

Stat HW1

Hui-Ying HSU

Table of contents

install packages	1
delete NA	3
summary statistic	6
Bar plot	8
Histogram	11
correlation test	20
chi_square_test	21

install packages

```
options(repos = c(CRAN = "https://cloud.r-project.org/"))  
install.packages("palmerpenguins")
```

Installing package into 'C:/Users/anya3/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)

package 'palmerpenguins' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\anya3\AppData\Local\Temp\RtmpIBGuFj\downloaded_packages

```
library(palmerpenguins)
```

Warning: package 'palmerpenguins' was built under R version 4.4.1

```
data(package = 'palmerpenguins')
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>
```

```
install.packages("DataExplorer")
```

Installing package into 'C:/Users/anya3/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)

package 'DataExplorer' successfully unpacked and MD5 sums checked

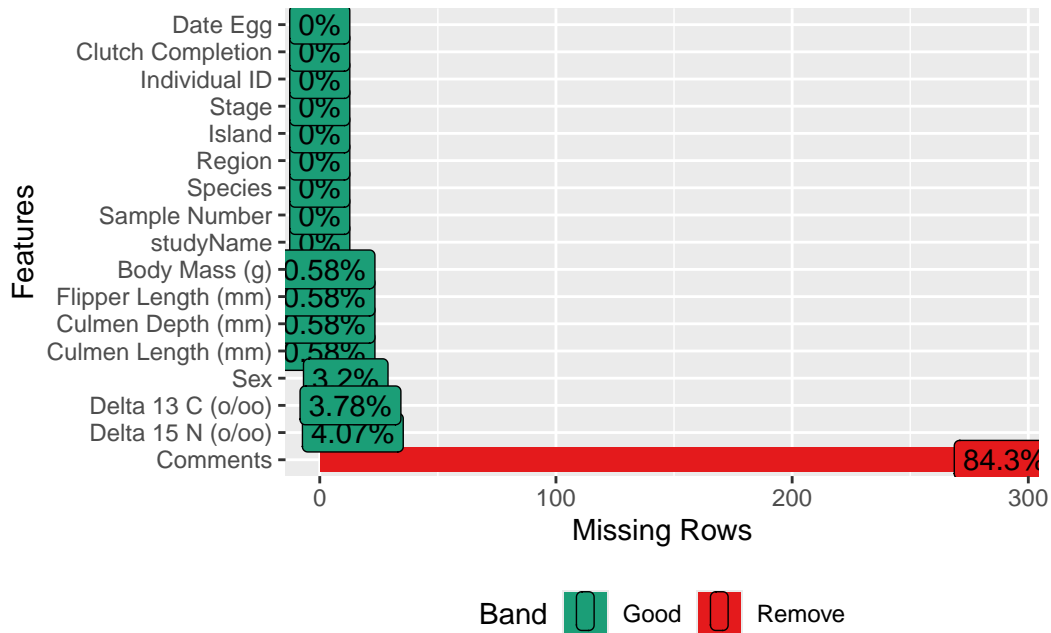
The downloaded binary packages are in
C:\Users\anya3\AppData\Local\Temp\RtmpIBGuFj\downloaded_packages

```
library("DataExplorer")
```

Warning: package 'DataExplorer' was built under R version 4.4.1

#plot missing

```
plot_missing(penguins_raw)
```



delete NA

```
library(dplyr)
```

Attaching package: 'dplyr'

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

filter, lag

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

intersect, setdiff, setequal, union

```
# delete NA
penguins_clean <- penguins_raw %>%filter(!is.na(`Body Mass (g)`), !is.na(`Culmen Length (mm)`))
penguins_clean
```

```
# A tibble: 324 x 17
```

```
  studyName `Sample Number` Species      Region Island Stage `Individual ID`
  <chr>          <dbl> <chr>          <chr> <chr> <chr> <chr>
1 PAL0708          2 Adelie Penguin~ Anvers Torge~ Adul~ N1A2
2 PAL0708          3 Adelie Penguin~ Anvers Torge~ Adul~ N2A1
3 PAL0708          5 Adelie Penguin~ Anvers Torge~ Adul~ N3A1
4 PAL0708          6 Adelie Penguin~ Anvers Torge~ Adul~ N3A2
5 PAL0708          7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
6 PAL0708          8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
7 PAL0708         15 Adelie Penguin~ Anvers Torge~ Adul~ N8A1
8 PAL0708         17 Adelie Penguin~ Anvers Torge~ Adul~ N9A1
9 PAL0708         18 Adelie Penguin~ Anvers Torge~ Adul~ N9A2
10 PAL0708         19 Adelie Penguin~ Anvers Torge~ Adul~ N10A1
# i 314 more rows
# 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>
```

```
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 (Pygoscelis adeliae)" ...
$ 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 Stage" "Adult, 1 Egg Stage" ...
$ 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"

```

```
library(Hmisc)
```

Attaching package: 'Hmisc'

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

```
src, summarize
```

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

```
format.pval, units
```

```
#latex(describe(penguins_raw), file = "", caption.placement = "top")
```

summary statistic

```
summary(penguins_raw)
```

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
	Mean : 63.15		
	3rd Qu.: 95.25		
	Max. :152.00		

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

Date Egg	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)
Min. :2007-11-09	Min. :32.10	Min. :13.10	Min. :172.0

1st Qu.:2007-11-28	1st Qu.:39.23	1st Qu.:15.60	1st Qu.:190.0
Median :2008-11-09	Median :44.45	Median :17.30	Median :197.0
Mean :2008-11-27	Mean :43.92	Mean :17.15	Mean :200.9
3rd Qu.:2009-11-16	3rd Qu.:48.50	3rd Qu.:18.70	3rd Qu.:213.0
Max. :2009-12-01	Max. :59.60	Max. :21.50	Max. :231.0
	NA's :2	NA's :2	NA's :2

Body Mass (g)	Sex	Delta 15 N (o/oo)	Delta 13 C (o/oo)
Min. :2700	Length:344	Min. : 7.632	Min. : -27.02
1st Qu.:3550	Class :character	1st Qu.: 8.300	1st Qu.: -26.32
Median :4050	Mode :character	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

Comments
Length:344
Class :character
Mode :character

```
summary(penguins_clean)
```

studyName	Sample Number	Species	Region
Length:324	Min. : 1.00	Length:324	Length:324
Class :character	1st Qu.: 31.00	Class :character	Class :character
Mode :character	Median : 59.50	Mode :character	Mode :character
	Mean : 64.60		
	3rd Qu.: 96.25		
	Max. :152.00		

Island	Stage	Individual ID	Clutch Completion
Length:324	Length:324	Length:324	Length:324
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

Date Egg	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)
Min. :2007-11-09	Min. :32.10	Min. :13.10	Min. :172.0
1st Qu.:2007-11-29	1st Qu.:39.50	1st Qu.:15.57	1st Qu.:190.0

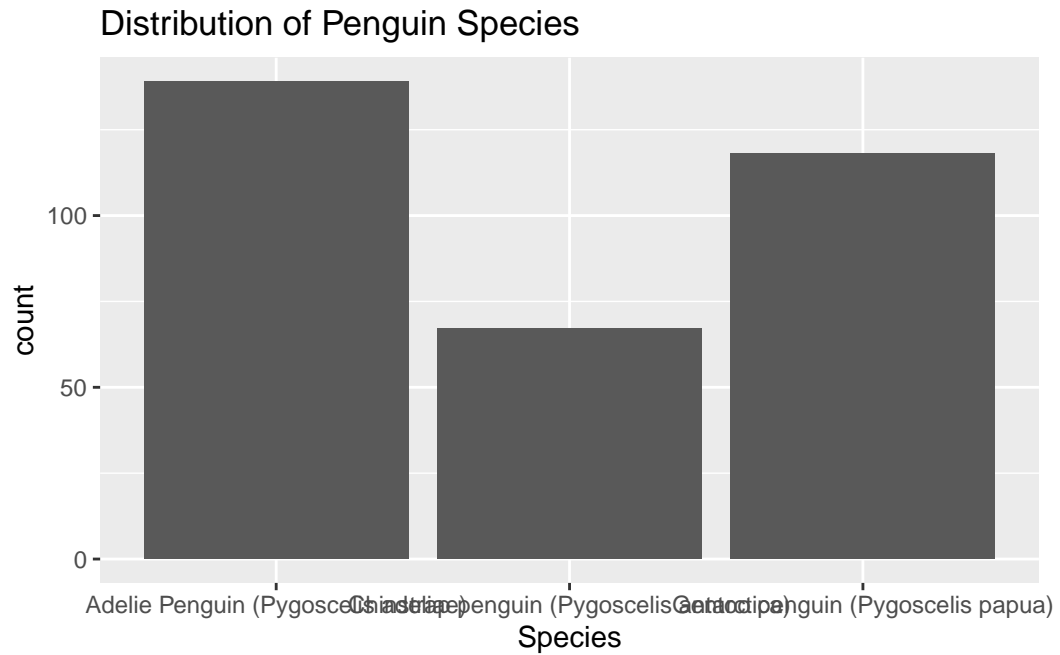
Median :2008-11-10	Median :44.95	Median :17.30	Median :197.0
Mean :2008-12-10	Mean :44.05	Mean :17.13	Mean :201.2
3rd Qu.:2009-11-16	3rd Qu.:48.70	3rd Qu.:18.60	3rd Qu.:213.0
Max. :2009-12-01	Max. :59.60	Max. :21.50	Max. :231.0
Body Mass (g)	Sex	Delta 15 N (o/oo)	Delta 13 C (o/oo)
Min. :2700	Length:324	Min. : 7.632	Min. : -27.02
1st Qu.:3550	Class :character	1st Qu.: 8.304	1st Qu.: -26.33
Median :4050	Mode :character	Median : 8.659	Median : -25.84
Mean :4214		Mean : 8.740	Mean : -25.69
3rd Qu.:4800		3rd Qu.: 9.181	3rd Qu.: -25.06
Max. :6300		Max. :10.025	Max. : -23.89
Comments			
Length:324			
Class :character			
Mode :character			

Bar plot

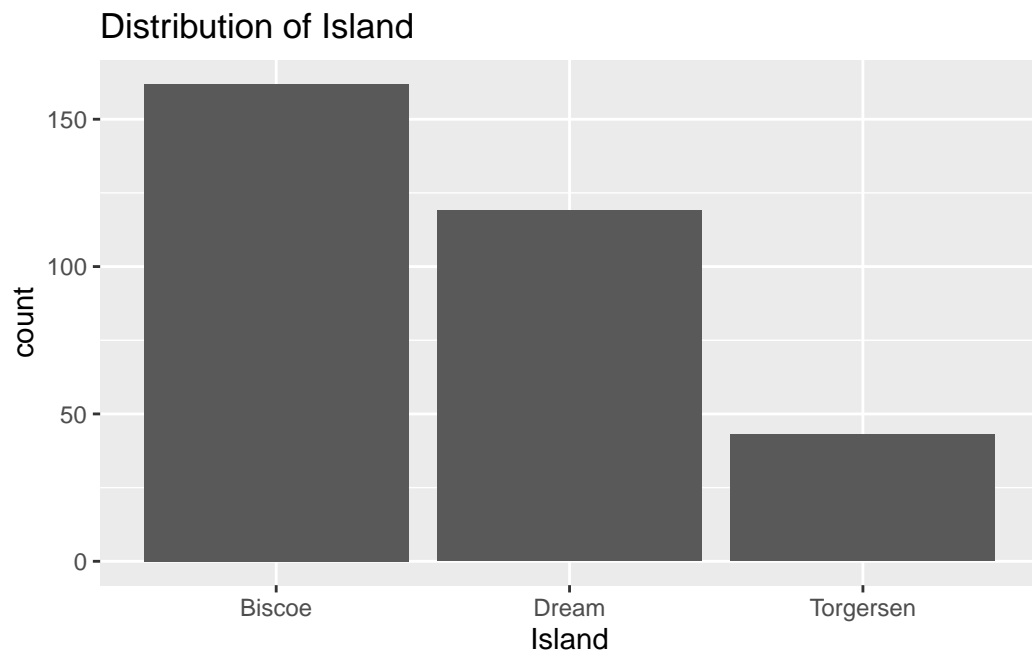
```
library(ggplot2)
```

Warning: package 'ggplot2' was built under R version 4.4.1

```
# Bar plot
# Species
ggplot(penguins_clean, aes(x = Species)) +geom_bar() +ggtitle("Distribution of Penguin Species")
```

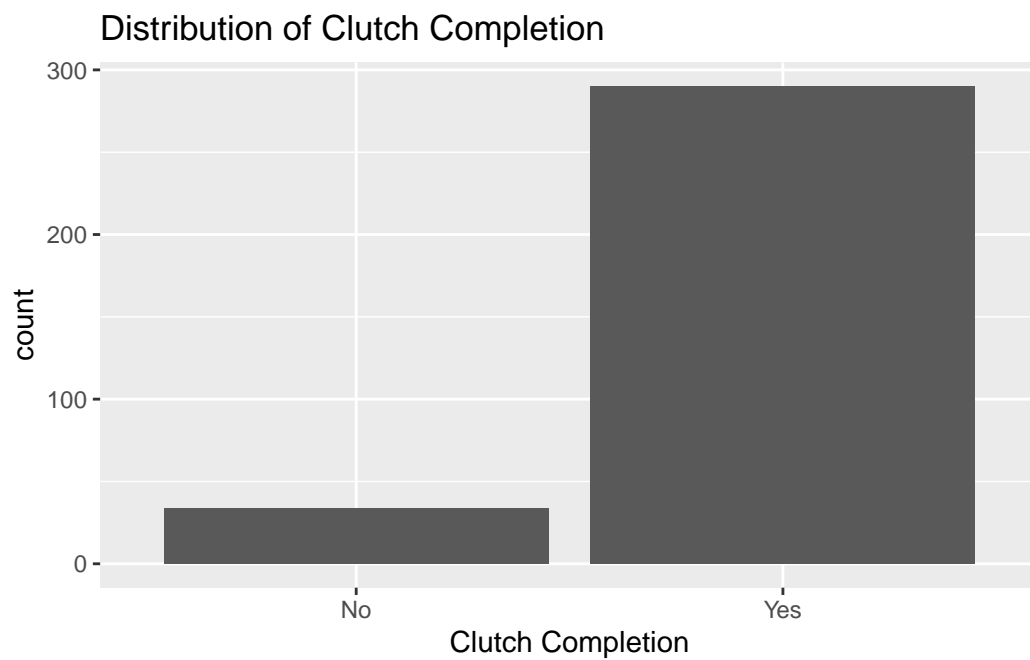



```
# Island
ggplot(penguins_clean, aes(x = Island )) +geom_bar() +ggtitle("Distribution of Island ")
```



```
#Clutch Completion
```

```
ggplot(penguins_clean, aes(x = `Clutch Completion`)) +geom_bar() +ggtitle("Distribution of C
```



```
#Sex
```

```
ggplot(penguins_clean, aes(x = Sex)) +geom_bar() +ggtitle("Distribution of Sex")
```

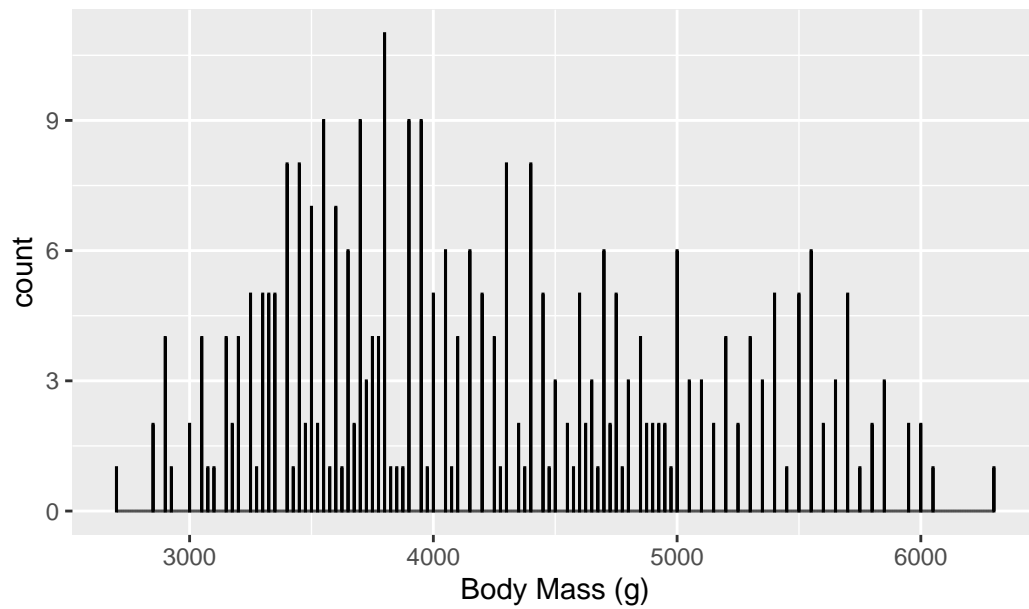


#From the plot,we can know that this data contains 3 species,3 islands, and the proportion of

Histogram

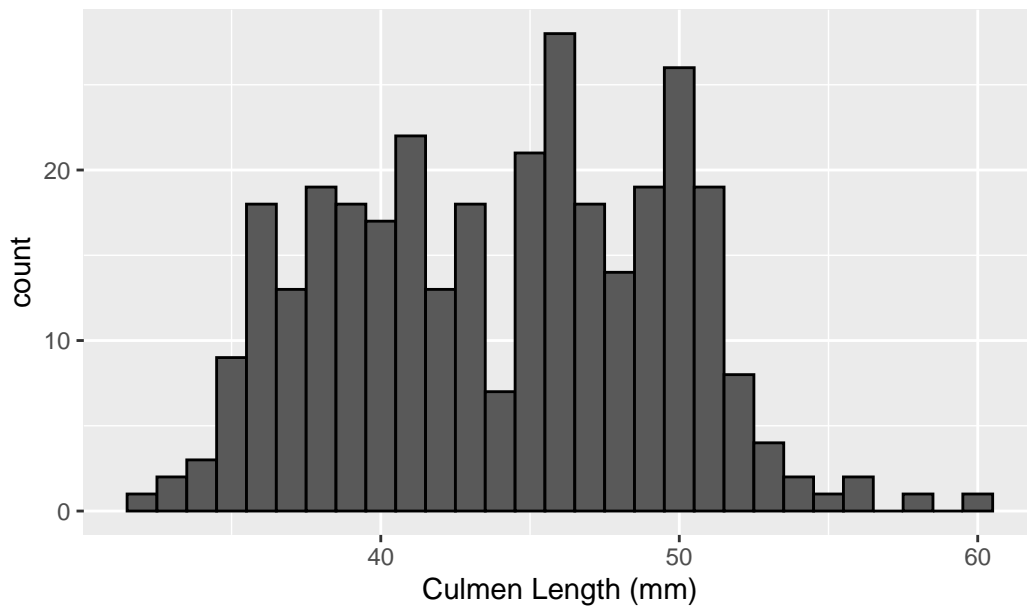
```
# Histogram
#Body Mass
ggplot(penguins_clean, aes(x = `Body Mass (g)`) +
  geom_histogram(binwidth = 1, color = "black") +
  ggtitle("Distribution of Body Mass"))
```

Distribution of Body Mass

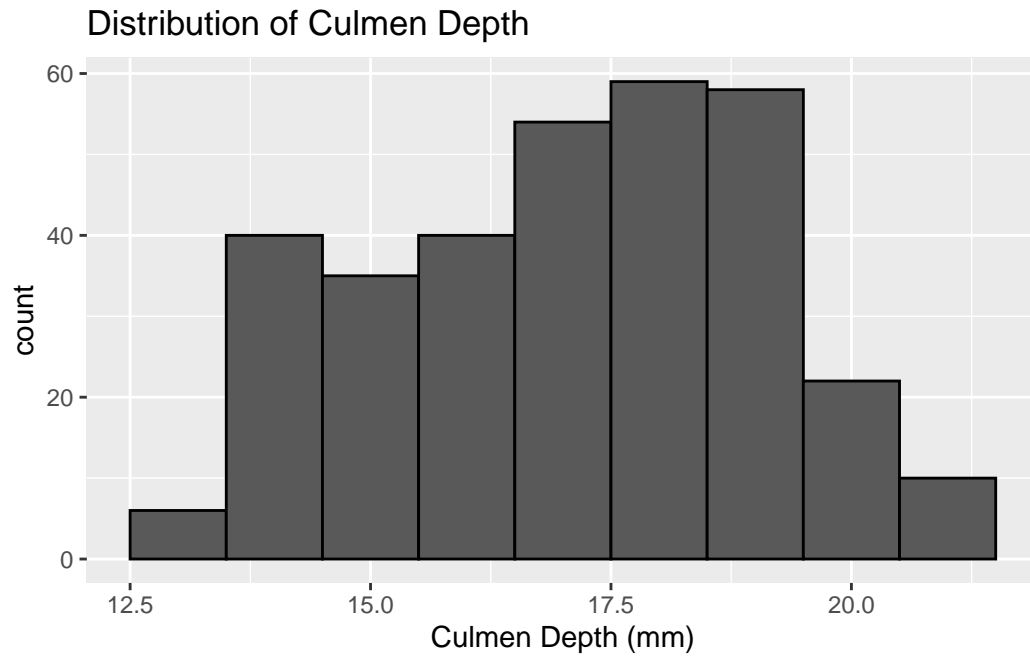


```
#Culmen Length
ggplot(penguins_clean, aes(x = `Culmen Length (mm)`)) +
  geom_histogram(binwidth = 1, color = "black") +
  ggtitle("Distribution of Culmen Length")
```

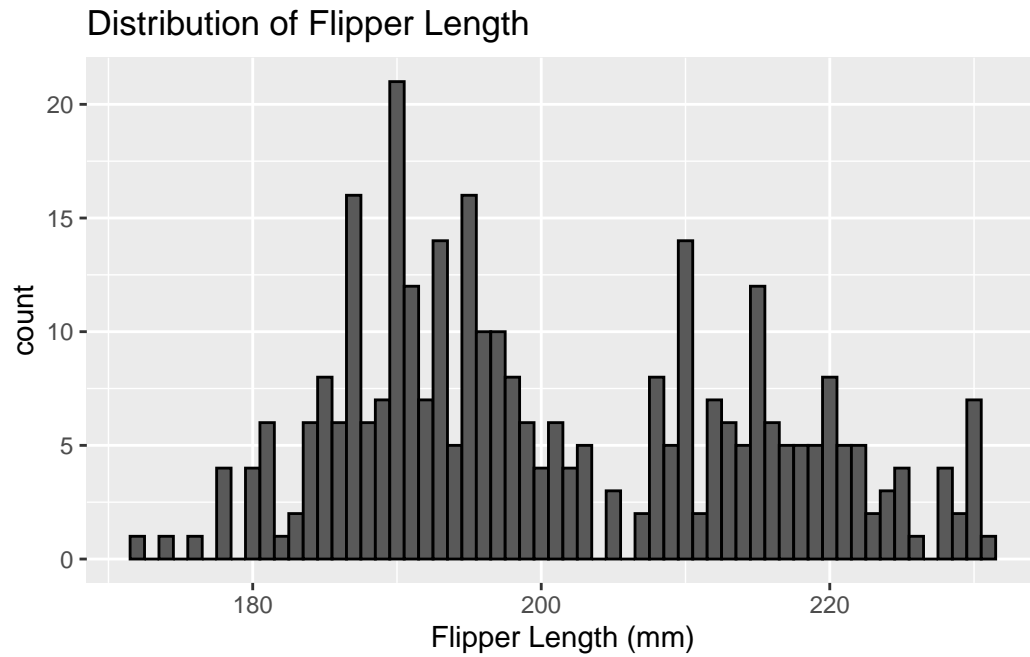
Distribution of Culmen Length



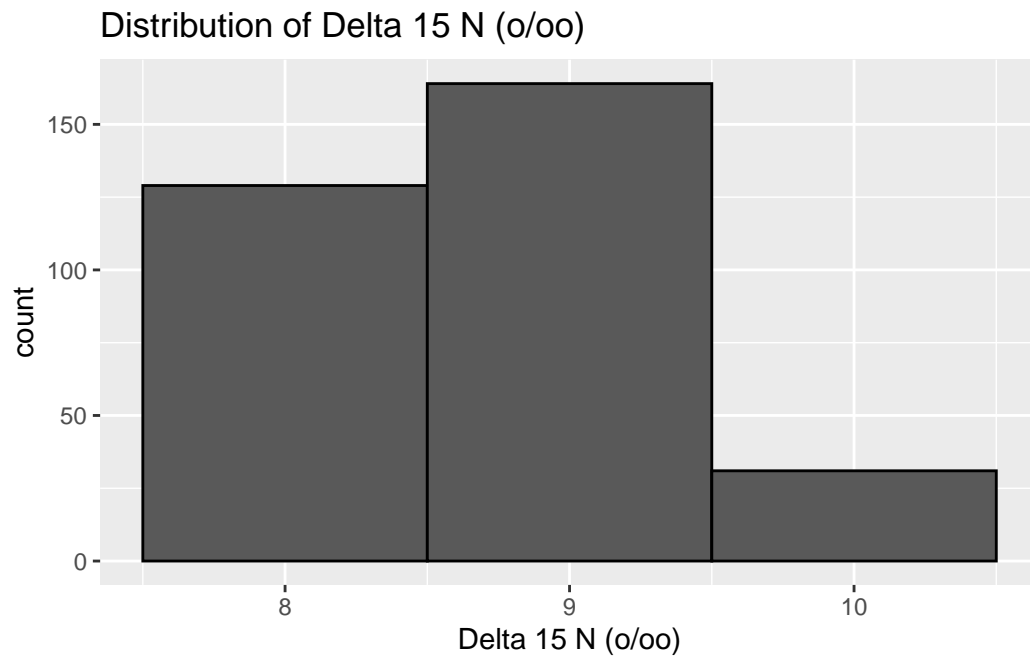
```
#Culmen Depth
ggplot(penguins_clean, aes(x = `Culmen Depth (mm)`)) +
  geom_histogram(binwidth = 1, color = "black") +
  ggtitle("Distribution of Culmen Depth")
```



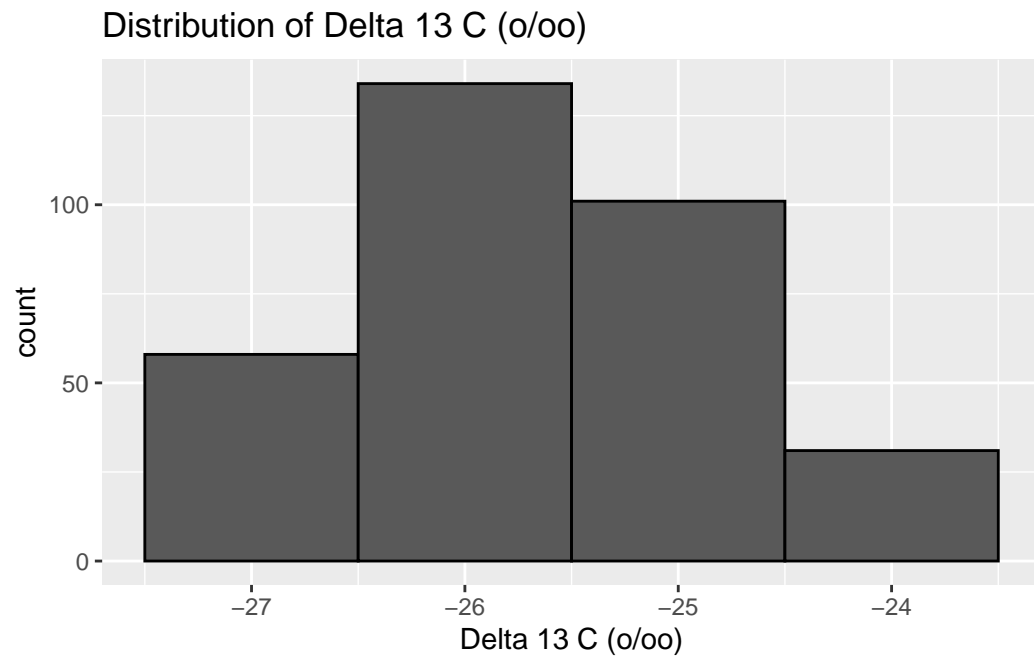
```
#Flipper Length
ggplot(penguins_clean, aes(x = `Flipper Length (mm)`)) +
  geom_histogram(binwidth = 1, color = "black") +
  ggtitle("Distribution of Flipper Length")
```



```
#Delta 15 N
ggplot(penguins_clean, aes(x = `Delta 15 N (o/oo)`)) +
  geom_histogram(binwidth = 1, color = "black") +
  ggtitle("Distribution of Delta 15 N (o/oo)")
```

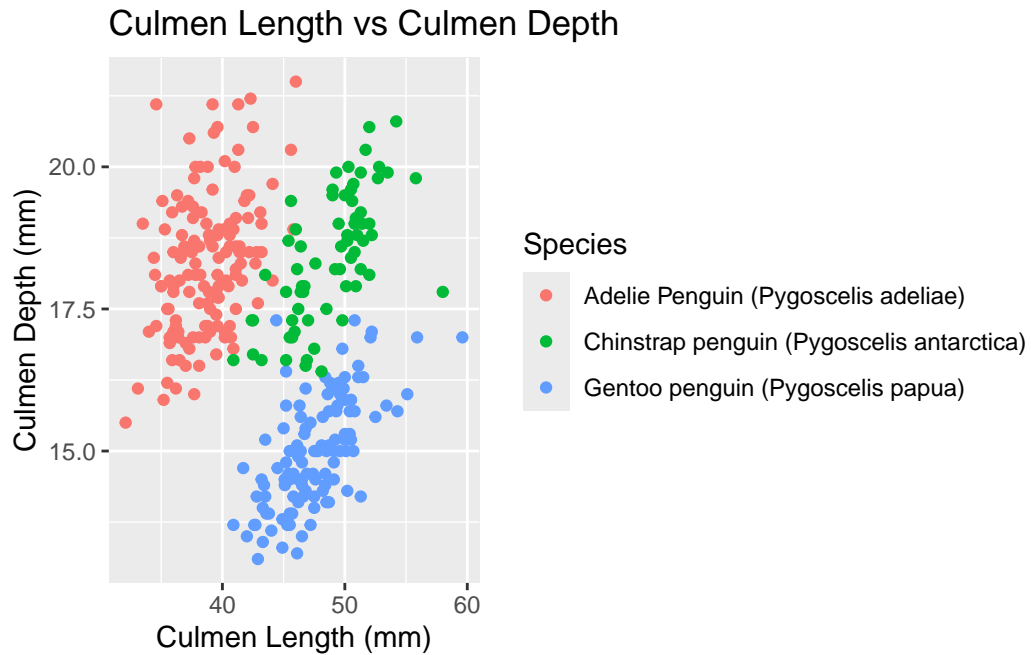


```
#Delta 13  
ggplot(penguins_clean, aes(x = `Delta 13 C (o/oo)`)) +  
  geom_histogram(binwidth = 1, color = "black") +  
  ggtitle("Distribution of Delta 13 C (o/oo)")
```

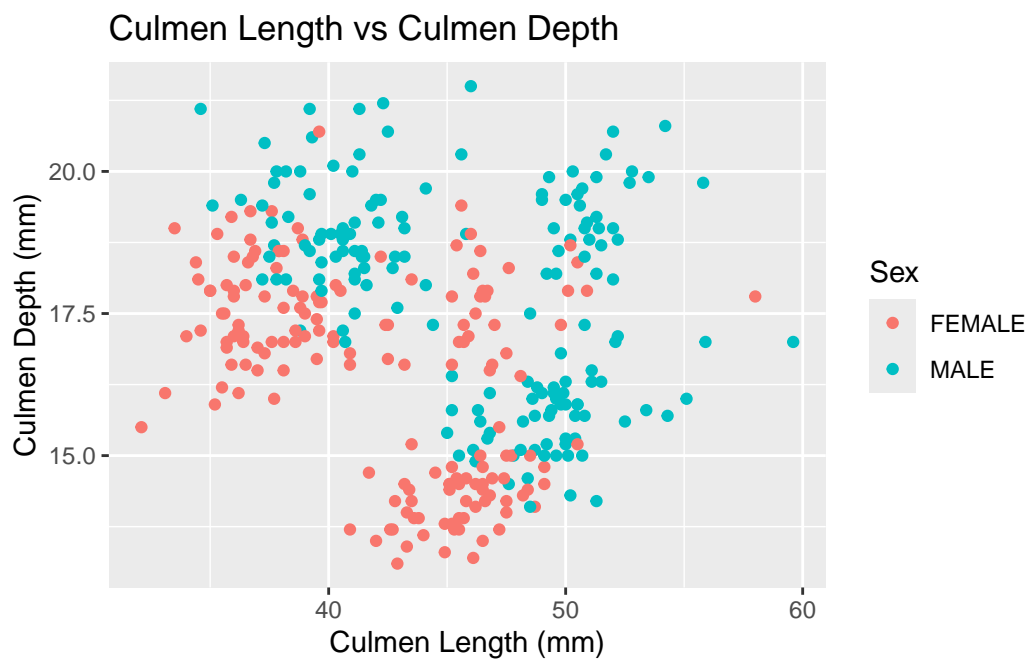
#Scatter plot

```
# Scatter plot of Culmen length vs Culmen depth
#Species
ggplot(penguins_clean, aes(x = `Culmen Length (mm)`, y = `Culmen Depth (mm)`, color = Species
```

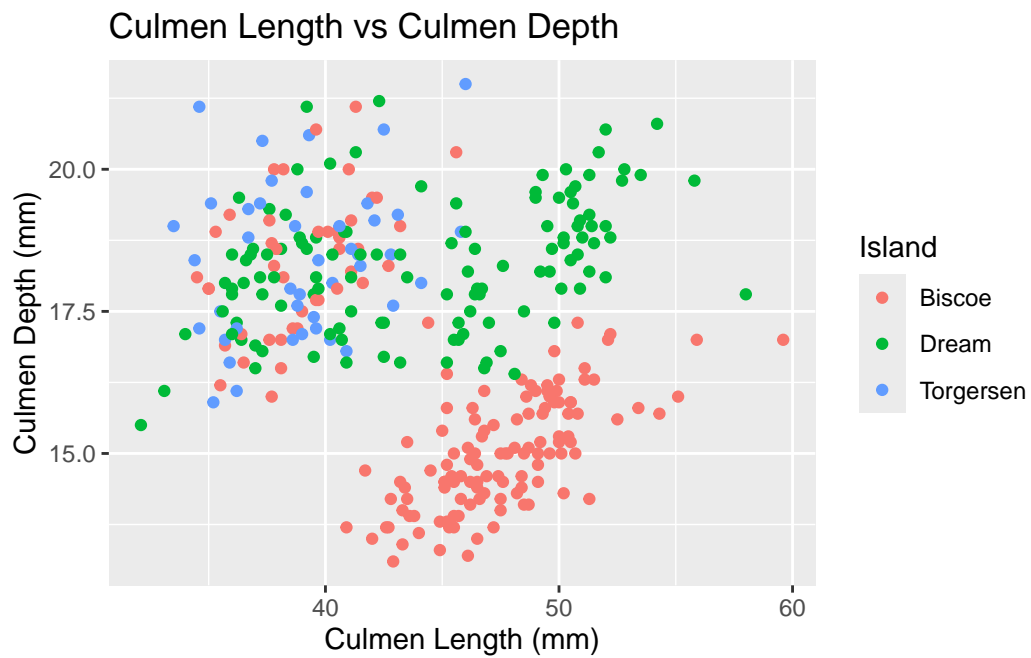


```
#Sex
```

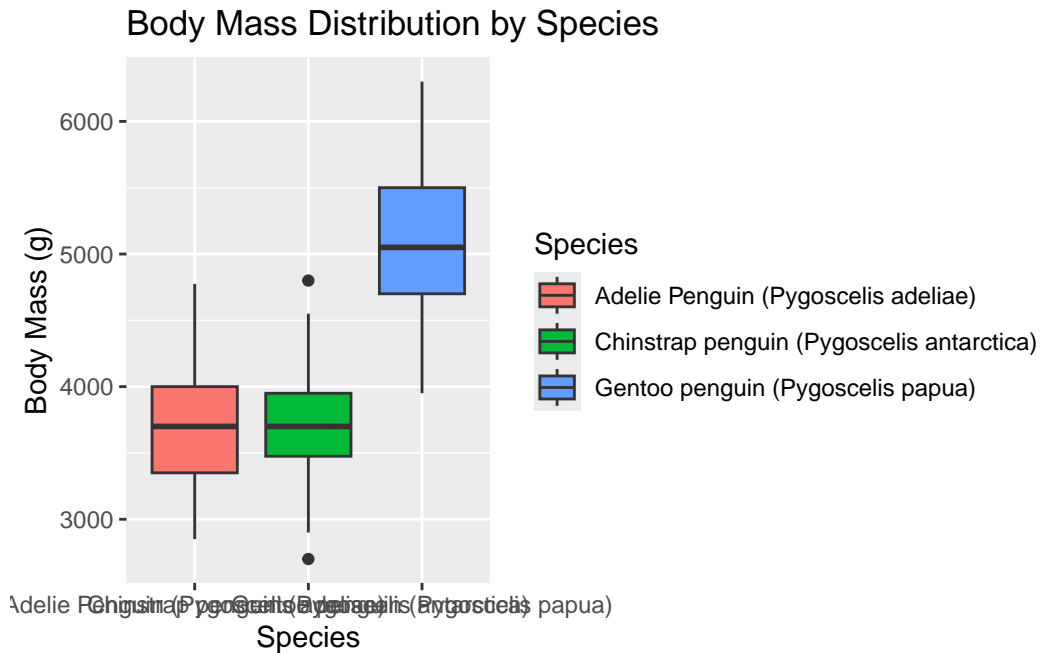
```
ggplot(penguins_clean, aes(x = `Culmen Length (mm)`, y = `Culmen Depth (mm)`, color = Sex))
```



```
#Island
ggplot(penguins_clean, aes(x = `Culmen Length (mm)`, y = `Culmen Depth (mm)`, color = Island))
```



```
#Species and Body Mass
ggplot(penguins_clean, aes(x = Species, y = `Body Mass (g)`, fill = Species)) + geom_boxplot()
```



```
#Gentoo penguin has higher body mass than the others.
```

correlation test

```
library(dplyr)
#correlation
correlation <- cor(penguins_clean$`Culmen Length (mm)`, penguins_clean$`Culmen Depth (mm)`)
correlation
```

```
[1] -0.2229228
```

```
#correlation test
cor_test_result <- cor.test(penguins_clean$`Culmen Length (mm)`, penguins_clean$`Culmen Depth (mm)`)
cor_test_result
```

Pearson's product-moment correlation

```
data: penguins_clean$`Culmen Length (mm)` and penguins_clean$`Culmen Depth (mm)`  
t = -4.1035, df = 322, p-value = 5.162e-05  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.3240128 -0.1167996  
sample estimates:  
cor  
-0.2229228
```

chi_square_test

```
#chi_square_test(Species/Island)  
species_island_table <- table(penguins_clean$Species,penguins_clean$Island)  
chi_test1<- chisq.test(species_island_table)  
chi_test1
```

Pearson's Chi-squared test

```
data: species_island_table  
X-squared = 275.47, df = 4, p-value < 2.2e-16
```

```
# Species and Island is correlated.  
  
#chi_square_test(Sex/Island)  
sex_island_table <- table(penguins_clean$Sex,penguins_clean$Island)  
chi_test2<- chisq.test(sex_island_table)  
chi_test2
```

Pearson's Chi-squared test

```
data: sex_island_table  
X-squared = 0.11124, df = 2, p-value = 0.9459
```

```
# Sex and Island is uncorrelated.
```