14 Base R plots

Graphical presentations in Statistics

- Graphs are helpful for presenting a summary of data and results of a statistical analysis
- *John W. Tukey*, the father of *exploratory data analysis* once said

> The greatest value of a picture is when it forces us to notice what we never expected to see.

Base R plot functions

Graphical presentations of data

- There are several graphs available to use for describing data, and the selection of the most appropriate graph depends on the data type and the research objectives
- **Quantitative data**
  - Histogram
  - Boxplot
  - Scatter plot
- **Qualitative data**
  - Bar chart
  - ~~Pie chart~~

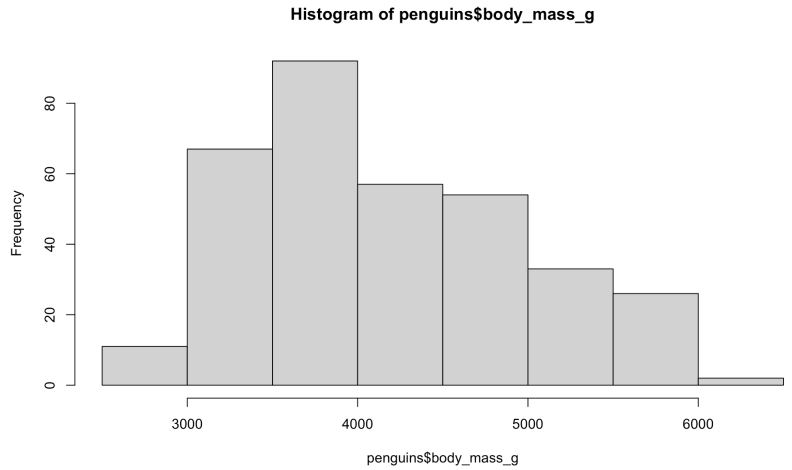
Graphical presentations of data

- Bivariate analysis involves two variables, depending on the combinations of the variables, i.e., qualitative or quantitative, there are different ways of presenting data graphically
- Quantitative-qualitative combination
  - Histogram and boxplot can be used for different levels of a qualitative variable
- Quantitative-quantitative and qualitative-qualitative combinations
  - Bar chart and scatter plot can be used

1 Histogram

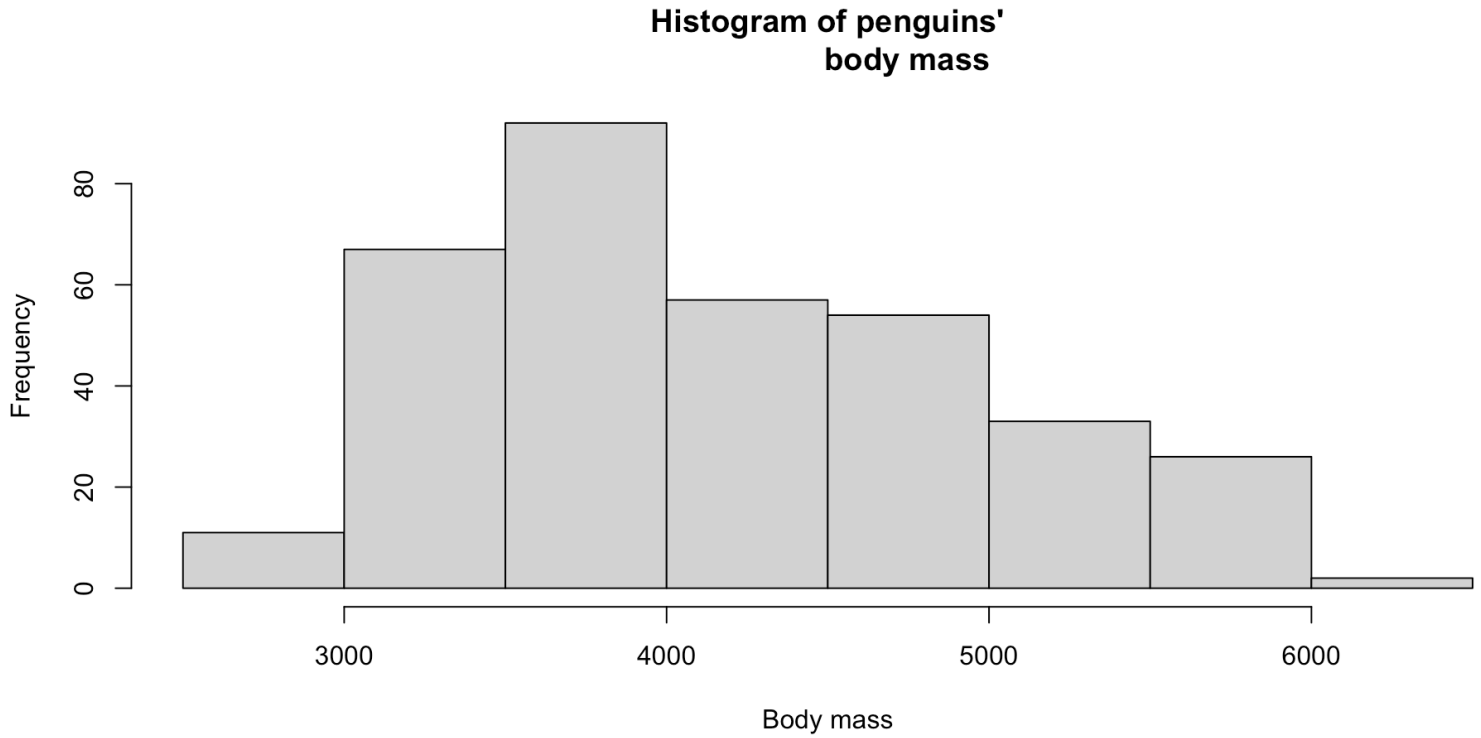
- The function `hist()` is used to obtain a histogram of a quantitative variable
- Syntax of `hist()`
- `x` is a quantitative vector

```
hist(x = penguins$body_mass_g)
```

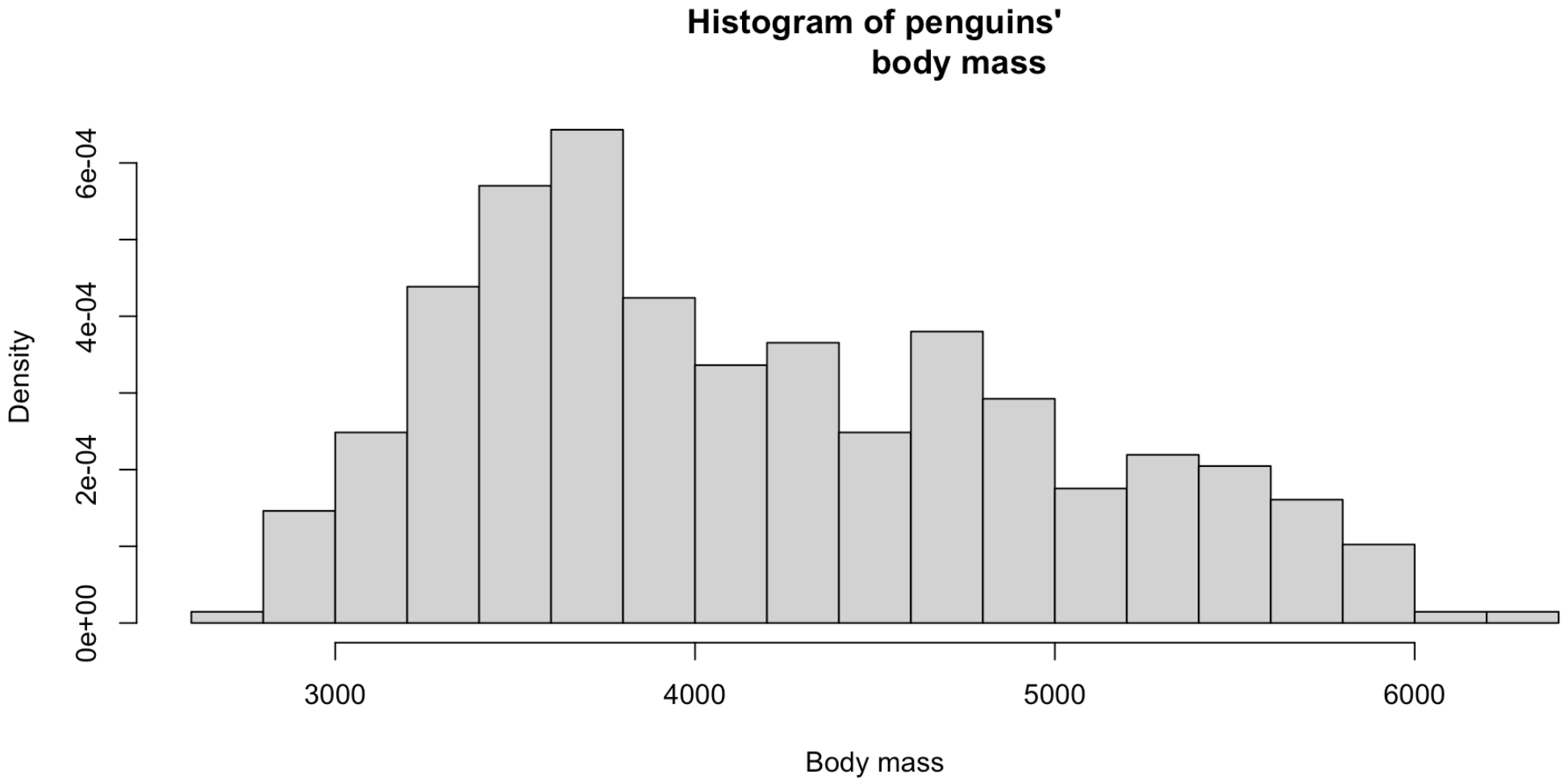


Graphical presentations of data

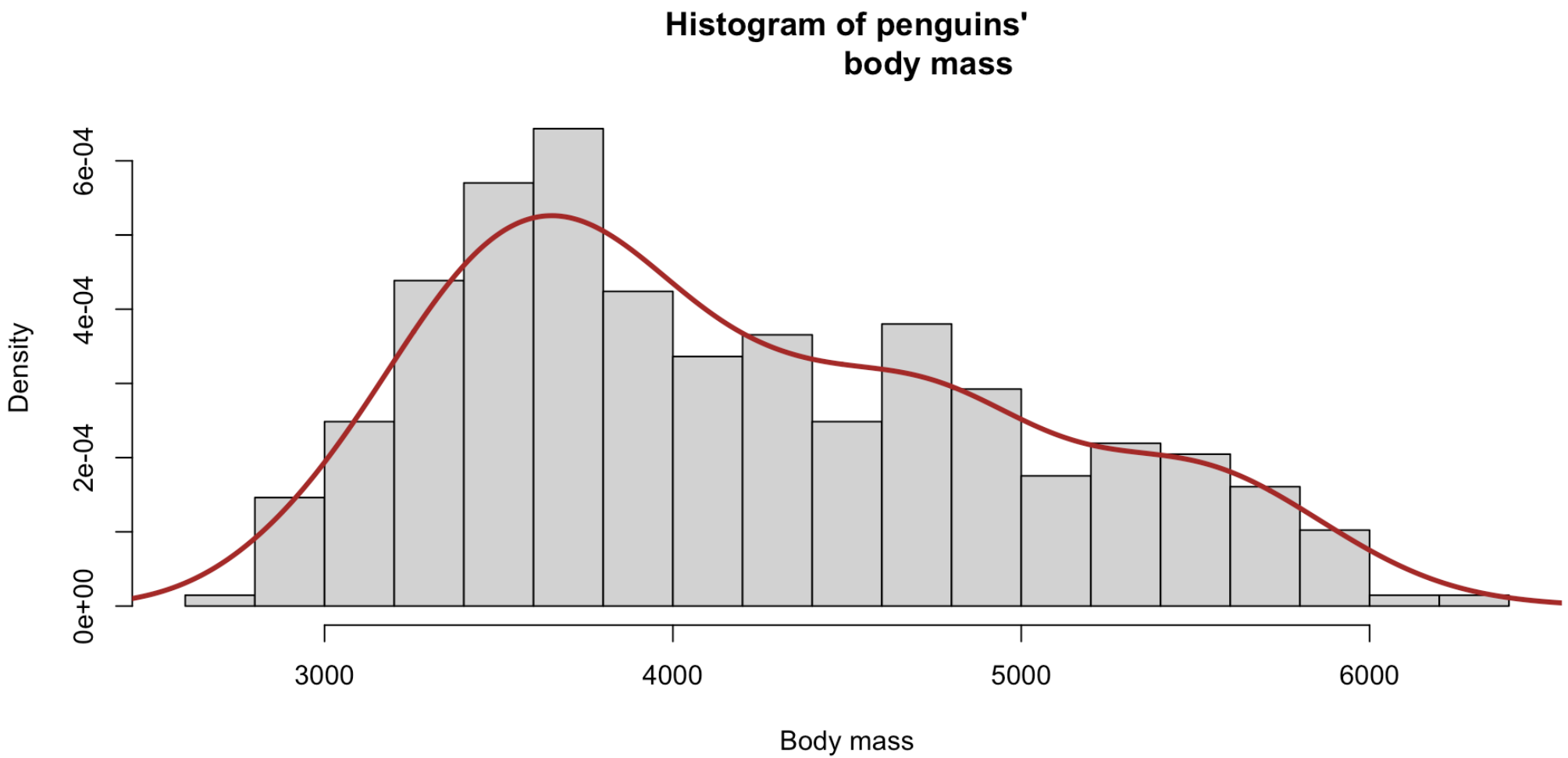
- Some useful arguments of `hist()`:
  - `xlab`, `main`, `probability`, etc.



Graphical presentations of data



Graphical presentations of data



Exercise 3.1.1

(use `mtcars` data frame to answer the followings)

- Create a histogram of `mpg` with appropriate labels
- Add density line to the plot obtained in Question 1.

2 Boxplot

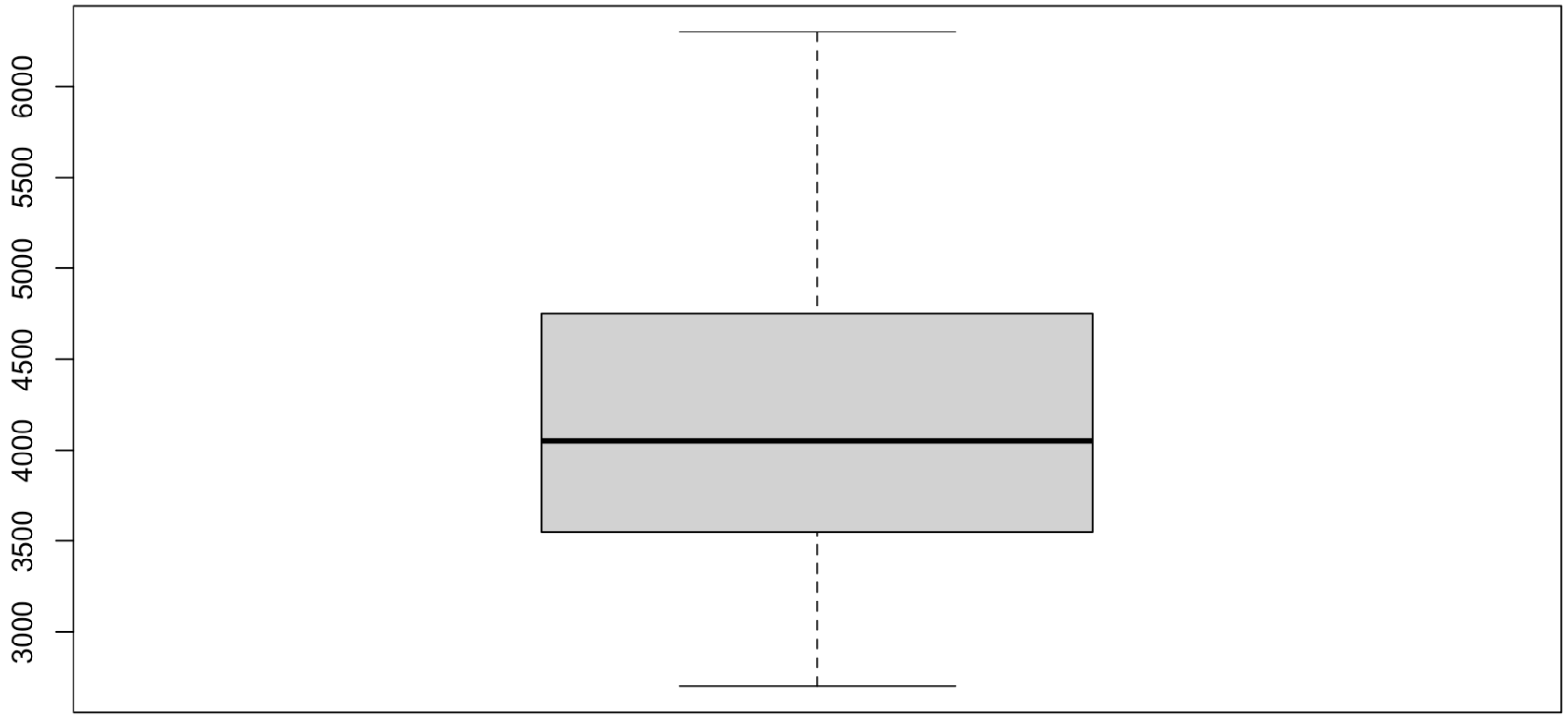
- Boxplot is a useful graphical tool that can be used to compare distribution of a quantitative variable at different levels of a qualitative variable
  - E.g. examine the distribution of body mass over different species of penguins

Exercise 3.1.1

- `boxplot()` function can be used for both univariate and bivariate analysis
  - `boxplot(x)` is used to obtain a boxplot of a single quantitative vector `x`
  - `boxplot(formula, data)` function is used for a bivariate analysis, where the formula species the quantitative and qualitative variables of interest
  - `formula = quant_var ~ qual_var`
  - `data` is a data frame that must contain `quant_var` and `qual_var`

Exercise 3.1.1

```
boxplot(x = penguins$body_mass_g)
```



Exercise 3.1.2

(use `mtcars` data frame to answer the followings)

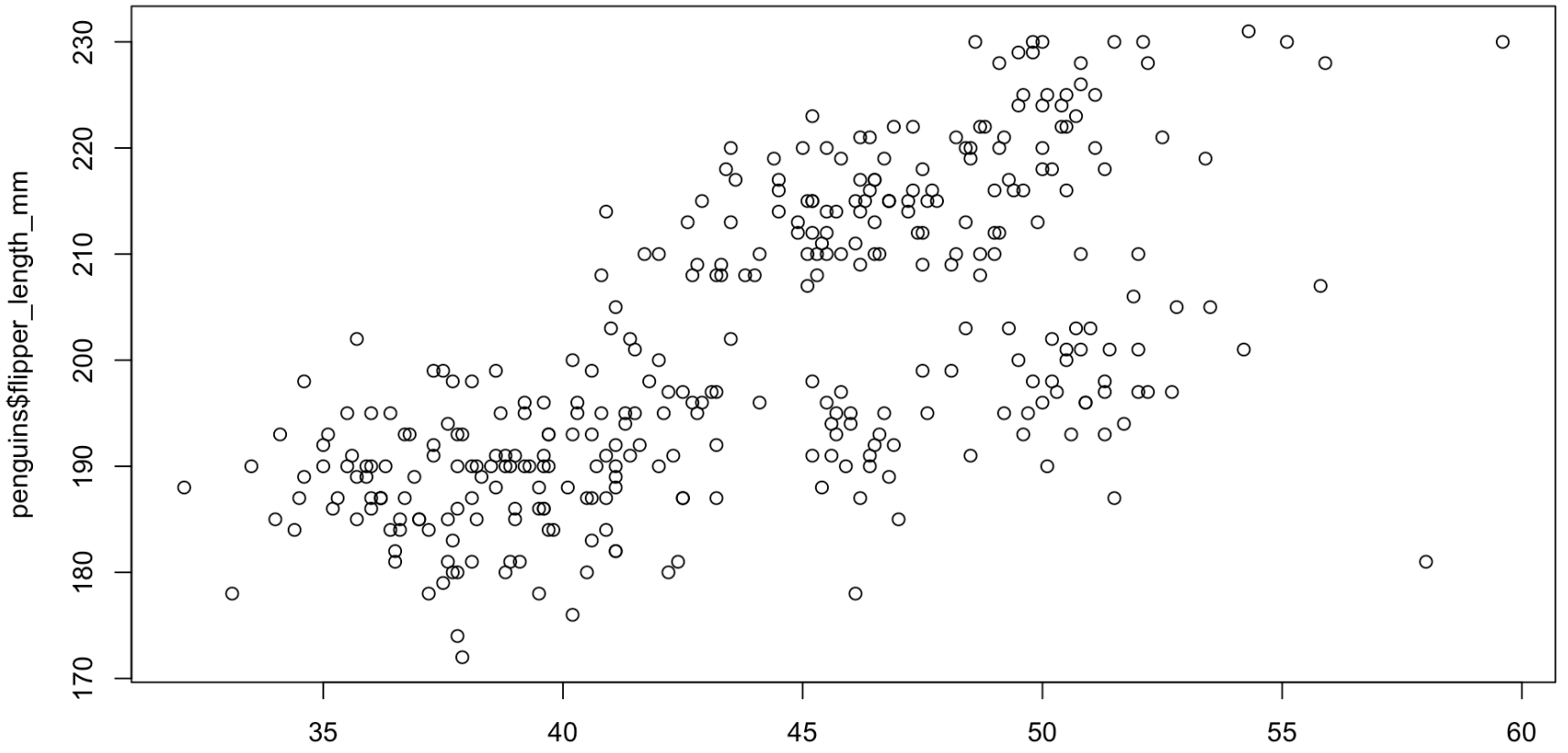
- Create a boxplot of `qsec` with appropriate labels.
- Create a boxplot of `mpg` to compare its distribution at different levels of `cyl`

3 Scatter plot

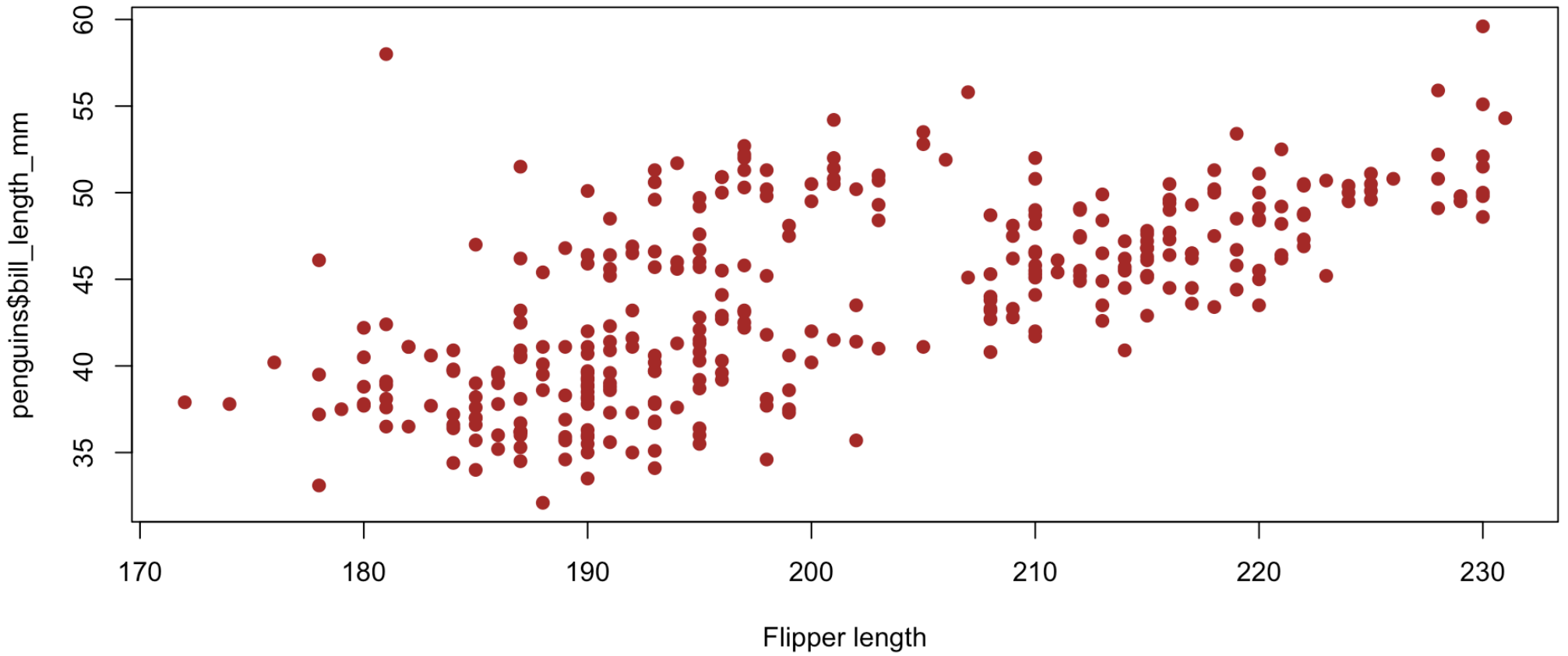
- The function `plot(x, y)` is used to obtain a scatter plot of two quantitative variables `x` and `y`
- Some useful arguments of `plot()` function
  - `xlab`, `ylab`, `main`
  - `pch` (point type)
  - `cex` (size of points), etc.

Exercise 3.1.2

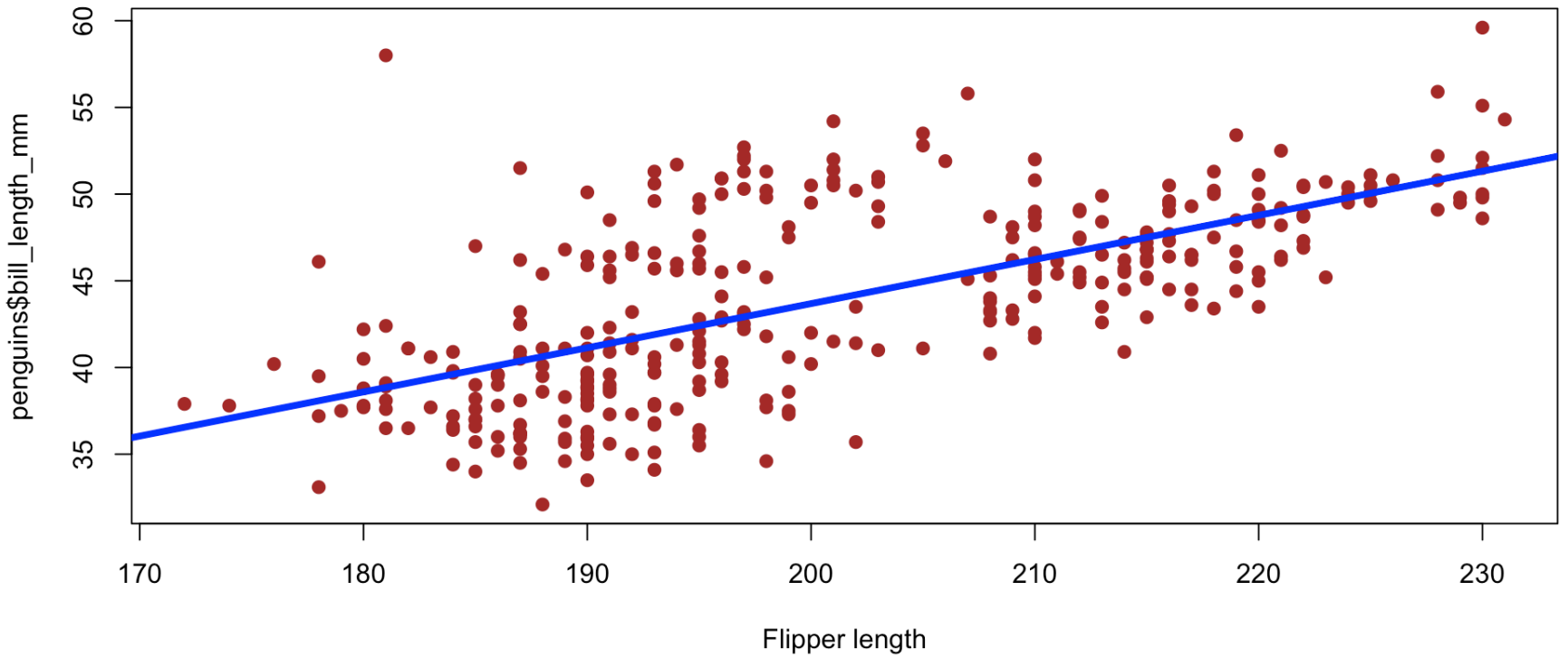
```
#
plot(x = penguins$bill_length_mm,
     y = penguins$flipper_length_mm)
```



Exercise 3.1.2



Scatter plot with a linear model fit



- `lm()` is for fitting a linear model

Exercise 3.1.3

(use `mtcars` data frame to answer the followings)

- Create a scatter plot to examine the association between `mpg` and `disp`
- Add the fit of a linear regression model `mpg` on `disp` to the plot obtained in Question 6

4 Bar chart

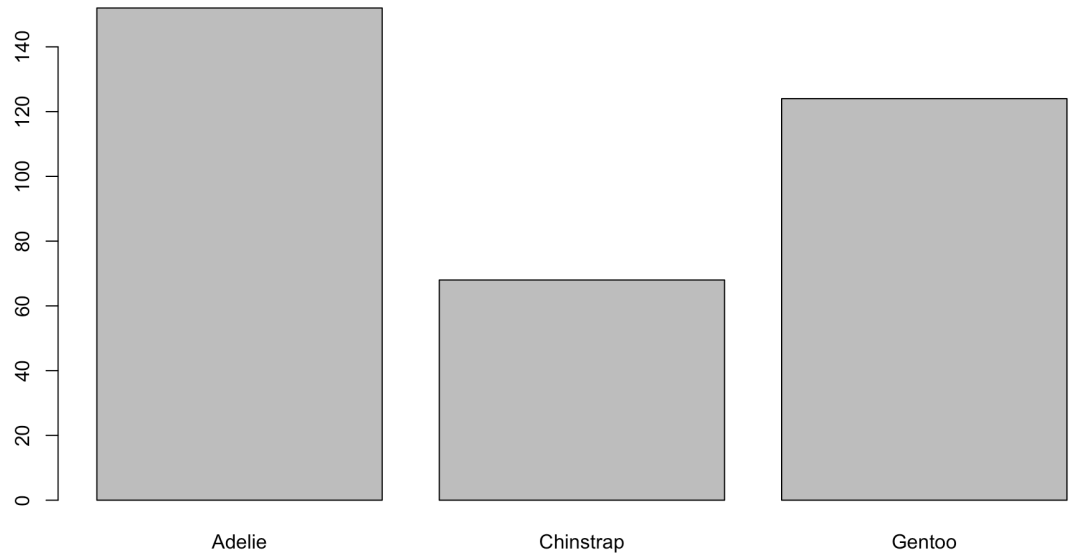
- Bar chart is used to examine the distribution of a qualitative variable
- The function `barplot(height, ...)` is used to obtain a bar chart in R, where `height` represents a frequency
- `table()` function takes a qualitative variable as an argument and returns `height`, the frequency corresponding to each level of the qualitative variable

Frequency distribution of `species`

```
table(penguins$species)
```

```
#>
#>    Adelie Chinstrap    Gentoo
#>       152        68       124
```

```
barplot(height = table(penguins$species))
```



Exercise 3.1.4

(use `mtcars` data frame to answer the followings)

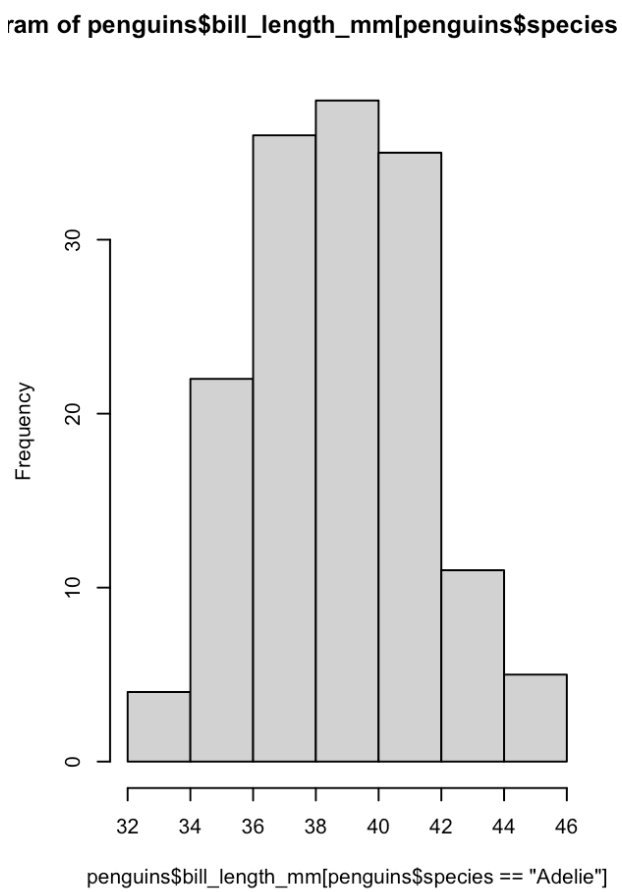
- Create a barchart if `cyl`

Bivariate analysis

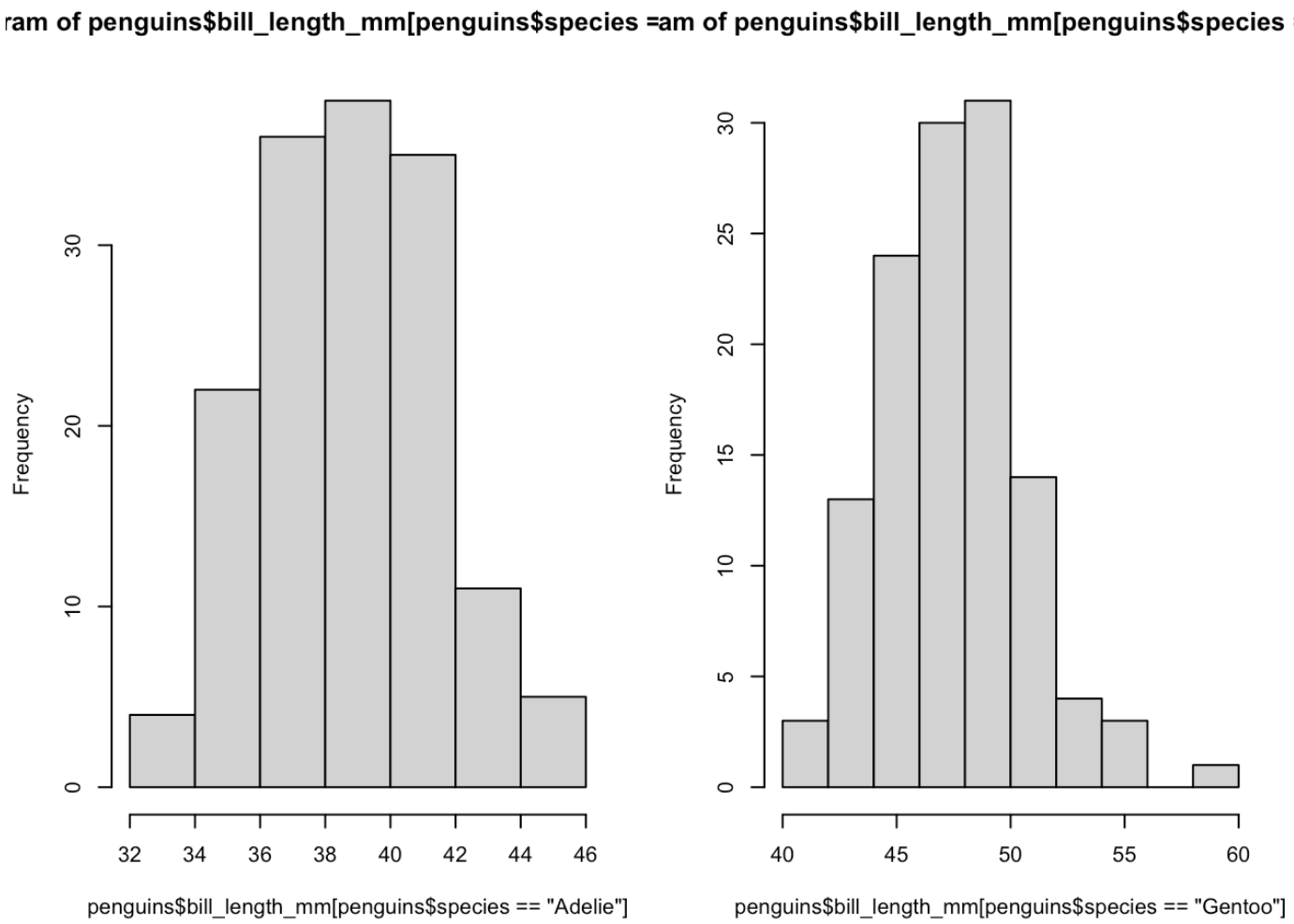
- The function `par()` has many arguments that can be used to produce high-quality graphs using base R plot functions
- `mfrow` argument of `par()` is used to split a figure layout into a number of rows and columns
  - E.g. `mfrow = c(2, 3)` will split the figure layout into two rows and three columns

Bivariate analysis

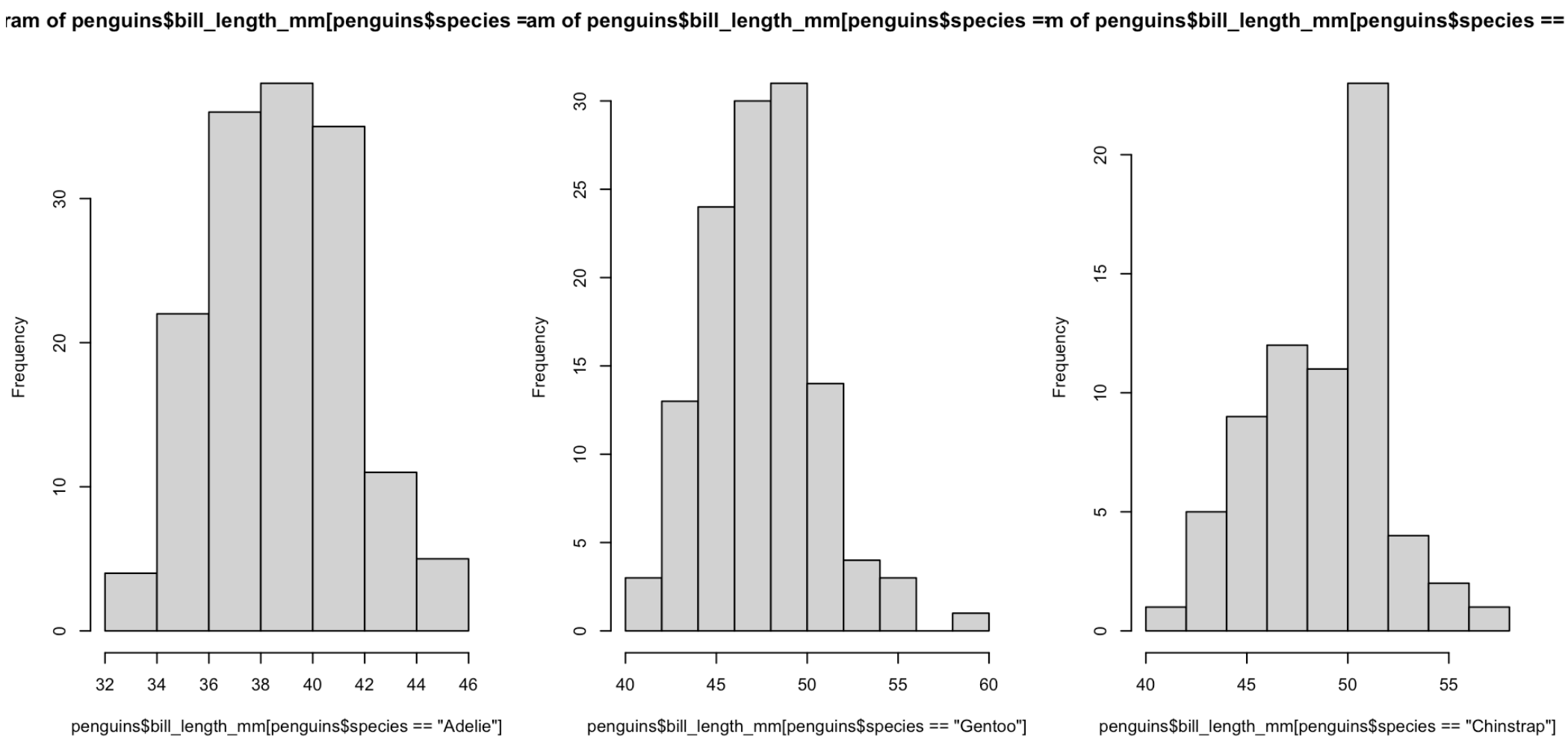
Distribution of `bill_length_mm` at different levels of `species`



Bivariate analysis

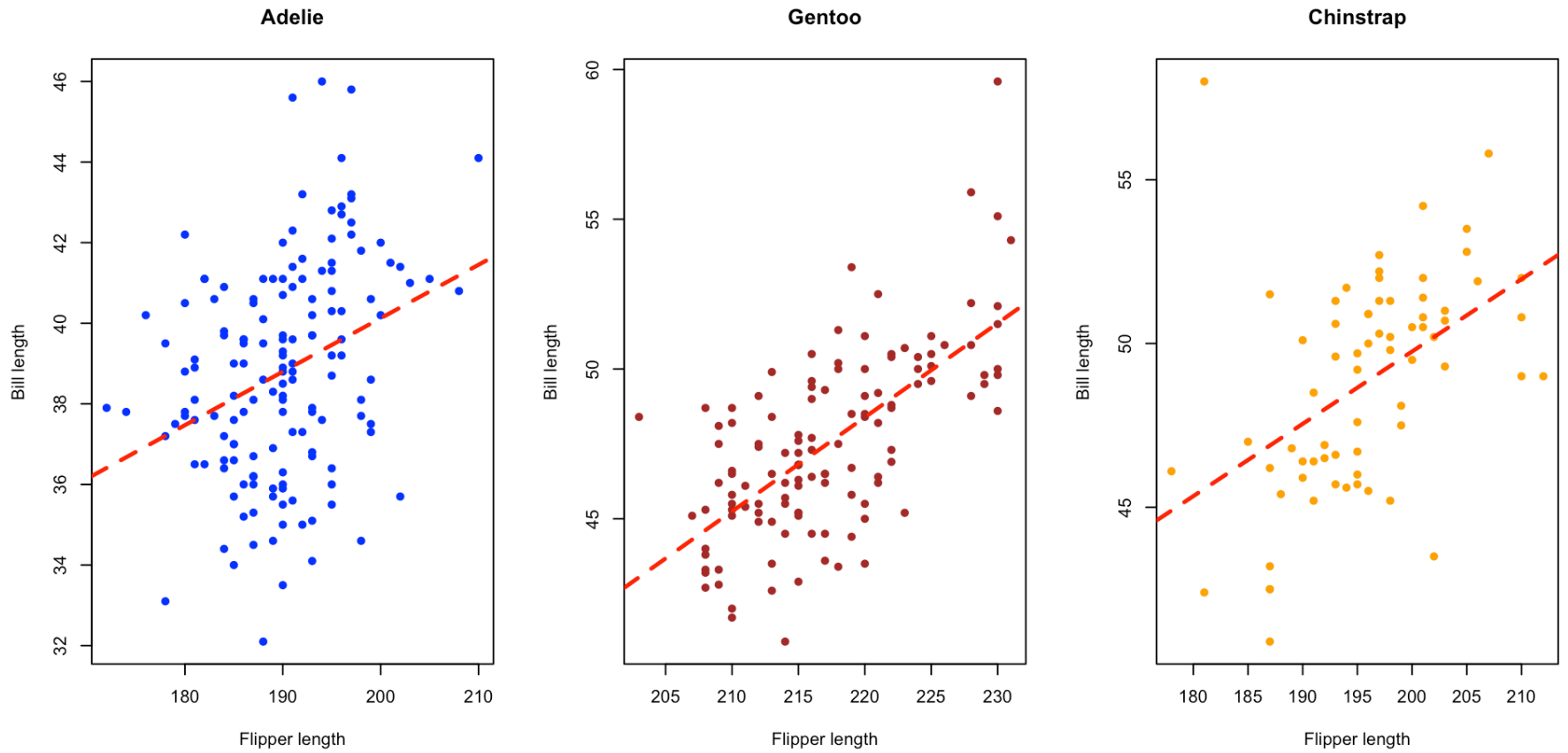


Bivariate analysis



Exercise 3.1.5

- Association between bill and flipper lengths by species



Exercise 3.1.5