개요

- 이상치(outlier)란 일부 관측치의 값이 전체 데이터의 범위에서 크게 벗어난 아주 작거나 큰 극단적인 값을 갖는 것을 의미
- 데이터의 모집단 평균이나 총합을 추정하는 것에 문제를 일으키며, 분산을 과도하게 증가시켜 분석이나 모델링의 정확도를 감소시키기 때문에 제거하는 것이 좋다
- 특히 자료 수집의 오류로 발생한 이상치의 경우 다른 관측치에 비해 극단적인 값이 들어가는 경우가 많아 분석 성능에 큰 영향을 미치기 때문에 가능한 한 제거해 주어야 함

(1) 관측값의 형식과 다른 형식의 값으로 표시된 결측치	(2) 관측값의 형식과 같은 형식의 값으로 표시된 결측치	(3) 자료 수집의 오류로 발생한 이상치	(4) 다른 관측치들과는 현저히 차이나는 실제 관측치
몸무게	성 적	거실 온도	일자별 게임시간
60.0	90	22.4	2
55.5	85	22.3	1
ERROR	999	2345	17
70.5	100	22.1	4

- 2,3번의 경우 실제 환경에서 발생하지 않을 값으로 저장된 이상치
- 모델 전체에 영향을 미칠 수 있는 극단값으로, 이를 정제하지 않는다면 평균 등의 연산을 하는 경우 관측값과 전혀 다른 값이 도출될 수 있다.
- 4번의 경우 실제로 수집된 값이지만 아주 특이한 값으로 모델 전체에 악영향을 줄 수 있기 때문에 정제해 주는것이 좋다

이상치 확인

- 데이터 정의서에 값의 범위가 정해진 경우 규칙에 따라서 필터링하여 이상치를 파악할 수 있음
- 시험 성적등 일반적으로 알려진 데이터의 구간이 있는 경우(0~100)에는 데이터의 구간이 벗어나는 값 파악을 통해 이상치를 확인할 수 있다
- 하지만 그렇지 않은 경우 분석가가 이상치를 판단하는 기준을 정해야 함
- 일반적으로 수치형 변수의 경우 IQR 방식을 이상치 판단 기준으로 사용할 수 있음

IQR

- Inter Quantile Range

- Box plot의 이상치 결정 방법을 그대로 이용