Overview of different tools and resources for data analysis

(v1.5)

Oscar Torres-Reyna

otorres@princeton.edu

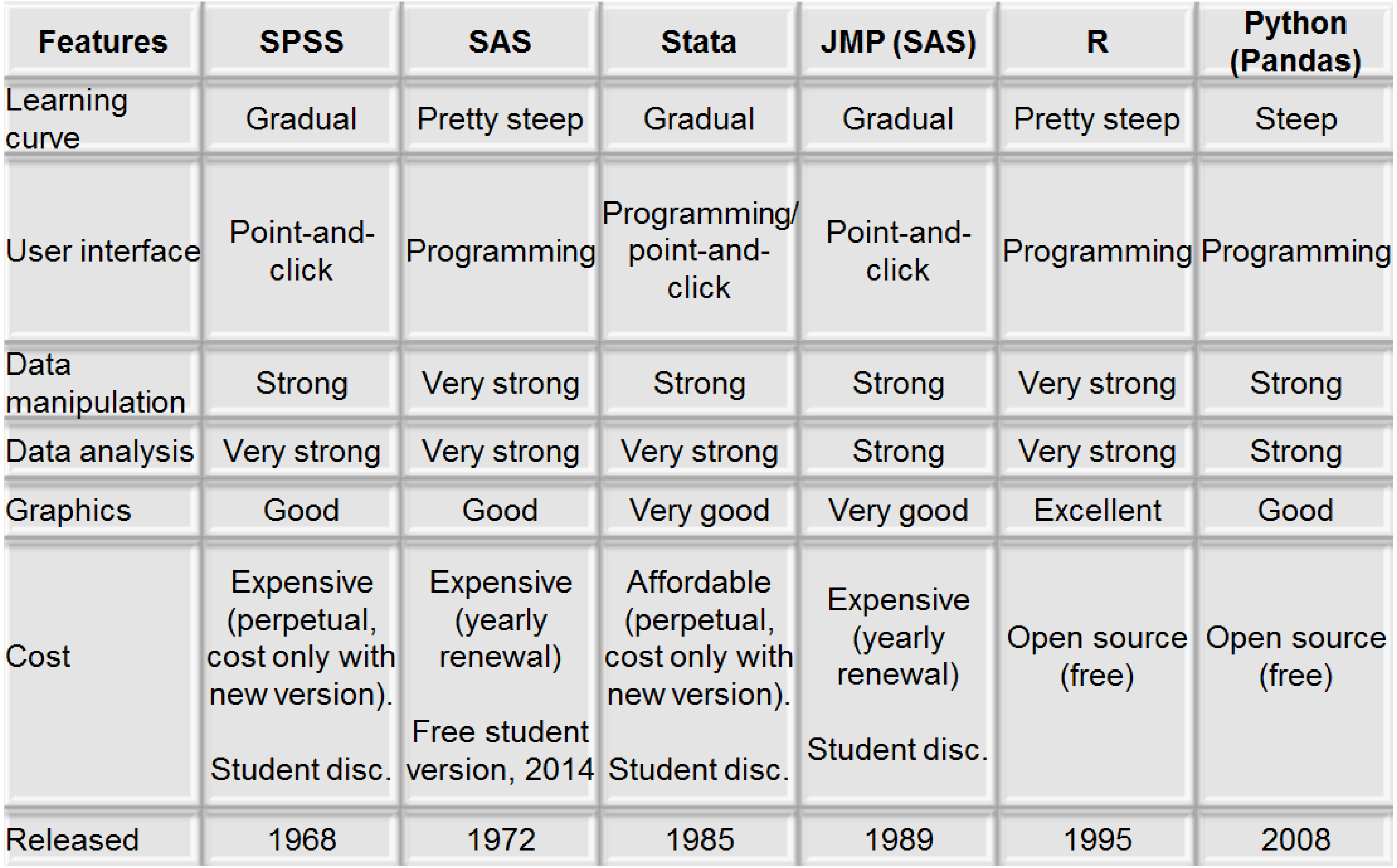
January 2019

http://dss.princeton.edu/training/

This document was presented at the
2019 Research Data Management Workshop
Lewis Library
January 28 - February 1, 2019

PRINCETON
UNIVERSITY

Once data is properly collected, need software to organize, clean, prepare and analyze data.



Features	SPSS	SAS	Stata	JMP (SAS)	R	Python (Pandas)
Learning curve	Gradual	Pretty steep	Gradual	Gradual	Pretty steep	Steep
User interface	Point-and-click	Programming	Programming/ point-and-click	Point-and-click	Programming	Programming
Data manipulation	Strong	Very strong	Strong	Strong	Very strong	Strong
Data analysis	Very strong	Very strong	Very strong	Strong	Very strong	Strong
Graphics	Good	Good	Very good	Very good	Excellent	Good
Cost	Expensive (perpetual, cost only with new version). Student disc.	Expensive (yearly renewal) Free student version, 2014	Affordable (perpetual, cost only with new version). Student disc.	Expensive (yearly renewal) Student disc.	Open source (free)	Open source (free)
Released	1968	1972	1985	1989	1995	2008

Source: http://dss.princeton.edu/training/RStata.pdf

PRINCETON UNIVERSITY

SPSS = Statistical Package for Social Sciences

SAS = Statistical Analysis System (original name)

Stata = Syllabic combination of the worlds '**sta**tistics' and 'da**ta**'

JMP = **J**ohn's **M**acintosh **P**roject (original name), pronounced 'jump'

R = From creators **R**oss Ihaka and **R**obert Gentleman and in line with the programing language S

Python Pandas = **P**ython **an**d **da**ta analy**s**is (some suggest a syllabic combo from **pan**el **da**ta)

Sources: Google, Wikipedia, Python sites

- Which package to use depends on your programming skills and willingness to learn a new software.
- Stata is easy to use and ideal for econometric analysis and survey research. Very little programming is required.
- R is ideal for those with, at least, intermediate statistical analysis skills and some programing experience.
- Those with programing skills needing to perform data analysis Python might be the way to go.
- The best package is the one you know.

- Text analysis, web scrapping, machine learning, data mining, neural networks → Python or R.
- Simulations → Matlab (free version *Octave*).
- PU software licensing → oitstore@princeton.edu
- If the research requires the analysis of different types of data: numeric, text, images, video, audio, surveys, or web sources, and no programing required, proprietary software like NVivo or Atlas.ti are an option (qualitative data analysis).
- Student license for Nvivo or Atlas.ti available:

-Atlas.ti https://atlasti.com/students/

-NVivo: https://www.qsrinternational.com/nvivo/products?pm=Student

PRINCETON UNIVERSITY

- R is a programing language used for data manipulation, data analysis, and data visualization.
- Can read almost any type of data in ASCII (*.csv, tab-delimited, fixed) or proprietary format (Stata, SPSS, SAS, Excel).
- Flexible statistical analysis capabilities.
- Strong publication quality capabilities.
- High degree of customization and strong community support that has developed almost 17,000 packages.

- R is available at the Comprehensive R Archive Network (CRAN) https://cran.r-project.org/
- CRAN Task Views provide a list of packages by subject area: https://cran.r-project.org/web/views/
- R documentation can be found in this site: https://www.rdocumentation.org/
- RStudio is a user-friendly interface to R: https://www.rstudio.com/
- Data visualization with –`ggplot2`: https://ggplot2.tidyverse.org/reference/

Where to get help?

- Google: https://www.google.com/
- Data and Statistical Services tutorials: https://dss.princeton.edu/training/
- UCLA: https://stats.idre.ucla.edu/
- Stackoverflow: https://stackoverflow.com/questions/tagged/r
- StackExchange: https://stats.stackexchange.com/questions/tagged/r

Some suggestions…

- Put all your data in one folder.
- Before sending data to R/RStudio simplify the name of the variables as much as you can and keep a mini codebook for your reference.
- It is a good practice to use lower case in the code (functions and variables).
- Select a standard file name system to keep track of versions of your data.
- Keep a backup of original data files.

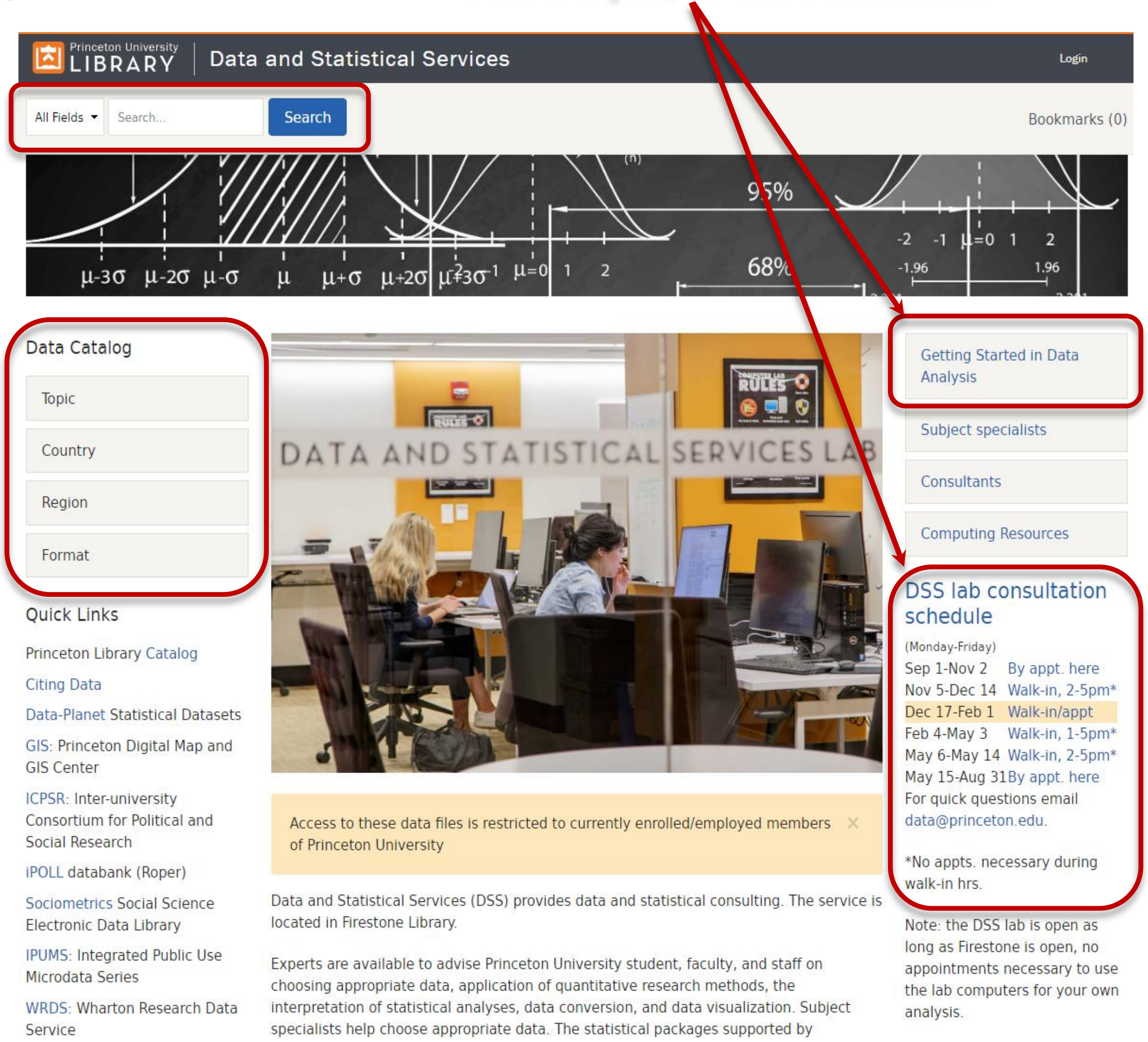
https://dss.princeton.edu/

Data analysis, software assistance

Finding data and access

Sources:
- Subscriptions
- Acquisitions
- Donations

Secure rooms for highly restricted data



Quick Links

Princeton Library Catalog

Citing Data

Data-Planet Statistical Datasets

GIS: Princeton Digital Map and GIS Center

ICPSR: Inter-university Consortium for Political and Social Research

iPOLL databank (Roper)

Sociometrics Social Science Electronic Data Library

IPUMS: Integrated Public Use Microdata Series

WRDS: Wharton Research Data Service

Access to these data files is restricted to currently enrolled/employed members of Princeton University ×

Data and Statistical Services (DSS) provides data and statistical consulting. The service is located in Firestone Library.

Experts are available to advise Princeton University student, faculty, and staff on choosing appropriate data, application of quantitative research methods, the interpretation of statistical analyses, data conversion, and data visualization. Subject specialists help choose appropriate data. The statistical packages supported by

DSS lab consultation schedule

(Monday-Friday)

Sep 1-Nov 2	By appt. here
Nov 5-Dec 14	Walk-in, 2-5pm*
Dec 17-Feb 1	Walk-in/appt
Feb 4-May 3	Walk-in, 1-5pm*
May 6-May 14	Walk-in, 2-5pm*
May 15-Aug 31	By appt. here

For quick questions email data@princeton.edu.

*No appts. necessary during walk-in hrs.

Note: the DSS lab is open as long as Firestone is open, no appointments necessary to use the lab computers for your own analysis.

Data:
- Management
- Cleaning
- Preparation
- Conversion
- Visualization
- Analysis

Statistical analysis:
- *From* descriptive statistics
- *To* advance models
- Model selection
- Output interpretation
- Presentation

Service
- Email
- Appointments
- Walk-ins
- Online tutorials
- Workshops (fall and by request)

Software: Stata, R, SPSS

data@princeton.edu

PRINCETON UNIVERSITY