📚 Diamonds 데이터 분석 캡스톤

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

본 과제는 Kaggle의 Diamonds 데이터셋을 활용하여 데이터 전처리부터 탐색적 데이터 분석, 통계적 추론, 회귀분석까지 데이터 분석의 전체 흐름을 경험하는 것을 목표로 하는 평가 과제 입니다.

아래 미션들은 “정답을 맞히는 과제”가 아니라, **데이터를 통해 하나의 이야기를 만들어 가는 탐구 과제**입니다.

각 미션은 서로 독립적이지만, 모두 합치면 하나의 질문으로 수렴합니다.

> **“다이아몬드 가격은 왜 이렇게 결정되는 걸까?”**

이 과제의 목적은 “회귀식 = 숫자”가 아니라 **“회귀식 = 설명 가능한 세계관”** 이라는 인식을 심어주는 데 목적이 있습니다.

※ 본 과제는 팀/개인 단위 모두 수행 가능합니다.

📘 데이터 불러오기

```
load_data("diamonds")
```

📘 데이터 설명

field	description
price	다이아몬드 가격 (USD, $326 ~ $18,823)
carat	중량 (0.2~5.01)
cut	컷 품질 (Fair, Good, Very Good, Premium, Ideal)
color	색상 등급 - J (worst) to D (best)
clarity	투명도 등급 (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	길이 mm (0~10.74)
y	너비 mm (0–58.9)
z	두께 mm (0–31.8)
depth	비율 정보 = z / mean(x, y) = 2 * z / (x + y) (43–79)
table	다이아몬드의 가장 넓은 지점에 비해 상단(테이블)의 너비 (43–95)

📘 과제 수행 미션

📝 미션 1. “이 데이터는 얼마나 믿을 수 있을까?”

본격적인 분석에 들어가기 전에, 우리가 사용할 데이터가 **얼마나 신뢰할 수 있는지**부터 점검해 보자.

데이터 안에는 - 비어 있는 값은 없는지, - 유난히 튀는 값은 없는지, - 현실적으로 말이 되지 않는 값은 없는지 차근차근 살펴볼 필요가 있다.

예를 들어, 다이아몬드의 크기를 나타내는 `x`, `y`, `z` 값이 0으로 기록된 경우가 있는지 직접 확인해 보자. 만약 있다면, 이런 값들을 그대로 분석에 사용해도 괜찮을지 팀에서 논의해 보자.

또한 `cut`, `color`, `clarity`와 같은 품질 변수들이 몇 개의 범주로 구성되어 있고, 특정 범주에 데이터가 지나치게 몰려 있지는 않은지도 함께 확인해 보자.

마지막으로, 전처리 전과 후를 비교하며 **어떤 기준으로 데이터를 제거하거나 수정했는지**를 정리해 보자.

> 📌 **출제 의도** 분석 전에 “이 값이 말이 되나?”를 먼저 묻는 습관을 들입니다. 결측·이상치 처리 기준을 말로 남겨야 이후 회귀 해석이 설득력을 가집니다.

📝 미션 2. “가격 데이터는 어떤 모습일까?”

`price` 변수를 처음 마주했을 때의 분포를 직접 그려보자. 가격이 고르게 분포되어 있는지, 아니면 특정 구간에 몰려 있는지 관찰해 보자.

이어서 `carat` 변수의 분포도 함께 살펴보고, 두 변수의 공통점과 차이점을 말로 설명해 보자.

분포를 살펴본 뒤에는 “이 상태로 회귀분석을 해도 괜찮을까?”라는 질문을 던져보자.

> 📌 **출제 의도** 목표 변수와 핵심 설명 변수의 생김새를 먼저 확인해, 지표 선택과 변환(로그 등) 필요성을 스스로 느끼게 하기 위함입니다.

📝 미션 3. “로그 변환은 왜 등장할까?”

`price`와 `carat`에 로그 변환을 적용해 보고, 변환 전과 후의 분포를 나란히 비교해 보자.

로그 변환을 하면 무엇이 달라졌는지, 그리고 왜 많은 분석에서 로그 변환을 사용하는지 이번 데이터셋을 기준으로 설명해 보자.

이때 중요한 것은 “로그를 썼다”가 아니라 **“왜 써야 했는가”**이다.

> 📌 **출제 의도** 단순한 기법 나열이 아니라 “왜 이 변환이 필요한가”를 설명하는 연습입니다. 분포와 해석이 어떻게 달라지는지 체감해 보세요.

📝 미션 4. “품질 등급은 가격을 얼마나 설명해 줄까?”

`cut`, `color`, `clarity`에 따라 다이아몬드 가격 분포가 어떻게 달라지는지 시각화해 보자.

중앙값, 분산, 분포의 겹침 정도를 관찰하며 다음 질문에 답해 보자.

- 어떤 품질 변수는 가격 차이가 뚜렷한가?
- 어떤 경우에는 등급이 달라도 가격이 크게 겹치는가?

숫자보다 **패턴과 느낌**을 중심으로 정리해 보자.

> 📌 **출제 의도** 그래프를 보고 “어떤 등급이 가격을 가르는지, 어디서 겹치는지”를 말로 설명하는 훈련입니다. 수치보다 패턴 읽기가 우선입니다.

📝 미션 5. “Premium 컷은 정말 더 비쌀까?”

다음과 같은 질문을 통계적으로 검증해 보자.

> Premium 컷 다이아몬드는 Ideal 컷보다 평균 가격이 높을까?

이를 위해 - 가설을 직접 세우고 - 어떤 검정 방법이 적절한지 고민한 뒤 - 실제로 검정을 수행해 보자.

검정 결과를 해석할 때는 “유의하다 / 유의하지 않다”에서 멈추지 말고, **이 차이가 얼마나 의미 있는 차이인지**까지 함께 생각해 보자.

> 📌 **출제 의도** 가설검정을 “있다/없다”로 끝내지 않고, 차이의 크기와 의미를 함께 이야기하는 연습입니다. 방법 선택 이유도 명확히 해보세요.

📝 미션 6. “모든 컷은 서로 다를까?”

이번에는 컷 등급 전체를 놓고 생각해 보자.

`cut`에 따른 가격 차이가 전반적으로 존재하는지 분산분석으로 확인해 보고, 차이가 있다면 **어떤 컷들 사이에서 차이가 발생하는지** 사후검정을 통해 살펴보자.

이 결과를 “가격 서열표”처럼 정리해 보는 것도 좋다.

> 📌 **출제 의도** 전체 차이 확인(ANOVA)에서 그치지 않고, 어떤 조합에서 차이가 나는지 구체화하는 연습입니다. 결과를 서열/지도처럼 정리해 보세요.

📝 미션 7. “품질 요인들은 서로 영향을 주고받을까?”

컷(cut)과 색(color), 혹은 컷과 투명도(clarity)를 함께 고려하면 가격 구조가 달라질까?

이원 분산분석을 통해 - 각각의 요인이 가격에 미치는 영향 - 두 요인이 함께 작용할 때의 효과 를 구분해 보자.

결과를 해석할 때는 “상호작용이 있다/없다”를 넘어 **그 의미가 무엇인지**를 설명해 보자.

> 📌 **출제 의도** 요인들이 단독으로만 작용하지 않을 수 있음을 체험시키려는 단계입니다. 상호작용이 의미하는 바를 말로 풀어내는 것이 핵심입니다.

📝 미션 8. “가격과 가장 가까운 물리적 변수는?”

연속형 변수들 사이의 상관관계를 계산해 보자.

`price`와 `carat`, `x`, `y`, `z`, `depth`, `table` 중 어떤 변수가 가장 강한 관계를 보이는지 확인해 보고, 왜 그런 결과가 나왔는지 팀의 언어로 설명해 보자.

이 과정에서 Pearson과 Spearman 중 어떤 상관계수가 더 적절한지도 함께 고민해 보자.

> 📌 **출제 의도** 연속형 변수 관계를 요약할 때 어떤 계수를 쓰고, 왜 그 선택이 타당한지 설명하는 연습입니다. 물리적 의미와 수치가 만나는 지점을 찾아보세요.

📝 미션 9. “상관관계는 곧 원인일까?”

상관분석 결과를 바탕으로 다음 질문에 답해 보자.

- 상관이 높다는 것은 무엇을 의미하는가?
- 이 결과를 그대로 “원인”이라고 말해도 될까?

분석 결과의 **한계**를 스스로 짚어보는 것이 이번 미션의 핵심이다.

> 📌 **출제 의도** “상관이 높다 = 원인이다”라는 착각을 피하도록, 데이터가 말해 주지 않는 부분과 한계를 스스로 적어 보게 합니다.

📝 미션 10. “가격을 설명하는 회귀모형을 만들어보자”

이제 지금까지 살펴본 변수들을 활용해 `log(price)`를 종속변수로 하는 다중선형회귀모형을 만들어보자.

변수를 선택할 때는 “넣을 수 있어서”가 아니라 **“설명하고 싶어서”** 선택했는지 스스로 점검해 보자.

회귀계수 하나하나가 가격 구조를 어떻게 설명해 주는지 말로 풀어 써보자.

> 📌 **출제 의도** 변수 선택과 모델 설계를 “설명하고 싶은 세계”에 맞춰보는 연습입니다. 계수를 단순 숫자가 아닌 가격 구조의 언어로 바꾸는 것이 목표입니다.

📝 미션 11. “이 회귀모형은 믿을 만할까?”

회귀분석 결과를 그대로 받아들이기 전에, 가정이 얼마나 잘 지켜졌는지 확인해 보자.

잔차 그림과 진단 지표를 통해 - 어떤 가정이 잘 지켜졌는지 - 어떤 부분이 아쉬운지 를 정리해 보자.

그리고 이 모형을 더 개선하려면 무엇을 해볼 수 있을지도 함께 제안해 보자.

> 📌 **출제 의도** 성능만 보는 것이 아니라, 회귀 가정과 진단을 통해 “이 모델을 어디까지 믿을 수 있는가”를 점검하고 개선 아이디어를 제시하게 합니다.

📝 미션 12. “같은 carat, 왜 가격은 다를까?”

마지막으로, 지금까지의 분석을 모두 모아 하나의 질문에 답해 보자.

> 같은 carat을 가진 두 다이아몬드의 가격이 다른 이유는 무엇일까?

가상의 사례를 하나 만들어, 통계 분석과 회귀 결과를 활용해 **비전공자에게 설명하듯** 풀어 써보자.

이 미션은 지금까지의 모든 분석을 하나의 이야기로 엮는 과정이다.

이 미션은 숫자가 아니라 **이야기와 설득력**을 평가합니다.

> 📌 **출제 의도** 지금까지의 수치를 사람의 언어로 엮어 설명하는 마무리입니다. 모델이 말하는 것과 말하지 못하는 것을 구분해 설득력 있는 이야기를 만들어 보세요.

📘 채점 기준 (총 100점)

미션	배점	평가 포인트	감점 기준
1. 데이터 신뢰도	8	결측·이상치·편향 확인, 처리 기준과 영향 서술	결측/이상치 언급 누락(-3), 처리 기준·영향 설명 없음(-2)
2. 가격/무게 분포	6	`price`/`carat` 분포 해석, 회귀 전 변환 필요성 언급	분포 시각화 부재(-2), 변환 필요성 논의 없음(-2), 해석 없이 그래프만 제시(-1)
3. 로그 변환	6	변환 전후 비교, 로그 사용 이유와 해석 변화 설명	변환 이유 미설명(-2), 전후 비교/해석 없음(-3)
4. 품질 등급 분포	8	cut/color/clarity별 패턴·겹침을 그래프로 읽어 설명	등급별 해석 없이 그래프만 제시(-3), 겹침/차이 언급 없음(-2)
5. Premium vs Ideal	7	가설 설정, 검정 방법 근거, 효과 크기·의미 해석	가설·방법 근거 누락(-2), 효과 크기/의미 해석 없음(-2), 검정 미실시(-2)
6. 컷 전체 비교	7	ANOVA+사후검정 수행, 어느 컷 간 차이인지 정리	ANOVA만 하고 사후검정 없음(-2), 차이 나는 컷 정리 미흡(-2)
7. 요인 상호작용	7	이원 분석·상호작용 결과를 의미 중심으로 해석	상호작용 해석 없음(-3), 결과를 수치만 제시(-2)
8. 연속형 상관	7	상관계수 선택 근거, 물리적 해석과 함께 제시	계수 선택 이유 미제시(-2), 물리적 해석 없이 수치만(-2), 잘못된 계수 선택(-2)
9. 상관 vs 인과	6	상관의 한계·누락 변수·해석 위험을 명시	한계/누락 요인 언급 없음(-3), 상관=원인으로 단정(-3)
10. 회귀모형 설계	14	변수 선택 논리, `log(price)` 계수 해석, 가격 구조 설명	선택/인코딩 논리 미기재(-3), 계수 해석 없음(-4), 가격 구조 설명 부재(-3)
11. 모형 진단	12	잔차·가정 점검, 문제 인식과 개선 아이디어 제시	잔차/가정 진단 누락(-4), 문제/개선안 언급 없음(-3)
12. 스토리텔링	12	사례 기반 설명, 비전공자 이해도, 모델 한계 포함	사례 미제시(-4), 가격 차이 이유 해석 없음(-3), 한계 언급 없음(-2)