데이터 전처리 개요

#01. 데이터 전처리의 이해

데이터를 본격적으로 분석하기 전에 분석에 적합하게 데이터를 가공하는 작업

데이터 가공(Data Manipulation), 데이터 핸들링(Data Handling)도 비슷한 의미로 사용되는 용어

데이터 전처리에서 수행되는 주요 작업

행,열 재배치

단일 데이터 프레임에 대한 작업

- 행 혹은 열(변수)에 대한 재배치, 이름 변경
- 정렬
- 특정 데이터 필터링
- 행 혹은 열 추가, 삭제

데이터 재배치

두 개 이상의 데이터프레임을 다루거나 새로운 데이터프레임이 생성되는 형태

- 다른 데이터 프레임과 데이터 합치기 (열)
- 다른 데이터 프레임과 데이터 합치기 (행)
- 피벗 테이블
- 교차표

#02. 변수의 이해

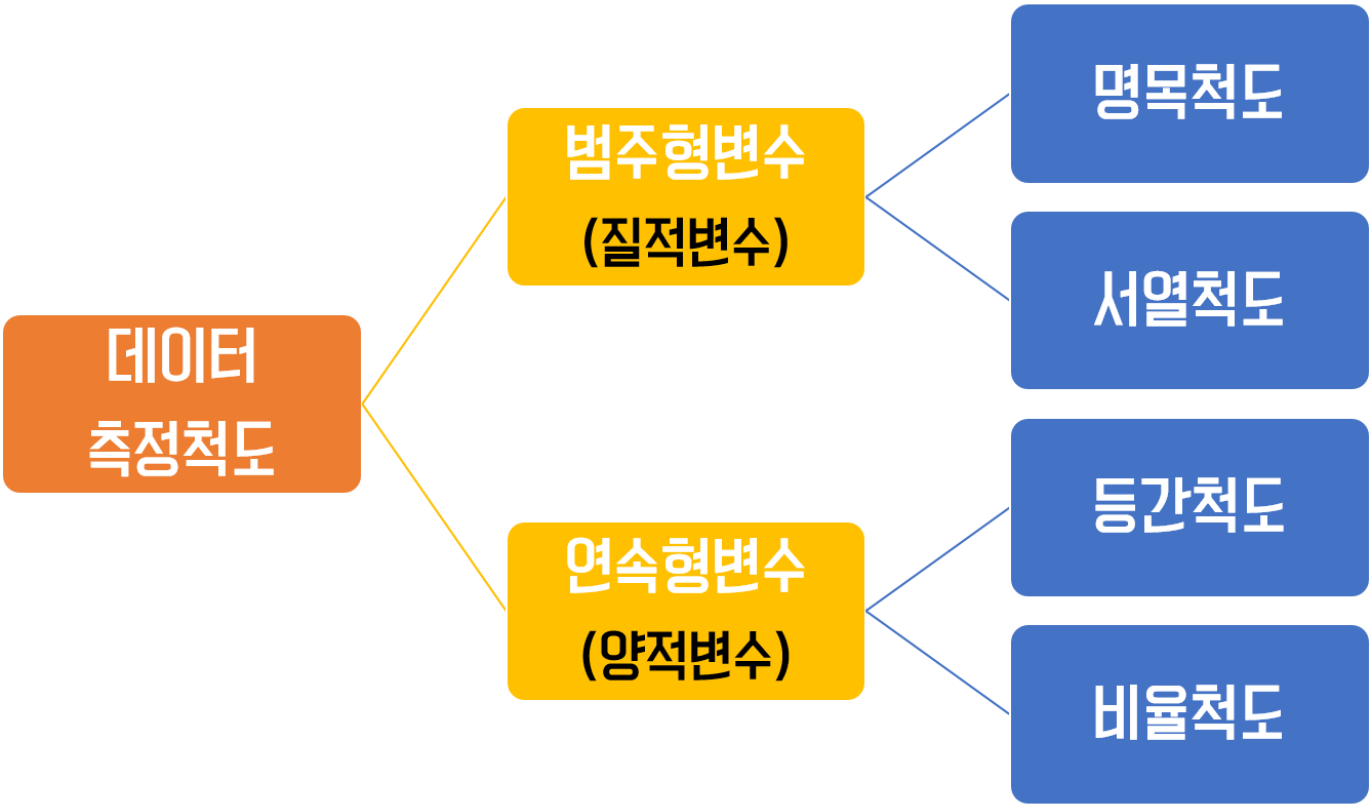
1) 변수와 데이터

용어	설명
변수	각 단위(Unit, row)에 대해 관측되는 특성
데이터	하나 이상의 변수에 대한 관찰값의 모음

2) 변수의 종류

척도

사물이나 사람 등 대상의 특성을 통계상의 수로 표현하기 위해 체계적으로 숫자를 부여한 것



범주형 자료 (categorical qualitative)

질적자료라고도 한다

명목척도

- 대상을 특성에 따라 카테고리로 분류하여 기호를 부여한 것.
- 측정이 이루어지는 항목들이 상호배타적인 특성만 가진척도
- 비교 방법 : 확인, 분류
- 연산 방법 : `=`
- 통계값 : 최빈치
- 적용 가능한 추론통계 방법 : 비모수 통계, 빈도 분석, 교차 분석
- 예시: 성별, 이름, 지역, 학년 등

서열척도

- 대상의 특성들을 구분할 수 있으며 이들 사이의 상대적인 크기를 나타낼 수 있고 서로 간 비교가 가능한 척도
- 명목척도들 중 항목들 간에 서열이나 순위가 존재하는 척도
- 비교방법: 순위비교
- 연산: `=`, `<`, `<=`, `>`, `>=`
- 통계값: 최빈값, 중앙값
- 적용 가능한 추론통계 방법 : 비모수 통계, 서열 상관관계
- 예시: 교육정도(중졸, 고졸, 대졸 이상), 선호도 순위, 학점

연속형자료(Numerical quantitative)

양적자료라고도 한다.

등간척도

- 상호간의 서열뿐 아니라 인접한 두 변수 값의 차이가 일정한 변수
- 서열척도들 중 항목들 간의 간격이 일정한 척도
- 비교방법 : 간격비교
- 연산 : `+`, `-`, `=`, `<`, `<=`, `>`, `>=`
- 통계값 : 최빈값, 중앙값, 산술평균
- 적용 가능한 추론통계 방법 : 모수 통계
- 예시: 온도, 연도(올림픽, 월드컵), IQ, 만족도(매우불만족,약간불만족,보통,약간만족,매우만족)

비율척도

- 상호간 서열, 크기 차이, 크기의 비교, 특성들 간의 계산까지 가능한 척도
- 등간척도 중 아무 것도 없는 상태를 `0`으로 정할 수 있는 척도
- 예시: 몸무게, 키, 길이, 임금, 나이(`20세 이하` , `21~30세` , `31~40세` , `41~50세` , `0` 이라는 개념은 아직 태어나지 않음을 뜻함)