머신러닝 데이터 전처리 '국룰' 5단계

1단계: 재료 상태 확인 (Data Inspection)

일단 데이터를 불러와서 상한 곳은 없는지, 어떻게 생겼는지 눈으로 봅니다.

- **뭐 하냐?**: `head()` (앞부분 5줄 보기), `info()` (데이터 타입 확인), `describe()` (통계 요약) 찍어보기.
- **목적**: "아, 날짜가 문자로 되어있네? 숫자가 비어있는 곳(Null)이 있네?" 하고 견적을 내는 단계입니다.

2단계: 맛보기 & 간 보기 (EDA - 탐색적 데이터 분석)

말씀하신 **"Boxplot 때리고, 그래프로 y 데이터 보고"**가 바로 이 단계입니다. 데이터의 특징을 시각적으로 파악합니다.

- **타겟(y) 확인**: 내가 맞춰야 할 정답(y)이 어떻게 분포되어 있는지 봅니다. (예: 히스토그램)
  - *한쪽으로 쏠려있으면 나중에 모델이 멍청해질 수 있어서 확인 필수!*
- **Boxplot**: 이상한 값(Outlier, 튀는 값)이 있는지 잡아냅니다. (아까 이미지에서 보신 그 작업)
- **상관관계 (Heatmap)**: "X랑 Y랑 얼마나 친해?"를 봅니다. (상관계수 확인)

3단계: 재료 다듬기 (Data Cleaning)

2단계에서 발견한 문제점들을 수정합니다.

- **결측치 처리**: 비어있는 값을 평균으로 채우거나(`fillna`), 그냥 지워버립니다(`dropna`).
- **이상치 처리**: Boxplot에서 발견한 말도 안 되는 값(예: 은 가격이 90달러)을 잘라내거나 조정합니다.

4단계: 양념 치기 & 모양 내기 (Feature Engineering - 특성 공학)

말씀하신 **"파생변수 만들기"** 단계입니다. 모델이 더 잘 학습하도록 데이터를 가공합니다.

- **파생변수 생성**: 기존 데이터로 새로운 힌트를 만듭니다.
  - 예: `날짜` → `요일` , `월` , `분기` 로 쪼개기 (주말엔 가격이 안 움직이니까 요일이 중요할 수 있음)
  - 예: `가격` → `이동평균(MA)` , `전일 대비 변동폭` 계산
- **스케일링 (Scaling)**: 숫자의 단위를 맞춥니다. (예: 가격은 20000원인데 거래량은 100만 주면 모델이 헷갈리니까 둘 다 0~1 사이로 압축)

5단계: 알짜배기 고르기 (Feature Selection)

만든 변수가 너무 많으면 오히려 방해가 됩니다. 중요한 것만 남깁니다.

- **상관관계 높은 것 제거**: A랑 B가 거의 똑같이 움직이면 둘 중 하나는 버립니다.
- **중요도 확인**: 모델한테 "뭐가 제일 도움 됐어?" 물어보고(Feature Importance) 쓸모없는 건 쳐냅니다.