# Data Analysis with pandas

Sirasit Lochanachit, PhD

# Outline

- What is pandas?

- How to install pandas?

- Using pandas

- Series & DataFrame

- Coding

# What is pandas?

# What is pandas?

- Open source library in Python

- For working with data (e.g. Data manipulation and analysis)

- Fast, powerful, flexible and easy to use

- Widely used by Data Analysts and Data Scientists

- Under active development

- Works well with other packages
  - Built on top of Numpy
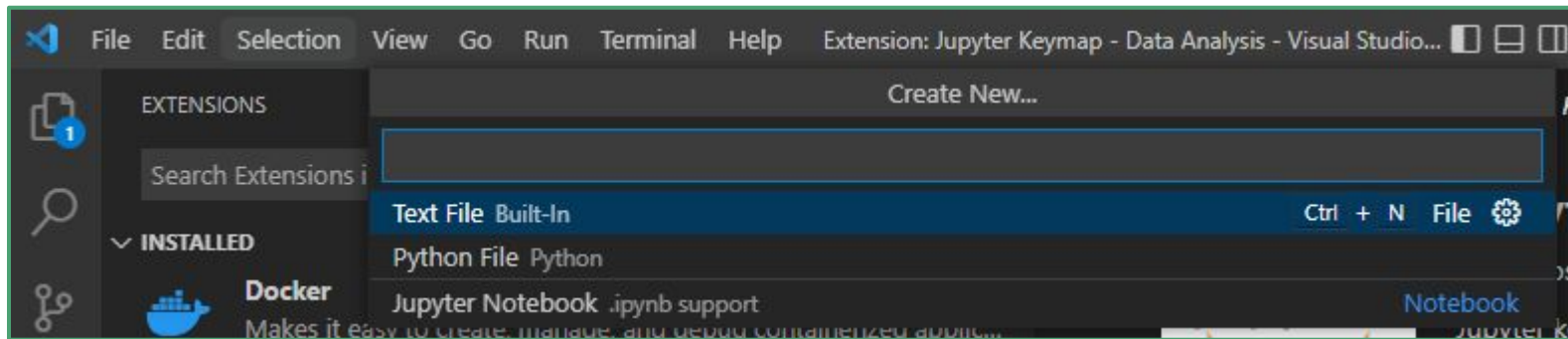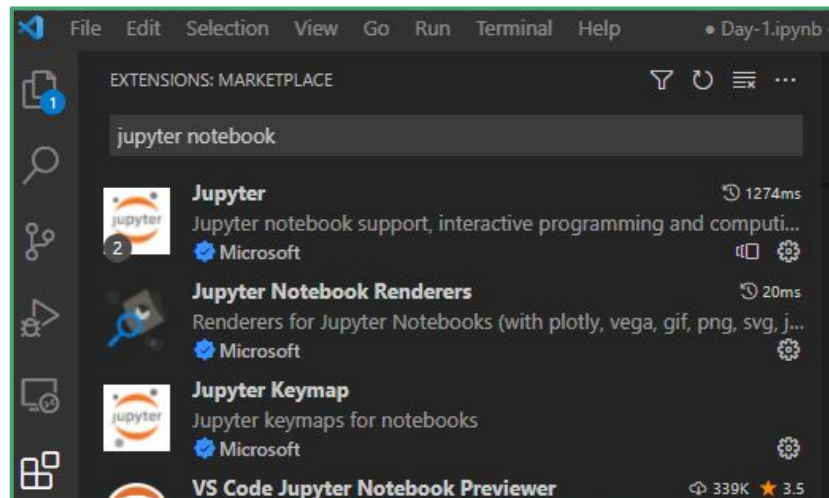  - scikit-learn

# How to install pandas?

- Easiest way is by downloading the Anaconda distribution of Python

  - https://docs.continuum.io/anaconda/

- pip install pandas

  - Try this in VS Code or other IDEs

# How to use pandas without installing?

- Google Colab provides pandas from the start (pre-install)

  - https://colab.research.google.com/

# Extra: How to use Jupyter Notebook in VS Code

- Install Jupyter Notebook Extension

# How do I read a tabular data file into pandas?

What is tabular data?

- Data that looks like a table

- Data with rows and columns

- Looks like excel spreadsheet

Common formats of tabular data:

- csv (comma, separated value)

```
import pandas as pd

pd.read_table(filename or URL)

pd.read_csv(filename or URL)


head()

tail()
```
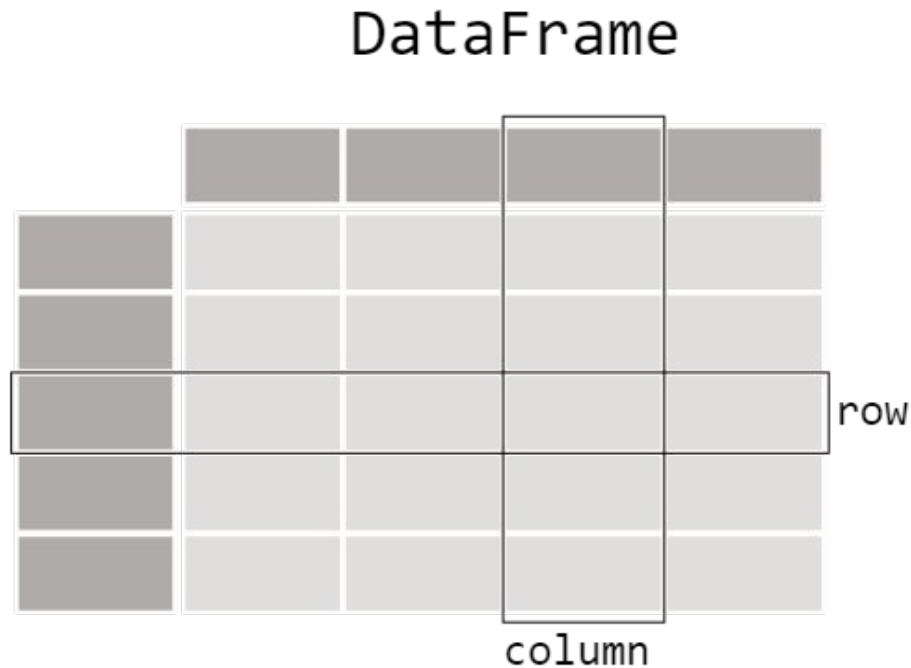
# What kind of data does pandas handle?

A **DataFrame** is a 2-dimensional data structure that can store data of different types (including characters, integers, floating point values, categorical data and more) in columns.

It is similar to a spreadsheet (i.e. Excel)
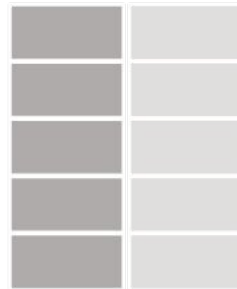
- Basically, it looks like a table.

DataFrame

row

column

# How to **select** a pandas Series from a DataFrame?
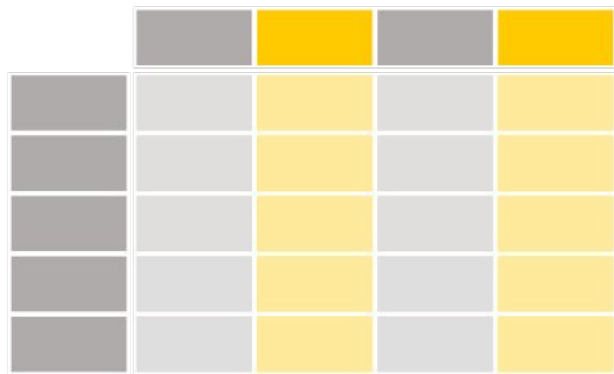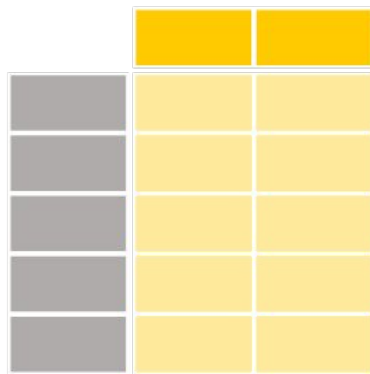
Bracket notation

- Can select multiple columns

Dot notation

# Basic DataFrame **Exploration**

shape

dtypes

info()

describe()

value_counts()

unique()

nunqiue()

mean()

std()

min()

max()

count()

# How to **rename** pandas DataFrame's columns?

1. rename()

2. re-assigned by list

3. read_csv()

# How to **remove** columns from a pandas DataFrame?

1. drop()

2. Select all other columns instead

# How to **filter rows** of a pandas DataFrame by column value?

- Need a Series of boolean values

- Use & and | to add filter criterias

# How to **sort** a pandas DataFrame or Series?

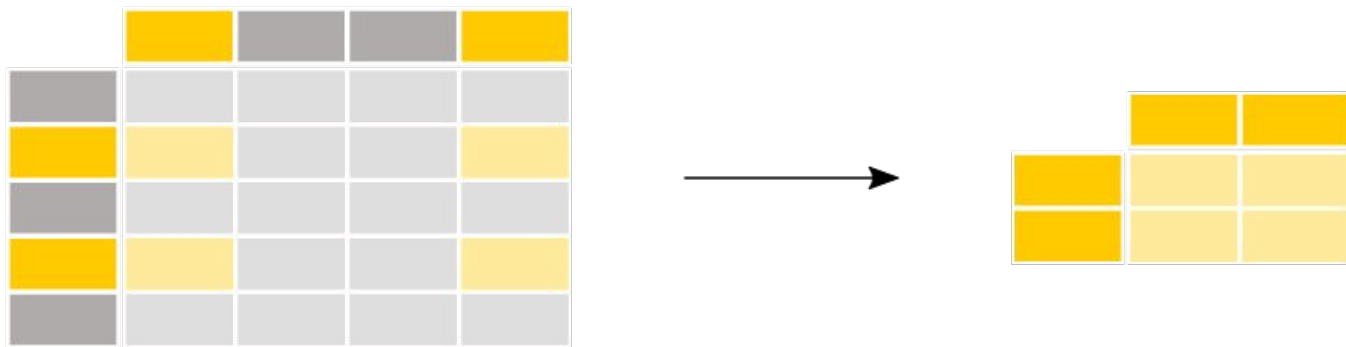- sort_values()

# String methods in pandas

- str.upper()

- str.contains('abc')

# Changing data type of a pandas Series

- astype()

# How to select specific **rows and columns** from a DataFrame?

- loc[*row*]

- loc[*row, column*]

- iloc[*row_number, column_number*]

# When to use a **groupby** command?

- groupby()

# How to handle **missing values**?

Check missing values in a DataFrame

- isnull() or isna()

- notnull() or notna()

1. Drop missing values with dropna()

Or

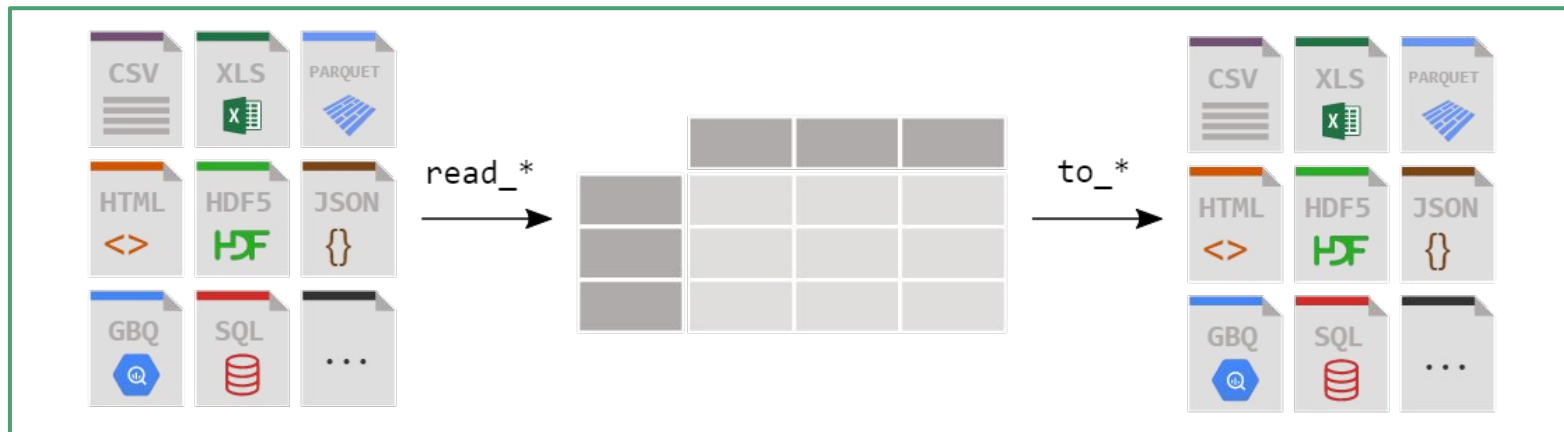2. Fill missing values with fillna()

# How to **export** a DataFrame?

CSV

- to_csv()

Excel

- to_excel()

# Exercise: SF Salaries

See SF Salaries Exercise.ipynb