**ESM 244 Assignment 3**

**All parts due Friday, 2023-03-03 at 5:00pm PST**

*PCA, hierarchical clustering, Shiny App progress and sign-ups for presentations*

## What you’ll submit for Assignment 3 by the due date

* Task 1: 10 pts - Your knitted html (not Rmd) for Task 1 **(on GauchoSpace):**

a3\_task1\_lastname\_firstname.html

* Task 2: 10 pts - Your knitted html (not Rmd) for Task 2 **(on GauchoSpace):**

a3\_task2\_lastname\_firstname.html

* Task 3: 5 pts - update Shiny app! If necessary, update your Github repository URL on the [Shiny App Information Page](https://docs.google.com/spreadsheets/d/1yaFF_qHK4CRt9qBDwsnkS1etnTkMcGpyEHIMyvA5D8s/edit?usp=sharing). If your repository is private, add Casey (GitHub: oharac) as a collaborator to your Shiny app repo.
* Task 4: [**Sign up for a time slot**](https://docs.google.com/spreadsheets/d/1mFRLRN80Slj0q92hk0mQyRIvec44nYJBE9Cht986-9I/edit?usp=sharing) for presenting your Shiny App (Friday of week 10, or Tuesday of finals week)

## Task 1 (individual): Principal components analysis (coder’s choice)

For this task, we’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 (**optional:** also include a nicely formatted screeplot)
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 (and screeplot, if included)

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 clearly describe how you are wrangling the data in your project, including how you are dealing with any NA values). **Alternately, 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](https://drive.google.com/file/d/177JGe8dYL3LUq0IHThqLhqg907gQqLJy/view?usp=share_link)
  + 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*](https://fdc.nal.usda.gov/index.html)*):*
  + The file: [usda\_nutrients.csv](https://drive.google.com/file/d/1TcvL3D0Bxk0fj_ofRV14F6CrCq0PBLR-/view?usp=share_link)
  + 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 1 knitted html on GauchoSpace.**

## Task 2: Agglomerative hierarchical clustering (submit html on GS)

For Task 2, use hierarchical clustering by **complete linkage** to create a dendrogram showing multivariate clustering for water chemistry by ***site***.

To perform hierarchical clustering by **site**, you’ll want to make a data frame that has a single summary row per site (e.g. based on means from all observations at that site), *then* calculate the Euclidean distance before performing complete linkage agglomerative hierarchical clustering.

Link to the data: [sbc\_lter\_registered\_stream\_chemistry.csv](https://drive.google.com/file/d/1qabHiI1ZOuamfPX6hJdGvW3YCQvoSsE9/view?usp=sharing)

DATA & METADATA SOURCE:

* SBC LTER: Stream chemistry in the Santa Barbara Coastal drainage area, ongoing since 2000
* Creators: Santa Barbara Coastal LTER, & Melack, John M
* Citation: Santa Barbara Coastal LTER and J. Melack. 2019. SBC LTER: Land: Stream chemistry in the Santa Barbara Coastal drainage area, ongoing since 2000 ver 16. Environmental Data Initiative. <https://doi.org/10.6073/pasta/67a558a24ceed9a0a5bf5e46ab841174>.

NOTES:

* Make sure you convert -999 values to NA.
* **If you have a dataset that you are *more interested in exploring with hierarchical clustering*, you are welcome to use an alternative dataset (and alternate linkage method) of your choosing instead.**

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 create the dendrogram, and to create a **professional looking** dendrogram that appears (with a caption) in your knitted HTML.
3. **Optional:** rerun the dendrogram using a single linkage method, and include a tanglegram (in addition to your pro-quality dendrogram) to show how the different linkage changes the clustering.
4. A brief summary (nicely formatted and professionally written bulletpoints are fine) highlighting some major takeaways from your hierarchical clustering analysis that can be gleaned from the dendrogram(s).

**Submit your Task 2 knitted html on GauchoSpace.**

## Task 3: Shiny app term project

For this task, update your Shiny app. **Make sure to push your changes to Github!**

* Work on finalizing the Introduction/Summary/Overview page of your app
* Update so that all widgets and reactive outputs (tables, graphs, maps, etc.) are functioning correctly
* Start finalizing the little things (citations, formatting graphs/tables & figures), work on usability for an app user, make this nice & professional to look at

That’s it. Casey should be added to your app repo already, so will check progress. The idea is that by the time you submit this assignment, you should have the last week and a half to work mostly on making it awesome (adding bells & whistles, fine-tuning, theming, etc.).

## Task 4: Sign up for a Shiny demos session

During Week 10 and Finals week, we’ll have three 2-hour sessions during which each group will have 15 minutes to present your Shiny app (about 10 minutes of presentation, then a little time for questions). You’ll get more details later on about what you’re expected to present on, but for now, **find a session that works for your group and sign up.** You are expected to stay for the 2-hour window to watch presentations from other groups on their apps (there is some flexibility - i.e. if one of your group members can only attend for 90 minutes of the session, you can still show up and they can attend 30 minutes of another session).

The sessions are scheduled for:

* Friday, 3/17, 9:30-11:30
* Friday, 3/17, 1:30-3:30
* Tuesday, 3/21, 4:00-6:00

**Add your name(s)** [**HERE**](https://docs.google.com/spreadsheets/d/1mFRLRN80Slj0q92hk0mQyRIvec44nYJBE9Cht986-9I/edit?usp=sharing)**, along with the name of your Shiny app.**