Homework Assignment - Week 1

Alison Lawyer

2024-09-12

```
# Homework Objectives In this homework, you will: - Import a dataset related to ecology
# and perform basic exploration.  - Manipulate the data using `tidyverse` functions.  -
# Reshape data and create new variables.  - Group and summarize data to gain ecological
# insights.  We will use the `palmerpenguins` dataset for this assignment, which contains
# data about three species of penguins in the Palmer Archipelago, Antarctica.

# Install/load and load the required packages Remember, if you have previously installed
# a package (like tidyverse) you do not need to install it again, just load it
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 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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 errors
```

```
library(palmerpenguins)

# Part A: Import and Explore the Dataset 1. Load the penguins dataset
data("penguins")

# 2. **Inspect the first few rows of the dataset**. What variables are included in the
# dataset?
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>
```

```
# Variables include: species, island, bill length and depth, flipper length, body mass,
# sex and year

# 3. **Check the dimensions of the dataset.** How many rows and columns are there?
dim(penguins)
```

```
## [1] 344   8
```

```
# There are 8 columns and 344 rows

# 4. **Get summary statistics** for each variable in the dataset.
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
# Part B: Basic Data Manipulation 1. **Select specific columns**.  Create a new dataset
# with only the following columns: `species`, `bill_length_mm`, `flipper_length_mm`, and
# `body_mass_g`.  Print the first rows of this dataset using the head() function.
penguins_selected <- penguins %>%
    select(species, bill_length_mm, flipper_length_mm, body_mass_g)
head(penguins_selected)
```

```
## # A tibble: 6 x 4
##   species bill_length_mm flipper_length_mm body_mass_g
##   <fct>            <dbl>             <int>       <int>
## 1 Adelie            39.1               181        3750
## 2 Adelie            39.5               186        3800
## 3 Adelie            40.3               195        3250
## 4 Adelie            NA                  NA          NA
## 5 Adelie            36.7               193        3450
## 6 Adelie            39.3               190        3650
```

```
# 2. **Filter the dataset** to only include penguins with a `body_mass_g` greater than
# 4000 grams. Print the first rows.
penguins_filtered <- penguins %>%
    filter(body_mass_g > 4000)
head(penguins_filtered)
```

```
## # 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           39.2          19.6               195        4675
## 2 Adelie  Torgersen           42            20.2               190        4250
## 3 Adelie  Torgersen           34.6          21.1               198        4400
## 4 Adelie  Torgersen           42.5          20.7               197        4500
## 5 Adelie  Torgersen           46            21.5               194        4200
## 6 Adelie  Dream               39.2          21.1               196        4150
## # i 2 more variables: sex <fct>, year <int>
```

```
# 3. **Arrange the dataset** by `bill_length_mm` in ascending order. Print the first
# rows.
penguins_arranged <- penguins %>%
    arrange(bill_length_mm)
head(penguins_arranged)
```

```
## # 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  Dream               32.1          15.5               188        3050
## 2 Adelie  Dream               33.1          16.1               178        2900
## 3 Adelie  Torgersen           33.5          19                 190        3600
## 4 Adelie  Dream               34            17.1               185        3400
## 5 Adelie  Torgersen           34.1          18.1               193        3475
## 6 Adelie  Torgersen           34.4          18.4               184        3325
## # i 2 more variables: sex <fct>, year <int>
```

```
# 4. **Create a new variable** that calculates the ratio of bill length
# (`bill_length_mm`) to flipper length (`flipper_length_mm`).  Call the new variable
# `bill_flipper_ratio`. Print the first rows.
penguins <- penguins %>%
    mutate(bill_flipper_ratio = bill_length_mm/flipper_length_mm)
head(penguins)
```

```
## # A tibble: 6 x 9
##   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 3 more variables: sex <fct>, year <int>, bill_flipper_ratio <dbl>
```

```
# Part C: Grouping and Summarizing Data 1. **Group the data by penguin species
# (`species`)** and calculate the following summaries: - The average body mass
# (`body_mass_g`) for each species.  - The maximum flipper length (`flipper_length_mm`)
# for each species.  Print the grouped dataframe

## initial attempt, produced summary table with NA for 2/3 species
```

```r
penguins_grouped <- penguins %>%
    group_by(species) %>%
    summarise(avg_body_mass = mean(body_mass_g), max_flipper_length = max(flipper_length_mm))
head(penguins_grouped)
```

```
## # A tibble: 3 x 3
##   species   avg_body_mass max_flipper_length
##   <fct>             <dbl>              <int>
## 1 Adelie               NA                 NA
## 2 Chinstrap          3733.               212
## 3 Gentoo               NA                 NA
```

```r
## second attempt, avoiding NA values
penguins_grouped <- penguins %>%
    group_by(species) %>%
    summarise(avg_body_mass = mean(body_mass_g, na.rm = TRUE), max_flipper_length = max(flipper_length_
        na.rm = TRUE))
head(penguins_grouped)
```

```
## # A tibble: 3 x 3
##   species   avg_body_mass max_flipper_length
##   <fct>             <dbl>              <int>
## 1 Adelie            3701.               210
## 2 Chinstrap         3733.               212
## 3 Gentoo            5076.               231
```

```r
# 2. **Interpret the results**. Which species has the highest average body mass? Which
# species has the longest maximum flipper length?

## Answer: The Gentoo species has the highest average body mass, as well as the longest
## max flipper length.


# Part D: Reshaping Data 1. **Convert the `penguins` dataset from wide to long format**
# using the `pivot_longer()` function.  Focus on the columns `bill_length_mm`,
# `flipper_length_mm`, and `body_mass_g`.  Print the first rows.
penguins_long <- penguins %>%
    pivot_longer(cols = c(bill_length_mm, flipper_length_mm, body_mass_g), names_to = "Measurement",
        values_to = "Values")
head(penguins_long)
```

```
## # A tibble: 6 x 8
##   species island bill_depth_mm sex    year bill_flipper_ratio Measurement Values
##   <fct>   <fct>          <dbl> <fct> <int>              <dbl> <chr>        <dbl>
## 1 Adelie  Torge~          18.7 male   2007              0.216 bill_lengt~   39.1
## 2 Adelie  Torge~          18.7 male   2007              0.216 flipper_le~  181
## 3 Adelie  Torge~          18.7 male   2007              0.216 body_mass_g 3750
## 4 Adelie  Torge~          17.4 fema~  2007              0.212 bill_lengt~   39.5
## 5 Adelie  Torge~          17.4 fema~  2007              0.212 flipper_le~  186
## 6 Adelie  Torge~          17.4 fema~  2007              0.212 body_mass_g 3800
```

```r
# 2. **Now, convert the data back to wide format** using the `pivot_wider()` function.
# Print the first rows.
penguins_wide <- penguins_long %>%
    pivot_wider(names_from = Measurement, values_from = Values)
head(penguins_wide)
```

```
## # A tibble: 6 x 9
##   species island    bill_depth_mm sex     year bill_flipper_ratio bill_length_mm
##   <fct>   <fct>             <dbl> <fct>  <int>              <dbl>          <dbl>
## 1 Adelie  Torgersen          18.7 male    2007              0.216           39.1
## 2 Adelie  Torgersen          17.4 female  2007              0.212           39.5
## 3 Adelie  Torgersen          18   female  2007              0.207           40.3
## 4 Adelie  Torgersen          NA   <NA>    2007             NA               NA
## 5 Adelie  Torgersen          19.3 female  2007              0.190           36.7
## 6 Adelie  Torgersen          20.6 male    2007              0.207           39.3
## # i 2 more variables: flipper_length_mm <dbl>, body_mass_g <dbl>
```

```r
# Step E: Handling Missing Data 1. **Identify rows with `NA` values** in any column.
# Print these rows.
penguins_na_rows <- penguins[!complete.cases(penguins), ]
print(penguins_na_rows)
```

```
## # A tibble: 11 x 9
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           NA            NA                NA          NA
##  2 Adelie  Torgersen           34.1          18.1              193        3475
##  3 Adelie  Torgersen           42            20.2              190        4250
##  4 Adelie  Torgersen           37.8          17.1              186        3300
##  5 Adelie  Torgersen           37.8          17.3              180        3700
##  6 Adelie  Dream               37.5          18.9              179        2975
##  7 Gentoo  Biscoe              44.5          14.3              216        4100
##  8 Gentoo  Biscoe              46.2          14.4              214        4650
##  9 Gentoo  Biscoe              47.3          13.8              216        4725
## 10 Gentoo  Biscoe              44.5          15.7              217        4875
## 11 Gentoo  Biscoe              NA            NA                NA          NA
## # i 3 more variables: sex <fct>, year <int>, bill_flipper_ratio <dbl>
```

```r
# 2. **Remove rows with missing values** from the dataset.  .  Print the first rows.
penguins_no_na <- penguins %>%
    na.omit()
head(penguins_no_na)
```

```
## # A tibble: 6 x 9
##   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           36.7          19.3              193        3450
## 5 Adelie  Torgersen           39.3          20.6              190        3650
## 6 Adelie  Torgersen           38.9          17.8              181        3625
## # i 3 more variables: sex <fct>, year <int>, bill_flipper_ratio <dbl>
```

```r
# 3. **Fill missing values** in the `body_mass_g` column with the mean body mass.  Print
# the first rows.

## first attempt, using `replace_na`
penguins_mass_filled <- penguins %>%
    replace_na(list(body_mass_g = as.integer(mean(penguins$body_mass_g, na.rm = TRUE))))
head(penguins_mass_filled)
```

```
## # A tibble: 6 x 9
##   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        4201
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 3 more variables: sex <fct>, year <int>, bill_flipper_ratio <dbl>
```

```r
## second attempt, using `ifelse`
penguins_mass_filled <- penguins %>%
    mutate(body_mass_g = ifelse(is.na(body_mass_g), mean(body_mass_g, na.rm = TRUE), body_mass_g))
head(penguins_mass_filled)
```

```
## # A tibble: 6 x 9
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <dbl>
## 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       4202.
## 5 Adelie  Torgersen           36.7          19.3               193       3450
## 6 Adelie  Torgersen           39.3          20.6               190       3650
## # i 3 more variables: sex <fct>, year <int>, bill_flipper_ratio <dbl>
```

```r
# Step F: Save the Modified Dataset 1. **Save the final cleaned and modified dataset** as
# a CSV file.
write.csv(penguins_mass_filled, "penguins_cleaned.csv")
```