Assessment 2

Dataset: STAT8010_Assignment_2022

Brian Higgins

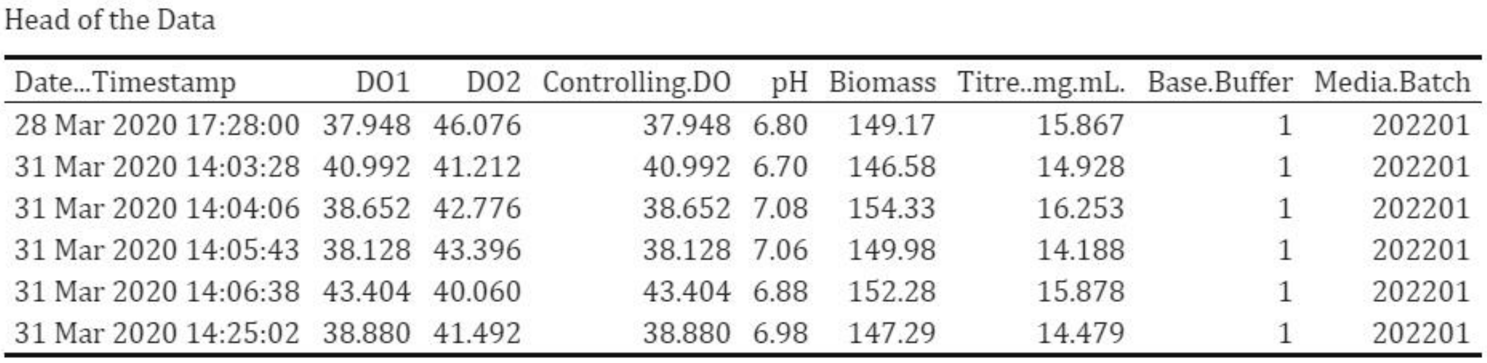
1. Introduction

1.1 Data Introduction.

The data set contains 670 entries over 9 columns, outlined below, for a period between 2020-03-28 and 2020-06-15. The columns are a mixture of Character and Numeric types. Below is a desciption for each column and unit it is in:

- **Date/Timestamp** – (Character) - Date and time of the reading.
- **D01/D02** – (Numeric) - Dissolved Oxygen readings from two probes.
- **Controlling DO** – (Numeric) - DO value being used to control the batch (note this will always be equal to one of DO1 or DO2)
- **pH – (Numeric)** - pH measured in the batch
- **Biomass (g/L)** – (Numeric) - Measured Biomass in the batch
- **Titre (mg/mL)** – (Numeric) - Concentration of product measured in mg/mL
- **Base Buffer** – (Numeric) - denotes which batch of base is used
- **Media Batch** – (Numeric) - denotes which batch of media is used

Table 1.1 shows the top 6 rows of the data, showing an example of what the data set looks like. The Date..Timestamp column is a date in Years, Months, Days, Hours, Minutes and Seconds. DO1, DO2, Controlling DO, pH, Biomass and Titre are in decimal numbers and Base Buffer and Media batch are in whole numbers.



Head of the Data

Date...Timestamp	DO1	DO2	Controlling.DO	pH	Biomass	Titre..mg.mL.	Base.Buffer	Media.Batch
28 Mar 2020 17:28:00	37.948	46.076	37.948	6.80	149.17	15.867	1	202201
31 Mar 2020 14:03:28	40.992	41.212	40.992	6.70	146.58	14.928	1	202201
31 Mar 2020 14:04:06	38.652	42.776	38.652	7.08	154.33	16.253	1	202201
31 Mar 2020 14:05:43	38.128	43.396	38.128	7.06	149.98	14.188	1	202201
31 Mar 2020 14:06:38	43.404	40.060	43.404	6.88	152.28	15.878	1	202201
31 Mar 2020 14:25:02	38.880	41.492	38.880	6.98	147.29	14.479	1	202201

Table 1.1: Head of Data

1.2 Report Outline

The report will look at how the Controlling Do dissolved Oxygen reading changes between the two DO probes available while also looking at Time Series to show the changes of pH, Titre, Controlling DO and Biomass over time. We will also look at the effect of the three batches on Titre and how the two Base Buffer groups affect Titre. Lastly we will do analysis on which of the column types has the most effect on the Titre and can be used to predict the value of Titre.

2. Exploratory Data Analysis

2.1 Missing / Duplicate Data

Investigating the data, there was no Missing Values in any of the columns as we can see below. The Date...Timestamp had no duplicates so we know we have unique entries. We are not concerned with duplicates in DO1, DO2, Controlling.DO, pH, Biomass or Titre as we would expect those entries to have some duplicated given the type of data recorded. Media Batch and Base Buffer had a large amount of duplicates as these columns are categorically columns but were originally saved as numeric columns.

```
Missing values in Date...Timestamp : 0
Missing values in DO1 : 0
Missing values in DO2 : 0
Missing values in Controlling.DO : 0
Missing values in pH : 0
Missing values in Biomass : 0
Missing values in Titre..mg.mL. : 0
Missing values in Base.Buffer : 0
Missing values in Media.Batch : 0
```

```
Duplicate values in Date...Timestamp : 0
Duplicate values in DO1 : 68
Duplicate values in DO2 : 95
Duplicate values in Controlling.DO : 82
Duplicate values in pH : 649
Duplicate values in Biomass : 109
Duplicate values in Titre..mg.mL. : 37
Duplicate values in Base.Buffer : 668
Duplicate values in Media.Batch : 667
```

2.2 First look at columns

Below in Fig 2.2, we see a quick look at the columns that shows for DO1 and DO2, the Controlling DO is strongly related in parts, which you can see with the solid straight lines at “A) Controlling Do”. Other parts of the DO1 and DO2 are not related to the Controlling.DO column so we can see other data points that are not within the lines. We will look at this in more detail later.

At “B) Titre/Biomass” you can also see that the two columns have some relationship to each other, we will look at this more later in the report as well. “C) Categorical” shows that the data is categorical as we can see with the clear lines and grouped data. When we investigate the data we see that Media Batch has three unique batches “2022201”, “202202” and “202203”. Base Buffer has two base batches, “1” and “2”.

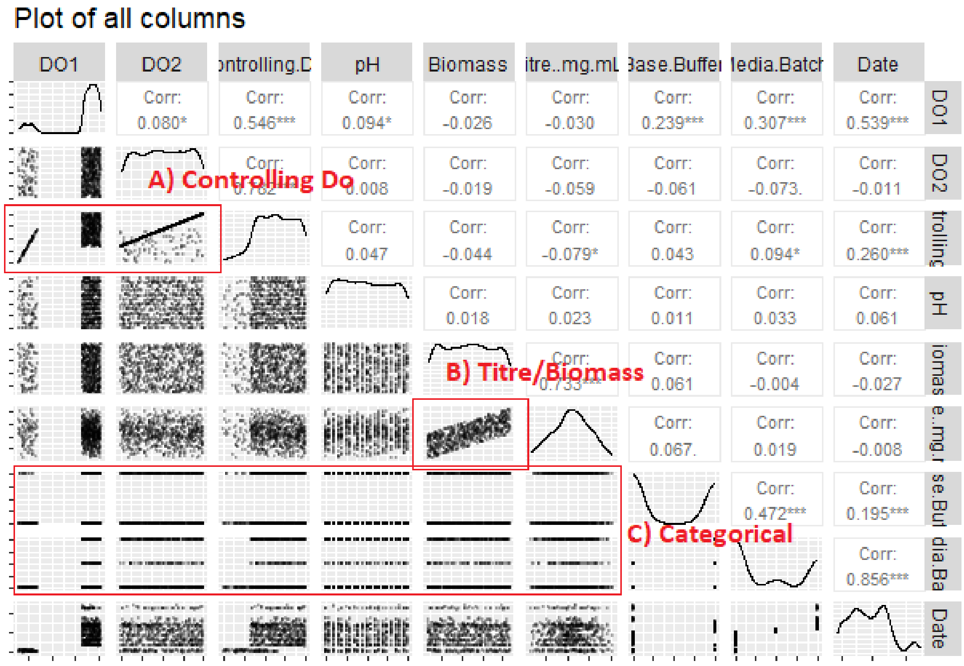


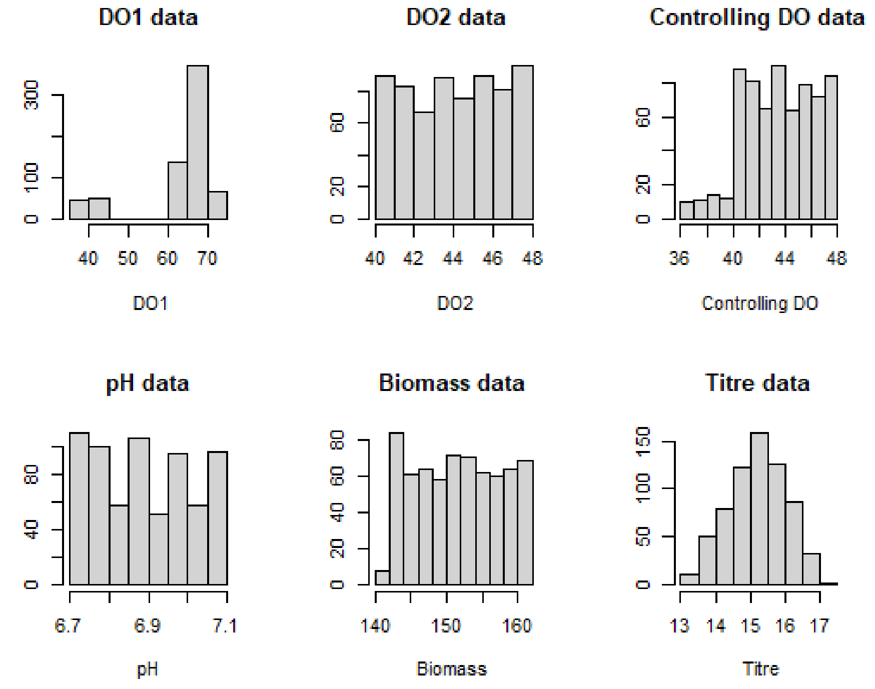
Fig 2.2: Quick EDA of columns

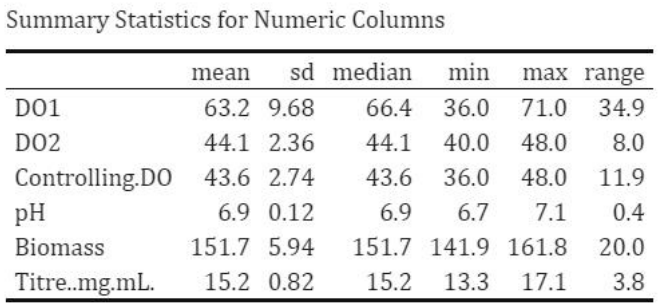
2.3 Distribution of Numeric Columns

In Fig 2.3 we can see the distribution of the numeric columns. The top row contains DO1, DO2 and Controlling DO which are all related. We can see that DO1 has low values and then a jumps to higher values, DO2 is fairly uniform and the Controlling DO has low values and then Jumps up for around the 40 value. We know from the data that the Controlling DO is equal to either DO1 or DO2 so we can see a point where it changes from DO1 to DO2.

Looking in more detail into the DO columns with Table 2.3. D01 has an min of 36 but has a max of 71, DO2 has a min of 40 and a much lower max of 48. We will see in the next section in more detail how when DO1 increase the controlling DO changes between DO1 and DO2.

pH has a very small range of values from 6.9 to 7.1, it looks mostly uniform. Biomass, as we can see from the plot has a low initial value but then increases to stay uniform for the rest of the data points. Lastly we have Titre which is normally distributed from the looks of the plot.





Summary Statistics for Numeric Columns

	mean	sd	median	min	max	range
DO1	63.2	9.68	66.4	36.0	71.0	34.9
DO2	44.1	2.36	44.1	40.0	48.0	8.0
Controlling.DO	43.6	2.74	43.6	36.0	48.0	11.9
pH	6.9	0.12	6.9	6.7	7.1	0.4
Biomass	151.7	5.94	151.7	141.9	161.8	20.0
Titre..mg.mL.	15.2	0.82	15.2	13.3	17.1	3.8

Table 2.3

Fig 2.3 Distribution of numeric columns

2.4. Data Manipulation

Before we do anymore Exploratory Data Analysis we need to fix two things:

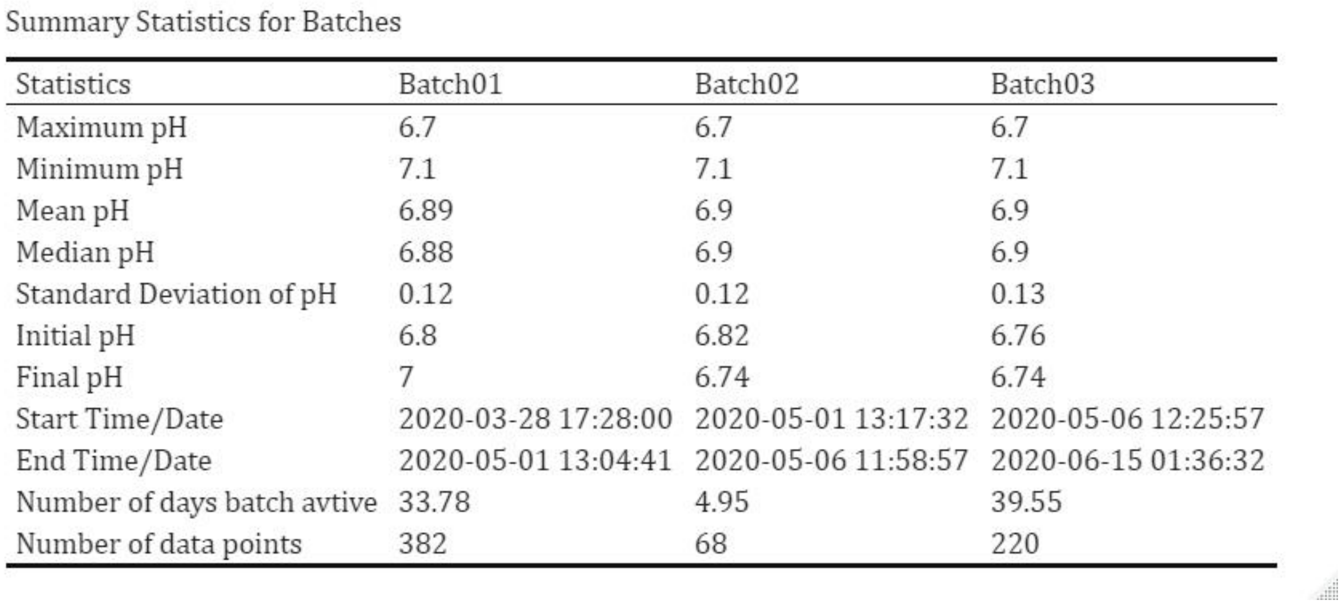
1) The Date...Timestamp column's first entry was in the wrong time format and this was changed to match the other rows. After that the Date...Timestamp was changed from a character variable type to a Date Type so the Date could be used in Time Series analysis.
2) As we saw both Media.Batch and Base.Buffer are categorical so we have changed them into Factor data types, which lets us group the different types for later use.

2.5 Summary Statistics for the Media Batches

Table 2.5 shows a summary of the three batches for the data. Across all three batches minimum(6.7) and maximum (7.1) pH do not change. There is slight difference in the mean (6.89) and median (6.88) pH for batch one and batches two and three, mean (6.9) and median (6.9). Standard deviation, is also very similar across all three batches at 0.12 and 0.13.

The initial pH for batch 01 is 6.8 and its final pH is 7. Batch 02 has an initial pH of 6.82 and a final pH of 6.74. Batch 03 has an initial pH of 6.76 and a final pH of 6.74. Batch 01 has an increase from the initial pH reading to the final pH reading, whereas Batch 02 and batch 03 have smaller decreases. We will look at this in more detail later at how the media changes per batch for Titre.

The most notable difference in the three batches is the length of time and the number of data points in each. Batch 02 had the shortest batch at just under 5 days and 68 data points. Batch 01 had the a batch length of less than 34 days and 382 data points and batch 03 had the longest length at just under 40 days but only 220 data points. We would expect batch three to have the most data points, but this is not the case. Looking into this we can see that batch three has no data from the 24th of May 2020 to the 6th of June, so this accounts for the lower number of data points with a 13 day gap for data points. Later we will see if this affects the results. We will see this more clearly later in the plots in the next section.



Summary Statistics for Batches

Statistics	Batch01	Batch02	Batch03
Maximum pH	6.7	6.7	6.7
Minimum pH	7.1	7.1	7.1
Mean pH	6.89	6.9	6.9
Median pH	6.88	6.9	6.9
Standard Deviation of pH	0.12	0.12	0.13
Initial pH	6.8	6.82	6.76
Final pH	7	6.74	6.74
Start Time/Date	2020-03-28 17:28:00	2020-05-01 13:17:32	2020-05-06 12:25:57
End Time/Date	2020-05-01 13:04:41	2020-05-06 11:58:57	2020-06-15 01:36:32
Number of days batch avtive	33.78	4.95	39.55
Number of data points	382	68	220

Fig 2.5: Summary of the Batches

2.6 Controlling Do with DO1 and DO2

As we saw in section 2.3 while exploring the data we could see that the Controlling DO had a strong correlation or relationship to DO1 and DO2 for different points as showed by the strong lines. Fig 2.6.1 shows a plot where you can see in red when the Controlling DO was equal to DO1 and in green when it switched to DO2. The Controlling.DO only matched DO1 for 95 entries and then on the 03rd of Apr 2020 at 15:17:51 it switched to DO2. You can see the blue dashed lines that show this change. The Controlling DO matched DO2 for the rest of the data for 575 entries. So There was only one change between DO1 and DO2.

The blue line also shows that there is a large gap in the data between the 03rd of Apr 2020 at 15:17:51 and the 07th of Apr 2020 at 11:09:04 for almost 4 days. If we look at the mean with a blue line we can see the mean has a jump at the start in DO1 but after that goes flat as the data is missing. When DO2 starts it climbs slowly and then levels off (but there is another data gap).

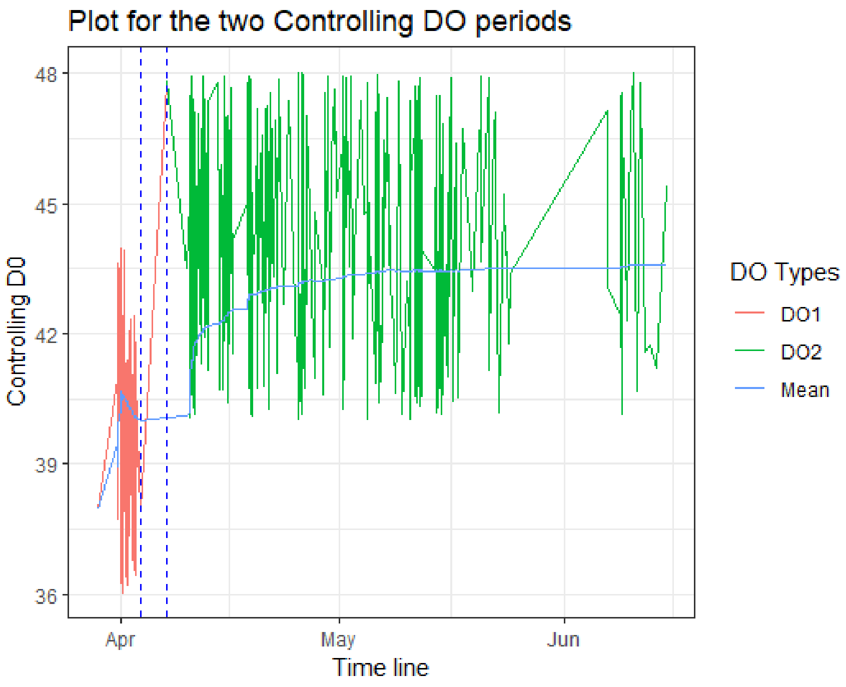


Fig 2.6.1 Plot for the two controlling DO periods

What is not clear from Fig 2.6.1 is what DO1 is doing after the Controlling.DO switched between DO1 and DO2. In Fig 2.6.2, once again we can see DO1 in red and DO2 in green with a blue line showing the Controlling.DO which matches the line plot in Fig 2.6.1. What we can see is the blue Controlling.DO line matches DO1 while its below DO2, once DO1 starts to rise above DO2, the blue controlling.DO line switches to DO2. The Controlling.DO line stays on the DO2 line for the rest of the data points and does not switch back to DO1 while it remains high.

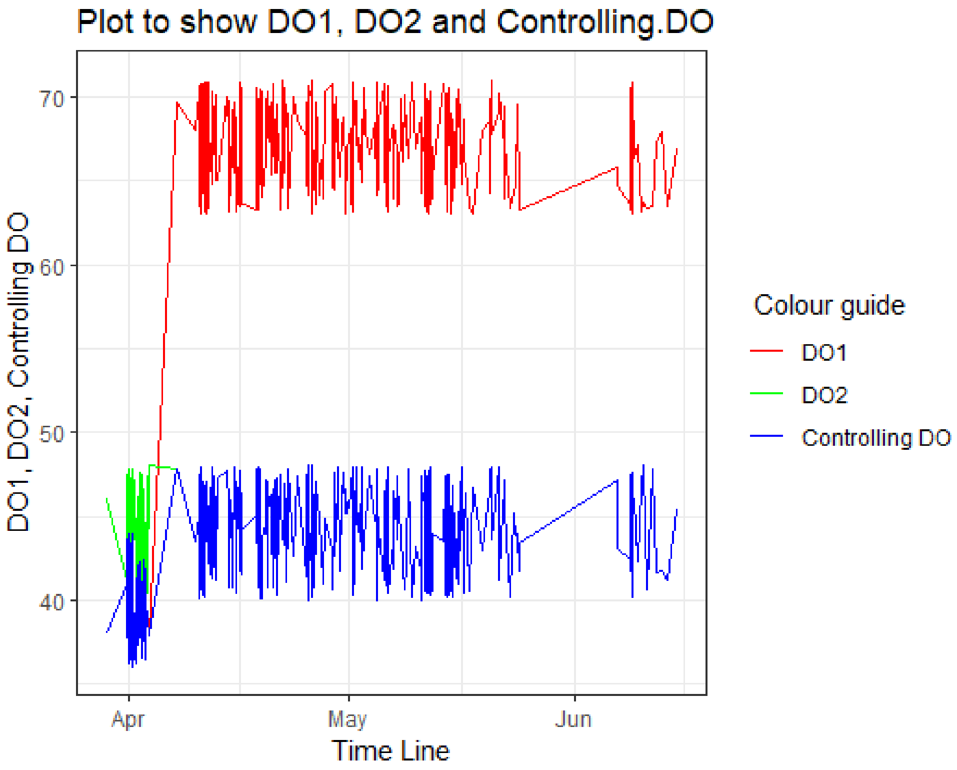


Fig 2.6.2 Plot to show when Controlling Do changes

3. Main Findings

3.1 Time Series Visualization of data

Fig 3.1 shows a time series with pH, Titre, Controlling DO and Biomass on it. You can see three colours used per line plot to show each of the three Media batch types. Batch 01 is in red, Batch 02 in Green and Batch 03 in Blue. As seen above and shown on the line plots, Batch 01 has the most data points (382), followed by a small amount for Batch two (68) in green and then blue (220) for Batch 03. Between the two red lines we can see the missing data points in the third batch for the gap between the 24th of May 2020 to the 6th of June.

This plot is good to see all four variable plotted together so we can see that pH has the lowest values followed by Titre, Controlling.DO and Biomass. We can still see the jump in values when the Controlling.DO switches between DO1 and DO2 but there is no corresponding jump for the other data. However, in this plot the pH values look quite flat. We have seen in the summary table above there is some variation in the results for pH so we will look at each of the variables separately.

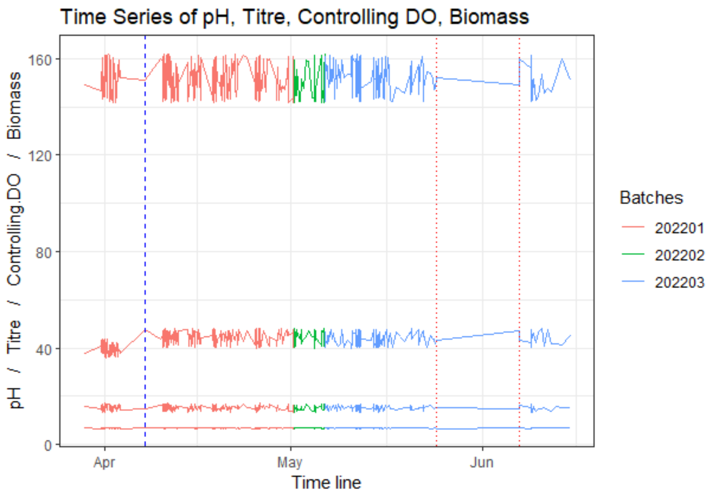
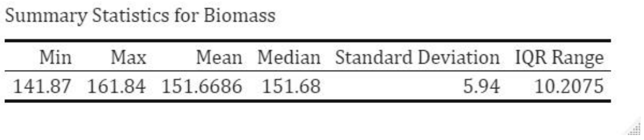


Fig 3.1 Time Series of pH, Titre, Controlling DO, Biomass

Below is a separate plot of each of the four columns individually.

3.1.1 Time Series of Biomass



Summary Statistics for Biomass

Min	Max	Mean	Median	Standard Deviation	IQR Range
141.87	161.84	151.6686	151.68	5.94	10.2075

Table 3.1.1 Biomass

Fig 3.1.1 shows the three batches for Biomass. From table 3.1.1,we can see that the minimum value is 141.87, the maximum value is 161.84 and the mean is 151.6686. The mean is also shown on the graph in purple which looks very stable except for a jump in DO1.

Biomass is similar for Batch 01 and Batch 02 but we can see that the third batch has values that has bigger gaps, more lower and higher values in parts and a big missing data section in red lines.

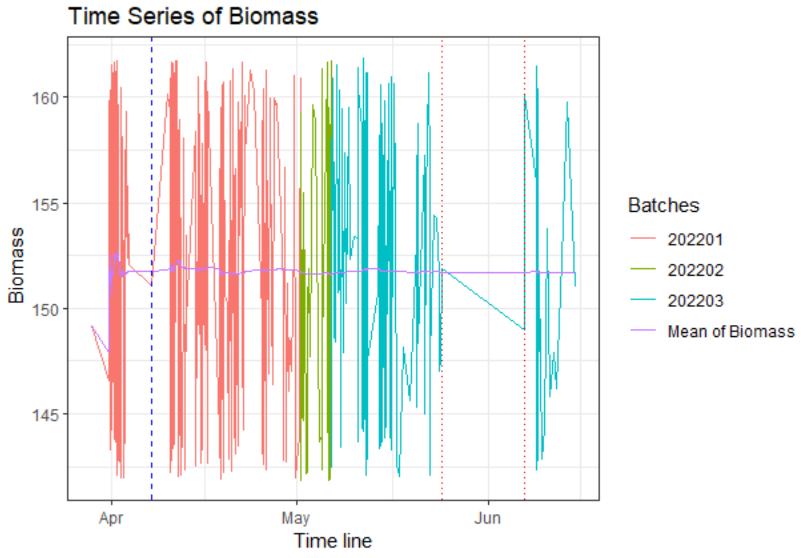


Fig 3.1.1 Time Series of Biomass

3.1.2 Time Series of Controlling DO

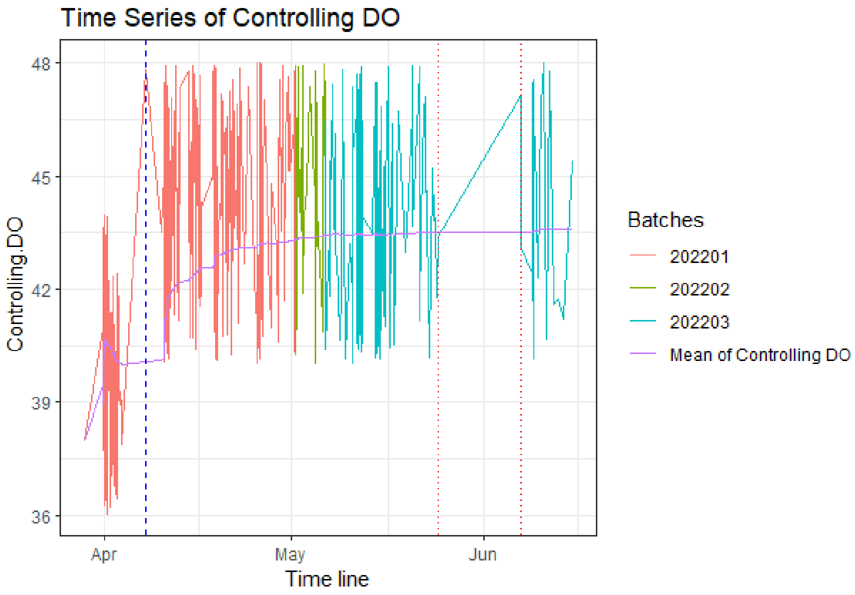
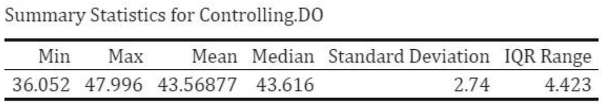


Fig 3.1.2 Time Series of Controlling DO



Summary Statistics for Controlling.DO

Min	Max	Mean	Median	Standard Deviation	IQR Range
36.052	47.996	43.56877	43.616	2.74	4.423

Table 3.1.2 Controlling DO

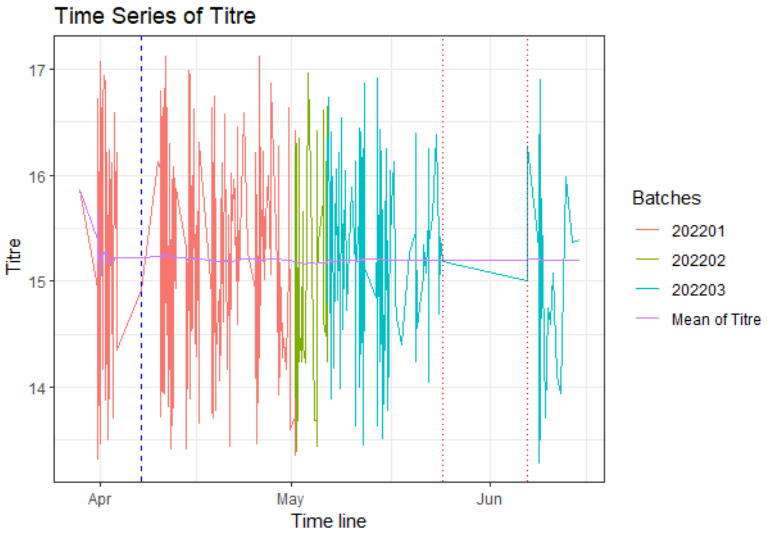
Fig 3.1.2 shows the three batches for the controlling DO. From table 3.1.2 we can see that the minimum value is 36.052, the maximum value is 47.996 and the mean is 43.56.

The mean is shown in purple on the graph which shows a large jump from DO1 to when DO2 becomes the controlling DO.

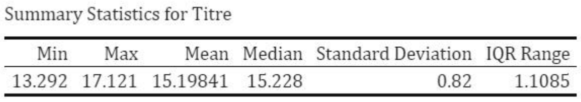
The change in controlling DO has been explored more in section 2.6 but what is clear is the mean value jumps between DO1 and DO2.

Again you can see the missing data in batch 3 between the red lines.

3.1.3 Time Series of Titre



3.1.3 Time Series of Titre



Summary Statistics for Titre

Min	Max	Mean	Median	Standard Deviation	IQR Range
13.292	17.121	15.19841	15.228	0.82	1.1085

Table 3.1.3 Titre

Fig 3.1.3 shows the three batches for the Titre. From table 3.1.3 we can see that the minimum value is 13.292, the maximum value is 17.121 and the mean is 15.19841. The mean is shown in purple on the graph.

Batch 03 looks less stable in batch 03. Its clear from the blue batch line that the values, ranges are lower and sections where there are bigger dips. What is interesting is even with these dips the mean stays very similar except for once again a small jump in the beginning.

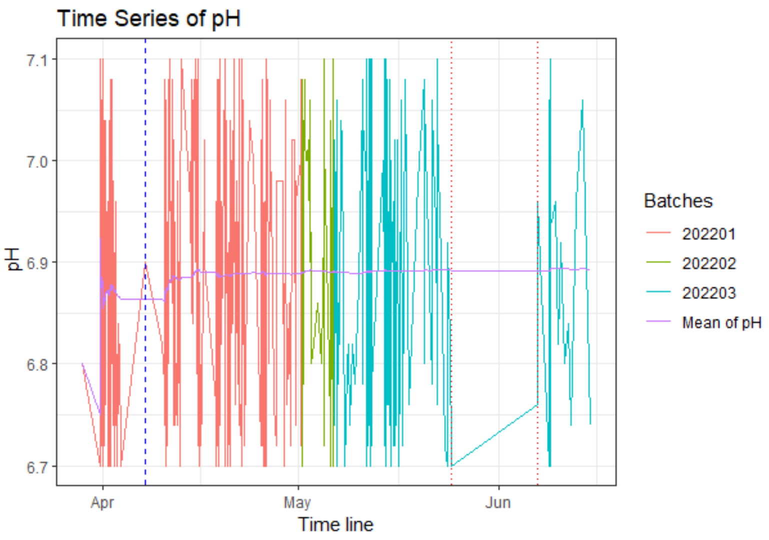
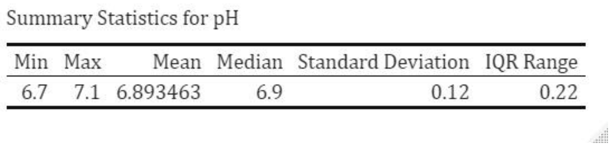


Fig 3.1.4 Time Series of pH

3.1.4 Time Series of pH



Summary Statistics for pH

Min	Max	Mean	Median	Standard Deviation	IQR Range
6.7	7.1	6.893463	6.9	0.12	0.22

Table 3.1.4 pH

Fig 3.1.4 shows the three batches for the pH. From table 3.1.4 we can see that the minimum value is 6.7, the maximum value is 7.1 and the mean is 6.89. The mean is shown in purple on the graph again where we can see a slight increase for DO1, a decrease when the data is missing and a an increase when DO2 becomes the controlling DO where after that it stays fairly flat.

Like previous Time series there is missing data and different ranges for batch 3 than compared to the other batches.

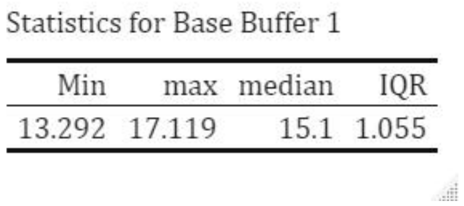
A main observation of the time series graphs above is that data is more condensed for batch 01, especially when DO1 is the controlling DO and batch 03 is missing a large amount of data. Batch 03 also has a lot less uniform data and more sections where values stay lower than Batch 1 and 2, this could be because of the missing data.

3.2 Visualizations of Base Buffer and Media Batch

There are two categorical columns; Media Batch which contains three batches and Base Buffer that contains two base batches. Next we will look at how these different Base Batches and Media Batches affect Titre.

3.2.1 Titre with Base Buffer

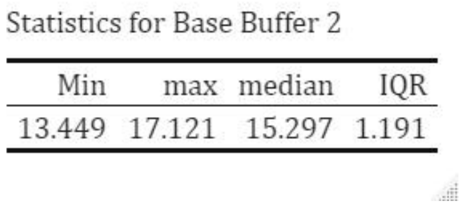
Fig 3.2.1 shows us a box plot of the Titre by each of the Two Base Buffer batches. First we can see from the red dots that the data points look evenly spread out. There are 367 data points in Base Buffer Batch 1 and 303 in Base Buffer Batch 2 so they are fairly evenly distributed. The skewness test for Batch 1 is -0.19 and for Batch 2 is 0.19 so both of then are symmetric with a very slight left and right skew, but not significant. Batch 2 has a larger range and a higher median so this has the most affect on Titre. We can see this clearer in the two tables below. You can see all the values, min, max, median and IQR are all larger.



Statistics for Base Buffer 1

Min	max	median	IQR
13.292	17.119	15.1	1.055

Table 3.2.1 Base Buffer 1



Statistics for Base Buffer 2

Min	max	median	IQR
13.449	17.121	15.297	1.191

Table 3.2.2 Base Buffer 2

In future reports we will look into how the mean is affected for Titre for each group by performing a T Tests.

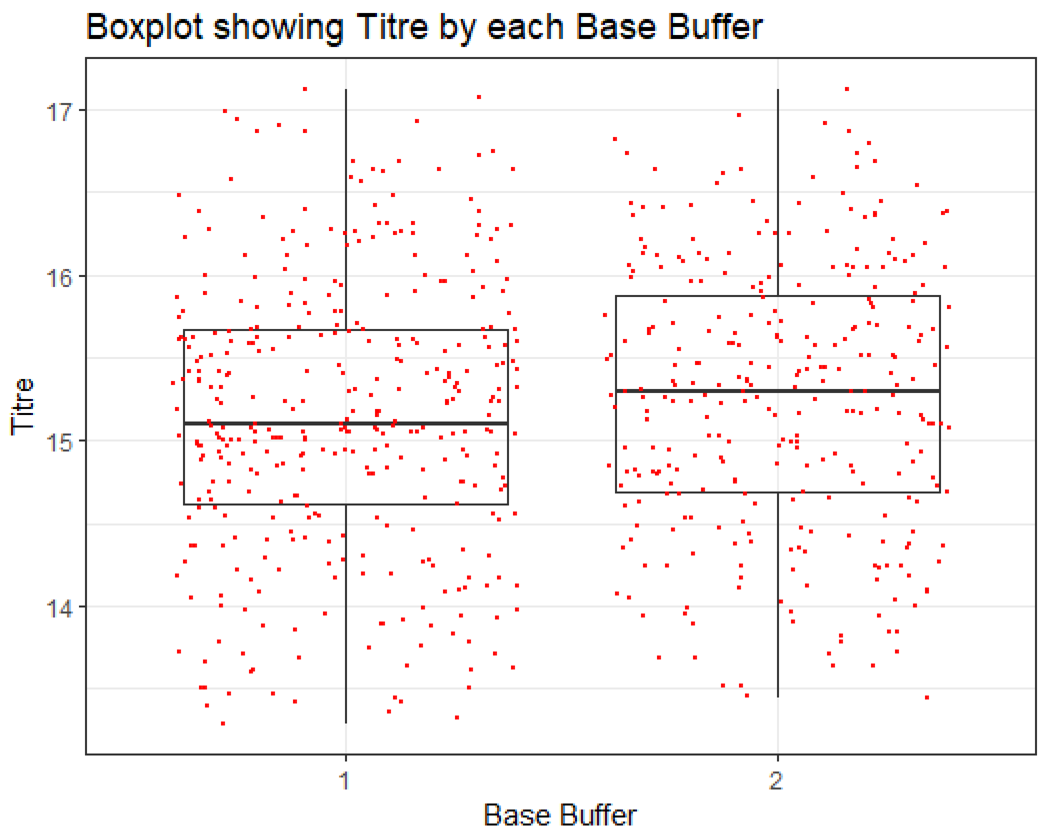


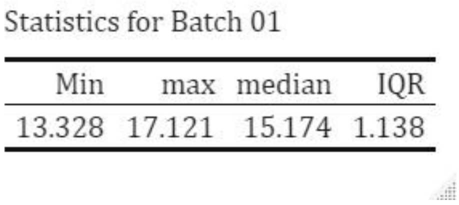
Fig 3.2.1 Boxplot showing Titre by each Base Buffer

3.2.2 Titre with each batch number

In Fig 3.2.2 we can see the three batches and the distribution of the Titre data points in blue dots. As mentioned previously there are 382 data points in batch 01, 68 in batch 02 and 220 in batch 03. Tables 2.4.3, 2.4.4 and 2.4.5 below show the values for Titre across all three batches.

Batch 01 looks evenly distributed and this is confirmed with a skewness value of -0.28 so its symmetric, batch 03 is similar with a value of 0.73 so they are very slightly left and right skewed. Batch 02 on the other hand has a higher left skew and you can see this in the boxplot as the median line is higher towards the top of the box and the skew value is 2.6.

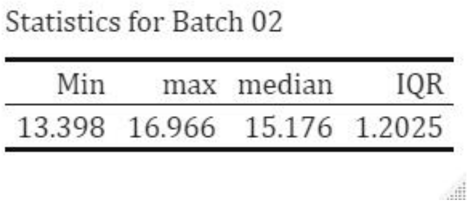
Batch 01 has data points that are evenly distributed around the median, whereas batch 02 with the least points has the biggest range. Batch 03 has the highest median. It appears that the more data points the data comes more evenly spread out.



Statistics for Batch 01

Min	max	median	IQR
13.328	17.121	15.174	1.138

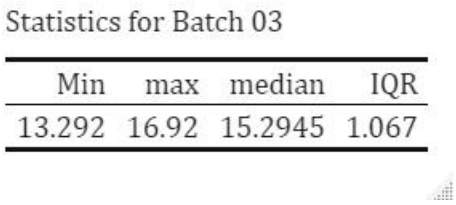
Table 3.2.2 Media Batch 01



Statistics for Batch 02

Min	max	median	IQR
13.398	16.966	15.176	1.2025

Table 2.2.3 Media Batch 02



Statistics for Batch 03

Min	max	median	IQR
13.292	16.92	15.2945	1.067

Table 2.3.4 Media Batch 03

In future reports we will look into how the mean is affected for all three groups with Titre by performing a Fishers F test.

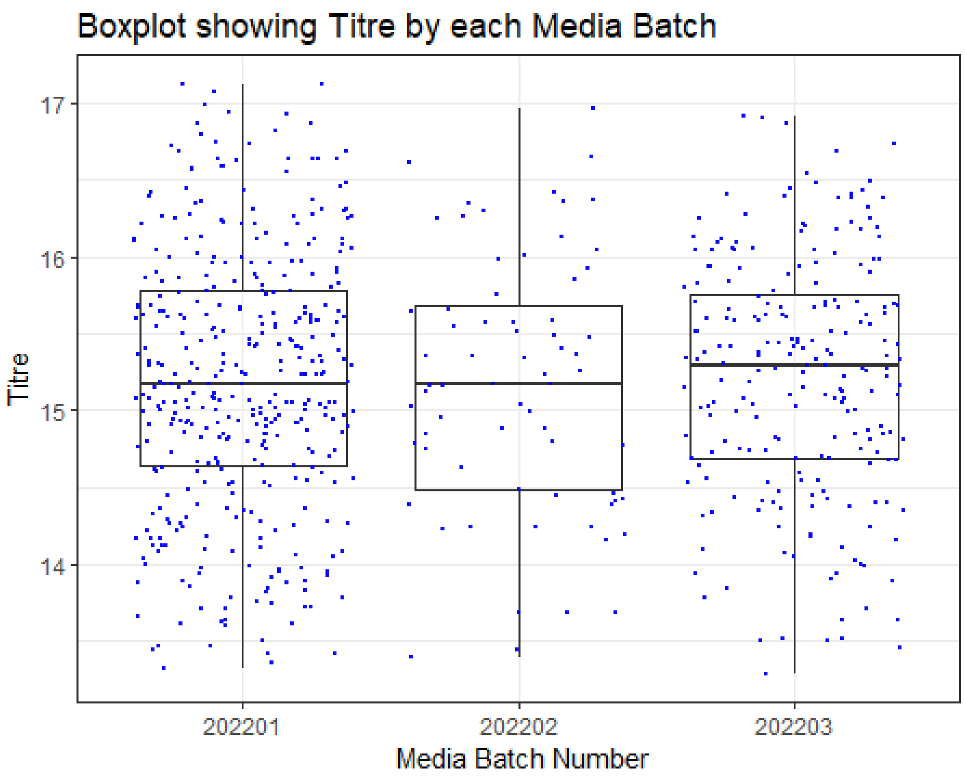


Fig 3.2.2 Boxplot showing Titre be each Media Batch

3.3 Linear Regression Model

Lastly we will look at the correlation of each of the numeric columns on Titre to see if any of them can be used to predict the Titre values.

3.3.1 Visualise Correlations

Fig 3.3.1 is similar to what we have seen before in the introduction, this time we are just looking at the numeric columns and how they affect the Titre column. The most significant column we can see from the plot is Biomass which has a value of 0.733 and three stars so we should expect this to have an effect on Titre. We will look at this in more detail once we run a linear regression model.

The other columns do not look like they have a relationship with Titre however the Controlling DO gives a low negative value but also includes one star which is interesting. We will look at this in more detail in the model. Perhaps there is something there.

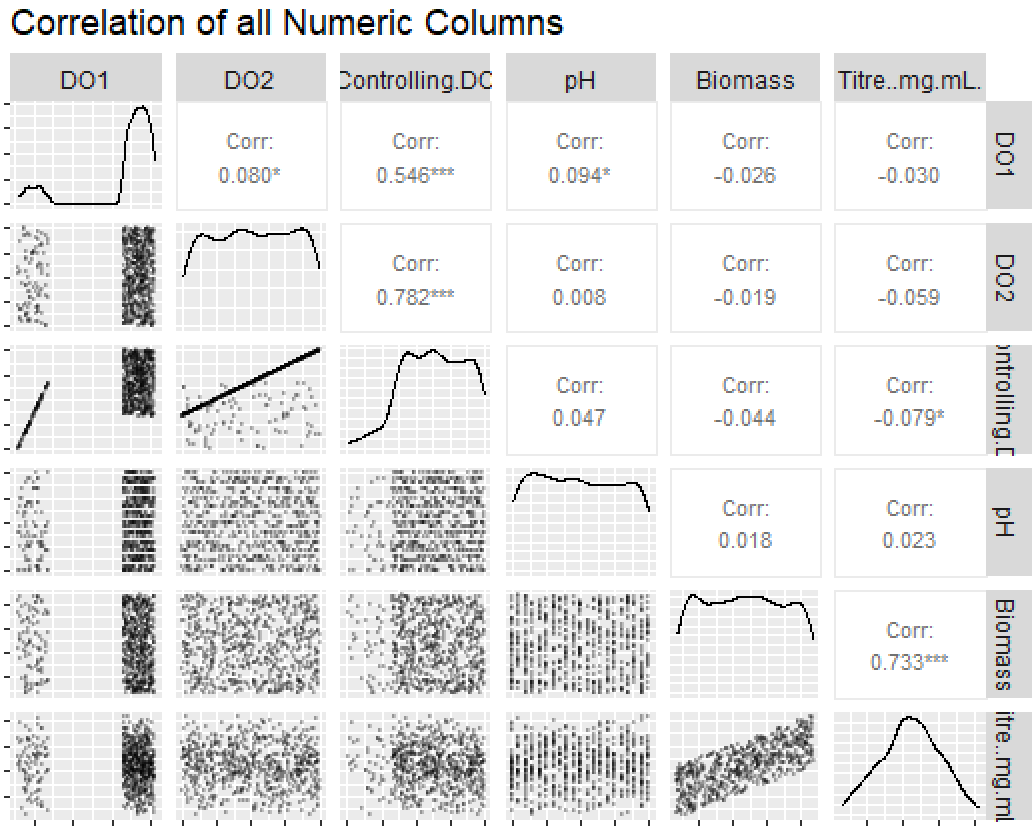
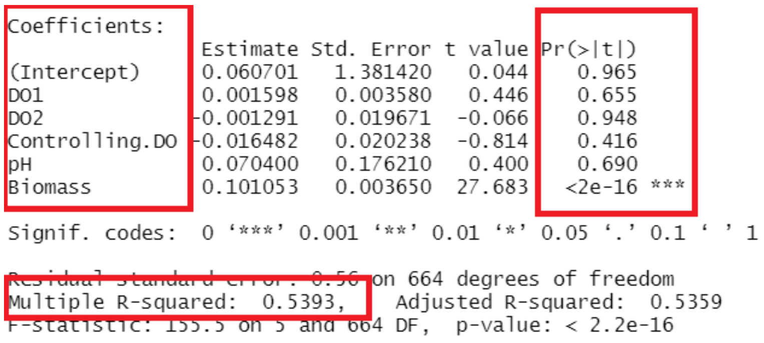


Fig 3.3.1 Correlation of all Numeric Columns.

3.3.2 Linear Regression Model

We will run a linear regression model in a Backward Stepwise to first take all the numeric columns and run a linear regression model on the numeric columns

First Linear Model: All columns



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.060701   1.381420   0.044    0.965
DO1             0.001598   0.003580   0.446    0.655
DO2            -0.001291   0.019671  -0.066    0.948
Controlling.DO -0.016482   0.020238  -0.814    0.416
pH              0.070400   0.176210   0.400    0.690
Biomass         0.101053   0.003650  27.683   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.56 on 664 degrees of freedom
Multiple R-squared:  0.5393,	Adjusted R-squared:  0.5359
F-statistic: 155.5 on 5 and 664 DF,  p-value: < 2.2e-16
```

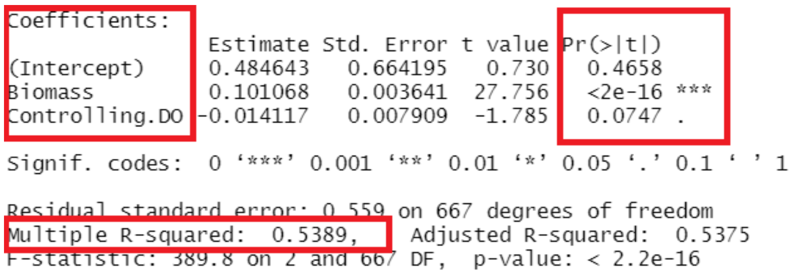
For the first model, we will use all the Numeric columns shown in the left red box.

We get values in the right box that show us a P-Value as a probability that the column is having an effect on Titre. The first four, DO1, DO2, Controlling DO and pH all have large values which shows they are not significant. Biomass has a very small number so this is showing a strong relationship.

Note, this time we do not see a star next to the Controlling DO like in the plot above.

The result for this model is 0.5393.

Second Linear Model: Remove columns



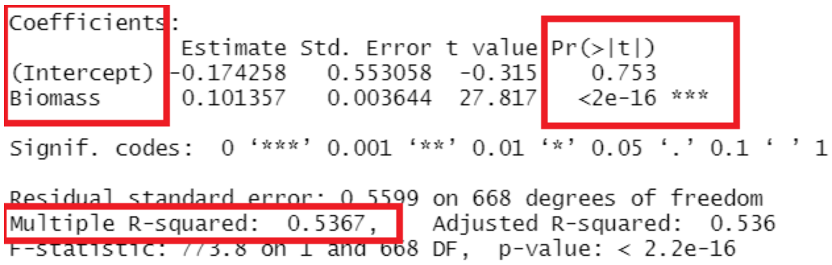
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.484643   0.664195   0.730   0.4658
Biomass        0.101068   0.003641  27.756   <2e-16 ***
Controlling.DO -0.014117   0.007909  -1.785   0.0747 .
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.559 on 667 degrees of freedom
Multiple R-squared:  0.5389,	Adjusted R-squared:  0.5375
F-statistic: 389.8 on 2 and 667 DF,  p-value: < 2.2e-16
```

For the second model we will remove DO1, DO2 and pH. I will leave in Biomass which shows a strong relationship. I am also interested in Controlling.DO seeing if there is a difference because of what was shown on the correlation plot above.

The results for this model is 0.5389. There is a very small drop but the Model is a lot simpler if we drop the three other columns and so is a positive step forward.

The p-Value for Controlling DO has also dropped a lot but is still well above our significant P-value of 0.05. We will run the model again without it to see how the model performs.

Third Linear Model: Only include Biomass



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.174258   0.553058  -0.315    0.753
Biomass      0.101357   0.003644  27.817   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 668 degrees of freedom
Multiple R-squared:  0.5367,	Adjusted R-squared:  0.536
F-statistic: 773.8 on 1 and 668 DF,  p-value: < 2.2e-16
```

For the last model we will only include Biomass. Once again the P-Value is very low.

The results for this model is 0.5367. Again there is a very small drop but we have once again removed a column so the best model has just the Biomass used to predict Titre.

This shows that we can use Biomass to predict Titre. Next we will plot these two columns separately to see the relationship.

3.3.3 Plot Biomass and Titre.

We now see from our models that Biomass is the best predicator for Titre. Lastly we will look at Titre and Biomass plotted together. Previously we have seen in a previous section that there is visual relationship between the two and in section 3.3.1 and we have showed a Linear Regression model that shows we can use Biomass to predict Titre in section 3.3.2. In 3.3.3 we plot both columns with each other and map the Linear Regression Model on to the plot so we see how we can use Biomass to predict values for Titre with this model. The blue line is our regression model we can use to predict Titre with Biomass.

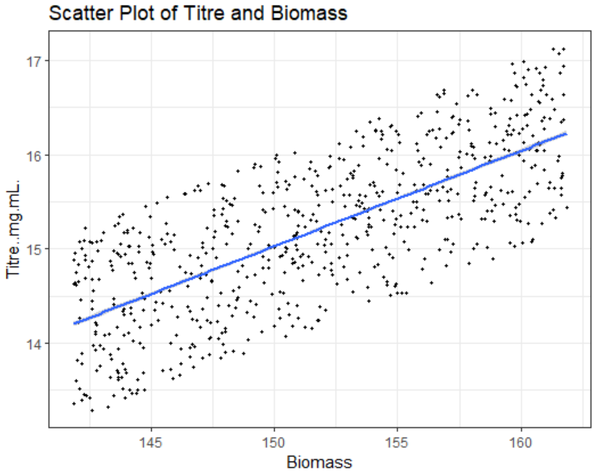


Fig 3.3.3 Scatter plot of Titre and Biomass

4. Conclusions

Conclusions are hard as I am not an expect in this type of data. However we can draw some conclusions

- When the DO1 value when above the DO2 value, the controlling DO switched from DO1 to DO2. This switch happened only once and after that DO1's values started above DO2.
- There is missing data points for 4 days between the 03rd of Apr 2020 at 15:17:51 and the 07th of Apr 2020 which resulted in an increase in the value for DO1 which caused a switch in the Controlling DO from DO1 to DO2.
- In the third batch there is 13 days period between 24th of May 2020 to the 6th of June which is also missing which results in Batch 03 having the lowest minimum and maximum Titre values. We would need to have this data to be sure.
- Base Buffer batch 1 had the lowest values for Titre over batch 2.
- Mostly importantly we can use a Linear Regression model to predict values for Titre using Biomass. No other column can be used as a predicator for Titre.
- In a later report we will use T-Tests and F-Tests to look into the relationships of Base Buffer and Media Batch on Titre.

5. Appendix

4.1 Cumulative Count in Minutes per DO

Plot Fig 4.1 shows a count in Minutes for each DO. We can see at the blue line it starts again when DO2 becomes the controlling Do.

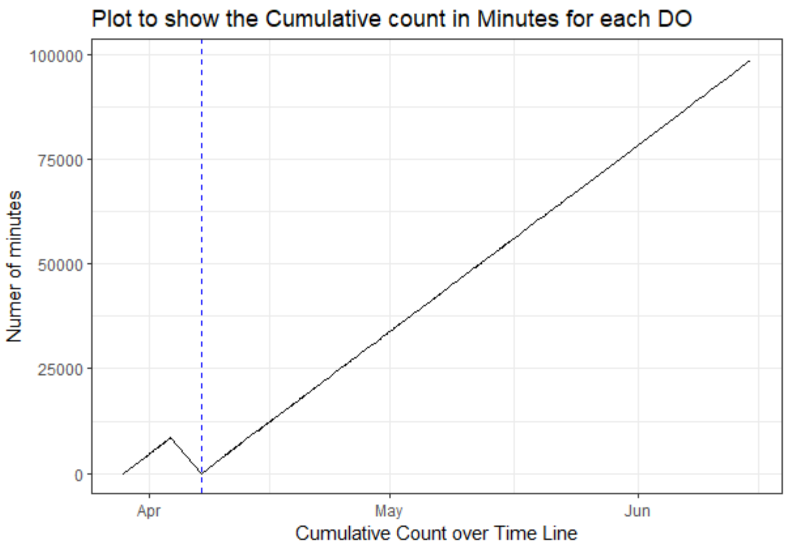


Fig 4.1 : Cumulative Count in Minutes per DO

4.2 Heatmap to show Correlation of all columns

Fig 4.2 shows another way of looking at the Correlation by using a heat map with colours. We can see the Controlling DO has a strong correlation with DO1 and DO2 as we have shown above in the report. Also Titre and Biomass have a strong correlation. Interesting there is also a correlation between Media batch and Base Buffer not seen in the above plots. This will be explored in later reports.

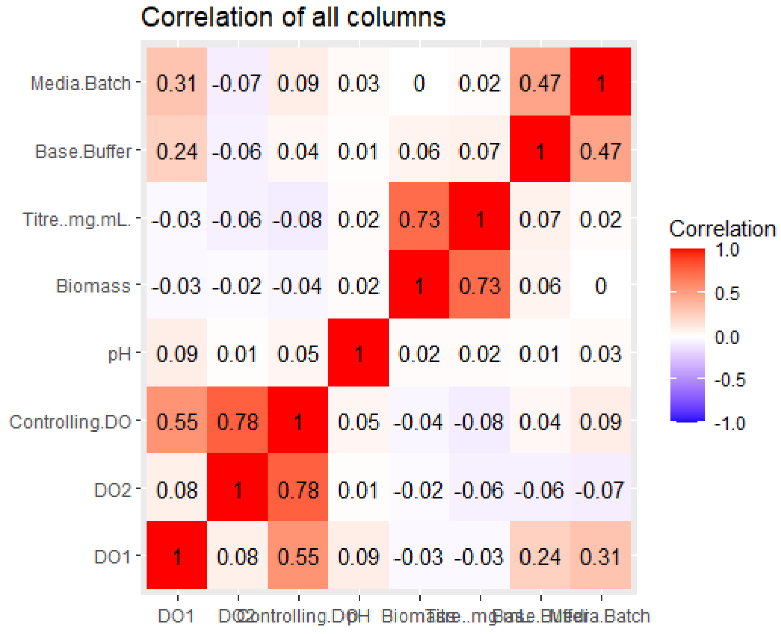


Fig 4.2: Heat map of all columns.