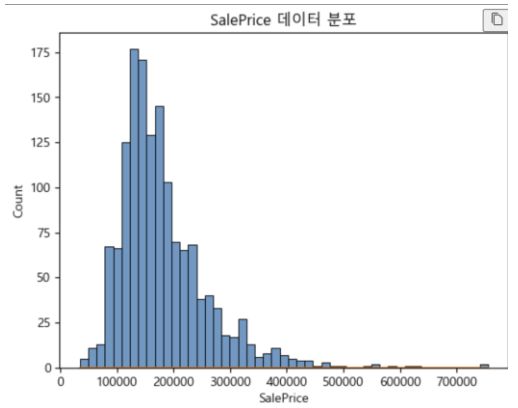
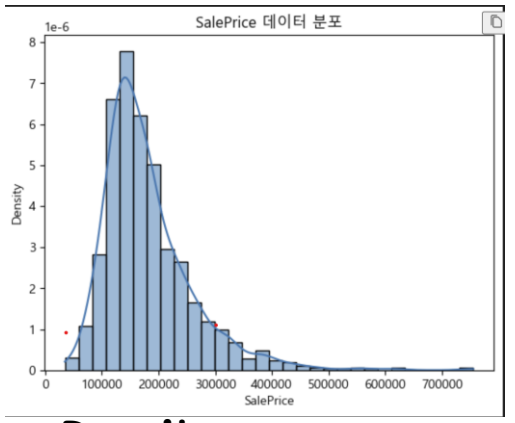
Housing Price Data

데이터 분포

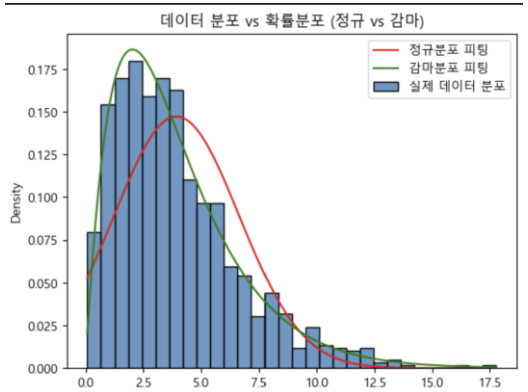
데이터를 설명하는 두 가지 방법

- **상관관계 기반**
  (ex. KDE기반 데이터 분포 모델링)
- **수학적 모델링**
  (ex. 정규분포, 감마분포)





Count -> Density

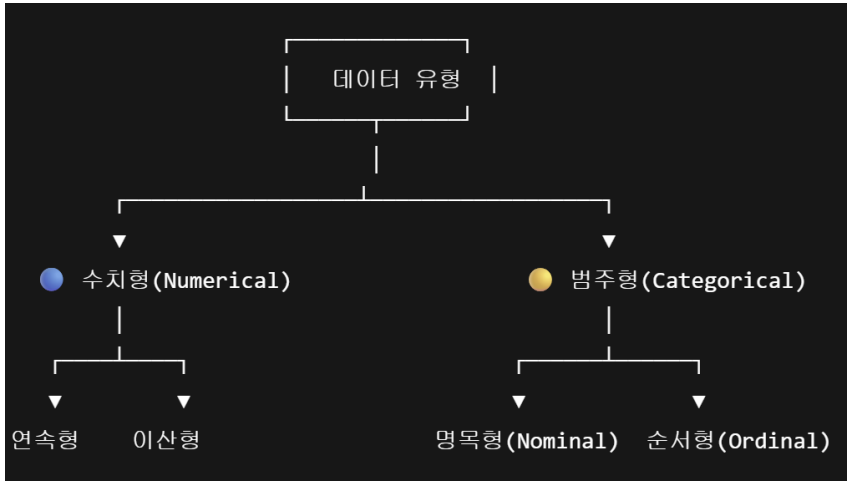


감마 / 정규 분포로 fitting

데이터를 예측하는 방법

- **상관관계 기반**
  : 독립변수(설명변수)들과의 상관관계를 이용해 회귀모델 등으로 종속변수(목표 변수) 예측
  (ex. 다중 선형회귀방법 등 지도학습 모델)
- **수학적 모델링**
  : 예측하려는 변수의 확률분포를 근사하고, 해당 분포를 기반으로 확률적으로 값을 추정하거나 시뮬레이션
  (ex. 감마분포를 따르는 고장 대기시간에서 특정 시간 초과 확률 예측 등)

데이터 유형



수치형 (Numerical)

숫자 형태로 나타나는 데이터. 연산(+, -, 평균 등)이 **의미 있음**.

1.1 연속형 (Continuous)

- 실수, 소수 포함 가능
- **무한한 값 사이**가 존재
- 예: 키(175.4cm), 몸무게(60.2kg), 온도, 시간, 매출액 등
- ➕ ➗ 평균·표준편차 등 통계 적용 가능

1.2 이산형 (Discrete)

- 정수, 개수 형태. 값의 개수가 **유한하거나 셀 수 있음**
- 예: 고객 수, 사고 횟수, 월별 접속 수 등
- 📌 연속형과 달리 **값 사이가 비어 있음**

범주형 (Categorical)

숫자/문자처럼 보여도, **연산이 무의미**하고 그룹/범주를 나타냄

2.1 명목형 (Nominal)

- 단순한 이름, 라벨. **서열 없음**
- 예: 성별(남/여), 국가, 혈액형(A/B/O), 색깔, 부서명
- 📌 one-hot encoding 대상

2.2 순서형 (Ordinal)

- **서열은 있지만**, 간격은 불명확
- 예: 학력(고졸 < 대졸 < 대학원), 만족도(1~5점), 등급(A/B/C)
- 📌 label encoding 가능 (단, 주의 필요)

✅ 예시 정리

변수명	데이터 값 예시	유형	설명
성별	남, 여	명목형	그룹에만 의미, 순서 없음
학력	고졸, 대졸, 대학원	순서형	순서만 있고 수치 간격은 불명확
키	170.2cm	연속형	실수, 평균 가능
자녀 수	0, 1, 2	이산형	정수, 셀 수 있음
국가코드	KR, US, JP	명목형	라벨 의미, 수학적 연산 불가
평점	1~5점	순서형	점수지만 보통 순서형 취급

데이터 유형

✅ 왜 중요할까?

- **분석 기법 선택**에 영향
  - 피어슨 상관계수 → **연속형**
  - 카이제곱검정 → **범주형**
- **시각화 방식**
  - 범주형: bar chart, pie chart
  - 수치형: histogram, boxplot, lineplot 등
- **전처리 방식**
  - 명목형 → one-hot encoding
  - 순서형 → label encoding
- **모델링 특성**
  - 범주형 변수는 직접 사용 불가 → 인코딩 필요

데이터 유형	사용 가능한 통계
연속형	평균, 표준편차, 회귀분석 등
이산형	빈도분석, 포아송 회귀 등
명목형	카이제곱검정, 비율분석
순서형	스피어만 상관, 순위회귀 등

Correlation

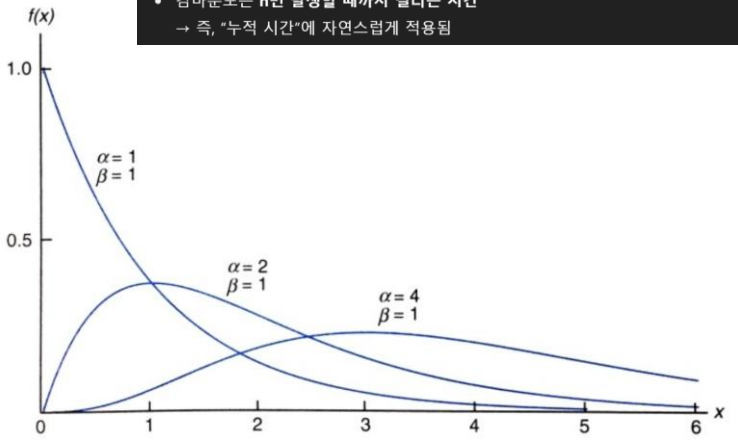
- **감마분포**
  -> 연속형 확률분포 중 하나
  -> 양의 실수값을 가지는 확률변수에 적용

$$f(x\,;\,\alpha,\beta)=\frac{1}{\beta^{\alpha}\cdot\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}=\frac{1}{\beta^{\alpha}\cdot(\alpha-1)!}x^{\alpha-1}e^{-\frac{x}{\beta}}$$

데이터 종류	공통 특징	감마분포와의 연결
대기시간, 생존시간	절대 0보다 작지 않음 (음수 불가)	감마는 0 이상의 연속형 분포
비용, 지출	대부분 0 근처에 많고 일부는 매우 큼	오른쪽 꼬리가 긴 비대칭 분포
누적 작업시간	반복된 사건의 누적합	감마는 포아송 간격의 누적 분포 (exponential 합)

✨ **감마분포 = 여러 개의 지수분포(Exponential)의 합**

- 지수분포는 "한 번 발생할 때까지 걸리는 시간"
- 감마분포는 **n번 발생할 때까지 걸리는 시간**
  → 즉, "누적 시간"에 자연스럽게 적용됨



감마분포를 간략히 표현하자면 **α 번째 사건이 일어날 때 까지 걸리는 시간에 대한 연속 확률분포** 입니다. 즉, 총 α 번의 사건이 발생할 때까지 걸린 시간에 대한 확률분포를 보여줍니다. 여기서 β는 포아송 분포의 모수와 비슷한 역할을 합니다.(감마 분포에서는 α, β 둘 다 모수(parameter)라고 부릅니다. 단지 이 둘의 역할이 다를 뿐이죠. **α**는 **'형태 모수(shape parameter)'**, **β**는 **'척도 모수(scale parameter)'**라고 합니다.)

- **상관관계 기반**
  -> pearson correlation 는 데이터 분포가 정규분포여야 함.
  -> 즉, 데이터 분포의 수학적 모델링이 필요함