Assignment 5: Data lifecycle planning (10% of overall credit score)

Due: Friday December 11, 2026 (by 11 pm ET)

Submission method: LMS

Please use the following file naming for electronic submission for any individual documents: DataScience_2026_A5_YOURFIRSTNAME_YOURLASTNAME.xxx

Late submission policy: If you are more than 2 days late it is likely that you will not have your grade for this assignment included in your final grade before they need to be submitted.

Assignment Instructions: You are the ranking data scientist on a committee planning a major project that is expected to collect and use large amounts of data as well as produce and publish various data and information products. You will be given two scenarios fitting that description, one from scientific research and the other from business. Each scenario will describe the goals of the project, specific requirements for the amount and types of data involved, how the data are expected to be used, and what outcomes should look like.

Use what you've learned about the data lifecycle, data management components, data/metadata standards and conventions to design a plan and workflow for your project.

Scenario A: A research project in astronomy will conduct a survey of exoplanets and synthesize a high-quality dataset for comparative planetology studies. The project will acquire data from **two** different sources. The first is an old institutional archive of data stored in **FITS** files. The second is a data collection you must design and manage for the project. The final combined dataset should be hosted on a **data portal** and made available to researchers. Any consequent papers and other derivative data products that make use of the dataset should also be hosted on the data portal.

You must create a data management plan for the data collection, taking into consideration the need to **integrate** the old data, and the requirement to host the final dataset **publicly**.

Scenario B: An IT infrastructure project will overhaul a company's information systems, replacing separate distributed software components with a unified solution. The project will migrate existing data where possible to new systems. Some past data were lost when an old IBM AS/400 mainframe failed catastrophically but **printed reports** of these data exist. The data in these reports must be made available to users of the new system. The company sells directly to customers and handles customer information including payment and other **sensitive information**. The new unified system will be used mostly by company employees with a limited public-facing component (website) for customers.

You must create a data management plan for the data migration, with a solution for making the **report outputs** available to users.

Use the following numbering to structure your submission:

1) Data Management Plan

- Use the 9 data management components to specify how you plan to manage the data for your project. Specify the solution and a brief comment on why it is suitable.
- Creation of logical collections: *how are the data organized logically?*
- Physical data handling: *file structure, format, backups, etc.*
- Interoperability support: *how can the data be used on a different platform?*
- Security support: *authorization/authentication, permissions, etc., who owns/approves?*
- Data ownership: *licenses, restrictions, etc., who is/are the point(s) of contact?*
- Metadata collection, management, and access: *who adds metadata, when, how and in what form (formats/standards)?*
- Persistence: *long-term storage/archival of the data, where/how/for how long?*
- Knowledge and information discovery: *how can others discover (find) the data (and the information/knowledge in the data)?*
- Data distribution and publication: *how to make available (usually publicly), how to notify of updates, etc.*

2) Data Lifecycle Workflow

- Create a workflow diagram demonstrating the data lifecycle in your project according to your plan. You may use any reasonable conventions in terms of the visuals, UML or similar would be suitable.
- The goal is to demonstrate what happens to data throughout the project. In scenario A for example, new data will be collected, processed, stored in the data portal, possibly analyzed by a researcher who produces a report and uploads it to the portal. In scenario B, data are extracted from legacy systems, processed and loaded onto the new system to be available to employees, with a subset available to customers.
- The diagram should contain the major applicable phases of a data lifecycle, at least “collect”, “process”, “analyze”, and “preserve”; include others where appropriate.
- You may split the workflow into separate diagrams.
- Be creative rather than exhaustive, cover the main steps and create branches where necessary, simplify wherever possible.

THE END