Answer Sketch and Rubric for Lab 04

431 Staff and Professor Love

Last Edited 2020-10-04 09:14:30

Contents

0.1 R Setup

Here's the complete R setup we used to build this answer sketch.

```
knitr::opts_chunk$set(comment=NA)
options(width = 60)

library(palmerpenguins)
library(broom)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)
## make sure these packages are installed in R

theme_set(theme_bw())
```

0.2 A Note on Tidyverse Conflicts

There are some packages we occasionally load (like `Hmisc` or `mosaic`) that have functions which conflict with the `tidyverse`. We can identify these issues directly with the `tidyverse_conflicted()` function.

```
tidyverse_conflicts()
```

```
-- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

Note that because we've loaded the `tidyverse` package last, functions like `filter` use the tidyverse (specifically the `dplyr` version) rather than the `stats` package's version.

Things behave as we expect them to. For instance, suppose we get some elementary summaries of the bill lengths by island with:

```
penguins %>% filter(complete.cases(bill_length_mm)) %>%
  group_by(island) %>%
  summarize(n = n(), median = median(bill_length_mm))
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
# A tibble: 3 x 3
  island        n median
  <fct>     <int>  <dbl>
1 Biscoe      167   45.8
2 Dream       124   44.7
3 Torgersen    51   38.9
```

No problem. Now, suppose we run `describe` from the `Hmisc` package without loading the package:

```
penguins %$% Hmisc::describe(bill_length_mm)
```

```
bill_length_mm
       n  missing distinct     Info     Mean      Gmd
     342        2      164        1    43.92    6.274
     .05      .10      .25      .50      .75      .90
   35.70    36.60    39.23    44.45    48.50    50.80
     .95
   51.99

lowest : 32.1 33.1 33.5 34.0 34.1, highest: 55.1 55.8 55.9 58.0 59.6
```

Now, suppose we check our conflicts...

```
tidyverse_conflicts()
```

```
-- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
```

```
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

No new problems. That's fine.

But suppose we actually load the Hmisc package at this stage (in other words, after we've loaded the tidyverse)…

```r
library(Hmisc)
```

```
Loading required package: lattice

Loading required package: survival

Loading required package: Formula


Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

Note the warnings.

Now, if we run our conflicts, we see that the `summarize()` function from `Hmisc` now supercedes the `dplyr` version, which will cause us problems.

```r
tidyverse_conflicts()
```

```
-- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
x Hmisc::src()       masks dplyr::src()
x Hmisc::summarize() masks dplyr::summarize()
```

So, now, if we try

```r
penguins %>% filter(complete.cases(bill_length_mm)) %>%
  group_by(island) %>%
  summarize(n = n(), median = median(bill_length_mm))
```

we will throw an error which says `Error in summarize(., n = n(), median = median(bill_length_mm)) : argument "by" is missing, with no default` which is actually telling us that the machine is using the summarize from `Hmisc` (for which this is a requirement) rather than the one from `dplyr` (for which this is accomplished through `group_by`)

How can we fix this?

1. Don't load `Hmisc`.
2. If you do load it, then you'll need to specify which `summarize` you want, with `dplyr::summarize`.

```r
penguins %>% filter(complete.cases(bill_length_mm)) %>%
  group_by(island) %>%
  dplyr::summarize(n = n(), median = median(bill_length_mm))
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
# A tibble: 3 x 3
  island        n median
  <fct>     <int>  <dbl>
1 Biscoe      167   45.8
2 Dream       124   44.7
3 Torgersen    51   38.9
```

The most common thing we're seeing people do at this point is reload the `dplyr` package.

```
library(dplyr)

tidyverse_conflicts()
```

```
-- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
x Hmisc::src()       masks dplyr::src()
x Hmisc::summarize() masks dplyr::summarize()
```

which doesn't fix the problem, or at least doesn't always fix the problem.

You could try

```
detach(name=package:Hmisc)

tidyverse_conflicts()
```

```
-- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

which works, but can cause other issues later.

0.3 Where do we see most of the tidyverse conflicts?

- when you load (or partially load) `Hmisc` or `rms` (which requires `Hmisc`)
- when you load `MASS` which has a version of `select` that conflicts with the tidyverse's `select`
- when you load `car` which masks `dplyr`'s version of `recode`.

The best solution is usually to load these packages prior to loading the tidyverse.

Thanks.

0.4 Note on Sketch for Questions 1-5

Because of the nature of Questions 1-5 in this Lab, this part is just a grading rubric. I hope you find it helpful. Note that the maximum score on this lab is 105 points, but we will treat it as if it was graded out of 100 points.

1 Question 1 (5 points)

> Specify the URL where we can see the headline and news story describing the findings of the study. You should provide a complete reference, including the names of the author(s) of the news story, and its full title, and source.

We accept any working reference that included the specified information.

1.1 Grading Notes

- Award 5 points if they have a working link that meets the specifications.
- Award no more than 2 points if they don't have the author names, title and source, or if the link doesn't work.

2 Question 2 (5 points)

> Specify a URL where we can see at least the abstract of the complete study. Again, provide a complete reference to the study, too.

We would accept any working reference that included the specified information.

2.1 Grading Notes

- Award 5 points if they meet specifications.
- Award no more than 2 points if they don't have the authors, title, journal name and so forth, or if the link doesn't work.

3 Question 3 (10 points)

> Describe your opinion (gut feeling) related to the conclusions of the study as summarized in the headline and news article, first in terms of a probability statement, and then calculate the appropriate odds. Motivate your internal prior probability, describing your relevant personal experiences or other factors that drove your gut feeling.*

We want to see an accurate calculation given your probability, and at least a reasonable attempt at motivation for your initial probability.

3.1 Grading Notes

Full 10 points awarded requires:

- Probability statement relating to gut feeling
- Calculate odds related to gut feeling
- Motivation/explanation of gut feeling
- Complete sentences with correct spelling and grammar
- Lose 3 points for each missing or problematic piece, down to a minimum of 0

4 Question 4 (20 points)

> Evaluate the study in terms of the six specifications proposed by Leek when evaluating study support.

We want to see a clear, motivated conclusion about each of the six specifications (each being worth 5 points), as well as direct quotes and evidence summaries to address the issues raised and justify conclusions.

The six specifications we are looking for are:

1. Was the study a clinical study in humans?
2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
3. Was the study a randomized, controlled trial (RCT)?
4. Was it a large study - at least hundreds of patients?
5. Did the treatment have a major impact on the outcome?
6. Did predictions hold up in at least two separate groups of people?

4.1 Grading Notes

Receive 3 points for each of the 6 parts listed above should be awarded for complete sentences with correct spelling and grammar, using quotations or paraphrasing the complete study appropriately, plus a bonus of an additional 2 points if they all 6 pieces get 3 points.

- They should score no higher than 1.5/3 for any part where they don't successfully generate a good response.

5 Question 5 (20 points)

> Incorporate the study support assessment into a Bayes' Rule calculation to obtain the final odds you should now be willing to give to the headline, and specify this value in terms of a probability statement, as well. Then react to the final conclusion specified by this approach in a few sentences. How does your subjective posterior probability that the headline is true match up with the formula's conclusions? Do you feel that the formulaic approach has yielded an appropriate conclusion for you in this case? Why or why not?

Grading here is broken up into two pieces.

First, up to 8 points are awarded for the calculations.

First, we wanted to see correct calculations of both the odds and probability in light of the prior probability established in Question 3 and the answers to the specifications described in Question 4. For 8 points they should provide **odds** and **probability** using Bayes' Rule, responding with **complete sentences with correct spelling and grammar**, losing 4 points for each missing item in bold, down to a minimum of 0.

Next, up to 12 points are awarded for the reaction.

For the additional 12 points, we then wanted to see you specify your subjective feeling about what the probability *should* be, and then match that up with the result of the calculation.

The student should reach a logical conclusion and summarize his/her reaction on whether or not this approach was accurate for assessing their initial gut feeling vs. their final conclusion after reading the article/study. No right or wrong answer, but using both logic and complete sentences are necessary.

- The default grade on this question is 10/12 points, which should be the grade for someone who uses logic and complete sentences, and meets the minimum standard for a complete answer.

- Somewhere between 8 and 11 points out of 12 should be awarded to most students for a good effort here.
- Give 7 or less if they don't manage to do all of these pieces.
- Give 12 points only to the top 6-8 responses, across all students.
- My prior assumption is that a moderate number of students will have a perfect score on Lab 04 for Questions 1-4, so we can grade a little tougher on Question 5.

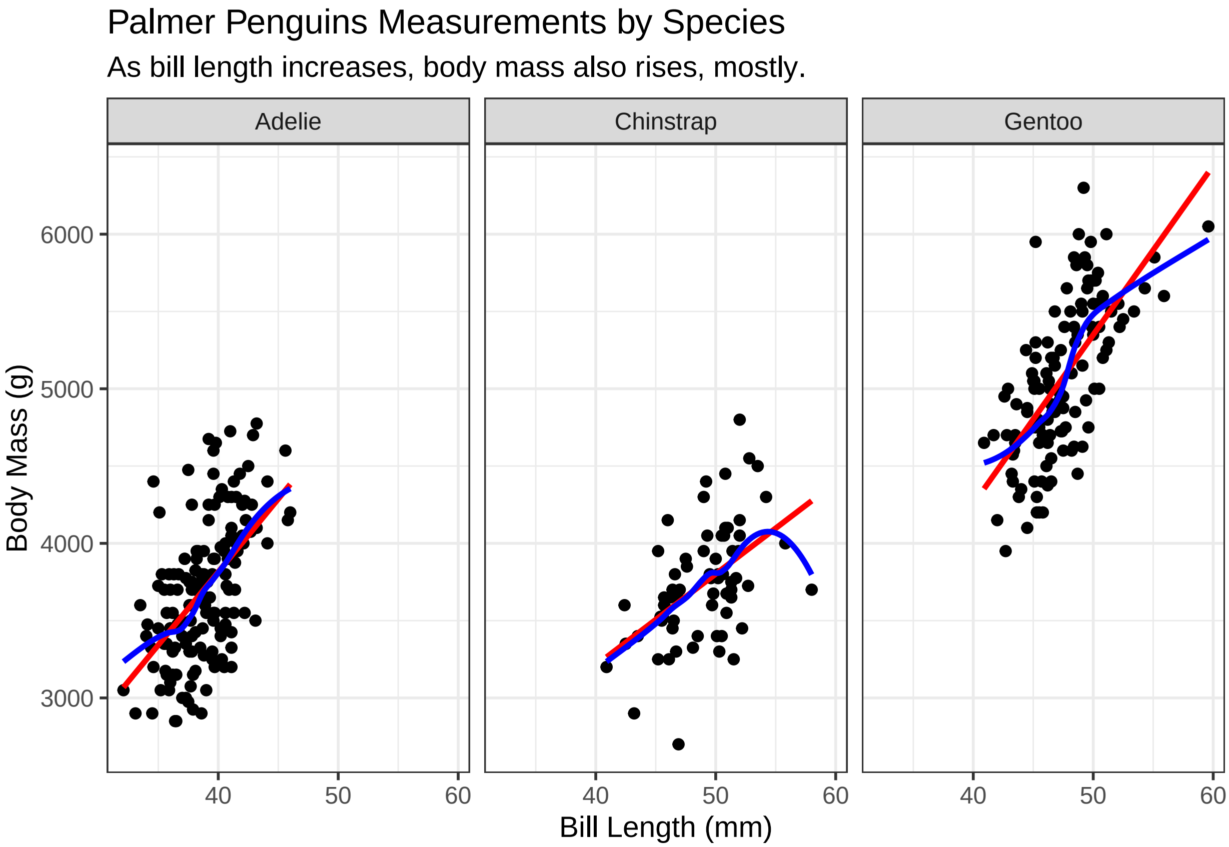
6 Question 6 (10 points, but see note below)

Suppose we are interested in which of the three species of penguin (Adelie, Chinstrap or Gentoo) shows the strongest linear relationship between bill length and body mass, using all 342 penguins with complete data on the two variables we're studying (bill length and body mass.)

Draw an attractive and thoughtfully labeled plot (including a title) to show the association of bill length (placed on the horizontal x axis) and body mass (on the vertical y axis) for these 342 penguins, and facet the plot by species (so that one facet shows the 151 Adelie, one shows the 68 Chinstrap and one shows the 123 Gentoo.) Use `geom_smooth()` to add an appropriate smooth curve (using `lm` or `loess` or both, as you prefer) to each facet of the plot.

Then write a sentence or two describing what your plot reveals.

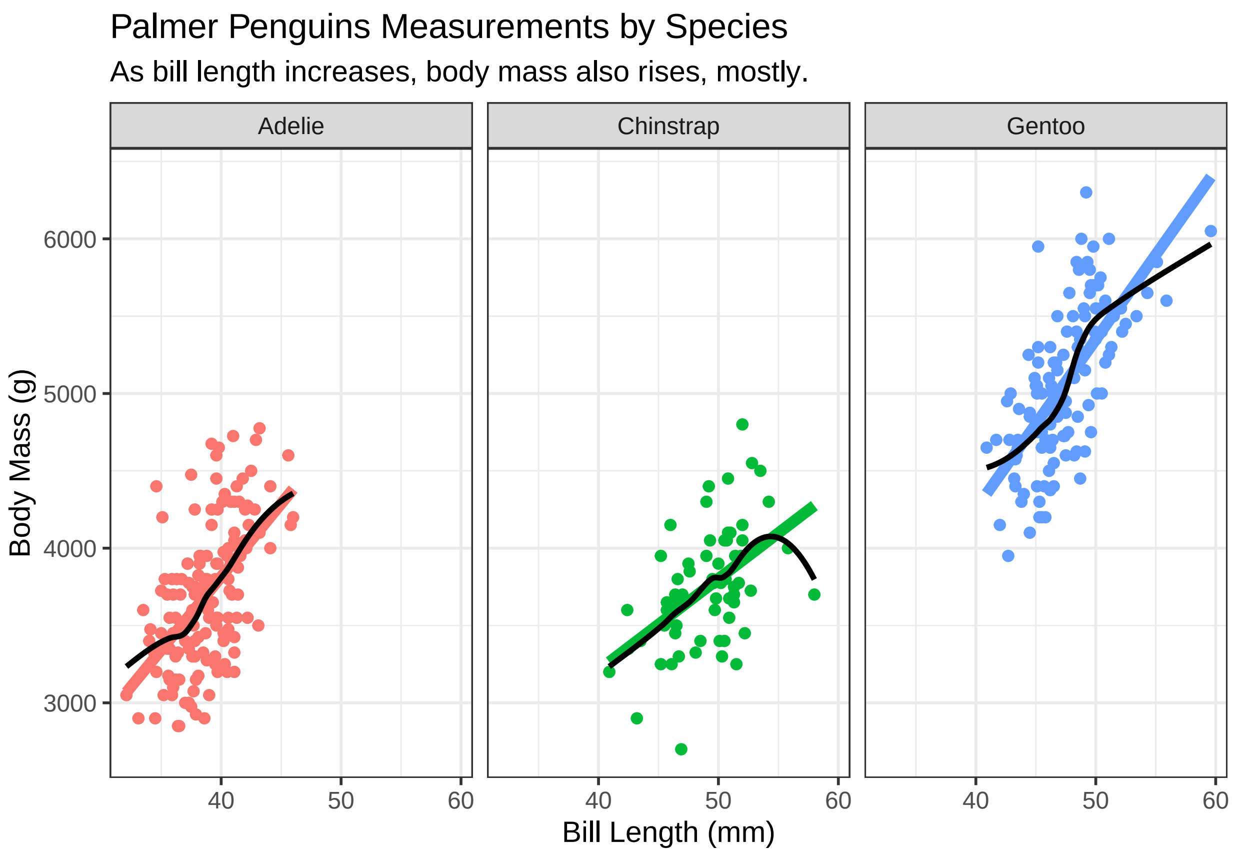
```r
lab04_penguins <- penguins %>%
    filter(complete.cases(bill_length_mm, body_mass_g))

ggplot(lab04_penguins, aes(x = bill_length_mm, y = body_mass_g)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, formula = y ~ x, col = "red") +
    geom_smooth(method = "loess", se = FALSE, formula = y ~ x, col = "blue") +
    facet_wrap(~ species) +
    labs(title = "Palmer Penguins Measurements by Species",
         subtitle = "As bill length increases, body mass also rises, mostly.",
         x = "Bill Length (mm)", y = "Body Mass (g)")
```



The primary conclusion here is that as bill length increases, body mass also rises. It is also worth noting that the sizes of the animals varies substantially by species, with Gentoo having the largest body mass, and generally larger bill lengths than the other two species, while Adelie definitely have the shortest bill lengths.

Here's a picture that changes the color of the points and regression lines (but not the smooths) within each facet.

```
ggplot(lab04_penguins, aes(x = bill_length_mm, y = body_mass_g,
                           group = species, col = species)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, formula = y ~ x, lwd = 2) +
    geom_smooth(method = "loess", se = FALSE, formula = y ~ x, col = "black") +
    facet_wrap(~ species) +
    guides(color = FALSE) +
    labs(title = "Palmer Penguins Measurements by Species",
         subtitle = "As bill length increases, body mass also rises, mostly.",
         x = "Bill Length (mm)", y = "Body Mass (g)")
```



Note that I was able to develop a complete response here just with the picture. Calculating a summary statistic, like the Pearson correlation or R^2 was not part of this question, although if you want to look at these values, you could…

```
lab04_penguins %>% group_by(species) %>%
    dplyr::summarize(r = cor(bill_length_mm, body_mass_g), r_square = r^2)
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
# A tibble: 3 x 3
  species       r r_square
  <fct>     <dbl>    <dbl>
1 Adelie    0.549    0.301
2 Chinstrap 0.514    0.264
3 Gentoo    0.669    0.448
```

6.1 Grading Question 6

Note that Question 6 was graded out of 10 points, even though it was originally intended to be graded out of 15 points. In computing your final score on the Quiz, you received 5 additional points if your R Markdown and HTML files were in and readable, on time, so that the total possible score on the Quiz was actually 105, although we'll still treat it as a 100 point Quiz for averaging purposes at the end of the term.

We award up to 5 points for appropriate code, and a correct plot (if they include either the `lm` or the `loess` smooth that is fine.)

We then award up to an additional 5 points for the a sentence (or two) describing the conclusions (they should mention the direction of the relationships *and* the differences in size between the species for full credit.)

We didn't worry in this section if you left the missing values in and generated the resulting warning, although you'll need to fix that for Question 7, and you really should fix it here.

7 Question 7 (15 points)

> Build a linear model predicting body mass using bill length for the species which has the largest body mass (on average) across the three species. Then use that equation to estimate the difference in the predicted body mass for two new penguins of that species (named Pingu and Skipper), where Pingu whom has a bill length at the 75th percentile of the original data for that species, and Skipper has a bill length at the 25th percentile of the original data for that species? In a summary sentence or two describing your results, be sure to specify which species you are studying and which of the two new penguins (Pingu or Skipper) would be expected to have a larger body mass (and by how much), according to your model.

Again, note that my `lab04_penguins` file has already removed the missing values.

The linear model will be built for the 123 Gentoo penguins, as they have the largest body mass, as we can see from the plot above, or table below.

```r
mosaic::favstats(body_mass_g ~ species, data = lab04_penguins) %>%
    kable(digits = 1)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

species	min	Q1	median	Q3	max	mean	sd	n	missing
Adelie	2850	3350.0	3700	4000	4775	3700.7	458.6	151	0
Chinstrap	2700	3487.5	3700	3950	4800	3733.1	384.3	68	0
Gentoo	3950	4700.0	5000	5500	6300	5076.0	504.1	123	0

Here is the resulting model:

```r
lab04_gentoo <- lab04_penguins %>%
    filter(species == "Gentoo")

m7 <- lm(body_mass_g ~ bill_length_mm, data = lab04_gentoo)

tidy(m7) %>% kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-123.83	526.05	-0.24	0.81
bill_length_mm	109.46	11.05	9.91	0.00

Now, we need to make predictions at the appropriate levels of `bill_length_mm` in the Gentoo penguins. First, let's find those quartiles:

```r
mosaic::favstats(~ bill_length_mm, data = lab04_gentoo) %>% kable(digits = 2)
```

min	Q1	median	Q3	max	mean	sd	n	missing
40.9	45.3	47.3	49.55	59.6	47.5	3.08	123	0

So, we need to predict body_mass_g for Pingu, whose bill is 49.55 mm long, and for Skipper, whose bill is 45.3 mm long, using our model m7.

```
newdat7 <- tibble(name = c("Pingu", "Skipper"), bill_length_mm = c(49.55, 45.3))

augment(m7, newdata = newdat7)
```

```
# A tibble: 2 x 3
  name    bill_length_mm .fitted
  <chr>            <dbl>   <dbl>
1 Pingu             49.6   5300.
2 Skipper           45.3   4835.
```

So the predicted body mass for Pingu is 5300 g, and for Skipper it is 4835 g, a difference of **465** grams. (note that our first draft of this sketch had this as 435, incorrectly.)

7.1 Grading Question 7

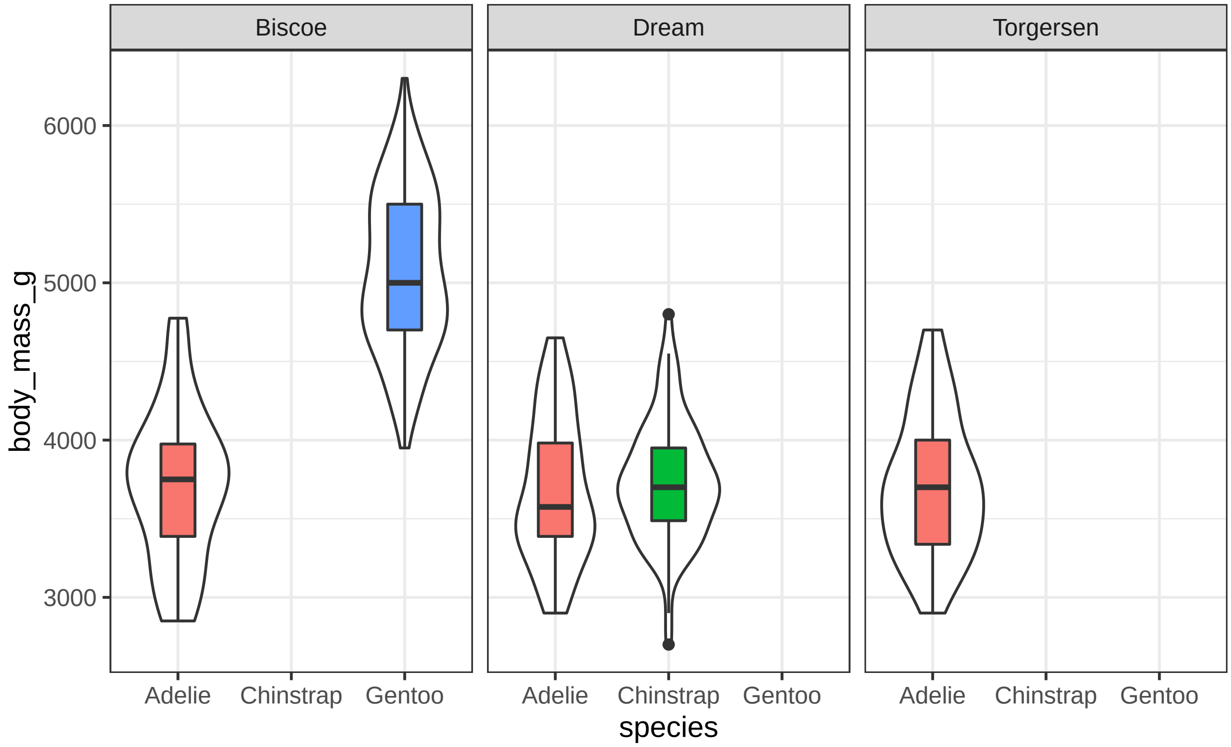
- Up to 5 points for a correct choice of species (Gentoo) and correct regression equation.
- Up to 5 points for working through the data and using whatever equation they come up with to obtain predictions for the correct values for Gentoo penguins at the 1st and 3rd quartiles.
- Up to 5 points for using their predictions (right or wrong) to answer the final question about which species would have a larger body mass and which species they are studying.

8 Question 8 (15 points)

> Now, let's focus just on our outcome, and how it is associated with species and island. Build a plot that shows the distribution of body mass (in grams) within each penguin species, and which also identifies the three islands (through the use of facets) in an attractive way. Your plot should have a meaningful title as well as useful axis and facet labels. Please restrict the plot to the 342 penguins with complete body mass information. Then, in a sentence or two, describe the results from your plot, and any reasonable conclusions you can draw from the plot about the relationships between island and species, and between body mass and the other two variables.

Here's my plot.

```
ggplot(lab04_penguins, aes(x = species, y = body_mass_g)) +
    geom_violin() +
    geom_boxplot(aes(fill = species), width = 0.3) +
    facet_wrap( ~ island) +
    guides(fill = FALSE) +
    labs()
```



A quick count of the combinations of species and island reveals why this plot has some blank spaces.

```
lab04_penguins %>% count(island, species)
```

```
# A tibble: 5 x 3
  island    species       n
  <fct>     <fct>     <int>
1 Biscoe    Adelie       44
2 Biscoe    Gentoo      123
3 Dream     Adelie       56
4 Dream     Chinstrap    68
5 Torgersen Adelie       51
```

An appropriate sentence will focus on the fact that all Gentoo penguins were found on Biscoe, all Chinstrap penguins were found on Dream, and only Adelie penguins were found on all three islands. The Adelie penguins show no particular changes in their distribution of body mass on the basis of what island they were associated with.

8.1 Grading Question 8

We award up to 8 points for appropriate code and a correct plot (which could also have been a facetted set of histograms, I suppose.)

We then award 7 additional points for the sentence (or two) describing the conclusions, which includes realizing that two of the species only appear on one island, and also recognizing that the only reasonable cross-island comparison is within the Adelie penguins.

9 Session Information

```r
sessionInfo()
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19041)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
 [1] ggridges_0.5.2       mosaicData_0.20.1
 [3] ggformula_0.9.4      ggstance_0.3.4
 [5] Matrix_1.2-18        Formula_1.2-3
 [7] survival_3.2-3       lattice_0.20-41
 [9] forcats_0.5.0        stringr_1.4.0
[11] dplyr_1.0.2          purrr_0.3.4
[13] readr_1.3.1          tidyr_1.1.2
[15] tibble_3.0.3         ggplot2_3.3.2
[17] tidyverse_1.3.0      patchwork_1.0.1
[19] magrittr_1.5         knitr_1.30
[21] broom_0.7.1          palmerpenguins_0.1.0

loaded via a namespace (and not attached):
 [1] nlme_3.1-148        fs_1.5.0
 [3] lubridate_1.7.9     RColorBrewer_1.1-2
 [5] httr_1.4.2          tools_4.0.2
 [7] backports_1.1.10    utf8_1.1.4
 [9] R6_2.4.1            rpart_4.1-15
[11] Hmisc_4.4-1         DBI_1.1.0
[13] mgcv_1.8-31         colorspace_1.4-1
[15] nnet_7.3-14         withr_2.3.0
[17] tidyselect_1.1.0    gridExtra_2.3
[19] leaflet_2.0.3       compiler_4.0.2
[21] cli_2.0.2           rvest_0.3.6
[23] htmlTable_2.1.0     xml2_1.3.2
[25] ggdendro_0.1.22     labeling_0.3
[27] mosaicCore_0.8.0    scales_1.1.1
[29] checkmate_2.0.0     digest_0.6.25
[31] foreign_0.8-80      rmarkdown_2.4
[33] base64enc_0.1-3     jpeg_0.1-8.1
[35] pkgconfig_2.0.3     htmltools_0.5.0
```

```
[37] highr_0.8           dbplyr_1.4.4
[39] htmlwidgets_1.5.1   rlang_0.4.7
[41] readxl_1.3.1        rstudioapi_0.11
[43] generics_0.0.2      farver_2.0.3
[45] jsonlite_1.7.1      crosstalk_1.1.0.1
[47] Rcpp_1.0.5          munsell_0.5.0
[49] fansi_0.4.1         lifecycle_0.2.0
[51] stringi_1.5.3       yaml_2.2.1
[53] MASS_7.3-53         plyr_1.8.6
[55] grid_4.0.2          blob_1.2.1
[57] ggrepel_0.8.2       crayon_1.3.4
[59] haven_2.3.1         splines_4.0.2
[61] hms_0.5.3           pillar_1.4.6
[63] reprex_0.3.0        glue_1.4.2
[65] evaluate_0.14       latticeExtra_0.6-29
[67] data.table_1.13.0   modelr_0.1.8
[69] tweenr_1.0.1        png_0.1-7
[71] vctrs_0.3.4         cellranger_1.1.0
[73] gtable_0.3.0        polyclip_1.10-0
[75] assertthat_0.2.1    xfun_0.17
[77] ggforce_0.3.2       cluster_2.1.0
[79] mosaic_1.8.2        ellipsis_0.3.1
```