📚 🏠 Boston Housing 회귀 분석 캡스톤

아이티윌 데이터분석 부트캠프 52기 이광호 강사 (leekh4232@gmail.com)

우리는 흔히 “방이 많을수록 집값은 비싸다”, “범죄율이 높은 지역은 집값이 낮다”라고 말한다.

그 말은 **어디까지가 사실이고, 어디부터가 단순한 인상**일까?

이번 캡스톤 과제에서는 Kaggle의 **Boston Housing 데이터셋**을 활용하여 지역·환경·교육·소득·범죄 요인이 **주택 가격을 어떤 구조로 설명하는지**를 데이터와 회귀모형을 통해 단계적으로 밝혀본다.

이 과제의 목표는 예측 정확도를 높이는 것이 아니라, **“보스턴 주택 가격은 어떤 논리로 형성되는가”를 수치와 언어로 설명하는 것**이다.

※ 본 과제는 팀 / 개인 단위 모두 수행 가능합니다.

📘 데이터 불러오기

```
load_data("boston_housing")
```

📘 데이터 설명

보스턴 인근 지역의 사회·환경·교육·교통 특성을 바탕으로 주택의 **중위 가격(MEDV)** 을 설명하기 위해 수집된 대표적인 회귀 분석용 데이터이다.

- 관측치: 약 500개

변수	설명
MEDV	주택 중위가격 (종속변수, $1,000 단위)
RM	주택당 평균 방 개수
LSTAT	저소득층 비율
CRIM	1인당 범죄율
PTRATIO	학생-교사 비율
NOX	일산화질소 농도
DIS	직업 중심지까지의 거리
CHAS	찰스강 인접 여부 (0/1)
TAX	재산세율
RAD	방사형 고속도로 접근성 지수
AGE	1940년 이전 건축 비율

📘 과제 수행 미션

📝 미션 1. “이 데이터는 믿을 만할까?”

- 결측·이상치·편향을 점검하고, 처리 기준을 제시한다.
- MEDV가 0이거나 극단값인지, RM/CRIM/LSTAT 등 주요 변수의 범위가 현실적인지 확인한다.
- 전처리 전·후가 어떻게 달라졌는지 한눈에 비교하는 표나 요약을 만든다.
- 단위·해석 주의: `MEDV`는 $1,000 단위입니다. 극단값 제거/변환이 해석에 주는 영향까지 짧게 기록하세요.

> 📌 **출제 의도** “이 값이 말이 되나?”를 먼저 묻고, 어떻게 처리했는지를 기록해 **나중 해석에 근거를 남기는 연습**입니다.

📝 미션 2. “가격과 핵심 변수의 첫인상”

- MEDV, RM, LSTAT, CRIM 분포를 히스토그램/KDE로 확인하고 알 수 있는 객관적 사실을 서술한다.
- 왜도/이상치가 회귀에 줄 수 있는 영향과 변환할 필요가 있는지 서술하시오.
- 분포 비교는 동일 축 스케일로 제시하고, 평균/중앙값/꼬리의 차이를 문장으로 요약하세요.

> 📌 **출제 의도** 목표변수·핵심 변수의 생김새를 먼저 읽고 **변환 필요성을 스스로 판단**하게 합니다. 숫자보다 해석 문장이 중요합니다.

📝 미션 3. “로그/비선형 변환을 고민해 보자”

- MEDV 혹은 주요 변수(RM, LSTAT, CRIM)에 로그/제곱근 등 변환을 적용해 전후 분포를 **나란히** 비교한다.
- 변환이 해석과 모델 적합에 주는 장단점, 해석이 어떻게 달라지는지 예상한다.
- “이 변환이 없으면 어떤 함정에 빠질까?”를 한 줄로 정리한다.
- 선택 기준을 명시하세요: 왜 `log(MEDV)`인지, 왜 `CRIM`에 제곱근인지 등 데이터 분포 근거로 설명합니다.
- 필요한 경우 Box-Cox 등의 변환을 비교해 가장 해석 친화적인 옵션을 선택하세요.

> 📌 **출제 의도** 단순히 변환을 쓰는 것이 아니라 **“왜 필요했는가, 해석이 어떻게 달라지는가”**를 설명하는 연습입니다.

📝 미션 4. “지역·환경 요인의 패턴 읽기”

- CHAS(강 인접), RAD(교통 접근성) 등 범주/이산 변수별 MEDV 분포를 시각화한다.
- 중앙값·분포 겹침을 근거로 “어떤 그룹이 비싸/겹쳐”를 문장으로 적는다.
- “왜 이런 서열/겹침이 생겼을까?”를 상식·데이터 근거로 추정한다.
- 범주 수가 많은 변수(RAD)는 그룹 묶음(예: 저·중·고 접근성) 시도 후 서열/겹침을 함께 보고하세요.

> 📌 **출제 의도** 범주형 요인이 가격을 **어디서 가르고 어디서 겹치는지**를 이야기로 풀어내게 합니다.

📝 미션 5. “강 인접 여부가 가격을 좌우할까?”

- CHAS(강 인접 여부)에 따라 MEDV가 다른지 시각화하고, 두 집단 평균 차이를 가설검정(예: t-test)으로 확인한다.
- 효과 크기(차이의 크기)를 함께 제시하고, “실제로 의미 있는 차이인가?”를 말로 해석한다.
- 정규성/등분산 가정 점검 후 필요 시 Welch’s t-test나 비모수 검정을 선택하세요.
- 효과 크기는 `Cohen’s d` 또는 `Hedges’ g`로 제시하고, 실무적 의미를 한 줄로 번역합니다.

> 📌 **출제 의도** 두 집단 비교에서 **방법 선택과 효과 크기 해석**을 연습하고, 숫자를 의미로 번역하게 합니다.

📝 미션 6. “교통 접근성 등급은 가격을 얼마나 가를까?”

- RAD(방사형 고속도로 접근성) 등급별 MEDV 차이를 시각화한다.
- 분산분석(ANOVA)로 전체 차이를 확인하고, 차이가 있다면 사후검정으로 어느 등급 사이에서 차이가 나는지 정리한다.
- 사후검정은 Tukey HSD 또는 Games-Howell(등분산 위반 시)을 사용하고, “가격 서열표” 형태로 가독성 있게 요약하세요.

> 📌 **출제 의도** 여러 범주를 **전체→어디서 차이** 순서로 해석하며, 결과를 서열/지도처럼 정리하는 훈련입니다.

📝 미션 7. “변수들은 서로 섞여 있을까?”

- 주요 연속형 변수 간 상관행렬/히트맵을 보고, Variance Inflation Factor(VIF)로 다중공선성을 점검한다.
- RM×LSTAT처럼 상호작용을 넣으면 해석이 달라질지 아이디어를 적고, 필요한 경우 간단히 시도해 본다.
- 공선성 완화 전략을 비교하세요: 변수 제거, 결합지표 설계, 정규화 회귀(Lasso/Ridge) 중 해석 우선 관점에서 선택합니다.

> 📌 **출제 의도 중복 정보와 상호작용**을 의식하며, 회귀에 넣을 변수를 더 깔끔히 설계하게 합니다.

📝 미션 8. “어떤 연속형 변수가 가격과 가장 가깝나?”

- MEDV와 RM, LSTAT, PTRATIO, CRIM, DIS, NOX 등 연속형 변수의 상관을 계산한다.
- Pearson vs Spearman 중 무엇이 더 적합한지 이유를 적고, 결과를 물리·사회적 언어로 해석한다.
- 상관이 높다고 바로 넣을지, 변환이 필요한지 간단히 논의한다.
- 비선형·이상치 영향 시 Spearman을 병행 제시하고, 선택 이유를 데이터 근거로 명시합니다.

> 📌 **출제 의도** 상관계수 선택 이유와 **수치→의미** 해석을 연습하며, 변수 설계의 근거를 쌓게 합니다.

📝 미션 9. “가격을 설명하는 회귀모형 설계”

- MEDV(또는 변환값)를 종속변수로 하는 다중선형회귀를 설계한다.
- 변수 선택·인코딩·스케일/변환 이유를 “**설명하고 싶어서**” 관점으로 글로 남긴다.
- 학습/검증 데이터 분리 방식(예: train/test split)을 명시한다.

- 검증 전략을 구체화하세요: `train/test split`에 더해 `K-fold` 혹은 `shuffle split`로 평균 성능과 분산을 함께 보고합니다.
- 과적합·공선성 위험 완화를 위해 Ridge/Lasso/Elastic Net을 후보로 비교하고, "해석-성능" 균형 관점에서 선택 이유를 기록합니다.

> 📌 **출제 의도** 설계 선택이 임의가 아닌 **설명 의도**에 근거하도록 훈련합니다.

📝 미션 10. "회귀계수는 무엇을 말해주나?"

- 계수(또는 표준화 계수)와 신뢰구간, 방향·크기를 해석한다.
- "방 1개 늘면 가격이 어떻게"처럼 **물리/사회적 의미**로 번역한다.
- 변환 변수가 있다면, 변환을 감안한 해석을 명확히 쓴다.
- 표준화 계수(베타)와 비표준화 계수를 병행 제시하고, 단위/변환을 고려한 해석 문장을 명확히 작성합니다.
- 상호작용항이 있을 경우, 조건부 효과(예: `RM`이 낮을 때 `LSTAT`의 영향)까지 사람 언어로 풀어 씁니다.

> 📌 **출제 의도** 숫자를 **가격 구조의 언어**로 바꾸는 연습입니다.

📝 미션 11. "모형 진단과 개선"

- 잔차 정규성/등분산/선형성, 영향력을 잔차플롯, Q-Q, Cook's distance 등으로 점검한다.
- 문제 지점과 개선 아이디어(변환, 변수 교체/제거, 강건 회귀 등)를 제안한다.
- R^2, RMSE/MAE 등 기본 지표를 보고하고, 이 숫자가 의미하는 바를 짧게 해석한다.
- 캘리브레이션(예측 vs. 실제) 산점도를 함께 제시하고, 체계적 과소/과대 예측 구간을 언급하세요.
- K-fold 평균 지표와 표준편차를 보고하여 안정성을 함께 판단합니다.

> 📌 **출제 의도** 점수보다 **가정·진단과 해석**을 통해 "얼마나 믿을 수 있는가"를 판단하게 합니다.

📝 미션 12. "같은 방 개수인데 왜 가격이 다를까?"

- RM이 같은 두 가상 주택을 설정하고, 다른 변수 차이로 가격 차이를 설명한다.
- 비전공자에게 이야기하듯, 모델이 설명하는 것과 못하는 것을 구분해 제시한다.
- 이야기의 흐름을 "데이터 관찰 → 회귀계수 → 가격 차이 설명" 순서로 연결한다.
- 예시 틀: "두 지역 모두 방은 6개지만, `LSTAT`가 낮고 `DIS`가 높으며 `NOX`가 낮은 A 지역은… 그래서 모델은 A의 가격을 더 높게 본다."
- 모델이 설명하지 못한 요인(건축 재료, 관리 상태, 리모델링 여부 등)의 가능성도 덧붙여 설득력을 높입니다.

> 📌 **출제 의도** 지금까지의 분석을 **스토리로 엮어** 설득력 있게 전달하는 마무리입니다. 모델이 말하는 것과 말하지 못하는 것을 구분해 주세요.

📘 미션별 채점 기준 (총 100점)

미션	배점	평가 포인트	감점 기준

1. 데이터 신뢰도	8	결측·이상치·편향 확인, 처리 기준·영향 서술	결측/이상치 언급 누락(-3), 처리 기준·영향 설명 없음(-2)
2. 분포 첫인상	6	MEDV/RM/LSTAT/CRIM 분포 해석, 변환 필요성 논의	분포 시각화 부재(-2), 변환 필요성 언급 없음(-2), 해석 없이 그래프만(-1)
3. 변환 비교	6	변환 전후 비교, 변환 이유와 해석 변화 설명	변환 이유 미설명(-2), 전후 비교/해석 없음(-3)
4. 지역·환경 요인	8	CHAS/RAD 등 범주 변수별 패턴·겹침 해석	범주별 시각화 없이 수치만(-2), 패턴/겹침 해석 없음(-3)
5. 강 인접 검정	7	두 집단 비교, 검정 방법·효과 크기 제시, 의미 해석	검정/효과 크기 미제시(-2), 해석 없음(-2), 시각화 부재(-1)
6. RAD 등급 비교	7	ANOVA+사후검정 수행, 어느 등급 차이인지 정리	ANOVA만 하고 사후검정 없음(-2), 차이 정리 미흡(-2)
7. 상호작용/다중공선성	7	상관/VIF 점검, 상호작용 아이디어 및 영향 설명	VIF/상관 없이 주장만(-3), 상호작용 논의 없음(-2)
8. 연속형 상관	7	상관계수 선택 근거, 물리·사회적 해석, 변환 필요 논의	계수 선택 이유 미제시(-2), 해석 없이 수치만(-2), 변환 논의 없음(-1)
9. 회귀 설계	14	변수 선택·인코딩·스케일/변환·검증 전략 근거, 설명 의도 명시	설계 이유 미기재(-3), 스케일/변환 설명 없음(-3), 검증 전략 누락(-3)
10. 계수 해석	12	방향·크기·신뢰구간 해석, 표준화·상호작용 포함, 물리/사회적 의미로 번역	계수 해석 없음(-4), 의미 번역 부재(-3), 변환/상호작용 해석 누락(-2)
11. 모형 진단·성능	8	잔차/가정/영향력·캘리브레이션 점검, K-fold 평균+분산, 기본 지표 보고·해석, 개선안	진단 그래프/지표 누락(-3), 안정성/캘리브레이션 언급 없음(-2), 개선안 부재(-2)
12. 스토리텔링	12	가상 사례 설명, 모델이 설명하는/못하는 요인 구분, 비전공자 친화	사례 미제시(-3), 가격 차이 이유 해석 없음(-3), 한계 언급 없음(-2)