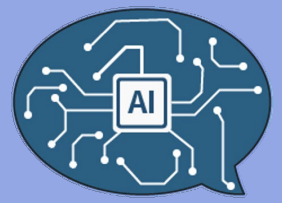
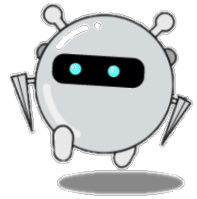


데이터 전처리 이론 2



진행자: 멘토 이창대

OUTTA

이전수업 복습

데이터 형태: 정형, 반정형, 비정형 데이터

데이터 분류: 수치 데이터, 범주형 데이터

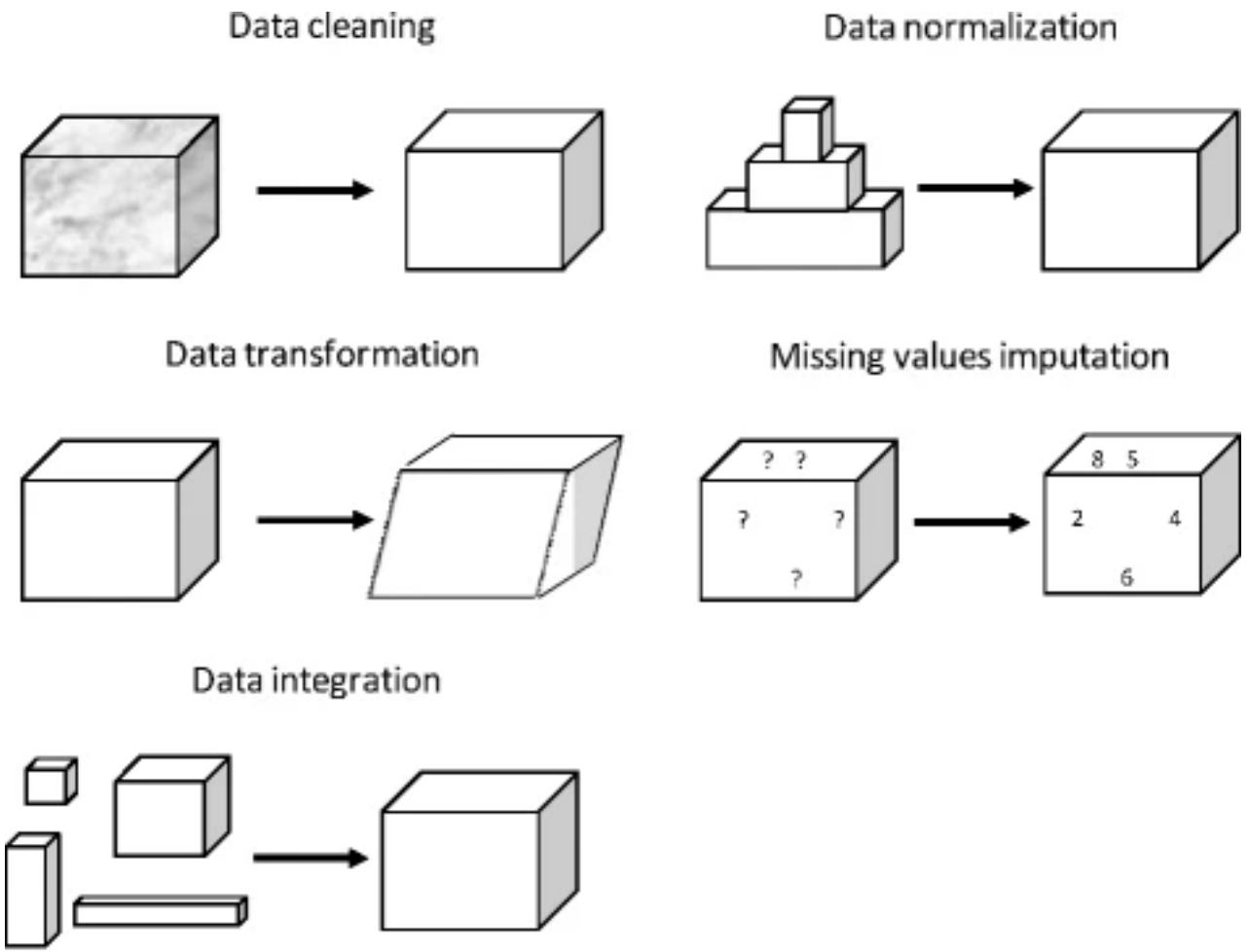
-> 컴퓨터가 데이터를 처리할 수 있도록 특정 형태로 만들어야 한다!

목차

- 데이터 전처리 방법론
  - Data Cleaning
  - Data Transformation
  - Data Integration
  - Data Normalization
  - Missing Values

OUTTA

데이터 전처리 방법론



출처: https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0

OUTTA

Data Cleaning

- 잘못 기입된 데이터
- 파손된 데이터
- 중복 데이터

A	B	C
1	2	3
1	2	3
1	5	9
	5	3
######	15	32
######	13	
15	23	51

OUTTA

Data Transformation

- 분석이 용이하도록 바꾸는 과정
- 기존 데이터 형태와 분석의 목적에 따라 다양하게 바꿀 수 있음.

OUTTA

Data Transformation

Country	Salary
Japan	12000
Korea	20000
China	10000
Japan	31000

Country	Salary
1	12000
2	20000
0	10000
1	31000

China	Japan	Korea	Salary
0	1	0	12000
0	0	1	20000
1	0	0	10000
0	1	0	31000

Label Encoding

- 알파벳 순서로 숫자를 할당해 주는 것
- 순서형 데이터에는 적합할 수 있으나 명목형 데이터에는 적합하지 않음

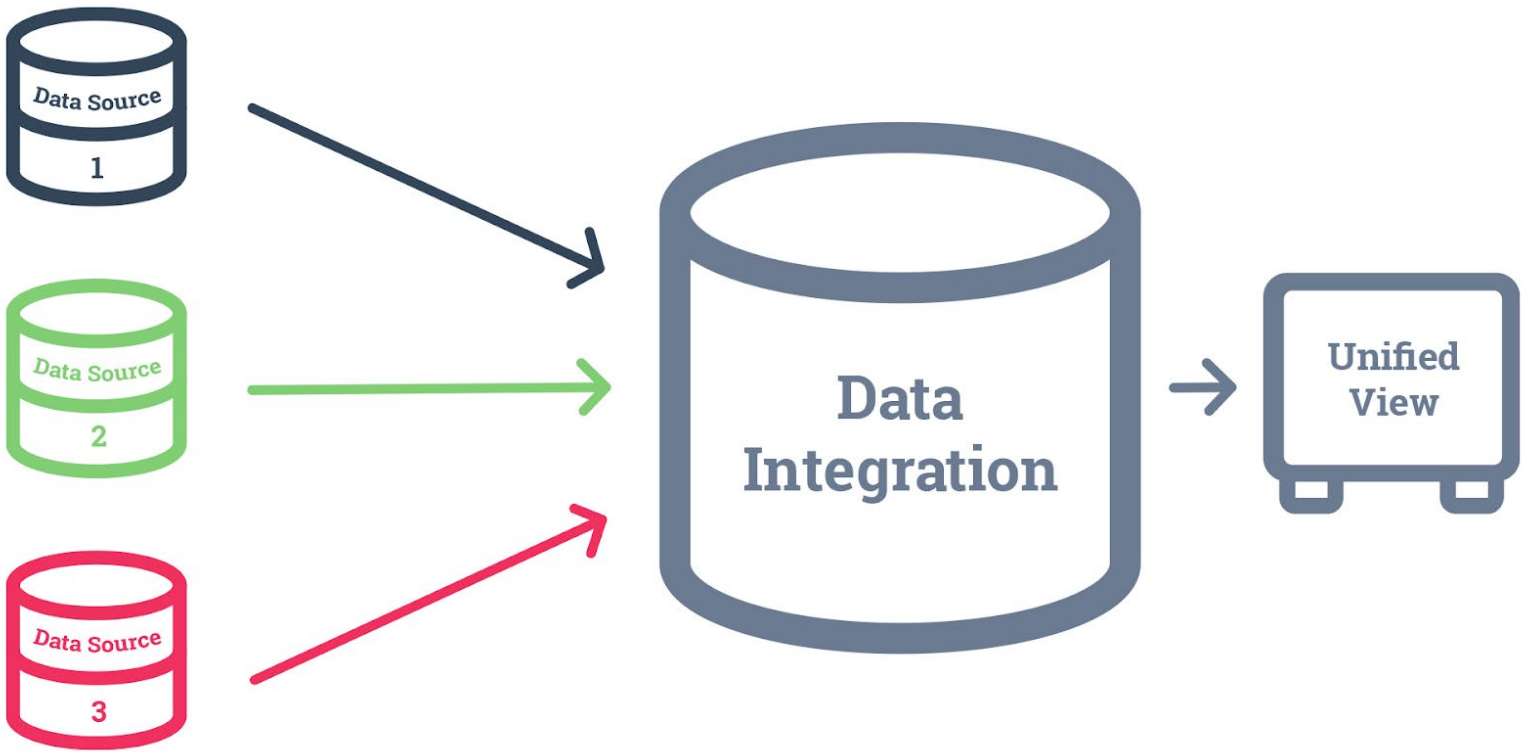
One-Hot Encoding

- 변수값 마다 변수를 추가해 이진값으로 만드는 방법
- 다중공선성 문제가 발생할 수 있음 (Dummy Variable Trap), 메모리 문제

OUTTA

Data Integration

- 다양한 소스에서의 데이터를 처리하기 용이하도록 한곳에 통합된 방식으로 모으는 방법



Data Normalization

- 데이터의 변수 간의 스케일이 심하게 차이가 나는 경우 적용
- 모든 데이터 포인트가 동일한 정도의 스케일로 반영되도록 해줌

Min-Max Normalization

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

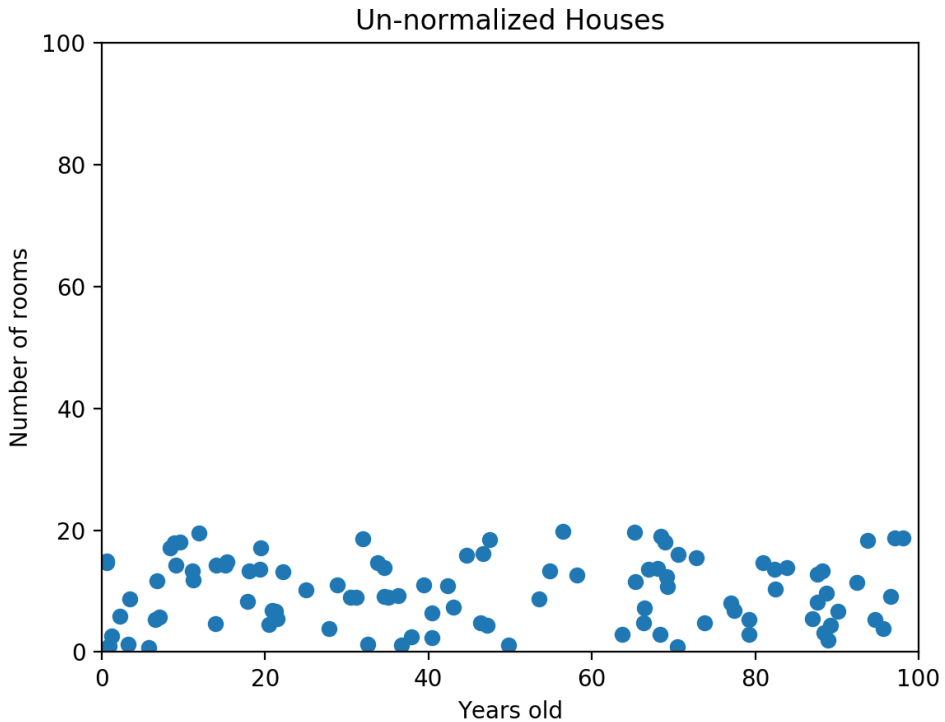
- 장점: 모든 feature의 스케일이 [0,1]로 동일
- 단점: 이상치를 잘 처리하지 못함

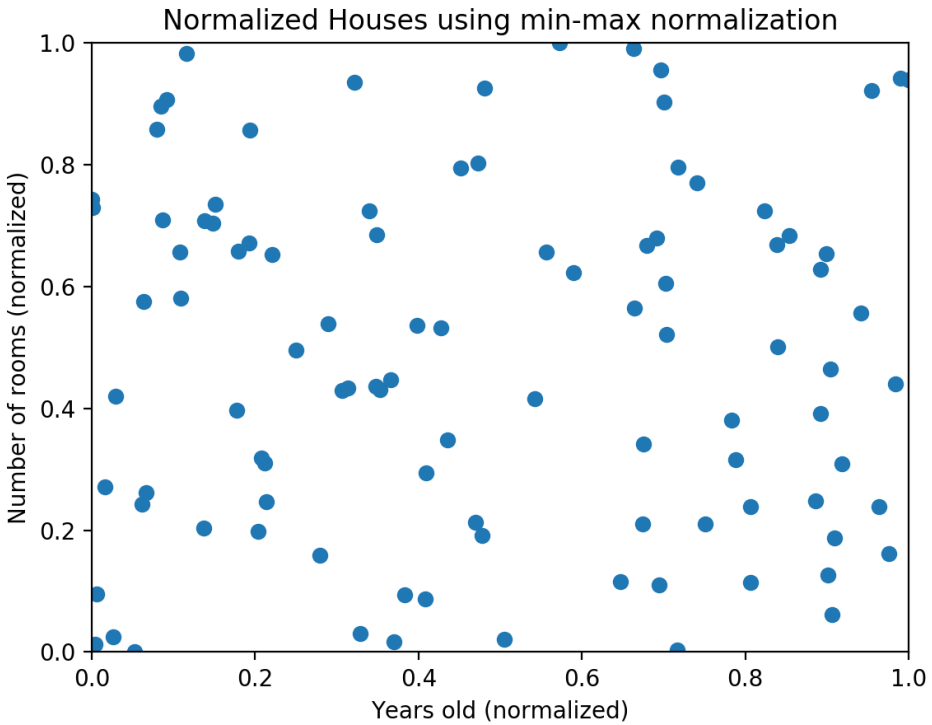
Z-Score Normalization

$$x_{scaled} = \frac{x - mean}{sd}$$

- 장점: 이상치를 잘 처리함
- 단점: 정확히 동일한 척도로 정규화 되지는 않음

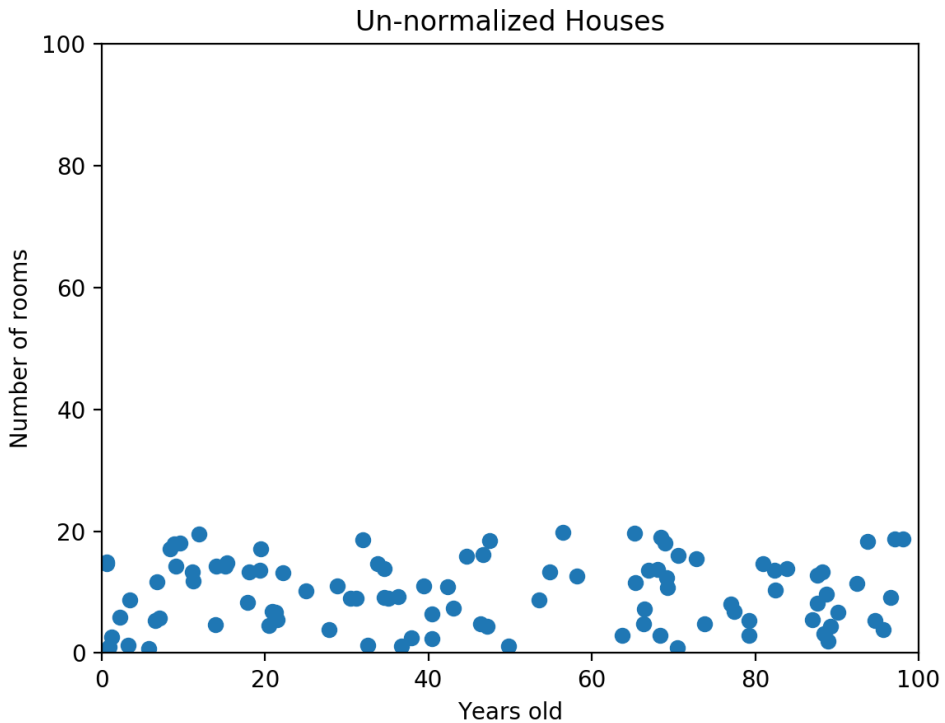
Data Normalization

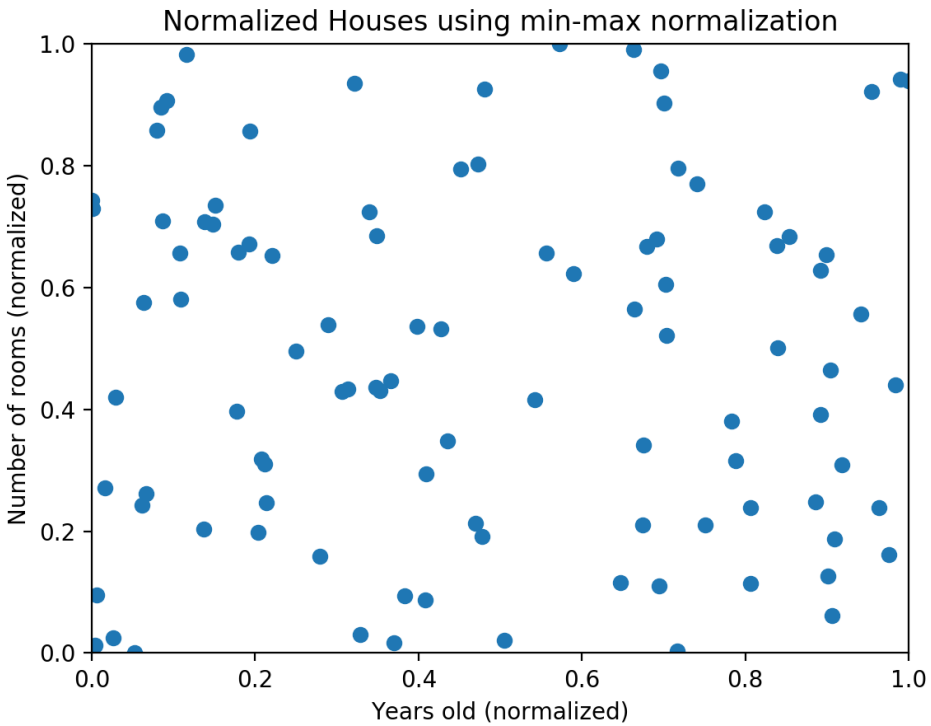




출처: https://hleecaster.com/ml-normalization-concept/

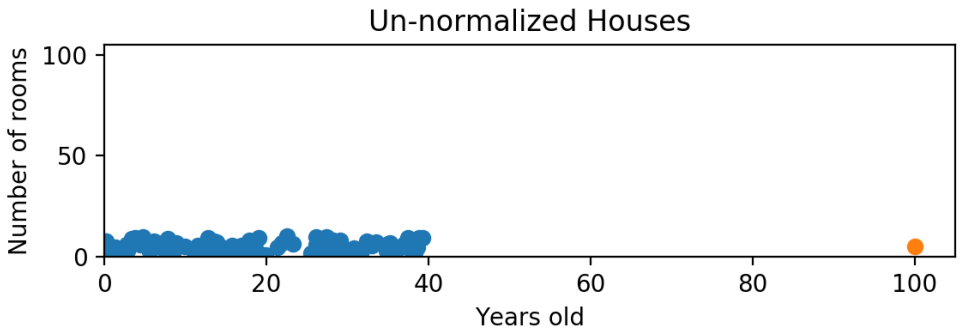
Data Normalization

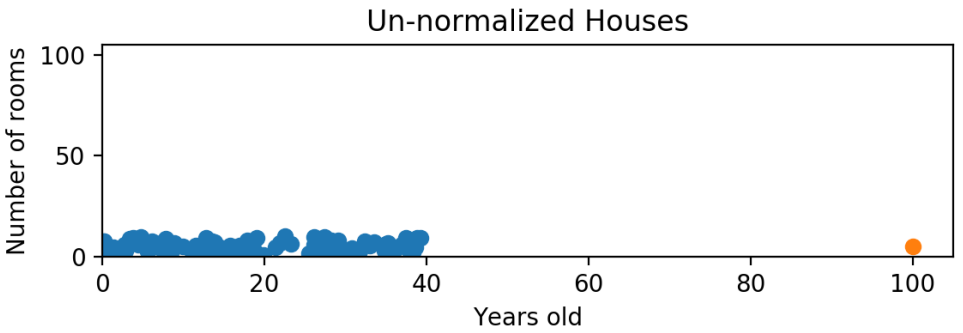


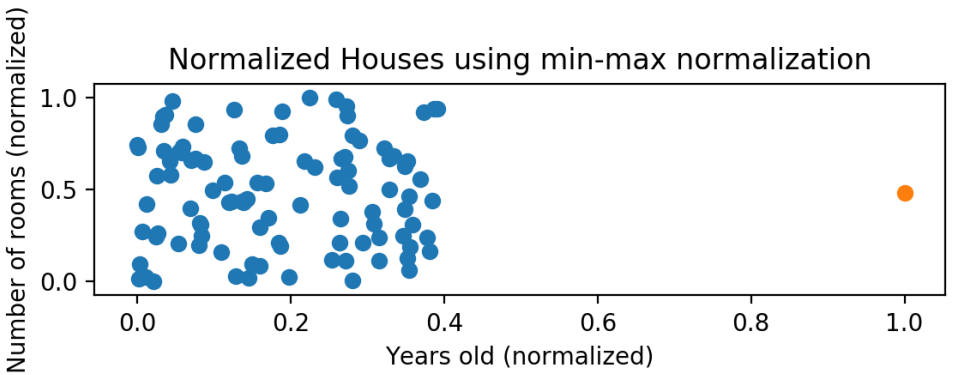


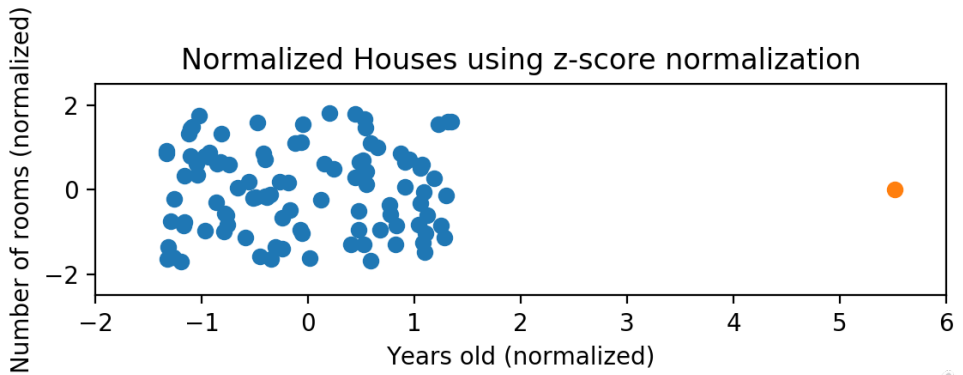
출처: https://hleecaster.com/ml-normalization-concept/

Data Normalization









출처: https://hleecaster.com/ml-normalization-concept/

Missing Values

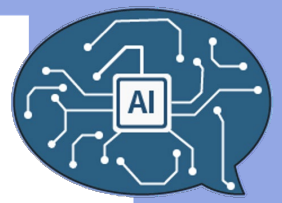
결측치 처리 방법

- 아무것도 하지 않기
- 누락된 데이터를 제거하기
- 누락된 값을 대체하기 (Imputation)

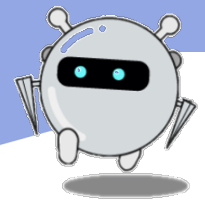
결측치 대체 방법

- 결측되지 않은 다른 값들의 평균이나 중앙값으로 대체 (숫자형)
- 최빈값/0/상수로 대체
- 이외 머신러닝/통계적 방법들

OUTTA



감사합니다



OUTTA