Stomatal conductance of *Sorghum bicolor*

This project stems from a larger study of *Sorghum bicolor*'s hydraulics under salt stress on the short term, this analysis is focused on the stomatal conductance, one of the many variables measured in the study. The stomata are small pores found in the epidermis of some plant leaves that allow the exit of water vapor and the entry of CO_2 (gas exchange), in most plants they are closed during the night and open during the day, when photosynthesis is carried out. In this case, the stomatal conductance is our response variable of interest, measured in mmol of water vapor per m^2 and second, it basically represents how many stomata are open and how many are closed.

Our hypothesis for this work was that the stomatal conductance decreases as the medium salt concentration increases and time passes (1, 4 and 24 hours were used in this study but for a shorter analysis only the 1 and 24 hours will be shown). And also, that this decline differs for each genotype, as seen in previous studies done under water stress.

First of all, we import the necessary libraries for the analysis and create a tiny function to calculate the standard error of the mean (S.E.M.).

```
library("readxl")
library("dplyr")
library("ggplot2")
se <- function(x) sd(x)/sqrt(length(x))
```

Then, we import the dataset and use the function `str()` to take a quick look at the dataset.

```
df <- read_excel("Data_gs.xlsx")
str(df)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    262 obs. of  6 variables:
##  $ Set     : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ Genotype: chr  "9530" "9530" "B2" "B2" ...
##  $ Plant   : num  4 5 4 5 4 5 4 5 4 5 ...
##  $ Time    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Salt    : num  0 0 0 0 100 100 100 100 200 200 ...
##  $ gs      : num  40 48.4 38.9 67.2 51.2 31.9 49.2 33.8 34.9 16.8 ...
```

As we can see, some columns were not imported in the most convenient type (such as the Time explanatory variable as numeric), we just convert them to factors accordingly:

```
df[1:5] <- lapply(df[1:5], as.factor)
```

Exploratory data analysis

Missing values, outliers and general description

We use the `summary()` built-in function to obtain some general statistic descriptive values for each column:

```
summary(df)
```

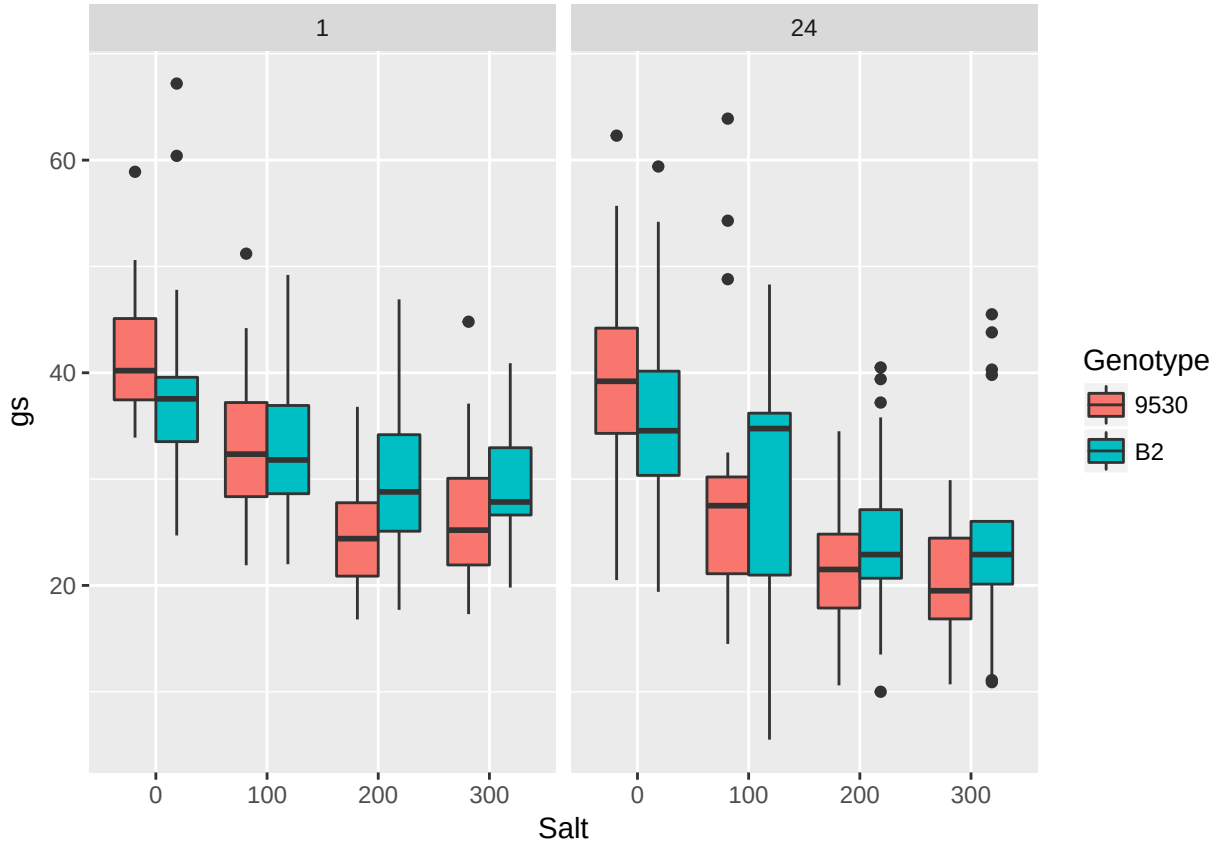
```
##  Set    Genotype       Plant     Time      Salt          gs
##  2:40   9530:130   19     : 17   1 :112   0  :65   Min.   : 5.50
##  3:41   B2  :132   5      : 16   24:150   100:63   1st Qu.:22.30
##  4:40              9      : 16            200:66   Median :28.20
##  5:40              13     : 16            300:68   Mean   :29.94
##  6:35              14     : 16                     3rd Qu.:36.80
```

```
##  7:33              8      : 15                    Max.    :67.20
##  8:33              (Other):166                    NA's    :7
```

As we can see, the data is considerably unbalanced (for example: time, 112 values for 1 hour, 135 for 4 and 150 for 24). In our response variable's column (gs), we detect 13 missing values and considering the median is approximately the same as the mean, we can assume outliers won't be a big deal (though we'll take care of them shortly).

Since this is a dataset coming from experiments, it wouldn't be correct to fill NAs with the mean because considering the small sample for each genotype, salt, time combination it would bias our results too much. So, we shall omit them and then use a boxplot to check outliers:

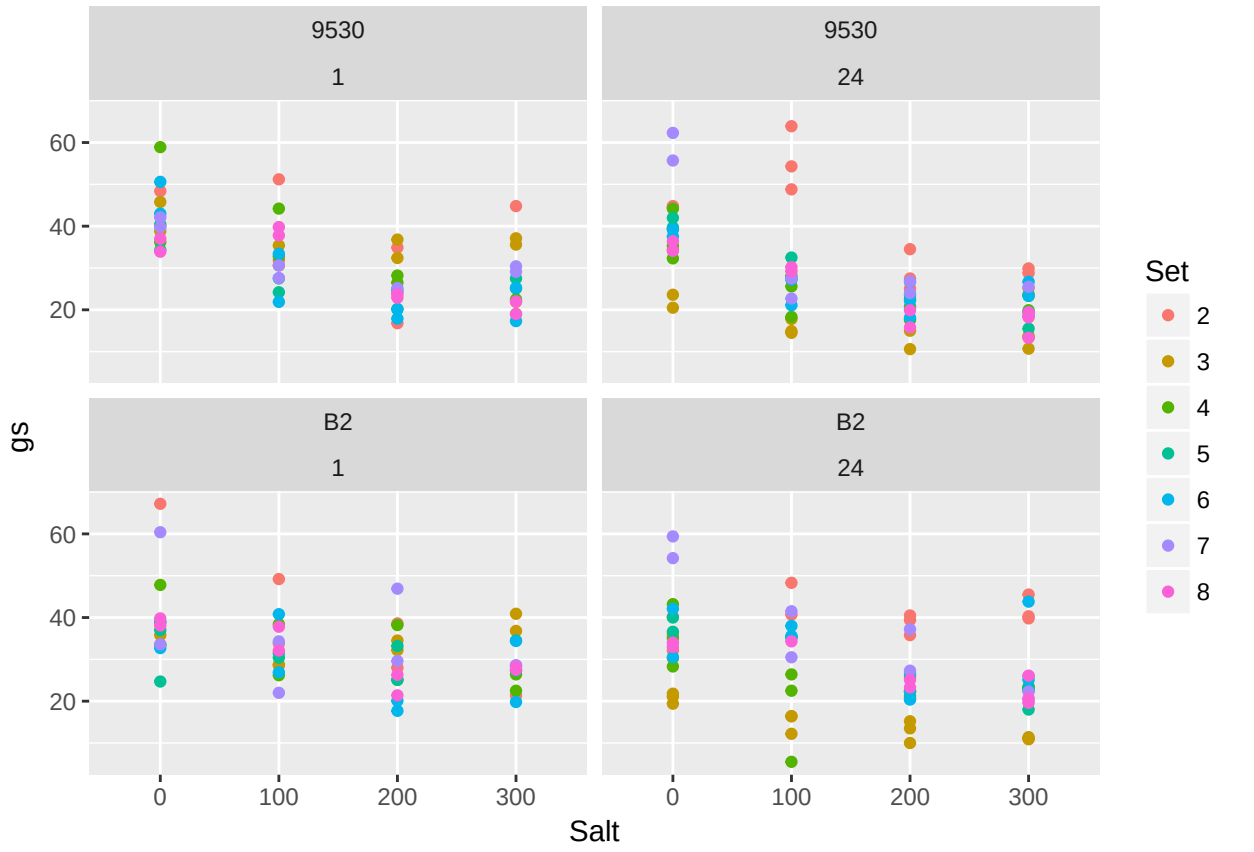
```
df = na.omit(df)

ggplot(df, aes(x = Salt, y = gs, fill = Genotype))+
  geom_boxplot()+
  facet_wrap(~Time)
```



For the 1 hour treatment, we see a declining trend until 200 mM where the values seem to be equal to those seen at 300 mM

There are a lot of outliers as we see in the boxplot above. But we'll go one step further and use the Set variable in the dataset to determine if a certain Set is more prone to outliers (the Set variable refers to each group of seeds sown in the same day and then taken to hydroponics on another same day, then measured altogether):

```
ggplot(df, aes(x = Salt, y = gs, col = Set))+
  geom_point()+
  facet_wrap(Genotype~Time)
```



First, we take a look at the 24 hour treatment data points. Comparing to the previous plot we can see that the Set 2 is where the extreme values at 100 and 300 mM stem from (and also the 200 ones but those won't affect the analysis that much). Considering this we'll choose to remove them from the analysis but only those, so the final N we'll be 6 for each combination of treatments.

```
df <- df[which(df$Set != '2'),]
```

Using the same criteria, we don't consider necessary to remove any further data points under the 1 hour treatment, we consider them to be genuine but extreme data points, not data input errors. Those in the 24 hour are considered to be genuine also, but we're not interested in dealing with such extreme cases, and as they stem from the same set, there was probably something special about that group or the experimental conditions during its measurement.

Finally, we take the mean of the response variable for each set, since the number of replicates was not the same for each combination of factors (although the number of independent experiments was).

```
df <- df %>% group_by(Set, Genotype, Time, Salt) %>% summarize(gs = mean(gs))
```

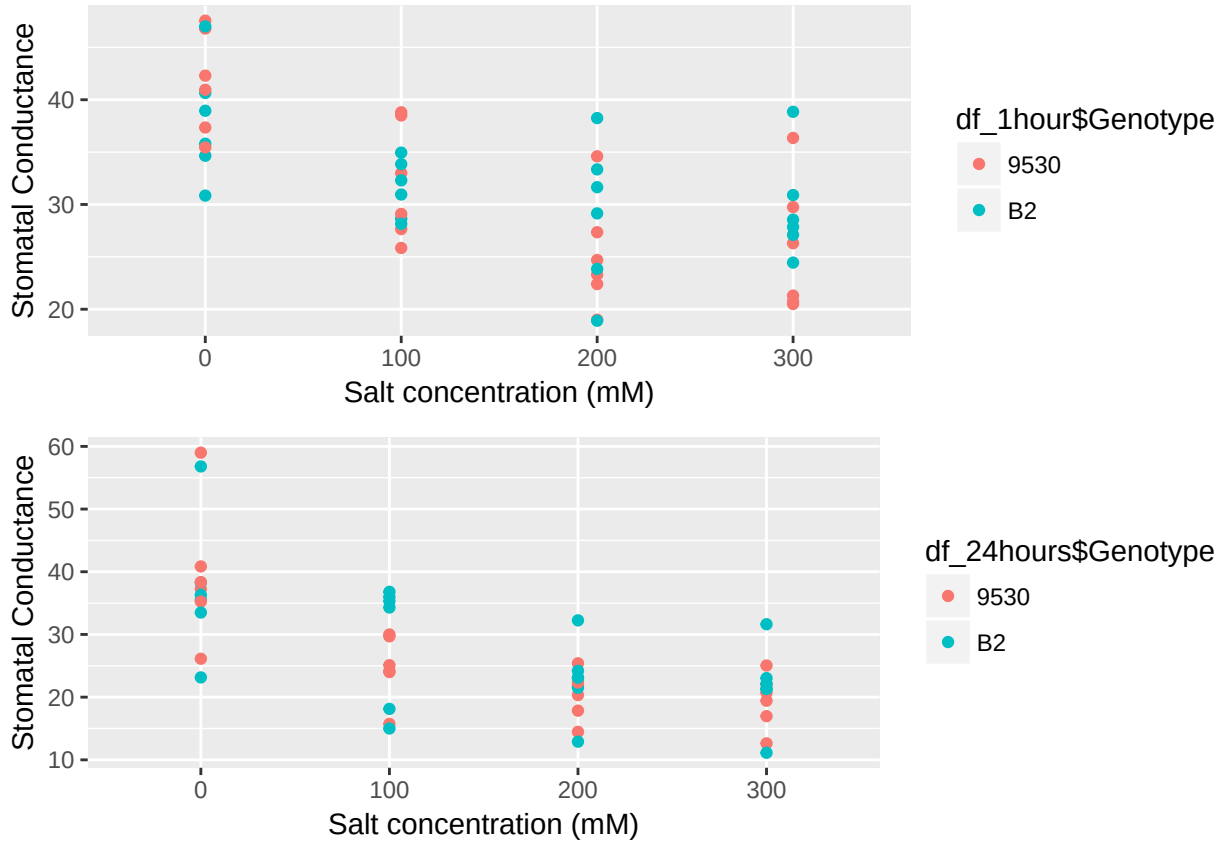
Time and salt treatment effect

Now we will play a closer look to each time treatment:

Using the `qplot` function from the ggplot2 package we can take a look at the data points when plants were subjected to 1 hour and 24 hours of salt treatment:

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
```

```
##
##     combine
```



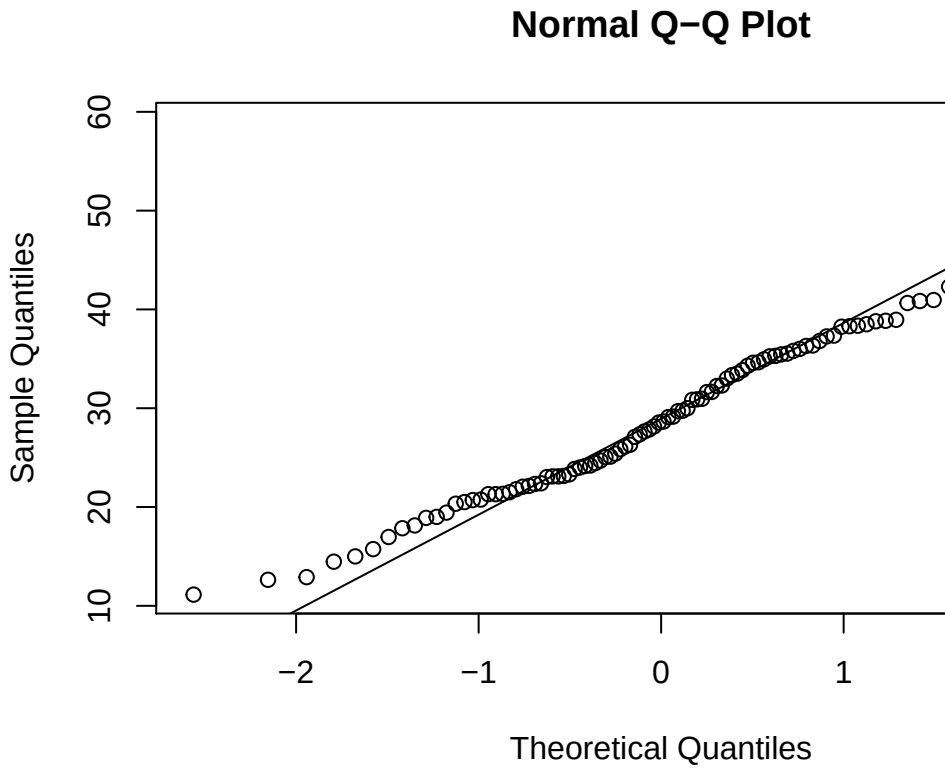
From both plots we can see that there is a declining trend for the stomatal conductance as salt increases (a common mechanism in most plants of closing stomata to avoid water loss) but in neither case seems to be a difference between genotypes (except maybe for 100 mM under 24 hours treatment).

Statistical Analysis and Hypothesis testing

Considering the variable type we are studying (a quantitative one) we will consider the assumptions we must make before modelling the data to determine if there are any statistical significant differences between treatments

Normal distribution

First we will use a qqplot to determine if our assumption of normal distribution is a plausible one, with the



`qqnorm()` and `qqline()` built-in functions:

We can see that for most of the data is a very plausible one but towards the higher quantiles the data departs from normality. For a discussion on how to interpret QQplots, this discussion results very useful.

Project still unfinished.