STAT394 / Project EDA

Group 7

12-08-2022

Contents


```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## Loading required package: MASS
```

```
## Loading required package: survival
```

1 Exploratory Data Analysis: Tasmania Lakes

1.1 Overview

The dataset features 63 variables (environmental indicators, spatial data, etc.) for 50 Tasmanian lakes. Understanding how natural environmental factors (rock type, precipitation, etc.) may drive environmental indicators of lakes (pH, temperature, etc.) is of particular interest. This is because Tasmania features a strong east-west binary division in many natural characteristics, including rock type. To address this preliminary area of interest, this EDA will focus on ten key variables: lake elevation, lake depth, lake pH, water temperature, precipitation, lake turbidity, lake area, the presence of glacial sediment, the presence of felsic rock and the presence of dolerite. The latter three variables are categorical, and the final two are proxies for two distinct rock classifications.

After loading the dataset, the data frame is formatted such that the variables of interest are extracted. The categorical variables are also reclassified as factors. The first five entries of each variable can be examined in Table 1 below.

Table 1: Tasmania Lakes Variables of Interest: First Five Entries

Elev	Depth	pH	Twater	Precip	Turbid	GlaSed	Felsic	Dolerite	Lake_Area
1066.0	0.95	5.2	18.10	2472	2.8	2	0	0	0.282251
39.7	0.13	5.8	29.15	1573	0.9	0	0	0	7.234741
576.7	1.90	4.7	18.55	3170	0.6	2	0	0	19.934649
70.6	1.48	4.8	27.10	1711	1.6	0	0	0	1.281947
149.3	10.07	4.9	21.25	1497	0.9	0	0	0	4.268296

1.2 Exclusivity of Rock Types: Interaction Table

Interactions between the factors (categorical variables) can be examined in Table 2 below.

Table 2: Interaction of Rock Types 'Felsic' and 'Dolerite'

Interaction	Frequency
Not Felsic + Not Dolerite	24
Felsic + Not Dolerite	4
Not Felsic + Dolerite	22
Felsic + Dolerite	0

Table 2 demonstrates that significant proportions of the lakes have either no felsic rock and no dolerite, or dolerite and no felsic rock. Importantly, Table 2 also shows that none of the 50 lakes have both felsic rock and dolerite. Given this exclusivity between the two distinct rock types, it could be interesting to further investigate how being one of these two rock types may drive variations in environmental indicators at the lakes.

1.3 Descriptive Summary Statistics

Table 3, below, presents a quantitative summary of the descriptive statistics for each of the seven numeric variables.

Table 3: Quantitative summary of numeric Tasmania lakes data

	Elev	Depth	pH	Twater	Precip	Turbid	Lake_Area
Sample Size	50	50	50	50	50	50	50
Minimum	9.40	0.10	4.10	9.55	1043.00	0.10	0.04
Lower Quartile	754.80	0.99	4.90	14.10	1576.25	0.50	1.10
Median	978.70	2.85	5.90	18.10	1809.50	0.75	2.64
Upper Quartile	1174.00	8.68	6.47	20.15	2462.75	1.00	4.26
Maximum	1451.00	35.20	7.00	29.45	3219.00	10.20	22.84
IQR	419.20	7.69	1.57	6.05	886.50	0.50	3.16
Range	1441.60	35.10	2.90	19.90	2176.00	10.10	22.80
Mean	869.31	5.05	5.72	17.45	2040.12	1.09	4.23
Standard Deviation	386.82	6.12	0.83	4.81	567.81	1.44	5.29
Skewness	-1.01	2.64	-0.24	0.45	0.21	5.14	2.22
Kurtosis	-0.09	9.56	-1.44	-0.12	-1.07	29.52	4.23

Table 3 indicates that the different numeric variables may exhibit different distributions. Turbidity ('Turbid'), Depth and Lake Area present as the most assymetric, with the highest 'skewness'. For each of these three

variables, means are greater than medians, suggesting that the assymetry follows a right-skew whereby a smaller number of highly turbid, deep and large lakes are in the sample. Turbidity ('Turbid') also has an exceptionally high 'kurtosis', suggesting that its distribution is very tail-heavy.

Precipitation ('Precip') and pH appear to be distributed the most symmetrically, with the smallest 'skewness' and means that more closely approximate their medians.

1.4 Covariance & Correlation Matrices

Estimates of the covariance matrix ($\hat{\mathbf{\Sigma}}$) and the correlation matrix ($\hat{\boldsymbol{\rho}}$) can be computed for each numeric variable. The correlation matrix is then visualised in Figure 1 (following page).

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 149628.37690 & -25.5461547 & 191.9572122 & -949.6211469 & -9841.09869 & 76.5131469 & -433.4444203 \\ -25.54615 & 37.5060776 & 0.1626029 & -5.1122086 & 570.53384 & -2.1255098 & 11.9763452 \\ 191.95721 & 0.1626029 & 0.6833837 & -0.9020286 & -206.39657 & -0.1284612 & -0.4749541 \\ -949.62115 & -5.1122086 & -0.9020286 & 23.1601878 & -156.80967 & 1.2311796 & 0.2883980 \\ -9841.09869 & 570.5338449 & -206.3965714 & -156.8096735 & 322403.94449 & -58.2411429 & 878.3889430 \\ 76.51315 & -2.1255098 & -0.1284612 & 1.2311796 & -58.24114 & 2.0628612 & -1.0662397 \\ -433.44442 & 11.9763452 & -0.4749541 & 0.2883980 & 878.38894 & -1.0662397 & 27.9921513 \end{pmatrix}$$

$$\hat{\boldsymbol{\rho}} = \begin{pmatrix} 1 & -0.01078370 & 0.60029559 & -0.51012001 & -0.04480603 & 0.13771907 & -0.21179139 \\ -0.01078370 & 1 & 0.03211775 & -0.17345495 & 0.16407047 & -0.24164470 & 0.36961990 \\ 0.60029559 & 0.03211775 & 1 & -0.22673416 & -0.43971371 & -0.10819439 & -0.10859283 \\ -0.51012001 & -0.17345495 & -0.22673416 & 1 & -0.05738543 & 0.17812110 & 0.01132669 \\ -0.04480603 & 0.16407047 & -0.43971371 & -0.05738543 & 1 & -0.07141587 & 0.29239406 \\ 0.13771907 & -0.24164470 & -0.10819439 & 0.17812110 & -0.07141587 & 1 & -0.14031423 \\ -0.21179139 & 0.36961990 & -0.10859283 & 0.01132669 & 0.29239406 & -0.14031423 & 1 \end{pmatrix}$$

1.5 Visualing the Correlation Matrix

Figure 1, below, provides a visual interpretation of the correlation matrix.

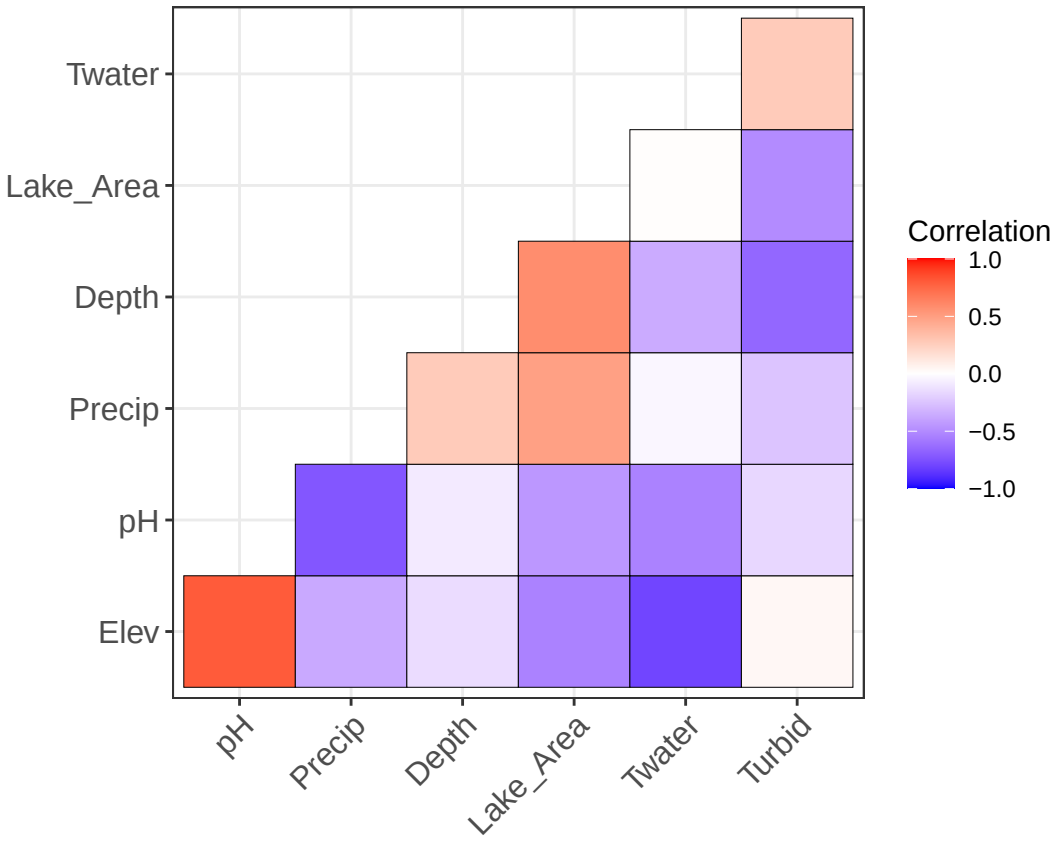

Figure 1: Plot of the numeric Tasmania lakes data correlation matrix

Figure 1 illustrates a range of correlations amongst the seven numeric variables. Of note, Elevation ('Elev') appears to have a strong positive correlation with pH and a strong negative correlation with Water Temperature ('Twater'). This implies that as the elevation of lakes increases, pH increases and water temperature decreases. pH also appears to have a strong negative correlation with Precipitation ('Precip'). Depth and Lake Area appear positively correlated, and Turbidity ('Turbid') and Depth appear negatively correlated, both with moderate strengths.

There appears to be little-to-no correlation between Water Temperature and Precipitation, or between Water Temperature and Lake Area. The pairs of Depth and pH, and Turbidity and Elevation, also appear to have very weak negative and positive correlations, respectively.

1.6 Pairs Plot

Figure 2, below, provides the pairs plot.

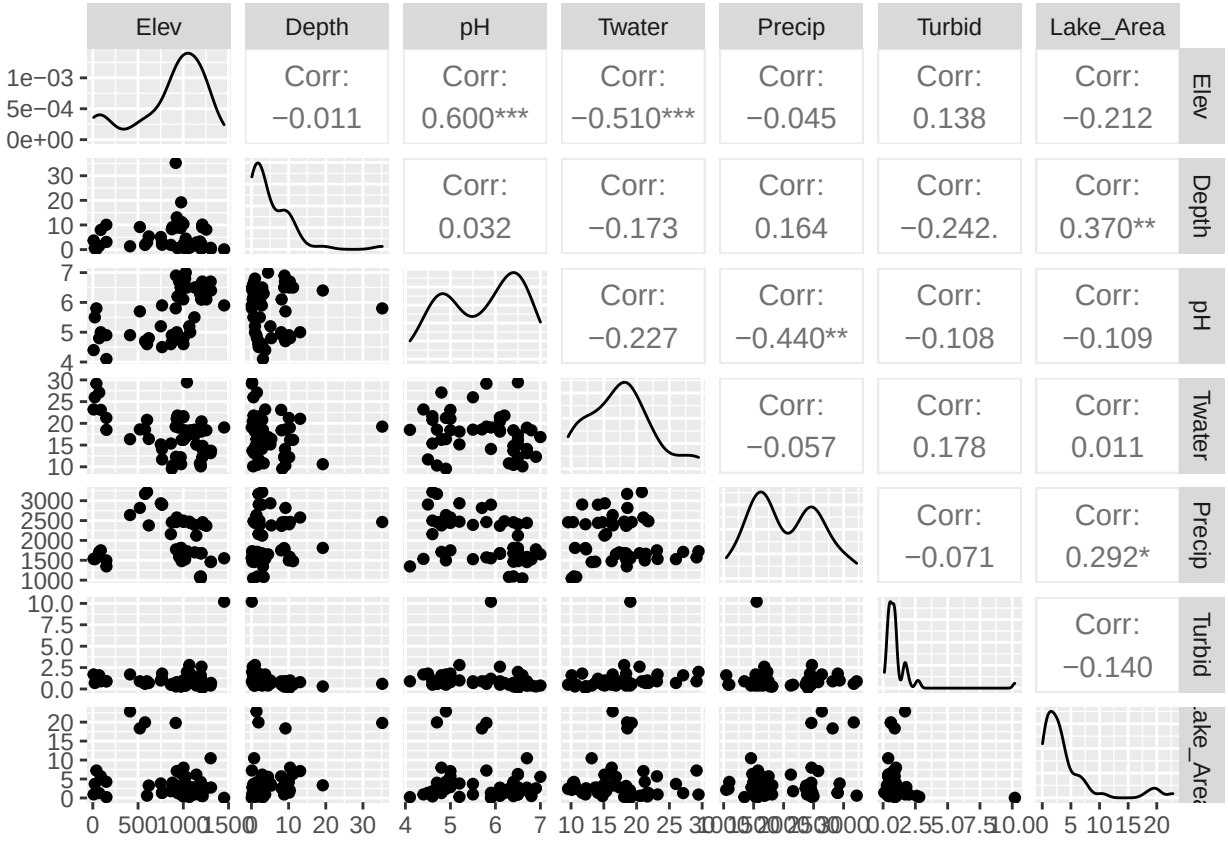

Figure 2: Pairs plot for numeric Tasmania lakes data

Figure 2 provides more insight into the distributions of each variable, after preliminary interpretations were made from the summary statistics. Strong asymmetry amongst many variables is confirmed, most strikingly for Turbidity, Depth and Lake Area. Though all other variables are less skewed, none of the seven appear to follow a Gaussian distribution.

Additionally, Precipitation and pH appear to follow a bimodal distribution. Both variables should be investigated in more detail, as their distributions may inform the presence of distinct sub-groups (i.e. sub-populations) of lakes that have been combined together in the sample. Referring back to Figure 1, Precipitation and pH exhibited a moderate-to-strong negative correlation. Given that relationship, the similarity in their distributions may imply that one variable is driving the other. This should be explored further.

1.7 Multiple Density Plots

Figure 3 and Figure 4 examine how pH and Precipitation, respectively, may vary depending on the categorical factor 'Dolerite', in an attempt to explain the bimodel distributions observed for pH and Precipitation in the sample.

Background research on the topic has suggested unique combinations of minerals are released when different rock type erode (e.g. Dolerite vs no Dolerite). These minerals then wash into lakes and possibly affect pH, such that rock type may drive lake pH. Background research also suggests that variations in precipitation around different lakes may also promote the abundance of different rock types.

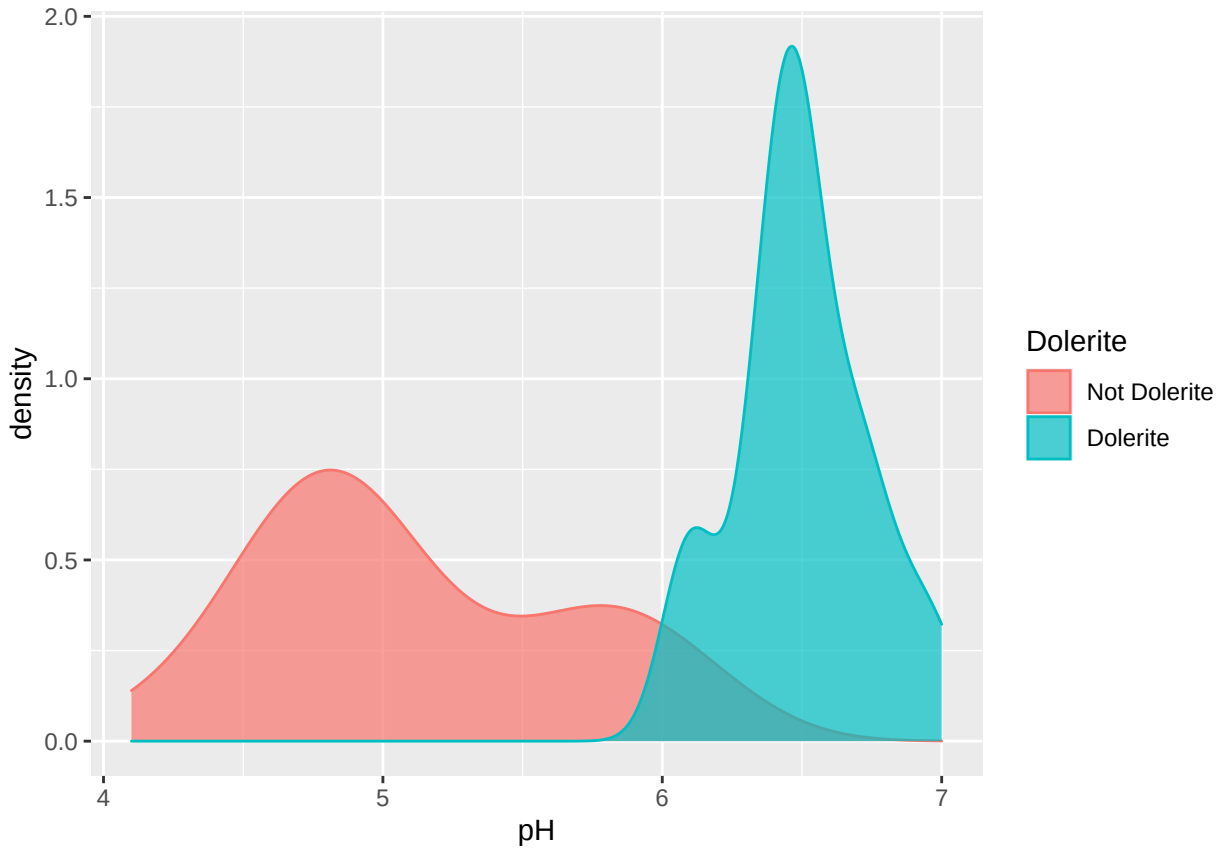

Figure 3: Multiple density plot of pH for the presence or absence of Dolerite

Figure 3 provides a strong indication that the two groups of lakes (lakes with Dolerite rock, lakes without Dolerite rock) have different mean pH. Specifically, lakes with Dolerite rock appear have have a higher pH with less variance than lakes without Dolerite. This may present an interesting area for future analyses.

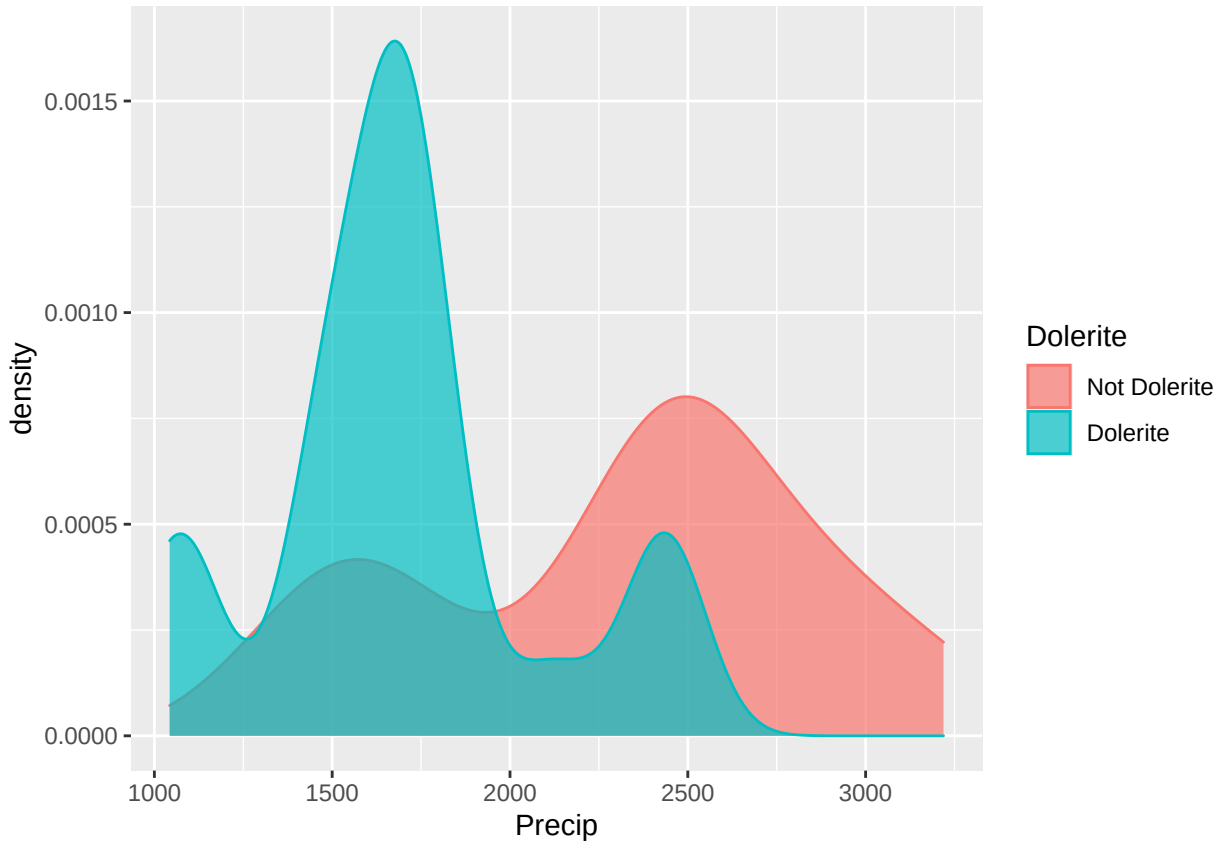

Figure 4: Multiple density plot of Precipitation for the presence or absence of Dolerite

Figure 4 demonstrates a more ambiguous relationship, though it appears that lakes with Dolerite rock may have lower mean Precipitation overall. Any assumptions on this particular relationship are made difficult, however, because the lakes without Dolerite still appear to have a bimodal Precipitation distribution, and the lakes with Dolerite appear to have a trimodal Precipitation distribution. This suggests that there may be futher sub-categories amongst the lakes to explore.

EDA continues on next page.

1.8 Scatterplot with Marginal Boxplot

Two pairs of numeric variables will be explored in more detail with scatterplots: Elevation and pH, and Elevation and Water Temperature. These two pairs capture strong positive and negative correlations in the Tasmania lakes dataset, respectively.

Figure 5 provides the scatterplot (with marginal boxplots) for Elevation and pH. Figure 6 provides the scatterplot (with marginal boxplots) for Elevation and Water Temperature.

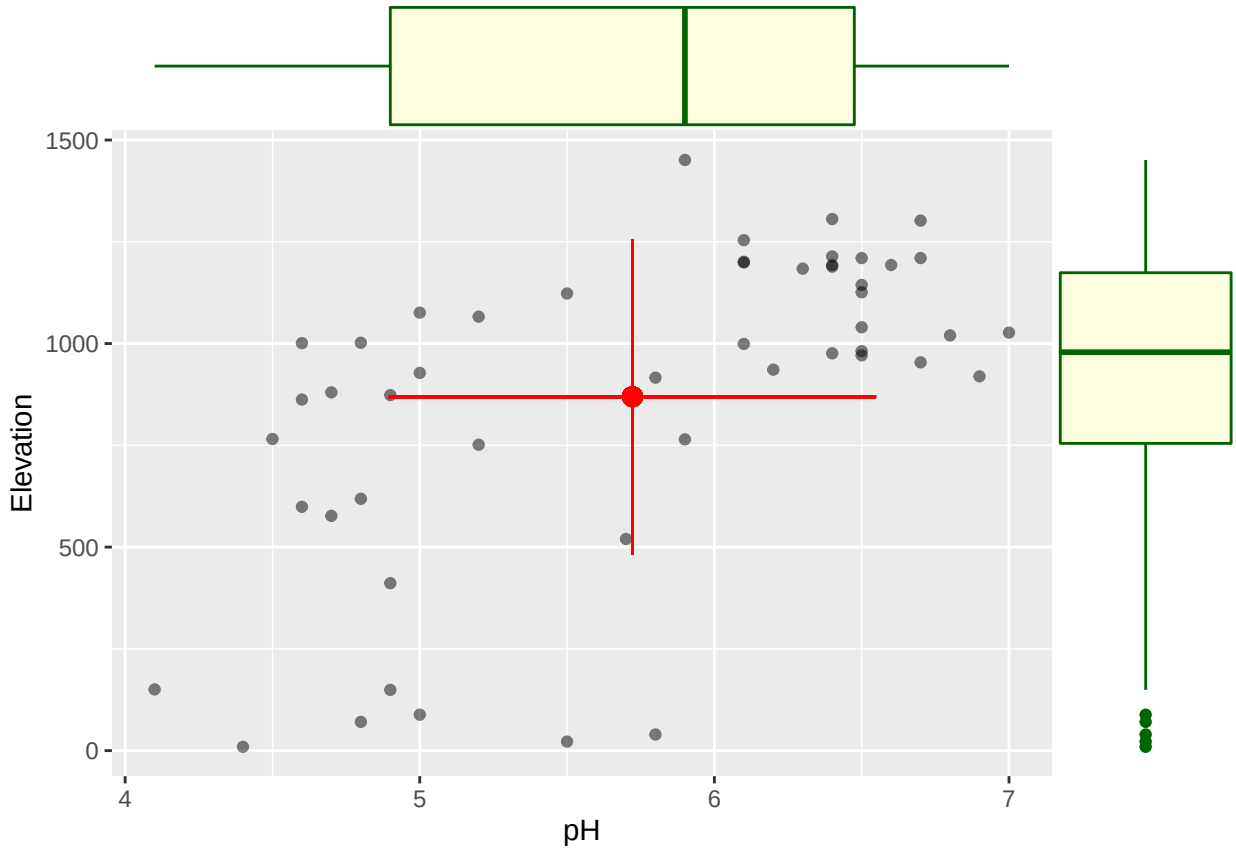

Figure 5: Scatterplot (with marginal boxplots) for 'Elevation' and 'pH' (Tasmania)

Figure 5 supports Figure 1, illustrating a positive correlation between Elevation and pH. The right-skew, inferred from the summary statistics, also appears in the boxplot distribution for Elevation

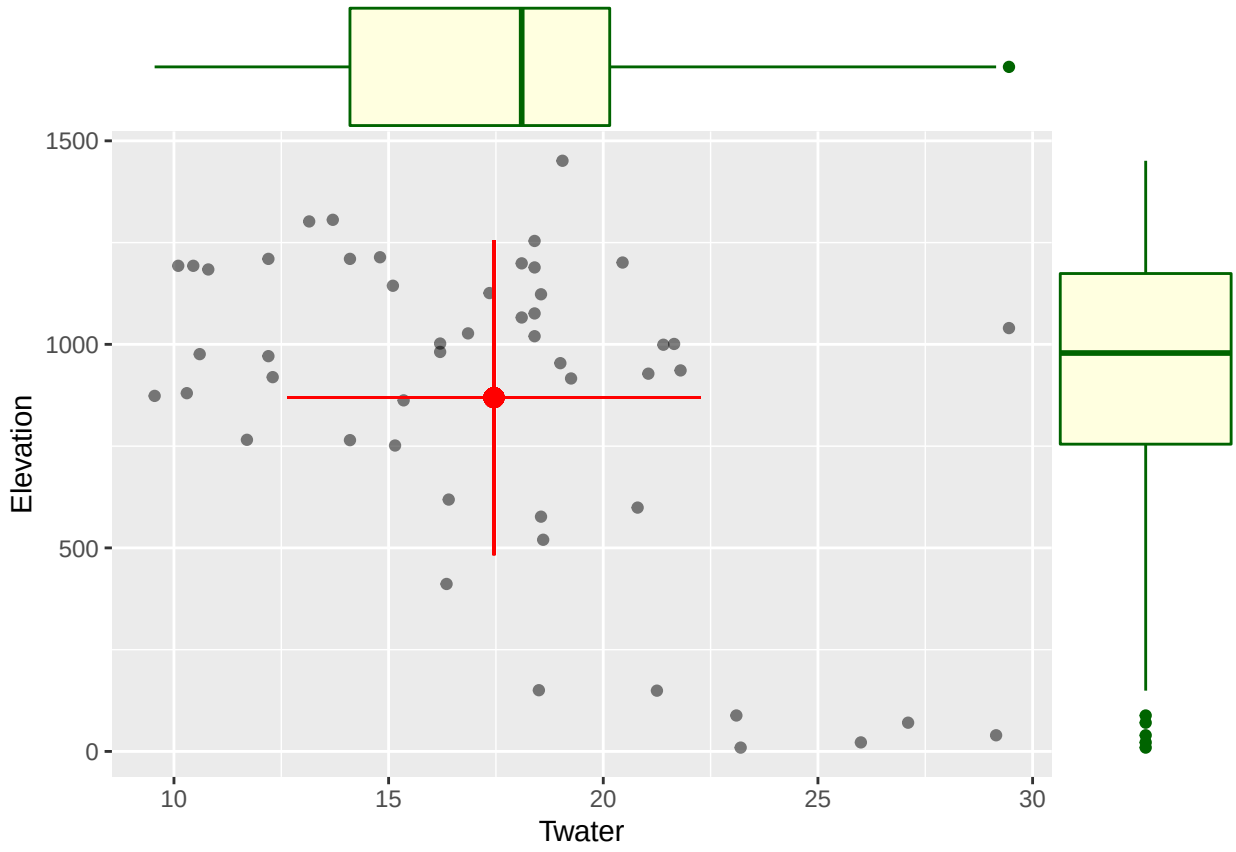

Figure 6: Scatterplot (with marginal boxplots) for ‘Elevation’ and ‘Water Temperature’ (Tasmania)

Figure 6 also supports Figure 1, illustrating a negative correlation between Elevation and Water Temperature The right-skew, inferred from the summary statistics, also appears in the boxplot distribution for both variables.

EDA continues on next page.

1.9 Cullen and Frey Plots

Figure 7, Figure 8 and Figure 9 (below) provide the Cullen and Frey plots for the numeric variables pH, Water Temperature and Turbidity. These three variables are selected as background research suggests that they function as response indicators, as opposed to drivers. The Cullen and Frey plots help to illustrate the potential distributions of the each variable.

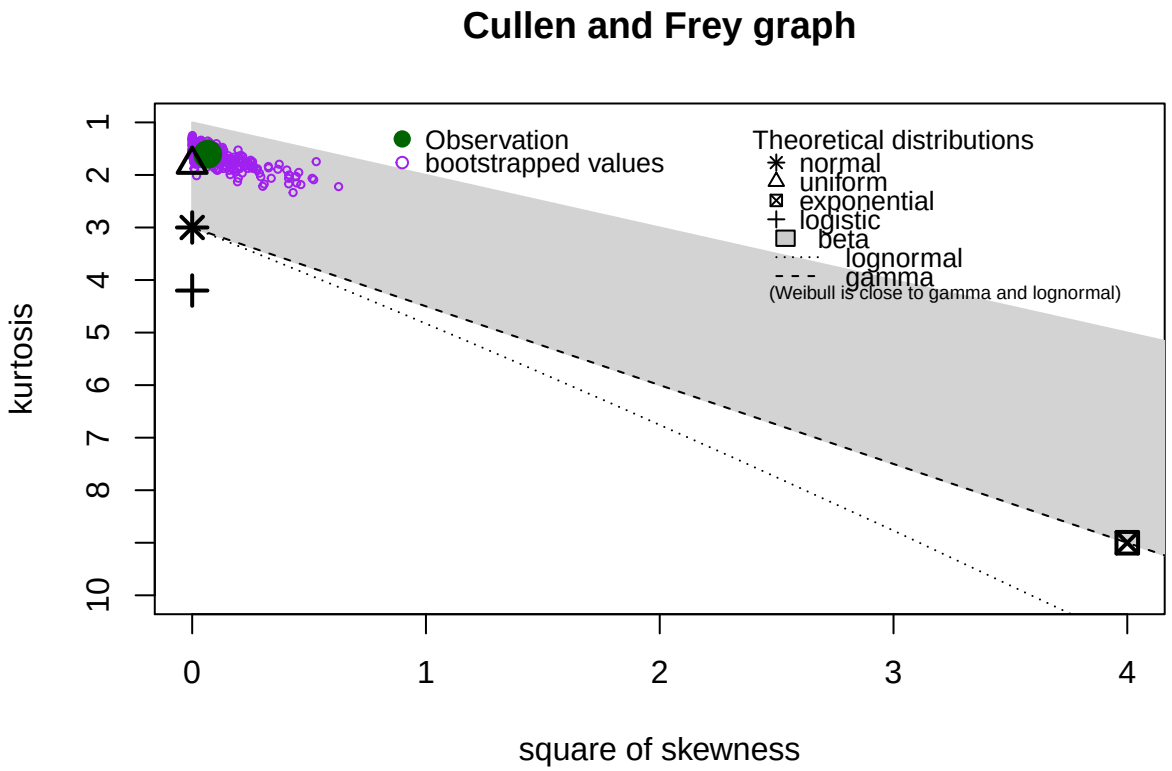

Figure 7: Cullen and Frey plot for pH of Tasmania lakes

```
## summary statistics
## ------
## min:  4.1   max:  7
## median:  5.9
## mean:  5.722
## estimated sd:  0.8266702
## estimated skewness:  -0.257977
## estimated kurtosis:  1.606766
```

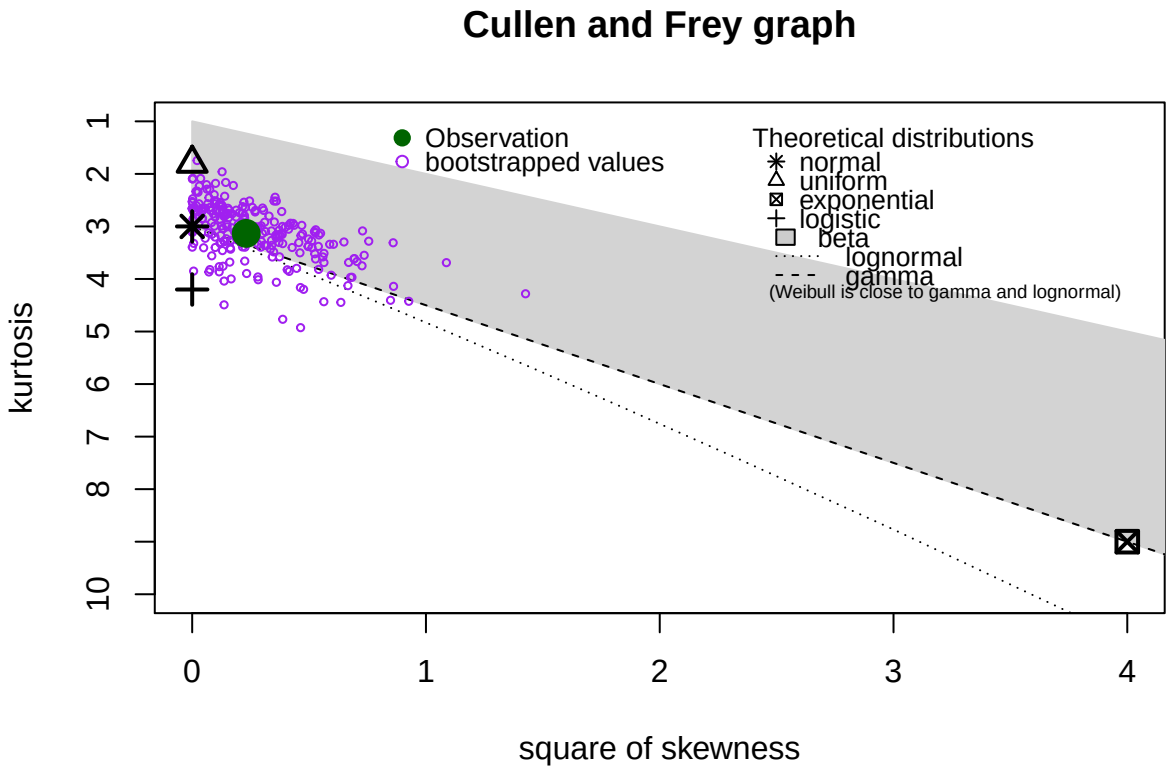

Figure 8: Cullen and Frey plot for Water Temperature of Tasmania lakes

```
## summary statistics
## ------
## min:  9.55   max:  29.45
## median:  18.1
## mean:  17.454
## estimated sd:  4.812503
## estimated skewness:  0.4806317
## estimated kurtosis:  3.132279
```

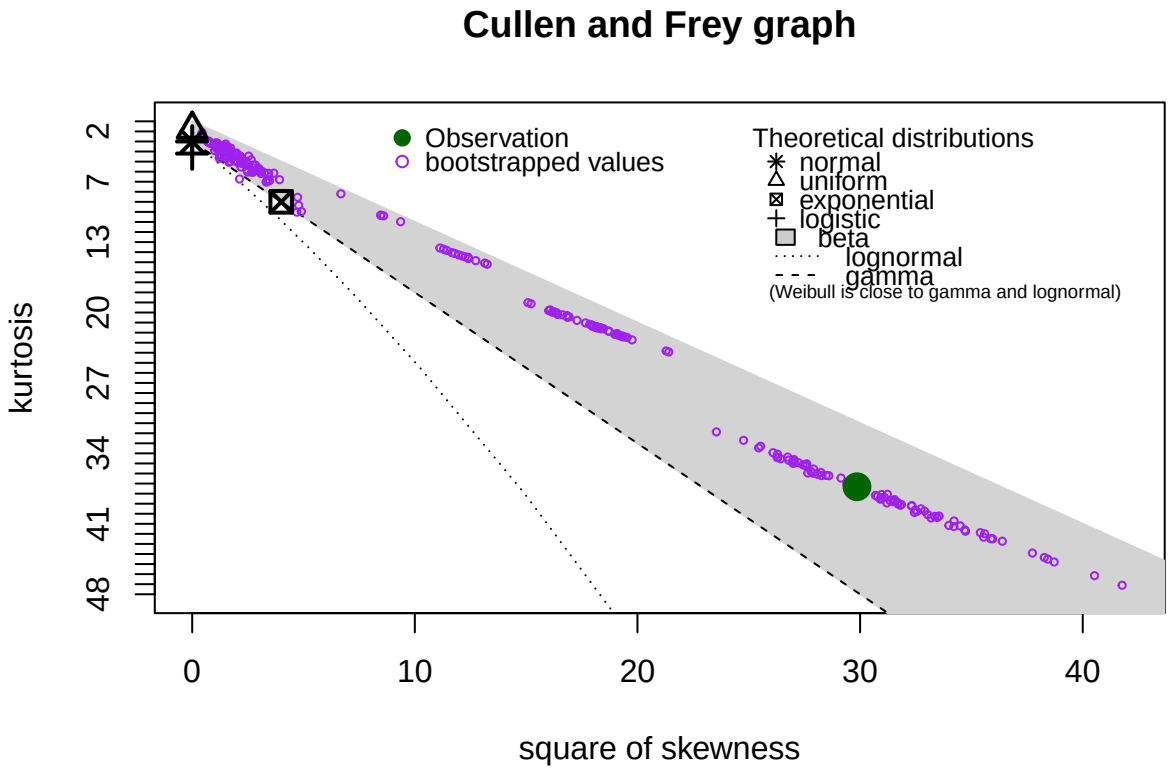

Figure 9: Cullen and Frey plot for Turbidity of Tasmania lakes

```
## summary statistics
## ------
## min:  0.1   max:  10.2
## median:  0.75
## mean:  1.086
## estimated sd:  1.436266
## estimated skewness:  5.464341
## estimated kurtosis:  37.31742
```

The plots suggest that each variable may follow a different distribution. Bootstrap values for pH in Figure 7 are clustered most tightly (all within the beta distribution region), with the observed value near the uniform distribution.

Water Temperature, by contrast, could follow a number of different distributions according to Figure 8. The bootstrap values range from within the beta distribution region, to across both the lognormal and gamma distribution lines.

Turbidity appears to follow the most peculiar distribution, with distinct elongated clusters of bootstrap values in Figure 9. As the summary statistics suggested, Turbidity (indicated here by the observed value) appears to be very far from the normal distribution.