Coursera: Python for data science

# Python Fundamentals

## Course content

* python is the choice of language: easy to learn and use, fully functional, popular, minimal templating, and has libraries for data science (scipy ecosystem)!
* first module: basics of Python language with some advanced Python
* second module: pandas toolkit, a dataframe library (similar to R)
* third module: advanced queries for pandas’ dataframes – Boolean masking and hierarchical indexing
* final module: course project, take some datasets, merge and clean, then process data to answer some questions.
* This course help you deal with data using python and perform basic stat test.

## What is Data Science

* apply data driven methodologies to aid in discovery.
* Data science is the intersection of hacking skills, math & statistical knowledge, and substantive expertise (Drew Conway).
  + skepticism, experimentation, simulation, and replication.
  + The Scientific Inquiry
* David Donoho’s 50 years of data science
  + Data Exploration and Preparation
  + Data Representation and Transformation
  + Computing with Data
  + Data Modeling
  + Data Visualization and Presentation (2nd course, this is a whole field of its own)
  + Science about Data Science (understand what works)
* Pipelining: use of many different programming languages to accomplish data science project.

## The Python Programming Language

* Jupyter Notebook System
  + Open Notebook for each lecture, and follow along.
  + Assignments are built in as well, some are right into the Jupyter Notebook.

MOST OF THE CONTENT AND WORK WILL BE TAKEN ON JUPYTER NOTEBOOK FOR FUTURE LECTURES. ONLY KEY CONCEPTS ARE OUTLINED HERE.

## Python Syntax

List

|  |
| --- |
| x = [1, 'a', 2, 'b']  type(x) |

Tuple

|  |
| --- |
| x = (1, 'a', 2, 'b')  type(x)  # unpacking tuple items  x = ('Christopher', 'Brooks', 'brooksch@umich.edu')  fname, lname, email = x  fname  lname |

Dictionaries

|  |
| --- |
| x = {'Christopher Brooks': 'brooksch@umich.edu', 'Bill Gates': 'billg@microsoft.com'}  x['Christopher Brooks'] # Retrieve a value by using the indexing operator |

# Working with Pandas

## The series data structure

* None is treated as “None” in a string list, whereas None is treated as NaN in number list.