Exploratory Data Analysis - Summary Report

1. Data Overview

Table 1: Composition of the sample

```
library(tidyverse)
library(palmerpenguins)
data(penguines)
summary(penguins)
```

```
     species          island    bill_length_mm  bill_depth_mm
 Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
 Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
 Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
                                 Mean   :43.92   Mean   :17.15
                                 3rd Qu.:48.50   3rd Qu.:18.70
                                 Max.   :59.60   Max.   :21.50
                                 NA's   :2       NA's   :2
 flipper_length_mm  body_mass_g       sex
 Min.   :172.0     Min.   :2700   female:165
 1st Qu.:190.0     1st Qu.:3550   male  :168
 Median :197.0     Median :4050   NA's  : 11
 Mean   :200.9     Mean   :4202
 3rd Qu.:213.0     3rd Qu.:4750
 Max.   :231.0     Max.   :6300
 NA's   :2         NA's   :2
```

According to the Table 1 the majority of penguins are Adelie and the majority of penguins are caught from Biscoe isaland. Except, species and island there are 2 missing values in other variables.

2. Composition of the sample

Sometimes table does not give the counts for the levels of the categorical variable. Then you can take bar charts to view individual counts. Here, is the command.

```
ggplot(penguins) + geom_bar(aes(species))
```

You can do the same for other qualitative variables if necessary.

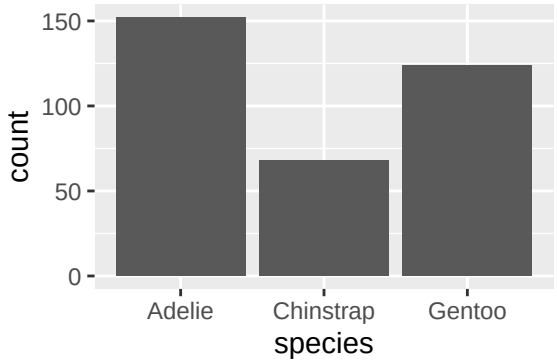


Figure 1: Composition of sample by Species

Important note:

Sometimes, you will see row names as a variable. Then you will get multiple bars like this (Figure 2). Here is an example. See Figure 2. Do not include those types of charts to the report.

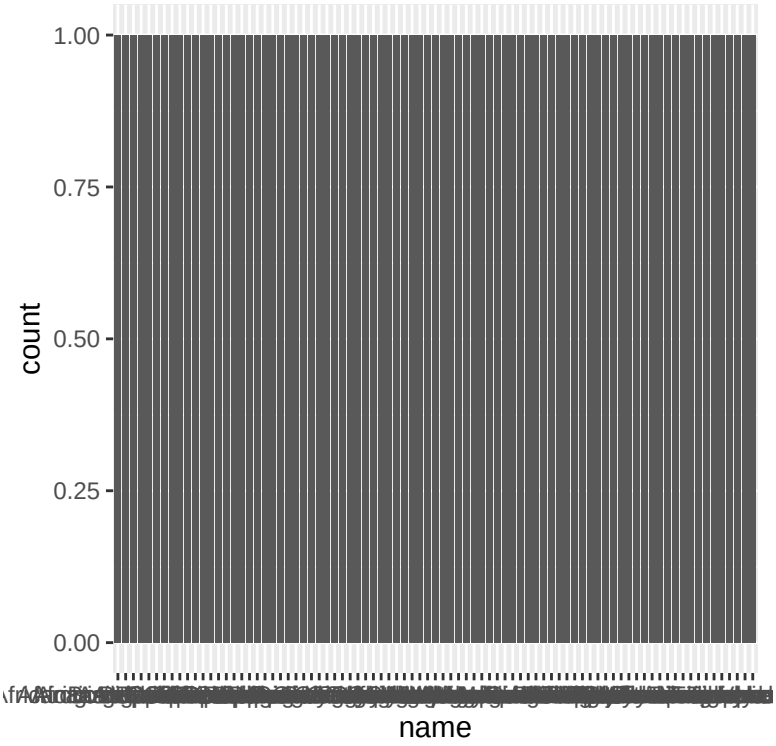


Figure 2: Do not include these types ofgraphs

Composition of the sample by sex and island

```
tablecount <- penguins %>%
  drop_na() %>%
  count(sex, species)
tablecount
```

```
# A tibble: 6 x 3
  sex    species       n
  <fct>  <fct>     <int>
1 female Adelie       73
2 female Chinstrap    34
3 female Gentoo       58
4 male   Adelie       73
5 male   Chinstrap    34
6 male   Gentoo       61
```

```
  ggplot(tablecount) + geom_col(aes(x = species, y = n, fill = species)) +
  geom_label(aes(x = species, y = n, label = n)) +
  scale_fill_manual(values = c("darkorange","purple","cyan4")) +
  facet_wrap(~sex) +
  labs(title = 'Penguins Species ~ Gender')
```

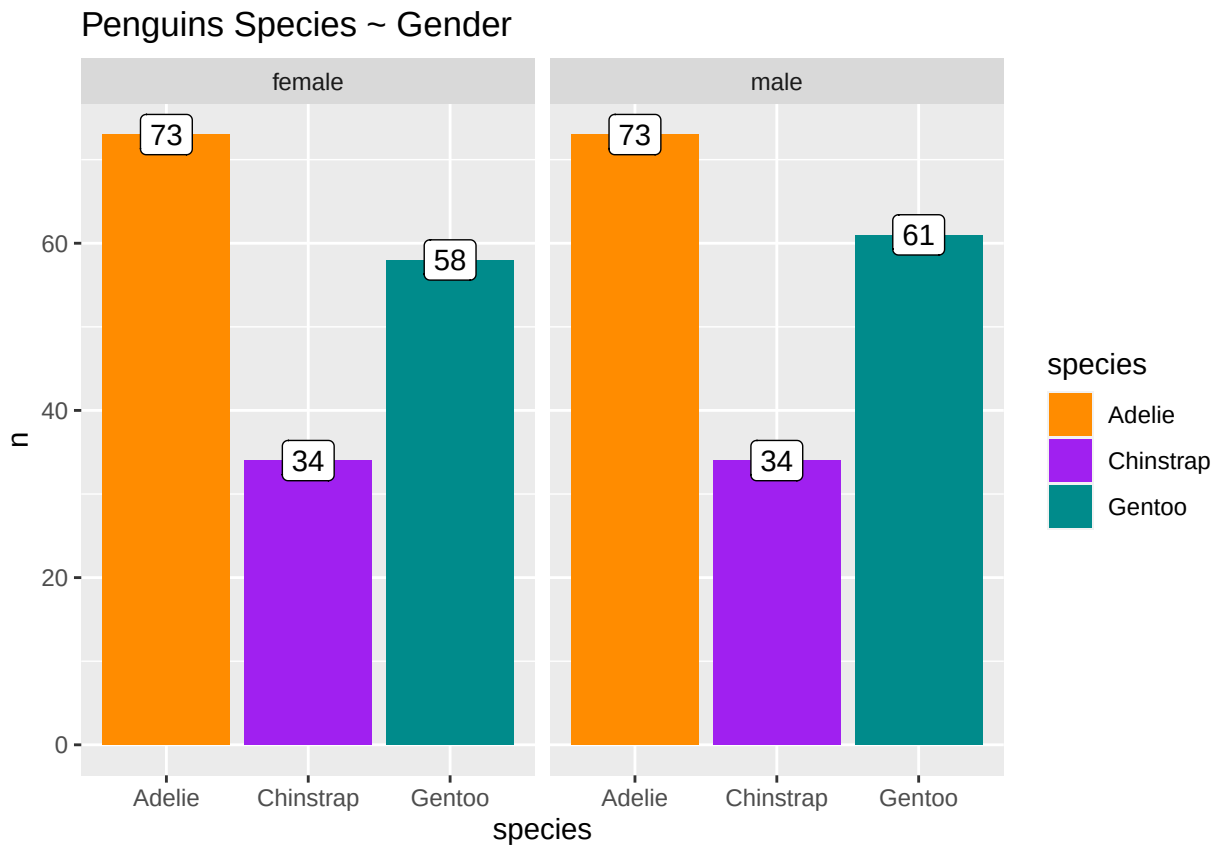


Figure 3: Composition of sample by species and gender

According to Figure 3, the distribution of species type within male and female groups are approximately same.

3. Distribution of body characteristics variables by qualitative variables

```
penguins2 <- penguins %>% drop_na()
ggplot(data = penguins2,
       aes(y = flipper_length_mm,
           x = sex,
           fill=sex)) +
  geom_boxplot()
```

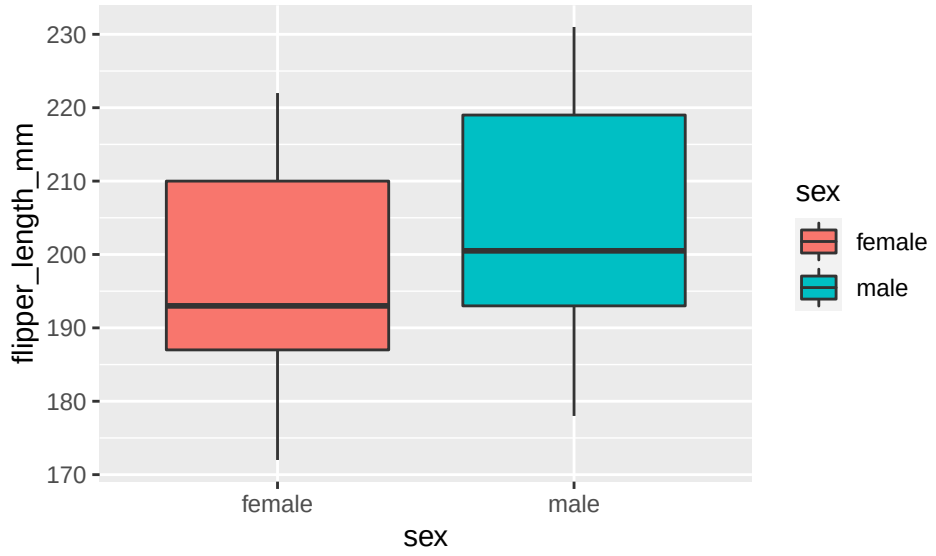


Figure 4: Distribution of flipper length by gender

According to figure 4 the flipper length of male penguins are higher than female penguins.

```
penguins2 <- penguins %>% drop_na()
ggplot(data = penguins2,
       aes(y = flipper_length_mm,
           x = island,
           fill=island)) +
  geom_boxplot()
```

According to figure 5 the flipper length of Biscoe is the highest when compared that with Dream and Torgersen penguins.

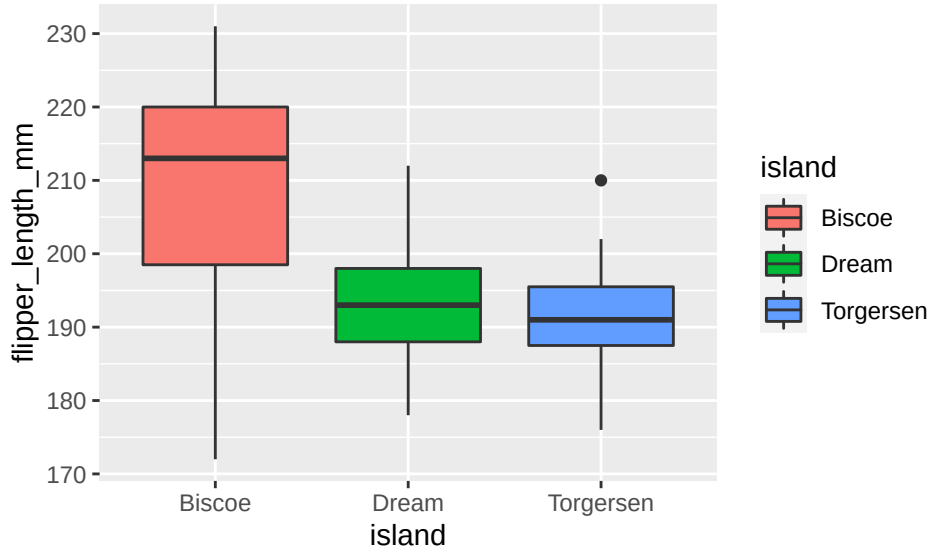


Figure 5: Distribution of flipper length by island

4. Relationship between body-characteristics

```
body1 <- select(penguins, c(bill_length_mm, bill_depth_mm,
                            flipper_length_mm,
                            body_mass_g))
library(GGally)
ggpairs(body1)
```

According to the figure 6, there is a strong positive relationship between body mass and flipper length.

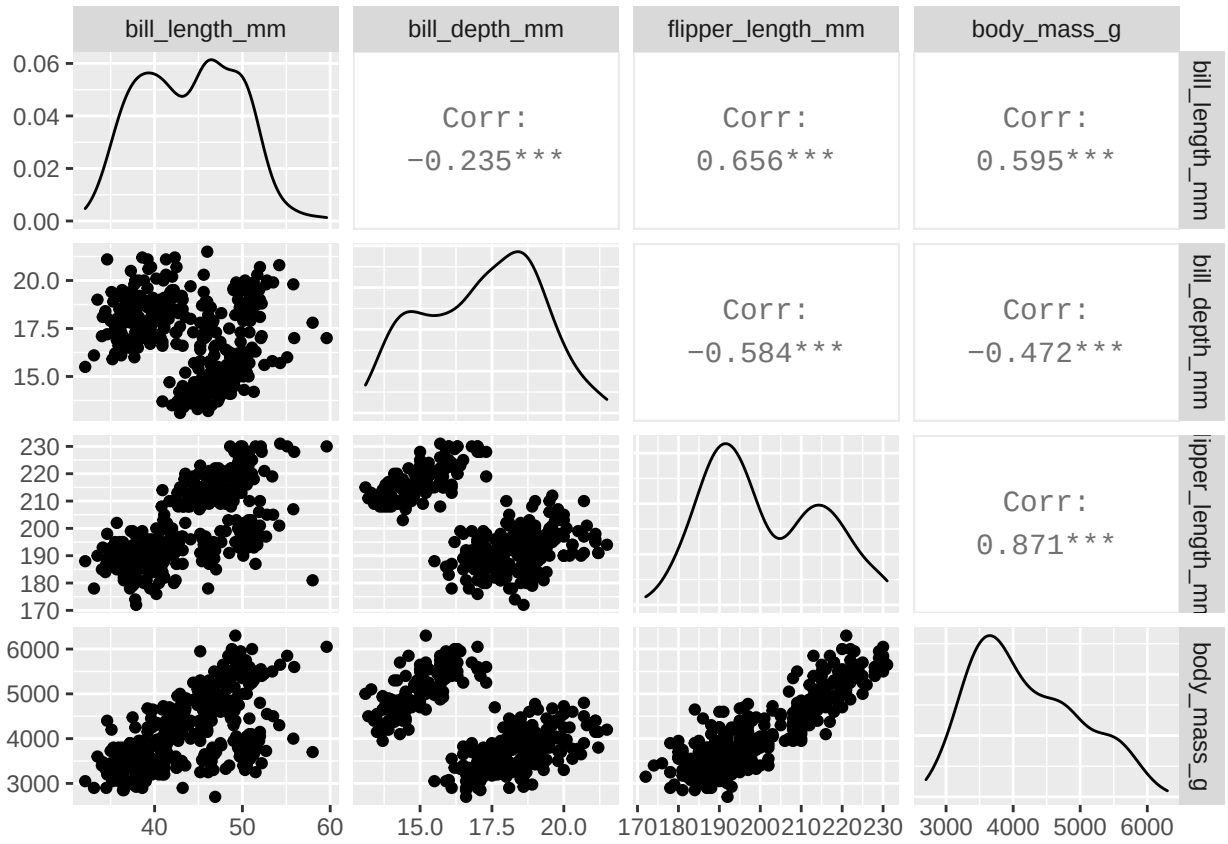


Figure 6: Relationship between body characteristics variables

```
body2 <- select(penguins, c(bill_length_mm, bill_depth_mm,
                            flipper_length_mm,
                            body_mass_g, species))
ggpairs(data=body2, aes(color = species),
          columns = c("flipper_length_mm", "body_mass_g",
                      "bill_length_mm", "bill_depth_mm"))
```

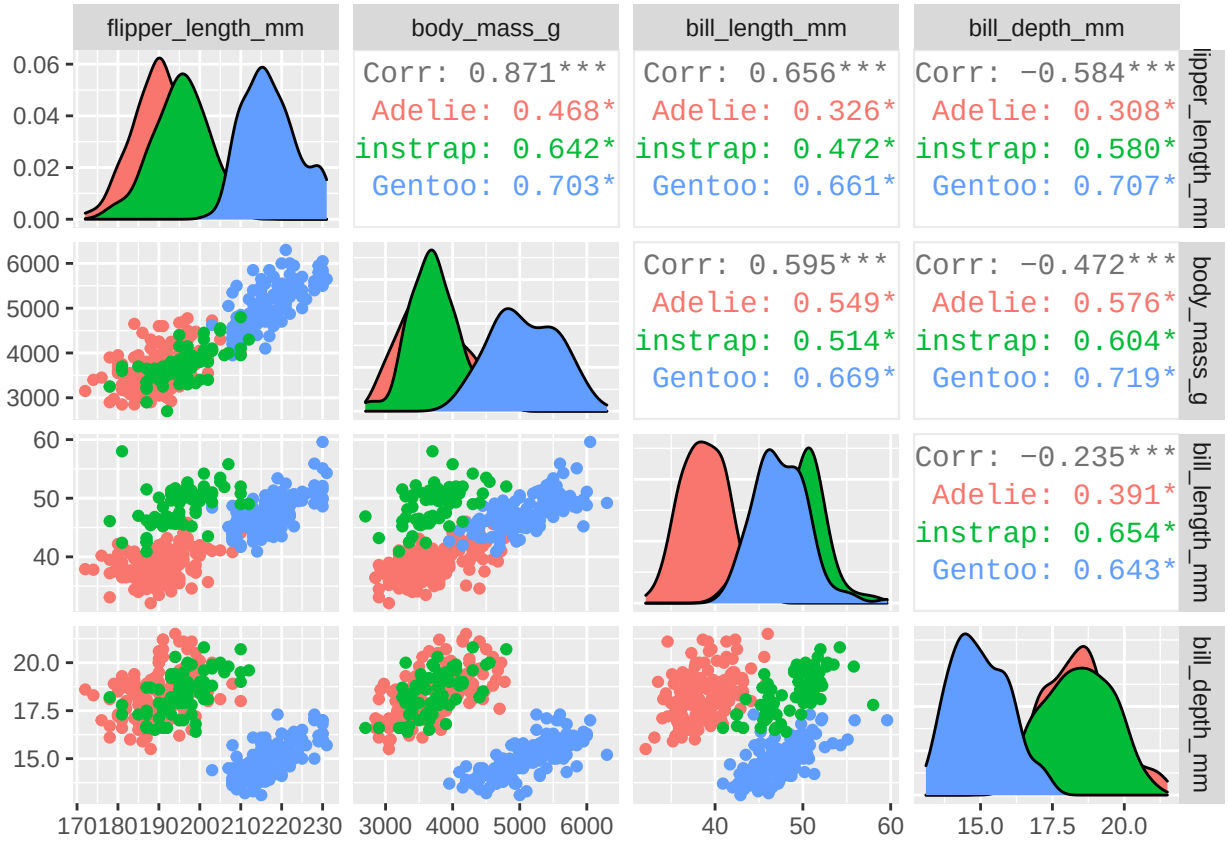


Figure 7: Relationship between body characteristics variables by species

According to the figure 7, the strength of the relationship between bill depth and flipper length vary according to the species type.

```
body3 <- select(penguins, c(bill_length_mm, bill_depth_mm,
                            flipper_length_mm,
                            body_mass_g, sex))
ggpairs(data=body3, aes(color = sex),
          columns = c("flipper_length_mm", "body_mass_g",
                      "bill_length_mm", "bill_depth_mm"))
```

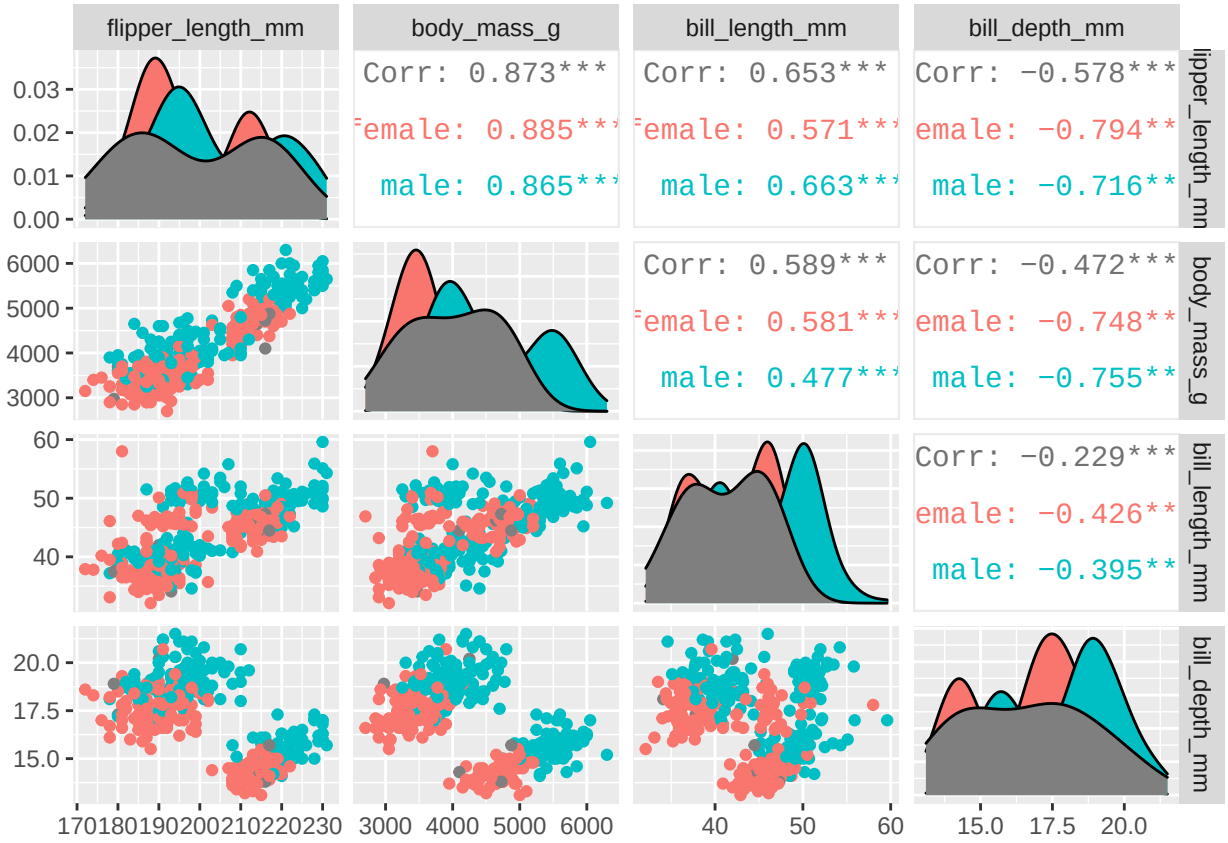


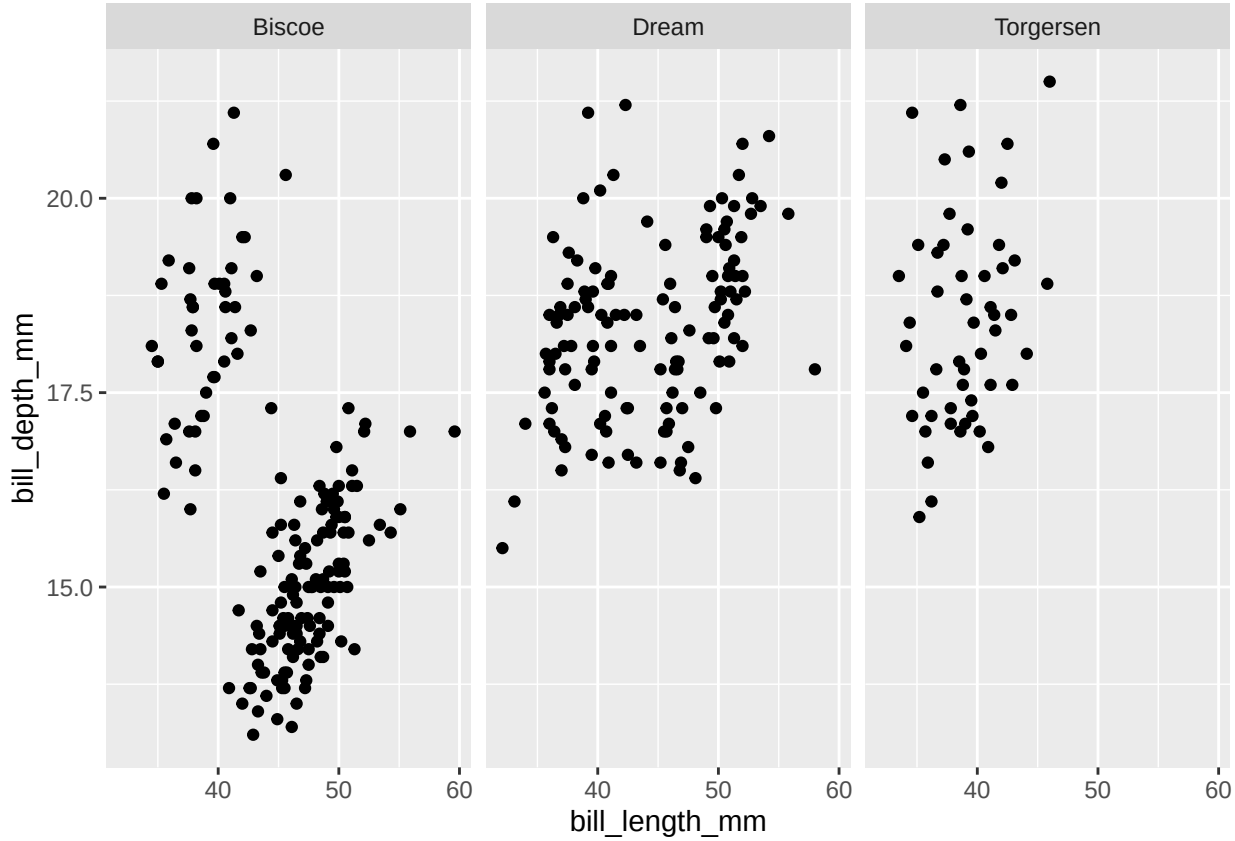
Figure 8: Relationship between body characteristics variables by sex

According to the Figure 8, there is a strong positive linear relationship between flipper length and body mass index. The pearson's correlation coefficients between flipper length and body mass index are approximately same for female and male penguins.

Few other useful commands for plots.

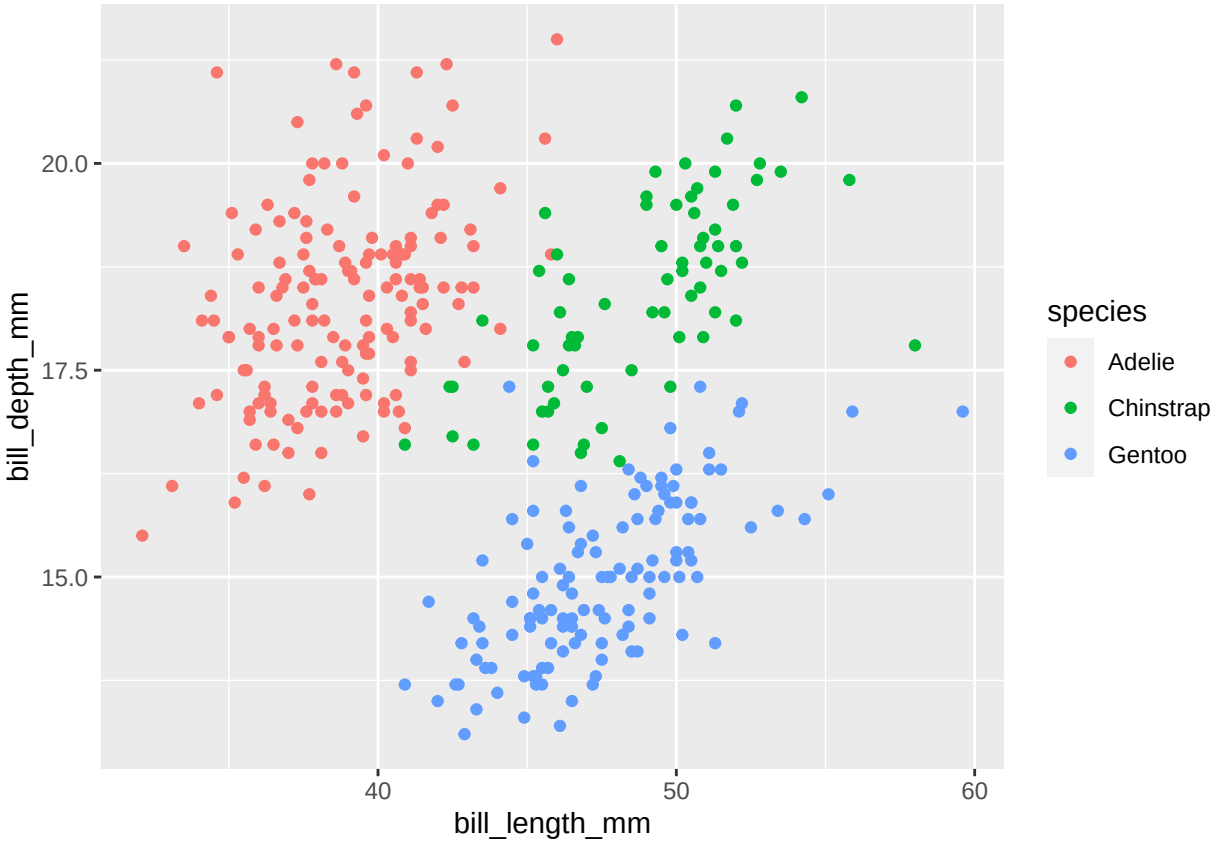
```
ggplot(penguins)+
  geom_point(aes(bill_length_mm, bill_depth_mm))+facet_wrap(~island)
```

```
Warning: Removed 2 rows containing missing values (geom_point).
```




```
ggplot(penguins)+
  geom_point(aes(bill_length_mm, bill_depth_mm, col=species))
```

```
Warning: Removed 2 rows containing missing values (geom_point).
```


bill_depth_mm
20.0
17.5
15.0
40
50
60
bill_length_mm
species
Adelie
Chinstrap
Gentoo