

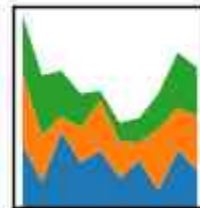
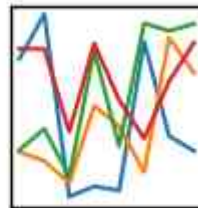
Pandas

Data processing in Python

Python and Pandas

pandas

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



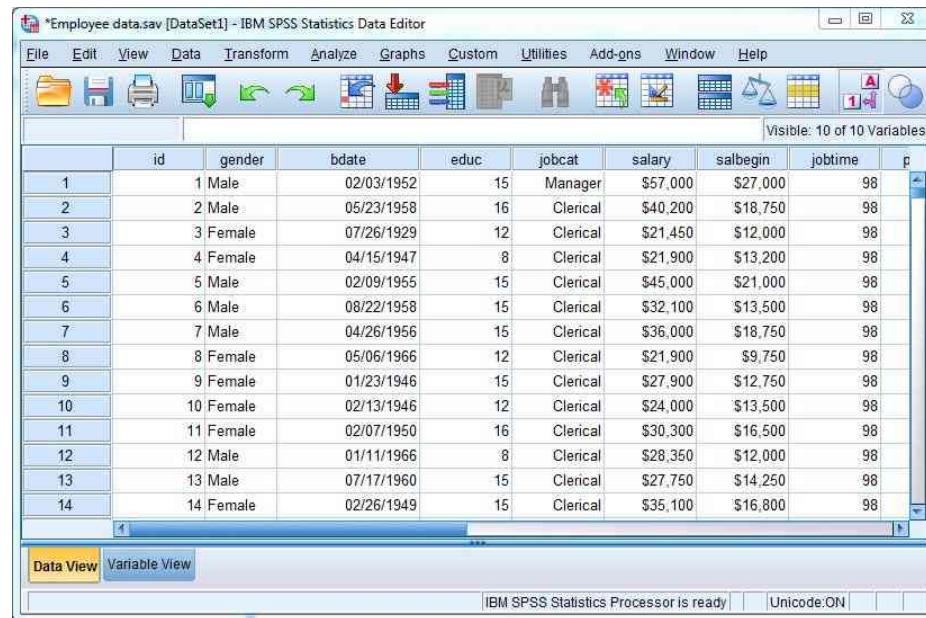
Pandas

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



Why?

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



The screenshot shows the IBM SPSS Statistics Data Editor window. The title bar reads '*Employee data.sav [DataSet1] - IBM SPSS Statistics Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Custom, Utilities, Add-ons, Window, and Help. The toolbar contains various icons for file operations, data manipulation, and analysis. The main data grid displays 14 rows of data with 10 columns: id, gender, bdate, educ, jobcat, salary, salbegin, jobtime, and p. The status bar at the bottom indicates 'IBM SPSS Statistics Processor is ready' and 'Unicode: ON'.

	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	

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

Employee data.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help

Visible: 10 of 10 Variables

	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	

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode: ON

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	8.77	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