데이터 다루기 I

데이터 Join, Combine, Reshape

- 데이터 분석과 모델링 작업은 데이터를 불러오고, 다듬고, 변형하고 재정렬하는데 많은 시간을 소요됨.
- 데이터가 저장된 형태를 다른 형태로 바꾸기 위해서 파이썬, 펄, R, 자바, awk, sed같은 유닉스의 텍스트 유틸리티를 사용함.
- Pandas는 이런 작업을 유연하고 빠르게 처리가능함.
- 고수준의 알고리즘등을 처리할수 있는 Pandas라이브러리를 제공함

데이터 합치기

- Pandas.merge는 하나 이상의 키를 기준으로 DataFrame의 로우를 합침
- SQL이나 다른 관계형 데이터베이스이 join과 유사함
- Pandas.concat은 하나의 축을 따라 객체를 이어 붙어감
- Combine-first인스턴스는 두객체를 포개서 한 객체에서 누락된 데이터를 다른 객체에 있는 값을 채울수 있도록함

- 데이터베이스 스타일로 DataFrame합치기
  - Merge나 join연산은 관계형 데이터베이스의 핵심 연산으로 키를 하나 이상 사용해서 데이터 집합의 행을 합침
- 색인 merge하기
  - Merge하려는 키가 DataFrame의 색인일 수 있음. Left-index=True 혹승 right_index=True옵션을 지정해 해당 색인을 Merge키로 사용함
- 축따라 이어붙이기
  - 데이터를 합치는 또 다른 방법으로 이어붙이기(concatenation), 연결(binding), 적층(stacking)등이 있음

데이터 변형

- 중복 제거
  - DataFrame에서 중복된 행을 변경할 수 있는 경우
- 함수와 매핑 이용해 데이터 변형하기
  - 데이터를 다루다 보면 DataFrame의 컬럼이나 Series, 배열 안의 값을 기반으로 데이터를 변경하고 싶을때가 있는 경우
- 값치환
  - Fillna메소드를 사용해서 누락된 밧을 채우는 일은 일반적인 값 치환 작업이라고 할 수 있음
- 축 색인 이름 바꾸기
  - Series의 값처럼 축 이름 역시 유사한 방식으로 함수를 새롭게 바꿀 갑으로 이용해서 병합할 수 있음.
- 개별화와 양자화
  - 연속성 데이터는 종종 개별로 분할하거나 분석을 위해 그룹별로 나누기도 함