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.

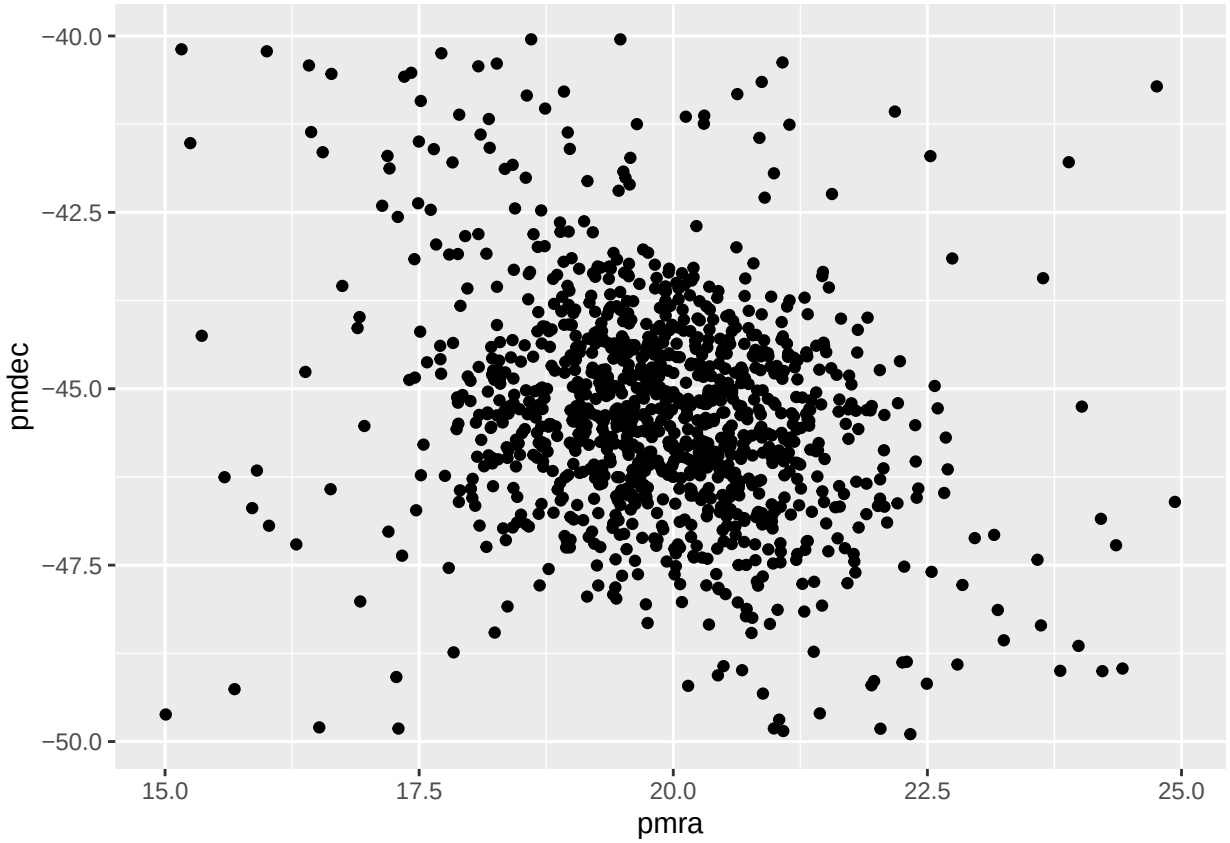
• Several summary tables and/or plots, each describing one, two or three variables in the data that you thought were interesting • Explain the definitions of the variables in each table/plot • Comment on the main features of each plot • Include suitable labels and keys for each plot • Make sure all plots 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

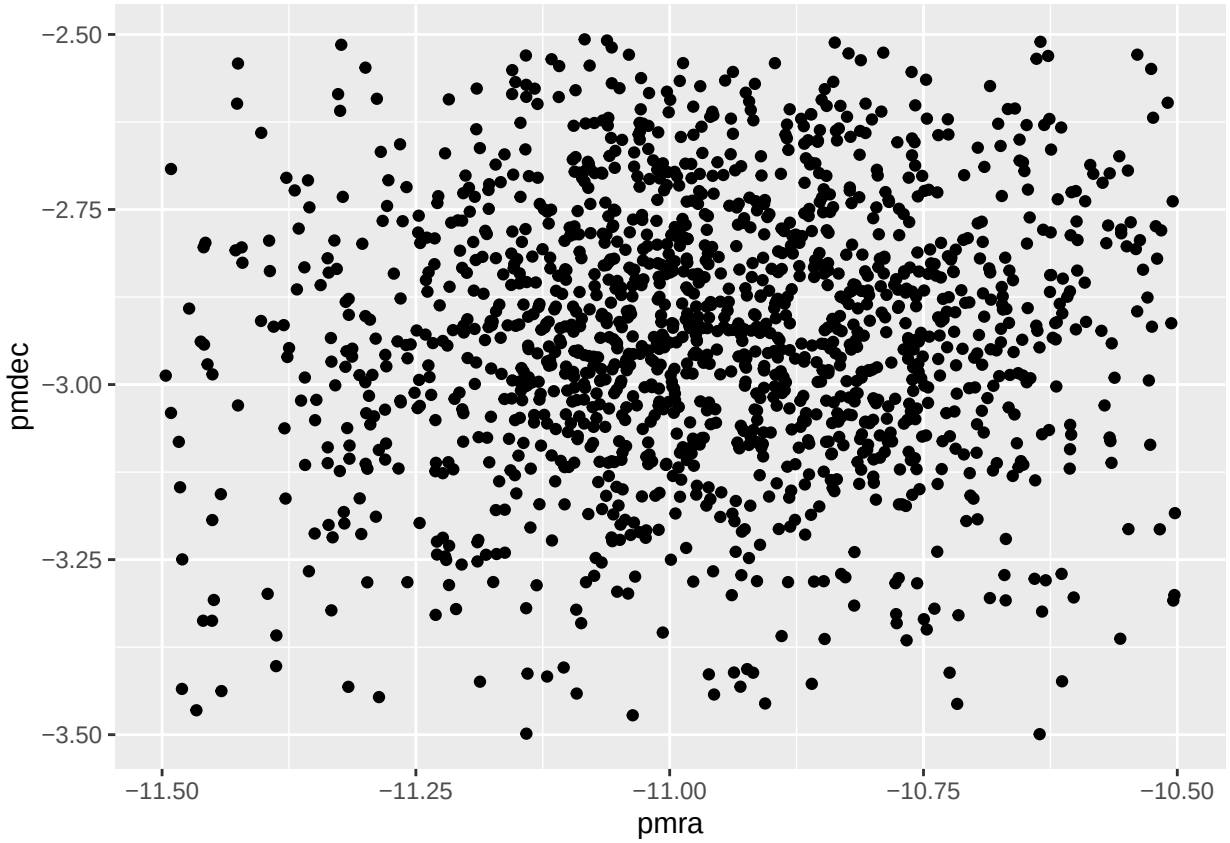
```
pleiades <- read.csv("Pleiades_4.csv")
pleiades$distparsecs <- 1/(pleiades$parallax/1000)
pleiades$abM <- pleiades$phot_g_mean_mag - (2.5 * log((pleiades$distparsecs/10)^2,10))


m67 <- read.csv("m67.csv")
m67$distparsecs <- 1/(m67$parallax/1000)
m67$abM <- m67$phot_g_mean_mag - (2.5 * log((m67$distparsecs/10)^2,10))
```

##Display pmra/pmdec plot

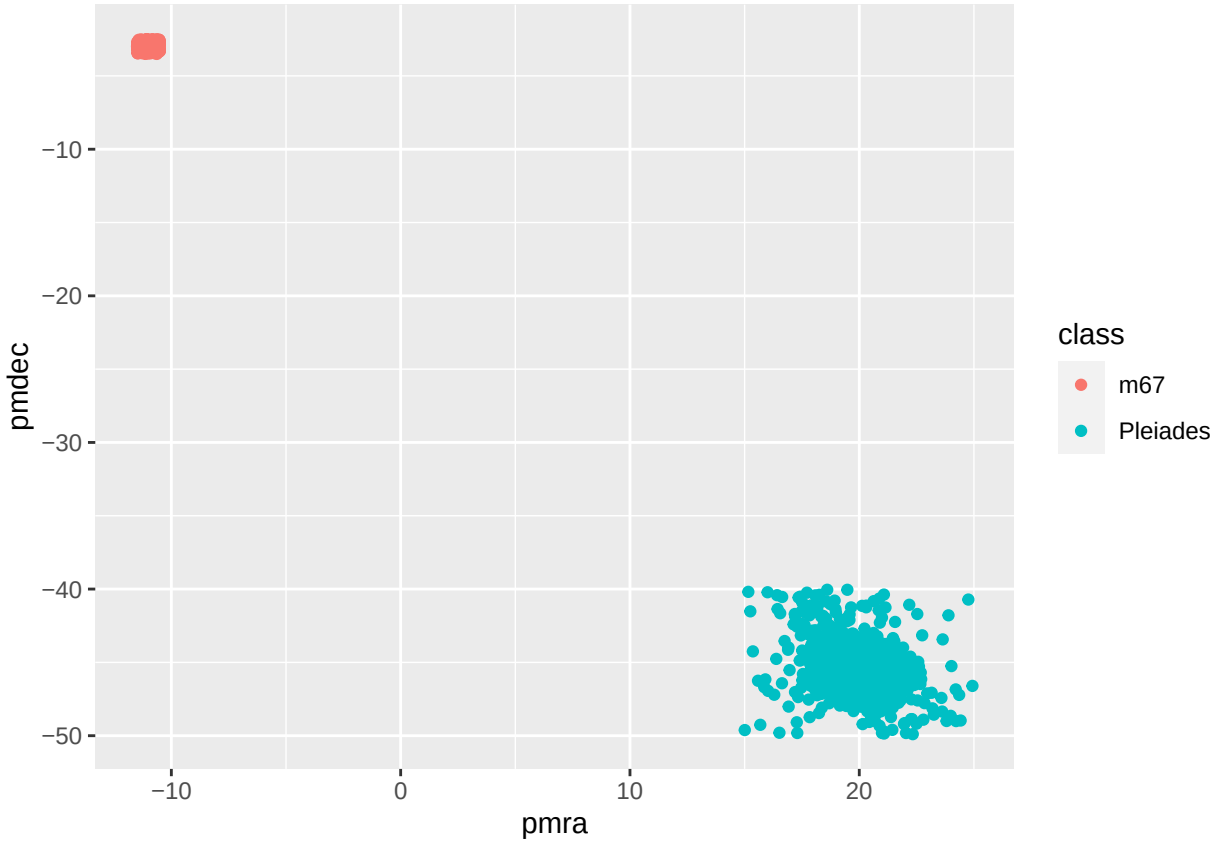
```
ggplot(pleiades, aes(x=pmra, y=pmdec )) +
  geom_point()
```




```
ggplot(m67, aes(x=pmra, y=pmdec )) +
  geom_point()
```




```
pleiades$class <- "Pleiades"
m67$class <-"m67"
merge <- rbind(pleiades, m67)

ggplot(merge, aes(x=pmra, y=pmdec )) +
  geom_point(aes(color =class ))
```




```
#ggplot(merge, aes(x=ra, y=dec )) +
  geom_point(aes(color =class ))
```

```
## mapping: colour = ~class
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
#ggplot(pleiades, aes(x = parallax)) + geom_histogram(position = "dodge")
```

TODO comparison of some means of variables

```
as.data.frame(apply(pleiades[][5:10],2,summary))
```

```
##         ref_epoch       ra   ra_error      dec   dec_error    parallax
## Min.         2016 54.50859 0.01201166 22.15381 0.007888133 -0.05737493
## 1st Qu.      2016 56.01916 0.02990496 23.56314 0.019903165  7.21938489
## Median       2016 56.60733 0.06329263 24.13237 0.043969543  7.35524115
## Mean         2016 56.61133 0.11058270 24.13899 0.077044271  7.18322200
## 3rd Qu.      2016 57.21096 0.11957908 24.67572 0.084959377  7.48571556
## Max.         2016 58.78746 1.25941940 26.09027 1.044534100 12.54750850
```

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

```
a <- ggplot(pleiades, aes(x=bp_rp, y=abM)) +
  geom_point(size= 0.5)+ scale_y_reverse()+ xlim(-1,4) + ylim(17,-5) + coord_fixed(0.25) + ggtitle("Ple
  xlab("bp_rp") + ylab("Absolute Magnitude")
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

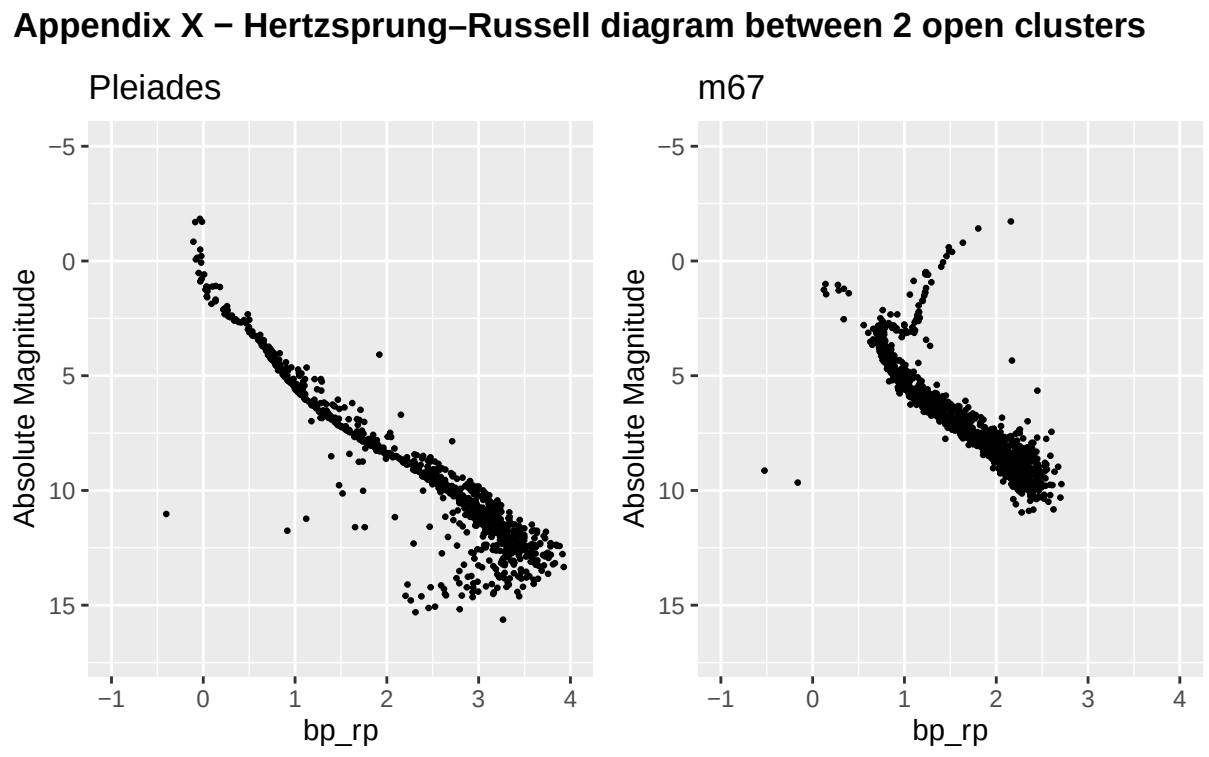
```
b <- ggplot(m67, aes(x=bp_rp, y=abM)) +
  geom_point(size= 0.5)+ scale_y_reverse() + xlim(-1,4) + ylim(17,-5) + coord_fixed(0.25) + ggtitle("m67
  xlab("bp_rp") + ylab("Absolute Magnitude")
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

```
t <- theme(plot.title = element_text(face="bold"))
a + b + plot_annotation(title = "Appendix X - Hertzsprung-Russell diagram between 2 open clusters", ther
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



```
#ggplot(merge, aes(x=bp_rp, y=abM)) +
#  geom_point(aes(color = class))+ scale_y_reverse()
```

IF MORE NEEDED:

- Matariki constellation subset

Different to the pleiades subset since some of the stars are missing proper motions/ parallax

- Stars vs Quasars vs Galaxies

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.