**Conclusion**

Summing this all up, some features that might be useful:

1. Number of genres a song has
2. Number of composers and lyricists a song has
3. Number of plays a song has
4. Number of tracks and plays the song's artist has
5. The songs language
6. The number of languages an artist sings in
7. The song's length
8. The genres a song has (these can be one-hot encoded, something has to be done for the genres that are in the test set but do not appear in the train set

Further exploration/feature ideas that I haven't tested, but that could be useful:

1. Split "Various Artists" into separate artists, based on language and/or genre of the tracks?
2. Song length compared to the average song length of songs by the artist?
3. Number of artists the songs composer and lyricist have worked with (a composer who writes 'hits' probably works with a lot of artists; but there seems to be some rubbish data in those fields, so maybe this feature is not that feasible after all for this data set)
4. Genre pairs (if a song has genres 1,2,3, it has genre pairs (1,2), (2,3), (1,3) - perhaps these are informative too)
5. Check whether the song is in an artists 'main' language (the most common language of the artists songs)
6. Check whether the song is an artists 'main' genre (and also perhaps check if it's of a genre the artist doesn't usually perform in)

<https://www.kaggle.com/kunstmord/exploring-the-songs>

more features:

At a glance I will be looking into the following issues:

* Merging Dataframes -> (*completed*)
* Handling Missing Values -> (*completed*)
  + Came across something interesting where number of unique values in column '**source\_screen\_name**' were different for both test and train set.
  + Columns like '**genre\_ids**' have a combination of more than one genre which must be handled appropriately.
* Feature Engineering -> (*in progress*)
  + number of genres per song
  + number of lyricists per song
  + number of composers per song
  + whether song has features artists
  + number of artists per song
  + whether artist and composer are the same
  + whether artist, composer and lyricist are all the same

Missing data imputation:

song length feature: use mean values to impute