



# ASTROCHALLENGE 2022

## SENIOR DATA ANALYSIS ROUND

Monday 14<sup>th</sup> March 2022

**PLEASE READ THESE INSTRUCTIONS CAREFULLY.**

This paper consists of **6** printed pages, including this cover page.

In this question you will receive a folder on Google Drive named "Astrochallenge 2022" with all the necessary documents needed. You will process this data set, analyse it, observe trends, and draw conclusions. **There are no right or wrong answers**; you will be marked solely on the quality of your analysis, even if your statistical methods are incorrect.

We **strongly** recommend you use the tools as highlighted within the paper to complete your task.

© National University of Singapore Astronomical Society  
© Nanyang Technological University Astronomical Society

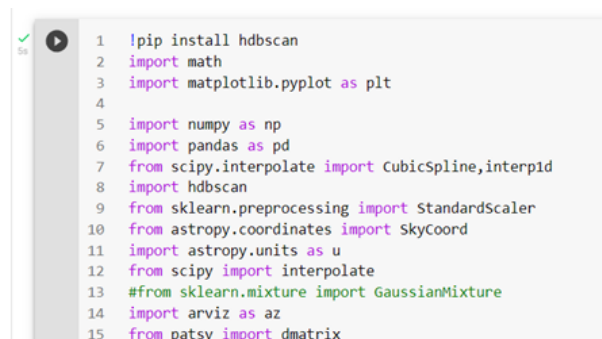
# 1 Instructions

## 1.1 How to Use the document

The Main document in this folder is called "Astrochallenge DAQ.ipynb". In order to open this file you will need to Right Click the file: **Select: Open With > Select: Connect to More Apps > Search For: "Google Colaboratory"** and install the corresponding app.

To run the code in this file you have to first copy the file to your own private google drive. DO NOT change the name of the folder or any file unless you are experienced. Open "Astrochallenge DAQ.ipynb" file using Google Colaboratory App installed earlier.

You should see some text which will explain to you the details of the question. There will be some blocks which does not look like plain text. These are coding blocks like you see below:

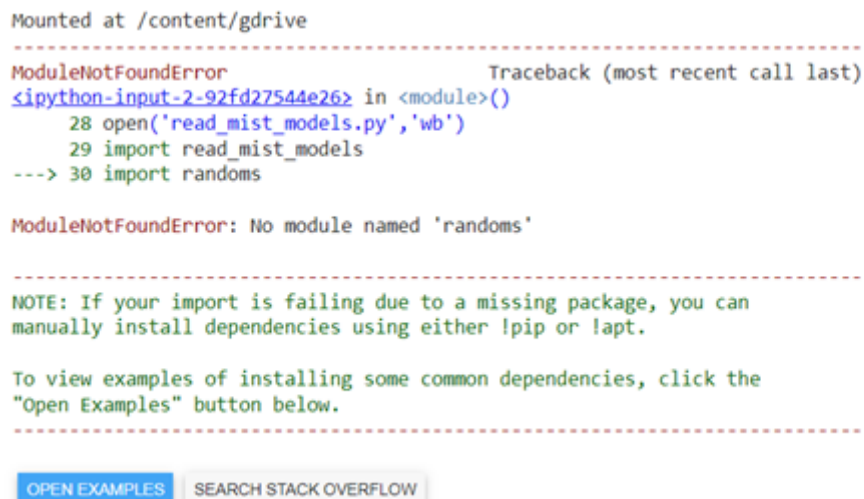


```
1 !pip install hdbscan
2 import math
3 import matplotlib.pyplot as plt
4
5 import numpy as np
6 import pandas as pd
7 from scipy.interpolate import CubicSpline,interp1d
8 import hdbscan
9 from sklearn.preprocessing import StandardScaler
10 from astropy.coordinates import SkyCoord
11 import astropy.units as u
12 from scipy import interpolate
13 #from sklearn.mixture import GaussianMixture
14 import arviz as az
15 from patsy import dmatrix
```

Figure 1: Example of code blocks

To run the code, press the play button beside each block of code. Do it in sequence from the top. When the code is done running you will get a tick mark beside each ran code.

If you see any message resembling the below image, email us for further assistance. (Alternatively, if you are particularly experienced in coding, you may troubleshoot and debug yourself)



```
Mounted at /content/gdrive
-----
ModuleNotFoundError                                Traceback (most recent call last)
<ipython-input-2-92fd27544e26> in <module>()
    28 open('read_mist_models.py','wb')
    29 import read_mist_models
--> 30 import randoms

ModuleNotFoundError: No module named 'randoms'

-----

NOTE: If your import is failing due to a missing package, you can
manually install dependencies using either !pip or !apt.

To view examples of installing some common dependencies, click the
"Open Examples" button below.
-----

OPEN EXAMPLES SEARCH STACK OVERFLOW
```

Figure 2: Example of an error you could experience

After running certain codes, Excel files will appear in your folder. You should download and use them plot the necessary graphs.

## 1.2 Deliverables

- Your answers to all questions in the ipynb file (Questions are in bold), either in the same file or in a separate pdf document;
- Three plots of HR Diagram as requested by the questions; and
- Any not mentioned points you have discovered while reading the explanation (for extra credit at the marker's discretion)

## 2 Data Analysis Question

In this question we will explore how an astronomer would use databases such as the Sloan Deep Sky Survey (SDSS) and the Gaia Deep Sky Survey to find out the age and distance of a star cluster.

Today, we will be using the star cluster M67, the oldest known open cluster, as our example.

### 2.1 Introduction

Star clusters have long been regarded as powerful tools for astronomers. It is often used by astronomers to determine the distance an object is away from us. Star clusters have also been long thought to be precursors to modern galaxies. Thus studying them can yield a better understanding of the formation of galaxies as well as their evolution.

The oldest star clusters are also used sometimes to gauge limits on the age and the evolution history of the universe. This is because many are formed close to the beginning of the universe. Thus star clusters are studied extensively in cosmology to verify the accuracy of our cosmological models of the universe.

In order to study star clusters, we first need to determine which stars are in that cluster itself. This is usually done using a clustering algorithm. In broad strokes, clustering algorithms assign stars into clusters. Should the algorithm identify multiple clusters within the area of sky specified, the cluster with the most members is defined as the star cluster. This method is often done using information from big surveys like the Gaia mission. Such datasets usually include parameters such as the star's proper motion, photometric magnitudes, their position relative to stars with similar right ascension and declination.

### 2.2 Objectives

The aims of this question are to:

- Filter the data for stars that belong to the cluster M67; and
- Determine the age of the cluster by using MISE isochrone fitting method

### 2.3 Theoretical Background

The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is one of the methods used by astronomers to determine if a particular star is in a cluster. This method is a variation of an algorithm called Friend of Friends (FOF) which decides that a star is in the cluster if the star is within a certain "distance" away from another star already in the cluster. The value of this "distance" is usually called the linking length.

**(a) What are the astrometric parameters you should consider when determining the linking length? (2 marks)**

### 2.4 Data

The data that will be used is from Gaia Early Data Release 3 ("Gaia eDR3"). This release from the ESA Gaia space mission is by far the deepest and most precise astrometric catalogue ever obtained.

The data you need is already downloaded in the code folder. You do not need to download any other data. The dataset contains data within 180 arcmins from the cluster center. Next, we select the sources that satisfy the following criteria:

- Each source must have the five astrometric parameters, positions, proper motions, and parallax as well as valid measurements in the three photometric passbands  $G$ ,  $G_{BP}$ , and  $G_{RP}$  in the Gaia eDR3 catalogue
- Each source's parallax value must be non-negative
- The errors in each source's G-mag must be less than 0.005

(b) Why are the above conditions necessary to ensure that the clustering of stars can be done properly? (3 marks)

#### 2.4.1 Determine the Center of the Open Cluster

To determine the membership of open cluster M67, we will use a module in python called HDBSCAN . The Collab Notebook contains more details on how exactly to do this.

#### 2.4.2 Data Preparation

To eliminate sources with high uncertainties and to retain a fraction of the sources down to  $G \sim 21$  mag, we need to select the sources with G-mag errors of less than 0.005. This is done by using the formulae given. Using the given formulae, calculate the error of  $G(|\sigma_G|)$ ,  $G_{BP}(|\sigma_{BP}|)$ , and  $G_{RP}(|\sigma_{RP}|)$ , which are the associated uncertainty for the G Band, BP band and RP band magnitude of the star's spectrum respectively.(each band refers to a specific wavelength range).

$$G(|\sigma_G|) = \frac{2.5}{\ln 10} \frac{\sigma_G}{F_G} \quad (1)$$

$$G(|\sigma_{RP}|) = \frac{2.5}{\ln 10} \frac{\sigma_{RP}}{F_{RP}} \quad (2)$$

$$G(|\sigma_{BP}|) = \frac{2.5}{\ln 10} \frac{\sigma_{BP}}{F_{BP}} \quad (3)$$

(c) Given the general formula for the error in the expression  $y = f(x)$  for some function  $f(x)$  is:

$$|\sigma_y| = \frac{df(x)}{dx} \sigma_x \quad (4)$$

show that the expressions (1)-(3) are the errors in magnitudes associated with the G band, BP band and RP band from the error of their respective fluxes  $F_G, F_{BP}$  and  $F_{RP}$ . (3 marks)

(d) Plot a color magnitude diagram for the file: *"M67\_colour\_magnitude\_unfiltered.xlsx"*. By explaining the features of a star cluster in a color magnitude diagram, determine if this diagram is representative of a star cluster. (2 marks)

Note: You can also edit the code in the previous cell to plot the color magnitude diagram in the given code file.

(e) Plot a color magnitude diagram of the file: *"M67\_colour\_magnitude\_hdbfiltered.xlsx"* and explain what is/are the reason(s) for the differences in appearance between this and the unfiltered data. (3 marks)

The code in the code file should enable you to find the center of the center of the Star Cluster.

#### 2.4.3 Isochrone Plotting

We are now in a position to try fit our filtered data into a model of the star cluster. In this case we use something called the isochrone model. This simulates a group of stars of different masses but with the same chemical composition and age. By comparing the isochrone of different ages with the color magnitude diagram of the cluster, we can determine the age of the cluster.

(f) Why is it that we use a group of stars with the same chemical composition but different masses as a model for the star cluster? What does it suggest about the nature of stellar formation? (2 marks)

To determine how good the isochrone model fits with the star cluster, we measure the Residual Mean Square Error(RMSE) given by:

$$\sigma = \sqrt{\sum_{i=1}^n \frac{(x_{\text{predicted}} - x_{\text{experimental}})_i^2}{n}} \quad (5)$$

where  $x_{\text{predicted}}$  is the predicted value of the variable (in this case the G band magnitude) from the isochrone model, and the  $x_{\text{experimental}}$  is the observed data from the filtered star cluster data.

(g) What does this RMSE value measure in terms of how good the model fits to the data? How do the RMSE values of models which fit very well to the data compare to those that do not fit very well? (2 marks)

(h) The file "M67\_RMSE\_age{age in log scale}.xlsx" should contain the model data and the experimental data for the same color index range for different times of evolution of the cluster. For isochrones of different ages, calculate the RMSE value using the formulae given and determine the cluster age of which best fits the cluster? (3 marks)

Note: Please show the RMSE value obtained in your final submission.