Homework Assignment 1

Trevor Klar and Blaine Quackenbush

October 08, 2021

We can start the analysis by loading into R the data from the "algaeBloom.txt" file (the training data, i.e. the data that will be used to obtain the predictive models). To read the data from the file it is sufficient to issue the following command:

```
algae <- read_table(
  "algaeBloom.txt",
  col_names=c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4',
              'oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),
  na="XXXXXXX")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   season = col_character(),
##   size = col_character(),
##   speed = col_character(),
##   mxPH = col_double(),
##   mnO2 = col_double(),
##   Cl = col_double(),
##   NO3 = col_double(),
##   NH4 = col_double(),
##   oPO4 = col_double(),
##   PO4 = col_double(),
##   Chla = col_double(),
##   a1 = col_double(),
##   a2 = col_double(),
##   a3 = col_double(),
##   a4 = col_double(),
##   a5 = col_double(),
##   a6 = col_double(),
##   a7 = col_double()
## )
```

```
glimpse(algae)
```

```
## Rows: 200
## Columns: 18
## $ season <chr> "winter", "spring", "autumn", "spring", "autumn", "winter", "su~
## $ size   <chr> "small", "small", "small", "small", "small", "small", "small", ~
## $ speed  <chr> "medium", "medium", "medium", "medium", "medium", "high", "high~
## $ mxPH   <dbl> 8.00, 8.35, 8.10, 8.07, 8.06, 8.25, 8.15, 8.05, 8.70, 7.93, 7.7~
## $ mnO2   <dbl> 9.8, 8.0, 11.4, 4.8, 9.0, 13.1, 10.3, 10.6, 3.4, 9.9, 10.2, 11.~
## $ Cl     <dbl> 60.80, 57.75, 40.02, 77.36, 55.35, 65.75, 73.25, 59.07, 21.95, ~
## $ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990, 0.886,~
```

```
## $ NH4    <dbl> 578.00, 370.00, 346.67, 98.18, 233.70, 430.00, 110.00, 205.67, ~
## $ oPO4   <dbl> 105.00, 428.75, 125.67, 61.18, 58.22, 18.25, 61.25, 44.67, 36.3~
## $ PO4    <dbl> 170.00, 558.75, 187.06, 138.70, 97.58, 56.67, 111.75, 77.43, 71~
## $ Chla   <dbl> 50.000, 1.300, 15.600, 1.400, 10.500, 28.400, 3.200, 6.900, 5.5~
## $ a1     <dbl> 0.0, 1.4, 3.3, 3.1, 9.2, 15.1, 2.4, 18.2, 25.4, 17.0, 16.6, 32.~
## $ a2     <dbl> 0.0, 7.6, 53.6, 41.0, 2.9, 14.6, 1.2, 1.6, 5.4, 0.0, 0.0, 0.0, ~
## $ a3     <dbl> 0.0, 4.8, 1.9, 18.9, 7.5, 1.4, 3.2, 0.0, 2.5, 0.0, 0.0, 0.0, 2.~
## $ a4     <dbl> 0.0, 1.9, 0.0, 0.0, 0.0, 0.0, 3.9, 0.0, 0.0, 2.9, 0.0, 0.0, 0.0~
## $ a5     <dbl> 34.2, 6.7, 0.0, 1.4, 7.5, 22.5, 5.8, 5.5, 0.0, 0.0, 1.2, 0.0, 1~
## $ a6     <dbl> 8.3, 0.0, 0.0, 0.0, 4.1, 12.6, 6.8, 8.7, 0.0, 0.0, 0.0, 0.0, 0.~
## $ a7     <dbl> 0.0, 2.1, 9.7, 1.4, 1.0, 2.9, 0.0, 0.0, 0.0, 1.7, 6.0, 1.5, 2.1~
```

1. ***Descriptive summary statistics*** (10 pts in total) Given the lack of further information on the problem domain, it is wise to investigate some of the statistical properties of the data, so as to get a better grasp of the problem. It is always a good idea to start our analysis with some kind of exploratory data analysis. A first idea of the statistical properties of the data can be obtained through a summary of its descriptive statistics.

a. (2 pts) Count the number of observations in each size using `summarise()` in `dplyr`.
b. (1 pts) Are there missing values? (2 pts) Calculate the mean and variance of each chemical (Ignore a1 through a7). (1 pts) What do you notice about the magnitude of the two quantities for different chemicals?
c.

---

* Estimator of standard deviation (SD):

    $$\text{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i} - \bar{x})^2}$$

    ```r
    s.d <- function(x){
        n <- length(x) # Sample size
        s2 <- sum((x - mean(x))^2)/(n-1) # sample variance
        s.d <- sqrt(s2) # sample standard deviation
        return(s.d)
    }
    ```

* Estimator of mean absolute deviation (MAD):

    $$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_{i} - \bar{x}|$$

    ```r
    mean.abs.d <- function(x){
        n <- length(x) # Sample size
        m <- sum(abs(x - mean(x)))/n # mean average deviation
        return(m)
    }
    ```