Penguin project

Tom Hardwicke

2024-09-25



# R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Most importantly, we can embed R code into our text! For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

# Load packages

In this first chunk, we load some useful packages. If you don’t have the packages installed already, you will get an error message. To install packages, use the install.packages() function.

# install.packages('tidyverse')  
library(tidyverse)  
# install.packages('here')  
library(here)

# Load data

Ok now we’re set up, let’s load in some data about penguins. The dataset contains observations about penguins collected at the Palmer Station in Antarctica by Gorman, Williams, and Fraser (2014).

penguins\_v1 <- read\_csv(here('data','raw','penguins\_v1.csv'))

Among the variables in penguins\_v1 are:

species: a penguin’s species (Adelie, Chinstrap, or Gentoo).

flipper\_length\_mm: length of a penguin’s flipper, in millimeters.

body\_mass\_g: body mass of a penguin, in grams.

# Make a plot

Let’s make a plot now to examine the relationship between flipper length and body mass.

ggplot(  
 data = penguins\_v1,  
 mapping = aes(x = flipper\_length\_mm, y = body\_mass\_g)  
) +  
 geom\_point(aes(color = species, shape = species)) +  
 geom\_smooth(method = "lm") +  
 labs(  
 title = "Body mass and flipper length",  
 subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",  
 x = "Flipper length (mm)", y = "Body mass (g)",  
 color = "Species", shape = "Species"  
 ) +  
 scale\_color\_viridis\_d() +  
 theme\_minimal()

## Warning: Removed 1 rows containing non-finite values (`stat\_smooth()`).

## Warning: Removed 1 rows containing missing values (`geom\_point()`).

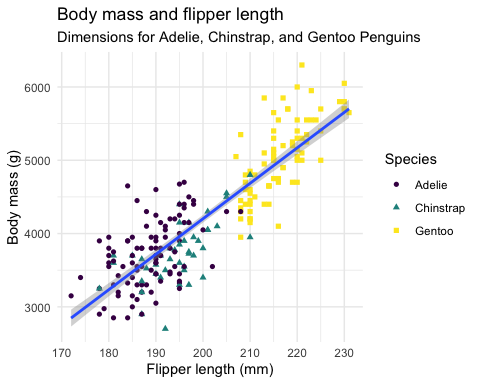


Figure 1: Relationship between flipper length and body mass.

# Updating a plot

But wait, it looks like we’re getting some warning messsages — let’s check out the source of those.

penguins\_v1 %>%   
 filter(is.na(flipper\_length\_mm))

## # A tibble: 1 × 9  
## id species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 94 Adelie Torgersen NA NA NA  
## # ℹ 3 more variables: body\_mass\_g <dbl>, sex <chr>, year <dbl>

It looks like penguin 94 is missing several measurements. You contact the scientist who supplied the data and they find the missing information in a notebook and send you an updated data file penguins\_v2.csv.

Let’s load in the new data.

penguins\_v2 <- read\_csv(here('data','raw','penguins\_v2.csv'))

Let’s check if the data for penguin 94 has been updated…

penguins\_v2 %>%   
 filter(id == 94)

## # A tibble: 1 × 9  
## id species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 94 Adelie Torgersen 38.1 20 191  
## # ℹ 3 more variables: body\_mass\_g <dbl>, sex <chr>, year <dbl>

It has — great!

Let’s try and make the plot again…

ggplot(  
 data = penguins\_v2,  
 mapping = aes(x = flipper\_length\_mm, y = body\_mass\_g)  
) +  
 geom\_point(aes(color = species, shape = species)) +  
 geom\_smooth(method = "lm") +  
 labs(  
 title = "Body mass and flipper length",  
 subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",  
 x = "Flipper length (mm)", y = "Body mass (g)",  
 color = "Species", shape = "Species"  
 ) +  
 scale\_color\_viridis\_d() +  
 theme\_minimal()

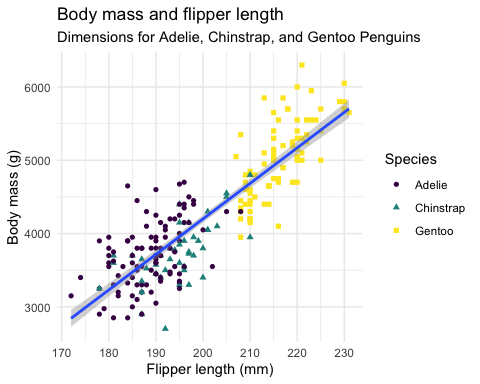


Figure 2: Relationship between flipper length and body mass with missing data corrected.

No warnings this time!

# Some new data comes in

Ok let’s look at a couple of other scenarios where we might want to update our analyses. The original penguin data was collected in 2007 and 2008. In 2009, your client tells you that some new penguin data come in and wants you to updated the analyses. Fortunately, we don’t have to rewrite everything from scratch because we have a reproducible analysis script!

So let’s load in the new data penguins\_v3.

penguins\_v3 <- read\_csv(here('data','raw','penguins\_v3.csv'))

We can make a table to see how many penguins we have each year…

penguins\_v3 %>%   
 count(year) %>%  
 kable(caption = "Number of penguins by year")

Table 1: Number of penguins by year

| year | n |
| --- | --- |
| 2007 | 110 |
| 2008 | 114 |
| 2009 | 120 |

Great! Now we have an additional 120 penguins for the year 2009.

Let’s make our plot again with the new data…

ggplot(  
 data = penguins\_v3,  
 mapping = aes(x = flipper\_length\_mm, y = body\_mass\_g)  
) +  
 geom\_point(aes(color = species, shape = species)) +  
 geom\_smooth(method = "lm") +  
 labs(  
 title = "Body mass and flipper length",  
 subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",  
 x = "Flipper length (mm)", y = "Body mass (g)",  
 color = "Species", shape = "Species"  
 ) +  
 scale\_color\_viridis\_d() +  
 theme\_minimal()

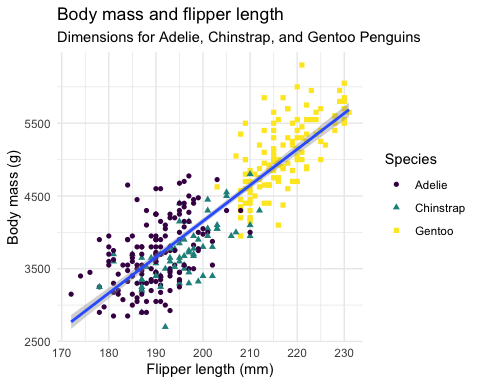


Figure 3: Relationship between flipper length and body mass with new data.

Excellent! In R Markdown, we don’t have to keep rewriting the analysis from scratch (which is inefficient and error prone). Instead we can just reuse our old code and update specific aspects that we want to change.

# Another request comes in…

Ok let’s look at one more scenario. You present your findings to your client. They are very impressed, but they have a question — does the relationship between flipper length and body mass vary between the three different islands that the data were collected from?

Fortunately, all we need to do is add a single line of code to our existing plot: facet\_wrap(. ~island) to stratify the results by island.

ggplot(  
 data = penguins\_v3,  
 mapping = aes(x = flipper\_length\_mm, y = body\_mass\_g)  
) +  
 facet\_wrap(. ~island) +  
 geom\_point(aes(color = species, shape = species)) +  
 geom\_smooth(method = "lm") +  
 labs(  
 title = "Body mass and flipper length",  
 subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",  
 x = "Flipper length (mm)", y = "Body mass (g)",  
 color = "Species", shape = "Species"  
 ) +  
 scale\_color\_viridis\_d() +  
 theme\_minimal()

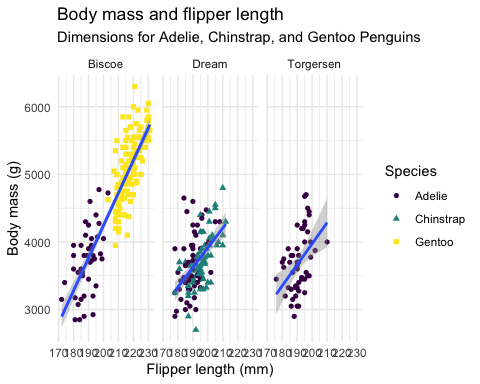


Figure 4: Relationship between flipper length and body mass stratified by island.

Great! You’ve learned how to update your analyses and generate a reproducible research report!

Let’s look at a few other nifty features of R Markdown…

# Citations

In our research reports, we often want to add citations. We can do this in R Markdown too!

For example, we can add a citation to Hadley Wickham’s paper about tidy data (Wickham 2014).

There’s also a helpful plugin for R Studio called citr. This allows you to search your Zotero library for references from within R Studio, which saves a lot of time.

# devtools::install\_github("crsh/citr")

# Cross-references

We can also add cross-references to tables and figures that automatically update their numbering if we change the order. For example, above we initially created Figure 1. We then corrected the missing data problem (Figure 2), and added in some new data (Figure 3). Finally, we made a plot that stratified the results by island (Figure 4).

We can also cross-reference to tables, for example Table 1.

# Inline code

As well as code chunks, we can embed short pieces of code directly into our text. This is helpful when we want to ensure that numbers embedded in text are reproducible. For example, when the sample size of penguins\_v1 is 224 and the sample size of penguins\_v3 is 344. If you take a look at the source code of this document, you can see that those numbers were generated by code.

# Setting the seed

How do we ensure reproducibility when our analyses involve randomness? We can use a function called set.seed() to ensure that such computations are repeatable, as shown below.

# this will generate different outputs  
sample(1:10, 5)

## [1] 4 7 5 3 10

sample(1:10, 5)

## [1] 2 4 7 9 6

# to get the same output, we need to set the seed each time  
set.seed(42)  
sample(1:10, 5)

## [1] 1 5 10 8 2

set.seed(42)  
sample(1:10, 5)

## [1] 1 5 10 8 2

# References

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. 2014. “Ecological Sexual Dimorphism and Environmental Variability Within a Community of Antarctic Penguins (Genus Pygoscelis).” Edited by André Chiaradia. *PLoS ONE* 9 (3): e90081. <https://doi.org/10.1371/journal.pone.0090081>.

Wickham, Hadley. 2014. “Tidy Data.” *Journal of Statistical Software* 59 (10). <https://doi.org/10.18637/jss.v059.i10>.