Week 7 Assignment

Ellen Bledsoe

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

Oct 14 at 2pm

Assignment Exercises

1. Set-Up (5 pts)

Load in the tidyverse to get started. If you haven't yet downloaded the three Portal data files for the week (found at the beginning of the lesson), you will need to do so.

2. Portal Data Review (25 points)

For this question, we are using the Portal data to review many of the dplyr, tidyr, and ggplot2 functions we have learned so far.

Load the three dataframes below from your data folder 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_types: Control, Long-term Krat Exclosure, and Short-term Krat Exclosure. The histogram should *not* be stacked, meaning each bin for each group should start at 0; modify the transparency so that any overlap is visible. Include good axis labels, a good legend label, and a theme for the plot. Save the plot as a png file.

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 (`mammal-size-data-clean.txt`) that has data on the mass of recently extinct mammals as well as extant mammals (i.e., those that are still alive today).

a. Read 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. Pay particular attention to the `continent`.

   You should do this *within* the `read_` function, not after you have read in the file.

   Hint: take a look at the Week 5 lesson (Data Viz) and the `na` *argument inside of* `read_tsv()`*.*

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. The histogram should *not* be stacked, meaning each bin for each group should start at 0; modify the transparency so that any overlap is visible. Use good axis and legend labels and add a theme.

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. The histogram should *not* be stacked, meaning each bin for each group should start at 0; modify the transparency so that any overlap is visible. Use good axis and legend labels and add a theme.

   Hint: look at the help file for `facet_wrap` *to figure out how to facet by more than one variable.*

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 a 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. The histogram should *not* be stacked, meaning each bin for each group should start at 0; modify the transparency so that any overlap is visible. Use good axis and legend labels and add a theme.

e. (Challenge: optional) Instead of excluding continent `EA` by name in your analysis (in part 3b), modify your code to determine from the data which continents had species that went extinct in the Pleistocene and only include those continents.

4. Palmer Penguins (35 points)

In this question, we are going to take some raw data and recreate a clean dataset.

The data are from the `palmerpenguins` R package, which has body size measurements from three 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.

```r
library(palmerpenguins)
```

```

Attaching package: 'palmerpenguins'


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

    penguins, penguins_raw
```

```r
# because the data is from a package, it doesnt automatically show up in our environment unle
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, read in the three original datasets (`adelie.csv`, `chinstrap.csv`, and `gentoo.csv`) that were used to create this cleaned version (`penguins`). They can be found in the `data` folder.

Take a look at the three original dataframes and compare them to the `penguins` dataframe.

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.

Coding

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

   - There is an instance in the sex column of one of the species where an unknown sex is marked with a `.` instead of `NA`
   - The year column is derived from the `Date Egg` column in the original dataframes
   - Culmen is basically a fancy word for a bird's bill
   - The regex for extracting the first word in a string using the `extract` function is `"(\\w*)"`
   - You do not need to match up most data types exactly (character and factors are mostly interchangeable; same with integer, numeric, and double). You might, however, run into the issue of character and integer data being incompatible. There are a number of different ways to address this, one of them being the `convert = TRUE` argument.

   If you are already familiar with the `stringr` and/or `lubridate` packages, I encourage you to challenge yourself and think about how you can use only the functions we have covered in class thus far to complete this question.

   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!

d. Create a scatterplot with bill length on the x-axis and bill depth on the y-axis. Add a line of best fit using another geom (remember the argument `method = "lm"` to make it linear). Make the line of best fit black instead of the default blue. Add descriptive column names and a theme.

e. Make the same plot as (d) above, except this time, both the points and the lines of best fit should be colored by species.

5. Preparing for Next Week

This is not graded homework, but I want us to be as prepared for next week as we can be.

That means that I would like you to have *both* R and RStudio downloaded on your own device. If you do not already have these downloaded, instructions for how to do so can be found in the description of "Weekly Materials > 8. Reproducibility" on our D2L site.