02.데이터 생성, 데이터 정제

- 데이터 전처리 : 데이터 생성, 데이터 정제

  데이터 마이닝 : 대용량의 데이터로부터 패턴을 파악하고.

  우리에게 유용한 의미 있는 정보를 만들어 내는 분야

  데이터 과학 발전 - 데이터 수집기술, ICT기술, 하드웨어발전, 다양한 학습알고리즘

  데이터 기반 의사결정 - 객관적, 합리적, 새로운 시각 제공

1. 데이터 생성
   a. 요약 변수
      i. 수집된 정보를 분석의 목적에 맞게 종합(aggregate)한 변수 (객관적인 변수여야 함)
      ii. 많은 모델에 공통으로 사용될 수 있어, 재활용성이 높음
      ex) 단어 빈도, 상품별 구매 금액, 상품별 구매량, 영화 매출액
   b. 파생 변수
      a. 특정한 의미를 갖는 작위적 정의에 의한 변수.
      b. 사용자가 특정조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수
      c. 매우 주관적일수 있으므로 논리적 타당성을 갖추어야함.

ex) 구매 상품 다양성 변수, 가격 선호대 변수, 라이프 스타일 변수, 영화 인기도 변수

2. 데이터 정제
    a. 결측값의 이해
        i. 기록누락, 미응답, 수집오류 등의 이유로 결측이 발생
        ii. 결측값이 포함된 자료라도 나머지 변수의 값들은 의미있는 정보이므로, 정보의 손실을 최소화 하도록 결측을 처리하는것이 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

→

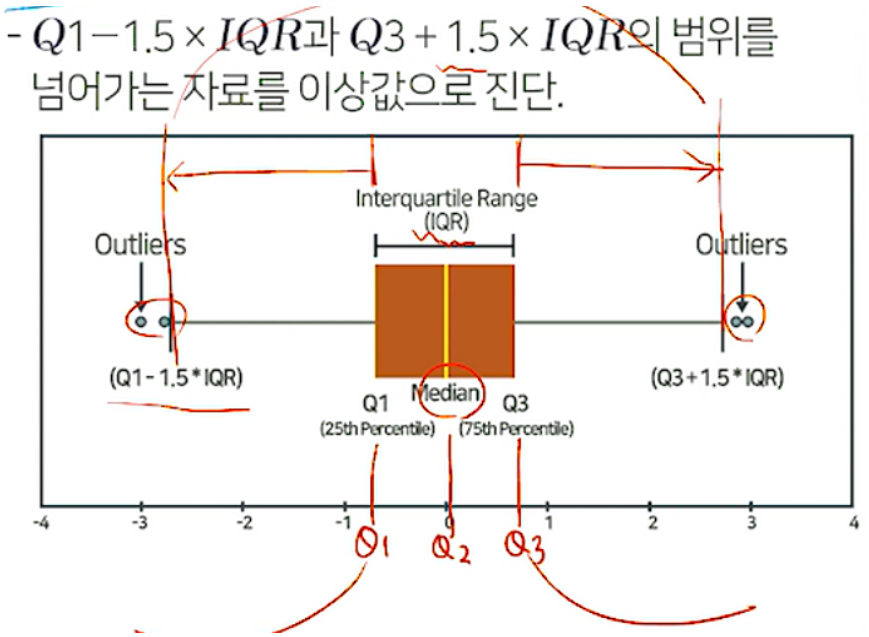
	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

    b. 결측값 처리법
        a. 완전제거법(list-wise deletion)
            - 결측값이 하나 이상 포함된 자료를 제거하는 방법
            - 정보의 손실로 분석결과가 왜곡될 수 있음
        b. 평균대체법(mean value imputation)
            a. 결측값을 해당 변수의 나머지 값들의 평균으로 대체하는 방법.
            b. 추정량의 표준오차가 과소추정되는 문제가 있음.

                (나머지 값들의 평균보다 크거나 작을 때 오차가 보다 작게 측정됨을 주의)
        c. 핫덱대처법(hot deck imputation)
            a. 동일한 데이터 내에서 결측값이 발생한 관찰지와 유사한 특성을 가진 다른 관찰지의 정보를 이용하여 대체하는 방법.
        d. 그 밖의 결측값 처리법
            a. Regression imputation, kNN imputation 등
    c. 이상값의 이해
        a. 이상값은 다른 데이터와 동떨어진 것을 말함
        b. 다른 자료값들에 비해 멀리 떨어져 있지만 의미가 있는 값일수도 있고, 단순히 입력 오류로 발생한 값일 수도 있음.

d. 이상값의 탐지

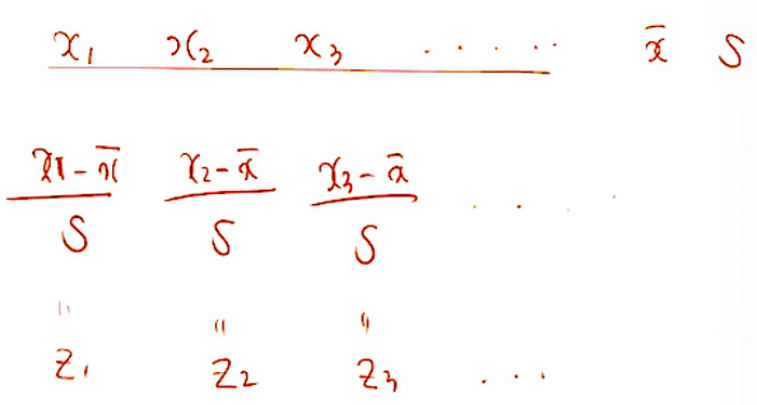
a. 상자그림 (1.5 : tukey의 계수)



b. 표준화 점수(Z-score)

- 표준화 점수의 절대값이 2, 3 보다 큰 경우를 이상값으로 진단.
- z1, z2, z3 를 표준화 점수라고 한다.
  표준화 점수의 특징 : 항상 평균값이 0이고 표준편차가 1이다.

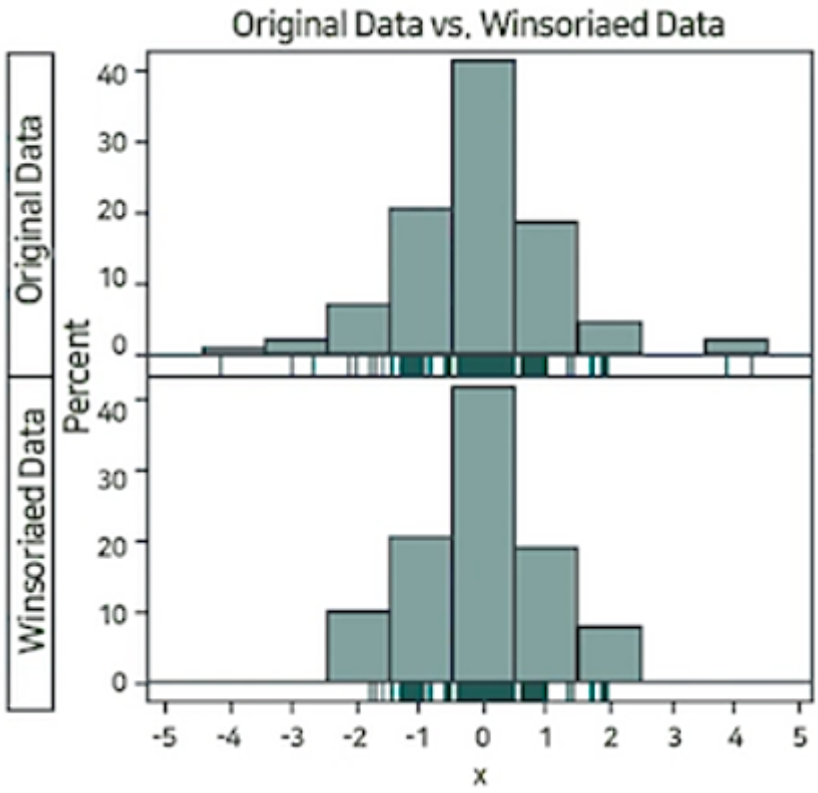
  따라서 표준화 점수가 1보다 크면 이상값이라고 할 수 있음
- x_var : 평균, S : 표준편차



e. 이상값 처리 방법

- 이상값 제외(trimming)

    - 처리는 간단하지만, 정보 손실이 발생하고 추정량 왜곡이 생길수 있음
- 이상값 대체(winsorization)
    - 이상값을 정상값 중 최대 또는 최소 등으로 대체하는 방식.



- 변수 변환
    - 자료값 전체에 로그변환 제곱근 변환 등을 적용.

f. 연속형 자료의 범주화

- 변수 구간화(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]

- 변수구간화(binning)의 효과
  - 이상치 문제를 완화
  - 결측치 처리 방법이 될 수 있음.
  - 변수간 관계가 단순화 되어 분석시 과적합을 방지할 수 있고, 결과 해석이 용이해짐.
  - 단, 범주화 후에는 정확도가 떨어질 수 있기 때문에 범주를 적절하게 나누는 것이 중요