Question 1

Both data sets will be addressed simultaneously as they were both survey results so the process for converting them to HDF5 was nearly identical.

a. The h5py python library was used along with the pandas library to convert the data to hdf5. The excel file was read into a pandas dataframe and then was exported using the pandas function to_hdf(). The original metadata was included as a json file that came with the excel file. The basic layout idea of the format can be found in the diagram below.

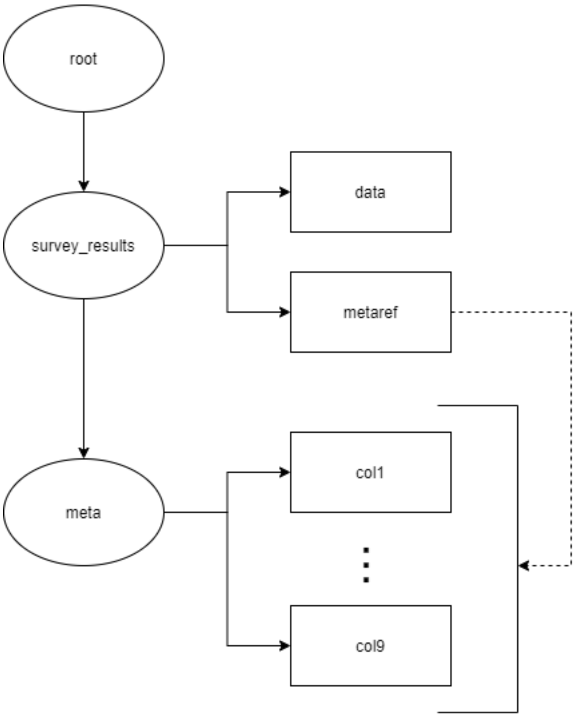

Figure 1: HDF5 Organizational Layout

The metadata describing the survey column was put in its own data set with the values being the accepted values for that column and then the data set 'metaref' was a set of HDF5 Reference objects to the the respective metadata data set. The json standard of name-value pair was maintained, but it was rearranged to use the attributes property to describe the name value pairs.

b. Metadata would need to be included with the file describing the file layout as the pandas library only fetches the data set and not the additional metadata. Other than that all metadata is self-contained.

c. Understanding HDF5 in the first place was very complicated given that its written in C, and C does not have any native description for the string datatype which made the conversion very difficult. This was solved by expliciting telling python to use the h5py string datatype instead of letting it automatically encode it at utf-8 which HDF5 did not recognize.

Question 2

The creation of the HDF5 files can be found in the python files creation.py and creation_2.py. The data can be retrieved from the HDF5 file by using the python file retrieval.py. The python virtual environment folder has been provided as well to simplify the process.