Module 3: Data Reproducibility Lab

Outcomes

Through this lab, students should be able to:

- Understand the purpose of data citation and cite the data source correctly
- Use information from documentation associated with an established and well-curated data set to reuse the data set and answer specific research questions
- Recognize the importance of good file naming convention and good folder structure and implement their own file naming rules and folder structure considering the data context and sharing needs
- Document data processing and analysis concisely with the sharing purpose in mind
- Identify errors and potential issues in peers' data management practices

Lab Design

Students will be divided into two Group A and Group B and form teams of 2-3 members within each group. Students in Group A and B will answer difference research questions using the given data sets.

Each team of students should:

- devise at least one figure or summary table in an Excel file
- create a data dictionary (txt file) to
    - describe general information of the data set
    - describe variables used in their data sheets
    - describe how variables evolved in the process
- create a summary (Word file) to
    - answer the research question with their analysis result
    - explain what's included in the folder they share
    - explain their data processing and analysis methods
    - cite the data source properly

Group A and Group B will exchange their results, data files, and required documentation; and review the other team's work based on a rubric provided.

Suggested datasets:
http://datadryad.org/resource/doi:10.5061/dryad.fj974
http://nsidc.org/data/NSIDC-0131

Instructions for students

Purpose: This lab aims at management of your data analysis process. You should focus on documenting your process well so that others can reproduce what you did. The scientific value of your answer and statistical validation are not the focus of this practice.

In this assignment, you will download a set of data from a repository and answer a research question through analyzing the data set or selected portion of the data set. Please use the documentation and metadata provided in the repositories to help you understand the scope and details of the dataset. Pay attention to how the data file structure, folder organization and variable definitions are described in the documentations. You are required to turn in your write up, data files, and documentations on your analysis.

1. Organize the data into an Excel file and incorporate any calculations you used to answer the questions into the spreadsheets.
2. Name your write up, excel file, and other documents appropriately so that your classmates can identify the content of the files and which group they belong to.
3. Organize your files into a folder. This folder should include the raw data file and documentations, your write up answering the research question, your excel file showing the data analysis process, and a README.txt file. The README.txt file should explain your project, cite the data source, list what's included in your folder, and provide a data dictionary for your spreadsheet.
4. The data dictionary (SEE LECTURE) should include, where appropriate, the following information for each variable:
    a. Variable name
    b. Variable meaning
    c. Variable units
    d. Variable format
    e. Variable coding values and meanings
    f. Known issues with the data (systematic errors, missing values, etc.)
    g. Relationship to other variables
    h. Null value indicator
    i. Anything else someone needs to know to better understand the data

Students in Group A and Group B will exchange their results and review the other team's work based on the Data Lab Documentation Rubric. Can you reproduce your peers' work based on the documentations they provided?

Group A

Please download the data package Life History Profiles for 27 strepsirrhine primate taxa generated using captive data from the Duke Lemur Center deposited in the DRYAD data repository: http://datadryad.org/resource/doi:10.5061/dryad.fj974

Research question:
Using the weight file, pick two species and at least two subjects per species, and test the hypothesis that adult males gain weight faster than adult females.

Guided questions:

1. How will you tell an adult subject from a juvenile?
2. What unit of time are you going to use?
3. You will need to create a plot - what measures need to be included? In the plot, what parameter(s) could represent the weight gain rate?

Group B

Please download the Methane and Carbonyl Sulfide Analysis of Siple Dome Ice Core Subsamples from the National Snow & Ice Data Center at:
http://nsidc.org/data/docs/agdc/nsidc0131_saltzman/index.html

[Tip: import text file into Excel, use File/Import… from the menu bar.]

Research question:
Was the concentration of Carbonyl Sulfide (OCS) in the preindustrial atmosphere significantly lower than the atmosphere of present day? (The global average OCS mixing ratio is approximately 500 pptv ± 50 pptv [1].)

Guided questions:

1. Which subset of data has the concentration of Carbonyl Sulfide (OCS) in the ice cores?
2. Which column(s) of the data would indicate the ages? Can you convert them into calendar years when the air were trapped based on the gas age (years) column? (Assume the gas age was determined in the year when the data was first published, i.e. 2002).
3. Make a plot to show how the concentration of the Carbonyl Sulfide fluctuated over the years observed from this set of data. Include the error bar in the plot if you know what it means and how to plot it in Excel.
4. What's the average concentration of the Carbonyl Sulfide during the time range measured? How does it compare to the present day? Explain any assumptions you made to reach your conclusion.

[1] Chin, M., and D. D. Davis, A reanalysis of carbonyl sulfide as a source of stratospheric background sulfur aerosol, *J. Geophys. Res.*, 100, 8993–9005,1995.