



# **[ProDS]** 머신러닝 이론 및 데이터 처리





# 데이터 전처리: 데이터 생성, 데이터 정제





# Key words

#요약변수 #파생변수

#이상치 #결측치 #Binning



# 데이터 생성

## I 요약변수 VS 파생변수

### ■ 요약변수

- 수집된 정보를 분석의 목적에 맞게 **종합(aggregate)**한 변수.
- 많은 모델에 공통으로 사용될 수 있어, 재사용성이 높음.

(예) 단어 빈도, 상품별 구매 금액, 상품별 구매량, 영화 매출액

객관적이고 누구나 동일하게 생성가능한 변수

재활용하기 쉽다.

# 데이터 생성

## I 요약변수 VS 파생변수

### ■ 파생변수

- 특정한 의미를 갖는 작위적 정의에 의한 변수.
- 사용자가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수.
- 주관적일 수 있으므로 논리적 타당성을 갖추어야함.  
(예) 구매상품 다양성 변수, 가격 선호대 변수, 라이프 스타일 변수, 영화 인기도 변수

# 데이터 정제

## I 결측값의 이해 missing value

- 기록누락, 미응답, 수집오류 등의 이유로 결측이 발생.
- 결측값이 포함된 자료라도 나머지 변수의 값들은 의미있는 정보이므로, 정보의 손실을 최소화 하도록 결측을 처리하는 것이 바람직함.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN



	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

NaN이 결측값이다.

결측값을 적절한 방식(평균값)으로 대체

# 데이터 정제

## I 결측값 처리법

- **완전제거법** (list-wise deletion)
  - 결측값이 하나 이상 포함된 자료를 제거하는 방법.
  - 정보의 손실로 분석 결과가 왜곡될 수 있음.
- **평균대체법** (mean value imputation)
  - 결측값을 해당 변수의 나머지 값들의 평균으로 대체하는 방법. 결측이 사라지긴 한다.
  - 추정량의 표준오차가 과소추정되는 문제가 있음.

평균값은 값들의 사이에 있다. 하지만 대체 값이 그 범위 안에 있다고 확정하지 못하기 때문이다.

자료의 변동성이 작아진다.

# 데이터 정제

## I 결측값 처리법

- 핫덱대체법 (hot deck imputation)

- 동일한 데이터 내에서 결측값이 발생한 관찰치와 유사한 특성을 가진 다른 관찰치의 정보를 이용하여 대체하는 방법.
- 정보의 손실로 분석 결과가 왜곡될 수 있음. 값을 여러개 선택하고 그 중에서 랜덤하게 대체해줌

- 그밖의 결측값 처리법

- Regression imputation, kNN imputation 등.

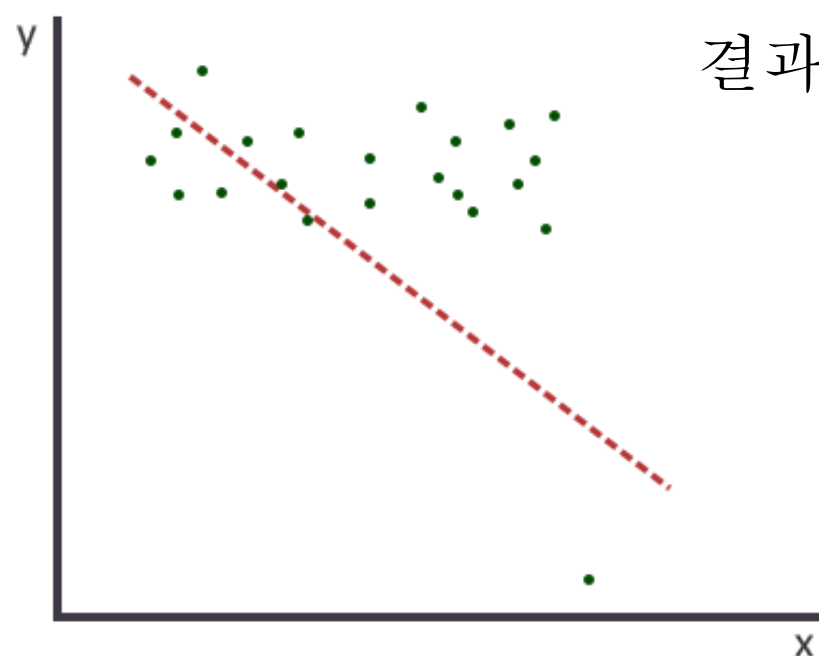


# 데이터 정제

## I 이상값의 이해

여러가지 경우가 존재한다.

- 이상값은 다른 데이터와 동떨어진 것을 말함.
- 다른 자료값들에 비해 멀리 떨어져 있지만 의미가 있는 값일 수도 있고, 단순히 입력 오류로 발생한 값일 수도 있음.



결과에 적용할 것인지 판단하는 것이 중요하다.

회귀 분석의 사례 -> 선형회귀

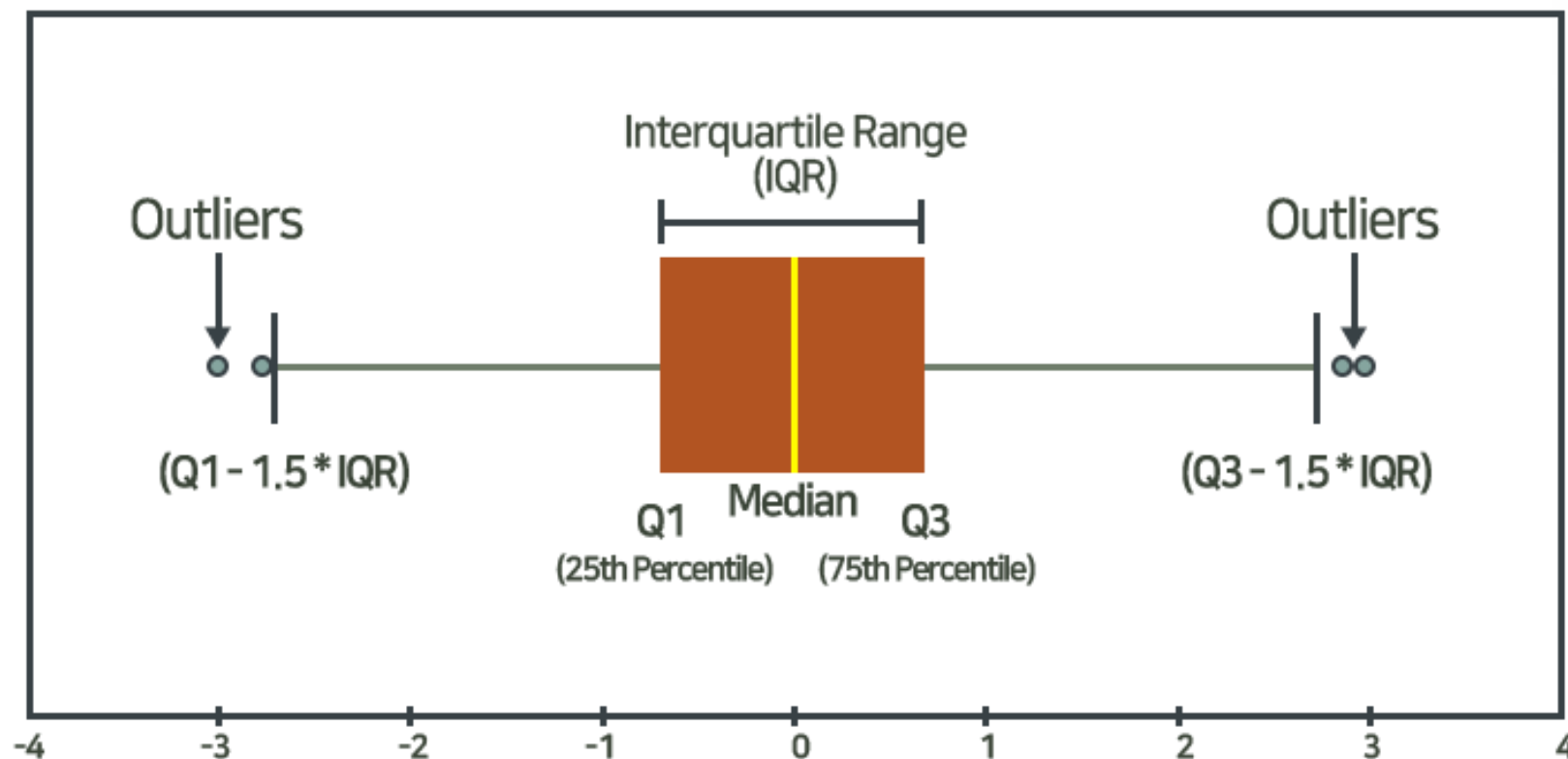
x 가 커지면 y의 값이 작아진다.

# 데이터 정제

**I 이상값의 탐지** 25%, 75% 를 기준으로 범위에 있는지를 판단해준다.

■ 상자그림 이 상자 범위 안에 있으면 정상값으로 판단해준다.

-  $Q1 - 1.5 \times IQR$  과  $Q3 + 1.5 \times IQR$  의 범위를  
넘어가는 자료를 이상값으로 진단. tuckey 의 계수





# 데이터 정제

## I 이상값의 탐지

- 표준화 점수(Z-score)

- 표준화 점수의 절대값이 2, 3다 큰 경우를 이상값으로 진단.

$Z\text{-score} : X - \text{표준편차} / \text{평균}$

보통 0과 1 사이의 값을 가진다.

# 데이터 정제

## I 이상값 처리 방법

- 이상값 제외 (trimming)

- 처리는 간단하지만, 정보 손실이 발생하고 추정량 왜곡이 생길 수 있음.

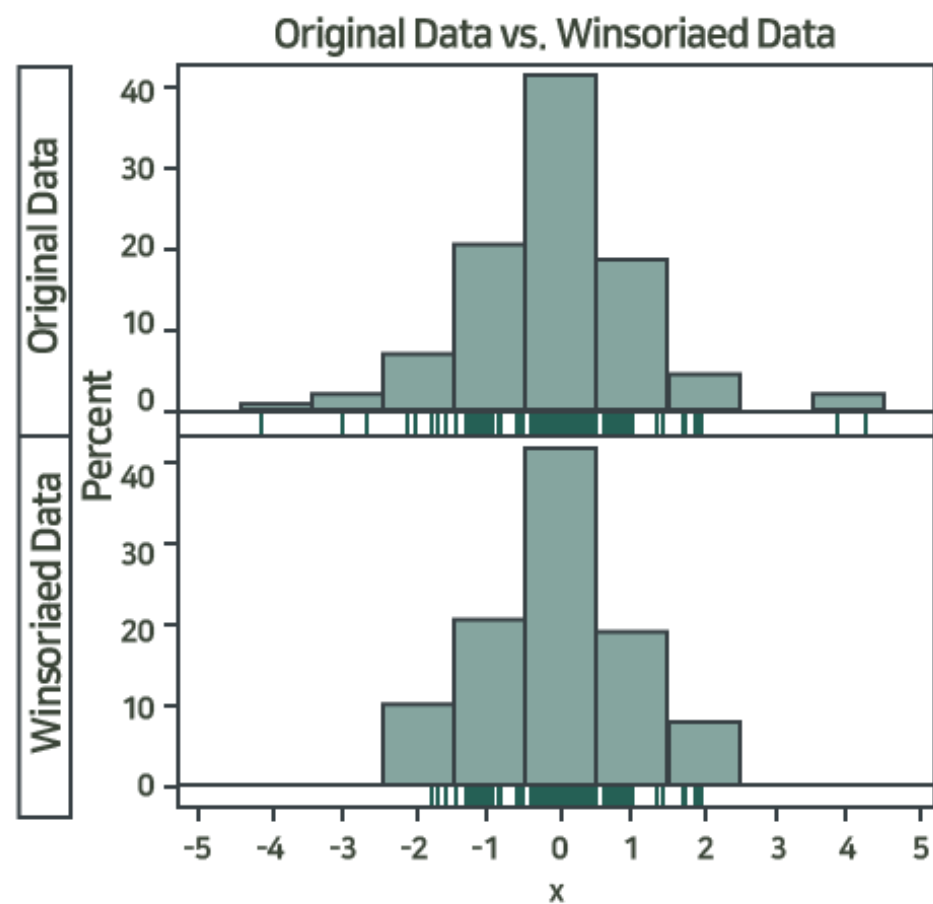
일정 자료들만으로도 정보를 해석하게 될 수도 있기 때문에



# 데이터 정제

## I 이상값 처리 방법

- 이상값 대체 (winsorization) 정상값의 범위를 정해주고 사용한다.
  - 이상값을 정상값 중 최대 또는 최소 등으로 대체하는 방식.



# 데이터 정제

## I 이상값 처리 방법

- 변수변환

- 자료값 전체에 로그변환, 제곱근 변환 등을 적용.



# 데이터 정제

## I 연속형 자료의 범주화

- 변수구간화(binning)

- 연속형 변수를 구간을 이용하여 범주화 하는 과정.

AGE	AGE_bins
10	[10, 21]
15	[10, 21]
16	[10, 21]
18	[10, 21]
20	[10, 21]
30	[22, 33]
35	[34, 45]
42	[34, 45]
48	[46, 55]
50	[46, 55]
52	[46, 55]
55	[46, 55]

# 데이터 정제

## I 연속형 자료의 범주화

- 변수구간화(binning)의 효과
  - 이상치 문제를 완화.
  - 결측치 처리 방법이 될 수 있음. overfit
  - 변수간 관계가 단순화 되어 분석 시 과적합을 방지할 수 있고, 결과해석이 용이해짐.

빈도를 통해 보완을 해줄 수 있다.

새로운 범주로 표현할 수 있다. 따로 관리하기 용이해진다.

범주의 특성 상 정확도가 떨어질 수 있지만, 범주를 잘 정해준다면 정확도를 높여줄 수 있다.





# 데이터 전처리: 데이터 변환, 데이터 결합





# Key words

#로그변환 #제공근변환 #박스콕스  
#이너조인 #레프트조인  
#라이트조인 #풀아우터조인



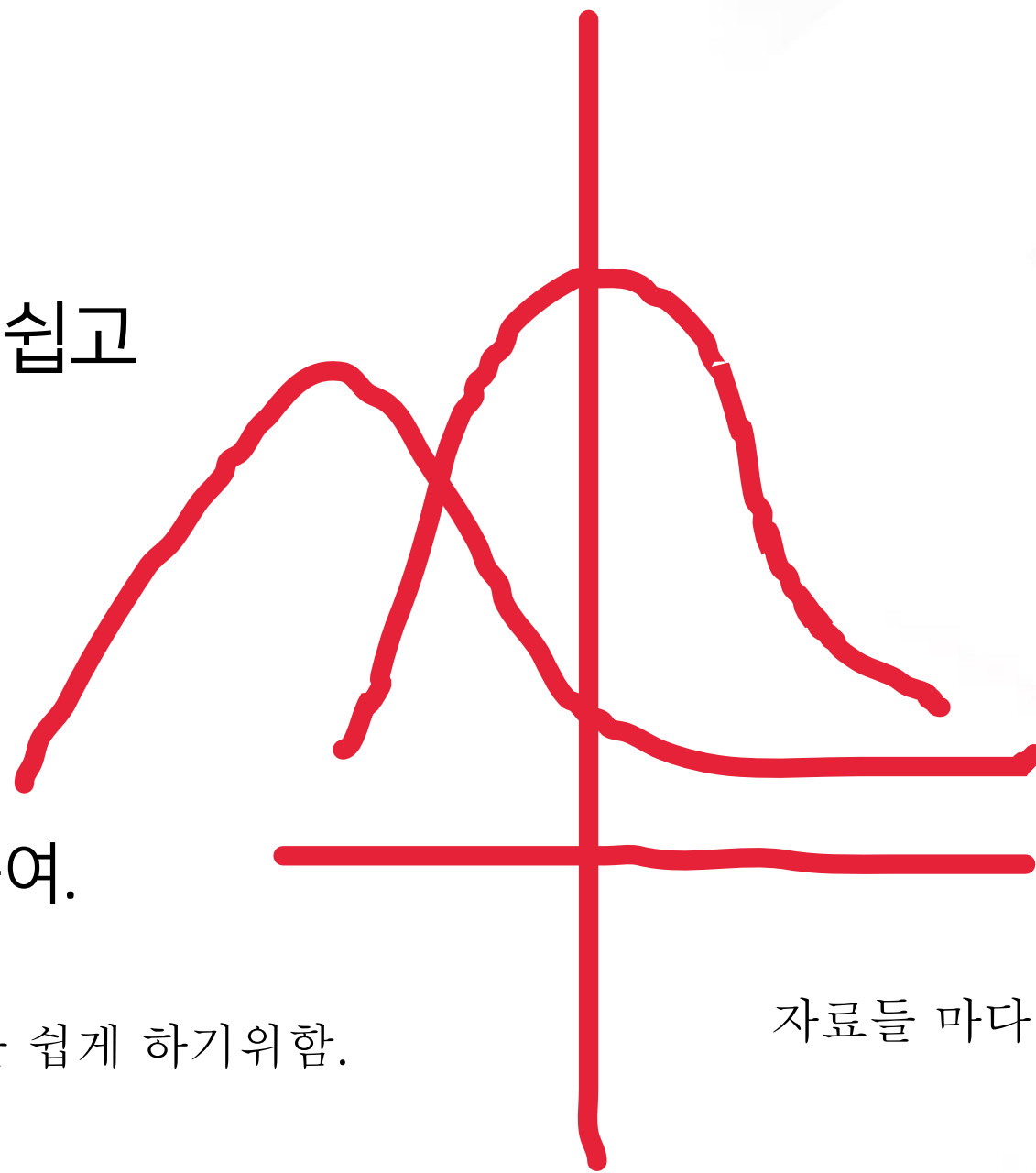
# 데이터 변환

## 데이터 변환

- 자료 변환을 통해 자료의 해석을 쉽고 풍부하게 하기 위한 과정.
- 데이터 변환 목적
  - 분포의 **대칭화**.
  - **산포**를 비슷하게 하기 위하여.
  - **변수 간 관계**를 **단순**하게 하기 위하여.

선형 -> 비선형으로 변경시켜 관계 파악을 쉽게 하기위함.

자료들마다 산포가 다르다.





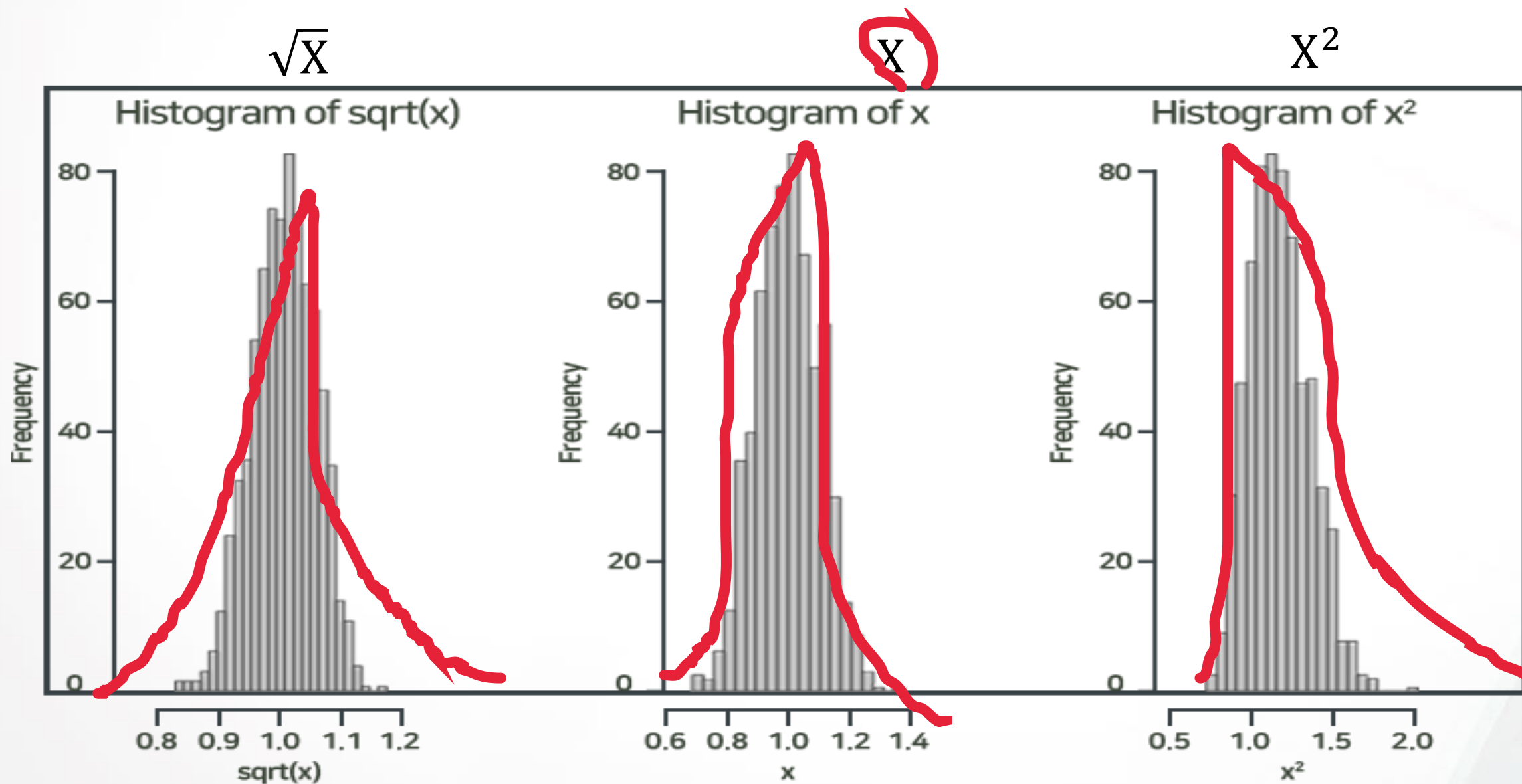
# 데이터 변환

## 데이터 변환

- 변환 유형 1 : 제곱근 변환 VS 제곱 변환



x 를 증가 (제곱) 또는 y 를 감소 (로그) 시키면  
선형으로 변경가능



대칭화작업

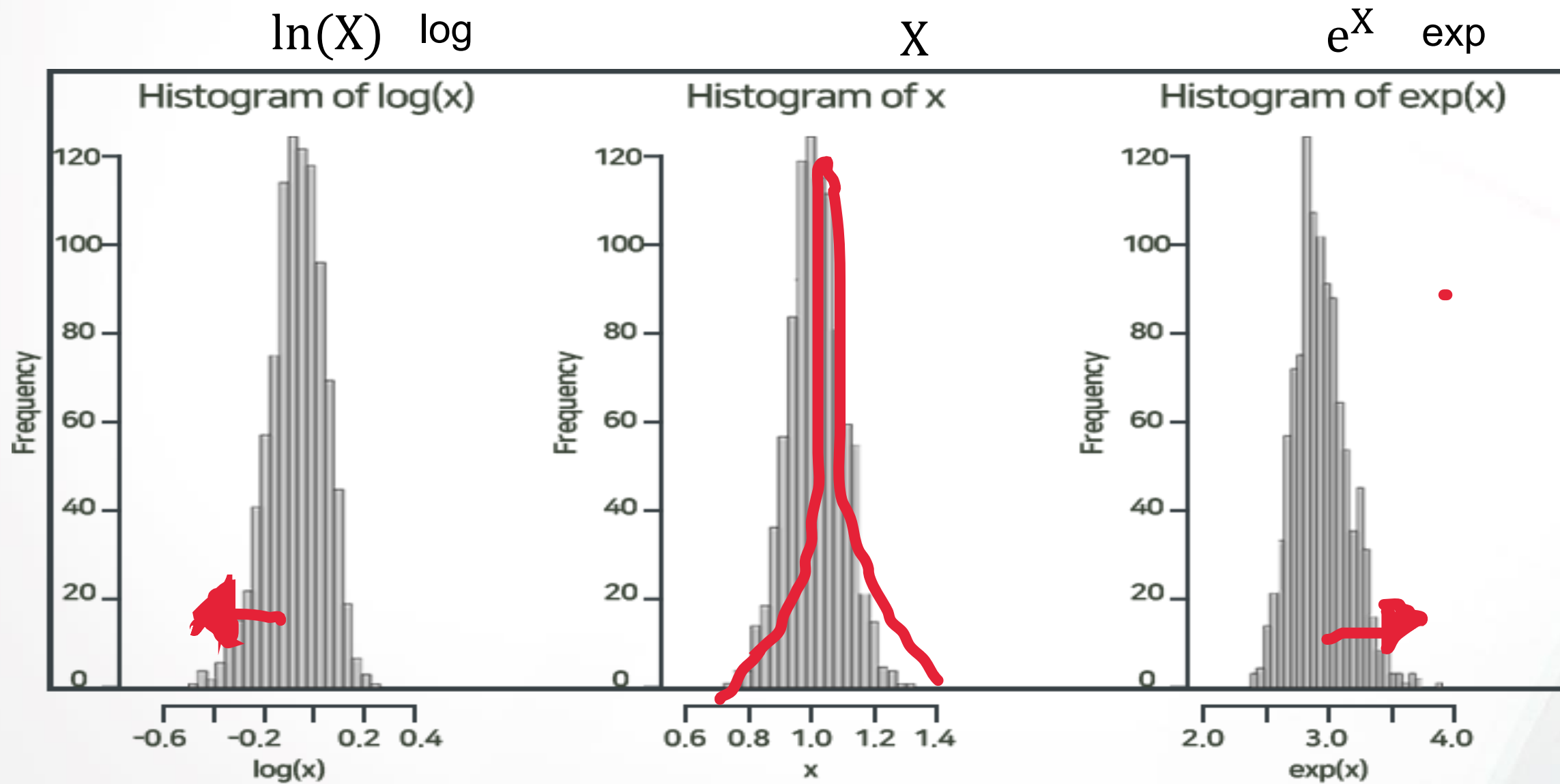
대칭형작업 가능해짐

# 데이터 변환

분포가 왼쪽이 길면 제곱  
오른쪽이 길면 로그변환

## 데이터 변환

- 변환 유형 2 : 로그 변환 VS 지수 변환



# 데이터 변환

## ■ 박스콕스 변환 (Box-Cox Transform)

-  $y = \frac{1}{p} ((x + 1)^p - 1)$ ,  $p = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$ : **제공된 유형의 변환을 일반화.**

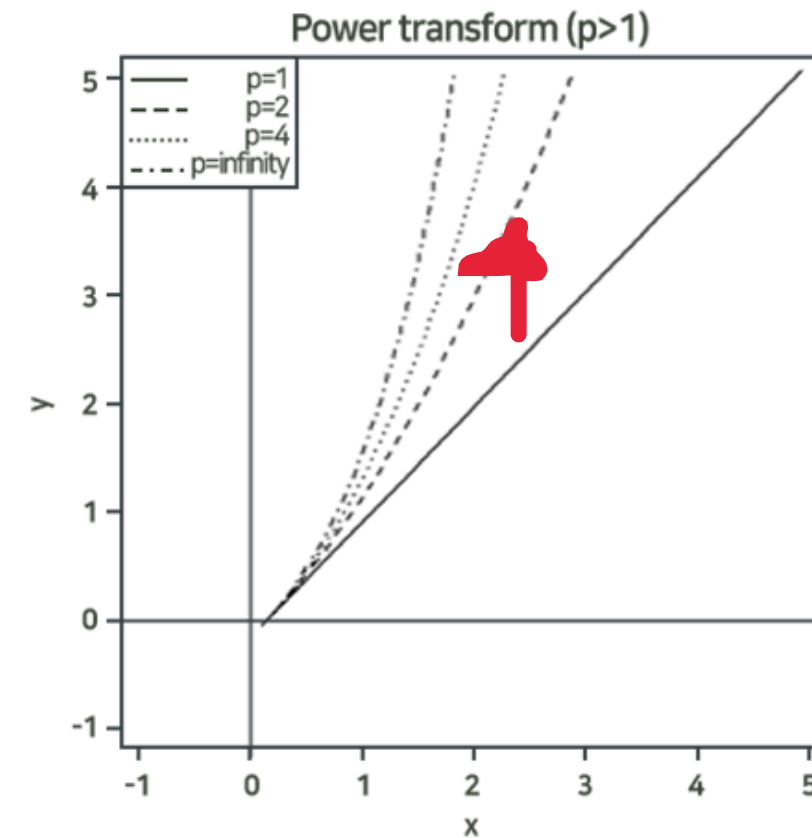
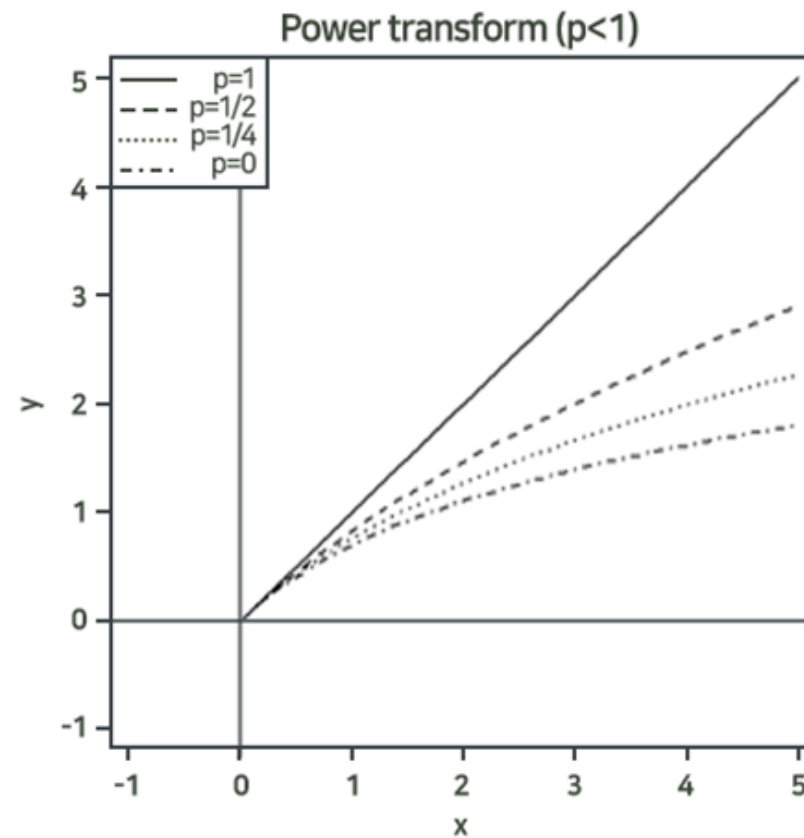
-  $y = \ln(x + 1)$ ,  $p = 0$

-  $y = (1 + \frac{x}{p})^p - 1$ ,  $p = 2, 4, 8, \dots$ : **제공된 유형의 변환을 일반화.**

$\exp(x+y) \rightarrow$  지수화

모두 원점을 통과한다.

시작하는 점의 기울기는  $y = x$  즉, 1이다.





# 데이터 결합

## I 데이터 결합

X1	X2
A	1
B	2
C	3

+

X1	X3
A	T
B	F
D	T

X1 : 키(key) 변수

- 이너조인(inner join)

- 두 테이블에 키(key)가 **공통으로** 존재하는 레코드(record)만 결합.

- (A, 1, T), (B, 2, F)

- 풀아우터조인(full outer join)

- 두 테이블 중 **어느 한쪽이라도** 존재하는 키에 대한 레코드를 모두 결합. 나머지는 또 채워준다.

- (A, 1, T), (B, 2, F), (C, 3, NA), (D, NA, T)



# 데이터 결합

## I 데이터 결합

X1	X2
A	1
B	2
C	3

+

X1	X3
A	T
B	F
D	T

X1 : 키(key) 변수

### ■ 레프트 조인(left join)

- 왼쪽 테이블에 존재하는 키에 대한 레코드를 결합. + 공통으로 존재

-(A, 1, T), (B, 2, F), (C, 3, NA)     a, b 가 남는다.

### ■ 라이트 조인(right join)

- 오른쪽 테이블에 존재하는 키에 대한 레코드를 결합. + 공통으로 존재

-(A, 1, T), (B, 2, F), (D, NA, T)     a, b 가 남는다.