데이터 정제

데이터 분석에 앞서 전처리가 완료된 데이터에 대해 빈값(결측치)이나 정상 범위를 벗어난 값(이상치)들을 제거하거나 다른 값으로 대체하는 처리

#01.결측치

- 비어있는 값 (DB에서의 NULL과 비슷한 의미)
- 현장에서 만들어진 실제 데이터는 수집 과정에서 발생한 오류로 인해 결측치를 포함하고 있는 경우가 많다.
- 결측치가 있으면 통계 처리 함수가 적용되지 않거나 분석 결과가 왜곡되는 문제가 발생한다.
- 결측값 자체의 의미가 있는 경우도 있는데 예를 들면 쇼핑몰 가입자 중 특정 거래 자체가 존재하지 않는 경우와 인구통계학적 데이터(demographic data) 에서 아주 부자이거나 아주 가난한 경우 자신의 정보를 잘 채워 넣지 않기 때문에 가입자의 특성을 유추하여 활용할 수 있다.
- 결측값 처리는 전체 작업속도에 많은 영향을 준다.

1. 결측치 처리 방법

1) 소거법

결측치가 포함된 행이나 열을 삭제

결측치의 비율이 미미하다면 가장 손쉬운 방법이지만 결측치가 중요한 위치를 차지하거나 비율이 높다면 결측치 소거로 인해 통계 결과가 의미 없어질 수 있다.

2) 대치법

평균 대치법 (Mean Imputation)

- 관측 또는 실험을 통해 얻어진 데이터의 **평균**으로 대치한다.
- 비조건부 평균 대치법 : 관측 데이터의 평균으로 대치
- 조건부 평균 대치법 : 회귀분석을 활용한 대치법

다중 대치법 (Multiple imputation)

- 단순대치법을 한번하지 않고 m번의 대치를 통해 m개의 가상적 완전 자료를 만드는 방법이다 .
- 1단계 : 대치 (imputation step), 2단계 : 분석 (Analysis step), 3단계 : 결합 (combination step)

#02. 이상치

정상 범주에서 크게 벗어난 값

1. 이상치로 판단하는 경우

이상값을 꼭 제거해야 하는 것은 아니기 때문에 분석의 목적이나 종류에 따라 분석가의 적절한 판단이 필요하다.

- 의도하지 않게 잘못 입력한 경우 (Bad data)
- 의도한바와 같이 입력되었으나 분석 목적에 부합되지 않아 제거해야 하는 경우 (Bad data)
- 의도하지 않은 현상이지만 분석에 포함해야 하는 경우
- 의도된 이상값 (fraud, 불량)인 경우

2. 극단치

이상치의 한 종류.

오류는 아니지만 굉장히 드물게 발생하는 극단적인 값.

> 💡 ex) 초등학생의 몸무게 변수에 200kg 이상의 값이 있다면, 존재할 가능성은 있지만 굉장히 드문 경우이므로 극단치라 볼 수 있다.

3. 이상치의 인식과 처리

먼저 어디까지를 정상 범위로 볼 것인가를 분석가의 주관에 따라 결정해야 한다.

논리적으로 판단하여 정하기

성인의 몸무게가 40~150kg를 벗어나는 경우는 상당히 드물 것으로 판단하고, 이 범위를 벗어나면 극단치로 간주하는 것이다.

통계적인 기준을 이용하기 : ESD (Extreme Studentized Deviation)

상하위 `0.3%` 또는 `+-3` 표준 편차에 해당할 만큼 극단적으로 크거나 작으면 극단치로 간주하는 방법

 img1

$기하평균 - 2.5 x 표준편차 < data < 기하평균 + 2.5 x 표준편차$

상자그림

- 중심에서 크게 벗어난 값을 극단치로 간주.
- 극단치가 원으로 표시된다.

 img2

사분위수 직접 계산

아래의 수식을 벗어나는 데이터

$Q1 - 1.5 x (Q3 - Q1) < data < Q3 + 1.5 x (Q3 - Q1)$

 img3