Homework Assignment

Yuhuan Lyu

19, 2025

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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)

```
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~
```

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

```
4 winter    62
```

- (b)

```
[1] 33
```

```
# A tibble: 8 x 3
  chemical   mean         var
  <chr>     <dbl>       <dbl>
1 mxPH       8.01       0.358
2 mnO2       9.12       5.72
3 Cl        43.6     2193.
4 NO3        3.28      14.3
5 NH4      501.   3851585.
6 oPO4      73.6     8306.
7 PO4      138.     16639.
8 Chla      14.0      420.
```

Based on the output above, we observe significant differences in the magnitude of means and variances across chemical variables.

- (c)

```
# A tibble: 8 x 5
  chemical   mean         var median   mad
  <chr>     <dbl>       <dbl>  <dbl> <dbl>
1 mxPH       8.01       0.358   8.06  0.34
2 mnO2       9.12       5.72    9.8   1.38
3 Cl        43.6     2193.     32.7  22.4
4 NO3        3.28      14.3     2.68  1.46
5 NH4      501.   3851585.    103.   75.3
6 oPO4      73.6     8306.     40.2  29.7
7 PO4      138.     16639.    103.   82.5
8 Chla      14.0      420.      5.48  4.5
```

Most chemicals (NH4, oPO4, PO4, Chla) exhibit **right-skewed distributions with extreme outliers**, as evidenced variance » MAD^2, indicating that **median and MAD are more robust and representative** measures for this dataset.

2. Data visualization (8 pts in total) Most of the time, the information in the data set is also well captured graphically. Histogram, scatter plot, boxplot, Q-Q plot are frequently used tools for data visualization. Use ggplot for all of these visualizations.

- (a)

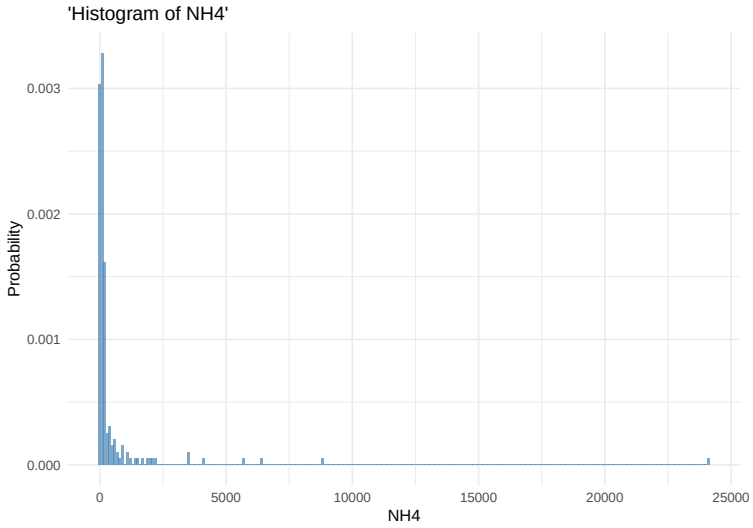

Figure 1: (a) Histogram of NH4

The distribution is left-skewed, as the tail extends more to the left side.

- (b)

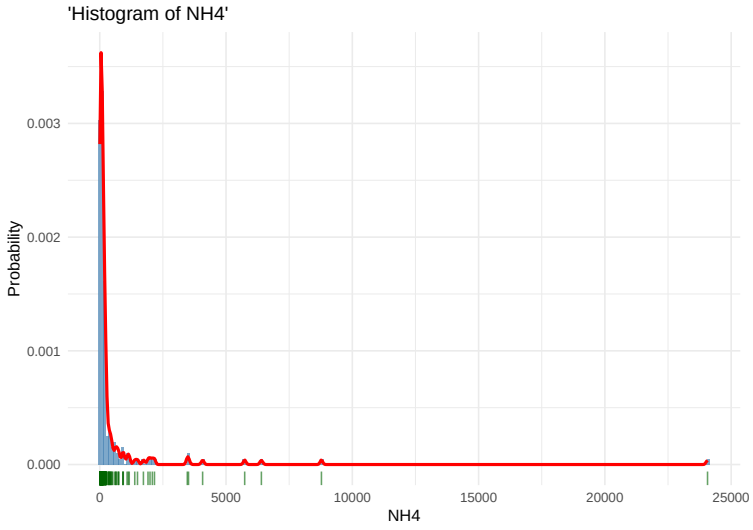

Figure 2: (b) Histogram of NH4 with density and rug

- (c)

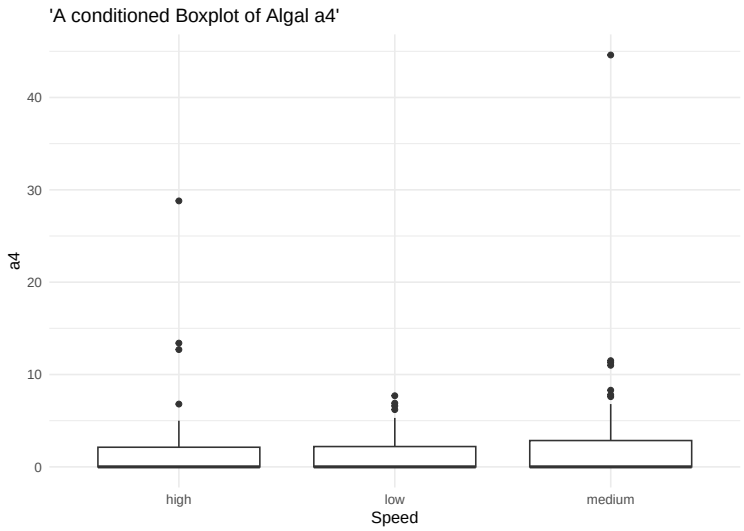

Figure 3: (c) Boxplot of a4 grouped by speed

Through observing the a4 boxplot grouped by speed, I notice that all speed groups have medians and quartiles very close to zero, with the box bodies almost touching the x-axis, indicating that most a4 values are concentrated near zero and algal a4 concentrations are generally very low. The main differences are reflected in the outliers, where the medium speed group has the highest outlier around 44, the high speed group has the second highest outlier around 29, and the low speed group has relatively lower outliers with the highest around 7. All groups contain multiple outliers, but the overall distribution patterns are similar, suggesting that in most cases, water flow speed has little effect on a4 concentration, but there may be differences in extreme circumstances.

3. Dealing with missing values

- (a)

```
(1) Number of observations containing missing values: 16

(2) Missing values in each variable:

# A tibble: 8 x 2
  Variable Missing_Count
  <chr>            <int>
1 mxPH                 1
2 mnO2                 2
3 Cl                  10
4 NO3                  2
5 NH4                  2
6 oPO4                 2
7 PO4                  2
8 Chla                12
```

- (b)

```
Number of observations in algae.del (after removing rows with missing values): 184
```

4. In lecture we present the bias-variance tradeoff that takes the form ···

- (a) The reducible error consists of the terms $\text{Var}\left(\hat{f}\left(\mathbf{x}_0\right)\right)$ and $\left[\text{Bias}\left(\hat{f}\left(\mathbf{x}_0\right)\right)\right]^2$. These terms are called "reducible" because they can be reduced by improving our model estimation $\hat{f}(\cdot)$

through better model selection, more training data, or improved estimation methods.

The irreducible error is represented by the term $\text{Var}(\varepsilon)$. This error is called "irreducible" because it represents the inherent randomness in the relationship between X and Y that cannot be eliminated by any model, no matter how sophisticated. It comes from unmeasured variables, natural variability, or measurement error in the response variable.

- (b)

Using the bias-variance tradeoff formula:

$$\text{E}\left[\left(y_0 - \hat{f}\left(\mathbf{x}_0\right)\right)^2\right] = \text{Var}\left(\hat{f}\left(\mathbf{x}_0\right)\right) + \left[\text{Bias}\left(\hat{f}\left(\mathbf{x}_0\right)\right)\right]^2 + \text{Var}(\varepsilon)$$

Proof:

Since variance is always non-negative, we have:

- $\text{Var}\left(\hat{f}\left(\mathbf{x}_0\right)\right) \geq 0$
- $\left[\text{Bias}\left(\hat{f}\left(\mathbf{x}_0\right)\right)\right]^2 \geq 0$ (since any real number squared is non-negative)

Therefore:

$$\text{E}\left[\left(y_0 - \hat{f}\left(\mathbf{x}_0\right)\right)^2\right] = \text{Var}\left(\hat{f}\left(\mathbf{x}_0\right)\right) + \left[\text{Bias}\left(\hat{f}\left(\mathbf{x}_0\right)\right)\right]^2 + \text{Var}(\varepsilon) \geq \text{Var}(\varepsilon)$$

This shows that the expected test error $\text{E}\left[\left(y_0 - \hat{f}\left(\mathbf{x}_0\right)\right)^2\right]$ is always at least as large as the irreducible error $\text{Var}(\varepsilon)$.

5. (231 Only) Prove the bias-variance tradeoff

Proof: Let $Y_0 = f(x_0) + \varepsilon$, with $E[\varepsilon] = 0$ and ε independent of $\hat{f}(x_0)$.

(1) Split the test error into signal and noise:

$$E[(Y_0 - \hat{f}(x_0))^2] = E[(f(x_0) + \varepsilon - \hat{f}(x_0))^2] = E[(f(x_0) - \hat{f}(x_0))^2] + E[\varepsilon^2] = E[(f(x_0) - \hat{f}(x_0))^2] + \text{Var}(\varepsilon).$$

(The cross term vanishes since $E[\varepsilon] = 0$ and is independent.)

(2) Decompose the first term into variance and squared bias. Let $a = E[\hat{f}(x_0)]$:

$$E[(f(x_0) - \hat{f}(x_0))^2] = E[(f(x_0) - a + a - \hat{f}(x_0))^2] = (f(x_0) - a)^2 + E[(a - \hat{f}(x_0))^2] + 2(f(x_0) - a)E[a - \hat{f}(x_0)].$$

Since $E[a - \hat{f}(x_0)] = a - E[\hat{f}(x_0)] = 0$, we obtain

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

(3) Combine (1) and (2):

$$E[(Y_0 - \hat{f}(x_0))^2] = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon).$$

This is the bias–variance tradeoff.

6. (231 Only) For $\hat{f}(x_0) = E[Y \mid X = x_0]$, show $\text{Bias}(\hat{f}(x_0)) = \text{Var}(\hat{f}(x_0)) = 0$.

Proof: Define the true regression function $f(x_0) := E[Y \mid X = x_0]$. If we take $\hat{f}(x_0)$ to be exactly this population quantity, then $\hat{f}(x_0) = f(x_0)$ is a deterministic constant (it does not depend on a training sample).

- Bias: $\text{Bias}(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0) = f(x_0) - f(x_0) = 0$.
- Variance: since $\hat{f}(x_0)$ is constant across training sets, $\text{Var}(\hat{f}(x_0)) = 0$.

Hence both the bias and the variance are zero.

7. (231 Only)Show that the following measures are distance metrics by showing the above properties hold:

- (1)$d(x,y) = \||x-y\|_2$ **Proof:**
  - Positivity: $d(x,y) = (\sum_{i=1}^{n}(x_i - y_i)^2)^{1/2} \geq 0$
  - Symmetry: $d(x,y) = (\sum_{i=1}^{n}(x_i - y_i)^2)^{1/2} = (\sum_{i=1}^{n}(y_i - x_i)^2)^{1/2} = d(y,x)$
  - Triangle inequality: Let $a = x - y$ and $b = y - z$. Then $x - z = a + b$ and

    $\|x-z\|_2 = \|a+b\|_2, \quad \|a+b\|_2^2 = (a+b)\cdot(a+b) = \|a\|_2^2 + \|b\|_2^2 + 2\,a\cdot b \leq \|a\|_2^2 + \|b\|_2^2 + 2\,\|a\|_2\,\|b\|_2$ (by Cauchy–Schwarz)

    Taking square roots gives $\|x-z\|_2 \leq \|x-y\|_2 + \|y-z\|_2$, i.e., $d(x,z) \leq d(x,y) + d(y,z)$.

- (2)$d(x,y) = \|x-y\|_\infty$ **Proof:**
  - Positivity: $d(x,y) = \max_{i=1}^{n}|x_i - y_i| \geq 0$
  - Symmetry: $d(x,y) = \max_{i=1}^{n}|x_i - y_i| = \max_{i=1}^{n}|y_i - x_i| = d(y,x)$
  - Triangle inequality: Let $a = x - y$ and $b = y - z$. Then $x - z = a + b$ and

    $\|x-z\|_\infty = \max\limits_{i=1}^{n}|x_i - z_i| = \max\limits_{i=1}^{n}|(x_i - y_i) + (y_i - z_i)| \leq \max\limits_{i=1}^{n}|x_i - y_i| + \max\limits_{i=1}^{n}|y_i - z_i| = \|x-y\|_\infty + \|y-z\|_\infty,$

    i.e., $d(x,z) \leq d(x,y) + d(y,z)$.