## 1.1 정형화된 데이터의 요소

정형 데이터의 종류

* 수치형(numeric) 데이터
  + 연속형(continuous) 데이터 – 온도, 풍속.
  + 이산(discrete) 데이터 – 횟수, 빈도
* 범주형(categorical) 데이터
  + 이진(binary) 데이터 – 0 or 1
  + 순서형(ordinal) 데이터 – 평점

데이터를 범주로 구분하는 이유:

* 데이터 분석을 할 때 데이터 종류가 결정적인 역할을 함. 종류마다 사용할 수 있는 기법(시각화, 통계 개념, 해석)이 다름.

## 1.2 테이블 데이터

테이블형 데이터

* 행(레코드, 사건)과 열(피쳐, 변수)로 구성된 2차원적인 데이터.

데이터 프레임과 인덱스

* Python: **pandas**
  + 기본적으로 숫자 인덱싱 있음. 다중/계층적 인덱싱 가능(pivot, reshape).
* R: **data.frame**
  + 마찬가지로 숫자 인덱싱 존재. 하지만 다중/계층적 인덱싱 불가.
    - 보완을 위해 dplyr, data.table 사용됨.

테이블형 이외의 데이터 구조

* 시계열 데이터, 공간 데이터, 그래프 데이터, …
  + 이 책에서는 다뤄지지 않음.

## 1.3 위치 추정

평균

* 가장 기본적인 위치 추정 방법. 모든 숫자들의 합을 숫자 개수로 나눔.
* 변형으로 절사평균, 가중평균 존재.
  + 절사평균: 값들을 크기 순서로 정렬한 후 양 끝 일정 개수의 값들을 제외하고 나머지 숫자들로 평균.
    - 극단값의 영향을 제거한다.
  + 가중평균: 숫자들에 가중치를 붙여 평균.
    - 특이 수치들이 존재할 때,
    - 데이터 그룹의 크기가 다를 때