# Group8_EDA

Russell and Frances

2022-08-16

## Contents



# 1 Exploratory Data Analysis of Lake Data

## 1.1 Introduction

In this EDA we investigate the relationship between ou the modelled median Ammoniacal Nitrogen, median Chlorophyll-A, median Clarity, median Total Nitrogen and median Total Phosphorus in New Zealand lakes over one hectare in area, for the period 2013 to 2017.

From the above variables, we examine their individual distributions, their correlations with one another, their relationship with Dominant Landcover, various sample statistics, and Regional differences.

We also examine Regional differences in lake areas, perimeters and depths.

This is a Stats NZ dataset from https://www.stats.govt.nz/indicators/modelled-lake-water-quality/.

## 1.2 Explanation of variables

*Ammoniacal Nitrogen* is a form of nitrogen that supports algae and plant growth, but in large concentrations can be toxic to aquatic life. This is measured in milligrams per litre. The national bottom line for this measure is 1.3mg/L, which none of the observations exceed.

*Chlorophyll-a* is an organic molecule found in plant cells that allows plants to photosynthesize. The variable Chlorophyll-a is a measure of the concentration of phytoplankton biomass in milligrams per cubic metre. High concentrations of chlorophyll is a symptom of degraded water quality. The national bottom line for this measure is 12.

*Total Phosphorus* is the sum of all phosphorus forms in the water, including phosphorus bound to sediment. Large amounts of phosphorus in lakes can reduce dissolved oxygen in the water. This can cause low oxygen

areas in the lake, where some aquatic life cannot survive. Total Phosphorus is measured in milligrams per cubic metre and has a national bottom line of 50mg/m3.

*Clarity* is measured in Secchi depth. This is the maximum depth (in metres) a black and white Secchi disk is visible from the surface of the lake.

*Dominant Landcover* is split into four types; Exotic Forest, Native, Pastoral and Urban area. There are 12 lakes with no entry for dominant land cover, however in the description of the dataset by Stats NZ, it states all lakes have been categorised, and indicated these empty entries should be another category called 'other' that includes 'Gorse and/or Broom', 'Surface mines and dumps', 'Mixed exotic shrubland', and 'Transport infrastructure' so we have assigned these to the other category. The category Urban area is applied if urban cover exceeds 15 percent of catchment area. Pastoral is applied if pastoral exceeds 25 percent of catchment area and not already assigned urban. The other three categories; Exotic forest, Native, or Other were assigned according to the largest land cover type by area, if not already assigned urban or pastoral.

*Regions* in this dataset are; Auckland, Bay of Plenty, Canterbury, Gisborne, Hawke's Bay, Whanganui, Marlborough, Northland, Otago, Southland, Taranaki, Tasman, Waikato, Wellington and West Coast. Each lake corresponds to the region it is located in.

### 1.3 Sample Statistics of Lake Health variables

Table 1 shows the summary statistics for each of these three measures.

Table 1: Table of Sample Statistics

| | Ammoniacal Nitrogen | Chloropyll-A | Phosphorus | Nitrogen | Clarity |
|---|---|---|---|---|---|
| Sample Size | 3802.0000000 | 3802.000000 | 3802.000000 | 3802.000000 | 3802.0000000 |
| Minimum | 0.0016940 | 0.473853 | 4.017657 | 35.444730 | 0.3553600 |
| 1st Quantile | 0.0073492 | 2.750785 | 12.158160 | 286.827100 | 2.5136188 |
| Median | 0.0096110 | 3.948234 | 17.896640 | 416.704400 | 4.4677300 |
| 3rd Quantile | 0.0140320 | 5.758621 | 22.802612 | 648.096175 | 6.2323935 |
| Maximum | 0.0614130 | 40.448870 | 150.416800 | 1883.172000 | 11.2488500 |
| Inter-quartile Range | 0.0066828 | 3.007836 | 10.644452 | 361.269075 | 3.7187747 |
| Standard Deviation | 0.0068358 | 2.807067 | 9.143676 | 277.994471 | 2.2553455 |
| Mean | 0.0119528 | 4.609290 | 18.720584 | 505.860630 | 4.4509687 |
| Median Absolute Deviation | 0.0039348 | 2.179829 | 8.055040 | 229.444359 | 2.8097827 |
| Kurtosis | 8.1157555 | 18.340809 | 26.244106 | 3.546338 | 1.9982754 |
| Skewness | 2.0338977 | 2.548402 | 2.872190 | 1.079944 | 0.2342007 |

Figure 1 shows the distribution of Ammoniacal Nitrogen. The fitted normal distribution (in red) differs significantly from the smoothed histogram (purple). The smoothed histogram is more skewed and the mode is well below the mean. The median 0.0096 and mean 0.0120. Table 1 confirms this with a skewness of 2.0339. The kurtosis is 8.1158, indicating the distribution of Ammoniacal Nitrogen has heavy tails. There are two extreme values, at around 0.06 mg per litre. There is a large amount of observations around the first quantile, 0.0073 mg per litre.

Figure 2 shows the distribution of Chlorophyll-A. We can see the fitted normal distribution differs slightly from the smoothed histogram. Table 1 shows the median is 3.9482 and the mean is 4.6093. The kurtosis is 18.3408, indicating the distribution of Chlorophyll-A is very heavy tailed, and the skewness is 2.5484, indicating the distribution is right skewed.

Figure 3 shows the distribution of Total Phosphorus. The fitted normal distribution (red) fits quite well to the smoothed histogram (purple), although the kurtosis is very high, 26.2763. We would expect the kurtosis of a normally distributed variable to be close to 3 and with a skewness of 0, however the sample statistics

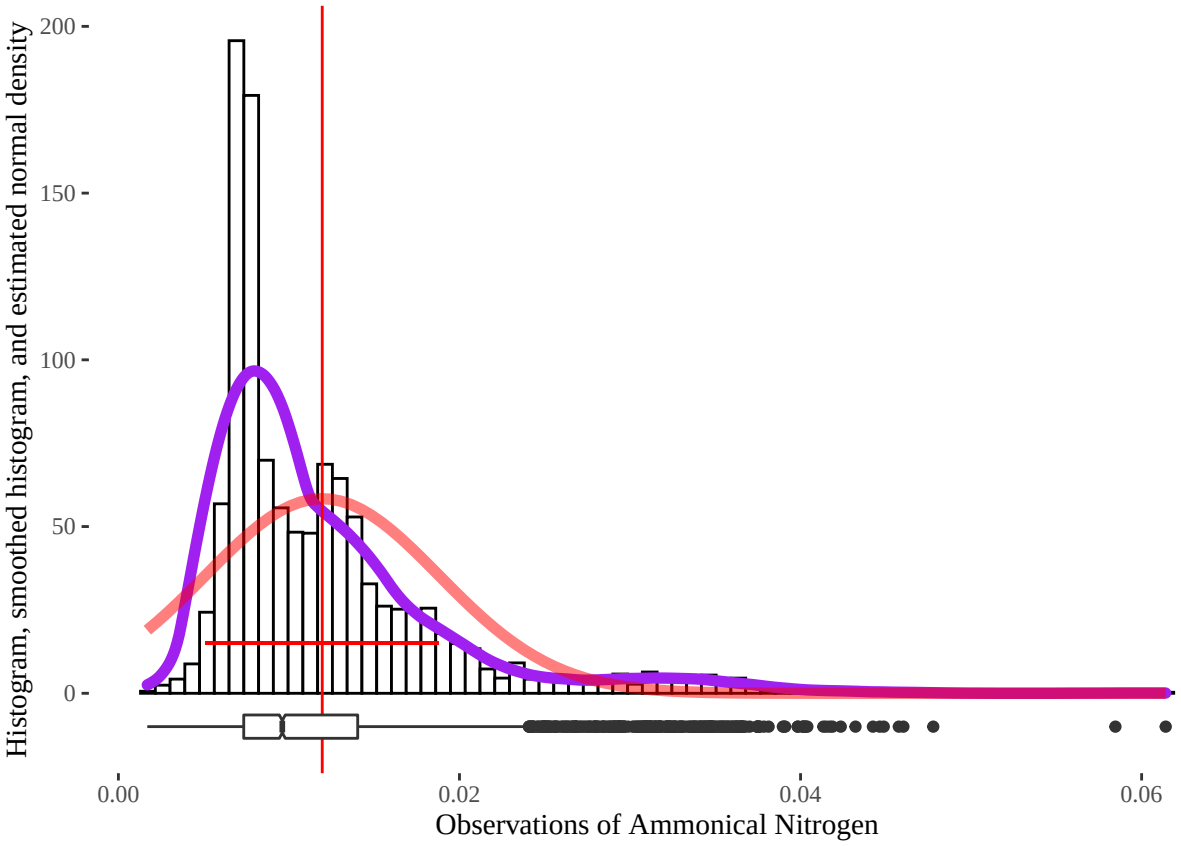



Figure 1: Histogram of Ammoniacal Nitrogen

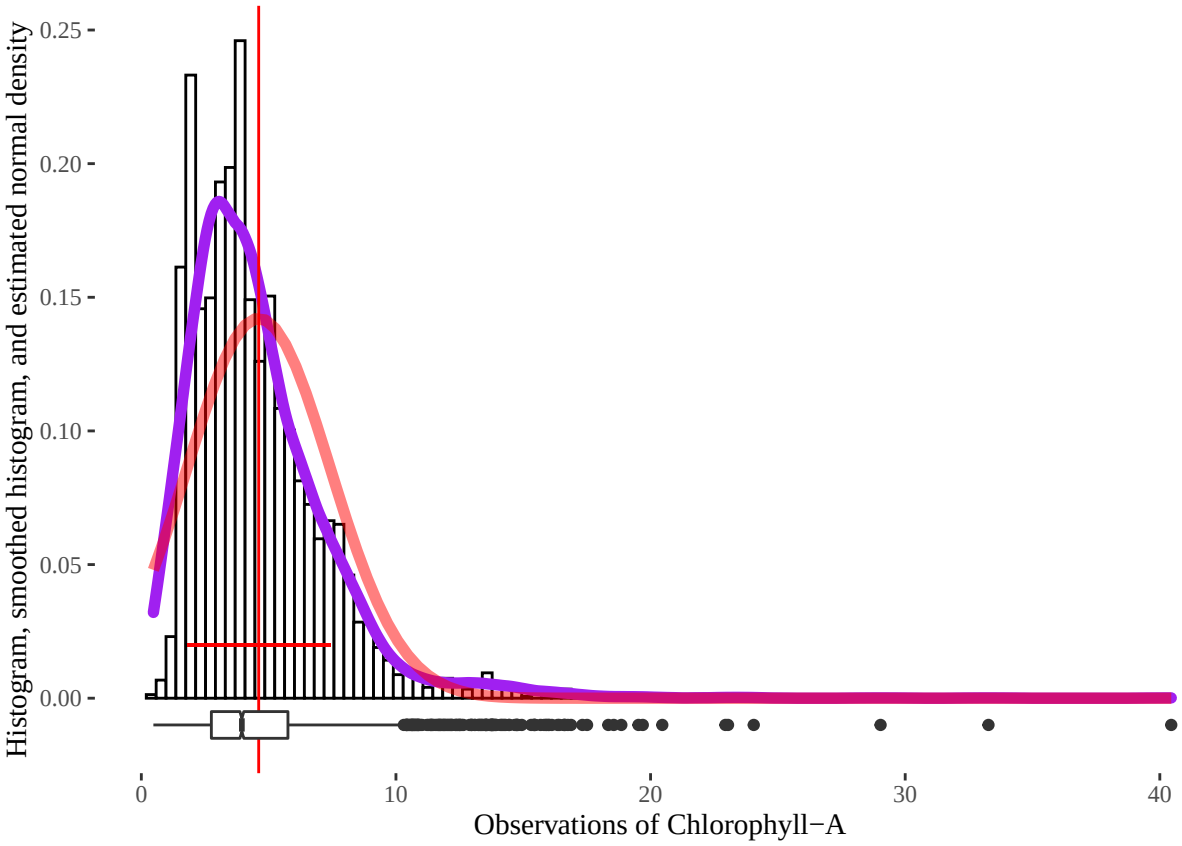



Figure 2: Histogram of Chlorophyll-A

of Total Phosphorus show the kurtosis much larger than 3 and the skewness 2.8722. This indicates the tails of this distribution are much heavier than a normal distribution, and it is right skewed. Table 1 shows the median of Total Phosphorus is 17.8966 mg per cubic meter, and the mean is 18.7206 mg per cubic meter.

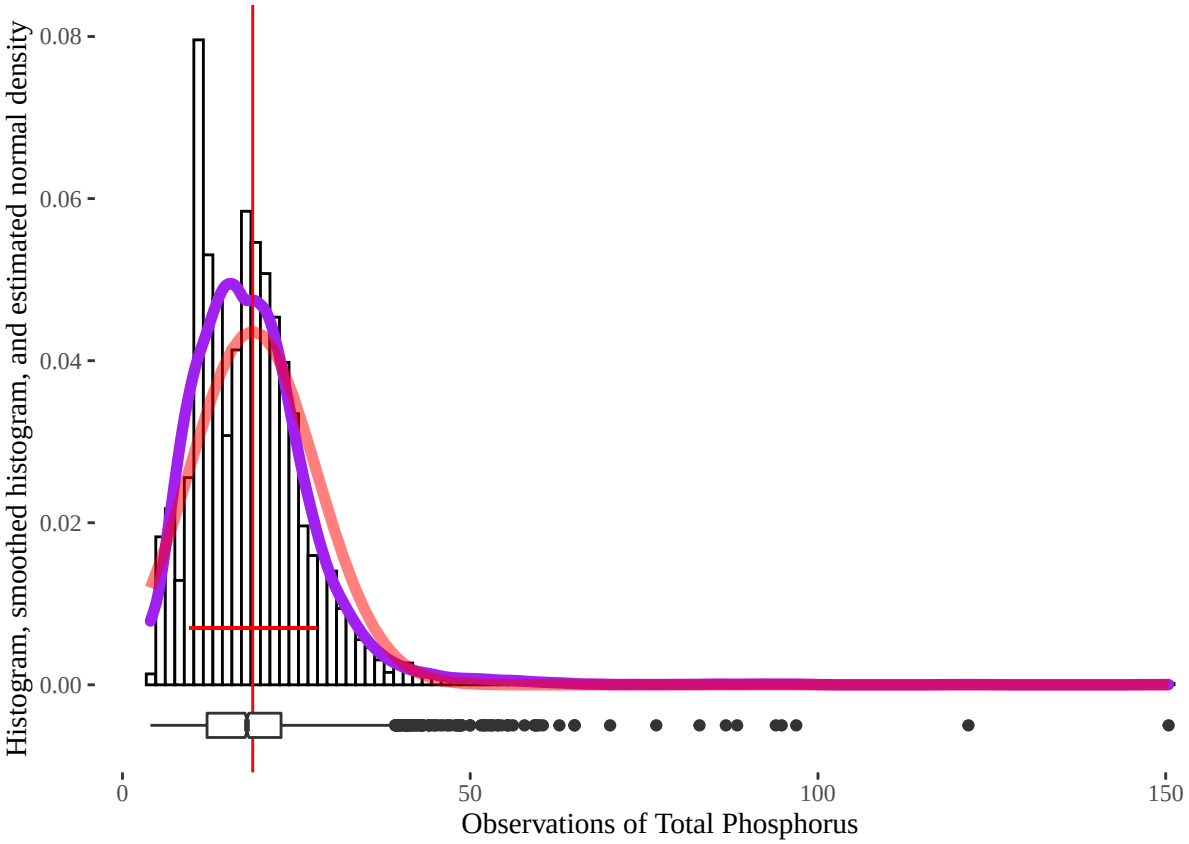



Figure 3: Histogram of Total Phosphorus

Figure 4 shows the distribution of Total Nitrogen. The fitted normal distribution (in red) differs significantly from the smoothed histogram (purple). The smoothed histogram is more skewed and the mode is well below the mean. The median 416.7044 and mean 505.8606. Table 1 confirms this with a skewness of 1.0799. The kurtosis is 3.5463, indicating the distribution of Total Nitrogen has reasonable tails.

Figure 5 shows the distribution of Clarity. The fitted normal distribution (in red) differs from the smoothed histogram (purple). The smoothed histogram is slightly asymmetrical but not skewed, with a median of 4.4677 and a mean of 4.4510. Table 1 confirms this with a skewness of 0.2342. The kurtosis is 1.9983, indicating the distribution of Clarity has slightly lighter tails than a normal distribution. These statistics indicate the distribution of Clarity is close to normal.

Table 2 shows the sample mean vector of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity.

The covariance matrix is shown in table 3. The variance of Ammoniacal Nitrogen is 0.000047, Chlorophyll-A is 7.8796, Phosphorus is 83.6068, Total Nitrogen is 77280.9259 and the variance of Clarity is 5.0866. The covariance of Ammoniacal Nitrogen and Chlorophyll-A is 0.0040, the covariance of Ammoniacal Nitrogen and Phosphorus is 0.0234, Ammoniacal Nitrogen and Total Nitrogen is 1.3861 and the covariance of Ammoniacal Nitrogen and Clarity is -0.0079. These indicate there is a Positive relationship between Ammoniacal Nitrogen and the other measures, but a negative relationship with Clarity. The covariance of Chlorophyll-A and Phosphorus is 17.8102, Chlorophyll-A and Nitrogen is 433.3090, and Chlorophyll-A and Clarity is -3.7018.

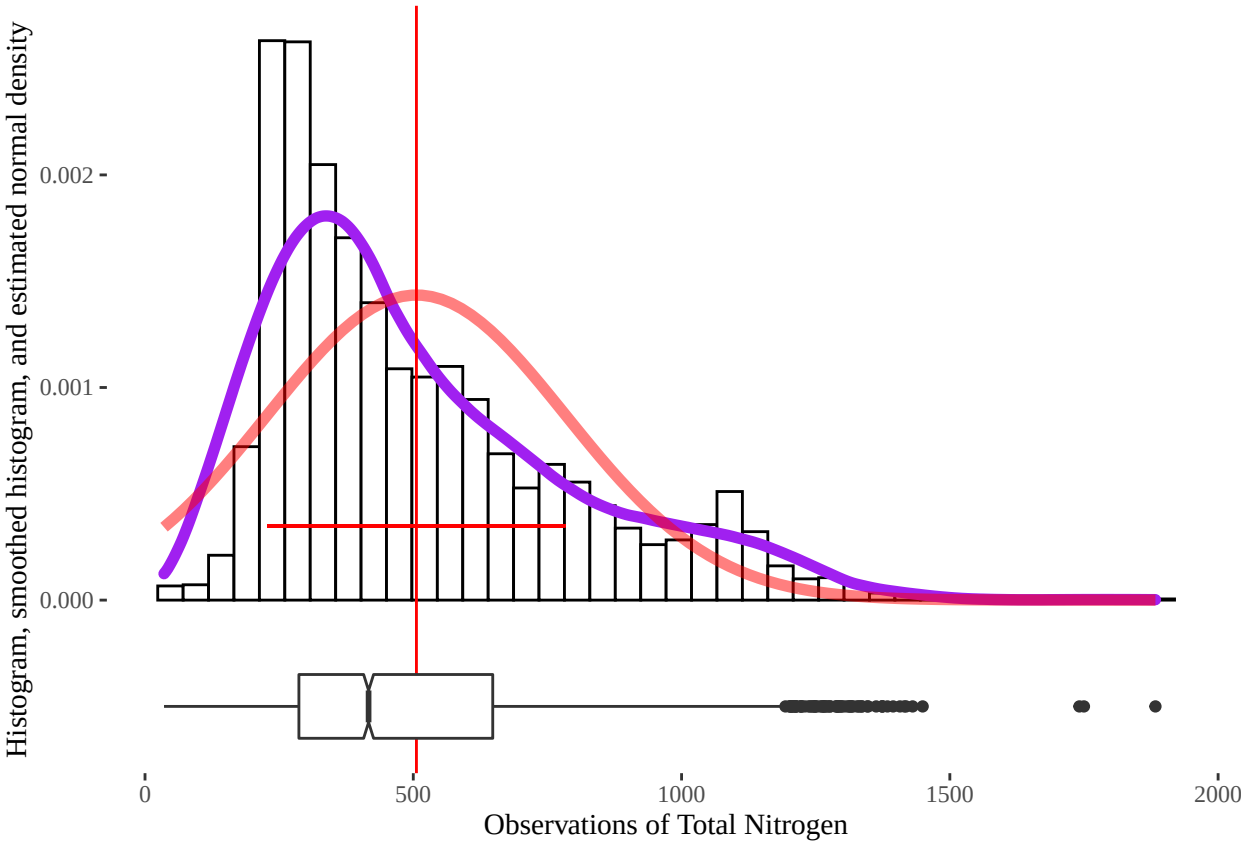



Figure 4: Histogram of Total Nitrogen

Table 2: Means Vector

| | Sample Means |
|---|---|
| Ammoniacal Nitrogen | 0.0119528 |
| Chlorophyll-A | 4.6092895 |
| Total Phosphorus | 18.7205843 |
| Total Nitrogen | 505.8606304 |
| Clarity | 4.4509687 |

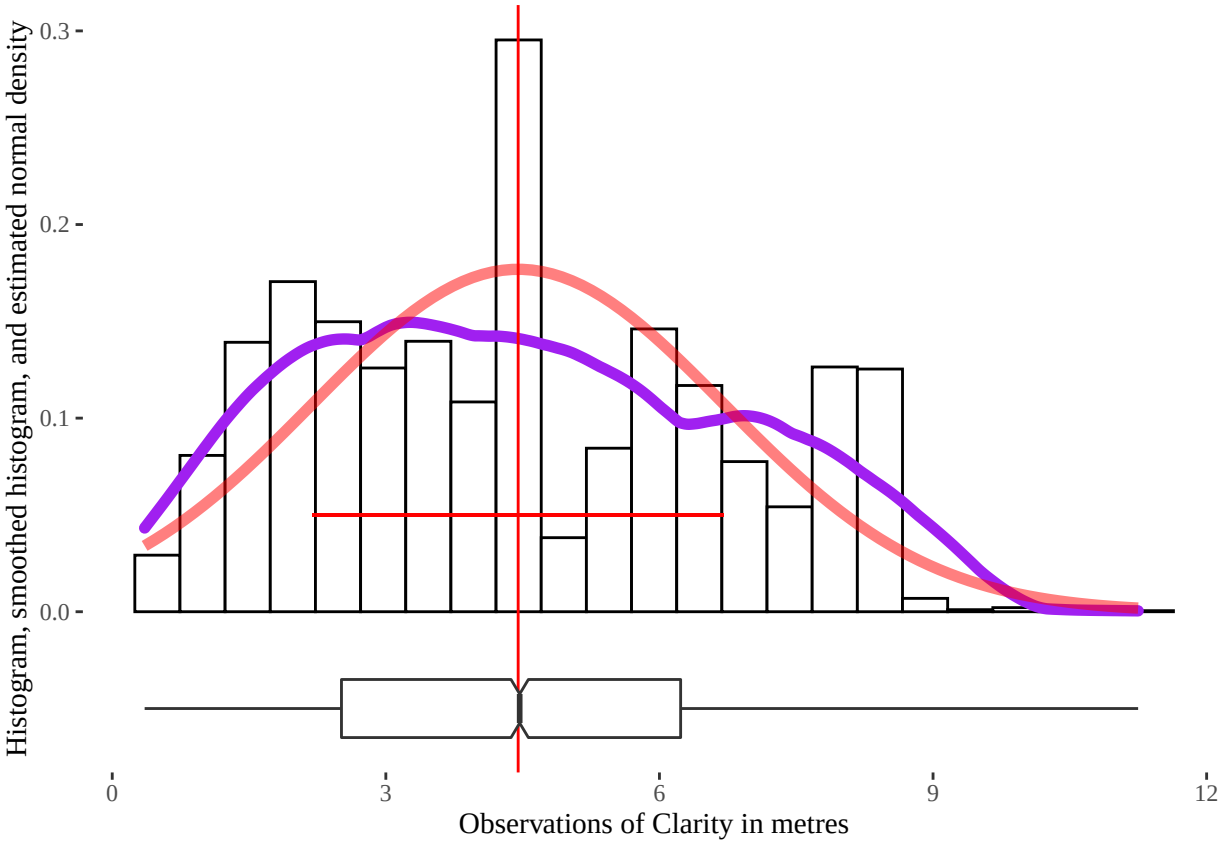



Figure 5: Histogram of Clarity (in metres)

Table 3: Covariance Matrix

| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Ammoniacal Nitrogen | 0.000047 | 0.003983 | 0.023413 | 1.386062 | -0.007853 |
| Chlorophyll-A | 0.003983 | 7.879623 | 17.810206 | 433.309014 | -3.701782 |
| Total Phosphorus | 0.023413 | 17.810206 | 83.606805 | 1640.359387 | -11.004913 |
| Total Nitrogen | 1.386062 | 433.309014 | 1640.359387 | 77280.925931 | -490.520745 |
| Clarity | -0.007853 | -3.701782 | -11.004913 | -490.520745 | 5.086583 |

Table 4: Correlation Matrix

| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Ammoniacal Nitrogen | 1.0000000 | 0.2075794 | 0.3745765 | 0.7293838 | -0.5093412 |
| Chlorophyll-A | 0.2075794 | 1.0000000 | 0.6938977 | 0.5552759 | -0.5847161 |
| Total Phosphorus | 0.3745765 | 0.6938977 | 1.0000000 | 0.6453303 | -0.5336453 |
| Total Nitrogen | 0.7293838 | 0.5552759 | 0.6453303 | 1.0000000 | -0.7823627 |
| Clarity | -0.5093412 | -0.5847161 | -0.5336453 | -0.7823627 | 1.0000000 |

Similar to Ammoniacal Nitrogen, all variables except Clarity have a positive relationship with Chloropyll-A. The covariance of Total Phosphorus and Total Nitrogen is 1640.3594, yet again indicating a positive relationship. However, the covariance of Total Phosphorus and Clarity is -11.0049 and Total Nitrogen and Clarity is -490.5207. These indicate a negative relationship between Phosphorus and Clarity, and Nitrogen and Clarity.

Table 4 shows a visualisation of the correlation matrix. The strength of relationship is interpreted the size of the circles with strong relationships having larger circles and weaker relationships having small circles. Strength is also communicated via colour with darker colours indicating stronger relationship. Colour also indicates the direction of the relationship via a colour spectrum with reddish indicating a positive relationship, blue indicating a negative relationship and white indicating no relationship.

Ammoniacal Nitrogen has a weak, positive relationship with Chlorophyll-A (0.2076) and a Moderate, positive relationship with Total Phosphorus (0.3746). It has quite a strong relationship with Total Nitrogen (0.7294, which is to be expected asTotal Nitrogen includes Ammoniacal Nitrogen). Chlorophyll-A has a reasonably strong relationship with both Total Phosphorus (0.6939), and a moderate relationship with Total Nitrogen (0.5553). Total Phosphorus has a reasonable strong relationship with Total Nitrogen of (0.64533). These statistics suggest that there may be a positive relationship between the different molecules in water.

Clarity is the only variable to have any negative relationship with the other variables. Clarity has a negative correlation with all of the other variables of at least -0.5. It has an especially strong relationship with Total Nitrogen of -0.7823. These negative relationships are tp be expected as it makes sense that an increase of other molecules in water would reduce the water's clarity. We can see this represented as dark blue circles in figure 6

The pairs plot is shown in figure 7.

Figure 8 shows the relationship between Ammoniacal Nitrogen and Chlorophyll-A. We can see a very weak relationship, as shown with the low correlation of 0.2076.

The linear relationship between Ammoniacal Nitrogen and Phosphorus is shown clearly in figure 9. The correlation of 0.3746 is still classified as a weak positive relationship but we can identify this relationship in the scatterplot.

The strong correlation between Ammoniacal Nitrogen and Total Nitrogen is apparent in the scatterplot of their relationship shown in figure 10. We can see a positive linear relationship, with very slight to no curvature.

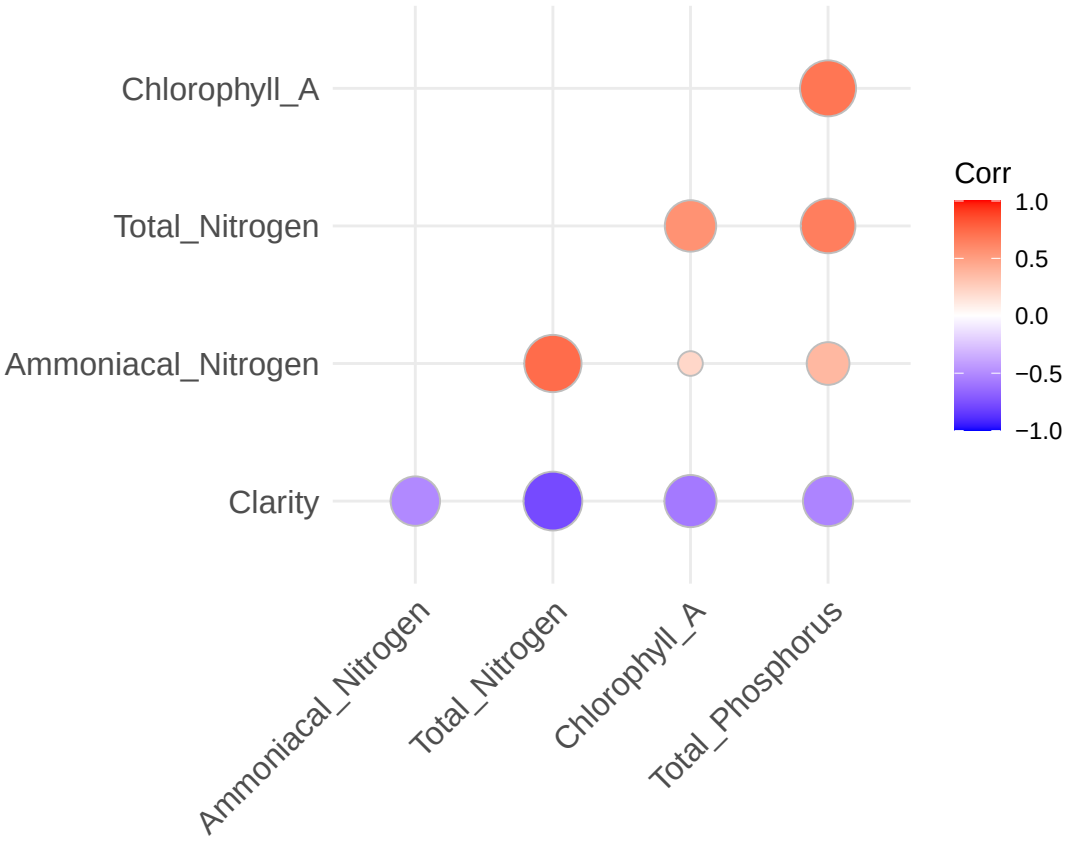



Figure 6: Visualisation of the Correlation Matrix

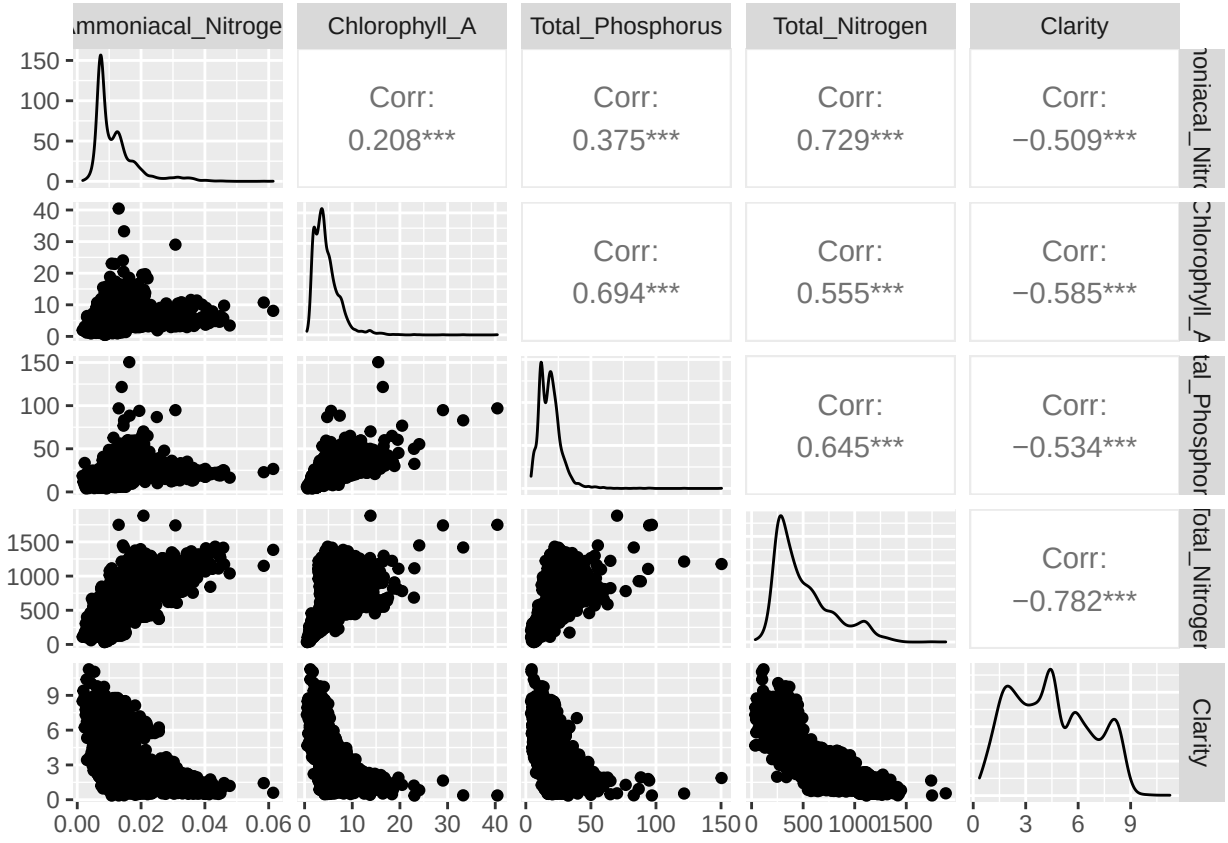



Figure 7: Pairs Plot

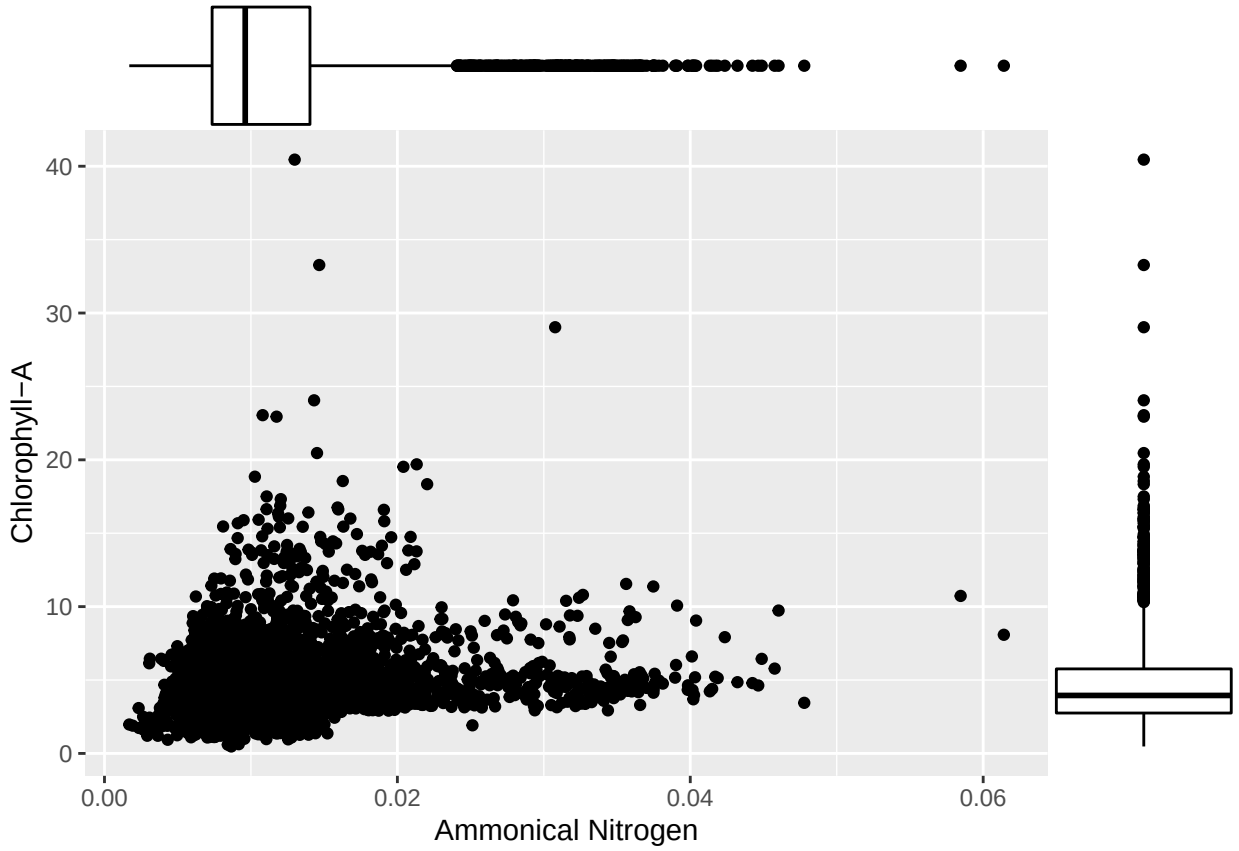



Figure 8: Pairs Plot of Ammoniacal Nitrogen and Chlorophyll-A

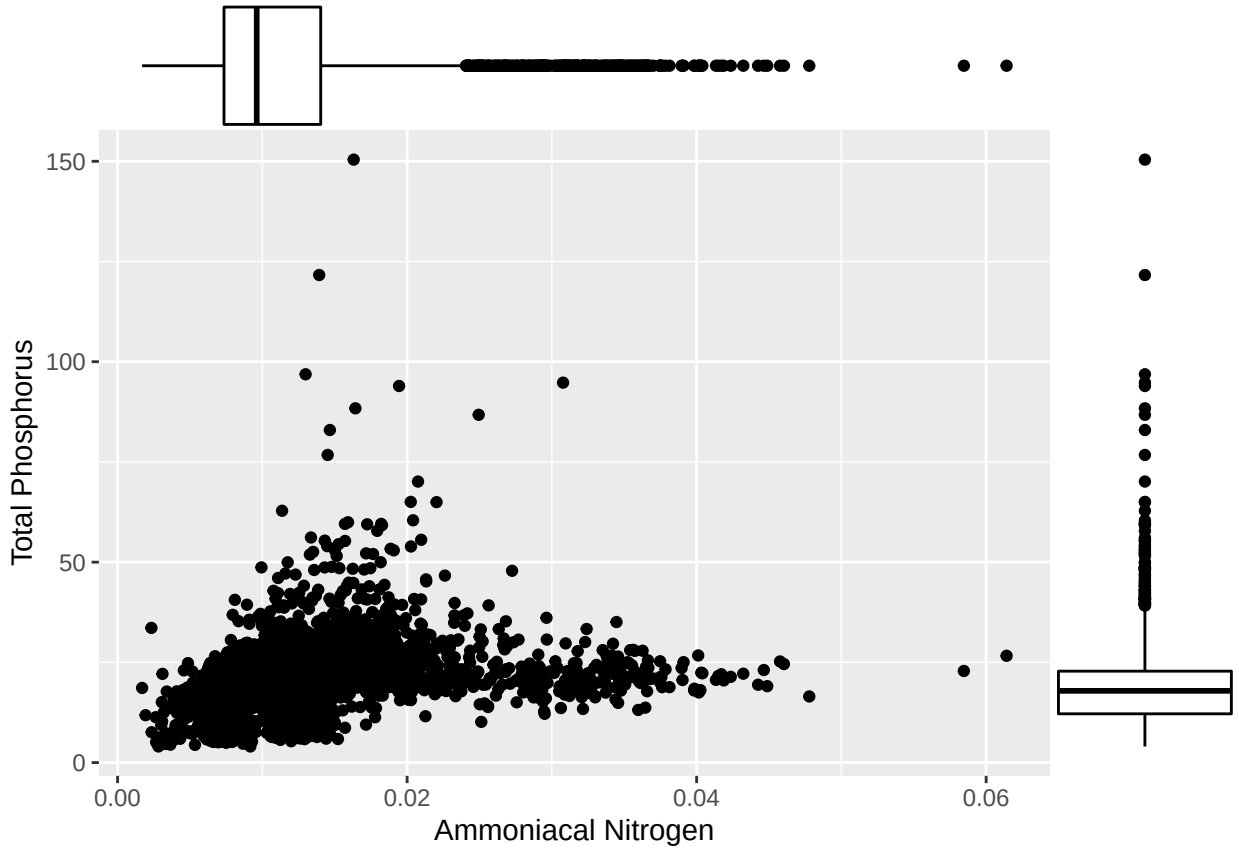



Figure 9: Pairs Plot of Ammoniacal Nitrogen and Total Phosphorus

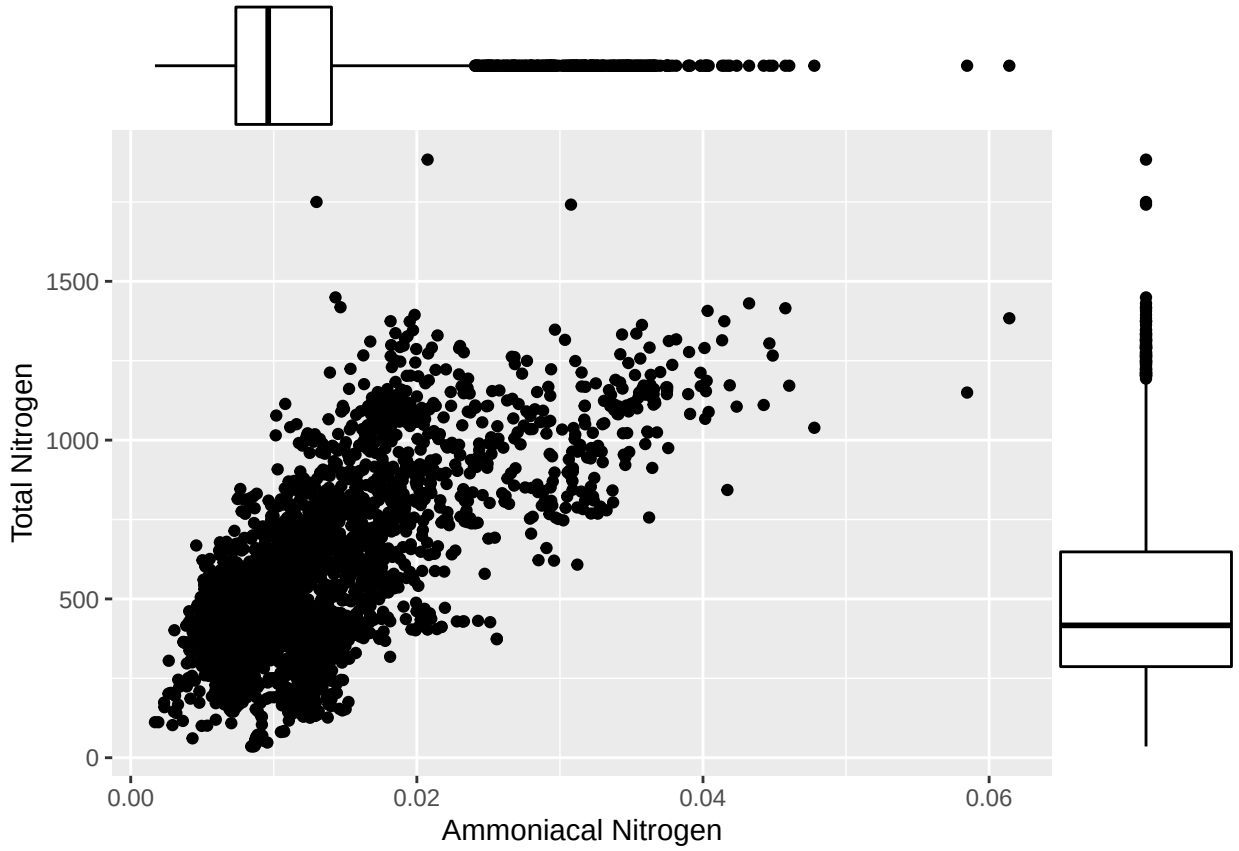



Figure 10: Pairs Plot of Ammoniacal Nitrogen and Total Nitrogen

Figure 11 shows the pairs plot for Ammoniacal Nitrogen and Clarity. We can see a clearly non-linear, negative relationship. This can be seen in the correlation of -0.5093.

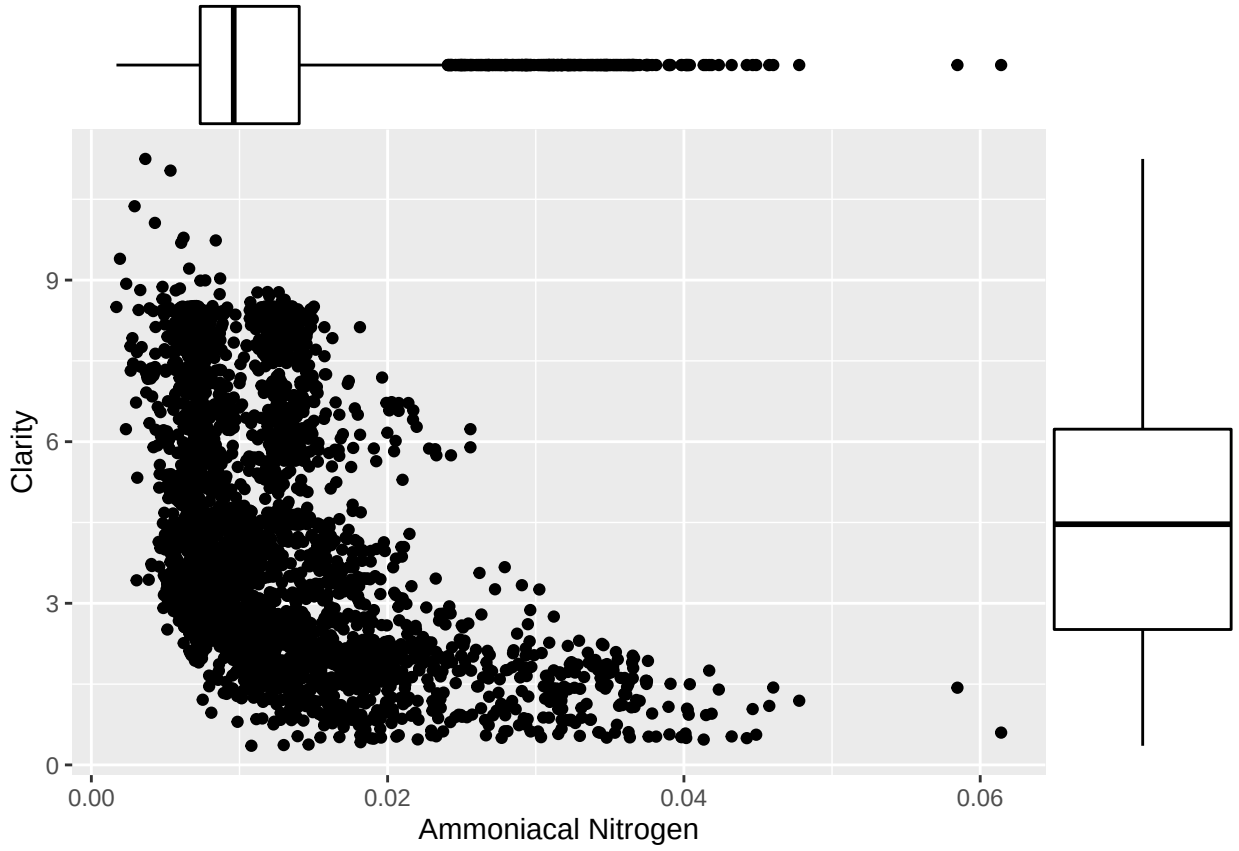



Figure 11: Pairs Plot of Ammoniacal Nitrogen and Clarity

Figure 12 shows a linear relationship between Chlorophyll-A and Total Phosphorus. However, we can see there is non-constant variance. As the amount of Chlorophyll-A and Phosphorus increases, the variance also increases. This can be seen in the funneling of the data. The relationship is strong in the lower values, but becomes weaker as Chlorophyll-A and Phosphorus increases, resulting in a reasonably strong correlation of 0.6939.

Figure 13 shows a similar relationship between Chlorophyll-A and Nitrogen as the relationship between Chlorophyll-A and Phosphorus in figure 12. Once again we can see funneling but a relatively strong linear relationship. This pairs plot shows an odd concentration of observations above the trend, around 1000mg/cubic metre of Total Nitrogen. The correlation is influenced by this, and the moderate sample correlation of 0.5553 reflects the strength of the trend, weakened by the odd concentration of observations away from the trend.

The pairs plot of Chlorophyll-A and Clarity is shown in figure 14 with correlation -0.5847. Similar to the pairs plot of Ammoniacal Nitrogen and Clarity, we can see a non-linear, negative trend. The clarity appears to decrease exponentially as Chlorophyll-A increases.

Figure 9 shows the pairs plot of Nitrogen and Phosphorus. The correlation is 0.6453 and this is shown in the graph by a strong increasing trend. Once again, we can see some non-constant scatter, which may affect the correlation.

The pairs plot of Total Phosphorus and Clarity is shown in figure 16. There is less curvature in this plot than in figures 11 and 14. The correlation of -0.5336 is seen in the moderate, negative trend. The trend appears to show the clarity of the water reduces as the level of total phosphorus increases.

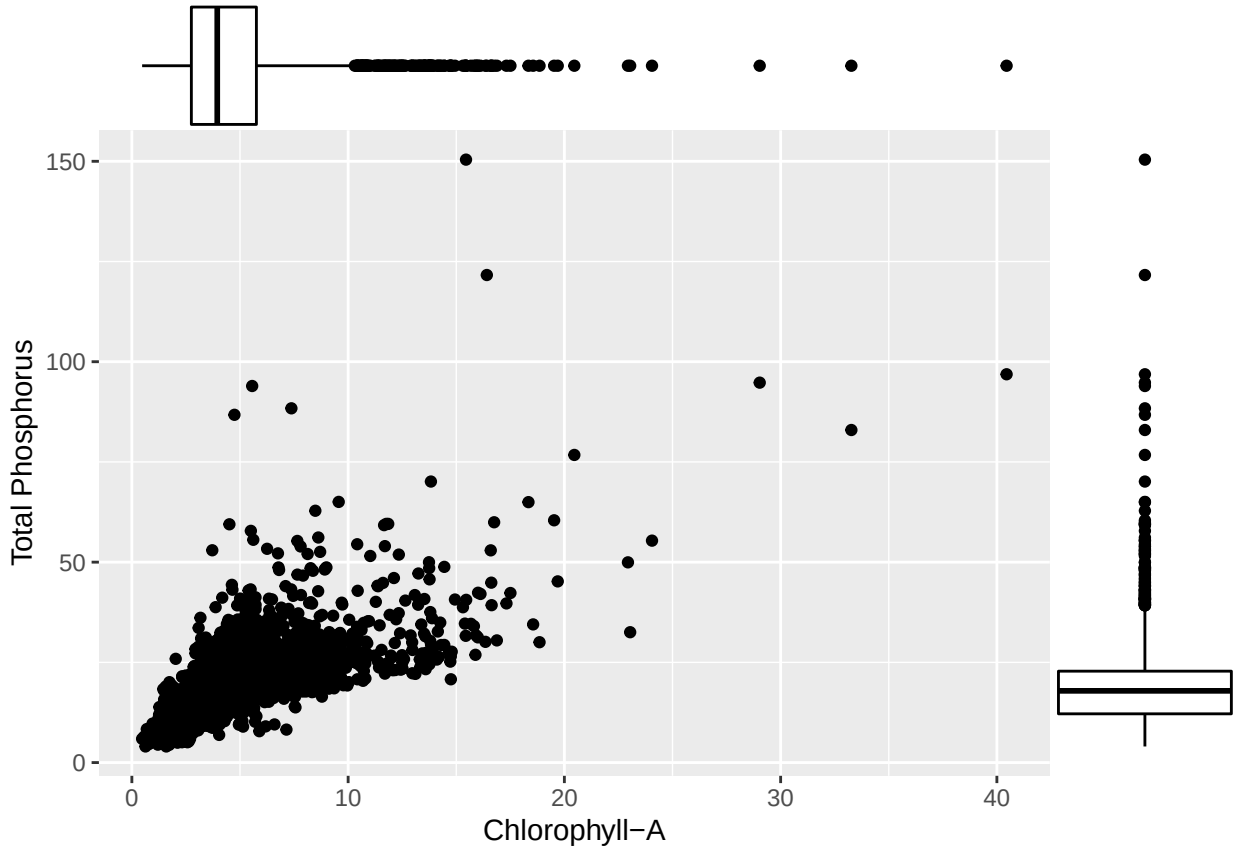



Figure 12: Pairs Plot of Chlorophyll-A and Total Phosphorus

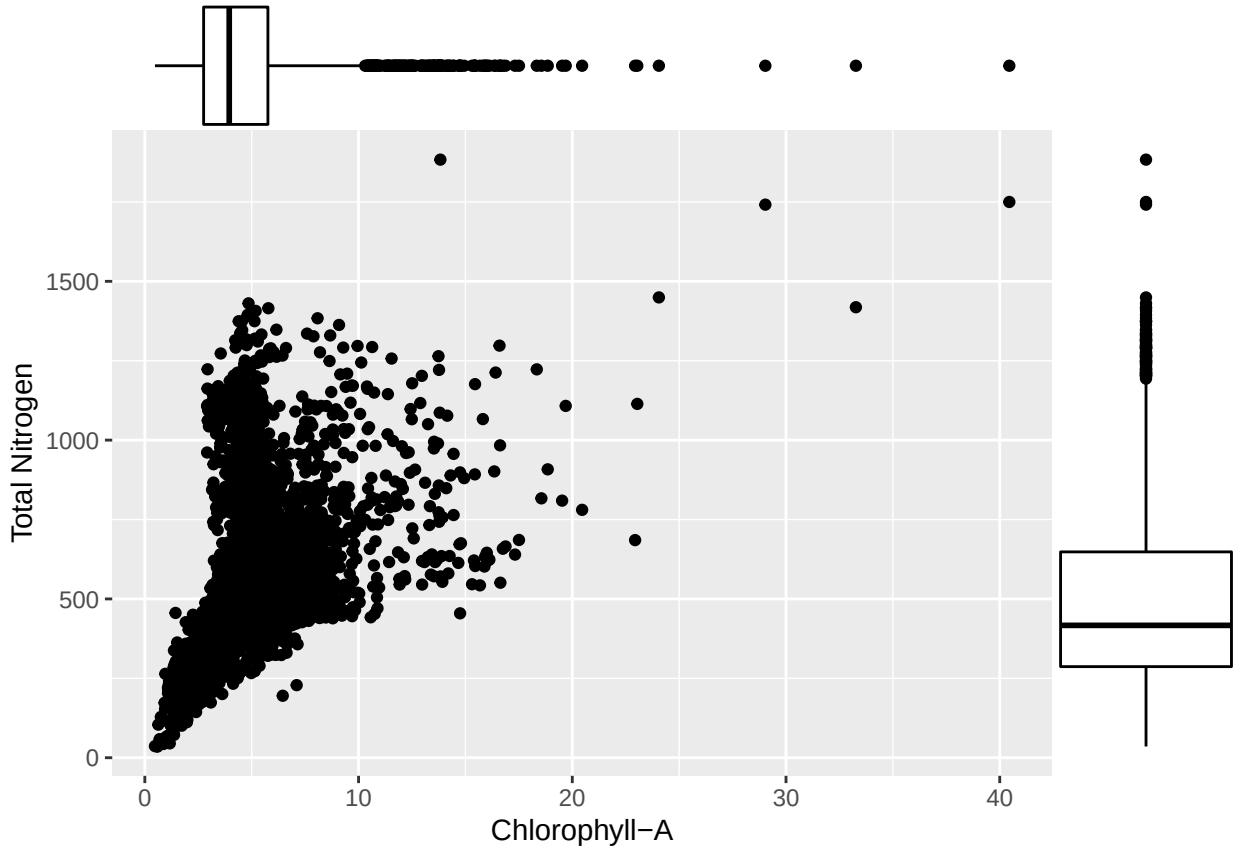



Figure 13: Pairs Plot of Chlorophyll-A and Total Nitrogen

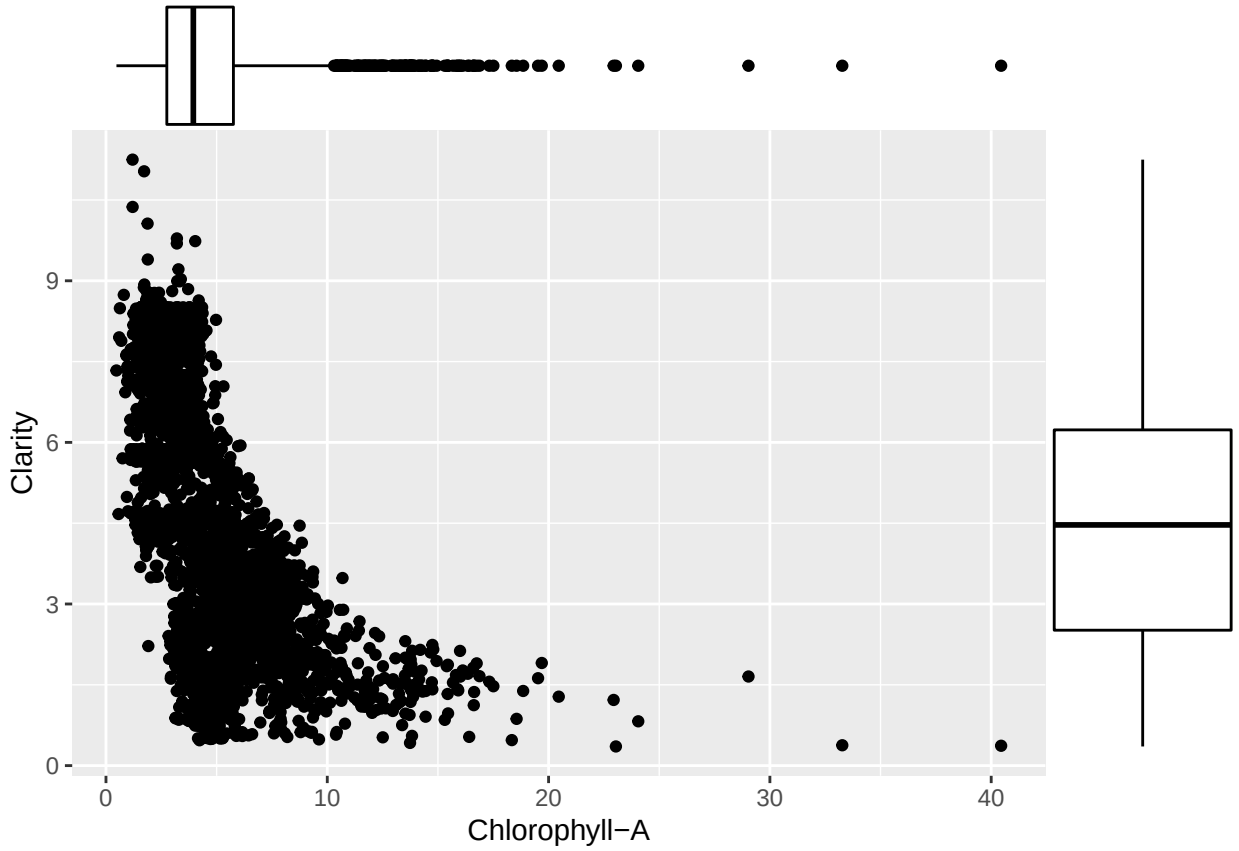



Figure 14: Pairs Plot of Chlorophyll-A and Clarity

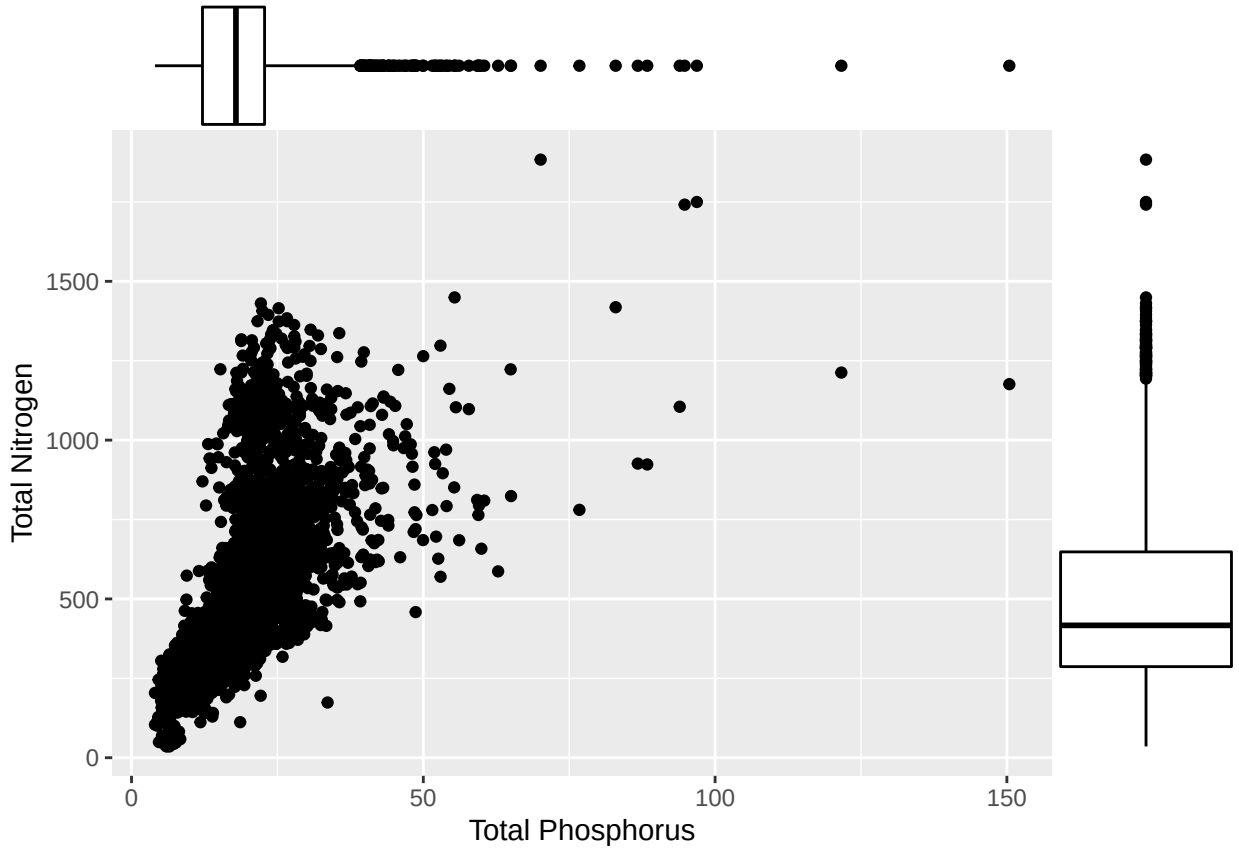



Figure 15: Pairs Plot of Total Phosphorus and Total Nitrogen

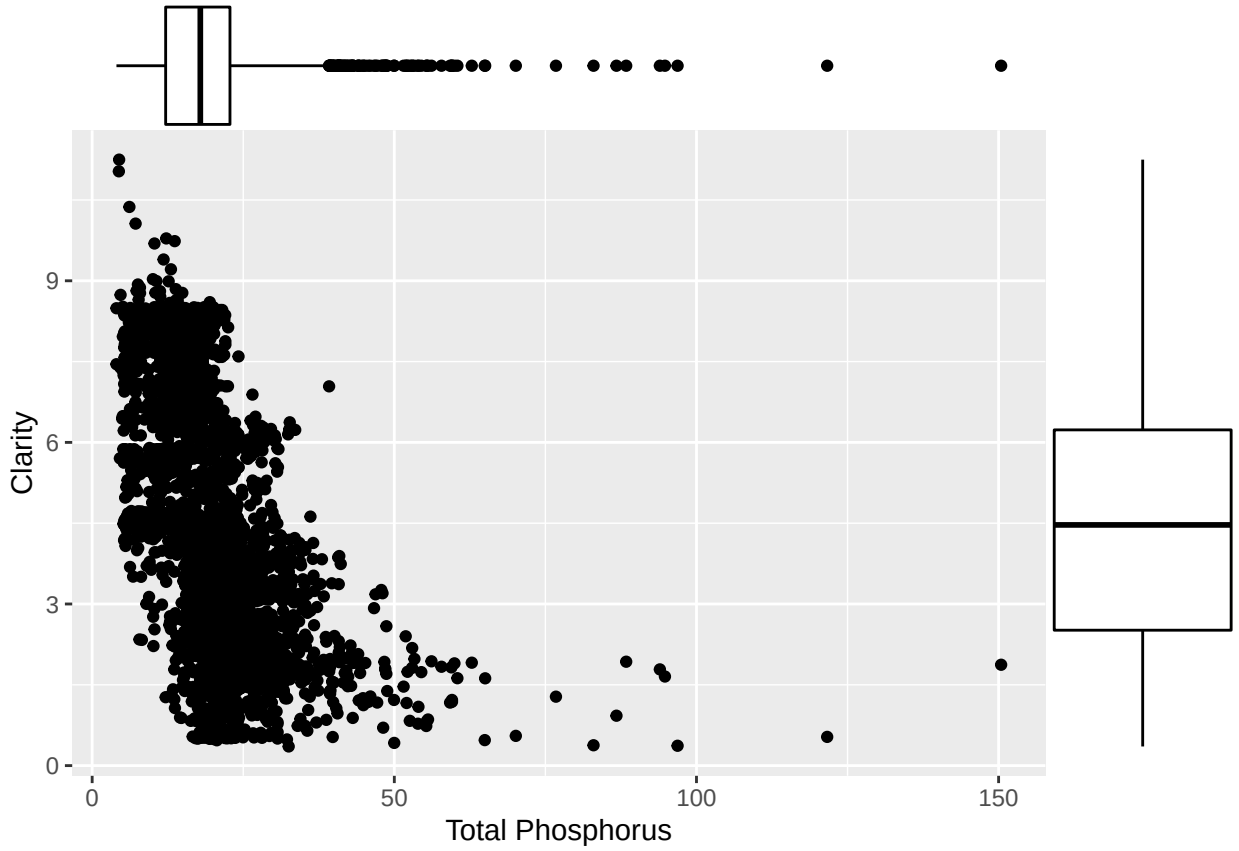



Figure 16: Pairs Plot of Total Phosphorus and Clarity

In comparison to figures 11, 14 and 16, figure 17 shows much more curvature in the relationship between Total Nitrogen and Clarity. We can clearly see a strong, non-linear, negative relationship between these two measures, supported by the sample correlation of -0.7824. This indicates as Total Nitrogen increases, the water becomes less clear.

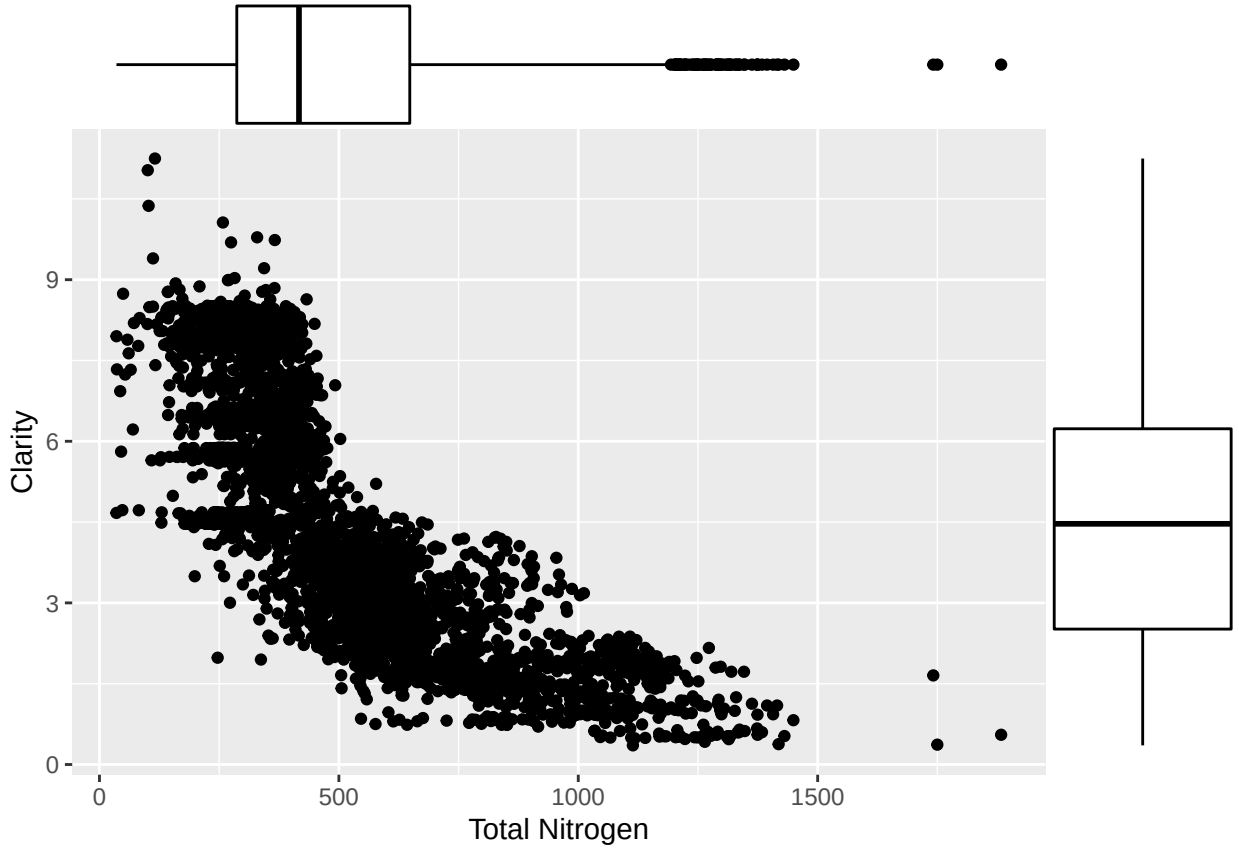



Figure 17: Pairs Plot of Total Nitrogen and Clarity

I have made Cullen and Frey plots for Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity, each with 1000 bootstrapped observations, shown in orange.

In figure 18, the Cullen and Frey graph of Ammoniacal Nitrogen, we can see the observed kurtosis and square of skewness was much larger than a normal distribution. The observation and all bootstrapped values were within the Beta distribution region, indicating the distribution of Ammoniacal Nitrogen in New Zealand lakes may follow a Beta distribution.

Figure 19 shows the Cullen and Frey graph of Chlorophyll-A. The observed kurtosis and square of skewness were both much larger than we would expect for a normal distribution. The observed value and the bootstrapped observations lie on or just below the line all lognormal distributions lie on. This could tell us the distribution of Chlorophyll-A in New Zealand lakes could follow a lognormal distribution.

The Cullen and Frey graph of Total Phosphorus is shown in figure 20. The observed kurtosis and square of skewness were larger than both Ammoniacal Nitrogen and Chlorophyll-A. Similar to the Cullen and Frey graph of Chlorophyll-A, the observed value and bootstrapped observations seem to lie close to the line that contains all lognormal distributions. However, very few of the bootstrapped observations lie on this line, indicating Total Phosphorus in New Zealand lakes likely does not follow a lognormal, or any other distribution illustrated on this graph.

The Cullen and Frey graph of Total Nitrogen is shown in figure 21. The observed value and the boostrapped

observations all lie within the grey area suggesting that Total Nitrogen follows a Beta distribution.

The Cullen and Frey graph of Clarity is shown in figure 22. The observed value and bootstrapped values of the skewness and kurtosis of Clarity lies within the area all beta distributions exist within, and very close to the Uniform distributions, indicating the distribution of Clarity could e Uniform or Beta.

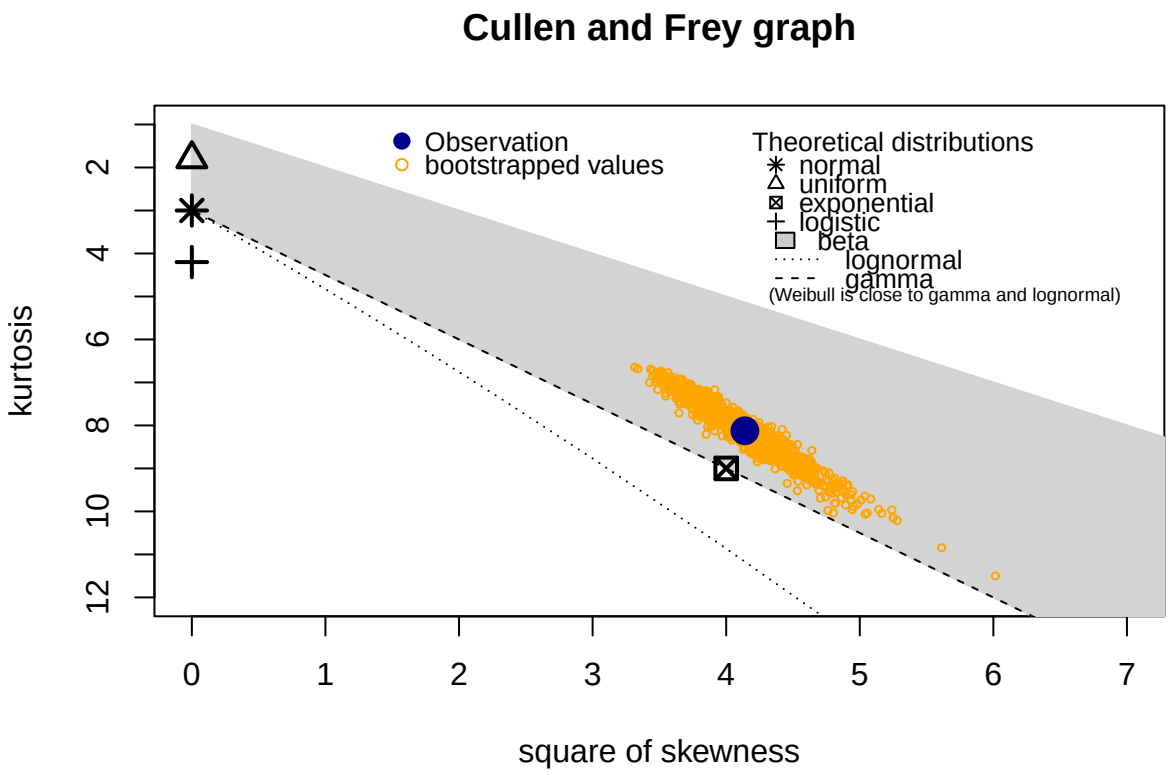



Figure 18: Cullen and Frey Graph of Ammoniacal Nitrogen

```
## summary statistics
## ------
## min:  0.001694   max:  0.061413
## median:  0.009611
## mean:  0.01195275
## estimated sd:  0.006835817
## estimated skewness:  2.034701
## estimated kurtosis:  8.124069
```

```
## summary statistics
## ------
## min:  0.473853   max:  40.44887
## median:  3.948234
## mean:  4.609289
## estimated sd:  2.807067
## estimated skewness:  2.549408
## estimated kurtosis:  18.36258
```

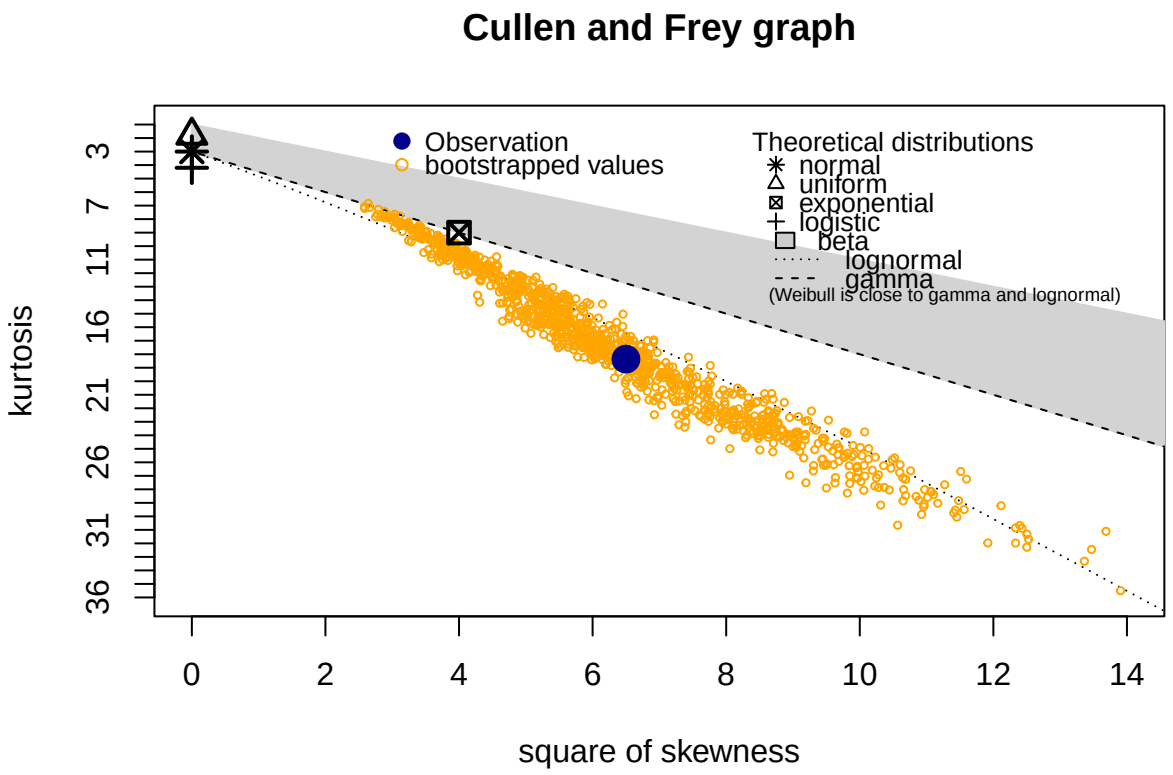



Figure 19: Cullen and Frey Graph of Chlorophyll-A

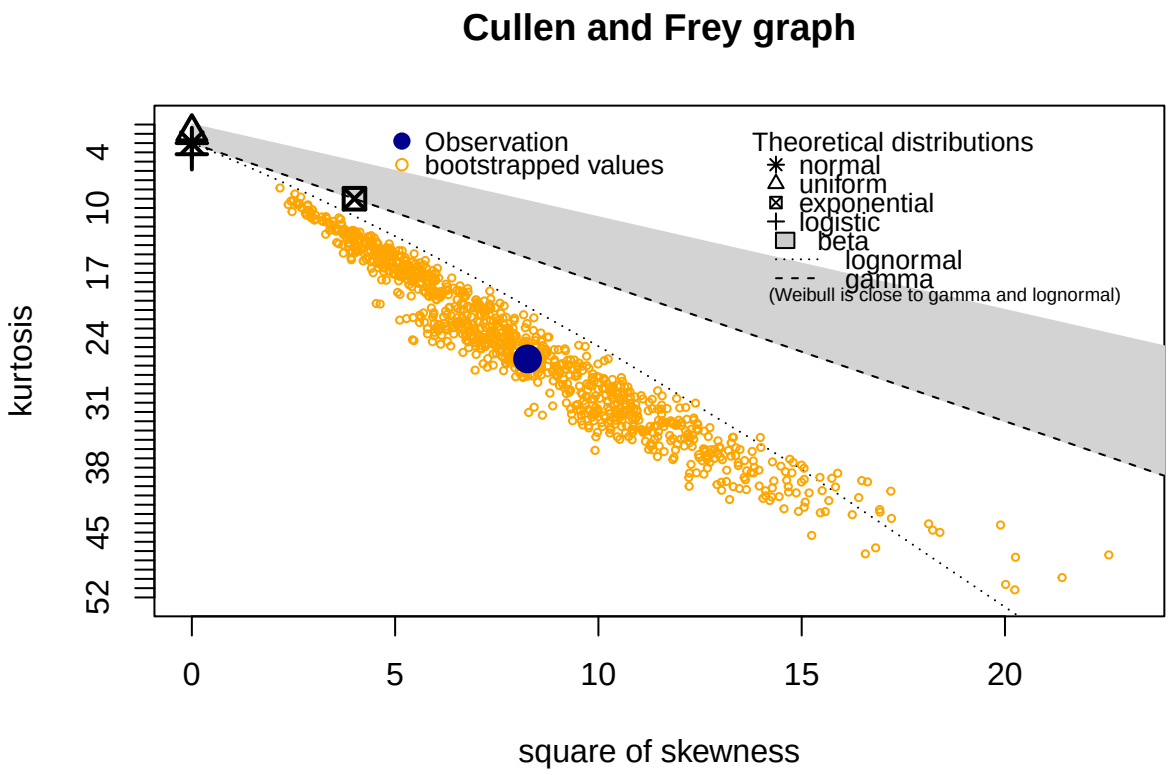



Figure 20: Cullen and Frey Graph of Total Phosphorus

```
## summary statistics
## ------
## min:  4.017657   max:  150.4168
## median:  17.89664
## mean:  18.72058
## estimated sd:  9.143676
## estimated skewness:  2.873323
## estimated kurtosis:  26.27628
```

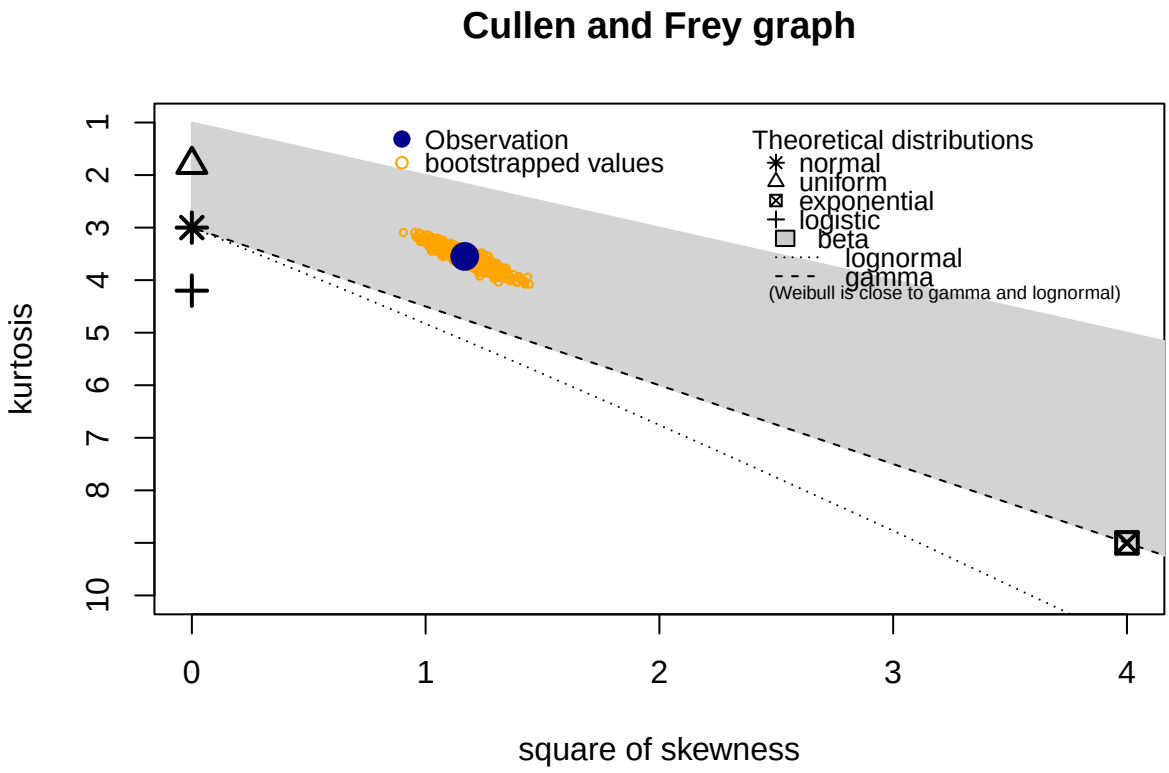



Figure 21: Cullen and Frey Graph of Total Nitrogen

```
## summary statistics
## ------
## min:  35.44473   max:  1883.172
## median:  416.7044
## mean:  505.8606
## estimated sd:  277.9945
## estimated skewness:  1.08037
## estimated kurtosis:  3.548637
```

```
## summary statistics
## ------
## min:  0.35536   max:  11.24885
## median:  4.46773
```

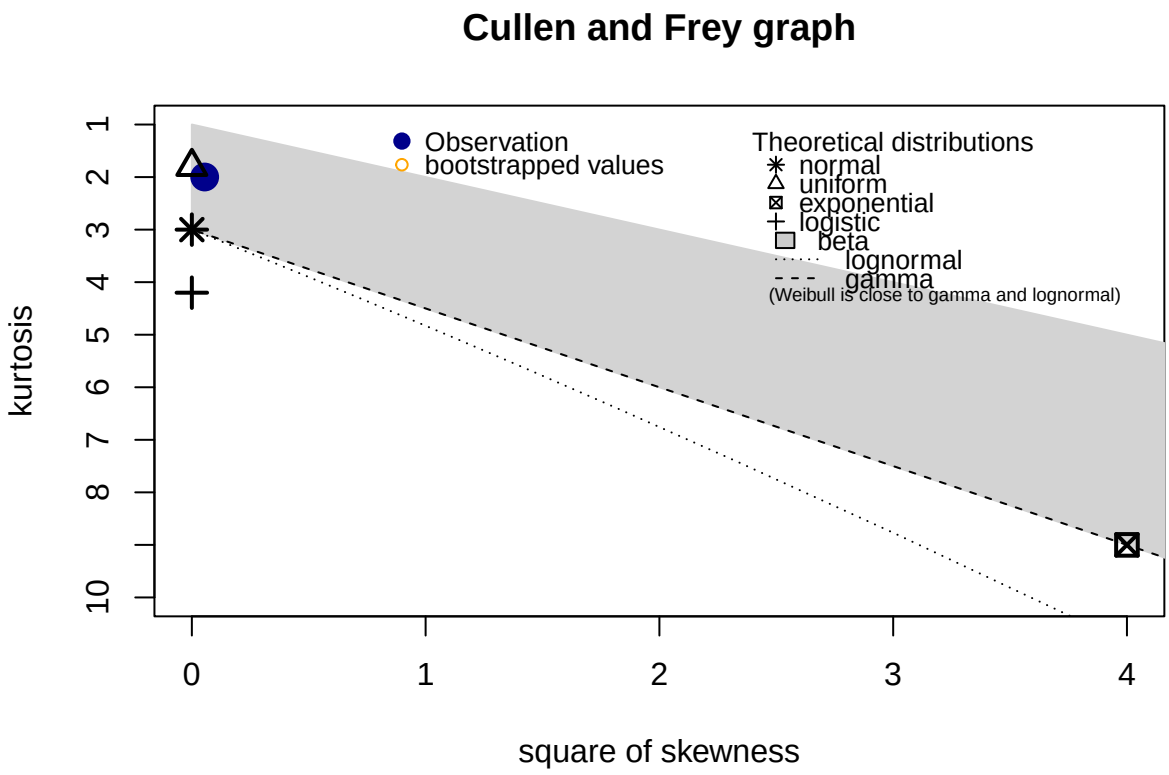



Figure 22: Cullen and Frey Graph of Clarity

```
## mean:  4.450969
## estimated sd:  2.255346
## estimated skewness:  0.2342932
## estimated kurtosis:  1.998537
```

## 1.4 Dominant Landcover

There are five types of dominant land cover; Exotic Forest, Native, Pastoral, Urban area and Other. 'Other' includes 'Gorse and/or Broom', 'Surface mines and dumps', 'Mixed exotic shrubland', and 'Transport infrastructure'. The category Urban area is applied if urban cover exceeds 15 percent of catchment area. Pastoral is applied if pastoral exceeds 25 percent of catchment area and not already assigned urban. The other three categories; Exotic forest, Native, or Other were assigned according to the largest land cover type by area, if not already assigned urban or pastoral.

Figure 23 shows side-by-side box plots of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity for each type of landcover. The observations have been log transformed to show the spread of the distributions more clearly.

Figures 24, 25, 26, 27 and 28 show comparisons for each type of landcover, for Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity, respectively.

```
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
```

Figure 24 illustrates the box plots for each type of landcover. The observations have been log transformed to show the distributions better. We can see the highest 75% of Ammoniacal Nitrogen measures in Urban areas are above the lower 75% of measures in Exotic forest and Native landcovers. This could inidcate a relationship between landcover and Ammoniacal Nitrogen. Exotic forest and Native landcovers tend to have lower amounts of Ammoniacal Nitrogen in the lake water than in Pastoral, Urban and Other landcovers. The medians are shown in tables 5, 6, 7, 8 and 9 show the median Ammoniacal Nitrogen for lakes with Exotic forest, Native, Other, Pastoral and Urban dominant landcover to be 0.0084040, 0.0075540, 0.0123870, 0.0124780 and 0.0150880, respectively. There is a clear difference between the medians, with lower medians in Exotic forest and Native landcovers.

```
## notch went outside hinges. Try setting notch=FALSE.
```

Figure 25 shows the distribution of Chlorophyll-A in each type of landcover. The observations have been log transformed to show the distributions better. We can see very similar distributions in Exotic forest, Other, Pastoral and Urban landcovers, however, lakes with Native landcover tended to have much lower levels of Chlorophyll-A than other landcovers. The medians support this, with the median Chlorophyll-A level of lakes with Native landcover being 2.8754, while all other types of landcover had medians above 5 mg per cubic meter (as shown in tables 5, 6, 7, 8 and 9).

Figure 26 shows the distribution of Total Phosphorus for each type of dominant landcover. The observations have been log transformed to show the distributions better. The Native landcover group appears to have lower levels of Phosphorus than the other landcover types, with the third quantile being below the first quantile of all other categories.

Tables 5, 6, 7, 8 and 9 show the median levels of Phosphorus are 22.610470, 12.176880, 21.836995, 21.650210 and 21.416020 for Exotic forest, Native, Other, Pastoral and Urban landcovers, respectively. We can clearly see the level of Phosphorus in lakes with Native landcover tended to be much lower than other landcover types.

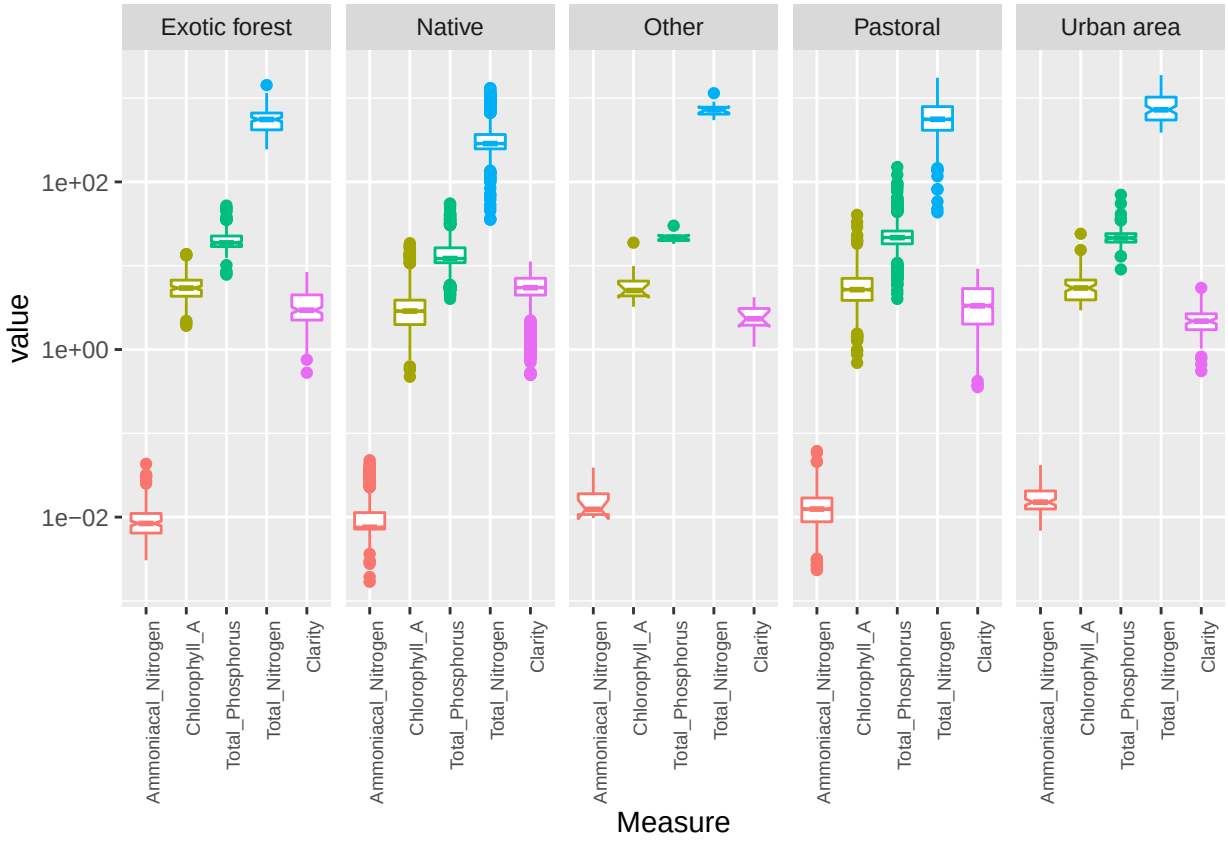



Figure 23: Box Plots of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity by Landcover

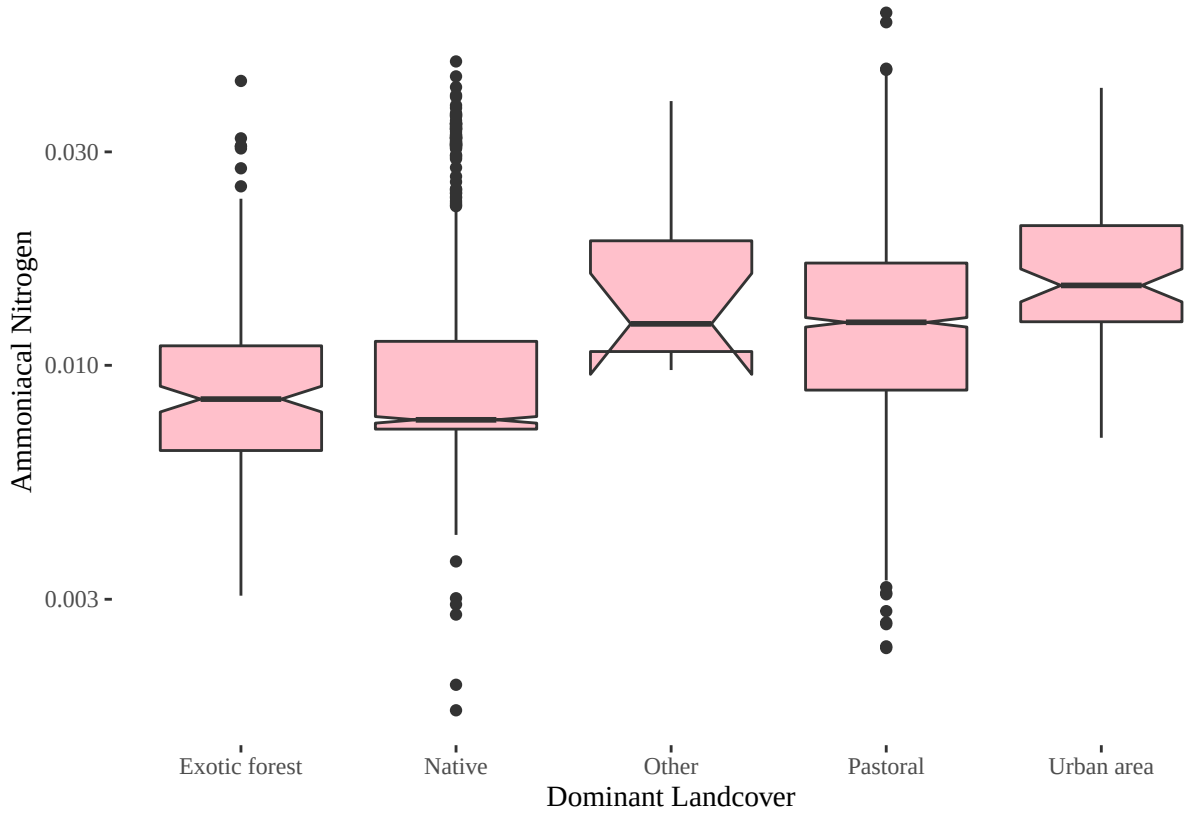



Figure 24: Box Plot of Landcover and Ammoniacal Nitrogen

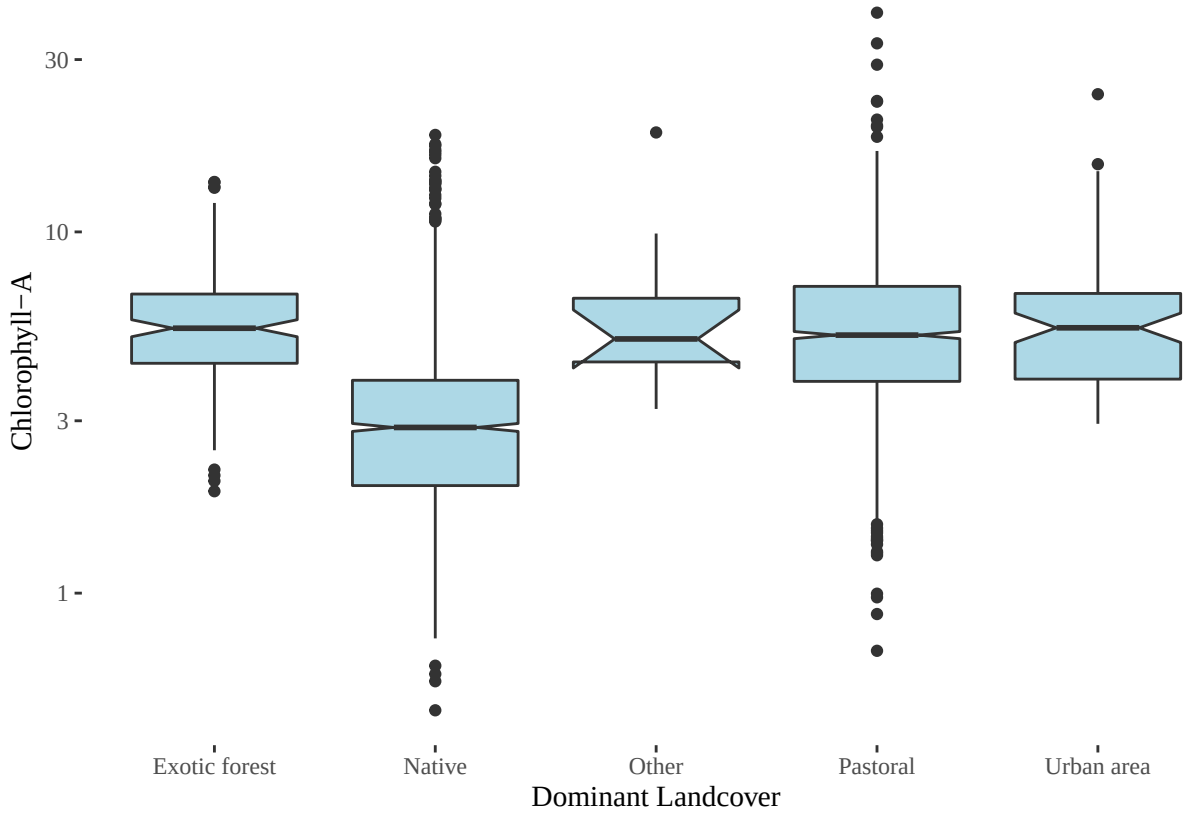



Figure 25: Box Plot of Landcover and Chlorophyll-A

```
## notch went outside hinges. Try setting notch=FALSE.
```

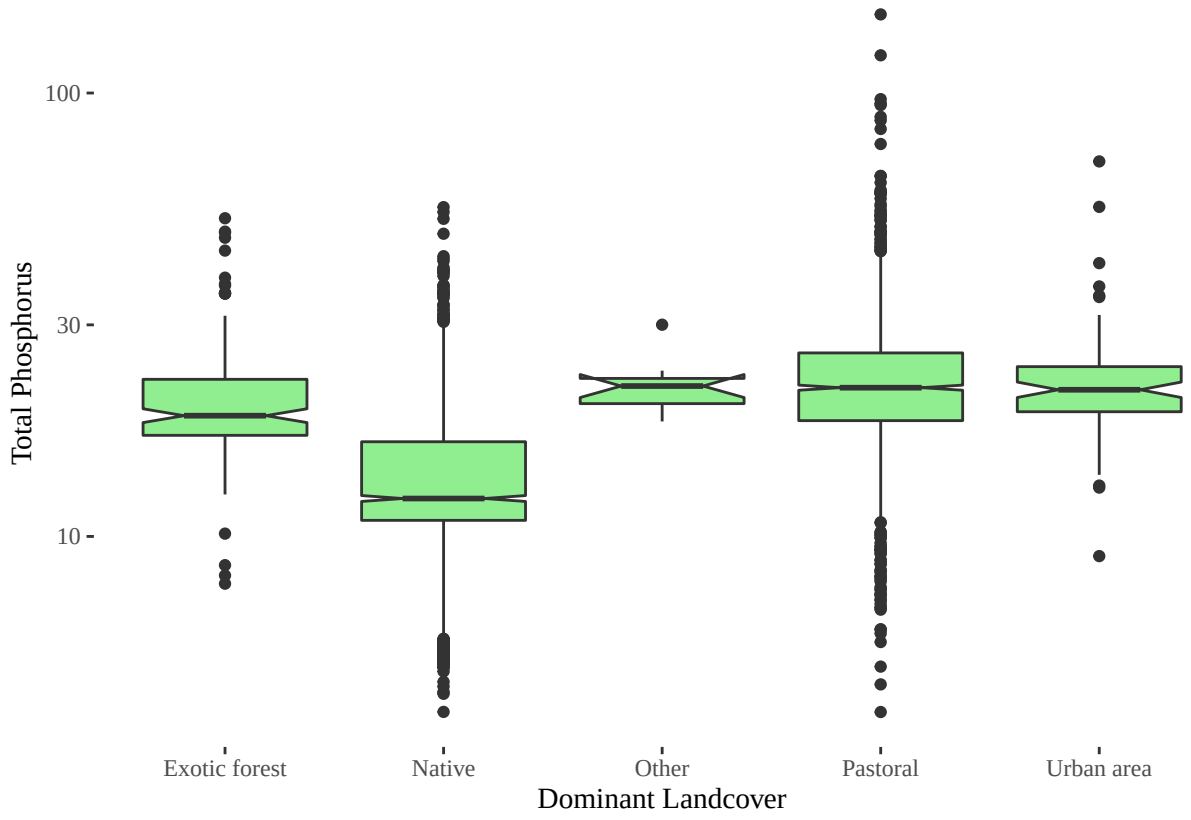



Figure 26: Box Plot of Landcover and Total Phosphorus

Figure 27 shows the distributions of Total Nitrogen for each type of dominant landcover. Similarly to Ammoniacal Nitrogen, Chlorophyll-A and Total Phosphorus, the lakes with native landcover appeared to have lower levels of Total Nitrogen. The other types of dominant landcover appear to have very similar distributions, with the distribution of Total Nitrogen in lakes with Urban landcover being slightly higher. Tables 5, 6, 7, 8 and 9 show the median levels of Nitrogen are 556.8908, 286.4589, 724.3681, 558.8725 and 725.8646 for Exotic forest, Native, Other, Pastoral and Urban landcovers, respectively. These statistics support the claim that lakes with Native landcover tend to have lower levels of Nitrogen than other types of dominant landcover.

```
## notch went outside hinges. Try setting notch=FALSE.
```

The distribution of Clarity by dominant landcover is shown in figure 28. The lakes with Native landcover appear to have a higher clarity, or appear clearer than lakes with other types of dominant landcover. Lakes with Urban or 'Other' landcover are less clear. This is to be expected as we have found that higher levels of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen are associated with lower levels of Clarity. Tables 5, 6, 7, 8 and 9 show the median levels of Clarity are 2.9450, 5.4663, 2.3282, 3.3240 and 2.1811 for Exotic forest, Native, Other, Pastoral and Urban landcovers, respectively. We can say that lakes with Native landcover tend to be clearer than lakes with other types of dominant landcover.

```
## notch went outside hinges. Try setting notch=FALSE.
```

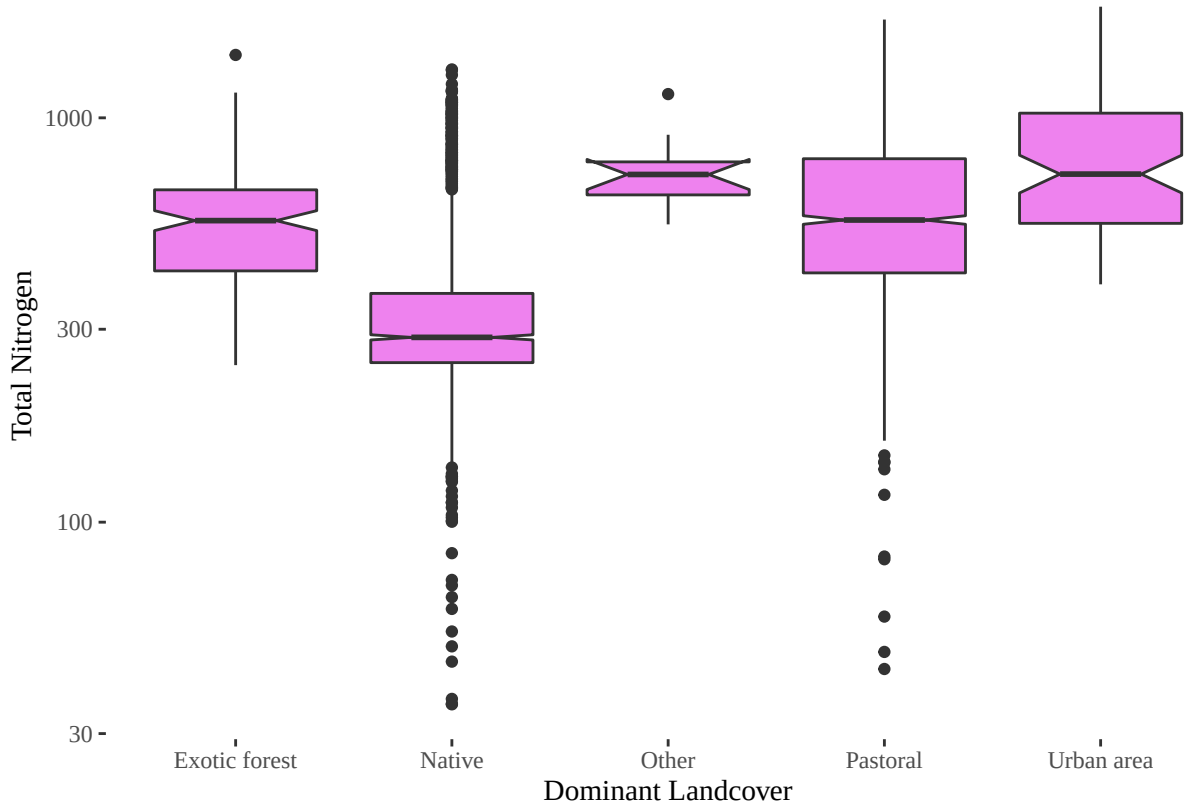



Figure 27: Box Plot of Landcover and Total Nitrogen

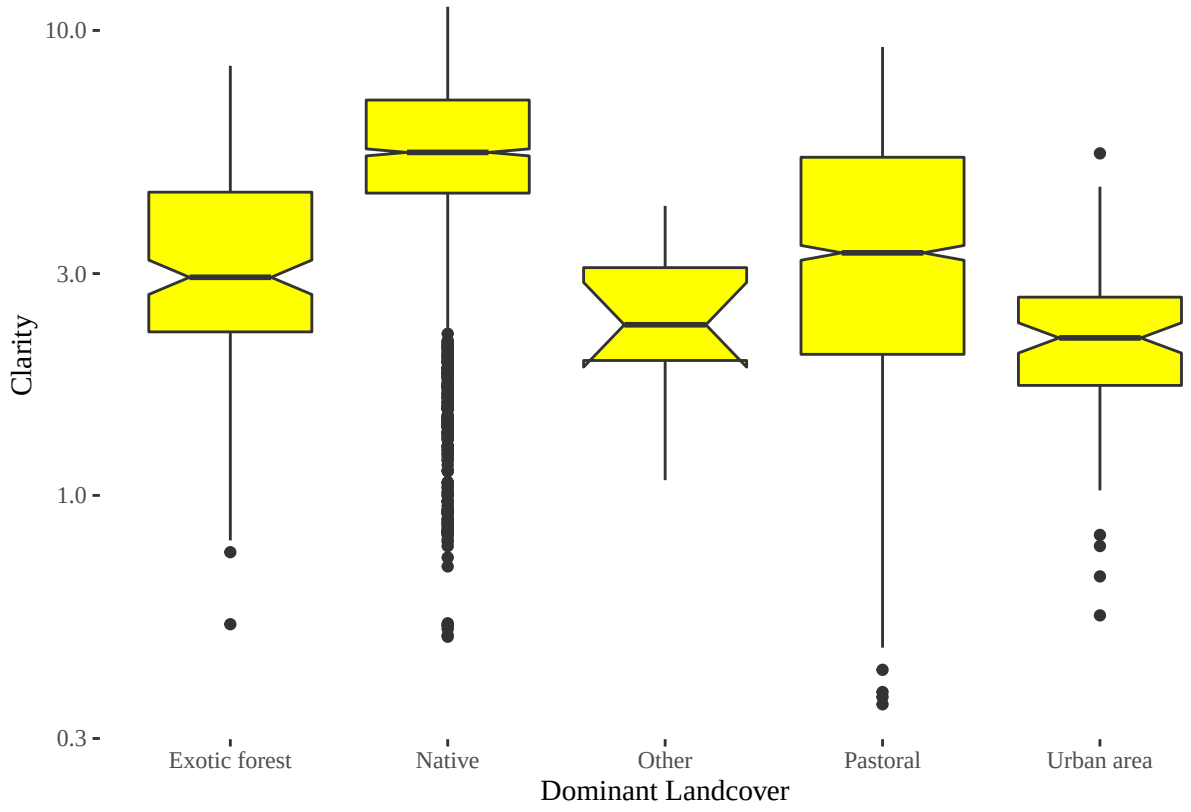



Figure 28: Box Plot of Landcover and Clarity

Table 5: Table of Sample Statistics for Exotic Forest Landcover

| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Sample Size | 165.0000000 | 165.000000 | 165.000000 | 165.000000 | 165.0000000 |
| Minimum | 0.0030560 | 1.914920 | 7.824604 | 244.552300 | 0.5283680 |
| 1st Quantile | 0.0064500 | 4.327840 | 16.903240 | 418.269100 | 2.2472620 |
| Median | 0.0084040 | 5.411983 | 18.720890 | 556.890800 | 2.9450360 |
| 3rd Quantile | 0.0110560 | 6.729298 | 22.610470 | 663.692100 | 4.4890560 |
| Maximum | 0.0432230 | 13.748320 | 52.192960 | 1430.616000 | 8.3944250 |
| Inter-quartile Range | 0.0046060 | 2.401458 | 5.707230 | 245.423000 | 2.2417940 |
| Standard Deviation | 0.0059558 | 2.251981 | 7.172119 | 205.723829 | 1.8382921 |
| Mean | 0.0101148 | 5.759535 | 20.640032 | 577.841100 | 3.4546657 |
| Median Absolute Deviation | 0.0031090 | 1.738199 | 3.746189 | 197.813978 | 1.4758809 |
| Kurtosis | 10.9583216 | 5.440503 | 8.390341 | 4.213177 | 3.0542778 |
| Skewness | 2.5252855 | 1.306315 | 2.056528 | 1.066055 | 0.8963324 |

Table 6: Table of Sample Statistics for Native Landcover

| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Sample Size | 1770.0000000 | 1770.000000 | 1770.000000 | 1770.000000 | 1770.0000000 |
| Minimum | 0.0016940 | 0.473853 | 4.020834 | 35.444730 | 0.4966520 |
| 1st Quantile | 0.0072020 | 1.984959 | 10.870083 | 248.086550 | 4.4677300 |
| Median | 0.0075540 | 2.875430 | 12.176880 | 286.458900 | 5.4663295 |
| 3rd Quantile | 0.0113143 | 3.884066 | 16.351820 | 367.934700 | 7.0846440 |
| Maximum | 0.0477830 | 18.554890 | 55.290970 | 1317.353000 | 11.2488500 |
| Inter-quartile Range | 0.0041123 | 1.899107 | 5.481738 | 119.848150 | 2.6169140 |
| Standard Deviation | 0.0053919 | 2.024765 | 6.085992 | 203.570216 | 2.0869312 |
| Mean | 0.0098571 | 3.310055 | 13.856583 | 360.579423 | 5.3440779 |
| Median Absolute Deviation | 0.0009919 | 1.376130 | 3.620259 | 69.888059 | 1.7089745 |
| Kurtosis | 14.5213007 | 15.904375 | 8.898851 | 6.930813 | 2.4458642 |
| Skewness | 3.1201691 | 2.918515 | 1.826233 | 2.035596 | -0.2370482 |

Table 7: Table of Sample Statistics for Other Landcover

| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Sample Size | 12.0000000 | 12.000000 | 12.000000 | 12.000000 | 12.0000000 |
| Minimum | 0.0097610 | 3.235230 | 18.167920 | 545.199500 | 1.0783810 |
| 1st Quantile | 0.0107397 | 4.381197 | 19.930630 | 644.760175 | 1.9537652 |
| Median | 0.0123870 | 5.055644 | 21.836995 | 724.368100 | 2.3281925 |
| 3rd Quantile | 0.0190660 | 6.590519 | 22.724430 | 778.679350 | 3.0892158 |
| Maximum | 0.0389650 | 18.851460 | 30.023000 | 1144.875000 | 4.1946350 |
| Inter-quartile Range | 0.0083262 | 2.209322 | 2.793800 | 133.919175 | 1.1354505 |
| Standard Deviation | 0.0085389 | 4.319832 | 3.147497 | 158.740455 | 0.9114810 |
| Mean | 0.0160571 | 6.517253 | 21.893479 | 745.837125 | 2.4852789 |
| Median Absolute Deviation | 0.0034952 | 1.780643 | 2.627234 | 120.245013 | 1.0849259 |
| Kurtosis | 5.3078064 | 6.712099 | 4.854776 | 4.404853 | 2.2352235 |
| Skewness | 1.7485667 | 2.135868 | 1.300069 | 1.324778 | 0.2319556 |

Table 8: Table of Sample Statistics for Pastoral Landcover

| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Sample Size | 1770.0000000 | 1770.000000 | 1770.000000 | 1770.0000000 | 1770.0000000 |
| Minimum | 0.0023340 | 0.692440 | 4.017657 | 43.3438700 | 0.3553600 |
| 1st Quantile | 0.0088040 | 3.856655 | 18.244765 | 413.5380500 | 2.0109795 |
| Median | 0.0124780 | 5.180067 | 21.650210 | 558.8725000 | 3.3239965 |
| 3rd Quantile | 0.0169287 | 7.067774 | 25.933367 | 792.1972000 | 5.3363585 |
| Maximum | 0.0614130 | 40.448870 | 150.416800 | 1749.8950000 | 9.2128900 |
| Inter-quartile Range | 0.0081247 | 3.211120 | 7.688603 | 378.6591500 | 3.3253790 |
| Standard Deviation | 0.0073402 | 2.920128 | 9.464692 | 275.1740646 | 2.1464743 |
| Mean | 0.0139015 | 5.715493 | 23.177153 | 629.6969317 | 3.7689957 |
| Median Absolute Deviation | 0.0058452 | 2.202454 | 5.377642 | 250.2305593 | 2.0776341 |
| Kurtosis | 6.6581435 | 23.631222 | 37.838715 | 2.9395897 | 2.3785691 |
| Skewness | 1.5961889 | 2.942548 | 4.067867 | 0.7974652 | 0.6671145 |

Table 9: Table of Sample Statistics for Urban Landcover

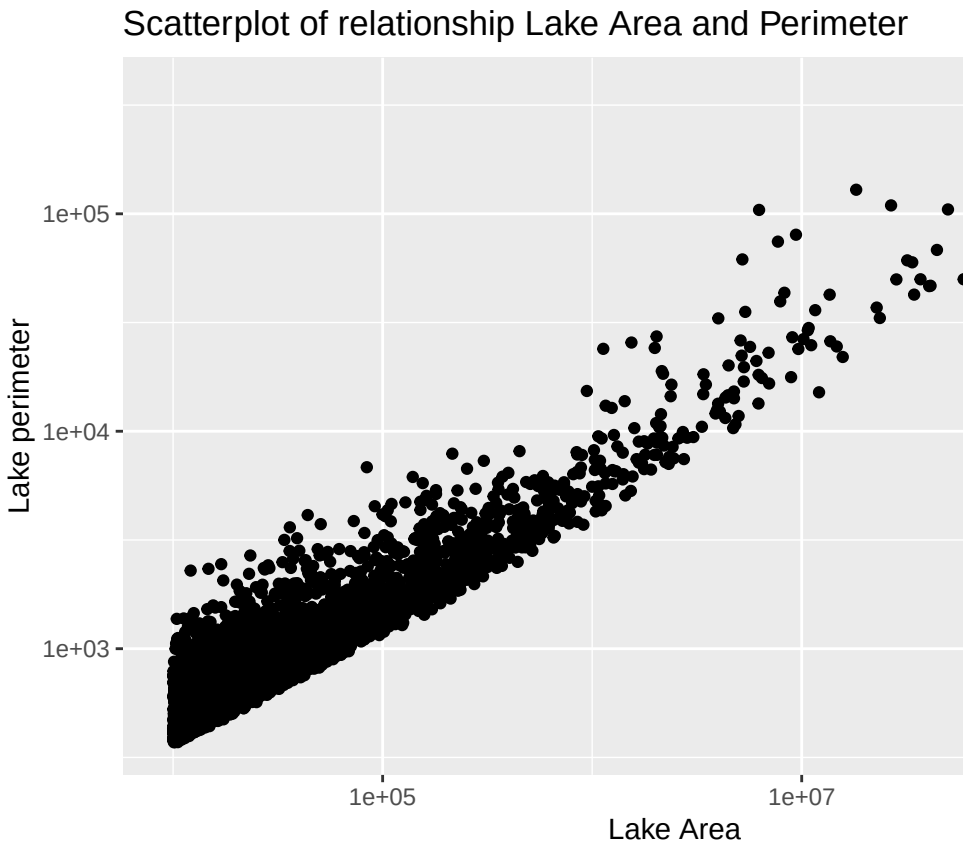
| | Ammoniacal Nitrogen | Chlorophyll-A | Total Phosphorus | Total Nitrogen | Clarity |
|---|---|---|---|---|---|
| Sample Size | 85.0000000 | 85.000000 | 85.000000 | 85.0000000 | 85.0000000 |
| Minimum | 0.0068880 | 2.942836 | 9.029183 | 387.4383000 | 0.5520370 |
| 1st Quantile | 0.0125150 | 3.912775 | 19.107640 | 548.3017000 | 1.7239980 |
| Median | 0.0150880 | 5.427137 | 21.416020 | 725.8646000 | 2.1811430 |
| 3rd Quantile | 0.0205250 | 6.755165 | 24.147820 | 1027.0020000 | 2.6690640 |
| Maximum | 0.0417070 | 24.052490 | 70.108590 | 1883.1720000 | 5.4404300 |
| Inter-quartile Range | 0.0080100 | 2.842390 | 5.040180 | 478.7003000 | 0.9450660 |
| Standard Deviation | 0.0083745 | 3.389789 | 8.392760 | 286.3296449 | 0.8785109 |
| Mean | 0.0180000 | 6.126680 | 23.030843 | 778.8134400 | 2.2658755 |
| Median Absolute Deviation | 0.0056546 | 2.205781 | 3.686455 | 317.4139853 | 0.7210136 |
| Kurtosis | 3.4495705 | 11.658444 | 15.327134 | 4.1074614 | 4.5139920 |
| Skewness | 1.1230235 | 2.512231 | 2.892799 | 0.9219365 | 0.8943855 |

### 1.4.1 Investigating lake area, depth, and perimeter

What we next wanted to investigate was if there was any noticeable difference between area and perimeter as these dimensions are strongly linked. We know from mathematical formulas such as the formula for a circle that there is a relationship between area and perimeter.

As such, we look at the correlation between these two variables. We have decided that if the correlation is greater than 0.95 than the variables are practically the same and can be used interchangeably. If not, then they will be used separately.

Because of the high asymmetry of lakes, we would expect that lakes with the same perimeter can have very different areas. As such, we are expecting there will be a very noticeable difference.




Below is a plot of the correlation between

It is difficult to see the distribution from a regular plot as we have some large outliers. As such, I plotted the above graph using x and y scaling of log10 to improve the display. The data shows a clear correlation however because of the log transformations the strength of the correlations may not be as strong as they appear in the above graph. What We can also tell is that there appears to be many "small" lakes relative to a handful of much much larger ones. This informs us for when we later examine Regions to expect large outliers to skew the distribution.

Correlation output for Lake area and perimeter:

```
##                      lake.lake_area lake.lake_perimeter
## lake.lake_area               1.000               0.798
## lake.lake_perimeter          0.798               1.000
```

The correlation matrix returns a value of 0.797 for the correlation between Lake area and perimeter. This result makes logical sense as lakes with the same perimeter can have drastically different areas.

From this analysis we know not to treat area and perimeter as equivalent.

### 1.4.2 Lake Data Summary Statistics:

Number of Lakes in each Region:

Table 10: Table of Sample Statistics for Urban Landcover

| LakeCat | Freq |
|---|---|
| Auckland | 74 |
| Bay of Plenty | 98 |
| Canterbury | 463 |
| Gisborne | 30 |
| Hawke's Bay | 285 |
| Marlborough district council | 50 |
| Northland | 262 |
| Otago | 378 |
| Southland | 991 |
| Taranaki | 90 |
| Tasman district council | 57 |
| Waikato | 233 |
| Wellington | 107 |
| West Coast | 457 |

Table of summary statistics of Area (m^2), Perimeter(m), Depth(m)

Table 11: Table of Sample Statistics for Urban Landcover

| | Area | Perimeter | Depth |
|---|---|---|---|
| Minimum | 10004.94 | 371.160 | 1.000 |
| Maximum | 613000000.00 | 369677.800 | 462.000 |
| Median | 24766.56 | 831.540 | 17.600 |
| First_Quartile | 14231.34 | 588.405 | 14.645 |
| Third_Quartile | 65144.91 | 1431.375 | 23.465 |

#### 1.4.2.1 Variance in our Area (m^2), Perimeter(m) and Depth(m)

#### 1.4.2.2 Second Moment of Variances

```
## [1] "Area:"           "220273515342803"

## [1] "Perimeter:"       "134064549.514477"

## [1] "Depth:"           "578.988707757268"
```

#### 1.4.2.3 Skewness in our Area (m^2), Perimeter(m) and Depth(m)

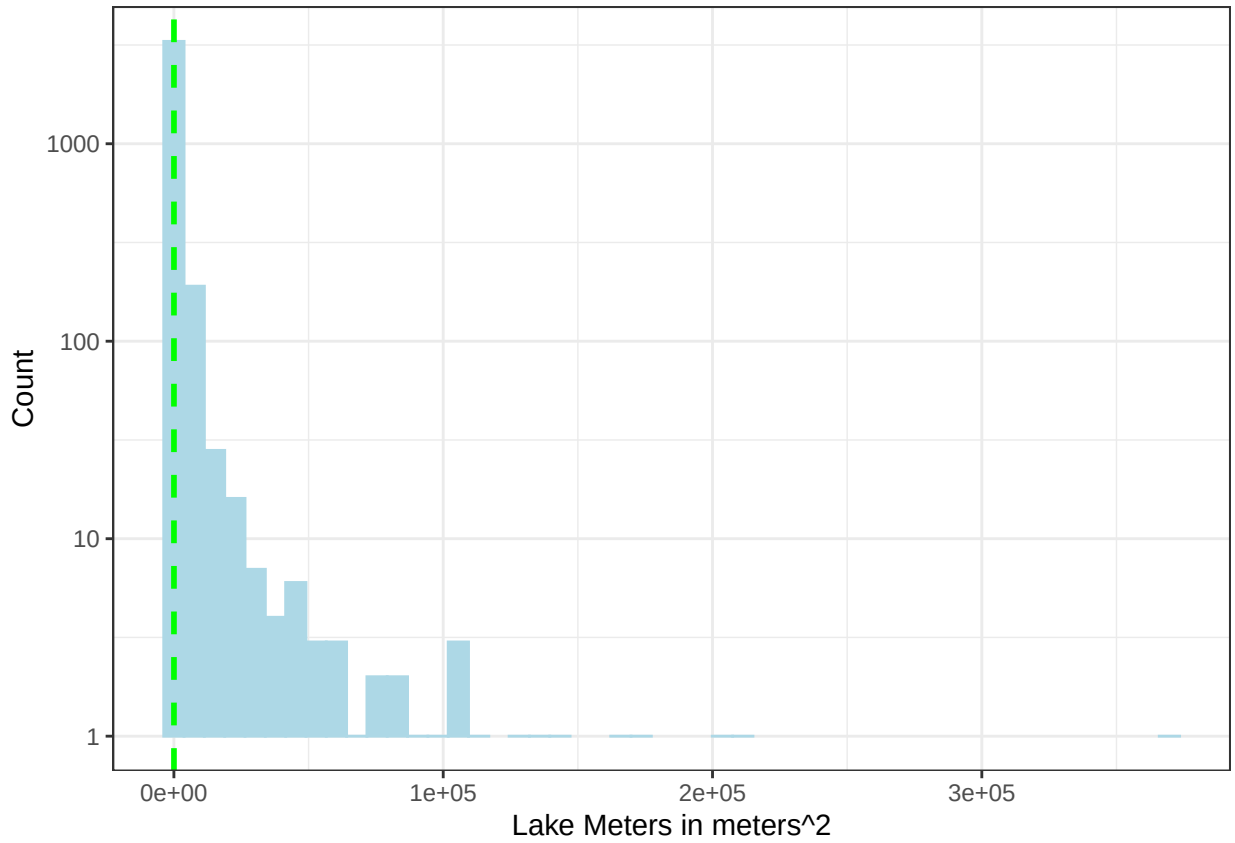
```
## lake_perimeter     lake_depth      lake_area
##      16.928070       9.622475      27.907339
```

Our positive skews indicate that data is skewed right and is therefore not normally distributed around the mean; where high skewness is greater than 1, moderate skewness is between 0.5 and 1, minor skewness is between 0 and <0.5. As such we see that all three variables have a huge amount of skew. To visualize this skew and to see the distribution of our data we will create histograms.

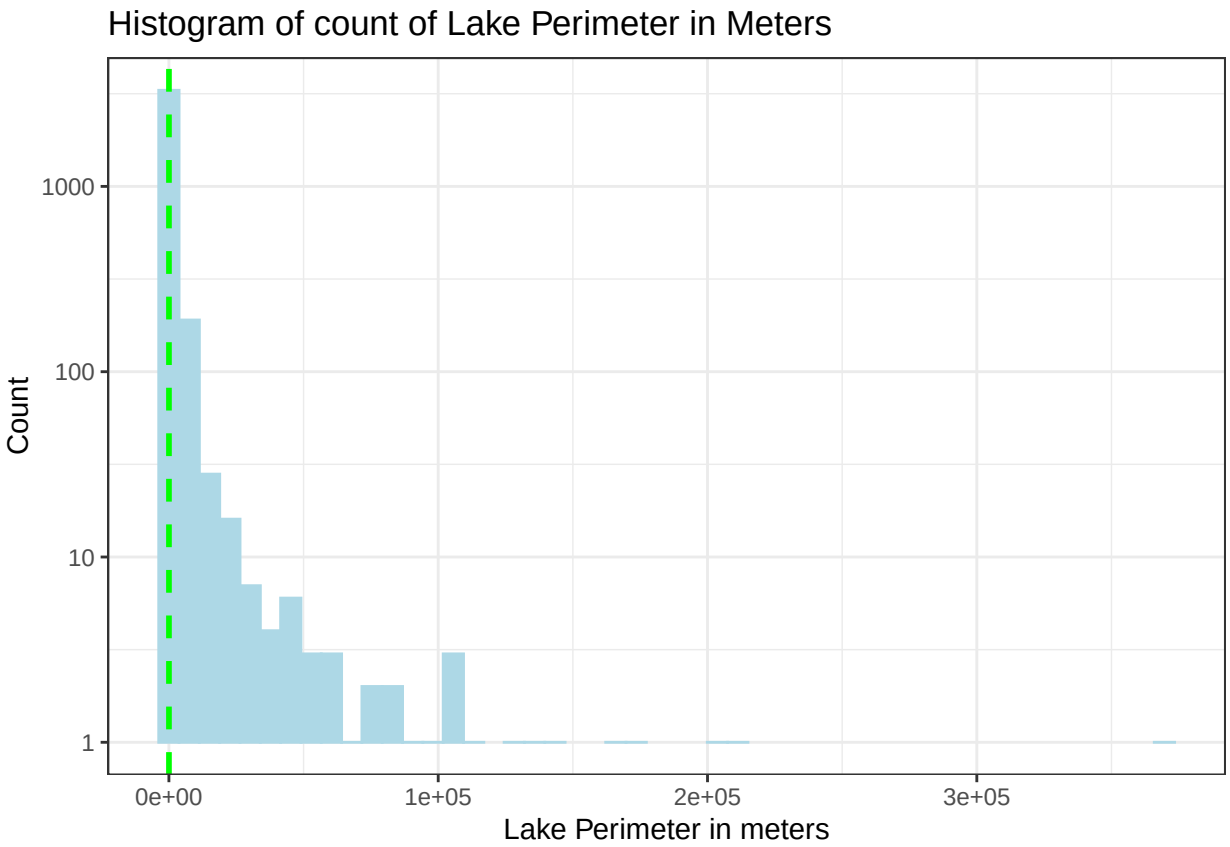
Histogram for Area figure **??** Histogram for Perimeter figure **??** Histogram for depth figure **??**

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

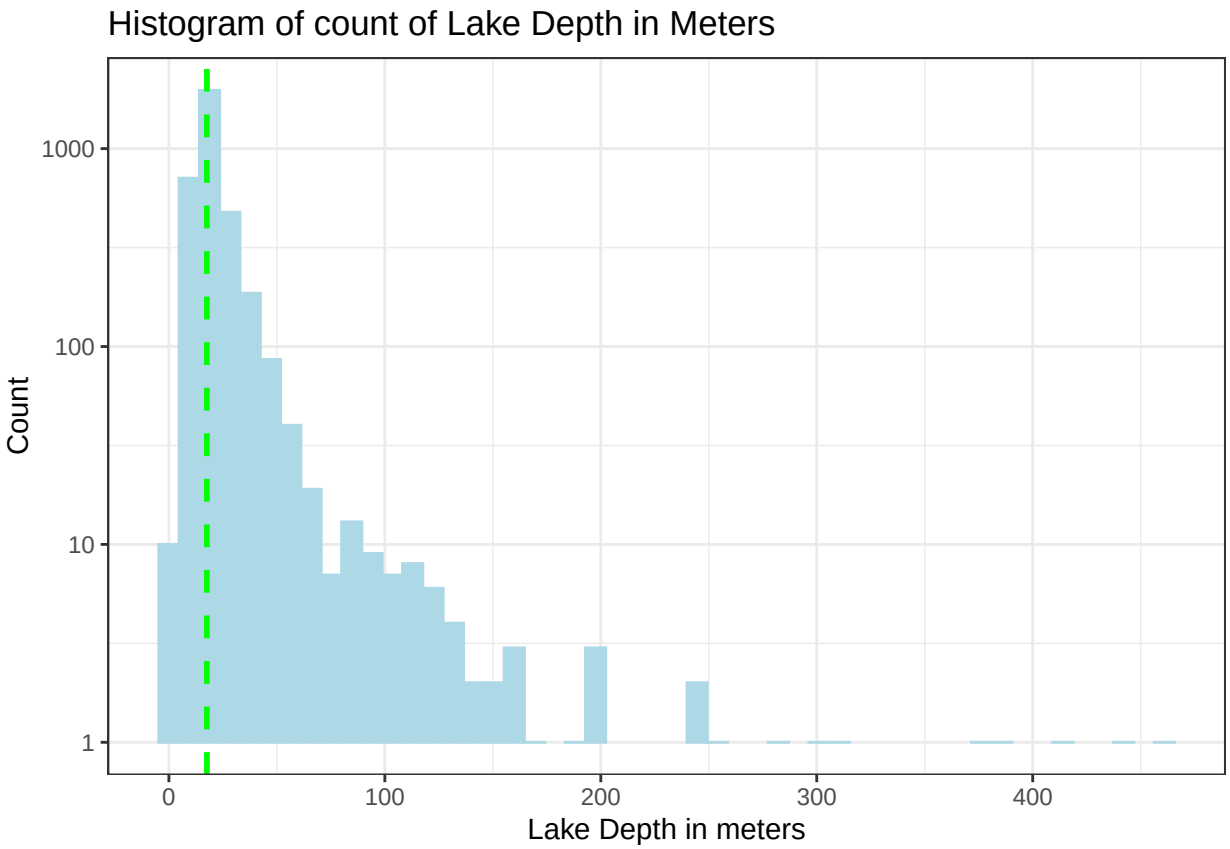
```
## Warning: Removed 26 rows containing missing values (geom_bar).
```




*Above Histogram includes y scaling of log10 (+scale_y_log10()) as it provides a better visual display of the data.




*Above Histogram includes y scaling of log10 (+scale_y_log10()) as it provides a better visual display of the data.



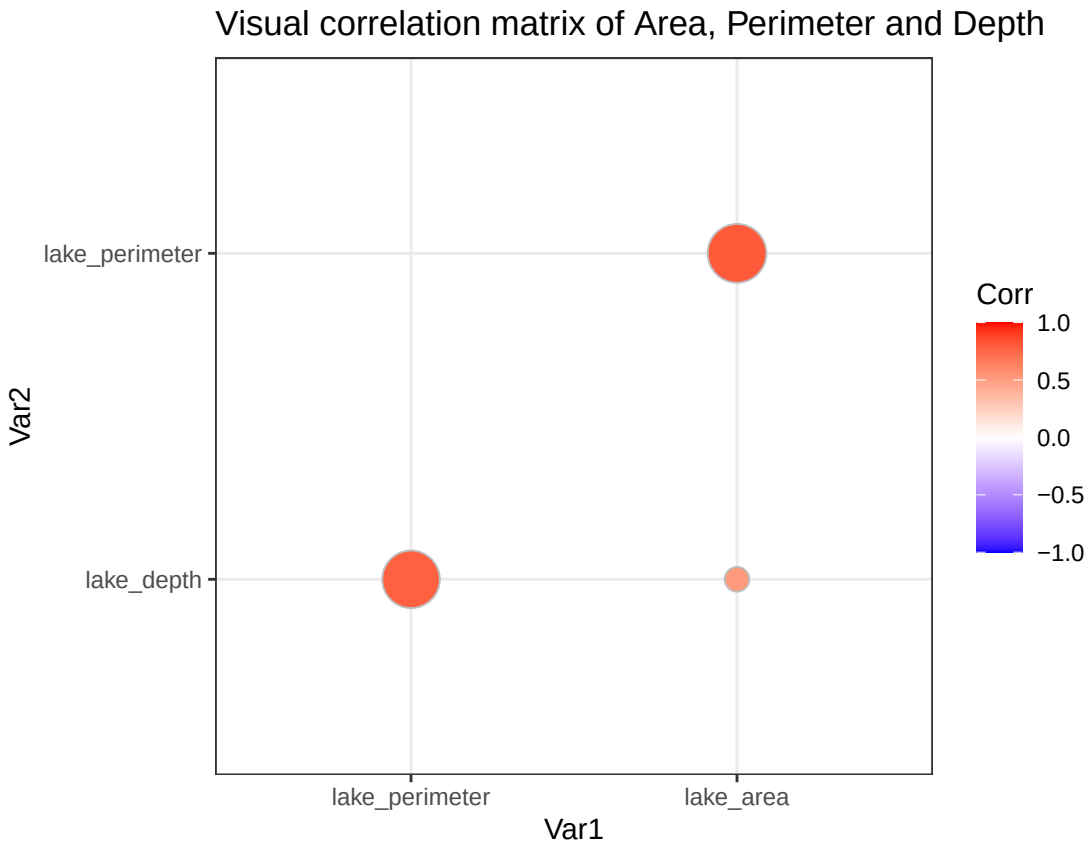

*Above Histogram includes y scaling of log10 (+scale_y_log10()) as it provides a better visual display of the data.

### 1.4.3 Examining the Correlation between Area, Depth and Perimeter

*Correlation matrix rounded to three decimal place

```
##                lake_perimeter lake_depth lake_area
## lake_perimeter          1.000      0.781     0.798
## lake_depth              0.781      1.000     0.520
## lake_area               0.798      0.520     1.000
```

Visualisation of the Correlation Matrix:




As to be expected, all three variables have positive correlations with one another. The correlation between lake area and lake depth isn't small but it is smaller than one may expect. Lake Perimeter has a rather strong relationship with both area and depth.

## 1.5 Investigating Regional differences

We first investigated Regional differences in our Lake Health variables.

Table 12: Table of Sample Statistics for Urban Landcov

| Region | Median_Trophic_Level3 | Median_Clarity_metres | Median_NH4N_mg | Median_T |
|---|---|---|---|---|
| Auckland | 3.939901 | 2.596739 | 0.0115450 | |
| Bay of Plenty | 4.050533 | 4.183662 | 0.0098540 | |
| Canterbury | 3.830216 | 5.743230 | 0.0131270 | |
| Gisborne | 4.359285 | 2.757868 | 0.0132765 | |
| Hawke's Bay | 4.224237 | 3.416575 | 0.0096720 | |
| Marlborough district council | 4.147844 | 4.020992 | 0.0097970 | |
| Northland | 4.023861 | 2.803300 | 0.0075225 | |
| Otago | 3.645382 | 5.727244 | 0.0120175 | |
| Southland | 3.539076 | 4.616884 | 0.0074620 | |
| Taranaki | 4.213852 | 3.109920 | 0.0101145 | |
| Tasman district council | 3.553968 | 5.709074 | 0.0072210 | |
| Waikato | 4.499246 | 2.461897 | 0.0123660 | |
| Wellington | 4.263893 | 1.734104 | 0.0134980 | |
| West Coast | 3.756785 | 4.601896 | 0.0087230 | |

| Region | Median_Trophic_Level3 | Median_Clarity_metres | Median_NH4N_mg | Median_T |
|---|---|---|---|---|
| Examining the above table there | are some interesting po | ints: | | |
| - Southland lakes have a low v | alue or the lowest value | on every metric (with t | he exception clar | ity where |
| - We see that the opposite is | true for Gisborne where | it is high in all variab | les except clarit | y w |
| - There is quite a wide range | of levels of clarity bet | ween regions suggesting | that region may h | ave a rel |

### 1.5.1 Investigating Regional differences of Lake Area, Perimeter and Depth

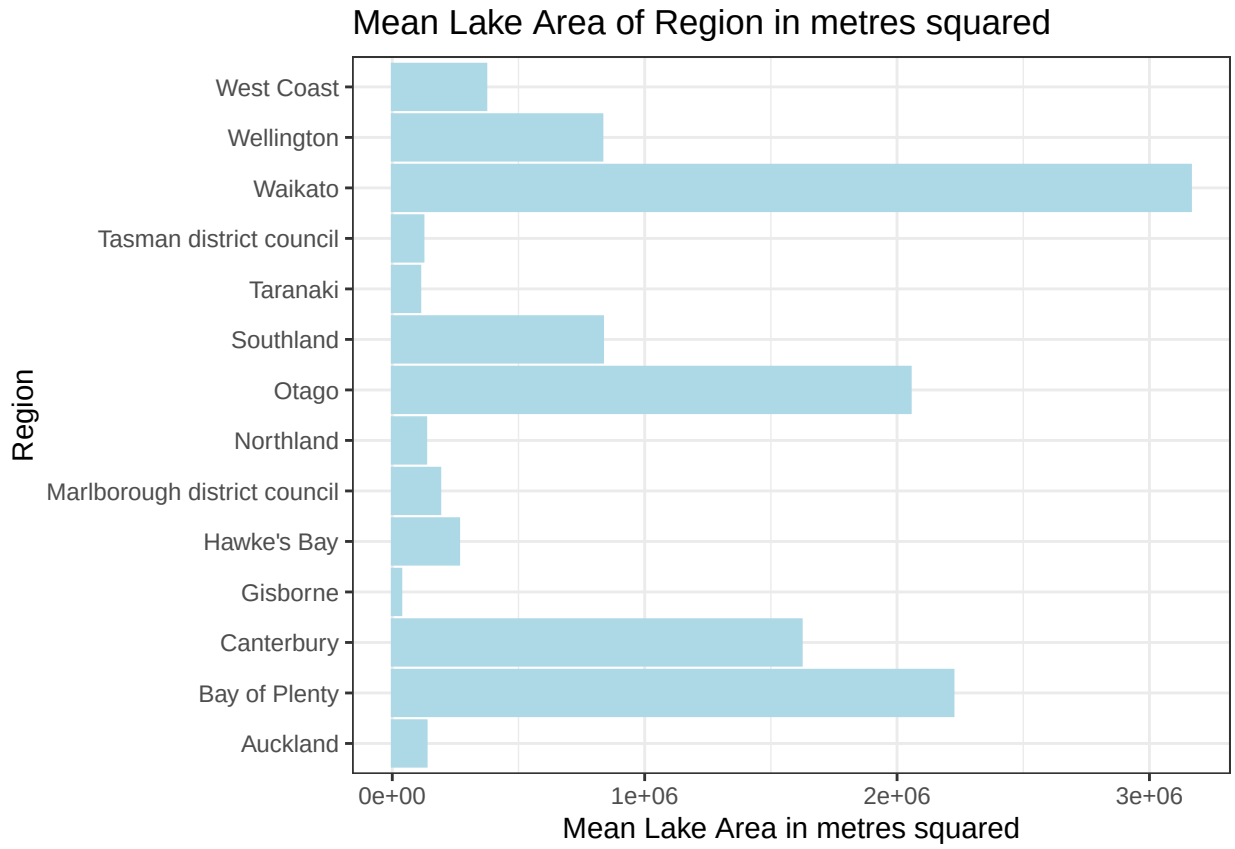



Figure 29: Barplot of regional mean lake areas

#### 1.5.1.1 View of the Mean and Median for Lake Area by Region:

**Values of NA have been removed.

**Values of NA have been removed.

Analysis of Lake Area: Of all of the Regional disparities we see the greatest in Area. There are some enormous Lakes in New Zealand and these skew the distribution. Judging by the median we would guess that the regions with the largest lakes by area are Southland and Bay of Plenty

#### 1.5.1.2 View of the Mean and Median for Lake Perimeter:

**Values of NA have been removed.

**Values of NA have been removed. Perimeter Analysis: Across all regions there is a disparity between the mean and the median. Most Regions have an median perimeter between 700 and 950 meters, however, most of their means are between 1000 and four thousand meters.

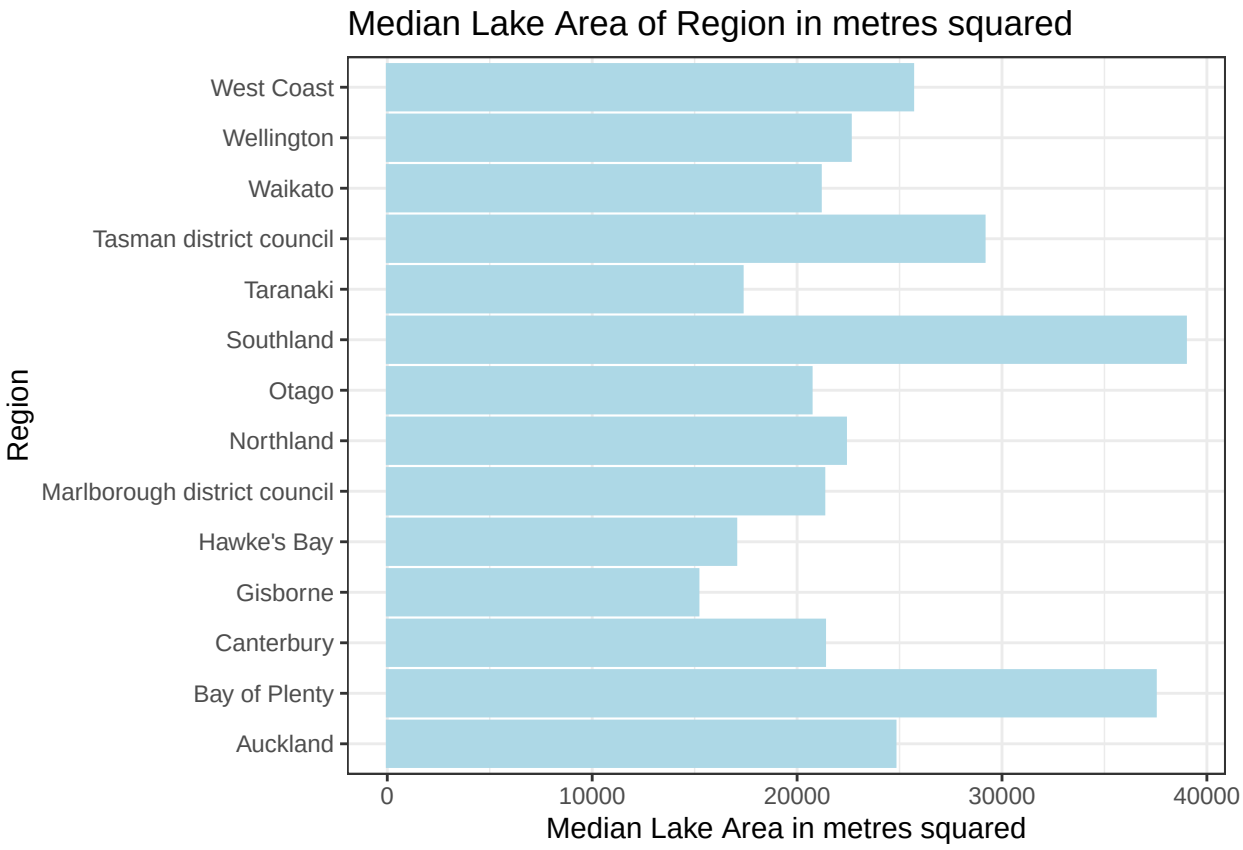



Figure 30: Barplot of regional median lake areas

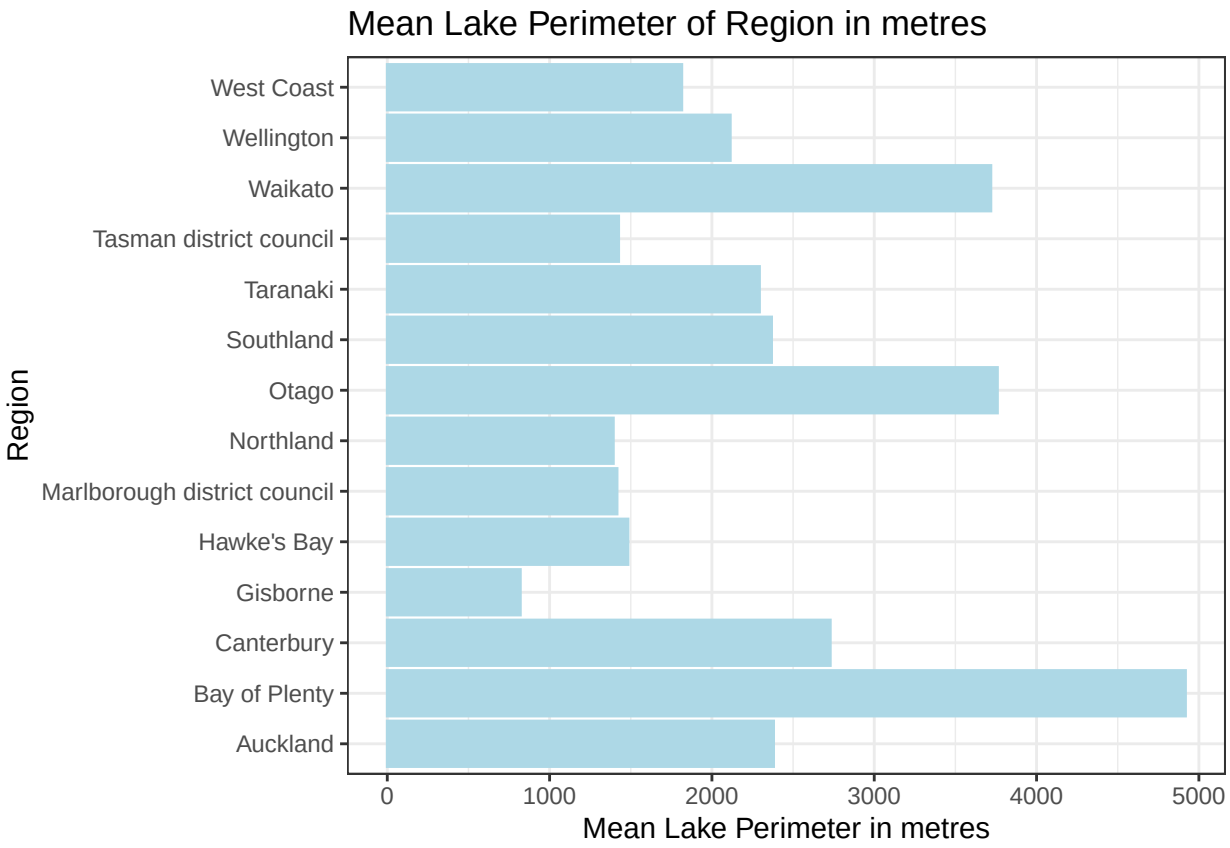



Figure 31: Barplot of regional mean lake perimeters

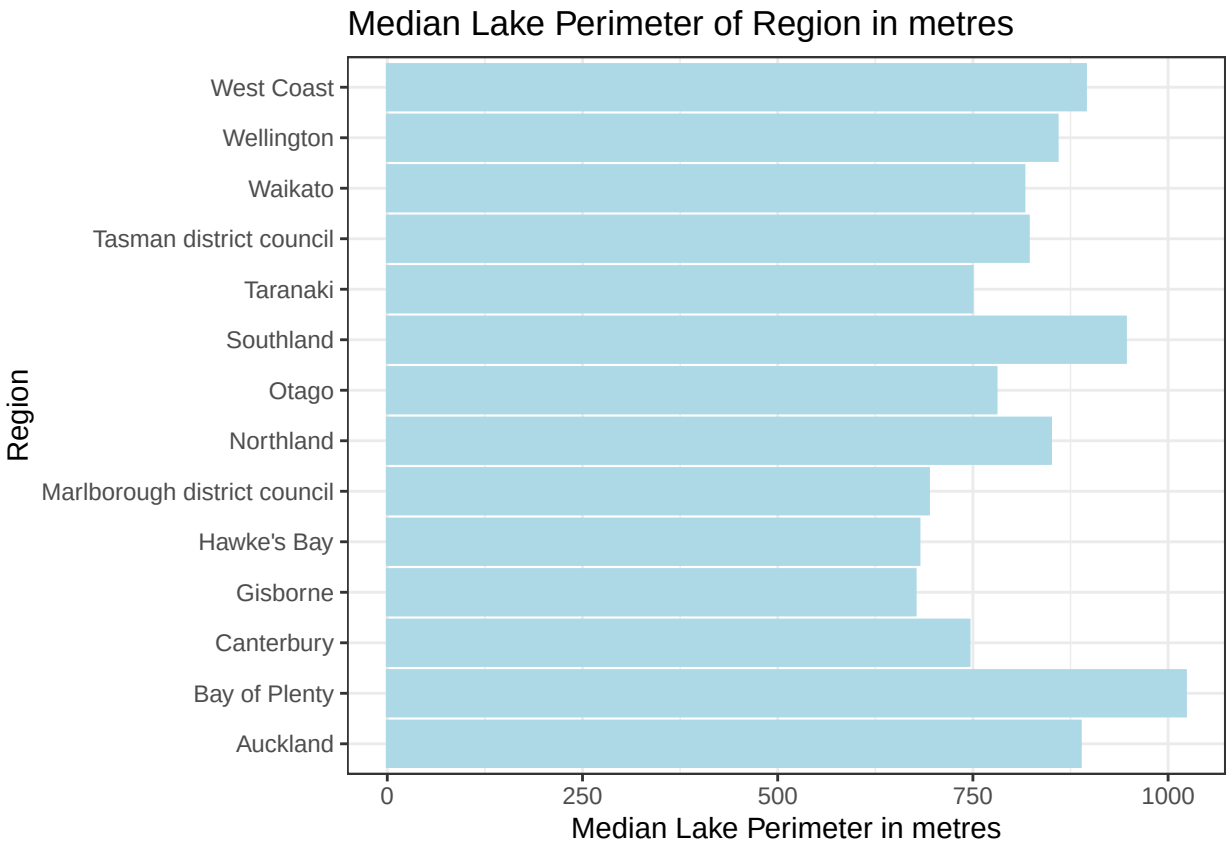



Figure 32: Barplot of regional median lake perimeters

We see that on by median and average, the Bay of Plenty has the largest perimeter with Gisborne having the smallest in both categories.

#### View of the Mean and Median for Lake Depth:

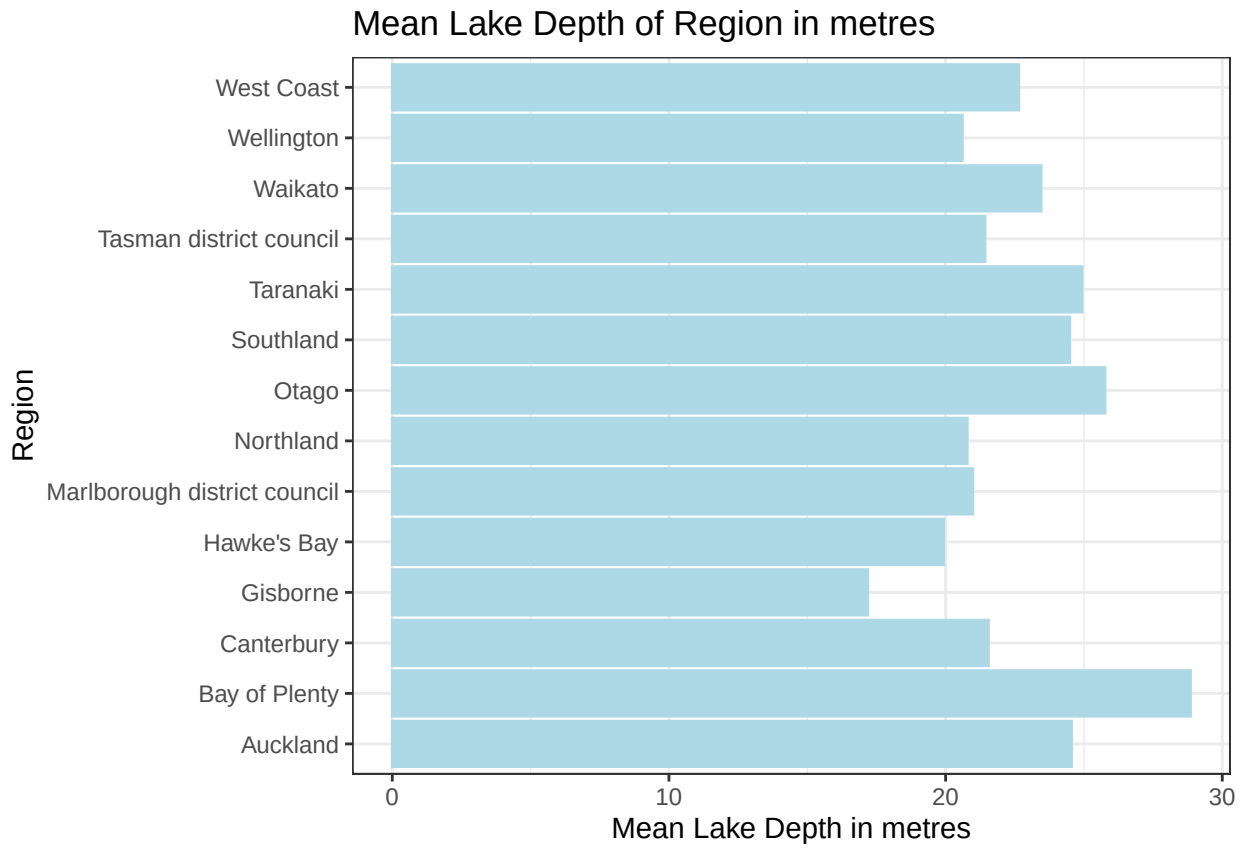



Figure 33: Barplot of regional mean lake Depth

**Values of NA have been removed.

```
## # A tibble: 14 x 3
##    Region                        mean median
##    <chr>                        <dbl>  <dbl>
##  1 Auckland                      24.6   18.1
##  2 Bay of Plenty                 28.9   19.6
##  3 Canterbury                    21.5   16.4
##  4 Gisborne                      17.2   15.9
##  5 Hawke's Bay                   19.9   16.0
##  6 Marlborough district council  21.0   16.1
##  7 Northland                     20.8   17.5
##  8 Otago                         25.8   17.1
##  9 Southland                     24.5   19.0
## 10 Taranaki                      24.9   16.7
## 11 Tasman district council       21.4   17.7
## 12 Waikato                       23.5   17.3
## 13 Wellington                    20.6   17.6
## 14 West Coast                    22.6   18.4
```

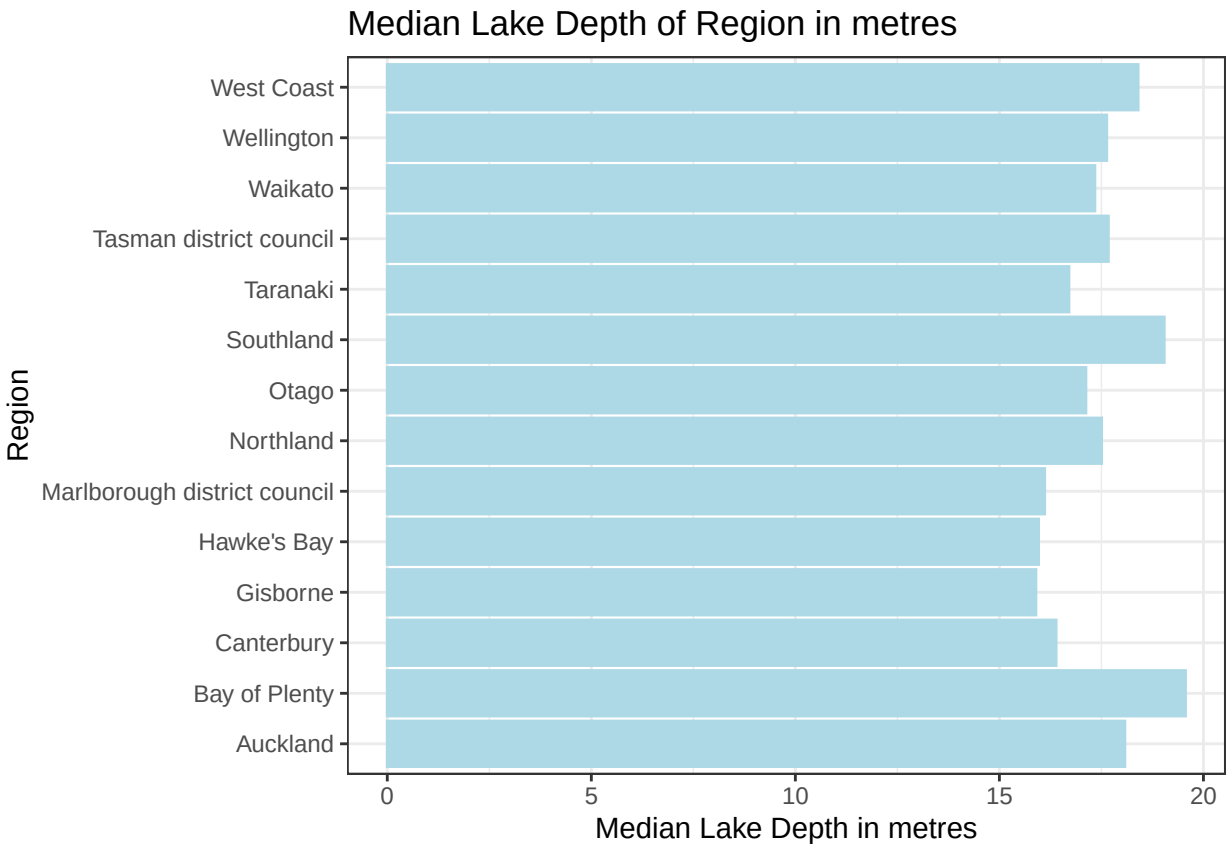



Figure 34: Barplot of regional median lake depths

**Values of NA have been removed.

Depth: What we can tell from the bargraphs that there is not much difference in distribution between the regions in Lake Depth, however, the mean for each region is approximately 5 meters deeper than the median. This suggests that across most regions there are some much deeper lakes. There are some exceptions to this such as Gibsborne.

## 1.6 Further Analysis we would like to do:

From this EDA we have identified the following items that we would like to investigate further:

- The relationship between Region and Lake Health indicators. We expect that we may find much stronger predicting ability if we include covariates when predicting Health indicators. This will be especially interesting if we can locate data from outside our dataset to explain why there may be such a disparity in Regional differences.
- The distribution of the Lake Health variables.
- The relationship between depth and the Lake Health variables. We suspect that depth may have more of an impact on the Lake Health variables than Area and Perimeter. We want to inspect this suspicion further.