Analyzing Palmer Penguins Dataset

Uma Negi

2022-04-24

1. Startng with installing the required packages for cleaning, analysis and visualization

```
library(tidyverse)      #for viewing data, visualizing data
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)           #referencing file easier
```

```
## here() starts at C:/Users/negiu/OneDrive/Documents/PROJECT/Analyzing-Palmer-Penguins-Dataset-using-R
```

```
library(skimr)          #data cleaning task (summarize and skim)
library(janitor)        #data cleaning (filter, sort)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

2. Installing Plamer Penguins dataset and have an idea of the dataset, the datatype and other things

```
library("palmerpenguins")
skim_without_charts(penguins) #brief summary
```

Table 1: Data summary

Name	penguins
Number of rows	344
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5
flipper_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0

```
glimpse(penguins)                 #summary of data set along with some starting values
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
## $ sex               <fct> male, female, female, NA, female, male, female, male~
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

```
head(penguins)                    #shows first 10 rows of data set
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3              193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6              190        3650 male
## # ... with 1 more variable: year <int>
```

```
colnames(penguins)            #shows the column names
```

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

3.DATA MANIPULATION USING SELECT

```
penguins %>%
  select(species, island)     #select column species & island
```

Use SELECT statement to select a particular column or exclude a column (CREATING SUBSET)

```
## # A tibble: 344 x 2
##    species island
##    <fct>   <fct>
##  1 Adelie  Torgersen
##  2 Adelie  Torgersen
##  3 Adelie  Torgersen
##  4 Adelie  Torgersen
##  5 Adelie  Torgersen
##  6 Adelie  Torgersen
##  7 Adelie  Torgersen
##  8 Adelie  Torgersen
##  9 Adelie  Torgersen
## 10 Adelie  Torgersen
## # ... with 334 more rows
```

```
penguins %>%
  select(-species,-island)    #select all column except species & island
```

```
## # A tibble: 344 x 6
##    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex     year
##             <dbl>         <dbl>             <int>       <int> <fct>  <int>
##  1           39.1          18.7               181        3750 male    2007
##  2           39.5          17.4               186        3800 female  2007
##  3           40.3          18                 195        3250 female  2007
##  4           NA            NA                  NA          NA <NA>    2007
##  5           36.7          19.3               193        3450 female  2007
##  6           39.3          20.6               190        3650 male    2007
##  7           38.9          17.8               181        3625 female  2007
##  8           39.2          19.6               195        4675 male    2007
##  9           34.1          18.1               193        3475 <NA>    2007
## 10           42            20.2               190        4250 <NA>    2007
## # ... with 334 more rows
```

```
new_set <- penguins %>%
  rename(island_new = island)  #rename the column
head(new_set)
```

```
## # A tibble: 6 x 8
##   species island_new bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>               <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen            39.1          18.7               181        3750
## 2 Adelie  Torgersen            39.5          17.4               186        3800
## 3 Adelie  Torgersen            40.3          18                 195        3250
## 4 Adelie  Torgersen            NA            NA                  NA          NA
## 5 Adelie  Torgersen            36.7          19.3               193        3450
## 6 Adelie  Torgersen            39.3          20.6               190        3650
## # ... with 2 more variables: sex <fct>, year <int>
```

```
clean_names(penguins)          #makes col-names unique & consistent: char, no & _
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 195        3250
##  4 Adelie  Torgersen           NA            NA                  NA          NA
##  5 Adelie  Torgersen           36.7          19.3               193        3450
##  6 Adelie  Torgersen           39.3          20.6               190        3650
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # ... with 334 more rows, and 2 more variables: sex <fct>, year <int>
```

4. Organizing Data

```
penguins %>%
  arrange(bill_length_mm)
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Dream               32.1          15.5               188        3050
##  2 Adelie  Dream               33.1          16.1               178        2900
##  3 Adelie  Torgersen           33.5          19                 190        3600
##  4 Adelie  Dream               34            17.1               185        3400
##  5 Adelie  Torgersen           34.1          18.1               193        3475
##  6 Adelie  Torgersen           34.4          18.4               184        3325
##  7 Adelie  Biscoe              34.5          18.1               187        2900
##  8 Adelie  Torgersen           34.6          21.1               198        4400
##  9 Adelie  Torgersen           34.6          17.2               189        3200
## 10 Adelie  Biscoe              35            17.9               190        3450
## # ... with 334 more rows, and 2 more variables: sex <fct>, year <int>
```

```r
penguins %>%
  arrange(-bill_depth_mm)
```

```
## # A tibble: 344 x 8
##    species   island    bill_length_mm bill_depth_mm flipper_length_~ body_mass_g
##    <fct>     <fct>              <dbl>         <dbl>            <int>       <int>
##  1 Adelie    Torgersen           46            21.5              194        4200
##  2 Adelie    Torgersen           38.6          21.2              191        3800
##  3 Adelie    Dream               42.3          21.2              191        4150
##  4 Adelie    Torgersen           34.6          21.1              198        4400
##  5 Adelie    Dream               39.2          21.1              196        4150
##  6 Adelie    Biscoe              41.3          21.1              195        4400
##  7 Chinstrap Dream               54.2          20.8              201        4300
##  8 Adelie    Torgersen           42.5          20.7              197        4500
##  9 Adelie    Biscoe              39.6          20.7              191        3900
## 10 Chinstrap Dream               52            20.7              210        4800
## # ... with 334 more rows, and 2 more variables: sex <fct>, year <int>
```

```r
penguins %>%
  group_by(island) %>% drop_na() %>%
  summarize(mean_bill_length_mm = mean(bill_length_mm),
            mean_bill_depth_mm = mean(bill_depth_mm))
```

```
## # A tibble: 3 x 3
##   island    mean_bill_length_mm mean_bill_depth_mm
##   <fct>                   <dbl>              <dbl>
## 1 Biscoe                   45.2               15.9
## 2 Dream                    44.2               18.3
## 3 Torgersen                39.0               18.5
```

```r
penguins %>%
  group_by(island) %>% drop_na() %>%
  summarize(max_bill_length_mm = max(bill_length_mm),
            min_bill_length_mm = min(bill_length_mm),
            max_bill_depth_mm = max(bill_depth_mm),
            min_bill_depth_mm = min(bill_depth_mm))
```

```
## # A tibble: 3 x 5
##   island    max_bill_length_~ min_bill_length~ max_bill_depth_~ min_bill_depth_~
##   <fct>                 <dbl>            <dbl>            <dbl>            <dbl>
## 1 Biscoe                 59.6             34.5             21.1             13.1
## 2 Dream                  58               32.1             21.2             15.5
## 3 Torgersen              46               33.5             21.5             15.9
```

```r
penguins %>%
  group_by(island, species) %>% drop_na() %>%
  summarize(max_bl = max(bill_length_mm),
            men_bl = mean(bill_length_mm))
```

```
## ‘summarise()‘ has grouped output by ’island’. You can override using the
## ‘.groups‘ argument.
## # A tibble: 5 x 4
## # Groups:   island [3]
##   island    species   max_bl men_bl
##   <fct>     <fct>      <dbl>  <dbl>
## 1 Biscoe    Adelie      45.6   39.0
## 2 Biscoe    Gentoo      59.6   47.6
## 3 Dream     Adelie      44.1   38.5
## 4 Dream     Chinstrap   58     48.8
## 5 Torgersen Adelie      46     39.0
```

```
penguins %>% filter(species == "Adelie")      #filter data
```

```
## # A tibble: 152 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 195        3250
##  4 Adelie  Torgersen           NA            NA                  NA          NA
##  5 Adelie  Torgersen           36.7          19.3               193        3450
##  6 Adelie  Torgersen           39.3          20.6               190        3650
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # ... with 142 more rows, and 2 more variables: sex <fct>, year <int>
```

```
#arranges data in asc order of beak length (for desc order use - sign before column)
```

5. TRANSFORMING DATA (combine, split, etc)

```
unite(penguins, 'specie_gender', species, sex, sep = "-" )
```

```
## # A tibble: 344 x 7
##    specie_gender island    bill_length_mm bill_depth_mm flipper_length_mm
##    <chr>         <fct>              <dbl>         <dbl>             <int>
##  1 Adelie-male   Torgersen           39.1          18.7               181
##  2 Adelie-female Torgersen           39.5          17.4               186
##  3 Adelie-female Torgersen           40.3          18                 195
##  4 Adelie-NA     Torgersen           NA            NA                  NA
##  5 Adelie-female Torgersen           36.7          19.3               193
##  6 Adelie-male   Torgersen           39.3          20.6               190
##  7 Adelie-female Torgersen           38.9          17.8               181
##  8 Adelie-male   Torgersen           39.2          19.6               195
##  9 Adelie-NA     Torgersen           34.1          18.1               193
## 10 Adelie-NA     Torgersen           42            20.2               190
## # ... with 334 more rows, and 2 more variables: body_mass_g <int>, year <int>
```

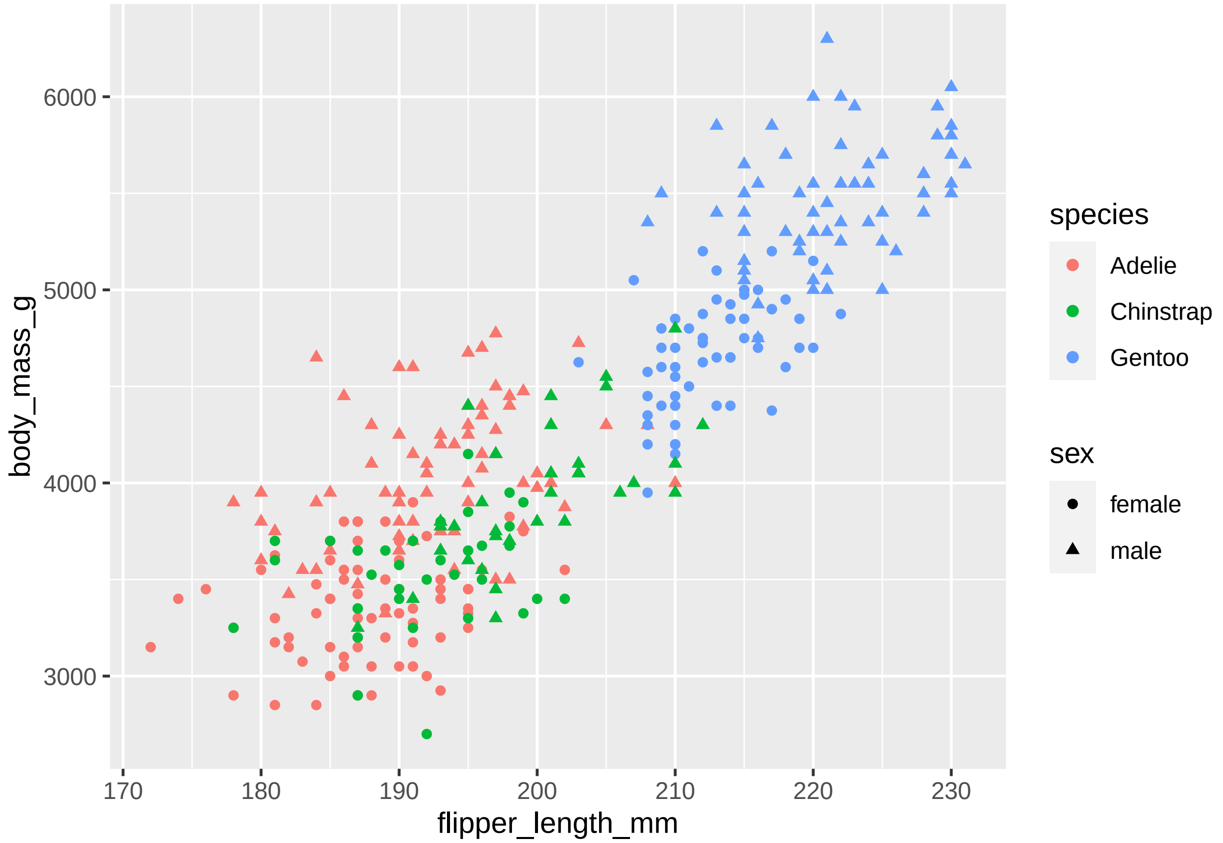
```
new <- penguins %>% mutate(body_mass_kg = body_mass_g/1000,
                           bill_length_m = bill_depth_mm/1000) %>% drop_na()
```

6. Using Visualization

Using ggplot2 to visualize data and share analysis

```
penguins_clean_data  <- penguins[complete.cases(penguins), ] # remove Null values

library(ggplot2)
ggplot(data = penguins_clean_data ) +
  geom_point(mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species,
                           shape = sex))
```



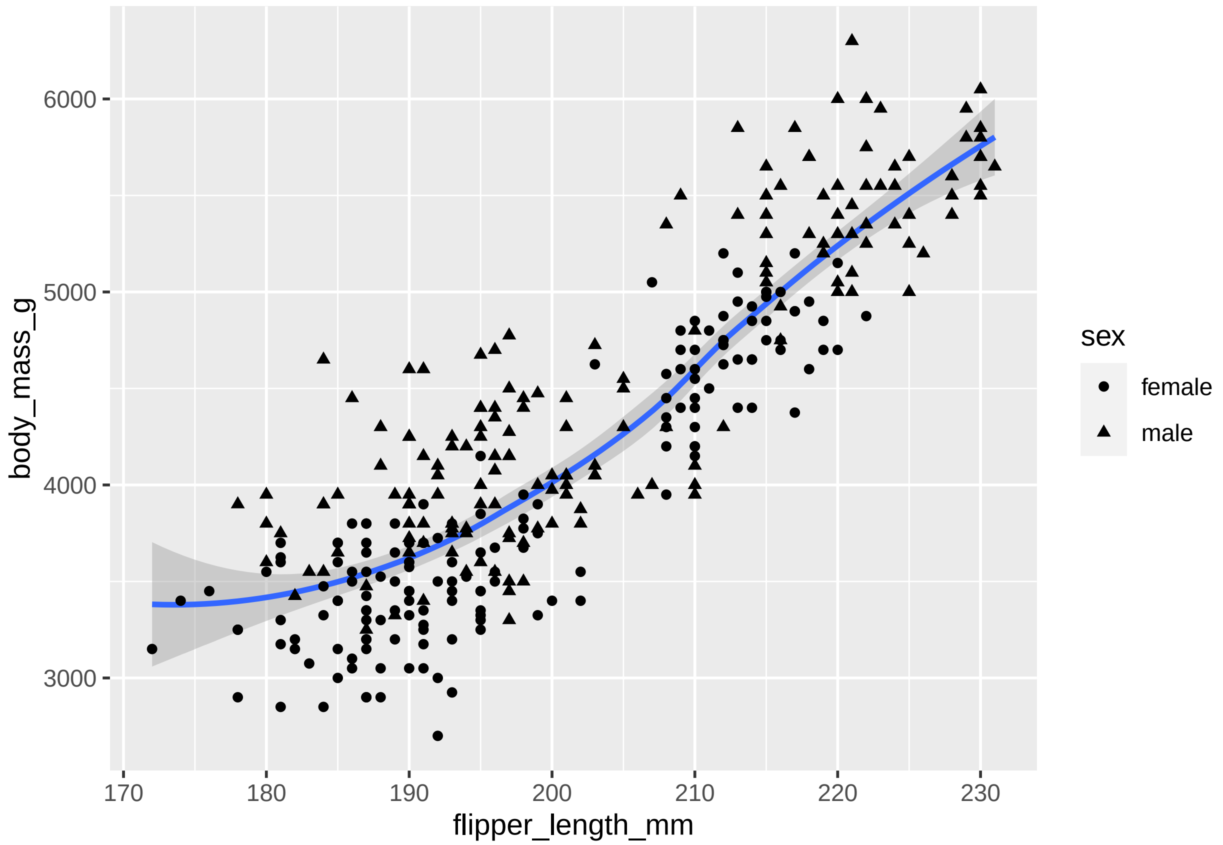
1. This graph shows a positive relation between flipper length and body mass.

2. It also shows that Gentoo species have the highest flipper length to mass ratio

3. It also shows that male penguins have high ratio compare to female ones in each penguin species

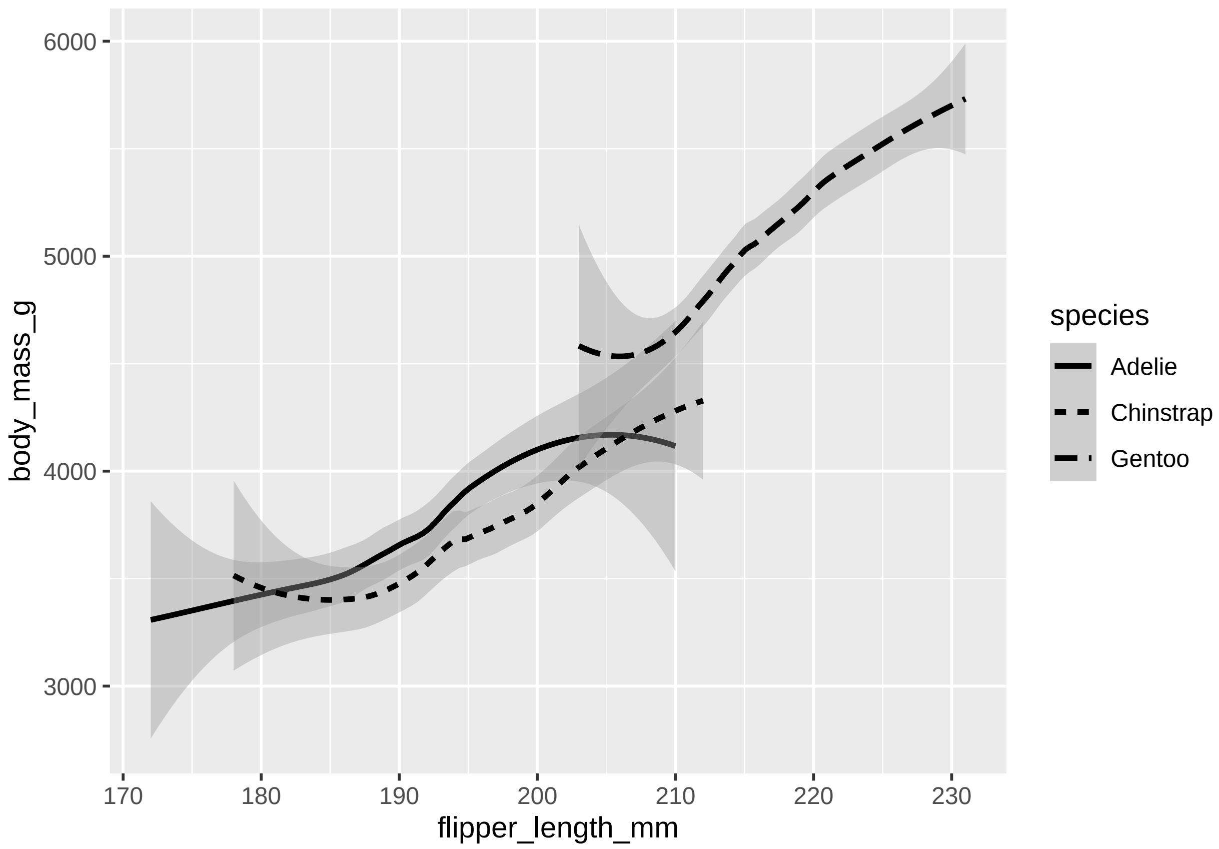
```
ggplot(data = penguins_clean_data ) +
  geom_smooth(mapping = aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point(mapping = aes(x = flipper_length_mm, y = body_mass_g, shape = sex))   #line smooth graph
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



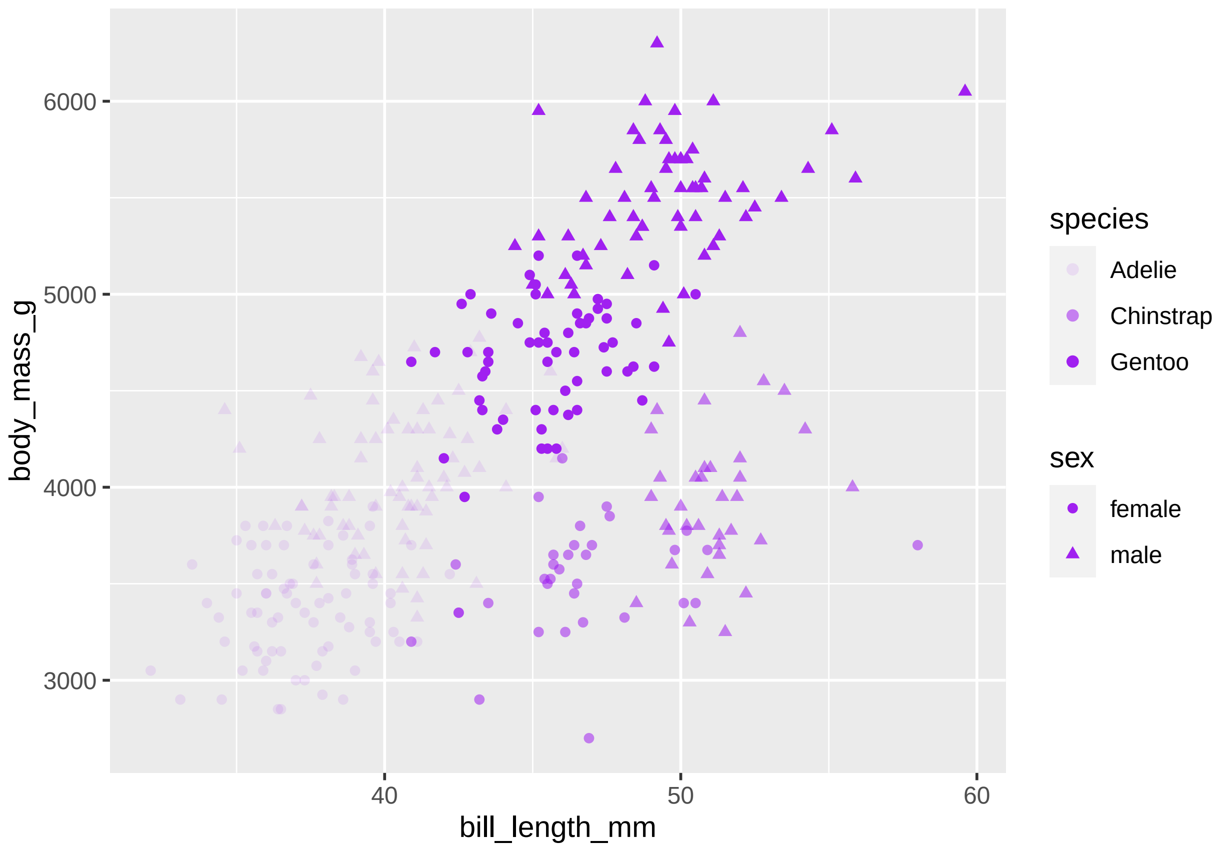
```
ggplot(data = penguins_clean_data ) +
  geom_smooth(mapping = aes(x = flipper_length_mm, y = body_mass_g,
                            linetype = species), color = "black")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

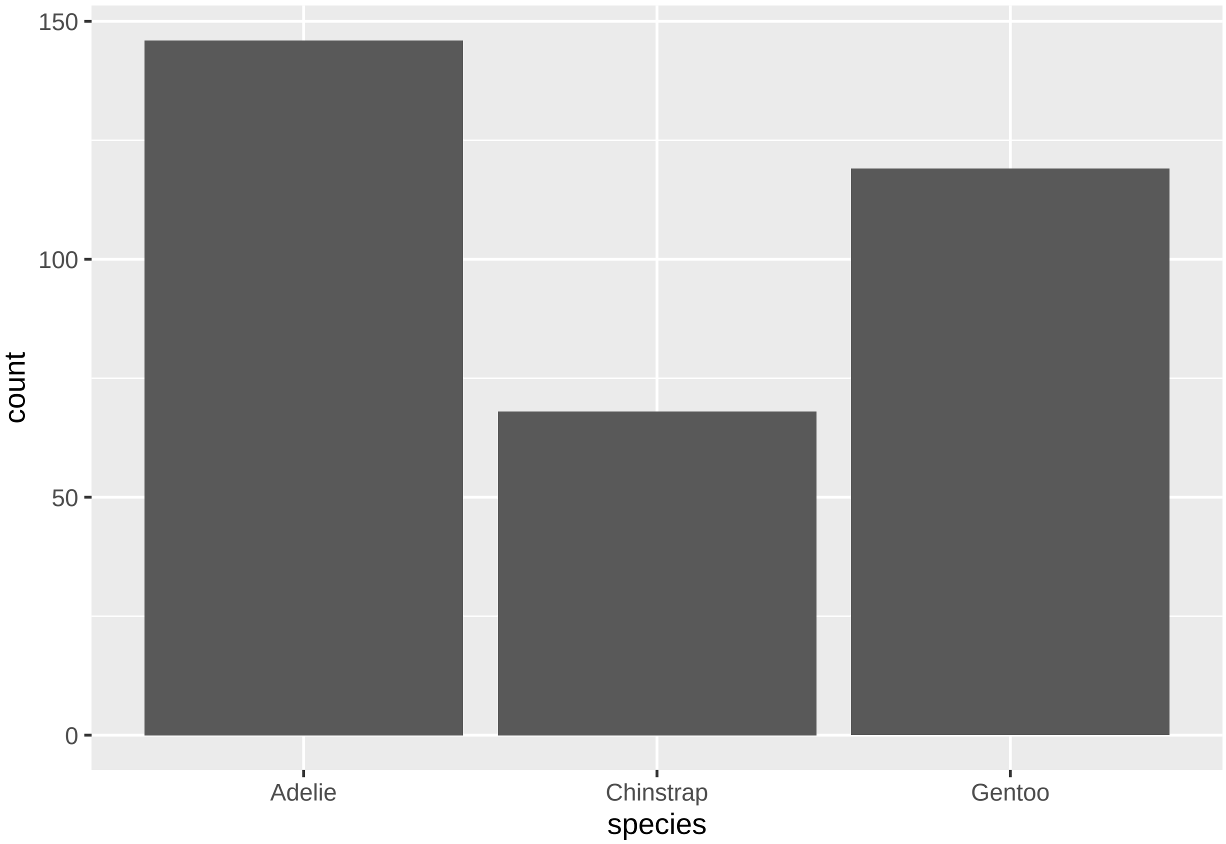


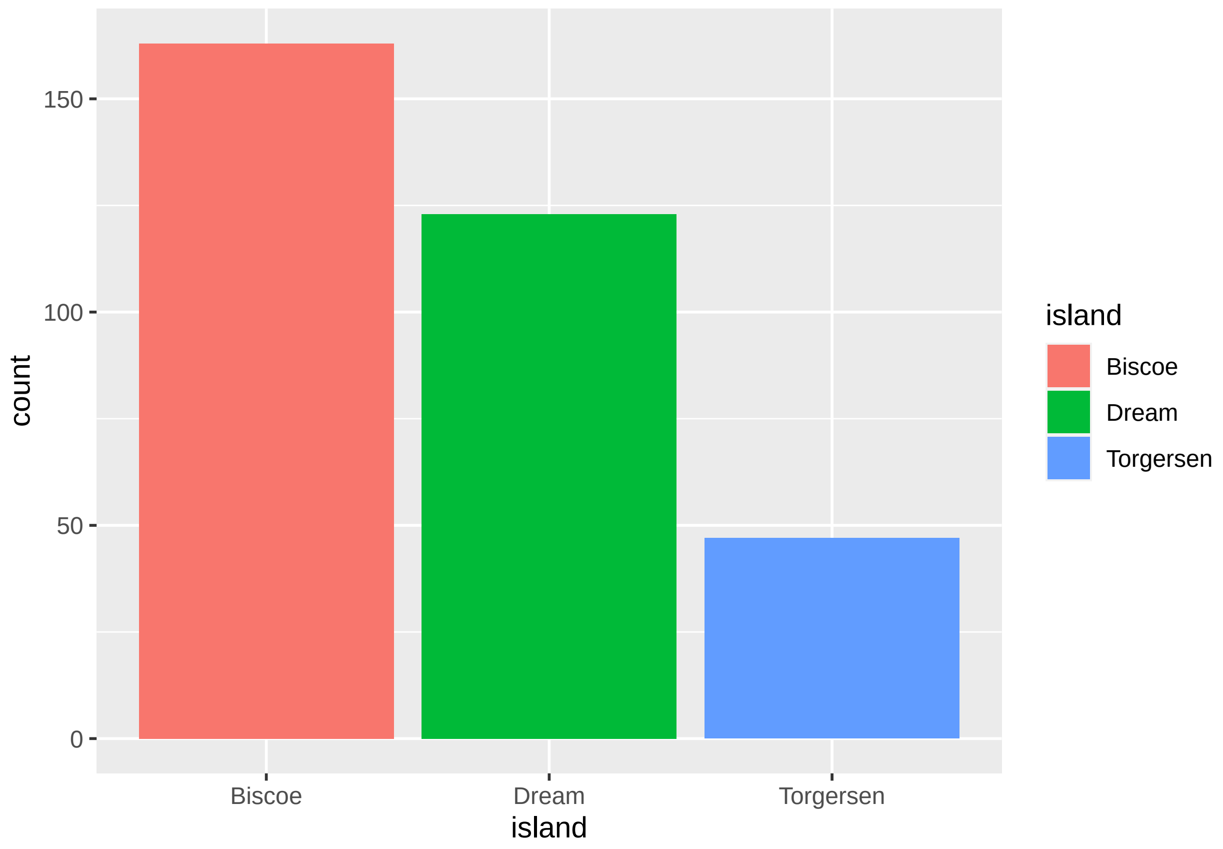
1. **This graph shows a clear relationship between the 3 penguin species**
2. **Gentoo having the highest followed by Adelie and then Chinstrap**

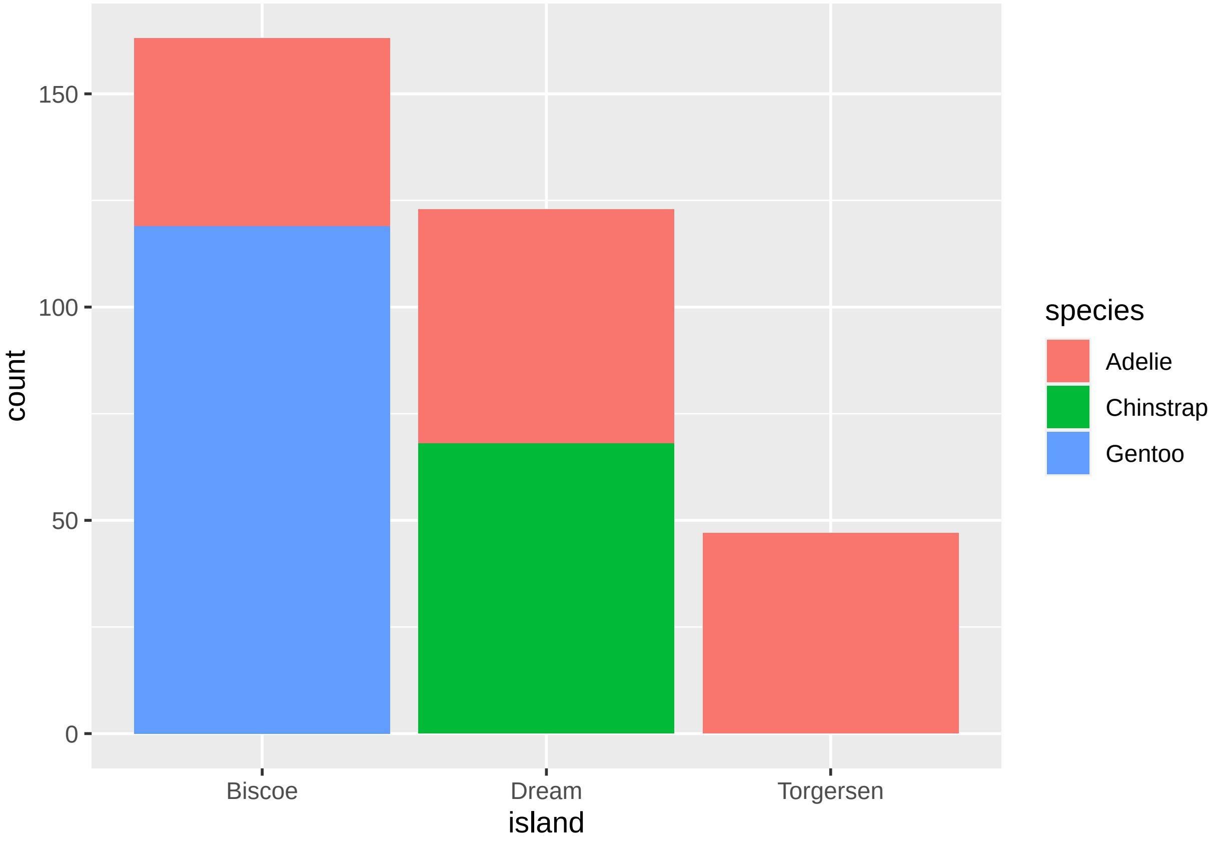
```
## Warning: Using alpha for a discrete variable is not advised.
```

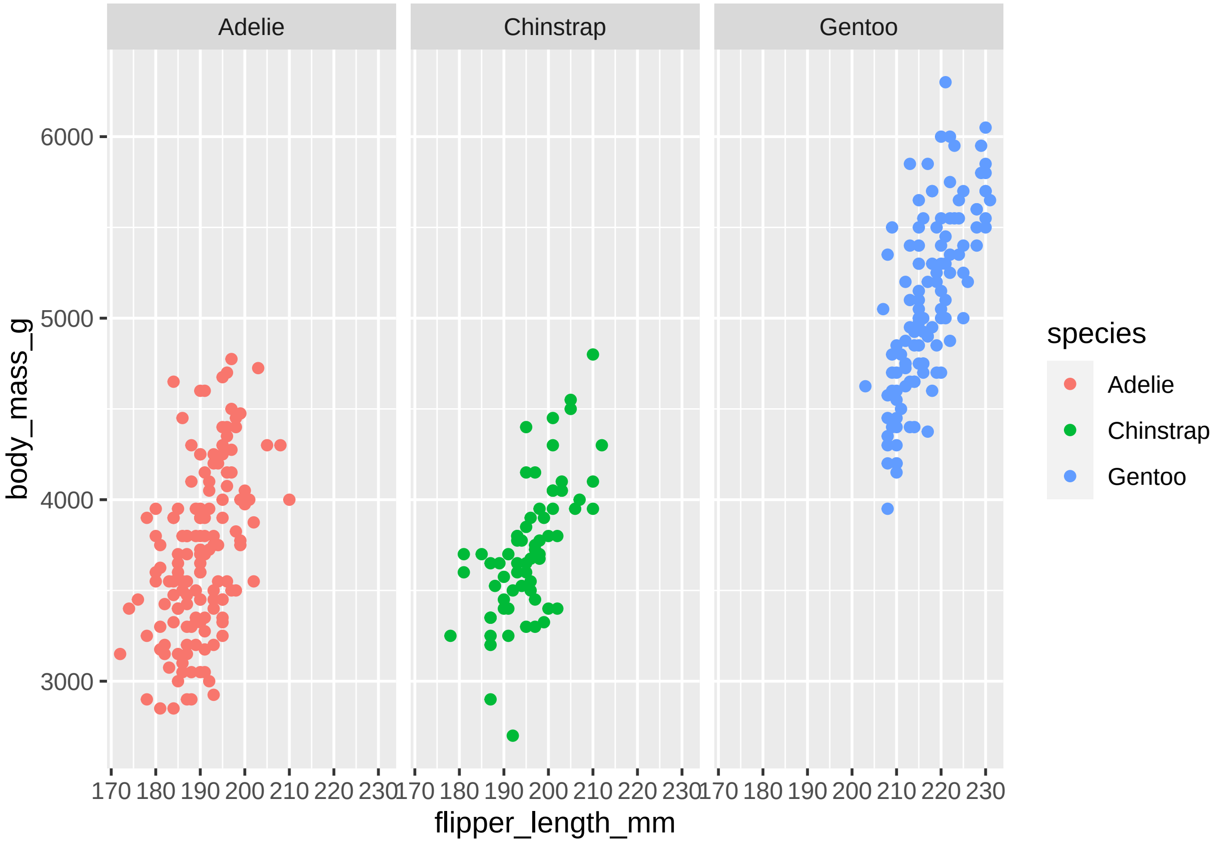


6000
5000
4000
3000
body_mass_g
40
50
60
bill_length_mm
species
Adelie
Chinstrap
Gentoo
sex
female
male


150
100
50
0
count
Adelie
Chinstrap
Gentoo
species


count
150
100
50
0
Biscoe
Dream
Torgersen
island
island
🟥 Biscoe
🟩 Dream
🟦 Torgersen


150
100
50
0
count
Biscoe
Dream
Torgersen
island
species
Adelie
Chinstrap
Gentoo



1. This graph shows us how different species of penguins are divided in different island

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.