Intro to Data Analysis in Python

PyLadies Vancouver Workshop

July 7, 2018

Instructor: Jennifer Walker

Agenda

- Navigating the Python world as a data geek
- Jupyter Lab orientation + quick recap of Python basics
- Working with spreadsheet data
    1. Reading and summarizing CSV files
    2. Basic calculations and graphs
    3. Text data and messy / missing data
    4. Sorting, aggregation, and subsets
- Data visualization: a brief tour of the Python landscape
- Next steps, ideas, and inspiration

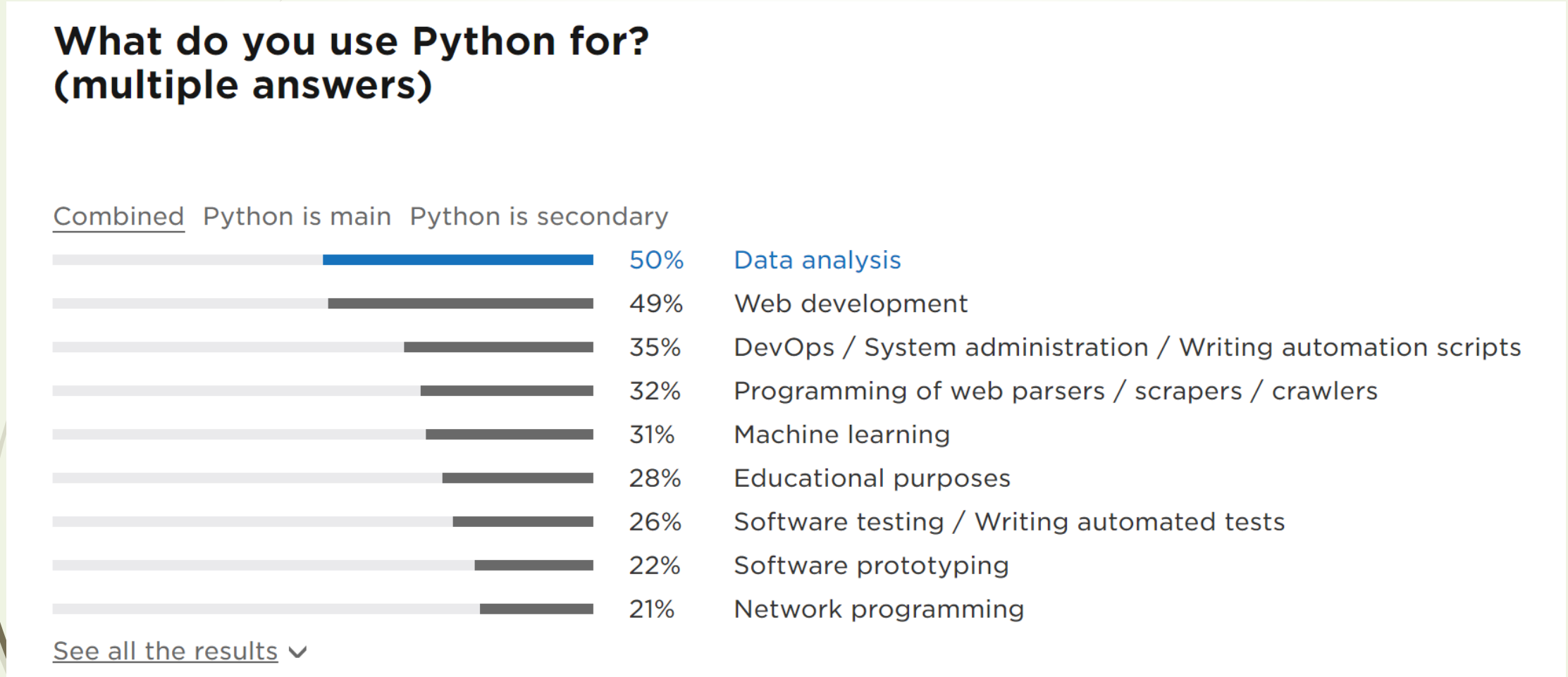
Navigating the Python world

... as a data geek

- Python is used in a huge variety of applications

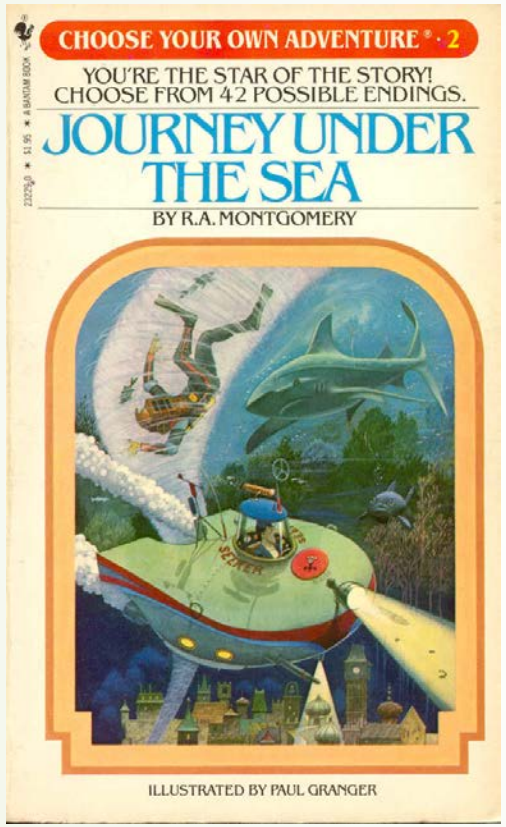
- It has recently become a powerhouse for data analysis

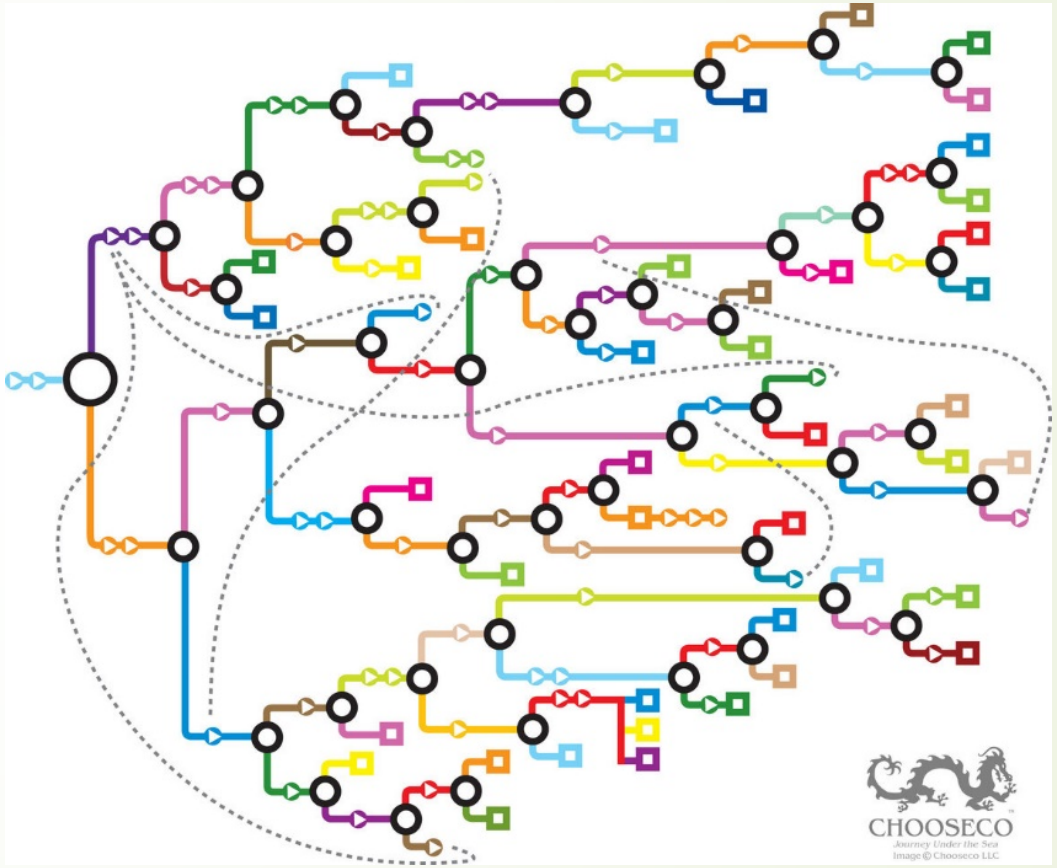
Python Developers Survey 2017



https://www.jetbrains.com/research/python-developers-survey-2017

Choose Your Own Adventure





https://www.atlasobscura.com/articles/cyoa-choose-your-own-adventure-maps

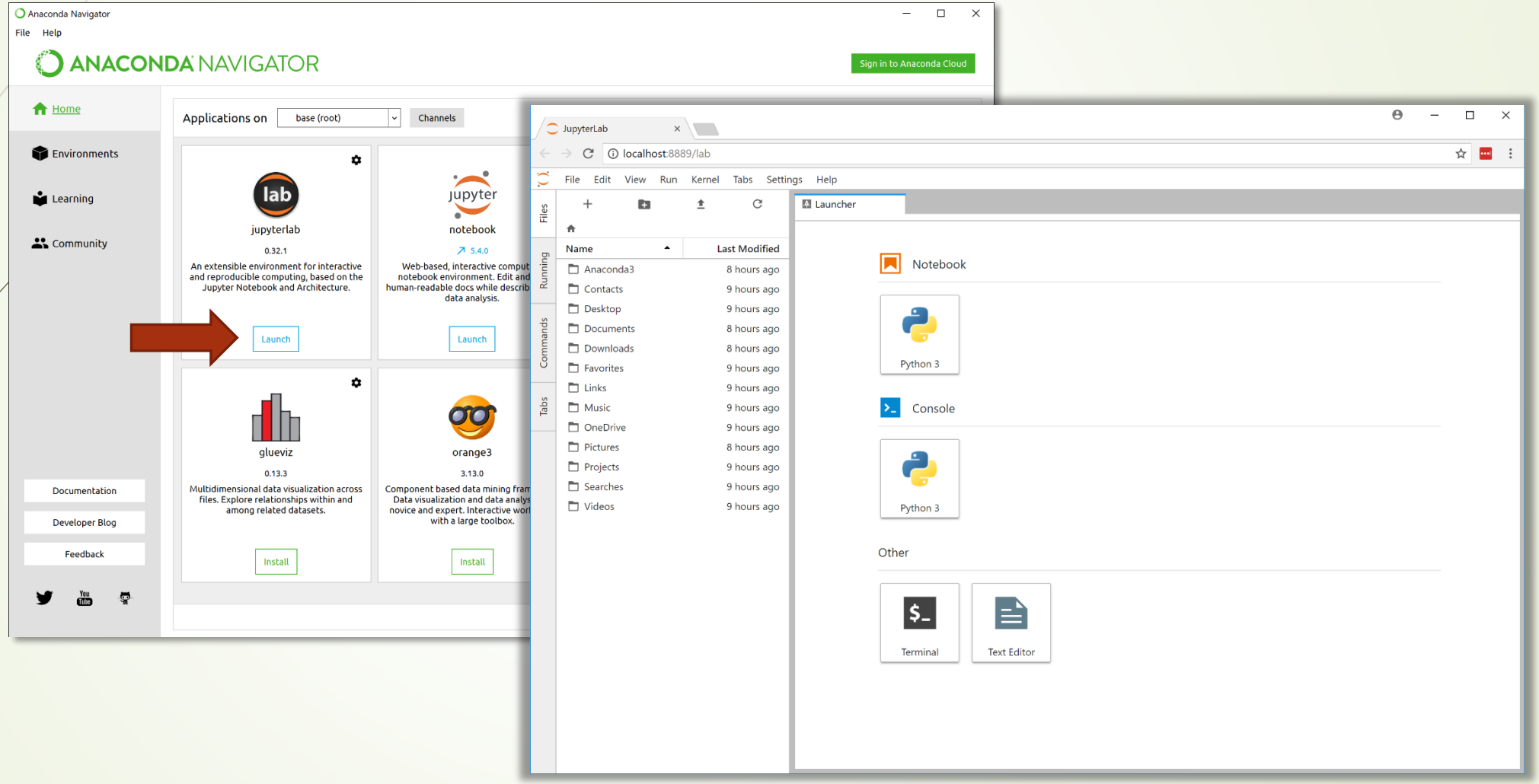
Jupyter Lab

- Development environment for working with data
  - Human-centered, interactive coding
- We'll be using Jupyter **notebooks:**
  - Code, graphs, formatted text, equations, etc. in a single document
    - Ideal for exploratory data analysis and for sharing your work with others who are interested in the entire workflow (step by step presentation of your code and results)
  - Uses an IPython kernel to run Python code (IPython = Interactive Python)
    - Many handy features for a much better interactive experience compared to the standard Python console
  - Also supports R, Julia, Perl, and over 100 other languages (and counting!)

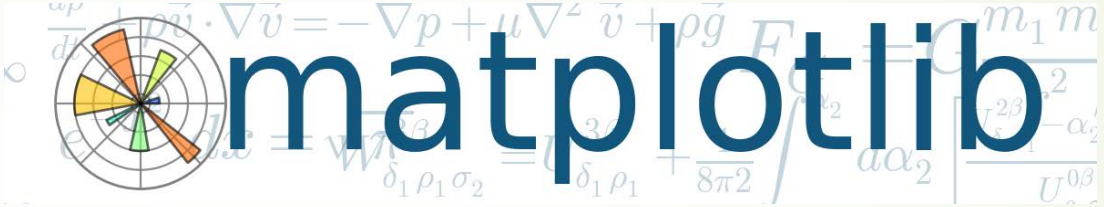
Example Jupyter Notebook

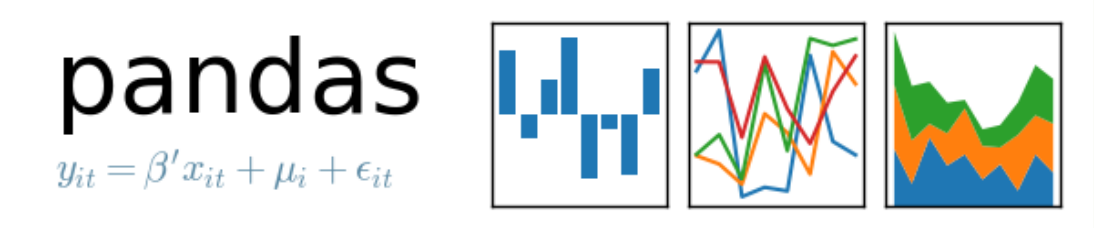
https://www.kaggle.com/arthurtok/generation-unemployed-interactive-plotly-visuals

Let's Get Coding!



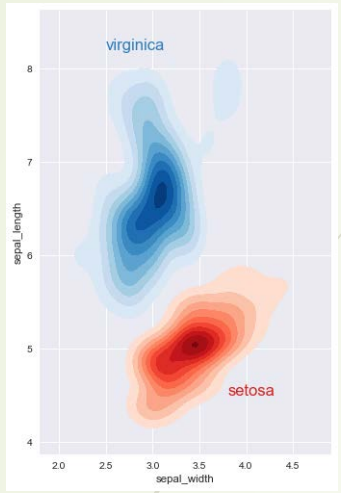
Data Visualization

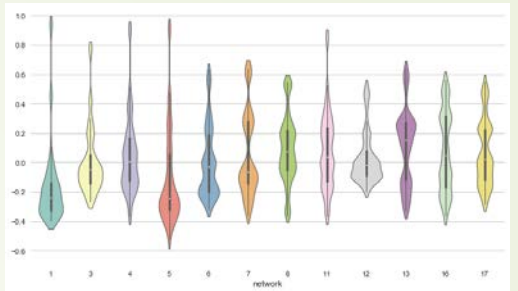


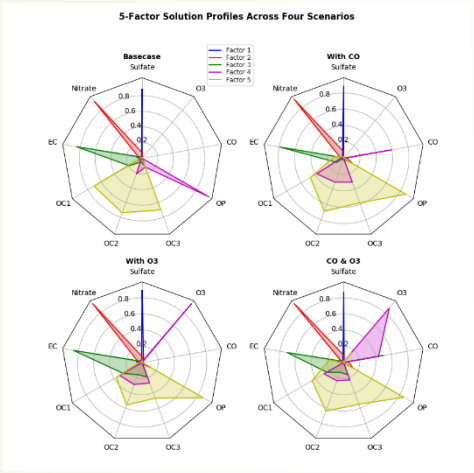


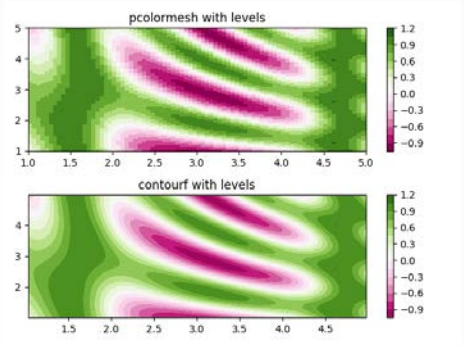
seaborn

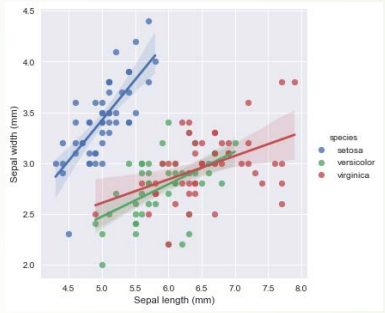
matplotlib & seaborn

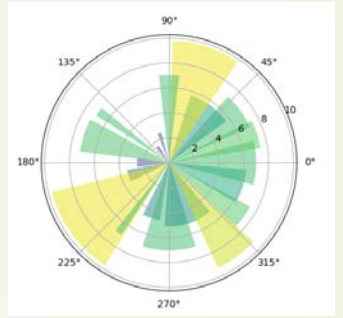


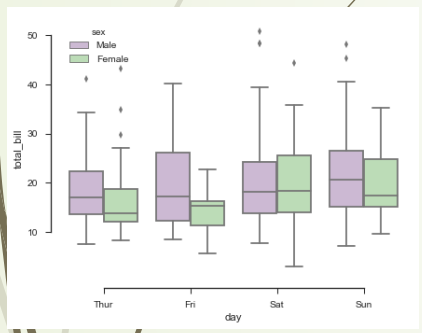


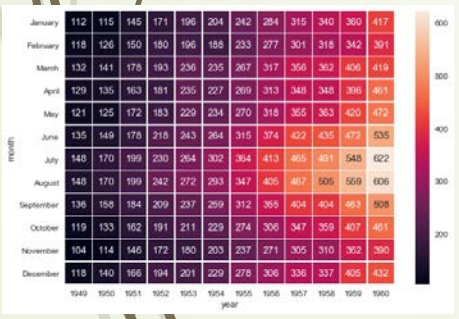


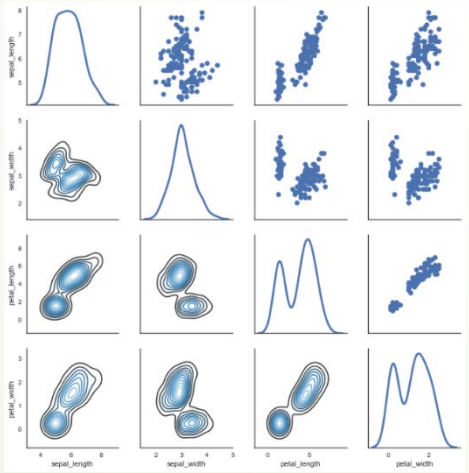


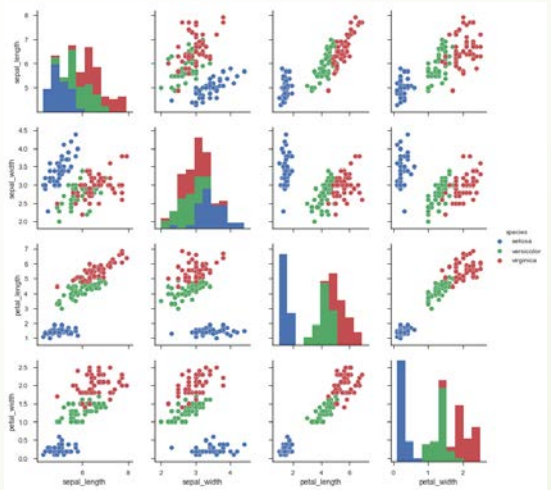


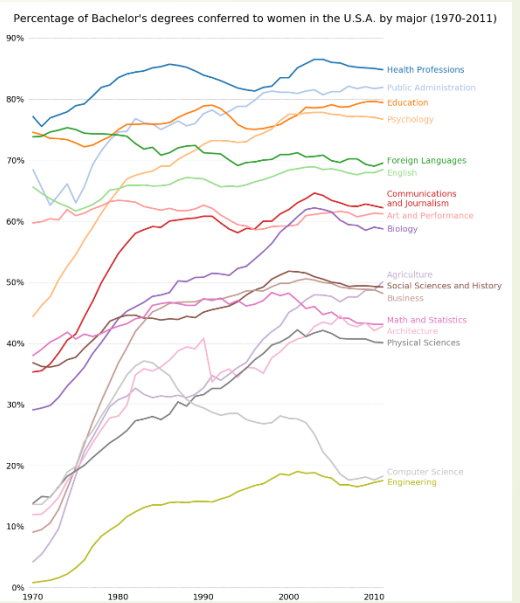




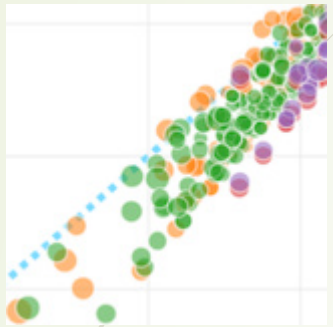


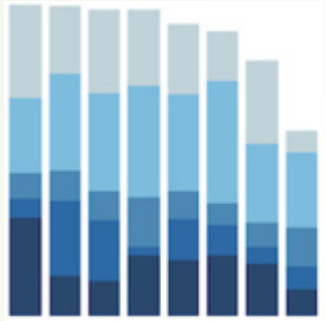


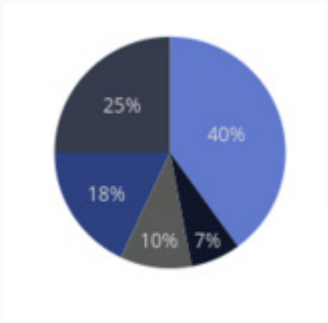


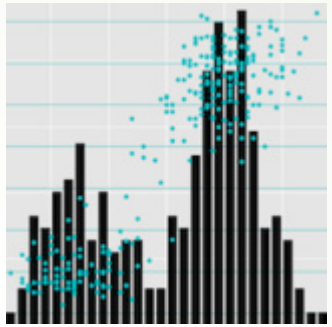


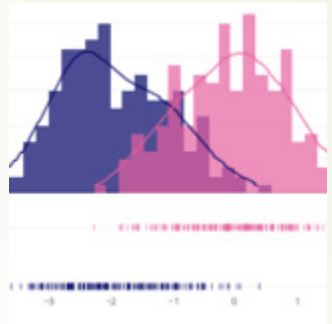
plotly

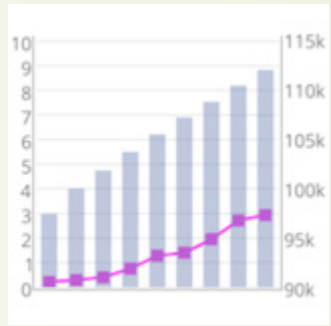


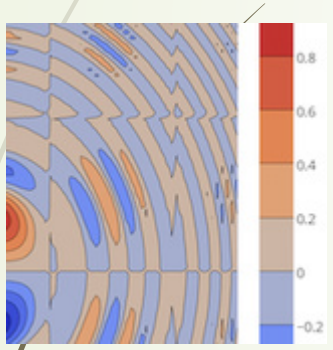


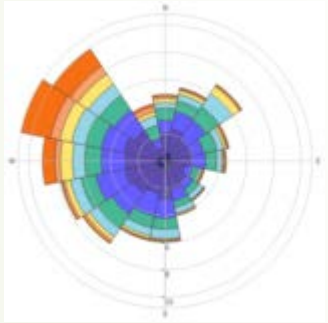

25%
40%
18%
10%
7%

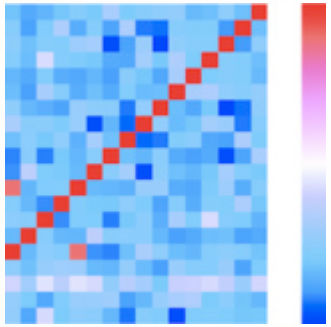





115k
110k
105k
100k
95k
90k

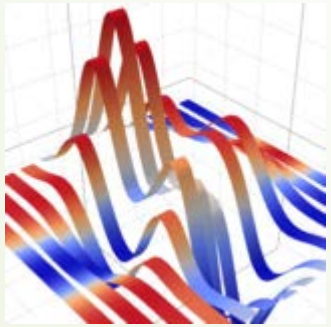

0.8
0.6
0.4
0.2
0
-0.2

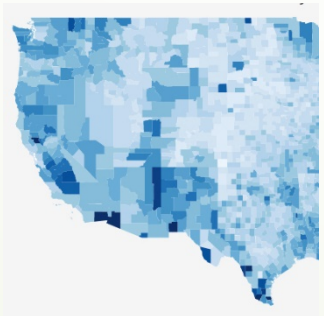




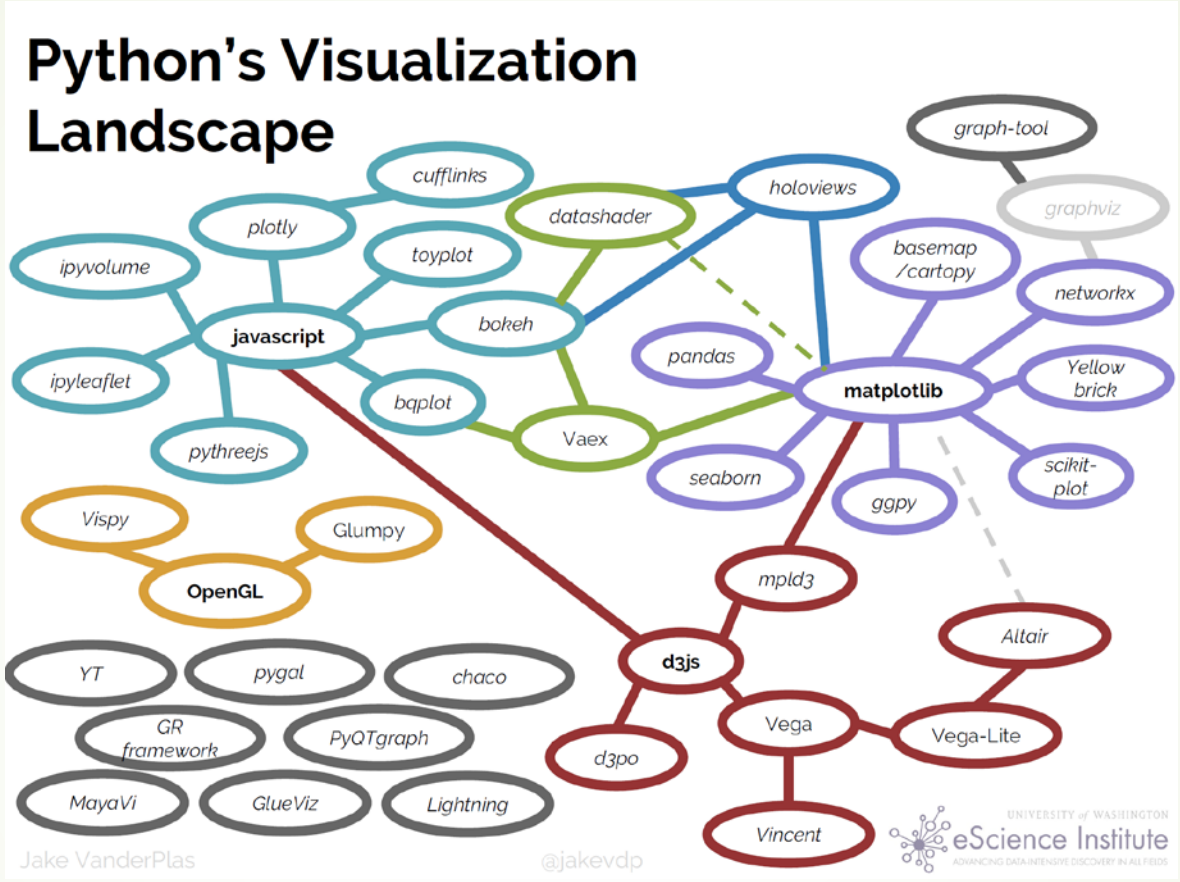

Aug 11
Aug 25
Sep 8
Sep 22
Oct 6


Funnel Chart






SAN FRANCISCO



- We've only looked at a tiny fraction of the visualization tools available in Python
- To learn about some of the other amazing visualization libraries and how they fit into the bigger picture, check out this fantastic talk by Jake VanderPlas from PyCon 2017: https://www.youtube.com/watch?v=FytuB8nFHPQ

Visualization Examples & Resources

- pandas: https://pandas.pydata.org/pandas-docs/stable/visualization.html
- matplotlib: https://matplotlib.org/gallery/index.html
- seaborn: https://seaborn.pydata.org/examples/index.html
- Interactive plots:
  - plotly: https://plot.ly/python/
  - bokeh: https://bokeh.pydata.org/en/latest/docs/gallery.html
- Maps:
  - plotly: https://plot.ly/python/#maps
  - cartopy: https://scitools.org.uk/cartopy/docs/v0.15/gallery.html
  - folium: http://folium.readthedocs.io/en/latest/

Where to go from here?

- Online resources and courses:
  - Data Carpentry: http://www.datacarpentry.org/python-ecology-lesson/
  - Data Camp: https://www.datacamp.com/courses/intro-to-python-for-data-science
  - Dataquest: https://www.dataquest.io/
    - Excellent blog with great tutorials and useful articles: https://www.dataquest.io/blog/
  - Kaggle: https://www.kaggle.com/learn/overview
    - Many more example Jupyter notebooks and tutorials: https://www.kaggle.com/kernels
    - Tons of datasets to play with: https://www.kaggle.com/datasets
  - plus Coursera, Udemy, and many others
- Book: Python for Data Analysis, by Wes McKinney. All data and code from the book is at https://github.com/wesm/pydata-book

Ideas & Inspiration

- PyData 101 – Jake VanderPlas
  - https://www.youtube.com/watch?v=DifMYH3iuFw
- Reproducible Data Analysis in Jupyter – Jake VanderPlas
  - http://jakevdp.github.io/blog/2017/03/03/reproducible-data-analysis-in-jupyter/
- Project Jupyter: From Interactive Python to Open Science - Fernando Perez
  - https://www.youtube.com/watch?v=xuNj5paMuow
- The Next Generation of Data Products – Hilary Mason
  - https://www.youtube.com/watch?v=OuRINNSDtlM

Thank You!