

# Project Instructions

*Environmental Data Analytics / Kateri Salk*

*Spring 2019*

## CREATE A REPOSITORY IN YOUR GITHUB ACCOUNT

1. Make sure you have completed the necessary software installations of:

**R**

**RStudio**

**Git**, with GitHub user account

**LaTeX**

2. Go to your user account on GitHub.
3. In the upper right corner, click the green “New” button.
4. Name your repository with recommended naming conventions. Write a short description of the purpose of the repository. Check the box to initialize the repository with a README. Add a .gitignore for R and add a GNU General Public License v3.0.

## LINK YOUR REPO TO YOUR LOCAL DRIVE WITH RSTUDIO

1. Click the “Clone or download” button for your repository and then the “copy” icon. Make sure the box header lists “Clone with HTTPS” rather than “Clone with SSH.” If not, click the “Use HTTPS” button and then copy the link.
2. Launch RStudio and select “New Project” from the File menu. Choose “Version Control” and “Git.”
3. Paste the repository URL and give your repository a name and a file path.

## CHOOSE A DATASET AND A RESEARCH QUESTION

1. Choose a dataset of interest. This can be one of the datasets we have analyzed in class this semester:
  - ECOTOX Neonicotinoids
  - EPA North Carolina Air Quality
  - North Temperate Lakes LTER
  - USGS Eno River Stream Monitoring

or a dataset of your choosing. If the latter, please consult with the course instructors and obtain their approval before proceeding.

2. Generate a research question that can be answered using data from your dataset. This question should be specific enough to be answered in the time span of the project but complex enough to require all steps in the data pipeline we have discussed in class.

## POPULATE YOUR REPO AND ANALYZE YOUR DATA

1. Populate your README with the necessary information for the project. Use the Environmental Data Analytics repository README and the associated README files for datasets as a guide.
2. Create folders for raw data, processed data, code, and output.
3. Work through the data pipeline with R (R script or RMarkdown document) to process, examine, wrangle, visualize, and model your data. Store this code and output in the appropriate sections in the “Project\_Template.Rmd” and rename the file as “Lastname\_ENV872\_Project.Rmd”.

### *Guidelines*

- Your **data exploration** section should contain **at least five lines** of code that generate summary information about your dataset (or components therein)
  - Your data exploration section should contain at least **three graphs**
  - Your **statistical modeling section** should contain at least **three tests**, including the **rationale** why you took the approach you did. Ensure you have met assumptions of tests or that you **justify** moving forward without meeting assumptions.
  - Your **data visualization** section should contain at least **three graphs**.
4. Commit and push your updates to Github (i.e., your Master repository) after each analysis session.

## CREATE AND SUBMIT A REPORT

1. Use the Project\_Template.Rmd file as a template for your report. Populate this document as you proceed with data analysis.
2. When your project is complete, knit your RMarkdown document into a PDF. Ensure that all sections have properly knitted (code, output, and messages all displaying as desired), and troubleshoot if necessary.
3. Make sure the PDF file has the file name “Lastname\_ENV872\_Project.pdf” and submit it to the dropbox in Sakai.