Homework Assignment

Parker Reedy

October 16, 2023

```
## -- 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 1.9.3     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 error
```

```
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. (a). 40 observations in Autumn, 53 in spring, 45 in summer, and 62 in winter

```
algae %>% group_by(season) %>% summarize(n = n())
```

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

(b).

```
colSums(is.na(algae))
```

```
## 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
```

There are missing values in this data frame.

```
sapply(algae[4:11], function(x) c(mean=mean(x, na.rm=TRUE), var=var(x, na.rm=TRUE)))
```

```
##       mxPH  mnO2      Cl    NO3       NH4    oPO4     PO4   Chla
## mean 8.012 9.118   43.64  3.282     501.3   73.59   137.9  13.97
## var  0.358 5.718 2193.17 14.262 3851584.7 8305.85 16639.4 420.08
```

I calculated the mean and variance of the chemicals after removing the missing values. The variance for chemicals such as NH4, oPO4, PO4, and Cl are extremely high compared to NO3 and Chlorophyll. NH4 clearly has the largest variance.

(c).

```
chem_median<- sapply(algae[5:11], function(x) c(median=median(x, na.rm=TRUE)))
chem_median
```

```
## mnO2.median   Cl.median  NO3.median  NH4.median oPO4.median  PO4.median
##       9.800      32.730       2.675     103.166      40.150     103.285
## Chla.median
##       5.475
```

```
MAD <- c()
for (x in 5:11){
  MAD <- append(MAD, mad(algae[x], na.rm=TRUE))
}

df <- data.frame(chem_median, MAD)
df
```

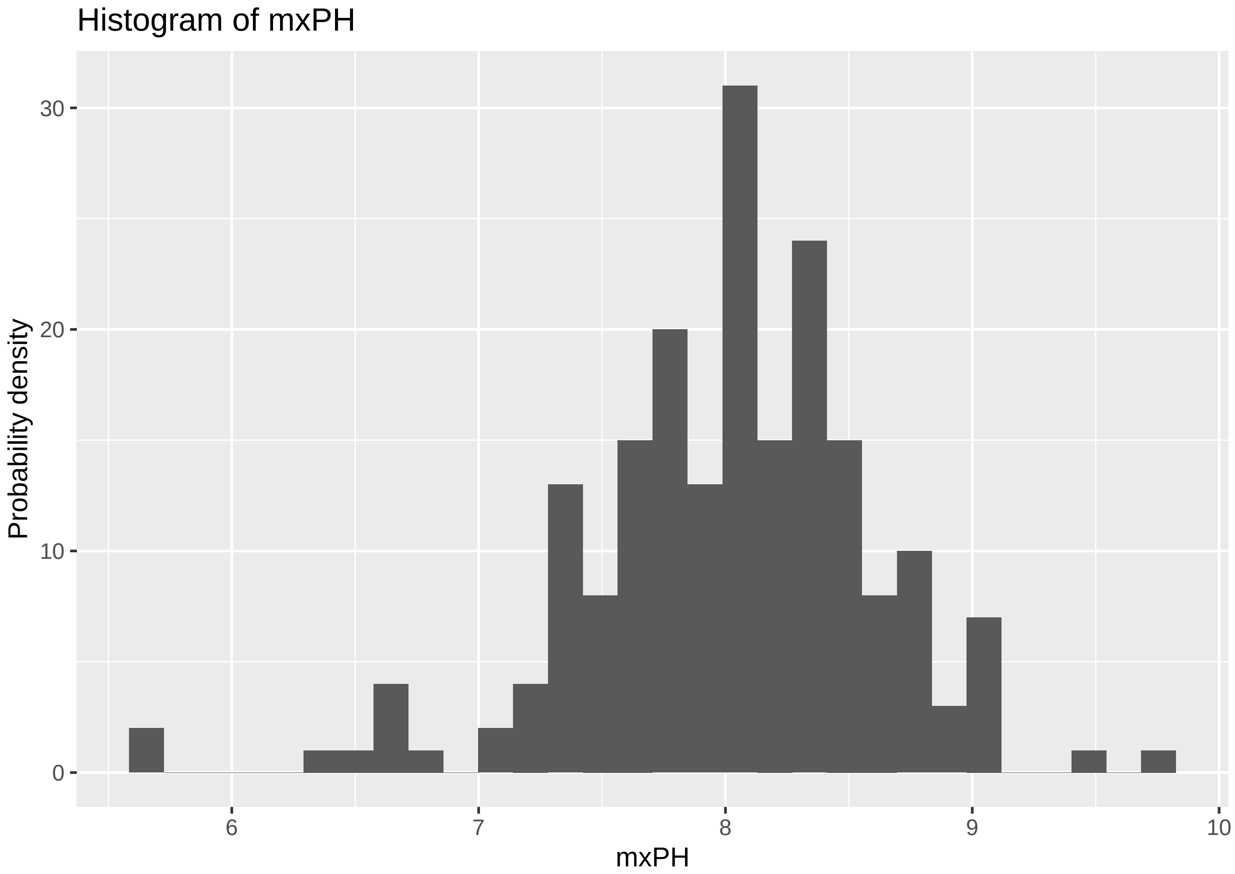
```
##             chem_median     MAD
## mnO2.median       9.800   2.053
## Cl.median        32.730  33.250
## NO3.median        2.675   2.172
## NH4.median      103.166 111.618
## oPO4.median      40.150  44.046
## PO4.median      103.285 122.321
## Chla.median       5.475   6.672
```

The Median & MAD are much more similar to each other than the Mean and Variance are. The variance is on a completely different magnitude than the Mean.

2. (a).

```
ggplot(algae, aes(mxPH), na.rm = TRUE) +
  geom_histogram(
    aes(y = after_stat(count)),
    bins=30
) + labs(
  title = "Histogram of mxPH", x = "mxPH", y = "Probability density"
)
```

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



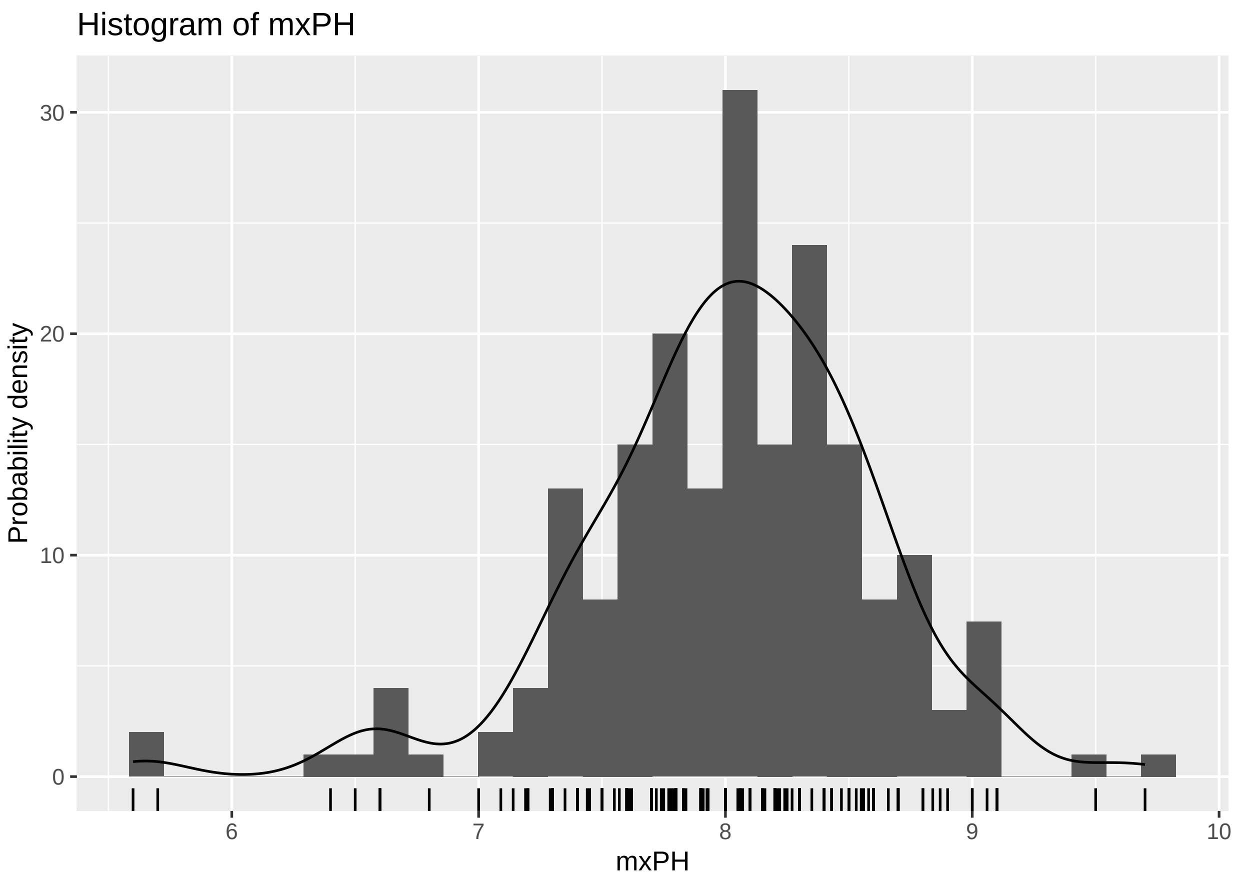
The distribution isn't skewed since there aren't enough observations concentrated in one direction to warrant labeling it as skewed.

(b).

```
ggplot(algae, aes(mxPH), na.rm = TRUE) +
  geom_histogram(
    aes(y = after_stat(count)),
    bins=30
) + labs(
  title = "Histogram of mxPH", x = "mxPH", y = "Probability density"
) + geom_density(aes(y = after_stat(density * 30)) # multiply density * the number of bins so the graph
) + geom_rug()
```

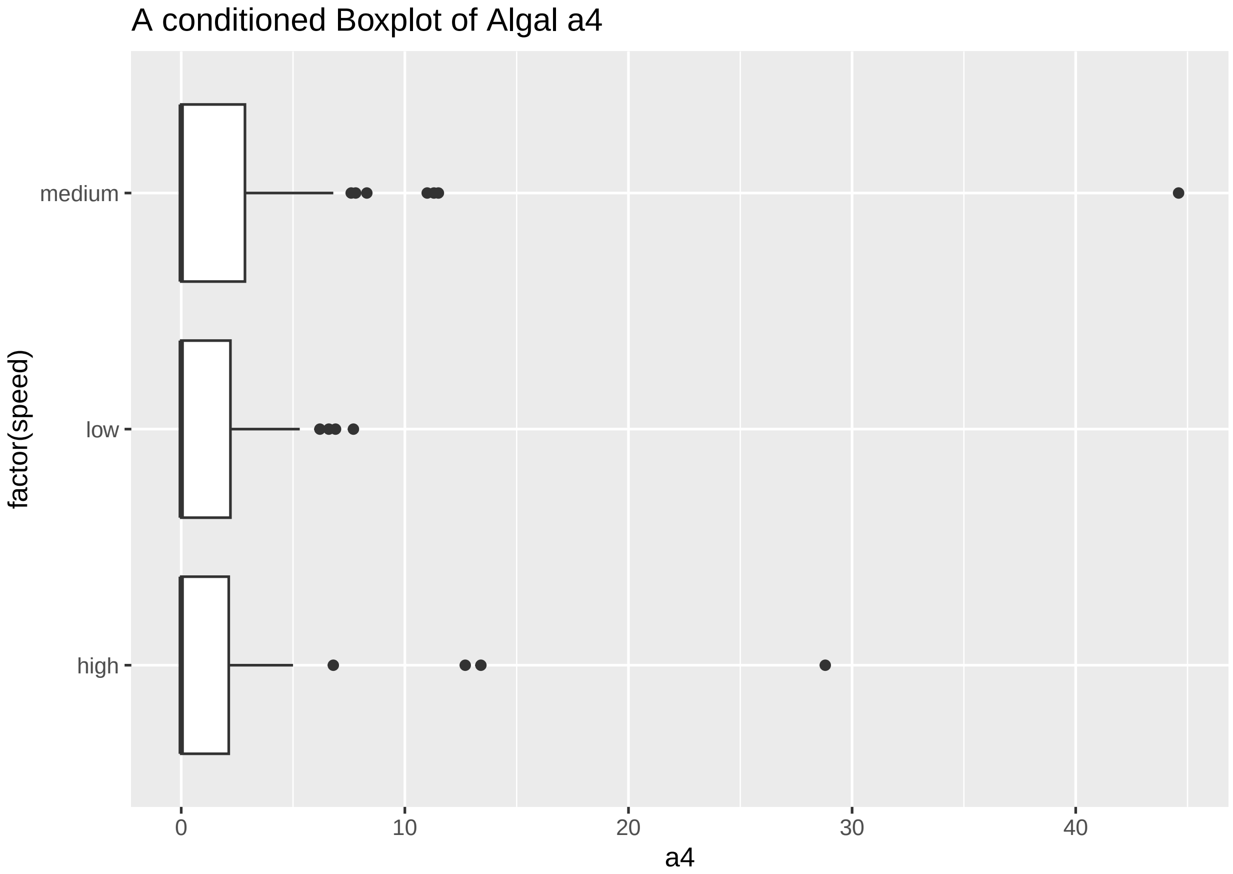
```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```

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



(c).

```
ggplot(algae, aes(a4, y = factor(speed)), na.rm = TRUE) +
  geom_boxplot() +
  labs(title = 'A conditioned Boxplot of Algal a4', ylab = 'speed')
```



I notice that there is not as much a4 present in slower rivers than there are in faster rivers.

3 (a).

```
sum(rowSums(is.na(algae)) > 0)
```

```
## [1] 16
```

```
colSums(is.na(algae))
```

```
## season   size  speed   mxPH   mn02     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
```

There are 16 observations with missing values in them. There is 1 missing value for mxPH, 2 for mn02, 10 for cl, 2 for NO3, 2 for NH4, 2 for oPO4, 2 for PO4, and 12 for Chla.

(b).

```
algae.del <- filter(algae, !is.na(mxPH & mn02 & Cl & NO3 & NH4 & oPO4 & PO4 & Chla))
all(complete.cases(algae.del))
```

```
## [1] TRUE
```

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

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

There are 184 observations remaining in algae.del

4 (a).

The terms $Var(\hat{f}(x_0))$ and $[Bias(f(\hat{x_0}))]^2$ are the reducible errors in the bias-variance tradeoff

the irreducible error is the $Var(\epsilon)$

(b).
bias-variance decomposition

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

If we take $\hat{f}(x_0) = E[Y|X = x_0]$ then $Var(x_0) = E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2]$ will be minimized

and $[Bias(\hat{f}(x_0))]^2 = [E[(\hat{f}(x_0)] - \hat{f}(x_0)]^2$ will also be minimized

leaving $Var(\epsilon)$ which can not be reduced because it is the random error.