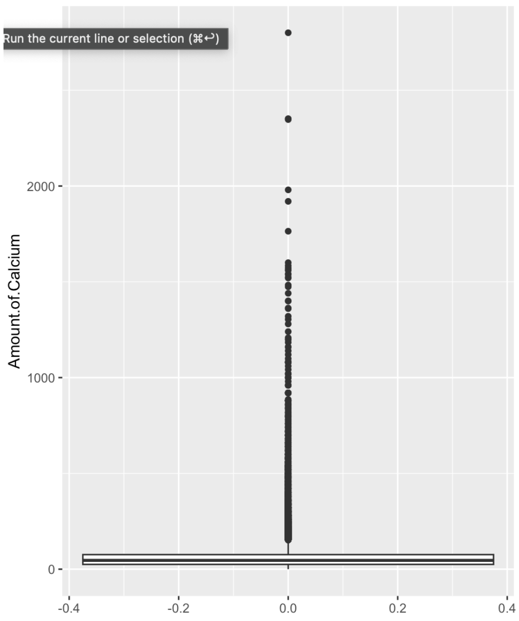


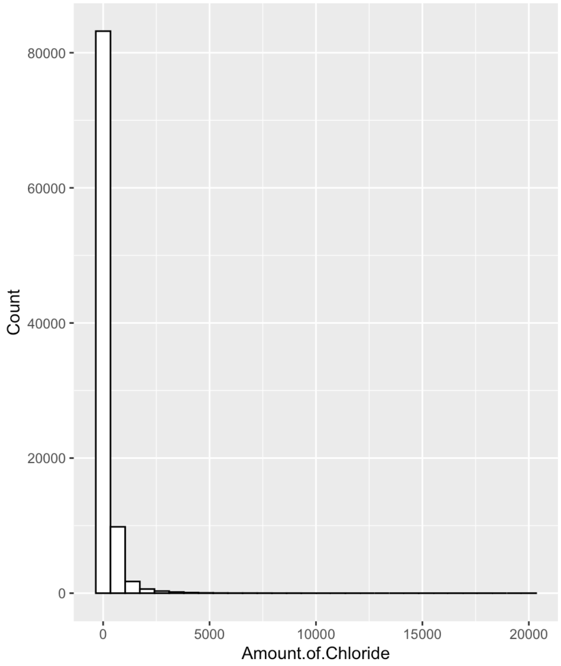


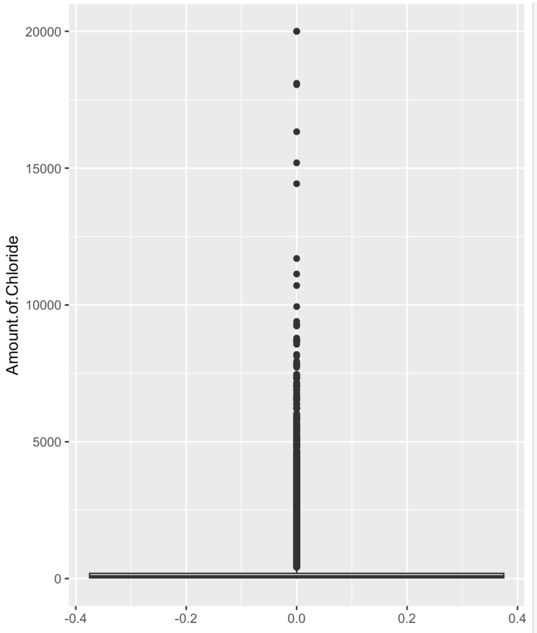


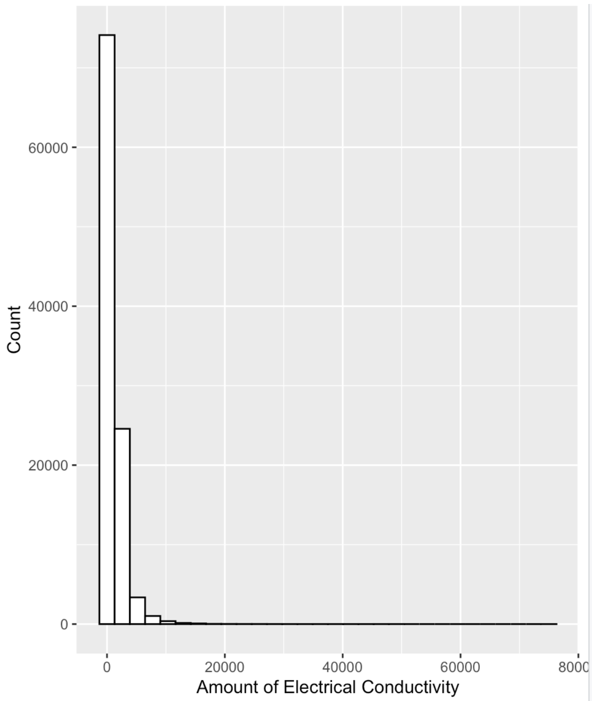


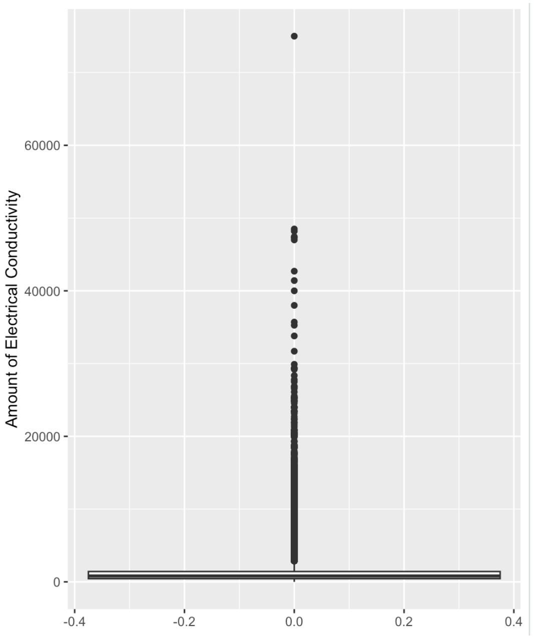

50000
40000
30000
20000
10000
0
Frequency
0
1000
2000
Amount.of.Calcium

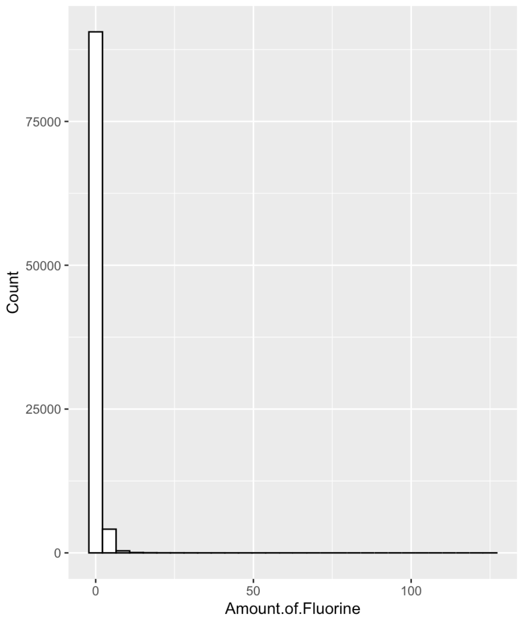

Run the current line or selection (⌘↩)
2000
1000
0
Amount.of.Calcium
-0.4
-0.2
0.0
0.2
0.4

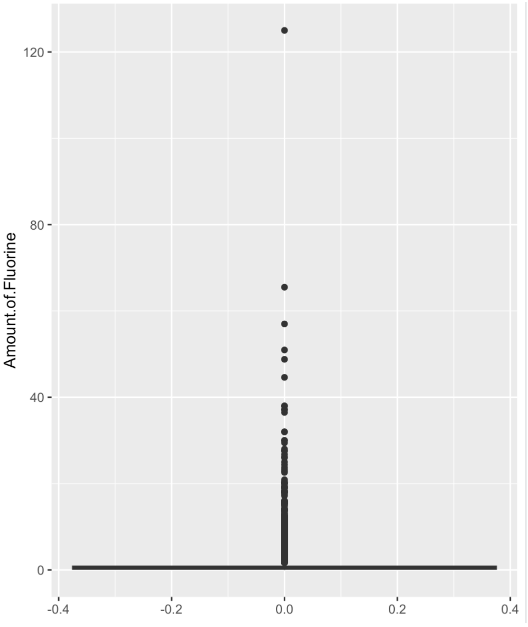

80000
60000
40000
20000
0
Count
0
5000
10000
15000
20000
Amount.of.Chloride

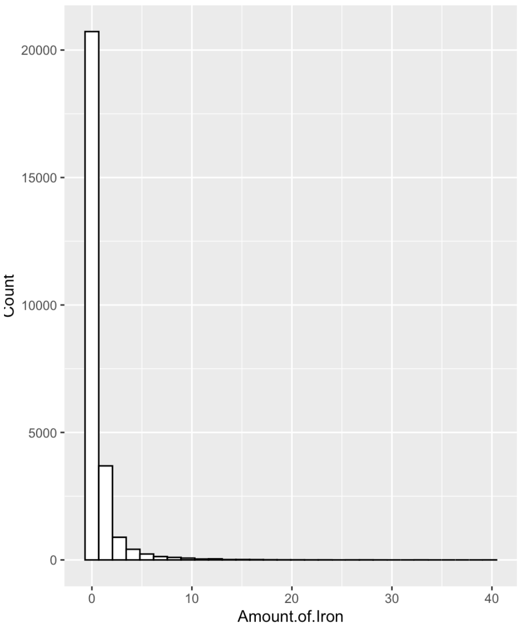

20000
15000
10000
5000
0
Amount.of.Chloride
-0.4
-0.2
0.0
0.2
0.4

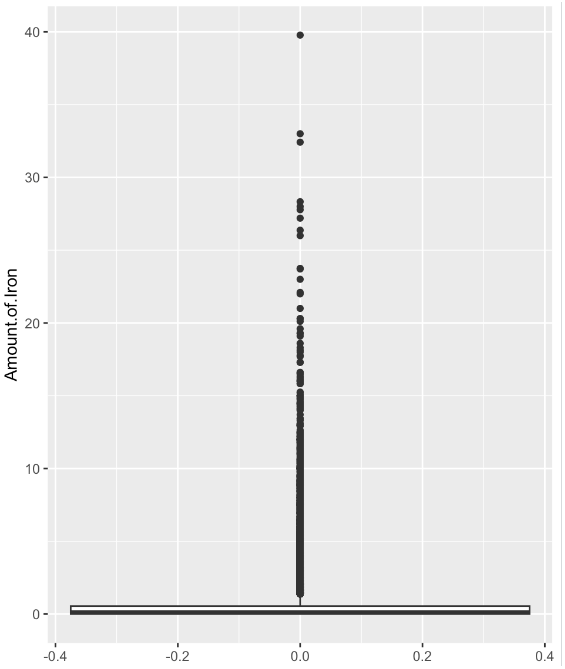

Count
60000
40000
20000
0
0
20000
40000
60000
8000
Amount of Electrical Conductivity

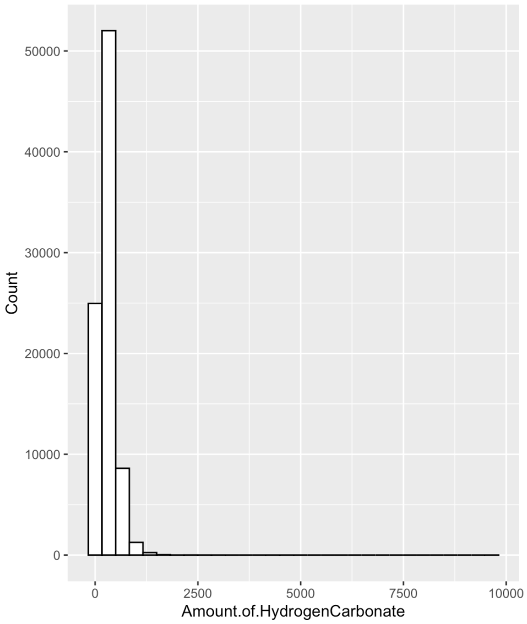

Amount of Electrical Conductivity
60000
40000
20000
0
-0.4
-0.2
0.0
0.2
0.4

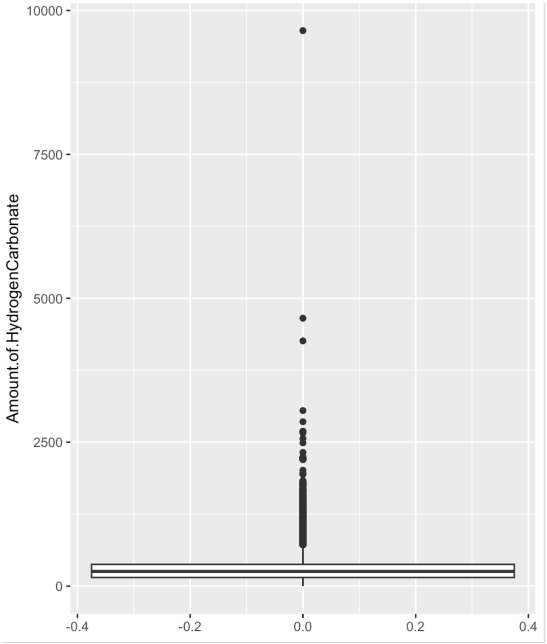

Count
75000
50000
25000
0
0
50
100
Amount.of.Fluorine

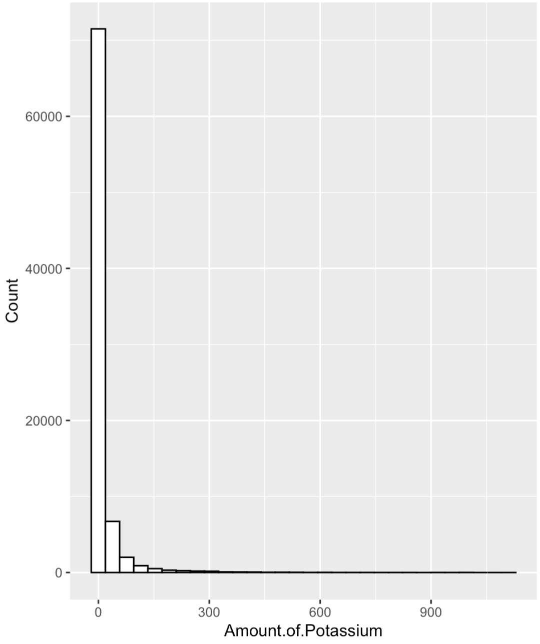

Amount.of.Fluorine
120
80
40
0
-0.4
-0.2
0.0
0.2
0.4

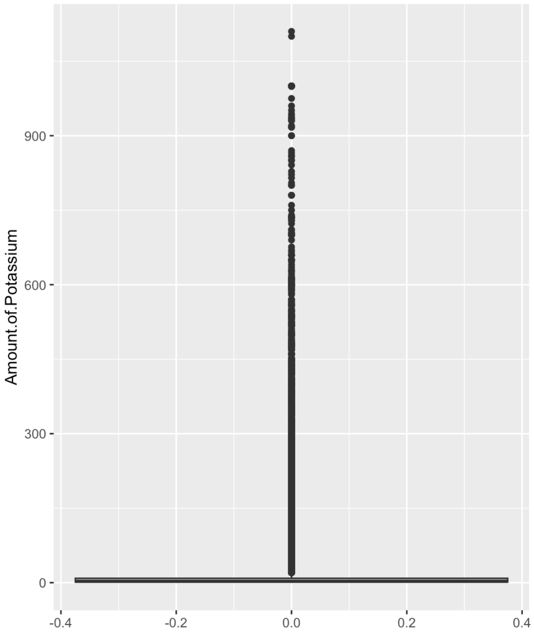

Count
20000
15000
10000
5000
0
0
10
20
30
40
Amount.of.Iron

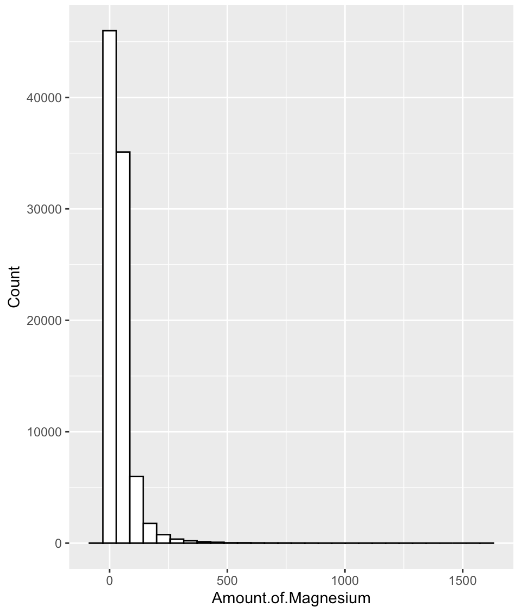

Amount.of.Iron
40
30
20
10
0
-0.4
-0.2
0.0
0.2
0.4

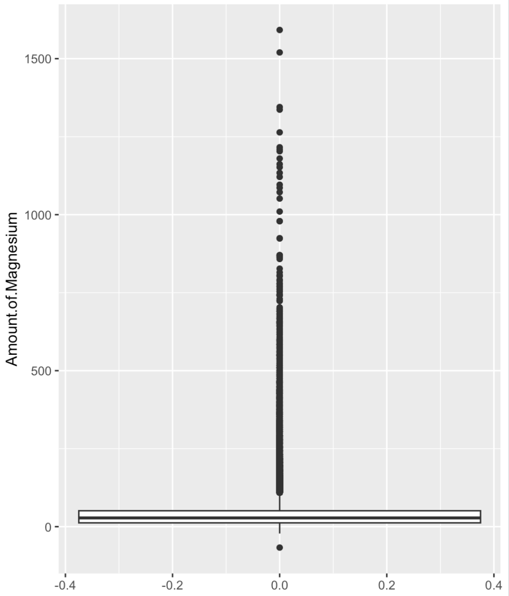

Count
50000
40000
30000
20000
10000
0
0
2500
5000
7500
10000
Amount.of.HydrogenCarbonate

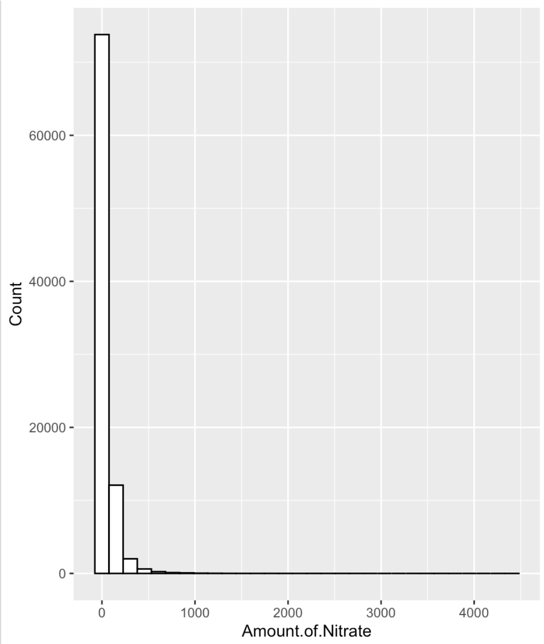

Amount.of.HydrogenCarbonate
10000
7500
5000
2500
0
-0.4
-0.2
0.0
0.2
0.4

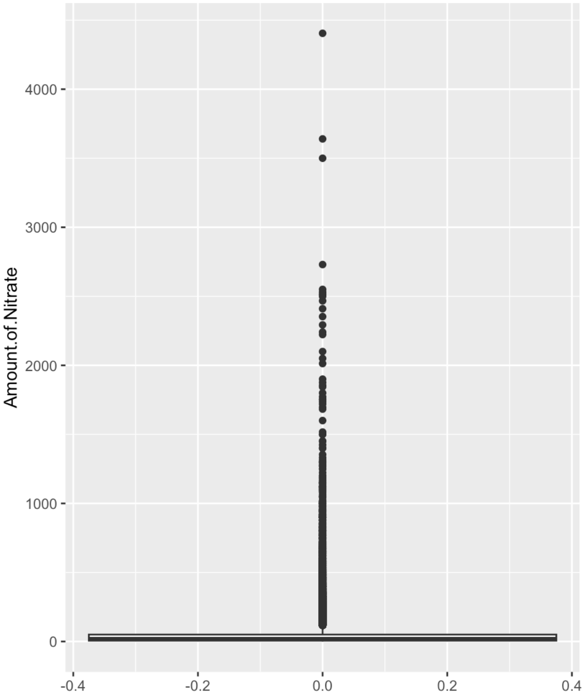

Count
60000
40000
20000
0
0
300
600
900
Amount.of.Potassium

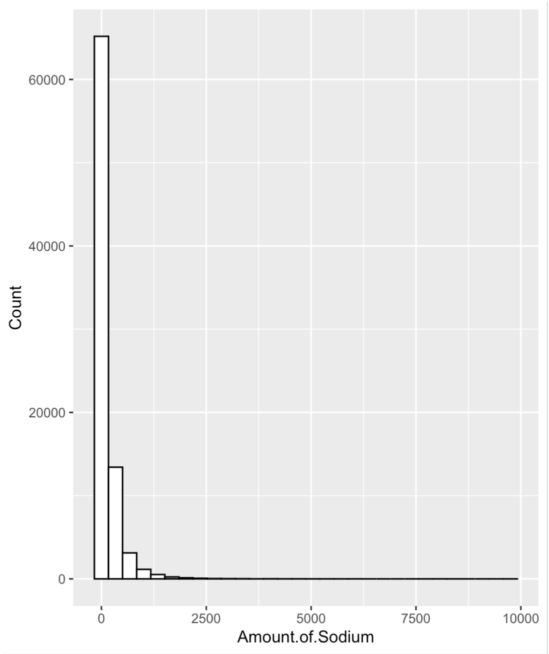

Amount.of.Potassium
900
600
300
0
-0.4
-0.2
0.0
0.2
0.4

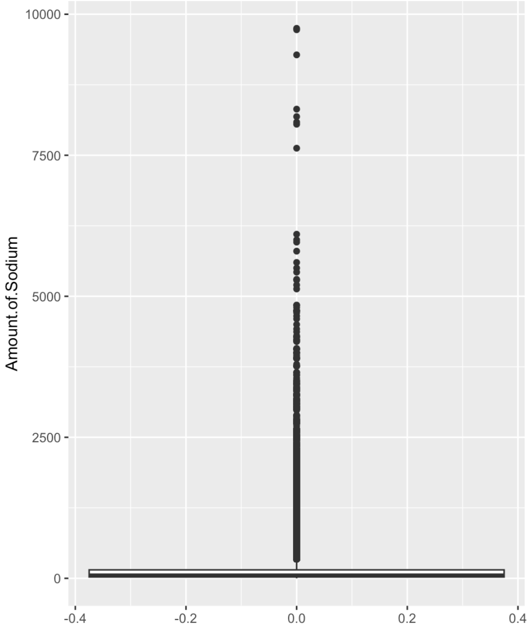

Count
40000
30000
20000
10000
0
0
500
1000
1500
Amount.of.Magnesium

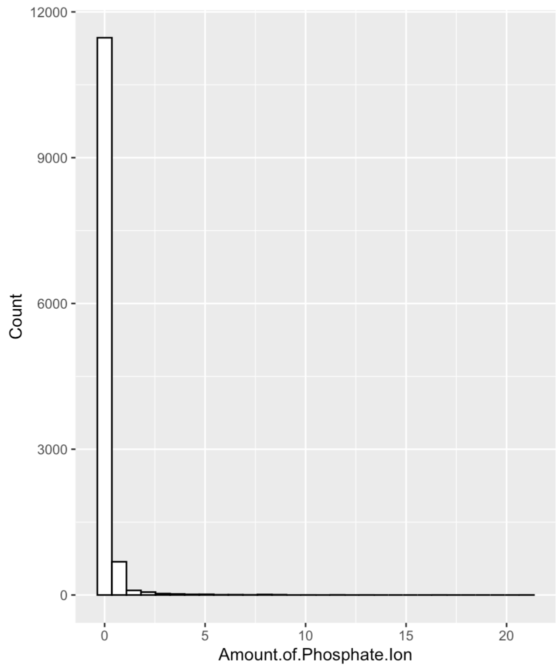

Amount.of.Magnesium
1500
1000
500
0
-0.4
-0.2
0.0
0.2
0.4

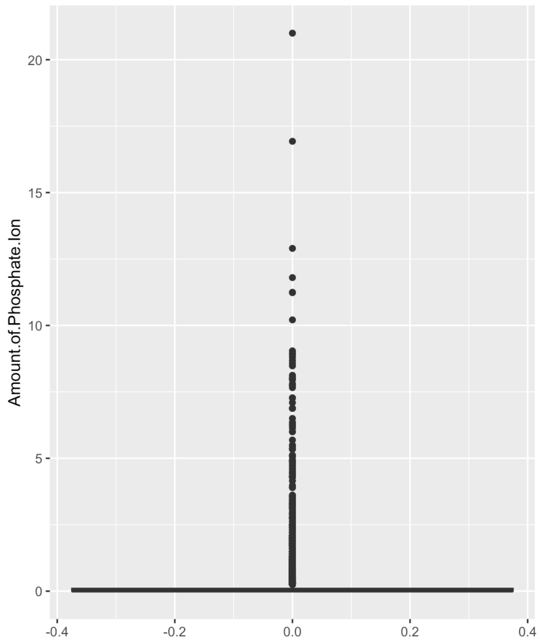

Count
60000
40000
20000
0
0
1000
2000
3000
4000
Amount.of.Nitrate

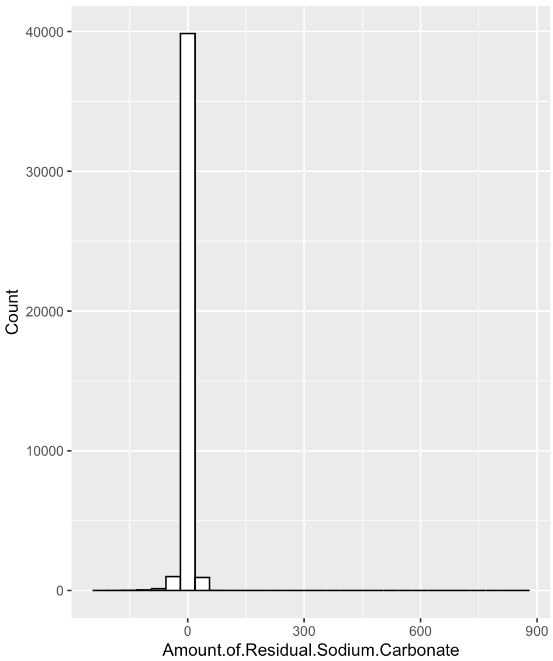

Amount.of.Nitrate
4000
3000
2000
1000
0
-0.4
-0.2
0.0
0.2
0.4

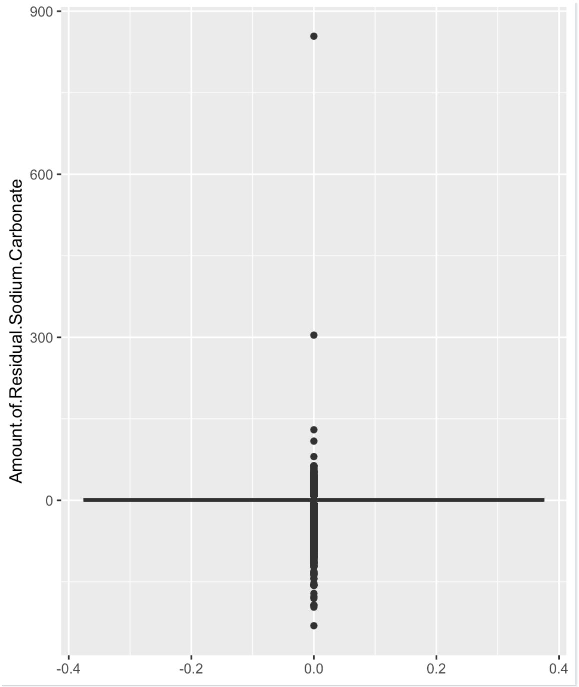

Count
60000
40000
20000
0
0
2500
5000
7500
10000
Amount.of.Sodium

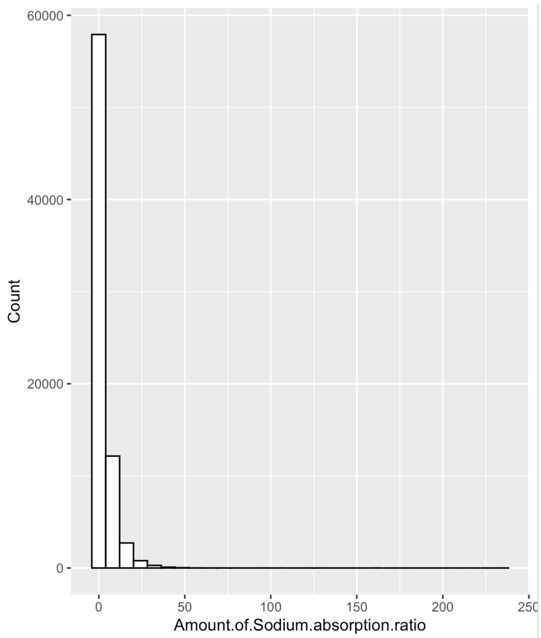

Amount.of.Sodium
10000
7500
5000
2500
0
-0.4
-0.2
0.0
0.2
0.4

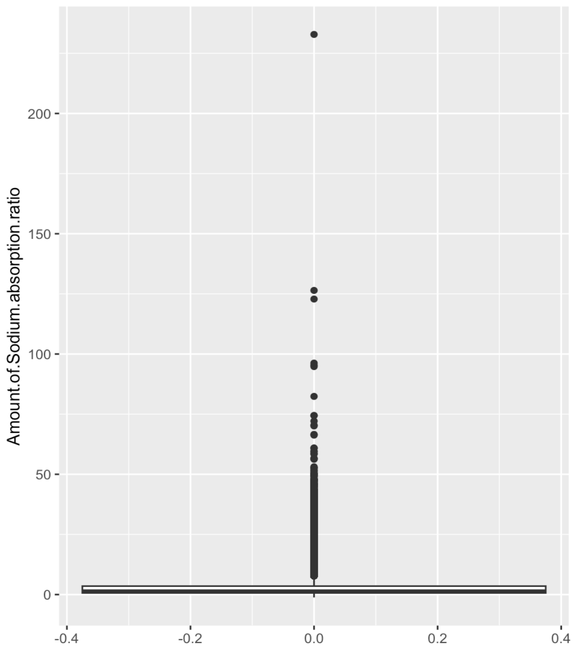

12000
9000
6000
3000
0
Count
0
5
10
15
20
Amount.of.Phosphate.Ion

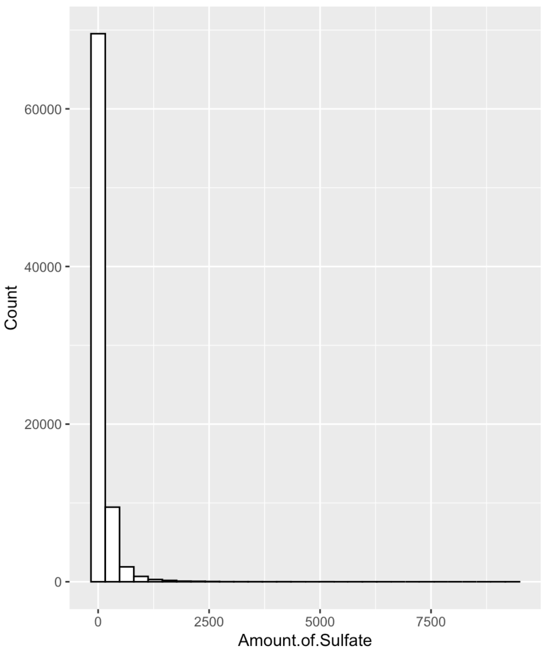

20
15
10
5
0
Amount.of.Phosphate.Ion
-0.4
-0.2
0.0
0.2
0.4

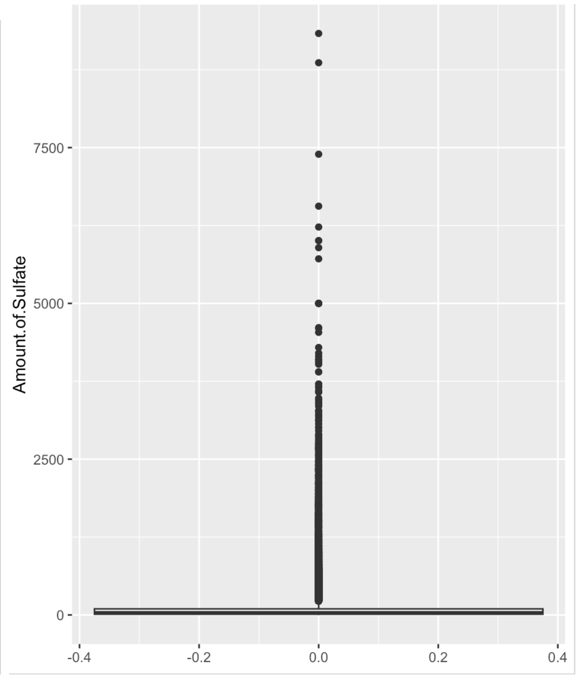

40000
30000
20000
10000
0
Count
0
300
600
900
Amount.of.Residual.Sodium.Carbonate

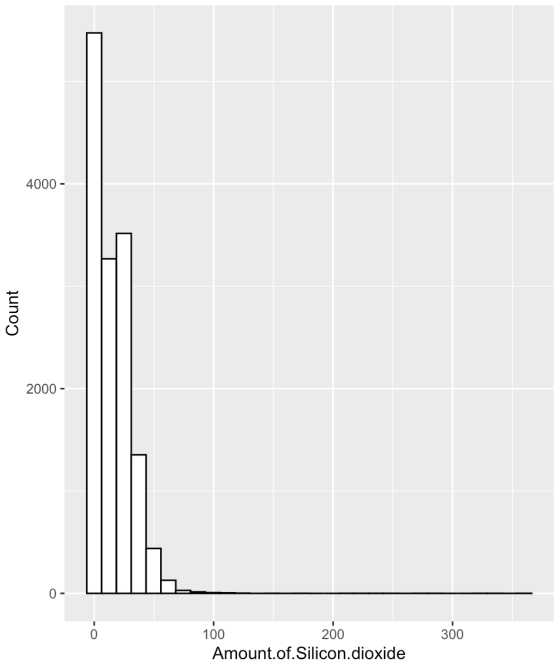

900
600
300
0
Amount.of.Residual.Sodium.Carbonate
-0.4
-0.2
0.0
0.2
0.4

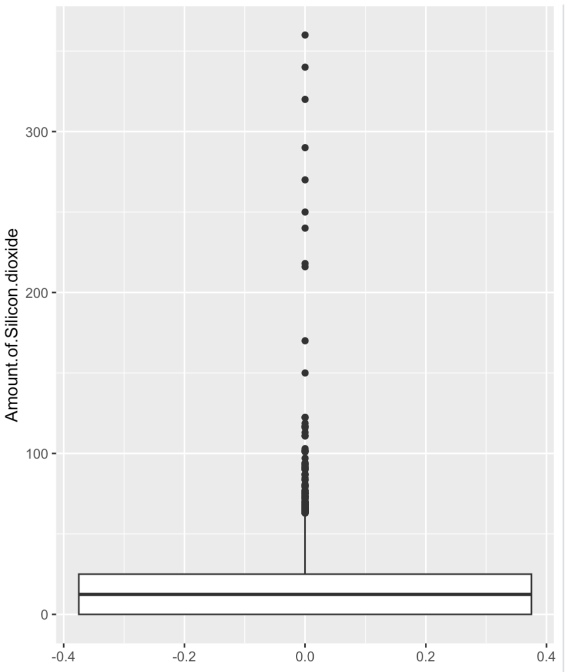

Count
60000
40000
20000
0
0
50
100
150
200
250
Amount.of.Sodium.absorption.ratio

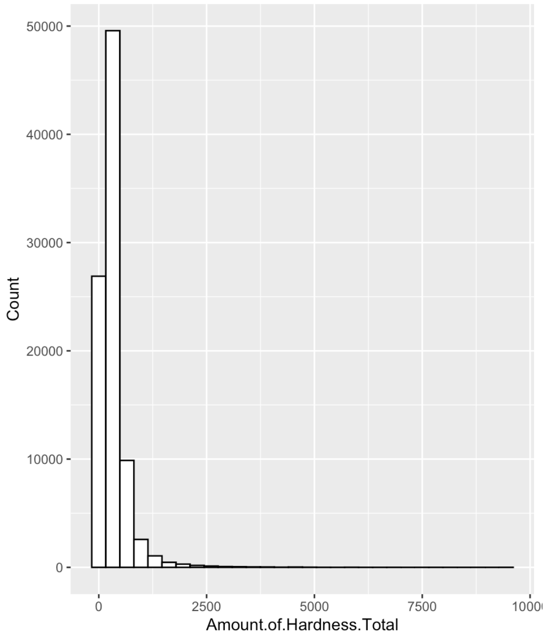

Amount.of.Sodium.absorption.ratio
200
150
100
50
0
-0.4
-0.2
0.0
0.2
0.4

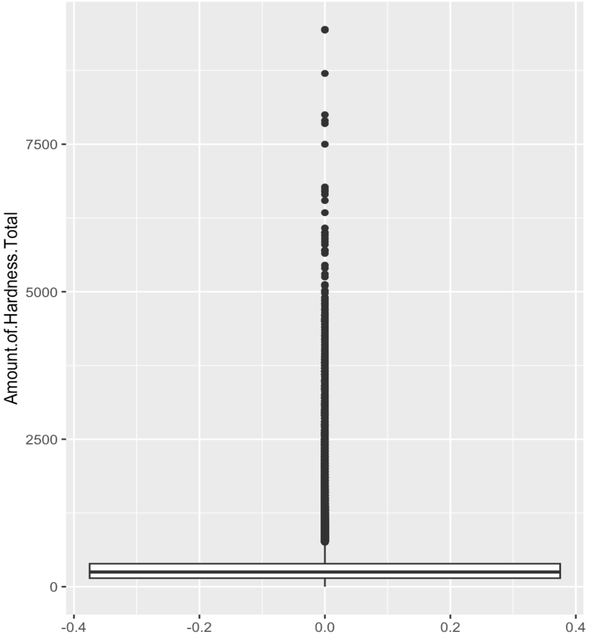

Count
60000
40000
20000
0
0
2500
5000
7500
Amount.of.Sulfate

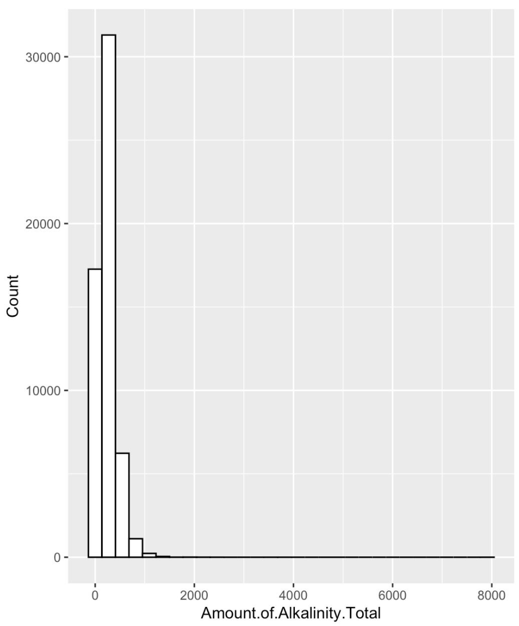

Amount.of.Sulfate
7500
5000
2500
0
-0.4
-0.2
0.0
0.2
0.4

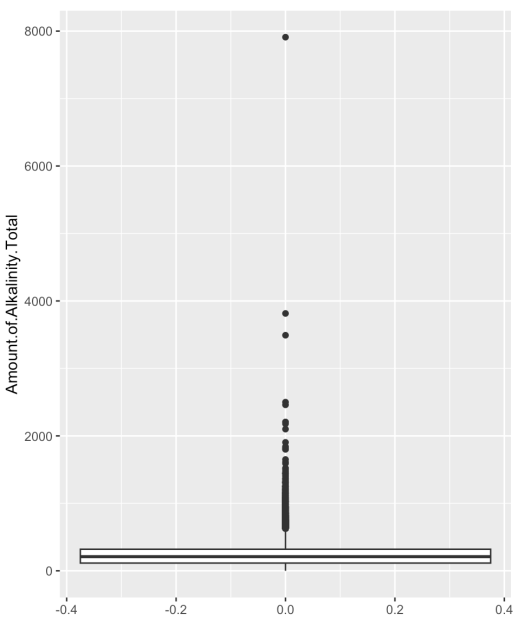

Count
4000
2000
0
0
100
200
300
Amount.of.Silicon.dioxide

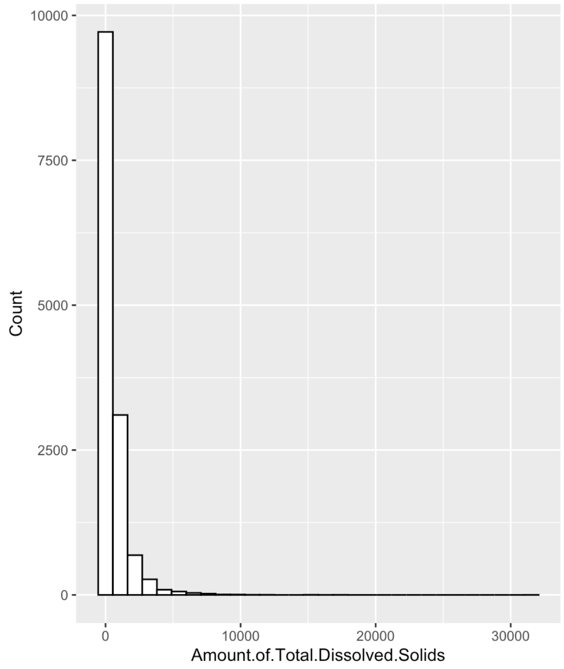

Amount.of.Silicon.dioxide
300
200
100
0
-0.4
-0.2
0.0
0.2
0.4

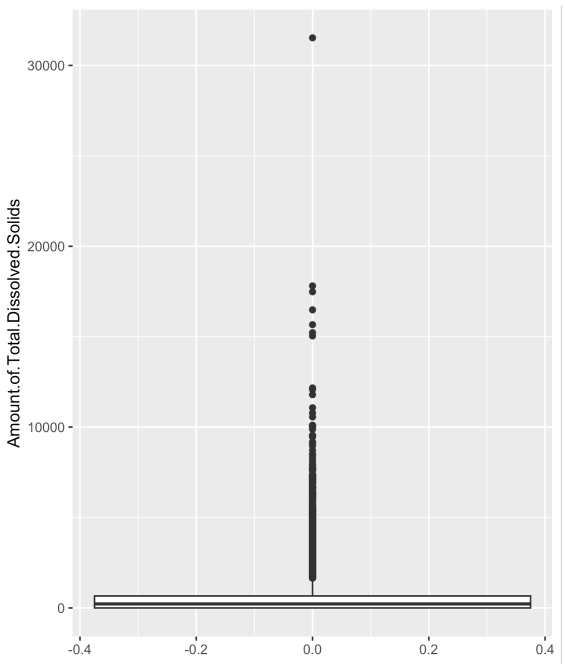

Count
50000
40000
30000
20000
10000
0
0
2500
5000
7500
10000
Amount.of.Hardness.Total

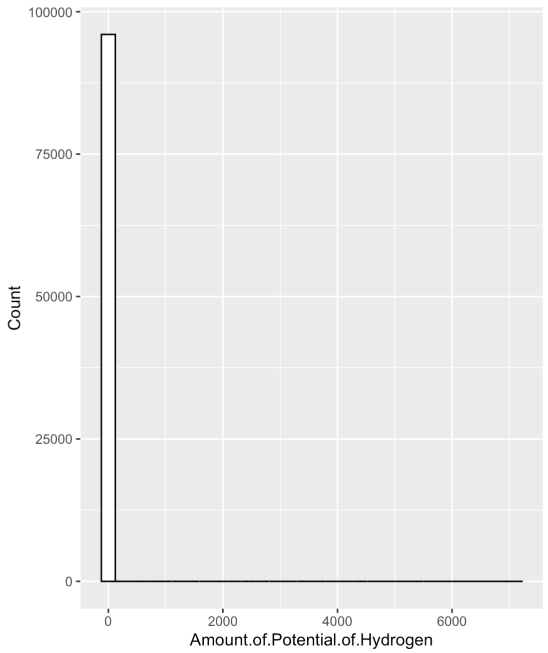

Amount.of.Hardness.Total
7500
5000
2500
0
-0.4
-0.2
0.0
0.2
0.4

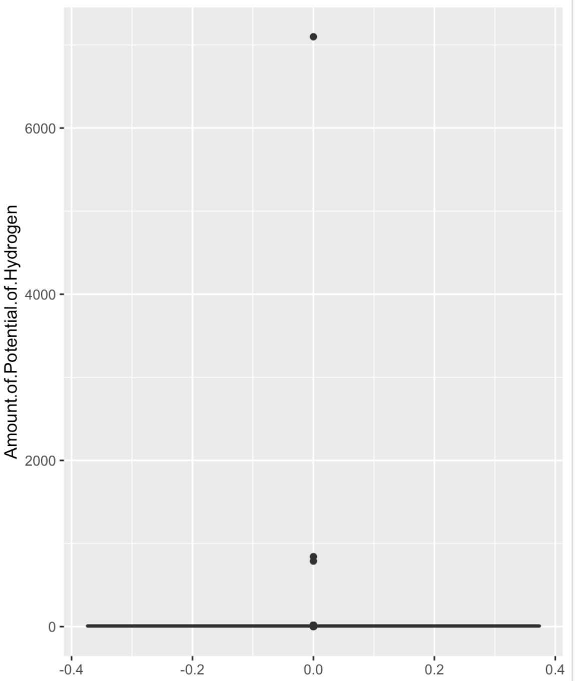

Count
30000
20000
10000
0
0
2000
4000
6000
8000
Amount.of.Alkalinity.Total


Amount.of.Alkalinity.Total
8000
6000
4000
2000
0
-0.4
-0.2
0.0
0.2
0.4


Count
10000
7500
5000
2500
0
0
10000
20000
30000
Amount.of.Total.Dissolved.Solids


Amount.of.Total.Dissolved.Solids
30000
20000
10000
0
-0.4
-0.2
0.0
0.2
0.4



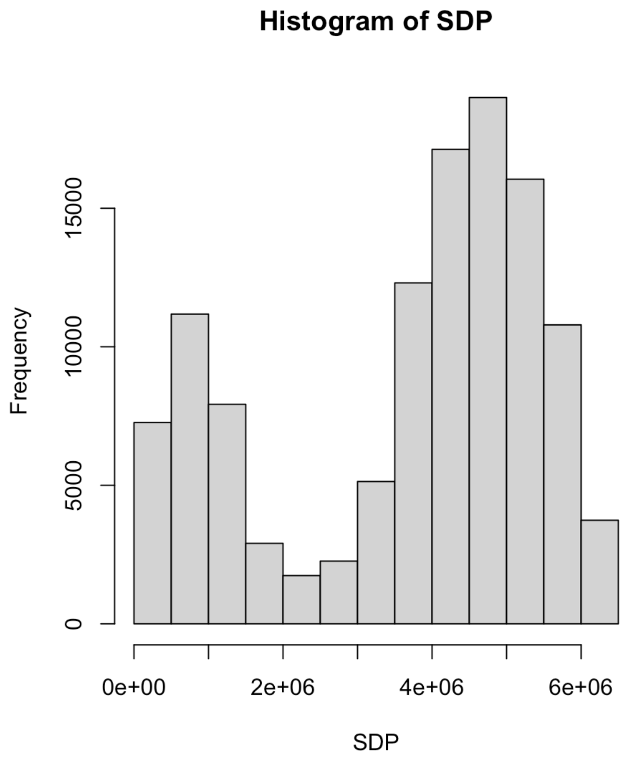


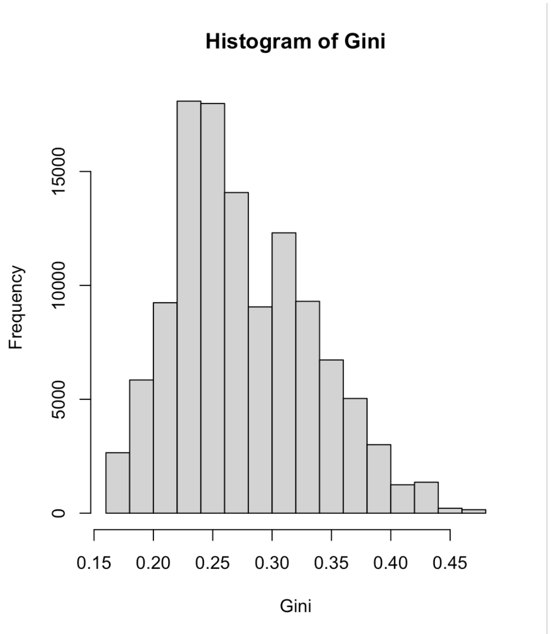
TABLE

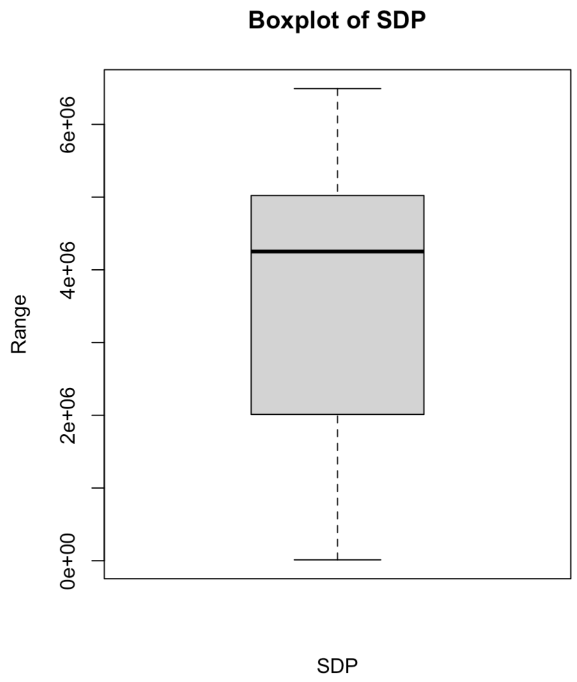
Parameters	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Amount.of.Potential.of.Hydrogen	0.00	7.60	7.91	7.99	8.20	7098.00	8912
Amount.of.Arsenic	0.00	0.00	0.00	0.28	0.00	345.46	98813
Amount.of.carbonate	0.000	0.000	0.000	6.156	0.000	555.000	30861
Amount.of.Calcium	0.00	25.00	46.00	62.33	76.00	2800.00	14936
Amount.of.Chloride	0.0	28.0	71.0	197.6	188.0	20000.0	8825
Amount.of.Electrical.Conductivity	0	444	807	1243	1421	75000	1143
Amount.of.Fluorine	0.000	0.220	0.450	0.715	0.830	125.000	9725
Amount.of.Iron	0.00	0.00	0.14	0.65	0.55	39.78	78495
Amount.of.HydrogenCarbonate	0.0	150.0	256.0	286.3	378.0	9650.0	17739
Amount.of.Potassium	0.00	1.20	3.10	15.36	8.80	1110.00	21910
Amount.of.Magnesium	-67.00	12.20	28.00	41.99	51.00	1592.00	14326

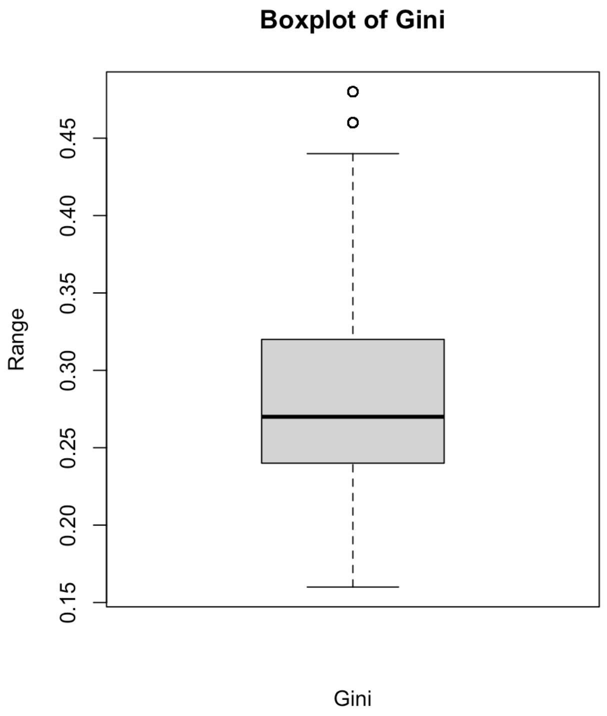
Parameters	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Amount.of.Potential.of.Hydrogen	0.00	7.60	7.91	7.99	8.20	7098.00	8912
Amount.of.Arsenic	0.00	0.00	0.00	0.28	0.00	345.46	98813
Amount.of.carbonate	0.000	0.000	0.000	6.156	0.000	555.000	30861
Amount.of.Calcium	0.00	25.00	46.00	62.33	76.00	2800.00	14936
Amount.of.Nitrate	0.00	5.58	20.00	48.26	50.00	4405.00	15860
Amount.of.Sodium	0.0	23.9	59.0	148.4	150.0	9750.0	20918
Amount.of.Phosphate.Ion	0.00	0.00	0.01	0.15	0.10	21.00	92471
Sodium.absorption.ratio	-1.200	0.660	0.660	1.480	3.182	232.856	30882
Amount.of.Sulfate	-0.466	12.000	37.000	102.506	97.000	9332.000	22637
Amount.of.Silicon.dioxide	0.00	0.00	12.40	15.15	25.00	360.00	90654
Amount.of.Hardness.Total	0.0	145.0	250.0	325.5	390.0	9445.0	13576
Amount.of.Alkalinity.Total	0.0	114.7	209.8	235.4	319.7	7909.8	48697
Amount.of.Total.Dissolved.Solids	0.0	0.0	222.8	549.1	665.5	31525.0	90893
Amount.of.Potential.of.Hydrogen	0.000	7.550	7.890	7.941	8.150	7098.000	8891

Histogram and Box-plot of SDP and Gini Index









There are no Outliers in the Box Plot of SDP as evident from the graph whereas, there are Outliers in Gini Index

SKEWNESS

Parameters	Skewness
SDP	-0.6876161
Gini.Index	0.5358346
Amount.of.Arsenic	50.97384
Amount.of.carbonate	6.328868
Amount.of.Calcium	7.459571
Amount.of.Chloride	10.53928
Amount.of.Electrical.Conductivity	6.974374
Amount.of.Fluorine	22.3695
Amount.of.Iron	6.798305
Amount.of.Hydrogencarbonate	2.854662
Amount.of.Potassium	8.761303
Amount.of.Magnesium	6.187738
Amount.of.Nitrate	9.042039
Amount.of.Sodium	7.704809
Percentage.of.Sodium	1.452093
Amount.of.Phosphate.Ion	12.57951
Amount.of.Residual.Sodium.Carbonate	7.216797
Amount.of.Sodium.absorption.ratio	5.358893

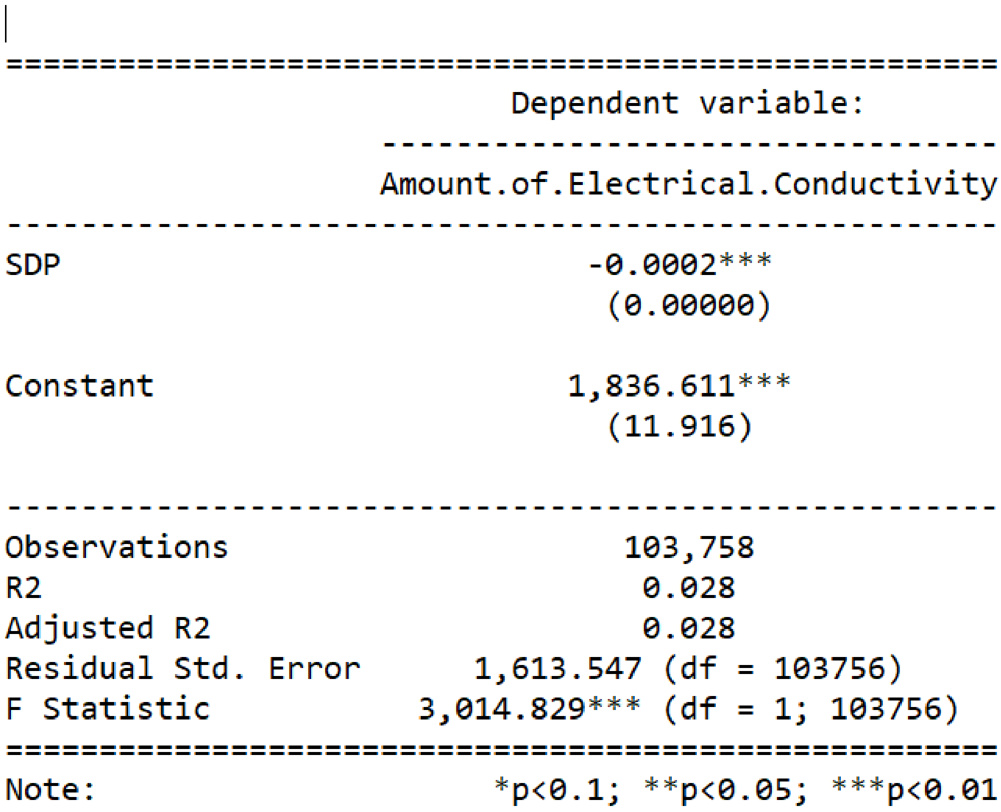
Amount.of.Sulfate	9.125552
Amount.of.Silicon.dioxide	3.641457
Amount.of.Hardness.Total	5.871705
Amount.of.Alkalinity.Total	2.939363
Amount.of.Total.Dissolved.Solids	6.487613
Amount.of.Potential.of.Hydrogen	316.2055

6. Here we are required to choose any one environment quality indicator of our choice and estimate the given regression for the chosen quality indicator. Next, summarise and interpret the results.
   We have chosen the amount of electrical conductivity as our quality indicator.

```
> #QUESTION 6
>
> #Reading the merged CSV file
> nf <- read.csv("/Users/sarvajeethuk/NN.csv")
> View(nf)
> # Omitting the NA values for regression analysis
> nf <- nf[!is.na(nf$Amount.of.Electrical.Conductivity),]
>
> #numeric check
> library(dplyr)
>
> nf <- nf %>%
+   mutate_at(vars(14:35, 37,38), as.numeric)
>
> # We have chosen Amount.of.Electrical.Conductivity as our environmental quality indicator
> #Running the RM
> rm <- lm(formula = merged_data_final$Amount.of.Electrical.Conductivity ~ merged_data_final$SDP, data = nf)
> summary(rm)

Call:
lm(formula = merged_data_final$Amount.of.Electrical.Conductivity ~ 
    merged_data_final$SDP, data = nf)

Residuals:
   Min     1Q Median     3Q    Max 
 -1715   -802   -405    213  73407 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.747e+03  1.075e+01  162.48   <2e-16 ***
merged_data_final$SDP -1.395e-04  2.627e-06  -53.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1608 on 116272 degrees of freedom
  (1148 observations deleted due to missingness)
Multiple R-squared:  0.02369,	Adjusted R-squared:  0.02369 
F-statistic:  2822 on 1 and 116272 DF,  p-value: < 2.2e-16
```

- This is a linear regression analysis of the relationship between the environmental quality indicator (Amount.of.Electrical.Conductivity) and the SDP (State Domestic Product) values.
- Intercept (β_0) : The intercept of the regression line is estimated to be 1747 with a standard error of 10.75. This means that when the SDP is zero, the predicted value of Amount.of.Electrical.Conductivity is 1747.
- SDP (β_1) : The coefficient of SDP variable is estimated to be $-1.395e^{-04}$ with a standard error of $2.627e^{-06}$. This means that for every one unit increase in the SDP, the predicted value of Amount.of.Electrical.Conductivity decreases by $-1.395e^{-04}$.
- Residuals: The model residuals (i.e., the differences between the predicted and actual values of Amount.of.Electrical.Conductivity) have a minimum value of -1715, 1st quartile of -802, median of -405, 3rd quartile of 213 and maximum value of 73407.

- R-squared: The R-squared value of the model is 0.02369, which means that only 2.36% of the variation in the Amount.of.Electrical.Conductivity can be explained by the SDP variable.
- F-statistic: The F-statistic value of the model is 2822 with a p-value of < $2.2e^{-16}$, which means that the regression model is statistically significant and SDP variable is a significant predictor of Amount.of.Electrical.Conductivity.



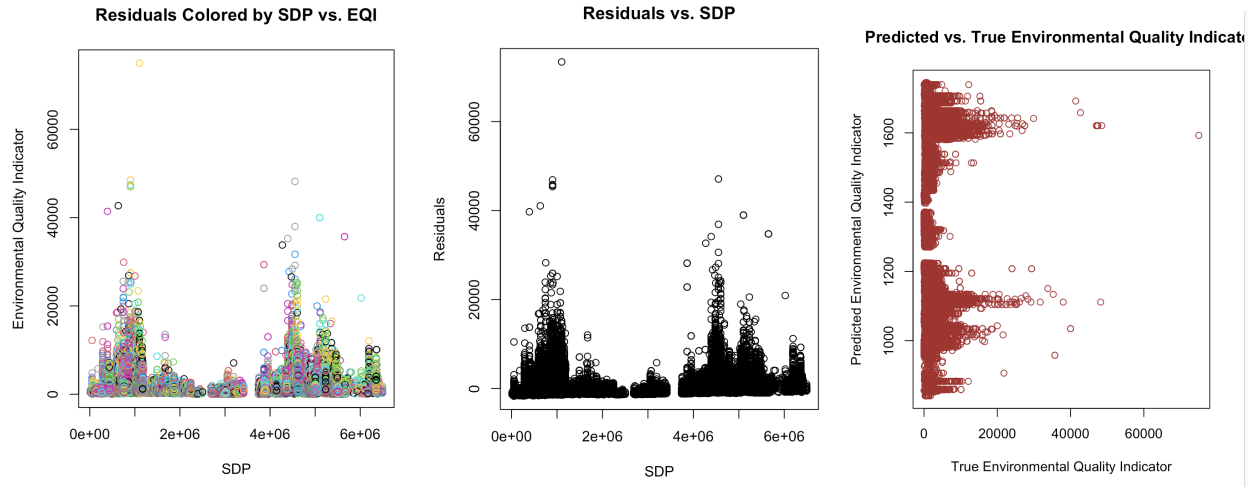
```
=====================================================
                          Dependent variable:
                    ---------------------------------
                    Amount.of.Electrical.Conductivity
-----------------------------------------------------
SDP                            -0.0002***
                                (0.00000)

Constant                       1,836.611***
                                 (11.916)

-----------------------------------------------------
Observations                      103,758
R2                                 0.028
Adjusted R2                        0.028
Residual Std. Error      1,613.547 (df = 103756)
F Statistic           3,014.829*** (df = 1; 103756)
=====================================================
Note:                     *p<0.1; **p<0.05; ***p<0.01
```

7. The first plot shows the relationship between the environmental quality indicator and the SDP. It helps to see if there is a linear or nonlinear relationship between the two variables.

   The second plot shows the relationship between the model residuals and the SDP. It helps to see if there is any pattern or trend in the residuals, which may suggest that the model is misspecified.

   The third plot shows the relationship between the predicted values of the environmental quality indicator and the true values. It helps to evaluate how well the model fits the data, as the closer the points are to the 45-degree line, the better the model fits the data.

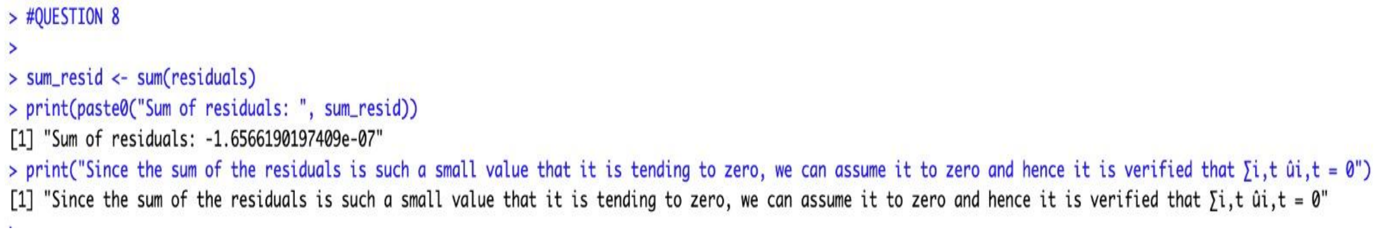
   All three plots are related as they help to evaluate different aspects of the model fit. The first plot helps to evaluate the linearity assumption, the second plot helps to evaluate the assumption of constant variance, and the third plot helps to evaluate the overall model fit.


Residuals Colored by SDP vs. EQI

Residuals vs. SDP

Predicted vs. True Environmental Quality Indicat

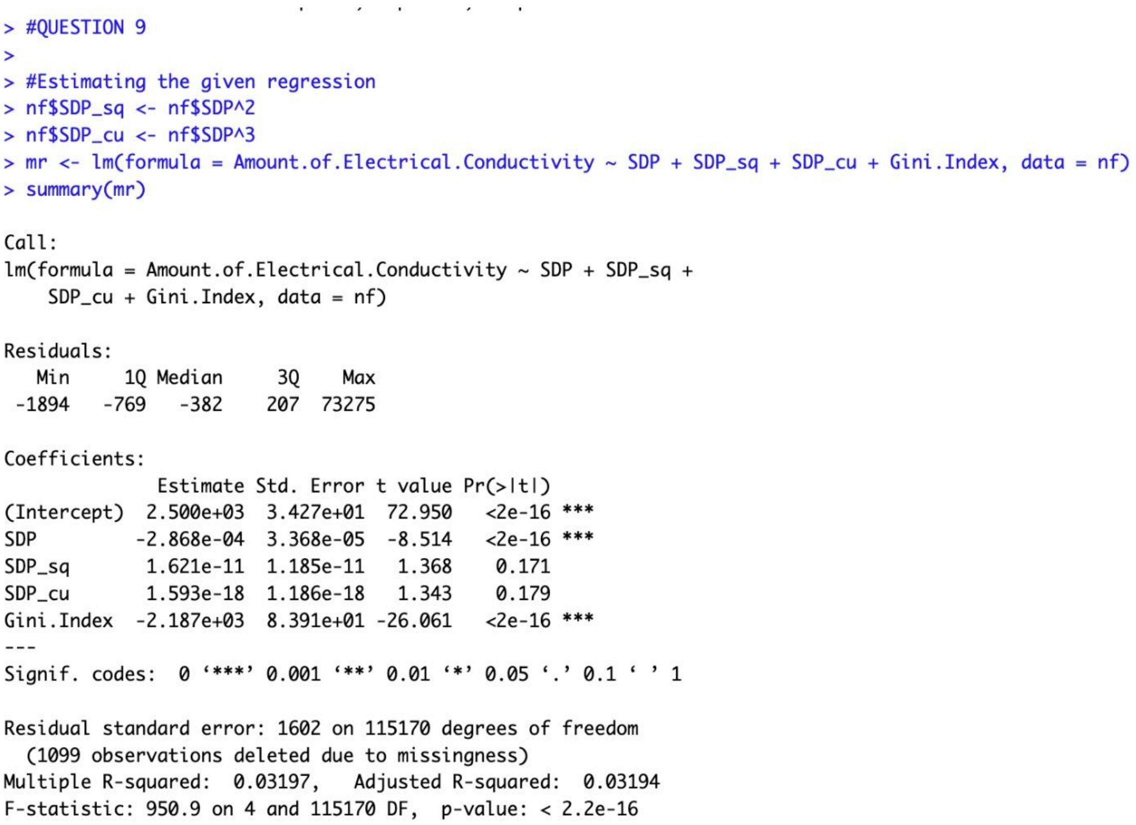
8. Here, we are required to plot a histogram of ûi,t and verify that ∑i,t ûi,t = 0.



```
> #QUESTION 8
>
> sum_resid <- sum(residuals)
> print(paste0("Sum of residuals: ", sum_resid))
[1] "Sum of residuals: -1.6566190197409e-07"
> print("Since the sum of the residuals is such a small value that it is tending to zero, we can assume it to zero and hence it is verified that ∑i,t ûi,t = 0")
[1] "Since the sum of the residuals is such a small value that it is tending to zero, we can assume it to zero and hence it is verified that ∑i,t ûi,t = 0"
```

After calculating the residuals, we obtained a sum of -1.86310575855941e-07. This value is very close to zero, indicating that we can assume that the sum of the residuals (∑i,t ûi,t) is equal to zero. In other words, the model's predictions are almost equal to the actual values, and any differences between them average out to zero. We can express this idea in a more elegant and concise manner by stating that the sum of the residuals is approximately zero, suggesting that the model's predictions are unbiased.

9. Here, we have to estimate the given regression equation and interpret the results. These are the results obtained.



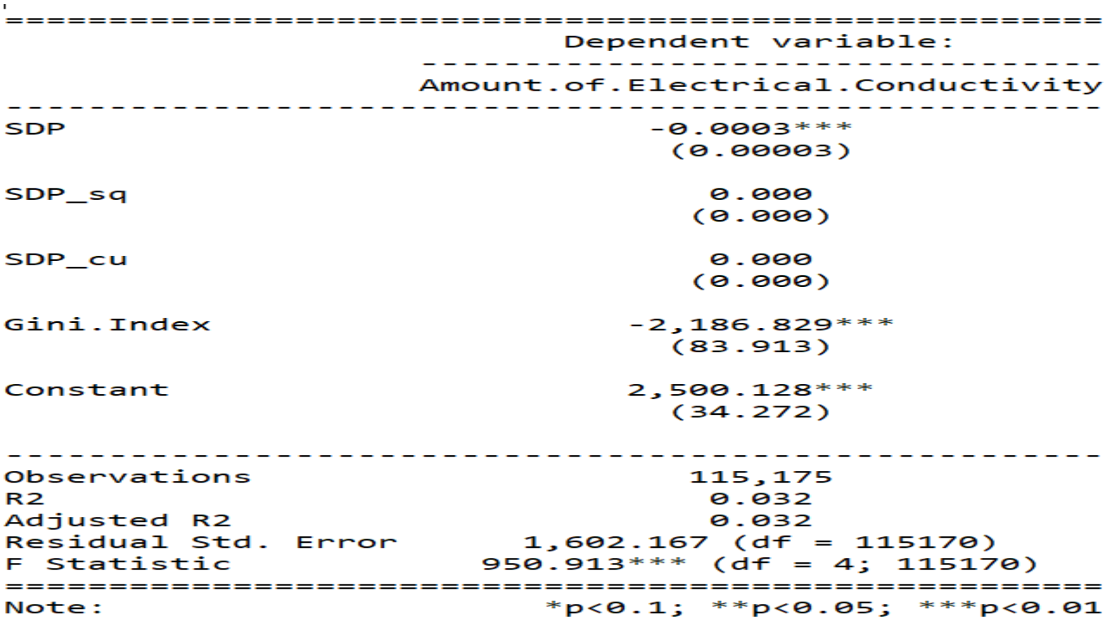
```
> #QUESTION 9
>
> #Estimating the given regression
> nf$SDP_sq <- nf$SDP^2
> nf$SDP_cu <- nf$SDP^3
> mr <- lm(formula = Amount.of.Electrical.Conductivity ~ SDP + SDP_sq + SDP_cu + Gini.Index, data = nf)
> summary(mr)

Call:
lm(formula = Amount.of.Electrical.Conductivity ~ SDP + SDP_sq +
    SDP_cu + Gini.Index, data = nf)

Residuals:
   Min     1Q Median     3Q    Max
 -1894   -769   -382    207  73275

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.500e+03  3.427e+01  72.950   <2e-16 ***
SDP         -2.868e-04  3.368e-05  -8.514   <2e-16 ***
SDP_sq       1.621e-11  1.185e-11   1.368    0.171
SDP_cu       1.593e-18  1.186e-18   1.343    0.179
Gini.Index  -2.187e+03  8.391e+01 -26.061   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1602 on 115170 degrees of freedom
  (1099 observations deleted due to missingness)
Multiple R-squared:  0.03197,   Adjusted R-squared:  0.03194
F-statistic: 950.9 on 4 and 115170 DF,  p-value: < 2.2e-16
```

The regression model estimates the relationship between the Amount of Electrical Conductivity and the SDP, SDP^2, SDP^3, as well as the Gini Index, which measures income inequality.

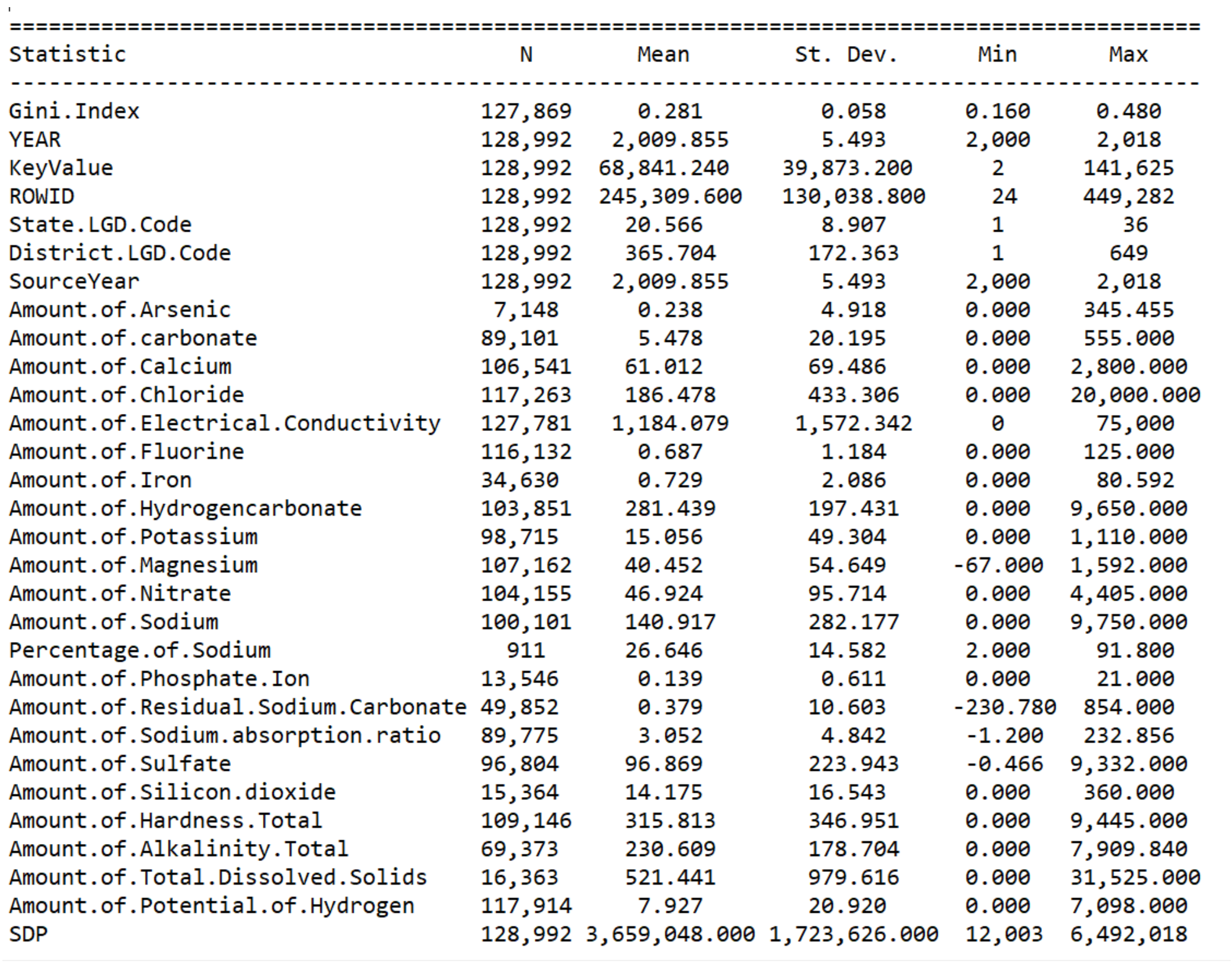
- Intercept (β_0): The intercept represents the expected mean value of the response variable (Amount of Electrical Conductivity) when all predictor variables (SDP, SDP_sq, SDP_cu, and Gini.Index) are equal to zero. In this case, the intercept is 2500, which means that when all predictor variables are equal to zero, we would expect the mean value of Amount of Electrical Conductivity to be 2500.
- SDP (β_1): The coefficient for SDP is -$2.868e^{-4}$, which means that for every one-unit increase in SDP, we would expect a decrease of $2.868e^{-4}$ units in Amount of Electrical Conductivity, holding all other predictor variables constant.
- SDP_sq (β_2): The coefficient for SDP_sq is $1.621e^{-11}$, which is a very small number in scientific notation. This means that for every one-unit increase in the square of SDP, we would expect a very small increase in Amount of Electrical Conductivity, holding all other predictor variables constant.
- SDP_cu (β_3): The coefficient for SDP_cu is $1.593e^{-18}$, which is also a very small number in scientific notation. This means that for every one-unit increase in the

cube of SDP, we would expect a very small increase in Amount of Electrical Conductivity, holding all other predictor variables constant.

- Gini.Index (β_4): The coefficient for Gini.Index is -2187, which means that for every one-unit increase in Gini.Index, we would expect a decrease of 2187 units in Amount of Electrical Conductivity, holding all other predictor variables constant.
- R-squared: The R-squared value is 0.03197, which means that the predictor variables explain 3.19% of the total variation in Amount of Electrical Conductivity.
- Residuals: The residuals represent the difference between the observed values of Amount of Electrical Conductivity and the predicted values based on the predictor variables. The minimum and maximum values of the residuals are -1894 and 73275, respectively. The residuals are used to assess the goodness of fit of the model, and to identify any outliers or influential observations.



```
====================================================
                           Dependent variable:
                    --------------------------------
                    Amount.of.Electrical.Conductivity
----------------------------------------------------
SDP                              -0.0003***
                                 (0.00003)

SDP_sq                              0.000
                                   (0.000)

SDP_cu                              0.000
                                   (0.000)

Gini.Index                      -2,186.829***
                                  (83.913)

Constant                         2,500.128***
                                  (34.272)

----------------------------------------------------
Observations                       115,175
R2                                  0.032
Adjusted R2                         0.032
Residual Std. Error       1,602.167 (df = 115170)
F Statistic             950.913*** (df = 4; 115170)
====================================================
Note:                      *p<0.1; **p<0.05; ***p<0.01
```



```
===========================================================================================
Statistic                              N       Mean         St. Dev.      Min       Max
-------------------------------------------------------------------------------------------
Gini.Index                          127,869     0.281         0.058      0.160     0.480
YEAR                                128,992   2,009.855       5.493      2,000     2,018
KeyValue                            128,992  68,841.240    39,873.200      2      141,625
ROWID                               128,992  245,309.600   130,038.800     24     449,282
State.LGD.Code                      128,992    20.566         8.907        1        36
District.LGD.Code                   128,992    365.704       172.363       1        649
SourceYear                          128,992   2,009.855       5.493      2,000     2,018
Amount.of.Arsenic                    7,148      0.238         4.918      0.000    345.455
Amount.of.carbonate                 89,101      5.478        20.195      0.000    555.000
Amount.of.Calcium                   106,541    61.012        69.486      0.000   2,800.000
Amount.of.Chloride                  117,263    186.478       433.306     0.000   20,000.000
Amount.of.Electrical.Conductivity   127,781   1,184.079     1,572.342      0       75,000
Amount.of.Fluorine                  116,132     0.687         1.184      0.000    125.000
Amount.of.Iron                      34,630      0.729         2.086      0.000     80.592
Amount.of.Hydrogencarbonate         103,851    281.439       197.431     0.000   9,650.000
Amount.of.Potassium                 98,715     15.056        49.304      0.000   1,110.000
Amount.of.Magnesium                 107,162    40.452        54.649     -67.000  1,592.000
Amount.of.Nitrate                   104,155    46.924        95.714      0.000   4,405.000
Amount.of.Sodium                    100,101    140.917       282.177     0.000   9,750.000
Percentage.of.Sodium                  911      26.646        14.582      2.000     91.800
Amount.of.Phosphate.Ion             13,546      0.139         0.611      0.000     21.000
Amount.of.Residual.Sodium.Carbonate 49,852      0.379        10.603     -230.780  854.000
Amount.of.Sodium.absorption.ratio   89,775      3.052         4.842      -1.200   232.856
Amount.of.Sulfate                   96,804     96.869        223.943     -0.466  9,332.000
Amount.of.Silicon.dioxide           15,364     14.175        16.543      0.000    360.000
Amount.of.Hardness.Total            109,146    315.813       346.951     0.000   9,445.000
Amount.of.Alkalinity.Total          69,373     230.609       178.704     0.000   7,909.840
Amount.of.Total.Dissolved.Solids    16,363     521.441       979.616     0.000   31,525.000
Amount.of.Potential.of.Hydrogen     117,914     7.927        20.920      0.000   7,098.000
SDP                                 128,992 3,659,048.000 1,723,626.000  12,003  6,492,018
```