# HW1

H24106036李祐君

2024-09-16

## Table of contents

## Read data

```r
library(palmerpenguins)
data <- penguins_raw
head(data)
```

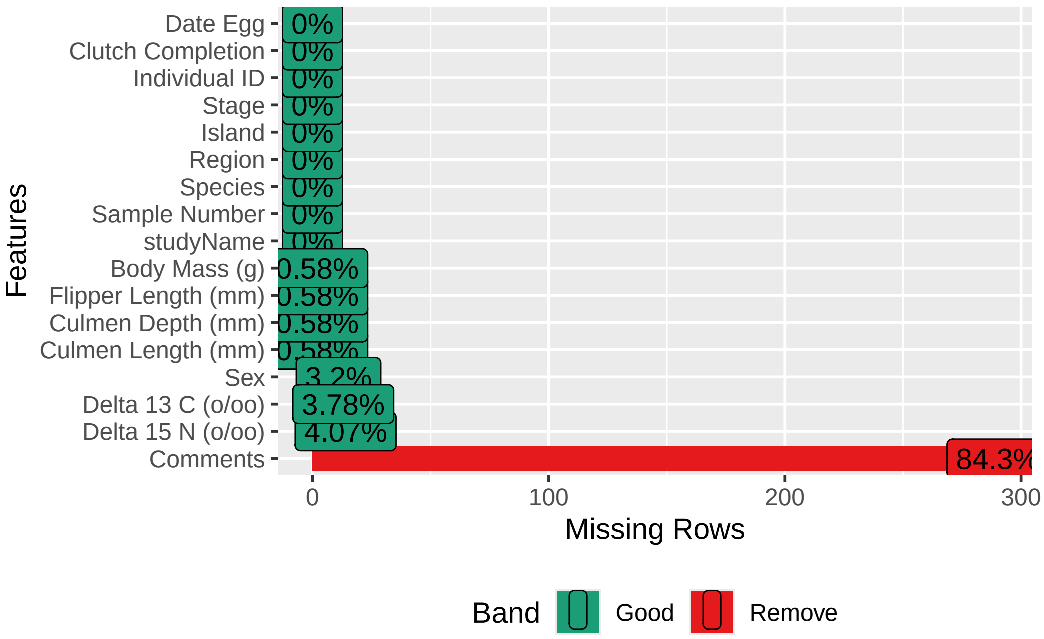
```
# A tibble: 6 x 17
  studyName `Sample Number` Species          Region Island Stage `Individual ID`
  <chr>               <dbl> <chr>            <chr>  <chr>  <chr> <chr>
1 PAL0708                 1 Adelie Penguin ~ Anvers Torge~ Adul~ N1A1
2 PAL0708                 2 Adelie Penguin ~ Anvers Torge~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

## Summary Staistic

### Missing Values

```
library(DataExplorer)
plot_missing(data)
```



我們得知在comments項有非常多的缺失值

### summary for whole data

使用三種方法 了解每個特徵的類別及分布情況

```r
library(Hmisc)
```

```
Attaching package: 'Hmisc'


The following objects are masked from 'package:base':

    format.pval, units
```

```r
latex(describe(data), file = "", caption.placement = "top")
```

## data
## 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
```

### △ 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
```

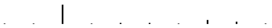
```
lowest : 7.6322  7.63452 7.63884 7.68528 7.6887 , highest: 9.93727 9.98044 10.0202 10.0237 10.0254
```

### △ 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.
```

```
library(pander)
pander(summary(data))
```

Table 1: Table continues below

| studyName | Sample Number | Species | Region |
|---|---|---|---|
| Length:344 | Min. : 1.00 | Length:344 | Length:344 |
| Class :character | 1st Qu.: 29.00 | Class :character | Class :character |
| Mode :character | Median : 58.00 | Mode :character | Mode :character |
| NA | Mean : 63.15 | NA | NA |
| NA | 3rd Qu.: 95.25 | NA | NA |
| NA | Max. :152.00 | NA | NA |
| NA | NA | NA | NA |

Table 2: Table continues below

| Island | Stage | Individual ID | Clutch Completion |
|---|---|---|---|
| Length:344 | Length:344 | Length:344 | Length:344 |
| Class :character | Class :character | Class :character | Class :character |
| Mode :character | Mode :character | Mode :character | Mode :character |
| NA | NA | NA | NA |
| NA | NA | NA | NA |
| NA | NA | NA | NA |
| NA | NA | NA | NA |

Table 3: Table continues below

| Date Egg | Culmen Length (mm) | Culmen Depth (mm) |
|---|---|---|
| Min. :2007-11-09 | Min. :32.10 | Min. :13.10 |
| 1st Qu.:2007-11-28 | 1st Qu.:39.23 | 1st Qu.:15.60 |
| Median :2008-11-09 | Median :44.45 | Median :17.30 |
| Mean :2008-11-27 | Mean :43.92 | Mean :17.15 |
| 3rd Qu.:2009-11-16 | 3rd Qu.:48.50 | 3rd Qu.:18.70 |
| Max. :2009-12-01 | Max. :59.60 | Max. :21.50 |
| NA | NA’s :2 | NA’s :2 |

Table 4: Table continues below

| Flipper Length (mm) | Body Mass (g) | Sex | Delta 15 N (o/oo) |
|---|---|---|---|
| Min. :172.0 | Min. :2700 | Length:344 | Min. : 7.632 |
| 1st Qu.:190.0 | 1st Qu.:3550 | Class :character | 1st Qu.: 8.300 |
| Median :197.0 | Median :4050 | Mode :character | Median : 8.652 |
| Mean :200.9 | Mean :4202 | NA | Mean : 8.733 |
| 3rd Qu.:213.0 | 3rd Qu.:4750 | NA | 3rd Qu.: 9.172 |
| Max. :231.0 | Max. :6300 | NA | Max. :10.025 |
| NA's :2 | NA's :2 | NA | NA's :14 |

| Delta 13 C (o/oo) | Comments |
|---|---|
| Min. :-27.02 | Length:344 |
| 1st Qu.:-26.32 | Class :character |
| Median :-25.83 | Mode :character |
| Mean :-25.69 | NA |
| 3rd Qu.:-25.06 | NA |
| Max. :-23.79 | NA |
| NA's :13 | NA |

```
library(summarytools)
```

```
Attaching package: 'summarytools'


The following objects are masked from 'package:Hmisc':

    label, label<-
```

```
dfSummary(data)
```

Data Frame Summary
data
Dimensions: 344 x 17
Duplicates: 0

| No | Variable | Stats / Values | Freqs (% of Valid) | Graph | Valid | Missing |
|---|---|---|---|---|---|---|
| 1 | studyName [character] | 1. PAL0708 2. PAL0809 3. PAL0910 | 110 (32.0%) 114 (33.1%) 120 (34.9%) | IIIIII IIIIII IIIIII | 344 (100.0%) | 0 (0.0%) |
| 2 | Sample Number [numeric] | Mean (sd) : 63.2 (40.4) min < med < max: 1 < 58 < 152 IQR (CV) : 66.2 (0.6) | 152 distinct values | : : : : : : : : . . : : : : : : : : : : : : : : : : : : : : : : : | 344 (100.0%) | 0 (0.0%) |
| 3 | Species [character] | 1. Adelie Penguin (Pygosceli 2. Chinstrap penguin (Pygosc 3. Gentoo penguin (Pygosceli | 152 (44.2%) 68 (19.8%) 124 (36.0%) | IIIIIIII III IIIIIII | 344 (100.0%) | 0 (0.0%) |
| 4 | Region [character] | 1. Anvers | 344 (100.0%) | IIIIIIIIIIIIIIIIIII | 344 (100.0%) | 0 (0.0%) |
| 5 | Island [character] | 1. Biscoe 2. Dream 3. Torgersen | 168 (48.8%) 124 (36.0%) 52 (15.1%) | IIIIIIIII IIIIIII III | 344 (100.0%) | 0 (0.0%) |
| 6 | Stage [character] | 1. Adult, 1 Egg Stage | 344 (100.0%) | IIIIIIIIIIIIIIIIIII | 344 (100.0%) | 0 (0.0%) |
| 7 | Individual ID [character] | 1. N13A1 2. N13A2 3. N18A1 4. N18A2 5. N21A1 6. N21A2 7. N22A1 8. N22A2 9. N23A1 10. N23A2 [ 180 others ] | 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 3 ( 0.9%) 314 (91.3%) |  | 344 (100.0%) | 0 (0.0%) |
| 8 | Clutch Completion [character] | 1. No 2. Yes | 36 (10.5%) 308 (89.5%) | II IIIIIIIIIIIIIIIII | 344 (100.0%) | 0 (0.0%) |
| 9 | Date Egg [Date] | min : 2007-11-09 med : 2008-11-09 max : 2009-12-01 range : 2y 0m 22d | 50 distinct values | . : : : : : : : : : : : : : : : | 344 (100.0%) | 0 (0.0%) |

| No | Variable | Stats / Values | Freqs (% of Valid) | Graph | Valid | Missing |
|---|---|---|---|---|---|---|
| 10 | Culmen Length (mm) [numeric] | Mean (sd) : 43.9 (5.5) min < med < max: 32.1 < 44.5 < 59.6 IQR (CV) : 9.3 (0.1) | 164 distinct values | . . : . : : : : : : : : : : : : : : : : : : . : : : : : : : : . | 342 (99.4%) | 2 (0.6%) |
| 11 | Culmen Depth (mm) [numeric] | Mean (sd) : 17.2 (2) min < med < max: 13.1 < 17.3 < 21.5 IQR (CV) : 3.1 (0.1) | 80 distinct values | : : : : . : : : . . : : : : : : : : : : : : : : . . | 342 (99.4%) | 2 (0.6%) |
| 12 | Flipper Length (mm) [numeric] | Mean (sd) : 200.9 (14.1) min < med < max: 172 < 197 < 231 IQR (CV) : 23 (0.1) | 55 distinct values | : . : : : : . . . : : : : : : : : : : : : : : : : | 342 (99.4%) | 2 (0.6%) |
| 13 | Body Mass (g) [numeric] | Mean (sd) : 4201.8 (802) min < med < max: 2700 < 4050 < 6300 IQR (CV) : 1200 (0.2) | 94 distinct values | : . : : : : : : : : : : . . : : : : : : | 342 (99.4%) | 2 (0.6%) |
| 14 | Sex [character] | 1. FEMALE 2. MALE | 165 (49.5%) 168 (50.5%) | IIIIIIIII IIIIIIIIII | 333 (96.8%) | 11 (3.2%) |
| 15 | Delta 15 N (o/oo) [numeric] | Mean (sd) : 8.7 (0.6) min < med < max: 7.6 < 8.7 < 10 IQR (CV) : 0.9 (0.1) | 330 distinct values | . : : : : . . . : : : : : : : : : : : : : . : : : : : : : : : : : | 330 (95.9%) | 14 (4.1%) |
| 16 | Delta 13 C (o/oo) [numeric] | Mean (sd) : -25.7 (0.8) min < med < max: -27 < -25.8 < -23.8 IQR (CV) : 1.3 (0) | 331 distinct values | : : . : : . : . : : : : : . : : : : : : | 331 (96.2%) | 13 (3.8%) |

| No | Variable | Stats / Values | Freqs (% of Valid) | Graph | Valid | Missing |
|---|---|---|---|---|---|---|
| 17 | Comments [character] | 1. Adult not sampled. 2. Adult not sampled. Nest n 3. Nest never observed with 4. Nest never observed with 5. No blood sample obtained 6. No blood sample obtained. 7. No delta15N data received 8. Not enough blood for isot 9. Sexing primers did not am 10. Sexing primers did not am | 1 ( 1.9%) 1 ( 1.9%) 34 (63.0%) 1 ( 1.9%) 2 ( 3.7%) 2 ( 3.7%) 1 ( 1.9%) 7 (13.0%) 4 ( 7.4%) 1 ( 1.9%) |  | 54 (15.7%) | 290 (84.3%) |

總共有344筆資料且有17個特徵，有些是類別型資料，有些則是連續型資料

大略瀏覽整體資料，接下來再細部觀察各個特徵並畫圖

### Observe and Plot some interesting features(in my opinion)

#### discrete(category) data

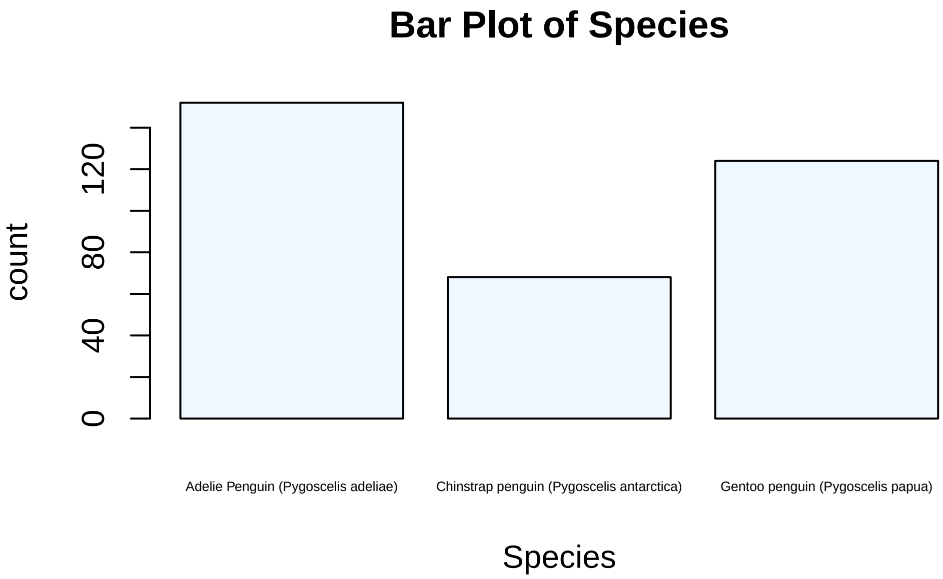
From above, we can see that we have three different species and island,but it just have one region and stage.

```r
table(data$Species)
```

```
      Adelie Penguin (Pygoscelis adeliae)
                                      152
Chinstrap penguin (Pygoscelis antarctica)
                                       68
```

```
        Gentoo penguin (Pygoscelis papua) 
                                     124 
```

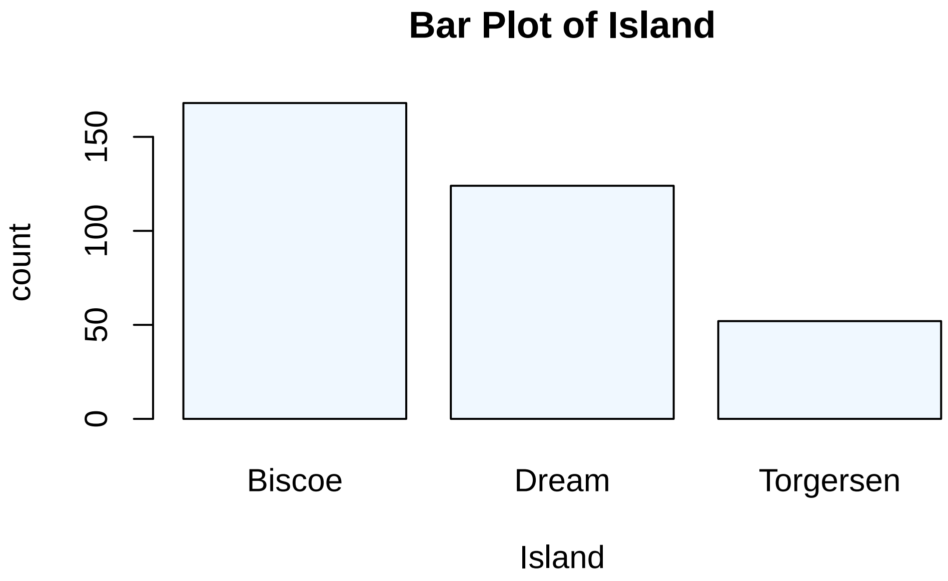
```
barplot(table(data$Species), main = "Bar Plot of Species",xlab = "Species", ylab = "count", c
        cex.names = 0.45)
```



```
table(data$Island)
```

```

   Biscoe      Dream Torgersen 
      168        124        52 
```

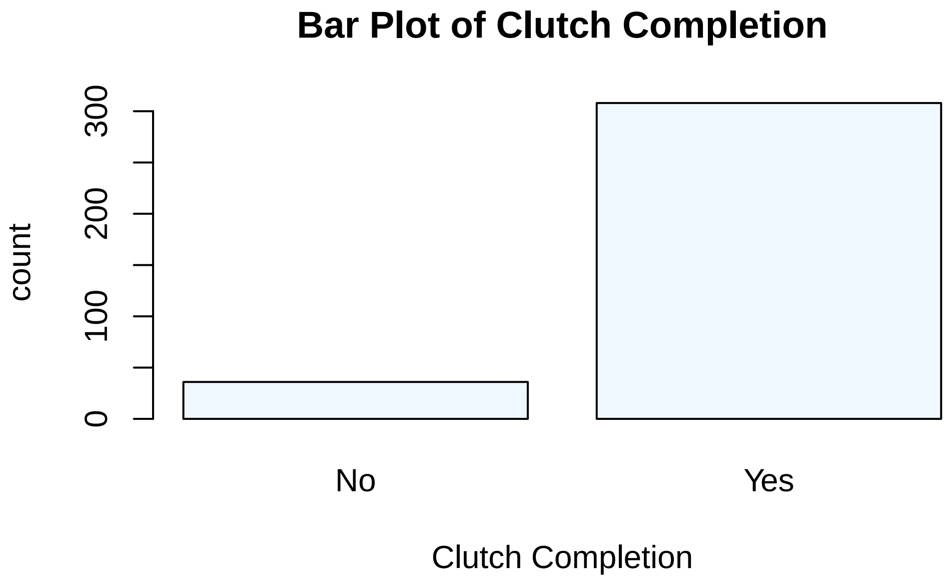
```
barplot(table(data$Island), main = "Bar Plot of Island",xlab = "Island", ylab = "count", col
```



```
table(data$`Clutch Completion`)
```

```
No Yes
36 308
```

```
barplot(table(data$`Clutch Completion`), main = "Bar Plot of Clutch Completion",xlab = "Clut
```

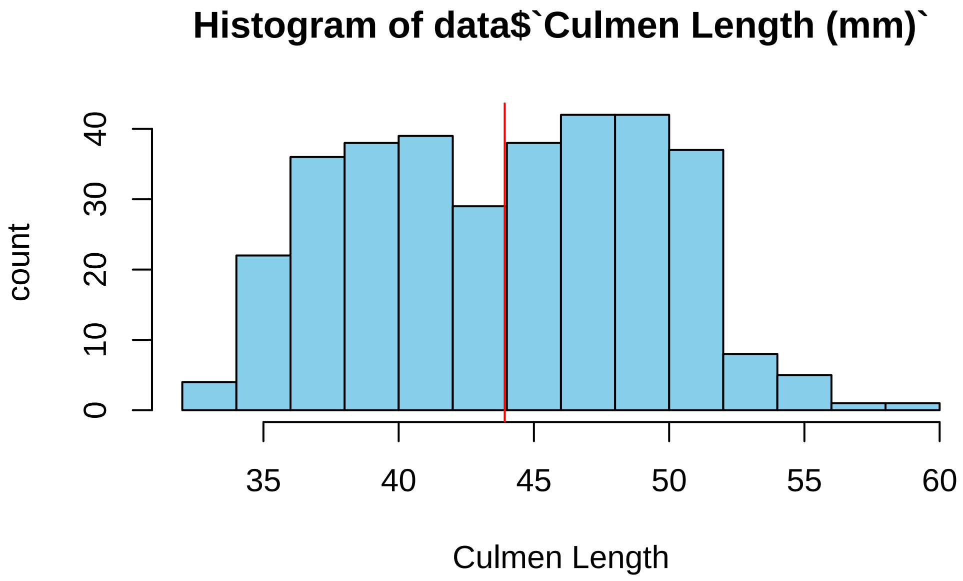


**continuous data**

```
summary(data$`Culmen Length (mm)`)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  32.10   39.23   44.45   43.92   48.50   59.60       2 
```

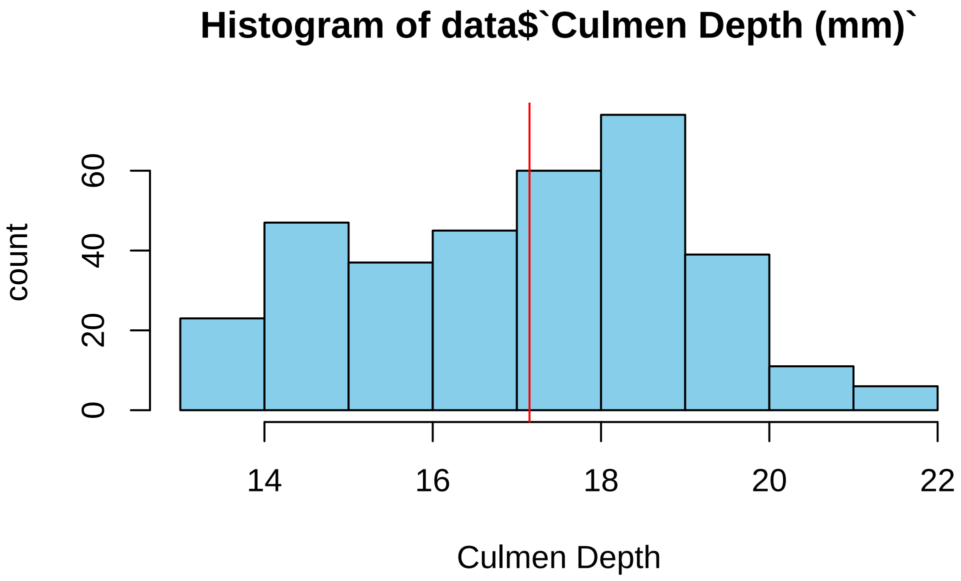
```
hist(data$`Culmen Length (mm)`,xlab="Culmen Length", ylab = "count",col="skyblue")
abline(v=43.92,col="red")
```



```
summary(data$`Culmen Depth (mm)`)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  13.10   15.60   17.30   17.15   18.70   21.50       2 
```

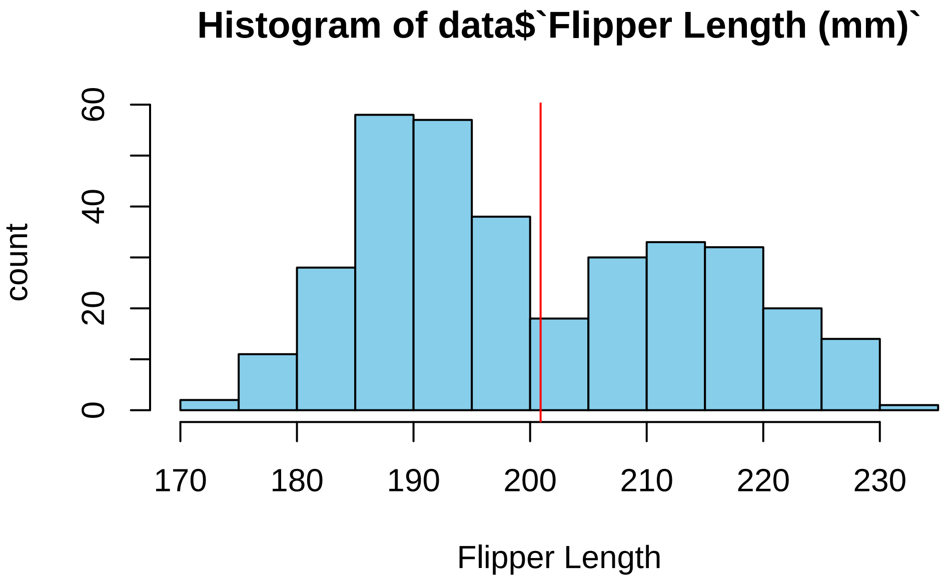
```
hist(data$`Culmen Depth (mm)`,xlab="Culmen Depth", ylab = "count",col="skyblue")
abline(v=17.15,col="red")
```



```r
summary(data$`Flipper Length (mm)`)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  172.0   190.0   197.0   200.9   213.0   231.0       2 
```

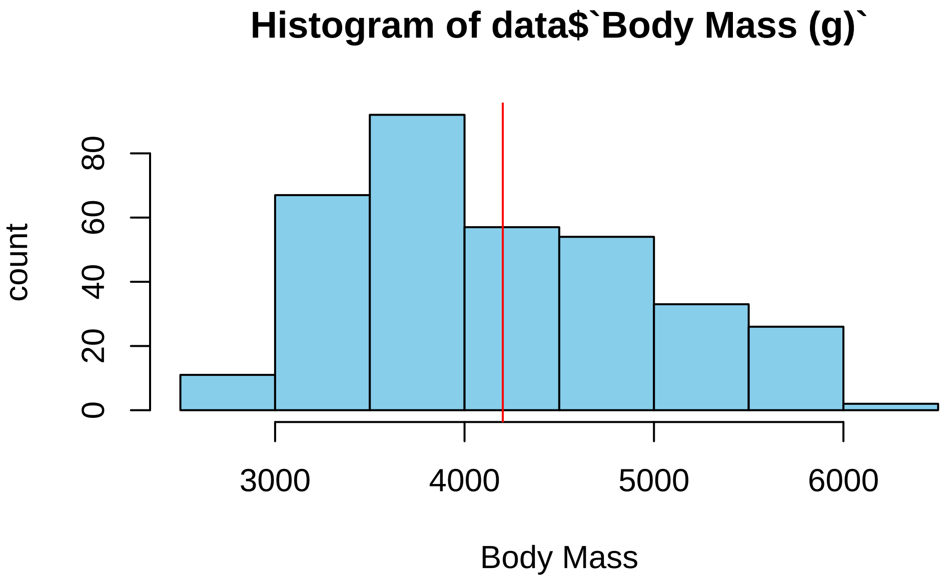
```r
hist(data$`Flipper Length (mm)`,xlab="Flipper Length", ylab = "count",col="skyblue")
abline(v=200.9,col="red")
```



```
summary(data$`Body Mass (g)`)
```

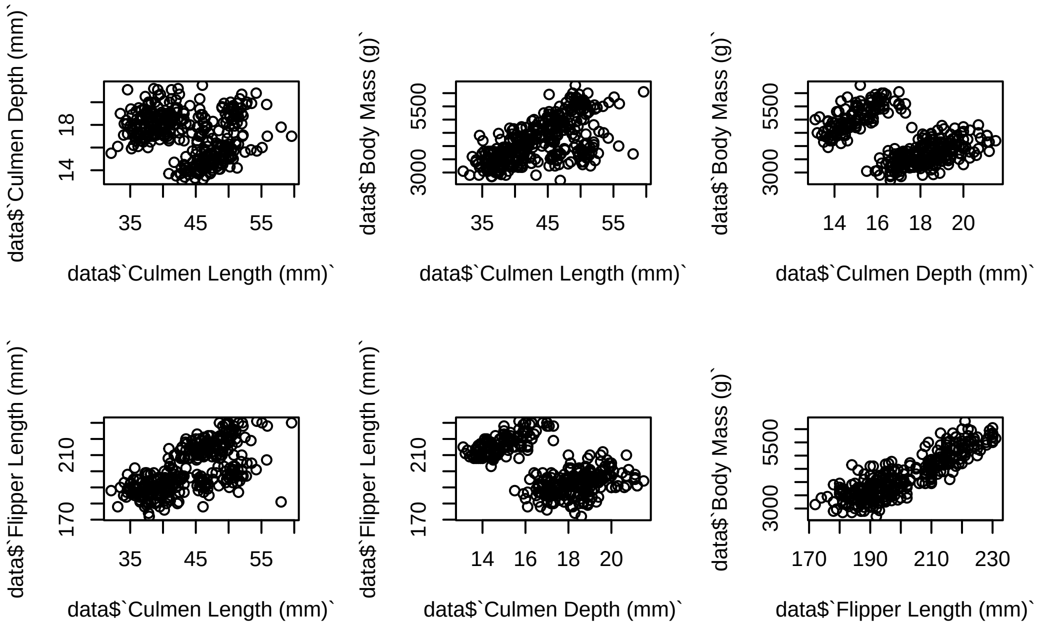
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   2700    3550    4050    4202    4750    6300       2 
```

```
hist(data$`Body Mass (g)`,xlab="Body Mass", ylab = "count",col="skyblue")
abline(v=4202,col="red")
```



看連續型資料相關性

```
layout(matrix(c(1,2,3,4,5,6),2,3))
plot(data$`Culmen Length (mm)`,data$`Culmen Depth (mm)`)
plot(data$`Culmen Length (mm)`,data$`Flipper Length (mm)`)
plot(data$`Culmen Length (mm)`,data$`Body Mass (g)`)
plot(data$`Culmen Depth (mm)`,data$`Flipper Length (mm)`)
plot(data$`Culmen Depth (mm)`,data$`Body Mass (g)`)
plot(data$`Flipper Length (mm)`,data$`Body Mass (g)`)
```



我們會發現這些資料幾乎都呈正相關，且有些有分群現象的感覺

因此我們將分群因素考慮再重新畫幾張分布圖

### groupby species

```
par(mfrow = c(1, 1))
library(table1)
```

```

Attaching package: 'table1'

The following objects are masked from 'package:summarytools':

    label, label<-

The following objects are masked from 'package:Hmisc':

    label, label<-, units
```

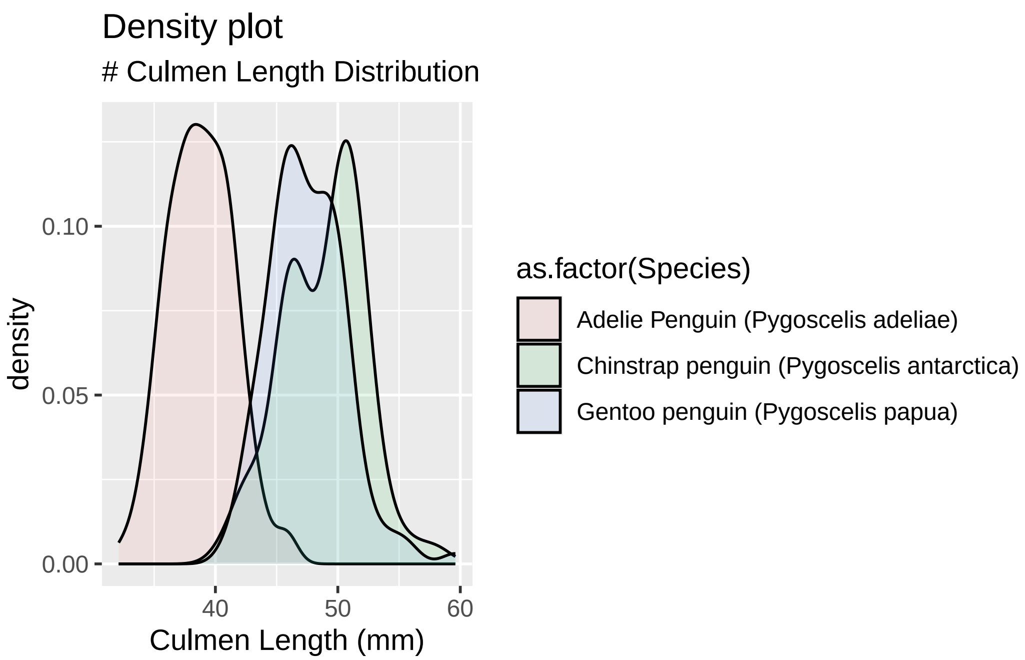
```
The following objects are masked from 'package:base':

    units, units<-
```

```r
library(kableExtra)
t1 <- table1(~ `Culmen Length (mm)` + `Culmen Depth (mm)` +
              `Flipper Length (mm)` + `Body Mass (g)` | Species, data)
# Output to PDF with LaTeX formatting
kable(as.data.frame(t1), "latex", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down")) %>%
  row_spec(0, bold = TRUE)
```

| | **Adelie Penguin (Pygoscelis adeliae)** | **Chinstrap penguin (Pygoscelis antarctica)** | **Gentoo penguin (Pygoscelis papua)** | **Overall** |
|---|---|---|---|---|
| | (N=152) | (N=68) | (N=124) | (N=344) |
| Culmen Length (mm) | | | | |
| Mean (SD) | 38.8 (2.66) | 48.8 (3.34) | 47.5 (3.08) | 43.9 (5.46) |
| Median [Min, Max] | 38.8 [32.1, 46.0] | 49.6 [40.9, 58.0] | 47.3 [40.9, 59.6] | 44.5 [32.1, 59.6] |
| Missing | 1 (0.7%) | 0 (0%) | 1 (0.8%) | 2 (0.6%) |
| Culmen Depth (mm) | | | | |
| Mean (SD) | 18.3 (1.22) | 18.4 (1.14) | 15.0 (0.981) | 17.2 (1.97) |
| Median [Min, Max] | 18.4 [15.5, 21.5] | 18.5 [16.4, 20.8] | 15.0 [13.1, 17.3] | 17.3 [13.1, 21.5] |
| Missing | 1 (0.7%) | 0 (0%) | 1 (0.8%) | 2 (0.6%) |
| Flipper Length (mm) | | | | |
| Mean (SD) | 190 (6.54) | 196 (7.13) | 217 (6.48) | 201 (14.1) |
| Median [Min, Max] | 190 [172, 210] | 196 [178, 212] | 216 [203, 231] | 197 [172, 231] |
| Missing | 1 (0.7%) | 0 (0%) | 1 (0.8%) | 2 (0.6%) |
| Body Mass (g) | | | | |
| Mean (SD) | 3700 (459) | 3730 (384) | 5080 (504) | 4200 (802) |
| Median [Min, Max] | 3700 [2850, 4780] | 3700 [2700, 4800] | 5000 [3950, 6300] | 4050 [2700, 6300] |
| Missing | 1 (0.7%) | 0 (0%) | 1 (0.8%) | 2 (0.6%) |

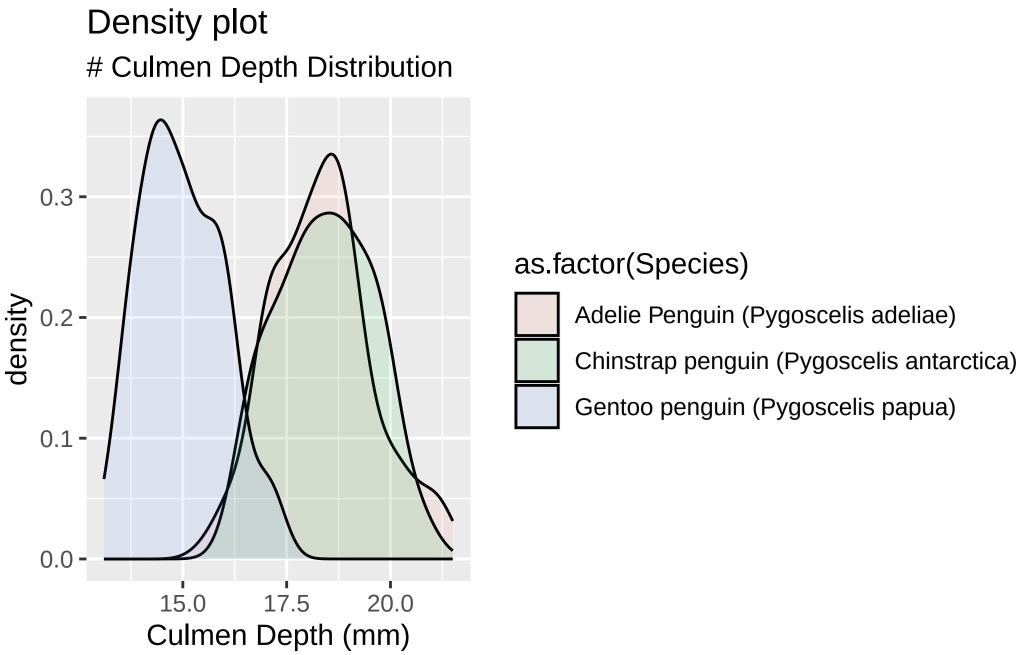
```r
library(ggplot2)
ggplot(data = data, aes(x = `Culmen Length (mm)`, fill = as.factor(Species))) +
  geom_density(alpha=0.1)+
      labs(title="Density plot",
          subtitle="# Culmen Length Distribution")
```

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



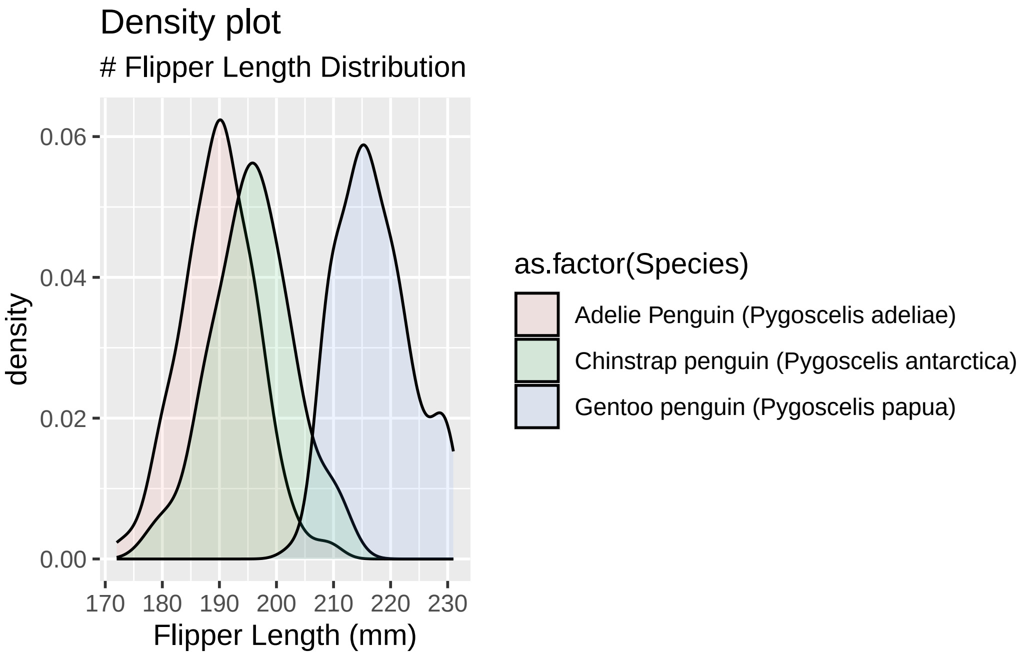
```r
ggplot(data = data, aes(x = `Culmen Depth (mm)`, fill = as.factor(Species))) +
  geom_density(alpha=0.1)+
      labs(title="Density plot",
          subtitle="# Culmen Depth Distribution")
```

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



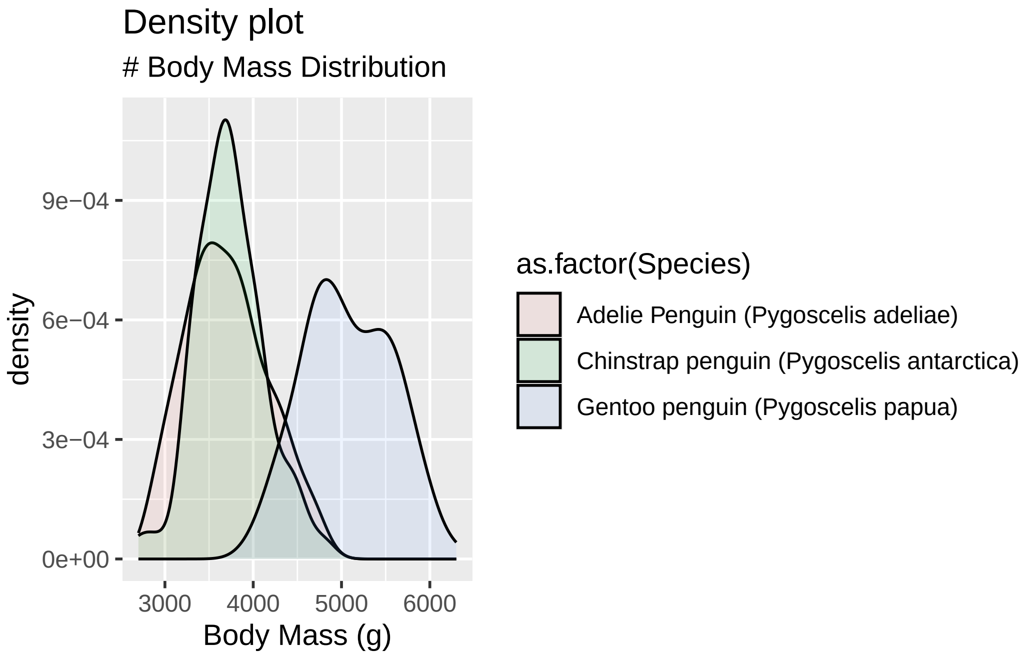
```
ggplot(data = data, aes(x = `Flipper Length (mm)`, fill = as.factor(Species))) +
  geom_density(alpha=0.1)+
      labs(title="Density plot",
          subtitle="# Flipper Length Distribution")
```

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



```
ggplot(data = data, aes(x = `Body Mass (g)`, fill = as.factor(Species))) +
  geom_density(alpha=0.1)+
      labs(title="Density plot",
          subtitle="# Body Mass Distribution")
```

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



我們可以發現Gentoo這個品種的企鵝跟另外兩種在我們擁有資料的這幾個特徵上有較顯著的差異
僅culmen length 是 Adelie企鵝與另外兩種差異較大