Data Wrangling Steps

The lyrical information, audio information and genre information are all present in different files. So before doing any sort of analysis I had to extract, clean and combine all this information.

The lyrical information is available in an sqlite database in a bag of words format for top 5000 words. Basically is row in sql table contains track id, word and word count.

The Genre information is available in a text file. Each line contains track id and its associated genre.

The audio information is present in a hdf5 table. Each row contains track id and the corresponding audio features such as energy, loudness, tempo, danceability etc.

As a first step I parsed the genre information and created a dictionary mapping track ids to genre labels. I also created an inverse dictionary mapping genre labels to a list of track ids.
As and example following were the number of tracks present for each genre:

1. Pop_Rock 95752
2. Vocal 1497
3. Rap 5291
4. Folk 1344
5. Country 5064
6. RnB 4689
7. Comedy_Spoken 215
8. Jazz 703
9. Reggae 927
10. Latin 5537
11. Electronic 3624
12. Religious 3501
13. International 2155
14. Easy_Listening 166
15. Blues 636
16. Classical 98
17. Avant_Garde 16
18. Children 41

19. Holiday 43
20. Stage 77

Creating a single table for all these tracks was giving me memory errors so I decided to work with a subset of this data for my analysis steps. I decided to randomly pick 500 tracks from each genre. Genres with less than 500 tracks were removed. These were Stage, Holiday, Children, Avant_Garde, Classical, Easy_Listening and Comedy_Spoken

For the remaining 13 genres for each of the 500 track ids picked I parsed to lyrical information and created a 5000 dimension vector with each element containing a number corresponding to a word count. Then a matrix containing these 500 vectors for each genre was saved as a numpy file for future use.

Analyzing the audio data some of the features present were 0 or same for all the tracks e.g danceability. And other features had a lot of missing values or did not seem related to genre e.g ‘artist_hotness’. Therefore, finally 3 features: duration, loudness and tempo were picked for each of the 500 track ids.