CLI Data Analysis

Rob Lass
Data Philly
11/17/2015

Background

- Rob Lass <rob.lass@gmail.com>
- Data Scientist at AWeber Communications
- Adjunct Professor at Drexel University

Original Inspiration

- Partially avoiding problems with RSI from poor mouse ergonomics.

- Partially about being cool (yes, I’m a nerd).

Tech Background

- Tried installing Red Hat (4.1?) in 1997. Took about a week to get X working.

- Went to college in 1998, started using screen, slrn, naim, mutt, vim, xmms-cli in fluxbox. Tried sticking with lynx, failed.

- Switched to OSX at home in 2008, switched at work in 2013.

Continued Motivation

- It’s just faster.
- Don’t have to spin up servers.
- Don’t have to navigate any poorly designed UIs.
- Can string together tools that do a job really, really well.

Relevant Quotes

- This is the Unix philosophy: Write programs that do one thing and do it well. Write programs to work together. Write programs to handle text streams, because that is a universal interface. *Doug McIllroy*

Tools

- awk
- cat
- csvkit (csvcut, csvgrep, csvjoin, csvsort, csvsql, csvstack, csvstat)
- curl
- grep
- head
- parallel
- scrape
- shell
- tail
- wc

Mac or Linux == Cool

- If you run Windows, install cygwin, maybe?

- If you run something else, you're probably crazy enough to figure it out yourself.

https://www.cms.gov/openpayments/

Sometimes, doctors and hospitals have financial relationships with health care manufacturing companies. These relationships can include money for research activities, gifts, speaking fees, meals, or travel. The Affordable Care Act requires CMS to collect information from applicable manufacturers and group purchasing organizations (GPOs) in order to report information about their financial relationships with physicians and hospitals. **Open Payments** is the federally run program that collects the information about these financial relationships and makes it available to you. View the summary data dashboard for an overview of the published data.

Payments Listed

- Types:
  - Research payments
  - “Supplemental” payments
  - Physician ownership details
- Self reported by industry, reviewed by physicians
- Disclaimer: Money != Bad

Basic Stuff

- cat: print a file to stdout
- grep: regex matching
- wc: count the number of words / lines / characters
- head/tail: print the first/last few lines

awk and sort

- AWK
  - Match patterns and print modified versions of input file.
    - I mainly use awk for stripping out individual fields from a file.
  - Great for stealing templates from: http://www.pement.org/awk/awk1line.txt (or just google “awk one liners”)

csvkit

- Toolkit for dealing with csv files.
- Probably not installed by default on your machine.
- Try: *pip install csvkit*

$156,388,338.20?

ICU Medical

Hospital supply company specializing in leakproof intravenous systems.

Headquarters: San Clemente

Founder, chairman and CEO: Dr. George Lopez

Employees: 1,700

Top-selling product: Clave, a one-piece, needleless intravenous connection device (38 percent of revenues)

Gregory Piskun

- Received $21,733,719 in 2014, second highest in value - investment (first was ungoogleable).
- Covidien Sales LLC “paid” him “$21M” for “promotional speaking”.
- Why all the quotes? Paid in shares:
  - 4,213,860 Class A Units
  - 3,410,981 Class E Units
  - HET Systems LLC,100000000352

Teresa De Marco

Laennec Young Clinician Award

Abstract 3080: A 22-Year-Old Female With Non-Hypertrophic Hypertrophic Cardiomyopathy

Rahul Sakhuja[1]; J. Eduardo Rame[2]; Colleen Brown[2]; Karen Ordovas[3]; Philip Ursell[4]; Teresa De Marco[5]

[1] Massachusetts General Hosp, Boston, MA

[2] Univ of California, San Francisco, San Francisco, CA

[3] Univ of Californa, San Francisco, San Francisco, CA

[4] Univ of California, San Francisco, San Francisco, CA

[5] Univ of San Francisco, California, San Francisco, CA

- Professor at UCSF.
- Not much else online, who is giving her money?

???


ENTER

More Information

- http://datascienceatthecommandline.com
- Google “awk one liners” or “sed one liners”.
- http://csvkit.readthedocs.org

Shoutouts

- Jeroen Janssens, Data Science at the Command Line (book & website)
- Chris Cera and Sven Guckes
- My Wife