Milestone 2: Data analysis

As an early-career researcher, the bulk of your time will be spent *analyzing data*. Exactly what this analysis entails will vary from project to project, but you'll almost always be doing some combination of writing code, interacting with data, and visualizing your results. No textbook can possibly cover everything you need to know or every situation you'll encounter, so the best way to develop your data analysis skills is just to practice *a lot.* For Milestone 2, we'll do exactly that by performing some simple analysis and visualization tasks on real astronomical data.

Due date: end-of-day on Thursday, 4/17

Learning objectives

This assignment is designed to help you:

1. **Develop familiarity with different kinds of astronomical data**, including how to retrieve it from online databases and how to load it into a Python script or notebook.
2. **Practice writing clear Python code** to clean, organize, and perform simple analyses of astronomical data.
3. **Practice creating meaningful visualizations** to present/interpret key research results.
4. **Synthesize results with the broader context** established during your literature review (Milestone 1) to understand the impact of your work on the field and potential next steps.

Task description

For each pre-defined project, I will provide a Jupyter Notebook with background information, a description of the relevant astronomical dataset, and a set of analysis tasks. The exact analysis tasks will vary, but each project will require the steps listed below.

If you proposed your own project: we'll work together to pick a dataset and define the analysis tasks you'll complete before Spring Break. Please get in touch ASAP to finalize plans!

Step 1: Acquire and explore necessary datasets

Using the information provided in the Jupyter Notebook for your project, locate and download any necessary astronomical data. Load the dataset(s) in Python. Perform an initial exploration of the data (e.g. by printing out tables, examining metadata, etc) and identify key columns or parameters related to your research question.

Step 2: Clean and prepare your data

Identify missing or problematic data and decide how to handle it (e.g., remove null entries, fill in missing values, ignore certain columns, etc). Filter or subset your data if you only need specific information. For instance, you might need to focus on a certain region of the sky or specific types of objects. Be sure to document your decisions (why you removed certain points, what thresholds you used, etc) in your notebook so others can follow your logic!

Step 3: Conduct your analysis

Write Python code to address each of the analysis tasks, each of which will culminate with some form of data visualization. Ensure that your code is clear and thoroughly commented. When making plots, make sure that you label them clearly (axis labels, legend, etc). If you generate intermediate results (like a quick histogram of a key parameter or a scatterplot), please also include those (with a brief explanation) in your final submission!

Step 4: Summarize and interpret your findings

Write a brief (1-2 paragraphs) interpretation of the patterns you observe in your plots or tables. Link it back to your original research question and key concepts from your literature review.

Then, write a brief (1-2 paragraphs) reflection on the limitations of your analysis. Are there any caveats or assumptions in your analysis? Could more data or a different method provide more robust results?

Bonus: Extend your findings

Can you think of ways to extend your analysis? I'll list some options for extension in each project notebook, and I'm also happy to discuss any ideas that you have! This is not required, but it's a great way to practice the elusive *asking questions* part of research, so I encourage you to spend at least some time brainstorming.

Submission

Please upload your code (as a script or Jupyter notebook) to Courseworks before the deadline. If your code requires other files to run (such as saved datasets or files containing intermediate results), please also upload those to the same Courseworks assignment. In order to receive full credit, I need to be able to run your code from start to finish without making significant changes (needing to update file paths is OK, of course).

Rubric

Criteria	**Excellent**	**Satisfactory**	**Needs improvement**
Completeness	Submission includes all required elements (steps 1-4 listed above).	Submission is either missing a required element, or some elements are not fully completed.	Submission has multiple elements that are missing or incomplete.
Technical accuracy	The analysis is correct. The code runs fully without modifications.	The analysis is mostly correct. The code runs fully with minor modifications.	The analysis is mostly incorrect. The code only runs fully with significant modifications.

Code organization	The code is organized logically, with clear headings, concise explanations, and thorough comments.	The overall structure of the code is coherent, but some sections may need more detailed comments or explanations.	The code is disorganized and/or under-explained, making it difficult to follow the logic of the analysis.
Visualizations	All figures/tables effectively convey key results. Visuals are properly labeled and easy to interpret.	Figures/tables are present but may lack complete labeling. Key results are conveyed, but clarity or formatting could be improved.	Figures/tables are missing, poorly designed, or not aligned with the analysis tasks. Labels and key explanations may be missing.