Group 8 Final Presentation

Russell and Frances

September 2022

Group 8

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

Textual Description of the Dataset

Our dataset contains information from 3801 lakes in New Zealand. This dataset was extracted from Stats NZ.
https://www.stats.govt.nz/indicators/modelled-lake-water-quality/

Variables

For analysis we split the dataset into two main categories; the lake health variables and the lake dimension variables. The lake health variables measure as a whole give an indication of the "health" of an individual lake. The five lake health variables are Clarity, Ammoniacal Nitrogen, Total Nitrogen, Total phosphorus, and Chlorophyll-A. Additional variables that we examined were:
- The lake dimension variables measure the dimensions of the lake. The three lake dimension variables are depth, area, perimeter.
- Region; which New Zealand region the lake was located in, and
- Dominant landcover; split into five types; Exotic Forest, Native, Pastoral, Urban area and other.

Lake Health Variables

Ammoniacal Nitrogen is a form of nitrogen that supports algae and plant growth, but in large concentrations can be toxic to aquatic life.

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.

Total Phosphorus is the sum of all phosphorus forms in the water. 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 Nitrogen is the sum of all nitrogens found in the water. An excess of nitrogen in lakes can cause an increase in algae and plant growth, possibly depriving the lake of oxygen.

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.

Exploratory Data Analysis

First we analysed the distribution of the lake health variables. Figure 1 shows the visualisation of the correlation matrix for the lake health variables.

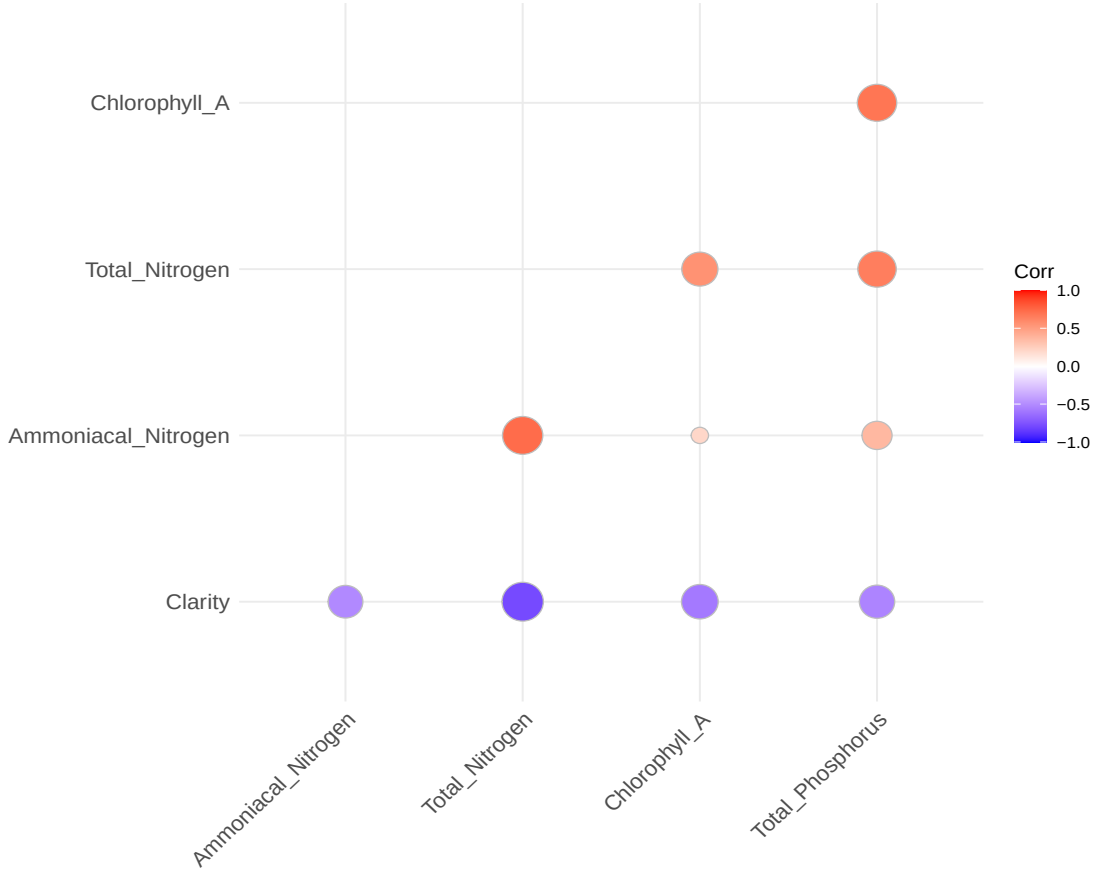


Figure 1: Visualisation of the Correlation Matrix

Exploratory Data Analysis

Next, we wanted to compare the distribution of the lake health variables by types of dominant landcover. The side-by-side boxplots are shown in figure 2.

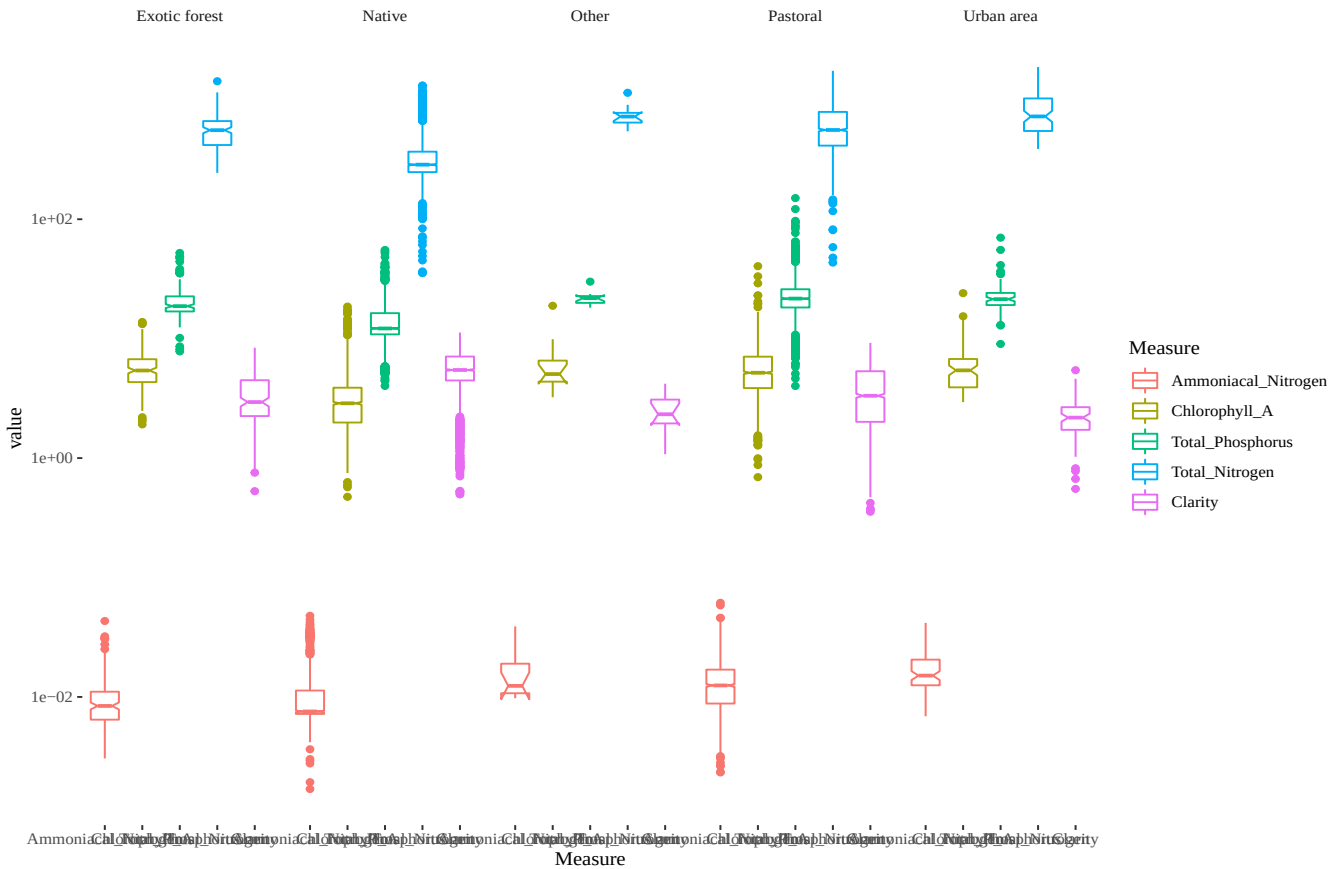


Figure 2: Box Plots of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity by Landcover

Exploratory Data Analysis

The pairs plot of the lake health and dimension variables, coloured by dominant landcover, is shown in figure 3.

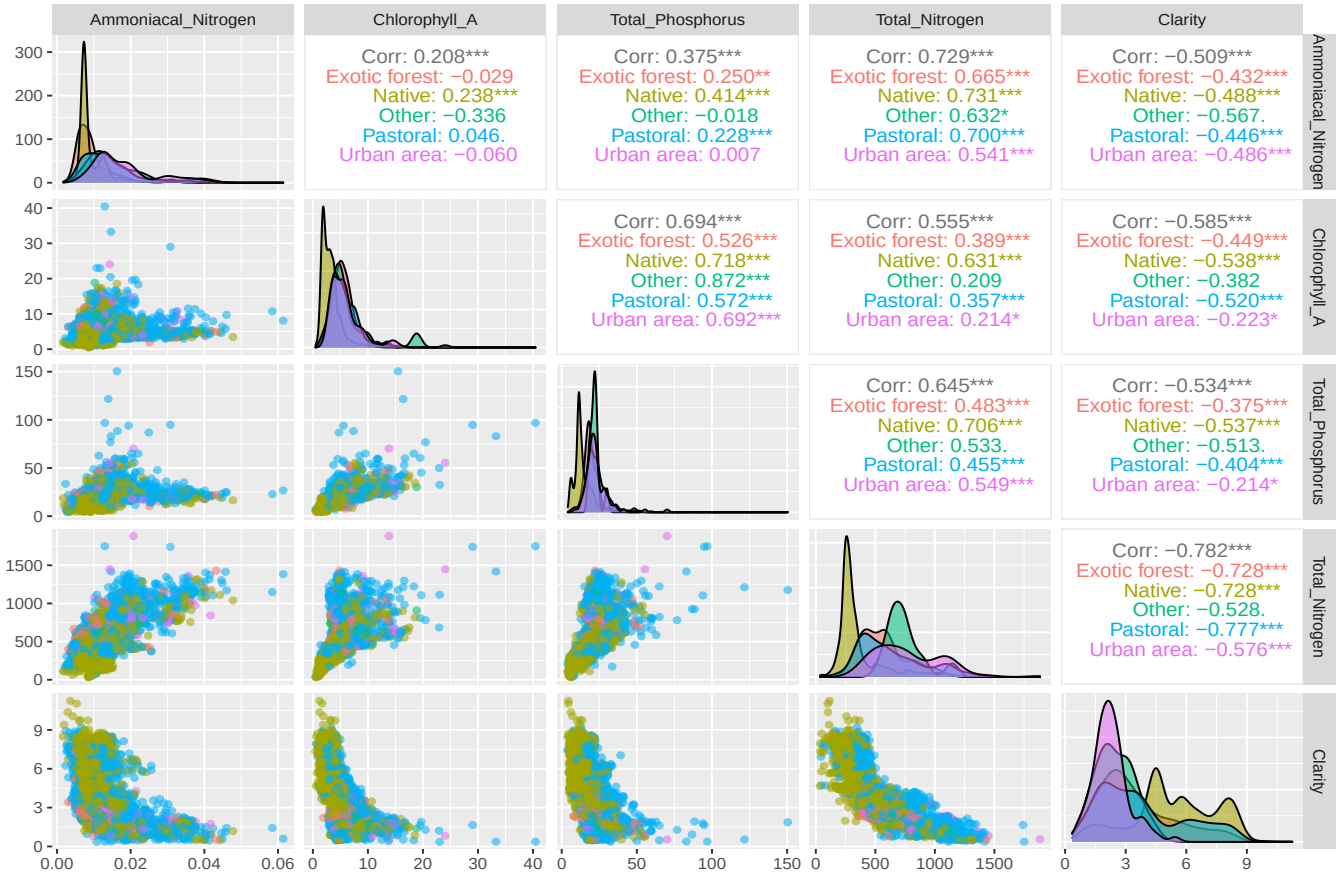


Figure 3: Pairs Plot

Exploratory Data Analysis

Next we wanted to investigate which regions in particular had the worst median lake health. We constructed the following tables, with Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen ordered in descending medians, and Clarity ordered in ascending medians.

Table 1: 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
ManawatÅ«-Whanganui	0.0143978	0.0110985	Waikato	7.115944	6.156554	ManawatÅ«-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	ManawatÅ«-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

Exploratory Data Analysis

Below is the tables for the regional means and medians of Total Nitrogen and Clarity.

Table 2: 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

Exploratory Data Analysis

We also looked into the relationship between the dimension variables and region, shown in figure 4.

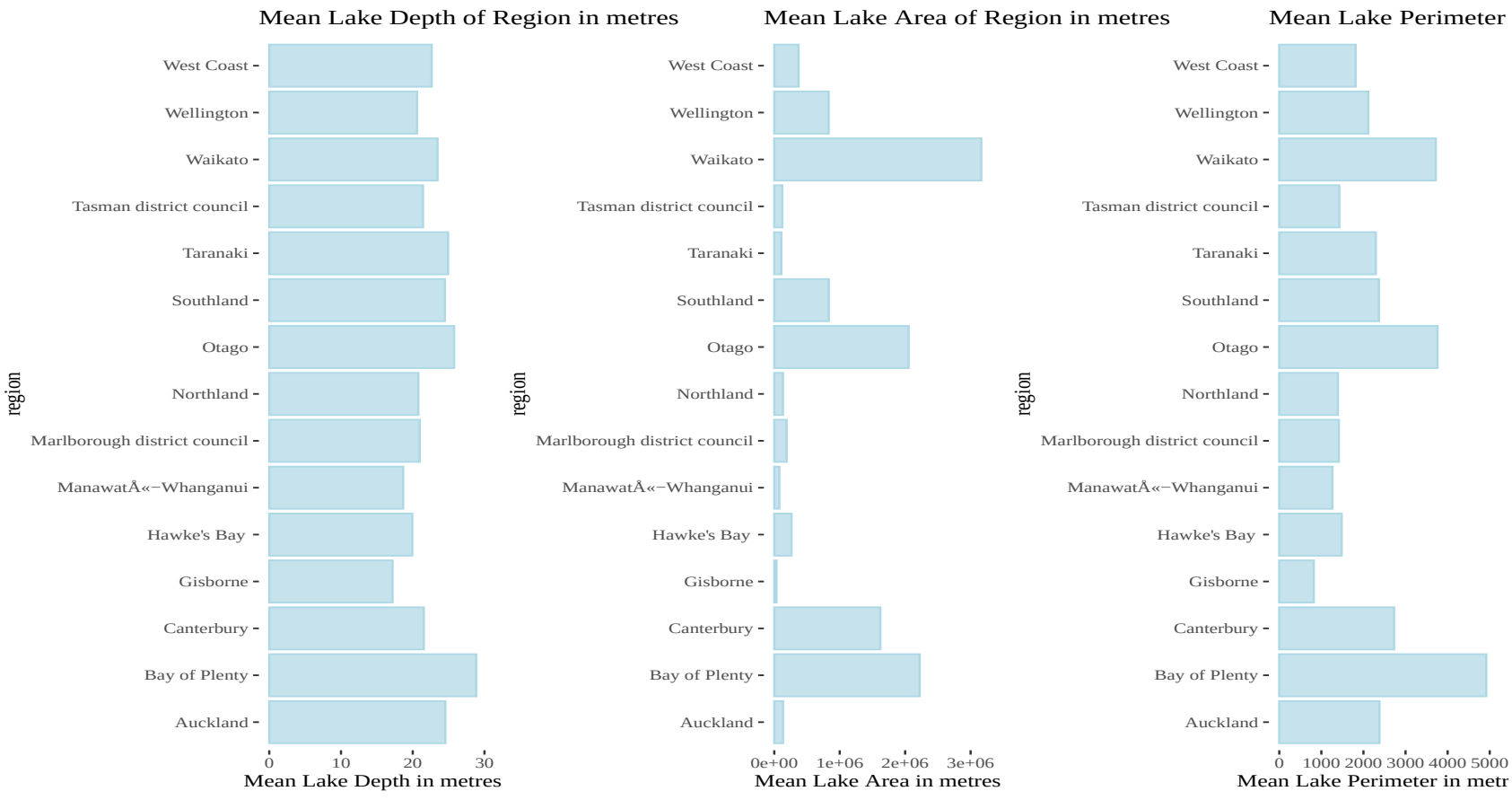


Figure 4: Bar Graphs of Lake Dimensions by Region

Exploratory Data Analysis

Figure 5 shows the pairs plot of the lake health variables by the lake dimension variables.

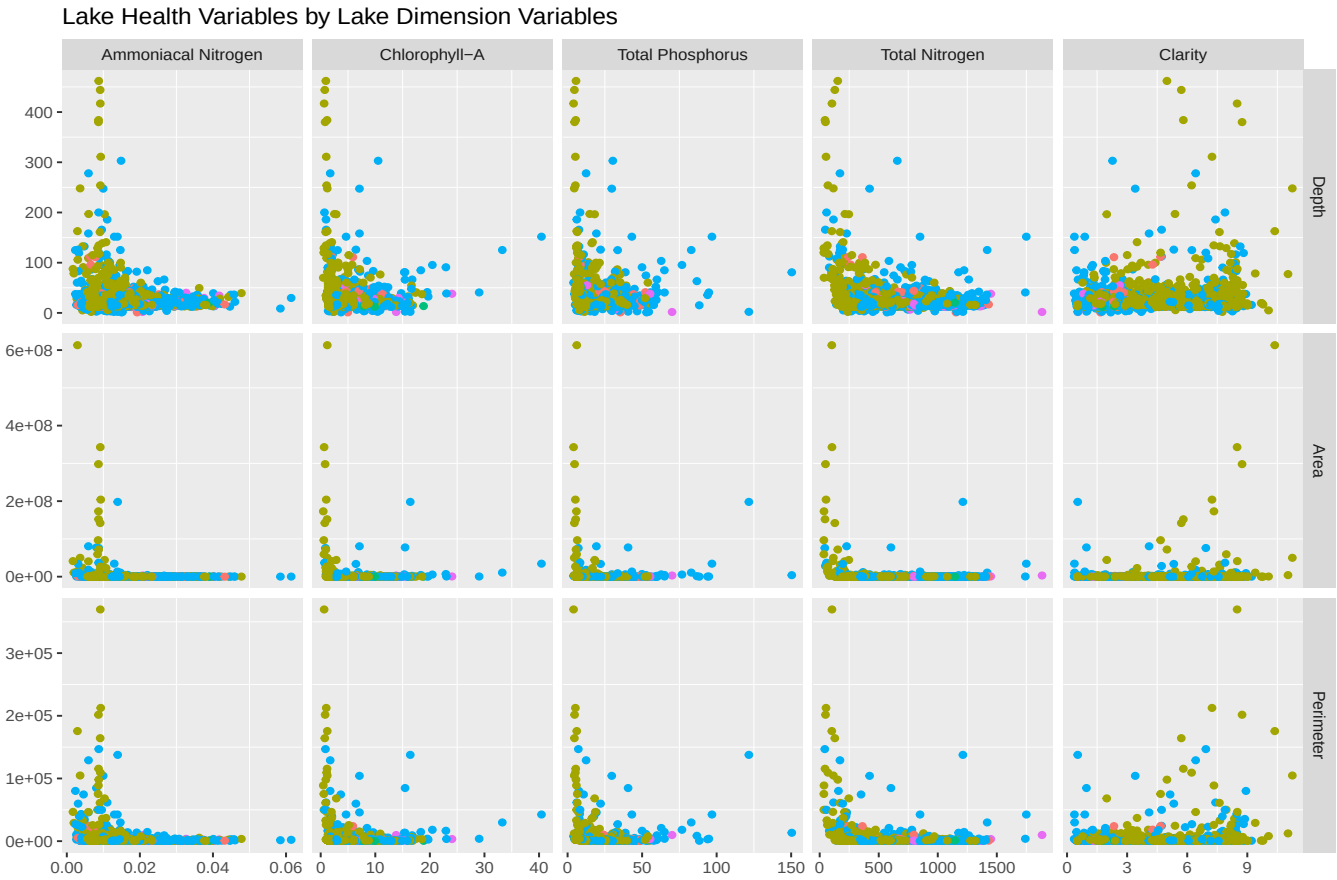


Figure 5: Pairs Plot of Lake Health Variables by Lake Dimension Variables

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?

Tools We Applied

In further analysis of these questions we applied the following tools;

- Cullen and Frey graphs
- Kruskal-Wallis tests with pairwise Wilcox tests
- Principal Component Analysis
- Factor analysis
- Linear discriminant analysis

Cullen and Frey

We used Cullen and Frey graphs on the lake health variables to try to determine what distribution they follow. An example for Ammoniacal Nitrogen is shown below in figure 6.

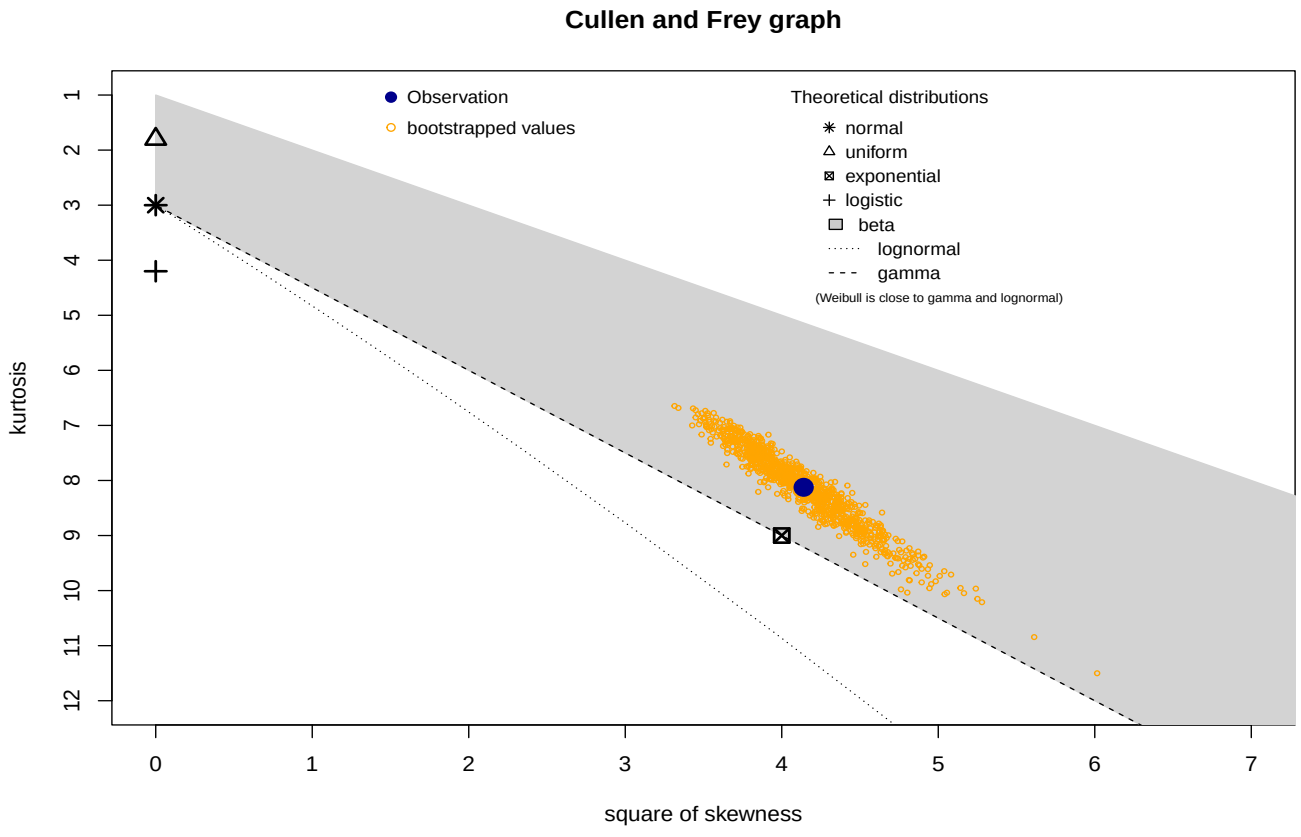


Figure 6: Cullen and Frey Graph of Ammoniacal Nitrogen

Kruskal-Wallis and Pairwise Wilcox Tests

We conducted Kruskal-Wallis tests for each lake health variable to determine whether there is a difference in those measures between types of dominant landcover. Following this, we conducted pairwise Wilcox tests. An example for clarity is shown below.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SECCHI by land
## Kruskal-Wallis chi-squared = 602.61, df = 4, p-value < 2.2e-1

##
##  Pairwise comparisons using Wilcoxon rank sum test with conti
##
## 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
```

Principal Component Analysis

We use Principal Component Analysis to try to simplify the complexity of the Lake Health variables, without losing trends and patterns. The summary of this analysis is shown below.

```
## 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
```

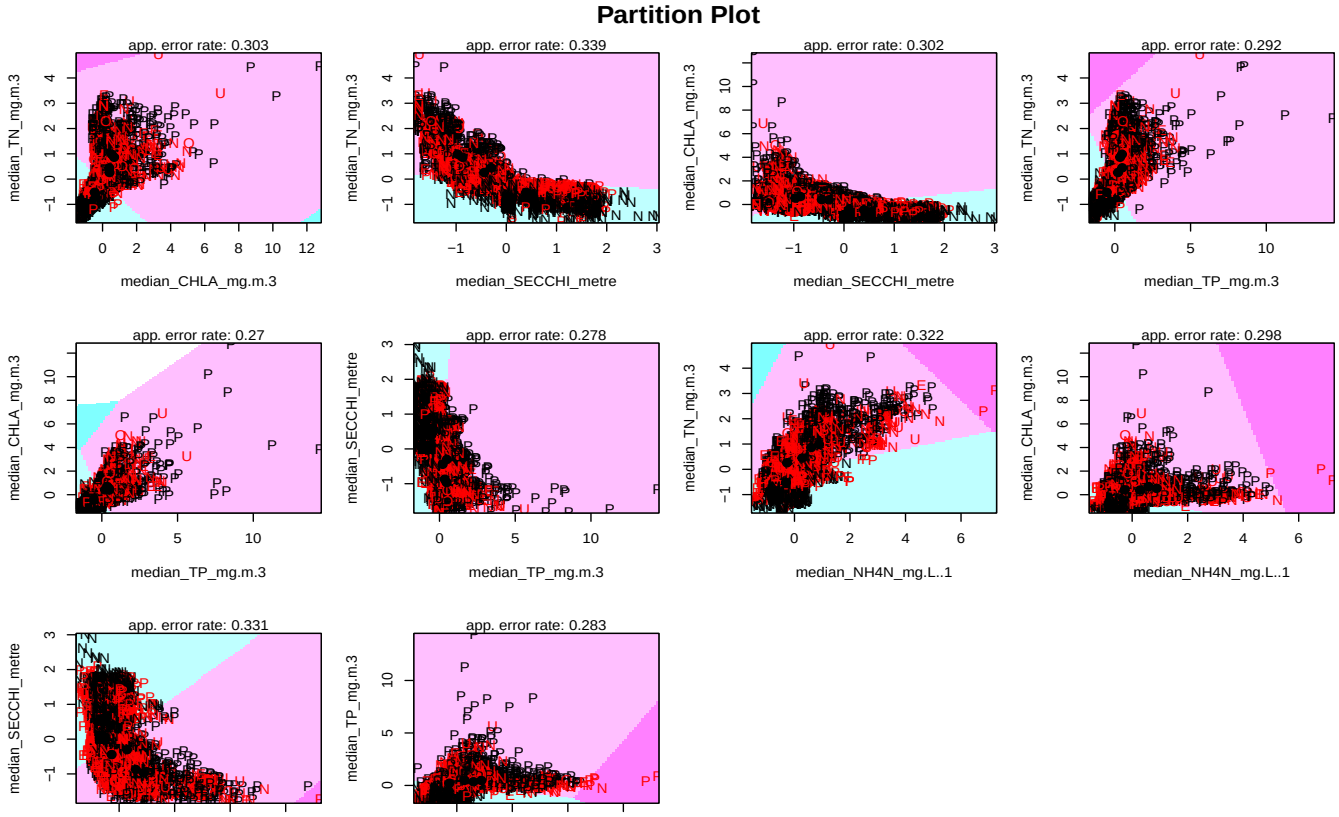
Factor Analysis

We conducted a factor analysis on the lake health variables with the null hypothesis that two factors is 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 less than 0.0001.

Linear Discriminant Analysis

We performed a Linear Discriminant Analysis on the Lake Health variables, to try make a partition to determine the type of dominant landcover, from the Lake Health variables. The model had about 73% accuracy.

We found that Pastoral Landcover and Native Landcover each explain about 46% of the prior probabilities. We then produced a partition plot to examine the separation.



Linear Discriminant Analysis

We decided to investigate just the Pastoral and Native classes as they accounted for about 93% of the observations in our sample so we produced a second partition plot with just these two classes. The accuracy was only slightly better at 76.7%.

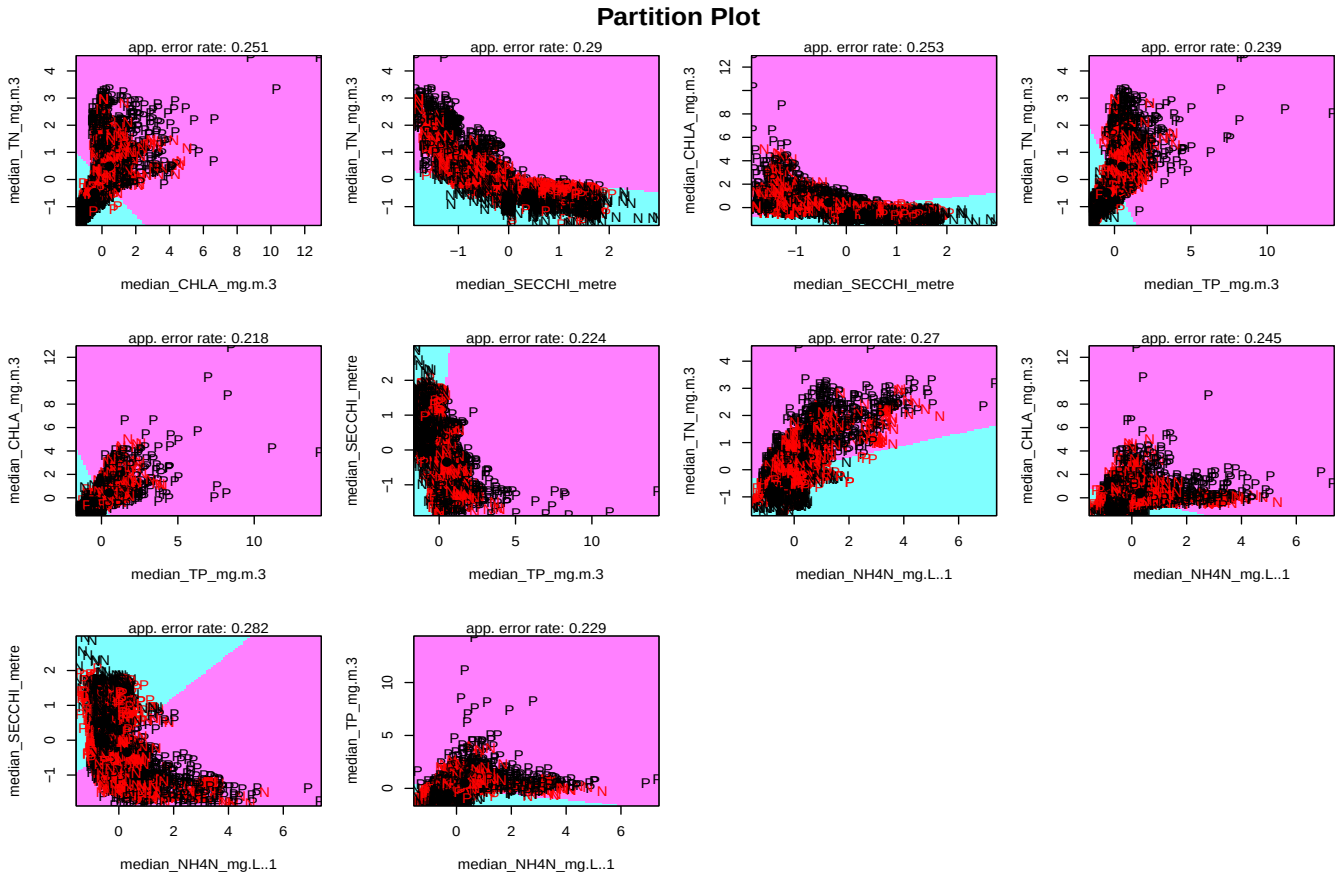


Figure 8: Partition Plot

Problems

The dataset contained 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.

We also found one lake with no entry for region so we excluded this observation when analysing region with the lake health or dimension variables further.

We wanted to conduct many more tests, such as ANOVA, which requires normally distributed data. However, our data was not normally distributed, so we used slightly different tests without the normality assumption.

We also tried to model relationships between the Lake Health variables using a linear model, however many of these relationships were not linear.

Conclusions

After our analysis, we can conclude the following:

- Waikato in particular, had higher levels of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen than other regions, and the lakes in Waikato tended to be murkier than many other regions.
- We could not come up with a model for the Lake Health or Lake Dimension variables so we cannot determine whether the Lake Health variables predict one another.
- Lake with predominantly Native landcover tended to have lower levels of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen and be clearer than other landcover types, while Pastoral and Urban tended to have higher levels of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen and were more murky.

Conclusion

Finally, after investigating Lake Health, Dimension, Region and Dominant Landcover, we concluded that, while lakes in the Waikato region tended to have poorer lake health than other regions, they tended to have a larger area, which could mean these lakes require more resources to treat. We noticed the lakes in the South Island tended to be healthier than lakes in the North Island, so we would recommend focusing resources to improve the health of New Zealand lakes to the North Island, in particular, in Waikato.