My Report

H24101052 洪嘉澤

2024-09-18

Table of contents

Statistical Thinking

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

Summary Staistic

```
library(Hmisc)
```

```
Warning: package 'Hmisc' was built under R version 4.3.3
```

```
Attaching package: 'Hmisc'
```

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

    format.pval, units
```

```
library(palmerpenguins)
```

```
Warning: package 'palmerpenguins' was built under R version 4.3.3
```

```
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

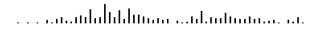
lowest : 32.1 33.1 33.5 34 34.1, highest: 55.1 55.8 55.9 58 59.6

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

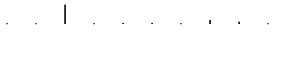
```
lowest : 7.6322  7.63452 7.63884 7.68528 7.6887 , highest: 9.93727 9.98044 10.0202 10.0237 10.0254
```

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.
```

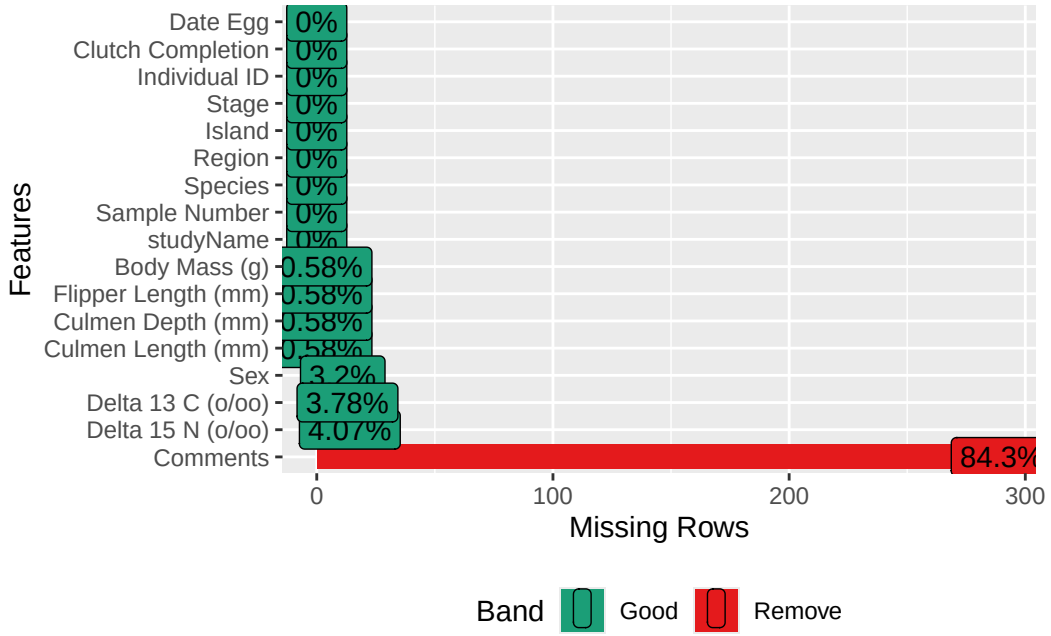
可以看出各變數的缺失數量、唯一值還有四分位數

Missing Values

```
library(DataExplorer)
```

```
Warning: package 'DataExplorer' was built under R version 4.3.3
```

```
plot_missing(penguins_raw)
```



可以看出缺失比例最多的是Comments，再來是氮同位素比值、碳同位素比值以及性別

Table 1

```
library(table1)
library(tidyverse)
A <- penguins_raw
A$Culmen_Length <- A$`Culmen Length (mm)`
A$Culmen_Depth <- A$`Culmen Depth (mm)`
A$Flipper_Length <- A$`Flipper Length (mm)`
A$Body_Mass <- A$`Body Mass (g)`
A$Species <- recode(A$Species,"Adelie Penguin (Pygoscelis adeliae)" = "Adelie",
                            "Gentoo penguin (Pygoscelis papua)" = "Gentoo",
                            "Chinstrap penguin (Pygoscelis antarctica)" = "Chinstrap")
str(A)
```

```
tibble [344 x 21] (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" "Adelie" "Adelie" "Adelie" ...
 $ 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
 $ Culmen_Length      : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ Culmen_Depth       : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ Flipper_Length     : num [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ Body_Mass          : num [1:344] 3750 3800 3250 NA 3450 ...
 - attr(*, "spec")=
  .. cols(
  ..   studyName = col_character(),
  ..   `Sample Number` = col_double(),
  ..   Species = col_character(),
```

```
..   Region = col_character(),
..   Island = col_character(),
..   Stage = col_character(),
..   `Individual ID` = col_character(),
..   `Clutch Completion` = col_character(),
..   `Date Egg` = col_date(format = ""),
..   `Culmen Length (mm)` = col_double(),
..   `Culmen Depth (mm)` = col_double(),
..   `Flipper Length (mm)` = col_double(),
..   `Body Mass (g)` = col_double(),
..   Sex = col_character(),
..   `Delta 15 N (o/oo)` = col_double(),
..   `Delta 13 C (o/oo)` = col_double(),
..   Comments = col_character()
.. )
```

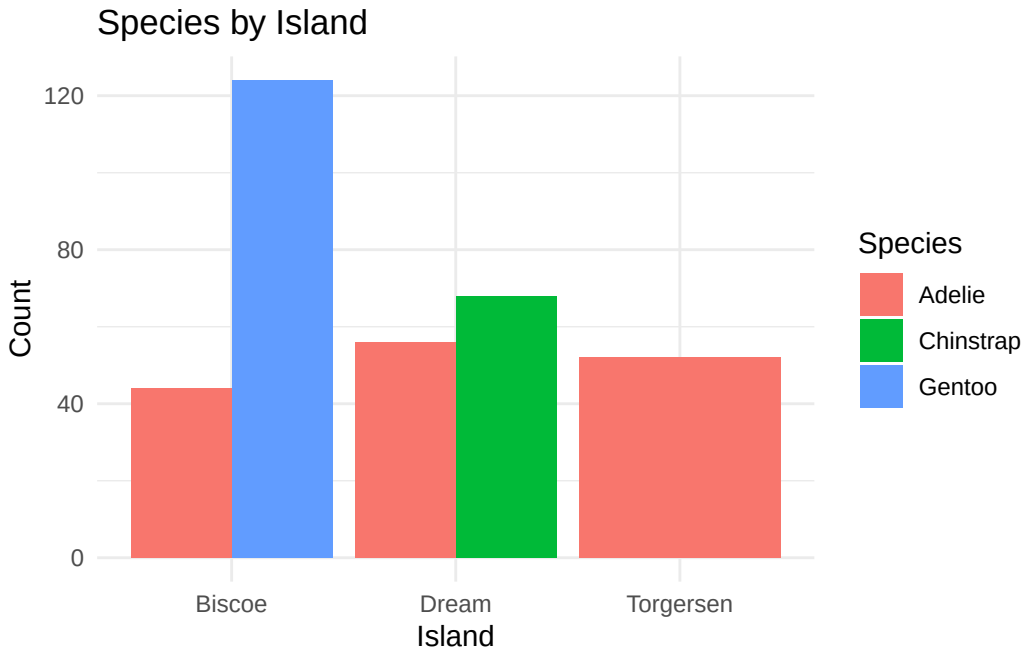
```
table1(~ Island + Culmen_Length + Culmen_Depth + Body_Mass| Species, data=A, topclass="Rtable
```

	Adelie	Chinstrap	Gentoo	Overall
	(N=152)	(N=68)	(N=124)	(N=344)
Island				
Biscoe	44 (28.9%)	0 (0%)	124 (100%)	168 (48.8%)
Dream	56 (36.8%)	68 (100%)	0 (0%)	124 (36.0%)
Torgersen	52 (34.2%)	0 (0%)	0 (0%)	52 (15.1%)
Culmen_Length				
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				
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%)
Body_Mass				
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%)

從表中能看出Adelie企鵝在三個島嶼上均存在，但Chinstrap以及Gentoo這兩種企鵝
分別只存在於Dream、Biscoe上。　其中以Chinstrap的平均喙長最長，中位數也位居第一。
從喙深來看可看出Adelie以及Chinstrap相當接近，而體重上則是以Gentoo高居第一。

Graph

```
library(ggplot2)
ggplot(A, aes(x = Island, fill = Species)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Species by Island",
       x = "Island",
       y = "Count")
```



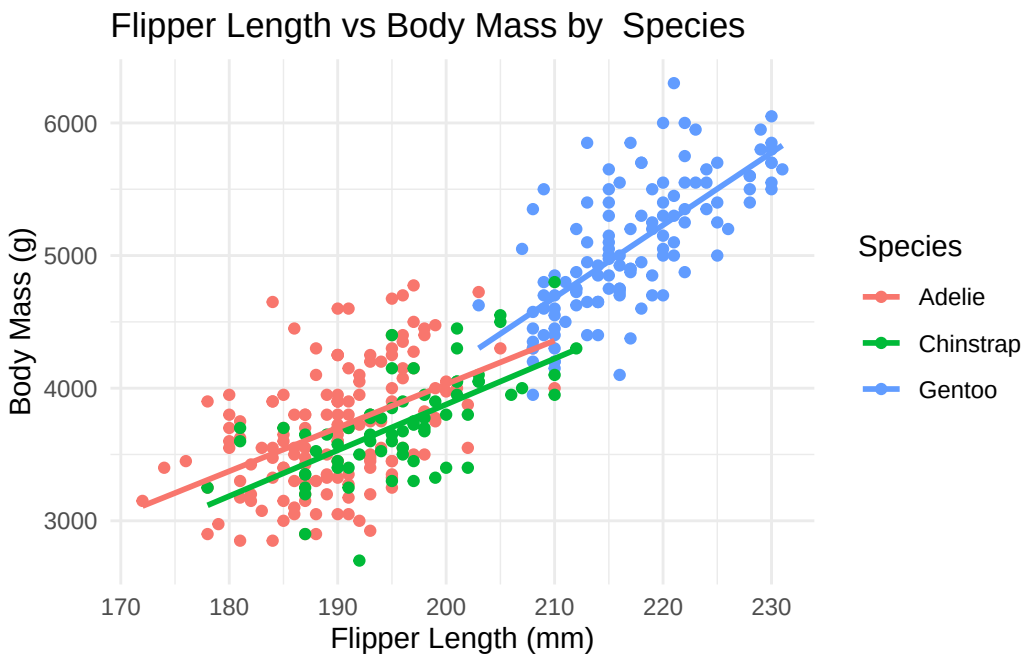
從此張圖可以看出三種企鵝於三個島嶼上的生存數量，其中Adelie平均存於三個島嶼中，而Gentoo只存於Biscoe，Chinstrap則只存於Dream中。

```
ggplot(A, aes(x = Flipper_Length, y = Body_Mass, color = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(title = "Flipper Length vs Body Mass by  Species",
       x = "Flipper Length (mm)",
       y = "Body Mass (g)",
       color = "Species")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

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

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



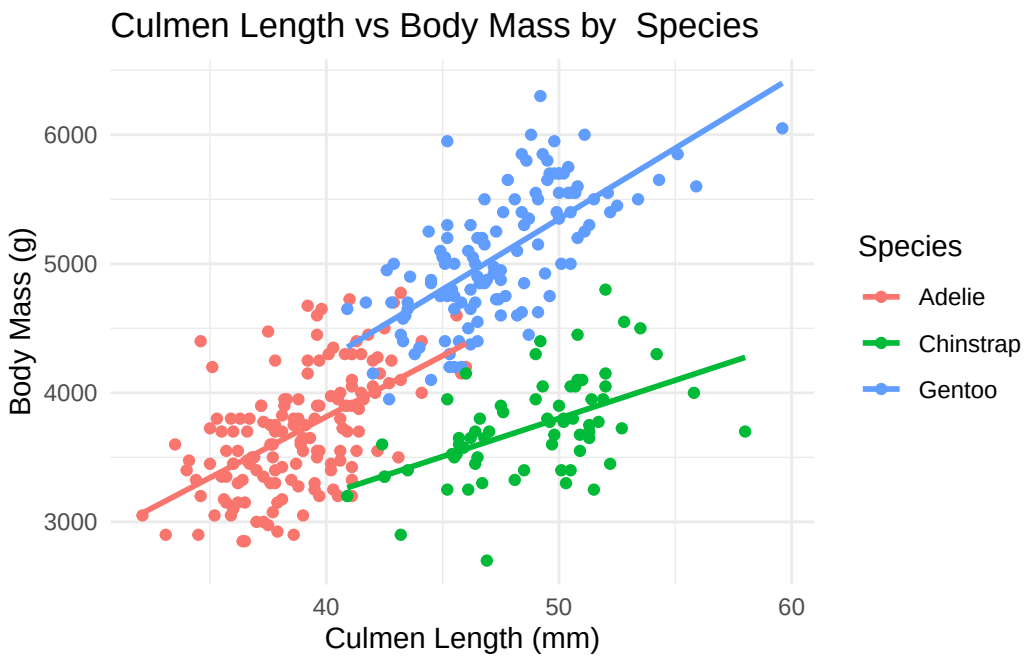
從此圖中可以看出Gentoo的翼展長度較長，且重量也比其他兩種企鵝重，翼展對於其重量影響較大。其餘兩種企鵝的翼展對於他們的重量影響較相近。

```
ggplot(A, aes(x = Culmen_Length, y = Body_Mass, color = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(title = "Culmen Length vs Body Mass by  Species",
       x = "Culmen Length (mm)",
       y = "Body Mass (g)",
       color = "Species")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

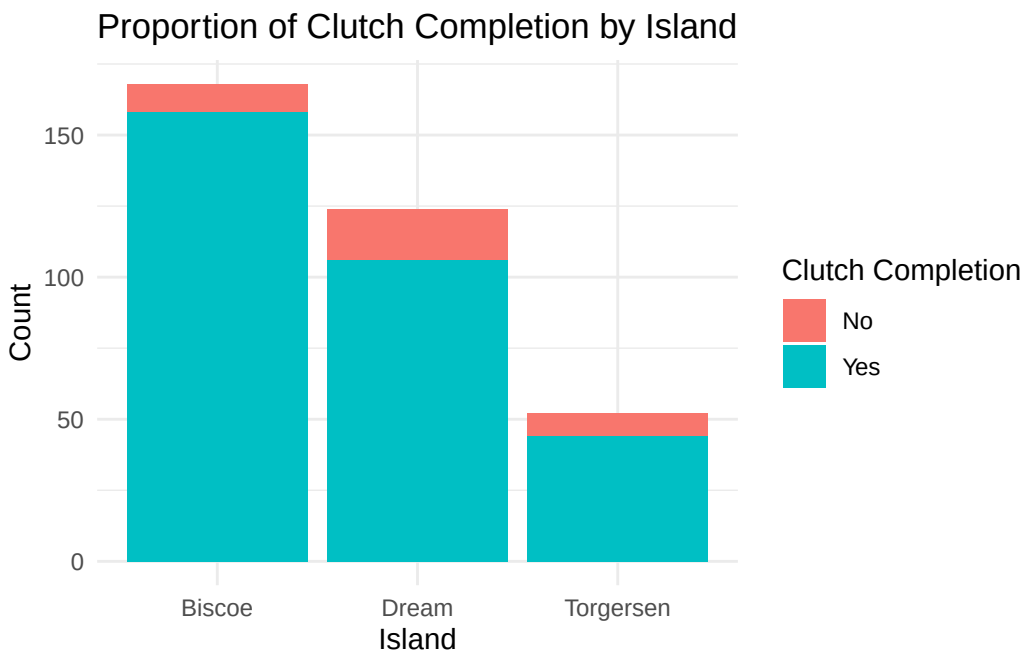
```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

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



從圖中可以看出嘴喙的長度對於Adelie以及Gentoo的身體重量影響程度較接近，但平均喙長是以Chinstrap以及Gentoo較接近，Adelie喙長則較短。

```r
ggplot(A, aes(x = Island, fill = `Clutch Completion`)) +
  geom_bar(position = "stack") +
  theme_minimal() +
  labs(title = "Proportion of Clutch Completion by Island",
       x = "Island",
       y = "Count",
       fill = "Clutch Completion")
```

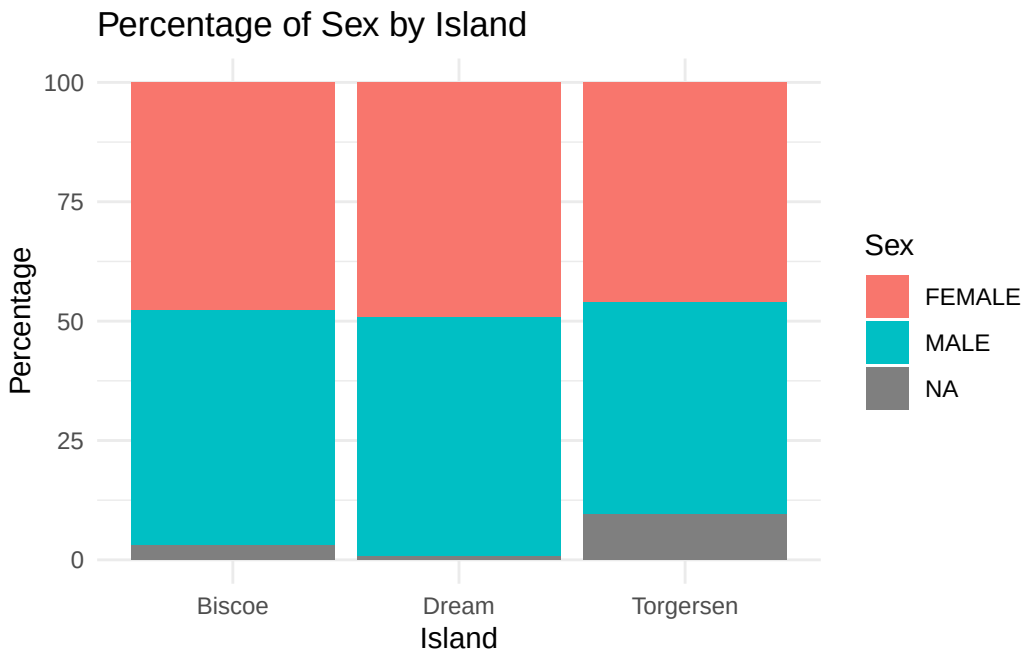


可看出每個島上大部分的企鵝皆有產下完整的一窩蛋。

```r
B <- A %>%
  group_by(Island, Sex) %>%
  summarise(Count = n()) %>%
  group_by(Island) %>%
  mutate(Percentage = Count / sum(Count) * 100)
```

```
`summarise()` has grouped output by 'Island'. You can override using the
`.groups` argument.
```

```r
ggplot(B, aes(x = Island, y = Percentage, fill = Sex)) +
  geom_bar(stat = "identity", position = "stack") +
  theme_minimal() +
  labs(title = "Percentage of Sex by Island",
       x = "Island",
       y = "Percentage",
       fill = "Sex")
```



可看出三個島嶼的企鵝性別比例相似，兩性別皆接近50%。