

<https://go.ncsu.edu/cleanpy-slides>

Data Cleaning with Python

NC State University Libraries

Topics Covered in this workshop

- **Pandas library for data cleaning**
- Tasks including:
 - load in python libraries for data cleaning (pandas) and graphing (matplotlib)
 - read csv files into Python from an internet source
 - examine the first and last few rows of the data
 - delete duplicates
 - filter the data to create subsets
 - sort the data
 - group the data for plotting
 - drop variables from the dataset
 - create new variables
 - generate and save summary statistics for a dataset

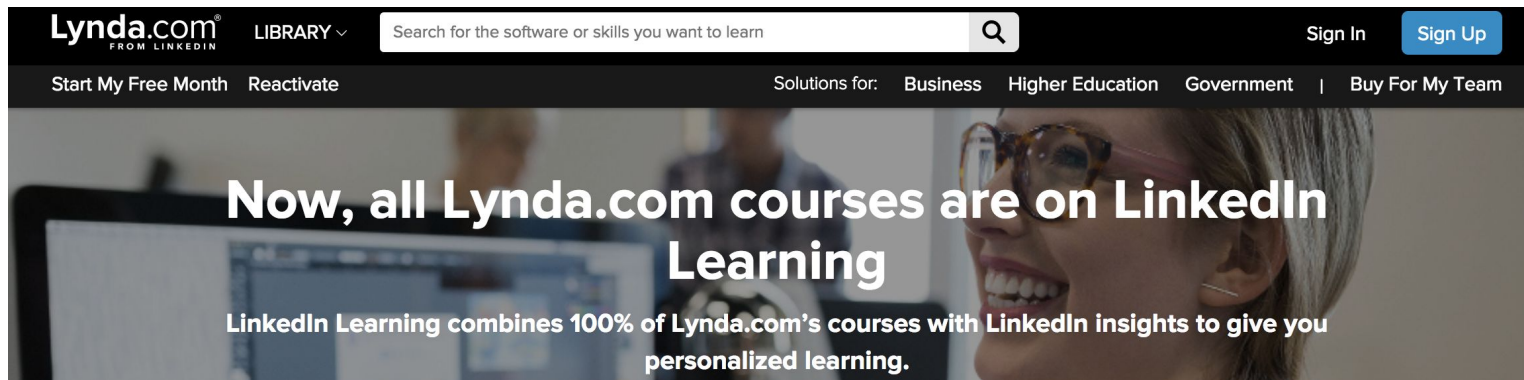
Continue learning in Python

Lynda.com video courses

Free to NCSU members with Unity ID

Sign-in instructions:

<https://www.lib.ncsu.edu/faq/can-i-access-lyndacom-at-ncsu-libraries>

A banner advertisement for Lynda.com. The top section is a dark navigation bar with the Lynda.com logo (with 'FROM LINKEDIN' in smaller text), a 'LIBRARY' dropdown menu, a search bar with the placeholder text 'Search for the software or skills you want to learn', and 'Sign In' and 'Sign Up' buttons. Below the navigation bar, there are links for 'Start My Free Month' and 'Reactivate', and a row of 'Solutions for:' categories: Business, Higher Education, Government, and Buy For My Team. The main part of the banner features a background image of a smiling woman with glasses. Overlaid on this image is the text 'Now, all Lynda.com courses are on LinkedIn Learning' in large white font, followed by 'LinkedIn Learning combines 100% of Lynda.com's courses with LinkedIn insights to give you personalized learning.' in a smaller white font.

Lynda.com
FROM LINKEDIN

LIBRARY

Search for the software or skills you want to learn

Sign In Sign Up

Start My Free Month Reactivate

Solutions for: Business Higher Education Government | Buy For My Team

Now, all Lynda.com courses are on LinkedIn Learning

LinkedIn Learning combines 100% of Lynda.com's courses with LinkedIn insights to give you personalized learning.

The Dataspace (at Hunt) and Data Point (at Hill)

Data visualization and statistical computing drop-in help

Python, R, Machine learning, SQL, Tableau, Java, visualization design, NVivo, etc.

3rd floor of Hunt Library, 7 days a week

1st floor of DH Hill library

<https://lib.ncsu.edu/spaces/dataspace>



Pandas (Python Data Analysis Library)

A python package designed for manipulating data for analysis.

Fast and free

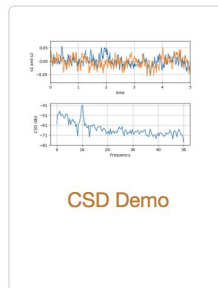
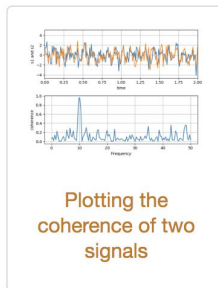
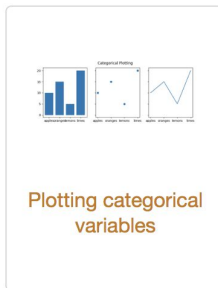
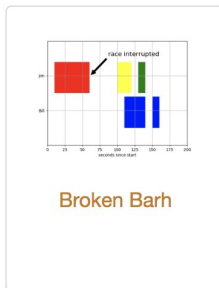
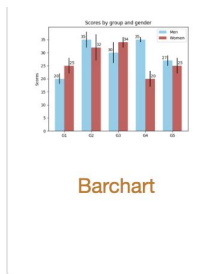
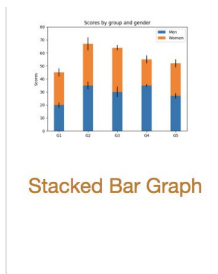
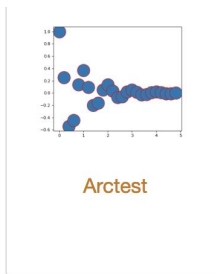
2 essential data structures:

- Series - one-dimensional labeled array
- Data frame - tabular 2-dimensional data structure (columns and rows)

User guide: http://pandas.pydata.org/pandas-docs/stable/user_guide/index.html

Matplotlib

Python library for making graphs with data.

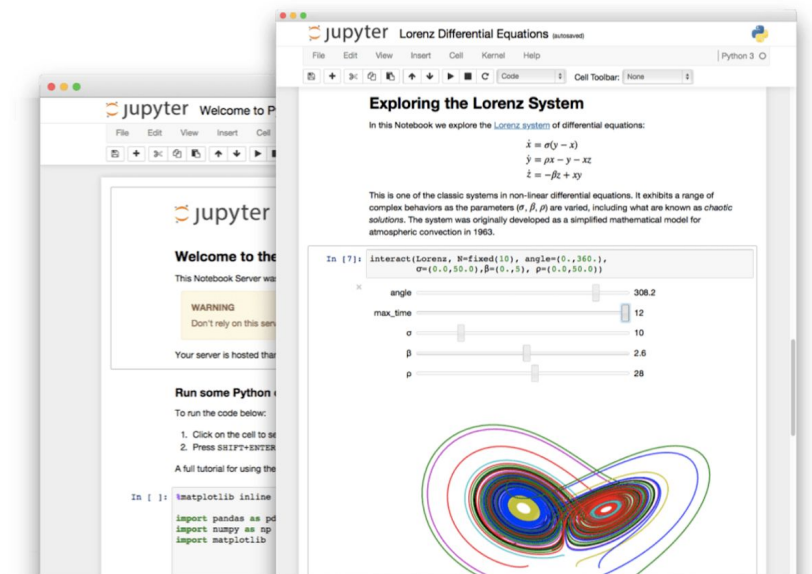


User guide: <https://matplotlib.org/users/index.html>

Jupyter Notebook - where we will type our code



[Install](#) [About Us](#) [Community](#) [Documentation](#) [NBViewer](#) [JupyterHub](#) [Widgets](#) [Blog](#)



The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

[Try it in your browser](#)

[Install the Notebook](#)

Google Colaboratory - where our project will live



Colaboratory

Frequently Asked Questions

What is Colaboratory?

Colaboratory is a research tool for machine learning education and research. It's a Jupyter notebook environment that requires no setup to use.

What browsers are supported?

Colaboratory works with most major browsers, and is most thoroughly tested with desktop versions of [Chrome](#) and [Firefox](#).

Is it free to use?

Yes. Colaboratory is a research project that is free to use.

What is the difference between Jupyter and Colaboratory?

[Jupyter](#) is the open source project on which Colaboratory is based. Colaboratory allows you to use and share Jupyter notebooks with others without having to download, install, or run anything on your own computer other than a browser.

Additional Resources

Pandas cheat sheet:

https://www.dataquest.io/blog/large_files/pandas-cheat-sheet.pdf

Pandas videos on Lynda.com: <https://www.lynda.com/search?q=pandas>

Online books @ NCSU Libraries:

[Python for data analysis : data wrangling with Pandas, NumPy, and IPython](#)

(McKinney, 2017)

[Python Data Analytics With Pandas, NumPy, and Matplotlib](#) (Nelli, 2018)

Step 1: Open the Google Drive folder at this link:

<https://go.ncsu.edu/cleanpy>

Step 2: Click the files to open them. Open in Colab button.



Step 3: File -> Save to Drive

Save a copy of each file to your Google Drive

Data Cleaning with Python

go.ncsu.edu/libeval