Data Managment Task

Jess Devine

2025-02-20

R Markdown

```
# import data
setwd("/home/jess/GIT/BIO1004W_DM/data")
data <- read_csv2("Chick_condition.csv")
```

```
## i Using "',' " as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 88 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr  (12): Year, Group, Nest, Groupsize, Afem, Amal, SA, Juv, Fledge date, R...
## dbl   (3): Ringingage, meanmaxTspecific, Rainfallspecific
## date  (2): Lay date, Hatch date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# checking data
head(data)
```

```
## # A tibble: 6 x 17
##   Year    Group Nest  Groupsize Afem  Amal  SA    Juv   `Lay date` `Fledge date`
##   <chr>   <chr> <chr> <chr>     <chr> <chr> <chr> <chr> <date>     <chr>
## 1 2004-2~ Addg~ Addg~ 5         1     3     0     1     2004-10-20 26/02/2005
## 2 2004-2~ Caro~ Caro~ 3         1     2     0     0     2004-11-28 02/04/2005
## 3 2004-2~ Herm~ Von ~ 2         1     1     0     0     2004-11-16 21/03/2005
## 4 2004-2~ John~ John~ 4         1     3     0     0     2004-10-20 25/02/2005
## 5 2004-2~ Keer~ Keer~ 4         1     2     1     0     2004-11-26 31/03/2005
## 6 2004-2~ Pitl~ Pitl~ 3         1     2     0     0     2004-11-08 12/03/2005
## # i 7 more variables: `Hatch date` <date>, Ringingdate <chr>, Chickmass <chr>,
## #   Tarsuslength <chr>, Ringingage <dbl>, meanmaxTspecific <dbl>,
## #   Rainfallspecific <dbl>
```

```
tail(data) # NOTE: `Fledge date` changes format
```

```
## # A tibble: 6 x 17
##   Year    Group Nest  Groupsize Afem  Amal  SA    Juv   `Lay date` `Fledge date`
##   <chr>   <chr> <chr> <chr>     <chr> <chr> <chr> <chr> <date>     <chr>
## 1 2020-2~ Kara~ Kara~ 5         1     3     0     1     2020-10-23 2021/02/24
## 2 2020-2~ Herm~ Von ~ 3         1     2     0     0     2020-10-23 2021/02/17
## 3 2020-2~ John~ McBr~ 6         1     2     1     2     2020-10-27 2021/03/02
## 4 2020-2~ Cope~ Char~ 4         1     3     0     0     2020-10-31 2021/03/08
## 5 2020-2~ Dover Pitl~ 3         1     2     0     0     2020-11-02 2021/03/06
## 6 2020-2~ York  York  2         1     1     0     0     2020-11-08 2021/03/08
```

```
## # i 7 more variables: `Hatch date` <date>, Ringingdate <chr>, Chickmass <chr>,
## #   Tarsuslength <chr>, Ringingage <dbl>, meanmaxTspecific <dbl>,
## #   Rainfallspecific <dbl>
```

```
summary(data)
```

```
##      Year              Group              Nest            Groupsize
##  Length:88          Length:88          Length:88          Length:88
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Afem               Amal               SA                Juv
##  Length:88          Length:88          Length:88          Length:88
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Lay date          Fledge date          Hatch date
##  Min.   :2004-10-15   Length:88          Min.   :2004-11-24
##  1st Qu.:2007-07-27   Class :character   1st Qu.:2007-09-05
##  Median :2013-10-29   Mode  :character   Median :2013-12-08
##  Mean   :2013-01-18                      Mean   :2013-02-26
##  3rd Qu.:2017-10-20                      3rd Qu.:2017-11-28
##  Max.   :2020-11-08                      Max.   :2020-12-18
##  Ringingdate         Chickmass          Tarsuslength         Ringingage
##  Length:88          Length:88          Length:88          Min.   :53.00
##  Class :character   Class :character   Class :character   1st Qu.:72.00
##  Mode  :character   Mode  :character   Mode  :character   Median :75.00
##                                                           Mean   :73.26
##                                                           3rd Qu.:76.00
##                                                           Max.   :81.00
##  meanmaxTspecific Rainfallspecific
##  Min.   :29.06    Min.   : 30.4
##  1st Qu.:30.75    1st Qu.:118.8
##  Median :31.52    Median :150.6
##  Mean   :31.56    Mean   :196.3
##  3rd Qu.:32.32    3rd Qu.:274.2
##  Max.   :34.71    Max.   :542.0
```

```
n_a_counts <- sapply(data, function(x) if (is.character(x)) sum(x == "n/a", na.rm = TRUE) else 0); n_a_
```

```
##             Year            Group             Nest        Groupsize
##                0                0                0                2
##             Afem             Amal               SA              Juv
##                6                6                6                6
##         Lay date      Fledge date       Hatch date      Ringingdate
##                0               23                0                0
##        Chickmass     Tarsuslength       Ringingage meanmaxTspecific
##                1                2                0                0
## Rainfallspecific
##                0
```

```r
# Replace "n/a" and empty strings with proper NA
data <- data %>%
  mutate(across(where(is.character), ~ na_if(.x, "n/a"))) %>%
  mutate(across(where(is.character), ~ na_if(.x, "")))

tidy_data <- data %>%
  mutate(
    # Convert to Date safely, replacing invalid values with NA
    Ringingdate = suppressWarnings(as.Date(Ringingdate, format = "%d/%m/%Y")),
    `Lay date` = suppressWarnings(as.Date(`Lay date`, format = "%Y-%m-%d")),
    `Hatch date` = suppressWarnings(as.Date(`Hatch date`, format = "%Y-%m-%d")),

    # Handle mixed Fledge date formats and invalid cases
    `Fledge date` = case_when(
      str_detect(`Fledge date`, "^\\d{4}/") ~ suppressWarnings(ymd(str_replace(`Fledge date`, "/", "-"))),
      str_detect(`Fledge date`, "^\\d{2}/\\d{2}/\\d{4}$") ~ suppressWarnings(dmy(`Fledge date`)),
      TRUE ~ NA_Date_
    ),

    # Convert to numeric safely, replacing non-numeric values with NA
    Groupsize = suppressWarnings(as.numeric(Groupsize)),
    Chickmass = suppressWarnings(as.numeric(Chickmass)),
    Tarsuslength = suppressWarnings(as.numeric(Tarsuslength)),

    # Convert to factor
    Year = as.factor(Year),
    Group = as.factor(Group),
    Nest = as.factor(Nest)
  )

# Identify problematic rows where NA was introduced
problematic_rows <- data %>%
  filter(
    is.na(tidy_data$`Fledge date`) & !is.na(`Fledge date`) |
    is.na(tidy_data$Tarsuslength) & !is.na(Tarsuslength)
  )

print(problematic_rows)
```

```
## # A tibble: 6 x 17
##   Year    Group Nest  Groupsize Afem  Amal  SA    Juv   `Lay date` `Fledge date`
##   <chr>   <chr> <chr> <chr>     <chr> <chr> <chr> <chr> <date>     <chr>
## 1 2020-2~ Kara~ Kara~ 5         1     3     0     1     2020-10-23 2021/02/24
## 2 2020-2~ Herm~ Von ~ 3         1     2     0     0     2020-10-23 2021/02/17
## 3 2020-2~ John~ McBr~ 6         1     2     1     2     2020-10-27 2021/03/02
## 4 2020-2~ Cope~ Char~ 4         1     3     0     0     2020-10-31 2021/03/08
## 5 2020-2~ Dover Pitl~ 3         1     2     0     0     2020-11-02 2021/03/06
## 6 2020-2~ York  York  2         1     1     0     0     2020-11-08 2021/03/08
## # i 7 more variables: `Hatch date` <date>, Ringingdate <chr>, Chickmass <chr>,
## #   Tarsuslength <chr>, Ringingage <dbl>, meanmaxTspecific <dbl>,
## #   Rainfallspecific <dbl>
```

```r
# demographic groups to long format and compute incubation period
tidy_data <- tidy_data %>%
```

```
  pivot_longer(
    cols = c(Afem, Amal, SA, Juv),
    names_to = "Age_Sex_Class",
    values_to = "Count") %>%
      mutate(
        Count = as.integer(Count), # Convert to integer
        Age_Sex_Class = as.factor(Age_Sex_Class), # convert to factor
        incubation_period = as.numeric(`Hatch date` - `Lay date`)  # Compute incubation period
         )

head(tidy_data)
```

```
## # A tibble: 6 x 16
##   Year   Group Nest  Groupsize `Lay date` `Fledge date` `Hatch date` Ringingdate
##   <fct>  <fct> <fct>     <dbl> <date>     <date>        <date>       <date>
## 1 2004-~ Addg~ Addg~         5 2004-10-20 2005-02-26    2004-11-29   2005-02-14
## 2 2004-~ Addg~ Addg~         5 2004-10-20 2005-02-26    2004-11-29   2005-02-14
## 3 2004-~ Addg~ Addg~         5 2004-10-20 2005-02-26    2004-11-29   2005-02-14
## 4 2004-~ Addg~ Addg~         5 2004-10-20 2005-02-26    2004-11-29   2005-02-14
## 5 2004-~ Caro~ Caro~         3 2004-11-28 2005-04-02    2005-01-06   2005-03-24
## 6 2004-~ Caro~ Caro~         3 2004-11-28 2005-04-02    2005-01-06   2005-03-24
## # i 8 more variables: Chickmass <dbl>, Tarsuslength <dbl>, Ringingage <dbl>,
## #   meanmaxTspecific <dbl>, Rainfallspecific <dbl>, Age_Sex_Class <fct>,
## #   Count <int>, incubation_period <dbl>
```

```
colSums(is.na(tidy_data))
```

```
##              Year             Group              Nest         Groupsize
##                 0                 0                 0                 8
##          Lay date       Fledge date        Hatch date       Ringingdate
##                 0                92                 0                 0
##         Chickmass      Tarsuslength        Ringingage  meanmaxTspecific
##                 4                32                 0                 0
##  Rainfallspecific     Age_Sex_Class             Count incubation_period
##                 0                 0                24                 0
```

```
# Function to create scatterplots
create_scatter <- function(data, x_var, y_var, x_label, y_label) {
  ggplot(data, aes(x = {{ x_var }}, y = {{ y_var }})) +
    geom_point(color = "black", alpha = 0.7, size = 0.5) +
    theme_minimal() +
    labs(x = x_label, y = y_label)
}

# Function to create boxplots
create_boxplot <- function(data, x_var, y_var, x_label, y_label) {
  ggplot(data, aes(x = as.factor({{ x_var }}), y = {{ y_var }})) +
    geom_boxplot() +
    theme_minimal() +
    labs(x = x_label, y = y_label)
}

# Incubation Period Plots
p1 <- create_scatter(tidy_data, Rainfallspecific, incubation_period, "", "Incubation Period (days)")
p2 <- create_scatter(tidy_data, meanmaxTspecific, incubation_period, "", "")
```

```
p3 <- create_boxplot(tidy_data, Groupsize, incubation_period, "", "")

# Chick Mass Plots
p4 <- create_scatter(tidy_data, Rainfallspecific, Chickmass, "Rainfall (mm)", "Chick Mass (g)")
p5 <- create_scatter(tidy_data, meanmaxTspecific, Chickmass, "Mean Max Temperature (°C)", "")
p6 <- create_boxplot(tidy_data, Groupsize, Chickmass, "Group Size", "")

# Arrange in 2x3 panel layout
(p1 + p2 + p3) / (p4 + p5 + p6)
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 4 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

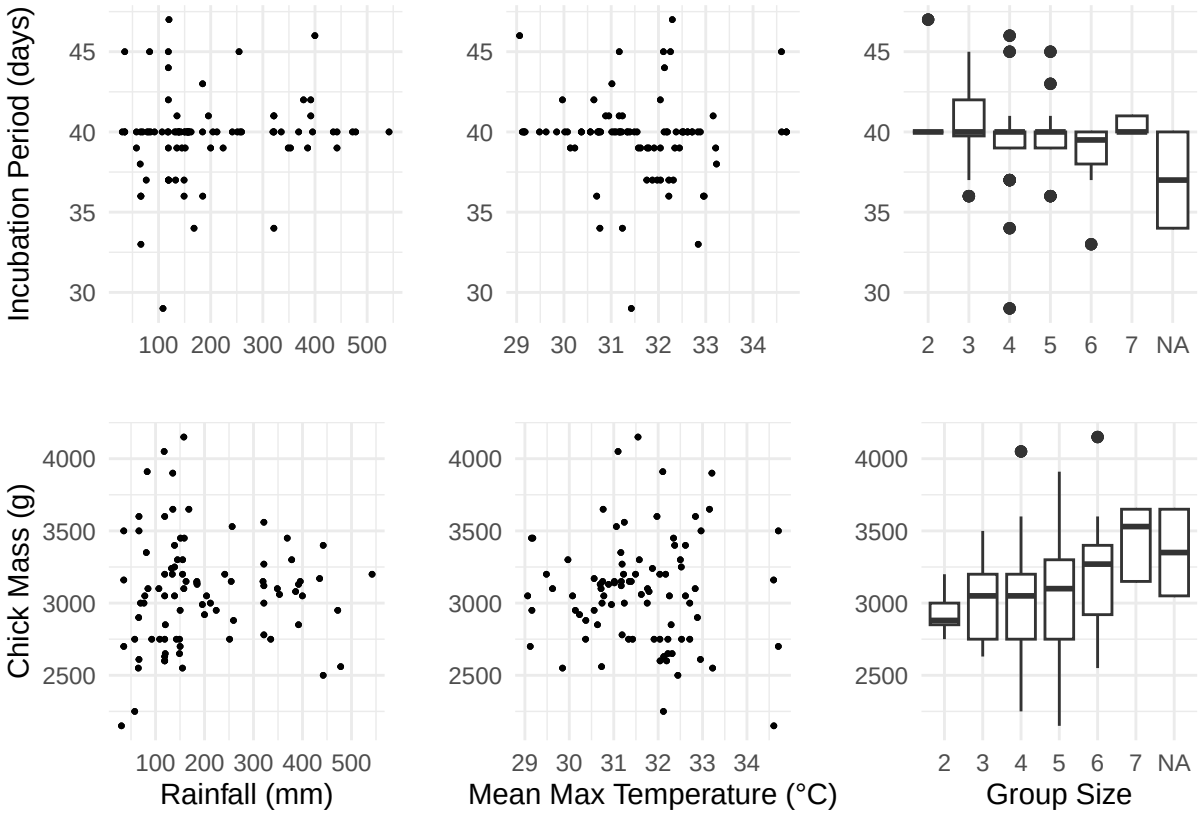

Figure 1: The incubation period of eggs (top panel) and mass of chicks (bottom panel) plotted against rainfall in mm (left panel), mean maximum temperature °C (middle panel) and group size (right panel).

```
# Shapiro-Wilk test for normality
shapiro.test(tidy_data$incubation_period)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tidy_data$incubation_period
## W = 0.82904, p-value < 2.2e-16
```

```r
shapiro.test(tidy_data$Chickmass)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tidy_data$Chickmass
## W = 0.97768, p-value = 3.2e-05
```

```r
# Kruskal-Wallis Test
kruskal.test(incubation_period ~ as.factor(Groupsize), data = tidy_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  incubation_period by as.factor(Groupsize)
## Kruskal-Wallis chi-squared = 25.572, df = 5, p-value = 0.000108
```

```r
kruskal.test(Chickmass ~ as.factor(Groupsize), data = tidy_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Chickmass by as.factor(Groupsize)
## Kruskal-Wallis chi-squared = 24.983, df = 5, p-value = 0.0001404
```

```r
# Dunn's test for pairwise comparisons
dunnTest(incubation_period ~ as.factor(Groupsize), data = tidy_data, method = "bonferroni")
```

```
## Warning: Some rows deleted from 'x' and 'g' because missing data.
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
##   p-values adjusted with the Bonferroni method.
```

```
##    Comparison           Z      P.unadj       P.adj
## 1       2 - 3  1.32325325 1.857512e-01 1.000000000
## 2       2 - 4  2.29936136 2.148443e-02 0.322266425
## 3       3 - 4  1.55192923 1.206792e-01 1.000000000
## 4       2 - 5  2.45009592 1.428182e-02 0.214227248
## 5       3 - 5  1.78457616 7.433009e-02 1.000000000
## 6       4 - 5  0.34274126 7.317931e-01 1.000000000
## 7       2 - 6  3.96560363 7.321044e-05 0.001098157
## 8       3 - 6  3.89989784 9.623329e-05 0.001443499
## 9       4 - 6  2.84336956 4.463928e-03 0.066958916
## 10      5 - 6  2.48000979 1.313788e-02 0.197068165
## 11      2 - 7 -0.08741812 9.303392e-01 1.000000000
## 12      3 - 7 -1.17173619 2.413030e-01 1.000000000
## 13      4 - 7 -1.94439083 5.184834e-02 0.777725091
## 14      5 - 7 -2.07876499 3.763896e-02 0.564584347
## 15      6 - 7 -3.39656336 6.823776e-04 0.010235664
```

```r
dunnTest(Chickmass ~ as.factor(Groupsize), data = tidy_data, method = "bonferroni")
```

```
## Warning: Some rows deleted from 'x' and 'g' because missing data.
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
##   p-values adjusted with the Bonferroni method.
```

```
##    Comparison          Z      P.unadj        P.adj
```

```
## 1       2 - 3 -1.1490368 2.505408e-01 1.0000000000
## 2       2 - 4 -1.2927252 1.961061e-01 1.0000000000
## 3       3 - 4 -0.1731448 8.625376e-01 1.0000000000
## 4       2 - 5 -1.8654996 6.211141e-02 0.9316711553
## 5       3 - 5 -1.1078233 2.679381e-01 1.0000000000
## 6       4 - 5 -1.0273522 3.042547e-01 1.0000000000
## 7       2 - 6 -2.6771682 7.424735e-03 0.1113710324
## 8       3 - 6 -2.2750308 2.290408e-02 0.3435612371
## 9       4 - 6 -2.2609319 2.376347e-02 0.3564520751
## 10      5 - 6 -1.4003924 1.613958e-01 1.0000000000
## 11      2 - 7 -4.1979484 2.693439e-05 0.0004040159
## 12      3 - 7 -4.0051038 6.199024e-05 0.0009298537
## 13      4 - 7 -4.0033580 6.244970e-05 0.0009367455
## 14      5 - 7 -3.4630657 5.340579e-04 0.0080108686
## 15      6 - 7 -2.4296671 1.511270e-02 0.2266904656
```