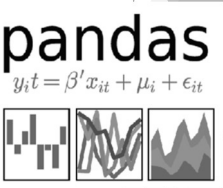


pandas

Pandas

- ▶ Pandas는 강력한 데이터 구조를 사용하여 고성능 데이터 조작 및 분석 도구를 제공하는 오픈소스 파이썬 라이브러리
- ▶ Pandas는 데이터 분석 라이브러리로 행과 열을 객체로 사용하여 데이터 분석 및 데이터들의 저장, 가공등을 도와줌
- ▶ 또한 데이터들의 대표성이나 평균값들 그리고 데이터병합 등을 가능하게 함



Pandas 특징

- ▶ 기본 인덱싱 및 사용자 정의 인덱싱이 포함된 빠르고 효율적인 DataFrame 개체
- ▶ 다른 파일 형식의 데이터를 메모리 내에 빠르게 적재하기 위한 도구 제공
- ▶ 데이터 정렬 및 누락된 데이터의 통합 처리 기능
- ▶ 날짜 집합의 피벗팅과 리셰이프 기능
- ▶ 대용량 데이터 집합에 대한 레이블 기반 슬라이싱, 인덱싱 및 서브네팅
- ▶ 데이터 구조의 컬럼을 삭제하거나 삽입
- ▶ 집계 및 변환을 위한 데이터 그룹핑
- ▶ 고성능 데이터 병합 및 결합
- ▶ 시계열 기능

Pandas 데이터 구조

- ▶ Pandas는 3가지 데이터 구조를 처리할 수 있음
- ▶ Series
- ▶ DataFrame
- ▶ Panel
- ▶ Numpy 배열 위에 구축되어 빠르게 수행

Series

- ▶ 1차원 배열
- ▶ 정수, 문자열, 실수, 파이썬 객체등의 데이터를 저장
- ▶ Array, Dict, 스칼라 값 또는 상수등으로 만들 수 있음
- ▶ 동일한 유형의 자료들로 구성
- ▶ 크기변경 불가
- ▶ 데이터값 변경 가능

Series

	Name	Team	Number
0	Avery Bradley	Boston Celtics	0.0
1	John Holland	Boston Celtics	30.0
2	Jonas Jerebko	Boston Celtics	8.0
3	Jordan Mickey	Boston Celtics	NaN
4	Terry Rozier	Boston Celtics	12.0
5	Jared Sullinger	Boston Celtics	7.0
6	Evan Turner	Boston Celtics	11.0

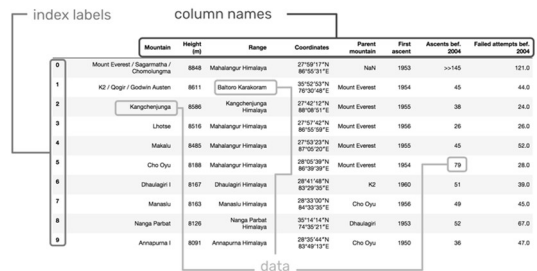
Series의 요소에 접근하기

- ▶ NumPy의 접근방식과 유사하게
- ▶ 인덱스로 접근하거나
- ▶ 슬라이싱 기법을 사용하거나
- ▶ 레이블 인덱스로 접근할 수 있음

DataFrame

- ▶ 2차원 테이블
- ▶ 서로 다른 유형의 자료들로 구성
- ▶ List, Dict, Series, NumPy배열, DataFrame으로 생성
- ▶ 크기변경 가능
- ▶ 데이터값 변경 가능

DataFrame



The diagram illustrates the structure of a DataFrame. It shows a table with 10 rows (index 0 to 9) and 8 columns (Mountain, Height, Range, Coordinates, Parent mountain, First ascent, Ascents bet. 2004, Failed attempts bet. 2004). The 'Mountain' column contains names like 'Mount Everest / Sagarmatha / Chomolungma', 'K2 / Qogir / Godwin Austen', 'Kangchenjunga', 'Lhotse', 'Makalu', 'Cho Oyu', 'Dhaulagiri I', 'Manaslu', 'Nanga Parbat', and 'Annapurna I'. The 'Height' column contains values like 8848, 8611, 8586, 8516, 8485, 8188, 8167, 8163, 8126, and 8091. The 'Range' column contains 'Mahalangur Himalaya', 'Sakura Karakoram', 'Kangchenjunga Himalaya', 'Mahalangur Himalaya', 'Mahalangur Himalaya', 'Mahalangur Himalaya', 'Dhaulagiri Himalaya', 'Manaslu Himalaya', 'Nanga Parbat Himalaya', and 'Annapurna Himalaya'. The 'Coordinates' column contains coordinates like '27°59'13"N, 86°50'23"E', '35°52'53"N, 76°50'48"E', '27°52'53"N, 86°50'55"E', '27°52'53"N, 86°50'55"E', '27°52'53"N, 86°50'55"E', '28°05'58"N, 85°39'39"E', '28°14'14"N, 83°39'39"E', '28°23'02"N, 84°39'39"E', '28°14'14"N, 83°39'39"E', and '28°23'02"N, 83°49'13"E'. The 'Parent mountain' column contains 'N/A', 'Mount Everest', 'Mount Everest', 'Mount Everest', 'Mount Everest', 'Mount Everest', 'K2', 'Cho Oyu', 'Dhaulagiri', and 'Cho Oyu'. The 'First ascent' column contains values like 1953, 1954, 1955, 1956, 1955, 1954, 1960, 1956, 1953, and 1950. The 'Ascents bet. 2004' column contains values like >=145, 45, 38, 26, 45, 79, 51, 49, 52, and 36. The 'Failed attempts bet. 2004' column contains values like 121.0, 44.0, 24.0, 26.0, 52.0, 28.0, 39.0, 45.0, 67.0, and 47.0.

DataFrame 컬럼 다루기

- ▶ 선택 : 컬럼 인덱스 사용
- ▶ 추가 : 데이터프레임[컬럼명] = 추가할컬럼
- ▶ 삭제 : del / pop 함수 사용

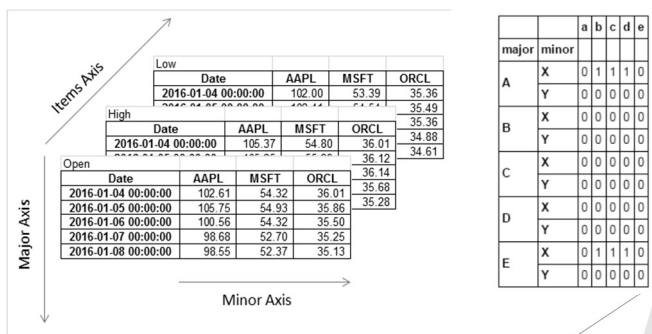
DataFrame 행 다루기

- ▶ loc : 레이블 인덱스 사용
- ▶ iloc : 정수 인덱스 사용
- ▶ ix : 레이블/정수 인덱스 사용
- ▶ 행 추가 : append 함수 사용
- ▶ 행 삭제 : drop 함수 사용

Panel

- ▶ 3차원 테이블
- ▶ 서로 다른 유형의 자료들로 구성
- ▶ 크기변경 가능
- ▶ 데이터값 변경 가능
- ▶ 0.20이후로 deprecated 됨

Panel



Panel

- ▶ 3차원 데이터 컨테이너
- ▶ 즉, 여러 데이터프레임을 저장하는 자료구조
- ▶ to_frame() 메소드를 이용해서 데이터프레임화한 뒤 사용
- ▶ Panel이라는 용어는 계량 경제학에서 파생
- ▶ pandas : pan (el) -da (ta) - s

Panel

- ▶ item - axis 0, 각 항목은 내부에 포함된 데이터 프레임 의미
- ▶ major_axis - axis 1, 각 DataFrames의 인덱스(행)
- ▶ minor_axis - axis 2, 각 DataFrames의 열

Pandas 기본 속성 및 메서드

- ▶ Series : axes dtype empty ndim size values
- ▶ DataFrame : T axes dtypes empty ndim shape size values
- ▶ head() tail()

Pandas 텍스트 다루기

- ▶ 문자열 데이터를 쉽게 조작할 수 있는 문자열 함수들을 제공함
- ▶ 이러한 함수들은 누락된/NaN 값을 기본적으로 무시(또는 제외) 함
- ▶ `lower()` `upper()` `len()` `strip()`
- ▶ `split()` `cat()` `get_dummies()`
- ▶ `contains()` `replace()`
- ▶ `repeat()` `count()` `startswith()`
- ▶ `endswith()` `find()` `findall()`
- ▶ `swapcase()` `islower()` `isupper()` `isnumeric()`

Pandas 기술 통계

- ▶ 기술통계를 위한 기본적인 메서드를 제공
- ▶ 누락된 데이터를 제외하고 기술통계를 작성할 수 있는 옵션이 제공됨
- ▶ `count()` `sum()` `mean()` `median()` `mode()` `std()`
- ▶ `min()` `max()` `abs()` `describe()`
- ▶ `prod()` `cumsum()` `cumprod()`
- ▶ `cov()`, `corr()`, `pct_change()`, `rank()`

Pandas 날짜 데이터 다루기

- ▶ 날짜 함수들은 재무 데이터 분석에서 중요한 역할 담당
- ▶ `datetime.now()` `Timestamp()`
- ▶ `date.range()` `to_datetime()`

DataFrame 테이블/행/열 단위 작업

- ▶ 데이터프레임에는 테이블 단위, 행 단위, 열 단위, 요소 단위로 작업을 수행하도록 도와주는 여러 함수를 제공함
- ▶ `pipe`, `apply`, `rename`

DataFrame Iteration

- ▶ Pandas 객체에 대한 반복작업은 유형에 따라 다름
- ▶ Series를 반복할 때는 배열처럼 취급
- ▶ DataFrame를 반복할때는 Dict 처럼 취급
- ▶ 반복은 읽기 전용이며, 반복작업 중에 객체의 값을 수정하면 사본을 반환

DataFrame 정렬

- ▶ Pandas에는 두 가지 종류의 정렬을 지원
- ▶ 레이블 기준 정렬 : `sort_index`
- ▶ 실제 값 기준 정렬 : `sort_values`

DataFrame 누락값 처리

- ▶ 기계학습 및 데이터 마이닝과 같은 영역에서는 누락값으로 인한 데이터 품질 저하 때문에 모델 예측의 정확성에 심각한 문제가 초래
- ▶ `isnull()` `notnull()` `fillna()`
- ▶ `dropna()`
- ▶ `replace()`

DataFrame 집계

- ▶ Pandas에는 집계를 수행하는 함수가 지원
- ▶ aggregation
- ▶ 집계를 수행하려면 해당 객체는 `rolling` 또는 `expanding` 객체가 되어야 함

DataFrame 그룹핑

- ▶ 기계학습 및 데이터 마이닝과 같은 영역에서는 누락값으로 인한 데이터 품질 저하 때문에 모델 예측의 정확성에 심각한 문제가 초래
- ▶ groupby
- ▶ agg
- ▶ filter

DataFrame joining

- ▶ Pandas는 SQL과 같은 관계형 데이터베이스와 매우 유사한 모든 기능을 갖춘 고성능 in-memory 조인 연산기능을 제공
- ▶ merge
- ▶ concat

Pandas 시각화

- ▶ Pandas의 Series와 DataFrame을 이용해서 데이터 시각화(Visualization) 기능을 수행할 수 있음
- ▶ 단, matplotlib의 plot() 메소드를 단순히 래핑한 것임
- ▶ plot().그래프유형

Pandas 시각화

- ▶ 선그래프
- ▶ 막대그래프 : bar, barh
- ▶ 히스토그램 : hist
- ▶ 상자수염 그래프 : box
- ▶ 산점도 : scatter
- ▶ 파이그래프 : pie

Pandas 외부데이터 파일 다루기

- ▶ Pandas I/O API를 이용해서 외부 데이터파일을 읽어
Pandas 객체에 불러오거나 Pandas 객체를 외부 파일에
저장할 수 있음
- ▶ `read_csv()` `to_csv()`