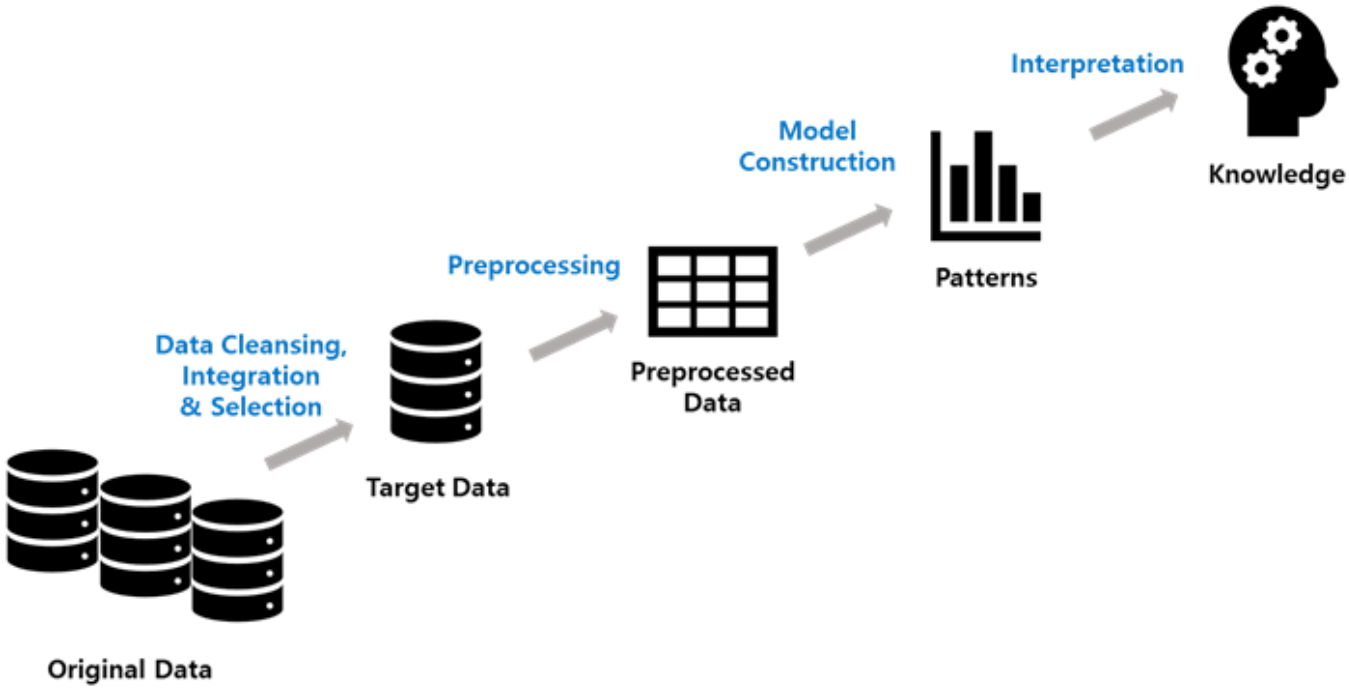
데이터 전처리

데이터 분석을 위한 데이터 정제 및 가공



왜 데이터 전처리를 하는가?

데이터의 품질을 보장하여 데이터의 신뢰도를 높이고, 분석결과의 질을 보장하기 위해서

데이터 품질을 낮추는 주요 원인

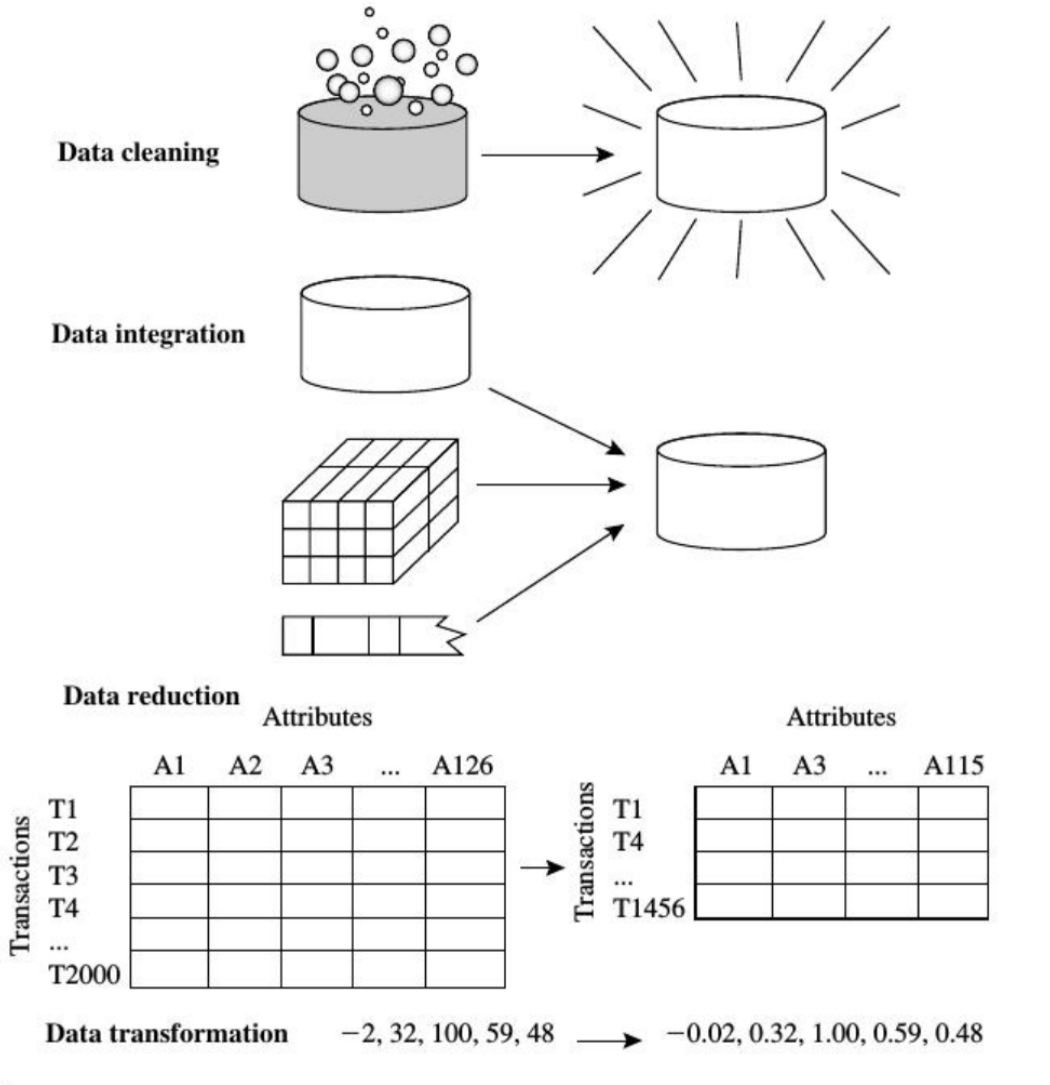
- 불완전(incomplete) : 데이터가 비어 있는 경우
- 잡음(noise) : 데이터에 오류가 포함된 경우
- 모순(inconsistency) : 데이터간 정합성, 일관성이 결여된 경우

데이터 품질을 비교하는 요소들

- 정밀성(accuarcy) : 오류나 예상치에서 벗어나는 값이 없음
- 완전성(completeness) : 속성의 값이나 관심있는 어떤 속성이 모두 존재함
- 일관성(consistency) : 값에 모순점이 없음
- 적시성(timeliness) : 모든 값이 필요한 시점에서 사용가능한 상태
- 신빙성(believability) : 자료에 대한 신뢰도
- 해석성(interpretability) : 데이터를 이해하기가 얼마나 쉬운가

- 출처 : R로 배우는 데이터과학, 한빛출판사

데이터 전처리의 주요 작업



출처:
https://www.potatogim.net/wiki/PGWiki:%EC%B1%85/%EB%8D%B0%EC%9D%B4%ED%84%B0_%EB%A7%8
(https://www.potatogim.net/wiki/PGWiki:%EC%B1%85/%EB%8D%B0%EC%9D%B4%ED%84%B0_%EB%A7%8

1. 데이터 정제(cleansing)

결측치(missing value)를 채우거나, 잡음값(noisy data) 완화, 이상치(outlier)를 발견하여 이를 제거하거나 적정하게 처리하여 데이터의 신뢰도를 높이는 작업

1) 결측치(missing value)

: 데이터 수집, 저장 과정에서 값을 얻지 못하여 발생(기록누락, 무응답, 수집오류)

- **결측치 처리**
  - 완전 제거
  - 평균값 대체
  - hot-deck 대체
    - 동일한 조사에서 다른 관측값으로부터 얻은 자료를 이용하여 대체
    - 관측값 중 결측치와 비슷한 특성을 가진 것을 무작위 추출하여 대체
  - 회귀분석

- pandas : pd.isna(df), df.dropna(조건), df.fillna(대체값)

2) 잡음값(noisy value)

: 측정된 값에서 임의의 오류나 변화가 발생한 것

- 잡음 요소
  - 중복된 관측값
  - 유사 속성이 데이터 내 포함 : 예. 월소득/년소득, 생년월일/나이
  - white noise
  - 오류, 오차값

잡음 사례

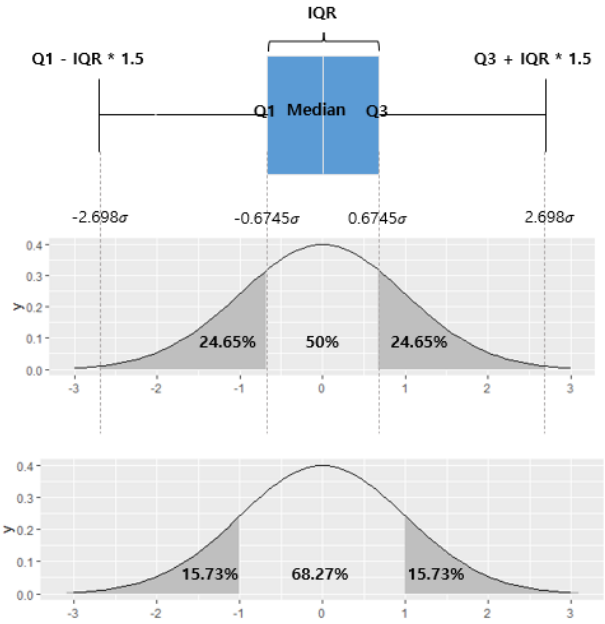
- **잡음 처리**
  - 구간화(binning) : 데이터 값 구간화
    - 평균값 평활화
    - 중앙값 평활화
    - 경계값 평활화
  - 연속형 변수 범주화
    - 이상치 문제완화, 결측치 처리 방법이 될 수 있음
    - 변수간의 관계가 단순화되어 분석 시 과적합 방지, 결과 해석이 용이해짐
  - 회귀 : 회귀함수에 의해 데이터 평활화
  - 군집화 : 유사한 값끼리 그룹화, 군집의 센터값 사용

참고: https://m.blog.naver.com/taewwon/221581853436 (https://m.blog.naver.com/taewwon/221581853436)

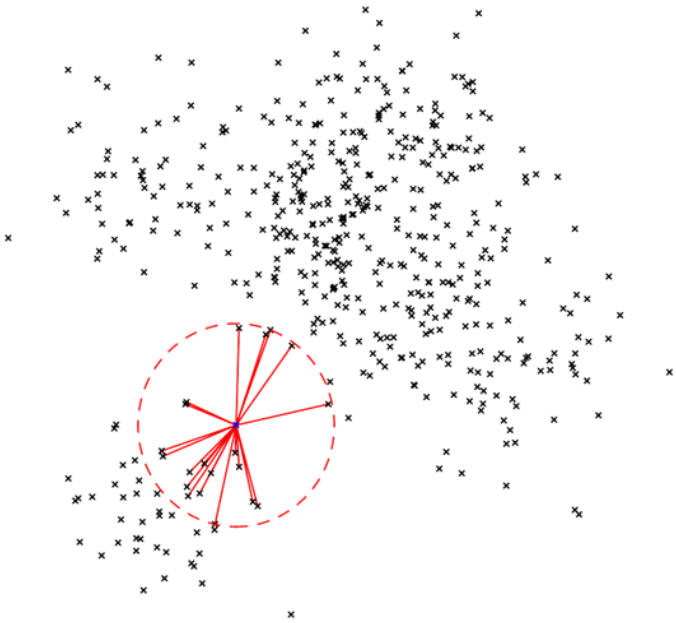
3) 이상치(outlier)

: 정상적이라고 생각되는 데이터이나, 기존 데이터와 매우 동떨어진 데이터

- **이상치 탐지**
  - 시각화 기법 : 히스토그램, 박스플롯, 산점도
  - 수치적 기법
    - IQR 기반
    - 거리기반 탐지 : 유클리드 거리, 마할라노비스 거리 등
    - 분석기법 : KNN(K Nearest Neighbor), 군집의 거리(clustering)



IQR



KNN

출처: https://jayhey.github.io/novelty%20detection/2018/01/29/Novelty_detection_KNN/ (https://jayhey.github.io/novelty%20detection/2018/01/29/Novelty_detection_KNN/)

- **이상치 처리**
  - 제거(trimming)
  - 대체(winsorization) : 이상치를 정상값 중 최소값 또는 최대값을 대체
  - 변수 변환 : log, 제곱, 지수, 제곱근 변환, 표준화
- pandas : quantile(), np.where()

2. 데이터 통합(integration)

: 다수의 근원지로부터 얻은 데이터를 합쳐서 하나의 데이터로 만듬

데이터 결합

- 조인(join)
- 결합(bind)
- 병합(merge)

3. 데이터 축소(reduction)

데이터 크기를 줄이지만 분석결과는 축소이전 데이터의 분석결과와 유사한 수준이 되도록 데이터를 구성함

1) Feature selection

: 데이터 행과 열의 핸들링을 통해 좋은 변수 조합 선택, 불필요한 변수 제거

- 데이터 선택(selection)

- 필터링(filtering)

2) Feature Extraction

: 알고리즘을 통한 축소로 주어진 변수들을 결합하여 유용한 변수로 생성

- 주성분 분석(PCA)
- 특이값 분해(SVD)
- 요인분석(FA)
- LDA(Linear Discriminant Analysis)

3) 샘플링(sampling)을 통한 축소

4. 데이터 변환(transformation)

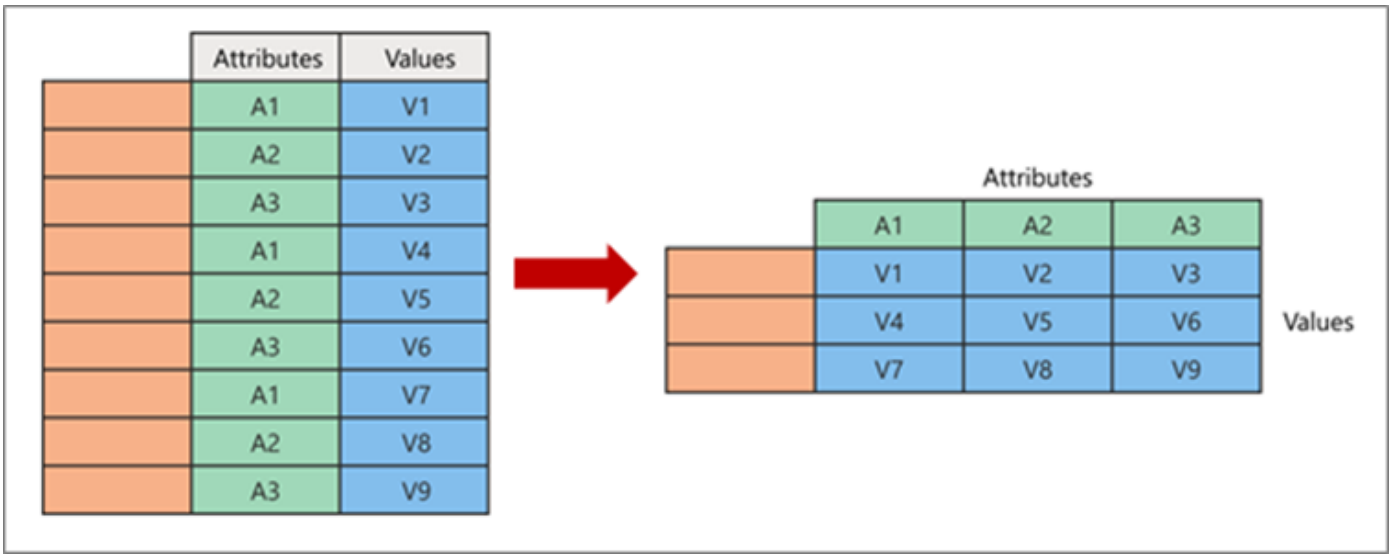
분석 알고리즘(모델)의 효율성을 극대화시키기 위해 임의의 변형을 줌

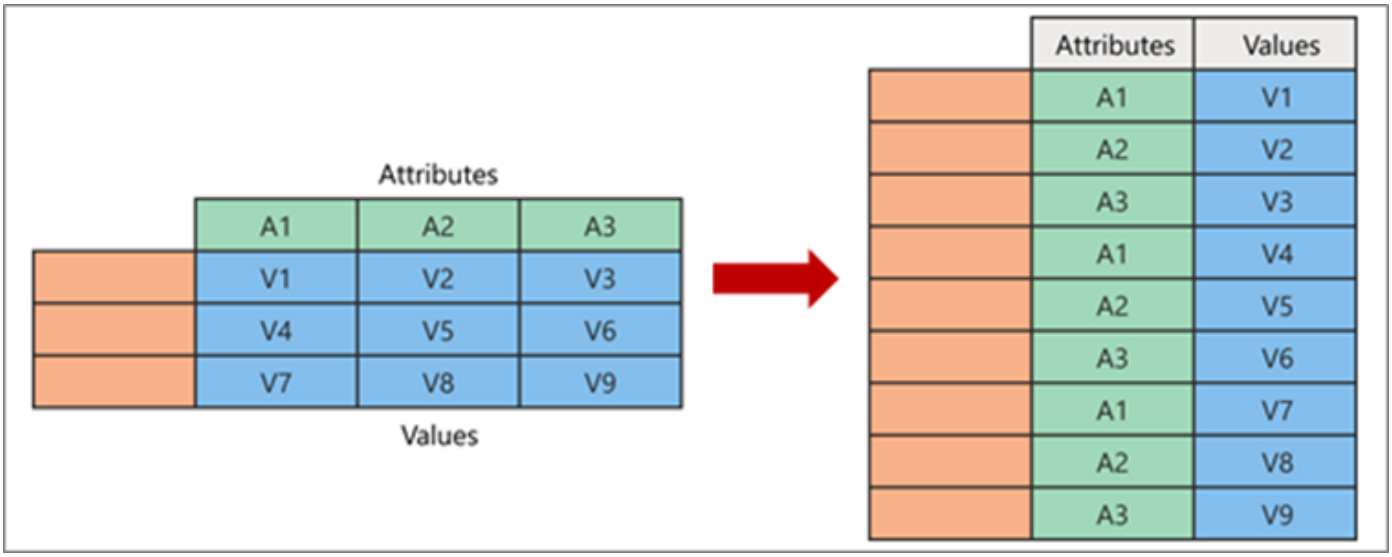
데이터 변환 목적은

- 분포의 대칭화
- 산포를 비슷하게
- 변수 간의 관계 단순화

1) 데이터 구조 변환: 재구조화(reshape)

- 데이터 전치(transpose)
- 피봇팅 : pivoting, unpivoting
- 스태킹 : stacking, unstacking





https://support.microsoft.com/ko-kr/office/%ED%94%BC%EB%B2%97-%EC%97%B4-%ED%8C%8C%EC%9B%8C-%EC%BF%BC%EB%A6%AC-abc9c8da-3be9-44c4-886e-0be331ab387a (https://support.microsoft.com/ko-kr/office/%ED%94%BC%EB%B2%97-%EC%97%B4-%ED%8C%8C%EC%9B%8C-%EC%BF%BC%EB%A6%AC-abc9c8da-3be9-44c4-886e-0be331ab387a)

https://support.microsoft.com/ko-kr/office/%EC%97%B4%EC%9D%84-%EC%96%B8pivot-power-query-0f7bad4b-9ea1-49c1-9d95-f588221c7098 (https://support.microsoft.com/ko-kr/office/%EC%97%B4%EC%9D%84-%EC%96%B8pivot-power-query-0f7bad4b-9ea1-49c1-9d95-f588221c7098)

2) 척도 변환(scaling)

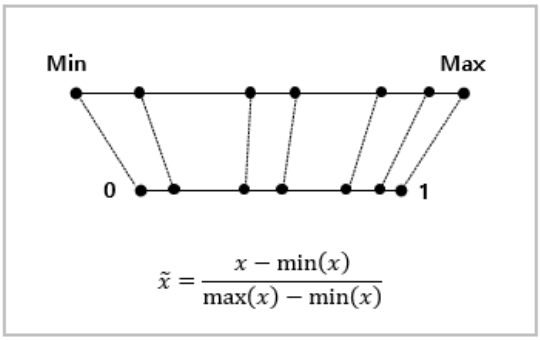
- 정규화(normalization)이라고도 부름
- 단위 차이, 극단값 등을 비교가 어렵거나 왜곡이 발생할 때 표준화하여 비교 가능하게 함
- 스케일이 다른 여러 변수에 대하여 스케일을 맞춰 도는 데이터를 동일한 중요도로 비교되도록 함

스케일링 방식

- Min-Max Scaling : $\frac{X-X_{min}}{X_{max}-X_{min}}$
- Standard Scaling(Z-score) : $\frac{X-\bar{X}}{std(X)}$
- Max Absolute Scaling : $\frac{X}{|X_{max}|}$
- Robust Scaling : $\frac{X-X_{2/4}}{X_{3/4}-X_{1/4}}$

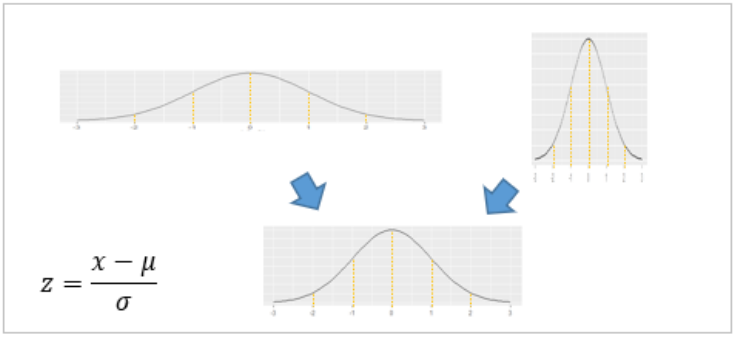
Min-Max Scaling

0~1 사이 값으로 변환



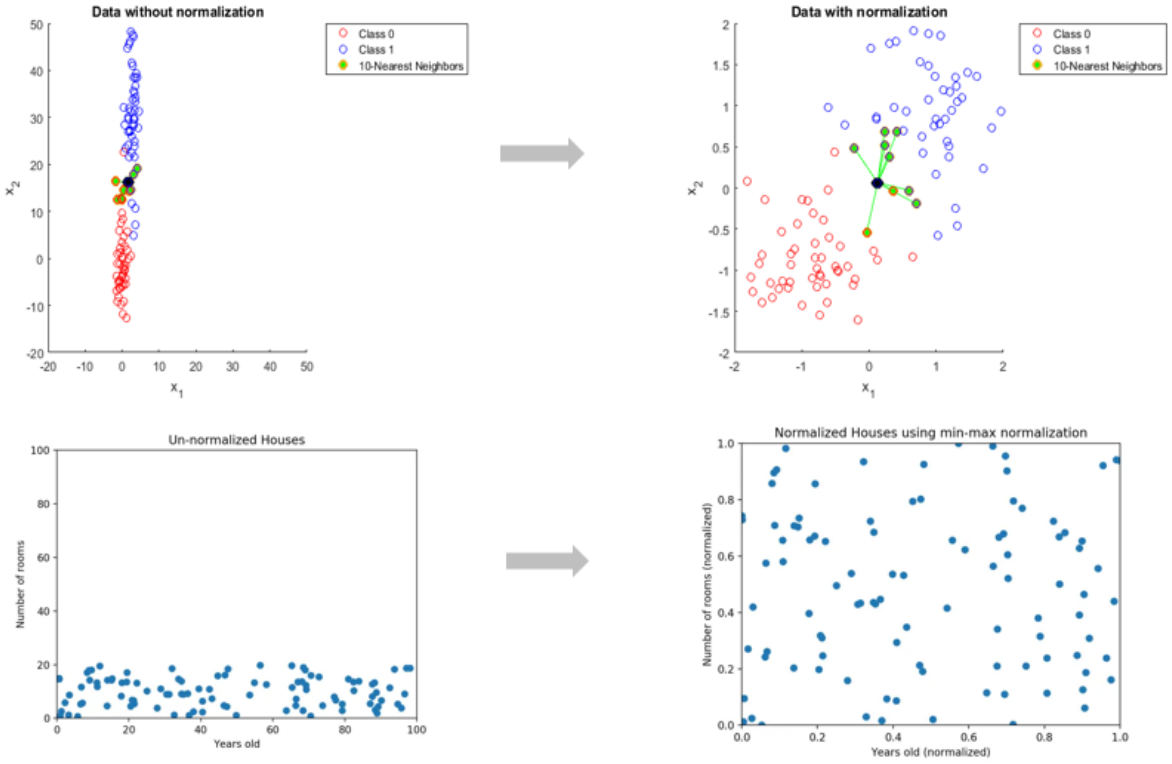
Standard Z Scaling

표준화 값으로 변환



Before Normalization

After Normalization



출처: https://stats.stackexchange.com/questions/287425/why-do-you-need-to-scale-data-in-knn (https://stats.stackexchange.com/questions/287425/why-do-you-need-to-scale-data-in-knn)

[http://hleecaster.com/ml-normalization-concept![image.png](attachment:image.png)](http://hleecaster.com/ml-normalization-concept!%5Bimage.png%5D(attachment:image.png)) (http://hleecaster.com/ml-normalization-concept!%5Bimage.png%5D(attachment:image.png))

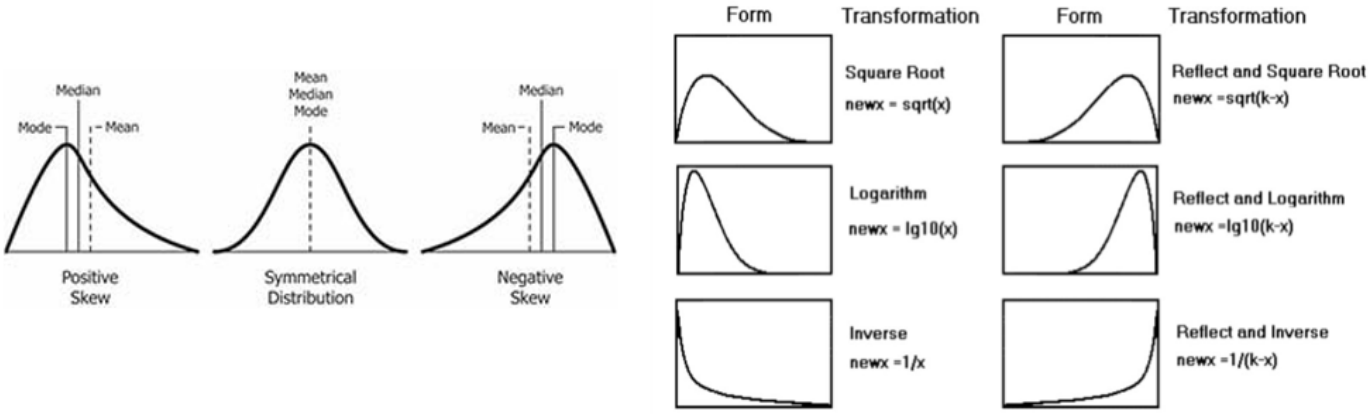
3) 데이터 분포 변환

: 분포 대칭화(정규분포에 가깝게 변환)

- **분포 비대칭 탐지**
  - 시각화 기법 사용 : 히스토그램, 박스플롯(boxplot)
- **변환 방식**
  - 제곱근 변환
  - 제곱변환
  - 지수변환

- 로그변환
- 박스콕스변환 : 정규성과 등분산성에 대한 문제 해결에 유용

- **오른쪽으로 꼬리가 긴 분포(positively skewed)의 경우**
  - sqrt(x) → log10(x) → 1 / x
- **오른쪽으로 꼬리가 긴 분포(negatively skewed)의 경우**
  - sqrt(max(x+1) – x) → log10(max(x+1) –x) → 1 / (max(x+1) - x)



출처: https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/ (https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/)

4) 변수 변환

- 연속형 변수의 범주화
- 범주형 변수의 가변수(dummy)화

5) 데이터 정렬

6) 파생변수, 요약변수

- **파생변수(derived variable)**
  - 이미 수집된 변수를 활용해 새로운 변수 생성하는 경우
  - 분석자가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수
  - 주관적일 수 있으며 논리적 타당성을 갖추어 개발해야 함
  - 예. 주구매 매장, 구매상품다양성, 가격선호대, 라이프스타일
- **요약변수(suumary variable)**
  - 데이터를 분석 요구에 맞게 종합한 변수
  - 데이터의 수준을 달리하여 종합하는 경우가 많음
  - 예. 총 구매금액, 매장별 방문횟수, 매장이용횟수, 구매상품목록
- **방식**
  - 데이터 요약(summary) : 기술통계, 빈도 등
  - 그룹별 집계(aggregate)

7) 데이터 행/열 핸들링을 통한 데이터 변환

- 타입 변환, 인덱스 변경
- 행/열이름 변경 등
- 데이터 행/열삭제

데이터전처리와 관련된 pandas의 API 함수들

전처리 구분	항목	내용	Pandas API
데이터 정제 (cleansing)	결측치 처리	• 완전제거 • 평균대체 • Hot-deck대체 등	• isna(), dropna() • fillna()
	잡음 제거	• 구간화(binning) : 평균/중앙값/경계값 평활 • 연속형변수 범주화 • 회귀, 군집화	• qcut() • cut()
	이상치 처리	• 이상치 탐지: 상자그림, 거리기반(IQR, KNN) • 이상치 처리 : 제거, 대체, 변수변환	
데이터 통합 (integration)	데이터 결합	• 조인(join), 결합(bind), 병합(merge)	• Join(), concat(), merge()
데이터 축소 (reduction)	Feature Selection	• 데이터 선택(selection) : 인덱싱과 슬라이싱 • 필터링(filtering) : 불린 인덱싱	• [ ], loc[], iloc[]
	Feature Extraction	• 주성분 분석(PCA) • 요인분석(Factor Analysis) • 특이값 분해(SVD) • LDA(Linear Discriminant Analysis)	
	표본추출 (sampling)	• 학습/검증/테스트 데이터 분할 • 통계적 표본추출법 : 단순임의, 계통, 층화, 군집추출	
데이터 변환 (transformation)	데이터 구조 변환(reshape)	• 데이터 전치(transpose) • 피봇팅과 언피봇팅 • 스태킹과 언스태킹	• T • pivot(), pivot_table(), melt() • stack(), unstack()
	척도 변환 (scaling)	• Min-max scaling • Standard scaling(Z-score) • Max Absolute scaling • Robust scaling	
	데이터 분포 변환(대칭화)	• 비대칭 탐지 : 시각화기법(histogram, boxplot) • 제곱근,로그, 제곱, 지수, Box-Cox 변환	• matplotlib , seaborn, pandas
	변수 변환	• 연속형 변수 변수 : 범주화 • 범주형 변수 변환 : 가변수(dummy)화 • 시계열 데이터 변환 : lead, lag	• cut(), qcut() • get_dummies()
	정렬	• 정렬	• sort_index(), sort_value()
	파생변수 /요약변수	• 데이터 요약(summary) • 그룹별 집계(aggregate) • 기술통계, 빈도 등 수학, 문자열, 날짜함수, 집계함수, 쿼리 등	• describe() , info(), shape() • groupby(), agg(), aggregate(), apply(), transform(), filter() • sum(), mean(), median(), min(), max(), std(), var(), quantile(), first(), last(), count(), value_counts()
	데이터핸들링을 통한 변환	• 데이터 열/행삭제 • 데이터 타입 변환 • 인덱스 변경 • 열/행 이름 변경	• drop() • astype() • set_index(), reset_index() • rename()