

# Pandas

# Pandas란 무엇인가?

- Pandas는 강력한 데이터 구조를 사용하여 고성능 데이터 조작 및 분석 도구를 제공하는 오픈 소스 파이썬 라이브러리임.
- Pandas라는 이름은 다차원 데이터의 계량 경제학 (Panel Data)이라는 단어에서 파생됨.
- Wes McKinney는 2008 년에 데이터 분석을위한 고성능의 유연한 도구가 필요할 때 팬더 개발에 착수함.
- Pandas 이전에는 Python이 data munging 및 preparation에 주로 사용되었음. Pandas는이 모두 문제를 해결함.
- Pandas를 사용하여 데이터의 적재, 준비, 조작, 모델링, 분석에 관계없이 데이터 처리 및 분석에서 5 가지 단계를 수행 할 수 있음.
- Python with Pandas는 금융, 경제, 통계, 분석 등을 포함한 학술 및 상업 도메인을 포함한 다양한 분야에서 사용됨.

# Pandas의 특징

- 기본 및 사용자 정의 인덱싱 기능을 갖춘 빠르고 효율적인 DataFrame 객체임.
- 다양한 파일 포맷을 메모리 내 데이터 객체에 데이터를 로드하는 곳에 사용
- 누락 된 데이터의 데이터 정렬 및 통합 처리
- 데이터 집합의 변형 및 피벗
- 대용량 데이터 세트의 레이블 기반 슬라이싱, 인덱싱 및 하위 집합
- 데이터 구조의 열을 삭제하거나 삽입 할 수 있음
- 집계 및 변환을 위해 데이터별로 그룹화함
- 고성능 병합 및 데이터 결합.
- 시계열 기능

# Data Structures

- Pandas는 세가지 데이터 구조를 가짐
  - Series
  - DataFrame
  - Panel
- 차원

데이터 구조	차원	기술
Series	1	1D 라벨 동질 배열, 크기 변경 불가능.
Data Frame	2	다양한 유형등이 데이터 형식이 가능함. 2D 크기 변형 테이블 구조.
Panel	3	3D 레이블, 크기 변경 가능한 배열.

# Data Structures

- 2 개 이상의 차원 배열을 작성하고 처리하는 것은 번거로운 작업이지만 함수를 작성할 때 사용자가 데이터 구조를 정할때 어려움이 있음. Pandas 데이터 구조를 사용하면 데이터 설정이 용이함
- 표 형식의 데이터 (DataFrame)에서는 축 0과 축 1 대신 인덱스(행)와 열 을 생각하는 것이 의미 적으로 도움을 줌.
- 모든 Pandas 데이터 구조는 값을 변경할 수 있으며 (변경 가능) Series는 모두 크기를 변경할 수 있음. 시리즈 크기는 변경할 수 없음.

# Series

- Series는 같은 데이터가 있는 구조. 1 차원 배열임. 예를 들어, 다음 시리즈는 정수 10, 23, 56, ...의 모음
  - 같은 데이터
  - 크기 변경 불가능
  - 데이터 값 변경 가능

# DataFrame

- DataFrame은 다른 타입의 데이터가있는 2 차원 배열

Name	Age	Gender	Rating
Steve	32	Male	3.45
Lia	28	Female	4.6
Vin	45	Male	3.9
Katie	38	Female	2.78

- 이 테이블은 조직의 판매 팀이 전체 성과 등급을 갖고있는 데이터를 나타냄.
- 데이터는 행과 열로 표시됨. 각 열은 특성을 나타내며 각 행은 사람을 나타냄.

# Columns Data Type

Column	Type
Name	String
Age	Integer
Gender	String
Rating	Float

- 다양한 데이터
- 크기 변경 가능
- 데이터 변경 가능



# Panel

- Panel은 다양한 데이터가 있는 3 차원 데이터 구조임.
  - Panel은 화면 형태로 표현하기는 어려움.
  - 패널은 DataFrame의 컨테이너로 설명 될 수 있음.
- 
- 다양한 데이터
  - 크기 변경 가능
  - 데이터 변경 가능

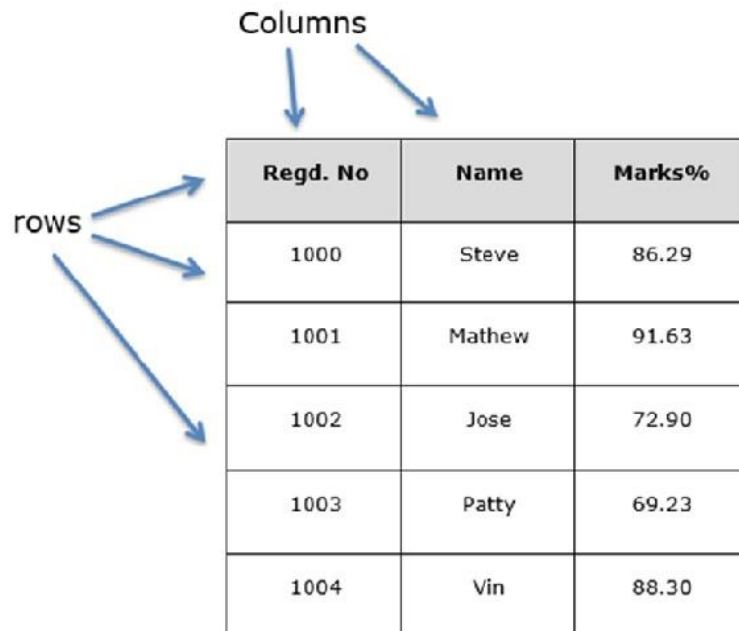
# Pandas - Series

- Series는 모든 유형 (정수, 문자열, 부동 소수점, 파이썬 객체 등)의 데이터를 보유 할 수 있는 일차원 레이블 배열임. 축 레이블을 index라고 함.
- `pandas.Series( data, index, dtype, copy)`

	매개 변수 및 설명
1	<b>data</b> 데이터는 ndarray, list, 상수와 같은 다양한 형식을 취함.
2	<b>index</b> 색인 값은 고유하고 해시 가능해야하며 데이터와 동일한 길이어야함. 인덱스가 전달되지 않으면 기본 <b>np.arange (n)</b> 임.
3	<b>dtype</b> dtype은 데이터 유형입니다. 없으면 데이터 유형이 유추됨.
4	<b>copy</b> 데이터를 복사함

# Pandas - DataFrame

- 데이터 프레임은 2 차원 데이터 구조임. 즉, 데이터는 행과 열로 표 형식으로 정렬됨.
  - DataFrame의 기능
    - 잠재적으로 열은 다른 유형임
    - 크기 - 변경 가능
    - 레이블이 지정된 축 (행과 열)
    - 행과 열에 대해 산술 연산을 수행할 수 있음



The diagram shows a table with 5 rows and 3 columns. Arrows point from the labels 'Columns' and 'rows' to the respective axes of the table.

Regd. No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

# Pandas - DataFrame

- `pandas.DataFrame(data, index, columns, dtype, copy)`

	매개 변수 및 설명
1	<b>data</b> 데이터는 ndarray, series, map, lists, dict, 상수 및 다른 DataFrame과 같은 다양한 형식을 취함.
2	<b>index</b> 행 레이블의 경우 결과 프레임에 사용할 인덱스는 선택 사항임. 인덱스가 전달되지 않으면 기본 <code>np.arange (n)</code> 임.
3	<b>columns</b> 열 레이블의 경우 선택적 기본 구문은 - <code>np.arange (n)</code> 임. 인덱스가 전달되지 않는 경우에만 해당됨.
4	<b>dtype</b> 각 열의 데이터 형식임.
5	<b>copy</b> 이 명령 (또는 그 무엇이든)은 기본값이 <code>False</code> 인 경우 데이터 복사에 사용됨.

# Pandas - Panel

- **Panel** 데이터의 3D 컨테이너임. **Panel 데이터**라는 용어는 계량 경제학에서 파생되었으며 부분적으로 pandas - **pan (el) -da (ta) -s** 이름을 정의함.
- 3 축의 이름은 패널 데이터가 포함된 작업을 설명하는데 의미적 의미를 부여하기 위한 것임.
  - **Items** - axis 0, 각 항목은 안에 포함 된 DataFrame에 해당함.
  - **Major\_axis** - 축 1, 각 DataFrames의 인덱스(행)임.
  - **minor\_axis** - 축 2, 각 DataFrames의 열임.
- `pandas.Panel(data, items, major_axis, minor_axis, dtype, copy)`

매개 변수	기술
Data	데이터는 ndarray, 시리즈, 지도, 목록, 사전, 상수 및 다른 DataFrame과 같은 다양한 형식을 취함.
Items	축 = 0
major_axis	축 = 1
Minor_axis	축 = 2
Dtype	각 열의 데이터 형식
copy	데이터를 복사. 기본값, 거짓

# Series 기본 기능

.	속성 또는 메소드	기술
1	axes	행 축 레이블 목록을 반환함.
2	dtype	객체의 dtype을 반환.
3	empty	series가 비어있는 경우 True를 반환함.
4	ndim	기본 데이터의 차원 수를 정의에 따라 반환함.
5	size	기본 데이터의 요소 수를 반환함.
6	values	Series를 ndarray로 반환함.
7	head()	최초의 n 개의 행을 반환함.
8	tail()	마지막 n 행을 리턴함.

# Pandas - Statistics

- 많은 수의 메서드가 DataFrame에서 설명 통계 및 기타 관련 작업을 집합적으로 계산함.
- 이들 중 대부분은 **sum ()**, **mean ()** 과 같은 집계이지만 **sumsum ()** 과 같은 일부는 같은 크기의 객체를 생성함.
- 일반적으로 말하면, 이 메서드는 *ndarray*. {*sum*, *std*, ...} 와 같이 축 인수를 취하지만 축은 이름 또는 정수로 지정할 수 있음
  - **DataFrame** – “index” (axis=0, default), “columns” (axis=1)