ZOO 800

Homework Week 8

Submission instructions

Submit a single URL to a public GitHub repository on Canvas. Please make sure it works – i.e., that you can clone the repo as a project yourself. Be sure to indicate in the submission who is in your group. Submit a single URL for each group, but if you're not the one submitting the URL, submit a comment mentioning the name of the person submitting for your group.

Problem

Online data repositories contain a wealth of information. However, it's easy to go wrong when working with someone else's data. An important second step after you've figured out how to query the data you need is to conduct exploratory data analysis (EDA). This is not hypothesis testing but an open ended exploration of characteristics and patterns in the data. These patterns may influence your analysis and be important to understand. For example, in a recent analysis of fisheries data in my lab, we realized that the apparent temporal patterns we saw when averaging across populations were confounded by the fact that a large and nonrandom portion of the data ended in 2000. So, analysis of trends after that time were essentially being conducted on a different data set. Other relevant changes in time series might include the sampling intensity, locations of sampling, and time of year that sampling is conducted.

Objective 1

A. Find an online dataset that is relevant to your research focus. We've used examples from GBIF and EDI. Fine to use those, but also feel free to branch out.

B. Write or modify code to connect to it remotely and query the dataset. No fair just downloading it and reading it into your workspace.

Objective 2

Conduct EDA of the data that you've downloaded. I know we haven't talked about this in class, but the basic concept is straightforward. See the link below for some examples, but also try to think more broadly about the kinds of issues mentioned in the problem statement above.

https://www.epa.gov/caddis/exploratory-data-analysis

A. Create at least three figures or tables describing the data set

B. In code comments and/or figure annotations, highlight any outliers or changes/imbalances in the dataset that might impact your interpretation of the data. Figure annotations will likely make this easier than a comment referring to row 325,678 of your data frame, for example.