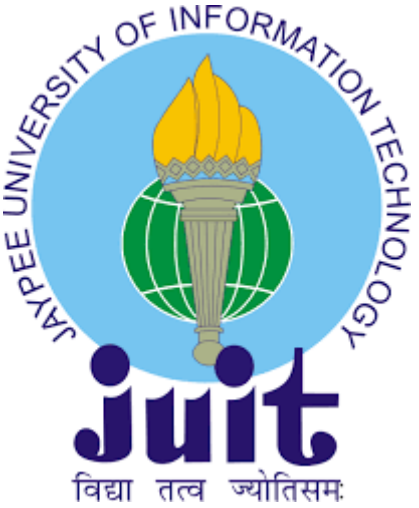
Jaypee University of Information Technology



Department of Computer Science and Engineering and Information Technology

\

PROJECT REPORT

Analysis on the Most Streamed Spotify Songs 2023

DATA SCIENCE AND VISUALISATION LAB

Submitted by:

Aadya Thakur 211184
Shagun 211170

Submitted to:

Dr. Diksha Hooda

1. Introduction

The aim of this report is to analyse the most streamed Spotify songs of 2023, utilising a comprehensive dataset with various features such as track name, artist information, release details, streaming metrics, and other relevant attributes. The analysis focuses on cleaning the data, identifying relationships between variables, and exploring how track metrics correlate with the number of streams, chart placement, and frequency in playlists.

1.1 Dataset Features Overview

The dataset comprises the following features:

- track_name: Name of the song
- artist(s)_name: Name of the artist(s) of the song
- artist_count: Number of artists contributing to the song
- released_year, released_month, released_day: Release details of the song
- in_spotify_playlists: Number of Spotify playlists the song is included in
- in_spotify_charts: Presence and rank of the song in Spotify charts
- streams: Total number of streams in Spotify
- in_apple_playlists: Number of Apple Music playlists the song is included in
- in_apple_charts: Presence and rank of the song in Apple Music charts
- in_deezer_playlists: Number of Deezer playlists the song is included in
- in_deezer_charts: Presence and rank of the song in Deezer charts
- in_shazam_charts: Presence and rank of the song in Shazam charts
- bpm: Beats per Minute, a measure of song tempo
- key: Key of the song
- mode: Mode of the song (Major or Minor)
- danceability_%: Percentage indicating how suitable the song is for dancing
- valence_%: Positivity of the song's musical content
- energy_%: Perceived energy level of the song
- acousticness_%: Amount of acoustic sound in the song
- instrumentalness_%: Amount of instrumental content in the song
- liveness_%: Presence of live performance elements
- speechiness_%: Amount of spoken words in the song

Dataset:

https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023/

1.2 Objectives

The main objectives of this analysis are as follows:

Data Cleaning:

- Identify and handle null values.
- Check and correct incorrect data types.
- Create additional columns if necessary.

Relationship Identification:

- Explore relationships between different variables.
- Investigate how track metrics (danceability, energy, valence, etc.) relate to the number of streams, chart placement, and playlist frequency.

Visualization and Observations:

- Utilize Seaborn and Matplotlib to visually represent relationships.
- Provide insights and observations based on the analysis results.

2. Methodology

2.1 Data Cleaning:

- Checked for null values and replaced or removed them.
- Ensured correct data types for each column.
- Created additional columns as needed for the analysis.

2.2 Exploration:

- Conducted exploratory data analysis to understand the distribution of variables.
- Investigated correlations between different features.

2.3 Visualization:

- Utilized Seaborn and Matplotlib for data visualization.
- Plotted relationships between variables to aid in analysis.

3. Libraries and Dataset

The analysis of the Most Streamed Spotify Songs 2023 project utilized several Python libraries to clean, explore, and visualize the dataset. Below are the libraries employed in the analysis, along with their respective purposes:

1.Numpy

It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays. In this analysis, NumPy is used for various array manipulations and mathematical operations.

2.Pandas

It provides data structures like DataFrame and Series, which are instrumental in handling and processing structured data. In this analysis, Pandas is used for loading and cleaning the dataset, creating additional columns, and conducting exploratory data analysis.

3.Matplotlib

Matplotlib is a versatile plotting library for creating static, animated, and interactive visualizations in Python. In this analysis, Matplotlib is utilized for generating various plots and visualizations, helping to illustrate relationships and patterns in the dataset.

4.Seaborn

Seaborn is built on top of Matplotlib and provides a high-level interface for drawing attractive and informative statistical graphics. It simplifies the process of creating complex visualizations and adds a layer of aesthetic appeal.

5.Warnings

The 'warnings' library is used to filter out specific warning messages that may not impact the analysis or may not be relevant at the moment. In this case, FutureWarnings are ignored to maintain a clean and concise output during the analysis.

4. Checking of Values in the Dataset

4.1 Null Values

The dataset was examined for null values, revealing occurrences in the columns in_shazam_charts and key. For in_shazam_charts, null values may indicate that certain songs, despite having a high number of streams, did not make it to the Shazam charts. Regarding the key column, which represents the key of the song, Spotify for Developers - Get Track's Audio Features suggests that null values can be replaced with -1 in the context of the standard Pitch Class Notation.

```
[ ]  ## Replace null values in 'in_shazam_charts' column into 0
     df['in_shazam_charts'].fillna(0, inplace=True)

     ## Replace invalid data in the 'stream' column into null
     df['streams'] = df['streams'] = pd.to_numeric(df['streams'], errors='coerce')

     ## Replace null values in 'key' column into 'none'
     df['key'] = df['key'].fillna(-1)
```

```
     ## Drop the null value in the 'streams' column
     df = df.dropna(how='any')

     print(df.isnull().sum())
```

4.2 Incorrect Datatypes

Upon inspection, the following columns were identified as having incorrect datatypes:

- streams
- in_deezer_playlists
- in_shazam_charts

To address this issue, the datatypes for these columns will be converted to int64.

```
## Change the datatype of 'streams' column into int64
df['streams'] = df['streams'].astype('int64')
```

4.3 Pitch Class Notation

In the context of Pitch Class Notation, the conversion of key values to integers is as follows:

Pitch Class	Tonal Counterparts
0	C
1	C#
2	D
3	D#
4	E
5	F
6	F#
7	G
8	G#
9	A
10	A#
11	B

```
## Change values in 'key' column into integers using Pitch Class Notation
pitch_class = {'C': 0,
               'C#': 1,
               'D': 2,
               'D#': 3,
               'E': 4,
               'F': 5,
               'F#': 6,
               'G': 7,
               'G#': 8,
               'A': 9,
               'A#': 10,
               'B': 11
               }

df['key'] = df['key'].map(pitch_class).fillna(-1)
```

4.4 Additional Column

To provide a clearer classification of a track's placement in the charts, four additional columns will be created. These columns will use the following metrics for rating a track's placement:

- 0: Uncharted
- 1-10: Top 10
- 11-50: Top 50
- 51-100: Top 100
- 101-200: Top 200
- >200: Charted, but beyond the top 200

```
## Create new columns for track's chart classification
def chart_cat(value):
    if value == 0:
        return 'Uncharted'
    elif 1 <= value <= 10:
        return 'Top 10'
    elif 11 <= value <= 50:
        return 'Top 50'
    elif 51 <= value <= 100:
        return 'Top 100'
    elif 101 <= value <= 200:
        return 'Top 200'
    else:
        return 'Charted'

for col_chart in ['in_spotify_charts','in_apple_charts','in_deezer_charts','in_shazam_charts']:
    new_col_chart_name = col_chart + '_category'
    df[new_col_chart_name] = df[col_chart].apply(chart_cat)
```

These modifications will enhance the dataset's clarity and assist in subsequent analyses.

5. DESCRIPTIVE STATISTICS

The **df.describe()** function was applied to gain insights into the descriptive statistics of the dataset. The observations are as follows:

Release Date:

- The dataset spans from songs released in 1930 to the latest releases in 2023.

Popularity:

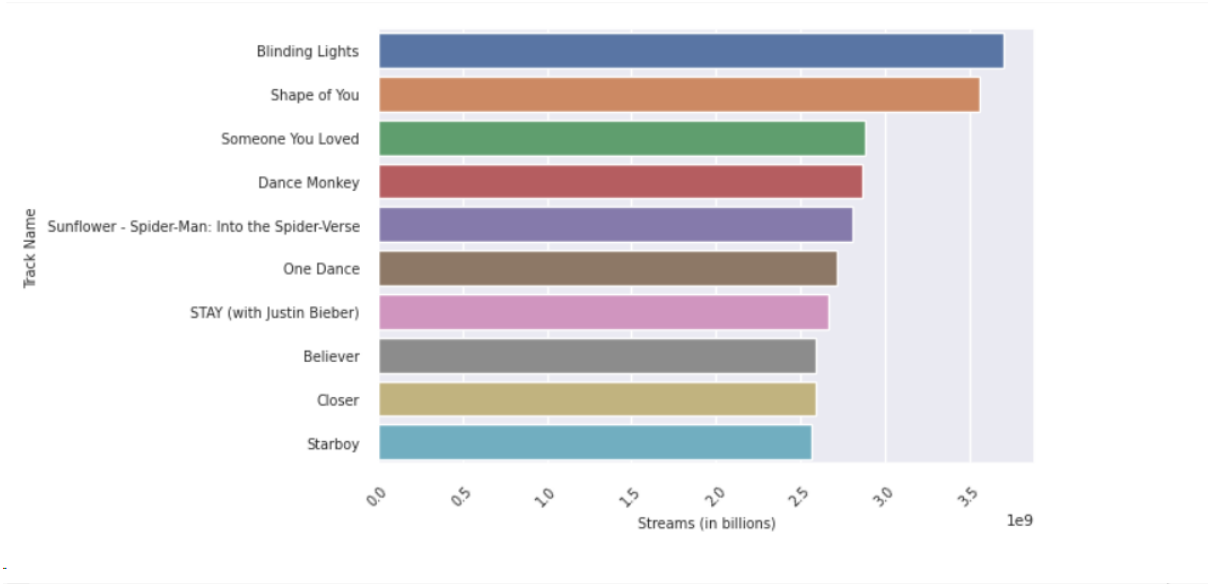
- Some of the most streamed songs did not chart.
- The number of streams ranges from a minimum of 2,762 to a maximum of 3,703,895,000.
- The mean number of streams is 514,137,400, with a significant standard deviation of 566,856,900, indicating a wide variability in the number of streams.

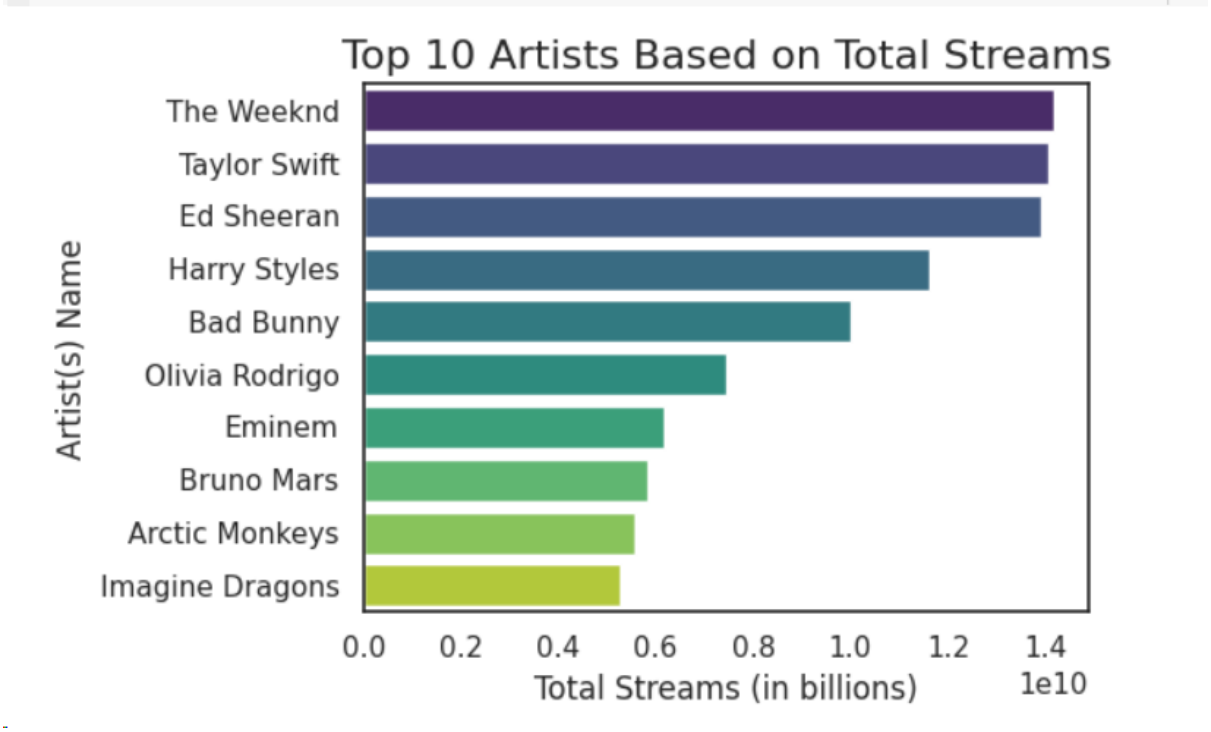
Song Properties / Characteristics:

- BPM (Beats per Minute) ranges from a minimum of 65 to a maximum of 206, with a mean of 122.55.
- Danceability ranges from a minimum of 23% to a maximum of 96%, with a mean of 66.98%.
- Valence (Positivity) ranges from a minimum of 4% to a maximum of 97%, with a mean of 51.41%.
- Energy ranges from a minimum of 9% to a maximum of 97%, with a mean of 64.27%.
- Acousticness ranges from a minimum of 0% to a maximum of 97%, with a mean of 27.08%.
- Instrumentalness ranges from a minimum of 0% to a maximum of 91%, with a mean of 1.58%.
- Liveness ranges from a minimum of 3% to a maximum of 97%, with a mean of 18.21%.
- Speechiness ranges from a minimum of 2% to a maximum of 64%, with a mean of 10.14%.

Data Visualization

Most Streamed songs visualized





4.2.1 Summary of Observations

Finding most underrated songs:

Following criterias to be considered for finding underrated songs

- There should be a limit on streams, to find underrated songs we will have to filter out the popular songs.

- The number of added in playlist should be high, which indicates listeners have liked that song.
- The song shouldn't be on any charts, which indicates the song wasn't discovered through any popular charts, but by word of mouth or exploration.

```
[ ] df.query("streams < 140000000 & in_spotify_charts == 0 & in_spotify_playlists > 2000")
```

	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams	i
420	Rumble	Skrillex, Flowdan, Fred again..	3	2022	1	17	2849	0	78489819	
475	Merry Christmas	Ed Sheeran, Elton John	2	2017	11	10	2209	0	135723538	
483	Deck The Hall - Remastered 1999	Nat King Cole	1	1959	1	1	3299	0	127027715	
509	Gasoline	The Weeknd	1	2022	1	7	2297	0	116903579	
512	Take My Breath	The Weeknd	1	2021	8	6	2597	0	130655803	

Artist Count:

- Tracks with only one artist tend to be more popular and receive higher streaming numbers.

Year of Release:

- Newer songs generally receive more streams, but there are older songs dating back to 1930 that have garnered significant streaming. These older songs could be classics or experienced a resurgence in popularity.

Presence in Platform Playlists:

- Histograms and scatterplots for playlist presence on Spotify, Apple Music, and Deezer show a leftward skew, indicating that most top-streamed songs are not frequently included in user playlists. This suggests that high streaming numbers do not necessarily correlate with playlist inclusion.

Presence in Platform Charts:

- Histograms and scatterplots for chart presence on Spotify, Apple Music, Deezer, and Shazam indicate that a significant number of top-streamed songs do not place on charts. New columns for chart classification reveal insights into the duration of streaming for different chart placements.

BPM (Beats per Minute):

- The most streamed music tends to have a BPM of 120, with 90 BPM following closely. High BPM values have fewer streams, suggesting a preference for moderate-speed music.

Key:

- The key with the most streams is 1 (C#), while the key 0 (C) has no streamed music. Other keys show small variations in the number of streams.

Track Metrics:

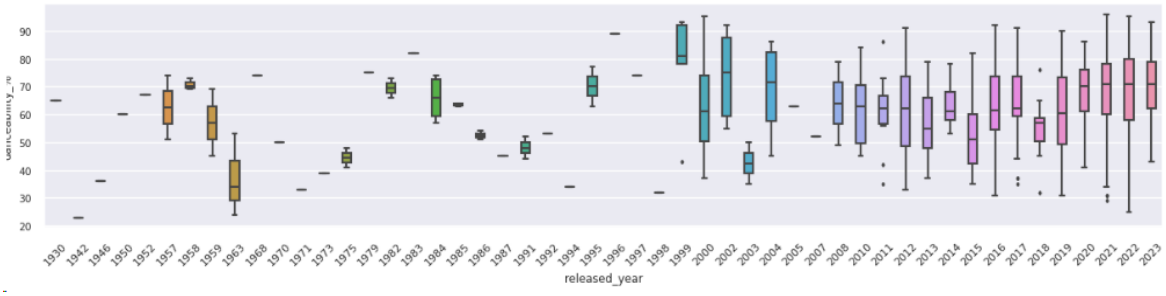
- **Danceability**: Most streams occur in tracks with 70% danceability, indicating a preference for danceable music.

  Tracks that are released from the year 2020 to 2023 have almost similar median danceability, and almost similar interquartile range.

  Tracks that are released from 2008 to 2023 have wide range of danceability. It could be due to the majority of the top tracks were released in these years.

  The most danceable track in the top streamed songs was released in 2021.

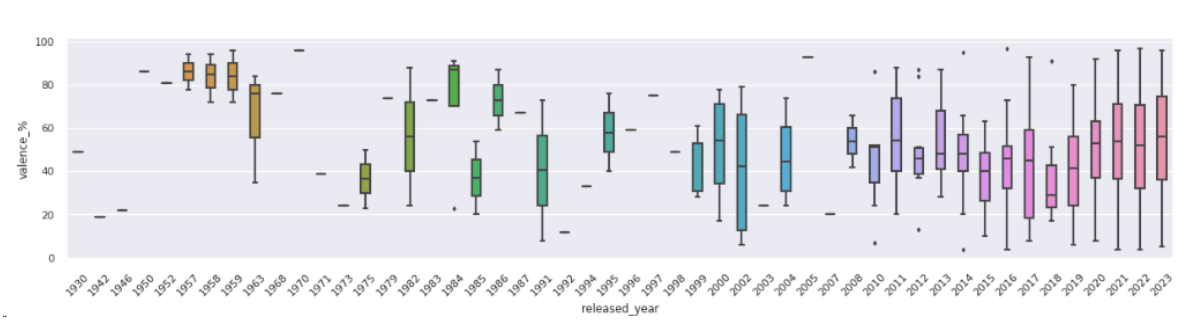
  The least danceable track in the top streamed songs was released in 1942.



Valence (Positivity):.

Tracks from 2020 to 2023 shows wide variety of moods in the top streamed songs, long whiskers extending from low valence to high valence, and median values at approximately 50%.

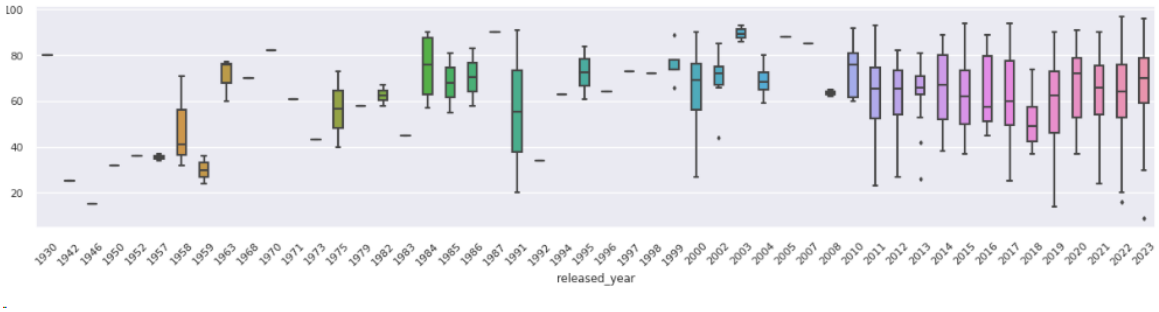
Tracks from 2011 to 2023 has median valence at approximately 40 to 50%, with the exception of 2018. The range is also at the middle of the chart, ranging 20 to 70%, which shows the neutrality of the mood in the top streamed songs.



- Energy:

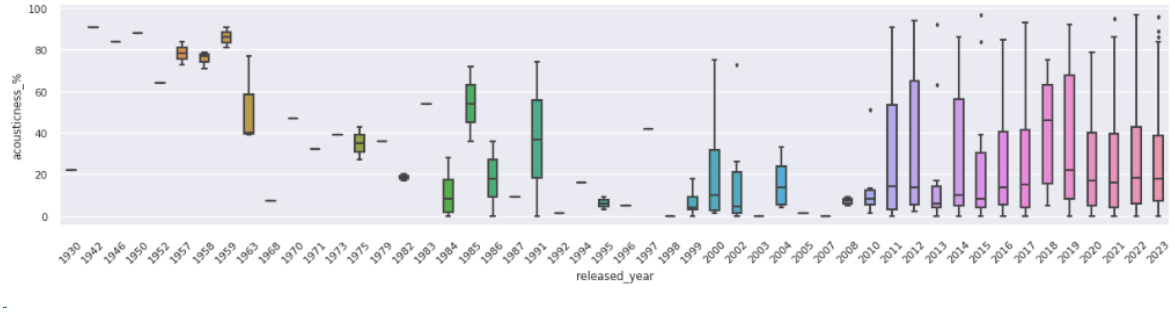
Top tracks from 2011 to 2023 contains a wide range of tracks from energetic to less energetic, but with median values at the 50 to 60% energy percentage.

The median values of the most streamed tracks are 50% or more, with the exception of 1958, 1959, and years with single tracks that made it to the most streamed. This shows that listeners prefer to listen to energetic tracks.



- Acousticness:

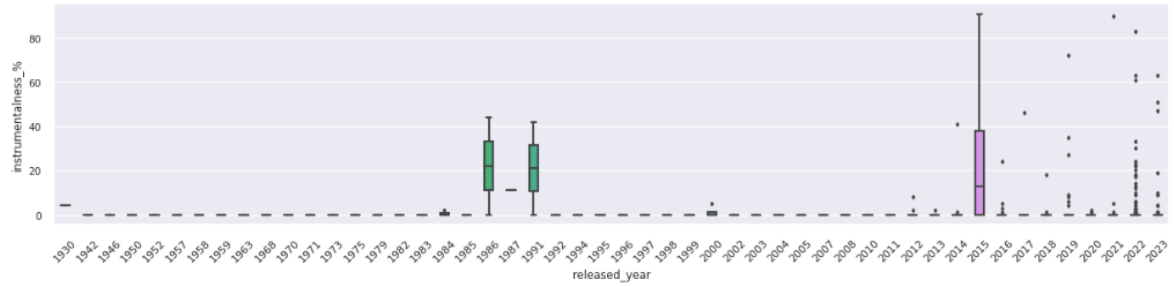
Left-skewed histogram indicates many top-streamed songs are not acoustic. However, the scatterplot reveals that partially acoustic tracks gain more listening time.



- Instrumentalness:

The boxplot shows that majority of the most streamed tracks contains less instrumentalness levels, with some tracks that falls under the instrumental category identifies as outliers.

Tracks released from 1986, 1991, and 2015, however, contains tracks that have considerably high instrumentalness levels, especially in 2015 where the boxplot whiskers reached the highest instrumentalness level.

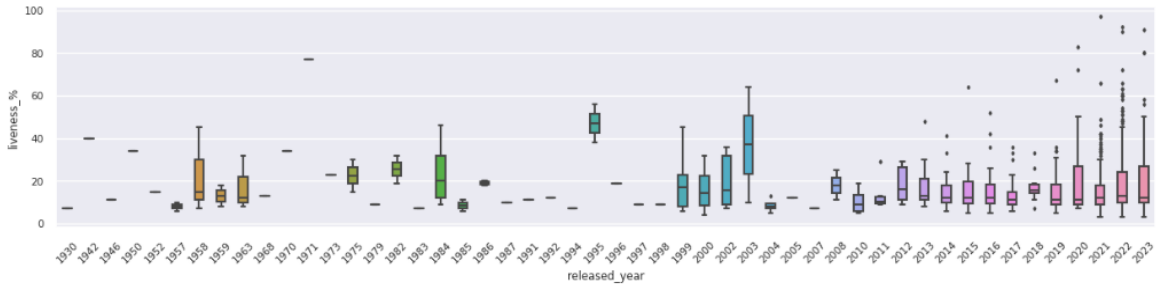


- Liveness:

Preference for recorded tracks is evident in left-skewed histogram and scatterplot.

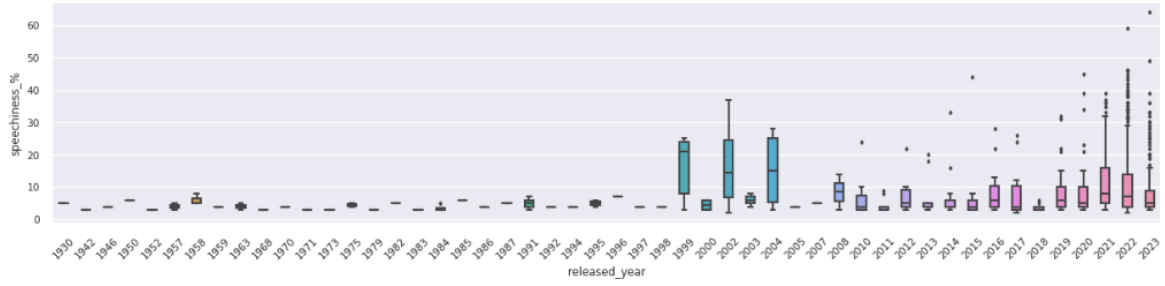
A huge number of top streamed tracks have values of less than 50% liveness, with whiskers and interquartile range falling below 50%.

Tracks which are performed live fall to the outliers, which means than listeners prefer to listen to recorded tracks.

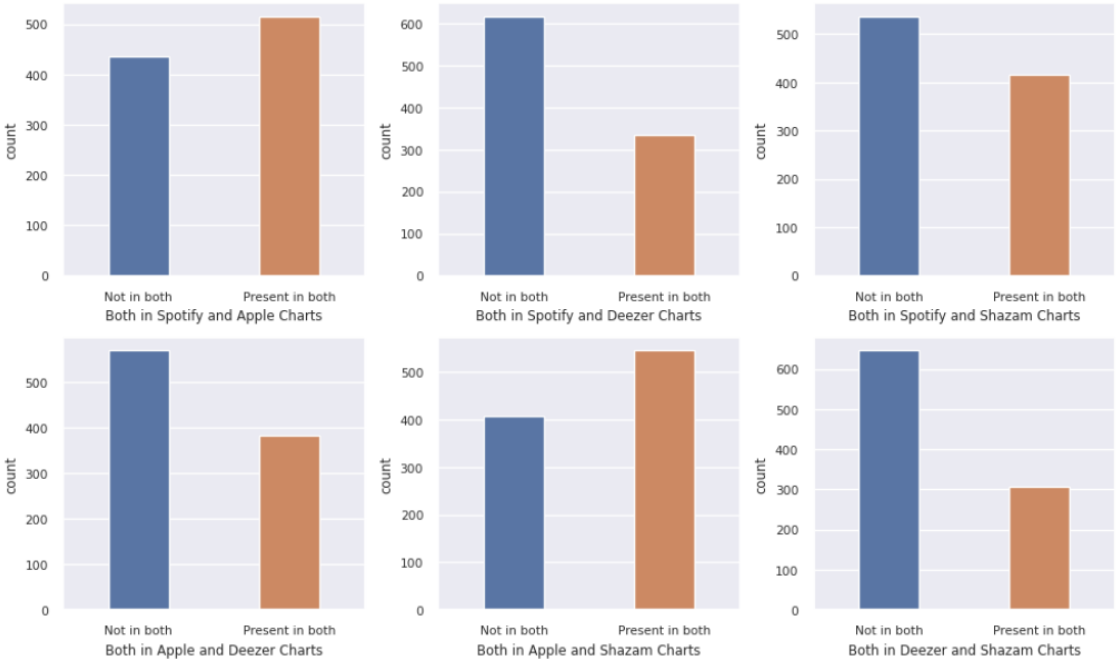


- Speechiness:

Listeners prefer to listen to tracks with less speechiness or spoken words, as shown in the boxplot, where ticks and interquartile range fall below 30 to 40%, and outliers are seldom be seen above 50%.

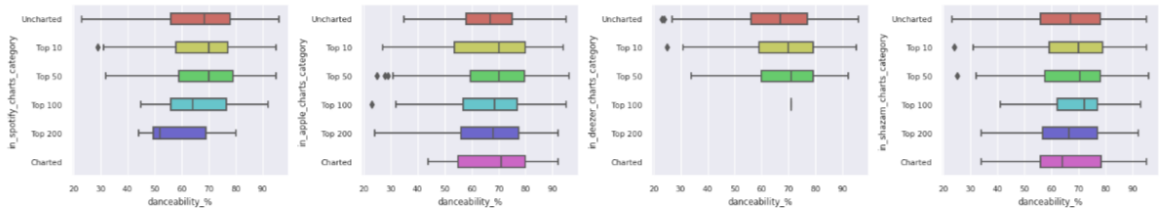


5. Distribution of Top Streamed Tracks in other Platforms

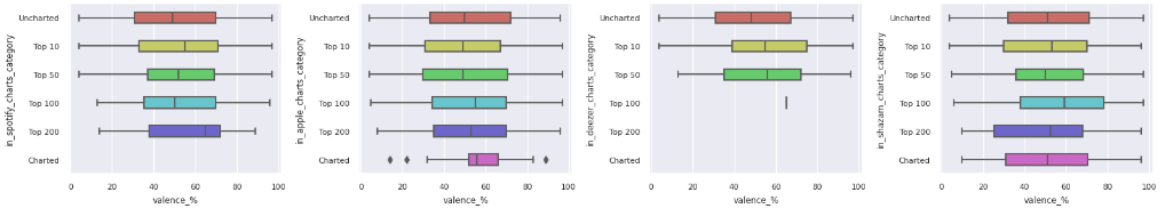


5.1 Chart Performance based on Track Metrics

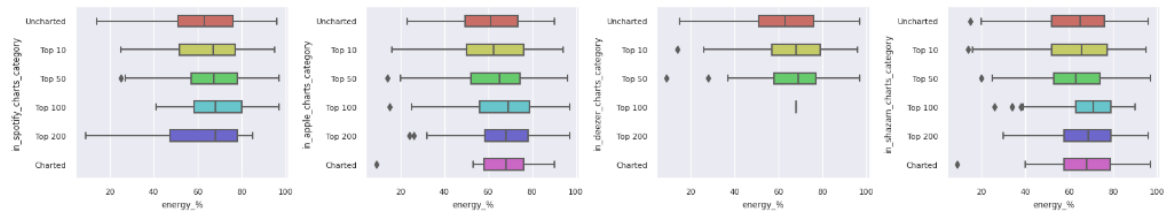
1. Danceability



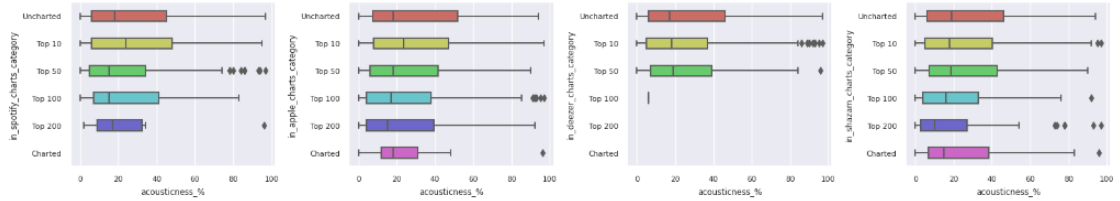
2. **Valence**



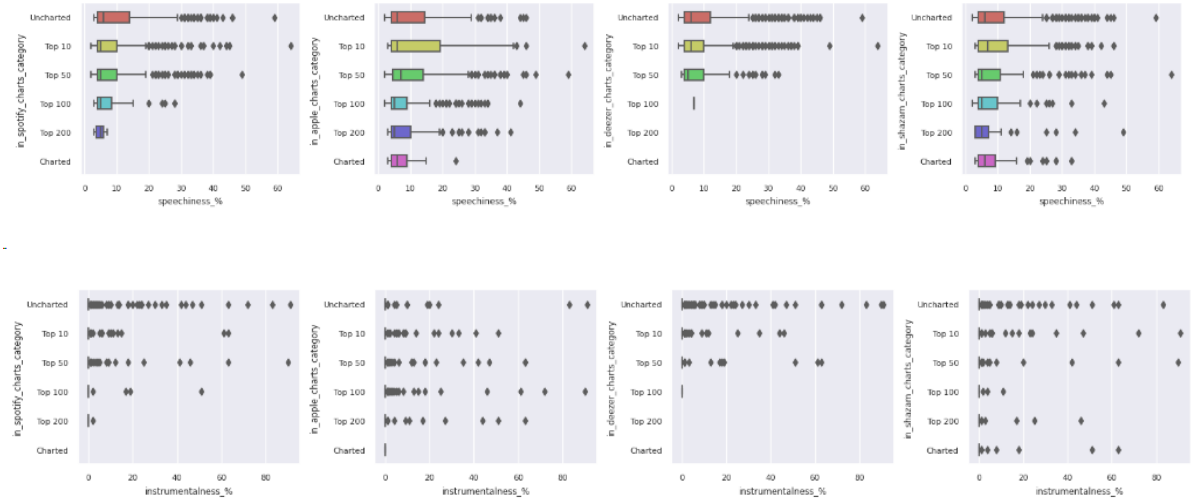
3. Energy



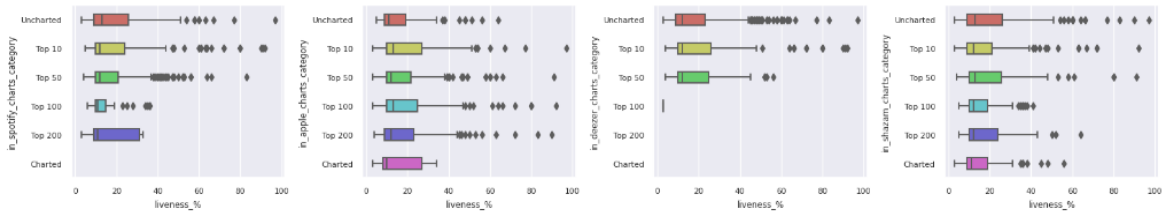
4. Acousticness



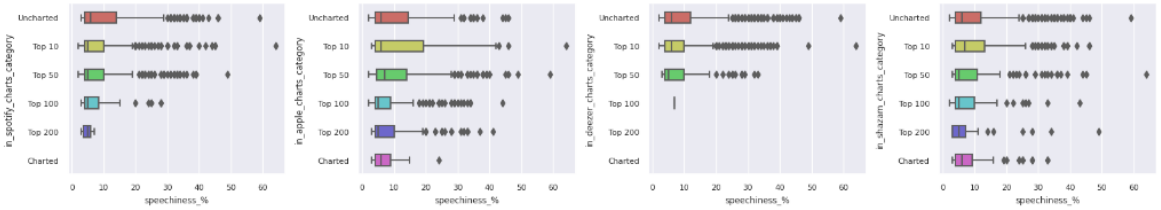
5. Instrumentalness



6. Liveness



7. Speechiness



6.Conclusion

The analysis of the Most Streamed Spotify Songs in 2023 reveals several key observations:

Chart Placement and Streaming Time:

- Most top-streamed tracks did not place in charts, and they exhibit shorter streaming durations compared to charting tracks. This is often indicative of newer tracks that haven't charted yet but are streamed heavily.

Playlist Presence:

- A significant number of top-streamed tracks are not present in many playlists across different platforms. This could be attributed to their recent release, indicating that tracks are streamed heavily even before being added to playlists.

Musical Characteristics:

- Top-streamed tracks include a variety of moods but generally consist of upbeat songs. Listeners prefer tracks with more singing than speech, acoustic elements, and instrumentals.

Release Date Impact:

- The release date significantly influences a track's total streaming time, with recently released tracks accumulating more streaming time compared to older tracks.

BPM and Key Influence:

- The track's BPM and key appear to be less influential factors in determining a track's total streaming time.

Correlation Insights:

- The dataset lacks highly correlated variables, suggesting that the provided variables alone may not be sufficient to predict a track's streaming success. Further analysis with additional historical records may reveal stronger correlations.

Summary: The analysis highlights the complex interplay of various factors contributing to a song's streaming success. While certain trends are observed, predicting a track's streaming time based solely on the available variables proves challenging. This underscores the need for more comprehensive data to unveil deeper insights into the dynamics of music streaming.