Final Report

Russell and Frances

5th October 2022

Contents

1 Health of New Zealand Lakes

1.1 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

1.1.1 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.

1.1.2 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?

1.2 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
    - Bivariate 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.

1.2.1 Exploratory Data Analysis

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
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.

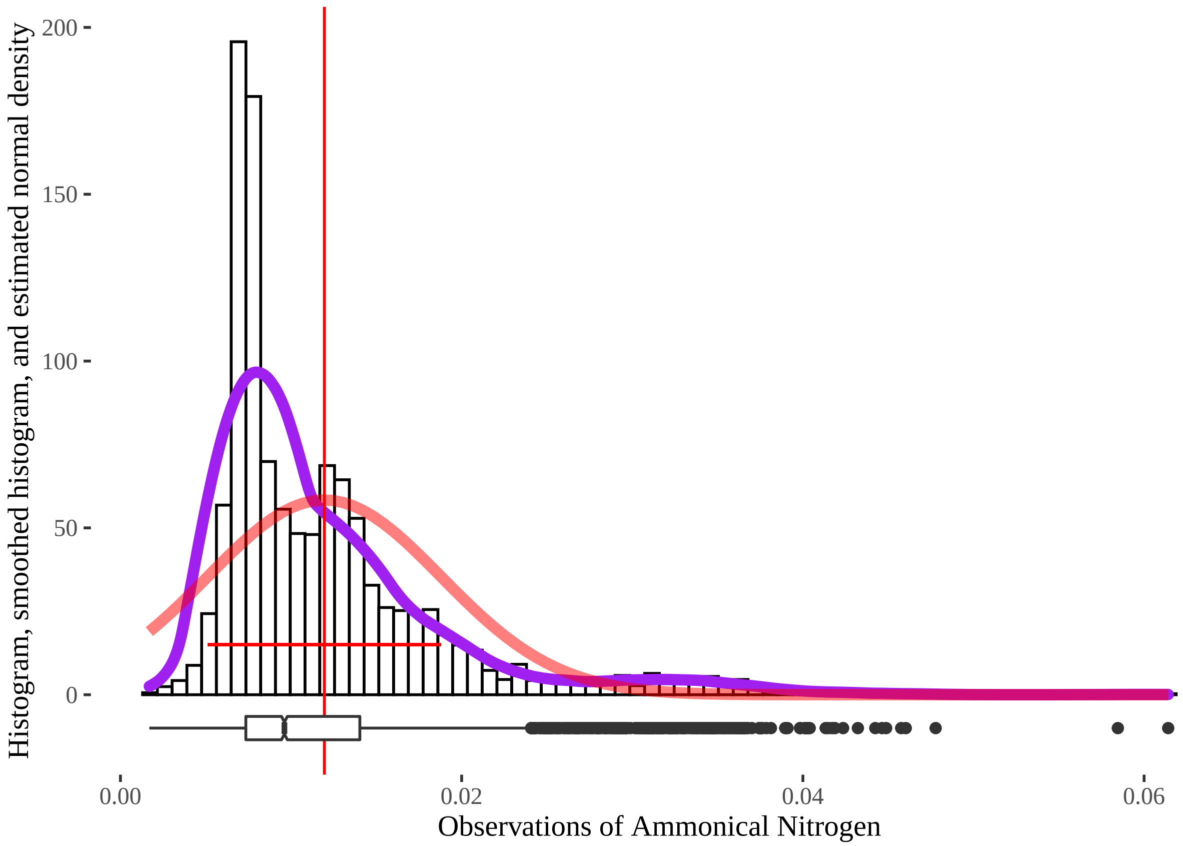

Figure 1: Histogram of Ammoniacal Nitrogen

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.

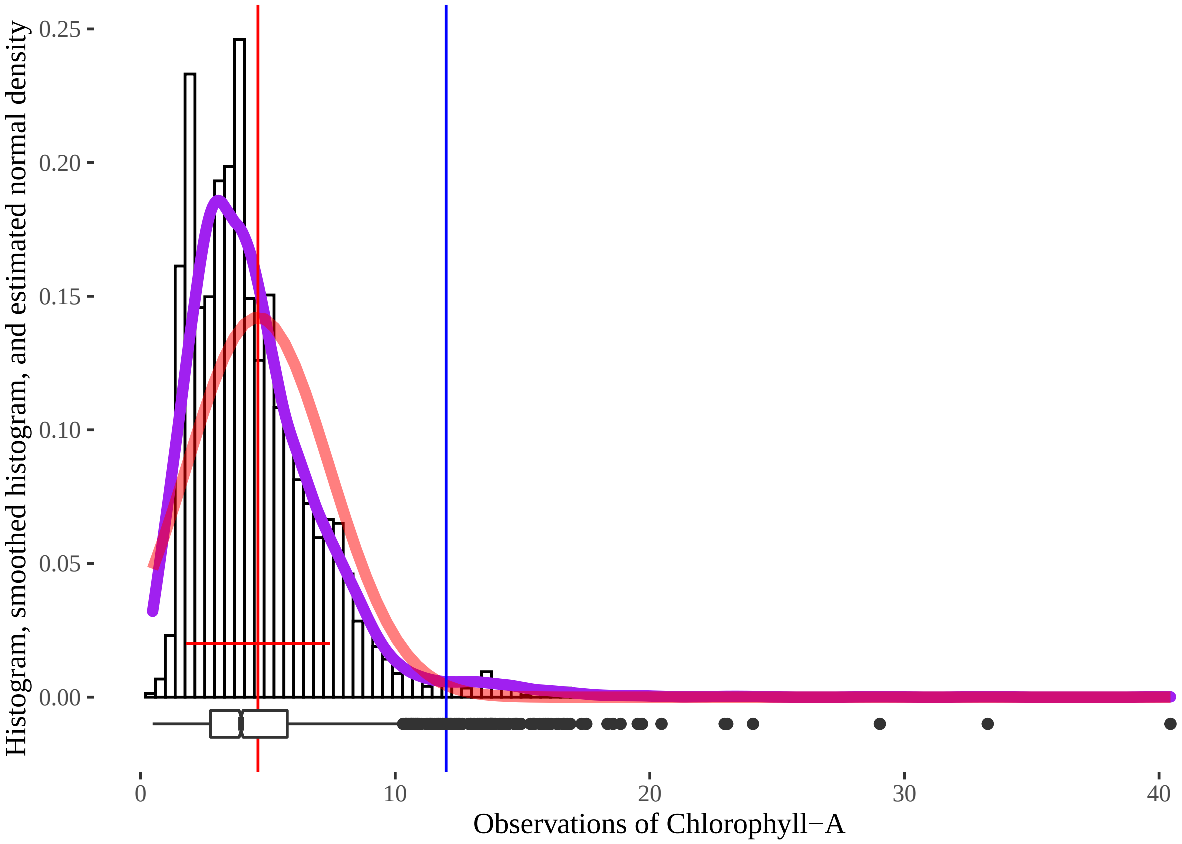

Figure 2: Histogram of Chlorophyll-A

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 natioal 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.

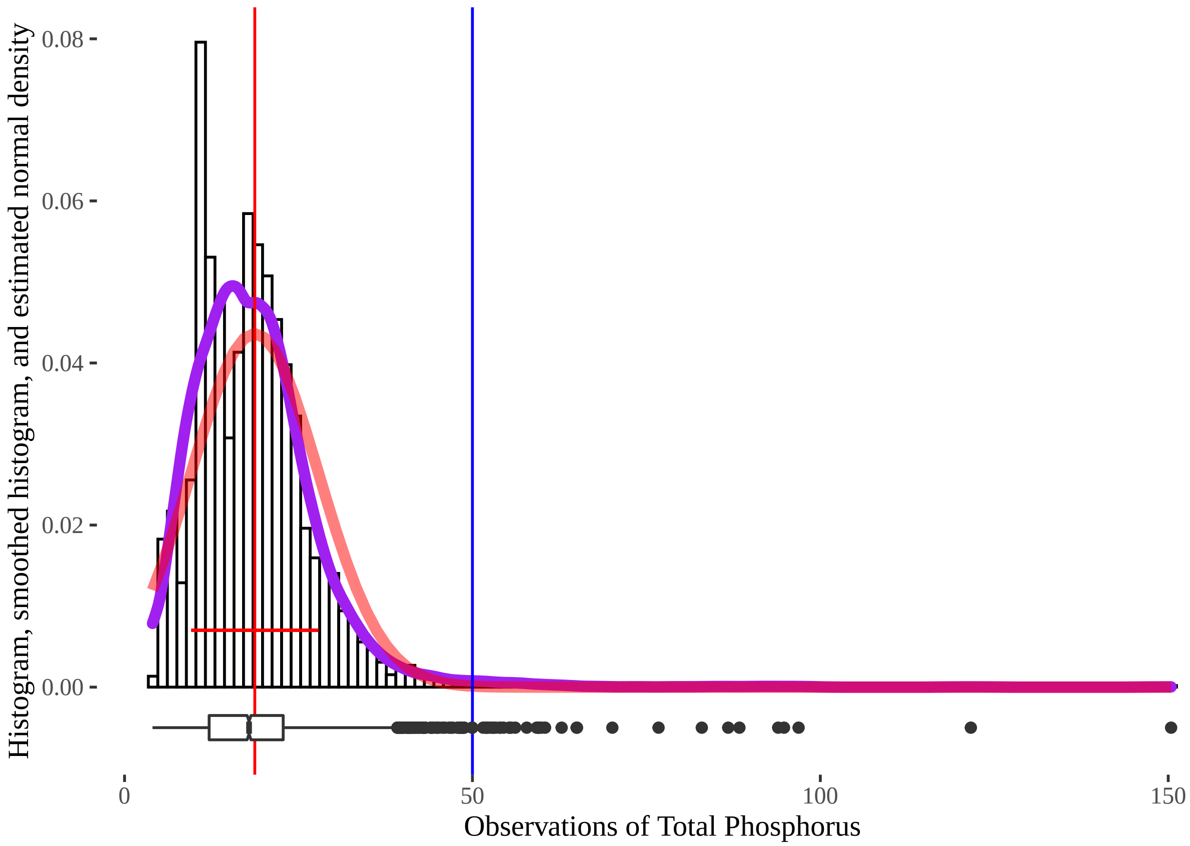

Figure 3: Histogram of Total Phosphorus

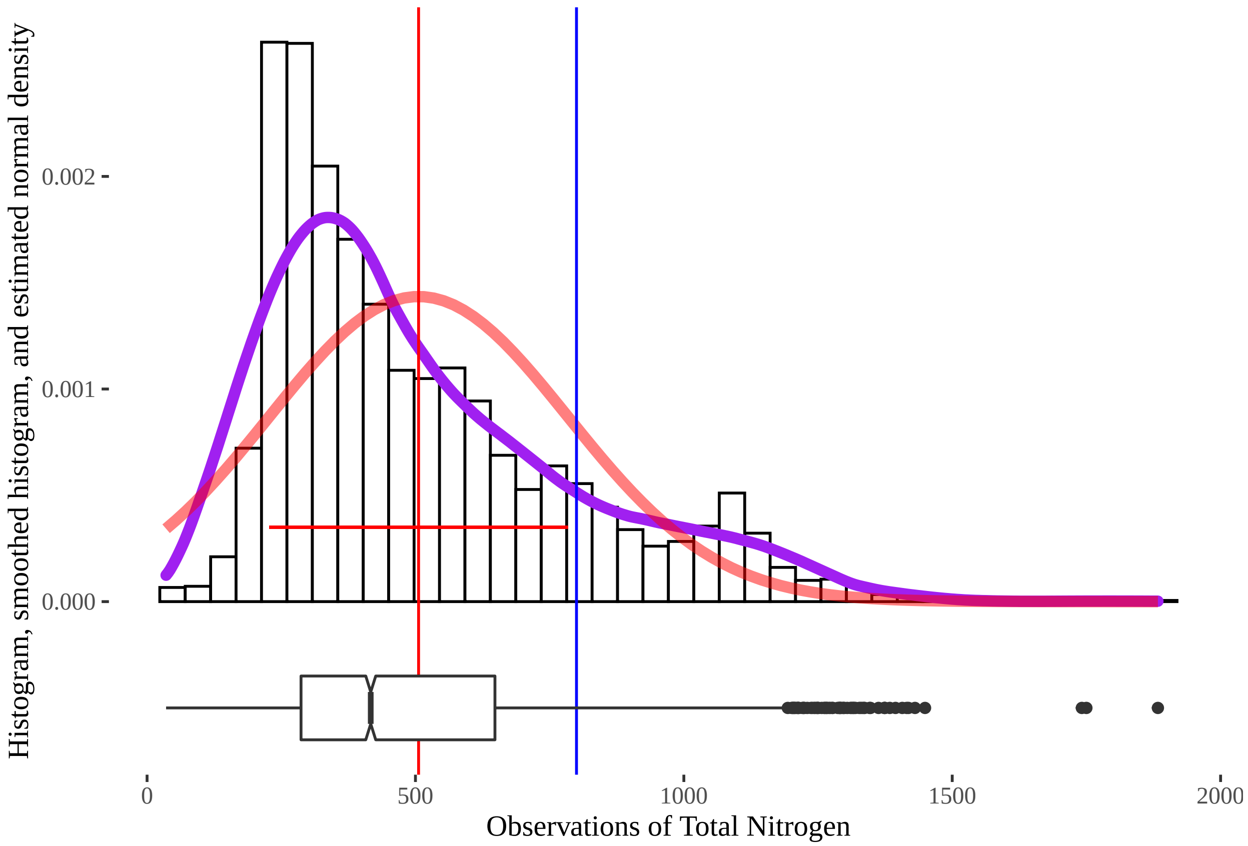

Figure 4: Histogram of Total Nitrogen

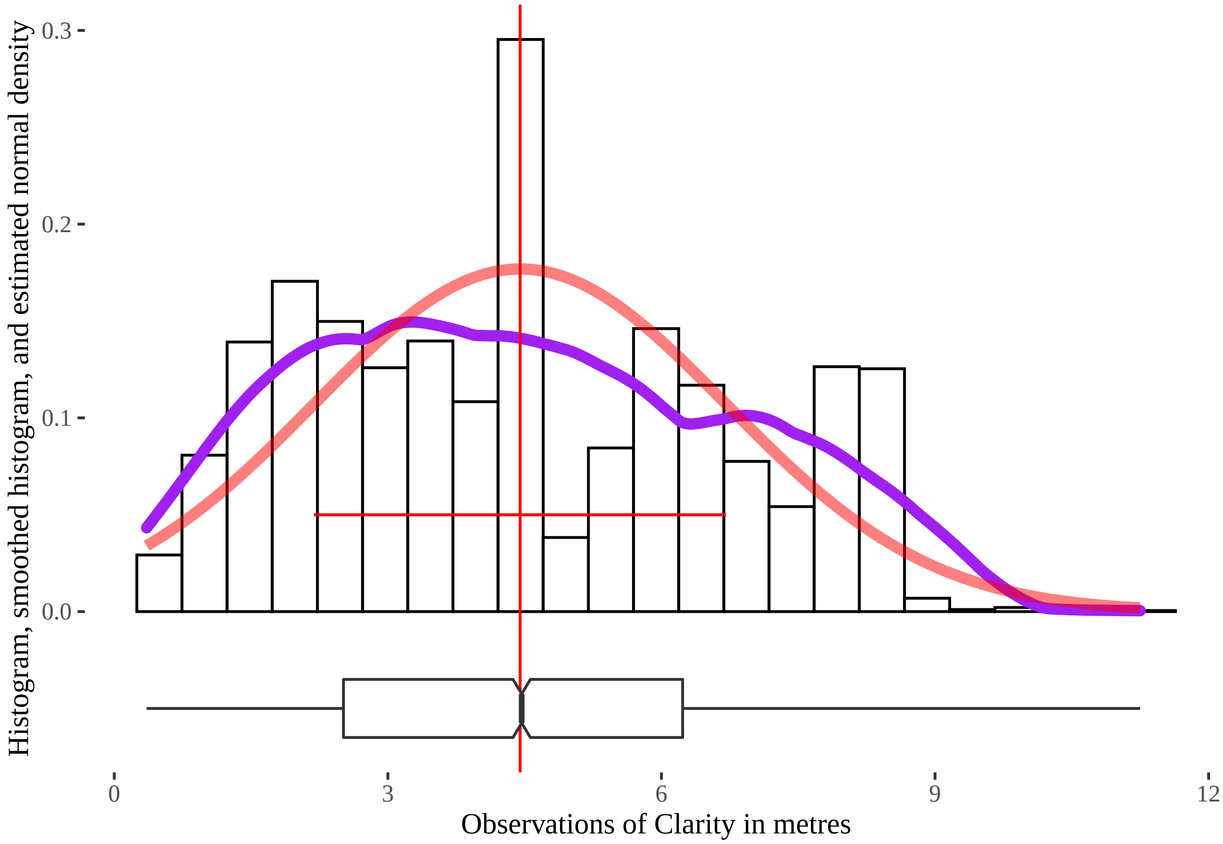

Figure 5: Histogram of Clarity (in metres)

1.4 Results

- test for difference in means health/land + pairwise tests
- PCA
- FA
- LDA

1.5 Discussion

- WHAT DOES IT MEEEAAAAAANNNNNNN
- problems
  - clarity (not always possible as you may just have a shallow lake).

1.6 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/