Milestone 3

Group 6

2022-08-19

The NZ river dataset includes three categorical variables (Dominant landcover, Trend and Percent of annual change) and six numerical variables (E.Coli, Phosphorus, Nitrogen, Turbidity, Latitude and Longitude), which we will be conducting exploratory data analyses on.

EDA for categorical variables

Values in each categorical variable:

```
unique(df$dominant_landcover)
```

```
## [1] "Native"        "Pastoral"      "Urban area"    "Exotic forest"
```

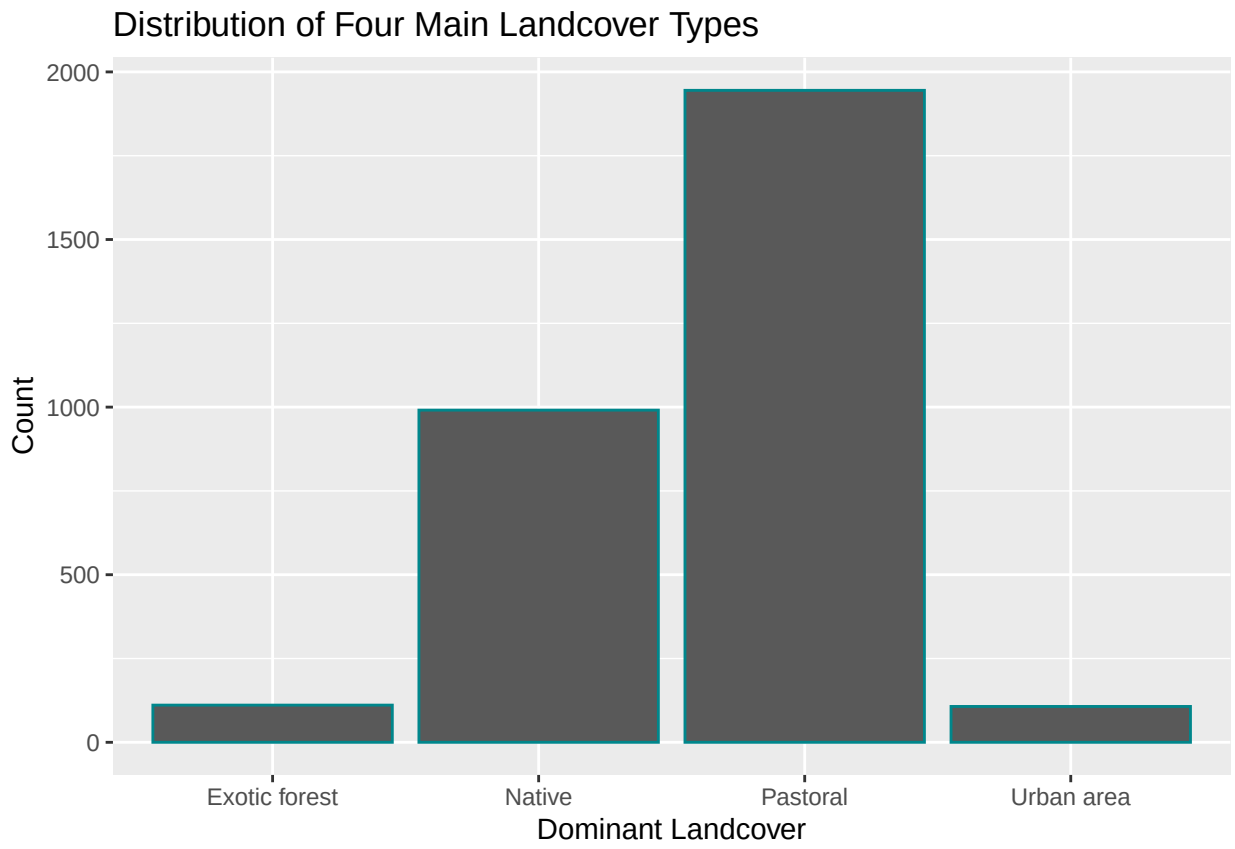
```
unique(df$Trend)
```

```
## [1] "Improving"     "Indeterminate" "Worsening"
```

```
unique(df$percent_annual_change)
```

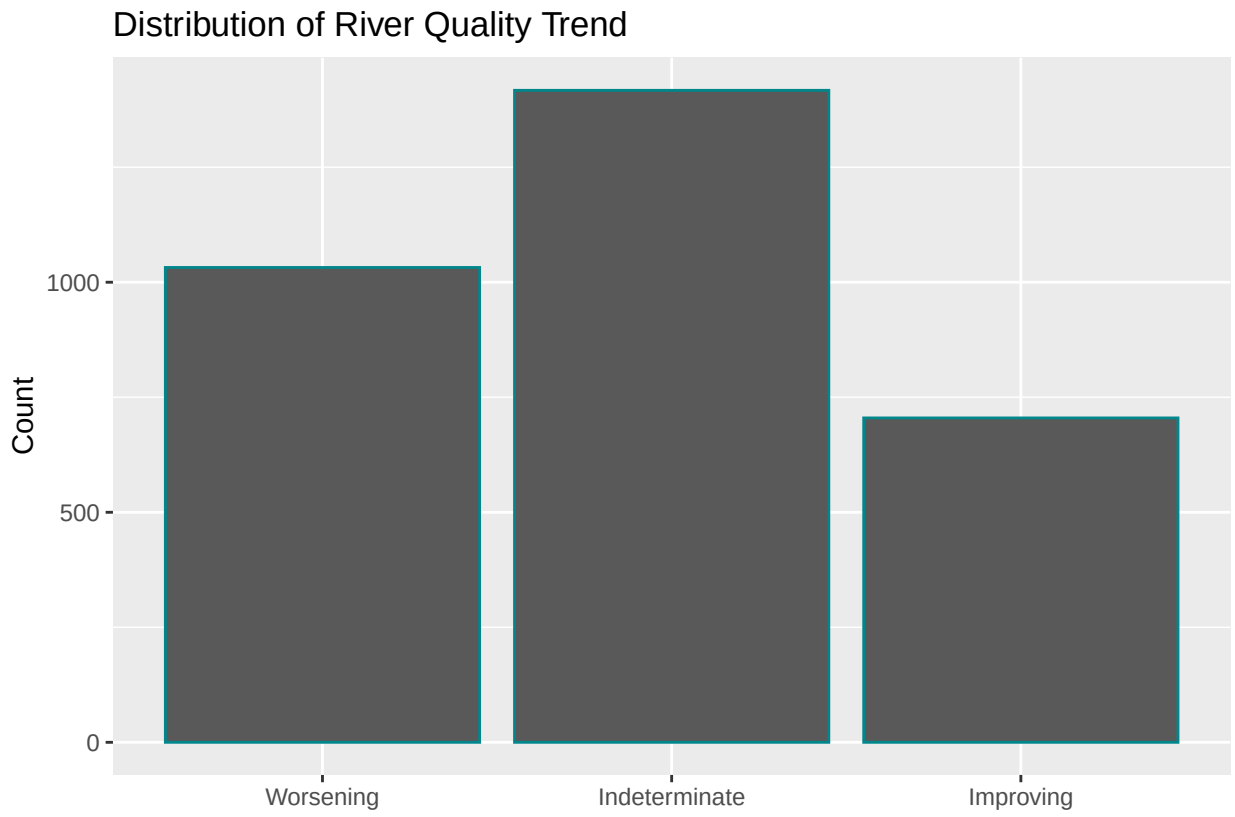
```
## [1] ">2% improving"  ">2% worsening"  "1-2% worsening" "0-1% worsening"
## [5] "0-1% improving" "1-2% improving"
```

Distribution of dominant land cover types



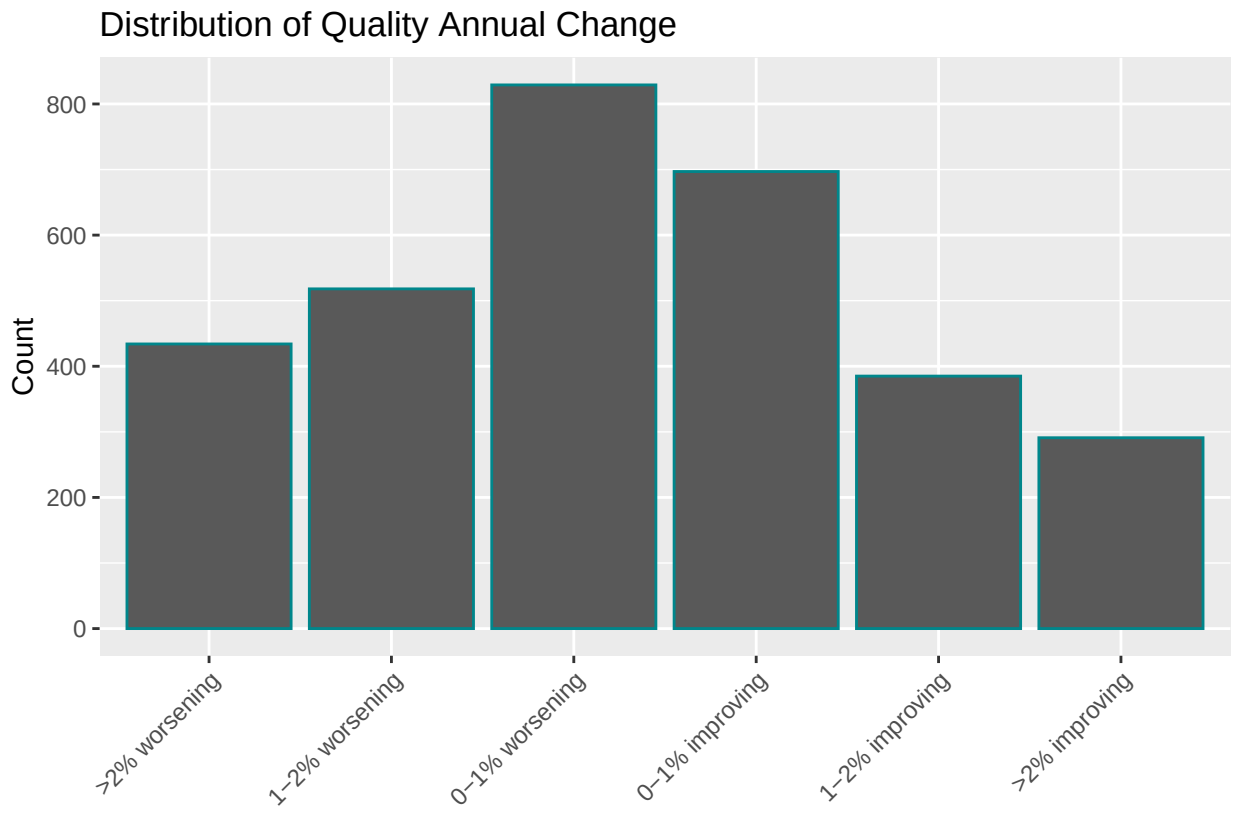
We can see in this graph, the distribution of dominant land cover over the monitored river sites throughout New Zealand. Pastoral land cover appeared to be the greatest among sites, followed by native land cover and comparably, very little land cover was exotic forest and urban areas.

Distribution of river quality trend



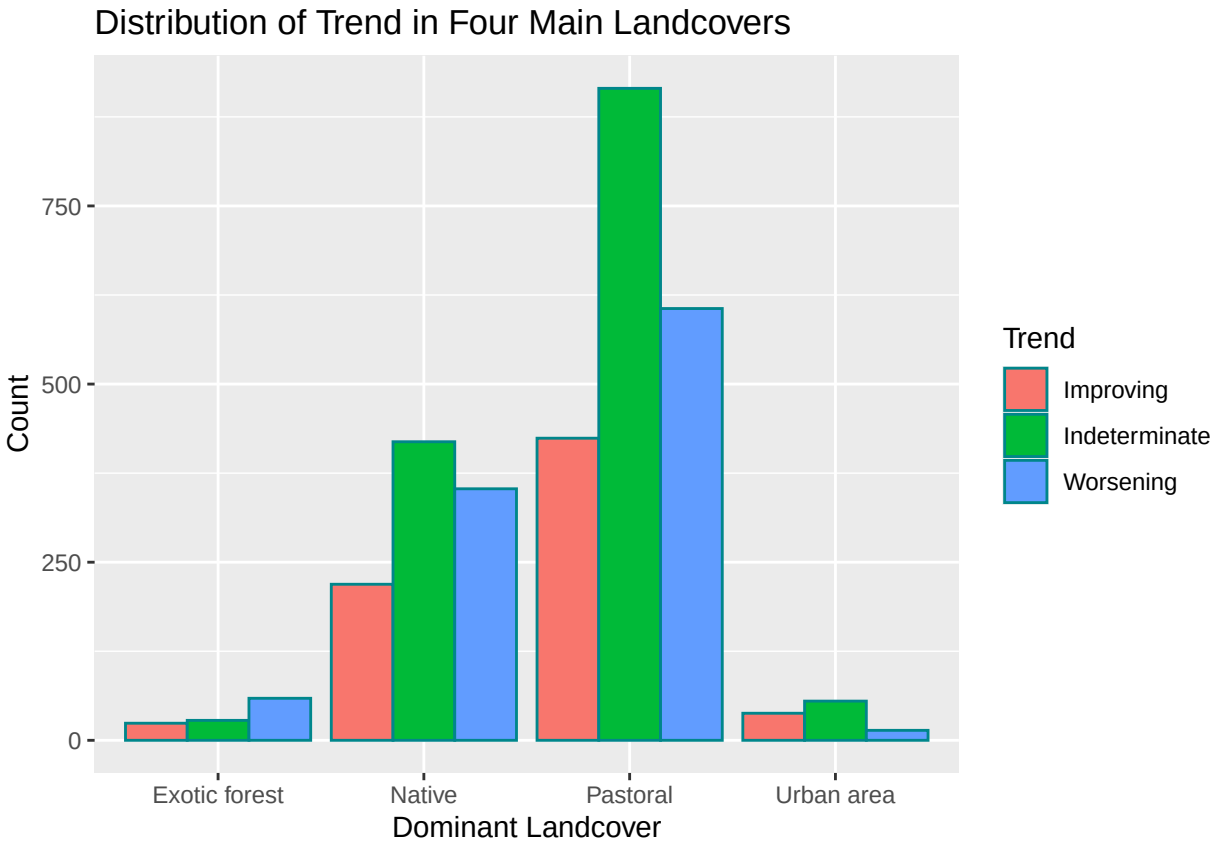
The graph above shows the distribution of trends in river quality, describing the direction of change of river health indicators. Trends in river health indicators were found to be worsening, indeterminate or improving. Most trends were shown to be indeterminate, closely followed by worsening trends. An improving trend was the least common for river health indicators across NZ.

Distribution of quality annual change



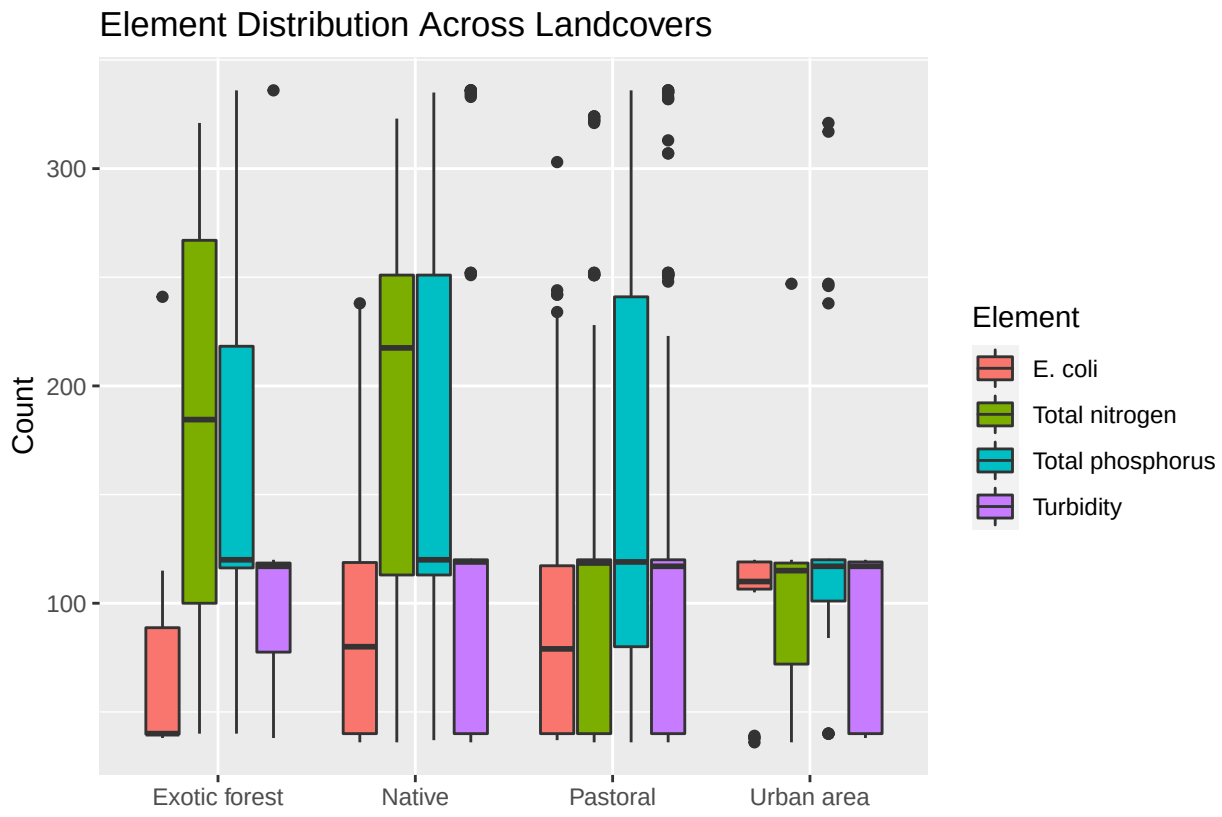
This graph depicts the distribution of annual change in river quality indicators, over the NZ river sites sampled. The % of annual change (either worsening or improving) was recorded for each indicator in each river site. We can see that the highest count of river indicators showed 0-1% annual change (worsening), followed by 0-1% improving, then by 1-2% and >2% worsening. 1-2% and >2% improving were the least common annual changes among river sites.

Distribution of trend in different land cover



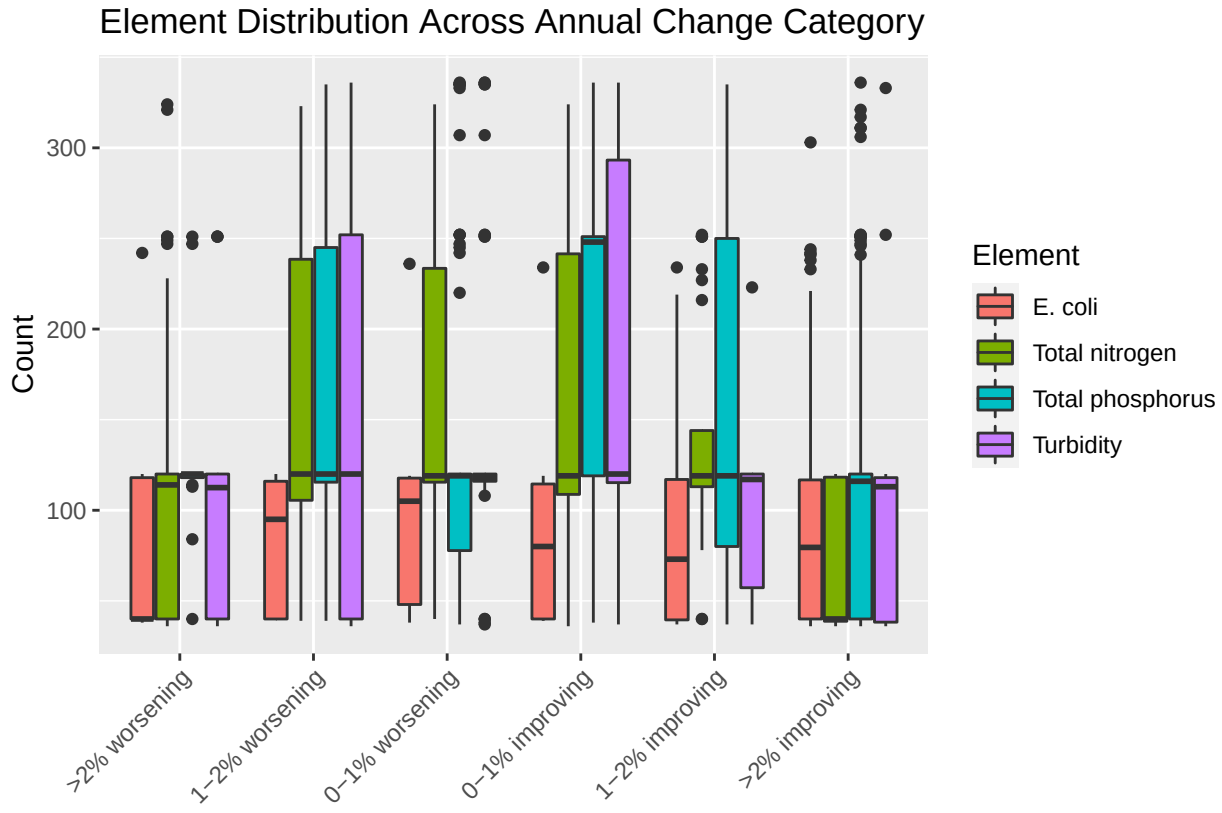
This graph shows the distribution of river quality trends over the four types of main landcover. Indeterminate trends appear to be the greatest in all landcover types, except for exotic forest for which it is mostly worsening. Trends of improvement appear to be the least common in all landcover types, except for urban areas, for which worsening trends are dominant.

Element Distribution Across Landcovers



The above graph depicts the distribution of river health elements/indicators over different land types. E.Coli tends to show low counts across land types, and much lower ranges than other indicators. Total nitrogen appears to be the greatest (mean/median) in exotic forest and native land, with large ranges. Total phosphorus shows to be greatly positively skewed and spread out for most land types. Conversely, turbidity appears to be greatly negatively skewed for all land types.

Element Distribution Across Annual Change Category



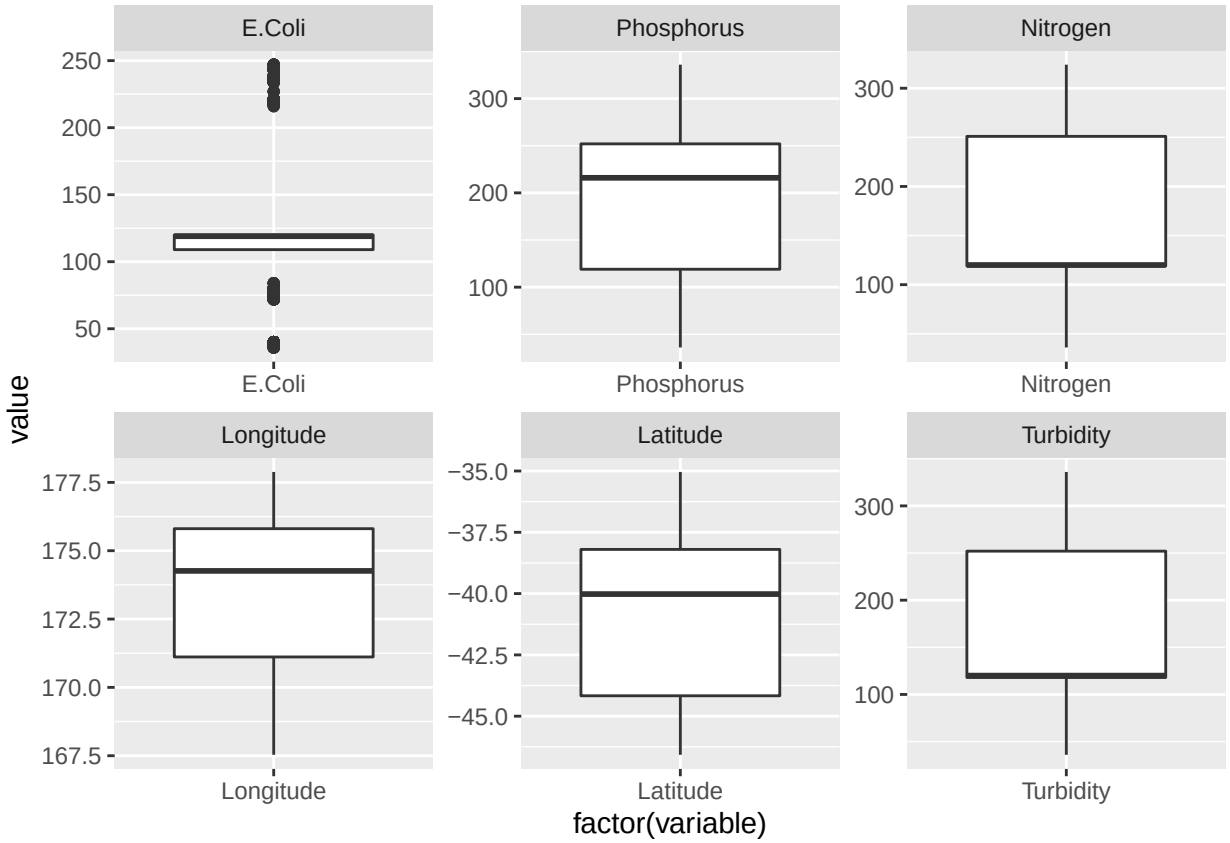
Finally, the above graph shows the distribution of annual change for each of the river health indicators. Data for E.Coli appears to be relatively similar across all levels of annual change. Total nitrogen, phosphorus and turbidity show high counts in 1-2% worsening and 0-1% improving categories, along with being greatly skewed.

EDA of the numerical variables

The following EDA investigates the numerical variables; the river health indicators of E.Coli, Phosphorus, Nitrogen, Turbidity and the respective locations (latitude and longitude) of the rivers.

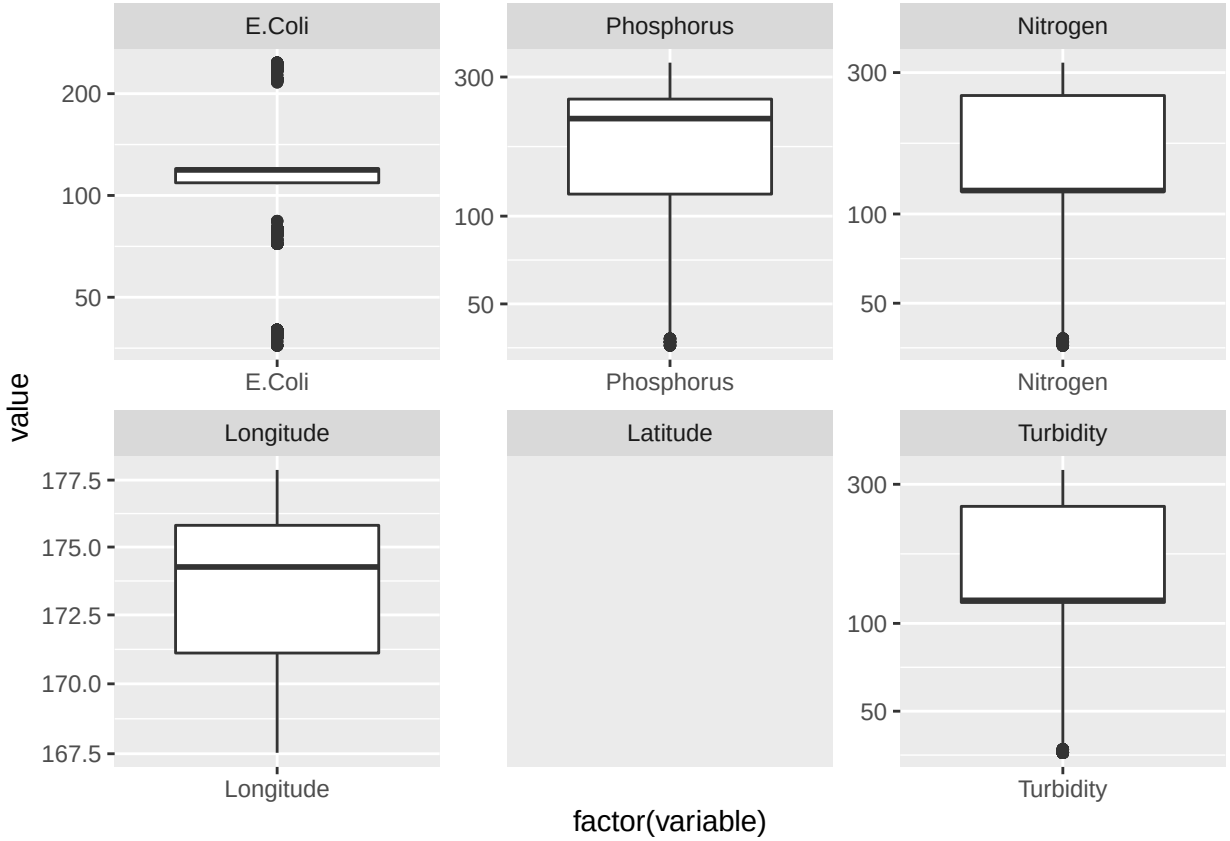
```
##      num_id          E.Coli        Phosphorus     Nitrogen       Longitude
##  Min.   :   1.0   Min.   : 36.0   Min.   : 36   Min.   : 36.0   Min.   :167.5
##  1st Qu.: 789.2   1st Qu.:109.0   1st Qu.:119   1st Qu.:119.0   1st Qu.:171.1
##  Median :1577.5   Median :119.0   Median :216   Median :120.0   Median :174.3
##  Mean   :1577.5   Mean   :109.2   Mean   :193   Mean   :181.3   Mean   :173.5
##  3rd Qu.:2365.8   3rd Qu.:120.0   3rd Qu.:252   3rd Qu.:251.0   3rd Qu.:175.8
##  Max.   :3154.0   Max.   :247.0   Max.   :336   Max.   :324.0   Max.   :177.9
##     Latitude        Turbidity
##  Min.   :-46.57   Min.   : 36.0
##  1st Qu.:-44.17   1st Qu.:118.0
##  Median :-40.02   Median :120.0
##  Mean   :-40.94   Mean   :178.6
##  3rd Qu.:-38.20   3rd Qu.:252.0
##  Max.   :-35.04   Max.   :336.0
```

Boxplots of river health indicators



The boxplots above show the distributions of data measured across the NZ river sites for river quality indicators and locations. E.Coli shows very concentrated data and the presence of outliers. Phosphorus appears to be evenly spread with a higher median. Nitrogen shows concentration of data in the upper quartile, and a lower median than phosphorus. Turbidity shows a similar distribution to nitrogen.

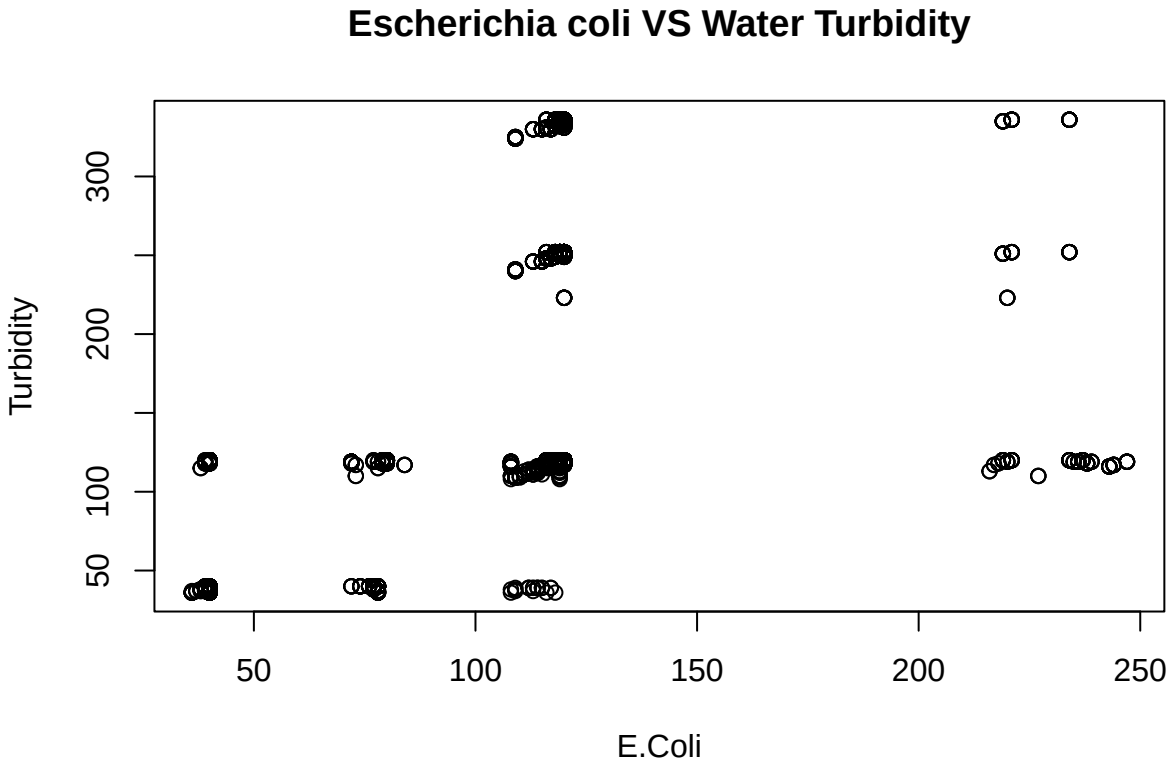
Boxplots with a logarithmic scale

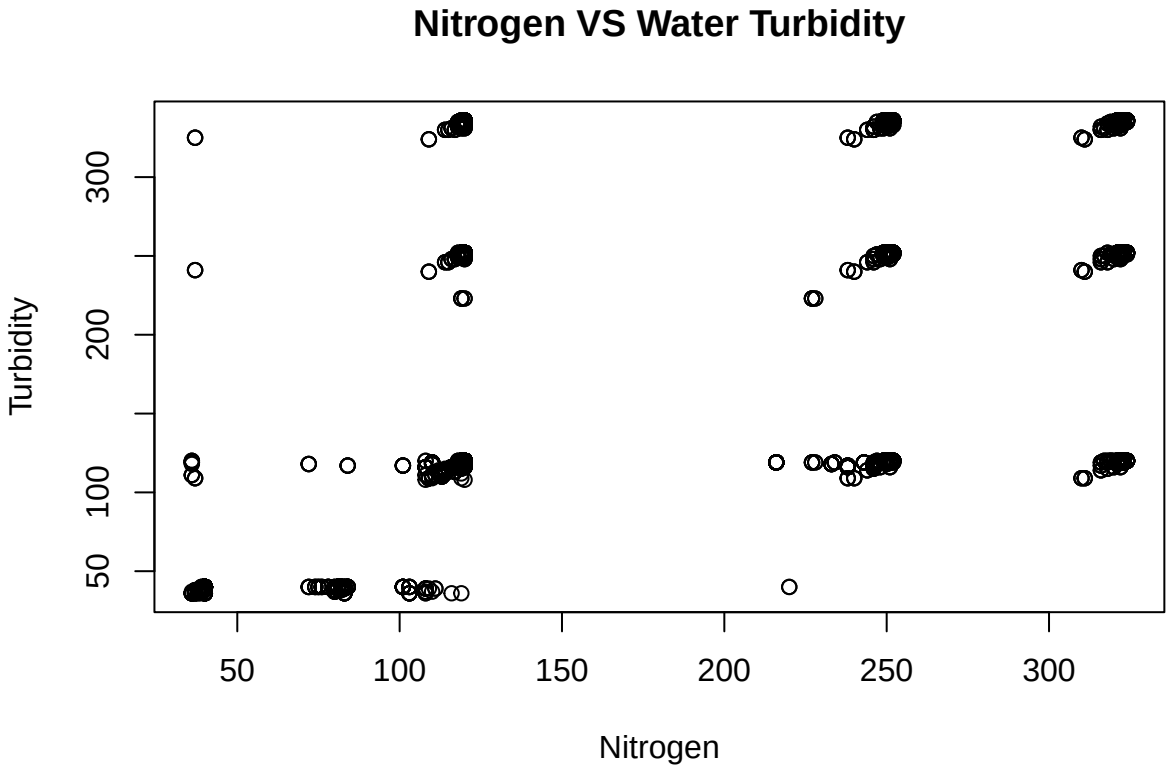


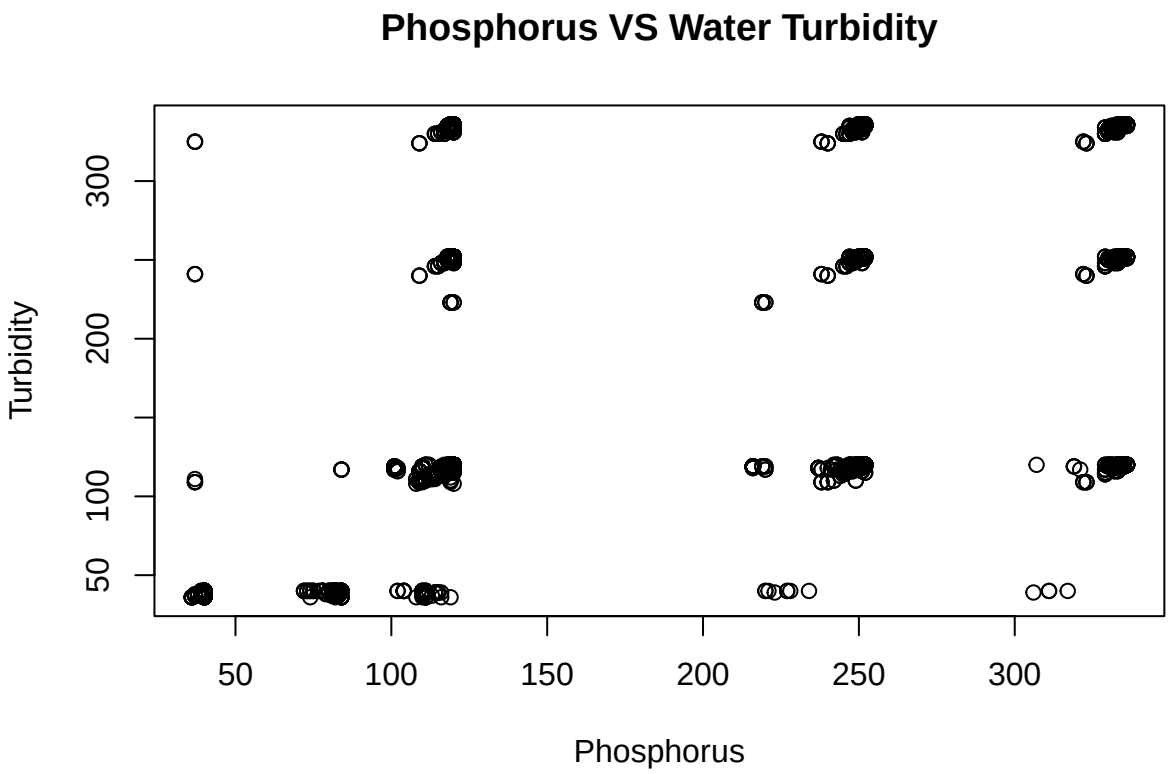
Correlation plot of the river quality indicators

The correlation plot above shows the correlations between each of the NZ river quality indicators. All pairs of variables produced positive correlations. The highest correlation is seen between nitrogen and turbidity (0.50).

Scatterplots of pair correlations

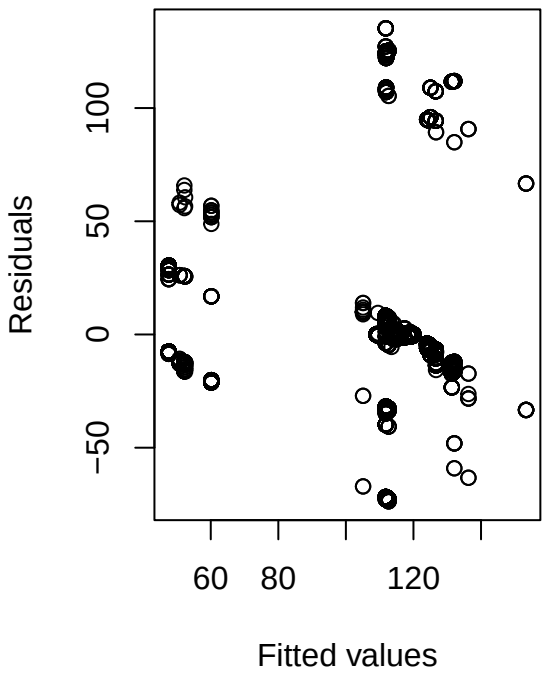


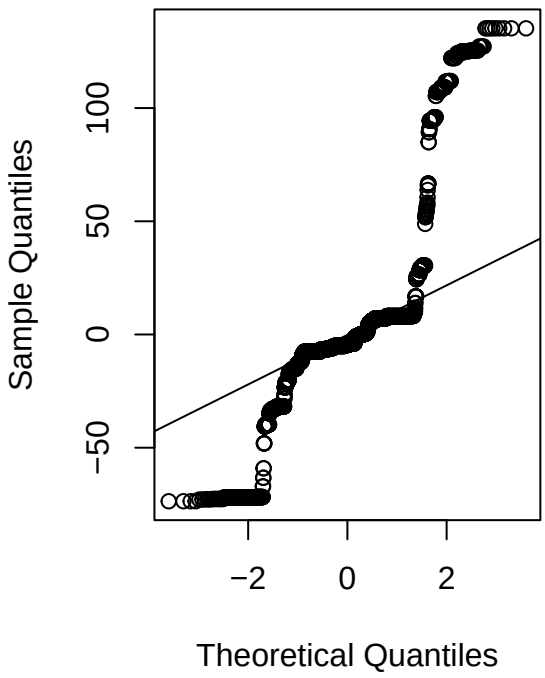




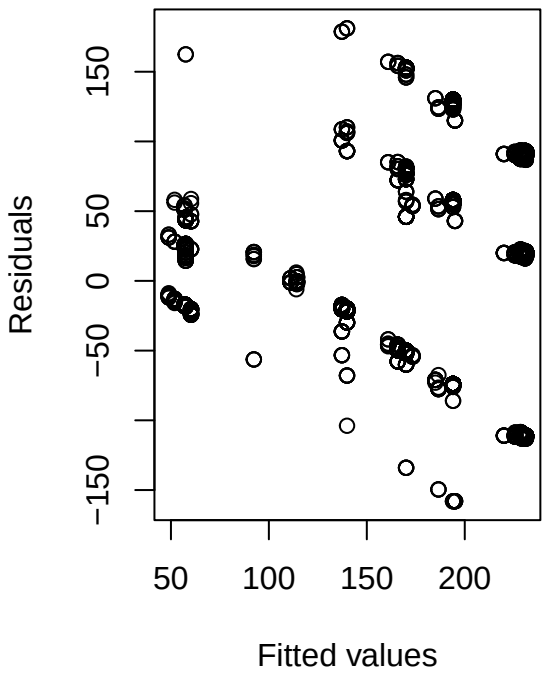
Normality and variance

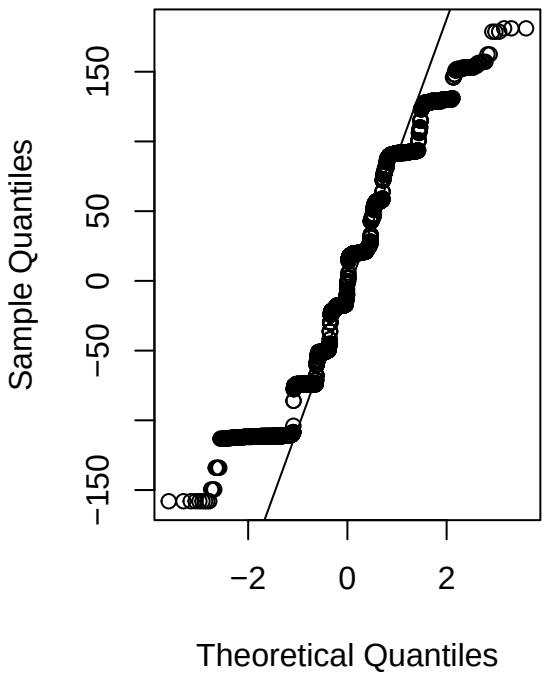
```
mod1 <- lm(wq1$E.Coli ~ as.factor(wq1$Turbidity), data = wq1)

mod1resid <- mod1$res
mod1fitted <- mod1$fit
par(mfrow=c(1,2))
plot(mod1resid~mod1fitted, xlab="Fitted values", ylab="Residuals")

qqnorm(mod1resid,main="")
qqline(mod1resid) #residual vs fitted: Bands not normally distributed about zero with differing 'height
```



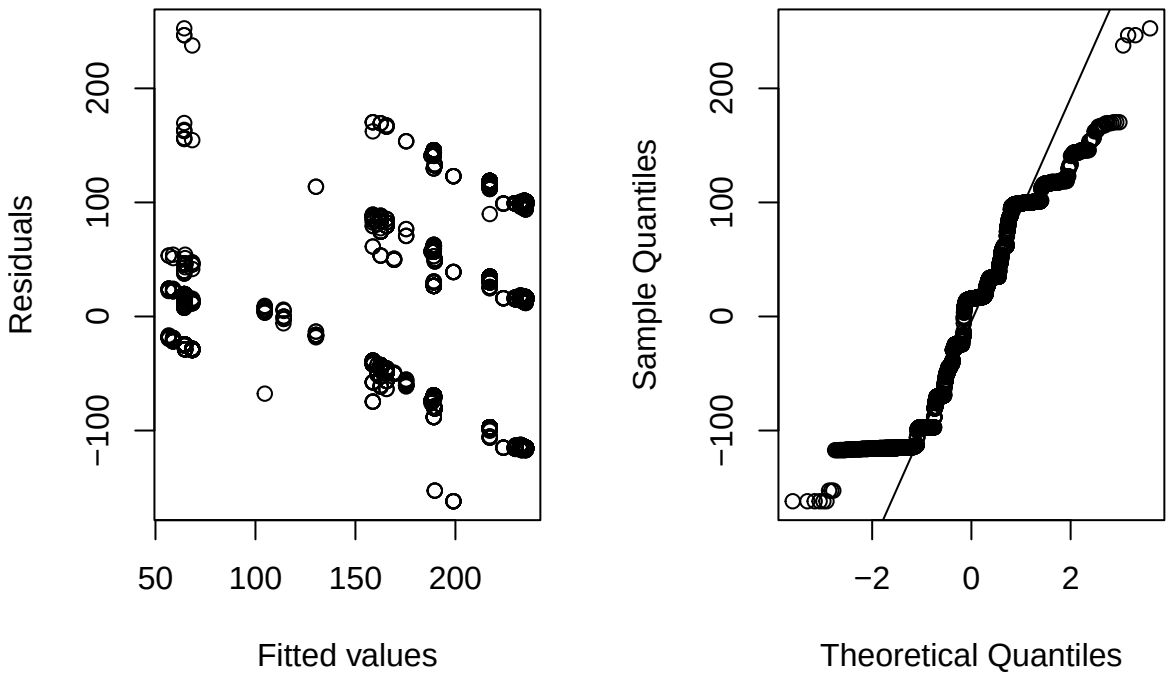



```
mod1 <- lm(wq1$Nitrogen ~ as.factor(wq1$Turbidity), data = wq1)

mod1resid <- mod1$res
mod1fitted <- mod1$fit
par(mfrow=c(1,2))
plot(mod1resid~mod1fitted, xlab="Fitted values", ylab="Residuals")

qqnorm(mod1resid,main="")
qqline(mod1resid) #residual vs fitted: Bands not normally distributed about zero with differing 'height
```



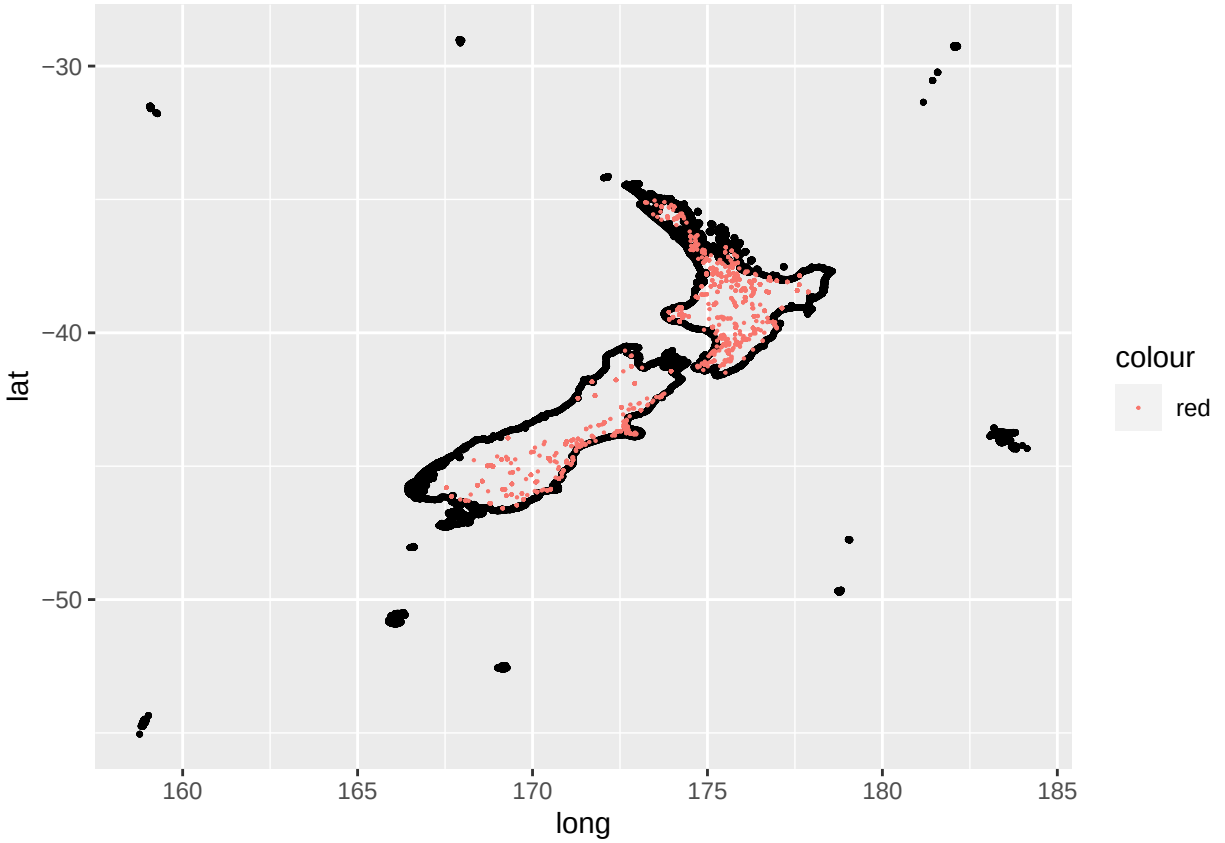



```r
mod1 <- lm(wq1$Phosphorus ~ as.factor(wq1$Turbidity), data = wq1)

mod1resid <- mod1$res
mod1fitted <- mod1$fit
par(mfrow=c(1,2))
plot(mod1resid~mod1fitted, xlab="Fitted values", ylab="Residuals")

qqnorm(mod1resid,main="")
qqline(mod1resid) #residual vs fitted: Bands not normally distributed about zero with differing 'height
```



The above plots show residuals vs fitted values and normal Q-Q plots between variables. For all pairs of variables; E.Coli and turbidity, nitrogen and turbidity and phosphorus and turbidity, bands are not normally distributed about zero, with differing 'heights'. This suggests non- constant variance. The Q-Q plots indicate that the data is not normally distributed, as it does not show a straight line.

Map of river sites monitored across NZ



The above map shows the coordinates (using variables of latitude and longitude) of river sites measured across New Zealand.