Summary report for the penguins_raw dataset

Pei-Yu Chen

2024-09-15

Table of contents

Statistical Thinking

Reference: https://www.fharrell.com/post/rflow/

Summary Staistic

```r
library(Hmisc)
```

```
Warning:   'Hmisc'   R   4.3.3
```

```
'Hmisc'

  'package:base':

 format.pval, units
```

```r
library(palmerpenguins)
latex(describe(penguins_raw), file = "", caption.placement = "top")
```

penguins_raw
17 Variables   344 Observations

studyName

```
  n   missing   distinct
344         0          3
```

```
Value      PAL0708 PAL0809 PAL0910
Frequency      110     114     120
Proportion   0.320   0.331   0.349
```

Sample Number

```
  n   missing   distinct   Info   Mean    Gmd    .05    .10    .25    .50    .75     .90     .95
344         0        152      1  63.15  46.35   6.15  12.00  29.00  58.00  95.25  121.00  134.85
```

```
lowest :   1   2   3   4   5, highest: 148 149 150 151 152
```

Species

```
  n   missing   distinct
344         0          3
```

```
Value           Adelie Penguin (Pygoscelis adeliae) Chinstrap penguin (Pygoscelis antarctica)
Frequency                                       152                                        68
Proportion                                    0.442                                     0.198

Value             Gentoo penguin (Pygoscelis papua)
Frequency                                       124
Proportion                                    0.360
```

Region

```
  n   missing   distinct    value
344         0          1   Anvers
```

```
Value      Anvers
Frequency     344
Proportion      1
```

Island

```
   n  missing  distinct
 344        0         3
```

```
Value         Biscoe     Dream Torgersen
Frequency        168       124        52
Proportion     0.488     0.360     0.151
```

Stage

```
   n  missing  distinct                 value
 344        0         1    Adult, 1 Egg Stage
```

```
Value      Adult, 1 Egg Stage
Frequency                 344
Proportion                  1
```

Individual ID

```
   n  missing  distinct
 344        0       190
```

```
lowest : N100A1 N100A2 N10A1  N10A2  N11A1 , highest: N98A2  N99A1  N99A2  N9A1   N9A2
```

Clutch Completion

```
   n  missing  distinct
 344        0         2
```

```
Value         No   Yes
Frequency     36   308
Proportion 0.105 0.895
```

Date Egg

```
         n    missing   distinct       Info       Mean        Gmd        .05        .10
       344          0         50      0.999 2008-11-27        328 2007-11-12 2007-11-16
       .25        .50        .75        .90        .95
2007-11-28 2008-11-09 2009-11-16 2009-11-22 2009-11-26

lowest : 2007-11-09 2007-11-10 2007-11-11 2007-11-12 2007-11-13
highest: 2009-11-22 2009-11-23 2009-11-25 2009-11-27 2009-12-01
```

Culmen Length (mm)

```
   n  missing  distinct   Info   Mean    Gmd    .05    .10    .25    .50    .75    .90    .95
 342        2       164      1  43.92  6.274  35.70  36.60  39.23  44.45  48.50  50.80  51.99
```

```
lowest : 32.1 33.1 33.5 34   34.1, highest: 55.1 55.8 55.9 58   59.6
```

Culmen Depth (mm)

```
  n   missing   distinct   Info   Mean    Gmd    .05    .10    .25    .50    .75    .90    .95
342         2         80      1  17.15  2.267   13.9   14.3   15.6   17.3   18.7   19.5   20.0
```

```
lowest : 13.1 13.2 13.3 13.4 13.5, highest: 20.7 20.8 21.1 21.2 21.5
```

Flipper Length (mm)

```
  n   missing   distinct    Info    Mean     Gmd     .05     .10     .25     .50     .75     .90     .95
342         2         55   0.999   200.9   16.03   181.0   185.0   190.0   197.0   213.0   220.9   225.0
```

```
lowest : 172 174 176 178 179, highest: 226 228 229 230 231
```

Body Mass (g)

```
  n   missing   distinct   Info   Mean    Gmd    .05    .10    .25    .50    .75    .90    .95
342         2         94      1   4202  911.8   3150   3300   3550   4050   4750   5400   5650
```

```
lowest : 2700 2850 2900 2925 2975, highest: 5850 5950 6000 6050 6300
```

Sex

```
  n   missing   distinct
333        11          2
```

```
Value      FEMALE   MALE
Frequency     165    168
Proportion  0.495  0.505
```

Δ 15 N (o/oo):

```
  n   missing   distinct   Info   Mean     Gmd     .05     .10     .25     .50     .75     .90     .95
330        14        330      1  8.733  0.6323   7.897   8.047   8.300   8.652   9.172   9.491   9.689
```

```
lowest : 7.6322  7.63452 7.63884 7.68528 7.6887 , highest: 9.93727 9.98044 10.0202 10.0237 10.0254
```

Δ 13 C (o/oo):

```
  n   missing   distinct   Info     Mean     Gmd      .05      .10      .25      .50      .75      .90      .95
331        13        331      1   −25.69  0.9093   −26.79   −26.69   −26.32   −25.83   −25.06   −24.53   −24.36
```

```
lowest : -27.0185 -26.9547 -26.8964 -26.8648 -26.8635, highest: -24.1657 -24.1026 -23.9031 -23.8902 -23.7877
```

Comments

```
  n   missing   distinct
 54       290         10
```

```
lowest : Adult not sampled.                                          Adult not sampled. Nest never observed with ful
highest: No blood sample obtained.                                   No delta15N data received from lab.
```

Data Structure

```r
library(table1)
library(palmerpenguins)
str(penguins_raw)
```

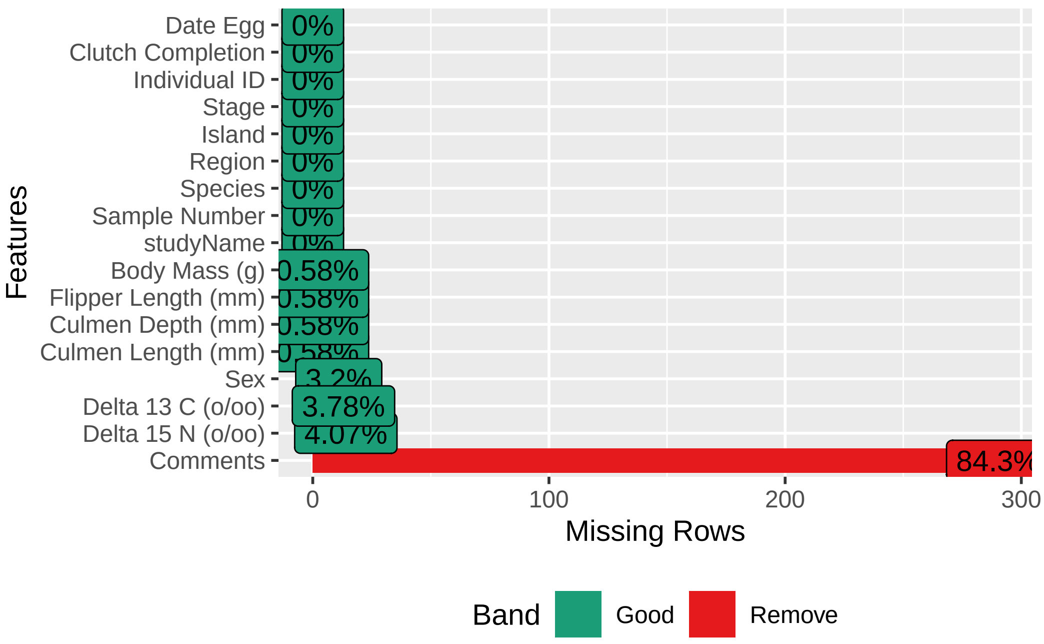
```
tibble [344 x 17] (S3: tbl_df/tbl/data.frame)
 $ studyName          : chr [1:344] "PAL0708" "PAL0708" "PAL0708" "PAL0708" ...
 $ Sample Number      : num [1:344] 1 2 3 4 5 6 7 8 9 10 ...
 $ Species            : chr [1:344] "Adelie Penguin (Pygoscelis adeliae)" "Adelie Penguin (Py
 $ Region             : chr [1:344] "Anvers" "Anvers" "Anvers" "Anvers" ...
 $ Island             : chr [1:344] "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
 $ Stage              : chr [1:344] "Adult, 1 Egg Stage" "Adult, 1 Egg Stage" "Adult, 1 Egg S
 $ Individual ID      : chr [1:344] "N1A1" "N1A2" "N2A1" "N2A2" ...
 $ Clutch Completion  : chr [1:344] "Yes" "Yes" "Yes" "Yes" ...
 $ Date Egg           : Date[1:344], format: "2007-11-11" "2007-11-11" ...
 $ Culmen Length (mm) : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ Culmen Depth (mm)  : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ Flipper Length (mm): num [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ Body Mass (g)      : num [1:344] 3750 3800 3250 NA 3450 ...
 $ Sex                : chr [1:344] "MALE" "FEMALE" "FEMALE" NA ...
 $ Delta 15 N (o/oo)  : num [1:344] NA 8.95 8.37 NA 8.77 ...
 $ Delta 13 C (o/oo)  : num [1:344] NA -24.7 -25.3 NA -25.3 ...
 $ Comments           : chr [1:344] "Not enough blood for isotopes." NA NA "Adult not sampled
 - attr(*, "spec")=List of 3
  ..$ cols   :List of 17
  .. ..$ studyName          : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Sample Number      : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Species            : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Region             : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Island             : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Stage              : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Individual ID      : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Clutch Completion  : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Date Egg           :List of 1
  .. .. ..$ format: chr ""
  .. .. ..- attr(*, "class")= chr [1:2] "collector_date" "collector"
  .. ..$ Culmen Length (mm) : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Culmen Depth (mm)  : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Flipper Length (mm): list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Body Mass (g)      : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Sex                : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Delta 15 N (o/oo)  : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Delta 13 C (o/oo)  : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Comments           : list()
  .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
  ..$ default: list()
  .. ..- attr(*, "class")= chr [1:2] "collector_guess" "collector"
  ..$ skip   : num 1
  ..- attr(*, "class")= chr "col_spec"
```

Missing Values

```r
library(Hmisc)
library(DataExplorer)
```

```
Warning:   'DataExplorer'   R   4.3.2
```

```r
plot_missing(penguins_raw)
```



Descriptive Statistics

Summary Statistics for Numeric Variables

```r
library(dplyr)
```

```
 'dplyr'

     'package:Hmisc':

  src, summarize

     'package:stats':

  filter, lag

     'package:base':

  intersect, setdiff, setequal, union
```

```r
numeric_vars <- data.frame(Variable_Name = character(), stringsAsFactors = FALSE)

for (i in 1:ncol(penguins_raw)) {
  if (is.numeric(penguins_raw[[i]])) {
    numeric_vars <- rbind(numeric_vars,
                          data.frame(Variable_Name = colnames(penguins_raw)[i],
                                     stringsAsFactors = FALSE))
  }
}

print(numeric_vars)
```

```
        Variable_Name
1       Sample Number
2  Culmen Length (mm)
3   Culmen Depth (mm)
4 Flipper Length (mm)
5       Body Mass (g)
6   Delta 15 N (o/oo)
7   Delta 13 C (o/oo)
```

```r
library(dplyr)
library(knitr)

calculate_stats <- function(x) {
  data.frame(
    Mean = mean(x, na.rm = TRUE),
    Median = median(x, na.rm = TRUE),
    SD = sd(x, na.rm = TRUE),
    Min = min(x, na.rm = TRUE),
    Max = max(x, na.rm = TRUE)
  )
}

all_stats <- data.frame(Variable = character(),
                         Mean = numeric(),
                         Median = numeric(),
                         SD = numeric(),
                         Min = numeric(),
                         Max = numeric(),
                         stringsAsFactors = FALSE)
```

```r
for (var in numeric_vars$Variable_Name) {
  stats <- calculate_stats(penguins_raw[[var]])
  stats <- cbind(Variable = var, stats)
  all_stats <- rbind(all_stats, stats)
}

print(kable(all_stats, caption = "Summary Statistics for Numeric Variables"))
```

```
Table: Summary Statistics for Numeric Variables

|Variable            |        Mean|      Median|          SD|         Min|         Max|
|:-------------------|-----------:|-----------:|-----------:|----------:|----------:|
|Sample Number       |   63.151163|   58.000000|  40.4301990|     1.00000|   152.00000|
|Culmen Length (mm)  |   43.921930|   44.450000|   5.4595837|    32.10000|    59.60000|
|Culmen Depth (mm)   |   17.151170|   17.300000|   1.9747932|    13.10000|    21.50000|
|Flipper Length (mm) |  200.915205|  197.000000|  14.0617137|   172.00000|   231.00000|
|Body Mass (g)       | 4201.754386| 4050.000000| 801.9545357| 2700.00000| 6300.00000|
|Delta 15 N (o/oo)   |    8.733382|    8.652405|   0.5517703|     7.63220|    10.02544|
|Delta 13 C (o/oo)   |  -25.686291|  -25.833520|   0.7939612|   -27.01854|   -23.78767|
```

```r
calculate_stats <- function(x) {
  data.frame(
    Q25 = quantile(x, 0.25, na.rm = TRUE),
    Q50 = quantile(x, 0.50, na.rm = TRUE),
    Q75 = quantile(x, 0.75, na.rm = TRUE),
    IQR = IQR(x, na.rm = TRUE)
  )
}
all_stats <- data.frame(Variable = character(),
                         Q25 = numeric(),
                         Q50 = numeric(),
                         Q75 = numeric(),
                         IQR = numeric(),
                         stringsAsFactors = FALSE)

for (var in numeric_vars$Variable_Name) {
  stats <- calculate_stats(penguins_raw[[var]])
  stats <- cbind(Variable = var, stats)
  all_stats <- rbind(all_stats, stats)
}

print(all_stats)
```

```
                 Variable        Q25         Q50         Q75          IQR
25%         Sample Number   29.00000   58.000000   95.250000   66.2500000
25%1  Culmen Length (mm)   39.22500   44.450000   48.500000    9.2750000
25%2   Culmen Depth (mm)   15.60000   17.300000   18.700000    3.1000000
25%3 Flipper Length (mm)  190.00000  197.000000  213.000000   23.0000000
25%4        Body Mass (g) 3550.00000 4050.000000 4750.000000 1200.0000000
25%5   Delta 15 N (o/oo)    8.29989    8.652405    9.172123    0.8722325
25%6   Delta 13 C (o/oo)  -26.32030  -25.833520  -25.062050    1.2582550
```

Data Visualization

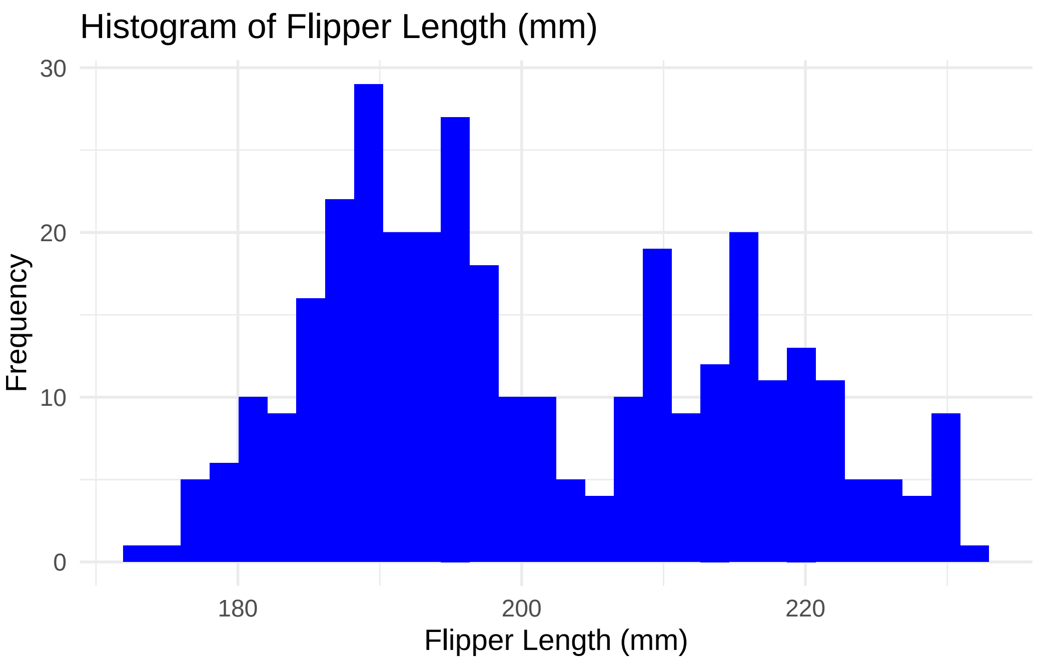
1.Histograms

```
library(ggplot2)

ggplot(penguins_raw, aes(x = `Flipper Length (mm)`)) +
  geom_histogram(fill = "blue") +
  labs(title = "Histogram of Flipper Length (mm)",
                          x = "Flipper Length (mm)", y = "Frequency") +
  theme_minimal()
```

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

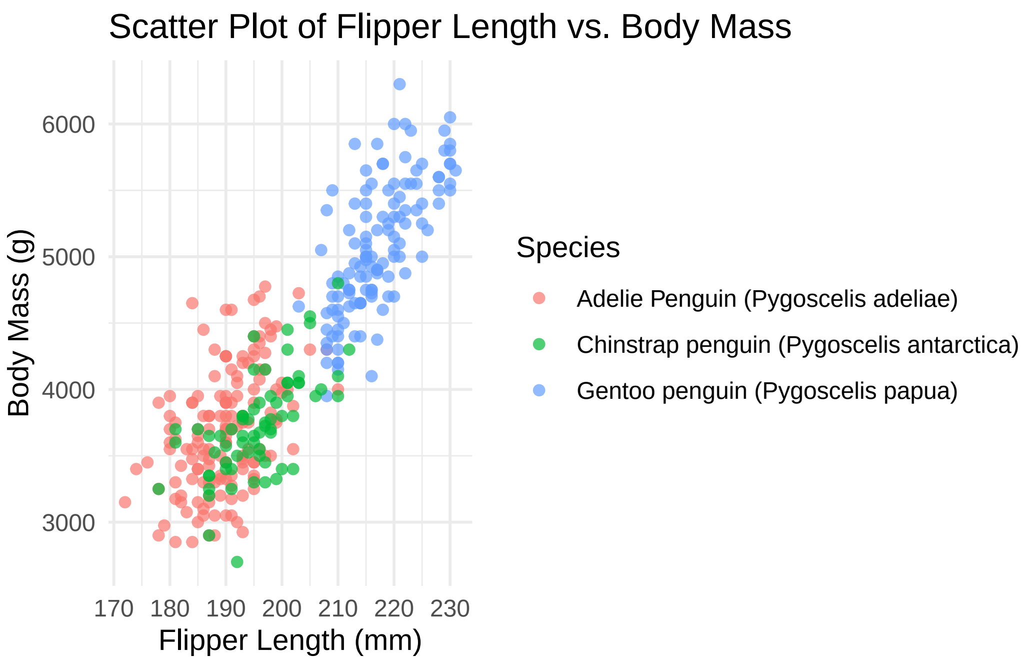
```
Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



2.Scatter Plots

```
ggplot(penguins_raw, aes(x = `Flipper Length (mm)`, y = `Body Mass (g)`)) +
  geom_point(aes(color = Species), alpha = 0.7) +
  labs(title = "Scatter Plot of Flipper Length vs. Body Mass", x = "Flipper Length (mm)", y =
  theme_minimal()
```

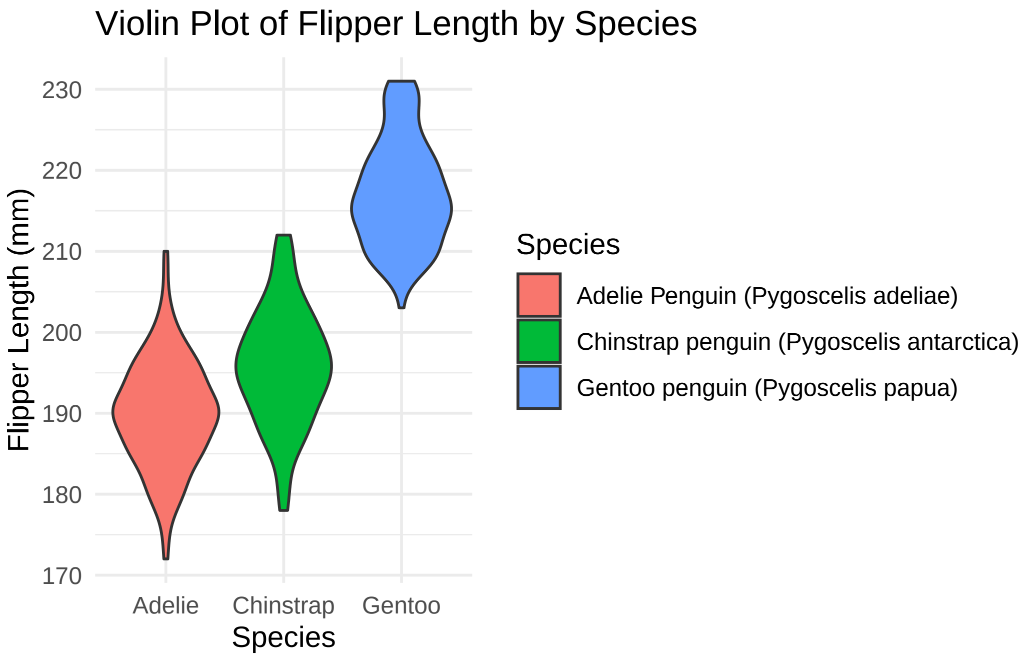
```
Warning: Removed 2 rows containing missing values (`geom_point()`).
```



3.Violin Plots

```
ggplot(penguins_raw, aes(x = Species, y = `Flipper Length (mm)`, fill = Species)) +
  geom_violin() +
  labs(title = "Violin Plot of Flipper Length by Species",
       x = "Species",
       y = "Flipper Length (mm)") +
  scale_x_discrete(labels = c("Adelie Penguin (Pygoscelis adeliae)" = "Adelie", "Chinstrap pe
  theme_minimal()
```

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



Violin Plot of Flipper Length by Species