📚🏠 California Housing 회귀 분석 캡스톤

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

우리가 집을 살 때 보는 가격은 정말로 “공정한” 가격일까?

1990년 캘리포니아 인구조사 기반 데이터를 보면, 같은 지역이어도 가구 소득에 따라 집값이 얼마나 달라질까? 집의 나이, 방의 수, 인구밀도는 정말 중요할까?

이번 캡스톤 과제에서는 **California Housing 데이터셋**을 활용하여 지리적 위치·가구 소득·주택 특성이 **주택 중위 가격을 어떤 구조로 설명하는지**를 데이터와 회귀모형을 통해 단계적으로 탐구합니다.

이 과제의 목표는 예측 정확도를 높이는 것이 아니라, **“캘리포니아 주택 가격은 어떤 논리와 불평등으로 형성되는가”를 수치와 언어로 설명하는 것**입니다.

※ 본 과제는 팀 / 개인 단위 모두 수행 가능합니다.

📘 데이터 불러오기

```
load_data("california_housing")
```

📘 데이터 설명

캘리포니아 각 지역의 1990년 인구조사를 기반으로 수집된 주택 특성 데이터입니다. **목표는 명목 주택 중위 가격을 설명하는 구조를 이해하는 것**입니다.

- 관측치: 약 20,640개 지역(블록 그룹)

변수	설명
MedHouseValue	주택 중위 가격 (종속변수, $1 단위)
MedInc	중위 가구 소득 (10,000 달러 단위)
HouseAge	주택 나이 (년, 1940년 이전은 52로 기록)
AveRooms	주택당 평균 방 개수
AveBedrms	주택당 평균 침실 개수
Population	지역 주민 수
AveOccup	평균 가구 인원 수
Latitude	지역의 위도
Longitude	지역의 경도

📘 과제 수행 미션

📝 미션 1. “이 데이터는 믿을 만할까?”

- 결측·이상치·편향을 점검하고, 처리 기준을 제시한다.
- MedHouseValue가 0이거나 극단값인지, MedInc/HouseAge/AveRooms 등 주요 변수의 범위가 현실적인지 확인한다.
- 1940년 이전 건축 주택이 52로 코딩되는 이유를 이해하고, 분석에 미칠 영향을 검토한다.
- 전처리 전·후가 어떻게 달라졌는지 한눈에 비교하는 표나 요약을 만든다.
- 단위·해석 주의: `MedHouseValue`는 실제 달러 단위입니다. 1990년 명목가를 참고해 현실성을 평가하세요.

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

📝 미션 2. “가격과 핵심 변수의 첫인상”

- MedHouseValue, MedInc, HouseAge, AveRooms 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.
- 지역별(Latitude/Longitude) 편향이나 시계열 문제(1990년 데이터만)가 해석에 미치는 영향도 짧게 언급하세요.

> 📌 **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”

- MedHouseValue 혹은 주요 변수(MedInc, Population, HouseAge)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(MedHouseValue)`인지, 왜 Population에 로그인지 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. “지리적 위치는 정말 중요할까?”

- 위도(Latitude)와 경도(Longitude)를 활용한 지리적 시각화를 수행한다.
- 산점도(경도 × 위도, 점의 색상 = MedHouseValue)를 그려 주택 가격의 지리적 불평등을 관찰한다.
- “어느 지역이 비싼가?”, “가격 차이가 얼마나 뚜렷한가?”를 데이터 기반으로 설명하세요.
- 위도/경도가 개별 변수보다 **지역 표상(Proxy)**이라는 점을 인식하고, 이것이 회귀에 주는 의미를 논의하세요.

> 📌 **출제 의도** 지리적 위치 데이터가 어떤 정보를 담고 있고, 회귀에서 어떤 역할을 하는지 체감하게 합니다.

📝 미션 5. “소득은 정말로 집값을 결정할까?”

- MedInc(중위 가구 소득)에 따라 MedHouseValue가 다른지 시각화하고, 선형성과 강도를 평가한다.
- 산점도와 함께 회귀선을 그려 관계의 형태를 관찰하세요.
- 상관계수(Pearson)를 계산하고, “정말로 소득이 집값의 주 결정 요인인가?”를 비판적으로 생각해 보자.
- 소득 수준에 따른 주택가격 불평등 문제를 해석 관점에서 논의하세요.

> 📌 **출제 의도** 개별 변수의 강한 관계를 먼저 인식하고, 이것이 모델에서 어떤 의미를 가지는지 생각하게 합니다.

📝 미션 6. “주택 특성은 소수일까, 다수일까?”

- AveRooms, AveBedrms, HouseAge, AveOccup, Population의 분포를 확인한다.
- 각 변수별로 생존자(정상) vs 이상치를 구분하고, 회귀에 미칠 영향을 평가하세요.
- 일부 극단값(예: 매우 작은 방의 수, 매우 높은 인구밀도)이 전체 분석을 왜곡할 가능성을 검토합니다.

> 📌 **출제 의도** 다양한 설명 변수의 생김새를 한눈에 이해하고, 각각이 가진 특성(분포, 이상치)을 인식하게 합니다.

📝 미션 7. “소득과 다른 특성은 독립적일까?”

- 주요 연속형 변수 간 상관행렬(MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, 위도, 경도)을 계산한다.
- 상관행렬을 히트맵으로 시각화하고, Variance Inflation Factor(VIF)로 다중공선성을 점검한다.
- 높은 상관이 보이는 변수 쌍(예: AveRooms vs AveBedrms)에 대해, 이것이 모델 설계에 주는 함의를 논의하세요.
- 공선성 완화 전략을 비교하세요: 변수 제거, 결합지표 설계, 정규화 회귀(Ridge/Lasso) 중 해석 우선 관점에서 선택합니다.

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

📝 미션 8. “어떤 변수가 가격과 가장 가까울까?”

- MedHouseValue와 MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude의 상관을 계산한다.
- Pearson과 Spearman을 모두 계산하고, 비교해 무엇이 더 적합한지 이유를 적으세요.

- 각 변수의 관계를 물리·사회적 언어로 해석하고(예: “소득이 높을수록 비싼 주택 지역”, “오래된 주택은 더 저렴”), 인과성을 말할 수 있는지 비판적으로 평가합니다.

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

📝 미션 9. “상관관계는 곧 원인일까?”

- 지금까지의 탐색 결과를 바탕으로 “상관이 높다 = 원인이다”라는 착각을 피하자.
- 예: 위도/경도와 주택가격의 높은 상관이 정말 위치 자체가 가격을 결정하는가, 아니면 위치가 **소득 수준이 높은 지역을 나타내는 대리변수**인가?
- 데이터 수집 시점(1990년)·지역 편향·측정 누락(학군 질, 교통 접근성, 범죄율 등)이 결론에 주는 한계를 서술하세요.

> 📌 **출제 의도** 데이터가 말해 주는 것과 말하지 못하는 것을 구분하며, 신중한 해석의 중요성을 체감하게 합니다.

📝 미션 10. “주택 가격을 설명하는 회귀모형 설계”

- MedHouseValue(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 80/20)을 명시한다.
- 검증 전략을 구체화하세요: `train/test split`에 더해 `K-fold`(예: 5-fold)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 11. “회귀계수는 무엇을 말해주나?”

- 계수(또는 표준화 계수)와 신뢰구간, 방향·크기를 해석한다.
- “소득이 1만 달러 늘면 주택 가격이 어떻게 바뀌는가”, “주택이 1년 오래될 때마다 가격이 얼마나 떨어지는가” 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.
- 위도/경도 계수가 통계적으로 유의하다면, 이것이 **지역의 숨은 특성**을 어떻게 대리하는지 해석하세요.

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

📝 미션 12. “모형 진단과 개선”

- 잔차 정규성/등분산/선형성, 영향력을 잔차플롯, Q-Q, Cook’s distance 등으로 점검한다.
- 문제 지점(예: 특정 가격대에서 체계적 오차, 극단값 지역)과 개선 아이디어(변환, 변수 교체/제거, 강건 회귀 등)를 제안한다.

- R^2, RMSE/MAE 등 기본 지표를 보고하고, 1990년 명목가 기준에서 이 오차가 얼마나 의미 있는지 해석하세요.
- 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 13. "같은 소득인데 왜 가격은 다를까?"

- MedInc이 같은 두 가상의 지역을 설정하고(예: A 지역은 부유한 해안, B 지역은 외곽), 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 중위 소득이 $50,000이지만, A 지역은 위도가 높고(해안) 주택이 신축이며 인구밀도가 낮아서… 그래서 모델은 A의 가격을 더 높게 본다. 하지만 이 모델은 학군이나 직장까지의 거리 같은 정보를 담지 못한다."
- 모델이 설명하지 못한 요인(학군 질, 교통 접근성, 범죄율, 경제 정책 변화 등)의 가능성도 덧붙여 설득력을 높입니다.

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

📘 채점 기준 (총 100점)

미션	배점	평가 포인트	감점 기준
1. 데이터 신뢰도	7	결측·이상치·편향·단위 확인, 처리 기준·영향 서술	결측/이상치 언급 누락(-2), 처리 기준·영향 설명 없음(-2), 단위/시점 고려 부족(-1)
2. 분포 첫인상	6	목표·핵심 변수 분포 해석, 변환 필요성·시간적 편향 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. 지리적 위치 분석	8	지도 시각화, 지리적 불평등 패턴, 대리변수 개념 이해	시각화 부재(-3), 지리적 패턴 해석 없음(-2), 대리변수 논의 부재(-1)
5. 소득-가격 관계	7	소득별 분포·산점도, 선형성 평가, 불평등 인식	시각화/상관 미제시(-2), 선형성 평가 없음(-2), 불평등 논의 부재(-1)
6. 주택 특성 분포	7	다양한 설명 변수 분포·이상치 확인, 회귀 영향 평가	변수별 분포 누락(-2), 이상치 언급 없음(-2), 영향 평가 부재(-1)
7. 다중공선성·상관	7	상관/VIF 점검, 상호작용 아이디어 및 영향 설명, 정규화 비교	VIF/상관 없이 주장만(-2), 정규화 비교 없음(-2), 상호작용 논의 부재(-1)
8. 연속형 상관	7	Pearson/Spearman 계산·선택 근거, 물리·사회적 해석, 인과성 비판	계수 선택 이유 미제시(-2), 해석 없이 수치만(-2), 인과성 비판 부재(-1)
9. 상관 vs 인과	6	상관의 한계, 대리변수·누락 변수·시점 편향 명시	한계/누락 요인 언급 없음(-2), 상관=원인 단정(-2), 시점 고려 부족(-1)
10. 회귀 설계	13	변수 선택·인코딩·스케일/변환·검증·정규화 전략 근거, 설명 의도 명시	설계 이유 미기재(-3), 검증 전략 부재(-2), 정규화 비교 미흡(-2)
11. 계수 해석	12	비표준화/표준화 계수, 신뢰구간, 물리/사회적 의미 번역, 지역 대리 설명	계수 해석 없음(-3), 의미 번역 부재(-2), 지리적 해석 누락(-2)
12. 모형 진단·성능	8	잔차/가정/영향력·캘리브레이션, K-fold 평균+분산, 1990년 가격 기준 오차 의미, 개선안	진단 그래프 누락(-2), 안정성/캘리브레이션 언급 없음(-1), 실무 의미 평가 부재(-1), 개선안 없음(-1)
13. 스토리텔링	12	가상 사례 설명, 모델 설명/미설명 요인 구분, 비전공자 친화, 사회적 맥락	사례 미제시(-3), 가격 차이 이유 해석 없음(-3), 모델 한계 언급 없음(-2), 맥락 부재(-2)