Ming Creekmore

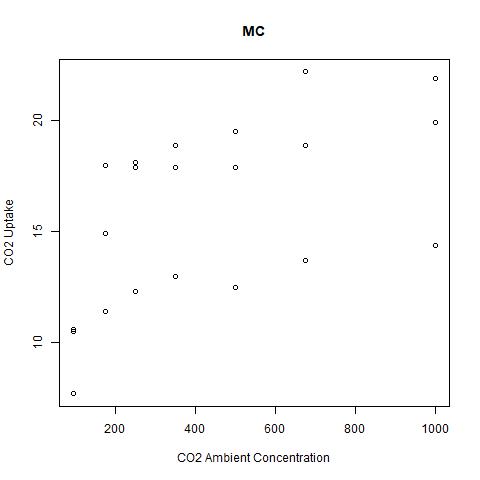
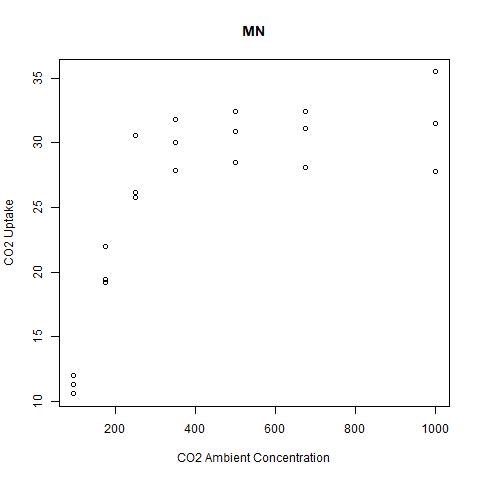
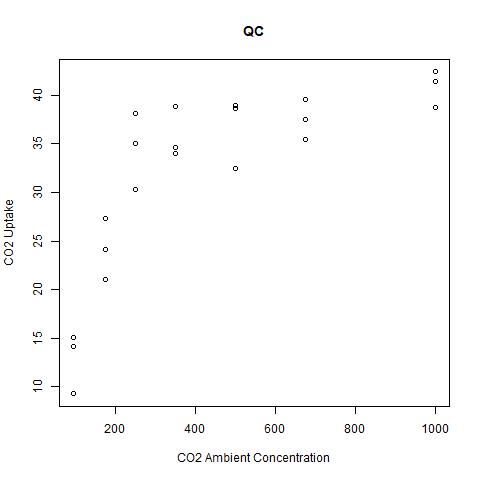
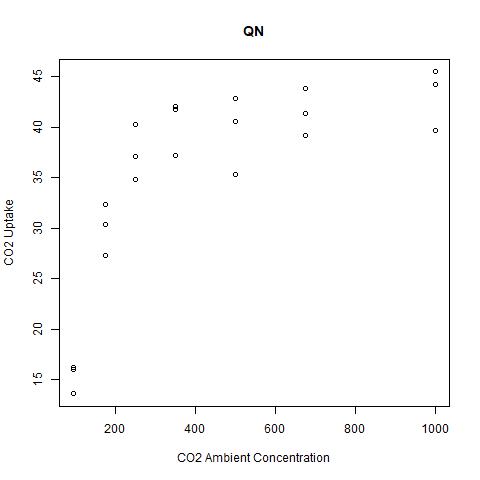
Unit1B Assignment

Statistical Analysis for Bioinformatics

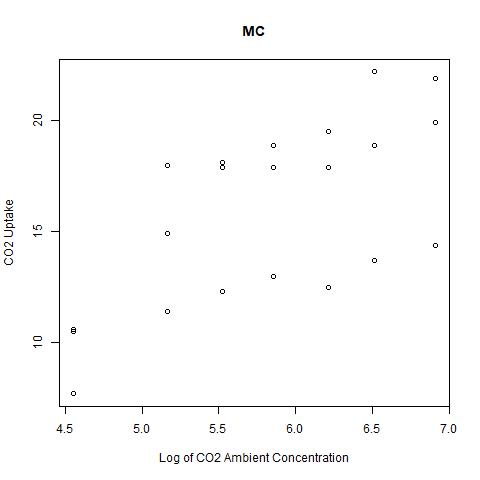
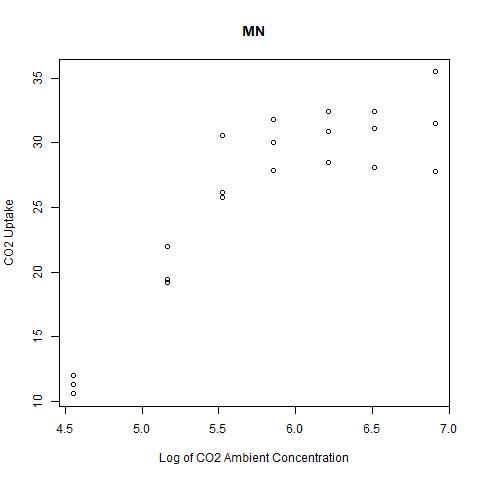
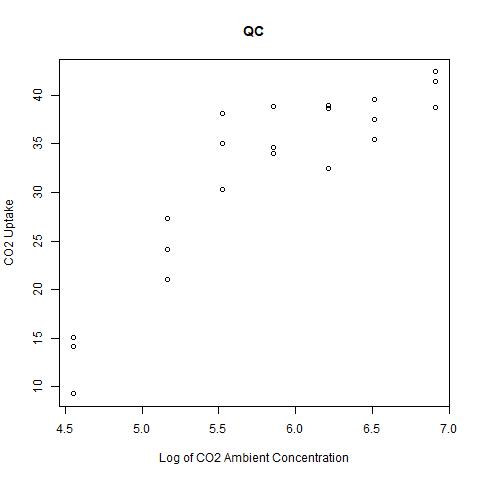
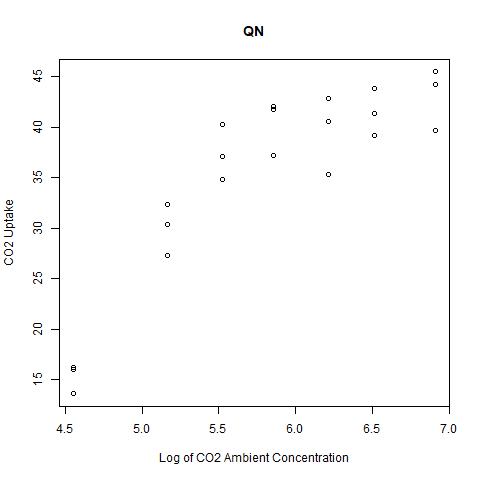
Dr. Babbit

I chose the R dataset on carbon dioxide uptake in grass plants. The data is made up of two kinds of plants: Quebec and Mississippi. Both plants are either chilled or unchilled, and the carbon dioxide ambient concentration as well as the carbon dioxide uptake in plants is recorded.

The first analysis I performed was a linear regression of concentration vs. uptake. I performed these on all plant type and treatment combinations: QN - Quebec Nonchilled, QC - Quebec Chilled, MN - Mississippi Nonchilled, MC - Mississippi Chilled.



These are the scatterplots of QN, QC, and MN, and MC. Seeing the shape of the data, I thought it might be a logarithmic function, minus MC. MC seems to not follow a pattern. I took the log of the CO2 ambient concentration to see if that would linearize it.



This graph still looks a little curved; however it does seem to have a linear pattern now, especially QC. MC still looks too scattered. MN still seems logarithmic. Even though, MC and MN doesn’t seem to fit the criteria, a linear regression was still performed on it.

**QN** - The linear regression model has a r^2 value of 0.7615, meaning that 76.15% of the data can be explained using the model. This isn’t that high of a number, but it isn’t extraordinarily low. The p-value is 2.492e-07. This is less than 0.001, so we can conclude that there is a positive relationship between the log of ambient CO2 concentration and CO2 uptake, with a 10.887 increase/decrease in CO2 uptake for every 1 unit log of ambient CO2 concentration.

Call:

lm(formula = data$uptake ~ data$conc)

Residuals:

Min 1Q Median 3Q Max

-7.9572 -3.7039 0.9632 3.2632 8.2092

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -28.019 8.202 -3.416 0.0029 \*\*

data$conc 10.887 1.398 7.788 2.49e-07 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.809 on 19 degrees of freedom

Multiple R-squared: 0.7615, Adjusted R-squared: 0.7489

F-statistic: 60.65 on 1 and 19 DF, p-value: 2.492e-07

**QC** - The linear regression model has a r^2 value of 0.7935, meaning that 79.35% of the data can be explained using the model. This is better than QN, so the relationship between the variables is stronger. The p-value is 6.22e-08. This is less than 0.001, so we can conclude that there is a positive relationship between the log of ambient CO2 concentration and CO2 uptake, with a 11.169 increase/decrease in CO2 uptake for every 1 unit log of ambient CO2 concentration.

Call:

lm(formula = data$uptake ~ data$conc)

Residuals:

Min 1Q Median 3Q Max

-8.3185 -3.4185 -0.3419 2.4324 9.6743

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -33.245 7.670 -4.334 0.000357 \*\*\*

data$conc 11.169 1.307 8.544 6.22e-08 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.497 on 19 degrees of freedom

Multiple R-squared: 0.7935, Adjusted R-squared: 0.7826

F-statistic: 73.01 on 1 and 19 DF, p-value: 6.22e-08

**MN** - The linear regression model has a r^2 value of 0.7631, meaning that 76.31% of the data can be explained using the model. This is similar to QN. The p-value is 2.334e-07. This is less than 0.001, so we can conclude that there is a positive relationship between the log of ambient CO2 concentration and CO2 uptake, with a 8.406 increase/decrease in CO2 uptake for every 1 unit log of ambient CO2 concentration.

Call:

lm(formula = data$uptake ~ data$conc)

Residuals:

Min 1Q Median 3Q Max

-7.3021 -3.3149 0.3979 2.3514 7.1514

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -22.966 6.305 -3.642 0.00173 \*\*

data$conc 8.406 1.075 7.823 2.33e-07 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.697 on 19 degrees of freedom

Multiple R-squared: 0.7631, Adjusted R-squared: 0.7506

F-statistic: 61.19 on 1 and 19 DF, p-value: 2.334e-07

**MC** - The linear regression model has a r^2 value of 0.4333, meaning that 43.33% of the data can be explained using the model. This is pretty low. The p-value is 0.001179. This is not less than 0.001, so we can conclude that there is not a relationship between the log of ambient CO2 concentration and CO2 uptake. This is to be expected, since the scatterplot of MC did not really look too linear.

Call:

lm(formula = data$uptake ~ data$conc)

Residuals:

Min 1Q Median 3Q Max

-5.1950 -2.4797 0.6702 2.3126 4.4592

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.3989 5.3473 -0.823 0.42091

data$conc 3.4735 0.9113 3.811 0.00118 \*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

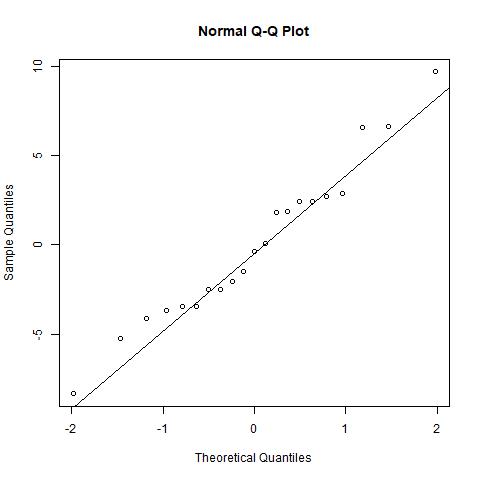
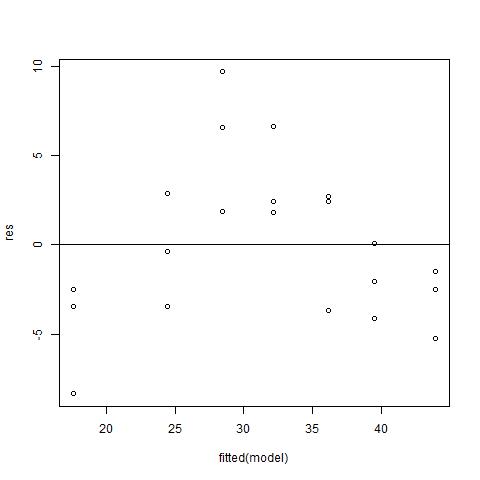
Residual standard error: 3.135 on 19 degrees of freedom

Multiple R-squared: 0.4333, Adjusted R-squared: 0.4035

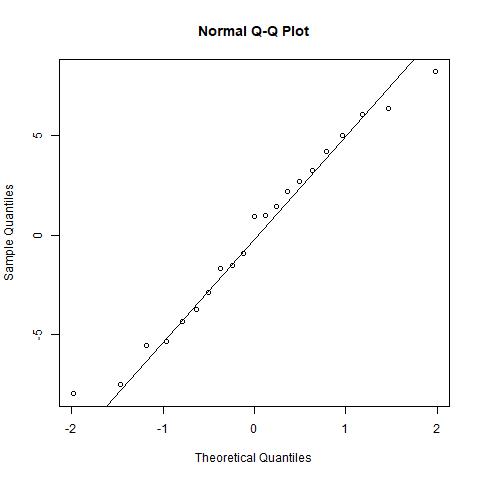
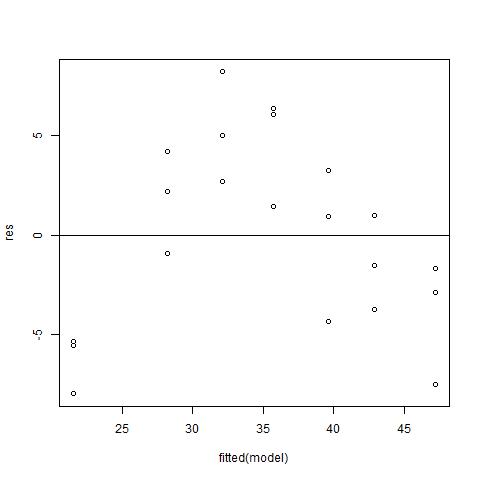
F-statistic: 14.53 on 1 and 19 DF, p-value: 0.001179

Furthermore, residual plots were graphed in order to see whether it looks how it is supposed to: scattered with no real direction and no real change in variance.

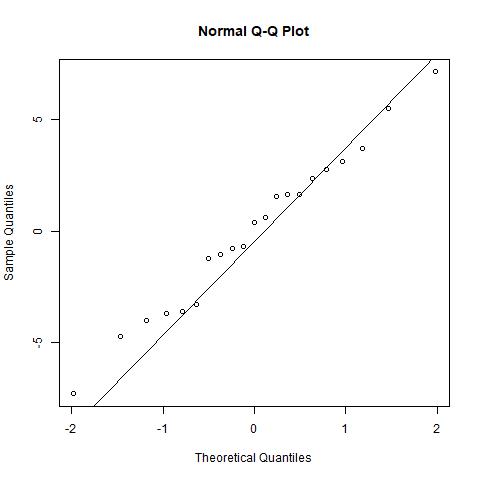
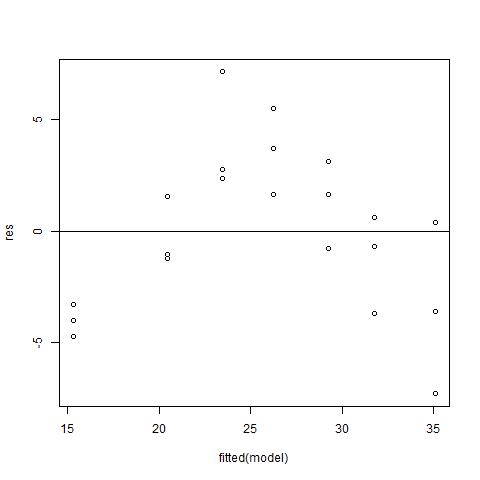
**QC** - The residual plot on the left seems to be alright, though I feel like there are some outliers that are quite far away. The normal Q-Q plot shows that most of the data follows the 45 degree angle, besides a few points that seem to be a bit farther away. From these graphs, I would say that there is some evidence to support that a linear regression (with the logged concentration) can be used to model this data, though it might be explained better with a different model.

****

**QN** - Although the Normal Q-Q plot seems alright besides a couple points on the end that stray, the residual plot seems to have a bit more of a pattern. This probably means that a linear model (with logged concentration) is not the best choice to analyze and predict this data.

****

**MN** - The pattern in the residuals is even clearer in MN. This definitely means that a linear model (with logged concentration) does not suit the data.

****

**MC** - There is no need to look at the residual and Q-Q plot for the MC data since we already determined that there was no relationship

Overall, the linear model with log of concentration values does not seem to be the best fit for this data.

The next part used Kolmogorov-Smirnov test to determine whether the data of QN and QC, MN and MC, QN and MN, and QC and MC come from the same distribution.

**QN and QC** - Between the two samples of QN and QC, the p-value is 0.09401, which is not less than 0.05. This means that there is not evidence to suggest that QN and QC come from different distrubutions. This suggests that chilling the Quebec plant does not make any difference in the CO2 uptake numbers.

Exact two-sample Kolmogorov-Smirnov test

data: QN$uptake and QC$uptake

D = 0.38095, p-value = 0.09401

alternative hypothesis: two-sided

**MN and MC** - Between the MN and MC samples, the p-value is 1.596e-05, which is less than 0.05. This means that there is significant reason to suggest that MN and MC come from different distributions. This suggests that there is a difference in CO2 uptake when a Mississippi plane is chilled vs nonchilled.

Exact two-sample Kolmogorov-Smirnov test

data: MN$uptake and MC$uptake

D = 0.71429, p-value = 1.596e-05

alternative hypothesis: two-sided

**QN and MN** - Between the QN and MN samples, the p-value is 8.928e-05, which is less than 0.05. This means that there is significant reason to suggest that QN and MN come from different distributions. There seems to be a difference in the CO2 uptake numbers in Quebec and Mississippi plants that are nonchilled.

Exact two-sample Kolmogorov-Smirnov test

data: QN$uptake and MN$uptake

D = 0.66667, p-value = 8.928e-05

alternative hypothesis: two-sided

**QC and MC** - Between the QC and MC samples, the p-value is 3.332e-07, which is less than 0.05, and an even smaller value than the other comparisons. This means that there is significant reason to suggest that QC and MC come from different distributions. There seems to be a difference in the CO2 uptake numbers in Quebec and Mississippi plants that are chilled.

Exact two-sample Kolmogorov-Smirnov test

data: QC$uptake and MC$uptake

D = 0.80952, p-value = 3.332e-07

alternative hypothesis: two-sided

Overall, there seems to be a difference in CO2 uptake between the two species. The Mississippi plant seems to be affected by chilled treatment, while Quebec doesn’t show.

To further determine whether there is a difference between the CO2 uptake in the groups based on plant type and treatment, a two-way annova was conducted. The p-value is 3.68e-10, which is less than 0.05. This means that there is significant evidence to reject the null hypothesis that there is no difference between the groups based on plant type and treatment. However, this could be bias due to skewed data. This can be checked by graphing histograms

Df Sum Sq Mean Sq F value Pr(>F)

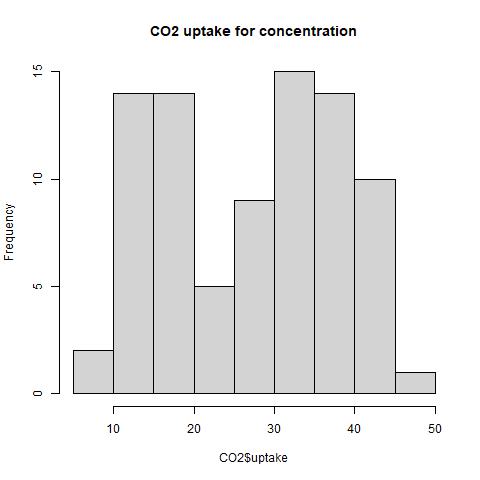
Type 1 3366 3366 50.92 3.68e-10 \*\*\*

Treatment 1 988 988 14.95 0.000222 \*\*\*

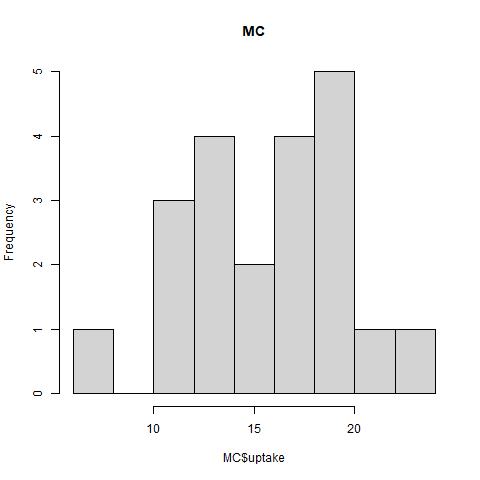
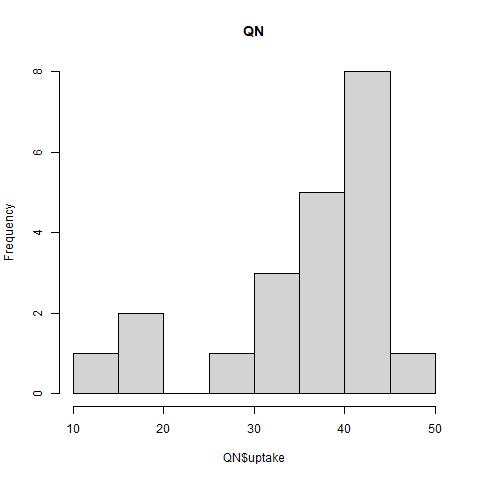
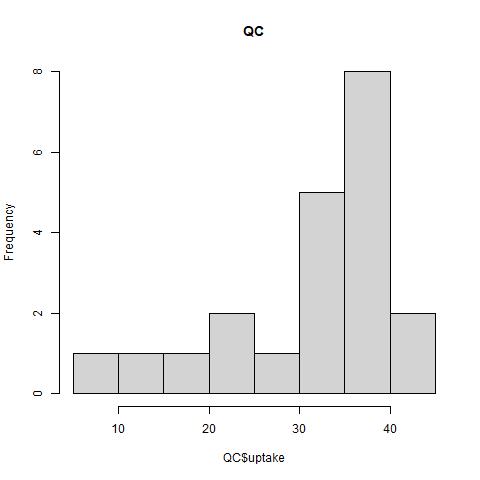
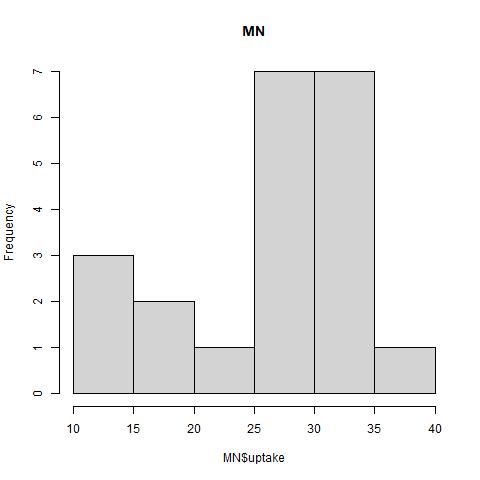
Residuals 81 5353 66

---

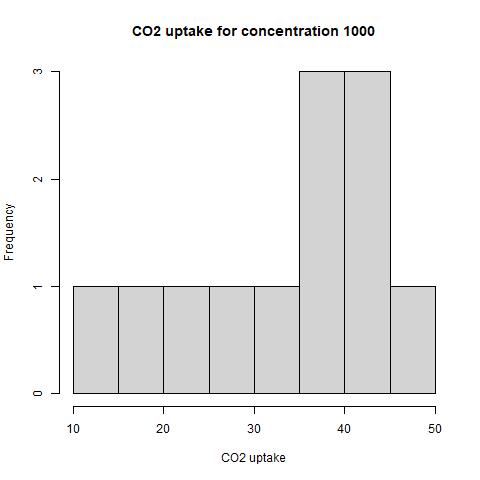
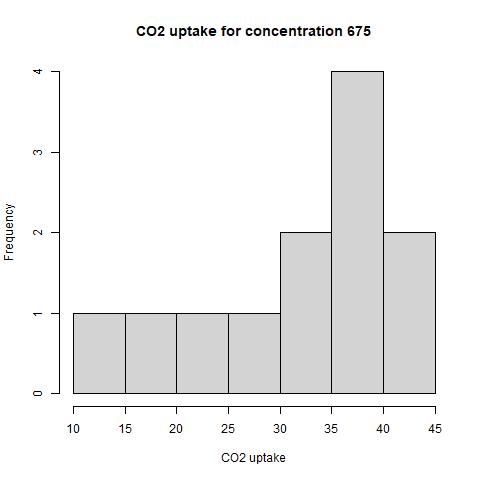
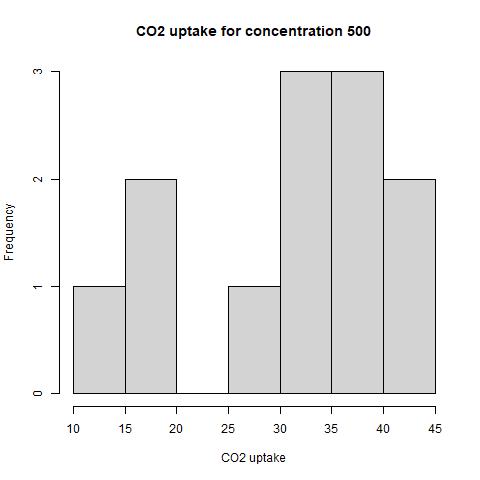
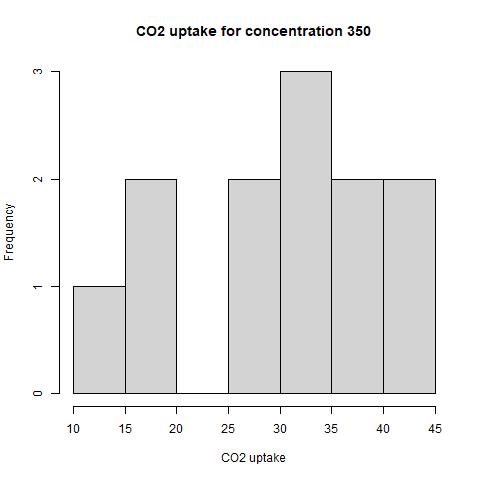
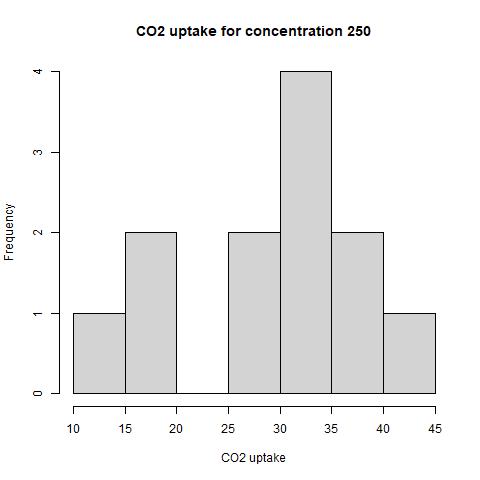
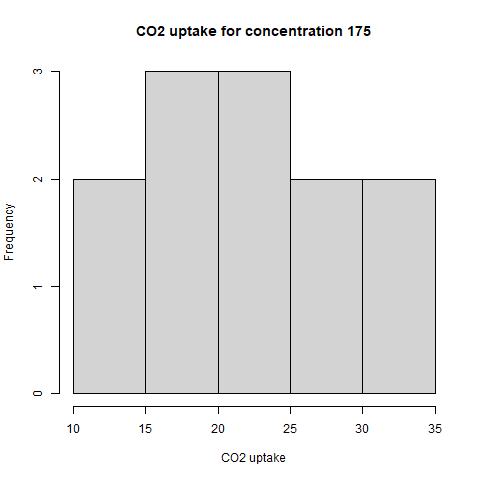
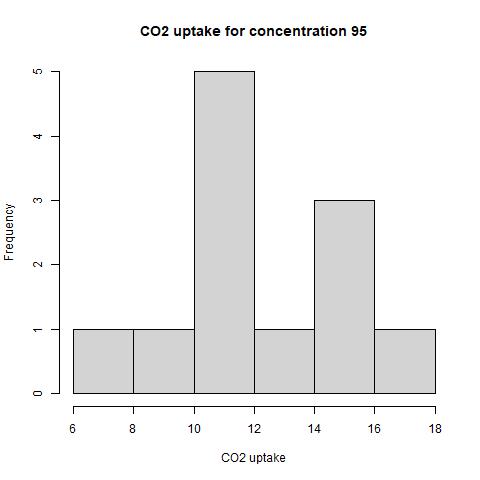
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The histogram that has all of the data together shows a bimodal distribution. One of the centers is at the range of 10-20, and the other is at the range of 30-45. This could indicate the differences in chilled vs. nonchilled, where chilled plants tend to have lower uptake. Or it could indicate the difference between ambient CO2 concentration since a lower ambient CO2 concentration correlates with a lower CO2 uptake. Or it could be the difference in the plant type, where Quebec tends to have higher CO2 rates. 

Other histograms were created to compare across consistent ambient CO2 concentrations. While MC might have a normal distribution, the others definitely do not. The others are sewed to the right. Both QC and QN have similar histograms, which reiterates the idea that QN and AC come from the same distribution. MN is not as badly skewed as the other two. It also shows a higher CO2 uptake average than MC, which relates to analysis done beforehand. Although MC does have a distribution that could be seen as normal, it seems like it could also be bimodal. More data would be needed to classify it. Especially since earlier analysis doesn’t show many relationships.

There were also histograms created between the different concentrations. These histograms also seem more skewed to the right on average. There also seems a higher chance of a bimodal distribution for uptake with ambient CO2 levels at 95, 250, 350, and 500. The histograms of 95 and 175 don’t seem to be as skewed as the others. 175 might even make it towards a uniform distribution. 350 could also be seen as a uniform distribution if there wasn’t a gap in there. These graphs seem to suggest that plants, regardless of type or treatment tend to try and uptake more CO2. But it also seems like there are quite a few plants that do hit the lower parts. This is probably the chilled Mississippi, since the cold seems to lower the Mississippi’s CO2 uptake. The bimodal distributions probably also came about due to the chilled Missisippi.



Finally, a Random Forest Machine Learning Algorithm was trained to classify the dataset. The random forest algorithm was specifically set to predict the plant type and treatment type based on the ambient CO2 concentration and the CO2 uptake. The dataset was separated into training data and testing data. This was done by setting every other two data rows to the training set and the skipped rows to the testing set. This assured that neither set was skewed towards the one plant type, treatment, or ambient CO2 concentration. The results of the random forest model was an accuracy of 78.57%. Below, you can see multiple scatter plots. All data scatterplot is the full data classification with x-axis as CO2 uptake and y-axis and ambient CO2 concentration. Based on human eyes, you can see the graphing patterns stated earlier. Since the axis is switched, the graph doesn’t look logarithmic, but exponential. Some plant types and treatments overlap, especially the Quebec plants. There is also a scatterplot of the training dataset, so you can see the variety in the data given to the machine learning algorithm to train. There is a scatterplot of the actual testing data and the predicted testing data. This shows what was classified right and what was classified wrong. There seems to be more overlap in the earlier concentrations, as well as overlap in the Quebec plant area. This is to be expected based on earlier analysis. The histogram distribution of specific concentrations show that there seems to be more normal or uniform distributions during the concentrations. And the Quebec plants were shown to not have significant evidence to suggest they are different.

