정규세션 1주차

EDA & 기술통계

데이터가 스스로 말하게 하는 법

파트 1

EDA 개념과 목적

01

탐색적 데이터 분석 · 개념 · 목적 · 파이프라인

평균 요금

32.2

VS

중앙값 요금

14.5

같은 891명의 타이타닉 데이터입니다.

두 숫자가 이렇게 다른 이유는 무엇일까요?

EDA의 정의: 모델링 전 데이터의 목소리 듣기

모델링 이전에 데이터의 구조와 특징을 탐색하는 과정

John W. Tukey (1977) — 가설을 세우기 이전에, 데이터가 무엇을 말하는지 들어보는 과정

🔍 데이터 모양 파악

통계기법 (평균·분산·표준편차)
이상치 발견·처리

🔗 속성·관계 파악

산점도·회귀선·상관분석

📊 시각화

Histogram·Box plot·Heatmap

💡 데이터 이해

다양한 방면에서 관찰하고 이해

분석 파이프라인과 EDA의 본질

⚠️ 일방통행 X 끊임없이 순환하는 정제 과정 O

⚠️ Train/Test Split → EDA·결측 대체·스케일링은 Train에서 배운 규칙을 Test에 적용 (Data Leakage 방지)

⚠️ EDA의 산출물은 그래프가 아니라 **의사결정**입니다

- deck → 삭제
- age → 결측 대체
- fare → log 변환
- alive → 즉시 제거

파트 2

데이터 이해

02

Row · Column · Feature · Target · 변수 유형 4분류 · Schema

데이터 구조의 기본 — Row · Column · Feature · Target

Passenger_Id	survived	pclass	sex	age
1	1	1	female	29
2	**0**	**3**	**male**	**22**
3	1	1	female	35

← 가로 한 줄 = 관측치 1건 (승객 1명)

Feature (X) — 모델 입력 변수

pclass · sex · age · fare · sibsp · parch · embarked … 14개

Target (y) — 예측 대상

survived

0 = 사망 / 1 = 생존

변수 유형 4분류 — 왜 구분해야 하는가

변수 유형에 따라 분석 도구 · 시각화 방법 · 인코딩 방식 이 모두 달라지기 때문입니다.

수치형 (Numerical)

연산이 의미 있는 숫자
크기를 비교할 수 있음

예: age, fare, sibsp, parch

도구: 평균·표준편차·히스토그램·Box plot

명목형 (Nominal)

순서 없는 범주
크기 비교 불가

예: sex, embarked, class

도구: 빈도·비율·막대그래프·파이차트

순서형 (Ordinal)

순서가 있는 범주
간격은 일정하지 않음

예: pclass (1>2>3등석)

도구: 단조 경향 확인·빈도

날짜형 (Datetime)

시간 축을 가진 값
타이타닉엔 없음

예: 날짜·시간

도구: 연/월/요일/경과시간으로 분해

dtype의 함정 ⭐ — 저장 방식 ≠ 변수의 의미

```
df['pclass'].dtype    →    int64
df['pclass'].mean()   →    2.309
```

계산은 됩니다.
But '평균 좌석등급 2.309등석'이 무슨 의미인가요?

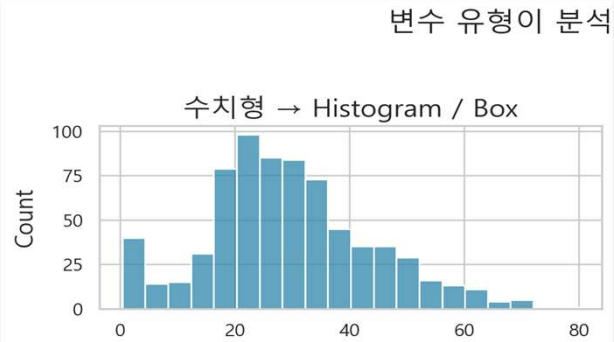
1등석 + 2등석 = 3등석?

dtype은 저장 방식일 뿐 — 변수의 의미(수치형 / 명목형 / 순서형)는 사람이 직접 판단해야 합니다.

수치형 변수의 분석 도구

Histogram
히스토그램

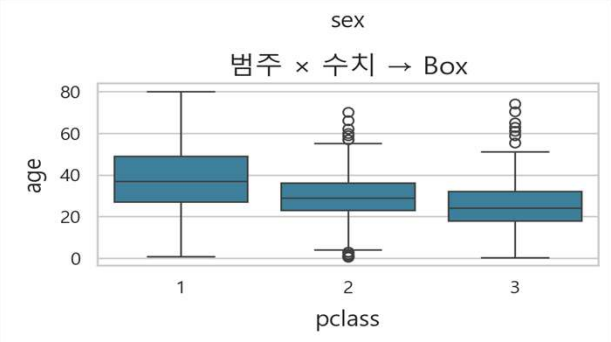
값의 분포 형태 확인
왜도·이상치·구간 확인



언제: 분포 형태를 처음 볼 때
→ 3파트 기술통계의 시작점

Box plot

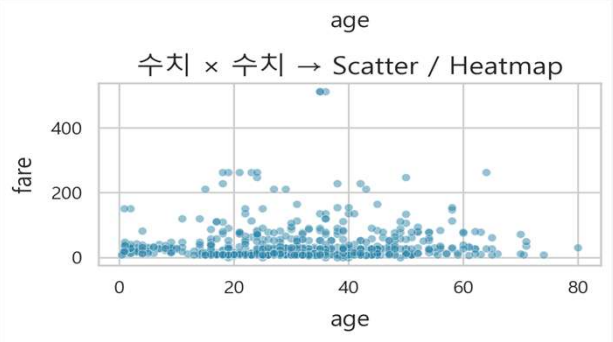
중앙값·IQR·이상치를 한 번에
Q1·Q3·Whisker 구조



언제: 이상치를 빠르게 탐지할 때
→ 4파트 이상치 분석의 핵심

Scatter plot
산점도

두 수치형 변수의 관계
양의·음의·무상관 파악

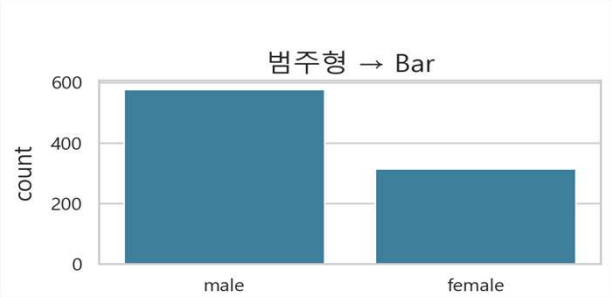


언제: 변수 간 관계를 볼 때

범주형 변수의 분석 도구

Bar chart
막대그래프

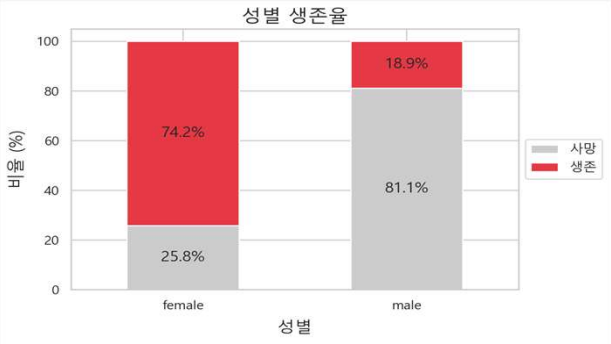
범주별 빈도·비율 비교
가장 기본적인 범주형 시각화



sex별 빈도 — male 577, female 314
가장 먼저 그려보는 그래프

Cross Tabulation
교차표

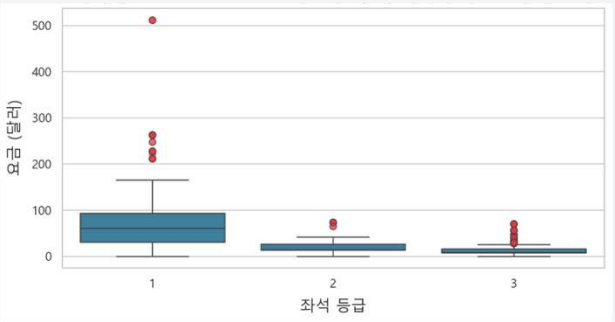
두 범주형 변수의
동시 빈도·비율 파악



sex × survived 교차
여성 74.2% vs 남성 18.9%

Box plot

범주별 수치형 분포를
나란히 비교



등급별 fare 분포
그룹 간 차이를 한 번에 파악

데이터 열어보기 — head() · shape · info()

df[cols8].head() — alive 컬럼을 기억해두세요

survived	pclass	sex	age	fare	embarked	alive	alone
0	3	male	22.0	7.25	S	no	False
1	1	female	38.0	71.2833	C	yes	False
1	3	female	26.0	7.925	S	yes	True
1	1	female	35.0	53.1	S	yes	False
0	3	male	35.0	8.05	S	no	True

df.shape = (891, 15) | 891명 × 15컬럼
결측: age 177건 deck 688건 embarked 2건

```
#   Column     Non-Null Count  Dtype
0   survived   891 non-null    int64
1   age        714 non-null    float64
2   deck       203 non-null    category
```

⚠ **info()에서 Non-Null Count가 891보다 적으면 결측치 존재 — age(714), deck(203)이 이미 보입니다.**

파트 3

기술통계의 핵심 개념

03

Mean · Median · Mode · Variance · Std · IQR · Skewness · Kurtosis

기술통계 (Descriptive Statistics) 란?

데이터의 특징을 중심 · 퍼짐 · 모양 세 가지 축으로 수치화하는 방법

중심 (Center)

데이터의 대표값은 어디에 있는가?

평균 / 중앙값 / 최빈값

퍼짐 (Spread)

값들이 얼마나 흩어져 있는가?

분산 / 표준편차 / IQR

모양 (Shape)

분포가 어떤 형태인가?

왜도(Skewness) / 첨도(Kurtosis)

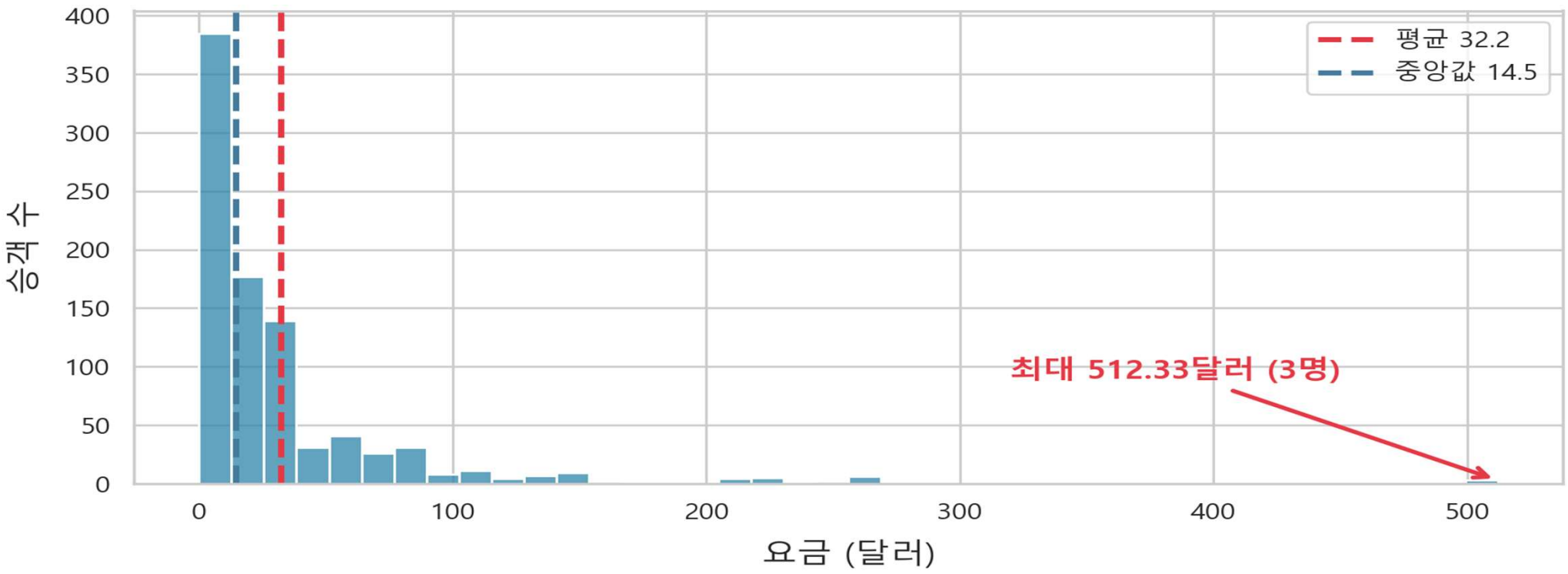
중심 경향 측정 — Mean · Median · Mode

Mean 평균	Median 중앙값	Mode 최빈값
$\Sigma x_i / n$ 모든 값의 합을 개수로 나눔	정렬 후 가운데 값 50번째 백분위수 = Q2	가장 자주 등장하는 값 복수 존재 가능
시소의 무게중심 이상치에 끌려감	이상치에 강건 왜곡 분포에 적합	범주형에도 사용 연속형엔 구간 설정 필요

셋 다 '대표값'이지만 대표 방식이 다릅니다. 이상치가 있는 데이터엔 평균 대신 중앙값을 봅니다.

떡밥 회수 ⭐

평균 32.2 vs 중앙값 14.5 vs 최빈값 8.05

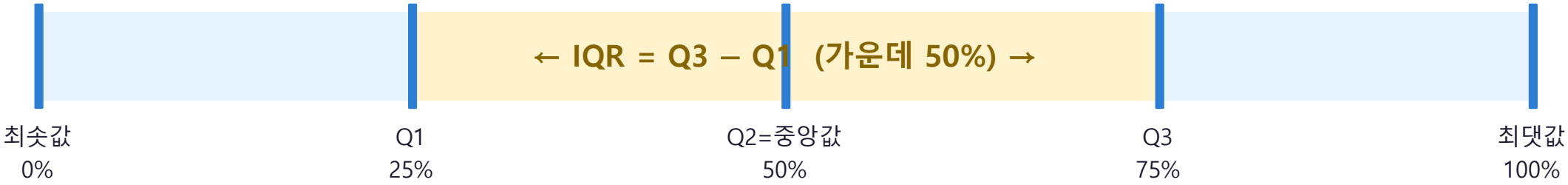


중앙값은 이상치에 강건 → 소득·집값·요금 중앙값 사용

산포도 — Range · Variance · Standard Deviation

Range 범위	Variance 분산	Std Dev 표준편차
max − min 최댓값 − 최솟값	$\Sigma(x_i - \bar{x})^2 / n$ 편차의 제곱 평균	√분산 분산에 루트를 씌워 원래 단위로
⚠️ 극단값 두 개에만 의존 이상치에 가장 취약	단위가 '달러²' 해석이 직관적이지 않음	✅ 가장 많이 사용 원래 단위로 해석 가능

Quantile · Quartile · IQR



Quantile
분위수

데이터를 정렬했을 때 위치
75%ile = 상위 25%

Quartile
사분위수

Q1(25%) · Q2(50%) · Q3(75%)

IQR

Q3 − Q1
가운데 50% 구간의 범위

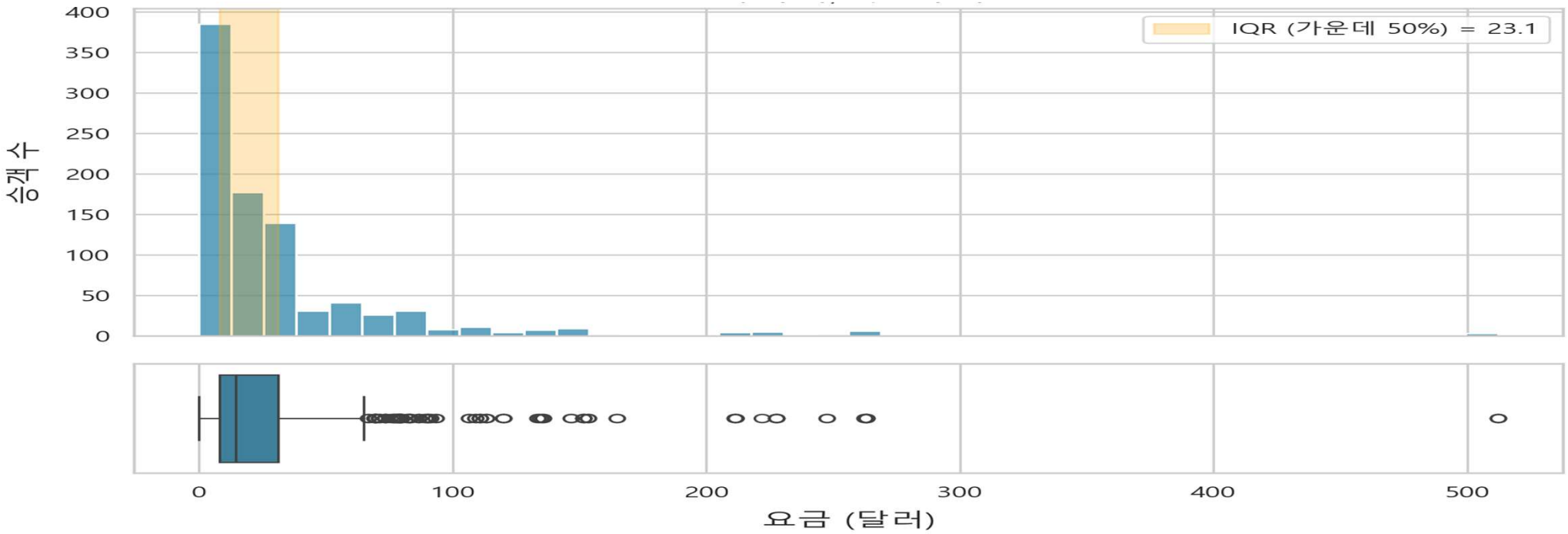
Q2 = Median

IQR의 중간점 = 중앙값

IQR은 이상치에 강건 → 이상치 판정에도 IQR이 쓰입니다.(Box-Plot)

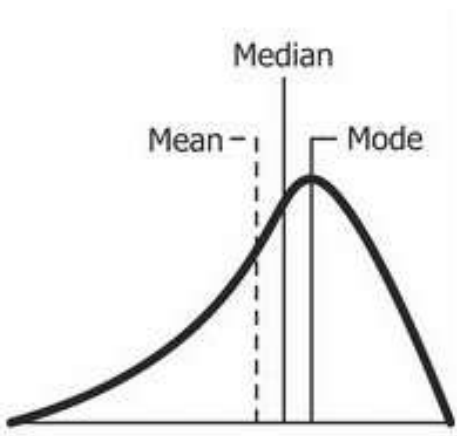
IQR을 눈으로 — Histogram + Box plot

같은 데이터, 두 가지 그림



박스의 폭 = IQR(23.09) / 박스 밖 점들 = 이상치 후보

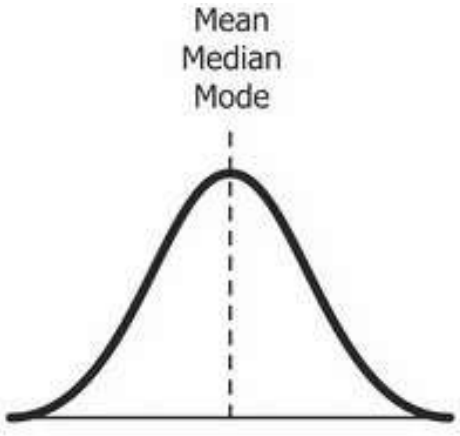
Skewness (왜도)



좌왜 (skew < 0)

꼬리가 왼쪽

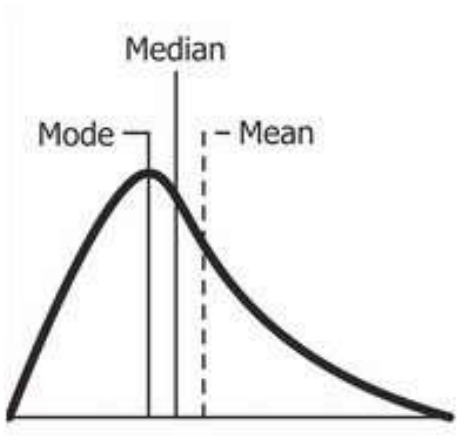
평균 < 중앙값



대칭 (skew ≈ 0)

정규분포

평균 ≈ 중앙값



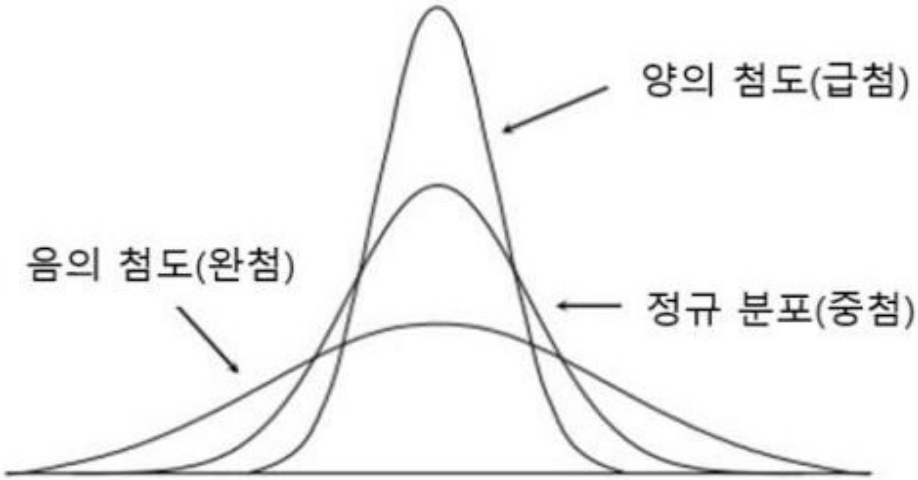
우왜 (skew > 0)

꼬리가 오른쪽

평균 > 중앙값

|skew| > 1 → 강한 비대칭 → 로그 변환 검토 | fare 왜도 4.79 → 강한 우왜 → 변환 필요

Kurtosis (첨도)



$$\text{Kurtosis} = \frac{E[(X-\mu)^4]}{\sigma^4} = \frac{\mu_4}{\sigma^4}$$

$$\text{Excess Kurtosis} = \frac{\mu_4}{\sigma^4} - 3$$

데이터 분포가 평균을 중심으로 얼마나 뾰족한지를 나타내는 통계적 척도
극단적인 이상치 ↑ 첨도 ↑
정규 분포의 첨도 = 3

describe() + 실측 — fare 왜도 4.79, age 왜도 0.39

df[["age","fare","sibsp","parch"]].describe() — fare 열 주목

	age	fare	sibsp	parch
count	714.0	891.0	891.0	891.0
mean	29.7	32.2	0.52	0.38
std	14.53	49.69	1.1	0.81
min	0.42	0.0	0.0	0.0
25%	20.12	7.91	0.0	0.0
50%	28.0	14.45	0.0	0.0
75%	38.0	31.0	1.0	0.0
max	80.0	512.33	8.0	6.0

fare 왜도 4.79 첨도 33.4

|skew| > 1 → 강한 비대칭
→ 로그 변환 필요

age 왜도 0.39 첨도 0.18

|skew| < 1 → 거의 대칭
→ 변환 불필요

파트 4

결측치 · 이상치 · 분포

04

Missing Value · Outlier · IQR · Z-Score · Log Transformation

Missing Value — 결측치의 3가지 원인

결측치: 기록되지 않은 칸 / pandas에서는 NaN(Not a Number)으로 저장 / isnull()로 탐지

입력 누락

기록 누락 또는 인식 불가
일부 데이터가 원래부터 없음

예: age
명부에 나이가 없던 승객

처방: 5~30% → 대체(Imputation)
평균·중앙값·그룹별·KNN 사용

구조적 결측

해당 값이 애초에 존재하지 않음
결측 자체가 의미 있는 정보!

예: deck
3등석엔 갑판 배정 자체가 없음

처방: 50% ↑ → 열 삭제
or 'Unknown' 별도 범주 생성

병합·오류

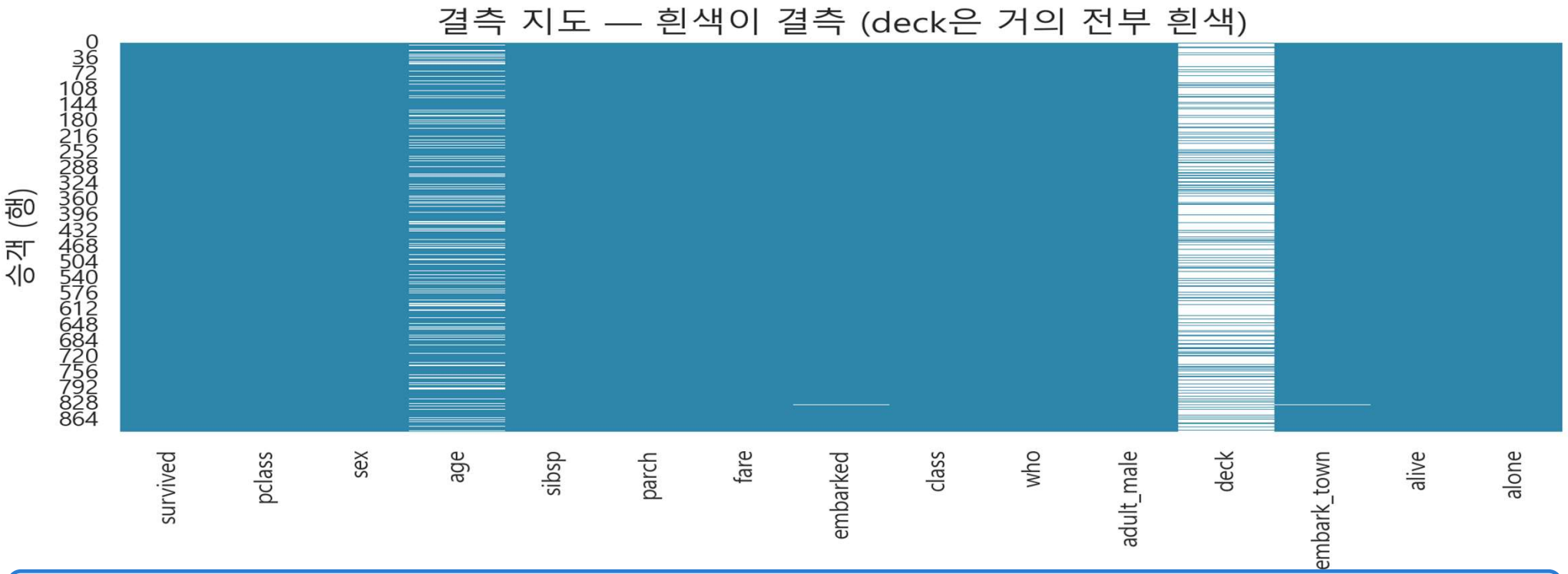
데이터 처리 과정의 실수
시스템 오류·병합 오류

예: embarked
2건 기록 오류

처방: ~5% → 행 삭제
(Row Deletion)

⚠️ **대체값(평균·중앙값)은 반드시 Train에서 계산 → Test에 그대로 적용. Test 통계를 쓰면 Data Leakage**

결측 지도 — isnull() 히트맵



흰색 = 결측 / deck 77.2%(구조적) · age 19.9%(입력 누락) · embarked 0.2%(오류) / ⚠️ alive 컬럼도 보입니다

Outlier — 이상치 탐지 : IQR 규칙 · Z-Score

IQR 규칙 (Tukey's Fence)

하한 = Q1 − 1.5 × IQR

상한 = Q3 + 1.5 × IQR

Z-Score (정규분포를 따를 때)

$z = (x - \mu) / \sigma$

$|z| > 3$ → 이상치

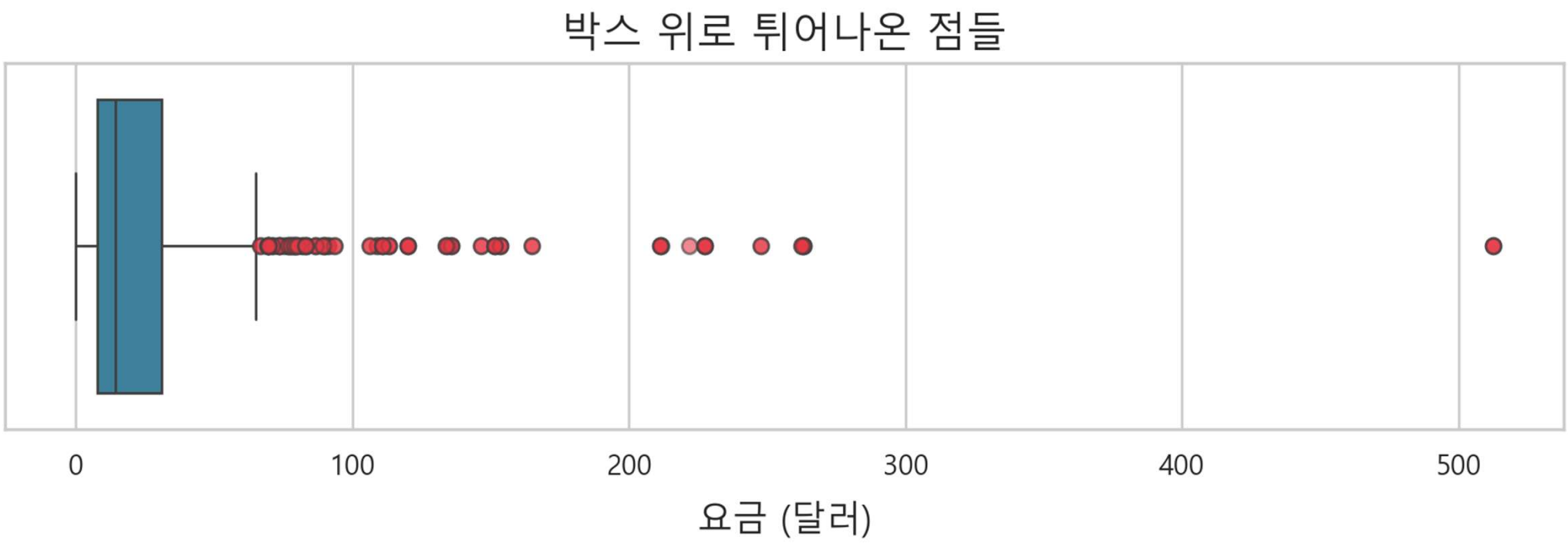
IQR 규칙
→ 정규분포 안 따를 때 (fare처럼 우왜인 경우)

Z-Score
→ 정규분포를 따를 때 (age 같은 경우)

타이타닉 fare 적용: Q1=7.91 Q3=31.00 IQR=23.09 상한=65.63
→ 초과 116건(13%)의 생존율 = 68% (전체 38%의 2배!)

이상치를 눈으로 — Box plot

박스 위로 튀어나온 점들 = IQR 규칙으로 판정된 이상치 후보



상한 65.63 초과 → 116건(13%) / 이 116명의 생존율이 68% → 이상치 제거는 기본값이 아닙니다.

이상치 처리 방향 — 4가지 선택지

오류성 이상치

존재 불가능한 값
나이=200세 요금=-50달러

즉시 제거 or 수정

의미 있는 이상치

실제 존재하는 극단값
=512달러 (1등석)

제거 금지
맥락이 정보

변환

log 변환으로 영향 완화
분포 보정 후 모델 사용

fare → log1p(fare)
왜도 4.79 → 0.39

Robust 모델

이상치에 강한 모델 선택
트리 계열 (RF, XGBoost)

스케일링 불필요
이상치에 덜 민감

이상치가 발견됐다고 바로 제거하면 안 됩니다. 맥락을 보고 판단해야 합니다.

변환(Transformation) vs 스케일링(Scaling)

Log Transformation
로그 변환

분포의 모양 변형

왜도를 줄여 비대칭 분포를 정규분포에 가깝게

np.log(x) → 그대로 로그에 대입(0이면 -∞)

np.log1p(x)→ 1을 더한 다음 로그 대입(0이어도 계산 가능)

Scaling
스케일링

값의 범위 변형

변수 간 단위 통일 → 특정 변수에 편향 X

1. Standard Scaler (표준화) – 평균 0, 표준편차 1로 맞춤
   → 로지스틱·SVM·KNN에 기본
2. Min-Max Scaler (정규화) - 0~1 범위로 압축
   → 이상치에 취약
3. Robust Scaler - 중앙값·IQR 기반
   → 이상치 있을 때 유리 (fare에 적합)

파트 5

변수 간 관계 탐색

05

Correlation · Heatmap · Groupby · Simpson's Paradox · Target Leakage

Correlation — 상관계수

두 수치형 변수 사이의 선형 관계 강도와 방향을 -1에서 +1 사이의 수로 나타낸 값
상관계수는 선형 관계만 측정 → r=0 '관계 없음'이 아니라 '선형 관계 없음'

r = -1	r = -0.5	r = 0	r = +0.5	r = +1
완벽한 음의 선형	중간 음의 상관	선형 관계 없음	중간 양의 상관	완벽한 양의 선형

피어슨 상관계수: $r = \Sigma(x_i - \bar{x})(y_i - \bar{y}) \;/\; \sqrt{[\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2]}$ ← 선형 관계를 측정

Correlation Matrix · Heatmap · 3가지 맹점

Correlation Matrix

여러 수치형 변수의 상관계수를 한 표로 정리

Heatmap: 색으로 표현 (붉을수록 양, 파랄수록 음)

⚠ **수치형 변수만**

df.corr(numeric_only=True)

→ 범주형은 표에 아예 없음

① 비선형

상관계수는 선형 관계만 측정

r=0 ≠ '관계 없음'

r=0 = '선형 관계 없음'

age↔survived = -0.077
→ 어린이만 생존율 59%
상관계수는 이 신호를 놓침

② 범주형 부재

df.corr()는 수치형만 포함

범주형 변수가 표에

아예 나오지 않음

sex·embarked가 히트맵에 없음
→ 가장 중요한 변수가
표에 없었던 것!

③ 숨은 변수

교란변수가 있으면

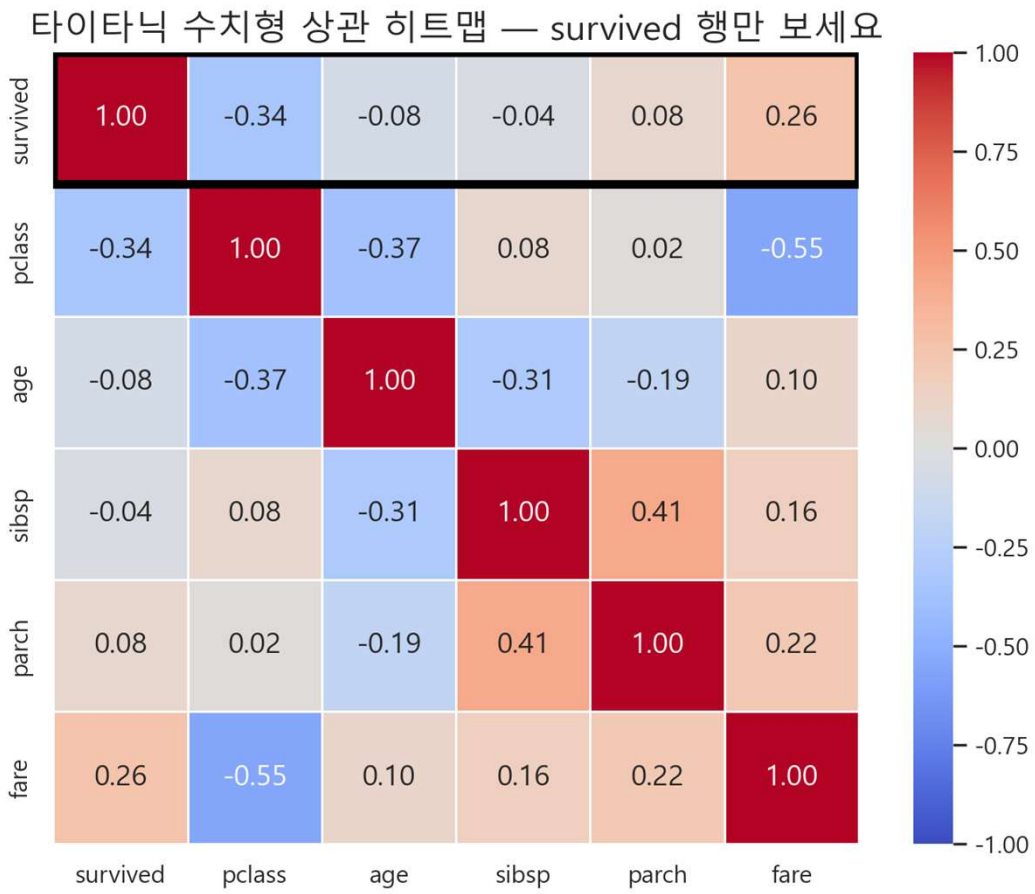
상관의 방향(부호)까지 뒤집힘

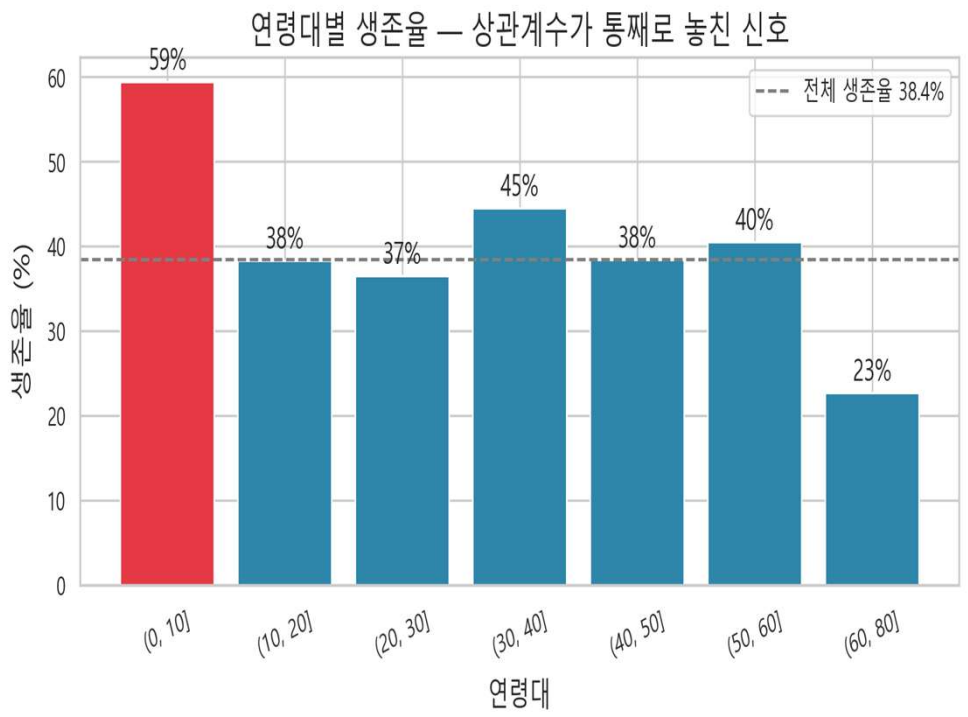
→ Simpson's Paradox

펭귄: 전체 r=-0.23(음)
→ 종별로 나누면
전부 양의 상관!

① 비선형

survived 행만 보세요 / 가장 큰 상관이 0.34입니다





age↔survived r=-0.08
0~10세만 생존율 59%로 급등!

② 범주형 부재

df.corr(numeric_only=True) 포함 컬럼
sex·embarked·alive는 빠집니다

컬럼	corr() 포함 여부
survived	포함
pclass	포함
sex	빠짐 (범주형)
age	포함
sibsp	포함
parch	포함
fare	포함
embarked	빠짐 (범주형)
class	빠짐 (범주형)
who	빠짐 (범주형)
adult_male	빠짐 (범주형)
deck	빠짐 (범주형)
embark_town	빠짐 (범주형)
alive	빠짐 (범주형)
alone	빠짐 (범주형)

히트맵에는 성별 X
가장 중요한 변수 = 성별

성별 피처 제거 시 -8.0%p / 나머지 6개 feature 전부 ±1%p 이내

범주형 관계 분석 — Groupby · Cross Tabulation · Grouped Box

Groupby

범주로 쪼개서 수치형 요약
→ 범주형 × 수치형 관계

```
df.groupby('pclass')['survived'
].mean()
```

히트맵이 못 잡는 범주별 패턴

Cross Tabulation

두 범주형 변수의 교차 빈도·비율
→ 범주형 × 범주형 관계

```
pd.crosstab(df['sex'],
df['survived'],
normalize='index')
```

히트맵에 없던 범주형 변수

Grouped Box plot

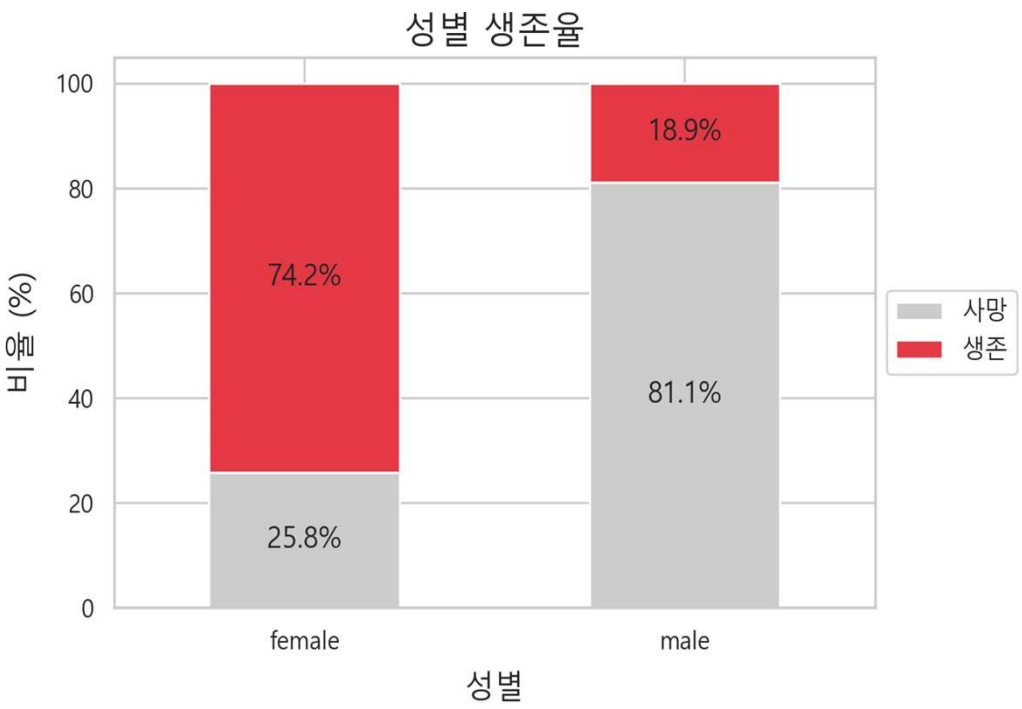
그룹별 분포를 나란히 비교
→ 범주형 × 수치형 분포 차이

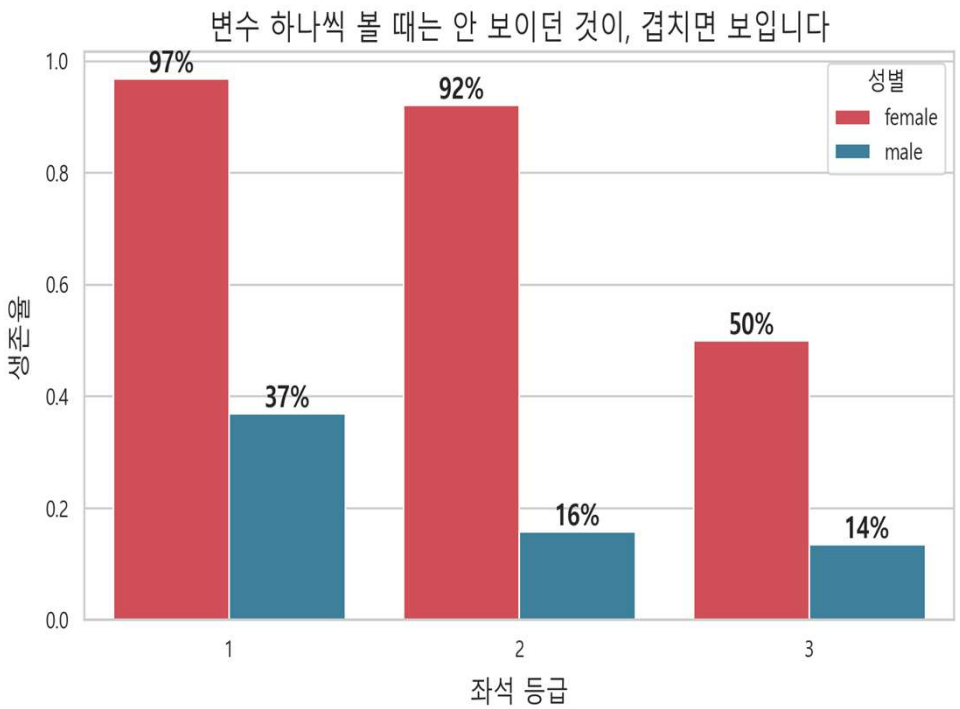
```
sns.boxplot(x='pclass',
y='fare',
data=df)
```

그룹 간 분포 비교 목적

💡 **변수 3개 이상 한눈에: sns.pairplot(df, hue='survived') — 전체 변수쌍의 산점도+분포를 한 번에**

범주형 관계 실측 — 성별 생존율 · 등급별 생존율

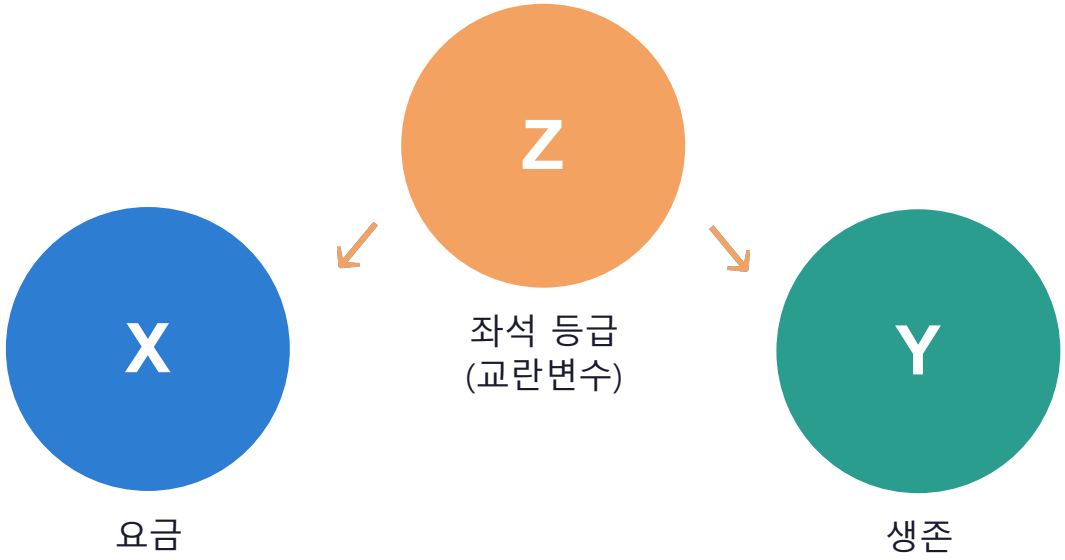




여성 74.2% vs 남성 18.9% / 1등석 여성 97% vs 3등석 남성 14%

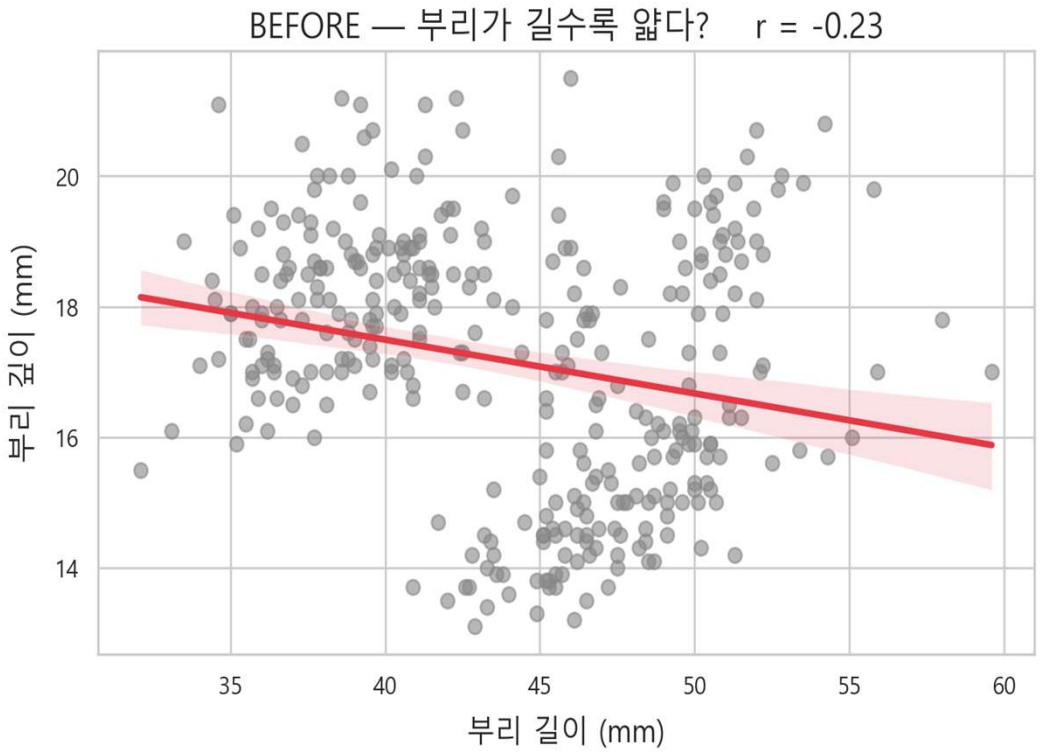
③ 숨은 변수 — 교란변수 (Confounder)

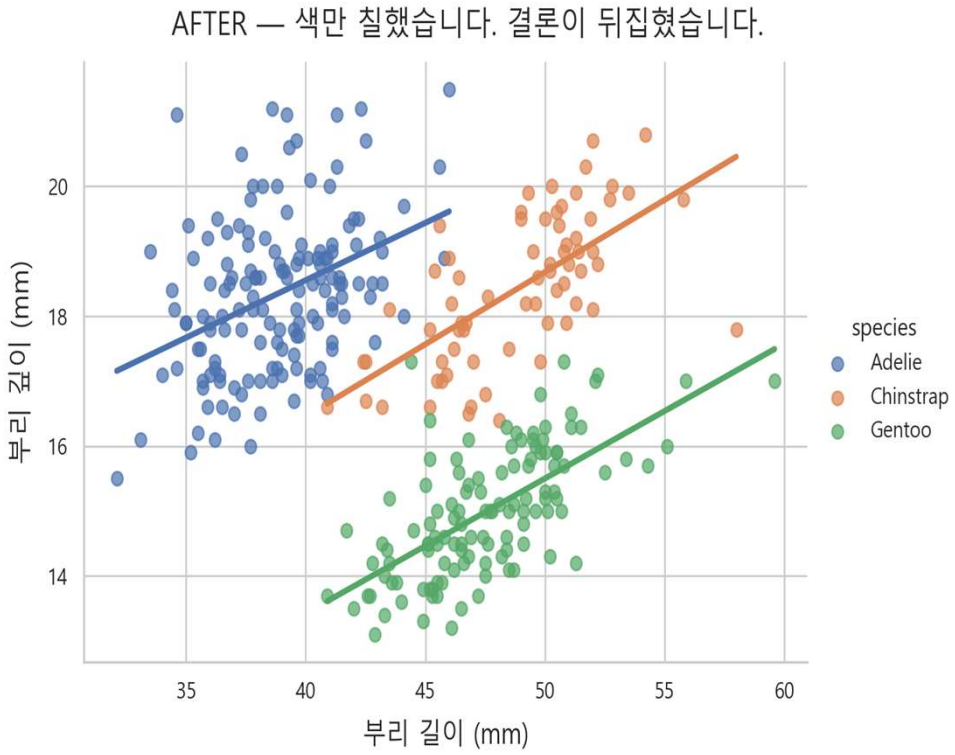
상관관계 ≠ 인과관계 (Correlation does not imply Causation)



교란변수: X와 Y 둘 다에 영향을 주는 제3의 변수 Z — 이것이 있으면 X↔Y의 상관 부호까지 뒤집힐 수 있습니다.

Simpson's Paradox





전체: r = -0.23 (부리 길수록 얇다?)

종별 분리: 전부 양의 상관 (+0.39~+0.65)

🎣 EDA 없이 컬럼 14개를 모두 넣고 모델을 돌리면?

정확도

100.0%

30번 반복해도 100.0% (±0.0)

우리는 방금 완벽한 모델을 만들었습니다. …정말요?

Target Leakage

학습 시점에는 존재할 수 없는 정보가 feature에 섞여 들어가는 것

탑승
(과거)

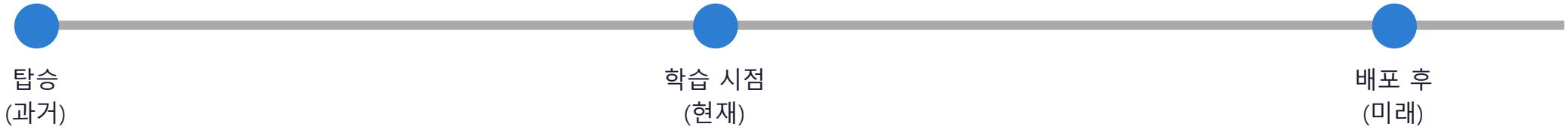
학습 시점
(현재)

배포 후
(미래)

survived	pclass	sex	age	fare	embarked	alive	alone
0	3	male	22.0	7.25	S	no	False
1	1	female	38.0	71.2833	C	yes	False
1	3	female	26.0	7.925	S	yes	True
1	1	female	35.0	53.1	S	yes	False
0	3	male	35.0	8.05	S	no	True

Target Leakage

모델 학습 시점에 사용할 수 없는 정보가 feature에 포함되는 문제



✅ **사용 가능**

pclass · sex · age · fare · embarked
학습 시점에도 알 수 있는 값

❌ **사용 불가 (미래 정보)**

alive → 생사 확인 후의 정보
학습 시점엔 존재하지 않는 값!

실측: pd.crosstab(alive, survived) → 일치율 100% | alive='no'는 전부 사망, 'yes'는 전부 생존

파트 6

전처리 · 모델링 연결

06

Encoding · Scaling · Feature Engineering · Modeling Strategy

복기 — EDA에서 내린 의사결정

EDA의 산출물은 그래프가 아니라 의사결정이다.

EDA Decision Table — The real output of EDA is not a graph, it's this table

Variable	Observation from EDA	Decision
alive	Target Leakage — 100% match with survived	Drop immediately
deck	Missing 77.2%, structural	Fill 'Unknown'
age	Missing 19.9%, input omission	Median impute (Train only)
fare	Skewness 4.79, strong right-skewed	log1p transform
sex	Nominal — no order	One-Hot Encoding
pclass	Ordinal — 1st > 2nd > 3rd class	Ordinal Encoding
embarked	Nominal, 2 missing rows	One-Hot + drop 2 rows

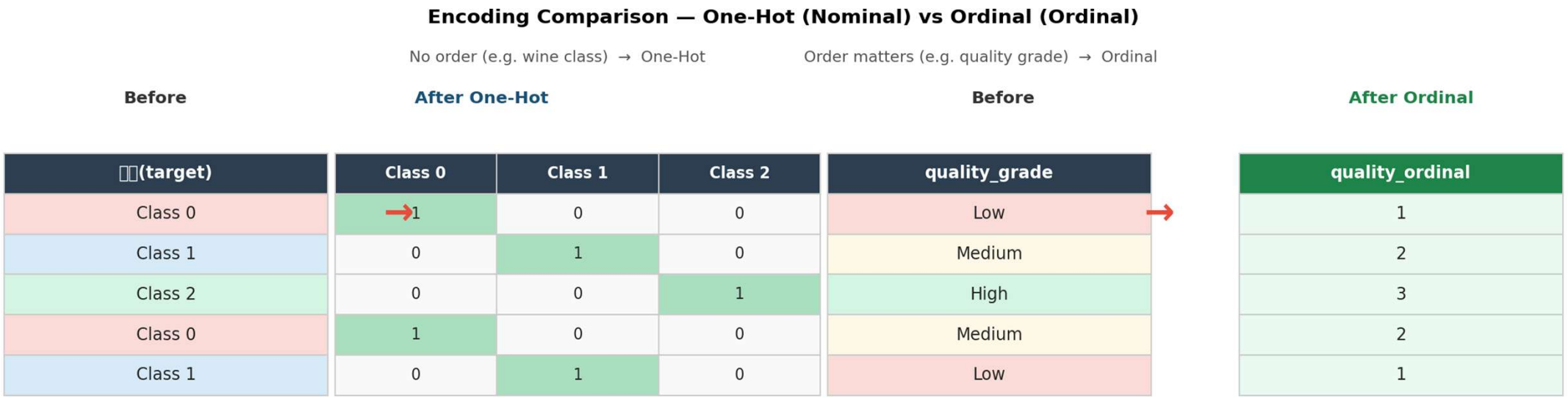
EDA 실습 — Wine Dataset

EDA의 산출물은 그래프가 아니라 의사결정이다.

class	Alcohol	Malic acid	Ash	Alcalinity	Magnesiu	Total pher	Flavanoids	Nonflavan	Proanthoc	Color inte	Hue	OD280/OD	Proline
1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
1	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
1	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
1	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
1	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045

인코딩 비교 — One-Hot(명목형) vs Ordinal(순서형)

왼쪽: 품종처럼 순서 없는 범주 → One-Hot | 오른쪽: 품질등급처럼 순서 있는 범주 → Ordinal

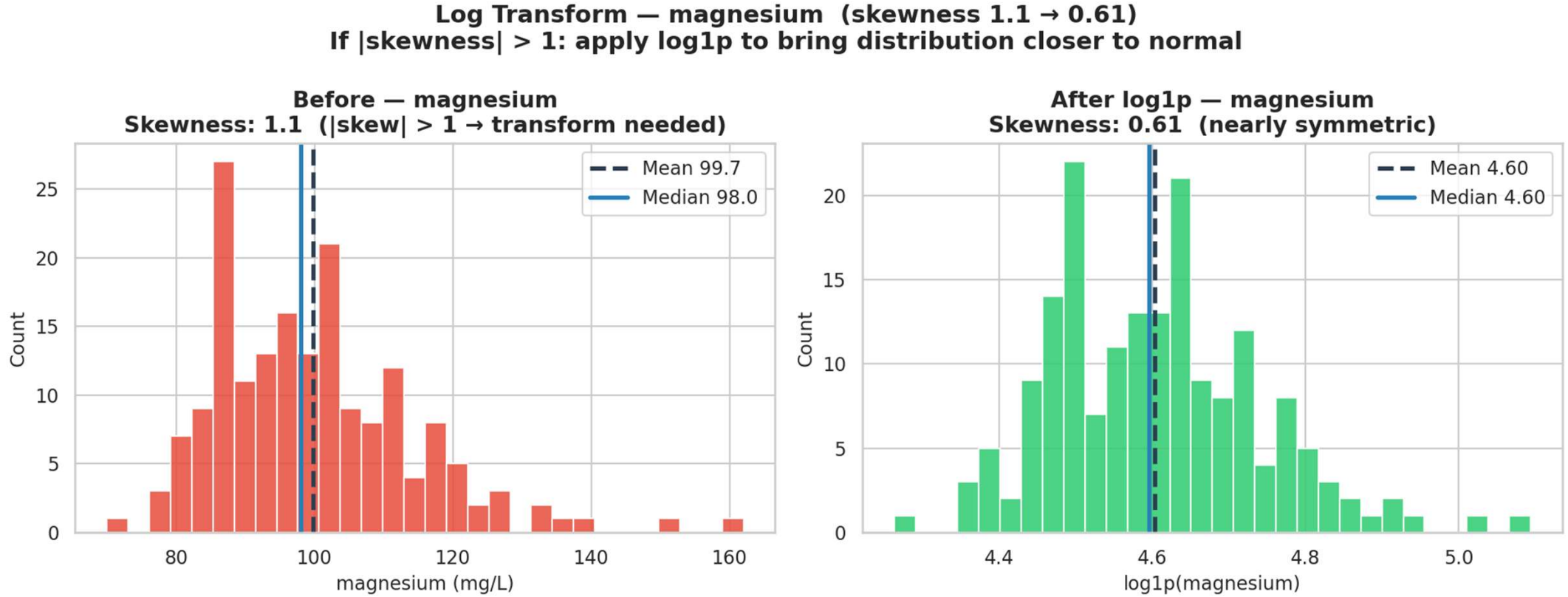


One-Hot 인코딩 명목형 변수에 사용. 품종처럼 Class 0/1/2가 크기 관계 없이 구분만 될 때, 각 범주를 별도 0/1 컬럼으로 만든다.

Ordinal 인코딩 순서형 변수에 사용. Low < Medium < High처럼 순서가 의미 있을 때, 직접 숫자 순서를 지정

로그 변환 — 왜도 1.10 → 0.31 (magnesium)

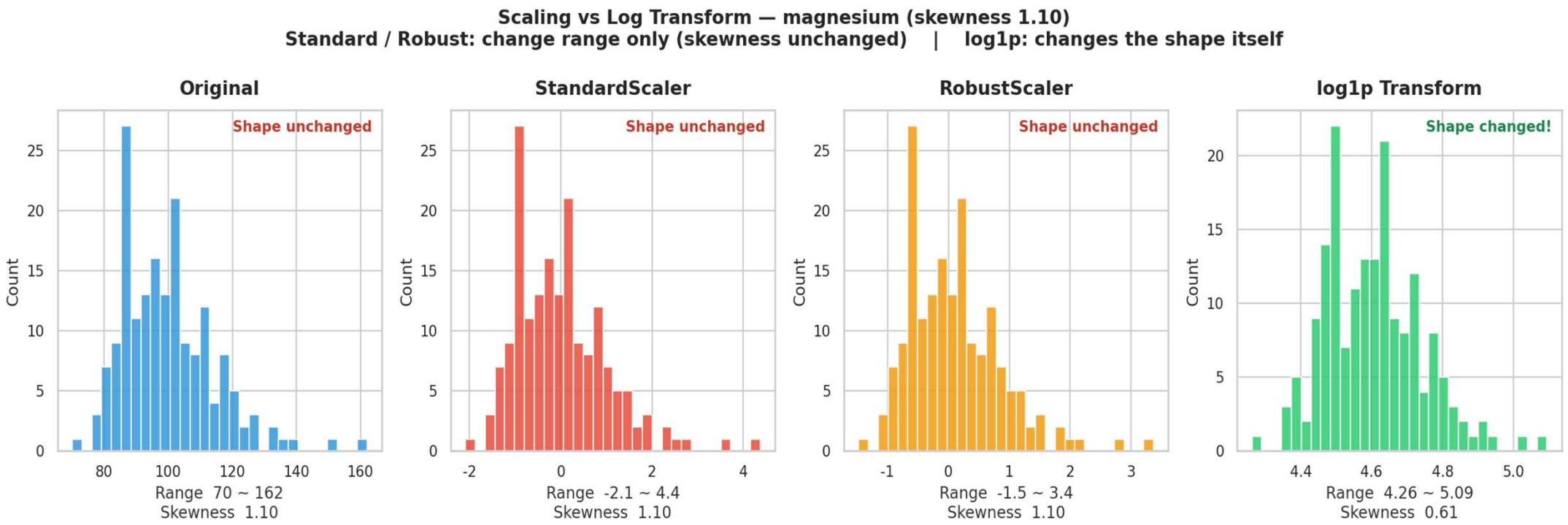
|왜도| > 1이면 강한 비대칭 → log1p 변환으로 정규분포에 가깝게



log1p(x) = log(1+x) x=0일 때 안전하게 사용 가능. 스케일링과 달리 분포의 "모양(왜도)" 자체를 바꾼다.

스케일링 vs 로그 변환 — magnesium (왜도 1.10)

StandardScaler / RobustScaler: 범위만 바꾼다 (왜도 유지) | log1p: 분포 모양 자체를 바꾼다



스케일링 : Standard와 Robust 모두 왜도가 1.10로 그대로유지, 분포 모양 변화 X. 단순히 값의 범위(scale)를 조정할 뿐이다.

log1p = 모양 변환 왜도가 1.10 → 0.61로 줄어들어 분포가 더 대칭적. 스케일링과 별개의 작업이며, 둘 다 적용하는 것도 가능하다.

모델별 성능 비교 (5-fold CV) — Wine Dataset

KNN & SVM: 스케일링 없이 ~0.67 → 스케일링 후 ~0.98 (+30%p) | RF: 변화 없음

Model Performance Comparison (5-fold CV) — Wine Dataset (scale range: 0.13 ~ 1,680)
KNN & SVM: ~0.67 without scaling → ~0.98 with scaling (+30%p) | RF: no change

Condition	Model	Scaler	Accuracy	F1(macro)	Note
① KNN (no scaling)	KNN	None	0.680	0.650	
② KNN + Standard	KNN	Standard	0.972	0.972	↑ distance-based: scaling critical
③ KNN + Robust	KNN	Robust	0.961	0.962	
④ SVM (no scaling)	SVM	None	0.674	0.625	
⑤ SVM + Standard	SVM	Standard	0.983	0.983	↑ kernel-based: scaling critical
⑥ SVM + Robust	SVM	Robust	0.978	0.977	
⑦ RF (no scaling)	RF	None	0.977	0.978	
⑧ RF + Standard	RF	Standard	0.977	0.978	= ⑦ tree: scaling irrelevant

KNN & SVM — 스케일링 없을 때

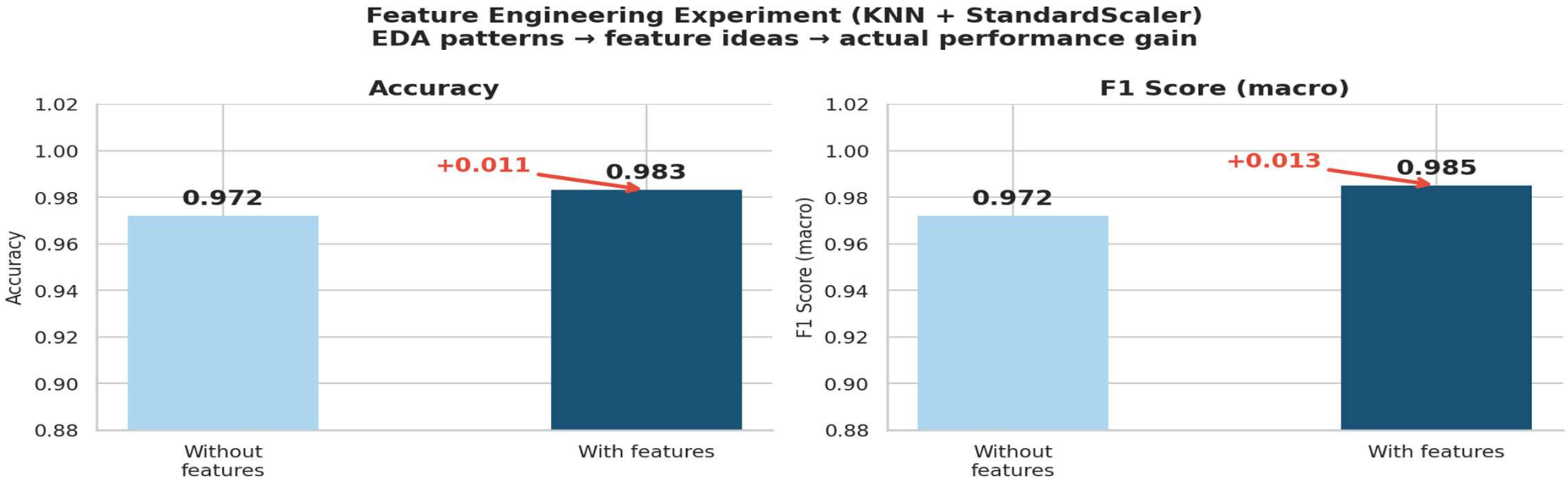
거리/커널 계산이 큰 값(proline ~700)에 지배당한다. 3개 클래스를 사실상 무작위로 맞히는 수준(~0.67).

RF — 스케일링 불필요

트리는 분기 기준값(threshold)으로 판단하므로 변수의 절대적 스케일이 영향을 주지 않는다. 0.977 = 0.977.

파생변수 추가 실험 (KNN + StandardScaler)

EDA에서 발견한 도메인 패턴 → 피처 아이디어 → 실제 성능 향상



phenol_ratio = total_phenols / flavanoids color_alcohol = color_intensity × alcohol mineral_index = magnesium + ash × 10

페놀_비율

total_phenols / flavanoids

페놀 구성 비율

색상_알코올

color_intensity x alcohol

색상-강도 교차항

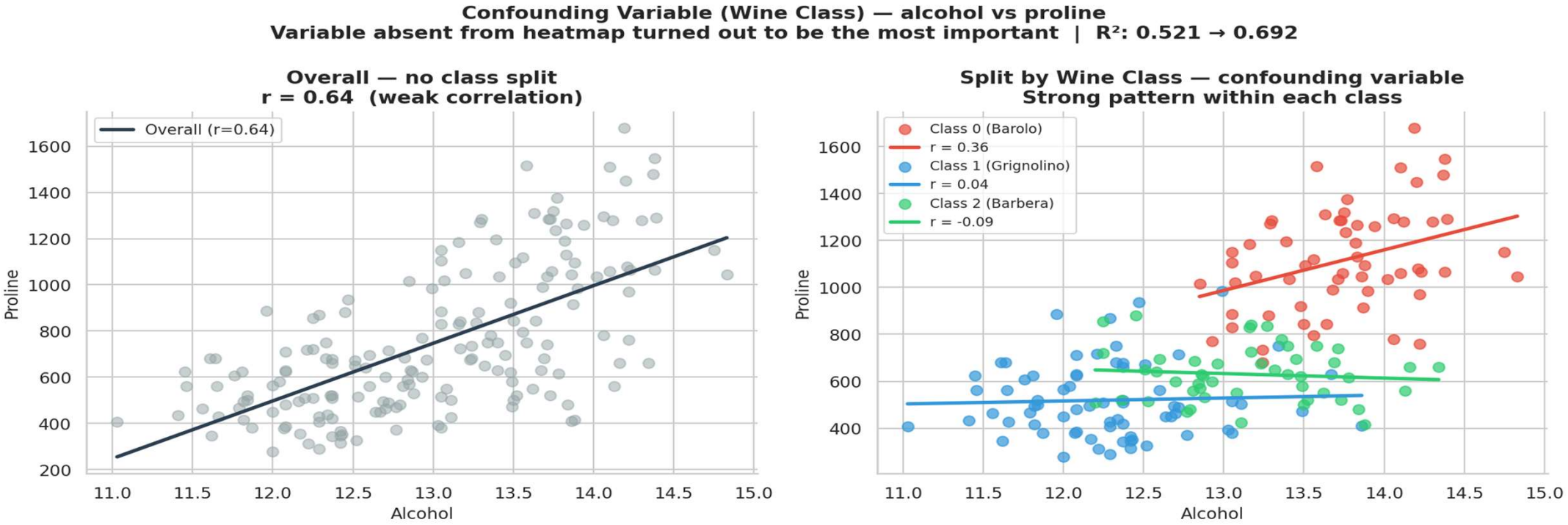
미네랄_지수

magnesium + ash x 10

미네랄 함량 지수

교란변수 — 품종(Wine Class)

alcohol vs proline: 전체 상관은 약하지만, 품종별로 나누면 패턴이 드러난다

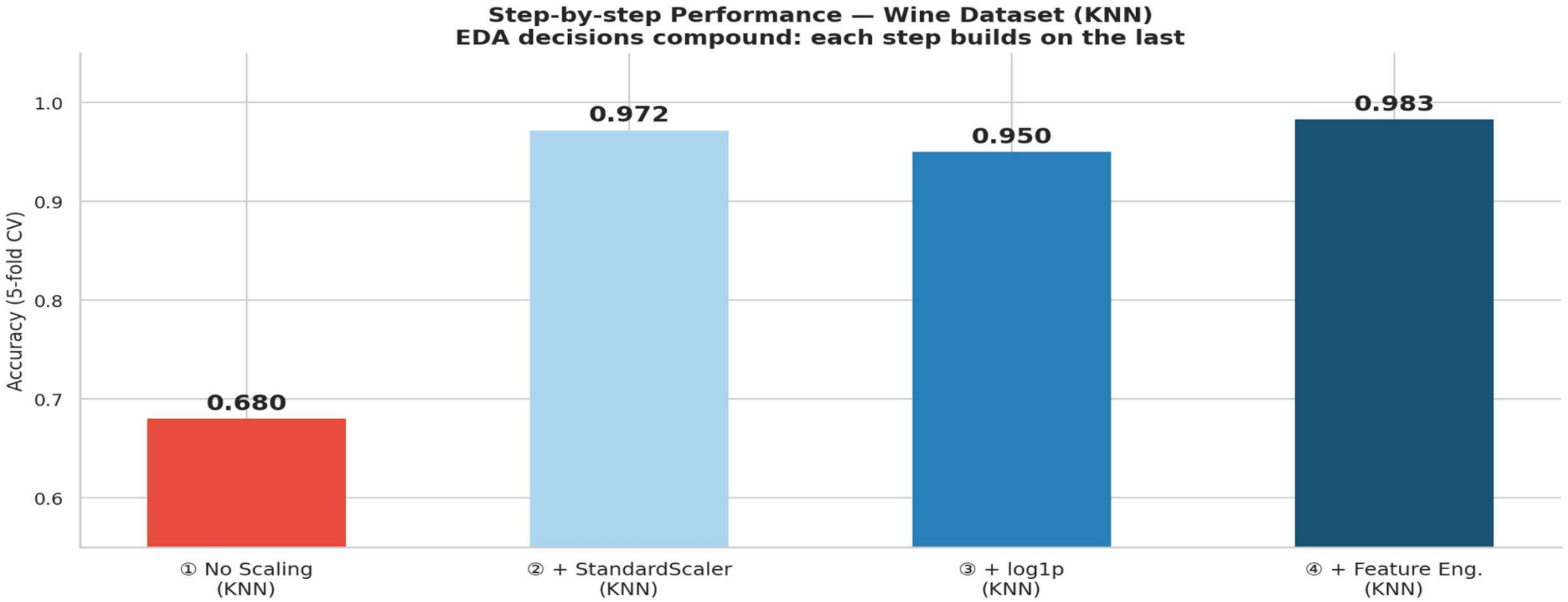


심슨의 역설 전체로 보면 alcohol과 proline의 상관 ↓

그러나 품종별로 나누면 → 뚜렷한 패턴 생성 / 품종은 히트맵에 나타나지 않는 범주형 교란변수(프롤린 예측 R^2: 품종 없이 0.521 → 품종 추가 후 0.692)

단계별 성능 변화 — Wine Dataset (KNN)

EDA 의사결정이 쌓여서 성능이 된다



"EDA의 의사결정은 결과물이 아니라, 더 나은 모델을 향한 입력값이다"

EDA 의사결정 목록 — Wine Dataset

EDA의 산출물은 그래프가 아니라 의사결정이다.

EDA Decision Table — Wine Dataset
The real output of EDA is not a graph, it's this table

Variable	Observation from EDA	Decision
magnesium	Skewness 1.10 — strong right skew	log1p transform
malic_acid	Skewness 1.04 — strong right skew	log1p transform
proline	Range 278~1,680 — largest scale	Scaling required
nonflavanoid_phenols	Range 0.13~0.66 — smallest scale	Scaling required
color_intensity	Range 1.28~13.0 — large variance	Scaling required
target (class)	3 classes (0,1,2) — nominal, no order	Classification target

마무리

EDA 및 기술통계

07

1. EDA는 데이터 분석 이전의 데이터 이해 과정
2. 기술통계는 데이터의 중심, 퍼짐, 분포를 요약하는 방법
3. 시각화는 결측치, 이상치, 변수 간의 패턴 및 관계를 확인하는 도구
4. 좋은 EDA는 전처리와 모델링 전략으로 연결

객관식

Q. 다음은 이번 강의 전체 내용을 요약한 설명들입니다. 이 중 **알맞지 않은 것을 모두 고르세요.**

1. EDA는 모델링에 앞서, 가설을 세우기 이전에 데이터가 무엇을 말하는지 들어보는 과정이다.
2. 타이타닉 fare의 평균(32.2)이 중앙값(14.5)보다 훨씬 큰 이유는 소수의 매우 큰 요금이 평균을 끌어올렸기 때문이며, 이런 경우 중앙값이 이상치에 더 강건한 대표값이다.
3. Deck 변수처럼 결측치 비율이 77.2%로 매우 높아도 무조건 평균이나 중앙값으로 대체하여 정보 손실을 최소화하는 것이 올바른 처리 방법이다.
4. fare 변수는 왜도(skewness) 4.79로 강한 우왜(right-skewed) 분포를 보이며, 이런 경우 log1p 변환을 통해 왜도를 줄여 정규분포에 가깝게 만들 수 있다.
5. 상관계수 히트맵은 범주형 변수만 포함하므로, 성별처럼 중요한 수치형 변수의 영향력이 표에 아예 드러나지 않을 수 있다.

서술형

Q. EDA 없이 14개 컬럼을 모두 넣고 모델을 돌렸더니 정확도가 30회 반복 모두 100.0%(±0.0)로 나왔다. 왜 이것이 "완벽한 모델"이 아니라 오히려 **문제가 있다는 신호**인지, Target Leakage 개념을 이용해 설명하시오.