측광 데이터 기반
적색편이 예측 모델 구축

2021212816 김이현

목차

- 연구 배경 및 목적
- 도메인 지식 설명
  - 적색편이
  - 측광/분광
- 데이터셋 소개
  - 컬럼 소개
- 데이터 전처리
- 데이터 분석
- 계획 수립

연구 배경 및 목적

- 우주의 구조와 진화 과정을 이해하기 위해 천체의 **거리**를 알아내는 것은 천문학의 가장 근본적인 과제

- 적색편이(Redshift)라는 물리량을 통해 천체까지의 거리 계산 가능

- 적색편이를 측정하는 대표적인 방법 2가지는 분광 분석/측광 분석

적색편이

적색편이(Redshift)
멀어지고 있는 천체에서 오는 빛의 파장이 늘어나
스펙트럼이 붉은색 쪽으로 치우쳐 보이는 현상

빛의 도플러 효과에 의해 발생

적색 편이 값(z)에 대한 거리 계산
1. 후퇴 속도(v) 계산
$v \approx z \times c$
$v \approx 0.1 \times 300{,}000\ km/s = 30{,}000\ km/s$

2. 메가파섹(Mpc) 단위 거리 계산
$d \approx v / H_0$
$d \approx 30{,}000\ km/s / 70\ km/s/Mpc =$ 약 $428.6\ Mpc$

3. 광년(Light-year)으로 변환
$428.6\ Mpc \times$ 326만 광년$/Mpc \approx 1{,}397{,}236{,}000$ 광년
약 14억 광년

측광/분광

항목	측광 데이터 (Photometry)	분광 데이터 (Spectroscopy)
핵심 개념	빛의 **총 밝기** (필터별)	빛의 **상세 스펙트럼** (파장별)
비유	색 셀로판지로 찍은 사진	프리즘으로 나눈 무지개
주요 정보	색깔, 밝기	**정확한 적색편이**, 화학 성분, 온도
효율성	**빠름** (대규모 관측에 유리)	**느림** (개별 천체 심층 분석에 유리)

Stellar Classification Dataset - SDSS17

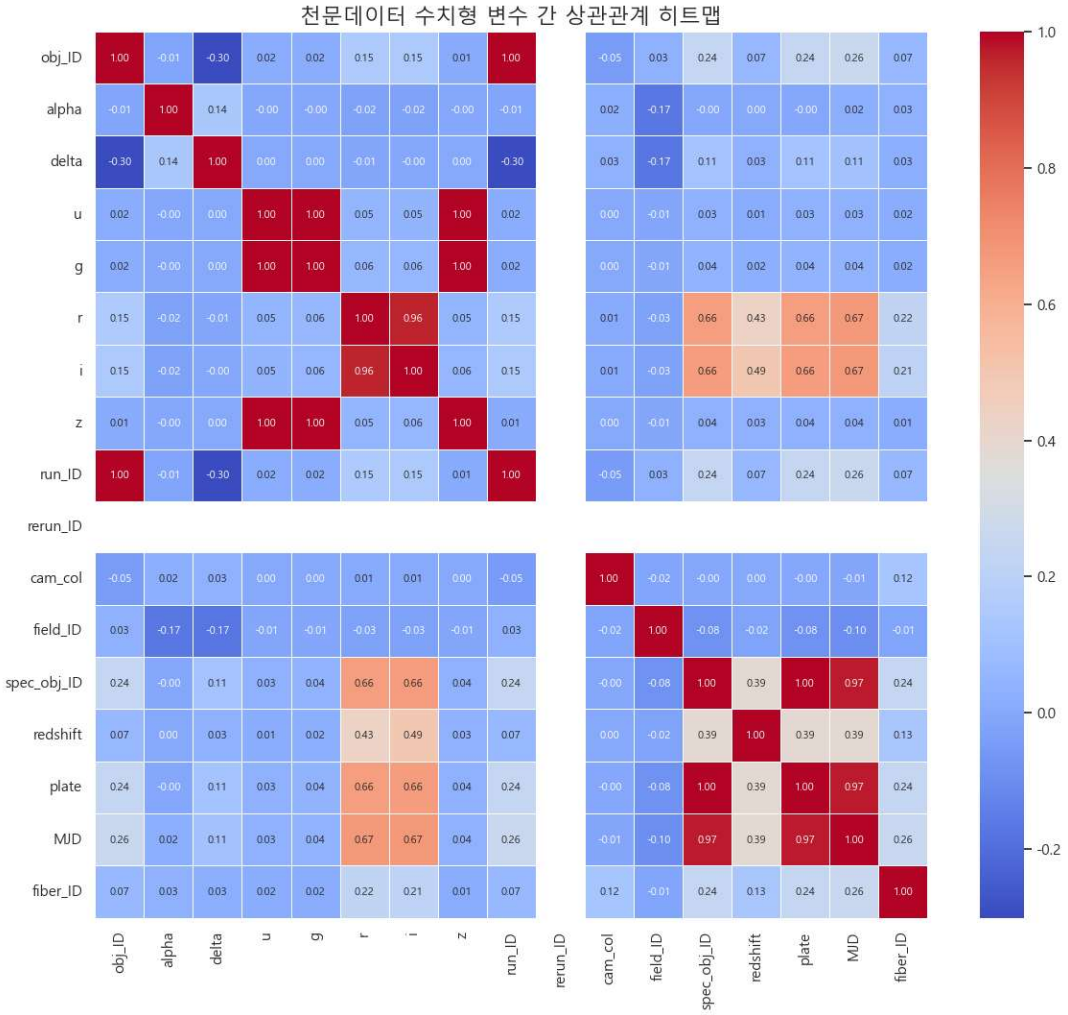
```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 18 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   obj_ID       100000 non-null  float64
 1   alpha        100000 non-null  float64
 2   delta        100000 non-null  float64
 3   u            100000 non-null  float64
 4   g            100000 non-null  float64
 5   r            100000 non-null  float64
 6   i            100000 non-null  float64
 7   z            100000 non-null  float64
 8   run_ID       100000 non-null  int64  
 9   rerun_ID     100000 non-null  int64  
 10  cam_col      100000 non-null  int64  
 11  field_ID     100000 non-null  int64  
 12  spec_obj_ID  100000 non-null  float64
 13  class        100000 non-null  object 
 14  redshift     100000 non-null  float64
 15  plate        100000 non-null  int64  
 16  MJD          100000 non-null  int64  
 17  fiber_ID     100000 non-null  int64  
dtypes: float64(10), int64(7), object(1)
memory usage: 13.7+ MB
```

1. obj_ID = Object Identifier, the unique value that identifies the object in the image catalog used by the CAS
2. alpha = Right Ascension angle (at J2000 epoch)
3. delta = Declination angle (at J2000 epoch)
4. u = Ultraviolet filter in the photometric system
5. g = Green filter in the photometric system
6. r = Red filter in the photometric system
7. i = Near Infrared filter in the photometric system
8. z = Infrared filter in the photometric system
9. run_ID = Run Number used to identify the specific scan
10. rereun_ID = Rerun Number to specify how the image was processed
11. cam_col = Camera column to identify the scanline within the run
12. field_ID = Field number to identify each field
13. spec_obj_ID = Unique ID used for optical spectroscopic objects (this means that 2 different observations with the same spec_obj_ID must share the output class)
14. class = object class (galaxy, star or quasar object)
15. redshift = redshift value based on the increase in wavelength
16. plate = plate ID, identifies each plate in SDSS
17. MJD = Modified Julian Date, used to indicate when a given piece of SDSS data was taken
18. fiber_ID = fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

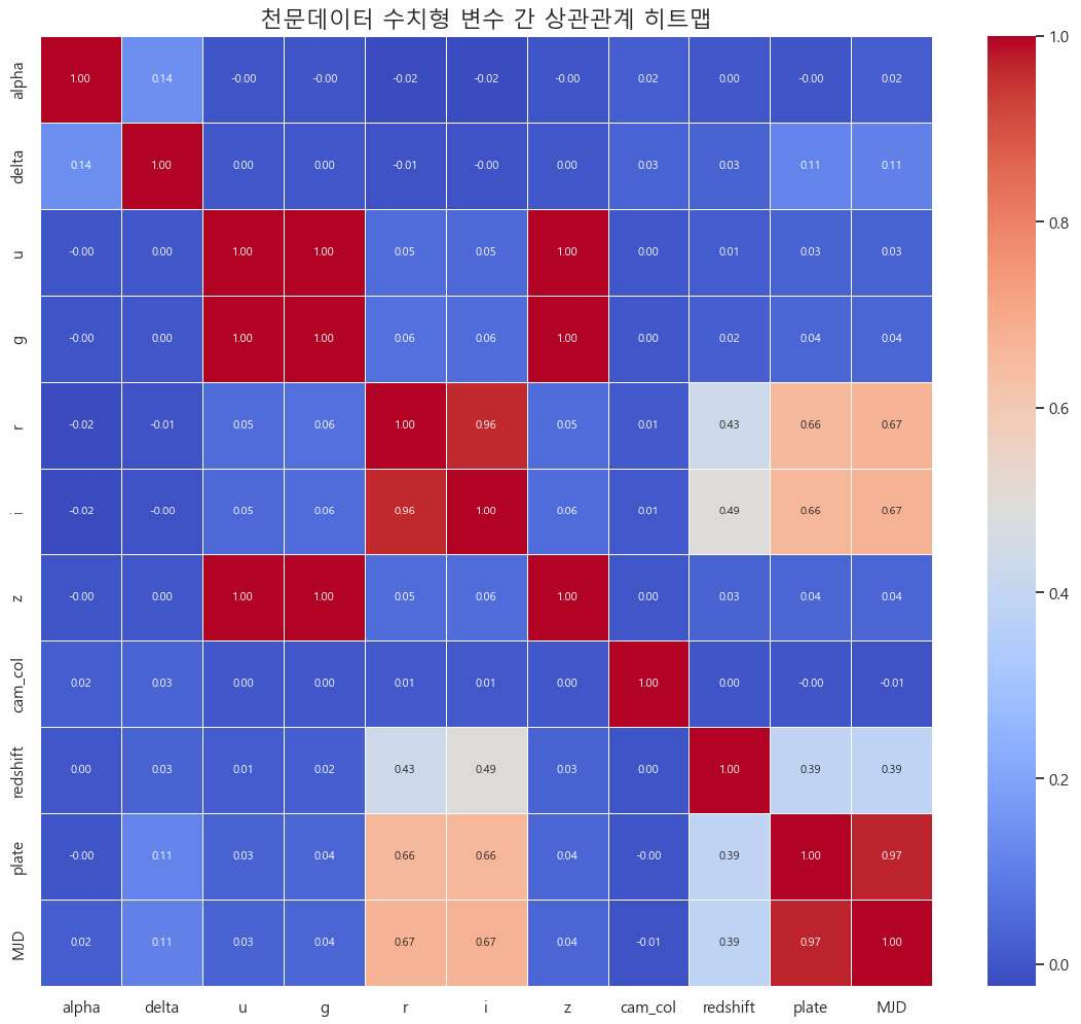
Stellar Classification Dataset - SDSS17

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. (자외선만 포함시킨 필터)
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

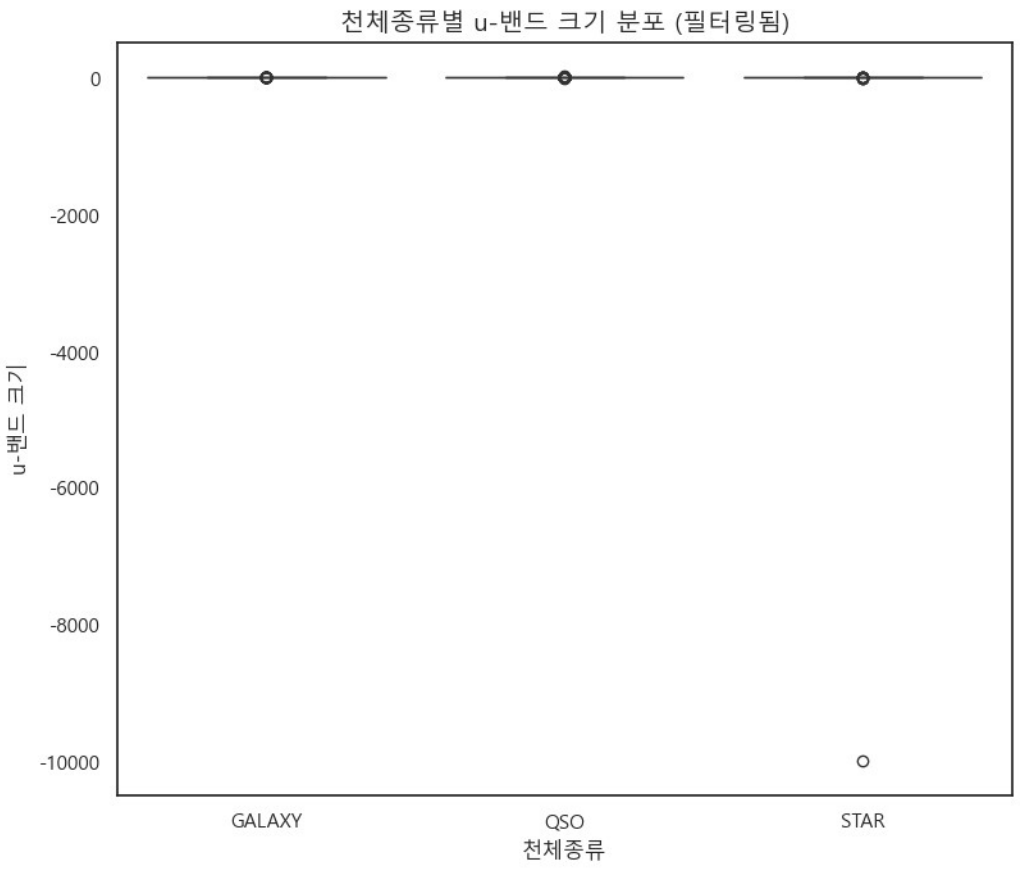
데이터 전처리 – 컬럼 제거 전

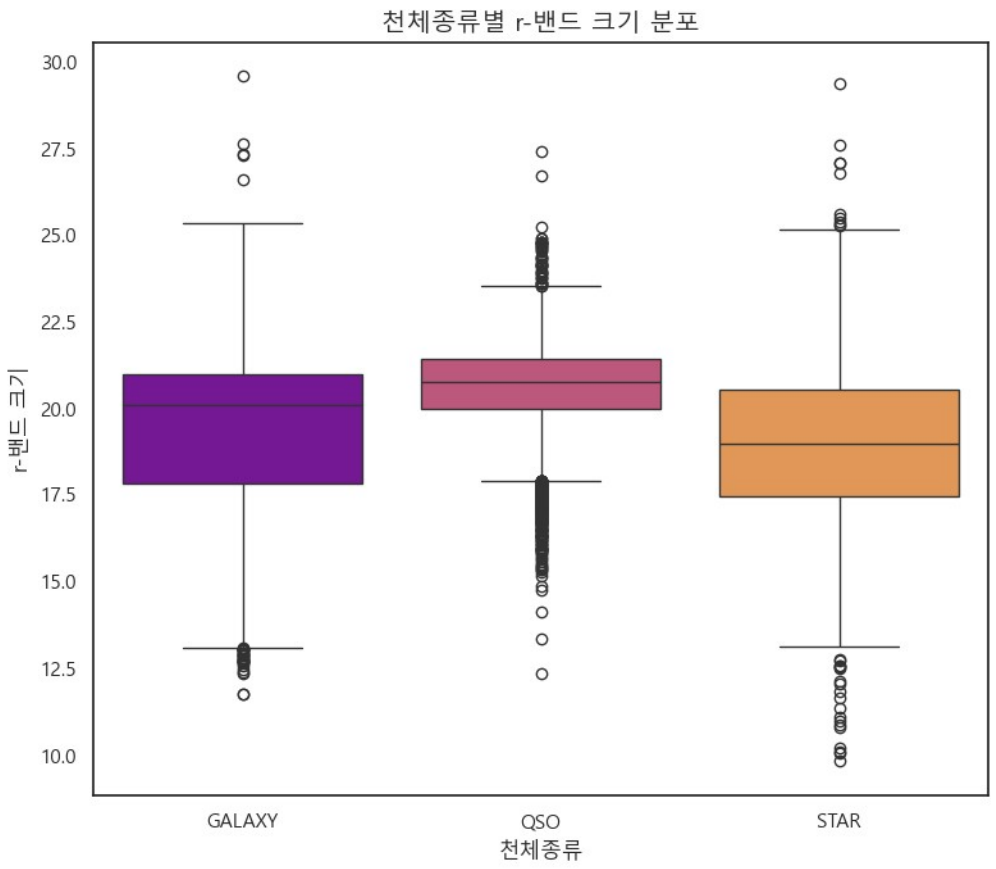


데이터 전처리 – 컬럼 제거 후

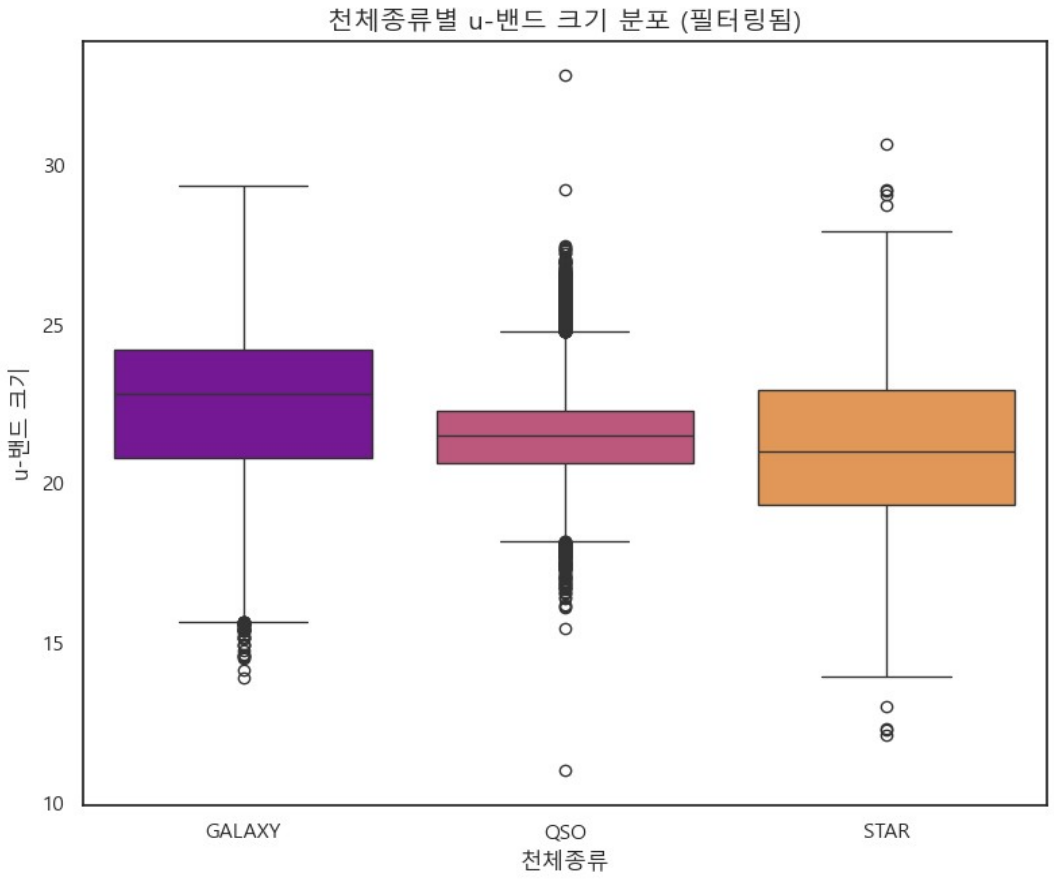


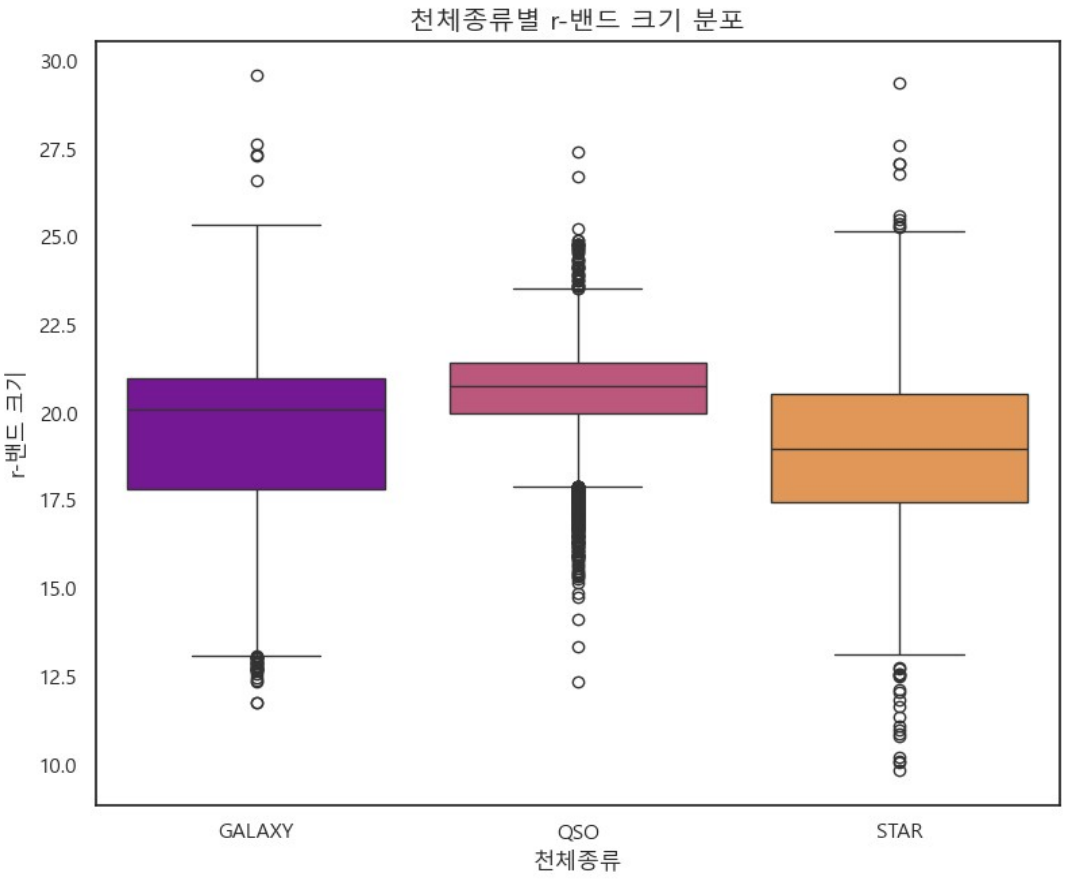
데이터 전처리 - 이상치 제거 전





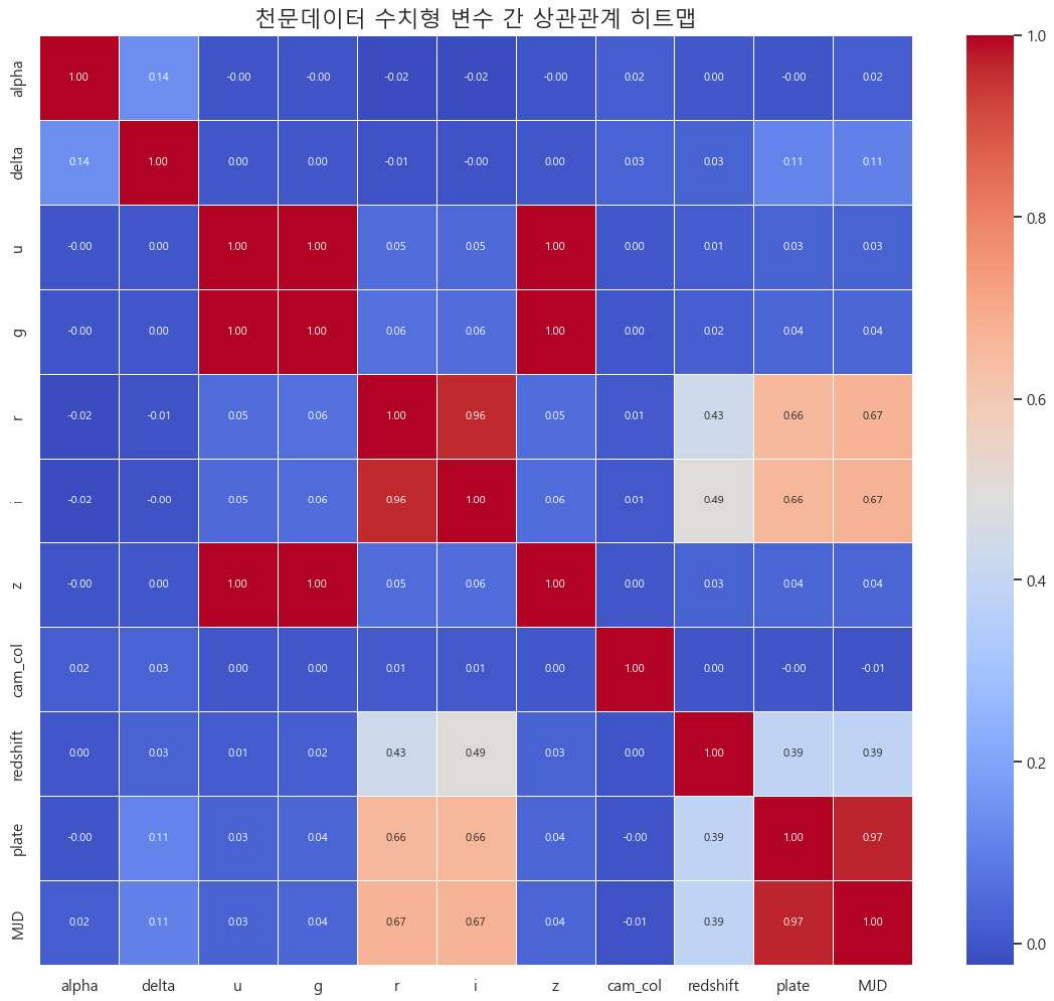
데이터 전처리 - 이상치 제거 후





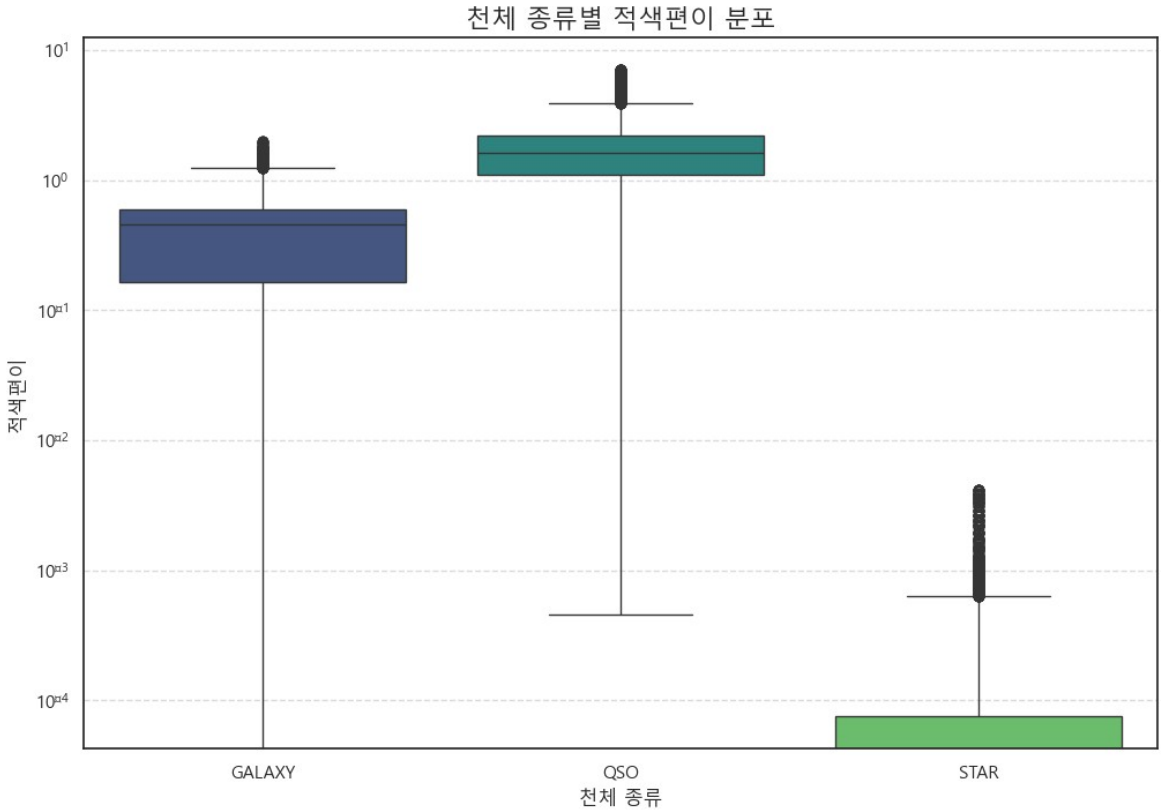
데이터 분석 - 전체 변수간 상관관계

- 여러 변수들 간의 상관관계를 직관적으로 파악하기 위해 히트맵을 사용하여 시각화
- redshift는 색 필터(r, i)와 강한 양의 상관관계(0.43 ~ 0.49)
- 천체의 색상 정보가 그 천체가 얼마나 멀리 있는지를 예측하는 데 매우 중요한 단서



데이터 분석 - **천체 종류에 따른 적색편이 분포**

- 특정 카테고리(천체 종류)에 따른 숫자형 데이터(적색편이)의 분포를 비교하는 데 가장 최적화된 box plot을 사용하여 시각화
- 별의 적색편이는 0에 가깝다
- 이 그래프를 통해 적색편이 값은 천체의 종류를 구분하는 데 매우 결정적인 특징임을 알 수 있다

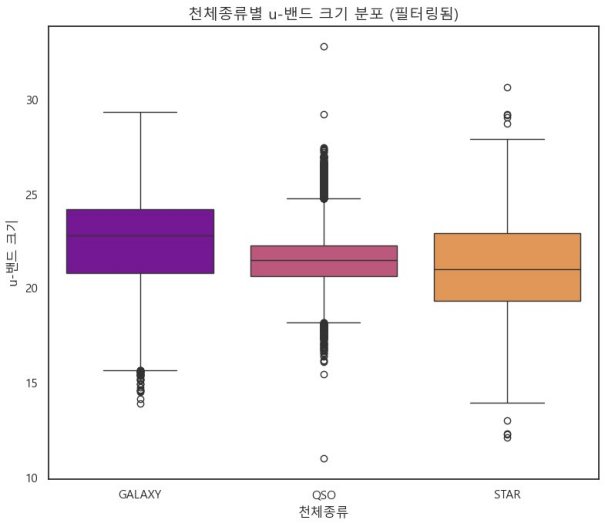


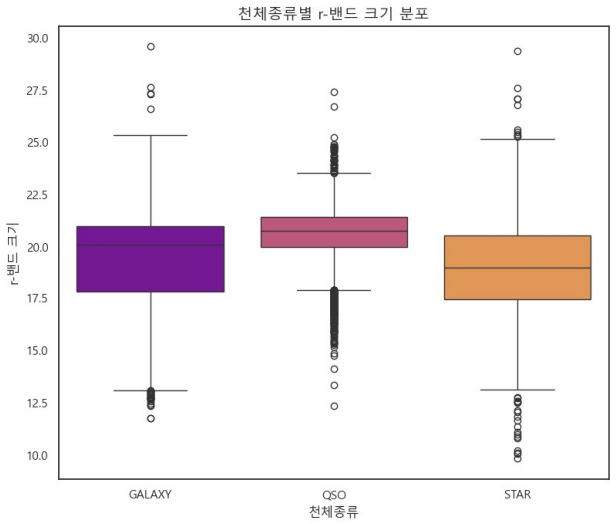
데이터 분석 - **천체 종류에 따른 밝기 분포**

- 여러 그룹(천체 종류)에 걸친 숫자형 데이터(밝기 등급)의 분포를 한눈에 비교하는 데 가장 효과적이기 때문에 가장 최적화된 box plot을 사용하여 시각화

- 자외선 필터에서는 별과 은하의 밝기 분포가 뚜렷하게 구분된다
- 적색 필터에서는 두 천체의 밝기 분포가 매우 유사하게 나타났다.

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

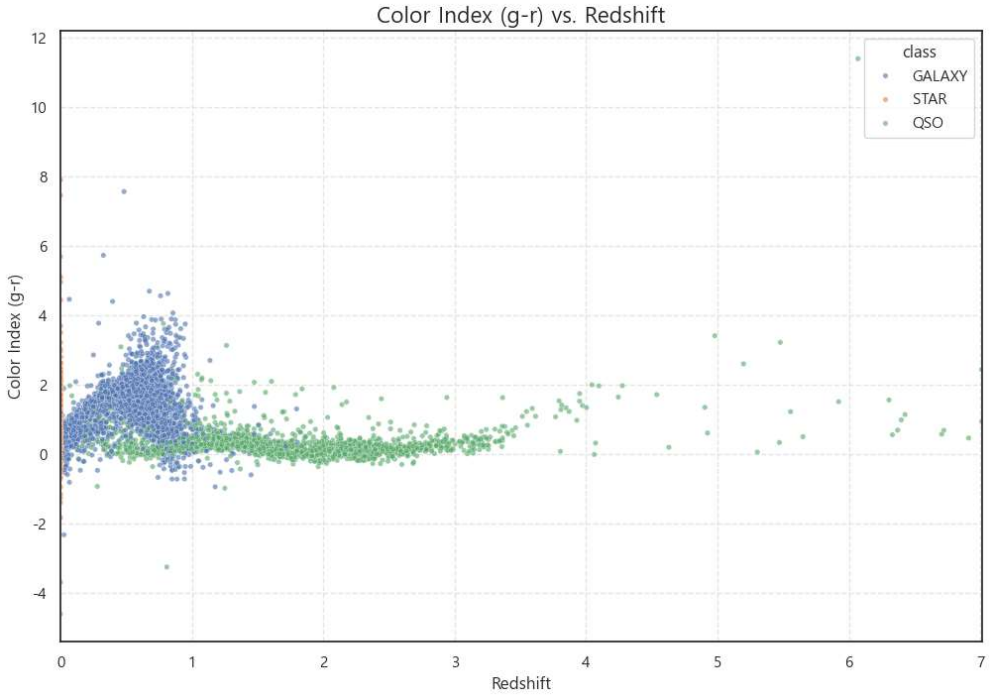




u-밴드 크기 : u 필터로 필터링 하여 측정한 밝기
r-밴드 크기 : r 필터로 필터링 하여 측정한 밝기

데이터 분석 - **적색편이에 따른 색 변화 분석**

- 두 개의 연속적인 숫자형 변수(적색편이, 색 지수) 사이의 관계, 경향성, 분포를 시각화하는 데 가장 직접적이고 효과적인 방법이기 때문에 Scatter를 이용하여 시각화
- 녹색 필터와 적색 필터의 밝기 차이(g-r)를 '색 지수(Color Index)'로 정의
- 적색편이 값이 커질수록 색 지수 값도 함께 증가한다
- 천체가 우리로부터 멀리 있을수록 빛의 파장이 길어져 더 붉게 보인다는 것을 의미하며, 우주가 팽창하고 있다는 사실을 보여주는 증거



계획 수립

주차	내용
1주차(10.20~10.26)	데이터 전처리(이상치 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

참고자료

- Dataset
  - https://www.kaggle.com/datasets/mirichoi0218/insurance/code(2025-10-8)
- Paper
  - https://arxiv.org/abs/2201.04391