Numpy and Pandas

Python for Economics Morning, Aug 13, 2024 Zhenhua He



The need for better number-crunching

Lists do not have many built-in methods for interacting with data.

The base Python data types also take up a lot of space compared to other languages.

The **Numpy** and **Pandas** modules offer powerful tools for improving performance when you're using lots of data and doing lots of operations on them.



Arrays and Series: Arrays

Numpy Arrays support common operations, such as arithmetic, on an element-by-element (or "vectorized") basis.

Example:

array C = array A + array B

This adds the elements of A and B pair-wise (Instead of concatenating the elements as would happen with lists).

Pandas Series and DataFrames further expand on this.



Arrays and Series: Series

Index

Α

В

Value

0

1

- One-dimensional labeled array
- Capable of holding any data type (integers, strings, floating point numbers, etc.)





Break Time Reminder Slide

10 minutes break





Matplotlib

Python for Economics Morning, Aug 13, 2024 Zhenhua He Richard Lawrence



This Module

1. Line Plots

ĀМ

2. Scatter plots

Matplotlib Setup

Matplotlib does lots of plotting for us – specifically using the "pyplot" submodule. Import a module with a nickname using as:

import matplotlib.pyplot as plt

import numpy as np

(The following slides show some plot style reference, but otherwise...) → Go to notebook 6 to practice

Line Plots Styles

Simple line styles can be defined using the strings "solid", "dotted", "dashed" or "dashdot".

Named linestyles

solid 'solid'			
dotted			
dashed 'dashed'			
dashdot 'dashdot'			
	Parametrized linestyles		
loosely dotted (0, (1, 10))			
dotted (0, (1, 1))			
densely dotted (0, (1, 1))			
loosely dashed			

Anatomy of a Plot

Marker

- style
- size
- color

Figure

ĀМ

- title
- xlabel
- ylabel



Lunch is from 11:30-12:30



Plot Marker symbols

marker	symbol	description
"."	•	point
","	£	pixel
"o"	٠	circle
"v"	▼	triangle_down
	▲	triangle_up
"<"	<	triangle_left
">"	•	triangle_right
"1"	Y	tri_down
"2"	٢	tri_up
"3"	\prec	tri_left
"4"	≻	tri_right
"8"	•	octagon
"s"		square
"p"	٠	pentagon
"P"	+	plus (filled)
"*"	*	star

AM

Lunch Break Reminder Slide



1 hour break

АМ



Data Manipulation

Python for Economics Morning and Afternoon, Aug 13, 2024 Wesley Brashear, Josh Winchell

A M

This Module

- 1. Array/Series data selection
- 2. DataFrames

Ā M

- 3. Columns and Filtering
- 4. DataFrame Methods

Intro: Pandas VS NumPy

NumPy	Pandas	
Faster mathematical operations 🔽	Slower mathematical operations	
Only supports integer index	Customized index	\checkmark
Must use structured arrays	Easily handles different data types	\checkmark
Better performance when number of rows is 500K or less	Better performance when number of rows is 500K or more	\checkmark
More complicated to read and write files	Simpler to read and write more file formats	\checkmark

Array/Series Data Selection

Say we have a lot of data-and now that we have matplotlib we want to plot it... but only some of it.

Arrays provide us with ways to select data that are more nuanced than the options provided by plain lists.

→ Go to notebook 7 to practice





extra BONUS break

High Performance Research Computing | hprc.tamu.edu

АМ

DataFrames

- Primary Pandas data structure
- A dict-like container for Series objects
- Two-dimensional, size-mutable
- Heterogeneous tabular data structure



→ Go to notebook to practice

Break Time Reminder Slide

10 minutes break





Columns and Filtering

Like arrays and series, DataFrames can be indexed, sliced, and filtered.

You can select specific rows and/or columns by name or based on some criteria. Say we only want columns 1 and 4 when column 4 is "True"...



→ Go to notebook to practice

DataFrame Methods

There's a lot we can do via DataFrame Methods:

- Selecting/slicing and filtering
- Sorting or grouping by index or values
- Reading or writing to files
- Plotting
- Data cleanup
- Data merging

→ Go to notebook to practice



Break Time Reminder Slide

10 minutes break





Pandas Cheat Sheet (continued learning)



https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf



DataFrame - Merging Data

Merge DataFrame with a database-style join.

- left join
- right join



DataFrame - Merging Data

Merge DataFrame with a database-style join.

- inner join
- outer join



