

Pandas

- PANel DAta
- 구조화된 데이터의 처리를 지원하는 python 라이브러리
- numpy와 통합해 강력한 spreadsheet 처리 기능 제공
- 인덱싱, 연산용 함수, 전처리 함수, 데이터 처리 및 통계 분석에 용이
- tabular data
 - attribute(field, feature, column)
 - instance(tuple, row)
 - data(value)
- Series
 - DataFrame 중 하나의 column에 해당하는 데이터의 모음
 - column vector를 표현하는 object
 - subclass of numpy.ndarray
 - index, data, dtype
- DataFrame
 - numpy array-like
 - each column can have a different type
 - row and col index
 - series를 모아서 만든 data table
- loc, iloc
 - loc: locate index (name)
 - iloc: locate index number(0, 1, 2, ...)
- selection with column names

```
df[['account', 'street', 'state']].head().T
```

- map

```
def change_sex(x):  
    return 0 if x == "male" else 1  
  
df.sex.map(change_sex)
```

- useful functions
 - unique()
 - describe(): (숫자형만) 통계치 한번에 출력
 - sort_values()
 - corr(): correlation
 - value_counts()