(a) Load the data into R, and create a data frame `water.df` to contain the data. Create a screen shot of the data frame after using the `View` command to indicate successful upload.

Use the command `summary(water.df)` to get a general overview of the contained information. Did the researchers collect the same amount of measurements over all seasons? Do we have missing data? Is the removal of observations with missing data necessary for exploratory analysis? Based on the quartiles what is remarkable about the algae proportions?

Data can be read in using the `read.csv` or `read.table` commands or simply using the drop down menu in RStudio. Fig. 1 shows a screenshot of the loaded data.

![Figure 1: The data frame water.df, a pairs plot of the chemical data, and a partial of the summary information for the data frame.](image)

Using the summary command on the data frame yields

```r
> summary(water.df)

season size velocity C1  C2
spring:53 small 71 low 33 Min. 5.600 Min. 1.500
summer:45 medium 84 medium 83 1st Qu. 7.700 1st Qu. 7.725
autumn:40 large 45 high 84 Median 8.060 Median 9.800
```

1
The summary output shows that the researchers do not have equal readings over all seasons. There are missing data in the chemicals, in particular C3 and C8 are missing 10 and 12 values respectively. Removing data is not necessary in exploratory analysis when done in R as it just omits the values from visualisations. The proportions indicate that the algal composition seems to be quite diverse. In particular, algae A4 and A6 are observed in less than 50% of the samples though in some samples A6 is the main alga contributing 77.6% of the overall algal diversity.

(b) We want to investigate the factor level combinations in the data. Create a mosaic plot providing a combined overview of the three categories. Are the combinations approximately equally represented? Can we make out any interactions of the different levels? Include the code you used to create the mosaic plot.

```r
water.table <- table(season=water.df$season,size=water.df$size,
```
velocity=water.df$velocity)
mosaicplot(water.table,main="",col=c("tomato","lightgoldenrodyellow","steelblue"))

The mosaic plot is shown in Fig. 2. We see that the researchers have tested rivers quite unequally, which could be due to the fact that the levels are quite clearly interacting. E.g., small rivers only have a low velocity during winter, and large rivers tend seldom have a high velocity.

(c) Use pairs plots to look at the chemical and algal data, respectively. Note any remarkable relationship between variables. Are there any outlier observations in the data? If so, provide means to identify these observations.

I am using my own interpretation of the pairs plot, called mypairs here. Since this function also computes the correlation between pairs of variables, we need to deal with the NA’s. I assume that an NA for a chemical concentration means it has not been detected, i.e. its concentration is likely to be 0.

water.df[is.na(water.df)] <- 0
mypairs(water.df[,4:11],histcol="tomato",pairscol="steelblue")
mypairs(water.df[,12:18],histcol="tomato",pairscol="steelblue")
Fig. 3 and 4 shows the respective pairs plots. We see that for the chemicals C6 and C7 appear to be strongly positively correlated ($\rho = .91$). Further, C1, C3 and C4 have remarkable outliers. The outlier for C1 occurs because of the replacement of an NA. We obtain the information about the outliers using the following commands.

```
> water.df[which(water.df$C1<3|water.df$C3>300|water.df$C4>20),]
season size velocity C1  C2  C3  C4  C5  C6  C7  C8  A1  A2
48 winter small    low 0.0 12.6  9.000 0.230  10  5  6   1.1 35.5  0.0
134 spring medium  medium 7.9  8.3 391.500  6.045  380 173 317  5.5  2.4 1.7
153 autumn medium  high 7.3 11.8 44.205 45.650  64 44  34  53.1  2.2  0.0
  A3  A4  A5  A6  A7
48  0.0  0.0  0.0  0.0  0.0
134  4.2  8.3  1.7  0.0  2.4
153  0.0 1.2  5.9  77.6  0.0
```

Figure 3: Pairs plot of chemicals.
For the algal proportions we observe that all algae distributions (diagonal) are gamma shaped, meaning most observations have only a small proportion of the respective alga but few have larger proportions. The pairs plots indicate that except for A5 and A6 ($\varrho = .4$) there seems to be no mixing of algal species. For any pair of algal species and any sample either neither is observed of high proportion or one of them, but rarely both. This seems to suggest that the algal communities are of relatively low diversity. There are no obvious outliers in the algal data.

Figure 4: Pairs plot of algae.

(d) We finally want to explore the relationship of chemicals and algae conditional on the given categories. How many panels would our table have, if we would include all three categorical variables? Select two of the three categorical variables as conditioning variables. As relationship variables randomly select one chemical and one alga. Build a trellis plot using these variables and describe your findings.
There are three categories; season has 4 levels, size and velocity have 3 levels each. Hence a trellis plot would consist of 36 cells. We reduce ourselves to 12 cells by selecting season and size as conditioning variables. We also randomly select an index for the chemical and alga using the following commands

```
chem.number <- sample(4:11,1)
alga.number <- sample(12:18,1)
```

We then build the trellis plot using the following command

```
chem.name <- colnames(water.df)[chem.number]
alga.name <- colnames(water.df)[alga.number]
xyplot(water.df[,alga.number]~water.df[,chem.number]|water.df[,1]*water.df[,2],
xlab=chem.name,ylab=alga.name,groups=water.df[,3],pch=16)
```

Fig. 5 shows the output. Using the groups-option we were able to include the velocity after all. My code randomly selected C8 and A7. There does not appear to be a clear relationship between chemical concentration and algal proportion. However, it appears that high A7 concentrations are observed in medium rivers rather than small or large, and that the chemical has a consistently low concentration in small rivers outside of winter.

Figure 5: Trellis plot relating a chemical to an alga conditional on season and size of river.