

데이터 다루기 III

시계열

- 시계열 데이터는 금융, 경제, 생태학, 물리학등의 여러분야에 사용되는 매우 중요한 데이터임.
- 시간상의 여러지점을 관측하거나 측정할 수 있는 모든것이 시계열에 포함됨.
- 대부분 시계열은 고정 빈도임
 - 시간 내에 특정 순간의 타임스탬프
 - 2008년 4월이나 2018년 전체 같은 고정된 기간
 - 시작과 끝 타임스탬프로 표시되는 시간 간격, 기간은 시간 간격의 특수한 경우로 생각할 수 있음
 - 실험 혹은 경과 기간, 매 타임스탬프는 특정 시작 시간에 대한 상대적인 시간의 측정값임. 예) 쿠키를 오픈에 넣은 시점부터 매 초가 지날때마다 변화되는 쿠키 반죽의 지름 값등등

시계열

- Pandas는 표준 시계열 도구와 데이터 알고리즘을 제공함
- 대량의 시계열 데이터를 효과적으로 다룰 수 있고 쉽게 나누거나 집계하여 불규칙적인 고정된 빈도를 갖는 시계열을 리샘플링할 수 있음
 - 날짜, 시간 자료형
 - 파이썬 표준 라이브러리에 datetime, time, calendar 모듈을 제공함
 - 날짜 범위, 빈도, 이동
 - Pandas에서는 일반적인 시계열은 불규칙적인 것으로 간주됨. 고정된 빈도를 갖지 않음.
 - 시계열 안에서도 누락된 값이 발생할 수 있음. 일별, 월별, 혹은 매 15분과 같이 상대적인 고정 빈도에서의 작업이 요구되는 경우가 종종있음.
 - Pandas에서는 리샘플링, 표준 시계열 빈도 모음, 빈도 추론, 고정된 빈도의 날짜 범위를 위한 도구가 있음.
 - 시간대 다루기
 - 시간대를 처리하는 일은 시계열을 다루는 작업 중에서 힘든 부분임. 일광절약시간(DST, 섬머타임)은 문제를 일으키는 중요한 요인중에 하나임
 - 시계열을 다루는 경우는 국제표준, UTC를 선택함

시계열

- 파이썬 시간대 정보는 전 세계의 시간대 정보를 모아둔 올슨 데이터베이스를 저장하고 있는 thirdy party 라이브러리임