Dataset I - Wine Dataset

1. Convert the datasets to HDF5. (total 10%)

a. Utilize any appropriate tools/libraries to convert the original data files to HDF. Ensure all original data are replicated with correct data types. Maintain data relationships and hierarchical structure. Transfer all original metadata to HDF5 attributes. (7%)

File attached separately, used HDFViewer to view the file

b. Briefly document the conversion process and include any code used. (3%)

The conversion process began with me reading some of the links provided in the assignment to figure out how to use python to convert the data. I then did some digging on my own to find functions that I could use to directly read the files given and format them accordingly. I was then able to use the functions to effectively get the data in a sensical format in which the relationships and hierarchical structure was maintained. I have submitted wine.py which is the code I ran to convert the dataset to HDF5 and wine_data_new.h5 is the dataset converted to HDF5. These can both be found in the wine folder of the file I submitted.

2. For BOTH datasets answer the following questions. (total 10%)

a. Describe how the logical organization and physical organization have changed in the transfer to HDF5. This includes an indication of whether all metadata will be encoded in the HDF5 file or not, i.e. externally and why, choices of file names, etc. (5%)

Originally, the wine dataset was split across two files: wine.data (the table) and wine.names (the documentation). After converting to HDF5, both data and metadata are now stored in one file, which makes it easier to view everything together. HDF5 uses a hierarchical structure, so I placed the table and all of its metadata in separate but organized levels inside the file. I added all the text from wine.names as attributes at the dataset level so the documentation appears when you click on the dataset in HDFView. This keeps the descriptive information directly tied to the data it explains. I stored structural metadata like column names directly on the table object, keeping it separate from the descriptive documentation. HDF5 also stores data in binary form instead of plain text, which preserves precision. Since HDF5 supports attributes, all metadata could be stored inside the file instead of needing an external companion file. I also renamed the file to show that it is the archival, cleaned HDF5 version.

b. Describe what additional metadata and/or information you would include for the cataloguing and preservation purposes. This is for the repository to be able to clearly present your dataset and for potential researchers to find it. (5% 3000-level / 1% 6000-level)

To make the dataset easier for future researchers to find and use, I would add more metadata about its background and technical details. This includes who created the dataset, where it came from, and how it was converted to HDF5. I would also note any cleaning steps or data-type changes made during conversion. Including technical details like units, value ranges,

and software versions would help others reproduce the process. Administrative information, such as rights, licensing, and who maintains the file, would also be useful. Finally, adding keywords and references to related research would make the dataset easier to discover in a digital repository.

Dataset II - Dataset from Assignment 2

1. Convert the datasets to HDF5. (total 10%)

a. Utilize any appropriate tools/libraries to convert the original data files to HDF. Ensure all original data are replicated with correct data types. Maintain data relationships and hierarchical structure. Transfer all original metadata to HDF5 attributes. (7%)

File attached separately, used HDFViewer to view the file

b. Briefly document the conversion process and include any code used. (3%)

The conversion process began with me reading some of the links provided in the assignment to figure out how to use python to convert the data. I then did some digging on my own to find functions that I could use to directly read the files given and format them accordingly. I was then able to use the functions to effectively get the data in a sensical format in which the relationships and hierarchical structure was maintained. I have submitted liquids.py which is the code I ran to convert the dataset to HDF5 and liquid_data_new.h5 is the dataset converted to HDF5. These can both be found in the liquid folder of the file I submitted.

2. For BOTH datasets answer the following questions. (total 10%)

a. Describe how the logical organization and physical organization have changed in the transfer to HDF5. This includes an indication of whether all metadata will be encoded in the HDF5 file or not, i.e. externally and why, choices of file names, etc. (5%)

In its original CSV form, the dataset was a single text file with no structure, so metadata had to be kept separately in Metadata.txt. After converting to HDF5, the dataset now has a hierarchical layout, separating the actual data from its documentation. The data are now stored in binary formats instead of plain text, which preserves data types and makes storage more efficient. I also added groups and attributes to show the relationships between the data and its metadata. All of the information from Metadata.txt is now stored inside the HDF5 file so that the file explains itself without needing anything external. The only external file kept is Metadata.txt for backup or readability. I renamed the final file liquid_data_new.h5 to make it clear that this is the archival HDF5 version.

b. Describe what additional metadata and/or information you would include for the cataloguing and preservation purposes. This is for the repository to be able to clearly present your dataset and for potential researchers to find it. (5% 3000-level / 1% 6000-level)

To support long-term use, I added metadata that helps researchers understand and locate the dataset. This includes descriptions of each variable, their units, and how measurements were taken. I also added context like how the data were collected and the purpose of the experiment. Provenance information such as creators, affiliations, and contact info were added for citation purposes. I included details about how temperatures were measured, how the file is structured, and what the original formats were. The full original Metadata.txt is saved inside the HDF5 file

so nothing is lost. I also included machine-readable metadata, like data types and missing values, to help digital repositories index the dataset properly.