측광 데이터 기반
적색편이 예측 모델 구축

학번	2021212816
학과	컴퓨터공학과
학년	3학년
이름	김이현
제출일	2025.10.15

목차

1. 주제

1. 전체 변수간 상관관계 히트맵
   - 모든 숫자형 관측 변수들 간의 선형적 관계 분석
   - 적색편이와 강한 양의 상관관계를 가지는 데이터 분석
2. 천체 종류에 따른 적색편이 분포
   - 천체 종류별로 적색편이 값의 분포가 어떻게 다른지 비교
   - 퀘이사가 다른 천체 간의 적색편이 값 비교
3. 천체 종류에 따른 밝기 분포
   - 천체 종류별로 각기 다른 색 필터에서 측정된 밝기 분포를 비교
   - 자외선 필터와 적색 필터에서의 밝기 분포의 차이 분석
4. 적색편이에 따른 색 변화 분석
   - g-r 값을 색 지수로 정의하여 천체의 색을 수치화
   - 색 지수와 적색편이 값의 관계 분석

2. 데이터 설명

해당 데이터셋은 SDSS를 통해 관측된 천체들의 분광 및 측광 데이터입니다. 각 천체의 위치(alpha, delta), 여러 색 필터에서의 밝기(u, g, r, i, z), 관측 정보, 그리고 물리적으로 측정된 적색편이 값이 포함되어 있습니다.
천문학에서 천체의 거리는 우주의 구조를 이해하는 데 가장 중요한 변수 중 하나입니다. 이 데이터셋의 분석을 통해 여러 필터에서 관측된 밝기 정보가 실제 적색편이 값에 어떤 영향을 미치는지, 그리고 그중에서 적색편이를 예측하는 데 가장 큰 영향을 주는 요인이 무엇인지 통계적으로 검증하고자 합니다.
나아가, 단순히 개별 필터의 영향만 살펴보는 데 그치지 않고, 여러 색상 정보가 결합된 상황에서 예측 정확도가 어떻게 달라지는지 분석하고자 합니다. 예를 들어, u-g 색 지수, g-r 색 지수가 동시에 주어졌을 때, 또는 특정 천체 종류라는 조건이 추가되었을 때 적색편이 예측이 어떻게 변화하는지를 다각적으로 분석하고자 합니다.
최종적으로는 천체의 색상 정보가 적색편이로 얼마나 이어지는지를 정량화하는 '측광 적색편이(photometric redshift)' 예측 모델을 구축하는 것이 본 분석의 목표입니다. 이를 통해, 직접적인 분광 관측이 어려운 수많은 천체에 대해 효율적으로 거리를 추정할 수 있는 객관적인 기준을 제시하고, 대규모 천문 데이터 처리의 기반을 마련하고자 합니다.

데이터셋 기본정보

데이터 이름	Stellar Classification Dataset - SDSS17
데이터 링크	https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17
데이터 형식	정형 csv파일
레코드 수	100000개
컬럼 수	18개
비고	결측지 없음

각 컬럼에 대한 소개

컬럼명	데이터타입	컬럼정보
obj_ID	float64	천체를 구별하는 고유 번호
alpha	float64	천구 상의 경도(가로)를 나타내는 좌표값.
delta	float64	천구 상의 위도(세로)를 나타내는 좌표값.
u	float64	자외선 필터로 측정한 밝기 등급.
g	float64	녹색 필터로 측정한 밝기 등급.
r	float64	적색 필터로 측정한 밝기 등급.
i	float64	근적외선 필터로 측정한 밝기 등급.
z	float64	적외선 필터로 측정한 밝기 등급.
run_ID	int64	실행 번호. 특정 관측(스캔)을 식별하는 번호
rerun_ID	int64	재실행 번호. 이미지가 어떻게 처리되었는지를 명시하는 번호
cam_col	int64	카메라 열. 특정 관측에서 스캔 라인을 식별하는 번호
field_ID	int64	필드 번호. 관측된 하늘의 각 영역을 식별하는 번호
spec_obj_ID	float64	분광 객체 식별자. 분광 분석을 통해 얻은 천체의 고유 ID. 이 ID가 같으면 같은 종류의 천체임을 의미함.
class	object(str)	천체 종류. 'GALAXY'(은하), 'STAR'(별), 'QSO'(퀘이사) 중 하나로 분류
redshift	float64	적색편이. 빛의 파장이 길어지는 현상을 측정한 값, 천체가 우리로부터 얼마나 멀리 떨어져 있고 얼마나 빠르게 멀어지는지를 나타내는 핵심 지표
plate	int64	플레이트 ID. SDSS(관측 프로젝트)에서 빛을 모으는 데 사용된 각 플레이트를 식별하는 고유 번호
MJD	int64	수정 율리우스일(Modified Julian Date). 특정 데이터가 관측된 날짜
fiber_ID	int64	광섬유 ID. 각 관측에서 빛을 초점면으로 향하게 한 광섬유를 식별하는 번호

데이터 사전작업

천문학 데이터이기 때문에 관측값의 차이가 클 수 있으므로, 0~1의 값으로 정규화하며, 많은 컬럼들은 PCA 차원 축소를 활용하여 머신러닝 효율성을 확보합니다.
Grid search CV을 통한 머신러닝 과정을 최적화합니다.
ID와 같은 예측과 관련 없는 컬럼을 제거합니다.

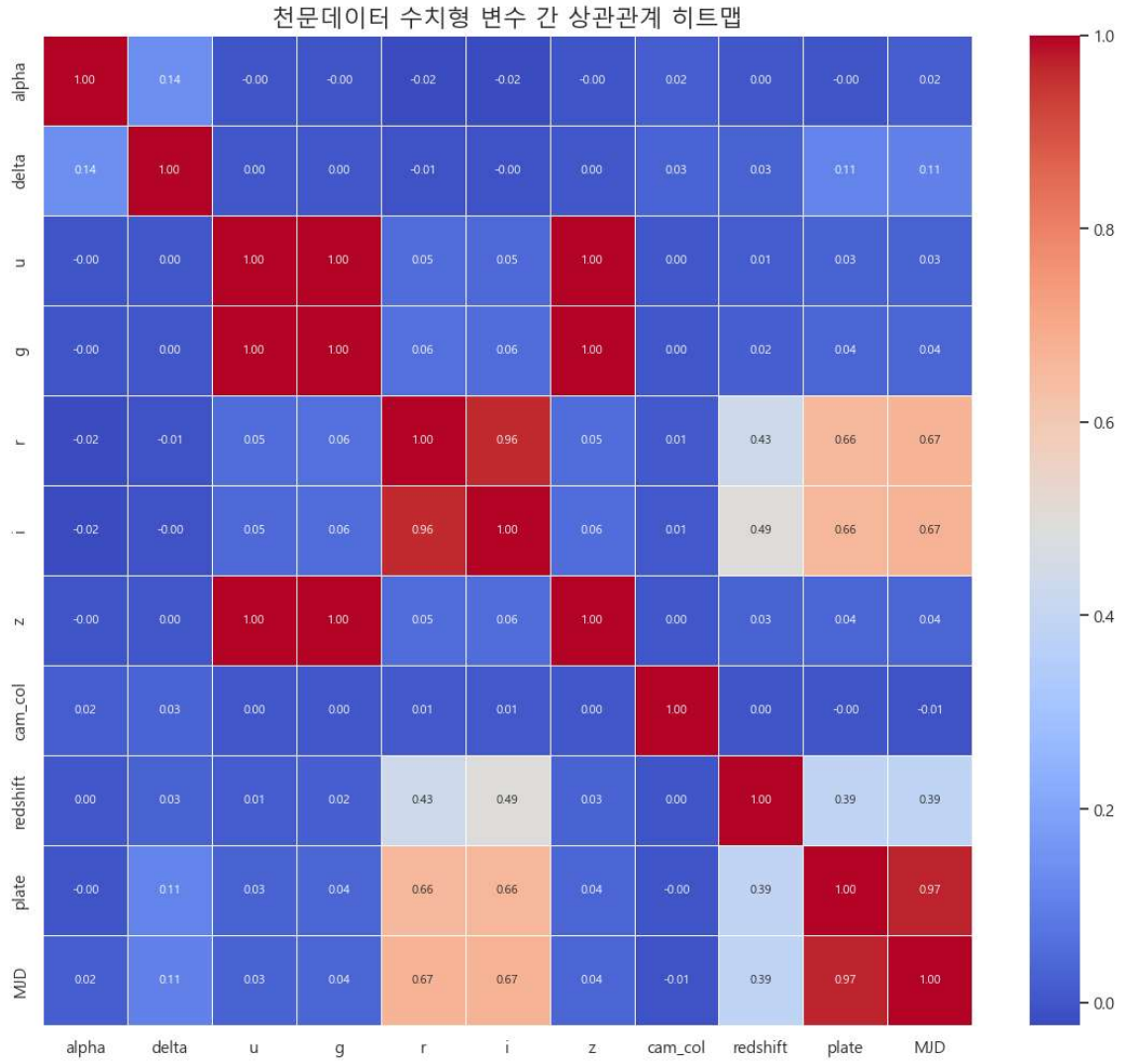
3. 분석방법 소개

저는 “어떤 관측값으로 적색편이(redshift)를 정확히 예측할 수 있는가?”라는 핵심 질문에 대한 통계적 해답을 찾기 위해, 예측 모델링과 심층 분석을 결합한 데이터 과학 접근법을 사용하고자 합니다. 회귀 모델을 기반으로 적색편이 값을 예측하는 주요 관측 변수들의 기여도를 정량화하는 것입니다. 이를 통해 색상 필터 값(u, g, r, i, z)과 이로부터 파생된 ‘색 지수(color index)’가 적색편이 예측에 얼마나 큰 영향을 미치는지 분석하고, 나아가 특정 색 지수들의 조합이 예측 정확도를 얼마나 향상시키는지 그 상호작용 효과를 구체적인 수치로 규명하고자 합니다.
이를 위해, 먼저 이상치를 정제하고 값의 범위가 다른 관측 데이터들을 정규화(Normalization)하며, 상관관계가 높은 변수들은 주성분 분석(PCA)을 통해 차원을 축소하는 등 체계적인 데이터 전처리 과정을 거칠 것입니다. 이후, 단순한 선형 회귀부터 성능이 뛰어난 XGBoost, LightGBM과 같은 앙상블 모델에 이르기까지 다양한 회귀 모델을 구축하여 성능을 비교 평가할 것입니다.
MAE, MSE, R2와 같은 평가지표를 통해 가장 예측력이 뛰어난 최적의 모델을 선정하고, 하이퍼파라미터 튜닝을 통해 모델의 성능을 극한으로 끌어올릴 계획입니다.
단순히 예측에서 그치는 것이 아니라, 최종적으로 선정된 모델을 분석하여 각 관측 변수가 적색편이 예측에 미치는 영향력을 정량적으로 측정하는 것이 이 분석의 핵심입니다. 특히, “천체의 색 지수는 적색편이와 강한 선형 관계를 가질 것이며, 이를 통해 거리를 효과적으로 추정할 수 있을 것이다”라는 가설을 통계적으로 검증하고, 나아가 심층 신경망(DNN) 모델까지 구축하여 더 높은 수준의 예측 정확도를 달성하고 의미 있는 결론을 도출하고자 합니다.

4. 데이터 기초분석 및 시각화

1. 변수 간 상관관계 히트맵

데이터셋의 여러 변수들이 적색편이와 어떤 관계가 있는지 전체적으로 상관관계 분석을 위해 히트맵을 통해 적색편이 값과 가장 밀접한 상관관계를 직관적으로 확인하였습니다.



히트맵을 통해 적색편이는 색 필터(r, i)와 매우 강한 양의 상관관계(0.43 ~ 0.49)를 보인다는 것을 알 수 있었습니다.

이는 천체의 색상 정보가 그 천체가 얼마나 멀리 있는지를 예측하는 데 매우 중요한 단서임을 의미합니다.

각 색 필터(u, g, r, i, z)들끼리도 서로 매우 강하게 연관되어 있습니다. 이는 특정 필터에서 밝게 관측된 천체는 다른 필터에서도 밝게 관측될 가능성이 높다는 것을 뜻합니다.

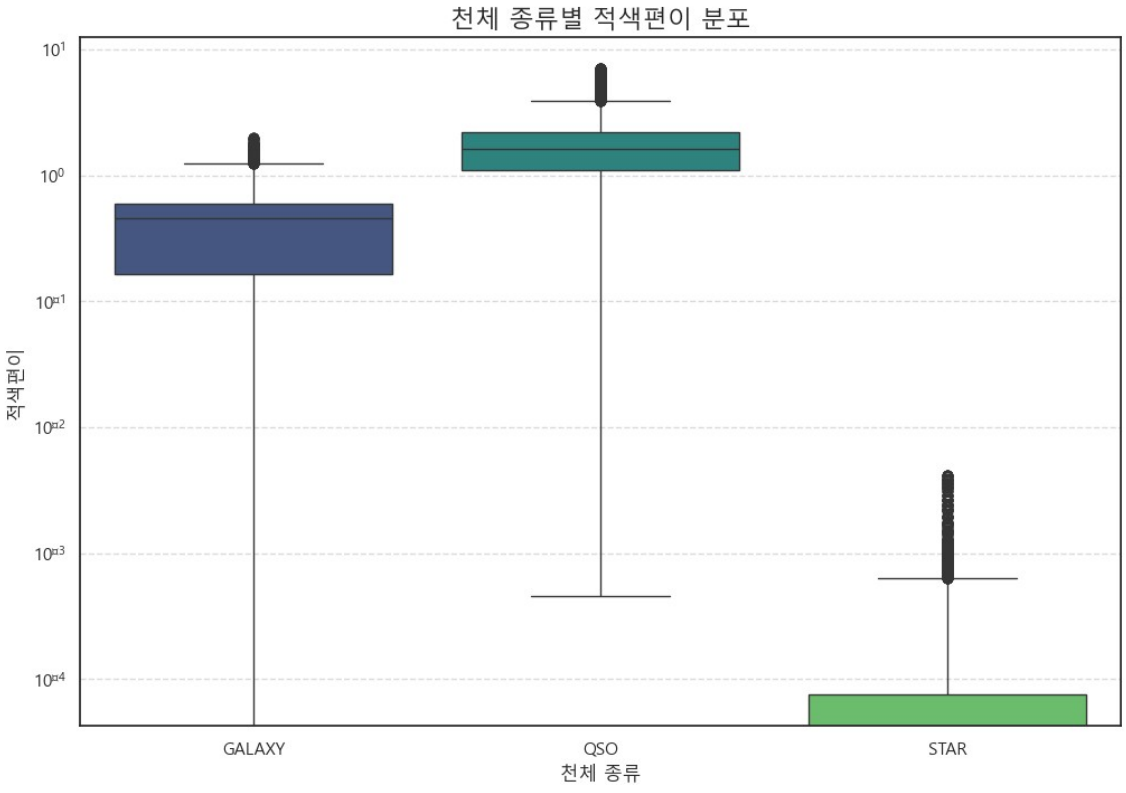
2. 천체 종류에 따른 적색편이 분포

이 박스 플롯은 세 가지 천체 종류(STAR, GALAXY, QSO)별로 적색편이 값의 분포를 비교합니다. 적색편이는 천체가 우리로부터 멀어지는 속도를 나타내며, 보통 천체까지의 거리에 비례합니다.

퀘이사의 적색편이 분포가 은하나 별에 비해 압도적으로 높습니다. 이는 퀘이사가 우주에서 관측 가능한 가장 멀리 있는 천체 중 하나라는 사실을 명확히 보여줍니다.

별의 적색편이는 거의 0에 가깝습니다. 이는 별들이 우리 은하 내, 즉 천문학적 관점에서는 상대적으로 아주 가까운 거리에 있기 때문입니다.

이 그래프 하나만으로도 적색편이 값은 천체의 종류를 구분하는 데 매우 결정적인 특징임을 알 수 있습니다.

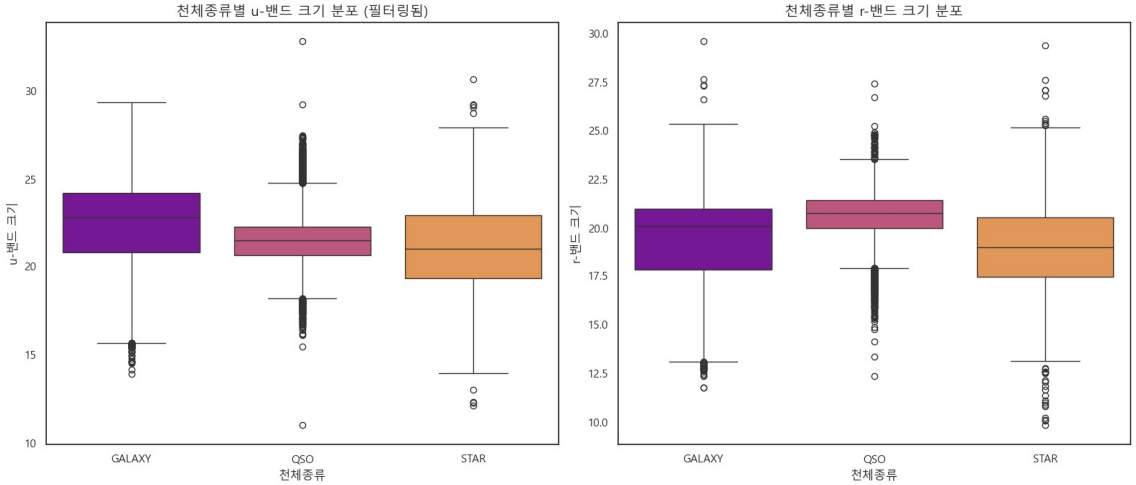


3. 천체 종류에 따른 밝기 분포

천체 종류별 고유한 색 특성을 파악하기 위해, 대표적인 두 색 필터인 자외선 필터(u-band)와 적색 필터(r-band)에서의 밝기 분포를 박스 플롯으로 비교 분석하였습니다.

분석 결과, 자외선 필터에서는 별과 은하의 밝기 분포가 뚜렷하게 구분되는 반면, 적색 필터에서는 두 천체의 밝기 분포가 매우 유사하게 나타났습니다.

이는 천체의 종류를 구별할 때, 단일 필터의 밝기보다 여러 필터 간의 밝기 차이, 즉 '색깔' 정보가 더 중요한 식별 기준이 될 수 있음을 나타냅니다.

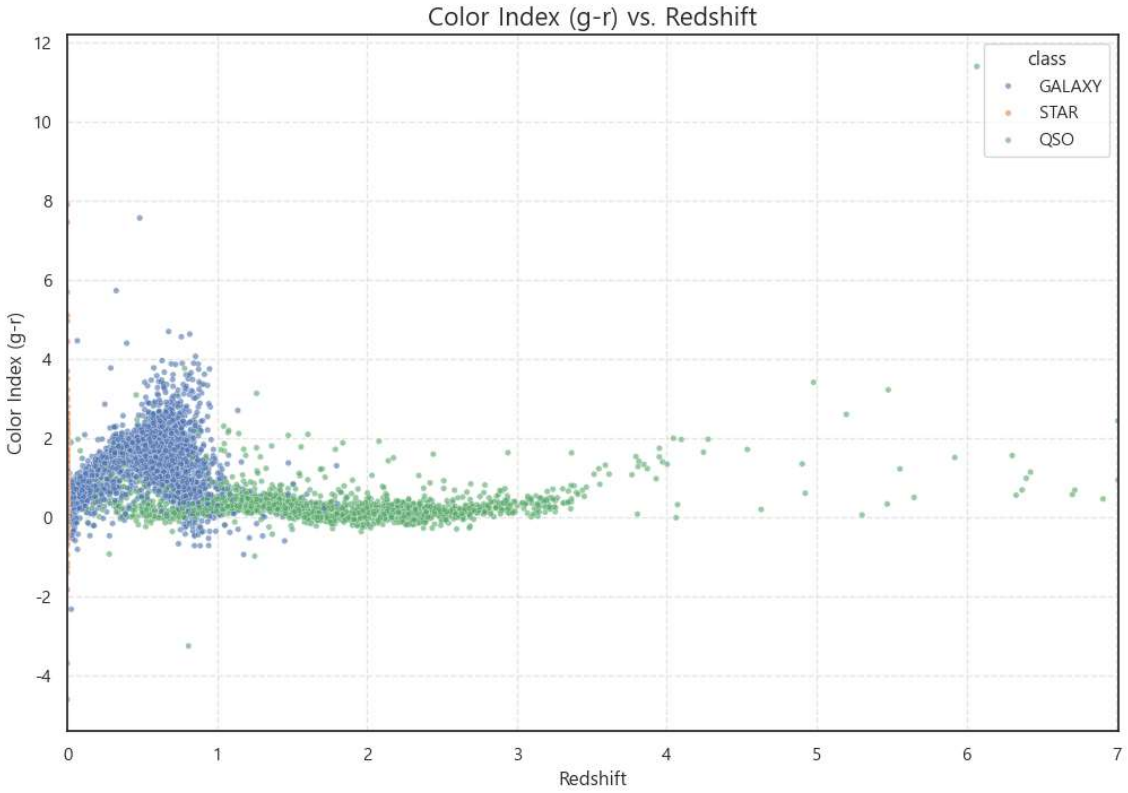


4. 적색편이에 따른 색 변화 분석

천체의 색을 수치화하기 위해 녹색 필터와 적색 필터의 밝기 차이(g-r)를 '색 지수(Color Index)'로 정의하고, 이 값이 적색편이와 어떤 관계를 갖는지 산점도를 통해 분석하였습니다.

분석 결과, 적색편이 값이 커질수록 색 지수 값도 함께 증가하는 뚜렷한 양의 상관관계를 확인할 수 있었습니다.

이는 천체가 우리로부터 멀리 있을수록 빛의 파장이 길어져 더 붉게 보인다는 것을 의미하며, 우주가 팽창하고 있다는 사실을 보여주는 매우 중요한 관측적 증거입니다.



지금까지 천체의 종류(class), 적색편이(redshift), 색상 필터(u, g, r, i, z) 등 다양한 관측 변수별로 분포와 상관관계를 시각적으로 분석하였습니다. 그 결과, 적색편이 예측에 가장 직접적으로 강한 영향을 미치는 변수는 각 색상 필터 값들임을 알 수 있었고, 이로부터 계산된 '색 지수(g-r)' 또한 적색편이와 뚜렷한 선형 관계를 갖는 보조 변수로 확인되었습니다. 반면, plate, MJD와 같은 관측 메타데이터는 적색편이와의 직접적인 연관성이 상대적으로 낮은 것으로 나타났습니다.

이제 단순 분포와 상관관계 분석을 넘어서, 여러 관측 변수들의 조합으로 적색편이 값을 정확하게 예측하는 통계적 모델을 구축하고자 합니다. 특히, 적색편이 값이 급격히 증가하는 퀘이사와 같은 특정 천체 그룹을 정확히 예측 해낼 계획입니다. 이를 통해, 시간이 많이 소요되는 분광 분석 없이도 측광 데이터만으로 천체의 거리를 효율적으로 추정할 수 있는 근거를 마련하고, 천문학 연구의 효율성을 높이는 데 기여하고자 합니다.

5. 향후 일정

1주차(10.20~10.26): 데이터 전처리(이상치 기준 정의 및 처리)
2주차(10.27~11.02): 데이터 전처리(정규화, 라벨인코딩)
3주차(11.03~11.09): 회귀 모델 성능 비교 및 평가
4주차(11.10~11.16): 하이퍼파라미터 튜닝 및 최적 모델 선정
5주차(11.17~11.23): 결과 분석 및 시각화
6주차(11.23~11.30): 결과보고서 작성

6. 참고 문헌

Estimating the Photometric Redshifts of Galaxies and QSOs Using Regression Techniques in Machine Learning(머신 러닝에서 회귀 기법을 사용하여 은하와 QSO의 광도 적색편이 추정)
https://arxiv.org/abs/2201.04391

사용한 데이터셋

https://www.kaggle.com/datasets/mirichoi0218/insurance/code(2025-10-8)

7. 부록

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings(action='ignore')
plt.rc("font", family = "Malgun Gothic")

file_path = 'star_classification.csv'

df = pd.read_csv(file_path)

df.info()
# 상관없는 컬럼 제거
df.drop(["rerun_ID","obj_ID","run_ID","field_ID","spec_obj_ID","fiber_ID"],axis=1,inplace=True)
# 모든 숫자형 관측 데이터를 선택
numeric_cols = df.select_dtypes(include=['float64', 'int64'])


# 상관관계 행렬 계산
correlation_matrix = numeric_cols.corr()

# 히트맵 시각화
plt.figure(figsize=(14, 12))
sns.heatmap(
    correlation_matrix,
    annot=True,
    cmap='coolwarm',
    fmt='.2f',
    linewidths=.5,
    annot_kws={"size": 8}
)
plt.title('천문데이터 수치형 변수 간 상관관계 히트맵', fontsize=16)
plt.show()
plt.figure(figsize=(12, 8))
sns.boxplot(x='class', y='redshift', data=df, palette='viridis',hue="class", legend=False)
plt.title('천체 종류별 적색편이 분포', fontsize=16)
plt.xlabel('천체 종류', fontsize=12)
plt.ylabel('적색편이', fontsize=12)
# y축을 로그 스케일로 변경하여 분포를 더 명확하게 확인
plt.yscale('log')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
# 이 플롯에서 이상치(plot 4분위 밖에 있는 값)는 머신러닝 할 때 처리함

df_filtered = df[df['u'] > 0]


fig, axes = plt.subplots(1, 2, figsize=(16, 7))
```

```
# u-band (자외선) 밝기 분포 (필터링된 데이터 사용)
sns.boxplot(ax=axes[0], x='class', y='u', data=df_filtered,
            hue='class', legend=False, palette='plasma')
axes[0].set_title('천체종류별 u-밴드 크기 분포 (필터링됨)', fontsize=14)
axes[0].set_xlabel('천체종류', fontsize=12)
axes[0].set_ylabel('u-밴드 크기', fontsize=12)

# r-band (적색) 밝기 분포
sns.boxplot(ax=axes[1], x='class', y='r', data=df,
            hue='class', legend=False, palette='plasma')
axes[1].set_title('천체종류별 r-밴드 크기 분포', fontsize=14)
axes[1].set_xlabel('천체종류', fontsize=12)
axes[1].set_ylabel('r-밴드 크기', fontsize=12)

plt.tight_layout()
plt.show()

# 색 지수 (g-r) 계산
df['color_index'] = df['g'] - df['r']

# 데이터가 많으므로 10000개만 샘플링하여 시각화
df_sample = df.sample(n=10000, random_state=42)
# df_sample = df

plt.figure(figsize=(12, 8))
sns.scatterplot(x='redshift', y='color_index', hue='class', data=df_sample,
alpha=0.6, s=15)
plt.title('Color Index (g-r) vs. Redshift', fontsize=16)
plt.xlabel('Redshift', fontsize=12)
plt.ylabel('Color Index (g-r)', fontsize=12)
plt.xlim(df_sample['redshift'].min(), df_sample['redshift'].max())
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```