CO2 Dataset EDA

Introduction/Dataset Description:

The dataset I chose to use was the CO2 dataset also called the Carbon Dioxide Uptake in Grass Plants dataset from the geeksforgeeks website. This dataset has 5 variables. It has 3 discrete variables which are:

Plant - Identifies each individual plant. It contains 12 unique plant identifiers such as Qn1 and Mc3. Q stands for Quebec and M stands for Minnesota. N stands for non-chilled, and C stands for chilled.

Type – Indicates the plant's origin whether it be Quebec or Minnesota.

Treatment – Says what treatment condition the plant was in. There are 2 options, and they are Non-Chilled and Chilled.

So, the other 2 remaining variables are continuous, and they are:

Conc – Conc stands for the concentration of co2 in parts per million. It has a range of 95-500ppm.

Uptake – This is the rate of co2 uptake by the plant in micromoles per square meter per second, with values ranging from approximately 1.4 to 37.3.

The target variable in this dataset is Uptake.

Look at the Data:

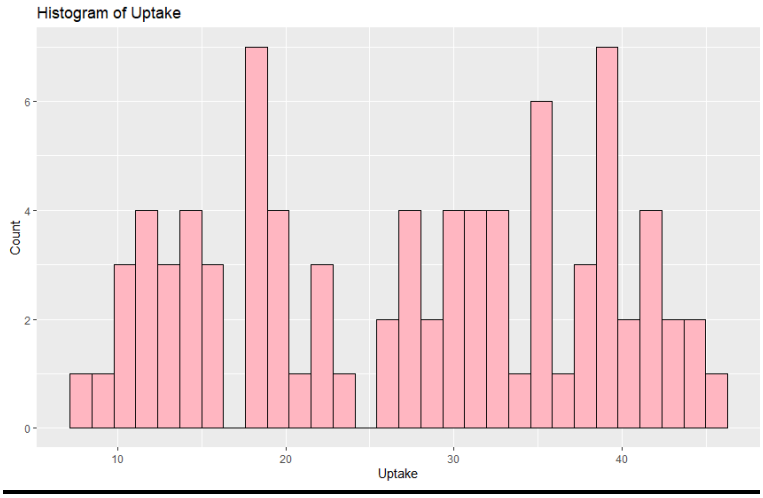
Below you will find the first 5 rows of the dataset, the structure and summary of the dataset, as well as if there are any missing values.

```
> str(data)
Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':	84 obs. of  5 variables:
 $ Plant    : Ord.factor w/ 12 levels "Qn1"<"Qn2"<"Qn3"<..: 1 1 1 1 1 1 1 2 2 2 ...
 $ Type     : Factor w/ 2 levels "Quebec","Mississippi": 1 1 1 1 1 1 1 1 1 1 ...
 $ Treatment: Factor w/ 2 levels "nonchilled","chilled": 1 1 1 1 1 1 1 1 1 1 ...
 $ conc     : num  95 175 250 350 500 675 1000 95 175 250 ...
 $ uptake   : num  16 30.4 34.8 37.2 35.3 39.2 39.7 13.6 27.3 37.1 ...
 - attr(*, "formula")=Class 'formula'  language uptake ~ conc | Plant
  .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "outer")=Class 'formula'  language ~Treatment * Type
  .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "labels")=List of 2
  ..$ x: chr "Ambient carbon dioxide concentration"
  ..$ y: chr "CO2 uptake rate"
 - attr(*, "units")=List of 2
  ..$ x: chr "(uL/L)"
  ..$ y: chr "(umol/m^2 s)"
```

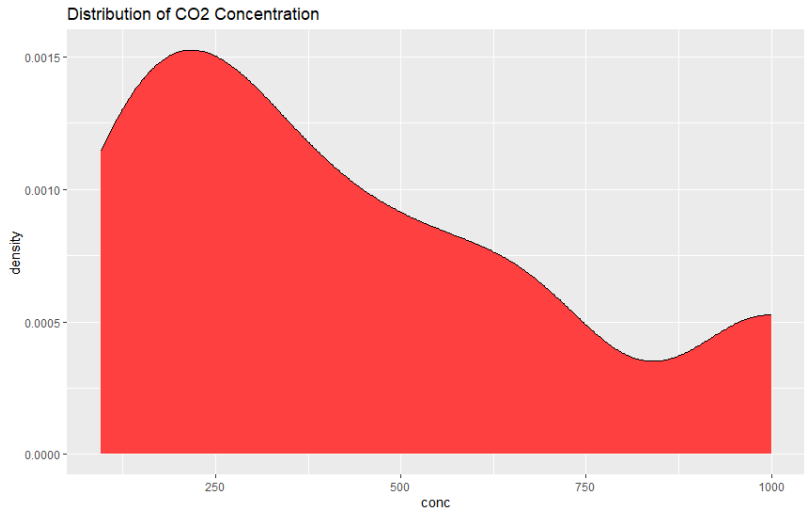
```
> summary(data)
     Plant          Type         Treatment       conc          uptake      
 Qn1    : 7   Quebec     :42   nonchilled:42   Min.   :  95   Min.   : 7.70  
 Qn2    : 7   Mississippi:42   chilled   :42   1st Qu.: 175   1st Qu.:17.90  
 Qn3    : 7                                    Median : 350   Median :28.30  
 Qc1    : 7                                    Mean   : 435   Mean   :27.21  
 Qc3    : 7                                    3rd Qu.: 675   3rd Qu.:37.12  
 Qc2    : 7                                    Max.   :1000   Max.   :45.50  
 (Other):42
```

```
> sapply(data, function(x) sum(is.na(x)))
    Plant      Type Treatment      conc    uptake 
        0         0         0         0         0 
```

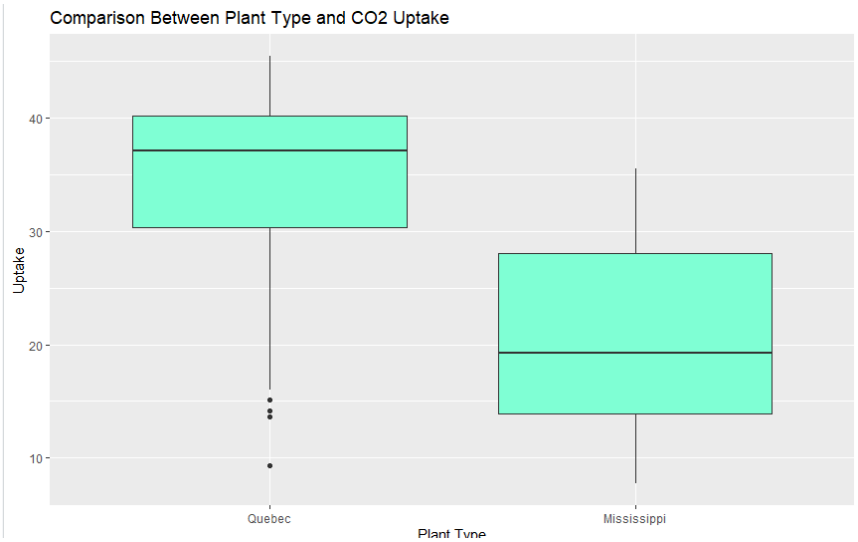
Graphical Exploration and Findings:



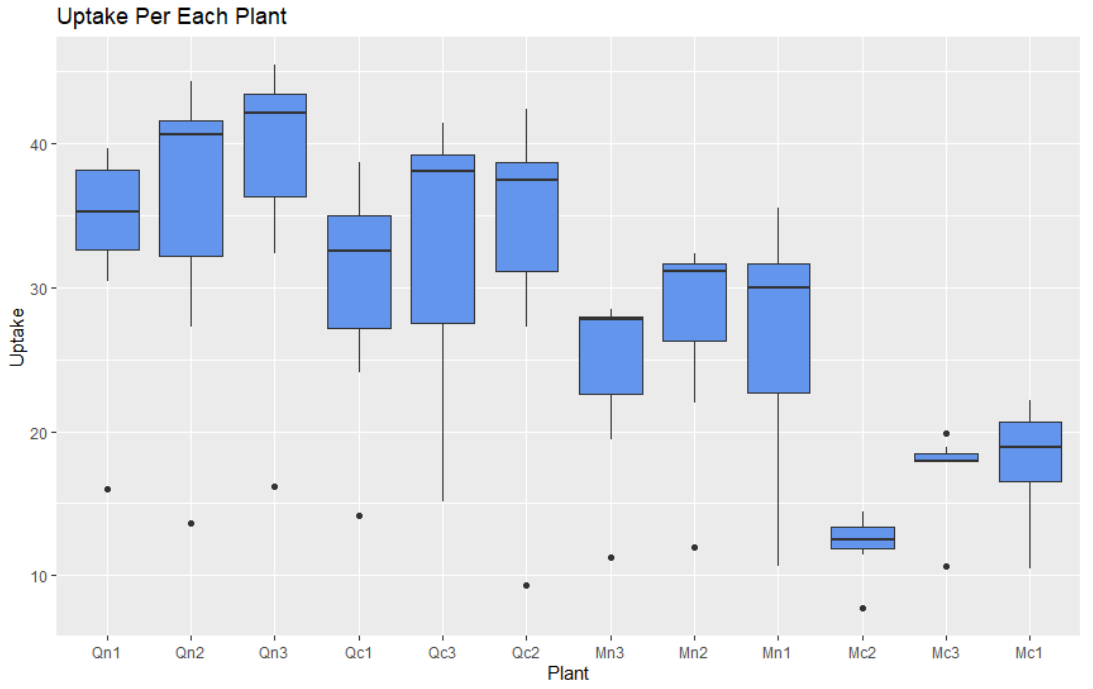
This first graph is just a histogram showing the distribution of Uptake in the plants. From this graph you can see the results are decently spread.



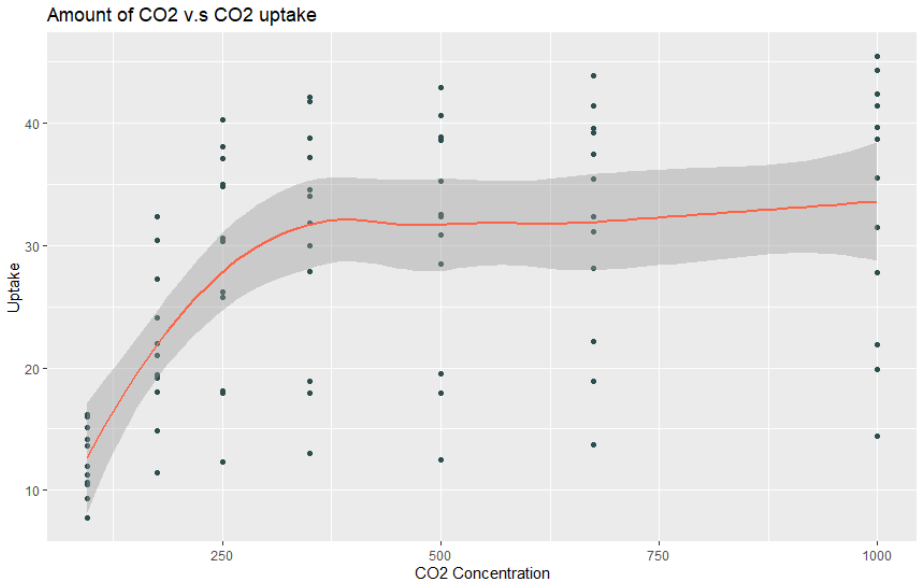
This next graph I want to show is a density plot showing Concentration of CO2. From this graph you can see that lower parts per million are more common.



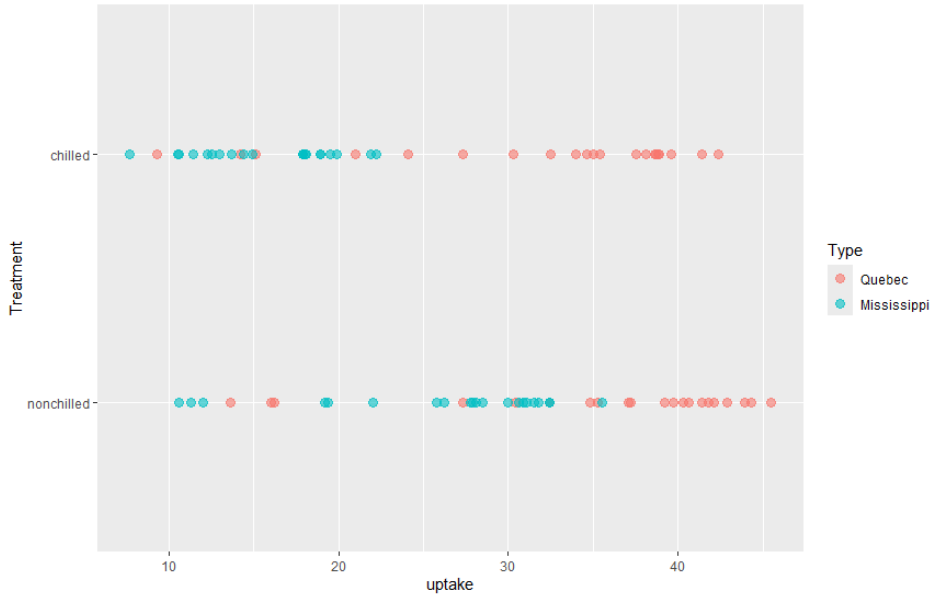
The first bivariate graph given above shows that grass plants from Quebec take in more co2 than grass plants from Mississippi. However, Quebec also has some outliers and Mississippi really doesn't.



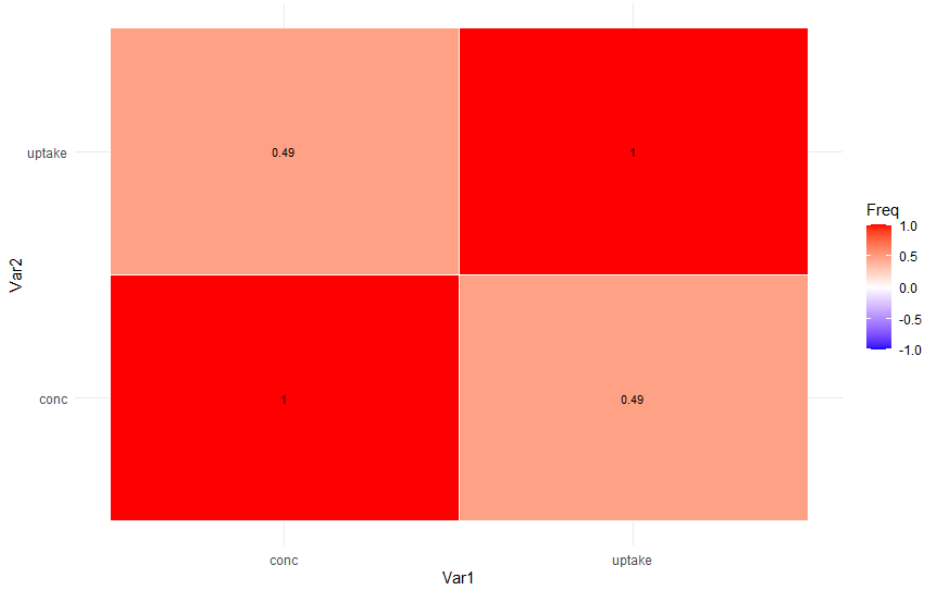
This graph is probably the most explanatory. I think this because it shows what plant has the highest uptake in co2. So, you can see from this that overall, the non-chilled plants from Quebec had the best result, specifically Qn3. You can also see that the chilled plants from Mississippi did the worst, specifically Mc2.



This plot shows the correlation between CO2 concentration and uptake. The slight trend shown from this is that the higher the concentration the higher the uptake. Again, it is not incredibly clear after around 250ppm.



This plot includes 3 variables where x is uptake and y are treatment, also the type is used for color. This graph gives a more in-depth view for Uptake and treatment for each plant type. It shows again that non-chilled Quebec grass plants generally have better uptake.



The heatmap that can be made from this dataset does not give a whole lot of insight, but it does show that uptake and co2 concentration have a moderate correlation.

Summary of Findings:

From the analysis done you can come to 2 main conclusions. One of these and probably the most important one is that grass plants from Quebec that aren't chilled have the best uptake. This can be seen from my second box plot as well as my multivariate point plot. Along with both showing what worked best they also showed chilled plants from Mississippi do the worst. The other conclusion you could come to is that the higher concentration of CO2 the better Uptake. This could be debated as both the heatmap and point plot with the trend show the correlation between them being moderate. So, it could go either way, however I would say the higher concentration the higher uptake. So, to recap grass plants from Quebec that aren't chilled have better results, and the more CO2 concentration the more likely Uptake will be higher as well.