data_wrangling

Jess Devine

2025-02-20

R Markdown

```r
# 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.
```

```r
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>
```

```r
tail(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 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
# Check unique values to see formatting inconsistencies
unique(data$`Fledge date`)
```

```
##  [1] "26/02/2005" "02/04/2005" "21/03/2005" "25/02/2005" "31/03/2005"
##  [6] "12/03/2005" "06/03/2005" "18/02/2005" "03/04/2006" "11/04/2006"
## [11] "12/03/2006" "18/02/2007" "11/02/2007" "07/03/2007" "19/02/2007"
## [16] "24/03/2007" "10/03/2007" "16/02/2007" "22/02/2007" "05/03/2007"
## [21] "10/02/2007" "21/02/2007" "n/a"        "2012/03/09" "2012/03/20"
## [26] "2012/03/28" "2012/02/18" "2013/03/01" "2013/02/25" "2013/02/19"
## [31] "2013/03/11" "2014/02/14" "2014/02/17" "2014/02/26" "2014/01/25"
## [36] "2014/03/16" "2014/04/10" "2014/02/08" "2014/04/02" "2014/03/04"
## [41] "2015/02/14" "2015/03/05" "2015/01/24" "2015/03/08" "2016/03/01"
## [46] "2016/03/06" "2016/03/11" "2018/02/26" "2018/02/24" "2018/02/21"
## [51] "2029/02/18" "2018/01/29" "2018/01/21" "2019/03/28" "2019/04/27"
## [56] "2019/04/20" "2019/04/10" "2021/02/24" "2021/02/17" "2021/03/02"
## [61] "2021/03/08" "2021/03/06"
```

```r
# correct data types
tidy_data <- data %>%
  mutate(
     # Convert to Date
    Ringingdate = as.Date(Ringingdate, format = "%d/%m/%Y"),
    `Lay date` = as.Date(`Lay date`, format = "%Y-%m-%d"),

    # Convert dates dynamically
    `Fledge date` = case_when(
      str_detect(`Fledge date`, "^\\d{4}/") ~ ymd(`Fledge date`),  # If starts with YYYY/, use YMD
      TRUE ~ dmy(`Fledge date`)  # Otherwise, assume DMY
    ),

    # Convert to integer
    Groupsize = as.integer(Groupsize),
    Chickmass = as.integer(Chickmass),
     Tarsuslength = as.integer(Tarsuslength),

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

```
## Warning: There were 5 warnings in `mutate()`.
## The first warning was:
## i In argument: `Fledge date = case_when(...)`.
## Caused by warning:
## !  45 failed to parse.
## i Run `dplyr::last_dplyr_warnings()` to see the 4 remaining warnings.
```

```r
head(tidy_data)
```

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

```
# demographic groups to long format
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
         )
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Count = as.integer(Count)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
head(tidy_data)
```

```
## # A tibble: 6 x 15
##   Year   Group Nest  Groupsize `Lay date` `Fledge date` `Hatch date` Ringingdate
##   <fct>  <fct> <fct>     <int> <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 7 more variables: Chickmass <int>, Tarsuslength <int>, Ringingage <dbl>,
## #   meanmaxTspecific <dbl>, Rainfallspecific <dbl>, Age_Sex_Class <fct>,
## #   Count <int>
```

```
summary(tidy_data)
```

```
##         Year               Group                  Nest       Groupsize
##  2006-2007: 44   Karan Khaya : 44   Karan Khaya : 44   Min.   :2.000
##  2007-2008: 36   Addger      : 36   Addger      : 40   1st Qu.:3.000
##  2013-2014: 36   Janovsky    : 32   Janovsky    : 32   Median :4.000
##  2004-2005: 32   Johnniesdale: 32   Hull        : 20   Mean   :4.233
##  2017-2018: 32   Copenhagen  : 28   Johnniesdale: 20   3rd Qu.:5.000
##  2019-2020: 28   Rhino Road  : 28   Rhino Road  : 20   Max.   :7.000
##  (Other)  :144   (Other)     :152   (Other)     :176   NA's   :8
##     Lay date           Fledge date           Hatch date
##  Min.   :2004-10-15   Min.   :2005-02-18   Min.   :2004-11-24
##  1st Qu.:2007-07-27   1st Qu.:2007-02-21   1st Qu.:2007-09-05
##  Median :2013-10-29   Median :2014-02-14   Median :2013-12-08
##  Mean   :2013-01-18   Mean   :2013-02-21   Mean   :2013-02-26
##  3rd Qu.:2017-10-20   3rd Qu.:2018-01-29   3rd Qu.:2017-11-28
##  Max.   :2020-11-08   Max.   :2029-02-18   Max.   :2020-12-18
##                        NA's   :92
##   Ringingdate           Chickmass     Tarsuslength     Ringingage
##  Min.   :2005-02-13   Min.   :2150   Min.   :128.0   Min.   :53.00
##  1st Qu.:2007-11-16   1st Qu.:2750   1st Qu.:143.8   1st Qu.:72.00
##  Median :2014-02-17   Median :3100   Median :152.0   Median :75.00
##  Mean   :2013-05-15   Mean   :3079   Mean   :151.0   Mean   :73.26
##  3rd Qu.:2018-02-10   3rd Qu.:3270   3rd Qu.:156.2   3rd Qu.:76.00
##  Max.   :2022-02-15   Max.   :4150   Max.   :179.0   Max.   :81.00
##                       NA's   :4      NA's   :32
##  meanmaxTspecific Rainfallspecific Age_Sex_Class     Count
##  Min.   :29.06    Min.   : 30.4    Afem:88        Min.   :0.000
##  1st Qu.:30.75    1st Qu.:118.8    Amal:88        1st Qu.:1.000
##  Median :31.52    Median :150.6    Juv :88        Median :1.000
##  Mean   :31.56    Mean   :196.3    SA  :88        Mean   :1.052
##  3rd Qu.:32.32    3rd Qu.:274.2                   3rd Qu.:1.000
##  Max.   :34.71    Max.   :542.0                   Max.   :4.000
##                                                   NA's   :24
```

```r
# trouble shooting with Flege date
sum(is.na(as.Date(data$`Fledge date`, format = "%d/%m/%Y")))
```

```
## [1] 66
```

```r
unique(data$`Fledge date`[is.na(as.Date(data$`Fledge date`, format = "%d/%m/%Y"))])
```

```
##  [1] "n/a"        "2012/03/09" "2012/03/20" "2012/03/28" "2012/02/18"
##  [6] "2013/03/01" "2013/02/25" "2013/02/19" "2013/03/11" "2014/02/14"
## [11] "2014/02/17" "2014/02/26" "2014/01/25" "2014/03/16" "2014/04/10"
## [16] "2014/02/08" "2014/04/02" "2014/03/04" "2015/02/14" "2015/03/05"
## [21] "2015/01/24" "2015/03/08" "2016/03/01" "2016/03/06" "2016/03/11"
## [26] "2018/02/26" "2018/02/24" "2018/02/21" "2029/02/18" "2018/01/29"
## [31] "2018/01/21" "2019/03/28" "2019/04/27" "2019/04/20" "2019/04/10"
## [36] "2021/02/24" "2021/02/17" "2021/03/02" "2021/03/08" "2021/03/06"
```