**Exploratory Data Analysis on the Million Song Data Set**

**Introduction**

The Million Song Dataset is a freely available collection of audio features and metadata for a million contemporary popular music tracks.

One of its purposes is to provide a reference dataset for evaluating research.

The dataset that this project uses is a sample of the main dataset and contains 10 000 rows of song data, i.e., 1%.

This EDA explores the following:

* Which words describing music genres occur most often in our dataset?
* Comparison of song duration, song tempo, and song loudness across four major genres.
* Correlation between song loudness vs song tempo for the four major genres.
* Select subgenres-comparison of song duration, song tempo, and song loudness

**DATA CLEANING & MISSING DATA**

Most of the columns were deleted, as they were not relevant to the analyses as per this report. The five remaining columns include:

* artist.name (name of artist);
* artist.terms (genres/subgenres of music);
* song.duration (total length of a song, measured in seconds);
* song.loudness (overall volume of a song, measure in decibels. Uses negative integers, the smaller the number, the softer the volume, i.e. -6dB is louder than -9dB); and
* song.tempo (measure in beats per minute, or bpm).

The data has been prepared and cleaned by the Million Song Dataset Organization. As a result, using Missingno map, the only issue found was 5 NaN values in the ‘artist.terms’ column. These NaN values were replaced with 0. This is justified as the missing data made up only 0.05% of the dataset and replacing the NaN values with 0 will not significantly impact subsequent calculations/data analyses.

DATA STORIES AND VISUALIZATIONS

1. **WHICH WORDS USED TO DESCRIBE A MUSIC GENRE OCCUR MOST FREQUENTLY IN THE DATASET?**

To find this out, WordCloud was utilised. The bigger the word in the word cloud is, the more frequent the occurrence of a particular word in the dataset. The words in these word clouds correspond to music genres/subgenres. There are 106867 words in combination in the ‘artist.terms’ column.

Two word clouds were implemented, the first for frequency of words over the entire dataset, and the second the top 50 words. The most frequently found words are:

* Rock;
* Pop;
* Blue;
* Jazz;
* Metal; and
* Hip Hop.

The word clouds show that the dataset contains a wide variety of genres/subgenres. Besides the abovementioned words, the dataset has such varying music genre words such as ‘country’, ‘brazilian’, ‘trance’, ‘progressive’, ‘soundtrack’, ‘punk’ and ‘comedy’. This wide selection indicates that there is opportunity to deep dive into the various genres/subgenres to glean insights.

1. **COMPARISON OF SONG DURATION, SONG TEMPO, AND SONG LOUDNESS ACROSS FOUR MAJOR GENRES**

As mentioned, the four most frequently occurring words/music genres are Rock (occurring 16.59%), Pop (10.71%), Jazz (5.90%), and Blues (8.26%). As these are broad, general genres, they are representative of their respective subgenres. So, it is worthwhile comparing the song duration, song tempo, and song loudness of these genres to one another. A bar graph has been utilised.

Song Duration

Jazz has the largest median song duration (247 seconds/4:07), followed by Rock (228 seconds/3:48), Pop (220 seconds/3:40), and Blues (208 seconds/3:28). The jazz median song duration is quite a bit larger than the other genres. This is expected as Jazz often contains long instrumental sections in the middle of the song, thus making the song length longer than other genres. However, Blues, while like Jazz in having long instrumental sections, has a median of only 3:28 minutes; one would expect Blues to have a larger median to Rock and Pop. Reasons for the small median could be the word ‘Blue’ occurs in the dataset in subgenres which have much shorter song durations (Rhythm & Blues for example), or the Blues genre/subgenres in the dataset includes songs that are considered early Blues from the beginning of the 20th Century. These songs are considerably short in length, usually lasting no longer than 3 minutes.

Song Tempo

Rock has the fastest tempo (123 bpm), followed by Pop (121 bpm), Jazz (118 bpm), and Blues (114). While Rock, Pop, and Jazz have similar medians, Blues is noticeably slower that Rock and Pop. This is indicative of the Blues form, music which characteristically swings to lower tempos.

Song Loudness

Pop is the loudest genres with a median of -8.7dB, followed by Rock (-9.1dB), Blues (-11.4dB), and Jazz (-12.9) being the softest. This makes sense as Rock and Pop songs utilise more sonic ‘punch’ in their production that results in louder volumes and a more intense listening experience. This is in contrast with Jazz, which requires much softer volumes to accentuate the nuances of the instruments and the subtleties of the musicians.

1. **CORRELATION BETWEEN SONG LOUDNESS VS SONG TEMPO FOR THE FOUR MAJOR GENRES**

Is there a correlation between song loudness and song tempo within the Rock, Pop, Jazz, and Blues genres? In other words, the hypothesis is that faster songs tend to be louder.

All four genres show weak correlation, with Jazz being the weakest. It will be interesting to see if there’s a stronger correlation for these genres in the bigger Million Song Dataset of a million songs.

1. **SELECT SUBGENRES-COMPARISON OF SONG DURATION, SONG TEMPO, AND SONG LOUDNESS**

INTRODUCTION

The subgenres below were selected for comparison due to there being enough songs in each subgenre to yield meaningful insights.

A note about song duration outliers. Very short songs can be explained as tracks that are ‘fillers’ on an album, i.e., very short tracks that serve to introduce the next track. Very long tracks can be due to a song having an extended instrumental section or being purely instrumental with many sections. In other words, song length can be easily account for.

HIP HOP (346 songs) VS GANGSTER RAP (134 songs) VS POP RAP (59 songs)

Song Duration

Most Hip Hop songs are between 180 and 300 seconds (3-5 minutes) in length, and quite a few songs less than 180 seconds. There are outliers above 400 seconds, a few over 9 minutes long. The Gangster Rap catplot also indicates most songs between 180 and 300 seconds, but the range of values is much narrower than Hip Hop, with the longest songs duration only at about 360 seconds. Pop Rap is like Gangster Rap, showing most songs from 180-240 seconds with a narrow range.

Song Tempo

In the Hip Hop genre, the range of tempos with the highest density is between 80-120bpm. There is a relatively even spread of faster tempos between the 120-190bpm. Even though they have considerably less songs than Hip Hop, both Gangster Rap and Pop Rap show similar trends to Hip Hop.

Song Loudness

The greatest density of volume for Hip Hop is between -3 and -10 dB, which is quite loud. From -10dB and softer, there is a thick tail that tapers off from -14dB. Once again, Gangster Rap shows similar trends to Hip Hop, while there is not enough data for Pop Rap to see if it mirrors the other two genres.

Analysis

The similarity of Hip Hop, Gangster Rap, and Pop Rap across song duration, song tempo, and song loudness is indicative of commercial nature of these genres, which is characterised by being quite homogenised in many aspects of music production and the music itself. Examples of this can be gleaned from our data. The mean song length(3-5 minutes), the slight slow to medium tempos (80-120bpm), and loud volume (-3 to -10dB) exemplify these genres.

HEAVY METAL (97 songs) VS ALTERNATIVE METAL (67 songs)

Song Duration

The majority of Heavy Metal songs are between 200 and 300 seconds long (3:20-4:00 minutes). There is a thin tail from about 320 seconds to 500 seconds, with a couple outliers between 650 and 700 seconds. On the other hand, most of the Alternative Metal songs are shorter in length, between 180 and 250 seconds (3:00-4:10 minutes).

Song Tempo

Both Heavy Metal and Alternative Metal show similar, relatively even distribution of songs tempos, from around 82 bpm to considerably fast tempos of around 170bpm.

Song Loudness

Again, Heavy Metal and Alternative Metal show similar song volume characteristics, with most songs between -3 and -10dB.

Analysis

Heavy Metal music is considered a commercial music genre, albeit less so than Rock or Pop. IN our dataset, most songs in the Heavy Metal genre are of similar length to Hip Hop and Rap songs as discussed above. On the other hand, on average, the song length of Alternative Metal songs is shorter than Heavy Metal. This might be explained by Alternative Metal generally being of a more commercial nature than Heavy Metal, which necessitates songs being of a short duration fitting for radio play, (usually between 3 and 4 minutes).

By their very nature, Heavy Metal and Alternative Metal have considerably faster tempos than the likes of Hip Hop and even Pop. This is consistent with our data. Heavy Metal has few songs under 100bpm (a medium to slow tempo) while both Heavy Metal and Alternative Metal have a spread of tempos from around 100bpm-165bpm.

Heavy Metal and Alternative Metal songs are loud. Our data shows this to be the case, with many songs having a volume on and around -5db. This is loud!

LATIN JAZZ (150 songs) VS SALSA (82 songs) VS JAZZ FUNK (80 songs) VS SMOOTH JAZZ (58 songs) VS FLAMENCO (56 songs)

A note: Latin Jazz, Jazz Funk, and Smooth Jazz are subgenres of a general nature, i.e., they are an overarching category for other subgenres. Salsa and Flamenco, in contrast, are specific music subgenres.

Song Duration

The general subgenres of Latin Jazz, Jazz Funk, and Smooth Jazz show a much wider spread of song length than the specialised subgenres, Salsa and Flamenco. This makes sense as the three general subgenres each consist of songs from numerous other specialised subgenres that vary greatly in song duration. In contrast, Salsa has a narrower range of song lengths, mostly from around 170-320 seconds (2:50-5:20 minutes). Flamenco has an even narrower range, from about 150-220 seconds (1:30-2:40 minutes). Intuitively this makes sense, as the small range in song length demonstrates that these specialised subgenres have songs of a specific song length, which is one of the factors that identify these specialised subgenres as Salsa and Flamenco.

Song Tempo

Amongst the generalised subgenres, Latin Jazz has a relatively even spread of song tempos, from very slow (approximately 60bpm) to quick (150bpm). There is also quite a thick tail from around 150bpm to above 200bpm. Again, this is indicative of Latin Jazz being a catchall subgenre for numerous other subgenres, whose tempo can vary greatly. Jazz Funk has most songs between 80 and 120bpm, a range many genres share. This includes commercial genres like Pop and Hip Hop, as seen above. Smooth Jazz has a reasonably even spread of song tempos, from approximately 75-140bpm. Again, like Latin Jazz, this large range reflects the catchall nature of the Smooth Jazz genre.

Regarding the specialised subgenres, the majority of Salsa songs have speeds from about 80bpm-125bpm. That there is a long tail is surprising. Salsa is a specific kind of Latin music that has a signature sound and tempo. Exploring the specific Salsa songs in the dataset is needed to ascertain whether labelling those songs as Salsa is accurate, or whether there is a hidden specific subgenre within the Salsa subgenre itself. The observations about Salsa are also true for Flamenco, although the relatively fewer number of Flamenco songs in the dataset may not truly representative.

Song Loudness

Generalised subgenres: Once again, Latin Jazz has a wide song volume range, with most songs falling roughly in the -7dB - -17dB range. This is a very wide range of volumes which, as mentioned above, makes sense. Jazz Funk has a large cluster of songs with song loudness between -12.5dB and -15dB. There is also a smaller (and louder) cluster between around -8dB and -10dB. Most Smooth Jazz songs fall in the -7.5dB - -15dB range.

The specialised song style of Salsa (and its distinctive qualities) is once again demonstrated here, with the majority of songs at a slightly louder and more narrow volume band of between -5dB and -9dB. Finally, the data shows that Flamenco songs have a wide spread of volumes (from around -4dB to -14dB).