Data frames

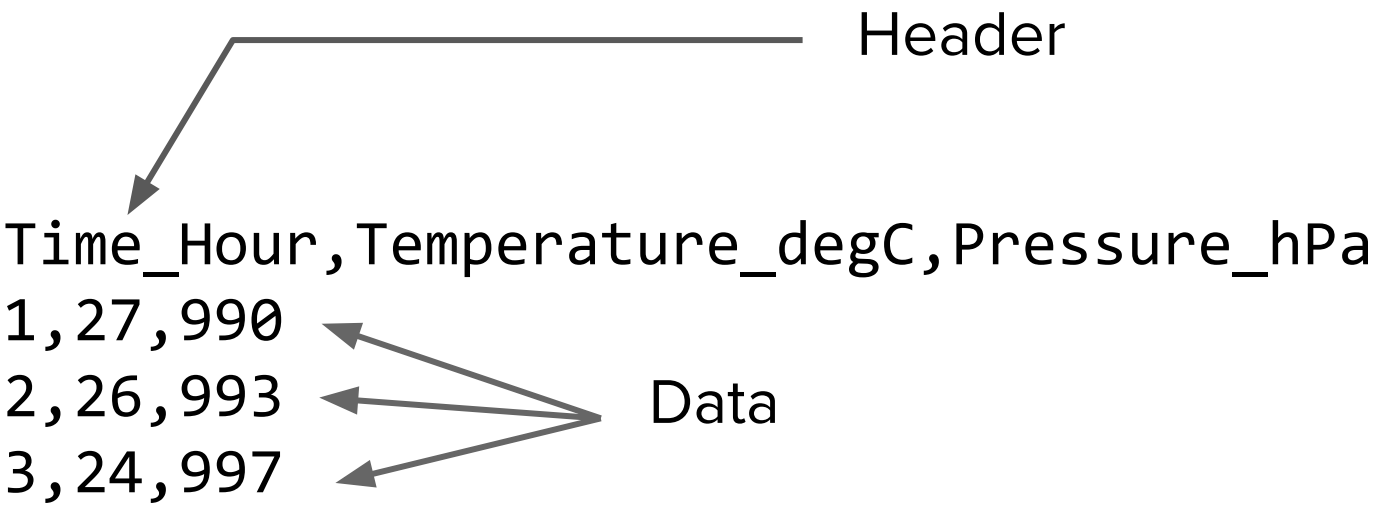
Salikh Zakirov

Outline

- CSV, data frames and ***tidy*** data frames
- Visualization with Plotly Express
- Simple data frame transformations

What is CSV format カンマ区切り形式

- A common format for representing tabular data

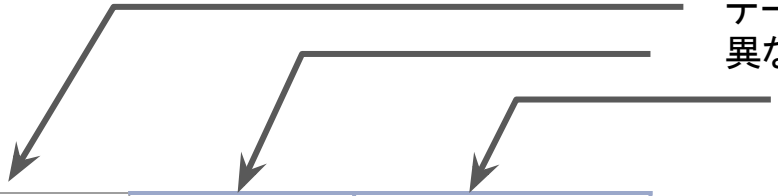


What is a data frame

- Columns have names
- Columns have types

列

データフレームの各列の型は
異なってもよい



	日付	降水量	風向
1	2019-08-08	50	南南東
2	2019-08-01	10	南

行

一行は一つの観測値と
認識してよい

Why data frame

- Provides useful tools for data manipulation
  - Vector arithmetics on columns
  - Split-apply-combine (not explained)
- Provides useful *thinking tool*
  - If you need to explore or visualize some data
  - ... and do not know how to organize it
  - use the data frame!

Data frame structure

日付	降水量	風向
2019-08-08	50	NE
2019-08-07	0	E

降水量.8/8	降水量.8/7	風向.8/8	風向.8/7
50	0	NE	E

日付	変数	値
2019-08-08	降水量	50
2019-08-08	風向	NE
2019-08-07	降水量	0
2019-08-07	風向	E

Tidy data frame

- One table = set of closely related measurements
- 1 column = 1 variable
  - i.e. one variable not spread across multiple columns
  - column header = variable name
- 1 row = 1 observation
  - easy to access **all** data about one measurement
- No data in column or row names

See https://vita.had.co.nz/papers/tidy-data.pdf [English]

Data frame structure

日付	降水量	風向
2019-08-08	50	NE
2019-08-07	0	E

Tidy

Not tidy: Data in column header

降水量.8/8	降水量.8/7	風向.8/8	風向.8/7
50	0	NE	E

日付	変数	値
2019-08-08	降水量	50
2019-08-08	風向	NE
2019-08-07	降水量	0
2019-08-07	風向	E

Not tidy: Multiple variables in one column

Why tidy data frame

- Consistent data structure
  - Fewer conversions and adapters
  - Standard tools
- Easy visualization
- *Thinking tool* (again)
  - What is "one observation"?

Visualization: how to describe a plot?

- Pick data
- Pick variables X, Y = mapping
- Pick scales
- Pick drawing style (dots/lines/bars/...) = geometry
- Other details (legend, ticks, ...)

Ref: "The Grammar of Graphics" by L. Wilkinson (2005)

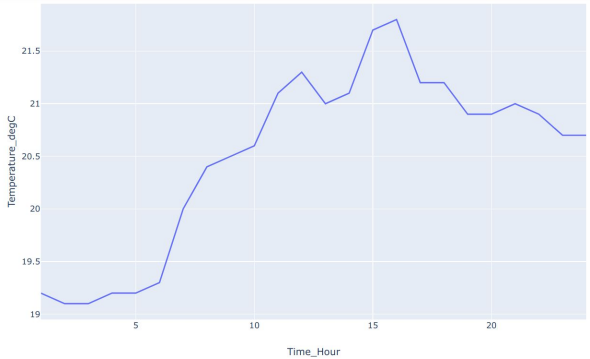
Visualization with tidy data frames

Plotting method

Data frame

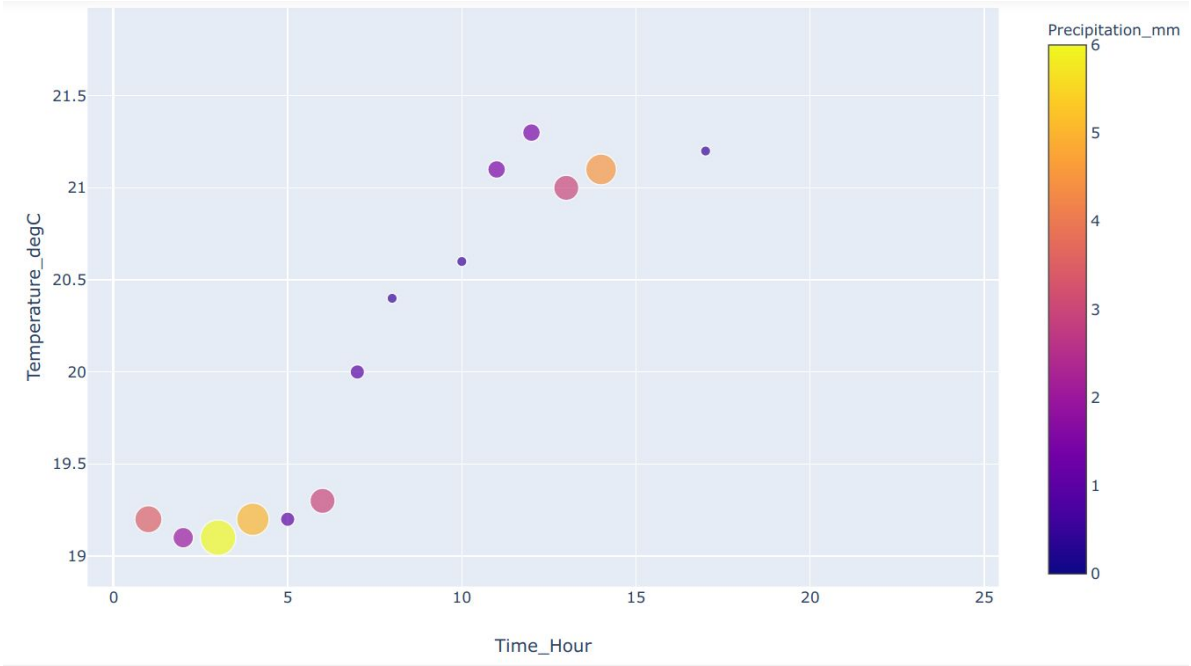
Mapping of variables

```
px.line(df, x='Time_Hour', y='Temperature_degC')
```



See: https://plot.ly/python/plotly-express/

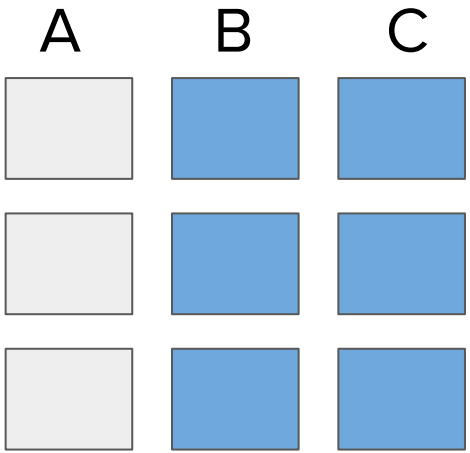
```
px.scatter(df, x='Time_Hour', y='Temperature_degC', size='Precipitation_mm', color='Precipitation_mm')
```



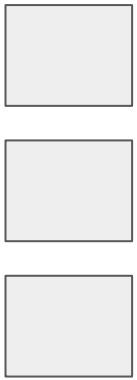
Data frame transformations

Extract column as vector

`df`



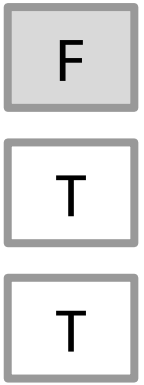
`df['A']`



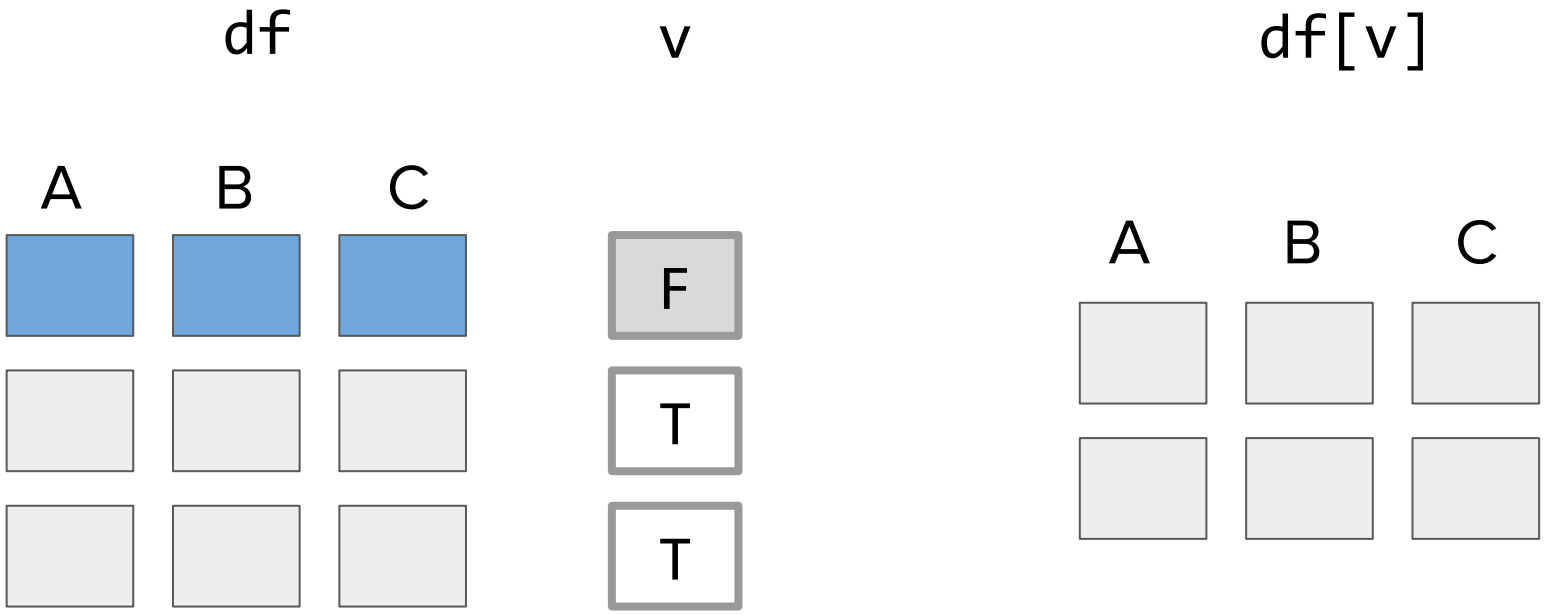
Vector operations

`df['A']`

`df['A'] > 0`

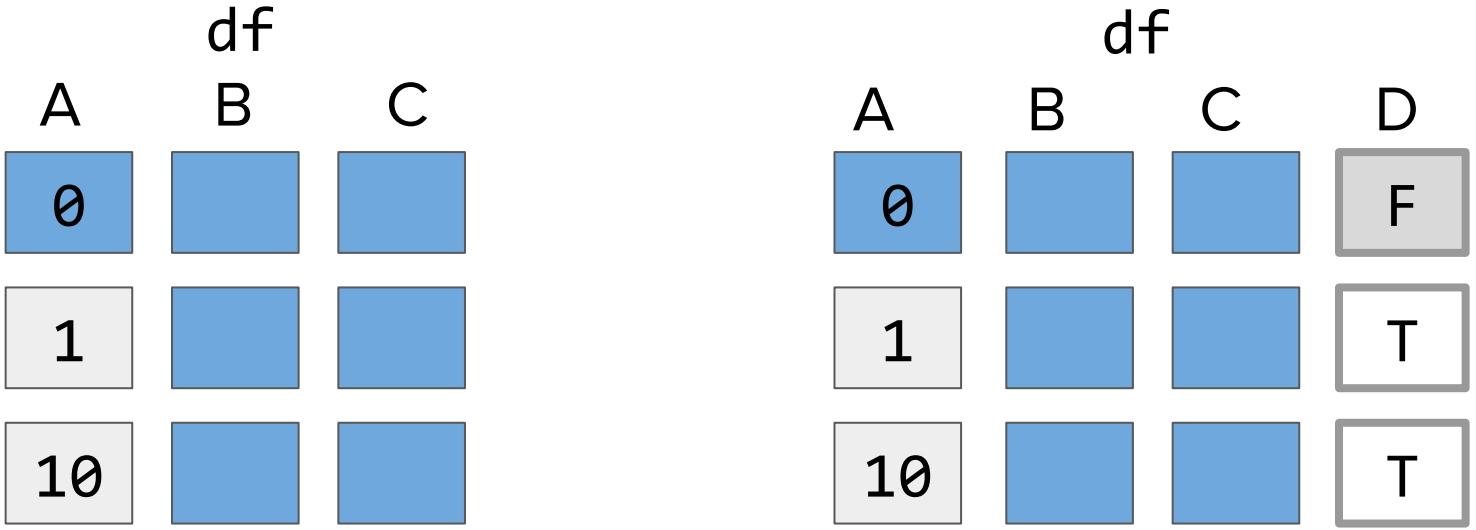


Filter rows



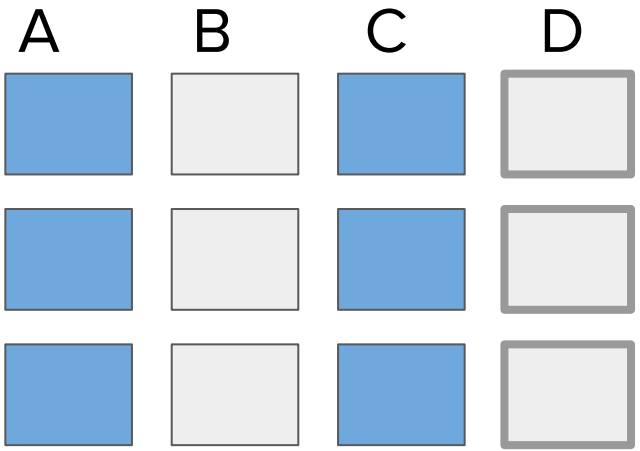
Insert new columns

```
df['D'] = df['A'] > 0
```



Select columns

`df`



`df[['B', 'D']]`