Homework 1 PSTAT 131

Nicholas Axl Andrian

October 17, 2023

```
#install.packages("tidyverse")
#install.packages("dplyr")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
```

Reading in the dataset

1. Descriptive summary statistics

```
algae <- read_table2("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")
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## i Please use `read_table()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
##
## -- 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> 60.800, 57.750, 40.020, 77.364, 55.350, 65.750, 73.250, 59.067,~
## $ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990, 0.886,~
## $ NH4    <dbl> 578.000, 370.000, 346.667, 98.182, 233.700, 430.000, 110.000, 2~
## $ oPO4   <dbl> 105.000, 428.750, 125.667, 61.182, 58.222, 18.250, 61.250, 44.6~
## $ PO4    <dbl> 170.000, 558.750, 187.057, 138.700, 97.580, 56.667, 111.750, 77~
## $ 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~
```

(a)

```
season_count <- algae %>%
  group_by(season) %>%
  summarise(count = n())

season_count
```

```
## # A tibble: 4 x 2
##   season count
##   <chr>  <int>
## 1 autumn    40
## 2 spring    53
## 3 summer    45
## 4 winter    62
```

(b)

```
total_null <- function(x)
  return(sum(is.na(x)))

null_values <- sapply(algae, total_null)
null_values
```

```
## season   size  speed   mxPH   mnO2     Cl    NO3    NH4   oPO4    PO4   Chla
##      0      0      0      1      2     10      2      2      2      2     12
##     a1     a2     a3     a4     a5     a6     a7
##      0      0      0      0      0      0      0
```

```
chem_means <- colMeans(select(algae, mxPH,mnO2,Cl,NO3,NH4,
oPO4,PO4,Chla), na.rm=TRUE)
chem_means
```

```
##       mxPH       mnO2         Cl        NO3        NH4       oPO4        PO4
##   8.011734   9.117778  43.636279   3.282389 501.295828  73.590596 137.882101
##       Chla
##  13.971197
```

```
chem_vars <- sapply(select(algae, mxPH,mnO2,Cl,NO3,NH4,
oPO4,PO4,Chla), var, na.rm = TRUE)
chem_vars
```

```
##         mxPH         mnO2           Cl          NO3          NH4         oPO4
## 3.579693e-01 5.718089e+00 2.193172e+03 1.426176e+01 3.851585e+06 8.305850e+03
##          PO4         Chla
## 1.663938e+04 4.200827e+02
```

I noticed that there is pretty high variance when compared to the mean. This suggests a wide spread/dispersion in the data values due to high variability.

(c)

```
calc_median <- function(x){
  med <- median(x, na.rm = TRUE)
  return(med)
}

calc_mad <- function(x) {
  med <- median(x, na.rm = TRUE)
  mad_val <- abs(x - med)
  mad_result <- median(mad_val, na.rm = TRUE)

  return(mad_result)
}
```

```
chem_med <- sapply(select(algae, mxPH,mnO2,Cl,NO3,NH4,
oPO4,PO4,Chla), calc_median)
chem_med
```

```
##     mxPH     mnO2       Cl      NO3      NH4     oPO4      PO4     Chla
##   8.0600   9.8000  32.7300   2.6750 103.1665  40.1500 103.2855   5.4750
```

```
chem_mad <- sapply(select(algae, mxPH,mnO2,Cl,NO3,NH4,
oPO4,PO4,Chla), calc_mad)
chem_mad
```

```
##    mxPH    mnO2      Cl     NO3     NH4    oPO4     PO4    Chla
##  0.3400  1.3850 22.4265  1.4650 75.2850 29.7085 82.5045  4.5000
```

testing the mad function instead, seeing how it is different since I believe R's mad function uses scaling

```
true_chem_mad <- sapply(select(algae, mxPH, mnO2, Cl, NO3, NH4, oPO4, PO4, Chla), mad, na.rm = TRUE)
print(chem_mad)
```

```
##    mxPH    mnO2      Cl     NO3     NH4    oPO4     PO4    Chla
##  0.3400  1.3850 22.4265  1.4650 75.2850 29.7085 82.5045  4.5000
```

I realised that the median and mad values are way lower than the mean and var. I assume this is because they are less sensitive to outliers. I also noticed that the chemicals with less null values happened to be similar to each other in terms of mean/var and med/mad.

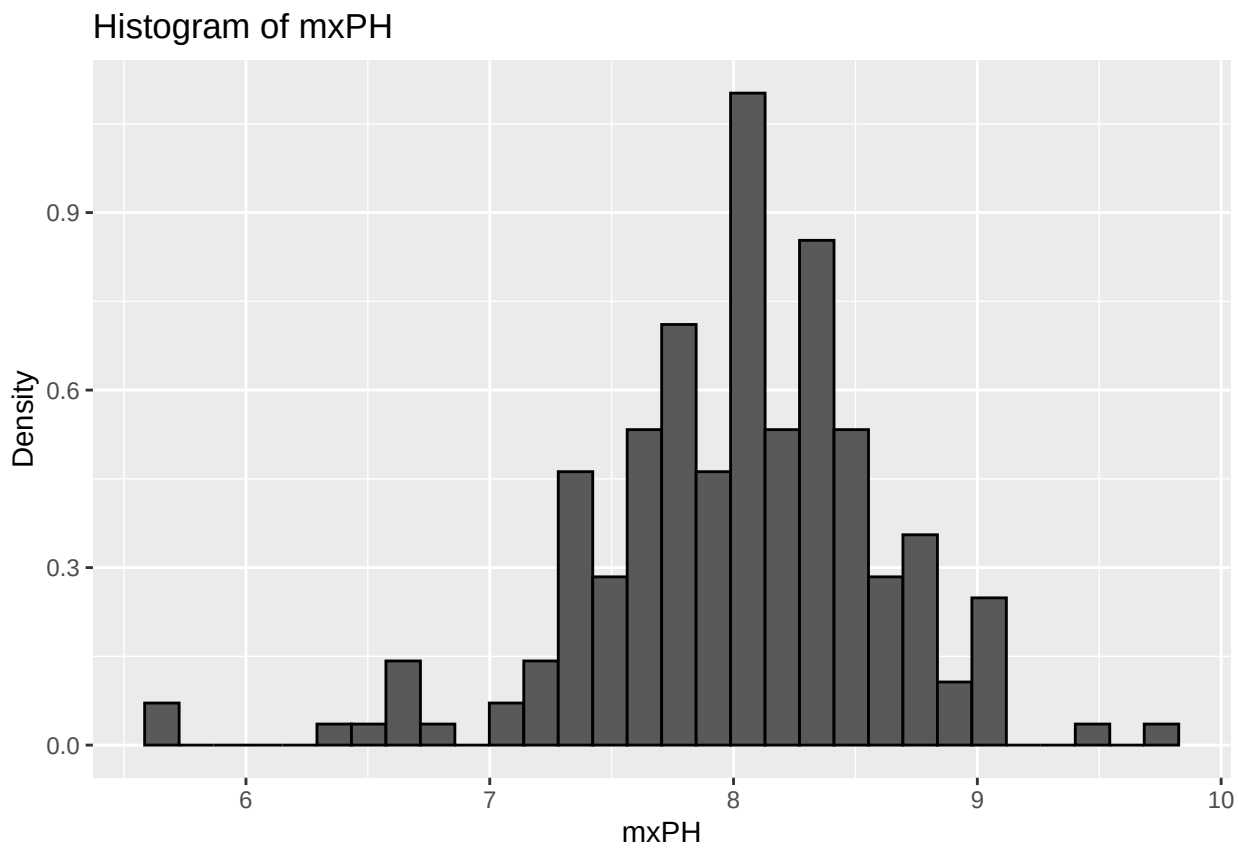
2. Data Visualization

(x)

```
ggplot(algae, aes(x = mxPH)) +
  geom_histogram(aes(y = after_stat(density)), color = "black") +
  labs(title = "Histogram of mxPH", x = "mxPH", y = "Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```

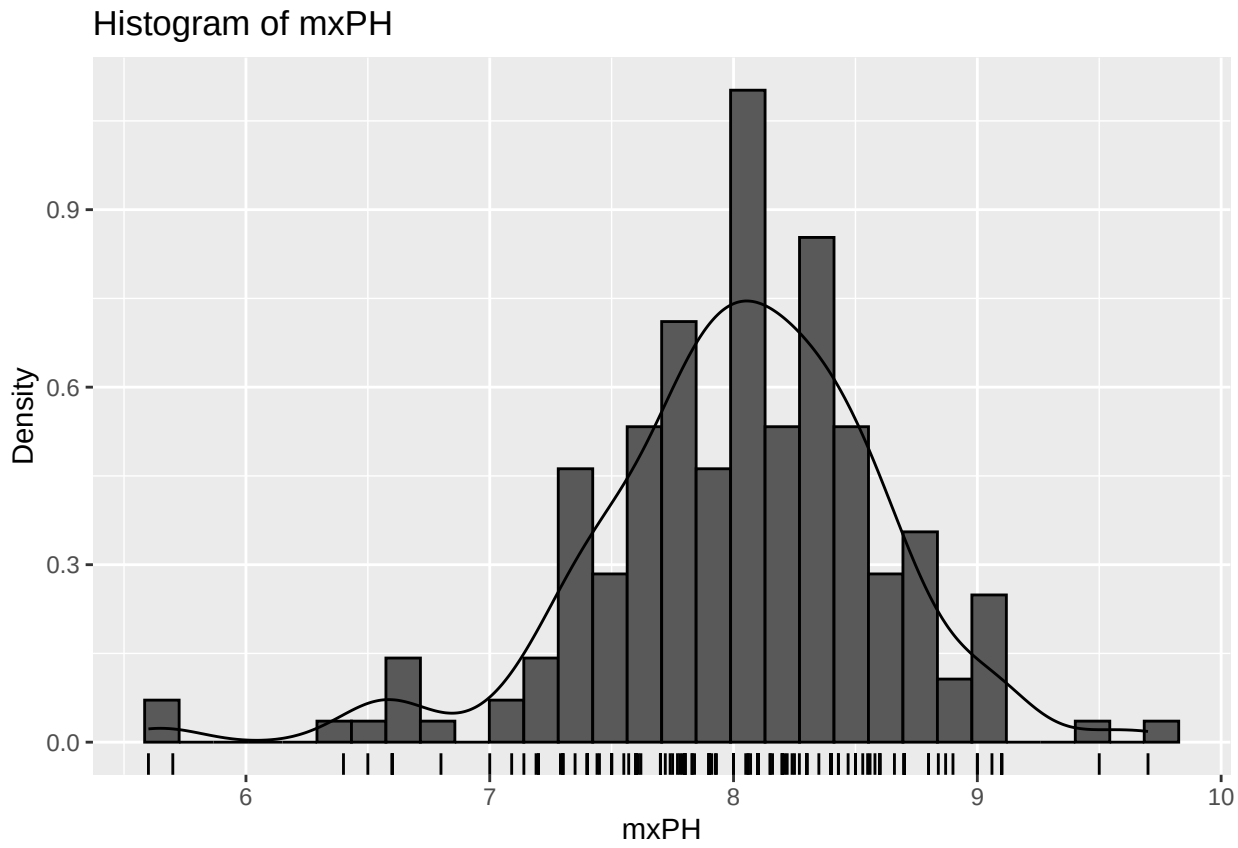


It does not look very skewed. The shape itself looks symmetrical if you ignore the small observation on the most left. Otherwise if you include the whole picture you can say that it is left skewed since it extends a bit more to the left. (b)

```
ggplot(algae, aes(x = mxPH)) +
  geom_histogram(aes(y = after_stat(density)), color = "black") +
  geom_density() +
  geom_rug() +
  labs(title = "Histogram of mxPH", x = "mxPH", y = "Density")
```

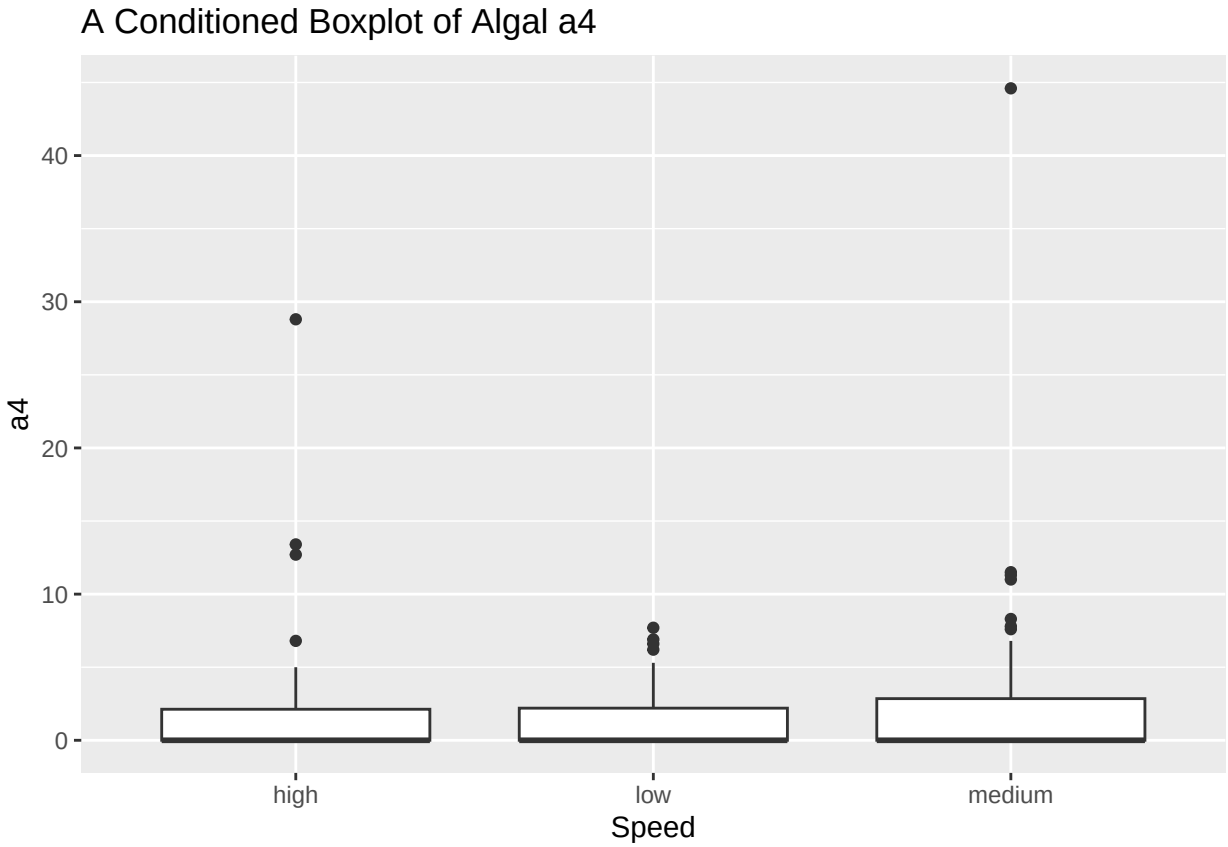
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
## Warning: Removed 1 rows containing non-finite values (`stat_density()`).
```



This makes it look more left skewed (c)

```
ggplot(algae, aes(x = speed, y = a4)) +
  geom_boxplot() +
  labs(title = "A Conditioned Boxplot of Algal a4", x = "Speed", y = "a4")
```



Majority of the observations are very close to 0

3 (a)

```
summary(algae)
```

```
##     season              size              speed               mxPH
##  Length:200         Length:200         Length:200         Min.   :5.600
##  Class :character   Class :character   Class :character   1st Qu.:7.700
##  Mode  :character   Mode  :character   Mode  :character   Median :8.060
##                                                           Mean   :8.012
##                                                           3rd Qu.:8.400
##                                                           Max.   :9.700
##                                                           NA's   :1
##       mnO2              Cl               NO3              NH4
##  Min.   : 1.500   Min.   :  0.222   Min.   : 0.050   Min.   :    5.00
##  1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296   1st Qu.:   38.33
##  Median : 9.800   Median : 32.730   Median : 2.675   Median :  103.17
##  Mean   : 9.118   Mean   : 43.636   Mean   : 3.282   Mean   :  501.30
##  3rd Qu.:10.800   3rd Qu.: 57.824   3rd Qu.: 4.446   3rd Qu.:  226.95
##  Max.   :13.400   Max.   :391.500   Max.   :45.650   Max.   :24064.00
##  NA's   :2        NA's   :10        NA's   :2        NA's   :2
##       oPO4             PO4              Chla              a1
##  Min.   :  1.00   Min.   :  1.00   Min.   :  0.200   Min.   : 0.00
##  1st Qu.: 15.70   1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50
##  Median : 40.15   Median :103.29   Median :  5.475   Median : 6.95
##  Mean   : 73.59   Mean   :137.88   Mean   : 13.971   Mean   :16.92
##  3rd Qu.: 99.33   3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80
```

```
##  Max.   :564.60   Max.   :771.60   Max.   :110.456   Max.   :89.80
##  NA's   :2        NA's   :2        NA's   :12
##        a2               a3               a4               a5
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 3.000   Median : 1.550   Median : 0.000   Median : 1.900
##  Mean   : 7.458   Mean   : 4.309   Mean   : 1.992   Mean   : 5.064
##  3rd Qu.:11.375   3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500
##  Max.   :72.600   Max.   :42.800   Max.   :44.600   Max.   :44.400
##
##        a6               a7
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 1.000
##  Mean   : 5.964   Mean   : 2.495
##  3rd Qu.: 6.925   3rd Qu.: 2.400
##  Max.   :77.600   Max.   :31.600
##
```

```
null_values
```

```
## season   size  speed   mxPH   mnO2     Cl    NO3    NH4   oPO4    PO4   Chla
##      0      0      0      1      2     10      2      2      2      2     12
##     a1     a2     a3     a4     a5     a6     a7
##      0      0      0      0      0      0      0
```

```
sum(is.na(algae))
```

```
## [1] 33
```

(b)

```
algae.del <- algae %>%
  filter(complete.cases(.))

sum(is.na(algae.del))
```

```
## [1] 0
```

```
nrow(algae.del)
```

```
## [1] 184
```