Pandas

Data processing in Python

Python and Pandas

pandas

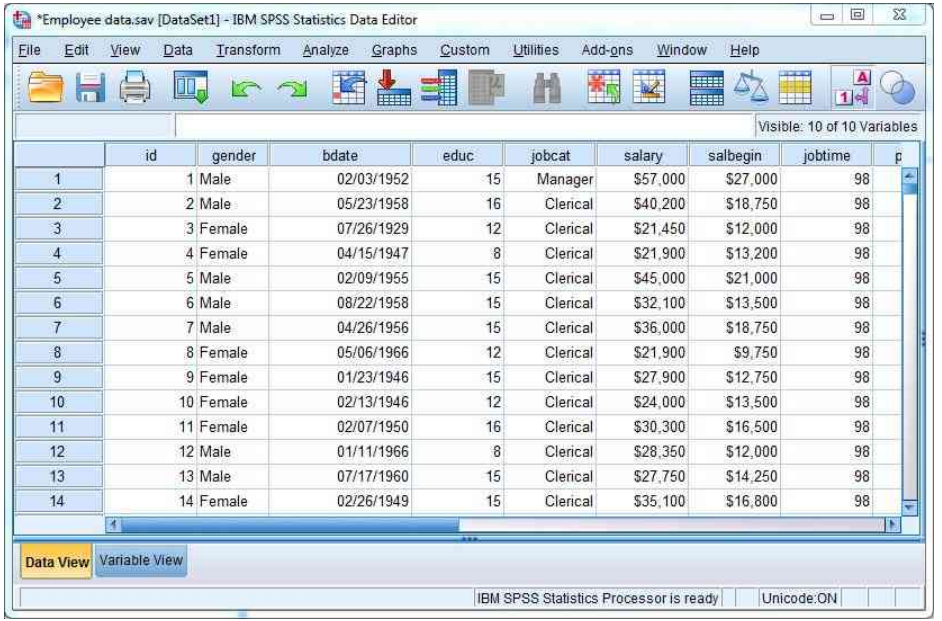
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

Pandas

- It is an extension/package/tool for Python to make data processing easier.

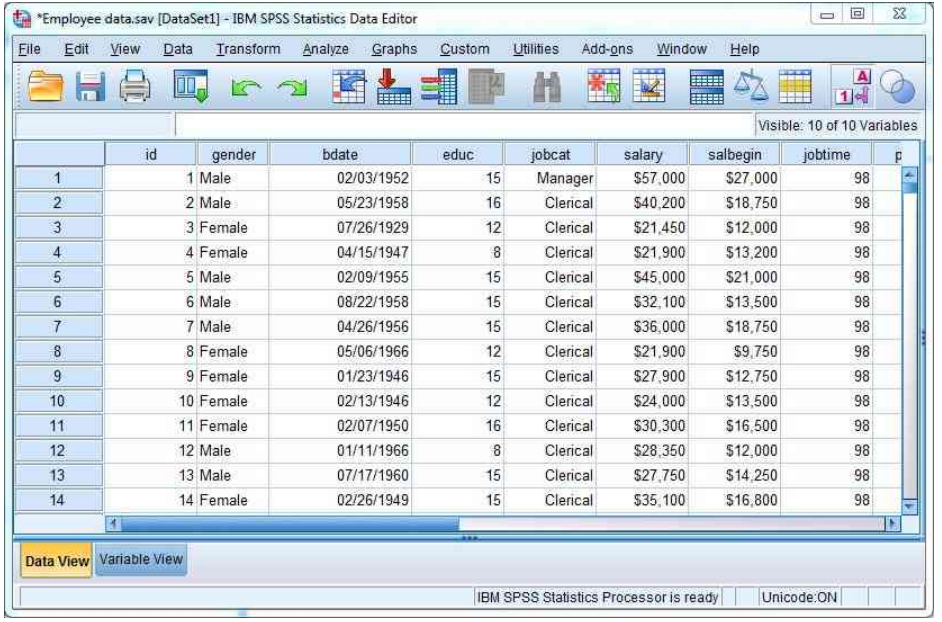
Why?

- Spreadsheet like data:
  - Records with multiple fields
  - Multiple formats



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98

- Select/Filter data
- Summarize per group
- Merging data
- Complex fields, e.g. dates



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	p
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98	
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98	
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98	
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98	
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98	
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98	
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98	
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98	

IBM SPSS

Pandas: DataFrame data format

- 2-dimensional
- labeled data
- Indexed data
- Columns different types
- Very much like excel spreadsheet

	total_bill	tip	sex	smoker	day	time	size
1	16.99	1.01	Female	No	Sun	Dinner	2
2	10.34	1.66	Male	No	Sun	Dinner	3
3	21.01	3.50	Male	No	Sun	Dinner	3
4	23.68	3.31	Male	No	Sun	Dinner	2
5	24.59	3.61	Female	No	Sun	Dinner	4
6	25.29	4.71	Male	No	Sun	Dinner	4
7	8.770	2.00	Male	No	Sun	Dinner	2
8	26.88	3.12	Male	No	Sun	Dinner	4
9	15.04	1.96	Male	No	Sun	Dinner	2
10	14.78	3.23	Male	No	Sun	Dinner	2

Sidenote: DataFrames in R and Matlab

- R: software for data analysis and statistics:
  - Dataframe is the bread and butter data format

- Matlab: software for numerical calculations
  - Uses arrays as default data structure
  - Has recently also added the dataframe

The anatomy of a Pandas DataFrame

Pandas DataFrame

- Both numerical and textual data

1	0.3	A	xyz	True
5	0.6	B	abc	False
6	0.7	C	ddf	True
7	0.8	D	qer	False
9	0.9	E	dft	False

Pandas DataFrame

lab1	data1	sx	age	dft	type
0	1	0.3	5	xyz	True
1	5	0.6	9	abc	False
2	6	0.7	3	ddf	True
3	7	0.8	6	qer	False
4	9	0.9	1	dft	False

- Numbers and text as data
- Names for the columns
- Names for the rows [index]
  - Rows names can be numbers

Pandas DataFrame

lab1	data1	sx	age	dft	type
rt	1	0.3	5	xyz	True
dt	5	0.6	9	abc	False
qt	6	0.7	3	ddf	True
my	7	0.8	6	qer	False
op	9	0.9	1	dft	False

- Numbers and text as data
- Names for the columns
- Names for the rows [index]
  - Rows names can be numbers
  - Or textual labels

Pandas DataFrame

nms	lab1	data1	sx	age	dft	type
a	1	1	0.3	5	xyz	True
a	2	5	0.6	9	abc	False
b	1	6	0.7	3	ddf	True
b	2	7	0.8	6	qer	False
b	3	9	0.9	1	dft	False

- Numbers and text as data
- Names for the columns
- Names for the rows [index]
  - Row labels can be numbers
  - Or textual labels
- It is possible to have a multi index:
  - each rows has a unique combination of names

Selecting data based on data properties

nms	lab1	data1	sx	age	dft	type
a	1	1	0.3	5	xyz	True
a	2	5	0.6	9	abc	False
b	1	6	0.7	3	ddf	True
b	2	7	0.8	6	qer	False
b	3	9	0.9	1	dft	False

- Selecting a number of rows
  - Based on a selecting criterion
  - Here: age < 5

Selecting data based on column labels

nms	lab1	data1	sx	age	dft	type
a	1	1	0.3	5	xyz	True
a	2	5	0.6	9	abc	False
b	1	6	0.7	3	ddf	True
b	2	7	0.8	6	qer	False
b	3	9	0.9	1	dft	False

- Often by column label

Selecting data based on column labels and row labels

nms	lab1	data1	sx	age	dft	type
a	1	1	0.3	5	xyz	True
a	2	5	0.6	9	abc	False
b	1	6	0.7	3	ddf	True
b	2	7	0.8	6	qer	False
b	3	9	0.9	1	dft	False

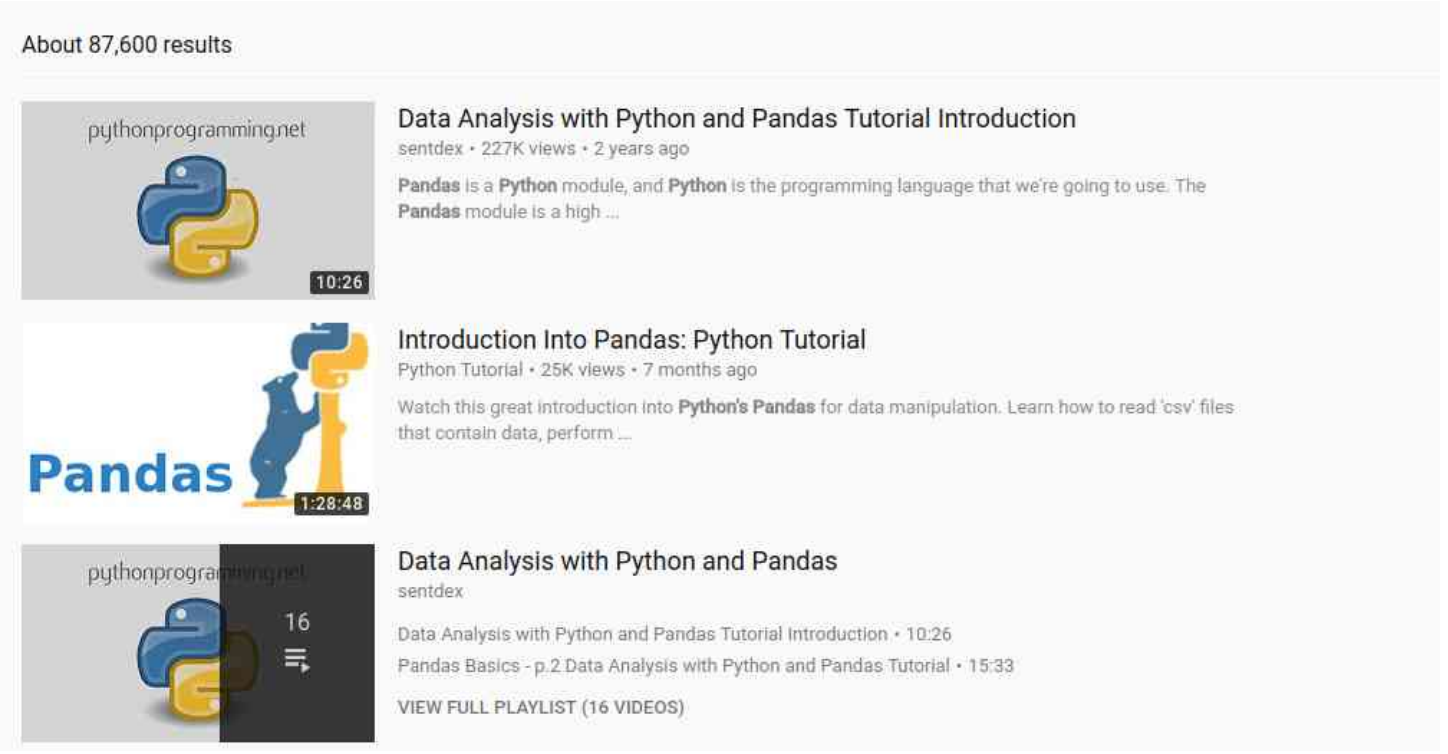
- Often by column label
- Sometimes: row & column labels
  - Rows: a,2 → b,2
  - Columns: sx → age

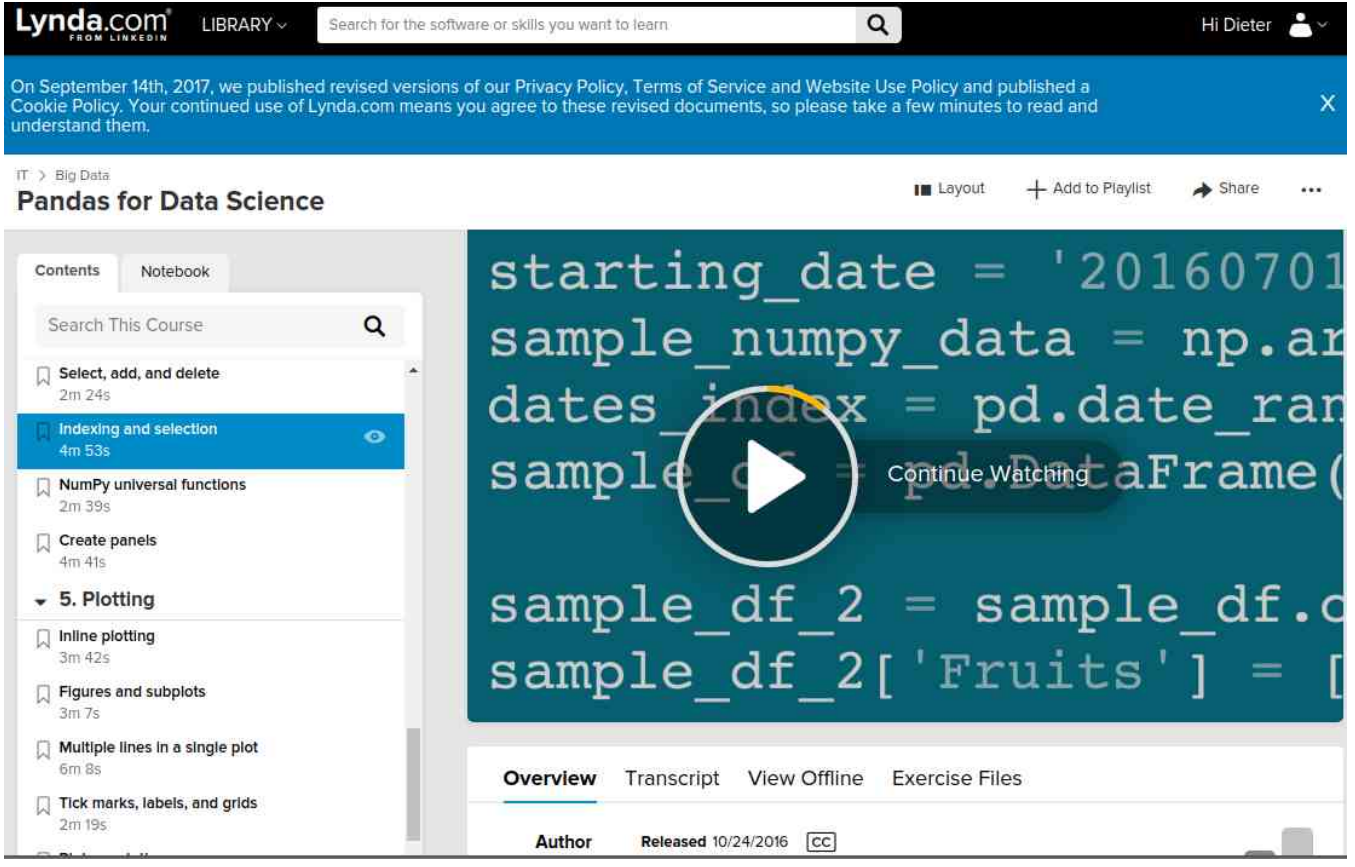
Summary

nms	lab1	data1	sx	age	dft	type
a	1	1	0.3	5	xyz	True
a	2	5	0.6	9	abc	False
b	1	6	0.7	3	ddf	True
b	2	7	0.8	6	qer	False
b	3	9	0.9	1	dft	False

- Different types of data
- Unique column names
- Unique combination of row names
- Selecting data:
  - Rows by criterion
    - Example: age < 5
  - By column and/or rows label:
    - Example:
      - Colums: sx → age
      - Rows: a,2 → b,2

By the way


About 87,600 results
pythonprogramming.net
10:26
Data Analysis with Python and Pandas Tutorial Introduction
sentdex • 227K views • 2 years ago
Pandas is a Python module, and Python is the programming language that we're going to use. The Pandas module is a high ...
Pandas
1:28:48
Introduction Into Pandas: Python Tutorial
Python Tutorial • 25K views • 7 months ago
Watch this great introduction into Python's Pandas for data manipulation. Learn how to read 'csv' files that contain data, perform ...
16
Data Analysis with Python and Pandas
sentdex
Data Analysis with Python and Pandas Tutorial Introduction • 10:26
Pandas Basics - p.2 Data Analysis with Python and Pandas Tutorial • 15:33
VIEW FULL PLAYLIST (16 VIDEOS)


Lynda.com FROM LINKEDIN
LIBRARY
Search for the software or skills you want to learn
Hi Dieter
On September 14th, 2017, we published revised versions of our Privacy Policy, Terms of Service and Website Use Policy and published a Cookie Policy. Your continued use of Lynda.com means you agree to these revised documents, so please take a few minutes to read and understand them.
IT > Big Data
Pandas for Data Science
Layout
Add to Playlist
Share
Contents
Notebook
Search This Course
Select, add, and delete
2m 24s
Indexing and selection
4m 53s
NumPy universal functions
2m 39s
Create panels
4m 41s
5. Plotting
Inline plotting
3m 42s
Figures and subplots
3m 7s
Multiple lines in a single plot
6m 8s
Tick marks, labels, and grids
2m 19s
starting_date = '20160701
sample_numpy_data = np.ar
dates_index = pd.date_ran
sample_df = pd.DataFrame(
sample_df_2 = sample_df.c
sample_df_2['Fruits'] = [
Continue Watching
Overview
Transcript
View Offline
Exercise Files
Author
Released 10/24/2016
CC