**Gaia 항성 데이터 분석 프로젝트 설계 보고서**

Jun Kwon

202376701

2025

**목차**

[**1. 프로젝트 개요** 2](#_Toc195972518)

[2. 데이터 분석 2](#_Toc195972519)

[**2-1. 데이터 수집 및 데이터 전처리** 2](#_Toc195972520)

[**2-2. 분석 방법** 3](#_Toc195972521)

[**2-3. 분석 결과** 3](#_Toc195972522)

[**2-4. 분석 신뢰성 및 검증** 4](#_Toc195972523)

[3. 결론 및 향후 계획 4](#_Toc195972524)

## **1. 프로젝트 개요**

본 프로젝트는 유럽우주국(ESA)의 Gaia DR3 데이터셋을 활용하여, 항성(별)들의 물리적 특성과 이들 간의 상관관계를 분석하는 것을 목표로 한다. 특히 항성의 거리, 광도, 표면 중력, 온도 등의 특성을 기반으로, 기존 천문학 이론에서 예측 가능한 관계를 검증하는 동시에, 데이터 기반으로 예상치 못한 패턴이나 새로운 통찰을 발견하는 것을 지향한다. 데이터는 고정밀 측정을 통해 확보된 것으로, 위치, 운동, 광도, 물리적 특성(온도, 중력 등)을 포함한다.

주요 분석 대상 변수는 거리(distance\_gspphot), 광도(phot\_g\_mean\_mag, phot\_bp\_mean\_mag, phot\_rp\_mean\_mag), 온도(teff\_gspphot), 표면중력(logg\_gspphot)이다.

# **2. 데이터 분석**

## **2-1. 데이터 수집 및 데이터 전처리**

Gaia DR3 데이터는 매우 정밀하게 수집되었으며, 관측 오차가 큰 경우 아예 데이터로 등록되지 않는 특성이 있다. 따라서 본 분석에서는 결측값만 제거하고 추가적인 이상치 처리는 실시하지 않았다. 분석을 위해 필요한 파생변수도 생성하였다. 상대등급(apparent magnitude) 데이터를 절대등급(absolute magnitude)으로 변환하였으며, 변환식은 M = m - 5 log10(d/10)이다. 여기서 M은 절대등급, m은 상대등급, d는 거리를 파섹 단위로 나타낸 값이다. 데이터 분포를 분석에 적합한 형태로 만들기 위해 Yeo-Johnson 정규화를 적용하여 데이터의 왜도를 개선하고, 이후 StandardScaler와 MinMaxScaler를 적용하여 스케일링하였다. 이 과정을 통해 분석 과정에서 변수 간 비교 및 관계 분석이 용이하도록 데이터 형태를 정비하였다.  
분석을 위한 데이터를 전처리 후 확인 하였을때 Figure 1 과 같이 HR도의 노란부분에 해당되는 데이터라는 것을 확인하였다. 이를 통해 해당 데이터가 주계열성과 준거성이 포함된 데이터라는 것을 알 수 있었고 이를 참고하여 분석을 진행하였다.

## **2-2. 분석 방법**

본 분석은 탐색적 데이터 분석(Exploratory Data Analysis, EDA)을 통한 상향식 접근 방식을 채택하였다. 이는 다양한 변수들 간의 관계를 사전에 정의하지 않고, 시각화와 상관관계 탐색을 통해 의미 있는 패턴을 직접 발견하는 방법이다. 변수 간 관계를 시각적으로 탐색하기 위해 산점도 행렬(Scatter Matrix)을 활용하였으며, 변수들 간의 선형적 상관성을 수치화하기 위해 Pearson 상관계수를 함께 계산하여 분석의 신뢰성을 보완하였다.

## **2-3. 분석 결과**

분석을 통해 여러 항목에서 예측 가능한 결과가 확인되었다. 첫째, [Figure 3 의 1] 연주시차와 거리는 예상대로 반비례 관계를 보였으며, 연주시차가 작을수록 거리가 먼 것으로 나타났다. 둘째, [Figure 3 의 3-1] 항성의 광도와 온도의 관계는 Figure 1의 HR 다이어그램(Hertzsprung-Russell Diagram)에서 기대할 수 있는 분포와 유사하였다. 즉, 온도가 높은 별일수록 광도가 높은 경향을 보였지만, 별의 종류에 따라 관계가 다르게 나타나는 부분도 관찰되었다. 셋째, [Figure 3 의 3-3] 광도 간 관계에서는 G밴드 광도가 높을수록 BP 및 RP밴드 광도도 높게 나타나, 이들 간에 강한 양의 상관관계가 존재함을 확인할 수 있었다. 넷째, [Figure 3 의 4] 항성의 표면중력과 온도 간에는 온도가 상승할수록 표면중력도 다소 상승하는 경향이 나타났으나, 항성의 종류별로 편차가 존재하여 단순한 선형 관계로 설명되지는 않았다. 다섯째, [Figure 3 의 5-2] 광도와 표면중력 사이에는 광도가 높을수록(절대등급이 낮을수록) 표면중력이 작아지는 경향이 있었다. 이는 밝은 별일수록 크기가 크고, 이로 인해 표면중력이 상대적으로 낮아지는 항성 물리학적 특성과 일치하였다.

한편, 분석 과정에서 일부 예상 밖의 결과도 확인되었다. [Figure 3 의 2] 거리가 가까운 별들만 선별하여 비교했을 때, 거리와 항성의 온도 간에 양의 상관관계가 나타나는 현상이 관찰되었다. 이는 가까운 거리에서는 광도가 낮은 별들이 주로 관측되는데, 이들은 주계열성(Main Sequence) 항성에 해당하며, 이들의 온도가 비교적 낮기 때문으로 해석할 수 있다. 또한, [Figure 3 의 3-2] 항성의 광도와 거리 간의 관계에서도 거리가 멀수록 절대등급이 낮아(광도가 높아)지는 경향이 관찰되었다. 이는 광학 관측의 한계로 인해 멀리서 관측 가능한 별들이 대체로 밝은 별이라는 관측편향에 기인한 것으로 보인다. 마지막으로, [Figure 3 의 5-1] 거리와 표면중력 간에도 거리가 멀수록 표면중력이 작아지는 패턴이 나타났는데, 이는 멀리 있는 별들이 광도가 높고 대형 항성일 가능성이 높으며, 이로 인해 표면중력이 낮아지는 경향을 보인다고 해석할 수 있다.

## **2-4. 분석 신뢰성 및 검증**

본 분석은 ESA Gaia DR3라는 신뢰성 높은 데이터를 기반으로 수행되었다. 데이터 전처리 과정에서는 결측값만 제거하고 불필요한 이상값 제거를 생략함으로써 원 데이터의 신뢰성을 최대한 유지하였다. 또한, 데이터 스케일링 및 정규화 과정을 투명하게 수행하고, 분석 방법론으로 객관적인 통계적 방법(산점도 행렬, Pearson 상관계수 계산)을 적용하였다. 분석 결과는 기존 천문학 이론(HR Diagram 등)과 일치하는 부분이 많아 분석 과정의 신뢰성을 높였으며, 예상 밖의 결과에 대해서도 논리적인 근거를 기반으로 해석을 제시하였다.

# **3. 결론 및 향후 계획**

본 프로젝트를 통해 Gaia DR3 데이터에 기반한 항성 물리적 특성 간의 관계를 성공적으로 분석하였다. 대부분 기존 이론에 부합하는 결과가 도출되었으나, 일부 예상 밖의 결과 또한 관찰되어 추가적인 심화 분석의 필요성이 제기되었다. 향후 계획으로는 별의 스펙트럼 종류 별로 세분화하여 분석을 진행하고, 관측편향된 데이터를 보정하는 방법을 적용하여 보다 정밀한 관계 분석을 시도할 예정이다. 또한, 단순 선형 분석을 넘어 기계학습 기법을 이용하여 비선형적 변수 간 관계를 탐색하는 연구도 병행할 계획이다.

Figures  
스크린샷, 텍스트, 멀티미디어 소프트웨어, 천문학이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.  
Figure 1. H-R도(Hertzsprung-Russell diagram) 와 분석 데이터가 포함된 범위   
항성들의 종류에 따라 밝기와 온도에 따라 나누어 놓은 등급도와 분석에 사용된 데이터(gaia dr3 데이터)가 포함되어있는 범위(노란색 범위)   
도표, 그래프, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.  
Figure 2. 분석에 사용된 전처리된 gaia dr3 데이터의 분포  
텍스트, 패턴, 라인, 직사각형이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.  
Figure 3. 전처리된 gaia dr3 데이터의 산점도 행렬 분석결과