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 variable. 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.

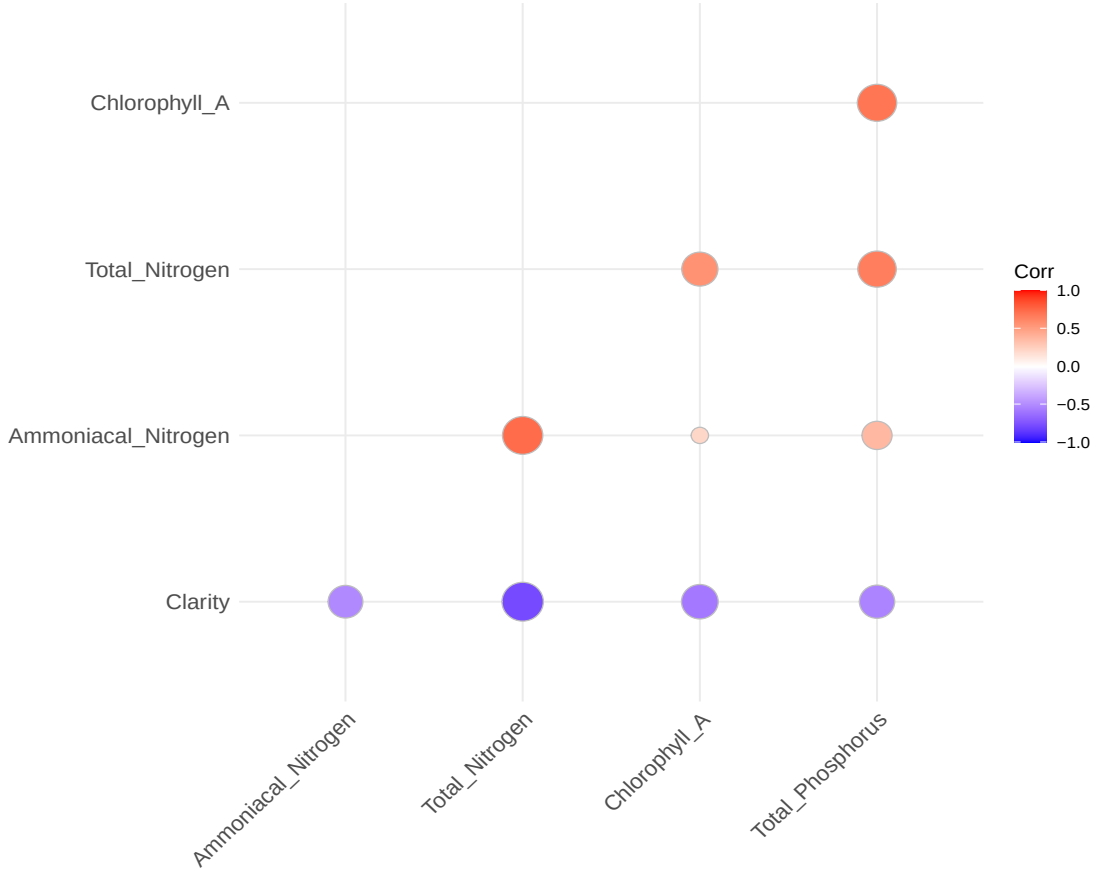

Figure 1: Visualisation of the Correlation Matrix

Exploratory Data Analysis

The pairs plot of the lake health and dimension variables, coloured by dominant landcover, is shown in figure 2.

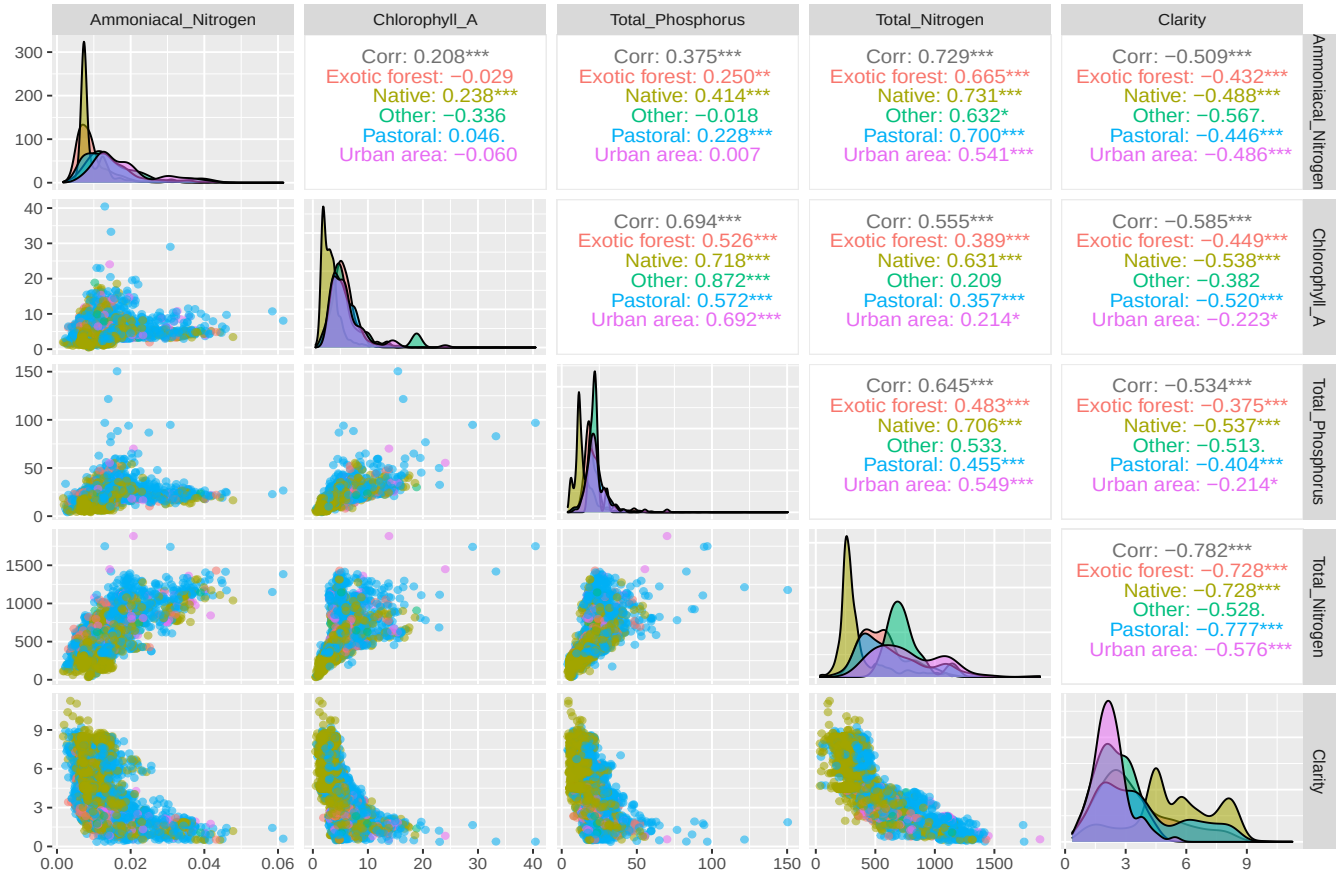

Figure 2: Pairs Plot

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 3.

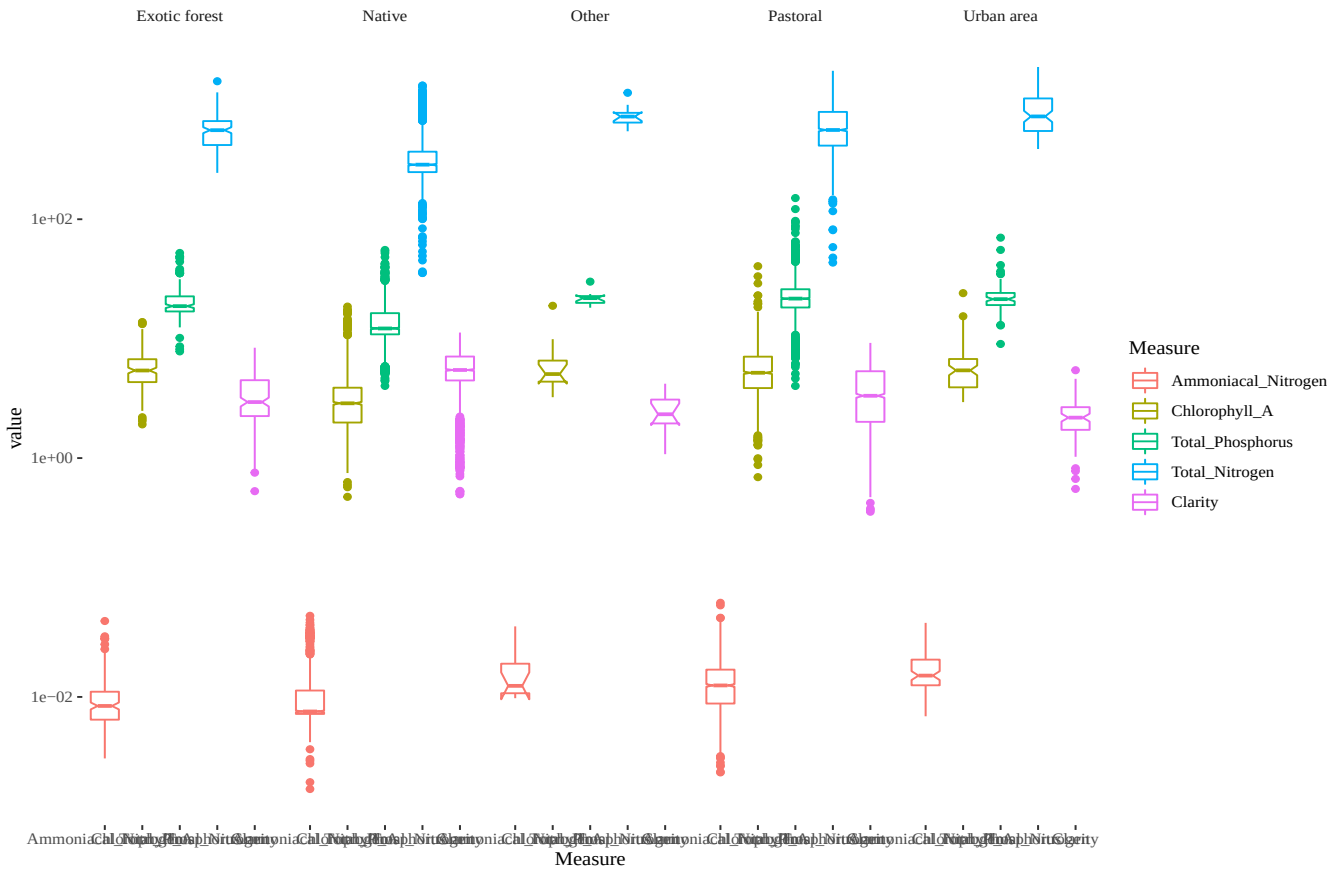

Figure 3: Box Plots of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity by Landcover

Exploratory Data Analysis

Next we analysed the distribution of the lake health variables by region. The box plots of Ammoniacal Nitrogen by region is shown in figure 4.

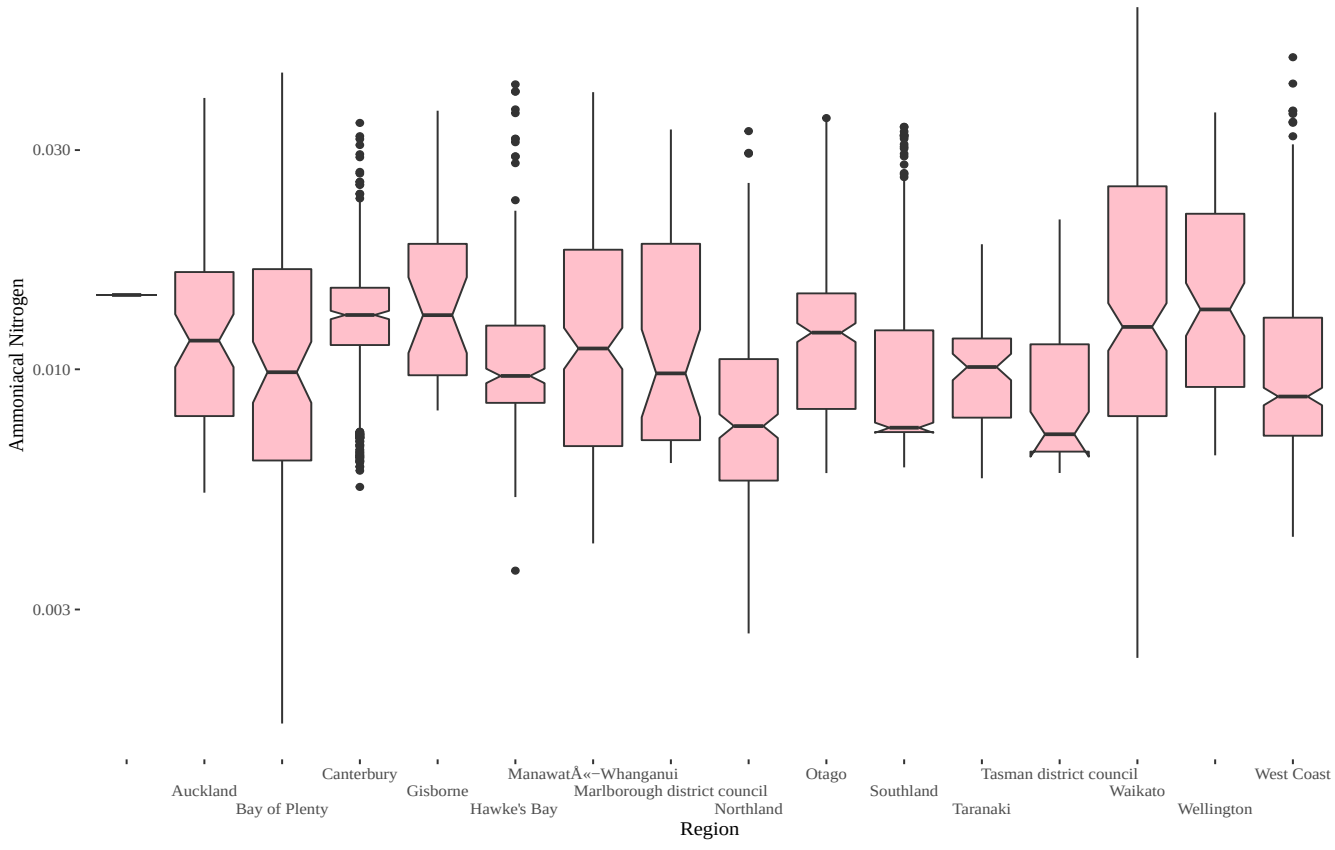

Figure 4: Box Plot of Region and Ammoniacal Nitrogen

Exploratory Data Analysis

The box plots of Chlorophyll-A by region is shown in figure 5.

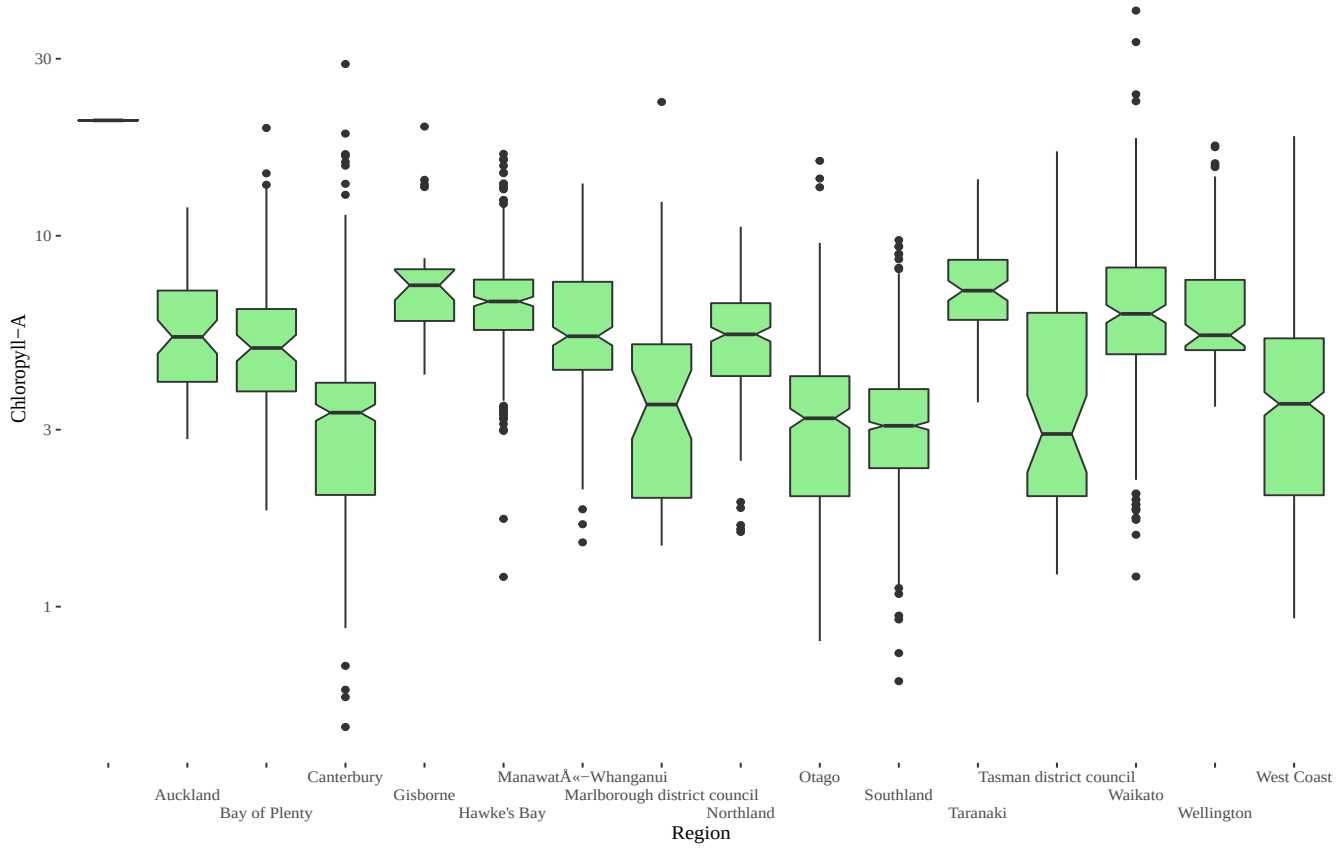

Figure 5: Box Plot of Region and Chloropyll-A

Exploratory Data Analysis

The box plots of Total Phosphorus by region is shown below in figure 6.

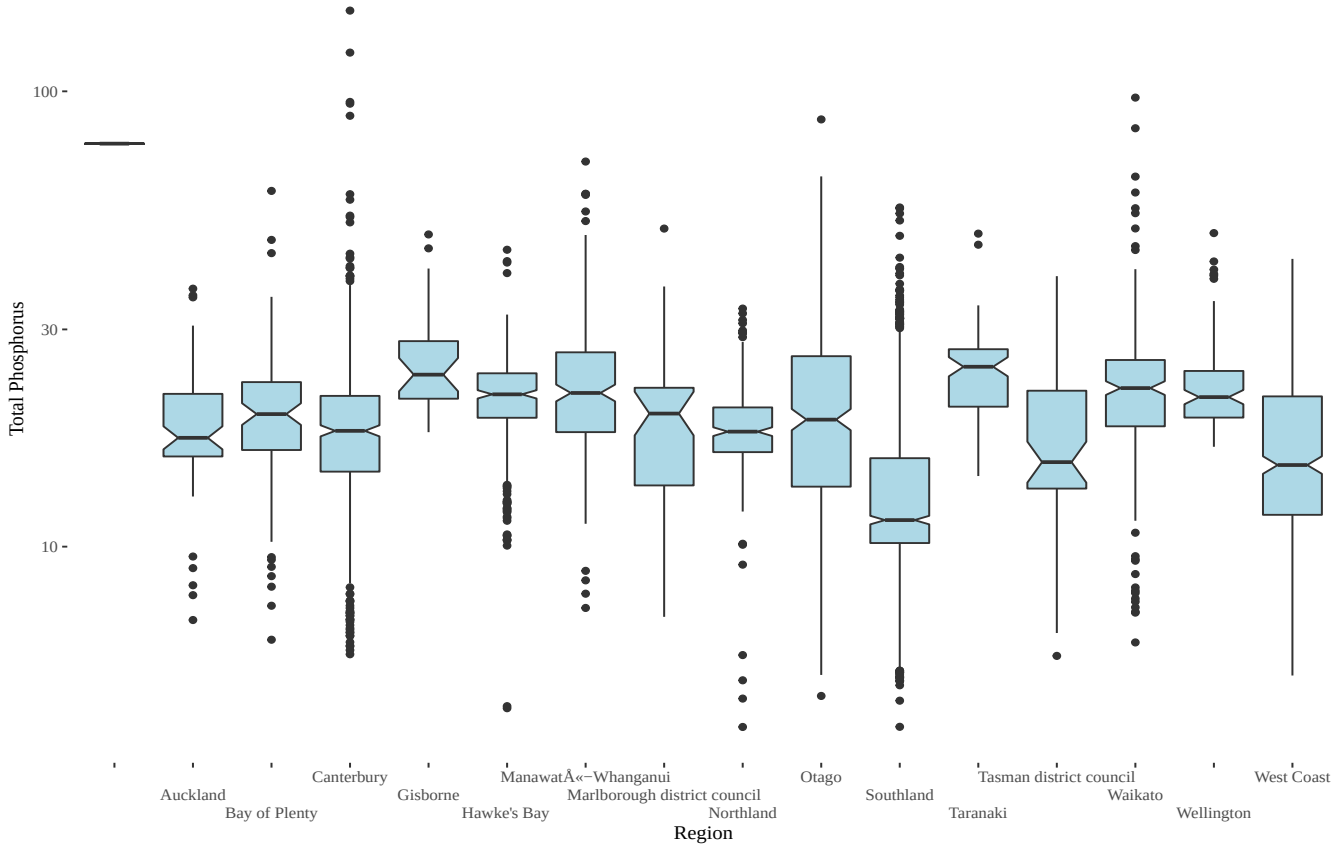

Figure 6: Box Plot of Region and Total Phosphorus

Exploratory Data Analysis

The box plots of Total Nitrogen by region is shown in figure 7.

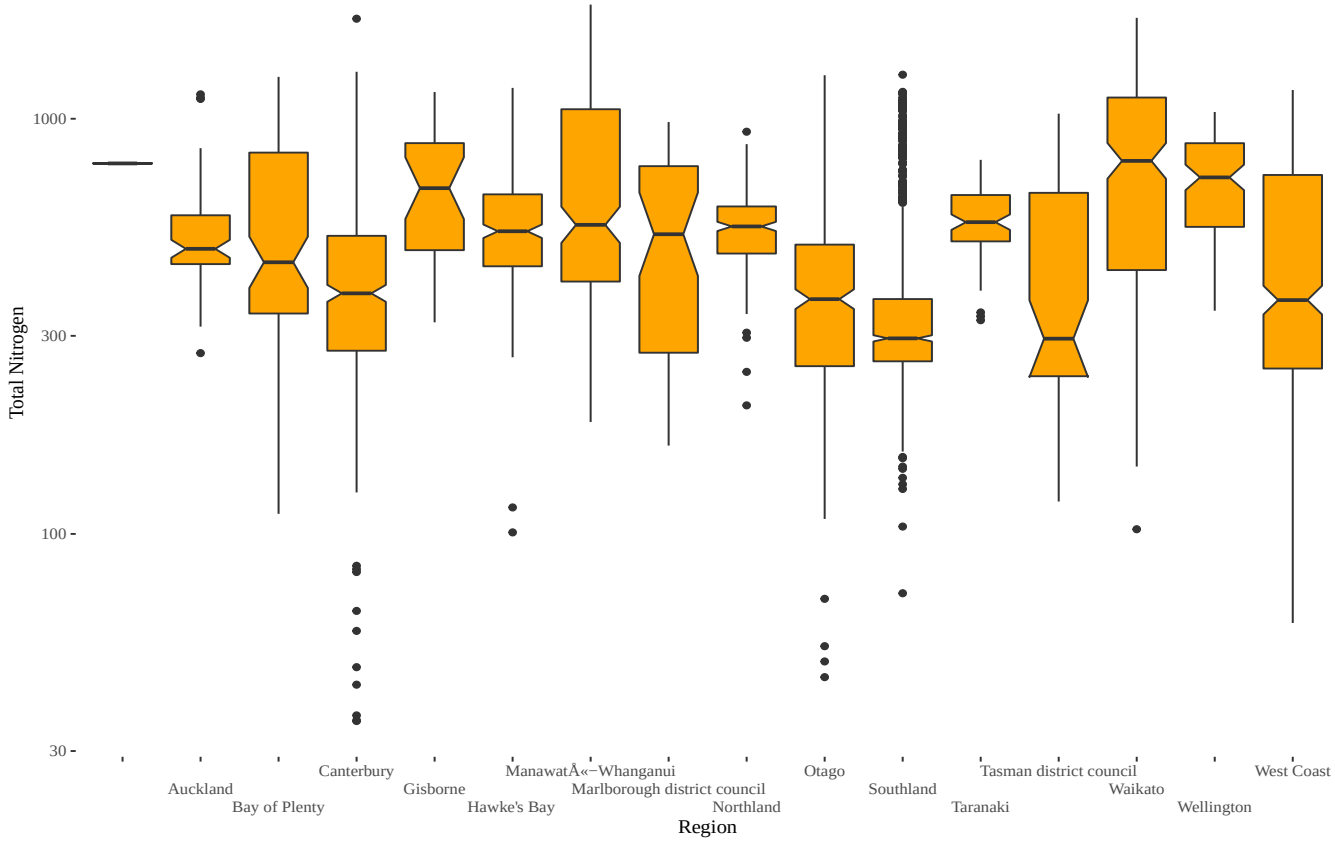

Figure 7: Box Plot of Region and Total Nitrogen

Exploratory Data Analysis

Figure 8 shows the box plots of clarity by region.

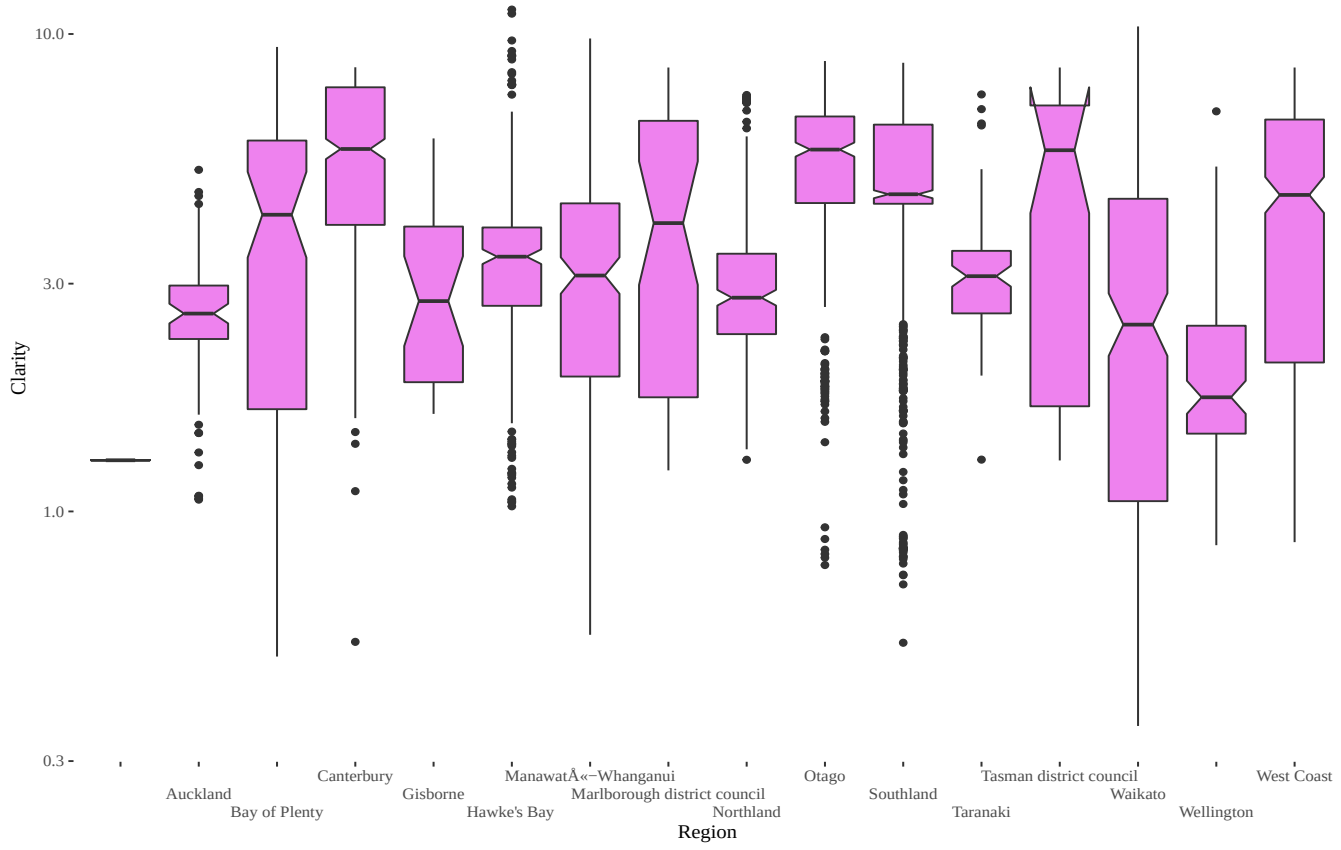

Figure 8: Box Plot of Region and Clarity

Exploratory Data Analysis

We also looked into the relationship between the dimension variables and region, shown in figure 9.

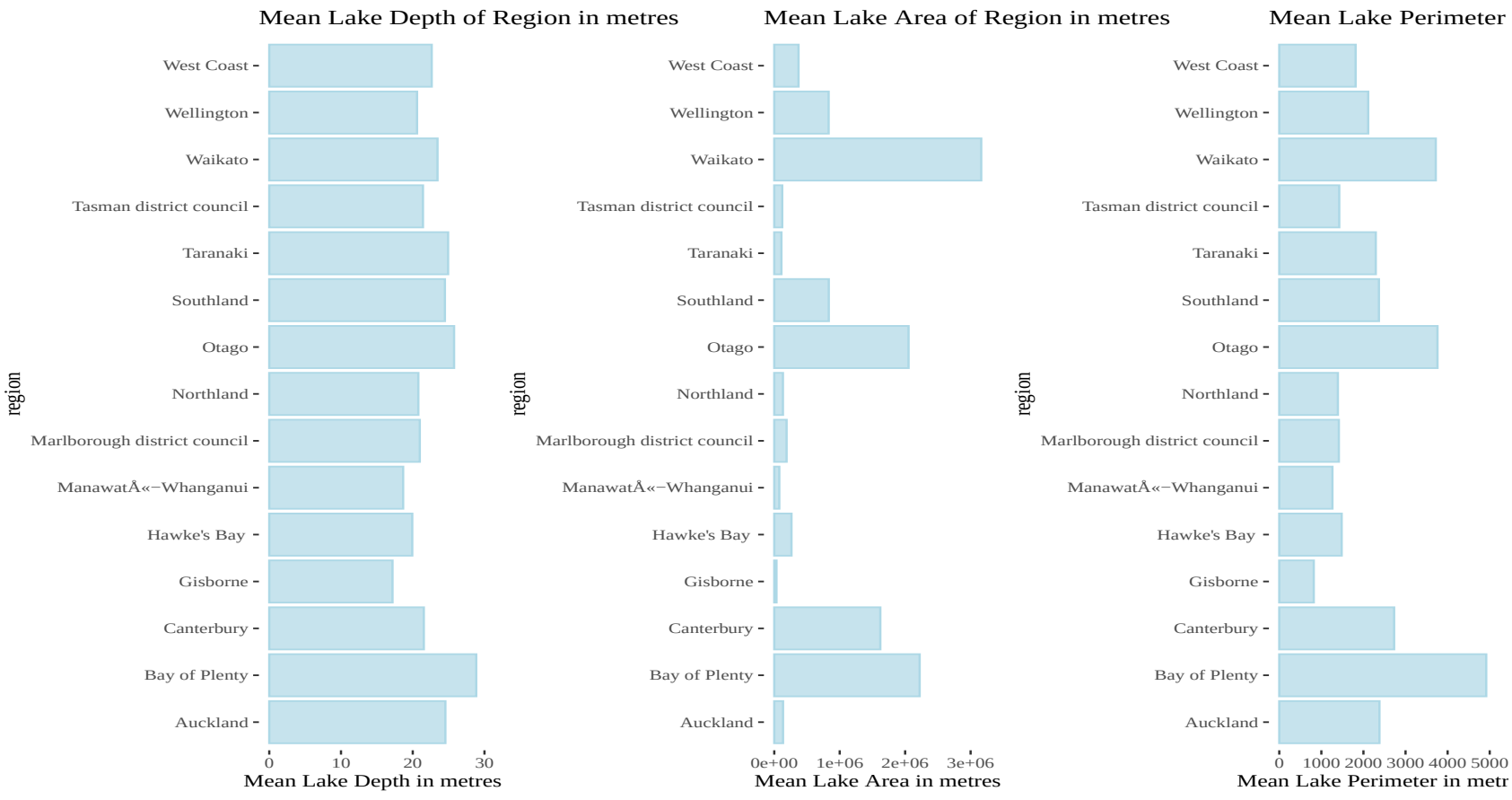

Figure 9: Bar Graphs of Lake Dimensions by Region

Exploratory Data Analysis

Figure 10 shows the pairs plot of the lake health variables by the lake dimension variables.

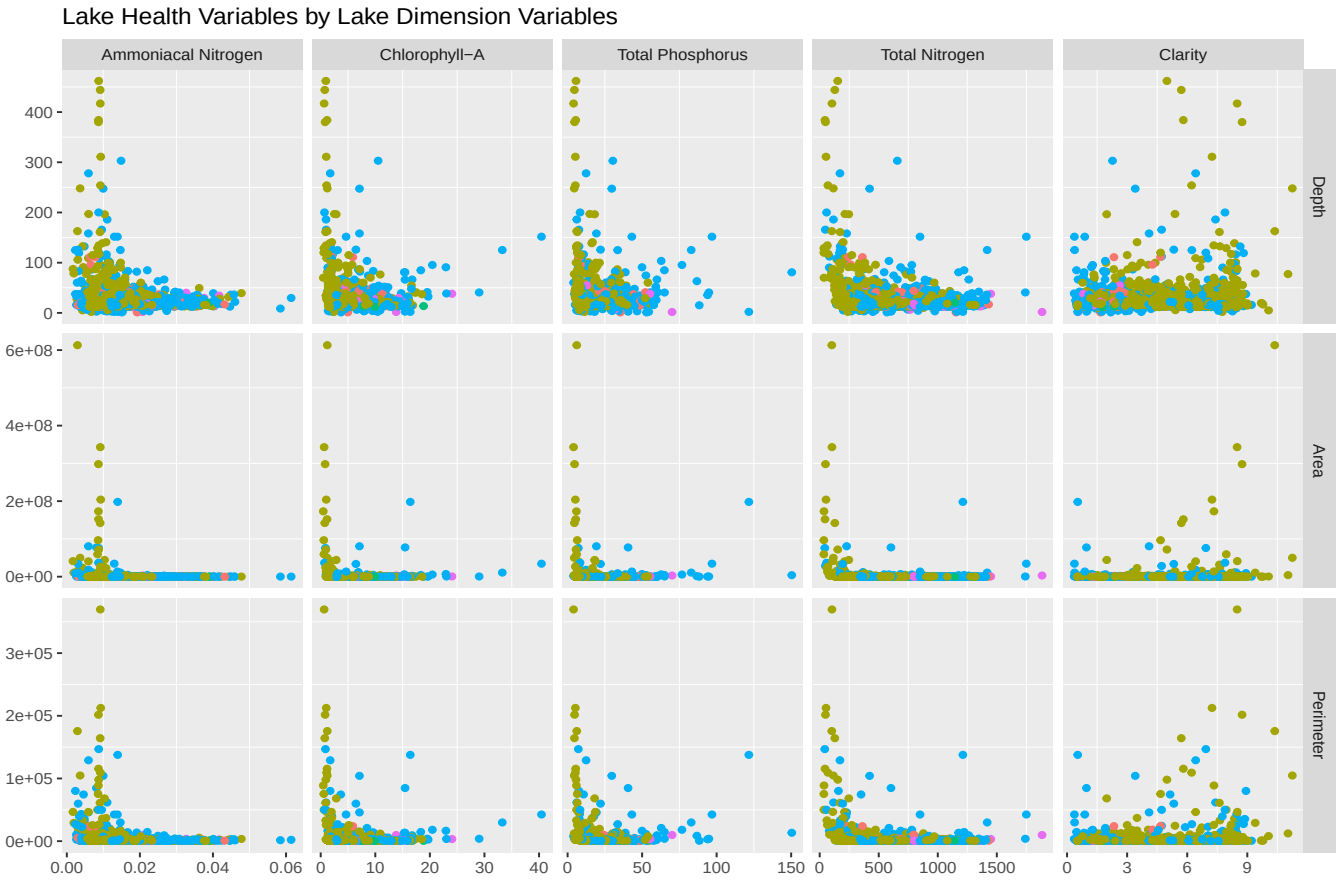

Figure 10: Pairs Plot of Lake Health Variables by Lake Dimension Variables

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
- 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 11.

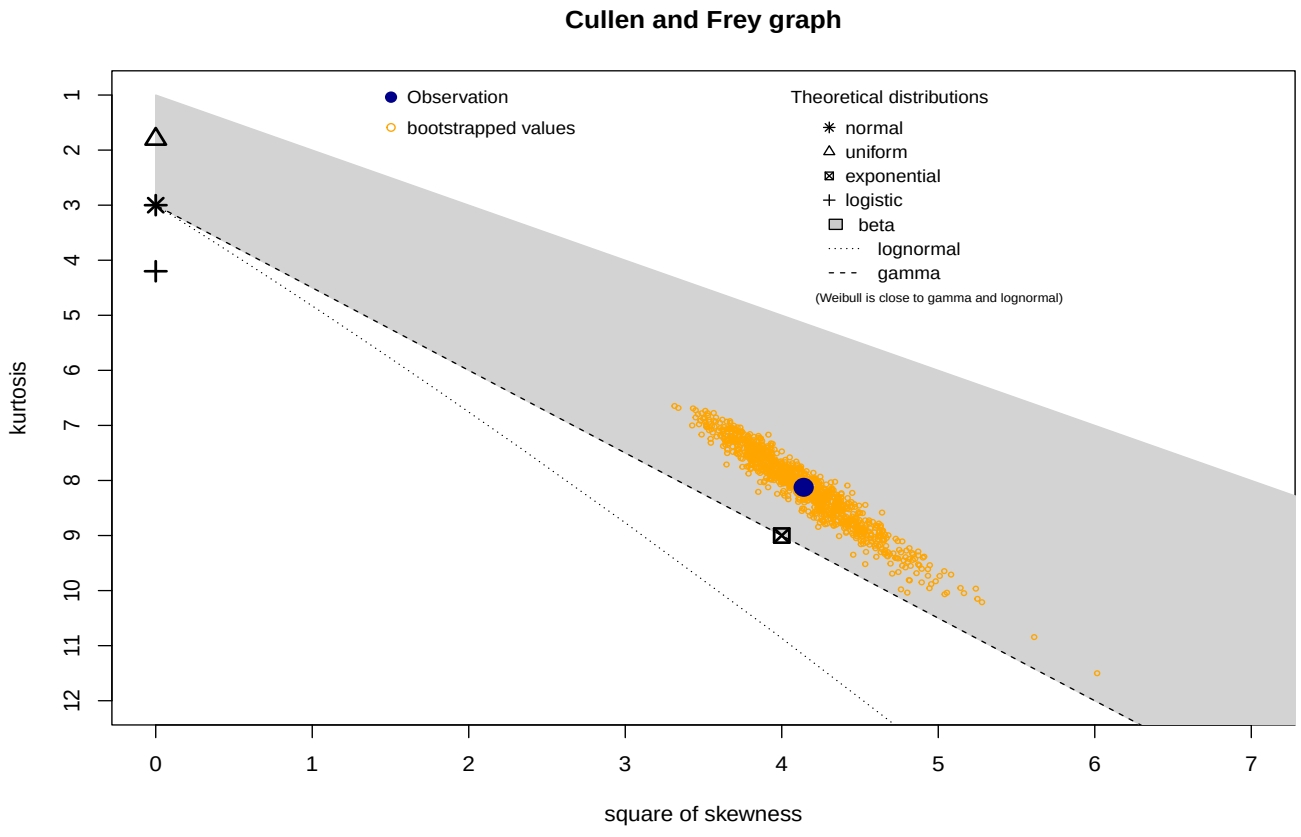

Figure 11: Cullen and Frey Graph of Ammoniacal Nitrogen

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
```

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

FILL THIS IN

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.

Conclusions

After our analysis, we can conclude that .....