Week 5 Practice Exercises

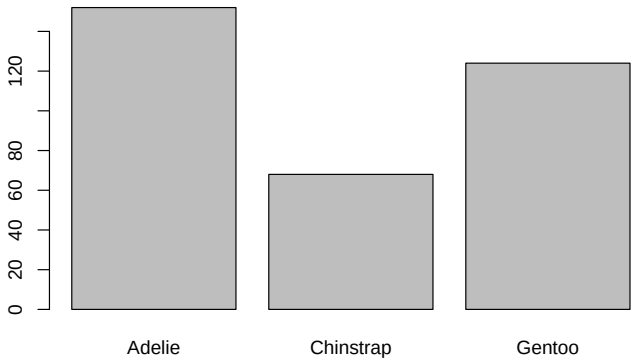
These exercises are an accompaniment to online lessons 5.1–5.5. We recommend trying them as they come up in the lessons, and completing them before moving on to the Week 5 Question Sheet. All of these questions can be answered using either base R graphics or ggplot, although some tasks are more straightforward in one or the other.

Exercise 1: Bar plots

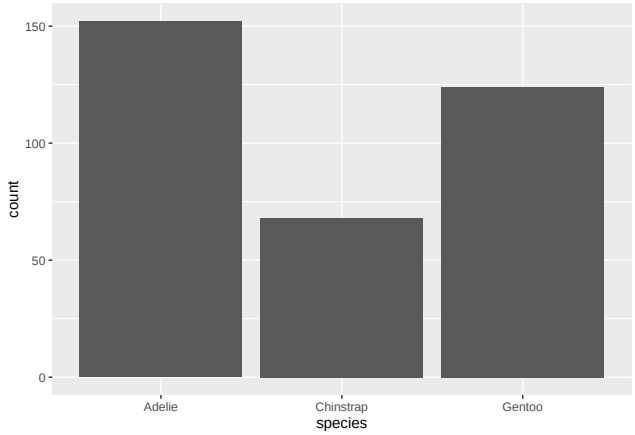
As setup, load the Palmer penguins data as in previous weeks.

(a) Produce a bar plot that lets you determine which species of penguin was sampled most often.

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

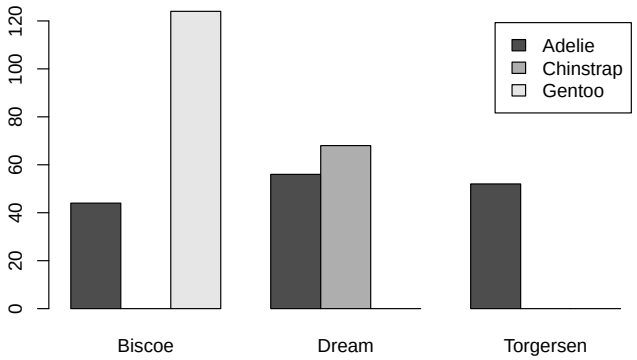



```
ggplot(penguins) +
  geom_bar(aes(species))
```

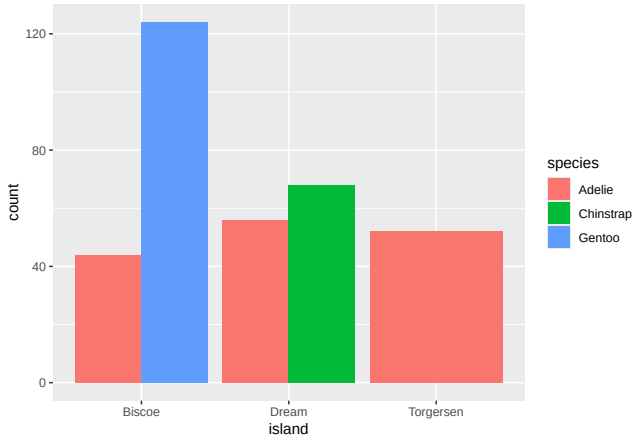


(b) Produce another version that shows the number of penguins of each species sampled on each island.

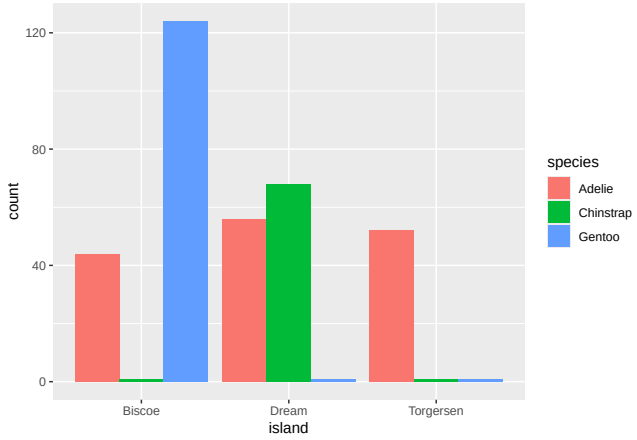
```
barplot(table(penguins$species, penguins$island), beside=TRUE, legend.text = TRUE)
```




```
ggplot(penguins) +
  geom_bar(aes(island, fill=species), position=position_dodge())
```

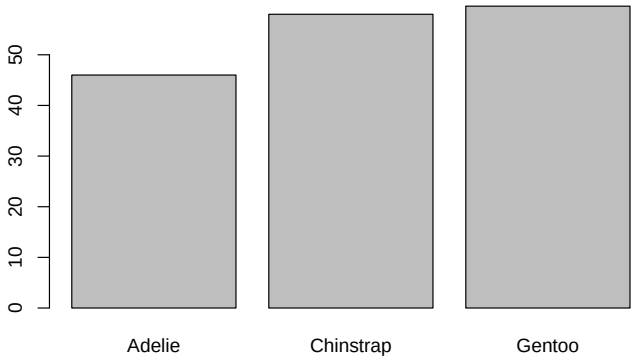



```
# The way ggplot expands the bars to fill the available width is not very nice.
# To overcome this, we can force all levels to appear by adding a row of data for
# each missing bar:
penguins %>%
  # Manually add one penguin for each of the missing species x island combinations
  bind_rows(data.frame(species=c("Chinstrap","Gentoo","Chinstrap","Gentoo"),
                       island=c("Biscoe","Dream","Torgersen","Torgersen"))) %>%
  ggplot() +
  geom_bar(aes(island, fill=species), position=position_dodge())
```

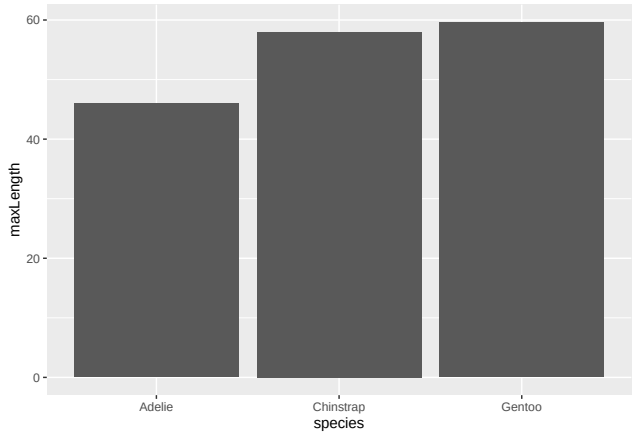


(c) Calculate the maximum bill length for each species of penguin and plot these on a bar plot.

```r
barplot(tapply(penguins$bill_length_mm, list(penguins$species), max, na.rm=TRUE))
```



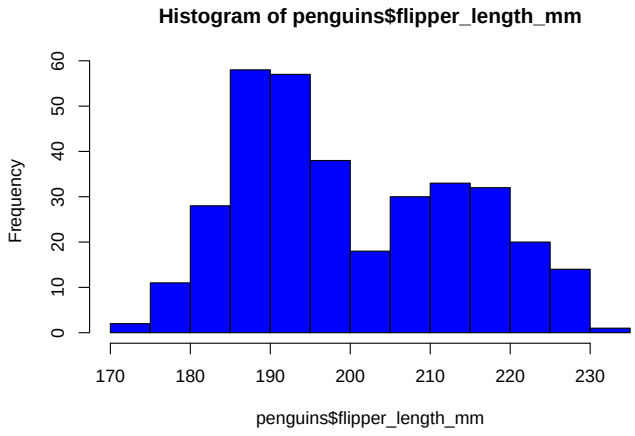

```r
penguins %>%
  group_by(species) %>%
  summarise(maxLength = max(bill_length_mm, na.rm=TRUE)) %>%
  ggplot() +
  geom_col(aes(species, maxLength))
```



Exercise 2: Histograms

(a) Make a histogram of the flipper length, and make the bars blue.

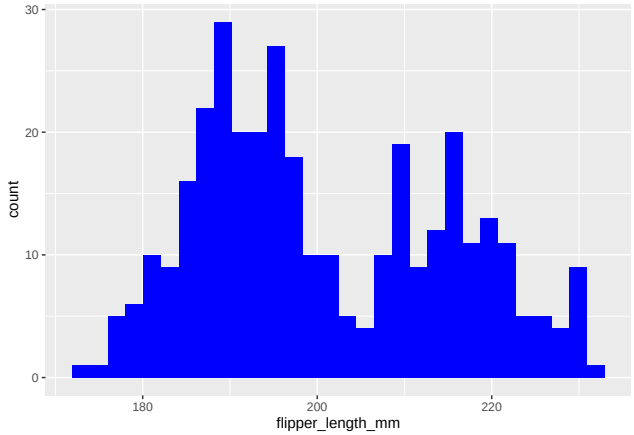
```r
hist(penguins$flipper_length_mm, col="blue")
```



```
ggplot(penguins) +
  geom_histogram(aes(flipper_length_mm), fill="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

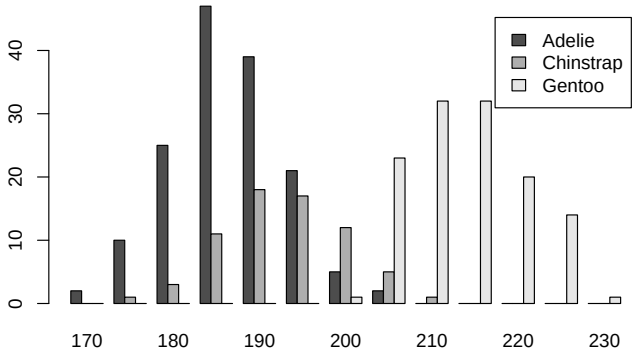


(b) The previous histogram has two "bumps" (the technical term for this is "multimodal"). it is often the case with multimodal data that the difference in modes is caused by some underlying grouping. To explore this, produce a histogram of flipper length by species.

```
require(plotrix)
```

```
## Loading required package: plotrix
```

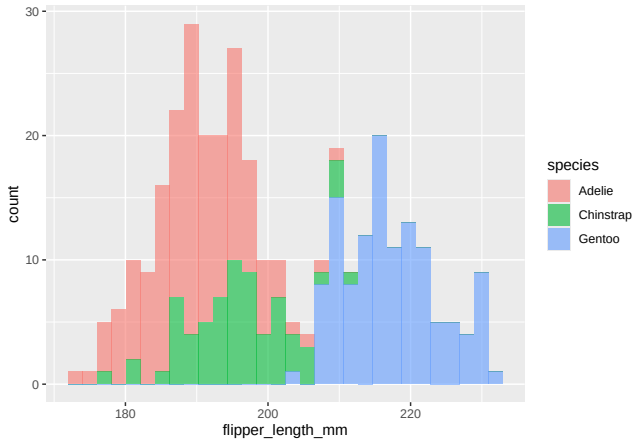
```
multhist(split(penguins$flipper_length_mm, penguins$species), legend.text=TRUE)
```




```
ggplot(penguins) +
  geom_histogram(aes(flipper_length_mm, fill=species), alpha=0.6)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

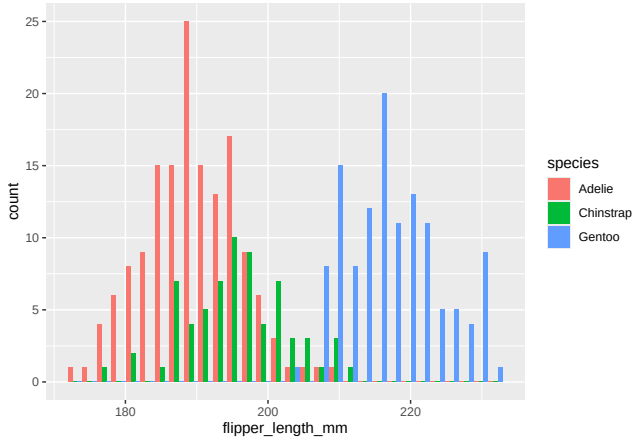



```
# Reducing alpha (the default is 1.0) makes the colours a bit less eye-burning

# for a version that is closer to the multhist version:
ggplot(penguins) +
  geom_histogram(aes(flipper_length_mm, fill=species), position=position_dodge())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

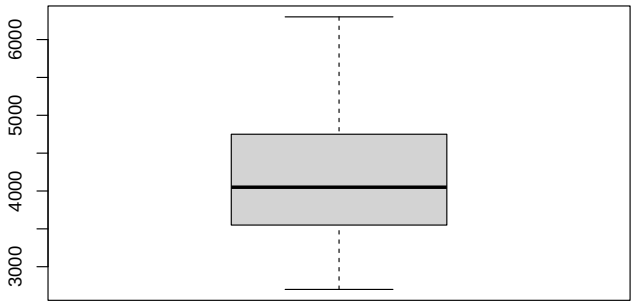
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Exercise 3: Boxplots

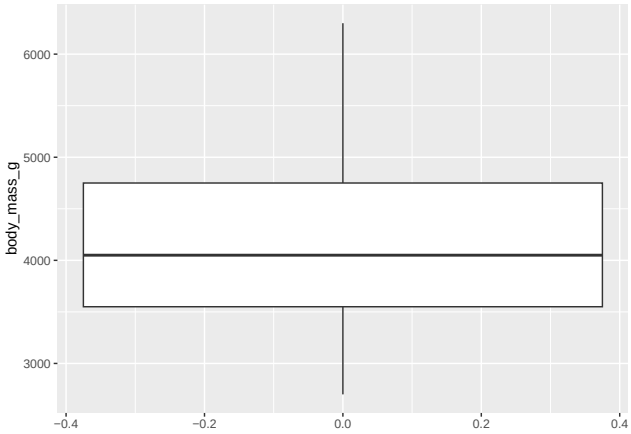
(a) Produce a box plot of penguin body mass.

```
boxplot(penguins$body_mass_g)
```



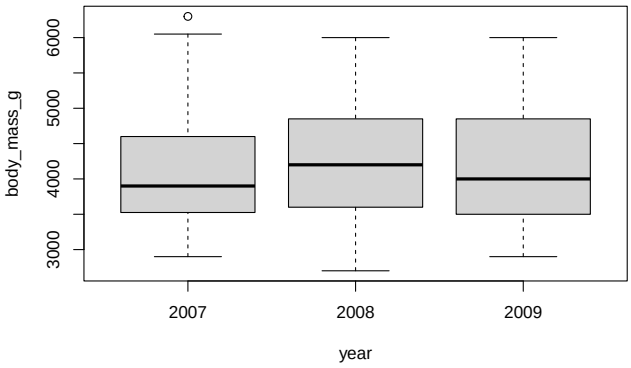
```
ggplot(penguins) +
  geom_boxplot(aes(y=body_mass_g))
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



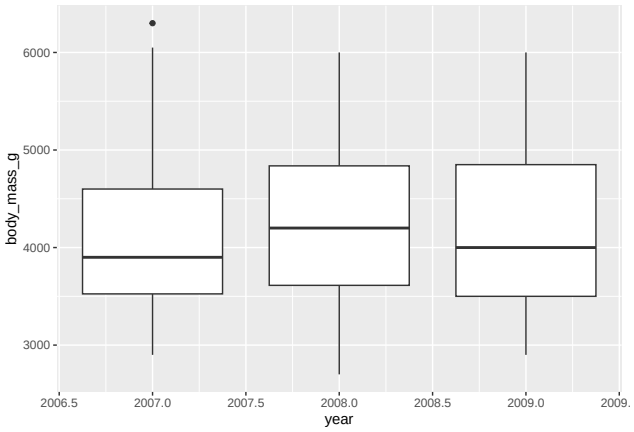
(b) Make a boxplot that lets you evaluate whether the distribution of body mass changed among years.

```
boxplot(body_mass_g~year, data=penguins)
```




```
penguins %>%
  ggplot() +
  geom_boxplot(aes(x=year, y=body_mass_g, group=year))
```

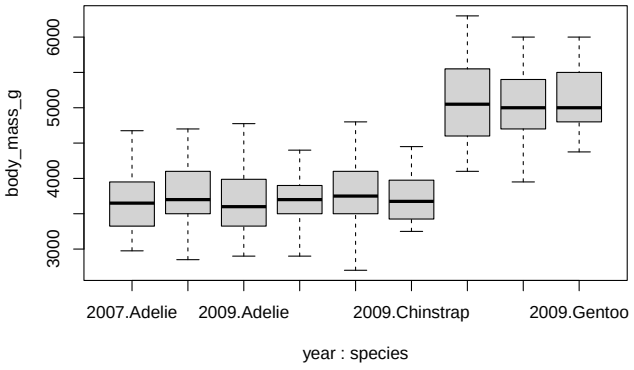
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
# group=year is necessary because year is a numeric variable and ggplot doesn't
# automatically understand that it can be interpreted as a categorical variable.
# another version that works is
#           ...aes(x=factor(year), y=body_mass_g)
# but if you try the simple version
#           ...aes(x=year, y=body_mass_g)
# you'll get a single boxplot (as in part (a)) and a warning message.
```

(c) Extend this analysis further: produce a boxplot that will describe the distribution of body mass for each species within each year. The observations within a species should appear next to one another.

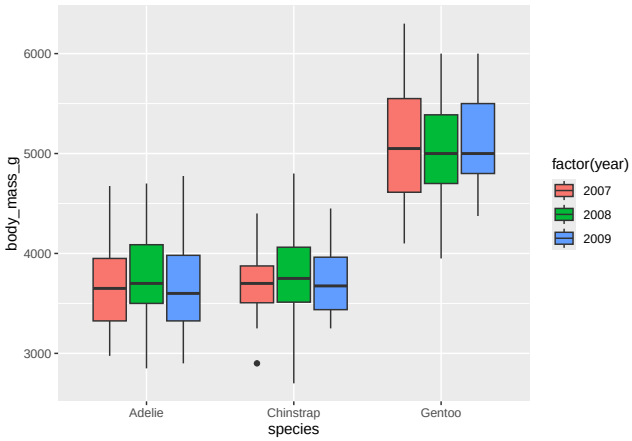
```
boxplot(body_mass_g~year+species, data=penguins)
```




```
# The default labelling isn't very good: see Exercise 5.

penguins %>%
  ggplot() +
  geom_boxplot(aes(x=species, y=body_mass_g, fill=factor(year)))
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

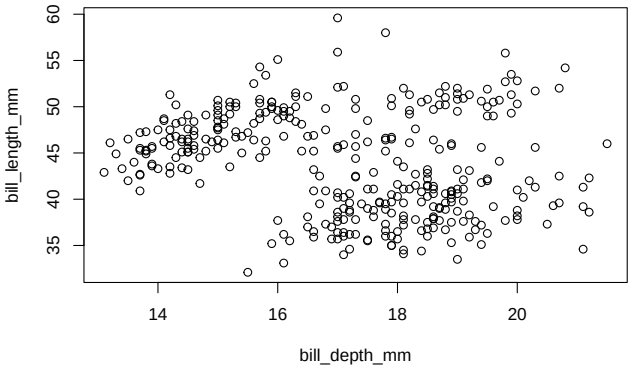



```
# Again, we have to tell ggplot to interpret year as a factor, not a continuous
# variable, for it to understand what we mean.
```

Exercise 4: Scatter plots

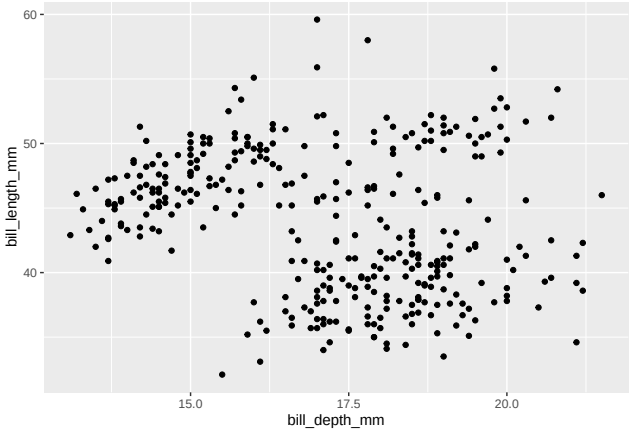
(a) Make a scatter plot of bill length vs. bill depth.

```
plot(bill_length_mm ~ bill_depth_mm, data=penguins)
```



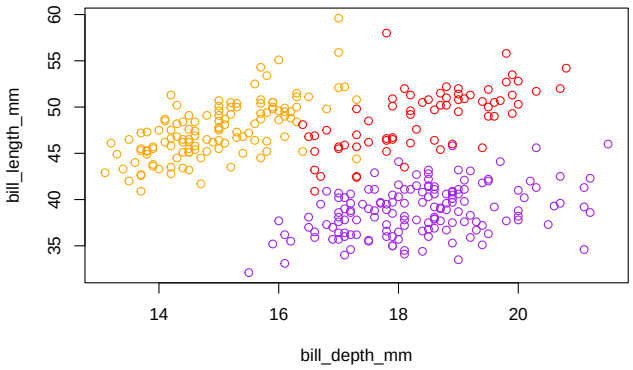

```
ggplot(penguins) +
  geom_point(aes(bill_depth_mm, bill_length_mm))
```

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



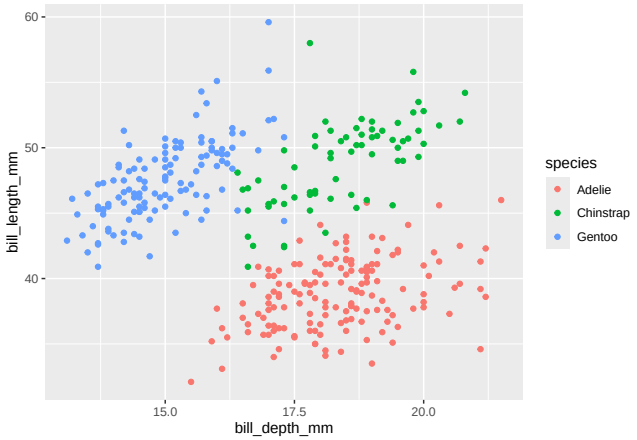
(b) Make this scatter plot again, colouring each data point by species.

```
plot(bill_length_mm~bill_depth_mm, data=penguins,
     col=ifelse(penguins$species=="Adelie","purple",
                ifelse(penguins$species=="Gentoo","orange","red")))
```




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

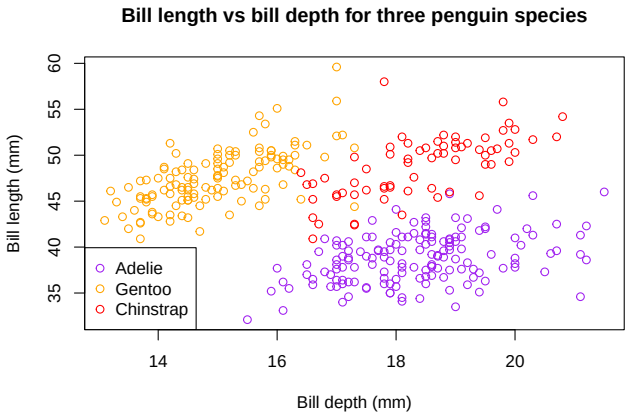
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Exercise 5: Additional customisations

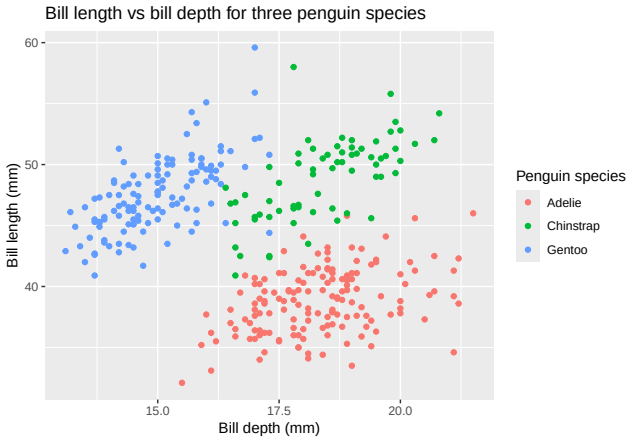
(a) Remake the three-colour scatter plot from Exercise 4b above, adding an appropriate title and better axis labels (i.e. "Bill length (mm)" instead of "bill_length_mm"). Add a legend if your plot doesn't already have one.

```
plot(bill_length_mm~bill_depth_mm, data=penguins,
     col=ifelse(penguins$species=="Adelie","purple",
                ifelse(penguins$species=="Gentoo","orange","red")),
    main="Bill length vs bill depth for three penguin species",
    xlab="Bill depth (mm)",
    ylab="Bill length (mm)")
legend("bottomleft",col=c("purple","orange","red"), pch=1,
                    legend=c("Adelie","Gentoo","Chinstrap"))
```



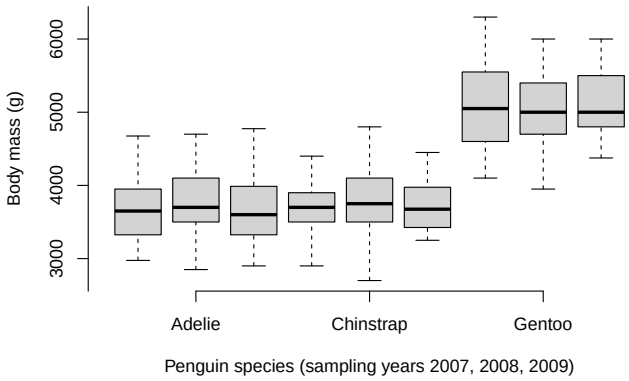

```
ggplot(penguins) +
  geom_point(aes(bill_depth_mm, bill_length_mm, colour=species)) +
  labs(title="Bill length vs bill depth for three penguin species",
       x="Bill depth (mm)",
       y="Bill length (mm)",
       colour="Penguin species") # the legend can be thought of as the colour axis
```

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



(b) (optional) Improve the species × year boxplot in Exercise 3c above. If you made this in base R, fix the labelling. If you made it in ggplot, try changing the colours used for the three species.

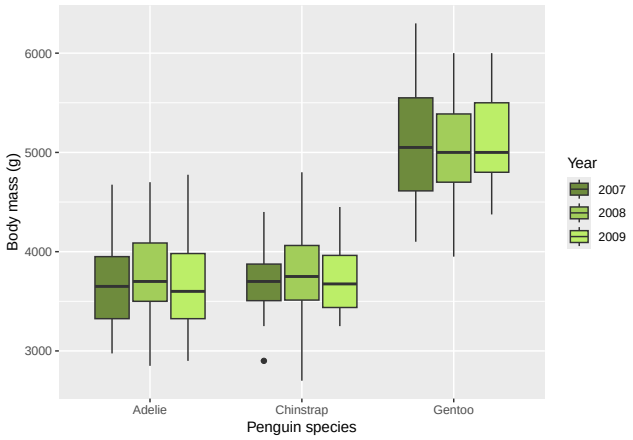
```
# base R: we'll make the boxplot with the axes=FALSE option so that the default
# axes (with poor labelling on the x axis) aren't drawn, and then add
# them back in with labels centered on the three species groups.
boxplot(body_mass_g~year+species, data=penguins,axes=FALSE,
        xlab="Penguin species (sampling years 2007, 2008, 2009)",
        ylab="Body mass (g)")
axis(side=2,at=seq(2000,7000,1000),labels=TRUE) # add the y axis back in
axis(side=1,at=c(2,5,8),labels=c("Adelie","Chinstrap","Gentoo"))
```




```
  # x axis, labelled by species only. Bars are at positions 1:9, so
  # 2,5,8 are the centres of each group of 3


penguins %>%
  ggplot() +
  geom_boxplot(aes(x=species, y=body_mass_g, fill=factor(year))) +
  labs(x="Penguin species",y="Body mass (g)",fill="Year") +
  scale_fill_discrete(type=c("darkolivegreen4","darkolivegreen3","darkolivegreen2"))
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
  # google "r color names" to see a table of these colours you can refer to by name
```