UC Berkeley Data8

JupyterHub 2016 mini-workshop

Background

- Foundations of Data Science (Stat/CS/Info 8)

    - has "connector" courses

    - for students with no knowledge of Statistics or Computer Science

- Fall 2015: ~100 students; Hub was on bare metal instance

- Spring 2016: ~480 students; Hub was on Azure with fallback to previous instance

- Fall 2016: ~480 students

- Future: 500-1500?

Deployment

- Copy Jess' compmodels deployment

  - https://github.com/compmodels/jupyterhub-deploy

  - Currently without nbgrader

- Tested on AWS; took a couple of months to learn about all of the components.

- Learning curve, large ecosystem.

  - ansible, docker, swarm, dockerspawner, swarmspawner, docker-py, jupyter, notebooks, node, nginx, web sockets.
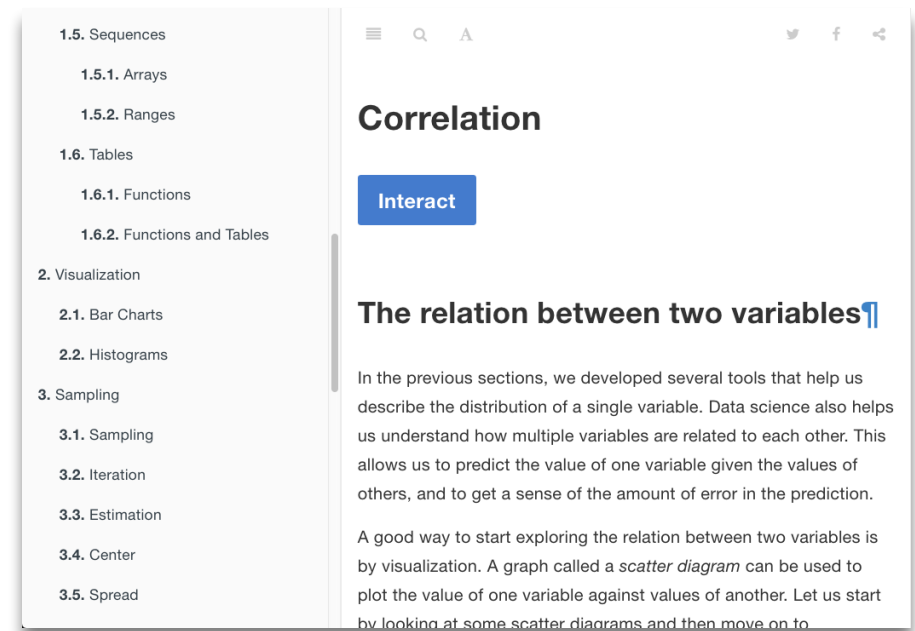
Versioning

- Versioning, dependencies, reproducibility…

  - A build process may break depending on when you execute it. Tools and libraries are always changing.

  - Be cautious about untagged "docker pull", "pip install". Wherever possible, embed git tags, docker image tags, conda and pip package versions, etc.

- Continuous Integration

- Reference deployments

  - https://github.com/jupyterhub/jupyterhub-deploy-teaching

  - https://github.com/jupyterhub/jupyterhub-deploy-docker

Content Distribution

How can faculty or graduate student instructors push notebooks or other files into student accounts?

- Interact: hub side service that pulls notebooks from github repos when students click on a textbook link.

- /hub/interact?repo=foo&path=a/b.ipynb

- Not all instructors are familiar with git.

- How to merge faculty changes with students' changes. Cell-centric contents manager? (Postgres)

- jupyter-drive: we can take advantage of UCB's Google Apps integration.

Correlation

Interact

The relation between two variables¶

In the previous sections, we developed several tools that help us describe the distribution of a single variable. Data science also helps us understand how multiple variables are related to each other. This allows us to predict the value of one variable given the values of others, and to get a sense of the amount of error in the prediction.

A good way to start exploring the relation between two variables is by visualization. A graph called a *scatter diagram* can be used to plot the value of one variable against values of another. Let us start by looking at some scatter diagrams and then move on to

http://www.inferentialthinking.com

Content Development

- Instructors want to add libraries to our system user container.

- How to test, given that users aren't root in the container?

- We have a derived system user container meant to be run on laptops with Docker Beta

- imagespawner: HTML form that allows the user to choose which container they want to run. Give instructors a choice between regular container and one where the user has elevated privileges

  - https://github.com/ryanlovett/imagespawner

Scaling

- How to support multiple courses

  - Data8 has "connectors"

- One hub with many courses or one course per hub instance?

- How to scale compute?

  - Spawners can do it: Swarm, kubernetes

  - Cloud APIs: AWS, Azure. Have spawner provision additional nodes on demand.

- How well does JupyterHub scale?

  - Start collecting data: statsd

Future Direction

- Keep going, knock on wood.

- Is our use case too large for JupyterHub? Should we use "NBaaS"?

- http://notebooks.azure.com

  - No need to create the infrastructure

  - Less control of server side

  - Requires each student have a cloud account

- http://mybinder.org

Links

- http://data8.org

- http://data.berkeley.edu

- http://inferentialthinking.com