My Report H24101094

YU FAN SHIU

2024-09-16

Table of contents

Summary Staistic

```
library(palmerpenguins)
```

```
Warning:   'palmerpenguins'   R   4.2.3
```

```
head(penguins_raw)
```

```
# A tibble: 6 x 17
  studyName `Sample Number` Species          Region Island Stage `Individual ID`
  <chr>               <dbl> <chr>            <chr>  <chr>  <chr> <chr>
1 PAL0708                 1 Adelie Penguin ~ Anvers Torge~ Adul~ N1A1
2 PAL0708                 2 Adelie Penguin ~ Anvers Torge~ Adul~ N1A2
3 PAL0708                 3 Adelie Penguin ~ Anvers Torge~ Adul~ N2A1
```

```
4 PAL0708                 4 Adelie Penguin ~ Anvers Torge~ Adul~ N2A2
5 PAL0708                 5 Adelie Penguin ~ Anvers Torge~ Adul~ N3A1
6 PAL0708                 6 Adelie Penguin ~ Anvers Torge~ Adul~ N3A2
# i 10 more variables: `Clutch Completion` <chr>, `Date Egg` <date>,
#   `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>,
#   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>,
#   `Delta 15 N (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

```
library(dplyr)
```

```
Warning:   'dplyr'   R   4.2.3
```

```
   'dplyr'

     'package:stats':

    filter, lag

     'package:base':

    intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
Warning:   'lubridate'   R   4.2.3
```

```
   'lubridate'

     'package:base':

    date, intersect, setdiff, union
```

```r
#
identify_var_types <- function(data, exclude = c()) {
  var_types <- sapply(data, class)
  var_types <- var_types[!names(var_types) %in% exclude]
  list(
    numeric = names(var_types[var_types %in% c("numeric", "integer")]),
    categorical = names(var_types[var_types %in% c("character", "factor")]),
    date = names(var_types[var_types == "Date"])
  )
}

#
summarize_numeric <- function(data, numeric_vars) {
  data %>%
    select(all_of(numeric_vars)) %>%
    summary()
}

#
summarize_categorical <- function(data, categorical_vars) {
  lapply(categorical_vars, function(var) {
    data %>%
      count(!!sym(var)) %>%
      mutate(percentage = n / sum(n) * 100)
  })
}

#
summarize_date <- function(data, date_vars) {
  lapply(date_vars, function(var) {
    data %>%
      summarise(
        min_date = min(!!sym(var), na.rm = TRUE),
        max_date = max(!!sym(var), na.rm = TRUE),
        mean_date = mean(!!sym(var), na.rm = TRUE),
        median_date = median(!!sym(var), na.rm = TRUE),
        n_unique_dates = n_distinct(!!sym(var)),
        most_common_year = names(which.max(table(year(!!sym(var))))),
        most_common_month = names(which.max(table(month(!!sym(var))))),
        most_common_day = names(which.max(table(day(!!sym(var)))))
      )
  })
```

```r
}

#
analyze_dataset <- function(data, exclude = c()) {
  var_types <- identify_var_types(data, exclude)

  results <- list(
    numeric_summary = summarize_numeric(data, var_types$numeric),
    categorical_summary = summarize_categorical(data, var_types$categorical),
    date_summary = summarize_date(data, var_types$date)
  )

  return(results)
}

#      "Individual ID"
results <- analyze_dataset(penguins_raw, exclude = c("Individual ID", "Comments"))

print(results$numeric_summary)
```

```
 Sample Number    Culmen Length (mm) Culmen Depth (mm) Flipper Length (mm)
 Min.   :  1.00   Min.   :32.10      Min.   :13.10     Min.   :172.0
 1st Qu.: 29.00   1st Qu.:39.23      1st Qu.:15.60     1st Qu.:190.0
 Median : 58.00   Median :44.45      Median :17.30     Median :197.0
 Mean   : 63.15   Mean   :43.92      Mean   :17.15     Mean   :200.9
 3rd Qu.: 95.25   3rd Qu.:48.50      3rd Qu.:18.70     3rd Qu.:213.0
 Max.   :152.00   Max.   :59.60      Max.   :21.50     Max.   :231.0
                  NA's   :2          NA's   :2         NA's   :2
 Body Mass (g)  Delta 15 N (o/oo) Delta 13 C (o/oo)
 Min.   :2700   Min.   : 7.632    Min.   :-27.02
 1st Qu.:3550   1st Qu.: 8.300    1st Qu.:-26.32
 Median :4050   Median : 8.652    Median :-25.83
 Mean   :4202   Mean   : 8.733    Mean   :-25.69
 3rd Qu.:4750   3rd Qu.: 9.172    3rd Qu.:-25.06
 Max.   :6300   Max.   :10.025    Max.   :-23.79
 NA's   :2      NA's   :14        NA's   :13
```

```r
print(results$categorical_summary)
```

```
[[1]]
# A tibble: 3 x 3
```

```
  studyName     n percentage
  <chr>     <int>      <dbl>
1 PAL0708     110       32.0
2 PAL0809     114       33.1
3 PAL0910     120       34.9

[[2]]
# A tibble: 3 x 3
  Species                                        n percentage
  <chr>                                      <int>      <dbl>
1 Adelie Penguin (Pygoscelis adeliae)          152       44.2
2 Chinstrap penguin (Pygoscelis antarctica)     68       19.8
3 Gentoo penguin (Pygoscelis papua)            124       36.0

[[3]]
# A tibble: 1 x 3
  Region     n percentage
  <chr>  <int>      <dbl>
1 Anvers   344        100

[[4]]
# A tibble: 3 x 3
  Island        n percentage
  <chr>     <int>      <dbl>
1 Biscoe      168       48.8
2 Dream       124       36.0
3 Torgersen    52       15.1

[[5]]
# A tibble: 1 x 3
  Stage                   n percentage
  <chr>               <int>      <dbl>
1 Adult, 1 Egg Stage    344        100

[[6]]
# A tibble: 2 x 3
  `Clutch Completion`     n percentage
  <chr>               <int>      <dbl>
1 No                     36       10.5
2 Yes                   308       89.5

[[7]]
# A tibble: 3 x 3
```

```
  Sex        n percentage
  <chr>  <int>      <dbl>
1 FEMALE   165       48.0
2 MALE     168       48.8
3 <NA>      11        3.20
```

```
print(results$date_summary)
```

```
[[1]]
# A tibble: 1 x 8
  min_date   max_date   mean_date  median_date n_unique_dates most_common_year
  <date>     <date>     <date>     <date>                <int> <chr>
1 2007-11-09 2009-12-01 2008-11-27 2008-11-09               50 2009
# i 2 more variables: most_common_month <chr>, most_common_day <chr>
```

Missing Values

```
library(DataExplorer)
```

```
Warning:   'DataExplorer'   R   4.2.3
```

```
library(Hmisc)
```

```
Warning:   'Hmisc'   R   4.2.2
```

```
    lattice
```

```
    survival
```

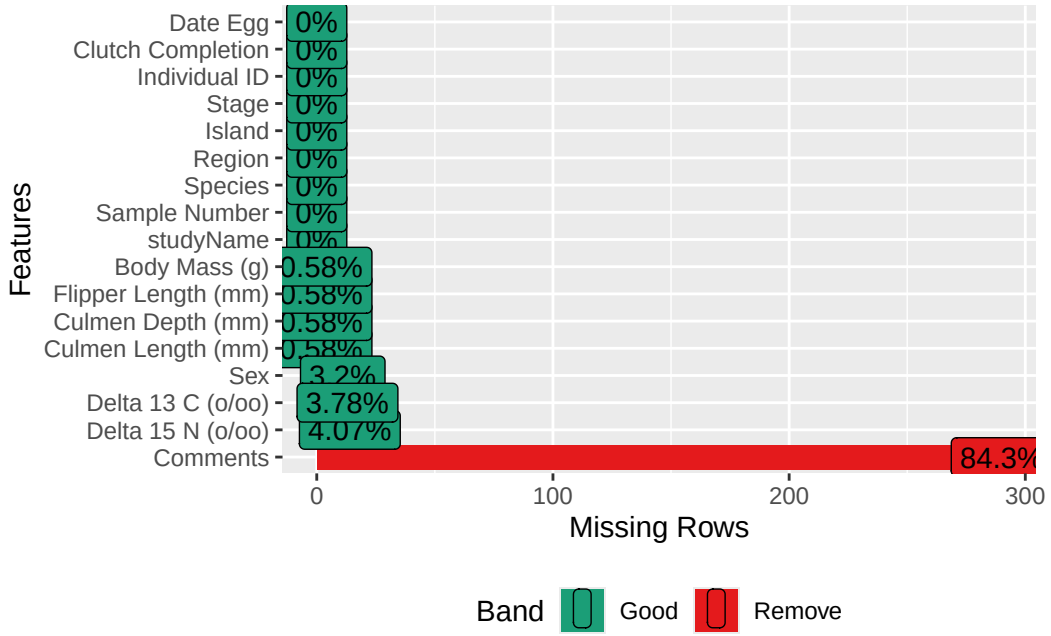
```
    Formula
```

```
    ggplot2
```

```
Warning:   'ggplot2'   R   4.2.3
```

```
  'Hmisc'

    'package:dplyr':

   src, summarize

    'package:base':

   format.pval, units
```

```
plot_missing(penguins_raw)
```



Penguin mass vs. flipper length

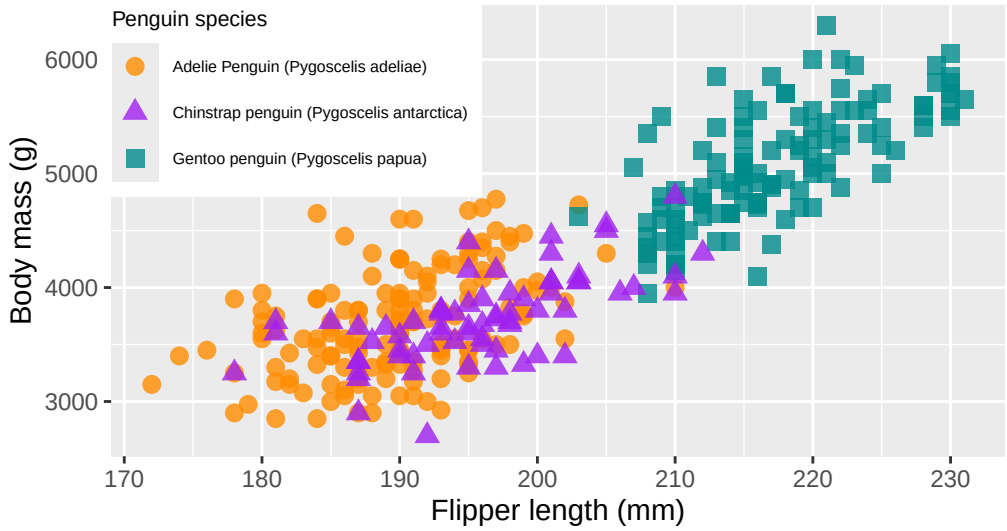
```
mass_flipper <- ggplot(data = penguins_raw,
                       aes(x = `Flipper Length (mm)`,
                           y = `Body Mass (g)`)) +
  geom_point(aes(color = Species,
```

```
                shape = Species),
            size = 3,
            alpha = 0.8) +
  scale_color_manual(values = c("darkorange","purple","cyan4")) +
  labs(title = "Penguin size, Palmer Station LTER",
       subtitle = "Flipper length and body mass for Adelie, Chinstrap and Gentoo Penguins",
       x = "Flipper length (mm)\n",
       y = "\nBody mass (g)",
       color = "Penguin species",
       shape = "Penguin species") +
  theme(legend.position = c(0.2, 0.8),
        plot.title.position = "plot",
        plot.caption = element_text(hjust = 0, face= "italic"),
        plot.caption.position = "plot",
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 6))
```

```
Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2
3.5.0.
i Please use the `legend.position.inside` argument of `theme()` instead.
```

```
mass_flipper
```

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```

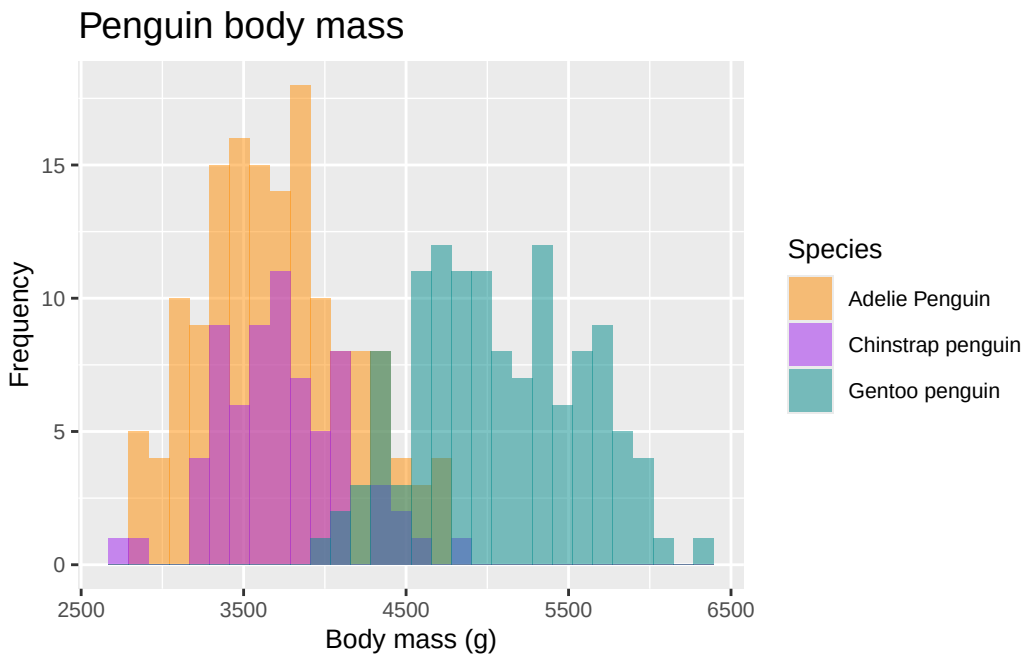


Body mass histograms

```
mass_hist <- ggplot(data = penguins_raw, aes(x = `Body Mass (g)`)) +
  geom_histogram(aes(fill = Species), alpha = 0.5, position = "identity") +
  scale_fill_manual(
    values = c("darkorange", "purple", "cyan4"),
    labels = c("Adelie Penguin", "Chinstrap penguin", "Gentoo penguin")
  ) +
  labs(x = "Body mass (g)", y = "Frequency", title = "Penguin body mass") +
  theme(plot.title = element_text(size = 14),
        axis.title = element_text(size = 10),
        axis.text = element_text(size = 8),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8))

mass_hist
```

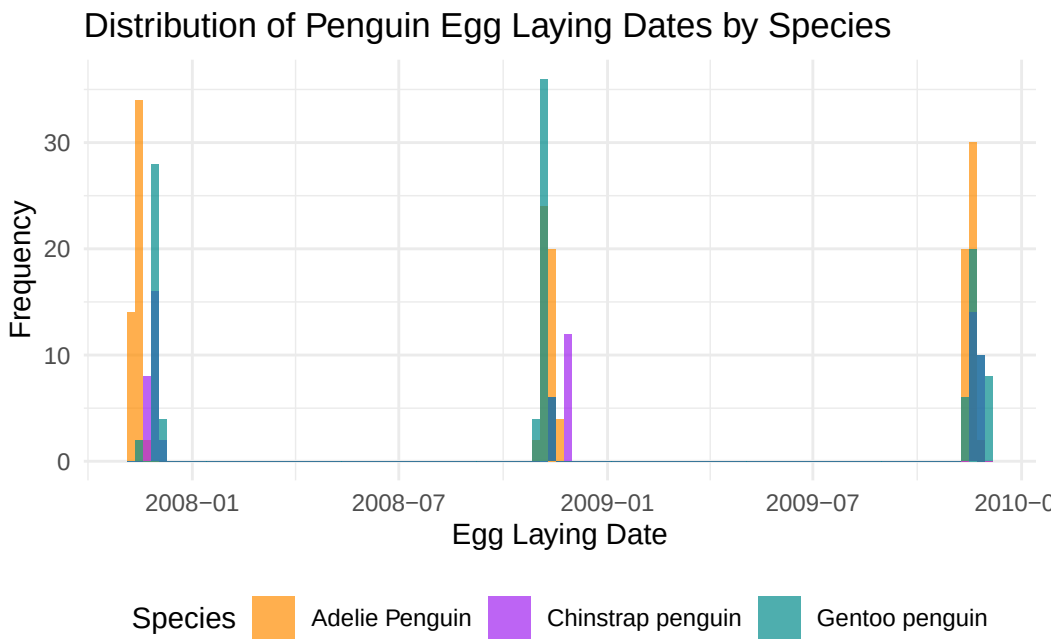
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_bin()`).
```

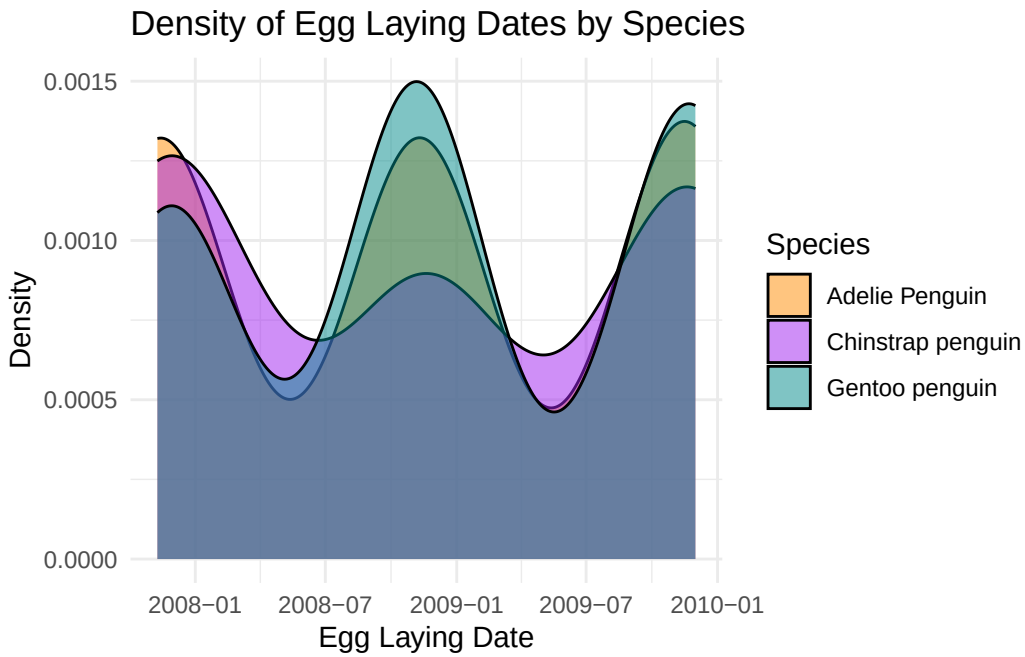


Date Egg

```
ggplot(penguins_raw, aes(x = `Date Egg`, fill = Species)) +
  geom_histogram(binwidth = 7, position = "identity", alpha = 0.7) +
  scale_fill_manual(
    values = c("darkorange", "purple", "cyan4"),
    labels = c("Adelie Penguin", "Chinstrap penguin", "Gentoo penguin")
  ) +
  labs(
    title = "Distribution of Penguin Egg Laying Dates by Species",
    x = "Egg Laying Date",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```




```
ggplot(penguins_raw, aes(x = `Date Egg`, fill = Species)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(
    values = c("darkorange", "purple", "cyan4"),
    labels = c("Adelie Penguin", "Chinstrap penguin", "Gentoo penguin")
  ) +
  labs(title = "Density of Egg Laying Dates by Species",
       x = "Egg Laying Date",
       y = "Density") +
  theme_minimal()
```


Density of Egg Laying Dates by Species
0.0015
0.0010
0.0005
0.0000
Density
2008−01
2008−07
2009−01
2009−07
2010−01
Egg Laying Date
Species
Adelie Penguin
Chinstrap penguin
Gentoo penguin