- 03/19 데이터 전처리

데이터 분석
Machine Learning
Deep Learning
→ ① 데이터의 양 ★
→ ② 데이터의 품질

★ → Missing Value (결측치)
★ → Outlier (이상치)
★ → 중복데이터
★ → Data의 타입처리
★ → 정규화 → 1억 ~ 5억 ∨ → 1000000000 ~ 5000000000 ∨
❶ 1점 40점 ∨ → 1~40

?

"Titanic" data set
★ Seaborn module이 제공하는 Titanic data set을 이용

① Missing Value (결측값)

DataFrame 안에 누락된 값이 존재. (데이터 입력시 실수로 누락 / 파일 변환시 처리 문제로 누락)

↓

NaN으로 표현

이런 Missing Value가 많으면 좋지 않아요!

처리해야 해요 ⇒

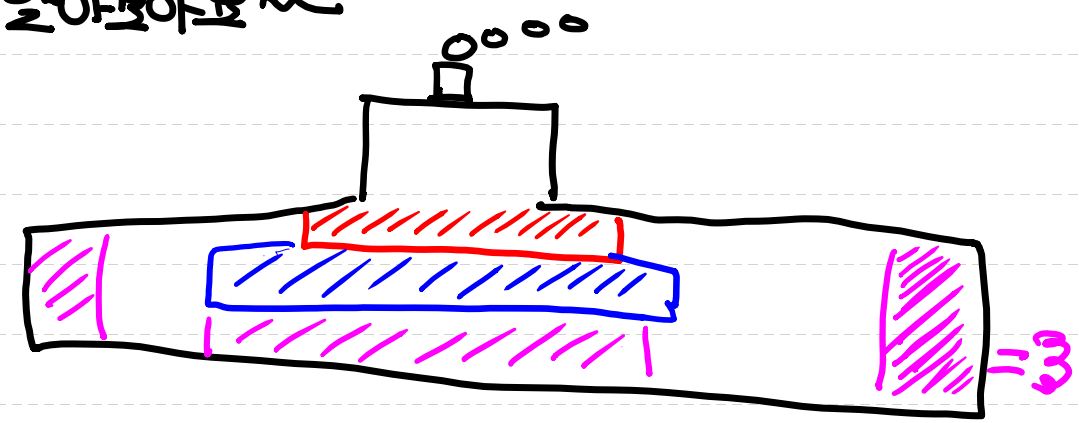
① 삭제 → 만약 데이터가 충분히 많다면 그리고 Missing Value가 전체의 3~4% 이내 일때

② 대체 → 평균, 중위, 최대, 최소, 빈도

* 머신러닝 기법으로 이값을 예측해서 채워요 *

↳ 절대 사용을 안되고 조건이 있어요!

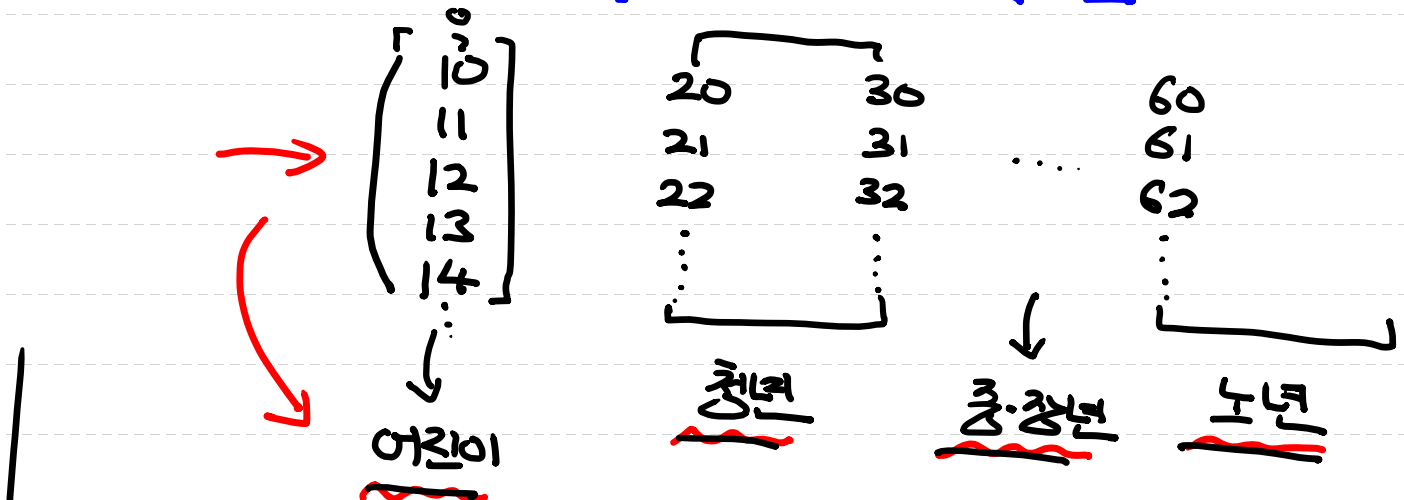
코드로 알아보아요~



② 이상치 처리 ~> "이상치를 찾아서 수정" ~> 나중에 다시 설명!

③ 중복 처리 ~> duplicated( )
drop_duplicates( )

★ ④ 자료형 변환 ~> ★ 만약 숫자가 문자열(Pandas Object)로 저장되어 있으면 숫자형으로 변환해달라고 말아요!



Code로 살펴보아요~

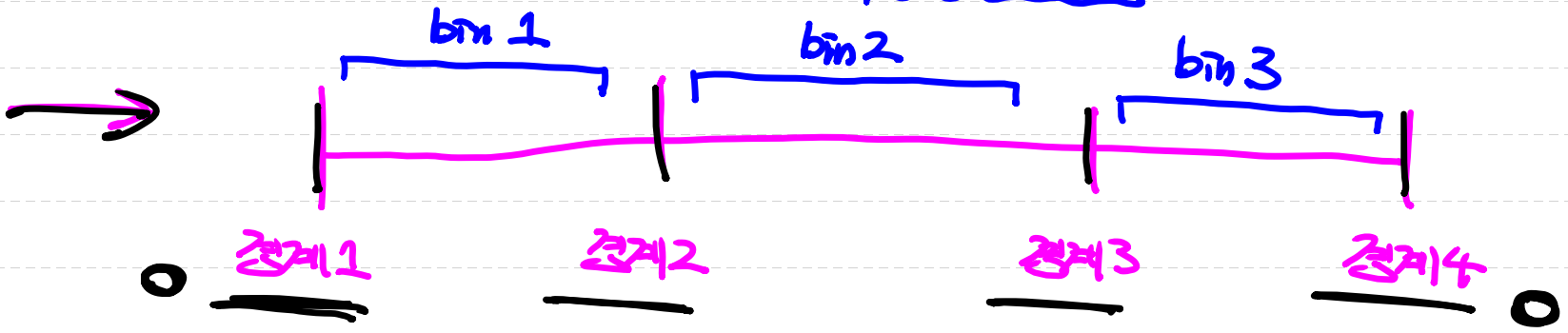
• 범주형 (Category) 데이터 처리

0세 ~ 15세 : 어린이
16세 ~ 25세 : 청소년
26세 ~ 40세 : 청년
⋮

→ 이런형태로 연속적인 데이터를 범주형 데이터로 처리하는게 훨씬 효율적인 경우가 있어요!

* 구간보환 ↘ bin

↳ 연속적인 데이터를 일정한 구간으로 나누기.



↓

고르고 구간보환을 통해 범주형 데이터를 만들어 보아요 ~~

* horsepower를 구간보환 → 저출력 / 보통출력 / 고출력 → 범주형 데이터로 보관

Category를 나타내는 범주형 data는 Machine Learning 알고리즘에서

바로 사용하기가 힘들어요!! → 컴퓨터가 인식할 수 있는 형태로 제공되어야 해요!

이 문제를 해결하기 위해 dummy variable (더미 변수)

이과 1로 표현하고 해당 특성이 있는지 없는지를 표현

* "

→ One-hot-encoding"

⑤ Normalization (정규화)

DataFrame의 각 열이 가지고 있는 숫자 데이터의 상대적인 차이 패턴이 머신러닝결과가 달라질수 있어요!!

→ 숫자 데이터의 상대적인 크기 차이를 제거해야해요!!

★★★★ <~

① 표준화 (Standardization) → "정확히 z-score" (추후에 배워요)

② Min-Max Scaling → 알아볼아요 ~

취약점이 있을까요??

이상치가 존재하면 이상해져요

★

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

$$\frac{2-1}{5-1} = \frac{1}{4}$$

	A	B	C
0	1000	2	20
1	3000	1	50
2	5000	3	70
3	2000	5	1000000

0	0.25
0.5	0
1	0.5
0.25	1