Introduction to R

Session 1 – Introduction

1. Using R as a calculator

1. Find the values of:

   (a) $1 + 4$

   ```
   1+4
   ```

   ```
   ## [1] 5
   ```

   (b) $2^3 + \frac{4}{\sqrt{34}}$

   ```
   2^3 + 4/sqrt(34)
   ```

   ```
   ## [1] 8.685994
   ```

   (c) $\log 30$

   ```
   log(30)
   ```

   ```
   ## [1] 3.401197
   ```

   (d) $\log_{10} 30$

   ```
   log(30)
   ```

   ```
   ## [1] 3.401197
   ```

   (e) $|-2|$ (Hint: $|x|$ denotes the *absolute value* of x. Search on Google if you're unsure about which `R` function to use.)

   ```
   abs(-2)
   ```

   ```
   ## [1] 2
   ```

2. Now open Rstudio, open an `R` script by clicking `File` → `New` → `R script`.
3. Save this script by clicking `File` → `Save As...`.
4. Select a directory/location and save the script.
5. Copy and paste (or just write out again) the code you used for question 1a – 1e into the script.
6. You can now submit your script line-by-line using Ctrl + Enter. You can also highlight the code you want to evaluate and press Ctrl + Enter. This will send the highlighted code in the script directly into the console.
7. From now on, type all of your code in your `R` script and submit it to the `R` Console using Ctrl + Enter.

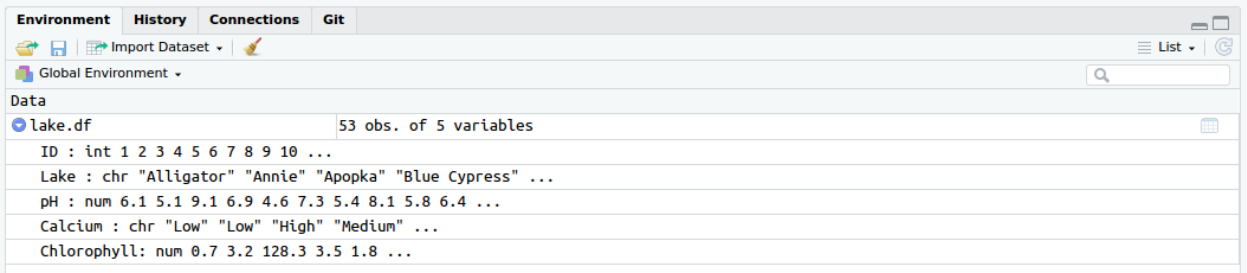

Figure 1: Environment tab in RStudio showing the properties for the `lake.df` data set

2. Reading data into R

1. `Lake.csv` contains data on 53 different lakes in Florida. The variable names and what has been measured are presented below.

   - `ID`: ID number of the lake
   - `Lake`: Name of the lake
   - `pH`: pH value
   - `Calcium`: concentration of Calcium
   - `Chlorophyll`: concentration of Chlorophyll (mg/L)

2. Read the CSV file into `R`, saving it as an object named `lake.df`. Make sure you don't read in the strings as factors (use `stringsAsFactors = FALSE`).

   Note that in the original CSV file, we have 'missing' cells to denote missing values. These cells actually contain "". We can code these cells as missing (NA's) in `R` by using the `na.strings = ""` argument.

```
lake.df = read.csv("location of your folder/Lake.csv",
                   stringsAsFactors = FALSE,
                   na.strings = "")
```

3. Use `str()` and `head()` to look at some of the properties of the dataset you have just read into `R`. *Always* perform this important step to check that your dataset is as it should be.

```
str(lake.df)
```

```
## 'data.frame':    53 obs. of  5 variables:
##  $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Lake       : chr  "Alligator" "Annie" "Apopka" "Blue Cypress" ...
##  $ pH         : num  6.1 5.1 9.1 6.9 4.6 7.3 5.4 8.1 5.8 6.4 ...
##  $ Calcium    : chr  "Low" "Low" "High" "Medium" ...
##  $ Chlorophyll: num  0.7 3.2 128.3 3.5 1.8 ...
```

```
head(lake.df)
```

```
##   ID         Lake  pH Calcium Chlorophyll
## 1  1    Alligator 6.1     Low         0.7
## 2  2        Annie 5.1     Low         3.2
## 3  3       Apopka 9.1    High       128.3
## 4  4 Blue Cypress 6.9  Medium         3.5
## 5  5        Brick 4.6     Low         1.8
## 6  6       Bryant 7.3    <NA>        44.1
```

Note that RStudio makes this step really easy. Check out the Environment tab in RStudio (see Figure 1).

4. Calculate the mean and standard deviation of both `pH` and `Chlorophyll`.

```r
mean(lake.df$pH, na.rm = TRUE)
```

```
## [1] 6.588
```

```r
mean(lake.df$Chlorophyll, na.rm = TRUE)
```

```
## [1] 23.83
```

```r
sd(lake.df$pH, na.rm = TRUE)
```

```
## [1] 1.317332
```

```r
sd(lake.df$Chlorophyll, na.rm = TRUE)
```

```
## [1] 31.52909
```

5. Check out what `summary()` does by running `summary(lake.df$pH)`.

```r
summary(lake.df$pH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   3.600   5.800   6.850   6.588   7.450   9.100       3
```

6. Check the frequency of each `Calcium` concentration.

```r
table(lake.df$Calcium)
```

```
##
##   High    Low Medium
##     15     17     19
```

7. Turn the frequency table from above into a table of proportions, keep only 2 decimal places.

```r
round(prop.table(table(lake.df$Calcium)) * 100, 1)
```

```
##
##   High    Low Medium
##   29.4   33.3   37.3
```