2. Metadata in the music industry is a mess. We have a number of data providers who do not communicate with each other and naming conventions are non-existent. For example, all of these artist/title combinations refer to the same song:

(Apple) The Kid LAROI & Justin Bieber – Stay
(Shazam) The Kid LAROI & Justin Bieber – STAY
(Spotify) The Kid LAROI – Stay (with Justin Bieber)
(Amazon) The Kid LAROI & Justin Bieber – Stay [Explicit]
(Youtube) The Kid LAROI, Justin Bieber – STAY (Official Video)
(Youtube) The Kid LAROI – STAY (Live Performance)
(tiktok) kid laroi – STAY BY KID LAROI leaked audio

The data provided has not been normalized or conformed to naming and conventions can cause many challenges for development and data analysis. With the data presented, we can see few similarities. That being what comes after the `-` character. After this character we can always find the name of the song (Stay or STAY). The goal here would be to normalize the data.

To normalize the data, it would be preferred to split each string by the '-' character, this will allow us to split the 'Artist' and the 'Song' (track). Upon performing this operation, with the data provided, we will have the artist along with the name of the song. The next concern will be the artist on the song. We do have one unique case (tiktok) where 'Kid Laroi' is included on the right side of the '-' character, for this we would have to trim the string for everything after 'stay' as we would for all of the other data sources. In addition to this on the left side of the '-' character, we have 'Justin Bieber' included in some cases, to normalize this, we would have to trim both the left and right side of 'KID LAROI' (the data transformed to .upper()).

Upon observing the data, we can see that the featured artist is not a dependency for the final dataset (tiktok and youtube - 2, Justin Bieber is not included). With this information, we can ignore the featured artist and create a primary key containing the song and artist. An example of this would be, STAYKIDLEROI. This primarykey or 'id' would exclude 'the' from 'The Kid Leroi' because 'THE' is not a substring that is utilized throughout all data sources we are ingesting from. Additionally you may see that the primaryket consist of all upper-case letters, this is used to create a standard across all data sources.

Final Primary_Key or ID:
SongId.toUpper()

Example:
STAYKIDLAROI