· 03/30

Ozone data → python 구현 (?) / Sklearn 구현 (O) → ?

내부적의 처리를 통해 사용

(100만건 이상)

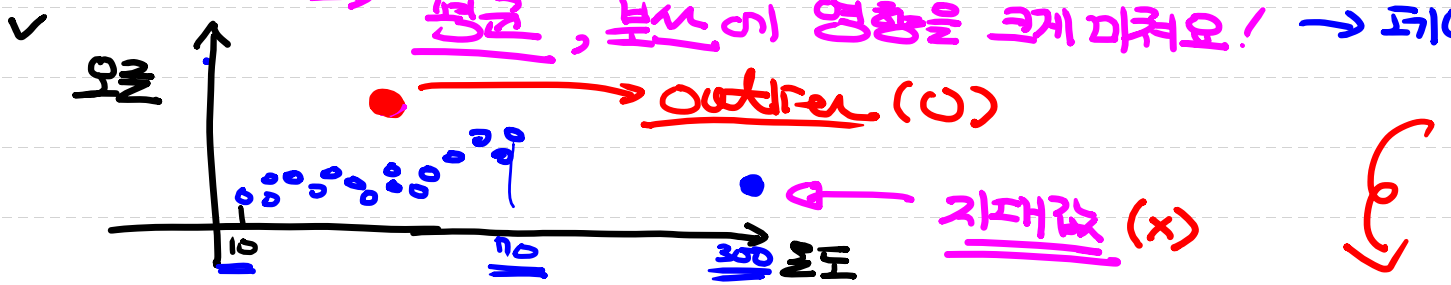
① Missing value 처리 (결측값)

- 삭제처리 (전체데이터가 충분히 많고 Missing value가 5% 이내)
- 대체처리
  - 평균, 중위, 최대, 최대, 최빈
  - 머신러닝 기법을 활용 (더 좋은 방식)
    - ↳ Missing value가 종속변수일때

② 이상치 처리

이상치는 값이 일반적의 다른 데이터에 비해 편차가 큰 데이터

↳ 평균, 분산에 영향을 크게 미쳐요! → 데이터를 신뢰하기 힘들게 만드는 요소!

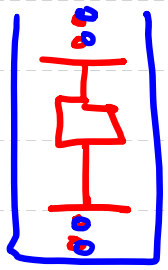


이상치 (O)

정상값 (X)

이상치 판단
(outlier)

- 눈으로 확인 (그래프로 그려서) → Boxplot, Scatter



- 계산을 통해 이상치 파악

말로 기법이 존재하지만

① Turkey Fence ✓

② 표준편차방법 (Z-score)

Boxplot

이상치 ○○○ ↑

☆ 3사분위 + (IQR value × 1.5)

IQR value = 3사분위값 − 1사분위값

IQR (Interquartile Range)

→ 3사분위

→ 2사분위

→ 1사분위

IQR value × 1.5

☆ 1사분위 − (IQR value × 1.5)

이상치 ○○○ ↓

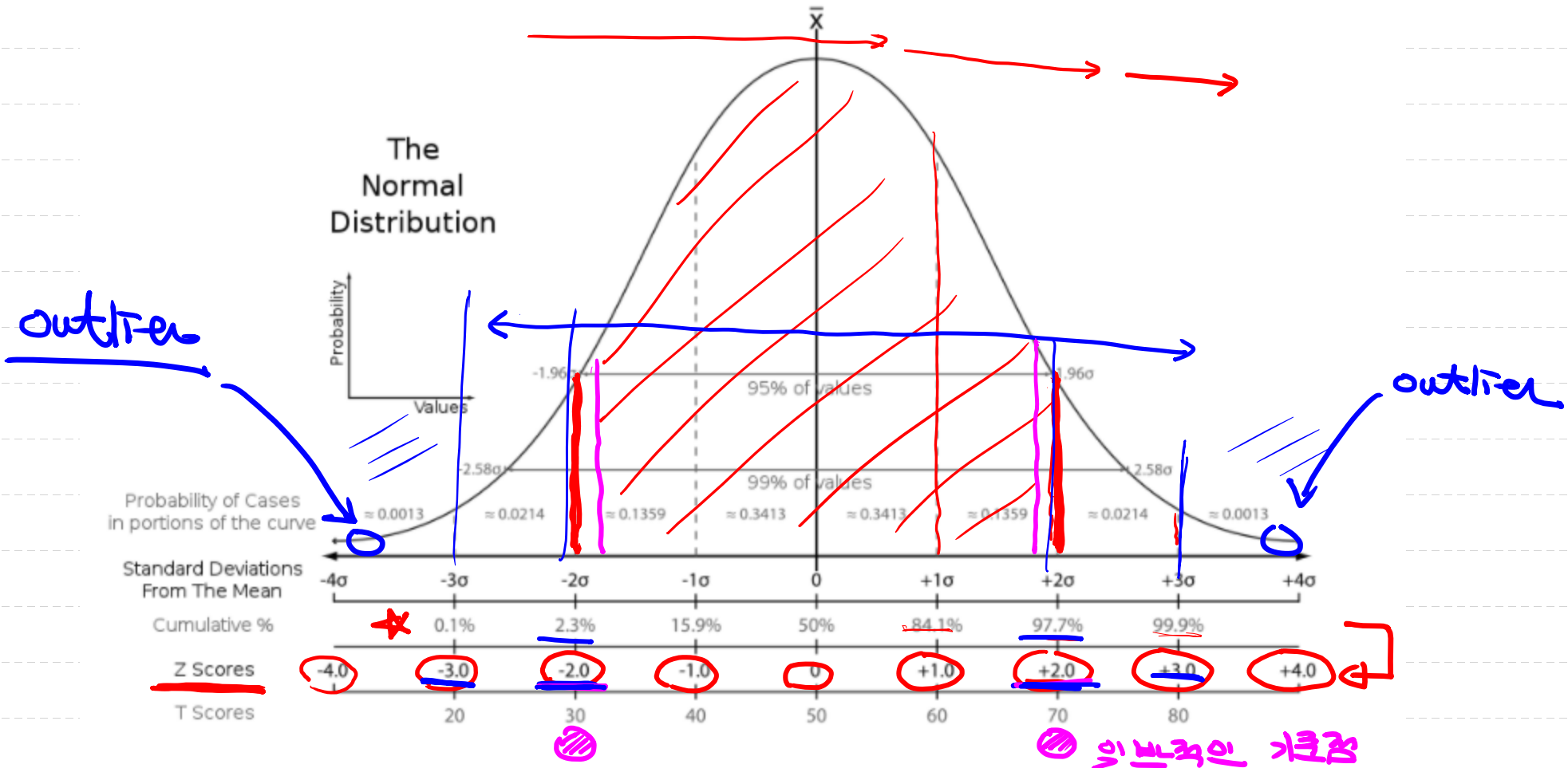
표준점수
(표준값, Z-score)

를 이용한 이상치 판별 방식

★ Z-score = (x − 평균) / 표준편차

[1, 2, 3, 4, … 22.1]

0 0 0 0 0 → Z-score



일반적인 기준점

⑧ 이상치를 판별할 수 있어요 !!

→ 하지만 그 값이 실제 "이상치" 인지는 조사의 필요 !!! (?)

그러면 우리의 목표 돌아온다 !!

- Python 구현.
- Sklearn 구현

구현결과의 차이로 "이상치" 때문인지를 추측해 보아요 !!

⇓

탐색결과

→ 아직도 Sklearn 구현 / Python 구현 차이가 심해요.

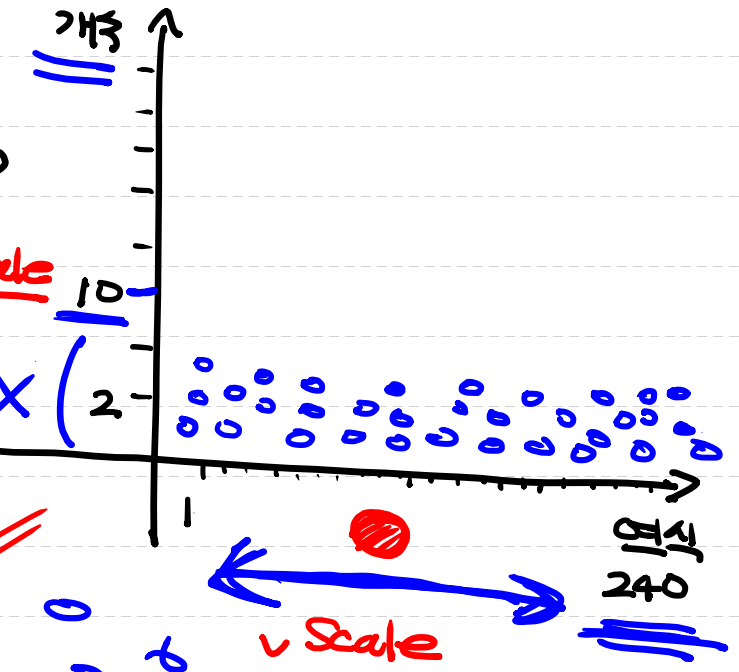
↳ 이상치때문에 문제가 된 것 아니었어요 ?

정규화 (Normalization)

Sklearn을 자체적으로 정규화 진행 (내부적으로)

"데이터가 가진 feature 들의 scale을 맞추는 작업"

집을 사러가요 (아파트)
- 방의 개수 2 ~ 10 개
- 연식(월) 1 ~ 240 개월



(Normalization) 정규화 기법

- Min-Max Scaling

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Standardization (표준화)
  - Z-score $\frac{X - \mu}{\sigma}$

최소 → 0
최대 → 1
] 최대값 최소값을 이용하기 때문에 "이상치에 민감"

"-2 ~ 2" 사이의값
양수, 음수 모두 있고 이상치에 영향을 덜 받아요