# MSDS 420

Atef Bader, PhD

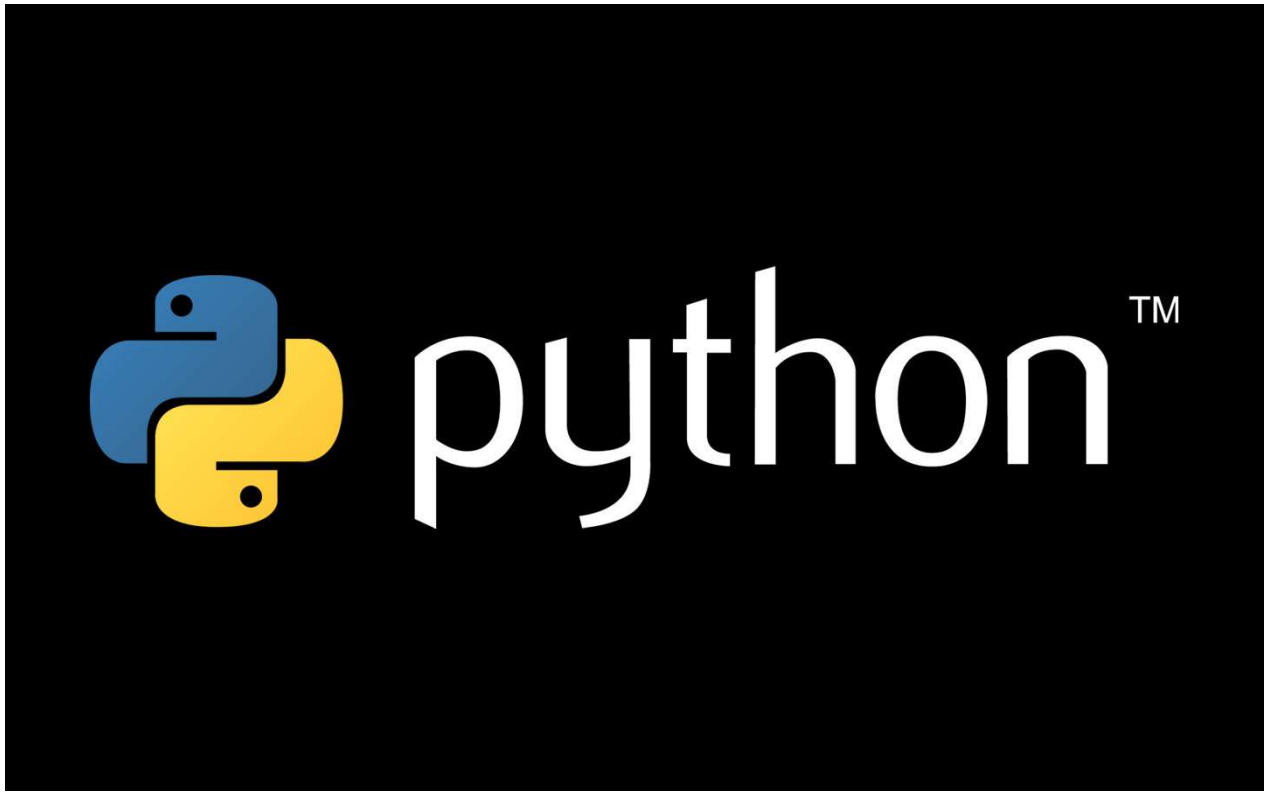
# Agenda

- Basics of Python programming language
- Pandas Data Analysis Library
- A-MUST readings for DataFrame
- Exercise #1 Walkthrough & Deliverable

# Basics of Python

1. Basic Data Types: Integer, Float, Boolean, etc.
2. Data Structures: DataFrame, Series, Lists, Dictionaries
3. Control Structures: Loops, if-statement

# Why Python ?



# https://insights.stackoverflow.com/survey/2018/

File Edit View History Bookmarks Tools Help

Stack Overflow Developer Surv

https://insights.stackoverflow.com/survey/2018/

Most Visited Getting Started

stackoverflow

- Overview
- Developer Profile
- Technology
  - I. Most Popular Technologies
  - II. Most Loved, Dreaded, and Wanted
  - III. Development Environments and Tools
  - IV. Top Paying Technologies
  - V. Correlated Technologies
  - VI. Technology and Society
- Work
- Community
- Methodology

Back to top

Take control of your job search.

Stack Overflow Jobs puts developers first. No recruiter spam or fake job listings.

Browse jobs

## Most Popular Technologies

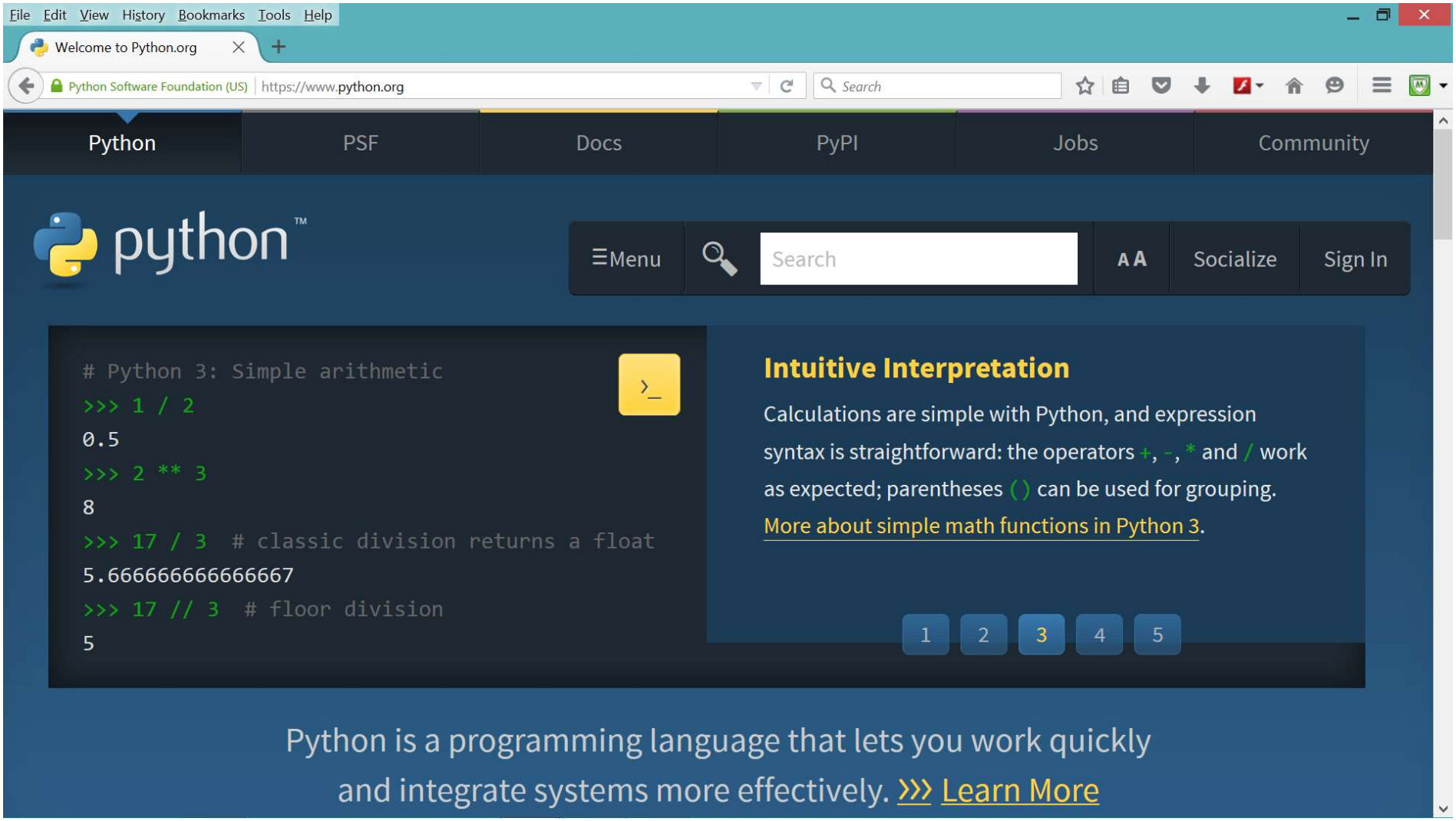
### Programming, Scripting, and Markup Languages

All Respondents | Professional Developers

| | |
|---|---|
| JavaScript | 69.8% |
| HTML | 68.5% |
| CSS | 65.1% |
| SQL | 57.0% |
| Java | 45.3% |
| Bash/Shell | 39.8% |
| Python | 38.8% |
| C# | 34.4% |
| PHP | 30.7% |
| C++ | 25.4% |
| C | 23.0% |
| TypeScript | 17.4% |
| Ruby | 10.1% |

python Highlight All Match Case Whole Words 2 of 10 matches

# Why Python ?


File Edit View History Bookmarks Tools Help
Welcome to Python.org
Python Software Foundation (US)
https://www.python.org
Search
Python
PSF
Docs
PyPI
Jobs
Community
python™
Menu
Search
A A
Socialize
Sign In
# Python 3: Simple arithmetic
>>> 1 / 2
0.5
>>> 2 ** 3
8
>>> 17 / 3  # classic division returns a float
5.666666666666667
>>> 17 // 3  # floor division
5
Intuitive Interpretation
Calculations are simple with Python, and expression syntax is straightforward: the operators +, -, * and / work as expected; parentheses ( ) can be used for grouping.
More about simple math functions in Python 3.
1
2
3
4
5
Python is a programming language that lets you work quickly and integrate systems more effectively. >>> Learn More

# Why Python (1): Simpler

- Python is a simpler language
- Simpler means:
  - Fewer alternatives (one way to do it)
  - Better alternatives (easier to accomplish common tasks)
- This allows us to focus less on the language and more on problem solving

# Why Python(2): Best Practices

- Many of the best parts of other languages are included in Python
  - data structures (lists, dictionaries)
  - control (iteration, exceptions)
  - many packages for common tasks

# Why Python(2): Best Practices

- Python is often described as "batteries included"

# Why Python(3): User base

- Python is Open Source:
  - freely available
  - large user base constantly contributing
  - new packages available to meet changing needs

# Pandas – Data Analysis Library

**Pandas is a library of pakcgaes for Data Analysis in Python**

File Edit View History Bookmarks Tools Help

Python Data Analysis Library ...

pandas.pydata.org

Search

Fork me on GitHub

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

overview // get pandas // documentation // community // talks

## Python Data Analysis Library

*pandas* is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

*pandas* is a NUMFocus sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project.

A Fiscally Sponsored Project of

NUMFOCUS

OPEN CODE = BETTER SCIENCE

### VERSIONS

**Release**

0.19.2 - December 2016

download // docs // pdf

**Development**

0.20.0 - 2017

github // docs

**Previous Releases**

0.19.1 - download // docs // pdf

0.19.0 - download // docs // pdf

0.18.1 - download // docs // pdf

0.18.0 - download // docs // pdf

# Pandas – Data Analysis Library

**Here is the main theme behind pandas package.**

**That is, to do data munging, preparation,modeling, and analysis without the need to switch to R language**

File Edit View History Bookmarks Tools Help

pandas 0.18.1 : Python Pac... | Python for Data Analysis -... | Python Data Analysis Library ...

pandas.pydata.org

## What problem does *pandas* solve?

Python has long been great for data munging and preparation, but less so for data analysis and modeling. *pandas* helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

Combined with the excellent IPython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate.

*pandas* does not implement significant modeling functionality outside of linear and panel regression; for this, look to statsmodels and scikit-learn. More work is still needed to make Python a first class statistical modeling environment, but we are well on our way toward that goal.

## What do our users have to say?

**Roni Israelov, PhD**
Portfolio Manager

AOR CAPITAL MANAGEMENT

# Pandas – Data Analysis Library

**All you need to work on for a given dataset: slice/dice/index/groupby/CSV/SQL/NoSQL**

File Edit View History Bookmarks Tools Help

Python Data Analysis Library ...

pandas.pydata.org

Search

Fork me on GitHub

## Library Highlights

- A fast and efficient **DataFrame** object for data manipulation with integrated indexing;
- Tools for **reading and writing data** between in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format;
- Intelligent **data alignment** and integrated handling of **missing data**: gain automatic label-based alignment in computations and easily manipulate messy data into an orderly form;
- Flexible **reshaping** and pivoting of data sets;
- Intelligent label-based **slicing**, **fancy indexing**, and **subsetting** of large data sets;
- Columns can be inserted and deleted from data structures for **size mutability**;
- Aggregating or transforming data with a powerful **group by** engine allowing split-apply-combine operations on data sets;

# DataFrame


**DataFrame is the Lamborghini of data structures in Python**

# What you need to submit for Exercise #1 on Canvas?

1. Run the IPython Notebook Script
2. Add your code for the requirements
3. Run the IPython Notebook Script Again
4. Save your updated IPython Notebook Script along with the OUTPUT for the cells
5. Submit your updated IPython Notebook Script on Canvas
6. Submit the  PDF document of your ipynb script on Canvas