GAIA_EDA

Taine Murphy - 300472954, Van Vo - 300520137, Max Tan - 300526544

2022-09-03

Acknowledgement:

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

Background and Data

This project has made use of the GAIA data set produced by the European Space Agency, from mission GAIA (https://www.cosmos.esa.int/gaia). Mission GAIA is launched to record the movement and position of billion objects in the sky, with a purpose to gain deeper understanding of the history and evolution of the Galaxy.

The GAIA data records information on location, proper motions, parallax, colour, magnitude, etc., of billion celestial objects, from which a three-dimensional mapping of the Galaxy is formed (Gaia Collaboration et al. (2016b): The Gaia mission). The GAIA instruments and payload setup allow the European Space Agency (ESA) to detect even the faintest objects and observe them a few times a year at different positions on the sky. With such information, GAIA has the power to unveil the Galaxy's history and also predict its future.

The team is intrigued by the idea of unravelling the Galaxy's origin and evolution. We are also motivated to obtain more astronomy knowledge. As a result, we choose the GAIA data set to base our project. Out of billions of objects in the Galaxy, we are especially interested in knowing more about open clusters. We hope to use the GAIA data to provide insights into some of the most famous open clusters, Pleiades and M67.

Data Summary

There are a total of 153 columns within the GAIA dataset "gaiadr3.gaia_source".

The range of data types include:

- char
- short
- float
- boolean
- double

Which cover a range of information about the dataset including, but not limited to:

- Position

- Movement (direction, speed)
- Distance
- Photometry (colour, brightness)
- Correlation values
- Classification probabilities (Quasar vs Galaxy vs Star)
- Measuring metrics (measurement error)

Data Gathering/Sampling Method

We ran two queries to gather two subsets of the dataset, each representing two open clusters (Pleiades and m67). This was gathered using the GAIA archive which filtered based on location, proper motion, present variables and error rates . These were collated using multiple different sources to construct an ADQL query within the archive to extract these subsets (Cánovas, 2022; Heyl et al., 2022).

Feature Selection

The 9 features we are looking at are "ra", "dec", "pmra", "pmdec", "parallax", "bp_rp", "phot_g_mean_mag", "distparsecs" and "abM".

Ra and Dec stand for Right Ascension and Declination respectively. They are to the sky what longitude and latitude are to the surface of the Earth. Right Ascension corresponds to east/west direction (like longitude), while Declination measures north/south directions, like latitude.

PmRA and PmDec are proper motion in right ascension direction and declination direction respectively. This is the local tangent plane projection of the proper motion vector in the direction of increasing right ascension and increasing declination.

Parallax is the apparent displacement of an object because of a change in the observer's point of view. Astronomers can use insights derived from the parallax measurements of the closer stars to estimate distances of those more distant.

Bp_rp is the colour of each star. More specifically, how blue it is. The more blue a star is, the higher its temperature.

Phot_g_mean_mag is the mean magnitude in the G band. This is computed from the G-band mean flux applying the magnitude zero-point in the Vega scale.

DistParsecs is the distance of a parsec which is 1 divided by the parallax divided by 1000. A parsec is a unit of distance used in astronomy.

Absolute magnitude is the intrinsic brightness of a star. This is calculated with the parallax and the observed colour from Earth.

We chose these paticular features because we wanted to isolate the most consistent, non-redundant, and relevant features to get the best insight into the GAIA dataset. We wanted a small number of features that are easy to explain because data that is too complex and unexplainable is not valuable. We especially wanted to look at position, distance, temperature, movement and brightness of the stars in the dataset and those are the features we thought would give us the best understanding.

Ethics, Privacy and Security

One of the most important ethical considerations we emphasise is properly acknowledging the authorship of the GAIA data in our project. Authorship and accreditation are ultimately important, especially in scientific research, as they proclaim the contribution one has made to broaden our understanding of a specific matter. Thus, we want to ensure ethical sourcing of the GAIA data by giving credit to the people and institutes that have worked hard to make the GAIA data accessible to the public. Therefore, throughout the report, a frequent acknowledgement will be dedicated to the European Space Agency, and there will be citations of their work.

Secondly, we are wary of possible bias in understanding the data. With a certain level of interest in astronomy, we have obtained a level of understanding of our Milky Way before this project. However, to objectively gain a new perception of open clusters, we do not desire manipulation/handling of data that results in changes in the data and pivot the findings towards our biased preferences. Thus, we are approaching the data set with an open mind and staying aware of the possible bias that might affect the conclusion of this project.

Thirdly, we are concerned about retaining the outcomes of this report until it is substantiated or approved by an authority. We believe it is ethical to wait for the validation of our findings because unvalidated scientific discoveries might result in misleading education to the public. Especially when dealing with a vast topic such as space, we must be mindful of the negative impacts of wrong interpretations and knowledge. As a result, we want to keep the project findings confidential. Team members must be aware of discussing this project with anyone outside of the team, and extra care must be taken regarding the devices and communication platforms used.

Lastly, preventing a wrong understanding of the GAIA data is ethical. This raises the question of whether the data and methods used can produce reliable results. According to the ESA, there are a few prevailing issues with the data set, most commonly seen as insufficient data (many missing values). Having this in mind helps us determine a better approach to the data by focusing on relevant variables that have plenty of entries.

There are few privacy and confidentiality concerns, as the data is not about people and has also been published by the data owner (ESA) for public scientific usage of the data .

Exploratory Data Analysis

Figure 1,2 - Check for missing values

```
##             ra            dec           pmra          pmdec       parallax
##              0              0              0              0              0
##          bp_rp phot_g_mean_mag    distparsecs            abM
##             18              0              0              0
```

There are 18 missing values in “bp_rp”. All other features in the pleiades dataset have no missing values.

```
##             ra            dec           pmra          pmdec       parallax
##              0              0              0              0              0
##          bp_rp phot_g_mean_mag    distparsecs            abM
##              1              0              0              0
```

There are 1 missing values in “bp_rp”. All other features in the m67 dataset have no missing values.

Figure 3 - Comparison of Means between Clusters

	Means of Pleiades Dataset	Means of M67 Dataset
ra	56.604875	132.851449
dec	24.133723	11.835734
pmra	19.914683	-10.962243

	Means of Pleiades Dataset	Means of M67 Dataset
pmdec	-45.360298	-2.926648
parallax	7.283866	1.157934
bp_rp	2.432774	1.129678
phot_g_mean_mag	15.157727	14.870049
distparsecs	143.780851	869.321879
abM	9.442023	5.181009

Figure 4 - Pleiades Correlation matrix

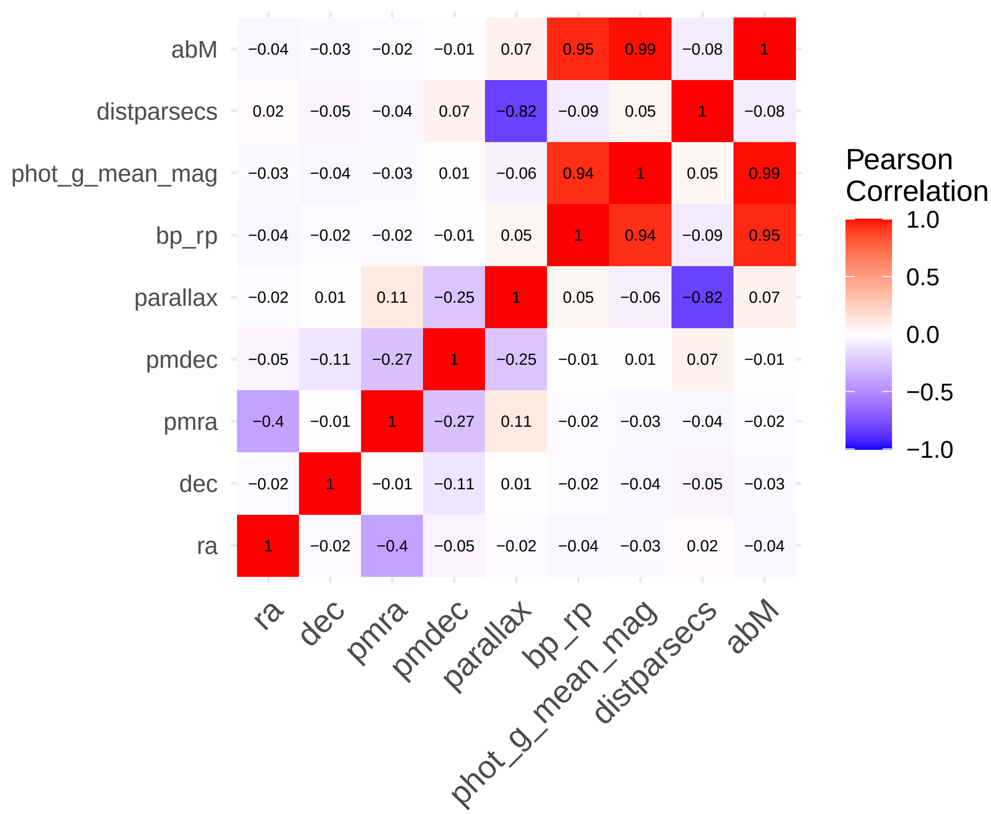
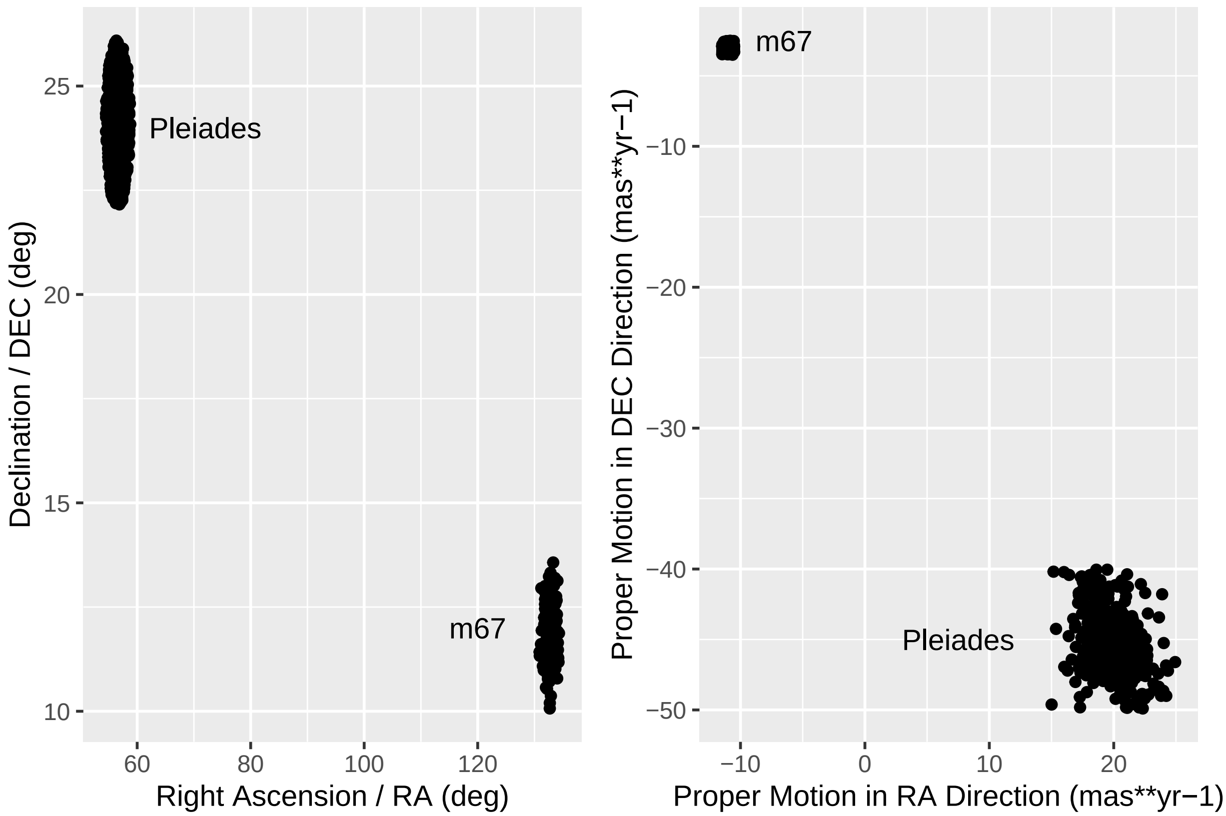


Figure 5 - Location and Proper Motion of Pleiades and m67

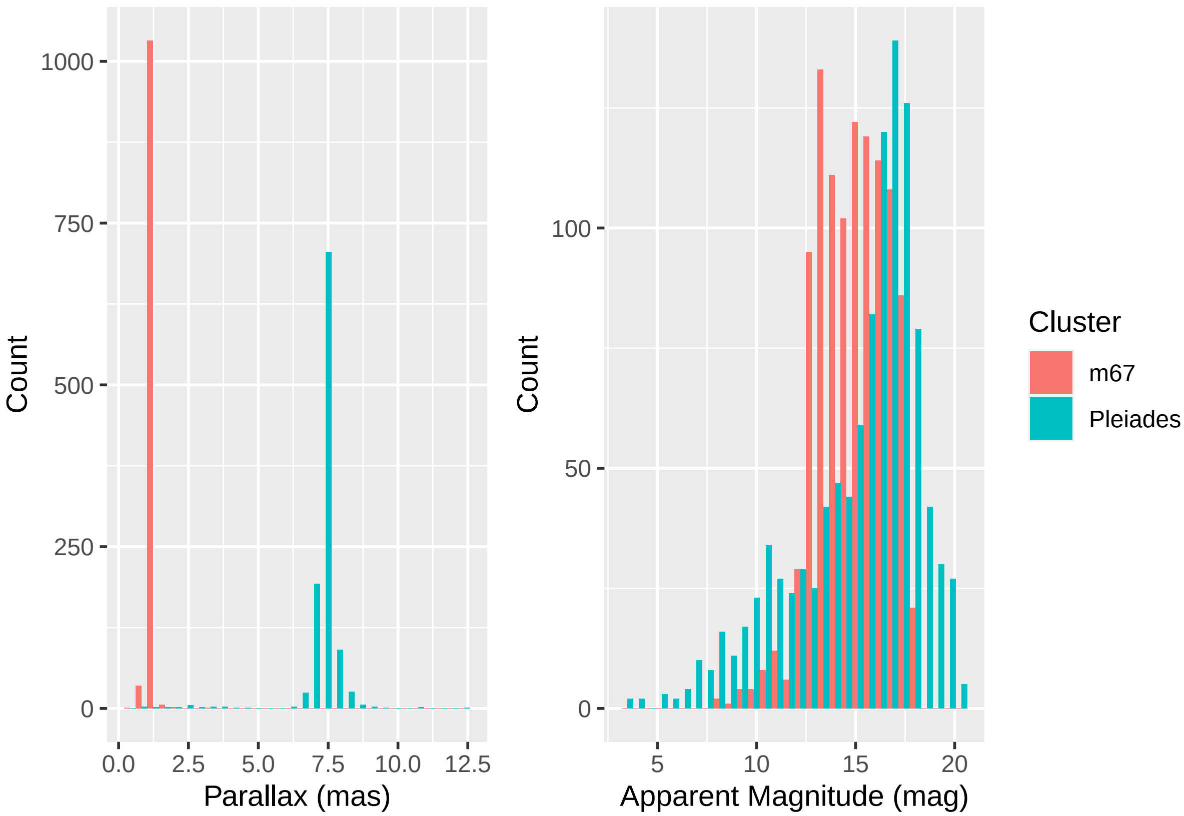


Right Ascension (RA) and Declination (DEC) are angular distances which are the astronomical equivalents of latitude and longitude. In other words, they are the position of stars within space.

The left-side scatterplot illustrate the position of the stars within space, using (ra) and (dec) variables. There is a clear difference between the two clusters on both RA and DEC, which shows that they are located in different positions in space. Through the distribution shown in space we can also see how Pleiades is bigger in size and is denser compared to M67.

Meanwhile, the right-side scatterplot illustrate the proper motion of the stars, using (pmra) and (pmdec) variables. The proper motions are astronomical equivalents of the movement of the stars in space. Looking at the right-side graph, there again is clear separation in the plots, representing the different directions the clusters are travelling in. The M67 cluster is travelling negatively in both RA and DEC, while the pleiades cluster is travelling positively in RA and more negatively in DEC (compared to the M67 cluster).

Figure 6 - Parallax and Apparent Magnitude of Pleiades and m67



Parallax is a measurement which is used to estimate the distance of the star from earth, using the observed displacement of the star caused by the change of the point of view.

Looking at the left-side histogram, we see that the m67 cluster has a large collection of stars with a parallax roughly around 1, which matches the mean value of 1.16 (2 d.p) in the summary table. The Pleiades cluster has a large collection of stars around 7.5, which is close to the mean value of 7.28 (2 d.p) in the summary table. The distribution of parallax measurements for the Pleiades cluster is wider than the m67 cluster, which might mean that the Pleiades cluster is widely spread or the data needs more filtering.

Apparent magnitude is the perceived brightness of a star, which relates to the brightness that we can see from earth. The lower the magnitude, the brighter the star. Both clusters appear to have quite similar Apparent magnitudes, with m67 possibly appearing slightly brighter. But accounting for the distance from earth of the stars, m67 will be even more intrinsically brighter compared to Pleiades, as the m67 cluster is further away and therefore will look less brighter than it actually is.

Figure 7 - Hertzsprung–Russell diagram between 2 open clusters

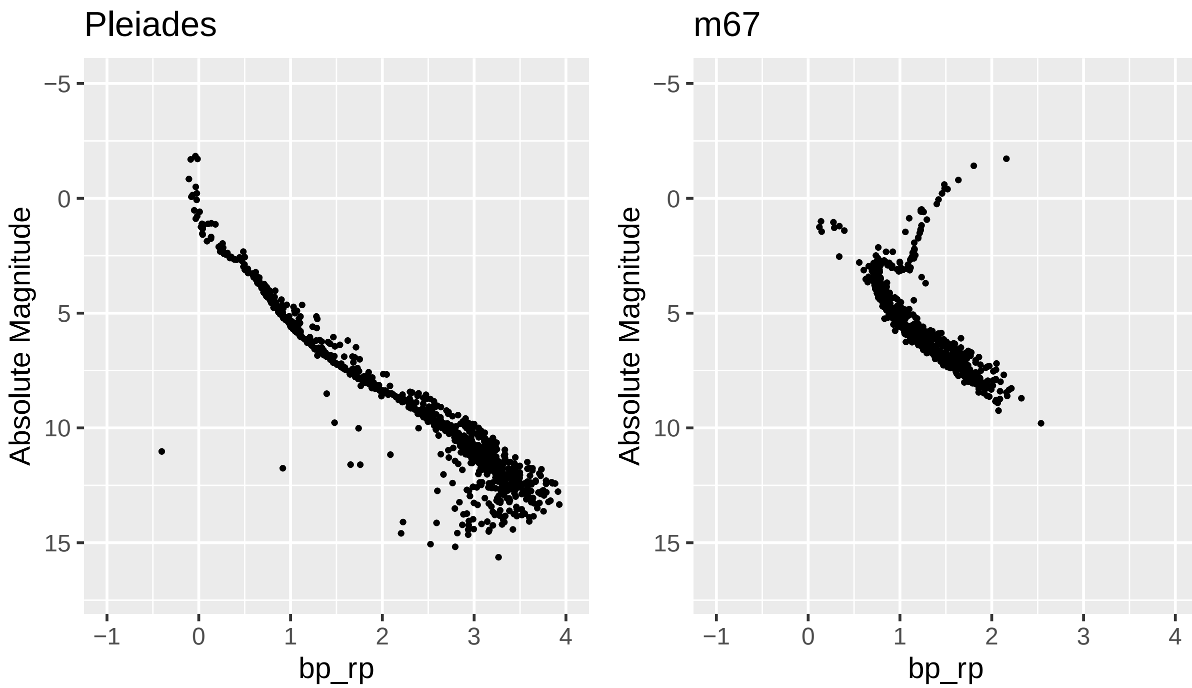


Figure X shows a colour-magnitude diagram of the two clusters. This shows the relationship between the Absolute Magnitude and BP-RP colours. Absolute magnitude is a term referring to the intrinsic brightness of a star, which was calculated using the distance of the star (parallax) with the observed colour from earth (phot_g_mean_mag or apparent magnitude). BP-RP colour represents the colors magnitude of the stars. A smaller BP-RP number indicates that the star appears bluer, and thus hotter; meanwhile a large BP-RP implies that the star has cooled off and is redder in appearance.

From this diagram we are able to estimate the age of the clusters (Palma, 2020). Pleiades has most of its stars on the 'main sequence', where they burn their hydrogen throughout their life time. On the other hand, many of the stars in M67 tails off of the main sequence, meaning they are dying stars that have already burnt out of hydrogen. As a consequence they move slower and fall behind. Therefore, we can conclude that the Pleiades is a much younger cluster that is still moving and burning; whereas M67 is older, and slowing down. Note that this age estimation is outside the scope of the EDA, but the plot is useful to guide future analysis.

Individual Contributions

Contribution across the team:

The team has had several meetings, in which we decided on our interest in the topic and identify the information we would like to explore. We also discussed suitable data handling methods and appropriate graphs. Moreover, as a team, we considered and agreed on our approach to our ethical and privacy consideration of our project. In detail, each member contributed by:

Taine assisted greatly in locating and extracting the data from the GAIA Archive. He is also the architect behind the scatter plots, histograms and HR diagram. Taine has also provided insightful comments on the plots, and supports other members to understand what the plots indicate.

Max conducted the summary view and missing value check for the data set. He also carried out the correlation matrix graphing and also presented comparison tables between the two clusters. Max has also provided thorough explanation on why we choose the used variables before he discussed missing values.

Van discussed the team's interest on the GAIA data set on behalf of the whole team, while she also elaborated on the GAIA missions and content briefs. She further reviewed the ethical consideration discussion, and further elaborated on some of Taine's interpretation on the plots. She also makes sure proper credits and acknowledgement is given to the ESA and ensure the reliability of the citations used.

References

Dr Christopher Palma. (2020). Measuring the Age of a Star Cluster. PennState. https://www.e-education.psu.edu/astro801/content/l7_p6.html

Héctor Cánovas. (2022). Use Cases. European Space Agency. https://www.cosmos.esa.int/web/gaia-users/archive/use-cases

Heyl, J., Caiazzo, I., & Richer, H. (2022). Reconstructing the Pleiades with Gaia EDR3. IOP. https://iopscience.iop.org/article/10.3847/1538-4357/ac45fc