Assignment 1: Reproducibility, Workflow, Version Control

Ying Wei Jong

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on reproducibility, workflow, and version control.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A01_Reproducibility.pdf") prior to submission.

The completed exercise is due on Thursday, 17 January, 2018 before class begins.

1) Discussion Questions

Question

Why are reproducible practices becoming the norm in data analytics?

> Answer: Because people need to understand how was the initial raw data transformed into the data they have on hand, and they should have to be able to reproduce the whole workflow of converting raw data into processed data to ensure consistency.

Question

What are your previous experiences with data analytics, R, and Git? Include both formal and informal training.

> Answer: I did a bit of data processing for my MP project, but I did not learn it in a systematic manner. In my MP, I read the raw csv and used codes to modify the dataframe to fix typos on diameters and species. I used Github desktop for my Advanced GIS class.

Question

Are there any components of the course about which you feel confident?

> Answer: The basic manipulation of R and Rstudio.

Question

Are there any components of the course about which you feel apprehensive?

> Answer: Pulling data from master. Troubleshooting in general.

2) GitHub

Your Repository

Provide a link below to your course repository in GitHub. Make sure you have pulled all recent changes from the course repository (https://github.com/KateriSalk/Environmental_Data_Analytics) and that you have updated your course README file.

> Answer: https://github.com/ywjong/Environmental_Data_Analytics
>
> I managed to pull recent changes from the course repository, just that there are merge conflicts because my name and Kateri's name are in the same line, and that is inevitable whenever the master and fork has different contents in the same line. I decided to keep the conflicted merge file that way (i.e. the one with both my name and Kateri's name and all instructions) and pushed it back to the cloud. I understand that additionally, I can fix it by editing the file the way I want it to be, and/or to and follow this page's instructions https://help.github.com/articles/resolving-a-merge-conflict-using-the-command-line/