Data Management and Reproducible Research Deliverable: Data Tidying

Jess Devine

2025-02-20

Import data

```
# import data
data <- read_csv2(here("data", "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
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
```

```
length(data)
```

```
## [1] 17
```

```
head(data[,1:8])
```

```
## # A tibble: 6 x 8
##   Year      Group        Nest         Groupsize Afem  Amal  SA    Juv
##   <chr>     <chr>        <chr>        <chr>     <chr> <chr> <chr> <chr>
## 1 2004-2005 Addger       Addger       5         1     3     0     1
## 2 2004-2005 Caroline     Caroline     3         1     2     0     0
## 3 2004-2005 Hermansburg  Von Tonder   2         1     1     0     0
## 4 2004-2005 Johnniesdale Johnniesdale 4         1     3     0     0
## 5 2004-2005 Keer Keer    Keer Keer    4         1     2     1     0
## 6 2004-2005 Pitlochry    Pitlochry    3         1     2     0     0
```

```
tail(data[,1:8])
```

```
## # A tibble: 6 x 8
##   Year      Group        Nest        Groupsize Afem  Amal  SA    Juv
##   <chr>     <chr>        <chr>       <chr>     <chr> <chr> <chr> <chr>
## 1 2020-2021 Karan Khaya  Karan Khaya 5         1     3     0     1
## 2 2020-2021 Hermansburg  Von Tonder  3         1     2     0     0
## 3 2020-2021 Johnniesdale McBride     6         1     2     1     2
## 4 2020-2021 Copenhagen   Charloscar  4         1     3     0     0
## 5 2020-2021 Dover        Pitlochry   3         1     2     0     0
## 6 2020-2021 York         York        2         1     1     0     0
```

```
head(data[,9:17])
```

```
## # A tibble: 6 x 9
##   `Lay date` `Fledge date` `Hatch date` Ringingdate Chickmass Tarsuslength
##   <date>     <chr>         <date>       <chr>       <chr>     <chr>
## 1 2004-10-20 26/02/2005    2004-11-29   14/02/2005  3400      n/a
## 2 2004-11-28 02/04/2005    2005-01-06   24/03/2005  2750      143
## 3 2004-11-16 21/03/2005    2004-12-26   14/03/2005  3000      153
## 4 2004-10-20 25/02/2005    2004-11-29   14/02/2005  3050      167
## 5 2004-11-26 31/03/2005    2005-01-05   24/03/2005  2250      149
## 6 2004-11-08 12/03/2005    2004-12-18   05/03/2005  2750      155
## # i 3 more variables: Ringingage <dbl>, meanmaxTspecific <dbl>,
## #   Rainfallspecific <dbl>
```

```r
tail(data[,9:17]) # NOTE: `Fledge date` changes format
```

```
## # A tibble: 6 x 9
##   `Lay date` `Fledge date` `Hatch date` Ringingdate Chickmass Tarsuslength
##   <date>     <chr>         <date>       <chr>       <chr>     <chr>
## 1 2020-10-23 2021/02/24    2020-12-02   15/02/2021  3170      147,37
## 2 2020-10-23 2021/02/17    2020-12-02   15/02/2022  2950      150,4
## 3 2020-10-27 2021/03/02    2020-12-06   19/02/2021  2560      151,66
## 4 2020-10-31 2021/03/08    2020-12-11   24/02/2021  3130      149,92
## 5 2020-11-02 2021/03/06    2020-12-14   27/02/2021  2850      152,32
## 6 2020-11-08 2021/03/08    2020-12-18   03/03/2021  2880      138,6
## # i 3 more variables: Ringingage <dbl>, meanmaxTspecific <dbl>,
## #   Rainfallspecific <dbl>
```

```r
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
```

Data tyding

```r
# Replace "n/a" and empty strings with proper NA
tidy_data  <- data %>%
  mutate(across(where(is.character), ~ na_if(.x, "n/a"))) %>%
  mutate(across(where(is.character), ~ na_if(.x, "")))
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
tidy_data <- tidy_data  %>%
  mutate(
    # Convert to Date safely, replacing invalid values with NA
    `Lay date` = suppressWarnings(as.Date(`Lay date`, format = "%Y-%m-%d")),
    `Hatch date` = suppressWarnings(as.Date(`Hatch date`, format = "%Y-%m-%d")),
    Ringingdate = suppressWarnings(as.Date(Ringingdate, format = "%d/%m/%Y")),

    # 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)
  )
length(problematic_rows)
```

```
## [1] 17
```

```
print(problematic_rows[,1:8])
```

```
## # A tibble: 31 x 8
##    Year      Group        Nest         Groupsize Afem  Amal  SA    Juv
##    <chr>     <chr>        <chr>        <chr>     <chr> <chr> <chr> <chr>
##  1 2004-2005 Addger       Addger       5         1     3     0     1
##  2 2004-2005 Rhino Road   Rhino Road   5         1     2     1     1
##  3 2007-2008 Addger       Addger       n/a       n/a   n/a   n/a   n/a
##  4 2007-2008 De Luca      De Luca      3         1     1     0     1
##  5 2007-2008 Giraffe      Giraffe      5         1     3     0     1
##  6 2007-2008 Janovsky     Janovsky     4         1     2     0     1
##  7 2007-2008 Johnniesdale Johnniesdale 4         1     2     0     1
##  8 2007-2008 Karan Khaya  Karan Khaya  5         1     2     0     2
##  9 2007-2008 Pitlochry    Pitlochry    5         1     2     1     1
## 10 2007-2008 Rhino Road   Rhino Road   4         1     2     0     1
## # i 21 more rows
```

```
print(problematic_rows[,9:17])
```

```
## # A tibble: 31 x 9
##    `Lay date` `Fledge date` `Hatch date` Ringingdate Chickmass Tarsuslength
##    <date>     <chr>         <date>       <chr>       <chr>     <chr>
##  1 2004-10-20 26/02/2005    2004-11-29   14/02/2005  3400      n/a
##  2 2004-10-30 06/03/2005    2004-12-09   27/02/2005  3300      n/a
##  3 2007-11-12 n/a           2007-12-22   11/03/2008  3050      154
##  4 2007-11-03 n/a           2007-12-13   29/02/2008  n/a       147
##  5 2007-11-26 n/a           2008-01-05   20/03/2008  3100      152
##  6 2007-10-27 n/a           2007-12-06   09/02/2008  2950      146
##  7 2007-10-29 n/a           2007-12-08   22/02/2008  3200      150
```

```
##  8 2007-10-27 n/a           2007-12-06   10/02/2008  2700      139
##  9 2007-10-31 n/a           2007-12-10   23/02/2008  2550      151
## 10 2007-10-20 n/a           2007-11-29   10/02/2008  3450      142
## # i 21 more rows
## # i 3 more variables: Ringingage <dbl>, meanmaxTspecific <dbl>,
## #   Rainfallspecific <dbl>
```

```
# 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
```