Assignment 1 - ESM 244 (Winter 2021)

All parts due by 5pm PST on Monday 1/25/2021

Data wrangling & viz, principal components analysis, Shiny app sign-ups, build your skeleton website

- **Tasks 1 & 2:** Submit **individual** knitted HTML for Tasks 1 & 2 through GauchoSpace
- **Task 3:** Add your Shiny app information HERE (**one entry per group**)
- **Task 4:** Start building your personal website in R with {distill} and {postcards}

Note: I recommend using Tasks 1 and 2 in this assignment as an opportunity to practice working in a single version-controlled R Project, but in separate branches (maybe one for each task), merging into main through pull requests as practiced in our discussions.

Task 1 (individual): Data wrangling & visualization (Sierra amphibians)

For Task 1, you will read in data an .xlsx file, do some data wrangling as needed, then create two data visualizations and put them together in a finalized compound figure (e.g. using the {patchwork} package introduced in the Week 2 Lab).

Data summary: You will explore amphibian abundance data recorded by the Sierra Lakes Inventory Project. From the Environmental Data Initiative repository: "The Sierra Lakes Inventory Project (SLIP) was a research endeavor that ran from 1995-2002 and has supported research and management of Sierra Nevada aquatic ecosystems and their terrestrial interfaces. We described the physical characteristics of and surveyed aquatic communities for > 8,000 lentic water bodies in the southern Sierra Nevada, including lakes, ponds, marshes, and meadows."

Data citation: Knapp, R.A., C. Pavelka, E.E. Hegeman, and T.C. Smith. 2020. The Sierra Lakes Inventory Project: Non-Native fish and community composition of lakes and ponds in the Sierra Nevada, California ver 2. Environmental Data Initiative. https://doi.org/10.6073/pasta/d835832d7fd00d9e4466e44eea87fab3

Data: Click here to download the Sierra amphibians dataset

Metadata & info: Metadata is available HERE (see Amphibian metadata under 'Data Entities' to interpret the variables and levels)

Complete Task 1 in a single well-organized .Rmd. Read in the amphibian data (sierra_amphibians.xlsx), then do any wrangling necessary to create two **finalized** data visualizations:

1) A graph of total **mountain yellow-legged frog (*Rana muscosa*) count** each year across all water bodies, by life stage excluding the 'EggMass' level. In other words, you should find the total number of adult, subadult, and tadpole yellow-legged frogs observed in the entire study region by life stage and year, but you will *not* use the lake ID or amphibian_location in your analyses as additional grouping variables (thanks Elmera Azadpour for clarifying). **Hint:** Convert the date to a date with the {lubridate} package, then pull just the year using the lubridate::year() function...then you can group by year to get counts.

2) A column graph containing total counts (over all years of the study) of combined adult and subadult endangered mountain yellow-legged frogs (*Rana muscosa*) observed in the 5 lakes with the greatest total observed counts. In other words, this graph will have at least 5 columns (OK to have more if there are ties - thanks Michelle Shteyn), with Lake ID (these aren't specified in the dataset for confidentiality) as the categorical label on one axis, and total yellow-legged frog counts (adult + subadult counts) in the dataset on the other axis. Make sure they're in high-to-low or low-to-high order by total yellow-legged frog counts. You should exclude tadpoles for this graph. **Note:** Consider reformatting the lake ID from just a number to a label like "Lake 10025"), then use fct_reorder to convert it to an ordered factor.

Combine your two graphs into a single compound figure using the {patchwork} package. Add a finalized figure caption that appears below the compound figure in your knitted html. **Note:** You might realize once you combine your graphs into a single compound figure, you need to update the formatting (e.g. move legends or instead directly label, etc.). You should customize as necessary to make it a professional final output.

For Task 1, your knitted html should show:

1. Your organized code, with clear subsections and any useful descriptive text / annotation (e.g. if you wanted to highlight this as a code example for a prospective employer)
2. Your finalized compound figure, with a figure caption
3. Make sure to suppress any messages & warnings

Submit your Task 1 knitted html on GauchoSpace.

Task 2 (individual): Principal components analysis (coder's choice)

For this task, I'll provide a couple of datasets that you *can* use for PCA exploration, but you are also welcome to find/choose a **different** dataset to use. You only need to use **one** dataset to perform PCA, create a biplot, and interpret the results. Whichever dataset you choose, create a finalized HTML (knitted from .Rmd) that includes:

1. A useful descriptive introductory summary (3 - 4 sentences) that helps the audience understand the data (include a citation as necessary) and what you'll be exploring
2. All of your organized and well-annotated code (with warnings/messages suppressed) you wrote to wrangle data then run PCA, and to create a **professional looking** PCA biplot that appears (with a caption) in your knitted HTML
3. A brief summary (nicely formatted and professionally written bulletpoints are fine) highlighting some major takeaways from your PCA that can be gleaned from the biplot

Here are two datasets that you **can** work with, either in whole or in part (i.e. you are welcome to wrangle in order to limit observations and/or variables, just make sure to describe how you are limiting things in your project), or you are encouraged to find and/or choose your own data that may benefit from exploration by PCA:

- *Miscellaneous environmental and climatic variables (country-level)*
    - The file: world_env_vars.csv
    - Compiled and provided by @zander_venter on Kaggle, described as: "This data is all acquired through Google Earth Engine (https://earthengine.google.com/) where publicly available remote sensing datasets have been uploaded...Most of the data is derived by calculating the mean for each country at a reduction scale of about 10km."

- *Food nutrient information for raw fruits and veggies from USDA (National Nutrient Database, now FoodData Central):*
    - The file: usda_nutrients.csv
    - Note: If you use this dataset, you'll probably want to narrow the scope of your PCA (e.g. by limiting the food types and/or nutrients explored)

Submit your Task 2 knitted html on GauchoSpace.

Task 3: Complete this Shiny app description spreadsheet for your ESM 244 Term Project

The guidelines and grading rubric for your Shiny app term project, which can be completed in groups of UP TO 3 people, are HERE. For Task 3, you should decide what you want to create your app on (which may require meeting with your teammates if you're working in a group), and add the requested information to THIS SPREADSHEET (see example entry for what is expected), including: a brief overview, where you'll get the data, the widgets you plan to include (this can change, but hopefully motivates you to think about it), etc. Each group should add their information to this spreadsheet by **5:00pm on Monday 1/25/2021**.

If you are looking for group members to join your project, or for a project to join, feel free to post / share your project ideas on the #shiny-apps channel in the 244 Slack workspace.

Task 4: Start building your personal website in R

Your second term project for ESM 244 is building a website with R. The rubric is found in the same document as above, HERE. A personal website featuring several data science projects is a powerful way to showcase your modern data science & coding skills with prospective employers, and is a great way to continue building and sharing a portfolio of your work.

For Task 4, follow along with THIS POST by Alison Hill to create and publish (with GitHub pages) a "skeleton" version of your personal website using the {distill} and {postcards} packages in R.

You do NOT need to add any actual content at this point. The **only thing** we will check for Assignment 1 is that your "skeleton" site is up-and-running, and that the Name on the postcard landing page is updated to your name. Beyond that, any customization is optional (though you might find making css updates for theming kind of fun…).

When you're done getting it set up, add your name and the link to your published skeleton site HERE.

What you'll submit for Assignment 1 (by 5pm PST Monday 1/25)

- Your knitted html for Task 1 (on GauchoSpace)
- Your knitted html for Task 2 (on GauchoSpace)
- Add your Shiny term project information (Task 3) to this spreadsheet
- Add the link to your skeleton website (Task 4) to this spreadsheet

END ASSIGNMENT 1