Assignment 8

Ellen Bledsoe

2025-03-18

Assignment Details

Purpose

The goal of this assignment is to practice problem decomposition and some best practices in reproducibility .

Task

Write R code to successfully answer each question below.

Criteria for Success

- Code is within the provided code chunks or new code chunks are created where necessary
- Code chunks run without errors
- Code chunks have brief comments indicating which code is answering which part of the question
- Code will be assessed as follows:
  - Produces the correct answer using the requested approach: 100%
  - Generally uses the right approach, but a minor mistake results in an incorrect answer: 90%
  - Attempts to solve the problem and makes some progress using the core concept, but returns the wrong answer and does not demonstrate comfort with the core concept: 50%
  - Answer demonstrates a lack of understanding of the core concept: 0%
- Any questions requiring written answers are answered with sufficient detail

Due Date

March 18 at midnight MST

Assignment Exercises

1. Set-Up (5 pts)

Load in the `tidyverse`

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

2. Portal Data Review (25 points)

Load them into R using `read_csv()`.

- `surveys.csv`
- `species.csv`
- `plots.csv`

a. Create a data frame with only data for the `species_id` `DO`, with the columns `year`, `month`, `day`, `species_id`, and `weight`.
b. Create a data frame with only data for species IDs `PP` and `PB` and for years starting in 1995, with the columns `year`, `species_id`, and `hindfoot_length`, with no null values for `hindfoot_length`.
c. Create a data frame with the average `hindfoot_length` for each `species_id` in each `year` with no null values.
d. Create a data frame with the `year`, `genus`, `species`, `weight` and `plot_type` for all cases where the `genus` is `"Dipodomys"`.
e. Make a scatter plot with `weight` on the x-axis and `hindfoot_length` on the y-axis. Use a `log10` scale on the x-axis. Color the points by `species_id`. Include good axis labels.
f. Make a histogram of weights with a separate subplot for each `species_id`. Do not include species with no weights. Set the `scales` argument in the `facet_wrap()` function to `"free_y"` so that the y-axes can vary. Include good axis labels.
g. (Challenge, optional) Make a plot with histograms of the weights of three species, `PP`, `PB`, and `DM`, colored by `species_id`, with a different facet (i.e., subplot) for each of three `plot_type`'s `Control`, `Long-term Krat Exclosure`, and `Short-term Krat Exclosure`. Include good axis labels and a title for the plot. Export the plot to a `png` file.

```r
surveys <- read_csv('surveys.csv')
species <- read_csv('species.csv')
plots <- read_csv('plots.csv')

# a. Create a data frame with only data for the `species_id` `DO`, with the columns `year`, `month`, `d
print("2a")
```

```
## [1] "2a"
```

```r
surveys %>%
  filter(species_id == "DO") %>%
  select(year, month, day, species_id, weight)
```

```
## # A tibble: 3,027 x 5
##     year month   day species_id weight
##    <dbl> <dbl> <dbl> <chr>       <dbl>
##  1  1977     8    19 DO             52
##  2  1977    10    17 DO             33
##  3  1977    10    17 DO             50
##  4  1977    10    17 DO             48
##  5  1977    10    17 DO             31
##  6  1977    10    18 DO             41
##  7  1977    11    12 DO             44
##  8  1977    11    12 DO             48
##  9  1977    11    14 DO             39
## 10  1977    12    10 DO             40
## # i 3,017 more rows
```

```
# b. Create a data frame with only data for species IDs `PP` and `PB` and for years starting in 1995, w
print("2b")
```

```
## [1] "2b"
```

```
surveys %>%
  filter(species_id == "PP" | species_id == "PB") %>%
  filter(year >= 1995) %>%
  select(year, species_id, hindfoot_length,) %>%
  filter(!is.na(hindfoot_length))
```

```
## # A tibble: 5,150 x 3
##     year species_id hindfoot_length
##    <dbl> <chr>                <dbl>
##  1  1995 PP                      23
##  2  1995 PP                      22
##  3  1995 PP                      22
##  4  1995 PP                      21
##  5  1995 PP                      21
##  6  1995 PP                      20
##  7  1995 PP                      22
##  8  1995 PP                      24
##  9  1995 PP                      22
## 10  1995 PP                      22
## # i 5,140 more rows
```

```
# c. Create a data frame with the average `hindfoot_length` for each `species_id` in each `year` with n
print("2c")
```

```
## [1] "2c"
```

```
surveys %>%
  filter(!is.na(hindfoot_length)) %>%
  group_by(species_id, year) %>%
  summarize(mean_hf = mean(hindfoot_length))
```

```
## # A tibble: 340 x 3
## # Groups:   species_id [25]
##    species_id  year mean_hf
##    <chr>      <dbl>   <dbl>
##  1 AH          1999    35
##  2 AH          2000    31
##  3 BA          1989    13
##  4 BA          1990    13.8
##  5 BA          1991    12.9
##  6 BA          1992    12
##  7 DM          1977    35.7
##  8 DM          1978    36.1
##  9 DM          1979    35.9
## 10 DM          1980    35.8
## # i 330 more rows
```

```
# d. Create a data frame with the `year`, `genus`, `species`, `weight` and `plot_type` for all cases wh
print("2d")
```

```
## [1] "2d"
```

```
surveys %>%
  inner_join(species, by = "species_id") %>%
  inner_join(plots, by = "plot_id") %>%
  select(year, genus, species, weight, plot_type) %>%
  filter(genus == "Dipodomys")
```

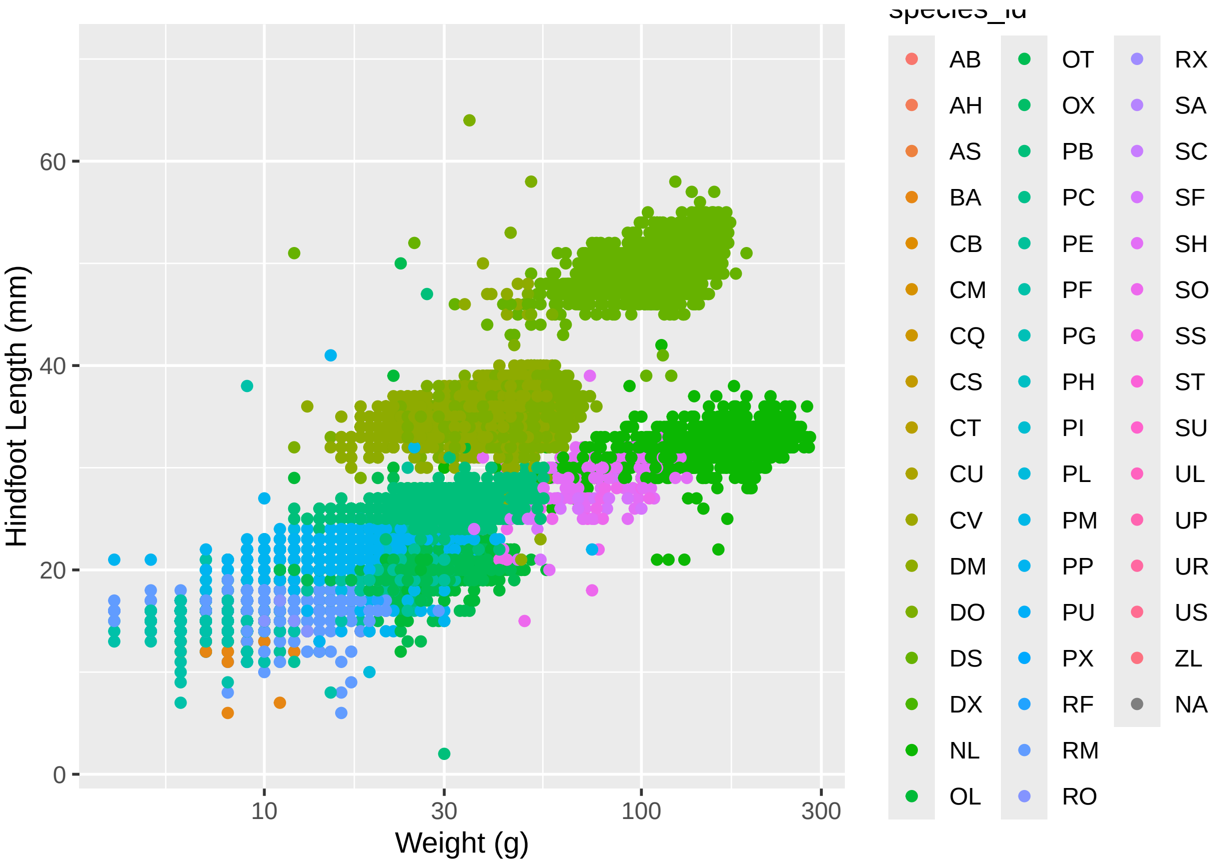
```
## # A tibble: 16,167 x 5
##     year genus     species     weight plot_type
##    <dbl> <chr>     <chr>        <dbl> <chr>
##  1  1977 Dipodomys merriami        NA Control
##  2  1977 Dipodomys merriami        NA Rodent Exclosure
##  3  1977 Dipodomys merriami        NA Long-term Krat Exclosure
##  4  1977 Dipodomys merriami        NA Spectab exclosure
##  5  1977 Dipodomys merriami        NA Spectab exclosure
##  6  1977 Dipodomys spectabilis     NA Rodent Exclosure
##  7  1977 Dipodomys merriami        NA Rodent Exclosure
##  8  1977 Dipodomys merriami        NA Long-term Krat Exclosure
##  9  1977 Dipodomys merriami        NA Control
## 10  1977 Dipodomys merriami        NA Short-term Krat Exclosure
## # i 16,157 more rows
```

```
# e. Make a scatter plot with `hindfoot_length` on the x-axis and `weight` on the y-axis. Color the poi
print("2e")
```

```
## [1] "2e"
```

```
ggplot(data = surveys, mapping = aes(x = weight, y = hindfoot_length, color = species_id)) +
  geom_point() +
  scale_x_log10() +
  labs(x = "Weight (g)", y = "Hindfoot Length (mm)")
```

```
## Warning: Removed 4811 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



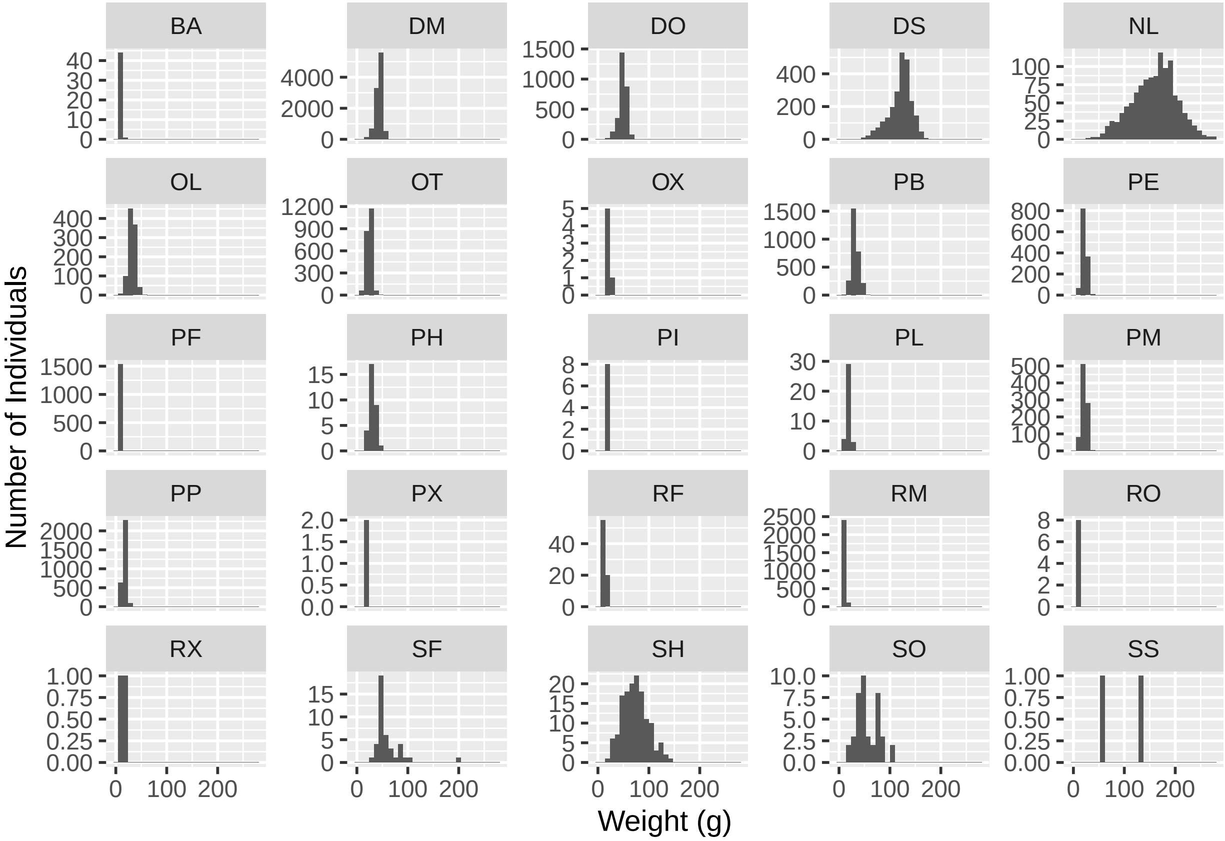
```r
# f. Make a histogram of weights with a separate subplot for each `species_id`.
# Do not include species with no weights.
# Set the `scales` argument to `"free_y"` so that the y-axes can vary.
# Include good axis labels.
print("2f")
```

```
## [1] "2f"
```

```r
surveys_with_weights <- filter(surveys, !is.na(weight))
surveys_with_weights
```

```
## # A tibble: 32,283 x 9
##    record_id month   day  year plot_id species_id sex   hindfoot_length weight
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>  <dbl>
##  1        63     8    19  1977       3 DM         M                  35     40
##  2        64     8    19  1977       7 DM         M                  37     48
##  3        65     8    19  1977       4 DM         F                  34     29
##  4        66     8    19  1977       4 DM         F                  35     46
##  5        67     8    19  1977       7 DM         M                  35     36
##  6        68     8    19  1977       8 DO         F                  32     52
##  7        69     8    19  1977       2 PF         M                  15      8
##  8        70     8    19  1977       3 OX         F                  21     22
##  9         71      8     19  1977        7 DM          F                    36      35
## 10         74      8     19  1977        8 PF          M                    12       7
## # i 32,273 more rows
```

```
ggplot(data = surveys_with_weights, mapping = aes(x = weight)) +
  geom_histogram() +
  facet_wrap(~species_id, scales = "free_y") +
  labs(x = "Weight (g)", y = "Number of Individuals")
```



```
# g. (Challenge) Make a plot with histograms of the weights of three species, `PP`, `PB`, and `DM`, wit
print("2g")
```

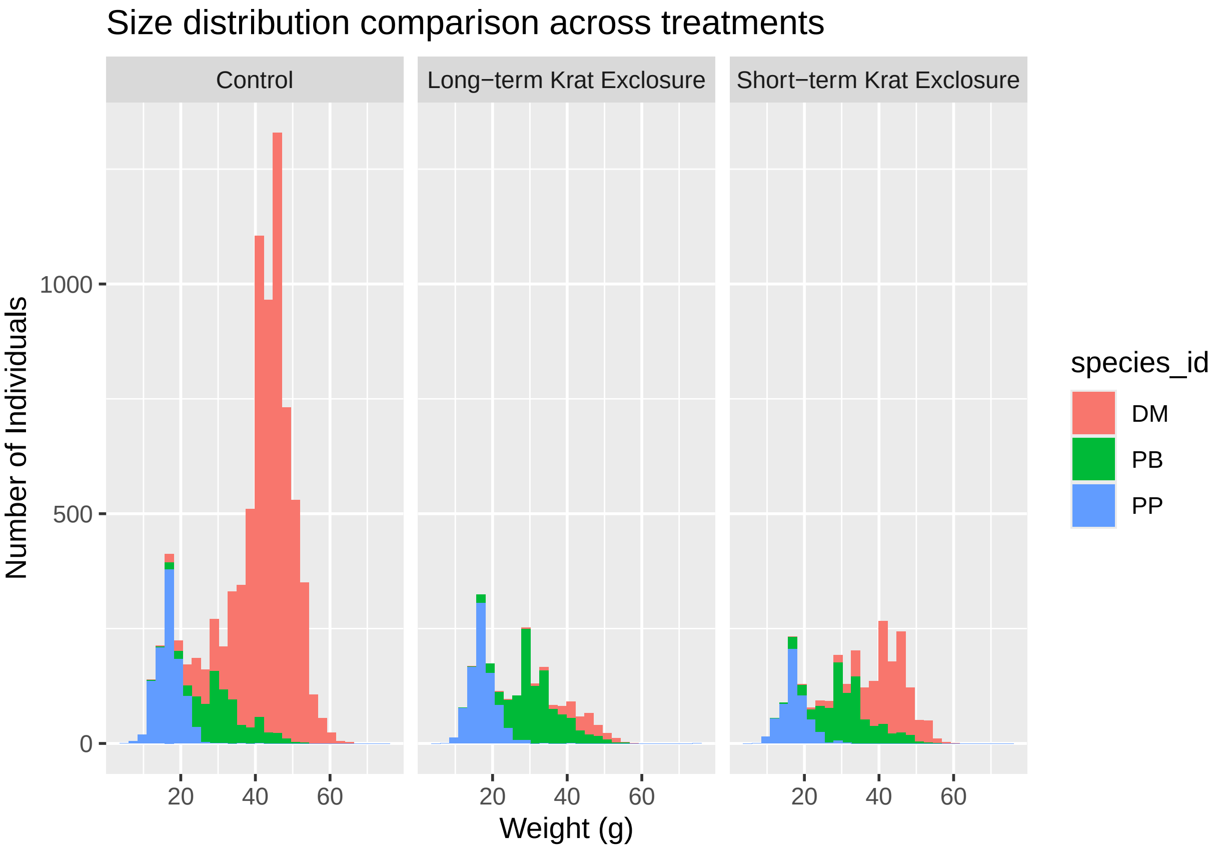
```
## [1] "2g"
```

```
plot_data <- surveys %>%
  inner_join(plots) %>%
  filter(species_id == "PP" | species_id == "PB" | species_id == "DM") %>%
  filter(plot_type == "Control" | plot_type == "Long-term Krat Exclosure" | plot_type == "Short-term Kra
plot_data
```

```
## # A tibble: 13,415 x 10
##    record_id month   day  year plot_id species_id sex   hindfoot_length weight
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>  <dbl>
##  1         3     7    16  1977       2 DM         F                  37     NA
##  2         5     7    16  1977       3 DM         M                  35     NA
##   3       13      7    16  1977       3 DM         M                  35      NA
##   4       14      7    16  1977       8 DM         <NA>               NA      NA
##   5       15      7    16  1977       6 DM         F                  36      NA
##   6       16      7    16  1977       4 DM         F                  36      NA
##   7       18      7    16  1977       2 PP         M                  22      NA
##   8       21      7    17  1977      14 DM         F                  34      NA
##   9       23      7    17  1977      13 DM         M                  36      NA
## 10       26      7    17  1977      15 DM         M                  31      NA
## # i 13,405 more rows
## # i 1 more variable: plot_type <chr>
```

```
ggplot(data = plot_data, aes(x = weight, fill = species_id)) +
  geom_histogram() +
  facet_wrap(~plot_type) +
  labs(x = "Weight (g)", y = "Number of Individuals", title = "Size distribution comparison across trea
```

```
## Warning: Removed 438 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



3. Megafaunal Extinction (35 points)

There were a relatively large number of extinctions of mammalian species roughly 10,000 years ago. To help understand why these extinctions happened scientists are interested in understanding if there were differences in the size of the species that went extinct and those that did not. You are going to reproduce the three main figures from one of the major papers on this topic Lyons et al. 2004.

You will do this using a large dataset of mammalian body sizes that has data on the mass of recently extinct mammals as well as extant mammals (i.e., those that are still alive today).

a. Import the data into R. As with most real world data there are a some things about the dataset that you'll need to identify and address during the import process. Print out the structure of the resulting data frame.
b. Create a plot showing histograms of masses for mammal species that are still present and those that went extinct during the pleistocene (`extant` and `extinct` in the `status` column). There should be one sub-plot for each continent and that sub-plot should show the histograms for both groups as a stacked histogram. To match the original analysis don't include islands (`Insular` and `Oceanic` in the `continent` column) and or the continent labeled `EA` (because `EA` had no species that went extinct in the pleistocene). Scale the x-axis logarithmically and use 25 bins to roughly match the original figure. Use good axis labels.
c. The 2nd figure in the original paper looks in more detail at two orders, *Xenarthra* and *Carnivora*, which showed extinctions in North and South America. Create a figure similar to the one in Part 2, but that shows 4 sub-plots, one for each order on each of the two continents. Still scale the x-axis logarithmically, but use 19 bins to roughly match the original figure.
d. The 3rd figure in the original paper explores Australia as a case study. Australia is interesting because there is good data on both Pleistocene extinctions (`extinct` in the `status` column) and more modern extinctions occurring over the last 300 years (`historical` in the `status` column). Make single stacked histogram that compares the sizes of `extinct`, `extant`, and `historical` statuses. Scale the x-axis logarithmically and use 25 bins to roughly match the original figure. Use good axis labels.
e. (Challenge, optional) Instead of excluding continent `EA` by name in your analysis (in part 2), modify your code to determine from the data which continents had species that went extinct in the pleistocene and only include those continents.

```r
print("3a")
```

```
## [1] "3a"
```

```r
mammal_sizes <- read_tsv("mammal-size-data-clean.txt", na = c("-999"))
str(mammal_sizes)
```

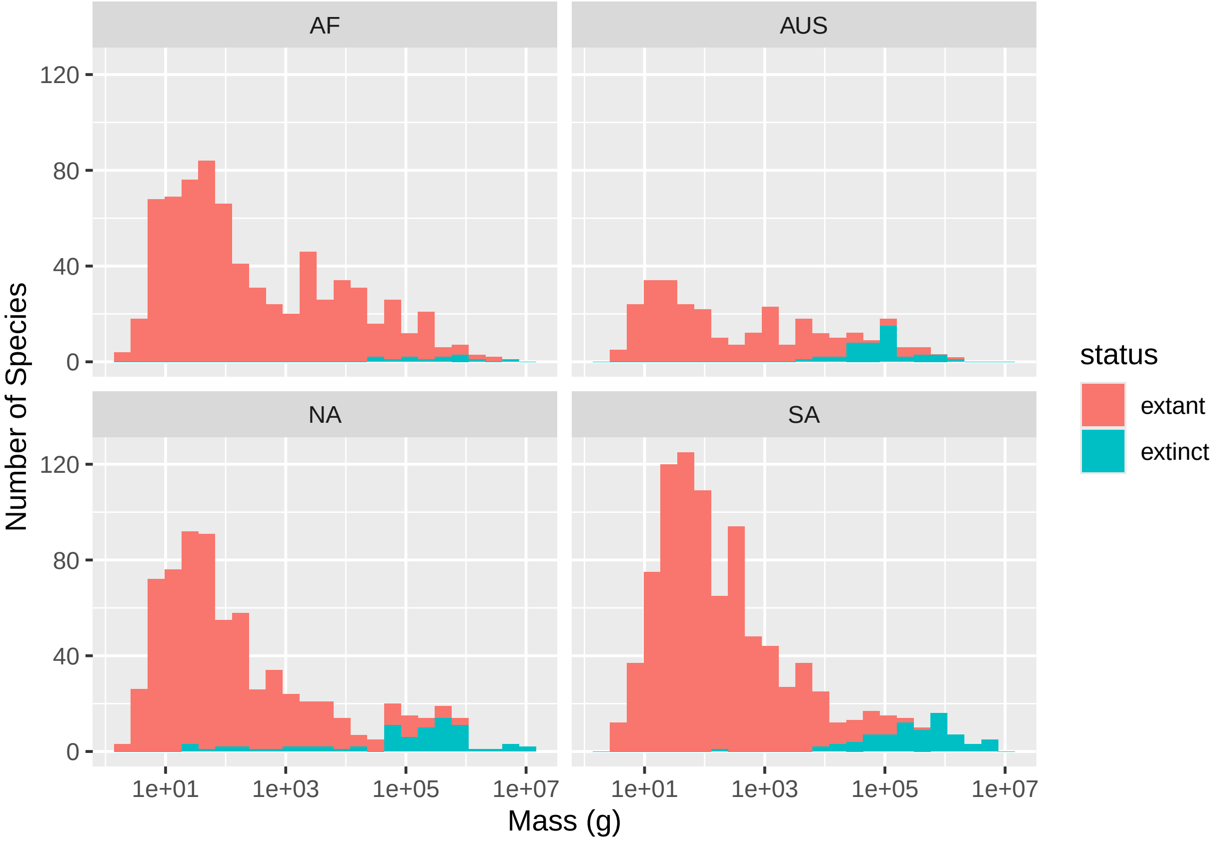
```
## spc_tbl_ [5,731 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ continent: chr [1:5731] "AF" "AF" "AF" "AF" ...
##  $ status   : chr [1:5731] "extant" "extant" "extant" "extant" ...
##  $ order    : chr [1:5731] "Artiodactyla" "Artiodactyla" "Artiodactyla" "Artiodactyla" ...
##  $ family   : chr [1:5731] "Bovidae" "Bovidae" "Bovidae" "Bovidae" ...
##  $ genus    : chr [1:5731] "Addax" "Aepyceros" "Alcelaphus" "Ammodorcas" ...
##  $ species  : chr [1:5731] "nasomaculatus" "melampus" "buselaphus" "clarkei" ...
##  $ mass     : num [1:5731] 70000 52500 171002 28050 48000 ...
##  $ reference: chr [1:5731] "60" "63, 70" "63, 70" "60" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   continent = col_character(),
##   ..   status = col_character(),
##   ..   order = col_character(),
##   ..   family = col_character(),
##   ..   genus = col_character(),
##   ..   species = col_character(),
##   ..   mass = col_double(),
##   ..   reference = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Figure 1
print("3b")
```

```
## [1] "3b"
```

```
mammal_sizes_cleaned <- mammal_sizes %>%
  filter(continent != "Insular", continent != "Oceanic", continent != "EA") %>%
  filter(status %in% c("extant", "extinct"))

ggplot(mammal_sizes_cleaned, aes(x = mass, fill = status)) +
  geom_histogram(bins = 25) +
  scale_x_log10() +
  facet_wrap(~continent) +
  labs(x = "Mass (g)", y = "Number of Species")
```

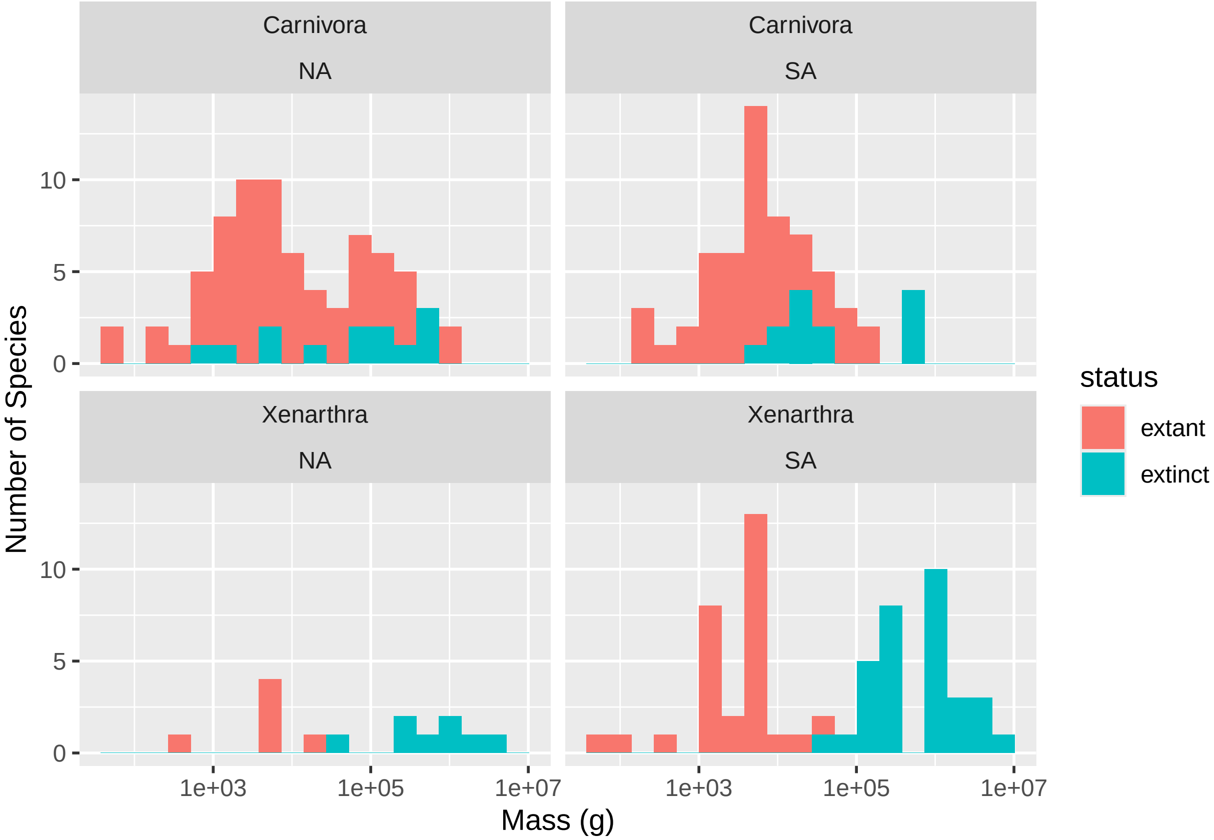


```
# Figure 2
print("3c")
```

```
## [1] "3c"
```

```
fig_2_data <- mammal_sizes %>%
  filter(continent %in% c("NA", "SA"), order %in% c("Xenarthra", "Carnivora"),
         status %in% c("extinct", "extant"))
```

```
ggplot(fig_2_data, aes(x = mass, fill = status)) +
  geom_histogram(bins = 19) +
  scale_x_log10() +
  facet_wrap(~order+continent) +
  labs(x = "Mass (g)", y = "Number of Species")
```

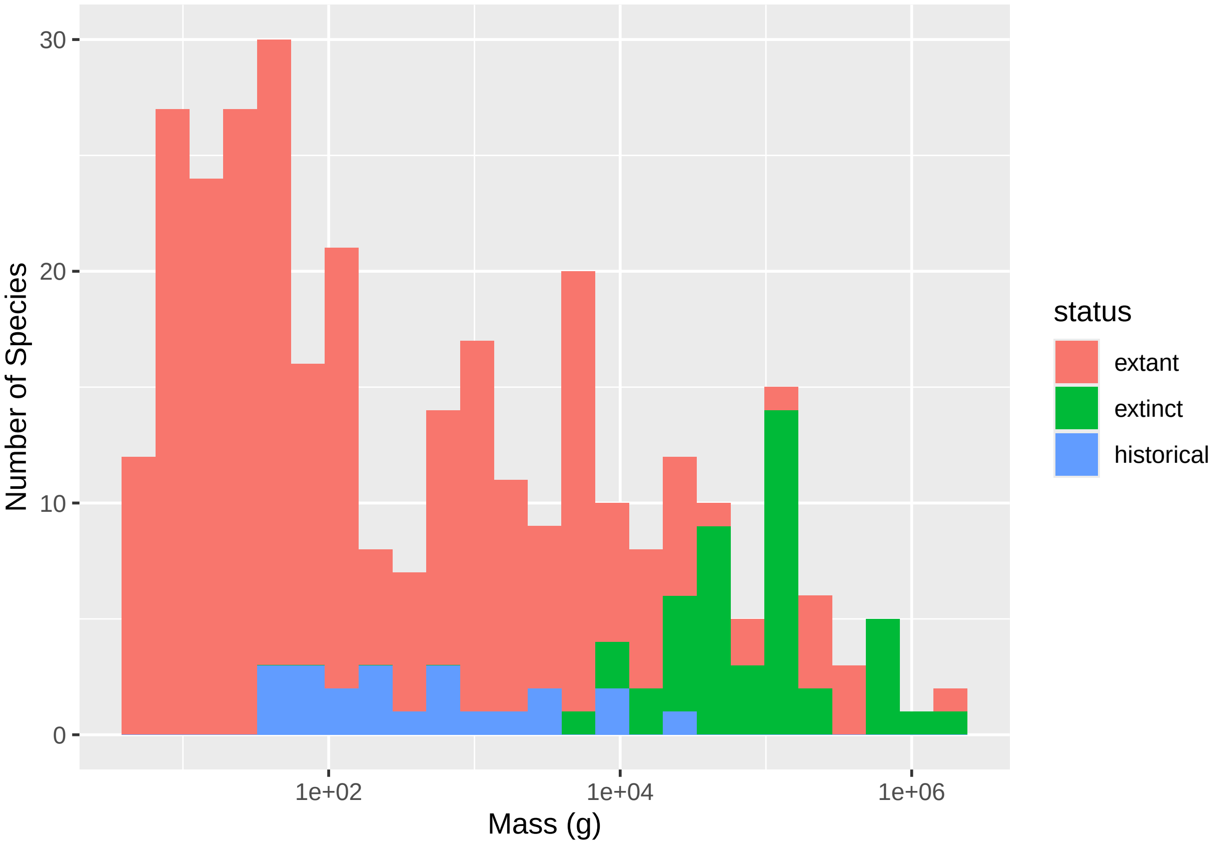


```
# Figure 3
print("3d")
```

```
## [1] "3d"
```

```
fig_3_data <- mammal_sizes %>%
  filter(continent == "AUS", status %in% c("extinct", "extant", "historical"))

ggplot(fig_3_data, aes(x = mass, fill = status)) +
  geom_histogram(bins = 25) +
  scale_x_log10() +
  labs(x = "Mass (g)", y = "Number of Species")
```



```
# Optional

# This is a fancy way to dynamically eliminate sites with no extinct species
# But EA can also just be eliminated manually by adding it to the continent
# filtering step below.
print("3e")
```

```
## [1] "3e"
```

```
extinct_rich_by_continent <- mammal_sizes %>%
  filter(status == "extinct") %>%
  distinct(continent)
extinct_rich_by_continent
```

```
## # A tibble: 5 x 1
##   continent
##   <chr>
## 1 AF
## 2 AUS
## 3 Insular
## 4 NA
## 5 SA
```

```
mammal_sizes_cleaned <- mammal_sizes %>%
  inner_join(extinct_rich_by_continent) %>%
  filter(continent != "Insular", continent != "Oceanic") %>%
  filter(status %in% c("extant", "extinct"))
mammal_sizes_cleaned
```

```
## # A tibble: 3,091 x 8
##    continent status  order        family  genus      species      mass reference
##    <chr>     <chr>   <chr>        <chr>   <chr>      <chr>       <dbl> <chr>
##  1 AF        extant  Artiodactyla Bovidae Addax      nasomacul~ 7.00e4 60
##  2 AF        extant  Artiodactyla Bovidae Aepyceros  melampus   5.25e4 63, 70
##  3 AF        extant  Artiodactyla Bovidae Alcelaphus buselaphus 1.71e5 63, 70
##  4 AF        extant  Artiodactyla Bovidae Ammodorcas clarkei    2.80e4 60
##  5 AF        extant  Artiodactyla Bovidae Ammotragus lervia     4.80e4 75
##  6 AF        extant  Artiodactyla Bovidae Antidorcas marsupial~ 3.90e4 60
##  7 AF        extinct Artiodactyla Bovidae Antidorcas bondi      3.4 e4 1
##  8 AF        extinct Artiodactyla Bovidae Antidorcas australis  4   e4 2
##  9 AF        extant  Artiodactyla Bovidae Bos        taurus     9   e5 <NA>
## 10 AF        extant  Artiodactyla Bovidae Capra      walie      1   e5 <NA>
## # i 3,081 more rows
```

4. Palmer Penguins (35 points)

In this question, we are going to take some raw data and recreate a clean dataset. This is from the `palmerpenguins` R package, which has body size measurements from 3 species of Antarctic penguins from 2007-2009. First, we need to load in the package and take a look at the clean version of the data that we are trying to recreate.

```
library(palmerpenguins)

# because the data is from a package, it doesnt automatically show up in our environment unless with us
penguins <- penguins
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

Now, let's bring in the original 3 datasets that were used to create this cleaned version (`penguins`)

```
# Adelie penguin data from: https://doi.org/10.6073/pasta/abc50eed9138b75f54eaada0841b9b86
url_adelie <- "https://portal.edirepository.org/nis/dataviewer?packageid=knb-lter-pal.219.3&entityid=002
adelie <- read_csv(url_adelie)

# Gentoo penguin data from: https://doi.org/10.6073/pasta/2b1cff60f81640f182433d23e68541ce
url_gentoo <- "https://portal.edirepository.org/nis/dataviewer?packageid=knb-lter-pal.220.3&entityid=e03
```

```r
gentoo <- read_csv(url_gentoo)

# Chinstrap penguin data from: https://doi.org/10.6073/pasta/409c808f8fc9899d02401bdb04580af7
url_chinstrap <- "https://portal.edirepository.org/nis/dataviewer?packageid=knb-lter-pal.221.2&entityid
chinstrap <- read_csv(url_chinstrap)
```

Problem breakdown (both `a` and `b` are graded for completion, not accuracy)

a. Start by breaking down the problem into plain language. This stage shouldn't include any specific functions but is allowing you to talk through the steps conceptually.

b. Make some predictions about the order in which you will want to accomplish this task, including which functions you will likely be using.

c. Recreate the clean dataset (`penguins`). Below are some tips (in no particular order) that will likely be helpful along the way

   - There is one instance in the sex column of one of the species where an unknown sex is marked with a `.` instead of `NA`
   - You do not need to match up data types exactly (character and factors are mostly interchangeable; same with integer, numeric, and double)
   - The year column is derived from the `Date Egg` column in the original 3 dataframes
   - Culmen is basically a fancy word for a bird's bill
   - I've taught you multiple ways to pull out a specific part of a character string. Additional helpful hints are that the regex for extracting the first word in a string is `'\\w*'`; there is also a function called `word()` that is part of the `stringr` package.

You will know that you have successfully completed the task at hand if you run the code `setdiff(your_clean_df, penguins)`, and the result has 0 rows.

The `setdiff()` function takes 2 dataframes and looks for any differences. The output is a dataframe with rows that do not match up. If you have 0 rows that don't match, that means all rows do match!

```r
penguins_raw <- bind_rows(adelie, gentoo, chinstrap)

penguins_clean <- penguins_raw %>%
  select(species = Species, island = Island,
         bill_length_mm = `Culmen Length (mm)`,
         bill_depth_mm = `Culmen Depth (mm)`,
         flipper_length_mm = `Flipper Length (mm)`,
         body_mass_g = `Body Mass (g)`,
         sex = Sex, year = `Date Egg`) %>%
  mutate(sex = na_if(sex, "."),
         species = str_extract(species, '\\w*'),
         sex = tolower(sex),
         year = lubridate::year(year))

setdiff(penguins_clean, penguins)
```

```
## # A tibble: 0 x 8
## # i 8 variables: species <chr>, island <chr>, bill_length_mm <dbl>,
## #   bill_depth_mm <dbl>, flipper_length_mm <dbl>, body_mass_g <dbl>, sex <chr>,
## #   year <dbl>
```