CO2 dataset from R

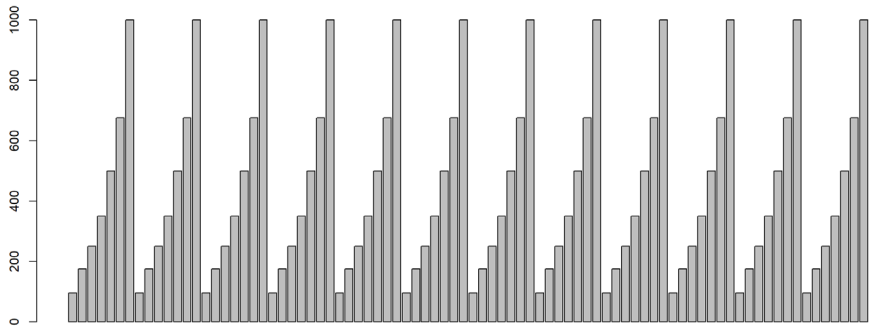
There are 5 columns in this dataset.
1 is the specific plant
2 is the type of plant
3 is if the plant is treated in chilled or nonchilled environment
4 is the concentration of CO2
5 is the actual uptake of CO2 into the plant.

The two numerical columns mean, median, mode and range value are calculated below

```
library(modeest)
attach(CO2)                  #to minimize the code and ease visual
mean(conc)                   #= 435

median(conc)                 #= 350

mfv(conc)                   #= 95, 175, 250, 350, 500, 675, 1000(mfv is mode from modeest package)


mean(uptake)                 #= 27.2131

median(uptake)               #=28.3

mfv(uptake)                  #=17.9 32.4

range(conc, na.rm = FALSE)   #= 95 to 1000
range(uptake, na.rm = FALSE) #= 7.7 to 45.5
```

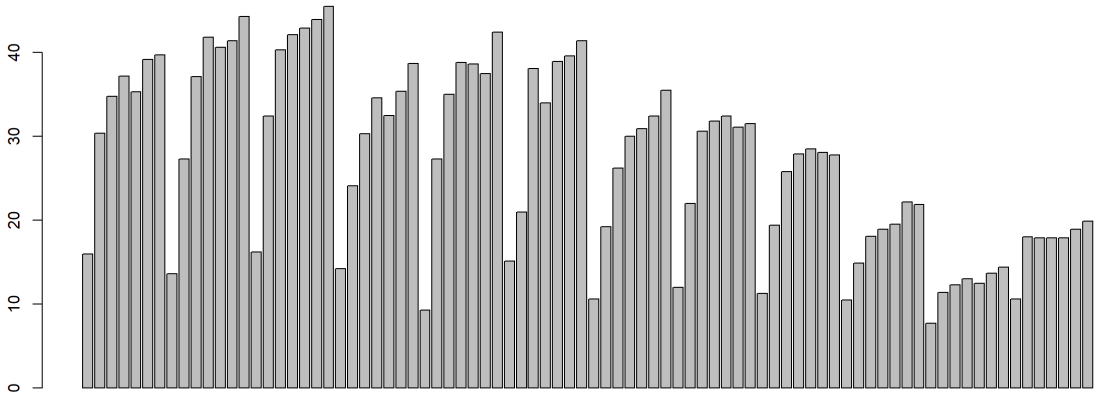
To see and understand the numerical values, a barplot is used.

```
barplot(conc)
```



With this we can clearly see that it's a systematic usage of concentrated CO2, so the dataset seem to be how much plants "uptake" from different concentration of CO2. This means also that mean, median and mode are not particular useful in the "conc" column. "conc" have discrete variable.

barplot(uptake)



Doing a barplot for the uptake the difference is not systematic.

class(CO2) #= "nfnGroupedData" "nfGroupedData" "groupedData" "data.frame"

class(Plant) #= "ordered" "factor"

class(Type) #="factor"

class(Treatment) #="factor"

class(conc) #="numeric"

class(uptake) #="numeric"

The dimensions of the dataset is
dim(CO2) #= 84 and 5. This means 84 rows, and 5 columns.

Colulmn 1-3 are factors, and column 1 seem to be based on code of the other 2 columns.

View(CO2) gives the first row
Qn1 Quebec nonchilled 95 16.0

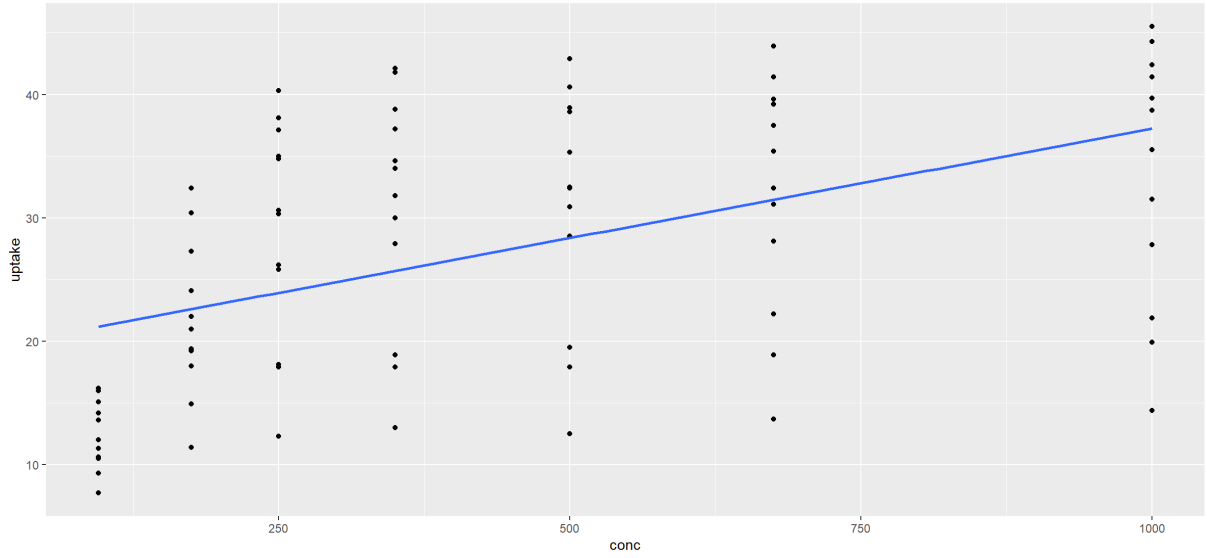
"Q" stands for Quebec, "n" for nonchilled 1 should be a plant type. In other Plants in the column "Plants" range from 1 – 3 with different combinations of column 2(Type) and 3(Treatment)
This is derived from
sapply(CO2$Plant,unique) # which gives Levels: Qn1 Qn2 Qn3 Qc1 Qc3 Qc2 Mn3 Mn2 Mn1 Mc2 Mc3 Mc1

The entire set of what sapply(CO2$Plant,unique) actually prints is not pasted here since what was interesting to see is the range of the Plant variable.

If, the correlation between the concentration of exposed CO2 to the amount that the plants take up is linear, then there should be no different based on what plant and/or if it is chilled or not. But since the barplot for the uptake clearly show thats not the case.

What type of plant, condition and CO2 concentration is most viable for future study. What can be used with higher probability to yield a better uptake.

Since Plant, Type and Treatment are non-numerical, and therefore can't be plotted in a Linear Regression model, we start with making the plots and calculation for the two remaining columns "conc" and "uptake" to see how they relate to each other.
qplot(conc,uptake)+geom_smooth(method=lm,se=F)



With…
summary(lm(formula=uptake~conc,data=CO2))
…we get
Residuals:

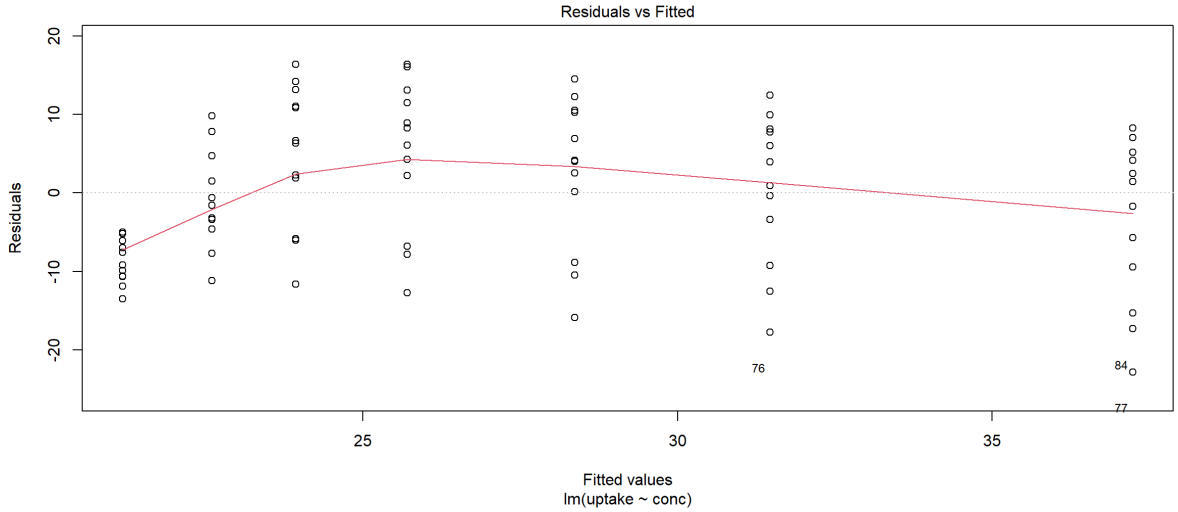
Min	1Q	Median	3Q	Max
-22.831	-7.729	1.483	7.748	16.394

So we have now, the point that is farthest below (-22.831), above(16.394), and this is pretty good, since the min and max should have the same magnitude. 25 % is below 1Q that is -7.729 and 25% that is above 3Q, that is 7.748. Same is for 1 Quartile and 3 Quartile, they too should have the same magnitude and it seem to fit well according to that. For this line to be a perfect match, as we have already established it isn't. The median should be 0. The number we have is 1,483, and that is all compared with the range of the values we work with.

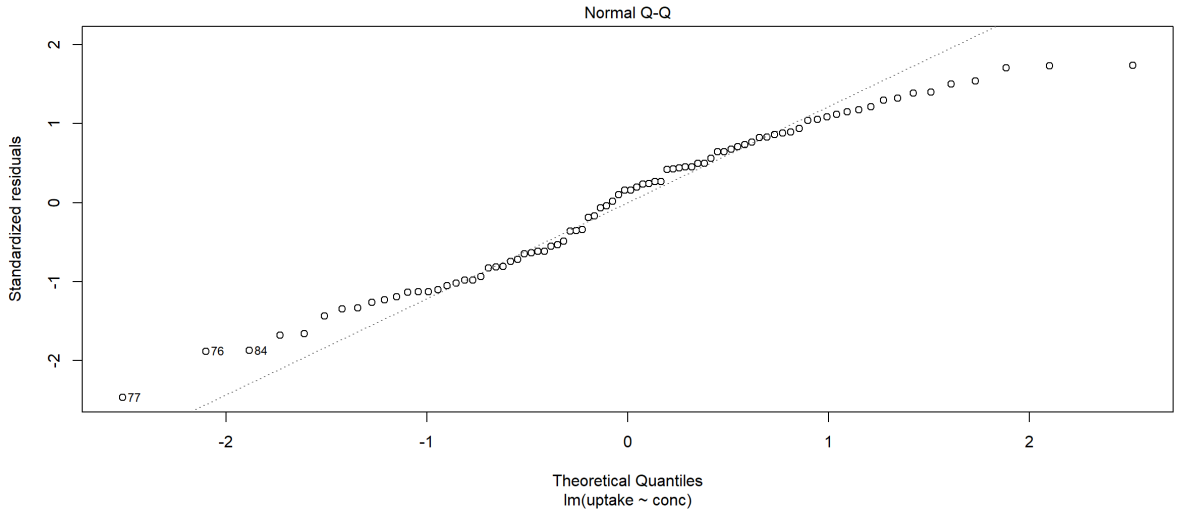
Residual values are the difference between the observed value and the predicted value.
The plot below illustrate this
summary(lm(formula=uptake~conc,data=CO2)) # plotting the residual values

summary(C)

plot(C)

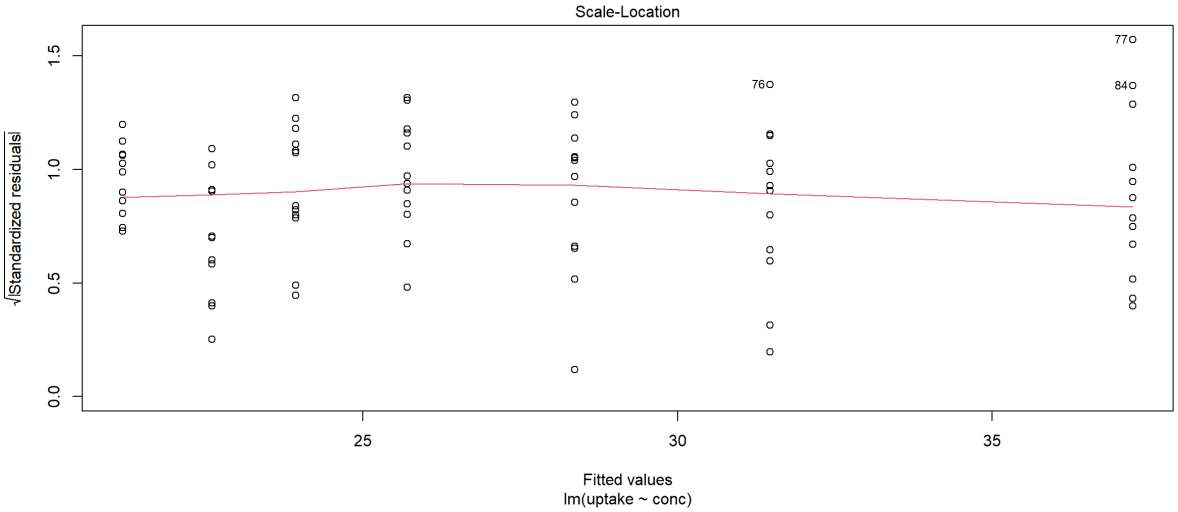


The predicted values should be both below and above, both on the line and around it. This graph shows that the residuals are initially close to 0, but the higher concentration of CO2, the residuals become more widespread (to the right of the graph), which means that with higher concentration, the accuracy is lower.

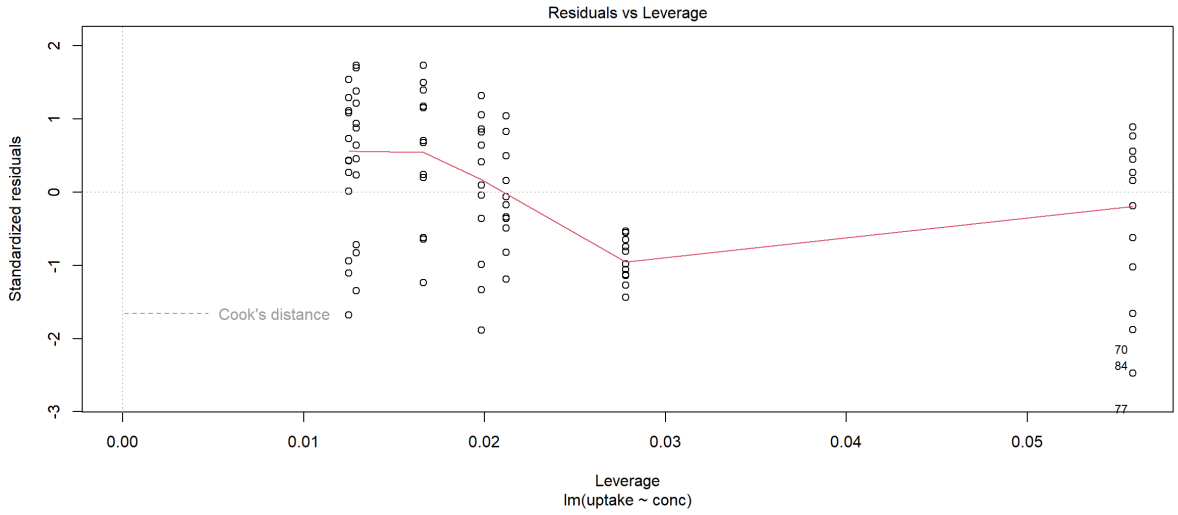


The Q-Q plot is showing if the residuals is distributed normally when they are compared to the normal distribution. Since the residuals are centered within -2Q and 2Q with only a few outliners, it shows that the data points are normally distributed.

The third plot shows Fitted Values vs Squared root of the Standardized residuals. A straight line shows equal value for the residuals



Forth plot from summary is a leverage plot. It means to see what data points which lay far to the right, but high above or below, since they would have an impact on the residual values. This is interesting, since it correlates to the first plot of summary that showed the sloping curve with higher concentration of CO2. In the graph below we can see, that datapoint 70, 77 and 84 have a high leverage on the residuals.

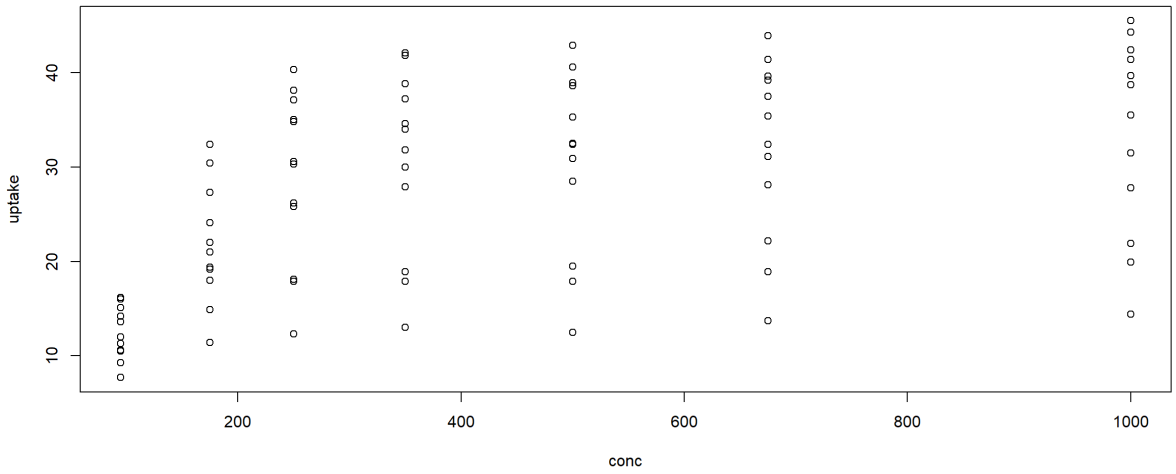


Looking at
View(CO2) we can see clearly that the concentration of CO2 is 1000 on all three. But this is only in Mc1, Mc2 and Mc3, which translate to Mississippi chilled 1, 2 and 3. All those 3 plants have a very low uptake. Since we are looking for a trend of high uptake and the correlation to concentration, this can be interesting sidenote only. Had it been the other way around that they would have a high uptake, this would have been even more interesting, and value to the question of this report.

Unfortunately, the Multiple R-squared: 0.2354(Coefficient of determination) is low, only 0,2354, which means that the model only explains the data at a level of 23,54%. The higher the better. Removing the row 70,77 and 84 on a test data.frame did not change the percent.

What we learned from plotting the summary and a linear regression graph is that higher concentration of CO2 does not correlate the continues uptake on all plants. Some take up less. But since 3 columns in this data set could not be used to summarize in the previous plots, they will help determinate which plants have a higher uptake here forth.

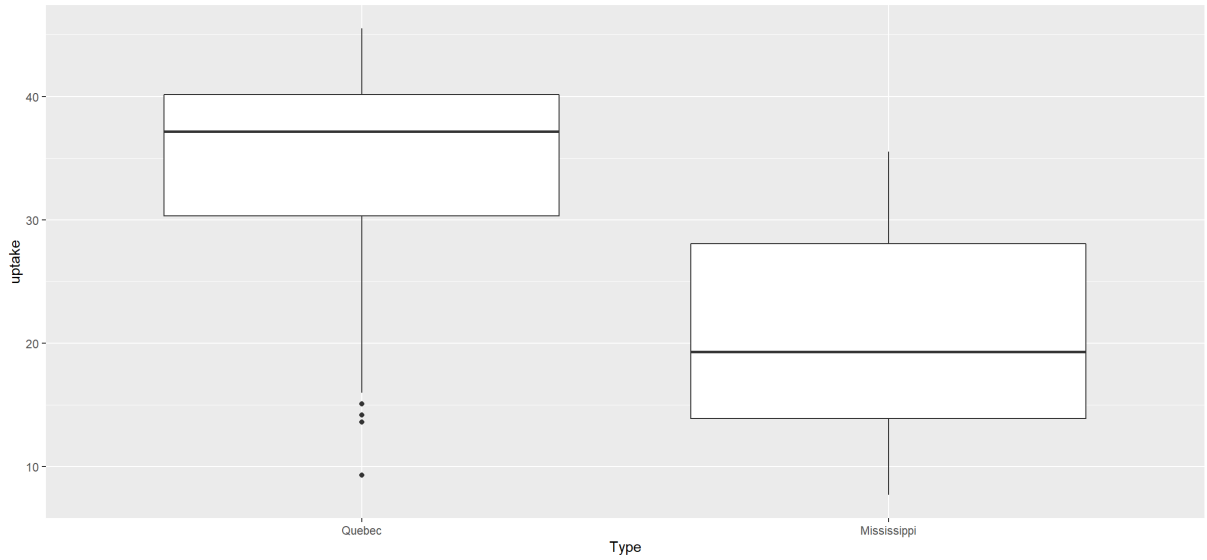
We start with a simple plot



This would indicate that they higher the concentration of CO2 the better uptake, but equally lower concentration show the same promising results. But this time around what type and if they are chilled or not needs to be in the “equation”.

C2 <- ggplot(CO2, aes(Type,uptake))

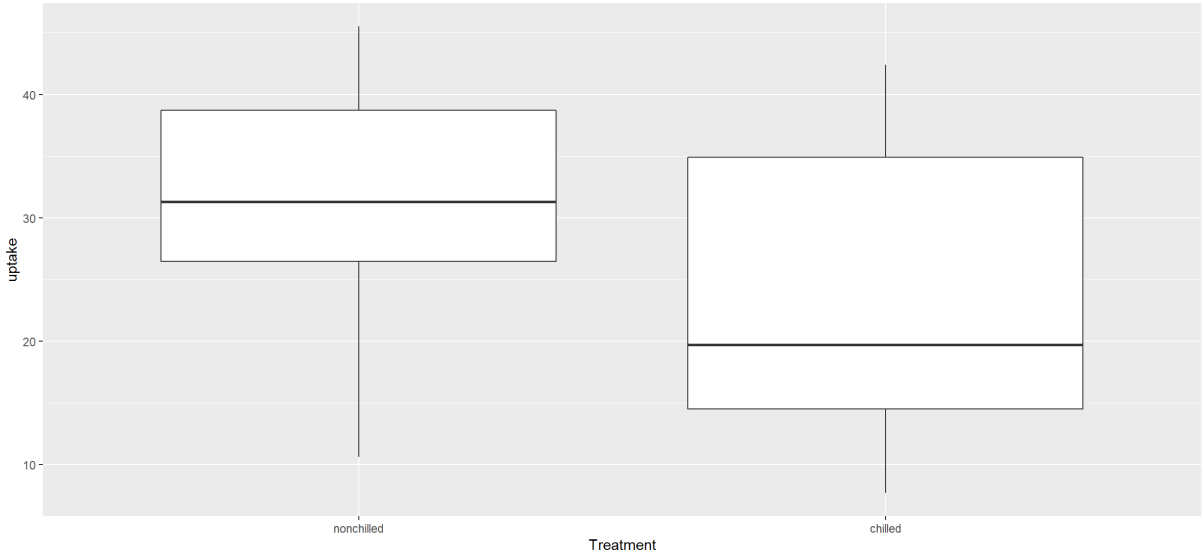
C2 + geom_boxplot()



This boxplot show that Mississippi plants do worse than all Quebec plants, when only the Type is compared to the actual uptake.

Similar on chilled vs nonchilled. The median for nonchilled is almost at the top of chilled number. The median is low on chilled. This means that the amount chilled that can actually compete with nonchilled are few. Since we are looking for plants that are more likely to have a higher uptake of CO2, nonchilled is generally more interesting.
C3 <- ggplot(CO2, aes(Treatment,uptake))

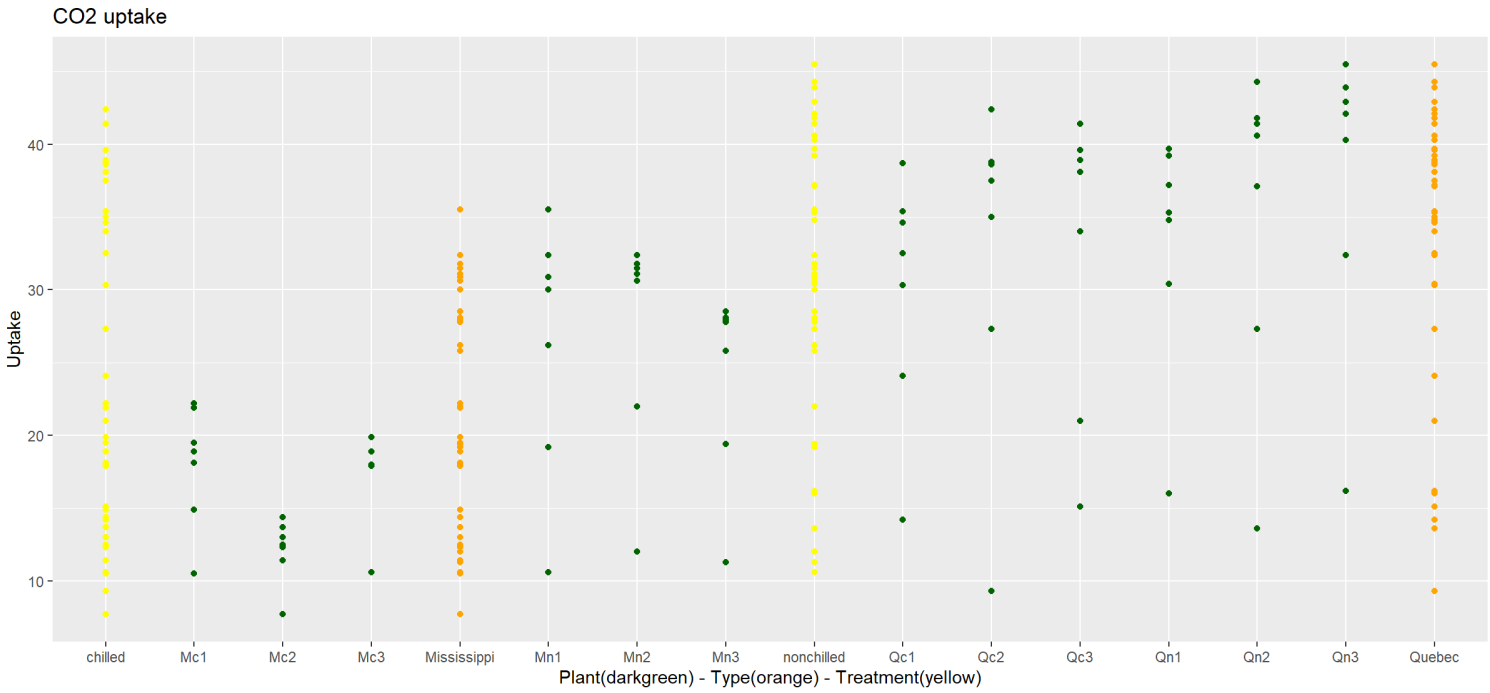
C3 + geom_boxplot()



This concluded that both Mississippi and chilled is not the plants we are looking for.

What is left is Quebec and nonchilled. What is the difference there when it comes to concentration and the type 1-3.
To find this:

```
library(ggplot2)

ggplot(CO2, aes(y=uptake)) +
  geom_point(aes(x = Plant), color = "darkgreen")+
  geom_point(aes(x = Type), color="orange")+
  geom_point(aes(x = Treatment), color="yellow")+
  labs(title="CO2 uptake",y= "Uptake",
  x= "Plant(darkgreen) - Type(orange) - Treatment(yellow)")
```



This graph that contains the variables Plants, Type and Treatment and is measured against Uptake. Chilled has a high value, but as we have seen, not on Mississippi. Nonchilled has an average of higher than chilled. Mc1, Mc2 and Mc3 has already been mentioned to have low score and according to this graph, that is at least true. Comparing Mississippi with Quebec, we can see that the range of Quebec is far, so from this graph Mississippi would still be viable, but again as we have seen in previous graphs, Mississippi does good with nonchilled. But not compared to Quebec, where they score higher. The conclusion of our question. What type of plant, condition and CO2 concentration is most viable for future study is Quebec, nonchilled, plant version 1, 2 and 3. To narrow it down farther to find only one plant. It would be Qn3