Data Wrangling - Exercise 1: Data manipulation Solution

Luna L Sanchez Reyes

2023-02-02

1. Load the data set from the file `surveys.csv` into R using the function `read.csv()`.

```r
surveys <- read.csv(file = "../data-raw/surveys.csv")
str(surveys)
```

```
## 'data.frame':    35549 obs. of  9 variables:
##  $ record_id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ month          : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ day            : int  16 16 16 16 16 16 16 16 16 16 ...
##  $ year           : int  1977 1977 1977 1977 1977 1977 1977 1977 1977 1977 ...
##  $ plot_id        : int  2 3 2 7 3 1 2 1 1 6 ...
##  $ species_id     : chr  "NL" "NL" "DM" "DM" ...
##  $ sex            : chr  "M" "M" "F" "M" ...
##  $ hindfoot_length: int  32 33 37 36 35 14 NA 37 34 20 ...
##  $ weight         : int  NA NA NA NA NA NA NA NA NA NA ...
```

---

2. Use `select()` to create a new data frame object called `surveys1` with just the `year`, `month`,

`day`, and `species_id` columns in that order.

- `select()` is from package `dplyr`, so we need to load the package, preferably in the "setup" R chunk, but can also be done here and commented out for knitting.

```r
# library(dplyr)
surveys1 <- select(surveys, year, month, day, species_id)
str(surveys1)
```

```
## 'data.frame':    35549 obs. of  4 variables:
##  $ year      : int  1977 1977 1977 1977 1977 1977 1977 1977 1977 1977 ...
##  $ month     : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ day       : int  16 16 16 16 16 16 16 16 16 16 ...
##  $ species_id: chr  "NL" "NL" "DM" "DM" ...
```

---

3. Create a new data frame called `surveys2` with the `year`, `species_id`, and weight in kilograms of each individual, with no null weights.

Use `mutate()`, `select()`, and `filter()` with `!is.na()`. The weight in the table is given in grams so you will need to create a new column called "weight_kg" for weight in kilograms by dividing the weight column by 1000.

- The goal of this question is to make them realize that they have to create intermediate data frames.

First, filter Na values in weight:

```
surveys_tmp <- filter(surveys, !is.na(weight))
str(surveys_tmp)
```

```
## 'data.frame':    32283 obs. of  9 variables:
##  $ record_id      : int  63 64 65 66 67 68 69 70 71 74 ...
##  $ month          : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ day            : int  19 19 19 19 19 19 19 19 19 19 ...
##  $ year           : int  1977 1977 1977 1977 1977 1977 1977 1977 1977 1977 ...
##  $ plot_id        : int  3 7 4 4 7 8 2 3 7 8 ...
##  $ species_id     : chr  "DM" "DM" "DM" "DM" ...
##  $ sex            : chr  "M" "M" "F" "F" ...
##  $ hindfoot_length: int  35 37 34 35 35 32 15 21 36 12 ...
##  $ weight         : int  40 48 29 46 36 52 8 22 35 7 ...
```

Second, create the new column with weight in Kg. Overwrite the object.

```
surveys_tmp <- mutate(surveys_tmp, weight_kg = weight/1000)
str(surveys_tmp)
```

```
## 'data.frame':    32283 obs. of  10 variables:
##  $ record_id      : int  63 64 65 66 67 68 69 70 71 74 ...
##  $ month          : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ day            : int  19 19 19 19 19 19 19 19 19 19 ...
##  $ year           : int  1977 1977 1977 1977 1977 1977 1977 1977 1977 1977 ...
##  $ plot_id        : int  3 7 4 4 7 8 2 3 7 8 ...
##  $ species_id     : chr  "DM" "DM" "DM" "DM" ...
##  $ sex            : chr  "M" "M" "F" "F" ...
##  $ hindfoot_length: int  35 37 34 35 35 32 15 21 36 12 ...
##  $ weight         : int  40 48 29 46 36 52 8 22 35 7 ...
##  $ weight_kg      : num  0.04 0.048 0.029 0.046 0.036 0.052 0.008 0.022 0.035 0.007 ...
```

Finally, select the columns that you want for the new `data frame`:

```
surveys2 <- select(surveys_tmp, year, species_id, weight_kg)
str(surveys2)
```

```
## 'data.frame':    32283 obs. of  3 variables:
##  $ year      : int  1977 1977 1977 1977 1977 1977 1977 1977 1977 1977 ...
##  $ species_id: chr  "DM" "DM" "DM" "DM" ...
##  $ weight_kg : num  0.04 0.048 0.029 0.046 0.036 0.052 0.008 0.022 0.035 0.007 ...
```

---

4. Use the `filter()` function to get all of the rows in the data frame `surveys2` for the species ID "SH".

The goal of this point is to make them practice the function `filter()` and logical statements.

For next iterations of the course:

- Tell them no to print the whole table in the knitted document. For this, there are a couple options:
  - You have to create a data frame object
  - You have to use r chunk options `results = 'hide`
- Ask how many rows does the filtered table have?
- Make them filter surveys1 instead, so you know that they did not overwrite it during question 2 and 3.

```
surveys_filtered <- filter(surveys2, species_id == "SH")
str(surveys_filtered)
```

```
## 'data.frame':    141 obs. of  3 variables:
##  $ year      : int  1978 1982 1982 1986 1987 1987 1987 1987 1987 1988 ...
##  $ species_id: chr  "SH" "SH" "SH" "SH" ...
##  $ weight_kg : num  0.089 0.106 0.052 0.055 0.077 0.078 0.104 0.058 0.052 0.06 ...
```