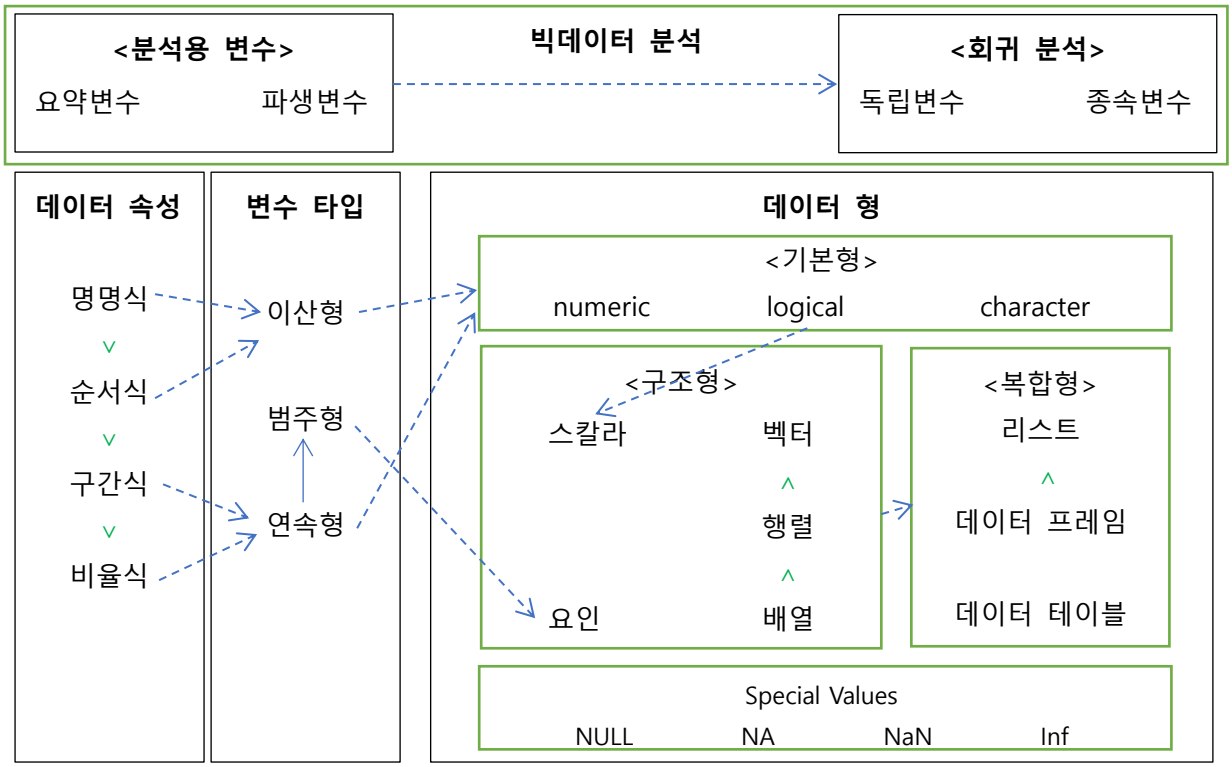
03. R 데이터 종류 및 구조 # src: 7_R/03_데이터종류및구조 참고



1) 데이터 종류

(1) R은 **데이터 속성**에 따라 명명식, 순서식, 구간식, 비율식으로 데이터 구분

① 명명식(Nominal): 명목척도. 이름으로 명명되는 자료 ex) 성별(남/녀)

② 순서식(Ordinal): 서열척도. 순서가 있는 명명식 ex) 소득 데이터(상/중/하)

③ 구간식(Interval): 간격척도. 순서의 간격을 측정할 수 있는 순서식 ex) 온도

④ 비율식(Ratio): 비율척도. 절대 영점이 존재해 비율이 의미있는 구간식 ex) 체중

(2) R은 **변수 타입**에 따라 연속형, 이산형, 범주형 변수 구분

① 연속형 변수(Continuous): 연속적인 값 ex) 체중

② 이산형 변수(Discrete): 개수가 정의된 변수 ex) 동전의 앞, 뒤

③ 범주형 변수(Categorical): 연속형 변수를 구간으로 묶어 이산형 변수로 만든 것 ex) 연령대

(3) R은 **데이터의 품질**에 따라 특이값과 결측값 구분

① 특이값(Outlier): 정상적이지 않은 데이터 ex) 잘못 측정된 값, 오차에 의한 값

② 결측값(Missing): 아직 측정되지 않은 값. NA로 표기

(4) R에서 **분석용 변수**는 요약변수와 파생변수로 구분

① 요약변수(Summary Variables): 데이터 분석을 위해 1차 가공한 변수

② 파생변수(Derived Variables): 분석자의 판단에 따라 특정 조건을 만족하는 변수

2) R 기본 데이터 타입: 문자(character), 숫자(numeric), 논리(logical)

3) Spevial Values: 기본형 외에 특별한 의미로 사용되는 예약어(NULL, NA, NaN, Inf)

4) **Factor**: 범주형(Categorical) 변수. 정해진 값 중 하나의 값으로 명명식이나 순서식 데이터 저장

- levels(): 범주의 목록 출력, nlevels(): 범주의 수 출력

5) **구조형 변수**와 **복합형 변수**

(1) **구조형 변수**: 변수가 한 가지 데이터 타입 값만 가질 수 있는 변수

스칼라(scalar), 요인(factor), 벡터(vector), 행렬(matrix), 배열(Array)

(2) **복합형 변수**: 서로 다른 데이터 타입 값을 가질 수 있는 변수

리스트(list), 데이터 프레임(data.frame), 데이터 테이블(data.table)

6) **벡터(Vector)**; 동일 형 데이터들의 집합 (인자들은 한가지 유형의 스칼라 타입이어야함)

(1) 문자, 숫자, 논리 타입이 섞여 있을 경우 문자 타입으로 자동 형변환 됨(문자>숫자>논리 순)

#숫자와 논리타입이 섞여 있을 경우 TRUE는 1, FALSE는 0으로 형변환

(2) 벡터의 결합: c() 함수 이용하여 여러 벡터를 하나의 벡터로 결합 가능, append() 벡터 추가

(3) 벡터 간 합집합, 교집합, 차집합, 비교: union(), intersect(), setdiff(), setequal()

7) **리스트(List)**; 복합 구조형의 벡터에 해당하는 데이터 타입

8) **행렬(Matrix)**; 행과 열을 가지는 2차원 배열, 행렬은 열 우선(행 우선: byrow=TRUE)

(1) 행렬 곱: %*%

(2) 전치행렬: **t(행렬)**

(3) 역행렬: **solve(행렬)**

(4) 행렬 이용한 선형방정식 풀이: ab <- solve(t(행렬)%*%행렬) %*% t(행렬) %*% matrix(y, ncol=1))

lines(x, x*ab[1]+ab[2])

9) **배열(Array)**; 3차원 이상의 데이터를 다룰 경우 사용 array(data, dim=c(행 수, 열 수 면 수))

10) **데이터 프레임(data.frame)**; 2차원 구조의 복합형 데이터 타입

(1) 데이터 프레임 합치기: cbind() 열단위 합침(행 수 동일), rbind() 행단위 합침(열 수 동일)

11) **타입 판별 및 타입 변환**

(1) 타입 판별: class(), is.*()

(2) 타입 변환: as.*(); as.numeric(), as.factor(), as.data.frame(), as.matrix() 등

12) 문자열과 날짜

(1) 문자열: nchar(), substr(), paste(), strsplit(), gsub()

(2) 날짜: Sys.Date(), as.Date() 등