Project Instructions

Environmental Data Analytics | John Fay and Luana Lima | Developed by Kateri Salk

Fall 2022

CREATE A REPOSITORY IN YOUR GITHUB ACCOUNT

1. Go to your user account on GitHub an navigate to the repositories tab.
2. In the upper right corner, click the green "New" button.
3. Name your repository with recommended naming conventions (suggestion: *Lastname1Lastname2Lastname3_ENV872_*
   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.
4. Invite other group members as collaborators to the repository.

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 or a dataset of your choosing. If the latter, I recommend consulting with the course instructors to make sure the scope of the dataset is appropriate.
2. Generate a research question(s) that can be answered using data from your dataset. This question(s) 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. This should include your name, information about the course project, and information about the topic of interest.
2. Create folders for Data/Raw, Data/Processed, Code, and Output.
3. (You will create a metadata file for the dataset as part of assignment 11, due March 31).

4. Work through the data pipeline with R (R script or RMarkdown documents). Create separate code files for data processing/wrangling, data exploration, and data analysis. Commit and push your updates to Github (i.e., your remote repository) after each analysis session. Use informative commit messages.
5. When you are ready to do so, copy information from your code files into the "Project_Template.Rmd" file (instructions below). Rename the project template as "Lastname_ENV872_Project.Rmd".

COMPLETE YOUR PROJECT REPORT

General Guidelines

1. Write in scientific style, not narrative style
2. Global options for R chunks should be set so that only relevant output is displayed
3. Set up autoreferencing for figures and tables in your document.
4. Make sure your final knitted PDF looks professional. Format tables appropriately, size figures appropriately, make sure bulleted and numbered lists appear as such, avoid awkwardly placed page breaks, etc.
5. Make sure the PDF file has the file name "Lastname_ENV872_Project.pdf" and submit it to the dropbox in Sakai.

Rationale and Research Questions

Write 1-2 paragraph(s) detailing the rationale for your study. This should include both the context of the topic as well as a rationale for your choice of dataset (reason for location, variables, etc.). You may choose to include citations if you like (optional).

At the end of your rationale, introduce a numbered list of your questions (or an overarching question and sub-questions).

Dataset Information

Provide information on how the dataset for this analysis were collected, the data contained in the dataset, and any important pieces of information that are relevant to your analyses. This section should contain much of same information as the metadata file for the dataset but formatted in a way that is more narrative.

Describe how you wrangled your dataset in a format similar to a methods section of a journal article.

Add a table that summarizes your data structure (variables, units, ranges and/or central tendencies, data source if multiple are used, etc.). This table can be made in markdown text or inserted as a `kable` function in an R chunk. If the latter, do not include the code used to generate your table.

Exploratory Analysis

Insert exploratory visualizations of your dataset. This may include, but is not limited to, graphs illustrating the distributions of variables of interest and/or maps of the spatial context of your dataset. Format your R chunks so that graphs are displayed but code is not displayed. Accompany these graphs with text sections that describe the visualizations and provide context for further analyses.

Each figure should be accompanied by a caption, and each figure should be referenced within the text.

Scope: think about what information someone might want to know about the dataset before analyzing it statistically. How might you visualize this information?

Analysis

Insert visualizations and text describing your main analyses. Format your R chunks so that graphs are displayed but code and other output is not displayed. Instead, describe the results of any statistical tests in the main text (e.g., "Variable x was significantly different among y groups (ANOVA; $df = 300$, $F = 5.55$, $p < 0.0001$)"). Each paragraph, accompanied by one or more visualizations, should describe the major findings and how they relate to the question and hypotheses. Divide this section into subsections, one for each research question.

Each figure should be accompanied by a caption, and each figure should be referenced within the text

Summary and Conclusions

Summarize your major findings from your analyses in a few paragraphs. What conclusions do you draw from your findings? Relate your findings back to the original research questions and rationale.