Programming for Data Science: Pandas

Instructor: Jakramate Bootkrajang

Outline

- Analysing Taiwan House pricing data using Pandas
- Visualising results with Matplotlib

The data

- Download the data from
- http://www.cs.science.cmu.ac.th/person/jakramate/courses/2018/ds223/real_estate.csv

Uploading to Colab

```
from google.colab import files

uploaded = files.upload()
```

[Q1] Exploring your data

- Upload ‘real_estate.csv’ to Colab, read the file and answer the following questions
  - How many data records are there ?
  - How many columns ? And what are they ?

[Q2] Data indexing

- Write a script to select and show house price

[Q3] Basic statistics

- How much is the most expensive house ?

[Q4] Basic statistics

- What’s the average price of houses in Taiwan ?

[Q5] Basic statistics

- Write a script to show houses whose prices are above the average price.
- How many are they ?

[Q6] Does number of stores affect house price ?

- Find the average prices of houses grouped by the number of store in that neighbourhood

[Q7] Visualising

- Write a script to plot ‘house price’ versus ‘number of stores’.

[Mini individual project]

- Use the remaining time to think about what other useful insight can be extracted from the data.
- Your project must make use of
  - Scipy or sklearn (next week)
  - Pandas
  - Matplotlib
- The project is worth 6 points