데이터 모순점 발견

- 잘못 설계된 데이터 입력 폼 존재
- 사람의 실수
- 응답자의 의도적 오류
- 만료된 데이터
- 데이터 표현 모순
- 일치하지 않는 코드
- 계측장치의 시스템 오류
- 의도적으로 부적절한 데이터의 사용
- 통합과정에서 잘못된 데이터의 통합

데이터 전처리

구분	수행 내용
데이터 유형 변환	데이터 유형을 변환하거나 데이터 분석에 용이한 형태로 변환
데이터 여과 (Filtering)	오류발견, 보정, 삭제 및 중복성 확인 등 데이터 품질 향상
데이터 정제 (Cleansing)	결측치 변환, 이상치 제거, 노이즈 데이터 교정 비정형 데이터를 수집할 때 반드시 수행

데이터 전처리 : 데이터 정제

결측값 (Missing values)

- 값이 존재하지 않고 비어있는 상태
- NA (Not Available) 또는 NULL 값
  - NA: 결측값
  - NULL: 값이 없다
- 분석 대상의 속성 값이 상당 부분 비게되면 대상 데이터가 충분하지 않으므로 분석을 제대로 수행하기 어렵다.
- 결측값 종류
  - MCAR(Missing Completely At Random)
    - 결측값이 데이터에 독립적, 무작위로 발생 이때는 편향이 없어 문제가 되지 않음
  - MAR(Missing At Random) :
    - 결측값이 다른 변수에 따라 조건부 무작위 발생. 결측값이 변수에 대해 설명가능 하기 때문에 데이터 분석에서 편향 발생 가능
  - MNAR(Missing Not At Random)
    - MCAR 또는 MAR이 아닌 데이터, 무시할 수 없는 무응답 데이터(누락 이유 존재) 결측값이 아니라 추가 조사가 필요

결측값 처리 방법

- 결측값 데이터 제거
  - 데이터가 충분하면 고려가능
  - 데이터 내 결측치 데이터가 많다면 대부분 정보가 제거될 수 있음
  - 실제로는 지양하는 방법
- 수동으로 결측값 입력
  - 결측값 발생한 데이터를 재조사 및 수집
  - 고비용, 소모적
  - 비현실적 방법
- 전역상수(global constant) 로 대체
  - 단순하며 명확함
  - 전역상수 값이 분석 결과 왜곡 가능
  - 보통 0 이나 평균값 등으로 대체
- 결측값 무시
  - 알고리즘 또는 모델이 결측치 값을 무시하고 분석 수행 할 수 있음
  - 한 속성이 없어도 다른 속성을 통해 알고리즘 조정
  - 속성이 적어 하나의 속성이라도 무시하기 힘들면 좋지 않음
- 결측값 추정
  - 일반적으로 사용되며 결측값이 발생한 데이터와 유사한 데이터를 사용하여 결측값 추정
  - 결측값 추정 방법에 따라 다양한 형태 존재

결측값 추정 방법

- 속성의 평균값 사용하여 추정
  - 평균값을 결측값에 대체
  - 분석 결과를 왜곡시킬 위험성 존재
- 같은 클래스에 속하는 속성의 평균값 사용
  - 주어진 데이터와 같은 클래스에 속하는 튜플 들의 속성 평균값 사용
  - 동일 유형에 속하는 데이터의 평균값을 사용하므로 왜곡 가능성 줄임
- 가장 가능성이 높은 값으로 추정
  - 회귀본석, 베이지안 등 머신러닝 통계 기법 활용하여 예측
  - 분석에 의해 가능성 높은 값을 찾아냄
  - 가장 효과적이며 높은 정확도의 예측 가능
  - 결측 값을 채우기 위한 분석 가설을 세우는 등의 복잡성 존재

이상값 발견 기법

- 개별 데이터 관찰: 데이터 값을 눈으로 보며 전체적인 추세와 특이사항 관찰
- **통계값** 활용: 요약 통계 지표(summary statics)
- **시각화**: 확률 밀도 함수, 히스토그램, 점플롯(dot plot), 워드 클라우드, 시계열차트 등
- 머신러닝 기법: 클러스터링(clustering) 등을 통한 이상치 확인
- 통계 기반 탐지 (Statistical-based detection): Distribution-based, depth-based
- 편차 기반 방법 (Deviation-based Method): Sequential exception, OLAP data cube
- 거리 기반 탐지 (Distance-based Detection): Index-based, Nested-loop, Cell-based

데이터 변환 (Data Transformation)

- 데이터 변환은 데이터 분석에 적절한 형태로 데이터를 바꾸는 전처리 작업을 의미합니다.
- 데이터 변환 방법
  - 평활화(Smoothing): 데이터의 잡음 제거, 구간화 회귀, 군집화 등
  - 집계: 그룹화 연산을 적용(일일 판매 데이터 -> 월별 -> 연도별 그룹화)
  - 속성구성: 주어진 속성 집합으로부터 새로운 속성 구성
  - **정규화**(Normalization): 데이터를 정해진 구간 내에 존재하도록 변환
  - **이산화**(Discretization): 수치형 속성을 구간 라벨이나 개념적 라벨로 대체
  - 개념계층(Conceptual Hierachy) : 도로명과 같은 속성을 시 국가와 같은 상위 레벨 개념으로 일반화

데이터 전처리 : 데이터 변환

정규화 (Normalization)

- 정규화는 -1 ~ 1 사이와 같이 정해진 구간 내에 들도록 하는 기법

- 종류
  - 최소-최대 (min-max normalization)
    - 원본 데이터 값들 간의 관계를 보존
    - 원본 데이터의 최소 최대값을 벗어난 값이 새로 들어오면 범위초과 오류 발생
  - Z-score 정규화 (Z-score normalization)
    - 속성 A에 대한 값을 A 평균과 표준편차를 기초로 정규화
    - 실제 최소값이나 최대값이 알려져 있지 않거나, 최소-최대 정규화 수행시 큰 영향을 끼치는 이상치가 데이터에 존재할 경우 유용

이산화 (numeric data discretization)

- 수치형 데이터의 이산화
  - 구간화 (binning)
  - 엔트로피-기반 이산화
  - 카이제곱 결합
  - 군집 분석
  - 직관적 분할에 의한 이산화
- 범주형 데이터의 이산화