# Assignment 1: Reproducibility, Workflow, Version Control

*Gaby Garcia*

## 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: Reproducibility is crucial in every scientific field, including data science. It is important to display each step of code; providing future data researchers with clear, concise, and correct code allows them to adapt our methods and results for their own research without having to attempt to reconstruct our methods. So in addition to facilitating future data collaborations, reproducible practices will save data scientists time and energy.

### Question

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

> Answer: In the summer of 2018, I completed a basic and free online course through DataCamp. It didn't cover much, but did get me used to how R works. I also took Env 710 with R this past fall semester, and was able to gain more experience with using statistical functions and graphs in R. I also have an assistantship as a water data analyst with Innovations in Infrastructure, which is a three year Duke based research project with the goal of investigating water infrastructure finance models. I worked with EPA SDWIS water systems data in R through this project.

**Question**

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

> Answer: I'm confident that I can import data, complete basic code (whether it is statistical functions or not), and visualize data using base R and GGplot.

**Question**

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

> Answer: In the past I have worked with very disorganized, complicated water data for my assistantship, and it was overwhelming to "wrangle" and clean up the data so anyone who opened the file could follow the code progression. I am also a little concerned about maintining reproducible workflows without running into errors, but hopefully experience will help me learn!

## 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/gdg12/Environmental_Data_Analytics