Pandas 를 사용한 데이터 분석 기초

분석할 데이터의 양(volume)이 커지고,
데이터의 입출력 속도(velocity)가 빨라지고
데이터의 종류가 다양해(variety) 짐에 따라
기존보다 데이터를 분석하기 어려워졌고
이로 인해 데이터 분석 분야가 업계의 주목을 받는 것

빅데이터 Volume, Velocity, Variety의 세 가지'V'를 가진 데이터

빅데이터 분석에는
다양한 프로그래밍 언어와 기술이 사용되고 있는데 파이썬도 그 중 하나.

파이썬이 오픈소스 기반의 통계언어인 R과 더불어
빅데이터 분석 분야에서 인기가 높아진 것은 여러 가지 이유가 있지만,
Pandas라는 라이브러리 덕이 크다.

Pandas Series

파이썬이 인기있는 이유 중 하나는
파이썬의 기본 자료 구조인 리스트, 튜플, 딕셔너리가 사용하기 편리하며
데이터를 다루는 데 효과적이기 때문.

Pandas 역시 효과적인 데이터 분석을 위한
고수준의 자료구조와 데이터 분석 도구를 제공.

Pandas의
Series는 1차원 데이터를 다루는 데 효과적인 자료구조이며,
DataFrame은 행과 열로 구성된 2차원 데이터를 다루는 데 효과적인 자료구조 이다.

Pandas를 이해하려면
가장 먼저 Pandas의 핵심 자료구조인 Series와 DataFrame을 알아야 한다.

파이썬 리스트, 튜플, 딕셔너리

- 리스트

```
mystock = ['kakao', 'naver']
print(mystock[0])
print(mystock[1])
for stock in mystock:
    print(stock)
```

- 튜플
    - 튜플은 리스트의 []와 달리 ()를 사용.
      수정이 가능한 리스트와 달리 수정할 수 없다.
      대신 리스트에 비해서 속도가 빠르다는 장점이 있다.
      그래서 원소를 한 번 넣은 후에
      수정할 필요가 없으며
      속도가 중요한 경우에 리스트 대신 튜플을 사용.

- 딕셔너리

```
exam_dic = {'key1':'room1', 'key2':'room2'}
print(exam_dic['key1'])
print(exam_dic['key2'])
```

- 리스트와 딕셔너리

```
kakao_daily_ending_prices = [92300, 94300, 92100, 92400, 92600]
for price in kakao_daily_ending_prices:
    print(price)

kakao_daily_ending_prices = {'2016-02-19':92600,
                             '2016-02-18':92400,
                             '2016-02-17':92100,
                             '2016-02-16':94300,
                             '2016-02-15':92300,}
print(kakao_daily_ending_prices['2016-02-19'])
```