Progress Report: Dataset Analysis

Provide a document that updates any details that were missing from the proposal and specifically describes, for each point in the timeline, what progress has been made, what issues cropped up, and what remains to be completed.

Standard Metadata: Title, Authors, Date

1. The Dataset I am working on is:

   Title: Million Song Dataset
   URL: https://labrosa.ee.columbia.edu/millionsong/

   I began working on the sample dataset which was of 1.8GB (Million Song Subset) available from the following website:
   https://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset

   The subset data files of 1.8 GB was in .h5 format. I began loading these files using pandas in Jupyter notebook. However, this step didn’t work well because of certain dependencies. The next step was on convert them in a .csv format. The MSD LABROSA website provided Wrappers in Python based on pytables and numpy along with another file: hdf5_getters.py, to access the fields in the HDF5 sing files is provided from which the file could be converted to .csv. For conversion from h5 to csv for the sample data, I faced the following challenges:

   →Jupyter notebook had its kernel restarted every time I was importing ‘tables’ from the pytables package

   →I then downloaded Canopy, which came with built in pytables and numpy installed. Conversion to .csv had another challenge. The code written by me for importing the hdf5_getters.py was not compatible with the pytables package version installed within Canopy.

   →While solving these dependencies, I referred to the following GitHub link: https://github.com/amgreenstreet/Million-Song-Dataset-HDF5-to-CSV which helped me convert .h5 to .csv files for the sample dataset

   With the sample .csv file obtained (10,000 records) I was able to analyze the data and answer the following two questions:

2. **Questions answered**:

   The questions I answered using this dataset as of now are:

   1. Which location most of the artists came from?
   →Displayed artists using location based on latitude and longitude

   2. What years had the highest number of album/song releases?
   →Displayed Counts as per descending order of years, the songs were released

3. **Next Steps to be carried out and Scalability Challenge Factor**:

   For the Scalability factor that I would be addressing, will be to further answer the other questions mentioned in the proposal.
   For the Scalability, the full data is available on AWS Public Datasets. Its size is 500GB. I will be answering the rest of the questions on my local machine sand then implement the same on the cloud.

4. **Papers Referenced**:

1. The Million Song Dataset, T. Bertin-Mahieux, D. Ellis, B. Whitman and P. Lamere, *ISMIR '11* http://ismir2011.ismir.net/papers/OS6-1.pdf

2. The Million Song Dataset Challenge, B. McFee, T. Bertin-Mahieux, D. Ellis and G. Lanckriet, *AdMIRe* https://pdfs.semanticscholar.org/73d2/16afb813f53134a3e635c3dcbecb612cb23e.pdf