Data Assignment 1 - Report
ECO221 (Winter 2023)
Group 8

(1) Environmental Quality Measure chosen - **Ground Water Quality.**

(2) Steps followed for transforming original dataset into a district - year level dataset are as follows -

- Removed rows in which District code is not available.
- Removed rows in which District code is 0 because No district is assigned to this District code.
- Changed state name "Tamilnadu" to "Tamil Nadu" because some of the states with name Tamil Nadu are already given.
- Changed state name "The Dadra And Nagar Haveli And Daman And Diu" to "Gujarat".
- Then, aggregated data to district-year level by taking mean environmental qualities indicated by grouping data into state, district and yearcode.
- At last, Created a unique district-year ID for each row in the format: "District.Code_YearCode".

(3) Now, we needed 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 (shared here) provided by the Reserve Bank of India accessed on the Database for the Indian Economy (DBIE) portal.

For that, we have the SDP excel file shared in question to a csv file which contains columns with column names State, Year and SDP. And then we have merged it with the district - year level dataset.

Code -

```
merged_data <- merge(df2, sdp_data, by = c("State", "YearCode"))
```

(4) Now , we needed to merge our dataset with the district-level Gini index from the following paper by Mohanty et al. (2016). PDF linked here.

For that, we have created a CSV file from the given pdf containing columns "district" and "GINI index". And then merged with the dataset generated from (3).

Code-

```
final_merged_data <- merge(merged_data, gini_data, by = c("District"),all.x = TRUE)
```

(5) Prepare detailed summary statistics for all the variables (tables; histogram; box-plot;shape of the distribution, skew). Are there any outliers?

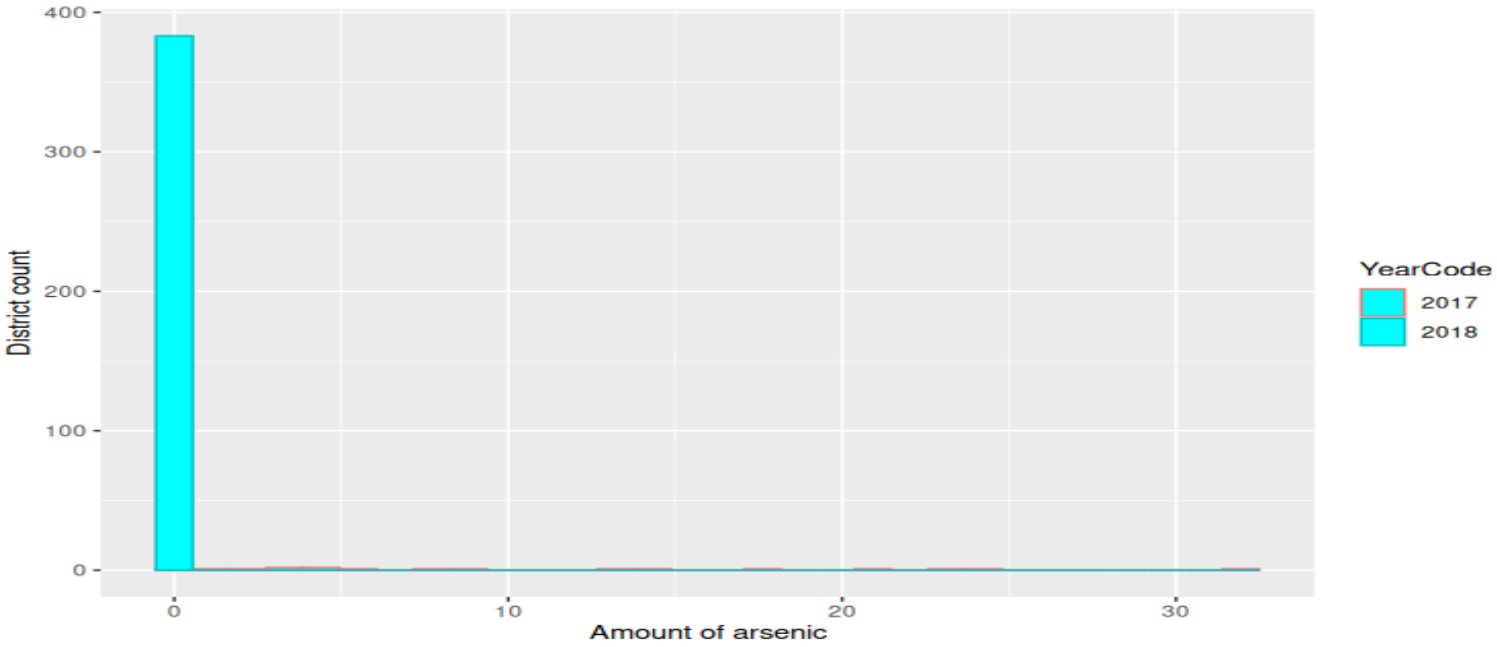
Detailed summary statistics for 22 environment variables along with SDP and gini index are below:

1) **Amount of Arsenic**

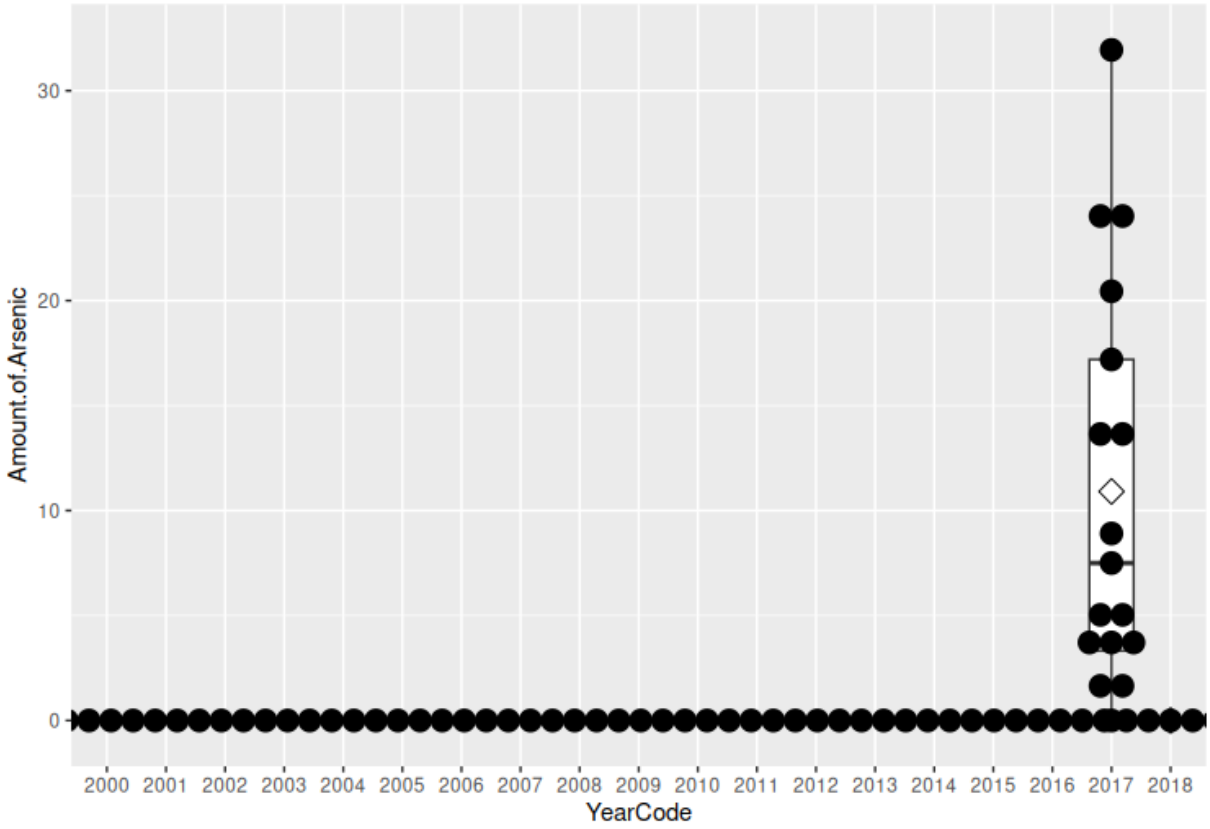
```
> summary(amountArsenic)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.000   0.000   0.464   0.000  31.944   11042
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's (Missing data)
Amount of Arsenic	0	0	0	0.46	0	31.94	11042

Histogram



Box plot



Skewness of the Graph:

```
> skewness(df3$Amount.of.Arsenic, na.rm = TRUE)
[1] 7.599911
```

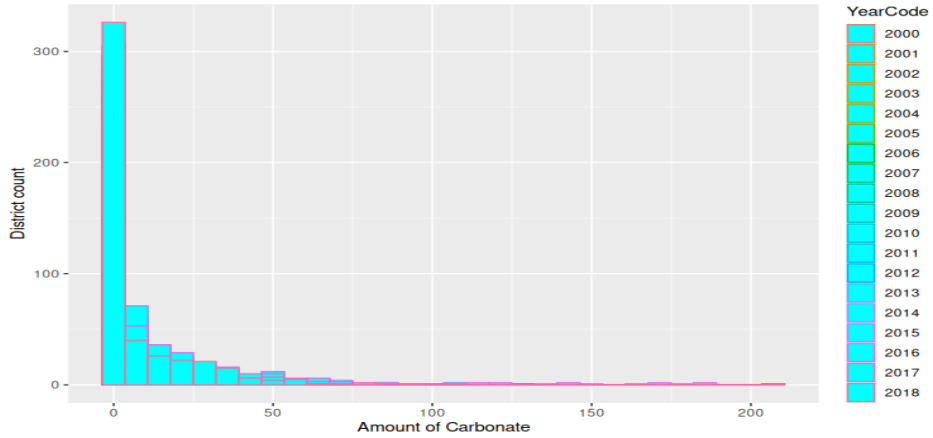
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- **An outlier:** outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile. All values > 0 are outliers.

2)Amount of carbonate

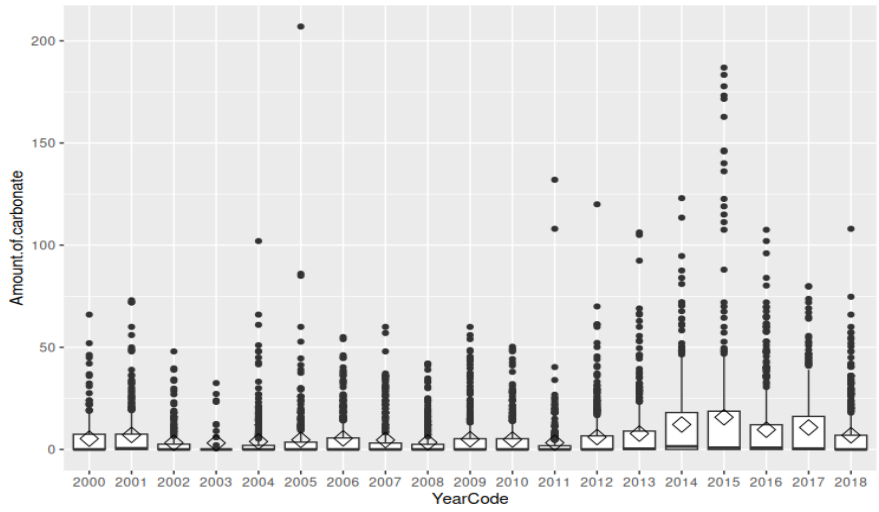
```
> summary(amountCarbonate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.000   0.000   6.783   6.410 207.000    5188
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Carbonate	0.000	0.000	0.000	6.783	6.410	207.000	5188

Histogram



Box- plot




```
> skewness(df3$Amount.of.carbonate, na.rm = TRUE)
[1] 4.508764
```

- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values greater than 9.615 are outliers.
-

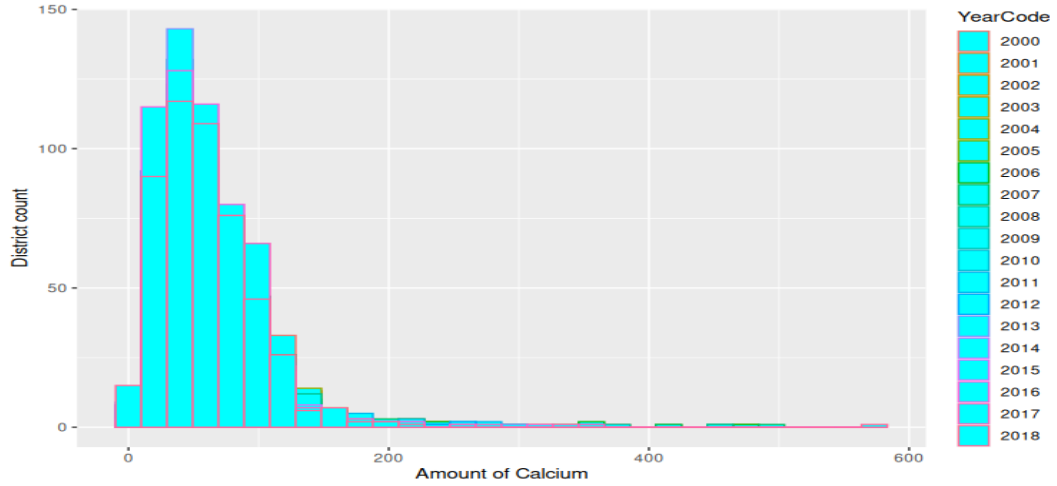
3)Amount of Calcium

```
> summary(amountCalcium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   35.23   54.93   62.16   80.60  573.33    4227
```

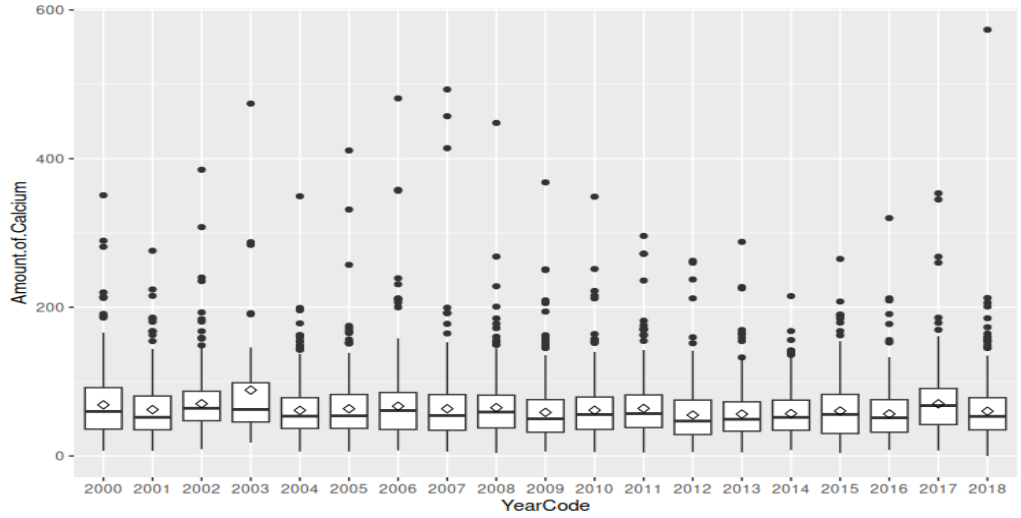
Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's

Amount of Calcium	0.00	35.23	54.93	62.16	80.60	573.33	4227

Histogram



Box Plot




```
> skewness(df$Amount.of.Calcium, na.rm = TRUE)
[1] 2.738365
```

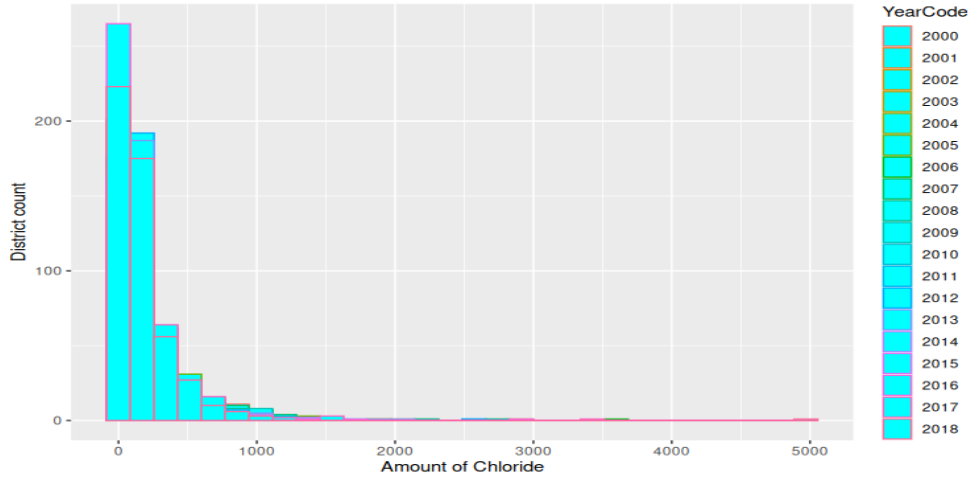
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 120.9 are outliers.

4)Amount of Chloride

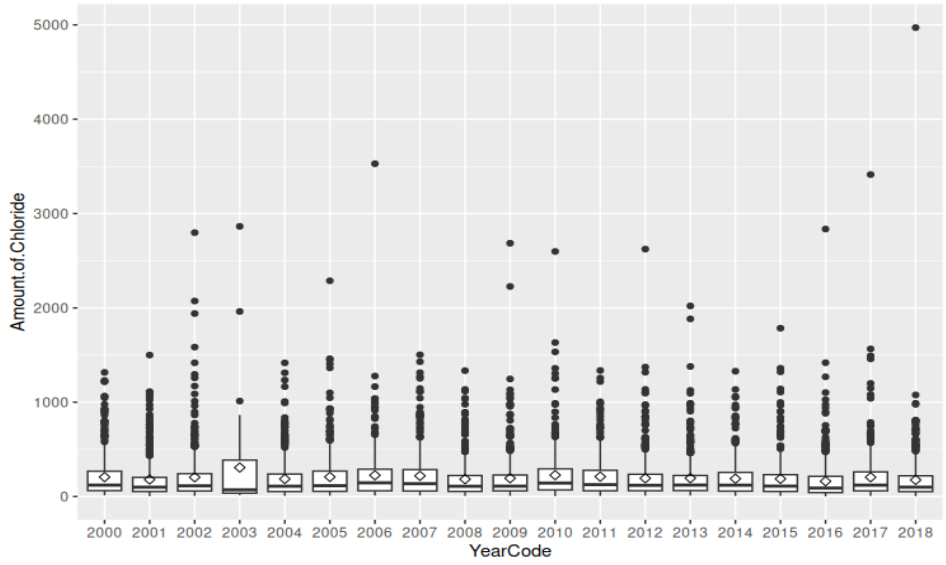
```
> summary(amountChloride)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   1.567   56.267  115.257  196.746  247.967 4971.333     3856
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Chloride	1.567	56.267	115.257	196.746	247.967	4971.333	3856

Histogram



Box Plot




```
> skewness(df$Amount.of.Chloride, na.rm = TRUE)
[1] 4.525621
```

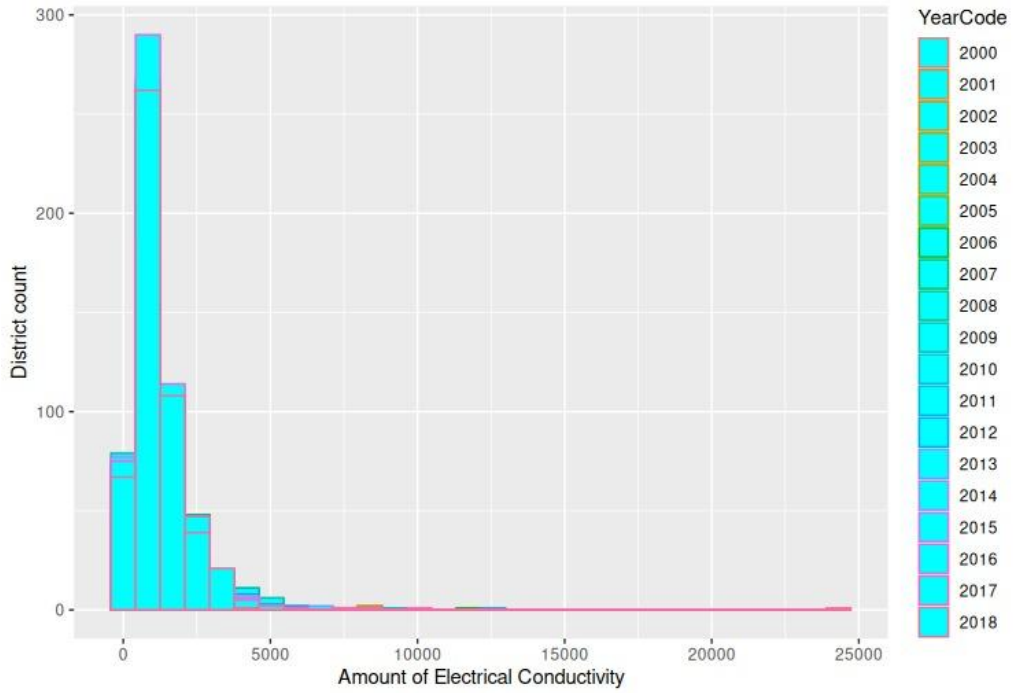
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 371.85 are outliers.

5)Amount of Electrical Conductivity

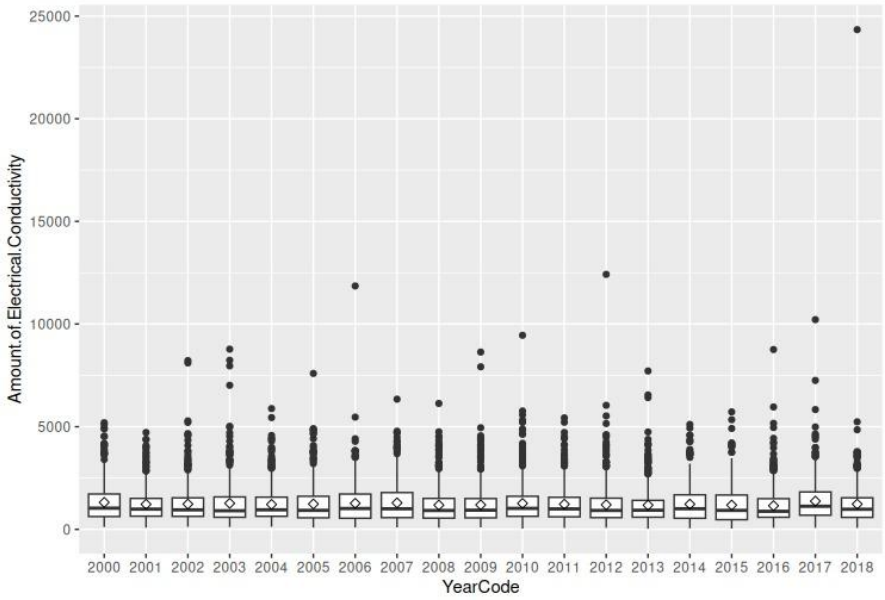
```
> summary(amountElecConductivity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   31.0   590.0   970.6  1231.7  1590.0 24340.0    2984
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Electrical Conductivity	31.0	590.0	970.6	1231.7	1590.0	24340.0	2984

Histogram



Box Plot




```
> skewness(df$Amount.of.Electrical.Conductivity, na.rm = TRUE)
[1] 3.590539
```

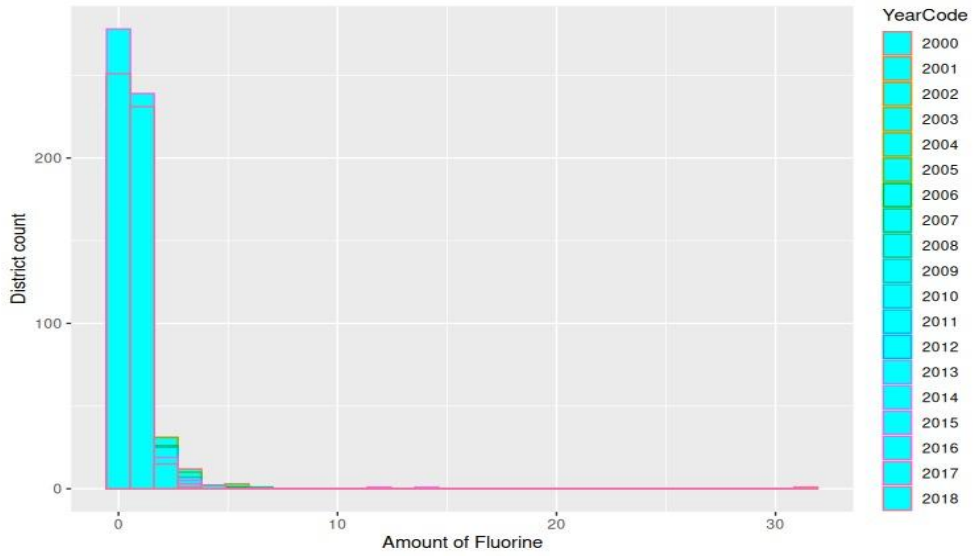
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 2385 are outliers.
-

6)Amount of fluorine

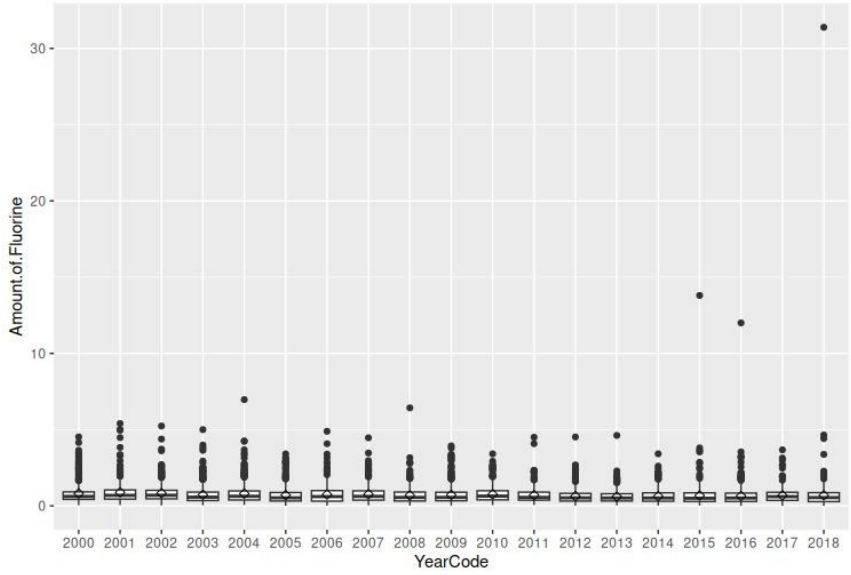
```
> summary(amountFluorine)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.347   0.571   0.718   0.903  31.385    3688
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Fluorine	0.000	0.347	0.571	0.718	0.903	31.385	3688

Histogram



Box Plot




```
> skewness(df$Amount.of.Fluorine, na.rm = TRUE)
[1] 13.67439
```

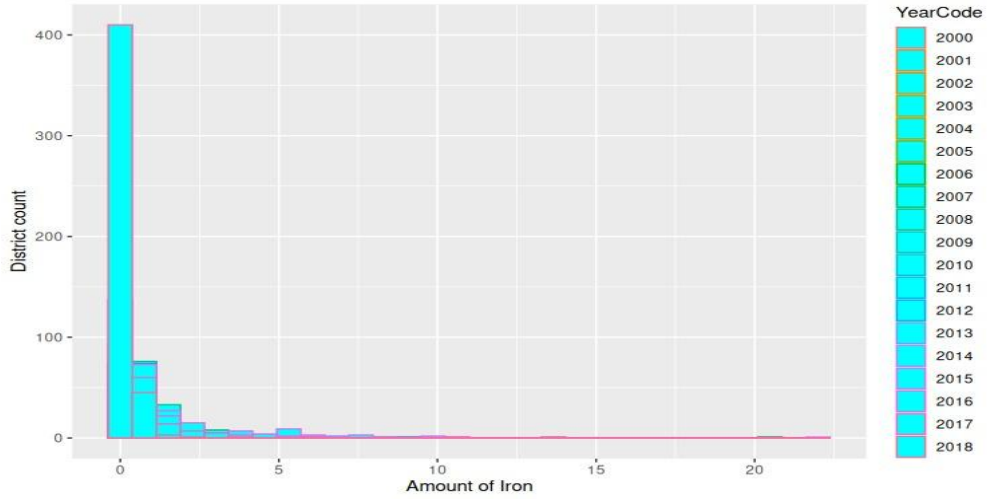
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 1.35 are outliers.

7) Amount of Iron

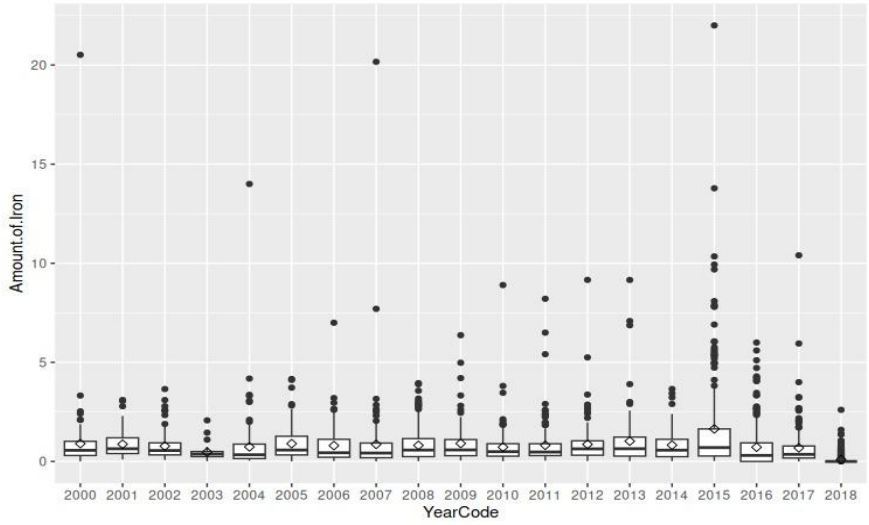
```
> summary(amountIron)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.120   0.406   0.755   0.912  22.000    8547
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Iron	0.000	0.120	0.406	0.755	0.912	22.000	8547

Histogram



Box Plot




```
> skewness(df$Amount.of.Iron, na.rm = TRUE)
[1] 6.964415
```

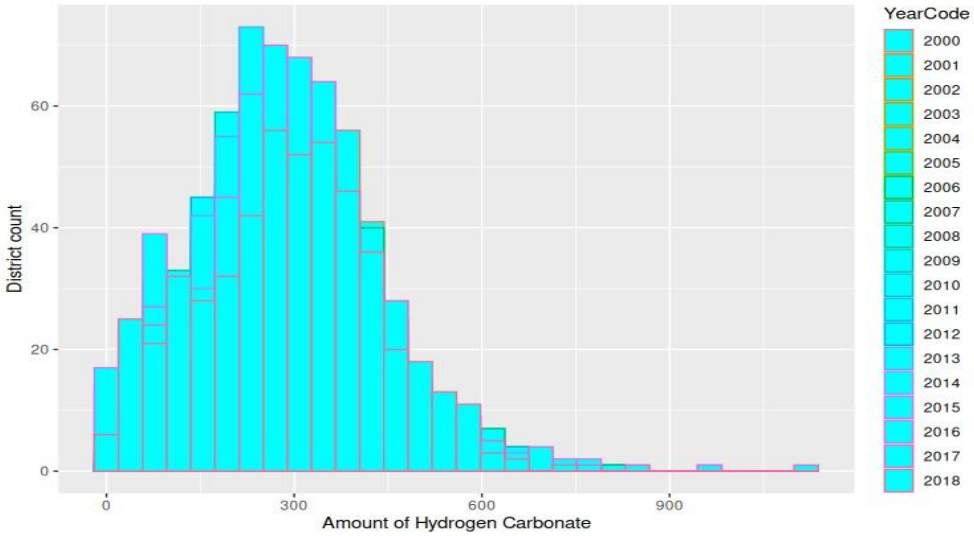
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 1.35 are outliers.

8) Amount of Hydrogen Carbonate

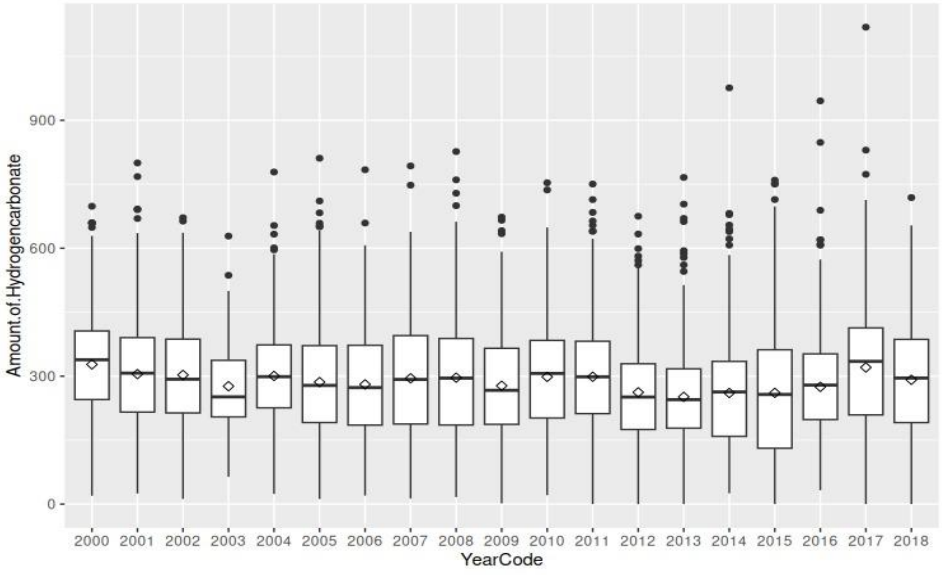
```
> summary(amountHydrogenCarbonate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0   193.4   285.2   287.6   375.5  1118.3    4216
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Hydrogen Carbonate	0.0	193.4	285.2	287.6	375.5	1118.3	4216

Histogram



Box plot



```
> skewness(df$Amount.of.Hydrogencarbonate, na.rm = TRUE)
[1] 0.3583631
```

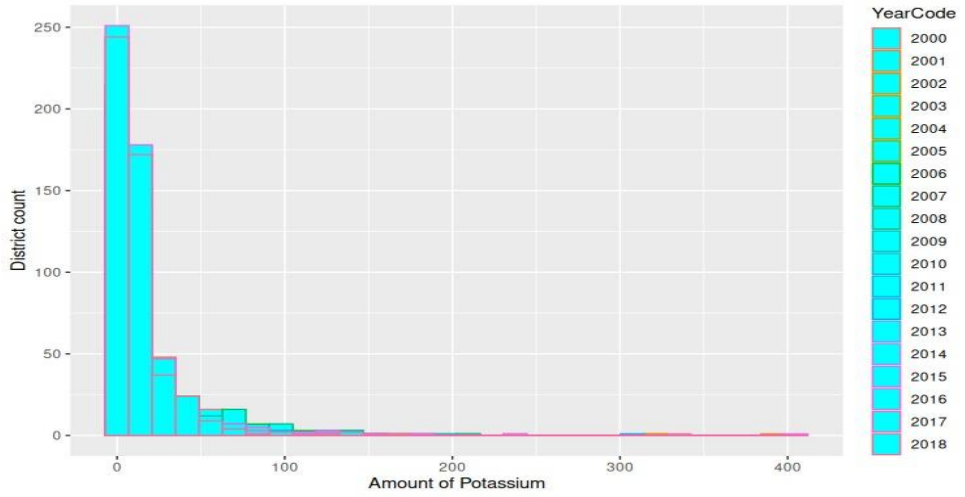
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left. But it is very close to normal distribution since skewness is < 0.5.
- All values > 563.25 are outliers.

9) Amount of Potassium

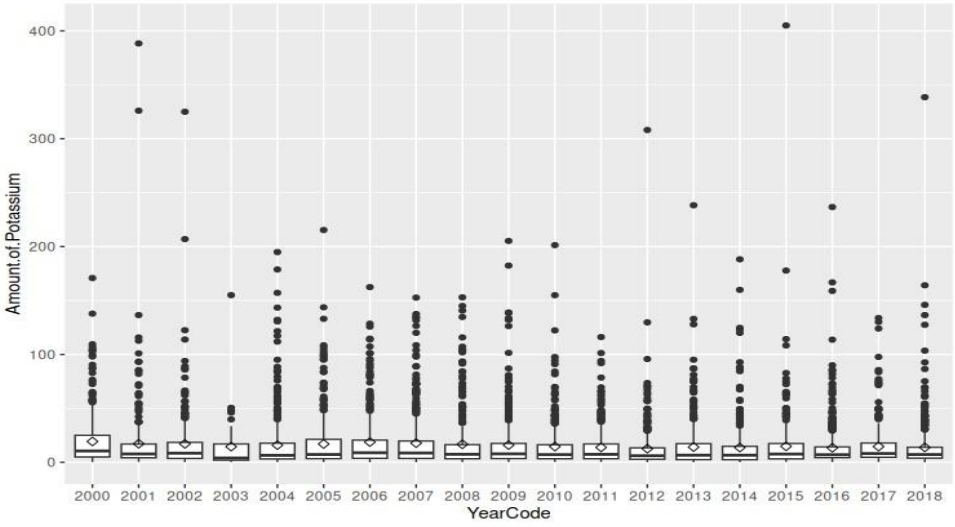
```
> summary(amountPotassium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   3.500   7.462  15.527  17.416 405.000    4574
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Potassium	0.000	3.500	7.462	15.527	17.416	405.000	4574

Histogram



Box Plot




```
> skewness(df$Amount.of.Potassium, na.rm = TRUE)
[1] 5.140091
```

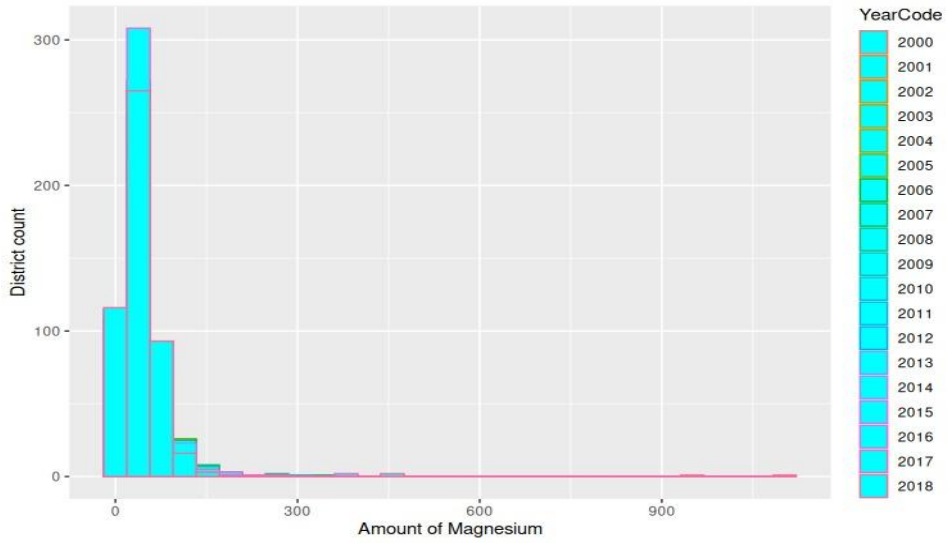
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 42.5 are outliers.

10) Amount of magnesium

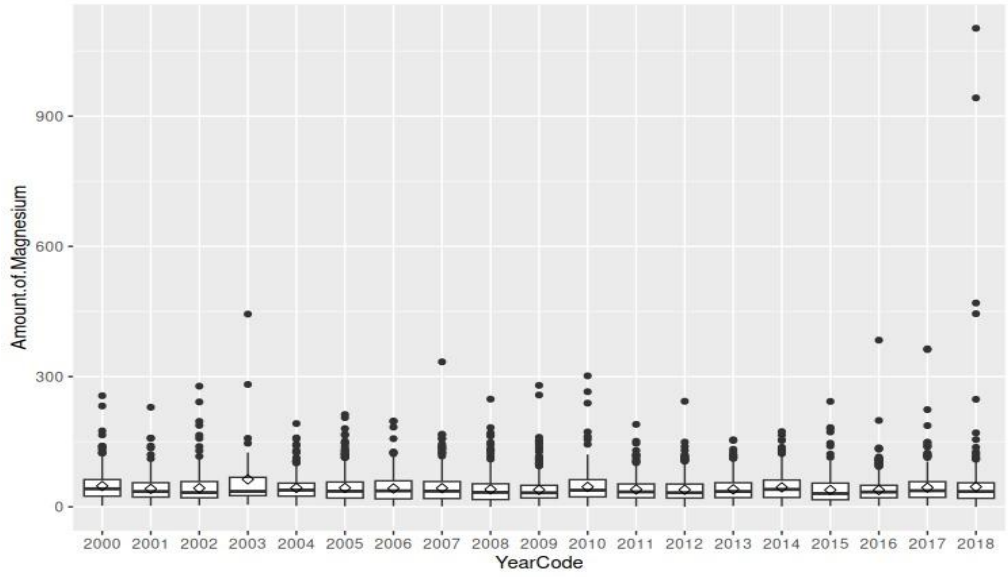
```
> summary(amountMagnesium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   20.69   35.55   42.59   56.34 1102.33    4188
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Magnesium	0.00	20.69	35.55	42.59	56.34	1102.33	4188

Histogram



Box Plot




```
> skewness(df$Amount.of.Magnesium, na.rm = TRUE)
[1] 7.435087
```

- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 84.51 are outliers.

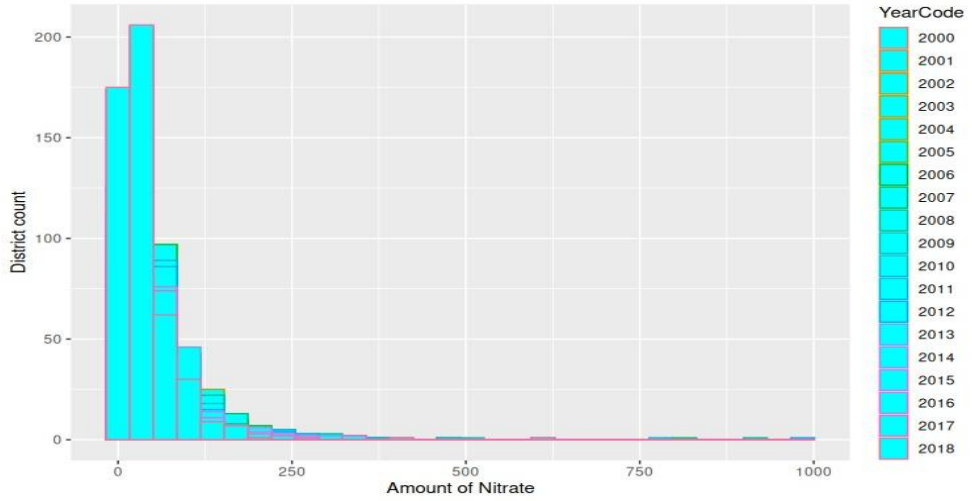
11) Amount of Nitrate

```
> summary(amountNitrate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   16.01   35.32   50.02   67.31  984.00    4580
```

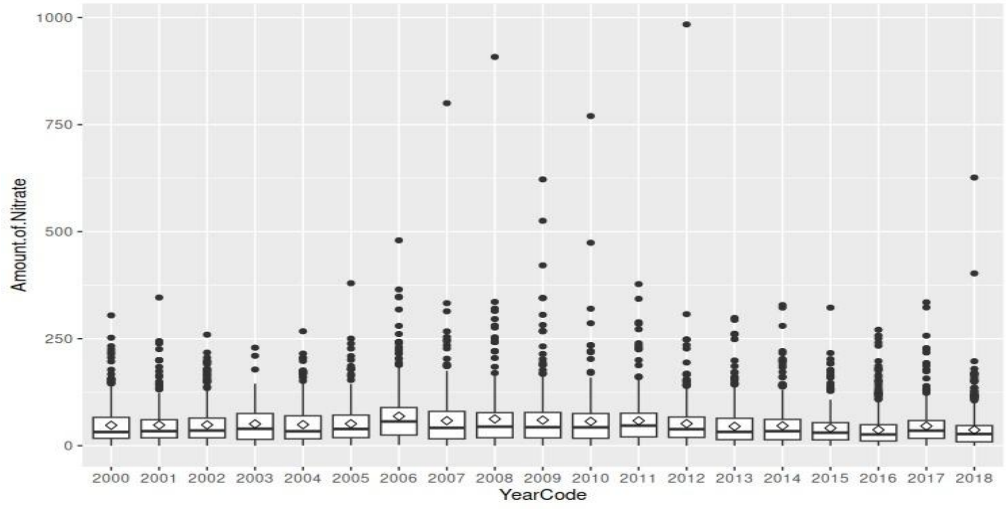
Variable	Min	1st	Median	Mean	3rd	Max	NA's

Name		Quartile			Quartile		
Amount of Nitrate	0.00	16.01	35.32	50.02	67.31	984.00	4580

Histogram



Box plot




```
> skewness(df$Amount.of.Nitrate, na.rm = TRUE)
[1] 4.258788
```

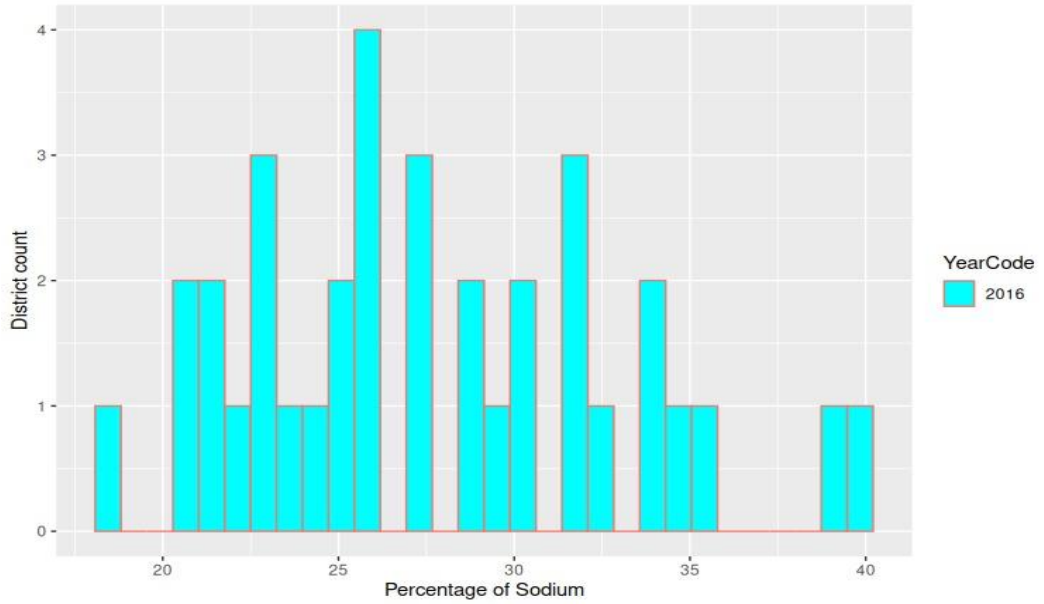
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 100.9 are outliers.

12) Amount of sodium

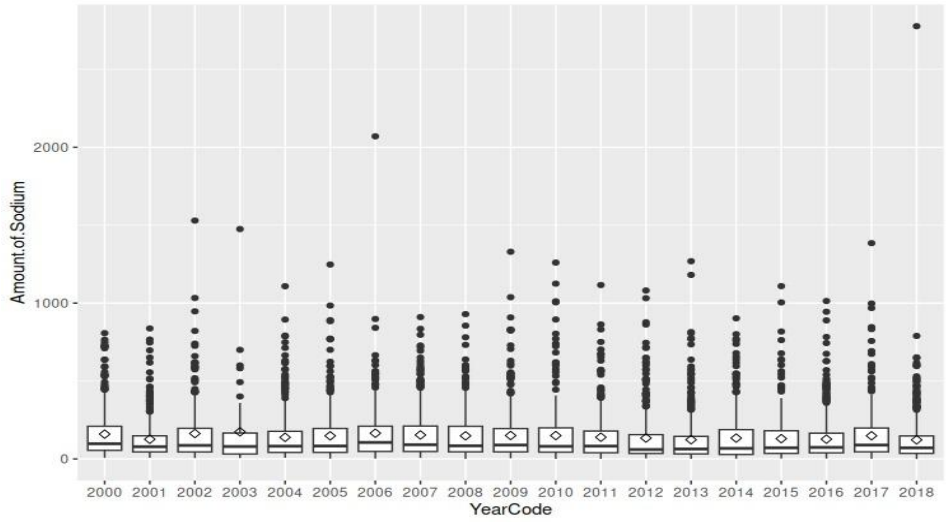
```
> summary(amountSodium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   41.14   81.17  141.73  183.67 2776.67    4573
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of sodium	0.00	41.14	81.17	141.73	183.67	2776.67	4573

Histogram



Box Plot




```
> skewness(df$Amount.of.Sodium, na.rm = TRUE)
[1] 3.14714
```

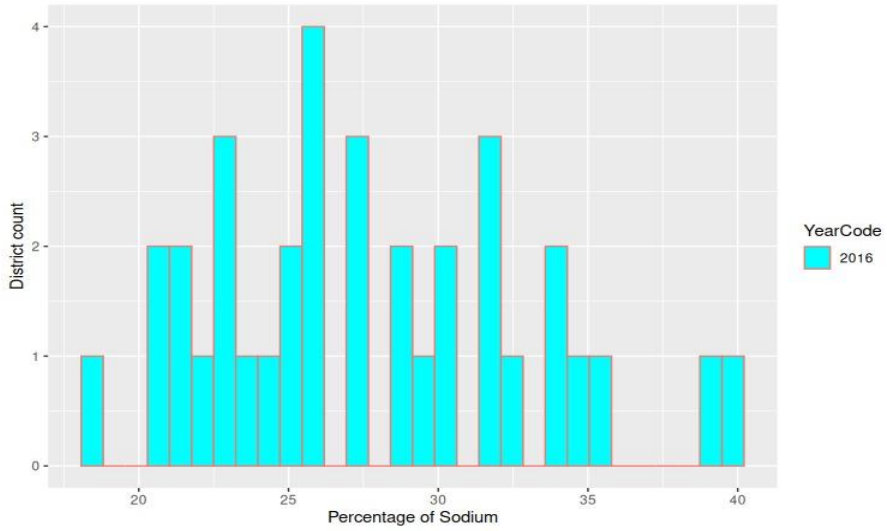
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 275.5 are outliers.

13) Percentage of sodium

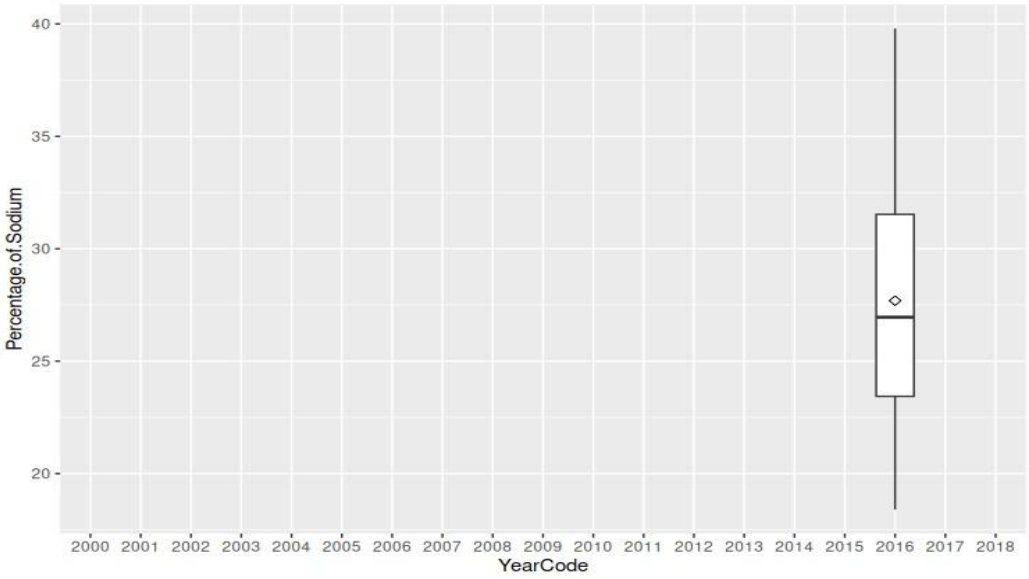
```
> summary(percentageSodium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  18.40   23.44   26.95   27.69   31.53   39.80   11407
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Percentage of sodium	18.40	23.44	26.95	27.69	31.53	39.80	11407

Histogram



Box Plot




```
> skewness(df$Percentage.of.Sodium, na.rm = TRUE)
[1] 0.452739
```

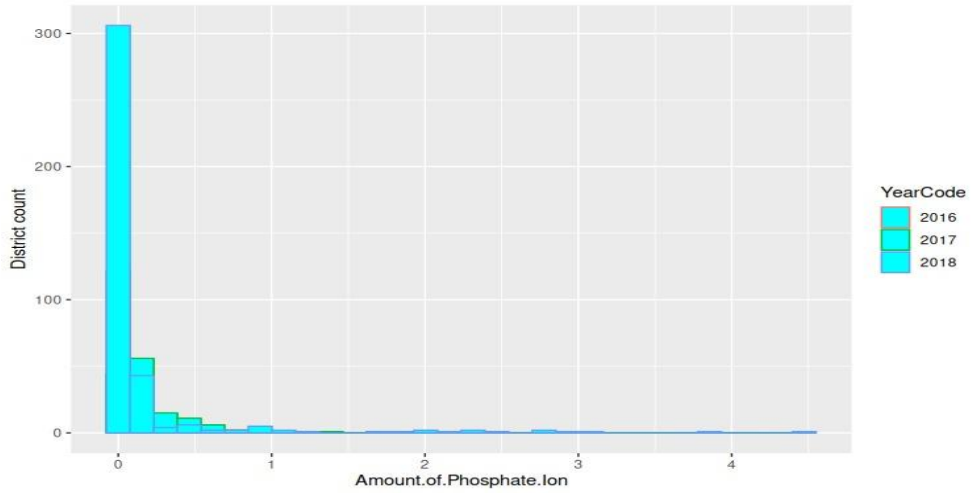
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 47.29 are outliers.

14) Amount of Phosphate ion

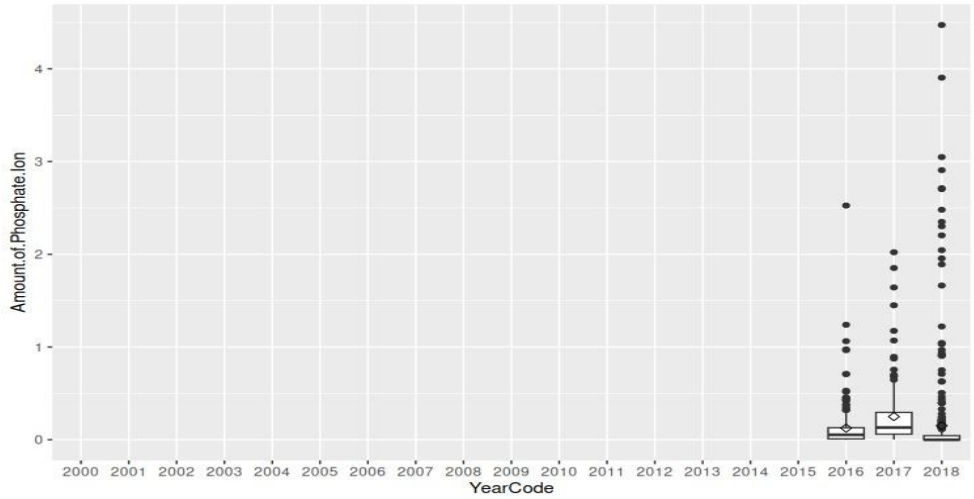
```
> summary(amountPhosphate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.000   0.025   0.164   0.130   4.473   10710
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Phosphate ion	0.000	0.000	0.025	0.164	0.130	4.473	10710

Histogram



Boxplot




```
> skewness(df$Amount.of.Phosphate.Ion, na.rm = TRUE)
[1] 5.246617
```

- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 0.195 are outliers.

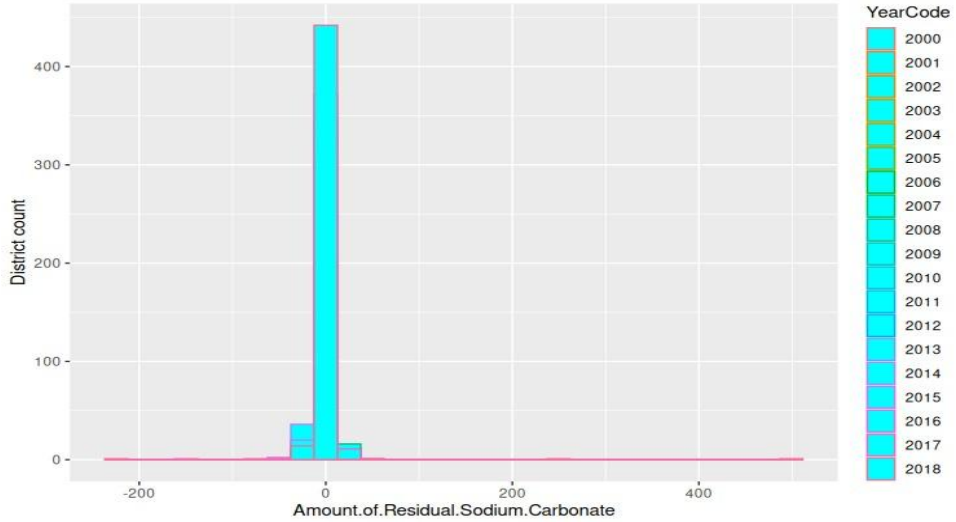
15) Amount of Residual Sodium Carbonate

```
> summary(amountSodiumCarbonate)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
-220.610    0.000    1.476    1.877    4.005  502.747     5987
```

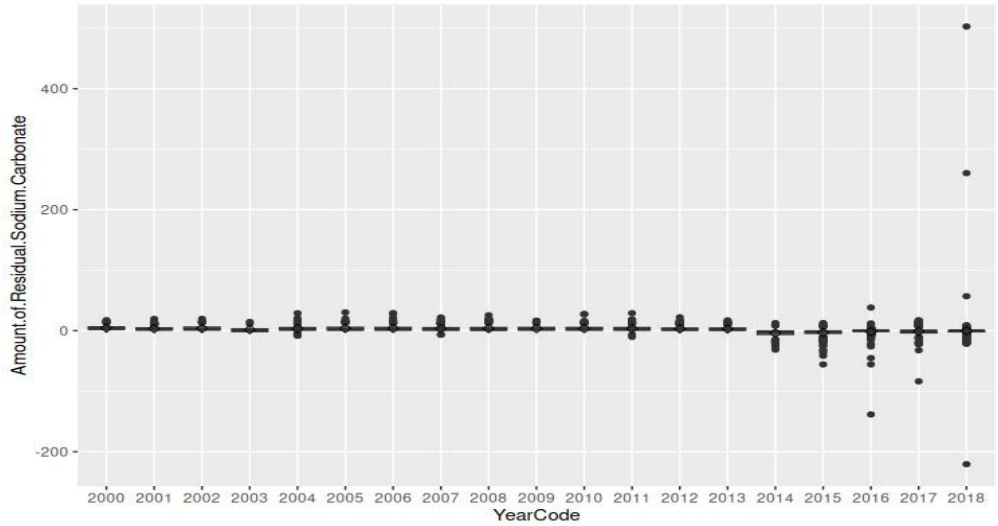
Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's

Amount of Residual Sodium Carbonate	-220.610	0.000	1.476	1.877	4.005	502.747	5987

Histogram



Box Plot




```
> skewness(df$Amount.of.Residual.Sodium.Carbonate, na.rm = TRUE)
[1] 22.98204
```

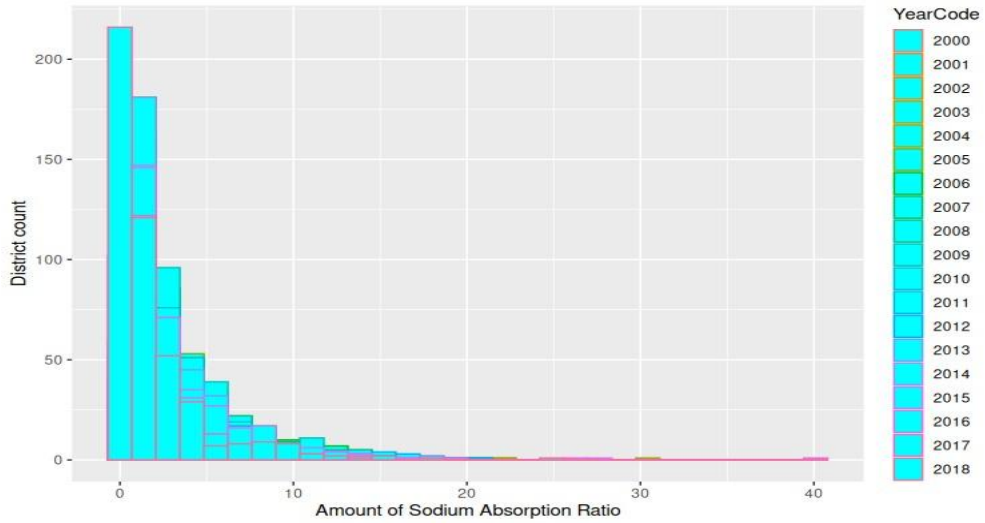
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 10 are outliers.

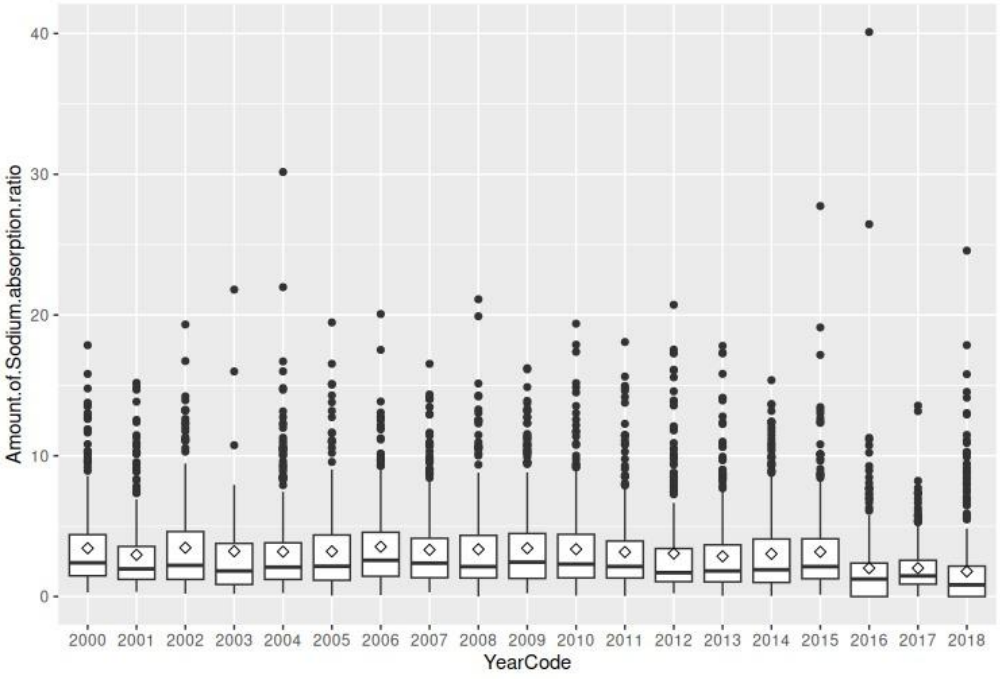
16) Amount of Sodium Absorption Ratio

```
> summary(amountSodiumAbsorptionRatio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.106   1.989   3.033   3.897  40.100    5016
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Sodium Absorption Ratio	0.000	1.106	1.989	3.033	3.897	40.100	5016

Histogram





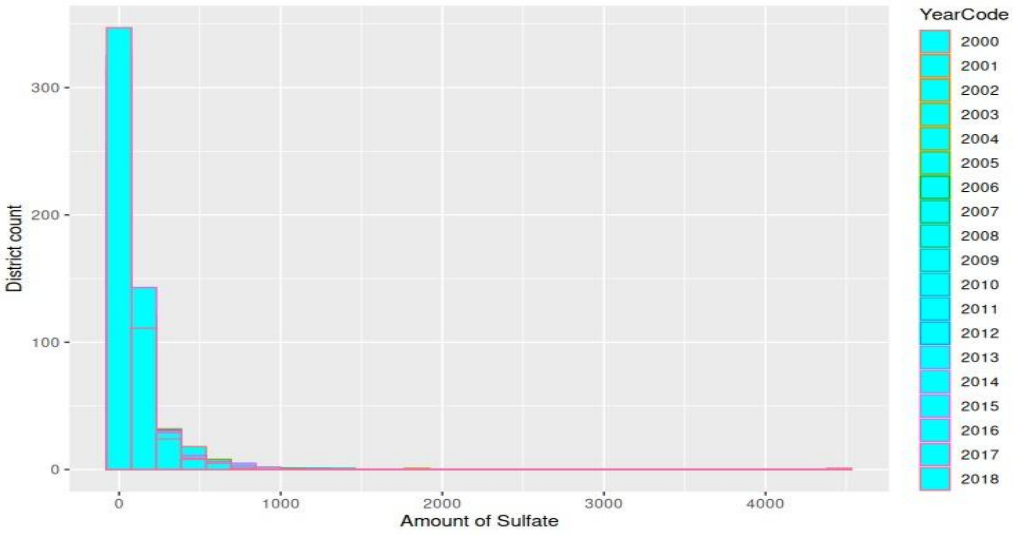

```
> skewness(df$Amount.of.Sodium.absorption.ratio, na.rm = TRUE)
[1] 2.436128
```

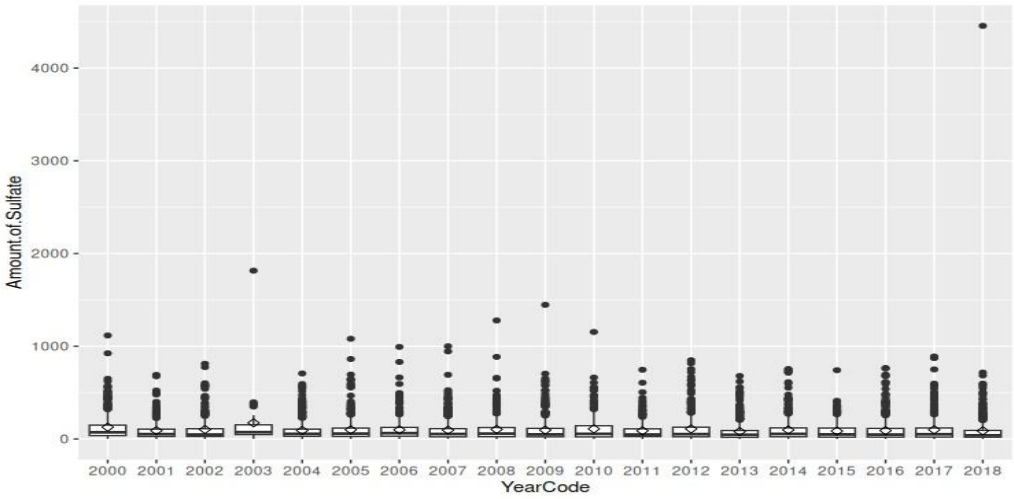
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 5.84 are outliers.

17) Amount of Sulfate

```
> summary(amountSulfate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   24.12   52.69   97.16  117.68 4455.33    4907
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Sulfate	0.00	24.12	52.69	97.16	117.68	4455.33	4907





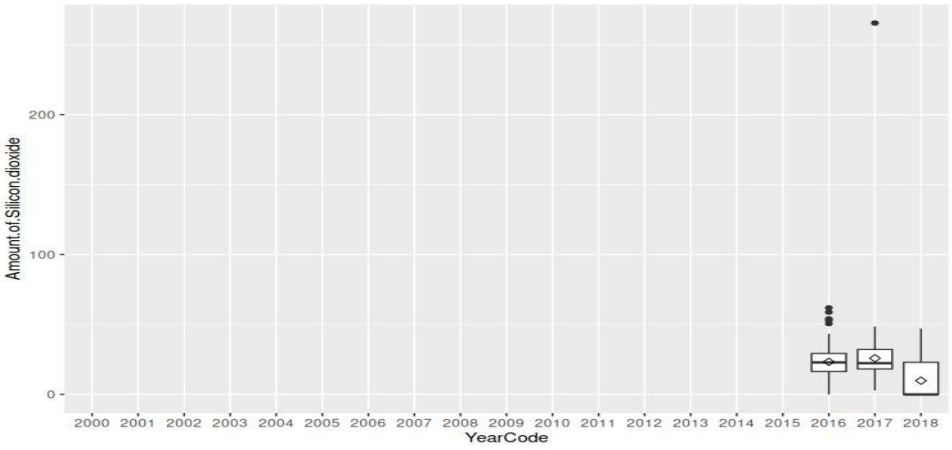

```
> skewness(df$Amount.of.Sulfate, na.rm = TRUE)
[1] 7.599526
```

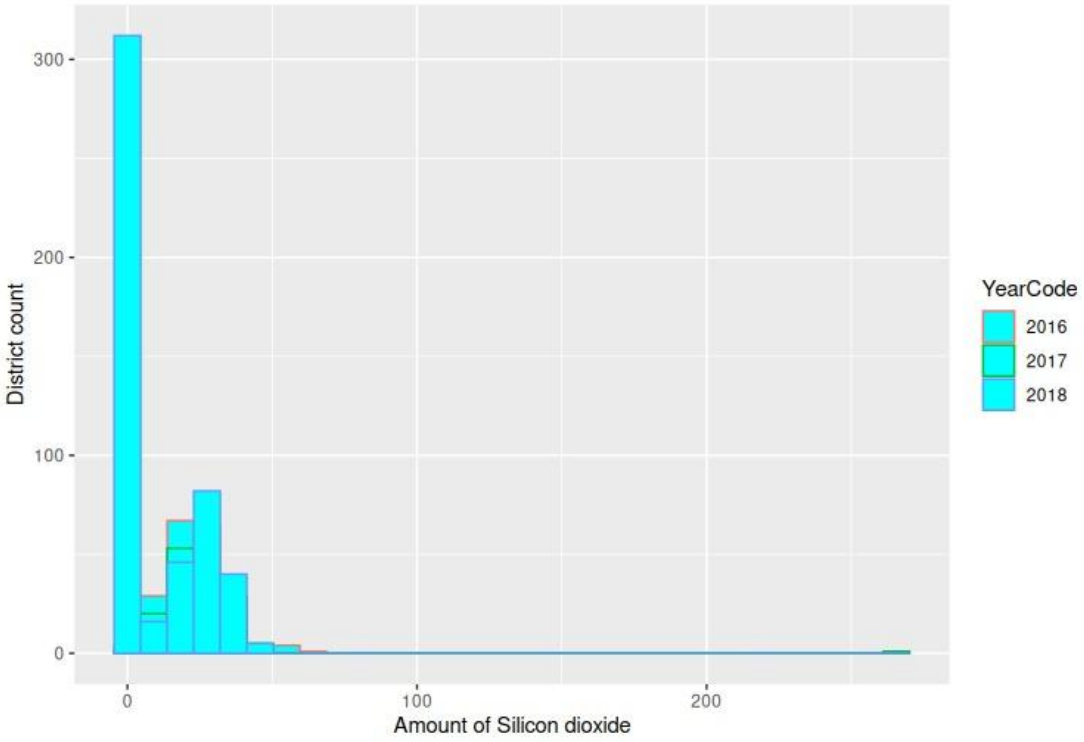
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 176.52 are outliers.

18) Amount of Silicon dioxide

```
> summary(amountSiliconDiOxide)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    0.00   16.77   15.71   26.89  265.56   10603
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Silicon dioxide	0.00	0.00	16.77	15.71	26.89	265.56	10603





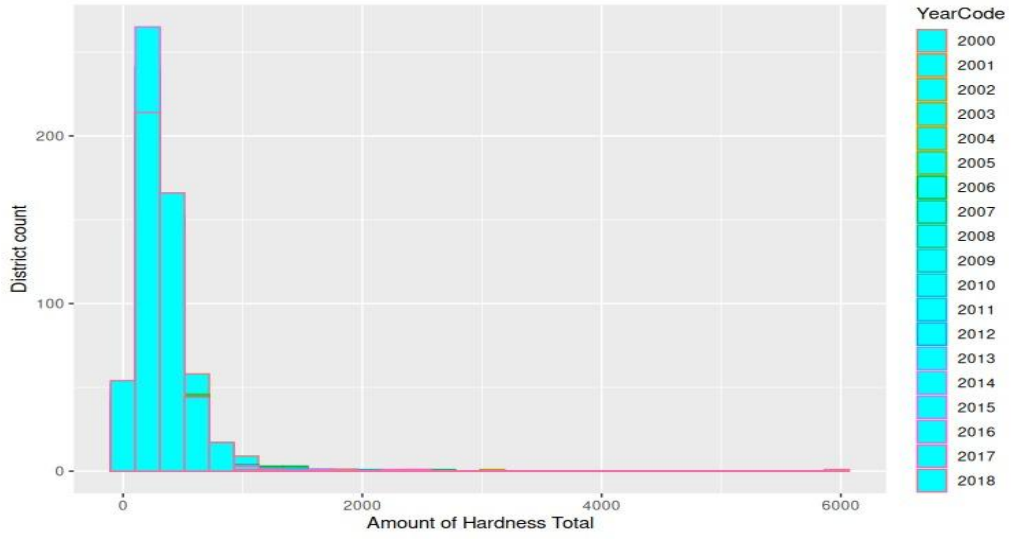

```
> skewness(df$Amount.of.Silicon.dioxide, na.rm = TRUE)
[1] 4.272341
```

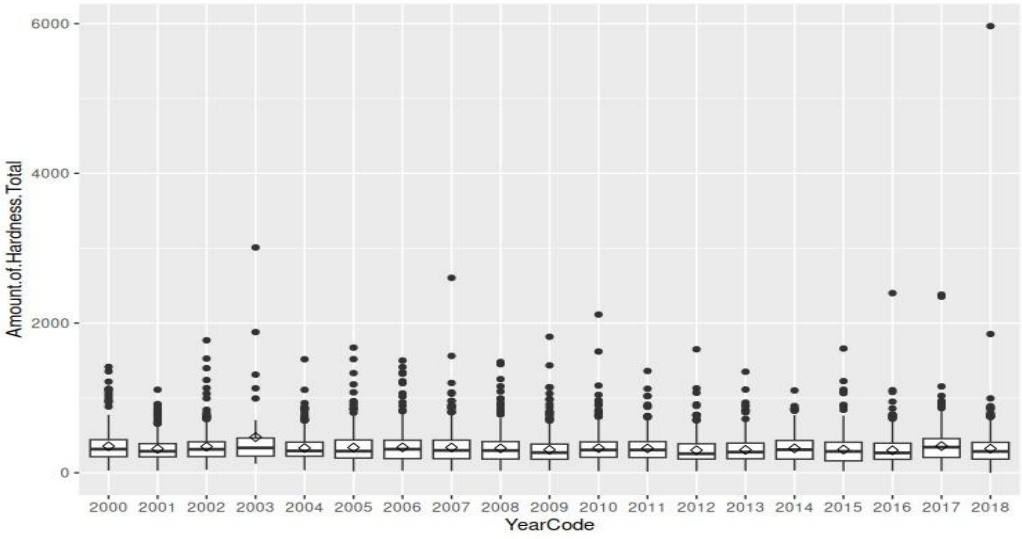
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 40.33 are outliers.

19) Amount of Hardness Total

```
> summary(totalHardness)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0   193.1   293.1   327.1   416.2  5966.7    4133
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Hardness Total	0.0	193.1	293.1	327.1	416.2	5966.7	4133





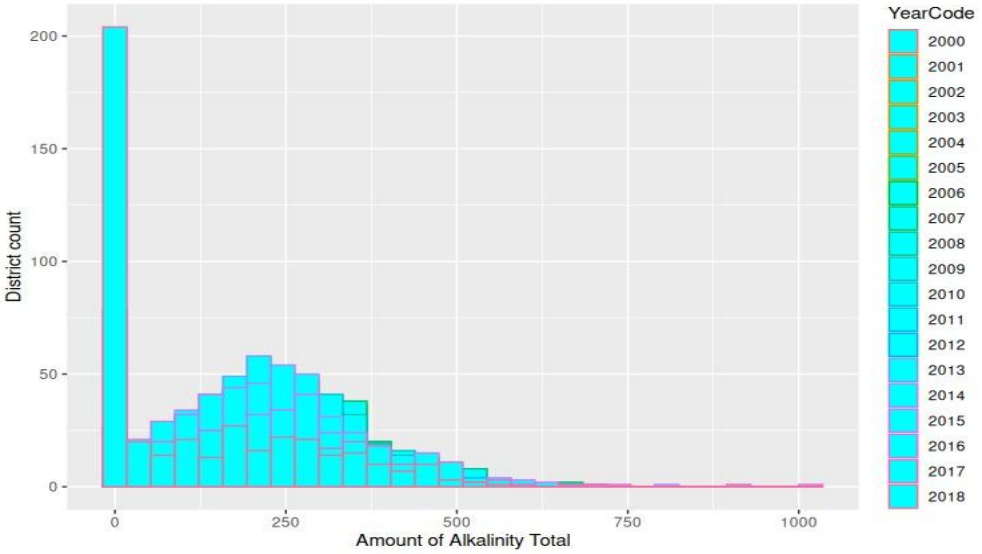

```
> skewness(df$Amount.of.Hardness.Total, na.rm = TRUE)
[1] 4.501191
```

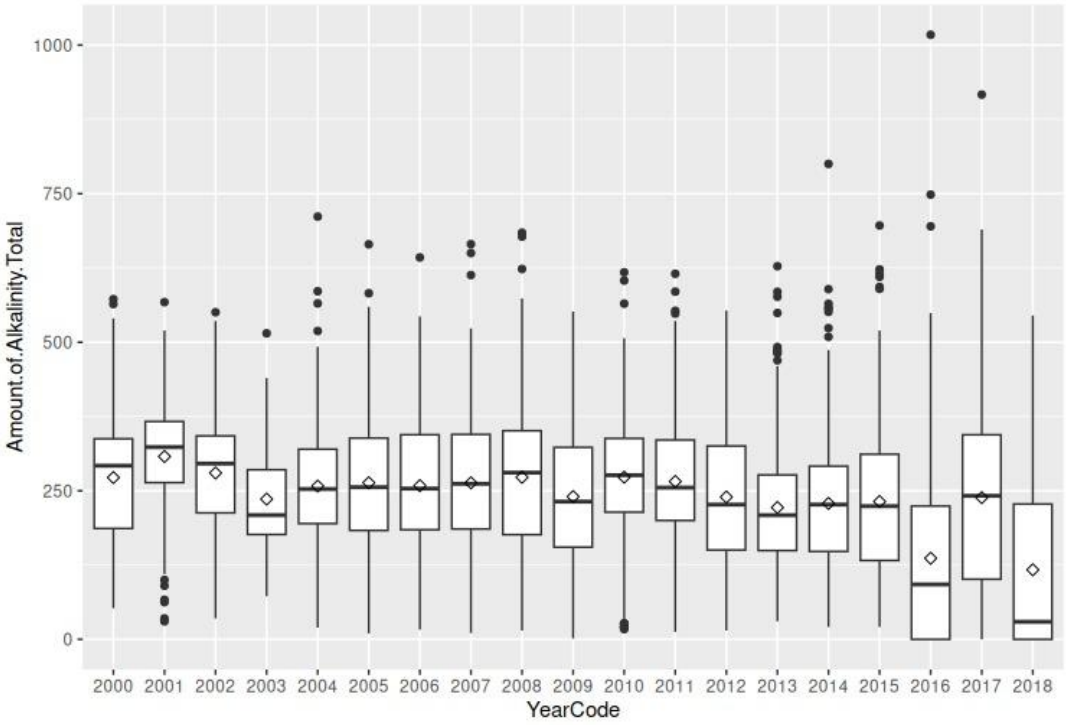
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 624 are outliers.

20) Amount of Alkalinity Total

```
> summary(totalAlkalinity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0   145.5   237.2   235.2   322.0  1017.4    6458
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Alkalinity Total	0.0	145.5	237.2	235.2	322.0	1017.4	6458





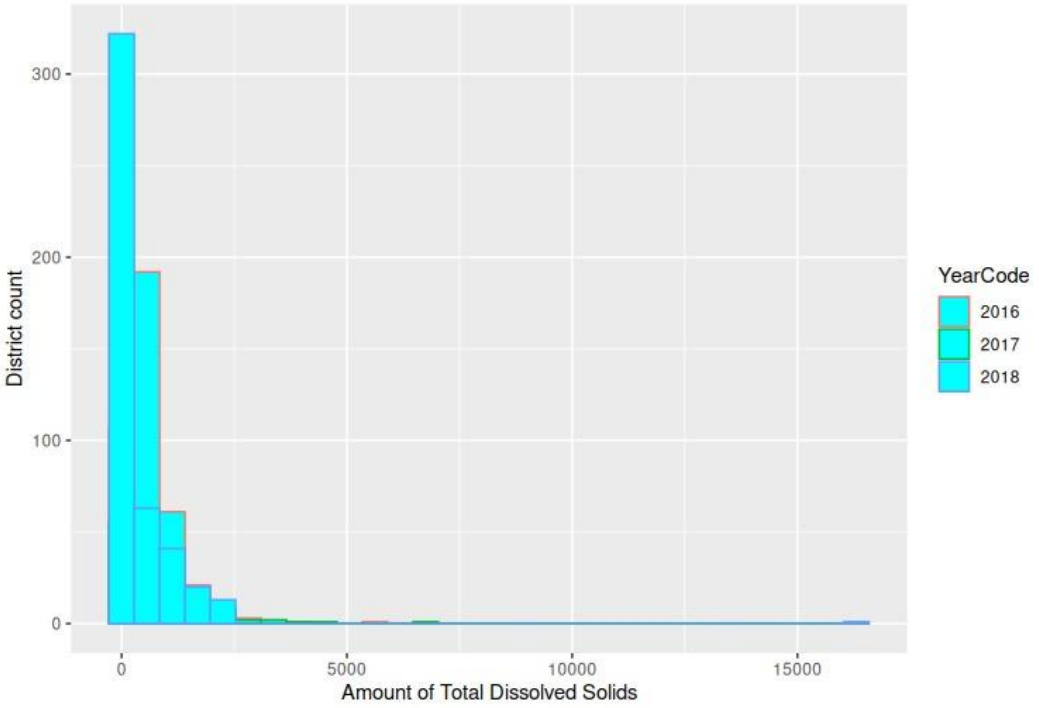

```
> skewness(df$Amount.of.Alkalinity.Total, na.rm = TRUE)
[1] 0.2290331
```

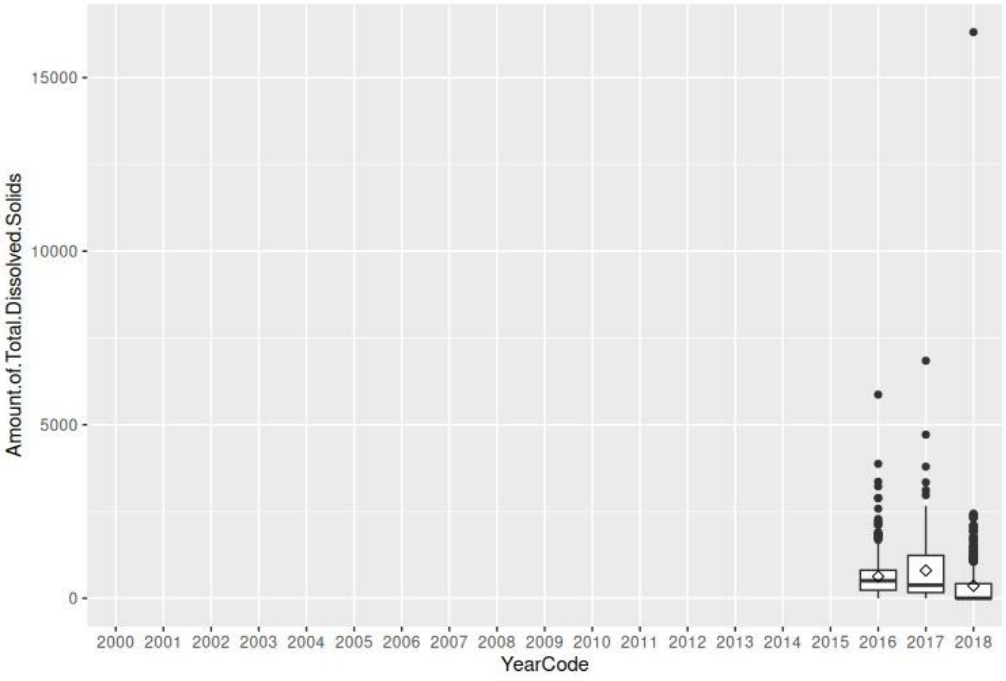
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left. But this is very close to normal distribution since skewness < 0.5.
- All values > 483 are outliers.

21) Amount of Total Dissolved Solids

```
> summary(totalDissolvedSolids)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     0.0   307.1   530.6   742.4 16307.7   10445
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Total Dissolved Solids	0.0	0.0	307.1	530.6	742.4	16307.7	10445





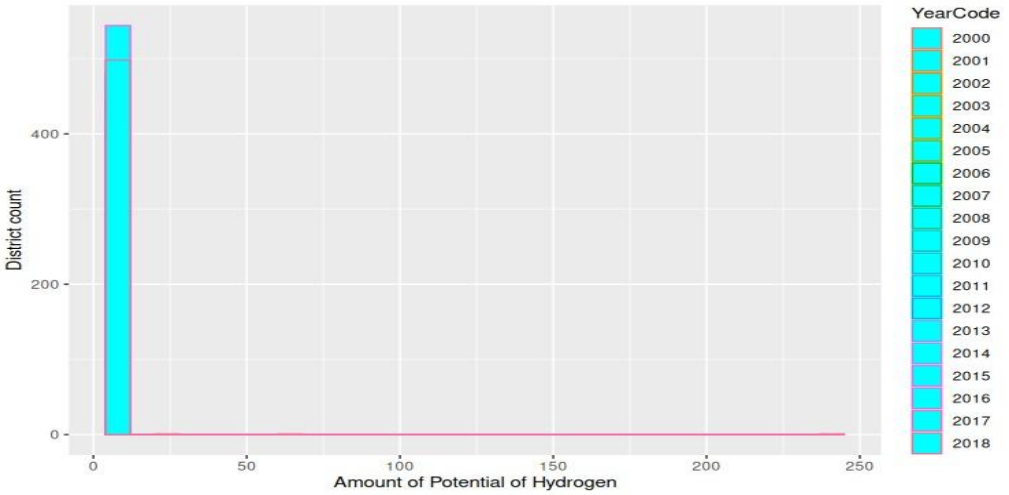

```
> skewness(df$Amount.of.Total.Dissolved.Solids, na.rm = TRUE)
[1] 7.724258
```

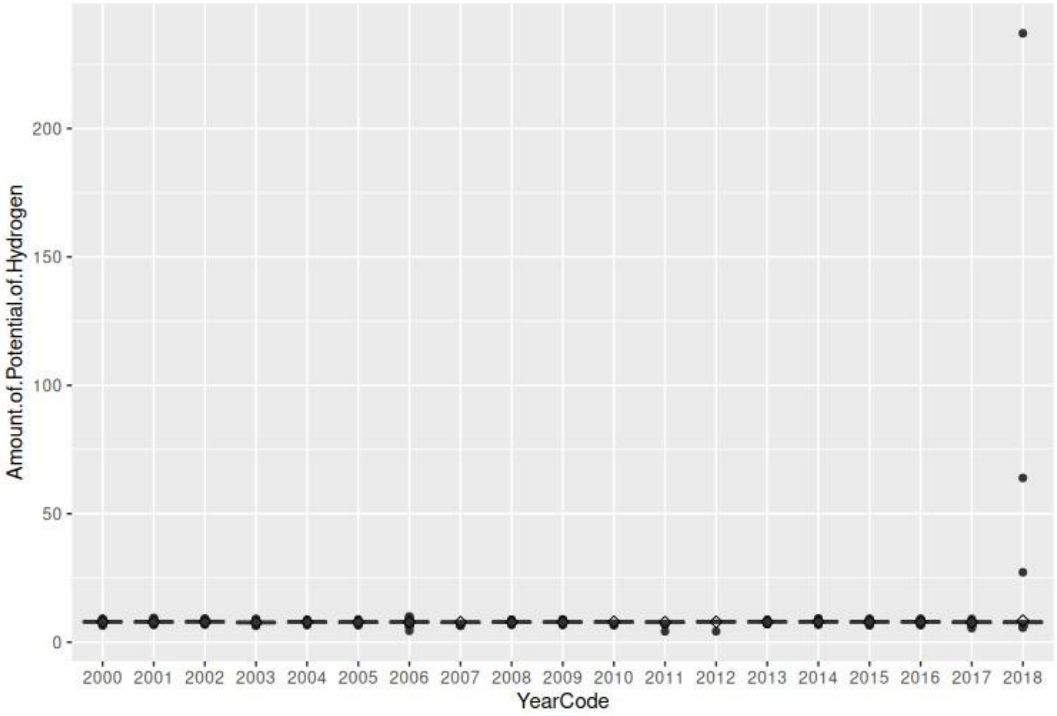
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 1113.6 are outliers.

22) Amount of Potential of Hydrogen

```
> summary(hydrogenPotential)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  4.200   7.608   7.872   7.883   8.084 237.060    3499
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Amount of Potential of Hydrogen	4.200	7.608	7.872	7.883	8.084	237.060	3499





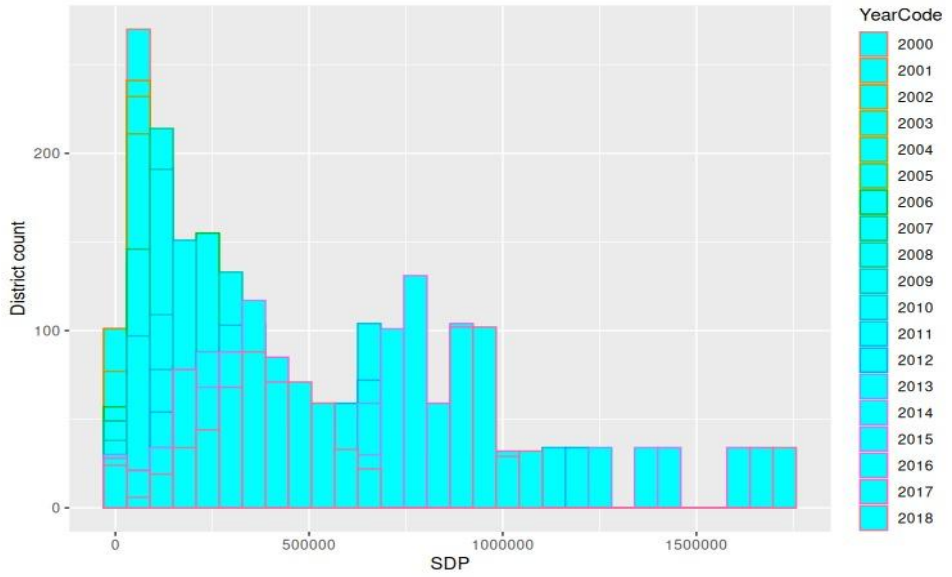

```
> skewness(df$Amount.of.Potential.of.Hydrogen, na.rm = TRUE)
[1] 79.6385
```

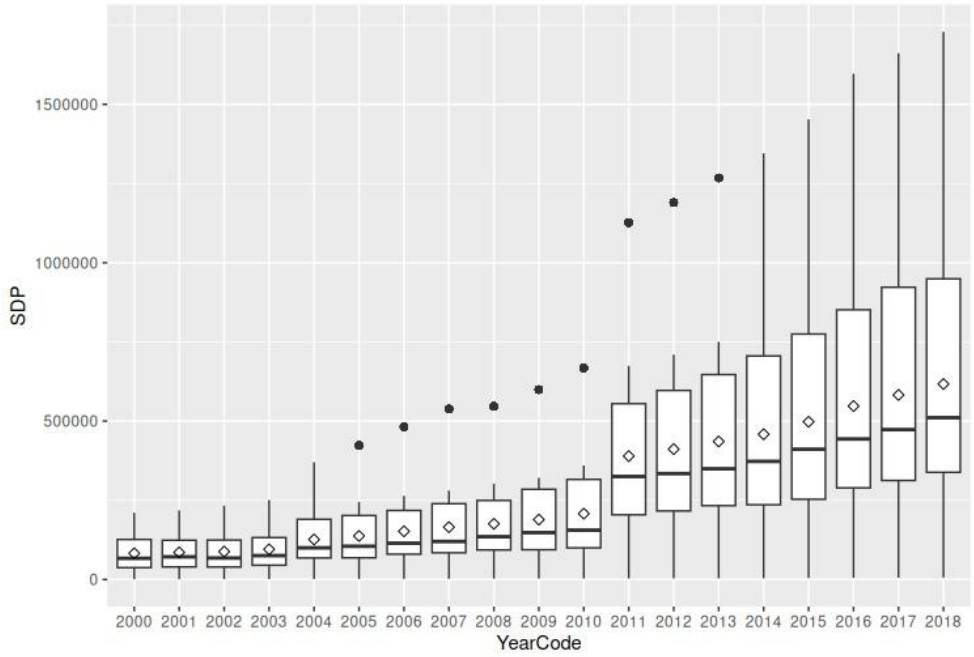
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 12.126 are outliers.

23) SDP

```
> summary(df$SDP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    838   84415  176363  287533  370023 1728578     128
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
SDP	838	84415	176363	287533	370023	1728578	128





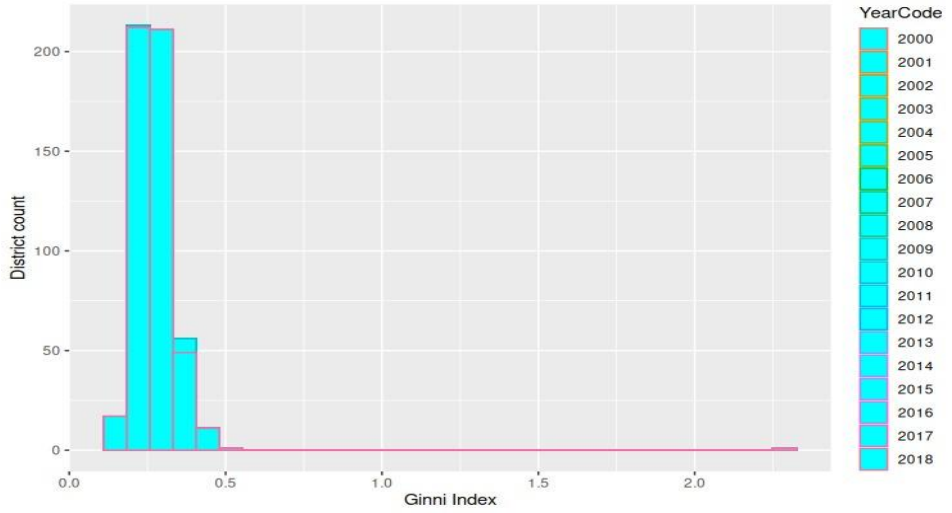

```
> skewness(df$SDP, na.rm = TRUE)
[1] 1.95294
```

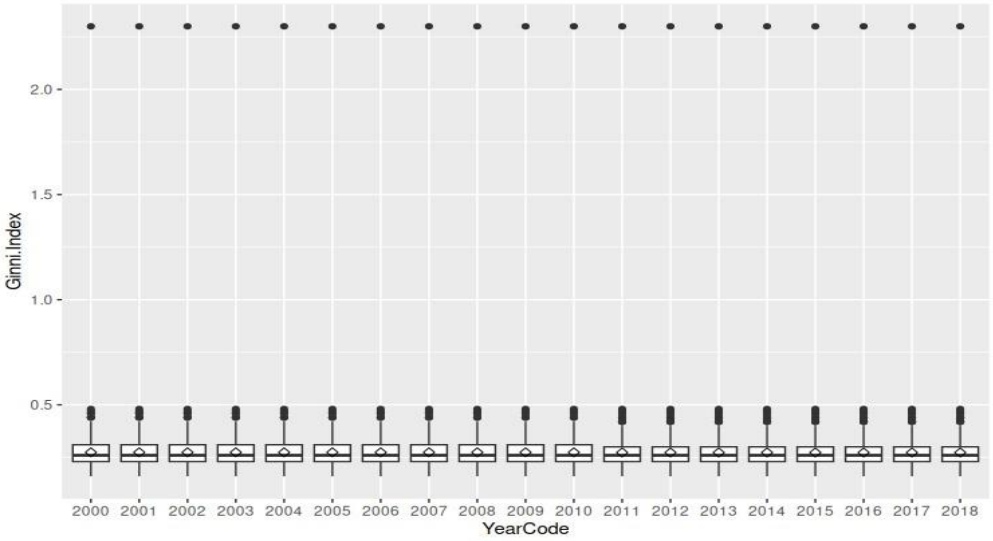
- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 555034.5 are outliers.

24) Ginni Index

```
> summary(df$Ginni.Index)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.1600  0.2300  0.2600  0.2737  0.3100  2.3000    1816
```

Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Ginni Index	0.16	0.23	0.26	0.27	0.31	2.30	1816






```
> skewness(df$Ginni.Index, na.rm = TRUE)
[1] 13.54968
```

- Positive skewness => Our graph is right skewed. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.
- All values > 0.465 are outliers.

(6) Now, using this data, estimate the following regression for any one environmental quality indicator of your choice. Summarize the results in a table and interpret in plain English. Note that i indexes districts, t indexes years, and ui,t is random error. Environmental Quality Indicator (EQI)i,t = β0 + β1SDPi,t + ui,t

Ans - Environmental Quality Chosen - **Amount of Hydrogen Carbonate**

Results after running the Linear Regression Model:

```
Call:
lm(formula = Amount.of.Hydrogencarbonate ~ SDP, data = df4)

Residuals:
    Min      1Q  Median      3Q     Max
-323.63  -94.22   -4.23   86.92  817.31

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.708e+02  2.237e+00  121.07   <2e-16 ***
SDP         5.419e-05  4.976e-06   10.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 136.2 on 7177 degrees of freedom
Multiple R-squared:  0.01626,   Adjusted R-squared:  0.01612
F-statistic: 118.6 on 1 and 7177 DF,  p-value: < 2.2e-16
```

Summarizing the results

Dependent Variable : EQI (i.e. Amount of Hydrogen Carbonate) , N = 7179, R2 = 0.0161	
Explanatory Variables	**Coefficient**
SDP	0.00005 ***
Intercept	270.8 ***

*** p-value is less than 0.001

Interpretation:

$\widehat{\beta 0}$ = 270.8

i.e the average value of the EQI when the SDP is 0.

$\widehat{\beta 1}$ = 0.00005

i.e the average increase in Amount of Hydrogen Carbonate with an unit increase in the SDP.

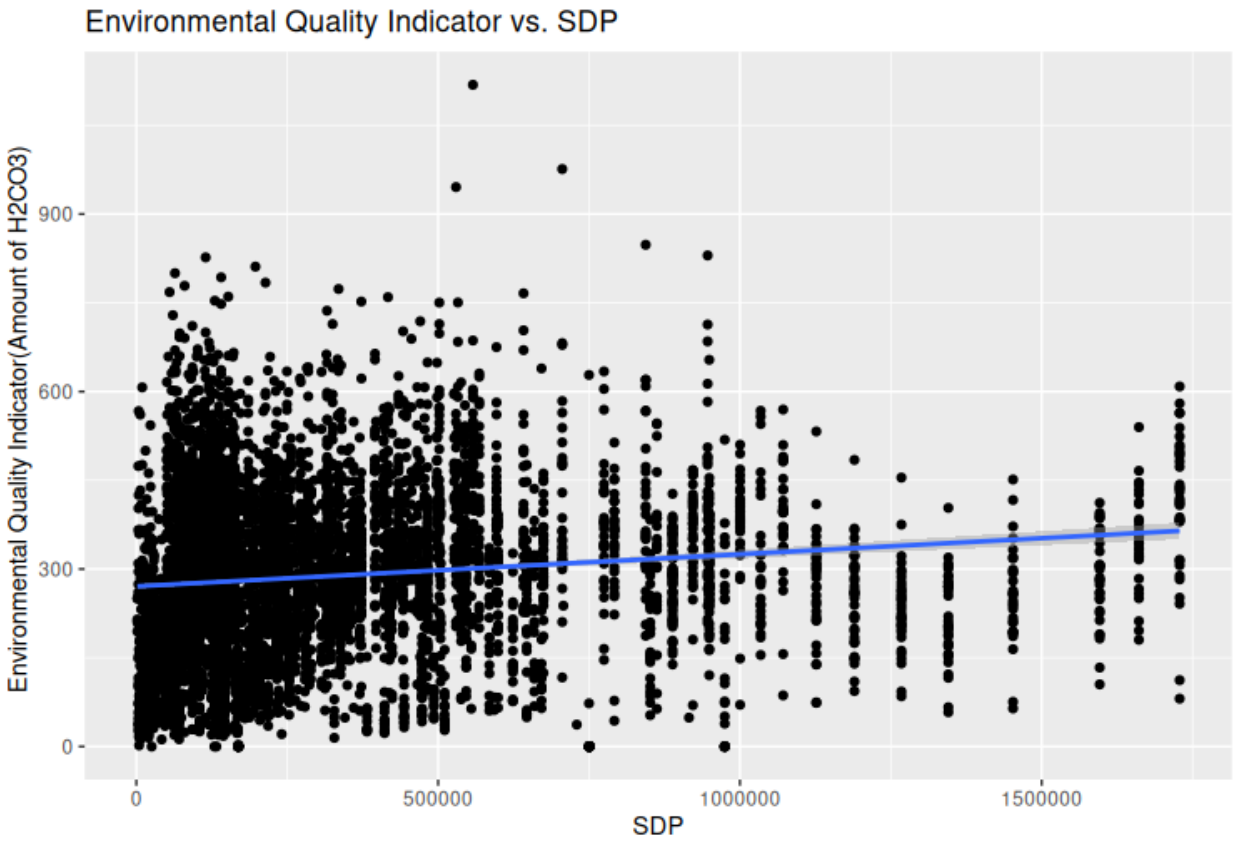
The estimated regression is EQI = 270.8 + 0.00005*SDP

It means that for every one unit increase in the SDP, the EQI is expected to increase by 0.00005 units, holding all other factors constant.

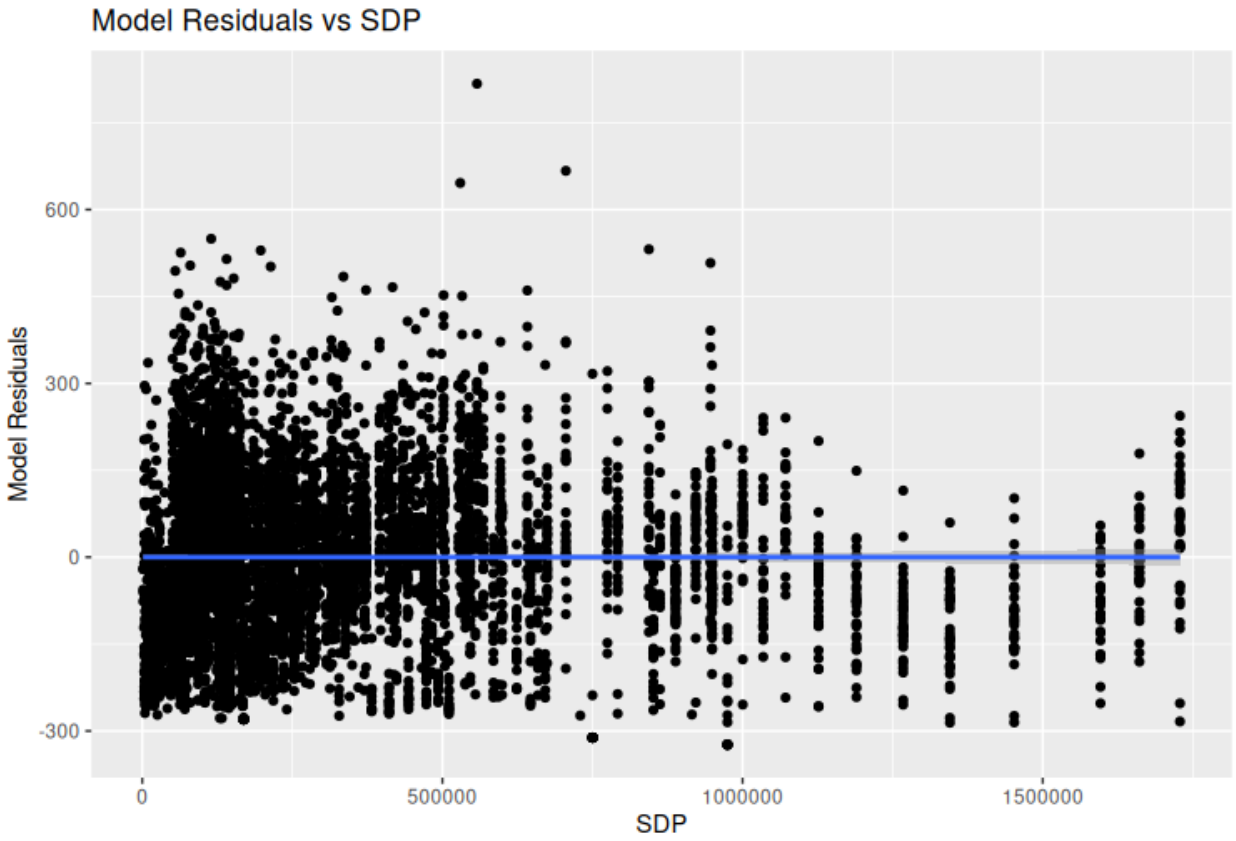
(7) Visualize the model residuals (i.e.,$\hat{u}_{i,t}$) on a plot having the environmental quality

indicator on Y-axis and SDP on the X-axis. Now, construct a second plot having ûi,t on the Y -axis and SDP on the X -axis. Finally, construct a third plot having predicted values of the environmental quality indicator on Y-axis and true values of the environmental quality indicator on X -axis. How are these three plots related, if at all? Explain.

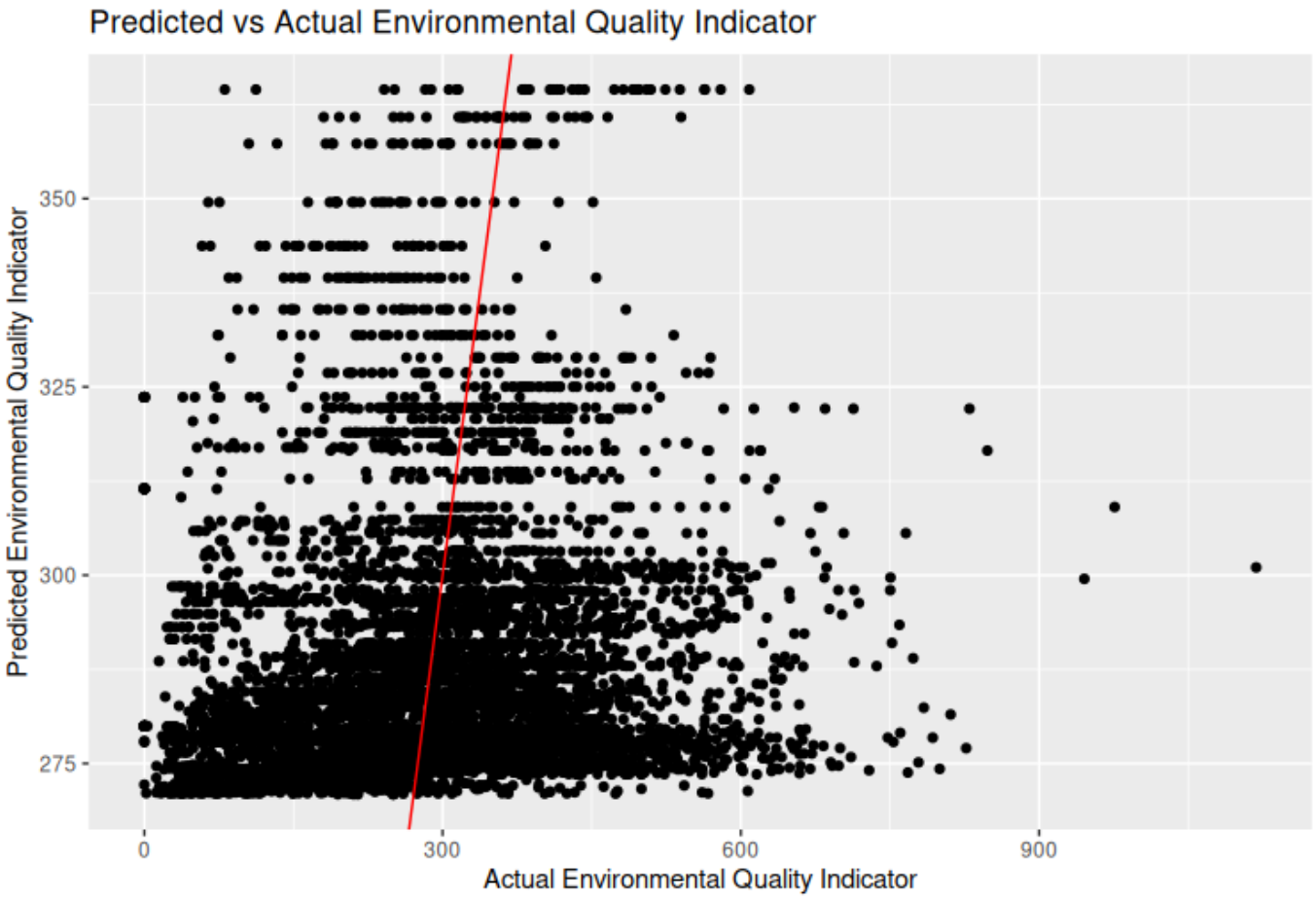
Ans -



Plot 1 : This plot shows the relationship between the environmental quality indicator (on Y-axis) and the net state domestic product SDP (on the X-axis). The regression line indicates the best linear fit between the two variables. The residuals are the vertical distance between the actual EQI values and the predicted EQI values on the line. The plot tells there is a positive relationship between SDP and EQI. Also, there is a possibility that SDP and EQI might be non-linearly related.



Plot 2: This plot shows the relationship between ûi,t (i.e. residuals on y axis) and SDP (on x-axis). The horizontal line at y=0 represents the mean of the residuals. The plot tells us that our given model is **Heteroskedastic .**



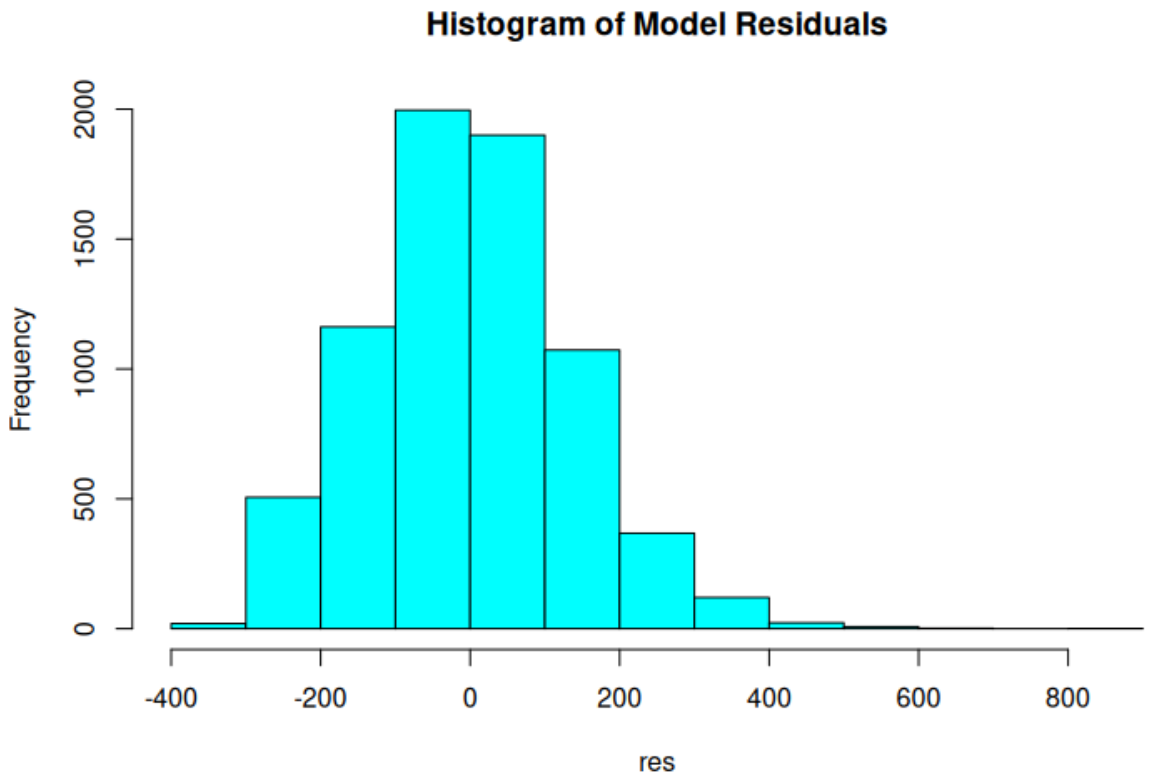
Plot 3: This plot shows predicted values of the environmental quality indicator on Y-axis and true values of the environmental quality indicator on X-axis. The closer the points are to the red line, the better the model fits the data. This scatter plot helps us visualize the accuracy of the regression model in predicting the values of our EQI. The values on the line are the points where our predicted value is **same** as the actual value.

All three plots above are related to each other and provide complementary information about our regression model.

The first plot, which is a scatter plot of EQI against SDP, helps us identify the relationship between the dependent variable (EQI) and the independent variable (SDP) and determine whether a linear or non-linear relationship is present. The second plot, which is a scatter plot of the predicted values (ûi,t) against SDP, helps us assess the goodness-of-fit of the model by showing how well the predicted values fit the actual data, it also tells us if our model is homoskedastic or hetereoskedastic. The third plot, which is a scatter plot of the predicted values of EQI against the true values of EQI, can help evaluate the accuracy of the model's predictions.

Together, these plots can help identify potential issues with our regression model, such as non-linearity, heteroscedasticity, or poor fit. They also provide a visual representation of the relationship between the independent and dependent variables.

8. Plot a histogram of $\hat{u}_{i,t}$ and verify that $\sum_{i,t} \hat{u}_{i,t} = 0$.




```
> print(paste0("Sum of Residuals: ", sum_residuals))
[1] "Sum of Residuals: 9.12470099478924e-11"
```

Sum of Residuals =~ 0.00000000009 which is very close to 0. Hence, verified.

(9) Finally, estimate the following regression, summarize the results in a table and interpret. Note that i indexes districts, t indexes years, and γ(i,t) is random error.

EQI(i,t) = α0 + α1SDP(i,t) + α2(SDP(i,t))^2 + α3(SDP(i,t))^3 + α4GINI(i) + γ(i,t)

Result of running the above linear regression model:

```
Call:
lm(formula = Amount.of.Hydrogencarbonate ~ SDP + sdp2 + sdp3 +
    Ginni.Index, data = df5)

Residuals:
    Min      1Q  Median      3Q     Max
-320.95  -91.75   -5.32   85.46  786.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.150e+02  7.338e+00  29.292   <2e-16 ***
SDP          5.029e-04  3.159e-05  15.919   <2e-16 ***
sdp2        -7.066e-10  5.356e-11 -13.191   <2e-16 ***
sdp3         2.688e-16  2.337e-17  11.503   <2e-16 ***
Ginni.Index  3.801e+01  2.256e+01   1.685    0.092 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.1 on 6029 degrees of freedom
  (1145 observations deleted due to missingness)
Multiple R-squared:  0.05014,   Adjusted R-squared:  0.04951
F-statistic: 79.56 on 4 and 6029 DF,  p-value: < 2.2e-16
```

sdp2 represents the square of SDP.
Sdp3 represents the cube of SDP.

Summarizing the results

Dependent Variable : EQI (i.e. Amount of Hydrogen carbonate) , N = 6034, R2 = 0.049	
Explanatory Variables	**Coefficient**
sdp	0.0005***
sdp2	-0.0000000007***
sdp3	0.0000000000000002 ***
ginni	38.01
Intercept	215.00 ***

*** p-value is less than 0.001

Interpretation:

The above model is a polynomial regression model where SDP is included up to the third power. The coefficient α0 represents the intercept or the value of EQI when all other independent variables are equal to zero. α1, α2, and α3 represent the effects of SDP on EQI, with α1 being the linear effect, α2 the quadratic effect, and α3 the cubic effect. α4 represents the effect of GINI on EQI. γ(i,t) represents the random variation that is not accounted for by the independent variables.

- The coefficient for SDP (α1) is 0.00005, which suggests a positive relationship between economic development and environmental quality, holding all other variables constant. For every unit increase in SDP, EQI is expected to increase by 0.00005 units.
- The coefficient for SDP squared (α2) is -0.0000000007, which suggests a non-linear relationship between economic development and environmental quality. The negative sign implies that the relationship between SDP and EQI is inverted U-shaped, with the maximum EQI occurring at some intermediate level of economic development.
- The coefficient for SDP cubed (α3) is 0.0000000000000002, which suggests that the effect of economic development on environmental quality is likely to be negligible beyond the point of maximum EQI.
- The coefficient for GINI (α4) is 38.01, which suggests a positive relationship between income inequality and environmental quality, holding all other variables constant. For every unit increase in GINI, EQI is expected to increase by 38.01 units.
- The intercept (α0) is 215, which represents the expected value of EQI when all independent variables are zero. In this case, it represents the expected value of EQI in a hypothetical scenario where economic development and income inequality are both absent.

R-squared = 0.0005: This indicates that only 0.05% of the variation in the dependent variable can be explained by the independent variables included in the analysis. In other words, the model does not fit the data very well.

Members:
Ashutosh Gera (2021026)
Anubhav Patel (2019148)
Razafiamiadana Jacqueline Precilia (2021411)
Dev mann (2021382)

Thank you :D