Assignment 2 - Pipes and csvkit - Due 9/15, 7pm

Use a jupyter bash notebook to demonstrate building up pipelines and using csvkit.

Objectives: Gain experience building up pipelines step by step, and narrating your work as you go. Begin to integrate tabular data into pipelines. Demonstrate an understanding of possibilities and limitations of these tools.

Use a single bash notebook in jupyter to complete both parts and document your work. Describe the steps you take and answer any questions by writing prose in markdown cells.

Part 1 - text and a pipeline

This first part is based on the review question at the end of Topic 7 - Finding Things in the Software Carpentry lesson on the Unix shell.

- Download the text of Alcott's Little Women from Project Gutenberg (use Plain Text UTF-8)
- Build up a pipeline, step by step, to count mentions of Jo, Meg, Beth, and Amy, one at a time, and do so
- Use a *for loop* to count mentions for all four at once
- Choose another popular text from Gutenberg and apply the same loop with new character names

Part 2 - xls, csv, and csvkit

The excellent csvkit is described in *Data Science at the Command Line* and has good documentation on its own site. Use its `in2csv`, `csvcut`, `csvlook`, `csvsort`, and `csvstat` tools along with other commands you've learned to do a brief summary of a dataset from data.gov.

- Find and select a dataset from data.gov that interests you and is available with an XLS download option, and download the XLS file
- Using csvkit commands and pipelines (do *not* use Excel to convert it to CSV), explore the data: how big is it, what types of values are contained, and which variables interest you?
- Use `csvlook` and `csvstat` to generate basic summary statistics about those variables
- What do the summary statistics tell you?

Give the notebook a clear name like "`assignment-02-mylastname`". Acknowledge any assistance you received. Download as PDF, and download as .ipynb as well, then zip these together. Upload your zipfile to blackboard by the deadline.

Deadline: Tuesday, September 15, 7pm (before class begins)