Data structures & Control Flow

Sugarkhuu Radnaa

Py4Econ in Ulaanbaatar

py4econ@gmail.com

Week 3: Learning objectives

Get to know:

1. Data structures in **numpy** and **pandas**

Main data structures in Python

- Array (numpy)
- Dataframe (pandas)

Numpy

Numpy and Scipy are extensively used for scientific computations and ML/AI algorithms

- random array and vector operations
- universal functions (sin, cos, max, argmax, ...)
- array handling (broadcast, tile, ravel, ...)
- matrix operations

Pandas

Powerful dataframe package in Python

- Can read data from many sources (csv, excel, json, html, url)
- Can handle large size of data
- Can sync well with other packages

Operations in Pandas

1. info, dtypes
2. head, tail
3. describe
4. df['field'] = df.field
5. loc, iloc
6. sort, filter
7. groupby, pivot table
8. type conversion
9. min, max, mean, sum
10. import/export dataframe from/to file (csv, json, xlsx, ...)

Homework

1. Task 1
2. Task 2

- Push your homework to your Homework repo
- Deadline: 1 week

Note: Commit your results step by step.

Task 1

1. Add rows with arbitrary values to “data.xlsx” so that you have 50 rows in total
2. Create a ’.csv’ file with data of only women older than 25 years old
3. Create a ’.json’ file with data of men under than 23 years old

Task 2

1. What are the values in row 17 and column 2-5 of dataframe created from "data.xlsx"
2. What are the values row 25-28 and column 'firstName, age' of dataframe created from "data.xlsx"
3. Find the lowest, the highest and mean salary and age for men and women separately using 'groupby'
4. Find the lowest, the highest and mean salary and age for men and women separately using 'pivot table'

Thank you!