Data Assignment 1

Q1.

Environmental Quality Measure: Ground Water Quality

Environmental Quality Indicators:

1. Amount of Arsenic
2. Amount of carbonate
3. Amount of Calcium
4. Amount of Chloride
5. Amount of Electrical Conductivity
6. Amount of Fluorine
7. Amount of Iron
8. Amount of Hydrogencarbonate
9. Amount of Potassium
10. Amount of Magnesium
11. Amount of Nitrate
12. Amount of Sodium
13. Percentage of Sodium
14. Amount of Phosphate Ion
15. Amount of Residual Sodium Carbonate
16. Amount of Sodium absorption ratio
17. Amount of Sulfate
18. Amount of Silicon dioxide
19. Amount of Hardness Total
20. Amount of Alkalinity Total
21. Amount of Total Dissolved Solids
22. Amount of Potential of Hydrogen

Range of Years: 2000 - 2018

Data Assignment 1

Q2. Done in R Script

Q3. Done in R Script

Q4. Done in R Script

IMPORTANT NOTE: There were several misspellings (of districts) in the Gini Coefficient Tables/Ground Water Quality data. Some changes were made to the Gini Coefficient Tables manually in order to match the spellings of the districts between the data.

Note: In the metadata file, no units have been given for most of the indicators.
Only:

Environmental Quality Indicator	Unit of Measurement
Amount of Residual Sodium Carbonate	mEq/L (Milliequivalents per Liter)
Amount of Hardness Total	MgCaCO3/L (calcium hardness)
Amount of Alkalinity Total	MgCaCO3/L (calcium hardness)
Amount of Potential of Hydrogen	pH

Have been given.
SDP is measured in Rs Crore.

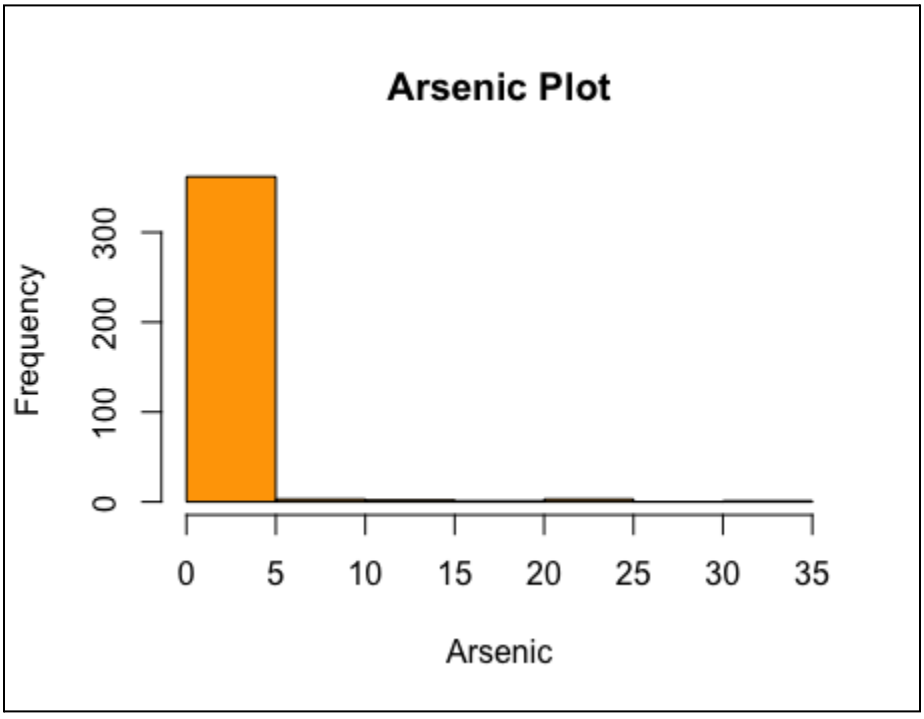
Data Assignment 1

Q5.

1. **Environmental Quality Indicator:** Arsenic

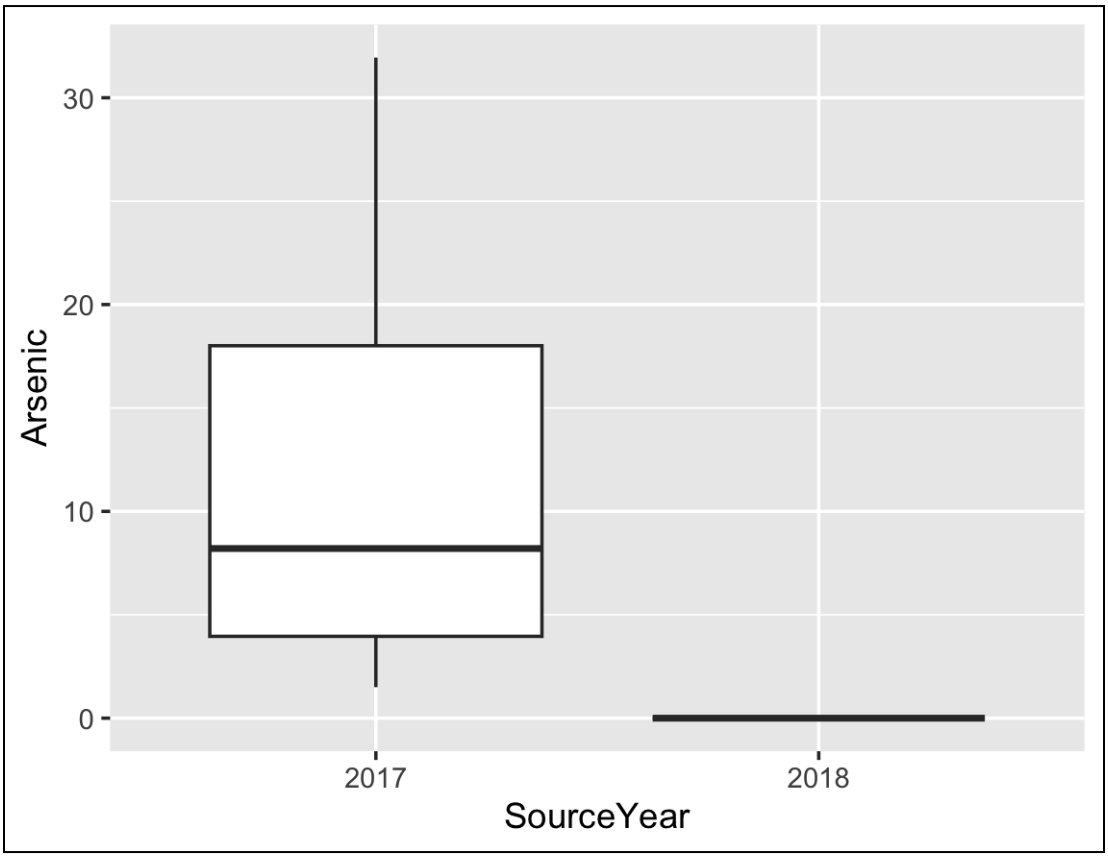
Statistic	N	Mean	St. Dev.	Min	Max
Arsenic	372	0.499	3.028	0.000	31.944

Histogram:



Data Assignment 1

Box Plot:



Skew: 7.343314

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

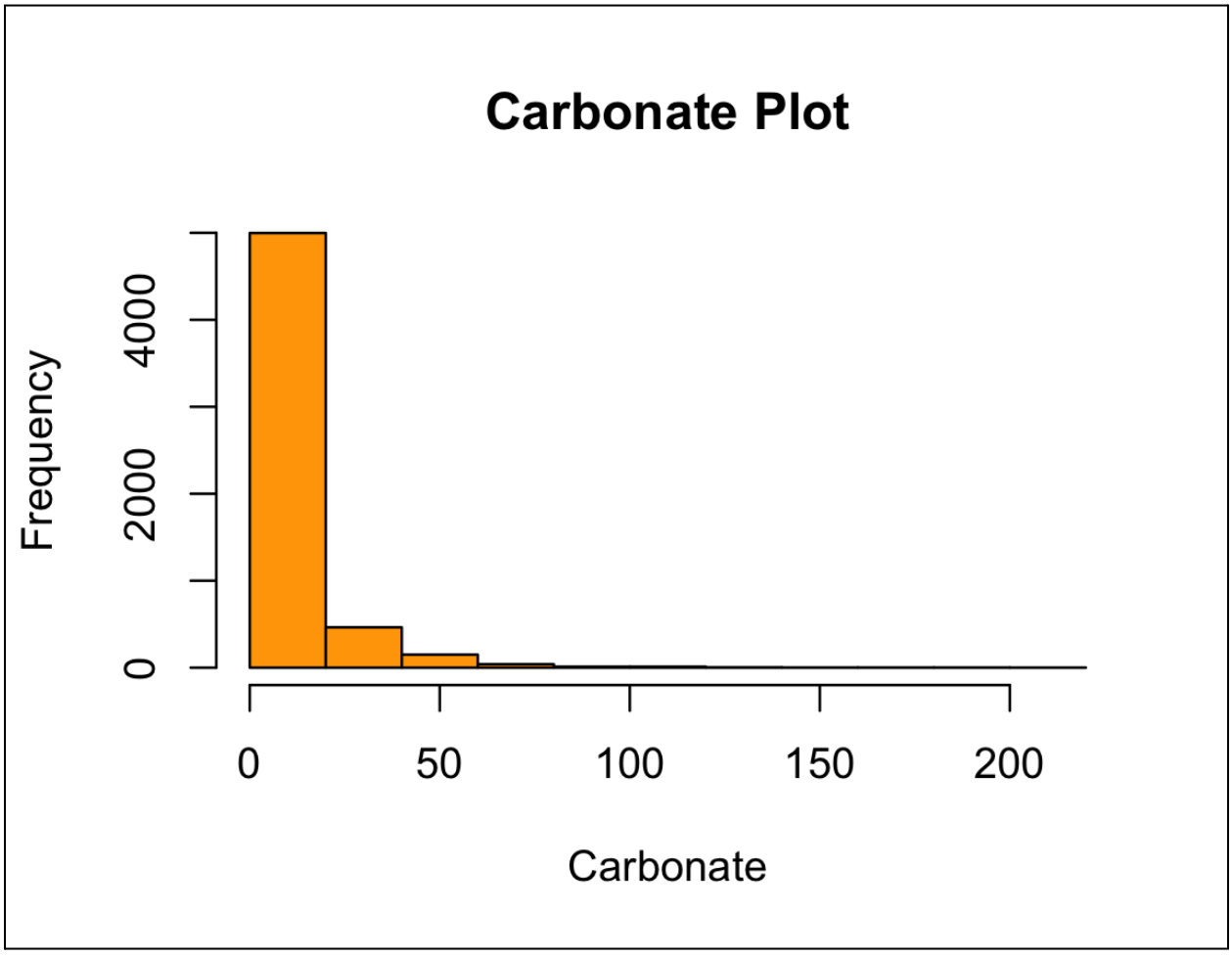
Outliers: Yes, there are some outliers. For example - there are some outliers between 20-25 in the histogram.

2. **Environmental Quality Indicator:** Carbonate

Data Assignment 1

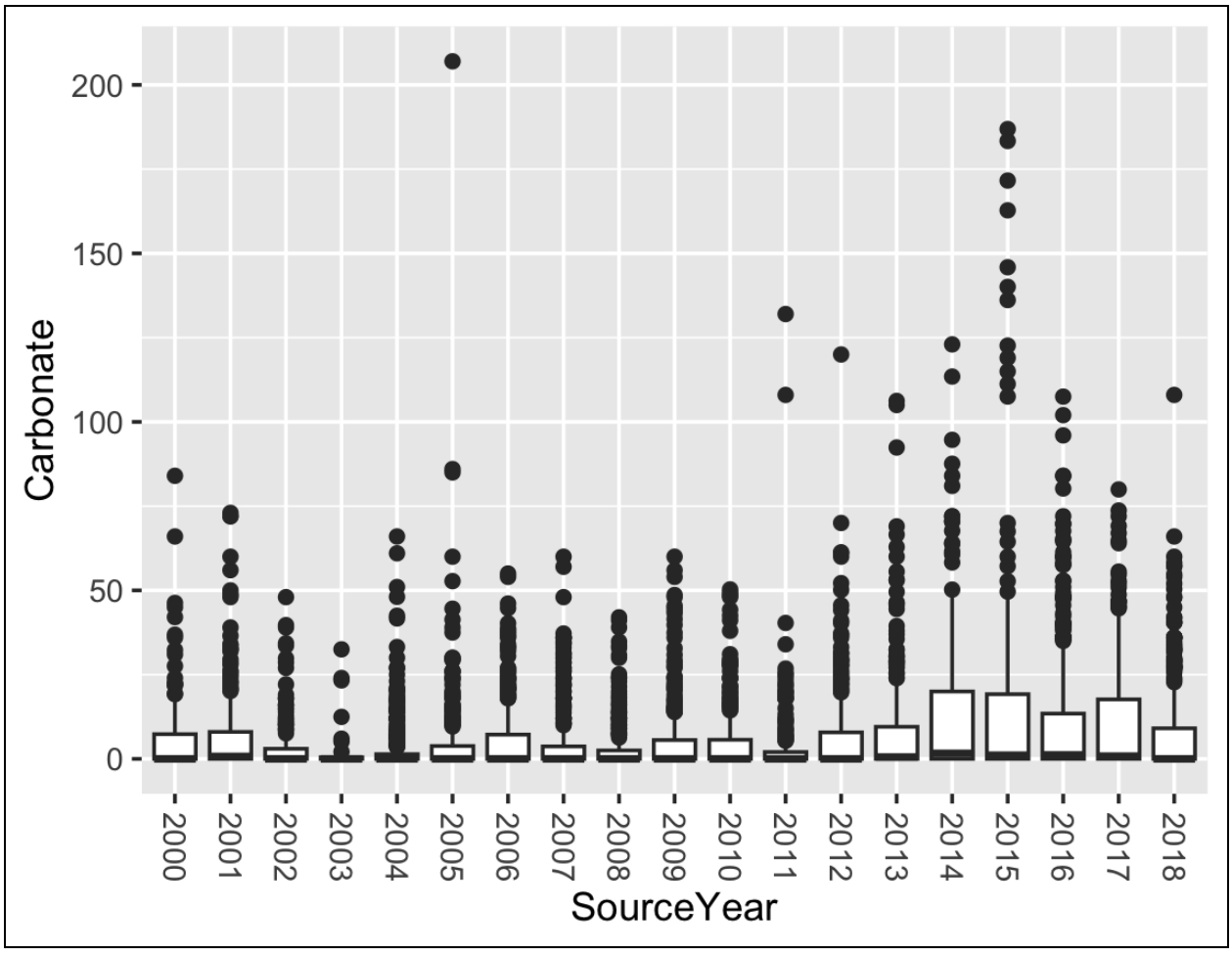
Statistic	N	Mean	St. Dev.	Min	Max
Carbonate	5,687	6.971	15.140	0.000	207.000

Histogram:



Data Assignment 1

Box Plot:



Skew: 4.264738

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

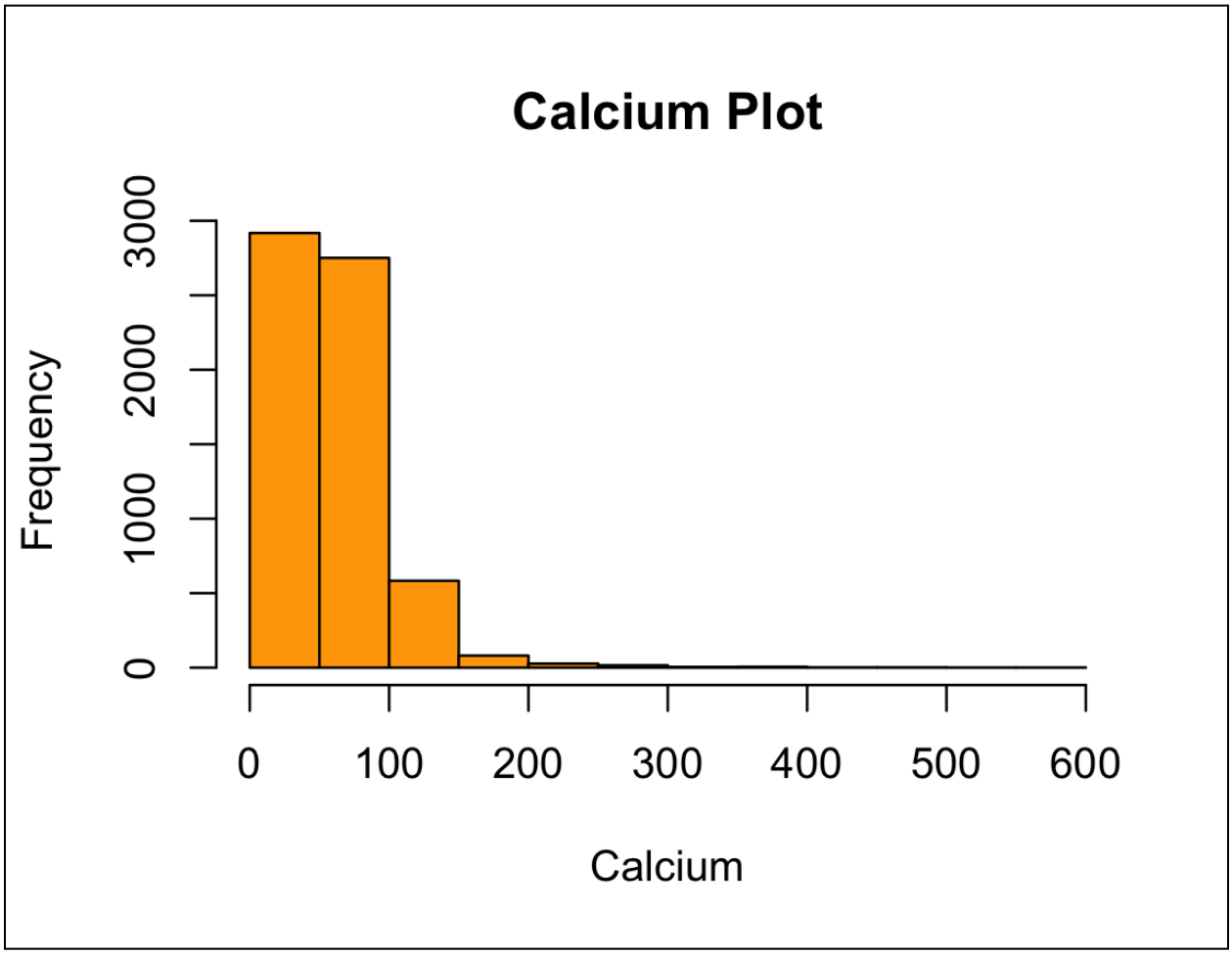
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 200 in 2015).

Data Assignment 1

3. **Environmental Quality Indicator:** Calcium

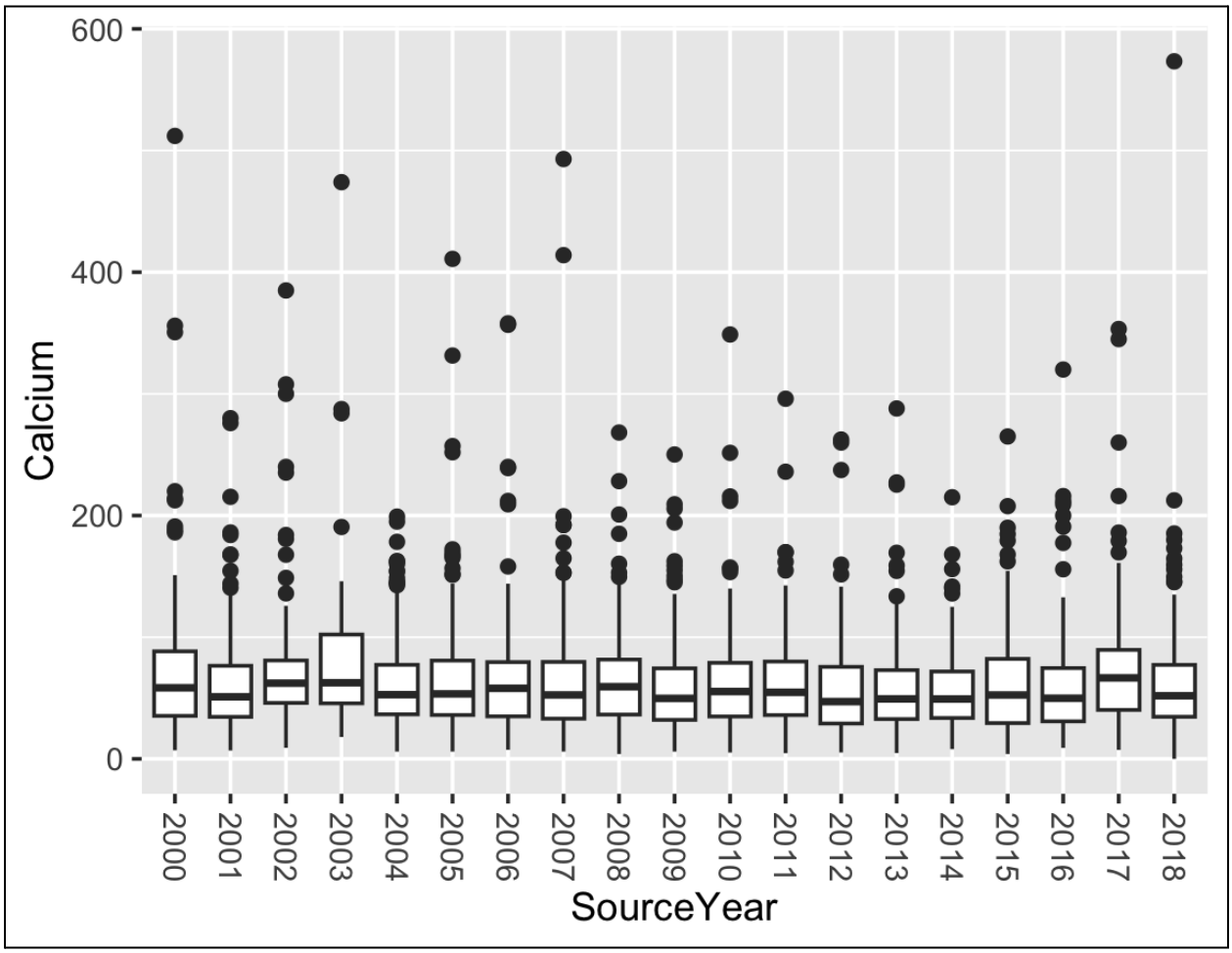
Statistic	N	Mean	St. Dev.	Min	Max
Calcium	6,393	60.495	39.533	0.000	573.333

Histogram:



Data Assignment 1

Box Plot:



Skew: 2.862771

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

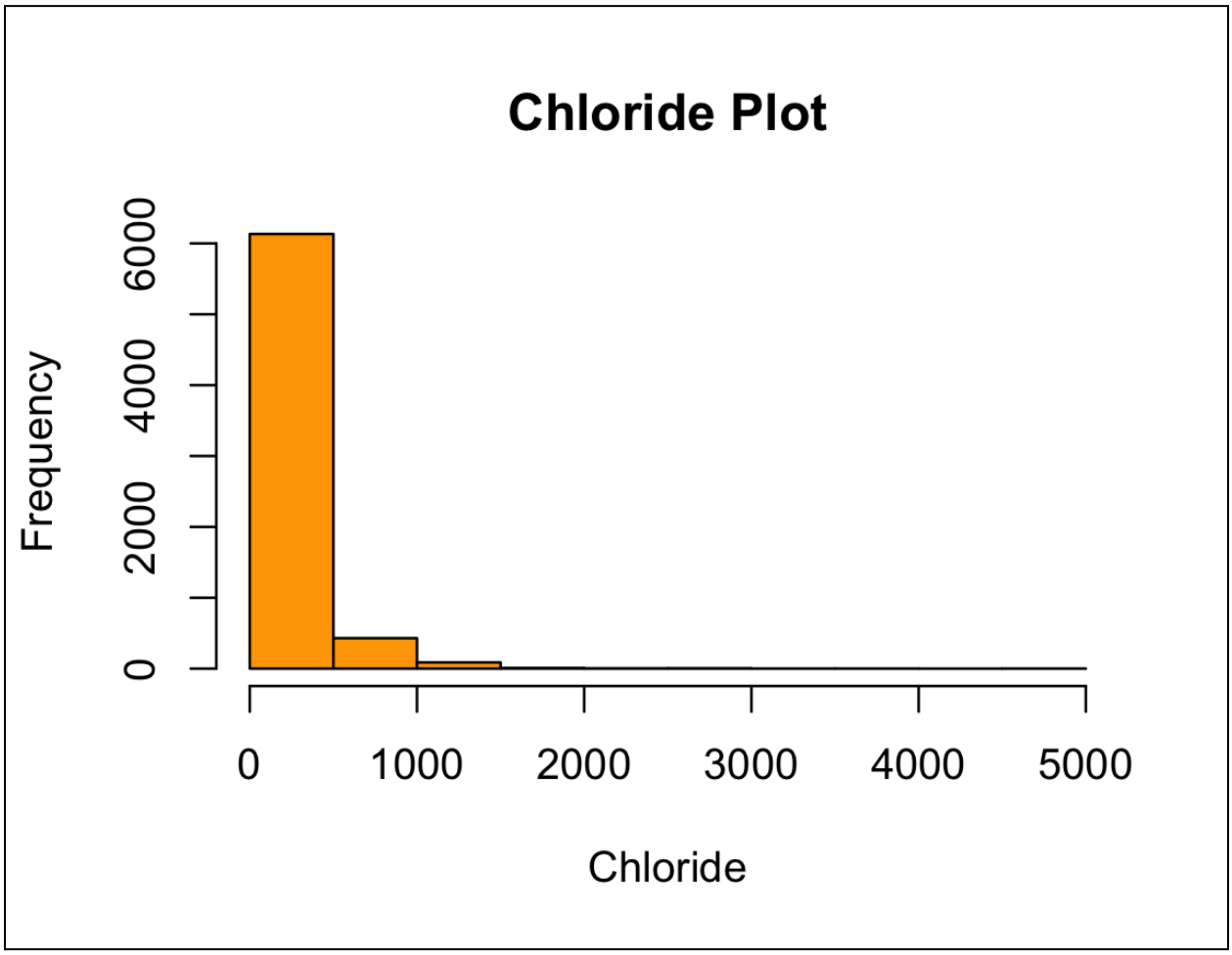
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 600 in 2018).

Data Assignment 1

4. **Environmental Quality Indicator:** Chloride

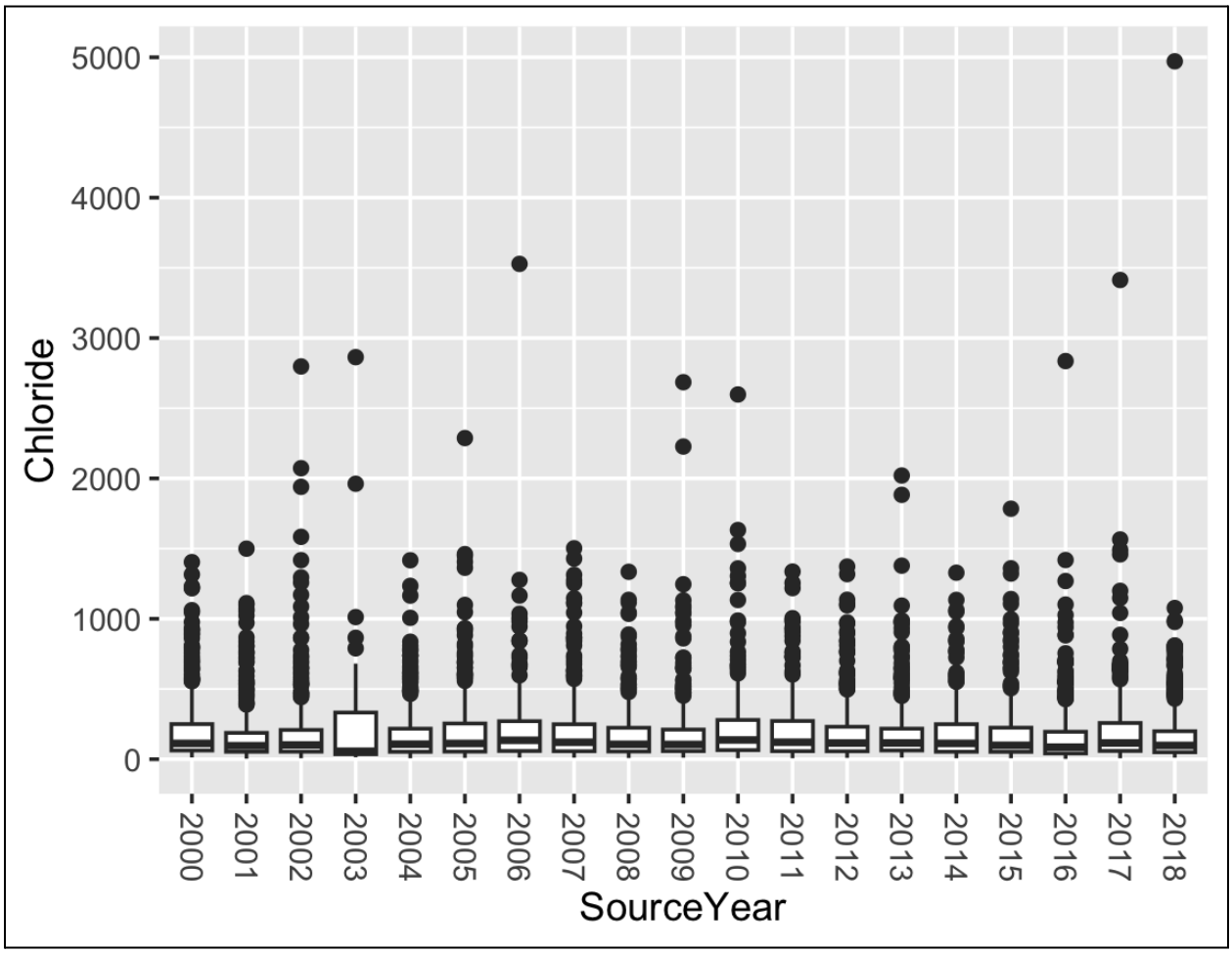
Statistic	N	Mean	St. Dev.	Min	Max
Chloride	6,672	191.942	248.511	2.351	4,971.333

Histogram:



Data Assignment 1

Box Plot:



Skew: 4.704272

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

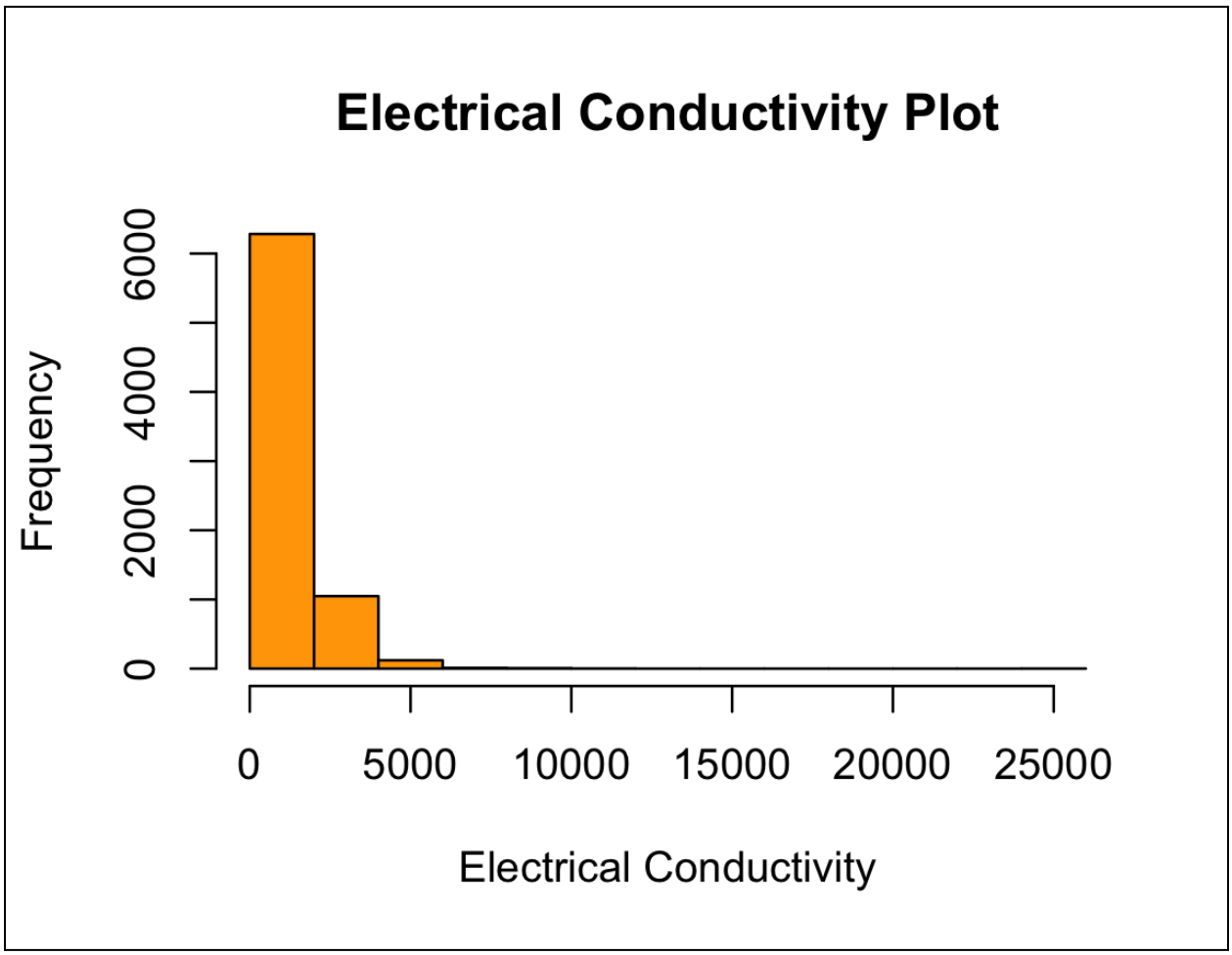
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 5000 in 2018).

Data Assignment 1

5. **Environmental Quality Indicator:** Electrical Conductivity

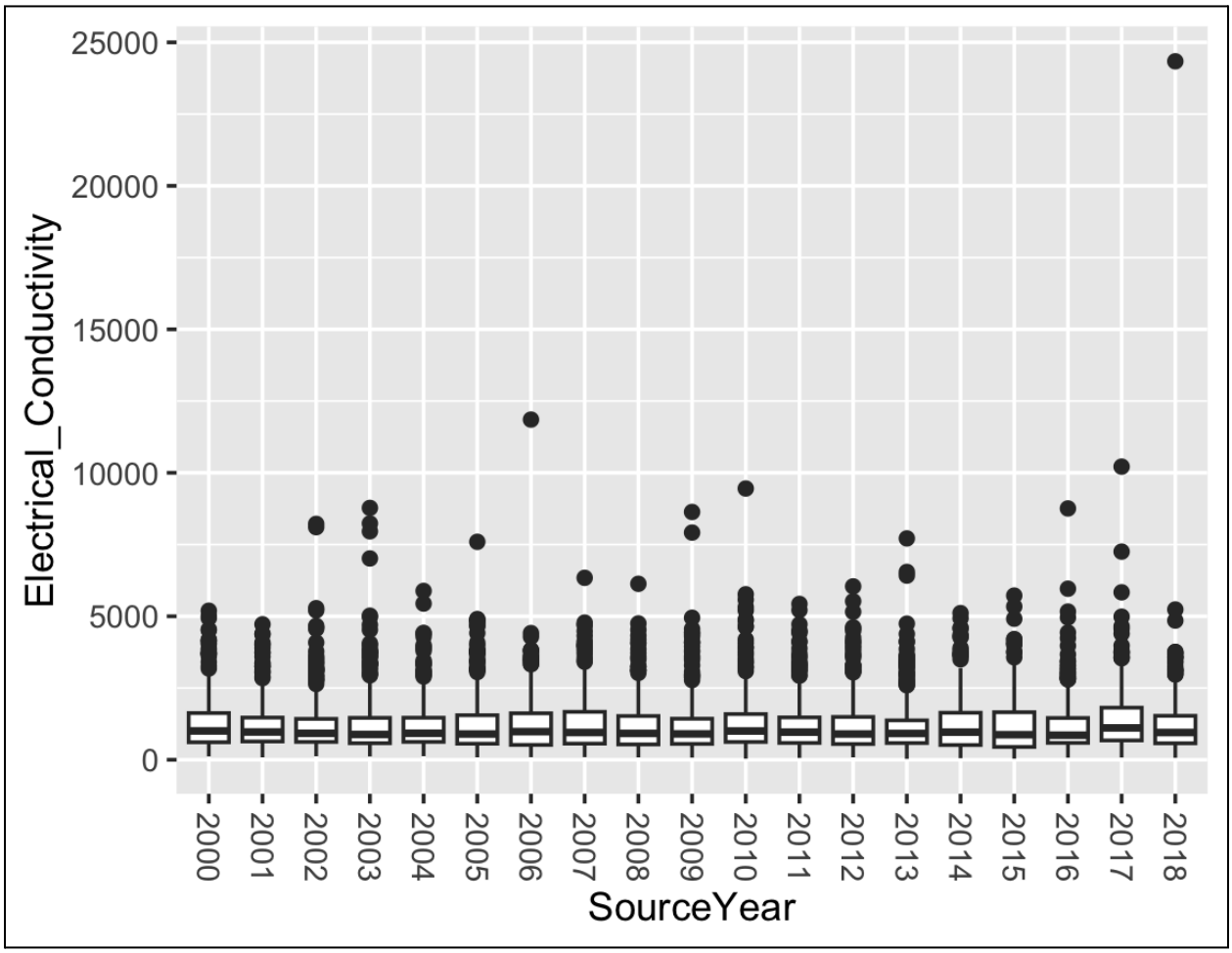
Statistic	N	Mean	St. Dev.	Min	Max
Electrical Conductivity	7,473	1,210.756	1,002.089	31.000	24,340.000

Histogram:



Data Assignment 1

Box Plot:



Skew: 3.702723

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

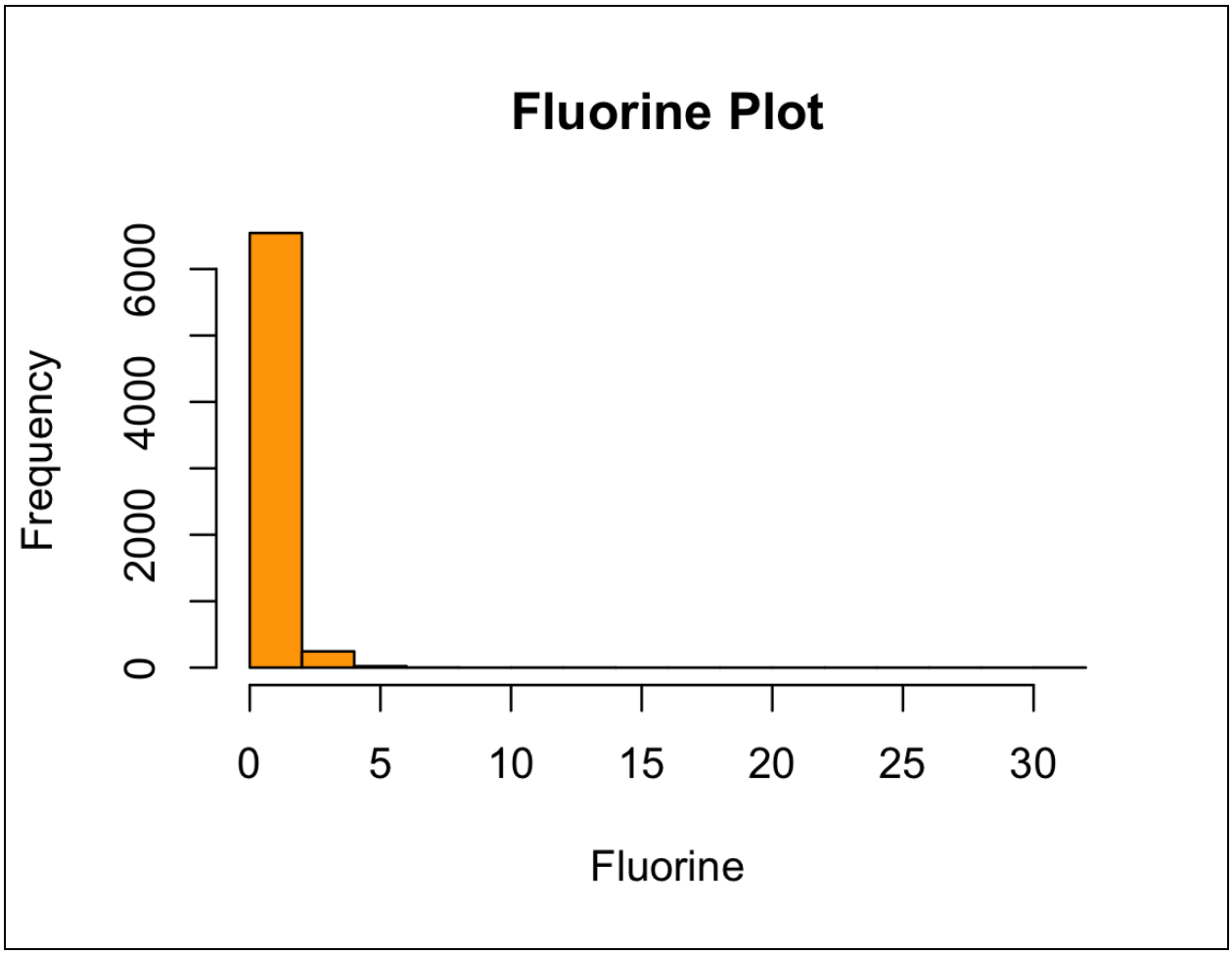
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 25000 in 2018).

Data Assignment 1

6. **Environmental Quality Indicator:** Fluorine

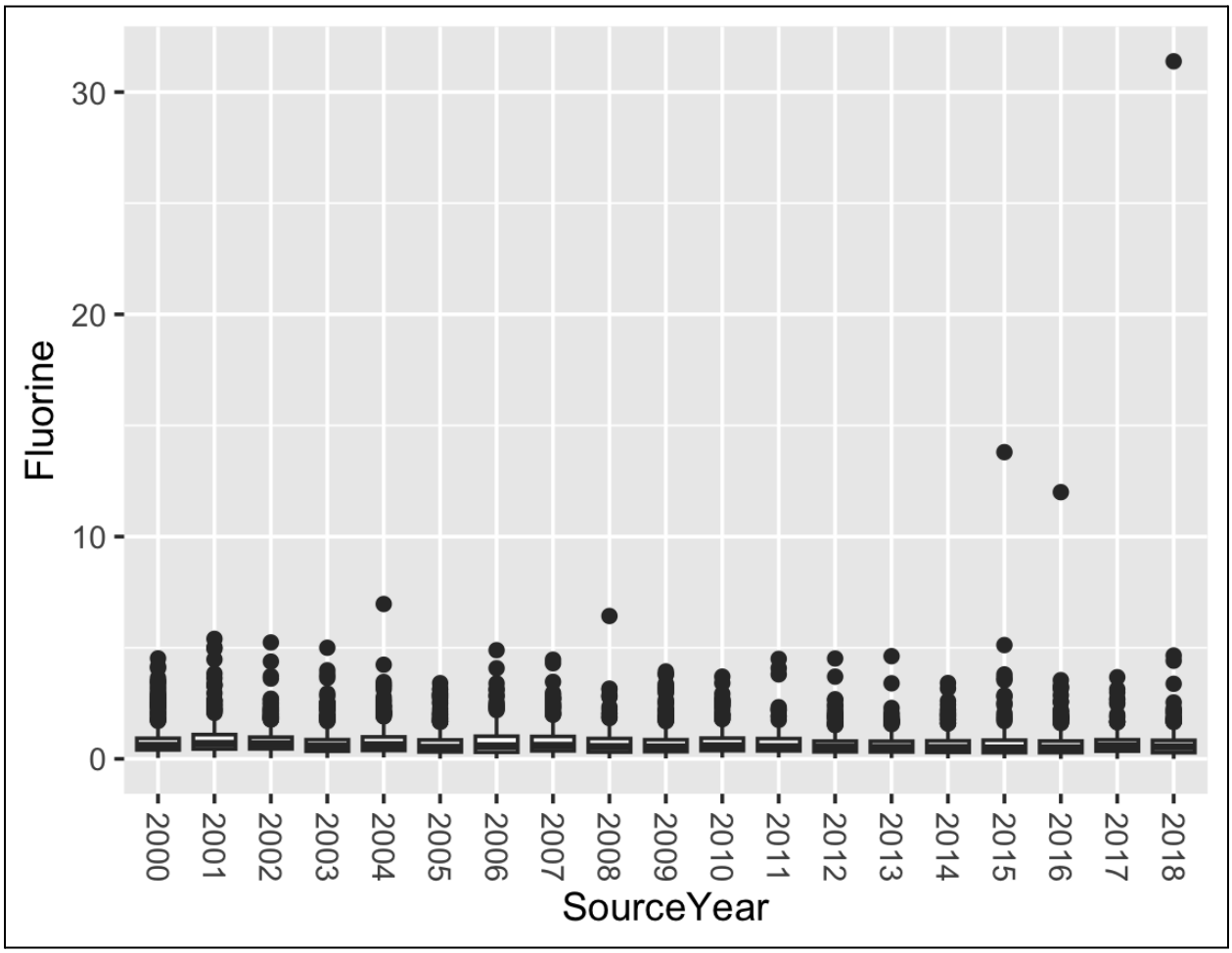
Statistic	N	Mean	St. Dev.	Min	Max
Fluorine	6,815	0.721	0.744	0.000	31.385

Histogram:



Data Assignment 1

Box Plot:



Skew: 13.01266

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

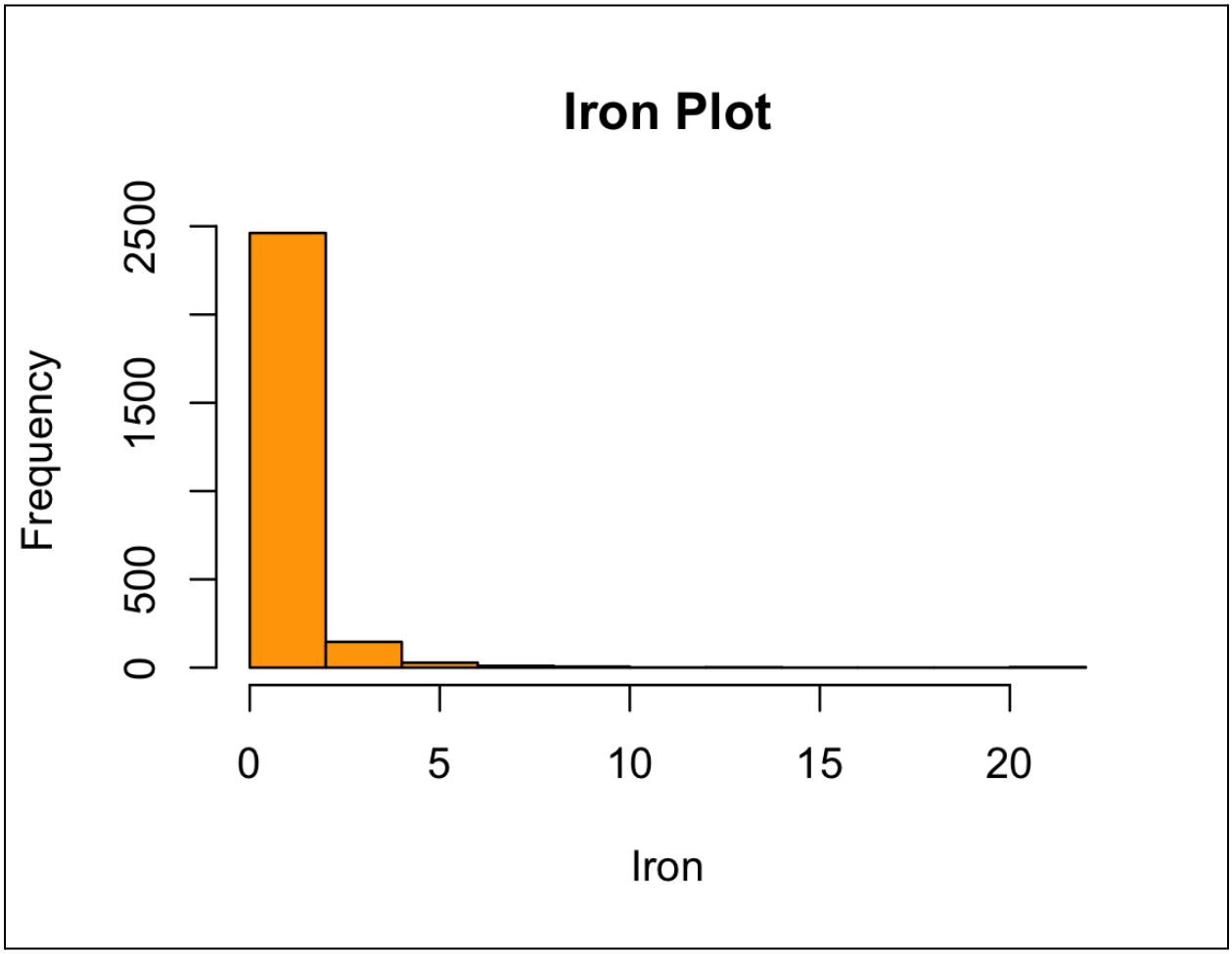
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 30 in 2018).

Data Assignment 1

7. **Environmental Quality Indicator:** Iron

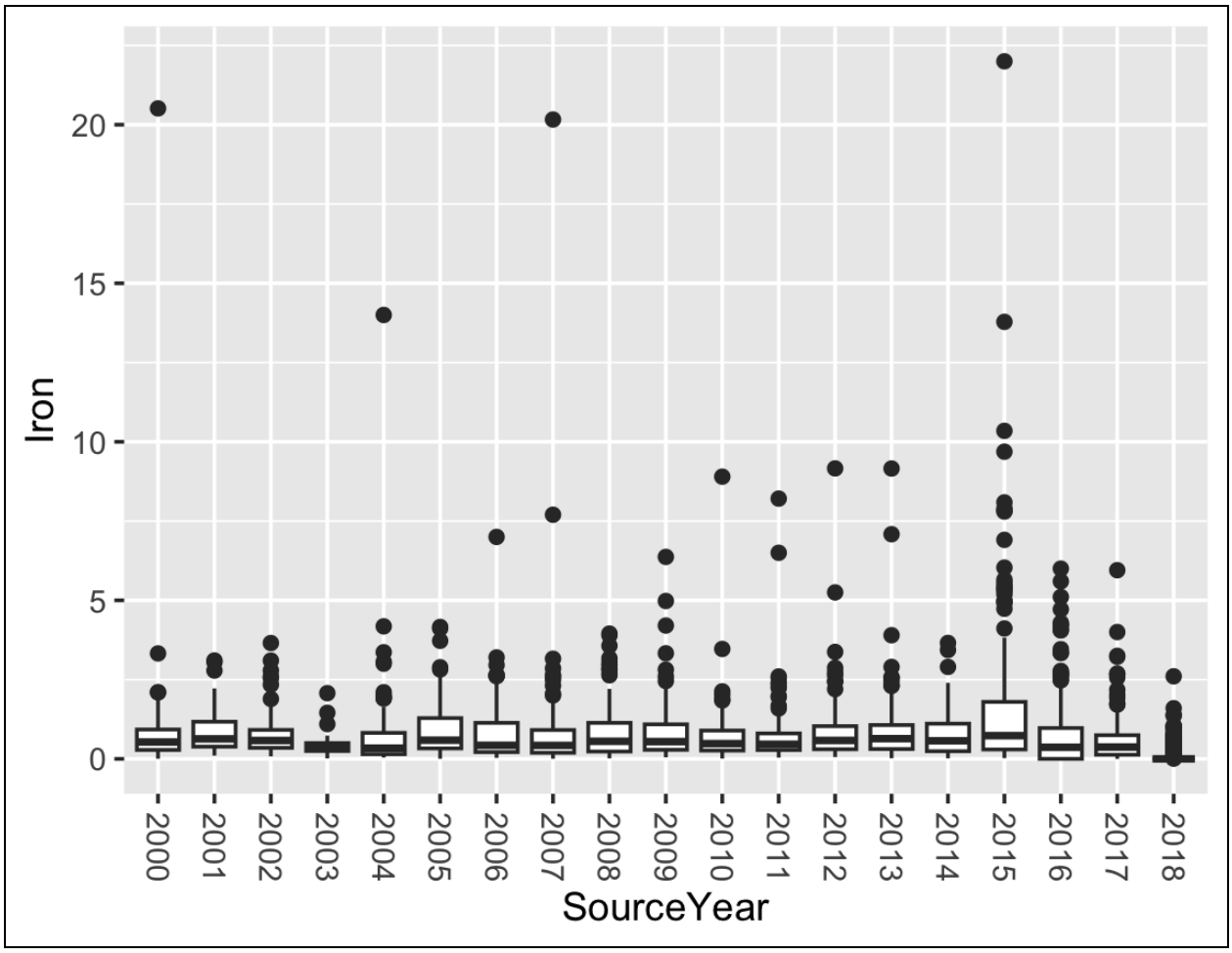
Statistic	N	Mean	St. Dev.	Min	Max
Iron	2,658	0.752	1.267	0.000	22.000

Histogram:



Data Assignment 1

Box Plot:



Skew: 7.255974

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

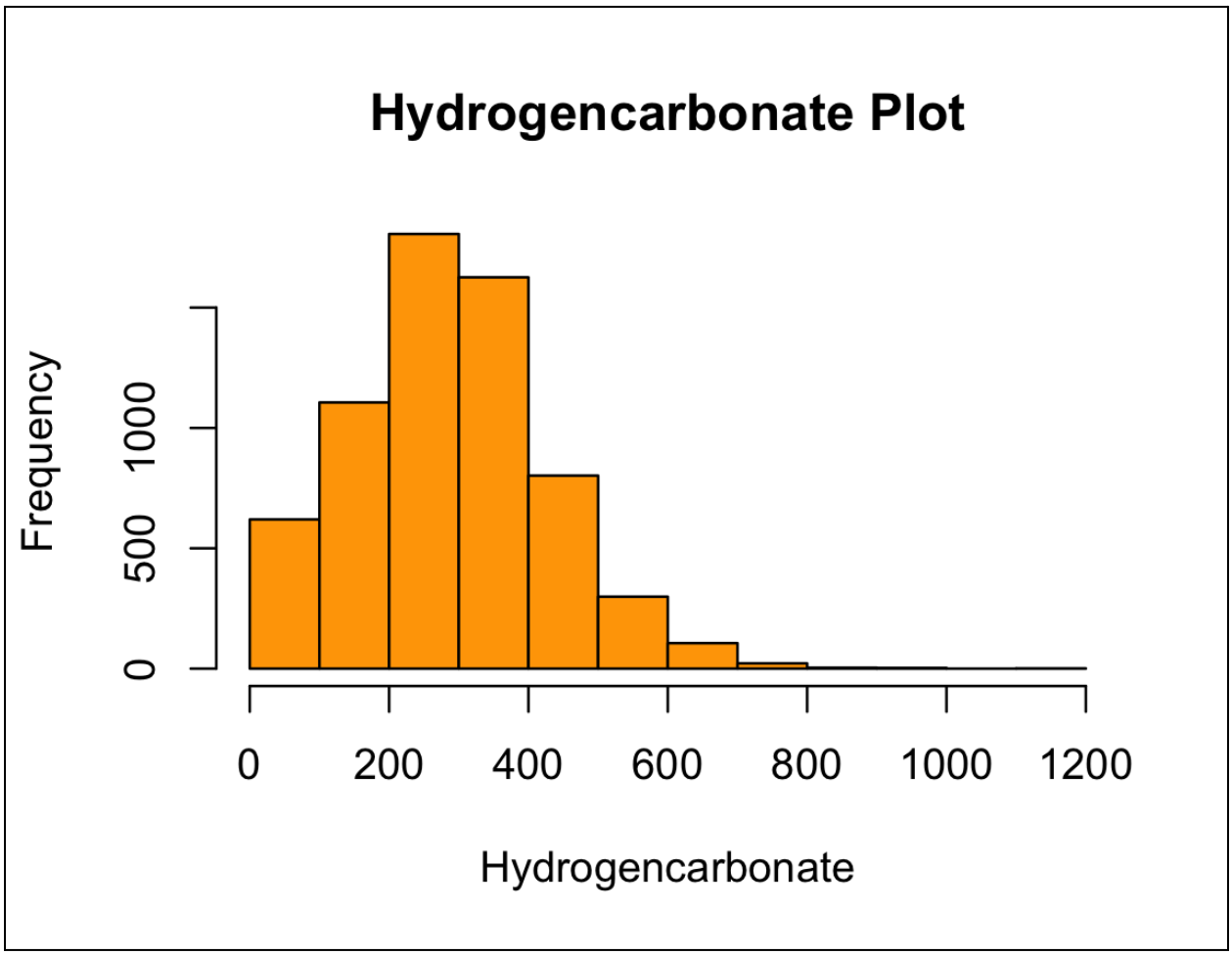
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points around value 20 in 2000, 2007 and 2015).

Data Assignment 1

8. **Environmental Quality Indicator:** Hydrogencarbonate

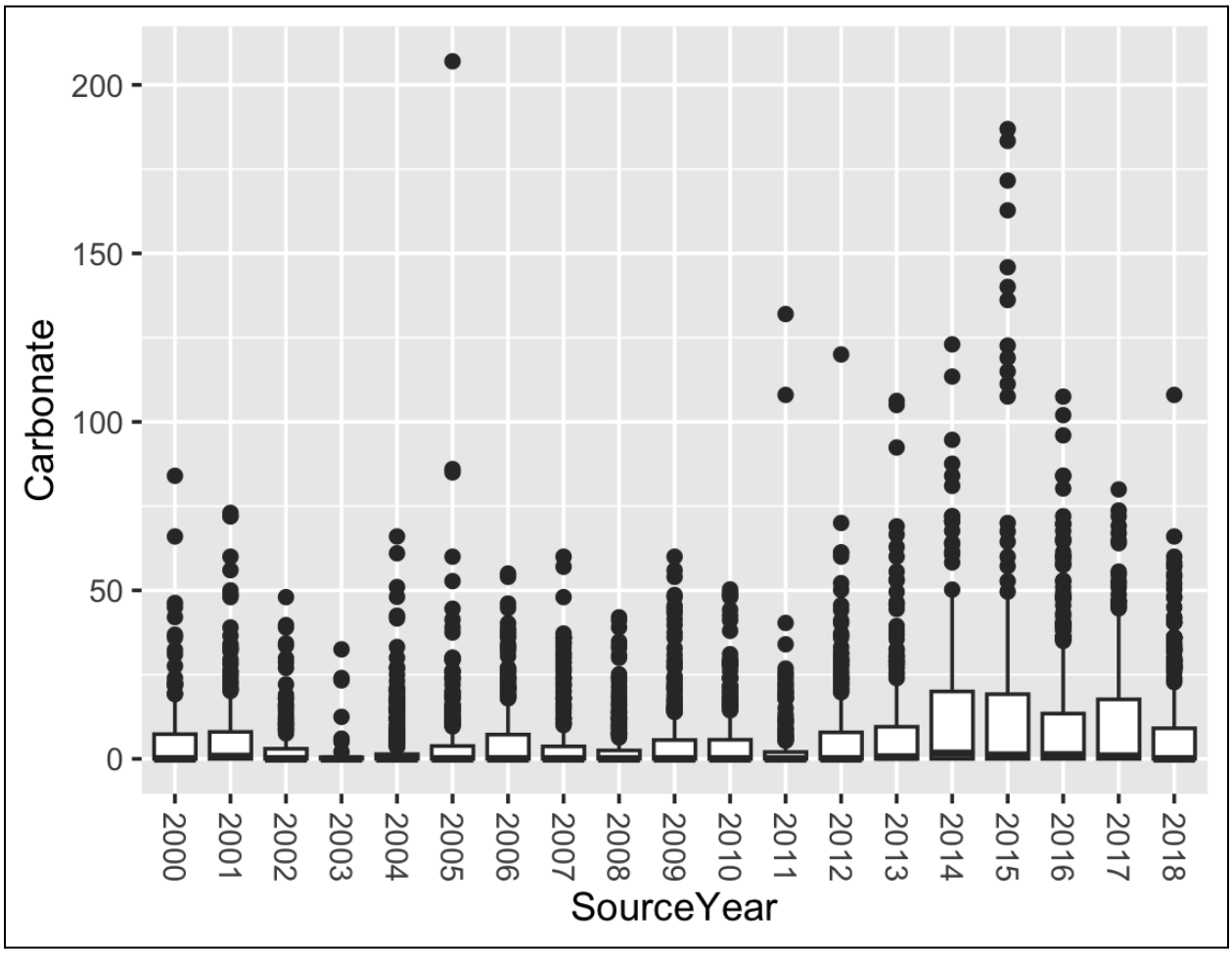
Statistic	N	Mean	St. Dev.	Min	Max
Hydrogencarbonate	6,395	286.566	139.174	0.000	1,118.333

Histogram:



Data Assignment 1

Box Plot:



Skew: 0.4290402

Shape of Distribution: Since the coefficient of skewness is slightly greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Although the data is distributed fairly normally, majority of the values are on the left side of the graph.

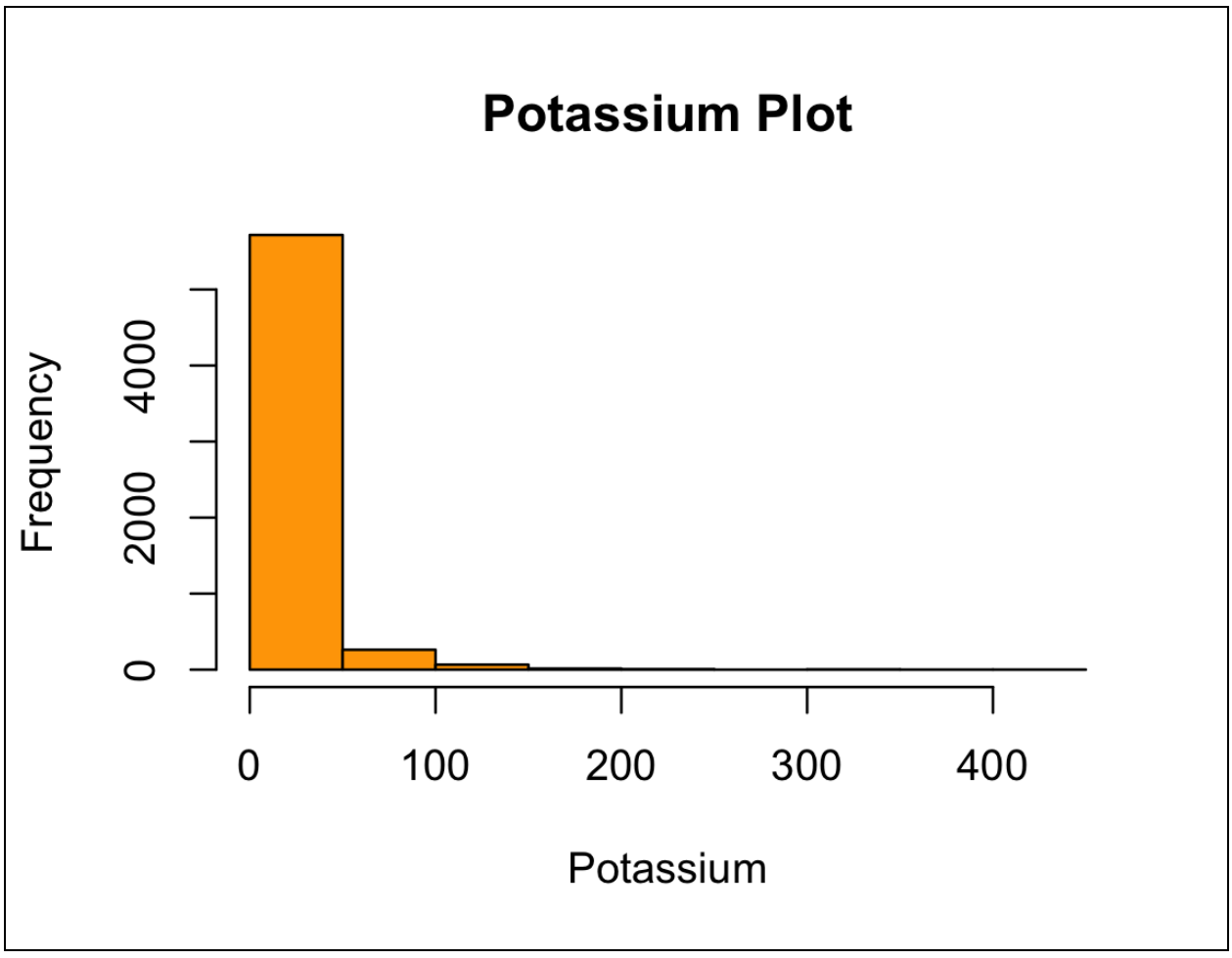
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 200 in 2005).

Data Assignment 1

9. **Environmental Quality Indicator:** Potassium

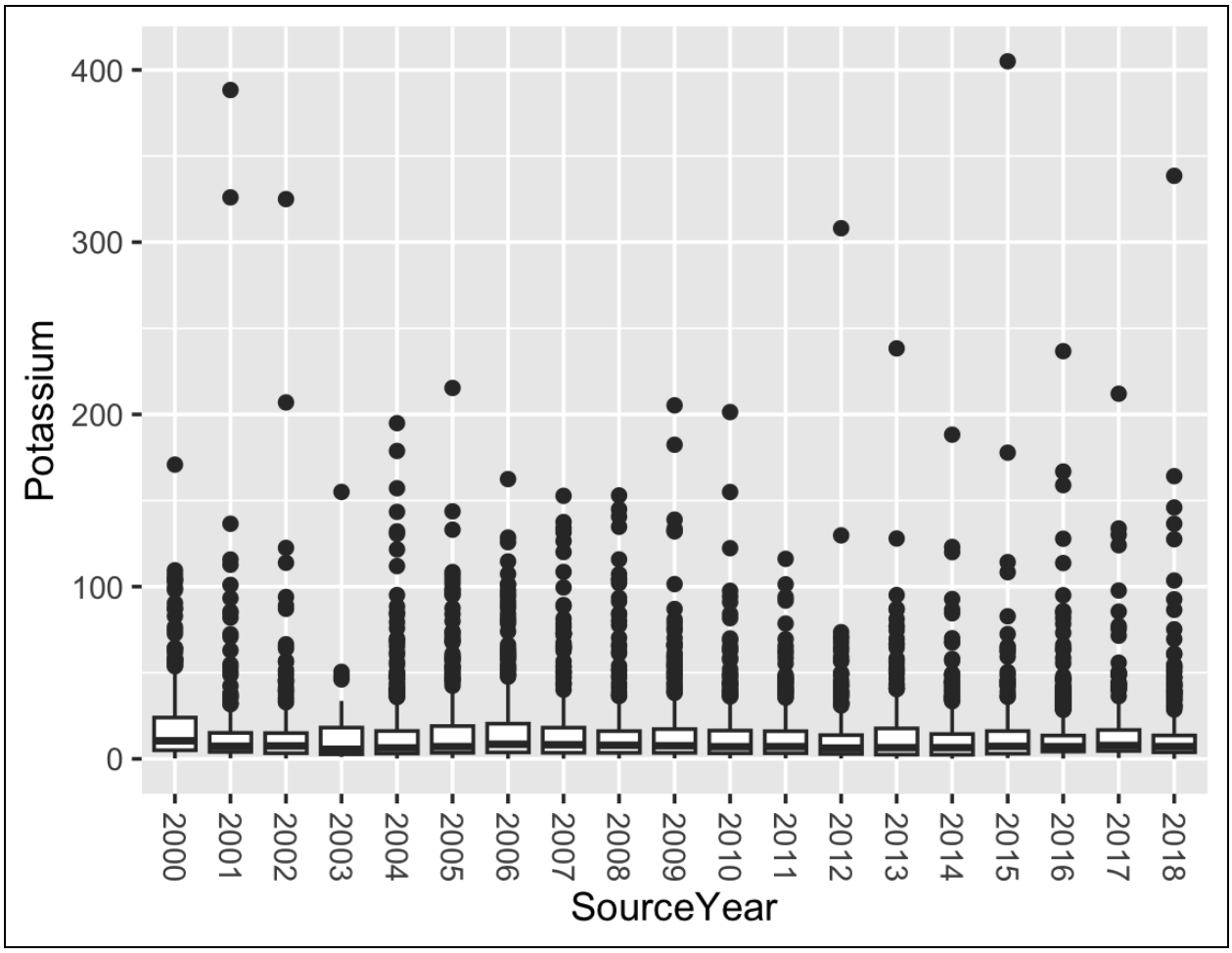
Statistic	N	Mean	St. Dev.	Min	Max
Potassium	6,073	15.306	24.390	0.000	405.000

Histogram:



Data Assignment 1

Box Plot:



Skew: 5.356444

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

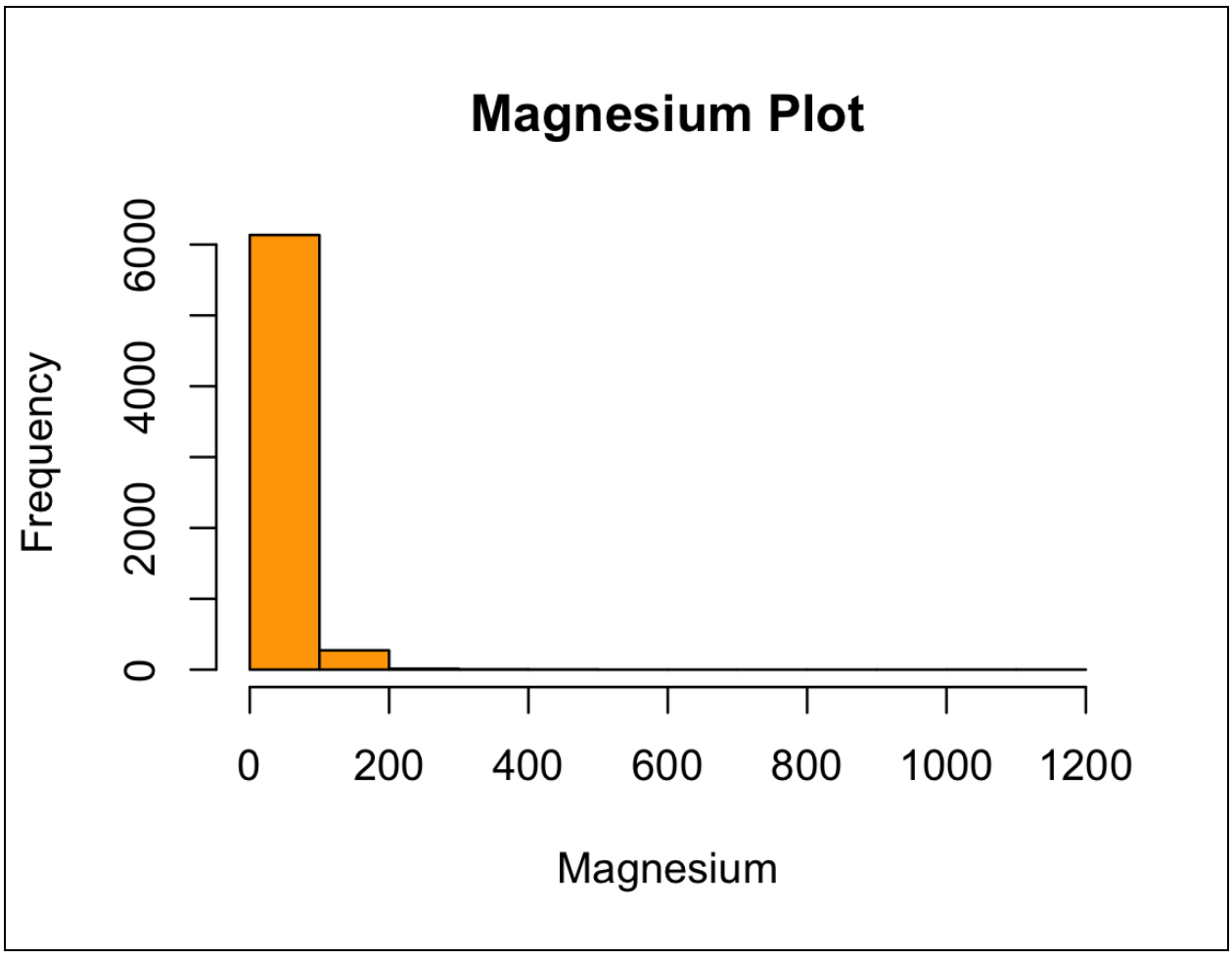
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points around value 400 in 2001 and 2015).

Data Assignment 1

10. **Environmental Quality Indicator:** Magnesium

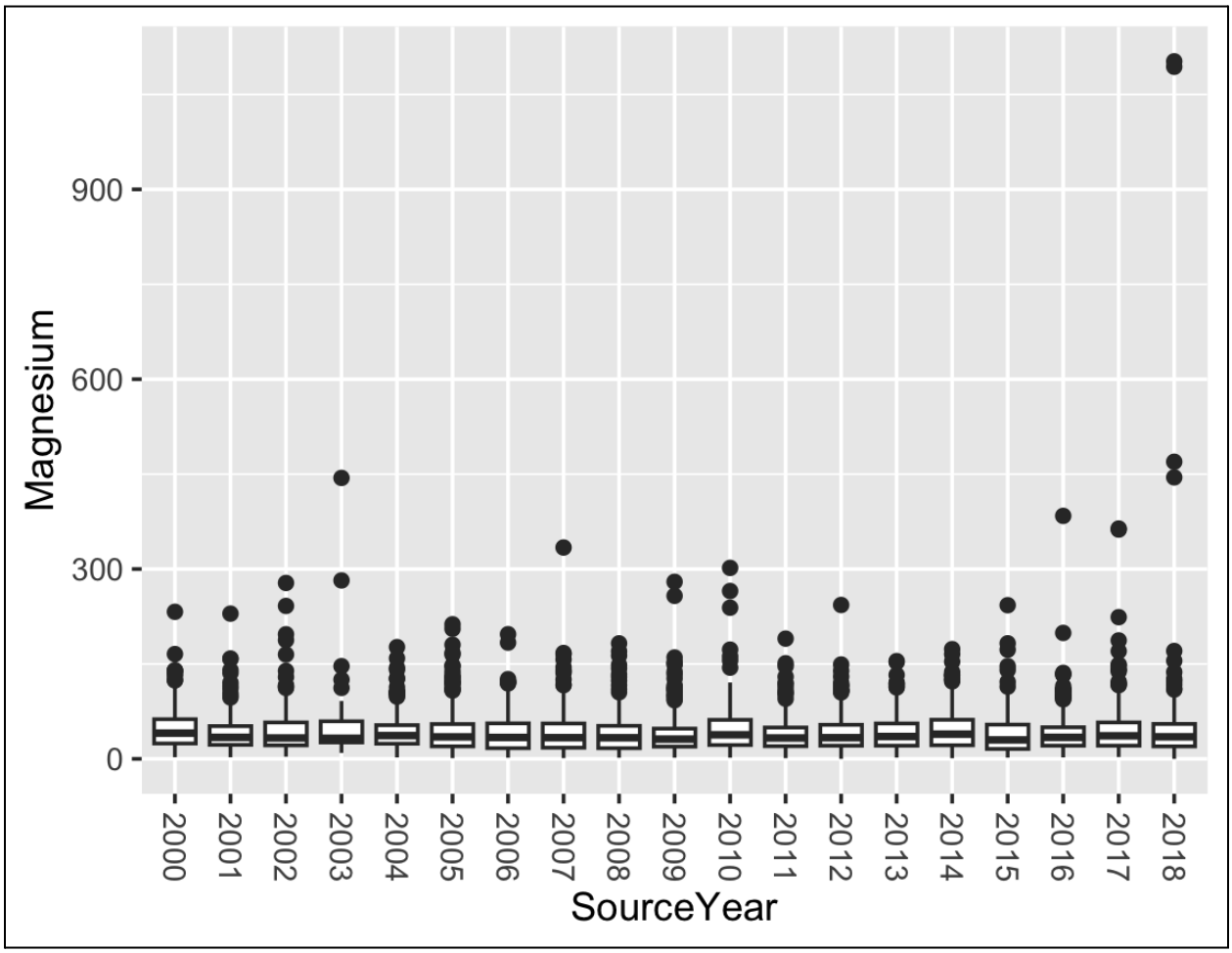
Statistic	N	Mean	St. Dev.	Min	Max
Magnesium	6,432	41.851	37.631	0.000	1,102.333

Histogram:



Data Assignment 1

Box Plot:



Skew: 8.834416

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

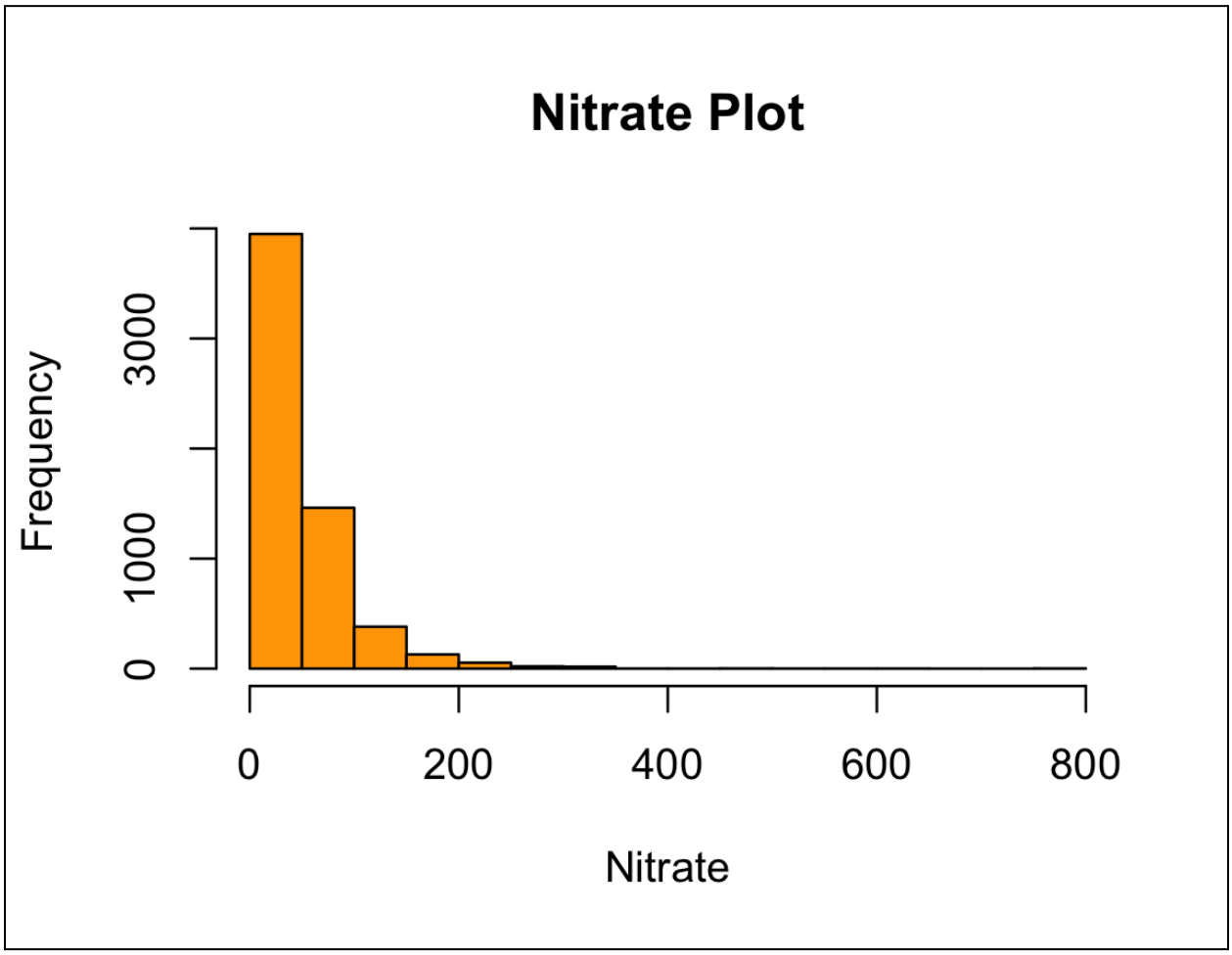
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points above value 900 in 2018).

Data Assignment 1

11. **Environmental Quality Indicator:** Nitrate

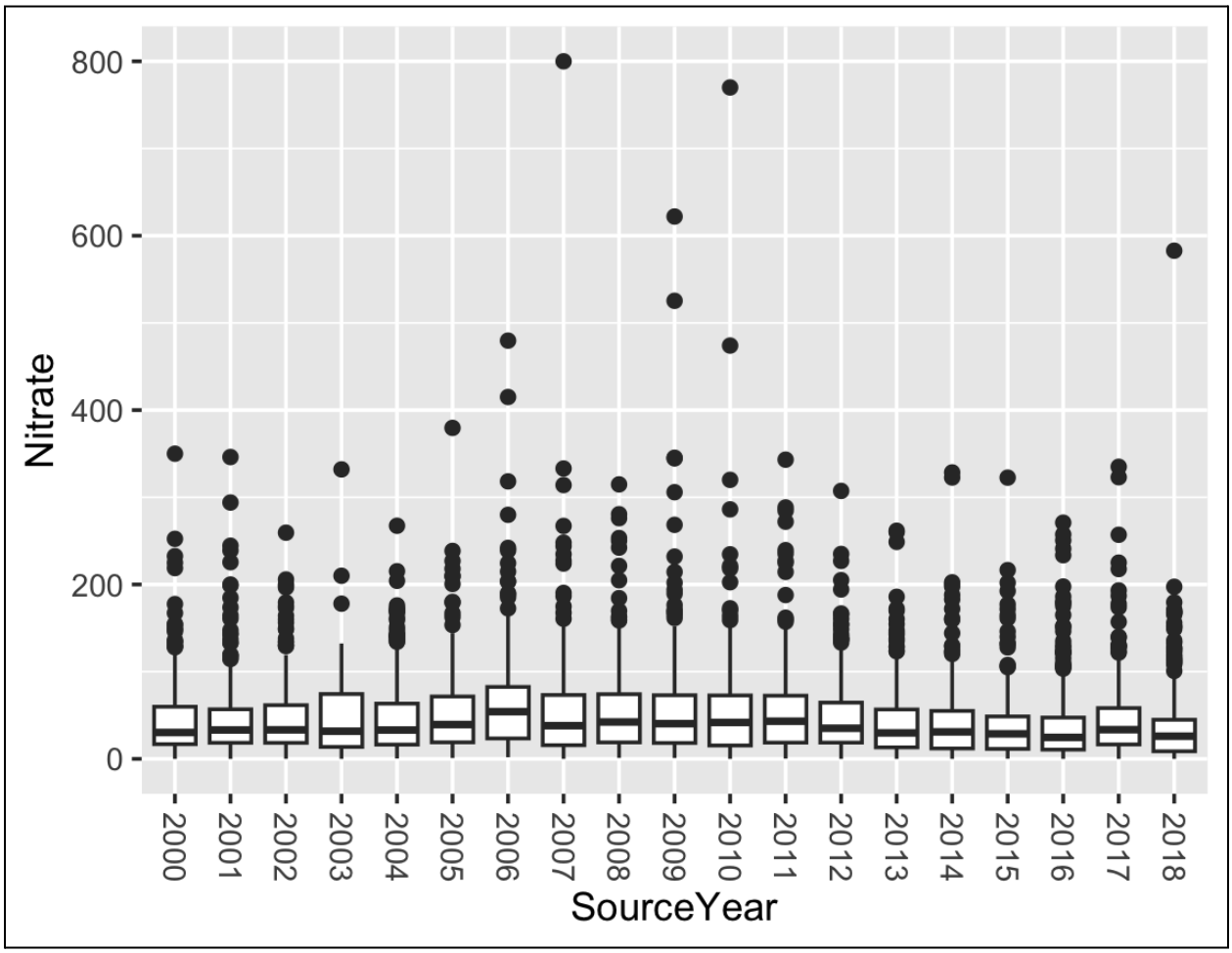
Statistic	N	Mean	St. Dev.	Min	Max
Nitrate	6,024	47.244	49.995	0.000	800.000

Histogram:



Data Assignment 1

Box Plot:



Skew: 3.655907

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

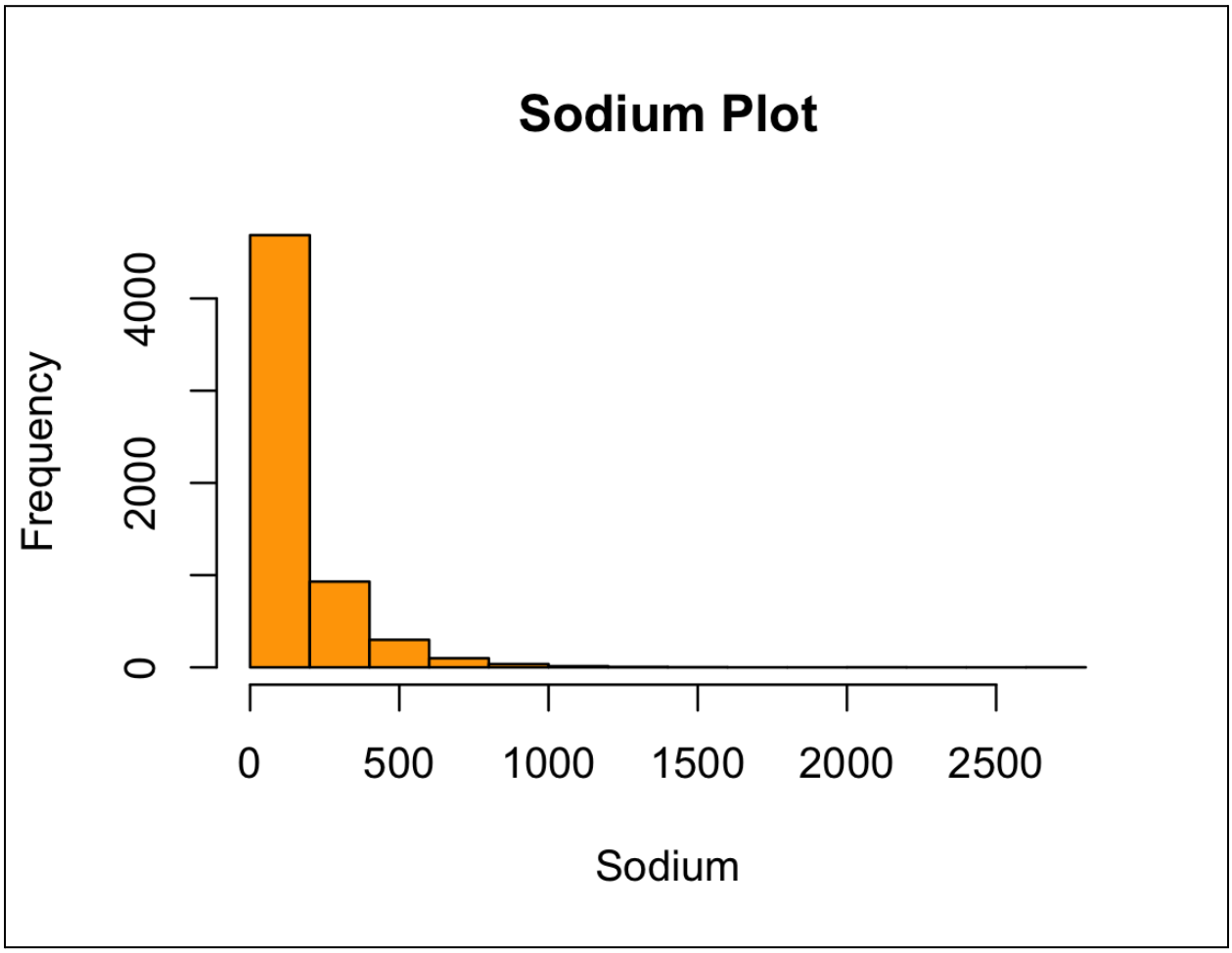
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points around value 800 in 2007 and 2010).

Data Assignment 1

12. **Environmental Quality Indicator:** Sodium

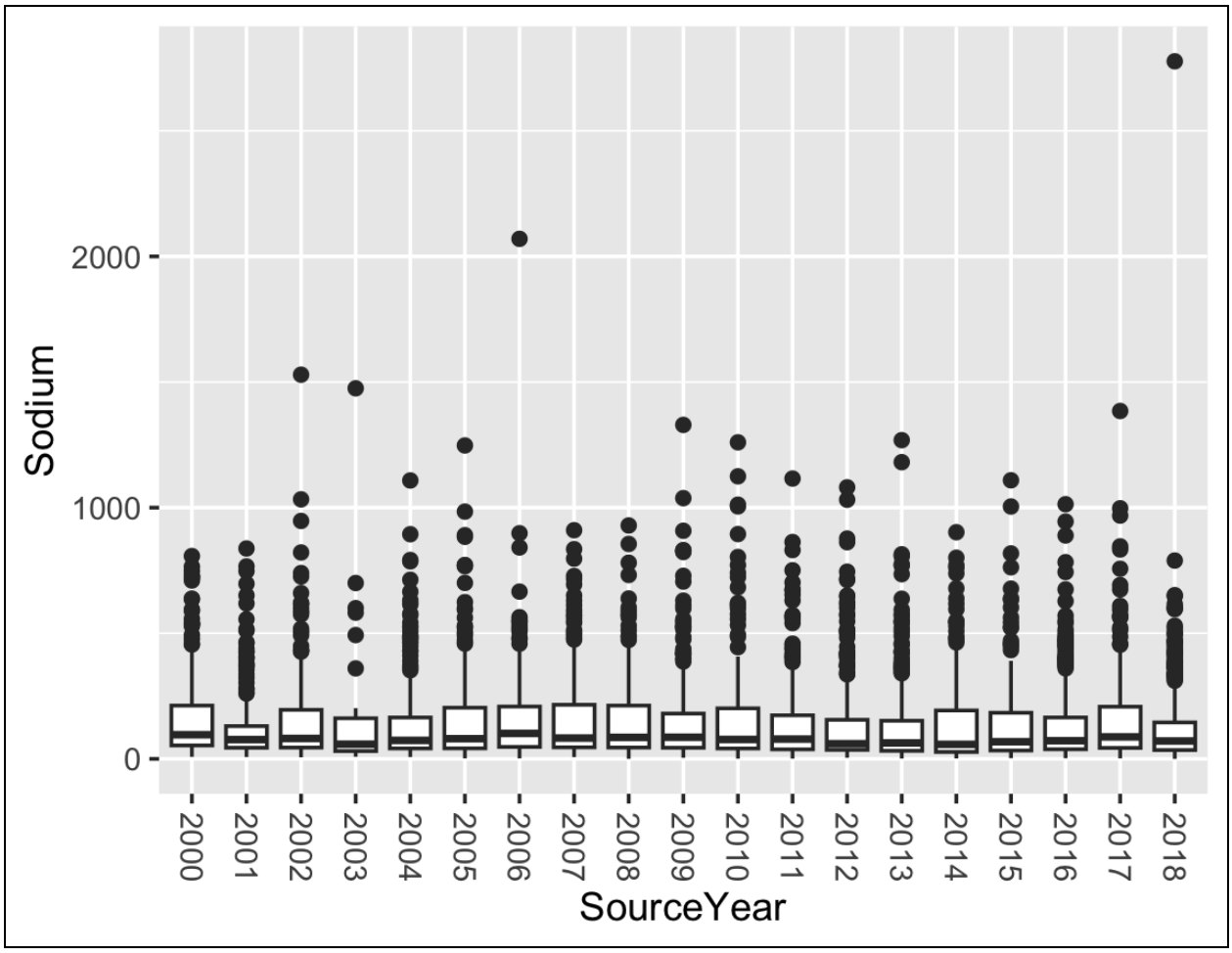
Statistic	N	Mean	St. Dev.	Min	Max
Sodium	6,073	142.041	169.322	0.000	2,776.667

Histogram:



Data Assignment 1

Box Plot:



Skew: 3.147331

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

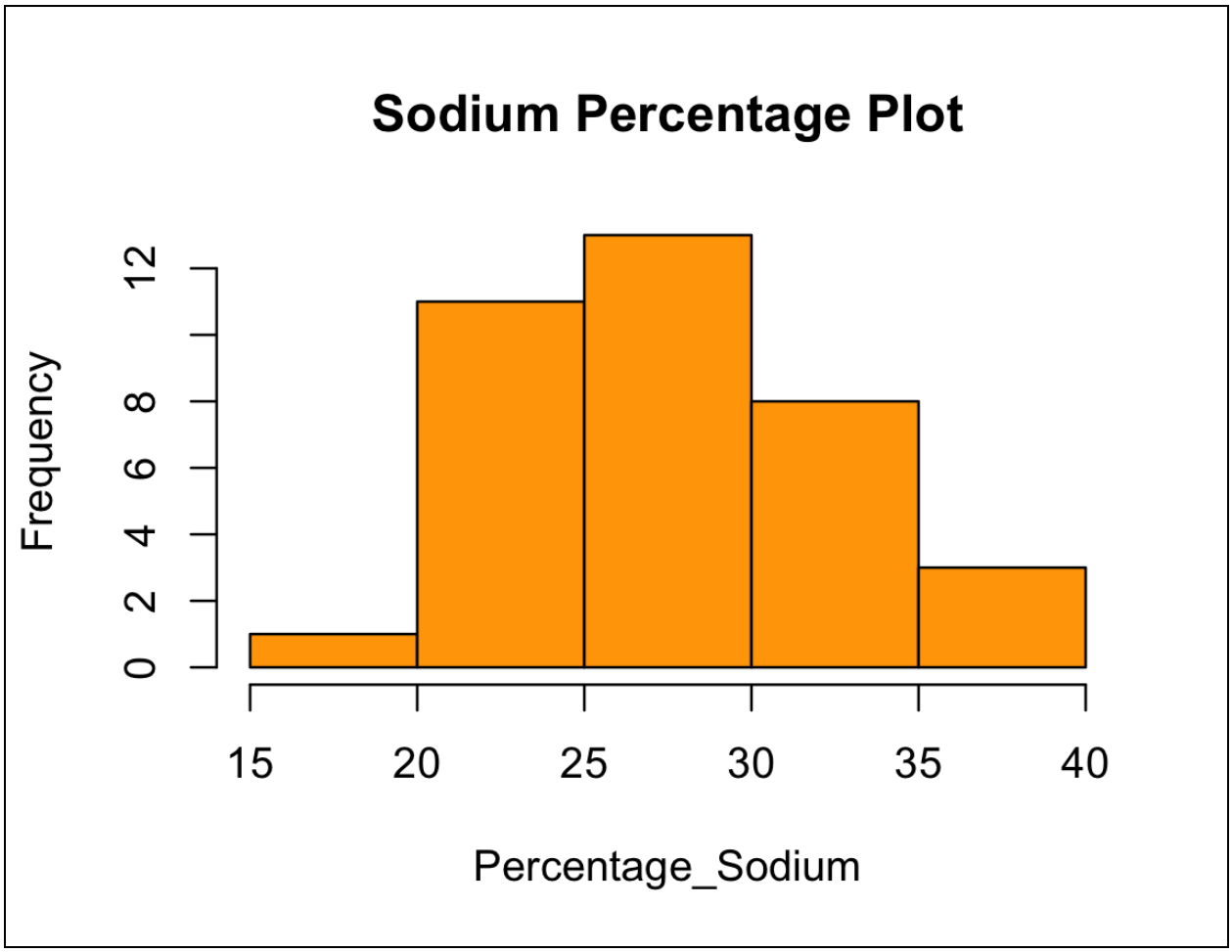
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points above value 2000 in 2006 and 2018).

Data Assignment 1

13. **Environmental Quality Indicator:** Percentage Sodium

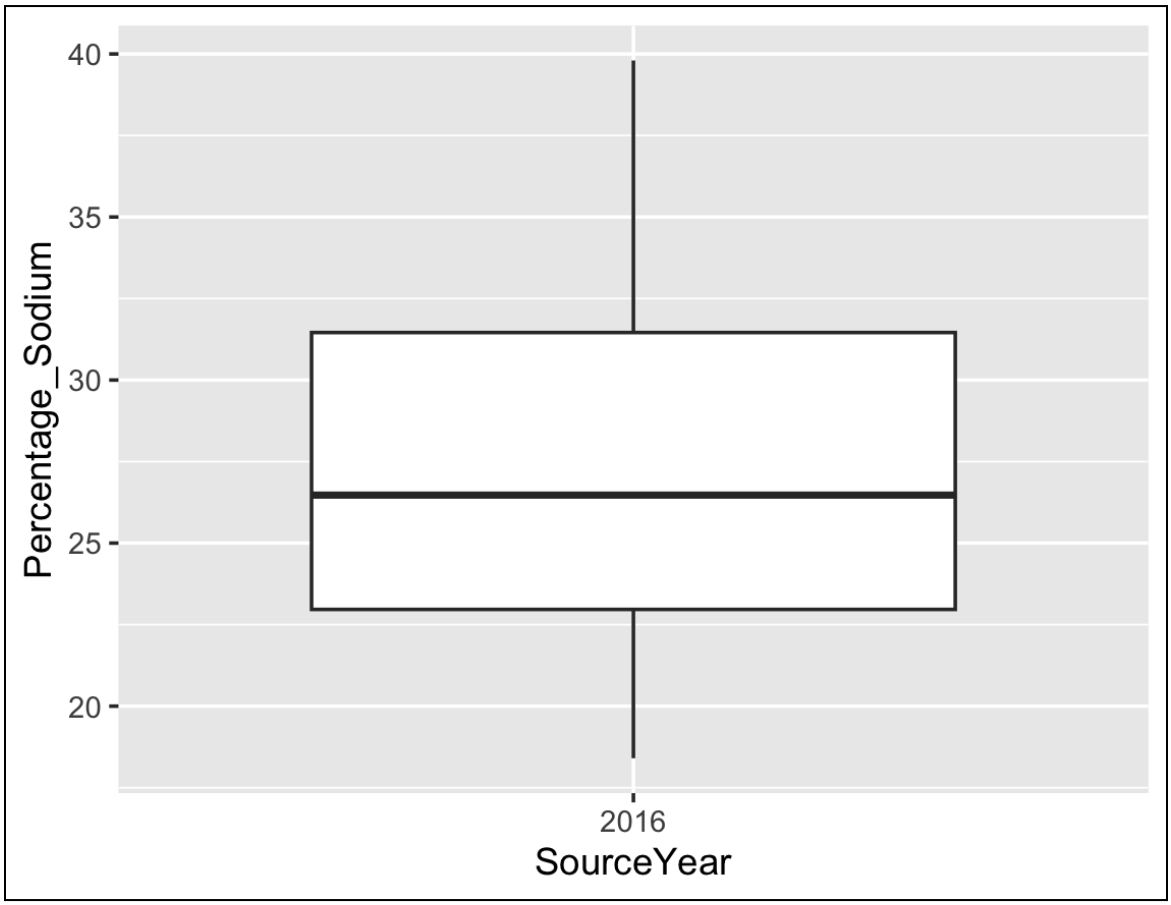
Statistic	N	Mean	St. Dev.	Min	Max
Percentage Sodium	36	27.493	5.394	18.403	39.800

Histogram:



Data Assignment 1

Box Plot:



Skew: 0.5476125

Shape of Distribution: Since the coefficient of skewness is slightly greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Although the data is distributed fairly normally, most of the values are on the left side of the graph.

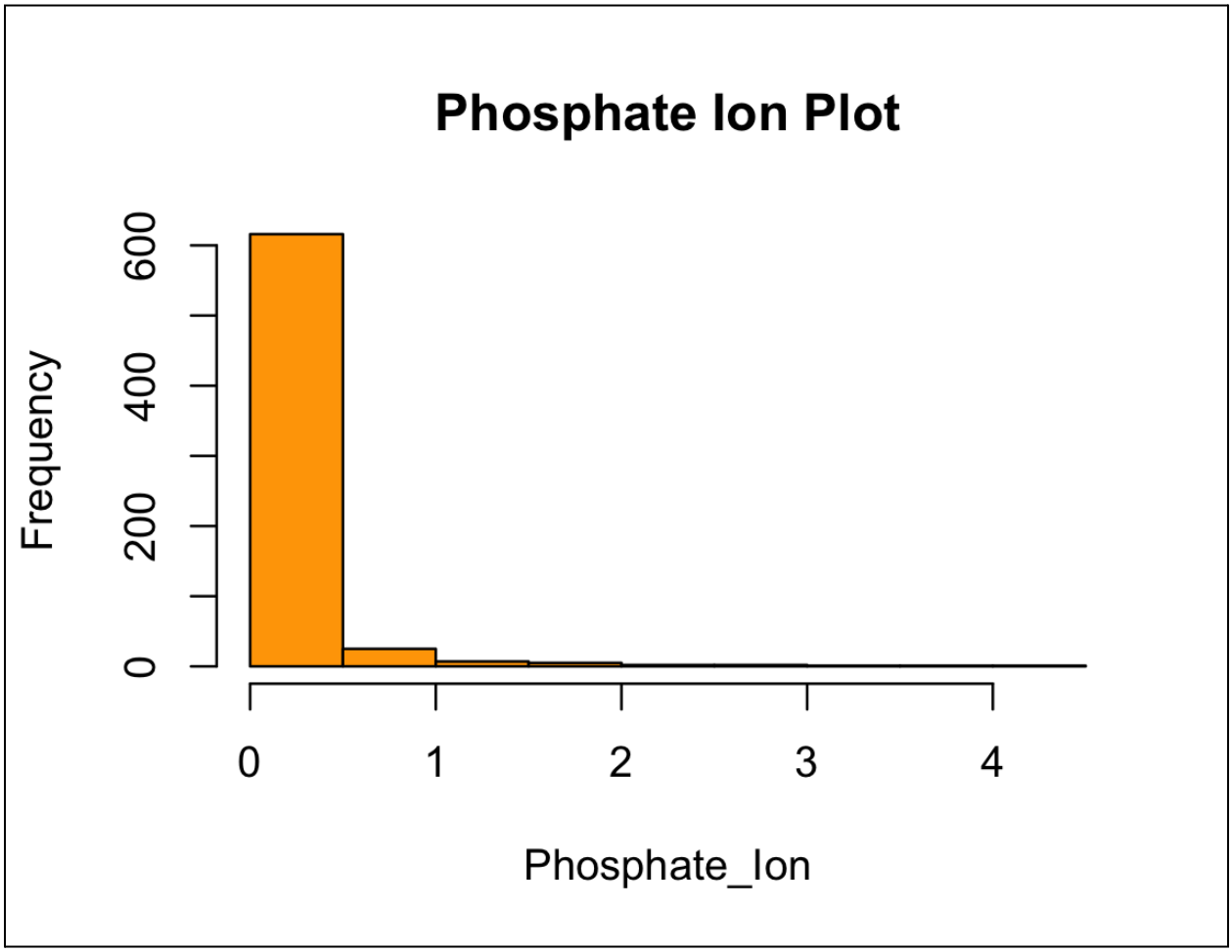
Outliers: No, there aren't particularly any outliers visible in the box plot. The distribution of the histogram also takes a normal shape and there aren't any outliers.

Data Assignment 1

14. **Environmental Quality Indicator:** Phosphate Ion

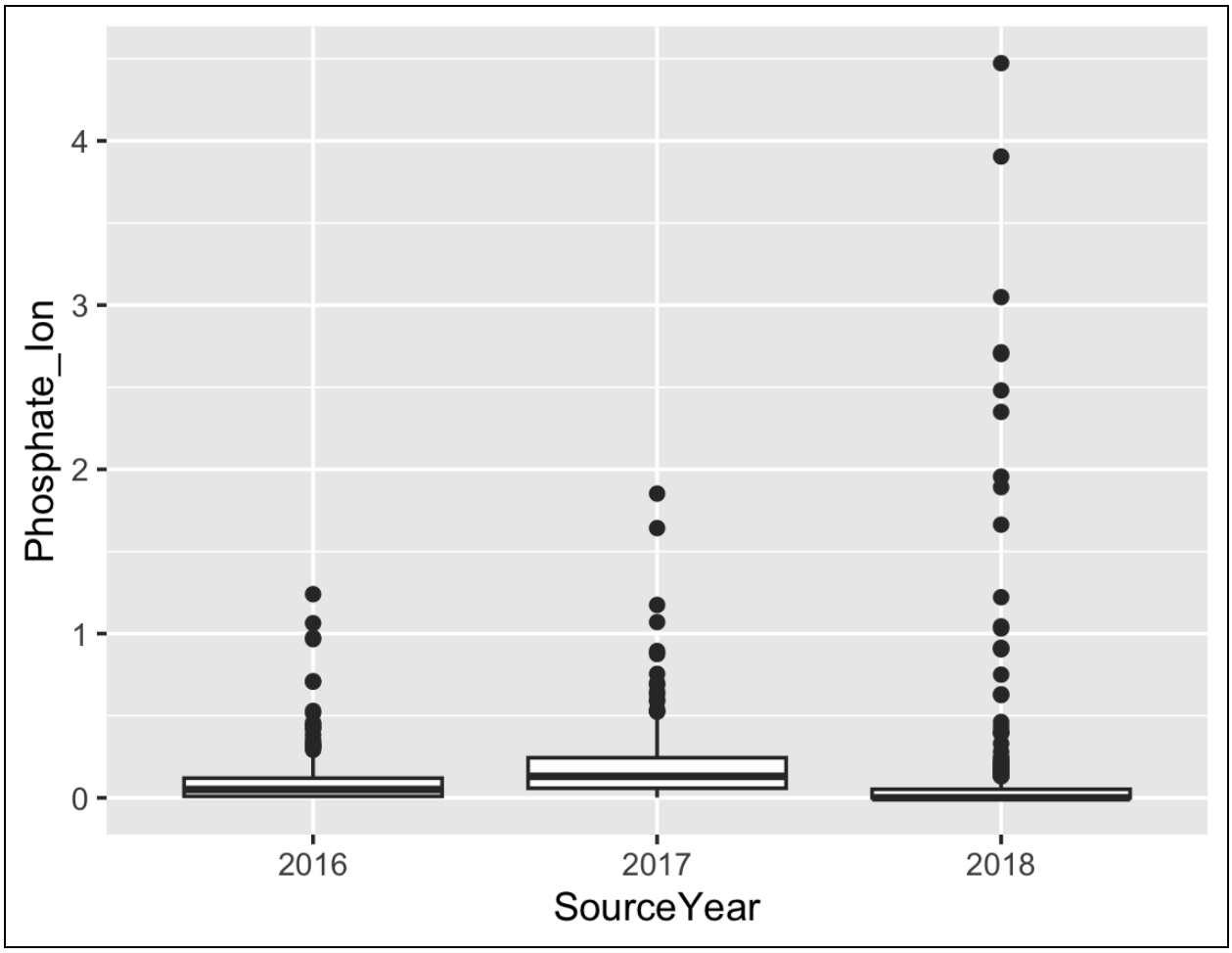
Statistic	N	Mean	St. Dev.	Min	Max
Phosphate Ion	660	0.147	0.396	0.000	4.473

Histogram:



Data Assignment 1

Box Plot:



Skew: 6.08999

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

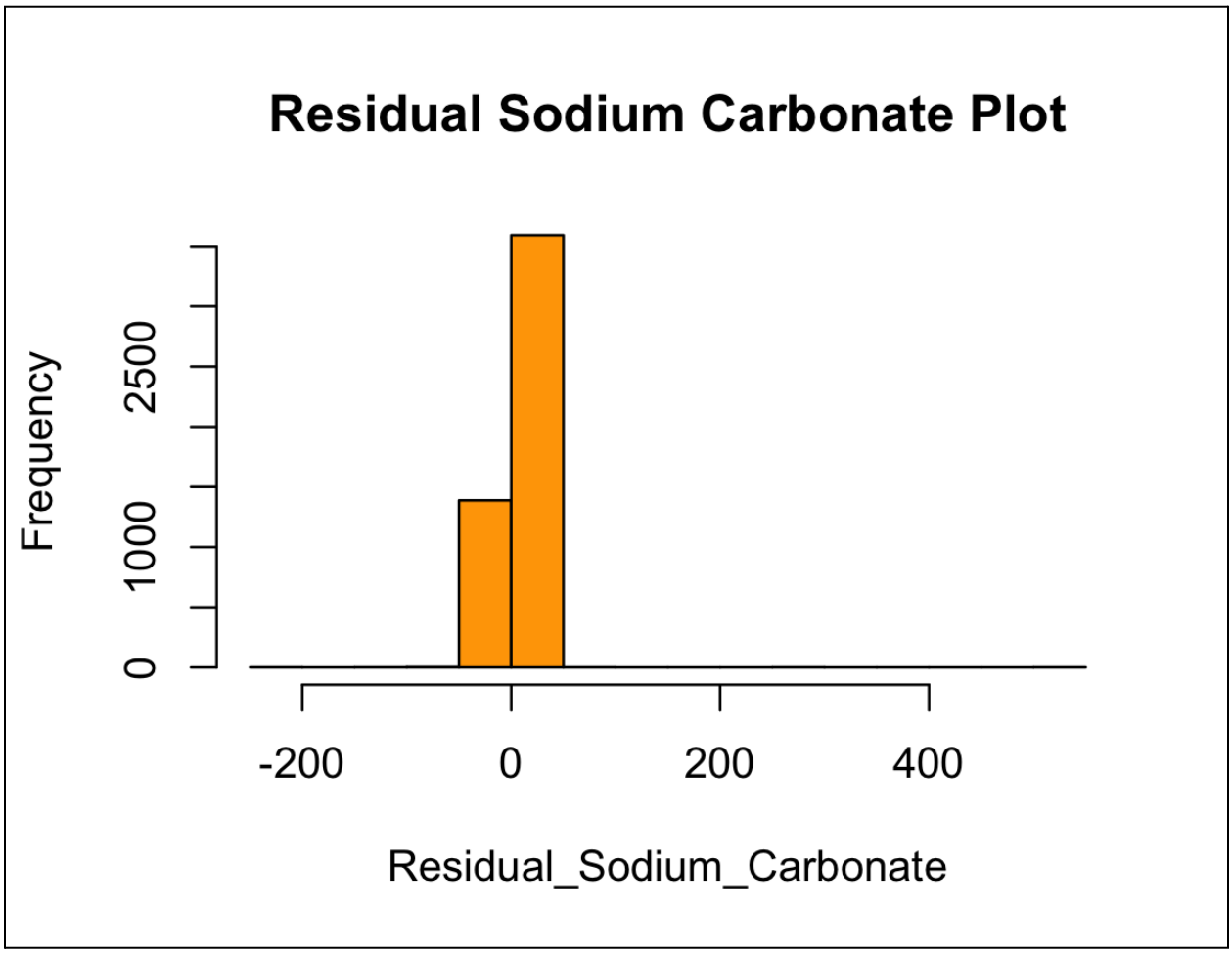
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points around/above value 4 in 2018).

Data Assignment 1

15. **Environmental Quality Indicator:** Residual Sodium Carbonate

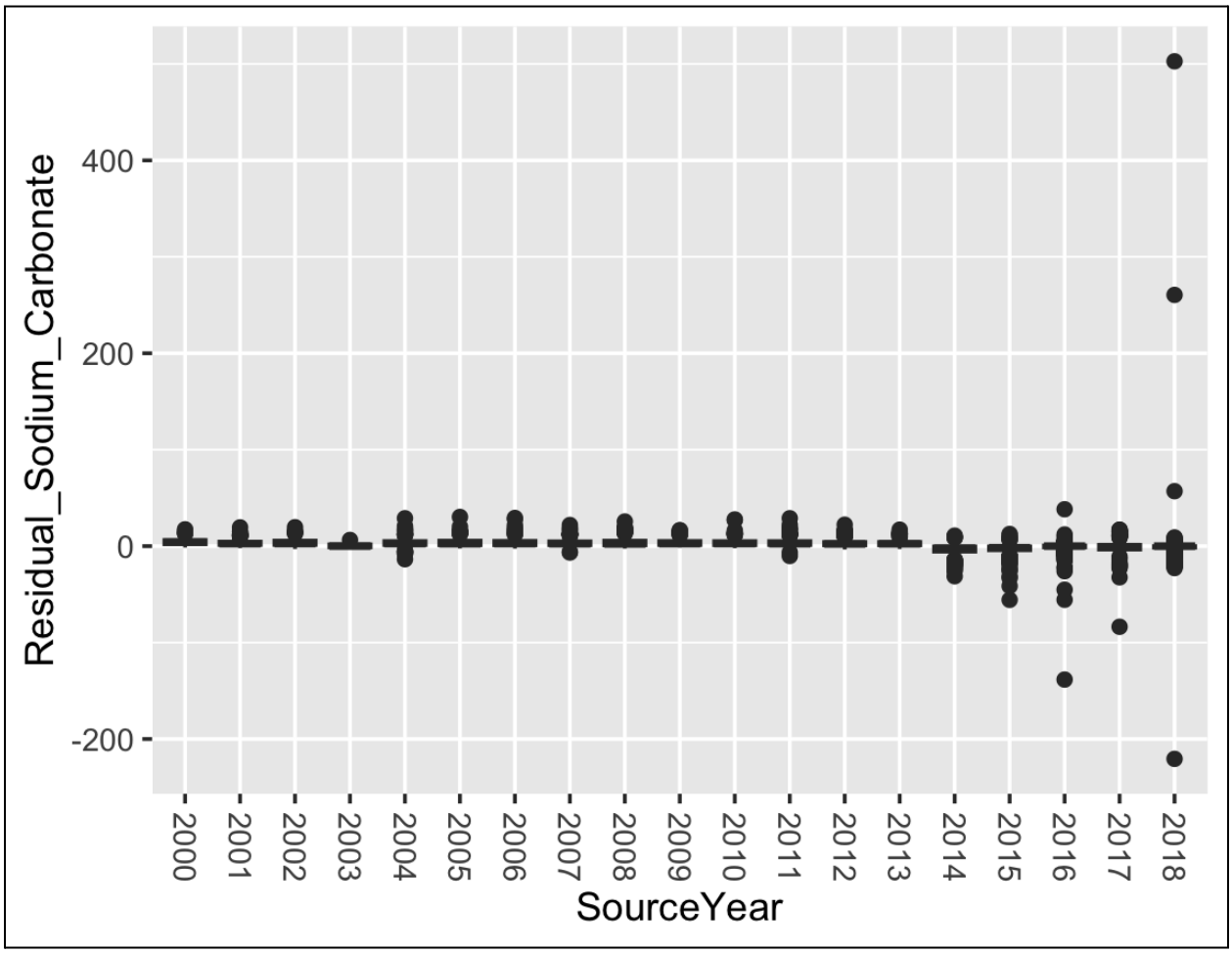
Statistic	N	Mean	St. Dev.	Min	Max
Residual Sodium Carbonate	4,988	1.976	10.401	-220.610	502.747

Histogram:



Data Assignment 1

Box Plot:



Skew: 22.77248

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

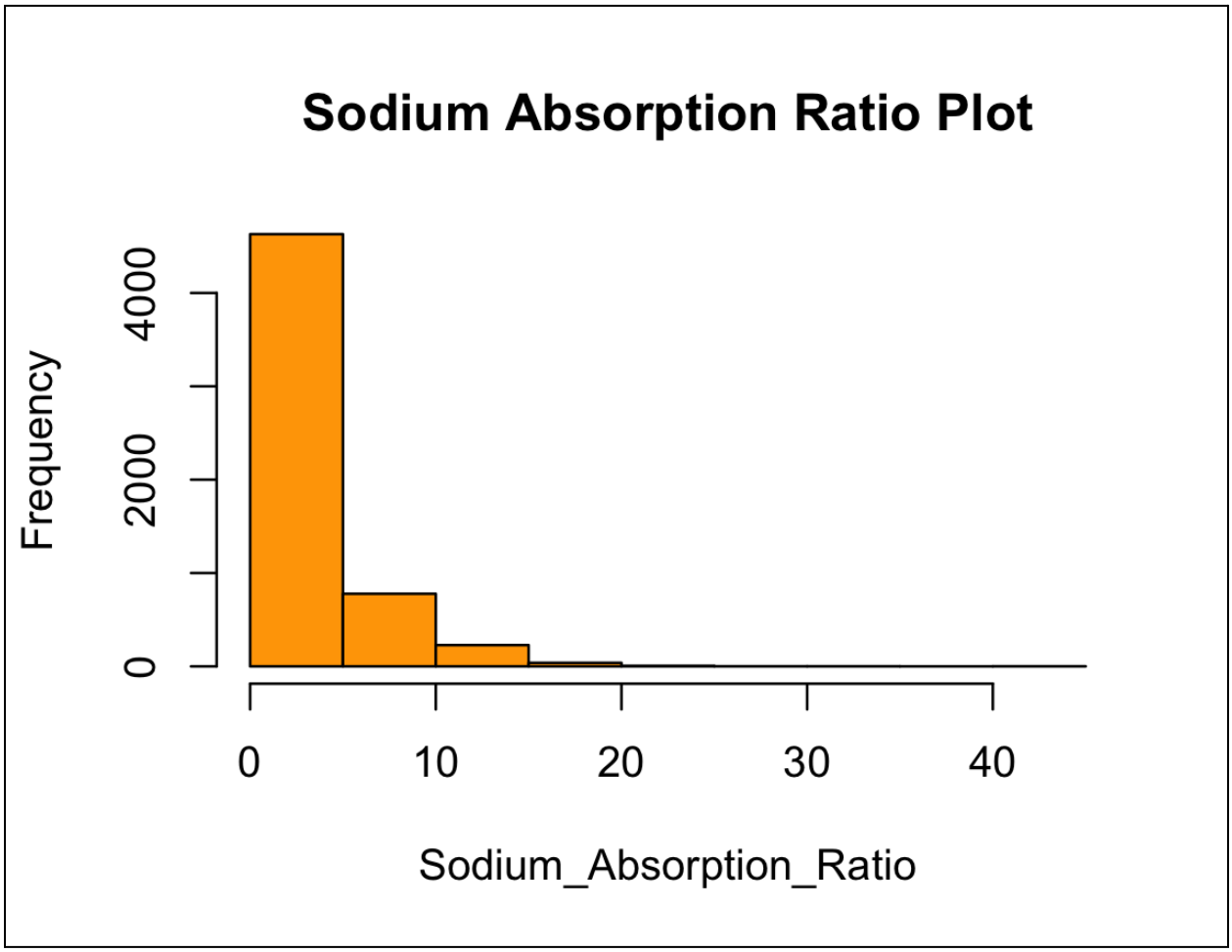
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points above value 200 in 2018).

Data Assignment 1

16. **Environmental Quality Indicator:** Sodium Absorption Ratio

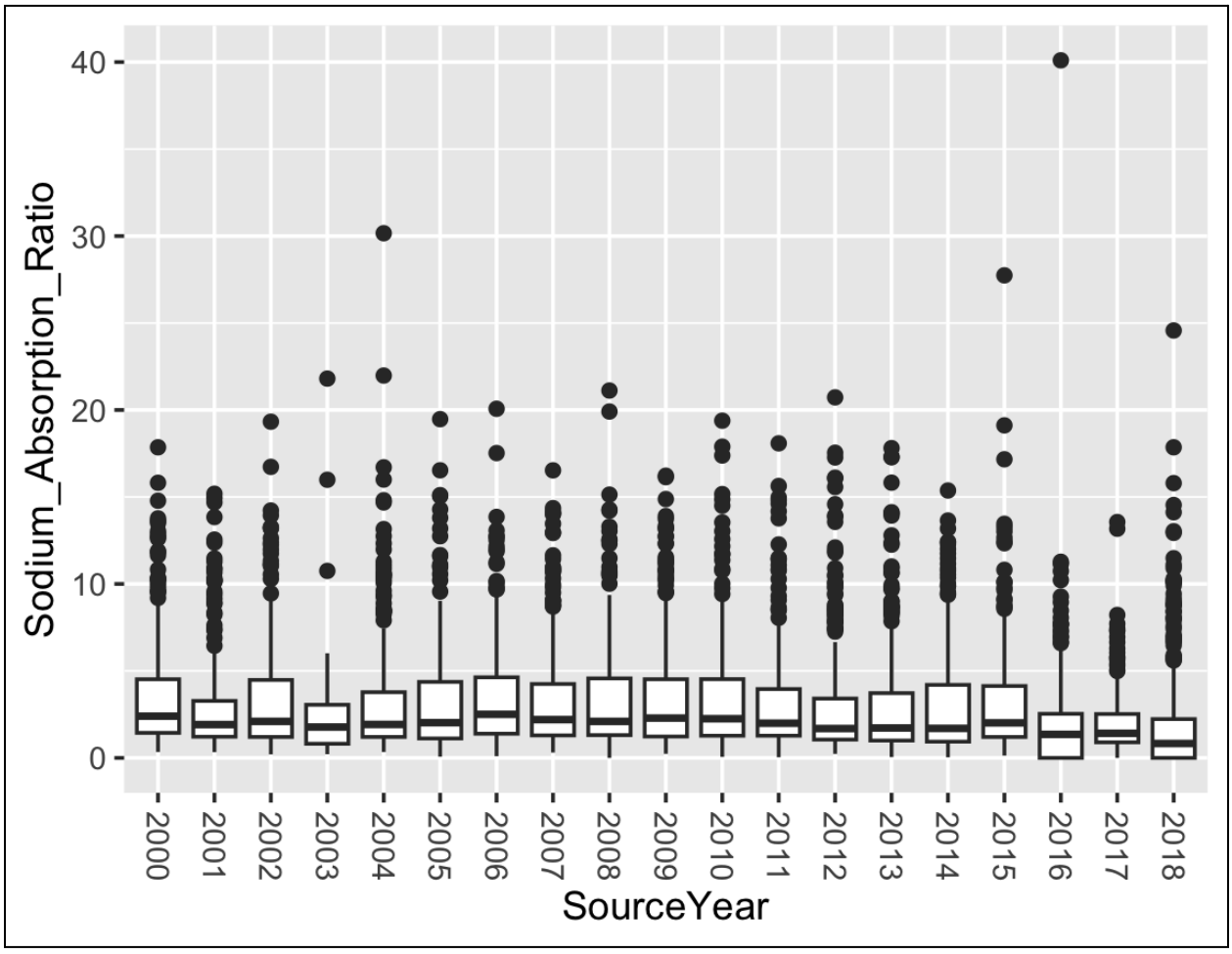
Statistic	N	Mean	St. Dev.	Min	Max
Sodium Absorption Ratio	5,683	3.062	3.178	0.000	40.100

Histogram:



Data Assignment 1

Box Plot:



Skew: 2.379569

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

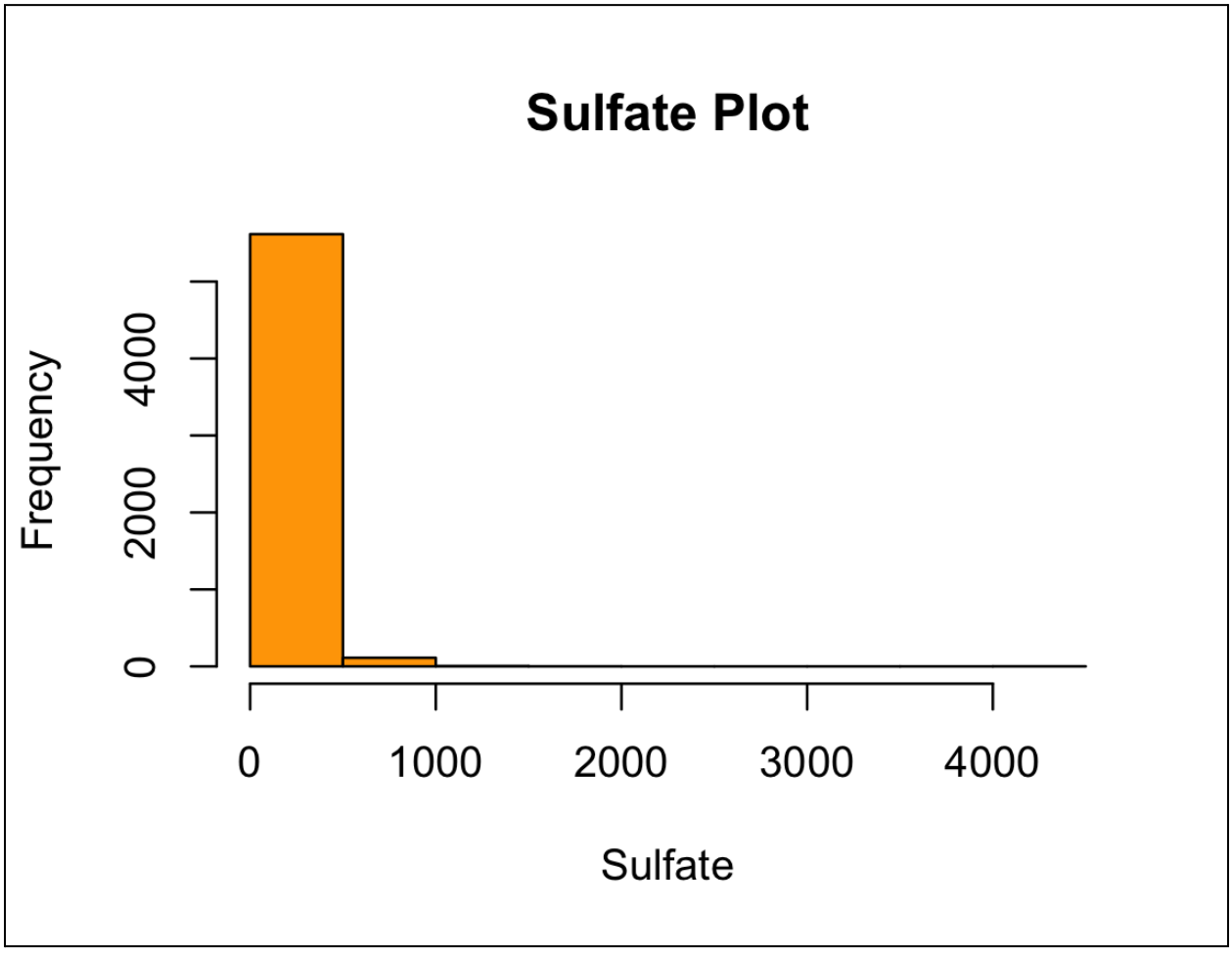
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points above value 30 in 2004 and 2016).

Data Assignment 1

17. **Environmental Quality Indicator:** Sulfate

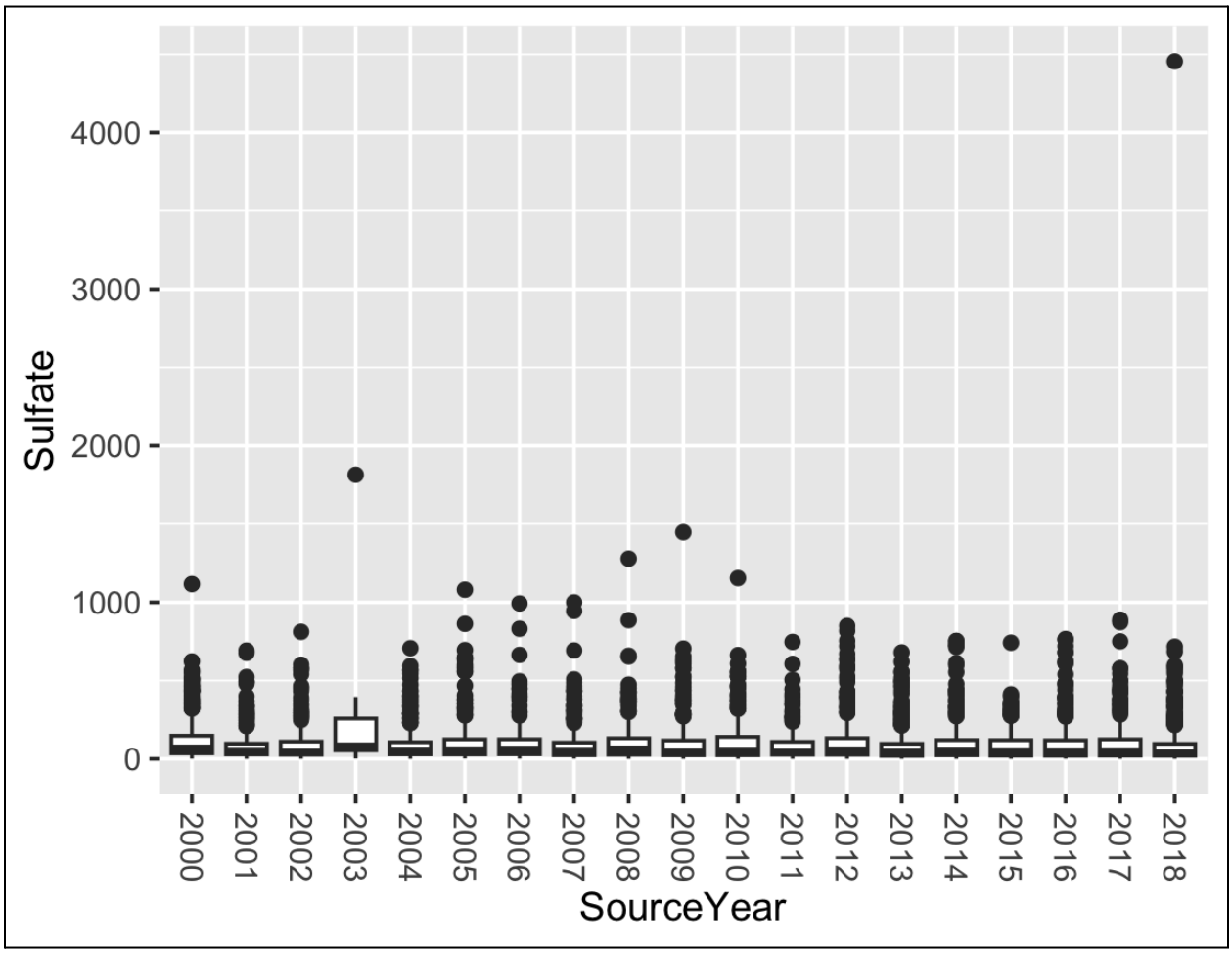
Statistic	N	Mean	St. Dev.	Min	Max
Sulfate	5,735	98.780	140.971	0.000	4,455.333

Histogram:



Data Assignment 1

Box Plot:



Skew: 7.606018

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

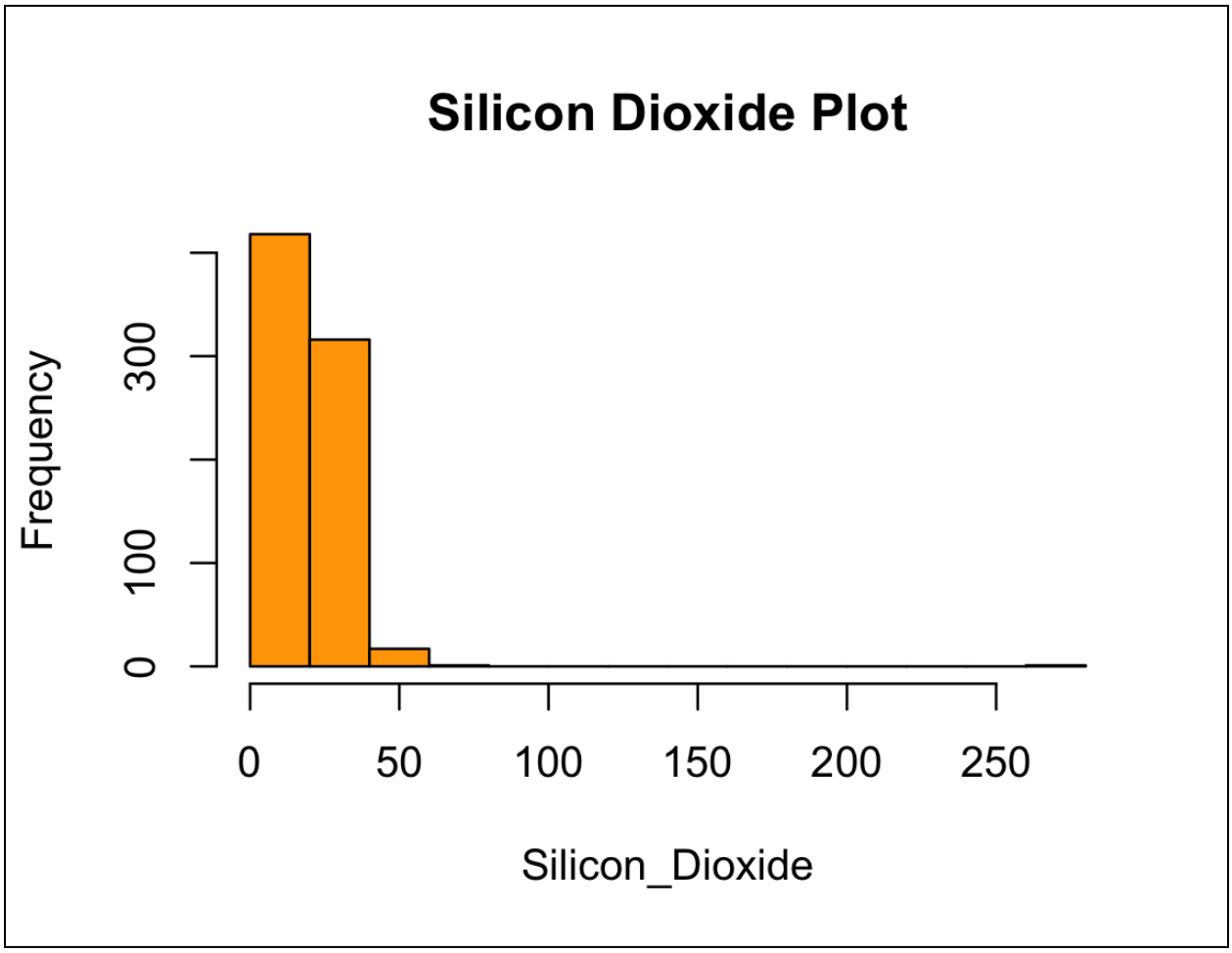
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around/above value 2000 in 2003 and 2018).

Data Assignment 1

18. **Environmental Quality Indicator:** Silicon Dioxide

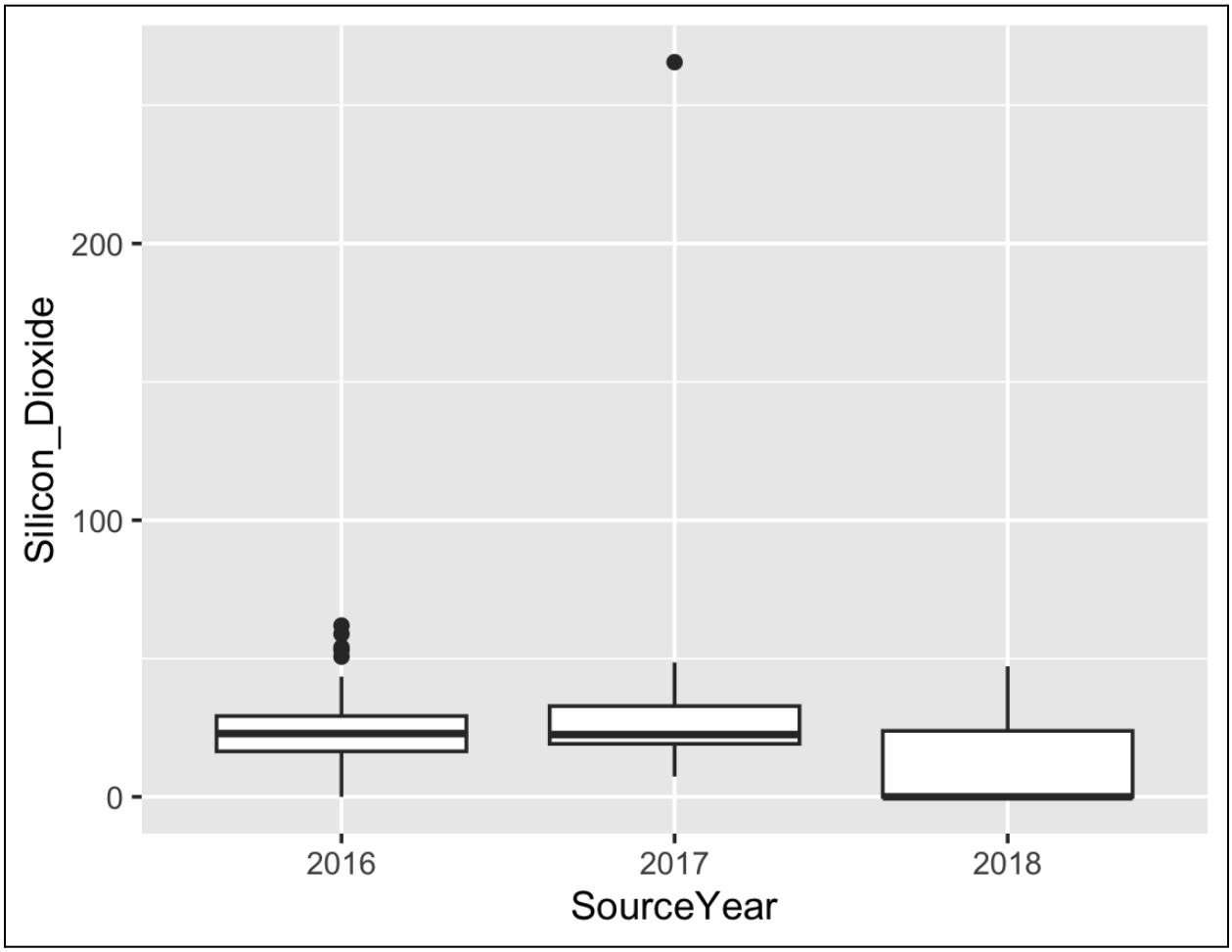
Statistic	N	Mean	St. Dev.	Min	Max
Silicon Dioxide	753	16.075	16.808	0.000	265.556

Histogram:



Box Plot:

Data Assignment 1



Skew: 4.462109

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

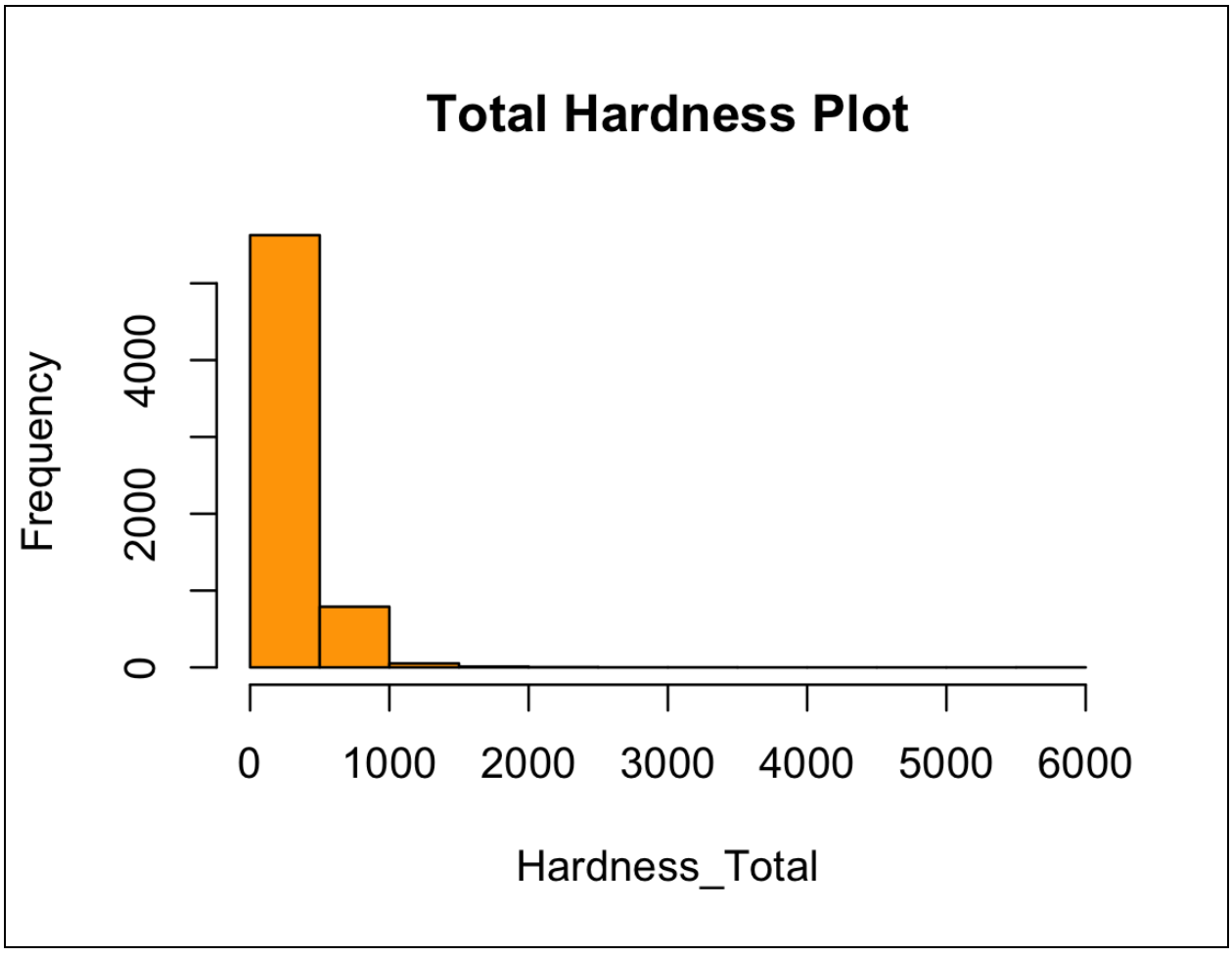
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point above value 200 in 2017).

19. **Environmental Quality Indicator:** Total Hardness

Data Assignment 1

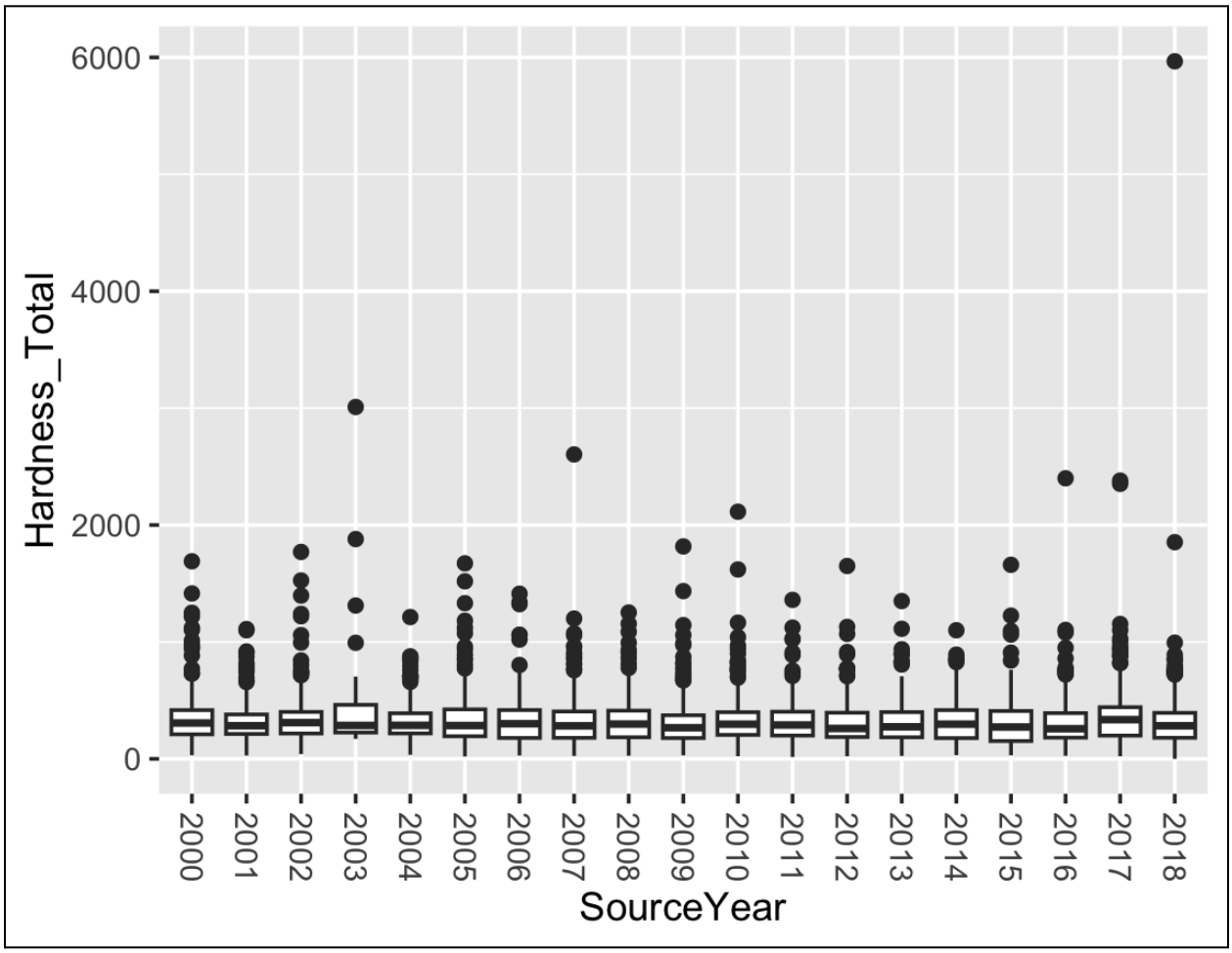
Statistic	N	Mean	St. Dev.	Min	Max
Total Hardness	6,487	319.834	216.228	0.000	5,966.667

Histogram:



Box Plot:

Data Assignment 1



Skew: 4.974096

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

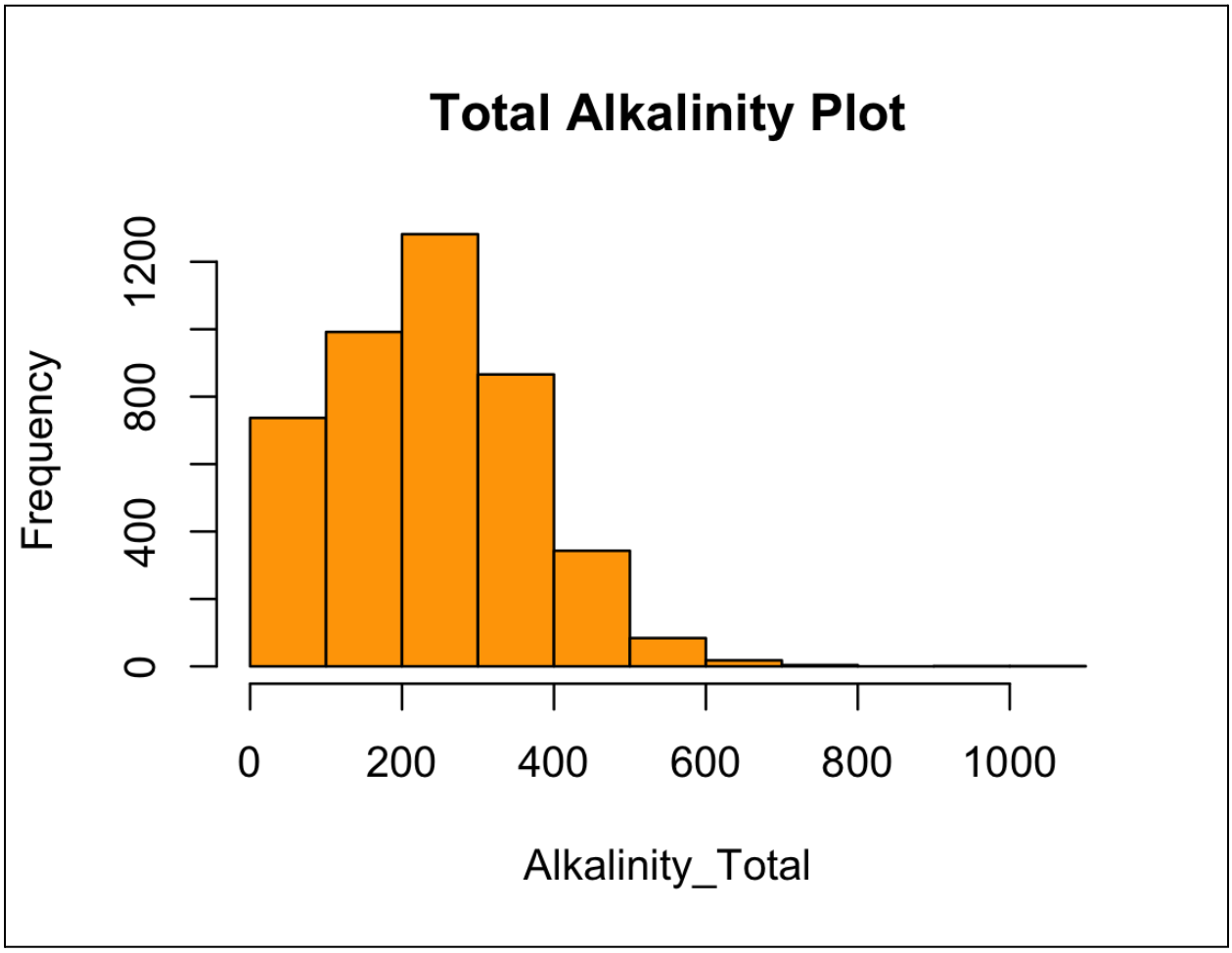
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point above/around value 3000 in 2003 and 2018).

20. **Environmental Quality Indicator:** Total Alkalinity

Data Assignment 1

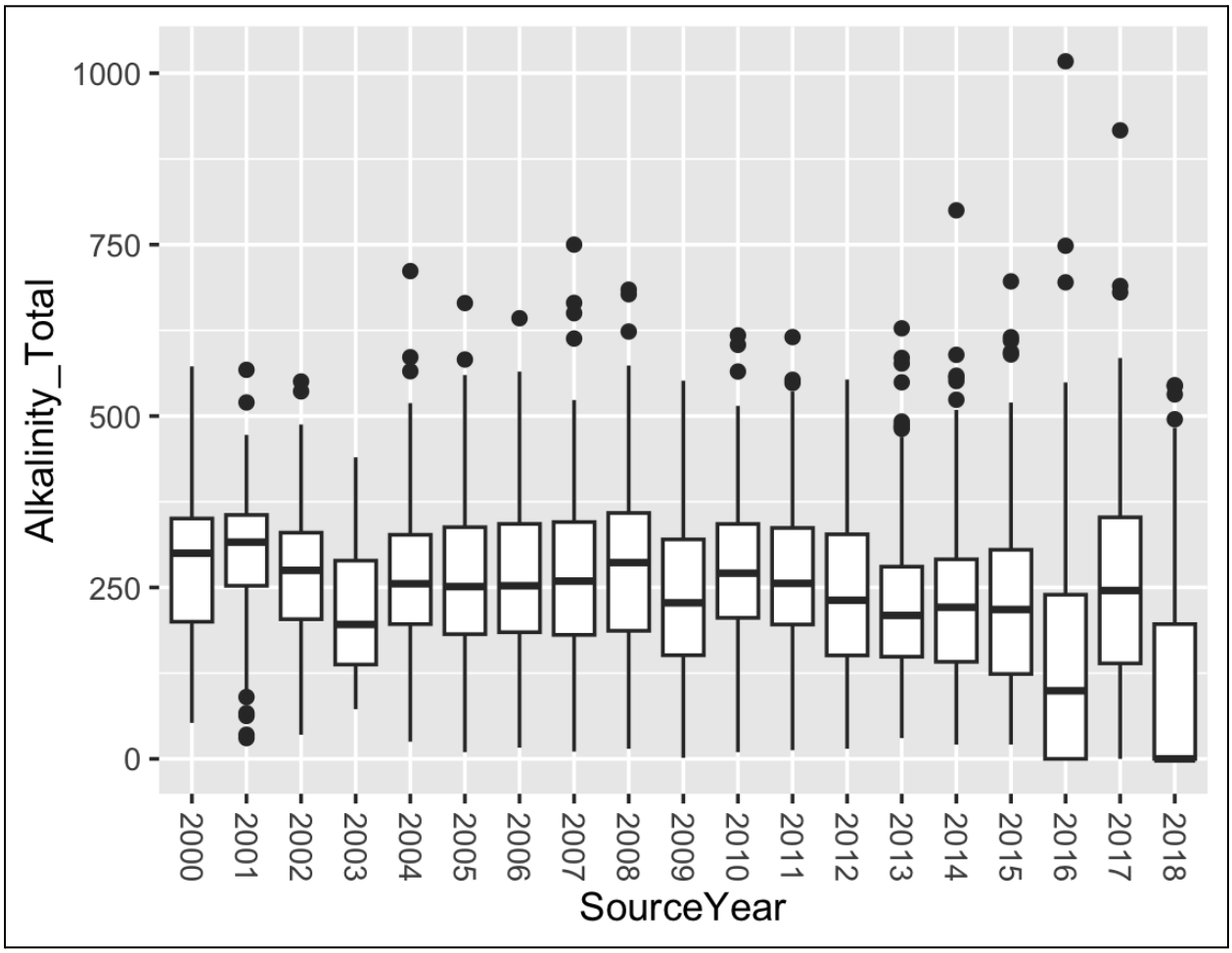
Statistic	N	Mean	St. Dev.	Min	Max
Total Alkalinity	4,328	234.223	133.405	0.000	1,017.350

Histogram:



Box Plot:

Data Assignment 1



Skew: 0.3058979

Shape of Distribution: Since the coefficient of skewness is slightly greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Although the data is distributed fairly normally, majority of the values are on the left side of the graph.

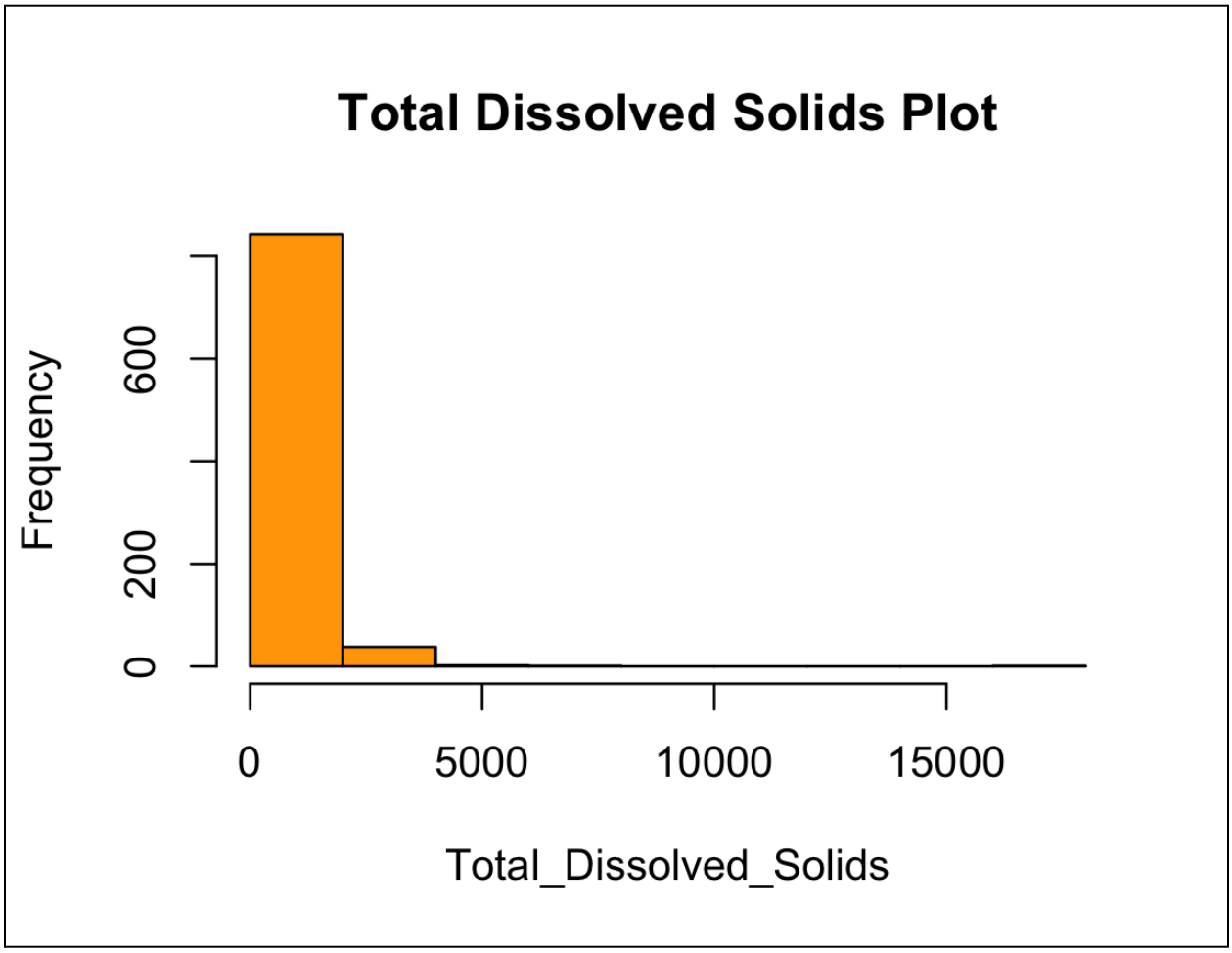
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 1000 in 2016).

21. **Environmental Quality Indicator:** Total Dissolved Solids

Data Assignment 1

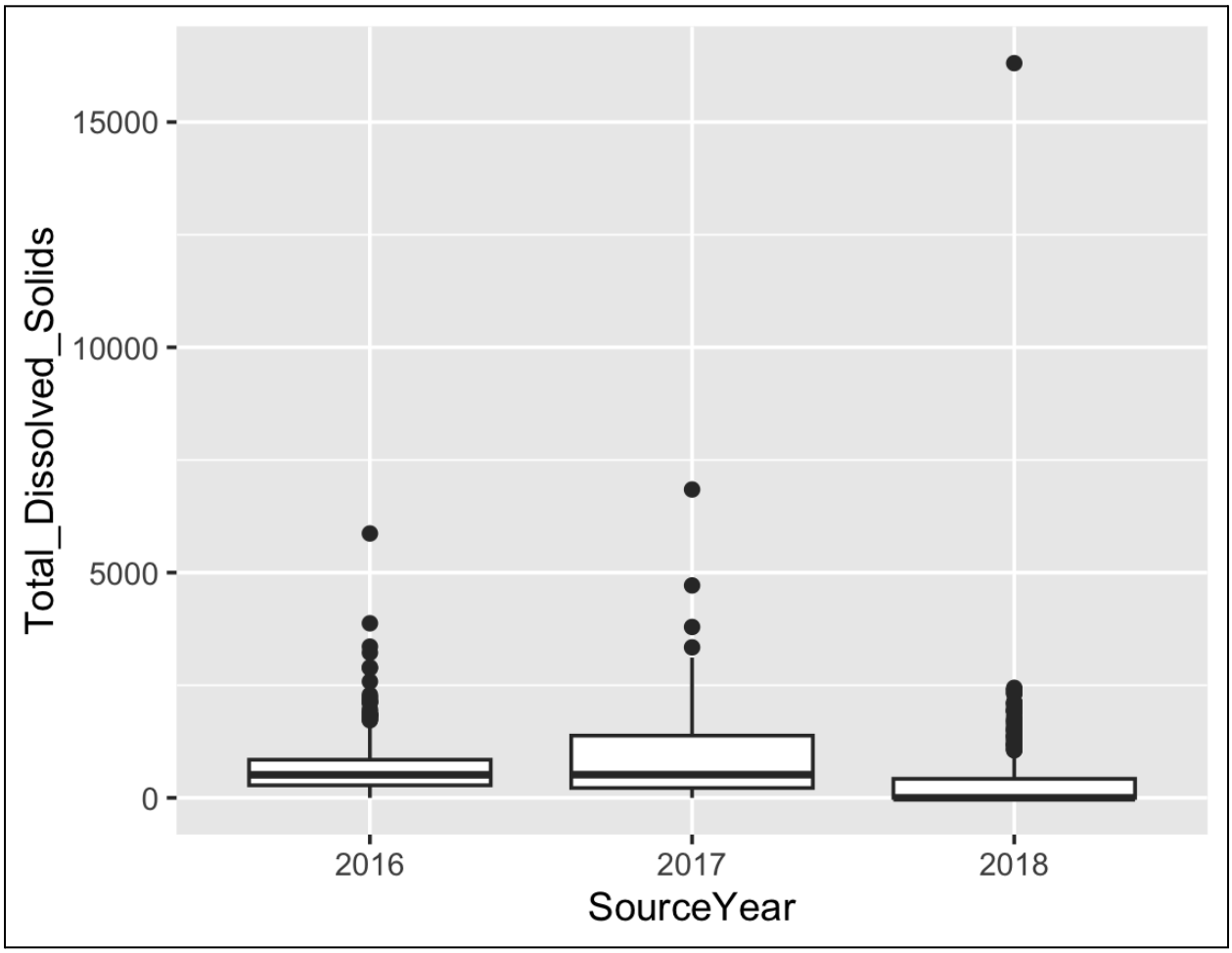
Statistic	N	Mean	St. Dev.	Min	Max
Total Dissolved Solids	885	562.370	896.116	0.000	16,307.670

Histogram:



Box Plot:

Data Assignment 1



Skew: 7.481956

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

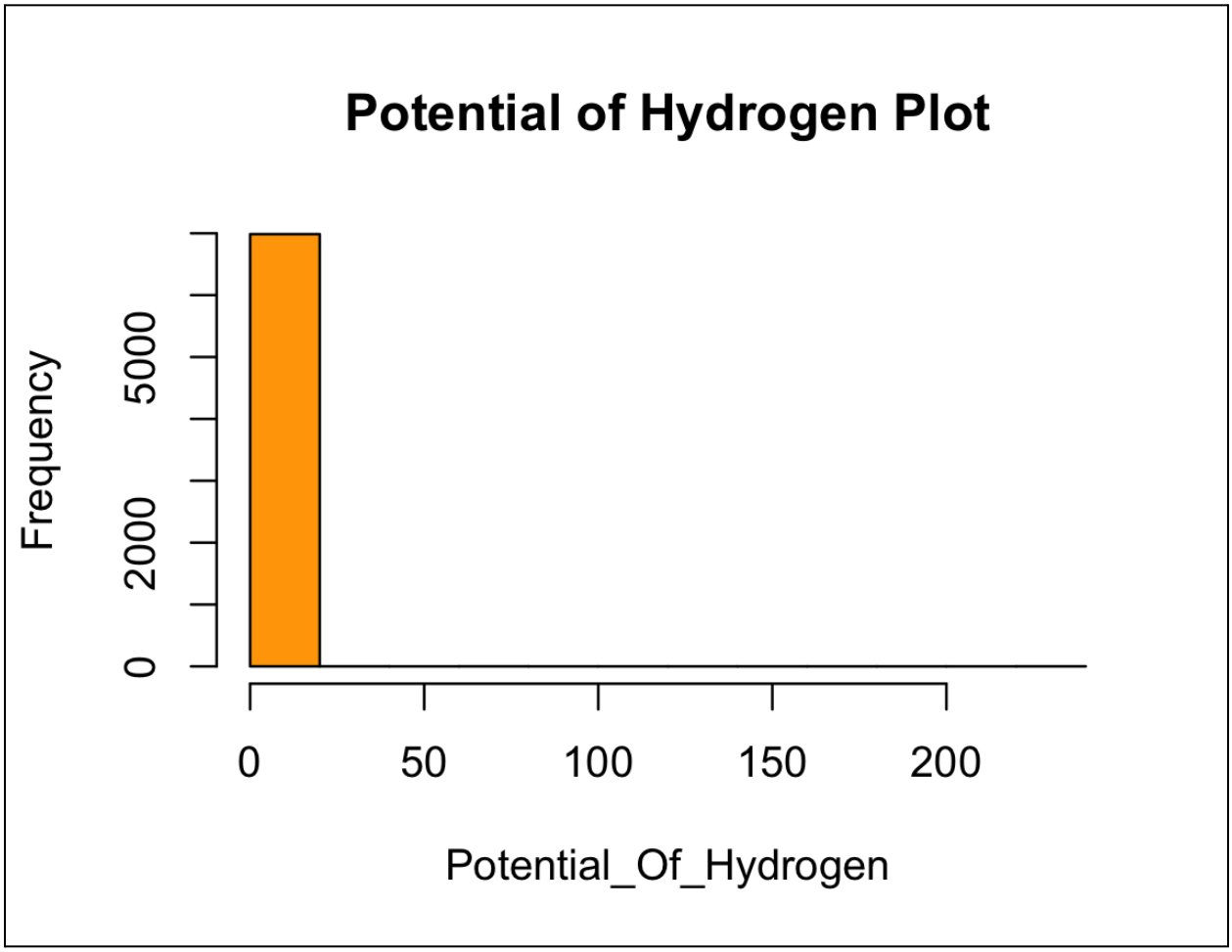
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point above value 15000 in 2018).

22. **Environmental Quality Indicator:** Potential of Hydrogen

Data Assignment 1

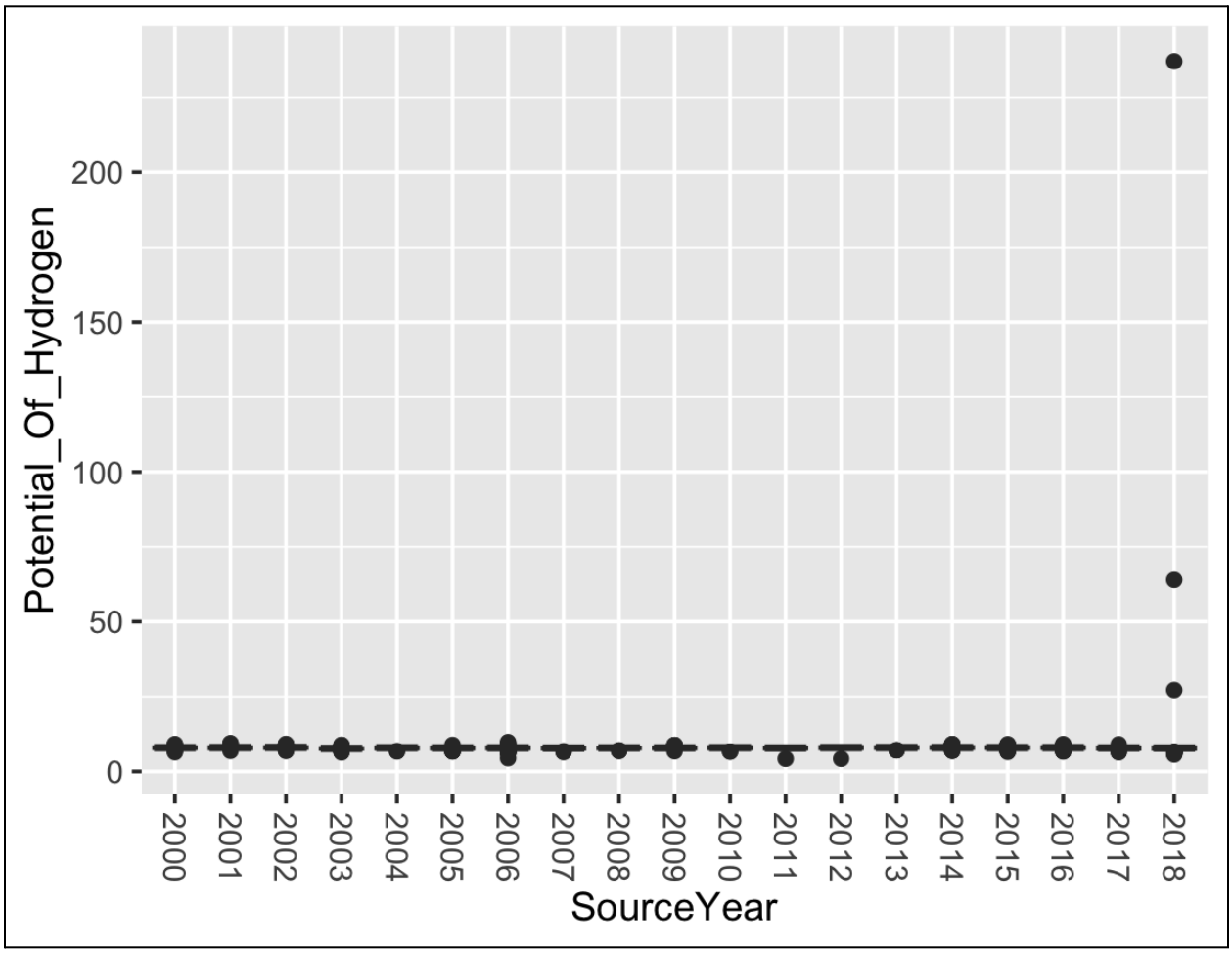
Statistic	N	Mean	St. Dev.	Min	Max
Potential of Hydrogen	6,989	7.898	2.857	4.200	237.060

Histogram:



Box Plot:

Data Assignment 1



Skew: 74.97018

Shape of Distribution: Since the coefficient of skewness is much much greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are heavily concentrated on the left side of the graph.

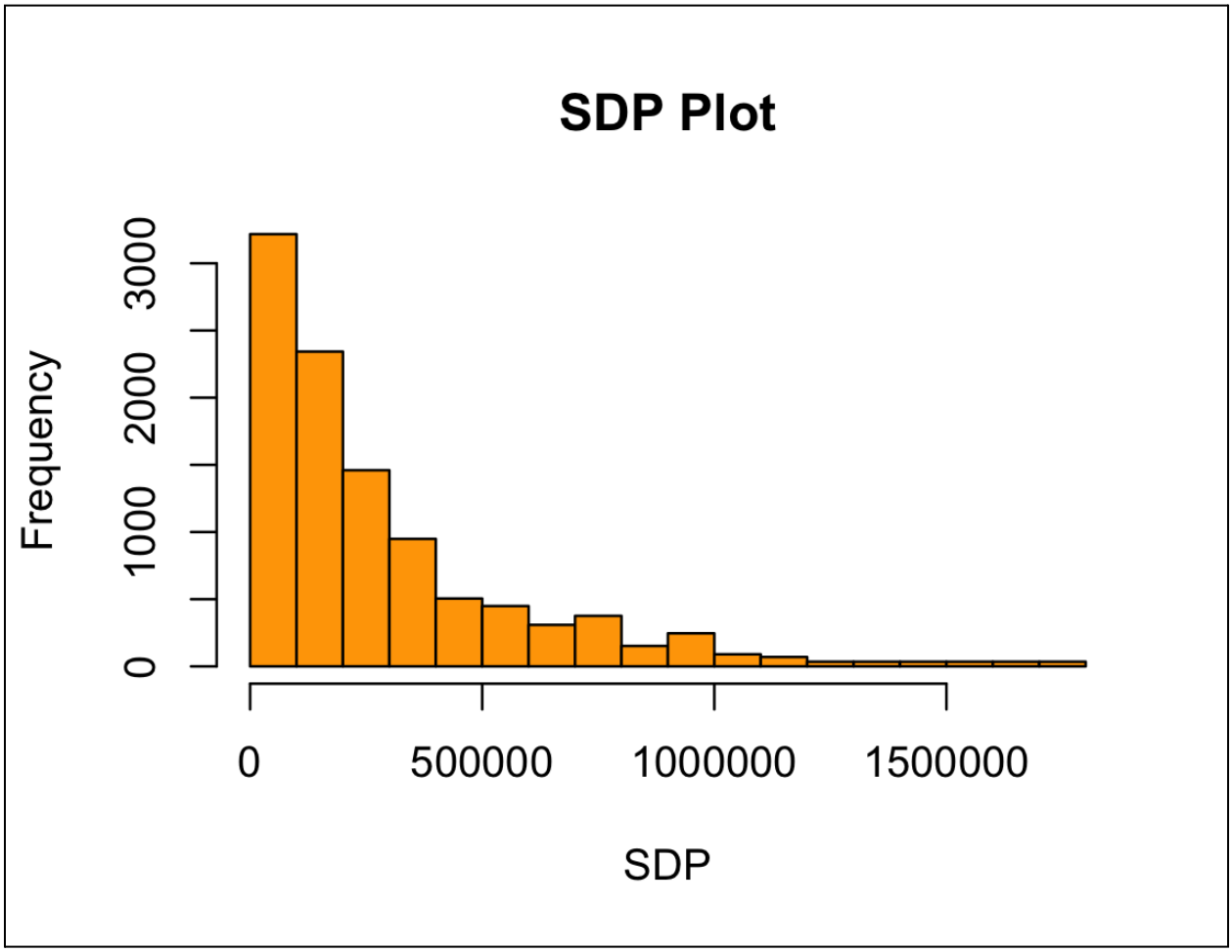
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data points above value 50 in 2018).

1. **Quantity:** SDP

Data Assignment 1

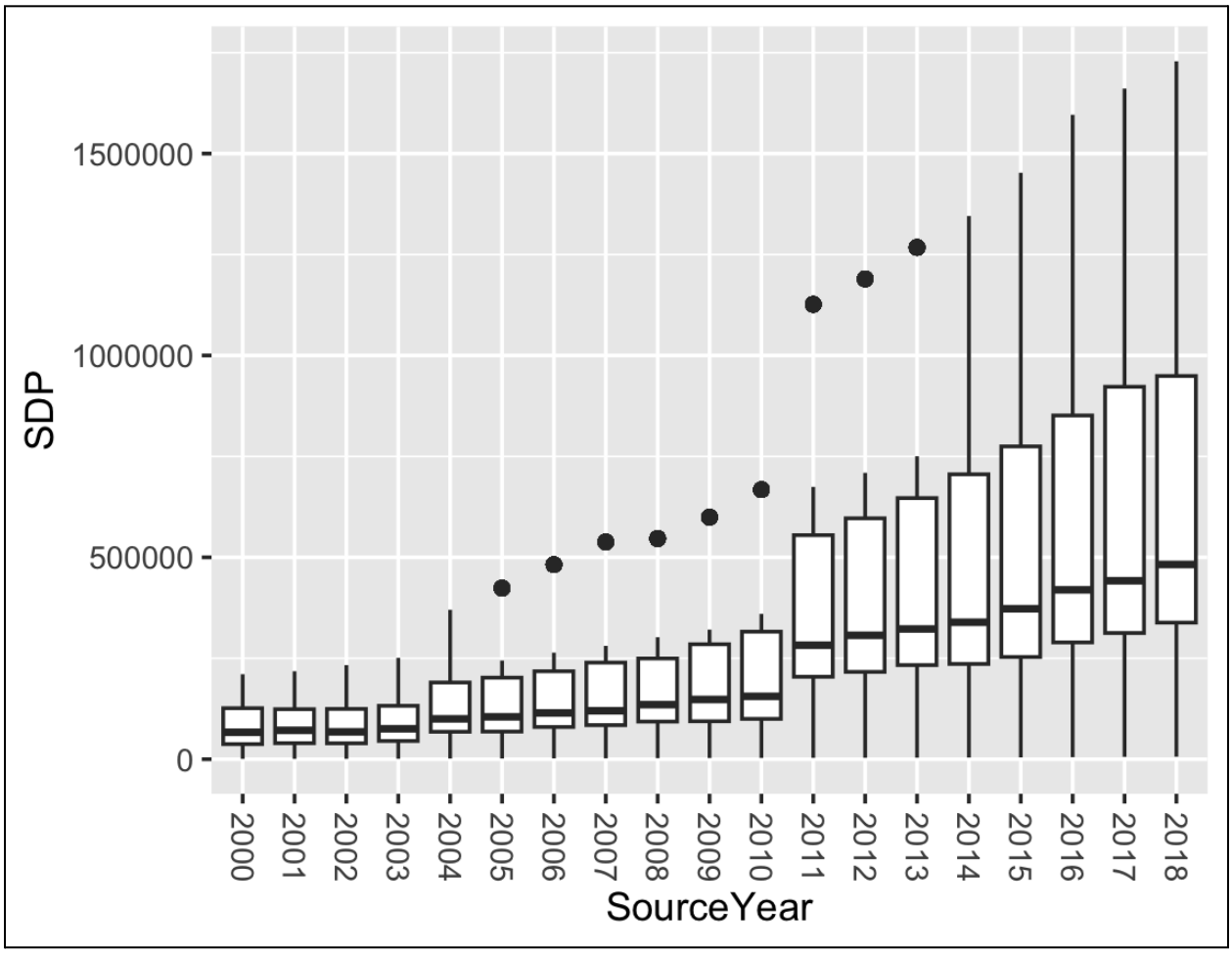
Statistic	N	Mean	St. Dev.	Min	Max
SDP	10,376	289,312.200	307,351.000	838	1,728,578

Histogram:



Box Plot:

Data Assignment 1



Skew: 1.966974

Shape of Distribution: Since the coefficient of skewness is greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Most of the values are concentrated on the left side of the graph.

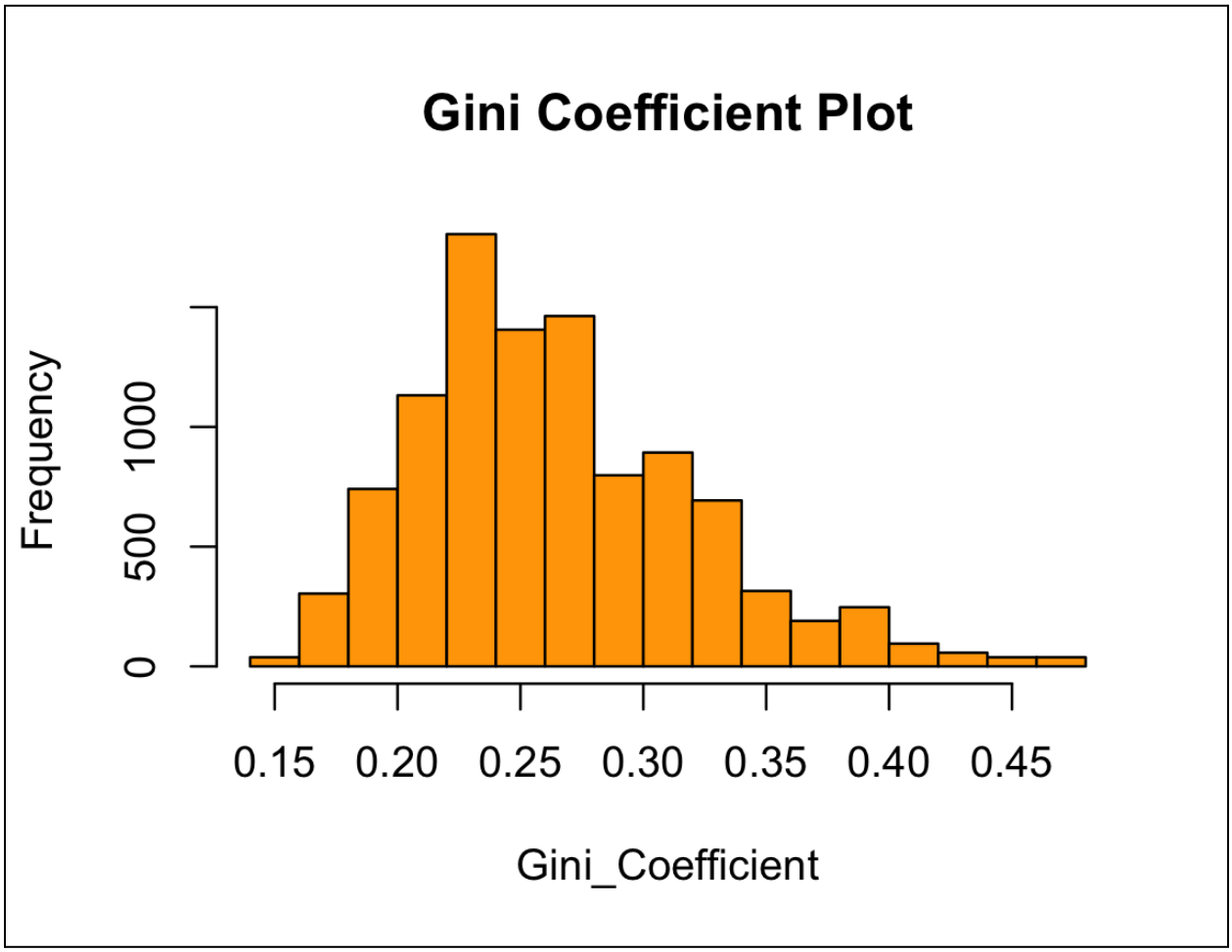
Outliers: Yes, there are some outliers quite visible in the box plot (for example, the data point around value 1250000 in 2013).

2. **Quantity:** Gini Coefficient

Data Assignment 1

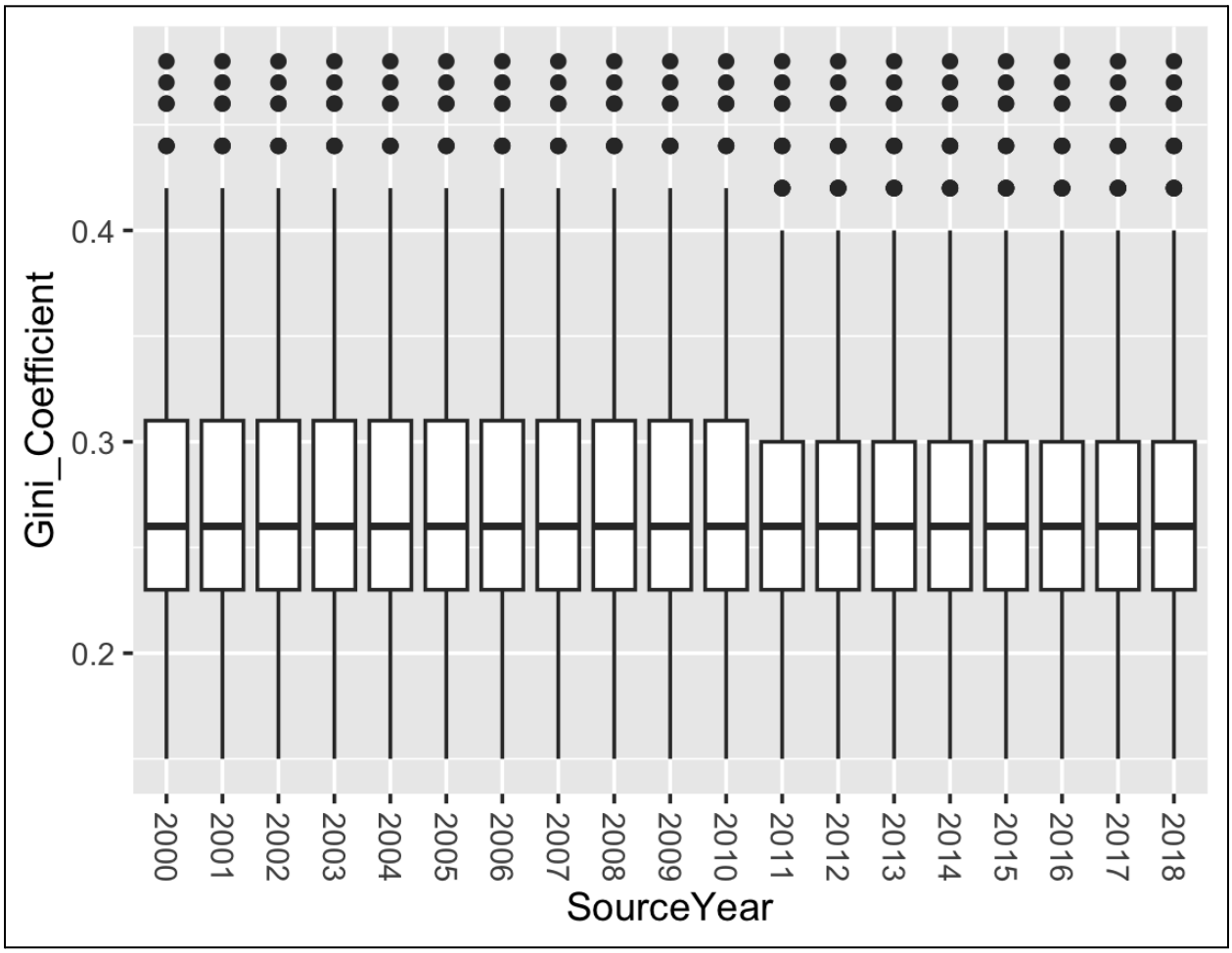
Statistic	N	Mean	St. Dev.	Min	Max
Gini Coefficient	10,253	0.269	0.057	0.150	0.480

Histogram:



Box Plot:

Data Assignment 1



Skew: 0.7797086

Shape of Distribution: Since the coefficient of skewness is slightly greater than 0, the graph is said to be positively skewed with the majority of data values less than mean.
Although the data is distributed fairly normally, majority of the values are on the left side of the graph.

Outliers: Yes, there are some outliers quite visible in the box plot.

Data Assignment 1

Q6.

Environmental Quality Indicator: Calcium

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

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
β_0	5.558e+01	6.588e-01	84.37	<2e-16
β_1	1.531e-05	1.469e-06	10.42	<2e-16

(where e+x = 10^x)

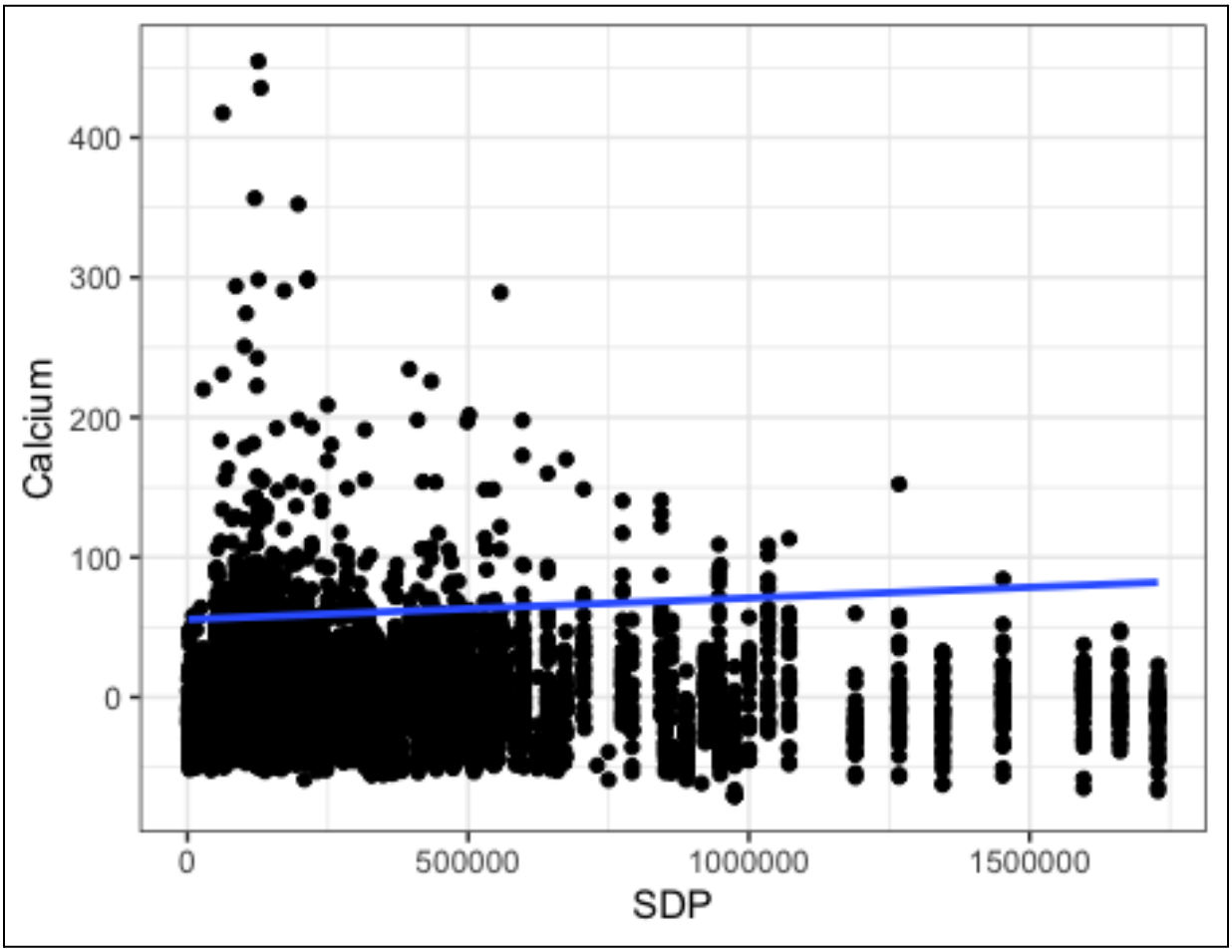
β_0 gives us the level of calcium in a district which has SDP = 0, that is, the expected level of calcium in groundwater in a district whose state has an SDP of Rs. 0 Cr, the level of calcium in the district's groundwater is expected to be 5.558e+01.

β_1 gives the expected increase in the EQI indicator (level of calcium in our case) in the groundwater of that district for an increase of Rs 1 Cr in the SDP value of the state which that the district belongs to. That is, if the SDP of a state increases by Rs 1 Cr, then the level of calcium in the district's groundwater will increase by 1.531e-05.

Data Assignment 1

Q7.

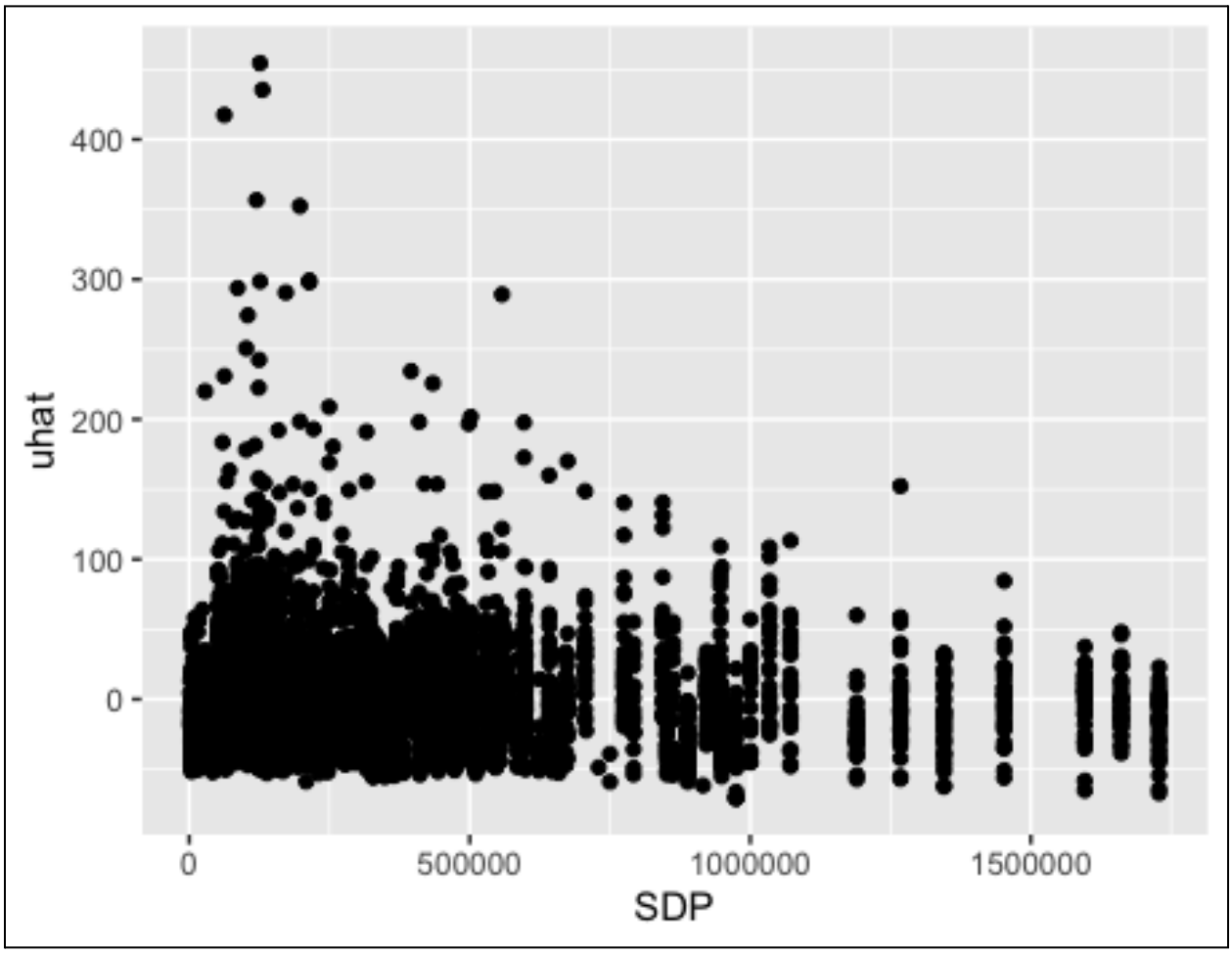
Plot 1:



Interpretation: The plot is shows how the residual values are distributed across the regression line. The closer the residuals are to the regression line, the better the line is able to estimate the predicted value of Caclcium given an SDP value.

Data Assignment 1

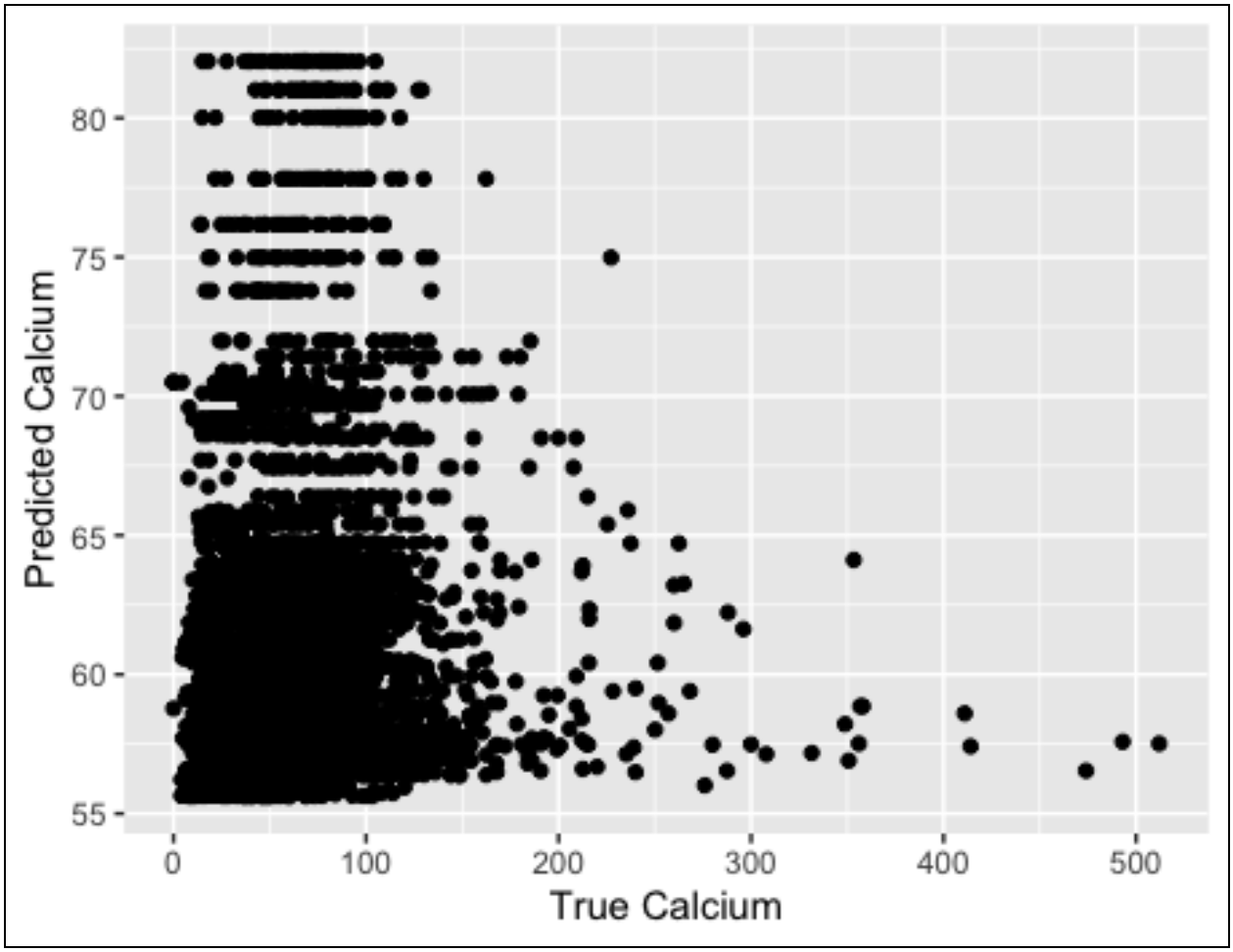
Plot 2:



Interpretation: This plot shows the relationship between uhat and SDP, that is, for a given SDP how well good are our predictions. It can help us see if the model is working well or as if no uhat is close to 0 then it means our predictions are not accurate.

Plot 3:

Data Assignment 1



Interpretation: This plot shows the relationship between the predicted values of calcium level and the true values of calcium levels. We can see if our model is a good predictor if the predicted value and the true value lie on the line y=x, that is, they are the same.

Relation between the Plot 1 and Plot 2: Plot 1 and Plot 2 are exactly the same except that we have plotted the regression line in Plot 1. This is obvious as our X-Axis is the same (SDP) and Y-Axis is also the same (level of Calcium).

Data Assignment 1

Relation between the Plot 2 and Plot 3:

If we notice, Plot 2 and Plot 3 have identical shapes when one is flipped across the line y = x. Therefore, we can establish a relationship between uhat, SDP, predicted_calcium and true_calcium.

First we add the most negative value of uhat to the second plot in order to ensure that the values are all above 0.

Then we can use the relationship

SDP/(uhat+mostnegative(uhat)) = K*(predicted_calcium/true_calcium)

We can find the value of K by trying some examples. We need this constant K since clearly the values aren't exactly the same and are instead scaled between Plot 2 and Plot 3.

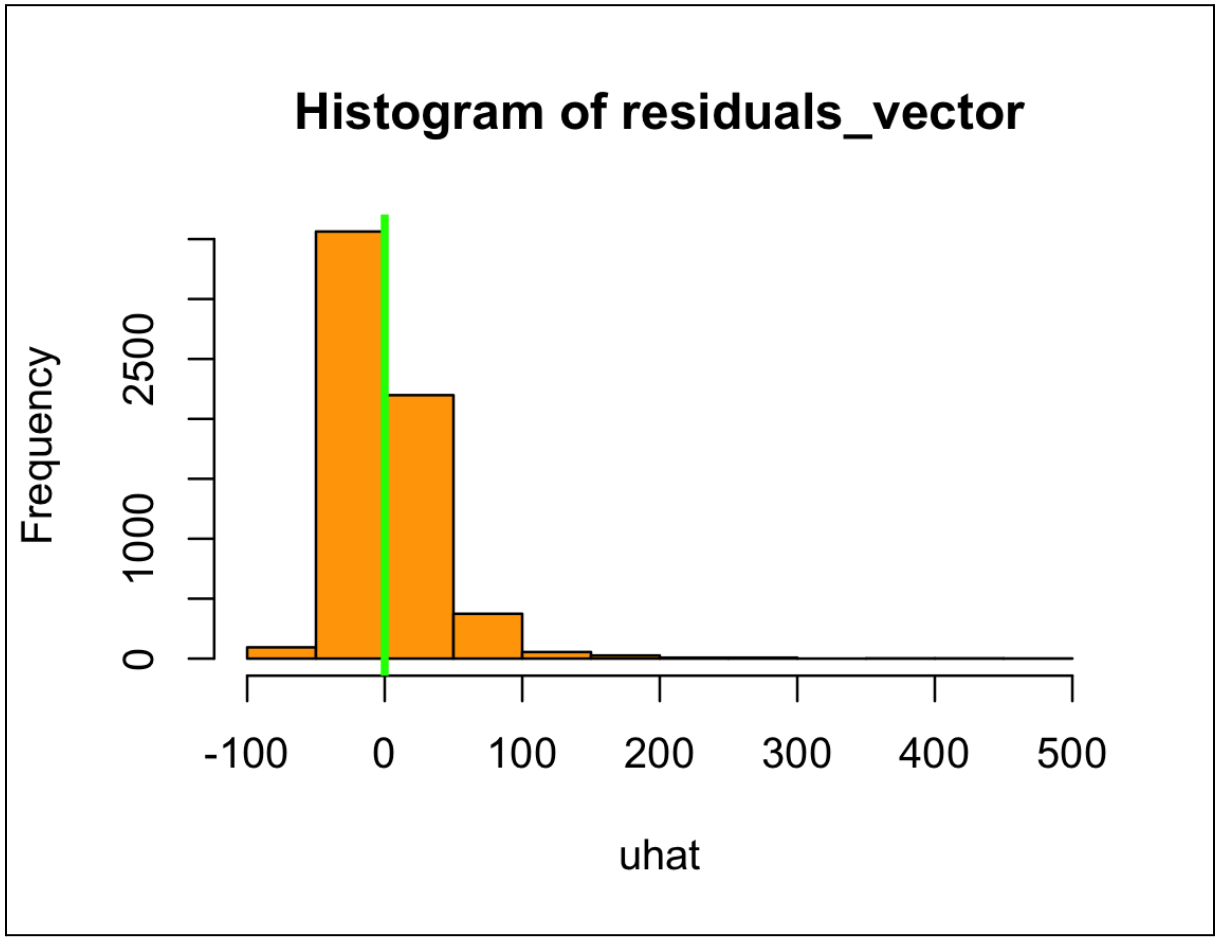
Data Assignment 1

Q8.

The green line is plotted using the command:

```
abline(v = mean(residuals_vector),col = "green", lwd = 3)
```

Since this line is at 0, we can graphically see that the mean of all uhats is 0.



Statistic	N	Mean	St. Dev.	Min	Max
SDP	6,331	307,628.500	326,506.600	1,633	1,728,578
Calcium	6,331	60.290	38.472	0.000	512.000
uhat	6,331	0.000	38.146	-70.502	454.485

Q9.

Data Assignment 1

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

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
α_0	5.183e+01	2.544e+00	20.374	< 2e-16
α_1	7.486e-05	8.823e-06	8.485	< 2e-16
α_2	-8.857e-11	1.531e-11	-5.785	7.58e-09
α_3	3.215e-17	6.715e-18	4.788	1.72e-06
α_4	-1.170e+01	8.440e+00	-1.387	0.166

(where e+x = 10^x)

α_0 gives us the level of calcium in a district which has SDP = 0 and Gini Coefficient = 0 (meaning zero income inequality) for that district, that is, the expected level of calcium in groundwater in a district whose state has an SDP of Rs. 0 Cr and no income inequality, the level of calcium in the district's groundwater is expected to be 5.183e+01.

Data Assignment 1

1. α_1 gives us the expected increase in the EQI level (calcium in our case) for a unit increase in ***SDP*** if we hold other factors to be constant (not possible since SDP^2 and SDP^3 will be affected).

   α_2 gives us the expected increase in the EQI level (calcium in our case) for a unit increase in SDP^2 if we hold other factors to be constant (not possible since ***SDP*** and SDP^3 will be affected).

   α_3 gives us the expected increase in the EQI level (calcium in our case) for a unit increase in SDP^3 if we hold other factors to be constant (not possible since ***SDP*** and SDP^2 will be affected).

2. The effect of unit increase in SDP will thus be dependent on the combined effect of α_1, α_2 and α_3.
   Since α_2 is negative and is greater than α_3, it would suggest that at some point in the graph of $EQI_{i,t} = \alpha_1 SDP_{i,t} + \alpha_2 SDP^2_{i,t} + \alpha_3 SDP^3_{i,t}$, we would find a small decrease in the value of EQI.
   Excluding this small section of decrease, a unit increase in SDP will have the combined effect of increase in EQI via ***SDP***, SDP^2, SDP^3.

3. Note that α_1, α_2 and α_3 are significant even at a level of 0.1% significance.

4. α_4 gives us the expected increase in the EQI level (calcium in our case) for a unit increase in the Gini Coefficient if the change $\alpha_1 SDP + \alpha_2 SDP2 + \alpha_3 SDP3$ Is constant (which can also happen when SDP=0).

5. Note that the change in the Gini Coefficient for a district level can only be equal to 1 under special circumstances (when the district goes from having no income inequality to total income inequality) since the value of Gini Coefficient is between 0 and 1. So generally, the actual change in EQI level in most cases will be some number between 0 and 1 multiplied by α_4.

Data Assignment 1

6. Also note that the value of α_4 (negative in our case, thus indicating that higher inequality leads to lower environmental degradation) is not significant even at 10% level of significance (as while running it on R it did not indicate that the Gini Coefficient was significant for determining the relationship). Thus, indicating that the effect of income inequality on the environmental quality indicator is not significant for this case.