CMP 464 Data Handling and Analysis

Liang Zhao
Fall 2021

Course Information

Instructor: Liang Zhao

Office: GI 101A

Office Hours: M/W 1:00 - 3:00PM

Email: Liang.Zhao1@lehman.cuny.edu

Prerequisites: Programming Methods II, Discrete Mathematics

Textbook: Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd Edition) Amazon O'Reilly

Data Handling

1. Import and export data in various formats
2. Modify and transform data
3. Handle outliers and missing values
4. Split, group, or combine data from multiple sources
5. Sorting, merging, ...

Data Analysis

1. Identify and extract useful data features
2. Visualize and summarize data distributions and statistics
3. Discover relationships among features
4. Provide estimates or predictions

Main Topics

1. A fast-paced introduction to Python programming
2. Arrays and vectorized computations with NumPy and SciPy
3. Data cleaning and processing with Pandas
4. Plotting and visualization with Matplotlib
5. Interacting with Data from Web or Databases
6. Managing time series data, images, texts
7. Introduction to Python modeling libraries
8. Data visualization with Tableau
9. Big data analytics with Apache Spark

Grading Policy

Homework - 20%

Tests - 20%

Midterm Project - 30%

Final Project - 30%

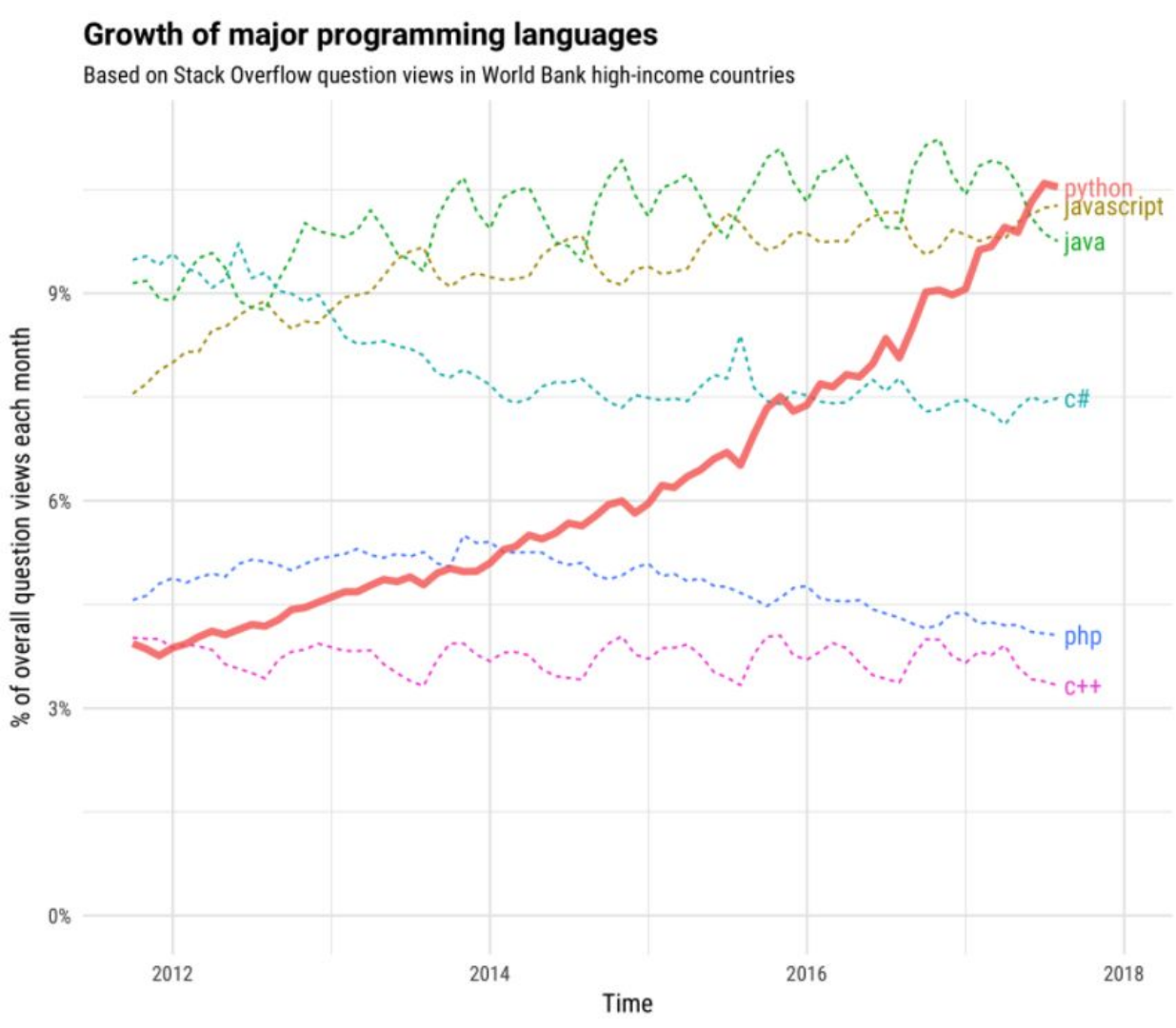
- Homework are submitted on Blackboard
- Late homework submissions will receive a 40% penalty
- Tests are held during class time. They are in-person tests.

Misc.

- Expectations
- Honor code
- Email
- Accommodating disabilities
- Consent to Zoom recording
- Additional Resources: Code and Data used in textbook:https://github.com/wesm/pydata-book
- Blackboard
- Github: to be announced
- Course videos: to be announced

Python

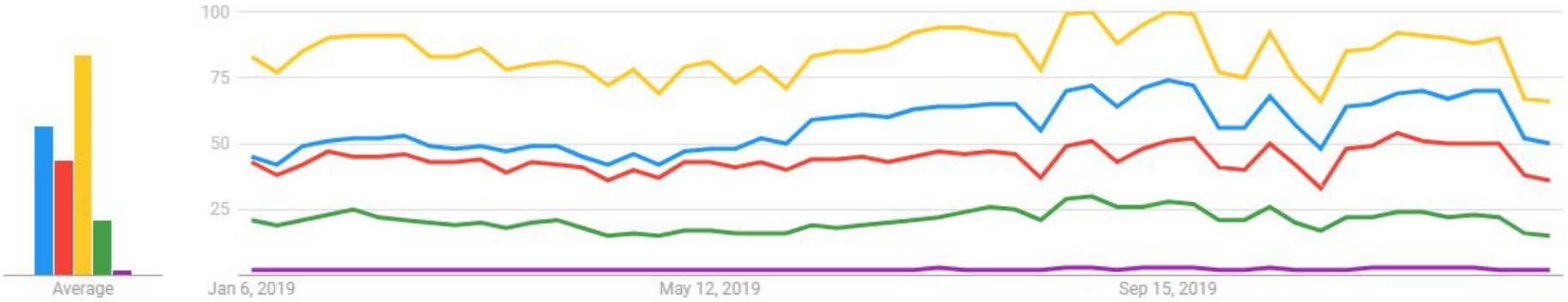
- Created by Guido van Rossum in 1991
- An interpreted, high-level, general-purpose programming language
- Open source and object oriented
- Has many data science libraries (NumPy, Pandas, Matplotlib, TensorFlow,...)
- Python 2.x (out-dated) vs. Python 3.x
+ Python shell
+ Python script
+ Jupyter notebook (Main programming environment for this class, installed through Anaconda)


Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
% of overall question views each month
9%
6%
3%
0%
2012
2014
2016
2018
Time
python
javascript
java
c#
php
c++

Python	JavaScript	Java	C++	Go
Programming language	High-level programmin...	Programming language	Programming language	Programming language

India | Past 12 months | All categories | Web Search

Interest over time



Jupyter Notebook

We will use Jupyter Notebook as our coding environment. It supports writing and executing python code in browser.

1. Install the Anaconda distribution of Python and start “Jupyter Notebook”.
2. Create a new notebook named “CMP464Week01”.
3. Type `print(“Hello World!”)` in the first cell.
4. Click the run button on the left to see execution results.

More on Jupyter Notebook

1. Click “+” button to add a new code cell.
2. Type `print(1+2)` and execute.
3. Add a new cell and convert its format from “code” to “markdown”
4. Click the up arrow to move the text cell to the top.
5. Type `# Week 1 Python Tutorial` and execute.
6. To print the notebook as a PDF file, click the “Print” option from the browser (not from the menu of jupyter notebook), and then choose “Print to PDF”. **All homework submissions for this class are required in PDF format.**

Python is simple

1. Print “Hello World”
2. Calculate the area of a square with length = 1.5
3. Add comments
4. Print sentence “The area is: ***”

Variable and Type

1. Calculate the Body Mass Index (BMI) of a person whose height is 1.73 and weight is 70.5.
2. Calculate the BMI of a person with height = 1.79 and weight = 68.7
3. What is the type of the variables used?

- Python types: int, float, bool, str
- “+” acts differently on numbers and strings
- Type conversion

Methods

1. Write a method called “square” that calculates the square of a given number.
2. Write a method called “print_square()” that calculates and prints the square of a given number.

Classes

Define a class Book with attributes:

1. Name
2. Author
3. Price

Define the constructor method __init__()

Define “print()” method that prints the book information

Language Design

- Indentation, not braces
- Everything is an object (including functions)
- And, or, not
- Name conventions:
  - Class - CapWords,
  - Variable/method - snake_case
  - Constants - ALL_CAPS
- Name conventions for private and protected methods