Final Report - Health of New Zealand Lakes

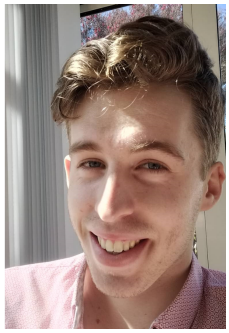
Russell and Frances

5th October 2022

Contents

1 Introduction

We are Russell and Frances and we form Group 8. Below are our pictures as well as our contact details and ORCID ID numbers.



Frances Smith
email: frances.j.smith.nz@gmail.com
ORCHID ID: 0000-0002-5168-3134

Russell Syder
email: russellsyder@gmail.com
ORCHID ID: 0000-0002-4582-5909

1.1 Introduction to our data

Our original dataset was extracted from Stats NZ. https://www.stats.govt.nz/indicators/modelled-lake-water-quality/

We made some manipulations to the original dataset such as removing extraneous variables (like Date which was the same for every lake) or variables that we were not interested in. We also reshaped some aspects of the dataset so that the data would be easier to analyse. We checked for missing values; there were some for the categorical variables, which we dealt with as described below, but there were no missing values for the numerical variables so imputation was not necessary.

Our final dataset contains information from 3802 lakes (exceeding one hectare) in New Zealand measured across 22 variables. Of these 22 variables we selected 10 that we would use for further analysis. They are as follows:

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. It acts as as measure of toxicity.

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.

Total Nitrogen is the sum of all nitrogens found in the water, including organic nitrogen from plant tissue. An excess of nitrogen in lakes can cause an increase in algae and plant growth, possibly depriving the lake of oxygen. Total Nitrogen is measured in milligrams per cubic metre and the national bottom line for stratified lakes is 750mg/m3, and for polymictic lakes is 800mg/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.

Area is the surface area of the lake measured in metres squared.

Perimeter is the overall perimeter of the lake, in metres.

Lake Depth is the maximum depth of the lake measured in metres.

Dominant Landcover is split into four types; Exotic Forest, Native, Pastoral and Urban area. There are 12 lakes with no entry for Dominant Landcover, 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 landcover exceeds 25 percent of catchment area, if the lake has not already been 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.

Upon first examining our data we thought that it would be prudent to group certain similar variables together for analysis. Specifically, the 4 variables that gave a measure of the levels of a given substance in a lake, and additionally clarity, we grouped as the "Lake Health variables" as for all of them, high levels of any of these variables can indicate poor lake health, with the exception of clarity, where, in general, higher values indicate better lake health.

We also grouped together lake Area, Perimeter, and Depth and classified this group as the Lake Dimension variables.

1.2 Leading Question

Our leading question was; What are some statistics that we can produce that may be beneficial for informing restorative actions that improve the health of lakes in New Zealand?

To investigate this we came up with the following questions;

- Are there any particular regions that have poor lake health?
- Do the Lake Health variables predict one another?
- How can we model the Lake Dimension variables?
- Do any types of Dominant Landcover have poorer lake health than others?

2 Methodology

To answer these questions, we will conduct the following investigations:

- An Exploratory Data Analysis on the Lake Health, Lake Dimension, Region and Dominant Landcover variables, which will consist of:
  - Univariate analysis of the Lake Health and Lake Dimension variables
  - Multivariate analysis with each of these four types of variables, specifically:
    - Relationship between the Lake Health and Lake Dimension variables, and
    - Comparisons of the Lake Health and Lake Dimension variables by Region and Dominant Landcover
- Tests for difference in means of the Lake Health variables by Dominant Landcover
- Principal Component Analysis on the Lake Health variables
- Factor Analysis on the Lake Health Variables
- Linear Discriminant Analysis on the Lake Health variables by Dominant Landcover

The latter four analyses will be in our Results section.

2.1 Exploratory Data Analysis

2.1.1 Univariate Analysis of the 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
Standard Deviation	0.0068358	2.807067	9.143676	277.994471	2.2553455
Mean	0.0119528	4.609290	18.720584	505.860630	4.4509687
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. A small proportion of the lakes exceeded the national bottomline of 12mg per cubic metre.

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 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. Few of the lakes exceeded the national bottomline of 50mg per cubic metre.

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. A significant proportion of the observations exceed the higher of the two national bottomlines, of 800mg per cubic metre.

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.

We 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.

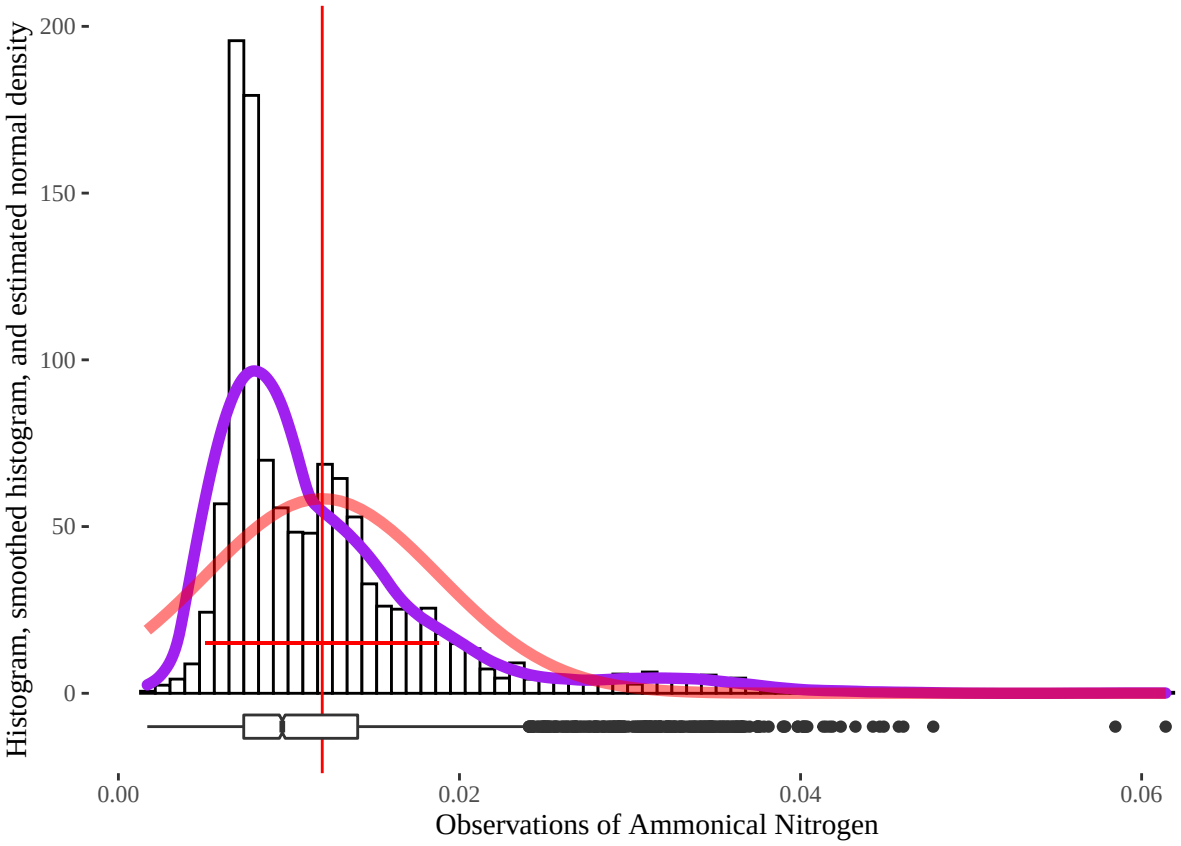


Figure 1: Histogram of Ammoniacal Nitrogen

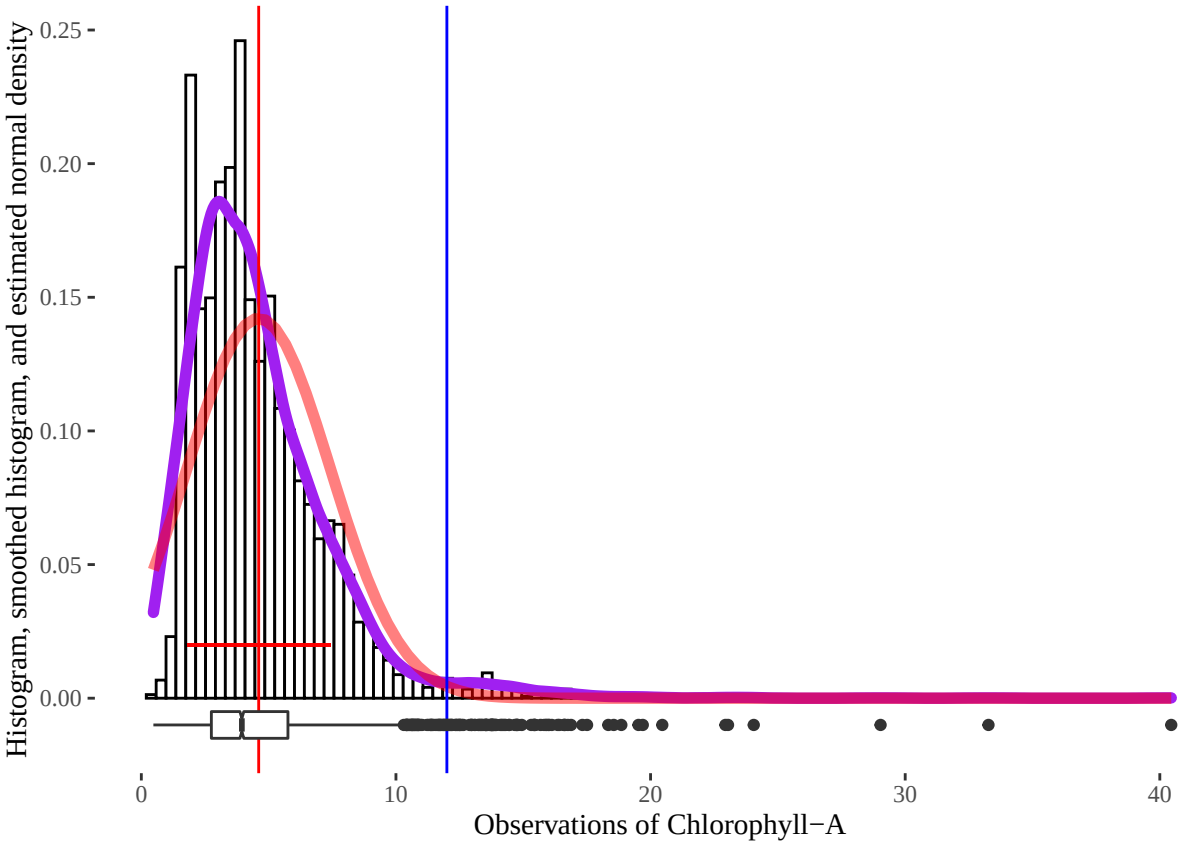


Figure 2: Histogram of Chlorophyll-A

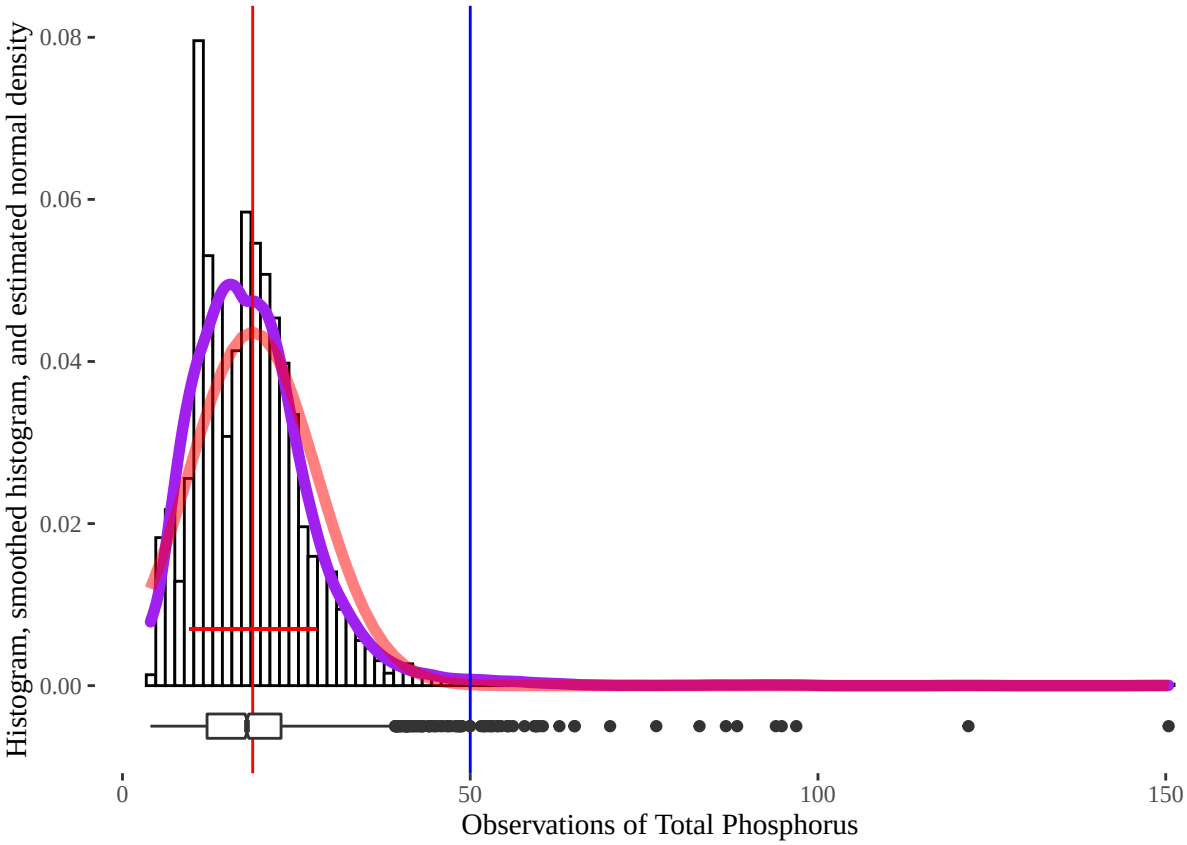


Figure 3: Histogram of Total Phosphorus

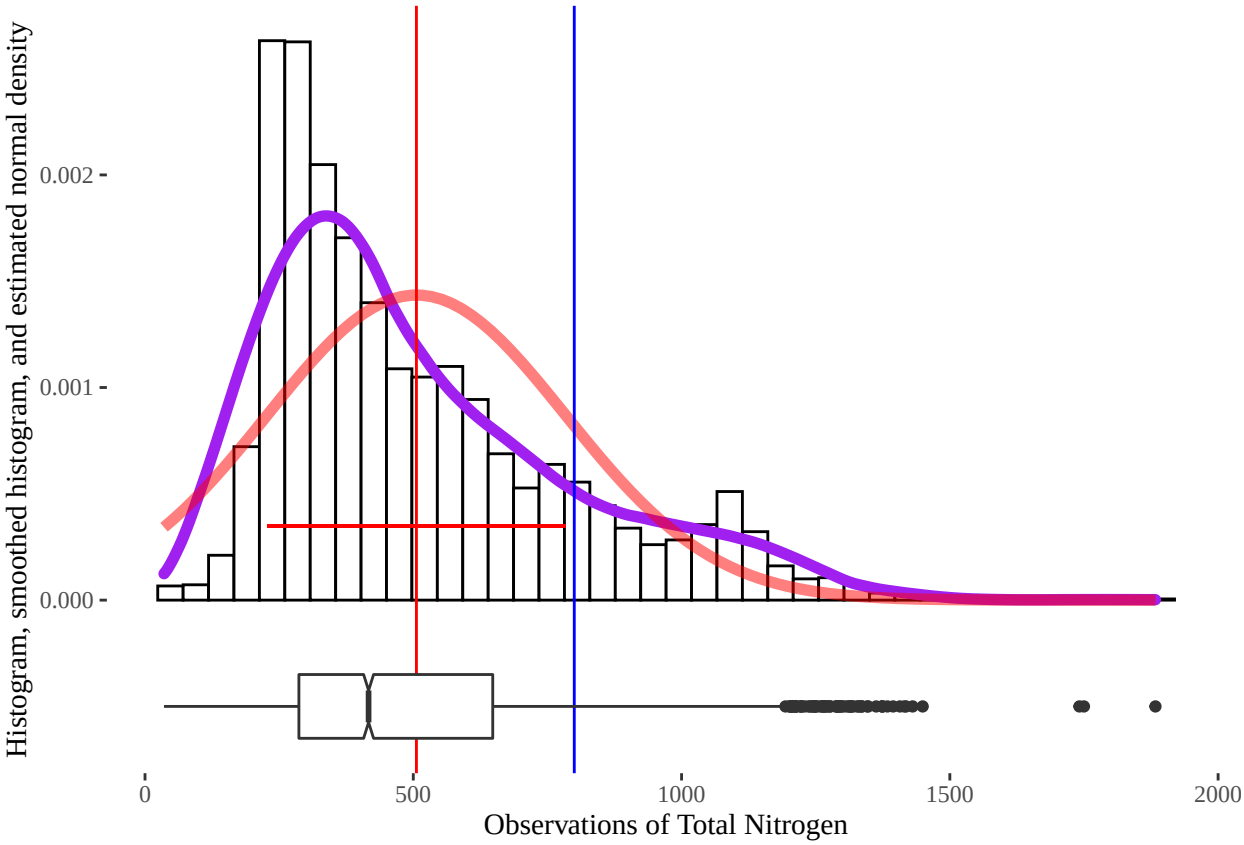


Figure 4: Histogram of Total Nitrogen

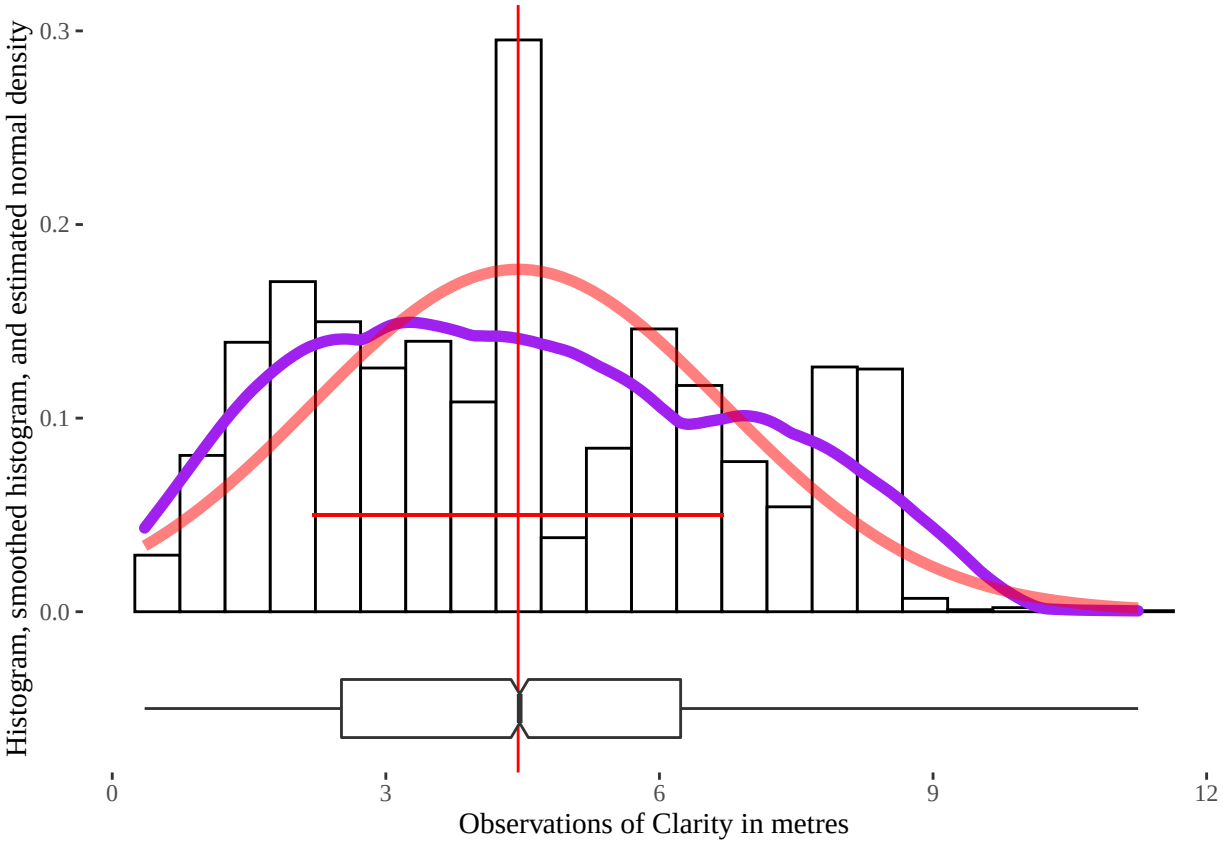


Figure 5: Histogram of Clarity (in metres)

In figure 6, 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 7 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 8. 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 9. 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 10. 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 be Uniform or Beta.

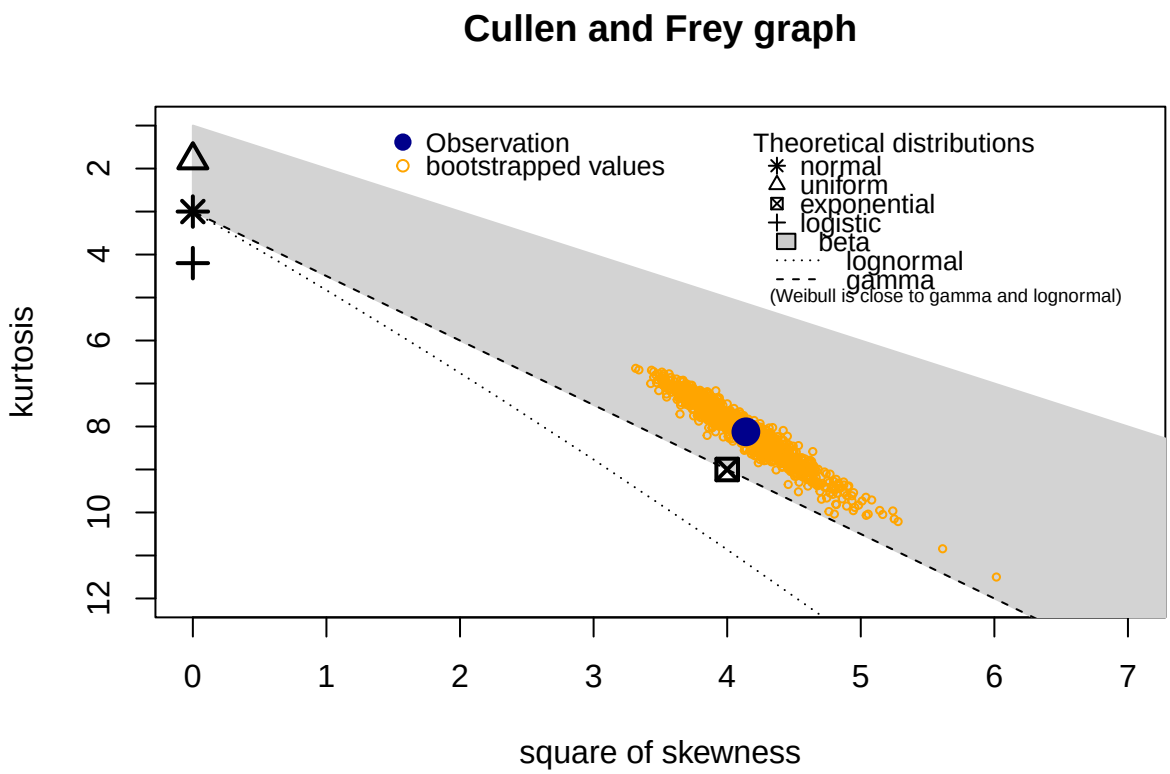


Figure 6: Cullen and Frey Graph of Ammoniacal Nitrogen

```
## 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
```

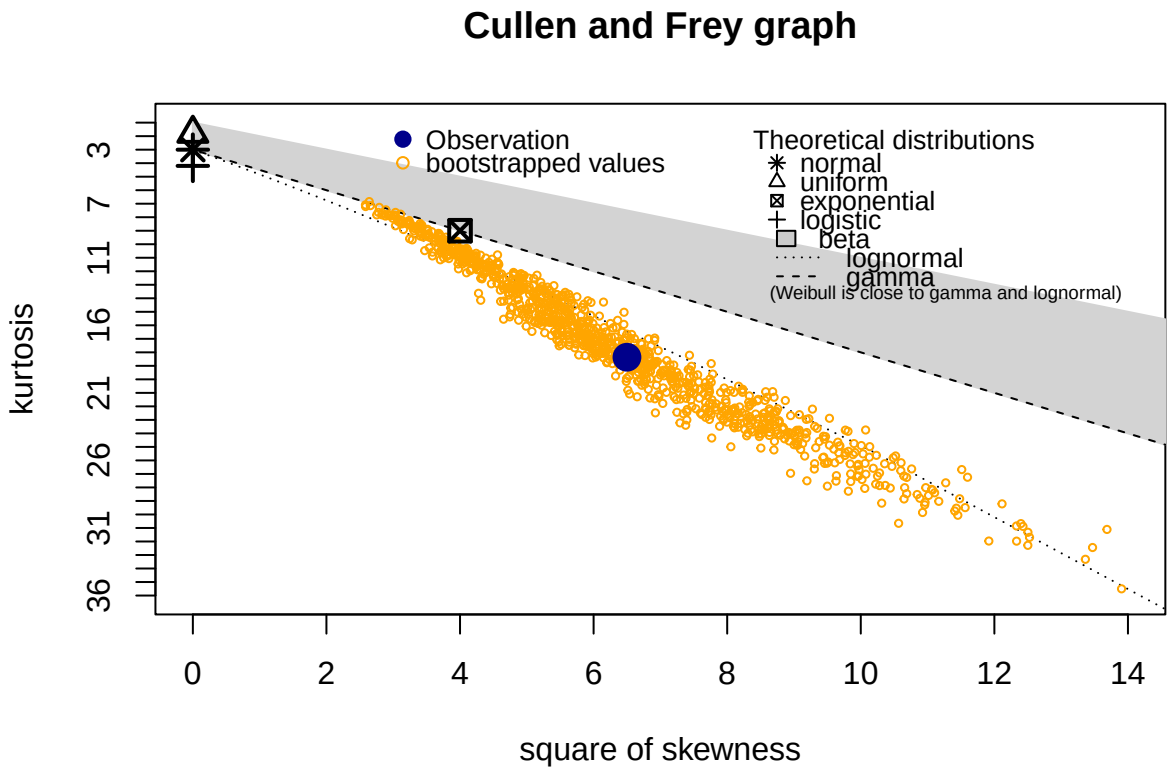


Figure 7: Cullen and Frey Graph of Chlorophyll-A

```
## 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
```

```
## summary statistics
## ------
## min:  4.017657   max:  150.4168
## median:  17.89664
## mean:  18.72058
```

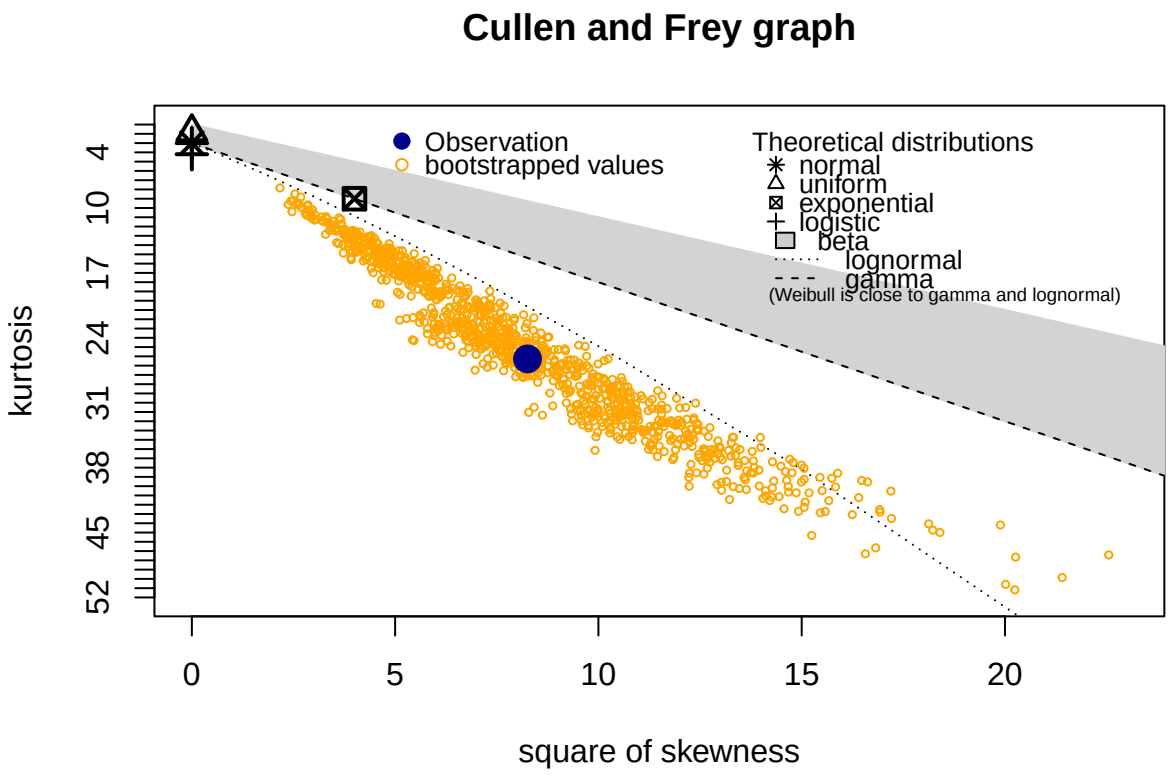


Figure 8: Cullen and Frey Graph of Total Phosphorus

```
## estimated sd:  9.143676
## estimated skewness:  2.873323
## estimated kurtosis:  26.27628
```

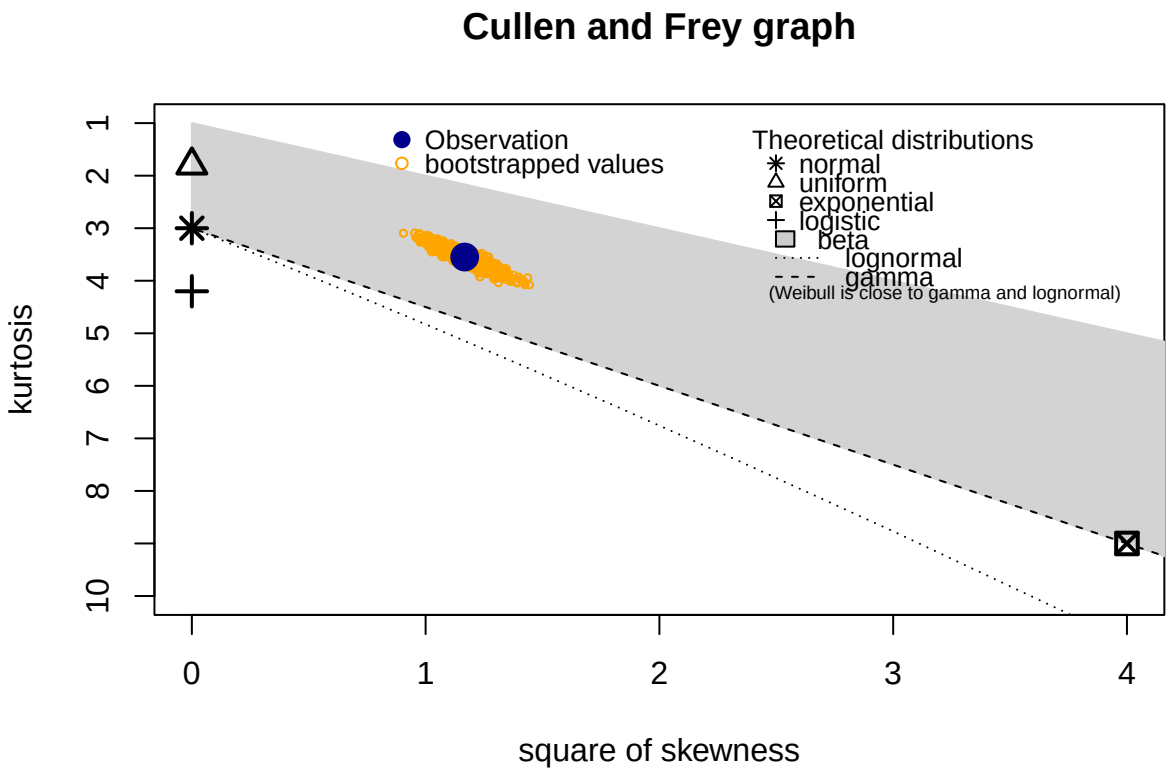


Figure 9: Cullen and Frey Graph of Total Nitrogen

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

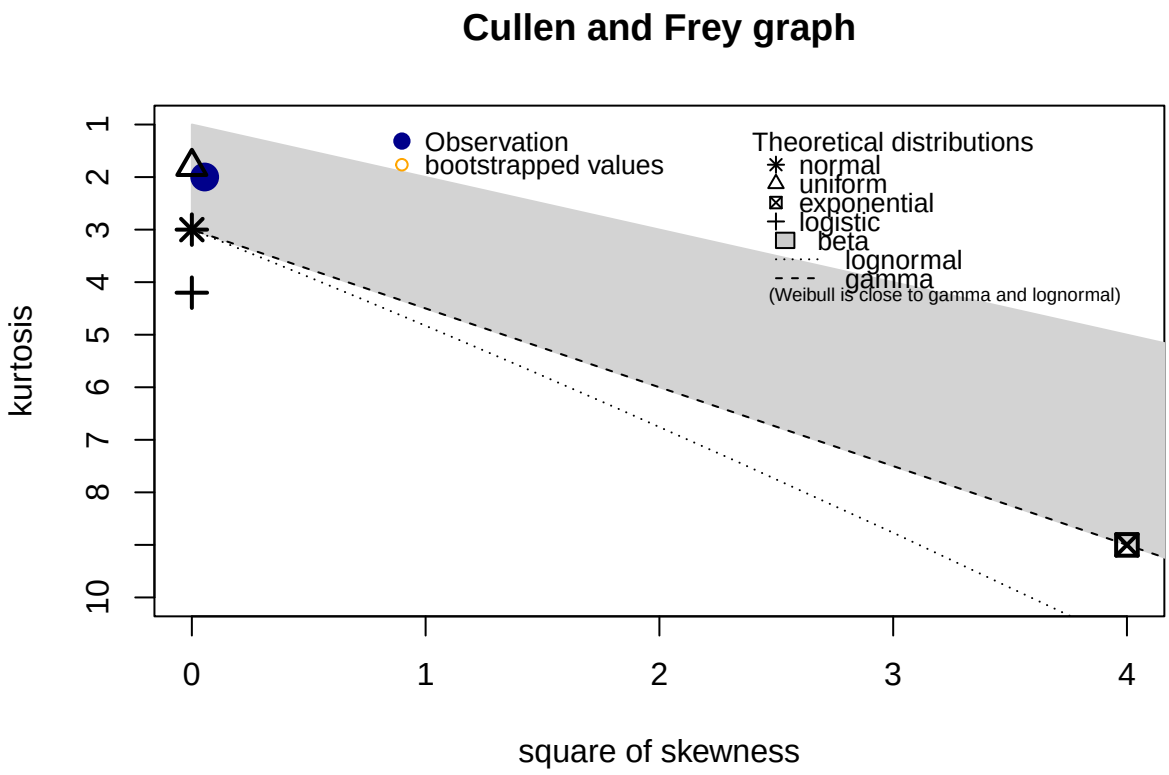


Figure 10: Cullen and Frey Graph of Clarity

Table 2: Anderson-Darling Test Statistic and P-value for the Lake Health Variables

	Ammoniacal Nitrogen	Chlorophyll-A	Total Phosphorus	Total Nitrogen	Clarity
Test Statistic	226.0701	111.2062	68.4039	133.6218	40.4676
P-Value	3.7e-24	3.7e-24	3.7e-24	3.7e-24	3.7e-24

We believe, based on the Cullen and Frey graphs, none of these distributions are normal. We confirmed this with Anderson Darlings tests of normality. Table 2 shows the test statistic and p-value for the Anderson Darling test on each Lake Health variable. The test statistics for Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity are 226, 111, 68, 134 and 40, respectively (to the nearest whole number). For each variable, the p-value is 3e-24, confirming our suspicions that these variables are not normally distributed.

2.1.2 Univariate Analysis of the Lake Dimension Variables

We gathered information on the Lake Dimension Variables: Depth, Area and Perimeter. In the univariate analysis we produced summary statistics as well as histograms of their distributions.

We first produced summary statistics as seen in table 3.

Table 3: Table of Sample Statistics

	Area	Perimeter	Depth
Sample Size	3.802000e+03	3802.0000	3802.000000
Minimum	1.000494e+04	371.1600	1.000000
1st Quantile	1.412159e+04	586.6325	14.610000
Median	2.416589e+04	823.9350	17.530000
3rd Quantile	6.299442e+04	1416.6125	23.310000
Maximum	6.130000e+08	369677.8000	462.000000
Standard Deviation	1.439392e+07	11246.8807	23.477025
Mean	9.822069e+05	2389.3750	22.910350
Kurtosis	1.023860e+03	415.0156	138.442310
Skewness	2.877753e+01	17.3889	9.785791

We produced histograms of each of the Lake Dimension Variables to examine their distributions.

Figure 11 shows the log 10 transformed histogram of the Lake Dimension Variable Lake Depth. We transformed it by log 10 to make the graph easier to interpret. The distribution of the graph is strongly skewed to the right, this is indicated by the boxplot at the bottom of the graph. Additionally from the boxplot we see that our data contains many outliers. The boxplot also suggests that with the removal of these outliers we may have a distribution that better approximates normal. We decided not to remove the outliers as these are still valid data points.

This distribution and interpretation of the above graph is very similar for all of the Lake Dimension variables; they all require log 10 transformation, they are all highly skewed, and appear to not approximate a normal distribution. As such we have not reproduced the histograms of the Lake Dimension variables Area and Perimeter here.

2.1.3 Multivariate Analysis of Lake Health Variables

Table 4 shows the correlation matrix and figure 12 shows the visualization of this 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,

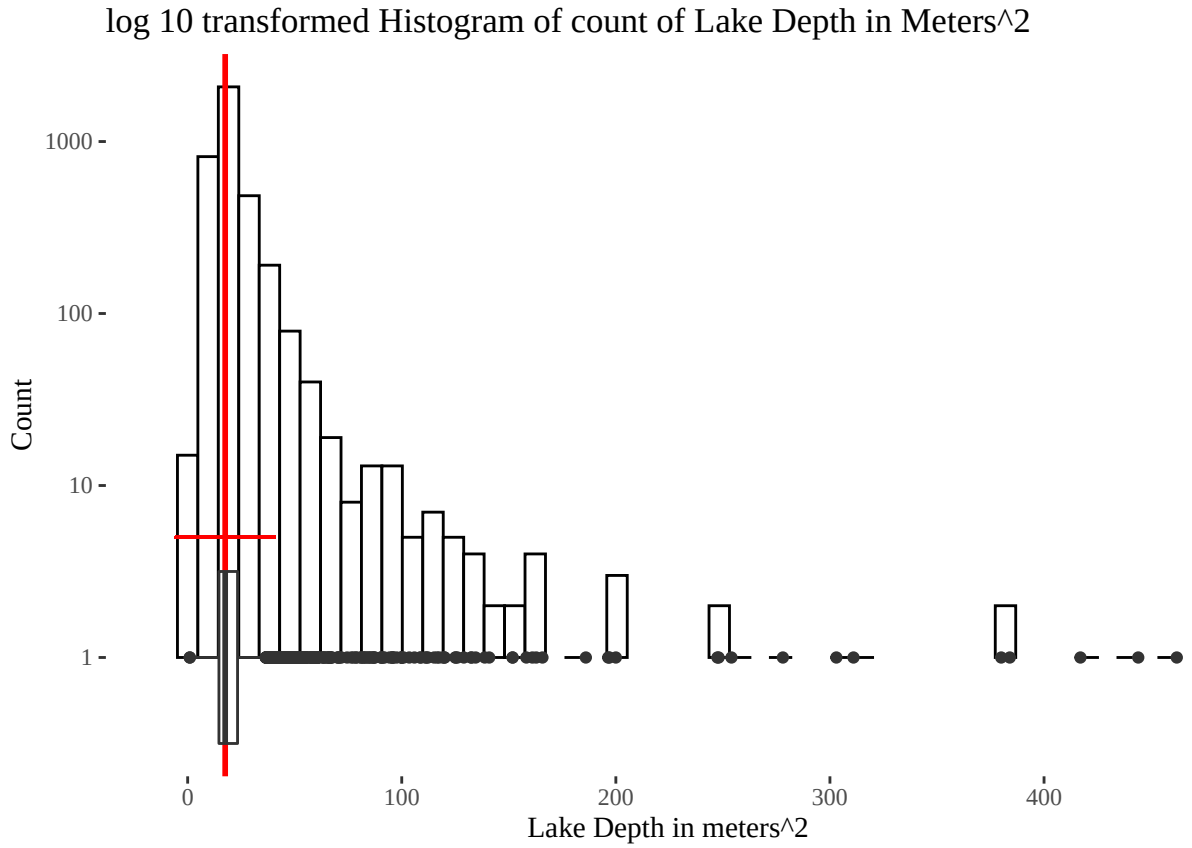


Figure 11: Log 10 transformed Histogram of Lake Depth in meters squared. Rectangles indicate skewed distribution, red vertical line indicates thge placement of the median value, red horizontal line indicates the placement of the median with one standard deviation.

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

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 to 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 12.

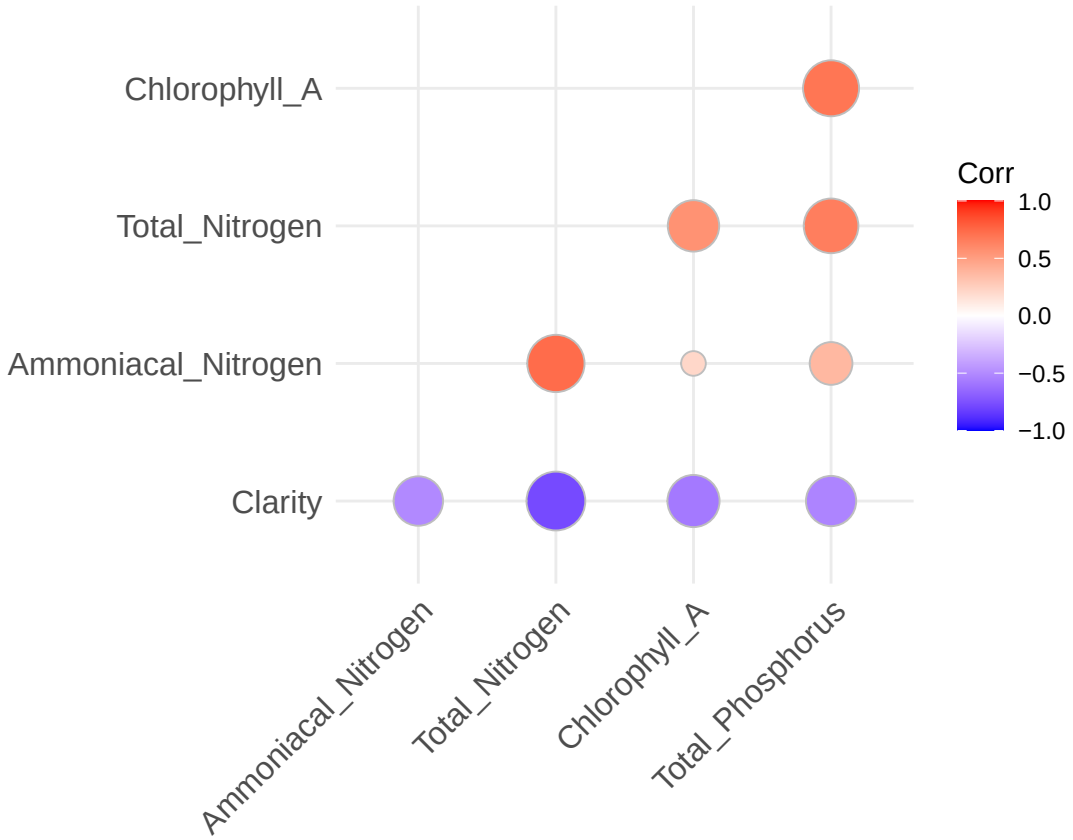


Figure 12: Visualisation of the Correlation Matrix

Figure 13 shows the pairs plot of the Lake Health variables, coloured by Dominant Landcover. We can see some separation between landcovers in the distribution of each variable. The relationships between Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen are positive, however all show

non-constant scatter. It is also difficult to see any separation between the landcovers in these scatterplots as the data is dominated by Native and Pastoral landcovers (both have 1770 observations). The relationship between these four measures and Clarity are all negative, and appear to be non-linear. Another observation from this graph is that the correlations differ greatly between landcovers. For Ammoniacal Nitrogen and Chlorophyll-A, Native and Pastoral landcover groups have a weak positive correlation, while the other three types of landcover have weak negative correlations. The correlations between Ammoniacal Nitrogen and Total Phosphorus is similar for Exotic Forest, Native and Pastoral lakes (weak to moderate positive correlation), while the correlations for Other and Urban Area are very weak. For the rest of the pairs, the correlations for each landcover was similar, with either all positive or all negative correlations, ranging from 0.3 less to 0.3 more than the collective correlation. The two types of landcover that tended to differ most are the Other and Urban Area landcovers. This may be due to the small size of these groups (12 observations for Other and 85 observations for Urban Area).

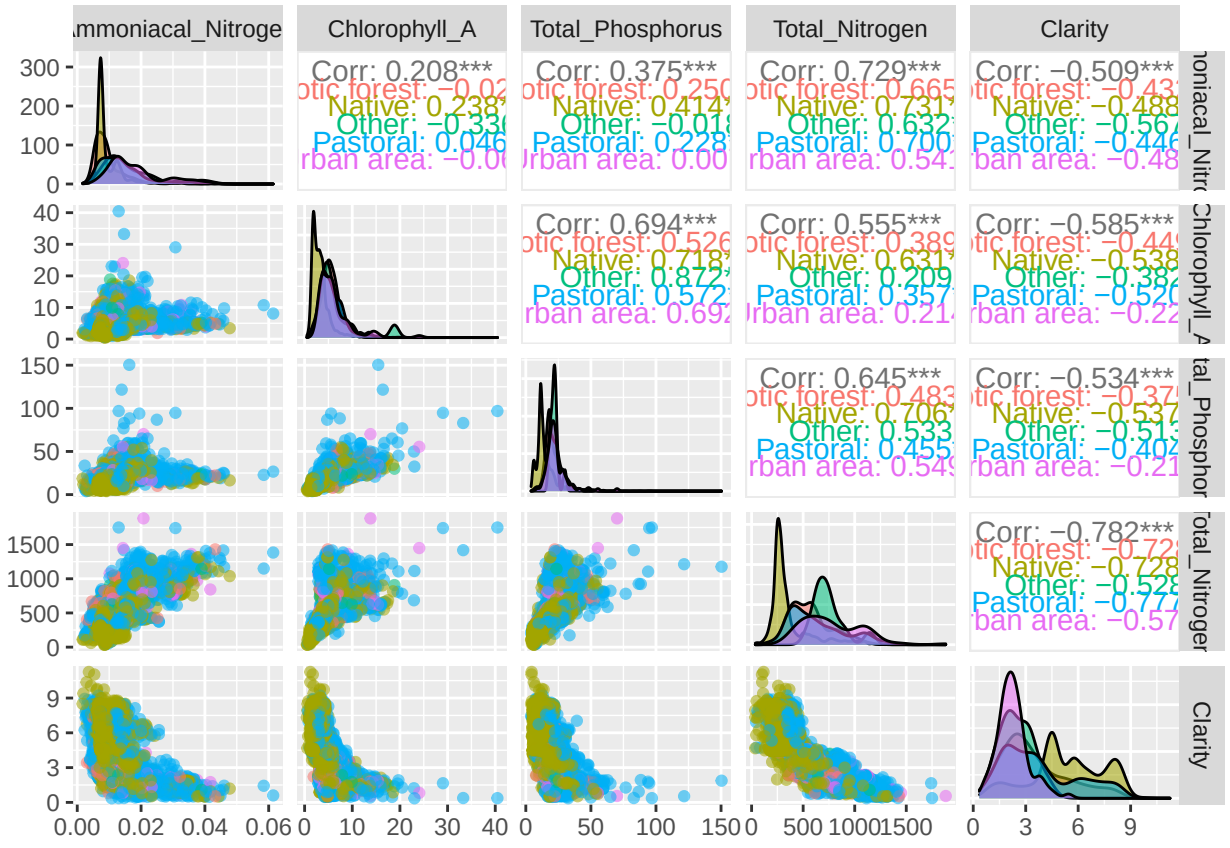


Figure 13: Pairs Plot of Lake Health

Next, we wanted to compare the distribution of the lake health variables by types of dominant landcover. Figure 14 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.

Figure 15 shows the distribution of Chlorophyll-A in each type of landcover. The observations have been log

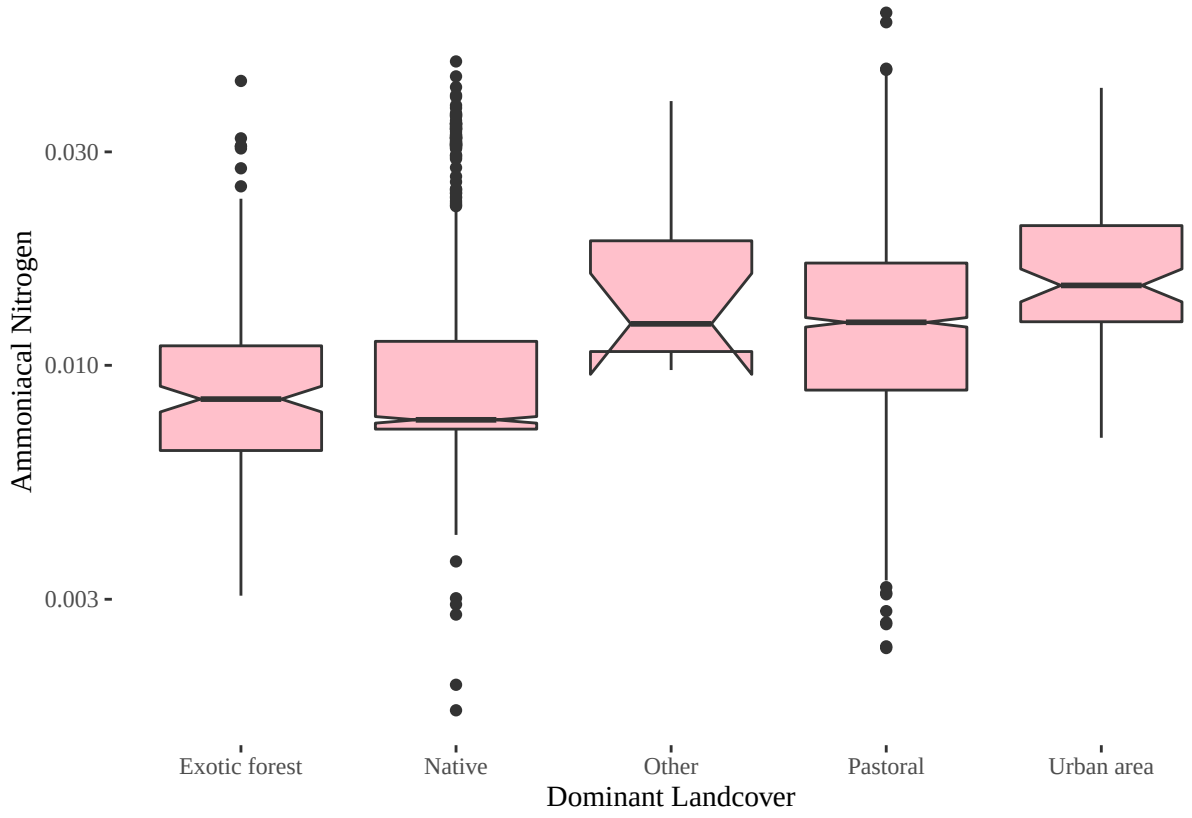


Figure 14: Box Plot of Landcover and Ammoniacal Nitrogen

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).

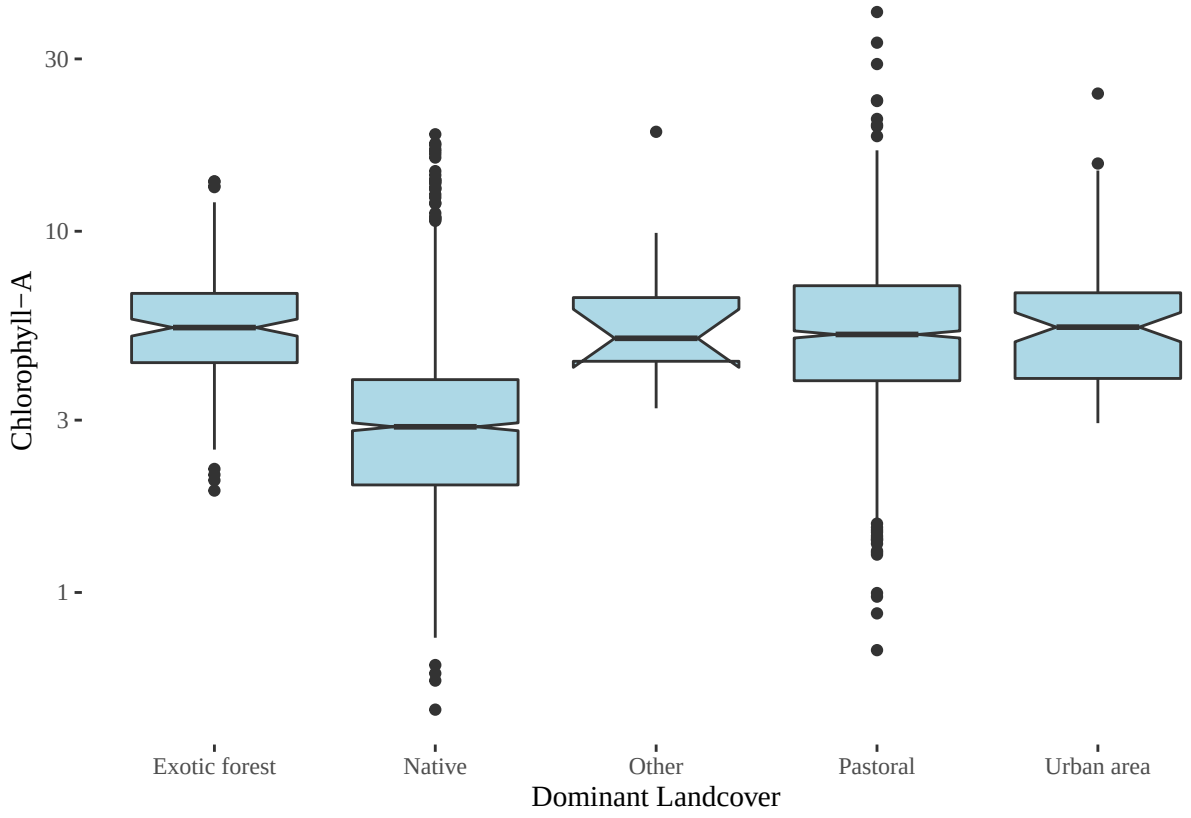


Figure 15: Box Plot of Landcover and Chlorophyll-A

Figure 16 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 17 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.

The distribution of Clarity by dominant landcover is shown in figure 18. The lakes with Native landcover

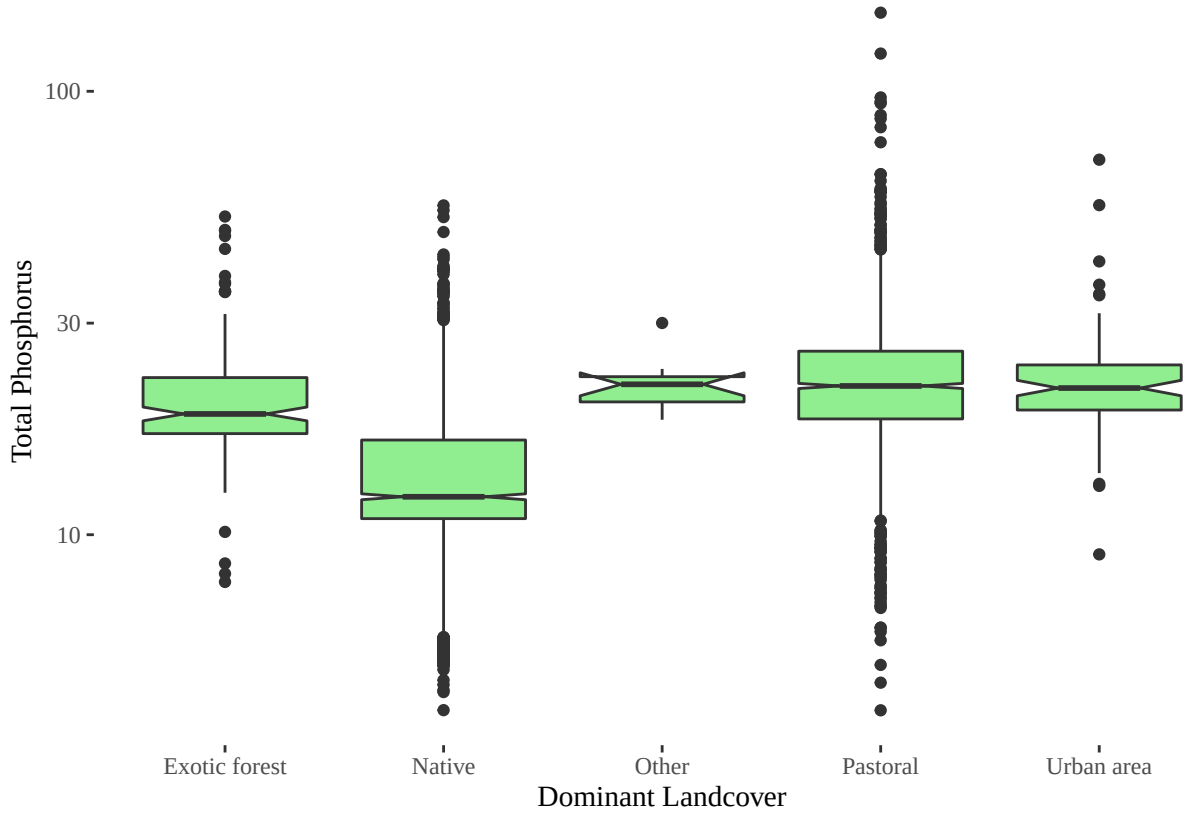


Figure 16: Box Plot of Landcover and Total Phosphorus

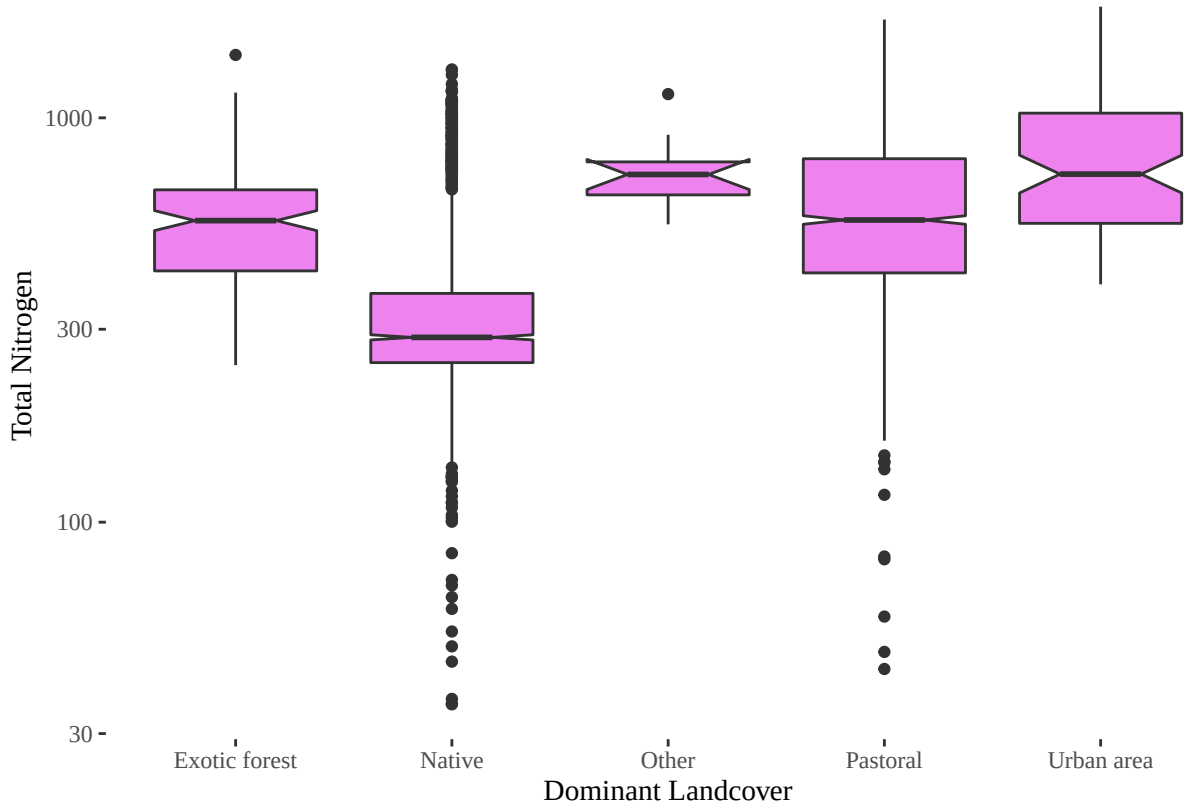


Figure 17: Box Plot of Landcover and Total Nitrogen

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.

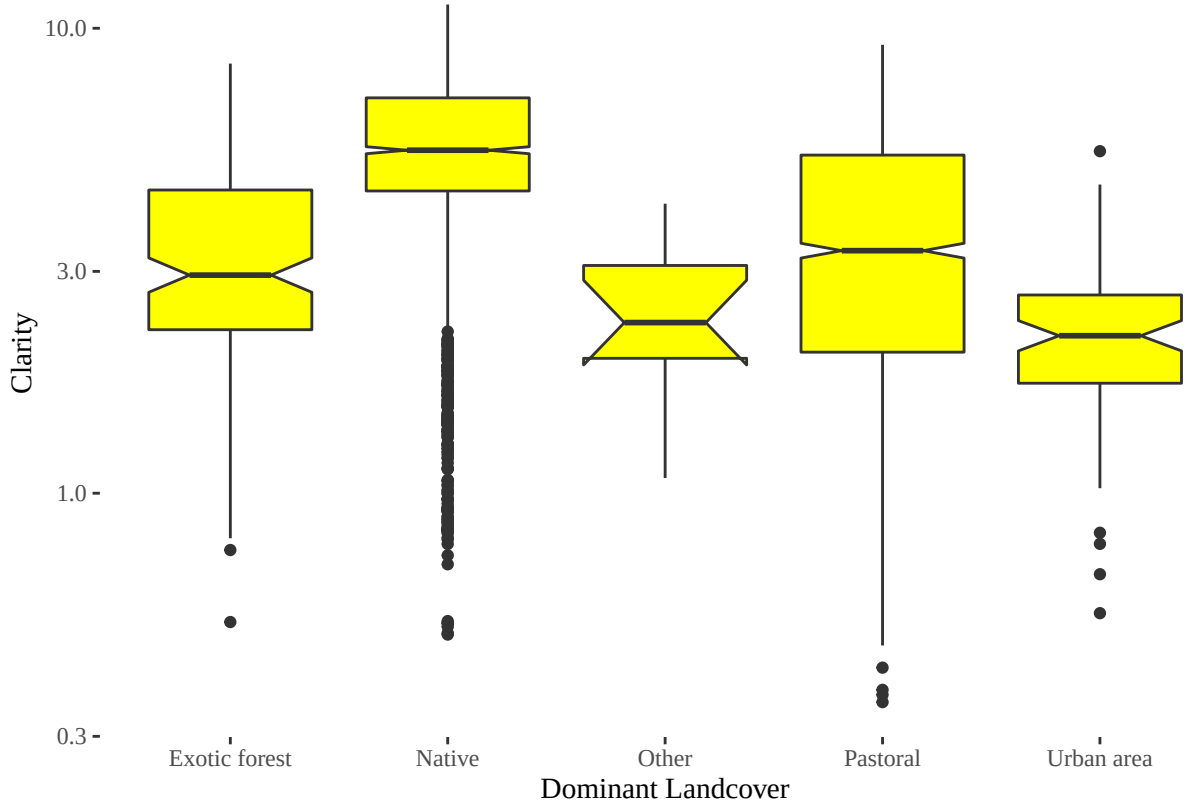


Figure 18: Box Plot of Landcover and Clarity

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
Standard Deviation	0.0059558	2.251981	7.172119	205.723829	1.8382921
Mean	0.0101148	5.759535	20.640032	577.841100	3.4546657
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
Standard Deviation	0.0053919	2.024765	6.085992	203.570216	2.0869312
Mean	0.0098571	3.310055	13.856583	360.579423	5.3440779
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
Standard Deviation	0.0085389	4.319832	3.147497	158.740455	0.9114810
Mean	0.0160571	6.517253	21.893479	745.837125	2.4852789
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
Standard Deviation	0.0073402	2.920128	9.464692	275.1740646	2.1464743
Mean	0.0139015	5.715493	23.177153	629.6969317	3.7689957
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

Table 10: Anderson-Darling Test Statistic and P-value for the Lake Health Variables by Native Landcover

	Ammoniacal Nitrogen	Chlorophyll-A	Total Phosphorus	Total Nitrogen	Clarity
Test Statistic	208.5689	91.6093	58.2436	167.8357	21.0911
P-Value	3.7e-24	3.7e-24	3.7e-24	3.7e-24	3.7e-24

Table 11: Regional Mean and Median Lake Health

Region	Mean_NH4N	Median_NH4N	Region	Mean_CHLA	Median_CHLA	Region	Mean_TP	Median_TP
Waikato	0.0169931	0.0123660	Gisborne	7.808367	7.351853	Taranaki	24.60256	24.83678
Wellington	0.0166320	0.0134980	Taranaki	7.176843	7.110607	Gisborne	26.26159	23.86373
Gisborne	0.0160306	0.0132765	Hawke's Bay	6.809741	6.649401	Waikato	23.23461	22.29681
ManawatÃ«-Whanganui	0.0143978	0.0110985	Waikato	7.115944	6.156554	ManawatÃ«-Whanganui	24.04704	21.75352
Marlborough district council	0.0138354	0.0097970	Northland	5.503763	5.421323	Hawke's Bay	21.36501	21.60229
Auckland	0.0134935	0.0115450	Wellington	6.890644	5.391158	Wellington	23.06120	21.30178
Canterbury	0.0133163	0.0131270	ManawatÃ«-Whanganui	5.896478	5.356666	Marlborough district council	19.68485	19.60149
Bay of Plenty	0.0128409	0.0098540	Auckland	5.925387	5.335608	Bay of Plenty	20.54074	19.54886
Otago	0.0127451	0.0120175	Bay of Plenty	5.670382	4.981374	Otago	20.69889	19.01017
Hawke's Bay	0.0114907	0.0096720	West Coast	4.128556	3.522108	Canterbury	20.03111	17.96068
West Coast	0.0111620	0.0087230	Marlborough district council	4.418469	3.506617	Northland	18.36308	17.88858
Taranaki	0.0101370	0.0101145	Canterbury	3.523742	3.334410	Auckland	19.11667	17.33723
Southland	0.0099627	0.0074620	Otago	3.326594	3.219797	Tasman district council	17.77295	15.33291
Tasman district council	0.0091943	0.0072210	Southland	3.258189	3.073008	West Coast	16.99558	15.10845
Northland	0.0089272	0.0075225	Tasman district council	5.150028	2.920793	Southland	13.71377	11.43682

	Ammoniacal Nitrogen	Chlorophyll-A	Total Phosphorus	Total Nitrogen	Clarity
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
Standard Deviation	0.0083745	3.389789	8.392760	286.3296449	0.8785109
Mean	0.0180000	6.126680	23.030843	778.8134400	2.2658755
Kurtosis	3.4495705	11.658444	15.327134	4.1074614	4.5139920
Skewness	1.1230235	2.512231	2.892799	0.9219365	0.8943855

We wanted to run tests for a difference in mean of each Lake Health variable for each type of landcover, but we first needed to check the normality assumptions to ascertain which tests we could conduct. If any one of these groups does not meet the normality assumption, normality cannot be assumed for these tests. To test this we conducted Anderson-Darling tests on just the Native landcover group.

Table 10 shows the test statistic and p-value for the Anderson Darling test on each Lake Health variable for the Native landcover. The test statistics for Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity are 209, 92, 58, 168 and 21, respectively (to the nearest whole number). For each variable, the p-value is 3e-24, confirming our suspicions that the distribution of the Lake Health variables for Native lakes are not normally distributed, violating the normality assumption. We conducted this test on every type of landcover with similar results.

We analysed the relationship between the Lake Health variables and Region. Table 11 shows the mean and median of the Ammoniacal Nitrogen, Chlorophyll-A and Total Phosphorus for each region, and table 12 shows Total Nitrogen and Clarity.

2.1.3.1 Multivariate analysis of the Lake Dimension variables

2.1.3.2 Examining the Correlation between Area, Depth and Perimeter NEED TO EDIT REFERENCING AND TITLES

*Correlation matrix rounded to three decimal place

```
##                lake_perimeter lake_depth lake_area
## lake_perimeter          1.000      0.780     0.797
```

Table 12: Regional Mean and Median Lake Health

Region	Mean_TN	Median_TN	Region	Mean_SECCHI	Median_SECCHI
Waikato	790.2603	791.5458	Wellington	2.182703	1.734104
Wellington	708.1088	722.2273	Waikato	3.108126	2.461897
Gisborne	700.7699	680.4638	Auckland	2.635606	2.596739
Taranaki	574.7697	563.5217	Gisborne	3.033813	2.757868
ManawatÅ«-Whanganui	709.0396	555.2100	Northland	3.121884	2.803300
Northland	553.4311	550.2800	Taranaki	3.216804	3.109920
Hawke's Bay	558.7630	535.7862	ManawatÅ«-Whanganui	3.582909	3.119695
Marlborough district council	522.9138	527.1625	Hawke's Bay	3.606981	3.416575
Auckland	534.9663	486.0566	Marlborough district council	4.336709	4.020992
Bay of Plenty	573.8112	451.1150	Bay of Plenty	4.182245	4.183662
Canterbury	481.9375	379.6930	West Coast	4.680724	4.601896
Otago	422.0329	367.6522	Southland	5.138127	4.616884
West Coast	492.6764	365.7440	Tasman district council	4.691611	5.709074
Southland	373.3848	295.8158	Otago	5.395194	5.727244
Tasman district council	448.2798	295.3265	Canterbury	5.546141	5.743230

```
## lake_depth              0.780      1.000      0.518
## lake_area               0.797      0.518      1.000
```

Visualisation of the Correlation Matrix:

As to be expected, the three Lake Dimension 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.

Below is a frequency table for the number of lakes in each region.

Table 13: Frequency table of number of lakes in each Region

region1	Freq
Auckland	74
Bay of Plenty	98
Canterbury	463
Gisborne	30
Hawke's Bay	285
ManawatÅ«-Whanganui	226
Marlborough district council	50
Northland	262
Otago	378
Southland	991
Taranaki	90
Tasman district council	57
Waikato	233
Wellington	107
West Coast	457

We then checked the relationship between the Lake Health variables and the Lake Dimension variables.

Figure 20 shows the pairs plot of the Lake Health variables by the Lake Dimension variables. We can see very little relationship between the Lake Health and Lake Dimension variables. The majority of the scatter we can see if due to outliers and large amounts of lakes with low values for depth, perimeter and area. We decided not to investigate these relationships any further.

3 Results

3.1 Tests for Difference in Mean Lake Health

As demonstrated in our Exploratory data Analysis, the distributions of the Lake Health variables for each landcover type are not normal. We performed Kruskal-Wallis tests for equality in means between landcover

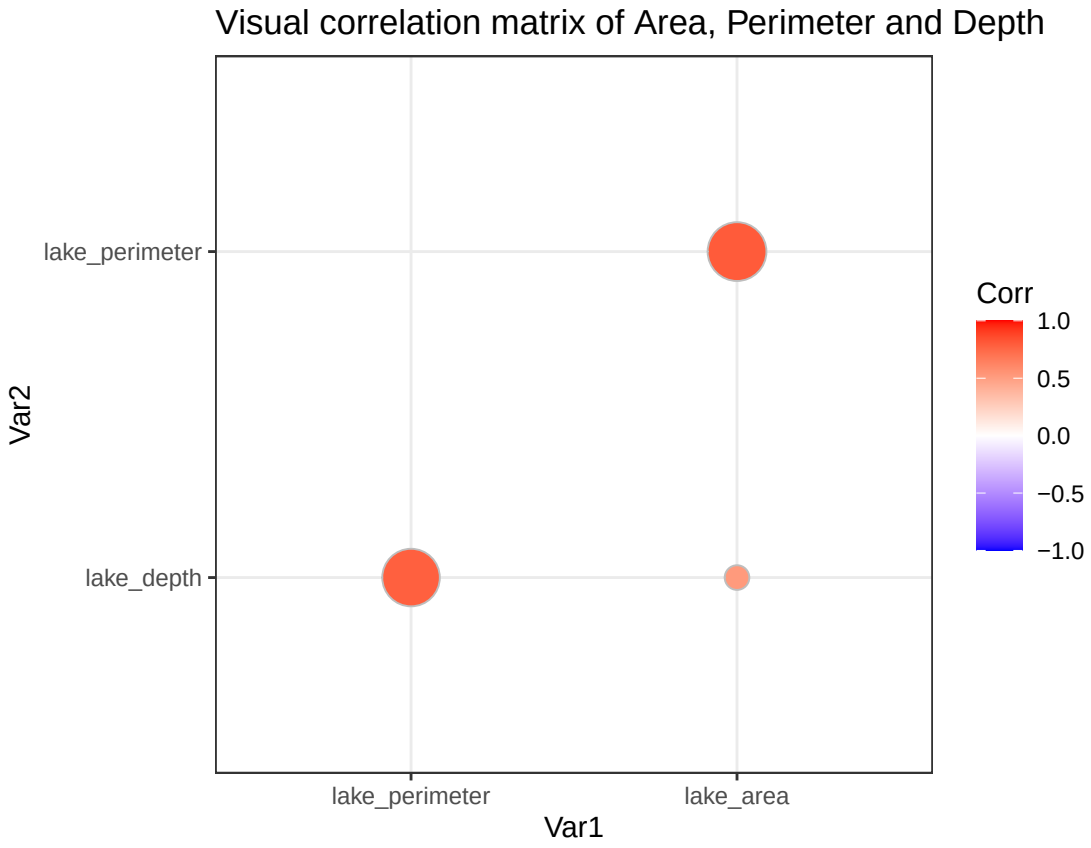


Figure 19: Visual of the correlation between, depth, perimeter and area where Strength and direction of relationship are indicated by circle size and colour

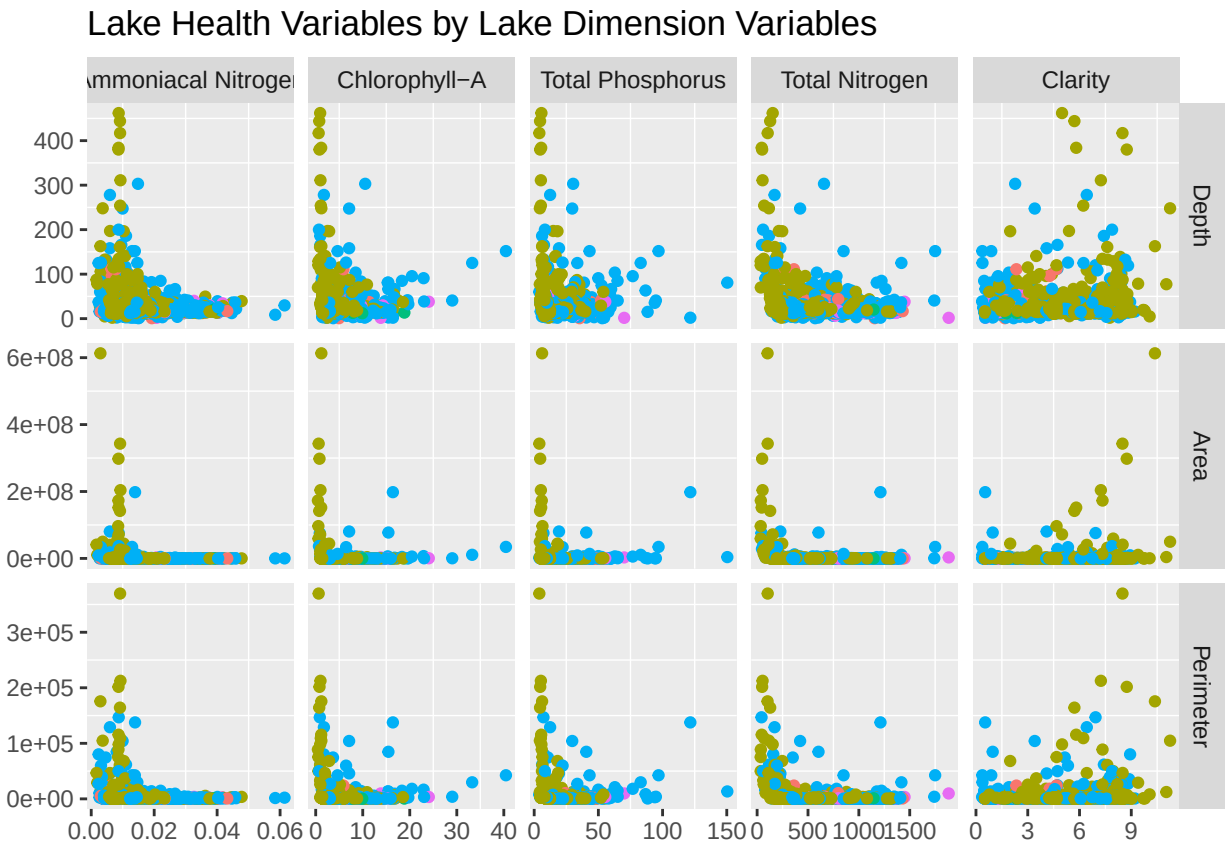


Figure 20: Pairs Plot of Lake Health Variables by Lake Dimension Variables

different mean Ammoniacal Nitrogen level than lakes with Exotic forest landcover, and lakes with Native landcover, both have a p-value of less than 2e-16. And lakes with Urban landcover have a different mean Ammoniacal Nitrogen level than lakes with Pastoral landcover, as the p-value is 9e-7. The pairs that failed to reject the null hypothesis were Exotic forest and Native landcovers, Pastoral and 'other' landcovers and Urban and 'other' landcovers so there could be no difference between the mean Ammoniacal Nitrogen level in lakes in these landcover pairs.

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  NH4N and land
##
##            Exotic forest Native  Other   Pastoral
## Native     0.74060       -       -       -
## Other      0.00041       0.00012 -       -
## Pastoral   < 2e-16       < 2e-16 0.36372 -
## Urban area < 2e-16       < 2e-16 0.33854 9e-07
##
## P value adjustment method: BH
```

Next we conducted a pairwise Wilcox test for difference in mean Chlorophyll-A levels between pairs. This concluded that there is a difference between mean Chlorophyll-A levels in lakes with Native landcover and every other type of landcover. The p-value for the Native and Exotic forest pair was less than 2e-16, Native and 'Other' pair was 6.4e-5, Native and Pastoral pair was less than 2e-16 and Native and Urban was less than 2e-16. No other pair had a p-value small enough to reject the null hypothesis that there is not a difference in mean Chlorophyll-A level between those two landcovers. We can conclude that lakes with Native landcover had a different mean Chlorophyll-A level than any other landcover type.

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  CHLA and land
##
##            Exotic forest Native  Other Pastoral
## Native     < 2e-16       -       -     -
## Other      0.98          6.4e-05 -     -
## Pastoral   0.43          < 2e-16 0.98  -
## Urban area 0.98          < 2e-16 1.00  0.88
##
## P value adjustment method: BH
```

We conducted a pairwise Wilcox test for difference in means between pairs. This concluded that there is a difference between mean Total Phosphorus levels in lakes with Native landcover and every other type of landcover. The p-value for the Native and Exotic forest pair was less than 2e-16, Native and 'Other' pair was 4.0e-6, Native and Pastoral pair was less than 2e-16 and Native and Urban was less than 2e-16. We can conclude that lakes with Native landcover had a different mean Total Phosphorus level than any other landcover type. The p-value for Pastoral and Exotic forest is 1.4e-7 and the p-value for Urban and Exotic forest was 0.001. So we can conclude lakes with Exotic forest landcover had a different mean Total Phosphorus level than lakes with Pastoral landcover or lakes with Urban landcover. No other pair had p-values small enough to conclude a difference in means.

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
```

```
##
## data:  TP and land
##
##            Exotic forest Native  Other Pastoral
## Native     < 2e-16       -       -     -
## Other      0.064         4.0e-06 -     -
## Pastoral   1.4e-07       < 2e-16 0.957 -
## Urban area 0.001         < 2e-16 0.957 0.957
##
## P value adjustment method: BH
```

Next we conducted a pairwise Wilcox test for difference in means between pairs. This found that there is a difference between mean Total Nitrogen levels in lakes with Native landcover and every other type of landcover. The p-value for the Native and Exotic forest pair was less than 2e-16, Native and 'Other' pair was 8.9e-7, Native and Pastoral pair was less than 2e-16 and Native and Urban was less than 2e-16. We can conclude that lakes with Native landcover had a different mean Total Nitrogen level than any other landcover type. The p-value for 'other' and Exotic forest is 0.0017 and the p-value for Urban and Exotic forest was 4.4e-8. So we can conclude lakes with Exotic forest landcover had a different mean Total Nitrogen level than lakes with 'Other' landcover or lakes with Urban landcover. The p-value for Pastoral and Urban lancovers was 8.9e-7. indicating there is a difference in mean Total Nitrogen levels in lakes with Pastoral landcover and lakes with Urban. No other pair had p-values small enough to conclude a difference in means.

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  TN and land
##
##            Exotic forest Native  Other  Pastoral
## Native     < 2e-16       -       -      -
## Other      0.0017        8.9e-07 -      -
## Pastoral   0.1418        < 2e-16 0.0294 -
## Urban area 4.4e-08       < 2e-16 0.8997 8.9e-07
##
## P value adjustment method: BH
```

Next we conducted a pairwise Wilcox test for difference in means between pairs. This concluded that there is a difference between mean Clarity in lakes with Native landcover and every other type of landcover. The p-value for the Native and Exotic forest pair was less than 2e-16, Native and 'Other' pair was 6.1e-6, Native and Pastoral pair was less than 2e-16 and Native and Urban was less than 2e-16. We can conclude that lakes with Native landcover had a different mean Clarity level than any other landcover type. The p-value for Pastoral and urban area is 1.9e-10 and the p-value for Urban and Exotic forest was 2.6e-7. So we can conclude lakes with Urban area landcover had a different mean level of Clarity than lakes with Pastoral landcover or lakes with Exotic forest landcover. No other pair had p-values small enough to conclude a difference in means.

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  SECCHI and land
##
##            Exotic forest Native  Other Pastoral
## Native     < 2e-16       -       -     -
## Other      0.103         6.1e-06 -     -
```

```
## Pastoral   0.183         < 2e-16 0.071 -
## Urban area 2.6e-07       < 2e-16 0.396 1.9e-10
##
## P value adjustment method: BH
```

3.2 Principal Component Analysis

We decided to conduct a Principal Component Analysis on the Lake Health variables to determine whether we can reduce the number of variables while still retaining most of the information. Figure 21 shows the scree graph of the five principal components. We can see from the summary below that the first principal component explains 65.6% of the variance, with the second principal component this becomes 83.4% and with the third they capture 92.4% of the variance.

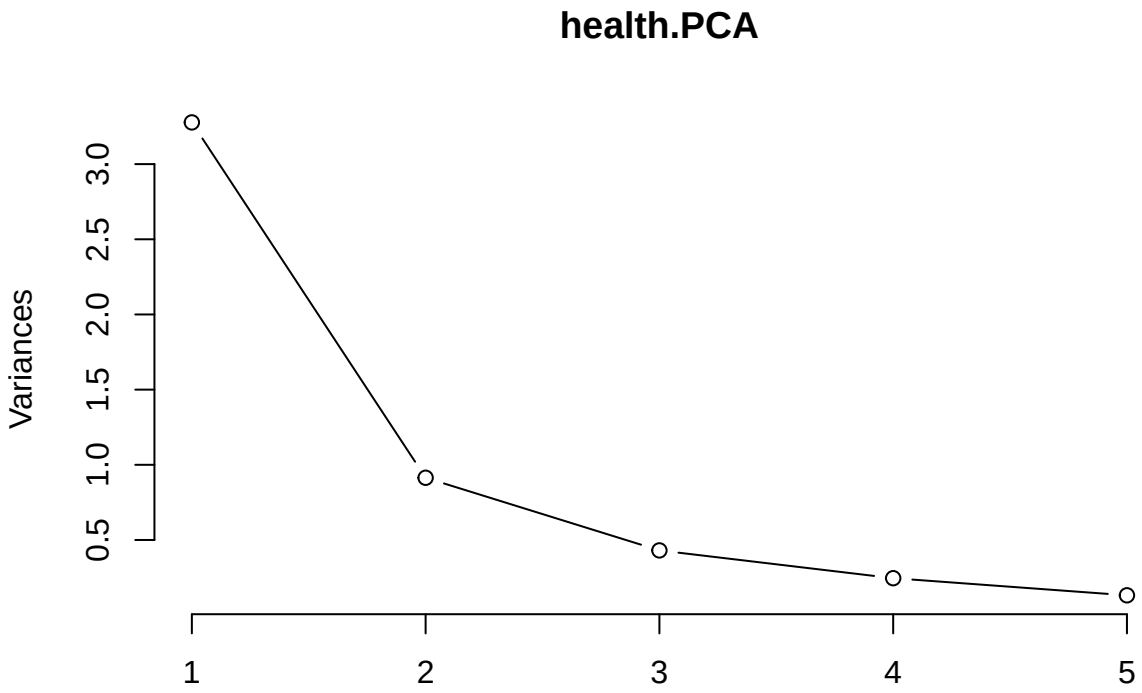


Figure 21: Scree Graph

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5
## Standard deviation     1.8103 0.9560 0.65632 0.49589 0.36343
## Proportion of Variance 0.6555 0.1828 0.08615 0.04918 0.02642
## Cumulative Proportion  0.6555 0.8383 0.92440 0.97358 1.00000
```

The eigenvectors are below, and shown visually in figure **??**. We can see the first principal component is approximately the average of the Lake Health variables. If we were to move forward with the principal components we would just retain principal components 1 and 2 as they capture 83.4% of the variance.

```
##               PC1         PC2         PC3         PC4         PC5
## NH4N    0.3804385 -0.68273926  0.32640483 -0.38001162  0.37173198
## CHLA    0.4160810  0.58414236 -0.08611696 -0.69136513  0.01588745
## TP      0.4448543  0.36615759  0.61121174  0.50549627  0.19729917
## TN      0.5114595 -0.23588357 -0.04521927  0.09530825 -0.81953627
## SECCHI -0.4718654  0.05415131  0.71443627 -0.33614772 -0.38858237
```

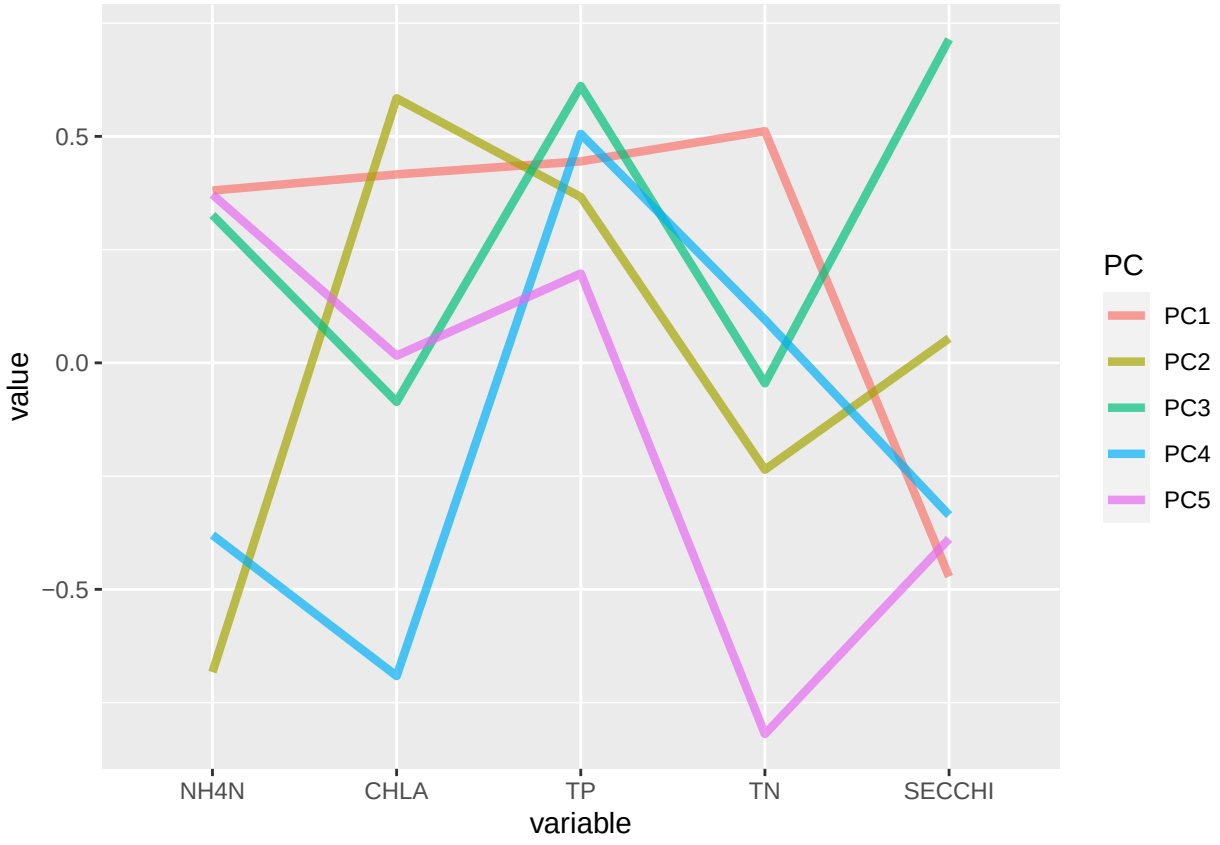


Figure 22: Visualisation of the Eigenvctors

3.3 Factor Analysis

We decided to perform a Factor analysis on the Lake Health variables to determine whether they could be explained using two factors. The null hypothesis for this test was that two factors are sufficient to capture the full dimensionality of the Lake Health variables. The output of this test showed a chi-square test statistic of 65.57 on one degree of freedom and a p-value of 5.59e-16. We concluded, at the 0.01 significance level, that two factors are not sufficient to capture the full dimensionality of these variables. The output of this test is shown below.

```
##
## Call:
## factanal(x = df1, factors = 2)
##
## Uniquenesses:
##   NH4N   CHLA     TP     TN SECCHI
```

```
##  0.408  0.005  0.419  0.005  0.353
##
## Loadings:
##        Factor1 Factor2
## NH4N    0.764
## CHLA    0.149   0.986
## TP      0.412   0.641
## TN      0.902   0.426
## SECCHI -0.632  -0.497
##
##                Factor1 Factor2
## SS loadings      1.988   1.821
## Proportion Var   0.398   0.364
## Cumulative Var   0.398   0.762
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 65.57 on 1 degree of freedom.
## The p-value is 5.59e-16
```

3.4 Linear Discriminant Analysis

For LDA, we wanted to see if there was potential to implement LDA if there was information loss. As such, we attempted to predict which Dominant Landcover a lake was, based on the Lake Health Variables.

Reminder - the categories that Dominant Landcover can take on are: Exotic Forest, Native, Other, Pastoral and Urban area.

We first scaled the Lake Health Variables. We then randomly created training and test sets with 70% of the data in the training set and 30% in the test set. We then fit the LDA model using the train data. The output was shown below.

```
## Call:
## lda(dominant_landcover ~ ., data = train)
##
## Prior probabilities of groups:
## Exotic forest        Native         Other      Pastoral    Urban area
##   0.045505829   0.461827755   0.003384731   0.464836405   0.024445280
##
## Group means:
##               median_TN_mg.m.3 median_CHLA_mg.m.3 median_SECCHI_metre
## Exotic forest        0.2730316          0.4435337          -0.4154415
## Native              -0.5062055         -0.4594529           0.3971934
## Other                1.0258008          0.6837159          -0.9124731
## Pastoral             0.4459826          0.4091680          -0.3274973
## Urban area           1.0525270          0.6213234          -0.9667699
##               median_TP_mg.m.3 median_NH4N_mg.L..1
## Exotic forest        0.2275124          -0.2646927
## Native              -0.5223730          -0.2933420
## Other                0.3777788           0.7550298
## Pastoral             0.4988880           0.2623736
## Urban area           0.5386756           0.8953922
##
## Coefficients of linear discriminants:
##                             LD1        LD2        LD3         LD4
```

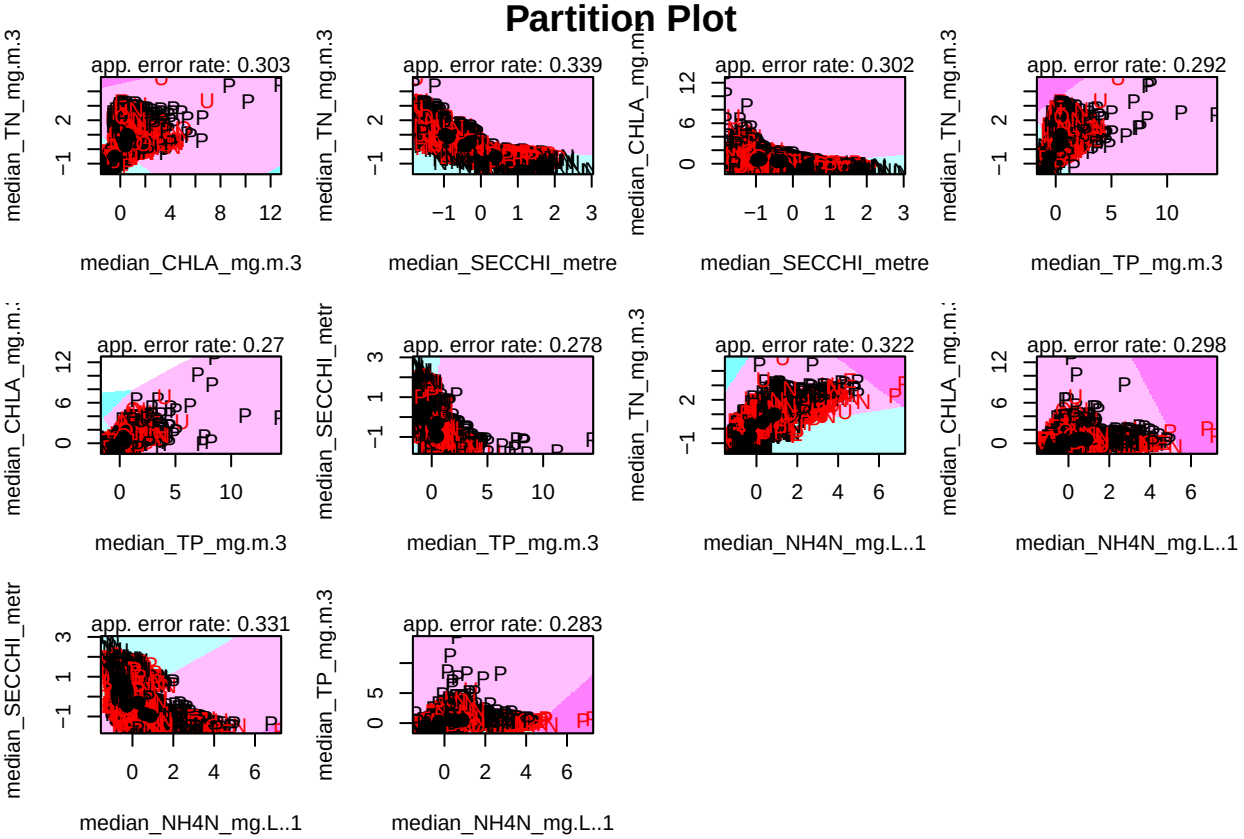
```
## median_TN_mg.m.3    -0.8323879 -0.7181879  0.1465040  1.21592998
## median_CHLA_mg.m.3  -0.3431264 -0.1201895  0.1163328  1.01510707
## median_SECCHI_metre -0.2016573  0.5174212 -0.6676406  1.41899817
## median_TP_mg.m.3    -0.4412428  0.4565318 -1.0312835 -0.95252709
## median_NH4N_mg.L..1  0.1122687  1.3228239  0.4631494 -0.06801566
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.9166 0.0483 0.0347 0.0004
```

From the Prior Probabilities we see that Pastoral Landcover and Native Landcover each explain about 46% of the prior probabilities in the training set.

From examining the coefficients of the linear discriminants, it seems that Total nitrogen (TN) impacts the first linear discriminant the most (-0.8323.).

Examining the proportion of traces we see that the LD1 accounts for approximately 91 percent of the separation. LD2 and LD3 account for approximately 4.8 and 3.5 percent of the separation respectively. The remainder is explained by LD4.

We produced a partition plot to examine the separation.



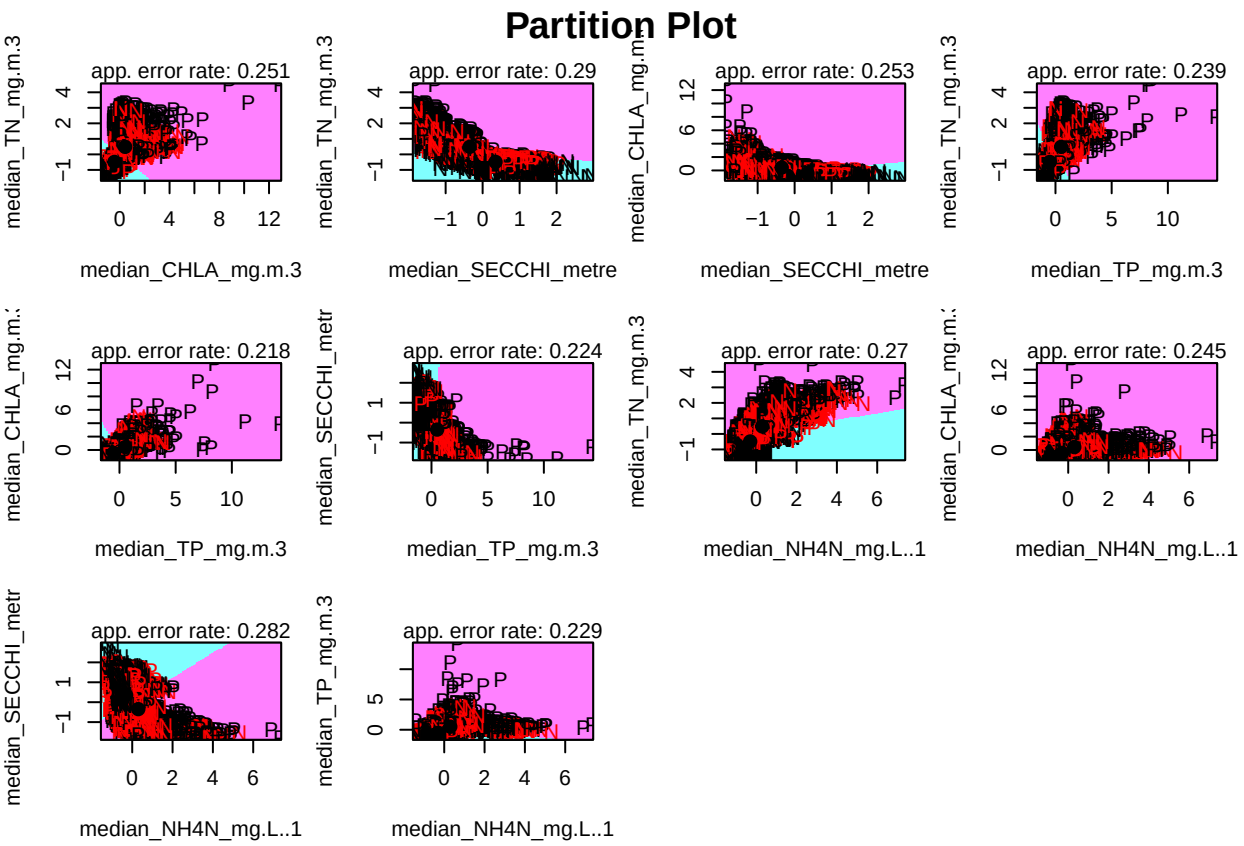
It is difficult to get much useful information from the the Partition Plot. What we can see is that there is a reasonable number of errors. We can also see is that most of the data seems to be in two of the partitions and some of the partitions have no observations in them.

We then looked at the accuracy of the model and found that the model had approximately 73% accuracy (with seed = 123). We were reasonably happy with this. However, one thing that bothered us was the partition plot and how difficult it was to decipher and the fact the prior probabilities for two of our classes

(Pastoral Landcover and Native Landcover) were so much higher than the others. We decided to investigate just these two classes as they accounted for about 93% of the observations in our sample.

```
## Call:
## lda(dominant_landcover ~ ., data = train)
##
## Prior probabilities of groups:
##   Native Pastoral
##      0.5      0.5
##
## Group means:
##          median_TN_mg.m.3 median_CHLA_mg.m.3 median_SECCHI_metre
## Native         -0.5084977         -0.4510662           0.3811648
## Pastoral        0.4825410          0.4131210          -0.3292880
##          median_TP_mg.m.3 median_NH4N_mg.L..1
## Native          -0.523458          -0.3217432
## Pastoral         0.504890           0.2886076
##
## Coefficients of linear discriminants:
##                              LD1
## median_TN_mg.m.3     0.89638023
## median_CHLA_mg.m.3   0.22175903
## median_SECCHI_metre  0.31530056
## median_TP_mg.m.3     0.52855329
## median_NH4N_mg.L..1 -0.08742443
```

In the second LDA both groups had prior probabilities of 0.5 (this is because the train happened to split them with an equal number of observations of 1242 each (seed = 123)).



The Partition plot while not any easier to interpret, sorted groups more appropriately and did not have unused partitions. We still see plenty of errors. The individual graphs give error rates between 0.218 and 0.29.

The accuracy was only slightly better at 76.7%.

The implications of this LDA investigation suggest are that Dominant Landcover may be an indicator of the levels of lake health variables in a given lake. Therefore, if we know whether a lake has native or pastoral landcover, this could give us an indication on the health of a lake. It also means that in the case that the Dominant Landcover observation is lost for a lake, the value may be able to be imputed (with a note mentioning the imputation) with some level of accuracy by using the values from the Lake Health variables.

4 Discussion

- WHAT DOES IT MEEEAAAAAANNNNNNN
- problems
  - clarity (not always possible as you may just have a shallow lake).
  - blank region
  - non-normality
  - landcover other

Native dff to every other landcover in terms of lake health pca first 2 83.4% factor 2 not suff

focus resources to north island, specifically wikato

5 Bibliography

- I have some references from research into health variables and where data from
- APA referencing

data: https://www.stats.govt.nz/indicators/modelled-lake-water-quality/