INTRODUCTION TO DATA SCIENCE IN PYTHON

Week 1

Python

• Why Python? 1. It's easy to learn • Now the language of choice for 8 of 10 top US computer science programs (Philip Guo, CACM) 2. Full featured

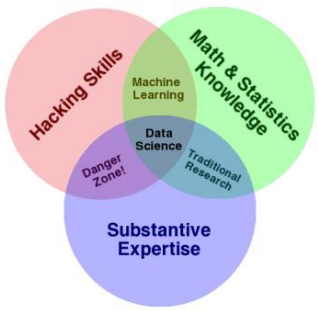
• Not just a statistics language, but has full capabilities for data acquisition, cleaning, databases, high performance computing, and more 3. Strong Data Science Libraries

• The SciPy Ecosystem

Course Outline

1. Prerequisite Python Knowledge
2. The pandas Toolkit
3. Advanced Querying and Manipulation with pandas
4. Basic Statistical Analysis with numpy and scipy, and project

Drew Conway perspective on data science:

– Hacking Skills – Math and Statistics Knowledge – Substantive Expertise • Other data science perspectives: – Skepticism, experimentation, simulation, and replication	

"50 Years of Data Science"

1. Data Exploration and Preparation
2. Data Representation and Transformation
3. Computing with Data
4. Data Modeling
5. Data Visualization and Presentation
6. Science about Data Science Week

The map() function

The `map()` function executes a specified function for each item in a iterable. The item is sent to the function as a parameter.

Week 2

Pandas
• Created in 2008 by Wes McKinney
• Open source New BSD license
• 100 different contributors

Stack Overflow
• http://stackoverflow.com
• Massive knowledge forum of python and pandas related content
• Free to join and participate in
• Heavily used by pandas developers instead of a mailing list

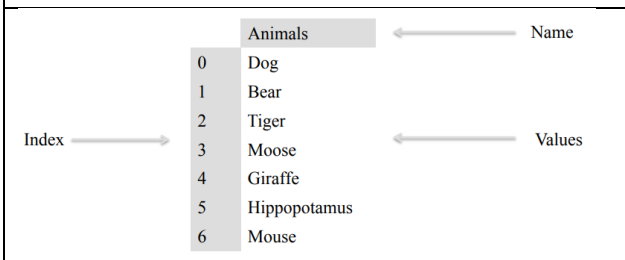
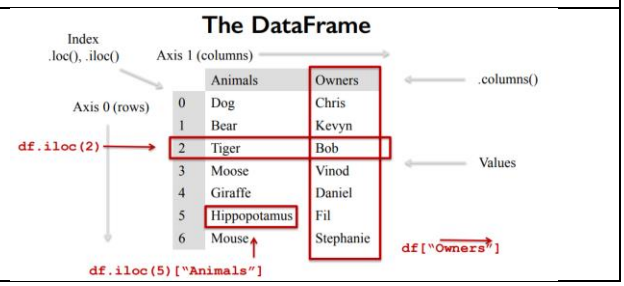
Books
Python for Data Analyst
Learning the Pandas Library *Matt Harrison)

Blogs
Planet Python • http://planetpython.org/

• Excellent blog aggregator for python related news
• Significant number of data science and python tutorials are posted
• Great blend of applied beginner and higher level python postings

Podcast
Data Skeptic Podcast

• http://dataskeptic.com/
• Kyle Polich, created in 2014 • Covers data science more generally, including: – Mini educational lessons – Interviews – Trends – Shared community project (OpenHouse)

The Series Animals	The DataFrame
	

NOC with Medals

Filer DF with bolean

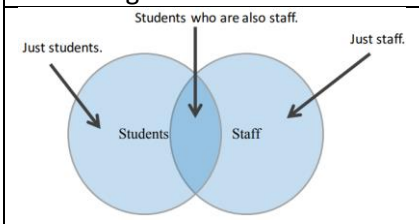
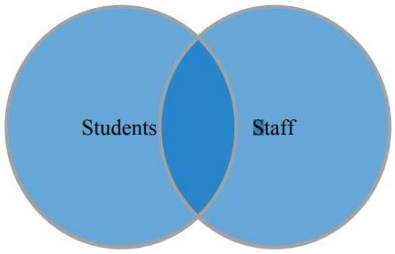
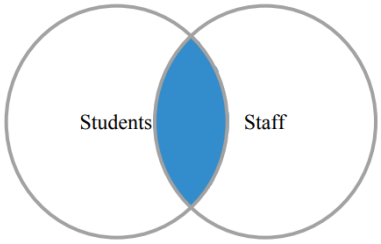
Week 3

Pandas Data Structures
• Series Object (1 dimensional, a row)
• DataFrame Object (2 dimensional, a table)
• Querying

– iloc[], for querying based on position
– loc[], for querying rows based on label
– Querying the DataFrame directly

• Projecting a subset of columns
• Using a boolean mask to filter data

Setting Data in Pandas
• To add new data
– df[column]=[a,b,c]
• To set default data (or overwrite all data):
– df[column]=2

Venn Diagram	Full outer join (union)	Inner join (intersection)
		

Chain Indexing:
– df.loc[“Washtenaw”][“Total Population”]
– Generally bad, pandas could return a copy of a view depending upon numpy
• Code smell
– If you see a][you should think carefully about what you are doing (Tom Augspurger)

(a,b) (c,d): Scales
• Ratio scale:

• units are equally spaced
• mathematical operations of +-/* are all valid
• E.g. height and weight

• Interval scale: • units are equally spaced, but there is no true zero
• Ordinal scale:

• the order of the units is important, but not evenly spaced.
• Letter grades such as A+, A are a good example

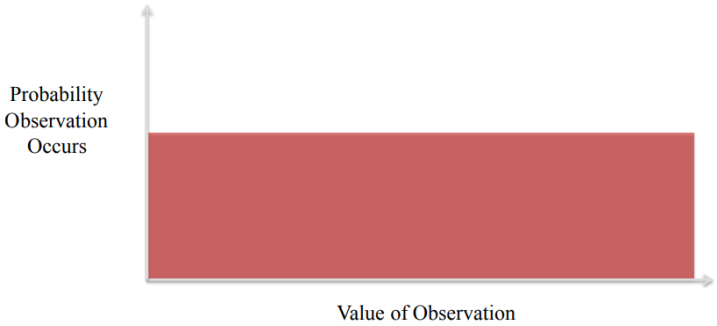
• Nominal scale:

• categories of data, but the categories have no order with respect to one another.
• E.g. Teams of a sport.

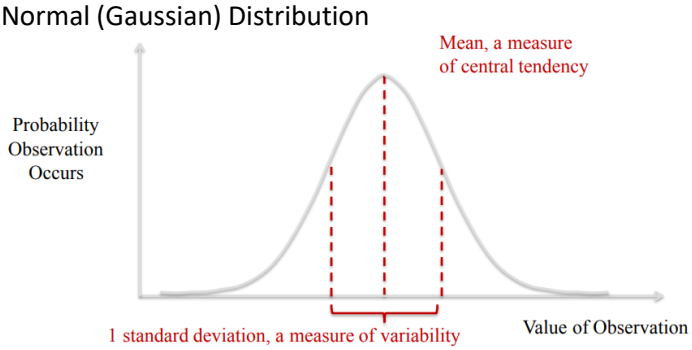
Week 4

- Distribution: Set of all possible random variables
- Example:
  - Flipping Coins for heads and tails
    - a binomial distribution (two possible outcomes)
    - discrete (categories of heads and tails, no real numbers)
    - evenly weighted (heads are just as likely as tails)
  - Tornado events in Ann Arbor
    - a binomial distribution
    - Discrete
    - evenly weighted (tornadoes are rare events)

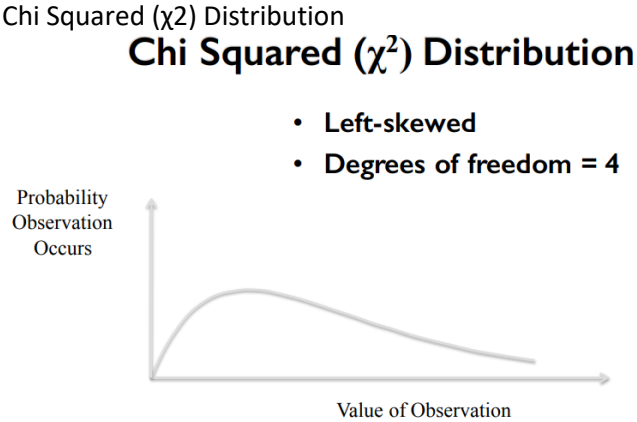
Uniform Distribution (Continuous)



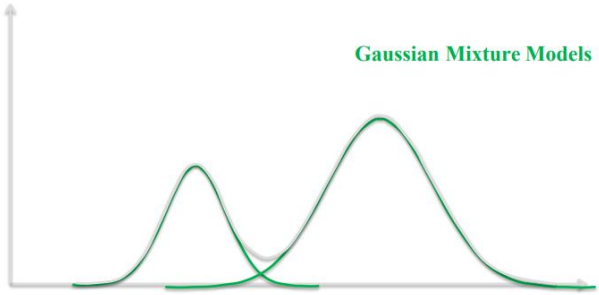
Normal (Gaussian) Distribution



Chi Squared (χ2) Distribution



Bimodal distributions



Probability and Statistics for Programmers
– Allen B. Downey – Available for free under CC license at: http://greenteapress.com/thinkstats2/index.html

Hypothesis Testing
• Hypothesis: A statement we can test

– Alternative hypothesis: Our idea, e.g. there is a difference between groups
– Null hypothesis: The alternative of our idea, e.g. there is no difference between groups

• Critical Value alpha (α)

– The threshold as to how much chance you are willing to accept
– Typical values in social sciences are 0.1, 0.05, or 0.01

p-hacking • P-hacking, or Dredging
– Doing many tests until you find one which is of statistical significance
– At a confidence level of 0.05, we expect to find one positive result 1 time out of 20 tests
– Remedies:

• Bonferroni correction
• Hold-out sets
• Investigation pre-registration