Dealing with Missing Data

Author: Rachel Hartnett, Oklahoma State University

Focus: The focus of this lesson is to help students learn common data cleaning procedures on a dataset, specifically with missing data.

Overview: The overall goal is to help students learn common data cleaning procedures on a dataset once they've collected measurements and before they are able to start their analysis. This swirl lesson is designed to provide a baseline of knowledge for what steps need to be taken to deal with missing data, and how to do that in R with the tidyverse package.

Learning objectives: By the end of the lesson, students should be able to provide a checklist for cleaning their datasets of missing data, be able to modify a dataset in R with missing data, and be able to export that dataset into a .csv file.

Lesson sequence: Provide a numbered, ordered list of the activities within your swirl lesson. This list can be taken from step 4 in your initial lesson design, with any modifications that were introduced.

1. Read instructions on Canvas
2. Complete the swirl lesson on dealing with missing data
3. Make their own flow chart of steps to deal with their data
4. Canvas Assessment which includes a checklist and completion email.

Pre-lesson activities: Please ***see the implementation notes below*** to learn more about previous experience students had with R before this lesson. The instructions posted on Canvas for students to learn about the lesson can be seen here:
Start Canvas Instructions

This week we will be learning about how to clean data.

Data cleaning is a critical step in good experimental design and analysis techniques. DO NOT SKIP DATA CLEANING STEPS when working with your own data. While tedious, cleaning your data appropriately is key to an honest analysis. This week you will complete a new lesson on cleaning data.

For this assignment, there are three parts:

1) Completing the lesson. I have built in the ability to email me at the end of the swirl lesson that you have completed the lesson.

2) Submitting the exported .csv file (exporting the file is part of the lesson, but you'll need to submit it to Canvas).

3) Submitting a Checklist of what steps you would take to deal with missing data. Be specific!!

Navigating to the Swirl lesson

Now, in order to navigate to the lesson, you will need to download this new Cleaning_Data.swc file, and save this in a folder that you remember. The default is typically the downloads folder. To install this new swirl course, use the following lines of code in the R studio terminal (see walk through slides with screenshots):

```
library(swirl)
install_course()
```

The line above will open a pop-up window of your file directories, and you should navigate to the .swc file that you downloaded. Then you'll get into swirl:

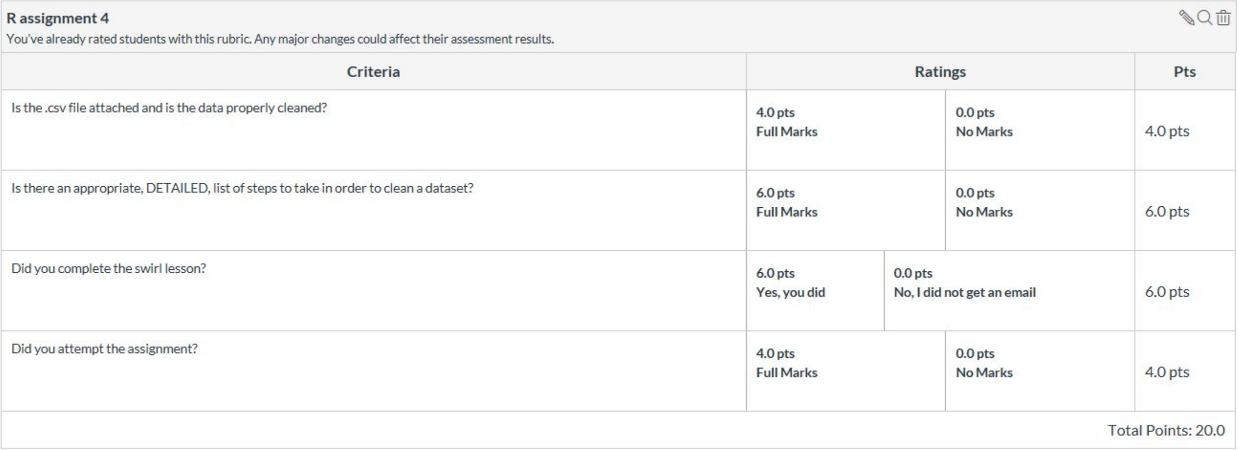
```
swirl()
Course: Cleaning_Data
Lesson: Missing_Data
```

If needed, here are instructions to manually install the required packages:

```
install.packages(c("tidyr", "dplyr"))
library(tidyr)
library(dplyr)
```

End Canvas instructions

Post-lesson activities: Below you can find a screenshot of the rubric used to assess this lesson:



R assignment 4

You've already rated students with this rubric. Any major changes could affect their assessment results.

Criteria	Ratings		Pts
Is the .csv file attached and is the data properly cleaned?	4.0 pts Full Marks	0.0 pts No Marks	4.0 pts
Is there an appropriate, DETAILED, list of steps to take in order to clean a dataset?	6.0 pts Full Marks	0.0 pts No Marks	6.0 pts
Did you complete the swirl lesson?	6.0 pts Yes, you did	0.0 pts No, I did not get an email	6.0 pts
Did you attempt the assignment?	4.0 pts Full Marks	0.0 pts No Marks	4.0 pts
			Total Points: 20.0

Implementation notes: This lesson was developed for an introductory biostatistics course, which included R basic programming E lessons from Github (Basic Building Blocks, Base Graphics, and Matrices and Data Frames) as well as additional lessons developed by me that you can find in QUBES OER. Specific functions referenced in this lesson that I assume prior knowledge of include read.csv(), View(), str().
At the end of this lesson, I have used the notify() function, which can be found within customTests.R. I recommend including your email in the yaml prompt.

This lesson is designed to be completed within ~35 minutes during class; however due to COVID-19 disruptions to in-person classes, this lesson was implemented online. The main troubleshooting issue was with the first use of the command line using the piping operator. This has been modified and hopefully resolved. Future modifications of this lesson could include more advanced visualizations and include different datasets that require different assumptions to be met or need to be cleaned in different ways.

References:

Hartnett, Rachel N. (2019), Data from: Variation in life-history traits among Daphnia and its relationship to species-level responses to phosphorus limitation, Dryad, Dataset, https://doi.org/10.5061/dryad.d4v7g74

Sullivan, J. (2019). “Data Cleaning with R and the Tidyverse: Detecting Missing Values.” March 21, 2019. Retrieved on May 13, 2020 from: https://towardsdatascience.com/data-cleaning-with-r-and-the-tidyverse-detecting-missing-values-ea23c519bc62

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). Cleaning up your Act. In, *Using multivariate statistics* (Vol. 5). Boston, MA: Pearson.