Exercise 2

(a) Setup

```
# install.packages("palmerpenguins")
library(palmerpenguins)
```

```
##
## Attaching package: 'palmerpenguins'

## The following objects are masked from 'package:datasets':
##
##     penguins, penguins_raw
```

```
penguins <- palmerpenguins::penguins
```

(b) Structure and dimensions

```
str(penguins)
```

```
## tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_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: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

```
dim(penguins)
```

```
## [1] 344   8
```

```
names(penguins)
```

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

```
sapply(penguins, class)
```

```
##           species            island    bill_length_mm     bill_depth_mm
##          "factor"          "factor"         "numeric"         "numeric"
## flipper_length_mm       body_mass_g               sex              year
##         "integer"         "integer"          "factor"         "integer"
```

#344 Observations. 8 variables, (3 categorical variables, 5 numeric)

(c) Summary

```
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           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

(d) Missing values

```
colSums(is.na(penguins))
```

```
##           species            island    bill_length_mm     bill_depth_mm
##                 0                 0                 2                 2
## flipper_length_mm       body_mass_g               sex              year
##                 2                 2                11                 0
```

#variables with missing data are bill_length_mm, bill_depth_mm, body_mass_g, sex #(e) Most variation

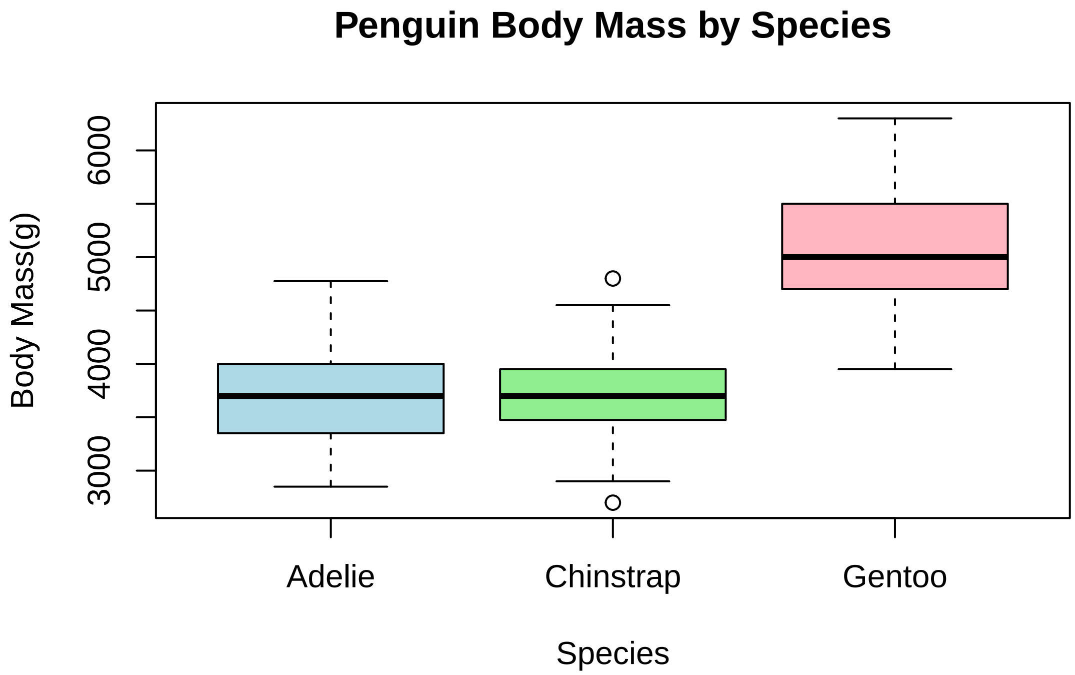
```
sapply(penguins[, sapply(penguins, is.numeric)], var, na.rm = TRUE)
```

```
##    bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
##      2.980705e+01      3.899808e+00      1.977318e+02      6.431311e+05
##              year
##      6.697064e-01
```

#body_mass_g has the most variation because its number is the highest

(F) Boxplot of body mass by species

```
boxplot(body_mass_g ~ species, data = penguins,
        main = "Penguin Body Mass by Species",
        ylab = "Body Mass(g)", xlab = "Species",
        col = c("lightblue", "lightgreen", "lightpink"))
```

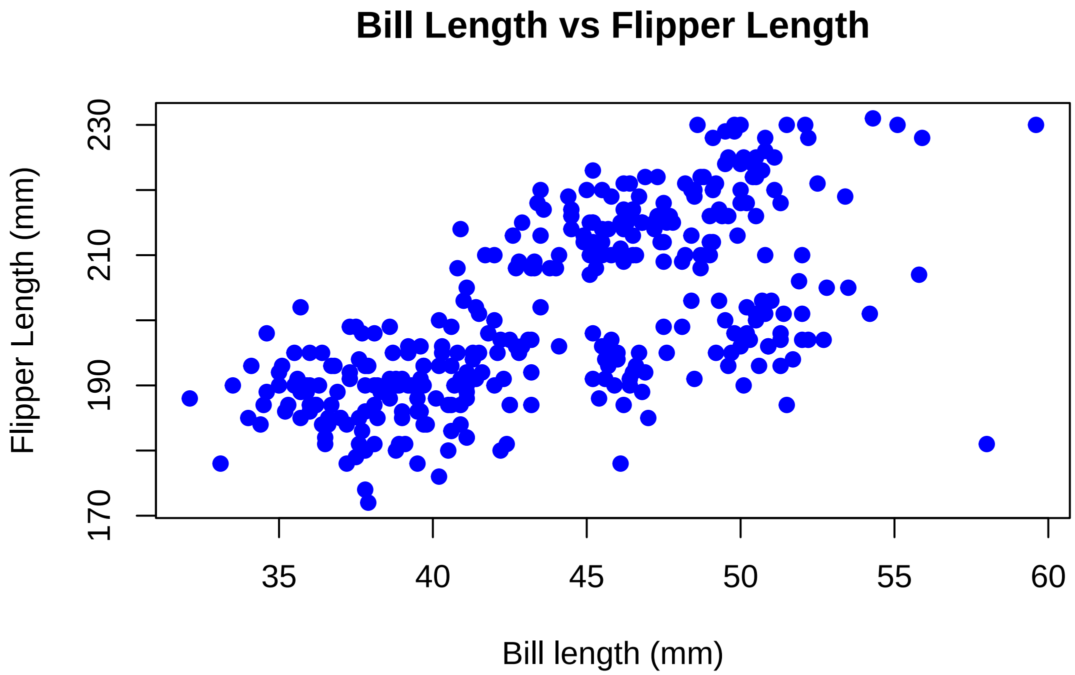


##

Choice ii -> “The species have different body mass.”

(g) Scatterplot

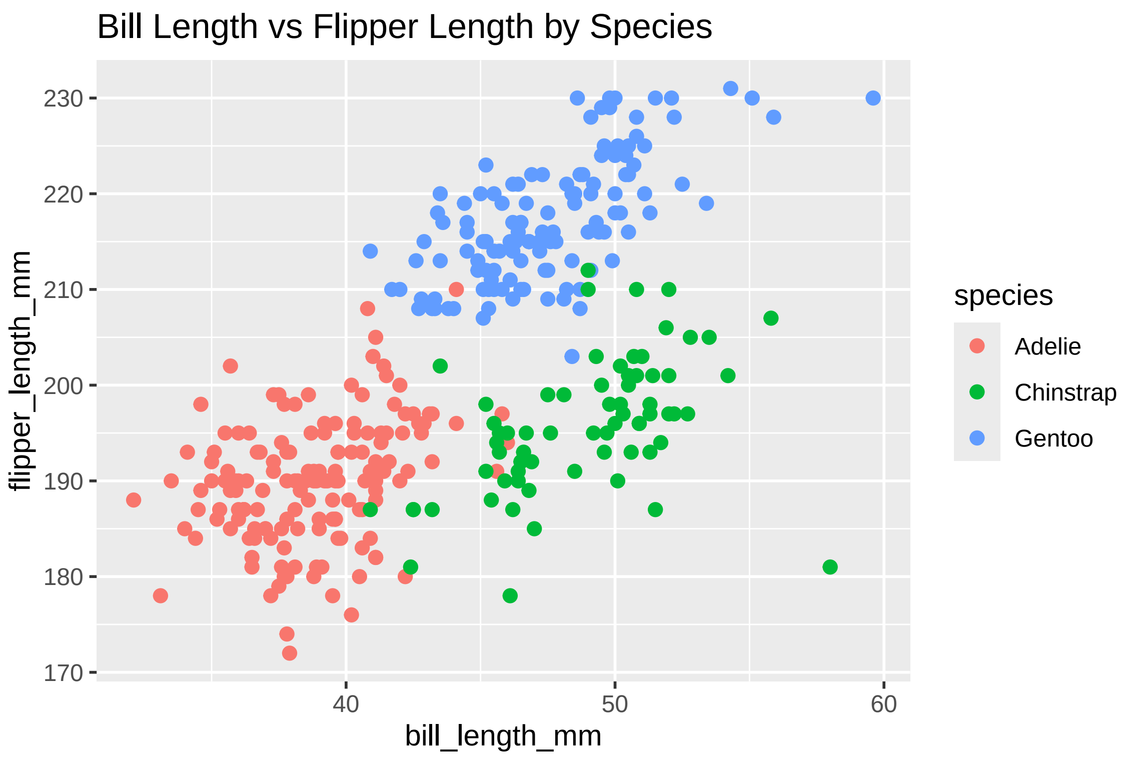
```
plot(penguins$bill_length_mm, penguins$flipper_length_mm,
     main = "Bill Length vs Flipper Length",
     xlab = "Bill length (mm)", ylab = "Flipper Length (mm)",
     pch = 19, col = "blue")
```



Extra credit (ggplot2)

```r
library(ggplot2)
ggplot(penguins, aes(x = bill_length_mm, y = flipper_length_mm, color = species)) +
  geom_point(size = 2) +
  labs(title = "Bill Length vs Flipper Length by Species")
```

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



#Choice i -> there is a positive relationship

(h) Which island has the largest number of penguins?

```
table(penguins$island)
```

```
##
##    Biscoe     Dream Torgersen
##       168       124        52
```

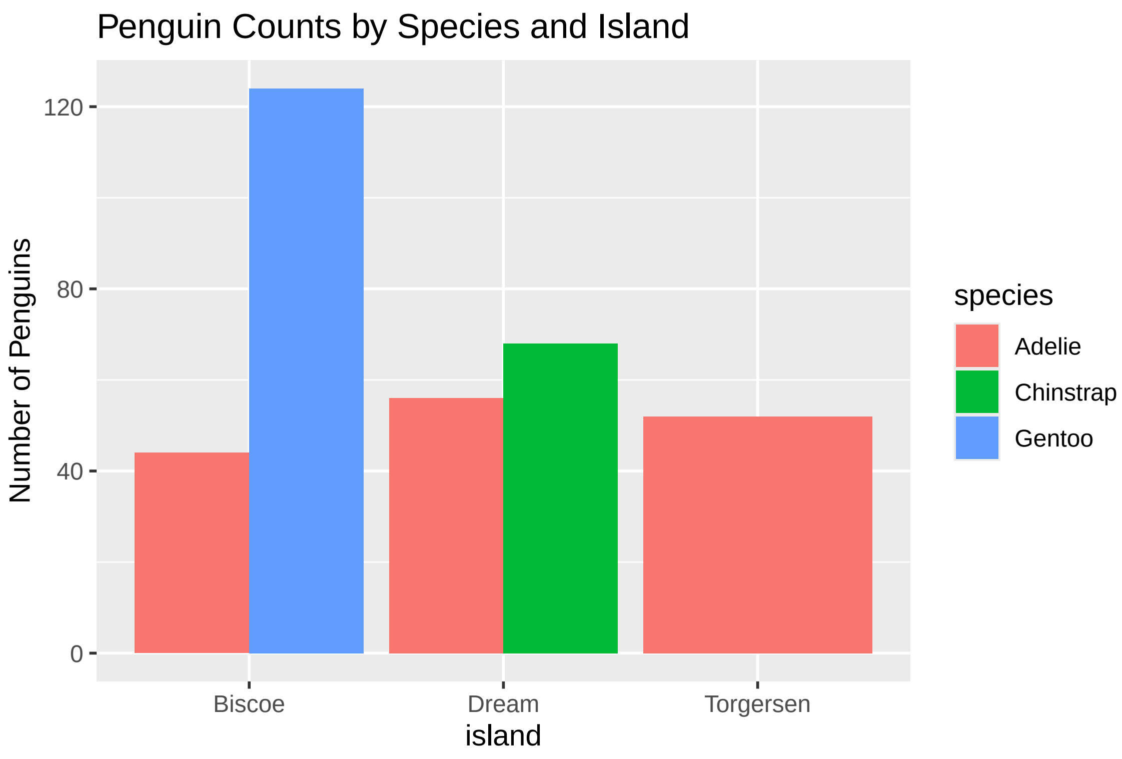
```
table(penguins$island, penguins$species)
```

```
##
##              Adelie Chinstrap Gentoo
##   Biscoe         44         0    124
##   Dream          56        68      0
##   Torgersen      52         0      0
```

#Biscoe has the largest number of penguins #There Are Adelie and Gentoo penguins on Biscoe #There are Adelie and Chinstrap penguins on Dream #There are only Adelie penguins on Torgersen

#(i) Penguins by island

```
ggplot(penguins, aes(x = island, fill = species)) +
  geom_bar(position = 'dodge') +
  labs(title = "Penguin Counts by Species and Island", y = "Number of Penguins")
```


Penguin Counts by Species and Island
Number of Penguins
120
80
40
0
Biscoe
Dream
Torgersen
island
species
Adelie
Chinstrap
Gentoo