

# Week 3 DS Practice Problems

SIBDS @ Columbia

2022-06-17

## Getting Started

1. Create a new project in your RWorkspace named `week_3_recitation`.
2. Click into the RProject and create a new RMarkdown document within the project named `week_3_recitation`.
3. Modify the YAML header to your preference and knit the RMarkdown document to make sure it runs as expected.
4. Finally, create a folder within the project folder named `data`. Put the data files from Jeff's lecture on Tuesday into this folder.

## Writing with Data

1. Write a named code chunk that creates a dataframe comprised of:
  - A numeric variable containing a random sample of size 500 from a normal variable with mean 1
  - A logical vector indicating whether each sampled value is greater than zero
  - A numeric vector containing the absolute value of each element
2. Produce a histogram of the absolute value variable just created.
3. Add an inline summary giving the median value rounded to two decimal places.
4. After the previous code chunk, write a bulleted list giving the mean, median, and standard deviation of the original random sample

## Data Import

1. Make sure you are able to load the `FAS_pups.csv` dataset. Use both absolute and relative paths.
2. Quit R Studio and move the directory containing your project, data, and RMarkdown document. Repeat the previous data import process; do both absolute and relative paths still work?
3. Import the dataset `FAS_pups.csv` using `dplyr::read_csv` and name the dataset `pups`. Make sure the column names are reasonable, and take some quick looks at the dataset.
  - a. What happens if your specifications for column parsing aren't reasonable (e.g. character instead of double, or vice versa)?
4. Now, we will use base R's `read.csv` function to import the `FAS_pups.csv` dataset. Compare the class of this dataset to the one produced by `read_csv`. Try printing both in the console – what happens?
5. After cleaning up the names, try accessing the `Sex` variable using `S` (e.g., `pups_data$S`). What happens?

## Data Manipulation

1. In the `pups` data, select the columns containing litter number, sex, and PD ears. Then, do the following:
  - a. Filter to include only pups with sex 1
  - b. Filter to include only pups with PD walk less than 11 and sex 2
  - c. Create a variable that subtracts 7 from PD pivot
  - d. Create a variable that is the sum of all the PD variables
2. Write a chain of commands that:
  - Loads the `pups` data
  - Cleans the variable names
  - Filters the data to include only pups with sex 1
  - Removes the PD ears variable
  - Creates a variable that indicates whether PD pivot is 7 or more days