다이아몬드 품질 분석 보고서

2021190002 장서현

[서론]

본 보고서의 목표는 주어진 다이아몬드 데이터셋을 활용하여 다이아몬드의 품질 등급(cut)에 따른 다이아몬드의 중량(carat)과 가격(price)의 유의미한 차이를 검증하는 것에 있다. 이를 통해 다이아몬드의 품질 등급이 다이아몬드의 중량과 가격에 어떠한 영향을 미치는지 확인하고자 한다. 다이아몬드의 품질 등급에 따라 그의 중량과 가격이 유의미하게 다를 것으로 예상되며, 품질 등급이 높을수록 더 무거운 중량과 비싼 가격을 가질 것으로 예상된다.

본 보고서는 총 네 가지의 통계검정법을 활용하여 위 내용을 분석하고 검증하고자 한다.

1. 정규성 검정: 각 등급의 데이터가 정규 분포를 따르는지 확인한다.
2. 등분산 검정: 각 등급의 분산이 동일한지 검정한다.
3. One-way ANOVA: 품질 등급에 따른 중량과 가격 각각의 평균 차이를 검정한다.
4. 사후 검정: 평균의 차이가 유의미할 경우, 구체적으로 어떤 등급 간에 차이가 있는지 확인한다.

[본론]

- 데이터셋 설명 및 기본 통계량

본 데이터셋은 403개의 다이아몬드 샘플로 구성되어 있으며, 각 샘플의 다이아몬드의 품질 등급, 중량, 가격, 색 등에 관한 정보를 제공한다. 그 중, 우리의 분석 대상이 되는 다이아몬드의 품질 등급은 "Good", "Ideal", "Premium"으로 나뉜다.

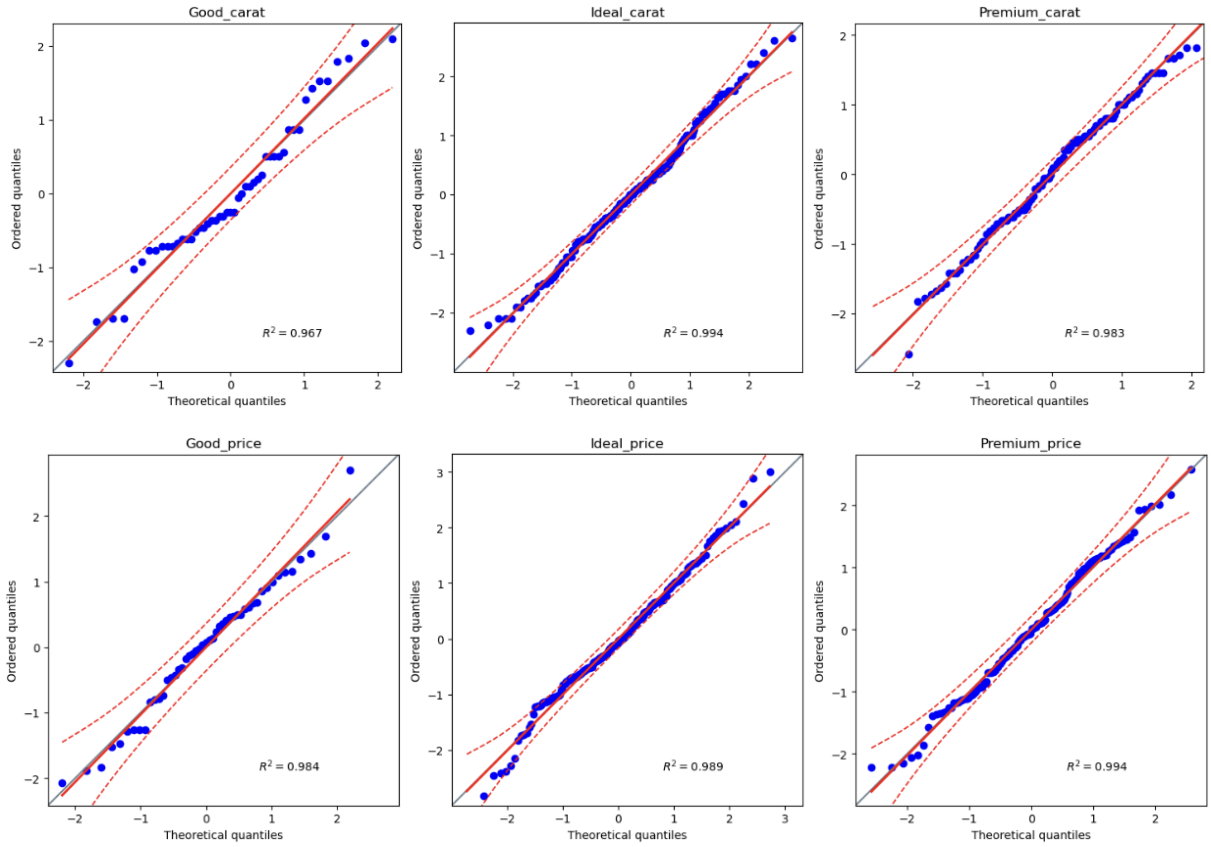
Cut	Carat (Mean ± Std)	Price (Mean ± Std)
Good	1.71 ± 0.20	4162.52 ± 4540.55
Ideal	1.50 ± 0.20	3237.53 ± 4519.30
Premium	1.72 ± 0.20	4542.46 ± 3834.93

다이아몬드의 품질 등급에 따른 중량과 가격의 평균은 위 표에 제시되어 있다. "Good"과 "Premium" 등급 사이의 중량의 평균 차이가 미세해 보여, 평균값 사이의 유의미한 차이가 있을지 더 자세히 검증해볼 필요성이 있다.

- 정규성 검정

데이터가 정규 분포를 따를 때 통계 검정의 신뢰도가 올라가기에, 각 품질 등급에 따라 나뉜 데이터셋이 각자 정규분포를 따르는지 확인하고자 한다.

우선, QQ-plot을 통해 각 품질 등급의 중량과 가격이 정규 분포를 따르는지 시각적으로 분석하였다. 그 결과로, 모든 다이아몬드 등급에서 데이터가 대각선에 가깝게 분포하였기에 정규성을 만족한다고 판단하였다.



다음으로, 더 정확한 수치를 기반으로 정규성을 판단하기 위해 Shapiro-Wilk Test를 진행하였다. 모든 경우의 수의 p-value가 0.05보다 크므로, 각 집단은 정규성을 만족하고 있다.

Cut	Variable	W-statistic	p-value
Good	Carat	0.965	0.148
Good	Price	0.984	0.739
Ideal	Carat	0.993	0.348
Ideal	Price	0.991	0.175
Premium	Carat	0.982	0.071
Premium	Price	0.992	0.586

- 등분산 검정

각 비교집단이 동일한 분산을 가지는 것은 정확한 선형 회귀 분석을 위한 또 다른 가정이다. 분산이 다를 경우, ANOVA 분석 결과의 신뢰성이 떨어질 위험 또한 있다. 따라서, Bartlett's test를 진행하여 해당 집단들이 등분산을 가지는지 검정하였다.

1. 중량에 대한 등분산 검정 가설:

귀무가설: 다이아몬드 품질 등급에 따른 다이아몬드의 중량 데이터의 분산이 동일하다.

대립가설: 다이아몬드 품질 등급에 따른 다이아몬드의 중량 데이터의 분산이 동일하지 않다.

2. 가격에 대한 등분산 검정 가설:

귀무가설: 다이아몬드 품질 등급에 따른 다이아몬드의 가격 데이터의 분산이 동일하다.

대립가설: 다이아몬드 품질 등급에 따른 다이아몬드의 가격 데이터의 분산이 동일하지 않다.

Bartlett's test의 결과는 다음과 같다. 중량과 가격, 두 경우에 모두 p-value가 0.05보다 크므로, 세 집단 사이의 모분산에 유의미한 차이가 없다고 판단되며, 등분산성 가정이 유지된다.

Variable	Bartlett's Statistic	p-value
Carat	0.018	0.991
Price	4.702	0.093

- One-way ANOVA

선형회귀 분석을 정확하기 진행하기 위한 두 가설이 만족되었기에, One-way ANOVA를 통해 다이아몬드의 품질 등급에 따른 중량과 가격의 평균을 비교하고 품질 등급에 따른 중량과 가격의 차이가 유의미한지 검정하고자 한다.

1. 중량에 대한 One-way ANOVA 가설:

귀무가설: 품질 등급에 따른 다이아몬드의 중량의 평균값에 차이가 없다. 즉, "Good", "Ideal", "Premium" 등급 사이의 다이아몬드 중량의 평균이 모두 동일하다.

대립가설: 품질 등급(cut)에 따른 다이아몬드의 중량의 평균값에 차이가 있다. 즉, 적어도 하나의 등급의 다이아몬드 중량 평균이 다른 등급과 다르다.

2. 가격에 대한 One-way ANOVA 가설:

귀무가설: 품질 등급에 따른 다이아몬드의 가격의 평균값에 차이가 없다. 즉, "Good", "Ideal", "Premium" 등급 사이의 다이아몬드 가격의 평균이 모두 동일하다.

대립가설: 품질 등급(cut)에 따른 다이아몬드의 가격의 평균값에 차이가 있다. 즉, 적어도 하나의 등급의 다이아몬드 가격 평균이 다른 등급과 다르다.

One-way ANOVA의 결과는 다음과 같다.

Variable	F-statistic	p-value
Carat	60.086	<0.001
Price	4.079	0.0176

두 경우 모두 p-value가 0.05보다 작기에, 귀무가설을 기각한다. 즉, 다이아몬드 중량과 가격의 평균값 중에서 적어도 하나의 등급과는 유의미한 차이가 있는 결론에 다다른다.

- 사후검정

ANOVA에서 평균값 사이 유의미한 차이를 발견함에 따라, 구체적으로 어떤 그룹 간에 차이가 있는지 확인하고자 Tukey's HSD Test를 통해 사후 검정을 시행한다. 사후 검정의 결과는 아래와 같다.

1. Tukey's HSD Test (carat)

Comparison	Mean Difference	p-value	Reject
Good vs Ideal	-0.2105	0.0	True
Good vs Premium	0.0103	0.9481	False
Ideal vs Premium	0.2208	0.0	True

- Good vs Ideal, Ideal vs Premium: p-value가 0.0으로 0.05보다 작으므로 평균값 사이 유의미한 차이가 있다.

- Good vs Premium: p-value가 0.948로 0.05에서 평균값의 유의미한 차이가 없다고 판단된다.

2. Tukey's HSD Test (price)

Comparison	Mean Difference	p-value	Reject
Good vs Ideal	-924.987	0.3633	False
Good vs Premium	379.9388	0.856	False
Ideal vs Premium	1304.9258	0.0155	True

- Ideal vs Premium: p-value가 0.0155으로 유의 수준 0.05에서 평균값의 유의미한 차이가 있다.

- Good vs Ideal, Good vs Premium: p-value가 각각 0.3633과 0.856으로 유의수준 0.05에서 평균값들의 유의미한 차이가 없다고 판단된다.

[결론]

중량의 경우, "Good"과 "Ideal", "Ideal"과 "Premium" 간의 중량에서 유의미한 차이가 있다. 이는 품질 등급에 따라 다이아몬드의 중량이 달라짐을 시사한다. "Ideal" 등급의 다이아몬드는 "Good" 등급의 다이아몬드보다 평균적으로 가벼운 경향이 있으며, "Premium" 등급의 다이아몬드보다도 가벼운 경향이 있다. 가격의 경우, "Ideal"과 "Premium" 간의 가격에서 유의미한 차이가 있다. "Ideal" 등급의 다이아몬드보다 "Premium" 등급의 다이아몬드가 평균적으로 더 비싼 경향이 있음을 드러낸다.