Homework 1

Amy Nguyen

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

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")
glimpse(algae)
```

Problem 1

(a)

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

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

There are 200 total observations. More specifically for each season: * Autumn = 40 * Spring = 53 * Summer = 45 * Winter = 62

(b)

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

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

There are 33 missing values.

```
chemicals_mean <- algae %>%
  summarise_at(vars(mxPH:Chla), mean, na.rm=TRUE)
chemicals_mean
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  8.01  9.12  43.6  3.28  501.  73.6  138.  14.0
```

```
chemicals_var <- algae %>%
  summarise_at(vars(mxPH:Chla), var, na.rm=TRUE)
chemicals_var
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3       NH4  oPO4    PO4  Chla
##   <dbl> <dbl> <dbl> <dbl>     <dbl> <dbl>  <dbl> <dbl>
## 1 0.358  5.72 2193.  14.3 3851585. 8306. 16639.  420.
```

NH4, oPO4, and PO4 have very large variances and means in comparison to the other chemicals. The large variances indicate that their respective means may not be very useful.

(c)

```
chemicals_med <- algae %>%
  dplyr::select(mxPH:Chla) %>%
  summarise_all(function(z) median(z, na.rm=TRUE))
chemicals_med
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  8.06   9.8  32.7  2.68  103.  40.2  103.  5.48
```

```
chemicals_MAD <- algae %>%
  summarise_at(vars(mxPH:Chla), funs(mad), na.rm = TRUE)
chemicals_MAD
```

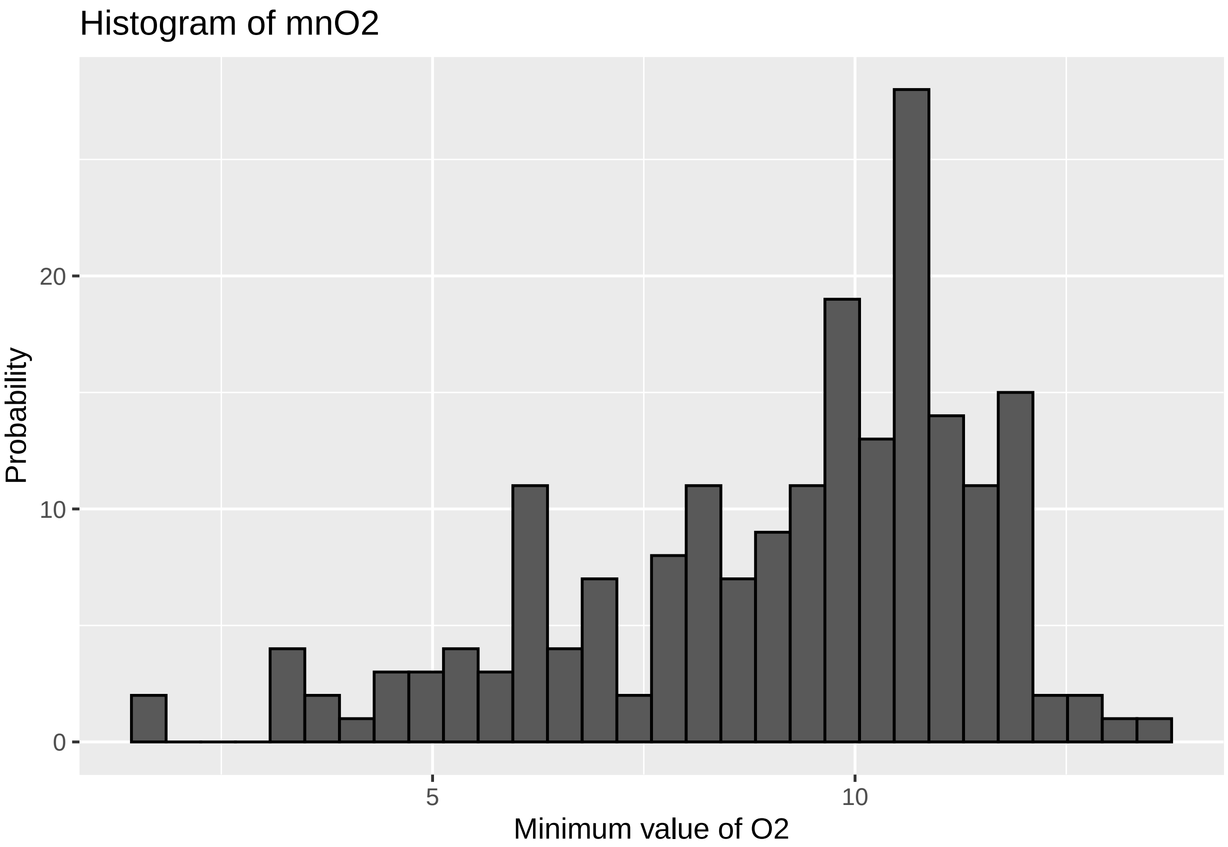
```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.504  2.05  33.2  2.17  112.  44.0  122.  6.67
```

The medians for the chemicals for the most part are smaller than the means. The MAD is for the most part, smaller in comparison to the variances. These differences can be explained by potential outliers that skewed the data.

Problem 2

(a)

```
algae %>%
  ggplot(aes(x=mnO2, stat = "density")) +
  geom_histogram(col = "black") +
  ggtitle("Histogram of mnO2") +
  labs(x = "Minimum value of O2", y = "Probability")
```



Based on the above histogram, it appears the distribution is left-skewed. The probability of obtaining a minimum O2 value increases as the minimum O2 value gets larger.

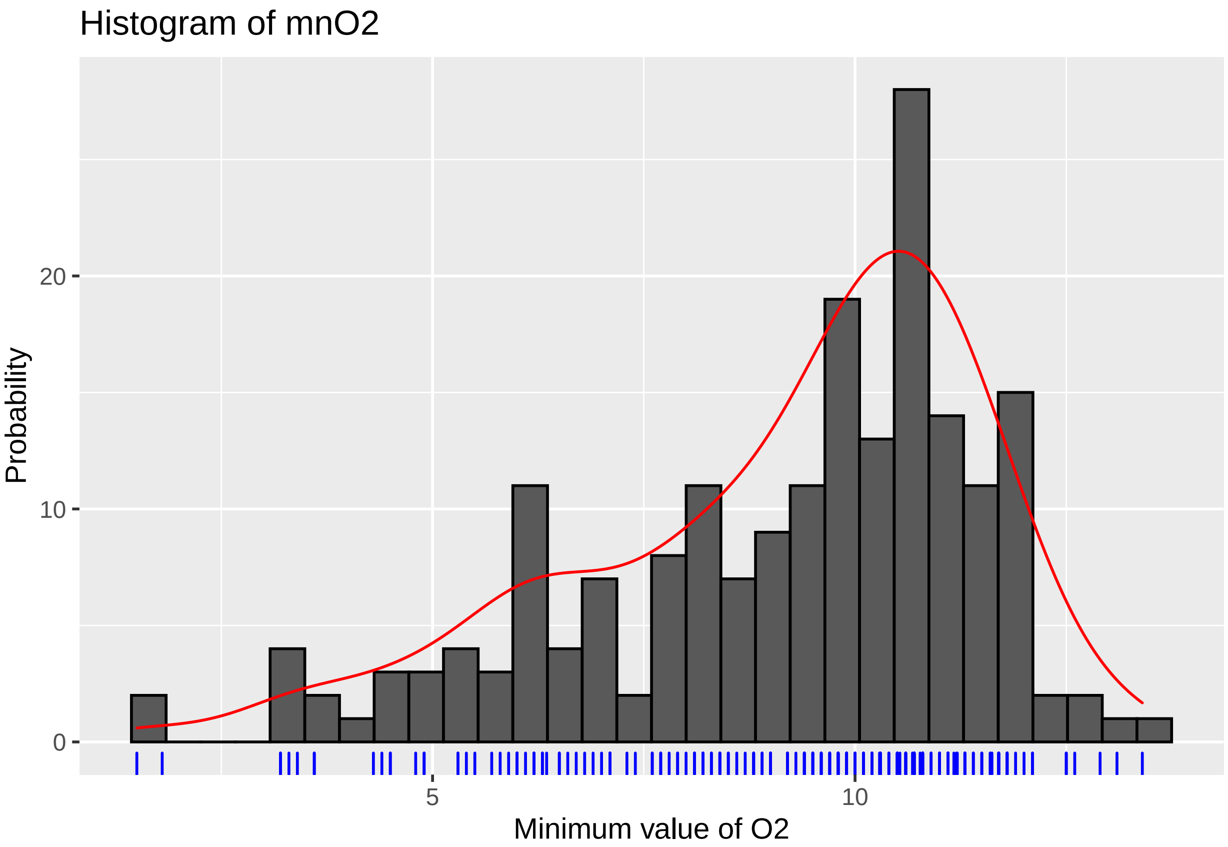
(b)

```r
algae %>%
  ggplot(aes(x=mnO2, stat = "density")) +
  geom_histogram(col = "black") +
  ggtitle("Histogram of mnO2") +
  labs(x = "Minimum value of O2", y = "Probability") +
  geom_density(aes(y= ..density..*(100)), col = "red") +
  geom_rug(col = "blue")
```

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

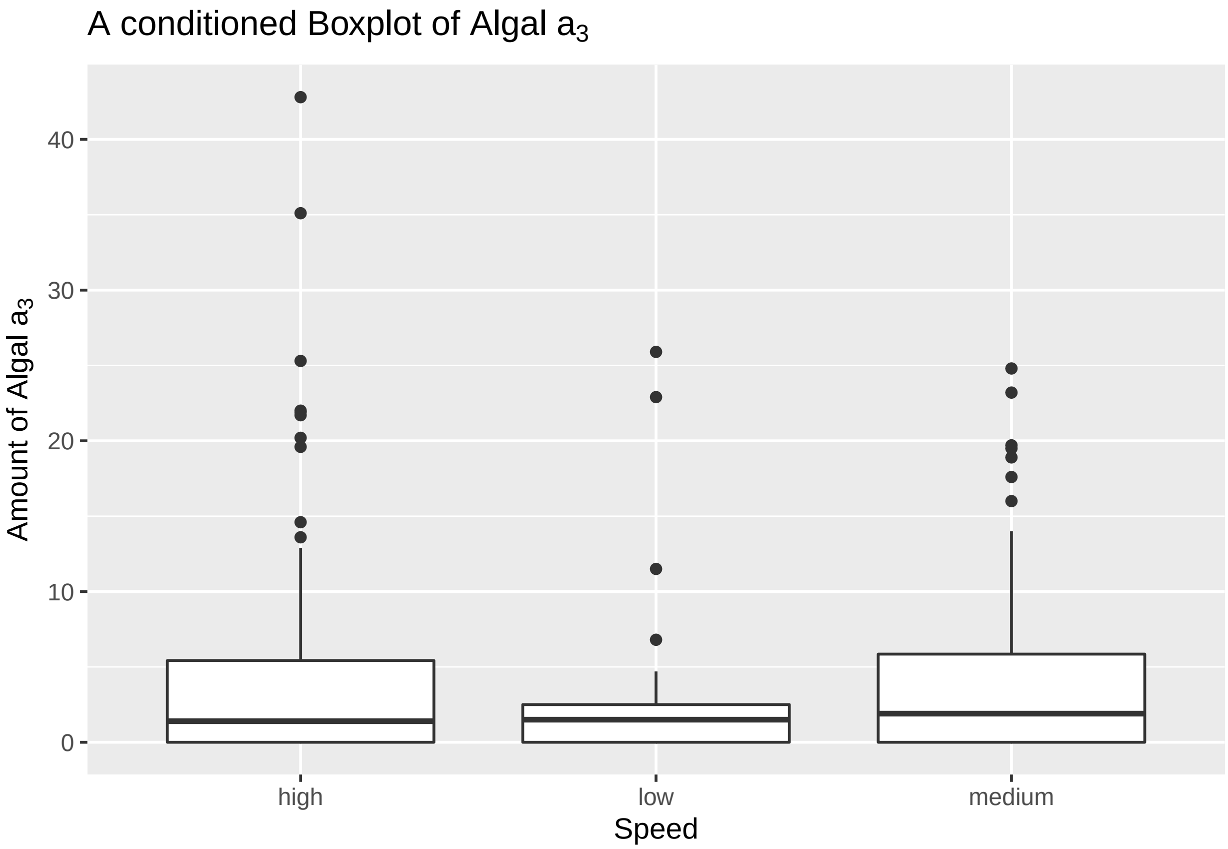
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



(c)

```
algae %>%
  ggplot(aes(x=speed, y=a3)) +
  geom_boxplot() +
  labs(title = expression("A conditioned Boxplot of Algal a"[3]), x = "Speed", y =
         expression("Amount of Algal a"[3]))
```



At low river speeds, there appears to be the lowest volumes of Algal a_3 with only 4 outliers. At high river speeds, there are more outliers with the largest volumes of Algal a_3. At medium river speeds, the interquartile range and top whisker is very similar to that of high river speeds, however, the outliers for the amount of algal $a3$ are lower than the outliers for high speeds.

Problem 3

(a)

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

There are 33 total missing values. More specifically for the following variables: * mxPH: 1 * mnO2: 2 * Cl: 10 * NO3: 2 * NH4: 2 * oPO4: 2 * PO4: 2 * Chla: 12 There are no missing values for a1 - a7.

```r
algae.del <- filter(algae, !is.na(mxPH)&!is.na(mnO2)&!is.na(Cl)&!is.na(NO3)&!is.na(NH4)
                    &!is.na(oPO4)&!is.na(PO4)&!is.na(Chla))
summary(algae.del)
str(algae.del)
```

There are 184 total observations in algae.del.

Problem 4

(a)

$Var(\hat{f}(x_0))$ and $[Bias(\hat{f}(x_0))]^2$ represent the reducible error terms, while $Var(\epsilon)$ represents the irreducible error.

(b)

$Var(\hat{f}(x_0)) \geq 0$ because variance is inherently nonnegative. $[Bias(\hat{f}(x_0))]^2 \geq 0$ must also be nonnegative because it is a squared term. Thus, the expected test error is always at least as large as the irreducible error.