GAIA_EDA

2022-09-03

Title • Project title, names and IDs of students in your project group, date of submission

Taine Murphy - 300472954
Van
Max

Background and Data [1-3 pages] • State which dataset(s) your group worked on, and their source

The dataset we used was the GAIA archive,

• Explain briefly why the dataset is of interest, or what questions it could be used to answer; assume that the reader has never heard of your dataset

• State the types of data in the dataset(s) and the structure of the dataset(s). Are the data numerical, categorical, or both? Time series? Coordinates? Diagnostic categories? This does NOT need to be an exhaustive list of every variable, just a few comments on the overall types.

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)

• State how complete the dataset(s) are (i.e. how many missing, any structure to the missing data, whether there are errors in the data)

• If you used more than one dataset, state what steps you had to take to integrate the datasets

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.

Ethics, Privacy and Security [1-2 pages]

• Brief discussion of any ethical considerations that apply to your project

- Brief discussion of any privacy concerns that might arise connected to your project

- Brief discussion of what steps you could take to keep your project data and results secure (you do NOT need to carry this out, you just need to talk about it in the report)

Exploratory Data Analysis [3-6 pages] For this section, do NOT try to summarize everything you can find in the dataset(s). Select a subset, highlighting features that you thought were interesting in the data. The plots do not have to be complicated; simple bar charts and scatter plots are fine.

- Several summary tables and/or plots, each describing one, two or three variables in the data that you thought were interesting • Explain the definitions of the variables in each table/plot • Comment on the main features of each plot • Include suitable labels and keys for each plot • Make sure all plots would be readable if printed in black & white, and adjust the point sizes and/or line thicknesses to improve readability • Lay out all tables so that they are clearly readable and clearly labelled, and do not use excessive significant figures

Open cluster subsets: m67 and Pleiades

#summary of datsets

```
##        ra             dec             pmra           pmdec       
##  Min.   :54.51   Min.   :22.16   Min.   :15.01   Min.   :-49.90  
##  1st Qu.:56.02   1st Qu.:23.57   1st Qu.:19.14   1st Qu.:-46.33  
##  Median :56.60   Median :24.12   Median :19.90   Median :-45.38  
##  Mean   :56.61   Mean   :24.14   Mean   :19.91   Mean   :-45.36  
##  3rd Qu.:57.20   3rd Qu.:24.67   3rd Qu.:20.67   3rd Qu.:-44.48  
##  Max.   :58.79   Max.   :26.09   Max.   :24.94   Max.   :-40.05  
##                                                                  
##     parallax           bp_rp         phot_g_mean_mag   distparsecs    
##  Min.   : 0.7815   Min.   :-0.4035   Min.   : 3.616   Min.   :  79.7  
##  1st Qu.: 7.2311   1st Qu.: 1.7362   1st Qu.:13.568   1st Qu.: 133.6  
##  Median : 7.3612   Median : 2.8521   Median :16.177   Median : 135.8  
##  Mean   : 7.2837   Mean   : 2.4328   Mean   :15.164   Mean   : 143.7  
##  3rd Qu.: 7.4878   3rd Qu.: 3.1662   3rd Qu.:17.309   3rd Qu.: 138.3  
##  Max.   :12.5475   Max.   : 4.2187   Max.   :20.471   Max.   :1279.6  
##                    NA's   :18                                         
##       abM        
##  Min.   :-1.837  
##  1st Qu.: 7.793  
##  Median :10.425  
##  Mean   : 9.449  
##  3rd Qu.:11.620  
##  Max.   :15.631  
## 
```

```
## 'data.frame':    1079 obs. of  9 variables:
##  $ ra             : num  58.3 58.3 58.1 58.5 58.5 ...
##  $ dec            : num  23 23.1 22.9 23.3 23.3 ...
##  $ pmra           : num  18.7 19.7 15 18.1 20.2 ...
##  $ pmdec          : num  -45.2 -46.5 -49.6 -42.8 -44.3 ...
##  $ parallax       : num  7.421 7.339 0.842 6.977 7.416 ...
##  $ bp_rp          : num  3.06 2.59 1.21 3.61 1.84 ...
##  $ phot_g_mean_mag: num  16.6 15.4 16.5 18.2 13.5 ...
##  $ distparsecs    : num  135 136 1188 143 135 ...
##  $ abM           : num  10.92 9.72 6.15 12.38 7.9 ...
```

```
##        ra             dec             pmra            pmdec        
##  Min.   :130.9   Min.   :10.07   Min.   :-11.47   Min.   :-3.499  
##  1st Qu.:132.7   1st Qu.:11.69   1st Qu.:-11.08   1st Qu.:-3.043  
##  Median :132.8   Median :11.82   Median :-10.97   Median :-2.926  
##  Mean   :132.9   Mean   :11.84   Mean   :-10.96   Mean   :-2.927  
##  3rd Qu.:133.0   3rd Qu.:11.96   3rd Qu.:-10.83   3rd Qu.:-2.804  
##  Max.   :134.3   Max.   :13.57   Max.   :-10.50   Max.   :-2.512  
##                                                                   
##     parallax         bp_rp        phot_g_mean_mag   distparsecs    
##  Min.   :0.591   Min.   :0.1226   Min.   : 7.948   Min.   : 309.2  
##  1st Qu.:1.124   1st Qu.:0.7884   1st Qu.:13.560   1st Qu.: 844.2  
##  Median :1.153   Median :1.0328   Median :14.996   Median : 867.3  
##  Mean   :1.158   Mean   :1.1297   Mean   :14.872   Mean   : 869.4  
##  3rd Qu.:1.185   3rd Qu.:1.4079   3rd Qu.:16.287   3rd Qu.: 890.0  
##  Max.   :3.234   Max.   :2.5372   Max.   :18.213   Max.   :1692.1  
##                  NA's   :1                                         
##       abM        
##  Min.   :-1.724  
##  1st Qu.: 3.872  
##  Median : 5.258  
##  Mean   : 5.183  
##  3rd Qu.: 6.538  
##  Max.   : 9.798  
## 
```

```
## 'data.frame':    1077 obs. of  9 variables:
##  $ ra             : num  133 133 132 132 133 ...
##  $ dec            : num  10.1 10.2 10.6 10.5 10.4 ...
##  $ pmra           : num  -10.7 -11 -10.5 -11.1 -11 ...
##  $ pmdec          : num  -2.62 -2.54 -2.53 -2.99 -2.83 ...
##  $ parallax       : num  1.17 1.07 1.13 1.31 1.14 ...
##  $ bp_rp          : num  1.071 1.715 0.822 1.906 1.163 ...
##  $ phot_g_mean_mag: num  15.4 17.3 14.4 17.5 15.7 ...
##  $ distparsecs    : num  854 934 889 763 877 ...
##  $ abM            : num  5.79 7.42 4.63 8.05 5.99 ...
```

```
##                 vars    n   mean    sd median trimmed   mad    min     max
## ra                 1 1077 132.85  0.35 132.85  132.85  0.21 130.91  134.32
## dec                2 1077  11.84  0.34  11.82   11.83  0.20  10.07   13.57
## pmra               3 1077 -10.96  0.18 -10.97  -10.96  0.19 -11.47  -10.50
## pmdec              4 1077  -2.93  0.18  -2.93   -2.92  0.18  -3.50   -2.51
## parallax           5 1077   1.16  0.11   1.15    1.15  0.04   0.59    3.23
## bp_rp              6 1076   1.13  0.38   1.03    1.09  0.39   0.12    2.54
## phot_g_mean_mag    7 1077  14.87  1.74  15.00   14.93  2.03   7.95   18.21
## distparsecs        8 1077 869.40 69.91 867.31  867.29 33.96 309.19 1692.12
## abM                9 1077   5.18  1.76   5.26    5.22  2.01  -1.72    9.80
##                   range  skew kurtosis   se
## ra                 3.41 -0.30     4.67 0.01
## dec                3.50  0.21     4.13 0.01
## pmra               0.97 -0.11    -0.19 0.01
## pmdec              0.99 -0.22    -0.02 0.01
## parallax           2.64  7.67   140.98 0.00
## bp_rp              2.41  0.70    -0.28 0.01
## phot_g_mean_mag   10.27 -0.42     0.03 0.05
## distparsecs     1382.92  2.15    30.68 2.13
## abM               11.52 -0.33     0.06 0.05
```

#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.

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
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

#correlation of pleiades

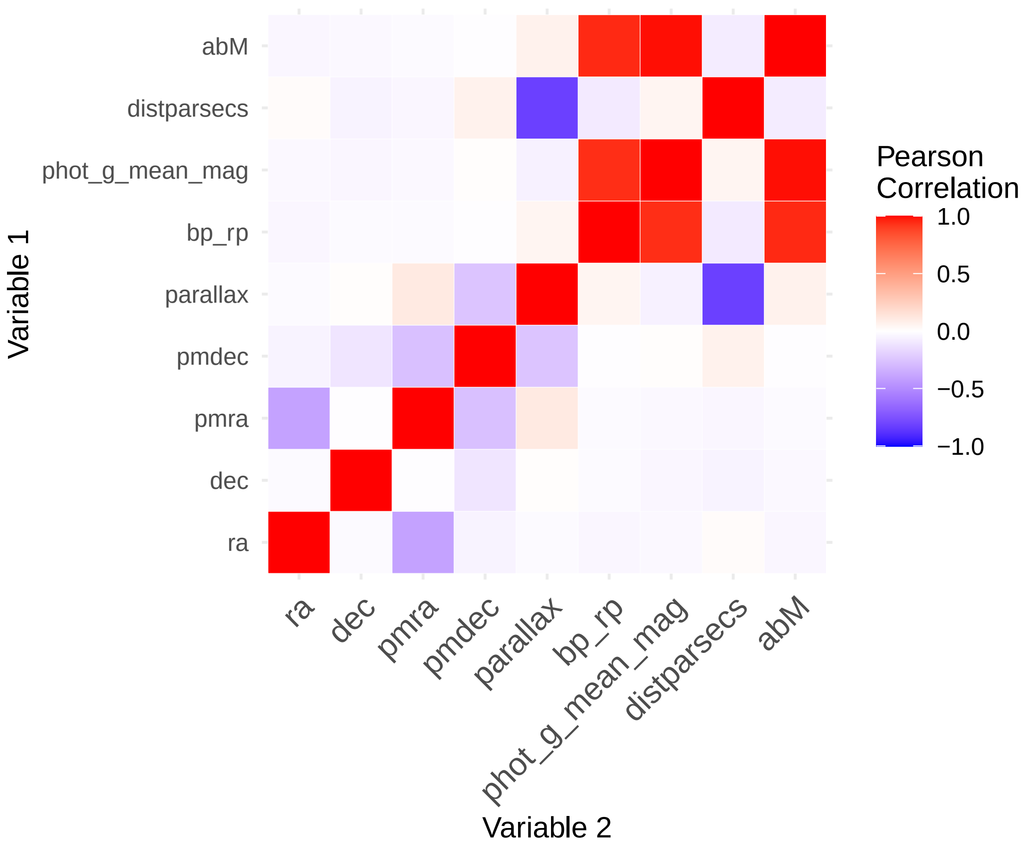
```
##
## -------------------------------------------------------------------------
##        &nbsp;           ra      dec     pmra    pmdec   parallax   bp_rp
## --------------------- ------- ------- ------- ------- ---------- -------
##        **ra**            1     -0.02   -0.4    -0.05    -0.02     -0.04
##
##        **dec**         -0.02     1     -0.01   -0.11     0.01     -0.02
##
##       **pmra**         -0.4    -0.01     1     -0.27     0.11     -0.02
##
##       **pmdec**        -0.05   -0.11   -0.27     1      -0.25     -0.01
##
##     **parallax**       -0.02   0.01    0.11    -0.25      1        0.05
```

```
##
##       **bp_rp**        -0.04   -0.02   -0.02   -0.01     0.05       1
##
##  **phot_g_mean_mag**   -0.03   -0.04   -0.03   0.01     -0.06     0.94
##
##    **distparsecs**     0.02    -0.05   -0.04   0.07     -0.82     -0.09
##
##        **abM**         -0.04   -0.03   -0.02   -0.01     0.07     0.95
## ------------------------------------------------------------------------
##
## Table: Table continues below
##
##
## -------------------------------------------------------------
##        &nbsp;          phot_g_mean_mag   distparsecs    abM
## --------------------- ----------------- ------------- -------
##        **ra**               -0.03           0.02       -0.04
##
##        **dec**              -0.04           -0.05      -0.03
##
##       **pmra**              -0.03           -0.04      -0.02
##
##       **pmdec**             0.01            0.07       -0.01
##
##     **parallax**            -0.06           -0.82      0.07
##
##       **bp_rp**             0.94            -0.09      0.95
##
##  **phot_g_mean_mag**          1             0.05       0.99
##
##    **distparsecs**          0.05              1        -0.08
##
##        **abM**              0.99            -0.08        1
## -------------------------------------------------------------
```

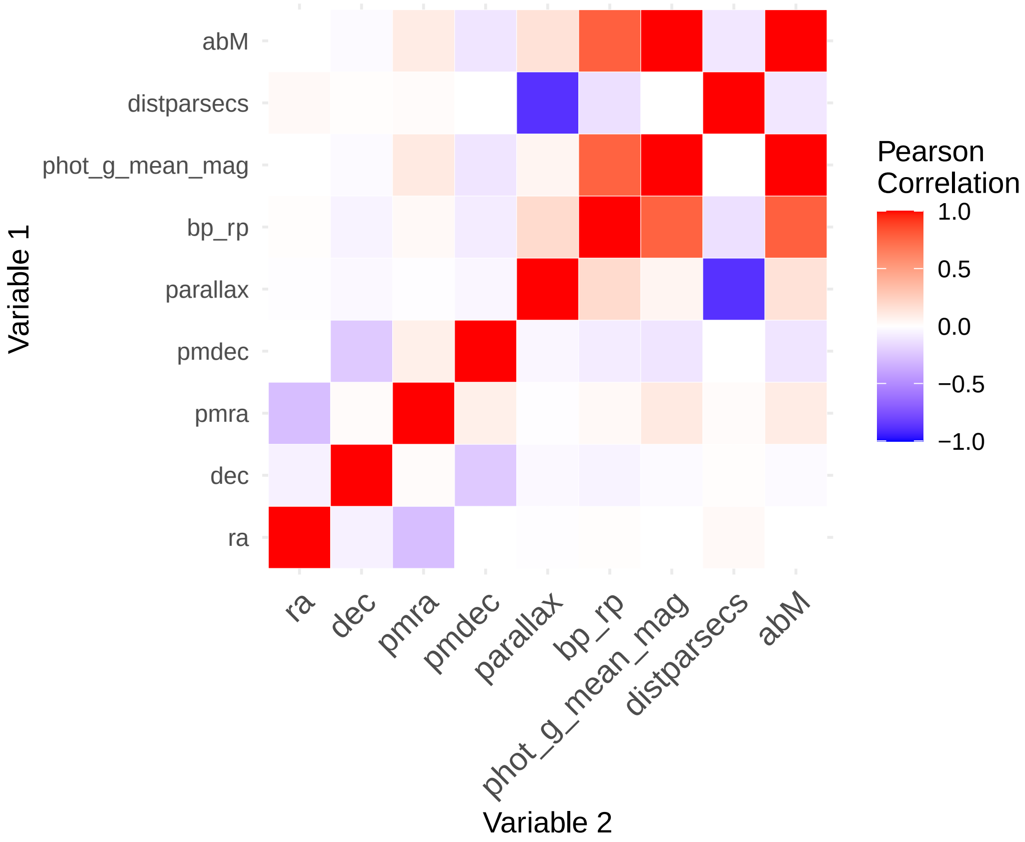
#correlation of m67

```
##
## ------------------------------------------------------------------------
##        &nbsp;           ra      dec    pmra    pmdec   parallax   bp_rp
## --------------------- ------- ------- ------- ------- ---------- -------
##        **ra**            1     -0.06   -0.28     0      -0.01     0.01
##
##        **dec**        -0.06     1     0.02    -0.23    -0.03     -0.05
##
##       **pmra**        -0.28   0.02      1     0.08     -0.01     0.03
##
##       **pmdec**         0     -0.23   0.08      1      -0.04     -0.08
##
##     **parallax**      -0.01   -0.03   -0.01   -0.04      1       0.19
##
##       **bp_rp**       0.01    -0.05   0.03    -0.08     0.19       1
##
##  **phot_g_mean_mag**     0     -0.02   0.11    -0.11     0.05     0.77
```

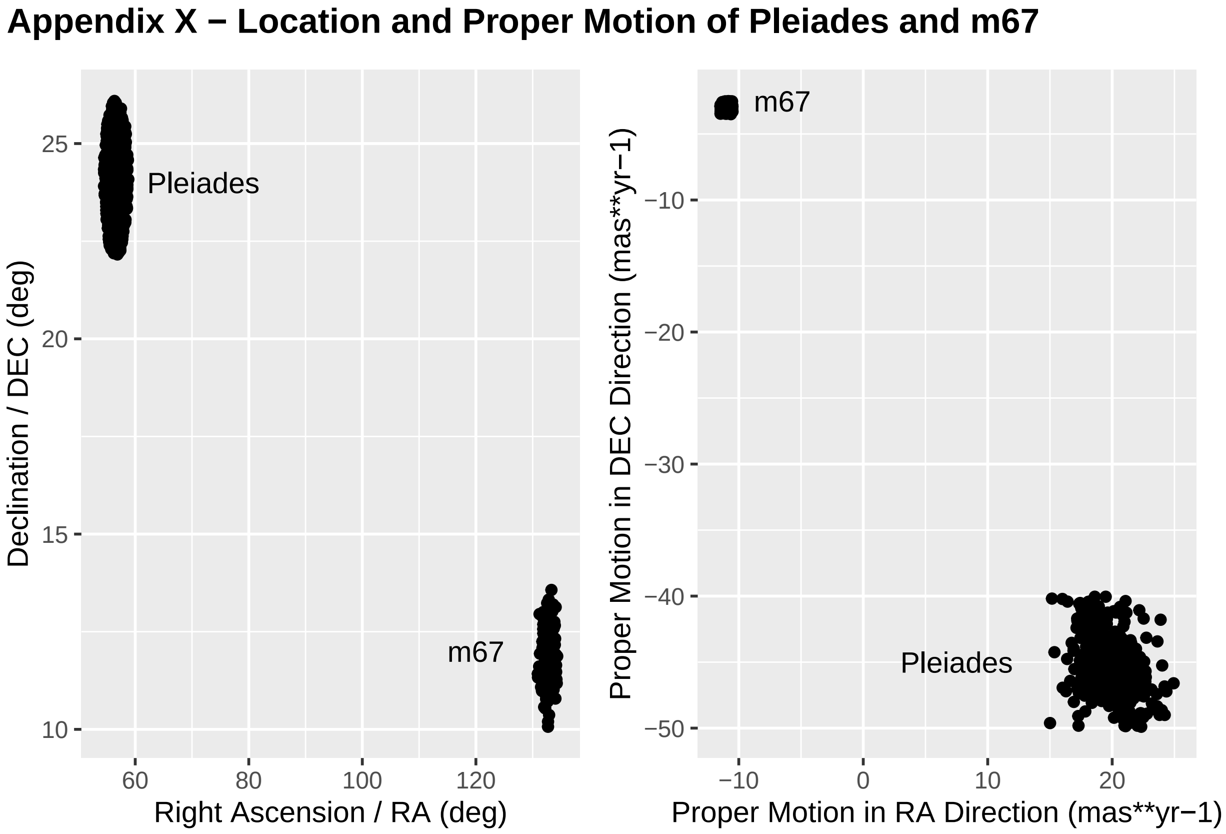
```
##
##    **distparsecs**     0.03    0.01    0.02      0      -0.88     -0.13
##
##        **abM**           0     -0.02    0.1    -0.11     0.15     0.78
## ------------------------------------------------------------------------
##
## Table: Table continues below
##
##
## -------------------------------------------------------------
##        &nbsp;          phot_g_mean_mag   distparsecs    abM
## --------------------- ----------------- ------------- -------
##        **ra**                 0             0.03         0
##
##        **dec**              -0.02           0.01       -0.02
##
##       **pmra**              0.11            0.02        0.1
##
##       **pmdec**             -0.11             0        -0.11
##
##     **parallax**            0.05            -0.88      0.15
##
##       **bp_rp**             0.77            -0.13      0.78
##
##  **phot_g_mean_mag**          1               0          1
##
##    **distparsecs**            0               1        -0.1
##
##        **abM**                1             -0.1         1
## -------------------------------------------------------------
```



#correlation matrices


abM
distparsecs
phot_g_mean_mag
bp_rp
parallax
pmdec
pmra
dec
ra
Variable 1
ra
dec
pmra
pmdec
parallax
bp_rp
phot_g_mean_mag
distparsecs
abM
Variable 2
Pearson
Correlation
1.0
0.5
0.0
−0.5
−1.0

Display pmra/pmdec plot

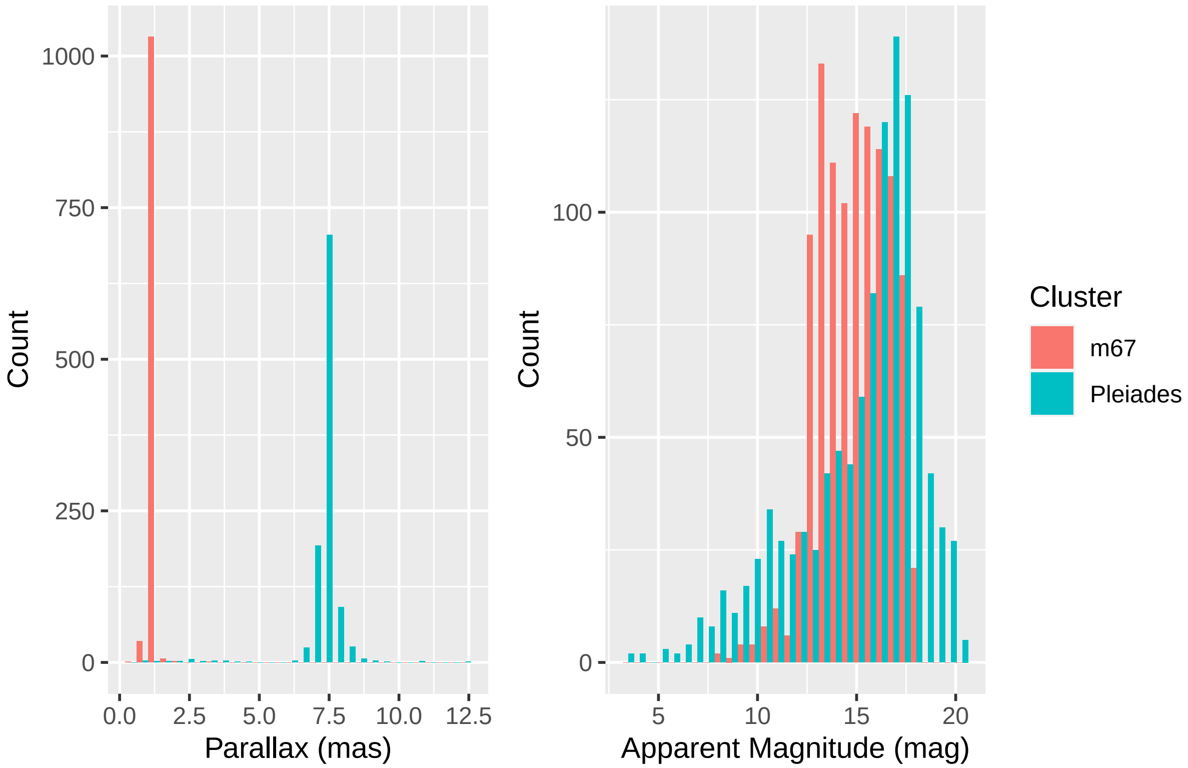


Right Ascension (RA) and Declination (DEC) are angular distances which are the astronomical equivalents of latitude and longitude. In other words, they are the position of stars within space.

Looking at the left-side graph, 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.

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).

Appendix X − 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

HR diagram - use ages of cluster with diagram showing the comparison

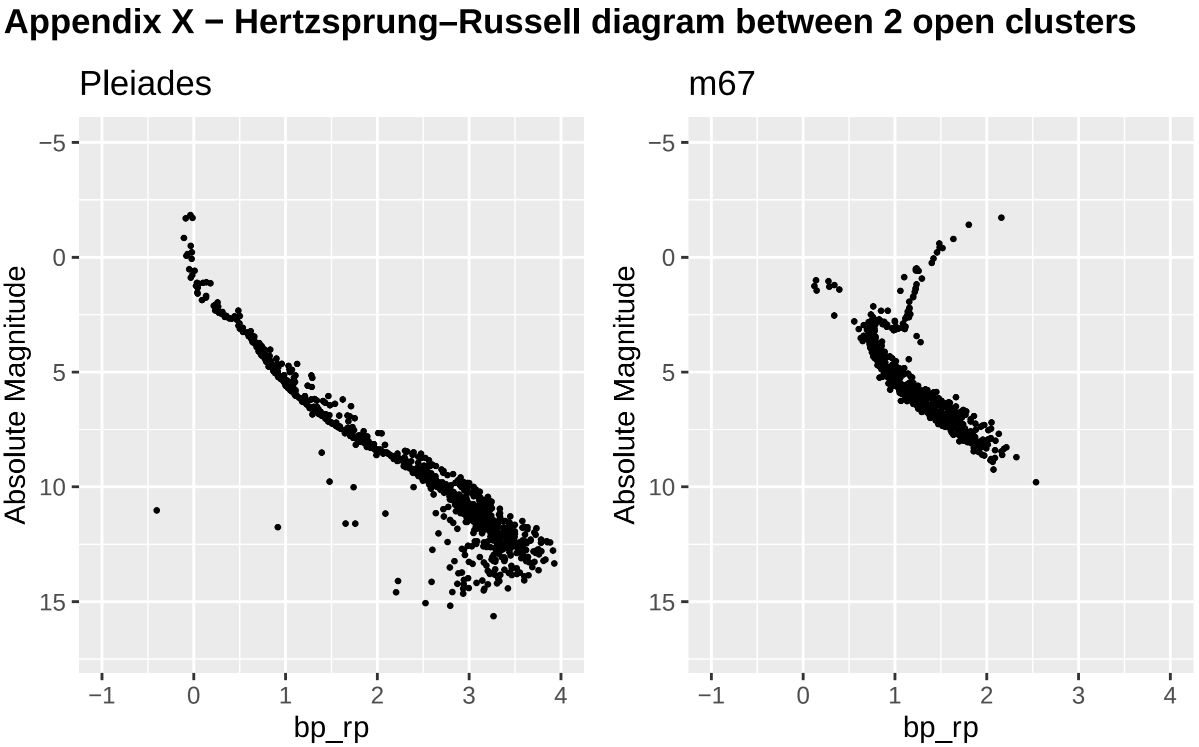


Figure X shows a colour-magnitude diagram of the two clusters. This shows the relationship between the Absolute Magnitude and BP-RP colours. Absolute magnitude is a term referring to the intrinsic brightness of a star, which was calculated using the distance of the star (parallax) with the observed colour from earth (phot_g_mean_mag or apparent magnitude). BP-RP colour represents the

From this diagram we are able to estimate the age of the clusters (reference). Pleiades is a young open cluster, meaning it has a lot of its stars on the 'main sequence'. On the other hand, m67 is a much older cluster, meaning the stars near the upper ends appear to tail off . Note that this is outside the scope of the EDA, but is useful to show to guide future analysis.

Individual Contributions [1 page]

- State what contribution each member of the group made to the data preparation, the analysis and the report

Overall Report These marks will be awarded for overall presentation, clarity and quality of the report. In particular, marks will be awarded for

- A clear logical layout
- Keeping to the page limits for each section, and using sensible font size
- Key facts being easily located
- Readability of tables and plots DATA 301 T2 2022 5
- Clarity of expression [Note: for non-native speakers of English: your English does not need to be perfect, it is the logic and correctness of your presentation that is most important. Nevertheless you are advised to get someone to proof-read your proposal.]

- Clear explanation of how your choice of exploratory plots and tables is relevant to your project, and how the ethical considerations apply to your project (i.e. not just a set of generalities)

- Make sure each time you use/refer to someone else's work you cite the source in the text, and include the reference in the list at the end. It does not need to be a long list; you may only need one or two references.

- Referencing should be correctly done: a complete list of references must be included. You can use any referencing style you wish; APA is fine if that's what you like.

Total: 35 marks

You will be expected to include a revised version of the Background, Ethics and EDA sections in the final project report; you do not have to rewrite those sections from scratch. You will be expected to consider any feedback you have received for this first report when revising it for the final report, and this will be taken into account when the final report is marked.