Group8_ReportPreliminary

Russell and Frances

5th September 2022

Contents

1 Report Preliminary

In this preliminary we include:

```
1. An introduction to our data.
2. A summary our previously conducted EDA.
3. A discussion of what we wish to explore further and why.
4. A show of our further exploration.
```

1.1 1. Introduction to our data:

Our dataset contains information from lakes in New Zealand.

Two groupings of variables are what we refer to as "Lake Health Variables" (median Ammoniacal Nitrogen, median Chlorophyll-A, median Clarity, median Total Nitrogen and median Total Phosphorus), and the "Lake Dimension Variables" (area, depth and perimeter).

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

1.1.1 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.

1.2 2. Summary of previously conducted EDA

In our previous EDA, we investigated the relationship between 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. We classified these variables as relating to lake health and thus refer to them as the “lake health variables”.

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

We also examined Regional differences in the Lake heath variables and Lake Dimension variables.

1.2.1 2.1 Summary of lake health variables investigation:

Please note that below is not all of the components of the EDA that we conducted, just a summary

Below is a table of the summary statistics of the lake health variables: table 1:

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

	Ammoniacal Nitrogen	Chloropyll-A	Phosphorus	Nitrogen	Clarity
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

We also produced histograms of all of the lake health variables. An example of this is seen below where where we model Chlorophyll-A in lakes

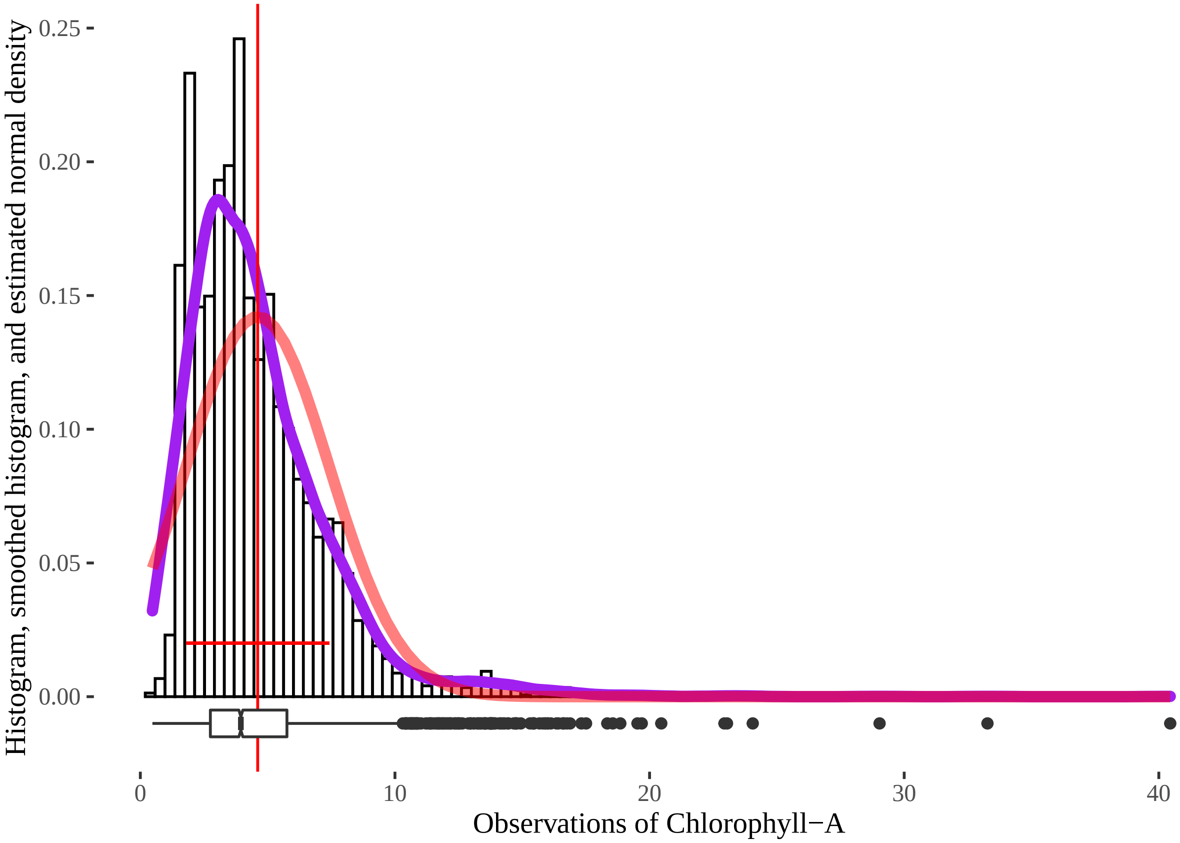


Figure 1: Histogram of Chlorophyll-A

Figure 1 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.

Figure 1 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.

We also examined the correlation between the lake health variables. A visualisation of this is shown below:

We then examined a pairs plot of the lake health variables

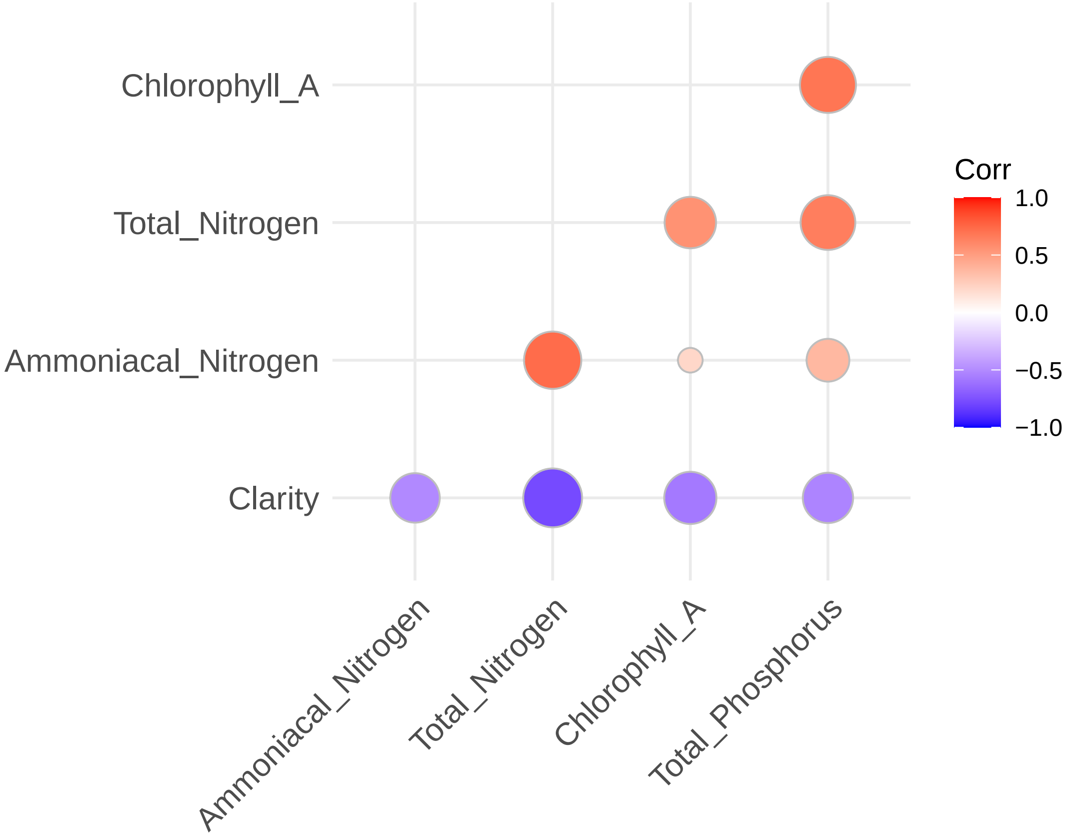
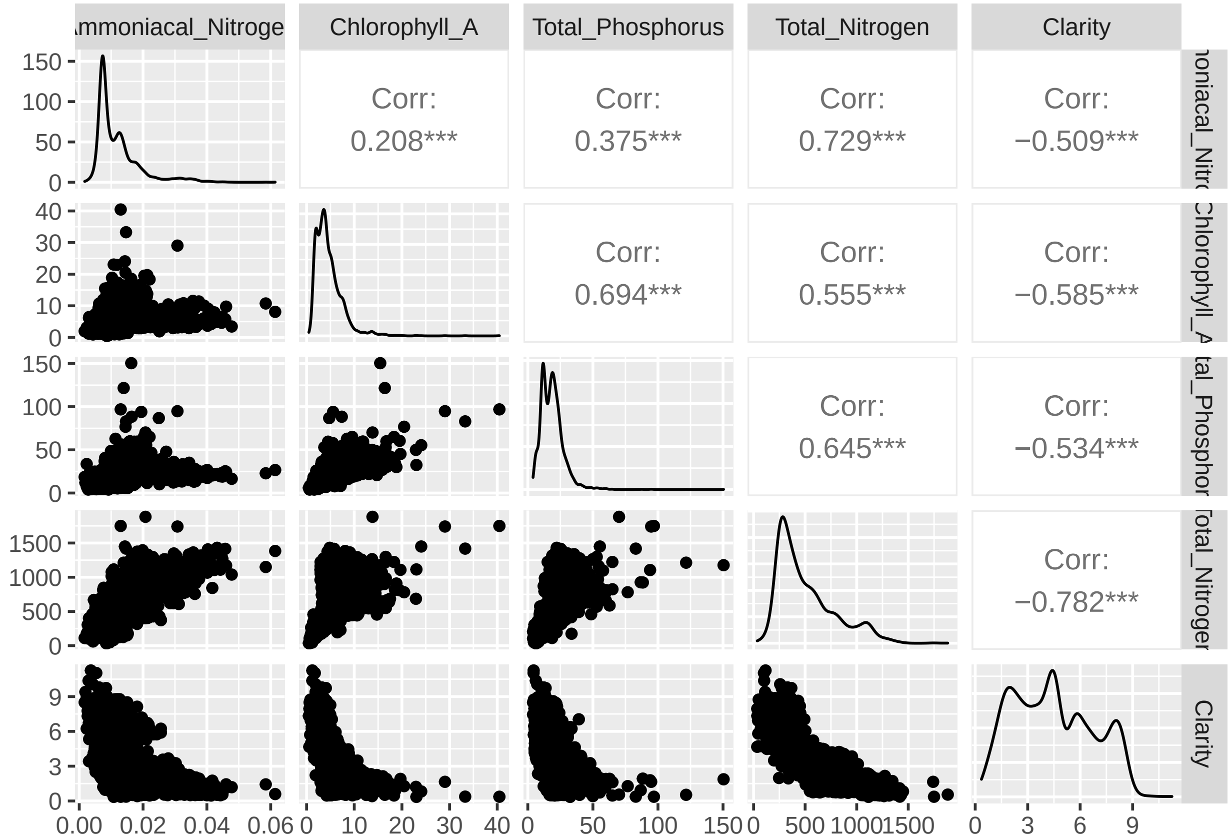
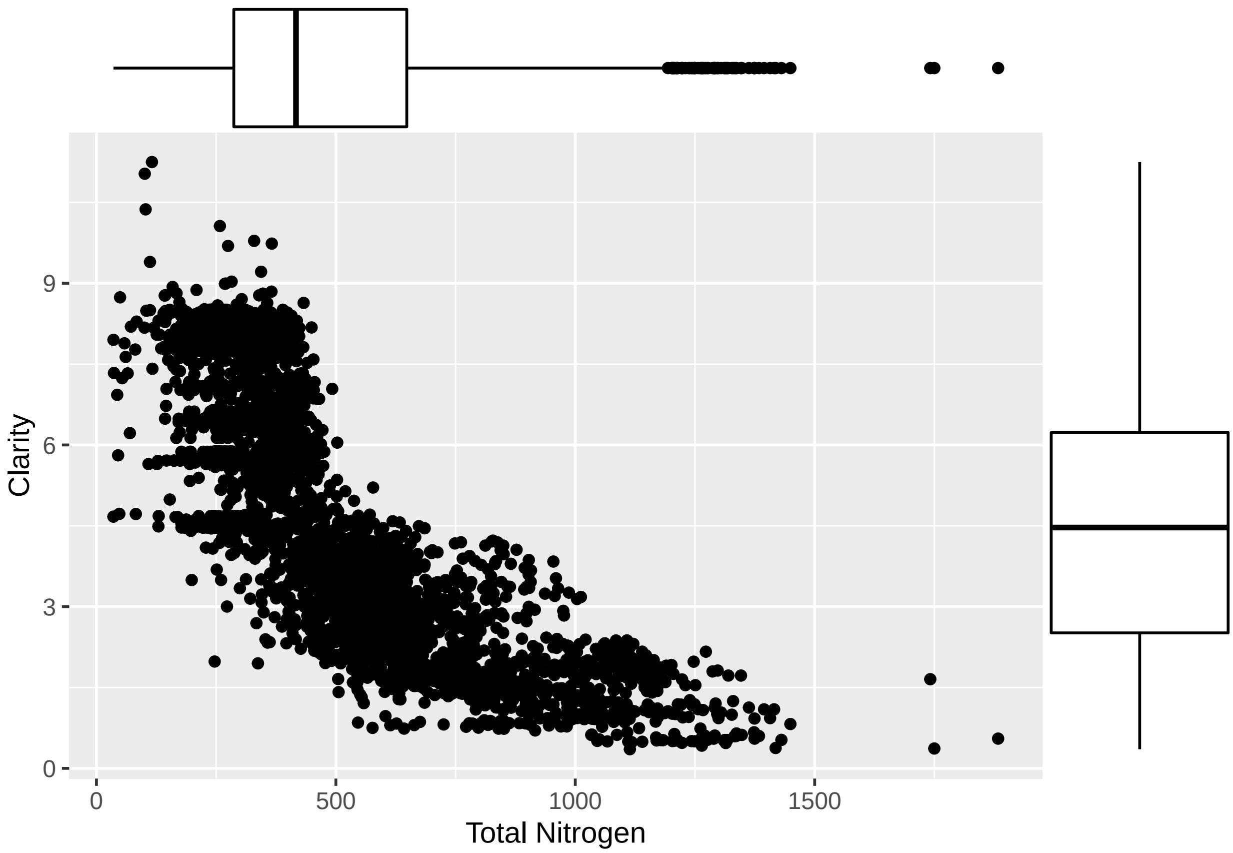


Figure 2: Visualisation of the Correlation Matrix



Pairs plot showed almost normal distributions but with tails long tails making the distributions right skewed.

We then produced scatterplots for each of the correlations. Below is an example of 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.

We then made Cullen and Frey Graphs to examine how each Lake Health variable is distributed. An example with clarity is shown below.

```
## 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
```

The Cullen and Frey graph of Clarity is shown in figure 3. 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.

1.2.2 2.2 Summary of Dominant Landcover investigation

A table of frequencies of the lakes' dominant land cover can be seen in table 2.

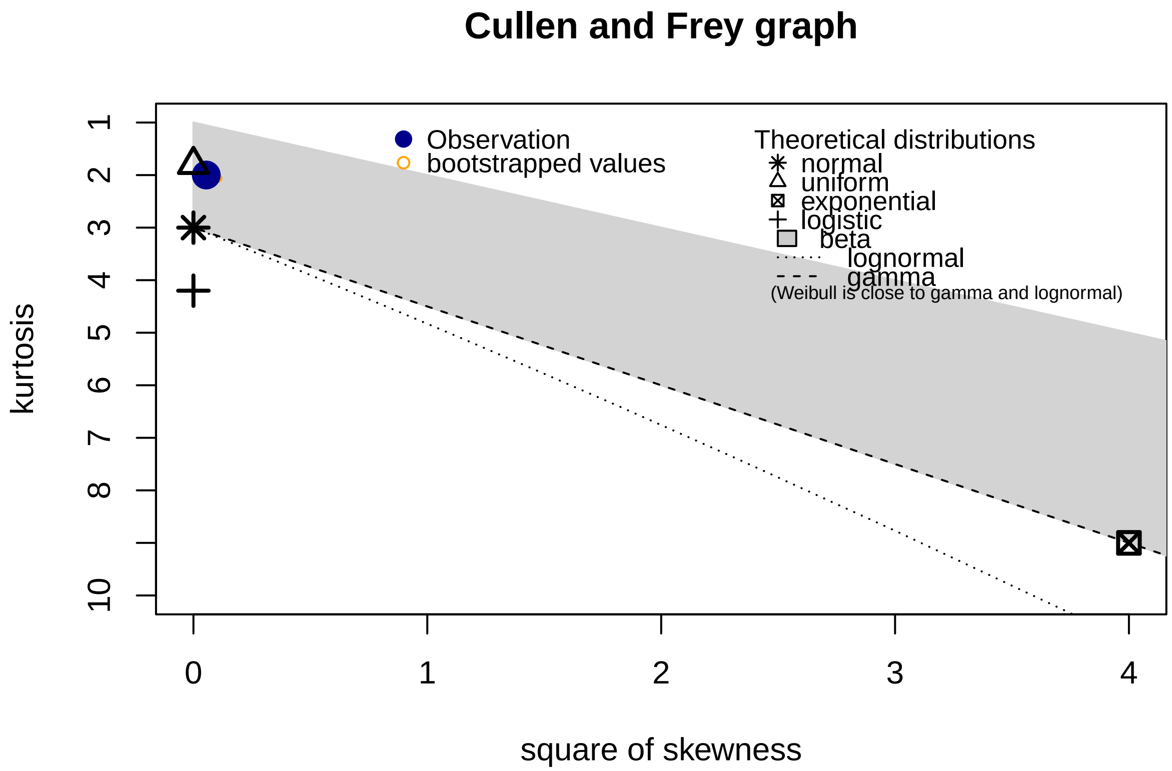


Figure 3: Cullen and Frey Graph of Clarity

Table 2: Table of Frequencies of Dominant Landcover

Landcover	Frequency
Exotic forest	165
Native	1770
Other	12
Pastoral	1770
Urban area	85

We wanted to examine the relationship between our health variables and dominant land cover so we began by producing side-by-side boxplots to display this as seen in figure 4.

```
## 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.
```

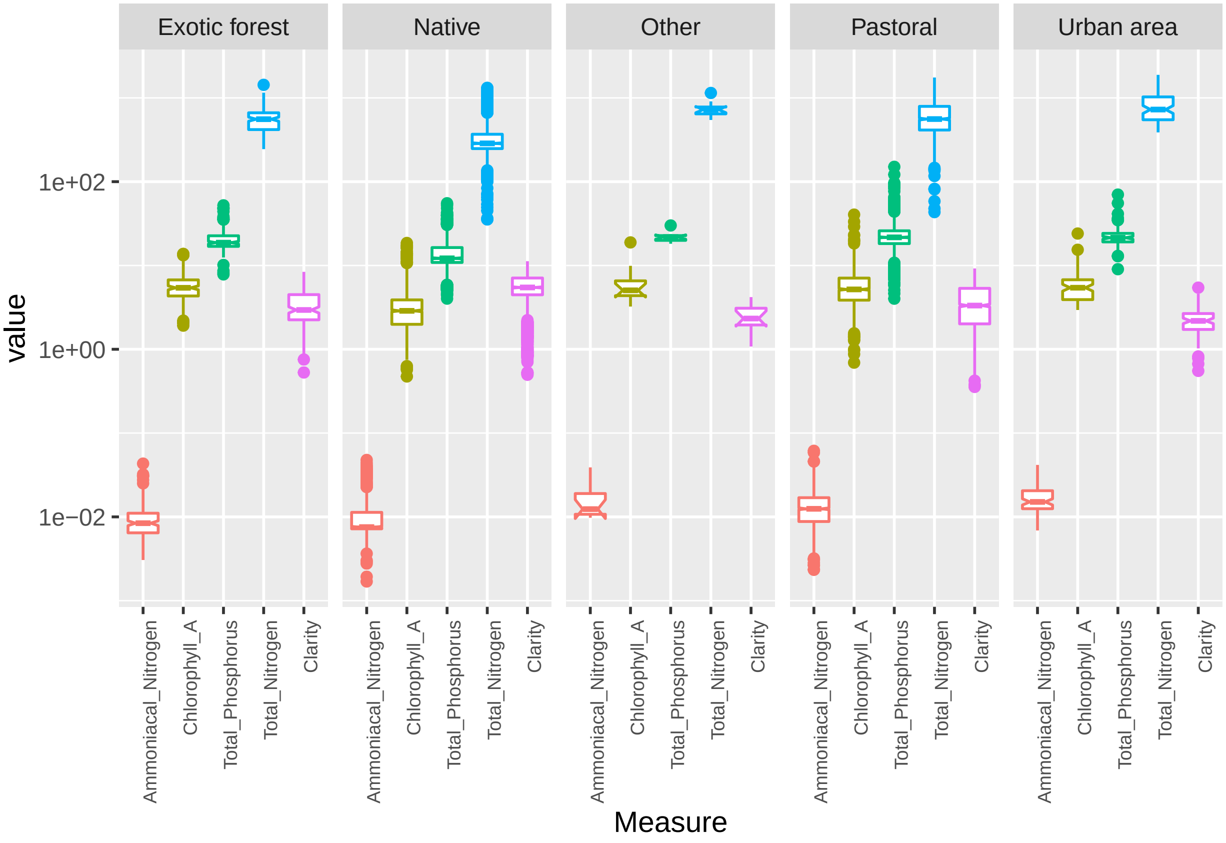


Figure 4: Box Plots of Ammoniacal Nitrogen, message=FALSE, warning=FALSE, Chlorophyll-A, Total Phosphorus, Total Nitrogen and Clarity by Landcover

We then examined each of the lake health variables and their relationship with dominant landcover more closely by looking at zoomed in boxplots. As an example, figure 5 is the boxplot of Landcover and Ammoniacal Nitrogen.

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

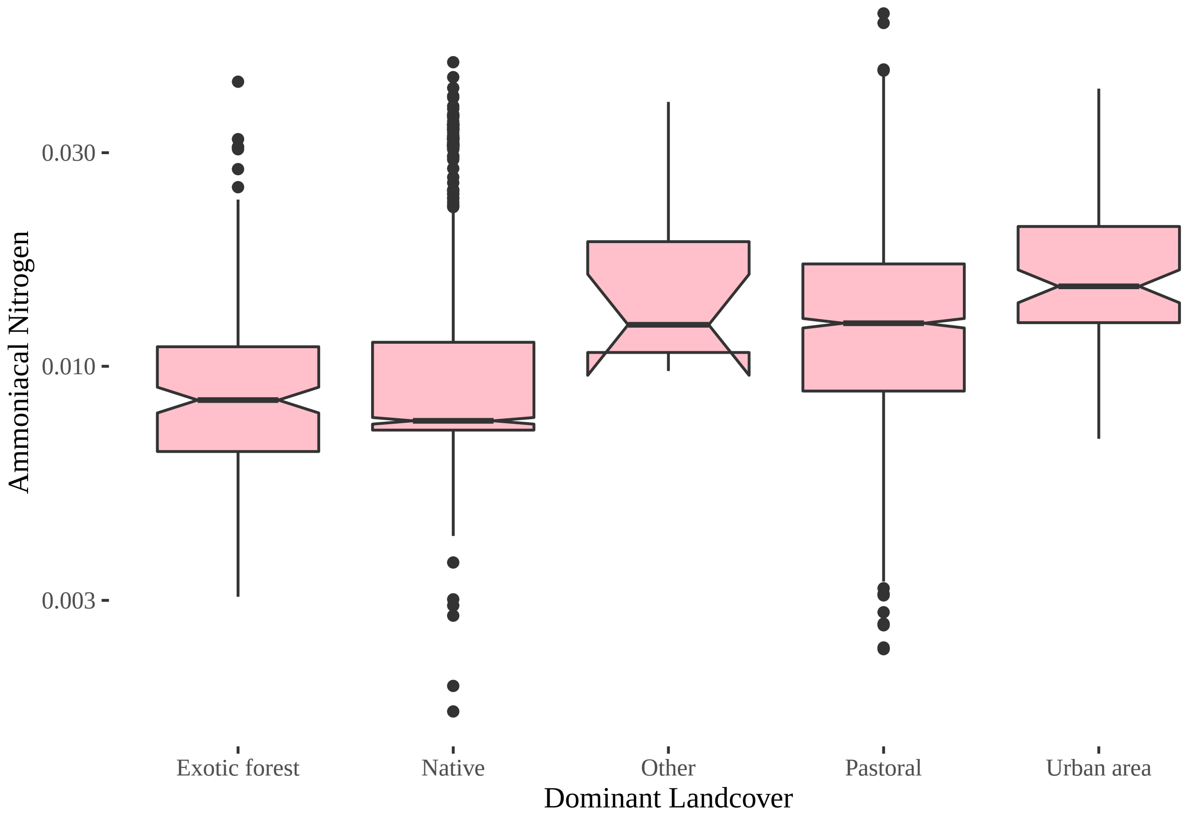


Figure 5: Box Plot of Landcover and Ammoniacal Nitrogen

Figure 5 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 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.

We also produced summary statistics of the Lake Health variables for each of the land covers. The summary statistics for Exotic forest is displayed below:

Table 3: 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

	Ammoniacal Nitrogen	Chlorophyll-A	Total Phosphorus	Total Nitrogen	Clarity
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

1.2.3 2.3 Summary of Depth, Area and Perimeter Investigation.

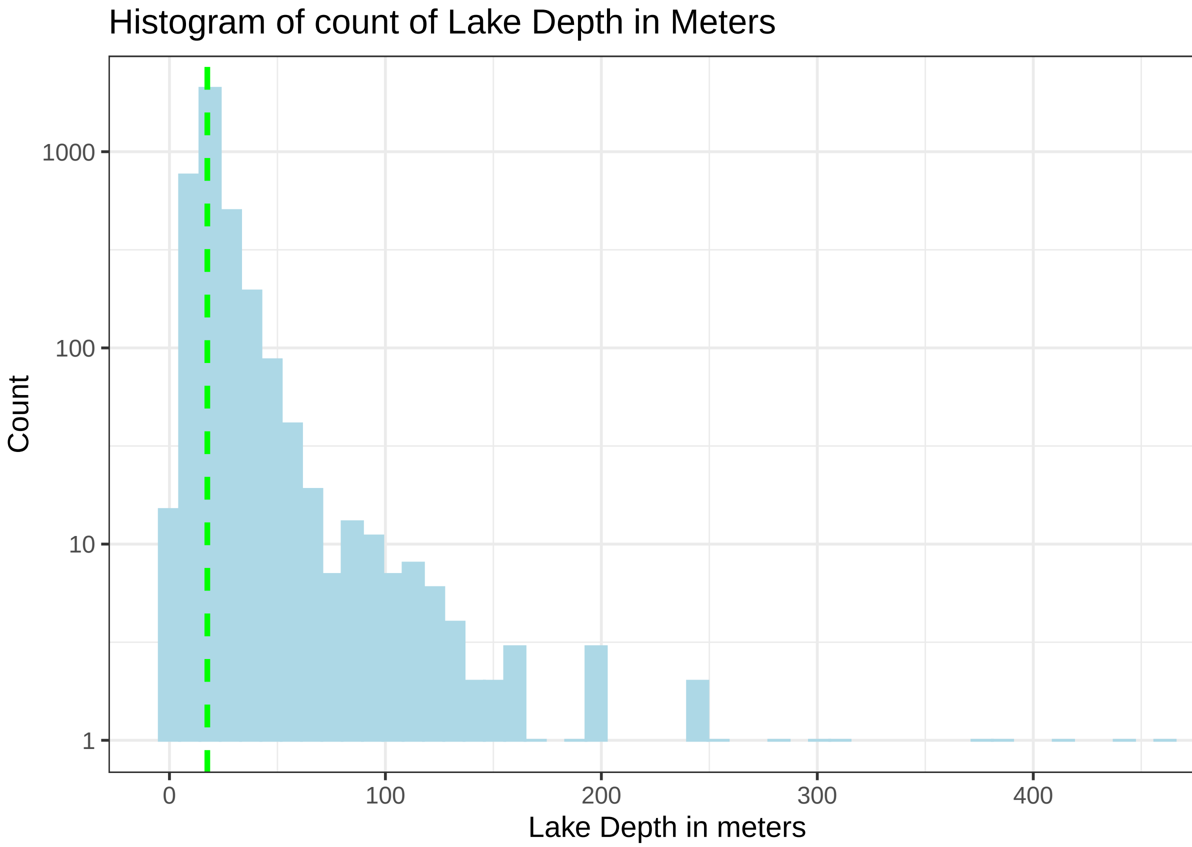
We next wanted to gather information on Lake Depth, Area and Perimeter.

We first produced summary statistics as seen in table 4.

Table 4: 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
Inter-quartile Range	4.887282e+04	829.9800	8.700000
Standard Deviation	1.439392e+07	11246.8807	23.477025
Mean	9.822069e+05	2389.3750	22.910350
Median Absolute Deviation	1.841853e+04	447.8786	5.411490
Kurtosis	1.023860e+03	415.0156	138.442310
Skewness	2.877753e+01	17.3889	9.785791

We produced histograms of the lake dimension variables. The histogram for lake depth is shown in figure



@ref{fig:histdepth}.
The histogram above suggests non-normal data.

All of the histograms for the lake dimensions produced skewed histograms. Because of the skewed data we examined the normality by by looking at skew and kurtosis.
Skewness in our Area (mˆ2), Perimeter(m) and Depth(m)

Lake Perimeter skew: 17.40
Lake Area skew: 28.78
Lake Depth skew: 9.82

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.

We found from the skew that there all of lake dimension variables had strong skews.

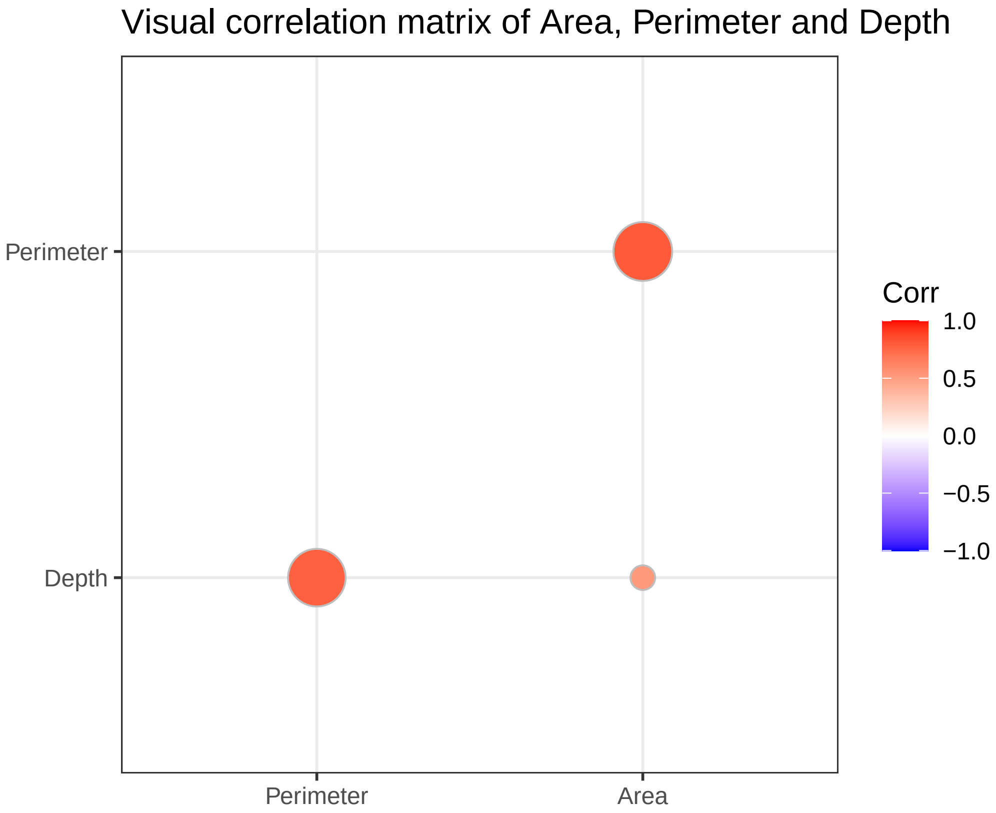
To further examine the normality, we examined kurtosis.

Lake Perimeter kurtosis: 415.36
Lake Area kurtosis: 1023.66
Lake Depth kurtosis: 139.11

All of the skews being much greater than 0 informs us that lake dimension variables are not normally distributed

We then produced calculated the correlation and covariance of the lake dimension variables.

A visualisation of the correlation between Area, Perimeter and Depth is shown in figure @ref{fig:viscorr1}



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.

1.2.4 2.4 Investigating Region

We next wanted to examine the relationship between Regions and the lake dimension variables, and the lake health variables.

We first produced a frequency table for the number of lakes in each region, shown in table 5

Table 5: Frequency table of number of lakes in each Region

region1
Auckland
Bay of Plenty
Canterbury
Gisborne
Hawke's Bay
ManawatÅ«-Whanganui
Marlborough district council
Northland
Otago
Southland
Taranaki

region1	
Tasman district council	
Waikato	
Wellington	
West Coast	
*We can see from the frequency	table that Southland has the greatest number of lakes with 991, and Gisborne the fewest

We then looked at the summary statistics of region, shown in table 6.

Table 6: Table o

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	
ManawatÅ«-Whanganui	4.264950	3.119695	0.0110985	
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	
*Examining the above table ther	e are some interesting p	oints:*		
- 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 where
- There is quite a wide range	of levels of clarity bet	ween regions suggesting	that region may h	ave a rel

We then examined the means and medians of the Lake Dimension variables for the Regions. We produced means and medians for each of the variables into a table and also produced braplot. The table, mean barplot, and median barplot for depth are shown in figure 6. View of the Mean and Median for Lake Depth

**Values of NA have been removed.

```
## # A tibble: 15 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 ManawatÅ«-Whanganui           18.7   16.7
##  7 Marlborough district council  21.0   16.1
##  8 Northland                     20.8   17.5
##  9 Otago                         25.8   17.1
## 10 Southland                     24.5   19.0
```

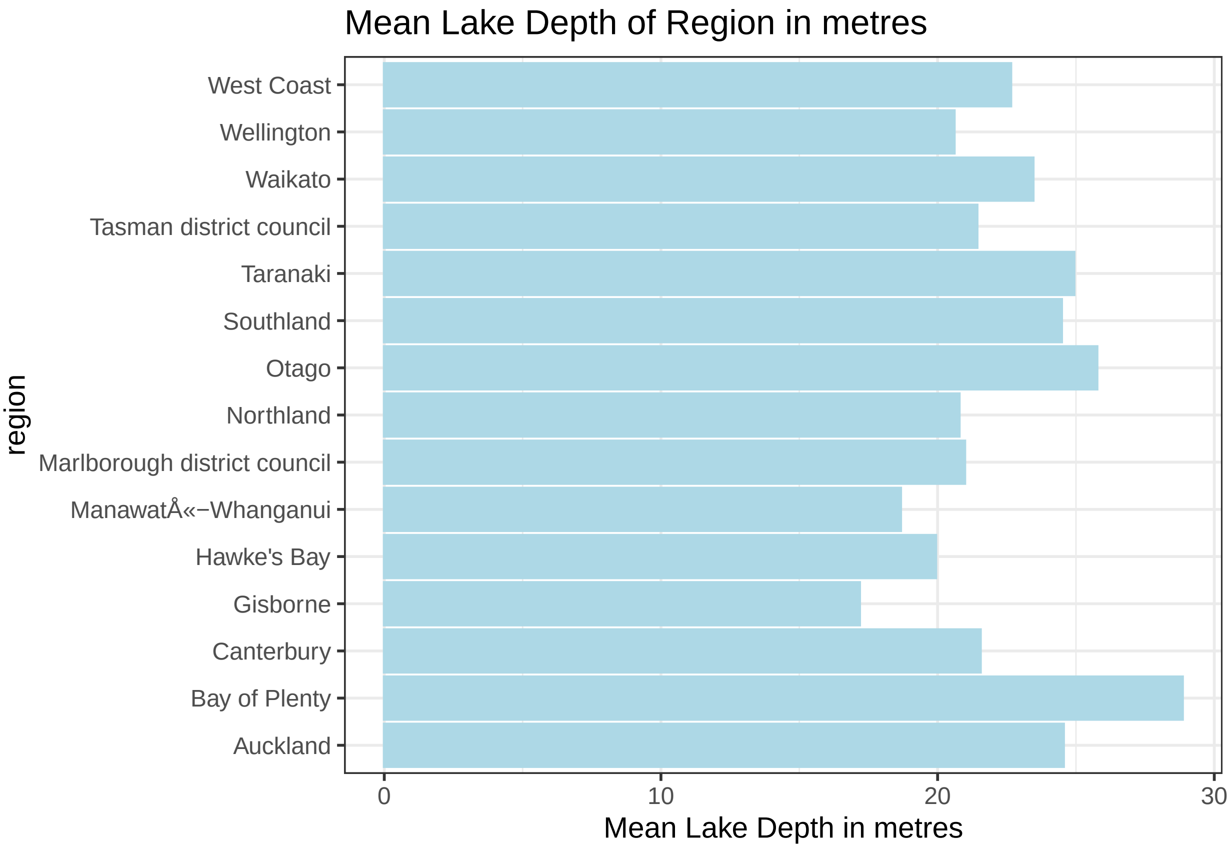


Figure 6: Barplot of regional mean lake Depth

```
## 11 Taranaki                        24.9   16.7
## 12 Tasman district council         21.4   17.7
## 13 Waikato                         23.5   17.3
## 14 Wellington                      20.6   17.6
## 15 West Coast                      22.6   18.4
```

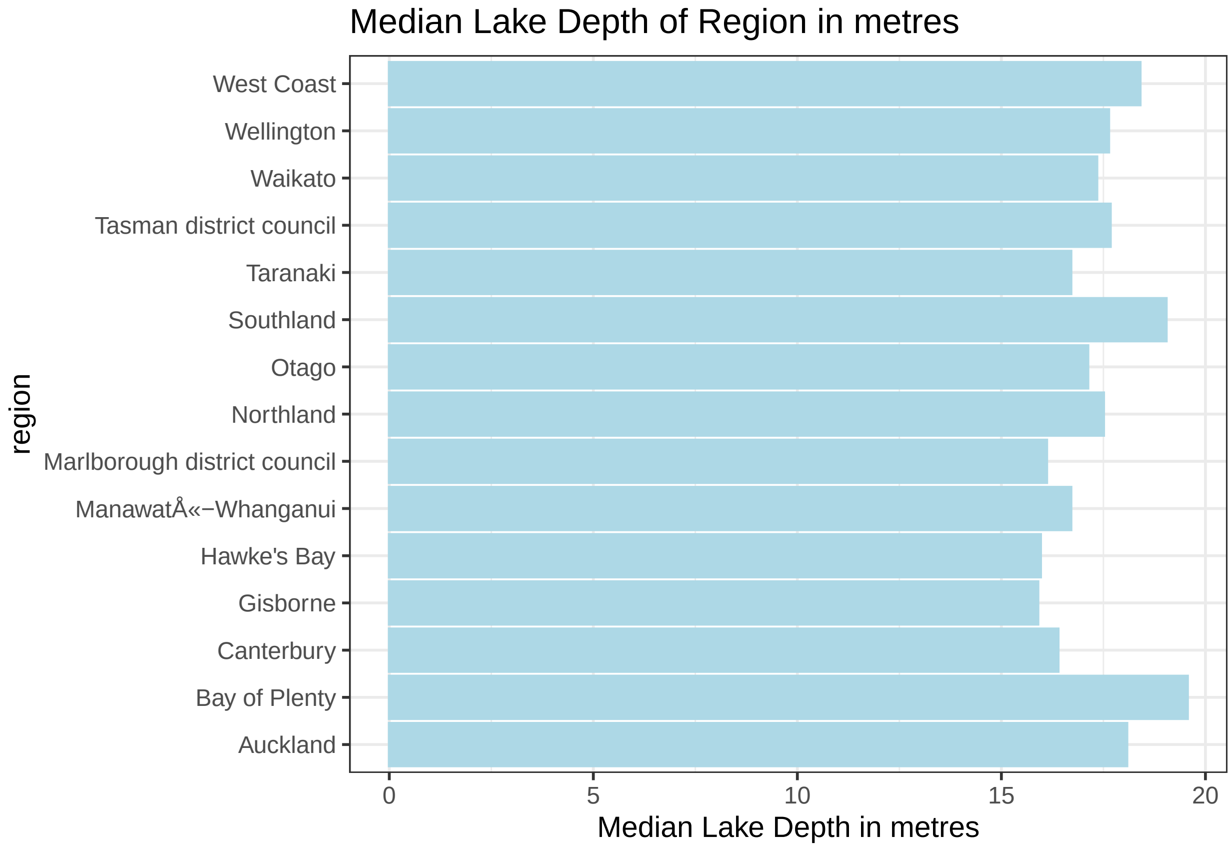


Figure 7: Barplot of regional median lake depths

**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.

1.3 3. The Main Takeaways from the EDA

Below is a summary of the main takeaways from the EDA.

1.3.1 3.1 Lake Health Variables Takeaways

Correlations:
Chlorophyll-A, Total Nitrogen, Ammoniacal Nitrogen and Total Phosphorus are all positively correlated with one another.

The strongest of these correlations are the relationships between Ammoniacal Nitrogen and Total Nitrogen, Total Phosphorus and Chlorophyll-A, and Total Phosphorus and Total Nitrogen. Clarity has a negative relationship with all of the other lake health variables with all correlations between moderate and strong in strength. The strongest of these correlations is between Clarity and Total Nitrogen.
These consistent correlations suggest that the Lake Health variables may be good predictors of one another.

Dominant Landcover The distribution of Lake Health variables differs between landcover categories. We found that lakes with predominantly native landcover tended to have lower levels of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen, and were clearer than lakes with other dominant landcovers. Pastoral and Urban landcovers tended to have higher median levels of Ammoniacal Nitrogen, Chlorophyll-A, Total Phosphorus and Total Nitrogen and tended to be less clear than the other landcovers.

1.3.2 3.2 Lake Dimension takeaways

- The Lake Dimensions are do not appear to be normally distributed
- The median lake areas are quite different and the means are drastically different.
- The median lake perimeters are quite different and the means are drastically different.
- The median lake depths are about the same but the means are quite different.

The considerable difference of lake dimensions in means and medians that some of the regions have suugests that some lakes have considerable influence on the data. This could examine this further with Cooke's Distance. As the Lake Dimensions do not appear to be normally distributed we could also attempt to see what model does fit them.

1.4 4. What we would like to explore further and why

After going through the results of our EDA we have identified several questions that we would like to further explore. These questions are outlined below.

1.4.1 4.1 Are there any particular regions that have poor lake health?

As we found that each of the lake health variables had considerable outliers, we are interested in examining whether these outliers are spread throughout the country or isolated to a particular region.
Knowing this information could increase the efficiency and assist in the planning of efforts to increase lake health.

1.4.2 4.2 Do the lake health variables predict one another?

Most of the correlations between the lake health variables were at least of moderate strength. This suggests that these variables may be able to predict each other. If we know these associations we may be able to reduce the costs of further field research by assuming presence. What we are most interested in seeing if it can be predicted is clarity as it is a measure that any lake observer could get an idea of by looking at any lake.

1.4.3 4.3 How can we model the Lake Dimension variables?

We have already gathered evidence to suggest that the lake dimension variables do not follow a normal distribution. We can further examine what distribution they do follow.

1.4.4 4.4 Do any types of dominant landcover have poorer lake health than others?

We have seen an association between landcover and the lake health variables, it would be interesting to test for difference in means between each type of landcover and each lake health variable.

1.5 5. Further exploration

1.5.1 5.1 Unhealthiest lakes by Region examination

What we want to examine from this is 1. which region has the greatest number of lakes with poor lake health and 2. Which region had the highest proportion of lakes with poor lake health. To do this we sorted we sorted the values for each of the lake health variables into descending order (with the exception of clarity (SECHHI) which is in ascending order). We then took the 100 highest values and, out of the 100 lakes, investigated which region had the highest number of lakes that were high in that variable and then calculated the proportion.

Poorer lake health is measured by having higher values on the lake health dimensions with the exception of clarity where lower values indicate poorer lake health.

We did not examine which Regions had the best lake health as this is harder to measure as it is not clear what a good measure of health is on these metrics. What we know is that high values are bad but low values are not necessarliy good, thus we should not examine the minimum lake health variables.

Frequency table, by region, of the 100 lakes with the highest values of Total Phosphorus

```
##                          Region No.Of.Lakes.Total No.High.TP.Lakes Proportion
## 1                      Auckland                74                0 0.00000000
## 2                 Bay of Plenty                98                3 0.03061224
## 3                    Canterbury               463               20 0.04319654
## 4                      Gisborne                30                3 0.10000000
## 5                    Hawke’s Bay               285                4 0.01403509
## 6            ManawatÅ«-Whanganui               226               19 0.08407080
## 7  Marlborough district council                50                1 0.02000000
## 8                     Northland               262                0 0.00000000
## 9                         Otago               378               13 0.03439153
## 10                    Southland               991               11 0.01109990
## 11                     Taranaki                90                2 0.02222222
## 12      Tasman district council                57                1 0.01754386
## 13                      Waikato               233               11 0.04721030
## 14                   Wellington               107                6 0.05607477
## 15                   West Coast               457                5 0.01094092
```

The investigation of 100 lakes containing the highest values of total phosphorus revealed that:
- The Canterbury region contained 20 of the 100 lakes with the highest total phosphorus.
- This was followed by the Manawatu-Whanagnui region with 19 of the 100 lakes.
- The third highest was the Otago region with 13 of the 100 lakes.

- of the 100 lakes with the highest total phosphorus, the region with the largest proportion of its lakes within the top 100 was Gisborne with 10%.
- This was followed by Manawatu-Whanagnui with 8.4%.
- The third highest was Wellington with 5.6%.

Frequency table, by region, of the 100 lakes with the highest values of Total Nitrogen

```
##                            Region No.Of.Lakes.Total No.High.TN.Lakes  Proportion
## 1                    Bay of Plenty                98                0 0.000000000
## 2                       Canterbury               463               12 0.025917927
## 3                         Gisborne                30                1 0.033333333
## 4                      Hawke’s Bay               285                1 0.003508772
## 5               ManawatÅ«-Whanganui               226               23 0.101769912
## 6  Marlborough district council                50                0 0.000000000
## 7                        Northland               262                0 0.000000000
## 8                            Otago               378                1 0.002645503
## 9                        Southland               991                2 0.002018163
## 10                        Taranaki                90                0 0.000000000
## 11          Tasman district council                57                0 0.000000000
## 12                         Waikato               233               54 0.231759657
## 13                      Wellington               107                0 0.000000000
## 14                      West Coast               457                3 0.006564551
```

The investigation of 100 lakes containing the highest values of total nitrogen revealed that:
- The Waikato region contained 54 of the 100 lakes with the highest total nitrogen.
- This was followed by the Manawatu-Whanagnui region with 23 of the 100 lakes.
- The third highest was the Canterbury region with 12 of the 100 lakes.

- Of the 100 lakes with the highest total nitrogen, the region with the largest proportion of its lakes within the top 100 was Waikato with 23.175%.
- This was followed by Manawatu-Whanagnui with 10.2%.
- The third highest was Canterbury with 2.6%.

Frequency table, by region, of the 100 lakes with the highest values of Ammoniacal Nitrogen

```
##                            Region No.Of.Lakes.Total No.High.NH4N.Lakes
## 1                         Auckland                74                  0
## 2                    Bay of Plenty                98                  5
## 3                       Canterbury               463                  1
## 4                         Gisborne                30                  3
## 5                      Hawke’s Bay               285                  5
## 6               ManawatÅ«-Whanganui               226                 22
## 7  Marlborough district council                50                  3
## 8                        Northland               262                  1
## 9                            Otago               378                 10
## 10                       Southland               991                  3
## 11                        Taranaki                90                  0
## 12         Tasman district council                57                  0
## 13                         Waikato               233                 32
## 14                      Wellington               107                  5
## 15                      West Coast               457                  7
##     Proportion
## 1  0.000000000
## 2  0.051020408
## 3  0.002159827
## 4  0.100000000
## 5  0.017543860
## 6  0.097345133
## 7  0.060000000
## 8  0.003816794
## 9  0.026455026
## 10 0.003027245
## 11 0.000000000
## 12 0.000000000
## 13 0.137339056
## 14 0.046728972
## 15 0.015317287
```

The investigation of 100 lakes containing the highest values of Ammoniacal Nitrogen revealed that:
- The Waikato region contained 32 of the 100 lakes with the highest levels of Ammoniacal Nitrogen
- This was followed by the Manawatū-Whanganui region with 22 of the 100 lakes.
- The third highest was the Otago region with 10 of the 100 lakes.

- Of the 100 lakes with the highest Ammoniacal Nitrogen, the region with the largest proportion of its lakes within the top 100 was Waikato with 13.7%.
- This was followed by Manawatū-Whanganui and Gisbone both with approximately 10%.

Frequency table, by region, of the 100 lakes with the highest values of Chlorophyll-A

```
##                            Region No.Of.Lakes.Total No.High.NH4N.Lakes
## 1                        Auckland                74                  1
## 2                    Bay of Plenty               98                  9
## 3                      Canterbury               463                  8
## 4                        Gisborne                30                  4
## 5                      Hawke's Bay              285                 14
## 6              ManawatÅ«-Whanganui              226                  4
## 7  Marlborough district council                50                  2
## 8                       Northland               262                  0
## 9                           Otago               378                  3
## 10                      Southland               991                  0
## 11                       Taranaki                90                  2
## 12       Tasman district council                57                  9
## 13                        Waikato               233                 19
## 14                     Wellington               107                 13
## 15                     West Coast               457                 11
##     Proportion
## 1  0.013513514
## 2  0.091836735
## 3  0.017278618
## 4  0.133333333
## 5  0.049122807
## 6  0.017699115
## 7  0.040000000
## 8  0.000000000
## 9  0.007936508
## 10 0.000000000
## 11 0.022222222
## 12 0.157894737
## 13 0.081545064
## 14 0.121495327
## 15 0.024070022
```

The investigation of 100 lakes containing the highest values of Chlorophyll-A revealed that:
- The Waikato region contained 54 of the 100 lakes with the highest levels of Chlorophyll-A.

- This was followed by the Manawatū-Whanganui region with 23 of the 100 lakes.
- The third highest was the Canterbury region with 12 of the 100 lakes.

- Of the 100 lakes with the highest levels of Chlorphyll-A, the region with the largest proportion of its lakes within the top 100 was Waikato with 23.2%.
- This was followed by Manawatū-Whanganui with 10.1%.
- The third highest was Gisborne with 3.3%.

Frequency table, by region, of the 100 lakes with the lowest values of Clarity

```
##                          Region No.Of.Lakes.Total No.Low.Clarity.Lakes
## 1                      Auckland                74                    0
## 2                 Bay of Plenty                98                    0
## 3                    Canterbury               463                    1
## 4                      Gisborne                30                    0
## 5                   Hawke’s Bay               285                    0
## 6           ManawatÅ«-Whanganui               226                    6
## 7  Marlborough district council                50                    0
## 8                     Northland               262                    0
## 9                         Otago               378                    5
## 10                    Southland               991                   30
## 11                     Taranaki                90                    0
## 12      Tasman district council                57                    0
## 13                      Waikato               233                   45
## 14                   Wellington               107                    1
## 15                   West Coast               457                    4
##     Proportion
## 1  0.000000000
## 2  0.000000000
## 3  0.002159827
## 4  0.000000000
## 5  0.000000000
## 6  0.026548673
## 7  0.000000000
## 8  0.000000000
## 9  0.013227513
## 10 0.030272452
## 11 0.000000000
## 12 0.000000000
## 13 0.193133047
## 14 0.009345794
## 15 0.008752735
```

The investigation of 100 lakes containing the lowest values of clarity revealed that:
- The Waikato region contained 45 of the 100 lakes with the lowest clarity
- This was followed by the Southland region with 30 of the 100 lakes.
- The third highest was the Manawatū-Whanganui region with 6 of the 100 lakes.

- Of the 100 lakes with the lowest clarity, the region with the largest proportion of its lakes within the top 100 was Waikato with 19.3%.
- This was followed by Southland with 3.0%.
- The third highest was Manawatū-Whanganui with 2.7%. **Summary of unhealthiest lakes analysis** This analysis gives evidence to say that the Waikato Region has the worst overall lake health, followed

by Manawatū-Whanganui (when examining lakes with the worst lake health). This is because both of these regions have many occurences having the highest proportions of their lakes within the top 100 'worst' scoring lakes.
Further analysis should examine these findings alongside the overall means and medians of the regions on the lake health variables. Further Analysis could also look at ranking regions based on their overall lake health scores.

1.5.2 5.2 Do the lake health variables predict one another?

Please note that this section is not complete, did not have time to examine whether all of the assumptions were met or to do AIC and BIC We learned from the EDA that the Lake Health variables are almost all correlated. Because of this, we guess that we will be able to make predictions using regression from these variables. What we most want to explore is if the other lake health variables are good predictors of clarity. The reason why we wish to explore this is that if the other lake variables are good predictors of clarity, then by knowing clarity, the average person could makes some assumptions about the other lake health variables.

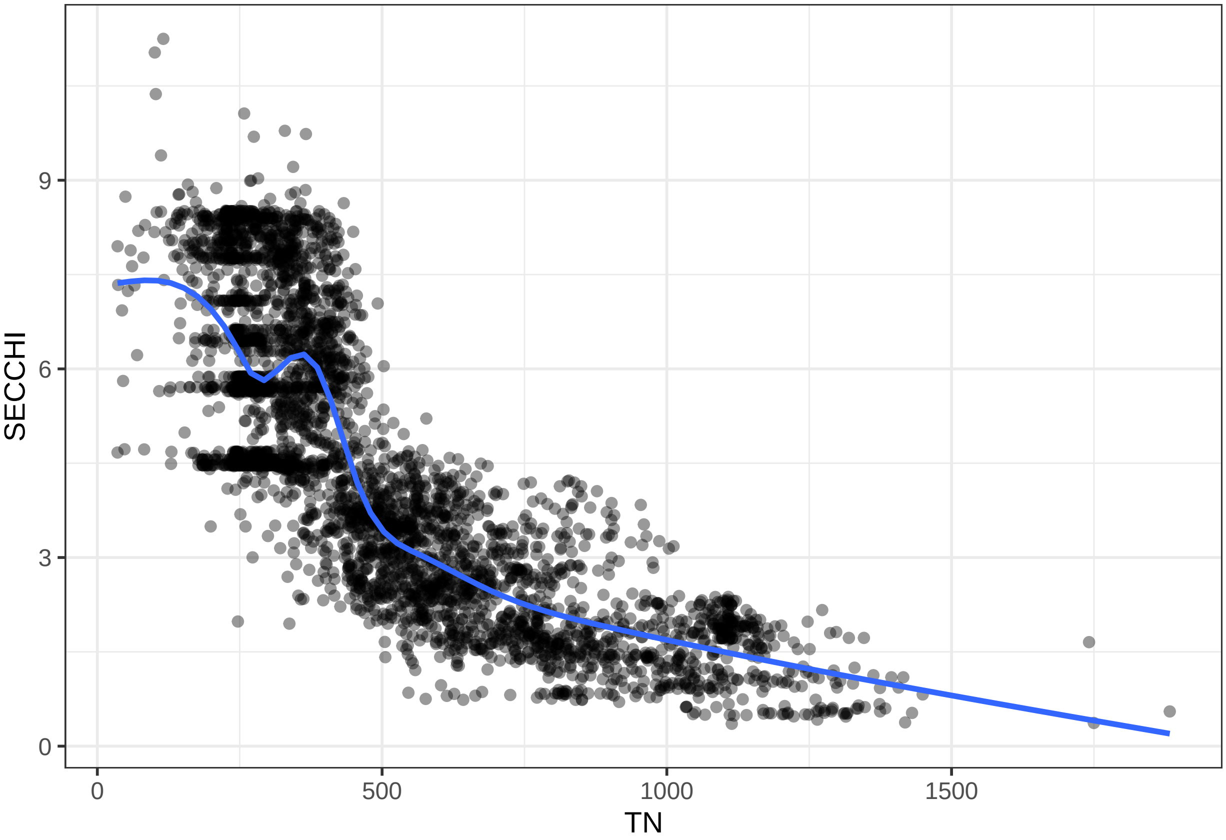
Before we can produce a multiple linear regression model, we must first check that the assumptions of multiple linear regression are met. These assumptions are:

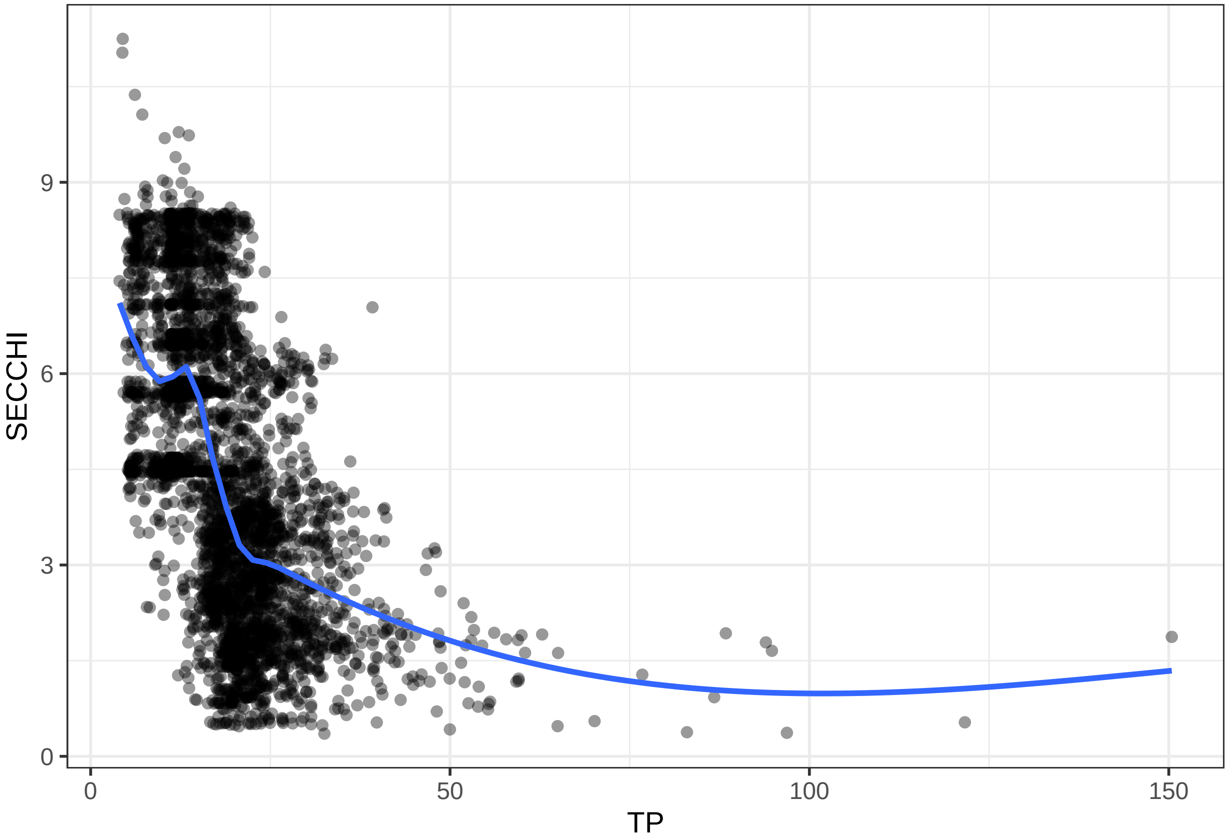
- That linear relationships between the outcome variable and the independent variables exist.
- Multivariate Normality such that the residuals are normally distributed.
- No Multicollinearity such that the independent variables are not highly correlated with each other.
- Homoscedasticity.

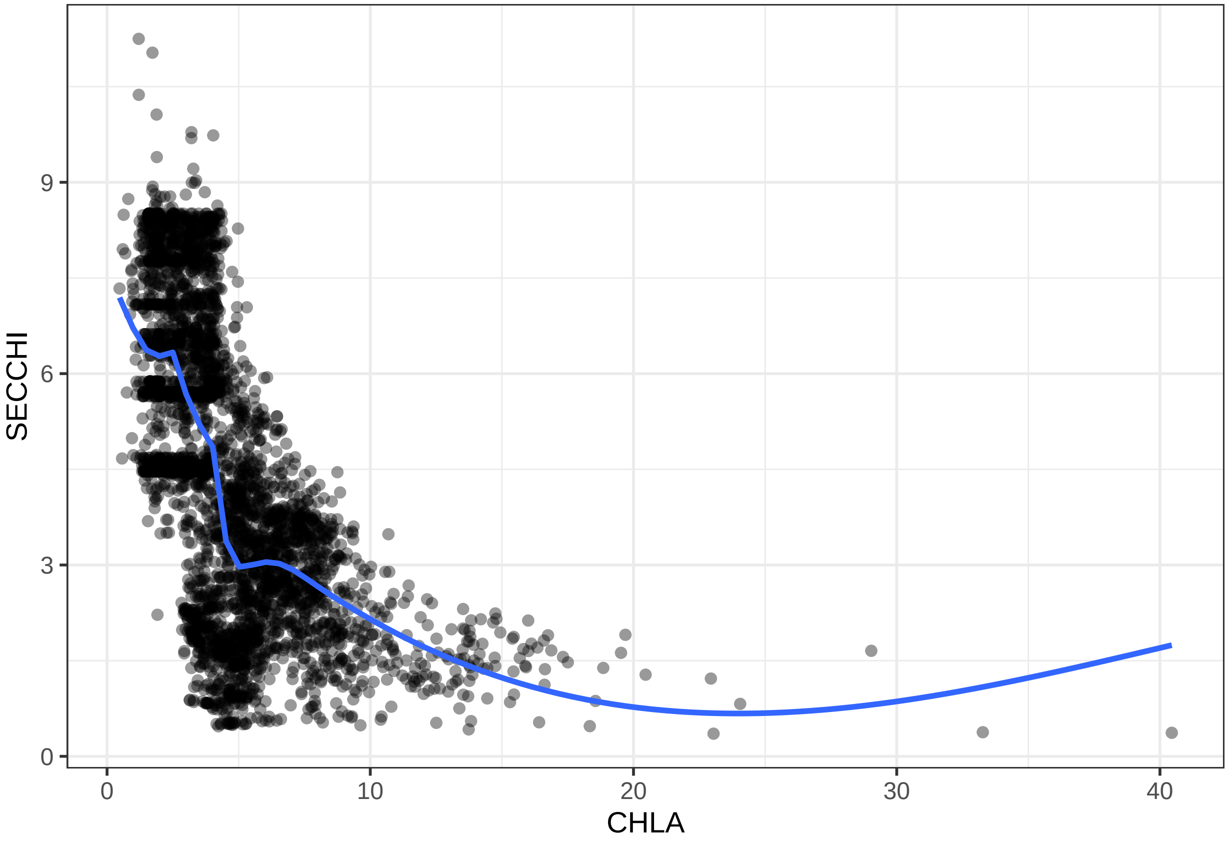
  To investigate whether the assumption of linear relationships have been met we have produced scatterplots.

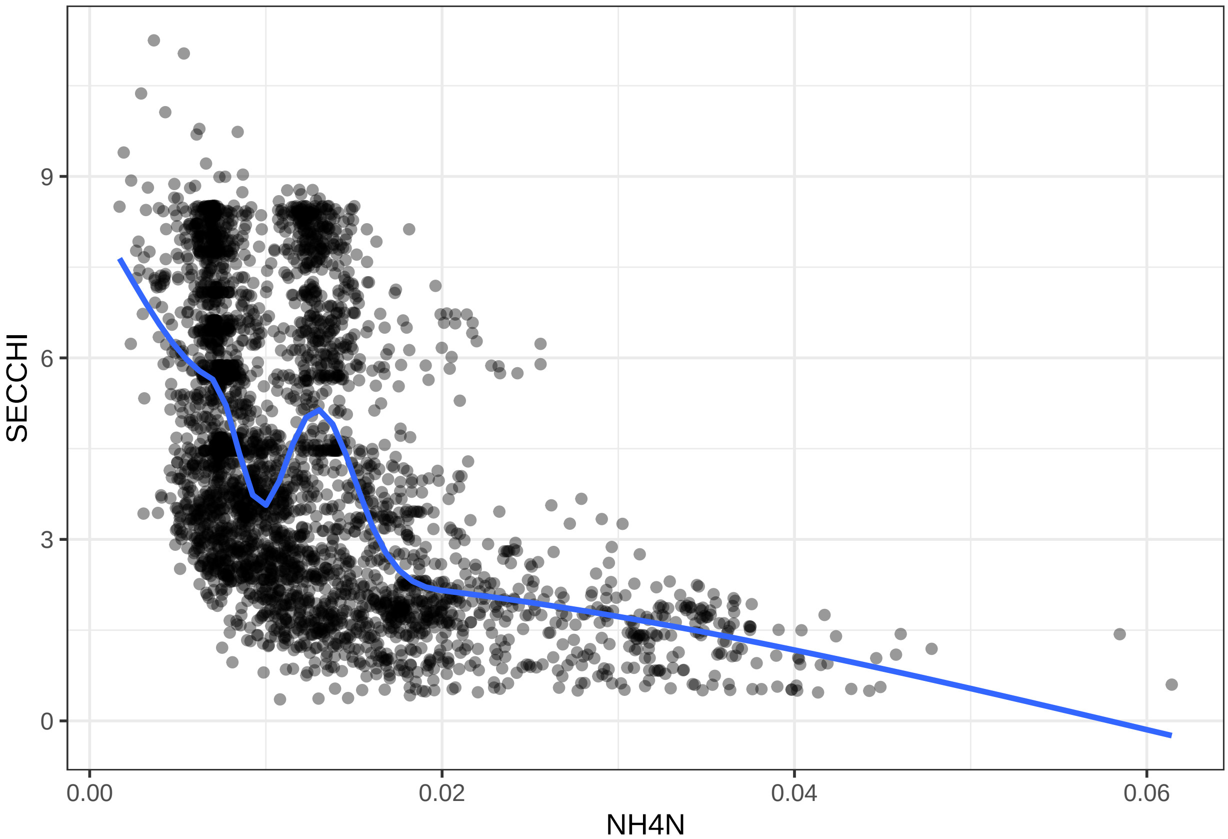
To investigate whether the assumption of multivariate normality has been met we have conducted a Mardia's Test.

To investigate whether the assumption of multicollinearity has been met we examine the a the correlation matrix


9
6
3
0
SECCHI
0
500
1000
1500
TN


SECCHI
0
3
6
9
TP
0
50
100
150


SECCHI
9
6
3
0
0
10
20
30
40
CHLA



Where TN is Total Nitrogen, CHLA is Chlorphyll-A, NH4N is ammoniacal NItrogen, TP is Total Phosphorus and SECCHI is clarity
The Scatterplots seem to show a non-linear relationship suggesting that the assumption of linearity is not met.

```
##         NH4N  CHLA    TP    TN SECCHI
## NH4N    1.00  0.21  0.37  0.73  -0.51
## CHLA    0.21  1.00  0.69  0.56  -0.58
## TP      0.37  0.69  1.00  0.65  -0.53
## TN      0.73  0.56  0.65  1.00  -0.78
## SECCHI -0.51 -0.58 -0.53 -0.78   1.00
```

As none of the correlations are greater than or equal to 0.9 or less than or equal to 0.9 we assume that the assumption of non-multicollinearity is met.

Multi-linear Regression model We mad an multiple linear regression model where the explanatory variables are Ammoniacal Nitrogen (NH4N), Total Phosphorus (TP), Total Nitrogen (TN), and Chlorophyll-A (CHLA) and the response variable is Clarity (SECCHI).

```
##
## Call:
## lm(formula = SECCHI ~ TN + CHLA + NH4N + TP, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3472 -1.0176 -0.1302  0.8563  8.7120
##
```

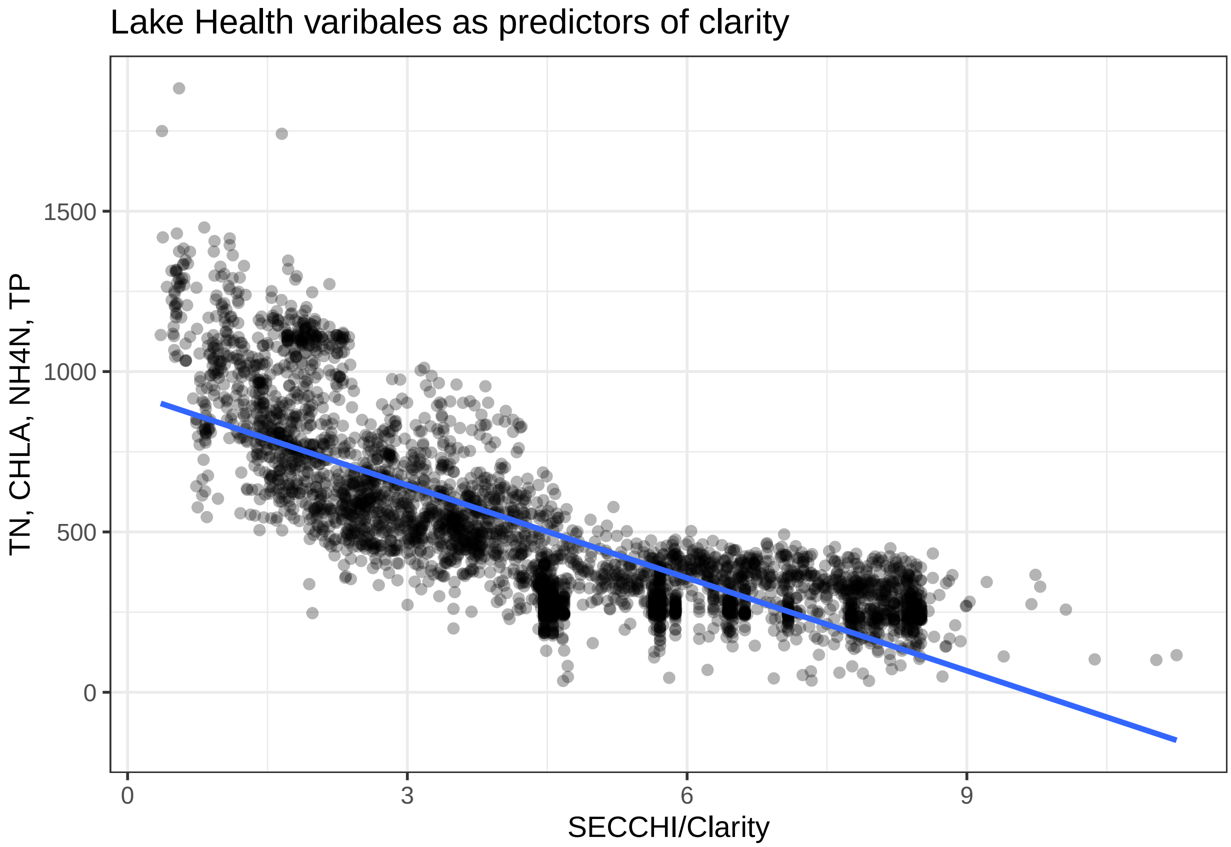
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.7763882  0.0553226 140.564  < 2e-16 ***
## TN          -0.0060112  0.0001497 -40.143  < 2e-16 ***
## CHLA        -0.2052203  0.0114449 -17.931  < 2e-16 ***
## NH4N        14.7780476  4.9356884   2.994  0.00277 **
## TP           0.0258903  0.0036498   7.094 1.55e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.335 on 3797 degrees of freedom
## Multiple R-squared:  0.6502, Adjusted R-squared:  0.6498
## F-statistic:  1764 on 4 and 3797 DF,  p-value: < 2.2e-16
```

At an alpha of 0.5, all of the explanatory variables are significant. At an alpha of 0.1, all of the explanatory variables except for Total Phosphorus (TP) are significant. We see from the model summary that our lake health variables (minus clarity) explain about 62% of the variance in clarity (SECCHI).

```
multReg1 <- ggplot(df1,aes(x=SECCHI,y= TN, CHLA, NH4N, TP))+
         geom_point(alpha = 0.3)+
 geom_smooth(method=lm,se=FALSE,fullrange=TRUE)+
  theme_bw()+
  labs( title = "Lake Health varibales as predictors of clarity", y = "TN, CHLA, NH4N, TP", x = "SECCHI,



# Plotting a single Regression Line
multReg1
```



The above graph for predicting clarity suggests that with the explanatory variables that we have included, they are not a good fit for a multi-linear regression model for predicting clarity. What we can see from the graph though is that higher values of the lake health variables predict reduced clarity; we cannot say necessarily though that lower values of the lake health variables predict high clarity.

After examining the evidence, we have decided that when predicting clarity, the other lake health variables should not follow a multi-linear relationship. Further Analysis could look into curvilinear regression. Further analysis should also take into account the depth of the lakes being examined.

1.5.3 5.3 How can we model the Lake Dimension variables?

We have already gathered evidence to suggest that the lake dimension variables do not follow a normal distribution. We can further examine what distribution they do follow.

We have conducted Anderson-Darling tests for normality for lake area, perimeter and depth along with constructing a Cullen and Frey graph to investigate what distribution these variables may follow.

Table 7 shows the test statistic and p-value from the Anderson-Darling test on lake area. The null hypothesis for this test is that the distribution of lake area is Gaussian, and the alternative hypothesis is that the distribution of lake area is not normal. The test statistic for this test is 1381.847, and the p-value is 3.7e-24. As the p-value is very small, we can reject the null hypothesis and conclude the distribution of lake area is not normal. Next we constructed a Cullen and Frey graph with 1000 bootstrapped observations. This is shown in figure 8. We can see the observation of the kurtosis and square of skewness of the observations of lake area, and all bootstrapped observations, lie within the beta region, indicating the distribution of lake area could be beta.

```
## summary statistics
```

Table 7: Anderson-Darling Test Statistic and P-value for Area

	Area
Test Statistic	1381.847
P-Value	3.7e-24

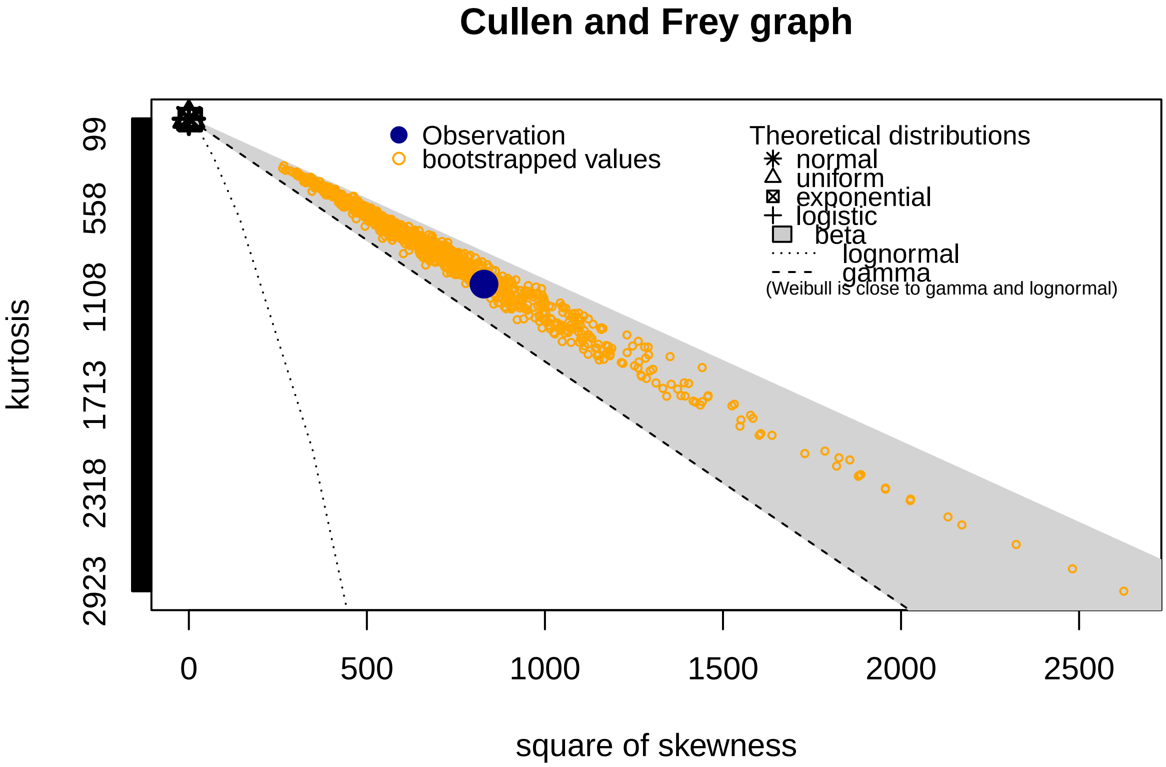


Figure 8: Cullen and Frey Graph of Lake Area

Table 8: Anderson-Darling Test Statistic and P-value for Perimeter

	Perimeter
Test Statistic	1130.8425
P-Value	3.7e-24

```
## ------
## min:  10004.94   max:  6.13e+08
## median:  24165.89
## mean:  982206.9
## estimated sd:  14393921
## estimated skewness:  28.78889
## estimated kurtosis:  1025.206
```

Table 8 shows the test statistic and p-value from the Anderson-Darling test on lake perimeter. The null hypothesis for this test is that the distribution of lake perimeter is Gaussian, and the alternative hypothesis is that the distribution of lake perimeter is not normal. The test statistic for this test is 1130.8425, and the p-value is 3.7e-24. As the p-value is very small, we can reject the null hypothesis and conclude the distribution of lake perimeter is not normal. Next we constructed a Cullen and Frey graph with 1000 bootstrapped observations. This is shown in figure 9. We can see the observation of the kurtosis and square of skewness of the observations of lake perimeter, and almost all bootstrapped observations, lie within the beta region, indicating the distribution of lake perimeter could be beta.

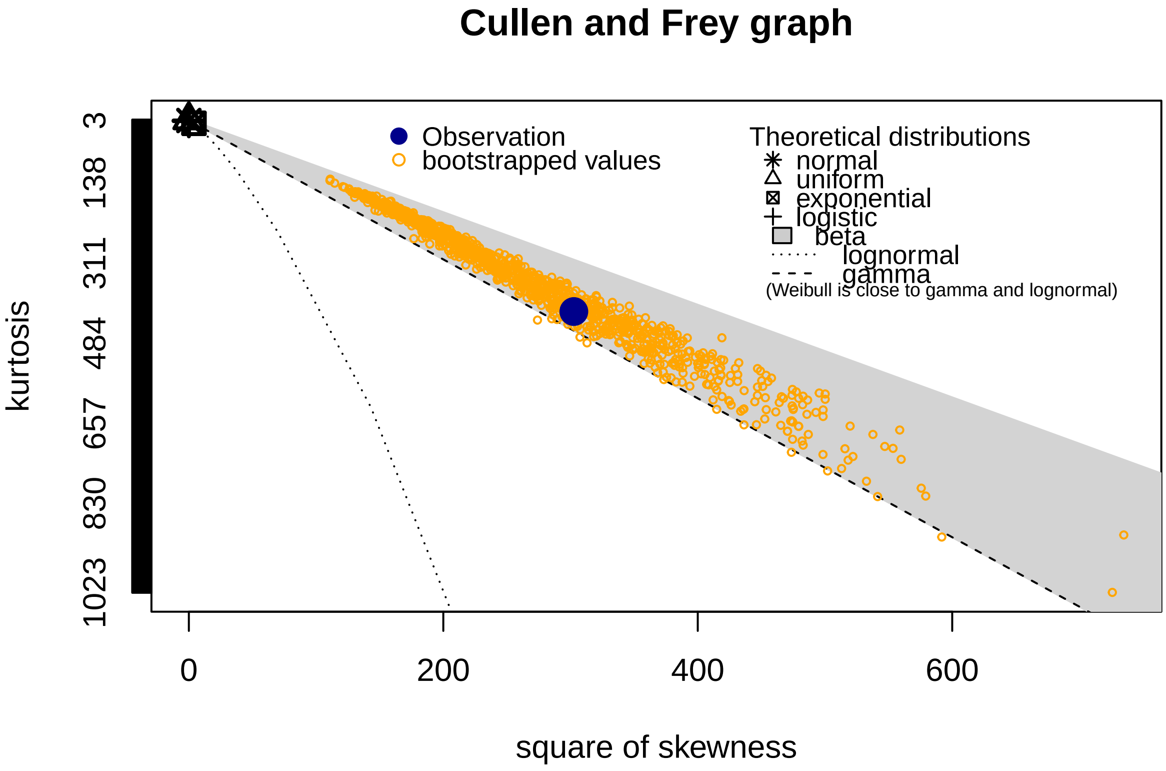


Figure 9: Cullen and Frey Graph of Lake Perimeter

Table 9: Anderson-Darling Test Statistic and P-value for Depth

	Depth
Test Statistic	621.2451
P-Value	3.7e-24

```
## summary statistics
## ------
## min:  371.16   max:  369677.8
## median:  823.935
## mean:  2389.375
## estimated sd:  11246.88
## estimated skewness:  17.39577
## estimated kurtosis:  415.5596
```

Table 9 shows the test statistic and p-value from the Anderson-Darling test on lake depth The null hypothesis for this test is that the distribution of lake depth is Gaussian, and the alternative hypothesis is that the distribution of lake depth is not normal. The test statistic for this test is 621.2451, and the p-value is 3.7e-24. As the p-value is very small, we can reject the null hypothesis and conclude the distribution of lake depth is not normal. Next we constructed a Cullen and Frey graph with 1000 bootstrapped observations. This is shown in figure 10. We can see the observation of the kurtosis and square of skewness of the observations of lake depth, and most of the bootstrapped observations, lie within the beta region, indicating the distribution of lake depth could be beta.

```
## summary statistics
## ------
## min:  1   max:  462
## median:  17.53
## mean:  22.91035
## estimated sd:  23.47703
## estimated skewness:  9.789654
## estimated kurtosis:  138.6222
```

1.5.4 5.4 Do any types of dominant landcover have poorer lake health than others?

We have seen an association between landcover and the lake health variables, it would be interesting to test for difference in means between each type of landcover and each lake health variable. We wanted to do a test for difference in means between each type of landcover for each lake health variable. Originally we wanted to do a one-way ANOVA test, however this test requires normality of each group. We did Anderson-Darling tests for normality of each landcover type in each lake health variable, and when these concluded the data was not normally distributed, we did a

Table 10 shows the test statistic and p-value from the Anderson-Darling test on Ammoniacal Nitrogen levels in lakes with predominantly Exotic Forest landcover. The null hypothesis for this test is that the distribution of Ammoniacal Nitrogen in lakes with exotic forest landcover is Gaussian, and the alternative hypothesis is that the distribution is not normal. The test statistic for this test is 12.2369, and the p-value is 3.7e-24. As the p-value is very small, we can reject the null hypothesis and conclude the distribution of Ammoniacal Nitrogen in lakes with predominantly exotic forest landcover is not normal. Therefore, we cannot use a one-way ANOVA test. We have also done this test on the other landcover types with similar results. For Native landcover, the test statistic was 208.5689 and the p-value was 3.7e-24. For “other” landcover, the test statistic was 1.1297 and the p-value was 3.5818e-3. For Pastoral landcover the test statistic is 50.8314 and the p-value was 3.7e-24. Finally, for Urban landcover, the test statistic was 3.7878 and the p-value was 1.5945. This information is shown in tables 11, 12, 13 and 14.

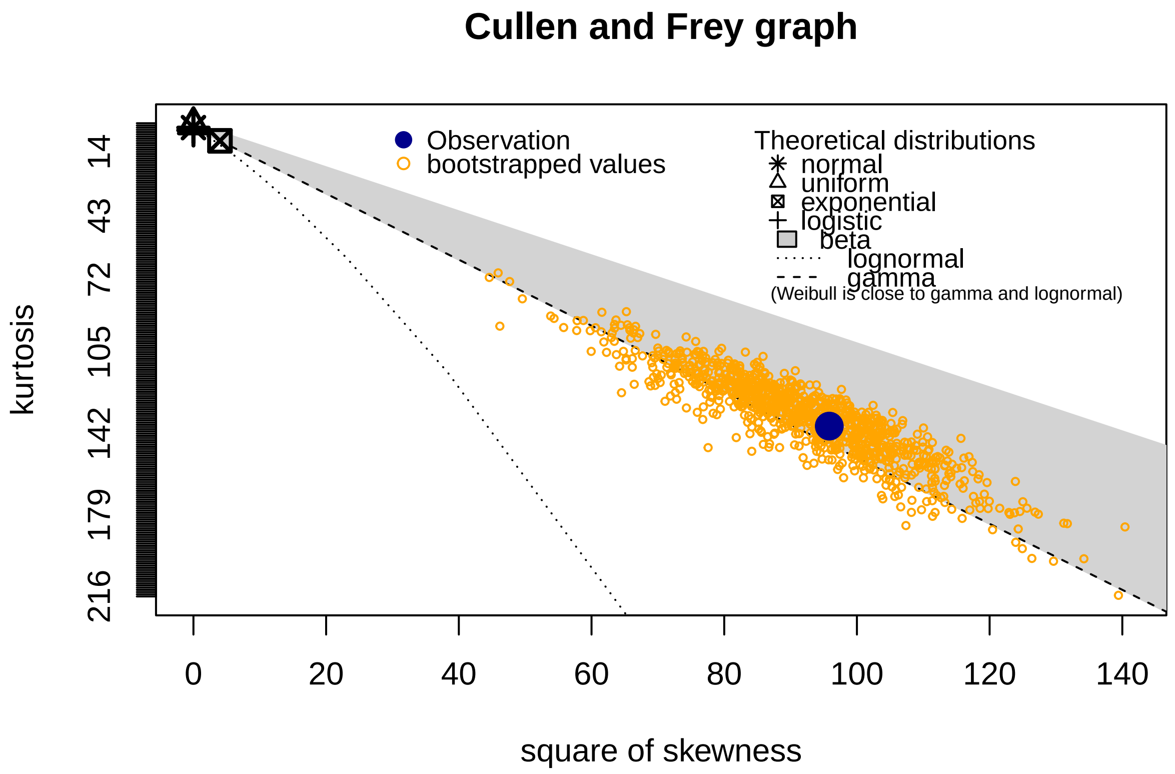


Figure 10: Cullen and Frey Graph of Lake Depth

Table 10: Anderson-Darling Test Statistic and P-value for Ammoniacal Nitrogen in Exotic Forest Landcovers

	Ammoniacal Nitrogen in Exotic Forest Landcovers
Test Statistic	12.2369
P-Value	3.7e-24

Table 11: Anderson-Darling Test Statistic and P-value for Ammoniacal Nitrogen in Native Landcovers

	Ammoniacal Nitrogen in Native Landcovers
Test Statistic	208.5689
P-Value	3.7e-24

Table 12: Anderson-Darling Test Statistic and P-value for Ammoniacal Nitrogen in Other Landcovers

	Ammoniacal Nitrogen in Other Landcovers
Test Statistic	1.1297
P-Value	3.581789e-03

Table 13: Anderson-Darling Test Statistic and P-value for Ammoniacal Nitrogen in Pastoral Landcovers

	Ammoniacal Nitrogen in Pastoral Landcovers
Test Statistic	50.8314
P-Value	3.7e-24

Table 14: Anderson-Darling Test Statistic and P-value for Ammoniacal Nitrogen in Urban Landcovers

	Ammoniacal Nitrogen in Urban Landcovers
Test Statistic	3.7878
P-Value	1.594502e-09

As the groups are not all normally distributed, we cannot perform a one-way ANOVA test, so we have performed a Kruskal-Wallis test for difference in means. The null hypothesis for this test is that the mean Ammoniacal Nitrogen level in each landcover type is equal, and the alternative hypothesis is that at least one of the means is different to the others. The test statistic for this test was 558.22 and the p-value was less than 2.2e-16. This means we reject the null hypothesis and conclude that at least one of the landcover types has a different mean Ammoniacal Nitrogen level than the others.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NH4N by land
## Kruskal-Wallis chi-squared = 558.22, df = 4, p-value < 2.2e-16
```

Next we conducted a pairwise Wilcox test for difference in means between pairs. This concluded that there is a difference between mean Ammoniacal Nitrogen levels in lakes with Exotic forest landcover and lakes with 'Other' landcover, as the p-value is 0.0004. There is also a difference in mean Ammoniacal Nitrogen levels in lakes with 'Other' landcover and lakes with Native landcover, with a p-value of 0.0001. Lakes with Pastoral landcover have a 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. Lakes with Urban landcover have a 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
```

For Chlorophyll-A, we have conducted an Anderson-Darling test for normality on the Exotic forest group. The test statistic was 3.5283 and the p-value was 7.4946e-9 (shown in table 15), rejecting the null hypothesis that the distribution of Chlorophyll-A in lakes with Exotic landcover is normal, and concluding the distribution of Chlorophyll-A in lakes with Exotic l=forest landcover is not normal, and therfore we cannot use a one-way ANOVA test.

We have performed a Kruskal-Wallis test for difference in means. The null hypothesis for this test is that the mean Chlorophyll-A level in each landcover type is equal, and the alternative hypothesis is that at least

Table 15: Anderson-Darling Test Statistic and P-value for Chlorophyll-A in Exotic Forest Landcovers

	Chlorophyll-A in Exotic Forest Landcovers
Test Statistic	3.5283
P-Value	7.494552e-09

Table 16: Anderson-Darling Test Statistic and P-value for Total Phosphorus in Exotic Forest Landcovers

	Total Phosphorus in Exotic Forest Landcovers
Test Statistic	9.7701
P-Value	1.00974e-23

one of the means is different to the others. The test statistic for this test was 1155.6 and the p-value was less than 2.2e-16. This means we reject the null hypothesis and conclude that at least one of the landcover types has a different mean Chlorophyll-A level than the others.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CHLA by land
## Kruskal-Wallis chi-squared = 1155.6, df = 4, p-value < 2.2e-16
```

Next we conducted a pairwise Wilcox test for difference in means 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
```

For Total Phosphorus, we have conducted an Anderson-Darling test for normality on the Exotic forest group. The test statistic was 9.7701 and the p-value was 1.0097e-23 (shown in table 16), rejecting the null hypothesis that the distribution of Total Phosphorus in lakes with Exotic landcover is normal, and concluding the distribution of Total Phosphorus in lakes with Exotic forest landcover is not normal, and therefore we cannot use a one-way ANOVA test.

We have performed a Kruskal-Wallis test for difference in means. The null hypothesis for this test is that the mean Total Phosphorus level in each landcover type is equal, and the alternative hypothesis is that at least one of the means is different to the others. The test statistic for this test was 1449.6 and the p-value was

Table 17: Anderson-Darling Test Statistic and P-value for Total Nitrogen in Exotic Forest Landcovers

	Total Nitrogen in Exotic Forest Landcovers
Test Statistic	3.5616
P-Value	6.220727e-09

less than 2.2e-16. This means we reject the null hypothesis and conclude that at least one of the landcover types has a different mean total Phosphorus level than the others.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TP by land
## Kruskal-Wallis chi-squared = 1449.6, df = 4, p-value < 2.2e-16
```

Next 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
```

For Total Nitrogen, we have conducted an Anderson-Darling test for normality on the Exotic forest group. The test statistic was 3.5616 and the p-value was 6.2207e-9 (shown in table 17), rejecting the null hypothesis that the distribution of Total Phosphorus in lakes with Exotic landcover is normal, and concluding the distribution of Total Nitrogen in lakes with Exotic forest landcover is not normal, and therefore we cannot use a one-way ANOVA test.

We have performed a Kruskal-Wallis test for difference in means. The null hypothesis for this test is that the mean Total Nitrogen level in each landcover type is equal, and the alternative hypothesis is that at least one of the means is different to the others. The test statistic for this test was 1340.3 and the p-value was less than 2.2e-16. This means we reject the null hypothesis and conclude that at least one of the landcover types has a different mean total Nitrogen level than the others.

```
##
##  Kruskal-Wallis rank sum test
```

Table 18: Anderson-Darling Test Statistic and P-value for Clarity in Exotic Forest Landcovers

	Clarity in Exotic Forest Landcovers
Test Statistic	4.3504
P-Value	7.582182e-11

```
##
## data:  TN by land
## Kruskal-Wallis chi-squared = 1340.3, df = 4, p-value < 2.2e-16
```

Next we conducted a pairwise Wilcox test for difference in means between pairs. This concluded 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
```

For Clarity, we have conducted an Anderson-Darling test for normality on the Exotic forest group. The test statistic was 4.3504 and the p-value was 7.522e-11 (shown in table 18), rejecting the null hypothesis that the distribution of Clarity in lakes with Exotic landcover is normal, and concluding the distribution of Clarity in lakes with Exotic forest landcover is not normal, and therefore we cannot use a one-way ANOVA test.

We have performed a Kruskal-Wallis test for difference in means. The null hypothesis for this test is that the mean Clarity level in each landcover type is equal, and the alternative hypothesis is that at least one of the means is different to the others. The test statistic for this test was 602.61 and the p-value was less than 2.2e-16. This means we reject the null hypothesis and conclude that at least one of the landcover types has a different mean Clarity level than the others.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SECCHI by land
## Kruskal-Wallis chi-squared = 602.61, df = 4, p-value < 2.2e-16
```

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
```